FLUXO DO TRABALHO COM DADOS: DO ZERO À PRÁTICA

ESCOLA DE DADOS

OPEN KNOWLEDGE BRASIL

Fluxo de trabalho com dados - Do zero à prática

Escola de Dados

Resumo

Esta publicação é um guia prático destinado a profissionais e estudantes interessados em trabalhar com dados no campo da comunicação, em especial no jornalismo e na produção de conteúdos para organizações da sociedade civil. O guia é baseado no fluxo de trabalho com dados (data pipeline), uma metodologia desenvolvida pela rede da Escola de Dados internacionalmente, que aborda todas etapas do trabalho, da definição das questões à visualização dos dados.

Sumário

Créditos

Esta é a primeira edição do guia **Fluxo de trabalho com dados - do zero à prática**, publicada em agosto de 2020. O ebook foi produzido graças aos recursos obtidos com o primeiro financiamento coletivo da Escola de Dados, o programa educacional da Open Knowledge Brasil, que alcançou sua meta e contou com o apoio de 282 pessoas. Feedbacks, sugestões e críticas são bem vindas e podem ser enviadas por email: escoladedados@ok.org.br.

Organização

Adriano Belisário

Revisão e edição de texto

Adriano Belisário

Marília Gehrke

Editoração

Adriano Belisário

Capa

Isis Reis

Autoria dos capítulos

Os capítulos do livro possuem a seguinte autoria:

Introdução: Adriano Belisário, Rodrigo Menegat e Marília Gehrke

Defina: Adriano Belisário

Obtenha: Adriano Belisário

Cheque: Marina Gama Cubas e Rodrigo Menegat

Limpe: Marina Gama Cubas e Adriano Belisário

Analise: Rodrigo Menegat e Marina Gama Cubas

Visualize: Rodrigo Menegat

Autores(as) do livro

Adriano Belisário

Coordenador da Escola de Dados e jornalista com mais de 10 anos de experiência com projetos na área de comunicação e tecnologias livres. É autor e organizador de diversas publicações, tutoriais e textos sobre o assunto, como o livro 'Copyfight - Pirataria & Cultura Livre' (2012) e dezenas de conteúdos publicados regularmente no site da Escola de Dados. Atualmente, também é pesquisador associado ao Medialab da Universidade Federal do Rio de Janeiro, onde desenvolve investigações baseadas em dados e técnicas de Open Source Intelligence (OSINT). Tem mestrado em Comunicação na mesma universidade e foi colaborador de veículos de jornalismo investigativo, como a Agência Pública, e organizações internacionais, como a UNESCO e Witness.

Marília Gehrke

Doutoranda em Comunicação e Informação pela Universidade Federal do Rio Grande do Sul (UFRGS), onde também fez mestrado. É jornalista graduada pela Universidade de Santa Cruz do Sul (Unisc), e na mesma instituição realizou aperfeiçoamento em Assessoria em Comunicação Política. Trabalhou como repórter multimídia no interior do RS e pesquisa temas como jornalismo guiado por dados, transparência e desinformação. Integra o Núcleo de Pesquisa em Jornalismo (Nupejor/UFRGS) e o Laboratório de Popularização da Ciência (Popcicom). Auxilia na gestão de cursos da Escola de Dados e integra a rede de pessoas embaixadoras da Open Knowledge. É uma das organizadoras do Open Data Day de Porto Alegre.

Marina Gama Cubas

Marina Gama Cubas é jornalista de dados. Colabora com investigações e análises de dados em projetos especiais do Aos Fatos. Trabalhou no DeltaFolha, editoria de dados do jornal Folha de S. Paulo, e no Estadão. No exterior, passou pelas seções de dados dos jornais El Mundo, na Espanha, e La Nación, na Argentina. Possui master em jornalismo investigativo e de dados pela Universidade Rey Juan Carlos, de Madri, e graduações em Comunicação Social e Letras.

Rodrigo Menegat

Rodrigo Menegat trabalha como jornalista na editoria de infografia digital do Estadão, onde produz reportagens e visualizações feitas a partir da exploração de bases de dados. Antes disso, trabalhou como redator e repórter na Folha de S.Paulo e teve conteúdo publicado em veículos como The Intercept e Agência Pública. É formado em jornalismo pela Universidade Estadual de Ponta Grossa, no Paraná, e fez uma especialização em jornalismo de dados na Universidade de Columbia, em Nova York.

Equipe - Escola de Dados

Direção-executiva da Open Knowledge Brasil

Fernanda Campagnucci

Coordenação da Escola de Dados

Adriano Belisário

Comunicação

Isis Reis

Apoio pedagógico

Marília Gehrke

Estágio

Edilaine dos Santos

Apresentação

A Escola de Dados é o programa da Open Knowledge Brasil que tem o objetivo de ajudar pessoas e organizações a transformarem dados em histórias, evidências e mudanças sociais, compartilhando conhecimento, encontros, plataformas e materiais didáticos gratuitos. De 2013 até 2020, já envolvemos mais de 18 mil pessoas em nossas atividades, em todas as regiões do Brasil, e realizamos quatro edições do Coda.Br, a maior conferência de jornalismo de dados e métodos digitais da América Latina. Reunindo esta experiência e aproveitando o acúmulo de diversas colaborações com profissionais externos, preparamos esta publicação, que foi viabilizada com o apoio de mais de 280 pessoas em nossa primeira campanha de financiamento coletivo.

Esta publicação é um guia prático destinado a profissionais e estudantes interessados em trabalhar com dados no campo da comunicação, em especial no jornalismo. **O guia é baseado no fluxo de trabalho com dados (*data pipeline*), uma metodologia desenvolvida pela rede da Escola de Dados internacionalmente, que aborda todas etapas do trabalho**.

O fluxo tem como ponto de partida a definição de uma pergunta ou hipótese, seguida da busca, localização e obtenção dos dados. O passo subsequente envolve a verificação e a limpeza das informações nas bases de dados, de modo que fiquem preparadas para a etapa de análise e, finalmente, para a visualização do conteúdo. Neste material, todas essas etapas serão detalhadas. Por isso, se você é iniciante na área, não se preocupe - este livro foi pensado para você.

Apesar de o apresentarmos de forma linear, tenha em mente que na prática estas etapas em geral costumam ser mais intercaladas. Às vezes, você chegará na etapa de visualização e perceberá que existe um erro nos dados e precisará refazer esta etapa. Como veremos, de fato, a checagem é uma etapa que deve ser transversal a todo o trabalho com dados. Do mesmo modo, um insight obtido com uma visualização pode gerar uma nova questão, que por sua vez irá demandar uma nova coleta e análise. Ainda assim, **a representação do trabalho com dados como um fluxo (pipeline) é útil, especialmente para quem está começando**.

Assim, inicialmente, apresentamos uma breve retrospectiva da comunicação baseada em dados. Falaremos sobre o jornalismo de dados e seus precursores, cujas raízes estão assentadas sobre o método científico, ou seja, o princípio da replicabilidade e a transparência na comunicação das informações. Apresentamos, ainda, as principais tecnologias utilizadas nessas práticas atualmente e uma introdu-

Figura 1: Foto com parte das pessoas participantes da quarta edição do Coda.Br, em 2019

ção às linguagens de programação mais relevantes da área.

No capítulo sobre **definição**, vamos discutir como formular uma boa pergunta, refletir sobre temas de interesse público e conhecer os princípios para identificar um bom escopo de investigação. Também veremos que os dados vão muito além do formato de tabelas e indicam não só a quantificação de fenômenos.

Em seguida, falaremos sobre a **obtenção dos dados**. Veremos as diferentes técnicas utilizadas, ferramentas úteis e problemas recorrentes. Nesta etapa, também aprenderemos como usar a Lei de Acesso à Informação para obter dados inéditos e conheceremos os principais formatos de dados abertos.

Posteriormente, trataremos da **checagem e limpeza dos dados**. Veremos que dados podem ser imperfeitos, assim como nós, e aprenderemos como podemos biografar os dados para, assim, identificar suas limitações e possíveis vieses. Na etapa de limpeza, revisaremos ferramentas importantes para deixar os dados prontos para serem analisados e destacaremos alguns pontos que merecem atenção especial. Essas etapas podem parecer menos empolgantes que as demais, mas acredite: elas são essenciais e você provavelmente irá passar bastante tempo nelas ao realizar seus trabalhos.

Depois, no capítulo de **análise de dados**, falaremos sobre operações matemáticas e estatísticas básicas, que são frequentemente empregadas. Também conheceremos operações de análise de dados fundamentais, que são utilizadas em todas as ferramentas e tecnologias, tais como filtragem, ordenação e agrupamento de registros.

No capítulo de **visualização de dados**, veremos as atenções que você deve ter na hora de apresentar o resultado do seu trabalho usando gráficos. Ao mesmo tempo em que esta técnica pode ampliar ampliar nossas capacidades cognitivas, se mal conduzida, pode ser fonte de erros e enganos. Por isso, você irá aprender alguns princípios práticos para elaborar gráficos e conhecerá ferramentas úteis para colocar a mão na massa.

Por fim, na conclusão, você encontra dicas de como seguir sua aprendizagem. Este livro funciona como um guia, e, por meio de inúmeros exemplos, poderá ser consultado inclusive durante a prática jornalística. Ao final desta publicação, separamos referências e dicas de comunidades e plataformas para você se conectar e continuar aprendendo.

Figura 2: Data pipeline ou o fluxo de trabalho com dados

Introdução

Durante a pandemia da Covid-19, gráficos representando o aumento do número de casos ao longo do tempo, por meio de curvas de distribuição, ganharam as notícias e o debate público. **A visualização de dados se tornou *mainstream*, com o apelo de salvar vidas**. Gráficos são debatidos por especialistas e leigos nas redes sociais, e a interpretação das estatísticas oficiais se revela uma habilidade crucial para a compreensão da pandemia.

A escala do coronavírus e suas implicações não têm precedentes, mas a visualização de dados sobre mortalidade tem um histórico que se confunde com os próprios primórdios da prática de entender e comunicar dados por meio de gráficos. Dois dos trabalhos mais icônicos da história da comunicação baseada em dados estão intimamente ligados à área de saúde e também buscavam, em última instância, salvar vidas. Revisitaremos estas duas visualizações de dados clássicas para em seguida tentar extrair daí algumas lições para o trabalho com dados nos dias de hoje.

Um deles é o gráfico de **Florence Nightingale** sobre as mortes durante a Guerra da Crimeia. Hoje considerada mãe da enfermagem moderna, ela esteve à frente dos cuidados aos soldados ingleses durante as batalhas e também compilou estatísticas sobre a causa de mortalidade dos combatentes.

Por meio de um gráfico elaborado em 1858, hoje considerado um clássico, ela mostrou que as mortes em combate eram minoritárias, se comparadas àquelas causadas por doenças preveníveis, como aquelas provocadas por má condições de higienes no hospital. Esta forma de representar os dados é conhecida como **diagrama de área polar** (ou *coxcomb,* em inglês) e é utilizada até os dias atuais. Na imagem, é possível ver como a proporção de mortes por causas evitáveis era constantemente maior ao longo dos meses.

Além de diversas guerras e batalhas, o século XIX teve também seis epidemias de cólera, que mataram milhões de pessoas em diversos continentes. E foi na mesma Inglaterra dos anos 1850 que o médico **John Snow** decidiu levantar dados "geolocalizados" sobre a doença, mapeando os casos e óbitos. Assim, ele elaborou o chamado **mapa da cólera**, que representou os casos da doença como pontos no mapa. A partir da análise dos padrões espaciais e uma intensiva investigação por meio de entrevistas, descobriu uma bomba d'água como vetor de disseminação da cólera no bairro de Soho. À época, Londres experimentava uma expansão urbana somada ao crescimento populacional em condições precárias de moradia e saneamento.

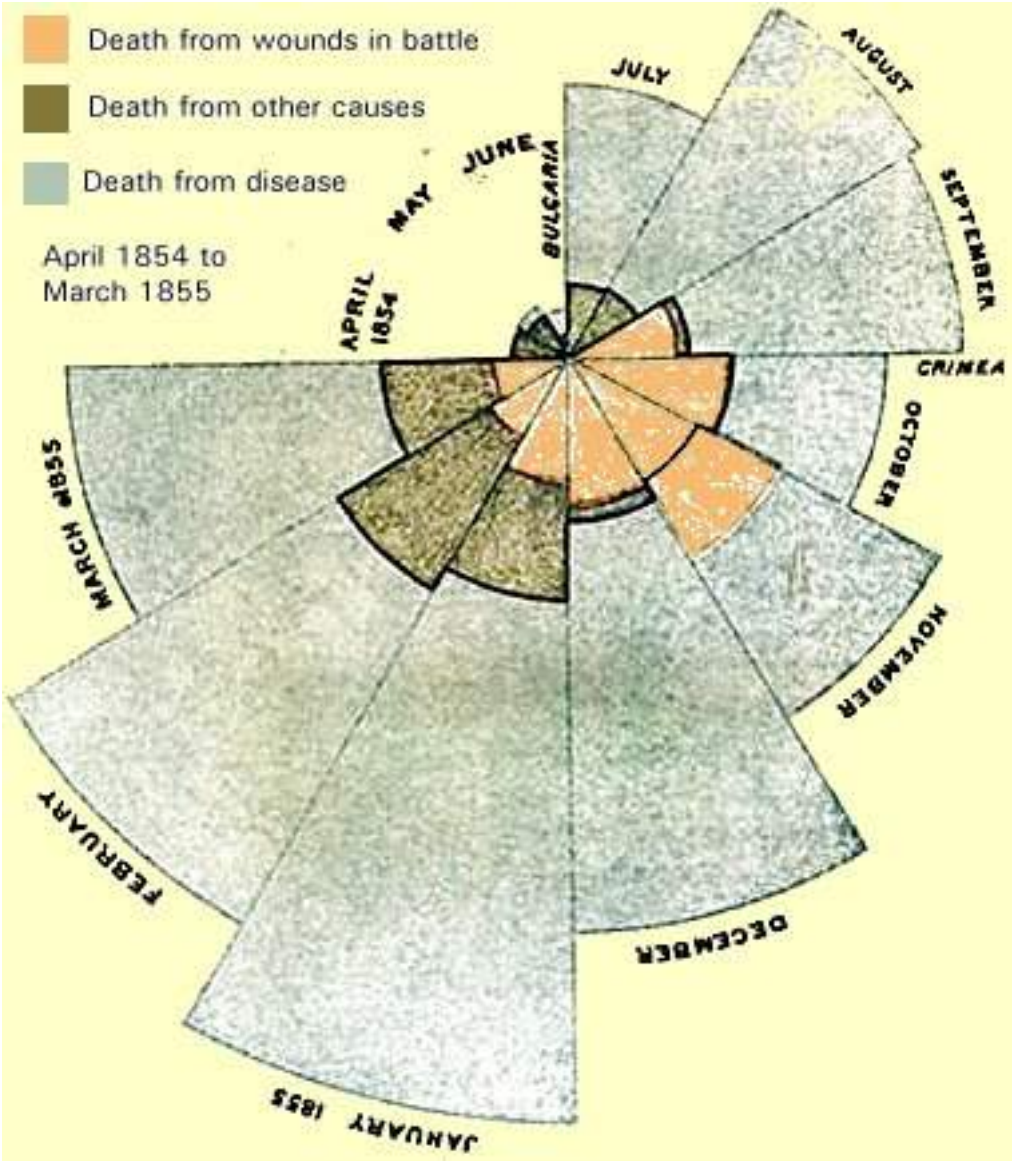


Figura 3: No gráfico, cada fatia representa um mês e a área coberta pelas cores indica as causas de mortalidade: em laranja, mortes em batalha; em azul, por doenças; em verde, outras causas. Fonte: *University of Houston.*

Com isso, John Snow reforçou sua teoria de que a água poderia ser um dos vetores de contaminação, na contramão da chamada teoria miasmática, que era predominante à época e supunha a transmissão por meio do ar. A história mostrou que ele estava certo. Hoje, Snow é considerado um dos nomes fundadores da epidemiologia.

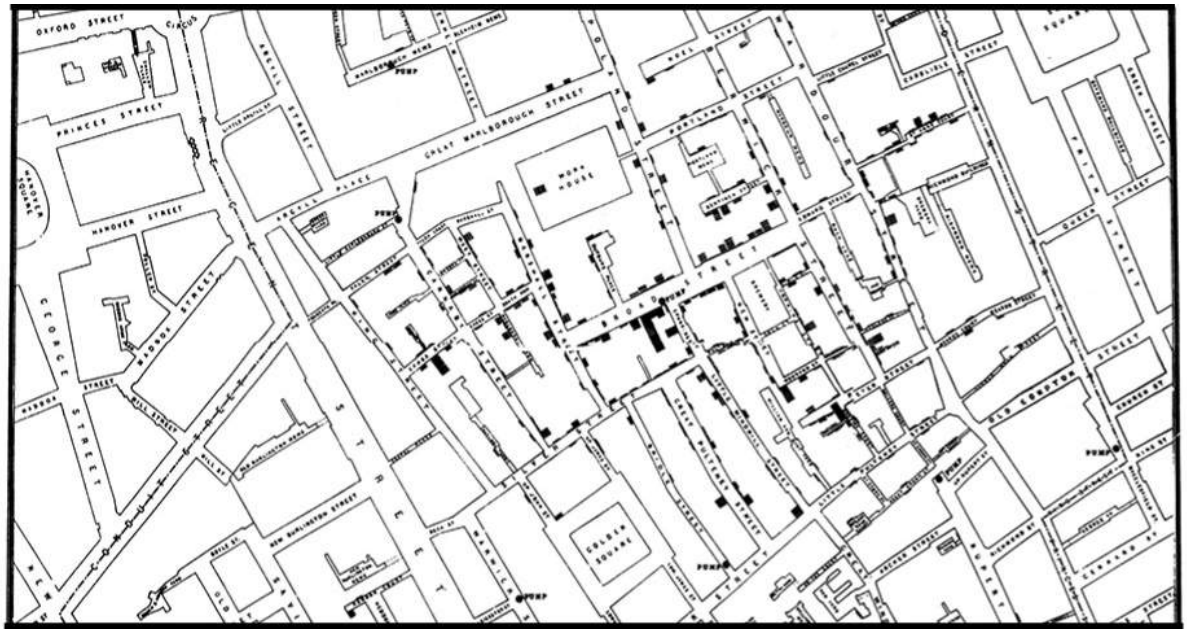


Figura 4: Mapa da cólera de John Snow na Inglaterra (1854) - Fonte: Wikimedia Commons

Retomamos os exemplos de Snow e Nightingale por uma dupla razão: primeiro, para desfazer um aura de novidade ("hype"), ainda que este tema fique mais evidente em épocas de crise em tópicos de interesse público, como a saúde. Lidar com dados não é exatamente algo recente na experiência humana, tampouco teve início com os computadores digitais. Na verdade, no fundo, **a utilização de dados é praticamente indissociável da experiência humana tal como a conhecemos**.

Os exemplos podem ser diversos. Podem ser até pré-históricos, como quando as contagens eram feitas em artefatos rudimentares, como o osso de Ishango, objeto datado do paleolítico que foi encontrado na África Central e hoje é considerado uma das primeiras formas de armazenamento de dados. Ou ainda o ábaco, inventado durante a Antiguidade no Oriente Médio para facilitar cálculos, e a máquina de Anticítera, um computador analógico para fazer previsões sobre eclipses e as posições dos astros, que foi inventado pelos gregos. Sem dúvida, as tecnologias digitais de armazenamento de dados atuais abrem horizontes novos e bastante amplos, mas é importante lembrar que lidamos com a quantificação da realidade cotidianamente e desde sempre.

Quanto tempo irei levar para chegar ao meu compromisso? Quantas xícaras de farinha preciso colocar no bolo? Qual a diferença de preço entre este e aquele produto? Quantos quilos eu ganhei depois do Natal? Como podemos dividir uma conta igualmente entre as pessoas? São perguntas banais, mas ilustram como sempre lidamos com dados no cotidiano em algum nível, mesmo sem refletir muito sobre isso.

Figura 5: Mapa do novo coronavírus em São Paulo (2020) - Fonte: Labcidade

Assim, por um lado, isso deixa claro que qualquer pessoa pode vir a trabalhar com dados. Esta prática não é e não precisa ser restrita a especialistas ou pessoas ligadas às chamadas "áreas de exatas" (matemática, estatística, desenvolvimento de software, etc). Por outro, a segunda razão para retomarmos o exemplo de Nightingale e Snow é para ressaltar um cuidado importante relacionado à comunicação baseada em dados: **qualquer pessoa poder trabalhar com dados, mas isso não significa que eles devam ser tratados de qualquer forma**.

Há alguns ensinamentos que podemos extrair das experiências do século XIX que foram mencionadas, particularmente no que diz respeito ao método científico. Esse tipo de procedimento nos ajuda a ter uma **leitura crítica dos dados para utilizá-los como uma ferramenta de conhecimento - e não como artifício retórico para reforçar crenças e opiniões pré-existentes**.

A ciência pressupõe o acúmulo de conhecimento. Cada cientista estuda uma pequena parte de um tema, ocupando-se de um problema muito específico. A partir de muita leitura - a revisão bibliográfica serve para conhecer o estado da arte de determinada área, ou seja, aquilo que outros pesquisadores já descreveram - e observação da realidade, o cientista está apto a formular e testar sua hipótese.

A descrição detalhada do método é fundamental para que a comunidade científica seja capaz de validar os resultados de uma pesquisa e eventualmente replicá-la integral ou parcialmente. Na academia, os artigos são submetidos à revisão cega por pares - em outras palavras, significa dizer que outros cientistas leram e avaliaram a consistência teórica e metodológica de um trabalho. Em projetos mais

complexos e longos, como teses e dissertações, o mesmo ocorre a partir de uma banca avaliadora.

No imaginário social, o estereótipo do cientista geralmente é formado por uma pessoa de jaleco branco que trabalha em um laboratório conduzindo experimentos químicos. A realidade, no entanto, sugere uma fotografia mais ampla: cientistas também usam roupas normais e trabalham apenas no computador em salas convencionais nas universidades ou mesmo de casa. Assim também se faz ciência, e com métodos igualmente rigorosos.

Neste guia, não iremos abordar a área de ciência de dados em si, ainda que muitos dos conceitos e técnicas que iremos abordar aqui sejam emprestados desta área. Nosso enfoque será a comunicação baseada em dados, especialmente na área do jornalismo. **Da ciência, iremos tomar emprestada principalmente a inspiração no método, para que possamos conduzir trabalhos mais sólidos e robustos**.

Embora o jornalismo não seja uma ciência propriamente dita, visto que trabalha com a ideia de pauta e de um conhecimento menos formal e sistemático, pode aproximar-se do rigor metodológico à medida que abre suas fontes, seus dados e o passo a passo empreendido para obter um resultado. É neste contexto que a ideia de replicabilidade entra em cena e pode ser apropriada tanto pelos leitores quanto por outros jornalistas. Ao longo deste guia, veremos como podemos nos inspirar no método científico para produzir conteúdos baseados em dados com mais rigor e qualidade.

Em tempos de debates acalorados na internet, é comum que as pessoas busquem os dados apenas quando querem legitimar seus argumentos. Há quem fale até em "torturar os dados" para que confessem o que o analista deseja. Aqui, porém, trabalharemos a perspectiva oposta.

É importante que você **seja cético em relação aos seus dados**. A respeito deste tema, vale conferir o artigo *Sou uma cientista de dados cética quanto aos dados* de Andrea Jones-Rooy, professora de Ciência de Dados na New York University. Ao trabalhar com investigações baseadas em dados, seu objetivo não é encontrar números que confirmem sua hipótese ou aquilo que você (acha que) sabe. Ao contrário, você deve olhar os dados sob uma perspectiva crítica, ao invés de adotá-los como uma representação objetiva da realidade.

Em especial se você não é um especialista no tema que está abordando, provavelmente, sua autocrítica talvez não seja capaz de avaliar todas as possíveis nuances dos dados ou erros de uma análise. Por isso, **para fazer sentido de tabelas ou dados, é importante também conhecer como eles foram produzidos e, se possível, entrevistar ou conversar com diferentes pessoas que entendam mais que você sobre o tema em questão**, seja o novo coronavírus, a performance de um clube em um campeonato esportivo ou mesmo dados sobre o mercado de trabalho.

No fim das contas, como na ciência, a **transparência também é fundamental na comunicação**. Então, sempre que possível, disponibilize os dados utilizados ou a referência de onde eles se encontram. Dessa forma, as pessoas poderão refazer o passo a passo da sua análise - remetendo ao princípio da replicabilidade científica - se assim desejar. Aprender a programar também ajuda, pois assim você consegue documentar e compartilhar todas as etapas de processamento, análise e até visualização dos

dados. Além de tornar seu trabalho mais confiável, a publicação de scripts e programas com código-aberto permite que outras pessoas possam revisar o trabalho feito, identificar possíveis erros (sim, isso é ótimo!), entender como ele foi feito e aprender com a sua experiência.

Visto que o jornalismo envolve a produção de conhecimento a partir da observação da realidade, semelhante ao que ocorre na ciência, desenvolveram-se diversas teorias e práticas para aproximar estes dois campos ao longo do século XX. O primeiro grande esforço sistemático neste sentido foi feito por Philip Meyer, criador do chamado **"jornalismo de precisão"**.

Meyer se destacou por aplicar métodos das ciências sociais no jornalismo. No ano de 1967, em meio à luta do movimento negro por direitos civis nos EUA, durante a cobertura de protestos em Detroit, ele decidiu aplicar técnicas de pesquisas amostrais (survey) para entender a opinião pública e as causas das manifestações a partir da perspectiva da população local. Algo com intuito parecido ao que um repórter ou uma organização faz ao coletar depoimentos de pessoas sobre determinado tema, mas com o objetivo de refletir a opinião coletiva de um grupo social em vez de opiniões individuais.

Ainda nos anos 1960, Philip Meyer e sua equipe criaram uma amostra das residências em regiões de manifestações, entrevistaram 437 pessoas e utilizaram um computador para testar hipóteses e processar os resultados, culminando nesta apresentação. Na contramão de discursos que circulavam à època, inclusive na própria imprensa, a pesquisa mostrou que os manifestantes não eram pessoas de baixa educação e que apoiavam atos violentos. Formação ou renda não eram fatores distintivos entre aqueles que participavam dos atos, mas o desemprego, sim. Além disso, os participantes da pesquisa apontaram possíveis causas para os atos violentos, que até então não recebiam a devida atenção, como violência policial ou condições de habitação precárias.

Figura 6: Prisioneiros dos protestos de Detroit em 1967 - Fonte: Wystan/Flickr

Ao defender uma guinada do jornalismo em direção à ciência, Meyer acreditava na ampliação do papel do repórter e na menor dependência de fontes oficiais. Dessa forma, como no caso de Detroit, acreditava que os próprios jornalistas deveriam fazer a coleta de dados para responder às questões levantadas no dia a dia da profissão. A cobertura do colapso social do final da década de 1960 rendeu a Meyer e sua equipe o primeiro lugar na categoria de jornalismo local no Prêmio Pulitzer.

Posteriormente, o jornalismo de precisão inspirou práticas como a Reportagem Assistida por Com-

putador (RAC) - ou Computer-Assisted Reporting (CAR), em inglês -, desenvolvida especialmente nos anos de 1980 nos Estados Unidos. Com a disseminação do uso de computadores no ambiente de trabalho dos jornalistas, surgiram novidades nos procedimentos de apuração. Nos primórdios da RAC, o uso de softwares simples para resolver problemas e obter informações já era motivo para comemorar.

No final dos anos de 1990, Meyer sugeriu que se pensasse em outro nome para substituir a RAC. Ele admitiu que o termo ficou datado por apresentar a palavra "computador" em sua composição, destacando uma ferramenta utilizada para a apuração jornalística. Tampouco o "jornalismo de precisão" obteve sucesso como conceito amplamente difundido.

Atualmente, porém, parte das premissas do jornalismo de precisão e da RAC são levadas a cabo por práticas do chamado jornalismo de dados. Dados são abundantes e habilidades básicas de análise são exigidas atualmente em diversos campos da comunicação. **O jornalismo de dados - também conhecido como jornalismo guiado por dados (data-driven journalism) - pode envolver todas as etapas do trabalho, desde a coleta até a visualização das informações**.

Quando publicou o artigo *A fundamental way newspaper sites need to change*, em 2006, o programador Adrian Holovaty defendia que os jornalistas se preocupassem com o uso de informações estruturadas. Isto é, que os repórteres não coletassem apenas informações para uma pauta específica, mas que refletissem sobre sua organização sistemática, facilitando a recuperação e o reuso do que já foi apurado. Em vez de somente divulgar informações sobre um incêndio específico, por exemplo, os jornalistas deveriam coletar aspectos que todos os incêndios têm em comum (localização, número de vítimas, tempo de atendimento, prejuízo etc.) para, com a formatação de um banco de dados, identificar padrões e obter informações mais abrangentes e contextuais.

Embora estes recursos ainda sejam utilizados hoje e as reflexões de Holovaty sigam atuais, planilhas e consultas online em portais da transparência remetem à primeira metade dos anos 2000, em que havia menos conhecimento técnico e disponibilidade de matéria-prima aberta e acessível aos jornalistas de dados. O professor Paul Bradshaw classifica a segunda década (2010-2020) do jornalismo de dados como uma "segunda onda", permeada por um movimento global pela abertura de dados, acesso às APIs e a inserção de algoritmos neste cenário, incluindo inteligência artificial e a internet das coisas.

Bradshaw defende que os jornalistas de dados abram seus códigos e submetam suas fontes e métodos ao escrutínio público, pois assim os cidadãos ganham conhecimento sobre sua própria realidade. **Ao obter informações sobre a sociedade onde vivem, as pessoas passam a ter condições de cobrar o poder público com mais veemência sobre suas demandas, por exemplo. Outro ponto destacado é o potencial que o jornalismo de dados tem em dar voz às minorias ao divulgar informações quantificáveis e que tendem a circular pouco**. Assim, é possível perceber que não só a tecnologia importa ao jornalismo de dados - as demais premissas da prática jornalística, relacionadas à produção de informação qualificada e à fiscalização do poder público, são mantidas e intensificadas.

Como mencionamos, uma das características do jornalismo de dados desta "segunda onda" é o uso de algoritmos de inteligência artificial (IA). Em um texto sobre o assunto, o artigo *How you're feeling*

when machine learning might help, Jeremy B. Merril elenca algumas características úteis deste tipo de tecnologia para jornalistas. Ele desta a possibilidade dessas técnicas processarem um volume enorme de documentos, que seriam impossíveis de ser lido por humanos, em busca de padrões.

Ele cita o exemplo do Atlanta Journal-Constitution, que usou o aprendizado de máquina para identificar relatórios de sanções contra médicos relacionados a abuso sexual. A equipe de dados do jornal desenvolveu cerca de 50 raspadores para diferentes sites, que permitiram a coleta de mais de 100 mil documentos de agências reguladoras. Como era impossível ler e categorizar manualmente todos os relatórios, eles selecionaram uma amostra relevante e fizeram esta categorização manual, escolhendo palavras-chave que identificavam relatos de abuso sexual dos demais casos.

Então, treinaram uma solução de aprendizado de máquina que baseado nessas palavras assinalava quais relatórios eram provavelmente mais interessantes para a investigação. Claro que a solução não nasce pronta. Foi necessário encontrar as melhores palavras-chaves por meio de tentativa e erro, observando os resultados das escolhas utilizadas. Porém, é um exemplo interessante do potencial da inteligência artificial para lidar com uma documentação muito vasta.

Neste sentido, uma solução simples, de código-aberto e interessante tanto para jornalistas quanto para organizações é o DocumentCloud, uma ferramenta que consegue "ler" dados em documentos, tabelas e PDFs para então realizar a extração de entidades automaticamente, identificando menções a pessoas, locais ou datas em cada documento.

Outros casos de uso mencionados por Merril incluem a chamada computação visual e a identificação de dados que possuam um certo padrão. As aplicações são várias e vale ler o artigo completo, caso você tenha interesse no tema. Por ora, o importante é saber que a **inteligência artificial e o aprendizado de máquina permitem converter uma quantidade incontrolável de informações em algo gerenciável, gerando taxonomias e classificações a partir dos dados fornecidos**.

Por ser introdutório, este guia não irá cobrir temas como o uso de algoritmos de IA na comunicação, uma área que começa a despontar agora. Por ora, o importante é saber que, mesmo com IA e independente da tecnologia utilizada, as conclusões de investigações baseadas em dados não devem ser encaradas como uma tradução objetiva da realidade.

Nada dado nos dados

O que são dados, afinal? Essa é uma pergunta importante. Existem várias definições possíveis. De modo geral, podemos dizer que a produção de dados é baseada na quantificação ou estruturação do mundo. Mencionamos "produção" não por acaso: **dados não dão em árvore, ou seja, não são entidades que não existem por si só na natureza**.

Dados são construções humanas para tentar abstrair (ou tentar "representar", com muitas aspas) fenômenos complexos da realidade através de elementos quantificáveis, que por sua vez são passíveis de

serem analisados. Um dado sempre implica um determinado escopo, um recorte. Eles são construídos baseados em conceitos e, por vezes, em preconceitos. Os dados não são neutros ou puramente objetivos, eles traduzem visões de mundo.

Em suma, não há nada dado nos dados. Dados são sempre construídos e, por serem construções humanas, assim como nós, estão sujeitos a erros de diferentes tipos. Devemos ter sempre ter isso em mente ao analisá-los.

Os dados podem ser encarados como uma fonte como outra qualquer, embora sejam "entrevistados" de forma diferente. Assim como você não considera automaticamente verdade tudo que uma pessoa diz, você também não deve "comprar pelo valor de face" tudo que os dados apontam.

Imagine esta cena: um repórter recebe a ligação de uma fonte, ou seja, uma pessoa que tem coisas interessantes para contar e que quer que essas coisas sejam publicadas no jornal. Essa tal fonte pode ser um assessor de imprensa, um político, um dissidente com documentos para vazar, um juiz, um promotor, um pesquisador. De todo modo, ela tem interesses pessoais – que podem ser completamente nobres – embutidos nas informações que divulga.

Do mesmo modo, como identificado pela pesquisadora Marília Gehrke em sua dissertação de mestrado, também podem ser fontes bancos de dados, estudos científicos, pesquisas das mais diversas naturezas, documentos públicos, legislações, sites de redes sociais e outros. A diferença é que essas fontes não podem ser acessadas a partir de pergunta e resposta diretas; elas pressupõem outro tipo de consulta, que não alteram sua natureza documental.

Ainda que apresentem caráter oficial, bancos de dados e outras fontes documentais podem servir para a obtenção de pautas exclusivas, visto que não estão estruturadas em formato de release nem são disparadas para a imprensa de modo geral. Exigem, sim, o esforço do repórter na elaboração de uma hipótese ou pergunta que dará início à investigação e, depois, na coleta e no tratamento das informações. Para valorizar ainda mais o caráter exclusivo, algumas redações optam por construir os seus próprios bancos de dados, realizando análises inéditas.

Em sua tese de doutorado, o pesquisador Marcelo Träsel entrevistou diversos jornalistas brasileiros que usam dados em suas tarefas cotidianas na redação.

Por meio dos procedimentos de apuração - que costumam ter etapas mais bem delineadas na execução do jornalismo de dados em comparação às práticas tradicionais -, esses profissionais acreditam estar mais próximos da noção de objetividade no método jornalístico, cuja concretização leva em conta a correspondência do relato jornalístico à realidade dos fatos.

O poder de mergulhar em números permite ao repórter contestar as narrativas mais óbvias oferecidas por fontes institucionais. Além disso, na contramão de simplesmente reproduzir declarações ou noticiar informações de bastidores, jornalistas de dados se utilizam de recursos (números e planilhas) que podem ser facilmente disponibilizados e verificados por terceiros (leitores e colegas da imprensa, por exemplo).

Como já dissemos, os bancos de dados entram no cardápio de fontes, e devem ser consideradas como tal - vistas de forma crítica, portanto. Para caracterizar seu trabalho como jornalista de dados, Kátia Brembatti, do jornal Gazeta do Povo, fez uma alusão a "entrevistar planilhas" enquanto participante da pesquisa, expressão que se tornaria o título da tese de Träsel.

Brembatti assinou uma série de reportagens conhecida como "Diários Secretos", que expôs um esquema de nomeação de funcionários fantasmas na Assembleia Legislativa do Paraná. Seu trabalho consistiu em organizar uma série de informações em arquivos de Excel para descobrir fatos de interesse público.

O esquema envolvia esconder alguns dos Diários Oficiais que continham referências às nomeações irregulares - e o trabalho de reportagem foi reunir documentos que expunham toda a trama. À medida que os dados eram sistematizados, a equipe envolvida na investigação conseguia respostas para suas perguntas, quase como se estivesse ganhando a confiança de um informante particularmente ruim de lidar.

Pensar nos dados como um entrevistado faz bastante sentido – principalmente porque desmistifica a ideia de que um dado é um fato encerrado, definitivo. Como qualquer entrevistado, um banco de dados tem sua história, seus vieses e um bom motivo para apresentar as informações da forma que está apresentando.

Por esses e outros motivos, há críticas quanto ao termo "dados brutos", já que em sua origem essas informações foram coletadas, selecionadas e disponibilizadas por pessoas. Dados nunca são "brutos", são sempre moldados por uma forma de abstrair uma realidade complexa em termos quantificáveis.

Por isso, é comum que, ao longo de um trabalho, os dados se tornem cada vez menos autoevidentes. Ao aprofundar suficientemente um determinado tema, os dados trazem consigo, muitas vezes, um emaranhado de questões sobre a forma como foram criados. Assim, vale consultar com atenção as notas metodológicas e os dicionários de variáveis que geralmente acompanham as bases de dados. A leitura da documentação é essencial para compreender o escopo e as limitações de determinado conjunto de informações, evitando problemas nas análises.

Trabalhando com dados

Saber lidar com dados abre caminhos para quem deseja ter mais autonomia e ser menos dependente de informes oficiais ou trabalhos de terceiros para analisar informações e obter suas próprias conclusões. O percurso de aprendizagem não é óbvio e muitas vezes não está disponível na grade curricular das universidades. Dessa forma, os profissionais costumam buscar por conta própria conhecimentos relacionados ao "letramento em dados" (data literacy) e outros tópicos.

Se este é o seu primeiro contato com o universo dos dados, não se preocupe. Muitas pessoas não se interessam em trilhar esses caminhos por se sentirem inaptas a trabalhar com números ou programa-

ção. No entanto, isso pode ser enganoso. **A facilidade em trabalhar com estatísticas ou escrever códigos no computador não precisa ser um ponto de partida, mas pode ser algo que você adquire durante o aprendizado**. Se você sente essa dificuldade, tudo que você precisa é ter cuidado na hora de comunicar os resultados das suas análises e pedir ajuda de pessoas com mais conhecimentos, quando necessário.

Além disso, atualmente, é possível realizar análises e visualizações bastante interessantes, fazendo tudo por meio de interfaces gráficas. De fato, em muitos trabalhos, especialmente aqueles que requerem pouca customização, importa menos a solução tecnológica utilizada e mais o conteúdo e o impacto da ação em si.

Existem diversos serviços e programas gratuitos que permitem a criação de trabalhos sofisticados sem a necessidade de escrever uma linha de código, como o Workbench, Tableau ou mesmo o Orange, que trabalha até com algoritmos de inteligência artificial. Neste guia, iremos citar alguns deles e dar algumas dicas de ferramentas, mas nosso foco será te passar uma base que te permitirá lidar com diferentes plataformas e tecnologias.

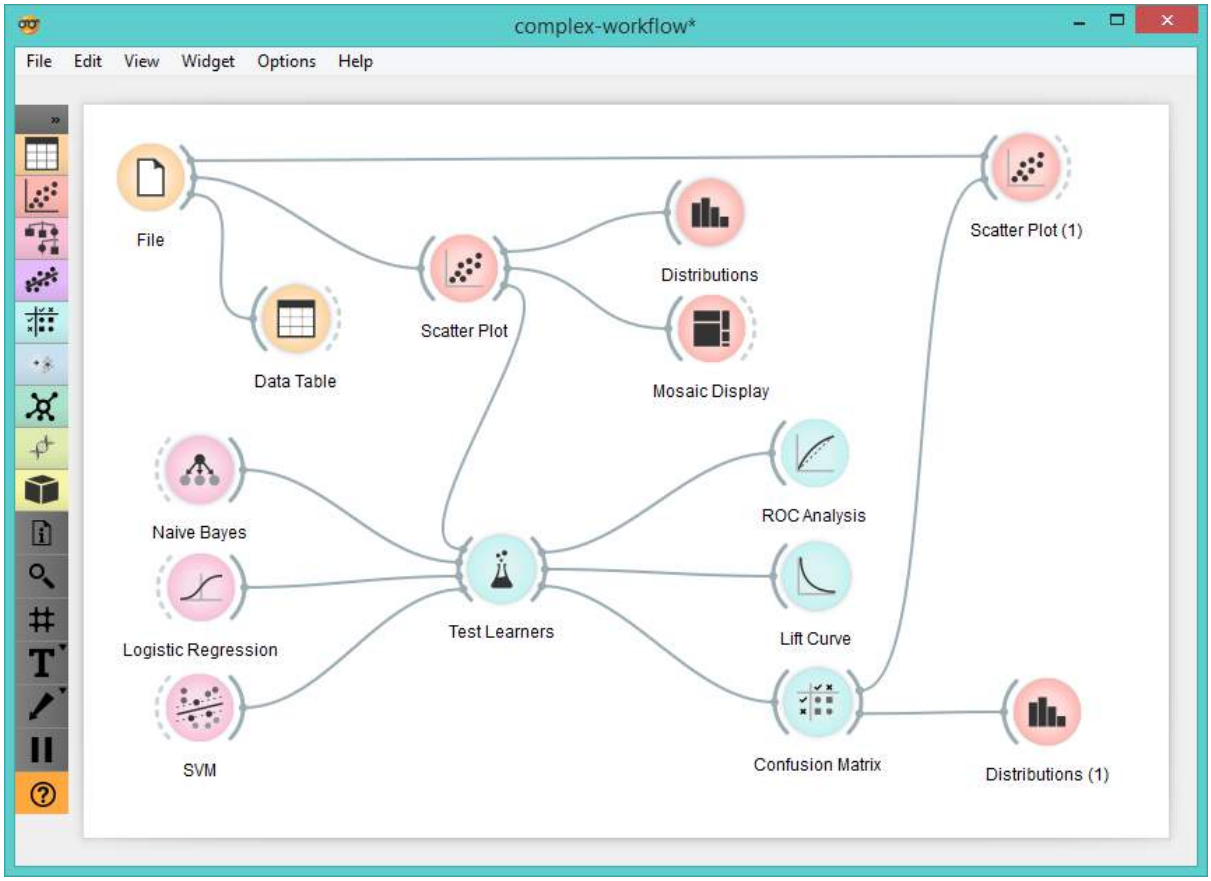


Figura 7: O Orange é um exemplo de programa com interface gráfica para trabalhar com dados, por meio de fluxos visuais. Fonte: Wikimedia Commons

De fato, as possibilidades são muitas, mas, se você está começando, vá com calma. Um passo de cada vez. Antes de se aventurar com visualizações interativas ou modelos estatísticos elaborados, é importante dominar o básico, como o uso de editores de planilha e a criação de gráficos simples.

Muito se ouve falar em *big data*, mas saiba que boas histórias também podem ser extraídas de planilhas de menor porte. A dica para começar a trabalhar com dados é encontrar algum tema que te motive por alguma razão, seja porque você tem paixão pelo assunto, seja porque é uma demanda do seu trabalho ou simplesmente algo que você gostaria aprender mais a fundo.

Uma possível definição é que a expressão big data indica um volume de dados tão grande que um computador só não consegue processar. Fique tranquilo: você provavelmente não precisará disso em suas análises. Para bases muito grandes, você pode importar os dados em um banco de dados SQL, seja no seu próprio computador ou utilizando serviços online como o BigQuery da Google.

A ideia aqui é aprender a trabalhar com dados a partir de um "mini-projeto"; real que te motive de alguma forma, buscando soluções para os problemas reais que enfrentar, em vez de tentar dar conta de todas as possibilidades de aprendizado na área de antemão. Na próxima seção, falaremos sobre uma outra dúvida comum para quem está começando na área: é preciso aprender a programar para trabalhar com dados?

Iniciação à programação

Ainda que não seja mais um obstáculo para começar a trabalhar com dados, saber programar de fato abre muitas possibilidades nessa área. Lidar com a obtenção, processamento, análise ou visualização de dados por meio de códigos te dá liberdade quase total para trabalhar. Você pode realizar as operações diretamente, sem as restrições de programas de terceiros, além de automatizar tarefas, documentar e compartilhar todas etapas de manipulação dos dados.

De modo geral, existem vantagens tanto nas soluções via interface gráfica quanto nas baseadas em linguagens de programação.

Vantagens de usar interfaces gráficas

- Curva de aprendizado baixa: para leigos, é mais fácil aprender a usar programas com interfaces gráficas;
- Poupa tempo: especialmente se você não é fluente em alguma linguagem de programação, muitas vezes, poderá conseguir executar tarefas de forma mais rápida com interface gráfica.

Vantagens de programar

- Alta customização: o software será totalmente adaptável às suas necessidades;
- Automação: é muito fácil automatizar tarefas com programação, mas nem todos programas gráficos suportam isso;
- Escalabilidade: do mesmo modo, é fácil aumentar a escala de um software próprio para que ele lide com um volume maior de dados, por exemplo.

É comum, por exemplo, a criação de “notebooks” (blocos de notas), que trazem não só os dados, mas também os códigos que foram utilizados em determinado trabalho, além de comentários adicionais. Esta prática permite um excelente grau de transparência e detalhamento da conclusão e do método utilizado em trabalhos baseados em dados.

Figura 8: Um notebook sobre análise de texto em ação, mostrando textos, códigos e o resultado dos códigos

É possível trabalhar com dados em qualquer linguagem de programação. Porém, iremos falar aqui especificamente sobre quatro linguagens que se destacam pela sua ampla utilização em projetos baseados em dados atualmente.

A primeira delas é o **SQL**, uma abreviação para linguagem de consulta estruturada (Structured Query Language, em inglês). O SQL é um excelente ponto de partida para quem ter um primeiro contato com códigos para qualquer tipo de profissional, devido ao uso de palavras comuns em inglês para as operações (“select”,“from”, “group by”, etc). Esta linguagem incrivelmente simples permite consultar tabelas enormes, que “travariam” qualquer editor de planilha.

O SQL pode ser uma porta de entrada para o mundo da programação, visto que demanda o uso de lógica computacional. Para começar a usar SQL, você pode experimentar softwares como o **DB Browser for SQLite** e **Dbeaver**. Você pode também utilizar o SQL em conjunto com linguagens como Python e R, que falaremos a seguir, ao lidar com um volume grande de dados. Por ser uma tecnologia voltada especificamente para o armazenamento e processamento de bases de dados, o SQL tem uma performance superior às soluções nativas daquelas linguagens para lidar com muitos registros.

Outra opção é o **JavaScript**, uma linguagem especialmente interessante para designers ou pessoas interessadas em se aprofundar para valer com visualização de dados na web. Ao lado do HTML e do CSS, o JavaScript é uma das linguagens que dão as bases para as páginas web tal como a conhecemos. Mais especificamente, ela é responsável por permitir interatividade e dinamicidade entre os elemen-

tos de uma página, como quando você clica no botão de fechar de uma propaganda e ele some de sua tela.

O JavaScript não costuma ser tão utilizado em projetos de raspagem ou visualização de analisar dados, mas é especialmente interessante para visualização. A biblioteca D3.JS destaca-se como uma das principais referências para trabalhos do tipo.

Porém, se você está interessado em linguagens mais flexíveis, com as quais seja possível realizar com facilidade operações das diversas etapas do trabalho com dados, então, precisa conhecer as duas estrelas dessa área: **Python** e **R**.

Ambas são consideradas o "estado da arte" para projetos de ciência de dados e são amplamente utilizadas não só por cientistas, estatísticos ou desenvolvedores, mas também por jornalistas, pesquisadores das ciências humanas e organizações da sociedade civil.

Cada uma delas conta com milhares de bibliotecas ou pacotes, que nada mais são que softwares escritos nas respectivas linguagens que expandem suas funcionalidades nativas para melhor performar tarefas específicas. A quantidade de pacotes e bibliotecas disponíveis é surpreendente. Suas funcionalidades são muitas: podem servir para estruturar dados a partir de páginas na web, criar visualizações de dados, modelos estatísticos, rotinas de automação e etc.

Para quem está começando, os comandos básicos de ambas são igualmente simples, mas a real complexidade de cada linguagem vai depender do seu objetivo, dos softwares já disponíveis para facilitar esta tarefa e do tempo/conhecimento disponível por você ou sua equipe.

Você pode escolher uma ou aprender ambas, o que te dará bastante flexibilidade. Mas vamos aqui destacar algumas características gerais sobre estas linguagens para você se situar.

De acordo com a revista do Institute of Electrical and Electronics Engineers, Python foi considerada a linguagem mais popular do mundo. Flexível e moderna, permite a utilização em aplicações diversas ao mesmo tempo em que mantém uma curva de aprendizado suave, com códigos simples e legíveis. Python é uma linguagem robusta, que pode ser utilizada não só para trabalhar com dados, como também para distintas aplicações (Web, administração de redes entre outras). O índice de pacotes em Python conta quase 250 mil projetos disponíveis.

Já o R costuma ser preferido por estatísticos e acadêmicos de diversas áreas. A linguagem leva a fama de ter bibliotecas que produzem gráficos e visualizações mais caprichadas e também conta com um rico universo de bibliotecas de softwares para trabalhar com dados. Atualmente, no CRAN, o principal repositório da linguagem, estão listadas mais de 16 mil bibliotecas - isso sem contar outras bibliotecas disponíveis em outra fontes.

Defina

Na prática, tudo começa com os dados ou um tema. Ou seja, em geral, no início de um projeto ou investigação baseada em dados, você irá se encontrar em uma destas duas situações: 1) você tem diante de si dados dos quais precisa extrair respostas ou 2) possui um tema, questão ou hipótese a ser estudada.

Na primeira situação, você já obteve os dados, mas precisa explorá-los para gerar novos conhecimentos ou responder a uma questão definida a priori. Ou seja, seu escopo inicial de certa forma já está definido. O recomendável aqui é compreender pelo menos o básico sobre análise exploratória e análise de dados para conseguir interpretar melhor os resultados - iremos falar sobre isso nos capítulos seguintes.

Já se você estiver abordando um tema bem específico, pode ser necessário também que você estruture sua própria base de dados. Organizar bases de dados inéditas pode ser um diferencial importante na sua proposta. Projetos premiados e reconhecidos no Brasil, como o Monitor da Violência e o Uma Por Uma, seguiram esta linha. Neste guia, não iremos cobrir o processo de publicação de dados inéditos, porém, de todo modo, as demais etapas do fluxo de trabalho com dados (para além da obtenção) seguem sendo úteis.

Mas nem sempre você terá os dados de antemão ou poderá criá-los. Especialmente se você está começando, talvez você tenha apenas uma questão ou um tema que gostaria de pesquisar. Aqui, vale ressaltar que qualquer assunto pode ser investigado a partir de dados. Para te inspirar, vamos listar aqui alguns temas e trabalhos inspiradores, que vão além dos assuntos mais óbvios (indicadores oficiais do governo, estatísticas de esportes, dados de mercado, etc):

- **Nutrição**: se você se preocupa com sua alimentação, você pode compilar e fazer análises sobre os dados da composição nutricional dos alimentos de sua dieta ou que estão na sua geladeira, por exemplo;
- **Cultura**: a produção musical e cinematográfica estão entre as áreas da cultura com bases de dados mais bem estruturadas (vide o Spotify ou OMDB) e você pode tirar proveito disso para investigar seu tipo de música ou filme favorito, por exemplo;
- **Moda**: neste excelente trabalho do The Pudding, eles investigaram como a indústria da moda lida com as questões de gênero ao desenhar o tamanho dos bolsos nas roupas, a partir de dados;

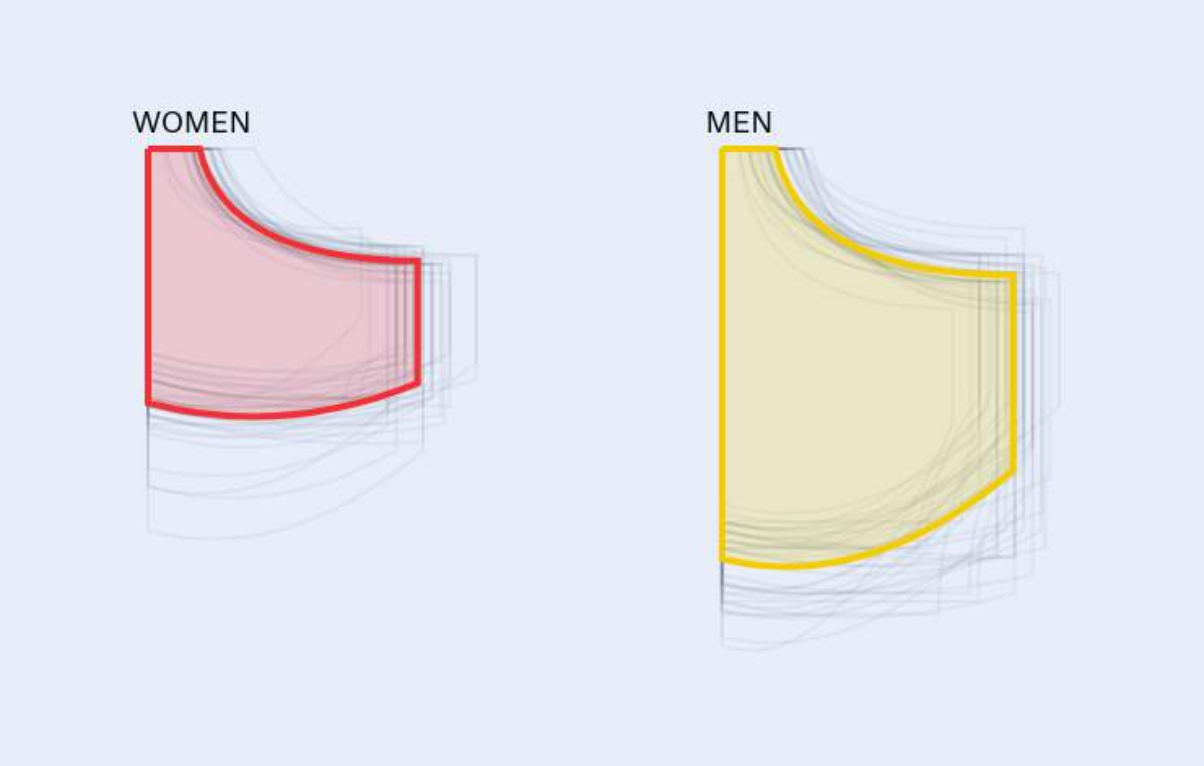


Figura 9: Para investigar o tema, os jornalistas pegaram o tamanho do bolso da frente de 80 calças de ambos os sexos para então calcular o tamanho médio. Fonte: The Pudding

- **Transporte por bicicletas**: no projeto Radmesser, o jornal alemão Der Tagesspiegel investigou a segurança no trânsito para os ciclistas em Berlim, criando um dispositivo que media a distância entre os carros e as bicicletas e distribuindo entre seus leitores, que coletaram dados e compartilharam com o jornal. O trabalho recebeu o prêmio de inovação do Data Journalism Awards em 2019.

Enfim, você consegue usar dados para abordar qualquer assunto. Depois de encontrar um tema de seu interesse, você precisa gerar algumas perguntas iniciais que possam ser investigadas e respondidas com auxílio de dados. A definição desta "pauta" inicial é bastante importante para o desenrolar da pesquisa. Vejamos, então, como criar boas questões ao iniciar um trabalho com dados.

Criando boas perguntas

O exercício de começar a investigação com uma pergunta é exatamente o oposto de já ter uma certeza e ir atrás dos dados que a confirmam. Reflita inicialmente sobre qual é a sua questão motivadora, esmiuçando-a o máximo possível. Por que essa é uma questão relevante? Quando/como/onde ela ocorre? Quem são as pessoas ou atores envolvidos? Qual é a definição exata de cada termo implicado na sua pergunta?

Vejamos tudo isso aplicado a um exemplo prático. Digamos que você queira investigar o transporte

público rodoviário em sua cidade. Uma primeira pergunta poderia ser: quais são as empresas que atuam neste mercado?

Esta pergunta, que parece simples e objetiva, pode ser ainda melhor detalhada.

Primeiro, o que entendemos como empresas atuante no mercado rodoviário da cidade? São aquelas que possuem concessões públicas com a Prefeitura no presente momento? Ou com concessões vencidas também? Os ônibus intermunicipais serão levados em conta? E se a firma dona da concessão for controlada por outra empresa, esta segunda empresa deve entrar em nossa listagem? O interesse é apenas saber o nome das empresas ou gostaríamos de obter outras informações, tais como o nome dos donos ou data de criação das mesmas?

Enquanto você formula e detalha sua pergunta, faça uma pesquisa intensa sobre o assunto do seu interesse. Quais outras publicações já abordaram o tema? Quais foram as fontes de dados utilizados? Quais foram as análises, seus pontos fortes e fracos?

Atenção: tenha um cuidado especial ao formular questões que envolvam causalidade. Estabelecer a causalidade entre dois fenômenos é algo bastante complexo. Falaremos mais sobre isso no capítulo sobre análise de dados.

A pergunta mencionada sobre o transporte rodoviário foi a questão seminal da investigação realizada por Adriano Belisário (coautor e organizador do ebook) e a equipe da Agência Pública, que resultou em uma série de reportagens publicadas no Especial Catraca. A partir de levantamentos de dados para responder àquela simples pergunta inicial, a investigação revelou relações inéditas entre deputados e empresários do ramo, por exemplo.

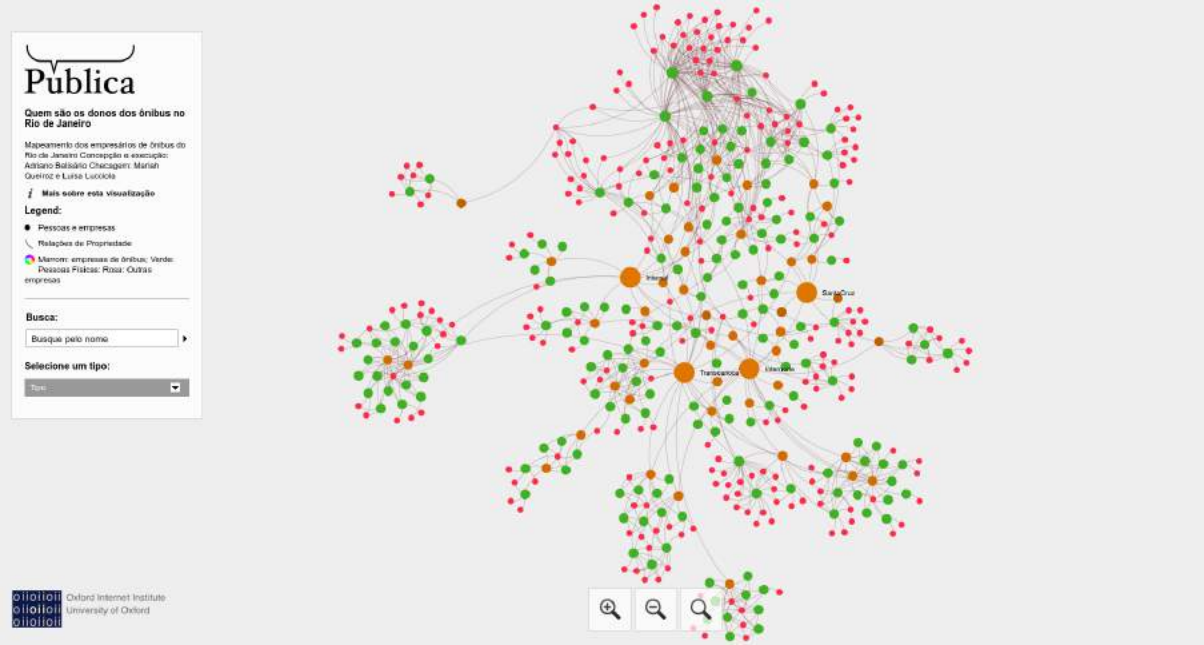


Figura 10: Visualização interativa da rede de empresas e empresários de ônibus no município do Rio de Janeiro

Obviamente, esboçar uma questão sempre requer algum grau de conhecimento sobre o tema e,

quanto mais você se aprofundar, melhores perguntas conseguirá fazer. Ao longo de sua pesquisa, é comum que você detalhe, aprofunde ou mesmo mude completamente a questão que irá te guiar. Não se preocupe em fazer a pergunta perfeita ou definitiva, mas esforce-se para que ela seja detalhada e o mais objetiva possível.

Com este processo, além de questões robustas, você provavelmente terá uma ou mais hipóteses, ou seja, possíveis respostas ou explicações para suas perguntas iniciais.

Depois de identificar e obter os dados, certifique-se que se você sabe suficientemente sobre o contexto no qual ele foi produzido. Quem produziu os dados? Quais são os interesses dessa organização ou instituição? Qual foi a metodologia de registro das informações? Ter respostas a essas perguntas é fundamental para você realizar uma leitura crítica dos dados.

Posteriormente, talvez seja preciso adaptar sua pergunta principal para questões que possam ser respondidas utilizando os dados obtidos. Neste momento, é possível não só testar suas hipóteses, como também explorar os dados para experimentar novas análises, que não estavam no foco da questão inicial, mas podem ser descobertas. Manter esta abertura te permite extrair novos insights na sua pesquisa. Fique atento a valores ou padrões inesperados e tente compreendê-los. Muitas vezes, isto o fará inclusive redefinir a sua pergunta inicial.

Reconhecendo um bom escopo

Para te ajudar neste processo, vamos elencar aqui algumas características de um escopo bem definido, no caso de pesquisas baseadas em dados. O objetivo aqui não é limitar sua abordagem, mas destacar aspectos importantes a serem levados em conta.

É objetivo

Busque definir seu escopo objetivamente e de forma concisa. Isto implica abordar um único problema ou tópico central por vez, mesmo que você tenha várias questões relacionadas a ele. Um escopo bem definido descreve a definição de cada termo ou conceito envolvido nas questões centrais, bem como explicita recortes temporais ou territoriais que serão aplicados. E também não depende de juízos subjetivos. Então, no geral, evite noções como “melhor” ou “pior” em perguntas tais como “isto é melhor que aquilo?”.

Pode ser abordado com dados

Pode ser redundante, mas não custa afirmar que suas questões precisam ser resolvidas com dados. Caso não seja possível obter ou coletar dados para tratar da pergunta que você abordou, talvez você precise refazê-la. É o caso, por exemplo, de um escopo sobre gastos governamentais que envolvam despesas sob sigilo.

É complexo

Ser objetivo não quer dizer que seu escopo deva ser simplista. Ao contrário, um bom escopo é complexo e não pode ser respondido simplesmente com um "sim" ou "não".

É viável de ser realizado

Outro desafio é definir um escopo que seja viável de ser realizado - de nada adianta um escopo perfeito e completo, porém inalcançável. Ou seja, um bom escopo é complexo, porém factível de ser abordado. Isso vai depender muito de seus conhecimentos, tempo, equipe, entre outros fatores. Se sua pergunta depende de dados que são pagos e pelos quais você não pode pagar ou que exige um tipo de análise mais complexa do as que você consegue fazer, simplifique o escopo.

É original

Se o seu intuito é publicar o resultado de uma investigação, é recomendável que você busque abordar um tema com alguma originalidade, nem que seja atualizar uma análise já feita para trazer números mais recentes, por exemplo. Caso você deseje apenas exercitar seus conhecimentos, não há problemas em refazer trabalhos alheios. Pelo contrário, é até recomendável tentar reproduzir análises - assim, você pode aprimorar seu conhecimento e compreender melhor abordagens feitas por outras pessoas.

É relevante

Este é outro caso que vale especialmente para quem deseja compartilhar o resultado do seu trabalho. Se você quer publicá-lo, tente definir temas e um escopo de investigação que seja relevante para as pessoas. Ao pensar sobre o recorte de sua pesquisa, pergunte-se: qual tipo de público terá acesso ao trabalho? Por que ele deveria se importar com isso?

Dados para além de tabelas

Para compreender as diferentes possibilidades de trabalhar com dados e pensar um escopo para sua pesquisa, também é interessante reconhecer as diferentes formas que os dados podem assumir. Geralmente, quando pensamos em dados, logo imaginamos números em tabelas. Porém, o formato de dado tabular é apenas um entre vários. Nosso guia tem foco justamente neles, os dados tabulares, mas nesta seção vamos te apresentar outros tipos de dados para que você conheça suas vantagens e tenha um panorama de seus principais formatos e tecnologias. Assim, você conseguirá pensar recortes para sua pesquisa de forma mais ampla, para além das tabelas.

Dados tabulares

Como dito, esta é a forma mais "tradicional" de estruturar dados. Ela consiste em organizar os registros em tabelas, com linhas e colunas. As extensões de arquivos mais comuns para **dados tabulares** são CSV, XLS, XLSX e ODS. Você pode usar editores de planilha para lê-los ou importá-los em bases de dados com SQL, por exemplo. Também pode acontecer de você encontrar tabelas em formato PDF, o

que exigirá um trabalho adicional. Falaremos mais sobre isso no capítulo seguinte.

Dados espaciais

É possível incluir coordenadas **geográficas em planilhas**, como informações de **latitude e longitude**, de modo que você consiga visualizar seus dados em um mapa. Porém, é justo considerar os dados geográficos ou espaciais como uma "família" ou tipo por si só. Isto porque há conceitos e técnicas de análise dos dados espaciais que são únicas e não se aplicam a outros tipos de dados tabulares. Há todo um universo próprio de dados geoespaciais no chamado **GIS** (Sistemas de Informação Geográficas ou Geographic Information System, em inglês).

Além disso, visualizar a distribuição territorial dos dados em um mapa nos permite revelar padrões espaciais, que são invisíveis em uma tabela. Vale a máxima do geógrafo Waldo Tobler, considerada a primeira lei da geografia: "todas as coisas estão relacionadas com todas as outras, mas coisas próximas estão mais relacionadas do que coisas distantes".

Os dados espaciais possuem uma variedade de formatos/extensões de arquivos bem particulares, por exemplo, KML, GeoJSON, SHP, entre outros. Há também ferramentas e programas específicos para lidar com eles e, neste caso, o destaque fica por conta do **QGIS**, que é o principal programa de código-aberto para lidar com dados geográficos e produzir mapas.

Redes

Outra forma de representar dados é em forma de **redes ou grafos**. Na prática, muitas vezes, você consegue "traduzir" dados tabulares em redes e vice-versa. Por exemplo, considere a tabela abaixo, com informações fictícias sobre os proprietários de algumas empresas de ônibus.

Empresa	Proprietário
Viação 1	Joaquim
Viação 1	Manoel
Viação 1	Luiza
Viação 2	Joaquim

Estes mesmos dados poderiam ser "traduzidos" para uma representação em redes com o grafo abaixo.

Ao invés de organizar a informação em uma matriz ou uma tabela, os dados são estruturados a partir de nós (os pontos nas imagens acima) e arestas (os traços que conectam os pontos). Essa estrutura é especialmente útil quando seu interesse são relações entre entidades. Não é à toa que os grafos se tornaram bastante populares para entender a disseminação da informação e relações entre pessoas na análise de redes sociais. Assim como os dados geográficos, as redes possuem conceitos, metodologias e técnicas de análises próprias, que seriam impossíveis ou muito difíceis de serem realizadas utilizando tabelas.

Para lidar com redes, vale a pena conhecer o **Gephi**, que é a principal solução em código-aberto para

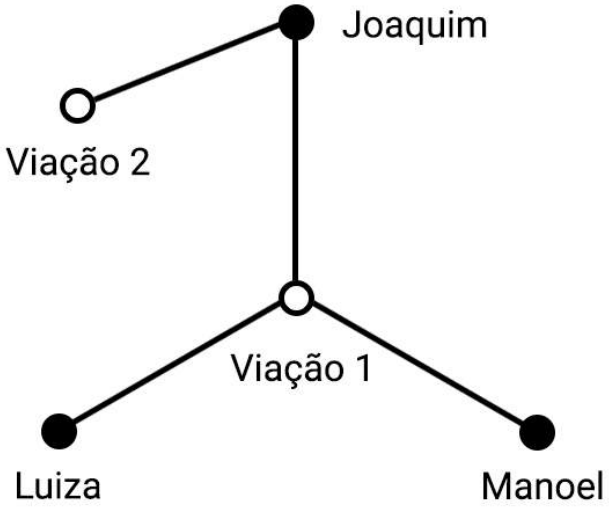


Figura 11: A imagem mostra em forma de grafo as relações entre empresas e empresários(as), que foram representadas como uma tabela anteriormente

visualizar e analisar grafos. Com ele, você consegue carregar tabelas e formatos nativos de grafos (como GEXF ou GML) ou até mesmo conectar-se diretamente a serviços como o Twitter, para extrair dados em tempo real.

Textos

Textos também podem ser estruturados como dados. Neste caso, as planilhas podem ser utilizadas em algum momento, mas há também uma série de conceitos, metodologias e técnicas de análise de texto próprias, que são bem específicas e não se confundem com as abordagens baseadas somente em dados tabulares. As **nuvens de palavras** talvez sejam o exemplo mais popular desta perspectiva de texto como dado. Nela, o tamanho das palavras-chaves é proporcional à frequência em determinado conjunto de textos.

Para lidar com análise de texto, você precisará aprender coisas como remoção de palavras frequentes (stopwords) para otimizar suas análises e padronização de palavras semelhantes (lematização ou stemização). Neste caso, a área do conhecimento que você deve se aprofundar é a de **Processamento de Linguagem Natural** (ou Natural Language Processing - NLP, na sigla em inglês). Com ela, você pode ir muito além das nuvens de palavras e realizar tarefas mais avançadas, como detectar automaticamente entidades (nomes de pessoas, empresas ou lugares) em um conjunto de texto. Estas técnicas são utilizadas em um volume muito grande texto, como no caso dos jornalistas que tiveram lidar com o Panamá Papers.

Não existem muitas opções de softwares de código-aberto com interface gráfica que facilitem a análise de textos, mas vale a menção ao Voyant Tools, que conta com uma versão web e permite análise rápidas, e o DocumentCloud, uma plataforma que facilite a organização de documentos e permite detecção de entidades até em PDFs.

Para entender as possibilidades de trabalhar com texto como dados, vale conferir trabalhos como este do The Pudding sobre o vocabulário de cantores de rap ou este do Nexo, que foi ganhador do primeiro Prêmio Cláudio Weber Abramo de Jornalismo de Dados no Brasil. Na reportagem premiada, os jornalistas analisaram as mudanças no texto da Constituição ao longo das três décadas desde sua promulgação.

Imagens

Este é um terreno mais difícil de ser trabalhado, se você está começando. De todo modo, é bom que você saiba que ele existe. Imagens e vídeos também podem ser trabalhados como dados, ainda que isto seja algo avançado e demande alguma familiaridade com programação. Neste caso, a chave é compreender as tecnologias de visão computacional. A principal ferramenta em código-aberto para isso é o **OpenCV** e você encontra implementações tanto em Python como em R.

A tecnologia de visão computacional permite que o computador consiga "ler" imagens para reconhecer entidades (animais, pessoas ou objetos, por exemplo) ou até mesmo expressões faciais. Um ótimo exemplo no jornalismo de dados deste tipo de abordagem é este trabalho realizado por Rodrigo Me-

negat (coautor deste ebook) no Estadão. Na reportagem, softwares de visão computacional foram utilizados para identificar e analisar os sentimentos dos candidatos à presidência durante um debate televisivo a partir de suas expressões faciais.

Obtenha

Agora que vimos alguns pontos importantes para definir o escopo de nossa pesquisa e a diversidade de formatos que os dados podem assumir, vamos voltar aos **dados tabulares** para entender como podemos obtê-los. Basicamente, existem três formas de se obter dados.

A primeira é coletá-los você mesmo. Apesar de comum em setores como o poder público e algumas organizações da sociedade civil, no campo da comunicação, em geral, você irá lidar com dados já coletados por alguém. Por isto, como dito no capítulo anterior, não iremos abordar a coleta de dados primários neste guia.

O segundo modo é obter dados inéditos. Neste livro, iremos focar no principal mecanismo legal para isso no Brasil. Quando você já esgotou suas possibilidades de busca, no caso do poder público, há uma alternativa bastante promissora para você obter os dados que precisa: a **Lei de Acesso à Informação**. Neste capítulo, veremos como utilizá-la a seu favor para conseguir novas bases de dados pelo Serviço de Informação ao Cidadão previsto pela lei.

Por último, o terceiro modo de se obter dados é simplesmente os encontrando. Muitas vezes, as informações que você precisa já estarão disponível na internet.

Porém, nem sempre é fácil encontrar a informação que você precisa, principalmente se você está lidando com um tema que não domina. Nestes casos, antes de ir atrás dos dados propriamente ditos, é preciso fazer uma pesquisa ou buscar ajuda com especialistas para entender melhor o campo que você desejar estudar e compreender quais bases de dados estão disponíveis.

Neste capítulo, veremos quais técnicas são utilizadas para lidar com situações como essa. Iremos revisar diversos conceitos e técnicas relativas à obtenção de dados, das mais simples às mais complexas.

Primeiramente, vamos abordar a busca avançada na web. Esta etapa é fundamental, pois antes de tudo você precisa definir quais termos ou palavras-chaves serão utilizadas na pesquisa. Por isso, veremos agora algumas técnicas importantes para fazer pesquisas precisas e obter os dados que você precisa.

Busca avançada na web

Não raro, os dados encontram-se "escondidos" ou são de difícil acesso. Por isso, é importante conhecer alguns operadores de busca avançada na web. Basicamente, esses operadores funcionam como uma espécie de filtro, que permite garimpar na vastidão da internet exatamente o que você precisa.

Talvez você já conheça alguns operadores de busca mais básicos. Por exemplo, se você apenas joga as palavras ou termos no buscador, ele irá vasculhar páginas que contenham os termos inseridos em qualquer lugar da página, mesmo que não estejam em sequência. Porém, se você colocar os termos entre aspas, a busca retornará apenas as expressões exatas, ou seja, quando os termos retornam exatamente na ordem em que foram inseridos. Para ver isso na prática, considere a palavra-chave "obtenção de dados" e primeiro faça uma busca no Google com elas entre aspas e depois sem aspas.

Este é apenas um entre vários operadores de busca disponíveis. Conhecer estas opções irá ampliar em muito a forma como você busca informações na internet. É possível fazer filtros por data, extensões de arquivo ou sites específicos da internet, entre várias outras opções. Além disso, você consegue combinar os operadores para fazer buscas realmente precisas. Vejamos o exemplo abaixo.

```
coronavirus filetype:xls site:gov.br
```

Se você digitar isso no Google, o buscador irá retornar todas as planilhas em Excel com extensão XLS que estão hospedados em endereços do governo brasileiro e possuem a palavra *coronavírus*.

Para conhecer os operadores de busca disponíveis e aperfeiçoar suas habilidades de busca na web, recomendamos a leitura deste tutorial introdutório e especialmente deste texto que publicamos sobre busca avançada. O texto é uma tradução de um artigo de Daniel M. Russell, pesquisador sênior do Google, que documentou todos operadores de busca da ferramenta.

Se você quiser ir fundo no tema, vale a pena pesquisar por técnicas de inteligência com fontes abertas (open source intelligence - OSINT). Os operadores de busca avançados são uma das técnicas utilizadas em OSINT. No entanto, o campo compreende várias outras abordagens, que inclusive vão além do trabalho com dados, como a geolocalização de imagens a partir de características visuais.

De todo modo, as metodologias e as abordagens de OSINT são bastante interessantes para melhorar suas habilidades de investigação na internet e com dados. Um dos grupos de mais destaque nesta área é o Bellingcat. Vale a pena conferir suas investigações e os tutoriais que disponibilizam. Além disso, o grupo tem o guia "Bellingcat's Online Investigation Toolkit", uma compilação fantástica de fontes e recursos para obtenção de dados online.

Lei de Acesso à Informação e dados abertos

Outro importante mecanismo de obtenção de dados são as requisições oficiais feitas a governos e entidades públicas. Este mecanismo existe em diversos países. Nos Estados Unidos, chama-se *Freedom*

Of Information Act (FOIA), por exemplo, mas no Brasil é conhecida como **Lei de Acesso à Informação** (LAI - Lei 12.527/2011).

Esta legislação que entrou em vigor em 2012 regulamentou o direito à informação garantido pela Constituição Federal brasileira em seu artigo quinto, reforçando o entendimento da publicidade como regra e o sigilo como exceção. Ela obriga órgãos públicos de diferentes esferas e poderes a fornecer ativamente e abertamente, na web, informações de interesse da sociedade. Assim, o caminho inicial para quem quer obter dados governamentais é verificar se eles já estão disponíveis no site do órgão sobre o qual se busca a informação.

Além de regulamentar esta transparência ativa, quando o próprio órgão disponibiliza a informação proativamente, a LAI também criou o Serviço de Informação ao Cidadão, que implementa na prática a chamada transparência passiva. É o SIC - ou eSIC, nos casos de atendimento online - que garante acesso do público aos dados que são de interesse coletivo, mas por algum motivo não estão disponíveis para consulta.

Mas a LAI vale para todos? Vejamos a quem se aplica esta legislação. **Tanto a União quanto os estados, o Distrito Federal e os municípios estão sujeitos à LAI. Isso vale no âmbito dos três poderes: legislativo, executivo e judiciário. Além deles, também devem obedecer a LAI as autarquias, fundações ou empresas públicas, sociedades de economia mista e entidades controladas direta ou indiretamente por entes da federação. Entidades privadas que recebam recursos públicos igualmente devem fornecer as informações relacionadas ao uso dessas verbas**.

Antes de recorrer à LAI, porém, é importante que você faça o dever de casa: busque os portais de transparência já existentes e verifique se o dado desejado não está disponível. Faça essa checagem antes de realizar o seu pedido para evitar sobrecarregar o serviço público com demandas que já foram resolvidas.

Além de buscar nos portais de transparência e nas fontes oficiais sobre o dado que você deseja, vale também consultar solicitações e respostas já prévias, no caso de pedidos submetidos ao governo federal. Pode ser que alguém já tenha solicitado o dado que você busca e, se for este o caso, você pode encontrar no site da Controladoria Geral da União tanto a solicitação quanto a resposta do órgão.

O prazo de resposta é de 20 dias, mas o órgão pode prorrogá-lo por mais 10. Em caso de pedidos negados, você tem direito de ter acesso à justificativa completa para esta decisão e pode questioná-la em até 10 dias. Depois disso, a autoridade hierarquicamente superior àquela que negou o acesso deve se manifestar em até 5 dias. O recurso pode ser usado tanto nos casos em que o acesso à informação não sigilosa for negado, ou procedimentos (como prazos) forem desrespeitados, quanto para pedir a revisão da classificação da informação sigilosa.

Para fazer bons recursos, vale buscar embasamento na legislação. Artigos de leis e decretos, de todas as esferas, podem ser aliados nesse processo. Também vá atrás de precedentes, não necessariamente do mesmo órgão. Pode ser de outras pastas com o mesmo tema. Nesse caso, o site Busca Precedentes, da CGU, pode ser um aliado. Em último caso, muitas vezes, a própria recusa em liberar a informação

ou a declaração do sigilo já podem ser informações suficientemente fortes para valer uma publicação.

Você não precisa dar nenhuma justificativa para realizar um pedido de acesso à informação. Ao solicitar dados, lembre-se também que a Lei de Acesso à Informação, especificamente no artigo 8º, explicita a importância de disponibilizar as informações em formatos abertos, estruturados e legíveis por máquina e de forma acessível para todos os cidadãos brasileiros. Para tentar minimizar os riscos de receber dados em formatos inadequados em seus pedidos, você pode inserir uma solicitação explícita para que sejam encaminhados como dados abertos, juntamente com um dicionário de dados (documento que traz informações sobre as "regras de negócios" por trás dados) ou alguma documentação básica que os explique.

Um exemplo de texto que você pode usar é o seguinte texto:

> "Solicito que os dados sejam fornecidos preferencialmente em formato aberto (planilha em .csv,.ods, etc) de acordo com o determinado no art. 8º, §3º da Lei Federal 12.527/11, o item V do art. 24 da Lei Federal 12.695/14 e a normatiza expressa no item 6.2 da Cartilha Técnica para Publicação de Dados Abertos no Brasil, elaborada pelo Governo Federal. Solicito ainda que seja encaminhado também o dicionário de dados, documento ou explicação equivalente que descreva o significado de cada variável/coluna".

Na sequência, entenderemos melhor o que significa na prática a noção de *dados abertos*. Mas antes vamos fazer uma revisão de um checklist para você conferir antes de fazer seus pedidos de LAI.

Dicas para consultas via Lei de Acesso à Informação

- **Faça o dever de casa**: é importante saber quem detém a informação solicitada para fazer o pedido ao órgão responsável;
- **Cada um no seu quadrado**: faça um pedido/protocolo para cada dado desejado, principalmente se forem informações diferentes;
- **Tenha foco**: pedidos muito genéricos costumam ser recusados, por isso é importante fazer recortes precisos de período, local e itens específicos. É importantes ser claro. Você pode fazer o pedido usando listas, explicar qual o recorte temporal que deseja ou mesmo quais os campos ou variáveis você gostaria de ter acesso, caso solicite dados;
- **Use artigos/precedentes**: para embasar o pedido é possível recorrer a artigos da lei ou outros precedentes, como decisões em outros estados, por exemplo;
- **Antecipe sigilos**: se os dados podem ter informações sigilosas, você já pode se antecipar e solicitar que eles sejam tarjados ou removidos para evitar possíveis negativas com essa justificativa.

Dados abertos

De acordo com a Open Knowledge, **dados abertos** são aqueles que qualquer um pode livre e gratuitamente acessar, usar, modificar e compartilhar para qualquer propósito, inclusive para uso de comercial. Eles estão sujeitos, no máximo, à exigência de menção da autoria e abertura de trabalhos derivados.

Isto diz respeito ao licenciamento dos dados, mas também é importante que o formato do arquivo no qual os dados estão disponibilizados seja um formato aberto. O formato dos arquivos são indicados pela sua extensão. Você deve saber que XLS ou XLSX é a extensão utilizada pelo Excel para planilhas, mas este é um formato de arquivo fechado, que é propriedade privada da Microsoft. Portanto, não é um formato aberto. Para saber mais sobre os diferentes níveis de abertura de dados, vale conferir o ranking da iniciativa 5-star Open Data, uma classificação criada por Tim Berners-Lee, inventor da web e defensor de dados interconectados (*linked data*).

Os **formatos mais comuns para dados abertos são o CSV, no caso de arquivos com planilhas já disponíveis para download, e o JSON ou o XML no caso de obtenção de dados por meio de APIs**. Enquanto o primeiro representa dados tabulares, com linhas e colunas, o XML e o JSON possuem uma estrutura hierárquica, onde os valores estão uns dentro dos outros, mais ou menos como as pastas de seu computador são organizadas.

No geral, dados abertos não tão comuns de encontrar. Você provavelmente encontrará alguns empecilhos antes de sequer começar a analisar ou visualizar a sua base. Os dados podem estar espalhados em várias planilhas diferentes, despadronizados, desatualizados e sem documentação suficiente para que se compreenda o que as informações significam. A lista de problemas é longa.

Um dos primeiros problemas nesta fase de obtenção é não encontrar os dados que você precisa como dados abertos. Eles podem estar disponibilizados por meio de páginas na web, que foram feitas para serem visualizadas por humanos, mas são inúteis se queremos fazer análises ou visualizações. Ou pior: podem ser fotos ou tabelas digitalizadas.

A seguir, veremos alguns problemas comuns na hora de obter dados em tabelas - e como você pode contorná-los.

Problemas comuns com tabelas

Mesmo quando você encontra uma tabela já estruturada em formatos abertos, é possível que você enfrente alguns problemas na hora de abrir os arquivos. Nesta seção, iremos ver três problemas muito comuns, especialmente entre iniciantes, que identificamos na Escola de Dados ao longo destes anos de treinamentos.

A solução deles é bastante simples, porém, se você nunca passou por isso antes, pode ser que encontre dificuldades. Todos os problemas são especialmente comuns ao lidar com tabelas, seja em editores

de planilhas ou utilizando linguagens como Python ou R.

Para quem está começando, apesar de serem de fácil resolução, estes problemas podem causar bastante confusão, pois a função padrão de *Abrir Arquivo* dos editores de planilha costumam passar por cima desses "detalhes" importantíssimos. Por isso, de modo geral, é recomendável utilizar a função de importar dados para abrir os CSV, pois assim você poderá fazer ajustes em configurações importantes, como as que veremos abaixo.

Além disso, no fim desta seção, também veremos como lidar com dados que estão em formato PDF ou como imagens/fotos.

Separador de campos

A sigla CSV significa *Comma Separated Values* ou, em bom português, valores separados por vírgulas. Isto porque um arquivo CSV nada mais é que um arquivo de texto onde um certo caractere é utilizado como separador das colunas. Ou seja, se você abrir seu bloco de notas e salvar um arquivo com extensão CSV contendo o texto abaixo poderá abrí-lo depois no editor de planilhas e ele será visualizado como uma tabela.

```
nome,aula,nota

rosa,matemática,10

walter,português,8

maria,português,5
```

	A	B	C
1	nome	aula	nota
2	rosa	matemática	10
3	walter	português	8
4	maria	português	5

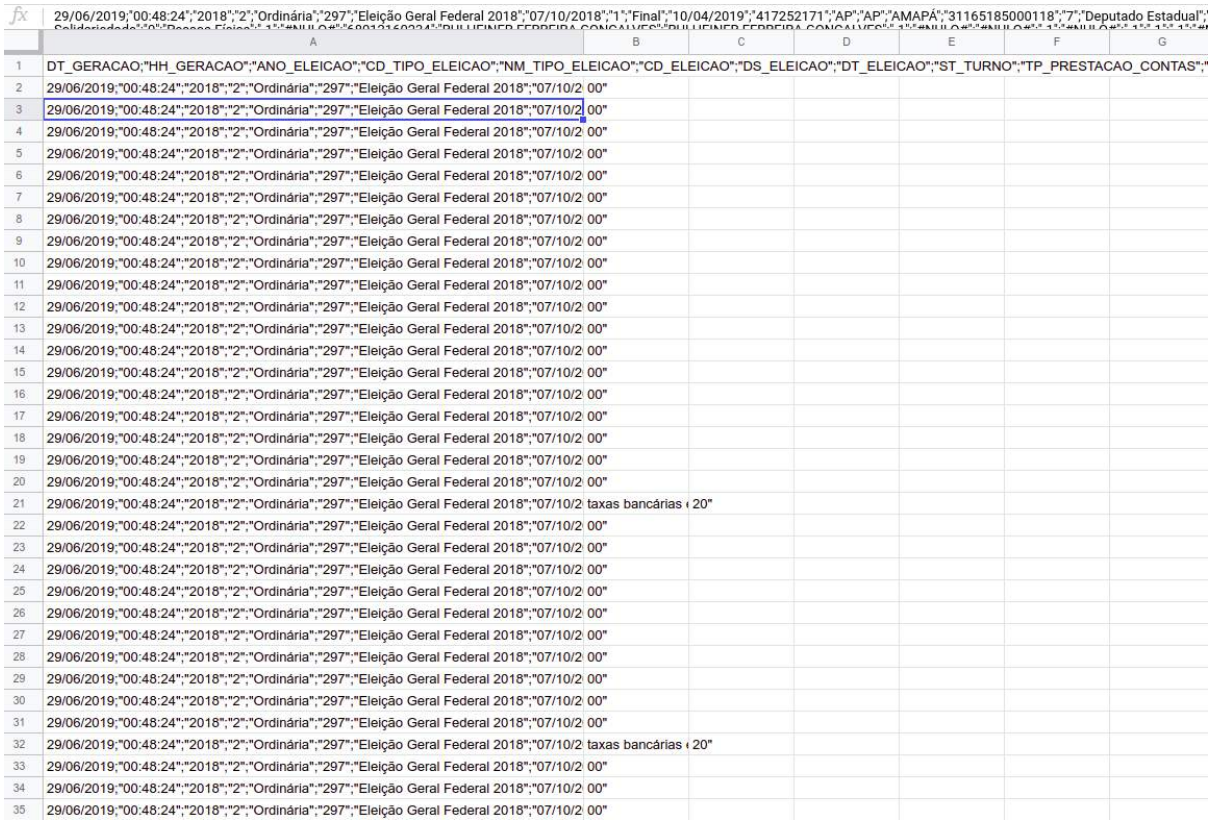
Figura 12: Texto estruturado em CSV, que será exibido como uma tabela, se aberto em um editor de planilha com o separador de campo adequado

No caso apresentado, os campos ou colunas são separados com vírgula. Essa é a base fundamental por trás de um CSV, mas existem particularidades. Podemos, por exemplo, utilizar aspas para delimitar o resultado de uma célula, além das vírgulas. Não iremos entrar nestes detalhes aqui. Para quem está começando, o importante não é compreender todas especificações de um CSV, mas simplesmente conseguir abrir adequadamente os dados quando necessário.

Como a sigla da extensão CSV indica, a vírgula é o separador de caractere padrão. Porém, é comum que você encontre tabelas que utilizem o ponto-e-vírgula como separador dos campos. É o caso, por exemplo, de algumas tabelas disponibilizadas pelo Tribunal Superior Eleitoral no Brasil (TSE).

Ao tentar abrir estes arquivos com um editor de planilha, sem especificar qual o separador de caractere utilizado, é possível que os campos e colunas apareçam desconfigurados. Por desconhecer o separador de campos adequado, o programa não conseguirá exibir os valores nas suas células adequadas.

No exemplo a seguir, você vê uma tabela em CSV do TSE aberta usando a função *Abrir*, sem a especificação do campo de separador de caractere. Repare que os valores das colunas são mostrados de forma inadequada.



	A	B	C	D	E	F	G
1	DT_GERACAO;"HH_GERACAO";"ANO_ELEICAO";"CD_TIPO_ELEICAO";"NM_TIPO_ELEICAO";"CD_ELEICAO";"DS_ELEICAO";"DT_ELEICAO";"ST_TURNO";"TP_PRESTACAO_CONTAS";"						
2	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
3	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
4	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
5	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
6	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
7	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
8	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
9	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
10	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
11	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
12	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
13	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
14	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
15	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
16	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
17	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
18	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
19	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
20	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
21	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	taxas bancárias (	20"				
22	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
23	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
24	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
25	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
26	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
27	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
28	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
29	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
30	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
31	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
32	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	taxas bancárias (	20"				
33	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
34	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					
35	29/06/2019;"00:48:24";"2018";"2";"Ordinária";"297";"Eleição Geral Federal 2018";"07/10/2	00"					

Figura 13: Exemplo de tabela com problema no separador de caractere.

Agora, veja a mesma tabela após ser aberta utilizando a função de *Importar*, que permite a especificação do separador de caractere correto. Neste caso, todos os valores aparecem nas colunas adequadas.

Na prática, é possível utilizar qualquer caractere como separador em um CSV. Porém, **a vírgula e o ponto-e-vírgula são de longe os mais comuns**. Portanto, saiba que, se você abriu uma tabela em

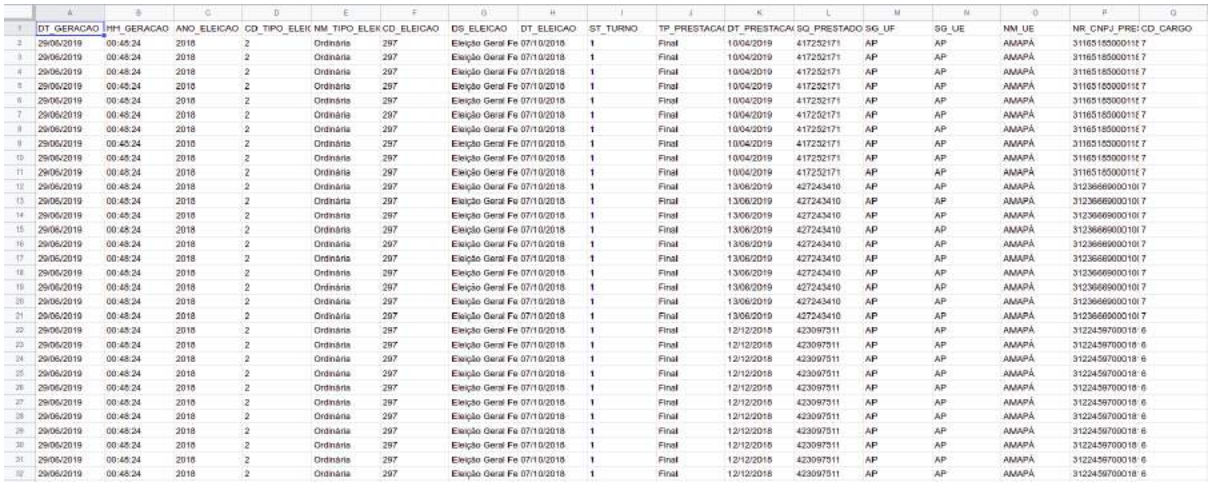


Figura 14: A mesma tabela com o problema no separador de caractere corrigido.

CSV e as colunas aparecem desformatadas, muito provavelmente você não selecionou o separador de campos adequado. Provavelmente, você precisará utilizar a função *Importar* e então especificar manualmente o ponto-e-vírgula como o separador de campos.

Codificação de caracteres

Outro problema bastante comum, não só com planilhas, mas com qualquer documento digital, são os erros de *encoding* ou codificação de caracteres. Você já tentou abrir algum arquivo e alguns caracteres especiais apareciam de forma estranha, mais ou menos como neste exemplo?

	A	B	C	D	E	F	G
1	DT_GERACAO	HH_GERACAO	ANO_ELEICAO	CD_TIPO_ELEICAO	NM_TIPO_ELEICAO	CD_ELEICAO	DS_ELEICAO
2	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
3	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
4	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
5	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
6	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
7	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
8	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
9	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
10	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
11	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
12	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
13	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
14	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
15	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
16	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
17	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018
18	29/06/2019	00:48:24	2018	2	Ordin�ria	297	Elei��o Geral Federal 2018

Figura 15: Exemplo de tabela com problema de codificação de caracteres

Repare como as letras com acentos na tabela não aparecem adequadamente. As codificações de caracteres que “traduzem” os caracteres que nos conhecemos em padrões que os computadores conseguem entender. Novamente, você não precisa se preocupar com os detalhes técnicos sobre isso para abrir suas tabelas. Basta saber que, se alguns caracteres da sua tabela aparecem de forma estranha, tal como na imagem acima, então, provavelmente, a codificação de caracteres selecionada não está correta.

Existem alguns programas que podem te ajudar a identificar a codificação de caractere utilizada em um arquivo, mas você pode consultar a documentação dos dados que obteve em busca de alguma especificação do tipo ou simplesmente tentar os dois tipos de *encoding* mais comuns.

O UTF-8 (Unicode) é de longe a codificação de caracteres mais comum da web. Quando você apenas lê algo, sem se preocupar com coisas como esta, provavelmente seu computador está usando UTF-8. Porém, se você encontrar problemas, vale a pena tentar abrir seus dados com o *encoding* ISO-8859-1 (Latin 1), uma codificação de caracteres latinos, que você pode encontrar em alguns documentos brasileiros, por exemplo. O caminho para corrigir a codificação varia conforme o software utilizado para manusear os dados, mas você conseguirá encontrar facilmente tutoriais e orientações na internet sobre o assunto.

Tipo de dados na coluna

Imagine que você se depare com uma planilha do Instituto Brasileiro de Geografia e Estatística (IBGE). No conteúdo organizado pelo IBGE, cada município possui um código único, viabilizado por meio de números. No entanto, você não os usará para fazer cálculos matemáticos. Não faz sentido, por exemplo, multiplicar o código de identificação de Fortaleza com o de Porto Alegre. Este número na realidade é um código, um conjunto de símbolos que tem um significado específico. Por isso, é importante que colunas que contenham códigos numéricos sejam convertidas para o formato texto.

Os números em formato de código podem ter diferentes significados dependendo da sua base: categorizações, identificadores (IDs), resposta afirmativa, positiva ou ausente. Isso pode ser esclarecido no dicionário de dados, caso sua base venha acompanhada dele.

Em geral, usuários iniciantes ignoram essa configuração do tipo de dados de cada coluna e acabam tendo dificuldades para realizar operações simples, como o ordenamento das linhas a partir de um valor específico, por exemplo. Outro problema ocorre quando os registros são documentos de identidade, Cadastro de Pessoa Física, Jurídica (CPF e CNPJs) ou outros identificadores que começam com 0, por exemplo. Quando esses registros estão no formato numérico, os editores de planilha podem eliminar os zeros à esquerda. Isso pode causar diversos problemas ao tentar cruzar bases de dados distintas ou mesmo ocasionar a perda dos valores originais da tabela, caso você seja sobrescreva o arquivo.

O contrário também pode acontecer. Há ocasiões em que os números - que realmente precisam ser números - estão em formato texto, impedindo que qualquer cálculo seja feito com eles. Por isso, antes de partir para a análise, verifique a formatação de cada coluna da sua planilha e, caso necessário, modifique-a.

Padrão de números e datas

Outro ponto para se ter atenção é a localidade onde os dados foram produzidos. Nos Estados Unidos, na representação de números, os pontos separam os decimais e as vírgulas, os centésimos. Já no Brasil e na Europa, é o inverso: usamos o ponto para separar os centésimos e a vírgula para separar os números decimais. Isto pode ser um problema na hora dos softwares "entenderem" e realizarem operações com números.

As datas também costumam ser elementos complicadores e precisam de atenção. Dependendo da origem da sua planilha, elas podem chegar a você no formato dia/mês/ano (padrão brasileiro) ou no formato mês/dia/ano (padrão norte-americano). Por isso, é importante saber qual a origem dos dados e certificar-se de que eles não foram produzidos em localidades distintas. Alguns editores de planilha possuem configurações que permitem indicar qual é a localidade do arquivo e, a partir disso, deduz as configurações de separadores, datas e moedas. Caso seus dados adotem padrões distintos, será necessário padronizá-los para realizar análises.

Lidando com PDFs

O primeiro passo para lidar com dados em PDF é identificar se há elementos de texto no documento ou se o arquivo é, na verdade, uma foto. No primeiro caso, o processo tende a ser mais tranquilo e menos suscetível a erro. No segundo, é preciso aplicar uma técnica especial chamada **reconhecimento ótico de caracteres** (OCR), que permite que o computador "veja" a imagem e "transcreva" os números e textos nela. Se você passa o mouse em cima da tabela e consegue selecionar os caracteres, então, trata-se do primeiro caso. Se não, então, você está vendo na verdade uma imagem.

Infelizmente, não existe uma solução única ou perfeita para libertar dados em PDF. No entanto, existem alguns fatores que ajudam a escolher a melhor ferramenta. Se você precisa realizar a conversão uma única vez e em um documento com poucas páginas, uma solução gratuita - inclusive com OCR - deve ser suficiente. Se você quiser automatizar esse procedimento, customizá-lo ou realizá-lo para um número maior de documentos e páginas, linguagens de programação como Python e R se apresentam como uma boa saída.

Para facilitar, fizemos um resumo das principais soluções para extração de dados em PDF.

PDF como imagens (OCR)

Caso você queira uma solução baseada em Python ou R, busque implementações do Tesseract:

- PyTesseract (Python)
- Tesseract (R)
- PyPDFOCR (Python)

Se preferir soluções que não requeiram conhecimento de programação, vale a pena conferir as seguintes alternativas:

- Abby
- Zamzar
- Cometdocs
- Yagf (open source)
- OnlineOCR (gratuito)

PDF como texto

Se você tiver um documento com elementos de texto, existe uma gama maior de opções. Usando linguagens de programação, destacamos os seguintes softwares:

- Rows (Python) (criada por Álvaro Justen, colaborador e membro da Escola de Dados)
- PDFMiner (Python)
- PDFTools (R)
- Tabula (Python)

Já se quiser opções com interface gráfica, então, separamos esta lista de opções:

- Tabula: provavelmente será sua primeira opção nestes casos, é de código aberto e você pode rodar no seu próprio computador.
- PDFTables (online)

Raspagem de dados e APIs

Quando as informações que você precisa não estão estruturadas como dados abertos, será necessário fazer uso de uma técnica conhecida como **raspagem de dados ou web scraping**, que envolve realizar cópias de informações de interesse disponibilizadas em páginas web, organizando-as como dados estruturados. Este processo é feito utilizando alguns programas específicos e, nesta seção, você verá algumas das principais ferramentas.

Para raspagens mais simples, você pode utilizar soluções com interface gráfica e resolver o que precisa com poucos cliques. Porém, a depender da complexidade da tarefa, talvez seja necessário que você desenvolva um programa ou script próprio, usando Python ou R, por exemplo.

Os detalhes sobre raspagens de dados variam imensamente de acordo com a fonte dos dados ou volume de requisições necessárias, já que alguns sites bloqueiam ou restringem o acesso quando identificam tentativas de raspagem dados. Outro problema comum é o uso de CAPTCHAS ou conteúdos carregados dinamicamente via JavaScript, que dificultam a raspagem.

Não iremos entrar no detalhe de cada situação. De modo geral, o que você precisa saber que é este método de coleta costuma envolver duas etapas fundamentais.

A primeira delas envolve descarregar as páginas e informações brutas. Independente de qual for a solução de raspagem adotada, ela precisará começar fazendo o download das páginas web ou arquivos que serão utilizados.

Em seguida, você precisará separar o joio do trigo. Como os dados não estão estruturados em sua origem, será preciso fazer isso. Esta etapa por vezes é chamada de *parsing*. Nela, você irá identificar quais elementos de uma página HTML que interessam na formação da sua tabela, por exemplo. Esta etapa já pode ser considerada também como parte do processo de limpeza de dados, que veremos adiante.

Todo processo de raspagem, *parsing* ou limpeza de dados envolve basicamente uma habilidade fundamental: reconhecimento de padrões. Seja na identificação dos elementos que precisam ser raspados, seja no processamento dos arquivos para "limpá-los" das informações desnecessários, será necessário identificar quais padrões irão permitir ao computador chegar aos dados tal como você deseja.

Para melhorar seus conhecimentos em raspagem de dados, você precisa conhecer pelo menos o básico de HTML. Também são úteis conhecimentos em CSS ou XPath. Abaixo, você encontra um resumo (não-exaustivo) de algumas das principais soluções para raspagens de dados.

Via interface gráfica

- WebScraper: é uma extensão do navegador Chrome, que ativa uma nova aba na seção do navegador voltada para desenvolvedores web. Com ela, você poderá descrever os caminhos de um site para um raspador. Apesar da interface não ser tão intuitiva a princípio, é bastante poderoso e pode ser usado com sucesso em casos como o de raspagem em páginas numeradas ou quando é exigido algum tipo de navegação pelo site para obter os dados desejados.
- Portia: é uma interface gráfica para o pacote em Python chamado Scrapy

Com Python

- Scrapy
- beautifulsoup4up4
- Selenium

Com R

- rvest: é parte do conjunto de bibliotecas conhecido como Tidyverse, bastante recomendado para quem deseja trabalhar com dados.

Comunicação com APIs

APIs são interfaces que facilitam a comunicação de dados entre computadores. A sigla significa **interface de programação de aplicações** (Application Programming Interface, em inglês) e esta forma de acessar dados está presente nas principais plataformas e redes sociais, mas também em portais de dados governamentais.

Se a raspagem de dados, como o nome indica, evoca um método um tanto bruto de se obter os dados, a API seria o oposto. APIs são interfaces que foram criadas justamente para isso: facilitar o fornecimento de dados estruturados, para que possam ser consumidos por terceiros.

Existem alguns programas com interface gráfica que são capazes de se comunicar com estas APIs e carregaram dados. É o caso do Gephi, por exemplo, cujas extensões (plugins) podem torná-lo capaz de acessar dados da API do Twitter. O OpenRefine, que veremos a seguir no capítulo sobre limpeza de dados, também pode consultar APIs.

Na Escola de Dados, você encontra um tutorial de como utilizá-lo para geolocalizar as coordenadas de latitude e longitude de endereços em texto, utilizando a API do OpenStreetMaps, por exemplo. Porém, no geral, quando utilizamos a API, é mais comum que ela seja acessada usando algum script.

Agora que vimos alguns aspectos e ferramentas da obtenção de dados, passaremos no próximo capítulo à etapa de checagem das informações recebidas.

Cheque

Até agora, você aprendeu, em linhas gerais, o que são dados e como encontrá-los ou obtê-los. Para começar a fazer análises, é preciso mergulhar na etapa de **verificação e limpeza** dos dados.

Você pode estar com uma planilha (aparentemente) bem estruturada na sua frente e ansioso para analisá-la, mas se você seguir direto para essa etapa e pular este capítulo poderá se frustrar, ter que refazer a análise ou, pior, publicar informações erradas.

É comum que as bases de dados não cheguem prontas para análise e precisem passar por uma série de verificações e limpezas. É bastante frequente detectar inconsistências ou retirar colunas para reduzir o tamanho de uma base de dados e lidar com elas mais facilmente.

A **verificação** consiste na busca de elementos que comprovem que os dados coletados estão corretos, são consistentes e que não há falta de informação que comprometa o seu trabalho posterior. Esse processo envolve desde as checagens mais básicas e aparentemente banais, como verificações de formato (texto ou numérico, por exemplo) a outros pontos mais complexos, que envolvam comparar os dados com outras fontes.

A checagem é uma etapa fundamental do trabalho com dados. Na prática, ela deve ocorrer de forma transversal, desde a obtenção até a visualização. Ou seja, ao realizar uma raspagem de dados, é importante que durante esta coleta você já verifique se as informações estão sendo capturadas corretamente. Na análise e visualização, o mesmo acontece. Faça a checagem dos dados e, se possível, compare-os com os resultados obtidos por outras fontes.

Já falamos sobre a importância de observar a documentação que acompanha as bases de dados. Este é o primeiro passo para começar a verificação: procure pelo dicionário de dados. Em alguns casos, esse documento pode vir com diferentes nomes: leia-me, nota técnica, nota informativa, anexo etc. Esse arquivo, que funciona como uma espécie de bula, incluirá explicações sobre as variáveis adotadas na estruturação das informações.

Nesse documento, idealmente, haverá também informações detalhadas sobre as unidades de medidas utilizadas (gramas, quilos ou toneladas, por exemplo) e os formatos adotados (número absoluto, índice, taxa, percentual etc.).

Dados, assim como humanos, são imperfeitos

Até mesmo uma fonte que conhecemos bem e tem as melhores credenciais possíveis pode se enganar: um especialista pode cometer um erro de avaliação, a testemunha de um evento pode se confundir e alguém que tenta relembrar um acontecimento histórico que vivenciou pode ser traído pela própria memória.

O mesmo vale para bancos de dados. Até algumas das melhores bases possíveis, mantidas e atualizadas por instituições relevantes e confiáveis, podem trazer problemas como esses. No caso das planilhas, essas falhas podem se apresentar como erros de digitação, entradas duplicadas e outros problemas do gênero. O papel do repórter é verificar se a informação que recebe é legítima antes de publicar, não importa se ela venha da boca de um especialista renomado ou de um banco de dados do governo.

Tomemos como exemplo um caso ocorrido durante a pandemia de Covid-19 no Brasil. Ao final da tarde, como de costume, o Ministério da Saúde divulgou uma imagem com os números atualizados de casos e mortes pela doença em todas os estados do país. No dia 20 de abril de 2020, foi publicado um dado que indicava crescimento substancial da epidemia em São Paulo: a tabela mostrava que as mortes passaram, em 24 horas, de 1.015 para 1.307 - um surreal crescimento de 30%, muito acima do que havia sido observado nos dias anteriores.

O número recorde chegou a aparecer na manchete de portais de notícias, mas estava errado. Por sorte, o crescimento fora do padrão chamou a atenção dos jornalistas mais atentos, que começaram a cobrar explicações do Ministério.

Resposta: o que aconteceu não foi um crescimento sem precedentes, mas um erro de digitação. O total de mortes não havia pulado de 1.015 para 1.307, mas sim para 1.037. O dado foi retificado.

Esse exemplo é excessivamente dramático, já que trata de uma estatística que estava no centro da agenda de notícias nacional e dificilmente passaria despercebido. Ainda assim, é ilustrativo de como apenas as credenciais oficiais não são suficientes para garantir precisão.

Há casos em que esses erros são mais difíceis de detectar. O *Basômetro*, ferramenta do Estadão, compreende dados de votos da Câmara dos Deputados para medir quanto apoio o Poder Executivo tem entre parlamentares. Em acesso às bases de dados das votações, desde 2003, foram percebidas algumas inconsistências: havia uma votação em que apareciam mais de 513 registros de deputados. Isso é impossível, já que a Câmara só tem esse número de membros. O que aconteceu, afinal? Alguns deputados, não se sabe o motivo, apareciam duplicados nos dados.

Pela lógica, cada sessão de votação precisava cumprir alguns requisitos para que os dados fossem válidos: não poderia ter registro de mais de 513 votos, não poderia ter mais de um voto registrado para um mesmo deputado e assim por diante. Apenas aplicando esses filtros à base de dados inteira, foi possível detectar e corrigir outras sessões com problemas semelhantes.

Biografando dados

Uma parte fundamental após a obtenção das informações é realizar uma espécie de **biografia dos dados**. Basicamente, você precisa entender minimamente a origem, a captação, o recorte e as limitações dos dados nas suas mãos. Isso pode parecer fácil, mas é preciso atenção. Um deslize nesta etapa e toda seu trabalho futuro com os dados estará comprometido. Por isso, nesta seção, iremos explicar um pouco deste método de verificação e checagem de dados.

Por que é importante biografar os dados antes de entrevistá-los?

Para seguir com a metáfora, vamos pensar no processo de apuração de uma reportagem tradicional, que não usa muitos dados quantitativos: o repórter, quando se encontra com a fonte, faz uma série de perguntas e usa as respostas para produzir a matéria. Isso é uma entrevista, obviamente.

No trabalho com dados, o equivalente acontece quando o jornalista começa a mexer em filtros e classificações para descobrir qual foi o maior gasto de um departamento do governo no ano, por exemplo.

Entretanto, a entrevista não é o começo de uma apuração jornalística. Antes de falar com qualquer pessoa, o bom repórter precisa fazer a lição de casa. Além de estudar o tema da conversa e preparar perguntas, ele precisa também descobrir o máximo sobre as características da pessoa com quem vai falar.

Quem é o entrevistado? Quais são seus interesses? Quais são suas áreas de especialidade e atuação? Ele tem alguma conexão com pessoas ou grupos que podem gerar conflitos de interesse? Existe alguma controvérsia relevante a seu respeito? Ele costuma ser ouvido por outros jornalistas? Ele tem um histórico confiável como informante? Ele é respeitado pelos seus pares? É preciso saber a resposta para essas perguntas antes de buscar informações com uma fonte.

O mesmo vale para um banco de dados. Tentar entrevistar uma planilha sem saber detalhes sobre ela é uma receita infalível para erros. Biografar dados é fazer essa pesquisa prévia, etapa necessária antes de qualquer tentativa de análise quantitativa.

E o que devemos procurar saber sobre uma planilha antes de partir para a entrevista? A cientista de dados Heather Krause dá um exemplo prático com base em um de seus temas de pesquisa mais recorrentes: violência contra a mulher.

O entrevistado da vez era um relatório público pela ONU em 2015. Uma fonte de confiança, certo?

Bem, vamos ver o que ela descobriu sobre a coleta de dados durante o trabalho de biografia e pensar um pouco sobre como esses detalhes podem impactar os números.

O primeiro passo foi analisar com atenção o **apêndice estatístico** do relatório. Esses documentos complementares são, geralmente, bem chatos de ler, mas reúnem informações metodológicas importantes sobre a coleta de dados de qualquer fonte.

Como dissemos anteriormente, você pode acabar esbarrando nele com outro nome: dicionário de da-

dos, metadados, anexo metodológico. Enfim, o que importa é que lá estão detalhes sobre os campos que aparecem na planilha, o que eles significam e como foram preenchidos. Exige força de vontade, mas é essencial superar a linguagem de bula de remédio desses documentos.

O que Krause descobriu nesse exercício de atenção e paciência? Que os métodos de coleta de informação variam de país para país e de ano para ano, o que dificulta comparações.

Exemplo: no Malawi, um país do sudeste africano, foi registrada uma variação inesperada nas taxas de violência de uma pesquisa para a outra, entre 2004 e 2005. O número subiu desproporcionalmente e, logo em seguida, voltou o cair.

O que aconteceu nesse intervalo? Será que houve um grande evento político, social e cultural que gerou esse pico? Parece uma boa história, não?

A explicação, porém, era bem menos empolgante. No ano do crescimento desproporcional, quem fez a coleta de dados foi uma instituição privada, com uma metodologia descrita apenas como "inovadora", sem mais detalhes.

Nos outros anos, a coleta foi feita por agências financiadas com recursos governamentais, que usavam uma metodologia mais tradicional e transparente.

A variação, portanto, é provavelmente efeito desse percalço na continuidade histórica dos dados e não consequência de uma mudança no mundo real. E os problemas não paravam por aí!

Além do uso de fontes diferentes, a data de coleta dos dados variava radicalmente entre os diferentes países. O relatório inclui informações de 2003 e 2013 sob o rótulo de "números mais recentes disponíveis", por exemplo.

O jornalista desavisado poderia, então, ao tentar contrastar dois países, acabar comparando uma realidade atual com outra que, provavelmente, não é mais relevante.

Essas limitações importantes estavam presentes em um relatório global da ONU, que é tida como uma fontes mais confiáveis para assuntos internacionais. Por sorte, os autores da pesquisa foram transparentes e incluíram essas possíveis falhas metodológicas no documento.

Entretanto, nem todo banco de dados é tão "honesto". Muitos dados públicos não vêm acompanhados de um detalhamento tão minucioso das próprias limitações. Muita vezes sequer vêm acompanhados de uma descrição decente do que significa cada um dos campos da planilha.

Diante de um cenário como esse, o trabalho do repórter inclui entrar em contato com os autores do levantamento para tirar todas as dúvidas possíveis. Como dissemos brevemente antes, a dica é nunca presumir o que um rótulo de coluna significa ou de que forma os dados foram reunidos. Sempre é preciso fazer esse trabalho prévio.

Krause resume o processo de biografia de dados com uma lista de perguntas que precisam de resposta antes da apuração seguir em frente:

Quem coletou os dados e por que os coletou? Fique de olho em possíveis conflitos de interesse e em outras formas que a autoria da pesquisa podem afetar os dados. Algumas são menos óbvias. Por exemplo, pesquisas sobre renda costumam ter resultados diferentes quando são feitas por instituições governamentais. Não é conspiração, mas sim desconfiança: as pessoas tendem a subestimar o quanto ganham, com medo de possíveis cobranças ou impostos.

Como os dados foram coletados? Os dados vêm de um censo que fala com todas as pessoas do país, batendo de porta em porta? São feitos com base em um recorte amostral da população? São coletados por telefone? São inseridos em um programa de banco de dados por policiais ou médicos? São consolidados automaticamente por um sistema automático? Tudo isso afeta a representatividade, a qualidade e as conclusões que podemos tirar dos números.

Quando os dados foram coletados ou atualizados? Uma pegadinha comum é que os dados divulgados em um ano mostram, na realidade, o cenário de outro momento. Um exemplo comum são os resultados de pesquisas eleitorais. Como, em época de eleição, a opinião pública costuma mudar rápido, a pesquisa representa o estado das coisas no dia de campo (ou seja, no dia em que os pesquisadores fizeram as entrevistas) e não na data de divulgação, que costuma ser até dois dias depois.

Tópicos de atenção

Para analisar os dados de uma planilha, é importante que você entenda o que significa cada linha e cada coluna da sua tabela. Imagine uma base de dados sobre Covid-19 no Brasil. O primeiro passo é entender o que ela registra: cada linha na tabela é um paciente atendido ou o total de casos de uma cidade, por exemplo? Depois de entender o que as linhas significam na sua tabela, é importante compreender cada coluna.

Uma coluna ou variável chamada “localidade”, por exemplo, pode representar pelo menos duas coisas: o município de residência de um cidadão ou a localidade onde ele foi atendido. Não raro, pessoas recebem atendimento de saúde fora do município onde vivem. Por isso, é preciso ficar atento à documentação e à descrição do significado das variáveis de uma planilha.

Se sua base for muito grande para você fazer checagens de todos os registros, uma dica é separar uma amostra aleatória dos seus dados e verificar se está tudo certo. É importante que a seleção seja aleatória. Por exemplo, se você checa apenas as primeiras linhas de uma tabela com registros em ordem cronológica, problemas que tenham ocorrido mais próximos ao fim do período podem passar despercebidos.

Completude e coerência

A identificação da eventual falta de registros em um banco de dados deve ser imediatamente verificada junto à fonte original. Exemplo: você obteve os dados do Ministério da Saúde sobre os contaminados pela Covid-19 nas 27 unidades federativas do Brasil e, ao checar se todas elas estão presentes

na base de dados, constata que um estado ou um município está ausente. Reflita: o que tal ausência significa? Significa que ninguém foi contaminado ou que os municípios ou estados não foram analisados?

Se você receber soma, dados agregados ou operações matemáticas já realizadas, mas também tiver acesso aos dados desagregados, a recomendação é refazer e checar os cálculos. Especialmente se a base de dados foi elaborada manualmente, erros neste sentido não são incomuns.

Para entender melhor a importância desse processo, aí vai uma história real. Em uma redação de jornal, dois profissionais estavam produzindo uma matéria sobre filiações partidárias e as migrações de pessoas entre partidos. Enquanto um jornalista de dados ficou responsável pela a obtenção dos dados na página do Tribunal Superior Eleitoral (TSE), via web scraping, outro tinha a missão de fazer as análises a partir dos dados coletados.

Quando as perguntas foram respondidas e o texto e a visualização estavam prontos, foi notada a ausência de dois partidos na análise. Consequentemente, as duas siglas estavam ausentes na reportagem e na infografia. O fato provocou estranheza: será que nenhuma pessoa dessas duas legendas políticas nunca trocaram de partido? Onde estão seus filiados?

Após uma conversa entre os profissionais, eles resolveram checar cada etapa do processo e concluíram que havia um erro no processo. Em função dos descuidos, a reportagem estava comprometida, uma vez que retratava as migrações de filiados de um partido ao outro utilizando tanto números absolutos quanto percentuais. Portanto, a ausência das duas legendas alterava, também, os cálculos realizados para os demais partido. Por sorte, porém, a matéria estava "pronta" com antecedência, o que permitiu refazê-la antes do envio do conteúdo para a gráfica responsável pela impressão do jornal.

Este caso permite refletir sobre o que poderia ser feito para evitar o erro e a consequente repetição de procedimentos. O desfecho seria diferente se houvesse uma verificação básica que mostrasse a presença de todos os 33 partidos na base de dados obtida via web scraping. O erro ocorreu duas vezes: na ausência de verificação da planilha recebida e, antes disso, na falta de detecção do problema na hora da captura dos dados, quando uma primeira verificação deveria ter sido feita.

A história se torna ainda mais complexa porque o código que havia sido desenvolvido para a raspagem e obtenção dos dados fora usado para uma reportagem anterior, que não apresentou erro em relação ao número de partidos.

O que mudou, então?

Os responsáveis pelo site do TSE alteraram a grafia das siglas desses partidos, fazendo com que o código, que já estava pronto, não conseguisse mais identificar as siglas. Por isso, **muito cuidado se for reutilizar um código**. Às vezes os responsáveis pelas páginas na web fazem pequenas alterações que podem comprometer a obtenção de dados recentes.

Este exemplo também mostra que a checagem das informações pode (e deve!) ser aplicada também na etapa de obtenção de dados, de modo a mitigar possíveis problemas posteriores. Ao fazer raspa-

gens de dados, faça sempre uma checagem dos dados em todas as etapas para garantir que todas informações estão sendo capturadas adequadamente.

E mesmo dados estruturados em planilhas podem vir acompanhados de muitos problemas. O jornalista de dados precisa ser minucioso e desconfiado. O trabalho com as bases de dados não admitem fios soltos e, por isso, o trabalho de verificação e limpeza inclui identificar potenciais problemas para eliminá-los imediatamente.

Valores anormais

Uma das primeiras etapas na checagem de dados com números ou valores é a identificação dos chamados **valores extremos ou outliers**. Números que destoam muito dos demais podem ter dois significados: serem de fato casos especiais que merecem investigação ou serem frutos de erros.

Em uma análise sobre a despesa de parlamentares com diárias, por exemplo, é interessante observar se um deles se sobressai em termos de gastos públicos em relação aos seus colegas. Em uma prestação de contas de campanha ou declarações de bens de políticos, o mesmo pode acontecer.

Números extremos, seja para mais ou para menos, podem ser chamativos. Lembre-se, porém, de verificar as informações antes de seguir uma possível história: o que ocorre, às vezes, é o preenchimento incorreto de informações. Na declaração de patrimônio de candidatos a um cargo eletivo, por exemplo, a simples adição do número zero infla o valor dos bens declarados. Por isso, a checagem dos valores extremos ou anormais nas planilhas é uma etapa fundamental não só da análise mas também da checagem da integridade dos dados.

Unidades de medida

Fique atento quanto às unidades de medida (milhas, quilômetros, pés) ou moedas (real, dólar, peso, libra e outros) utilizadas em um banco de dados. Nem sempre essas informações são padronizadas. Para descobrir qual é o caso, leia cuidadosamente a documentação. Não presuma, por exemplo, que todas as planilhas criadas no Brasil estão estruturadas em reais.

Se estiver trabalhando com dados de diversas fontes, escolha a unidade de medida ou moeda com a qual pretende trabalhar e realize as conversões necessárias. Se possível, acrescente colunas e sinalize as alterações.

Verifique, sobretudo, em que formato estão os números da base de dados com a qual pretende trabalhar. Em alguns casos, não haverá casas decimais. É o que ocorre com o número de nascimentos, mortes e outros: não é possível que haja 3,2 óbitos ou 10,8 nascimentos, por exemplo. Se uma coluna reúne informações sobre *pageviews* (número de acessos), da mesma forma, não pode conter números decimais.

Limpe

A **limpeza de dados** é uma etapa fundamental de qualquer trabalho com dados. Em geral, ela não é muito empolgante, mas pode tomar um tempo considerável do seu trabalho. Aqui, você irá organizar os dados para que seja possível posteriormente realizar análises, operações matemáticas, filtros, ordenações, enfim, tudo que você precisa para responder às perguntas desejadas.

Ao limpar seus dados, você irá alterar as informações originais da planilha. Por isso, é importante tomar nota das alterações, bem com manter uma cópia da versão original. Em caso de problemas com seu trabalho de verificação e limpeza, você sempre poderá recomeçar do zero.

Caso você utilize linguagens de programação, o próprio código já serve como esta documentação. Além disso, há os chamados notebooks, que costumam ser ótimas ferramentas para documentar códigos e processos de limpeza e análise dos dados. Eles também permitem que você acrescente comentários e lembretes para si ou para seus colegas. Entre os notebooks mais conhecidos estão o Jupyter Notebook e o Colab do Google.

Antes de começarmos a listar os problemas mais comuns para limpeza de dados, é importante que você conheça o conceito de **dados organizados** (tidy data). Trata-se de um forma de organizar dados tabulares, proposto por **Hadley Wickham**, um dos desenvolvedores de maior destaque da comunidade de R.

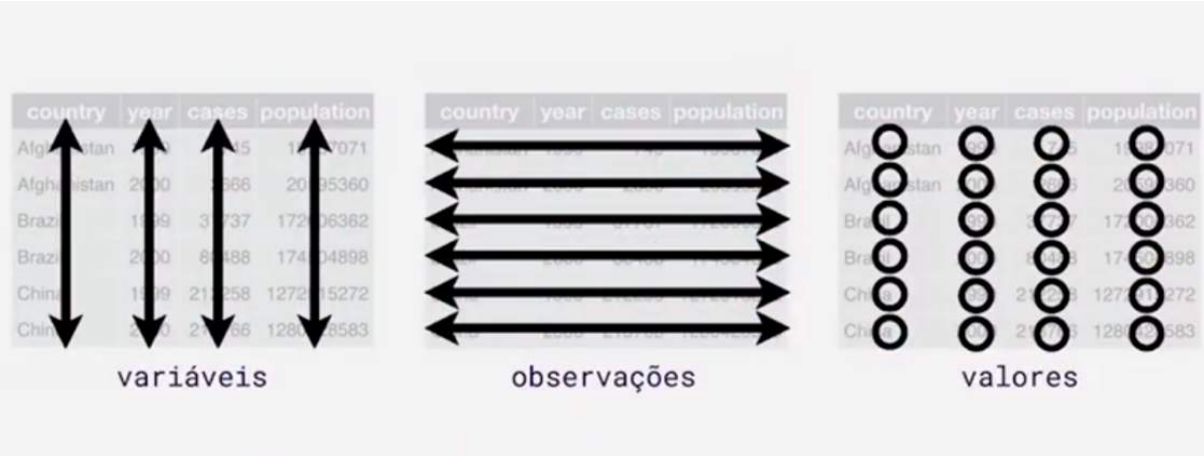


Figura 16: No modelo do “tidy data”, cada variável deve ter sua própria coluna, cada observação deve ter sua própria linha e cada valor deve ter sua própria célula.

Apesar do tempo que cientistas de dados passam limpando a matéria-prima do seu trabalho, ele notou que há pouca reflexão sobre este processo, ao contrário do que acontece com etapas como a análise e visualização de dados, onde há um largo acúmulo de reflexões teóricas e práticas. Por isso, ele concebeu a ideia de tidy data.

De acordo com esta noção, cada tabela registra um certo fenômeno (despesas de um órgão do governo, as notas de alunos, resultados de jogos de futebol ou da situação de pacientes de um hospital, por exemplo). Cada ocorrência será registrada em uma linha ou em uma observação. Cada característica deste fenômeno será uma variável ou coluna e, por fim, a descrição destas características serão chamadas de valores e ocuparão as células.

A seguir, um exemplo de uma tabela fora do formato tidy data.

Nome/Registro	**Turma**	**Biologia**	**Química**	**Física**
Marcela Nunes - 20351	A	7	6	6
Ricardo Fernandes - 20589	C	5	4	7

Na tabela, o fenômeno capturado são notas de alunos. Mas repare que a variável "disciplina" ocupa diferentes colunas. No formato tidy data, essa planilha deveria ter uma coluna chamada "disciplina", onde as matérias estivessem declaradas nas células abaixo dessa coluna.

Outro problema que pode ser visto na tabela são as linhas referentes à coluna Nome/RA. Nesse caso, há duas informações (valores) em uma mesma célula: o nome do estudante e o registro do aluno (um código identificador).

Porém, percebemos que os nomes dos alunos estão separados dos registros por um hífen. Deste modo, podemos usar esse padrão para separar estes valores em colunas ou variáveis distintas, utilizando editores de planilha ou linguagens de programação, por exemplo. Como dissemos, o reconhecimento de padrões que permitam a limpeza dos dados será o seu maior aliado nesta etapa.

Para Hadley Wickham, todas estas questões precisam ser corrigidas antes de qualquer trabalho de análise ou visualização de dados. Assim, a tabela deveria ficar assim para atender o formato do tidy data. Com a organização dos dados, as disciplinas se tornaram uma única coluna, com diferentes valores. Além disso, há uma linha para cada observação (ou seja, notas de aluno em disciplinas) e o nome e número de registro do aluno também foram separados.

Nome	**Registro**	**Turma**	**Disciplina**	**Nota**
Marcela Nunes	20351	A	Biologia	7
Marcela Nunes	20351	A	Química	6
Marcela Nunes	20351	A	Física	6
Ricardo Fernandes	20589	C	Biologia	5
Ricardo Fernandes	20589	C	Química	4

Nome	Registro	Turma	Disciplina	Nota
Ricardo Fernandes	20589	C	Física	7

Ferramentas e técnicas de limpeza

Na etapa de limpeza dos dados, mais uma vez, é possível usar tanto linguagens de programação ou softwares com interfaces gráficas. Para trabalhar com planilhas simples e constituídas por poucas linhas, softwares de editores de planilhas (Excel, Google Sheets ou LibreOffice, por exemplo) correspondem à principal solução para quem não domina linguagens de programação.

Mas há também programas úteis para a limpeza de dados, que são gratuitos e possuem interfaces gráficas. Eles irão ampliar em muito suas habilidades na padronização de dados. Nesta seção, falaremos mais sobre estas soluções e, a seguir, conheceremos uma linguagem para descrição de padrões, que podem ser aplicados em diferentes programas e plataformas: as expressões regulares.

Expressões regulares

Uma expressão regular, também conhecida como Regex, é uma forma comum e concisa de representar algum tipo de padrão em texto. Isso ocorre em uma variedade de aplicações e linguagens, seja pesquisando em um documento, seja fazendo operações de *Localizar e Substituir* ou limpando um conjunto de dados. Muitas linguagens de programação e editores de texto/planilha permitem o uso de expressões regulares. As sintaxes, por sua vez, podem diferir levemente de acordo com a linguagem ou o programa em uso.

As expressões regulares podem parecer enigmáticas e confusas à primeira vista, mas não se assuste. O importante é entender que as letras e caracteres podem ter significados especiais. Um exemplo bastante simples é:

```
[0-9]
```

Esta é uma expressão regular bem simples, que seleciona todos os números entre 0 e 9 em uma sequência de caracteres. Poderia ser útil se você tem um campo com textos e números e precisa separá-los, por exemplo.

Se você precisa descrever padrões para obter informações específicas de sequência de texto, provavelmente é o caso de começar a aprender expressões regulares. Explorar a fundo esta sintaxe exigiria um guia por si só. Mas se você quiser ter uma introdução ao funcionamento de expressões regulares, vale conferir este tutorial da Escola de Dados e utilizar sites como Regex ou Regex101 para testar as operações na prática. Eles podem ser úteis também para "traduzir" expressões regulares que você encontrar na internet, indicando o significado de cada caractere.

Programas úteis

A **expressão regular** é uma linguagem "universal", que você poderá utilizar tanto em editores de planilha, quanto ao escrever códigos com Python e R. Ela também pode ser utilizada em outros tipos de programas, que possuem funções específicas para a limpeza de dados.

Nesta seção, iremos abordar duas destas ferramentas: o OpenRefine e o Workbench.

O **OpenRefine** é uma ferramenta voltada à limpeza de dados. Ele consegue carregar arquivos em diferentes formatos (inclusive JSON), transformar para tabelas e então aplicar diversas operações, como a remoção de duplicatas, criação ou remoção de linhas e colunas, normatizações, correção de ortografia entre outras funções.

Além disso, o OpenRefine mantém um "log" com todas as operações que foram feitas nos dados originais. Este registro é extremamente útil, caso você precise refazer uma operação ou reaplicar as transformações realizadas em outros dados que seguem os mesmos padrões.

O OpenRefine também conta funcionalidades muito práticas, que permitem padronizar textos com erros de digitação ou grafias distintas. Ele permite até mesmo a consulta a APIs mais simples, que você pode utilizar para preencher novas colunas a partir de dados existentes. O OpenRefineé uma ferramenta de código-aberto, que conta com versões para Windows, Mac e GNU/Linux. Para utilizá-lo, você irá precisar instalá-lo no seu computador.

Se você preferir utilizar uma solução online, pode utilizar o **Workbench**. Na realidade, esta plataforma é uma ferramenta útil para todas etapas do trabalho com dados, não só a limpeza.

Nela, você pode carregar dados em tabelas, realizar raspagens de dados ou carregar informações do Twitter, por exemplo, para em seguida realizar transformações de limpeza ou operações de análise e visualização dos dados. Na área de limpeza, atualmente, o Workbench oferece funções para remoção de colunas duplicadas, preenchimento de células nulas, geração automática de colunas com identificadores únicos, união e separação de colunas, extração de valores com expressão regular, entre outras funcionalidades.

Tópicos de atenção

Nesta seção, veremos alguns tópicos que merecem atenção especial na hora de limpar dados, independente da ferramenta ou software que você irá utilizar. Para uma listagem mais completa sobre o tema, vale consultar o Guia Quartz para Limpeza de Dados, que traduzimos para o português na Escola de Dados.

Ausência de valor

Um problema comum nos conjuntos de dados são a ausência de valor, ou um valor vazio, que pode ser representado de diferentes formas. O nulo (null), a célula em branco, o NA (*not available*) e o número

zero podem significar coisas completamente diferentes, dependendo de quem coletou esses dados. Ainda que na programação cada uma dessas representações tenha significados distintos, é preciso verificar.

Para o aprendizado introdutório, é importante entender que qualquer um desses valores em uma planilha pode significar 0. Em outra ocasião, porém, o 0 ou a ausência do valor representa que aquele dado não foi coletado ou que foi perdido. Busque mais informações contextuais junto à fonte original. Diante disso, será possível avaliar se a ausência de valor será mantida ou excluída da análise.

Entradas duplicadas

Um dos erros mais comuns em bases de dados é quando uma entrada aparece duas vezes por engano. Verifique se há duplicatas exatas no arquivo: em caso positivo, pode ser que você esteja diante de um erro. Porém, só é possível saber se um registro duplicado é um erro ou não conhecendo bem as "regras de negócio" por trás da base de dados.

Considerando um exemplo sobre a Covid-19, poderíamos pensar que determinado estado ou cidade apresenta duas linhas de informação, quando o correto é apenas uma. Se a linha aparecer duas vezes em um conjunto de dados, é preciso entender o motivo. É um erro ou uma atualização/complementação? Atente também às planilhas que envolvem dados financeiros e transações. Vale contatar a fonte primária dos dados e entender o que houve.

Cabeçalho com informações inúteis

Alguns órgãos públicos oferecem, de praxe, informações técnicas no cabeçalho e no rodapé das planilhas (exemplo: nome do órgão responsável, data da coleta, explicações sobre a metodologia). Embora esse conteúdo auxilie a dar contexto ao material coletado, prejudica a análise (ao fazer uma simples ordenação, por exemplo) e precisa ser removido. Por isso, antes de começar a editar as informações, faça uma cópia do arquivo original - as informações extras podem servir para tirar dúvidas e complementar o material.

Grafia distinta

Ao comparar dados, certifique-se de que nomes de cidades estão registrados com a mesma grafia. Por exemplo, diferentes linhas podem conter a seguinte variação da cidade mais famosa do mundo: Nova Iorque, Nova York, New York, NY - ou, ainda, erros de digitação, como "Nova Yok" e tantas outras possibilidades. Para que possa fazer a análise, é preciso corrigir as grafias estabelecendo uma única versão (as expressões regulares podem ser particularmente úteis aqui). Do contrário, operações realizadas com esses registros podem ficar comprometidas. Não se esqueça de documentar as mudanças.

Atenção para nomes de pessoas: nunca faça correções em nomes de pessoas caso não tenha certeza absoluta de que se trata de um erro. Nomes podem ter variações de grafias que podem não acompanhar sua imaginação. Por isso, não altere se não estiver absolutamente seguro.

Analise

Você já checou como os dados foram coletados e divulgados e não encontrou nenhum problema. Também já entrou em contato com os responsáveis pelo levantamento e os números parecem responder aos mais rigorosos critérios de qualidade. Além disso, você verificou se os arquivos tinham alguns erros comuns e tudo parece estar conforme o esperado. Agora é só sucesso, certo? Nem sempre.

O maior risco de erro vem de onde menos prestamos atenção: nós mesmos. Nossos próprios vieses e eventuais falhas nas premissas que adotamos na hora de analisar um banco de dados são um estoque inesgotável de cascas de banana.

Philip Meyer, criador do jornalismo de dados, propõe uma reflexão que ajuda a entender esse problema.

Em seu livro de 1973, "*Precision Journalism*", ele discute a maneira como repórteres encaram a missão de reportar fatos com objetividade. De acordo com Meyer, jornalistas costumam perceber a si próprios como pessoas de mente aberta, capazes de embarcar em investigações como "tábulas rasas", sem qualquer prejulgamento. O autor aponta, porém, que isso é impossível. Na prática, não dá para começar a pensar sobre qualquer problema sem partir de um enquadramento teórico ou de uma hipótese. Nós começamos a investigar um assunto porque achamos que pode existir algo de interesse ali, afinal.

Paul Bradshaw, professor da Birmingham City University, publicou recentemente um artigo relacionado ao tema. O texto explora como diversos vieses cognitivos afetam o trabalho jornalístico. Sem jargão, isso significa o seguinte: existem vários mecanismos programados na mente humana que fazem com que nossa avaliação da realidade não seja tão isenta e objetiva como gostaríamos que fosse.

Um dos vieses listados é tão importante que mereceu um texto só para ele: o viés de confirmação. Em resumo, humanos tendem a prestar mais atenção em informações que reafirmam suas próprias opiniões sobre um tema, enquanto ignoram informações que possam colocar essas perspectivas em xeque. O viés de confirmação também é frequentemente acionado nas crenças de fatos duvidosos, como ocorre no circuito da desinformação.

Essa questão não surge exclusivamente no jornalismo de dados, mas também pode ser observada nas reportagens que não se utilizam de quantificações. Muitas vezes, na ânsia de ver uma investigação se confirmar, repórteres direcionam a atenção apenas para os elementos que corroboram o resultado

que querem alcançar e ignoram evidências contrárias. Isso acontece porque o repórter tem estímulos positivos para confirmar as próprias suposições, desde a satisfação de ver seu faro estar certo até a pressão que envolve apurar, escrever e entregar uma matéria relevante dentro de um prazo apertado.

Assim como acontece com as pessoas do outro lado do balcão, que produzem dados para nosso consumo, jornalistas de dados também podem falhar: erros de digitação, fórmulas mal aplicadas, a mão que escorrega na hora de aplicar uma função na planilha… Tudo isso pode acontecer e passar despercebido.

Entretanto, o próprio Philip Meyer propõe uma saída para o problema que levantou - e ela funciona também para minimizar tanto os efeitos dos vieses cognitivos quanto os deslizes mundanos. Para ele, o *jornalismo de precisão* deveria adotar, na medida do possível, os ideais, os métodos e o conceito de objetividade dos cientistas.

O que isso significa, porém?

De saída, significa formalizar, enunciar e tomar consciência das hipóteses, teorias e premissas que assumimos na hora de apurar uma matéria - ou, no nosso caso, de fazer uma análise de dados. Só assim é possível fazer uma análise mais criteriosa dos pressupostos que envolvem nossa forma de pensar e as conclusões que derivam dela.

Além disso, é importante adotar o **maior rigor metodológico possível.** Antes de mergulhar nos números, vale listar quais são os elementos que você procura, que evidências seriam necessárias para comprovar a hipótese que você investiga e, em contraste, o que seria necessário para admitir que não há nada ali. A ciência estatística pode ser aliada nesse processo, como veremos adiante.

Por fim, Meyer também destaca a importância de assumir uma postura de **transparência radical** , semelhante àquela que os bons cientistas adotam ao divulgar seus estudos de forma que possam ser meticulosamente avaliados e até reproduzidos pelos pares.

Explico melhor: cientistas estão acostumados a registrar todos os seus passos de forma detalhada. Junto com a conclusão de seus experimentos, publicam também uma descrição detalhada do método utilizado, os dados relacionados e um passo a passo de como os resultados foram obtidos.

Assim, um trabalho científico é também uma espécie de "mapa" que serve para que colegas (na revisão entre pares, por exemplo) e o público em geral possam verificar se tudo está como deveria estar ou se há alguma falha no estudo.

O jornalismo investigativo tradicional faz isso em certa medida, quando reúne documentos e registros concretos que podem ser verificados: atas de reunião, contratos, editais… Uma reportagem de impacto geralmente cita e mostra elementos como esses, o que ajuda a conferir credibilidade ao relato.

A coisa fica mais complicada, porém, quando o trabalho envolve informações confidenciais, fontes anônimas ou eventos que o repórter testemunhou em primeira pessoa. Algumas reportagens, pela própria natureza, não podem vir acompanhadas da demonstração de evidências.

O jornalismo de dados, por sua vez, apresenta condições de oferecer **transparência** ao leitor. Como o trabalho costuma ser mais metódico e controlado, é possível documentar cada passo de uma apuração ou análise.

Quando o repórter sabe programar, ele pode divulgar o **código-fonte** de uma matéria, ou seja, as linhas de instruções que foram executadas pelo computador para chegar ao resultado. Assim, pessoas que também entendem de código podem revisar o processo de elaboração do material em busca de erros. Este processo já é adotado na redação de grandes jornais no Brasil e no mundo, como o Estadão ou o The New York Times.

Mesmo quando o trabalho não envolve programação, é possível fazer algo semelhante através do registro sistemático de todas as ações que foram tomadas no processo de produção.

Vale a pena anotar tudo o que você fizer, desde a origem dos arquivos que baixou até o passo a passo da limpeza de dados e das fórmulas e filtros que aplicou. Além de servir para organizar melhor o próprio fluxo de trabalho, esse tipo de material pode ser divulgado junto com a matéria.

Ao permitir que terceiros avaliem se seu métodos são sólidos, você também minimiza a chance de que um erro passe despercebido por todos. Mais do que isso, você demonstra que está trabalhando de boa fé e constrói uma relação de confiança com os leitores. Quanto mais gente prestar atenção e discutir os méritos e falhas do jornalismo, mais a qualidade dos conteúdos tende a crescer.

Dicas para começar uma análise

Desenvolva um método: em vez de mergulhar na base de dados de forma livre, olhando para todos os lados em busca de algo que possa ser interessante, é importante elaborar um plano de ação. Assim, nos obrigamos a tomar consciência das hipóteses e premissas que envolvem o trabalho. O que exatamente você espera encontrar? De que maneira vai sistematizar essa busca? Que evidências seriam suficientes para corroborar suas hipóteses?

Reporte contra suas convicções: quando começamos uma análise de dados, é natural que o esforço de investigação enfatize elementos que corroborem a história que queremos contar. Entretanto, precisamos pensar também naquilo que pode derrubar a matéria. Que evidências fariam você desistir da reportagem? Procure por elas também.

Seja transparente: desenvolver um método claro de trabalho não é útil apenas para guiar nossos próprios esforços de apuração. Quando a metodologia é publicada junto com uma reportagem, você ajuda o público a entender quais foram os passos da investigação, o que aumenta a credibilidade do material e permite que interessados verifiquem se os resultados são mesmo sólidos.

Vieses e limitações

Agora, já sabemos que dados podem trazer erros e vieses e que esses problemas podem ter várias causas, desde falhas metodológicos até pequenos deslizes humanos.

Falta entender, entretanto, como esses dados ruins podem afetar a sociedade de forma negativa - e o que podemos fazer diante disso.

Em 2016, o veículo americano *ProPublica* investigou como um programa de inteligência artificial usado pelo sistema de justiça americano discriminava pessoas negras.

Essa inteligência artificial supostamente era capaz de dizer qual era a chance de um prisioneiro reincidir no crime caso fosse solto e, assim, ajudar juízes a tomarem decisões melhores. O programa avalia dados sobre o réu e emite uma nota, em uma escala de 1 a 10, para mensurar o risco. O juiz, depois de consultar o valor, decide o que fazer: concede liberdade condicional ou mantém o sujeito na prisão?

A promessa de um sistema como esse é reduzir a influência do preconceito inerente aos humanos no processo de decisão. Entretanto, a reportagem mostrou que isso não aconteceu.

De acordo com a análise, o programa sistematicamente classifica americanos negros como mais propensos a reincidir do que americanos brancos, ainda que tenham cometidos crimes menos violentos.

Por que isso acontece? Porque uma inteligência artificial precisa de dados para operar. Um sistema como esse funciona ao analisar quantidades enormes de números, buscando padrões neles. É com base em registros e métodos elaborados por humanos que o computador cospe seu veredito.

Agora o problema fica mais aparente: de acordo com um dos procurados citados na matéria, a sociedade americana, e o sistema prisional em especial, tem um histórico grave de racismo.

O programa não era alimentado com dados sobre a raça dos réus, mas se debruçava sobre dados sócio-econômicos e comportamentais, como renda e desemprego. Esses problemas não afetam pessoas de diferentes raças igualmente: negros sofrem mais com eles. Assim, a máquina passou a enxergar uma relação indireta - e inexistente, como a reportagem demonstra - entre raça e risco de reincidência.

Quando um sistema analisa automaticamente números produzidos por uma sociedade que discrimina, o resultado da análise vai trazer consigo o mesmo tipo de discriminação. Os dados são, em sua origem, essencialmente humanos.

A reportagem da ProPublica, além de revelar essa dinâmica que perpetua desigualdades, mostra como o jornalismo de dados pode agir diante de um contexto como esse.

O trabalho usou técnicas do jornalismo de dados não apenas para encontrar tendências em uma base, mas para entender melhor os efeitos que ela tem sobre o mundo. Aqui, o dado não é encarado como espelho da realidade, mas como uma construção que tanto reflete quanto afeta a sociedade.

Além disso, as melhores práticas foram seguidas e o resultado foi publicado junto com uma descrição detalhada da metodologia. A matéria serve para ilustrar como é possível tratar os dados não apenas

como uma fonte de informação, mas como um tema que merece ser investigado por si só.

Outro exemplo dessa postura vem do trabalho da Folha de São Paulo na cobertura da epidemia de Covid-19: a equipe do núcleo de dados do jornal (Delta Folha), ao analisar dados sobre mortalidade publicado pelos cartórios do Brasil, percebeu uma série de erros que impediam fazer uma análise sólida.

Em vez de simplesmente desistir da reportagem e começar outra matéria, o jornal publicou um texto em que expôs os diversos problemas daquelas estatísticas, que vinham sendo usadas cada vez mais por outros participantes do debate público.

O The New York Times também tem se destacado nesse aspecto, com uma cobertura dedicada aos efeitos que a coleta em massa de dados pode ter para a privacidade dos cidadãos e suas consequências para a democracia.

Essa atitude inquisitiva é especialmente importante porque, cada vez mais, dados e programas de computador têm efeitos na vida das pessoas. O tema tem se tornado mais caro ao jornalismo, tanto que alguns repórteres já se propõem a fazer a **cobertura de algoritmos**. Em vez de se especializar em reportar sobre política, cultura ou saúde, eles se dedicam a entender melhor e discutir de forma crítica como esses mecanismos atuam. O tema não cabe no escopo do livro, mas vale registrar que técnicas do jornalismo de dados podem ser úteis para quem quer se aventurar por esse campo.

Agora que já dedicamos uma boa atenção às preocupações que precisamos ter durante a análise de uma base dados, vamos começar a parte mais prática deste capítulo. **Como, afinal, fazemos para descobrir quais são as histórias que se escondem em uma série de números? Quais são as técnicas e saberes que eu preciso mobilizar para produzir uma boa reportagem guiada por dados?**

A primeira recomendação é não esquecer daquilo que importa para todos os repórteres do universo: seu trabalho é encontrar e comunicar o que há de mais novo, relevante e interessante no mundo. O que muda é que um bom jornalista de dados precisa desenvolver habilidades para encontrar informações em um novo ambiente social.

Uma das melhores explicações sobre o que isso significa veio de Tim Berners Lee, pesquisador e criador da internet moderna (World Wide Web).

Em uma entrevista de 2010 para o jornal britânico The Guardian, Lee resumiu a questão: segundo ele, jornalistas estão acostumados a encontrar histórias conversando com pessoas, e vão continuar fazendo isso. Entretanto, em mundo cada vez mais digitalizado, eles também vão encontrar histórias em bancos de dados, e portanto precisam estar equipados para analisá-los.

Esse pensamento se materializa quando percebemos que, hoje, praticamente todas as ações humanas acabam se tornado um registro numérico, no final das contas. Quando um parlamentar compra uma refeição, o registro vai parar em uma planilha em algum canto do site da Câmara. Quando o Presidente da República publica um tweet, um banco de dados é atualizado. Quando alguém faz uma busca no Google, um algoritmo é posto para funcionar.

E qual é o equipamento que um jornalista de dados precisa ter em mãos para compreender e usar esses dados, afinal? **O mais essencial é desenvolver uma maneira quantitativa de pensar. Em vez de enxergar anedotas e histórias individuais, um repórter treinado para trabalhar com informação quantitativa busca ver padrões, tendências, repetições**.

Para deixar mais claro como essa mentalidade funciona, vamos pensar em um acontecimento frequente: um roubo de celular que aconteceu em uma rua movimentada de uma grande capital.

Um único roubo, por si só, provavelmente não seria noticiado por um jornal da cidade. É corriqueiro demais, comum demais. Entretanto, o bom repórter de dados não para em um roubo só. Ele percebe que esse fato singular está inserido dentro de vários outros. Quantos celulares são roubados por dia em São Paulo, por exemplo?

Além disso, ele percebe que o roubo é um fenômeno quantificável: a ação aconteceu em determinado dia da semana e em determinado horário, em um local que pode ser representado como coordenadas geográficas. A vítima tem uma idade, uma raça, uma profissão. O ladrão também. Tudo isso cabe muito bem em uma planilha, não acha?

Quando jornalistas e editores percebem essa possibilidade, a história sobre roubo de celular fica interessante. Vira até capa de jornal: o exemplo é real, e resultou em uma matéria que foi publicada pelo Estadão em setembro de 2017. A planilha com todas essas informações sobre os roubos existe, de fato: é o registro dos boletins de ocorrência publicado pela Secretaria de Segurança Pública de São Paulo.

A matéria ecoa o texto clássico sobre jornalismo de dados, publicado em 2006 por Adrian Holovaty e sobre o qual falamos na introdução deste livro.

Depois que você aprende a enxergar números em todos os lugares, precisa aprender também a entender o que eles revelam. E é aí que entra no jogo algo que parece assustador: a matemática.

Dificilmente um jornalista ou comunicador entrou na profissão porque gosta de fazer contas. Na verdade, o que acontece com frequência é o contrário: muita gente escolhe virar repórter porque quer evitar os números, geralmente depois de passar anos sofrendo para superar provas de matemática.

O resultado dessa aversão generalizada à matemática é perceptível no ambiente de trabalho, visto que algumas pessoas entram em desespero ao precisar fazer uma simples conta de regra de três.

O fenômeno da ansiedade causada pela matemáticajá foi descrito em pesquisas acadêmicas e a conclusão assusta: muitos estudantes deixam de cursar disciplinas de jornalismo de dados porque acham que não gostam ou não sabem o suficiente de matemática.

Esse medo, contudo, não faz muito sentido: quase tudo que um jornalista de dados faz envolve matemática básica. Além disso, ao contrário da escola, podemos usar calculadoras ou, como é mais frequente, programas como o Google Sheets e o Microsoft Excel.

Antes de aprender como mexer com essas ferramentas, porém, precisamos relembram algumas tare-

fas simples dos tempos de escola, como taxas, porcentagem, estatística básica e afins.

Operações matemáticas básicas

Nesta seção, aprenderemos um pouco sobre como calcular porcentagem e taxas, além de ter algumas dicas relacionadas à correção de valores pela inflação e análise de séries temporais. Para uma descrição mais detalhada das operações matemáticas básicas importantes para o trabalho com dados vale conferir o livro de Sarah Cohen, *"Numbers In The Newsroom"*. Trata-se de um manual clássico escrito por uma ex-editora de jornalismo de dados do New York Times, que apresenta os conceitos da matemática que mais são usados nas redações de forma didática.

Porcentagem

Calcular a porcentagem é, provavelmente, a tarefa mais comum de um jornalista. Manchetes como "65% dos entrevistados se preocupam com o desemprego" são extremamente comuns e não costumamos pensar muito no que elas significam. É óbvio, certo?

Pode até ser, mas vale a pena esmiuçar o conceito. Quando alguém menciona 65% dos entrevistados, na prática está dizendo que 65 de cada 100 deles demonstrou preocupação com o desemprego.

Entretanto, a descrição metodológica dessa pesquisa diz que foram entrevistados, na verdade, 6.734 brasileiros e não 100. Desses, 4.377 disseram que estavam preocupados, em vez de 65. Nesse contexto, por que estamos falando de "65 de cada 100" em vez de "4.377 de cada 6.734"?

A explicação é simples: é mais fácil entender o que significa "65 de 100" do que "4.377 de 6.734". No primeiro caso, fica bem mais fácil entender que estamos falando de mais da metade do total. Ao falar em "65%", a proporção dos preocupados fica mais clara.

É para isso que serve a porcentagem: o cálculo parte de um valor arbitrário e difícil de compreender e o transforma em algo mais palatável. O passo a passo é o seguinte:

Então, chegamos ao resultado de 64,9% (na prática, os 65% mencionados)

Agora que já entendemos o que é uma porcentagem, podemos dar um passo adiante e falar de variação percentual. Veja este exemplo real de título: "Com Bolsonaro, liberação de agrotóxicos cresceu 42%, diz estudo".

Vamos aos números: de acordo com a reportagem, até abril de 2019, o governo de Jair Bolsonaro havia aprovado o uso de 166 novos agrotóxicos. No mesmo período de 2018, sob outro presidente, haviam sido 117. Como fazemos para chegar nessa variação de 42%, então?

Primeiro, precisamos calcular a **diferença** entre a aprovação de agrotóxicos em 2019 e 2018. Depois, queremos saber quanto essa diferença representa do **valor original**. Na prática, é muito parecido

Figura 17: Para calcular o percentual neste exemplo, primeiro, efetuamos esta divisão

Figura 18: Em seguida, multiplicamos a taxa obtida por cem, afinal, queremos o percentual

com o cálculo simples de porcentagem, mas com um passo a mais.

O que todas essas contas fazem, em sua essência, é descobrir **o quanto a mudança nos números brutos** representa do valor inicial. **Discernir isso é importante para não cair em um erro comum: confundir variação percentual e diferença em pontos percentuais**.

Esse último é mais fácil de entender: ele não envolve as transformações de números em percentuais que acabamos de fazer, mas sim a comparação já entre duas porcentagens.

O exemplo desta vez é a pesquisa Datafolha de 28 de maio de 2020, que mostrou que a reprovação do presidente Jair Bolsonaro cresceu de 38% para 43% desde o mês anterior.

Alguém desavisado poderia dizer que a variação foi de 5%, já que 43 menos 38 é igual a cinco. Entretanto, calcular a variação em porcentagem envolve os passos que acabamos de mostrar. Quando a

conta é simples assim, comparando a diferença entre duas porcentagens com uma simples subtração, o correto é dizer que a queda foi de cinco pontos percentuais.

Calcular a **variação percentual**, por outro lado, envolveria pegar essa diferença de 5 pontos percentuais e dividir pelos 38 originais. O resultado seria uma queda de aproximadamente 13%.

Taxas

Ao entender como funciona o cálculo de uma porcentagem, você automaticamente aprende também a usar outra ferramenta importante no cinto de utilidades de um jornalista de dados: o cálculo de **taxas**.

Como vimos, uma porcentagem consiste em transformar um número absoluto, muitas vezes difícil de colocar em perspectiva, em algo mais palatável. Assim, é mais fácil entender o que aquele valor representa.

Calcular uma porcentagem nada mais é que calcular uma taxa. Lembre-se do passo final da operação, quando o resultado da divisão da parte pelo todo é multiplicado por 100. A multiplicação, na prática, está colocando o resultado em uma **base** diferente.

Isso significa que, em vez de pensarmos em um número arbitrário, podemos usar uma quantidade que é mais fácil de compreender, mas representa a mesma proporção de casos.

A novidade é que essa base **não precisa ser 100**, necessariamente. Um caso frequente é a taxa de homicídios, que geralmente é calculada na base de **100 mil**.

Além de tornar mais fácil entender o que um número representa de fato, o cálculo da taxa serve, principalmente, para comparar fenômenos que acontecem em populações de tamanho diferente.

Veremos como isso funciona na prática, ainda com os dados de assassinatos de 2016. Em Trinidad e Tobago, 420 pessoas foram vítimas de homicídio naquele ano. No Brasil, no mesmo ano, foram aproximadamente 61 mil. **Qual país é mais violento?**

A resposta simples é que mais gente é assassinada no Brasil do que no país caribenho. Entretanto, a população do Brasil é cerca de 200 vezes maior. É natural que a quantidade total de mortes seja maior aqui.

Como podemos responder a essa pergunta de forma justa? **Usando uma taxa que coloque esses números na mesma base.** Assim, conseguimos saber o quão comuns são as ocorrências de assassinato dentro de uma população. Veja abaixo o passo a passo e repare como o processo se assemelha ao da porcentagem:

420
assassinatos

1,365
milhões de habitantes

61.000
assassinatos

207,7
milhões de habitantes

420
assassinatos

1,365
milhões de habitantes

→ **0.000308**
assassinatos por habitante

61.000
assassinatos

207,7
milhões de habitantes

→ **0.000293**
assassinatos por habitante

Os assassinatos, ainda que por pouco, são **menos comuns** no Brasil que em Trinidad e Tobago, mesmo que a contagem de mortes seja maior.

É uma boa prática calcular taxas sempre que formos comparar a realidade de universos de tamanhos diferentes. Mesmo assim, é bom nunca perder os números absolutos de vista, também.

De acordo com as circunstâncias, controlar os valores por população pode induzir leituras erradas. Precisamos prestar atenção especial na hora de comparar universos de escalas muito diferentes. Quando a população total é muito pequena, poucas ocorrências podem inflar uma taxa artificialmente.

Para entender melhor, vale olhar para os números de mortes por Covid-19 para cada milhão de habitantes em dois países. Na Islândia, essa taxa era de 27 ao final de maio de 2020. Na China, era de 3. A situação do país asiático parece muito melhor sob essa métrica, mas sabemos que a realidade é diferente.

Uma cidade chinesa foi o epicentro do primeiro surto da doença, com milhares de mortes que levaram a medidas compulsórias de isolamento social. No país nórdico, morreram apenas 10 pessoas.

As taxas de ambas as nações são puxadas para extremos opostos porque eles estão em extremos opostos: a China tem bilhões de habitantes, enquanto a Islândia tem menos de 500 mil.

Um único caso da doença teria grande impacto na taxa islandesa, enquanto mesmo uma quantidade de mortes capaz de colocar um sistema de saúde em colapso pouco alteraria a taxa chinesa.

Inflação

A análise de dados que envolve valores monetários deve, necessariamente, considerar a inflação do período. Antes de utilizar números de uma série histórica, por exemplo, é preciso verificar se os va-

lores foram corrigidos. Depois de verificar esse detalhe na documentação do banco de dados que pretende utilizar, é hora de fazer a conversão. Para tanto, existem ferramentas gratuitas que automatizam esse processo, como a calculadora do Banco Central e do IBGE.

Tomemos como ponto de partida o poder de compra de um cidadão que, em janeiro de 2005, recebeu R$ 1 mil pelo seu trabalho. Em valores corrigidos pelo Índice Nacional de Preços ao Consumidor Amplo (IPCA), esses R$ 1 mil tornaram-se R$ 2,13 mil em janeiro de 2019. Em outras palavras, este cidadão não teria o mesmo poder de compra caso continuasse recebendo R$ 1 mil em 2019. Por isso, a inflação do período precisa ser observada. Como afirma a jornalista Mariana Segala neste trecho do livro Siga os Números, atualizar valores de acordo com a inflação ou com os juros do período é uma das habilidades demandadas especialmente dos jornalistas que cobrem economia e finanças. A ausência de correção de valores provocaria uma análise equivocada.

Séries históricas

A análise temporal de um fenômeno exige dados de série histórica, isto é, informações correspondentes a um grande período de tempo. Podemos pensar, por exemplo, sobre dados de emprego e desemprego, distribuição de renda entre a população, produção agrícola e números de importação e exportação em determinado país.

Se você obtiver acesso apenas às informações referentes a um curto espaço de tempo (poucos meses ou poucos anos), é preciso redobrar os cuidados quanto aos resultados e afirmações contundentes e que remetem ao "melhor" ou "pior" desempenho de um setor. Também é importante observar se a coleta de dados foi feita pela mesma instituição ao longo dos anos e/ou se a metodologia de coleta e apresentação de dados passou por mudanças.

Outro ponto que exige atenção é a sazonalidade, muito importante na comparação entre períodos. Em períodos regulares do ano, por exemplo, o trânsito das cidades costuma ser bastante movimentado. Assim, não é possível comparar a circulação de veículos em períodos "normais" com os meses de janeiro e fevereiro. É natural que a circulação de veículos na cidade caia nos meses de férias escolares.

Produção agrícola e até mesmo criminalidade são outras áreas que bem exemplificam a necessidade de escolher períodos comparáveis entre eles, visto que há oscilação dos dados de acordo com os meses. Por isso, é fundamental que qualquer análise seja feita a partir do mesmo período correspondente no ano ou nos anos anteriores. Na dúvida, consulte um especialista.

Daqui a alguns anos, pense bem antes de comparar qualquer tempo com períodos de 2020: a Covid-19 marcou o ano e precisa ser vista como um fenômeno à parte, que deve ser levado em conta nas análises futuras sobre este ano.

Estatística para leigos

Depois de entrar em alguns conceitos de matemática básica, precisamos discutir as principais noções de estatística. Na linguagem do dia a dia, costumamos chamar de estatística tudo aquilo que é número. Quando um repórter fala de "trazer estatísticas" para uma matéria, geralmente está falando de rechear o texto com exemplos, contagens, porcentagens e afins.

Estatística é, na verdade, uma ciência que se dedica a analisar e interpretar bancos de dados. Geralmente, o jornalista ou comunicador trabalha com um campo específico da área: **a estatística descritiva** , que se preocupa em resumir de forma significativa as características de um conjunto de informações.

O conceito pode parecer estranho quando descrito dessa forma, mas é algo que está muito presente em nossa cotidiano. Os valores de média, moda, e mediana, por exemplo, fazem parte desse campo, assim como os conceitos de desvio padrão, outliers e padrões de distribuição. Vamos entender para o que serve cada um deles.

Moda, média e mediana

Você provavelmente já calculou a **média** de alguma coisa, nem que seja a média das notas do seu boletim escolar. O procedimento é simples: somar todos os números de uma série e dividir pela quantidade total de elementos. Você já parou para pensar qual é o objetivo desse cálculo?

Na prática, o cálculo da média tenta encontrar um número que seja capaz de representar todas as entradas de uma série de dados de forma simplificada. Quando calculamos a média das notas de todas as provas que um aluno fez em um ano, esperamos que o resultado seja um número que resuma todas as avaliações em uma só.

Não é algo tão simples e justo quanto parece, como qualquer estudante que tenha ficado insatisfeito com sua nota final já percebeu. Existem casos em que usar a média é ainda mais complicado.

Veja, na tabela, a distribuição dos salários de uma empresa:

Empregado	Salários (em mil R$)
Jéssica	1.5
Miguel	2.5
Bernardo	2.5
Juliana	2.5
Adriana	2.5
Carlos	3.5
Antônia	3.5
Gabriel	3.5

Empregado	Salários (em mil R$)
Letícia	4.0
Sandra	4.5
José	5.0
Bianca	5.5
João	6.5
Camila	28.0
Amanda	150.0

Ao calcular a média salarial dos funcionários, chegamos a um número de R$ 15 mil por mês. Entretanto, **não há uma única pessoa que receba esse valor**. Nesse caso, a média não oferece um bom resumo da realidade.

Isso acontece porque os salários de Camila e Amanda são muito maiores que os demais. Como o cálculo da média simples dá um mesmo peso para todos os elementos, o resultado acaba sendo muito sensível a extremos como esses.

Por sorte, existem outras medidas que podemos usar em situações com essa.

A **mediana**, por exemplo, usa uma tática diferente para encontrar um número representativo do banco de dados: ela divide a série em duas metades e pega o valor que está **exatamente no meio**.

Consideremos esta série hipotética:

[1, 2, **3** , 4, 8]

Neste caso, o valor seria o **3**, já que há uma mesma quantidade de valores maiores e menores do que ele.

No caso dos salários que mostramos acima, a mediana é um dos salários de R$ 3,5 mil.

[1.5, 2.5, 2.5, 2.5, 2.5, 3.5, 3.5, **3.5** , 4.0, 4.5, 5.0, 5.5, 6.5, 28.0, 150.0]

Note como há exatamente sete itens acima e sete itens abaixo do valor selecionado. O valor não é sensível a valores exagerados e oferece um dado mais real.

A **moda**, outra métrica de centralidade, usa uma estratégia mais simples: selecionar, simplesmente, o número que mais se repete. No exemplo acima, é o salário de R$ 2,5 mil, que ocorre com quatro dos 15 funcionários.

Para escolher qual é a melhor métrica para o banco de dados que você tem em mãos, é preciso entender um pouco mais sobre outro conceito de estatística: **a distribuição**.

Distribuição: dispersão, desvio padrão e outliers

Ainda seguindo com o exemplo dos salários, vamos prestar mais atenção na característica dos dados que tornou o cálculo da média pouco representativo: havia elementos extremos na série, ou seja, salários muito distantes dos demais e que distorciam o cálculo.

Esse tipo de característica pode ser medida usando estatísticas descritivas de **dispersão**. Para entender melhor quais são as informações de um banco de dados - se ele tem muitos extremos ou é mais uniforme, por exemplo - é muito útil usar um tipo de gráfico chamado histograma. Em resumo, esse gráfico permite observar a distribuição dos dados e identificar pontos fora da curva. Falaremos mais sobre ele no capítulo sobre visualização de dados.

Agora, vamos falar de medidas que, assim como as estatísticas de centralidade que vimos anteriormente, tentam resumir as características de um banco de dados em um número único.

Já vimos como esse tipo de abordagem pode ser perigoso mas, ainda assim, servem para ter uma ideia rápida do tipo de dados com os quais estamos lidando.

O **desvio padrão** , por exemplo, é útil para saber se estamos diante de um banco de dados com muitos valores extremos. Esse número é, na prática, **uma média da distância de cada ponto da média de todos os bancos de dados.**

Confuso? Nem tanto. Vamos voltar aos dados de salários com que trabalhamos, cuja média é R$ 15 mil.

Para calcular o desvio padrão, precisamos calcular a diferença do salário de cada funcionário da média salarial: Jéssica, por exemplo, ganha R$ 1,5 mil, ou seja, está R$ 14,5 mil distante da média. Amanda ganha R$ 150 mil, então está R$ 135 mil distante da média.

Ao somar todas essas diferenças e dividir pelo total de funcionários, temos o desvio padrão, que tenta resumir o quão uniforme é um banco de dados. Um desvio padrão maior indica dados mais díspares, com a presença de um ou mais extremos.

Nesse caso, o valor é de R$ 37,9 mil. Para saber se ele é significativo, vale comparar com as outras estatísticas que vimos: trata-se de mais do que o dobro da média e é mais de dez vezes maior do que a moda e a mediana.

Também é possível usar o desvio padrão para descobrir o quão extremo é um ponto de dados. Para isso, é preciso dividir o valor pelo desvio padrão.

Por exemplo, vamos analisar novamente o salário de Amanda: R$ 150 mil, divididos pelo desvio padrão de R$ 37,9 mil, resultam em aproximadamente 4 desvios padrão acima da média.

Usualmente, dados que estejam a três desvios padrão ou mais da média, em um banco de dados que segue uma distribuição normal, podem ser considerados extremos - ou seja, são outliers.

Ainda que um outlier possa ser de interesse jornalístico, esse tipo de dado pode prejudicar a análise dos demais. Sempre verifique se esse não é o caso antes de interpretar estatísticas descritivas simples.

Correlação

Outra técnica estatística útil para o jornalismo é o cálculo da taxa de correlação. Essa medida, que é bastante complexa e quase sempre é calculada com o uso de programas de planilha ou tecnologias equivalentes, estima como o comportamento de duas variáveis está interligado.

Vamos, de novo, deixar o jargão de lado e tentar entender o conceito com um exemplo mais próximo da realidade: temos um banco de dados que mostra a nota que cada estudante tirou no Exame Nacional do Ensino Médio (Enem) e a renda média de sua família.

Com um cálculo de correlação, podemos verificar que, quanto maior é a renda da família, melhor a nota tende a ser. Trata-se de uma correlação positiva: quando um dos valores cresce, o outro provavelmente vai crescer também.

Existem as correlações negativas, em que ocorre algo oposto: quando uma das variáveis aumenta, a outra diminui. Um exemplo possível é a relação entre renda média e taxa de analfabetismo. Quanto maior for a renda média entre os moradores de uma cidade, por exemplo, menor tende a ser a prevalência de analfabetismo entre a população.

E o quão forte são essas ligações, afinal? Para mensurar exatamente o quanto duas variáveis estão conectadas, é possível calcular um valor chamado **coeficiente de correlação.**

Existem vários tipos, mas o de uso mais frequente no jornalismo é coeficiente de **correlação de Pearson**. Ela mede justamente o tipo de relação aqui apresentado e classifica a intensidade do fenômeno de -1 até 1.

Uma correlação de 1 é perfeitamente positiva, em que cada unidade a mais em uma variável gera uma unidade a mais na outra. Uma correlação de -1 é perfeitamente negativa, e uma unidade a mais em uma variável gera uma unidade a menos na outra. Quando o valor é 0, não há correlação alguma e as variáveis não se influenciam.

Entretanto, na prática, é raríssimo encontrar correlações perfeitas ou a total ausência de correlação. O coeficiente se apresenta, na vida real, em frações: 0,8 indica uma correlação forte, por exemplo, enquanto 0,2 indica uma correlação fraca.

Esse tipo de cálculo é útil para fortalecer estatisticamente matérias que mostram como dois fatores estão relacionados. É possível calcular a correlação entre pobreza e acesso à educação ou pobreza e violência, por exemplo.

Contudo, mesmo diante de um coeficiente de correlação alto, é preciso tomar alguns cuidados.

Vamos voltar para o exemplo de notas no Enem. Agora, em vez de comparar a nota de cada estudante

com a renda de sua família, você resolveu medir o efeito da arborização na performance escolar. E existe uma correlação clara entre conseguir notas melhores e viver em um bairro cheio de árvores.

Hora de escrever uma matéria espetacular sobre isso, certo? Errado. É hora de pensar um pouquinho mais a fundo.

É muito provável que a variável de renda influencie tanto a quantidade de árvores no bairro de cada estudante como a sua nota na prova. Gente com mais dinheiro mora em locais com melhor estrutura urbana e também tende a ter resultados melhores no Enem.

Não é o excesso de árvores que influencia o desempenho na prova. É a renda que influencia tanto a arborização quanto a nota no exame. Quando não nos atentamos para essa relação escondida, na verdade, estamos medindo uma terceira variável sem nos darmos conta.

Além disso, é possível encontrar dados que se correlacionam de forma perfeita, mas que não tem nenhuma conexão entre si. Existe até um livro apenas sobre isso. Entre os exemplos, está até uma correlação de 0.95 (portanto, super forte) entre o consumo de queijo muçarela e a quantidade de pessoas formadas em um curso superior de Engenharia Civil.

É por motivos como esses que existe um mantra entre quem trabalha com análise de dados: **correlação não é causalidade.** Só porque dois fenômenos ocorrem de forma paralela, não podemos afirmar que um **causa** o outro, mesmo que o coeficiente de correlação seja absurdamente alto.

Para afirmar que um fenômeno causou outro, é preciso fazer testes estatísticos mais específicos e poderosos, que não conseguiremos cobrir nesse livro.

Na dúvida, é melhor não avançar o sinal: você não quer ser o cara que disse que viver perto de árvores faz alunos irem melhor na prova, não é mesmo?

Margem de erro

Imagine que você queira compreender os hábitos de consumo dos brasileiros ou mesmo quais são suas preferências de voto em uma eleição iminente. É improvável que alguém tenha tempo ou recursos suficientes para ouvir cada um dos 210 milhões de brasileiros. Por isso, trabalha-se com a ideia de amostra, ou seja, um conjunto de pessoas cujas características sociais e demográficas sejam capazes de representar a população como um todo. Dessa forma, as respostas obtidas em uma amostra de pesquisa podem ser atribuídos a um universo (neste caso, a população inteira).

Ao trabalhar com dados amostrais, portanto, fique atento à margem de erro, que corresponde a quantos percentuais (para mais ou para menos) as informações podem variar. Se 90% dos respondentes de uma pesquisa concordaram com uma afirmação, por exemplo, e a margem de erro for de cinco pontos percentuais, estima-se, então, que entre 85% e 95% pessoas responderam afirmativamente a esta questão.

Pesquisas com margens de erro muito altas (a partir de 10 pontos percentuais) dificilmente trazem dados que representam de forma qualificada uma determinada realidade, e por vezes indicam problemas na metodologia. A margem de erro em pontos percentuais é uma característica importante a ser levada em conta, e deve, inclusive, ser comunicada claramente na reportagem. Em geral, as pesquisas amostrais também divulgam o intervalo de confiança, normalmente de 95%. Isso significa que, se a pesquisa for repetida com amostras aleatórias 100 vezes, em 95 delas o resultado será mesmo.

Operações de análise de dados

Existem algumas operações básicas que podem ajudar a encontrar padrões e tendências nos dados. De acordo com o software que você usa para trabalhar, os procedimentos podem ter algumas diferenças, mas o raciocínio segue o mesmo.

De início, tente **ordenar os dados** do maior para o menor ou do menor para o maior. Se você está mexendo com dados de gastos governamentais, faz sentido ver qual foi a maior compra de todas, por exemplo.

Filtrar a planilha é interessante para ver se o padrão de comportamento dos dados é o mesmo quando olhamos apenas para um segmento deles. Ainda usando a mesma metáfora, pode ser que o gasto médio de um ministério específico seja muito diferente do gasto médio do governo como um todo.

Da mesma forma, **agregar** os dados pode revelar padrões que não seriam vistos de outra forma: em vez de olhar apenas para cada uma das compras governamentais, individualmente, é possível reuni-las em categorias. Quantas foram feitas por cada departamento? Quanto gastou cada ministério? Você conseguirá fazer esta operação usando operadores de linguagens de programação, como o GROUP BY, ou através das tabelas dinâmicas nos editores de planilha.

Com as tabelas dinâmicas, você conseguirá de fato utilizar os editores de planilha para entrevistar seus dados. Imagine que você tenha uma tabela onde cada linha representa repasses de financiamento de campanha e você gostaria de responder a seguinte pergunta: quais doadores financiaram a maior soma de recursos?

Existem várias operações de análise dos dados que são necessárias para respondê-la: primeiro, você precisa agrupar os doadores de acordo com uma coluna. Ao realizar agrupamentos, sempre dê preferência a usar campos identificadores como CPF ou CNPJ ao invés de campos de texto, como o nome, que são mais suscetíveis a erros de digitação.

Ao agrupá-los, será necessário, então, fazer uma soma na coluna com os valores financiados para, em seguida, ordenar os resultados de forma decrescente e, assim, descobrir os 10 principais financiadores.

Na tabela dinâmica, você conseguirá realizar todas estas operações. Basicamente, a ideia por trás deste recurso é que você deve criar uma nova tabela, que irá “moldar” os dados da planilha original

para que eles assumam a forma que você deseja.

Em geral, as linhas dessa nova tabela representam uma entidade, como um doador, nesse exemplo. Nas colunas, geralmente, aparecem os valores agregados por categoria. No exemplo, revelariam a soma total de doações feitas por cada pessoa .

Assim, independente do software utilizado, o importante ao entrevistar dados com tabelas dinâmicas é saber identificar qual é a forma que essa nova tabela precisa assumir.

A	B	C
Doador	**Valor pago**	**Data**
Pedro	R$ 120,00	01/10/2019
Maria	R$ 160,00	05/11/2019
Pedro	R$ 80,00	07/12/2019
Pedro	R$ 60,00	13/12/2019
Júlia	R$ 30,00	17/12/2019
João	R$ 10,00	19/12/2019
Júlia	R$ 40,00	21/12/2019

Figura 19: Exemplo de tabela com as informações originais.

A	B
Doador	SUM de Valor pago
João	R$ 10,00
Júlia	R$ 70,00
Maria	R$ 160,00
Pedro	R$ 260,00
Total geral	**R$ 500,00**

Figura 20: Tabela dinâmica gerada automaticamente com os valores da tabela anterior.

As duas imagens acima mostram como funciona uma tabela dinâmica: o valor pago por cado doador, em diferentes datas, foi agregado em uma única soma.

Por fim, **cruzar** dados também pode ser interessante: o termo significa trazer uma segunda base de dados que possa enriquecer o trabalho de análise. Um exemplo clássico é reunir dados de resultados eleitorais por município com os indicadores de desenvolvimento de cada cidade, por exemplo.

Existem diferentes formas de cruzar tabelas diferentes. Nos editores de planilha, há funções como a procura vertical (PROCV ou, em inglês, VLOOKUP, vertical lookup), que podem resolver cruzamentos

mais simples. Porém, à medida que você avançar no trabalho com dados, vai perceber que realizar cruzamentos com editores de planilha não é o ideal. Nestes casos, o recomendável é utilizar linguagens de consulta ou programação, que contam com a poderosa função **join**, responsável por unir tabelas diferentes.

Seja qual for a ferramenta que você adote, o princípio permanece: **para fazer um cruzamento, você precisa de um campo que sirva como identificador, ou seja, uma coluna ou variável que esteja presente nas tabelas que serão cruzadas**. A existência de campo identificador serve como uma "ponte" para conectar bases diferentes e é fundamental para realizar cruzamento.

Além disso, também vale atentar para padrões temporais (os gastos do governo crescem em algum momento do ano?), para outliers suspeitos (por que essa compra específica é muito mais cara que as demais?) e para tendências de crescimento (por que os gastos desse ministério subiram tanto a partir do ano passado?).

No final das contas, a chave para uma boa matéria é a curiosidade e a capacidade de fazer boas perguntas para a planilha. Como vimos, é possível fazer muitas coisas com editores de planilha, mas se você quer realmente progredir na análise de dados é altamente recomendável aprender uma linguagem de programação. Assim, você conseguirá documentar e checar melhor suas análises, além de não ficar limitado às funcionalidades dos editores de planilha.

Como visto no primeiro capítulo, a linguagem SQL pode ser um bom começo. Ao contrário das planilhas, trabalhar com bases de dados em SQL permitirá lidar com volumes grandes de informações. Com ele, você conseguirá fazer operações de análise, mas tenha em mente que esta é antes de tudo uma linguagem de consulta, como o seu nome indica.

Se você quer entrar a fundo na análise de dados, vá para o **Python ou R**. Na linguagem R, a dica para quem quer trabalhar com análise é a biblioteca Tidyverse, uma biblioteca que reúne diversos recursos que facilitam e muito a lida com os dados. Ele é relativamente simples de ser compreendido e guarda muitas semelhanças com o SQL, de modo que pode ser uma opção interessante para quem quer começar com aquela linguagem para depois alçar voos maiores. O Tidyverse também já conta com o ggplot2, uma poderosa ferramenta visualização de dados.

Já em Python, um componente fundamental para manuseio de dados é o **pandas**. Trata-se de um pacote bastante robusto, considerado a principal referência para análise de dados nesta linguagem. Em vez de lidar com loops e condicionais, como é comum no mundo da programação, o pandas implementa uma estrutura de dados chamada *dataframe*, de modo semelhante ao Tidyverse. Assim, na prática, a análise de dados se assemelha bastante ao trabalho com planilhas, com uma estrutura de colunas e linhas.Ele pode ainda ser combinado com outras bibliotecas, como o *matplotlib*, parar criar gráficos simples de maneira rápida. Além disso, ele se destaca pela performance ao lidar com grandes volumes de dados ou realizar operações complexas.

Visualize

Ainda que a visualização de dados se insira no fluxo de trabalho com dados, esta área se trata de um campo do conhecimento que merece atenção por si só. É impossível falar de forma exaustiva sobre o tema em um capítulo de livro. Assim, nas próximas páginas, a abordagem será pragmática e bastante resumida.

Antes de mais nada, vamos definir o que é visualização de dados: como reconhecer uma visualização de dados na página de uma publicação? Como não confundir com outras formas de representação visual comuns na imprensa?

Na introdução de "*The Functional Art*", livro publicado em 2013, o pesquisador **Alberto Cairo** descreve uma cena que testemunhou várias vezes quando trabalhava em redações: quando editores precisavam de um gráfico para acompanhar uma reportagem, dirigiam-se para a equipe de infografia do veículo para encomendar "uma arte".

O termo, na rotina de trabalho, faz algum sentido. Na hierarquia da maioria dos jornais, a equipe responsável por produzir gráficos está, de algum modo, subordinado à **editoria de Arte**. Nessas editorias, muitas vezes, a maior parte da força de trabalho não tem formação em jornalismo: são designers e artistas visuais, não repórteres e redatores.

Segundo Cairo, porém, o uso desse termo gera um problema. É frequente que, na hora de produzir um jornal, as peças visuais que acompanham o conteúdo da reportagem sejam vistas como mero complemento ou ilustração. Sob essa perspectiva, o conteúdo realmente importante de uma matéria ficaria no texto, enquanto o gráfico serve para resumir o assunto para um leitor pouco interessado ou, simplesmente, para preencher o espaço em branco ou tornar uma página mais agradável esteticamente.

Entretanto, a função de um gráfico não é decorativa. A função primordial de um gráfico é informar - tanto quanto um bloco de texto. Uma visualização de dados impactante concentra tanta informação quanto o *lead* (o que, quem, quando, onde, como e por que) de uma reportagem.

Para Cairo, a relação entre visualizações e arte é como a relação entre jornalismo e literatura. Segundo ele, um jornalista pode se inspirar nas técnicas e no estilo de um grande ficcionista. Entretanto, seu trabalho nunca deve virar literatura, porque é de outra natureza.

O mesmo vale para a produção de gráficos. Ainda que um bom profissional de visualização tenha

apreço pela beleza e elegância de uma peça artística, essa é uma aspiração secundária. O foco é transmitir informação da forma mais precisa possível.

Há uma confusão comum entre visualização de dados e infografia. Existe todo um debate a respeito das definições exatas de cada termo e às vezes os termos são usados de forma intercambiável. Porém, de modo geral, podemos destacar algumas características que distinguem um conceito do outro.

De acordo com a definição proposta por Tatiana Teixeira, pesquisadora da UFSC (Universidade Federal de Santa Catarina), por infográfico, compreendemos uma peça visual com um objetivo narrativo, que explica algo a quem está lendo. São trabalhos que mostram como ocorreu um acidente de avião ou uma operação da polícia, por exemplo. O importante aqui é que os fatos se sucedem de forma mais ou menos contínua no tempo.

Por outro lado, uma visualização de dados estrutura ou tenta traduzir dados quantitativos, codificando visualmente algumas de suas características em dimensões, como cor, forma, tamanho, posição e etc. Como nós humanos somos melhores observando padrões visuais do que numéricos, a visualização de dados nos ajuda a reconhecer padrões e identificar características importantes das bases, que seriam impossíveis de serem enxergadas em tabelas. É o caso dos gráficos de barra, pizza, dispersão entre outros.

Assim, de acordo com esta divisão, um infográfico poderia compreender uma visualização de dados, mas não o inverso. Como dito, a definição explorada neste tópico não é consensual, mas serve para delimitar o escopo desta seção. Neste guia, não iremos abordar as práticas de infografia, mas sim de visualização de dados.

Visualização como expansão cognitiva

Alberto Cairo oferece uma definição bastante útil de visualização de dados em "*The Truthful Art*", seu livro de 2016. Ele afirma que uma visualização de dados é uma forma de exibir informação quantitativa que é **"desenhada para permitir análises, explorações e descobertas".**

Cairo destaca também que uma visualização de dados não tem como objetivo principal enviar uma mensagem pronta e hermética para o leitor. A ideia desse tipo de gráfico, na verdade, é servir para que as pessoas tirem suas próprias conclusões a partir dos números exibidos.

Assim, visualizações de dados são formas de **aumentar a capacidade de compreensão** humana. A prática funciona porque valores quantitativos, com sua natureza abstrata, são difíceis de analisar e interpretar.

Quando a informação é representada de forma visual, nossa mente é capaz de realizar tarefas como comparação e ordenação de forma mais eficiente.

Um exemplo ajuda a deixar essa realidade mais palpável. O que é mais fácil: fazer uma conta de divisão de cabeça ou comparar o tamanho de uma série de barras em um gráfico?

A segunda opção certamente é menos cansativa que a primeira, a não ser que você seja alguém com uma aptidão fora do comum para a matemática. Entretanto, na prática, ambas as atividades são a realização de uma mesma tarefa: estimar quantas vezes um valor é maior que outro.

Fazer essa operação com um gráfico de barras é mais fácil porque essa visualização de dados foi elaborada para tirar proveito de algumas características da cognição humana - nesse caso, a facilidade em perceber e comparar o tamanho de objetos.

O exemplo do gráfico de barras é o mais simples possível, mas mesmo as visualizações de dados mais complexas partem de um princípio semelhante.

Ainda que seja desejável produzir algo bonito, memorável e divertido, o principal objetivo do bom criador de gráficos é permitir que o seu leitor obtenha uma percepção mais aguçada de fenômenos complexos.

Conforme as formas de mensurar a realidade se tornam mais sofisticadas, as representações visuais também precisam se transformar. Os fundamentos, porém, são relativamente constantes há alguns séculos.

Breve histórico da visualização de dados

Uma retrospectiva histórica ajuda a demonstrar como o desenvolvimento de formas de representação visual está intimamente ligada à necessidade de realizar tarefas intelectuais e analisar valores quantitativos de forma sistemática. Um dos trabalhos mais detalhados sobre o passado do campo foi feito por **Michael Friendly**, professor de psicologia da Universidade de York, em Toronto, no Canadá.

Ele elaborou uma lista de marcos na história da visualização de dados que começa mais de 6 mil anos antes de Cristo, com mapas primitivos. A compilação passa pelo surgimento de tabelas que tentavam representar a posição dos planetas, diagramas trigonométricos e afins, mas vê um ponto de inflexão a partir do século XVII.

É nesse momento que se consolidam áreas do saber que preocupavam-se em mensurar e analisar fenômenos físicos, como a astronomia e a cartografia. Além disso, nasciam os rudimentos das teorias da probabilidade e os primeiros censos demográficos.

Existia, assim, um conjunto de dados relevante e o interesse coletivo por novos métodos de análise. Entretanto, uma quantidade maior de números exigia novas ferramentas para compreendê-los.

É nesse momento histórico que surgem os primeiros gráficos estatísticos modernos, como o gráfico de pizza, de barras e de linha. Boa parte deles foi inventada ou popularizada por **William Playfair**, um engenheiro escocês.

Em 1786, ele publicou uma obra chamada "*The Commercial and Political Atlas*", que foi pioneira em usar elementos visuais para representar tendências políticas e econômicas de forma sistemática.

Nos séculos seguintes, o campo da visualização de dados viu desenvolvimentos significativos, com impactos efetivos no discurso público.

Apenas neste livro, já foram citados dois exemplos de trabalhos de visualização de dados que mudaram consensos sociais nessa época: o mapa de cólera de John Snow e os gráficos feitos por Florence Nightingale sobre as causas de morte durante a Guerra da Crimeia.

Tanto Snow como Nightingale não se esforçaram apenas para reunir informações quantitativas. O que fez com que seus levantamentos tivessem efeitos significativos sobre a sociedade foi a apresentação visual.

O mapa de Snow mostra casos de cólera se distribuindo ao redor de um poço de aǵua contaminada, revelando uma relação espacial que não seria fácil de notar a partir de uma simples tabela de endereços. Já os gráficos circulares de NIghtingale revelam que a maior parte das mortes no front do confronto eram causadas por doenças infecciosas e preveníveis, sem que o leitor precise fazer contas e calcular percentuais.

Na prática, os dois se esforçaram para mostrar de forma mais palpável fenômenos que seriam difíceis de perceber e avaliar a olho nu - o que, segundo Cairo, é justamente o que a visualização de dados faz de melhor até hoje, algumas centenas de anos depois.

Princípios práticos

Agora, já temos uma definição mais precisa do que significa o termo **visualização de dados**. Já vimos também que um gráfico serve como ferramenta que ajuda a mente humana a interpretar valores quantitativos abstratos. Além disso, vimos com exemplos históricos como essa prática é antiga e pode ter impactos significativos na sociedade. Resta, agora, aprender conceitos práticos. Nas próximas páginas, vamos ver algumas das ideias básicas que estão por trás de visualizações de dados efetivas.

Um bom gráfico tem uma missão principal: possibilitar uma comunicação efetiva. De pouco vale uma ideia criativa, uma apresentação bonita e uma boa pauta se a visualização de dados é confusa ou poluída, falhando nessa tarefa básica.

Para garantir que um gráfico seja compreensível, é preciso pensar mais a fundo na **tarefa perceptiva** que ele ajuda a fazer - ou seja, no tipo de análise que o leitor vai tentar realizar.

Seguindo com a metáfora de que a visualização de dados é uma ferramenta, precisamos escolher a peça ideal para o propósito do trabalho.

Assim como não faz sentido algum usar um martelo para apertar um parafuso ou um alicate para prender um prego, há tipos de gráficos específicos para vários tipos de operações mentais.

No final deste capítulo, vamos ver um inventário de formatos que pode ser útil como ferramenta de consulta na hora de escolher a forma de representação adequada para uma reportagem.

Antes disso, porém, precisamos entender **por que** há gráficos que são melhores para algumas coisas do que outros. A resposta tem tudo a ver com a capacidade da mente humana de visualizar e comparar formas.

Hierarquia da percepção

Em 1984, dois pesquisadores decidiram testar quais tipos de gráficos são mais **precisos** para transmitir informações quantitativas.

William S. Cleveland e Robert McGill organizaram um experimento em que mostraram para dezenas de colaboradores uma série de visualizações de dados em diferentes formatos: gráficos de barras, de dispersão, de pizza, de bolhas e mapas de cor.

Além disso, eles definiram que tipo de elemento visual era usado em cada um dos gráficos. Em um gráfico de barras, por exemplo, o **comprimento** de cada item é usado para representar um número. Em um gráfico de pizza, os valores são representados pelo **ângulo** de cada fatia.

Os gráficos também foram mostrados com ou sem o auxílio de um **eixo** - ou seja, de uma escala que mostrasse os números representados.

O objetivo do teste era descobrir quais formas gráficas permitiam uma estimativa mais adequada da diferença entre os valores exibidos. O resultado mostrou que alguns formatos de gráfico foram lidos consistentemente de forma mais precisa.

Ao analisar os resultados, os estudiosos propuseram um ranking de formatos, do mais preciso para o menos.

1. Posição sobre um eixo comum
2. Posição sobre eixos não alinhados
3. Comprimento
4. Direção
5. Ângulo
6. Área
7. Volume
8. Curvatura
9. Tom de cor
10. Saturação de cor

A imagem a seguir, preparada por Alberto Cairo para o blog Periodismo con futuro, do El País, mostra de que tipo de gráfico falamos em cada item.

Quanto mais alto na escala está a forma de representação, mais fácil é usá-la para fazer comparações precisas. Quanto mais para baixo, mas difícil.

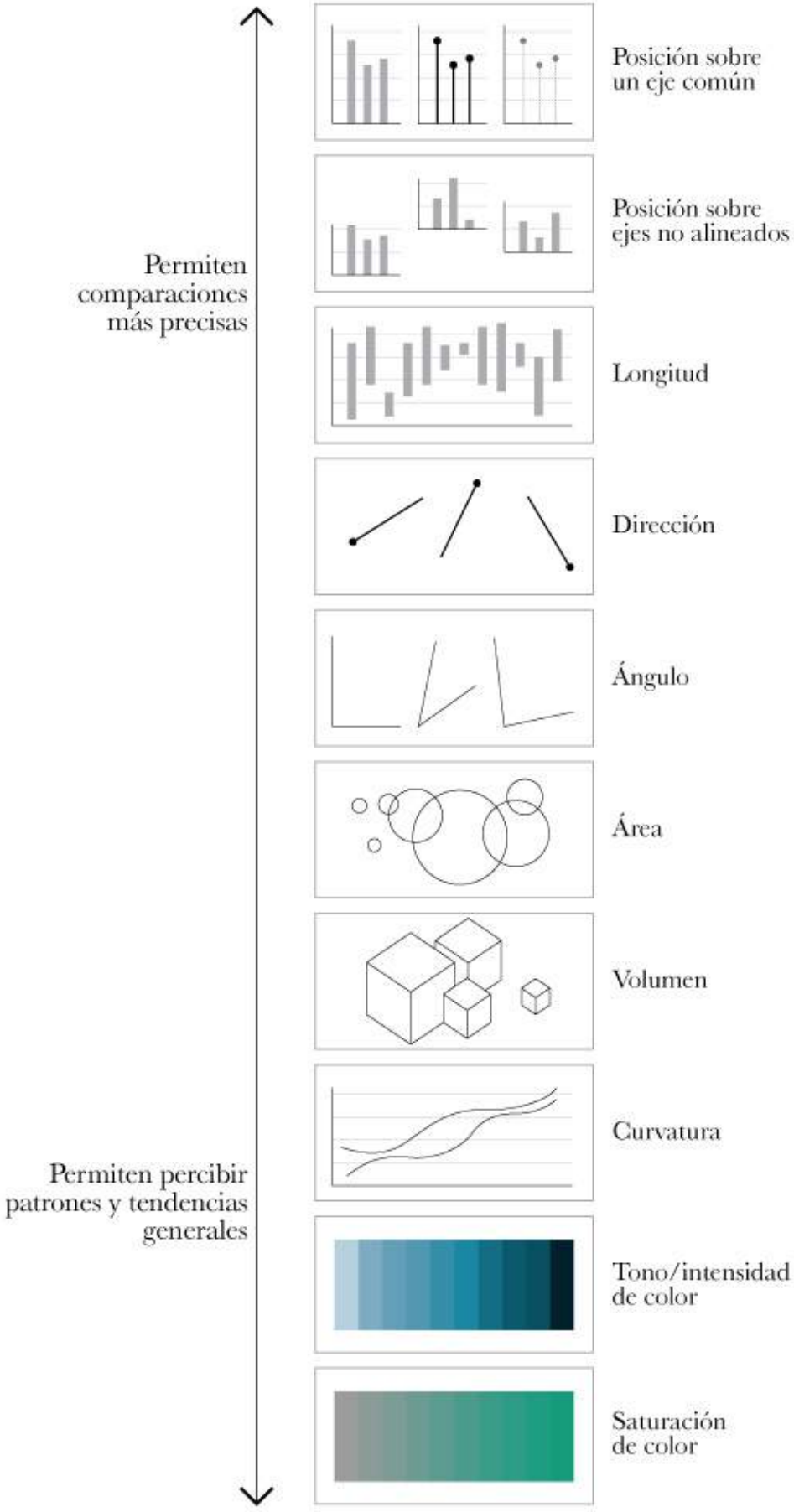


Figura 21: A hierarquia da percepção segundo experimento de Cleveland e McGill. Fonte: Alberto Cairo.

Esta outra imagem, também elaborada por Cairo para o mesmo blog, mostra como isso acontece na prática.

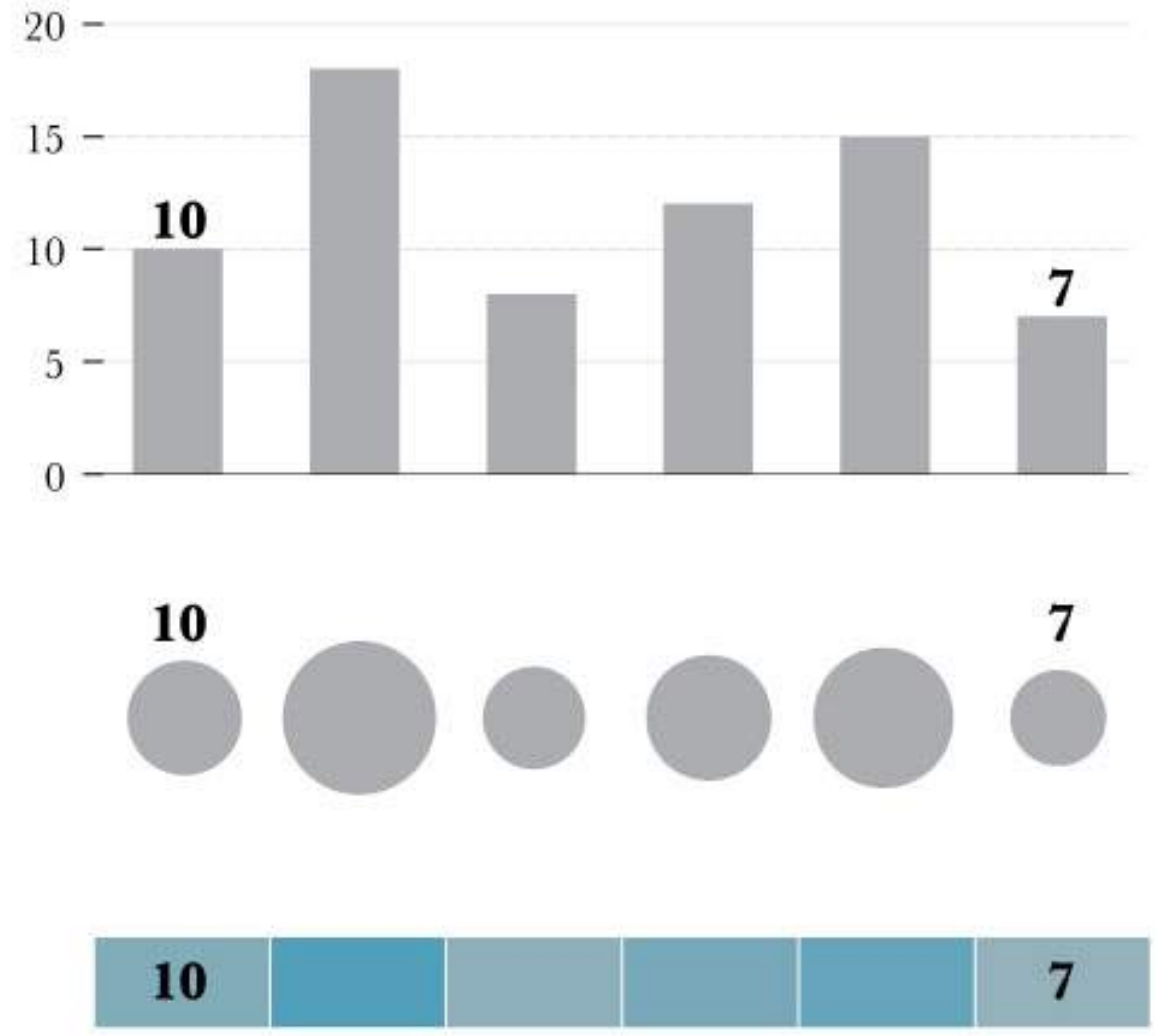


Figura 22: Qual destes gráficos permite uma leitura mais precisa? - Fonte: Alberto Cairo (2012)

Todos os gráficos representam os mesmos valores, mas de diferentes formas. Todavia, os gráficos de barras permitem uma leitura mais precisa e rápida.

É evidente que a segunda barra, por exemplo, é mais alta que a quinta. Contudo, é bem mais difícil dizer qual dos círculos é o maior ou qual das cores é a mais escura.

Assim, é possível afirmar, objetivamente, que os gráficos na parte de cima do espectro permitem comparações mais **precisas**.

Nem sempre a forma gráfica mais precisa é a mais adequada. De vez em quando, não queremos que os leitores façam comparações muito específicas, mas sim que sejam capazes de ver **tendências gerais e padrões**.

Nesses casos, formas próximas da parte inferior da escala passam a ser mais úteis, especialmente quando são usadas em conjunto com outras dimensões, como, por exemplo, em um mapa.

Veja este gráfico sobre o resultado das eleições presidenciais no Brasil em 2018:

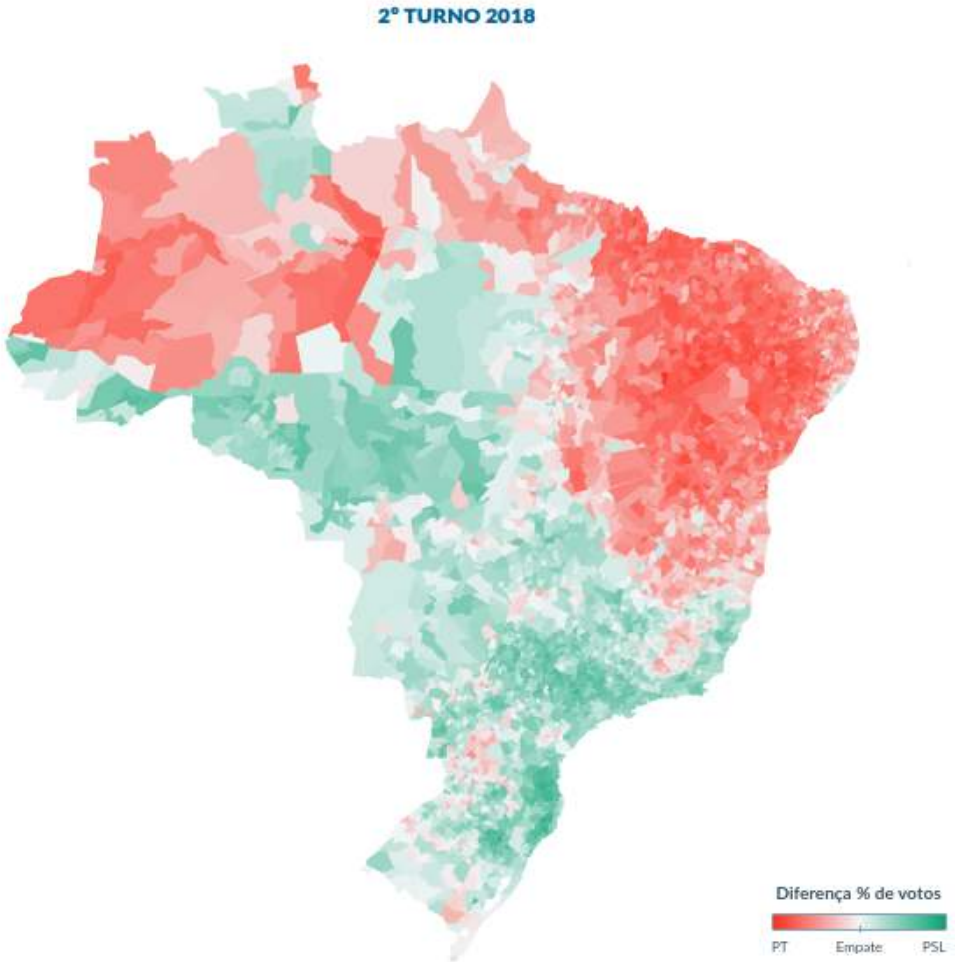


Figura 23: Mapa do resultado das eleições presidenciais de 2018 no Brasil, produzido pelo Estadão

O objetivo dessa visualização não é fazer com que o leitor descubra qual é a cidade onde Bolsonaro teve mais votos ou comparar precisamente a quantidade de votos recebidos por Haddad em dois municípios distantes. **A ideia é mostrar a distribuição geográfica do voto**, que tende a ser PT na região Nordeste e PSL no Sul. Um gráfico de barras não seria efetivo nesse caso.

De novo, cabe ao autor da visualização determinar qual é a tarefa que seu gráfico vai ajudar a cumprir e transformá-lo na ferramenta mais adequada possível para tal. Vamos ver, agora, um princípio básico das boas visualizações de dados: **ater-se ao que é essencial**.

Clareza vs emoção

Outro pesquisador fundamental do campo da visualização de dados é **Edward Tufte**. Seu livro *"The Visual Display of Quantitative Information"*, publicado em 1983, é uma das obras mais influentes da área.

Talvez o conceito mais marcante do trabalho de Tufte seja a ideia de **data-ink ratio**, ou a **razão dados-tinta**.

Em resumo, a ideia é que um **gráfico é mais efetivo quando usa a menor quantidade de elementos visuais para comunicar um valor** - ou seja, quando a razão entre os dados comunicados e a quantidade de "tinta" usada para tal for a menor possível.

Na prática, isso significa que Tufte defende a produção de gráficos minimalistas, em que a presença de cada elemento no papel tem uma clara função informativa. Assim, a "densidade de dados" do gráfico aumenta.

Como consequência, surge o conceito de "*chart junk*", que pode ser compreendido como a poluição de elementos em um gráfico. São justamente os elementos desnecessários, redundantes, e que não agregam densidade informativa ou que não representam dado algum. Servem apenas como decoração, para efeitos estéticos - e mesmo assim, na maioria das vezes, deixam a apresentação do conteúdo mais feia em vez de interessante.

Veja o gráfico a seguir, gerado a partir de dados ficcionais, sem muito esforço nem ou atenção ao design no Google Sheets.

Vamos analisá-lo a partir dos princípios de Tufte que acabamos de conhecer.

Primeiro, percebemos que há uma série de elementos redundantes nessa visualização: o eixo, na esquerda, exibe os mesmos valores que os rótulos em cima de cada uma das barras.

O título do gráfico, "gastos mensais", aparece também no eixo esquerdo. E uma legenda completamente descartável revela que as barras de cor vermelha representam "gastos".

Além disso, existem elementos que não agregam dado algum. Qual é a razão das barras estarem em 3D, quando a única dimensão que importa é a altura, e não a profundidade ou a largura? Por que tantas linhas no fundo, por que tantos marcadores no eixo vertical, por que aplicar uma sombra cinza na base das barras?

Removendo todos estes elementos, ficamos com um gráfico mais próximo daquilo que Tufte considera o ideal. A visualização a seguir foi feita em ainda menos tempo no Datawrapper, uma ferramenta gratuita disponível online.

Perceba como o gráfico fica mais fácil de ler. O motivo é que não há mais elementos redundantes ou pouco informativos disputando atenção. Assim, nossos olhos e cérebro conseguem se concentrar apenas naquilo que importa.

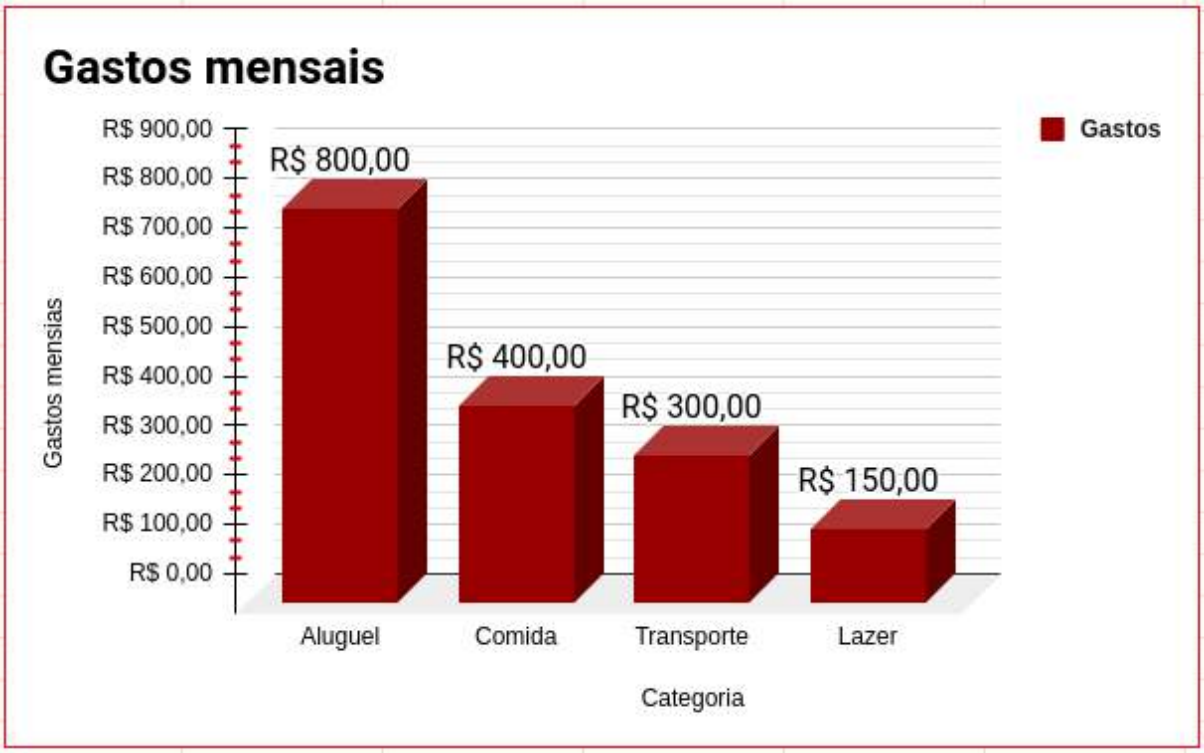


Figura 24: Exemplo de gráfico de barras com elementos visuais desnecessários.

Figura 25: O mesmo gráfico, porém com uma apresentação mais minimalista

Agora, faça o mesmo exercício com o próximo gráfico.

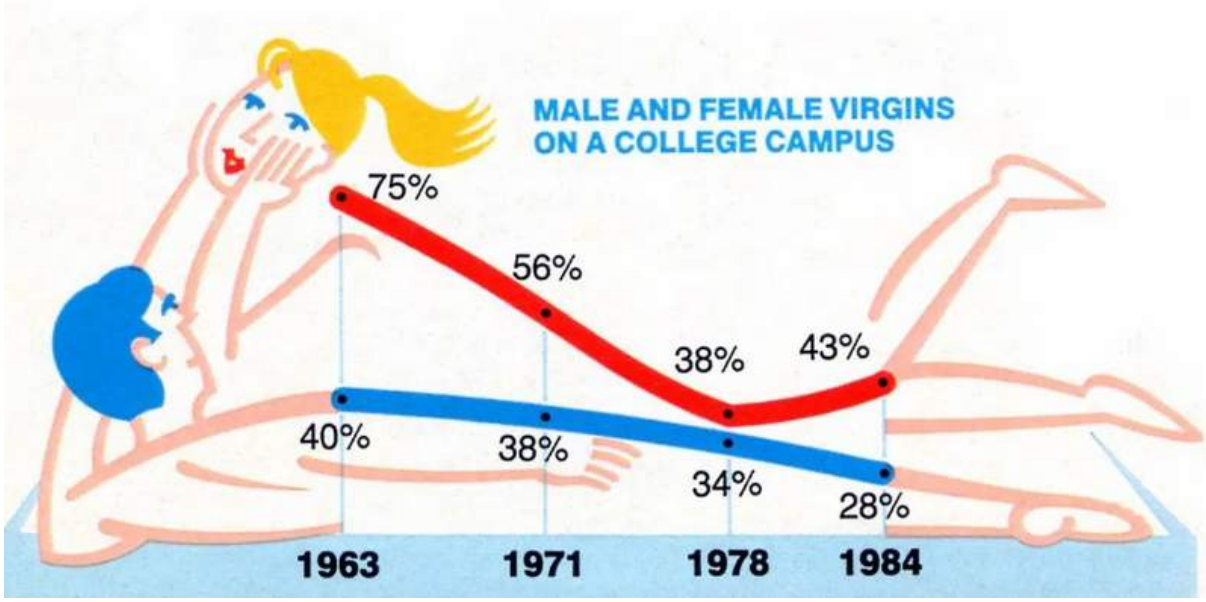


Figura 26: Gráfico publicada na Glamour Magazine na década de 1980, feito por Nigel Holmes.

Segundo as regras dispostas por Tufte, o que pensar deste trabalho pouco convencional?

Conectando-se com o público

O gráfico anterior foi produzido por **Nigel Holmes**, um designer gráfico que ficou conhecido por ter feito carreira na revista Time. Praticamente toda a sua produção é marcada por uma mistura entre ilustração e visualização de dados, praticamente a antítese de todos os conceitos que expusemos até aqui.

O trabalho de Holmes foi alvo de críticas ferozes por parte de Tufte, que destacou um gráfico em especial, mostrado a seguir, para um achincalhamento público.

Para Tufte, as peças apresentadas anteriormente são ultrajantes.

Em *Envisioning Information (1990)*, outro de seus livros, o pesquisador usa esse gráfico como exemplo de "chartjunk" que menospreza a audiência, que é tratada com obtusa e desinteressada.

Além disso, Tufte afirma que quem faz esse tipo de gráfico promove a ideia de que informação é algo chato, que precisa de decoração para ficar interessante.

Diante de uma crítica tão pesada de um nome tão relevante para o campo, o trabalho de Holmes se transformou em um exemplo do que **não fazer** na hora de desenhar gráficos.

Entretanto, apesar das críticas acadêmicas, Nigel Holmes seguiu fazendo sucesso, tornou-se diretor de arte da revista em que trabalhava e, posteriormente, um consultor independente. Alguns de seus gráficos, por bem ou por mal, estão entre os mais memoráveis da história da imprensa e do design.

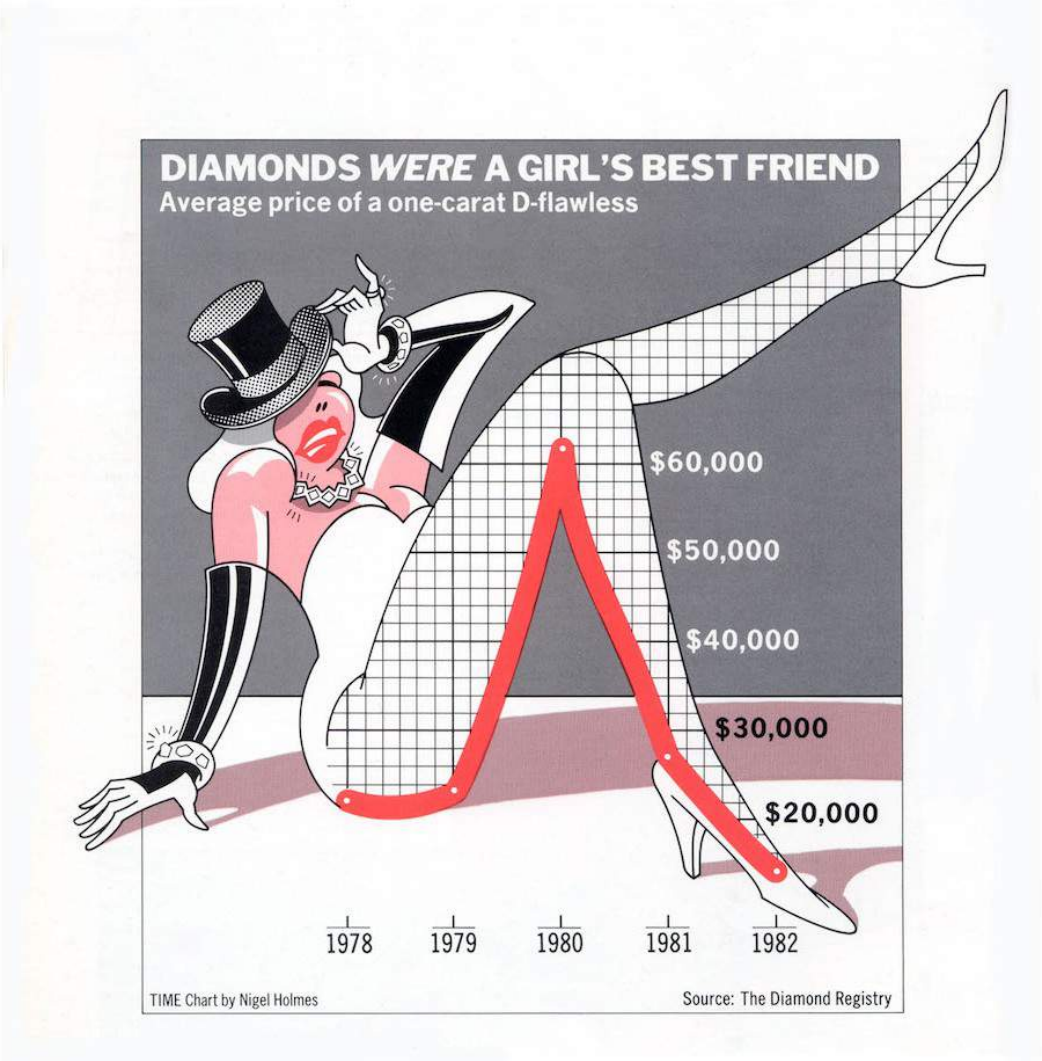


Figura 27: Gráfico publicado na revista Time em agosto de 1982, feito por Nigel Holmes.

Assim, vale a pena olhar também para essa perspectiva sobre a visualização de dados, diametralmente oposta aos pontos que elucidamos até aqui.

O que Holmes tem a dizer sobre as críticas que recebeu? Como justifica suas decisões criativas? O que pesquisas revelam sobre esse tipo de linguagem hoje, 30 anos depois da crítica violenta de Tufte?

O pesquisador Ricardo Cunha Lima, professor da Universidade Federal de Pernambuco (UFPE), analisou em um artigo de 2019 quais era os elementos típicos da linguagem de Holmes. Segundo ele, Holmes adotou o uso de metáforas visuais marcantes porque acreditava que esse tipo de conteúdo tornaria mais palatáveis temas complexos, com os quais os leitores provavelmente nunca tiveram contato algum antes.

Em uma palestra de 2015 na conferência OpenVis, em Boston, o próprio Holmes defende o uso de humor como forma de **conectar-se com a audiência** e explicar assuntos complexos de forma quase conversacional.

Desta forma, percebe-se que o objetivo do designer era comunicar informação relativamente complexa de forma mais aprazível para um público leigo.

O argumento central dessa posição, portanto, é que os elementos redundantes e decorativos não são descartáveis, mas sim formas de facilitar a compreensão sobre os temas tratados usando uma linguagem figurativa, explicativa e amigável.

Uma pesquisa feita por Scott Bateman e mais cinco acadêmicos da Universidade de Saskatchewan, no Canadá, colocou o conceito a prova.

Os estudiosos pediram para que voluntários avaliassem gráficos que representavam os mesmos dados, mas em duas versões diferentes: uma no estilo decorado de Holmes e outro em um formato minimalista que se aproximava mais das convicções de Tufte.

Os leitores eram questionados, logo depois de ler o gráfico, sobre os números que viram e outras questões de ordem mais técnica, como a tendência demonstrada pelos dados.

Além disso, os participantes responderam a perguntas sobre o tipo de gráfico que preferiam e se avaliavam que a imagem mostrava apenas dados objetivos ou se exibia também alguma forma de interpretação ou opinião.

Por fim, depois de um intervalo de cerca de duas semanas, os voluntários foram questionados sobre o que **lembravam** sobre cada um dos gráficos.

O estudo descobriu que não havia diferenças significativas na precisão da leitura dos gráficos no estilo de Holmes e no estilo minimalista - ou seja, os voluntários conseguiram ler os gráficos da mesma maneira, interpretando tanto gráficos decorados quanto gráfico sem “chartjunk” de forma correta na mesma proporção.

Entretanto, os gráficos no estilo cartum foram mais **lembrados** depois do intervalo longo, o que sugere que eles são mais **memoráveis** - ou seja, que a informação comunicada por visualizações nesse

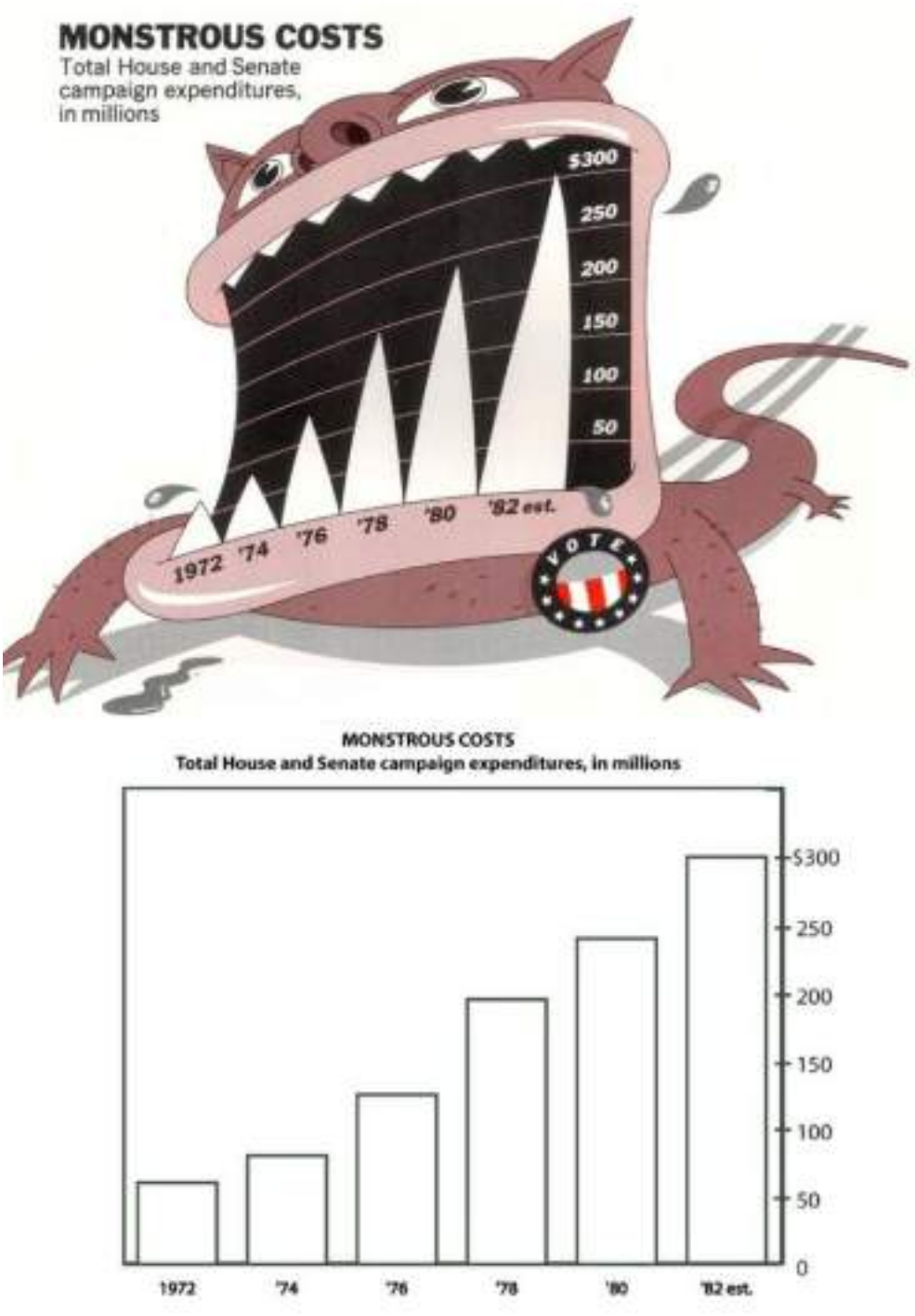


Figura 28: Exemplo de um gráfico com *chart junk* e um gráfico minimalista com os mesmos dados

formato são mais marcantes para o público.

Além disso, os participantes disseram gostar mais dos gráficos com elementos decorativos. Por fim, eles também notaram mais elementos interpretativos e opinativos nos trabalhos de Holmes - o que, pode-se argumentar, torna essas visualizações menos úteis como ferramentas que aumentam as capacidades de percepção do leitor.

Essa pesquisa, claro, não resolve debate algum. O texto, inclusive, recebeu pesadas críticas metodológicas de pesquisadores contemporâneos, como Stephen Few. Os achados servem, sobretudo, para mostrar que há espaço para discussão em torno do tema.

Tarefas perceptivas e gráficos

Depois de visitar essa quantidade monumental de conceitos, podemos finalmente listar alguns dos formatos mais comuns de gráficos.

Neste momento, vamos inverter o processo. Em vez de começar com uma lista de visualizações, vamos elencar algumas das **tarefas perceptivas** mais comuns na leitura de uma visualização de dados.

A apresentação segue essa ordem porque, para elaborar uma boa visualização de dados, é preciso ao menos ter alguma ideia prévia de sua **função** - ou seja, de que tipo de ferramenta mental o gráfico pretende ser.

Depois dessa reflexão, é necessário escolher um formato de representação funcional. Assim, **vamos descobrir quais são os formatos de gráficos que funcionam melhor para fazer comparações precisas ou ver a evolução de fenômenos ao longo, entre outras possibilidades**.

Vamos começar por formatos mais comuns, que costumam aparecer com frequência na televisão e nos jornais, em programas e notícias generalistas. Conforme a lista avança, porém, surgirão alguns formatos que não são tão usuais fora da análise de dados e da estatística.

Comparar valores de forma precisa

Essa, talvez, seja a tarefa perceptiva mais simples que uma visualização de dados pode ajudar a executar. Aqui, **o objetivo é fazer com que o leitor consiga contrastar de forma precisa o tamanho de poucas grandezas**.

Assim, faz sentido escolher formas visuais que tenham maior precisão na escala elaborada por Cleveland e McGill. Entre as formas de representação mais perto do topo, como vimos, estão aquelas que posicionam os dados em um eixo fixo.

Dessa forma, os formatos mais indicados para essa tarefa são **gráficos de barra** ou alguma de suas variantes. Veja a seguir uma visualização que compara a quantidade de imposto pago em relação a

renda em alguns países.

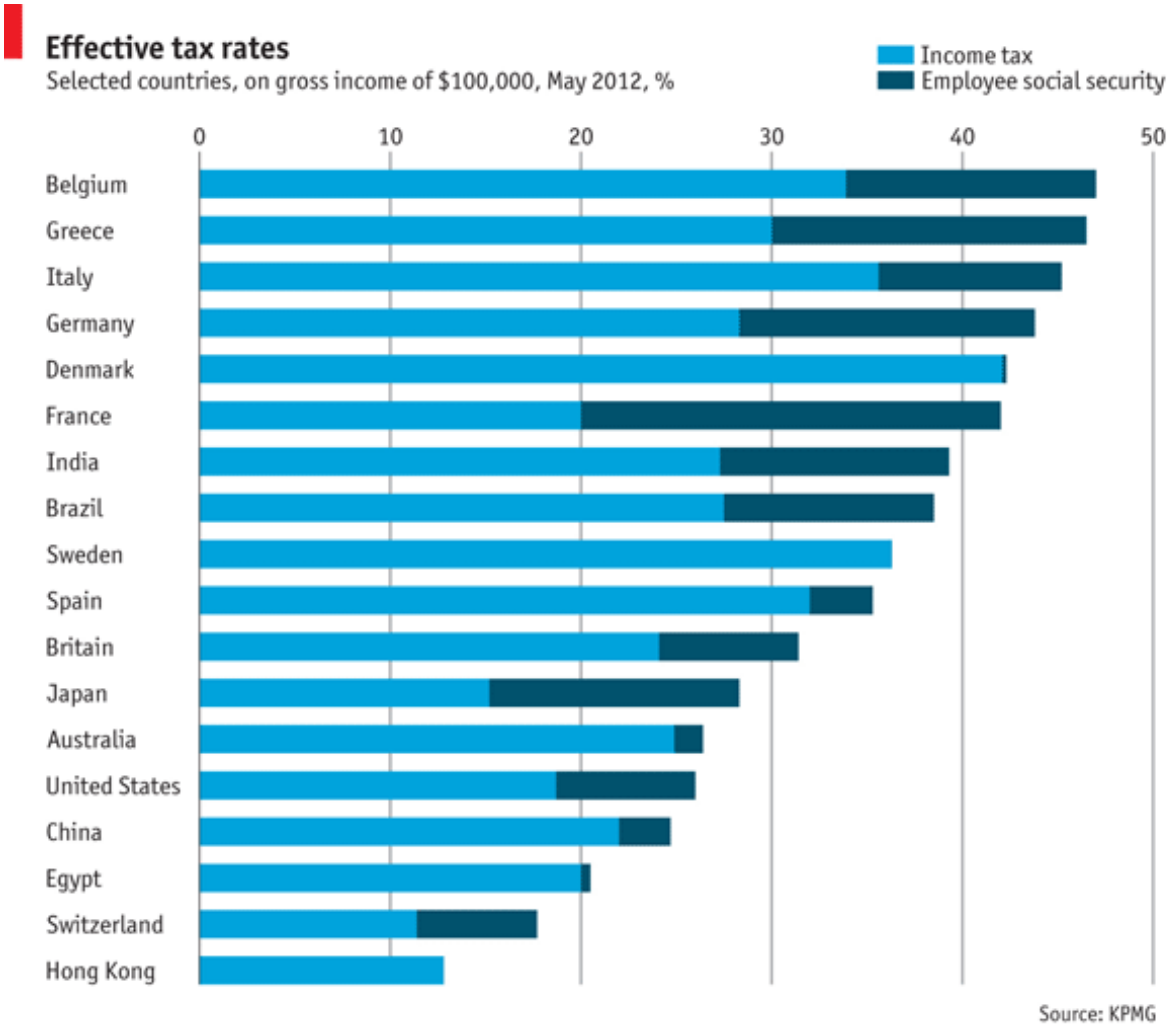


Figura 29: Gráfico de barras publicado na The Economist. Note que aqui a barra é composta por dois valores, cada um representado por uma cor. Na prática, eles se somam para representar uma barra única.

Para decodificá-la, o leitor precisa localizar o ponto final da barra e comparar a posição dela com a das demais, que estão na mesma escala. É fácil perceber, por exemplo, que Hong Kong tem uma carga tributária de perto de 12% e que a Suíça se aproxima dos 18%.

O exemplo é particularmente interessante porque funciona para fazer uma **comparação entre os diferentes países**, mas também permite enxergar a distribuição do imposto **dentro** de cada um deles. Vamos falar mais sobre essa segunda tarefa a seguir.

Comparar partes do todo

Aqui, a tarefa é parecida com a anterior, mas não estamos mais comparando valores independentes. Em vez disso, queremos saber como se distribuem, por exemplo, opiniões dentro de um grupo de pessoas.

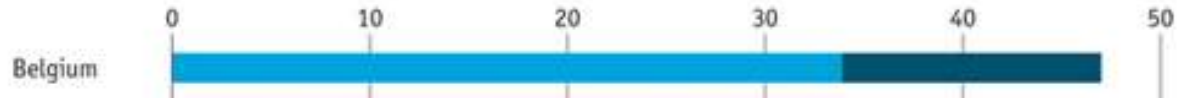
Talvez a forma mais usual de representar dados como estes seja um gráfico de pizza. Entretanto, ao

consultar o ranking da percepção visual, vemos que representações com base em ângulos não são boas para nem para comparações precisas nem para enxergar tendências gerais.

Esse tipo de gráfico fica bem no meio da escala, o que indica que não é muito bom para nenhum tipo de tarefa. Assim, não faz sentido usá-lo.

Por sorte, existem alternativas melhores. Vamos voltar para o gráfico de barras que mostramos acima, mas agora olhando para uma linha de cada vez.

Recapitulando: estamos vendo qual é a composição da carga tributária de um país. A parte **azul clara** mostra o **percentual referente ao imposto de renda**. A parte **azul escura** , o **percentual referente à seguridade social**.



Agora, para descobrir qual parte é maior, em vez de comparar ângulos, estamos comparando **comprimento**.

Ainda que os dois segmentos da barra não estejam alinhados verticalmente, é relativamente fácil perceber que a parte escura é cerca de três vezes menor que a parte clara.

Caso você tenha ficado curioso, o nome desse tipo de gráfico é **barra empilhada**, ou **stacked bar**.

A ideia é que as diferentes partes de um mesmo item sejam representadas por pedaços de reta colocados um em cima do outro. Assim, a barra pode ser lida a partir de cada segmento individual e também como a soma de todos eles.

Esse formato traz a vantagem, como vimos, de permitir a comparação da distribuição interna de diferentes grupos.

Se essa não for uma preocupação, porém, talvez o melhor a se fazer seja usar um gráfico de barras simples, separadas e alinhadas. Assim, ganha-se mais precisão.

Nesse caso, o formato não sugeriria que todos os itens fazem parte de um mesmo grupo, mas esse é um problema que pode ser superado facilmente com informações de contexto, com em um título ou legenda.

Enxergar tendências temporais

Na prática, quando queremos ver a evolução de um valor ao longo do tempo, poderíamos usar gráficos de barra, com uma barra para cada data. No final das contas, a tarefa executada é uma **comparação entre valores**.

Geralmente, ao analisar uma série temporal, não queremos que o foco de leitor seja na diferença **exata** entre uma data e outra. Usualmente, queremos que ele perceba **se o movimento dos dados é**

de subida ou de queda.

Assim, faz sentido usar um formato de visualização que realça a tendência e, de quebra, mostra que os pontos de dados estão conectados de forma linear: gráficos de linha.

Vamos analisar o gráfico abaixo, publicado pelo The New York Times em 2006.

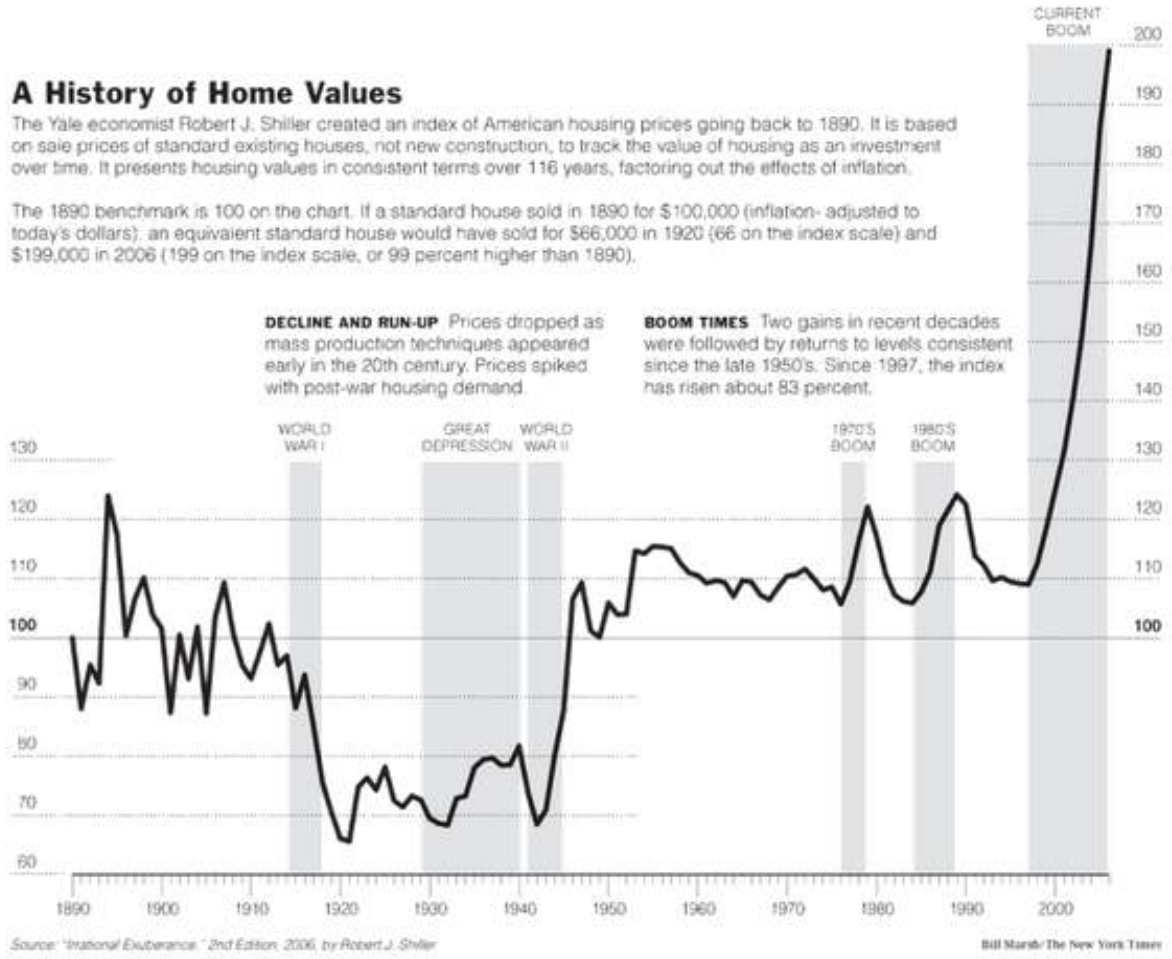


Figura 30: Gráfico de linhas apresenta tendências de ascensão e queda

Na prática, poderíamos substituir essa linha por uma série de barras ou pontos: o valor de cada ano é representado, afinal, pela posição do ponto mais alta da linha.

Ao conectar esses pontos, porém, o **tracejado sugere conexão e movimento**. Além disso, a inclinação gerada pelas ligações enfatiza a ideia de **queda e subida**, ainda que seja mais difícil perceber a variação exata de valor entre um ano e outro.

Perceba que aqui começamos a passar para a outra ponta da hierarquia da percepção, onde a preocupação passa a ser a percepção de tendências mais amplas e não a comparação exata entre valores.

No próximo item, vamos falar de forma mais genérica sobre a utilidade de representações desse tipo.

Enxergar tendências gerais

Parece um contrassenso escolher formas menos precisas quando temos à disposição gráficos que permitem fazer comparações precisas.

Quando, afinal, faz sentido usar as representações visuais que ocupam a parte de baixo do ranking elaborado por Cleveland e McGill?

Geralmente, isso acontece quando um olhar mais distante **revela padrões que não podem ser detectados comparando poucos pontos de dados**.

De vez em quando, é melhor perder precisão e ganhar escopo, permitindo assim que a **visualização de dados revele uma realidade mais ampla**.

Vamos olhar para um exemplos prático e marcante que mostra como uma abordagem menos precisa pode ser útil. A visualização de dados usa cores, tidas como a mais imprecisa das formas de representação, para oferecer um olhar original sobre um tema já muito explorado.

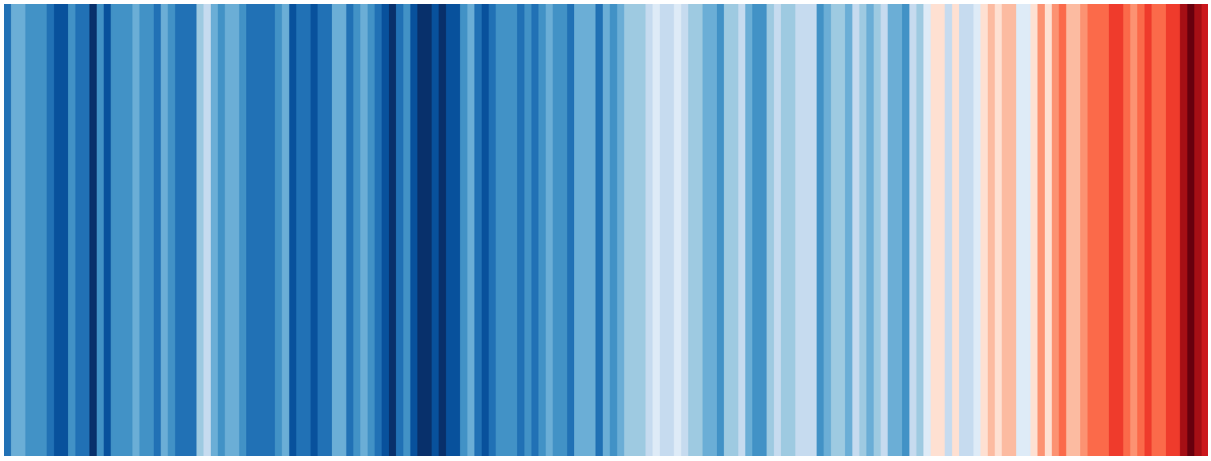


Figura 31: *Warming stripes*, listras do aquecimento, publicadas pelo climatologista Ed Hawkins em 2018.

O gráfico acima representa a evolução da temperatura média do planeta entre 1850 e 2018. Quando a temperatura de um ano está acima do valor médio registrado entre 1970 e 2000, ele é pintado em um tom de vermelho. Quando está abaixo, em um tom de azul. Quanto maior for a diferença, mais escura é a cor.

Como esperado, é muito difícil dizer **quantas vezes** a temperatura de um ano específico é maior do que a de qualquer outro. Contudo, o gráfico revela como a tendência de aquecimento é evidente.

Essa visualização funciona não apenas porque é esteticamente agradável, mas porque o autor reconheceu que o mais central em sua mensagem era realçar as **linhas gerais** do fenômeno.

Pouco importa quantos graus a mais o ano de 2008 teve em comparação ao ano de 1987, afinal. O que importa, em resumo, é que a década de 1980 foi mais fria que a década de 2000.

Logicamente, seria possível produzir vários outros gráfico que passe a mesma mensagem de forma

mais precisa, mas é difícil imaginar um formato que tenha o mesmo efeito intuitivo. Ao ver essa imagem, o leitor de imediato entende o que está representado.

Enxergar a distribuição

Por fim, vamos apresentar um formato muito parecido com o gráfico de barras, mas que não é exatamente a mesma coisa: o **histograma.**

Trata-se de um tipo de visualização que mostra qual é a **distribuição** dos dados, o que é essencial para entender melhor as características do fenômeno que estamos analisando.

Figura 32: Histograma com distribuição das notas no Enem

O gráfico anterior mostra a distribuição das notas no Enem 2018: muitos alunos tiraram entre 400 e 600. Poucos tiraram um valor perto de 800. No exemplo, o eixo horizontal representa a variedade possível de notas no Enem, de 0 a 800. O eixo vertical mostra quantos alunos tiraram cada nota nesse intervalo.

A barra mais alta, por exemplo, mostra que cerca de 200 mil estudantes tiraram uma nota próxima de 500. Essa é a **moda** , o valor mais comum do banco de dados.

O gráfico realça essa diferença e mostra, ainda, que uma quantidade muito pequena de alunos conseguiu notas próximas de 800 – ou seja, que essas altas pontuações são eventos raros que distorcem o cálculo das estatísticas descritivas.

Isso é evidenciado pelo valor da **média** , que está representada pela **linha preta pontilhada** , com um valor um pouco superior ao da moda.

Depois de ver um gráfico como esse, o leitor consegue uma percepção mais detalhada **de todo o banco de dados** e pode entender melhor quais são as tendências e quantidades envolvidas.

De teor um pouco mais técnico, esse tipo de gráfico não costuma aparecer muito em reportagens, apesar de ser muito útil na hora de analisar dados.

Com cuidado e atenção especial para as explicações, ele pode ser uma ferramenta útil para explicar fenômenos complexos para o leitor, como no exemplo abaixo, retirado de uma reportagem do Estadão sobre o impacto da desigualdade socioeconômica nos resultados do Enem.

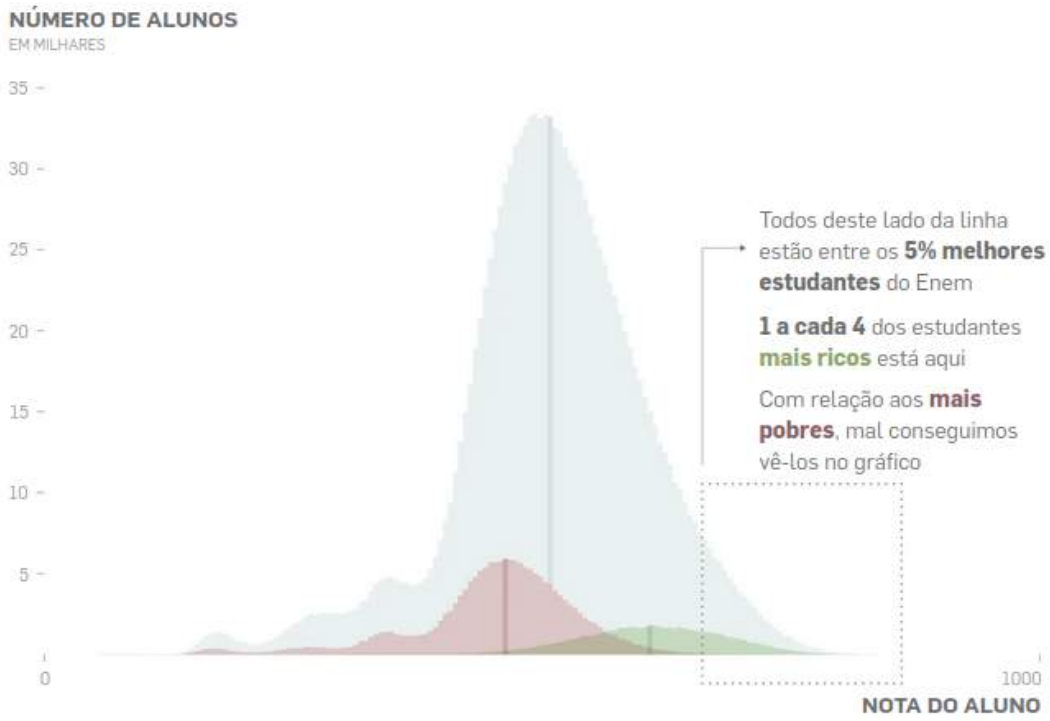


Figura 33: Histogramas sobrepostos para a comparação de distribuições

A sobreposição de três histogramas, no caso, revela que a distribuição de notas dos alunos mais ricos tende muito mais para a direita – e para as notas mais altas – do que a dos alunos mais pobres.

Ferramentas para visualização de dados

Depois da exposição teórica, é hora de criar os próprios gráficos. Existem inúmeros *softwares* para fazer visualizações de forma fácil e rápida.

Todavia, é importante dizer que esse tipo de ferramenta importa menos do que os princípios básicos

expostos anteriormente. O crucial é entender como gráficos funcionam e quais são os formas mais efetivas de usá-los.

Para isso, não há aprendizado melhor que **desenhar** com lápis e papel. Não é necessário riscar gráficos perfeitos e precisos, mas é útil tentar esboçar o que você espera construir antes de começar a clicar nos botões de algum aplicativo.

Ao fazer isso, é inevitável pensar nas tarefas perceptivas que a sua visualização de dados pretende auxiliar. Além disso, esse processo analógico ajuda a entender melhor as características de sua base de dados.

Sobre este último ponto, vale destacar algo que, de tão óbvio, muitas vezes acaba ignorado. Visualizações de dados são ferramentas que ajudam pessoas a entender melhor valores quantitativos, como vimos ao longo das páginas anteriores.

Separamos uma seleção de 10 ferramentas para visualização de dados, que são úteis tanto para quem está começando quanto para pessoas com mais conhecimento.

- **Flourish** (web): tem uma interface amigável e permite a criação de gráficos interativos, que podem ser unidos em histórias;
- **Datawrapper** (web): produz gráficos elegantes facilmente, com atualização em tempo real e edição colaborativa;
- **Raw Graphs** (web): bom para gráficos pouco usuais, é baseado em D3 (JavaScript) e tem código aberto;
- **GGplot2** (R): tem uma "gramática" para visualização de dados que é considerada referência, permitindo produzir gráficos muito elegantes em poucas linhas;
- **Tableau** (Windows/MacOS): fácil de usar, bom para análises exploratórias;
- **Matplotlib** (Python): bom para gráficos simples e análises exploratórias;
- **Seaborn** (Python): permite a criação de gráficos mais complexos;
- **Google Charts** (JavaScript): conta com diversos modelos/templates, além de gráficos responsivos;
- **D3** (JavaScript): permite a criação de isualizações complexas na web, conta com ampla documentação e exemplos;
- **C3** (JavaScript): baseado em D3, porém, mais simples, permite gerar visualizações interativas e responsivas.

No jornalismo, é comum tratar um gráfico como ponto final do processo de apuração, etapa em que aquilo que foi descoberto durante a análise de dados é comunicado para o leitor.

Pensar dessa maneira, porém, é ignorar um dos maiores potenciais da visualização de dados. Ainda durante a fase da análise de dados, vale a pena elaborar gráficos para entender melhor os números

com os quais você está trabalhando. Com paciência, tentativa e erro, é bem possível que tendências até então ignoradas aparecem, revelando novos ângulos para uma reportagem.

Conclusão

Ainda que este livro possa te dar um pontapé para começar a trabalhar com dados, nem de longe iremos esgotar todo o universo sobre o tema. Tanto por conta de sua complexidade própria quanto pelo ritmo incessante de novidades na área, a utilização de de dados na comunicação hoje é um campo em constante crescimento. E, como em muitas outras áreas, o melhor aprendizado se dá na prática.

Por isso, para terminar este guia, iremos compartilhar algumas dicas de como você pode continuar aprendendo. Na Escola de Dados, acreditamos que este processo de aprendizagem é impulsionado por experiências coletivas e comunitárias. Por isso, estimulamos encontros (como o Cerveja com Dados e a Conferência de Jornalismo de Dados e Métodos Digitais, o Coda.Br) e, em parceria com a Associação Brasileira de Jornalismo Investigativo (ABRAJI), mantemos o jornalismodedados.org.

Além de reconhecer trabalhos de excelência que podem servir de inspiração para outras pessoas com o Prêmio Cláudio Weber Abramo de Jornalismo de Dados, esta plataforma mantém um fórum aberto e gratuito para a resolução de dúvidas, discussões sobre tópicos de interesse e divulgação de oportunidades.

No Brasil, há uma comunidade de jornalistas de dados bastante colaborativa e disposta a ajudar na disseminação do conhecimento. Espaços como o Fórum de Jornalismo de Dados reúnem profissionais experientes e iniciantes em um mesmo espaço de troca de conhecimento. Há outros canais, como grupos abertos no Telegram ou no Facebook.

E o mesmo vale para linguagens de programação como o R e o Python, que possuem comunidades bastante ativas, como os grupos no Telegram R Brasil ou Pt-BR Data Science & Python. Há também canais como o Stack Overflow, um lugar que provavelmente terá a resposta para a sua dúvida com linguagens de programação e, caso não tenha, você pode postar sua questão para que outras pessoas possam te ajudar. Há um Stack Overflow em português, porém, assim como toda documentação da área de ciência de dados, a quantidade de conteúdos em inglês é infinitamente maior.

As comunidades das linguagens R e Python também realizam encontros no Brasil e no mundo. Esses encontros locais são uma excelente porta de entrada para você se aventurar no mundo da programação. Conhecendo e compartilhando experiências com as pessoas localmente, você irá se sentir muito mais motivado a aprender do que lutando contra as dificuldades sozinho com seu computador. Procure os encontros (meetups), grupos e ações das comunidades na sua cidade ou no seu estado.

Além dos grupos regionais, há também diversas comunidades voltadas a grupos específicos, com recorte de gênero ou raça, por exemplo. Infelizmente, apesar da contribuição decisiva de mulheres para o desenvolvimento da computação, a área tecnológica ao longo do século XX se tornou um campo predominantemente masculino, o que gera dificuldades na inserção de mulheres ainda hoje. Para superar esse desafio, foram criadas redes como **PyLadies** e **R-Ladies**.

A PyLadies é uma rede internacional sem fins lucrativos, com grupos locais em várias regiões do mundo. Seu objetivo é aumentar a atividade e a liderança das mulheres (cis ou trans) dentro da comunidade Python, por meio de atividades de divulgação, ensino, eventos, entre outras. E a R-Ladies é uma organização mundial que promove a diversidade de gênero na comunidade da linguagem R, por meio de encontros (meetups) e mentorias em ambientes seguros e amigáveis. O objetivo é promover a linguagem computacional R compartilhando conhecimento entre pessoas que se identifiquem com o gênero feminino (mulheres cis, mulheres trans, e também pessoas não-binárias e queer), independente do nível de conhecimento.

Ambas atuam hoje internacionalmente para fomentar comunidades de tecnologia mais plurais. E nos dois casos o Brasil ocupa um lugar de destaque. Em 2019, o país tinha o maior número de capítulos regionais da PyLadies no mundo e a R-Ladies foi fundada por uma brasileira nos idos de 2012, a estatística Gabriela de Queiroz, que atualmente trabalha na IBM. Também no Brasil surgiram experiências que buscam fomentar uma maior diversidade racial na área, como **AfroPython**.

Essas comunidades servem como canais para a troca de experiência e colaboração entre as pessoas. Além de estudar e praticar, participar delas sem dúvida vai ter ajudar a quebrar barreiras e superar dificuldades. Esses espaços servem para tirar dúvidas e tomar conhecimento sobre eventos e atividades da área, mas também existem diversos cursos que podem te dar um empurrão inicial. Em Python, há opções em português como o Python para Zumbis e, no R, as formações do Curso-R.

A Escola de Dados também oferece cursos e conta com diversos tutoriais gratuitos. Outras referências interessante são os cursos da Abraji e do Knight Center for Journalism in the Americas, que frequentemente abordam temas relacionados ao jornalismo de dados.

Por fim, se você gostou deste guia e quer apoiar o trabalho da Escola de Dados, fica aqui nosso convite para o nosso programa de membresia. Nele, você terá acesso a uma newsletter exclusiva para ficar por dentro das principais atualizações da área, além de receber descontos nos cursos e participar de canais exclusivos com nossa equipe.

Referências

- 5-star Open Data - Tim Berners-Lee
- 30 anos: o quanto a Constituição preserva de seu texto original - Nexo
- A fundamental way newspaper sites need to change - Adrian Holovaty
- A journalist's guide to cognitive bias (and how to avoid it) - Paul Bradshaw
- Bad Data Handbook - Q. Ethan McCallum
- Base de dados de cartórios traz falhas que impedem calcular efeito real do coronavírus no Brasill - Folha de S.Paulo
- Basômetro - Estadão
- Bellingcat's Online Investigation Toolkit - Bellingcat
- Busca Precedentes - Controladoria Geral da União
- Calculadora de correção de valores - Banco Central
- Cartilha Técnica para Publicação de Dados Abertos no Brasil - Governo Federal
- Com Bolsonaro, liberação de agrotóxicos cresceu 42%, diz estudo - Uol
- Consulta E-sic - Controladoria Geral da União
- Entrevistando planilhas: estudo das crenças e do ethos de um grupo de profissionais de jornalismo guiado por dados no Brasil - Marcelo Träsel
- Expressão regular pode melhorar sua vida - Escola de Dados
- Envisioning Information - Edward Tufte
- Guia para Quartz para Limpeza de Dados
- How to prevent confirmation bias affecting your journalism - Paul Bradshaw
- How you're feeling when machine learning might help - Jeremy Merrill
- 'I Don't Like Maths, That's Why I am in Journalism': Journalism Student Perceptions and Myths about Data Journalism - Amy Schmitz Weiss e Jéssica Retis
- Inflação - IBGE
- Machine bias risk assessments in criminal sentencing - ProPublica
- Metáforas e gráficos pictórico-esquemáticos de Nigel Holmes - Ricardo Cunha Lima
- Monitor da Violência - G1
- No Enem, 1 a cada 4 alunos de classe média triunfa. Pobres são 1 a cada 600 - Estadão
- Numbers in the newsroom - Sarah Cohen
- Operadores de busca avançada - Escola de Dados

- O que revela uma análise das emoções dos candidatos durante o debate - Estadão
- O uso de fontes documentais no jornalismo guiado por dados - Marília Gehrke
- Panama Papers - ICIJ
- Pockets - The Pudding
- Precision Journalism: a reporter's introduction to social science methods - Philip Meyer
- Radmesser - Tagesspiegel
- Roubos de celular atingem metade das ruas de São Paulo - Estadão
- Siga os Números - Mariana Segala
- Spurious Correlations - Tyler Vigen
- The chartjunk debate: a close examination of recent findings - Stephen Few
- The Commercial and Political Atlas - William Playfair
- The Functional Art - Alberto Cairo
- The Truthful Art - Alberto Cairo
- The Largest Vocabulary In Hip Hop - The Pudding
- The next wave of data journalism? - Paul Bradshaw
- The New York Times Privacy Project
- The Visual Display of Quantitative Information - Edward Tufte
- Tutoriais da Escola de Dados
- Uma Por Uma - Uol
- Sou uma cientista de dados cética quanto aos dados - Andrea Jones-Rooy
- What is open data? - Open Knowledge

Apoiadores

Esta publicação foi realizada graças ao apoio de 282 pessoas, que colaboraram com a primeira campanha de financiamento colaborativo da Escola de Dados em 2019.

Adriana Farias

Adriana Ramos Guarani Kaiowá

Ailma Teixeira

Aisten Baldan

Alberto Oliveira

Alessandra Ferreira

Alessandra Midori

Amanda Carvalho

Amanda Lemos

Amanda Moura

Amanda Rossi

Ana Beatriz Assam

Ana Carolina Castro

Ana Freitas

Ana Laura Azevedo

Ana Paula de Oliveira

Ana Paula Ferreira

Andre Spritzer

André Tamura

Andre Mota

Ângela Prestes

Angélica Martins

Anna Livia Arida

Anne Clinio

Anselmo Dias

Antonio Junior

Antônio Prestes

Augusto Conconi

Augusto Fadel

Barbara Blum

Barbara Castro

Barbara Jambwisch

Bárbara Libório

Bianca Berti

Bruno Luiz

Bruno Teixeira

Bruno Lemos

Bruno Evangelista

Bruno Silva

Bruno De Barros

Bruno Nunes

Café.art.br

Caio do Nascimento

Carol Moreno

Carol Pacobahyba

Carol Queiroz

Carolina Assunção

Carolina Ribeiro

Caroline Franco

Cecília Do Lago

Celia Beatriz Rosemblum

Cesar de Oliveira

Christian Moryah

Christina Brentano

Clara Sacco

Clarissa Mendes

Colaboradados

Cristian Weiss

Cristiane Paião

Cristiano Pavini

Daniel Silveira

Daniel Mariani

Daniel Bramatti

Daniel de Augustinis

Daniel Dieb

Daniel Fireman

Daniel Portugal

Daniel Schwabe

Daniel Torquato

Danilo Torini

Dênis Tavares

Denise Businaro

Denise Neumann

Diana Yukari

Diego Oliveira

Diogo do Carmo

Diogo Mafra

Diogo Pires

Edna Simao de Souza

Edson Alves Jr

Eduardo Belo

Eduardo Almeida

Elisângela Mendonça

Emanuele dos Santos

Eric Daher

Érico Lima

Everton Alvarenga

Fabiana Cambricoli

Fabiana Pulcineli

Fabiano Silos

Fábio de Araujo

Fabio Takahashi

Felipe Grandin

Felippe Mercurio

Fernanda Campagnucci

Fernanda Távora

Fernando Barbalho

Filipe da Silva

Filipe Augusto

Filipe José Quintans

Filipe Menezes

Flavia Alemi

Flavio Dal Pozzo

Flávio Passos

Flavio Shirahige

Francisco de Oliveira Júnior

Francisco de Lima Moacyr Filho

Gabriel Santo

Gabriel Cortilio

Gabriela Caesar

Géssica Gonçalves

Gilberto Vieira

Gisele Barros

Giulia Afiune

Glaucia Campregher

Guilherme Duarte

Guilherme Mendes

Guilherme Storck

Gustavo Macedo

Gustavo Tosello

Guto Schultz

Haruan Mossato

Helder da Rocha

Helio Miguel

Henrique Dentzien

Henrique França

Henrique Parra

Iago Bolívar

Ígor Passarini

Ilito Torquato

Isabella Sander

Isis Reis

Ivan Lemos

Ivana Bentes

Jamile Santana

Jardel Duque

Jean Prado

João Vitor Freitas

João Vitor Rodrigues

Josir Cardoso Gomes

Joyce Heurich

Juan Torres

Juciano

Julia Mente

Juliana Marques

Juliana Lopes

Juliana Almeida

Juliana Vitor

Júlio Boaro

Júlio Oliveira

Karina Barcellos

Karla Mendes

Katia Brembatti

Keila Guimarães

Larissa Brainer

Leandro Filippi

Leonardo Germani

Leonardo Mendes Júnior

Ligia Guimarães

Liraucio Girardi Júnior

Lucas Gelape

Lucas Palma

Lucas Rizzi

Luciana Junqueira

Lui Pillmann

Luís Guilherme Julião

Luiz Claudio Reis

Luiz Fernando Teixeira

Luiz Fernando Toledo

Luiz Farias

Luiza Bodenmüller

Manoel Galdino

Marcela Miller

Marcelo da Fontoura

Marcelo Dias

Marcelo Granja

Marcelo Lima

Marcelo Oliveira

Marcelo Soares

Marcelo Träsel

Márcio Viana

Marco Antonio Konopacki

Marco Túlio Pires

Marcos André

Marcos Côrtes

Marcos de Salles

Maria Rosa

Mariah Queiroz

Mariana Tiemi

Marianna Araujo

Marilene Oliveira

Marília Gehrke

Marina Merlo

Mário Sérgio

Marlise

Marlos Pereira

Mateus Sousa

Matheus de Araújo

Max Stabile

Melissa Lüdeman

Michele Vieira

Moisés de Melo

Nitai da Silva

Patricia Bado

Patrícia da Silva

Pedro Burgos

Pedro Capetti

Pedro Maia

Pedro Hazan

Pedro Markun

Pedro Renaux Wanderley

Pedro Franco

Pedro Sarvat

Pedro dos Reis

Philippe Watanabe

Priscila Santos

Rachel Domingues

Rafael Vazquez

Rafael Moreira

Rafael Kenski

Rayan Alves

Reinaldo Silva

Renata Hirota

Renata Montechiare

Renato Rebelo

Revista AzMina

Ricardo Rossetto

Robson da Rosa

Rodolfo Almeida

Rodrigo Cunha

Rodrigo Guedes

Rodrigo Coelho

Rodrigo Schuinski

Rodrigo Takenouchi

Rosangela Lotfi

Rose do Nascimento

Rosental Alves

Rui Barros

Rute Pina

Sérgio Lüdtke

Sérgio Seabra

Sérgio Spagnuolo

Silvana Fernandes

Silvia Follador

Stefano Wrobleski

Stephanie de Paula

Suzana Barbosa

Tadeu Junior

Tadeu Teixeira

Taís Seibt

Talita Duvanel

Tatiana Balachova

Tatiana Coelho

testecolab

Thays Lavor

Thiago de Moraes

Thiago Corte

Thomaz Barbosa

Thomaz Rezende

Tiago Pereira

Tiago Rogero

Tomas Martinez

Verónica Goyzueta

Victor Grinberg

Vinícius Valle

Vitor Mafra

Vitor Paulos Bellini

Vívian Vieira

Viviane Machado

Yan Dutra Hill

Yuri Almeida