UN PASO A PASO BÁSICO PARA PERIODISTAS

TUTORIAL: LIMPIEZA Y ANÁLISIS DE DATOS

CON EJEMPLO PRÁCTICO
#PERIODISMODATOS

Elaborado por: Hassel Fallas @HasselFallas,

Editora Inteligencia de **Datos de La Nación**, Costa Rica

datacamp combiná transformá construí

28 al 30 de Julio
Montevideo

HasselFallas

LA NACIÓN DATA
Costa Rica

Introducción:

"Hubo una época en que todo lo que se requería para ser periodista era dedicación a la verdad, mucha energía y algo de talento para escribir. Todavía se necesita de eso, pero ya no es suficiente. Hoy el mundo se ha complicado, es tan explosivo el crecimiento de la información disponible que el periodismo necesita ser tanto un filtro como un transmisor; un organizador y un intérprete tanto como quien recopila y entrega los hechos (…) En resumen, un periodista debe ser un gerente de bases de datos, procesar y ser analista de data", Phillip Meyer, The New Precision Journalism

¿Qué es Periodismo de datos? En mi experiencia es construir una investigación de interés público a partir del análisis de bases de datos, contengan estas millones, miles o cientos de registros. Los resultados se evidencian en publicaciones que pueden incluir visualizaciones de datos (interactivas primordialmente, pero no exclusivamente) aplicaciones de noticias y el acceso al público de la matriz de datos del proyecto.

Afortunadamente, cada vez es más frecuente encontrarme con periodistas interesados en aprender a analizar bases de datos para producir reportajes. Ese proceso, ciertamente, es exigente, no solo en tiempo y dedicación al proceso sino en conocimientos sobre Estadística, Minería de Datos, técnicas de análisis, extracción de datos, Business Intelligence y herramientas (software), entre otros.

Aprender a hacer Periodismo de Datos también es un proceso, uno de aprendizaje continuo, de entender tanto el valor de un número como su significado y su contexto. De no olvidar que siempre lo más importante es y será el periodismo, el interés público y las historias humanas detrás de los datos.

Siempre he creído en que la mejor manera de empezar a hacer Periodismo de Datos es simplemente empezando, sin poner excusas, ni dejar para mañana lo que puede iniciarse hoy.

Esta guía pretende ser eso: un punto de partida en el camino por el que hoy usted decidió transitar. Incluye respuestas a algunas cosas que, en su momento, fueron un dolor de cabeza en mi propia experiencia con bases de datos en Excel.

En este documento encontrará: Cómo abrir un archivo CSV en Excel; cómo limpiar datos con los comandos Buscar y Reemplazar, cómo separar fechas de horas y cómo desagregar las fechas en día, mes y año. También cómo emplear filtros, crear una tabla pivote para entrevistar sus datos y combinar variables con el objetivo de ampliar la profundidad de las respuestas que buscará. Finalmente, algunas herramientas gratuitas para visualización de datos.

¡Feliz inicio del viaje!

[Hassel Fallas](#)
www.hasselfallas.com

¿Qué es Periodismo de Datos?

"**El Periodismo de Datos es**: **traspiración** en el 80%, **buenas ideas** en un 10% y **producción** en otro 10%"

Simon Rogers, exdirector de Data de The Guardian, hoy editor de Data en Google

#Periodismodatos

Tutorial de limpieza de datos y análisis básico

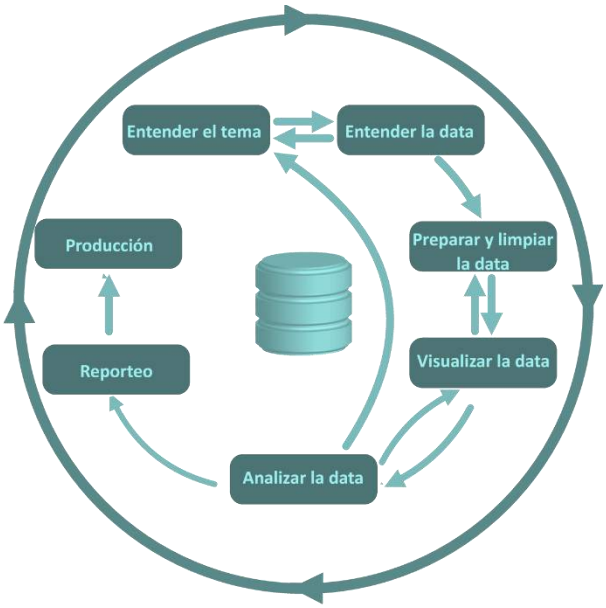
Elaborado por: Hassel Fallas @HasselFallas, Editora Inteligencia de Datos de La Nación, Costa Rica

¿Qué es un dato?

«Un dato es una representación simbólica (numérica, alfabética, algorítmica, espacial, etc.) de un atributo o variable cuantitativa o cualitativa»

Añado: debe ser estructurado, comprensible para ser procesado por una computadora. Un dato por sí solo no dice mucho, comparado y en contexto es capaz de contar la mejor de todas las historias.

El ciclo del trabajo con datos:



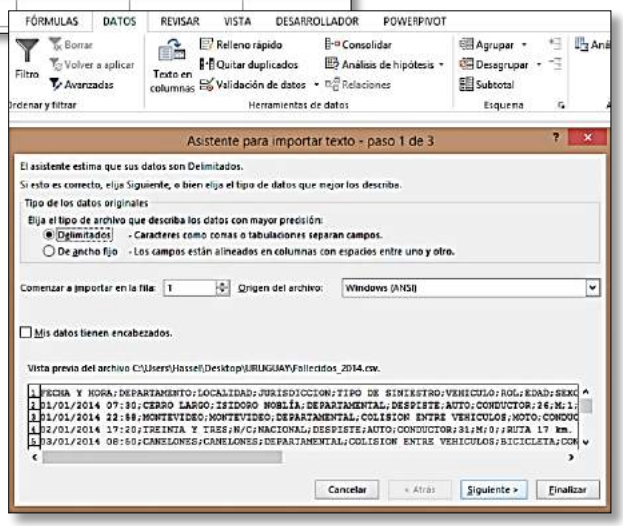
Método adaptado para periodismo usando como base el **Cross Industry Standard Process for Data Mining**, (**CRISP-DM)**

Paso a paso:

1. **Descargue** el archivo para el ejemplo y también sus metadatos: https://catalogodatos.gub.uy/dataset/personas-fallecidas-en-siniestros-de-transito
   (Si no funciona utilice este link: http://unasev.gub.uy/inicio/sinatran/datos_abiertos/2014/ o este otro: https://www.dropbox.com/s/covxi58rn5bv1rz/Fallecidos_2014.csv?dl=0)

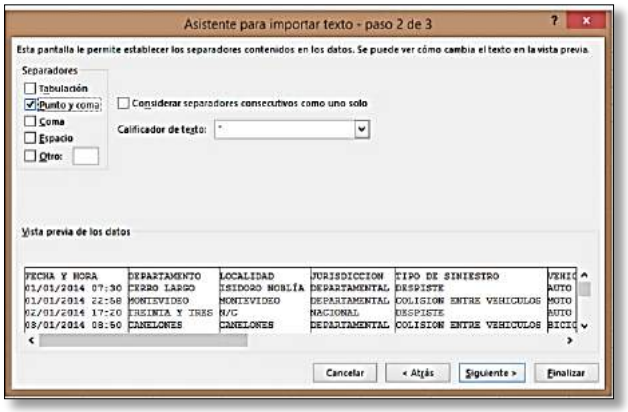
2. Se trata de un archivo en formato **CSV** (Comma Separated Values o valores separados por comas)

3. Para acceder al CSV, abra **Excel**, en la pestaña Datos seleccione, **Obtener datos externos desde texto**



4. Seleccione el archivo ***Fallecidos 2014*** desde la carpeta donde lo descargó.

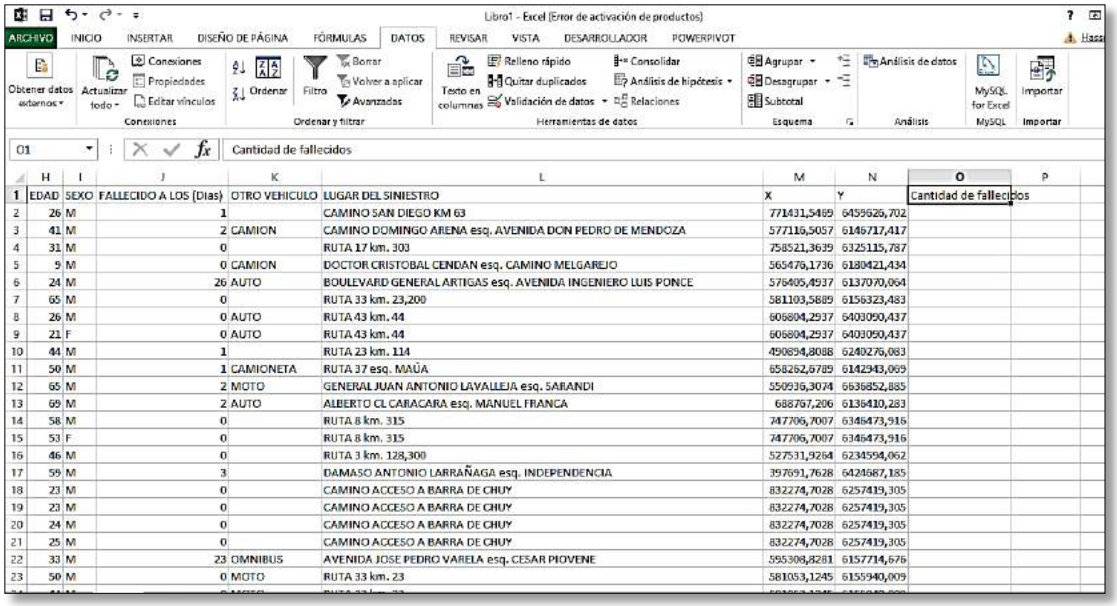
5. Cuando salga esta pantalla, de clic en **siguiente**:
6. Elija el separador por **punto y coma**



7. Su archivo está listo para empezar la **exploración** en Excel.

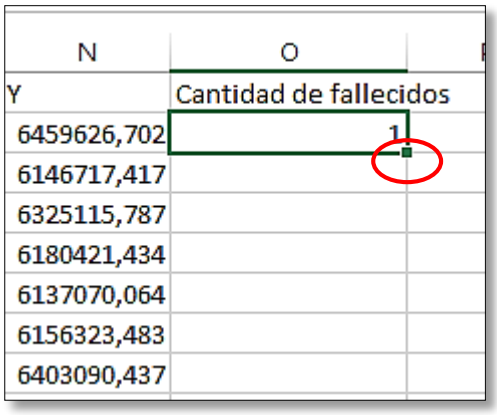
8. **Exploración**: Abra **la *metadata*** de su archivo o el diccionario de datos. Recuerde **siempre solicitarlos**, le facilitarán la comprensión de las variables incluidas en la **base de datos**.

9. **Primer paso:** Aunque fácilmente podríamos saber la cantidad de fallecidos que hubo en accidentes de tránsito, pues cada registro en la base de datos identifica a una única persona, para efectos de este ejercicio crearemos una variable que se llame: Cantidad de fallecidos. Para hacerlo, colóquese sobre la columna O y escriba en la primera celda el rótulo: **cantidad de fallecidos.**



	H	I	J	K	L	M	N	O	P
1	EDAD	SEXO	FALLECIDO A LOS (Dias)	OTRO VEHICULO	LUGAR DEL SINIESTRO	X	Y	Cantidad de fallecidos	
2	26	M	1		CAMINO SAN DIEGO KM 63	771431,5469	6459626,702		
3	41	M	2	CAMION	CAMINO DOMINGO ARENA esq. AVENIDA DON PEDRO DE MENDOZA	577116,5057	6146717,417		
4	31	M	0		RUTA 17 km. 303	758521,3639	6325115,787		
5	9	M	0	CAMION	DOCTOR CRISTOBAL CENDAN esq. CAMINO MELGAREJO	565476,1736	6180421,434		
6	24	M	26	AUTO	BOULEVARD GENERAL ARTIGAS esq. AVENIDA INGENIERO LUIS PONCE	576405,4937	6137070,064		
7	65	M	0		RUTA 33 km. 23,200	581103,5889	6156323,483		
8	26	M	0	AUTO	RUTA 43 km. 44	606804,2937	6403050,437		
9	21	F	0	AUTO	RUTA 43 km. 44	606804,2937	6403050,437		
10	44	M	1		RUTA 23 km. 114	490894,8088	6249276,083		
11	50	M	1	CAMIONETA	RUTA 37 esq. MAÚA	658262,6789	6142343,069		
12	65	M	2	MOTO	GENERAL JUAN ANTONIO LAVALLEJA esq. SARANDI	550936,3074	6636852,885		
13	69	M	2	AUTO	ALBERTO CL CARACARA esq. MANUEL FRANCA	688767,206	6136410,283		
14	58	M	0		RUTA 8 km. 315	747706,7007	6346473,916		
15	53	F	0		RUTA 8 km. 315	747706,7007	6346473,916		
16	46	M	0		RUTA 3 km. 128,300	527531,9264	6234594,062		
17	59	M	3		DAMASO ANTONIO LARRAÑAGA esq. INDEPENDENCIA	397691,7628	6424687,185		
18	23	M	0		CAMINO ACCESO A BARRA DE CHUY	832274,7028	6257419,305		
19	23	M	0		CAMINO ACCESO A BARRA DE CHUY	832274,7028	6257419,305		
20	24	M	0		CAMINO ACCESO A BARRA DE CHUY	832274,7028	6257419,305		
21	25	M	0		CAMINO ACCESO A BARRA DE CHUY	832274,7028	6257419,305		
22	33	M	23	OMNIBUS	AVENIDA JOSE PEDRO VARELA esq. CESAR PLOVENE	595308,8281	6157714,676		
23	50	M	0	MOTO	RUTA 33 km. 23	581053,1245	6155940,009		

10. Luego, en la celda O2, escriba un 1



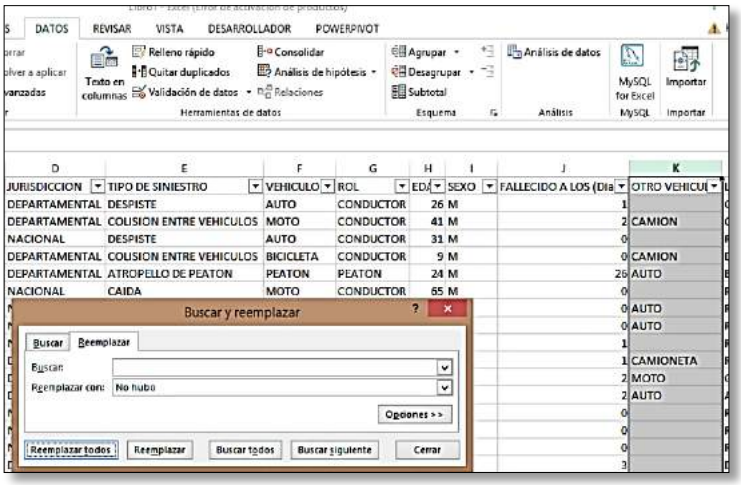
N	O
Y	Cantidad de fallecidos
6459626,702	1
6146717,417	
6325115,787	
6180421,434	
6137070,064	
6156323,483	
6403090,437	

Coloque el cursor sobre el pequeño cuadro verde o negro que aparece al final de la celda, de doble clic. Eso le permitirá registrar con un uno a cada persona fallecida. De esta forma se nos facilitará el conteo del total de muertos y el cruce con otras variables cuando lleguemos al final del ejercicio, en la parte de crear tablas dinámicas.

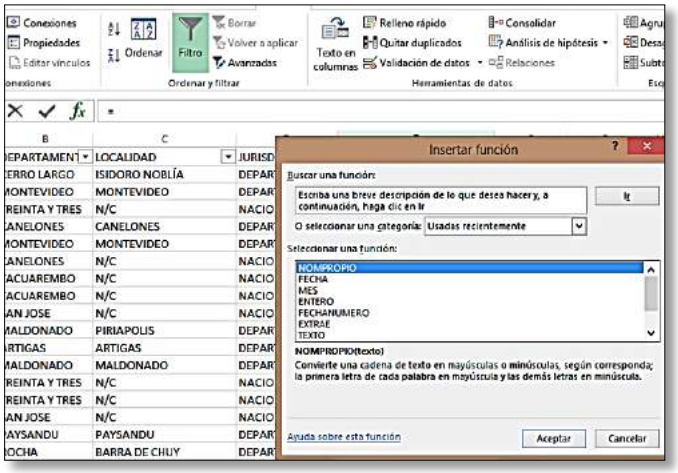
11. Limpieza:

¿En qué variables hay campos vacíos?

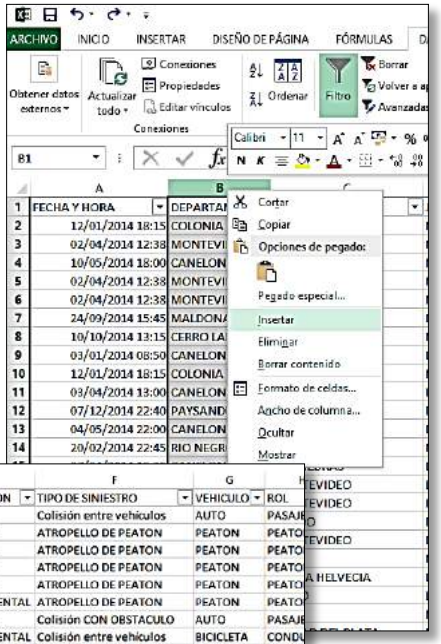
- ✓ **Variable otro vehículo**.
- ✓ Seleccione toda la columna K.
- ✓ Comando CTRL B: **REEMPLAZAR**.
- ✓ **Buscar:** espacio en blanco. Reemplazar: **No hubo**.
- ✓ Clicar sobre **Reemplazar todos**



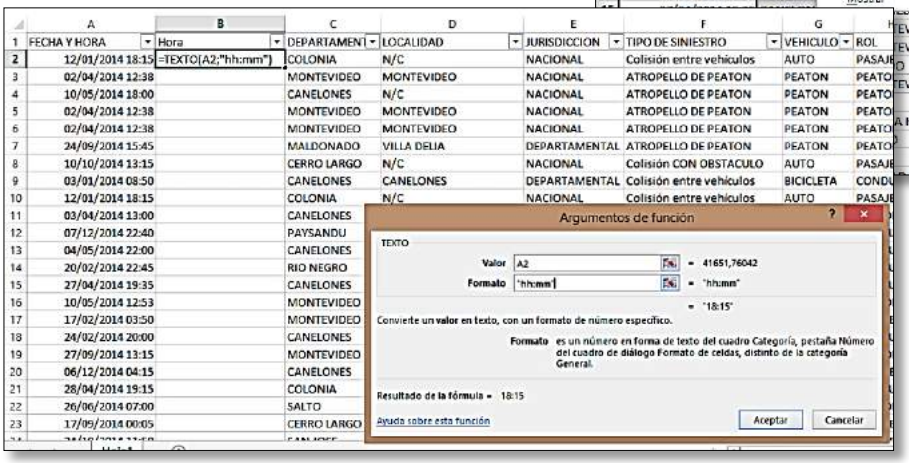
12. Convertir texto en mayúsculas en un formato de altas y bajas
    Marque toda la columna que desea cambiar. **Use la función: NOMPROPIO**



13. **Fechas:** Algunas veces nos conviene analizar por separado fechas y las horas. Para separarlas añadimos una columna más al lado de la que queremos dividir.



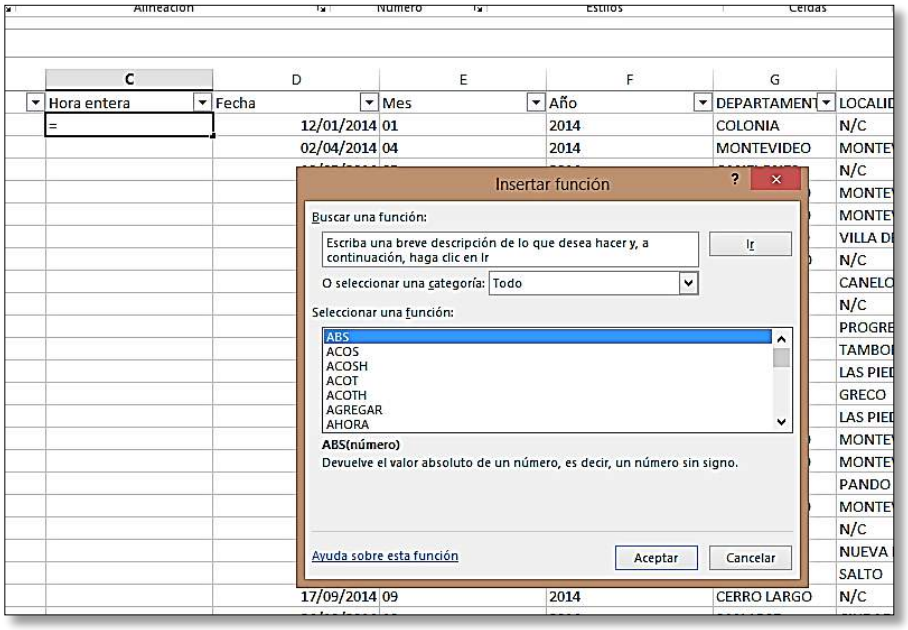
14. **Separar la hora**. Use la función: **=TEXTO(A2;"hh:mm")**



15. Es probable que nos interese analizar las horas en que más accidentes se producen y para hacerlo requerimos la hora entera, sin sus minutos. Para conseguirlo emplearemos la función **Extrae**

En el panel superior izquierdo, haremos clic sobre este ícono:

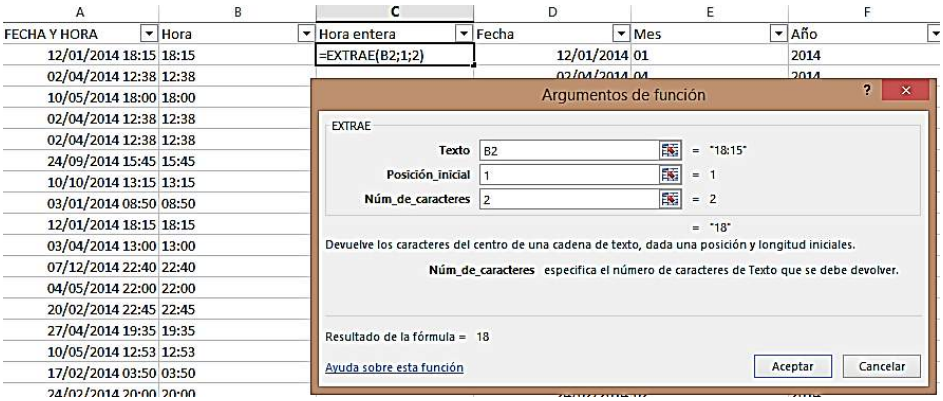
De inmediato se abrirá una ventana como esta:



Allí, en el espacio debajo de la instrucción **Buscar una función**, digite el nombre de la que necesitamos: **EXTRAE** y de inmediato pulse sobre Ir. Excel lo llevará directamente a la pantalla donde deberá digitar la orden para cumplir la función.

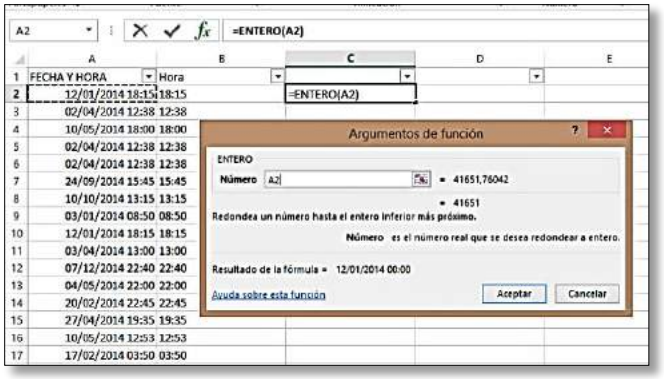
Si su versión de Excel no le retorna la función que necesita, búsquela usando la categoría TODO en el menú desplegable ubicado debajo del espacio **para buscar función**.

16. **Ejecutando la función:** Elija la celda que contiene el primer dato del que extraeremos la información. En **Texto**, ubique esa celda. En posición inicial digite 1, en número de caracteres digite 2, que corresponden a la cantidad de números que necesita extraer (también es posible hacerlo con letras).

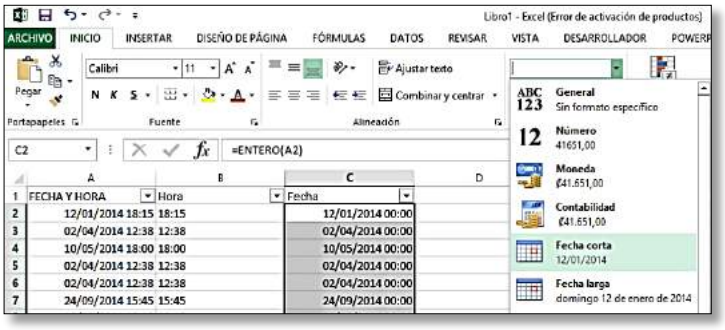


Como resultado obtendrá un número entero para las horas del 00 al 23.

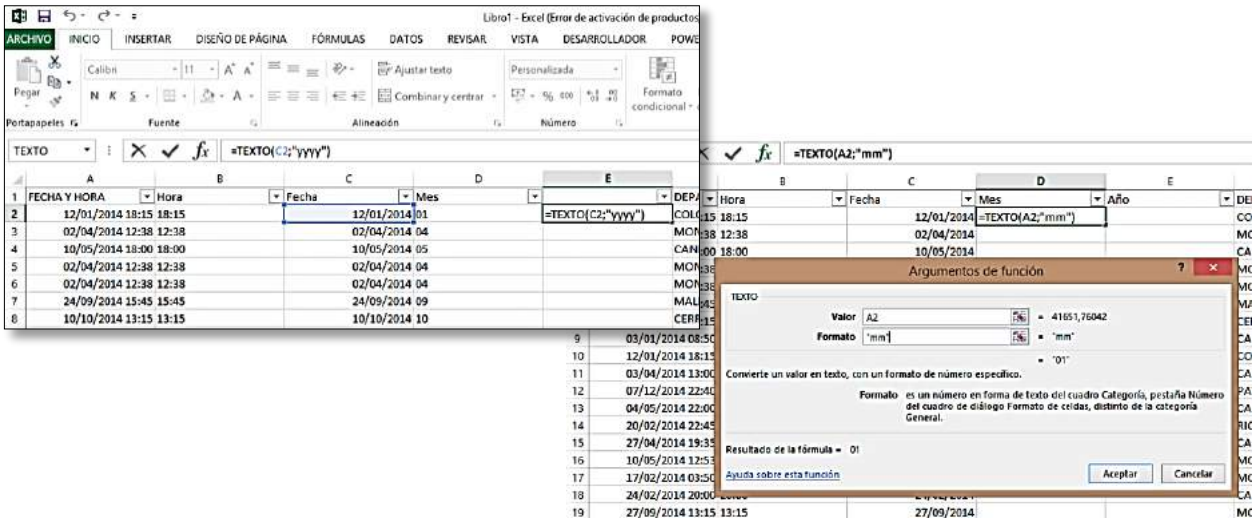
17. Para separar la fecha usamos la función **Entero**



18. Una vez que tenemos la fecha como un número entero, le damos **Formato de Fecha Corta** para eliminar la hora



19. Para extraer de la fecha el **Mes** y el **Año**, y convertirlas en variables para análisis, usamos, también, la función **Texto**



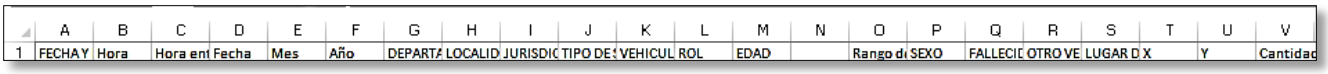
20. Usando los filtros

Los filtros son indispensables para encontrar información de interés (subconjuntos de datos) dentro de la gran masa de cifras en nuestra hoja de cálculo. Con ellos podemos ordenar y determinar, fácilmente, los valores más altos o más bajos en la tabla.
Asimismo, filtrar por un único nombre o número sobre el que tengamos alguna curiosidad particular.

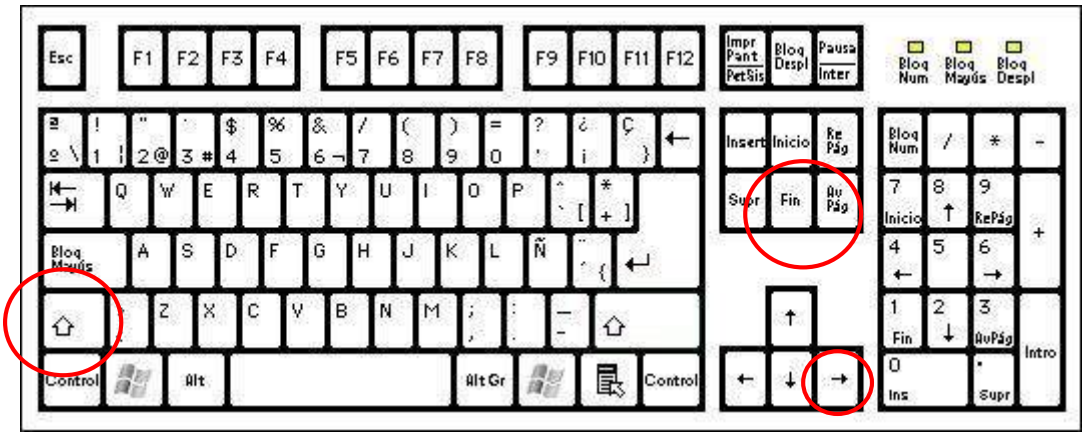
El siguiente ejemplo le dará más luz al respecto.

Supongamos que usted precisa de saber **cuántos fallecidos** hubo en el 2014 únicamente en el departamento de **Montevideo** y cuántos de ellos murieron en una **colisión entre vehículos**.

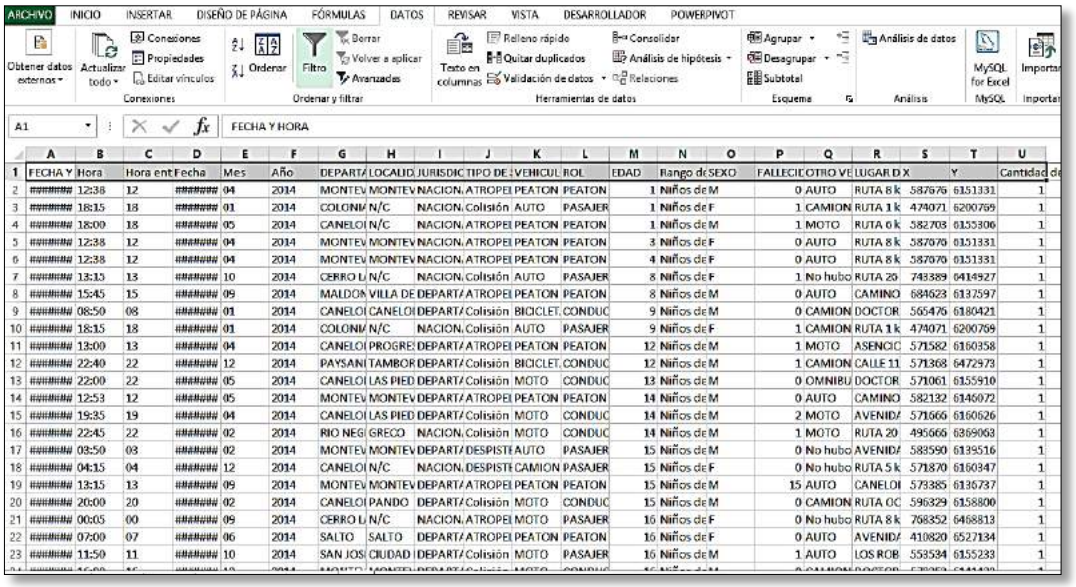
Lo resolveremos creando un filtro. Para hacerlo es indispensable que posicione el cursor de su mouse en la **celda A1**, donde inician sus datos y además seleccionar todos los rótulos de datos que componen la tabla.



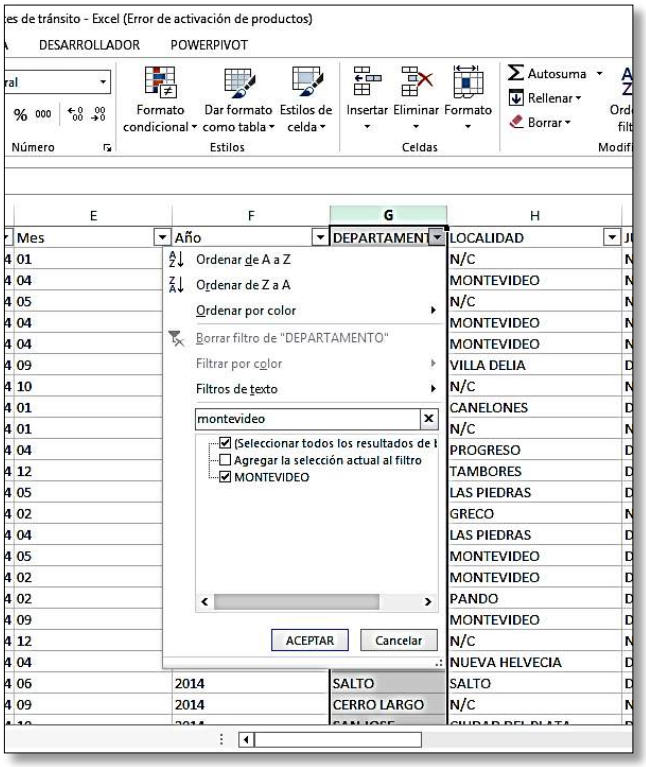
Una vez en la **celda A1**, presione y mantenga presionada la tecla SHIFT, luego presione la tecla FIN, suéltela y finalmente, clique sobre la flecha hacia la derecha.



Una vez marcados todos los rótulos de datos en su base, vaya a la pestaña Datos y elija la opción Filtro.



Vaya a la columna rotulada como **Departamento**; de clic en el pequeño triángulo negro del filtro, seguidamente, en el campo **Buscar** digite Montevideo. Pulse sobre **Aceptar** y a continuación, su filtro le permitirá ver los datos, únicamente de Montevideo.



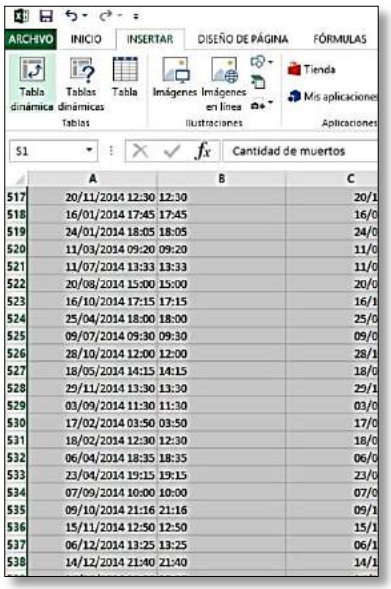
Repita el procedimiento solo que esta vez para la columna rotulada como Tipo de Siniestro, en la opción buscar del filtro digite: Colisión entre Vehículos.

Vaya a la columna Fallecidos y con el comando explicado en el paso 18 **(SHIFT-FIN-FLECHA HACIA ABAJO, esta vez)** cuente la cifra de fallecidos en colisiones de vehículos en el Departamento de Montevideo en 2014: 90.

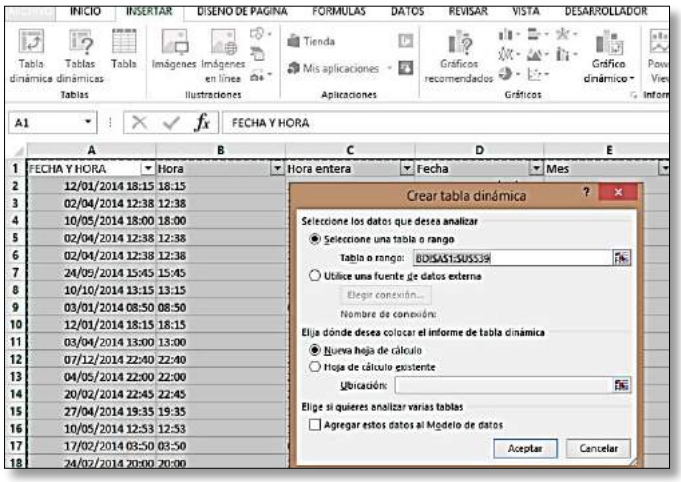
21. Análisis

Tablas Pivote: Construyamos una para analizar datos

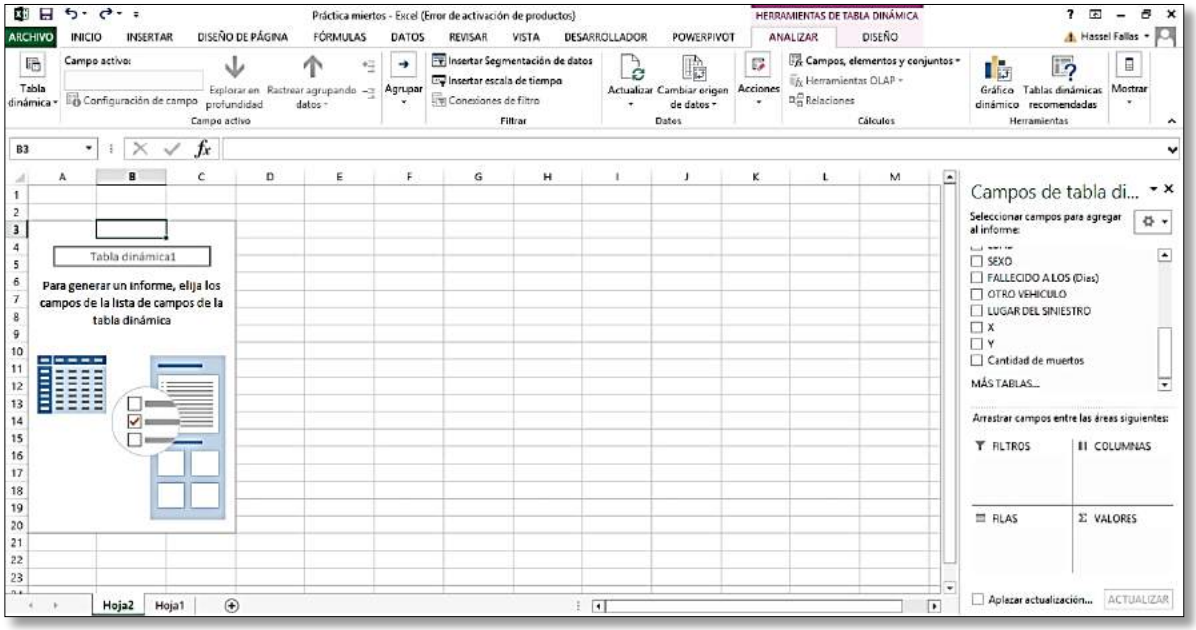
- ✓ Marque todas las variables que componen su base
- ✓ **Consejo práctico:** Para seleccionar toda la data: SHIFT (LA MANTIENE MARCADA), FIN (suelta la techa), finalmente: FLECHA HACIA ABAJO



22. Una vez seleccionada toda la data, elija en la pestaña **Insertar** la opción **Tabla dinámica**



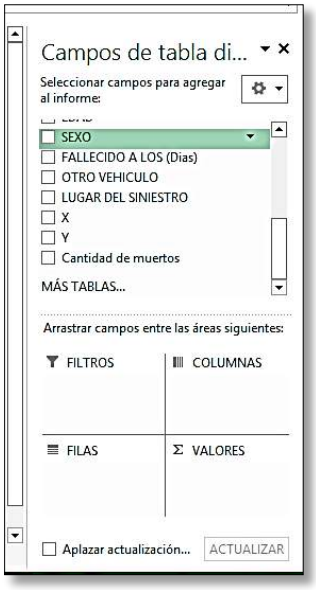
23.**Así luce una tabla dinámica**, una herramienta indispensable para encontrar respuestas a nuestras preguntas sobre datos, **rápidamente**.



24. Entrevista a los datos

Las tablas dinámicas nos permiten **arrastrar y colocar** las variables para analizarlas.

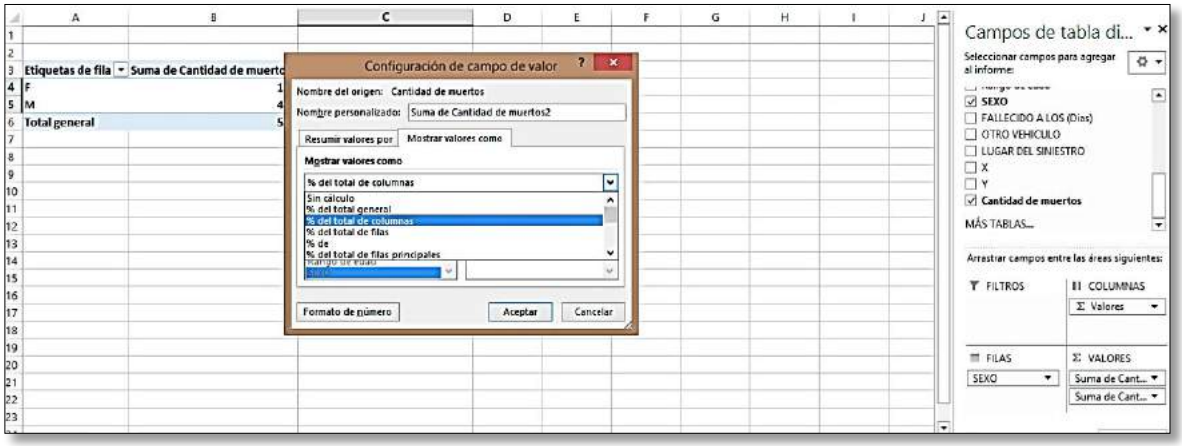
- ✓ Por ejemplo si queremos saber **la cantidad de hombres y mujeres** que fallecieron en accidentes de tránsito en Uruguay en el 2014; arrastramos **la variable Sexo** al **campo de Filas** y la **variable Cantidad de fallecidos** hasta el **campo de valores.**
- ✓ Obtendrá una tabla como esta:



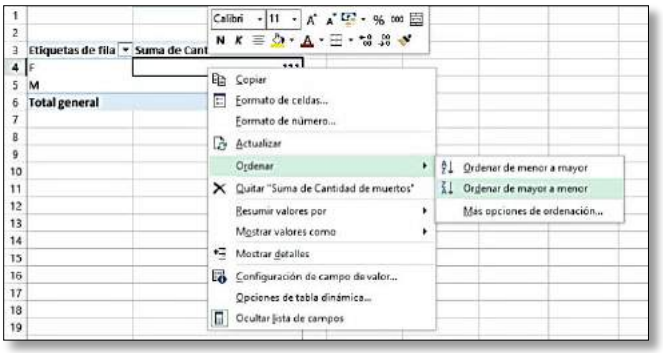
Etiquetas de fila	Suma de Cantidad de muertos
F	111
M	427
Total general	538

25. Si desea **ver la proporción (%)** de los datos respecto del total, arrastre nuevamente la **variable Cantidad de muertos** al campo **Valores**, de clic sobre el pequeño triángulo negro a la derecha y elija la opción: **Configuración de campo de valor**.

26. En la pestaña **Mostrar valores como** escoja: **% del total de la columna**.



27. **Un truco para ordenar los valores** (+ a -) rápidamente: **colóquese sobre la celda** que contiene el primer dato, con **botón derecho** del mouse busque ordenar y elija la opción que más le convenga (ordenar de menor a mayor u ordenar de mayor a menor).



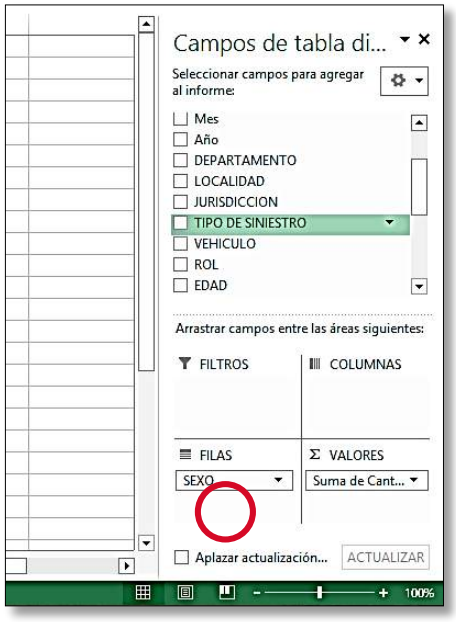
28. Combinando variables

En una tabla dinámica podemos ampliar la profundidad de las respuestas que buscamos por medio de la combinación de variables.

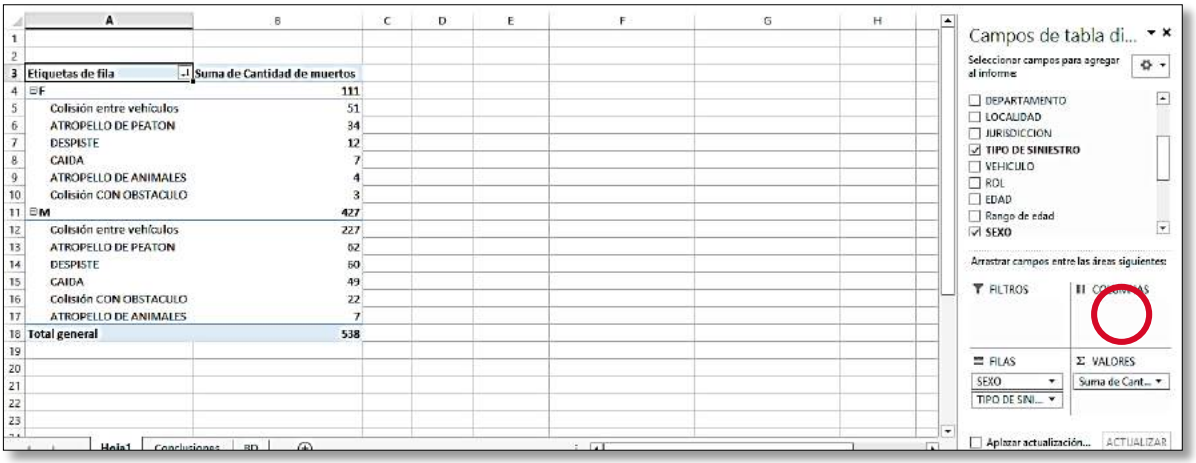
29. Por ejemplo: ya sabemos la distribución de muertos en accidentes de tránsito según su género, ahora queremos saber:

- ✓ ¿En qué tipo de accidente es más común que mueran los hombres y en cuál las mujeres?
- ✓ Además, ¿cuál era el rol de la víctima al momento del percance?

30. Arrastramos **la variable Tipo de siniestro** al **campo de Filas** y la colocamos **debajo de la variable Sexo**.



31.Luego, arrastramos la **variable ROL** al **campo de columnas**



32. Obtendrá una tabla como esta:

Suma de Cantidad de muertos	Etiquetas de columna			
Etiquetas de fila	CONDUCTOR	PEATON	PASAJERO	Total general
F	38	34	39	111
Colisión entre vehículos	28		23	51
ATROPELLO DE PEATON		34		34
DESPISTE	3		9	12
CAIDA	6		1	7
ATROPELLO DE ANIMALES			4	4
Colisión CON OBSTACULO	1		2	3
M	325	59	43	427
Colisión entre vehículos	208		19	227
ATROPELLO DE PEATON	2	59	1	62
DESPISTE	40		20	60
CAIDA	48		1	49
Colisión CON OBSTACULO	20		2	22
ATROPELLO DE ANIMALES	7			7
Total general	363	93	82	538

33. Herramientas para visualización

- ✓ http://www-969.ibm.com/software/analytics/manyeyes/
- ✓ https://datawrapper.de/

¡Gracias!

Para más información sobre periodismo de datos y herramientas, puede visitar: www.hasselfallas.com

¡Extra!

Para más comprensibles los porcentajes a la hora de hacer periodismo de datos

¿Cómo se puede expresar un porcentaje como una tasa?:

Por ejemplo:
0,05 es uno de cada veinte (5%)
0,10 es uno de cada diez (10%)
0,20 es uno de cada cinco (20%)
0,25 es uno de cada cuatro; o un cuarto (25%)
0,33 es uno de cada tres (33%)
0,40 es dos de cada cinco (40%)
0,5 es uno de cada dos (50%)
0,6 es tres de cada cinco (60%)
0,66 es dos de cada tres (66%)
0,75 es tres cuartos, o tres de cada cuatro (75%)
0,8 es cuatro de cada cinco (80%)
0,9 es nueve de cada diez (90%)
0,95 es diecinueve de cada veinte (95%)

0,49 es "casi la mitad"

Cualquiera que sea su cifra, busque la proporción más cercana a ella y úsela como referencia apoyándose en expresiones como "más de" o "casi".

Por ejemplo:
0,06 es "más de uno de cada veinte"
0,09 es "casi uno de cada diez"
0,23 es "más de uno de cada cinco"
0,27 es "más de la cuarta parte"
0,36 es "más de uno de cada tres"
0,39 es "poco menos de dos de cada cinco"

Tomado del libro: Finding stories in spreadsheets de Paul Bradshaw

Extra 2
Nombres de las funciones en Excel en Inglés y Español
http://excel.facilparami.com/2012/09/nombre-en-espaol-ingles-de-las-funciones-de-excel/