

Összefüggések statisztikai elemzése

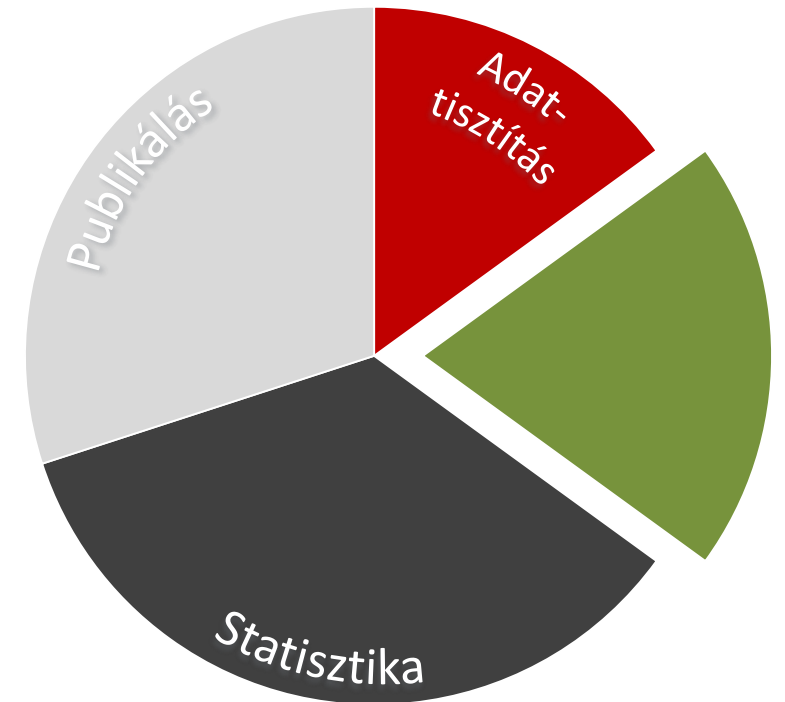
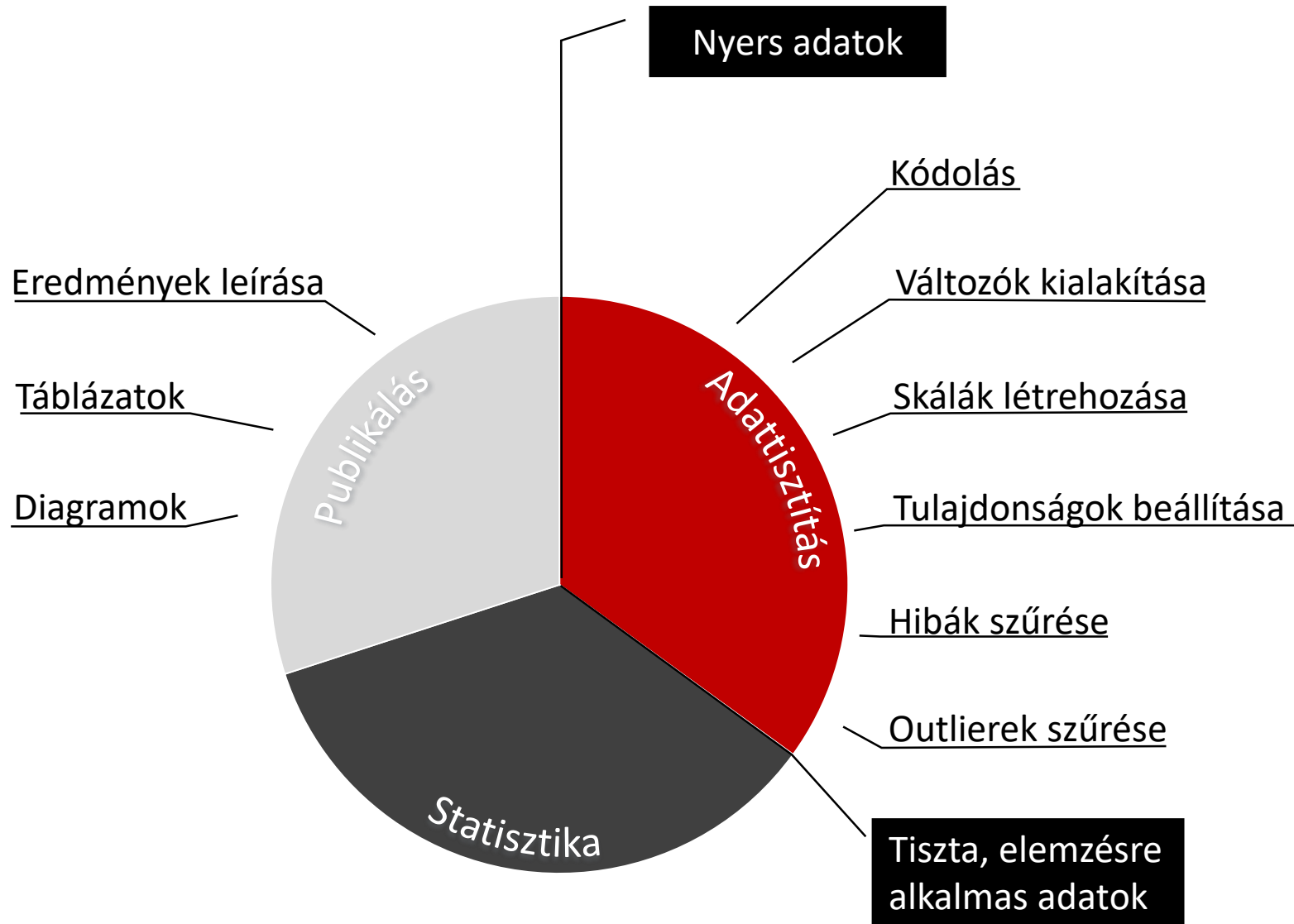
BBNPS01000

Készítette: Soltész-Várhelyi Klára

Adatfeldolgozás

Adattisztítás – hibák és outlierok

Adattisztítás



Egyéni tapasztalataim alapján az adatok Google formról való letöltése és az első hipotézis tesztelése közötti munkák teszik ki általában egy statisztikai elemzés egyharmadát. Jól átgondolt kísérlettervezéssel ez az idő rövidíthető le.

Az órán használt adatbázishoz tartozó kérdőív

Jelige: _____

Szemek teszt pontszáma (a vizsgálat nevezője tölti ki): _____

Kérjük adja meg korát: _____

Nem:

férfi

nő

Legmagasabb iskolai végzettség:

általános iskola

középiskola vagy gimnázium

főiskola, egyetem vagy magasabb

Családi állapot:

egyedülálló

párkapcsolatban élő

házas

özvegy

egyéb: _____

Facebook ismerősök száma: _____

Kérjük, tekintse át Facebook ismerőseit, és sorolja fel azon személyeket, akikre igaz a következő állítások valamelyike! Elég, ha a nevek Ön számára azonosíthatók. Olyan személyt is megemlíthet, aki nincs Facebook ismerősei között!

1. Merne tőle nagyobb összeget kérni. Elmondaná neki, ha súlyos betegsége lenne. Kérne tanácsot tőle egy súlyos probléma esetén:

2. Legalább néhány hetente tartják a kapcsolatot, és életét viszonylag jól ismeri, illetve ők ismerik az ön életét:

Kérjük, karikázza be azt a válaszlehetőséget, mely legjobban leírja, hogyan érezte magát az elmúlt 2 hét során!

Az elmúlt két hét során érezte-e magát...	egyáltalán nem jellemző	alig jellemző	jellemző	teljesen jellemző
1. vidámnak és jókedvűnek?	0	1	2	3
2. nyugodtnak és ellazultnak?	0	1	2	3
3. aktívnak és élénknek?	0	1	2	3
4. ébredéskor frissnek és élénknek?	0	1	2	3
5. A napjai tele voltak számára érdekes dolgokkal?	0	1	2	3

Beolvasás, változók tulajdonságainak beállítása

- Az előző adatbázis segítségével megismerkedtünk az SPSS kezelésének alapjaival, most a megszerzett tudást felhasználva végig haladunk az adatfeldolgozás lépésein úgy, ahogy azt egy kutatás során érdemes tenni.
- 1. Olvasd be az adatbázist!
 - Fájl: statgyak03_adatfeldolgozás2_adattisztitas.xlsx
- Az Excelfájlban található kódkönyv segítségével állítsd be a változók tulajdonságait!
 - Töltsd ki a címkéket (Labels)!
 - Kategoriális változóknál adj értékcímkét (Value labels)!
 - A nagykör változónál van hiánykód, ezt állítsd be (Missing)!
 - Állítsd be a változók mérési szintjét (Measures)!

Hiba

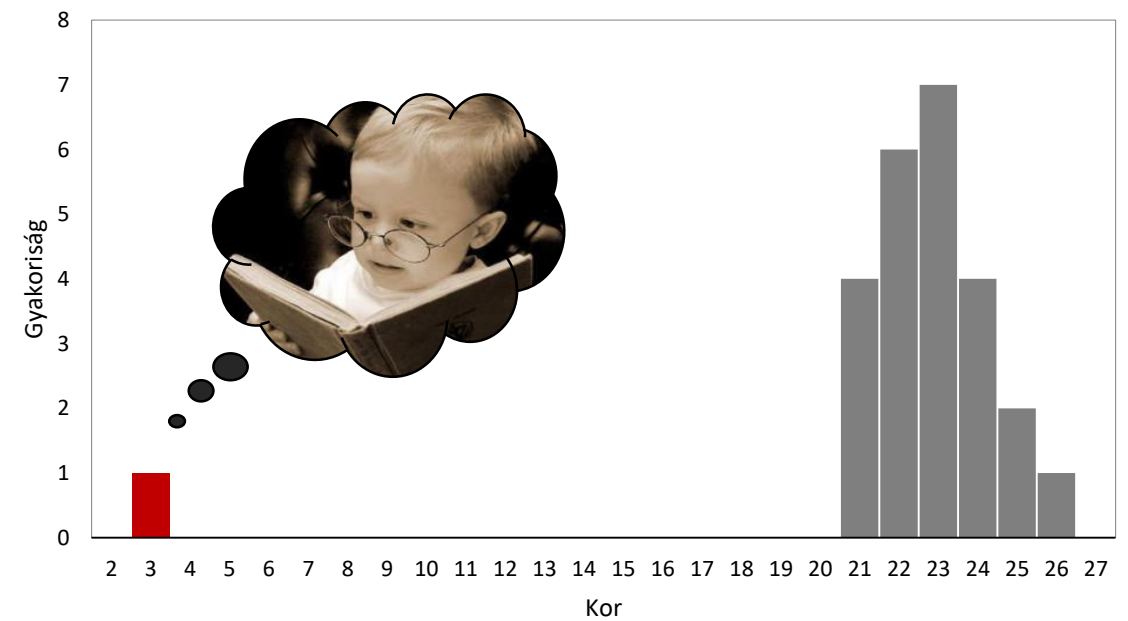
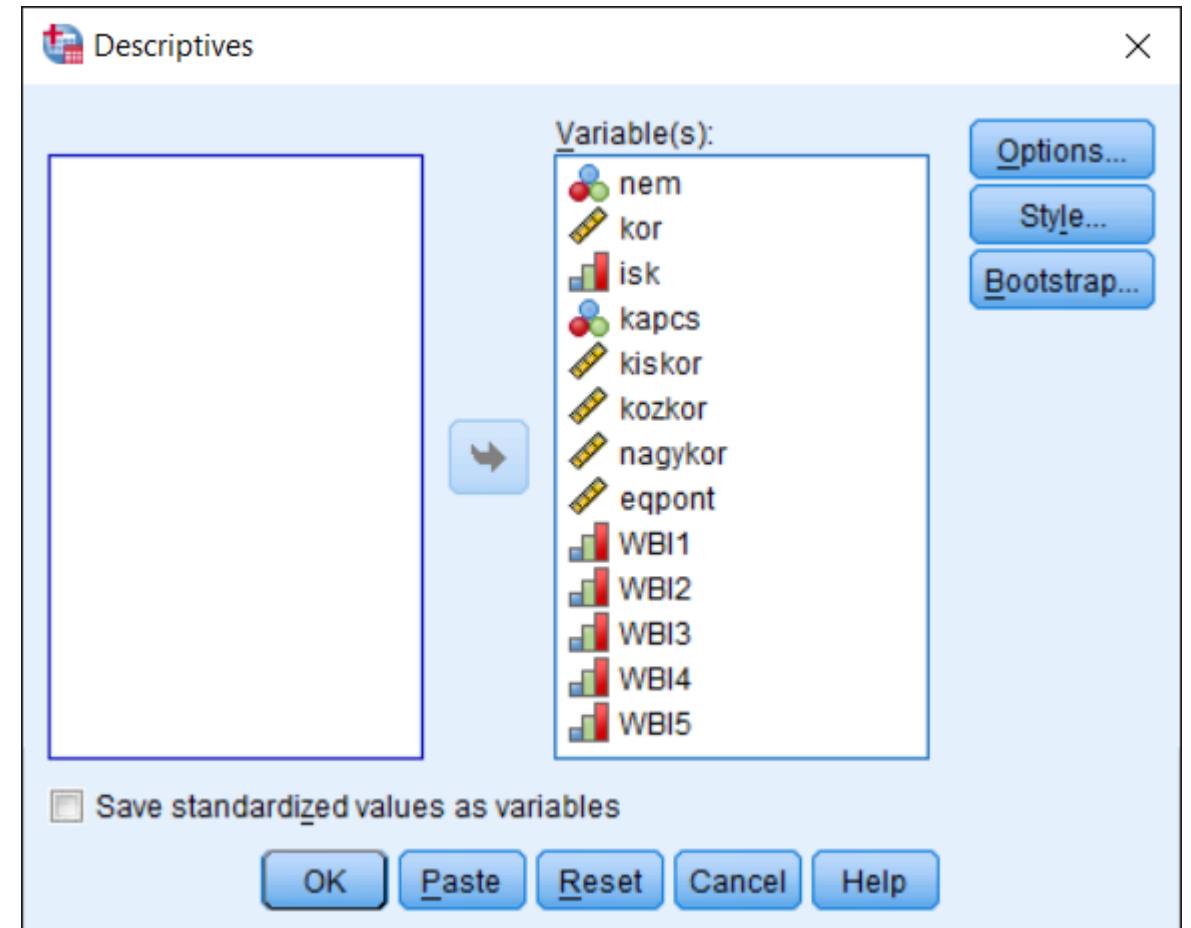


Fig. 1. A statisztika kurzus hallgatóinak életkori eloszlása

- A „hiba” szó több értelemben is előfordul a statisztikában, az adattisztítás során az adathibákkal foglalkozunk
 - **Adathiba**
 - Téves adatok
 - Legtöbbször félregépelésekből, mérési problémákból, esetleg nem őszinte válaszadókból adódnak. (Pl. ha valaki 165 cm magas, de az adatok bevitelekor véletlenül 175-öt ütöttél be, az hiba. De az is hiba, ha meg volt nyúlva a mérőszalag, és mindenkit 2 centivel rövidebbnek mértél.)
 - Törekedni kell arra, hogy adathiba ne legyen az elemzésben.
 - Adathiba leggyakoribb forrásai:
 - Félregépelések adatbevitelkor
 - Félregépelés, tizedesvessző rossz helyen
 - (Számítási) hibák kiértékeléskor
 - Elszámolások „fejben” számolásakor
 - Rosszul megírt kiértékelő, rossz megoldó kulcs
 - Rosszindulatú / komolytalan kitöltők
 - Beleunt, és csupa 1-est nyom, vagy mintázat szerint tölti ki
 - Visszajelzést adsz, ezért megpróbálja egy „alternatív személyisége” szerint is kitölteni

Adatok áttekintése

- Az adattisztítás mindig az adatok áttekintésével kezdődik
 - általános képet kapunk az adatokról
 - „furcsaságokat” keresünk (nem lehetséges értékeket, gyanúsan kicsi/nagy szórásokat stb.)
- Analyze / Descriptive / Descriptive
 - Jeligét kivéve tegyünk be minden változót!
 - Persze, nem minden változónál lehet majd minden statisztikai mutatót értelmezni (például a kategoriális változóknál nem nézzük majd az átlagot, szórást)



- Mit tudunk leolvasni?
 - 1. Vannak hiányzó értékek
 - Onnan látható, hogy az elemszám nem mindenhol 51.
 - A tág kapcsolati körnél számítottunk erre, hiszen akinek nincs FB-ja, annál ez az érték hiányzik.
 - Kornál, Családi állapotnál érdemes az adott személy lapját megkeresni, és ellenőrizni.
 - 2. Vannak hibás értékek
 - Onnan látható, hogy a kor maximuma 666 év, ami halandó ember esetén nem elképzelhető érték.
 - Azonnal látható egy másik hibás érték is, a WBI5 változóban, mely egy 0-tól 3-ig terjedő Likert-skála, a táblázat alapján a maximuma mégis 4.

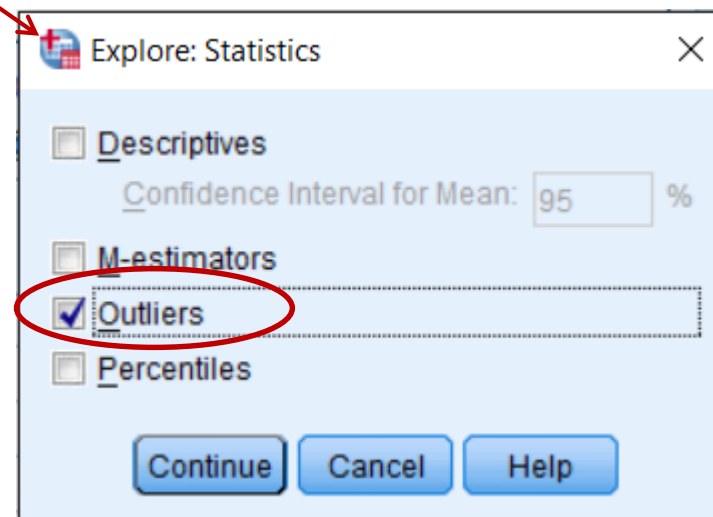
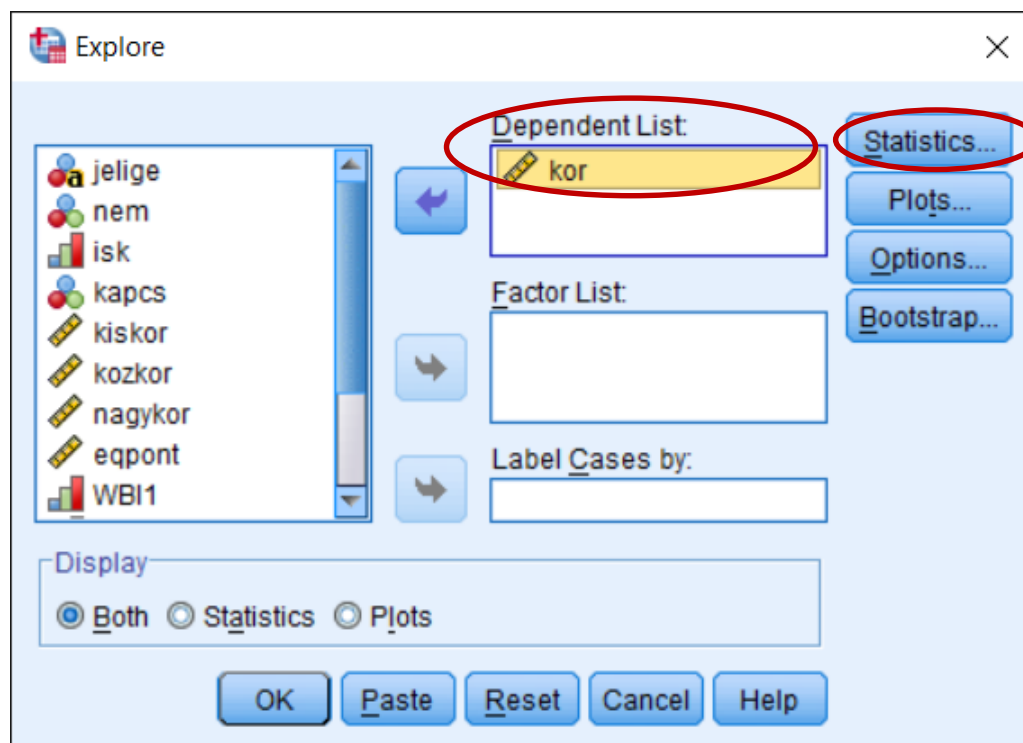
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Nem	51	1	2	1,51	,505
Kor	50	18	666	49,86	90,929
Legmagasabb iskolai végzettség	51	1	3	1,98	,883
Családi állapot	50	1	3	1,96	,832
Szűk kapcsolati kör mérete	51	0	8	3,76	2,286
Közepes kapcsolati kör mérete	51	5	25	16,04	5,306
Tág kapcsolati kör mérete	38	45	1910	463,24	384,974
Érzelmi intelligencia	51	9	36	28,37	7,962
WBI1 - vidámnak és jókedvűnek?	51	0	3	1,86	,530
WBI2 - nyugodtnak és ellazultnak?	51	1	3	2,00	,632
WBI3 - aktívnak és élénknek?	51	0	3	1,90	,608
WBI4 - ébredéskor frissnek és élénknek?	51	1	3	2,04	,564
WBI5 - A napjai tele voltak számára érdekes dolgokkal?	51	0	4	1,55	,901
Valid N (listwise)	37				

Hibás értékek megtalálása

- **Hibák megtalálása:**

- Analyze / Descriptive Statistics / Explore
- A hibás változókat érdemes nem egyszerre, hanem egyenként megvizsgálni, ezért először csak a kor változót tegyük be.
- A Statistics beállításban pipáljuk be az Outliers pontot!
 - Az Outliers elnevezés nagyon félrevezető. A menü **NEM az outliereket** (később megtanuljuk mi is az) listázza, **hanem az öt legmagasabb és öt legalacsonyabb értéket**, mely segít a hibák megtalálásában.



Hibás értékek megtalálása

Extreme Values

		Case Number		Value
Kor	Highest	1	38	666
		2	27	126
		3	31	77
		4	21	65
		5	2	60
	Lowest	1	50	18
		2	22	18
		3	14	18
		4	23	19
		5	40	20 ^a

a. Only a partial list of cases with the value 20 are shown in the table of lower extremes.

27	piramis	2	126
28	csakafradi	2	29
29	madár	2	37
30	esztike	1	29
31	kőbánya	1	77
32	saci	2	25
33	háziárkány	2	57
34	cocacola	2	50
35	szivárvány	2	31
36	zafir	2	37
37	lukács	1	26
38	noname	1	666
39	huligán	2	30

Az Extreme Values táblázatban az öt legmagasabb és öt legalacsonyabb értéket látjuk, és annak a sornak számát, melyben az érték található.

A táblázatból látjuk, hogy két hibás érték is van, a 666-os a 38. sorban, de van egy 126-os is a 27. sorban.

• Mit kezdjünk a hibás értékekkel?

- Ha vissza tudjuk keresni a helyes adatot (pl. papírról), akkor javíthatjuk, minden más esetben **TÖRÖLNI KELL**.
- Nézzünk a 27. sorban lévő „piramis” jeligét! Az adatbázisban hozzá 126 került, az adatlapról viszont látszik, hogy egyszerűen elgépettünk. Az ő **értékét javíthatjuk** az adatbázisban is 26-ra.
- Nézzük a 38. sorban lévő „noname” jeligét! Látható, hogy nem elgépelésről van szó, a kitöltő maga írt 666-ot a korához. Ilyenkor érdemes megnézni, hogy a többi adata validnak tűnik-e, és ha nem, **az egész sort törölni** érdemes.
- Ha nem tudjuk ehhez hasonlóan visszakeresni az adatlapot, vagy nincs okunk feltételezni, hogy a többi adata hamis lenne, akkor **nem kell az egész sort, elég csak a hibás értéket törölni**.

Jelige: PIRAMIS

Szemek teszt pontszáma (a vizsgálat nevezője tölti ki): 36

Kérjük adja meg korát: 26

Nem:

férfi

nő

Legmagasabb iskolai végzettség:

általános iskola

középiskola vagy gimnázium

főiskola, egyetem vagy magasabb

Családi állapot:

egyedülálló

párkapcsolatban élő

házas

özvegy

egyéb:

Facebook ismerősök száma: 756

Kérjük, tekintse át Facebook ismerőseit, és sorolja fel azon személyeket, akikre igaz a következő állítások valamelyike! Elég, ha a nevek Ön számára azonosíthatók. Olyan személyt is megemlíthet, aki nincs Facebook ismerősei között!

1. Merne tőle nagyobb összeget kérni. Elmondaná neki, ha súlyos betegsége lenne. Kérne tanácsot tőle egy súlyos probléma esetén:

ANYU, APU, JULI, PETI

2. Legalább néhány hetente tartják a kapcsolatot, és életét viszonylag jól ismeri, illetve ők ismerik az ön életét:

ANYU, APU, JULI, PETI, ANNA, LACIKA, ANDRIS, BALU, BENCE, PISTA, ISTI, CSENGE, ESZTER, ROBI, ZSOLTI, FERI, GÖLYŐ, K. ANDRIS, KRISTI, F. ZOLI, DANÓ

Kérjük, karikázza be azt a válaszlehetőséget, mely legjobban leírja, hogyan érezte magát az elmúlt 2 hét során!

Az elmúlt két hét során érezte-e magát...	egyáltalán nem jellemző	alig jellemző	jellemző	teljesen jellemző
1. vidámnak és jókedvűnek?	0	1	☒	3
2. nyugodtnak és ellazultnak?	0	1	☒	3
3. aktívnak és élénknek?	0	1	2	☒
4. ébredéskor frissnek és élénknek?	0	1	☒	3
5. A napjai tele voltak számára érdekes dolgokkal?	0	☒	2	3

Jelige: noname

Szemek teszt pontszáma (a vizsgálat nevezője tölti ki): 36

Kérjük adja meg korát: 666

Nem:

férfi

nő

Legmagasabb iskolai végzettség:

általános iskola

középiskola vagy gimnázium

főiskola, egyetem vagy magasabb

Családi állapot:

egyedülálló

párkapcsolatban élő

házas

özvegy

egyéb:

Facebook ismerősök száma: 69

Kérjük, tekintse át Facebook ismerőseit, és sorolja fel azon személyeket, akikre igaz a következő állítások valamelyike! Elég, ha a nevek Ön számára azonosíthatók. Olyan személyt is megemlíthet, aki nincs Facebook ismerősei között!

1. Merne tőle nagyobb összeget kérni. Elmondaná neki, ha súlyos betegsége lenne. Kérne tanácsot tőle egy súlyos probléma esetén:

Andi, Steve, Kata, Kótya, Laci, Feri

2. Legalább néhány hetente tartják a kapcsolatot, és életét viszonylag jól ismeri, illetve ők ismerik az ön életét:

Andi, Steve, Kata, Kótya, Laci, Feri, anyám, apám, nagyapám

Kérjük, karikázza be azt a válaszlehetőséget, mely legjobban leírja, hogyan érezte magát az elmúlt 2 hét során!

Az elmúlt két hét során érezte-e magát...	egyáltalán nem jellemző	alig jellemző	jellemző	teljesen jellemző
1. vidámnak és jókedvűnek?	0	1	2	☒
2. nyugodtnak és ellazultnak?	0	1	2	☒
3. aktívnak és élénknek?	0	1	2	☒
4. ébredéskor frissnek és élénknek?	0	1	2	☒
5. A napjai tele voltak számára érdekes dolgokkal?	0	1	2	☒

- Ugyanezzel a technikával keressük meg a WBI5-ben lévő hibás értéket! A 25. sorban lévő „süssfelnap” adatlapján megnézve, megint csak elgépeltük az értéket, a helyes érték 3, ezért ezt is javíthatjuk.
- Végül, ha úgy érezzük, kezeltünk minden hibát, ellenőrizzük újra a leíró statisztikákat! Figyeld meg, milyen sokat változott a kor átlaga és szórása.
Ezért fontos a hibák kiszűrése!

Extreme Values

			Case Number	Value
WBI5 - A napjai tele voltak számára érdekes dolgokkal?	Highest	1	25	4
		2	5	3
		3	38	3
		4	50	3
		5	51	3
	Lowest	1	47	0
		2	45	0
		3	41	0
		4	36	0
		5	17	0 ^a

a. Only a partial list of cases with the value 0 are shown in the table of lower extremes.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Nem	50	1	2	1,52	,505
Kor	49	18	77	35,24	14,297
Legmagasabb iskolai végzettség	50	1	3	2,00	,881
Családi állapot	49	1	3	1,98	,829
Szűk kapcsolati kör mérete	50	0	8	3,72	2,286
Közepes kapcsolati kör mérete	50	5	25	16,18	5,263
Tág kapcsolati kör mérete	37	45	1910	473,89	384,562
Érzelmi intelligencia	50	9	36	28,22	7,968
WBI1 - vidámnak és jókedvűnek?	50	0	3	1,84	,510
WBI2 - nyugodtnak és ellazultnak?	50	1	3	1,98	,622
WBI3 - aktívnak és élénknek?	50	0	3	1,88	,594
WBI4 - ébredéskor frissnek és élénknek?	50	1	3	2,02	,553
WBI5 - A napjai tele voltak számára érdekes dolgokkal?	50	0	3	1,50	,839
Valid N (listwise)	36				

Gyakoriságok

- Érdemes ilyenkor kikérni a gyakoriság táblákat is a kategoriális változókra (Analyze / Descriptive / Frequencies), hogy leellenőrizzük, van-e olyan kategória, melyben értelmetlenül kevés személy van, nincs-e szükség összevonásokra vagy egy-egy kategória kihagyására.

Láthatjuk, most minden kategoriális változó minden kategóriája megfelelő elemszámú.

- Ha tudjuk, hogy további alcsoportokat (például házas férfiak) fogunk majd vizsgálni, érdemes ellenőrizni az adott tulajdonságok keresztmetszetét is (Analyze / Descriptive / Crosstabs), mert az, hogy a mintában van elég férfi és elég házas, nem jelenti azt is, hogy házas férfiból is van elegendő (lehet, hogy a mintában az összes férfi egyedülálló, és az összes házas nő)

		Nem			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	férfi	24	48,0	48,0	48,0
	nő	26	52,0	52,0	100,0
	Total	50	100,0	100,0	

Legmagasabb iskolai végzettség

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	általános iskola	19	38,0	38,0	38,0
	középiskola, gimnázium	12	24,0	24,0	62,0
	főiskola, egyetem vagy magasabb	19	38,0	38,0	100,0
	Total	50	100,0	100,0	

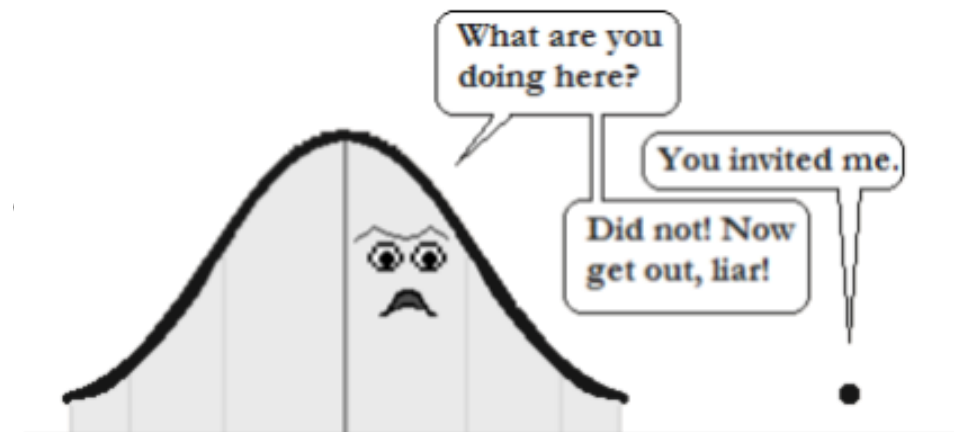
Családi állapot

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	egyedülálló	17	34,0	34,7	34,7
	párkapcsolatban élő	16	32,0	32,7	67,3
	házas	16	32,0	32,7	100,0
	Total	49	98,0	100,0	
Missing	System	1	2,0		
Total		50	100,0		

További transzformációk

- Végül hozzuk létre a WBI öt iteméből a WBI skálapontszámot (Transform / Compute Variable) segítségével.
 - A WBI teszt kiértékelési útmutatójából tudjuk, hogy nincs fordított tétel, és a skálapontszám az itemek összegéből adódik. Mivel tudjuk azt is, hogy nincsenek hiányzó értékek, ezért az összeadás megfelelő eljárás lesz, nincs szükség az átlagolásra.
 - A létrehozott változónak állítsd be a tulajdonságait is!

Outlier



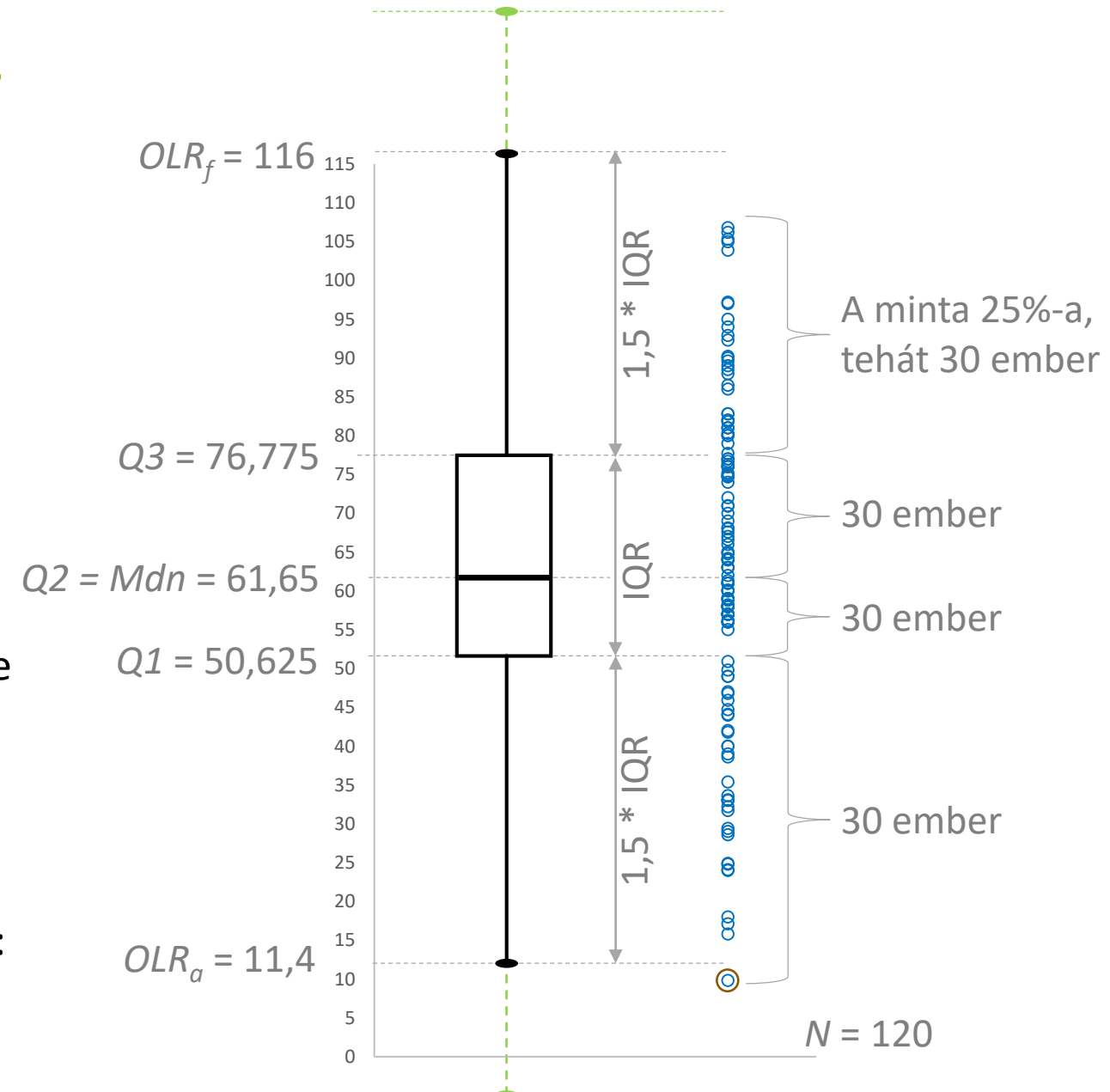
Outlier definíciója

- Általános definíció: outlier az,
 - „ami valid érték, de nagyon eltér a többi adatunktól”
 - Miért probléma:
 - Bár valid érték, de nem jól reprezentálja a populációt – ez főleg kis minták esetén probléma (egy 15 fős mintában van egy 2m magas nő – a valóságban nincs a nők közel 7%-a 2m körül)
 - Torzítja a statisztikai mutatókat: elhúzza az átlagot, megnöveli a szórást.
 - Skála típusú változók (nagyon ritkán ordinális változók) esetén van értelme outlierről beszélni, nominális változóban nem lehet outlier!
 - Hogy találjuk meg őket:
 - **Outlier labelling rule**
 - Olvasni lehet néha az „átlag $\pm$ 2 szórás” szabályról is, de az elavult, ne használjátok!

• Outlier labelling rule

- **Outlier az, ami távolabb az alsó/felső negyedről, mint a középső 50% nagyságának másfélszerese**
- Kvartiliseknek azokat az értékeket nevezzük, melyek a mintát pont 4 egyforma részre bontják.
 - Q1 – alsó kvartilis alatt van a minta egynegyede, felette háromnegyede
 - Q2 – a medián, azaz pont félbeveszi a mintát
 - Q3 – felső kvartilis alatt van a minta háromnegyede, felette egynegyede.
 - IQR – interkvartilis terjedelem, a Q1 és Q3 közötti rész, azaz a minta középső 50%-a
- Azaz outliernek tekintjük azt, aki magasabb illetve alacsonyabb értéket vesz fel, mint a következő határok.
 - **Felső határ: $Q3 + 1,5 * IQR$**
 - **Alsó határ: $Q1 - 1,5 * IQR$**
- Később módosították a határok számítási módját:
 - $Q1 - 2,2 * IQR$ és $Q3 + 2,2 * IQR$

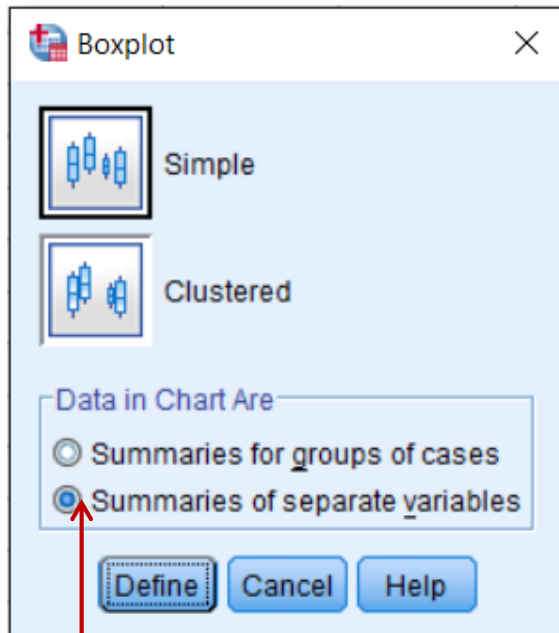
Outlier labelling rule



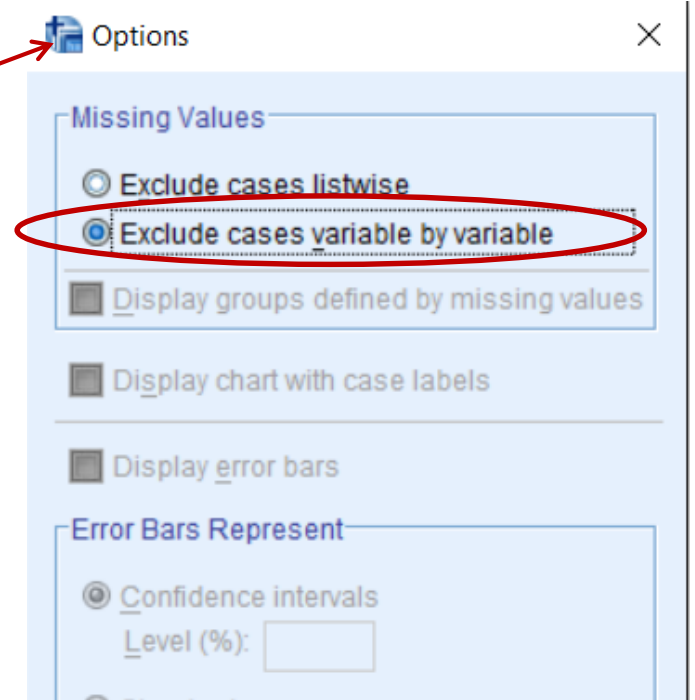
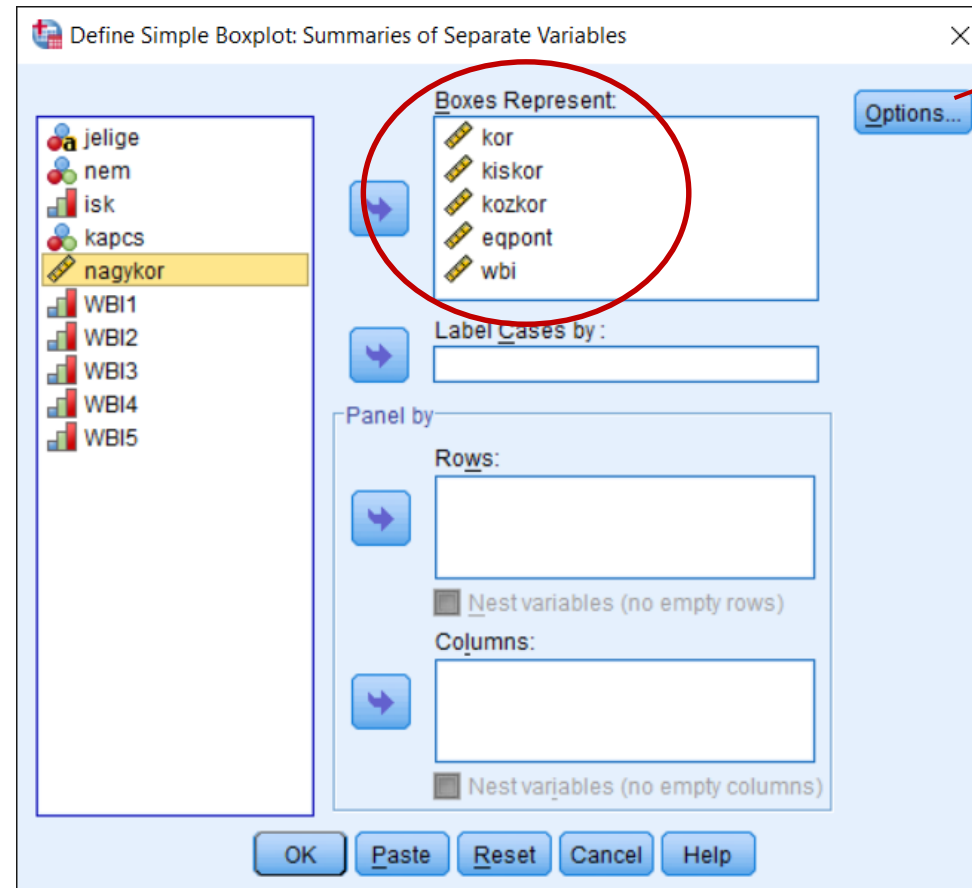
Outlierek megtalálása boxplottal

- **Outlierek megtalálása:**

- Graphs / Legacy Dialogs / Boxplot
- Outlier csak skála típusú változóban (vagy nagyon sok szinttel rendelkező ordinális változóban lehet)
- Mivel a nagykor sokkal nagyobb értékeket tartalmaz, mint a többi változó, nem lenne jól látható a grafikon, ezért azt érdemes külön vizsgálni



Groups of cases, ha több csoportot akarsz megjeleníteni
Separate variables, ha több változót akarsz megjeleníteni

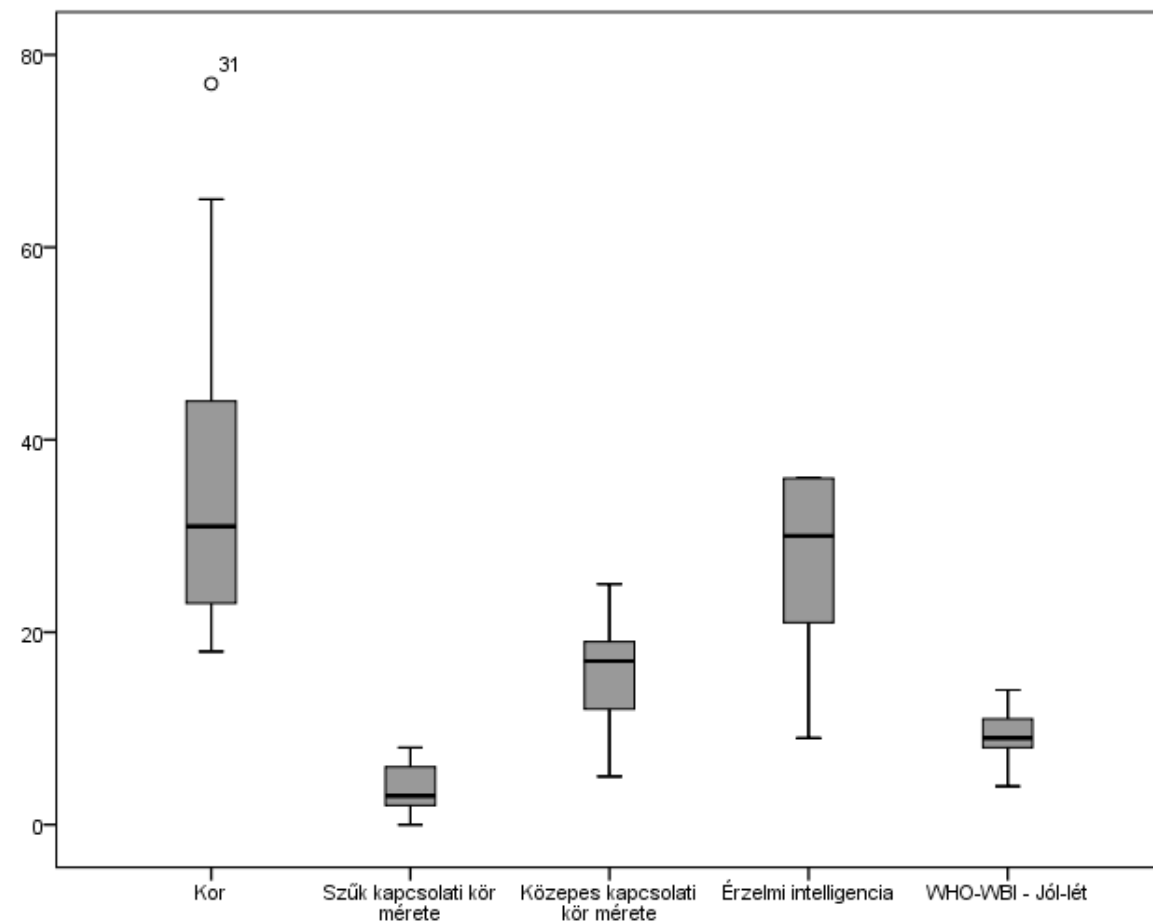


Olvasd el a hiányzó értékekről szóló részt is!

- Először ellenőrizd, hogy jól állítottad-e be a hiányzó értékeket, minden változónál látod-e a változóhoz tartozó teljes elemszámot!
- Majd nézd meg a Boxplot grafikont!
 - A téglalap tartalmazza a IQR-t, azaz minta középső 50%-át (pl. a WBI-nél, melyben 50 fő van, a középső 25 embert).
 - Az úgynevezett whiskers végei jelölik az outlier labelling rule alapján számolt alsó és felső határt
 - A **köröcskék** jelölik az outliereket, tehát azokat az értékeket, akik $1,5 \cdot \text{IQR}$ -nél távolabb vannak az alsó/felső határtól.
 - A köröcskék mellett a szám az adatbázisban a sor számát jelöli
 - Néha láthatsz **csillagokat** is, azok jelentése, hogy az egyed $3 \cdot \text{IQR}$ -nél távolabb van az alsó/felső határtól. Ezek gyakran hibák (de nem feltétlenül), illetve annyira szélsőséges értékek, amikre érdemes kiemelten figyelni

Case Processing Summary

	Valid		Cases Missing		Total	
	N	Percent	N	Percent	N	Percent
Kor	49	98,0%	1	2,0%	50	100,0%
Szűk kapcsolati kör mérete	50	100,0%	0	0,0%	50	100,0%
Közepes kapcsolati kör mérete	50	100,0%	0	0,0%	50	100,0%
Ézelmi intelligencia	50	100,0%	0	0,0%	50	100,0%
WHO-WBI - Jól-lét	50	100,0%	0	0,0%	50	100,0%





Módosított outlier labelling rule

- Az outlier labelling rule-nak van egy módosított verziója, ahol a szorzó 1,5 helyett 2,2
 - A gyakorlati tapasztalatok alapján a 2,2-es sokkal jobban működik, de az SPSS-ben nem lehet ezt kérni, ezért kézzel kell számolni a határt.
- Analyze / Descriptive / Frequencies / menüben Statistics almenü, és a Quartiles-t kattintjuk be.
 - (a főoldalon a Display Frequency Tables mellől kiszedhető a pipa, mert hosszúak lennének a táblázatok és nincs is rájuk szükség)
- A negyedelő pontok ismeretében számológéppel kiszámoljuk a módosított alsó és felső határt
 - Az alsó határ: $Q1 - (Q3 - Q1) * 2,2 = 23 - (44,5 - 23) * 2,2 = - 24,3$
 - A felső határ: $Q3 + (Q3 - Q1) * 2,2 = 44,5 + (44,5 - 23) * 2,2 = 91,8$

Statistics

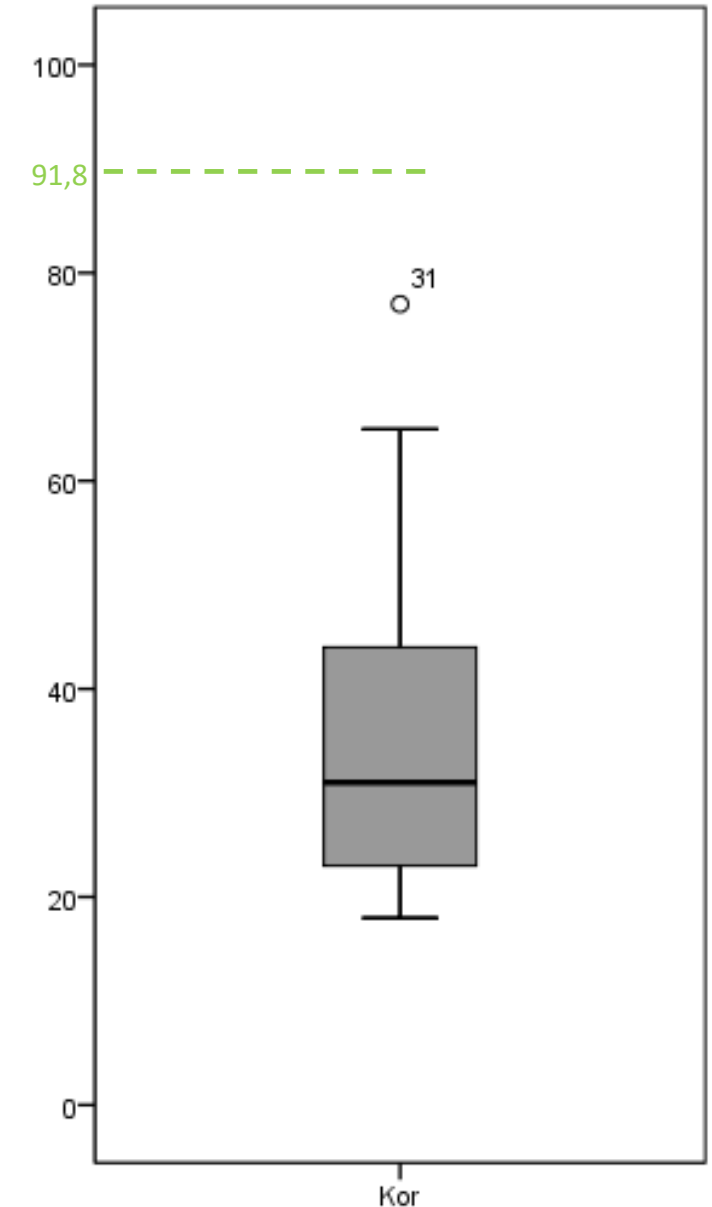
Kor

N	Valid	49
	Missing	1
Percentiles	25	23,00
	50	31,00
	75	44,50

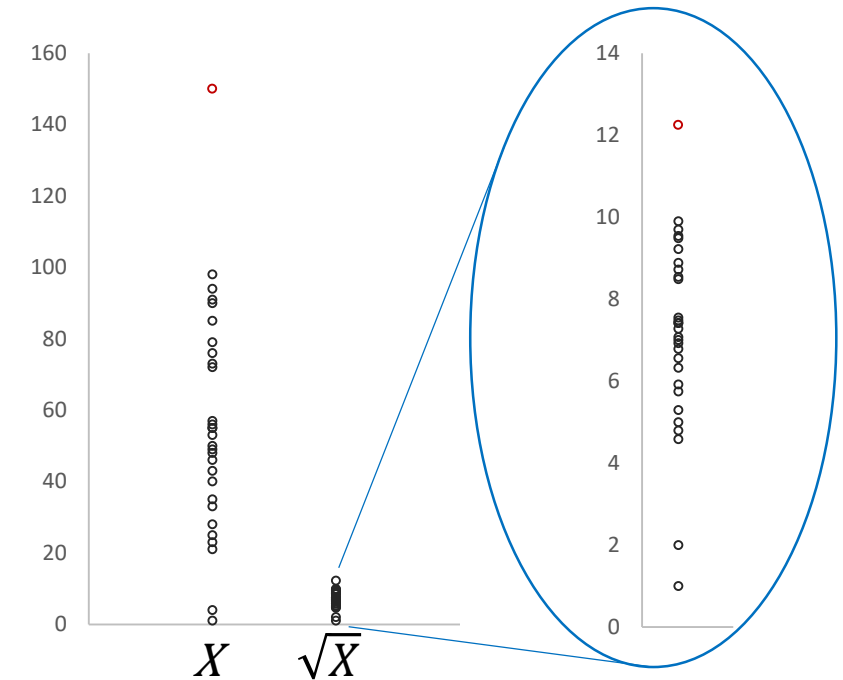
Q1 az alsó negyedelő
Q3 pedig a felső negyedelő

Módosított outlier labelling rule

- A mi esetünkben a felső határ az érdekes, mert azon az oldalon van outlier.
- A felső határ: $Q3 + (Q3 - Q1) * 2,2 = 44,5 + (44,5 - 23) * 2,2 = 91,8$
- Látható, hogy a módosított outlier labelling rule alapján a 31. sorban lévő személy kor adatát nem kell outliernek tekintenünk, és a mintában hagyhatjuk.
- Megjegyzés boxplotokhoz:
 - Ne lepődjünk meg, ha az SPSS nem mindig olyan hosszúra húzza a whiskers szárát, mint ahova azt mi számoltuk, az áttekinthetőbb ábrázolás érdekében néha módosít a szárak hosszán. Például a mintában a legfiatalabb személy 18 éves, tehát hiába van még a másfeles szorzóval is alacsonyabban az elméleti határ, gyakorlatban nincs értelme a szárat 18 alá húzni.



- Mit kezdünk velük?
 - **-1. Hiba vagy outlier?**
 - Ha hiba, akkor nincs helye a mérlegelésnek, mennie kell
 - **0. Elemezzük a helyzetet**
 - Pl. nem normális eloszlásnál mit jelentenek?
 - 2,2-es szorzóval is outliereknek tekinthetők?
 - **A. Töröljük őket**
 - Ésszel!!
 - PRO: torzító hatás megszűnik KON: információt veszünk
 - Leggyakrabban használt, kis minta esetén bajos
 - Gondoljuk át, miből jöhet egy szélsőséges érték
 - **B. Transzformáljuk az adatokat**
 - PRO: nincs információvesztés KON: nehéz értelmezni
 - Lineáris transzformációkkal (pl. gyökvonással) az outlierek hatásának csökkentése— elméletileg legjobb megoldás, gyakorlatilag minél jobban transzformáljuk az adatokat, annál nehezebb pszichológiailag releváns következtetéseket levonni majd
 - **C. Winzorizálás - Átírjuk az értéket valami olyanra, ami már nem outlier**
 - PRO: torzító hatás megszűnik kevés információvesztéssel KON: sokan (feleslegesen) idegenkednek tőle
 - Csalás lenne? De ha benne hagyjuk, az is torzítja az adatokat, a két rossz közül ez a kisebb
 - Mire? A következő legnagyobb -már nem outlier- érték plusz 1 egység (a következő legkisebb -1 egység)



- A tág kapcsolati körre a boxplotot kikérve látható, hogy itt is van egy outlier.
- 1. ellenőrizzük, hogy ez inkább hiba vagy outlier. Az 47. sorban lévő vegas jeligéjű személy facebook ismerőseinek száma 1910, ami elképzelhető érték, hiszen a Facebook ismerős-limitje 5000 főnél van
- 0. a módosított outlier labelling rule-lal ellenőrizzük, mennyire tekinthető szélsőségesnek. A módosított felső határ $(709,5 - 175,5) * 2,2 + 709,5 = 1884,3$, tehát igen, a módosított szabály alapján is outlier.
- A. Törölhetjük az értéket. Ekkor elegendő csak ezt az egy adatot törölni, nem kell az egész személyt, hisz a többi adata rendben van
- B. Használhatunk transzformációt – a gyökvonás például jól működni
- C. Winzorizálhatunk. Az öt legmagasabb értéket kilistázva látjuk, hogy a következő legmagasabb érték 1217, tehát vegas barátainak számát átírjuk 1218-ra.

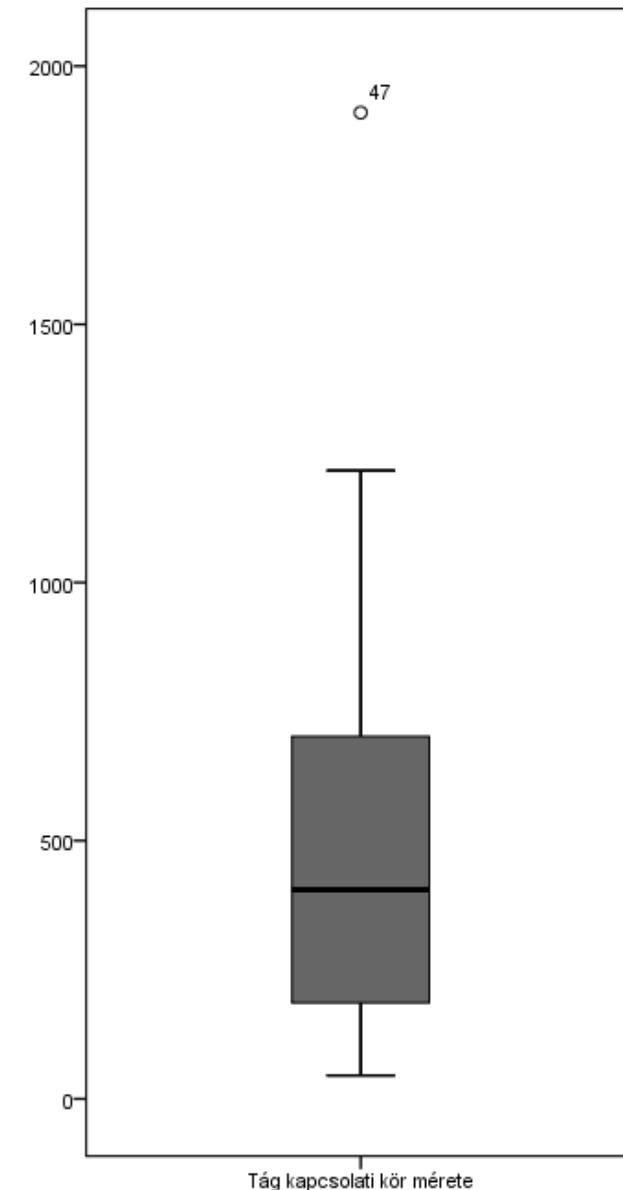
Statistics

Tág kapcsolati kör mérete

N	Valid	37
	Missing	13
Percentiles	25	175,50
	50	405,00
	75	709,50

Extreme Values

		Case Number		Value
Tág kapcsolati kör mérete	Highest	1	47	1910
		2	22	1217
		3	49	1113
		4	3	829
		5	48	758
	Lowest	1	2	45
		2	45	56
		3	1	75
		4	42	92
		5	16	102



Hiba vagy outlier?

- **Hiba**

- Mi hiba?
 - Nincs információ értéke, például valamilyen elgépelésből származik. A megtalálható hibák gyakran az adott skálán nem képzelhetők el. Például nemi adatoknál 3 vagy 0-5-ig terjedő skálán 7 vagy magasságban adatokban 257 (oké, kivéve, ha a Guinness rekordok könyvéből veszed az adatokat 😊)
- Honnan tudom, hogy vannak-e hibás értékek, és ha igen, hol?
 - A leíró statisztikában minimum és maximum nem megfelelő
 - A Extreme Value-kat érdemes megnézni
 - A boxploton ÁLTALÁBAN a csillagos értékek valamilyen elírásból származnak (de a csillag tényleges jelentése más, lásd a következő dia)
- Mit kell velük kezdeni?
 - Hibás értékeknél nincs kérdés, ha tudod, javítod őket, ha nem tudod javítani, törlöd

- **Outlier**

- Mi outlier?
 - Olyan érték, mely elképzelhető az adott skálán, de annyira szélsőséges, hogy torzíthatja a statisztikákat
- Honnan tudom, hogy vannak-e outlierok, és ha igen, hol?
 - A bloxploton ÁLTALÁBAN a körökkel jelzett értékek outlierok
- Mit kell velük kezdeni?
 - Eldöntheted, lásd az előző diát!

Boxplot csillagok és körök

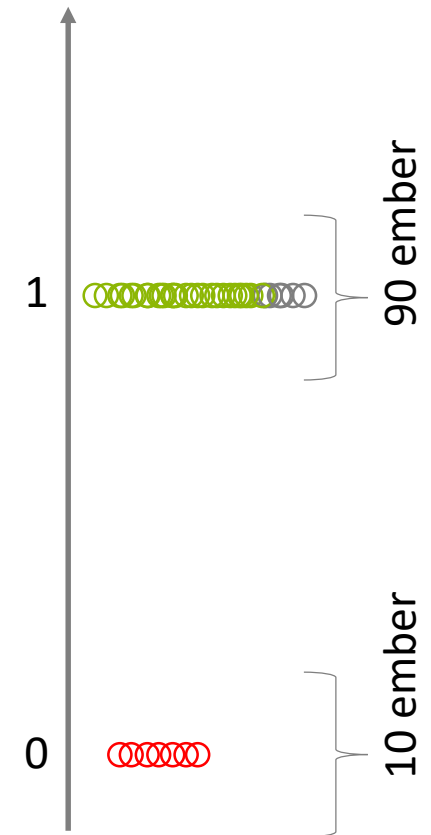
- **Körök:** ami távolabb az alsó/felső negyedről, mint a középső 50% nagyságának **másfélszerese**
- **Csillagok:** ami távolabb az alsó/felső negyedről, mint a középső 50% nagyságának **háromszorosa**
- Nagyon gyakran a csillagok valamilyen elírásból származnak, míg a körök outlierok a mintában, de ez nem feltétlenül igaz:
 - Példa1: nemi adatoknál a 0 (ami egyértelműen elírás) nem jelenik meg csillagként, de még körként sem.
 - Példa2: 5 fokú skálán a 6-osték (ami elírás), lehet, hogy nem jelenik meg csillagként, csak körként
- Megjelenhetnek a skálán elképzelhető értékek is csillagként (például sikerült beválogatnod egy NBA játékost). Ezek bár elképzelhető értékek, de annyira szélsőségesek, hogy mindenképp kezdeni kell velük valamit.



Miért nincs outlier kategoriális változóban?

Miért nincs outlier kategoriális változóban?

- Mert kategoriális változó értékei nem sorba rendezhetők
 - Ha 1-gyel jelölöm a szőkét, 2-vel a vörösöket, 3-mal a barnákat, annak nincs olyan jelentése, hogy a barnák többek lennének, mint a szőkék.
 - Ezért a negyedelő pontok sem értelmezhetők (nincs olyan, hogy a minta alsó vagy felső negyede, mert a mintának nincs alsó és felső része)
- Miért jelölhet az SPSS mégis outliert, ha kategoriális változóra kérünk ki boxplotot.
 - Az SPSS azonban nem „tudja”, hogy milyen butaságra készülsz, és ha kéred, készít neked boxplotot kategoriális adatokhoz. Ahol a kategóriákat jelölő értékeket úgy veszi, mintha azok sorba rendezhetők lennének.
- Otthon gondold át, mi történik a következő esetben:
 - Van egy dichotóm változó, például az, hogy átmegy-e a hallható a statisztika vizsgán. 0-val jelölve a nem, 1-gyel jelölve az igen. Az évfolyam (a minta) 100 fős, ebből 10 bukik, 90 átmegy.
 - Ha kikéred a boxplotot, akkor a 10 bukók mind outlierként lesznek jelölve.
 - Miért? Gondold végig az outlier labelling rule számítását! Hol lesznek ebben a tévesen értelmezett helyzetben a „kvartilisek”



A adatokban a hagyományos outlier labelling rule (Tukey, 1977) alapján két szélsőséges értéket is találtunk, a kor adatokban egy 77 éves személyt és a tág kapcsolati körben egy személyt, akinek 1910 Facebook ismerőse van.

Az így megtalált outlierok közül a felülvizsgált, 2.2-es szorzót használó outlier labelling szabály (Hoaglin és Iglewicz, 1987) alapján csak a tág kapcsolati kör érték minősül outlieroknak.

A megtalált szélsőséges értéket a winzorizálás szabályainak megfelelően átírtuk a rangsorban következő legnagyobb már nem szélsőséges értéknél egy egységgel nagyobbra, azaz 1218-ra.

- Hivatkozások:
- Tukey, J. W. (1977). Exploratory data analysis.
- Hoaglin, D. C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. Journal of the American Statistical Association, 82(400), 1147-1149.
- Hoaglin, D.C., Iglewicz, B., and Tukey, J.W. (1986). Performance of some resistant rules for outlier labeling, Journal of American Statistical Association, 81, 991-999.