

Poligámok, nyugdíjmilliomosok és óvodás vállalkozók — avagy az adattisztításról

Máté Fanni
tanársegéd

ELTE Társadalomtudományi Kar

Mi az adattisztítás?

- Adatbázisok hibáinak kijavítása
- Adatok ellenőrzése
- Adatok elemezhető formára hozása
- Az elemezhető adatok pontosítása
- „Menteni a menthetőt”
- NEM az adatok kozmetikázása

Milyen az adattisztítás?

- Időigényes folyamat
- Folyamatos
- Elkerülhetetlen
- Fárasztó
- Érdekes

Miért fontos?

- Elemzés, döntések alapját képezik az adatok
- Hiba mindig van
- Téves adatok az elemzés közben
- Nem biztos, hogy feltűnik a hiba
- Kérdezőbiztosok ellenőrzése
- Kérdőív fejlesztése

Hol fordul elő?

Papíros kérdezés (PAPI)

- A kérdezőbiztos rosszul írja le
- A rögzítő viszi be tévesen — kérdezőbiztos csúnyán ír; elgépelés
- Nem tartja be a kérdező az ugratásokat
- Válaszadó téveszti el

Hol fordul elő?

Tablettel, számítógéppel történő kérdezés (CAPI)

- Elgépelés
- Félrekattintás
- Programozási hiba
- Válaszadó téveszti el

Hol fordul elő?

Rendelkezésre álló adatok

- Karakterkódolás
- Rossz minőségű adatok, hiányos, nem jól rögzített
- Programhiba
- Adatbázisok összekapcsolása

Mitől kell megtisztítani az adatokat?

- Adatbázisok
 - Eltérő formátum (pl. dátumok, tizedesjegy)
 - Változónevek
- Nyitott kérdésnél, szöveges válasznál
 - Karakterkódolás
 - Eltérő alakok
- Logikai tisztítás
 - Válaszok nem konzisztensek
 - Háztartástáblánál gyakori
- Egyéb válaszlehetőség

m1
M1
Magyar 1
Magyar 1 tv
magyar1
magyar1tv
MTV 1
mtv1
Mtv1
MTV1

Ki kell hozzá?

- Precizitás
- Programozási hajlandóság, programismeret
- Adat előállításának ismerete
- Társadalom, kérdés, téma ismerete
- Kérdőív vagy program ismerete

Hogyan tisztítunk adatot?

Hiba megkeresése

- Eloszlások vizsgálata — kiugró értékek, nem a kérdésre adott válaszok
- Szöveges válaszok
 - Karakterkódolás
 - Egységes alakra hozás
 - Elgépelések
 - Értelmezés
- Logikai tisztítás
 - Fontos kérdéseknél keresztvalidálás
 - Előre megadott kérdések, szabályok

Hogyan tisztítunk adatot?

Hiba kijavítása

- Új adatbázisba
- Jól dokumentálva
- Minél több információ megőrzése, pontosítása — de nem minden áron
- Parancsok írása
- Visszakeresés — adatbázisban vagy kérdőívben
- Következtetés
- Kérdezés

SPSS Statistics-ben hogyan?

- **Menüből**

- Kiugró értékek (Data – Identify unusual cases)
- Validálási szabályok (Data – Validate)
- Duplikátumok (Data – Identify duplicate cases)

- **SYNTAX Editor**

- Visszakövethető
- Jól kommentált
- Gyors
- Segédprogramok

SPSS Statistics-ben mivel? — jó tudni

- TEMPORARY — csak ideiglenesen hajtja végre
- IF — feltételek megadása
- SELECT IF — esetek kiválasztása
- LIST — kilistázza az eseteket
- \$CASENUM — esetszám - változik
- COMPUTE — változó létrehozása, módosítása
- LOWER() — kisbetűssé tesz
- SORT CASES BY — adatbázis rendezése

13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

*** Háztartásjövedelem hol kisebb, mint a személyi? .

TEMPORARY.

SELECT IF haztjov < szemjov.

LIST id haztjov szemjov htnagysag.

*** Tájékoztató csatorna kisbetűssé tétele.

COMPUTE csatorna=lower(csatorna).

EXECUTE.

FREQUENCIES csatorna.

Hivatkozások

- <https://blog.signifai.io/data-cleansing-age-big-data/>
- <https://towardsdatascience.com/data-cleaning-101-948d22a92e4>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1198040/>