Információkeresés

Labor 5-6.

A valós adathalmazok gyakran zajosak, hiányosak, avagy éppen redundáns információt vagy duplikátum egyedeket tartalmaznak. Ezért a tudásfeltárás folyamatában az adattisztítás és adatintegrálással kezdődik.

Az adattisztítás szerepe javítani az adatok minőségén azáltal, hogy kiszűri és eltávolítja az adatokban fellépő hibákat és inkonzisztenciákat.

Az adattisztítás során:

- felmérjük a hibákat
  - ellenőrizzük az adatfájl szerkezeti épségét
  - a zajt, felesleges információt tartalmazó mezőket javítjuk
  - felmérjük a hiányzó értékeket és amennyiben lehet ezeket pótoljuk
  - felmérjük az adatközlési és adatbeviteli hibákat
    - megvizsgáljuk az egyes változók eloszlását
      - az eloszlások szélein elhelyezkedő extrém értékeket ellenőrizzük
      - felmérjük, hogy az eloszlások megfelelnek-e az előzetes elvárásainknak, vannak-e nem várt sűrűsödések, ritkulások egyes értéktartományokban (például durva kerekítés vagy eltérő mértékegység használata az adatszolgáltatók egy részénél)
    - megvizsgáljuk, hogy a változók közötti triviális összefüggések teljesülnek-e
- a hibásnak tűnő adatokat felülvizsgáljuk, javítjuk.

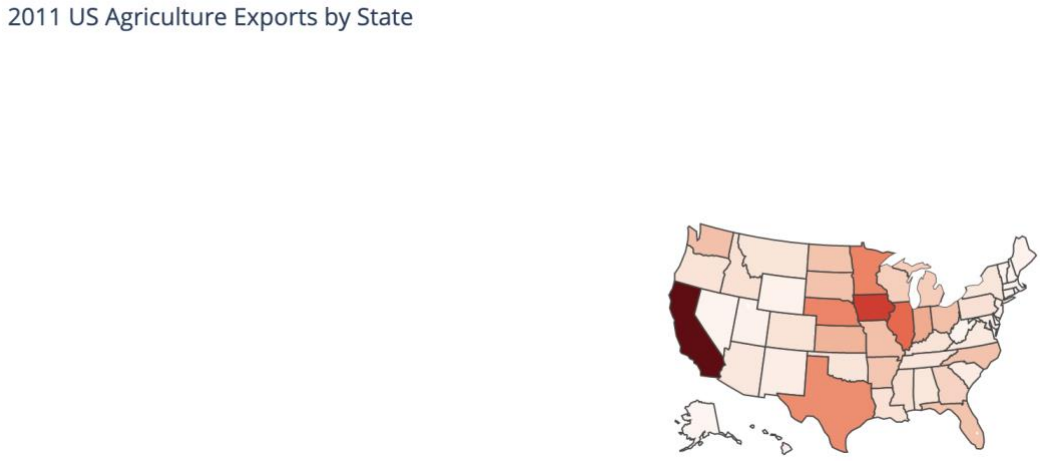
Feladatok

1. Az „`egyetemek.txt`” fájlból szűrjük ki az államokat és azon belül a városokat, melyben egyetemek találhatóak. Ha vannak duplikátumok, helytelen adatok (pl. számokat tartalmazó államnév), ezeket javítsuk. Vizsgáljuk meg az egyetemek eloszlását államok szerint. Melyik államban van a legtöbb, legkevesebb egyetem?
2. Bővítsük ki az adatbázisunkat egy oszloppal, mely tartalmazza az államok rövidítését is (pl. Texas - TX, California - CA stb.). Forrás: https://en.wikipedia.org/wiki/List_of_U.S._state_abbreviations
3. Az adatbázist integráljuk a [https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_population] linken szereplő népszámlálási adatokkal és számoljuk ki államonként hány főre jut egy egyetemi város.
4. Hasonlóan, a [https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_area] linken szereplő területi adatok integrálásával, számoljuk ki államonként átlagban hány négyzetkilométerre jut egy egyetemi város.

5. Az alábbi példa, a kivitelt, exportmennyiséget ábrázolja térképen, államonként. Készítsünk hasonló ábrákat az egyetemek abszolút, lakosság és terület szerinti eloszlásáról is. Az ábrákat exportáljuk kép formájában.

```
import plotly.graph_objects as go
import pandas as pd
df = pd.read_csv('https://raw.githubusercontent.com/plotly/datasets/master/2011_us_ag_exports.csv')

fig = go.Figure(data=go.Choropleth(
    locations=df['code'], # Spatial coordinates
    z = df['total exports'].astype(float), # Data to be color-coded
    locationmode = 'USA-states', # set of locations match entries in `locations`
    colorscale = 'Reds',
    colorbar_title = "Millions USD",
))
fig.update_layout(
    title_text = '2011 US Agriculture Exports by State',
    geo_scope='usa', # limite map scope to USA
)
fig.show()
```



6. Számoljuk ki, hány egyetem van városonként, majd összesítve államonként. Ehhez az „egyetemek.txt” fájlban miután megkapunk egy várost, az utána következő kerek zárójelek között megszámoljuk, hány vesszővel elválasztott karakterlánc található. Pl.:

   Claremont (Claremont McKenna College, Pomona College, Harvey Mudd College, Scripps College, Pitzer College, Keck Graduate Institute, Claremont Graduate University)[5] → 6 vessző van, tehát 7 egyetem.

   Melyik államban és melyik városban van a legtöbb egyetem?

7. Ha marad időnk, folytassuk a Datacamp modulokat.