Mérési adatok vizuális elemzése

Házi feladat specifikáció

YouTube trending videók statisztikáinak elemzése

Készítette: Gao Tong (I2SVOS) és Sike Ádám (E8Z277)

Verzió: 1.0

2022

Budapesti Műszaki és Gazdaságtudományi Egyetem

Méréstechnika és Információs Rendszerek Tanszék

# A specifikáció célja

A specifikáció célja a tárgy házi feladatának megalapozása, a választott adatkészlet kiértékelése és kérdések megfogalmazása. A házi feladat feltételeit a vonatkozó előadás tartalmazza.

Javasolt, de nem kötelező jelen a mintának a használata. Egy megfelelő szöveges magyarázattal ellátott Jupyter Notebook, a belőle generált HTML állománnyal, vagy tetszőleges markdown állomány, esetleg kezdeti dashboard is elfogadható, ha a kötelező elemeket tartalmazza.

Metaadatok:Feladat címe (pl. *Fluxuskondenzátor hatékonysági adatainak vizsgálata*), hallgató neve, Neptun-kódja (ebben a sablonban a címlapon).

# Kiválasztott adatkészlet

Az általunk választott adatkészlet (Kaggle-en): <https://www.kaggle.com/datasets/datasnaek/youtube-new>

Ez az adatkészlet [CC0: Public Domain](https://creativecommons.org/publicdomain/zero/1.0/) licensz alatt érhető el.

Mérete: Az adathalmaz egy 205MB-os tömörített állományban érhető el, összesen 20 fájlban. Tíz ország adatait tartalmazza. Egy ország adathalmaza 16 oszloppal rendelkezik (.csv-ben), illetve tartozik hozzá egy kategóriákat leíró fájl (JSON-ben).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ország** | Kanada | Németország | Franciaország | Egyesült Királyság | India |
| **Sorok** | 40881 | 40840 | 40724 | 38916 | 37352 |
| **Ország** | Japán | Dél-Korea | Mexikó | Oroszország | Egyesült Államok |
| **Sorok** | 20511 | 34476 | 40692 | 40449 | 40949 |

# Adatkészlet alapvető statisztikai kiértékelése

Változók:

Az adatkészlet lényegesebb változóinak azonosítása, és (változóként értelmezett, leíró jellegű) statisztikai kiértékelése:

* Változó típusa (felsorolt: értelmezhető-e rendezés, numerikus)
* Minimum, maximum,átlag, Q1, Q2, Q3, (tapasztalati) szórás
* Felsorolt változó esetében a különböző értékek száma (cél: kiderüljön, ha egy változó lényegében konstans)
* Üres (NA) értékek aránya.

Ehhez javasolt pl. describe() a Python nyelvben, summary R nyelvben, ill. a pandas.profiling.

# Kérdések/célok

Az adatkészlettel kapcsolatos alapvető kérdések, melyeket elsősorban feltárú vizuális módszerekkel meg szeretne válaszolni.

Pl. igaz-e, hogy a napkitörések idején a fluxuskondenzátor hatékonysága jobb? A felhasználók csoportosíthatóak-e a rendszerbe küldött kérések darabszáma/mérete szerint? Stb.

Ha az adatkészlethez tartozik meglévő publikus elemzés (pl. Kaggle felületén elérhető kernel), akkor egyértelműsíteni kell, miben szeretne mást csinálni.

# Vizualizáció jellege, technológiája

Milyen jellegű vizualizációt szeretne készíteni? **Dashboard**

Milyen technológiát választ ehhez? **dash+plotly**

Ettől, mind gyakorlatilag a specifikáció összes többi részétől is, egyeztetés mellett természetesen el lehet a későbbiekben térni.