Mérési adatok vizuális elemzése

Házi feladat specifikáció

YouTube trending videók statisztikáinak elemzése

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

Verzió: 1.2

2022

Budapesti Műszaki és Gazdaságtudományi Egyetem

Méréstechnika és Információs Rendszerek Tanszék

# 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 |

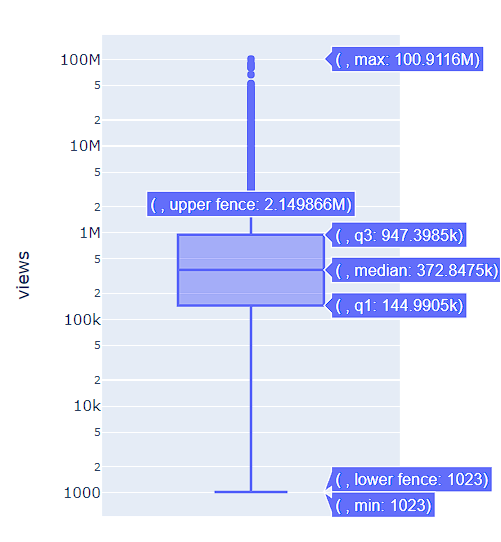
Ez az adathalmaz 2017-ből származik és 205 nap adata találtató meg benne.

# Adatkészlet alapvető statisztikai kiértékelése

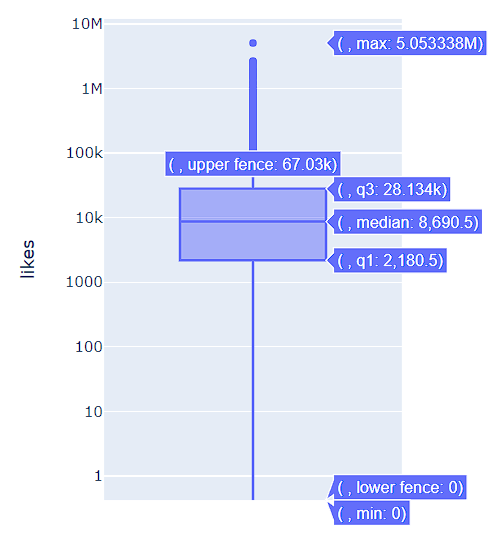
Az adathalmaz a már említett 10 részer van bontva, mindegyikben ugyanazokkal a sorokkal (példaként a kanadai adatokat hozom):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Oszlop neve** | **Típusa** | **Különböző (nem null) értékek** | **Null értékek aránya:** | **Értelmezés** |
| **video\_id** | szöveges | 24427 | *0%* | *A videó egyedi alfanumerikus azonosító* |
| **trending\_date** | **dátum** | **205** | *0%* | *Milyen dátumokon volt a videó a ’’trending’’ fülön* |
| **title** | **szöveges** | **24573** | *0%* | *A videó címe* |
| **channel\_title** | **szöveges** | **5076** | *0%* | *A csatorna neve* |
| **category\_id** | **kategóriaváltozó (nem sorrendezett)** | **17** | *0%* | *A videó kategóriájának azonosítója (kategória külön fájlban* |
| **publish\_time** | **dátum és pontos idő** | **23613** | *0%* | *A videó feltöltésének ideje* |
| **tags** | **szöveges tömb** | **20157** | *0%* | *A videóra illesztett címke, tipikusan a tematikához konkrétabban kötődik* |
| **views** | **egész szám** | **40170** | *0%* | *Nézettségek száma, amikor ’’trending’’ volt a videó* |
| **likes** | **egész szám** | **24676** | *0%* | *Kedvelések száma, amikor ’’trending’’ volt a videó* |
| **dislikes** | **egész szám** | **6241** | *0%* | *’’Dislike’’-ok száma, amikor ’’trending’’ volt a videó* |
| **comment\_count** | **egész szám** | **11172** | *0%* | *Kommentek száma, amikor trending volt* |
| thumbnail\_link | URL | 24422 | *0%* | *Link a videó előnézetére* |
| comments\_disabled | Boolean változó | 2 | *0%* | *Ki voltak-e kapcsolva a kommentek* |
| ratings\_disabled | Boolean változó | 2 | *0%* | *Ki voltak-e kapcsolva az értékelések* |
| video\_error\_or\_removed | Boolean változó | 2 | *0%* | *Törölve lett-e, vagy egyéb hiba* |
| description | szöveges | 22345 | 3,27% | A videó leírása |

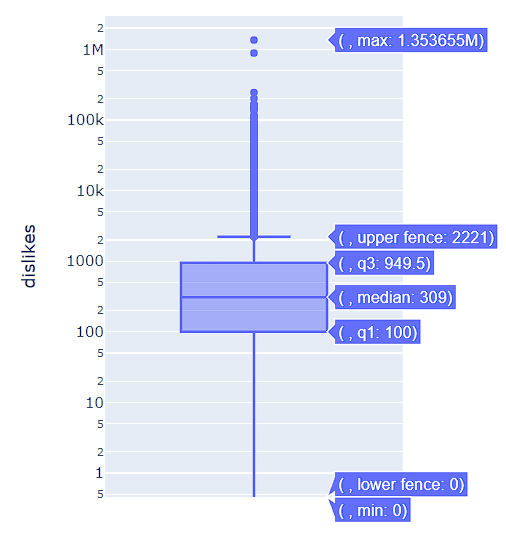
**Megtekintések**:

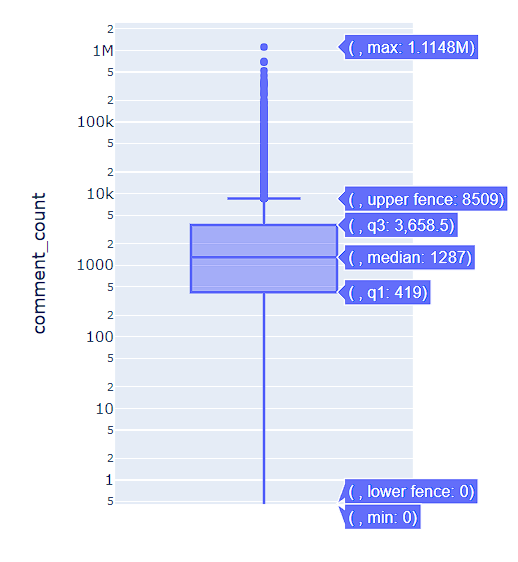


**Kedvelések:**



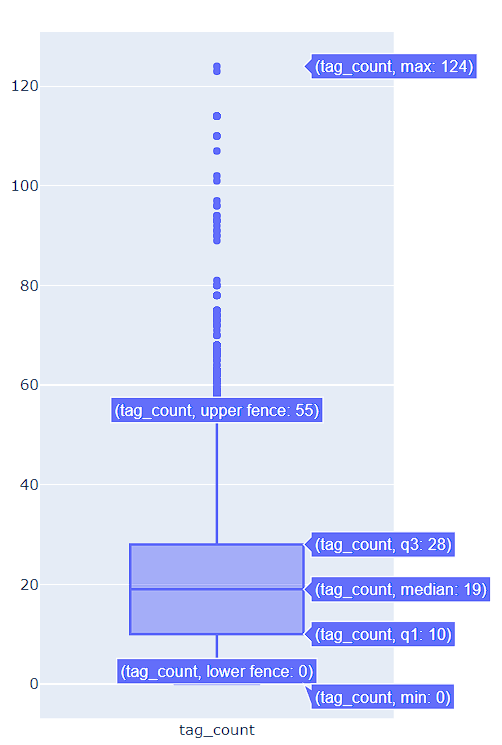
**Dislike-ok:**

****

**Kommentek száma:** ****

Láthatóan a numerikus változók értékei eléggé nagy szórásúak, feltételezzük, hogy Maxwell-Boltzmann vagy exponenciális eloszlásúak (erről később részletesen.)

**A címkék jellemzése (ez egy utólag általunk hozzáadott oszlop, ami a videókra illesztett ’’tag’’-ek számát adja meg):**



# Kérdések/célok

***Mik a legmegosztóbb témák/kategóriák/csatornák? – Gao Tong***

Itt előzetesen azt feltételezzük, hogy egy videó ’’megosztóságát’’ lehet jellemezni. Minél megosztóbb egy videó annál több komment fog tartozni adott nézettséghez, annál alacsonyabb a like/dislike arány és annál valószínűbb, hogy a kommentek és a like/dislike ki van kapcsolva.

***Kiugróan magas nézettségű videók a csatornához képest? – Gao Tong***

Tekinthető-e egy videó a csatornához képest kiugrónak, ha a csatorna kevésszer szerepel a trendi videók között, vagy többször szerepel a csatorna, de a nézettségek és a like-ok kiugróan magasak.

***A különböző kultúrákban mik a legnépszerűbb témák, kategóriák? Milyen különbségek vannak ebben? – Sike Ádám***

**A feldolgozást főleg Én (Sike Ádám) tervezem megcsinálni, de Gao is segíteni fog néha.**

# Vizualizáció jellege, technológiája

Milyen jellegű vizualizációt szeretne készíteni? **Dashboard**

Milyen technológiát választ ehhez? **Pandas, Bamboolib, Plotly, Dash**

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

# Javítás, pontosítás

Amennyiben hibák vagy hiányosságok merülnek fel a specifikációban, úgy azt frissíteni fogjuk; a hibákat kijavítjuk, a hiányosságokat pótoljuk.