* A feladat végrehajtásához létrehoztunk egy saját repository-t. Ebben a szekesztők név szerint: Nagy Máté (BigM19), Potoniec Vencel (PVencel2003) és Polgár Benedek (YTH40)
* Mindannyian klónoztuk a repository-t egy a saját gépünkön található mappába
* Létrehoztuk a tiktok.py-t, amiben a kódunkat írtuk a feladathoz
* Kezdetnek importáltuk a szükséges csomagokat: pandas, numpy
* Beolvastuk az adathalmazt a pandas read\_csv metódusával
* Megvizsgáltuk az adatainkat és típusaikat, majd kiszűrtük az elemzéshez nem szükséges adatokat
  + A „#” oszlopra nem volt szükség, mert az mégegy ID oszlop lett volna csak
  + Ugyanez igaz volt a „video\_id” oszlopra is
  + A „video\_transcription\_text” oszlop sem hordozott amost számunkra többletinformációt, mert nem készítettünk belőle NLP elemzést
  + A „verified\_status” és „author\_ban\_status” sem volt szükséges számunkra, mert ezek szöveges változók voltak
* Megvizsgáltuk az adathalmazt a hiányzó értékek szempontjából is. Az volt megfigyelhető, hogy ahol a „claim\_status” érték hiányzott, ott hiányoztak a cideóra érkező interakciók adatai is, ezért úgy döntöttünk, hogy ezeket a sorokat töröljük az adathalmazból
* Végül az esetleges duplikátumokat is kiszűrtük
* A mutatószámok kiszámolását a videók hosszának átlagának, mediánjának és szórásának kiszámításával kezdtük
* Ezután hozzáadtunk egy új oszlopot az adathalmazunkhoz „like\_view\_ratio” néven, ami a megtekintések és lájkok arányát tárolja a videókra. Ezt később a megosztó videók szűrésére használtuk.
* A „claim” és „opinion” típusú videók elemzésére ketté szedtük az adathalmazunkat, majd az előbb létrehozott „like\_view\_ratio” változóra számoltunk átlagot, mediánt és szórást
* A megosztó videók kiszűrésére is a „like\_view\_ratio” változót használtuk. Ehhez először megnéztük az átlagot az egész adathalmazra nézve, hogy átlagosan hány kedvelés esik az egyes videókra, és ugyanezt megtettük a szórással is. Ezek alapján létrehoztunk egy intervallumot, és ha ezen kívül esett a videók „like\_view\_ratio” értéke, akkor megosztónak ítéltük őket
* A videók hosszának és népszerűségének vizsgálatára több módszert is alkalmaztunk
  + Az első módszer egy korrelációs mátrix volt, amiben jó látható, hogy a videók hossza és mind a megtekintések, kedvelések és kommentek hossza között nagyon kicsi az összefüggés
  + Készítettünk egy hőtérképet is, amin szintén látszik, hogy nincs összefüggés a megtekintések száma és a viedók hossza között
  + Végül egy lineáris regressziót csináltunk az R2 mutató vizsgálatára. Először csak a videó hossza és a megtekintések számára, aztán a többi videókkal kapcsolatos interakciókra is. Mindegyik arra utalt, hogy nincs összefüggés.