* 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