Intermediate report

./notebooks/data_cleaning.ipynb: contains process of fixing a problem of .csv file having **\n** characters in **text_original column** without quoting – this caused that **\n ' s** were treated as starts of new rows for pandas.

./notebooks/EDA.ipynb: contains EDA but only for TikTok platform as for other platforms – EDA will be the same in terms of general process.

Results:
from **text_original** column using regular expressions were extracted: phone numbers, emojis, links, hashtags, clean description, for emojis was used special dataset and notebook from some person kaggle account. That was mainly done to determine accounts, posts – language. 3 main languages – italian, portugese, polish. 10 accounts – italian, 10 accounts – polish, one – portugese.

For binned by post count polish accounts: accounts with 0-100 posts show best average (by accounts) 90th percentile engagement-to-views ratio, median weighted engagement, 90th percentile community-to-views ratio, 200+ - best 90th percentile views number.

For binned by post count italian accounts: accounts with 0-25 posts show best average (by accounts) median weighted engagement, 90th percentile views, 90th percentile community-to-views ratio, 25-50 posts - best 90th percentile community-to-views ratio, 200+ posts - 90th percentile engagement-to-views ratio.

Posts with text gain higher views with no residuals for all languages.

For multilingual TikTok dataframe hypothesis that tags boost engagement-to-views ratio was true, interestingly, posts with phone numbers have a higher mean ratio (0.19) compared to those without (0.17), suggesting that including phone numbers is associated with higher engagement relative to views.

Using Friedman test no significant trend connected with day of week of post was found. For polish accounts seems better to post at 7 am.