wrangle_report

September 14, 2022

0.1 Reporting: wragle_report

The data preparation was done in three steps: *data collection*, *data assessement* and the *data cleaning* based on the detected errors. > For the **data collection** we had first the file twitter_archive_enhanced.csv with the ids of the tweets and other information such as the ratings. We were also able to retrieve the image_predictions.tsv file containing the results of predictions made on the images of each tweet by using python `requests` library. Using the ids at our disposal, the `Tweepy` library of python and Tweeter API we were able to retrieve other information on the tweets (the number of retweets and favorites among others). At this level, it was found that there are tweets that were probably deleted and for which we could not get data.

> Then follows the step of **assessement** of the three datasets in order to detect errors. It was first done visually with excel and pandas. Then we moved on to the programmatic analysis using some Pandas functions. I used the `info` function to check the types of the columns and the number of missing values then the `describe` function to check mainly the minimum and maximum values of the numerical types columns. After that, I checked for each column the list of unique values to check if there are some aberrations, inconsistencies or if there are duplicated values for columns like id which should be unique.
>
> To finish, having the list of detected errors, I moved on the **cleaning** of these errors. I started by correcting the missing data, then the tidiness issues, then the quality problems and finally I merged the three datasets to get one. For the merging I based myself on the ids of the tweets. I took the ids that they have in common in order not to have any missing data at any level.

```
In [ ]:
```