wrangle_report

August 29, 2022

Project 1 - Report on data wranging

0.0.1 Data gathering

The data has been gathered from three sources: CSV File, TSV File on the server and through Twitter API. To gather data form CSV File, I only had to use a proper padndas method, to gather tsv file, I used a context manager and pandas. Gathering data from Twitter API was more challenging due to rate limits. I created a class that helped me to wait the necessary amount of time. Then I realized there are proper methods to facilitate that more easily, but I went with my class.

0.0.2 Data Assessment

All three databasess were assessed programaticly and visually. I took my own turn on what issuss I would like to tackle and with which I am comforable being unaddressed. The process was circular. Some of the issues I have noticed from the beginning and some I discovered later down the path.

0.0.3 Data cleaning

Data cleaning process was performed as instruced in Udemy-NanoDegree course. I started tackling tideness issues and later the problems with quality dimensions. I probably should have dealt with all the issues of one table before I started cleaning the next one. However, as mentioned previously, the process was circular. It happened that I dealt with one issue, but decided to assess data anew and add another point to a ‘to do list’. Therfore, I cleant tables without a specific order. I also found helpfull the data quality dimension. They helped me both with encountering new issues with data and categorizing the ones that I previously had found.

0.0.4 Data Visualization and Assesment

It took some time of experimentation to find the data visualization that I liked. On their basis, I decided to create insights. Maybe a better idea would be to ask myself an interesting question before analysis and then try to find an answer. But I guess this was one of the reason why I made those particular visualizaitons. Also they look nice.