Wrangling Report

This report briefly describes the wrangling steps of the WeRateDogs twitter archive for data analysis through gathering, assessing, cleaning the data:

1) Gathering:

We have three main sources of data which are:

- **The given twitter archive of WeRateDogs:** using Pandas library, a dataframe is made from the archive CSV file.
- **TSV file with data from a Neural Network that identified the dog breed of each tweet:** using the Requests library, the file was accessed on the cloud and its data was then written to a local file. After that, another dataframe is made from that given TSV file
- **Twitter API (Tweepy) that provides more data for the tweets of WeRateDogs archive:** firstly, I got access to Tweepy through my twitter developer account access keys, then I used the tweets' IDs from given archive to extract more info about each tweet (as favorites count, tweets count). Finally, these data were also used to make the third dataframe in the project.

2) Assessing:

This was the most time-consuming part. However, both visual (using defined panadas methods and excel) and programmatic assessments were used to assess the data. The purpose of this step was to extract some issues, so as to get a cleaner dataset. By the end of this step, I have detected:

- **Quality issues:** such as wrong puppies' names, wrong denominator values (as zero), duplicated data and undescriptive column names
- **Tidiness issues:** such as unnecessary columns which could be replaced by only one, unnecessary columns for the analysis process and more than dataframe that could be merged down.

3) Cleaning

Not as time-consuming as the previous two parts. However, it required no little time to solve code issues and to discover new techniques of cleaning the data (mainly from Stackoverflow and Towardsdatascience), so mainly this step required more effort.

For each of the detected issues from the previous part, I used the "define, code, test" technique to solve these issues, but only on copies of each of the dataframes (so as to iterate back , in case of unrecoverable error on the data). By the end of this step, I got cleaner data for the insights and visualization part and only on dataframe that combined all the required data from the previous three dataframes.

4) Visualizations

For this part, I used Matplotlib, Seaborn, and NumPy to visualize the data and to be able to make four insights with their visualizations using somehow different technique for each. I started by asking some question that could be answered by analyzing the data, then I started to extract the answers and provide visual aids.