WeRateDog project
Rim Issaad

The project was challenging for me. The worst part was to use the API to extract the data and then to put it in the Json file. After that, using python to clean the data frame was not very easy and I faced a lot of issues. I followed the methodology learned in class : first I inspected the data frames both manually and programmatically. I determined the issues, and then started working on them one by one : first on the tidiness issues, and then on the quality issues.
I merged 2 data frames that I think should be one: the archive and the twitter API data frame, so I added the 2 column : retweet_count and favorite_count to the archive data frame.
The tidiness issues were multiple, I separated the URL from the text variable and created a new column and then dropped the expanded_urls column that was full of empty urls. And then I regrouped the 4 columns with dog stages: puppo, floofer, doggo and pupper into one column called stage.
The quality issues were also multiple. I decided to work on the most important for the analysis,

- I changed the source column into simple and understandable values.
- I deleted all the retweets and removed the column indicating it was a retweet because they were not useful anymore.
- Changed the type of the column where it was necessary: tweet_id should not be an integer so I changed it to a string , and changed the other float columns to integer.
- Replaced the NaN values with zeros in order to be able to do the analysis on the retweet_count and favorite_count.
- And finally, I normalized the numerator and denominator rating, by replacing the wrong values