Data Wrangling

In order to complete this project I followed three steps, first I gathered the data by downloading files manually and programmatically. Them, I assessed all the data visually and programmatically in order to detect quality and tidiness issues. Finally I cleaned the data by resolving the assessed issues.

Gathering the data:

This project was based on data from 3 sources:
twitter_archive_enhanced.csv: downloaded from the Udacity classroom which contained the main data of the tweets of WeRateDogs' Twitter account.
image_predictions.tsv: downloaded programmatically from Udacity's server using the request library.
tweet_json.txt: downloaded from the Twitter API using python Tweepy library

Assesing the data:

After I had all the necessary data I proceed to asses it using numpy and pandas methods like sample(), describe(), info(), and value_counts()

Cleaning the data:

To conclude the wrangling I continued to fix or clean all the previous issues. Following a process where first I stated the previously defined problems in English, then I coded the solution to resolve both quality and tidiness issues and to conclude I wrote more python code to determine if the problems were solved.

Issues:

1. The first quality issue was: twitter_archive: contained retweets, and to solve it I dropped the rows where the column retweet_status wasn't null

2. The second quality issue was that the df had columns related to retweets, and since I dropped the rows related to retweets I also dropped the columns that contained information about those retweets

3. The third quality issue was: twitter_archive: source has 4 values, however it has a complete < a > tag, which I solved by replacing them with only the text of the anchor tags

4. The fourth quality issue was that some of the ratings were incorrect, so I extracted the correct ratings set them in their respective columns

5. The fifth quality issue was that some of the dog names were incorrect, and to resolve it I extracted the correct dog names and replaced the incorrect ones

6. The sixth quality issue was that the data had some entries without rankings and consequently weren't of dogs, for that reason I kept only rows that contained a ranking.

7. The seventh quality issue was that some of the dog stages were incorrect, also the value of that stage was into 3 columns, to resolve it I dropped those columns and extracted the correct stage and set it on a new column called dog_stage.

8. The last quality issue was in the image_predictions dataframe, were the name of the predicted dog breed was inconsistent with upper and lower cases, and also it had hyphen and underscores mixed as word separators. For that reason I converted all the prediction names to lowercase and replaced the hyphens for underscores

Once the data what cleaned I proceed to make a quick analysis of some of the variables by creating 4 plots and analyzing the summary statistics which helped me define 4 insights about the collected data.