Wrangle Report

WeRateDogs Twitter Data

Jonathan Obise

Data Gathering

I gathered the data for the analysis from the given csv, json zip file and a website. I was unable to query Twitter's API as permission was not granted at the time of writing this report. I extracted the data from the given files and stored them into dataframes.

Process

To process the data, I performed checks on the various dataframes, searched for missing data, null values, among others. I also created new columns such as the month and day data to help with my analysis and visualization. I was able to extract this information from the timestamp column in the dataframe.

For tweets that had multiple dog stages/breeds, I was able to delete the duplicates and replaced same with a multiple dog stage. I also converted a number of columns such as tweet_ids, timestamp, dog stages, etc to their appropriate data types.

Furthermore, I identified a number of anomalies with the names given to dogs and manually replaced those. The replaced names includes 'a', 'an', 'getting', 'unacceptable ', 'the', 'space', 'officially', 'just', 'one', 'very', 'quite', 'not', 'actually', 'mad', 'space', 'infuriating', 'all', 'officially', '0', 'old', 'life', 'unacceptable', 'my', 'incredibly', 'by', 'his', and 'such'. This is in addition to dropping off columns that had retweets as my interest was on the original tweets.

Storing Data

I stored the cleaned dataframe in a csv file.

Additional Notes

A number of the variables and columns were not useful to my analysis, so I simply ignored them.