Wrangling data Report

The purpose of this project is to gather data using Tweepy python library and with gathered data and given archived dataset, wrangle WeRAteDogs Twitter data to create analyses and visualizations.

The gather part of this project was simple except using tweepy library to get retweet count and favorite count of each ID. When I tried to retrieve data using simple api.get_status function, the resulted json file only contained last ID with retweet count and favorite count of that ID. To fix this problem, I had to search online and ask mentor as well as ask people on slack. At the end, I was able to solve the problem by making a list of IDs and use loop to loop over the IDs and get twitter status and append it into new list. With the list, I appended each value into the dictionaries and created JSON file.

Assessing data was not hard to do, but sometimes I had to add more points since I found out some things that needed to get changed while cleaning. I was able to find total of 12 quality issues with 3 tidiness issue. For the rating numerator, I forgot to use .describe to see if there are any abnormal outliers so I had to go back to change after trying to see mean value of numerators. Another problem was that replies were also included in the data set. I did not know what extended urls mean until I decided to copy tweet_id of the ones that doesn't have urls and actually go into the twitter webpage. Since I thought replies are not actual ratings, I have removed ones without extended urls.

Cleaning part was not hard since it was simple programming exercises.

After gathering and assessing and cleaning data was performed, I moved on to analysis and visualization.

Overall, assessing the data took most effort even though cleaning part took most time.