For Data Wrangling of WeRateDogs, I first create Developer Account on Twitter to fetch Twitter Archive Tweets Data. Then I checked all libraries present with my Jupyter.

Gathering Data:

I installed all required libraries and read Twitter archive CSV file, tweet image predictions TSV file. Create a logic to fetch tweets' data from Twitter using API with start and end time.

I then created list of dictionaries to read tweet's JSON data line by line and later convert to a DataFrame.

Assessing Data:

I followed below checks on the data gathered:

1) If there are any records in tweets_df which are retweets

2) If there are any records in tweets_df whose corresponding record with same tweet_id is missing in img_df table

3) If there are any records in tweets_df whose corresponding record with same tweet_id is missing in status_df table

4) Sort by rating_denominator values

5) Sort by rating_numerator values

6) Sort by names values

After Analysing, I got the following Quality and Tidy issues:

Quality:

1) contains retweets and therefore, duplicates

2) many tweet_id(s) of tweets_df table are missing in img_df (image predictions) table

3) many tweet_id(s) of tweets_df table are missing in status_df (Twitter Archive) table

4) erroneous datatypes (in_reply_to_status_id, in_reply_to_user_id and timestamp columns)

5) unnecessary html tags in source column in place of utility name e.g. Twitter for iPhone

6) some records have more than one dog stage

7) erroneous dog names starting with lowercase characters (e.g. a, an, actually, by)

8) extract dog breed from predection data into img_df

Tidiness:

1) tweets_df table without any duplicates (i.e. retweets) have empty retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp columns, which can be dropped

2) doggo, floofer, pupper and puppo columns should be merged into one column named "stage"

3) merge all 3 dataframes to get single master file in archive_clean

4) Dropping all extra columns which are not useful

Cleaning Data:

Take a copy of tweets_df on which the cleaning tasks will be performed and start removing the quality and tidiness issues.After cleaning and dropping the unrequired columed, I merged the dataframe into single dataframe for further storing the file.

Storing Data:

Dataframe is stored in twitter_archive_master.csv as a single file for further Analysis and Visualisation.

Analysing and Visualizing Data:

For analysis, I created a copy of the cleaned twitter archive data. Following are the analysis and visualisations done:

1) Analysis of rating of dogs

2) Most used Twitter source

3) Analysis of famous Dog's Names