Describing the wrangling effort:

Gathering

The requirement was to extract the data for the wrangling project in 3 different ways- Downloading the twitter_archive_enhanced file manually, requests library on the URL link and using Tweepy library on json data and store it in a file. I was able to retrieve the data mentioned in the first two methods but I had internet issues with tweepy so I had to use the given Json file provided by the instructor. I even tried to use the code to extract the data, it is commented in the wrangle_act.ipynb file.

I opened the tweet_json.txt file and made a new dataframe which had tweet_id, favorites, retweets and timestamp columns.

Assessment

For the assessing the data visually, I used Microsoft excel and sublime to explore how the data is arranged and entries in the columns. Assessment programmatically was done only in jupyter notebook by using different python functions to check out the quality on different dataframes. Different functions used - info(), describe(), null(), value_counts(), duplicated(). After assessment, I added a markdown cell explaining all the issues I found.

Cleaning

It was the hardest part of the project, for easiness I merged all the 3 dataframes (after making copies of all the dataframes) into one to make it a singular effort.

- Dropped columns like "in_reply_to_status_id", "in_reply_to_user_id", etc. these had a lot of missing values and were not contributing much to our dataframe.
- Remove retweets
- Get rid of different columns of different dog stages and merged them into 1.
- Checked if the tweet_id : 835246439529840640 is still in the dataframe.
- Created new lists to store the entries for prediction and confidence level, dropped the extra columns afterwards.
- Capitalization was done to make the first letter of each predicted breed capital.
- Corrected the datatypes.
- Naming under the name column was fixed and changed the irrelevant names to NaN.
- Renamed all the columns at the end.

Testing was done to verify each change to the dataframe. After finalizing the changes, a csv file was created called **twitter_master.csv**

Visualization

Visualized the predicted breed using a histogram and compared the favorite and retweet.