Overview:

WeRateDogs on twitter is a user that rates people's dogs with a humorous comment about the dog. WeRateDogs downloaded their twitter archive and send to us. I will need to access my twitter api to gather more data regarding retweets and favorite counts for tweet_id in this archive file. Also, I'm provided with an image prediction file which classify breeds of dogs.

I would like to first gather all three datasets from various data sources, then wrangle them with a few data wrangling techniques in Jupyter Notebook.

Data and Wrangling Approach:

Data:

- twitter-archive-enhanced.csv (downloaded from left panel)
- tweet_json.txt (connect to twitter developer api, query the data and write to tweet_json)
- image-predictions.tsv (fetch this dataset from url provided)

Data Assessing:

I identified a few quality issues and tidiness issues.

Quality

df table

- timestamp is a string not a datetime
- tweet_id is an int instead of a string
- 181 retweets are identified
- For tweet_id 854010172552949760, a floofer is marked as both floofer and doggo
- A few dog names are weird. Eg. a, an, all, actually, by, getting, his, etc.

df_api table

- id is an int instead of a string
- Missing data in df_api compared to df
- retweet_count and favorite_count should always be int

df_image table

- tweet_id is an int instead of a string
- Missing data in df_image campared to df

Tidiness

- doggo, floofer, puppo, pupper in `df` should be merged into one column stage
- `df`,`df_api`,`df_image` should be in one table

Wrangling Approach:

- Firstly, clean the tidiness issues
    - Make copies for all three dataframes
    - Merge the four columns doggo, floofer, puppo, pupper in `df` into one column stage.
    - Combine three dataframes by joining by tweet_id.
- Secondly, clean the quality issues
    - Change the data type for timestamp from string to datetime.
    - Change datatype of tweet_id from int to string
    - Delete the 181 retweets from df_all.
    - Change the stage for tweet_id 854010172552949760 to the right one.
    - For dogs with weird names, mark them all as 'None'.
    - Truncate data to only include rows whose image prediction data and retweet data is not null
    - Convert retweet_count and favorite_count from float64 to int
    - Change p1_dog, p2_dog,p3_dog to bool value

Result

A clean dataframe combining all three datasets are created: twitter_archive_master.csv.