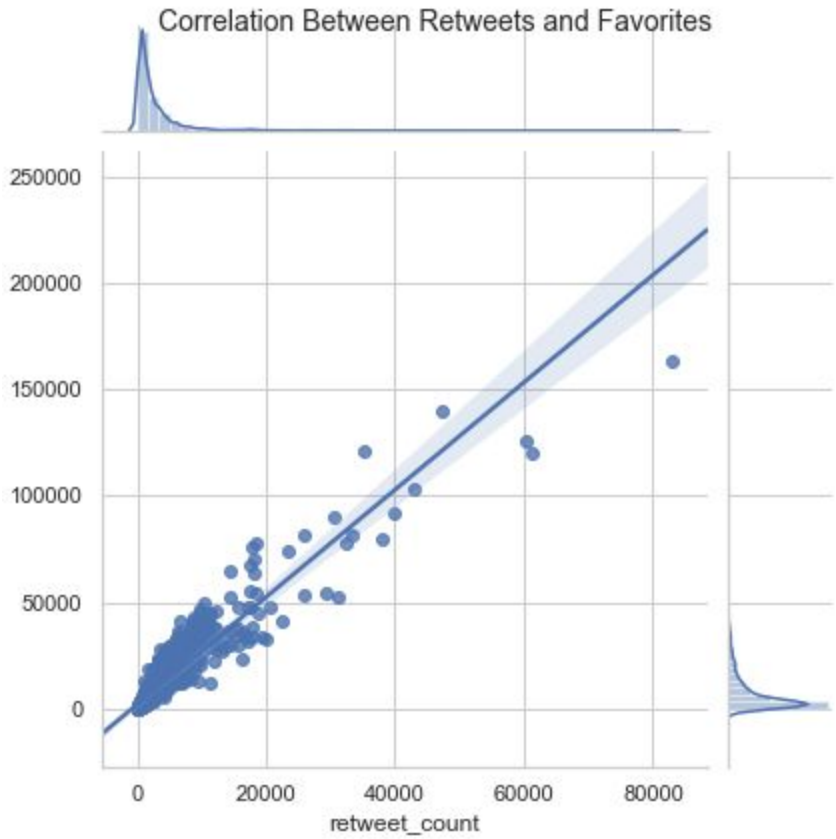
DATA WRANGLING

WRANGLING EFFORT

LINSON IGOCHE ABAH



Data Gathering

As the first step in Data Wrangling, I collected all the needed data from different sources.

1. twitter-archive-enhanced.csv: This file holds the tweet archive for @WeRateDogs and has been pre-downloaded from the classroom to my local machine.
2. image-predictions.tsv: This file will be downloaded from here.
3. tweet_json.txt: This is the extra file to be scraped from twitter.

Packages

I import the needed packages (requests, tweepy, pandas , etc) to send a get request, read files, scrape twitter, analyze and visualize the data.

Data Accessing (Data Quality and Tidying Messy Data)

1. I removed retweets from the data as only original tweets are needed.
2. I converted the timestamp column to datetime format.
3. I created 4 new variables year, month, day, and hour.
4. The wrong words used as name of dogs were replaced
5. The different stages of dog were merged into one column to form a single variable called stage.
6. Tweed_id, name, and stage, were changed to string datatype.
7. Underscores were removed from name column.
8. Values in the name columns were set as title to better present the name of dogs.
9. Renamed “id” variable in the scraped data to “tweet_id” to make it align with other dataframes.
10. Merged the three dataframes into one Master Dataframe

Analysis and Visualization

After merging the three datasets into one master, I went on to analyze and visualize the data. I started by computing a summary statistics for selected variables. Plotted graphs for dog rating, and explored the relationship between retweets and favourites. I also explored the time and day that had the most and least tweets.

Further information about this project can be found in the wrangle_act and act_report files.