The Data wrangle project was about how data analysts clean up dirty data and create insights and visualizations from the cleaned data. For this project, the tweet archive of twitter user data WeRateDogs is wrangled and analyzed. We rate dogs is a twitter account that rates people's dogs. Using the data available I tried to investigate which dog was the most predicted, the most popular dog name and other analysis was carried out for more insights.

I gathered data from 3 different sources.

Data Gathering:

The first data, 'twitter_archive_enhanced.csv' was provided which I downloaded, uploaded it to my notebook and read it into a pandas DataFrame.

The second data was the image predictions, 'image_predictions.tsv' which I downloaded programmatically using the Requests library and the url 'https://d17h27t6h515a5.cloudfront.net/to-predictions/image-predictions.tsv'

The third data was gathered from twitter. Using the tweet IDs in the WeRateDogs twitter archive, I queried the twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of data in a file called 'tweet.json.txt'. this text file was then read line by line into a pandas DataFrame with the following columns: tweet ID, retweet count, favorite count,

Data Cleaning:

As a data analyst, I have learnt that most times data provided is often dirty and must be cleaned before analysis are carried out. The data gathered were accessed in two ways: physically and programmatically. The following issues were observed in the data and cleaned.

Quality issues

1. There were NAN in reply-to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_timestamp columns of the twitter_archive.csv.
2. There was an unwanted <a tag in the source column.
3. The name column was not descriptive enough.
4. The datatype of the Timestamp column was not a DateTime datatype.
5. Tweet_ID datatype was integer and had to be changed to string.
6. Some of the names in the dog name column were not actually dog names like 'a'.
7. Capitalization inconsistencies in the name column.
8. Tweet_ID in the image prediction dataset should be changed from int to string datatype.

Tidiness Issues

1. The columns 'doggo', 'puppo', 'floofer', 'pupper' were merged to form a new column called Dog_type
2. The three tables were merged to form one table using the tweet_ID.

After cleaning the datasets, the new merged dataset was stored in a csv file as 'twitter_master.csv'.

Insights:

Taking the value_counts of the name column, it was discovered that the most popular dogname is Charlie.

Using the p1 prediction, it can be deduced that the most predicted dog was the Golden retriever.

Using a scatter plot, it was discovered that there was a strong positive relationship between retweet count and favorite count.