WRANGLE AND ANALYSE DATA

To wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations.

DATA WRANGLING :

1. Gathering Data

Data for this project was collected from three different sources.

- The first set of data was gathered from the “twitter-archive-enhanced.csv” file. This csv file was imported into pandas dataframe and it was named “archive”.

- The second set of data was extracted programmatically from a URL: ‘https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_imagepredictions/image-predictions.tsv’. In order to extract this data, the request library from Python was used. This extracted file was imported as a dataframe in pandas by using tab as the separator and it was named “images”.

- The third set of data was extracted from the ‘Twitter API’ using python’s tweepy library.The favourites and retweet counts for each tweet was extracted. Then this data was saved as a JSON file.

2. Assessing Data

- info() method from pandas was used to gain information about the data. The timestamp column was found to be an object type which was then converted to datetime type.

- There were many empty values in many columns of the data such as : in_reply_to_status, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.

- The ‘name’ column had several entries which did not look like name and We can also find that the most frequent value in the name column is "a", which is not a name..

- Some of the ratings did not seem like right format of ratings.

- The expected value for numerator and denominator was found to be around 10, but there were many values above 100.

- The number of rows in archive data and images data did not match, which has to be wrangled for better analysis.

- Several columns in the data, null values are treated as non-null values and some entries seem to contain “Nan” value as string.

- The Unnamed: 0 column is not necessary for data analysis, so it should be removed.

- The columns for dog breed predictions may be condensed and made into a single column. The dog stages values in the data is named as columns instead of one column containing all the dog stages values.

3. Cleaning Data

- The extra Unnamed: 0 column was removed, as it was not needed and the timestamp column was converted to datetime object format.

- Retweets were removed as well as the tweets which did not include images were also removed.

- Dog_type column was created which can be used to identify the type of dog.

- The dog breed predictions values were converted into a single column.

- Ratings for dogs was extracted from the text column. The values in the text column is used to create the ‘names’ column. Finally, the data was exported to a CSV file named “twitter_archive_master.csv”