**Data Wrangling Project Act Report Case Study Twitter WeRateDog Post**

In this Data wrangling data was gathered from three source which are:

Gathering Data.

1. Downloading twitter\_archive csv file from Udacity website and reading the data on Jupiter notebook environment.
2. Downloading image prediction post using request.get method from twitter WeRateDog website from link below <https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv>' which was given by Udacity.
3. Th third source is the use of Twitter API which the JSON file was downloaded from Udacity website and read on Jupiter notebook environment using pandas.

Accessing

The dataset was accessed so as to get more information about the dataset using programmatical and Visual approach.

Cleaning

During cleaning the following data quality where cleaned.

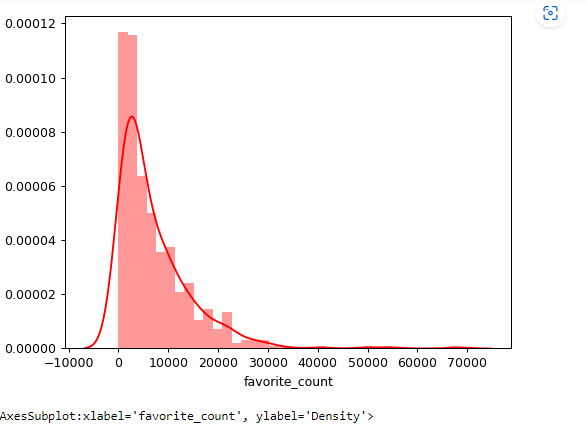
1. Removal of in reply to status id, in reply to status user id, retweeted status id, retweeted status user id, retweeted status timestamp, lang, retweeted\_status, quoted\_status\_id, quoted\_status\_id\_str, quoted\_status, id str, entities extended\_entities, display\_text\_range and truncated column
2. Dropping of the null values in the 3 datasets
3. Data type of Timestamp and created\_at is object which supposed to be datetime
4. Renaming timestamp column to Date, create\_at to Date and id to tweet\_id, column header p1, p1\_conf, p2\_conf, p2, p3, p3\_conf, p1\_dog, p2\_dog and p3\_dog
5. Stripping off of href infront of source column needs to be removed from Tweet\_API\_clean and Enhance\_twitter\_clean
6. Stripping off of Time from date in both datasets Enhance\_twitter\_clean and Tweet\_twitter\_clean
7. Arranging the date and text column in Enhance\_twitter\_clean and Tweet\_API\_clean dataset
8. Dropping of the null values in the 3 datasets
9. Renaming missing values in name column from None to NaN and remove all rows with any NaN values
10. Some tweet has no pictures which needs to be removed
11. Some rating numerators are incorrect because they are below the range usually given (lower than 10 and greater than 13)
12. Remove retweet from the json file by removing all rows where retweet status is not null

I also try to make the dataset to be tidier by carrying out the following process

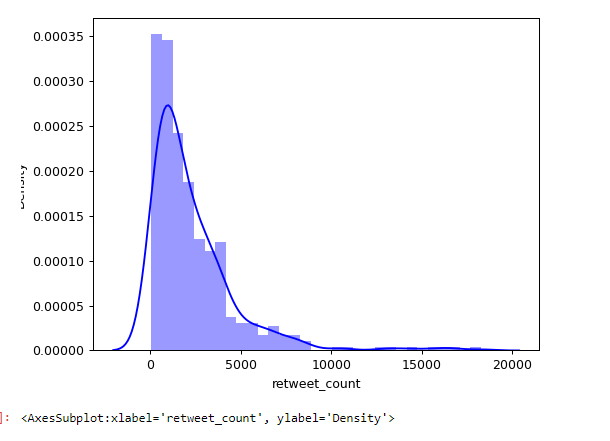
1. Merging the three datasets together to be a master dataset.
2. Merge each dog status column into a one column and named it Stage and remove the duplicated values
3. The rating numerators as well as the denominator column need to be a single column called Ratings

The following Insight was derived from the dataset.

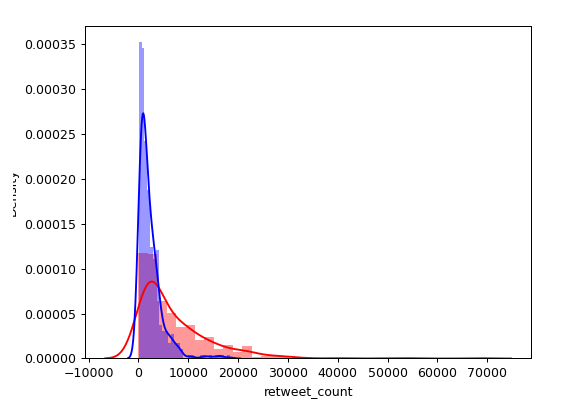
Distribution of Favorite count in the dataset



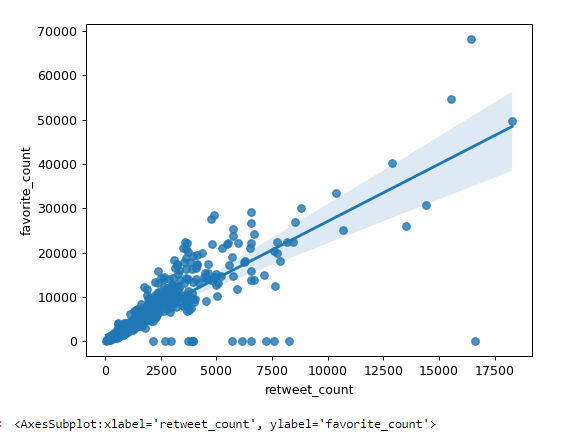
***fig 1.0 the plot above shows the distribution of favourite\_count in the dataset which indicate a gradual increase of the count.***



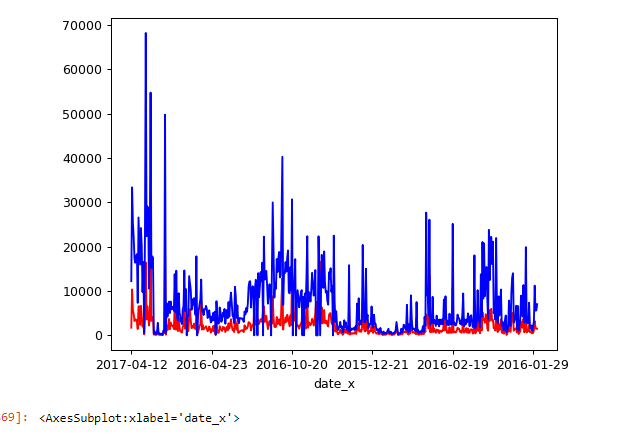
***fig 2.0 the plot above shows the distribution of retweet\_count in the dataset.***



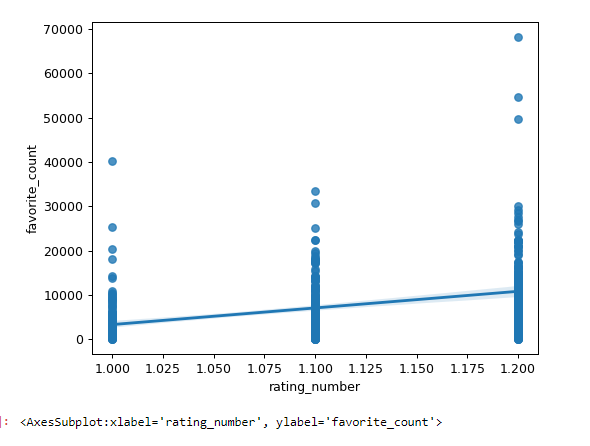
***fig 3.0 the plot above shows the distribution of retweet\_count and favourite\_count in the dataset***



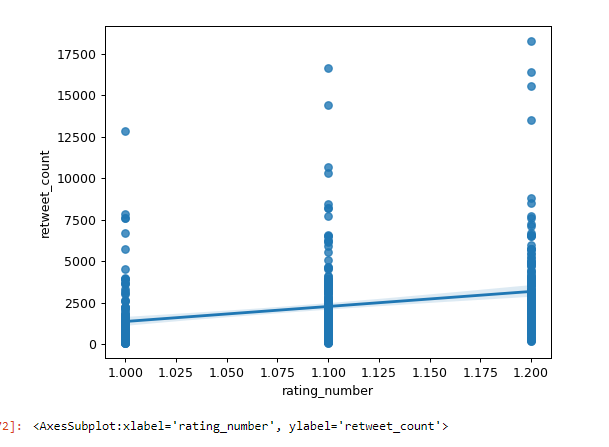
***fig 4.0 the plot above shows the linear relationship between retweet\_count and favourite\_count***



***fig 5.0 the plot above shows the impact of favourite\_count and retweet-count over a period of time***

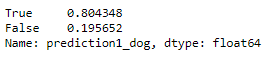
**

***fig 6.0 plot above show the linear relationship between favourite\_count and rating\_number***

******

***fig 7.0 plot above show the linear relationship between retweet\_count and rating\_numbers***

1. Which dog prediction has the highest prediction confidence

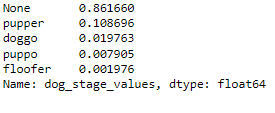






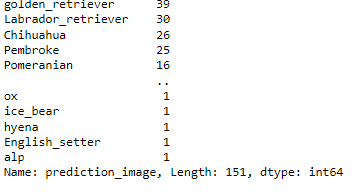
*From the above analysis it can be concluded that Dog prediction 2 has the highest Prediction confidence followed by Dog prediction 1*

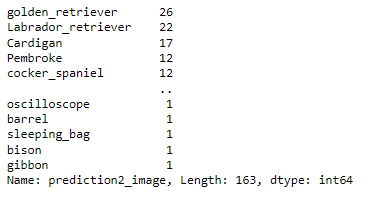
1. Which dog type has the highest Rating

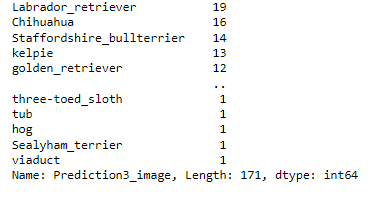


*From above analysis Pupper has the highest rating count followed by Doggo while Puppo has the lowest rating count*

1. Which image has the highest favorite count







*On prediction image 1-3 the followings are the top 2 image predicted:*

1. *Labrador retriever*
2. *Golden retrieve*
3. When do we have the highest rating in the dataset?

