# Wrangle and Analyze Data

As a part of the Data Wrangling project, We did analyses of the tweet archive of Twitter user @dog_rates, also known as WeRateDogs and few other . We Download the data set from the https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv using requests python module. And download the retweets and favorites count for the twitter id present in the data set provide in the Udacity pages.

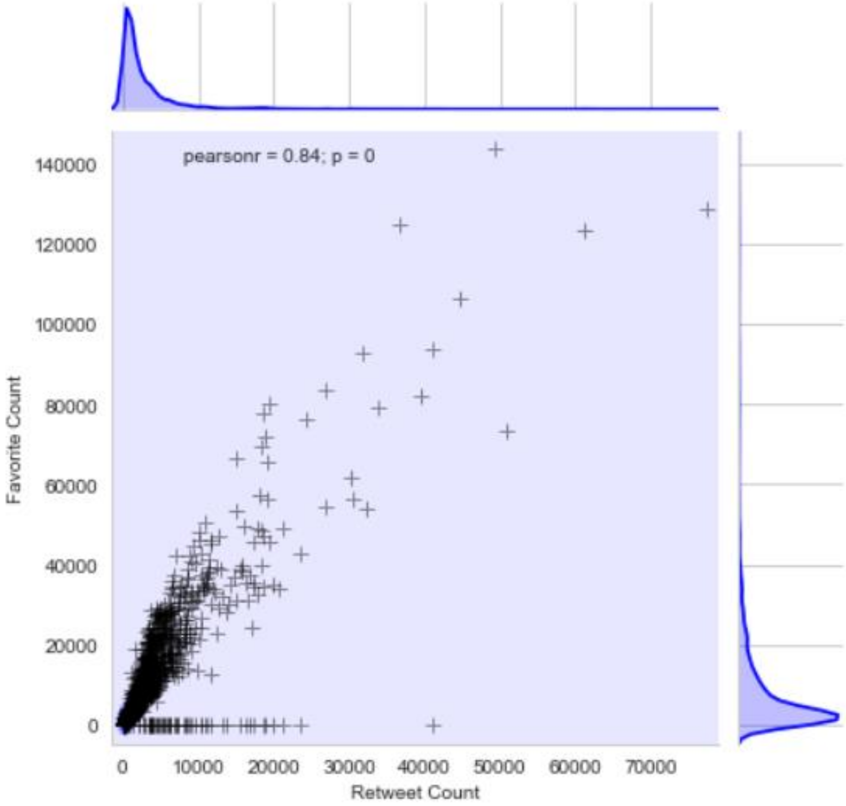
Next we clean the data frames and merge it and dropped the unnecessary columns to ease our analysis.

After Cleaning we performed few Statistical analysis and visualizations.

**Statistics:**

| | tweet_id | rating | favorites | retweets |
|---|---|---|---|---|
| count | 2.06E+03 | 2063 | 2063 | 2063 |
| mean | 7.38E+17 | 1.059913 | 8533.434804 | 2874.03684 |
| std | 6.77E+16 | 0.216818 | 12470.20846 | 4866.529065 |
| min | 6.66E+17 | 0 | 0 | 13 |
| 25% | 6.76E+17 | 1 | 1633.5 | 612.5 |
| 50% | 7.12E+17 | 1.1 | 3777 | 1359 |
| 75% | 7.93E+17 | 1.2 | 10757 | 3343 |
| max | 8.92E+17 | 1.5 | 143551 | 77468 |

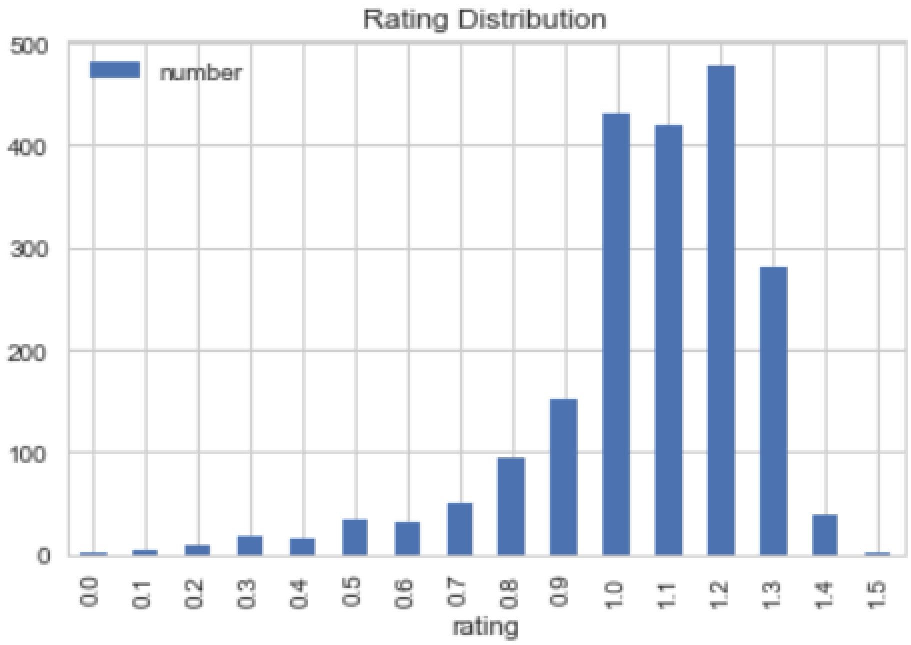
**Plot 1: Retweet count vs Favorite Count**



# Wrangle and Analyze Data

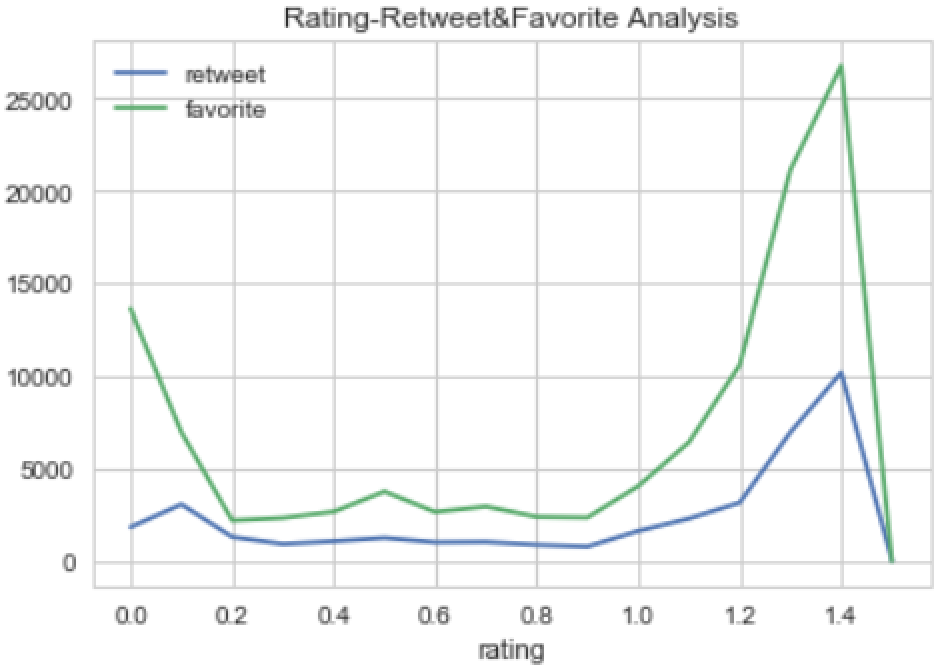
The above plot shows that there is strong correlation exists between favorites and retweets count (as expected). The Correlation between retweet count and favorite count is 0.84.

**Plot 2: Rating Distribution in data set.**



This Plot show the rate distribution in the cleaned data set. As can be seen 475 dogs records in data set has 1.2 rating.

**Plot 3: Rating-Retweet&Favorite Analysis.**

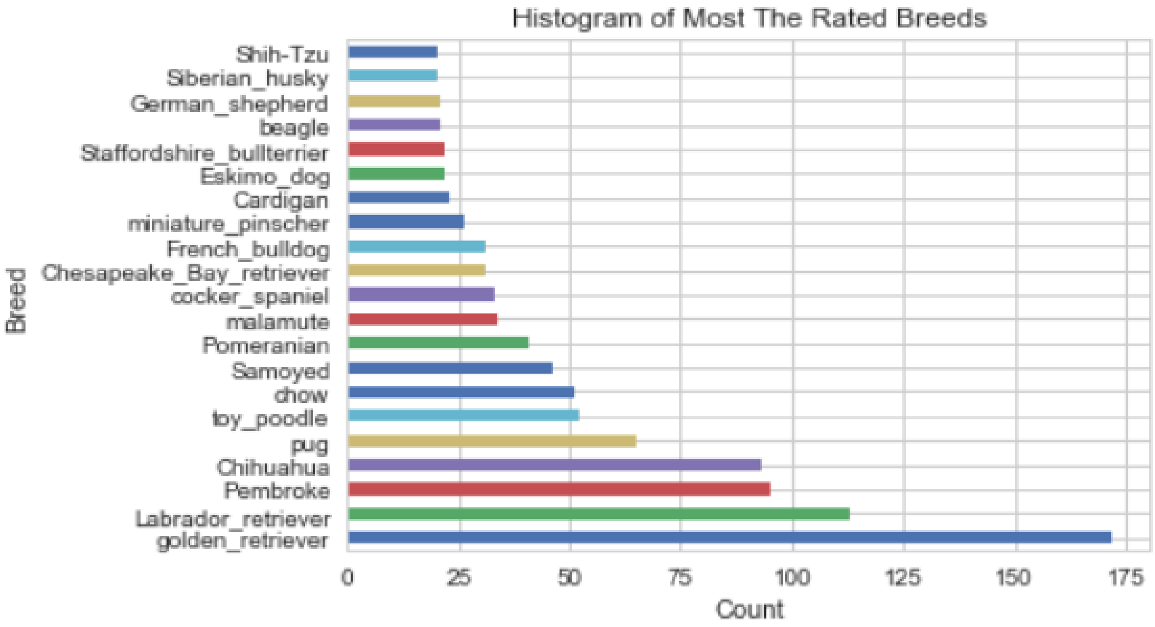


# Wrangle and Analyze Data

From the above plot we can see the relationship between rating vs (retweets count and favorites count).

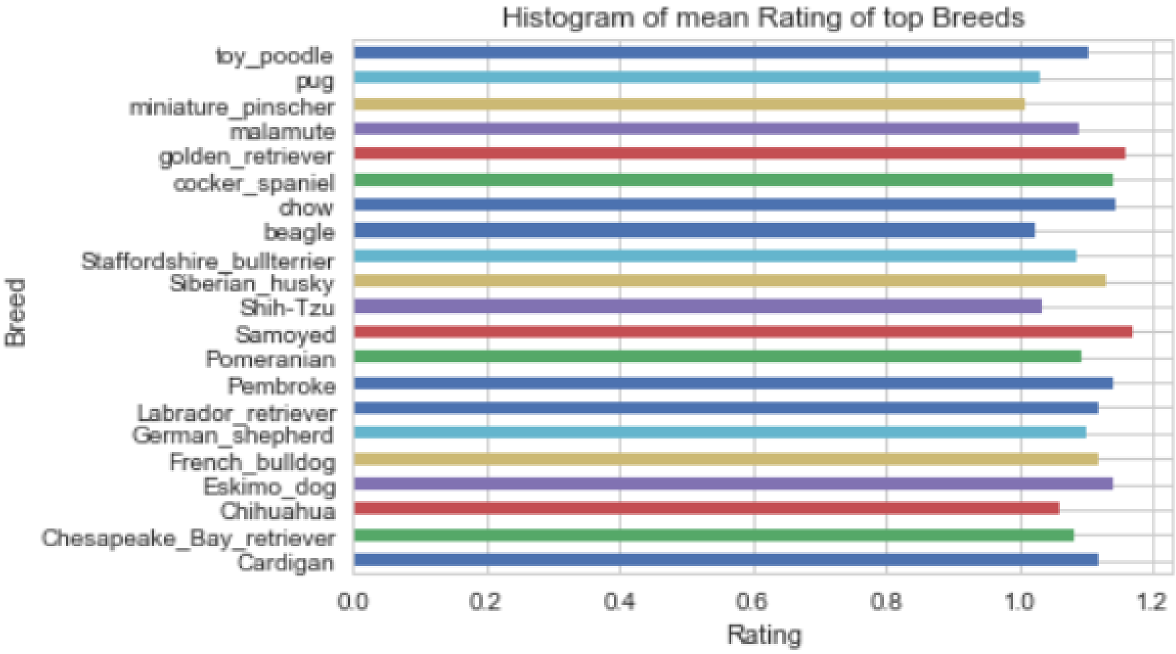
In above we can see as the rating increases there is increase in retweet count and favorite count.

**Plot 4: Histogram of Most The Rated Breeds**



From the above plot we can conclude that, In the cleaned data set the most rated dog breed is ‘Golden_Retriever’.

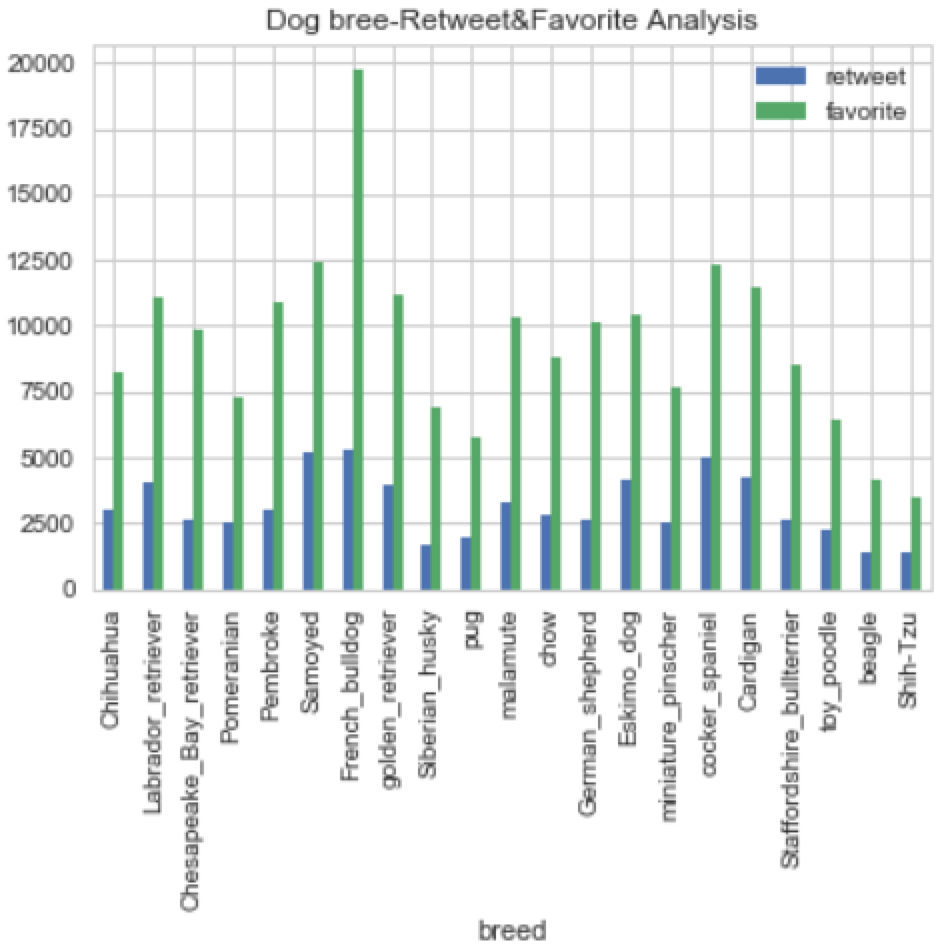
**Plot 5: Histogram of mean Rating of top Breeds.**



# Wrangle and Analyze Data

Though 'Golden_retriever' is most rated dog but the highest rated dog is 'Samoyed' as per above plot. And after that 'Golden_retriever' .

**Plot 6: breed-Retweet&Favorite Analysis.**



From the above plot we can see that the highest favorites is French bulldog instead of Golden Retriever or Samoyed. But the Highest retweet count position is shared by French bulldog and Samoyed breed.

**Summary:**

- There is positive correlation of 0.84 between retweets and favorites.
- There is an increase in no of retweets and no. of favorites with rating.
- I have analyzed the dataset by grouping breeds. Most common breed was Golden Retriever.
- But Golden Retriever does not have highest mean rating .The highest mean rating dog was samoyed.
- The favorites and retweets was highest for french bulldog instead of Golden Retriever and Samoyed.