Project 4: Wrangle and Analyze Data

Act Report

Project is based on Twitter account for **WeRateDogs** that rates people's dogs with a humorous comment about the dog.
As per the project, I performed data wrangling activities and created analysis.

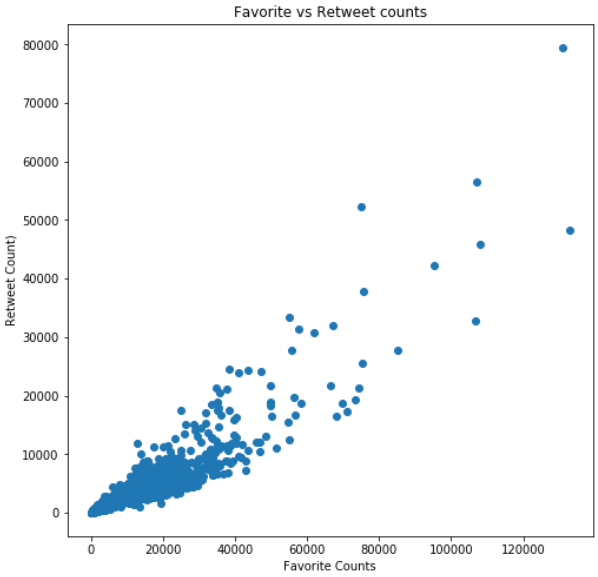
First Step of Data Wrangling, **Data Gathering**:
The WeRateDogs Twitter archive (provided by WeRateDogs and hosted on Udacity server) contains 2356 filtered tweets with ratings. This archive contains basic tweet data like tweet ID, rating, dog stage, timestamp, text, etc. Retweet count and favourite count corresponding to each tweet were not included in the source data. I extracted this data from file (tweet_json.txt) downloaded from Udacity server). I downloaded an image prediction algorithm file from Udacity server containing categorized top 3 predictions for dog breeds based on the images from the tweets.

After gathering of data, I moved to next step: **Data Assessment**.
Here, I assessed the data by both visually and programmatically approach for quality issues (based on completeness, validity, accuracy and consistency) and tidiness issues (structural issues).

After assessing the data, I performed final step of Data Wrangling, **Data Cleaning**.
Here, at first, I addressed completeness issues, tidiness issues and then proceed to clean remaining tidiness issues. After cleaning issues addressed during the assessment, I had 2075 observations obtained after few iterations of assessment and cleaning.

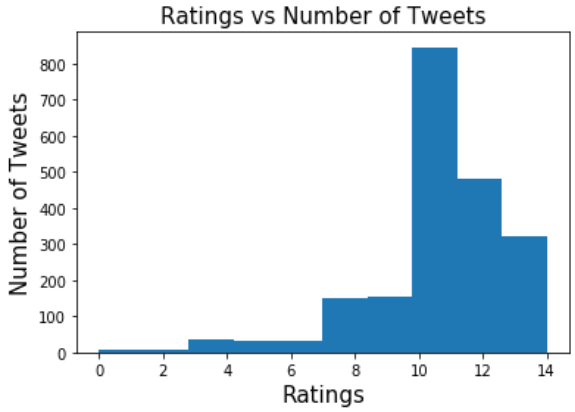
Analysis:

During analysis I found retweet count and favorite count have positive and strong correlation (i.e., increase in retweets would lead to increase in favorites) as shown below:



Tweets Ratings

Below distribution illustrate how many tweets there are with high ratings:



Top 5 breeds are:

Golden Retriever,
Labrador retriever,
Pembroke,
Chihuahua,
Pug

Mean Rating of dogs:
10.6

Most common dog stage:
Pupper