WE RATE DOGS: TWITTER RATING ANALYSIS

Table of Contents

Introduction:

The following dataset is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. **WeRateDogs** is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators are almost always greater than 10 like 11/10, 12/10, 13/10, etc. The reason why the numerators are greater than denominator is that "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively to use it as a part of analysis for their student project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

Gathering Data:

- **Twitter archive file**: Provided by udacity. There are 17 columns and 2356 rows in the data set.

  Original columns of dataframe are as follows:

  1. tweet_id: the unique identifier for each tweet
  2. in_reply_to_status_id: if the represented Tweet is a reply, this field will contain the integer representation of the original Tweet’s ID
  3. in_reply_to_user_id: if the represented Tweet is a reply, this field will contain the integer representation of the original Tweet’s author ID
  4. timestamp: time when this Tweet was created
  5. source: utility used to post the Tweet, as an HTML-formatted string. e.g. Twitter for Android, Twitter for iPhone, Twitter Web Client
  6. text: actual UTF-8 text of the status update
  7. retweeted_status_id: if the represented Tweet is a retweet, this field will contain the integer representation of the original Tweet’s ID
  8. retweeted_status_user_id: if the represented Tweet is a retweet, this field will contain the integer representation of the original Tweet’s author ID
  9. retweeted_status_timestamp: time of retweet
  10. expanded_urls: tweet URL
  11. rating_numerator: numerator of the rating of a dog. Note: ratings almost always greater than 10
  12. rating_denominator: denominator of the rating of a dog. Note: ratings almost always have a denominator of 10

13. name: name of the dog
14. doggo: one of the 4 dog "stage"
15. floofer: one of the 4 dog "stage"
16. pupper: one of the 4 dog "stage"
17. puppo: one of the 4 dog "stage

- **Twitter download**: I had issues with getting access. Hence, used the JSON file provided by Udacity

  Columns of the dataframe are as follows:

  1. tweet_id: the unique identifier for each tweet
  2. retweet_count: counts of retweets
  3. favorite_count: counts of likes by people

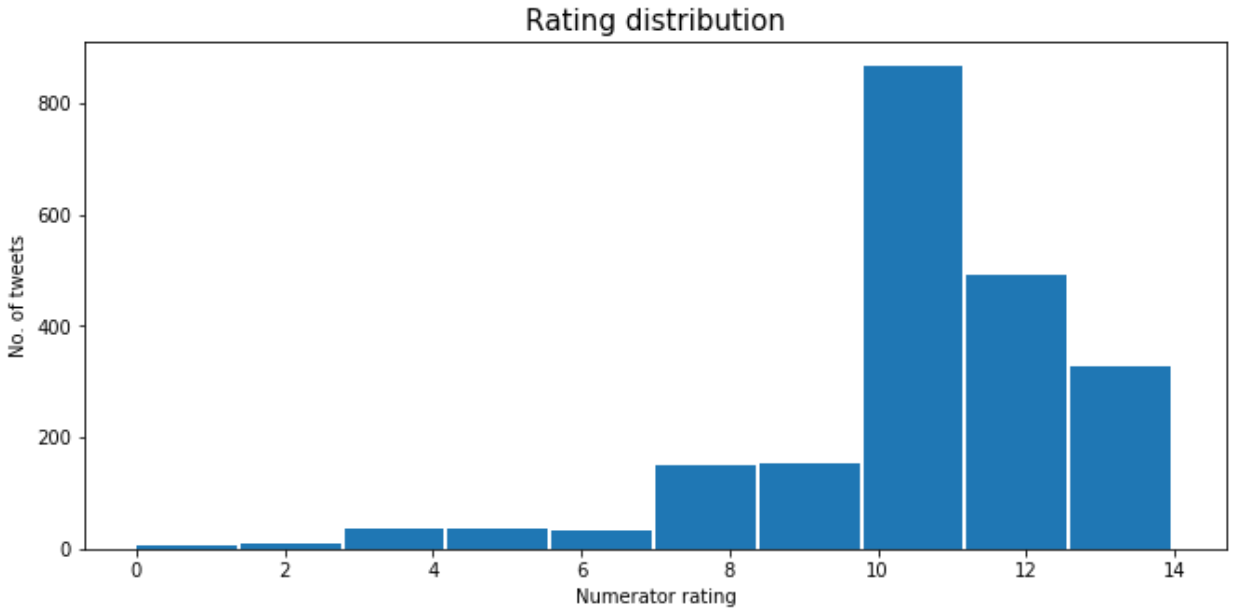
- **Image Prediction file:** This file is downloaded from the servers of Udacity. It has 2075 rows and 12 columns.

  Columns of the dataframe are as follows:

1. tweet_id: the unique identifier for each tweet
2. jpg_url: url of each tweet
3. img_num: No. of images
4. p1: the algorithm's first prediction for the image in the tweet
5. p1_conf: how confident the algorithm is in its first prediction
6. p1_dog: whether or not the first prediction is a breed of dog
7. p2: the algorithm's second most likely prediction
8. p2_conf: how confident the algorithm is in its second prediction
9. p2_dog: whether or not the seconf prediction is a breed of dog
10. p3: the algorithm's third prediction for the image in the tweet
11. p3_conf: how confident the algorithm is in its third prediction
12. p3_dog: whether or not the thir prediction is a breed of dog

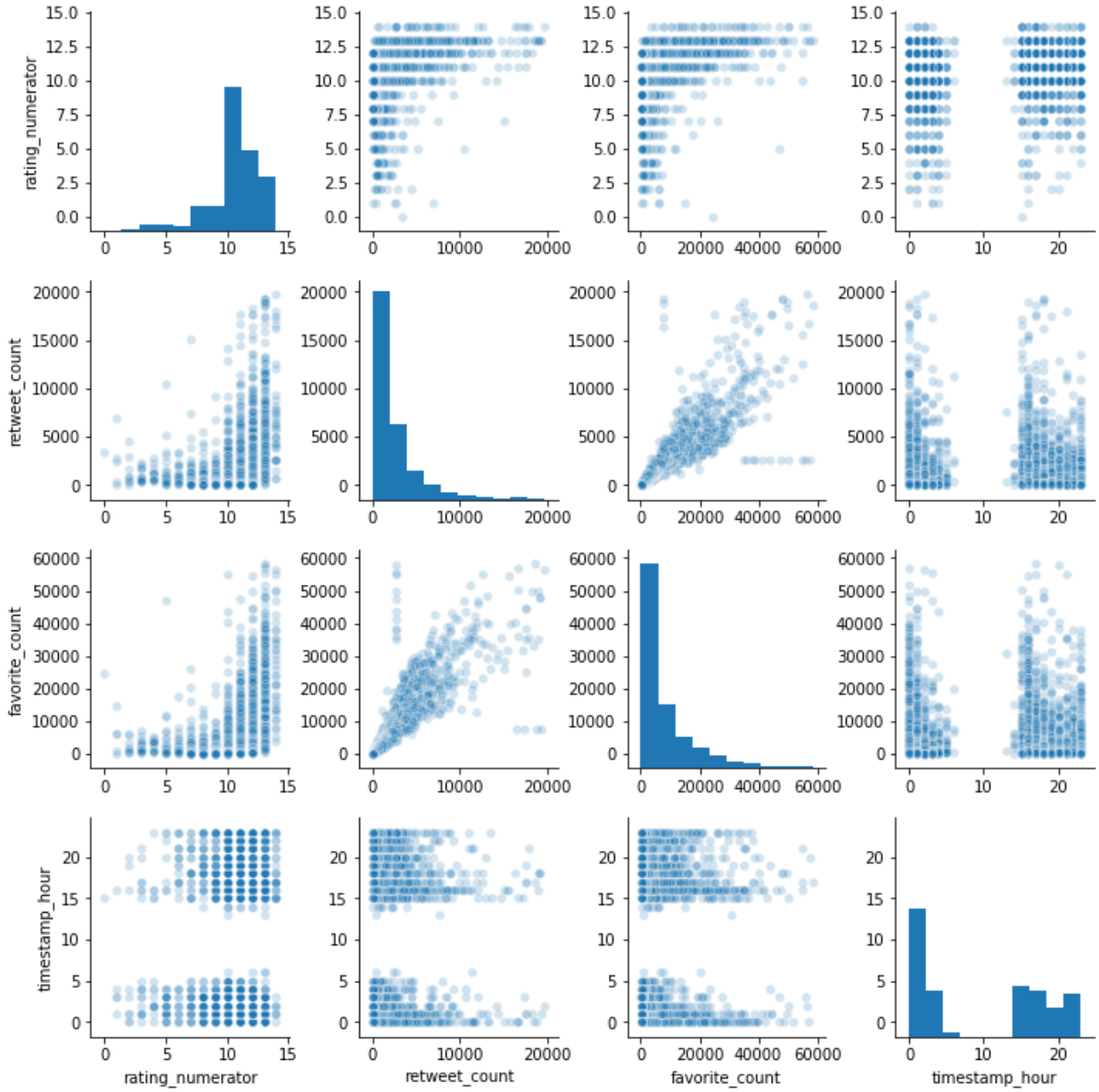
Analysis:

1. **Distribution of the ratings:**



Observation 1: The ratings are negatively skewed. One of the key reasons is that we have capped the ratings at 15 as there were lot of outliers. Although the ratings are out of 10, there are hardly any posts where the ratings are lesser than 10. Pet owners love their dog so much that they rate their dogs 10 or higher.

2. **Relationship between different variables**



Observation 2: Favorite counts and retweet counts are positively correlated.

Favorite counts and ratings are also positively correlated.

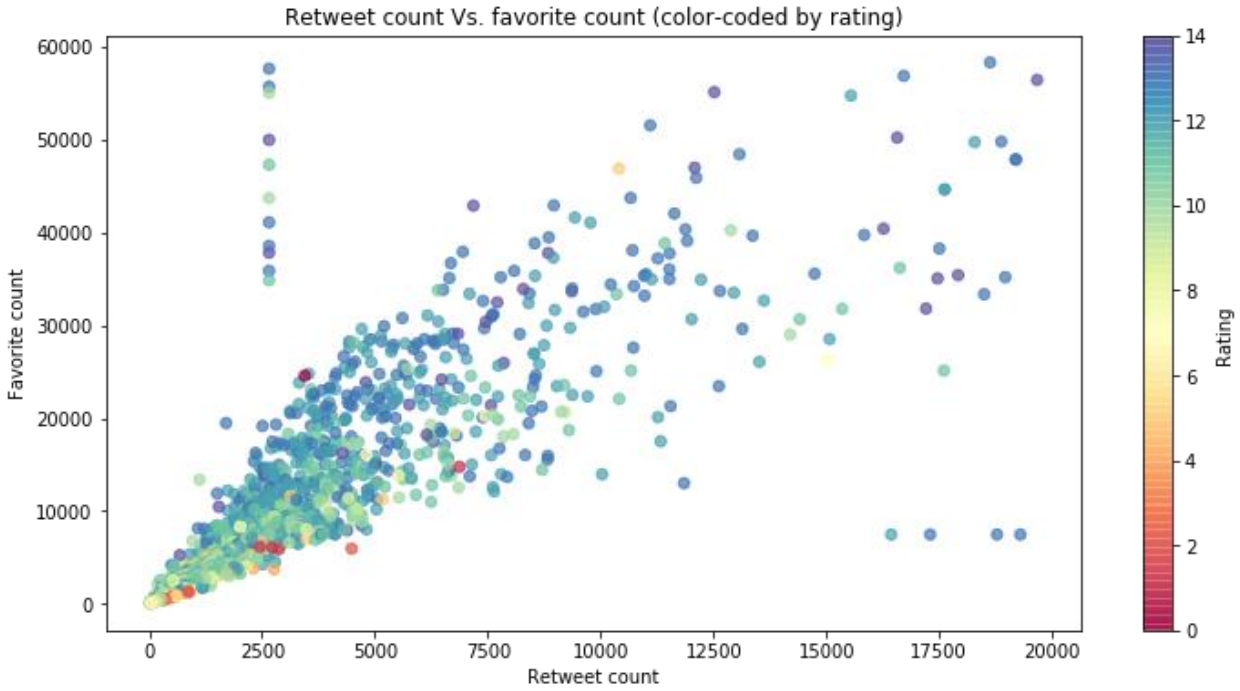
Retweet counts and ratings are positively correlated.

Favorite count it rightly skewed distribution as likes can't be negative.

Retweet count is also rightly skewed distribution as retweets can't be negative.

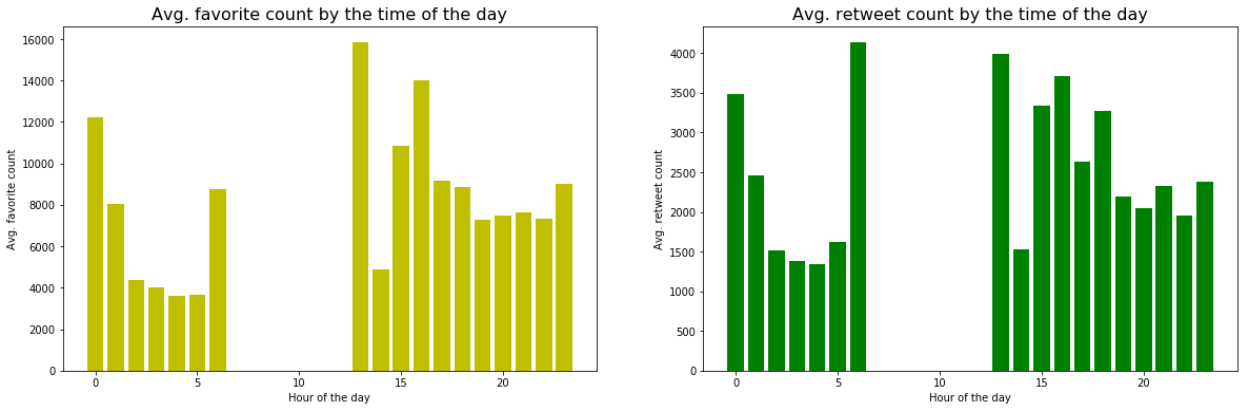
Hardly any activity is seen during peak office hours. The ratings, favorite counts and retweet counts are higher post lunch.

3. **Retweet count Vs. favorite count**



Observation 3: The retweet count and favorite count are positively correlated. The correlation between them seems to be strong. It seems that the tweet with higher rating is re-tweeted or marked as favorite more.

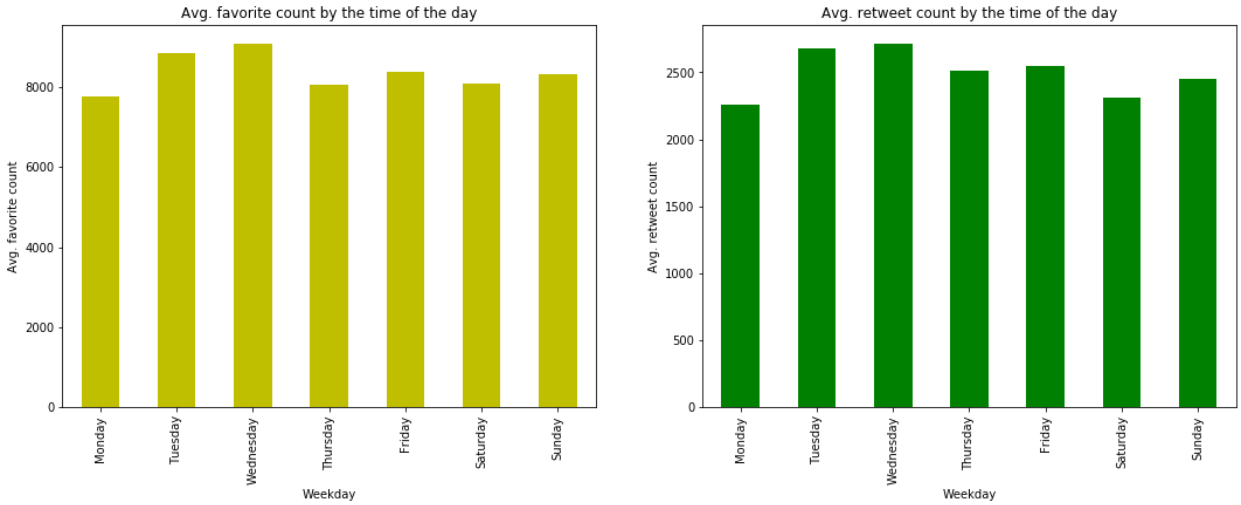
4. Favorite count and retweet count by hour of the day



Observation 4: Most of the tweets happen before 7 am or starting 1 pm. There are no tweets between 7 am and 1 pm. These are peak office hours office hours.The favorite count of the tweet posted at 1 PM is higher than the tweets posted at 6 AM. It might be the time of the day where the boredom sets in and people would be posting and viewing more stuff than usual.

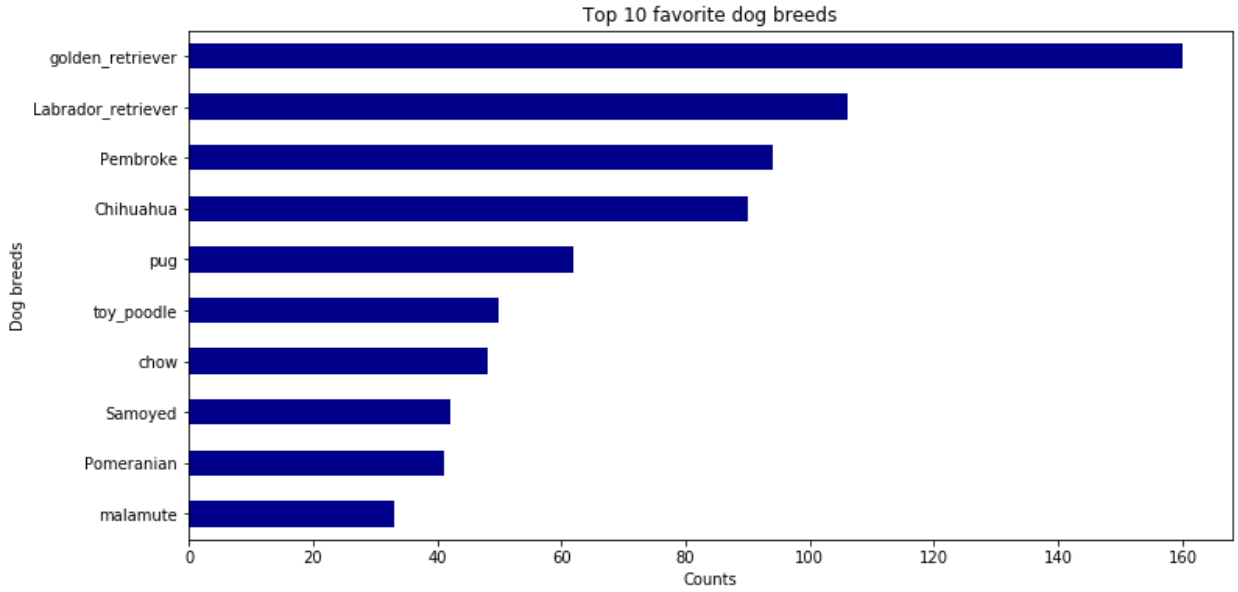
The retweet count of the tweet posted at 6 AM and 1 PM is almost similar.

5. **Favorite count and retweet count by day of the week.**



Observation 5: Average favorite and retweet counts are highest on Tuesdays and Wednesdays.

6. **Top 10 favorite breeds of dog.**



Observation 6: Golden retriever, Labrador retriever and Pembroke are top 3 favorite dog breeds.

Conclusion:

1. The ratings are negatively skewed. One of the key reasons is that we have capped the ratings at 15 as there were lot of outliers. Although the ratings are out of 10, there are hardly any posts where the ratings are lesser than 10. Pet owners love their dog so much that they rate their dogs 10 or higher.

2. Favorite counts and retweet counts are positively correlated.
    - Favorite counts and ratings are also positively correlated.
    - Retweet counts and ratings are positively correlated.
    - Favorite count it rightly skewed distribution as likes can't be negative.
    - Retweet count is also rightly skewed distribution as retweets can't be negative.
    - Hardly any activity is seen during peak office hours. The ratings, favorite counts and retweet counts are higher post lunch.
3. The retweet count and favorite count are positively correlated. The correlation between them seems to be strong. It seems that the tweet with higher rating is re-tweeted or marked as favorite more.
4. Most of the tweets happen before 7 am or starting 1 pm. There are no tweets between 7 am and 1 pm. These are peak office hours office hours.The favorite count of the tweet posted at 1 PM is higher than the tweets posted at 6 AM. It might be the time of the day where the boredom sets in and people would be posting and viewing more stuff than usual.
    - The retweet count of the tweet posted at 6 AM and 1 PM is almost similar.
5. Average favorite and retweet counts are highest on Tuesdays and Wednesdays.
6. Golden retriever, Labrador retriever and Pembroke are top 3 favorite dog breeds.