Manal Almehmadi

DAND Project 8

30 December 2019

WeRateDogs Twitter Archive Act Report

WeRateDogs Data

The WeRateDogs Enhanced Twitter archive contains data extracted from 2356 of the 5000+ tweets from the **@dog_rates** twitter account, posted between November 15, 2015 and August 1, 217.

The retweet count and favourite count for each tweet were not included in the enhanced archive, and so I had to download this additional data from the twitter account using the tweet ID from the archive.

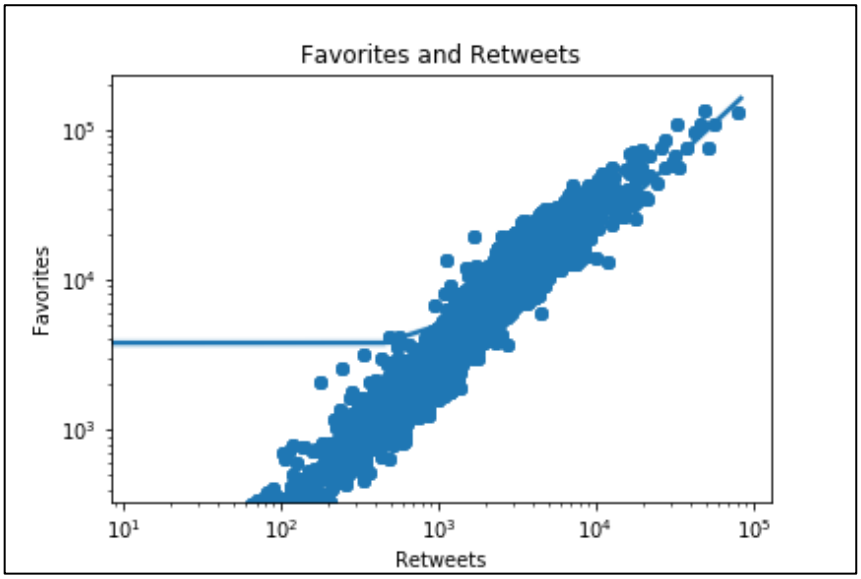
Along with the Twitter data, I also downloaded an image predictions file from Udacity servers containing the top 3 predictions for dog breeds based on the images from the tweets.

Wrangling Data

Before I could begin the analysis. I assessed the data both visually and programmatically for quality and tidiness; the quality of data is determined mainly by looking at several aspects or dimensions to ensure that it is complete, valid, accurate and consistent. In addition, the data had to be each variable forms a column, each observation forms a row and each type of observational unit forms a table.

Insights

Strong relationship between Favorite and Retweet account variables



I used scatter plot to identify the type of relationship (if any) between two quantitively variables.

The above scatter plot shows that there is a positive, strong relationship between favorite and retweet counts with correlation (r= 0.79). in other words, if a tweet has a high favorite count it will also have a high retweet count. Also, both of them are increasing over time. To identify the correlation between the two variables, I used corrcoef function from Numby library.

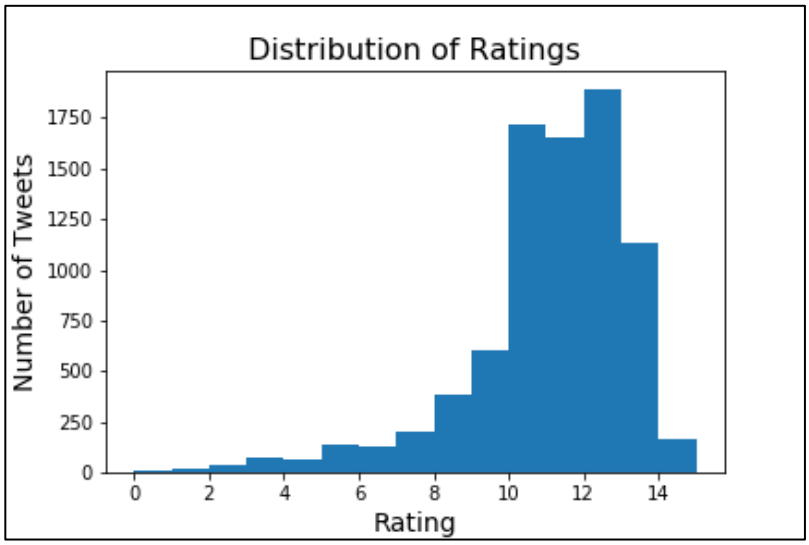
Favorite counts are higher than retweet count

```
In [134]: archive_clean.retweet_count.describe()

Out[134]: count     8292.000000
          mean      2976.089243
          std       5053.982915
          min         16.000000
          25%        634.000000
          50%       1408.000000
          75%       3443.000000
          max      79515.000000
          Name: retweet_count, dtype: float64

In [135]: archive_clean.favorite_count.describe()

Out[135]: count     8292.000000
          mean      8556.718283
          std      12096.451921
          min          0.000000
          25%       1674.000000
          50%       3864.000000
          75%      10937.000000
          max     132810.000000
          Name: favorite_count, dtype: float64
```

Above statistic shows that the median number of favorite and retweet for most tweets are about 3864 and 1408 respectically.

The max value in retweet account was 79515 in other hand the max value in favorite account was 132810. That make me conclued that people are more likely save tweets for themselves rather than retweet it share it with other.

Most of tweets in WeRateDog account with high rating




```
archive_clean['rating'].describe()
```

```
count    8216.000000
mean       10.601266
std         2.174879
min         0.000000
25%        10.000000
50%        11.000000
75%        12.000000
max        15.000000
Name: rating, dtype: float64
```

The distribution of ratings is clearly left skewed distribution. From the descriptive statistics above we see that 75% of all ratings is 12 (the IQR is from 10 to 12).