Act Report - Project: WeRateDogs

Introduction

This is a data wrangling project (also including analysis and visualization) for record which belongs to a Twitter user - @dog_rates, and his nickname of Twitter is 'WeRateDogs'. WeRateDogs is a blogger who rates for various of pet dogs by a humorous method. These rates use 10 as denominator; however, the numerators of these rates are larger than 10 normally, like 11/10, 12/10 or 13/10.

Most of the insights in this report come from the visualization, and some are from further programmatic actions based on the cleaning stage. And all the data I use in this stage is from the ‘twitter_clean.csv’ file, which is created by storing the last stage’s final dataframe - ‘df_combined’.

By ZhenwenXu

Act Report - Project: WeRateDogs

Insights

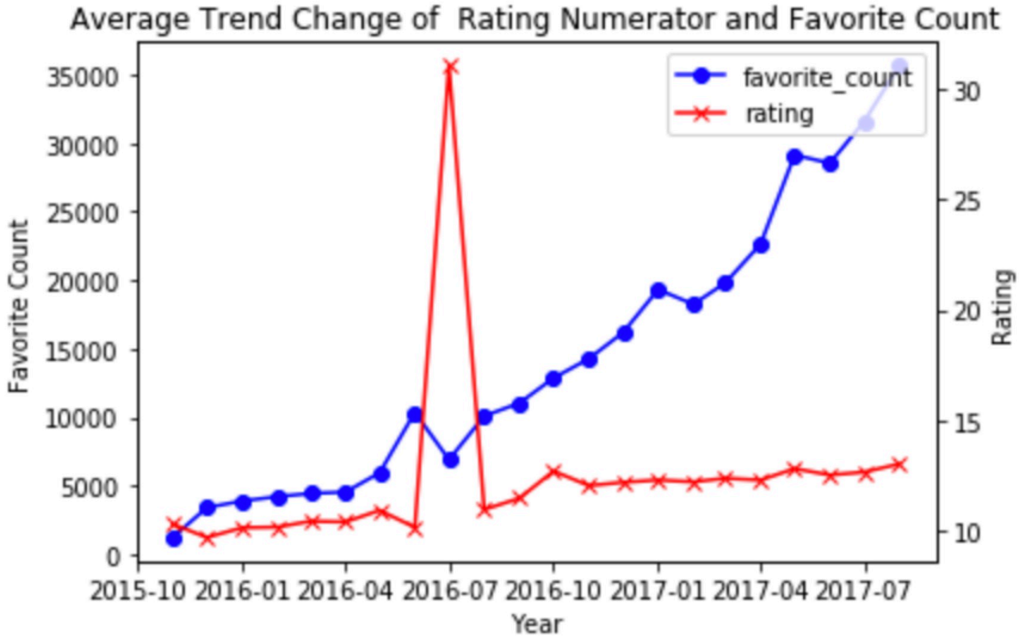
Insight 1:

1. The average rating numerator for dogs keeps a wavy but steady rise except the period between 2016-06 to 2016-08, which means the rate of dogs might becomes relaxed gradually.

2. The trend of favorite count also increases with a relatively steady rate with time goes by.

3. The average rate of dogs reach to about 31 in July, 2016, and this means in this month, the rate might have some outliers.

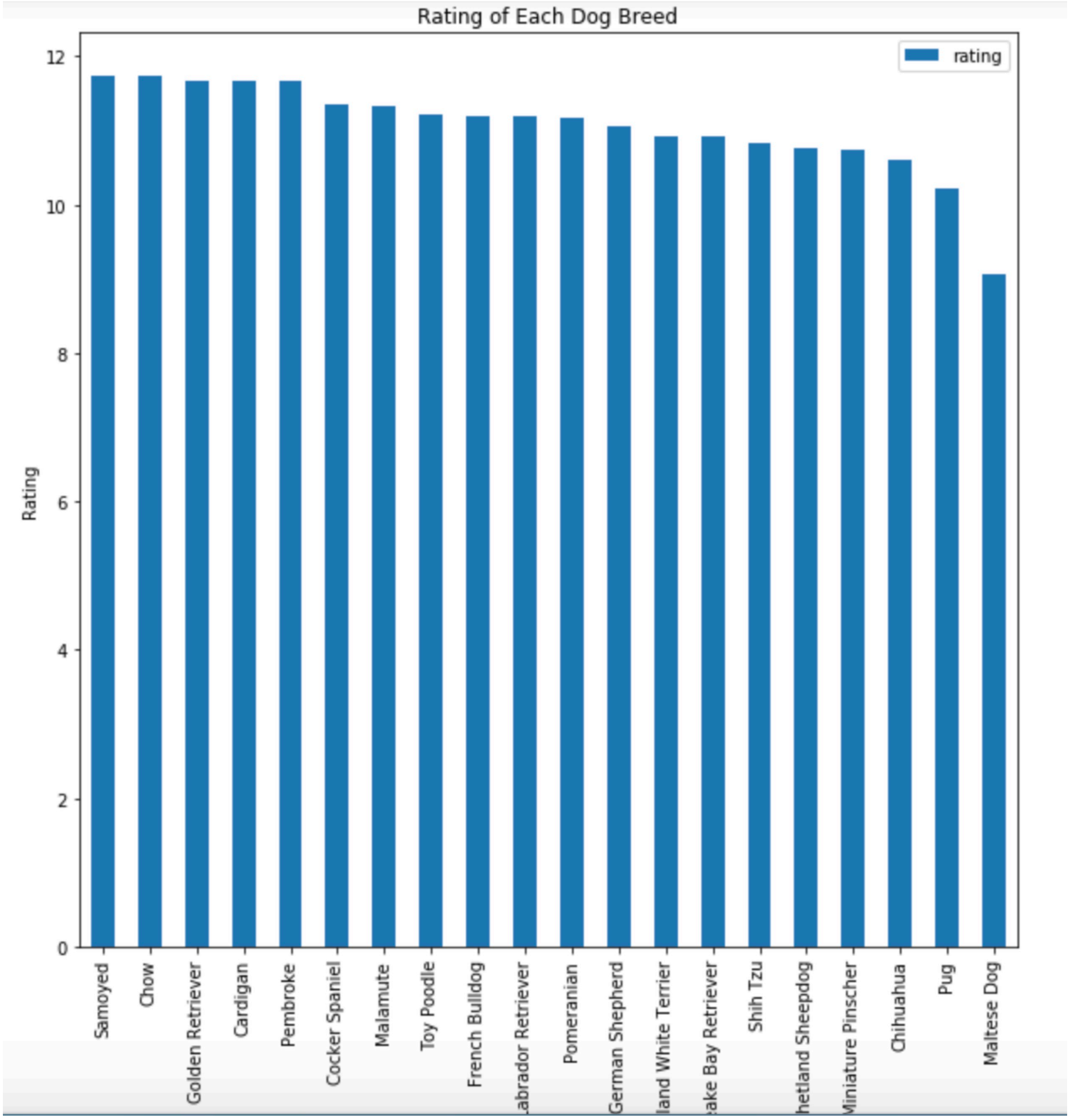
4. The prediction confidence remains a relatively consist range which is from 0.5 to 0.7, and because this range is always over than 0.5, the prediction might be ok.



By ZhenwenXu

Insight 2:

5. Excluding impact of the prediction confidence, missing dog breed, and the outlierthe based on the tweet id, the top three dog breed of rating are 'Samoyed', 'Chow', and 'Golden Retriever', and the bottom third dog breed of rating are 'Chihuahua', 'Pug', and 'Maltese Dog'.

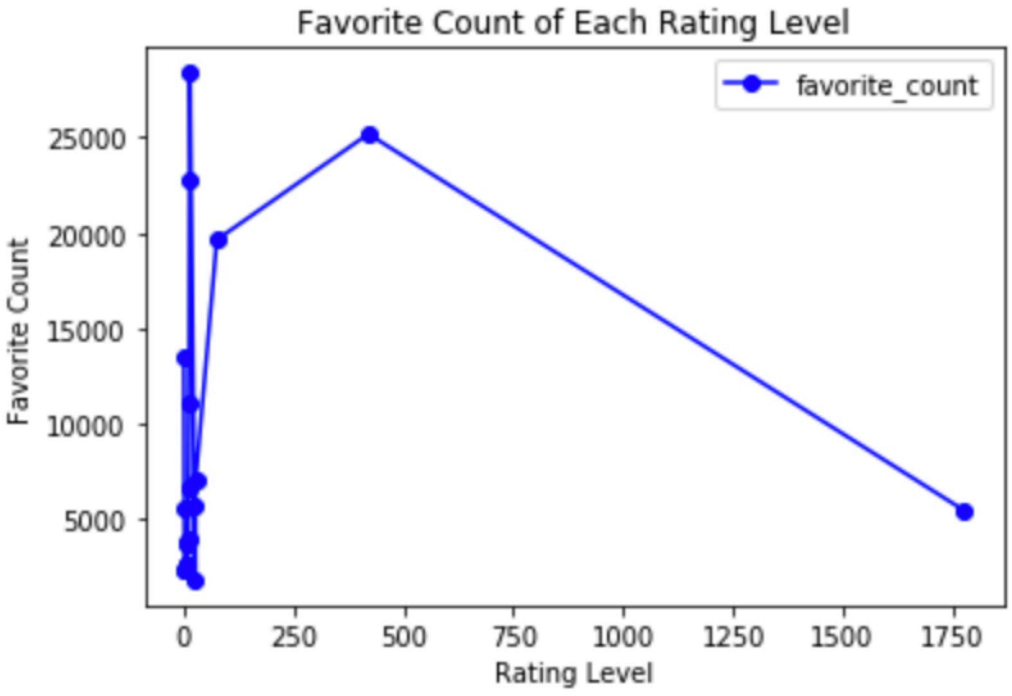


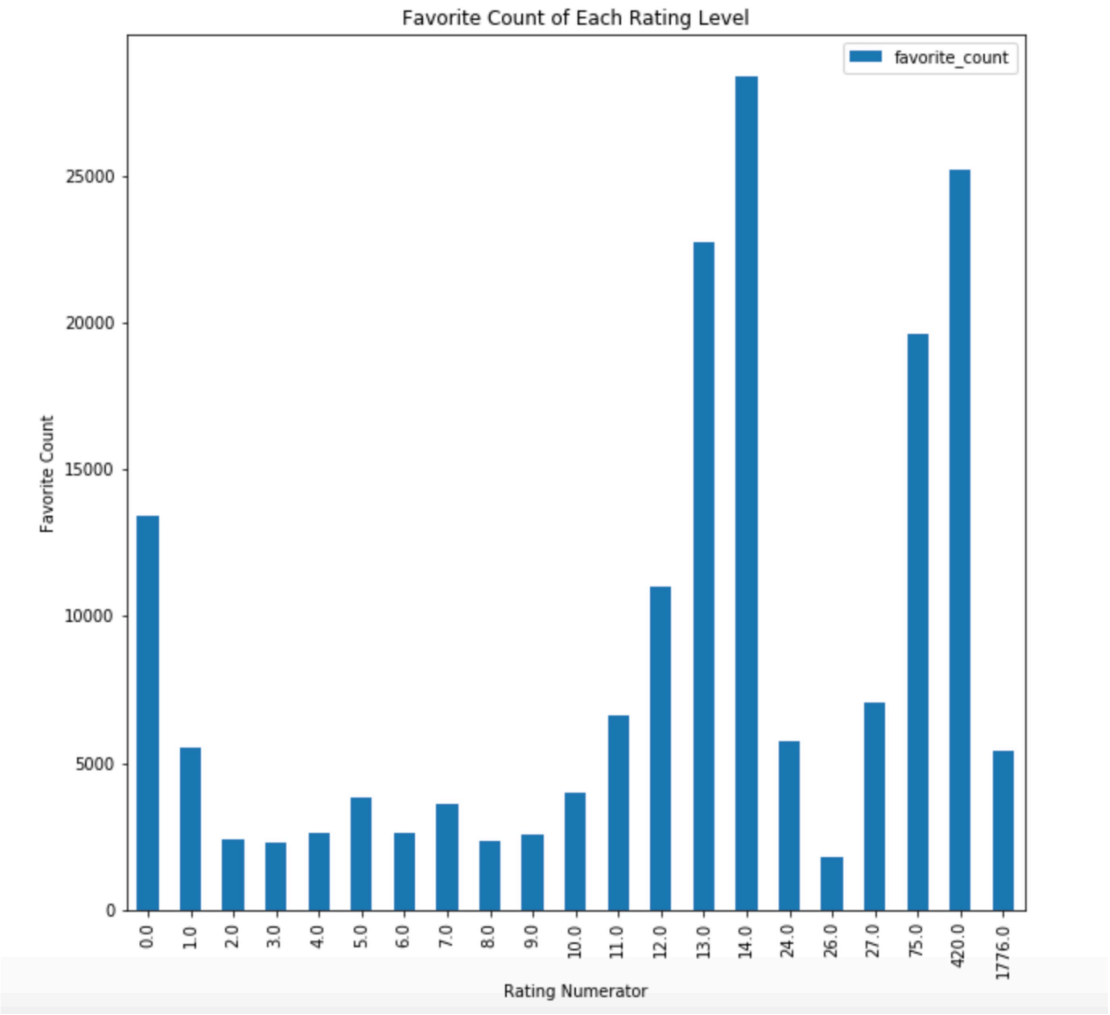
Insight 3:

6. From these two diagrams, it is hard to say these two variables have a direct correlation, but the result shows that it doesn't make sense to say the higher the rating is, the more public will like this twitter.

7. From the first figure, it looks like that most of the ratings focus on the level below to 250,which means the ratings over 250 might be outliers or mistaken records.

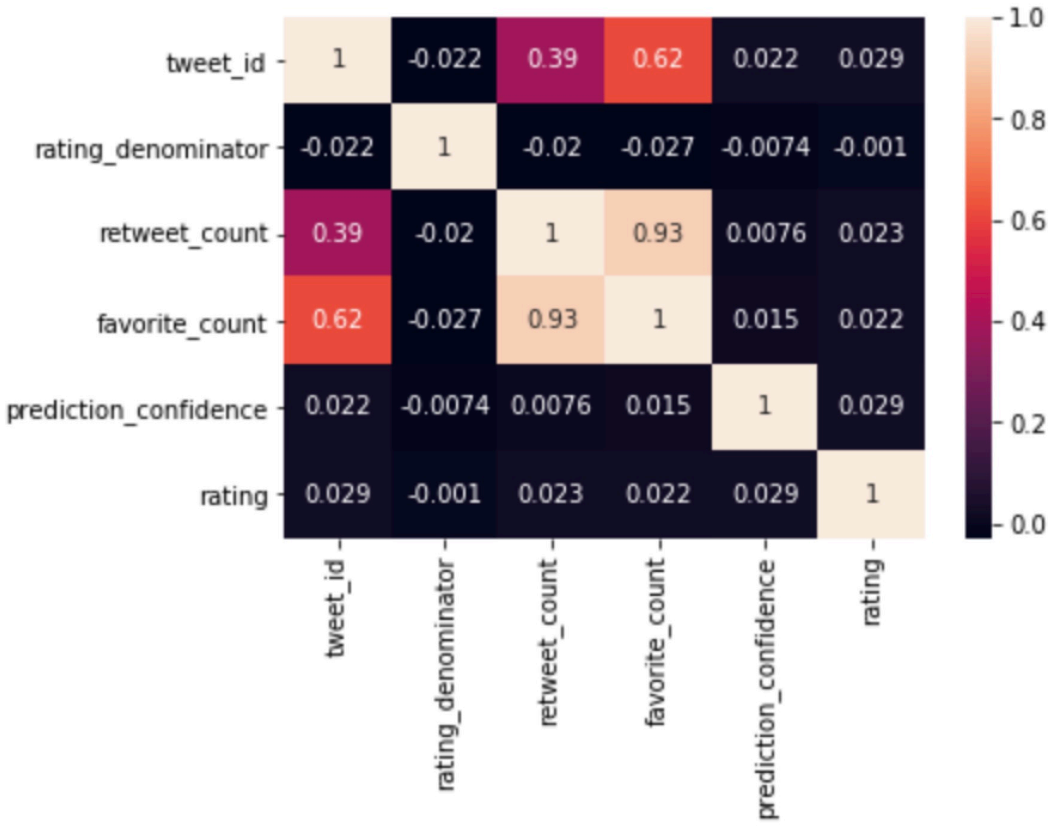
8. From the second figure, it is clear that the rate in 14 owns the most count in public favorite twitter, and the rate in 26 gets the least favor in twitter.




Favorite Count of Each Rating Level
favorite_count
Favorite Count
0
5000
10000
15000
20000
25000
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0
24.0
26.0
27.0
75.0
420.0
1776.0
Rating Numerator

Visualization

Heatmap: It is useful to create heatmap to get the correlation coefficient of each varible, and in this project, I choose to use pearson method as standard.



Insight 4:

9. The most powerful correlation is between 'retweet_count' and 'favorite_count', whose correlation coefficient is 0.93.

10. The weakest correlation is between 'rating_denominator' and 'rating', whose correlation coefficient is -0.001.

Act Report - Project: WeRateDogs

Conclustion

This project is primarily about how to gather, assess, and clean a dataset using Twitter API, and it chooses a famous Twitter host - 'WeRateDogs' to get the rates and text about the dogs attached with some surrounded information, like .jpg image URL, dogs' name, and so on. After the analysis and visualization, it gets many result which shows various of laws based on the cleaned datatset which is called 'twitter_clean.csv' file.

By ZhenwenXu