Udacity Data Analyst Nanodegree Project 7: Analyses and Visualizations Report

Oliver Kröning

October 24, 2018

Abstract

In this project, the process of data wrangling is performed on a dataset consisting of tweets of Twitter user *WeRateDogs*. Once the dataset is wrangled and cleaned, analyses and visualizations are conducted. This report documents the insights and plots made to find relationships and characteristics within the dataset.

1 Introduction

The purpose of this report is to describe the analyses and visualizations, that are performed on a wrangled dataset within the framework of Udacity's Data Analyst Nanodegree project *Data Wrangling*. Beforehand, a data wrangling process came to pass, which is documented in the Wrangle Report. Here, the tweet archive of Twitter user *WeRateDogs* (@dog_rates) as well as other sources has been prepared. In this work, we want to get insights in the dataset using analysis methods to identify relations between different aspects of the dog rating system.

2 Tasks

The aim of this project is to analyze and visulize the WeRateDogs Twitter data. Therefor, we have to perform some filtering and processing tasks to handle the data.
At least three insights and one visualization have to be produced. The actual work is documented and coded in the wrangle_act.ipynb Jupyter Notebook file.

3 Analyses and Visualizations

3.1 Insight #1: Count of Dog Types

The image prediction enables an assignment of the tweets' images to certain dog types. This value is stored in the data base within the columns p1 to p3. The numbers distinguish the order of which dog type is best predicted. This assignment is given with a certain confidence level. We assume that the first prediction p1 is correct.
Thus, we count the numbers for each dog type and visualize them in a histogram.

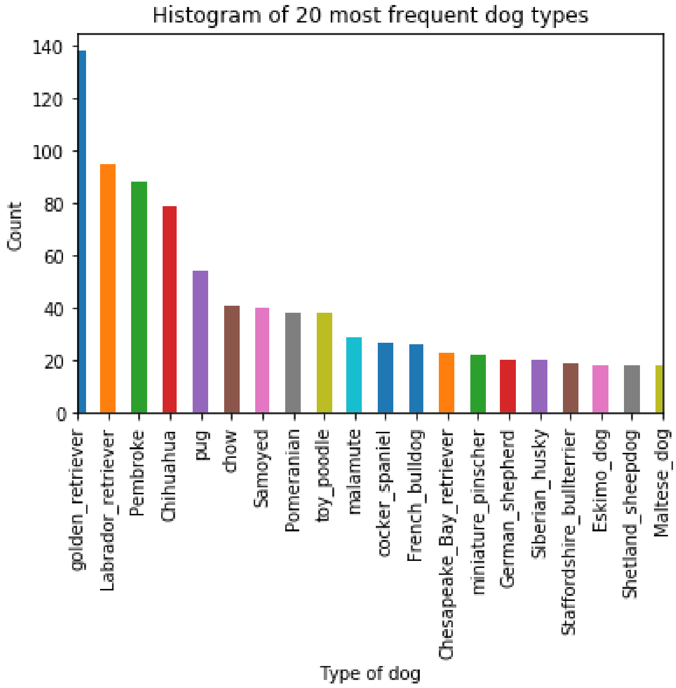
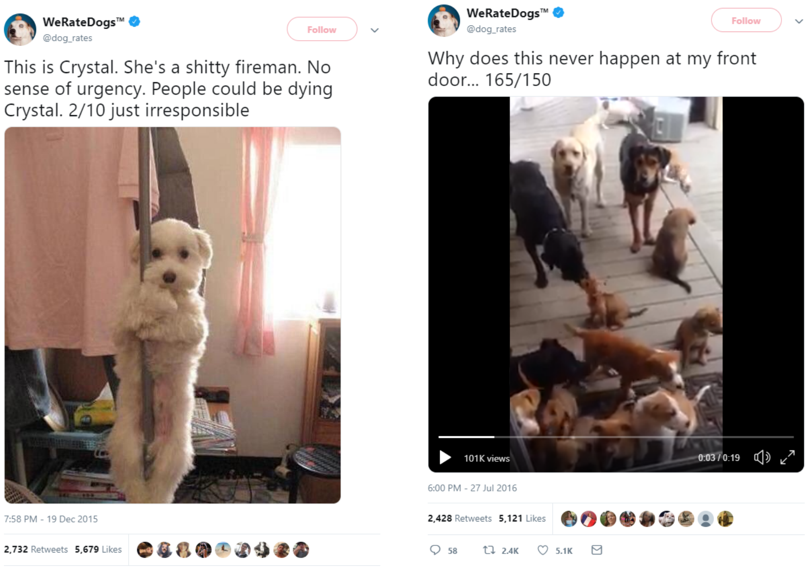


Figure 1: Histogram of the 20 most frequent dog types

Therefor, we only take the 20 most frequent dog types into account. Beforehand, we have to filter the images that actually contain dogs (p1_dog = True). The histogram is displayed in figure 1.

3.2 Insight #2: Mean Rating of Dog Types

Besides the count of dog types, it is interesting to get information about their rating. Are there certain differences between the dog types? Are some dog types cuter than others and, hence, have a higher rating? We also want to know, which dog has the highest and which dog has the lowest rating.
To do so, we have to calculate the mean of the rating numerator for each dog type. Pandas enables to group a dataframe by a certain variable, in our case the predicted dog type. Again, we only consider tweets with dog images. After that, we sort the mean values to identify the dog types with lowest and highest rating.
As a result, we obtain that Japanese spaniels got the lowest mean rating with 5/10. The dog type soft-coated wheaten terrier achieved the highest mean rating with 25.45/10.
Additionally, the tweets with the lowest and highest single rating are displayed in figure 2.



WeRateDogs™ @dog_rates

This is Crystal. She's a shitty fireman. No sense of urgency. People could be dying Crystal. 2/10 just irresponsible

7:58 PM - 19 Dec 2015

2,732 Retweets 5,679 Likes

WeRateDogs™ @dog_rates

Why does this never happen at my front door... 165/150

101K views 0:03 / 0:19

6:00 PM - 27 Jul 2016

2,428 Retweets 5,121 Likes

58 2.4K 5.1K

(a) (b)

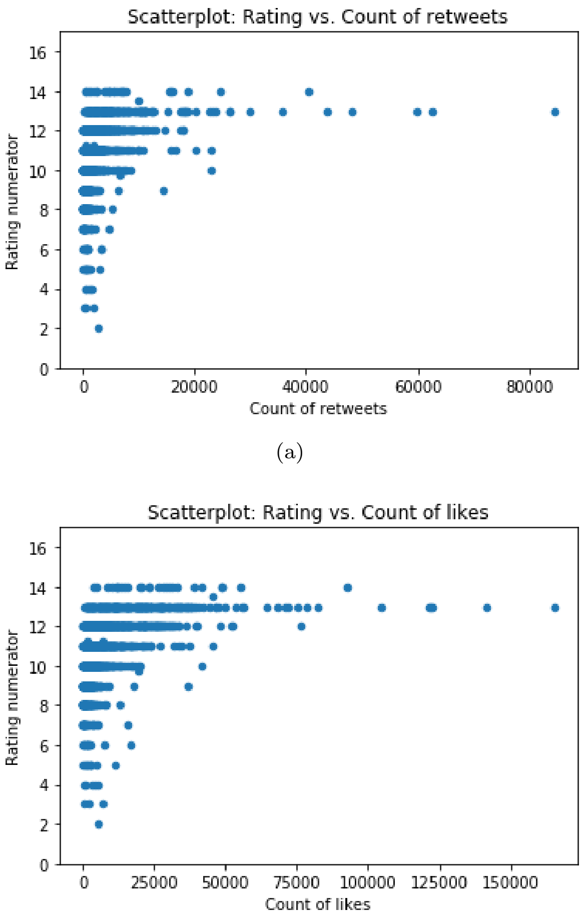
Figure 2: Tweets with the lowest (a) single rating (2/10) and highest (b) single rating (165/10)

3.3 Insight #3: Number of Retweets and Likes vs. Ratings

In the last insight, we want to explore the correlation between ratings given by the publisher and the counts of retweets and likes. Are dog images only rated subjectively or is there a relationship between rating and the objective opinion of the community with regard to that image? As an indicator, the number of retweets and likes of the tweet are chosen.
To analyze this relationship, two scatter plots are created. The first plot (see figure 3(a)) displays the relation of the rating numerator to the count of retweets and the second plot (see figure 3(b)) shows how the rating numerator is correlated to the number of likes.
Since, outliers can be detected at high rating numerators and low retweet and and like counts, we focused on rating numerators between 0 and 17. The majority of the data points can be found within this range.

4 Conclusion

In this report, we described and documented some insights and visualizations of the dataset, we gathered, assessed and cleaned from Twitter user *WeRateDogs*. Statistical methods have been performed to show and analyze interesting features of the dog rating system. The aim of this work was to put the learned aspects of the data wrangling part into practice.
In the first insight, we analyzed the distribution of the dataset related to the dog types we determined using a machine learning image prediction. The most



(a)

(b)

Figure 3: Scatterplots showing the rating vs. number of retweets (a) and rating vs. number of likes (b).

frequent dog types are the golden retriever followed by the Labrador retriever and Pembroke.

The second insight gave information about the mean ratings of the dog types. Obviously, there are huge differences between the ratings for certain dog types. The Japanese spaniel only achieved a mean rating of 5/10, while the soft-coated wheaten terrier got a mean rating of 25.45.

The last analysis considers the relationship between the rating numerator and the number of retweets as well as the number of likes. In a scatter plot visualizations, both relationships were displayed. Both plots show a positive correlation, although the relation between the rating numerator and the number of likes seems to be slighly stronger correlated.