Project 4: Act Report

Omar Bougacha

April 2020

1 Introduction

The data wrangling process allows us to have a clean and tidy data to analyze. However, the analysis part of the data could actually start in the wrangling process. A first rapid analysis of obtained data could help us define whether or not we need some extra cleaning and or gathering. In this report, I document the insights I drew from the previously wrangled data. All the details about the codes can be found in the **wrangle_act.html** file.

This report presents some of the insights I drew after a first rapid analysis pass. In section 2, I present some visualizations about the tweets data and their respective analysis. In section 3, I present the insights I drew from the obtained dogs' data. Section 4 contains a combination of the data to answer an analysis question. Finally, I conclude this report with a summary of the results and what I have learned during this project.

2 Tweets Insights

After taking a small gaze at the next step of the data analysis course, I found that it is a good idea to start the investigation/analysis of data by doing some uni-variate exploration. So I looked into the tweets dataframe and found two ideas about this exploration. In the dataframe we have two continuous numeric variables:

- The number of retweets, and
- The number of favorited/likes on the tweet.

So I decided to take a closer look at these features. First I perform a uni-variate exploration for each of these variable. Then, I study the relationship between the two.

Since the variable are continuous and numeric, I plotted the histogram of each variable. The figures 1 and 2 present the histogram of respectively the number of likes and the number of retweets.

From these histograms, we can easily note that the variables are skewed to the right. We can also see some points that are quite distant from the others (outliers maybe?).

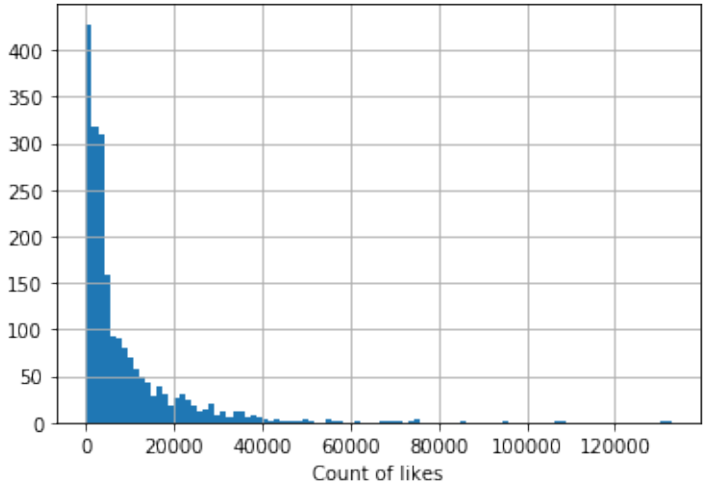

Figure 1: Histogram of Number of Likes

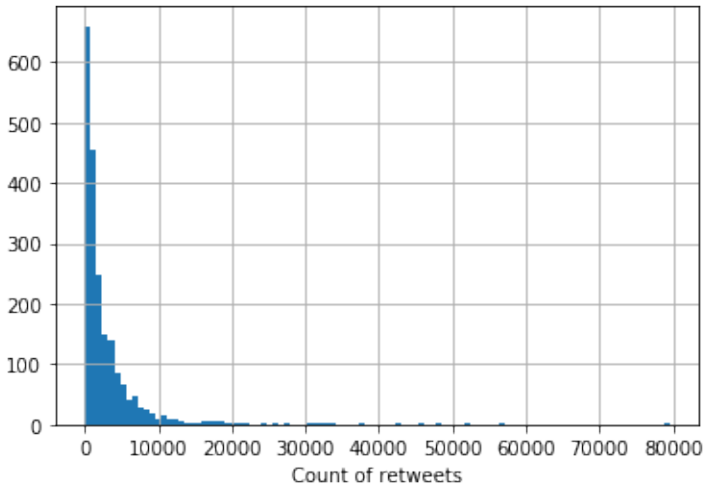

Figure 2: Histogram of Number of Retweets

Let's see what's the relationship between these two variables. Let's plot the points of retweets as a function of the number of likes. Figure 3 presents this visual.

We can note that the retweets and the likes are positively correlated. The obtained results are quite logical. The high the number of likes reflects that people like this post. Which means that they are more likely to retweet the picture. However, we can see that a large number of twitter posts have zero

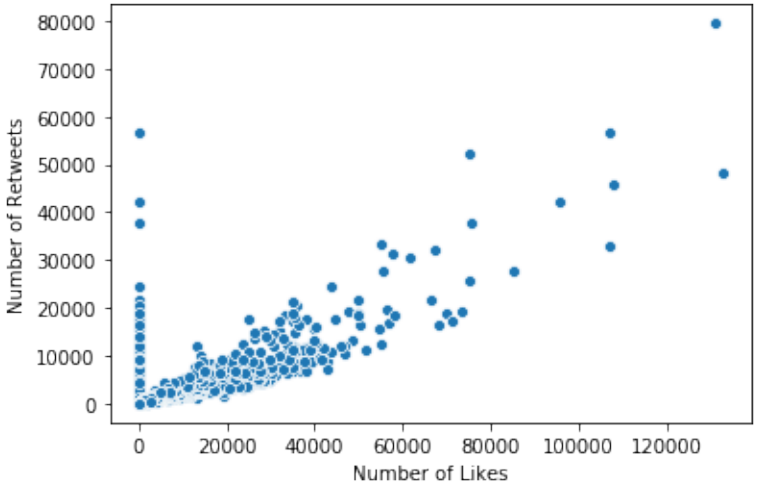

Figure 3: Scatter-plot of the Likes-Retweet Relationship

likes and yet they are retweeted largely. Is this a problem in the data wrangling process?

3 Dogs Insights

For the Dogs insights, I choose to go for a bi-variate analysis. At first, I choose to bar-plot the average numerator rating to each dog stage. I obtained the figure 4.

We can see that the rating of no-stage dogs is the highest on average. However, if we take a look at the numerator rating we can easily note that the numbers are quite subjective. On the other hand, the denominator is not the same for all ratings. So the comparison is not valid. Therefore, I created a new feature called rating_value in which I computed the rating by dividing the numerator by the denominator. And then I repeated the visualization.

Now, we can see on figure 5 that now the results changed:

- Puppos is now the highly rated stage instead of the no stage dogs

However, pupper is always the lowest rated stage. Moreover, we can see that the mean rating is always above 100

4 Combination

One interesting thing we can do with these tables is to combine them to find which dog stage had the highest mean of likes and retweets. For this we bar plot the mean of likes (respectively retweets) in function of the dog stage. We

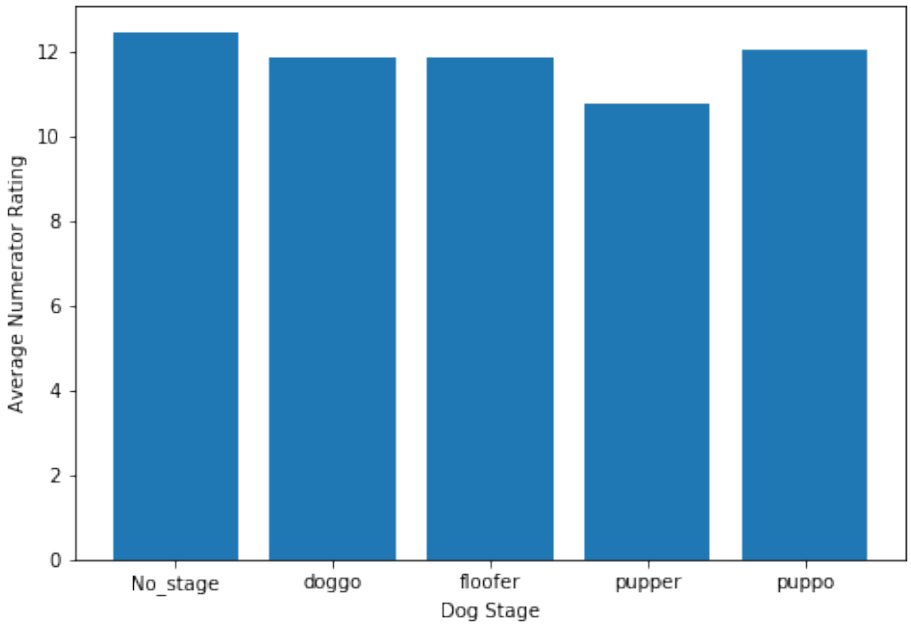

Figure 4: Bar-Plot of the Average Numerator Rating per Dog Stage

obtain figures 6 and 7. I used the mean because we don't have the same number of dogs in each stage so the sum will favor the stage with the highest number of dogs.

From these figures, we can see that:

- Doggo is the stage having the highest average of retweets. Puppo is the next stage. Still we have pupper with the lowest average number of retweets.
- In this case, puppo and doggo are still the two stages with the highest averages. But, it's the other way around between them. Puppo is the one with the highest average number of likes. Doggo is the second in the rating. However, pupper is always at last.

5 Conclusion

These are small insights. I'm not so comfortable with the rating used by WeRateDogs because it's not that objective. I think I should take a closer look on the extracted ratings especially the denominator. Maybe there is a data quality issue there but I didn't focus on that.

In this project, I had the chance to start a basic application on data wrangling process. I worked on several techniques of data gathering, assessment, and

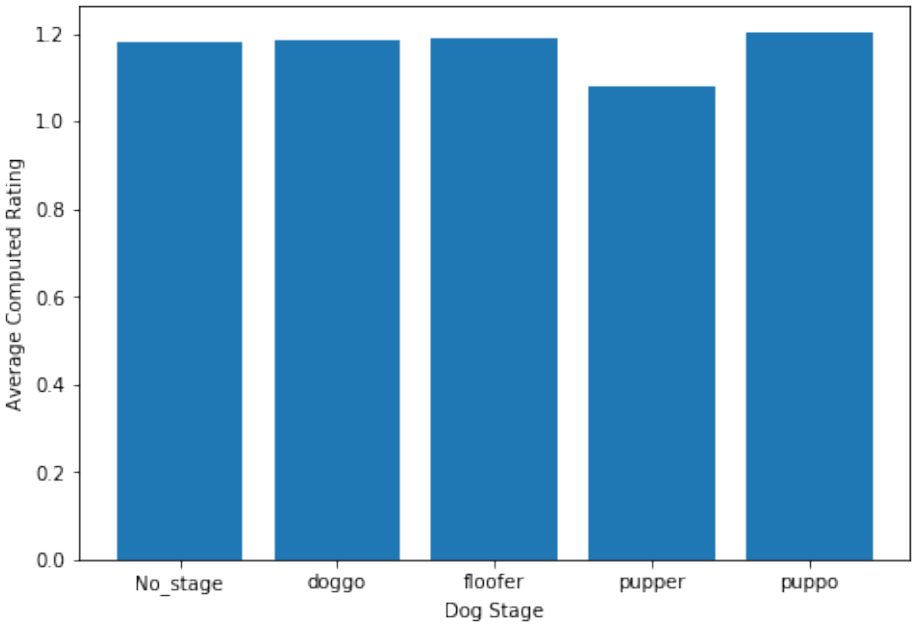

Figure 5: Bar-Plot of the Average Computed Rating per Dog Stage

cleaning. I had the chance to practice most of the new learned skills. And I'm excited to get to use them in real life applications.

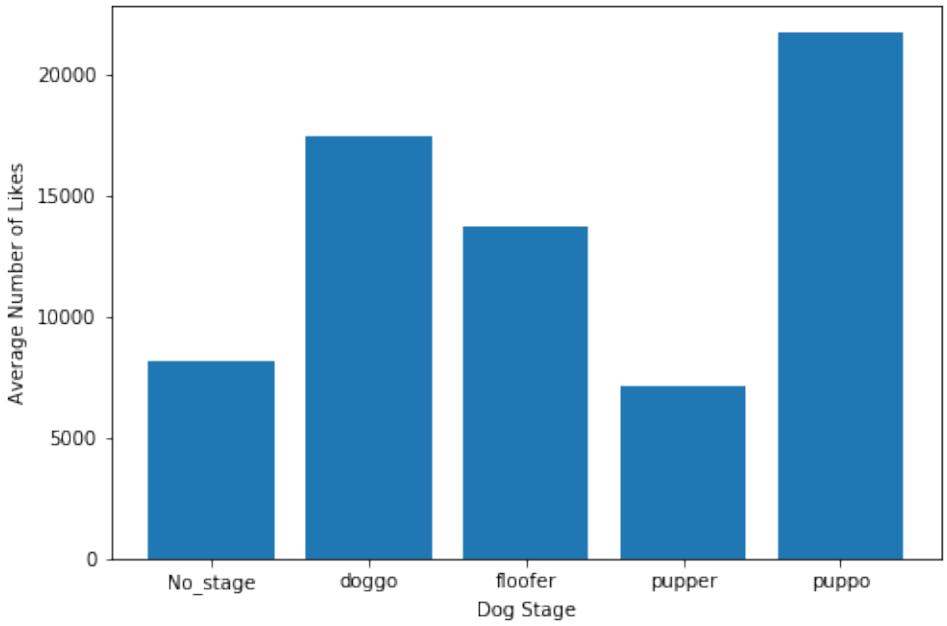

Figure 6: Bar-Plot of the Average Number of Likes per Dog Stage

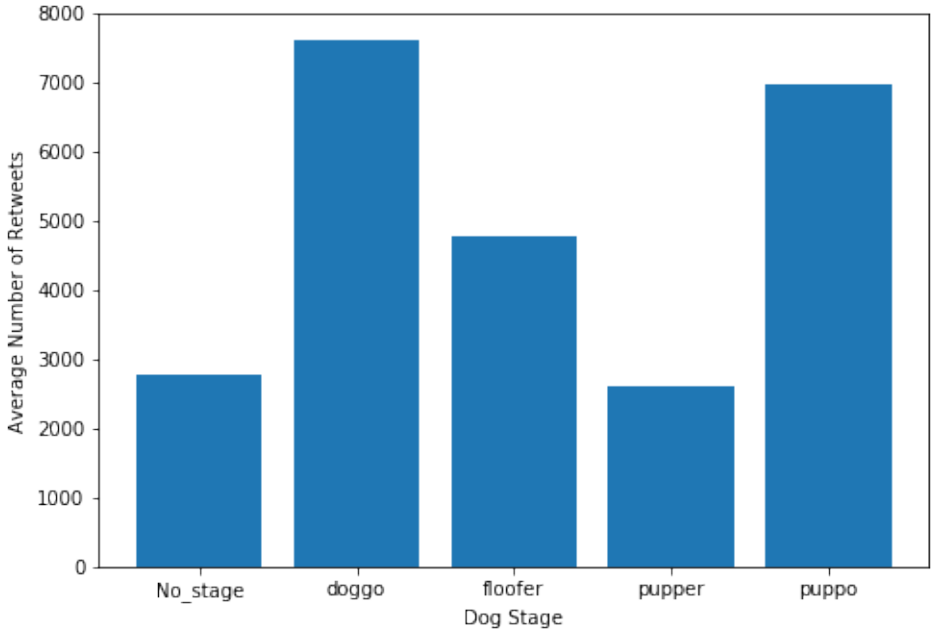

Figure 7: Bar-Plot of the Average Number of Retweets per Dog Stage