Wrangling is a crucial and perhaps the most tedious step in data analysis. In this project, we gathered data about a twitter page, called WeRateDogs, in multiple formats. The data was then assessed, cleaned, and analyzed. Some of our key findings and their discussion are summarized in this report.

<u>Observations and Insights</u>:

- Charlie is the most popular **dog name** observed. Cooper, Lucy and Oliver are tied for second place. Penny, Lola and Tucker are tied for third place. Winston and Bo are tied for fourth. Sadie is the fifth most common with no ties.
- There are four **source apps** found in the data, namely "Twitter for iPhone", "Vine - Make a Scene", "Twitter Web Client" and "TweetDeck". The most frequently observed app is "Twitter for iPhone" with a percentage occurrence of 94.26%.
- The dog **rating scores** are marked out of 10. But the scores can be higher than 10 because of the unique rating system followed by the page. 99.1% of the rating values are either below or equal to 20, with the max value being 1776. A bar chart for the quartiles of rating scores is shown in Fig. 1.

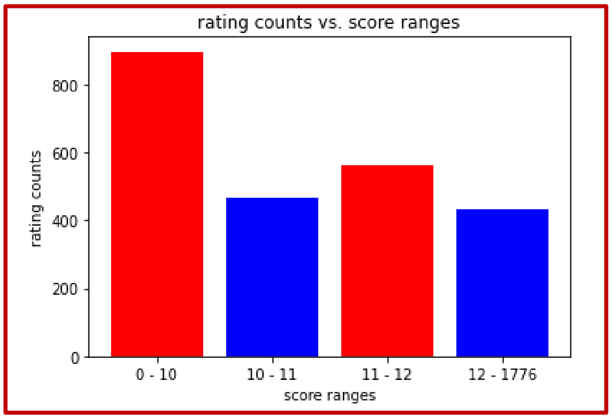

Figure 1: Bar chart for rating score quartiles

For purposes of analyzing plots and patterns we choose to classify rating values greater than 20 as outliers.

- A histogram for rating scores is presented in Fig. 2. It can be visibly observed that the ratings are not normally distributed. The distribution is **skewed to the left**.

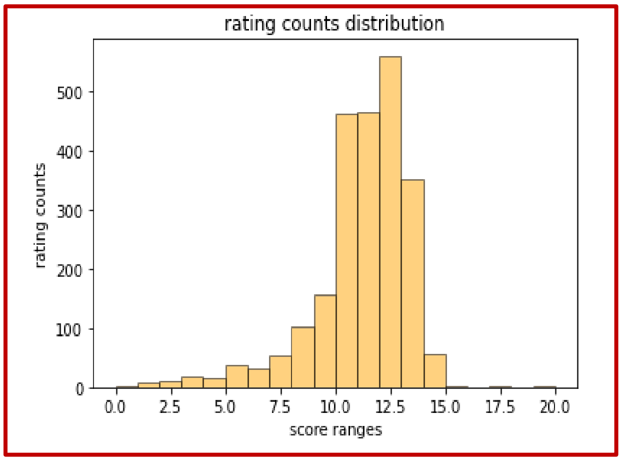

Figure 2: rating scores distribution

- Analysis was done to see the expected ratings for various **dog stages**. It was found that the "puppo" stage has the highest expected rating of 12.1. "Doggo" and "floofer" stages showed a nearly same value of around 11.8. "Pupper" stage has the lowest expectation of 10.8. In cases with more than one stages (doggo, puppo) combination has the highest expected rating of 13.
- The trends for the **retweet** and **favorite** count behavior vs. the ratings. are shown in Figs. 3 and 4. Both the retweet and favorite counts seem to exponentially increase with the rating score, eventually maxing out at a value close to 13 and then start to decline sharply. It seems like people start viewing scores as "unfair" after a certain point. We also looked at the data for outliers and found out that the almost all the retweet and favorite counts were below even than the average numbers.

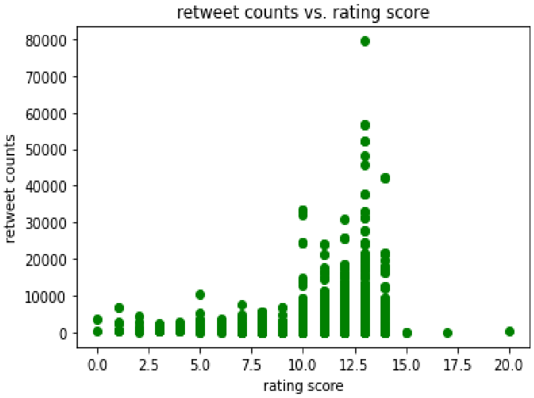

Figure 3: Retweet count vs. ratings

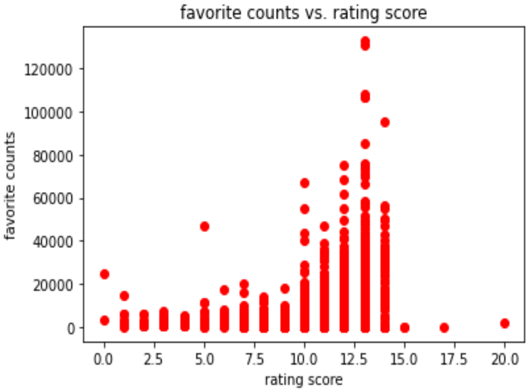

Figure 4: Favorite count vs. ratings

- The top five most popular breeds of dogs featured on WeRateDogs page are Golden Retriever, Labrador Retriever, Pembroke, Chihuahua, and Pug. The breed with the highest expected rating score is Clumber, with Soft Coated Wheaten Terrier and West Highland White Terrier at second and third places, respectively.
- A very interesting behavior is revealed for ratings and favorite counts with respect to the dog breed rarity. The plots can be seen in Figs. 4 and 5. In both cases, the curve smooths out as the rarity decreases; ratings become more consistent with respect to rarity as it decreases. Similarly, people seem to like a breed more consistently with respect to rarity. The more a breed is seen, easier it is for people to make up their minds.

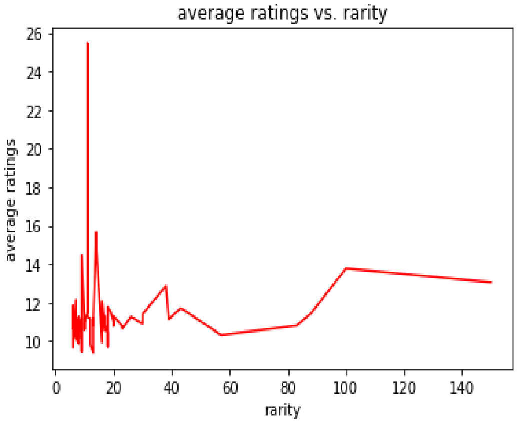

Figure 5: Mean retweet count vs. dog breed rarity

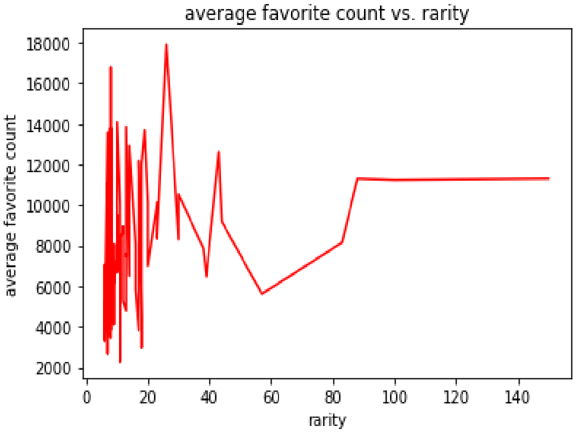

Figure 6: Mean favorite count vs. dog breed rarity

Limitations:

- There was some missing data that limited our ability to produce accurate analysis.
- For dog breeds we assumed that algorithm’s first choice “was the right prediction”.
- We only used basic statistical functions and probability theory for analysis. More sophisticated statistical techniques such as hypothesis testing, and regressions can be used.