Data Wrangling – Insights & Findings

By: Khaled Salem

Communicating insights

After implementing various data wrangling efforts and making partial amendments to downloaded files, some insights can be drawn such as being able to understand underlying relationships between variables (if any) and discovering patterns from WeRateDogs archive.

1. Most popular dog names

Once name column is cleaned and regular text is set to appear as None values, the most common dog names can be determined by examining the name column.

Table 1 – Common dog names

Dog name	Count
Charlie	11
Lucy	11
Oliver	10
Cooper	10
Tucker	9
Penny	9
Lola	8
Sadie	8
Winston	8
Toby	7

Table 1 show the top ten counts of most common dog names where the highest and therefore most common dog names is a tie between Charlie and Lucy.

2. Retweets and favorites

Extracted information obtained by querying Twitter API includes the counts of retweets and favorites pertaining to each tweet. Analyzing top ten favorite counts alongside top ten retweet counts gives indicates that the most retweeted tweet is also the most favorited. Moreover, there are seven entries shared by top ten tweets and top ten favorites.

Table 2 – Shared top 10 retweets and top 10 favorites

Index number	Tweet ID	Favorite count	Retweet count
824	744234799360020481	152176	75240
326	822872901745569793	129797	42367
418	807106840509214720	117508	54882
115	866450705531457537	113731	31995
60	879415818425184262	96996	39532
350	819004803107983360	86141	36411
1523	678399652199309312	77148	30390

An insight is gained from table 2 and that is there may be a correlation between the two variables; favorite count and retweet count. To substantiate this, a correlation heat map is created and displayed in figure 1 below which confirms primary observations made.

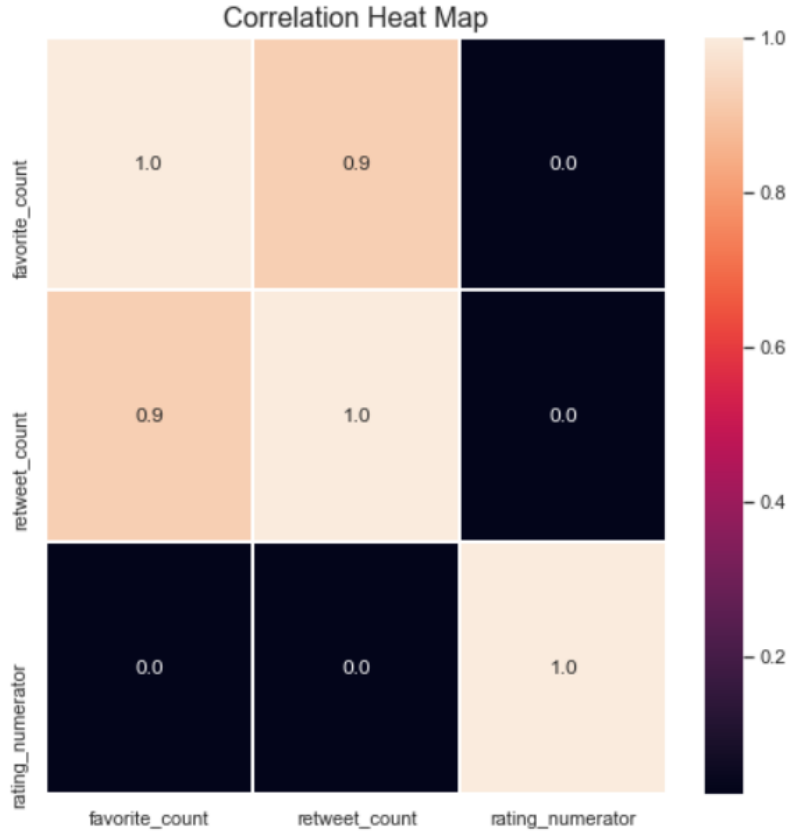

Figure 1: Correlation heat map

To visualize this correlation, a scatter plot of retweet counts versus favorite counts is plotted in figure 2 and shows a strong positive correlation.

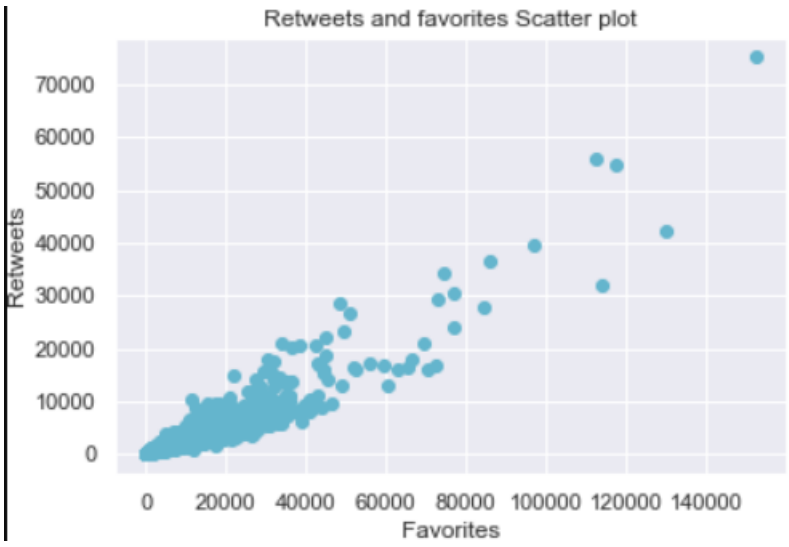

Figure 2: Scatter plot illustrating correlation

3. Most popular dog stages

Melting the four types of dog stages into a single column with each classification of puppo, floofer, doggo and pupper enables further analysis to be conducted. The most common dog stage classification was found to be pupper as portrayed by the horizontal bar chart in figure 3.

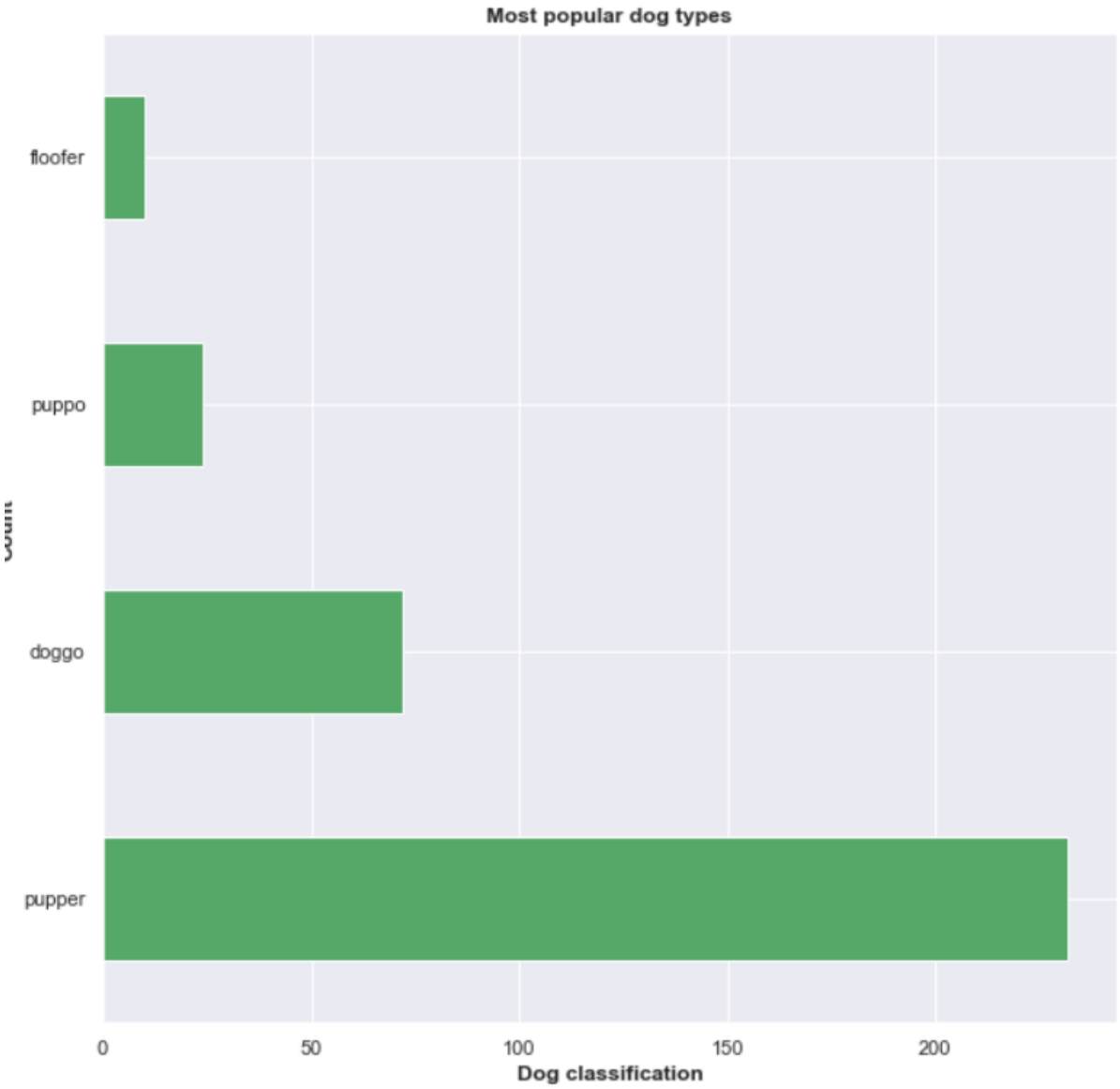

Figure 3: Horizontal bar chart of dog classifications

4. Tweet source

Upon extraction of relevant information from original source column that was stored in html format found in enhanced Twitter archive, an obvious discovery was made. The discovery made was that tweets originated from four sources only; twitter for iPhone, Vine – Make a scene, Twitter Web client and TweetDeck. Further visual analysis (bar chart in figure 4) made it possible to identify the most popular tweet source, which was made from Twitter for iPhone onto the social media platform.

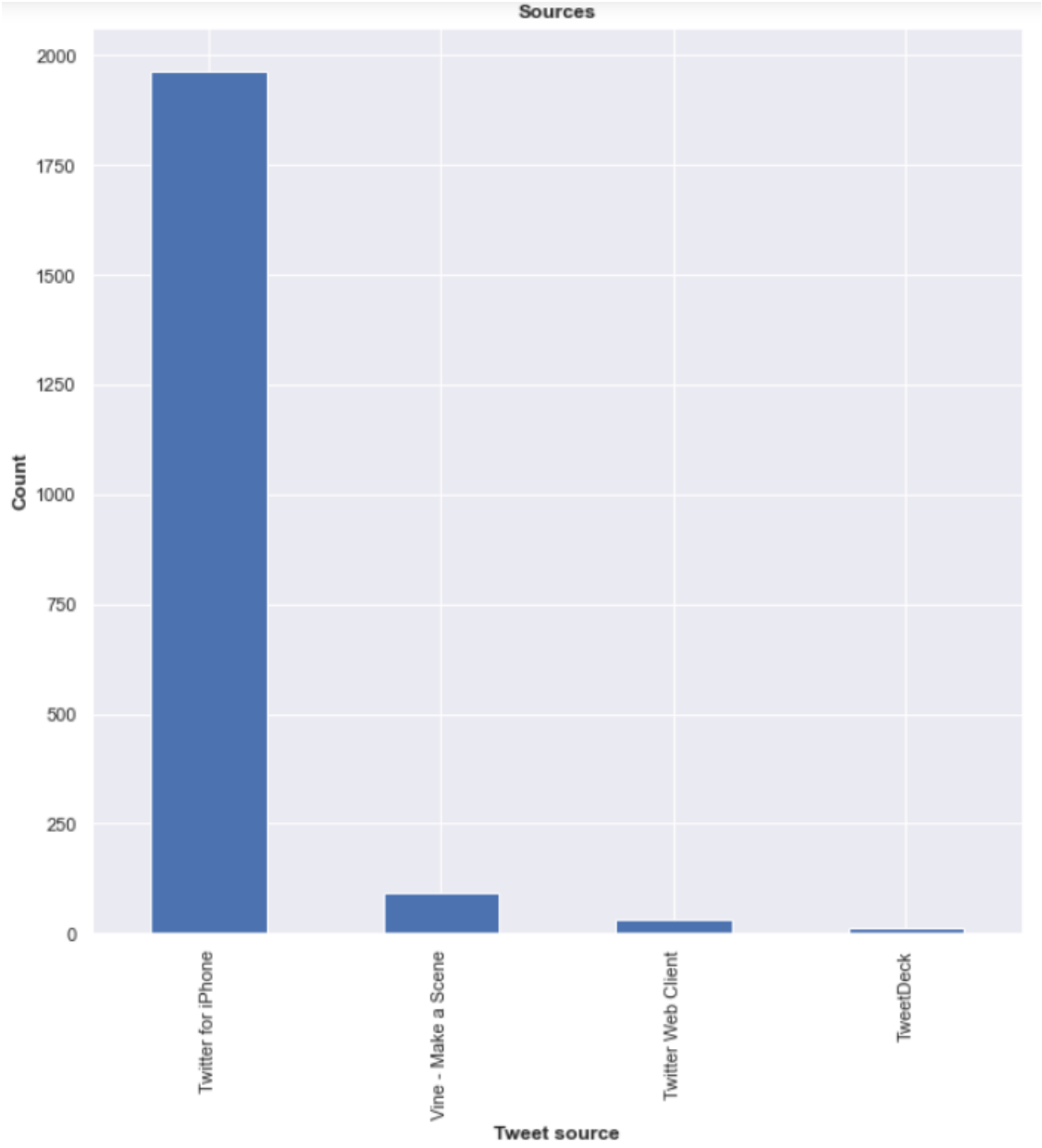

Figure 4: Bar chart of tweet sources

5. Numerator ratings

Numerical ratings made by users in tweets were benchmarked to a filtered denominator value of 10 to ensure consistency throughout. Additionally, numerator ratings were filtered to eliminate unrealistic numerator ratings (quality issue resolved exclusively for the sake of this analysis) by removing all values above the value of 14.

A histogram was plotted to display the distribution of (numerator) ratings among tweets, this is shown by figure 5. The graphical distribution represents clearly that ratings are extremely high with most tweets exceeding the denominator value of 10. A reasonable explanation behind those very high ratings would be that these people adore their dogs.

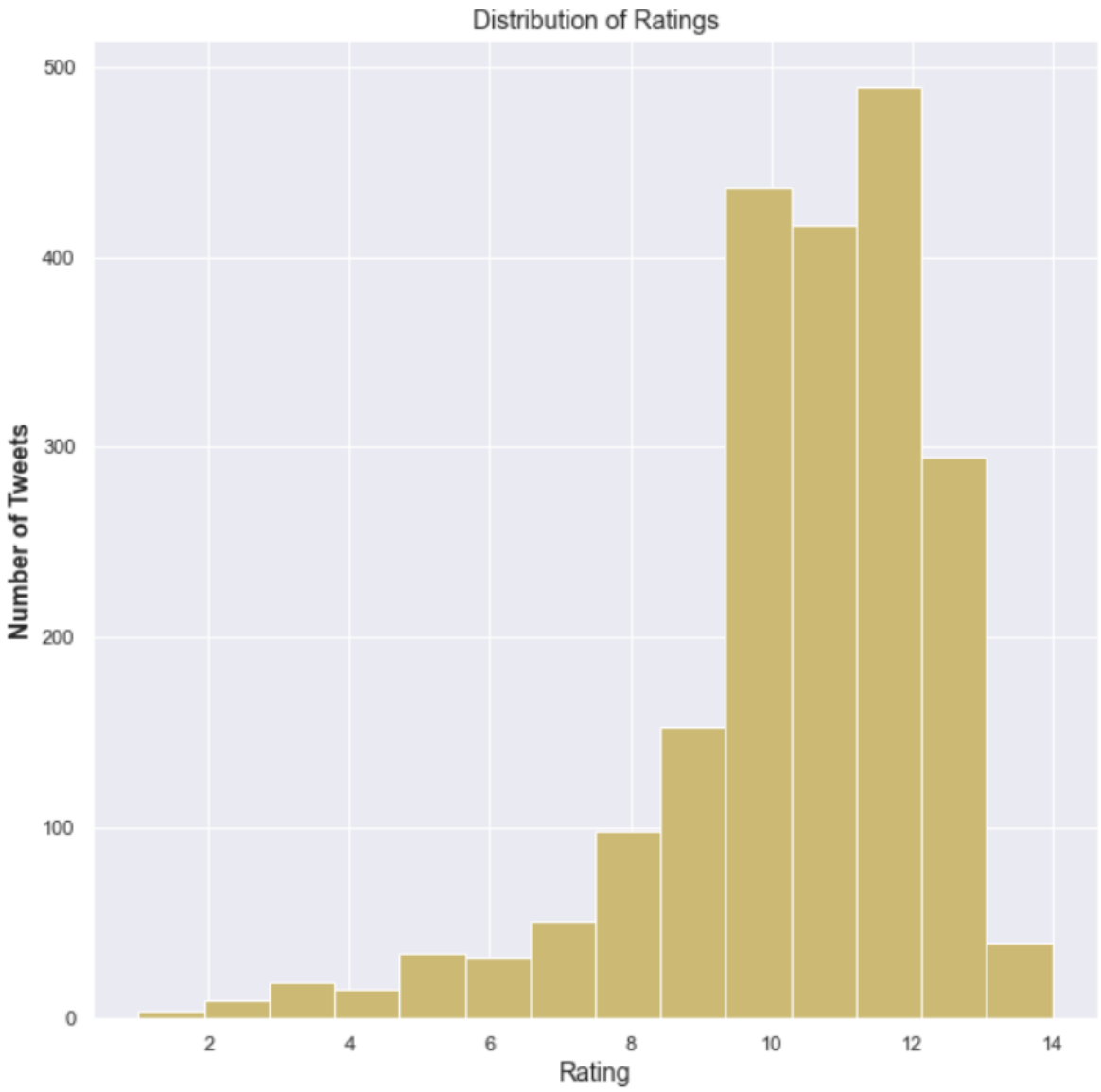

Figure 5: Distribution of ratings