Insights from 'WeRateDogs' tweets
by Michael Bong

OVERVIEW

In this project we wrangled multiple data sets relating to 'WeRateDogs' Twitter data and created a combined data set that we can use to analyse and derive insights. We focused on generating 3 insights and at least 1 visualization and explaining each in turn below.

It must be noted that the scope of the data set used to generate each insight are slightly different and will be specified for each insight.

INSIGHT 1: RATINGS ARE LENIENT

Each tweet has an associated rating (calculated as rating_numerator / rating_denominator). Normally, a typical rating scale would go from 0 to 1 (assuming a max of 1). In the case of 'WeRateDogs' tweets, ratings are very lenient. Although it is expected that ratings have a max score of 1, a large majority of tweets have a rating of greater than 1.

After removing records where ratings are outliers from the data set, and generating the box and histogram plot below, a few things become apparent.

Distribution of actual ratings
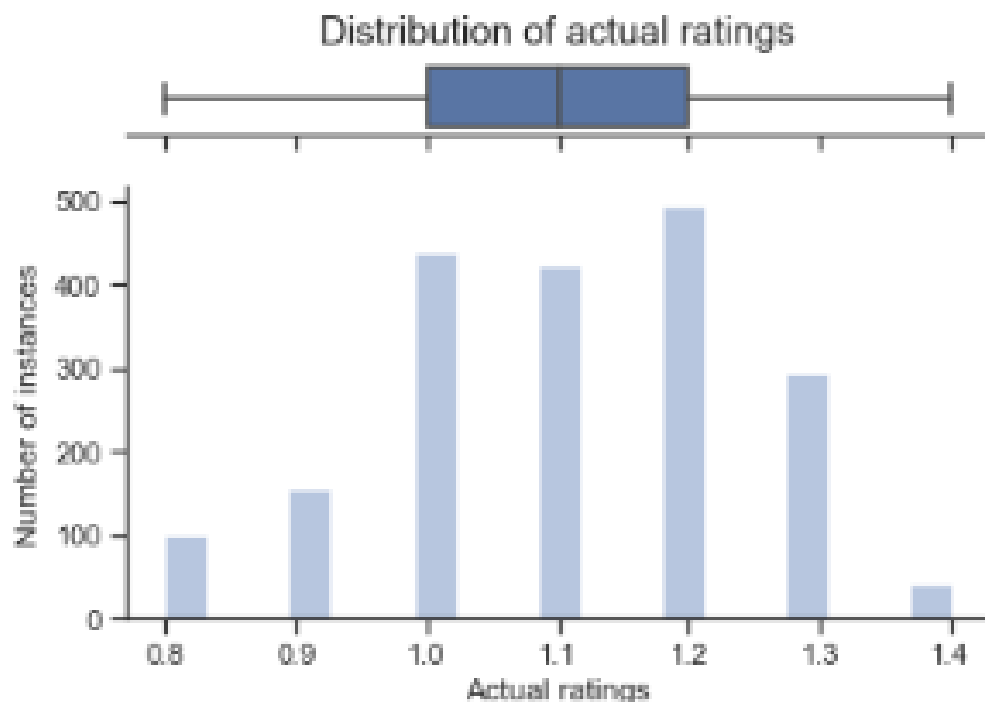
Approximately 75% of ratings are greater than 1.0 (ie 100/100). More than 50% are >= 1.1 and more than 25% of the ratings are >= 1.2. This shows that the ratings are lenient.

After all, who could possibly rate these awesome dogs any other way!

INSIGHT 2: RATINGS ARE CORRELATED WITH RETWEET AND FAVORITE COUNTS

Intuitively, one would expect that a tweet with higher ratings would drive a greater number of retweets and favorites.

We focused on studying whether there is a relationship between the average number of retweets and favorites for each rating, instead of looking at the sum/total of retweets and favorites for each rating.

After removing records where the ratings are outliers from the data set, a line chart was created below to see if a there are any relationships between ratings and retweet and favorite counts.
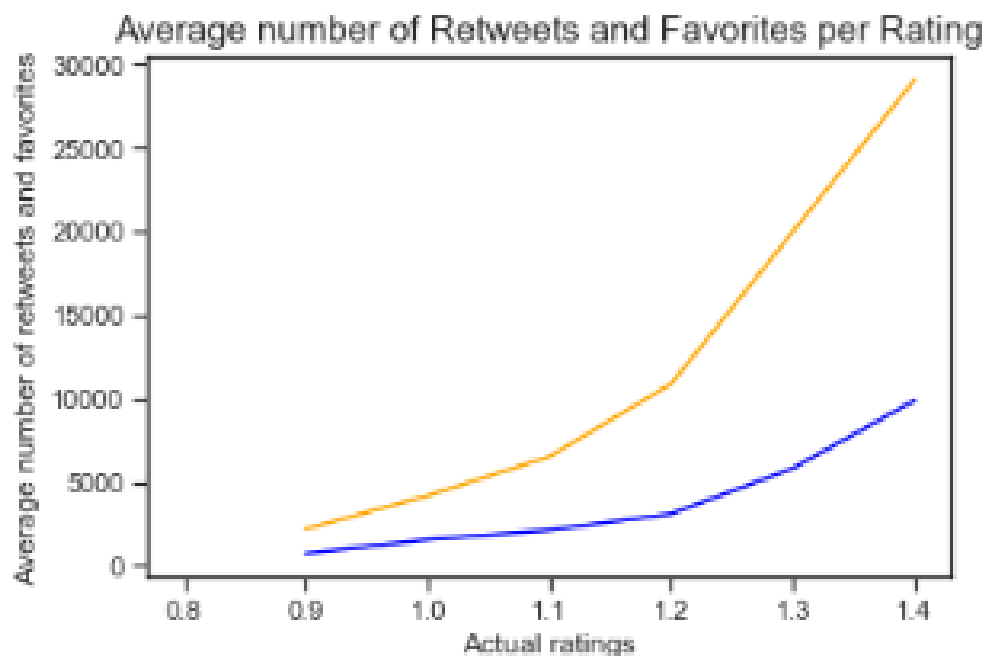

We find that there is a positive correlation between the ratings of a given post and the average number of retweets and favorites. the higher the ratings of a post, the higher the average number of retweets and favorites.

INSIGHT 3: NUMBER OF TWEETS DO NOT DETERMINE THE POPULARITY OF DOG TYPES

We sought to determine whether the most common dog type will also be the most popular dog type. We measure how common a dog type is by the number of tweets against each dog type. We measure how popular a dog type is by the average rating, retweets and favorites per tweet post against each dog type.

For this analysis, we limited the data set to only contain tweets where a dog type has been specified. A series of line charts are then created to help with our analysis.

Analysis of dog types
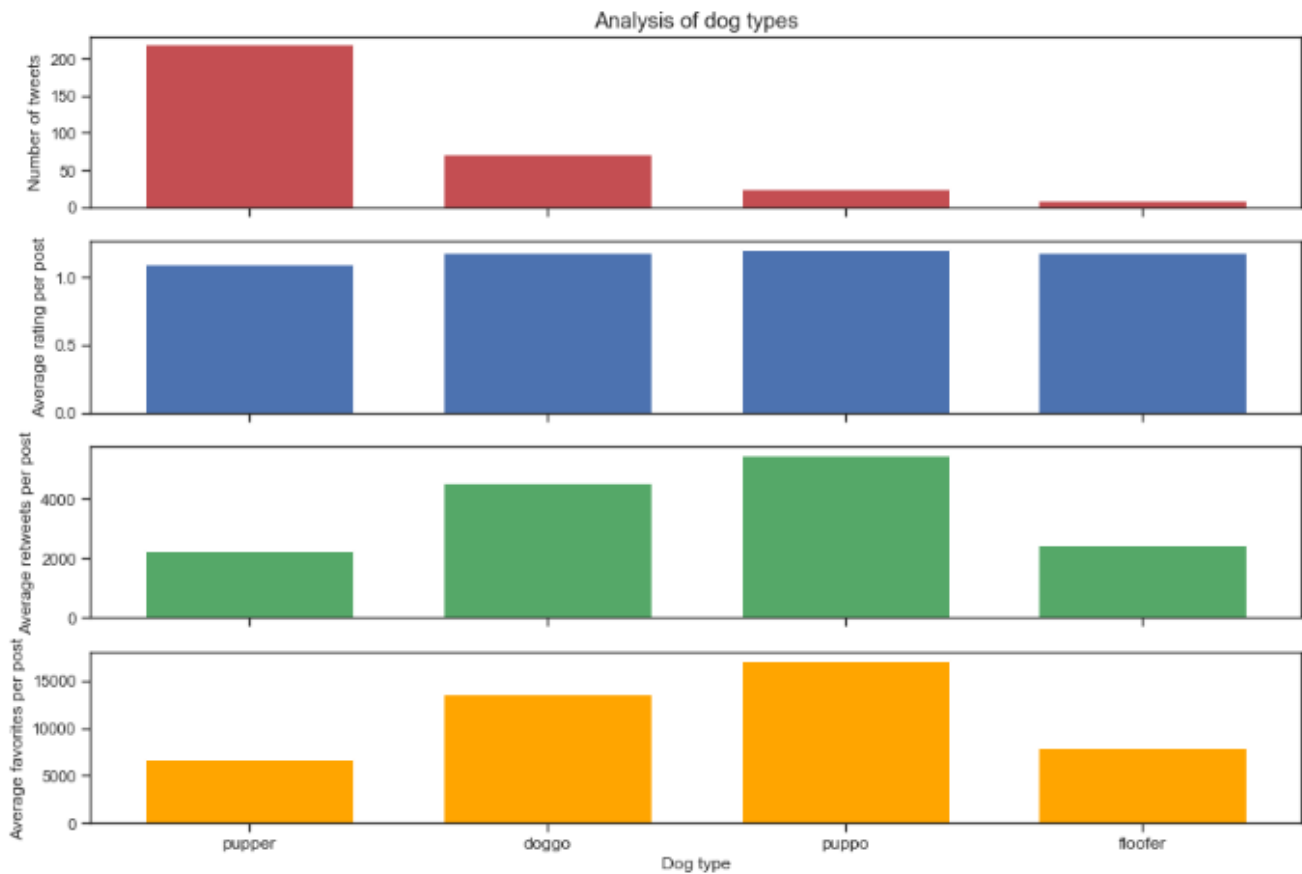
Pupper has the highest number of tweets, thus is the most common dog type. However, it has the lowest average rating, retweet count and favorite count per post, making it the least popular dog type. Puppo and Doggo have the highest average rating, retweet count and favorite count per post despite having low number of tweets.

Therefore, it is unlikely that the most common dog type will also be the most popular dog type.

CONCLUSION

These insights are only 3 of the many that exist and are waiting to be derived from this rich data set! These insights show that we have a high-level bearing of how ratings and retweets and favorites counts interact. They also show the nature of the rating system in the twitter feed.