Wrangling_report

The dataset is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that makes a humerous comment about a dog and provides a rating. The goal here is to wrangle the data and extract interesting insight from the data. There were three separate datasets to gather first: the basic tweet_data, image predication of dog breeds previously collected by a neural network classifier and other tweet-related information containing favorite and retweet counts. The data were both messey and dirty. We made them tidy by combining all three tables and structuring columns. We then cleaned the data for example correcting datatypes, correcting the tweets manually that contained more than one dog stage, and correcting rating numerator and denominator for non-standard values.

We asked three questions about this dataset: Which are the first 10 dog breed in terms of retweet, count, and rating? Is there any relationship between retweet count and favorite count? Is there any relationship between dog stages and rating? We observed the top 10 dog breeds in terms retweet_count, favorite_count, and rating. Among them, Saluki stands in the first place in terms of favorite_count and rating. Standard poodle is in the first place in terms of retweet_count. There is a positive relationship between retweet_count and favorite_count which suggests people have tendency to retweet the ones that are their favorite and vice versa. Puppo has the highest rating and pupper has the lowest.