

# wrangle\_report

December 16, 2018

Real-world data can be disorderly and unclear, making analyses difficult. In this project, I have collected tweet data and details for the twitter archive of the user @dog\_rates, also known as WeRateDogs. WeRateDogs rates dogs with some humorous content about them. The ratings are out of 10, but these fun dogs all manage to score above 10. Since the twitter data consists primarily of tweets, additional data including favorite/retweet count data, and image predictions to identify dog breeds, were collected by querying Twitter's API and from alternate sources.

**Data gathering:** Sources: The data were collected from three sources. The basic tweet texts from 2015 to 2017, along with the information gleaned from it including the dog's name, stage and ratings were downloaded from the Udacity server as a csv file and saved as twitter-archive-enhanced.csv, and loaded into a dataframe using Pandas. Additional data were obtained by querying Twitter's API using the Tweepy library after creating a developer account, using the associated secure keys. The tweet\_id indicated json files were downloaded and stored in a text file tweet\_json.txt, and parsed to extract the favorite and retweet counts into a dataframe. Image-based breed predictions for the dogs, openly available from a url was downloaded as a tsv (image-predictions.tsv) and loaded into a third dataframe.

**Data wrangling:** Upon assessment using the Pandas and Numpy functions, the following data tidiness and quality issues were identified. The prospective solutions are noted alongside.

Tidiness: 1. Type of dogs split into 4 columns in df\_archive; join into one 2. jpg\_url and img\_num columns in df\_impreds are not necessary for the planned analyses; remove these columns 3. df\_addl should be part of df\_archive; merge to the archive 4. Add df\_impreds also to the df\_archive for easier analyses

Quality: 1. in\_reply\_to\_status\_id and in\_reply\_to\_user\_id columns are reply tweets; remove rows 2. retweeted\_status\_id, retweeted\_user\_id and retweeted\_time\_stamps columns represent retweets; remove rows 3. Timestamp as string; change to datetime format 4. Source column has 4 unique values which represent the actual sources, lost in the html format; Pick out the source from the html link 5. expanded\_urls column doesn't make sense; remove 6. Rating denominator should be 10 for all; change to 10 7. Rating numerator has low and high values; begin at 11, cap at 15 8. Datatype of tweet\_id is int; change to string

**Data Analyses:** After merging the three dataframes and fixing the cleanliness issues, it was seen that there were 2356 instances and 19 attributes. These were the tweet\_id, timestamp, source, text, rating\_numerator, rating\_denominator, name, dog\_type, favorite\_count, retweet\_count and the 9 image prediction details columns. Of the 4 sources used to render the images, iPhone, Vine, Web\_client and Tweetdeck, iPhone was found to be the most frequent, with 2221 source listings.

There were 690, 1183 and 483 tweets for the years 2015, 2016 and 2017 respectively. The maximum number of tweets during 2016 is expected since the tweets were collected from November 2015 to August 2017. From the information contained in the texts, the dogs were classified into dog stages, ranging from pupper to doggo to floofer, and alternative intermediate or composite stages. The pupper was seen to be the most frequent stage among the classifiable dogs (from the tweet texts) of the categories encountered viz. 'pupper', 'doggo', 'puppo', 'doggopupper', 'floofer', 'doggopuppo' and 'doggofloofer'. The highest favorite count was 140764, for an unnamed puppo, apparently a super supportive one participating in the Women's march in Toronto. The highest retweet count was 61150, for Stephan, a helpful dog predicted to be a Chihuahua with 50% confidence level. A scatter plot analysis showed that there was a strong correlation between the favorite and retweet counts for the dogs. The dogs received ratings between 11 and 15. The most frequent ratings were in the range of 11-12, numbering around 900. The average confidence levels for dog predictions based on their images were 60% for their first prediction, and lower for the subsequent two breed predictions.