# Wrangle Report

**By Natalie Urban**

**Data Gathering**

The data for this report was collected from three different data files with three different methods. The main file with most of the data on WeRateDogs tweets was the twitter archive that was downloaded using panda's csv reading function. Then the image predictions file, which holds predictions of the dog breeds from the WeRateDogs tweets, was downloaded using the requests package and written to the operating system. Finally the third file, which contained twitter API data on WeRateDogs tweets, was collected through the tweepy package by looping through and writing each tweet id match to a file. Then the file was queried to collect the tweet ids, retweet counts, and favorite counts.

**Assessing Data**

The three files were each assessed for their quality and tidiness. The dog tweet archive table had some nondescriptive columns that needed to be more descriptive of what the columns contained, the dog's name column contained incorrect dog names like 'the', incorrect rating thresholds, the data type of the column timestamp (renamed to 'date') was incorrect and needed to be datetime, retweets and reply tweets were present, but not needed. In the retweet and like table there was a nondescriptive column id that needed to be more specific. The dog image predictions table had multiple nondescriptive columns that needed to be fixed and one row that only compared the images to food, so the prediction was unnecessary. The dog image prediction table also had all columns combined into one and three of the dog type columns needed to be combined into one since the data in each were all directly related. All three tables also needed to be combined into one.

**Data Cleaning**

Each piece of dirty and untidy data was cleaned using the process of defining how the problem would be cleaned, coding the problem, then testing the result. Nondescriptive columns were renamed to be more specific, rows with incorrect values were changed to None, incorrect rating threshold tweets were corrected, incorrect data types were changed to

appropriate types, unnecessary rows and columns were dropped, columns were tidied into separate columns for each value, the 'dog type' column was created and the three columns that made it up deleted, and finally, the three tables were combined on 'tweet id' to form one clean table.

**Data Storing**

Once the data had been cleaned and tidied into one main dataset, it was stored into a file. The final cleaned master dataset was saved to a csv file called 'twitter archive master' by using pandas 'to csv' function.

**Analyzing and Visualizing Data**

The data was analyzed, and insights were discovered through analyzing the dataset with three questions. The first question was: what is the difference in percentage of ratings over the threshold and percentage of ratings under the threshold? It was discovered that 57.92% of ratings were over 10 and 20.91% were under 10. The second question was: which dog name was the most popular? It was discovered that besides WeRateDogs tweets providing no name being the highest name, the name Charlie was the most popular dog name. The last question analyzed was: what is the difference between likes and retweets from WeRateDogs tweets? It was found that the total WeRateDogs retweets were 5,479,297 and their total likes were 17,599,222. With this analyzed data, I created a pie chart for visualization on the most popular dog name and how it compared to the other dog names. The visualization showed that there was a large variety in dog names with None exceeding in number all other names.