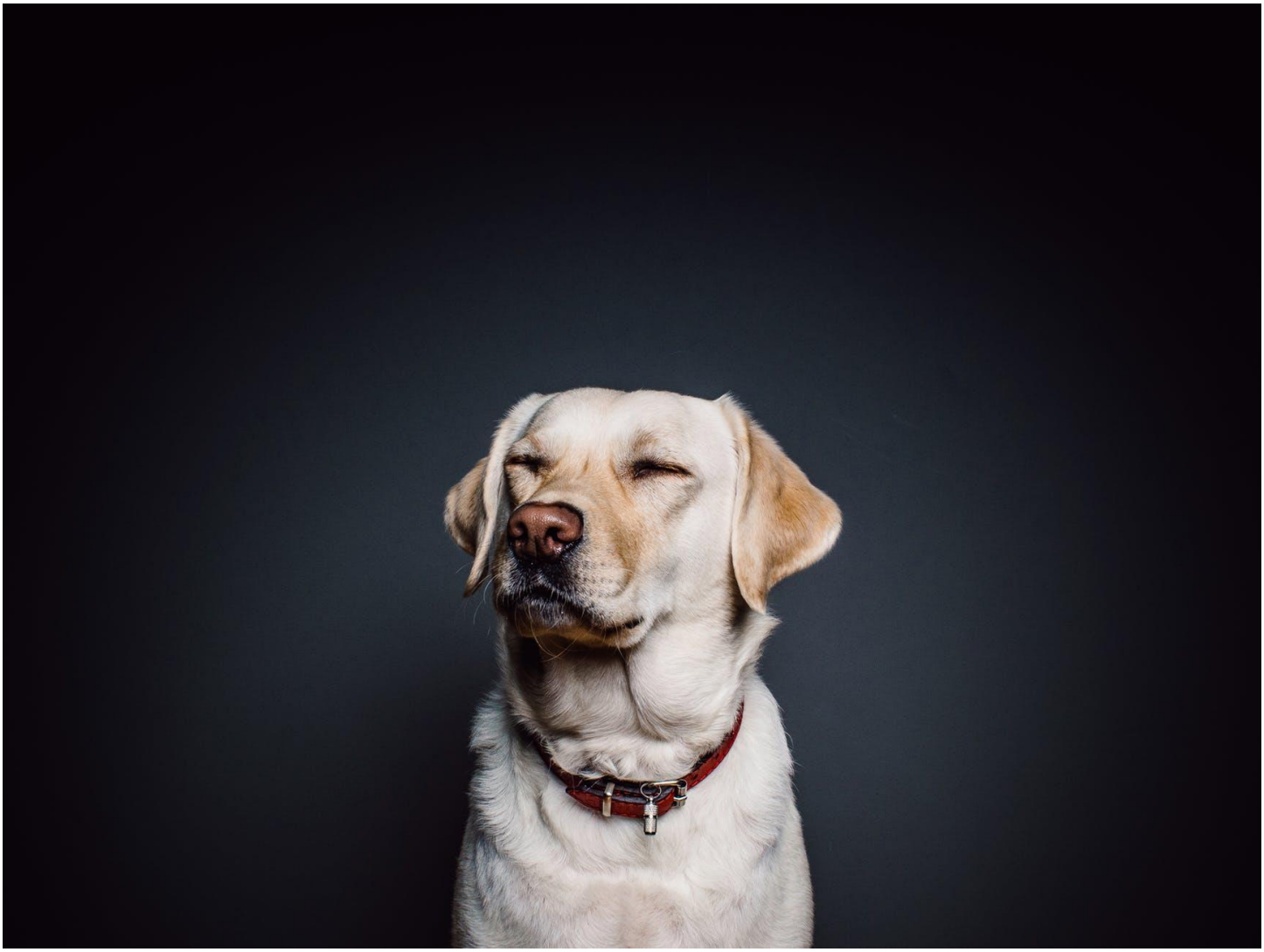
We Rate Dogs | Data Wrangling

Sarah Aranda



In this project, I wrangled Twitter Data from WeRateDogs to make trustworthy analyses and visualizations of the information. WeRateDogs (https://en.wikipedia.org/wiki/WeRateDogs) is a Twitter account that posts, rates dog photos, and provides comical descriptions of them.

There are three general stages to this wrangling project: gathering, assessing, and cleaning.

Gathering of the information involved three datasets that would later be merged to create a single data frame. The first one contained basic information from the Twitter archives such as tweet text, rating, favorite and retweet counts. The second dataset contained image predictions of the dog breed and corresponding information that were derived using a neural network. The third dataset was gathered by querying Twitter's API. The tweet IDs from the first dataset were used to get each tweet's JSON data using the Tweepy library. The information was later stored in a text file and read into a Pandas Data Frame.

In next stage of the project was assessment. This involved both visual and programmatic evaluation of the data for quality and tidiness issues. The majority of the problems were ones of quality and were often found in the information obtained from the Twitter archive. Quality issues include problems with the data content itself

whereas tidiness issues are related to the structure. Only a few tidiness issues were uncovered. Each dataset was assessed individually, and a list of problems were compiled for cleaning at the next stage.

In the cleaning stage of the assignment, I systematically went through the list of problems previously defined. Before attempts at cleaning were made, a description of how each of the problems would be cleaned was described. Afterwards cleaning at each step, a check was done to ensure it was done properly.

After all of this, the three datasets were merged into a single data frame labeled “twitter_archive_master.csv”.