Report

I were provided tweet detail data from Twitter for WeRateDogs page. This page is created so that people can show their appreciation for dogs and upload dog pictures. The archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

Once the data was loaded from various sources, it was evident that cleaning needs to performed on the dataset before analysis.

Since the data was present in 3 different tables. Wrangling the data happened in 3 sections:

1. Enhanced Twitter Archive: Since the archive consisted of tweets and retweets, the first step was to remove the retweets since they would be redundant data and might provide incorrect data points. Posts of dog pictures are generally rated by the post owner and its generally above 10 but out of 10. (Ex: 13/10 rating) Some of the numerator values were clearly out the range and some of the denominators were not 10. Once that was fixed, I removed the rows that contained retweet data. It was a possibility that were either duplicate rows in the database or just repeated tweet ids. On checking for the duplicates, I found none. The project administrator mentions that he has used ML/AI to identify names of the dogs from the text of the post that the owners made. Some of the names were not identified, but there were also some of the names that were incorrectly identified such as values like a,an,then, quite etc. It was important to check whether all the tweets had the right tweet id format. Each of them did have equal number of characters. A lot of times data cells starting with 0s are cut off in excel.
2. In one of the tables there also exists multiple columns for string id. I made a comparison to check the number of rows having the same string id to understand the significance of both fields.
3. There were the tweet URL links in the Text column of the twitter archived file which can be split into two different columns or just removed entirely since the link is already present.

These were the general quality and tidiness wrangling I could come up with on visually and programmatically assessing the dataset.