Wrangle Report – WeRateDogs Project

By Jordan Mazza

While visually and programmatically assessing the data, the first thing I noticed was the dataset included retweets and replies as well as original tweets and had unnecessary columns to depict this data. I removed this data by saving the rows that had null values in the columns for retweet and reply status id. I also noticed many rows had denominators other than 10, as well as very large and very small numerators. After filtering through this data, I found it unnecessary to remove the strange variables in the numerator column for two reasons, the first being some tweets included many dogs and so the numerator rating was incredibly high. The second reason is this account's rating system developed over time to what it is today, so some of the early tweets included a standard rating system out of ten. For both of these instances, the data isn't erroneous and therefore, I found no need to remove it.

For the denominator, however, there were only 17 rows with strange data, so I went through them manually and found some of the rating incorrectly interpreted from the text. For example, one tweet listed the date 7/11 in the text, and the denominator rating was 11. For rows with similar issues, I removed them using the .drop() function to drop their specific indices. I then used the .astype() function to change the column object types that needed to be changed in each dataframe. I also noticed during my assessing that some of the missing values in the name column were represented as 'a' while others were represented as 'None'. For uniformity's sake, I replaced all the rows 'a' with 'None' using the .replace() function.

Finished with the quality issues for this dataframe, I moved to the tidiness issues, firstly being the irrelevant columns. I removed all the columns (five in total) relating to retweets and replies using the .drop() function. Finally, to solve the tidiness problem of the dog stages being in four separate columns, I found a solution from a mentor in response to a Udacity Knowledge question posted. In this solution, I first removed the text in all the rows with 'None' values in all four columns using the .replace() function. Then, I created a new column by combining all the dog stage columns. I used the str.replace() function again to fix the formatting on the rows with multiple dog stages. After checking to ensure this method worked by querying the old dog stage columns to make sure they matched the new column, I dropped the old columns.

For the twitter image dataframe, I used the .str.replace() function to replace the underscores with spaces in the p1, p2, and p3 columns. The only other issue I found with this dataframe was the fact there was only 2075 rows in this dataframe, where there were 2356 rows in the twitter archive dataframe. I found a similar issue with the df_tweet dataframe, where there were only 2331 rows; however, we do not have access to retrieve any missing data for this project, so there is not anything I can do about this missing data for this project.

I also noticed in the df_tweet dataframe there were some tweets with low retweet and favorite counts. Given the popularity of this account, I was concerned some retweets and replies were in this dataframe, so I wanted to investigate more. After querying a few tweets in the twitter archive dataframe, its seemed tweets with low retweet and favorite counts were from when the twitter page first began in 2015. Therefore, this data was not problematic data and no change needed to be made. Now that all the dataframe have been assessed and cleaned, I combined them all to create one master dataframe for my analyses and stored this new dataframe to a CSV file.