wrangle_report

January 14, 2023

0.1 Reporting: wragle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

I began my data wrangling efforts by first gathering all the necessary data. I did this by creating three dataframes using Pandas. The first dataframe I created was from a csv file created using data from the twitter archive. This file included information about each tweet from the We Rate Dogs (@dog_rates) twitter account. Information such as the tweet_id, source, text, dog name/stage, and url were included in this file. The next dataframe was used to predict the breed of each dog based on the image in the tweet. I downloaded this from Udacity's servers. The last dataframe had each tweet and the favorite and retweet count. For this dataframe I had to get the data from Twitter's API.

After gathering the data and creating the three dataframes in Pandas, I assessed the dataframes both visually and programatically to find quality and tidiness issues, of which there were many. After spotting these issues I needed to take action and use code to clean all three dataframes. My strategy was first identifying that the twitter archive dataframe, which had all the information of each tweet, was the central dataframe. The other dataframes contained information that would ultimately be joined to this dataframe. Next, I began cleaning each dataframe of the quality issues to make the data consistent and accurate. Most of this was done on the central dataframe. After cleaning the dataframes of quality issues I removed the "pupper" "puppo" "floofer" and "doggo" columns from the twitter archive dataframe, because these all represented the same variable, Dog Stage, and thus should not have been seperate columns. Classic data tidiness issue! Following this step I LEFT JOINED the retweet/favorite count dataframe and the image prediction dataframe to the twitter archive dataframe to create one master dataframe. Once I had this master dataframe it was time to analyze the dataframe and generate some insights.

In [ ]: