Gather

I gathered data from three sources: the WeRateDogs Twitter Archive CSV, the Twitter API, and the image prediction neural network hosted by Udacity. I used Pandas read_csv() to create a dataframe (df) from the WeRateDogs Twitter Archive. From the Twitter API, I collected each tweet’s JSON and saved them as lines in a .txt file, then created a dataframe (tweet_selected_attr) with tweet id, favorites, retweets, and timestamps from the .txt file. I then used requests to get the image predictions tsv file from Udacity’s servers and used read_csv() to create a third dataframe (image_prediction).

Assess

For the WeRateDogs Twitter Archive (df), my visual assessment showed that there are Reply and Retweet rows in the dataframe. This project is focused on original tweets, so I determined the Reply and Retweet entries needed to be removed to make sure they do not skew the analysis. I also observed that the word “None” was written in the columns “doggo,” “floofer,” “pupper,” and “puppo” - which meant there were many rows for which a dog stage was not assigned, but programmatic assessment would miss this because the entries have “None” typed in (as opposed to a blank entry). Further visual assessment in google sheets revealed that some numerator ratings were decimals and not extracting correctly. Furthermore, there were 2 tidiness issues visible: dog_stage should be in one column (as opposed to separate columns for “doggo,” “floofer,” “pupper,” and “puppo”) and numerator_rating and denominator_rating should be in one “score” column. Programmatic assessment also revealed long urls as source entries (found using value_counts(), and incorrect datatypes for source, tweet_id, numerator_rating, and denominator_rating (found using info()). Using sample() also revealed a nonsensical name, which upon further investigation using str.contains() and value_counts() revealed more nonsensical names, all starting with lowercase letters.

For the Tweet Image Predictions (image_prediction), my visual assessment revealed that the dataframe had rows for each tweet, and from a tidiness perspective, it would make the most sense to combine it with the other dataframes to make one combined dataframe where tweets are the unit of analysis. Programmatic assessment using info() showed that the predicted breeds were object datatypes rather than category datatypes.

For the dataframe tweet_selected_attr, the visual assessment revealed the same need to combine it with the other dataframes. Programmatic assessment using info() showed that retweets and favorites were integer datatypes when they should be floats.

Clean

I first addressed missing data by removing retweet and reply rows using isnull() and drop(). Then I combined the WeRateDogs Twitter Archive dataframe (df) with the Image Prediction dataframe (image_prediction) and the tweet_Selected_attr dataframe using Pandas merge() to produce one combined dataframe that only included rows with image predictions, retweets, and favorites.

I then addressed tidiness issues, starting with changing the “None” entries for “doggo,” “floofer,” “pupper,” and “puppo” to blanks using replace(), creating a new dog_stage column combining entries from those columns and replacing blanks with NaN using Numpy’s NaN method, and dropping the old separated columns. Afterwards, I collected the correct numerator ratings using str.extract(),changed the numerator and denominator rating datatypes to float using astype(), and combined numerator_rating and denominator_rating into a new column, dropping the separate numerator and denominator columns.

Then I addressed quality issues, including making the most likely prediction a category datatype using astype() (and dropping the other prediction columns not being used), and making the tweet source a category and easier to read using replace() and astype(), making the tweet_id an object using astype(), making retweets and favorites integers, and removing rows with nonsensical names using index and .str.islower().