Wrangle Report

1. Gathering Data

a. I first gathered the Archive dataset by downloading it directly from Udacity. This was given in a CSV file, and I uploaded it into the Jupyter Notebook via the pd.read_csv function. I named the file “archive”.

b. The second data set was downloaded from the Udacity servers programmatically. I downloaded it using the requests library, created a file for the dataset to be stored, I saved it in that file, and I uploaded it to the Jupyer Notebook by using the pd.read_table function. I named the file “images”.

c. The third dataset was gathered as JSON data via the Twitter API using the Tweepy library. The JSON data was saved line by line in a text file. I named the file “api_df”.

2. Assessing Data

a. I assessed the data visually by opening up each file in a Pandas DataFrame.

b. I also assessed the data programmatically by using the .info(), .value_counts(), and other functions on a few of the different variables.

c. I was able to identify the following issues that needed to be addressed:

i. Tidiness Issues:

1. Merge 3 DataFrames.
2. Drop unnecessary columns.
3. Combine "doggo, fluffer, pupper, and puppo" columns.

ii. Cleanliness Issues:

1. Delete retweets.
2. Timestap column in archive dataset needs to be changed to a datetime data type.
3. Some of the animals listed in the p_1, p_2, and p_3 columns are not dogs.
4. Denominator values in the archives dataset are not all 10.
5. Dog names of "a", "an", "the", etc.
6. Change tweet_id to a string instead of an integer.
7. Some numerator values appear to be off.
8. Change the tweet_id data type to a string.

3. Cleaning Data

a. To start the cleaning process, I first merged the 3 datasets together on the "tweet_id" columns. I then copied the dataframe and started solving the steps listed above. I consulted several sources online, and I was able to complete each step listed above.

4. Storing and Acting on Wrangled Data

a. I saved the dataset as a CSV file using the pd.to_csv() function.

b. I then created 3 visualizations to and wrote down some insights into the data.