Data Wrangling Report

By Chase Eby

Introduction:

The art of data wrangling is a skill that every data analyst needs to possess. Without the ability to wrangle and clean data, analysts wouldn't be able to provide as accurate insight into many real-world problems. The three main steps to data wrangling include the following.

- Gathering
- Assessing
- Cleaning

Gathering:

We gathered the data from three different sources. Two of which were provided by Udacity and were fairly easy.

We imported the 'twitter-archive-enhanced.csv' into our dataset using Panda's 'read_csv('')' function.

The next dataframe we imported was the 'image_predictions.tsv' that we gathered programmatically using the Request library.

The last dataframe 'tweet_data' was gathered by using a loop and the twitter api to programmatically query the tweets JSON.

Assessing:

After gathering the necessary data, we have to assess it and see what needs to be cleaned and tidied. I took the following steps to identify what needed to be cleaned.

Quality issues in the datasets.

Twitter archive dataset¶

- 1: Variables timestamp and retweeted_status_timestamp should be a datetime instead of String: Convert in cleaning process.
- 2: Variables in_reply_status_id, in_reply_to_user_id, retweeted_status_user_id, retweet_status_id: These variables should be integers instead of floats.
- 3: Name column has no missing data or incomplete data example: 'none'.
- 4: Many Nulls represented as 'None' in the columns 'doggo', 'floofer', 'pupper', 'puppo'
- 5: Rating_numerator and rating_denominator variables contain numerous errors.

- 6: Delete all unnecessary columns that won't be used
- 7: Remove retweets

image_prediction dataset

- 7: We have missing values from the prediction datase 2075 observations in the image prediction dataset and 2356 from the Twitter_archive dataset.
- 8: Split image prediction and confidence levels into their own columns

Tidiness Issues

- We need to combine the last 4 variables 'doggo', 'floofer', 'pupper', 'puppo'
- We need to clean up tweet_id into str and merge the dataframes

CLEANING:

The final process that we needed to do on the data is cleaning all of the issues found in the assessing the quality of the datasets. After cleaning the datasets, we merged them into the final cleaned dataset 'twitter-archive-master.csv'