Data Wrangling Report

About the Dataset(s)

The dataset I'll be working with is the archive of tweets from WeRateDogs, also known as Twitter user @dog rates (https://twitter.com/dog rates). 2356 basic tweets from November 2015 to August 2017 comprise this archive/dataset. WeRateDogs is a Twitter account that rates users' pets along with a lighthearted comment about the dog.

1. Gathering Data

WeRateDogs Twitter archive

I manually downloaded the WeRateDogs Twitter archive using the link provided by Udacity as Twitter archive enhanced.csv, and then I imported this file into a pandas DataFrame called arc_df

Tweet image Prediction

I used Python's Requests package and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad image-predictions/image-predictions.tsv to programmatically download the tweet image predictions file stored on Udacity's servers. I saved it locally to image_predictions.tsv file. Then, I imported this file into a Python Pandas DataFrame called image_df

Additional Data from Twitter API

I got the complete set of JSON data for each tweet via the Twitter API using the tweet IDs from the Twitter archive, and I saved it all in a file called tweet json.txt. Only the tweet id, retweet count, and favorite count were included in the DataFrame tweet df I created from this JSON.

2. Assessing Data

Assessing of data was done via:

Visual Assessment

I open the Twitter_archive_enhanced.csv manually in excel and scrolled through the CSV file. I was able to spot some quality and tidiness issues

Quality Issues

- Unnecessary HTML tag in the source column
- The text column includes a link
- Strange names such as a, an, the, all, very, quite etc in the names dataset

Tidiness Issues

- Doggo, floofer, pupper, and puppo columns should be merged into one column

Programmatic Assessment

I used the pandas.info method on the arc_df, image_df, and tweet_df. I also use the value counts method on the rating_numerator,

rating_denominator, and names columns to count the unique values. This process helped me to identify some issues which include

Quality Issues

- tweet_ids are stored as an integer
- The DateTime datatype is a string
- Values for the rating numerator are significantly higher than 10. eg. 420, 666, 1776
- rating_denominator has values other than 10. e.g 420, 666, etc

3. Cleaning Data

I created a copy of three DataFrame before I started cleaning. For each quality/tidiness issue, I performed the programmatic data cleaning process in 3 stages - Define, Code & Test.

Storing Data

After completion of the cleaning process, I merged the three datasets into one DataFrame and I stored the clean datasets in twitter_archive_master.CSV file