Wrangle Report

I did this project as part of Udacity Data Analyst Nanodegree Program, in order to practice and examine my understanding of the Data Wrangling section.

This project entails two parts 1/data wrangling, which consists of gathering data, assessing data, cleaning data; 2/storing, analyzing, and visualizing the wrangled data, whereby I would try to draw some insights based on the datasets I have on hand.

This report will detail my data wrangling efforts.

Gathering data

There are three datasets in three different formats

- WeRateDogs Twitter archive, which can be obtained manually from course website
- Tweet image predictions, which can be downloaded programmatically via a given URL
- Twitter JSON data, which can be downloaded via Twitter API

To acquire the Twitter archive file, I manually downloaded the Twitter archive file from course website and saved it as twitter_archive_enhanced.csv. I opened it in a dataframe called twt.

To acquire the image predictions, I used the Requests library to download the file from the given URL. I saved the file as image_predictions.tsv and opened it in a dataframe called images.

To acquire the Twitter JSON data, I used tweet IDs in the Twitter archive file, queried Twitter APIs for each tweet's JSON data using Python's Tweepy library, and stored each tweet's entire set of JSON data in a file called tweet_json.txt. I opened the file in a dataframe called twt_jsons.

Accessing data

The fact that this analysis has three datasets means that there is a tidiness issue, because observation should form one single row. I also checked duplicate columns of the three datasets.

I took a holistic view on each dataset via 1/.info() and 2/.describe(). I also looked into the content of individual columns by value_counts() and sample(). I also checked to see if the columns have the right data type.

Specific to the Twitter archive file, as many details are extracted from the text field, I took a close look at the content of the text column, and spot-checked if 1/name, 2/rating details are extracted properly.

Cleaning data

I made a copy of the three datasets before making any changes.

To start, I first identified and removed the retweets and tweets with missing pictures.

I also ensured the columns have the right datatypes (timestamp, rating_numerator, rating_denominator), removed the 1/empty columns, which include the ones that detailed retweet status, 2/duplicate columns within the same table and across the three datasets.

Meanwhile, I compared the text field with the dog name and rating_numerator, rating_denominator columns to correct wrong extraction. Some of the rating details were not properly extracted, and I manually corrected those. I also removed the tweet that doesn't have any rating details.

In addition, I also converted the four dog type columns (doggo, floofer etc) into one column.

Lastly, I merged the three cleaned datasets and removed the 1/duplicate and 2/empty columns

For each step, I followed the procedure of 1/identify issue 2/come up with a solution, and 3/test if the solution has been resolved.