Data Wrangle Report

About the dataset:

Dataset 1

The dataset that i will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc.

Dataset 2

This dataset consists of the image prediction with tweet ID, image number consisting of the most confident prediction (numbered 1 to 4 since tweets can have up to four images), this dataset will have additional data for the tweet archive dataset, The prediction is made by a neural network that can classify breeds of dogs.

Dataset 3

This dataset contains additional data beyond the data included in the WeRateDogs Twitter archive.This dataset will have retweet count and favourite count.

Data Gathering

Twitter archive csv file

Udacity provided the link to this dataset

url ---→ https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv

which i downloaded and imported into a dataframe in my notebook.

Tweet image prediction

Using python's request library, I downloaded the tweet image prediction file on Udacity's page and saved it to my machine as image_prediction.tsv file and imported the file into a pandas dataframe.

Data from the Twitter API

I downloaded the twitter_json.txt text file from udacity's server because i wasn't given twitter developer account privilege, I then created a dataframe from this json text file with columns id, retweet_columns and favourite_count

Assessment

Visual

I opened the first file (twitter_archive_enhanced.csv) in my excel spreadsheet and spotted some quality and tidiness issues including:

Tidiness:

1. The columns floofer,doggo,puppo,pupper which are the dog stages according to the dogtionary should be in a column going by dog_stage.
2. The three datasets are parts of the same observational unit and sould be merged into one.

Quality:

1. **The html tags in source column are not necessary and should be cleaned**
2. **Some of the names in the name column is in an inappropriate and missing**

Programmatic

After calling the info pandas function on the first dataset, I saw that some columns are in an inappropriate datatype amongst other quality issue, I worked around the quality issues and noticed that there are some instances where some dogs have 2 dog_stages which shouldn't be. Some other quality issues i noticed are:

1. There is retweets column in twitter archive dataset that are duplicates of actual tweets
2. Many tweet_id(s) in the archive datasets are missing in the image prediction datasets and the tweet.json dataframe
3. The in_reply_to_status_id and in_reply_to_user_id are in a inappropriate datatypes, Timestamp column is not in a datetime datatype
4. TimeStamp column is not in datetime datatype
5. none string should be in nan format
6. columns p1,p2,p3 in image_prediction datasets should be converted to a categorical datatype
7. retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp will become empty after archive dataframe has dropped duplicates which will have to be dropped.
8. drop the rows with missing tweets and dog name
9. tweet_id should be converted to object datatype(string)

Other tidiness issues:

- The three datasets are part of the same observational unit and should be merged into one (dataset: twitter_archive_master.csv as instructed)

Data cleaning

After i created a copy of the twitter archive dataframe,i then performed the programmatic cleaning process needed in define, code and test format.

Data storing

The cleaned dataframe was stored in twitter_archive_master.csv file as instructed by udacity after I was done with the cleaning process.