Udacity - Data Analyst Nanodegree

Project: Wrangling and Analyze Data

1. Summary

The goal of this project is to gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling.

2. Gathering Data

The data was gathered from three sources:

- The WeRateDogs Twitter archive (twitter-archive-enhanced.csv) was provided in the starter kit of the project. This file contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.
- Tweet image predictions archive (image-predictions.tsv). In order to get this data, we had to make a request to an url provided in the project description. This data is about what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network.
- Data from Twitter API to get each tweet's retweet count and favorite ("like") count at minimum, and any additional data.

3. Assessing Data

3.1 Archive Data

- Quality Issue:
  1. Null values in [in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls]
  2. tweet_id column has int type, but the other "id" columns have float type, so to be consistent we have to convert all "id" columns to float type.
  3. timestamp and retweeted_status_timestamp columns have object type, but that columns must have timestamp type.
  4. name column has 745 rows with "None" value and 55 values with "a" value.
- Tidiness Issue:
  1. One variable is spread across four different columns (doggo, floofer, pupper and puppo columns). These four columns should be combined into a single column as this is one variable that identify stage of dog.

3.2 Image Data

- Quality Issue:
  1. p1, p2 and p3 have invalid data, there are rows with cases such as laptop, restaurant, basketball, tricycle, etc.
  2. p1_conf, p2_conf and p3_conf have lower values (near to zero), so this indicates that there are predictions made with underestimation.
  3. In order to merge with the other dataframes, we have to change the tweet_id column's type to float.

3.3 Tweet's retweet info

- Quality Issue:

1. There are 161 retweets.
2. In order to merge with the other dataframes, we have to change the tweet_id column's type to float.

- Tidiness Issue:

1. We have to merge the three datasets.

4. Cleaning Data

Define, code and test the following items:

- Copying of the original pieces of data
- Merge the three datasets
- Drop duplicated columns
- Combine doggo, floofer, pupper, puppo columns
- Drop duplicated tweets
- Remove null values
- Convert column types
- Quality issue