Data Wrangling Report

1. Gathering Data

In this project, I worked with the following three datasets.

i). Enhanced Twitter Archive

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets. The dataset for this has been sourced by downloading the file manually from twitter_archive_enhanced.csv

ii). Image predictions File

This dataset is hosted on Udacity's servers and was downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

iii). Twitter API

I also used Twitter API to get additional data for the Project.

2. Assessing the Data

After gathering all pieces of data, I assessed them visually and programmatically for quality and tidiness issues. I detected and documented **eight quality issues** and **three tidiness issues**.

Visual Assessment

By examining the Twitter Archive dataset, I was able to identify one quality issue and one tidiness issue.

- **Quality**: The retweets were also in the Twitter Archive dataset, yet I was only interested in the original tweets.
- **Tidiness**: The columns 'doggo', 'floofer', 'pupper', and 'puppo' in the Twitter archive data frame

Programmatic Assessment

Through the program code, I was able to identify eight quality issues and two tidiness issues in the datasets.

- **Quality**:
    - ♦ **Twitter Archive Dataset**
        - We are interested with the original tweets only and not including retweets.
        - We have so many missing values in the 'in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id' and 'retweeted_status_timestamp' columns
        - The rating denominator is supposed to be always 10.
        - We have long html names in the source column rather than the device name i.e http://twitter.com/download/iphone
        - The 'timestamp' and 'tweet_id' columns are in wrong data types respectively. Also in the 'twitter_image' and 'tweet_api' dataframes.
        - We have so many dogs without names and also improper names like 'a','an', 'the','just', etc
    - ♦ **Image Prediction File dataset**
        - The p1, p2 and p3 columns have both upper and lower cases for the individual records.
        - We have some false predictions for the dogs.
    - ♦ **Twitter API**
        - The columns 'doggo', 'floofer', 'pupper' and 'puppo' in the twitter_archive dataframe can be reshaped to one column to avoid redundancy.

- The twitter_archive dataframe and twitter_api_df dataframe can be merged for easy analysis.

3. Cleaning the Data

I created copies for each original dataset and named the new datasets twitter_archive_clean, twitter_image_clean, and tweets_api_clean respectively. I cleaned the datasets for each of the issues I had identified during the assessment stage. The process of cleaning involved, Defining, Coding, and Testing respectively. Eventually, I merged the three datasets for easy analysis.

4. Storing the Data

After the cleaning process, I stored the data as a CSV file and as an SQLite database by the names twitter_archive_master.csv and twitter_archive_master.db respectively.