Report on Data Wrangling

Project-4
Udacity Data Analyst Nanodegree

By: Jahnavi Pinnamraju

This project involves wrangling the data that include the following steps.

1. Data Gathering
2. Data Assessment
3. Data Cleaning
4. Data Storage
5. Data Visualization

Data Gathering:

Data was collected from the twitter-archive-enhanced which is a csv file. This CSV file contains 2356 columns and 17 rows. It consists of columns like tweet_id, timestamp, source, text, etc.

Another file called image-predictions.tsv was provided to analyze the type of dogs present in tweets. The data such as retweet count, friends count, followers count, and retweet count was gathered using Twitter API and JSON library. A dictionary was formed to store the data list from twitter API. Later a data frame with the significant data elements including url, tweet id, followers count, friends count, favorite count and retweet count. Later I have converted to a data frame using the pandas library.

Data Assessing and Data Cleaning:

After data collection step, I assessed the data and identified some issues. I have framed the following objectives after assessing the data and then cleaned the data. The issues were defined, coded, and tested.

Quality issues:

1.To convert timestamp to date-time variable.

2. ID fields should be objects.

3.To delete the retweets.

4.To standardize the breed names of the dogs.

5.Issue with dog names.

6.Dropping unnecessary columns.

7.To remove the duplicated images.

8.Source column must be categorical.

9.Ratings.

Tidiness issues:

1.To merge tweet_json_clean and df_clean tables.

2.To merge the 4 dogtypes into a single column.

Data Storage:

Later I have stored the dataframe into csv file. The name of the file is. twitter_archive_master.csv'.

Data Visualization:

Three visualizations were plotted. They include correlation between the data elements in the data frame. I found that there is a strong correlation between the tweet_id and in_reply_to_status_id. I have extracted mostly repeated words in the tweets. Japanese, Irish, Bigfan, Setter and Lost were some of the mostly used words. From the above barplot, we can observe that a, Charlie, Lucy, Oliver, Cooper, Penny was the mostly used dog names.