WRANGLE REPORT

By Abhilasha Dwivedi

1.Gathering Data

I gathered the first data by downloading the twitter-archive-enhanced.csv file from the Udacity that is provided in the project section.

The second dataset was programmatically downloaded from Udacity's server using the requests function (URL = https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_imagepredictions/image-predictions.tsv).

Then , as I didn't have twitter API therefore I used static twitter data "twitter.json" which was provided by the udacity.

2.Accessing Data

Here in this ,I found the quality and tidiness issues. Following are these:

Quality issues :

(Quality issue in twitter_archive_data dataframe)

1. Name column contain some invalid name.
2. In this data tweet_id datatype should be String instead of Int.
3. Timestamp is not of datetime format.
4. Name column contains a total of 745 of name which was showing none.
5. Tweet rating contains values as in decimal and also some tweets contain more than 2 rating because it is rating two dogs in that column.
6. It contain columns which are not needed such as in_reply_to_status_id , in_reply_to_user_id .
7. More than two dogs are rated in one tweet which can bring inconsistency in that data.

(Quality Issue in prediction_data Dataframe)

1. In columns p1,p2,p3 dog breeds are not valid everywhere, contains some invalid dog breed such as window etc.
2. Data in img_num is not of much use because not much information can be inferred.

(Quality issue in additional_tweet Dataframe)

1. tweet_id can be of string type.

Tidiness issue

Tidness Issues in DataFrame

1. In twitter_archive_data, it has 4 columns namely dogger, floofer, pupper, puppo which all signify dog stage only
2. Column of Dataframe such as additional_tweet , prediction_data and twitter_archive_data can be joined

3.Cleaning Data

For this part ,firstly I made the copies of the dataframe . Next I cleaned the column values.There were some retweets which was needed to be deleted.Then few columns like “retweeted_status_id” and “retweeted_status_user_id” were dropped which was not needed .There were decimal values which was corrected.Then dataframes were merged using type=’inner’.Next was to format the timestamp column and then tweet_id type was changed.

4.Storing data

After clearing the data I stored the final cleaned dataset as:

merged_df.to_csv('twitter_archive_master.csv', encoding='utf-8', index=False)