WRANGLING REPORT

OLADEJO OLANREWAJU OLAWALE

Data Gathering

Before gathering the datasets the libraries needed in the project such as numpy, panda, json, request, tweepy and matplotlib.pyplot were imported, then data was gathered from 3 different sources separately into the workspace as follows:

1. The WeRateDogs Twitter archive data (twitter_archive_enhanced.csv), which was manually downloaded from the Udacity classroom and uploaded into the project workspace.

2. The tweet image prediction (image_predictions.tsv), programmatically downloaded from the Udacity provided url using the request library.

3. The Twitter API (tweet_json.txt), downloaded from the udacity classroom and the 'tweet_id', 'retweet_count', 'favorite_count' columns were read line by line into a panda dataframe.

The 3 data files gathered were loaded into 3 separate data-frames namely: twitter_archive_df, image_predictions_df and json_df.

Data Assessment

The twitter_archive_df was assessed first both visually and programmatically, followed by the image_predictions_df and the json_df by viewing the information of each dataframe and identifying several quality and tidiness issues as follows in the dataframes:

Quality issues

Twitter archive data

1. Wrong datatype for the tweet_id column, timestamp and source column
2. Invalid names found in the name column such as (a,an, o, my, his, this, all, old, the, not, one, quite).

3. Null entries for name represented as None instead of NaN.
4. The entries of source is dirty.
5. The retweets are not needed as specified in the project specification.
6. Ratings should be in one columnn

Image prediction data

1. (66) jpg_urls are duplicated.
2. Entries in column P1 P2 P3 is not consistent case wise of dog breeds and underscore is used instead of space.

Tidiness issues

1. columns containing the dog stages (floofer, doggo, pupper, puppo) should be put into one column
2. The three data sets should be merged into one.

Data Cleaning

Before commencing on data cleaning a copy of each data-frames was made in other not to loss the original data.

First of all the quality issues were addressed, starting with the wrong datatypes for the tweet_id column, timestamp and source column in the twitter_archive_df which were converted to the right datatypes as follows: tweet_id(string), timestamp(datetime) and source(category).

The invalid names found in the name column such as (a, an, o, my, his, this, all, old, the, not, one, quite) in the twitter_archive_df were changed to NaN, some null entries for name represented as none were also converted to NaN.

The sources were extracted into a clean text from the original html string in the dataframe.

The rows containing the retweets column (retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp) were all dropped as we are interested only in the original tweet.

A rating column was also created from the rating_numerator and rating_denominator column by dividing the rating_numerator by the rating_denominator, and both columns were dropped.

Some jpg_url duplicates were found in the data and the rows containing the duplicates were dropped.

In the image_predictions_df columns p1, p2 and p3 had entries that were not in consistent case-wise and also underscore was used to separate text instead of space, the issues were corrected making the entries all lowercase and the underscore was replaced by space.

The tweer_id column in the image_predictions_df had the wrong datatype and was converted to the string datatype. The json_df was found clean.

The tidiness issue was addressed next, the columns containing the dog stages (floofer, doggo, pupper, puppo) were melted into one column as they are datatype of categories, The twitter_archive_df and json_df were merged first and the resulting data was merged with the image_predictions_df, columns such as ‘in_reply_to_status_id' and 'in_reply_to_user_id' were dropped from the data.

The merged clean dataset was then stored as a new csv file named **“twitter_archive_master.csv”** file and analysis was carried out.