DATA WRANGLING REPORT

Wrangle and Analyze – WeRateDogs tweets

Udacity – Data Analyst Nanodegree

Chinelo Cynthia Ezenwafor

INTRODUCTION

Data wrangling is part of the courses in the Data Analyst Nanodegree offered by Udacity. It follows three major steps;

- Data Gathering
- Data Assessment
- Data Cleaning

Data Gathering

The twitter_archive_enhanced.csv and image_predictions.tsv files of the WeRateDogs tweets were gathered from udacity servers using the requests package. However, some information needed for analysis such as favorite_count and retweet_count was not available in the dataset and needed to be queried from twitter. The tweepy package was used for querying tweet_ids present in the twitter_archive_enhanced.csv and the result was stored a tweet_json.txt.

The data gathered were stored in three different pandas dataframe:

- twitter_archive_data: loaded from twitter_archive_enhanced.csv
- image_prediction: loaded from image_predictions.tsv
- tweets: loaded from tweet_json.txt

Data Assessment

The data was assessed using the two types of assessment, visual and programmatic assessment. The following quality and tidiness issues were disovered:

Quality issues

Twitter_archive_data

1. Some of the names in the name column are not proper dog names
2. Dataset contains retweets and replies, but we need only original tweets, not retweets or replies
3. Incorrect datatypes
4. Irrelevant columns that would not be used for analysis
5. The source column contains data that should be split to give more meaningful information

Image predictions dataset

6. Some column names are not descriptive (p1, p1_conf, p2, p2_conf, p3, p3_conf)
7. Some of the p1_dog an p2_dog columns have false values which means they are not identified as dogs by the neural network.
8. Tweet_id column has an incorrect datatype
9. Use of both lower and upper case in the predictions columns

Tweets dataset

10. Tweet_id column has an incorrect datatype.

Tidiness issues

Twitter_archived_data dataset

1. The doggo, floofer, pupper and puppo columns which are dog stages should be a single column.

Image_prediction dataset

2. it contains repeated predictions of dog type along with the confidence interval, the prediction with the highest confidence interval should be used for analysis and others dropped.
3. The three datasets should be merged together to give one master dataset.

Data Cleaning
Before cleaning, a copy of each dataset was made using the pandas .copy() method.

Cleaning steps:

- Tweets that are retweets or replies were dropped using the .drop() method by rows.
- Irrelevant columns were also dropped using the .drop() method by columns
- The four dog stages (doggo, pupper, puppo and floofer) were merged into one column using the pandas melt function, and the stages that had None were converted to Nan type.
- The invalid dog names issue was resolved by replacing them with 'None' and then converting it to the Nan type.
- The .astype() method was used to correct datatype for the tweet_id columns and the pd.to_datetime() method for the timestamp column.
- The non-descriptive column names were renamed using the pandas rename function.
- The str.lower() method was used to change all the cases in the prediction columns to lower case.
- The columns which were not identified as dogs in the image_prediction dataset were identified and dropped from the dataset.

- The image_prediction dataset had three different dog predictions with confidence level, the first predictions that were 'True' for dogs were taken and the remaining columns were dropped.
- Finally, the three datasets were merged on the tweet_id column using the pandas merge function and stored in the master_df data frame.

CONCLUSION

A total of thirteen issues were identified and cleaned in this wrangling process, however, this does not mean that the dataset is completely free from issues as Data wrangling is a continuous process. The dataset id finally ready for analysis and visualization.