WeRateDogs Twitter Archive – Wrangle Report

Introduction

This report describes the wrangling processes involved to assemble and clean the data required for analysis of the WeRateDogs Twitter Archive.

Data Gathering

Data was gathered from 3 different sources;
df_twitter_archive: “twitter_archive_enhanced.csv” file was provided by
Udacity and downloaded manually then imported into the working environment using Pandas library.
df_image_predictions: “image_prediction.tsv” file was downloaded programmatically using Requests library from a provided URL. This file consists of image predictions results for the dogs’ breeds obtained through a neural network on most of the tweets in the “twitter_archive_enhanced.csv” file.
df_twitter_extra: I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data
in a file called “tweet_json.txt file” obtain extra information pertinent
to the tweets’ ids in the “twitter_archive_enhanced.csv” file. I read this .txt file line by line into a pandas data frame with tweet ID, favourite count, retweet count and create date.

Data Assessing

I assessed the three saved data frames both visually and programmatically.
Visual assessment: this was done by opening the “twitter_archive_enhanced.csv” file in Microsoft Excel and also opening the three data frames separately in Jupyter notebook.
Programmatical assessment: this was done strictly on Jupyter notebook using the .head() to check the top rows of the data, .shape() to check the number of rows and columns, .info() to check data characteristics, .describe() to check data description, .dtypes, to check the datatypes of the columns, isnull() to check for null values and .duplicated() to check for duplicates on the three data frames.
The datasets were assessed under two criteria, quality and tidiness. All issue detected were documented as one of these two.
Quality refers to issues related to the content of the data, sometimes called dirty
data. The standard criteria of completeness, validity, accuracy, and consistency of the data were used to identify quality issues
Tidiness refers to issues related to the structure of the data, sometimes called
messy data. The basis for assessment is that each variable forms a column, each observation forms a row and each type of observational unit forms a table.

Data Cleaning

This part of the data wrangling is divided into Define, code and test
the code. These three steps were on each of the issues described in the assess
section.

On the Twitter archive data frame, the following cleaning processes occurred;

- Retweets and replies were deleted by dropping not null in column 'retweeted_status_user_id' and 'in_reply_to_user_id'

- Columns like 'source', 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'expanded_urls' that are not needed were dropped.
- Change datatype of ‘tweet_id’ and ‘timestamp’to strings and datetime respectively.
- Change all Null objects as represented as 'None' to NaN.
- Change Incorrect names in the name column, names weren't successfully extracted from the text. i.e. (a, an, the, very) to NaN
- Solve the Invalid rating data is both rating_numerator and rating_denominator
- Create a date stage column and fill with columns containing stages names.

On the Image predictions data frame ;

- The datatype of ‘tweet_id’ was changed to strings
- Change columns names that are not informative.
-

On the Twitter extra data frame;

- Change datatype of ‘create_date’ and ‘tweet_id’ to datetime and strings respectively

Finally, I merged the 3 cleaned data frames on tweet_id into one dataframe called ‘twitter_archive_master’

Conclusions

Data wrangling processes help to provides a clean data frame for exploratory data analysis and visualization.