WERATEDOGS DATA WRANGLING REPORT

PRESENTED BY OBI UCHECHI LYDIA

This report summarized the process used in this project. The main goal of this report is to put into practice what was taught as a student for the Udacity Data Analyst Nanodegree program.

The project is about gathering of data from web with the use of API to analyze the data to generate insights and make a data driven decisions based on the extracted information.

The data used was driven from a Twitter account WeRateDogs @dogs_rates which is a twitter account. That information was extracted from the rating of people's dogs.

The objective of this project is to query and wrangle the data derived from the twitter accounts.

The processes used are gathering of the data, accessing the data and cleaning the data. We further stored the data.

The data were gathered from three sources. The tweetarchive - tweetarchive(twitter-archive-enhanced.csv) was manually downloaded. I programmatically downloaded the image prediction data – imgpred (image-predictions.tsv) using the python request library. I used the Twitter API - tweety to gather data about favorite count and retweet count. The twitter-archive-enhanced.csv contains data about tweet from WeRateDogs from 2015-2017.

Assessment was carried out on the three data set provided. I accessed them virtually and programmatically for quality and tidiness.

Assessing virtually was easy. I looked through the files withing the Jupyter notebook by scrolling thoroughly. The tweetarchive data sets- twitter-archive-enhanced.csv contains 2,356 columns and 15 rows while the imgpred data sets

- image-predictions.tsv contains 2,075 columns and 12 rows and the tweety-Twitter API data set contains 2,327 columns and 5 rows.

Programmatical assessment was also carried out by using python functions and methods such as .info(), .sample(), .shape(), dtypes(), .describe(), .value_counts(). They were used to access each of the dataset.

Quality issues:

1. In tweetarchive, there are missing values found in the following columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp. Also, removing the retweet_counts on the tweety data as requested.

2. In tweetarchive, time stamp is in object and needs to be changed in date time

3. In tweetarchive, doggo, floofer, pupper, poppy should be classified in one table as dog names.

4. In tweetarchive, under the dog names, some names are missing.

5. The tweet ID in tweetarchive should be changed from an integer to a string.

6. In imgpred, there are case sensitive dog names in p1, p2, and p3

7. Tweet ID in impred should be changed to a string

8. Tweet ID in tweety should also be changed to string

Tidiness issues:

1. In tweety, the id should be changed to tweet ID for easy merger
2. Tweetarchieve, imgpred, and tweety should be merged as instructed for ease of reading.

Cleaning of Data

Before cleaning the data, I made original copies of each dataset

- tweetarchive_clean = tweetarchive.copy()
- imgpred_clean = imgpred.copy()
- tweety_clean = tweety.copy()

I used the three steps cleaning process which is to define the issues raised, code, and test within the data wrangling.

Storing of Data

After I have gathered, assessed and cleaned the three datasets, the datasets were merged and saved in a csv file called twitter_archive_master.csv.

Analyzing and Visualization of the Data

Finally, the saved dataset was analyzed and visualized to answer certain questions concerning the dataset.