WERATEDOGS DATA WRANGLING REPORT

This report describes briefly all the processes I used in my project.

The goal of this report is putting into practice what was taught and learned from the Udacity Data Analyst Nanodegree program, Lesson 3 (Data Wrangling). The project is about gathering data from the web with the use of API to analyze the data, generate insights and make data-driven decisions based on the information extracted. The tweet from WeRateDogs @dog_rates which is a Twitter account from people's dog rate with the humorous comment about dogs contains basic tweet data for all 5000+ of their tweets.

OBJECTIVE OF THE PROJECT

To query and wrangle data sourced from the Twitter account @WeRateDogs effectively. The dataset can be analyzed after gathering, assessing and cleaning.

PROCESS USED

1. GATHERING THE DATA

Three different datasets gathered were all obtained as follows.

Twitter_archive_file: It was provided and I downloaded manually, uploaded in my Jupyter Notebook and imported the required libraries. Using the python pandas I read the file *[i.e pd.read_csv()]* into a dataframe which I called **Twitter.**

Tweet-image-prediction-file: Using the python *request and os libraries* imported, I used the *request.get() function* to get the data through the URL and a response of 200 was returned showing that the request is successful. Using with open function, I wrote the response content to a tsv file within the same file and read the downloaded tsv file into a dataframe I called **Image**.

Tweet_Json text: I signed up for a Twitter account to get a developer account for access to Twitter API. The account was created and gave me access to app credentials. Using the tweet_id from the given dataset twitter_archive_file I query the Twitter API, writing the content into a text file which was converted to a JSON file and read in a dataframe using the *read_json pandas function* and I called it **Tweet**.

2. ACCESSING THE DATA

I assessed the three datasets both visually and programmatically for quality and tidiness issues.

Accessing visually: I assessed the csv files using MS Excel, and also through each table within a Jupyter Notebook by scrolling thoroughly.

Accessing Programmatically: Using python functions and methods such as .info(), I accessed each of the dataset.

Here are the issues observed.

Quality issues

Twitter:
1. 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.
2. The timestamp column is an integer instead of a date time
3. The tweet_id column is an integer instead of a string.
4. Some Dog names in name column have invalid entries.
5. Missing addresses were found in the expanded_urls column

Image:
6. Inconsistent alphabetic format in columns P1, P2 and P3
7. The tweet_id column is an integer instead of a string

Tweet:
8. The id column is an integer instead of a string

Tidiness issues

Twitter:
1. The Categories of the Dog stage (doggo, floofer, pupper and puppo) were in different columns and need to be one column.

Tweet:
1. The column containing friends_count has only 1 value and only 24 values in the followers_count column.

General issues:
1. The column name tweet_id is in the three dataset and should be left that way.
2. The three dataset should be merged as a single dataset.

3. CLEANING THE DATA

I made original copies of each of the dataset and used the three-step cleaning process (Define, Code and Test), to clean the stated issues observed.

4. STORING THE DATA

I merged and saved three datasets in a csv file called **twitter_archive_master.csv.**

ANALYZING AND VISUALIZATION THE DATA

The saved dataset was furthermore analyzed and visualized to answer certain questions.