Data Wrangling Report

Introduction:

Wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. The project contains:

Data Gathering:

Gathering data for the project from 3 different locations as below:

A- Download the **file twitter_archive_enhanced.csv** manually then read into a dataframe with **df_1** name using the pandas **library pd.read_csv** method.

B- Download the file **tweet_image_predictions.tsv** programmatically using Requests library then read into a dataframe named **df_2**.

C- Using the **tweet IDs** in the WeRateDogs Twitter **archive**, query the **Twitter API for each tweet's JSON data** using Python's **Tweepy** library and store each tweet's entire set of JSON data in a file called **tweet_json.txt** file. Then read this **.txt** file line by line into a **pandas DataFrame** with **tweet ID, retweet count**, and **favorite count** into a dataframe named **df_3.**

Data Assessment:

Investigating the imported datasets by both visually and programmatically for the Quality and Tidiness issues to meet specifications as below:

A- Visual assessment: by Excel and text editor's software and each piece of gathered data is displayed in the Jupyter Notebook for visual assessment purposes.
B- Programmatic assessment: by coding like pandas' functions and/or methods are used to assess the data.

Data Cleaning:

Doing some cleaning efforts with coding to get high quality and tidy master pandas DataFrame named

tweet_master_data.csv.

The assessment and cleaning efforts as below:

No.	Investigation	Solution	Location	Type
1	retweats needs to be deleted	Drop **retweeted_status_id** column	Df_1	Quality
2	**in_reply_to_status_id, in_reply_to_user_id**, **retweeted_status_id, retweeted_status_user_id** and **retweeted_status_timestamp** are not important columns will be dropped	Drop these columns	Df_1	Quality
3	**expanded_urls** has 59 missing values	Drop missing values	Df_1	Quality
4	**tweet_id** is stored as integer. change it to string	Change type to string	Df_1	Quality
5	**timestamp** is stored as string. change it to datetime	Change type to datetime	Df_1	Quality
6	In **name** column, 745 are stored as "none", 55 are stored as "a", 8 are stored as "the", 7 are stored as "an" 1 is stored as "my" and 1 is stored as "by"	Replace 'none', 'a', "an", "the", "my", "by" to NaN	Df_1	Quality
7	change **tweet_id** to string	Change type to string	Df_2	Quality
8	Remove duplicates from **jpg_url** Column	Drop duplicates in **jpg_url**	Df_2	Quality
9	Change **id** column name to **tweet_id** to be unique	Change the name		
10	- make **tweet_id** type as string	Change type to string	Df_3	Quality
11	- Adjust `doggo` `floofer` `puppo` `pupper` Dog Stage columns to be one column **dog_stage** and make values unique.	Replace "None" with empty string then Combine the columns into one column and Modify the un-defined named	Df_1	Tidiness
12	- merge the image predictions dataframe to the twitter Achive dataframe	Merge to one dataframe	Df_2	Tidiness
13	- merge the Twitter Api dataframe to the twitter Achive dataframe	Merge to one dataframe	Df_3	Tidiness

Storing Data:

Store the clean DataFrame in a **CSV file** with the main one named **tweet_master_data.csv**. The dataframe has **1987** records.

Data Analysis and Visual Reporting:

Analyze and visualize the cleaned wrangled data to get some insights visualization about the data will be discussed in the next report.