DATA WRANGLING REPORT

OVERVIEW

In order to realize project number 2 which purpose is to wrangle and analyze **WeRateDogs Tweeter** data, we started by collecting data from various sources and then by cleaning them. This report summarises our efforts in this process.

GATHERING THE DATA

In the gathering phase, 3 files were required: ***twitter-archive-enhanced.csv*** which we downloaded directly from the internet, ***tweet_json.txt*** that we downloaded from the tweeter api, and ***image-predictions.tsv*** that we downloaded programmatically using the ***requests*** library.Then we extracted retweet count and favorite count from the ***tweet_json.txt*** file and merged to the ***twitter-archive-enhanced.csv*** dataframe.

ASSESSING THE DATA

After gathering the data, we have assess it programmatically using the ***pandas*** library fucntions and visually using *Microsoft Excel.*

CLEANNING THE DATA

Assessing the data revealed many issues in both ***twitter-archive-enhanced.csv*** and ***image-predictions.tsv*** dataframes as follow:

- Quality issues
    - In the ***twitter-archive-enhanced.csv*** dataframe:
        - some cells in the *expended_urls* column have duplicated image url and unterminated images urls and some links lead to video and external webpages, so we removed duplicated and unterminated images urls from the expended_urls column, same as images urls that lead to video or external webpages.

        - there are cells in the *expanded_urls* column with null values: we removed rows that do not have image url.

- some cells in the *text* column contain hashtags and mentions:we removed any tag and mention from the text column.

- the *timestamp* column has unnecessary characters: +0000: we removed the trailing "+0000" from the timestamp column.

- the data type of the *tweet_id* column is *int* which is supposed to be *object:*we converted the type of tweet_id column to string

- the data type of the *timestamp* column is supposed to be *datetime*, not *object:* converted the data type of the timestamp column from object to datetime.

- the cells in the *text* column contain short images urls: we deleted short image urls from the text column.

- the *text* column contains ratings: we removed ratings from the text columns.

- in the *name* column, the abscence of value is represented by the string 'None' which is confusing: we replaced the string 'None' with the python None type in the name column.

- Tidyness issues:
    - In the ***twitter-archive-enhanced.csv*** dataframe:
        - the columns doggo, floofer, pupper, puppo, represent the same value which is dog stage: we combined the columns doggo, floofer, pupper, and puppo into a single column named dog_stage.

        - there are many useless columns: in_reply_to_status_id, in_reply_to_user_id, source, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp: we removed unnecessary columns and reorganised columns.

        - the timestamp column contains time, day, month, and year at once: we splited the timestamp column into time, day, month, and year.

    - In the ***image-predictions.tsv*** dataframe:
        - the image_predictions dataset contains mutliple predictions for each jpg image: we filtered and preserved only the best prediction data for each image.

STORING THE DATA

After cleaning the issues that we detected, we the merge resulting the ***twitter-archive-enhanced.csv*** dataframe and ***image-predictions.tsv*** dataframe into a single master dataframe that we stored in a file called ***twitter_archive_master.csv.***