WeRateDogs - Twitter Data

I. Gather Data

Instructions were given by the udacity instructor on how to proceed in gathering data.

- I downloaded the data, which is a given CSV file and named as **twitter-archive-enhanced.csv**.
- Next, I programmatically downloaded the file image predictions file, which is in the .tsv format.
- Then, I downloaded 'tweet-json.txt' from the udacity platform as I was having issues confirming my twitter developer account. I read the API pseudo-code and understood it before proceeding to the next step

A dataframe was created using pandas for the three files;

- *archive_df* - this is a dataset "twitter-archive-enhanced.csv" which was converted into a dataframe and gives information on basic tweet data such as tweet id, timestamp, and the tweet itself; where various other details were extracted from it, such as the dog’s name and image.
- *predictions_df* - This dataset will contain information about predictions about the image, such as the co-efficient for each prediction. This dataframe was gotten from the image prediction file, which was in.tsv
- *tweetsinfo_df* - This dataset will contain information like tweet_id, no of retweets and no of favorites etc., it was gotten from the twitter-json file

II. Assessing the data

Each table was displayed in its entirety by displaying the pandas DataFrame that it was gathered into. This task is the mechanical part of the visual assessment in pandas.
Steps taken while assessing dataframes include:

- The first five rows of the dataframe were viewed to see if any anomaly such as column names and misspelling could be seen easily.
- Then null values were checked
- Duplicate rows were also investigated.
- The numerical values were then described to check for outliers and weird values.
- Then the info of each column was investigated.
- Lastly, we checked the datatypes for irregularities.
- Then based on some view observations, various columns were investigated.

1.Enhanced Twitter Archive

The columns were well explained https://sfm.readthedocs.io/en/1.4.3/data_dictionary.html . for better understanding of the datasets

Quality

- Missing values in some columns from archive_df
- outrageous and inconsistent values in rating numerator and denominator
- one rating has a zero denominator
- weird names found for dogs - 'infuriating', 'just', 'life', 'mad', 'my', 'not', 'officially', 'old', 'one', 'quite', 'space', 'such', 'the', 'this', 'unacceptable', 'very'
- timestamp and retweeted_status_timestamp must be of datetime instead of the object
- comparing both prediction_df and tweetsinfo_df to archive_df we see that they both don’t have complete tweeter id like archive_df
- The columns which have missing values in doggo, floofer, pupper , puppo are written as None instead of NaN
- We see that the information on text is truncated to 50 characters. Anything in excess is ellipsized

Tidiness

- Plenty of columns explaining dog stage in archive_df when they could easily be merged as one
- in prediction_df p1,p2,p3 could have been given a more and self-explanatory name

III. Cleaning

The following steps were carried out when cleaning the data;

- All three datas were copied to a different dataframe so a not to deal or mess with the first one
- Row with zero”0” value ranking denominator was removed
- Timestamp and retweeted_status_timestamp were converted to datetime datatype

- Renaming columns dealing with prediction in the prediction_df to a more self-explanatory name
- None value in the dog stage columns was converted to an empty string for easy concatenation
- After concatenation, the same rows were noticed to have more than one stage. For these rows, the stages were splitted using a dash
- The three dataframes were merged together using inner on the tweeter id as common ground.

IV. Store

I stored the final dataframe into csv file with name **twitter_archive_master.csv** with final data of 2073rows and 29columns