WeRateDogs - Twitter Data

This project was majorly to test the data wrangling skills of students taking the Udacity data analyst nanodegree, in gathering data, cleaning gathered data, accessing cleaned data and making visual analysis of the data to make meaning to the gathered data.

The following steps were taken for this project:

- Gathering data
- Cleaning data
- Accessing data
- Storing data
- Visual analysis of stored data.

Gather Data

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

- First step was to download the archived 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 extension.
- Lastly, I downloaded 'tweet-json.txt' from the udacity platform as I had issues confirming my twitter developer account. I read the API pseudo-code to completely understand the code before I proceeded with the project..

Three different data frames were created using pandas for the three files described above. They are:

- ***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; and other details were extracted from it, such as information like the dog’s name and image.

- ***img_predictions_df*** **-** This dataset contains information about predictions about the image, such as the precision(in range between 0 and 1) for each prediction. This dataframe was obtained from the image prediction file.
- ***infotweet_df*** **-** This dataset contains information like tweet_id, no of retweets and no of favorites etc., it was gotten from the twitter-json file.

Assessing the data

Each table was displayed in its entirety by displaying the pandas DataFrame that it was gathered into. The steps taken while assessing the gathered datasets include but not limited to:

- The first five rows of the dataframe were viewed to see if any anomaly such as column names and mis-spelling could be seen easily.
- The null values were checked using .isnull().sum()
- Duplicate rows were also investigated using .duplicated().sum()
- The numerical values were then described to check for outliers and weird values.
- Then the info of each column was investigated to check for more information on the various columns.
- Lastly, we checked the datatype for each column for irregularities.
- Further assessment was carried out on some columns based on findings from the steps above.

The columns of the **enhanced archived data** were well explained in this link here for better understanding of the datasets.

The data post assessing was scrutinized to figure out issues around quality and tidiness that would later be cleaned, they are listed below.

Quality

- Some of the records are without names and also have a zero numerator rating, these records are irrelevant to our data and to avoid skewness of data, should be removed.

- Timestamp and retweeted_status_timestamp in archived_df table are of datatype object instead of datetime

- Columns which have missing values in doggo, floofer, pupper, puppo are written as None instead of NaN hence their representation seems like they have values when they don't

- Img_prediction_df column names - p1,p2,p3 could be given better explanatory names

- The archive_df table has some values in the retweet columns, which is not to be considered in this project as a user can retweet on their tweet. This means records(rows) with values in these columns will be removed.

- Once the above issue is sorted and we don’t have values in the retweet columns anymore, we can remove the retweet columns as they become redundant in our analysis

- The rating_numerator and rating_denominator columns in archive_df table could have inconsistent values from how they were extracted, these values need to be properly extracted and replaced

- Records with rating denominator values not equal to 10 could be taken as inaccurate (as the ideal situation, rating_denominator = 10), and would be removed to ensure data integrity

Tidiness

- The columns explaining the dog stages in archive_df could have easily been merged into one to give comprehensive information on the current dogstage of that record

- The three datasets currently exist in silos, we need to merge the datasets into one to have a master dataset.

Cleaning (Quality Issues)

The following steps were carried out to clean the quality issues in the data;

- All the datasets were copied to a different dataframe so as not to deal or mess with the original datasets.
- Records without names and zero numerator ratings were removed from the dataset
- Timestamp and retweeted_status_timestamp were initially recorded as strings, but were converted to datetimes.
- Replaced all None values in the datasets with NaNs to properly represent they are missing values.
- Renaming columns dealing with prediction in the prediction_df to a more self-explanatory name
- Records with values in the retweet columns were removed, as we don’t want inaccurate data in our analysis, and a retweet record is not needed as it can skew our data.
- The retweet columns were also removed as they became redundant columns
- To avoid inconsistencies with rating data (numerator and denominator), I properly extracted these values as floats from the dataset and replaced current numerator and denominator columns with them
- The ideal rating denominator value is 10, so every record that doesn't obey this rule are exempted going forward in the analysis
- The three dataframe were merged together using inner on the tweet_id as common ground.

Cleaning (Tidiness Issues)

The following steps were carried out to clean the tidiness issues in the data;

- Applied pandas method to create a column that is a concatenation of the previous dog stages column names and standardize the names
- Merge the three datasets to have one single master dataset using the column “tweet id”

Storage

I stored the final master data frame into a csv file with the name “**twitter_archive_master.csv**'' having the number of rows and columns as 1974 rows and 26 columns.