WeRateDogs - Twitter Data

This project is majorly to test the data wrangling skills of students taking the Udacity data analyst nanodegree in; data gathering, cleaning, accessing cleaned data and creating visualization of the data to get meaningful insights.

The following steps were taken for this project:

● Gathering data

● Cleaning data

● Accessing data

● Storing data

● Visual analysis of stored data.

Gathering Data

The First step in this phase was to download the archived data, which is a given csv file, then I named it **“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. I read through the API pseudo-code to completely understand the code before I proceeded with the project. After which I created three different dataframes using pandas for the three files described above. The dataframes are:

● **archived_df** - from "twitter-archive-enhanced.csv". It gives information on basic tweet data such as tweet id, timestamp, and the tweet itself; and other details were extracted from it, such as the dog’s name and image.

● **img_prediction_df** - from image predictions file. It contains information such as the precision in range between 0 and 1 for each prediction).

● **infotweets_df** - from the twitter-json file. It contains information like tweet_id, no of retweets and no of favorites, etc.

Assessing the data

The tables were assessed 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 misspelling 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, the datatype for each column were checked for irregularities.

● Further assessment was carried out on some columns based on findings from the steps above.

The data after being assessed was scrutinized to figure out issues around quality and tidiness and then cleaned, they are listed below

Quality

● The archived_df table has columns with missing values

● Errors in dog names. Some dogs had names like - 'just', 'life', 'mad', 'my', 'not', 'officially', 'old', 'one', 'quite', 'space', 'such', 'the', 'this', 'unacceptable', 'very', 'infuriating

● Timestamp in archived_df table is of datatype object instead of datetime

● Outrageous and inconsistent values in rating numerator and denominator

● Some ratings have a zero numerator and no name

● The archived_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.

● 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

● Comparing both img_prediction_df and infotweets_df to archived_df, we can see that they both have incomplete tweet ids unlike archive_df

● Retweeted_status_timestamp in archived_df table is of datatype object instead of datetime

Tidiness

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

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

● All redundant columns are removed (e.g the single dog stage columns once I have the merged column, and the retweet columns after they are cleaned).

Cleaning

The following steps were carried out to clean the data;

- All the datasets were copied to a different dataframe to avoid mistakes in the original datasets.
- Row with zero ”0” value ranking denominator was removed.
- Timestamp and retweeted_status_timestamp were initially recorded as strings, but were converted to datetimes.
- Columns dealing with prediction in the prediction_df were renamed to a more self-explanatory name
- Records with no names and zero numerator rating were removed
- All none values in the datasets were replaced with NaNs to properly represent they are missing values.
- Dog stage columns were concatenated to have a single column having the information in one.
- 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 dataframe were merged together using inner on the tweet_id as common ground.

Storage

I stored the final master dataframe into a csv file with the name “twitter_archive_master.csv" having 1992 rows and 26 colums.