Introduction

The idea behind this project is to work with provided datasets using the appropriate analytical tool, python; to generate insights.

Data Description

- Enhanced Twitter Archive (twitter-archive-enhanced.csv): This is a downloaded copy of WeRateDogs Twitter archive dataset that is made available by Udacity. It contains about 5000+ tweets that rates people's dog with a humorous comment about the dog but only a total of 2356 tweets with ratings are in the dataset.
- The tweet image predictions (image-predictions.tsv): using neural network, this dataset holds the outcome of the model classification of the breeds of dogs with different degree of confidence of the predictions.
- Twitter Data Api(tweet_json.txt): Using python's Tweepy library, I was able to extract via API each tweets' entire set as a JSON file and store them in my working directory as tweet_json.txt.

Of the back of this analysis, this report is prepared to briefly outline the steps followed to get this data analytics task done.

Project overview details

The tasks of this project are as follows:

- **Gathering data:**

  Using Python's Pandas packages, I was able to read the saved files (twitter-archive-enhanced.csv and image-predictions.tsv) into memory using the appropriate methods (pandas.read_csv).

  Using python's Tweepy library, I was able to extract via API each tweets' retweet count and likes for the entire set of tweets_ids provided as a JSON file and store them in my working directory as tweet_json.txt. With python's Pandas I was able to read the file line by line into memory and save them as pandas DataFrame. This data contains tweets_ids, retweet_counts and favourite_counts.

- **Assessing data:**

  While assessing the provided data using both visual inspection and programmatical analysis using different methods such as info, value_counts etc. , I was able to detect the following quality and tidiness issues which would greatly impact my analysis if not properly taken care of:

  - Enhanced Twitter Archive: The Enhanced has the following data quality issues.
    - The name's column has 745 dog names missing.
    - 88 records have stop words as name some which are 'a' occurs 55 times etc.
    - There are 1976 observations where there is no dog stage (doggo, floofer, pupper and puppo)
    - 14 observations have dog stage into 2 groups.
    - 380 observations have dog stage not missing and, in this case, doggo is doggo where other classes are missing etc.

        - The **expanded_urls** column has a lot of duplicates.
        - The **expanded_urls** has 362 cases where observations are either duplicates, missing or URL provided is not a twitter link.
        - The **expanded_urls** has 303 observations where tweet_id is different from the id in the provided URL link.
        - Missing observations in the dog stage were captured as a string with 'None' as an entry rather 'NaN'.
        - There are 5 cases where **rating_numerator** column is different from what was provided in the text.
        - There are 23 cases where observations in the **rating_denominator** column are not equal to 10.
        - There are cases in the data where **rating_numerator** are outliers, i.e., the data point is either larger or smaller than what is expected. A good example is where an entry has 1776 which is more than thrice the expected value.
        - There are cases where one **jpg_url** is linked to more than one tweet_id.
    - Image-predictions
        - There are cases where one jpg_url is linked to more than one tweet_id
        - The prediction columns are not consistent while this is expected as the neural network would behave differently, we need to have a way to filter out instances where the given observations are predicted to be non-dog.
    - Twitter Data Api
        - No data issues discovered with this dataset.
- **Data Cleaning:**

  The Define, code and test approach approaches were followed to clean the data quality and tidiness issues after creating a copy of the original given datasets.

        - The observations with missing names were removed.
        - Duplicated expanded_urls, missing URL links or where URL is not a Twitter link were removed.
        - To collapse the 4 (doggo, floofer, pupper and puppo) columns into a dog stage column.
        - Extracted the rating_numerator where what was given was different from what is in the text to ensure consistency.
        - Convert timestamp to datetime format instead of integer.
        - Ensuring that all entries in the rating_denominator equals 10.
        - Converting the tweet_id to string.
        - Remove stop words from the name column.
        - Removing unwanted columns that may not be useful to the analysis.
        - Removing observation that are too large or too small to the other data points (outliers).
        - Created a logic to filter out non-dog images.
        - Created a logic to create a new column for the predicted images.
        - Merge the datasets into a useful dataset.
- **Storing the cleaned data**
    - A copy of the cleaned and merged dataset has now been saved to working directory as a flat file with a .csv format (twitter_archive_master.csv).