Wrangle and Analyze Data

The purpose of this project is to demonstrate the knowledge and skills acquired in the Data Wrangling and Analysis chapter of the Data Analyst Nanodegree offered by Udacity. The dataset used for this exercise is the tweet archive of Twitter user known as WeRateDogs.

Gathering Data

- The “twitter-archive-enhanced.csv” file was downloaded from the Udacity Data Analyst Nanodegree Project website. The file was read and saved as “twitter_archive”.
- The “image_predictions.tsv” file was downloaded using the requests library in Python using the url provided in the project. The file was read and saved as image_predictions.
- The tweet IDs from the WeRateDogs archive was queried using the Twitter API for each tweet’s JSON data using Python’s Tweepy library. Each tweet’s entire JSON data was stored in a filed called “tweet_json.txt” which comprised of id, retweet_count, and favourite_count. This part of the data gathering process was the most challenging and consumed the maximum amount of time. After the simulation was completed, the personal API keys, secrets and tokens were removed from the code.

Assessing Data

- Each of the three dataframes, “twitter_archive, “image_predictions” and “tweet_info” was analysed visually with the head(), tail(),info(), describe(), value_counts(), null(), duplicated() functions etc.

- The following Quality issues were detected in the “twitter_archive” dataframe, which needs to be cleaned.
  (i) Change timestamp in strings format to datetime format
  (ii) Remove rows with NaN values in the names column
  (iii) Change the missing values in name column from 'None' to 'NaN'
  (iv) Replace the dog name "O" with "O'Malley"; Incorrect names such as 'a', 'an', 'this', 'not', 'one' etc were also replaced with 'None'.
  (v) Change datatype of rating numerator and denominator to float; Reextract values in text for numerator in the situation where the numeral after the decimal point has been extracted in the text.
  (vi) Remove extra characters after '&' in the text column
  (vii) tweet_id is defined as an 'integer' whereas it should be a string.
  (viii) Dog Stage is converted to category.
  (ix) In row 385, the rating is erroneously recorded as 24/7.
  (x) Replace the underscore in p1, p2 and p3 columns with a space.

- The following tidiness issues were detected which needs to be cleaned.
  Duplicated data is present in the form of retweets in the twitter archive table (ii) Dog stages in the twitter archive table is present in 4 columns (doggo, floofer, pupper and puppo), and should be merged into one column. (iii) Each of 'tweet_info' and 'image predictions' tables should be merged into the 'twitter_archive' dataframe.

Other issues which were also assesed and cleaned:
(i) Remove html tags in source column
(ii) Display full width of the text column for entire text to be seen

- A copy of each of the files was created, keeping the original file intact.

Cleaning Data

- This part of the project comprised of Programmatic cleaning of data using Python in Jupyter Notebook. The steps in the cleaning part of this process were divided into Define, Code and Test. Each of the issues identified in the Assessing part of the project were first defined as to what the issue was and what needed to be done to rectify it, a code in Python was written to remove the issue identified, and once the code ran successfully, a test was done to see of the issue identified had been removed. This is an iterative process and was run several times till the desired outcome was achieved.
- One of the challenges in the cleaning process was to melt the dog stages into one column instead of four columns as presented in the original twitter_archive table.

Finally, when the data was successfully cleaned, the merged twitter_archive dataframe was saved as “twitter_archive_master.csv” and used for visual analysis using matplotlib.pyplot package in Python.