Udacity Micro-degree Project 2: Wrangle report of the WeRateDogs! Twitter dataset analysis

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings always have a denominator of ten. The archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. In this exercise, I followed the normal flow of steps in data wrangling, beginning with gathering the data, assessing the data, and then cleaning the data. The steps are further explained in the headings below.

Before gathering the data, all the important libraries that will be used in the analysis were loaded into the notebook.

Step 1: Gathering the data

Three important pieces of data were gathered from various sources unto the notebook. These included the **'Twitter_archive_enhanced.csv'** file**,** which I downloaded from the Udacity website into my machine, which I then read into the notebook using Pandas' '**read_csv'** function by making a call to the file path. A glance (using '**df.shape**' and '**df.head**') revealed that the data frame had 2,356 rows and seventeen columns. I named the file '**dogs_archive**.'

The second file I downloaded was the '**image_predictions'** file, which was downloaded programmatically, using the Requests library, with the provided URL. I then created a new directory, where I saved the tsv file, using the **os.mkdir()** function of the OS library. I then read the tsv file into a new data frame which I called '**df_predictions**' using the '**read_csv'** function.

The third file I loaded was the json.txt file, which I downloaded into my device and saved into a data frame which I had named '**tweets_json**.'

Step 2: Visual and programmatic assessment

This involved the following sub-steps:

- As part of visual assessment, I made a call to the head of each data-frame, to physically check the type of data in each column, the number of columns, columns which need to be merged, those whose data are in the wrong format and those which needed to be deleted.
- Programmatic assessment involved checking the less obvious information about the data in each data-frame, concerning the value counts of each variable, checking for duplicated values, unique values, and the descriptive statistics. Both steps also involved looking for quality and tidiness related issues that would be corrected in the data cleaning stage.

Step 3: Cleaning the data

- Data cleaning involved making copies of the data-frames, checking for and eliminating null values, removing duplicates in the form of retweets, dropping NAs, replacing wrong values, and merging the numerator and denominator rating columns to form the cuteness_rating column.
- This part also involved merging of the three cleaned data frames and saving the resultant data frame in a csv file as "**twitter_archive_master.csv**."