Data Analyst Nanodegree

Wrangle Report

- **Introduction:**

In this project, I wrangled data from twitter account which is @dog_rates and known as WeRateDogs. I used the file which Udacity provided. There are three files with different formats (csv, tsv, text, and json).

These are steps that describe how I wrangle data:

- **Gathering Data:**
  These are the files that I used.
  - twitter-archive-enhanced.csv
  - image-predictions.tsv
  - tweet_json.text

- **Assessing Data:**
  I used two types of assessment:
  - Visual assessment: I displayed data by using Excel.
  - Programmatic assessment: I used pandas' functions such as head, tail, describe, info, shape, methods, and libraries such as tweepy, numpy, re, json, timeit, and matplotlib.

  I defined quality issues and tidiness issues.

  Quality:

  - Missing value in these columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.
  - Inaccurate datatype for tweet_id, timestamp, and dog_stage columns.
  - Tweets without images.
  - Text of the source column is unreadable.
  - Inaccurate names of name column.
  - Uppercase given in p1, p2, and p3 columns.

- Unneeded columns for anayst (rating_numerator, rating_denominator, img_num, p1_dog, p1_conf, p2_dog, p2_conf, p3_dog, p3_conf, and jpg_url).
- Text column contains links.

Tidiness:

- Dog stage variable in 4 columns.
- Join tweet_df_clean and image-predictions.tsv with twitter-archive-enhanced.csv
- Rating variable in 2 columns.

- **Cleaning Data:**
  I defined every issue, wrote solving code, and tested the solving to make sure the issues have solved.

- **Storing, Analyzing, and Visualizing Data for this Project:**
  I saved the clean DataFrame(s) in a CSV file to twitter_archive_master.csv. I provided 4 insights with visualizations.

- **Conclusion:**
  In this project, I used Jupyter Notebook and Python programming language to wrangle data. These are output files:
  - wrangle_act.ipynb: contains the codes of gathering, assessing, cleaning, storing, analyzing, and visualizing data.
  - wrangle_report.pdf: contains the documentation of data wrangling steps: gathering, assessing, cleaning, storing, analyzing, and visualizing data.
  - act_report.pdf: contains the documentation of analysis, insights, and visualizations.
  - twitter-archive-enhanced.csv: given file.
  - image-predictions.tsv: given file.
  - tweet_json.text: given file.
  - twitter_archive_master.csv: contains the cleaned data.

- **Resources:**
  - https://www.w3schools.com/python/python_file_open.asp
  - https://chrisalbon.com/python/data_wrangling/pandas_saving_data frame_as_csv/
  - https://stackoverflow.com/questions/24848925/how-to-install-tweepy-with-anacondas-and-easy-install
  - https://stackoverflow.com/questions/45899613/divide-certain-columns-by-another-column-in-pandas
  - https://stackoverflow.com/questions/35439613/python-pandas-dividing-column-by-another-column
  - https://pandas.pydata.org/pandas-docs/stable/user_guide/merging.html
  - https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.str.contains.html
  - https://stackoverflow.com/questions/29337123/how-to-change-the-range-of-the-x-axis-and-y-axis-in-matlibplot
  - https://stackoverflow.com/questions/40699778/query-a-data-frame-by-multiple-columns
  - https://www.youtube.com/watch?v=YPItfQ87qjM
  - https://www.youtube.com/watch?v=xvpNA7bC8cs