Introduction

The wrangle and analyze project is part of Udacity Data Analysis Nanodegree program. It is about wrangling data from multiple sources related to tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The goal is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations, Additional gathering, then assessing and cleaning is required for worthy analyses and visualizations.

Body

- ### Gathering Data for this Project
- Gather each of the three pieces of data as described below in a Jupyter Notebook:
- The WeRateDogs Twitter archive. Which was a giving file ready to download.
- This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- I was supposed to use API from twitter to get the data from Twitter but my request wasn't approved and I used the tweet_json.txt file.
- ### Assessing Data for this Project

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues.

Quality issues

- Remove retweets.
- Remove tweets without images.
- tweet_id is an integer which doesn't need calculation.
- timestamp is object which should be datatime type.
- Remove HTML tags from source.
- Fix denominators other than 10.
- Drop unneeded columns .
- Calculate Rating value.

Tidiness Issues

- Combine dog stage columns to one column.
- combine all data frames.
- There are some columns in this dataframe will not be needed for analysis.

Cleaning Data for this Project

- Clean each of the issues documented while assessing. This cleaning is performed in `wrangle_act.ipynb` as well. The result should be a high quality and tidy master pandas DataFrame .

Conclusion

In conclusion, the cleaned DataFrame Stored in a CSV file with the main one named twitter_archive_master.csv.