WeRateDogs Wrangle Report

This project was focussed on gathering, assessing, and cleaning data through python. It involved gathering data from three different sources by using manual and programmatic methods. Once the data was gathered it needed to be assessed and cleaned. Using pandas, numpy, and other python libraries the data was loaded into data frames where it could be analyzed for quality and tidiness issues. After identifying these issues they were cleaned and tested.

Gather

The following data was gathered for this project.

- twitter_archive_enhanced.csv - This file was provided through the project details by Udacity as a manual download.
- image_predictions.tsv - This file was provided by Udacity as a url. It required using the Requests library to programmatically download the file.
- tweet_json.txt - This data was available through the Twitter archive using Tweepy. It required the following:
    - Create a Twitter account and sign up for a developer account.
    - Use the Tweepy library to programmatically access the api's.
    - Authenticate to the Twitter archive API.
    - Add logic to work within the API's request limits for a developer account.
    - Parse the json returned and write the information to a file.

Assess

The following steps were performed on the gathered data.

- Load the three files gathered into data frames.
- Perform visual analysis on the data frames using pandas, excel, and sublime.
- Document areas to cleanup from the visual inspection.
- Perform programmatic inspection of the data using pandas.
- Document areas to cleanup from the programmatic inspection.

Issues found:

Clean df_twitter_final data frame

- Define - Quality: Update the data type of timestamp to a timestamp instead of string
- Define - Quality: Remove rows that are replies to other tweets
- Define - Quality: Remove rows that are retweets
- Define - Quality: Remove reply and retweet columns from the dataframe

- Define - Tidiness: Dog stages need to be combined into one column
- Define - Quality: Change rating_numerator and rating_denominator to floats
- Define - Quality: Incorrect values in rating numerators
- Define - Tidiness: Join all tweet information into one dataframe
- Define - Quality: Drop unused twitter columns
- Define - Quality: Drop all rows that do not have retweet or favorite values

Clean df_image_pred_final data frame

- Define - Quality: Remove _'s from the p1, p2, and p3 columns
- Define - Quality: Capitalize the first letter in each word
- Define - Tidiness: Determine the best prediction of dog type
- Define - Quality: Remove unused image prediction columns
- Merge final datasets into one dataframe for graphs
- Create prediction categories
- Define - Quality: Drop NA rows in dataframe

Clean

The following steps were performed to cleanup the issues found.

- Write tests for all cleanup items.
- Write code to correct the issues.
- Leverage classroom materials and various websites for the correct syntax to use in cleaning the issues.
- Once the issues were cleaned up the data frames were saved to files.

Resources:

- http://docs.python-requests.org/en/master/user/quickstart/#response-content
- https://stackabuse.com/reading-and-writing-json-to-a-file-in-python/
- https://stackoverflow.com/questions/7370801/measure-time-elapsed-in-python
- http://docs.tweepy.org/en/v3.2.0/api.html#API
- https://developer.twitter.com/en/docs/basics/rate-limiting
- https://stackoverflow.com/questions/4432208/how-does-work-in-python
- https://www.epochconverter.com/
- https://www.pythoncentral.io/pythons-time-sleep-pause-wait-sleep-stop-your-code/
- https://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_csv.html
- https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html
- https://pandas.pydata.org/pandas-docs/stable/generated/pandas.melt.html
- https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.merge.html
- https://chrisalbon.com/python/data_wrangling/pandas_join_merge_dataframe/
- https://pandas.pydata.org/pandas-docs/stable/options.html

- https://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.str.contains.html
- https://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.str.title.html
- http://jonathansoma.com/lede/foundations/classes/pandas%20columns%20and%20functions/apply-a-function-to-every-row-in-a-pandas-dataframe/
- https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.apply.html
- https://pandas.pydata.org/pandas-docs/stable/generated/pandas.to_numeric.html
- https://matplotlib.org/gallery/statistics/barchart_demo.html
- https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.sort_values.html
- https://stackoverflow.com/questions/25146121/extracting-just-month-and-year-from-pandas-datetime-column-python
- https://stackoverflow.com/questions/19377969/combine-two-columns-of-text-in-dataframe-in-pandas-python
- https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.replace.html
- https://stackoverflow.com/questions/12625636/python-string-function-to-strip-the-last-comma
- https://pandas.pydata.org/pandas-docs/stable/options.html
- https://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.str.extract.html
- https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.astype.html