Data Wrangling Report

Project Overview: This report consists of my data wrangling steps that I performed in a Jupyter Notebook. It demonstrates why and how I wrangled my data, leading to my analyses and visualizations using Python and its libraries.

Gathering Data

- Twitter Archive Enhanced
  - This dataset contains the tweet archive of Twitter account WeRateDogs. The file contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. It was downloaded and saved as *twitter archive enhanced.csv* from Udacity's server.
- Image Predictions
  - This dataset contains tweet image predictions and was utilized to determine what kind of breed a dog is. In detail, neural network technique was utilized to decipher each tweet and categorize them accordingly. The file, *image_predictions.tsv*, is hosted on Udacity's server and I downloaded it programmatically using the Requests library.
- Twitter API Data
  - This dataset contains additional data such as retweet count and favorite count which were omitted in the Twitter Archive Enhanced dataset. I created a Twitter developer account to get the credentials for my API calls and it was done programmatically (Tweepy) in Python. With the tweet_id contained in the Twitter Archive Enhanced file, I queried the API to get the entire stored JSON data for those tweets. The query took about 30 minutes. The gathered data was stored in JSON by Twitter so I wrote the JSON data to a *tweet_json.txt*

Assessing Data

- Visual Assessment

  - I assessed the three files that I gathered (underlined above) first by using Jupyter Notebook. Prior to looking at the files, I named them as *archive*, *images*, and *api_data* for easy access and usage. Looking at the dataframes allowed me to get a glimpse of what I'm dealing with and how I can plan for my cleaning process.
- Programmatic Assessment
  - This component followed the visual assessment where I used both Python and Pandas functions such as groupby, info, value_counts, duplicated, query, etc. to see the details of these dataframes. Further, this component is where I first saw many quality and tidiness issues and I noted them in my notebook. These are well documented in the Assessment Overview where I state both the quality and tidiness issues and how I plan to fix them.

Cleaning Data

- Before jumping into cleaning my datasets, I first separated the type of issue (quality or tidiness) respectively and listed them as an agenda. I also made copies of the original datasets that I gathered so that I can revert changes easily when I run into programming issues.
- Each issue follows the following format.
  - 1. Address the issue and its type
  - 2. Define the issue
  - 3. Provide code that will fix the issue
  - 4. Test to demonstrate that it's fixed and save

Conclusion

This project gave me the opportunity to utilize data wrangling skills I've acquired and use them in a dynamic environment. It gave me an interesting perspective on how data analysis skills can be utilized to answer interesting questions. I learned that in real-world examples, datasets are often very messy and even daunting to attempt analyzing. And with that said, the data wrangling part is the most challenging portion of the data analysis process. However, once you clean your data well enough, you are able to be more creative with your insights and visualizations.