Garett Hansen – Data Wrangling Project

The purpose of the data wrangling project was to leverage python and several different libraries (pandas, numpy, etc) in order to gather, assess, and clean three sets of Twitter data, specifically tweets from the group WeRateDogs.

Gathering

1. The first dataset, Twitter Archive, is a .csv file provided by Udacity. This was imported manually into the notebook.
2. The second dataset, Image Predictions, is a .tsv file and was downloaded programmatically from a link provided by Udacity.
3. The third dataset, Twitter API Data, was accessed by querying the Twitter API by using the Python Tweepy library. Using the tweet id as a reference, the json Twitter data for each tweet was downloaded and stored in the file tweet_json.txt. This data was then stored as a pandas dataframe.

Assessing

Next, each of the three datasets was assessed, both manually and programmatically. This included looking at sample rows in Jupyter Notebook and reviewing the full dataset in Excel as viewing an entire data set in a notebook can be difficult. Each dataset was assessed programmatically by using commands such as .info() (to inspect data types), .duplicated() (to check for duplicates), and .describe() (to view summary statistics).

I then compiled a list of both quality and tidiness issues with each dataset. This list helped me approach the cleaning portion of project in a clear, systematic manner.

Cleaning

Before actually cleaning any data, the first step I took was to make copies of the original dataframes. Doing this made sure I kept the original datasets intact and provided me an insurance policy if I needed to revert back to the original layout.

I then started at the top of my list of quality issues and worked my way down. A large part of my cleaning efforts came from removing rows and columns that were unnecessary, inaccurate, or otherwise. Despite hundreds of rows and several columns being removed, we were still left with over 2,000 rows to analyze.

One of the most challenging, yet rewarding, aspects of the cleaning process was combining the four dog type columns into a single column. This was challenging as some rows had multiple values; however, after completing the process it was very satisfying to see the one column with its appropriate value(s) listed for each row.

Other Remarks

In my opinion, this was a fairly difficult project. Having no prior experience with querying data via API, this was also a very challenging part of the project. It took me several attempts to get the code right, but the moment where I realized the call was working was a very cool moment to experience.

Overall, I enjoyed working on this project and appreciate the skills I learned along the way.