Data Wrangling Report
By James Cajuste

February 10, 2019

I found the data wrangling project to be very challenging. I decided to complete my project in the Project Workspace using the Jupyter Notebook provided there for simplicity. After setting up a developer account, I was having trouble accessing the data so I followed the directions for accessing the Twitter data without actually creating a Twitter account. I manually downloaded the two Twitter text files, CSV and TSV files then uploaded the files to the Jupyter Notebook Workspace. Once I opened the Jupyter Notebook file called wrangle_act.ipynb, I imported all the necessary libraries to continue the data gathering phase.

During the data gathering phase, I had no problem reading the Twitter archive and image prediction as well as querying the Twitter API data; however, storing each tweet's entire set of JSON data in a file called tweet_json.txt file was very frustrated because my Jupyter Notebook kept crashing as I was running the provided code from the file called twitter-api.py.txt, which kept overwriting the file called tweet-json.txt. However, my Mentor, Septi Rito T., informed me to avoid running the provided code so that the tweet-json.txt file doesn't get overwritten, which might crash my Jupyter Notebook. As a result, I had to "Reset Data" several times in the Jupyter Notebook Workspace in order to successfully complete the data gathering phase.

Once completed the data gathering phase, I proceed to the data assessment phase, which was the least challenging phase. I did visual assessment of all three dataframes then a programmatic assessment of them again in order to determine quality and tidiness issues. And then I documented quality and tidiness issues for the three datasets in the Jupyter Notebook as basis for my data cleaning phase.

I started the data cleaning phase by making cleaned copies of the three datasets. I highlighted the dataset name, defined the quality issue to fix, coded the fix then tested it for success. The cleaning consists mostly of addressing missing data and mislabeled information, which was predominantly found in the WeRateDogs Twitter archive. I had to convert columns to a proper data format, primarily changing the timestamp data into datetime objects, tweet_id from a number into a string and the rating columns into float objects. I also had to address quality issues in the Prediction columns of the Image Prediction dataset. I had to use pandas functions like str.replace() and str.title() to remove the underscore between the words and capitalize the letter in each word to make a more cohesive table. Lastly, I had to inner join all three datasets into a final document containing all relevant information. For this task I used the pandas library using the pd.merge() function.

In summary, this project may have been my biggest challenge so far if you include the data storing, analyzing, visualization as well as the two required reports. Overall, I am pleased with what I have learned so far so I intend to keep practicing after completing the Data Analyst course.