Data Wrangling

Project 2

We Rate Dogs Data.

Gathering Data Phase:

I started my project by downloading the ‘twitter-archive-enhanced.csv’ file manually. Then created a folder named 'image predictions' before I downloaded ‘image predictions tsv programmatically from Udacity's server using the requests library. Next I wrote it into image predictions tsv and later convert it into a pandas DataFrame . I couldn’t obtain twitter API, so I have to make use of the link provided on the additional resources page, tweet_json.txt.

I later wrote code to ensure each tweet’s data was written in a new line. After the query was completed and all the data was written in the text file, I read the text file line by line, obtained each tweet’s information (tweet ID, retweet count, favorite count, and followers count) using the json library, and appended the information into an empty list.
Finally, I convert the list of dictionaries to a pandas DataFrame and saved it into ‘twitter_df’.

The datasets were visually and programmatically inspected to discover both the quality ad tidiness issues.

Now I had 3 datasets gathered in total which I had to assess for quality and tidiness issues and clean. Undoubtedly the twitter archives table had most of the issues. This was because it was the largest of the 3 tables and had a lot of features regarding the tweets like tweet text, date and time of the tweet, etc. A lot of the issues could just be spotted by the naked eye thanks to pandas functions like head, tail and sample. Others required a little more analysis, mainly through summaries or filtering out certain sections of the data and evaluating the features. The info and value_counts functions were frequently used for the same. Most of the tidiness issues involved joining of the tables and melting certain features into a single column.

The final part was the most code-oriented part which was data cleaning. For each operation, I used the define, code and test format wherein I first defined the operation I wanted to perform, then wrote the code for it and later confirmed the changes on the dataset to establish the success of the operation. Pandas library was handy at this face, some in isolation and some were chained. A lot of the operations involved extracting data from certain features and replacing the inaccurate or incomplete column data with that data. A lot of unwanted rows and columns were dropped.

Although this project was somehow difficult, I was able to make sense from a very dirty data. The project got me to work on major data quality and tidiness related issues and then cleaning up those issues to make the datasets ready for analysis.