Wrangle Report

By Necmi KILIC

The data wrangling project was very challenging and I got a lot information about the data gathering process and the Twitter API.

I gathered data from three different sources for this data analysis.

1) WeRateDogs csv twitter archieve given by Udacity exclusive Access to their Twitter archive. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets.
2) Each tweet image was run through a convolutional neural network to analyze the images of dogs and correctly identify their breeds. I downloaded the tsv file by using Pyhton REquests library.
3) And finally, using the tweet IDs from the WeRateDogs archive I queried the Twitter API for each tweet's JSON data using the Python's Tweepy library I stored each tweet's entire set of JSON data, which I would later use to analyze the tweet's retweet and favorite counts.

The data gathering process for this project was my greatest challenge, particularly querying the Twitter API. I spent a few days to creating consumer information to connect to twitter account. It was very new to me. I searched for a lot of websites from Google to accomplish it. I completd it after a several fails.

Once I had successfully gathered all the data, I copied the files for the assessment and data cleaning processes. I evaluated the dataframes looking for quality and tidiness issues and then set about fixing them. I began the cleaning process by addressing missing data and odd column values(like dog names. I then converted the timestamp column to datatype type because it was necessary to analyze and plot it.

For tidiness, firstly I merged 4 columns in archieve which represents dogs stages. 1 column was enough for it.

The final step in the data cleaning process was 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.

The next stop was storing the cleaned data by usind Pandas to_csv command.

And finally I analyzed and plotted data after calculating rating_ratio (rating_numerator/rating_denoinator) to see rating over time.The other analyze was based on retweets and favourites counts over time.

In summary, this project was my biggest challenge to date, specifically using the Twitter API and Python Requests library to gather the JSON data. Overall, this project was completed successfully and I'm extremely pleased with the new skills I acquired.