Wrangle Report

By Aisulu Omar

The goal of this project was to gather real-world data from different sources, clean it and provide exploratory analysis. The datasets are the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. I had the access of Twitter archive file in csv, enhanced twitter archive with images and additional data via the Twitter API.

My project included 3 steps:

- Gathering Data
- Cleaning Data
- Analyzing Data

Step 1: Gathering Data

I had to gather three different datasets. First dataset (df) was provided by Udacity. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets. The dataset was provided in csv file and it was easy to gather it.
Second dataset (tweet_df) was gathered through querying Twitter API for each tweet's JSON data using Tweepy library. Gathering second dataset was the most challenging part. Luckily, Udacity provided supporting materials where I had an example of this work.
The third dataset (images) contained images and image predictions. The file was downloaded programmatically using Requests library and URL information.

Step 2: Cleaning

Once I gathered all the datasets, I created copies of dataset to perform my cleaning process. I evaluated datasets and before making any changes, I documented all the quality and tidiness issues. I define every issues and tested it after to make sure issues are corrected. I defined 8 quality issues and 8 tidiness issues. Quality issues cleaning included: removed retweets, since those tweets were not original, correcting misspelling, dropping unnecessary columns, renaming columns. Tidiness issues cleaning included: converting into correct format, separating dates and time and combining different stages of dogs into one column. Once I finished cleaning of all three datasets, I performed inner join to combine all three datasets into one.

Conclusion

Data wrangling is really important skill for Data Analysts and Data Scientists. In real world, data usually comes raw, with a lot of quality and tidiness issues. Accessing data through API gives a lot freedom in data exploration. There are many sources that can be interesting to analyze. I

used stackoverflow and other resources to finish this project. Accessing JSON data was a really challenging part, I had to rerun my code and fix several times.