Wrangle Report

Introduction

The project purpose is to get practice with real data, the data was provided by Udacity, this data consists of many WeRateDogs tweets, WeRateDogs is a page on Twitter that people rates dogs with a curious rate, as 13/10, 14/10, 15/10, but like many other datasets, this data has many problems of tidiness and quality, so to solve that I followed some steps that I learned on Udacity.

Gathering Data

The data for this project consists of three different dataset that were obtained as following:

1. twitter-archive-enhanced.csv: This file was provided by Udacity and founded in Udacity's resource, I downloaded and storage manually.
2. Image_predict.tsv: This file is storage in Udacity's servers and I get using the 'request' that is in Python library to download the file programmatically using a URL.
3. tweet_json.txt: This file was gathered using Twitter API, and using library tweepy that allow us access some resources of API, in my cause focused to get the number of favorite and number of retweets using tweet ids that are on twitter-archive-enhanced.csv.

Accessing Data

In access, step is a part of the project that you need to be focused to analyze and understand what you have in your hands, and identify problems. I spent a lot of time in this part, but it's an essential part, I found tidiness problems like unnecessary three datasets so join their twitter-archive-enhanced with my tweet_json is a good way. Some quality problems like many names aren't really names and some expanded URLs are missing.

Cleaning Data

Finally, the final step of wrangling, now every problem is fixed up, to make changes with safety is a good practice to make a copy of the dataset, after that, I could try to fix, in most of the problems I solved with methods of Pandas like merge, to_datetime, dtype and was successfully fixed up, but in one case I needed to make a function to handle dogs stages and all done.

Conclusions

How I said wrangling data is an essential aspect of data analysis, I gained many skills with this project, the programming language that I use is Python because python has a lot of libraries to data analysis and because I feel more comfortable with.

In gathering data was great to learn about tweepy, reading how this works, spent time but open your eyes to possibilities.

Other steps I used a lot of forums like stackoverflow to get some help.

With this skills gain, I got to do a good analysis and visualize confident data to make insights.