Wrangle Report

By Ahmed Sayed

September,2020

This report illustrates the Wrangle and Analyze Udacity project, we could classify this project into 4 steps

Data Gathering

At this step I gathered data from different sources and formats:

The first file is the twitter-achive-enhanced.csv file, I downloaded this file manually, uploaded it to the jupyter notebook in the project workspace and then stored it in a pandas data frame.

The second file is the image-prediction.tsv file which I downloaded programmatically and then opened it in a panda's data frame.

The third file was a json file that I downloaded using twitter's API, then extracted three columns from it and made a data frame from them to later use them in my analysis.

Data Assessing

At this step I searched visually and programmatically for any potential issues in the three data frames that I have using various python functions like head(), info(), describe(), I followed the rule of detect and document at the three different data frames, but actually I didn't document all the issues I saw I only documented a number of issues that satisfy the project requirements.

Data Cleaning

Mainly this step is to fix the issues that I documented at the previous step in the data wrangling process, I looked at each issue, then I defined a solution for it, then I converted this solution from English language to python language which the compiler could understand , and then finally I tested the solution to see whether the problem is solved or not, I iterated the data assessing step more than one time at this step because I figured out some issues that I have to fix for the sake of the analysis but I forget to document, I fixed some issues manually like changing a number in a specific location, and others programmatically like dropping a column from a data frame or change its data type.

Data Storing, Analysis and Visualization

After cleaning all the issues in the data I merged all the three data frames to have only one clean and tidy data frame and I stored it in a csv flat file with the name twitter_archive_master.csv, after storing this file I used its data to analyse it and look for patterns, or get an insight or even helps us answer some questions about the data.