Achraf Safsafi
DSC 540
project summary

In my project, I learned three fundamental approaches to extract data on your own and then create my dataset from three different sources of information. So, after selecting my sources. I did the first extracting task that was the easiest part of the project. I imported the flat file into Python using the Pandas library and then store it in a data frame. Next, I used web scraping to access a website's content and create my second dataset using some Python libraries like urllib and BeautifulSoup. Then, I generated my third dataset using the Application Programming Interface (API) and some Python Libraries such as requests and JSON.

After completing each data extracting task, I needed to clean those collected data using proper formatting and removing unnecessary signs and values in order to create a uniform and usable dataset before any analysis can be performed. for instance, if I find some data are missing, I could delete them if they don't represent an important percentage in the data set. However, I can calculate either mean, mode, or median of the feature and replace it with missing values.

After loading each cleaned dataset into SQL Lite as an individual table, I merged all of them in Python into one dataset, then I stored the resultant table in the database. Finally, I did my last part of the project, that is data visualization. I created some basic plots such as scatter plot, KDE plot, and heatmap using Seaborn.