Achraf Safsafi
DSC540
Project: Milestone 1

.

In my project, three different data sources are selected. Data sources have different file types of information. The first dataset will be extracted from https://ucr.fbi.gov/crime-in-the-u.s/2018/crime-in-the-u.s.-2018/tables/table-10/table-10.xls/view. Data are in the form of an Excel file. The data was collected, in 2018, by the FBI through the Uniform Crime Reporting (UCR) Program. The dataset contains 2357 rows of 12 variables. The data is a collection of various crimes by states and counties. There are two main types of crimes.and each type includes subtypes. According to the website, Violent crime is composed of four offenses: murder and nonnegligent manslaughter, rape, robbery, and aggravated assault. Then, property crime includes the offenses of burglary, larceny-theft, motor vehicle theft, and arson.

The second data set is 2018 data published on the US Bureau of Labor Statistics website,https://data.bls.gov/cew/apps/table_maker/v4/table_maker.htm#type=1&year=2018&qtr=A&own=0&ind=10&supp=0, the data is stored in the form of an HTML file. The table has 8 columns: County(all states), Annual Establishments, Annual Average Employment, Total Annual Wages, Annual Average Weekly Wage, Annual Wages per Employee, Annual Average Employment Location Quotient, and Total Annual Wages Location Quotient.

The last source will be an API source, I selected the US census bureau website, https://www.census.gov/data/developers/data-sets.html . It provides many data sets that are currently available via API. The website includes a large range of topics about social, economic, demographic, and housing characteristics of the U.S. population. So I can use these features by county later.

So, As we see, All three of the selected data sources are related by the county variable. In the next steps, I will clean the data that I will take from the various sources above. I will convert and merge it into a consistent format for improving its quality to be prepared for its use in business data analysis.