Project 2: IS 296 UMBC

Requirements:

1. Please identify a dataset from the public sources discussed in class (UCI ML repository; Physionet.org; Data.gov; Kaggle). Please choose a simpler dataset with a clearly available .csv file
2. Please identify a question you would like to investigate.
3. Provide data exploration steps in your jupyternotebook (such as extracting columns, manipulating arrays, grouping, histograms)
4. Provide an example function that performs a prediction task on the data
5. Provide a connection to the question you started investigating and your final evidence to prove or disprove your initial question

Example Task:

1. Two datasets:

Diabetes data by county: https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html

COVID 19 data by county: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports (Look at the last data file for the most recent deaths and cases)

2. Is there any association or correlation of deaths by county and diabetes cases by counties? How does population play a role (you could also look for population data and study the correlation between cases/deaths to population)?

3. For this task the exploratory analysis and processing you will do, is the key outcome.
   If you are able to conclude any findings that is also very useful. While studying correlations please look out for spurious correlations.

You can pick any other datasets. This is just one example.