DSC540 – Final Project Summary

I am using three different datasets which has relationship between them by type of heart decease or Attack, age groups & sex. It means that the dataset can be joined together by the standard columns and explored together.

In other milestones of the project & , I explored pull the data from Kaggle API/ chronic data.cdc.gov and apply data wrangling techniques that I have learned throughout the course.

Project MileStone 1 : Identified all the datasources and their relationship.

Project Milestone 2 : Understood and loaded Flat file datasource and explored it using various data cleaning techniques techniques.

Also Performed Deduplicating techniques, removing outliers, Data Formatting techniques.

Tried Renaming columns and that also helped me for my final assignment.

Project Milestone 3 : This milestone was all about exploring website data.

1. Used pandas, numpy, requests, bs4, urllib.request & urlopen python libraries. Each one served a very specific purpose for extracting data from a website (url = “https://www.sciencedaily.com/news/health_medicine/heart_disease/“)
2. How to extract information from webpages and use protocols.
3. How to use response and convert in table format.

4. How to handle error shows up as it thinks its a bot when you are exploring webpage.
5. I converted my data into a csv file and performed various operations on it to clean data.
6. Also applied grouping methods to aggregate and summarize data.

Project Milestone4:

1. For this milestone I have used Kaggle API to extract the data using API.

2. I have converted the returned data into JSON format for ease of access.

3. I have used renaming of column and also ran profile report to deep dive into data.

Project Milestone 5:

1.I am loading the data from all three sources into separate tables using sqllite.

2.I have to create summary tables to reduce the size of the dataset and I have also used ranking.

3. I have created few charts as described below:

-

DSC540 – Final Project Summary

- HighBP Vs SleepTime - This uses two datasets
- Scatter plot for BMI and MentalHealth
- Histogram of HeartDiseaseorAttack or SleepTime
- Boxplots of the Sex and SleepTime
- BMI Desnsity plot