Capstone Two - Project Proposal

Sam Domeier

6/16/2020

The idea that I have selected for the second capstone project is “Predicting heart disease rate in the US”. This dataset is provided by kaggle and can be located here. There are (3) files associated with this dataset, all of which are of type CSV. Originally the dataset was meant for a competition, so there is one csv file that will not be used during this project since there is no target column for the rows.

The overall goal of the project is to predict the rate of heart disease (per 100,000 individuals) across the United States at the county-level from other socioeconomic indicators. To predict this, the target column is identified as ‘heart_disease_mortality_per_100k’ and there are 34 features that will be used for creating this model. The original shape of the dataset is (3198, 35) - which includes the target column.

This data was originally sourced from the United States Department of Agriculture Economic Research Service (USDA ERS) by someone in the kaggle community. Each row in the dataset represents a US county, and the data covers two particular years (which are denoted in a categorical column by ‘a’ and ‘b’). Each county is identified by a number in the ‘row_id’ column.