Michael Wienczkowski

July 18, 2018

Data Analytics

Final Project

**Prevalence of Heart Disease Mortality in the United States**

**Problem**

Heart disease is typically ranked as the number one cause of death in the United States. Heart disease costs the United States USD 200 billion each year, which would include the cost of medications and health care services.[[1]](#endnote-1) Identifying who is primarily affected by heart disease mortality within the United States is one of the main problems that this project seeks to address. Health care is considered an important topic and heart disease is one of the main diseases affecting the lives of many people. Identifying those affected by the disease could help to hone the focus of treatment and improving health before it becomes an issue. The problem is also considered to be the leading cause of death worldwide, which is why it is a topic that has caught my interest. According to the CDC, heart disease is the leading cause of death for most racial/ethnic groups in the United States, which includes Caucasians, Hispanics, and African Americans.[[2]](#endnote-2) There are also several factors or lifestyle choices that can increase the risk of having the disease. One of the correlating factors that caught my interest is obesity/weight gain which is another primary issue in the United States. Other important risk factors that increase the chances of developing heart disease are diabetes, a poor diet, excessive alcohol use, smoking, high blood pressure, and physical inactivity.[[3]](#endnote-3)

**Approach**

This project would also provide practice regarding data manipulation using python and Jupyter notebook. A data set that includes the results of heart disease mortality in the United States could offer the opportunity to practice cleaning data, creating charts, and answering important questions after proper data analysis. I am also interested in using the datasets to identify who is more likely to be affected by the disease based on gender, race/ethnicity, and location. I am also interested to confirm the high rates of heart disease mortality based on specific ethnic groups. This project provides the opportunity to address the problem of heart disease mortality in the United States and to find correlations that could result in a successful predictive analysis.

**Dataset**

The data that used for the project illustrates the number of heart disease related deaths based upon location within the United States. The data can be viewed by race/ethnicity and gender. The races that have been included within the dataset are the following: White, Hispanic, Black, Asian and Pacific Islander, and American Indian and Alaskan Native. The type of values provided indicate the amount of deaths per 100,000 over the course of a three year average from 2013 to 2015. The data value itself is averaged for the year at a geographical level of counties rather than cities as a means to contain the amount of data to be manipulated. The data also has the ability to be manipulated by each of the states within the US. The age of the fatalities are standardized with the age being over 35 years. The dataset includes several static variables such as the county geographic level, same data source, involves cardiovascular diseases within the topic of heart disease mortality, and the population restriction to create the averages. The data provides enough to use data analytics to solve the above problems. The data was provided by the Center for Disease Control Division for Heart Disease and Stroke Prevention. The data source is from the National Vital Statistics System.[[4]](#endnote-4)

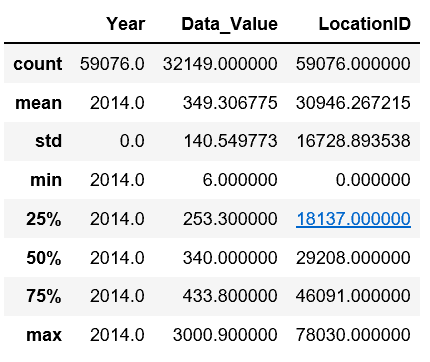
**Data Analysis and Evaluation**

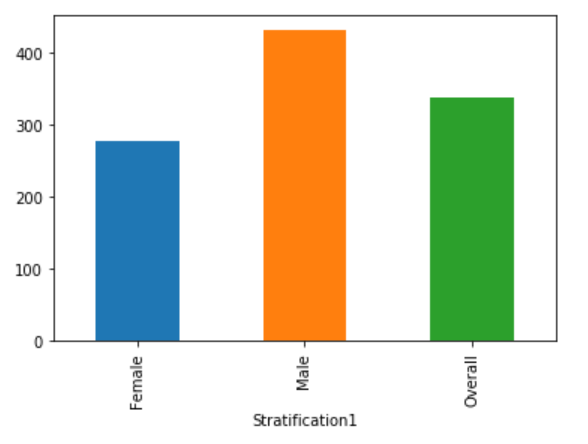
The necessary python packages and loading the file were the main steps to begin data analysis of the heart mortality data. A statistical summary suggests that the average mortalities are around 350 per 100,000 in the year of 2014. The lower end amount of averaged mortalities are at 253 while the higher averaged mortalities are around 433. Cleaning the data allowed to eliminate several unnecessary rows of data and to help with the process of data manipulation. Averaging the data into different categories such as state, gender, year, and race help to portray where heart mortality is primarily an issue. For instance, it is apparent that men are nearly twice as likely to die from heart disease as compared to women. The data manipulation also resulted in blacks, American natives, and whites, to be way more likely to die from heart disease compared to Hispanic and Asian and Pacific Islander ethnicities. One of the issues that I ran in to is creating a correlation between numbers because of this data set. For instance, I cannot compare obesity rates, smoking, drinking, or diabetes to the amount of deaths simply because the data does not include the rates. Since the data does not have many numbers to manipulate, there is a significant limitation to look for additional correlations. The data is primarily based on descriptive factors as opposed to numerical data. This causes limitations for comparisons. For instance the only numbers used include the year and the data value of individual mortalities averaged per 100,000. However, I did apply new techniques that were learned in class, such as a density plot that would compare the heart mortality rates between men and women. The density plot indicates that the probability of the mortality average belonging to a male rises as there are more deaths.

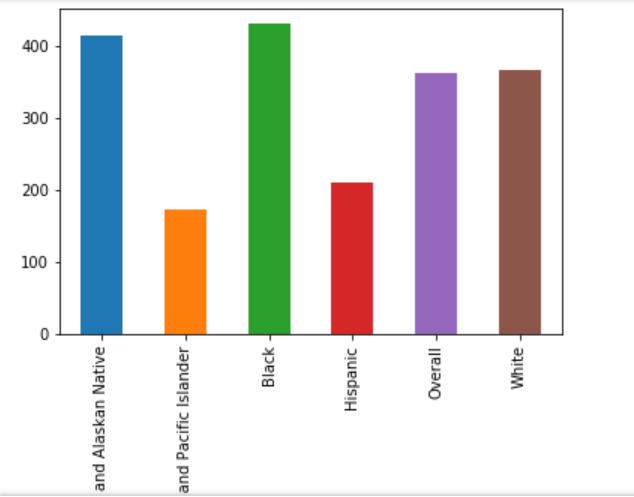
**Data manipulation**

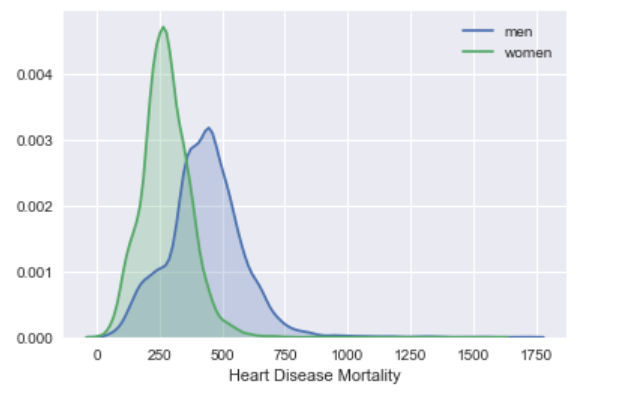
Jupyter notebook file “Prevalence of Heart Disease” illustrates a preliminary trial of data manipulation of the data from the CDC. The file was renamed to HeartMortality.csv for convenience. It illustrates part of the process involved to create the charts and tables for the project.

**Charts and Tables**

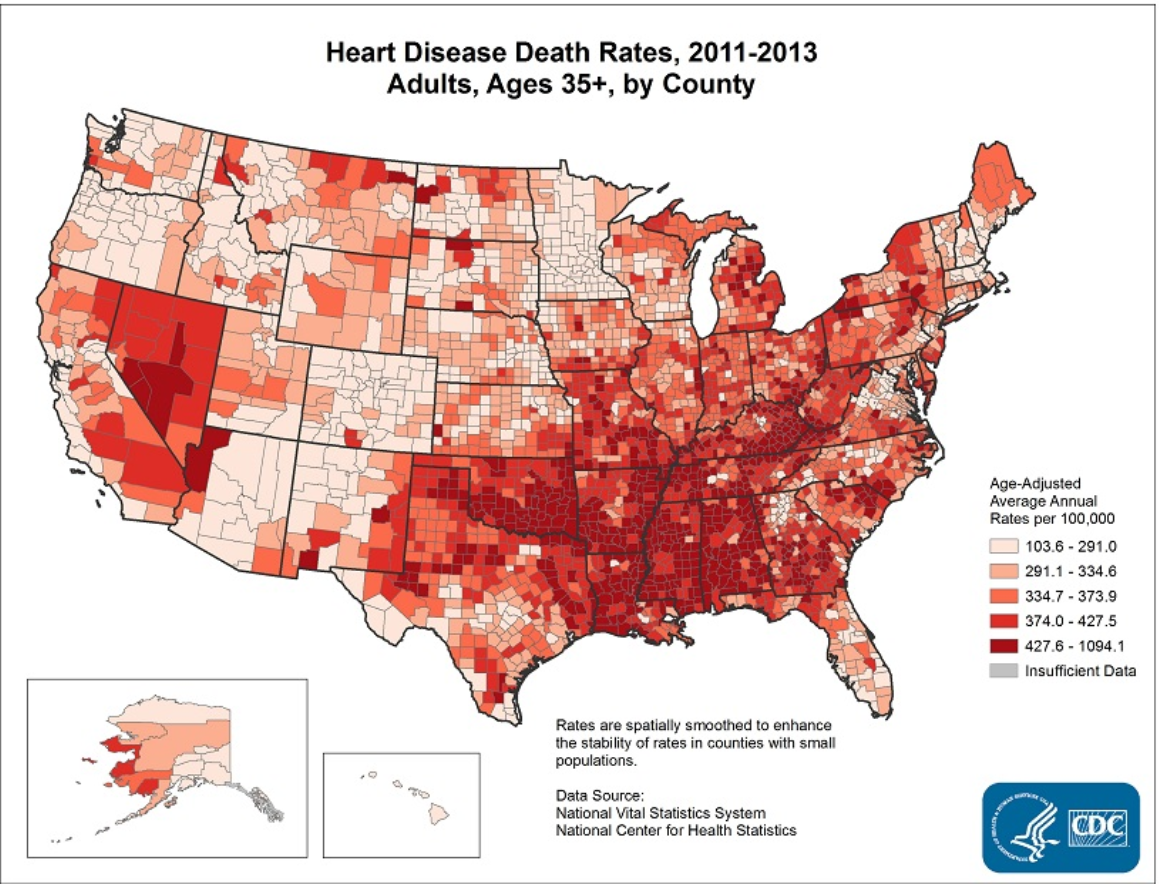








More Descriptive Chart of Heart Mortality by the CDC



**References**

1. https://www.cdc.gov/dhdsp/data\_statistics/fact\_sheets/fs\_heart\_disease.htm [↑](#endnote-ref-1)
2. Ibid. [↑](#endnote-ref-2)
3. Ibid. [↑](#endnote-ref-3)
4. https://data.cdc.gov/Heart-Disease-Stroke-Prevention/Heart-Disease-Mortality-Data-Among-US-Adults-35-by/i2vk-mgdh [↑](#endnote-ref-4)