**Introduction**

The Heart Disease mortality rate within the U.S is constantly increasing. According to the CDC in 2021 and 2020 Heart disease was the considered the #1 killer, it out beat cancer and covid-19. Not only is it number one it is still growing every year. Heart disease covered 20% of the United States Deaths for 2021. The CDC defined heart disease as major cardiovascular disease, heart disease, acute myocardial infarction, coronary heart disease, heart failure and strokes.

The data set we chose discusses heart disease mortality rate for people over the age of 35. The data set is set for the year of 2014 and discusses the mortality rate for each county in the United States per 100,000 people. This included age adjusted data and people who had 3 years average of dealing with heart disease. Each county also discussed the gender, race and longitude and latitude of each county. This study goes into the data set and tries learning the impact, gender, race, and the state you chose to live can impact you having heart disease. The model focuses specifically on the individual data provided instead of the overall gathered inside the dataset. The goal is to raise awareness to how your gender, race and the state you live in can impact your life and chances to have heart disease and die from it.

**Data Cleaning/Preparation**

The dataset itself had 59077 unique rows and 19 columns within itself. We initially started the project by reading in the dataset. From there we noticed a lot of columns had blank/invalid data and had a column with the label that discusses which rows had insufficient data. Removed those columns. Next, we had to remove the rows that had the overall label within stratification 1 and 2 (this is for gender and race). We wanted to focus on the individual data itself for our test instead of focusing on the overall data for all races and genders for better inferential testing. Next, we focused only on keeping the county data specifically. We did not want the overall data for states or the nation as this was going to broad for the test. Finally, we renamed the columns to better readability and easier testing.

Another dataset we did clean up as well was for overall. We wanted to keep an overall dataset for exploratory data/visual analysis. This helps keep reference on what data looked like compared to when combined. This followed the same process as discussed before but reversed it to where we only kept overall for stratification 1 and 2.

Next, we had to identify outliers within each set. The outliers we focused on were within the column that discussed heart disease mortality rate. We removed the outliers by creating a function to get the interquartile range, quarter 1 and quarter 3 to get the lower/upper bound with the formulas:

lower\_bound = Q1 - 1.5\*IQR

upper\_bound = Q3 + 1.5\*IQR

This covers over 95% of the data and removed outliers that went well beyond the scope of the lower/upper bound. This function ran against both datasets we cleaned up. After all this was done, we had two clean datasets to analyze and interpret.

Exploratory Data Analysis

\*code notes before you start this section

Clean up visual analysis (basically combine the charts and make the charts look the same)

Clean up zscore. This is just making it use the right dataset and not the overall dataset

Prove you have a normal distribution with a histogram and line with it.

Model Selection

Model Analysis

Conclusion and Recommendations.

(overall example)

<https://ieeexplore-ieee-org.sandiego.idm.oclc.org/document/10040352?arnumber=10040352>

^ Example I recommend we try to follow for formatting and all that

Links for introduction section

<https://www.cdc.gov/nchs/nvss/deaths.htm>

<https://www.cdc.gov/nchs/data/databriefs/db456-tables.pdf#4>

<https://www.cdc.gov/dhdsp/maps/dtm/data_sources.htm>