**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 overall dataset is included in the research to see how close the overall data and individual data is related to each other. 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

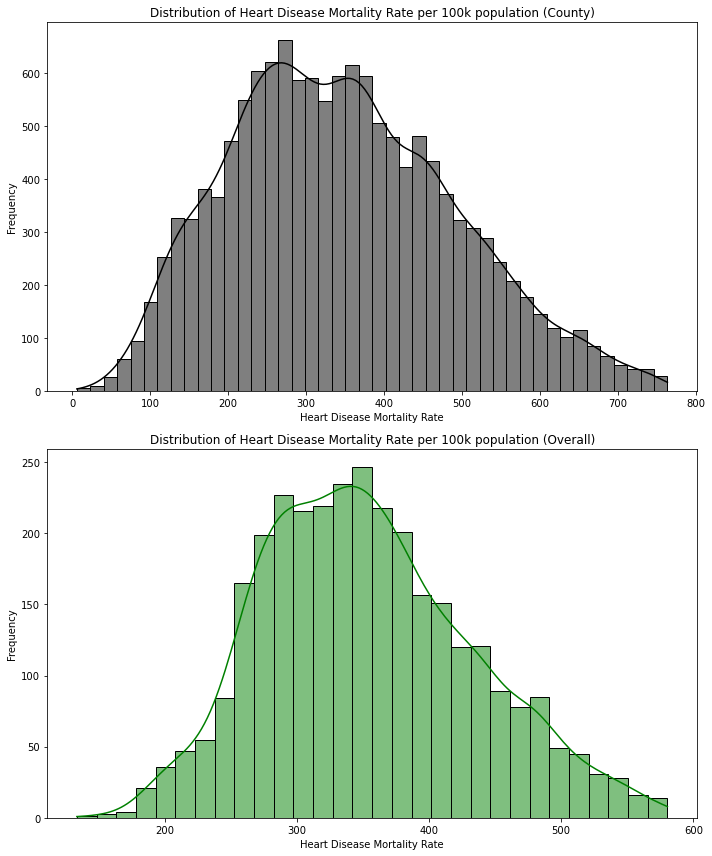
upper\_bound = Q3 + 1.5\*IQR

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**

As we started to graph the data, first started to look at the five number summary for our cleaned-up data. We see from our cleaned data that the mean and is around 347 people out of 100,000 population and the overall data showing about 353 people. Looking at the standard error rate, it is .68 units (people), which is low. This means that our individual data is very similar to our overall data.

Next, we looked at the distribution for heart disease mortality rate. Below are the results.



The two graphs highlight that both are of normal distribution. This mean that our data is following the central limit theorem can be used for our testing and modeling in determining the cause for high/low heart disease mortality rates for people over 35 years old.

Next we started to look at our independent variables and see how they are associated with the mortality rate in bar graphs

A graph of different colored squares

Description automatically generated

This graph above breaks out the count of amount of people for each race who were part of the statistic. As highlighted above, White has the most count for the mortality rate. The lowest is American Indian and Alaskan Native. This will allow us to use White as the default race for future testing.

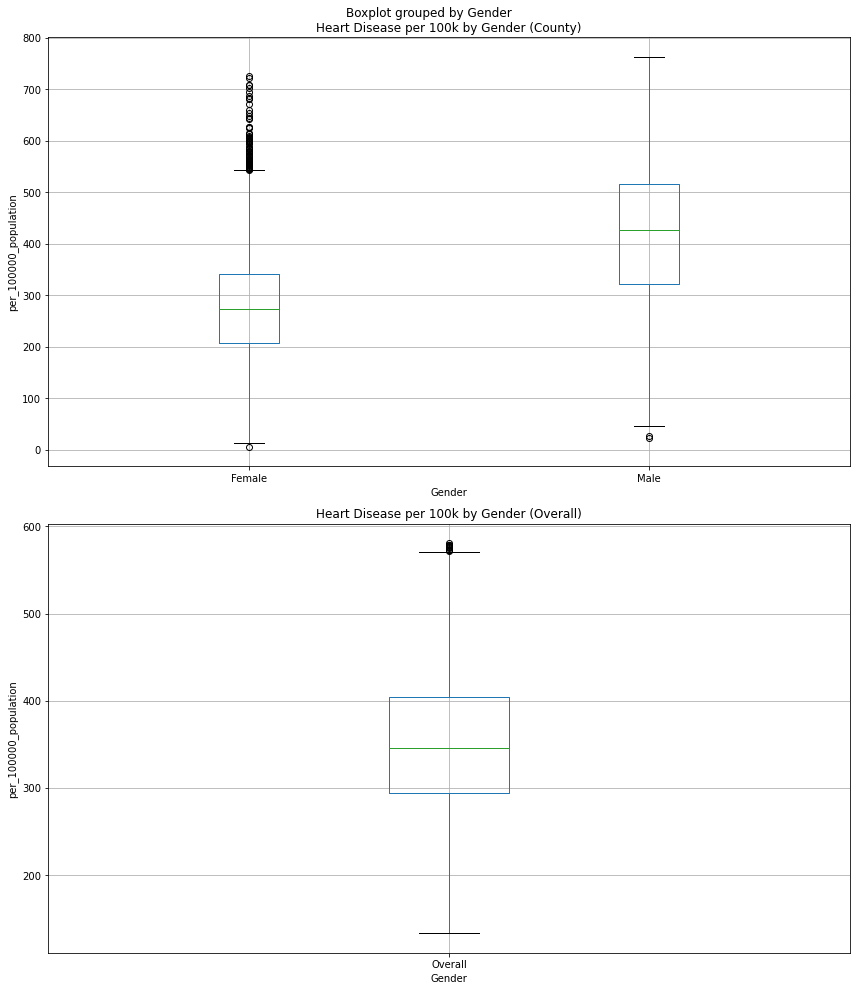
A blue and orange rectangles

Description automatically generatedThe graph above highlights the count for males and females for the mortality rate. As you can see the count is very close to each other showing we have a good representation for each gender. A screenshot of a screen

Description automatically generated

Next, we looked at the count by the states. We included the overall to see if there was any associated difference from states. Looking at this, this shows there are some states not included in the individual data. This was due to insufficient data collected for the independent variables and the mortality rate. This does correlate well with general populations of each state. This is highlighted with Texas on the high side having bigger populations and same on the lower side with Hawaii not having a lot of counts.

Next we created boxplots to look at the independent data related to the mortality rate and to see if there are any outliers and generally how the distribution will look. This included the overall dataset as well as the overall for each independent variable. This is to see how the individual statistics compared to the combined stat.



Looking at the bar graph, we can see that males have an overall higher amount of people affected than females. The females do have a considerable amount that are outliers. Comparing it to the overall statistic. We can see that the ceiling is considered lower compared to the individuals. The ceiling is only 600 instead of being around 750.

A screenshot of a computer screen

Description automatically generated

Looking at race, we can see Black has the highest average amount out of all the races. It is interesting to not that they do not have any outliers and neither do American Indian and Alaskan Natives. In fact, American Indians and Alaskan Natives show very similar results to Black which is considerable since they had the lowest count. Next, we see White falling in the middle. With minimal outliers. Hispanic is actually the race that has the most outliers including some counties reaching almost the ceiling. Finally looking at Asian and Pacific Islanders, we can see they have the lowest average. Interesting note is they do have one outlier that stands above the rest with one county having about an average of 720. Comparing this to overall, we see White having the closets average. The ceiling is lower as well for this with it being closer to 600 than 800.

After the visual analysis, we ran chi square test to test for association between the assumed independent variables and for creating hypotheses for our testing. We also ran z testing as well to see if there was any significant impact on the heart rate with the other categories.

Category Chi-square statistic p-value

0 Gender 7.171499e+03 3.097805e-69

1 Ethnicity 2.560000e+04 2.355235e-111

2 County 9.512658e+06 1.000000e+00

3 State 2.557907e+05 9.999865e-01

Looking at the Chi square testing suggest that gender and ethnicity have highly significant association with the mortality rate. The large statistic and miniscule p value show the strong relationship. Looking at state and county, there is not statistical association. County has a perfect 1.00 for the p-value which indicates no association. With state having a value extremely close to 1.00 it falls similar to county.

For z testing for gender we saw that females have a p-value of less that .05 with a z-statistic of -56 which shows that there is a significant impact with being a female and that the mortality rate is lower. For male even with having 39 this p-value was 2.0 which indicates there was no association. Since females did show significant impacts, this suggest that there is a noticeable difference which will lead to some hypothesis testing once we start looking at modeling and model analysis.

Looking at the z testing for race, there was significance for Hispanic and Asian and Pacific Islander. For Hispanic the z-statistic was -58 and Asian and Pacific Islander -81. Both show that the significant impact will affect the average to be lower. This indicates that that some hypothesis testing can happen on this variable when we start looking at models.

There was z testing for state as well. To see if there was any significant impacts on the mortality rate. It is worth noting that with such a small p-value with chi square test the results could be due to random chance and not due to statistical significance.

\*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.~~

~~Look through all modules and see what your missing~~

~~Histogram with line curve~~

~~5 number summary (Kinda did with IQR) (.describe)~~

~~Confidence interval (kinda did with zscore) (Did with OLS .025 - .975)~~

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>