Factors Influencing Heart Disease Deaths

Outhai Xayavongsa and Aaron Ramirez

University of San Diego

Masters of Applied Artificial Intelligence

AAI-500: Probability and Statistics for Artificial Intelligence

Professor Leon Shpaner

February 26, 2024

Factors Influencing Heart Disease Deaths

Heart disease is the predominant cause of mortality in the U.S. indicated by CDC's analysis, overshadowing other significant causes such as cancer and the recent COVID-19 pandemic. In the year 2021 alone, heart disease was responsible for one-fifth of all deaths, reflecting its significant impact on public health. The CDC employs a comprehensive definition for heart disease, including various cardiovascular conditions: Heart attacks, known as Acute myocardial infarction, arise when a segment of the heart muscle is deprived of blood; coronary artery disease is when an accumulation of fatty deposits inside the coronary arteries leads to reduction of blood flow; heart failure describes the heart cannot pump blood efficiently; and strokes occur when the blood supply to the brain is obstructed (Xu et al., 2022).

The dataset selected for analysis offers a nuanced view of heart disease mortality for the year 2014, specifically targeting the demographic aged 35 and older across different U.S. counties (U.S. Department of Health & Human Services, 2023). It employs age adjustment to mortality rates, a critical statistical method allowing equitable comparisons across diverse population age structures. By averaging mortality data over three years, the dataset aims to eliminate year-to-year fluctuations, providing a more consistent and dependable portrayal of the heart disease mortality landscape.

Demographic details such as gender and race are included within each county's data, which are vital for recognizing mortality trends and patterns among distinct population groups. Including geographical coordinates for each county opens avenues for spatial analysis, potentially linking heart disease mortality rates to environmental, economic, and healthcare accessibility factors. The study's purpose is twofold: exploratory and preventive. It intends to investigate how demographic and geographical factors contribute to the risk and outcomes of heart disease. The overarching goal is to inform public health strategies that could mitigate the risk factors associated with heart disease, improve health outcomes, and pinpoint populations that might benefit from enhanced preventive healthcare services.

Data Cleaning and Preparation

The initial dataset contained 59,077 unique rows and 19 columns. The project commenced with the importation of the dataset, followed by the observation of numerous columns containing blank or invalid data, along with a column labeling row with insufficient data. Subsequently, these columns were eliminated. Next, rows labeled with the "overall" tag within stratifications 1 and 2 (pertaining to gender and race) were excluded, with the focus redirected towards individual-level data for improved inferential testing. Moreover, emphasis was placed solely on retaining county-specific data, disregarding broader state or national data for the test. Finally, column names were modified for enhanced readability and ease of testing.

A parallel dataset was also subjected to cleanup, intended for overall exploratory data and visual analysis. The process mirrored that described previously, albeit with a reverse approach, retaining only overall data for stratifications 1 and 2. The subsequent step involved outlier identification within each dataset, focusing on the column detailing heart disease mortality rates. Outliers were addressed by employing a function to calculate the interquartile range, as well as the first and third quartiles, subsequently determining lower and upper bounds using the formulas:

$$\textbf{Lower bound} = \textbf{Q1} - \textbf{1.5IQR}$$

$$\textbf{Upper bound} = \textbf{Q3} + \textbf{1.5IQR}$$

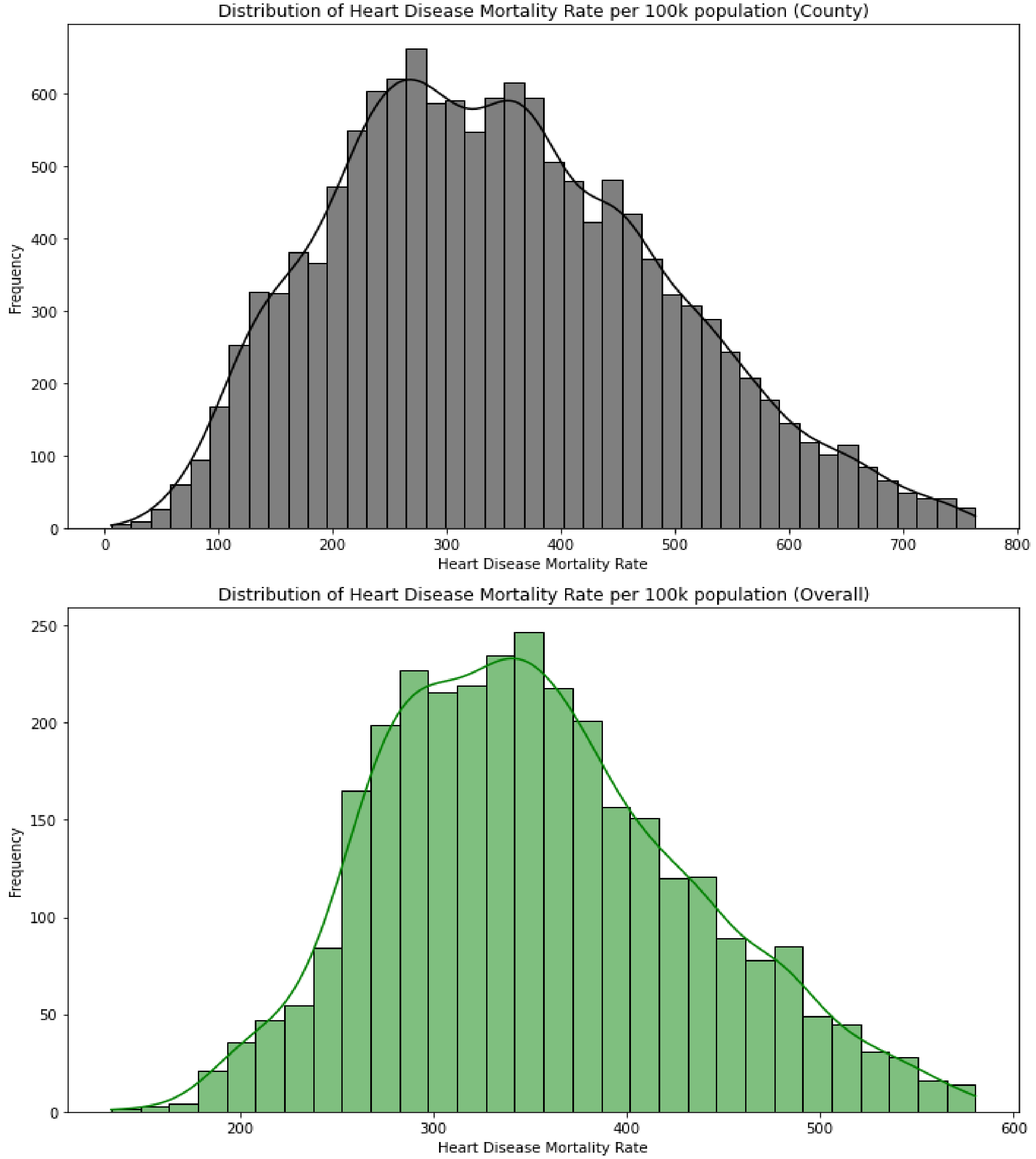
This methodology effectively captured over 95% of the data while eliminating outliers beyond the specified lower and upper bounds. The outlier removal process was applied to both datasets, resulting in two clean datasets poised for analysis and interpretation.

Exploratory Data Analysis

As the data analysis commenced, the focus initially turned to examining the five-number summary for the cleaned-up datasets. It was observed that the mean heart disease mortality rate for the individual-level data stood at approximately 347 individuals per 100,000 population, while the overall dataset indicated a rate of about 353 individuals. Additionally, the standard error rate was calculated to be 0.68 units (individuals), denoting a low level of variability. These findings suggest a notable similarity between the individual-level and overall datasets. Subsequently, attention was directed towards assessing the distribution of heart disease mortality rates. The results are presented below.

Figure 1

Histogram Distributions on Heart Disease Mortality Rate per 100k for the County and Overall



Distribution of Heart Disease Mortality Rate per 100k population (County)

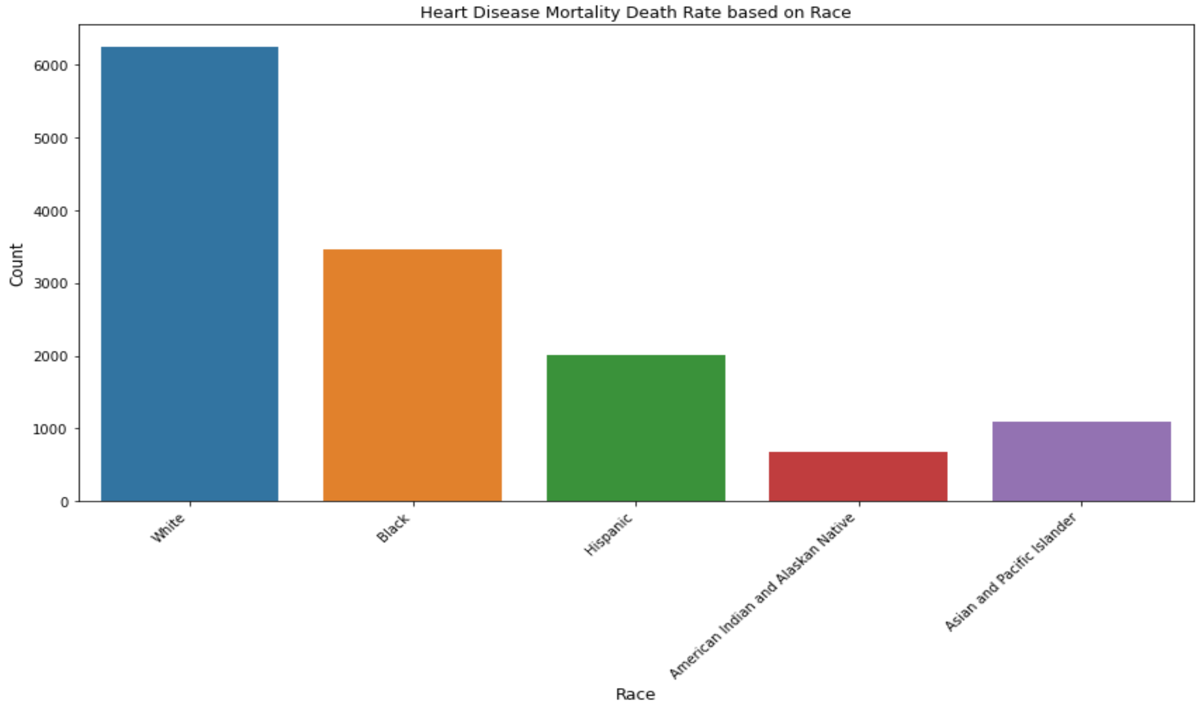
Distribution of Heart Disease Mortality Rate per 100k population (Overall)

Note. The two graphs illustrate that both datasets exhibit a normal distribution, indicating adherence to the central limit theorem. This suggests that our data can be effectively utilized for testing and modeling purposes in discerning the factors influencing high or low heart disease mortality rates among individuals aged over 35 years.

Subsequently, the examination delved into assessing the association between independent variables and mortality rates through bar graphs.

Figure 2

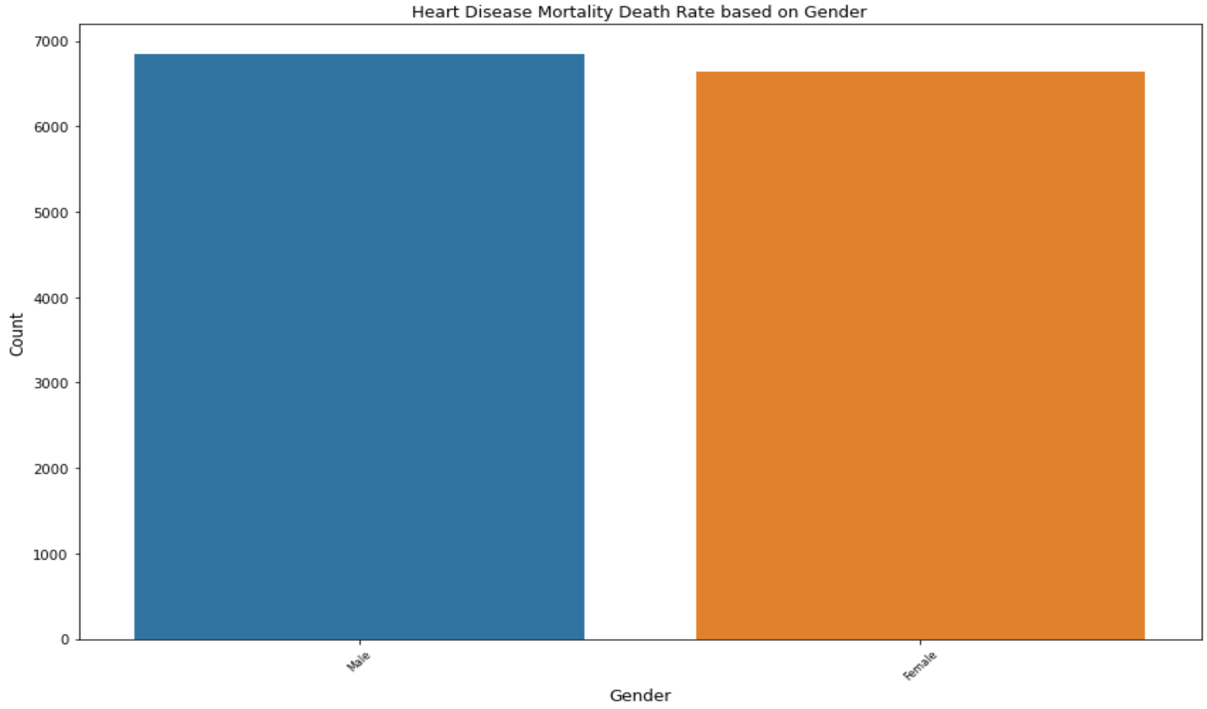
Bar Plot on Heart Disease Mortality Death Rate based on Race.



Note. The graph presented above delineates the distribution of individuals across different racial categories within the dataset. As indicated, individuals identifying as White constitute the highest count in the mortality rate statistics, whereas the count is lowest for those identifying as American Indian and Alaskan Native. Consequently, White race will be utilized as the default category for subsequent testing purposes.

Figure 3

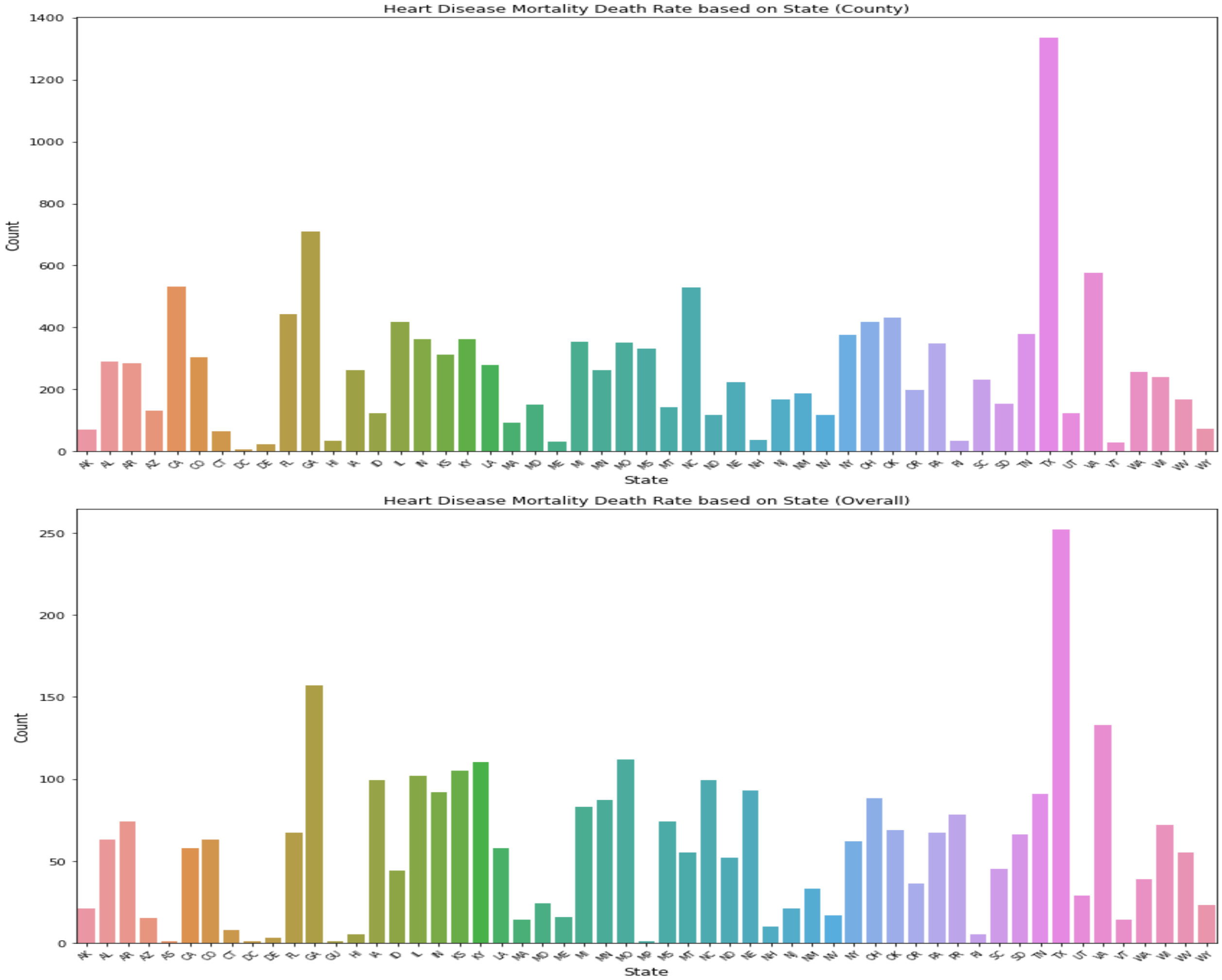
Bar Plot of Heart Disease Mortality Death Rate on Gender



Note. The graph depicted above showcases the count of individuals categorized by gender in relation to the mortality rate. It is evident from the graph that the counts for males and females are closely balanced, indicating a satisfactory representation of each gender within the dataset.

Figure 4

Heart Disease Mortality Death Rate by State Comparison

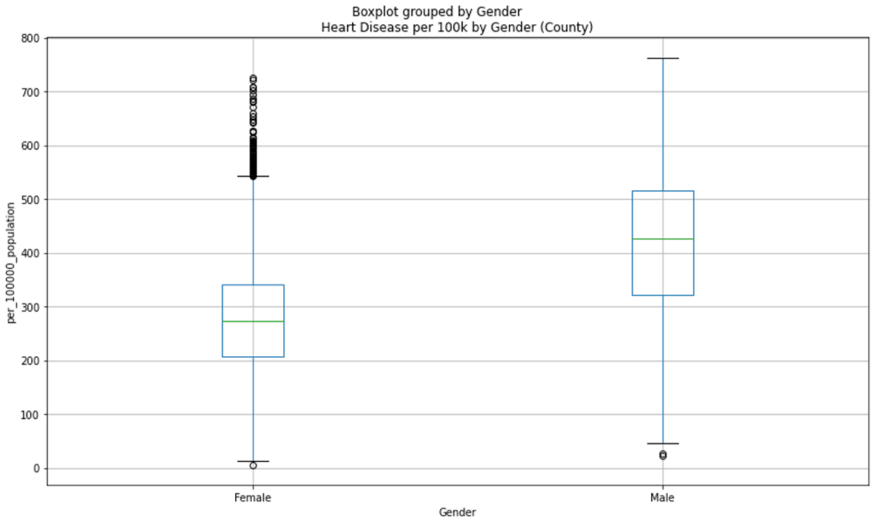


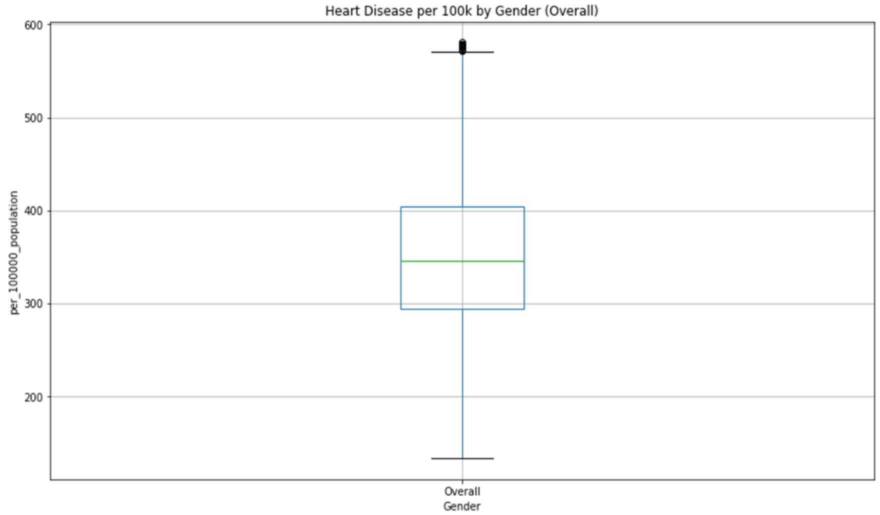
Note. An analysis was conducted to examine the count of individuals by state, incorporating both individual and overall datasets to discern any discrepancies among states. The observation revealed that certain states were not adequately represented in the individual-level data, attributed to insufficient data collection concerning independent variables and mortality rates. Notably, this discrepancy appeared to align with the general population distribution of each state, exemplified by larger states such as Texas exhibiting higher counts, while smaller states like Hawaii demonstrated lower counts.

Furthermore, boxplots were generated to explore the relationship between independent variables and mortality rates, aiming to identify outliers and discern the distribution patterns. Both individual and overall datasets were utilized for comparison, facilitating an assessment of how individual-level statistics contrasted with aggregated data.

Figure 5

Box Plots on Heart Disease per 100k by Gender based on County and Overall



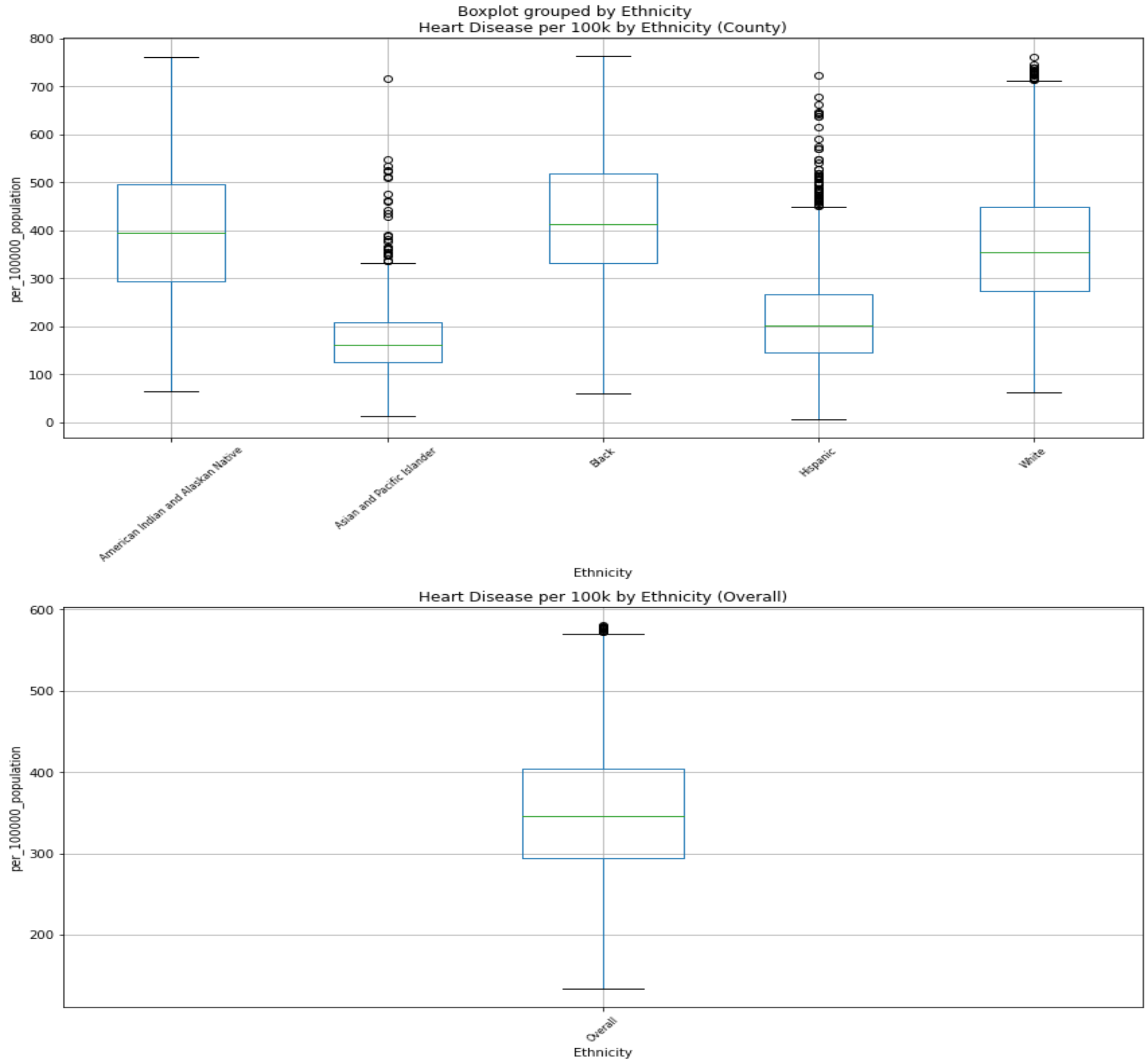


Note. The boxplot analysis reveals that the number of heart disease cases is greater among males than females. Additionally, the data for females includes a significant number of outliers.

A comparison with the aggregated data highlights that the maximum value in the overall dataset is around 600, which is lower than the individual-level data where the maximum reaches approximately 750. This suggests a discrepancy between the individual and aggregated data points, with individual records showing higher extremes.

Figure 6

Box Plots of Heart Disease per 100k by Each Ethnicity and Overall



Note. It becomes apparent that Black individuals exhibit the highest average mortality rate among all races.

Notably, Black individuals do not have any outliers, a trend shared with American Indian and Alaskan Natives, despite their initially low count. Surprisingly, American Indians and

Alaskan Natives display similar results to Black individuals, which is noteworthy given their comparatively lower count. Subsequently, White individuals fall within the middle range with minimal outliers. Conversely, Hispanic individuals emerge with the highest number of outliers, with some counties nearly reaching the ceiling. Lastly, Asian and Pacific Islanders present the lowest average mortality rate, albeit with one outlier that significantly exceeds the others, with one county registering an average of approximately 720. Comparing these findings to the overall statistics, White individuals display the closest average, while the ceiling is notably lower, approaching 600 rather than 800 as observed in the individual-level data.

Following the visual analysis, chi-square tests were conducted to examine the association between the assumed independent variables and to formulate hypotheses for subsequent testing. Additionally, z-tests were employed to assess whether there was any significant impact on heart rate in relation to the other categories.

Table 1

Chi-Square test results

Category	**Chi-square statistic**	**p-value**
Gender	7.171499e+03	3.097805e-69
Ethnicity	2.560000e+04	2.355235e-111
County	9.512658e+06	1.000000e+00
State	2.557907e+05	9.999865e-01

The results indicate that gender and ethnicity exhibit highly significant associations with mortality rate. The substantial chi-square statistics and minuscule p-values underscore the strength of these relationships. Conversely, there is no statistical association observed for county, as evidenced by a p-value of 1.00. Similarly, state demonstrates a p-value extremely close to 1.00, indicating a lack of association akin to county.

In the z-testing conducted for gender, it was observed that females exhibited a p-value of less than 0.05, with a z-statistic of -56, indicating a significant impact on mortality rate, resulting in lower rates for females. Conversely, for males, despite a count of 39, the p-value was 2.0, suggesting no significant association. The significant impact observed for females implies a notable difference that warrants further hypothesis testing during subsequent modeling and analysis phases.

Similarly, in the z-testing for race, significance was observed for Hispanic and Asian and Pacific Islander categories. Specifically, Hispanic individuals displayed a z-statistic of -58, while Asian and Pacific Islanders exhibited -81, both indicating significant impacts due to p-value being less than .05 resulting in lower averages. This suggests the potential for hypothesis testing on these variables in subsequent modeling endeavors.

Additionally, z-testing was conducted for states to assess any significant impacts on mortality rate. However, it is worth noting that the small p-values obtained from the chi-square test could potentially be attributed to random chance rather than statistical significance. Nonetheless, the z-testing revealed statistical significance for numerous states including Arizona, California, Colorado, Connecticut, Delaware, Florida, Iowa, Idaho, Louisiana, Massachusetts, Maryland, Maine, Minnesota, North Carolina, North Dakota, Nebraska, New Hampshire, New

Jersey, New Mexico, Nevada, New York, Oregon, Pennsylvania, Rhode Island, Utah, Virginia, Vermont, Washington, and Wisconsin.

As the analysis transitions to the next phase of modeling, it is evident that gender and race emerge as the most promising variables due to significant associations identified through the chi-square test, coupled with the significant impacts observed in certain categories within each variable. While testing on state variables is plausible, its significance is somewhat diminished by the chi-square p-value. Nonetheless, exploring the modeling of the states could still yield valuable insights.

Model Selection

When considering suitable models for analyzing gender and race in relation to heart disease mortality rate, several factors come into play. Heart disease mortality rate constitutes continuous float data, while ethnicity, gender, and state are categorical variables. Given these considerations, the models under consideration included linear regression/multilinear regression, logistic regression, clustering, and naïve Bayes.

Eliminating naïve Bayes from consideration was due to it primarily focus being on categorical data, which does not align with our target variable of continuous data. Similarly, logistic regression, which emphasizes classification outcomes, was deemed less applicable to our goal of analyzing continuous data.

Multilinear regression emerged as a promising choice, as it facilitates an examination of the impact of gender and ethnicity on the target variable (mortality rate), allowing for predictions based on these independent variables. This model enables the assessment of the combined effects of both independent variables on the target variable.

Furthermore, clustering was retained as a potential model to explore patterns and identify grouped observations that may contribute to variations in mortality rate. This approach aims to uncover any underlying patterns driving mortality rate trends, thereby providing insights beyond what is currently known.

Model Analysis

Upon selecting multi/linear regression and clustering as our models, we proceeded to format our dataset, accordingly, necessitating the hot encoding of all categorical variables. Subsequently, hypotheses were formulated to align with the significant data previously identified and the chosen models.

Hypothesis 1: There is a significant difference in heart disease mortality rates between genders.

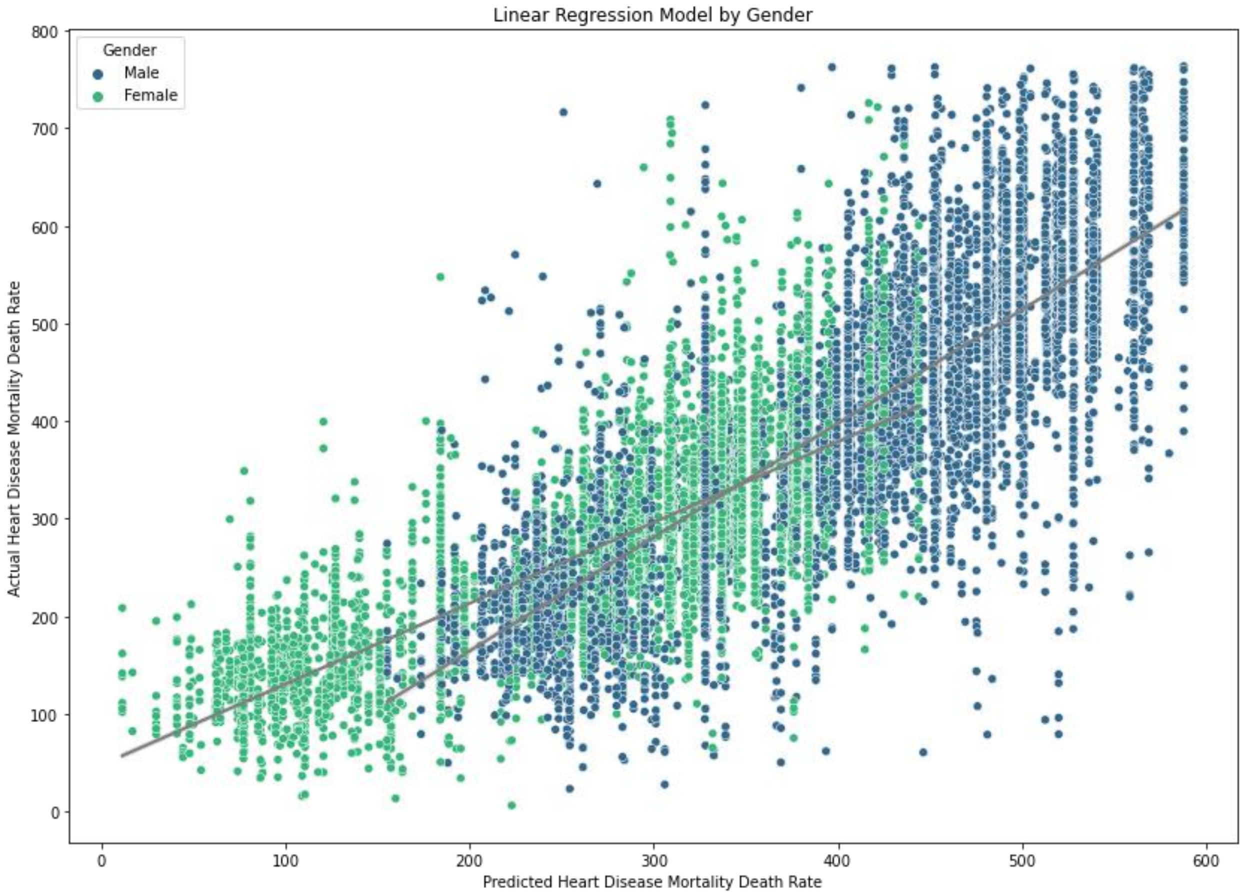
- Null Hypothesis (H0): There is no significant difference in heart disease mortality rates between genders.
- Alternative Hypothesis (H1): There is a significant difference in heart disease mortality rates between genders.

**All genders are tested but females were the primary subject to see their significant impact*

Linear regression was employed to test this hypothesis. Analysis of the results revealed that the model accounted for 23.6% of the variation in predicting the dependent mortality rate, indicating a relatively low explanatory power. The model indicated a default mortality rate of 276 individuals per 100,000 population for females (const), with males exhibiting a significantly higher rate, reflecting an increase of 140 individuals. Furthermore, the p-values for each gender were found to be significant, and the condition number indicated the absence of multicollinearity. Consequently, the results supported the rejection of the null hypothesis and the acceptance of the alternative hypothesis, signifying a significant difference in heart disease mortality rates between genders. Females are the lower sloped line.

Figure 7

Linear Regression Model by Gender



Hypothesis 2: There is a significant difference in heart disease mortality rates between ethnicities.

- Null Hypothesis (H0): There is no significant difference in heart disease mortality rates between ethnicities.

- Alternative Hypothesis (H1): There is a significant difference in heart disease mortality rates between ethnicities.

**All ethnicities are tested but the main subjects were Hispanic and Asian and Pacific Islanders*

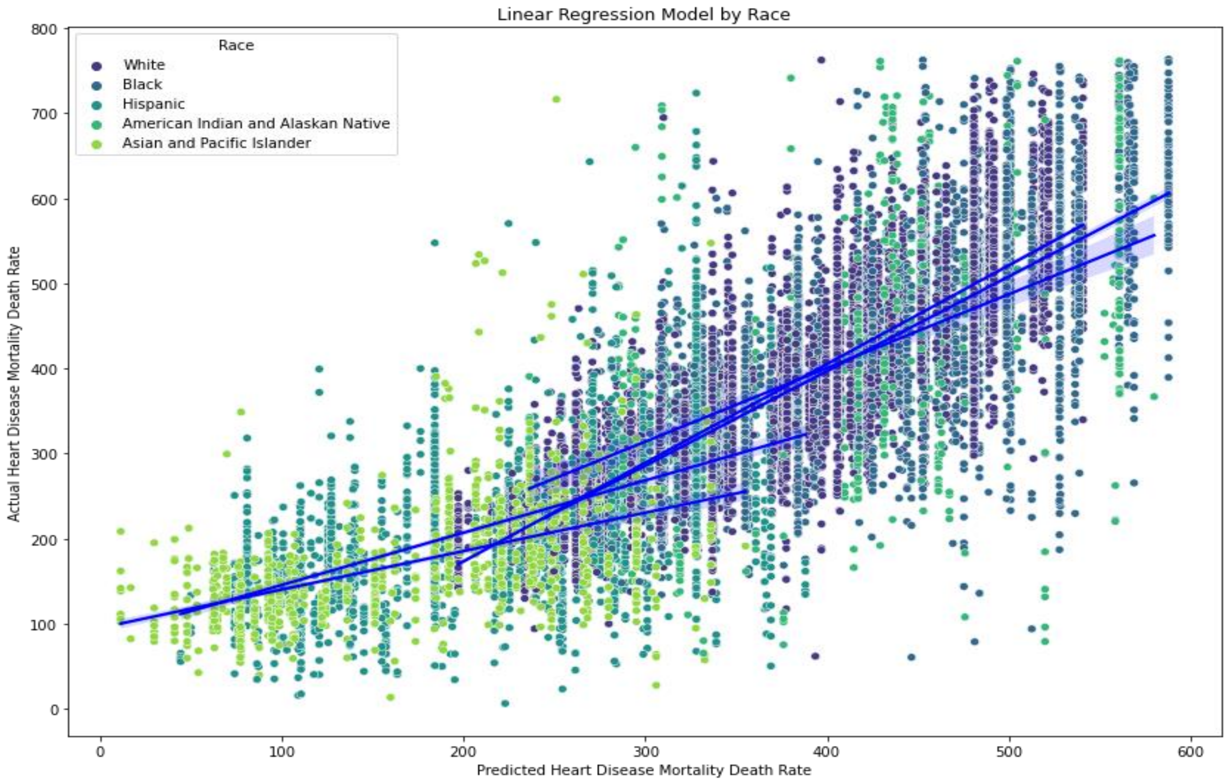
In examining the initial results, indications of multicollinearity emerged, evidenced by a condition number of 1.44e+15. To address this issue, the test was rerun after eliminating multicollinearity through several methods. Firstly, the 'const' column was replaced with 'White', aiding in the initial hot encoding process. Subsequently, the variance inflation factor (VIF) was calculated to detect additional multicollinearity among ethnicities. Any VIF values exceeding 10 were removed to prevent unstable and unreliable estimates.

The revised model, which accounted for 34% of the variation in predicting the dependent mortality rate, demonstrated moderate explanatory power, explaining a considerable portion of the mortality rate. Notably, the 'White' ethnicity exhibited a default mortality rate of 368 individuals per 100,000 population, while Black and American Indian and Alaskan Native displayed increased rates of 60 and 35 individuals, respectively. Furthermore, significant impacts were observed for Hispanic and Asian and Pacific Islander, with both ethnicities exhibiting decreases of 151 and 195 individuals, respectively. These reductions in mortality rates for Hispanic and Asian and Pacific Islander ethnicities signify significantly lower rates compared to the other three.

Additionally, the p-values for each ethnicity were significant, further supporting the rejection of the null hypothesis and acceptance of the alternative hypothesis. Notably, the condition number indicated the absence of multicollinearity, thus resolving the initial issue. These findings underscore the significant impacts of Hispanic and Asian and Pacific Islander ethnicities on heart disease mortality rates, aligning with the earlier z-test results and providing valuable insights for further exploration.

Figure 8

Linear Regression Model by Race



Note. The two lower sloped lines are Hispanic and Asian and Pacific Islander. The graph shows how they are significantly different from the other ethnicities.

Hypothesis 3: There is a significant interaction effect between gender and ethnicity on heart disease mortality rates.

- Null Hypothesis (H0): There is no significant interaction effect between gender and ethnicity on heart disease mortality rates.

- Alternative Hypothesis (H1): There is a significant interaction effect between gender and ethnicity on heart disease mortality rates.

The selected model for this hypothesis was multilinear regression. Upon examination of the results, the model demonstrated an explanatory power of 58.6%, indicating a moderate ability to predict the dependent mortality rate. Like the approach used for ethnicity, multicollinearity was addressed.

The model revealed that white females (const) had a default mortality rate of 297 individuals per 100,000 population. Males exhibited a significantly higher rate, with an increase of 140 individuals, with black males showing the highest among all groups. Conversely, Asian or Pacific Islander females displayed the lowest mortality rate. Additionally, Hispanic ethnicity continued to exhibit a significant impact, with a coefficient of -155.

The p-values for each variable were found to be significant, confirming their individual contributions to the model. Moreover, the condition number indicated the absence of multicollinearity, affirming the reliability of the results. Notably, the variables demonstrating significant impact trended towards lower mortality rates, as observed in the preceding graphs. These findings support the rejection of the null hypothesis and the acceptance of the alternative hypothesis, reinforcing the significant differences observed in heart disease mortality rates among genders and ethnicities.

Hypothesis 4: There is a significant difference in heart disease mortality rates between states.

- Null Hypothesis (H0): There is no significant difference in heart disease mortality rates between states.
- Alternative Hypothesis (H1): There is a significant difference in heart disease mortality rates between states.

**All states are test, but the subjects were Arizona, California, Colorado, Connecticut, Delaware, Florida, Iowa, Idaho, Louisiana, Massachusetts, Maryland, Maine, Minnesota, North Carolina, North Dakota, Nebraska, New Hampshire, New Jersey, New Mexico, Nevada, New York, Oregon, Pennsylvania, Rhode Island, Utah, Virginia, Vermont, Washington, and Wisconsin primary subjects for significant impact.*

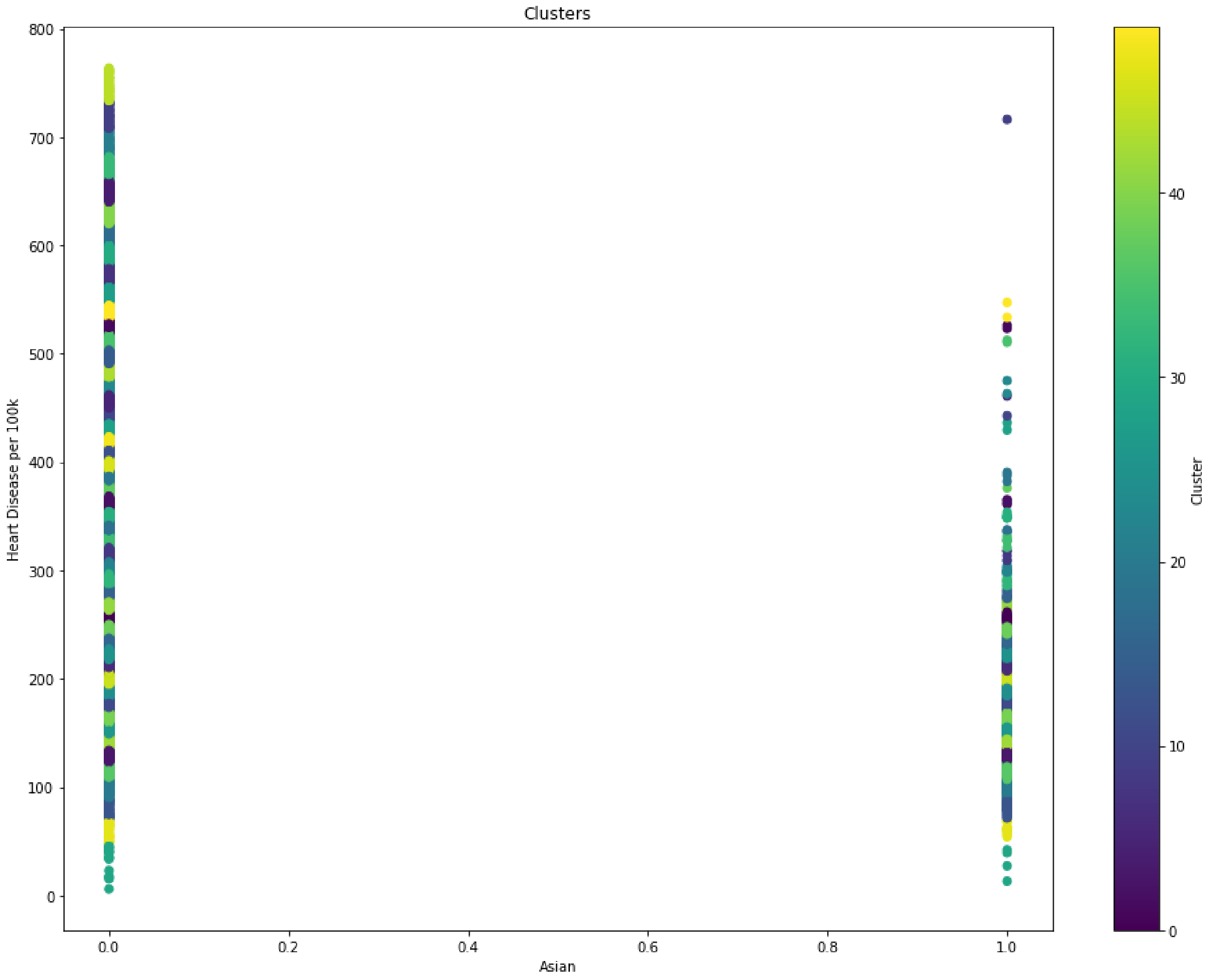
Regarding the fourth hypothesis, which assesses the influence of states on mortality rates, the results are inconclusive. The Chi-Square test indicated a lack of association, whereas the linear regression suggested a minor influence of certain states, though they accounted for a small portion (16.6%) of the model's predictive capacity. The hypothesis was tested using linear regression, focusing primarily on states identified as statistically significant in the preceding z-test. Attempted measures to address multicollinearity involved applying the method utilized for ethnicity variables. While this approach successfully removed Georgia and Texas, multicollinearity persisted, indicated by a condition number exceeding the threshold of 30, reaching 41.6.

Analysis of the results revealed that the model accounted for 16.6% of the variation in predicting the dependent mortality rate, signifying a low explanatory power. Notably, states identified as significant in the z-test exhibited significant findings based on their p-values. However, given the observed multicollinearity and the potential for unreliable results, further testing is warranted to make informed decisions regarding this hypothesis. The considerable variability observed suggests the need for additional exploration before drawing definitive conclusions.

Regarding clustering analysis, notable patterns emerged that could be associated with the findings discussed above. These patterns appeared to be aligned with our earlier observations and analyses. Further exploration of these patterns could provide valuable insights into the factors influencing heart disease mortality rates.

Figure 9

Cluster chart of the Asian Ethnicity

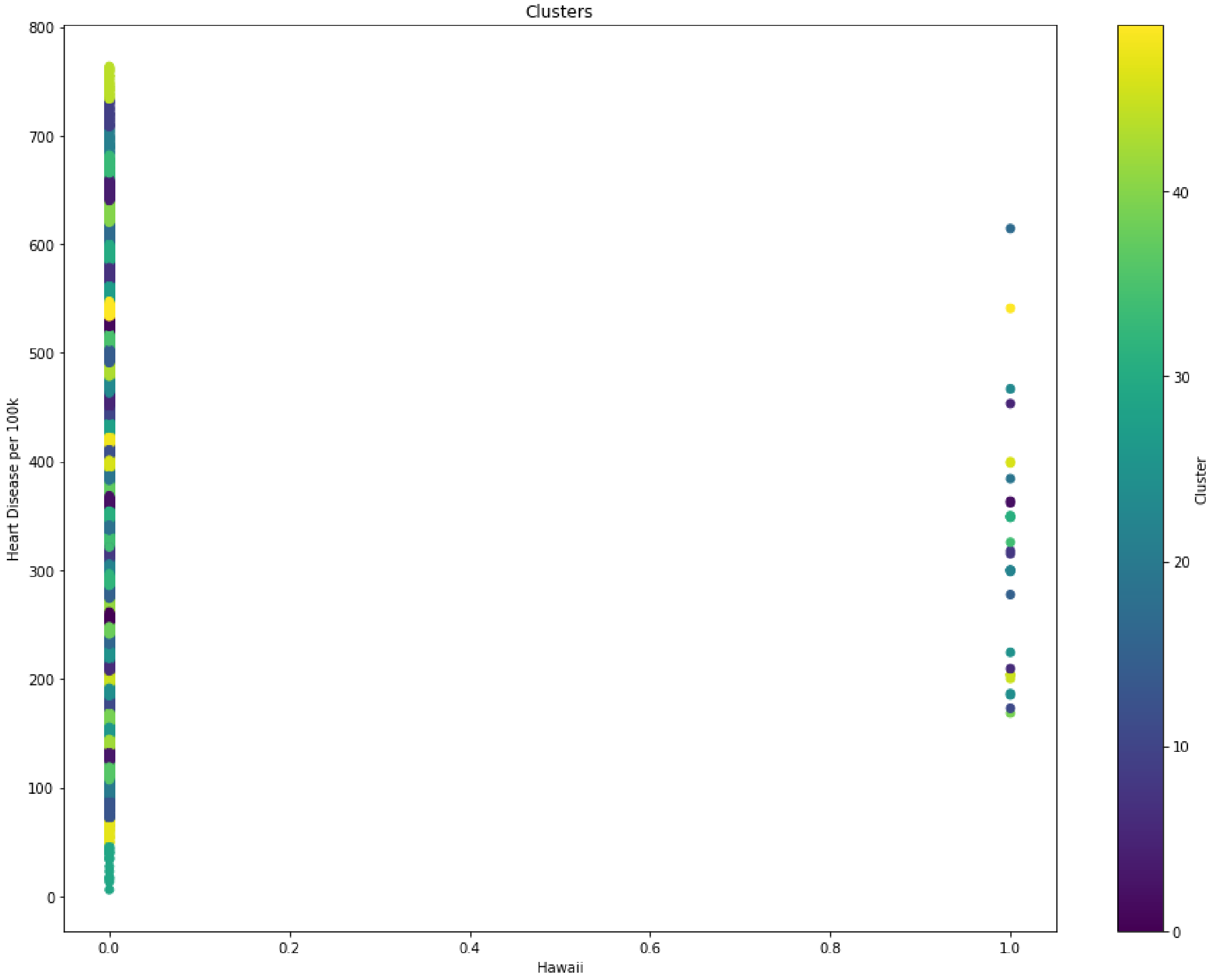


Note. Asians exhibited the lowest ceiling, as evidenced by both the box plot and the cluster chart.

Moreover, the cluster chart indicated a majority clustered closer to their floor and ceiling. These findings align with observations made for other ethnicities, suggesting a consistent pattern across different analyses and visualizations.

Figure 10

Cluster chart of the State of Hawaii



Note. The data from Hawaii appears to be sparse and widely spread out. The clustering pattern is not clearly discernible, and there exists a significant gap between the floor and ceiling values.

Conclusion

In the conclusion of the study, we observe a distinct pattern in the mortality rates due to heart disease that correlates with gender and ethnicity. Women, as well as individuals from Asian and Pacific Islander and Hispanic backgrounds, tend to show lower mortality rates. Conversely, African American males are highlighted as the group with the highest mortality rates. To gain a deeper understanding of these patterns, it is proposed that future research should delve into the significant variables that contribute to these lower rates. Cultural practices, including dietary customs among different ethnic groups, warrant further examination. For women, it would be beneficial to explore how their diet differs from that of men, along with the potential influences of living conditions and cultural expectations.

For a more comprehensive analysis, future studies could compare ethnic groups globally to discern any patterns that might correlate with environmental conditions. Examining the impact of different climates, such as colder regions or areas with high humidity, on heart disease mortality could be particularly revealing. Furthermore, socio-economic factors, healthcare access, and lifestyle choices, such as smoking and physical activity levels, across states could provide a clearer picture of the underlying causes of heart disease. By addressing these additional factors, researchers may uncover more nuanced insights into the prevalence of heart disease among adults over the age of 35 and the disparities seen across different demographic groups. This would not only contribute to the academic understanding of heart disease but could also inform public health policies and interventions aimed at reducing the burden of this disease (Nagar et al., 2023).

Recommendations

Based on the findings of the study, it is recommended that public health interventions should be tailored to address the specific needs of high-risk groups identified in the analysis. Given that African American males have the highest mortality rates, targeted prevention strategies such as community-based health education programs, improved access to early screening, and culturally sensitive healthcare services should be prioritized.

Moreover, the lower mortality rates observed among women, Asians Pacific Islanders, and Hispanics suggest that there may be protective cultural, dietary, or lifestyle factors at play. It is recommended that these factors be researched further to isolate the beneficial practices that could be promoted more widely. Public health campaigns might focus on the adoption of heart-healthy diets and lifestyles that mirror the positive aspects found within these communities.

Considering the inconclusive impacts of state-level factors, it would be prudent to conduct a more granular investigation into the socioeconomic and environmental factors that vary across states, which could be influencing heart disease mortality rates. This might include an in-depth analysis of healthcare infrastructure, the prevalence of risk factors like smoking and obesity, and even state-specific policies on healthcare. The goal would be to identify actionable policy levers that state governments can pull to improve heart health outcomes among their populations.

References

Nagar, K., Darji, J., Christian, A., & Patel, N. (2023). A Household Survey To Assess Prevalence of Communicable and Non-Communicable Disease and Standard of Living Patterns among Rural Peoples Residing in Rural Area of Kheda District, Gujarat. *Journal of Coastal Life Medicine*, *11*(1). https://www.jclmm.com/index.php/journal/article/view/558

U.S. Department of Health & Human Services (2023, August 26). *Heart Disease Mortality Data Among US Adults (35+) by State/Territory and County*. Data.gov. Retrieved January 27, 2024, from https://catalog.data.gov/dataset/heart-disease-mortality-data-among-us-adults-35-by-state-territory-and-countyhttps://catalog.data.gov/dataset/heart-disease-mortality-data-among-us-adults-35-by-state-territory-and-county

Xu, J., Murphy, S. L., Kochanek, K. D., & Arias, E. (2022, December 22). *Mortality in the United States, 2021*. Center for Disease Control and Prevention. Retrieved February 24, 2024, from https://stacks.cdc.gov/view/cdc/122516