**Introduction**

The mortality rate attributed to heart disease in the United States is persistently rising. According to the Centers for Disease Control and Prevention (CDC), heart disease has consistently held the position of the primary cause of death, surpassing both cancer and COVID-19 in 2021 and 2020. Notably, the mortality rate due to heart disease continues to escalate annually, covering 20% of all deaths in the United States for the year 2021. The CDC defines heart disease broadly, encompassing major cardiovascular diseases such as acute myocardial infarction, coronary heart disease, heart failure, and strokes.

The selected dataset focuses on heart disease mortality rates among individuals aged 35 and above in the United States for the year 2014. The dataset provides mortality rates per 100,000 people for each county, incorporating age-adjusted data and a three-year average of individuals dealing with heart disease. Additionally, demographic variables including gender, race, and geographical coordinates are included for each county. This study aims to investigate the impact of gender, race, and geographical location on the likelihood of developing heart disease. By analyzing individual-level data within the dataset, the research seeks to raise awareness of how factors such as gender, race, and state of residence can influence an individual's susceptibility to heart disease and mortality outcomes.

**Data Cleaning/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 rows 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:

lower\_bound = Q1 - 1.5IQR

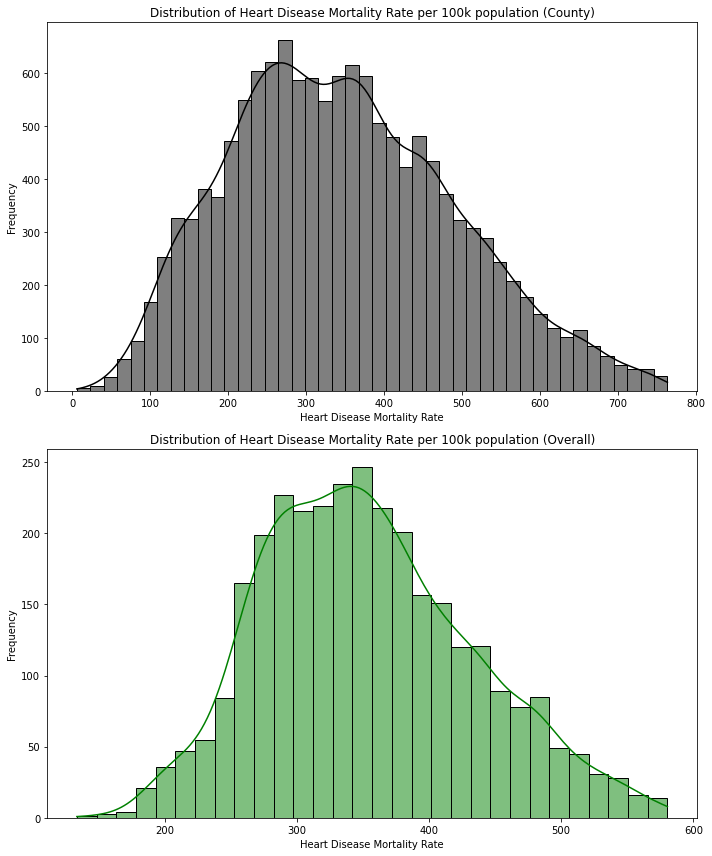
upper\_bound = Q3 + 1.5IQR

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.



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.A graph of different colored squares

Description automatically generated

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.

A blue and orange rectangles

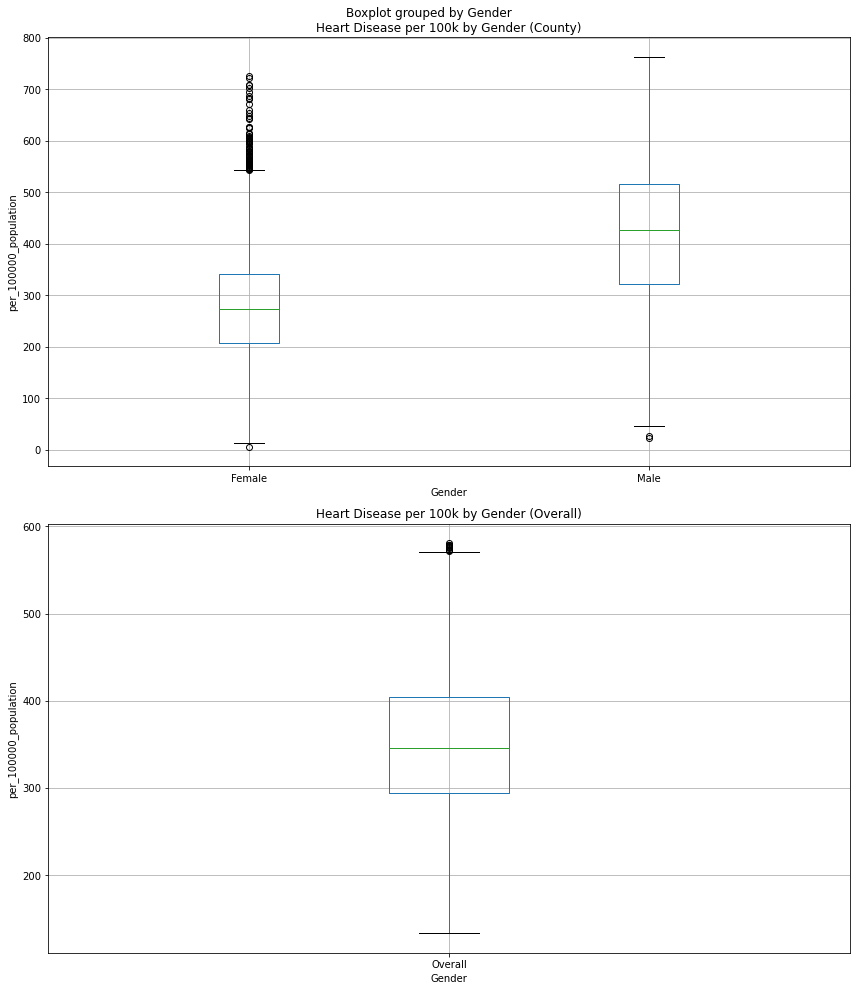
Description automatically generatedThe graph depicted above showcases the count of individuals categorized by gender in relation to the mortality rate. It is evident from the graph that the counts for males and females are closely balanced, indicating a satisfactory representation of each gender within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

A screenshot of a screen

Description automatically generated

Subsequently, 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.



Examining the bar graph, it is evident that males exhibit a higher overall count of individuals affected by heart disease compared to females. Notably, females display a notable number of outliers. When comparing these findings to the overall statistic, it becomes apparent that the upper limit is comparatively lower in the overall dataset compared to the individual-level data. Specifically, the ceiling is observed to be around 600 in the overall dataset, contrasting with a higher ceiling of approximately 750 in the individual-level data.

A screenshot of a computer screen

Description automatically generated

Upon examination of the race variable, 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.

The results of the chi-square tests are presented in the table below:

| **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 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 their 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 aforementioned 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.A screen shot of a graph

Description automatically generatedFemales is the lower sloped line.

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.

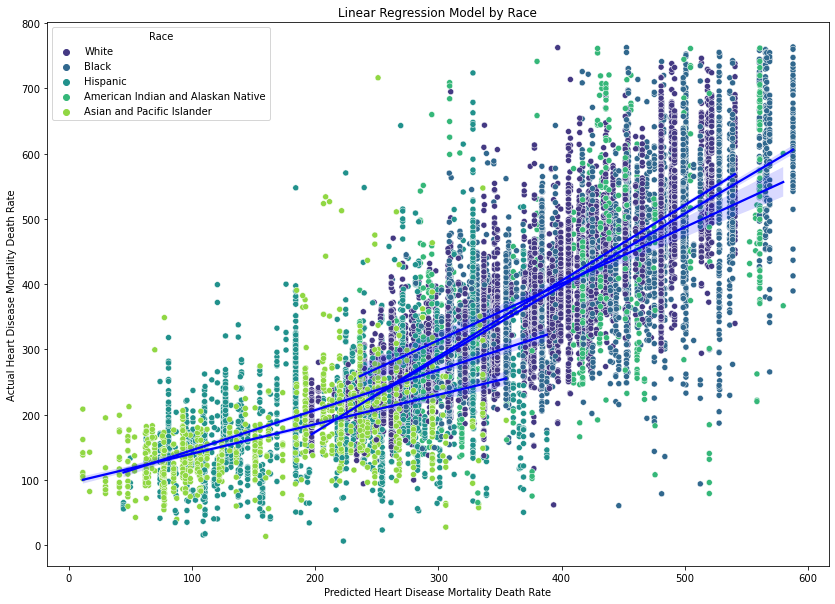
Alternative Hypothesis (H1): There is a significant difference in heart disease mortality rates between ethnicities.

\*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.

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. Similar to 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.

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.A white background with colorful dots

Description automatically generated

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.

A white background with colorful dots

Description automatically generated

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 and Recommendations**

Overall, significant impacts are observed within gender and ethnicity categories, with both contributing to patterns indicative of lower mortality rates. Specifically, females, Asian and Pacific Islanders, and Hispanics emerge as key subjects demonstrating these results, while black males exhibit the highest mortality rates. Further investigation into variables exhibiting significant impacts is recommended to fully understand the reasons behind their lower rates. Potential areas for exploration include cultural factors and dietary habits among different ethnicities. For females, examination of dietary differences compared to males, as well as variations in living standards and societal norms, could provide valuable insights.

As for Hypothesis 4, aimed at identifying significant impacts from states, additional testing is warranted. While the Chi-Square test suggested no association, linear regression revealed some states influencing the study, albeit with minor impact, representing only 16.6% of the model. To advance this study, a broader approach could involve exploring connections between the tested ethnicities worldwide and investigating environmental factors in different locations. This could shed light on whether climate conditions, such as colder or more humid climates, contribute to heart disease mortality rates among individuals over 35. Answering these questions could provide deeper insights into the factors contributing to middle-aged heart disease mortality.

Citations

(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>