The State of Our Hearts

Project Link:
https://chiemeka-nwakama.github.io/Heart-Disease-Awareness/

Chiemeka Nwakama, Julia Christenson, Lily Li, Ibrahim Ismail-Adebiyi, Emmanuel Mongare

INTRODUCTION

Heart disease is the leading cause of death in the U.S. and globally. There are many factors that play a role in cardiovascular health, and the goal of this project was to find what conclusions could be drawn from the various factors. Through an interactive webpage, the data findings were displayed through various visualizations.

DATA ANALYSIS

Data Preparation

In selecting datasets, we decided to search for both microdata (individual-level data) and aggregated state-level data. This was decided as a group to get both a specific and broad scope of the heart disease problem in the U.S. Since we also wanted to explore not just individual factors such as pre-existing health conditions and lifestyle, we did some literature review on these factors, as well as socio-economic factors which correlate with heart disease. According to the research, BMI, smoking, physical fitness, and diet all impacted cardiovascular health (https://pmc.ncbi.nlm.nih.gov/articles/PMC6378495/#section2-1559827618812395). This was expected. We found that heart disease prevalence increases after 35 years of age in both men and women and that specific ethnic groups with an increased risk of coronary artery disease (https://www.ncbi.nlm.nih.gov/books/NBK554410/). Lastly, doing some review on socio-economic factors, a 2018 publication stated that there have been four measures that are consistently associated with cardiovascular disease in high-income countries: income level, educational attainment, employment status, and local socioeconomic factors (https://pmc.ncbi.nlm.nih.gov/articles/PMC5958918/#:~:text=Socioeconomic%20status%20%28SES%29%20has%20a,a%20role%20in%20select%20areas). Given this background information on our topic, we decided to search for some datasets which capture these lifestyle factors and socio-economic factors. For simplicity's sake, we decided to search for aggregate state data to capture the socio-economic factors.

Our first dataset is from Kaggle, named “Indicators of Heart Disease (2022 UPDATE)”, which originally comes from the 2022 annual CDC survey data of 400k+ adults related to their health status. This one is tabular data with ordinal, nominal, and quantitative-ratio attribute types. This dataset includes over 18 columns of variables, but we decided to only keep a subset of these. The second dataset, “CDC Heart Disease Deaths by State 2014 - 2023,” is from the CDC

website. This data is tabular with quantitative-interval and quantitative-ratio attribute types. It contains the age-adjusted heart disease mortality rates and the total number of heart disease deaths per state per year. Our third dataset, “U.S Census Bureau Median Household Income by State 1984 - 2024,” comes from the United States Census Bureau. The data contains median household income data for each state for each year since 1984, complete with standard error. The data come in two groups, one with current (as of 2025) dollar amounts and another with income adjusted for the 2024 dollar. Lastly, we gathered aggregated data from IPUMS NHGIS. This publicly available data is at the state-level and includes tables: Educational Attainment for the Population 25 Years and Over, Employment Status for the Population 16 Years and Over, and Types of Health Insurance Coverage by Age.

After downloading this tabular data, we preprocessed the data before loading it or extracting key information into our visualization. For all rows, we removed any with null values. This mostly applied to the Kaggle dataset, which was a large dataset that had many null values for the factors we were looking at. We also only included certain columns in our dataset due to size and factor analysis. This includes age, race, angina status, stroke status, COPD status, depression, Body Mass Index, sleep quality, smoker status, alcohol usage, diabetes, depression, and whether or not the individual has had COVID or a heart attack. For simplicity's sake, we determined an individual as heart disease positive if they have angina, have had a stroke, or have had a heart attack. For state-level data, we merged it into one large dataset so we could easily make it into an interactive visualization.

Exploration & Analytical Methods

Initially, we looked at the data distributions with Excel and Jupyter Notebooks. This includes viewing the data by state and looking for any potential outliers.

Top Causes of Death: Line Chart

To emphasize the importance of heart disease awareness in the U.S., we wanted to show how prevalent it is compared to the top causes of death nationwide. We specifically chose a dataset from the National Vital Statistics System covering the top causes of death in the U.S. from 2017 to 2022 so it would align with the time period of our other data. This dataset includes the number of deaths for the top 12 causes of death in the U.S. during those years.

For clarity and comparison, we decided that a line chart was the best way to display how these causes of death changed over time. To make the visualization more engaging, we added a scrolly feature that shows year by year how the rankings of these causes shift. We also included a tooltip that displays each cause’s ranking out of 12, along with the specific death toll, allowing users to take a closer look at which causes were most prevalent.

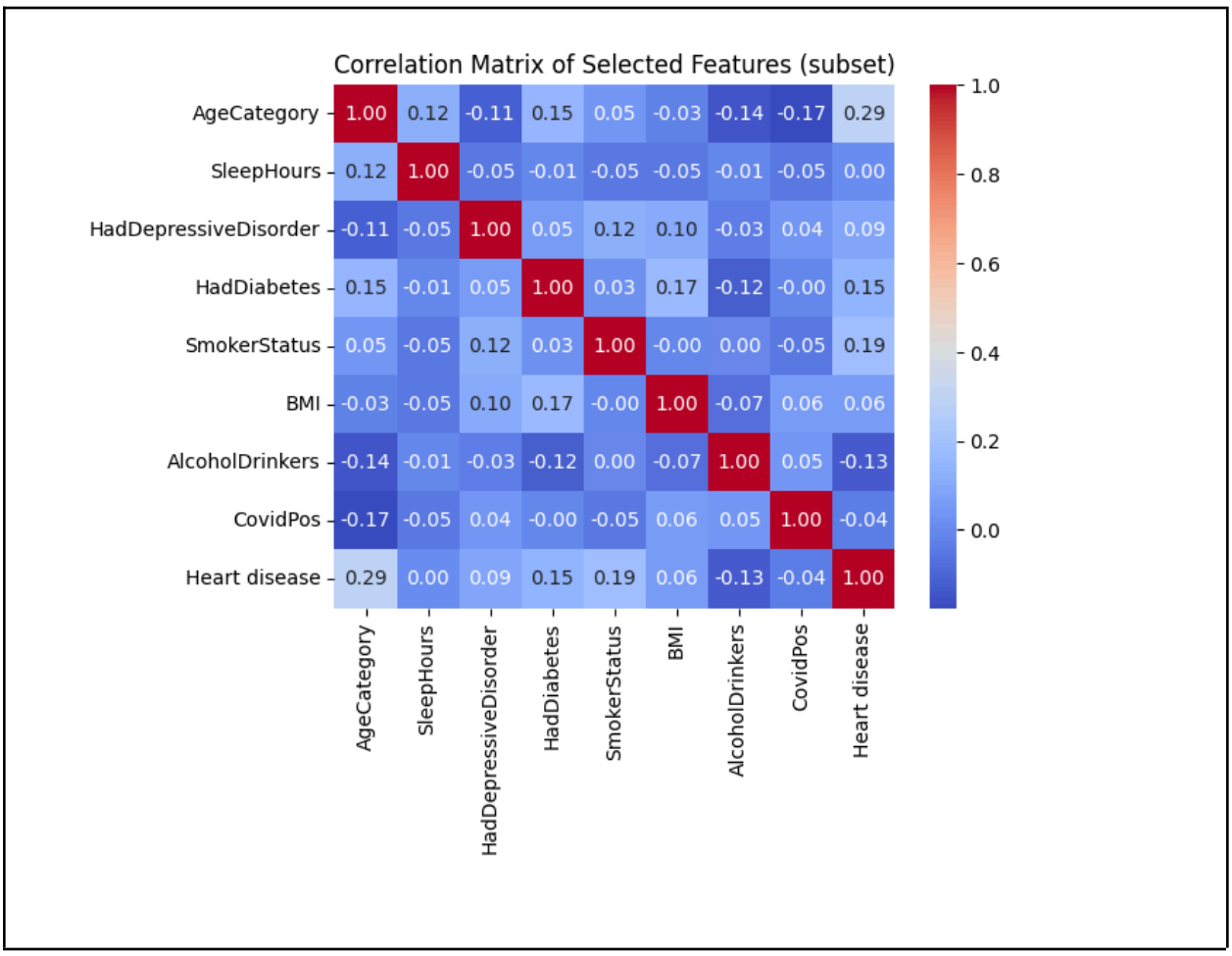
In cases where data was missing, we displayed the value as 0 but clearly explained in the tooltip that this represents zero reported deaths, not that there were no deaths from that cause during that year. To further highlight how severe heart disease deaths are in the U.S., we added a card on the left side that shows the top cause of death for each year in bold. This turns out to be heart disease every year. Alongside this, we included brief descriptions explaining what was

happening during each year, such as the onset of COVID, and short analyses of what changed from the previous year or what trends were occurring among the leading causes of death.

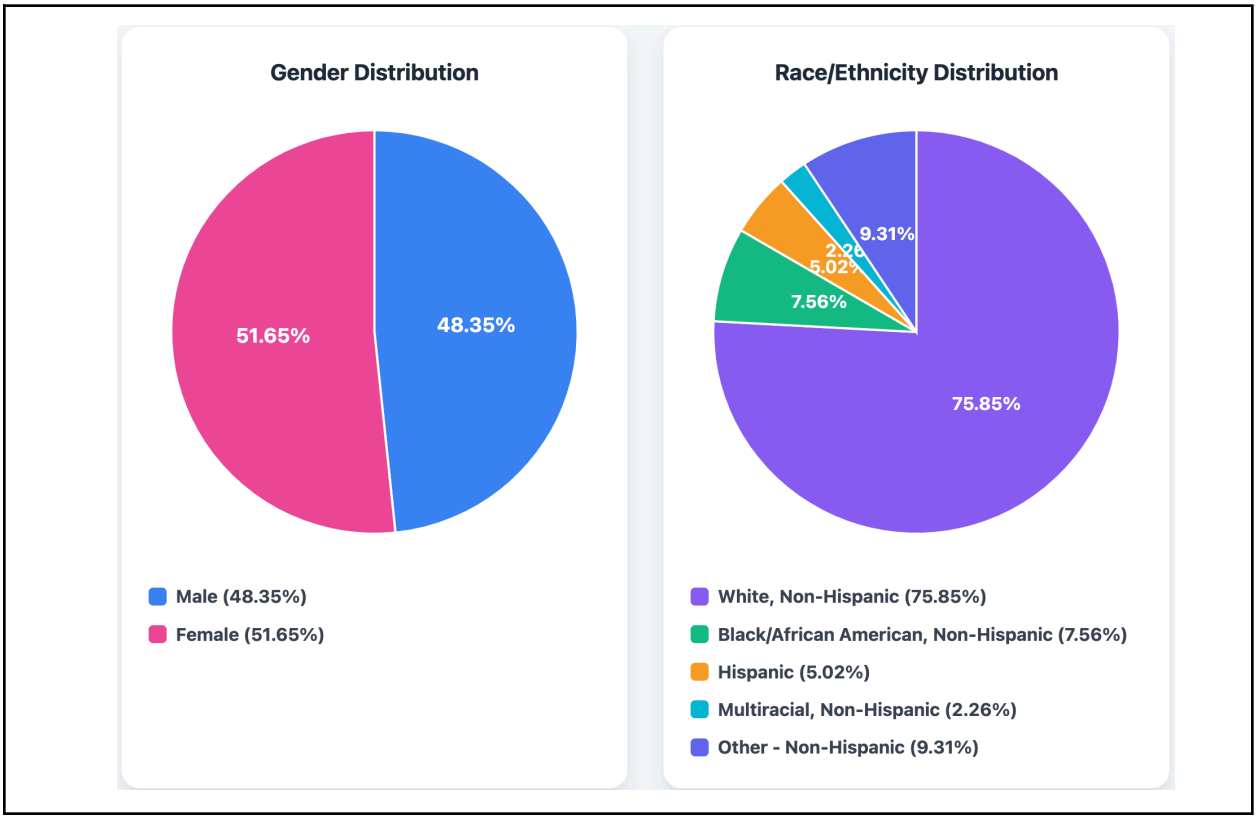
Overall, this visualization serves as a strong introduction to why we are examining the causes of heart disease and who is most at risk. It clearly shows that heart disease-related death has been and still is consistently relevant throughout the years, which makes the topic immediately meaningful to the viewer through the data presented.

Demographic Variables: Heatmap and Pie Chart

When looking at the state data for heart-disease-related mortality rate and the median income level, we found there to be a correlation between the two. We also used Seaborn in Python to visualize pairplots between variables within the Kaggle dataset. Since the dataset was so large, we decided to use a randomized sub-sample of 2,000 individuals from each state to load. We also decided to only keep a subset of the variables for this specific visualization and keep other variables for a different part of the website.

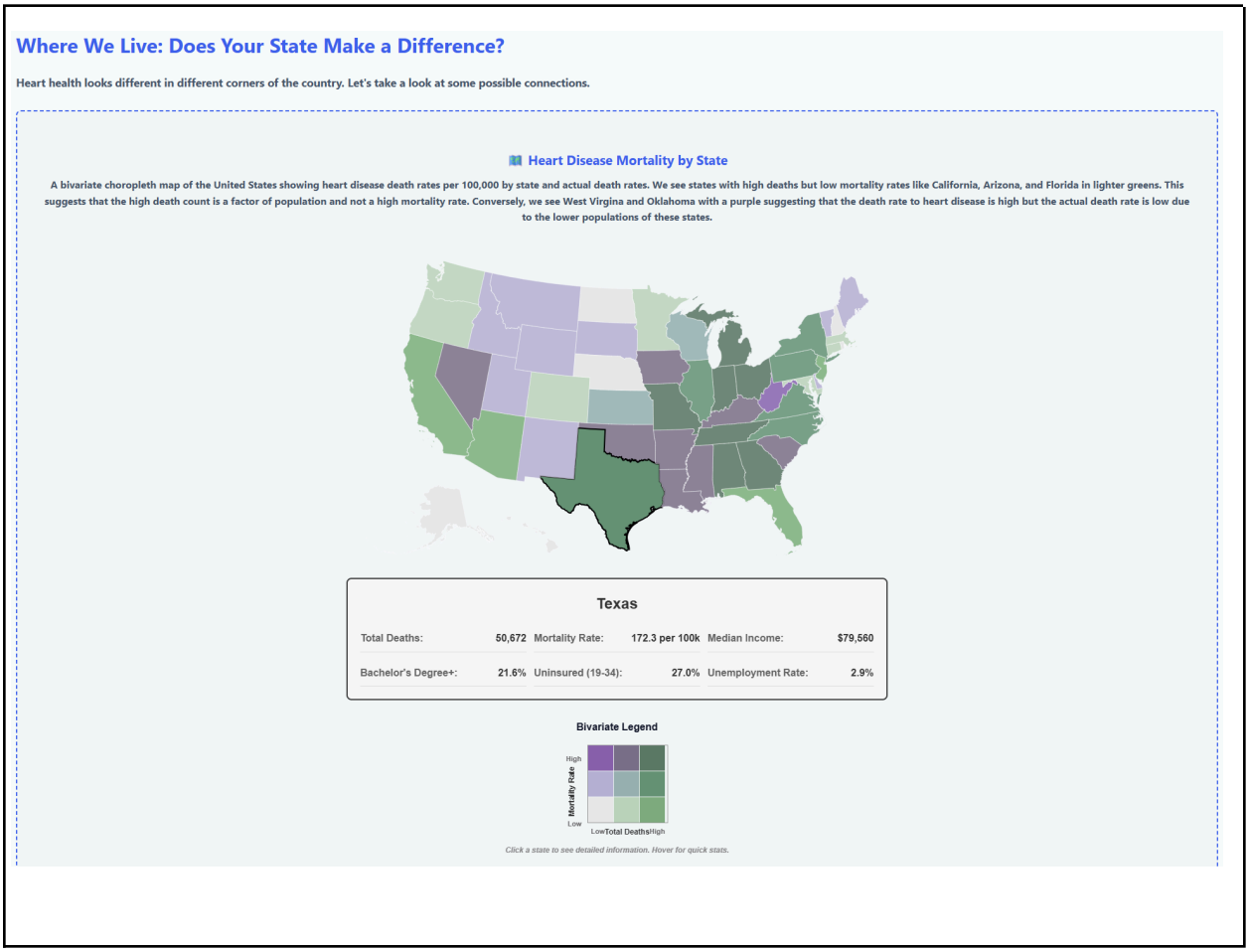


As we continued to explore the Kaggle dataset for more insights, we were curious to see if demographics played a significant role in heart disease. Using python we were able to pre-process our entire kaggle dataset and found distributions for both race and gender in our data. These values were stored and passed to our visualization in order to save time and resources when the site is first loaded. Once these distributions were found, it was decided that simplicity would be the most effective approach, and we decided to create two pie charts showing the gender and race/ethnicity percentages of our data.



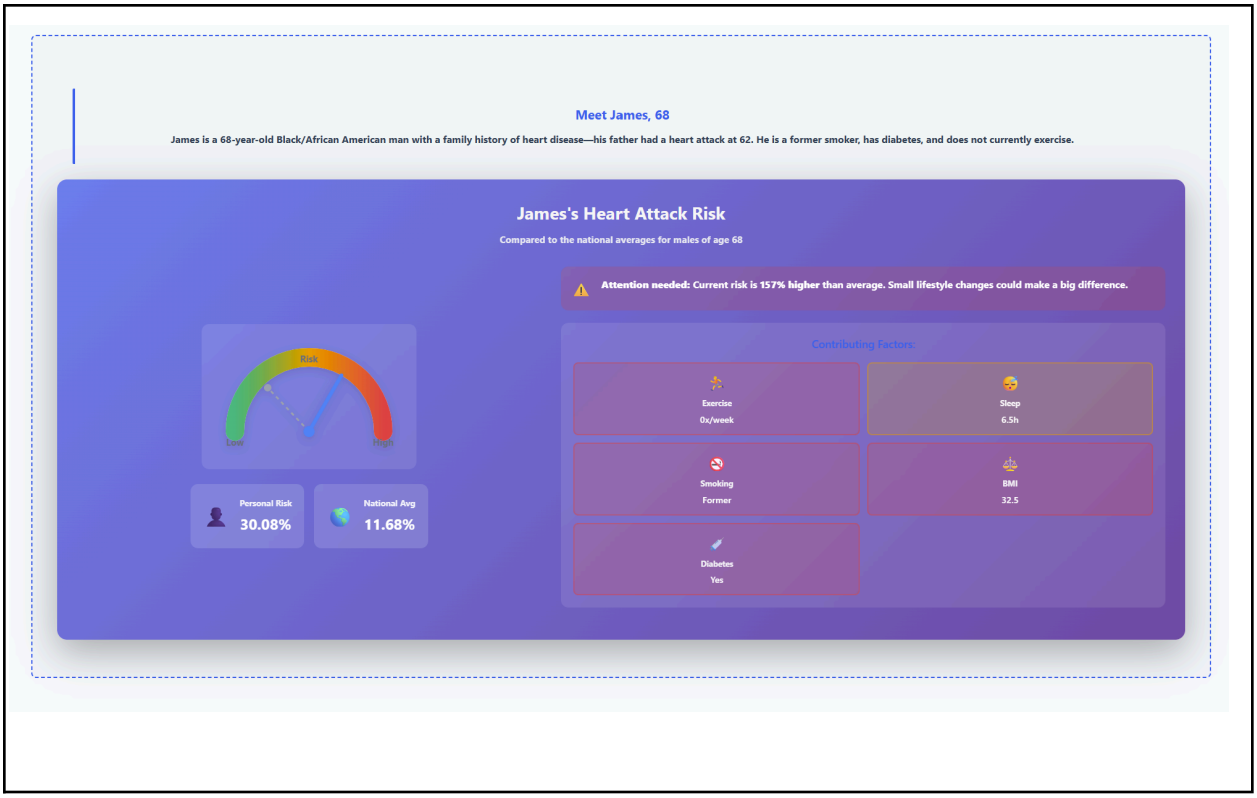
Geographical Factors: Bivariate Map

To analyze the differences in heart disease outcomes across geographic regions, we combined state-level data about mortality rates and actual death counts. We visualized this data using a bivariate choropleth map to separate the effects of death rates from the size of the state's population. By displaying the mortality rate and total deaths simultaneously, we could find states where high death counts were driven primarily by the large population of the state, as opposed to states where smaller populations still experience disproportionately high mortality rates. We chose this visualization because it answers our story's central question of what things affect a person's risk of heart disease. What we saw was that geography might have structurally differing ways of affecting heart disease rates and outcomes.



Risk Calculator: Radial Gauge

For our individual “risk-calculators”, we aggregated data on demographic and lifestyle factors and compared them to national averages in different categories for the individual being assessed. For our analysis, we considered established risk factors, including diabetes, BMI, and smoking history. We chose this visual because it allowed us to translate abstract statistics into a real-world example. This would allow us to make our data more interpretable and impactful. We also highlighted the gap between our fictional characters and national averages to ground the narrative about risk factors.



Correlations: Scatterplot

To assess how daily habits relate to heart disease outcomes, we made an exploratory visual to show correlations between state-level lifestyle indicators and mortality/disease rates. We did this with a scatter plot that we made interactive with a fitted trend line. We also quantified and displayed correlations so users could see the strength of correlations between factors. We chose this because it allowed us to communicate how some lifestyle factors were important, but not the main drivers of heart disease, in a visual way.

Selecting Insights

The primary objective of this project was to explore the connections between heart disease indicators/factors. After the initial data analysis, it was decided to refine this question further to cover three questions:

1. How do lifestyle factors correlate with heart disease prevalence?
2. Do demographic factors correlate with the likelihood of a heart disease diagnosis?
3. Which states have the highest and lowest heart disease death rates, and what socioeconomic factors may contribute to them?

DATA COMMUNICATION

Alternative Designs

Sunburst Diagram to Heatmap

Initially, we considered using a sunburst diagram to represent the hierarchical structure of the data. However, during development, we found that the sunburst did not convey comparisons and patterns as clearly as intended, particularly across categories. To improve clarity and reduce cognitive load, we transitioned to a heatmap. The heatmap provided a simpler and more intuitive way to identify trends, variations, and outliers, allowing users to quickly compare values across dimensions.

Adding Interactivity to All Visualizations

The original design relied primarily on static visualizations. As the project evolved, we recognized that interactivity could significantly enhance user engagement and understanding. We redesigned each visualization to include interactive elements, such as hovering over or selecting a state to view detailed information. This allowed users to explore the data at their own pace, focus on areas of interest, and gain deeper insights without being overwhelmed by all the information at once.

Expanding the Variety of Visualizations

Early iterations of the project used a smaller number of visualization types. Over time, we realized that different aspects of the data were better communicated through different visual forms. By incorporating a wider variety of visualizations, we were able to present results more effectively and highlight multiple perspectives within the data. This approach improved overall comprehension and helped convey complex findings in a more accessible and meaningful way.

Final Design Decisions

Decision 1: Creating Personas

Personas were created to make the data more relatable by grounding abstract statistics in real-world examples. This approach helped users better understand how the data could represent individuals and situations in everyday life, strengthening the overall narrative of the project.

Decision 2: Interactive Visualizations and Pre-Computed Metrics

We incorporated interactive visualizations to give users control over what aspects of the data they wanted to explore, helping reduce information overload. In addition, calculating certain values ahead of time rather than dynamically loading the entire dataset improved site performance, resulting in faster load times and more responsive interactions.

Decision 3: Breaking the Primary Question into Smaller, Focused Questions

Rather than addressing one broad question, we divided it into smaller, more manageable questions that allowed for more focused exploration. This decision improved the readability and organization of the site, making the content easier to follow and understand at a high level.

REFLECTION

Work Process and Distribution

Our initial milestone/schedule planning was split into a few main phases: initial proposal, data analysis and visualization, proof of concept, adjustments, and then the final delivery. We didn't follow that plan exactly, but it served as a good general structure to ensure that we were all on the same page and were on pace. In hindsight, this was a pretty good plan for us, but it could still be improved. We had good communication within our group, but it would have been helpful to receive more external feedback, so that we could better determine what needed to be adjusted.

After deciding on the topic and the key questions we wanted to explore, we collectively selected the datasets. From there, we discussed how we could answer those questions and what visualizations would best represent the answers. We had an open discussion on how we can best visualize the data. Once we had selected our main visuals, many of our design choices were made individually, and we worked on our own visuals throughout the process. However, nearing the end of the project with the final slides, we grouped to discuss the entire website and what can be improved.

When we were lost on which factors to group together, reviewing the literature helped us understand *why* the data might look a certain way. It gave us some insight and some additional validation of the data. We did notice that the subset of the Kaggle data did not completely agree with the statistics we found in the published papers or reliable sources like the American Heart Association. This was a bit frustrating because it was contradictory, but when presenting, we made sure to talk about this and how the dataset does not capture the entire population.

Some challenges we encountered include data loading and miscommunication. The original plan was to load all of the data into the site. This quickly proved to cause some issues, since we had a lot of data, and it was very slow to load in. We had two solutions to this: aggregating the data and calculating some of the numbers beforehand, and visualizing the pre-calculated results. There was some miscommunication on what factors were being explored per visualization. Since we worked on our individual visualizations and then re-grouped, we realized some factors were being double-covered on the website, when it was not necessary. After sitting down and discussing each question during our meetings, we all gained a better understanding of what should be included in each section, such that it was relevant and wouldn't be duplicated in other areas.

Evaluation

For an evaluation of our final product, we asked three friends about how much they knew about heart disease. To our surprise, many of them didn't know that heart disease has been the leading cause of death in the US for many years. We started by asking them a few general questions about heart health/disease: what did they know about it? What impact do they think it has? What factors, both controllable and noncontrollable, impact heart health? One friend stated they knew that physical activity and cholesterol made an impact. The other friend mentioned smoking and drinking. The last friend mentioned smoking, physical activity, and diet. When asked to briefly answer these questions, they were prompted to conduct some of their own research. All three of them searched for heart disease factors on a search engine like Google and found many of the factors that are shown in our visualization, with others like diet, stress, high blood pressure, high cholesterol, and family history. Their conclusion was very general and vague; there's a lot of information and data on the internet, but the feedback was that it's difficult to parse through all of it to find the relevant answers and how/if they correlated with each other.

Then, we showed them our site with the interactive data visualizations. They could easily follow the story, especially with the three big overarching questions. They could explore the data themselves and easily make their own connections and conclusions through our interactive visualizations. The overall feedback was that our visualizations were helpful and did a great job conveying the answers to the questions, and it gave them a better idea of what heart disease factors there are and how they may be correlated. One of them mentioned that they would want to see more data on how family history and genetics contribute to heart disease. They enjoyed playing with the calculator, since it very quickly demonstrated how even small changes can change how much people are at risk of heart disease. However, another user mentioned that they would have wanted to see the accuracy rate of the calculator.

CONCLUSION

Limitations and Future Improvements

This analysis was limited by the scope of the available data. We focused exclusively on the United States and relied on data from only a few selected years. Expanding the dataset to include additional years and locations outside of the U.S. could reveal broader trends and improve the generalizability of the findings. Additionally, the calculator used in this project was relatively simplistic. Future work could involve adopting a more sophisticated model to produce more accurate and reliable estimates.

Final Thoughts

Heart disease is a critical topic with direct implications for human life. Through our visualizations, we were able to better understand the relationships among various contributing factors and gain a more holistic view of the state of heart disease in the United States. Throughout this project, we also strengthened our ability to select, process, analyze, and visualize data in efficient and effective ways, reinforcing the importance of thoughtful data-driven analysis.

REFERENCES

Datasets

Dataset 1: Indicators of Heart Disease on Kaggle
Dataset 2: CDC Heart Disease Deaths by State 2014 - 2023
Dataset 3: U.S Census Bureau Median Household Income by State 1984 - 2024
Dataset 4: NHGIS Datasets from 2022, grouped by state-level data
    Educational Attainment for the Population 25 Years and Over
    Employment Status for the Population 16 Years and Over
    Types of Health Insurance Coverage by Age
Dataset 5: National Vital Statistics System U.S. Leading Causes of Death 2017–2022

Literature Review

Risk Factors of Heart Disease
Lifestyle Strategies
Socioeconomic Factors and Cardiovascular Outcomes

Website Citations

https://my.clevelandclinic.org/health/articles/heart-problems-after-covid
https://www.cdc.gov/diabetes/diabetes-complications/diabetes-and-your-heart.html
https://www.hopkinsmedicine.org/health/conditions-and-diseases/depression-and-heart-disease
https://pmc.ncbi.nlm.nih.gov/articles/PMC6378495/#section2-1559827618812395
https://www.heart.org/en/healthy-living/healthy-lifestyle/how-to-help-prevent-heart-disease-at-any-age