University of Virginia

Project 1 - Factors that affect Access to Healthcare

Divya Kuruvilla and Glory Gurrola

DS 3001

Professor Johnson

26 October 2023

Summary

The goal of this project was to practice data wrangling, exploratory data analysis and visualization to examine data pulled from the General Social Survey and derive conclusions for the research question at hand. The General Social Survey (GSS) is a widely used sociological survey conducted to collect data on a wide range of social, demographic, and personal attitude topics. The survey has been conducted regularly since 1972 and it provides insights to trends and changes in society over the years. For this project our research question was "How does social class, gender, or income affect access to healthcare?", so the relationship between income, social class, and gender was observed to see how these different factors affect access and quality to healthcare.

Access to healthcare is a fundamental right, yet there exist disparities based on certain socioeconomic factors and gender persist in our society. Our aim was to reveal the trends existing between income, social class, gender, and their influence on access to and the quality of healthcare services. In order to analyze trends of healthcare access and opinions on healthcare access, we collected variables related to income, social classs, and gender from the GSS and performed operations on them to see trends and relationships between people's social status or income and their opinions on different healthcare factors. The data was cleaned to handle any discrepancies, and then different visualizations, such as histograms, heatmaps, stacked bar charts, and kernel density plots, were created to illustrate relationships between the different variables.

Income plays a crucial role in a person's ability to afford health care and oftentimes those with more disposable income are able to easily purchase private care that offers better coverage and access to a broader network of healthcare. By examining public perceptions and experiences related to health care, we wanted to gain a new understanding of the issues that are present in the healthcare industry, especially relating to those less fortunate than others, and aim to promote equitable access to healthcare by shedding light on the issues present in society.

Data

In this section we will be talking about the process of collecting data from the GSS and the steps taken to clean the data and proceed to do analysis on the variables and trends present in what was collected. Through the process of data wrangling, we were able to draw meaningful insights from the collected information. Since the GSS dataset has over 5,000 variables, we started our data collection by using the Pandas library to process each chunk of data separately and only saving the relevant variables that we chose for our project. We decided on keeping 25 variables that were related to our research question of how income or social status affects access to health care. We changed the names of the variables in order to make it easier to understand what each is representing. Before changing the names, lots of the variables had unclear naming such as "doc13," "helpsick," or "satfin." Renaming variables is very important in data wrangling because it enhances the readability of the code so that if others are reading or reviewing the code, it is more comprehensible. Having clearer names also prevents errors that could occur because of ambiguous or inconsistent naming.

After saving all of the variables from the GSS Data, we took steps to start cleaning all of the variables. To ensure the clarity and effectiveness of our visualizations and graphs for our analysis, we applied specific data cleaning techniques to handle the missing values in each of our categorical variables. Lots of the categorical data had missing values, NaN, so we opted to replace these with 'no response' instead of having an abundance of Not a Number data types. This decision was driven by the fact that having missing values in our categorical variables would not be accommodating for creating our visualizations. This data cleaning approach not only enhances our data graphics but also ensures that our analysis is based on a more complete dataset, allowing us to better understand the relationships and patterns between the data relating to social status, income, and wealth, and health related data points.

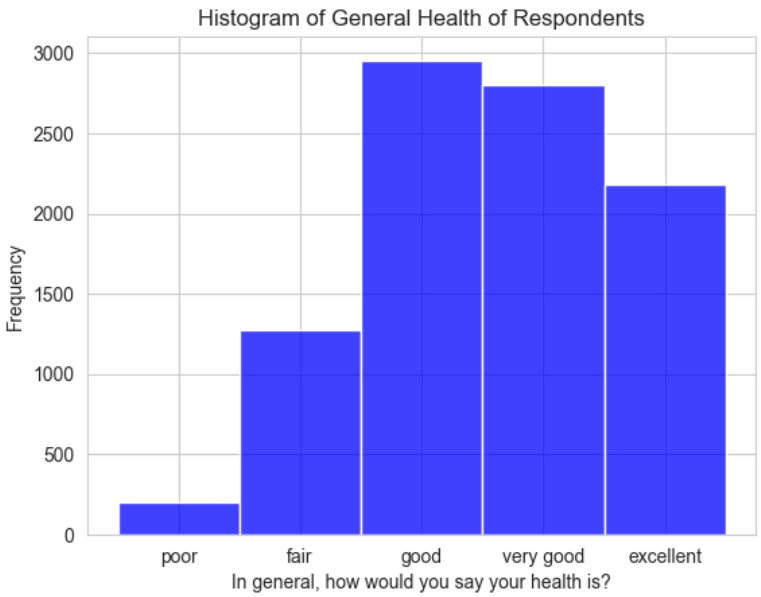
Another step that was taken in order to clean the data was dropping certain observations from the data frame. Since our research question was centered on the investigation of how income and social status can affect access to healthcare, it was very important to have a complete dataset relating to certain variables. Therefore, we took the step of excluding all observations that contained missing values within the income variable. Since it was not useful to keep the NaN values for our analysis and making visualizations, we dropped them from the dataset.

Since the GSS data has over 30 years of responses, and the survey has evolved and changed over the years, some of the variables only have responses from certain years, while others have lots more observations because the same question was asked every single year. For example, the question asking about annual income was asked every single year which made it have over 60,000 observations, while a question about if the person was worried about having limits to their insurance was only asked in 2002. This appeared to be a challenge at first because we had an abundance of missing values for certain variables because the dataset had varying amounts of observations for different columns which left us with lots of NaN data types in the data frame. To address this issue and maintain the completeness of our dataset, we implemented the approach of filling in the missing values with 'no response' so that we were still able to produce appropriate graphs and visual representations of the data.

Having successfully addressed missing values and ensured data consistency, our dataset stands as a reliable foundation for generating insights about the relationship between income, social status, and how it affects access to health care. With our dataset now thoroughly cleaned and complete, we turned our attention to creating visuals to represent the connections in our data.

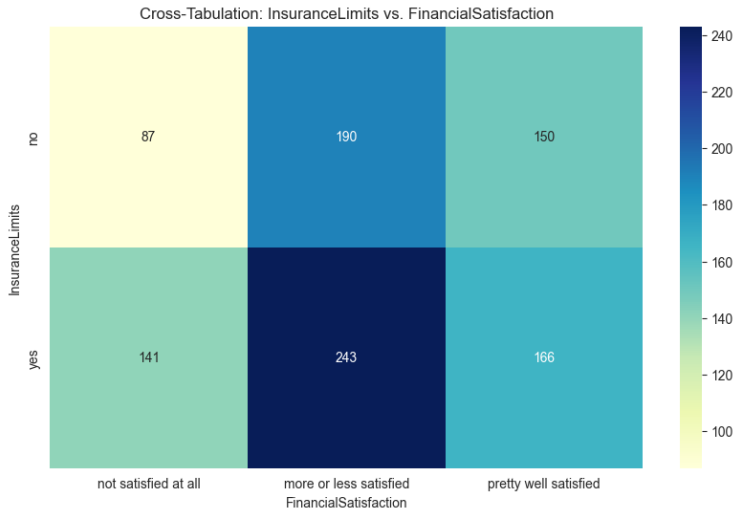
Results

Histogram of General Health of Respondents Analysis



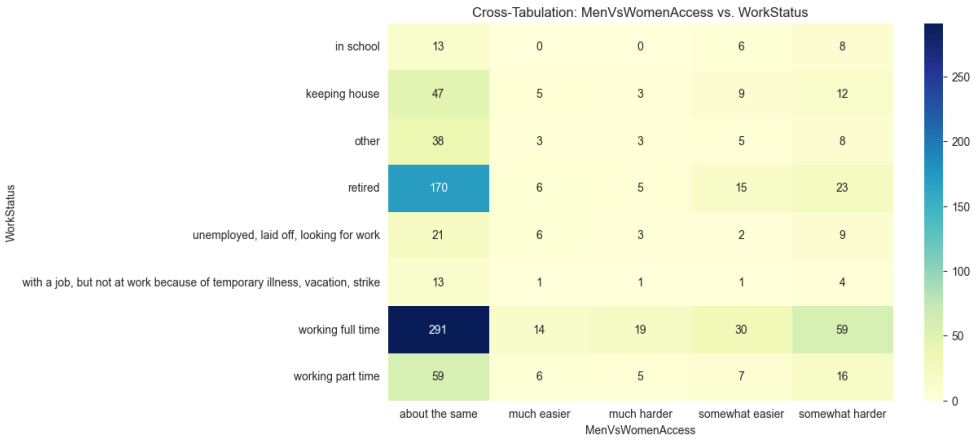
This histogram illustrated the distribution of participant responses when asked how they would evaluate their general health. Responses were categorized from ‘poor’, ‘fair’, ‘good’, ‘very good’, and ‘excellent.’ A small number of respondents rated their health as ‘poor’, while the frequency for those who described their health as ‘fair’ was slightly larger. A majority of the respondents described their health as ‘good’, making it the prominent category, but a large number of respondents considered their health to be ‘very good’ as well. Lastly, those who viewed their health as ‘excellent’ formed a smaller group compared to the ‘good’ category, but still represented a substantial portion of the overall responses. In summary, the majority of participants seem to have a positive perception of their health, with ‘good’ and ‘very good’ being the most prevalent categories. Additionally, a positive trend, from ‘poor’ to ‘good’, was reflected in self-perceived health among the participants.

Heatmap for InsuranceLimits vs. FinancialSatisfaction



This cross-tabulation graph presents the relationship between respondents' belief regarding their health insurance limitations and their financial satisfaction. On the vertical axis, individuals are categorized based on whether their health insurance has limitations by being represented with 'yes' or 'no.' On the horizontal axis, individuals' financial satisfaction was measured and represented with 'not satisfied at all', 'more or less satisfied', and 'pretty well satisfied.' Of respondents without insurance limitations, 87 said they were not satisfied with their financial status, while 190 respondents believe they are 'more or less satisfied.' Additionally, 150 of these respondents believed that they were 'pretty well satisfied' with their financial situation. On the other hand, individuals with insurance limitations, 141 respondents reported that they were 'not satisfied at all' financially, 243 responded that they were 'more or less satisfied' financially, and 166 respondents believed that they were 'pretty well satisfied' with their financial situation. In summary, it was observed that the largest group of respondents consists of those with insurance limitations who were 'more or less satisfied' with their financial standing. Even though these respondents had limitations with their insurance, they still expressed moderate satisfaction.

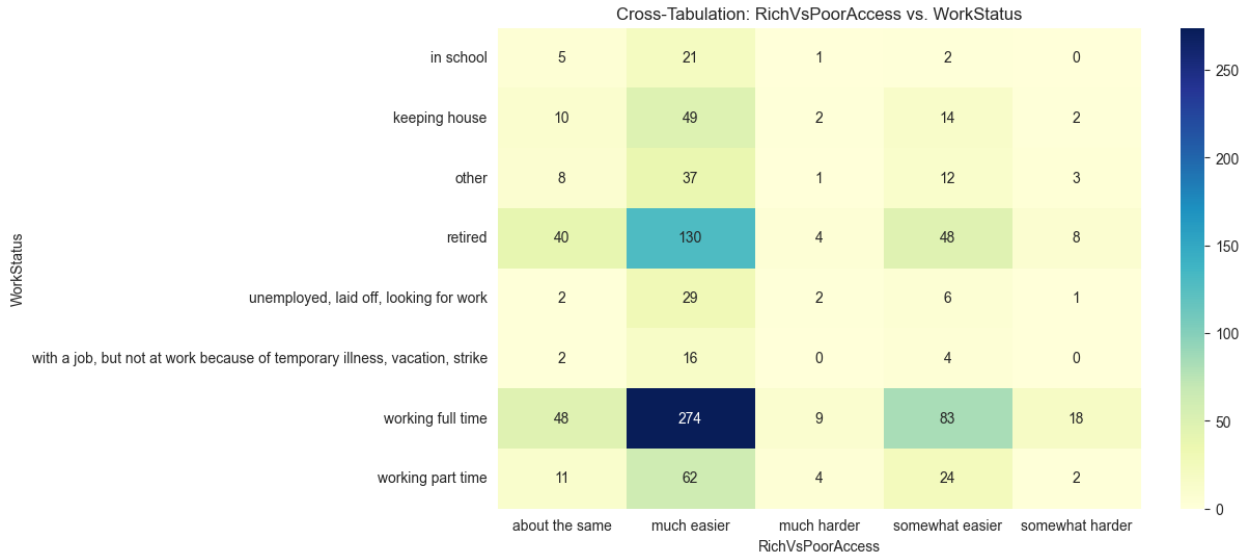
Heatmap for MenVsWomenAccess vs. WorkStatus



This heatmap is a cross-tabulation visualizing the relationship between respondents' work status and their perceptions concerning the ease of healthcare access for men versus women. The vertical axis outlines the various employment categories, such as 'in school', 'keeping house', 'other', 'retired', 'unemployed, laid off, looking for work', 'with a job, but not at work because of temporary illness, vacation, strike', 'working full time', and 'working part time.' 'The horizontal access represents the ease of healthcare access for men versus women, with possible values 'about the same', 'much easier', 'much harder', 'somewhat harder', and 'somewhat easier.' The color intensity in the heatmap indicates the count of responses, with darker shades representing higher counts. Interestingly, the majority of respondents in the 'working full time' category, amounting to 291, perceive that access for both genders is 'about the same.' In the United States of America, it is common for employers to sponsor health insurance for their employees. Thus, respondents who are working full time, might have had the perception that they have similar health benefits as their male and female coworkers, leading them to think that

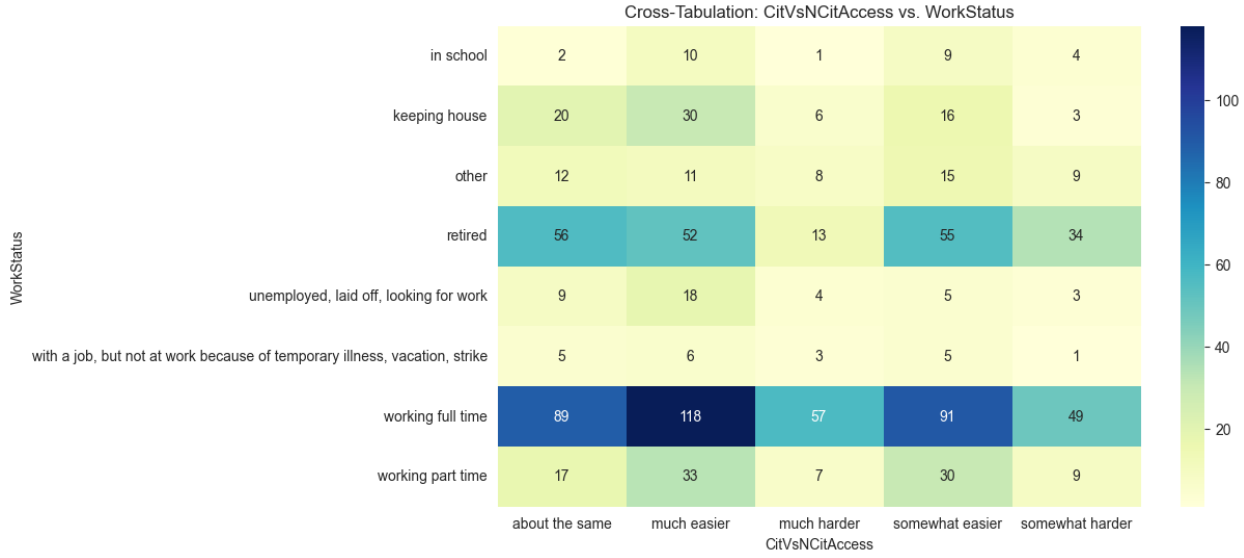
men and women have equal access. Similarly, about 170 respondents in the 'retired' group also feel that access to healthcare for both genders is 'about the same.' The retired group might have respondents who are most likely eligible for Medicare, which is a national healthcare program that provides health insurance for Americans older than 65. Since access to this healthcare program is available to all eligible seniors, irrespective of gender, retirees might think that access to healthcare is equal for both genders. Additionally, there may be nuanced disparities in healthcare access related to gender in the USA, and not everyone is aware of or has experienced these disparities. Therefore, respondents believe access to healthcare is the same for both genders. However, it is important to recognize that these are just possible speculations for the observed findings. Overall, the overwhelming belief among all work status categories leans towards the belief that men and women have 'about the same' access to healthcare.

Heatmap for RichVsPoorAccess vs. WorkStatus



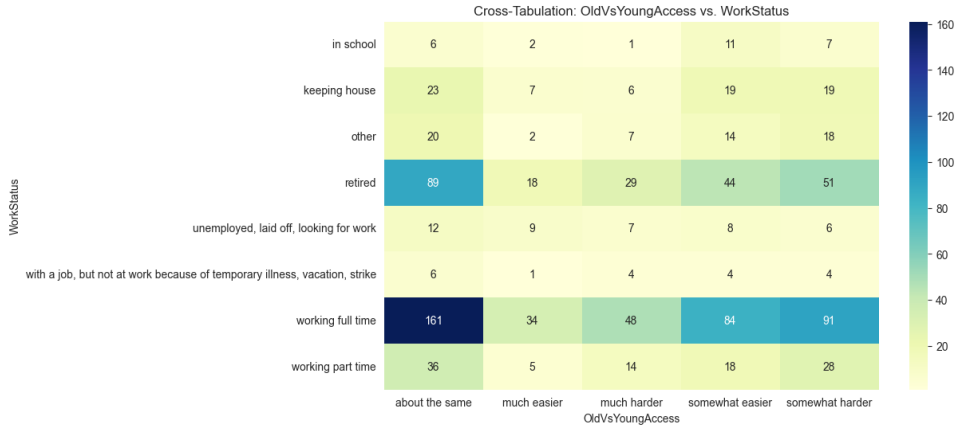
This heatmap provides a cross-tabulation that depicts the respondents' perceptions of healthcare access disparities between rich and poor individuals in the USA, categorized by their work status. The horizontal axis represents the perceived ease of access to healthcare for the rich vs. poor, with values ranging from 'about the same', 'much easier', 'much harder', 'somewhat harder', and 'somewhat easier.' The vertical axis outlines various employment categories, such as 'keeping house', 'retired', 'working full time', among others. The majority of those 'working full time' as indicated by the 274 respondents, believe healthcare access for both socioeconomic groups are 'about the same.' Similarly, 130 respondents from the 'retired' category believe that healthcare access for both socioeconomic groups are 'about the same.' However, it is important to note that there are varying opinions in the other work categories. For example, 49 respondents who identify as 'keeping house', as their work status, believe that healthcare access is 'much easier' for those who are rich, while 2 respondents believe it is 'much harder.' In summary, the overwhelming belief among all work categories leans towards the belief that the rich have it 'much easier' to access healthcare.

Heatmap for CitVsNCitAccess vs. WorkStatus



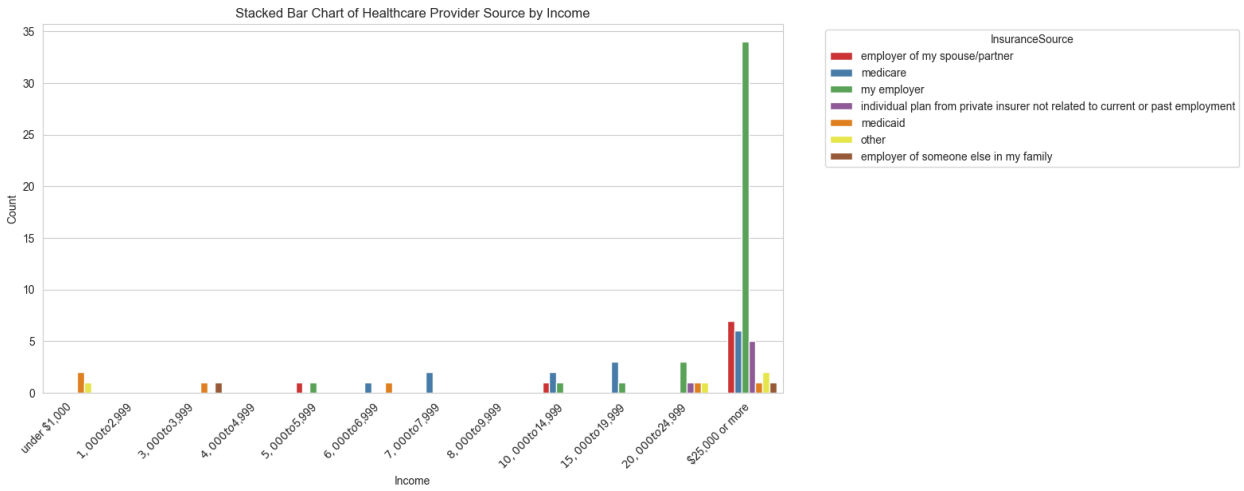
This cross-tabulation graph illustrates the relationship of respondents' work status and their perspectives concerning the ease (or difficulty) of healthcare access for citizens versus non-citizens. Similarly to the two previous heatmaps, the vertical axis classifies individuals by their employment status while the horizontal axis represents perceptions about healthcare access disparities between citizens and non-citizens. For respondents' 'working full time', 118 respondents believe that citizens have 'much easier' access to healthcare compared to non-citizens. On the other hand, 49 respondents think that it is 'somewhat harder' for citizens to have access to healthcare. It is important to note that in the 'retired' category, the opinion for ease of healthcare access for citizens and non-citizens appears divided; 56 respondents believe that access is 'about the same' for both groups while 55 respondents feel it is 'somewhat easier' for citizens. Additionally, there is a variety of perspectives in other employment categories as well. For instance, 30 respondents 'keeping house' believe that healthcare access is 'much easier' for citizens while 16 respondents believe that it is 'somewhat easier' for citizens. In summary, a large majority of respondents from various work statuses perceives that citizens have 'somewhat easier' or 'much easier' access to healthcare than non-citizens.

Heatmap for OldVsYoungAccess vs. WorkStatus



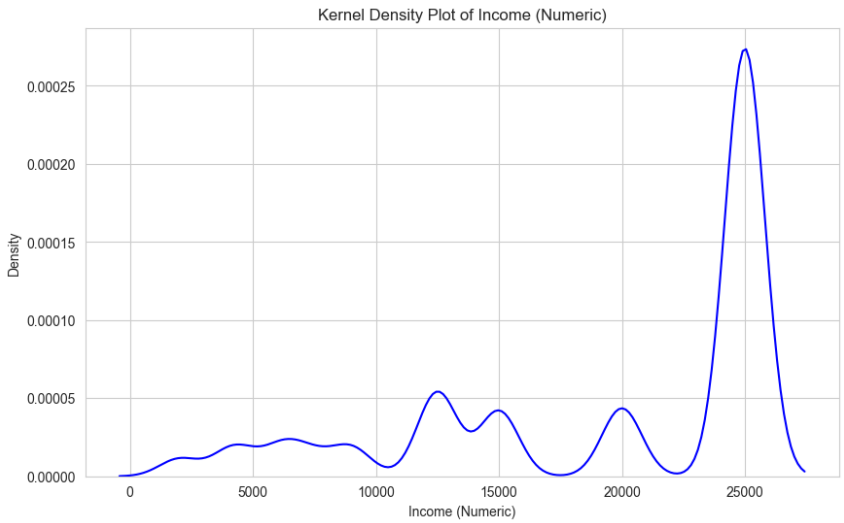
This heatmap presents the relationship between individuals' work status and their perceptions concerning the ease of healthcare access for older individuals versus younger individuals. Like the other heatmaps, the vertical axis represents the current work status of respondents while the horizontal axis represents perceptions about health care accessibility disparities between the elderly and the youth. Among respondents who were 'working full time', 161 respondents think that healthcare access remains relatively constant and 'about the same' for both age groups. However, 91 respondents who were 'working full time' believe that it is 'somewhat harder' for the older generation to access healthcare, while 84 respondents believe that it is 'somewhat easier' for older individuals. It was also observed that the 'retired' group believes that healthcare access is equitable, with 89 respondents believing that access is 'about the same.' Regardless, 51 respondents from the 'retired' group believe it is 'somewhat harder' for older people while 44 respondents feel that it is 'somewhat easier.' Overall, the responses display mixed beliefs across different work statuses. A large portion of participants from diverse work backgrounds perceive equal healthcare access for both the elderly and young; however, there is a substantial number of respondents that believe healthcare access is either 'somewhat easier' or 'somewhat harder' for the older population.

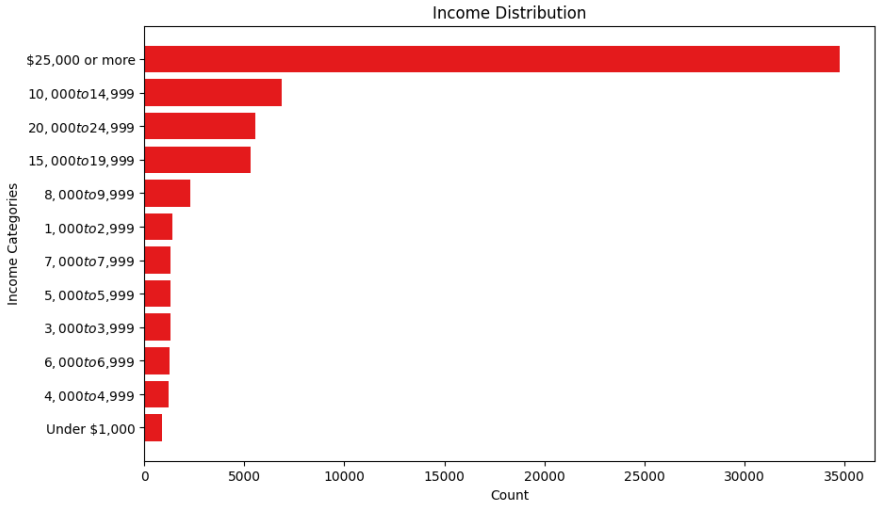
Stacked Bar Chart for Income vs. InsuranceSource



This stacked bar chart illustrates the distribution of healthcare provider sources based on different income brackets. Each income bracket, ranging from ‘under $1,000’ to ‘$25,000 or more’, is represented on the x-axis, while the count of respondents is represented on the y-axis. Within each bar, the different colors represent various sources of insurance. For respondents with incomes under $1,000, the majority rely on Medicaid as their primary healthcare provider, followed by a smaller number who have other plans as their insurance source. As income brackets increase, there is an increase in reliance on Medicare, employer-based insurance, individual plans, and plans of employer of spouse/partner. Notably, in the ‘$25,000 or more’ bracket, the dominant insurance source is provided by the respondents’ employer. In summary, this graph highlights the diversity of healthcare insurance sources across the different income levels, emphasizing the relationship between personal income and the choice (or availability) of healthcare provider sources.

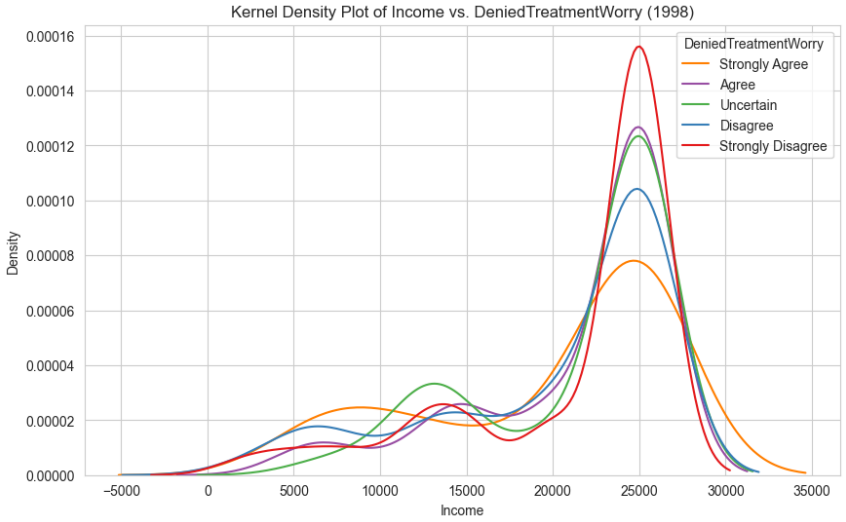
Income Distribution





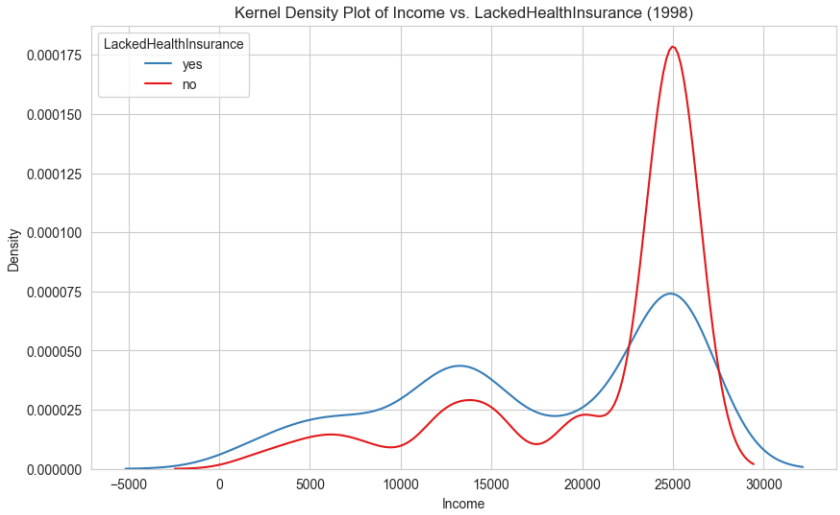
The kernel density plot illustrates the distribution of incomes. On the x-axis, numeric values represent the income levels, ranging from 0 to 25,000, and the y-axis denotes the density, which is a measure of how many respondents fall within a particular income range. The plot depicts several peaks, suggesting multiple modes of commonly reported income levels. However, most notably, there is a prominent peak around the 25,000 income range; this suggests that a significant portion of respondents fall within this income bracket. Smaller peaks at lower income levels indicate that there are also groups of respondents in those income categories, but they are not as prominent as the higher income. The bar chart depicts the same relationship and supports the fact that the most populous category, where a large number of respondents fall, is the '$25,000 or more' bracket. In summary, the kernel density plot gives a continuous view of the income distribution, highlighting peaks at certain income levels, while the bar graph depicts a categorical breakdown, indicating the count of respondents in each income range.

Kernel Density Plot of Income vs. DeniedTreatmentWorry (1998)



This kernel density plot showcases the distribution of incomes against the respondents' level of concern regarding being denied treatment in 1998. The x-axis represents varying income ranges, while the y-axis represents the density of respondents. The different colored curves represent the different sentiments: 'strongly agree', 'agree', 'uncertain', 'disagree', and 'strongly disagree.' The curves peak around the 20,000 to 25,000 income range, suggesting that a significant number of respondents from this income bracket voiced their concerns (or lack thereof) of being denied treatment. Each of the five different sentiments peaked between the $20,000 to $25,000 income range, with the 'strongly disagree' curve exhibiting the highest density, and the 'strongly agree' curve exhibiting the lowest density. The remaining sentiment curves show diverse patterns, with the 'agree' curve closely following the trend of the 'uncertain' curve. In summary, the majority of respondents, of which fall between the $20,000 to $25,000 range as depicted by the bar graph and kernel density plot for income, strongly disagree that they are denied treatment from the year 1998, while the minority strongly agree that they have been denied treatment.

Kernel Density Plot for Income vs. LackedHealthInsurance



This kernel density plot illustrates the relationship between income ranges and the status of respondents' health insurance in 1998. The x-axis represents varying income ranges, while the y-axis represents the density of respondents. Two distinct curves are depicted: one for individuals who lacked health insurance (blue), and one for individuals who have health insurance (red). The red curve peaks notably within the income range of $20,000 to $25,000; this illustrates that a significant proportion of individuals within this income bracket had health insurance. On the other hand, the blue curve peaks at a lower density, but stretches across a wider span of income levels. This visual representation suggests that, in 1998, a large portion of people in the middle-income range, particularly around the $20,000 to $25,000 range, had health insurance, whereas the lack of insurance was more evenly distributed across the income levels.

Conclusion

With the research conducted and visualizations created, conclusions can be drawn to answer the research question at hand. Each of the graphs depicted a relationship between either social class, gender, or income related variables. When respondents were asked to describe how their health is generally, a majority of the population stated that they were in good health, signifying a positive trend in self-perceived health among the participants. It was also found that there was a significant portion of respondents with insurance limitations that were more or less satisfied with their financial standing. Even though these respondents had limitations with their insurance, they still expressed moderate satisfaction with their financial situation.

When looking at ease of healthcare access for men versus. women, the overwhelming majority among all work statuses leans towards the belief that men and women have about the same amount of access to healthcare. This is interesting to note as it seems that the respondents believe there isn't any significant disparities between access to healthcare for men versus women; thus, answering our research question as most respondents believe that gender doesn't

have a big impact on access to healthcare. However, there was an overwhelming belief among all work statuses that the rich have it much easier to have access to healthcare. This answers our question as respondents believe that socioeconomic status does in fact seem to be a factor when it comes to having access to quality healthcare. When looking at the access for healthcare for older individuals versus younger individuals, mixed beliefs, of whether healthcare access was easier or harder for the elderly, was found across different work statuses. A large portion of participants from diverse work backgrounds perceive equal healthcare access for both the elderly and young; however, there is a substantial number of respondents that believe healthcare access is either 'somewhat easier' or 'somewhat harder' for the older population.This sheds light on the relationship between age and social class and how those can affect access to healthcare.

Additionally, diversity of healthcare insurance sources across different income levels was also observed, emphasizing the relationship between personal income and the choice (or lack thereof) of healthcare provider sources. For respondents with incomes under $1,000, the majority relied on Medicaid as their primary provider; however, as income brackets increase, there is an increase in reliance on Medicare, employer-based insurance, and individual plans. Notably, in the $25,000 or more bracket, the dominant insurance source is provided by the respondents' employer. This helps answer how income affects access to healthcare, as respondents with a larger income are able to be on insurance plans provided by their employers, while those with a smaller income rely on government funded healthcare.

When analyzing the distribution of incomes with respondents' level of concern regarding being denied treatment, the majority of respondents strongly disagreed that they have been denied treatment at income level $25,000. Conversely, the minority of respondents strongly agree that they have been denied treatment at the same income level $25,000. This sheds light on our research question as it answers how access to healthcare, such as healthcare treatment, may vary even at the same income level. In addition, a large portion of respondents in the middle-income range, particularly around the $20,000 to $25,000 range, had health insurance, whereas the lack of insurance was more evenly distributed across the income levels. This also helps answer our question as it seems that having a higher level of income leads to a majority of the population having health insurance.

This research project set out to examine the intricate relationships between social class, gender, and their impact on access to healthcare, all through the lens of data obtained from the General Social Survey. Through the process of data wrangling, we gained invaluable insights into the importance of data preparation and cleaning to ensure the reliability and accuracy of our analysis. Throughout the stages of data wrangling, exploratory data analysis, and data visualization, we have made an effort to reveal the disparities and subtleties within this crucial subject and our findings and observations revealed several compelling insights.

One of the strengths from our project lies in the thorough steps taken for data wrangling and analysis processes. Through meticulous data cleaning, transformation, and visualizations, we navigated the complexities of the GSS dataset and addressed issues such as missing values, variations and inconsistencies of survey questions, and vast amounts of years to parse through.

These efforts ensured that our data was well prepared and made our findings and observations more reliable and valid.

Some criticisms that may be pointed out are limitations such as the potential for response bias in the survey data, or the limitation of available variables, and the inconsistencies of the years the data was collected. Surveys with closed-ended questions can result in less validity than open response and having lots of ‘no response’ can create a bias as well. Since these limitations and criticisms are inherent in survey-based studies, it was important for us to pick certain questions/variables and cross tabulate them with others in order to draw conclusions and explore why there were certain trends and answers present in the data.

In the future, this project could set the stage for further exploration of the interplay of social class, gender, and income and how those affect health care, but there are other factors that could be explored such as regional location, and cultural influences. This project, with its data-driven insights, sets the stage for further exploration and development of policies based on evidence to promote equitable healthcare access.

Works Cited

General Social Survey. 1972-2022. NORC at University of Chicago. https://gss.norc.org/

“Variable Filter.” *GSS Data Explorer*, NORC at University of Chicago, https://gssdataexplorer.norc.org/variables/vfilter