Summary

In this study, we aimed to investigate the most influential socioeconomic factors on family income. To do so, we employed various data preparation techniques in Python, including data filtering, response variable cleaning, and exploration of missing data in the General Social Survey dataset. We also performed data cleaning and transformation, renaming columns for clarity, and filtered the data to include relevant years for our analysis. Using graphical methods, we conducted exploratory data analysis to understand the relationships between explanatory variables and family income. Our results revealed three notable relationships: 'HouseholdPopulation' showed a weak positive relationship with family income, 'HoursWorkedLastWeek' exhibited a slightly stronger positive relationship, and 'NumberOfEarners' displayed a substantial positive correlation with family income. Subsequently, we explored qualitative variables with the help of violin plots, providing a comprehensive view of how race, education, health, and newspaper readership influence income disparities. Our analysis unveiled that 'Race' depicted significant income disparities, 'Health' suggested a positive association between better health and higher family income, and 'News' exhibited an intriguing pattern with individuals who read the news daily having a notably higher median income. Lastly, 'Education' or 'Degree' displayed a hierarchy of median income levels, emphasizing the impact of educational attainment on income. While 'Number of Earners' showed a strong correlation, 'Race' and 'Education' emerged as contenders, each offering unique insights into the complex dynamics of family income disparities.

Data Source and Key Variables

The data for our analysis is sourced from the General Social Survey (GSS), a longitudinal survey conducted since 1972 that provides invaluable insights into Americans' social and economic views. Accessible at the [GSS Website](https://gss.norc.org/), this dataset offers a unique opportunity to track changes in individuals' perspectives over time. Notably, despite the relatively modest sample size of about 3,000 individuals surveyed each year, the cumulative dataset size is substantial, approximately 40 MB zipped and 520MB unzipped, owing to its extensive historical coverage. Moreover, the GSS encompasses both constant and time-varying questions, necessitating thoughtful variable selection for analytical purposes.

In our study, we focus on key variables to investigate the socioeconomic factors affecting family income. These variables include race, household size, education, labor force status, health, newspaper readership, and the number of earners in a household. 'Race' categorizes respondents into White, Black, and Other, enabling us to explore income disparities across different racial and ethnic groups. 'Household Size' records the number of individuals within a household, and we aim to understand its correlation with family income. 'Education' categorizes respondents based on their highest level of education, providing insights into the relationship between education and income. 'Labor Force Status' variables capture work hours and the employment status of respondents and their spouses, allowing us to examine how work patterns affect family income. 'Health' assesses the self-reported health status of respondents, and we intend to investigate its correlation with earning potential and family income. 'Newspaper Readership' indicates the frequency of reading newspapers and allows us to explore potential connections with education and income. 'Household Earners' counts the number of individuals

in the family who earned money from employment in the previous year, and we seek to understand its relationship with overall household income.

Response Variable

Our response variable, 'Income,' represents family income in constant dollars and spans the years 2014 to 2022, with data for the year 2020 being notably missing due to the COVID-19 pandemic's impact.

Data Preparation Methods

To ensure that the data is fit for analysis, we implemented various data preparation techniques in Python. This process began by filtering the dataset to include only the response and explanatory variables. Subsequently, we undertook the essential task of cleaning the response variable by excluding missing values, which initially numbered 7,477 cases. Missing data in longitudinal survey research can arise due to various factors such as attrition, non-response, or changes in respondents' status. Further steps involved cleaning the data by removing non-numeric characters from the quantitative response variable. To enhance the interpretability of the family income distribution, we transformed it by taking the logarithm. Each explanatory variable was also subjected to cleaning procedures, including the exclusion of cases with missing values where necessary. We filtered the data to include only years from 2000 onwards and ensured column names were clear and consistent. Exploratory data analysis using graphical methods was employed to understand the relationships between explanatory variables. Lastly, we calculated the correlation between each explanatory variable and family income to identify which variable exhibited the strongest correlation.

Challenges of Missing Data and Their Impact on Results

Missing data is a common issue in survey-based research, and it can introduce challenges and potential biases into the analysis. In our study using the General Social Survey (GSS) data, we encountered missing data in several variables, particularly in the response variable 'Income' and some of the explanatory variables. These missing values, which initially amounted to 7,477 cases, are the result of various factors inherent to longitudinal survey studies and can significantly affect the robustness of our analysis.

One significant challenge posed by missing data is the potential for selection bias. If the missing data is not missing completely at random (MCAR), it can introduce bias into our findings. For instance, in a longitudinal survey like the GSS, missing data might arise due to attrition, where respondents drop out of the study over time. If certain groups are more likely to drop out, it can skew our results, as we may be left with a sample that does not accurately represent the population.

Moreover, the missing data might not be missing at random (MAR), meaning that the probability of missing values depends on observed data. For example, respondents with lower income might be less likely to disclose their earnings. In such cases, the missing data can lead to underestimating the true extent of income disparities.

Addressing missing data inappropriately or insufficiently can also result in information loss and reduced statistical power. In our study, we applied data cleaning techniques to exclude cases with missing values. However, this approach may result in sample reduction and could potentially bias our results if the missing data is related to the variables of interest.

To mitigate these challenges and potential biases arising from missing data, we took several steps. We carefully considered the nature of missing data and ensured that our data preparation methods were transparent and well-documented. We also applied statistical techniques, such as imputation, to estimate missing values when appropriate. Imputation methods like mean imputation or regression imputation can help to preserve the sample size and reduce potential bias.

It's essential for researchers to acknowledge the presence of missing data, describe the reasons for missingness, and assess its potential impact on the results. Transparent reporting of how missing data was handled is crucial for the interpretation and reliability of the findings. In our study, the missing data challenge was addressed rigorously to provide valid and informative results despite the complexities introduced by missing values.

Results

EDA & Visualizations for Univariate Analysis

Response Variable Histogram:

The distribution of family income, as represented by the 'coninc' variable, exhibits notable characteristics. The initial histogram of family income shows a skewed and non-normal distribution, which suggests income disparities within the dataset. However, the histogram of 'LogFamilyIncome,' where the income values have been transformed using the logarithm, presents a more balanced distribution. This transformation is particularly useful for handling skewed data. It is essential to recognize that this distribution has left-skewness with some low outliers, which may carry valuable information. The statistics reveal a mean income of approximately $57,143 and a standard deviation of $44,163. The median, at $45,360, is a better measure of central tendency due to the skewed distribution. The presence of high outliers becomes evident when examining the boxplot and kernel density plot. These outliers likely stem from significant income disparities, resulting in a smaller upper class with incomes several standard deviations above the mean.

In our univariate exploratory data analysis (EDA), we created histograms of the count of each category within each explanatory variable. The 'race' variable reflects the self-identified race of respondents. Its histogram showed out of 5,824 responses, the top identifier was white with 4,302, followed by black with around 1000, and the category other with 522 responses. The 'Degree' variable, which indicates the highest level of education attained by respondents, revealed several noteworthy findings. Among the 5,824 responses, five unique categories emerged. The most frequently occurring category was 'High School' education, with 2,887 respondents. 'Bachelors' degree was the second most common, followed by 'Graduate,' 'Associates/Junior College,' and 'Less than High School.' These results provide a glimpse into the educational distribution of the dataset and set the stage for further analysis on the

relationship between education levels and family income. Exploring the 'Household Population' variable, which records the number of individuals within each respondent's household, we found several essential statistics based on 5,824 responses. The mean household size was approximately 2.05, with a standard deviation of about 1.62. The minimum household size was 0, signifying single-person households, while the maximum reached 11 individuals. The distribution is skewed to the right, implying that smaller households are more typical in the dataset, a trend that may have implications for family income patterns. In our analysis of the 'Health' variable, representing self-reported health status, we identified four unique categories among 5,824 responses. 'Good' health was the most frequently reported status, with 3,034 responses, followed by 'Excellent,' 'Fair,' and 'Poor.' This distribution might be influenced by the likelihood that individuals with poor health are less inclined to participate in a longitudinal study, especially if they are of an age less likely to have annual income outside of social security funds. Examining the 'Hours Worked Last Week' variable, which records the number of hours worked in the last week, we derived critical statistics from 5,824 responses. The mean hours worked was approximately 41.90, with a standard deviation of roughly 14.10. The distribution demonstrated normal characteristics, being unimodal with a prominent mode at 40 hours. The insights into work hours provide a foundation for investigating the connection between working hours and family income. The 'News' variable, reflecting the frequency of newspaper readership, revealed five unique categories. 'Never' was the most frequent response, with 1,477 respondents, followed by 'Everyday,' 'A Few Times a Week,', 'Less Than Once a Week', and 'Once a Week.' This distribution suggests that many individuals fall into the categories of either consistent or non-readers of newspapers. Lastly, our examination of the 'Number of Earners' variable, which counts the individuals in the family who earned money, brought out several key statistics based on 5,824 responses. The mean number of earners in households was approximately 1.71, with a standard deviation of about 0.88. The distribution exhibited a bimodal, non-normal pattern, with modes at 1 and 2 earners, indicating that some households have up to 8 earners. This finding sets the stage for understanding the dynamics of multiple earners in relation to family income. If interested in the appendix section, one can view the histograms of each explanatory variable.

These univariate analyses provide valuable insights into the distribution and characteristics of each explanatory variable, allowing for a more comprehensive exploration of the relationships between these variables and family income in subsequent analyses.

EDA & Visualizations for Bivariate Analysis

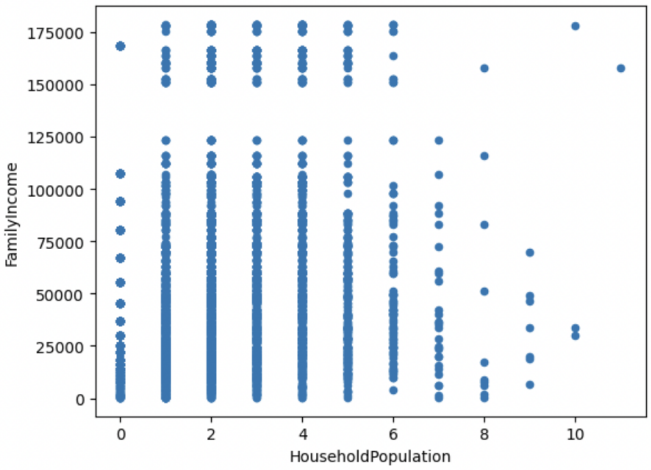
In the subsequent section of our results paper, we delve into the relationships between explanatory variables and the response variable, 'FamilyIncome.' These relationships provide critical insights into how various factors, both quantitative and qualitative, impact family income. To assess the quantitative relationships, we explore 'HousePopulation' versus 'FamilyIncome,' 'NumberHoursWorkedLastWeek' versus 'FamilyIncome,' and 'NumberOfEarners' versus 'FamilyIncome.' These analyses will help us understand how household size, work hours, and the number of earners influence family income.

In parallel, we examine qualitative relationships, particularly how categorical variables are linked to 'FamilyIncome.' The variables 'Race,' 'Degree,' 'Health,' and 'News' are scrutinized to comprehend how self-identified race, education level, health status, and newspaper readership correlate with family income. These explorations will contribute to a comprehensive understanding of the socioeconomic dynamics that drive variations in family income within the dataset. The subsequent sections will present the findings and insights gained from these quantitative and qualitative relationships, shedding light on the intricate interplay of these variables with family income.

First, the relationships between the quantitative variables ‘HousePopulation’, ‘NumberHoursWorkedLastWeek’ and ‘NumberofEarners’ versus family income are explored. These investigations were facilitated through scatter plots and the calculation of correlation coefficients.

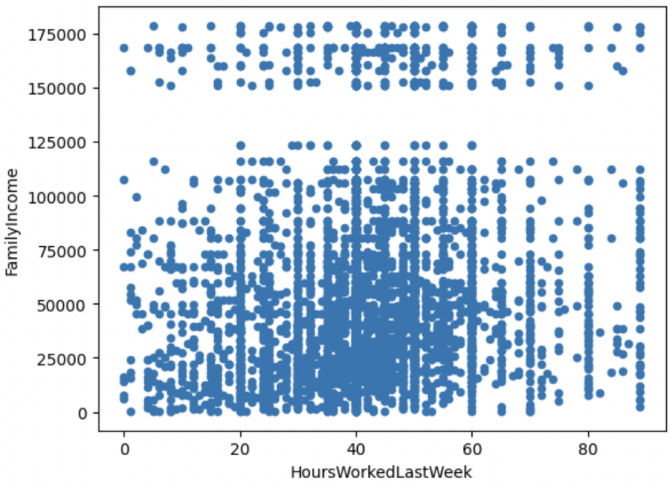
The correlation between 'HouseholdPopulation' and 'FamilyIncome' is quite low, with a coefficient of approximately 0.049. This indicates a weak positive relationship between the size of a household and family income, suggesting that, on average, larger households do not necessarily have higher incomes. The scatter plot, shown in **Figure 1**, illustrates this relationship, showing a scattering of data points without a clear linear trend.

Figure 1: The Relationship between Household Size and Family Income



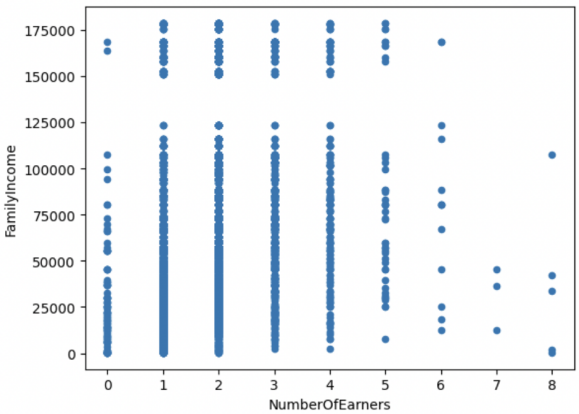
In contrast, the correlation between 'HoursWorkedLastWeek' and 'FamilyIncome' exhibits a slightly stronger positive relationship, with a coefficient of around 0.142. This implies that, in general, individuals who worked more hours in the last week tend to have higher family incomes. The scatter plot, **Figure 2**, displays a discernible upward trend, indicating that increased work hours are associated with higher family income. However, this trend is not clearly visible because the correlation coefficient is still relatively low.

Figure 2: The Relationship between Number of Hours Worked in Previous Week and Family Income



The 'NumberOfEarners' variable showcases a more substantial positive correlation with 'FamilyIncome,' with a correlation coefficient of approximately 0.307. This suggests that households with a greater number of earners tend to have higher family incomes. The scatter plot, **Figure 3**, for this relationship emphasizes this positive association, displaying more data points in the higher range of income for families with a greater number of earners. One may also hypothesize that families with an uncommonly large number of earners have relatively lower incomes, as seen in the scatter plot. This may be because they are in need of more family members to work, and may have younger teenagers working in minimum wage jobs.

Figure 3: The Relationship between Number of Earners per Household and Family Income



The 'Year' variable, though not a typical quantitative variable, is also included in the analysis to assess its relationship with 'FamilyIncome.' The correlation coefficient is close to zero, indicating a weak relationship. This demonstrates that the year in which the data was collected does not strongly affect family income.

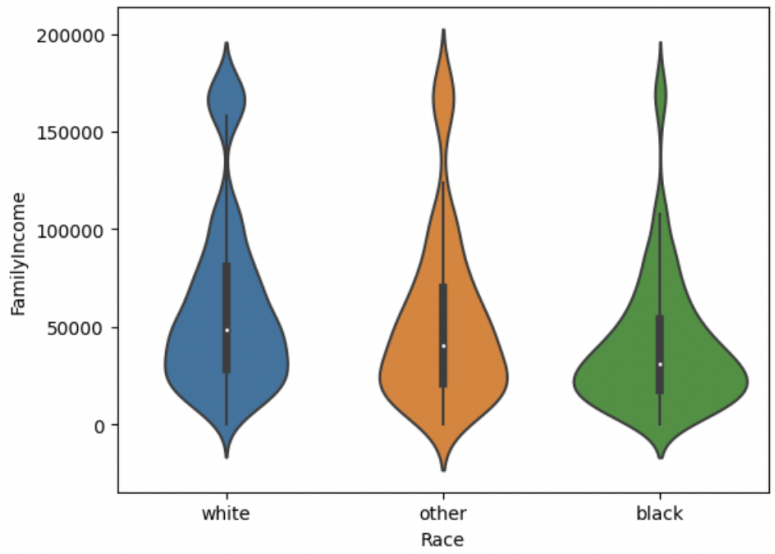
To investigate the connections between the qualitative variables and the response variable, we employed violin plots. These visualizations offer a valuable approach for examining

the impact of qualitative variables on a quantitative response variable. The advantages of using violin plots stem from their ability to provide a comprehensive view of the distribution of the response variable within each category of the qualitative variable. Unlike simple box plots, violin plots not only display summary statistics but also show the density of the data, making them particularly useful for assessing the relationship between categorical and numerical data. By visualizing the distribution of family income across different categories of qualitative variables, we can gain a deeper understanding of how factors such as race, education, health, and newspaper readership influence income disparities. This approach allows for a more holistic exploration of the impact of qualitative variables on family income.

The effect of race on family income is a result of a complex interplay of historical legacies, socio-economic factors, educational opportunities, and systemic disparities. Analyzing this relationship requires a comprehensive understanding of the numerous variables at play and a commitment to addressing the structural inequalities that perpetuate income disparities among racial groups. As one can see in the violin plot, **Figure 4**, white people have the highest median average income, followed by the category other, and lastly blacks have the lowest median family income.

However, it is important to note that within racial groups, there is considerable variation in income, and many individuals break through barriers to achieve high incomes. For example, in the plot one can see that the ranges between the different categories of race are relatively the same, meaning there are exceptions to the norm of one's particular race and their family income distribution. Additionally, the intersectionality of race with other factors like gender, age, and geographic location further complicates the relationship between race and income.

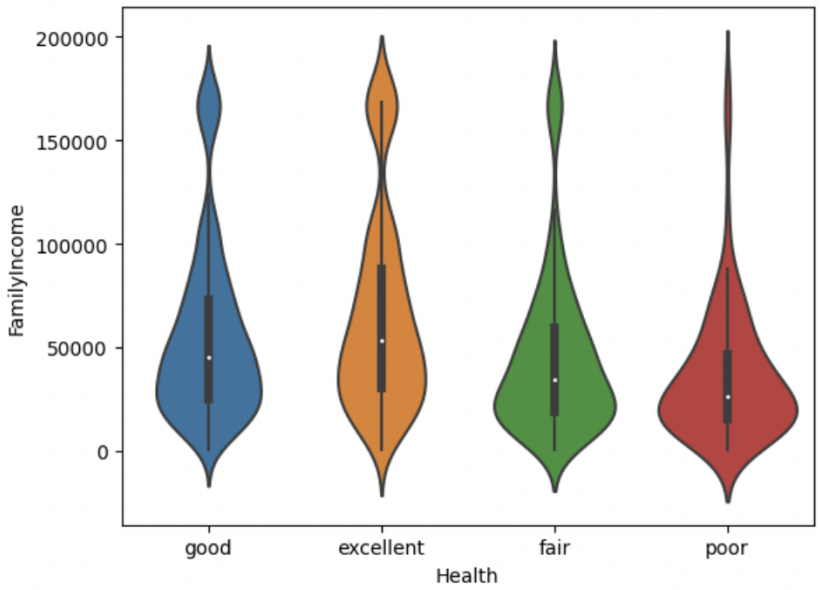
Figure 4: The Violin Plot of Race versus Family Income



The violin plot, **Figure 5,** illustrating the income differences among individuals with various health statuses suggests that better health is associated with higher family income. However, this relationship is influenced by a wide array of factors, including healthcare costs,

productivity, access to social safety nets, and emotional well-being. Emotional well-being, which encompasses mental health and overall life satisfaction, can motivate individuals to work harder and pursue higher-paying opportunities or, conversely, may lead to reduced work capacity if one's mental health is compromised. The interplay of these physical and emotional health factors underscores the complexity of the health-income relationship. While good health can contribute to greater earning potential, a comprehensive understanding of this relationship must consider the holistic well-being of individuals and the intersections with socio-economic and demographic factors.

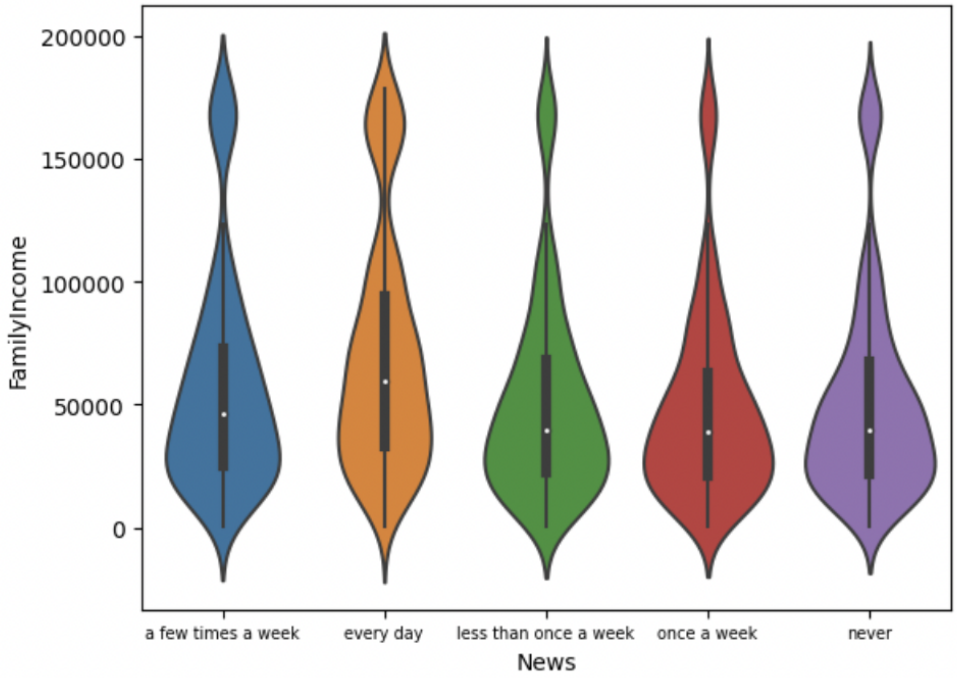
Figure 5: The Violin Plot of Health versus Family Income



The relationship between reading the news and family income, as indicated by the violin plot, **Figure 6,** reveals an intriguing pattern. While categories such as 'a few times a week,' 'less than once a week,' and 'once a week' exhibit similar income distributions with comparable ranges and medians, the category of individuals who read the news 'every day' stands out with a notably higher median income. Several hypotheses can be considered to explain this observation, including their occupations, which may require up-to-date information (e.g., finance professionals), higher levels of education and knowledge, greater financial literacy and investment activity, increased access to networking and career-enhancing opportunities, and a personal drive for success and aspiration.

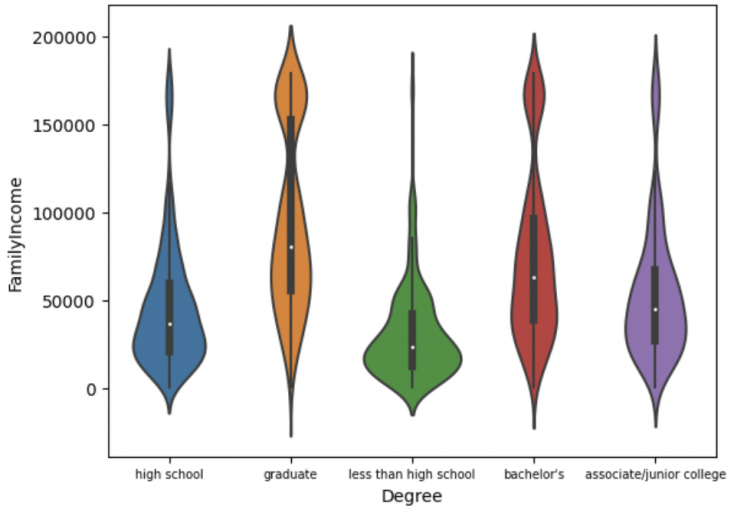
It's important to note that these are hypotheses, and causation cannot be inferred from this observation alone. The relationship between daily news consumption and higher median income may be influenced by various factors, including individual motivations, career choices, and access to resources. Further research and analysis would be needed to establish a causal relationship and identify the specific factors driving this pattern.

Figure 6: The Violin Plot of News Read versus Family Income



The relationship between the highest degree obtained or education level and family income is marked by significant variations. The plot, **Figure 7,** reveals a hierarchy of median income levels, with the highest median associated with those holding graduate degrees, followed by individuals with bachelor's degrees, associate or junior college degrees, and, lastly, those with less than a high school education. These findings underscore the critical influence of educational attainment on income. Factors contributing to this disparity include advanced skills and specialization associated with higher degrees, access to well-paying job opportunities, career advancement, job stability, and the potential for upward social and economic mobility. Higher education is often viewed as an investment in long-term career and income growth, offering opportunities for improved living standards and financial security, with graduate degrees leading the way in income potential. Individual circumstances and other variables, such as occupation and location, also play a role in determining family income.

Figure 7: The Violin Plot of Highest Degree Obtained versus Family Income



Number of Earners After an in-depth analysis of the relationships between our variables and Family Income, we conclude that 'Number of Earners' from the quantitative variables exhibits the strongest correlation with Family Income. This conclusion is supported by a high correlation coefficient, indicating a clear and direct relationship between the number of earners in a household and the overall family income. As more individuals contribute to the household's earnings, the total income is likely to increase, making 'Number of Earners' a powerful predictor of financial well-being for families.

Race Alternatively, we find compelling evidence to suggest that 'Race' from the qualitative variables could hold the strongest correlation with Family Income. This conclusion arises from the distinct differences in median incomes among racial and ethnic categories. The variations in income levels imply a deeper-rooted relationship between race and income disparities, reflecting historical and systemic influences. Extensive research outside of our analysis also underscores the persistence of racial income gaps. Therefore, while 'Number of Earners' is a strong predictor, the racial disparities in income observed here and in broader contexts make 'Race' a prominent variable to consider.

Education Lastly, 'Education' or 'Degree' from the qualitative variables emerges as a contender for the variable with the strongest correlation with Family Income. The medians among different education levels show significant disparities, highlighting the influence of educational attainment on income. The link between higher education and increased earning potential is well-documented, as individuals with advanced degrees often access better job opportunities. This is consistent with broader research and understanding of the impact of education on income, making 'Education' a robust predictor of Family Income.

While 'Number of Earners' stands out due to its quantitative correlation coefficient, 'Race' and 'Education' from the qualitative variables demonstrate strong relationships with Family Income based on their significant variations in income medians. The choice among these variables depends on the specific focus and research objectives, as each provides valuable insights into the complexities of income disparities.

Conclusion

In this project, we have delved into the intricate web of socioeconomic factors influencing family income. Through extensive data preparation, exploratory data analysis, and visualization techniques, we sought to uncover the variables that hold the strongest correlation with family income. The project aimed to contribute to a deeper understanding of income disparities and provide insights into the most relevant factors affecting financial well-being.

Our findings have shed light on several key relationships. Firstly, the quantitative variable 'Number of Earners' exhibited a strong positive correlation with family income. As more individuals contribute to a household's earnings, total income is likely to increase, making it a powerful predictor of financial well-being for families. Secondly, the qualitative variable 'Race' demonstrated substantial income disparities among racial and ethnic categories, highlighting the

influence of historical and systemic factors on income inequalities. Finally, 'Education' or 'Degree' also emerged as a robust predictor of family income, with significant differences in median income levels across different education levels.

Despite the project's comprehensive nature, it is not without its limitations. One potential criticism is the lack of consideration for other influential variables not included in this analysis. Factors such as occupation, location, and specific demographics may play critical roles in determining family income. Future research could incorporate these variables for a more comprehensive understanding of income disparities. Additionally, the use of a single dataset, the General Social Survey, may limit the generalizability of our findings. It would be valuable to validate the results using other datasets or longitudinal studies.

Moreover, while the project explored the relationships between variables and family income, it did not establish causation. Correlation does not imply causation, and further research would be necessary to unravel the underlying mechanisms driving these relationships. Qualitative research, such as interviews or surveys, could provide a more in-depth understanding of the experiences and challenges faced by individuals of different races or educational backgrounds in the context of income.

Furthermore, additional work could focus on understanding the intersectionality of these variables. For instance, exploring how race, education, and other demographic factors interact and influence income disparities could provide a richer and more nuanced perspective. Additionally, it would be informative to investigate the temporal aspects of these relationships. How have these correlations evolved over time, and what external factors have shaped them?

In conclusion, this project has successfully provided valuable insights into the complex web of socioeconomic factors influencing family income. While it has limitations, it lays the groundwork for future research to delve deeper into the intricacies of income disparities and the multifaceted dynamics at play. The robust correlations found in this project between 'Number of Earners,' 'Race,' and 'Education' offer valuable insights into the key determinants of family income. This work contributes to the broader discussion on income disparities and opens the door for further investigation into the factors shaping financial well-being in contemporary society.

Appendix

Figure 8: Box Plot of Family Income

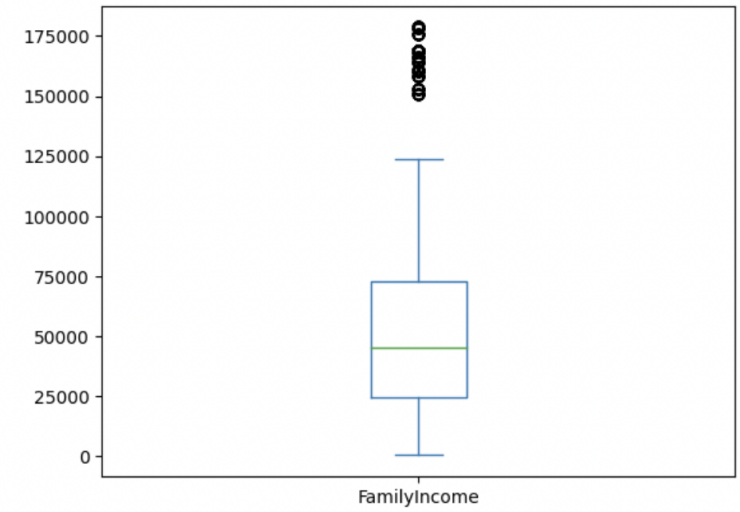
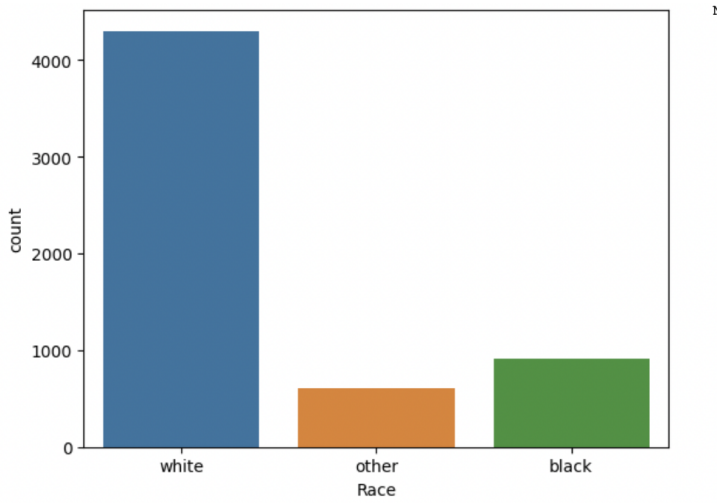
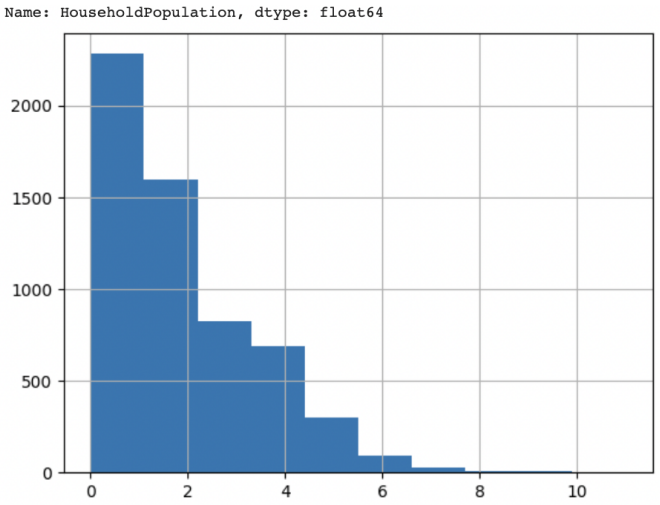
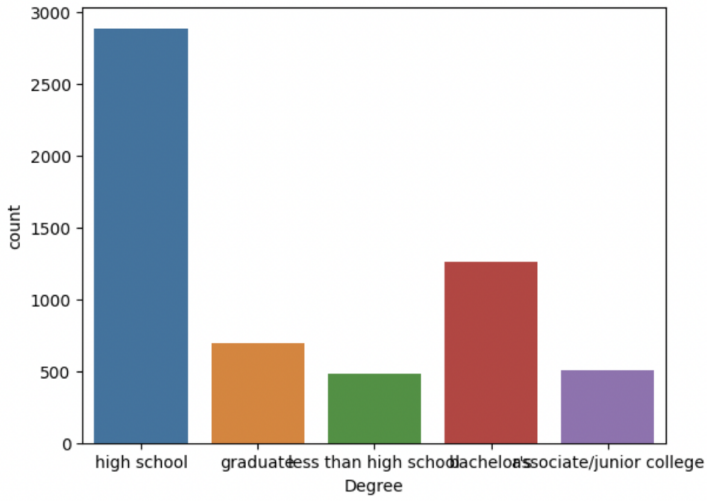
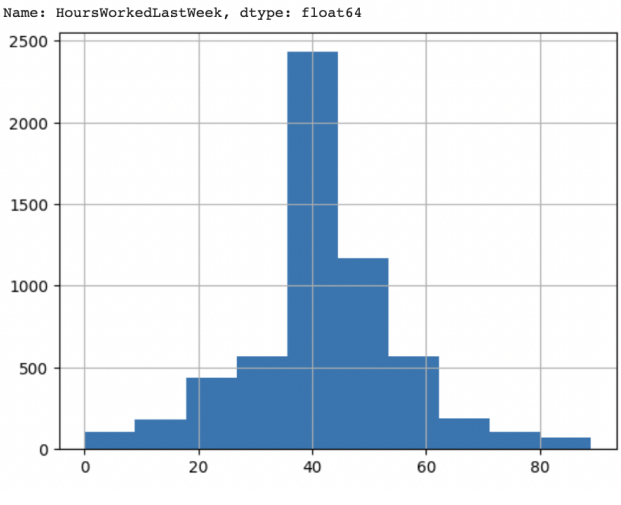


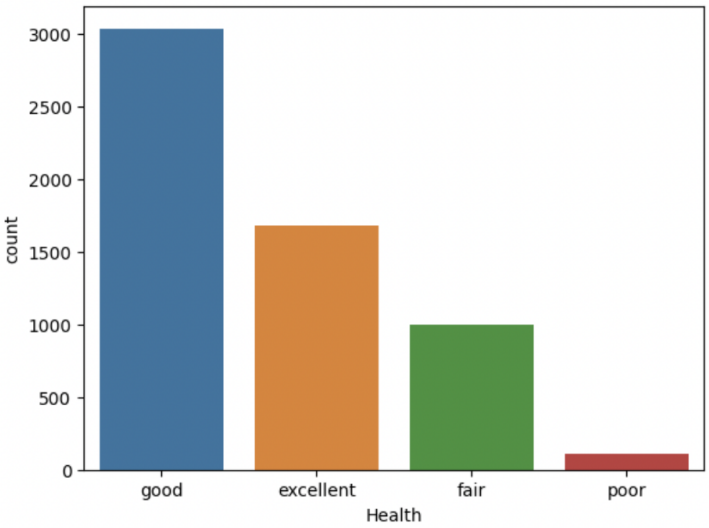
Figure 9: Univariate Histograms of Explanatory Variable Responses

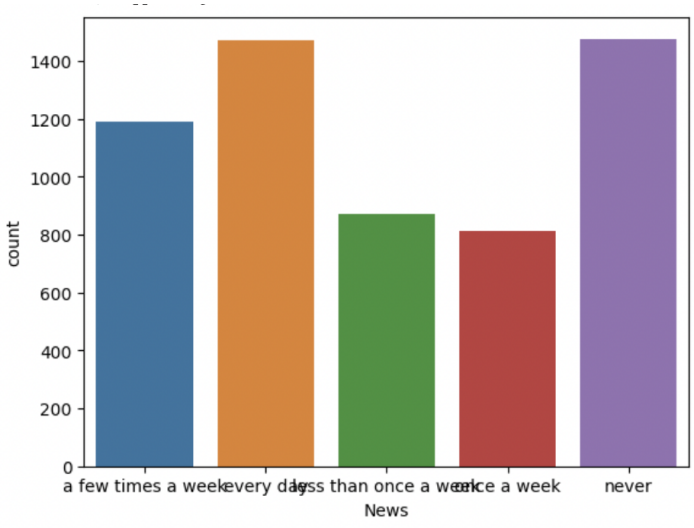













```
Name: NumberOfEarners, dtype: float64
```