Investigating the Impact of Various Variables on the Quality of Life in US States

Upperstatsmen

Pledge

Please type your names in the appropriate space below. Failing to do so will result in a 0 on this assignment.

"We have neither given nor received unauthorized aid on this assignment"

- Member 1: Leah Germain
- Member 2: Kayla Denoo
- Member 3: Sana Rehan
- Member 4: Starr Mark

Introduction

Why is our topic relevant to study?

The topic of our project is how different variables impact the overall quality of life in states in the United States. The overall quality of life in this project is gauged by assessing the total average number of physically unhealthy and mentally unhealthy days per month in each state. This topic is important because it allows interested parties to determine the factors that play the biggest role in improving the quality of life of American citizens. This data could be used to target increasing access to healthcare, developing a more healthy food system, increasing access to higher education, etc. This topic could also be useful to local, state, and national politicians, as it would show them what factors are contributing to the overall health and wellbeing of their citizens and the people voting for them. Then, they could choose to prioritize certain issues over others on their campaign trail based on which state they are campaigning in (Burdina 2014). In addition, this topic could be of interest to insurance companies. By examining the number of physically and mentally unhealthy days, along with factors like uninsured rates, diabetes prevalence, and access to healthy foods, insurers can gain insights into the health profiles of different states. By analyzing these relationships, insurance companies can refine risk models, adjust premiums, and create targeted health plans that address specific health and economic challenges in high-risk areas. During the COVID-19 pandemic, insurers innovated new technologies to collect and analyze diverse information for risk modeling (Lanfranchi & Grassi, 2022), showing how insurance companies value risk analysis of diverse information during public health crises, and would therefore be interested in our topic. Finally, this topic could be relevant to crisis decision-makers, such as government response agencies. Whether there is a public health crisis, a natural disaster, or some other immediate crisis, it is helpful to have as much information beforehand as possible. For example, after Hurricane Maria in Puerto Rico, response agencies realized that the overall death count was highly influenced by people who had chronic conditions before the crisis occurred. Having prior information about the prevalence of chronic conditions in different geographical areas could have aided in prevention and recovery (Roeder, 2021).

Motivation for variables of interest

Our motivation for choosing to look at education, specifically people who have a bachelor's degree or higher, is that education typically correlates with more job opportunities and a more stable lifestyle. If a state is more educated, then we will anticipate that the state will also have a higher quality of life. Our motivation for choosing to look at access to healthy

foods is that many states struggle with obesity and health issues relating to diet, which can be attributed to access to healthy foods. Therefore, a greater access to healthier food options may contribute to an overall higher quality of life. Our third variable of interest is the percentage of households in a state with severe housing problems, which includes overcrowding, housing costs, and kitchen and plumbing facilities. Housing problems can cause stress and physical effects, which therefore impacts a given person's quality of life.

Research Questions:

Our three research questions are:

1. Does having more than 35% of a state's population as college graduates suggest a higher overall quality of life in that state?
2. Do people who live in states with limited access to healthy foods have a lower overall quality of life?
3. Do people in states with a higher percentage of households with severe housing problems have a lower overall quality of life?

Data Description

Most of our data was compiled from the Social Explorer data finder website. We used the 2023 Health Data to compile this, and used all of the US states. We excluded Washington DC. The variables we acquired from the Social Explore website were number of physically unhealthy days per month, number of mentally unhealthy days per month, percentage of households with severe housing problems, percent of people with limited access to healthy foods, percent of low birthweight births, percent of people without insurance (population under 65 years), and percent of adults with diabetes. Our quality of life variable is the sum of the number of physically unhealthy days per month and the number of mentally unhealthy days per month for each state.

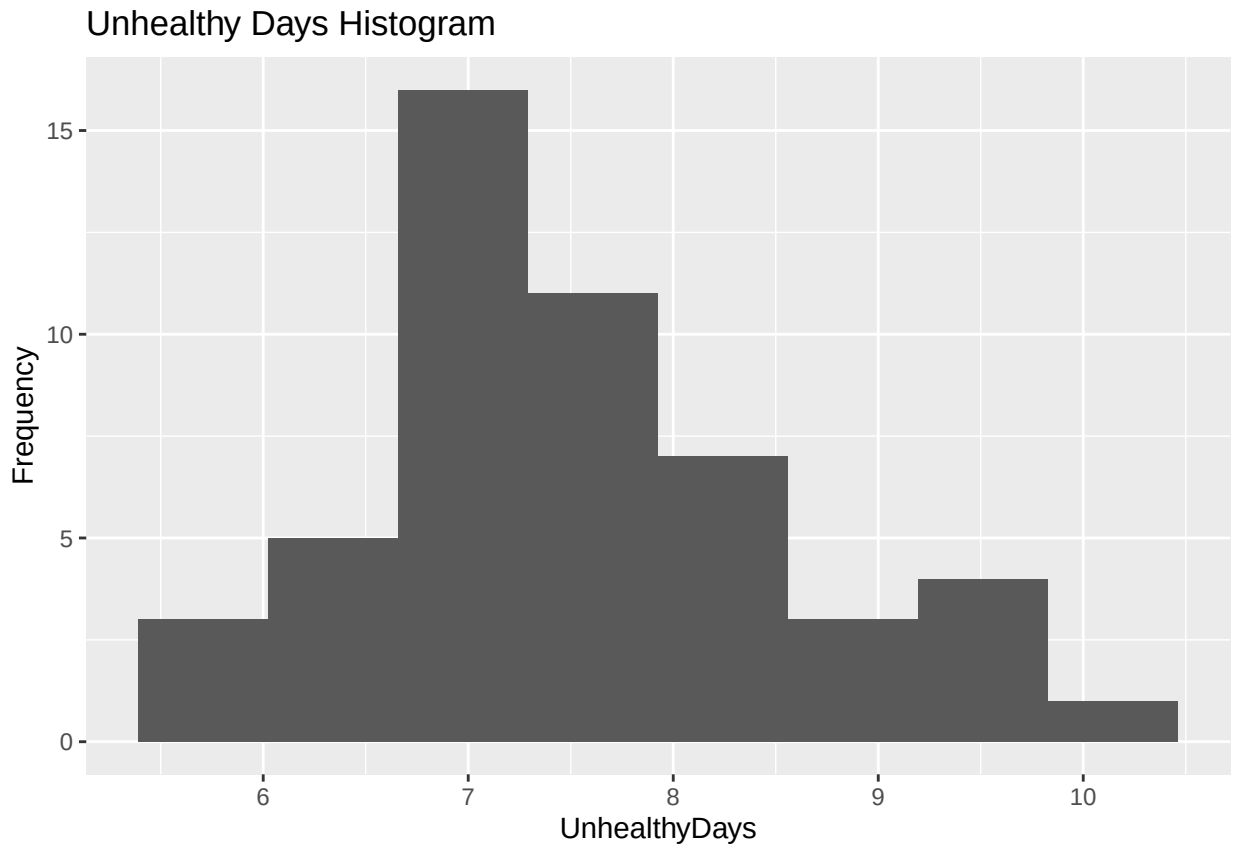
Our next variable is gathered from the US Census Bureau 2022 report. The variable is qualitative, and it is whether or not a specific state's median income is equal to or greater than the United State's median income. If it is above, the state is "yes" and if it is below, the state is "no". The Census Bureau did a survey in 2022 called the American Community Survey (ACS) which is the data used in this report. This data was collected in order to gather information about people's income across the country. From this report, we used

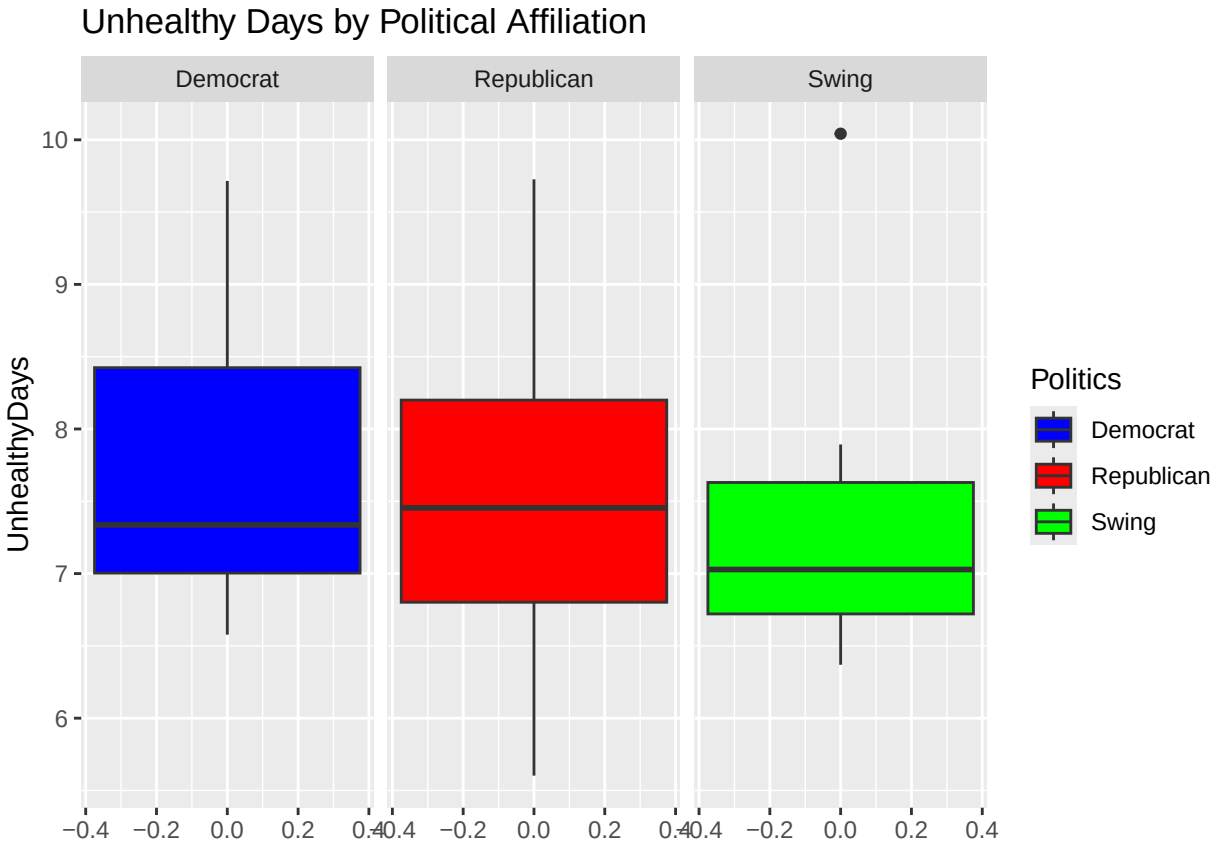
the household income data, which was collected to "include income of the householder and all other people 15 years and older in the household, whether or not they are related to the householder". Based on this data, the median household income for the United States was $74,755 in 2022. This survey also listed each state's median household income, so we compared those values to the $74,755 and gave each state a "yes" if it was higher, or a "no" if it was lower.

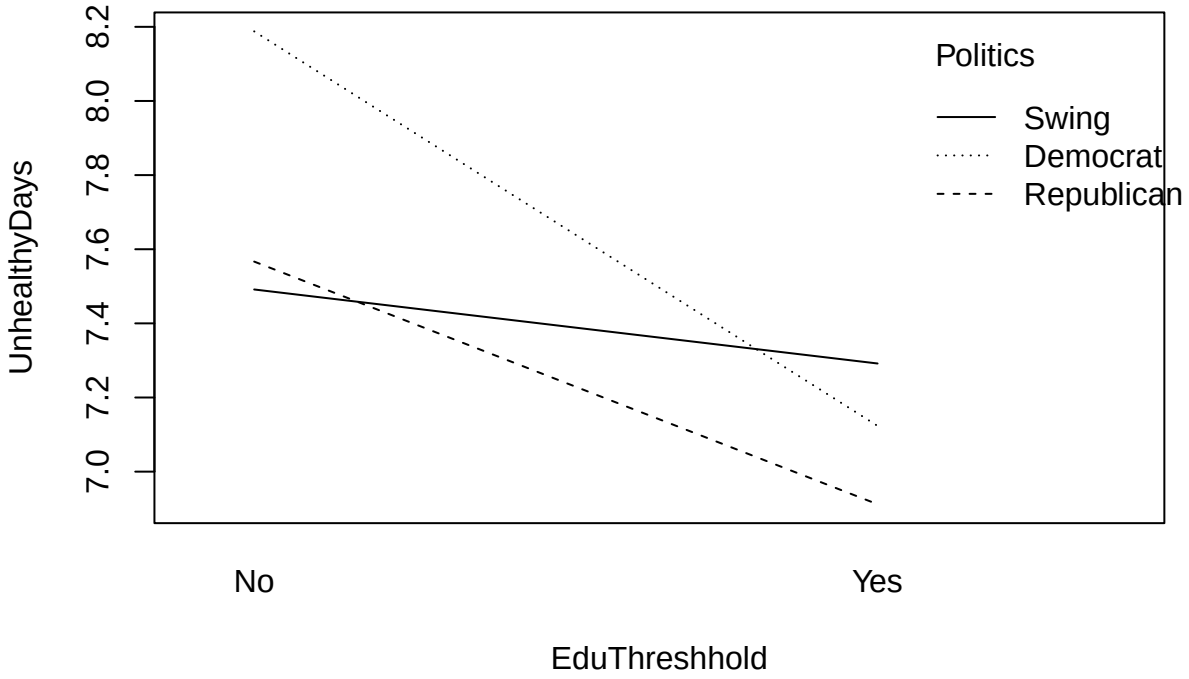
Another variable is each state's typical political affiliation in presidential races. This is a qualitative variable, and the categories are "Republican", "Democratic", and "Swing". To determine each state's political affiliation, we looked at the 2024 Presidential Election Consensus Electoral Map from 270towin, which shows the typically Republican, Democratic, and toss-up states, otherwise known as swing states in our dataset.

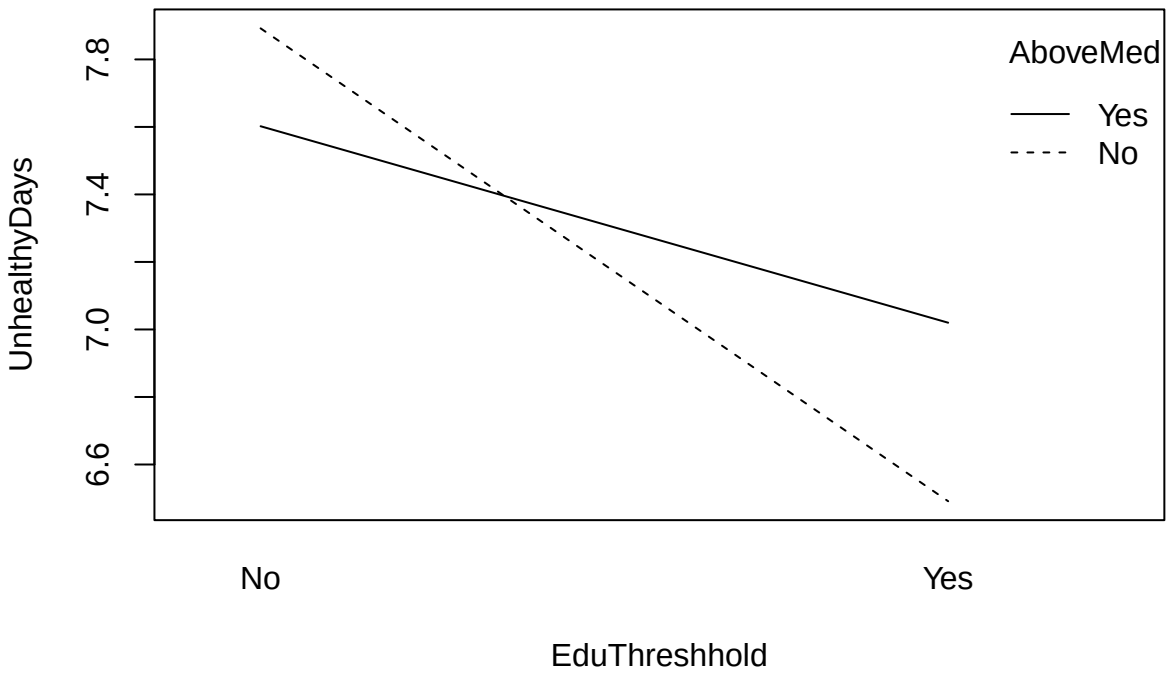
Exploratory Data Analysis:

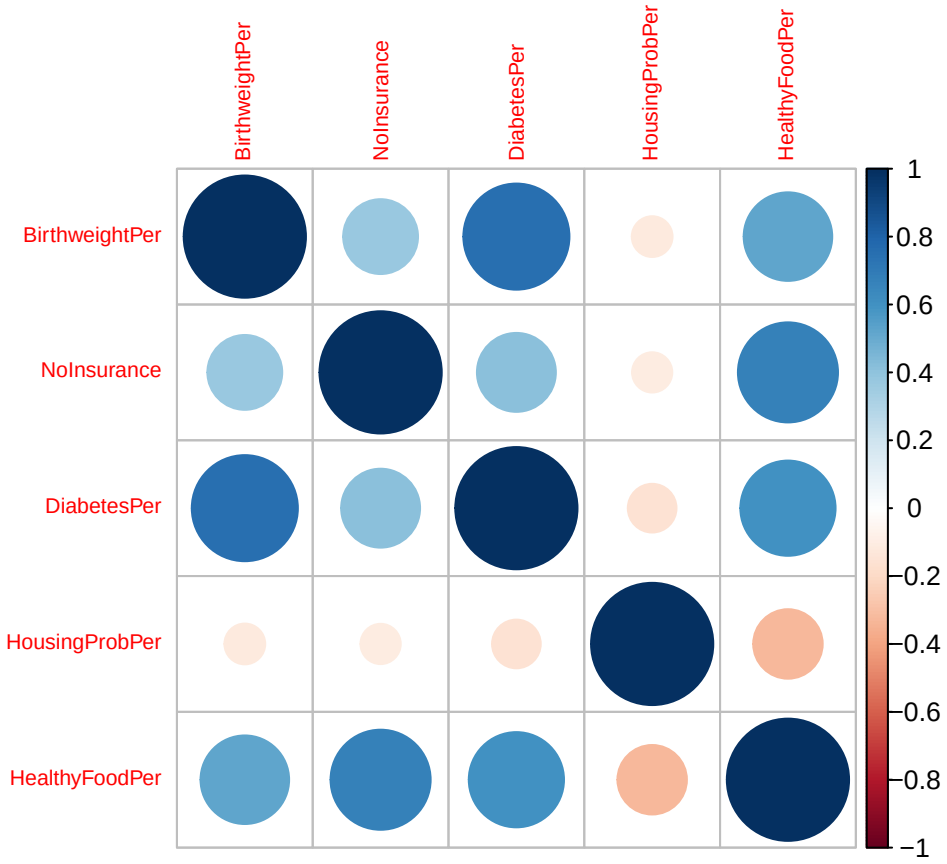
```
Warning: Removed 2 rows containing non-finite outside the scale range
(‘stat_bin()‘).
```

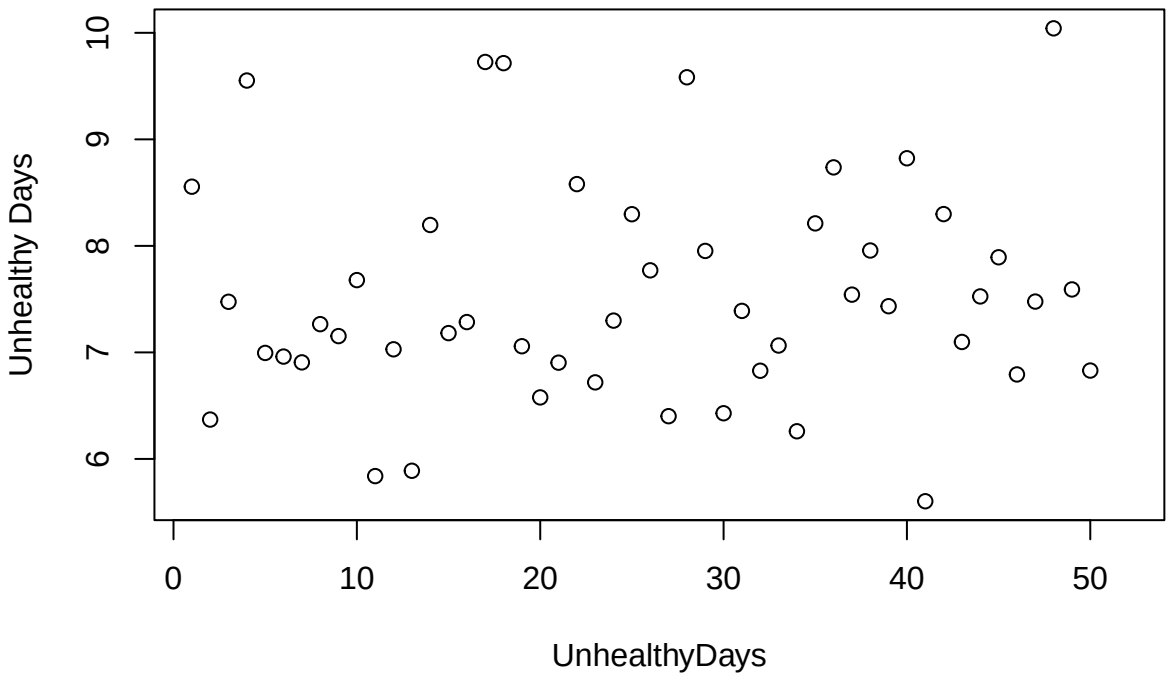


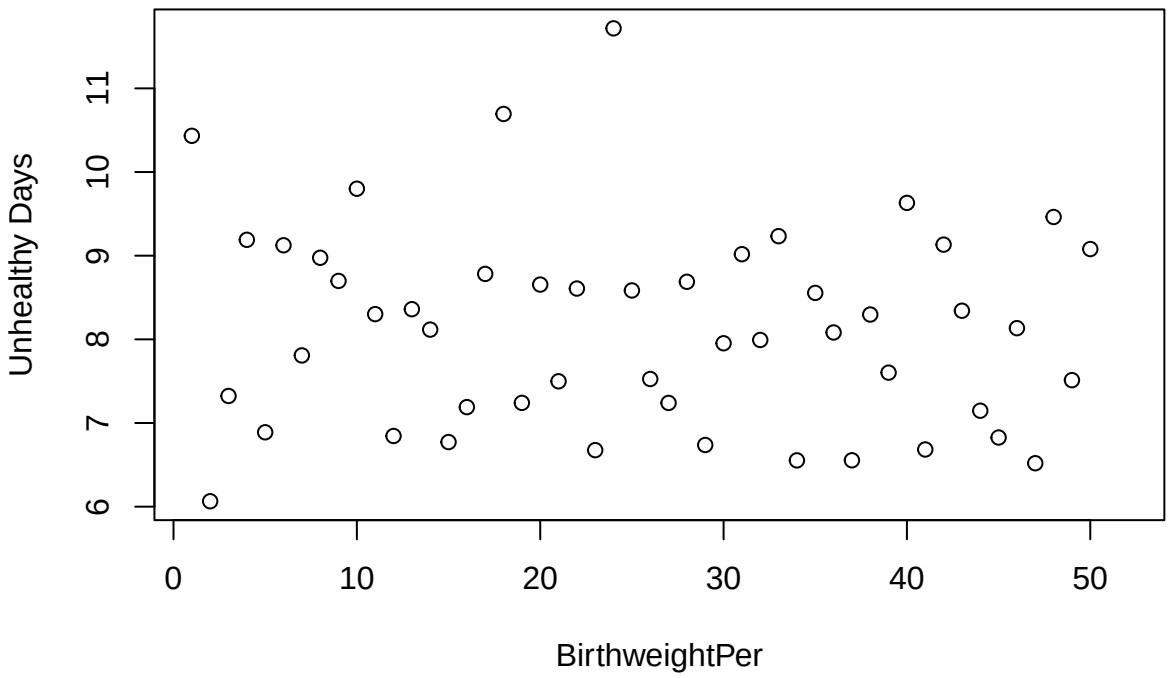

Unhealthy Days by Political Affiliation
Democrat
Republican
Swing
UnhealthyDays
10
9
8
7
6
−0.4
−0.2
0.0
0.2
0.4
Politics
Democrat
Republican
Swing

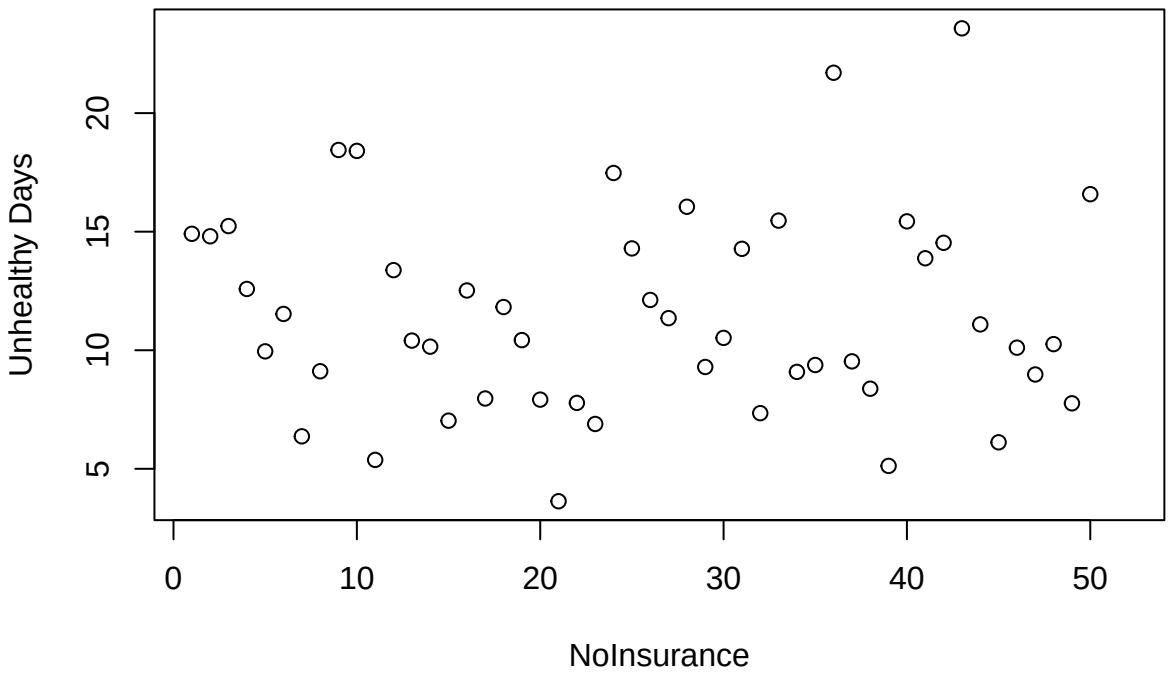

UnhealthyDays
8.2
8.0
7.8
7.6
7.4
7.2
7.0
Politics
Swing
Democrat
Republican
No
Yes
EduThreshhold

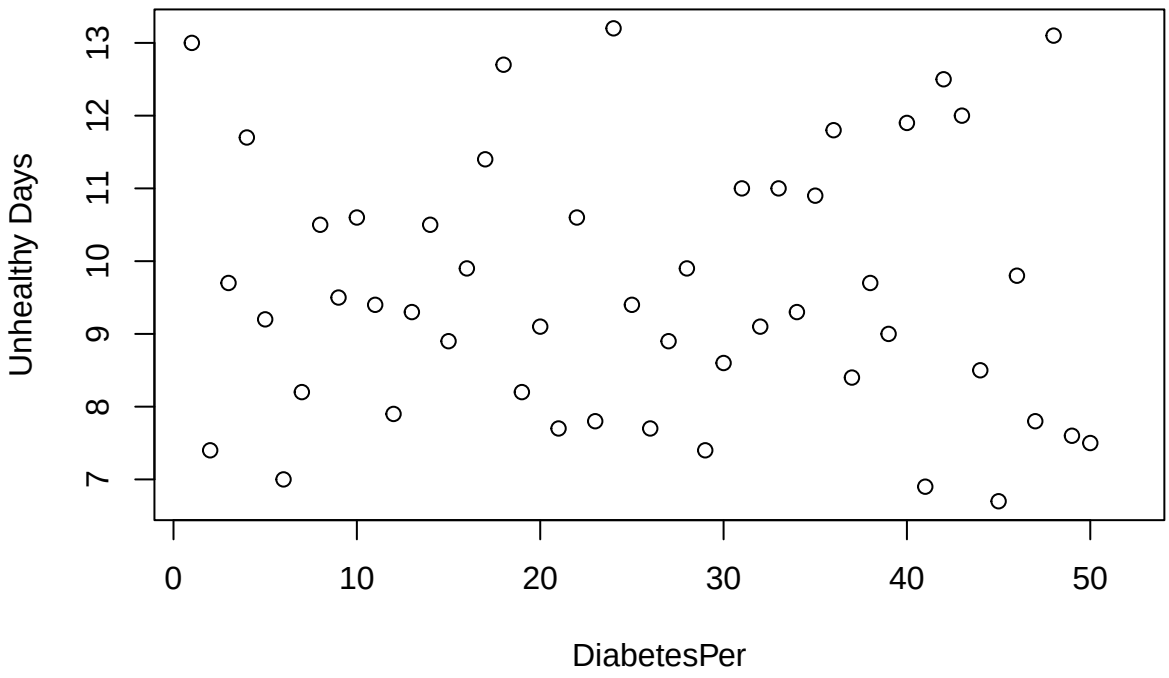

UnhealthyDays
7.8
7.4
7.0
6.6
AboveMed
Yes
No
No
Yes
EduThreshhold


BirthweightPer
NoInsurance
DiabetesPer
HousingProbPer
HealthyFoodPer
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1


Unhealthy Days
6
7
8
9
10
0
10
20
30
40
50
UnhealthyDays


Unhealthy Days
6
7
8
9
10
11
0
10
20
30
40
50
BirthweightPer


Unhealthy Days
5
10
15
20
0
10
20
30
40
50
NoInsurance



Conclusions

Quality of Life Unhealthy Days Histogram

Numerical Summary:

Mean	Median	SD	Max	Min
7.534726	7.343983	1.037236	10.042045	5.602752

The histogram displays a unimodal and right-skewed distribution of unhealthy days, indicating that there are fewer data points at the lower end of unhealthy days. This right skew suggests that most states report a higher number of unhealthy days, with the data centered around a mean of 7.53 and a median of 7.34. The mean being greater than the median reinforces the right-skewness of the distribution, as it indicates that some states have notably high unhealthy day counts that are pulling the mean upward. The maximum value, noted at approximately 10 (10.04 in the numerical summary), reflects the highest reported number

of unhealthy days, while the minimum value of around 5.5 (5.60 in the numerical summary) indicates the lowest count of unhealthy days. The absence of outliers in this dataset suggests that the values fall within a consistent range, without any extreme counts that could skew the results. Overall, this distribution highlights that most states experience a low to moderate number of unhealthy days, with a few states reporting particularly high counts, which may indicate a lower overall quality of life in those areas. Going forward, it may be useful to log transform this variable to control for skewness.

Quality of Life Boxplots Faceted by Political Affiliation

Numerical summary:

Politics	median	sd	max	min	IQR
Democrat	7.34	1.03	9.72	6.58	1.42
Republican	7.45	1.00	9.73	5.60	1.40
Swing	7.03	1.19	10.0	6.37	0.909

Looking at the boxplot and numerical summary, the medians show that Republican states experience slightly more unhealthy days, with a median of 7.45 compared to 7.34 for Democrat and 7.03 for Swing states, suggesting a marginally lower quality of life in Republican states overall. The boxplot also highlights greater variability in unhealthy days for Democrat and Republican states, as reflected in their interquartile ranges (IQRs) of 1.42 and 1.40, respectively. In contrast, Swing states show a narrower IQR of 0.909, indicating more consistency around the lower median. This pattern suggests that unhealthy days in Democrat and Republican states are more widely spread, while Swing states tend to cluster more tightly around their median. The whiskers on the boxplot further illustrate these differences, with the minimum values showing a slightly lower number of unhealthy days in some Republican states (5.60) compared to Democrat (6.58) and Swing (6.37) states. The maximum values extend close to 9.72 in Democrat states and 9.73 in Republican states; however, Swing states have an outlier at 10.0 unhealthy days, represented by West Virginia. This outlier indicates a particularly low quality of life within that state compared to other Swing states, underscoring unique factors likely affecting its population's health. Going forward, it may be useful to include political affiliation in any predictive modeling for unhealthy days.

Explanatory Variable Correlation Dot Plot

The maximum absolute correlation amongst the pairwise sets was between Percent_Diabetics and Percent_Low_Birthweight_Births (r=0.7532). This strong correlation between these two variables highlights the relationship between Gestational Diabetes and Premature Birth & Low Birthweights. This may be a sign of multicollinearity, however, it is the only relationship that has an r value of above 0.70. This shows that the majority of variables have either moderate or weak correlations with one another, and is not to be worried about in terms of multicollinearity in our model. The lowest absolute correlation was between Percentage_Households_with_Severe_Housing_Problems and Percent_Persons_Without_Insurance (r=0.1084), meaning that the lack of insurance has little to do with whether one will have severe housing problems. The absolute mean and median of this data are 0.4049 & 0.3959 respectively. These values show that, on average, the relationships between variables are moderate, which is a good sign in terms of multicollinearity. Additionally, the range of this data was 0.6448, showing significant variability amongst the correlations of the explanatory variables. Going forward, it would be a good idea to monitor any low correlations & multicollinearity and look at the VIFs, R-squared & RMSE values.

Quantitative Explanatory Variable x Quality of Life Scatterplots

The plot showing the percent of low birth weight births versus the total number of unhealthy days per month in every US state shows a positive correlation. As the percent of low birth weight births increases, so does the amount of unhealthy days. The plot showing the percent of people under age 65 without insurance per state versus the total number of unhealthy days per month has a pretty equal variance across insurance percentages. The plot showing the percent of diabetics in a given state has a positive correlation with the total number of unhealthy days. As the percentage of diabetics increases, so does the amount of unhealthy days. The plot showing the relationship between the percent housing problems in a US state versus the unhealthy days gives an overall moderate negative relationship. These plots suggest that certain health-related variables may contribute to increased unhealthy days per month. This supports using the total number of unhealthy days as an appropriate response variable for regression, as it captures varying health conditions across states in the US.

Moving forward, predictors like diabetes rates and birth weight percentages appear to be significant factors, and they will be prioritized in the regression model. Transformations may be necessary if further analysis reveals skewness or nonlinearity. One of our research questions

was: do people in states with a higher percentage of households with severe housing problems have a lower overall quality of life? From the scatterplot produced, the preliminary answer to this question is no, as there is a moderate negative correlation of percentage of households with severe housing problems and number of unhealthy days per month. However, it should still be included in regression to test whether or not it is significant.

Quality of Life: Qualitative x Qualitative Interaction Plots

The first graphical summary depicts a qualitative x qualitative interaction between a state meeting the fixed educational threshold and its political affiliation. The average number of unhealthy days depends on whether or not a state attained the fixed educational threshold. Or the average quality of life score per state depends on the state's political affiliation. The plot shows

The second graphical summary visualizes a qualitative x qualitative interaction between a state meeting the education threshold versus and if the state's median household income falls above the national median household income level. The plot shows that either the average number of unhealthy days is dependent on the national median income level or on whether the state met a certain level of education.

Appendix A: Data Dictionary

Variable name	Abbreviated name	Description
Quality_of_Life_Unhealthy_Days	UnhealthyDays	(Quantitative) Total sum of physically unhealthy days per month and mentally unhealthy days per month in a given state (2023)
Education_Threshold	EduThreshhold	(Qualitative) Whether 35% or more of a given state's population has a Bachelor's Degree or Higher (2019)
Political_Affilitation	Politics	(Qualitative) Whether a given state is a Republican, Democratic or Swing state (2024)
Above_US_Median_ Household_Income	AboveMed	(Qualitative) Whether a given state's Median Household Income was above the total US Median Household Income of $80,610 (2023)
Percent_Low_Birthweight_Births	BirthweightPer	Percentage of Low Birthweight Births (<2.5kg) in a given state (2023)
Percent_Persons_Without_Insurance	NoInsurance	Percentage of a given state's population under 65 without insurance in (2013)
Percent_Diabetics	DiabetesPer	Percentage of diabetics in a given state (2023)
Percentage_Households_with _Severe_Housing_Problems	HousingProbPer	Percentage of households with severe housing problems in a given state (2023)

Variable name	Abbreviated name	Description
Percent_People_Limited_Access_To_Healthy_Foods	HealthyFoodPer	Percentage of households in a given state that are susceptible to food insecurity (2023)

Appendix B: Data Rows

	X	State	EduThreshhold	Politics	AboveMed	UnhealthyDays	BirthweightPer
1	1	Alabama	No	Republican	No	8.555693	10.432760
2	2	Alaska	No	Swing	Yes	6.369383	6.064884
3	3	Arizona	No	Republican	No	7.475239	7.323263
4	4	Arkansas	No	Democrat	No	9.552272	9.189877
5	5	California	Yes	Democrat	Yes	6.994943	6.889022
6	6	Colorado	Yes	Democrat	Yes	6.960719	9.124500

	NoInsurance	DiabetesPer	HousingProbPer	HealthyFoodPer
1	14.910001	13.0	13.15679	8.760549
2	14.806005	7.4	20.42274	7.868789
3	15.236259	9.7	17.41027	8.679974
4	12.584366	11.7	13.78932	10.015975
5	9.953589	9.2	25.86903	3.234381
6	11.530297	7.0	15.96653	5.044387

Appendix C: References

Introduction

Burdina, M. (2014). WAVE RIDING OR OWNING THE ISSUE: HOW DO CANDIDATES DETERMINE CAMPAIGN AGENDAS? The American Economist, 59(2), 139–152. http://www.jstor.org/stable/43664832

Lanfranchi, D., & Grassi, L. (2022). Examining insurance companies' use of technology for innovation. The Geneva papers on risk and insurance. Issues and practice, 47(3), 520–537. https://doi.org/10.1057/s41288-021-00258-y

Roeder, A. (2021). The power of data in a crisis. Harvard T.H. Chan School of Public Health. https://www.hsph.harvard.edu/news/features/the-power-of-data-in-a-crisis/

Data

270toWin. (2024). Consensus 2024 presidential election forecast. 270toWin. https://www.270towin.com/maps/consensus-2024-presidential-election-forecast.

Federal Reserve Bank of St. Louis. (2019). Educational Attainment, Annual: Bachelor's Degree or Higher by State. [Data set]. Retrieved October 27, 2024, from https://fred.stlouisfed.org/release/tables?rid=330&eid=391444&od=2019-01-01#.

U.S. Census Bureau. (2023). Health data. [Data set]. Prepared by Social Explorer. Retrieved October 27, 2024, from https://www.socialexplorer.com/data/HD2023/metadata/?ds=SE.

U.S. Census Bureau. (2023). Median household income and Gini index in the past 12 months by state and Puerto Rico: 2021 and 2022 [Data set]. U.S. Census Bureau. Retrieved October 27, 2024, from https://www.census.gov/content/dam/Census/library/publications/2023/acs/acsbr-017.pdf.