Research Project Questions

DATA5207: Data Analysis in the social sciences

NAME

March 03, 2025

Table of Contents

# Introduction

The monitoring of health statistics forms creates a strong foundation for the tracking of the general population health (Bchir et al., 2006) and guides policy formulation around matters health (Krieger et al., 1997). Studies have been undertaken to map out the relation between health metrics and socio-economic aspects, including (Braveman et al., 2010) checking on education, income levels and race and found that the least educated and lower income groupings were often associated with lower health status.

The main objective is pointing out the predictors for better health outcomes at the county level in the United States. The relationship existing between various variables is assessed and how they generally influence the number of deaths among United States of America county residents under age 75 per 100,000 population (age-adjusted). Regression analysis with the dependent variable, premature age-adjusted mortality variable, an important health indicator (Mansfield et al., 1999) and predictor variables are selected based on previous research (Cheng & Kindig, 2012) is undertaken.

The predictors based on strong association to premature age-adjusted mortality variable by previous studies include variables with information on income and income inequality (Adler & Newman, 2002), population, demography aspects (Cooper et al., 2001) including race, health care costs and associated level of access (Han et al., 2024), and numbers on primary care providers, preventable hospitalizations (Bhanot et al., 2024), high school graduation rates and college education levels (Roy et al., 2020), the percentage of single-parent households, and children living in poverty guidelines, and finally percentages of adult obesity and smoking (Rosella & Buajitti, 2024).

# Methods

The data on different health outcomes within United States counties, is obtained from the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute. A multiple regression model is used to assess the linear relation, based on its ease of interpretation (Anandhi & Nathiya, 2023), with no pre-processing of data undertaken before fitting.

## Data Information

The data contains 3142 observations and 20 features.

### Measures

Dependent Variables: Premature age-adjusted mortality

Independent Variables:

1. Median household income: Small Area Income and Poverty Estimates
2. Income inequality: Ratio of household income at the 80th percentile to income at the 20th percentile
3. Census Population Estimates 2016
4. Population Size
5. Population percentages by Race
   * % Non-Hispanic African American:
   * % American Indian and Alaskan Native:
   * % Asian:
   * % Native Hawaiian/Other Pacific Islander:
   * % Hispanic:
   * % Non-Hispanic white:
6. Uninsured adults:
7. Health care costs
8. Other primary care providers
9. Preventable hospital stays: Number of hospital stays for ambulatory-care sensitive conditions per 1,000 Medicare enrollees
10. High school graduation: Percentage of ninth-grade cohort that graduates in four years
11. Some college: Percentage of adults ages 25-44 with some post-secondary education
12. Children in single-parent households: Percentage of children that live in a household headed by single parent
13. Children in poverty: Percentage of children under age 18 in poverty
14. Adult smoking: Percentage of adults who are current smokers
15. Adult obesity: Percentage of adults that report a BMI of 30 or more

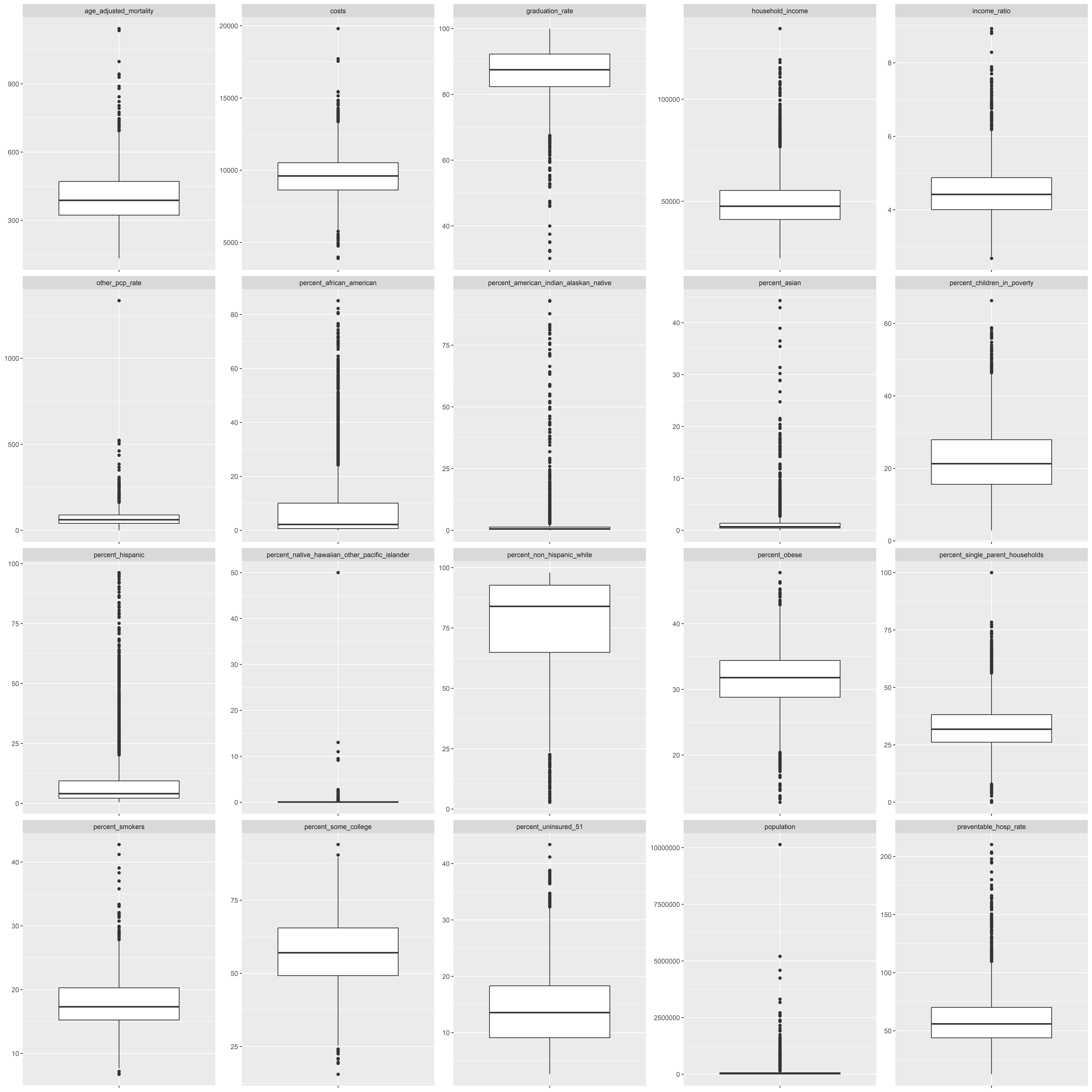
## Descriptive Statistics

Summary statistics of the data is as below:

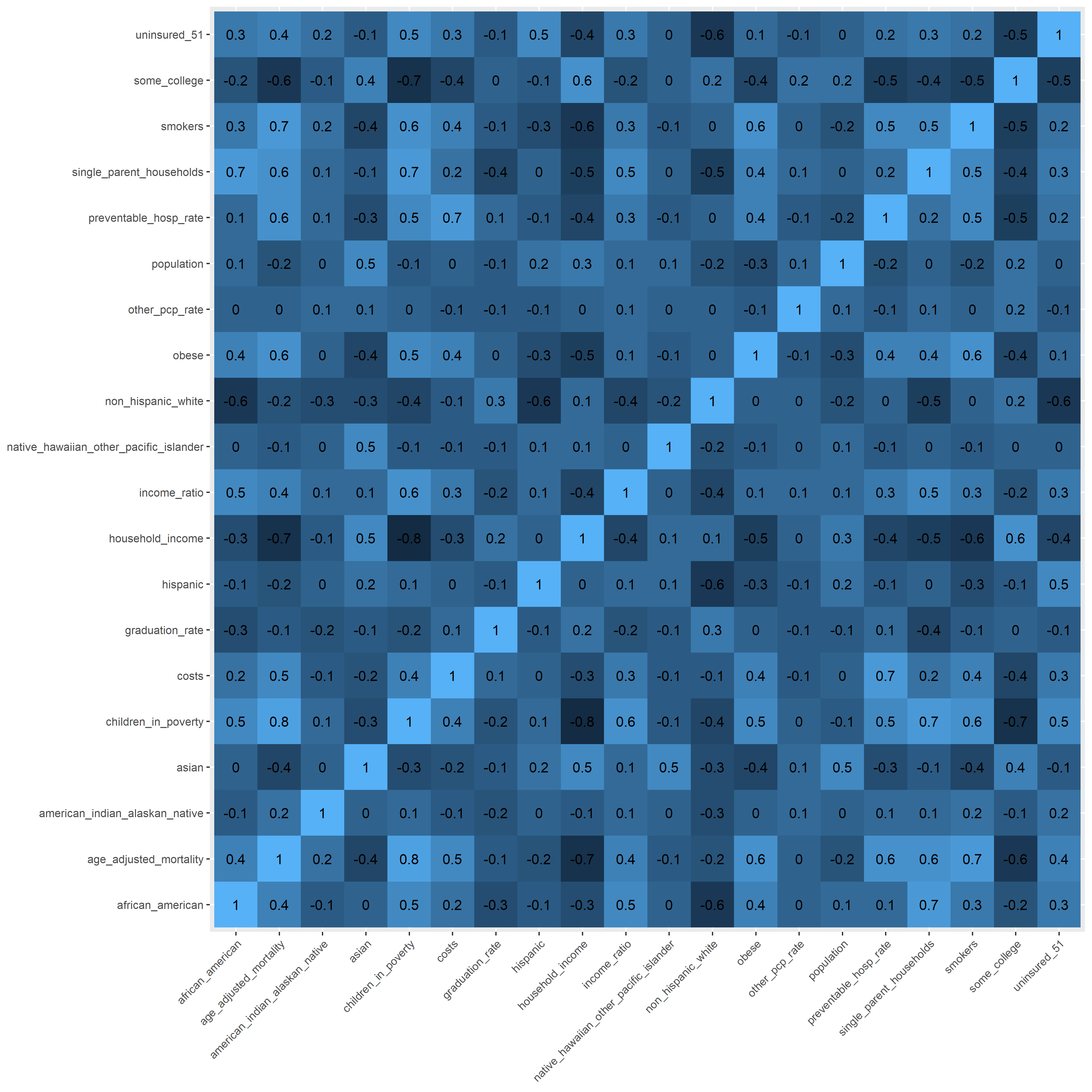
|  | min | mean | median | max | sd |
| --- | --- | --- | --- | --- | --- |
| age\_adjusted\_mortality | 133.0000000 | 401.3173871 | 388.3000000 | 1142.600000 | 109.7046107 |
| household\_income | 22045.0000000 | 49522.1139764 | 47589.0000000 | 134609.000000 | 12887.3096222 |
| population | 88.0000000 | 102841.3472311 | 25771.0000000 | 10137915.000000 | 330319.1419040 |
| percent\_african\_american | 0.0000000 | 8.9426385 | 2.1734134 | 85.151548 | 14.3111752 |
| percent\_american\_indian\_alaskan\_native | 0.0000000 | 2.3093548 | 0.6104830 | 93.067462 | 7.7206458 |
| percent\_asian | 0.0000000 | 1.4856343 | 0.6958972 | 44.265777 | 2.8592849 |
| percent\_native\_hawaiian\_other\_pacific\_islander | 0.0000000 | 0.1349949 | 0.0573143 | 50.000000 | 0.9849561 |
| percent\_hispanic | 0.5017485 | 9.2896091 | 4.0994985 | 96.254016 | 13.6617415 |
| percent\_non\_hispanic\_white | 2.8118228 | 76.5842976 | 83.9893060 | 97.977244 | 20.0874461 |
| percent\_uninsured\_51 | 2.6159435 | 14.2886153 | 13.5673012 | 43.395037 | 6.2494419 |
| costs | 3895.8500000 | 9630.1742424 | 9602.5800000 | 19802.560000 | 1500.4725958 |
| other\_pcp\_rate | 0.0000000 | 71.3990510 | 61.7753300 | 1335.659380 | 53.5738398 |
| income\_ratio | 2.6821628 | 4.5220269 | 4.4206096 | 8.928957 | 0.7354973 |
| preventable\_hosp\_rate | 12.5700000 | 59.8828510 | 55.9250000 | 210.320000 | 23.8996288 |
| graduation\_rate | 30.1369863 | 86.1949627 | 87.5000000 | 100.000000 | 8.2227711 |
| percent\_some\_college | 15.5103402 | 57.2305738 | 57.0647221 | 94.047619 | 11.5503299 |
| percent\_single\_parent\_households | 0.0000000 | 32.6598644 | 31.8126223 | 100.000000 | 10.4310825 |
| percent\_children\_in\_poverty | 2.9000000 | 22.3723018 | 21.3000000 | 66.300000 | 9.2092091 |
| percent\_obese | 12.8000000 | 31.4736792 | 31.8000000 | 47.800000 | 4.5141390 |
| percent\_smokers | 6.7354328 | 17.8726124 | 17.3208934 | 42.754056 | 3.6600786 |

### Exploratory Analysis

Majority of the variables depicting skewness and presence of outlier points in the data.



### Correlation Analysis



## Model Fitting

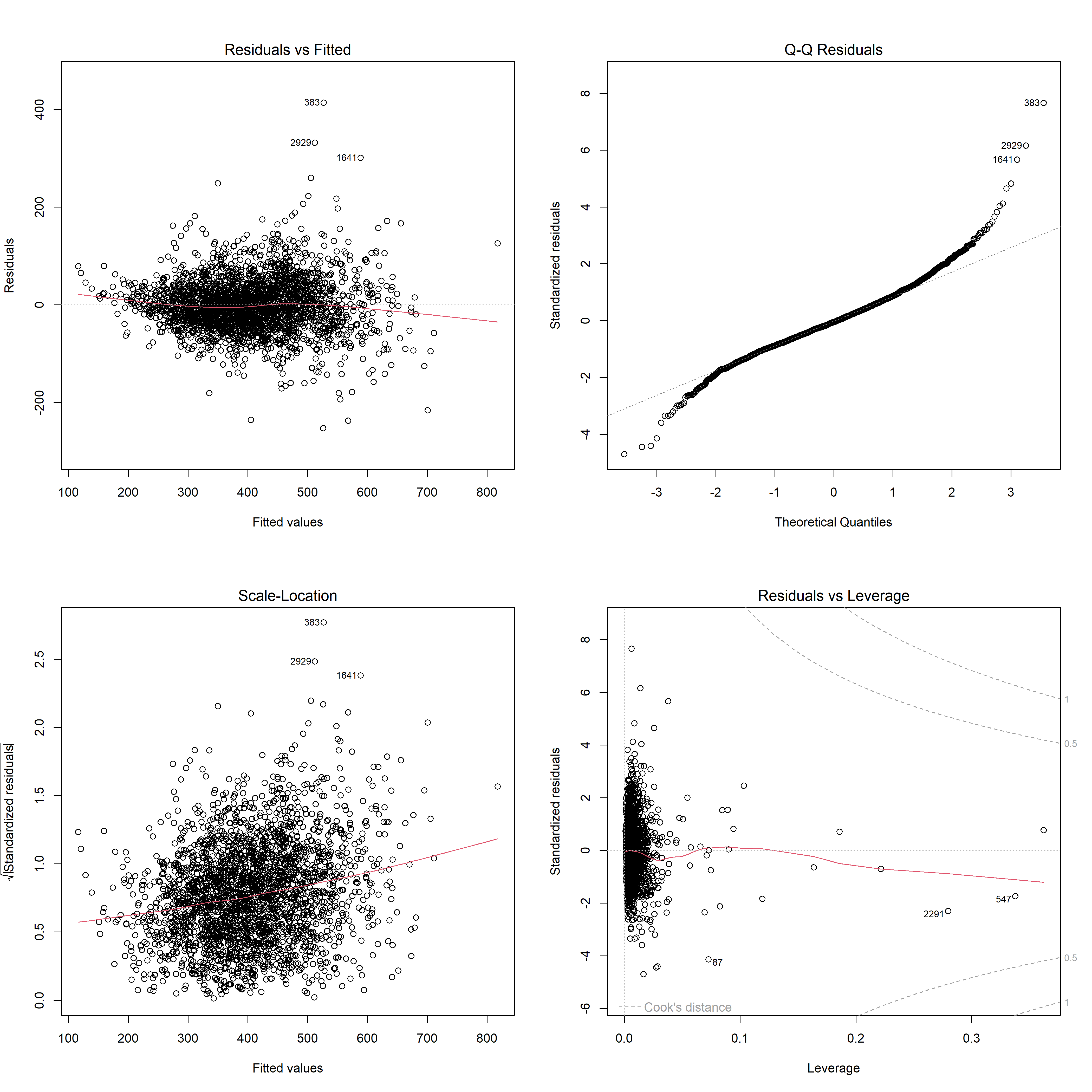
Model Coefficients

| term | estimate | p.value |
| --- | --- | --- |
| (Intercept) | 583.3315756 | 0.0000047 |
| household\_income | -0.0005119 | 0.0056913 |
| population | -0.0000075 | 0.0368519 |
| percent\_african\_american | -5.5871238 | 0.0000045 |
| percent\_american\_indian\_alaskan\_native | -4.5741249 | 0.0003098 |
| percent\_asian | -6.7136490 | 0.0000029 |
| percent\_native\_hawaiian\_other\_pacific\_islander | -4.5601726 | 0.2536166 |
| percent\_hispanic | -6.2382143 | 0.0000001 |
| percent\_non\_hispanic\_white | -5.2450414 | 0.0000184 |
| percent\_uninsured\_51 | 1.3832088 | 0.0000005 |
| costs | 0.0112580 | 0.0000000 |
| other\_pcp\_rate | 0.0960971 | 0.0000197 |
| income\_ratio | 5.5144600 | 0.0074830 |
| preventable\_hosp\_rate | 0.4208420 | 0.0000000 |
| graduation\_rate | -0.0875965 | 0.5631576 |
| percent\_some\_college | -0.9917633 | 0.0000000 |
| percent\_single\_parent\_households | 1.2294889 | 0.0000000 |
| percent\_children\_in\_poverty | 3.4167606 | 0.0000000 |
| percent\_obese | 2.2109633 | 0.0000000 |
| percent\_smokers | 3.7975199 | 0.0000000 |

Model Summary

| r\_squared | adj\_r\_squared | sigma | p\_value | nobs |
| --- | --- | --- | --- | --- |
| 0.7447771 | 0.7429113 | 54.11637 | 0 | 2619 |

## Model Diagnostics



There is no clear pattern in the Residuals vs Fitted plot, leading to assumption that the relation between predictor and dependent variable is linear. The Normal Q-Q depicts normal residual closely following the straight 4-degree dashed line. There is homogeneity of variance of the residuals based on the spread of data points in Scale-Location plot. There are 3 influential observations, 87, 547 and 2291 on Residuals vs Leverage, which may have insignificant influence on the model (Anandhi & Nathiya, 2023).

# Results

The observations with missing data points were automatically excluded from the model. The correlation plot depicts strong positive linear relationship between Premature age-adjusted mortality and Percentage of children under age 18 in poverty at 0.8, and negative with Median household income at -0.7.

The r-squared of 0.7447771 implies that the model explains 74.5 % variance in Premature age-adjusted mortality at the county level. All the independent variables, but Percent Native Hawaiian other Pacific Islander and graduation\_rate are statistically significant at significance level , implying that the significant variables are associated to and can be used as determinants of premature age mortality. The results are consistent with findings of previous studies on race, education and female led households (Mansfield et al., 1999), income and inequlity and race (Cooper et al., 2001) and percentages of adult obesity and smoking (Rosella & Buajitti, 2024) as predictors of Premature age-adjusted mortality. An approach of multilevel analysis (Shavers,2007) can be undertaken to further understand the influence of insignificant variables.

The analysis undertaken is limited to only the data provided and extensive analysis can be undertaken to include other variables, datasets or models of interest that may capture the influence of socio-economic attributes on premature age mortality.

# References

Adler, N. E., & Newman, K. (2002). Socioeconomic disparities in health: Pathways and policies. Health Affairs, 21(2), 60–76. <https://doi.org/10.1377/hlthaff.21.2.60>

Anandhi, P., & Nathiya, E. (2023). Application of linear regression with their advantages, disadvantages, assumption and limitations.

Bchir, A., Bhutta, Z., Binka, F., Black, R., Bradshaw, D., Garnett, G., Hayashi, K., Jha, P., Peto, R., Sawyer, C., Schwartländer, B., Walker, N., Wolfson, M., Yach, D., & Zaba, B. (2006). Better health statistics are possible. The Lancet, 367(9506), 190–193. <https://doi.org/10.1016/s0140-6736(06)68011-2>

Bhanot, K., Erickson, J. S., & Bennett, K. P. (2024). Mortalityminder: Visualization and AI interpretations of social determinants of premature mortality in the United States. Information, 15(5), 254. <https://doi.org/10.3390/info15050254>

Braveman, P. A., Cubbin, C., Egerter, S., Williams, D. R., & Pamuk, E. (2010). Socioeconomic disparities in health in the United States: What the patterns tell us. American Journal of Public Health, 100(S1). <https://doi.org/10.2105/ajph.2009.166082>

Cheng, E., & Kindig, D. (2012). Disparities in premature mortality between high- and low-income US counties. Preventing Chronic Disease. <https://doi.org/10.5888/pcd9.110120>

Cooper, R. S., Kennelly, J. F., Durazo-Arvizu, R., Oh, H.-J., Kaplan, G., & Lynch, J. (2001). Relationship between premature mortality and socioeconomic factors in black and white populations of US metropolitan areas. Public Health Reports, 116(5), 464–473. <https://doi.org/10.1016/s0033-3549(04)50074-2>

Han, X., Hu, X., Zheng, Z., Shi, K. S., & Yabroff, K. R. (2024). Associations of medical debt with health status, premature death, and mortality in the US. JAMA Network Open, 7(3). <https://doi.org/10.1001/jamanetworkopen.2023.54766>

Jobson, J. D., & Jobson, J. D. (1991). Multiple linear regression. Applied multivariate data analysis: regression and experimental design, 219-398.

Krieger, N., Chen, J. T., & Ebel, G. (1997). Can we monitor socioeconomic inequalities in health? A survey of US health departments’ data collection and reporting practices. Public health reports, 112(6), 481.

Mansfield, C. J., Wilson, J. L., Kobrinski, E. J., & Mitchell, J. (1999b). Premature mortality in the United States: The roles of geographic area, socioeconomic status, household type, and availability of medical care. American Journal of Public Health, 89(6), 893–898. <https://doi.org/10.2105/ajph.89.6.893>

Premature age-adjusted mortality\*. County Health Rankings & Roadmaps. (n.d.-a). <https://www.countyhealthrankings.org/health-data/health-outcomes/length-of-life/premature-age-adjusted-mortality?year=2024>

Rosella, L. C., & Buajitti, E. (2024). Risk of premature mortality due to smoking, alcohol use, obesity and physical activity varies by income: A population-based Cohort Study. SSM - Population Health, 25, 101638. <https://doi.org/10.1016/j.ssmph.2024.101638>

Roy, B., Kiefe, C. I., Jacobs, D. R., Goff, D. C., Lloyd-Jones, D., Shikany, J. M., Reis, J. P., Gordon-Larsen, P., & Lewis, C. E. (2020). Education, race/ethnicity, and causes of premature mortality among middle-aged adults in 4 US urban communities: Results from Cardia, 1985–2017. American Journal of Public Health, 110(4), 530–536. <https://doi.org/10.2105/ajph.2019.305506>

Shavers, V. L. (2007). Measurement of socioeconomic status in health disparities research. Journal of the national medical association, 99(9), 1013.

Sorlie, P. D., Backlund, E., & Keller, J. B. (1995). US mortality by economic, demographic, and social characteristics: The National Longitudinal Mortality Study. American Journal of Public Health, 85(7), 949–956. <https://doi.org/10.2105/ajph.85.7.949>