632_paper

County-Level Unemployment Rates in the United States

Akhil Sachar, Arash Ahmadi

Table of contents

1. Introduction

Unemployment is one of the most widely watched measures of economic health, but behind national averages, local realities can vary dramatically. It represents the percentage of the labor force that is actively looking for work but is unable to find employment. By focusing at the county level, we can uncover deeper patterns and better understand the social, economic, and environmental factors that drive unemployment in different parts of the country. Understanding unemployment at the county level is especially important because it highlights regional differences in economic performance and social well-being. Local governments rely on this information for decision-making related to budgets, housing, and social services. The goal of this project is to build a statistical model to identify which county-level factors are most associated with unemployment rates in the United States in 2023. We explore various predictors, including economic indicators (e.g., poverty, cost-to-income ratio), health statistics (e.g., life expectancy, death rates), and quality-of-life metrics (e.g., air quality and the presence of national parks). By applying linear regression modeling techniques, we aim to better understand the drivers of unemployment and provide insights that could support more effective regional planning and policy.

2. Data Description and Sources

The dataset for this analysis was compiled from several reputable U.S. government and health-related sources. Below is a list of the data sources and the specific variables we used from each, along with a short description of each variable:

US Department of Agriculture (USDA)
Website: County-level Data Sets - County-level Data Sets: Download Data | Economic Research Service

- Poverty Rate – Percentage of people living below the poverty line in each county.
- Median Household Income – The middle value of household income distribution, used as an indicator of local wealth.
- Total Population – The number of residents in each county.

US Bureau of Labor Statistics (BLS)
Website: https://www.bls.gov/lau/

- Unemployment Rate – The percentage of the labor force actively seeking but unable to find work (response variable).

County Health Rankings & Roadmaps
Website:https://www.countyhealthrankings.org

- Life Expectancy – Average expected lifespan of residents in each county.
- Uninsured Rate – Percentage of people without health insurance.
- Community Deaths – Total number of deaths (excluding COVID-19) in the county.
- COVID-19 Deaths – Deaths due to COVID-19, used as a public health impact metric.

US Census Bureau
Website: County Population Totals: 2020-2024

- Urban/Rural Classification – A categorical variable indicating if a county is urban or rural.
- National Parks Presence – Indicator of whether the county includes any portion of a national park.
- Educational Attainment (Bachelor's Degree Rate) – Percentage of adults age 25+ with a bachelor's degree or higher.

Derived Variables (Calculated Internally)

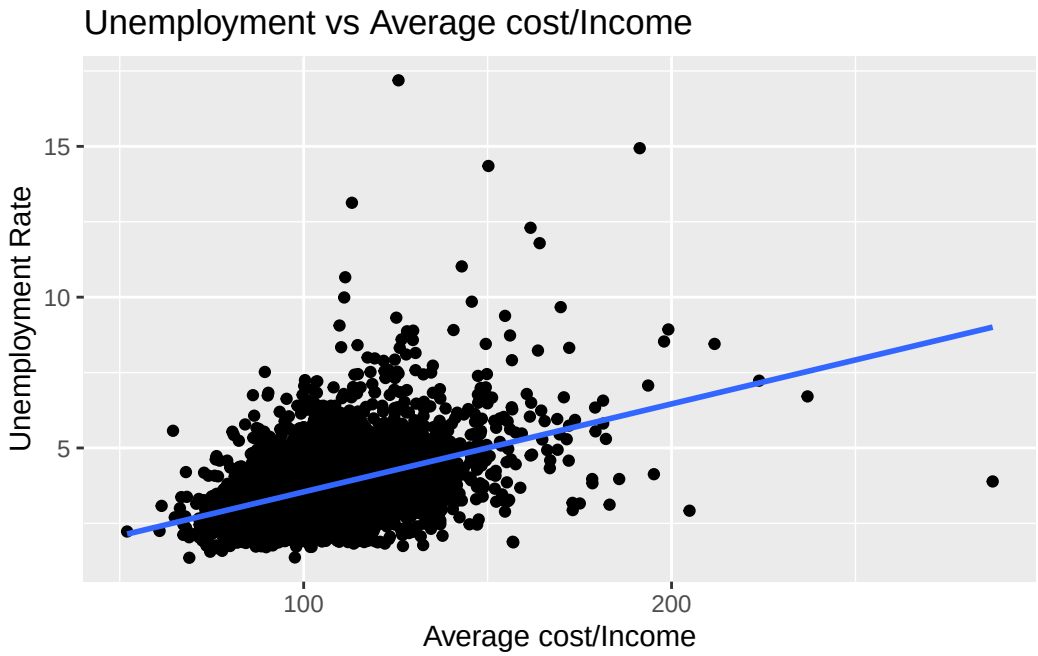
- Cost-to-Income Ratio (C2I) – Calculated by dividing average cost of living by median income. Reflects affordability in each county.
- Air Quality Index (AQI) – A measure of air pollution where a higher score means poorer air quality.

We grouped our variables into three categories:

- Economic: Median income, poverty rate, cost of living
- Social: Percentage uninsured, crime rate, education level
- Environmental: Water quality, air quality, number of national parks

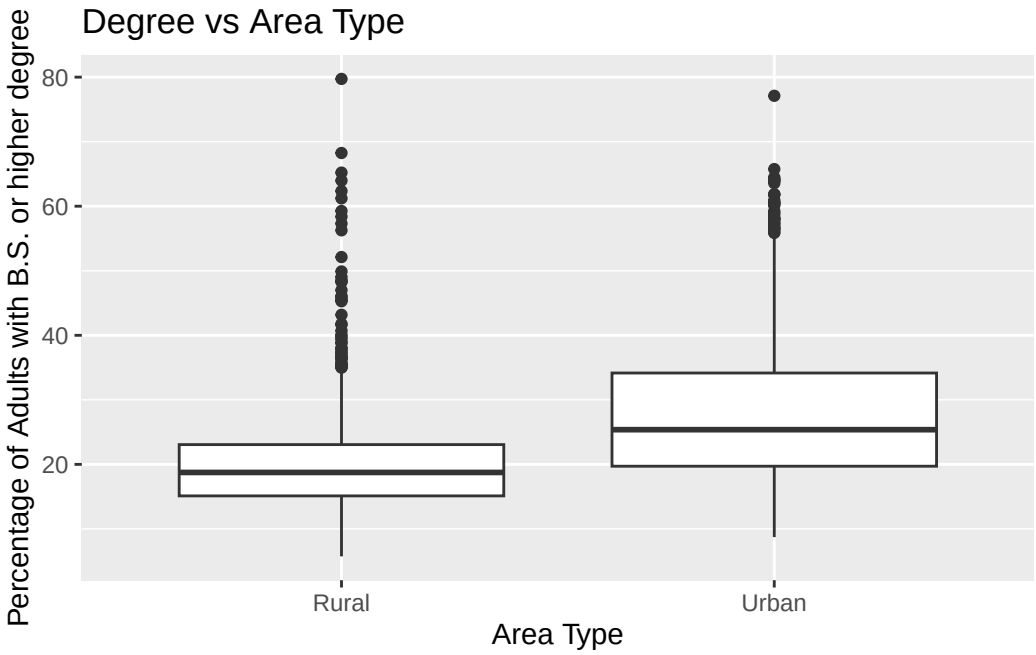
3. Exploratory Data Analysis

```
`geom_smooth()` using formula = 'y ~ x'
```

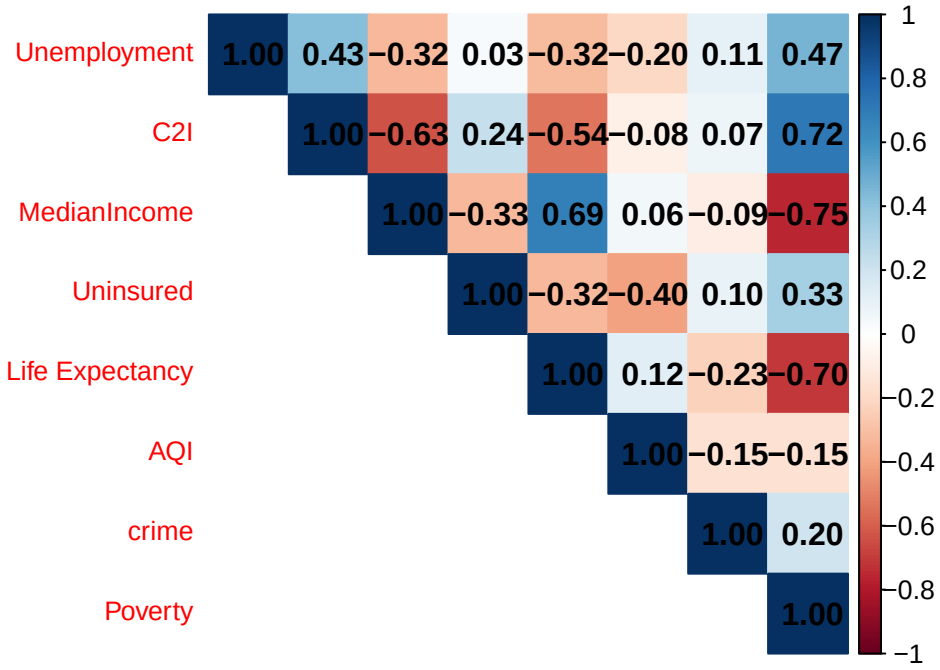


The scatter plot illustrating the relationship between the cost-to-income ratio and unemployment rate serves as a key component of our exploratory data analysis (EDA). The cost-to-income ratio, calculated by dividing average living costs by median income, is used here to reflect local affordability and to address multicollinearity concerns between cost and income. A positive linear trend is evident: as the cost-to-income ratio increases, the unemployment rate also tends to rise. This trend suggests that areas where living costs significantly outpace

incomes are more likely to face elevated unemployment levels, potentially due to heightened economic stress, job displacement, or worker migration. The observed pattern supports the idea that cost burdens act as a predictive factor for unemployment, emphasizing the importance of affordability in local labor market dynamics.



The boxplot comparing urban and rural counties on the percentage of adults with a bachelor's degree or higher provides valuable context in our exploratory data analysis (EDA) related to regional educational attainment. The data show a clear disparity: urban counties have a higher median percentage (27%) of residents with at least a bachelor's degree compared to rural counties (19%), and this difference is reflected in their respective interquartile ranges (urban: 20–33%, rural: 17–25%). This visualization underscores the broader trend that urban areas tend to have more highly educated populations, which can influence a range of socioeconomic outcomes. Given that educational attainment is often associated with lower unemployment rates and greater economic resilience, these disparities may help explain some of the regional variation observed in unemployment patterns. In particular, the lower levels of higher education in rural areas may contribute to more limited employment opportunities and reduced adaptability in times of economic change.



We used the correlation matrix to provide a comprehensive overview of the linear relationships among key socioeconomic and health-related variables in the dataset, serving as a foundational element of our exploratory data analysis (EDA). In this matrix, unemployment, our response variable, exhibits moderate positive correlations with both poverty (0.47) and the cost-to-income ratio (C2I, 0.43). This suggests that areas facing higher poverty rates and greater affordability challenges tend to experience elevated levels of unemployment. In contrast, median income shows strong negative correlations with both poverty (−0.75) and C2I (−0.63), indicating that higher-income regions are generally more affordable and less impoverished. Additionally, median income is positively correlated with life expectancy (0.69), further highlighting its role as an indicator of broader socioeconomic well-being. Together, these relationships point to a tightly interwoven set of conditions—where poverty, income, affordability, and health outcomes are all linked to labor market dynamics. This underscores the importance of accounting for these variables when modeling unemployment patterns.

4. Methods and Results

In this linear regression project, we are analyzing unemployment rates at the county level, considering a range of factors that may affect unemployment. We focused on quality of life factors such as the Air Quality Index (AQI), the presence of parks, and health-related metrics like community deaths, COVID deaths, life expectancy, and the percentage of people insured. We also examined economic factors, including median income, cost of living, and poverty rates.

To refine our model, we used stepwise AIC to eliminate predictors that were not statistically significant, while ensuring the integrity of the model. From this process, we removed median income and cost of living, as both were represented by the C2I variable. The stepwise AIC also removed population, student-teacher ratio, and life expectancy as predictors.

After stepwise AIC, we checked the assumptions of normality and found that the residuals did not meet normality requirements, as indicated by a small p-value in the Shapiro-Wilk test and a W value below 0.95. As a result, we applied a log transformation to the dependent variable.

log(Unemployment) = 1.57809 + -0.00784 * AQI + 0.02856 * NationalPark + 0.00206 * C2I + 0.11401 * Urban.RuralUrban + -0.01383 * Uninsured + -0.00605 * Bachelors + 0.000002753 * Deaths + 0.02332 * Poverty

We also checked for multicollinearity using the Variance Inflation Factor (VIF), and the values were all below 5, with poverty and C2I being the highest at 2.38 and 2.20, respectively—well below the threshold for concern. This suggests there is no significant impact of correlation on the model. When examining the residuals versus fitted values, we found no discernible pattern, indicating that the assumption of constant variance was met. To verify the results of the model, a plot of the standardized residuals against the model's fitted values was made in addition to a Q-Q plot of the standardized residuals. The QQ plot showed some slight deviations from the normal distribution, but given the large sample size, we decided to assume normality despite a small p-value in the Shapiro-Wilk test.

The final model reveals that unemployment increases with higher National Park presence, higher C2I, living in urban areas, higher poverty rates, and slightly with higher death rates. On the other hand, unemployment decreases with better air quality (higher AQI), a higher percentage of uninsured individuals, and a greater proportion of people with Bachelor's degrees.

The R-squared value of our model is 0.358, meaning that approximately 35.8% of the variability in unemployment rates across counties is explained by the included factors. While the model captures some of the variation, there remains a significant amount of unexplained variability, suggesting that there may be other influential factors not accounted for in this model.

5. Conclusion

Our findings suggest that county-level unemployment in the United States is influenced by a combination of economic, health, and geographic factors. Higher poverty rates and a greater cost-to-income ratio (C2I) are both associated with increased unemployment, indicating that financial strain and local affordability challenges significantly impact labor markets. Urban areas, despite offering more services and opportunities, tend to have higher unemployment rates, possibly due to greater job competition, economic inequality, or regional disparities.

In contrast, better air quality and higher educational attainment are linked to lower unemployment, supporting the idea that healthier and more educated populations tend to have more stable employment. One unexpected result is that a higher percentage of uninsured individuals is associated with lower unemployment, a pattern that may reflect complex socio-economic dynamics or hidden variables. Additionally, counties containing national parks show a small increase in unemployment, which may be due to seasonal work patterns in tourism-dependent areas.

While the model accounts for 35.8% of the variation in unemployment, it has several important limitations. Merging datasets from multiple sources required extensive cleaning and introduced inconsistencies. Some variables—such as minimum wage—were unavailable at the county level, and about 2% of the data had to be removed due to missing values. These issues may affect the model's generalizability. Moreover, the moderate R-squared value highlights that unemployment is a complex outcome shaped by many factors beyond those included in our analysis. To improve predictive power, future work could explore additional variables (e.g., housing prices, industry composition) and use more advanced modeling techniques such as LASSO, Ridge regression, or machine learning approaches.

6. Appendix

For supplementary R script, visit github link.