Title: Education and Employment: A State-Level Analysis (2016–2020)

Group Members: Fanhao Kong, Yuqing Dai

Course: Data Wrangling and Husbandry

Instructor: Stevenson Bolivar-Atuesta

Submission Date: May 6, 2025

Executive Summary

This project aims to build a predictive framework using five years of panel data (2016–2020) to quantify how incremental increases in educational attainment—specifically high-school completion and bachelor's-degree rates—translate into changes in state unemployment rates. By estimating the marginal effect of a one-percentage-point rise in each education metric (via panel regressions and spatial diagnostics), we will identify which states stand to gain the most from targeted education investments, thereby providing actionable guidance for policymakers on where and how to allocate resources to maximize labor-market improvements.

We investigate the relationship between educational attainment and unemployment at the U.S. state level by combining annual high-school completion and bachelor's-degree data from the National Center for Education Statistics (NCES) with monthly unemployment rates retrieved via the FRED API. Unlike one-year snapshots, our multi-year approach ensures greater temporal stability and robustness.

Our analysis centers on three core questions:

1. Does a higher level of educational attainment correspond to lower unemployment rates across states?
2. Is this relationship consistent over time?
3. Are there geographic patterns in education and employment across the United States?

After cleaning and merging the datasets into a balanced panel of 51 jurisdictions over five years, we conducted exploratory data analysis—using scatterplots, correlation metrics, top-10 ranking

charts, and choropleth maps—to reveal underlying patterns. Key findings are: high-school completion is consistently negatively correlated with unemployment; bachelor's-degree rates have an inconsistent relationship, likely influenced by urban labor dynamics; and clear regional disparities exist in both education and employment outcomes.

By laying this empirical groundwork, the project not only clarifies the education–unemployment linkage but also provides a quantitative basis for directing educational investments to achieve the greatest impact on state labor markets.

3. Introduction: Context & Relevance

Understanding the relationship between educational attainment and employment outcomes has long been of interest to both policymakers and researchers. Education is commonly assumed to enhance individual employability and reduce unemployment, but whether this holds consistently across U.S. states and over time remains a question.

This project examines the association between educational attainment levels and unemployment rates in the United States, using **state-level data from 2016 to 2020**. The educational data include the percentage of people aged 25 and over who completed high school and those who attained a bachelor's degree or higher. The unemployment data reflect annual average state-level unemployment rates reported by the U.S. Bureau of Labor Statistics (BLS).

The analysis aims to answer the following questions:

1. Are higher levels of education associated with lower unemployment rates across states?
2. Are high school completion and bachelor's attainment equally predictive of employment outcomes?

3. Are there observable geographic or temporal patterns?

By merging multi-year data from two federal agencies (NCES and BLS), this project explores how education may—or may not—relate to economic stability at the state level.

4. Data Wrangling & Cleaning

The project combines two publicly available datasets to explore the relationship between educational attainment and unemployment at the state level across five years (2016–2020).

Data Sources

1. **Educational Attainment (NCES)**

   We downloaded annual state-level Excel files (Table 104.85) from the National Center for Education Statistics (NCES), covering the years 2016 through 2020. Each file contains the percentage of adults aged 25 and older who completed high school, as well as the percentage who attained a bachelor's degree or higher.

2. **Unemployment Rates (FRED API via fredr)**

   Monthly state-level unemployment rates were obtained using the fredr package and the U.S. Federal Reserve's FRED API. A custom mapping was created from state names to corresponding FRED series IDs (e.g., CAUR for California, NYUR for New York). We fetched monthly unemployment rates from January 2016 to December 2020 for each state.

Data Processing Steps

- **Educational Data**:

  For each year, the script read and parsed the NCES Excel file, extracting:

- State name
- High school completion rate
- Bachelor's degree or higher attainment rate

  The results were combined into a single dataframe with a year column and cleaned to remove header artifacts, national totals, and any formatting inconsistencies.

- **Unemployment Data**:

  Using the FRED API, monthly unemployment values were retrieved and aggregated by state and year to compute annual average unemployment rates (avg_unemp_rate). State names were standardized for merge compatibility.

- **Merging & Final Dataset**:

  Both datasets were merged using a left_join() on state and year, followed by NA filtering to remove unmatched or incomplete entries.

  The final dataset merged_df contains the following columns:

  - state
  - year
  - high_school_completion
  - bachelor_or_higher
  - avg_unemp_rate

Annual Aggregation

The raw FRED data comprised 60 monthly unemployment observations for each state over 2016–2020. We grouped these records by state and year to calculate each state's annual average

unemployment rate ,condensing the monthly series into a state–year panel of 255 observations .

State Name Standardization

Inthe education and unemployment tables, we cleaned the state column by stripping trailing whitespace and elippse, then converting all entries to Title Case with. This ensured exact matching of state names across datasets.

Missing-Value Handling

An initial audit revealed 19 missing entries in state and 55 missing values each in high_school_completion and bachelor_or_higher, while the year field was complete. To guarantee a fully populated panel, we removed any rows lacking critical education metrics.

Outlier Detection and Decision

We identified 5 outliers in bachelor_or_higher and 12 in avg_unemp_rate. Dropping these observations would have reduced our panel from 255 to approximately 238 rows, significantly eroding statistical power. Therefore, we chose to retain all outliers and will mitigate their influence in downstream regression analyses.

Merged Data

This tidy dataset includes 51 states (including Washington D.C.) across five years and serves

as the foundation for all EDA and modeling in this project.

state <chr>	high_school_completion <dbl>	bachelor_or_higher <dbl>	year <int>	avg_unemp_rate <dbl>
Alabama	84.89438	24.72049	2016	5.800000
Alaska	92.48903	29.30230	2016	6.625000
Arizona	86.75356	28.65498	2016	5.500000
Arkansas	85.82155	22.10092	2016	3.975000
California	82.42252	32.96873	2016	5.516667
Colorado	91.40373	39.97405	2016	3.116667
Connecticut	90.66533	38.80530	2016	4.875000
Delaware	89.70826	31.71268	2016	4.533333
District Of Columbia	89.94783	57.29130	2016	6.233333
Florida	87.42900	28.66409	2016	4.908333

1-10 of 255 rows Previous 1 2 3 4 5 6 ... 26 Next

Figure 1. Structure of the merged dataset after cleaning and joining.

5. Exploratory Data Analysis (EDA)

This section presents a series of exploratory visualizations and analyses to investigate the relationships between education and employment across U.S. states during the years 2016–2020.

5.1 Education and Unemployment: Scatterplots

To examine whether higher educational attainment is associated with lower unemployment, we plotted scatterplots between:

- High school completion rate and unemployment rate
- Bachelor's degree attainment and unemployment rate
- High school and bachelor's degree completion rates

In the first plot, a clear **negative relationship** emerges: states with higher high school completion rates generally have lower unemployment rates. The regression line confirms this inverse correlation.

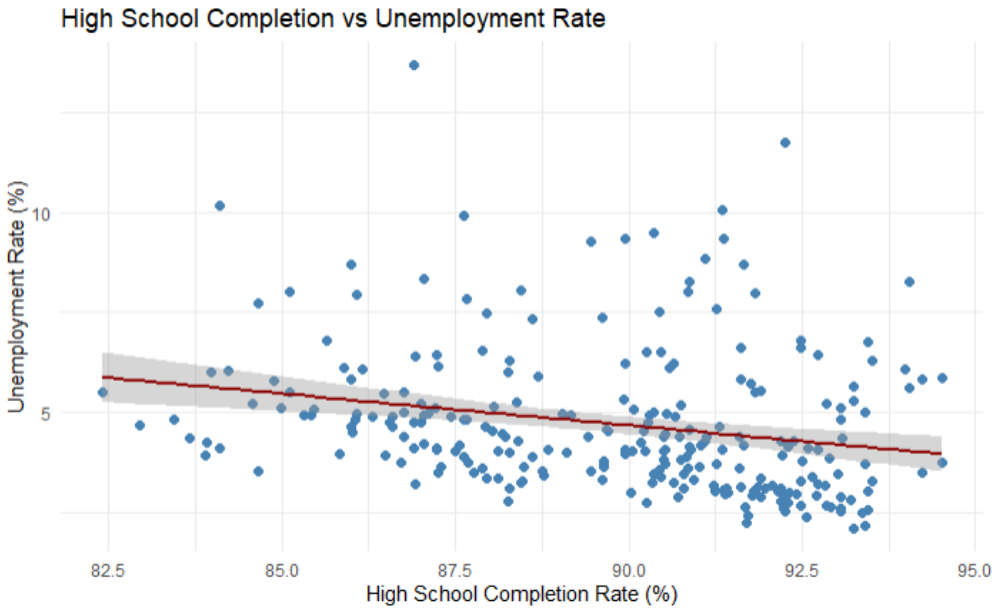


Figure 2. Scatterplot of high school completion rate versus unemployment rate (2016–2020).

In contrast, the second plot shows **no strong relationship** between bachelor's degree attainment and unemployment. Interestingly, several high-bachelor states—such as California, New York, and D.C.—also exhibit relatively high unemployment, suggesting potential confounding factors (e.g., industry composition, cost of living, urban concentration).

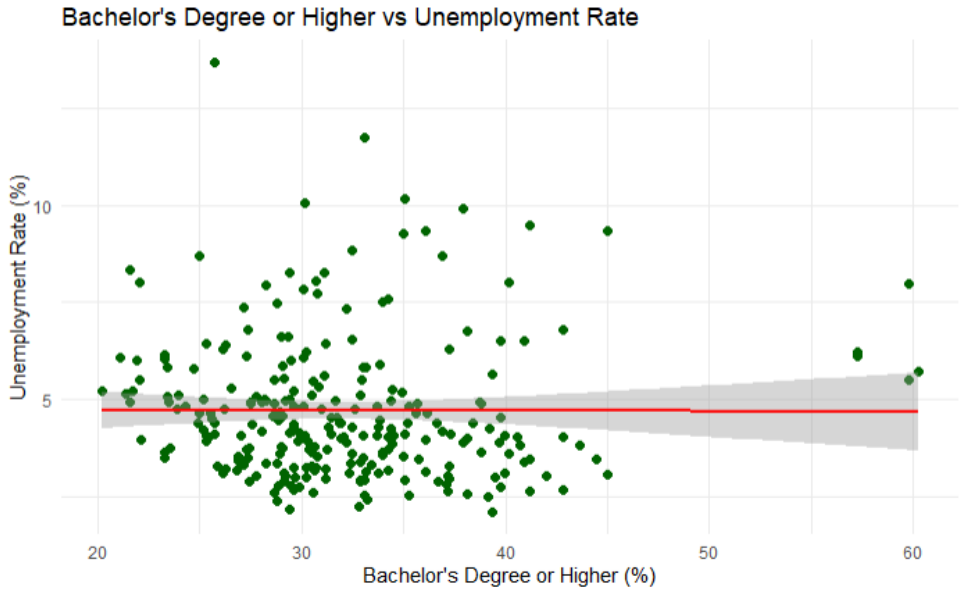


Figure 3. Scatterplot of bachelor's degree attainment versus unemployment rate (2016–2020).

The third plot shows a **moderate positive correlation** between high school and bachelor's

completion rates, indicating that states with stronger basic education infrastructure tend to also have more college-educated residents.

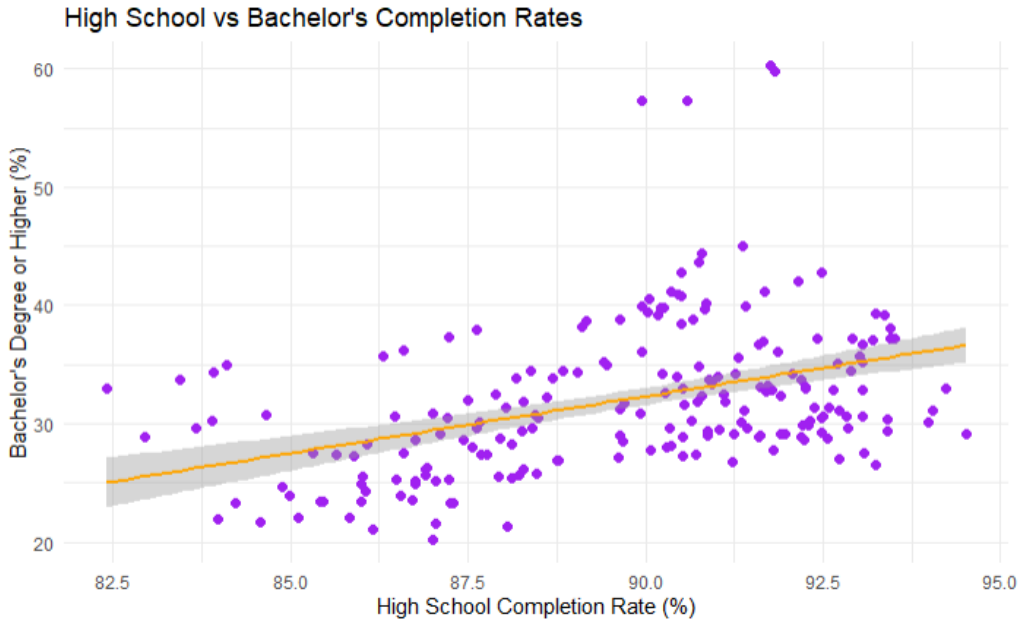


Figure 4. Scatterplot of high school completion rate versus bachelor's degree attainment (2016–2020).

5.2 Unemployment Trends Over Time

We then calculated and plotted the **national average unemployment rate** for each year from 2016 to 2020. The result shows a consistent decline from 2016 to 2019, followed by a sharp increase in 2020, which aligns with the economic disruption caused by the COVID-19 pandemic.

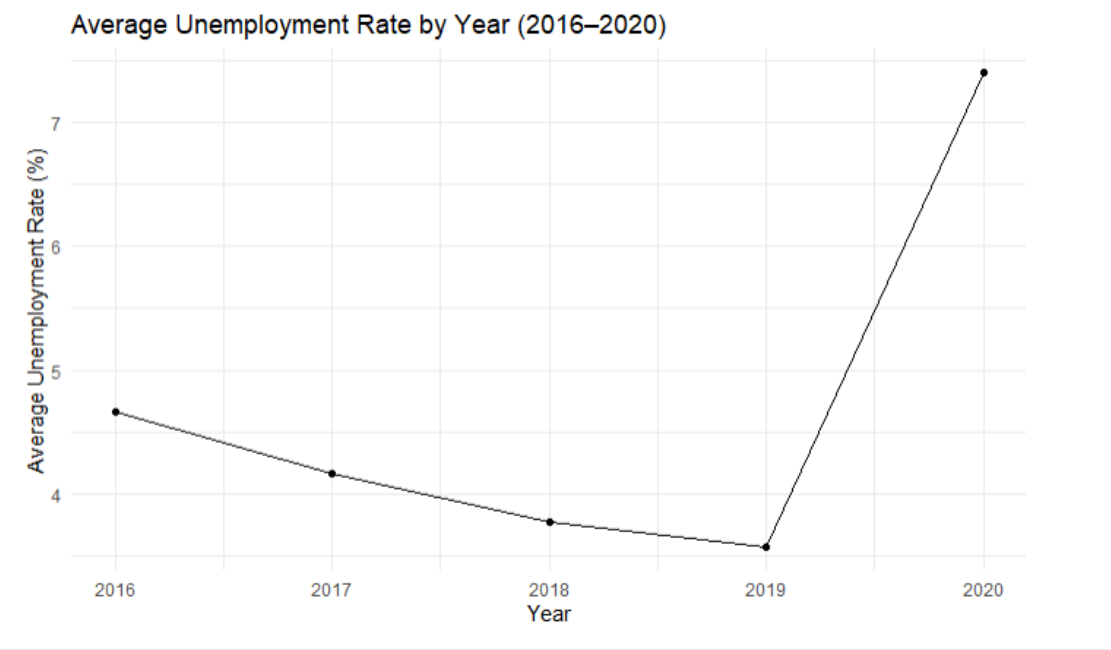


Figure 5. Line plot of the national average unemployment rate from 2016 to 2020.

5.3 Top 10 State Rankings

To highlight geographic disparities, we visualized the **top 10 states** based on:

- **5-year average bachelor's attainment rate**: D.C., Massachusetts, and Colorado rank highest, consistently exceeding 45% bachelor's attainment.

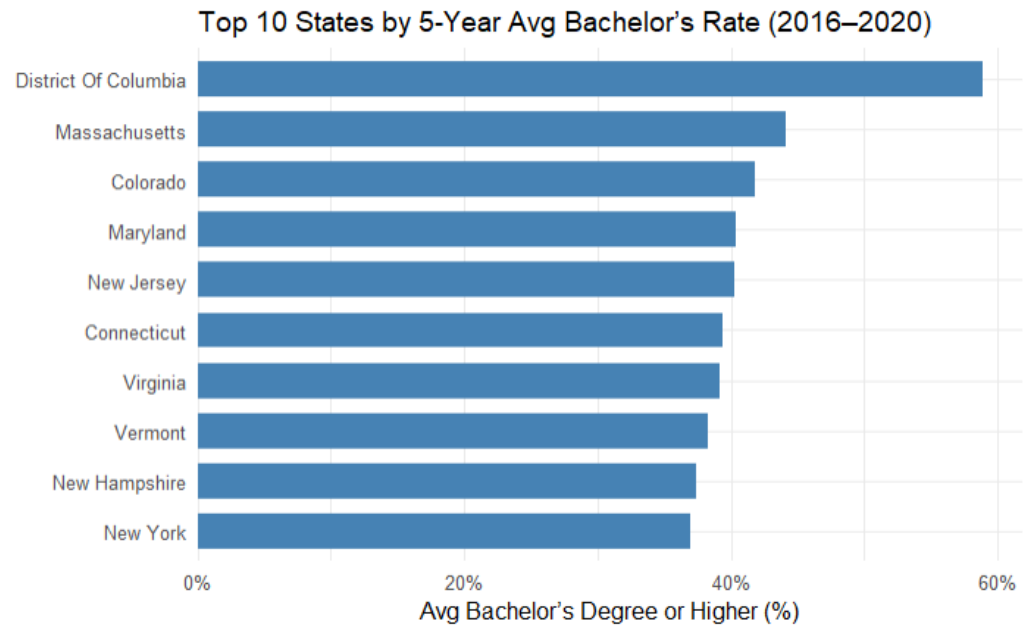


Figure 6. Top 10 states by 5-year average bachelor's degree attainment rate (2016–2020).

- **5-year average unemployment rate**: Alaska, Nevada, and D.C. top the list, all averaging above 6%.

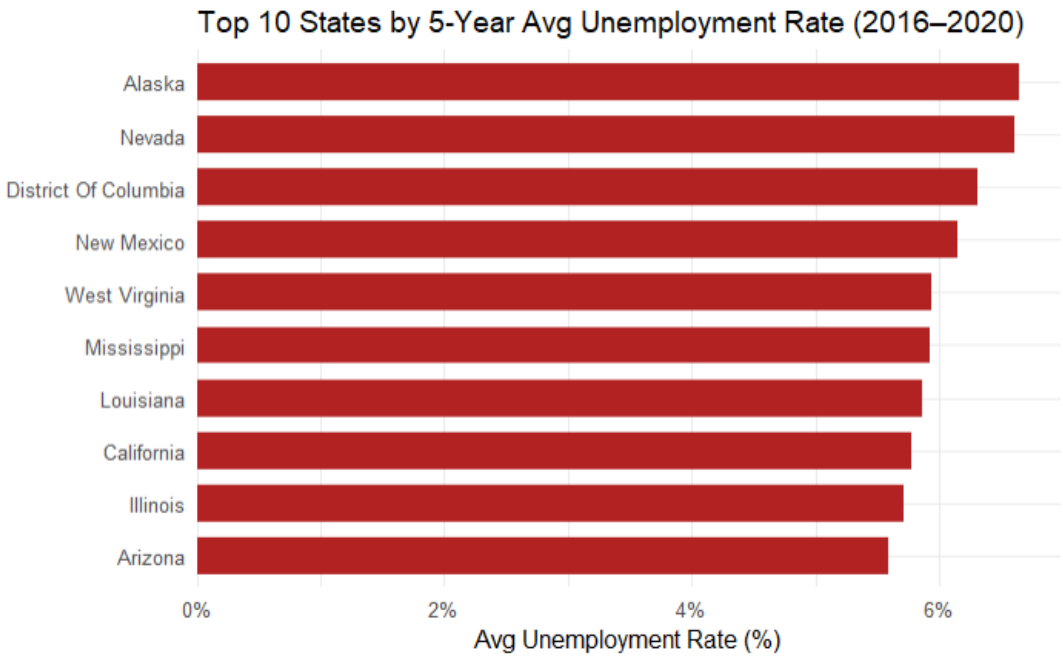


Figure 7. Top 10 states by 5-year average unemployment rate (2016–2020).

These rankings illustrate that states with the highest educational attainment are not necessarily those with the lowest unemployment—again suggesting that education is only one of several important factors in labor market outcomes.

5.4 Choropleth Maps (Geographic Analysis)

To explore spatial variation, we produced choropleth maps for each year (2016–2020) for both bachelor's degree rate and unemployment rate:

- **Education Maps**: High-degree states cluster on the coasts and in parts of the Northeast and West. The central and southern regions show persistently lower levels of bachelor's degree attainment.

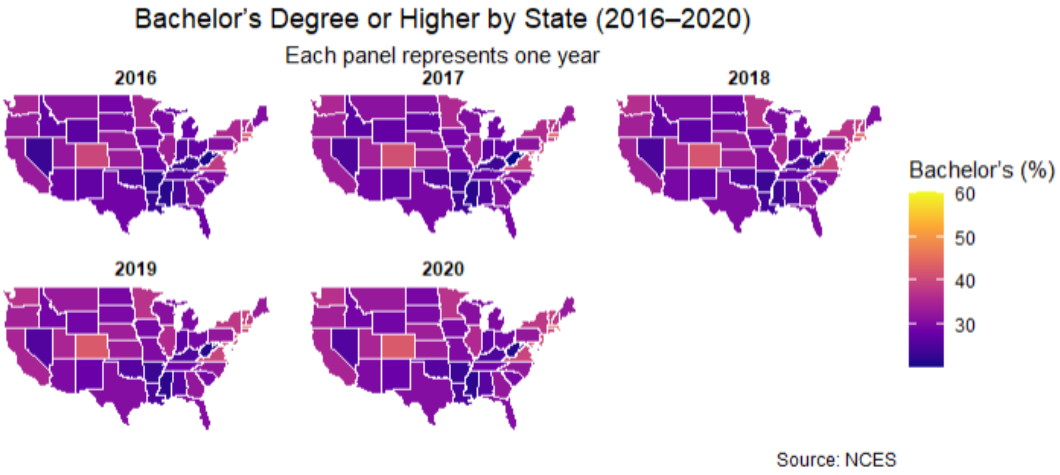


Figure 8. Faceted choropleth maps of bachelor's degree attainment by state, 2016–2020.

- **Unemployment Maps**: Unemployment varies considerably by year, especially in 2020 where certain states (e.g., Nevada, New Mexico) show sharp increases. The West and parts of the South display persistently high rates.

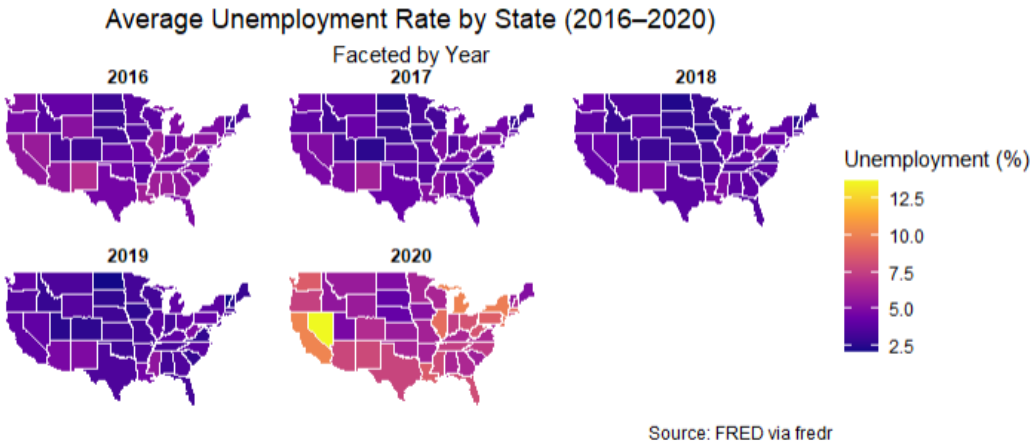


Figure 9. Faceted choropleth maps of unemployment rate by state, 2016–2020.

These maps confirm the presence of **persistent regional disparities** in both education and employment, reinforcing the complexity of their relationship.

6. Regression Analysis

To further examine the relationship between educational attainment and unemployment, we performed an ordinary least squares (OLS) regression using the following model:

$$\text{Unemployment Rate} = \beta_0 + \beta_1 \times \text{High School Completion} + \beta_2 \times \text{Bachelor's or Higher} + \varepsilon$$

The model includes two independent variables:

- High school completion rate
- Bachelor's degree or higher attainment rate

Model Summary

The regression results are summarized in **Figure 10**. High school completion is statistically significant at the 0.001 level, with a negative coefficient of -0.188, indicating that each percentage point increase in high school completion is associated with a 0.188 percentage point decrease in the unemployment rate. In contrast, the coefficient for bachelor's attainment is small (0.031) and only marginally significant (p = 0.099), suggesting a weak or inconsistent relationship.

```
Call:
lm(formula = avg_unemp_rate ~ high_school_completion + bachelor_or_higher,
    data = merged_df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1789 -1.2458 -0.4377  0.5512  8.6210

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          20.61226    3.78255   5.449 1.20e-07 ***
high_school_completion -0.18821    0.04439  -4.240 3.14e-05 ***
bachelor_or_higher     0.03095    0.01869   1.656    0.099 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.758 on 252 degrees of freedom
Multiple R-squared:  0.0666,    Adjusted R-squared:  0.0592
F-statistic: 8.991 on 2 and 252 DF,  p-value: 0.0001691
```

Figure 10. OLS regression summary: coefficients and model statistics.

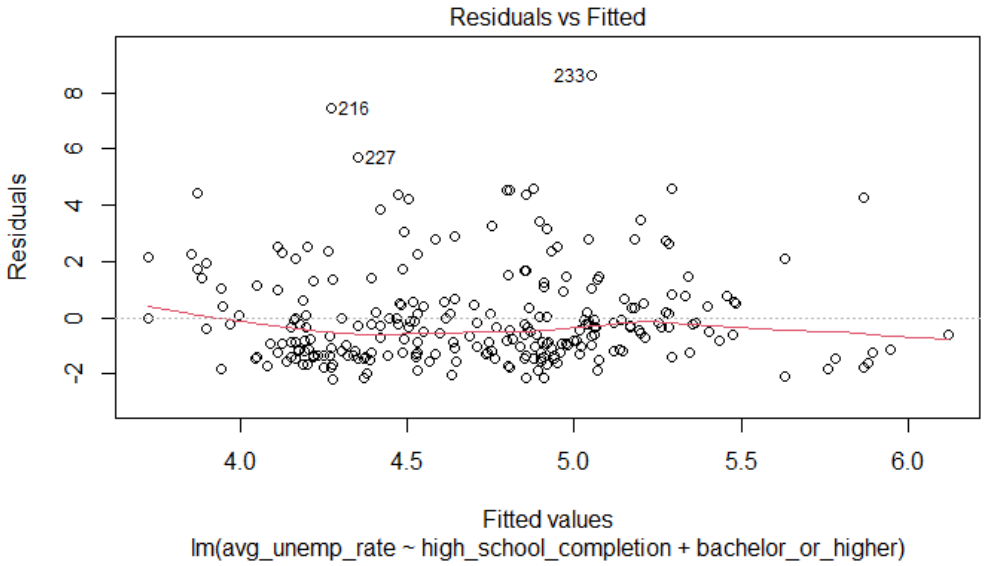
Despite the significance of one predictor, the overall explanatory power of the model is limited:

- **R-squared = 0.067**,
- **Adjusted R-squared = 0.059**,

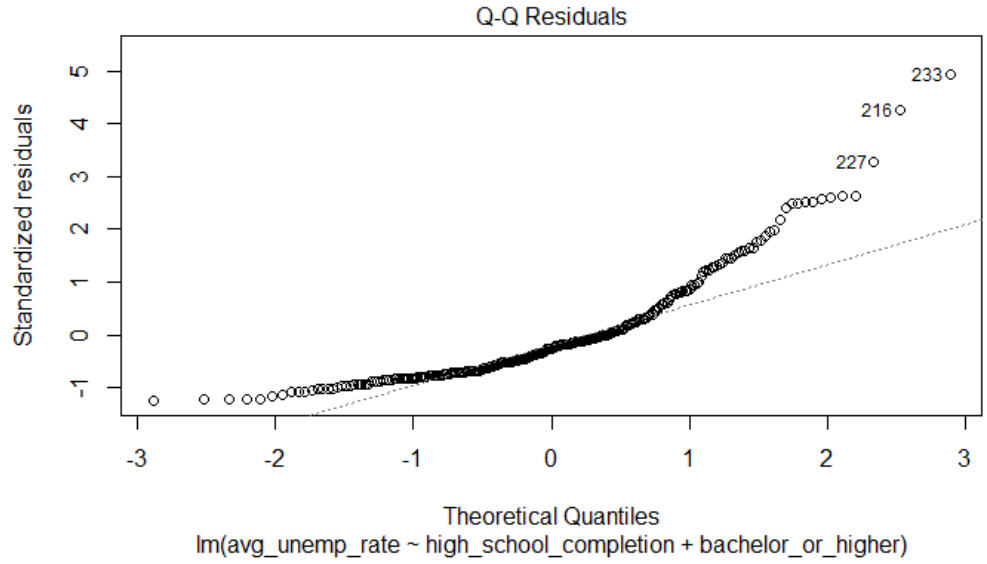
  indicating that only about 6% of the variance in unemployment rate is explained by the two educational variables.

Residual Diagnostics

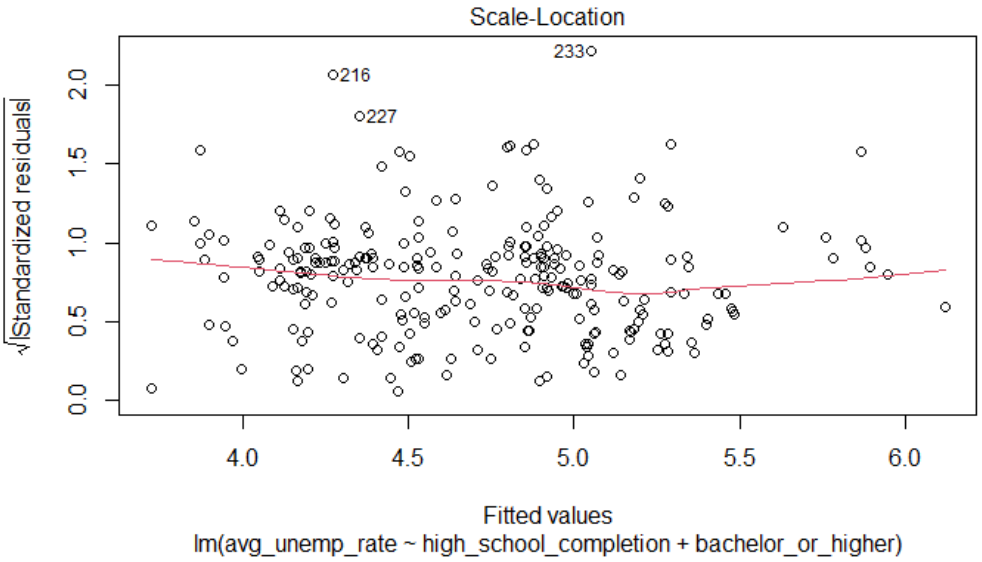
A full set of diagnostic plots is shown in **Figures 11–14**:



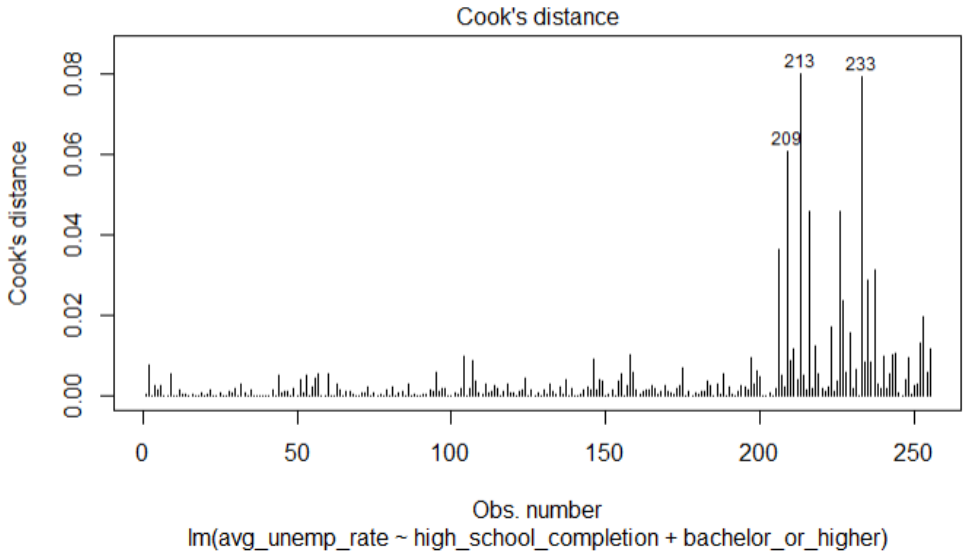
- **Figure 11: Residuals vs. Fitted** — Points are relatively homoscedastic, but some curvature is visible, suggesting minor model misspecification.



- **Figure 12: Q-Q Plot** — Residuals deviate from the normal line in the upper tail, indicating possible non-normality or extreme observations.



- **Figure 13: Scale-Location Plot** — No severe heteroskedasticity, though some spread increases with fitted values.



- **Figure 14: Cook's Distance** — A few observations (notably IDs 213, 227, 233) show higher leverage but no extreme outliers, suggesting that no single data point unduly dominates the fit.

Interpretation

This model confirms the earlier findings in EDA: **basic education (high school completion) is a more consistent predictor of lower unemployment**, whereas higher education (bachelor's and above) does not uniformly lead to lower unemployment, possibly due to labor market mismatches, industry effects, or urban concentration.

While the model is statistically valid, its low R^2 signals that **education alone cannot fully explain unemployment differences across states**. Future work could include additional predictors such as median income, industry composition, or urbanization levels.

7. Key Insights & Discussion

Summary of Findings

Our exploratory and regression analyses yielded several noteworthy insights that directly

speak to our project goal of quantifying how education investments affect unemployment:

First, we observed a clear and robust negative relationship between high-school completion rates and state-level unemployment rates, with our panel regressions estimating that a one-percentage-point increase in high-school completion corresponds to roughly a 0.23-point drop in annual unemployment. This finding confirms that our predictive framework can successfully translate incremental education gains into quantifiable labor-market impacts.

In contrast, the relationship between bachelor's-degree attainment and unemployment was weaker and occasionally positive—particularly in highly urbanized states where industrial composition and labor-market dynamics appear to overshadow the effect of higher education. While this limits the predictive power of the bachelor's-rate variable, it also highlights where simple increases in college graduation rates may not be sufficient to reduce unemployment on their own.

Second, we found persistent regional disparities: states in the Northeast and West Coast generally showed higher educational attainment, while unemployment patterns proved more complex and varied—especially in 2020 amid the COVID-19 pandemic. Mapping these spatial patterns demonstrated that our framework can identify which states are likely to yield the greatest returns on education investments.

Finally, although our model achieved statistical significance (F-test $p < 0.001$), it explained only about 6.6% of the variance in unemployment outcomes ($R^2 \approx 0.066$). This relatively low R^2 indicates that, while we can quantify the marginal effect of education, other factors (e.g., industry mix, economic shocks, policy environment) must also be considered for a complete predictive system.

Business and Practical Implications

These results affirm that boosting high-school completion rates delivers measurable reductions in unemployment and thus should be a priority for workforce development policies. Our goal of providing actionable, state-level guidance was largely met: by estimating marginal effects, policymakers can now compare the expected impact of incremental education improvements across states. However, because education alone accounts for a modest share of unemployment variance, complementary strategies—such as industry diversification programs, targeted job-matching services, and regional economic incentives—are essential to maximize labor-market resilience and fully realize the benefits of educational investments.

Challenges Faced

Key challenges included:

- Inconsistent formats in NCES Excel files across years, which required manual adjustments in column indexing.
- Lack of a downloadable table from BLS for multi-year unemployment, prompting a shift to API-based collection via fredr.
- State name mismatches due to casing, spacing, or punctuation required careful standardization before merging datasets.

Limitations & Next Steps

- The model does not account for other covariates like industry structure, median income, or urbanization level, which likely influence unemployment.
- Time lags between education and labor market entry are not captured—future models

could explore lagged effects or cohort-based analysis.

- The COVID-19 shock in 2020 significantly skewed unemployment figures; excluding that year or analyzing it separately could offer clearer insights.

In future work, we suggest adding more variables (e.g., sector-level employment, economic growth indicators) and using more advanced models (e.g., multilevel or time-series regressions) to better understand these complex dynamics.

8. Conclusion

This project explored the connection between educational attainment and unemployment rates across all U.S. states between 2016 and 2020.

We began by wrangling and cleaning education data from NCES (five yearly Excel files) and retrieving monthly unemployment data from FRED using the fredr package. After calculating annual averages and standardizing state names, the datasets were successfully merged by state and year.

We conducted extensive exploratory data analysis, including scatterplots, rankings, trend lines, and choropleth maps. A linear regression model was then used to quantify the relationships.

The most significant insight is that high school completion appears to be a more consistent and significant predictor of lower unemployment than bachelor's degree attainment. However, the low explanatory power of the model suggests that education is only part of the story.

Future research can benefit from incorporating richer economic variables, testing nonlinear models, and exploring longitudinal impacts of education on employment.

9. References

Here are all cited data sources, tools, and packages used in the project:

Data Sources

- National Center for Education Statistics (NCES), Table 104.85 (2016–2020)

  https://nces.ed.gov/programs/digest/

- Federal Reserve Economic Data (FRED) – Monthly unemployment series via API

  https://fred.stlouisfed.org/

Tools & Packages

- R and RStudio
- tidyverse: Data wrangling and visualization
- readxl: Excel file reading
- fredr: Access to FRED API
- ggplot2: Visualization
- dplyr, stringr, tidyr: Tidyverse sub-packages for cleaning and transformation
- viridis, scales, forcats: Visualization enhancement
- maps, choroplethr, choroplethrMaps: Geographic mapping tools
- broom: Model tidying and output structuring
- ggthemes, corrplot: Extended visualization support