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 investigates the relationship between educational attainment and unemployment at the U.S. state level.

We use state-level data from the **National Center for Education Statistics (NCES)** on high school completion and bachelor's degree attainment, combined with unemployment data from **the U.S. Bureau of Labor Statistics (BLS)**.

Unlike analyses that rely on data from a single year, our study uses five years of data (2016–2020) to ensure greater temporal stability and robustness.

The core questions we explore are:

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, we performed exploratory data analysis (EDA) across 51 states (including Washington D.C.), incorporating scatterplots, correlation analysis, top-10 ranking charts, and geographic choropleth maps.

Key findings include:

- High school completion rates are negatively correlated with unemployment, suggesting that broader basic education access may help stabilize labor markets.
- Bachelor's degree attainment does not consistently predict lower unemployment. In fact, several states with higher proportions of college-educated residents also experience higher unemployment, possibly due to urban labor market dynamics or industrial composition.
- There are clear regional disparities in both education and employment outcomes.

This project lays the groundwork for further modeling and policy-focused analysis and demonstrates a structured approach to merging, transforming, and visualizing multi-source real-world data.

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

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.

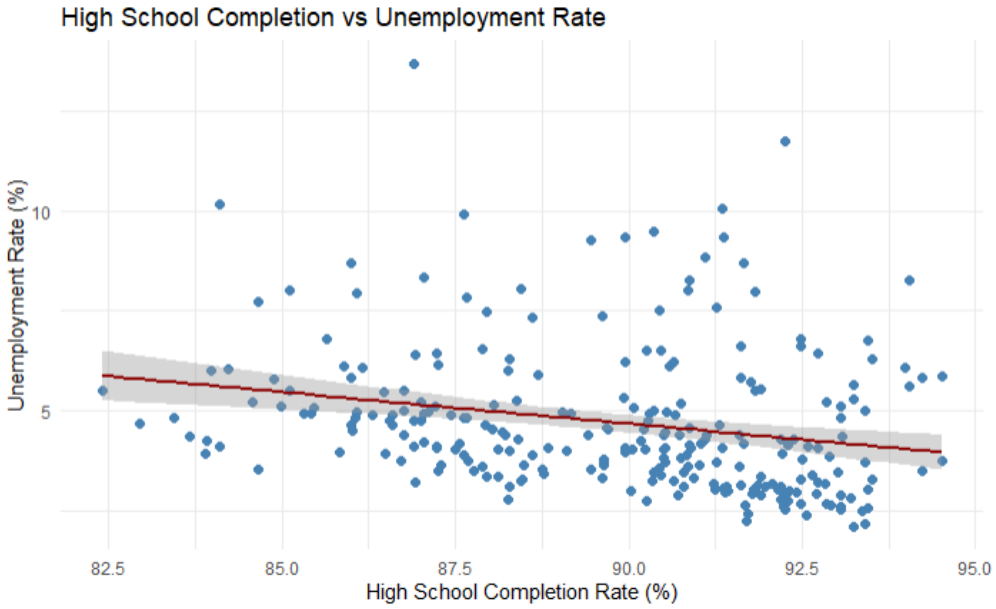


Figure 2. Scatterplot of high school completion rate versus unemployment rate (2016–2020).

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).

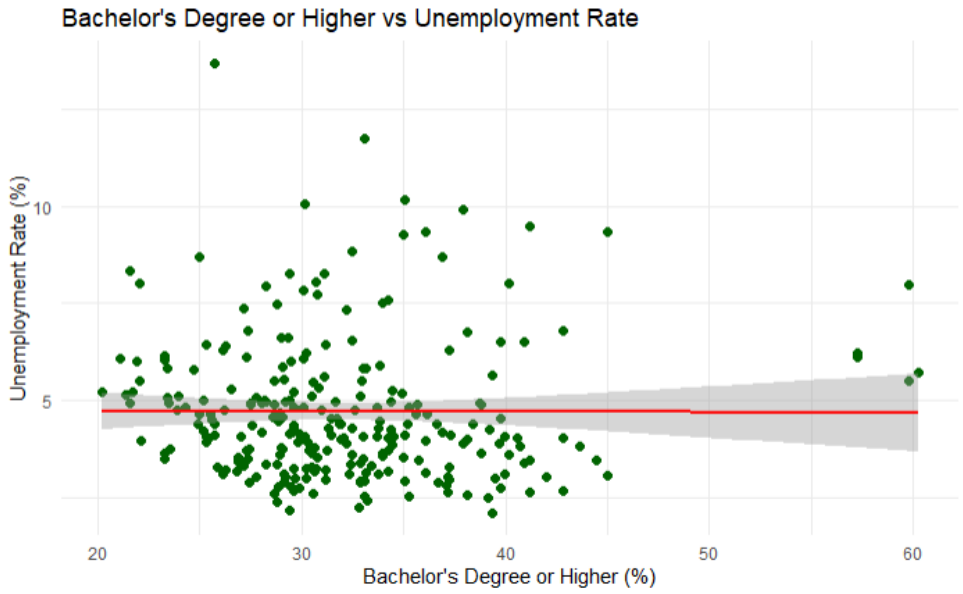


Figure 3. Scatterplot of bachelor's degree attainment versus unemployment rate (2016–2020).

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.

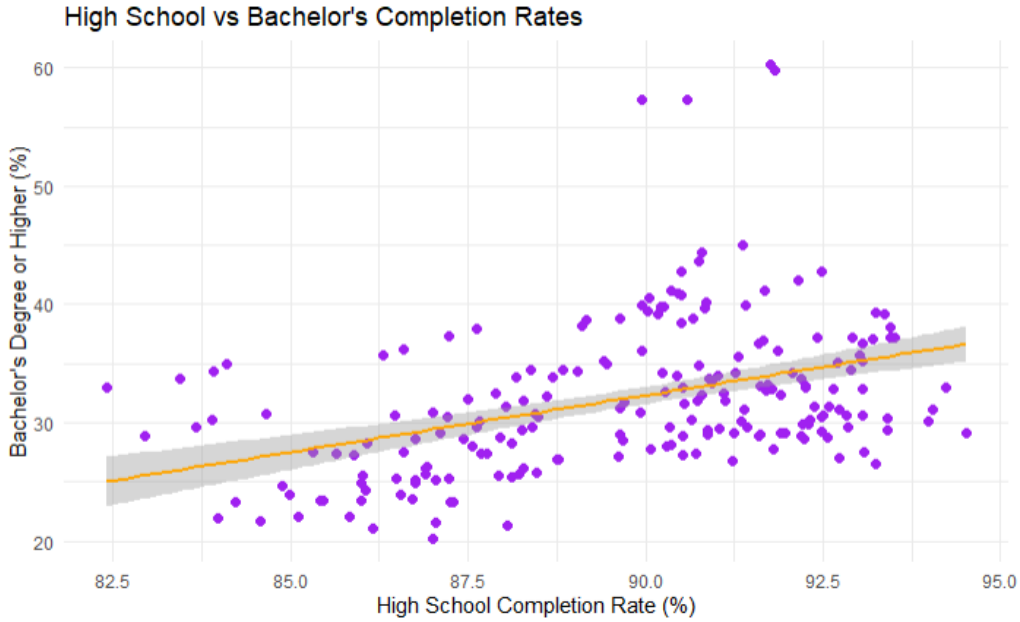


Figure 4. Scatterplot of high school completion rate versus bachelor's degree attainment (2016–2020).

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.

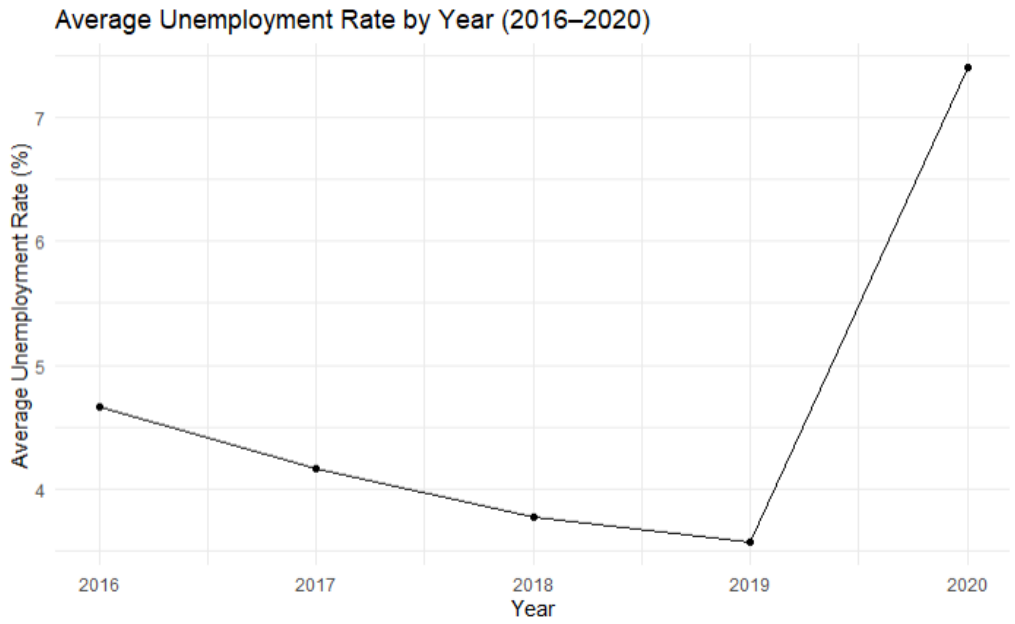


Figure 5. Line plot of the national average unemployment rate from 2016 to 2020.

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.

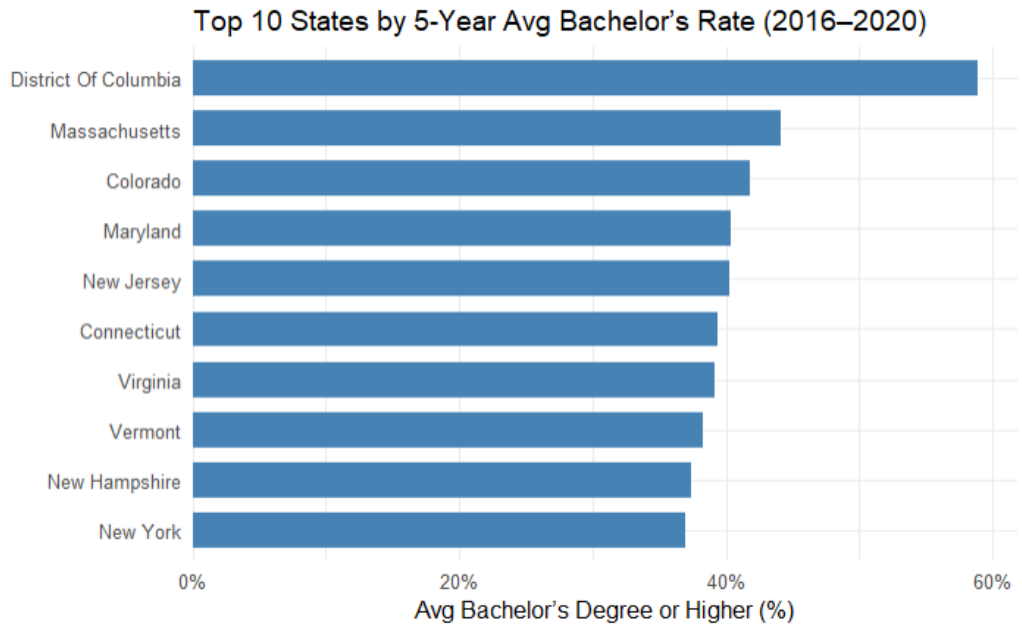


Figure 6. Top 10 states by 5-year average bachelor's degree attainment rate (2016–2020).

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

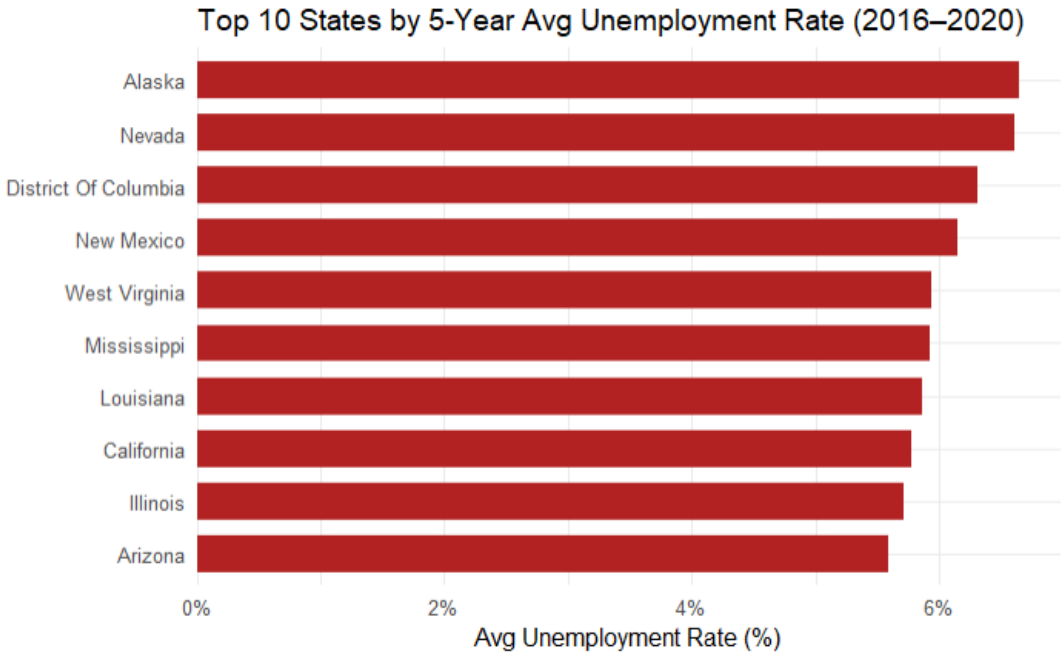


Figure 7. Top 10 states by 5-year average unemployment rate (2016–2020).

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.

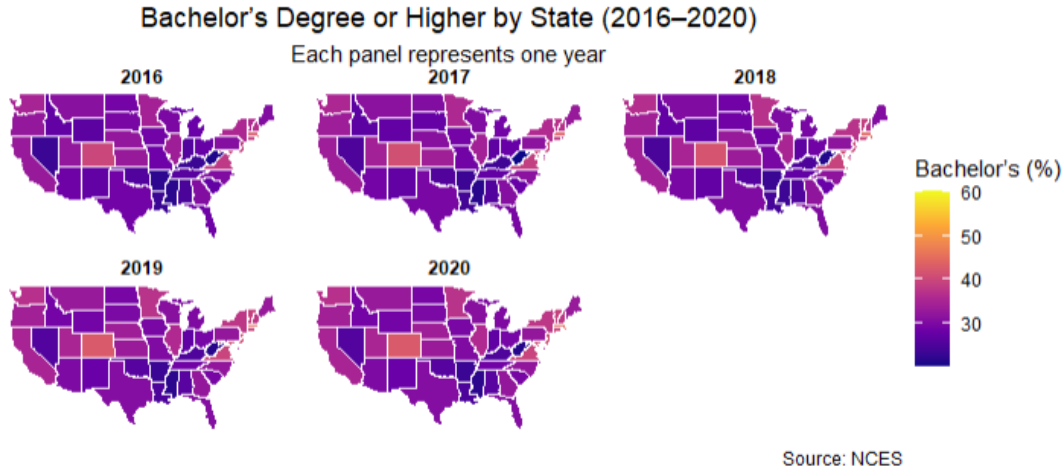


Figure 8. Faceted choropleth maps of bachelor's degree attainment by state, 2016–2020.

- **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.

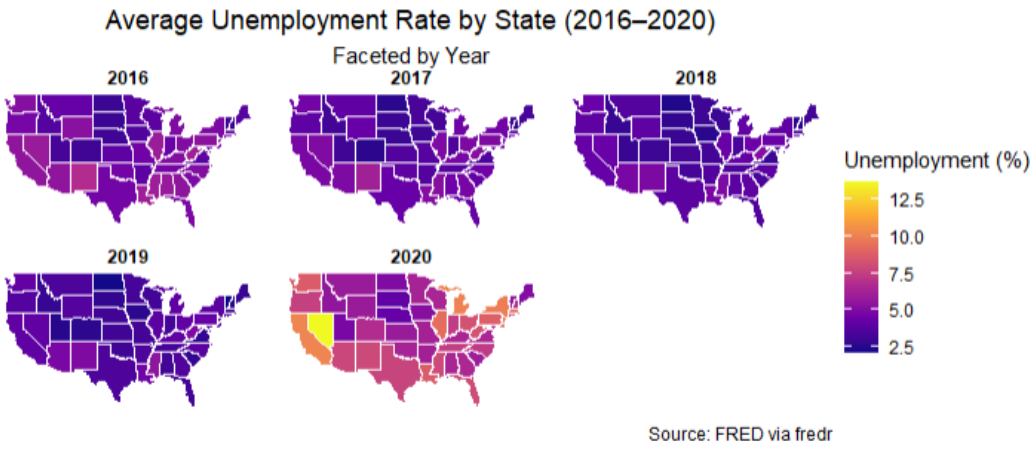


Figure 9. Faceted choropleth maps of unemployment rate by state, 2016–2020.

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:

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

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
```

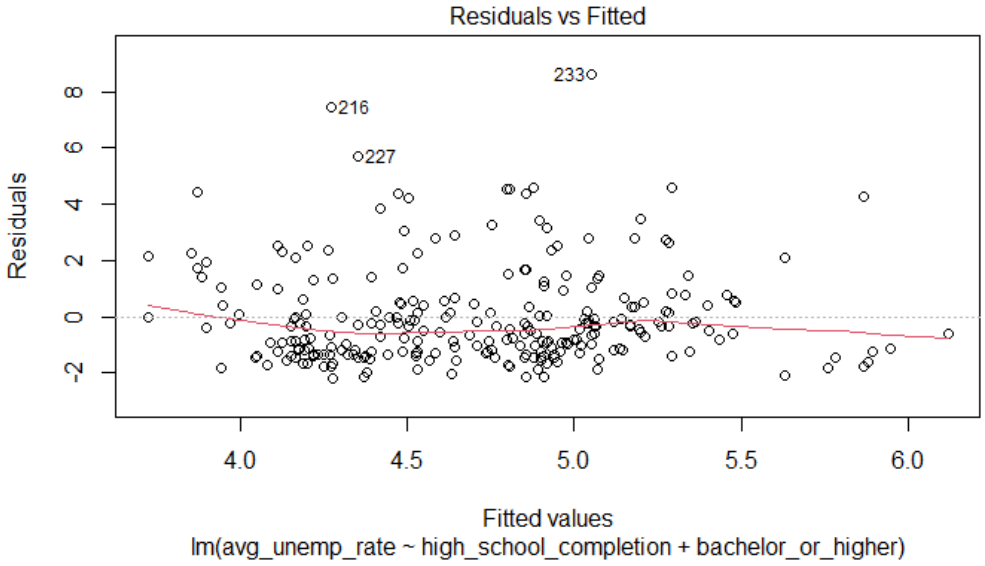
Figure 10. OLS regression summary: coefficients and model statistics.

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

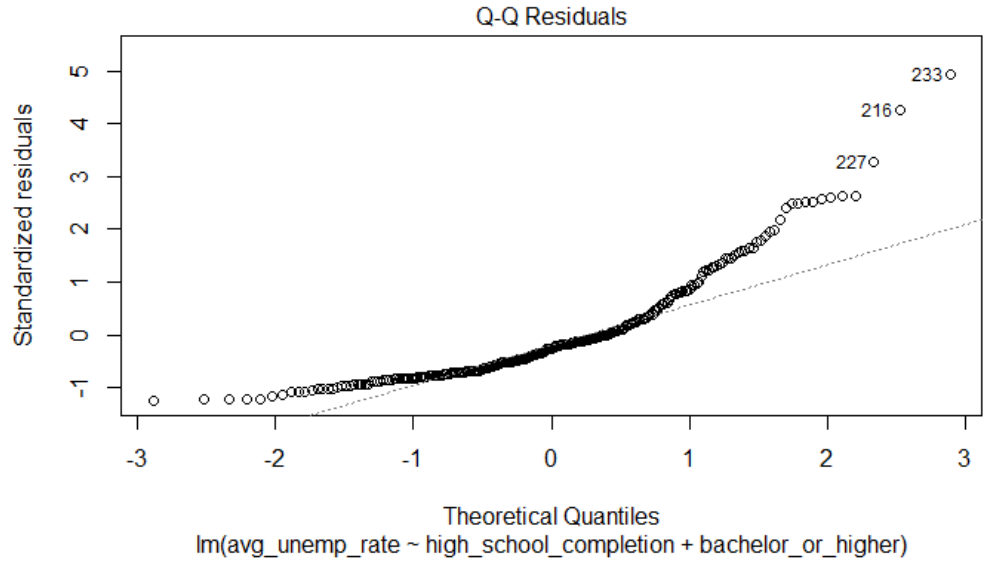
- **R-squared = 0.067**,
- **Adjusted R-squared = 0.059**,
  indicating that only about 6% of the variance in unemployment rate is explained by the two educational variables.

Residual Diagnostics

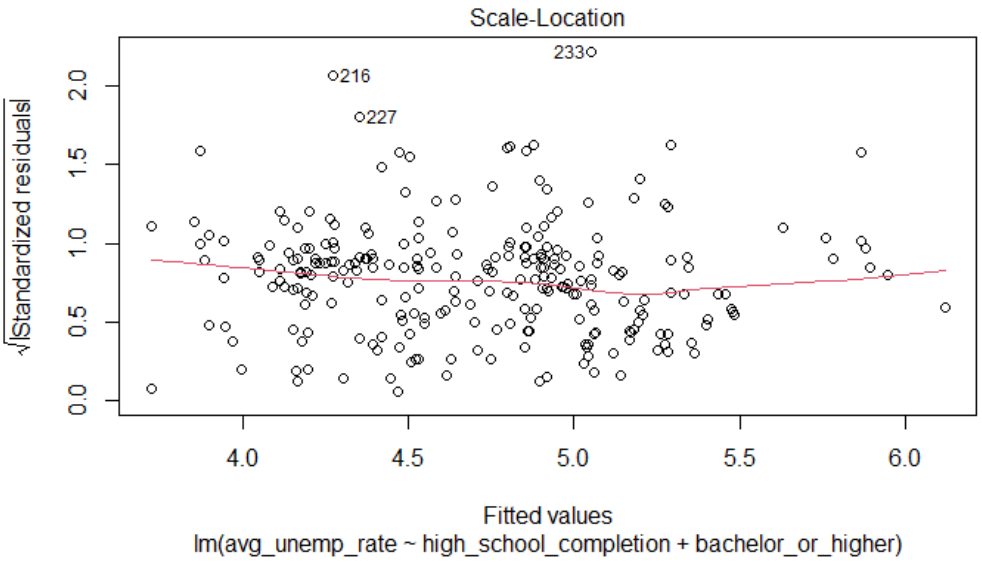
A full set of diagnostic plots is shown in **Figures 11–14**:



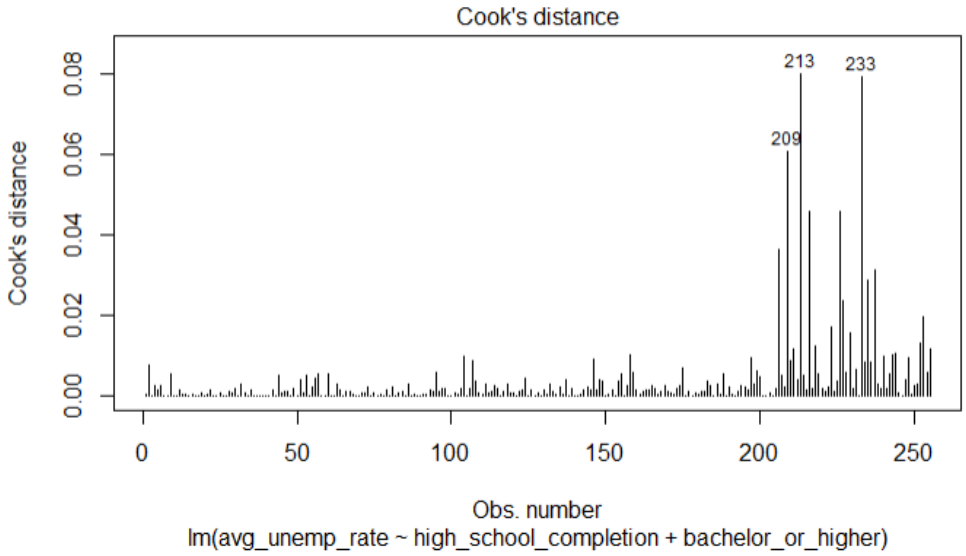
- **Figure 11: Residuals vs. Fitted** — Points are relatively homoscedastic, but some curvature is visible, suggesting minor model misspecification.



- **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 findings. First, we observed a clear negative relationship between high school completion rates and state-level unemployment rates, consistent across five years of data. In contrast, the relationship between bachelor's degree attainment and unemployment was weaker and occasionally positive, particularly in states with highly urbanized labor markets.

Second, we found persistent regional disparities: states in the Northeast and West Coast generally showed higher educational attainment, while unemployment patterns were more complex and varied, particularly in the aftermath of the COVID-19 pandemic in

2020.

Finally, while education did explain some variance in unemployment outcomes, our regression model had a low R^2 (~6.6%), reinforcing the idea that education alone cannot fully explain employment differences.

Business or Practical Implications

These findings have important implications for policy and workforce development. While promoting higher education remains important, our results suggest that expanding basic education access (i.e., high school completion) may yield more immediate returns in reducing unemployment.

Furthermore, strategies to improve employment outcomes may need to go beyond education—for example, by investing in industry diversification, job matching programs, or regional economic support.

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

10. Appendix

A: Work Contributions

Fanhao Kong: Collected and processed education data from NCES; led exploratory data analysis (EDA) and data visualization.

Yuqing Dai: Retrieved and processed unemployment data from FRED; conducted regression modeling and diagnostics.

Both team members collaborated on data cleaning, report writing, and final presentation.

B. Data Cleaning Process

- **Education data**:
  - Read from NCES .xls files for each year (2016–2020)
  - Used read_excel() with skip = 5, extracted columns manually
  - Renamed and standardized columns for state, high school %, and bachelor's %
  - Removed national aggregate rows (e.g., "United States")
  - Standardized state formatting (trimming, title case)
- **Unemployment data**:
  - Retrieved monthly unemployment from FRED API using fredr
  - Constructed state_map of series IDs manually
  - Calculated yearly averages per state
  - Matched to education data by state and year
- **Merging**:
  - Used left_join() on state and year
  - Verified matches for all state-year combinations
  - Removed unmatched NA rows (e.g., Puerto Rico)

C. Code by Section

Section 4 (Wrangling)

```
# PART 1: Read education data for each year and combine
years <- 2016:2020

edu_df <- map_dfr(years, function(yr) {
  path <- paste0("NCES", yr, ".xls")

  read_excel(path, skip = 5, col_names = FALSE) %>%
    select(
      state                          = ...1,     # state name
```

```
      high_school_completion = ...2,     # high school completion rate Total (%)
      bachelor_or_higher     = ...20     # bachelor's degree or higher Total (%)
    ) %>%
    filter(
      !is.na(high_school_completion),
      !is.na(bachelor_or_higher)
    ) %>%
    mutate(
      state                  = str_trim(state),
      high_school_completion = as.numeric(high_school_completion),
      bachelor_or_higher     = as.numeric(bachelor_or_higher),
      year                   = yr
    )
})

# Check the structure
glimpse(edu_df)
# PART 2: Clean and merge education + unemployment data

# 2.1 Clean education data: remove trailing whitespace/punctuation, standardize case
edu_clean <- edu_df %>%
  mutate(
    state = str_remove(state, "[\\s[:punct:]]+$"),
    state = str_to_title(state)
  ) %>%
  filter(!state %in% c("1", "United States"))

# 2.2 Clean unemployment data: same trimming
unemp_clean <- yearly_unemp %>%
  mutate(
    state = str_remove(state, "[\\s[:punct:]]+$"),
    state = str_to_title(state)
  )

# 2.3 Verify that all states for 2016–17 match
```

```
edu_clean %>%
  filter(year %in% 2016:2017) %>%
  anti_join(unemp_clean, by = c("state","year")) %>%
  distinct(state, year) %>%
  print(n = Inf)
# If no output appears, all states for those years have matched.

# 2.4 Perform the full left join and drop any missing values
merged_df <- left_join(edu_clean, unemp_clean, by = c("state","year")) %>%
  filter(!is.na(avg_unemp_rate))

print(merged_df)
```

Section 5.1 (Scatterplots)

```
# PART 3: Scatter plots (Education vs. Unemployment)

# High school completion vs. unemployment rate
ggplot(merged_df, aes(x = high_school_completion, y = avg_unemp_rate)) +
  geom_point(color = "steelblue", size = 2) +
  geom_smooth(method = "lm", se = TRUE, color = "darkred") +
  labs(title = "High School Completion vs Unemployment Rate ",
       x     = "High School Completion Rate (%)",
       y     = "Unemployment Rate (%)") +
  theme_minimal()

# Bachelor's degree rate vs. unemployment rate
ggplot(merged_df, aes(x = bachelor_or_higher, y = avg_unemp_rate)) +
  geom_point(color = "darkgreen", size = 2) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "Bachelor's Degree or Higher vs Unemployment Rate",
       x     = "Bachelor's Degree or Higher (%)",
       y     = "Unemployment Rate (%)") +
  theme_minimal()

# High school vs. bachelor's degree relationship
```

```
ggplot(merged_df, aes(x = high_school_completion, y = bachelor_or_higher)) +
  geom_point(color = "purple", size = 2) +
  geom_smooth(method = "lm", se = TRUE, color = "orange") +
  labs(title = "High School vs Bachelor's Completion Rates",
       x     = "High School Completion Rate (%)",
       y     = "Bachelor's Degree or Higher (%)") +
  theme_minimal()
```

Section 5.2 (Trends)

```
# PART 4: Trend of average unemployment rate over years
yearly_trend <- merged_df %>%
  group_by(year) %>%
  summarise(
    avg_unemp_rate = mean(avg_unemp_rate, na.rm = TRUE)
  )

# Plot line chart
ggplot(yearly_trend, aes(x = year, y = avg_unemp_rate)) +
  geom_line() +
  geom_point() +
  labs(
    title = "Average Unemployment Rate by Year (2016–2020)",
    x     = "Year",
    y     = "Average Unemployment Rate (%)"
  ) +
  scale_x_continuous(breaks = 2016:2020) +
  theme_minimal()
```

Section 5.3 (Top/Bottom States)

```
# PART 5: Bar charts – Top/Bottom 10 States

# Top 10 by 5-year avg bachelor's rate
avg_bach_states <- merged_df %>%
  group_by(state) %>%
  summarise(mean_bach = mean(bachelor_or_higher, na.rm = TRUE)) %>%
```

```
  slice_max(mean_bach, n = 10)

ggplot(avg_bach_states, aes(
    x = fct_reorder(state, mean_bach),
    y = mean_bach
  )) +
  geom_col(fill = "steelblue", width = 0.7) +
  coord_flip() +
  scale_y_continuous(
    labels = percent_format(scale = 1),
    expand = expansion(mult = c(0, 0.05))
  ) +
  labs(
    title = "Top 10 States by 5-Year Avg Bachelor's Rate (2016–2020)",
    x     = NULL,
    y     = "Avg Bachelor's Degree or Higher (%)"
  ) +
  theme_minimal(base_size = 12)
# Top 10 by 5-year avg unemployment rate
avg_unemp_states <- merged_df %>%
  group_by(state) %>%
  summarise(mean_unemp = mean(avg_unemp_rate, na.rm = TRUE)) %>%
  slice_max(mean_unemp, n = 10)

ggplot(avg_unemp_states, aes(
    x = fct_reorder(state, mean_unemp),
    y = mean_unemp
  )) +
  geom_col(fill = "firebrick", width = 0.7) +
  coord_flip() +
  scale_y_continuous(
    labels = percent_format(scale = 1),
    expand = expansion(mult = c(0, 0.05))
  ) +
  labs(
```

```
    title = "Top 10 States by 5-Year Avg Unemployment Rate (2016–2020)",
    x     = NULL,
    y     = "Avg Unemployment Rate (%)"
  ) +
  theme_minimal(base_size = 12)
```

Section 5.4 (Maps)

```
# PART 6: Choropleth maps for each year

# 1. Prepare data in long format for bachelor's degree rate
df_bach_long <- merged_df %>%
  filter(year %in% 2016:2020) %>%
  transmute(
    region = tolower(state),
    year,
    value  = bachelor_or_higher
  )

# 2. Retrieve state boundary data
states_map <- map_data("state")

# 3. Merge map data with education data
map_df <- left_join(states_map, df_bach_long, by = "region")

# 4. Plot all years in one faceted map
ggplot(map_df, aes(x = long, y = lat, group = group, fill = value)) +
  geom_polygon(color = "white", size = 0.2) +
  coord_fixed(1.3) +
  scale_fill_viridis_c(
    option    = "plasma",
    na.value  = "grey90",
    name      = "Bachelor's (%)"
  ) +
  facet_wrap(~ year, ncol = 3) +
  labs(
```

```
    title    = "Bachelor's Degree or Higher by State (2016–2020)",
    subtitle = "Each panel represents one year",
    caption  = "Source: NCES"
  ) +
  theme_void() +
  theme(
    strip.text    = element_text(face = "bold"),
    plot.title    = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5)
  )
# part 7: Map of average unemployment rate
df_unemp_long <- merged_df %>%
  filter(year %in% 2016:2020) %>%
  transmute(
    region = tolower(state),
    year,
    value  = avg_unemp_rate
  )

# 2. Get US state boundaries
states_map <- map_data("state")

# 3. Merge map data with unemployment data
map_unemp_df <- left_join(states_map, df_unemp_long, by = "region")

# 4. Plot all years in one faceted map
ggplot(map_unemp_df, aes(x = long, y = lat, group = group, fill = value)) +
  geom_polygon(color = "white", size = 0.2) +
  coord_fixed(1.3) +
  scale_fill_viridis_c(
    option   = "plasma",
    na.value = "grey90",
    name     = "Unemployment (%)"
  ) +
  facet_wrap(~ year, ncol = 3) +
```

```
  labs(
    title       = "Average Unemployment Rate by State (2016–2020)",
    subtitle = "Faceted by Year",
    caption    = "Source: FRED via fredr"
  ) +
  theme_void() +
  theme(
    strip.text       = element_text(face = "bold"),
    plot.title       = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5)
  )
```

Section 6 (Regression)

```
# PART 8: OLS regression and diagnostics
fit <- lm(avg_unemp_rate ~ high_school_completion + bachelor_or_higher,
          data = merged_df)

# Coefficients, R², and p-values
summary(fit)

# Residual plots, leverage points, etc.
plot(fit, which = 1:4)
```