Education and Inequality: A State-Level Analysis of the U.S. (1990–2000)

Carlijn Calori (2860228) Leah Delikát (2813155) Fadhil Dhafir (2847129)
Sten Groen (2813781) Julia Koeleman (xxx) Anne Schrama (2812834)
Marie-Louise Stevens (2814083) Sophia Zentgraf (2853242)

2025-06-24

Title Page

Carlijn Calori, Leah Delikát, Fadhil Dhafir, Sten Groen, Julia Koeleman, Anne Schrama, Marie-Louise Stevens & Sophia Zentgraf

Tutorial: Group 1

Tutor: J.F. Fitzgerald

1 Identification of the Social Problem

1.1 Describe the Social Problem

Income inequality is a serious social issue in the United States. Over the past few decades, the gap between the highest and lowest earners has grown a lot, leading to problems on social, economic, and political levels. This widening gap creates unequal opportunities, adds pressure to communities, and makes it harder for people to succeed in the job market. Education plays an important role in this situation. People with higher levels of education tend to earn more money than those with lower levels of education. But how straightforward is that relationship?

Several sources have pointed out that income inequality is a major concern. The Organisation for Economic Co-operation and Development (OECD) has stated in multiple reports that the U.S. has one of the highest levels of income inequality among developed countries (Denk et al., 2013). They point out that this can slow down economic growth and limit social mobility. Horowitz et al. (2020) from the Pew Research Center has also found that income gaps keep growing and that education is a key factor—people without higher education are falling further behind in the labor market. Basu and Stiglitz (2014) point out that individuals in the top income decile tend to have relatively high levels of education, whereas those in the bottom four deciles often have limited or poor-quality schooling. Denk et al. (2013) and Horowitz et al. (2020) show that income inequality, and the role education plays in it, is a long-term issue that needs real attention.

Even tough this topic has been researched before, this analysis takes a closer look than many past studies. We focus on differences between U.S. states to see how the link between education and income changes from one region to another. By doing this, we can offer a more detailed and complete picture of income inequality. With this more in-depth analysis, we hope to find new insights that haven't been explored in other research.

2 Data Sourcing

2.1 Description and limitations

The dataset "Inequality and Growth in the United States: Evidence from a New State-Level Panel of Income Inequality Measure" of the researcher Frank has been used for our project. This dataset provides historical coverage (1990-2000) of the highest income group in the US. Another reason why this data set has been used is because it specifically shows the top income groups per state. This way, the differences per state can be easily compared to one another. The dataset shows multiple variables that can measure inequality. The data is not from a government website, but from a researcher himself. However, the researcher did use data from the World Inequality Database to calculate the different top income groups. The limitation of this dataset is that it only provides the top income percentages, so there is not a possibility to compare the bottom 10 percent with the top 10, which makes it harder to accurately measure inequality.

The second dataset is "The Educational attainment of persons 25 years old and over, by race/ethnicity and state: April 1990 and April 2000". This dataset focuses on education level and has been retrieved from the National Center form Education Statistics (NCES). We chose this dataset because the level of education is a key indicator of inequality. The higher the education, the higher the income. Therefore big differences between education level will lead to a greater inequality. Furthermore, the dataset specifies the education level across all U.S. states. Which we can combine with the dataset of the top income percentages. Differences across racial and ethnic groups are also shown in the data, but will not be considered here, as they fall outside the scope of this analysis. The limitation of the dataset is that it includes data about only two points in time: April 1990 and April 2000. Unfortunately, there is no data available for the years in between.

2.2 Load in the data

To load the datasets, a link is provided to a Google Drive folder, from which a ZIP file containing the two datasets can be downloaded.

```
#Load the files
Distribution <- read_excel("datasets_income_inequality/Frank_WID_2020.xls", sheet = 3)
Education <- read_excel("datasets_income_inequality/tabn012.xls", col_types = c("text", "skip", "text",
```

```
## New names:
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
```

3 Quantifying

3.1 Data Cleaning

Only the years 1990 to 2000 are kept in the Distribution dataset. This is because the Education dataset includes data only for 1990 and 2000. To ensure the datasets can be merged and compared properly, just 1990, 2000, and the years in between are included.

In addition, two rows are removed, as they are not a U.S. state. One of the rows corresponds to the entire United States, while the other represents the District of Columbia.

```r
Distribution <- Distribution %>%
  filter(Year %in% 1990:2000,
         State != c("United States", "District of Columbia"))
```

For the Education dataset, unnecessary rows are removed. These rows contain no data and consist only of blank spaces between entries. Also, two rows representing the entire United States and the District of Columbia are removed.

To simplify the datatset, columns are renamed to reflect the type of data they contain, and any dots following state names are removed.

For the final step of cleaning, before merging the two datasets, the Education dataset needs to be converted to a long format. Currently, it contains four separate variables for the years 1990 and 2000, but these need to be combined into a single column with the years listed vertically. This transformation allows the dataset to be merged properly with the Distribution dataset, which is already in long format.

```r
Education <- Education[ -c(1:13, 14, 20, 24, 26, 32, 38, 44, 50, 56, 62, 68, 75:79), ]
```

```r
names(Education)[1:5] <- c("State", "Highschool1990", "Highschool2000", "Bachelor1990", "Bachelor2000")
```

```r
Education <- Education %>%
  mutate(State = State %>%
           str_replace_all("...+", "") %>%
           str_replace_all("\\.+", "") %>%
           str_trim()) # to delete all the dots after the state
```

```r
df1990 <- data.frame(State = Education$State, Year = 1990, Highschool = Education$Highschool1990, Bachel

df2000 <- data.frame(State = Education$State,Year = 2000,Highschool = Education$Highschool2000,Bachelor

Education <- rbind(df1990, df2000)

Education <- Education %>%
  mutate(across(3:4, ~ as.numeric(.)))
```

To merge the datasets, the full_join() function is applied. This function is appropriate because it performs an outer join, retaining all observations from both datasets based on the specific join keys (e.g., State and Year). Since the distribution dataset includes data for each year from 1990 to 2000, and the Education dataset only includes data for 1990 and 2000, full_join() ensures that no rows are lost during the merge. Missing values are introduced where data is not available in one of the datasets, allowing for a complete and alinged structure for subsequent analysis. After the merge, all numeric values are rounded to two decimals.

```r
Inequality <- Distribution %>%
  full_join(Education, by = c("State", "Year"))

Inequality <- Inequality %>%
  select(-st) %>%
  mutate(across(where(is.numeric), ~ round(.x, 2)))
```

3.2 Describing the type of variables in the datasets

```r
head(Inequality)
```

```
## # A tibble: 6 x 10
##    Year State      Top10_adj Top5_adj Top1_adj Top05_adj Top01_adj Top001_adj
##   <dbl> <chr>          <dbl>    <dbl>    <dbl>     <dbl>     <dbl>      <dbl>
## 1  1990 Alabama         39.9     26.0     12.4      9.1       4.74       1.78
## 2  1990 Alaska          34.0     22.7     11.2      8.74      4.93       2.15
## 3  1990 Arizona         38.9     26.5     12.3      9.25      4.56       1.64
## 4  1990 Arkansas        38.2     25.6     12.4      8.84      4.72       1.8
## 5  1990 California      42.8     30.6     16.6     12.5       6.82       2.84
## 6  1990 Colorado        36.4     25.4     12.0      9.16      4.74       1.82
## # i 2 more variables: Highschool <dbl>, Bachelor <dbl>
```

This is the merged dataset, named "Inequality". The first variable indicates the year (i.e., from 1990 to 2000), and the second column identifies the state from which the data originates. Each state appears eleven times in the dataset - once for every year.

The next five variables represent the percentage of total income earned by specific income groups. For example, the first of these columns shows that the top 10 percent of earners in Alabama in 1990 earned 39.9% of the total income. The following four columns follow the same structure, with each representing increasingly smaller income groups.

The final two columns capture educational attainment levels, expressed in percentages. For instance, in 1990, 66.9% of the population in Alabama had a high school diploma or higher and 15.7% held a bachelor's degree or higher.

3.2 Creating new variables

To measure income inequality, a new variable is constructed, modeled after the Palma Ratio. This ratio is defined as "the ratio of the income share of the top 10 percent over that of the bottom 40 percent" (Basu & Stiglitz, 2016, p. xxvi), and is designed to capture inequality at the extremes of the income distribution. However, the Distribution dataset does not include the income share of the bottom 40 percent. Therefore, as a practical and intuitive alternative, inequality is measured by dividing the income share of the top 10 percent by that of the bottom 90 percent (i.e., 100% minus the top 10%).

Using this measure, inequality will be visualized over time by presenting the mean and median values of all U.S. states. This visualization provides a broad overview of the national trend in inequality across the country, showing whether inequality is generally increasing or decreasing over time.

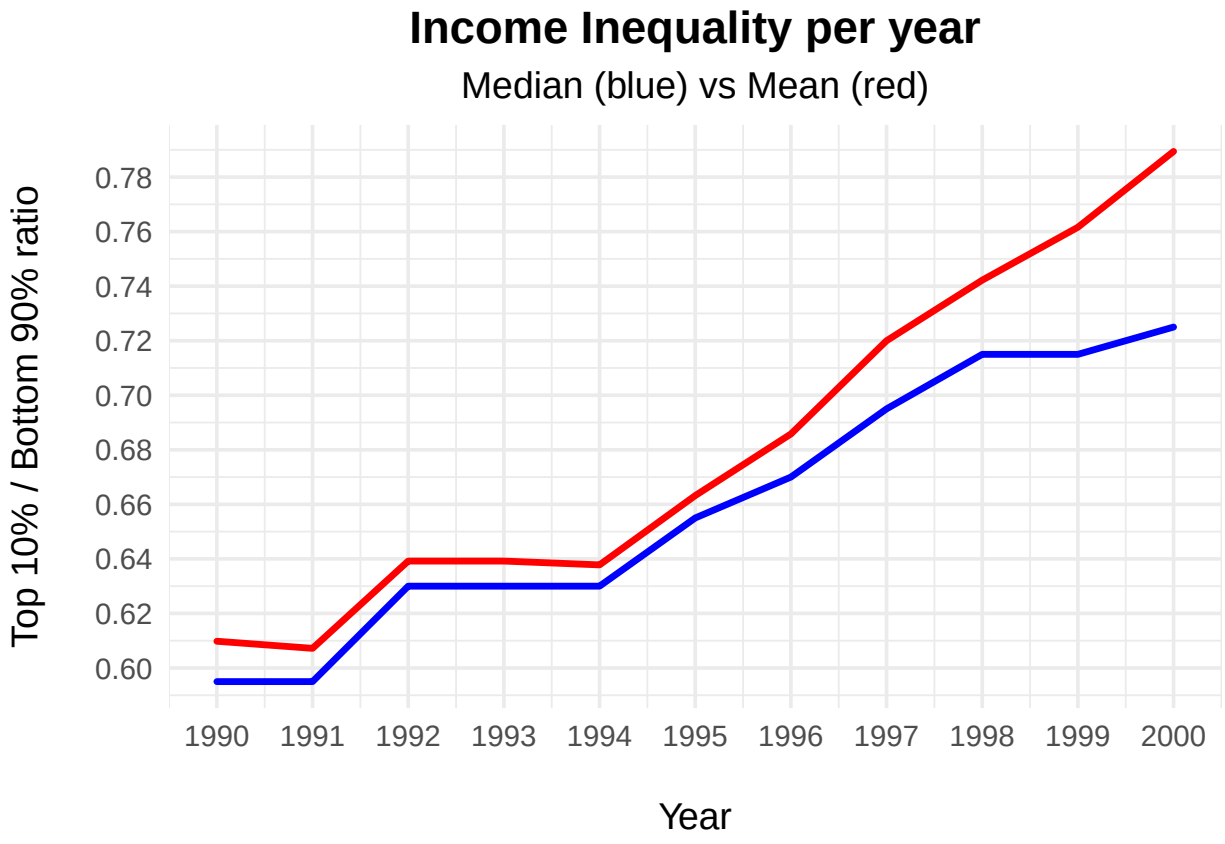
Additionally, a second new variable groups states into four categories - Low, Medium Low, Medium High, and High - based on the percentage of the population with a bachelor's degree or higher, allowing inequality to be examined across sub-populations. This approach helps reveal whether states with a higher share of highly educated residents tend to have a higher or lower inequality.

Finally, inequality will be visualized across the 50 states through a third new variable that shows the absolute change in inequality between 1990 and 2000. Not all states experience inequality changes equally; some show significant increases while others remain stable. This variation can be attributed to factors such as local institutions, tax policies, or demographic shifts (Wei, 2015; Arcabic et al., 2021).

```r
Inequality <- Inequality %>%
  mutate(Top10_vs_bottom90 = round(Top10_adj/(100-Top10_adj), 2))
```

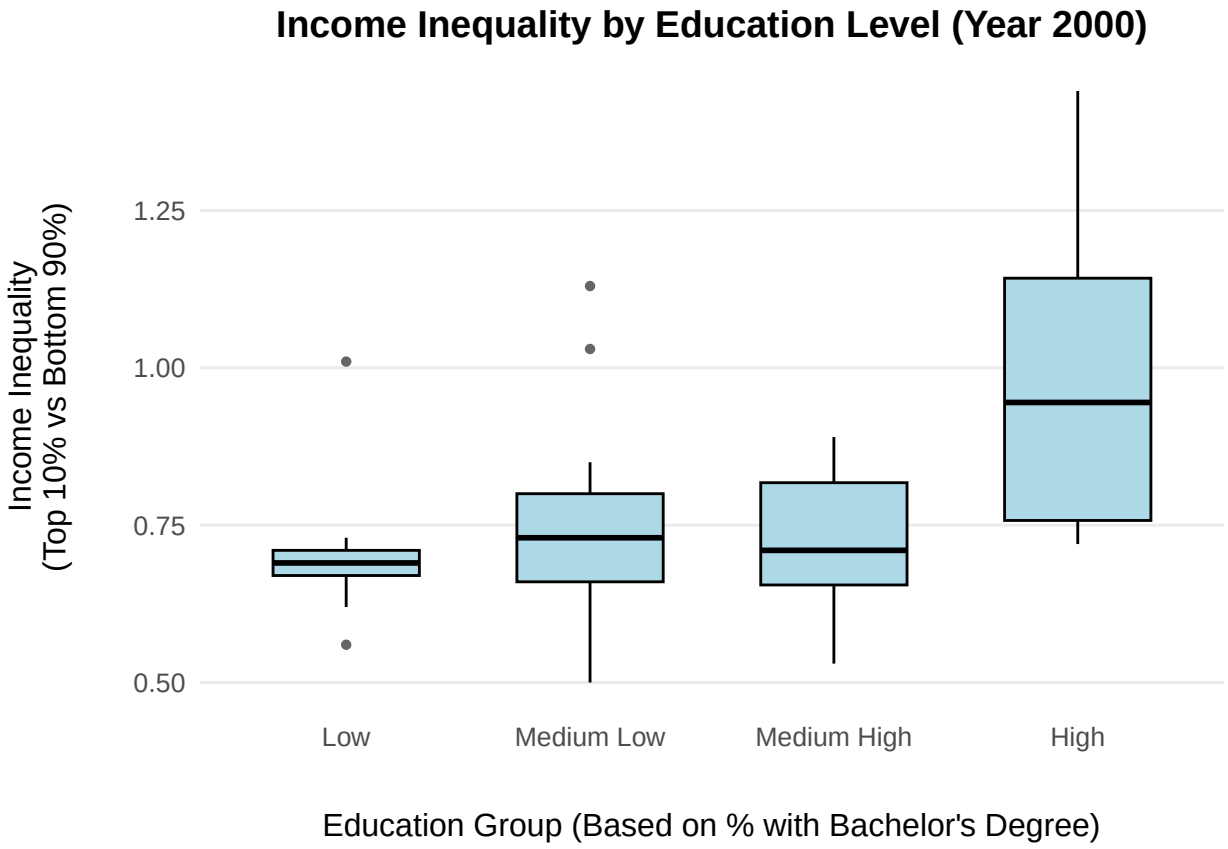
```r
Inequality <- Inequality %>%
  group_by(Year) %>%
  mutate(Quartile = ntile(Bachelor, 4), Bachelor_level = case_when(
    Quartile == 1 ~ "Low",
    Quartile == 2 ~ "Medium Low",
    Quartile == 3 ~ "Medium High",
    Quartile == 4 ~ "High"))

# Make sure Education is a factor with the right order
Inequality$Bachelor_level <- factor(
  Inequality$Bachelor_level,
  levels = c("Low", "Medium Low", "Medium High", "High"))
```

3.3 Temporal Variation



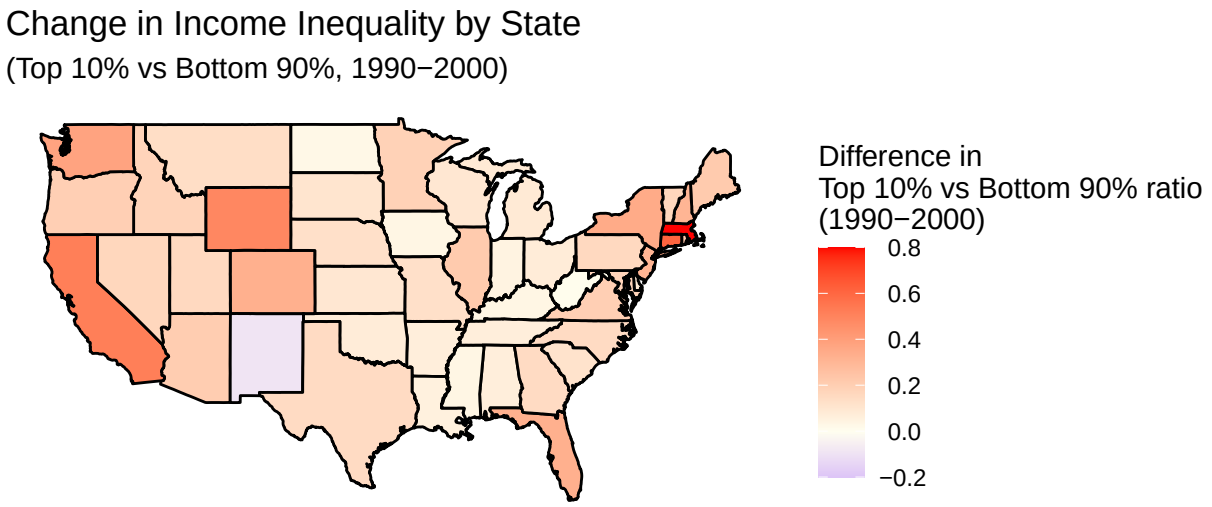
The graph illustrates the development of income inequality in the United States from 1990 to 2000, using the ratio between the income share of the top 10 percent and the bottom 90 percent. Both the median (blue) and mean (red) values show a clear upward trend over time, with a notable acceleration after 1994. The increasing gap between mean and median also suggest that the highest earners experienced disproportionate income growth compared to the general population.

3.4 Sub-population Variation



The box-and-whisker plot displays income inequality in the year 2000 across four groups of states, categorized by the percentage of the population with a bachelor's degree or higher. The results show a positive correlation: states with a higher level of educational attainment tend to exhibit greater income inequality. This suggests that while higher education leads to better-paying jobs, it also contributes to a wider income gap, as the earnings of highly educated individuals pull away from those with lower education levels.

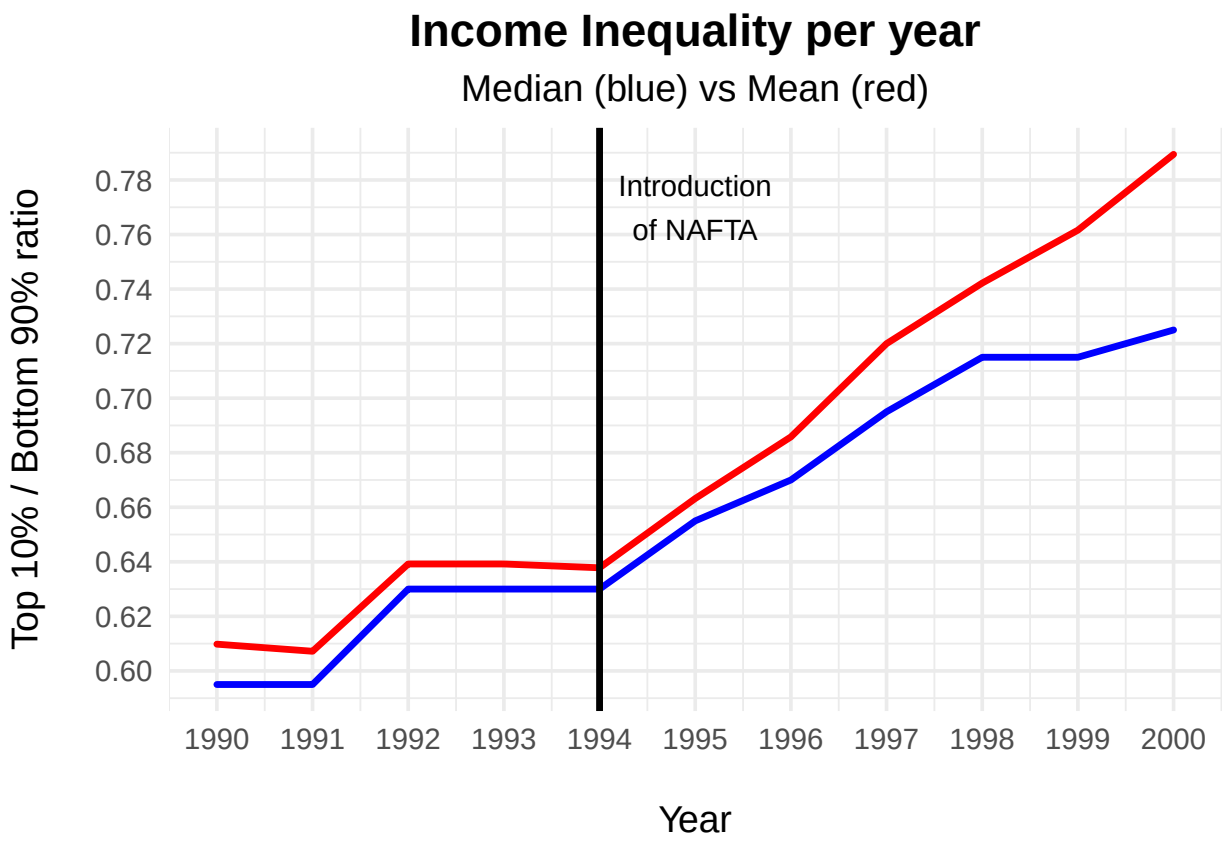
3.5 Spatial Variation



The spatial variation graph shows the change in income inequality between 1990 and 2000 across all U.S. states, measured by the Top10_vs_bottom90 ratio. Most states experienced an increase in inequality, though the magnitude of change varies. The most significant increases are observed in states such as Massachusetts, California, and Arizona, indicating that regional economic developments may have amplified income disparities. Conversely, several Midwestern and Southern states show smaller changes, and a few appear relatively stable or even slightly declining.

This map effectively highlights regional disparities, allowing for geographic patterns in inequality growth to become visible. However, the use of a continuous color gradient lacks categorical thresholds, making it harder to distinguish what constitutes a mild versus a substantial increase in inequality. For example, grouping values into low, medium, and high changes could make regional patterns more interpretable. Additionally adding regional labels or reference zones ("Northeast", "Southwest", etc) could further support spatial variation.

3.5 Event Analysis



The temporal visualization shows that inequality rose significantly after 1994, coinciding with the implementation of the North American Free Trade Agreement (NAFTA). NAFTA, established between the U.S., Mexico and Canada, came into effect in January 1994, aiming to boost trade and economic integration by removing trade barriers and import tariffs (Floyd, 2025). As a result, exports and imports became cheaper, and a significant share of U.S. manufacturing was relocated to Mexico due to lower labor costs (Scott, 2011).

While NAFTA established the world's largest free trade zone, the benefits were uneven. Higher-income groups gained the most, as corporate profits increased and primarily flowed to shareholders and high-income individuals (This trend explains the rise in the mean income, while the median grew more slowly).

At the same time, many low-skilled workers were negatively affected. The U.S. lost an estimated 682,900 manufacturing jobs by 2010, as outsourcing intensified (Scott, 2011). As some workers faced stagnating or declining incomes, those in high-skill sectors saw gains, widening the income gap.

However, inequality cannot be contributed to NAFTA alone. Technological change, labor market shifts, and globalization also contributed significantly (IMF, 2017). Therefore, while the timing suggests a potential link, the analysis does not confirm a causal relationship.

4 Discussion

4.1 Discussion of the findings

The analysis has shown that inequality in the US is structural growing and a more urgent social issue. By measuring inequality through the top 10% versus the bottom 90% ratio, we find a general increase

across nearly all U.S states. The most significant increase occurred in Massachusetts, California, Arizona, and Nebraska. These changes may be linked to regional economic factors in specific areas. For example California and Massachusetts had a rapid growth in tech and finance industries during the 1990s, where highly skilled workers (often in the top 10%) were paid significantly more than lower income groups.

The sharp increase in inequality after 1994 coincides with the implementation of NAFTA. NAFTA, as discussed earlier, led to job outsourcing that likely benefited higher earners while disadvantaging low-skilled workers.

The sub-population analysis, based on the share of people with a bachelor's degree or higher, showed that states who have a higher education level, often tend to have higher income inequality. Highly educated workers have more access to better-paying jobs, which further widens the income gap. These findings confirm previous research indicating that populations with higher levels of education are more unequal (Horowitz et al., 2020; Basu & Stiglitz, 2016). This means simply raising education levels may not reduce inequality. To close the gap, other measures, such as fair wages and better job opportunities for lower-educated workers, are also needed.

To conclude, our analysis has shown that income inequality is widespread in the United States. It is most prominent in states with higher levels of educational attainment and increased significantly following the implementation of NAFTA.

5 Reproducibility

5.1 Github Repository Link

5.2 Reference List

Arčabić, V., Kim, K. T., You, Y., & Lee, J. (2021). Century-long dynamics and convergence of income inequality among the US states. Economic Modelling, 101, 105526. https://doi.org/10.1016/j.econmod.2021.105526

Basu, K., & Stiglitz, J. E. (2016). Inequality and Growth: Patterns and Policy Volume II: Regions and Regularities. Palgrave Macmillan. https://doi.org/10.1057/9781137554598

Denk, O., Hagemann, R. P., Lenain, P., Somma, V. (2013). Inequality and Poverty in the United States: Public Policies for Inclusive Growth. In *OECD Economics Department Working Papers* (No. 1052). https://dx.doi.org/10.1787/5k46957cwv8q-en

Floyd, D. (2025). *How Did NAFTA Affect the Economies of Participating Countries?* Investopedia. https://www.investopedia.com/articles/economics/08/north-american-free-trade-agreement.asp#:~:text=Cons%20of%20NAFTA%20Many%20jobs%20shifted%20from,suppress%20wages%20and%20opportunities%20for%20several%20parties.

Horowitz, J., Igielnik, R., Kochhar, R. (2020). *Most Americans Say There Is Too Much Economic Inequality in the U.S., but Fewer Than Half Call It a Top Priority.* https://www.pewsocialtrends.org/wp-content/uploads/sites/3/2020/01/PSDT_01.09.20_economic-inequailty_FULL.pdf

International Monetary Fund. (2017). Fiscal monitor: tackling inequality. In *World Economic And Financial Surveys.*

Scott, R. E. (2011). Heading South: U.S.-Mexico trade and job displacement after NAFTA (Briefing Paper #308). Economic Policy Institute.

Wei, Y. D. (2015). Spatiality of regional inequality. Applied Geography, 61, 1–10. https://dx.doi.org/10.1016/j.apgeog.2015.03.013