Final Project

INFO 5371 - Spring 2023

Blair Yu, Xiaohan Wang

4/30/23

Overview

This project analyzes the relationship between the number of college degree holders and social inequality across states in the United States from 1980 to 2022. By examining the trends in higher education attainment, we aim to provide valuable insights into the evolving landscape of education and its potential implications for social inequality.

Our team's visualizations focus on three key aspects: the trends in college degree attainment by sex, the distribution of college degree holders across states, and the top 10 states with the most college degree holders over time. These visualizations offer a comprehensive view of how higher education has evolved in the United States and how it may relate to social inequality.

Understanding the dynamics of higher education and its associations with social inequality can help inform future research and discussions surrounding education policy, access to higher education, and potential interventions to promote social mobility and reduce disparities across different social groups and geographic locations.

Research Question

Our research question is: **"What is the relationship between the number of college degree holders and social inequality across states in the United States over time?"**

Target Population

The target population for this study consists of individuals living in the United States who have completed a college degree. By focusing on this population, we aim to explore the differences in the distribution of college degree holders across various subpopulations, including sex and states, over time.

Descriptive Goal

This research question has a descriptive goal, as it seeks to characterize the differences among subpopulations rather than making claims about the effect or influence of a given variable. While this research question does not directly assess the impact of such an intervention, examining the differences in distribution can provide valuable context for understanding the potential effectiveness of policies aimed at promoting equitable access to higher education.

Importance of This Research Question

The importance of this research question lies in its potential to provide insights into the educational landscape of the United States, particularly concerning higher education attainment among different subpopulations. Understanding these differences can inform discussions on the accessibility and inclusivity of higher education, provide context for interpreting broader trends in higher education and workforce development, and contribute to a more nuanced understanding of social inequality in the United States. By identifying differences in college degree attainment among various subpopulations, this study can help inform policy discussions and guide future research in this area.

Dataset Introduction

The dataset used for this project was obtained from the Integrated Public Use Microdata Series (IPUMS), a reputable source that provides access to high-quality, harmonized data from the U.S. Census Bureau and other sources. We chose this dataset because it offers comprehensive and consistent information on the variables of interest, including year, sex, state, and education level, across a significant time period (from 1980 to 2022).

The dataset contains the following variables:

- **`YEAR`:** The year of the data, ranging from 1980 to 2022.
- **`SEX`:** The sex of the individual, coded as 1 for male and 2 for female.
- **`STATEFIP`:** The state Federal Information Processing Standard (FIPS) code, ranging from 1 to 56.
- **`EDUC`:** The education level of the individual, coded using IPUMS' harmonized coding scheme.

These data are particularly suitable for our research question as they allow us to examine the distribution of college degree holders across various subpopulations over time, thereby shedding light on patterns of social inequality in the United States.

Sample Restrictions

To ensure the relevance and accuracy of our analysis, we applied the following sample restrictions:

1. We selected data from the years 1980 to 2022 to provide a comprehensive view of trends over time.
2. We included only two sex categories: male (coded as 1) and female (coded as 2). This decision was made to focus on the most common sex categories and to facilitate comparability of our findings.
3. We filtered the dataset for individuals with an education level above a college degree (EDUC > 111) and excluded cases with unknown or missing education data (EDUC < 999). This step was taken to focus on the population of interest: college degree holders.
4. We included only state codes ranging from 1 to 56 to focus on the United States' states and to ensure that our analysis remains consistent and comparable across the entire sample.

By applying these sample restrictions, we aimed to create a dataset that is both relevant and manageable, allowing us to conduct a meaningful analysis of the distribution of college degree holders across various subpopulations in the United States over time.

Load Packages and Dataset

```r
# Load packages
library(tidyverse)
library(haven)
library(viridis)
library(gganimate)

# Load dataset
data <- read.csv("cps_00010.csv")
```

Visualization 1: College Degree Attainment by Sex over Time

Purpose and Goal

The first visualization aims to illustrate the difference between the number of male and female college degree holders over time. By showcasing this difference, we can gain insights into the changing dynamics of gender and education across the years, allowing us to better understand the evolution of gender inequality in higher education.

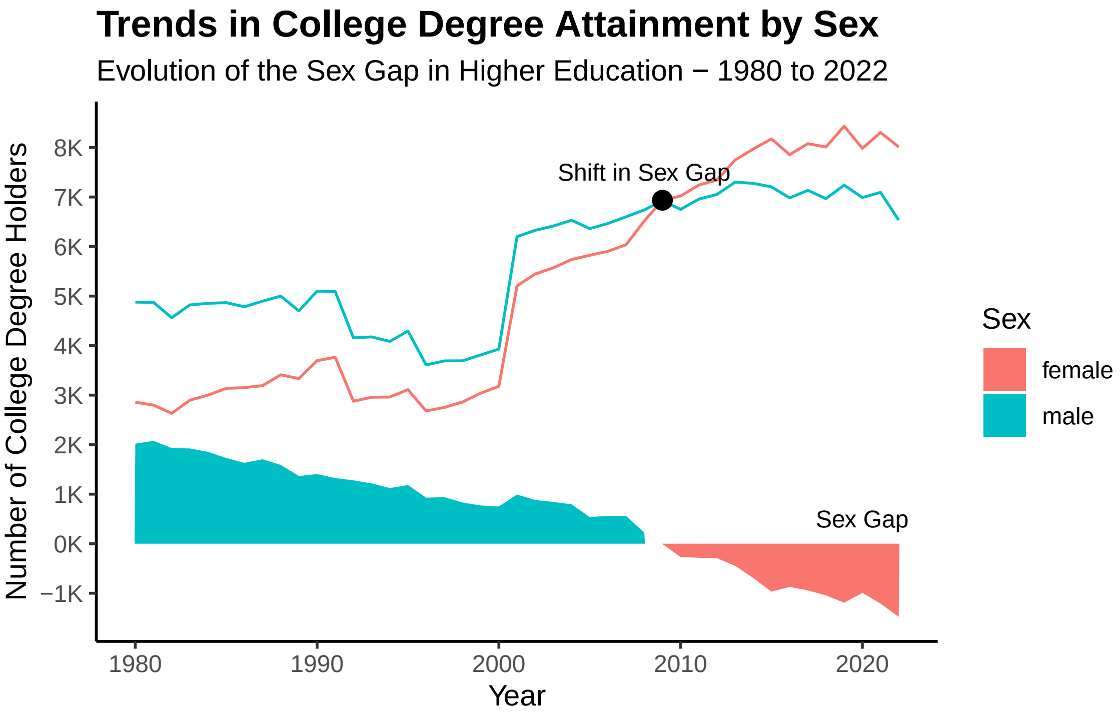
Approach

To create this visualization, we first filtered the dataset to include only college degree holders, and then grouped the data by year and sex. We calculated the total number of college degree holders for each group (males and females) and computed the gap between them for each year.

We used **`geom_line()`** to plot two lines showing the trend of male and female college degree holders over time. The lines are color-coded, with blue representing males and red representing females. Additionally, we used a **`geom_area()`** to visually highlight the gap between the number of male and female college degree holders. We also included **`geom_point()`** and annotations to emphasize the shift in the sex gap and its implications.

Figure

```r
# Group the data by year and sex, and calculate the total number of people
# having college degree for each group
college_degree_sum <- data %>%
  filter(EDUC > 111 & EDUC < 999) %>%
  mutate(sex = case_when(SEX == 1 ~ "male",
                         SEX == 2 ~ "female")) %>%
  group_by(YEAR, sex) %>%
  summarize(count = n())%>%
  rename(year = YEAR)

# Separate the data for males and females
male_college <- filter(college_degree_sum, sex == "male")
female_college <- filter(college_degree_sum, sex == "female")

# Calculate the difference between male and female for each year
sex_gap <- college_degree_sum %>%
  spread(sex, count) %>%
  mutate(gap = male - female)

# Plot a line chart showing the trend of male and female babies
ggplot() +
  geom_line(data = male_college,
            aes(x = year, y = count, color = "male"),
            show.legend = FALSE) +
  geom_line(data = female_college,
            aes(x = year, y = count, color = "female"),
            show.legend = FALSE) +
  geom_area(data = sex_gap,
            aes(x = year, y = gap, fill = ifelse(gap > 0, "male", "female"))) +
  geom_point(aes(x = 2009,
                 y = 6935),
             color = "black", size = 3) +
  annotate("text", x = 2020, y = 500, label = "Sex Gap", size = 3) +
  annotate("text", x = 2008, y = 7500, label = "Shift in Sex Gap", size = 3) +
  labs(title = "Trends in College Degree Attainment by Sex",
       subtitle = "Evolution of the Sex Gap in Higher Education - 1980 to 2022",
       x = "Year",
       y = "Number of College Degree Holders",
       fill = "Sex",
       caption = "Data Source: Integrated Public Use Microdata Series (IPUMS)")+
  scale_y_continuous(breaks = seq(-2000, 9000, by = 1000),
                     labels = function(x) paste0(x/1000, "K")) +
  theme_classic() +
  theme(plot.title = element_text(size = 14, face = "bold"))
```



Data Source: Integrated Public Use Microdata Series (IPUMS)

Interpretation of the Visualization

The visualization reveals interesting trends in college degree attainment by sex over time. From 1980 to 2009, the number of male college degree holders was consistently higher than that of female degree holders, but the gap gradually decreased. This trend can be attributed to the implementation of Title IX in 1972, which prohibits sex-based discrimination in educational programs and activities receiving federal financial assistance. Over the years, this policy has led to an increase in women's access to higher education.

Starting from 2009, the trend reversed, and the number of female college degree holders surpassed that of males. The gap between the two has been increasing ever since. One possible reason for this shift is the increasing demand for a college-educated workforce in sectors that have traditionally been more appealing to women, such as healthcare, education, and social services.

Between 1990 and 1997, the numbers of college degree holders for both sexes decreased, which could be a result of the economic recession during the early 1990s that affected both higher education funding and job prospects for college graduates. However, from 2000 to 2001, there was a significant increase in the number of degree holders for both sexes, which might be a result of the economic recovery, increased availability of financial aid, and a renewed focus on higher education as a pathway to better job opportunities.

The implications of these trends for social inequality are complex. On one hand, the closing and eventual reversal of the gender gap in higher education could be seen as a sign of progress towards greater gender equality. However, the growing dominance of women in higher education may lead to new forms of inequality, as some men may be left behind in terms of educational attainment. Policymakers and educational institutions must continue to monitor these trends and strive to create inclusive and equitable learning environments for all.

Visualization 2: College Degree Attainment by State

Purpose and Goal

The second visualization aims to provide a comprehensive view of the geographical distribution of college degree holders across the United States. By displaying this information in the form of a map, we aim to identify states with the highest and lowest numbers of individuals holding college degrees. This will allow us to better understand the regional differences in educational attainment and explore potential factors contributing to these disparities.

Approach

To create this visualization, we first grouped the data by state and calculated the total number of people with a college degree for each state. We then merged this data with a dataset containing the geographical boundaries of US states using the **`left_join`** function.

We used ggplot2 to generate the map, employing **`geom_polygon`** to create the polygons representing each state. The fill color of each state represents the number of college degree holders in that state, with a color scale (using the **`scale_fill_viridis`** function) corresponding to the magnitude of the count. We applied the **`coord_map()`** function to adjust the map projection for better visualization.

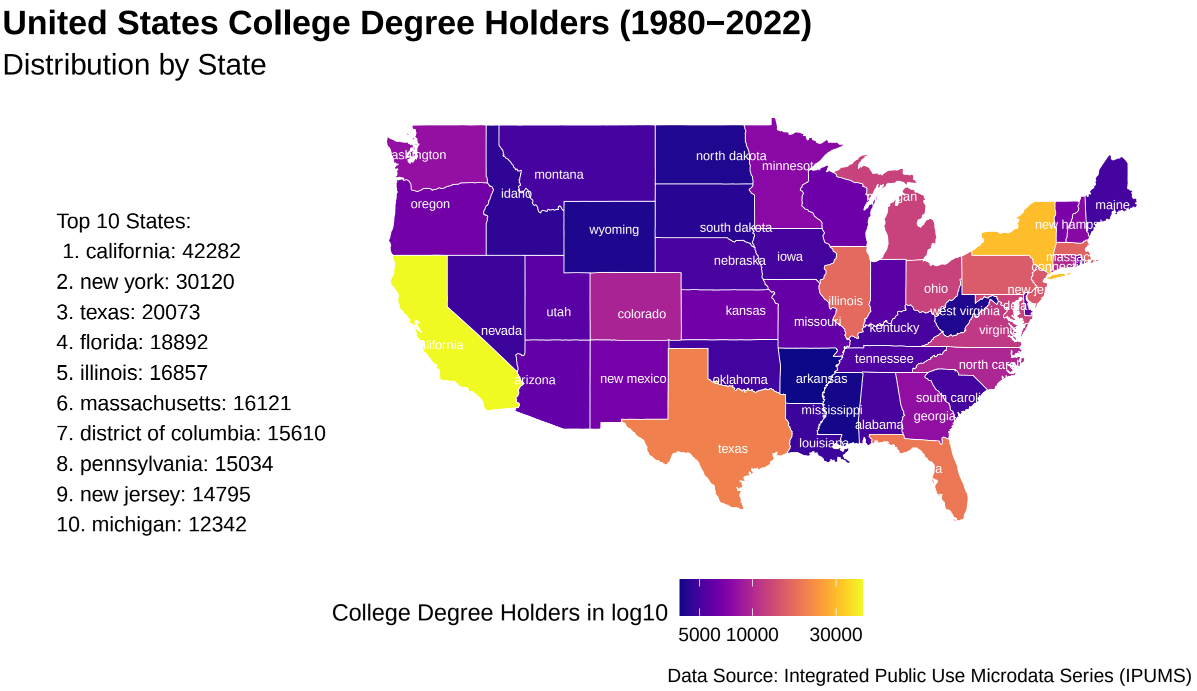
To add state labels, we used **`geom_text`** with the **`check_overlap`** parameter set to **`TRUE`** to ensure that labels do not overlap. For highlighting the top 10 states with the highest number of college degree holders, we first identified these states and their respective counts, then ranked them using the **`rank`** function. We combined the ranks and state names in a separate text box using the **`annotate`** function, and placed this box in a suitable location on the map.

The resulting map provides a clear visual representation of the geographical distribution of college degree holders, making it easy to identify regions with higher or lower levels of educational attainment.

Figure

```
# Data Cleaning
college_degree_state_sumorigi <- data %>%
  # Filter the data for college degree holders, exclude missing values
  filter(EDUC > 111 & EDUC < 999) %>%
  rename(year = YEAR) %>%
  # Change the state codes to be state names
  mutate(state = case_when(
    STATEFIP == 01 ~ "alabama",
    STATEFIP == 02 ~ "alaska",
    STATEFIP == 04 ~ "arizona",
    STATEFIP == 05 ~ "arkansas",
    STATEFIP == 06 ~ "california",
    STATEFIP == 08 ~ "colorado",
    STATEFIP == 09 ~ "connecticut",
    STATEFIP == 10 ~ "delaware",
    STATEFIP == 11 ~ "district of columbia",
    STATEFIP == 12 ~ "florida",
    STATEFIP == 13 ~ "georgia",
    STATEFIP == 15 ~ "hawaii",
    STATEFIP == 16 ~ "idaho",
    STATEFIP == 17 ~ "illinois",
    STATEFIP == 18 ~ "indiana",
    STATEFIP == 19 ~ "iowa",
    STATEFIP == 20 ~ "kansas",
    STATEFIP == 21 ~ "kentucky",
    STATEFIP == 22 ~ "louisiana",
    STATEFIP == 23 ~ "maine",
    STATEFIP == 24 ~ "maryland",
    STATEFIP == 25 ~ "massachusetts",
    STATEFIP == 26 ~ "michigan",
    STATEFIP == 27 ~ "minnesota",
    STATEFIP == 28 ~ "mississippi",
    STATEFIP == 29 ~ "missouri",
    STATEFIP == 30 ~ "montana",
    STATEFIP == 31 ~ "nebraska",
    STATEFIP == 32 ~ "nevada",
    STATEFIP == 33 ~ "new hampshire",
    STATEFIP == 34 ~ "new jersey",
    STATEFIP == 35 ~ "new mexico",
    STATEFIP == 36 ~ "new york",
    STATEFIP == 37 ~ "north carolina",
    STATEFIP == 38 ~ "north dakota",
    STATEFIP == 39 ~ "ohio",
    STATEFIP == 40 ~ "oklahoma",
    STATEFIP == 41 ~ "oregon",
    STATEFIP == 42 ~ "pennsylvania",
    STATEFIP == 44 ~ "rhode island",
    STATEFIP == 45 ~ "south carolina",
    STATEFIP == 46 ~ "south dakota",
    STATEFIP == 47 ~ "tennessee",
    STATEFIP == 48 ~ "texas",
    STATEFIP == 49 ~ "utah",
    STATEFIP == 50 ~ "vermont",
    STATEFIP == 51 ~ "virginia",
    STATEFIP == 53 ~ "washington",
    STATEFIP == 54 ~ "west virginia",
    STATEFIP == 55 ~ "wisconsin",
    STATEFIP == 56 ~ "wyoming",
    TRUE ~ "unknown"
  ))

college_degree_state_sum=college_degree_state_sumorigi%>%
  group_by(state) %>%
  summarize(count = n())

# Load state map data
states_map <- map_data("state")

# Calculate state centroids for Map
state_centroids <- states_map %>%
  group_by(region) %>%
  summarize(
    long = mean(long, na.rm = TRUE),
    lat = mean(lat, na.rm = TRUE))

# Merge the college degree data with the map data
state_centroids <- left_join(state_centroids,
                             college_degree_state_sum,
                             by = c( "region" = "state"))
```

```r
# Merge the college degree data with the map data
map <- left_join(college_degree_state_sum, states_map,  by = c("state" = "region"))

# Find top 10 states of the most college degree holders
top_10_states <- state_centroids %>%
  arrange(desc(count)) %>%
  head(10) %>%
  mutate(label = paste0(
    rank(-count, ties.method = "first"),
    ". ",
    region,
    ": ",
    round(count, 1)
  )) %>%
  pull(label) %>%
  paste(collapse = "\n")

# Create the map
ggplot(data = map) +
  geom_polygon(aes(x = long, y = lat, group = group, fill = count),
               color = "white",
               linewidth = 0.2) +
  coord_map() +
  scale_fill_viridis(option = "plasma",
                     name = "College Degree Holders in log10",
                     trans = "log10") +
  labs(title = "United States College Degree Holders (1980-2022)",
       subtitle = "Distribution by State",
       caption = "Data Source: Integrated Public Use Microdata Series (IPUMS)") +
  theme_minimal() +
  theme(panel.grid = element_blank(),
        axis.text = element_blank(),
        axis.title = element_blank(),
        axis.ticks = element_blank(),
        legend.position = "bottom",
        plot.title = element_text(size = 16, face = "bold"),
        plot.subtitle = element_text(size = 14))+
  # Add state names
  geom_text(
    data = state_centroids,
    aes(x=long, y=lat,label = region),
    size = 2,
    color = "white",
    hjust = 0.5,
    vjust = 0.5,
    check_overlap = TRUE
  ) +
  # Add the top 10 state list
  annotate(
    "text",
    x = -150,
    y = 35,
    label = paste("Top 10 States:\n", top_10_states),
    hjust = 0,
    size = 3.5
  )
```



United States College Degree Holders (1980–2022)

Distribution by State

Top 10 States:
1. california: 42282
2. new york: 30120
3. texas: 20073
4. florida: 18892
5. illinois: 16857
6. massachusetts: 16121
7. district of columbia: 15610
8. pennsylvania: 15034
9. new jersey: 14795
10. michigan: 12342

College Degree Holders in log10

Data Source: Integrated Public Use Microdata Series (IPUMS)

Interpretation

The visualization demonstrates a notable pattern in the distribution of college degree holders across the United States. In general, the East Coast appears to have a higher concentration of college degree holders, with California standing out as the state with the highest number.

This trend could be attributed to the presence of renowned educational institutions in these regions, as well as the availability of job opportunities for college graduates.

The top 10 states with the highest number of college degree holders are:

1. California: 42,282
2. New York: 30,120
3. Texas: 20,073
4. Florida: 18,892
5. Illinois: 16,857
6. Massachusetts: 16,121
7. District of Columbia: 15,610
8. Pennsylvania: 15,034
9. New Jersey: 14,795
10. Michigan: 12,342

It's worth noting that many of the top-ranking states are coastal states, which often have major urban centers and stronger economies. These factors may contribute to the higher number of college degree holders in those areas.

On the other hand, the majority of the states with lower numbers of college degree holders are found in the central and western parts of the country. One possible reason for this could be the presence of more rural areas and smaller populations, leading to fewer opportunities for higher education and lower demand for college-educated workers.

This visualization highlights the geographical disparities in educational attainment, which could have significant implications for social inequality. Regions with lower levels of education may face challenges in terms of economic development, access to resources, and overall quality of life. As a result, it is crucial for policymakers and stakeholders to consider these geographic differences when developing strategies to promote education and address social inequality.

Visualization 3: Changes of College Degree Attainments across Top 10 States over Time

Purpose and Goal

The third visualization is an animated visualization aims to illustrate the changes in the number of college degree holders over time for the top 10 US states. Clicking on the "play" button below the visualization can play the animation. This animation feature can work well with specific pdf editors (e.g. Adobe Acrobat).

The primary goal is to showcase the trends in higher education attainment in these states, providing valuable insights into how the landscape of education has evolved and potentially identifying factors contributing to their success in terms of educational achievement.

Approach

To create this visualization, we first identify the top 10 states with the highest number of college degree holders using the aggregated data. We then filter the original dataset to include only these states and group the data by year and state. This allows us to observe changes in the number of college degree holders for each of the top 10 states over time.

We use **`geom_bar()`** to generate a bar chart, with states on the x-axis (ordered by the number of college degree holders) and the number of college degree holders on the y-axis. The bars are filled with different colors to represent each state. We apply the **`coord_flip()`** function to create a horizontal bar chart for better readability.

To make the visualization more dynamic, we use the **`transition_reveal()`** function from the gganimate package to create an animation of the bar chart that reveals changes in the number of college degree holders over time. This helps in effectively illustrating the evolution of the top 10 states' educational attainment.

The resulting animated bar chart is an engaging visualization that highlights the progression of the top 10 US states in terms of the number of college degree holders over time.

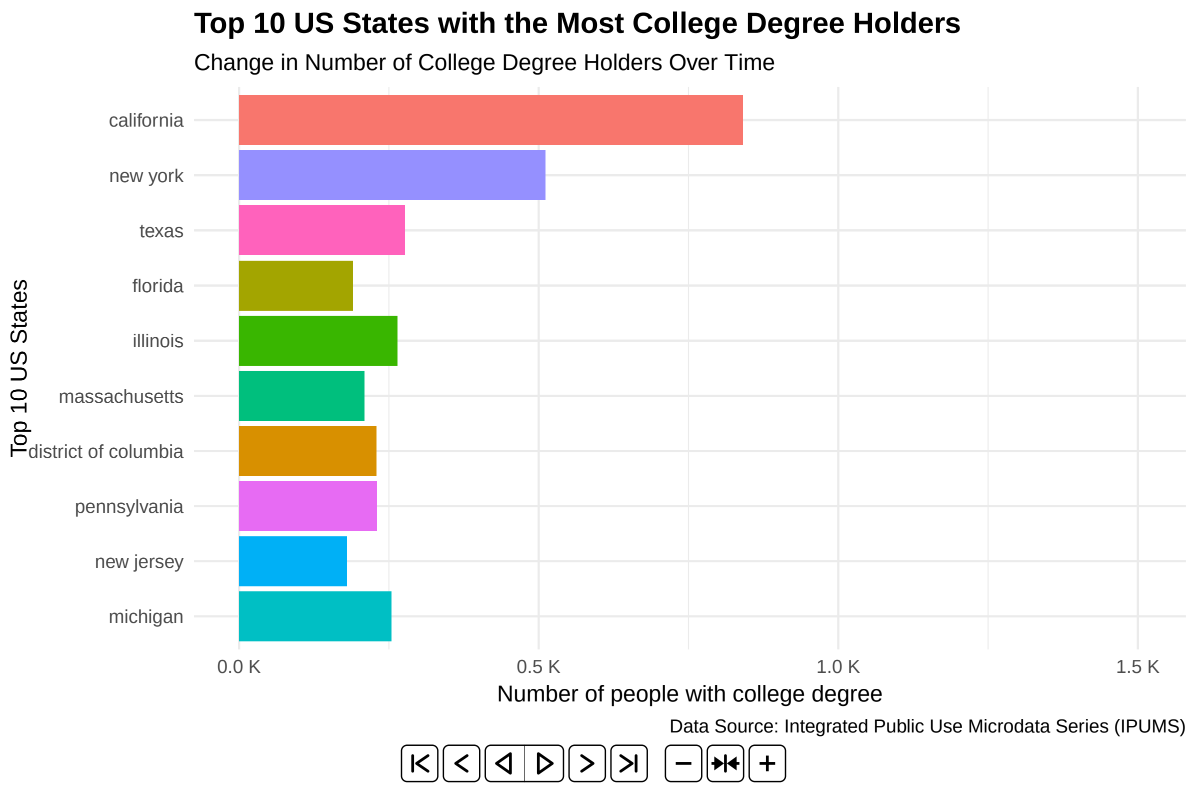
Figure

```r
# Data Cleaning
college_degree_state_new=college_degree_state_sumorigi %>%
  group_by(year,state) %>%
  summarize(count = n())
```

```
`summarise()` has grouped output by 'year'. You can override using the
`.groups` argument.
```

```r
# find the top 10 countries
top_10 <- college_degree_state_sum %>%
  arrange(desc(count)) %>%
  head(10)
```

```
# Subset the original data to include only the top 10 countries
data_sub <- college_degree_state_new %>%
  filter(state %in% top_10$state)

# Create the animation plot
ggplot(data = data_sub, aes(x = reorder(state, count), y = count, fill = state)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Top 10 US States with the Most College Degree Holders",
       subtitle = "Change in Number of College Degree Holders Over Time",
       x = "Top 10 US States",
       y = "Number of people with college degree",
       caption = "Data Source: Integrated Public Use Microdata Series (IPUMS)") +
  scale_y_continuous(labels = scales::unit_format(unit = "K", scale = 1e-3)) +
  theme_minimal()+
  theme(legend.position = "none",
        plot.title = element_text(size = 14, face = "bold"))+
  transition_reveal(year)
```


Top 10 US States with the Most College Degree Holders

Change in Number of College Degree Holders Over Time

Data Source: Integrated Public Use Microdata Series (IPUMS)

Interpretation

From the animated bar chart, we can observe several key trends in the top 10 US states with the most college degree holders.

In the early years, California and New York had significantly higher numbers of college degree holders compared to other states. This can be attributed to the large population in these states, as well as the presence of prestigious universities and robust higher education systems that attract and retain students.

For a period, the gap between the states appeared to narrow, potentially indicating a more equal distribution of educational attainment. However, this trend did not persist, as California experienced a rapid increase in the number of college degree holders, leaving even New York behind. One possible reason for this surge is the growth of the tech industry in California, which attracted highly educated individuals seeking job opportunities in the sector.

Other states, such as Texas, the District of Columbia, and Florida, also maintain relatively high numbers of college degree holders. These states may have benefited from their diverse economies, urbanization, and investment in education infrastructure, which contributed to their higher educational attainment levels.

The visualization highlights social inequality implications in terms of higher education attainment across the US states. It suggests that access to education and the number of college degree holders are not uniformly distributed across the country. Factors such as historical investment in education, the presence of prestigious institutions, and economic opportunities in specific industries may contribute to the disparities observed in the visualization.

Conclusion

In conclusion, our analysis of the data reveals a significant evolution in the gender gap in higher education over time. Women have surpassed men in terms of college degree attainment, which may be attributed to factors such as changing social norms, increased access to education for women, and their growing representation in various fields.

Furthermore, the data show considerable disparities in the distribution of college degree holders across the United States. States like California, New York, Texas, the District of Columbia, and Florida stand out with notably higher numbers of individuals with a college degree. This may be linked to historical investment in education, the presence of prestigious institutions, and economic opportunities in specific industries.

The visualizations underscore the need for policymakers to address the social inequalities in higher education distribution and promote equitable access to quality education across the country. Efforts to invest in education infrastructure, increase access in underprivileged areas, and foster economic development that attracts educated individuals to a wider range of states will be essential in creating a more balanced landscape for higher education in the United States. By understanding these trends and their underlying causes, we can work towards a more inclusive and equitable society that empowers all individuals to reach their full potential.