Project 2: US High School Graduation Rates

Serena Amro Gazze, Tatyanna Caputo, Evangeline Najarro, Elizabeth Techeira

```
# Load all necessary libraries

suppressWarnings(
  suppressPackageStartupMessages({
    library(tidyverse)
    library(ggplot2)
    library(dplyr)
    library(tidyr)
    library(readr)
    library(readxl)
    library(maps)
    library(ggthemes)
    library(usmap)
    library(ggbeeswarm)
    library(networkD3)
    library(htmlwidgets)
    library(htmltools)
    library(waffle)
  })
)
```

Original NCES Dataset: Notes and Issues

```
# Code arranged by Serena

# Data from National Center for Education Statistics (NCES)

# Load graduation dataset
graduation_nums <- read_excel("C:/Users/samro/Downloads/US High School Graduation Rates.xls")
```

```
## New names:
## • `` -> `...2`
## • `` -> `...3`
## • `` -> `...4`
## • `` -> `...5`
## • `` -> `...6`
## • `` -> `...7`
## • `` -> `...8`
## • `` -> `...9`
## • `` -> `...10`
## • `` -> `...11`
## • `` -> `...12`
## • `` -> `...13`
## • `` -> `...14`
## • `` -> `...15`
## • `` -> `...16`
## • `` -> `...17`
## • `` -> `...18`
## • `` -> `...19`
## • `` -> `...20`
## • `` -> `...21`
## • `` -> `...22`
## • `` -> `...23`
## • `` -> `...24`
## • `` -> `...25`
## • `` -> `...26`
## • `` -> `...27`
## • `` -> `...28`
```

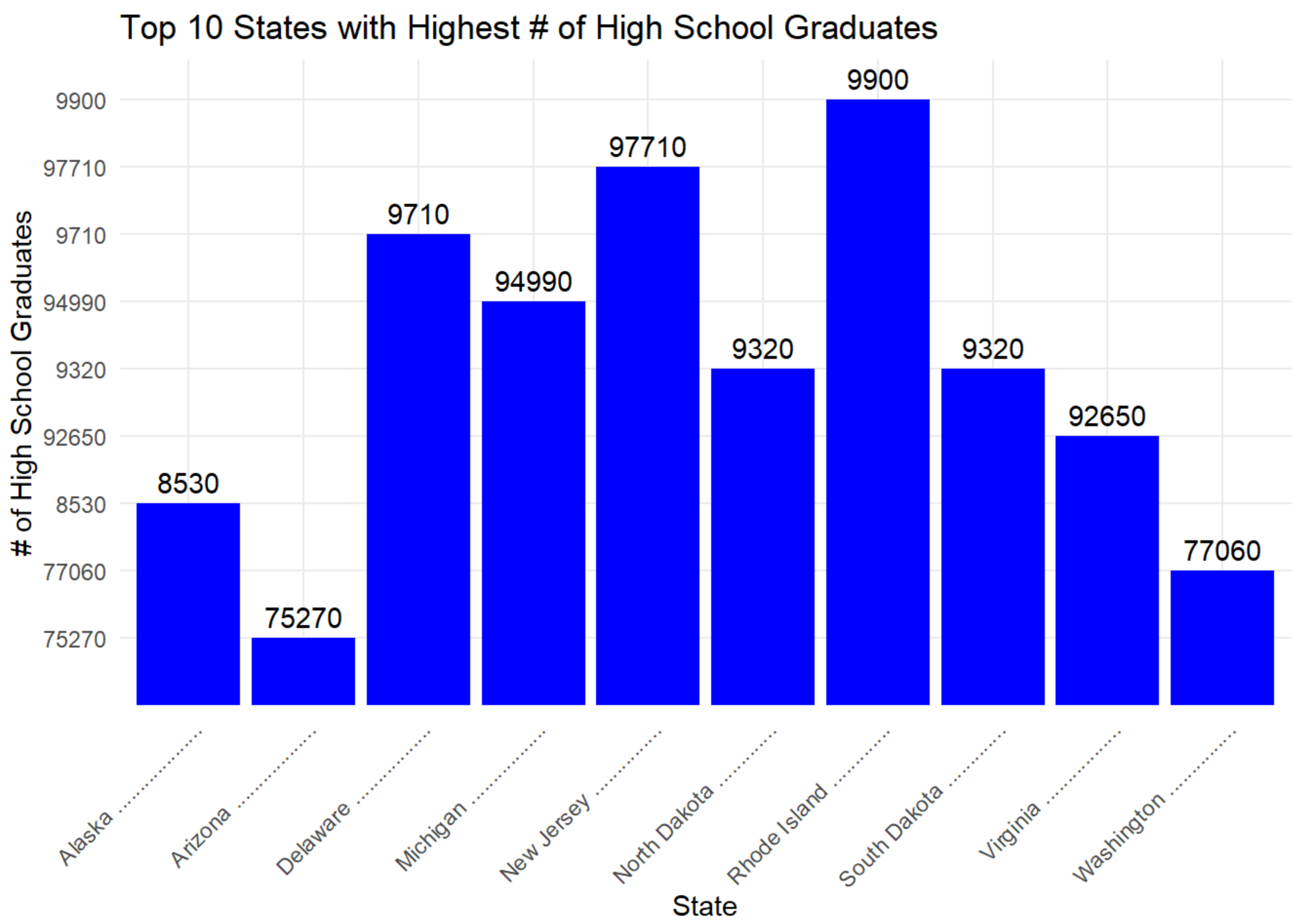
```
# Load dropouts dataset
dropout_nums <- read_excel("C:/Users/samro/Downloads/US High School Dropout Rates 1960 2022.xlsx")
```

```
## New names:
## • `` -> `...2`
## • `` -> `...3`
## • `` -> `...4`
## • `` -> `...5`
## • `` -> `...6`
## • `` -> `...7`
## • `` -> `...8`
## • `` -> `...9`
## • `` -> `...10`
## • `` -> `...11`
## • `` -> `...12`
## • `` -> `...13`
## • `` -> `...14`
## • `` -> `...15`
## • `` -> `...16`
## • `` -> `...17`
## • `` -> `...18`
## • `` -> `...19`
## • `` -> `...20`
## • `` -> `...21`
## • `` -> `...22`
## • `` -> `...23`
## • `` -> `...24`
## • `` -> `...25`
```

```
# Slice data frame to include only rows with states
states0 <- graduation_nums %>% slice(14:73)

# Make new data frame of top 10 states with highest graduation numbers
top_graduation_states <- states0 %>%
  arrange(desc(...25)) %>%
  select('Table 219.20. Public high school graduates, by region, state, and jurisdiction: Selected years, 1980-81 through 2026-27', ...25) %>%
  slice_head(n = 10)

head(top_graduation_states)
```

```
## # A tibble: 6 × 2
##   Table 219.20. Public high school graduates, by region, state, and juri…¹ ...25
##   <chr>                                                                    <chr>
## 1 Rhode Island .............                                               9900
## 2 New Jersey ...............                                               97710
## 3 Delaware .................                                               9710
## 4 Michigan .................                                               94990
## 5 North Dakota .............                                               9320
## 6 South Dakota .............                                               9320
## # ℹ abbreviated name:
## #   ¹`Table 219.20. Public high school graduates, by region, state, and jurisdiction: Selected years, 1980-81 through 2026-27`
```

```
# Make bar plot of top 10 states with highest graduation numbers
top_state_plot <- ggplot(top_graduation_states, aes(x = `Table 219.20. Public high school gradua
tes, by region, state, and jurisdiction: Selected years, 1980-81 through 2026-27`, y = ...25)) +
  geom_bar(stat = "identity", fill = "blue") +
  geom_text(aes(label = ...25), vjust = -0.5, color = "black") +
  labs(title = "Top 10 States with Highest # of High School Graduates",
       x = "State",
       y = "# of High School Graduates") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

top_state_plot
```



Notes and Issues

We found this dataset from the National Center for Education Statistics (NCES), which is an official governmental website. So, we thought it would be the perfect dataset to use.

However, the dataset had so many issues that (after trying to tidy/clean it for almost 2 hours) we decided to find a new, different dataset to use instead. The issues with our original dataset included the following: **the data wasn’t current data, it was projected data**; **the dataset didn’t include percentages, only raw numbers**; the column names were labeled in an unusable manner; their headers messed up the column numbering; and the dataset was in an outdated Excel file format.

Overall, we were able to make one plot using the original dataset, but we quickly decided we wouldn’t be able to use it for our whole project.

```
# Data from https://eddataexpress.ed.gov/download/data-library?field_year_target_id=All&field_po
pulation_value=&field_data_topic_target_id=All&field_reporting_level_target_id=All&field_program
_target_id=All&field_file_spec_target_id=All&field_data_group_id_target_id=All&combine=graduatio
n

# Load Graduations Dataset
graduations <- read_csv("C:/Users/samro/Downloads/SY2122_FS150_FS151_DG695_DG696_SEA.csv") %>%
  # Only load School Year, State, Value, and Subgroup columns
  select(`School Year`, State, Value, Subgroup)
```

```
## Rows: 701 Columns: 18
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): School Year, State, Data Group, Data Description, Value, Populatio...
## dbl  (1): Denominator
## lgl (10): NCES LEA ID, LEA, School, NCES SCH ID, Numerator, Characteristics,...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
  # Value = Graduation Percentage

# Remove % from Value column & Convert column from chr to dbl
graduations$Value <- as.numeric(gsub("%", "", graduations$Value))
```

```
## Warning: NAs introduced by coercion
```

```
head(graduations)
```

```
## # A tibble: 6 × 4
##   `School Year` State         Value Subgroup
##   <chr>         <chr>         <dbl> <chr>
## 1 2021-2022     UNITED STATES  86.6 All Students
## 2 2021-2022     UNITED STATES  73.9 American Indian/Alaska Native/Native Americ…
## 3 2021-2022     UNITED STATES  93.7 Asian/Pacific Islander
## 4 2021-2022     UNITED STATES  NA   Asian
## 5 2021-2022     UNITED STATES  NA   Native Hawaiian or Other Pacific Islander
## 6 2021-2022     UNITED STATES  81   Black (not Hispanic) African American
```

How do high school graduation rates vary by state?

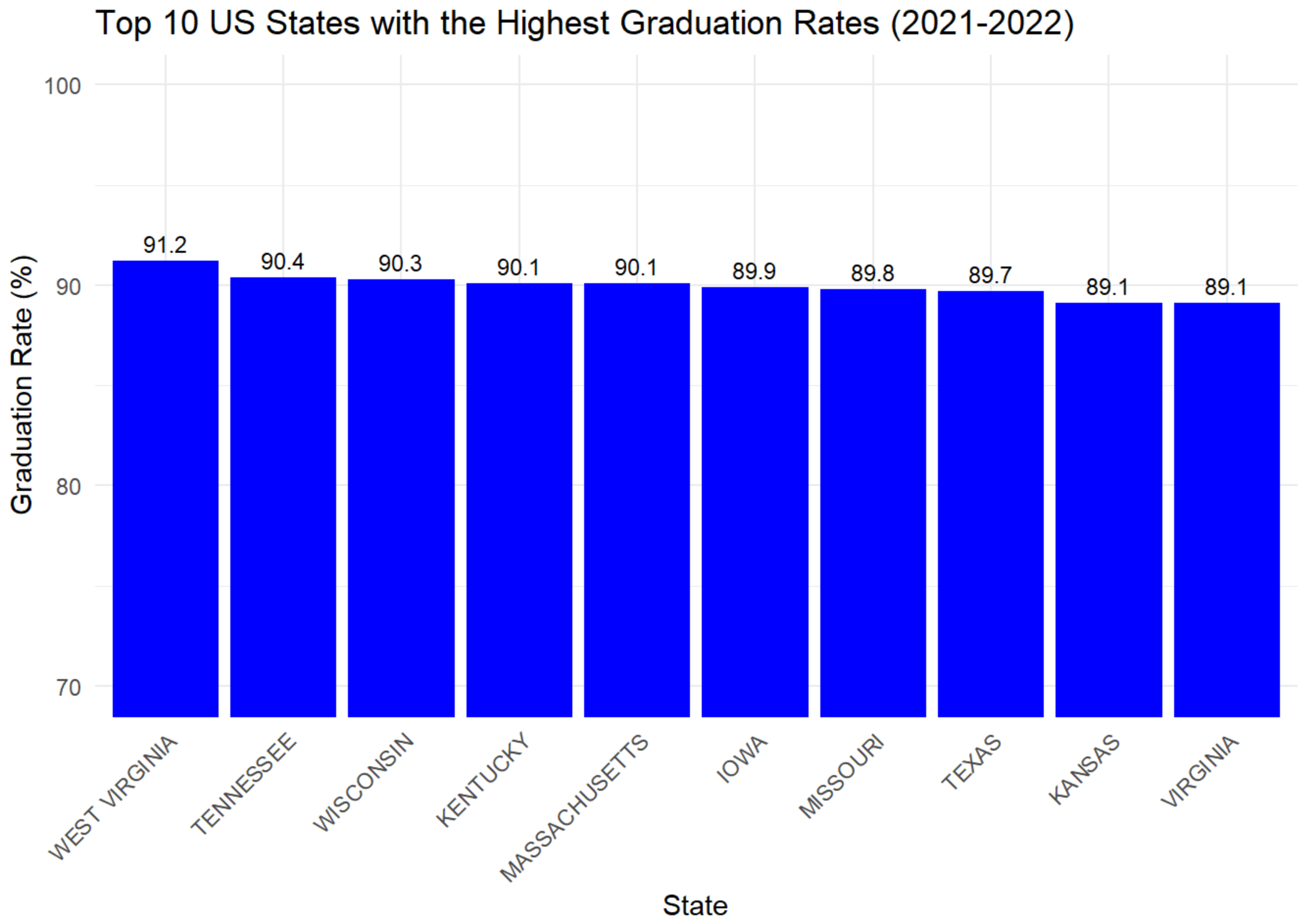
```
# Filter dataset for only the rows that include all students from each state
states <- graduations %>%
  filter(Subgroup == "All Students in SEA")

# Add row for New Mexico data
# Data from https://nmeducation.org/new-mexico-graduation-rates-show-slight-dip-in-2022/
new_mexico <- data.frame(`School Year` = "2021-2022", State = "NEW MEXICO", Value = 76.8, Subgro
up = "All Students in SEA")

colnames(new_mexico) <- colnames(states)

states <- rbind(states, new_mexico)

# Add row for Oklahoma data
# Data from https://wisevoter.com/state-rankings/high-school-graduation-rates-by-state/#:~:text=
Oklahoma%20and%20Wyoming%20both%20have,a%20graduation%20rate%20of%2082%25.
oklahoma <- data.frame(`School Year` = "2021-2022", State = "OKLAHOMA", Value = 81, Subgroup =
"All Students in SEA")

colnames(oklahoma) <- colnames(states)

states <- rbind(states, oklahoma)

# Remove non-state rows from data frame
states <- states[!states$State %in% c("PUERTO RICO", "DISTRICT OF COLUMBIA", "BUREAU OF INDIAN E
DUCATION"), ]

head(states)
```

```
## # A tibble: 6 × 4
##   `School Year` State      Value Subgroup
##   <chr>         <chr>      <dbl> <chr>
## 1 2021-2022     ALABAMA     88.2 All Students in SEA
## 2 2021-2022     ALASKA      77.8 All Students in SEA
## 3 2021-2022     ARIZONA     77.3 All Students in SEA
## 4 2021-2022     ARKANSAS    88.2 All Students in SEA
## 5 2021-2022     CALIFORNIA  87   All Students in SEA
## 6 2021-2022     COLORADO    82.3 All Students in SEA
```

```
# Code arranged by Serena

# Arrange states by Graduation Percentage in descending order, and then cut for only the top 10
states
top_states <- states %>%
  arrange(desc(Value)) %>%
  slice_head(n = 10)

top_states
```

```
## # A tibble: 10 × 4
##    `School Year` State         Value Subgroup
##    <chr>         <chr>         <dbl> <chr>
##  1 2021-2022     WEST VIRGINIA  91.2 All Students in SEA
##  2 2021-2022     TENNESSEE      90.4 All Students in SEA
##  3 2021-2022     WISCONSIN      90.3 All Students in SEA
##  4 2021-2022     KENTUCKY       90.1 All Students in SEA
##  5 2021-2022     MASSACHUSETTS  90.1 All Students in SEA
##  6 2021-2022     IOWA           89.9 All Students in SEA
##  7 2021-2022     MISSOURI       89.8 All Students in SEA
##  8 2021-2022     TEXAS          89.7 All Students in SEA
##  9 2021-2022     KANSAS         89.1 All Students in SEA
## 10 2021-2022     VIRGINIA       89.1 All Students in SEA
```

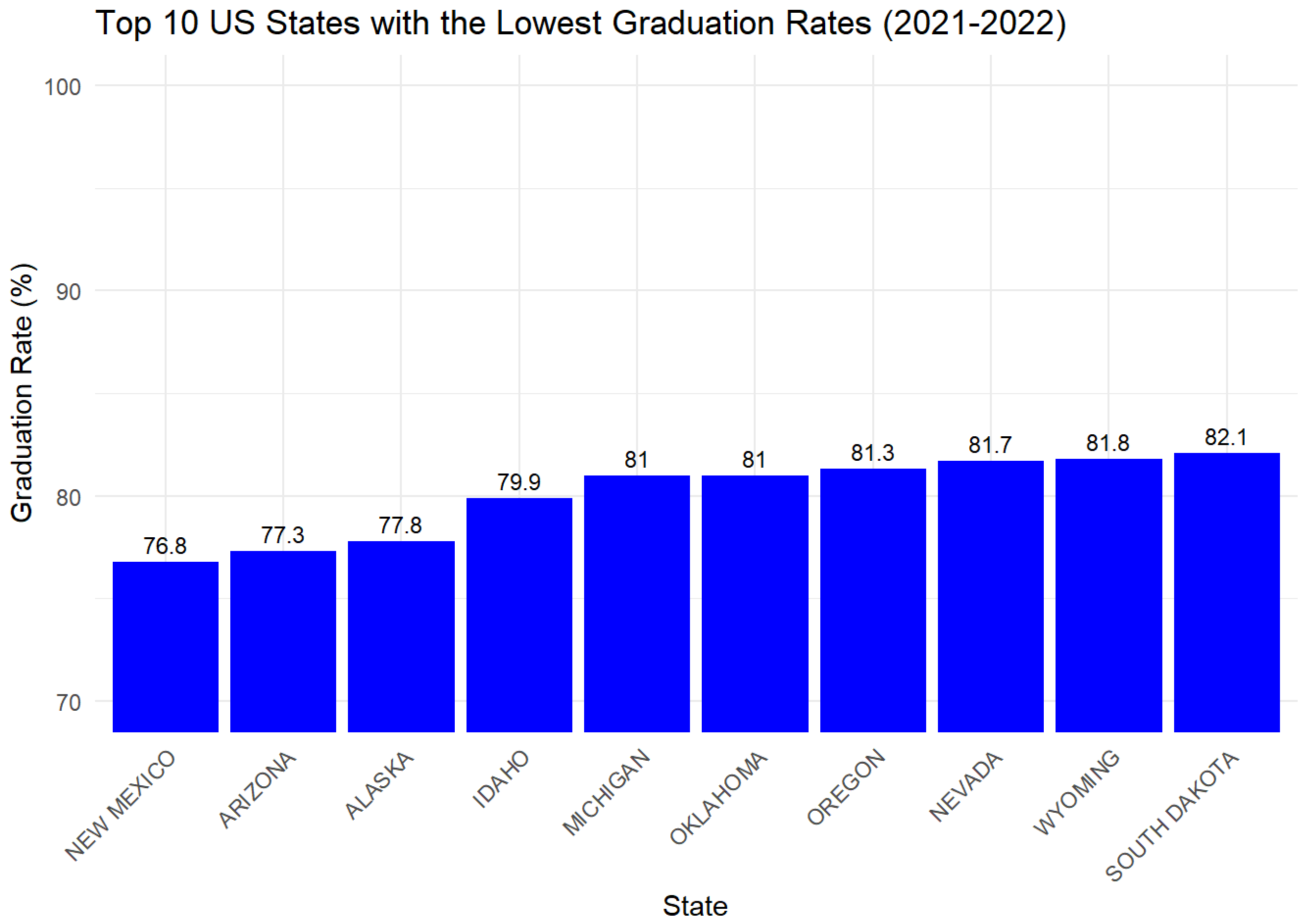
```r
# Plot bar graph of top 10 states with highest Graduation Percentage
top_plot <- ggplot(top_states, aes(x = reorder(State, -Value), y = Value)) +
  geom_bar(stat = "identity", fill = "blue") +
  geom_text(aes(label = Value), vjust = -0.5, size = 3) +
  labs(title = "Top 10 US States with the Highest Graduation Rates (2021-2022)",
       x = "State",
       y = "Graduation Rate (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_cartesian(ylim = c(70, 100))

top_plot
```



```
# Code arranged by Serena

# Arrange states by Graduation Percentage in ascending order, and then cut for only the top 10 s
tates
bottom_states <- states %>%
  arrange(Value) %>%
  slice_head(n = 10)

bottom_states
```

```
## # A tibble: 10 × 4
##    `School Year` State        Value Subgroup
##    <chr>         <chr>        <dbl> <chr>
##  1 2021-2022     NEW MEXICO    76.8 All Students in SEA
##  2 2021-2022     ARIZONA       77.3 All Students in SEA
##  3 2021-2022     ALASKA        77.8 All Students in SEA
##  4 2021-2022     IDAHO         79.9 All Students in SEA
##  5 2021-2022     MICHIGAN      81   All Students in SEA
##  6 2021-2022     OKLAHOMA      81   All Students in SEA
##  7 2021-2022     OREGON        81.3 All Students in SEA
##  8 2021-2022     NEVADA        81.7 All Students in SEA
##  9 2021-2022     WYOMING       81.8 All Students in SEA
## 10 2021-2022     SOUTH DAKOTA  82.1 All Students in SEA
```

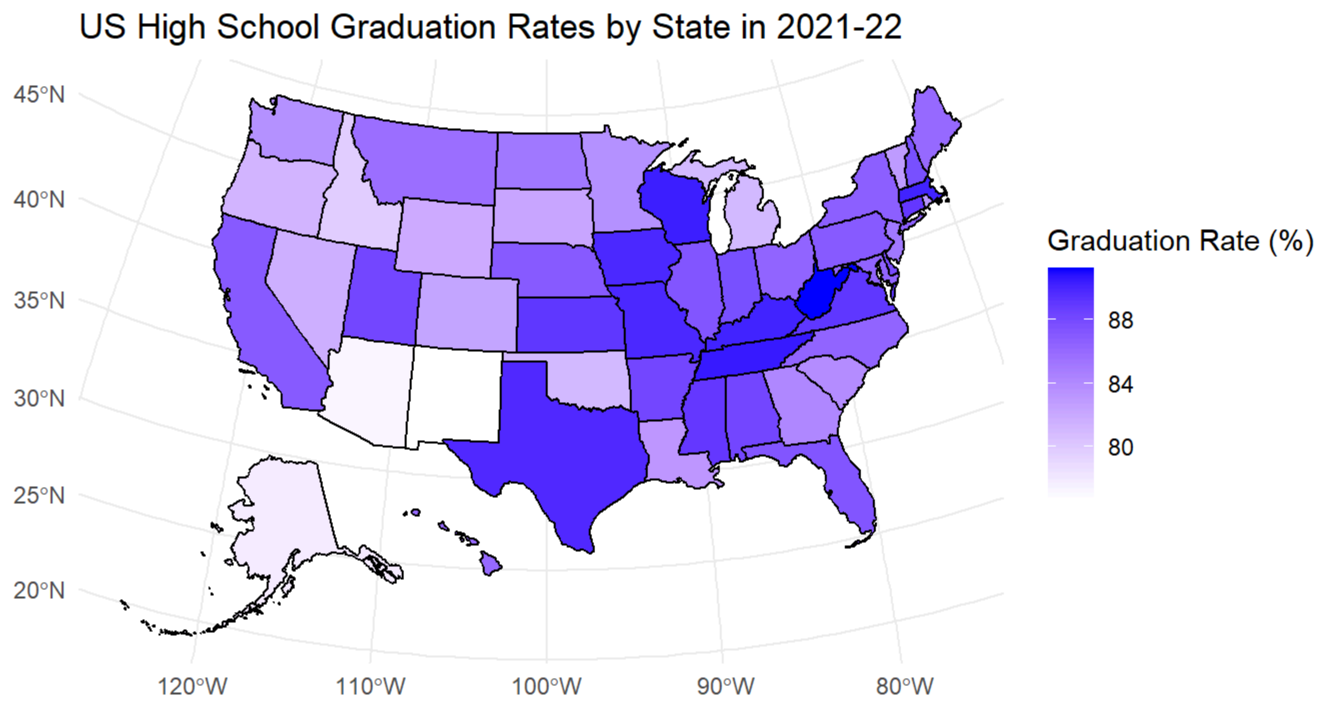
```
# Plot bar graph of top 10 states with lowest Graduation Percentage
bottom_plot <- ggplot(bottom_states, aes(x = reorder(State, Value), y = Value)) +
  geom_bar(stat = "identity", fill = "blue") +
  geom_text(aes(label = Value), vjust = -0.5, size = 3) +
  labs(title = "Top 10 US States with the Lowest Graduation Rates (2021-2022)",
       x = "State",
       y = "Graduation Rate (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_cartesian(ylim = c(70, 100))

bottom_plot
```



```
# Code arranged by Serena, assisted by Tatyanna

states1 <- states

# Change column name (from State to state) to work with plot_usmap
colnames(states1)[colnames(states1) == "State"] <- "state"

head(states1)
```

```
## # A tibble: 6 × 4
##   `School Year` state      Value Subgroup
##   <chr>         <chr>      <dbl> <chr>
## 1 2021-2022     ALABAMA     88.2 All Students in SEA
## 2 2021-2022     ALASKA      77.8 All Students in SEA
## 3 2021-2022     ARIZONA     77.3 All Students in SEA
## 4 2021-2022     ARKANSAS    88.2 All Students in SEA
## 5 2021-2022     CALIFORNIA  87   All Students in SEA
## 6 2021-2022     COLORADO    82.3 All Students in SEA
```

```
# Make heatmap of high school graduation rates by state
heatmap_plot <- plot_usmap(data = states1, values = "Value", regions = "states") +
  scale_fill_gradient(low = "white", high = "blue", name = "Graduation Rate (%)") +
  labs(title = "US High School Graduation Rates by State in 2021-22") +
  theme_minimal()

heatmap_plot
```

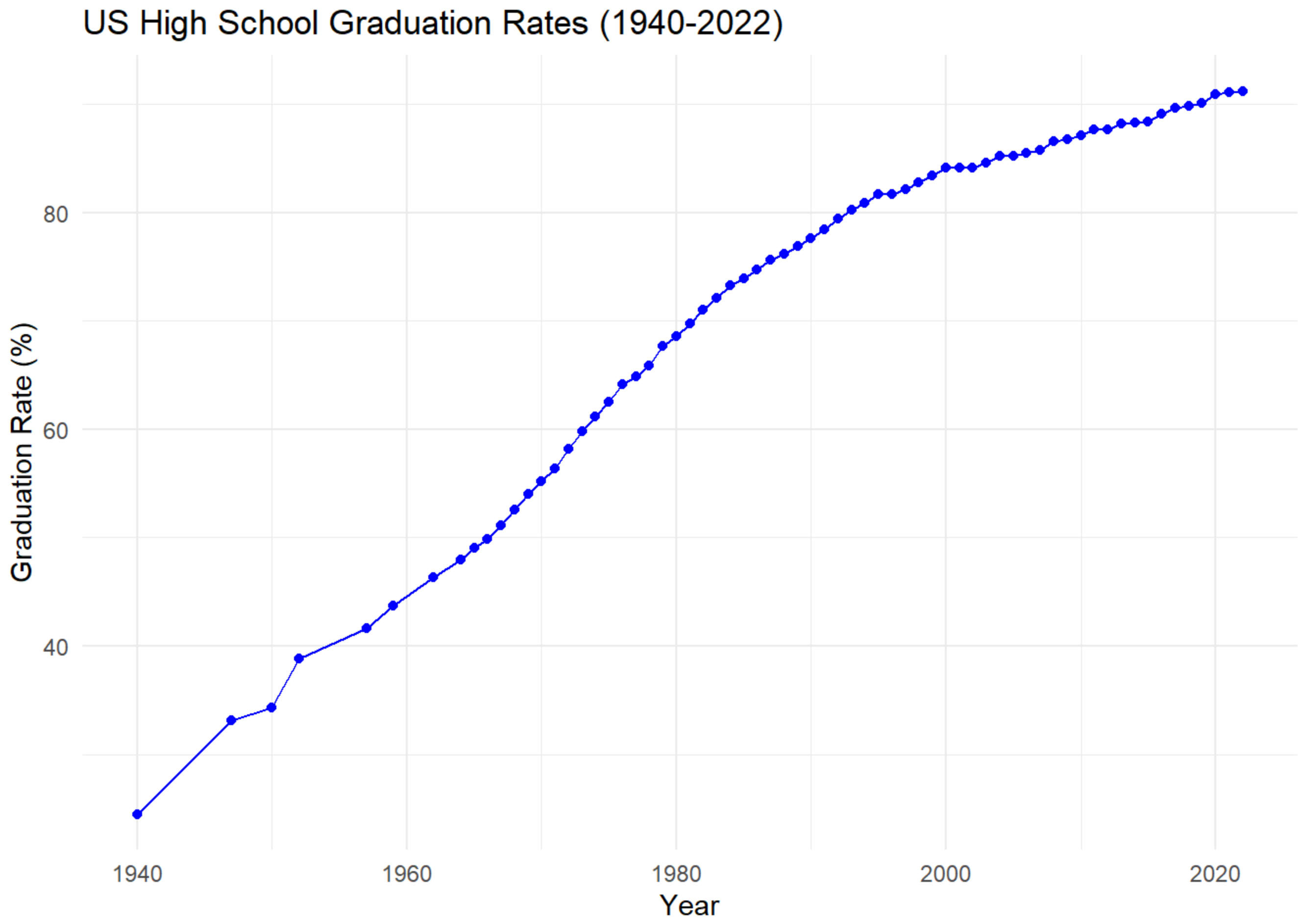


What are the trends in US high school graduation rates over time?

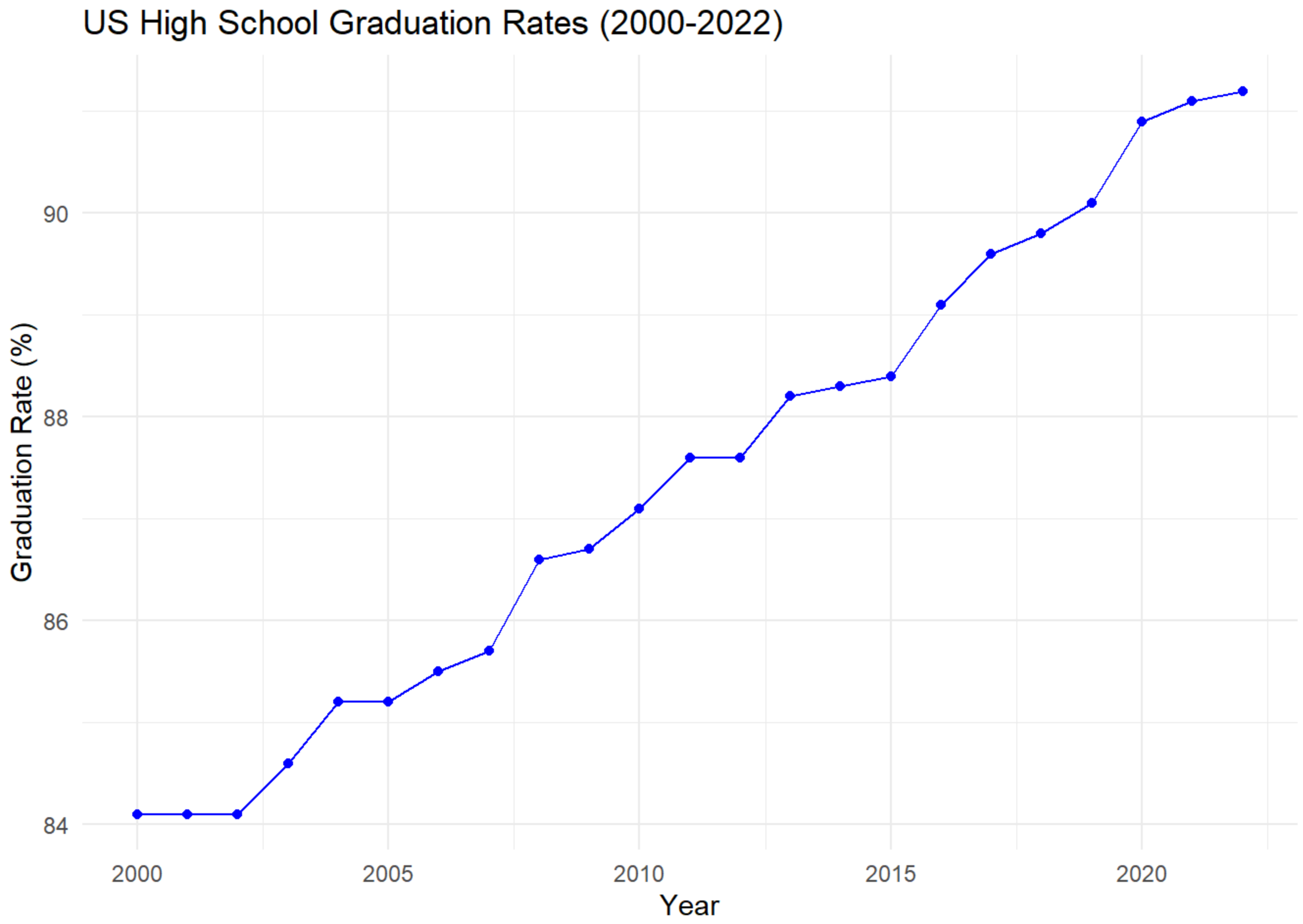
```
#Plot code arranged by Serena

# Data from https://www.census.gov/data/tables/time-series/demo/educational-attainment/cps-histo
rical-time-series.html
# "Percent of People 25 Years and Over Who Have Completed High School or College, by Race, Hispa
nic Origin and gender: Selected Years 1940 to 2022" dataset

# Make data frame of high school graduation rates by year
hs_grad_years <- data.frame(
  Year = c(2022, 2021, 2020, 2019, 2018, 2017, 2016, 2015, 2014, 2013, 2012, 2011, 2010, 2009, 2
008, 2007, 2006, 2005, 2004, 2003, 2002, 2001, 2000, 1999, 1998, 1997, 1996, 1995, 1994, 1993, 1
992, 1991, 1990, 1989, 1988, 1987, 1986, 1985, 1984, 1983, 1982, 1981, 1980, 1979, 1978, 1977, 1
976, 1975, 1974, 1973, 1972, 1971, 1970, 1969, 1968, 1967, 1966, 1965, 1964, 1962, 1959, 1957, 1
952, 1950, 1947, 1940),
  GraduationRate = c(91.2, 91.1, 90.9, 90.1, 89.8, 89.6, 89.1, 88.4, 88.3, 88.2, 87.6, 87.6, 87.
1, 86.7, 86.6, 85.7, 85.5, 85.2, 85.2, 84.6, 84.1, 84.1, 84.1, 83.4, 82.8, 82.1, 81.7, 81.7, 80.
9, 80.2, 79.4, 78.4, 77.6, 76.9, 76.2, 75.6, 74.7, 73.9, 73.3, 72.1, 71.0, 69.7, 68.6, 67.7, 65.
9, 64.9, 64.1, 62.5, 61.2, 59.8, 58.2, 56.4, 55.2, 54.0, 52.6, 51.1, 49.9, 49.0, 48.0, 46.3, 43.
7, 41.6, 38.8, 34.3, 33.1, 24.5))

head(hs_grad_years)
```

```
##   Year GraduationRate
## 1 2022           91.2
## 2 2021           91.1
## 3 2020           90.9
## 4 2019           90.1
## 5 2018           89.8
## 6 2017           89.6
```

```
# Make line plot of high school graduation rates by year from 1940-2022
hs_grad_plot <- ggplot(hs_grad_years, aes(x = Year, y = GraduationRate)) +
  geom_line(color = "blue") +
  geom_point(color = "blue") +
  labs(title = "US High School Graduation Rates (1940-2022)",
       x = "Year",
       y = "Graduation Rate (%)") +
  theme_minimal()

hs_grad_plot
```



```
#Plot code arranged by Serena

# Filter dataset to only include years between 2000 and 2022
hs_grad_00_22 <- hs_grad_years %>%
  filter(Year >= 2000 & Year <= 2022)

# Make line plot of high school graduation rates by year from 2000-2022
hs_grad_00_22_plot <- ggplot(hs_grad_00_22, aes(x = Year, y = GraduationRate)) +
  geom_line(color = "blue") +
  geom_point(color = "blue") +
  labs(title = "US High School Graduation Rates (2000-2022)",
       x = "Year",
       y = "Graduation Rate (%)") +
  theme_minimal()

hs_grad_00_22_plot
```

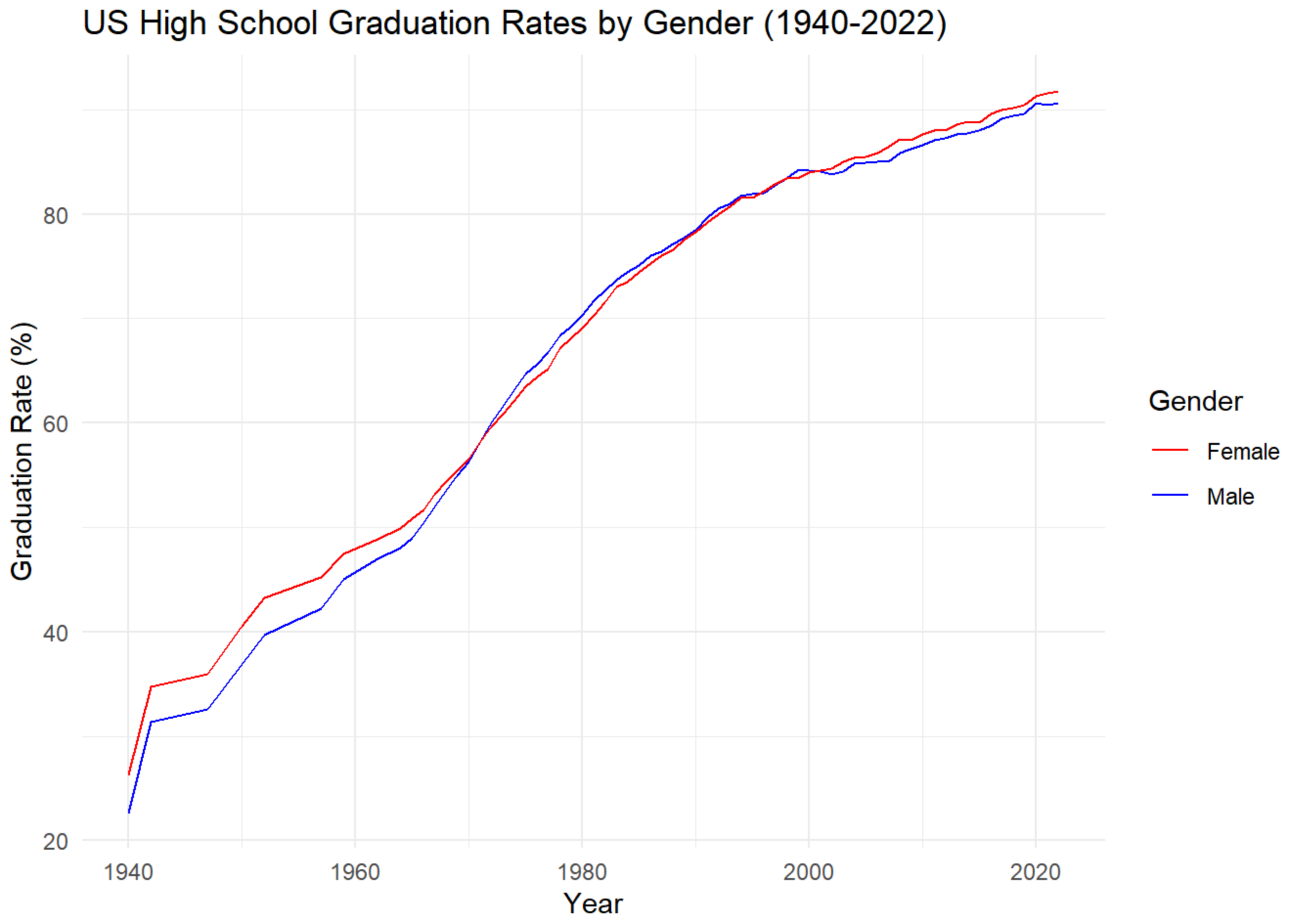


How do high school graduation rates vary for different genders over time?

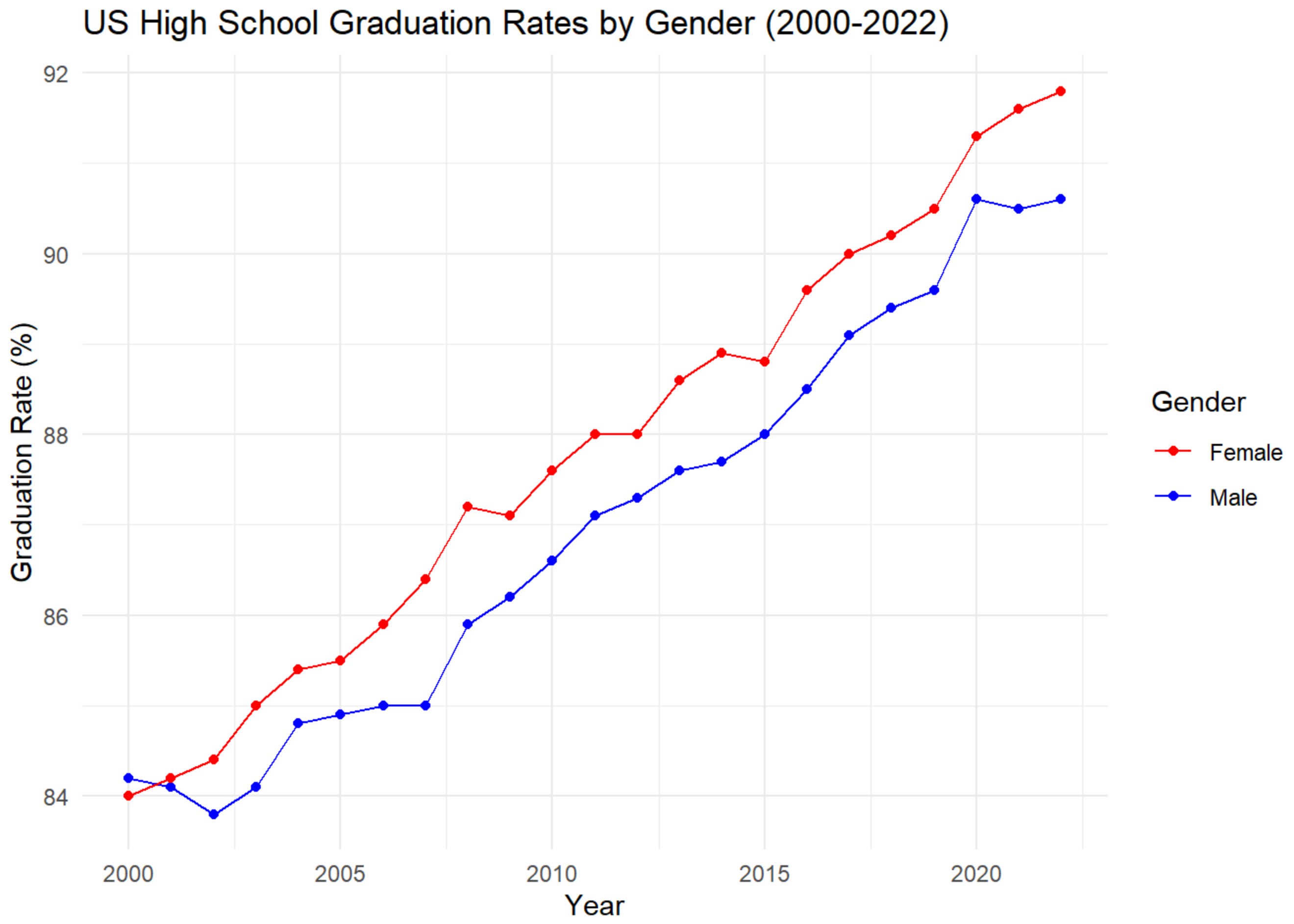
```
# Code arranged by Serena

# Make data frame of graduation rates by gender
hs_grad_gender <- data.frame(
  Year = c(2022, 2021, 2020, 2019, 2018, 2017, 2016, 2015, 2014, 2013, 2012, 2011, 2010, 2009, 2
008, 2007, 2006, 2005, 2004, 2003, 2002, 2001, 2000, 1999, 1998, 1997, 1996, 1995, 1994, 1993, 1
992, 1991, 1990, 1989, 1988, 1987, 1986, 1985, 1984, 1983, 1982, 1981, 1980, 1979, 1978, 1977, 1
976, 1975, 1974, 1973, 1972, 1971, 1970, 1969, 1968, 1967, 1966, 1965, 1964, 1962, 1959, 1957, 1
952, 1950, 1947, 1942, 1940),
  Male = c(90.6, 90.5, 90.6, 89.6, 89.4, 89.1, 88.5, 88, 87.7, 87.6, 87.3, 87.1, 86.6, 86.2, 85.
9, 85, 85, 84.9, 84.8, 84.1, 83.8, 84.1, 84.2, 84.2, 83.4, 82.8, 82, 81.9, 81.7, 81, 80.5, 79.7,
78.5, 77.7, 77.2, 76.4, 76, 75.1, 74.4, 73.7, 72.7, 71.7, 70.3, 69.2, 68.4, 66.8, 65.6, 64.7, 6
3.1, 61.6, 60, 58.2, 56.3, 55, 53.6, 52, 50.5, 49, 48, 47, 45, 42.2, 39.7, 36.9, 32.6, 31.4, 22.
7),
  Female = c(91.8, 91.6, 91.3, 90.5, 90.2, 90, 89.6, 88.8, 88.9, 88.6, 88, 88, 87.6, 87.1, 87.2,
86.4, 85.9, 85.5, 85.4, 85, 84.4, 84.2, 84, 83.4, 83.4, 82.9, 82.2, 81.6, 81.6, 80.7, 80, 79.2,
78.3, 77.5, 76.6, 76, 75.3, 74.4, 73.5, 73, 71.5, 70.3, 69.1, 68.1, 67.1, 65.2, 64.4, 63.5, 62.
1, 60.9, 59.6, 58.2, 56.6, 55.4, 54.4, 53.2, 51.7, 50.8, 49.9, 48.9, 47.5, 45.2, 43.3, 40.5, 36,
34.7, 26.3))

head(hs_grad_gender)
```

```
##   Year Male Female
## 1 2022 90.6   91.8
## 2 2021 90.5   91.6
## 3 2020 90.6   91.3
## 4 2019 89.6   90.5
## 5 2018 89.4   90.2
## 6 2017 89.1   90.0
```

```
# Make line plot of high school graduation rates by gender from 1940-2022
hs_grad_gender_plot <- ggplot(hs_grad_gender, aes(x = Year)) +
  geom_line(aes(y = Male, color = "Male")) +
  #geom_point(aes(y = Male, color = "Male")) +
  geom_line(aes(y = Female, color = "Female")) +
  #geom_point(aes(y = Female, color = "Female")) +
  labs(title = "US High School Graduation Rates by Gender (1940-2022)",
       x = "Year",
       y = "Graduation Rate (%)",
       color = "Gender") +
  scale_color_manual(values = c("Male" = "blue", "Female" = "red")) +
  theme_minimal()

hs_grad_gender_plot
```



```
# Code arranged by Serena

# Filter dataset to only include years between 2000 and 2022
hs_grad_gender_00_22 <- hs_grad_gender %>%
  filter(Year >= 2000 & Year <= 2022)

# Make line plot of high school graduation rates by gender from 2000-2022
hs_grad_gender_00_22_plot <- ggplot(hs_grad_gender_00_22, aes(x = Year)) +
  geom_line(aes(y = Male, color = "Male")) +
  geom_point(aes(y = Male, color = "Male")) +
  geom_line(aes(y = Female, color = "Female")) +
  geom_point(aes(y = Female, color = "Female")) +
  labs(title = "US High School Graduation Rates by Gender (2000-2022)",
       x = "Year",
       y = "Graduation Rate (%)",
       color = "Gender") +
  scale_color_manual(values = c("Male" = "blue", "Female" = "red")) +
  theme_minimal()

hs_grad_gender_00_22_plot
```



How do high school graduation rates vary for different races over time?

```
# Make data frame of graduation rates by race
# Races = White, Black, Hispanic, Asian
hs_grad_race <- data.frame(
  Year = c(2022, 2021, 2020, 2019, 2018, 2017, 2016, 2015, 2014, 2013, 2012, 2011, 2010, 2009, 2
008, 2007, 2006, 2005, 2004, 2003),
  White = c(91.4, 91.2, 91.3, 90.5, 90.2, 90.1, 89.5, 88.8, 88.8, 88.6, 88.1, 88.1, 87.6, 87.1,
87.1, 86.2, 86.1, 85.8, 85.8, 85.1),
  Black = c(90.1, 90.3, 89.4, 87.9, 87.9, 87.3, 87.1, 87, 85.8, 85.1, 85, 84.5, 84.2, 84.1, 83,
82.3, 80.7, 81.1, 80.6, 80),
  Hispanic = c(75.2, 74.2, 74.3, 71.8, 71.6, 70.5, 68.5, 66.7, 66.5, 66.2, 65, 64.3, 62.9, 61.9,
62.3, 60.3, 59.3, 58.5, 58.4, 57),
  Asian = c(92.3, 92.9, 91.6, 91.2, 90.5, 90.9, 90.3, 89.1, 89.5, 90.1, 88.9, 88.6, 88.9, 88.2,
88.7, 87.8, 87.4, 87.6, 86.8, 87.6)
)

head(hs_grad_race)
```

```
##   Year White Black Hispanic Asian
## 1 2022  91.4  90.1     75.2  92.3
## 2 2021  91.2  90.3     74.2  92.9
## 3 2020  91.3  89.4     74.3  91.6
## 4 2019  90.5  87.9     71.8  91.2
## 5 2018  90.2  87.9     71.6  90.5
## 6 2017  90.1  87.3     70.5  90.9
```

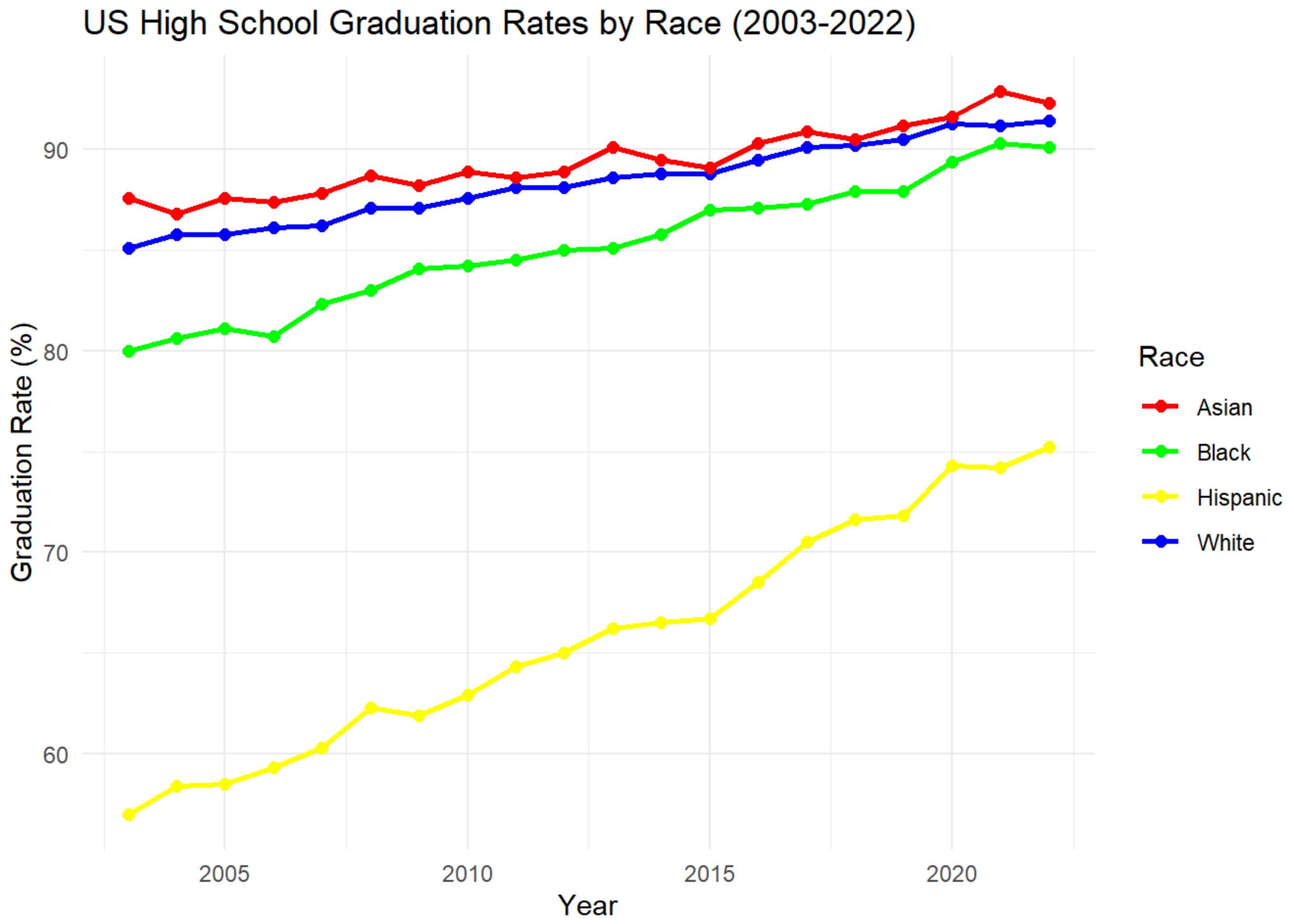
```
# Make race data frame long to make plotting the line plot easier
long_hs_grad_race <- hs_grad_race %>%
  gather(key = "Race", value = "GraduationRate", White, Black, Hispanic, Asian)

head(long_hs_grad_race)
```

```
##   Year  Race GraduationRate
## 1 2022 White           91.4
## 2 2021 White           91.2
## 3 2020 White           91.3
## 4 2019 White           90.5
## 5 2018 White           90.2
## 6 2017 White           90.1
```

```
# Code arranged by Serena

# Make line plot of high school graduation rates by race from 2003-2022
# Dataset didn't include all race data from 2002 and before
hs_grad_race_plot <- ggplot(long_hs_grad_race, aes(x = Year, y = GraduationRate, color = Race))
+
  geom_line(size = 1) +
  geom_point(size = 2) +
  labs(title = "US High School Graduation Rates by Race (2003-2022)",
       x = "Year",
       y = "Graduation Rate (%)",
       color = "Race") +
  theme_minimal() +
  scale_color_manual(values = c("White" = "blue", "Black" = "green", "Hispanic" = "yellow", "Asi
an" = "red"))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
hs_grad_race_plot
```



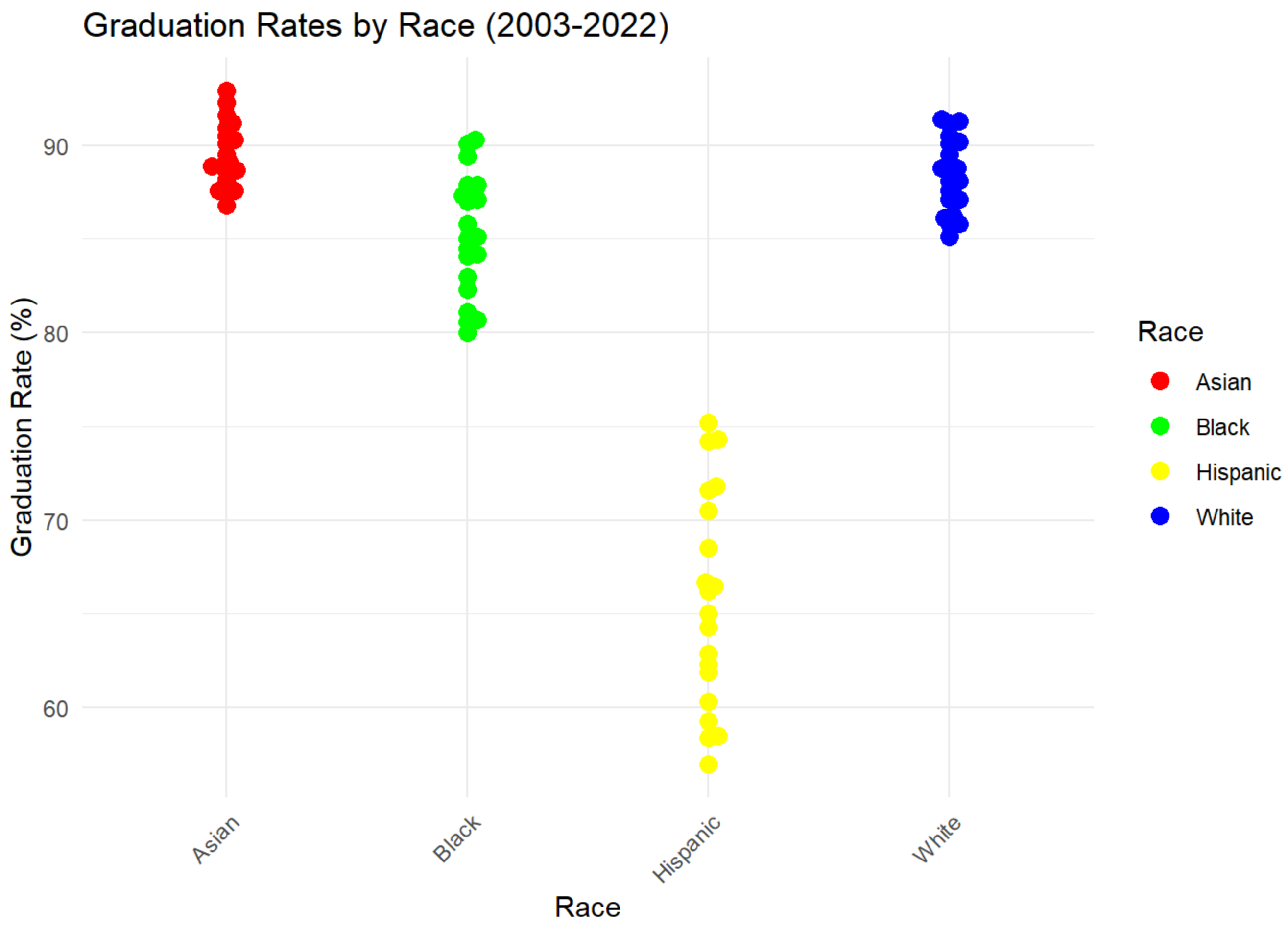
Beeswarm Plot vs. Violin Plot

For the visualizations of the high school graduation rates for different genders, we thought that the line graph distinctly showed that the Hispanic group was the lowest performing group. However, we wanted to include another visualization to show the distribution of the graduation rates within each gender.

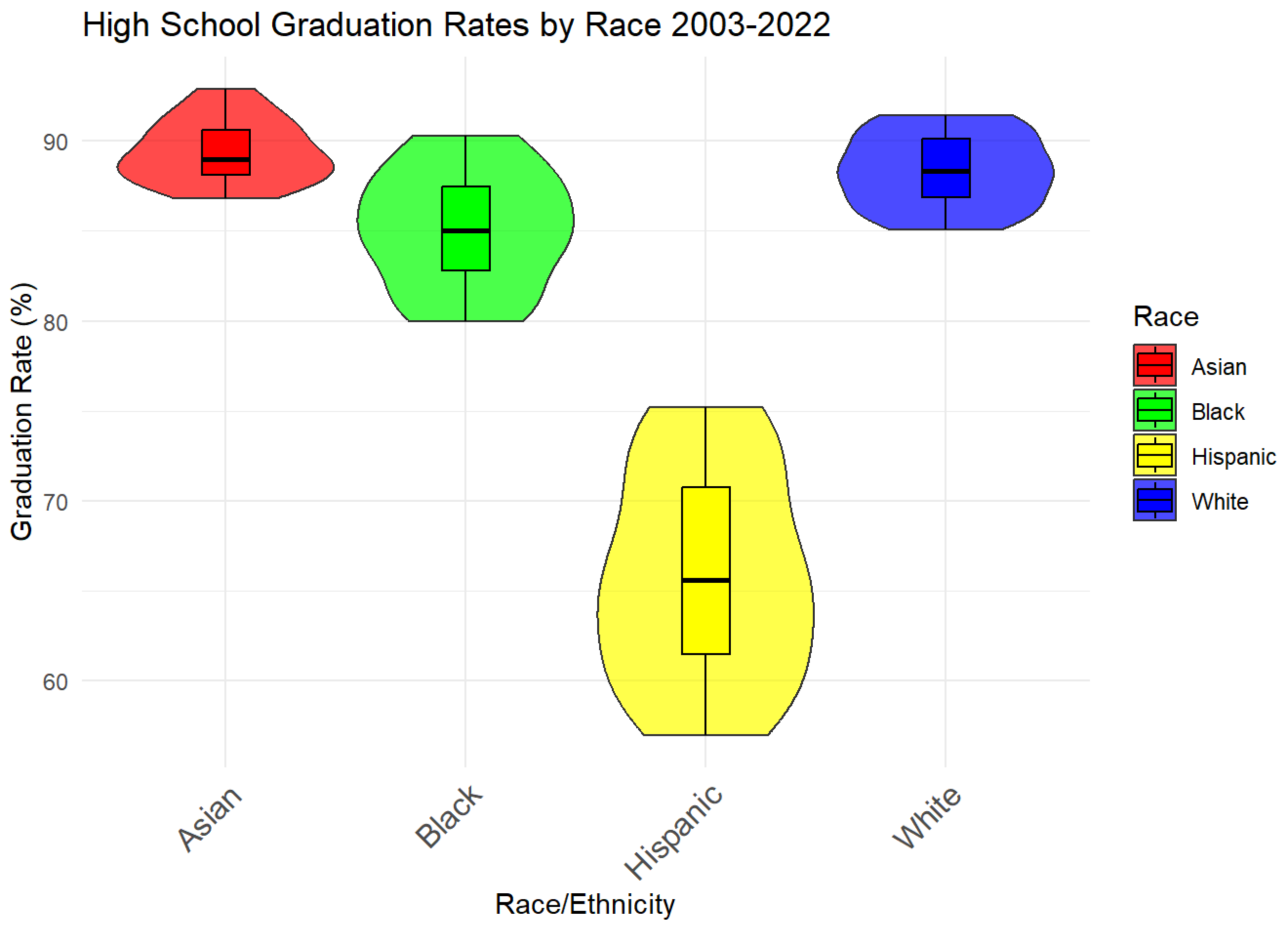
We originally decided to make a beeswarm graph to show the distribution, but our limited years data couldn’t clearly show the distribution using a beeswarm graph.

So, we looked to find a different kind of graph to use to show the distribution, and we landed on the violin plot! With our data, you can still clearly see that the Hispanic group is the lowest performing group. But, you can also now clearly see that the Hispanic group had the widest distribution of graduation rates over the years.

```
# Code arranged by Serena

hs_grad_race_long <- hs_grad_race %>%
  pivot_longer(cols = -Year, names_to = "Race", values_to = "Graduation_Rate")

race_swarm_plot <- ggplot(hs_grad_race_long, aes(x = Race, y = Graduation_Rate, color = Race)) +
  geom_beeswarm(size = 3) +
  labs(title = "Graduation Rates by Race (2003-2022)",
       x = "Race",
       y = "Graduation Rate (%)",
       color = "Race") +
  scale_color_manual(values = c("White" = "blue", "Black" = "green", "Hispanic" = "yellow", "Asi
an" = "red")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

race_swarm_plot
```



```
# Code arranged by Tatyanna

# Make race data frame long to make plotting the violin plot easier
hs_grad_race_long <- hs_grad_race %>%
    pivot_longer(cols = -Year,
                 names_to = "Race",
                 values_to = "Graduation_Rate")

# Make violin plot of high school graduation rates by race from 2003-2022
hs_grad_race_violin_plot <- ggplot(hs_grad_race_long, aes(x = Race, y = Graduation_Rate, fill =
Race)) +
  geom_violin(scale = "width", alpha = 0.7) +
  geom_boxplot(width = 0.2, outlier.shape = NA, color = "black") +
  scale_fill_manual(values = c("White" = "blue", "Black" = "green", "Hispanic" = "yellow", "Asia
n" = "red")) +
  theme_minimal() +
  labs(title = "High School Graduation Rates by Race 2003-2022",
       x = "Race/Ethnicity",
       y = "Graduation Rate (%)",
       fill = "Race") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 12))

hs_grad_race_violin_plot
```

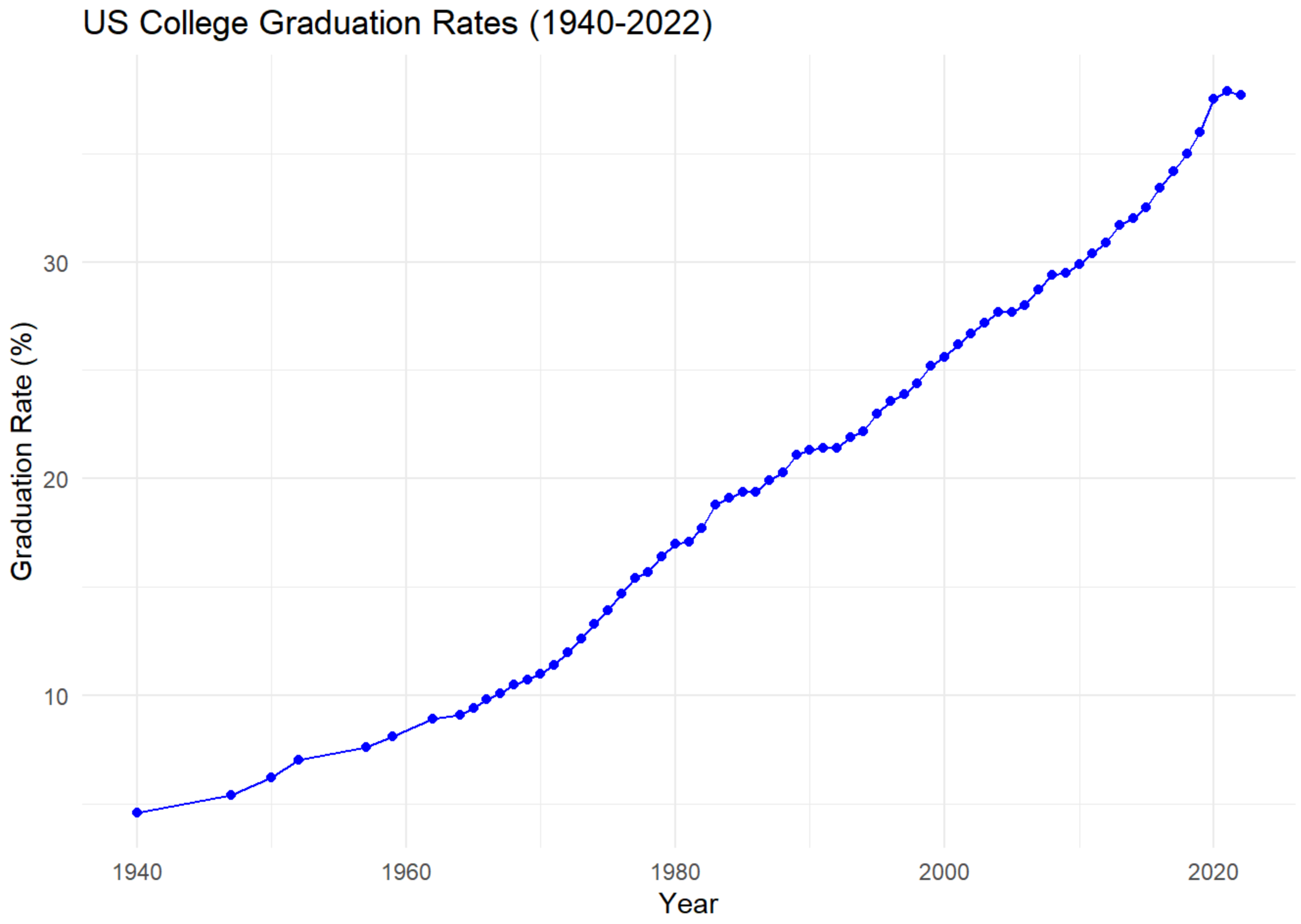


What are the trends in US college graduation rates over time?

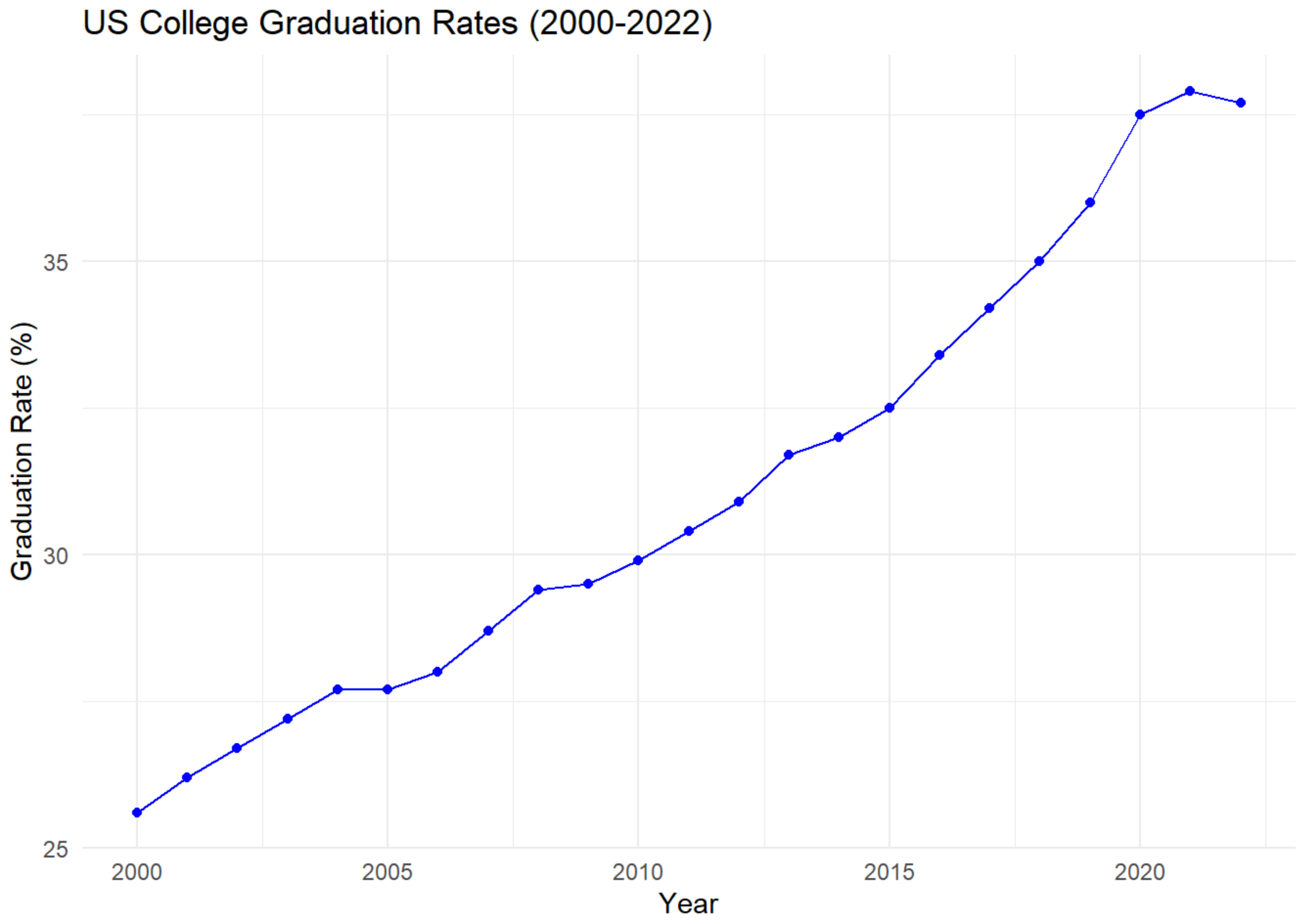
```
# Code arranged by Serena

# Make data frame of college graduation rates by year from 1940-2022
college_grads <- data.frame(
  Year = c(2022, 2021, 2020, 2019, 2018, 2017, 2016, 2015, 2014, 2013, 2012, 2011, 2010, 2009, 2
008, 2007, 2006, 2005, 2004, 2003, 2002, 2001, 2000, 1999, 1998, 1997, 1996, 1995, 1994, 1993, 1
992, 1991, 1990, 1989, 1988, 1987, 1986, 1985, 1984, 1983, 1982, 1981, 1980, 1979, 1978, 1977, 1
976, 1975, 1974, 1973, 1972, 1971, 1970, 1969, 1968, 1967, 1966, 1965, 1964, 1962, 1959, 1957, 1
952, 1950, 1947, 1940),
  GraduationRate = c(37.7, 37.9, 37.5, 36.0, 35.0, 34.2, 33.4, 32.5, 32.0, 31.7, 30.9, 30.4, 29.
9, 29.5, 29.4, 28.7, 28.0, 27.7, 27.7, 27.2, 26.7, 26.2, 25.6, 25.2, 24.4, 23.9, 23.6, 23.0, 22.
2, 21.9, 21.4, 21.4, 21.3, 21.1, 20.3, 19.9, 19.4, 19.4, 19.1, 18.8, 17.7, 17.1, 17.0, 16.4, 15.
7, 15.4, 14.7, 13.9, 13.3, 12.6, 12.0, 11.4, 11.0, 10.7, 10.5, 10.1, 9.8, 9.4, 9.1, 8.9, 8.1, 7.
6, 7.0, 6.2, 5.4, 4.6))

head(college_grads)
```

```
##   Year GraduationRate
## 1 2022           37.7
## 2 2021           37.9
## 3 2020           37.5
## 4 2019           36.0
## 5 2018           35.0
## 6 2017           34.2
```

```
# Make line plot of college graduation rates from 1940-2022
college_grad_plot <- ggplot(college_grads, aes(x = Year, y = GraduationRate)) +
  geom_line(color = "blue") +
  geom_point(color = "blue") +
  labs(title = "US College Graduation Rates (1940-2022)",
       x = "Year",
       y = "Graduation Rate (%)") +
  theme_minimal()

college_grad_plot
```



```
# Code arranged by Serena

# Filter dataset to only include years between 2000 and 2022
college_grad_00_22 <- college_grads %>%
  filter(Year >= 2000 & Year <= 2022)

# Make line plot of college graduation rates from 2000-2022
college_grad_00_22_plot <- ggplot(college_grad_00_22, aes(x = Year, y = GraduationRate)) +
  geom_line(color = "blue") +
  geom_point(color = "blue") +
  labs(title = "US College Graduation Rates (2000-2022)",
       x = "Year",
       y = "Graduation Rate (%)") +
  theme_minimal()

college_grad_00_22_plot
```



How do poverty rates impact US high school graduation rates?

```
# Data from USDA "All people in poverty (2021)" dataset

# Make data frame of poverty rates by state
state_poverty_21 <- data.frame(
  State = c("ALABAMA", "ALASKA", "ARIZONA", "ARKANSAS", "CALIFORNIA", "COLORADO", "CONNECTICUT",
"DELAWARE", "DISTRICT OF COLUMBIA", "FLORIDA", "GEORGIA", "HAWAII", "IDAHO", "ILLINOIS", "INDIAN
A", "IOWA", "KANSAS", "KENTUCKY", "LOUISIANA", "MAINE", "MARYLAND", "MASSACHUSETTS", "MICHIGAN",
"MINNESOTA", "MISSISSIPPI", "MISSOURI", "MONTANA", "NEBRASKA", "NEVADA", "NEW HAMPSHIRE", "NEW J
ERSEY", "NEW MEXICO", "NEW YORK", "NORTH CAROLINA", "NORTH DAKOTA", "OHIO", "OKLAHOMA", "OREGO
N", "PENNSYLVANIA", "RHODE ISLAND", "SOUTH CAROLINA", "SOUTH DAKOTA", "TENNESSEE", "TEXAS", "UTA
H", "VERMONT", "VIRGINIA", "WASHINGTON", "WEST VIRGINIA", "WISCONSIN", "WYOMING"), PovertyPercen
t = c(16.3, 10.8, 12.9, 16, 12.3, 9.7, 10.1, 11.5, 16.8, 13.2, 14.2, 10.9, 10.8, 12.1, 12.1, 11,
11.6, 16.3, 19.5, 11.2, 10.3, 10.4, 13, 9.3, 19.2, 12.8, 12, 10.5, 14, 7.4, 10.2, 17.7, 14, 13.
5, 10.9, 13.3, 15.4, 12.2, 12, 12.1, 14.5, 11.9, 13.7, 14.2, 8.7, 10.2, 10.3, 9.9, 16.8, 10.8, 1
0.6))

head(state_poverty_21)
```

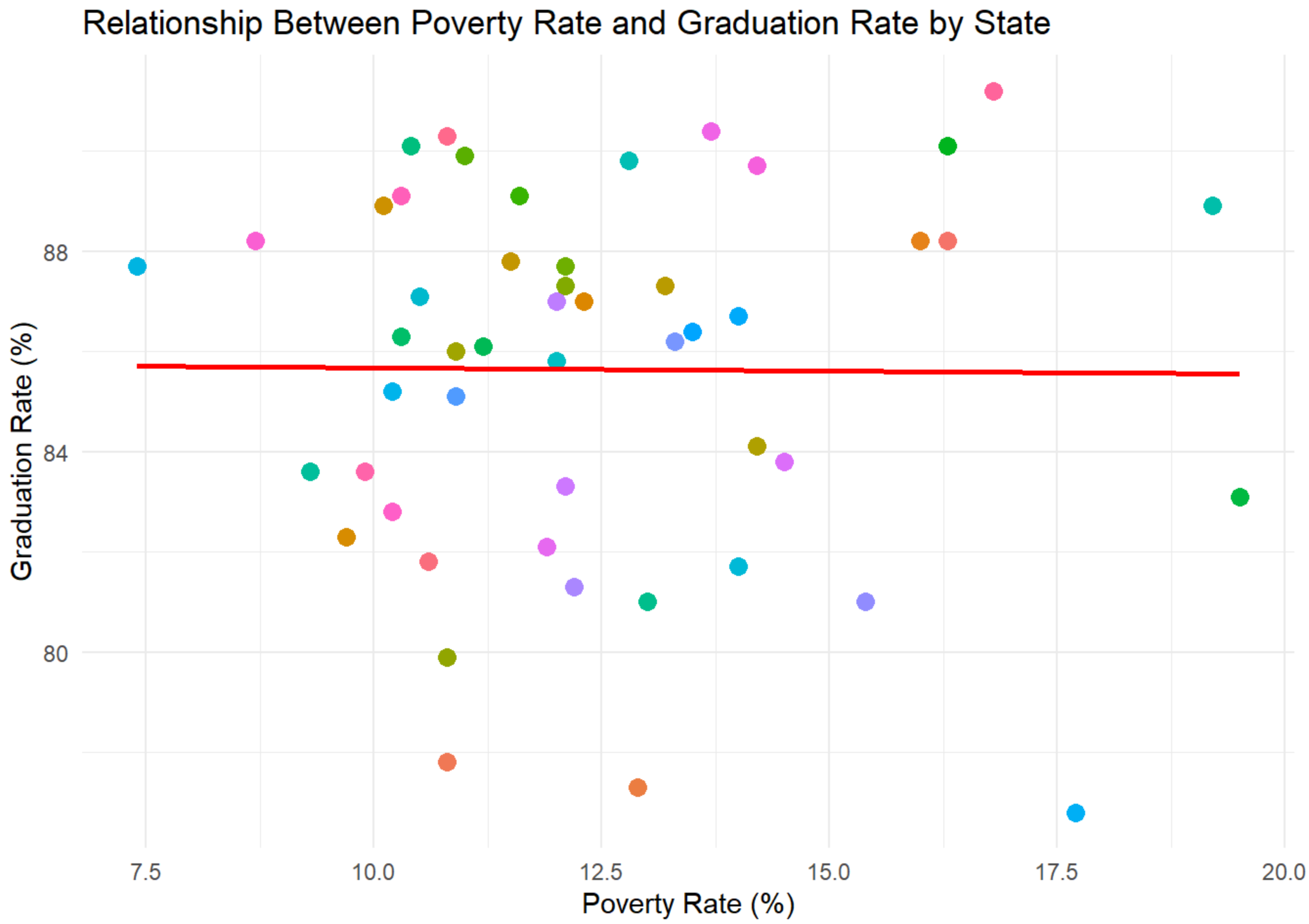
```
##         State PovertyPercent
## 1    ALABAMA           16.3
## 2     ALASKA           10.8
## 3    ARIZONA           12.9
## 4   ARKANSAS           16.0
## 5 CALIFORNIA           12.3
## 6   COLORADO            9.7
```

```
# Make a new states data frame that includes a poverty column
states2 <- states %>%
  left_join(state_poverty_21, by = "State")

states2 <- states2[!is.na(states2$PovertyPercent), ]

head(states2)
```

```
## # A tibble: 6 × 5
##   `School Year` State      Value Subgroup            PovertyPercent
##   <chr>         <chr>      <dbl> <chr>                        <dbl>
## 1 2021-2022     ALABAMA     88.2 All Students in SEA           16.3
## 2 2021-2022     ALASKA      77.8 All Students in SEA           10.8
## 3 2021-2022     ARIZONA     77.3 All Students in SEA           12.9
## 4 2021-2022     ARKANSAS    88.2 All Students in SEA           16  
## 5 2021-2022     CALIFORNIA  87   All Students in SEA           12.3
## 6 2021-2022     COLORADO    82.3 All Students in SEA            9.7
```

```
# Code arranged by Serena

# Make scatter plot of Poverty Rate vs Graduation Rate by state
# Each colored dot is a state
poverty_scatter_plot <- ggplot(states2, aes(x = PovertyPercent, y = Value)) +
  geom_point(aes(color = State), size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Relationship Between Poverty Rate and Graduation Rate by State",
       x = "Poverty Rate (%)", y = "Graduation Rate (%)") +
  theme_minimal() +
  theme(legend.position = "none")

poverty_scatter_plot
```

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



```
# Calculate linear regression model from scatter plot
lm_model <- lm(Value ~ PovertyPercent, data = states2)

summary(lm_model)
```

```
##
## Call:
## lm(formula = Value ~ PovertyPercent, data = states2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7758 -2.4264  0.7006  2.6055  5.6128
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     85.8005     2.5794  33.263   <2e-16 ***
## PovertyPercent  -0.0127     0.2023  -0.063     0.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.685 on 48 degrees of freedom
## Multiple R-squared:  8.203e-05,  Adjusted R-squared:  -0.02075
## F-statistic: 0.003938 on 1 and 48 DF,  p-value: 0.9502
```

Linear Regression Model Results

This linear regression model shows that poverty rate appears to have an almost nonexistent effect on graduation rates, with a **coefficient of -0.0127**, meaning a 1% increase in poverty is predicted to decrease graduation rates by just 0.0127%.

The **p-value of 0.95** shows that this effect is extremely weak and statistically insignificant.

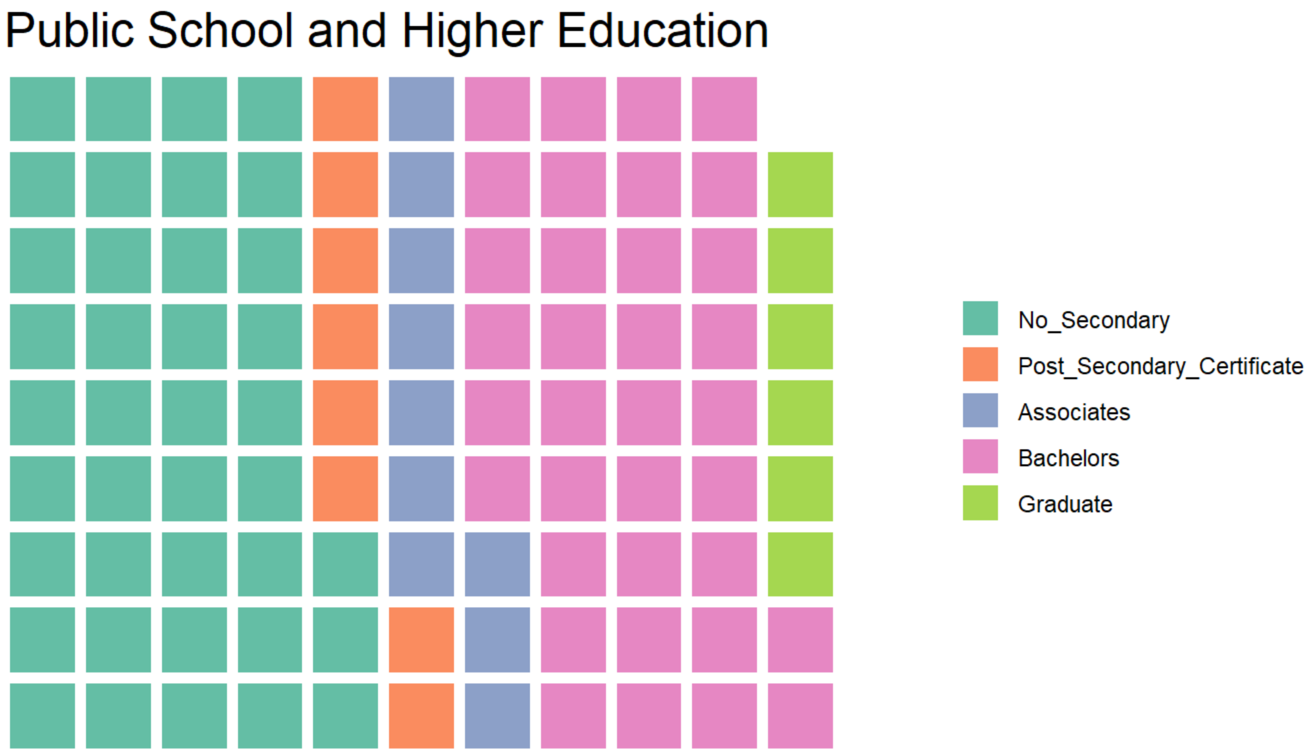
The model has virtually no predictive power, with an **R-squared of 0.00008**, indicating that poverty rate explains almost none of the variation in graduation rates.

Based on the linear regression model of the scatter plot, poverty rate **does not** appear to significantly affect graduation rates.

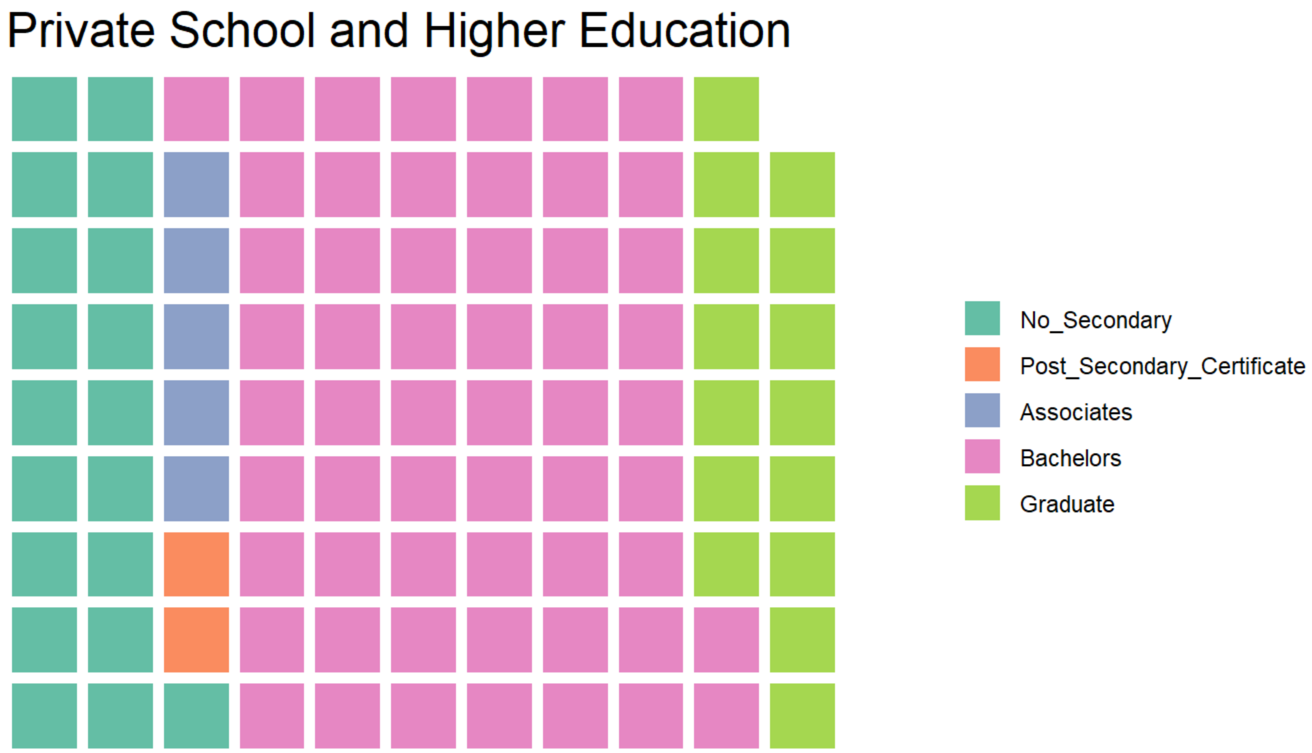
How does learning in private vs public schools affect students' higher education rates?

```
PublicData <- c("No_Secondary" = 39.1, "Post_Secondary_Certificate" = 8.1, "Associates" = 10.5,
"Bachelors" = 35.5, "Graduate" = 6.8)

PrivateData <- c("No_Secondary" = 19.3, "Post_Secondary_Certificate" = 2.5, "Associates" = 5.2,
"Bachelors" = 57.4, "Graduate" = 15.7)
```

```
# Code arranged by Elizabeth

waffle(PublicData,
       rows= 9,
       keep = TRUE,
       title = "Public School and Higher Education",
       )
```



```
# Code arranged by Elizabeth

waffle(PrivateData,
       rows= 9,
       keep = TRUE,
       title = "Private School and Higher Education",
       )
```



Private School and Higher Education

How does the type of area students live in affect their post-high school plans?

Sankey Graph: Notes and Issues

```
# This code did not work while trying to make sankey plot

# Data from https://nces.ed.gov/pubs2024/2024022.pdf

# sankey_manual_data <- "Location,no_postsecondary, postsecondary_certificate, associates, bache
lors, graduate City, 39.7, 7.6, 7.8, 37.7, 7.2 Suburb, 34.5, 7.1, 8.6, 41.4, 8.4 Town, 39.3, 7.
5, 13.9, 32.7, 6.6 Rural, 37.3, 8.2, 12.4, 34.6, 7.5"

#making sure the data works
#df <- read_csv(I(sankey_manual_data))

#because making my own csv file did not work, I am going to import a csv file of the same data t
hrough google sheets/ numbers on mac

# failed use of highcharter package
#custom stuff for the sankey
#pl <- highchart() %>%
 # hc_add_series(
  #  type = "sankey",
   # data = sankey_list,
    #name = "What Degree did they Receive?"
#  ) %>%
 # hc_title(text = "What Degree did they Receive?") %>%
  #hc_subtitle(text = "by: Elizabeth") %>%
  #hc_add_theme(hc_theme_economist())
```

```
# Data from https://nces.ed.gov/pubs2024/2024022.pdf

# Load sankey csv data
sankey_data <- read_csv("C:/Users/samro/Downloads/graduation.final.2.0-1.csv")
```

```
## Rows: 4 Columns: 6
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): Location
## dbl (5): no_postsecondary, postsecondary_certificate, associates, bachelors,...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Rename columns for easier readability
sankey_data <- sankey_data %>% rename(No_Post_Secondary = no_postsecondary, Post_Secondary_Certi
ficate = postsecondary_certificate, Associates = associates, Bachelors = bachelors, Graduate = g
raduate)

head(sankey_data)
```

```
## # A tibble: 4 × 6
##   Location No_Post_Secondary Post_Secondary_Certificate Associates Bachelors
##   <chr>                <dbl>                      <dbl>      <dbl>     <dbl>
## 1 City                  7940                       1520       1560      7540
## 2 Suburb                6900                       1420       1720      8280
## 3 Town                  6900                       1500       2780      6540
## 4 Rural                 7460                       1640       2480      6920
## # ℹ 1 more variable: Graduate <dbl>
```

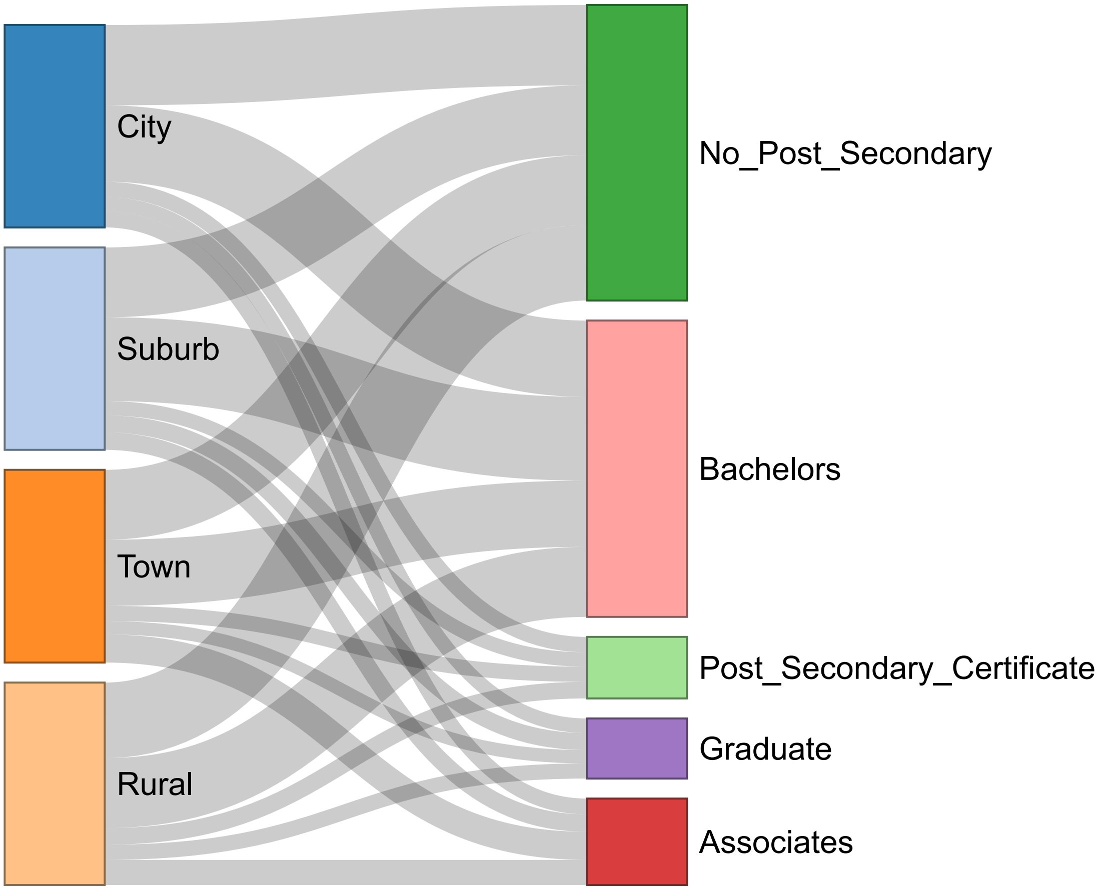
```
# Assign estimated sample size based on article numbers
total_sample_size <- 20000

# Make sankey data frame long to make plotting the sankey plot easier
long_df <- sankey_data %>%
  pivot_longer(-Location, names_to = "Education", values_to = "Value") %>%
  mutate(Value = (Value / total_sample_size) * 100)

# Rename columns for sankey
long_df <- long_df %>%
  rename(Source = Location, Target = Education)
```

```
# Code arranged by Elizabeth

# Make nodes data frame
nodes <- data.frame(name = trimws(unique(c(long_df$Source, long_df$Target))))

# Make links data frame
links <- long_df %>%
  mutate(
    SourceID = match(trimws(Source), nodes$name) - 1,
    TargetID = match(trimws(Target), nodes$name) - 1,
    Tooltip = paste0(Source, " → ", Target, ": ", round(Value, 2), "%"),
    LinkGroup = Source
  ) %>%
  select(SourceID, TargetID, Value, LinkGroup, Tooltip)

# Make actual sankey diagram
sankey <- sankeyNetwork(
  Links = links,
  Nodes = nodes,
  Source = "SourceID",
  Target = "TargetID",
  Value = "Value",
  NodeID = "name",
  sinksRight = FALSE,
  fontSize = 16,
  nodeWidth = 50)
```

```
## Links is a tbl_df. Converting to a plain data frame.
```

```
# Adding tooltip for both links and nodes
sankey_plot <- htmlwidgets:: onRender(
  sankey,
  "
  function(el, x) {
    // percent values on the links
    d3.selectAll('.link')
      .append('title')  // Create a tooltip for each link
      .text(function(d) {
        return x.nodes[d.source].name + ' → ' + x.nodes[d.target].name + ': ' + d.value + '%';
      });

    // percentage text in the center of the link
    d3.selectAll('.link')
      .append('text')
      .attr('text-anchor', 'middle')
      .attr('dy', '0.3em')  // Vertical alignment of the text
      .style('fill', 'blue')  // text color
      .style('font-size', '14px')  // font size
      .text(function(d) {
        return d.value.toFixed(2) + '%';  // Show value as percentage (2 decimals)
      });
  }
  "
)

# Using html tools to add a title
sankey_plot <- htmlwidgets::prependContent(sankey_plot, htmltools::tags$h2("School Location and Postsecondary Plans"))

sankey_plot
```

School Location and Postsecondary Plans



This sankey graph shows us that by far, living in a Suburb means that this cohort is more likely to be more educating with them having the highest number of post secondary degrees. Showing that where you live may have an effects on how likely you are to pursue and complete higher education.

Conclusion

In summary, we analyzed trends in US high school and college graduation rates over time. We showed how factors such as residence location, race, poverty, private or pubic education, and gender have affected these graduation rates.

Our data showed that girls are statistically more likely to complete high school and that Hispanics are the least likely to graduate high school. The data also showed that the rate of poverty does not significantly impact high school graduation rates. Also, living in the suburbs means a higher likelihood of being more educated, especially with a post-secondary degree. Finally, receiving a private school education gives a higher chance of getting a post-secondary degree in Florida, and dropout rates are lower in private schools compared to public schools.

The data we have analyzed is extremely significant for students all across America, as being able to understand all of these educational factors can play a huge role in the trajectory of students’ career plans and overall paths in life.

This analysis provides insights into the factors that influence educational success in the US, which helps government officials and educators understand disparities in graduation rates. Understanding these disparities can help inform policies on resource allocation and other factors to improve educational outcomes for everyone, including disadvantaged students.