STAT 184 - Course Project

Ryan Sulaiman, Luke Bislig

Introduction

Education is a global asset that promotes potential, learning, reasoning, empathy, and understanding of the world. This is the process gaining of knowledge, skills, values, and building character. Over the past decades, there have been major transformation of tuition costs, graduation rates, and years to graduate. These changes have influenced how students learn, institutions operate, and how society's understanding of education has changed. Tuition costs have been skyrocketing, which prompt discussion of higher education, affordability, and equity. Redefining educational access, flexibility, and skills for academic success.

This project is going to explore different variable correlations on education.

1. Key factors
    1. Graduation rates,
    2. Tuition costs,
    3. Years spent,
    4. Year taken,
    5. State.

Investigating these questions to illuminate patterns of structures and economy. Obtaining insights to understand the long-term changes of educational progress and challenges.

Data Provenance

This project uses two datasets for the key factors that we looked into for education. These datasets were taken from reliable online sources on Kaggle for analysis.

College Completion Dataset

Source: Kaggle

Description: The dataset contains data on student, success, graduation rates, race, and gender demographics. Measure to compare colleges across states and more. Source of information to help better understand college completion and student success in the United States.

Purpose: Exploring trends in college completion

Case: Each case in the dataset corresponds to a student and state respectively

Average Cost of Tuition Dataset

Source: Kaggle

Description: Constructed using data from the National Center for Education Statistics Annual Digest. Average undergaduate tuition and fees, charged to full-time students at degree-granting postsecondary institutions. By level of institution, state.

Purpose: Analyze the average cost of undergrad students by state in the US

Case: Each case in the dataset corresponds to a student and state respectively

Data Wrangling

In the College Completion dataset, data columns were removed based on the relevance to our research questions.

In the Average Cost of Tuition dataset, data rows were removed based on the data required for the research questions.

FAIR Principles

Findable:

Both datasets are hosted publicly on Kaggle with clear titles, descriptions, and metadata, making them easy to locate.

Accessible:

The datasets can be downloaded by standard HTTP(s) protocols, and metadata remains accessible even if the dataset changes.

Interoperable:

Data is provided in standard CSV format with commonly used variable definitions, allowing integration with other datasets and tools.

Reusable:

The data can be reused because it has a clear license and information about where it came from.

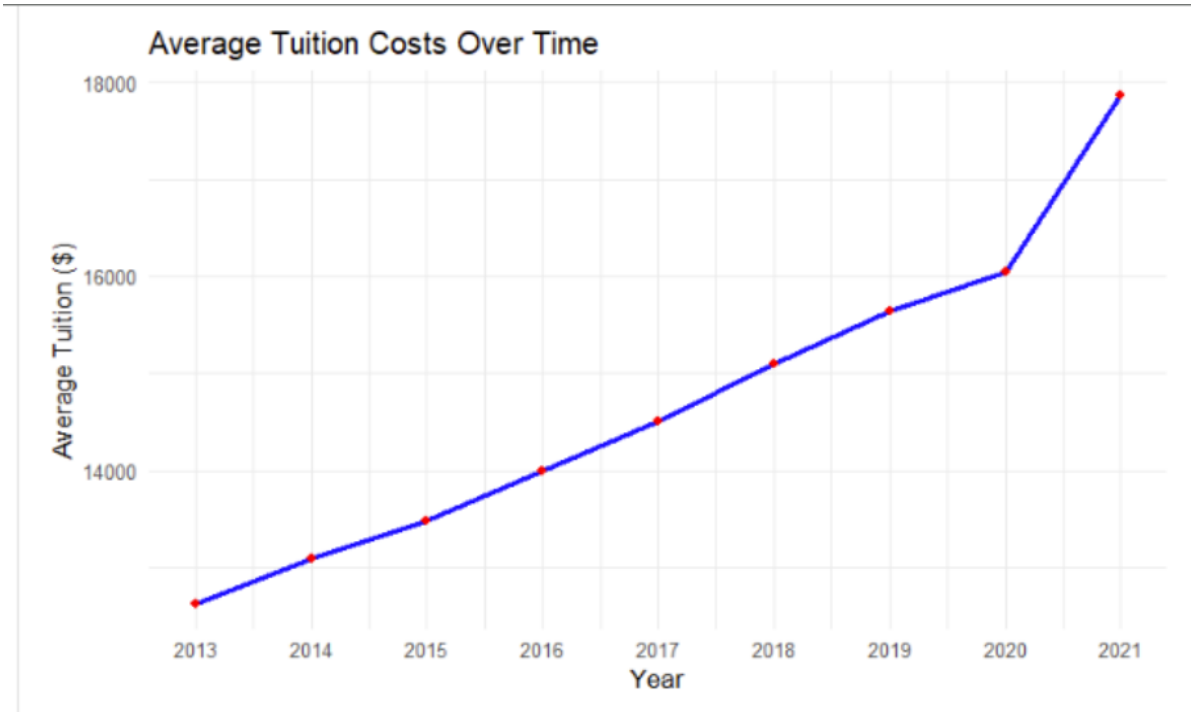
Exploratory Data Analysis

Figure 1 Visualization

This visualization shows how average undergraduate tuition costs have changed from 2013 to 2021.

The line increases over the entire period, showing that tuition prices rise consistently year after year.

From 2013 to 2020, the increase is gradual but steady, and this suggests a predictable upward trend caused by factors like inflation and rising education costs.



This table shows the wide range of tuition costs in the dataset. The mean tuition is about \$13,027.72, but the median is lower at \$10,203.50, suggesting that most schools charge less than the average. The large gap between the minimum (\$1,225) and maximum (\$49,152) shows a lot of variation between institutions, with a few very expensive schools pulling the average upward. Overall, these statistics highlight that tuition costs are unevenly distributed and vary greatly across different colleges.

Statistic <chr>	Value <dbl>
Mean	13027.72
Median	10203.50
Minimum	1225.00
Maximum	49152.00

Key Insights

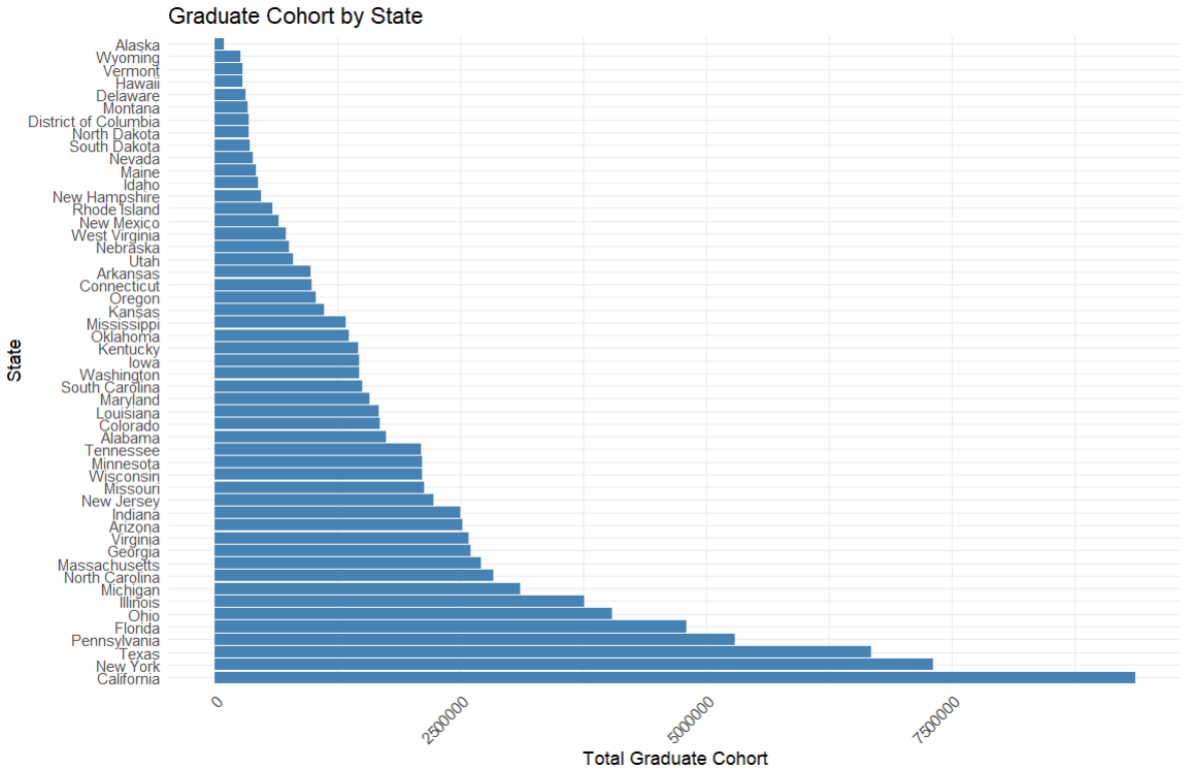
- Tuition costs have risen constantly.
- The upward trend of tuition increase can introduce additional factors that can be explored.

Figure 2 Visualization

This bar chart shows the total number of college graduates by state.

Larger states like California, New York, and Texas have the biggest graduate cohorts

Smaller states like Alaska, Wyoming, and Vermont have far fewer graduates.



Key Insights

- States with higher population density have a higher graduation cohort.
- These states have more post secondary school attending population

Figure 3 Visualization

The visualization emphasizes the significant differences in graduate numbers across states, which reflects differences in population size, number of colleges, and higher education enrollment.

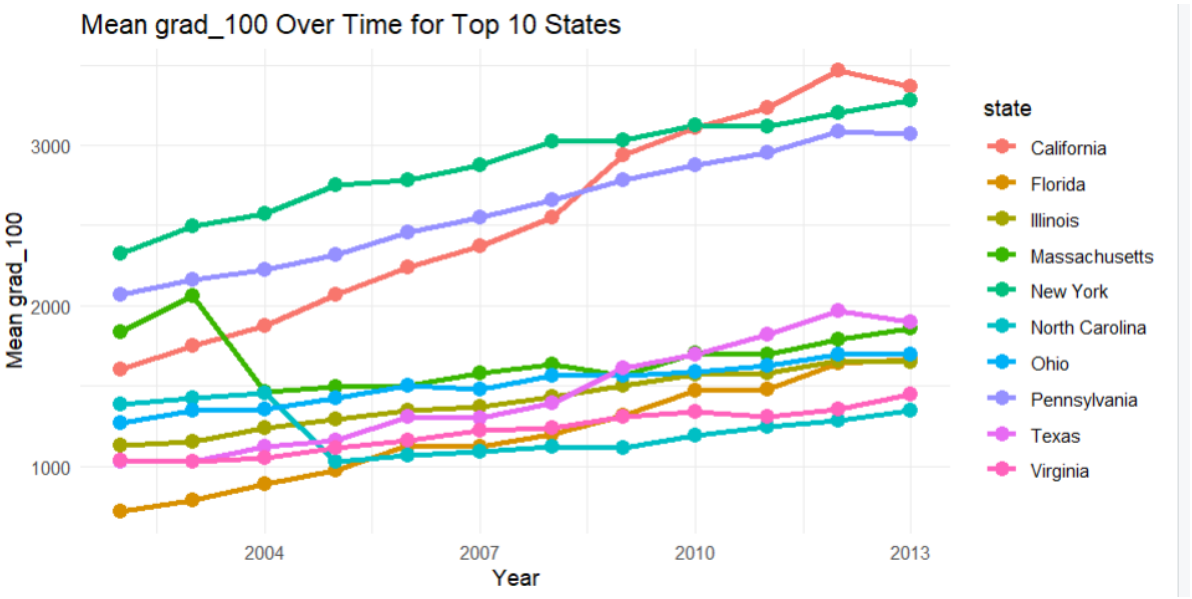
The graduation rates for the ten largest U.S. states by student cohort size showed mixed results

The top two remained pretty constant throughout the 10 year period.

California and Florida experienced a lot of variation.

Most of the rest of the states stayed constant as well.

Texas, while ranking on the lower end, actually had a big increase in graduate rates.



Key Insights

- Overall, there is a steady graduation rate throughout the span over this decade. However, there are several states that have substantial increase and decreases that can be explored further.

Conclusion:

From our findings, we found that the price of college tuition has been steadily increasing and continues to rise.

Tuition costs vary widely across colleges, with most schools charging below the average of $13,027.72 and a few expensive outliers driving up the mean.

We took a look at college completion rates and the top three states were California, New York, and Pennsylvania.

We also analyzed college completion rates for the top 10 states over time and got mixed results. Some rates went up, some went down, and some stayed the same.

Overall, these findings suggest that while tuition continues to rise and varies greatly across institutions, college completion rates show diverse trends across states, highlighting the complex relationship between cost and educational outcomes.

References/Sources

Chirumamilla, B. (2023, February 9). Average cost of undergraduate student by State USA. Kaggle.

https://www.kaggle.com/datasets/bhargavchirumamilla/average-cost-of-undergraduate-student-by-state-usa/data

Devastator, T. (2022, December 6). College completion dataset. Kaggle.

https://www.kaggle.com/datasets/thedevastator/boost-student-success-with-college-completion-da?select=cc_institution_grads.csv

Code Appendix

```
library(tidyverse)
```

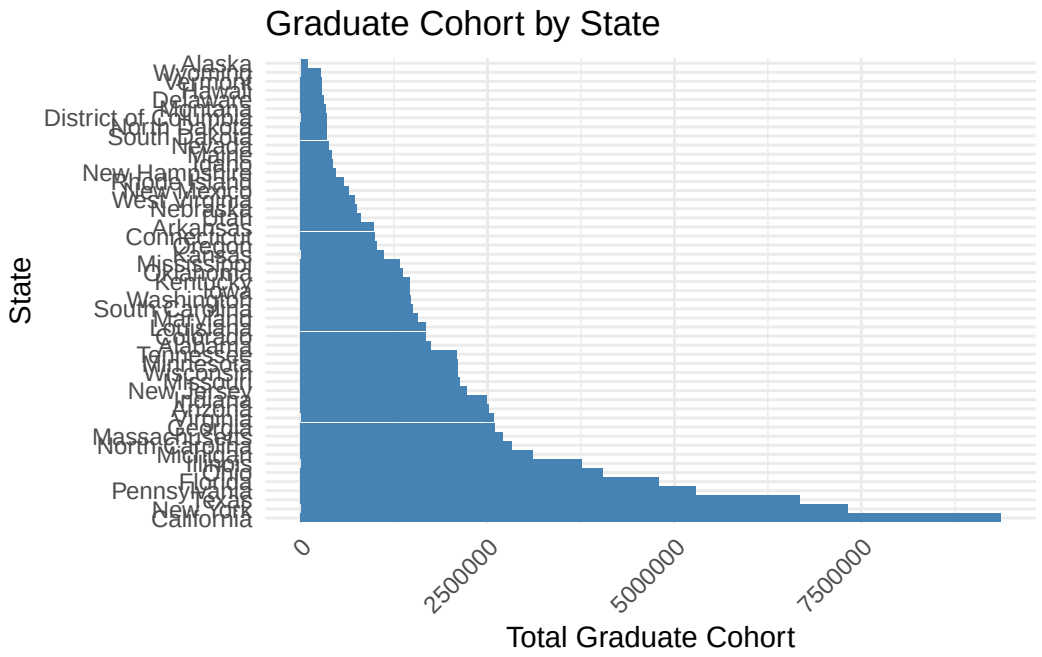
```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
```

```
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```r
library(ggplot2)
echo=FALSE
GradSectors <- read.csv("C:/Users/ryanw/Downloads/GradSectors.csv", row.names=NULL)


state_summary <- GradSectors %>%
    group_by(state) %>%
    summarise(total_grad_cohort = sum(grad_cohort, na.rm = TRUE)) %>%
    arrange(desc(total_grad_cohort))

# View the summary
print(state_summary)
```

```
# A tibble: 51 x 2
   state          total_grad_cohort
   <chr>                      <int>
 1 California               9366070
 2 New York                 7313694
 3 Texas                    6682776
 4 Pennsylvania             5292298
 5 Florida                  4795470
 6 Ohio                     4039884
 7 Illinois                 3757246
 8 Michigan                 3110316
 9 North Carolina           2833390
10 Massachusetts            2708542
# i 41 more rows
```

```r
ggplot(state_summary, aes(x = reorder(state, -total_grad_cohort), y = total_grad_cohort)) +
    geom_col(fill = "steelblue") +
    coord_flip() +
    labs(title = "Graduate Cohort by State",
         x = "State",
         y = "Total Graduate Cohort") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

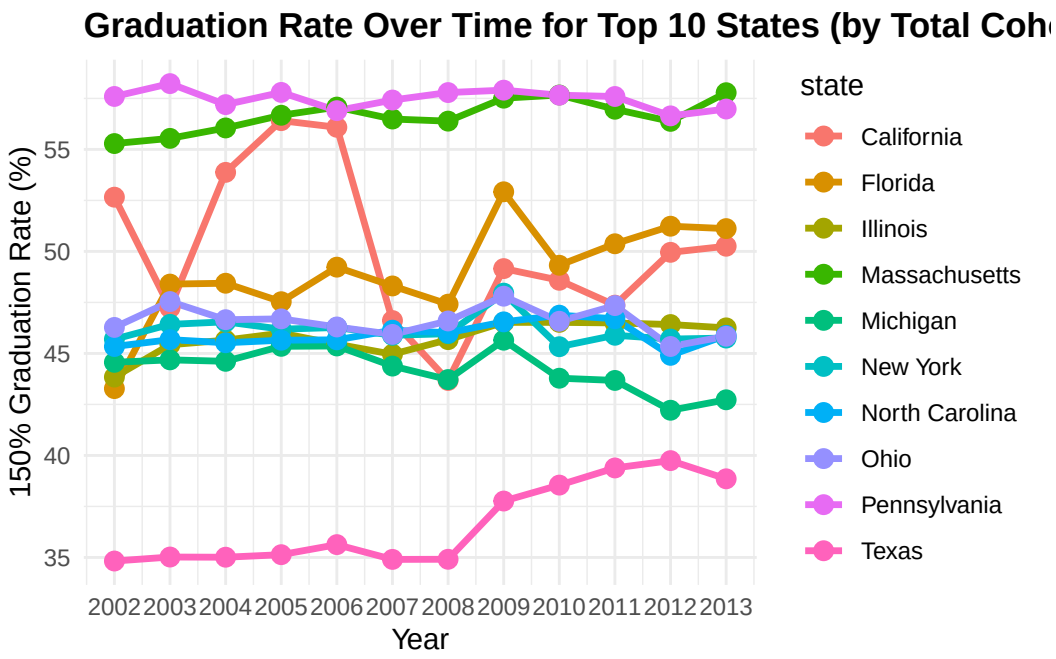



```
state_totals <- GradSectors %>%
    filter(!is.na(grad_cohort), !is.na(grad_150), grad_cohort > 0) %>%
    group_by(state) %>%
    summarise(
        total_grad_cohort = sum(grad_cohort, na.rm = TRUE),
        .groups = 'drop'
    )

top_10_states <- state_totals %>%
    arrange(desc(total_grad_cohort)) %>%
    slice_head(n = 10) %>%
    pull(state)

# Calculate Yearly Graduation Rate for Top 10 States
yearly_rates_top_10 <- GradSectors %>%
    filter(state %in% top_10_states) %>%
    filter(!is.na(grad_cohort), !is.na(grad_150), grad_cohort > 0) %>%
    group_by(state, year) %>%
    summarise(
        total_yearly_cohort = sum(grad_cohort, na.rm = TRUE),
        total_yearly_grad_150 = sum(grad_150, na.rm = TRUE),
        .groups = 'drop'
    ) %>%
```

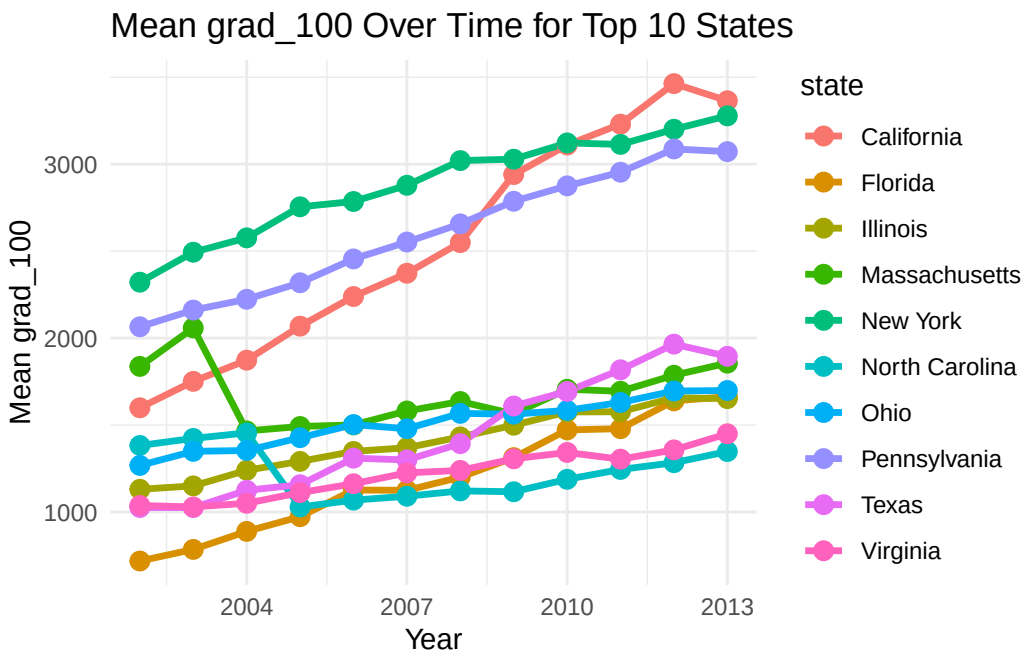
```
    # Calculate the aggregate 150% graduation rate
    mutate(
        yearly_grad_rate = total_yearly_grad_150 / total_yearly_cohort * 100
    )

# 4. Create Plot
plot <- yearly_rates_top_10 %>%
    ggplot(aes(x = year, color = state, y = yearly_grad_rate)) +
    geom_line(size = 1.2) +
    geom_point(size = 3) +
    labs(
        title = "Graduation Rate Over Time for Top 10 States (by Total Cohort Size)",
        x = "Year",
        y = "150% Graduation Rate (%)",
    ) +
    scale_x_continuous(breaks = unique(yearly_rates_top_10$year)) + # Ensure all years are di
    theme_minimal() +
    theme(
        plot.title = element_text(face = "bold"),
        legend.position = "right"
    )
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
# Print the plot
print(plot)
```




```r
top_10_states <- GradSectors %>%
    group_by(state) %>%
    summarise(overall_mean_grad_100 = mean(grad_100, na.rm = TRUE)) %>%
    arrange(desc(overall_mean_grad_100)) %>%
    slice_head(n = 10) %>%
    pull(state)

# 3. Calculate Yearly Means for the Top 10 States
plot_data <- GradSectors %>%
    filter(state %in% top_10_states) %>%
    group_by(state, year) %>%
    summarise(mean_grad_100 = mean(grad_100, na.rm = TRUE), .groups = 'drop')

# 4. Create the Line Plot
ggplot(plot_data, aes(x = year, y = mean_grad_100, color = state)) +
    geom_line(size = 1.2) +
    geom_point(size = 3) +
    labs(
        title = "Mean grad_100 Over Time for Top 10 States",
        x = "Year",
        y = "Mean grad_100"
    ) +
    theme_minimal()
```




```
library(ggplot2)
library(tidyverse)

data <- read.csv("C:/Users/ryanw/Downloads/nces330_20.csv")
n_rows <- nrow(data)

head(data)
```

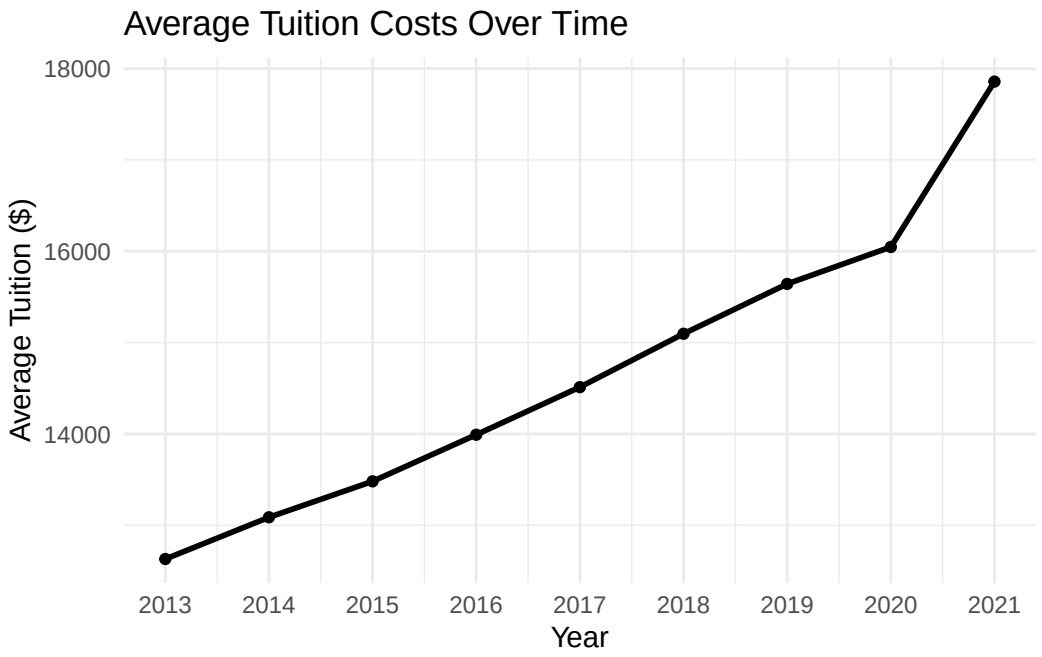
```
  Year   State                Type Length      Expense Value
1 2013 Alabama             Private 4-year Fees/Tuition 13983
2 2013 Alabama             Private 4-year   Room/Board  8503
3 2013 Alabama     Public In-State 2-year Fees/Tuition  4048
4 2013 Alabama     Public In-State 4-year Fees/Tuition  8073
5 2013 Alabama     Public In-State 4-year   Room/Board  8473
6 2013 Alabama Public Out-of-State 2-year Fees/Tuition  7736
```

```
data_tuition <- data %>%
  filter(Expense == "Fees/Tuition")

head(data_tuition)
```

```
  Year   State                Type Length      Expense Value
```

```
1 2013 Alabama             Private 4-year Fees/Tuition 13983
2 2013 Alabama     Public In-State 2-year Fees/Tuition  4048
3 2013 Alabama     Public In-State 4-year Fees/Tuition  8073
4 2013 Alabama Public Out-of-State 2-year Fees/Tuition  7736
5 2013 Alabama Public Out-of-State 4-year Fees/Tuition 20380
6 2013  Alaska             Private 4-year Fees/Tuition 21496
```

```r
data_tuition$year <- as.numeric(data_tuition$Year)

avg_cost_by_year <- data_tuition %>%
  group_by(Year) %>%
  summarise(avg_tuition = mean(Value, na.rm = TRUE))

ggplot(avg_cost_by_year, aes(x = Year, y = avg_tuition)) +
  geom_line(size = 1) +
  geom_point() +
  theme_minimal() +
  labs(
    title = "Average Tuition Costs Over Time",
    x = "Year",
    y = "Average Tuition ($)"
  ) +
  scale_x_continuous(
    breaks = seq(
      min(avg_cost_by_year$Year),
      max(avg_cost_by_year$Year),
      by = 1
    )
  )
```




```
stats <- data.frame(
  Statistic = c("Mean", "Median", "Minimum", "Maximum"),
  Value = c(
    mean(data$Value, na.rm = TRUE),
    median(data$Value, na.rm = TRUE),
    min(data$Value, na.rm = TRUE),
    max(data$Value, na.rm = TRUE)
  )
)

stats
```

```
  Statistic    Value
1      Mean 13027.72
2    Median 10203.50
3   Minimum  1225.00
4   Maximum 49152.00
```