Tuition Cost in College and Analysis

Data Bootcamp - CU Boulder

Xingyu Chen

August 06, 2021

Abstract

The cost of college tuition has been one of the most important topics for both high school students and pre-college programs. It is necessary to have a data product that shows how tuition costs change between US universities. In my project, I want to perform research on the cost of college tuition based on several different categories such as in-state tuition vs out-state tuition; School type: public, private,and for profit. The dataset will also include degree length and state name so that I can plot a map view of the data product and compare cost by year. I added another dataset contains Date and Time data so I can plot a graph about Tuition Difference based on year. I will use different charts generated with R along with related R library and analyze the result. At the end of my project, I will come to a conclusion about tuition statistics for US colleges and make recommendations for high school students and pre-college students.

Contents

List of Figures

1 Introduction

1.1 Motivation

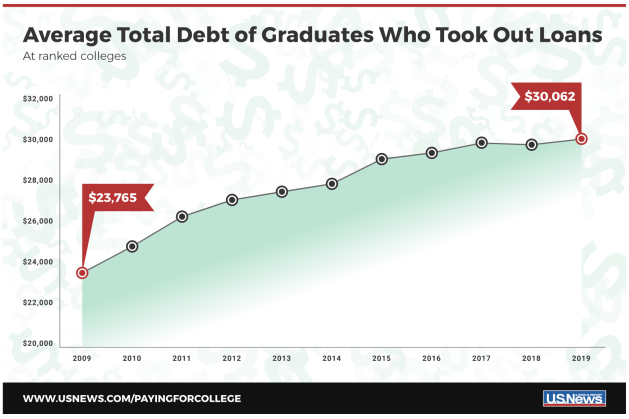

Figure 1: Average Total Debt of Graduates Who Took Out Loans

The outstanding college student loan debt has reached an all-time high of $1.41 trillion in 2019, average $30,000 for each college students according to **Figure 1** from U.S News report on student loan, which demonstrated by the graph above showing 10 Years of average total student loan debt. There is no clear guideline application for pre-college student or transfer students to search state-wide college tuition data as an important reference before applying for U.S. College.

The main motivation behind this project is to give a reference for them when they starting to apply for U.S. College, potentially avoid their regrets of endure high student loan debt after entered dreaming college for years. I wants to build an interactive application and report on U.S. college tuition which present clean and clear graphs and conclusion based on user's options such as Degree-length, School-Type, and In/Out-State.

Secondary motivation in hindsight will hopefully be able to grab attention from United States Department of Education (ED) so that they can make some policies to slow down the increasing trend of college tuition fee and flat the curve of outstanding college student loan debt, which is definitely help a lot to the sustainable development of U.S. college.

1.2 Related Applications

There are several open source tables and application that related to my project. Thanks them for these data models, they provide some inspirations to my subtopic, like **Figure 2** from U.S News report on Average College Tuition, which show me a view of average college Tuition in 2020-2021. In **EducationData** website, I am able to view a list of table about average cost of college with analysis such as "why expensive?", "Cost by State", and "Room and Board On and Off Campus". More importantly, it has a nice work flow to discuss all the plots based on research which I can learn from it.

The **TuitionTracker** web application is powered by U.S. Department of Education data from

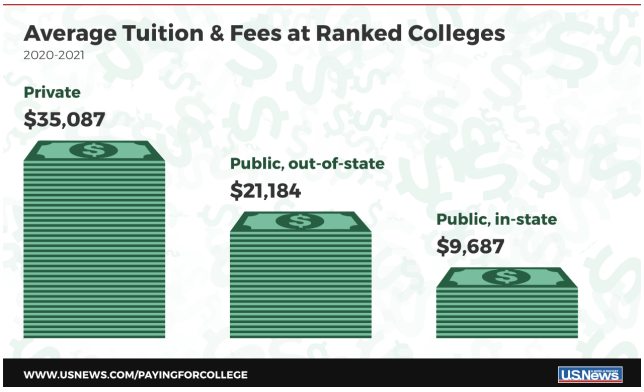

Figure 2: Average Tuition & Fees at Ranked Colleges

IPEDS(Integrated Postsecondary Education Data System, a service provided by the National Center for Education Statistics). This interactive web tool shows what students really pay for college based on their family income with more than 3,800 colleges and universities in the United States.

Another online application called **CollegeTuitionCompare** provide a nice bar plot about in-state vs out-of-state tuition comparison. The application also include the graduation rates and diversity which help parents and students compare colleges using data that shows what schools actually cost the year student expect to enroll. They have highlighted the essential data such as financial aid and application stats so important data are transparent to users.

1.3 Observations and Questions

From the existing web applications and tables, I find most of these data products having too much information even somewhat messy. The main problem is that they do not have options for users to choose what kind of information they want to view and compare so they just pop up a long form report including all the details.

For example, it do give essential data such as Tuition difference between in-state and out-of-state, graduation rate, admission stats, and etc. However, it also include Not that important data such as student to faculty ratio, dormitory capacity, and average earning after 10 years of graduation with salary range.

In my project, my analysis will focus on essential information for pre-college student and transfer student such as tuition difference between in-state and out-of-state, tuition increasing graph based on yearly report, and tuition difference between school type, and etc. In this way, my report will be more efficient and useful compared to the existing research based on U.S. college tuition.

My hypothesis is to explore which state has the most expensive tuition and which region has the most selective for pre-college student and college student to apply for college across the country.

2 Design of Data and Methodology

2.1 Data Resource and Explanation of Variables

The aim of data collection definitely focus on if the dataset contains state and tuition amount column, school-type column, and tuition-type column. Ideally, I want to a dataset contains yearly report for each college so I can plot a date trend series graph based on that dataset. College tuition data is somewhat difficult to find - with many sites limiting it to online tools.

Thanks to **Chronicle of Higher Education**, I were able to get a table of information about tuition and fees of nearly 3,000 colleges/universities from the 1998-99 to the 2019-20 academic year, along with school type, degree length, state, in-state vs out-of-state as shown in Table 1.

Table 1: The related-variables in my data set.

Related-Variable	Description
name	School name
state	State name
state_code	State abbreviation
type	School type: public, private, for-profit
degree_length	4 year or 2 year degree
room_and_board	Room and board in USD
in_state_tuition	Tuition for in-state residents in USD
in_state_total	Total cost for in-state residents in USD
out_of_state_tuition	Tuition for out-of-state residents in USD
out_of_state_total	Total cost for out-of-state residents in USD
year	Academic year
tuition_type	Tuition Type: All Constant/Current, 2/4 Year Degree Constant/Current
tuition_cost	Tuition cost in USD

These related-variables are documented and will be used for my research after reshaping and analyzing the original dataset. Having a data product on college tuition is useful and necessary for pre-college or high school students who want to have a point of reference when they apply for college.

The data for this project was obtained via the websites mentioned above and compiled into one CSV file to be read into R. Although the data set is good, it is missing some values and state information. So, after I read the CSV file, I use the complete.case function to remove empty spots inside the data set and use the usmap library to filter the state name column of the data set.

The data used in this project to produce data plots can be found in the **Kaggle** dataset website. This website provides the yearly college tuition data for the United States. This website provides reshaped data to some extent which is originally from the National Center

for Education Statistics (**NCES**) - spanning the years 1985 - 2016. From this site, enough data was obtained to create two interactive and informative data plots.

2.2 Preparing the Data

Reshape original data to readable data set is a major part of this report. I spend a lot of time to extract data using several different functions of library and package in R.

The data set originally comes from the US Department of Education. The most comprehensive and easily accessible data cames from TuitionTracker.org who allows for a .csv download! Unfortunately it's in a very wide format that is not ready for analysis, but tidyr R library can make quick work of that with pivot_longer().

It has a massive amount of data, I have filtered it down to a few tables as seen in the attached .csv files. Tuition data can be quickly joined by dplyr::left_join(tuition_cost, diversity_school, by = c("name", "state")). Some of the other tables can also be joined but there may be some fuzzy matching needed.

The Tuition and fees by college/university for 2018-2019, along with school type, degree length, state, in-state vs out-of-state from the Chronicle of Higher Education.

The Historical averages from the National Center for Education Statistics (NCES) - spanning the years 1985 - 2016.

Before I use this data set to explore the costs of college tuition in the US on their own, by geographic area, degree type, I have applied functions to reshape the data set. Here is the summary of the process:

1. Drill in on the seemingly most popular regions using the "include" parameter in the plot_usmap() function. Regional divisions can be found in the doc of us_regdiv.pdf here.

2. There were lots of missing data for room and board fee which will be added to total tuition fee in the end so I remove these rows when I want to compare these in-state tuition total and out-of-state tuition total.

3. The original dataset do have a column to showing the states name. In order to simplify the output, I use merge anther column called Abbreviation to make room for the map view of US college tuition in state wise.

4. Due to I use two different data set in this report, I only use couple columns in the history data set in order to not duplicate the information such as room and board, college name, state name are duplicated in these two data sets.

5. Original history data set do not use typically academic year but in the format such as 2008-09 format, I change them to 20xx year format so that I can plot data year by year.

6. In order to compare the tuition data in state wise, I have to calculate the average tuition for each state because there might have more than 50 college in one state but less than 10 college in another state.

7. I have a docs for each columns field for these two data sets. In order to illustrate them to readers, I create a table for each columns names and descriptions for that to help understand.

8. In order to zoom in the area of my findings, I use filter function to only display the corresponding information. For example, the highest tuition region I only display specific region instead of whole US map.

9. Showing all the column names and explain each column to readers is hard but I use colnames function to get all column names first. Then I use cbind to append explanation and description. Finally, showing the table using kable function.

2.3 R Library Foundations of the Project

This project may be imported into the RStudio environment and compiled by researchers wishing to reproduce this work for newest plot with future data sets, and having new findings or discussions from that.

> The Core of Statistics were done using R 4.1.0 (R Core Team, 2021-05-18), the ggplot2 (v3.3.5; RStudio Team, 2021-06-25), and the knitr (v1.33; Yihui, 2021-04-24) packages.

For **ggplot2** package, this package has been used for creating graphics such as box plot, line plot, bar plot, and density plot from my reshaped data sets. With built-in theme, I are able to generate decent plots map variables to aesthetics with details to present in this report.

From **knitr** package, this report is constructed to have reproducibility that it can regenerate the plot based on the latest dataset contains yearly report in the future. Using literate programming techniques for dynamic report generation in R.

> The Initial Scenarios package is usmap 0.5.2 (Paolo Di Lorenzo, 2021-01-21).

For **usmap** package, I use plot_usmap(based on ggplot object) to plot state wise US map in convenient. The map data frames include Alaska and Hawaii conveniently placed to the bottom left, as they appear in most maps of the US. More over, it built-in function from U.S. Census Bureau help us to filter and divide the whole US map into four regions, which help us to zoom in specific area to analyze.

> The Most Frequently Used package is dplyr (v1.0.7; RStudio Team, 2021-06-18).

For **dplyr** package, I use a lot of functions to reshape my data set and working with data frame through mutate, filter, arrange, group_by, summarize_all functions. Most of modification to the data set were using these functions list above.

- **Note**: There are few functions has been used to reshape the data set in other packages. Due to limited usage, I will not list all these packages in this report such as melt function in reshape2 package; ggarange function in ggpubr package; readPNG function in png package.

3 Exploration on US College Tuition

Several data plots were generated in order to effectively convey tuition information in different ways. The data plots allow the user to easily view the data.

3.1 State Wise Tuition Scenarios

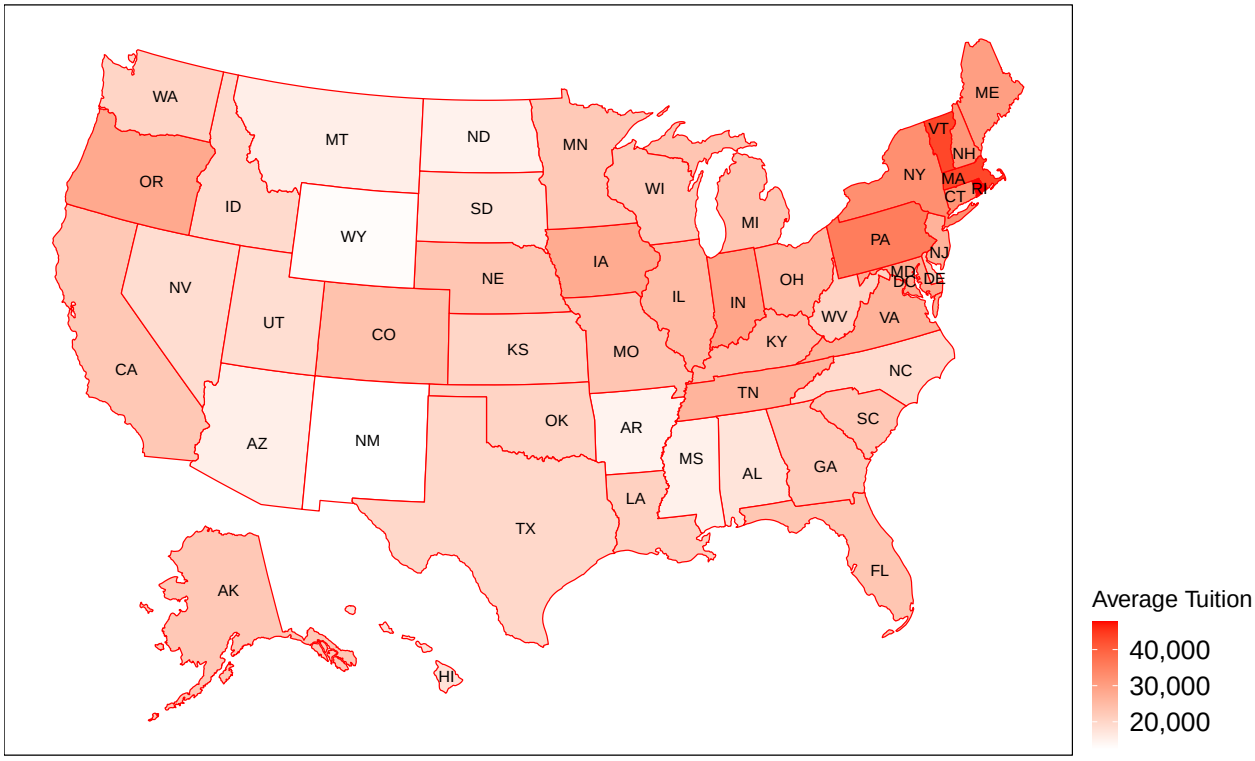

Figure 3: Average tuition per state. The spectrum from white to red indicates an increasing (Expensive) tuition

In order to get a nice model for my data set, I use **plot_usmap** function from usmap package to generate US map by state geographically and add settings for theme to help users to compare tuition between two states by color. The spectrum from white to red indicates an increasing (Expensive) pattern from lower to higher tuition.

Note that this is the image at one instance of time (in this case, the 2018-19 academic year). There are 50-state (including Alaska and Hawaii) United States thematic map in the Figure 3, with map scale and with state abbreviations.

For the given variables: state abbreviation and tuition amount, I use these two column as x-value and y-value for plot_usmap function. The user can view average tuition totals by state name instead of in-state or out-of-state only.

From this plot, I have a more **intuitive** perception to the data set I are trying to analyze. This is the starting point for the next couple of plots.

3.2 Region Wise Tuition Scenarios

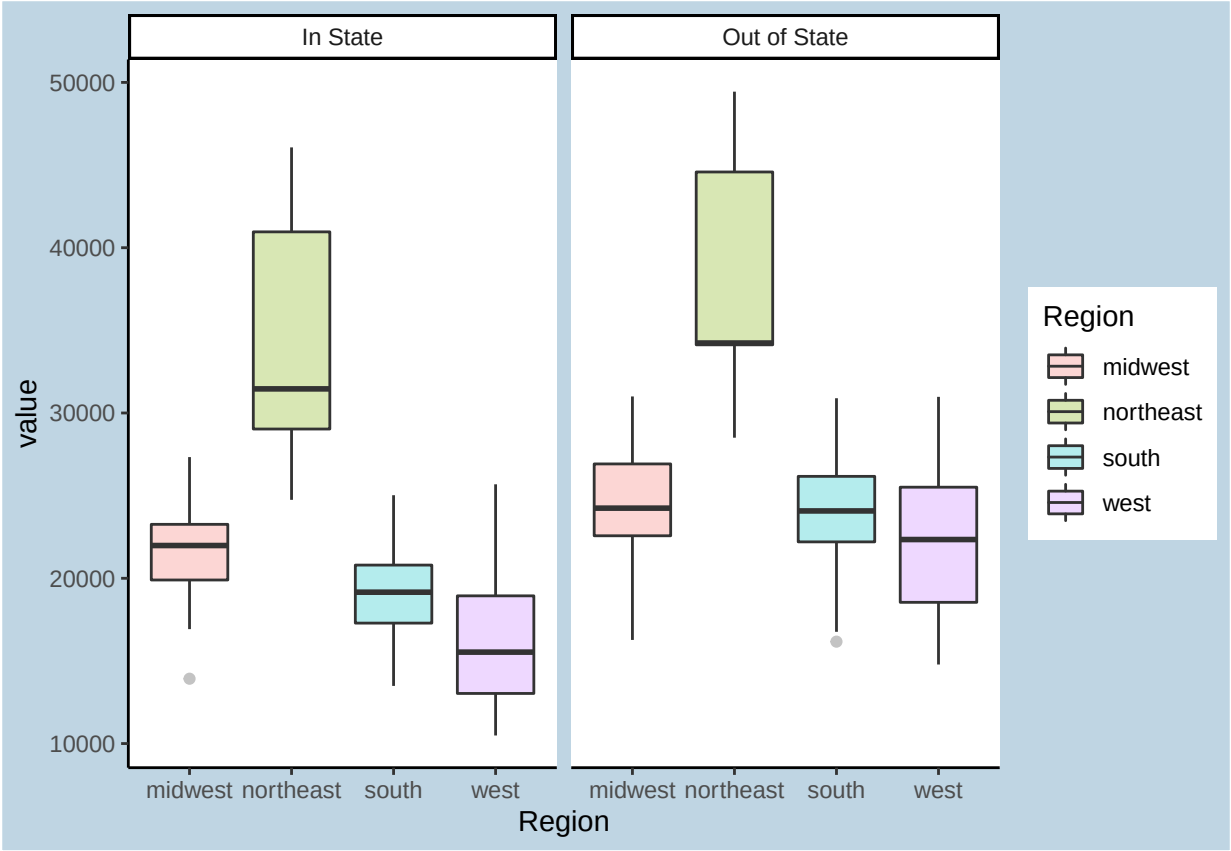

Figure 4: Average tuition per region. Show the lowest/highest/Median value in region

In general, people would like to know the tuition data in a specific area (most likely near their home) instead of viewing 50 states' data in one graph. In this box plot, I separate the whole data into four areas: NorthEast, West, MidWest, and South for better comparison. I use built-in function defined by U.S. Census Bureau in the usmap library to filter the original data set.

This box plot of college tuition data allows the user to compare a State's average in-state tuition to its average out-state tuition. This plot makes it easy for the user to see the **mean, maximum and minimum** values of college tuition. More over, with facet with in In-state and Out-of-state, it is easy to see that out-of-state tuition is generally higher than in-state tuition.

Besides the difference between in-state and out-of-state tuition, I find that the **West** region has relatively the lowest average tuition (around **$16,248** in-state) for college meanwhile **NorthEast** is the most expensive region (around **$34,001** in-state) in general, even the tuition fees vary widely with the biggest gap in northeast region.

3.3 Comparison between In-state and Out-of-State

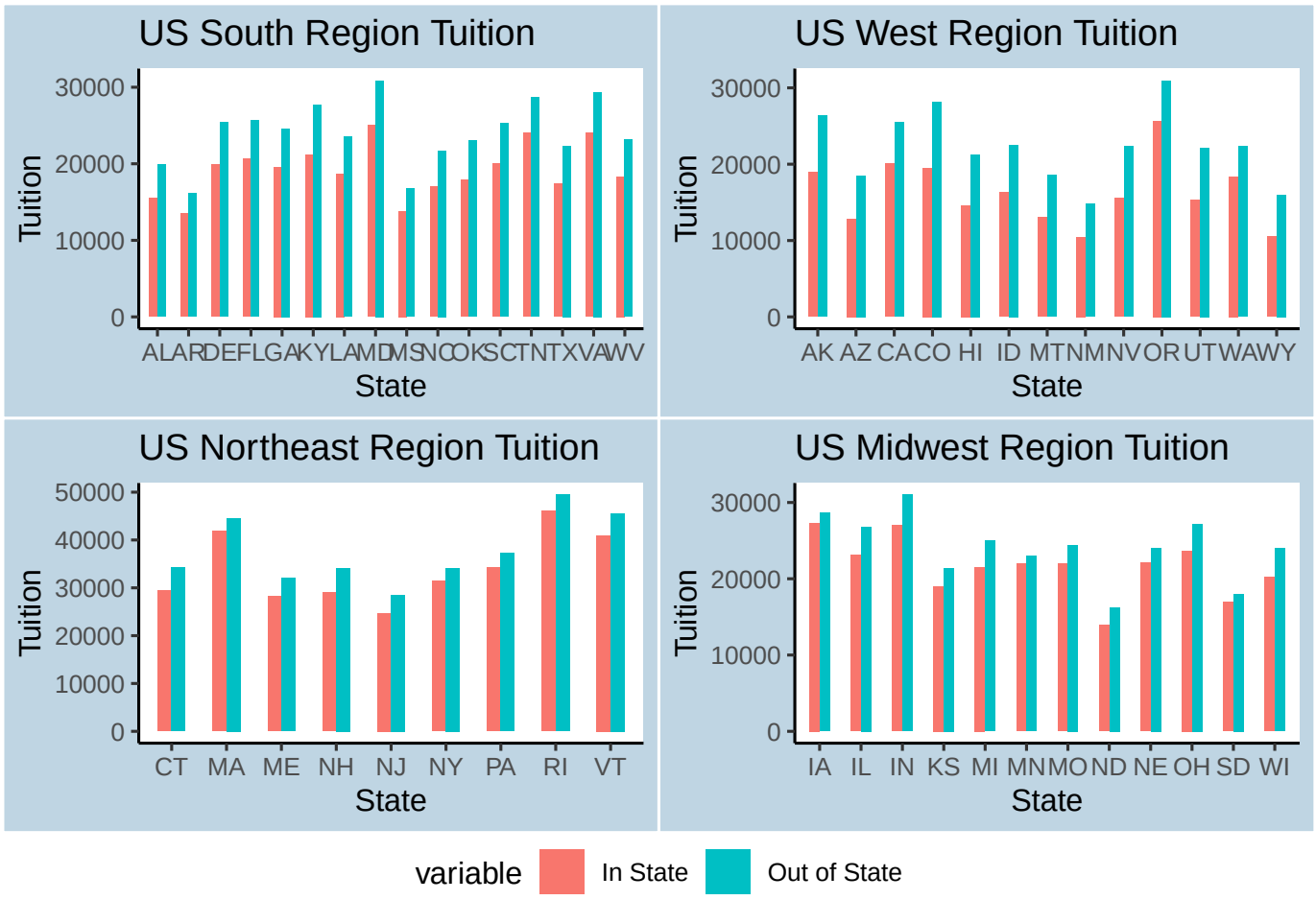

Figure 5: Tuition Difference for In-state vs Out-State. The red bar in-state and blue bar out-state

Last plot I divide whole country to four regions and make analyze for each region. I continue my study on data sets with this facet but focus on the difference between in-state and out-of-state tuition. I want to find which state has the most and the least gap between these two variables.

From this bar plot, people can easily view the difference for specific state and compare states among each region directly. The biggest difference is **Colorado** state with **$8,637** difference ($19,523 in-state vs $28,161 out-state) and minimal difference is **Minnesota** state with **$985** difference ($21,975 in-state vs $22,960 out-state).

This plot is not just benefit U.S citizen who cares about in-state tuition (Red Bar) for specific state, but also a good reference for the international student who applying for US college to see the differentiation between in-state and out-of-state. There are also some features in this plot that I have already found in previous scenarios.

3.4 Trends of Tuition based on school-type and tuition-type

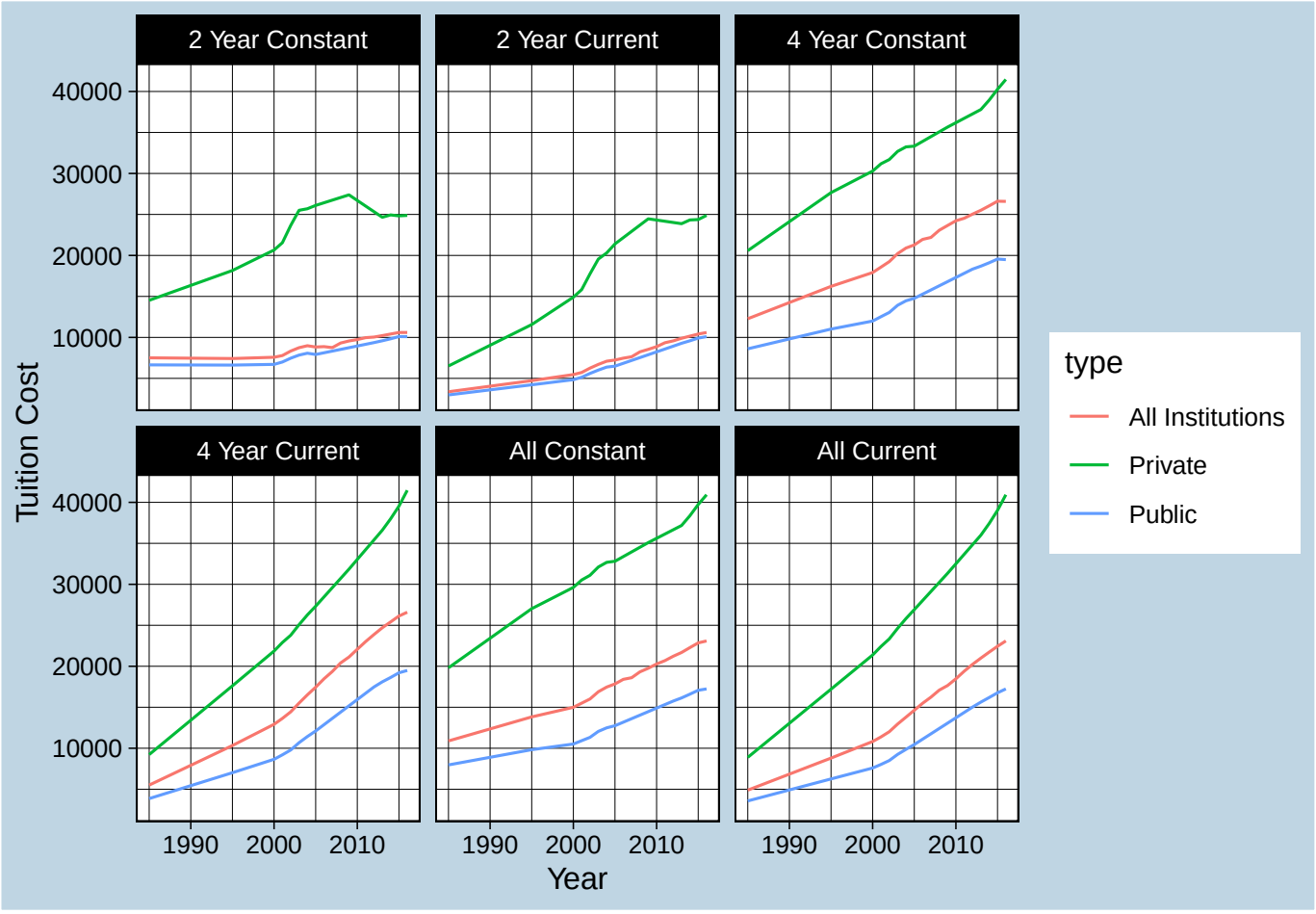

Figure 6: US College Tuition from 1980 to 2017. Observe the rise in college tuition.

When I accomplish the analysis of the tuition data set on the 2017-18 academic year, I want to generate a date series plot to show the time trend for tuition change over years based on history_tuition data set. I facet with school_type and color with tuition_type to show different tuition levels in this date and time series plot.

According to the chart, it is not difficult to see the increasing trend of college tuition. In the numerical experiment, the tuition for private college has been **numerical double** from **$21,373** (in 2000) to **$40,925** (in 2016) for only about sixteen years. Although all-institution and public college-type is not as crazy as private college, it rise from **$21,373** (in 2000) to **$40,925** (in 2016).

Constant dollars based on the Consumer Price Index, prepared by the Bureau of Labor Statistics, U.S. Department of Labor, adjusted to an academic-year basis. In general, constant tuition is inferior to current tuition. When it terms to tuition cost per academic year, 2-year tuition is less than 4-year and all tuition type. I would explore the college type at following discussion.

4 Result and Discussion

4.1 The Highest Tuition Region

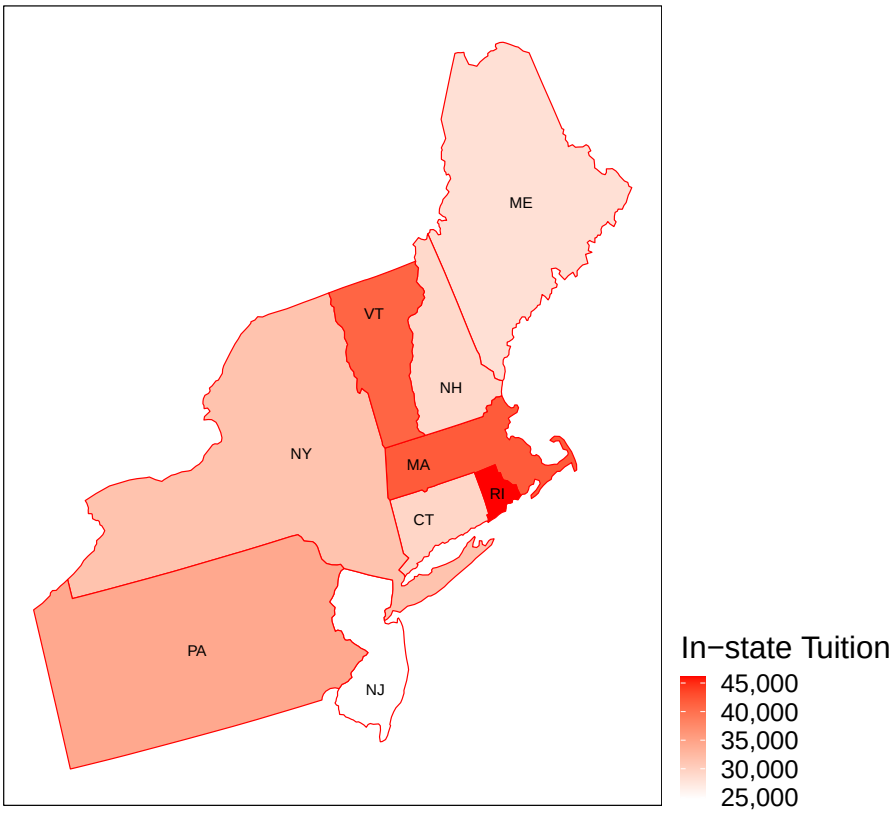

Figure 7: Average In-state tuition in NorthEast region. The highest in-state tuition area is RI, MA, and VT states

The State Wise Tuition Scenarios (Figure 3) clearly illustrates the variation in tuition costs. However, I want to find out which region has relatively high tuition cost in general so I use built-in function to only present NorthEast region. In this way, I can figure out which state in this region has high tuition cost which then lead to the highest tuition region in U.S.

Despite what I originally thought about some major cities and States could have relatively high tuition cost such as Seattle, New York City, and California, the college tuition there does not seem to be strongly higher than that of other major cities or States. Instead, **Rhode Island, Massachusetts, and Vermont** have most deepest color in this region.

Based on my research on these specific three states, there are several private college with expensive tuition cost in these three areas such as Brown University in Rhode Island with **$70,326**, Amherst College in Massachusetts with **$71,166**, Middlebury College in Vermont with **$69,980**. These private college have a long history and liberal arts college type make them much more expensive in tuition cost.

4.2 The most selective college region

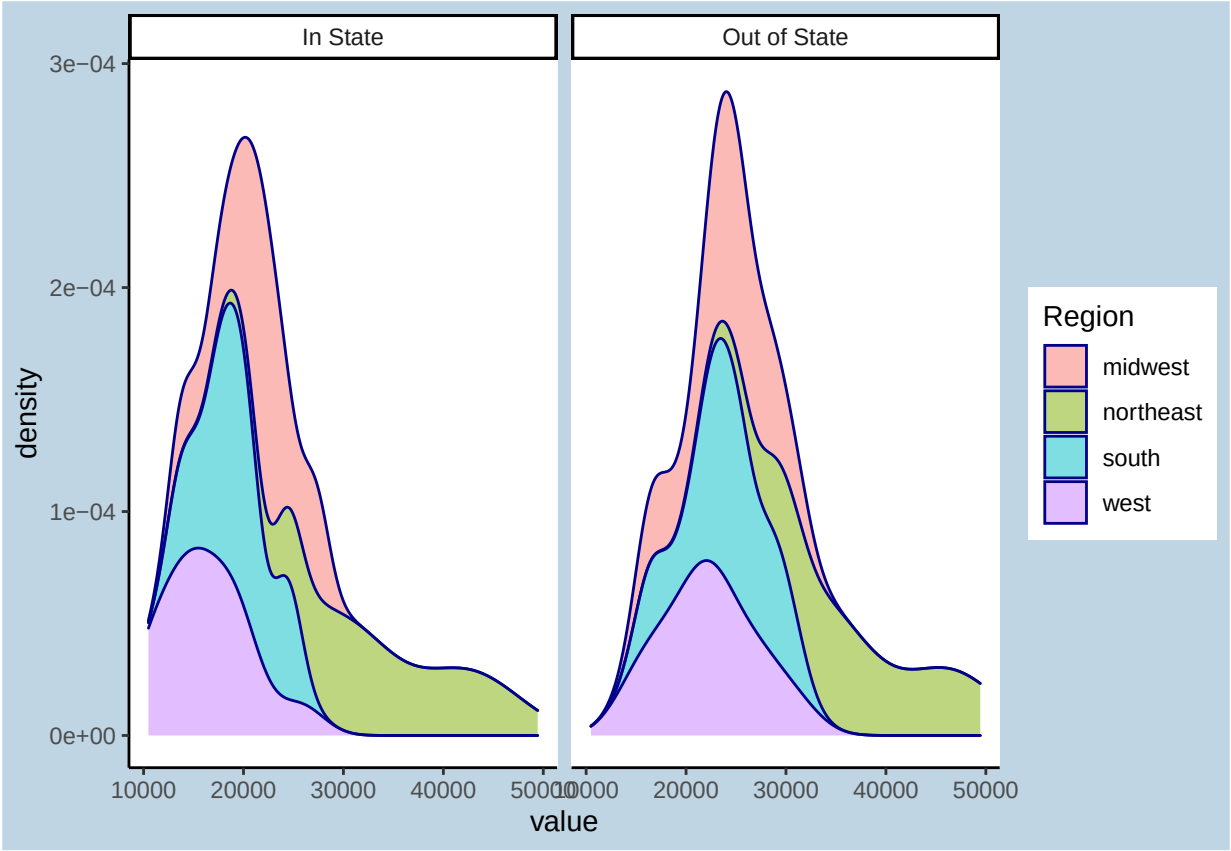

Figure 8: US college tuition density distribution. The midwest is the most selective region

The Region Wise Tuition Scenarios (Figure 4) is my first try to divide whole state into four different regions, which present the min/max/median value for each region. In this density plot, I want to figure out the distribution of tuition cost and features for each region so I choose geom_density from ggplot2 package.

Clearly, the most selective college region is MidWest because it is at the top of the chart, which means it has more college options in this region and tuition gap between each college is minimal all over the country.

On the other hand, the West region has relative low tuition cost but limited options and NorthEast has relative high tuition cost and large tuition gap between each college in this region.

4.3 The top 10 expensive US college state

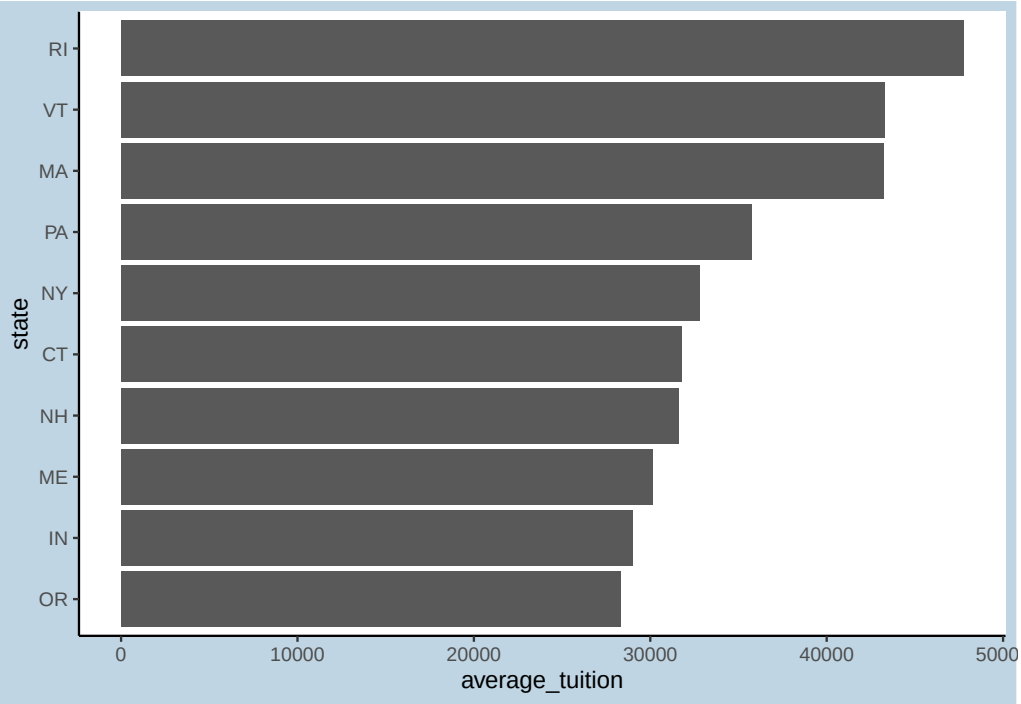

Figure 9: Top 10 expensive US college state. Check Average Tuition Cost for each State

The In-state vs Out-of-State tuition different for each state (Figure 5) show users which state has the most and the least gap between these two variables. I want to continue analyzing my reshaped data but focus on the Top 10 expensive state report, which help readers to take note of that when they search college in specific state.

I first observe that from above bar plot that Rhode Island (RI), Massachusetts (MA), and Vermont (VT) appears the top 3 expensive state and more exaggerated than others in terms of tuition cost and which I already discuss in previous chapter sections.

The rest of states in the plot is for reference only because it does not mean that there is no affordable college tuition in these stats. From this plot, I do not find enough features so I choose to create a model based on history data set in the following section.

4.4 Trend and distribution of US college tuition

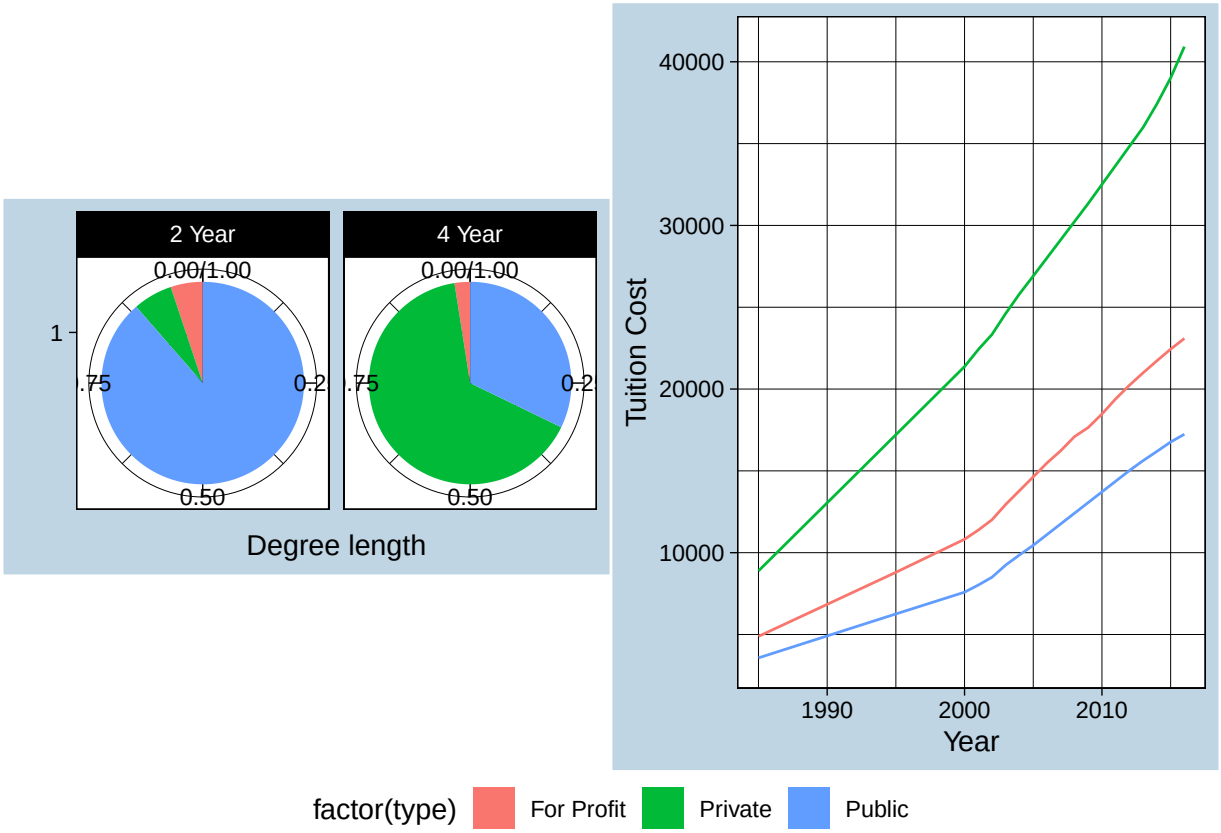

Figure 10: Trend and distribution of US college tuition. Left shows school-type classification meanwhile Right shows general trend

I analyze some features in The Trends of Tuition based on school-type and tuition-type (Figure 6), especially for tuition-type. In this pie plot and time series plot, I want to focus on history data set on school type so that I are able to tell the difference among them.

The left plot is school-type classification type(Private, Public, and All Institution). I can see most of public college provide two-year degree length study program and Private college provide more four-year degree length study program than the other two. From this pie plot, I can see college type proportion contribute to my data set.

Similarly, I have an conclusion that the tuition cost keep increasing from 1980 to 2017 academic year. More over, I observe that private tuition is two times bigger than public school in general (Private in 1985: $8,885 vs Public in 1985: $3,571; Private in 2016: $40,925 vs Public in 2016: $17,237).

The All-institution college type is not much difference between public college type. And I see a inflection point at 2002 academic year then the college tuition is increasing in linear and it is likely to have a exponential increase for private college-type.

5 Problems Tackled and Conclusion

Achievement Adjust:
I have learned the shiny application, plotly library, D3.js to make some of my graphs interactive and deploy the package into external server, which in the course material in Week 10th. However, due to time limits and other technique unfamiliar, I decide to abandon this idea for this semester currently and to see if I can accomplish this function in the future.

These plots are useful references for pre-college programs or senior students in high school who want to apply for US college based on tuition data. It will be also useful for college students who want to transfer to other college at the end of sophomore year. I have also learn a lot technique about RStudio and Data Analysis when I generate this report emblem with graphs and tables.

For the 2017–18 academic year, annual current dollar prices for undergraduate tuition, fees, room, and board were estimated to be $17,797 at public institutions, $46,014 at private nonprofit institutions, and $26,261 at private for-profit institutions. Between 2007–08 and 2017–18, prices for undergraduate tuition, fees, room, and board at public institutions rose 31 percent, and prices at private nonprofit institutions rose 23 percent, after adjustment for inflation.

Besides the trends for U.S. college tuition, I find that the most selective region is MidWest meanwhile the most expensive region is NorthEast. The Top three expensive states are Rhode Island (RI), Massachusetts (MA), and Vermont (VT) because of private college with a long history and liberal arts college type make them much more expensive in tuition cost.

The future extension (if someone in community want to develop more features and write discussion based on it) is that I wish to have a interactive map using the Shiny application, Google APIs and the D3.js, so that users can click on the state name to view more details about college costs, such as room and board fees and degree length. Thus, making this report interactive just like a search engine (TuitionTracker.org) is my plan for future implementation.

6 References

[1] College Tuition and Fees 1998-99 through 2018-19 - Chronicle of Higher Education, https://www.chronicle.com/article/tuition-and-fees-1998-99-through-2018-19/

[2] Historical Tuition for college - Kaggle Data website, https://www.kaggle.com/jessemostipak/college-tuition-diversity-and-pay?select=historical_tuition.csv

[3] Tuition costs of colleges and universities - National Center for Education Statistics, https://nces.ed.gov/fastfacts/display.asp?id=76

[4] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/, 2020

[5] Yihui Xie knitr: A general-purpose package for dynamic report generation in R, http://yihui.name/knitr/, 2020

[6] Compare colleges using data - Tuition Tracker, https://www.tuitiontracker.org/, 2020

[7] Census Regions and Divisions of the United States - U.S. Census Bureau, https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf, 2020

[8] Easy way to mix multiple graphs on the same page - ggplot2 package, http://www.sthda.com/english/articles/24-ggpubr-publication-ready-plots/81-ggplot2-easy-way-to-mix-multiple-graphs-on-the-same-page/

[9] Tuition costs of colleges and universities - U.S. Census Bureau, https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf, 2020

[10] Trends in the cost of college education - National Center for Education Statistics, https://nces.ed.gov/fastfacts/display.asp?id=76, 2020

[11] Figures, Tables, Captions - R Markdown for Scientists, https://rmd4sci.njtierney.com/figures-tables-captions-.html, 2020

[12] Yang Liu ggplot US state heatmap - usmap package, https://liuyanguu.github.io/post/2020/06/12/ggplot-us-state-and-china-province-heatmap/, 2020