Data151 Final Writeup

Tyler Bontrager, Ganesh Singh

2022-12-13

DATA151 Final Writeup:

An Analysis of Trends in College Tuition

Abstract

Introduction

For our Final Project Submission we decided to submit a write up of our analysis of a Data Set that showcased the cost of attendance varying with different states, types of institutes and some other factors. The Data Set got its data from tuitiontracker.org and we picked it up from the tidy tuesday github repository. Our Project is a combination of data frames and graphics which showcase what we learned not only about R, but about the Data set in general.

Dataset

Our dataset, linked here, consisted of 5 CSV tables related mostly by school name. It contained information ranging from tuition and housing prices, salary potential for graduates from different schools, student body makeup (race, gender, STEM major status, etc.), and other information about the school itself.

Interest

As college students, the interest was primarily comparing the cost of attendance given certain factors to determine which factor might be the most influential to how much a student could expect to be paying for higher education. Tidy Tuesday makes an explicit disclaimer that the compilation of their datasets do not inherently assume correlation or causation, though this investigation may shed light on interesting trends that were discovered throughout the course of this data analysis and visualization project.

High-Engagement Variables

High-Engagement variables are variables that have been most interesting throughout this project and have yielded the most insight during our examination or would lend themselves well to future examination. These variables and their explanations may be found below:

Variable	Explanation and Interpretation
name	The name of the school. This variable is generally what was used to relate each table to other tables when necessary.
type	Whether the school was public/private or for/non profit.
degree length	Whether the school offered two- or four-year programs.
in/out state tuition	Two variables: what each student at a school could expect to pay in tuition
room and board	What each student at a school could expect to pay in room and board
in/out state total	Consists of tuition, room and board, and applicable fees.
early/mid career pay	Two variables: What graduates of a school could expect to make in their early and mid careers.
stem_percent	The percentage of the student body enrolled in a STEM major.
category	Demographic information like race and gender.
enrollment	Percentage of student body enrolled who identify with above variable.

Initial Thoughts

Overall, our initial thoughts were based on our own experience and perception of the reality of prices of higher education in the current day. Given the dataset, certain information and data were unavailable to us. Specifically, whether a school is has ivy-league status, school rating, and information about how much aid each student could expect as an award. It might be a stretch to expect that academic information such as GPA statistics be made available, but it would have been interesting to see what kind of insights we could derive from information like school-specific statistics (graduation rates, drop-out rates, percentage of students graduating magna cum laude, etc.) and other information.

Generally speaking, high school counselors will advise that public education at an in-state institution will be the cheapest any student could expect to pay for higher education. Further exploration will show that this may be the case, yet the question remains whether there may be other factors for students to consider.

After a preliminary and cursory glance through the variables and the dataset itself, more questions arose: would a school's degree length or the state where a school is based show any insightful trends? Might there be any influence between how much a school costs and factors like how much graduates are expected to make? Perhaps even enrollment percentage in STEM could correlate to a generally higher cost of attendance?

Exploration

Delving Deeper

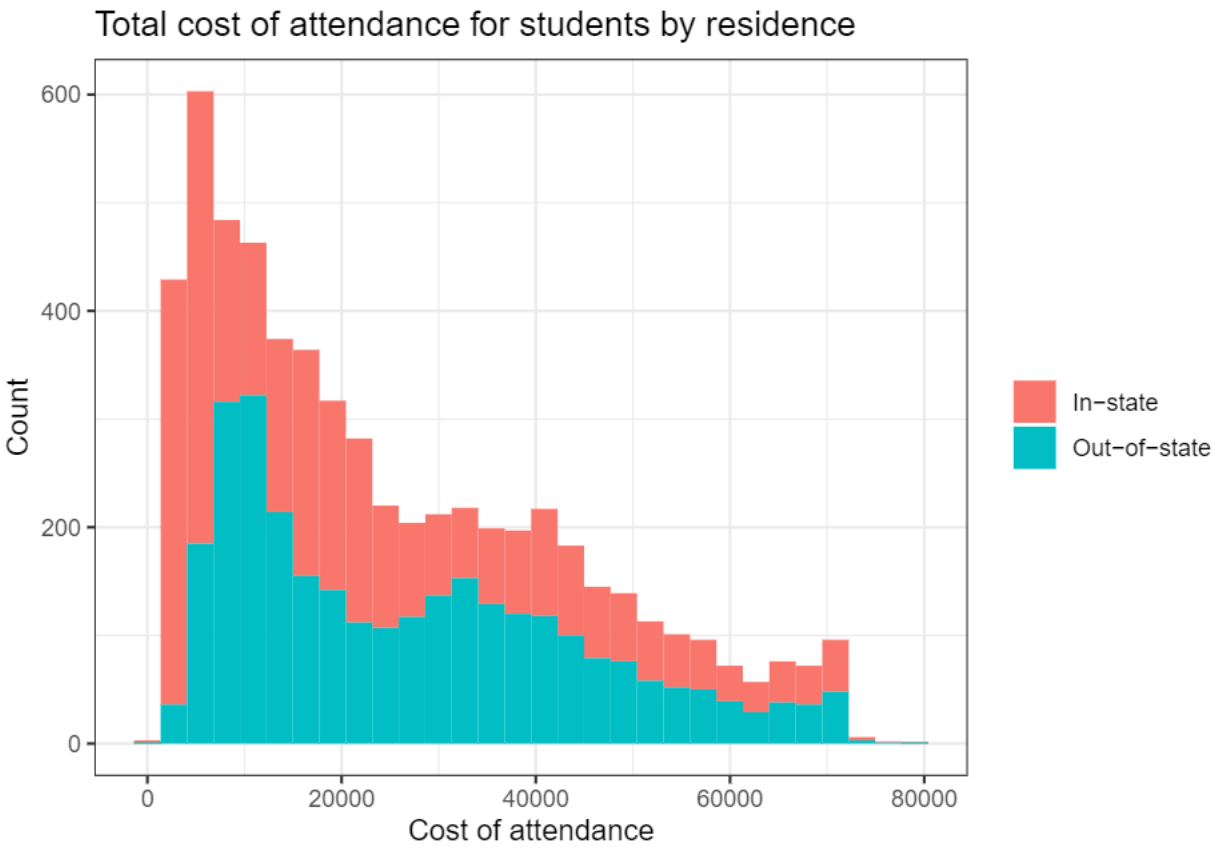
`str()` and `head()` were used on our imported datasets to learn more about them before working on wrangling or visualizing the data.`tuition_cost` was the most-referenced. Each row consists of a record of information pertaining to a unique school. Some of these schools were part of a university, that is each

campus of the same affiliated school is considered a unique school. However, there is no repeated information for any school.

Our first task was to tabulate trivial observations into tables. Our one-way table showed us that California has the most number of schools proportionally by state, where approximately 21% of the records in our `tuition_cost` dataset are schools in California, and the next state with the second-most school prevalence at 16% is New York. Table 1 contains all of this information and can be viewed in the appendix. Table 2 notes the number of schools

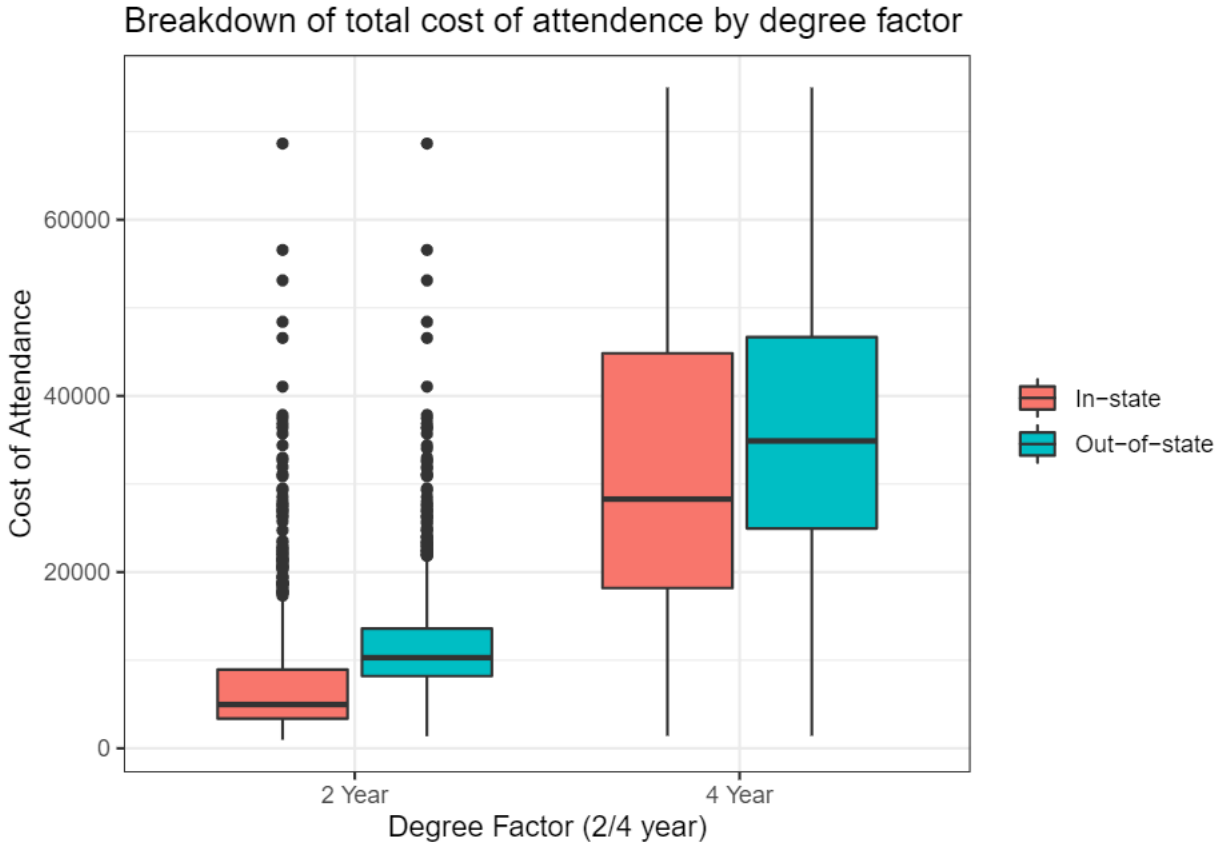
Analysis

Plot 1



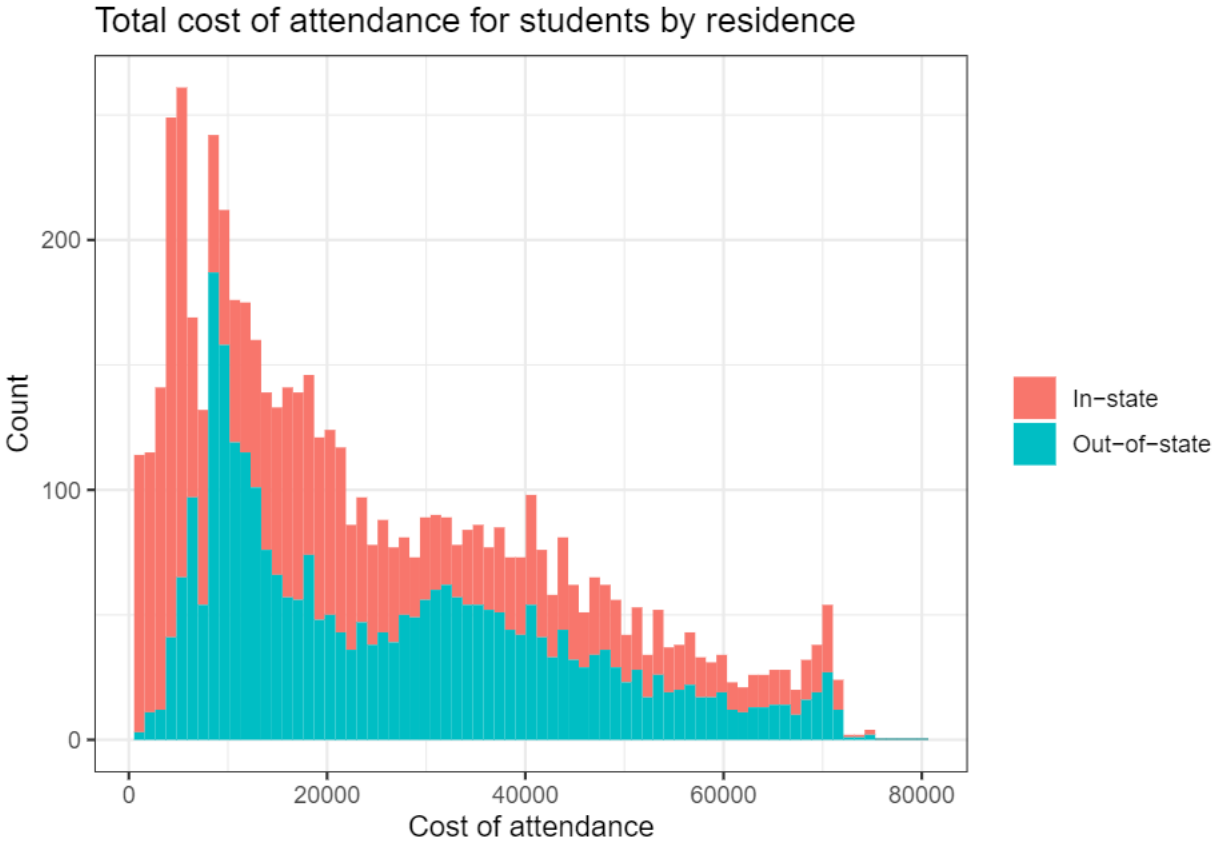
We noticed that the more common cost of attendance students could expect is generally lower than what we had initially thought. We hypothesized that the mode would be closer to $20,000. We then wanted to create a boxplot to compare what students at two- and four-year institutions generally pay and visualize if there are any differences. The plot is intuitively skewed to the right, as we expected that there are fewer schools that will cost students an arm and a leg to attend.

Plot 2



For 4-year institutions, it does not seem to be the case that there is a significant difference between costs of attendance for in-state and out-of-state students. This contradicts the insights derived from the left side of the plot where we see that there is a significance between in-state and out-of-state residence and its effect on cost of attendance for students. We wondered why this could be the case, and suspect that the type of 4-year institution might be more influential to the price students see than whether they are in-state or out-of-state residents. We omit the 5-number summaries for these boxplots for the purpose of clarity and brevity, but they can be found in the appendix.

Plot 3

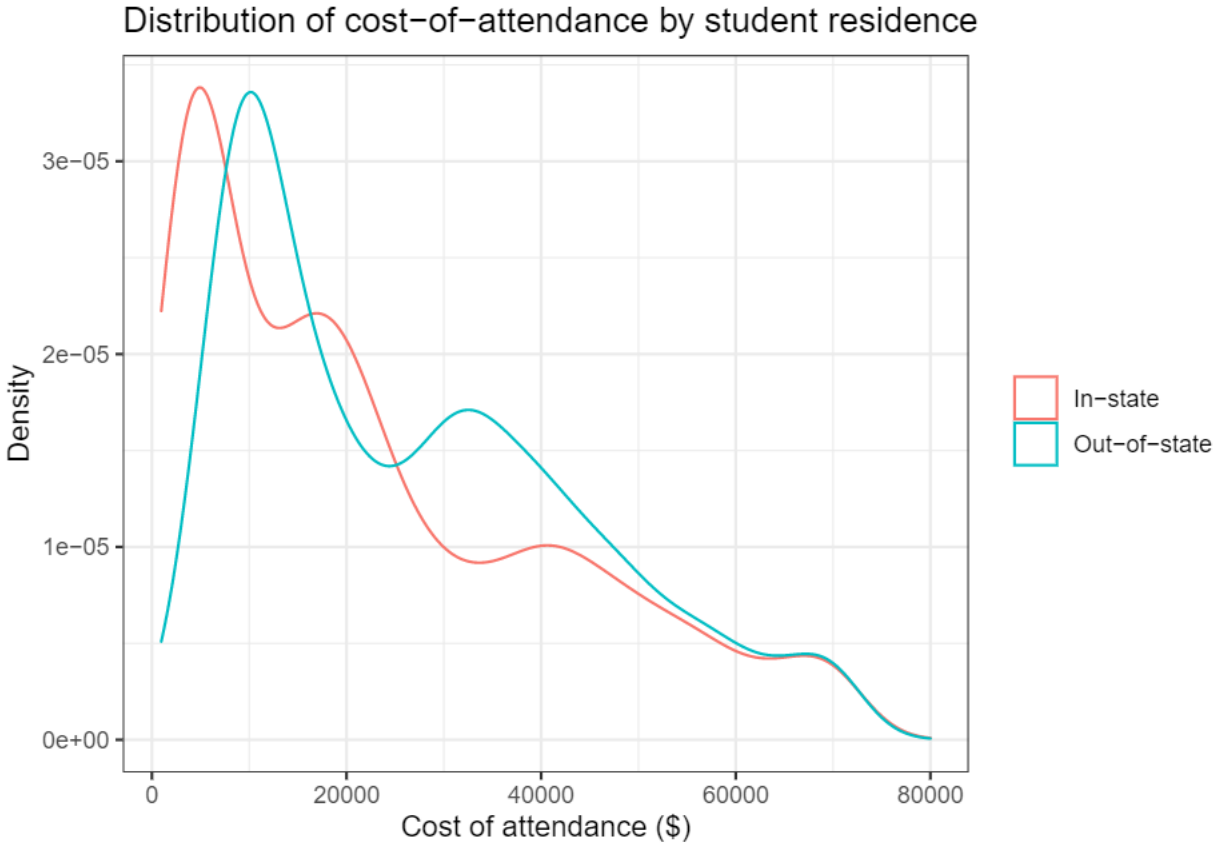


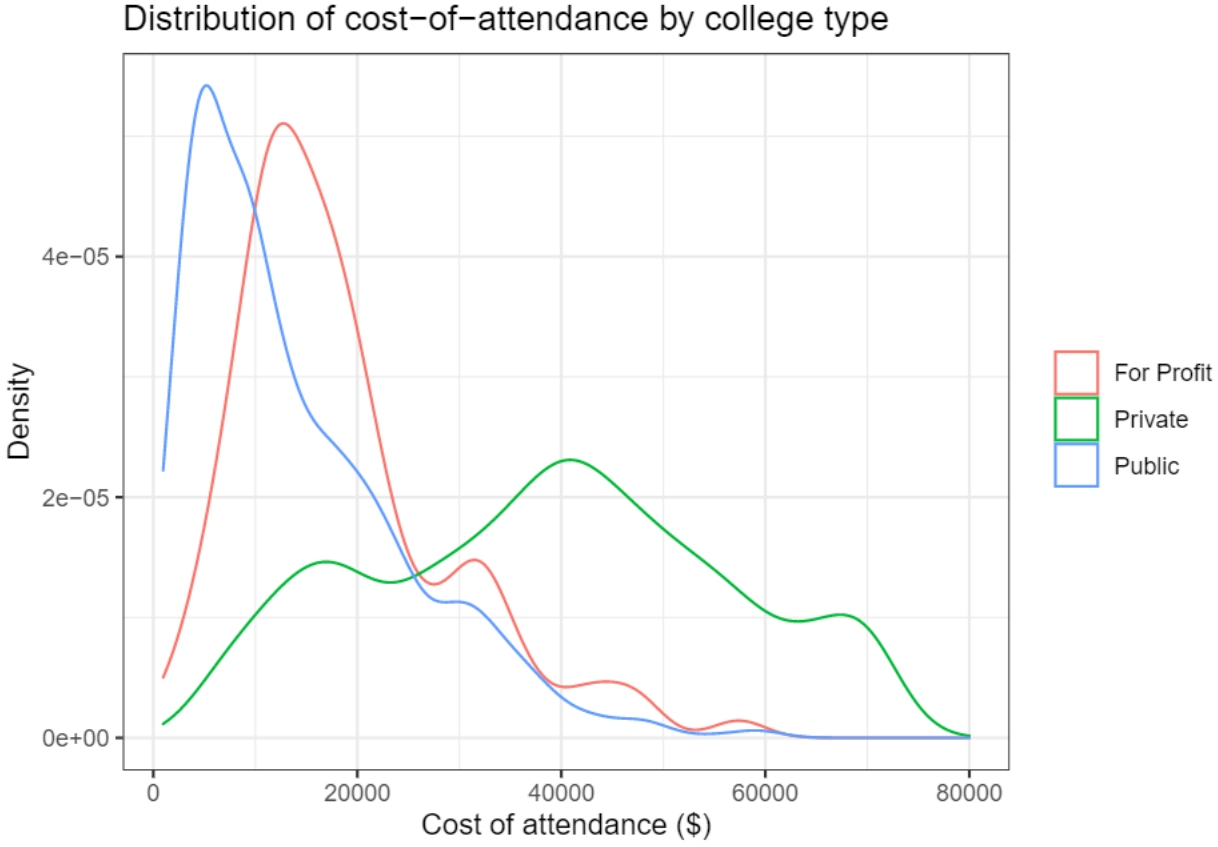
Upon closer inspection of plot 1, we noticed a slight bimodality phenomenon. When we increased the bin size, we were better able to see the shape of the plot.

We exaggerated a bit to better understand this distribution and the second mode became clearer. We started to wonder what the cause of this bimodality could be, and hypothesized that it must be due to school type. Exploring this route became crucial to making good progress on this investigation.

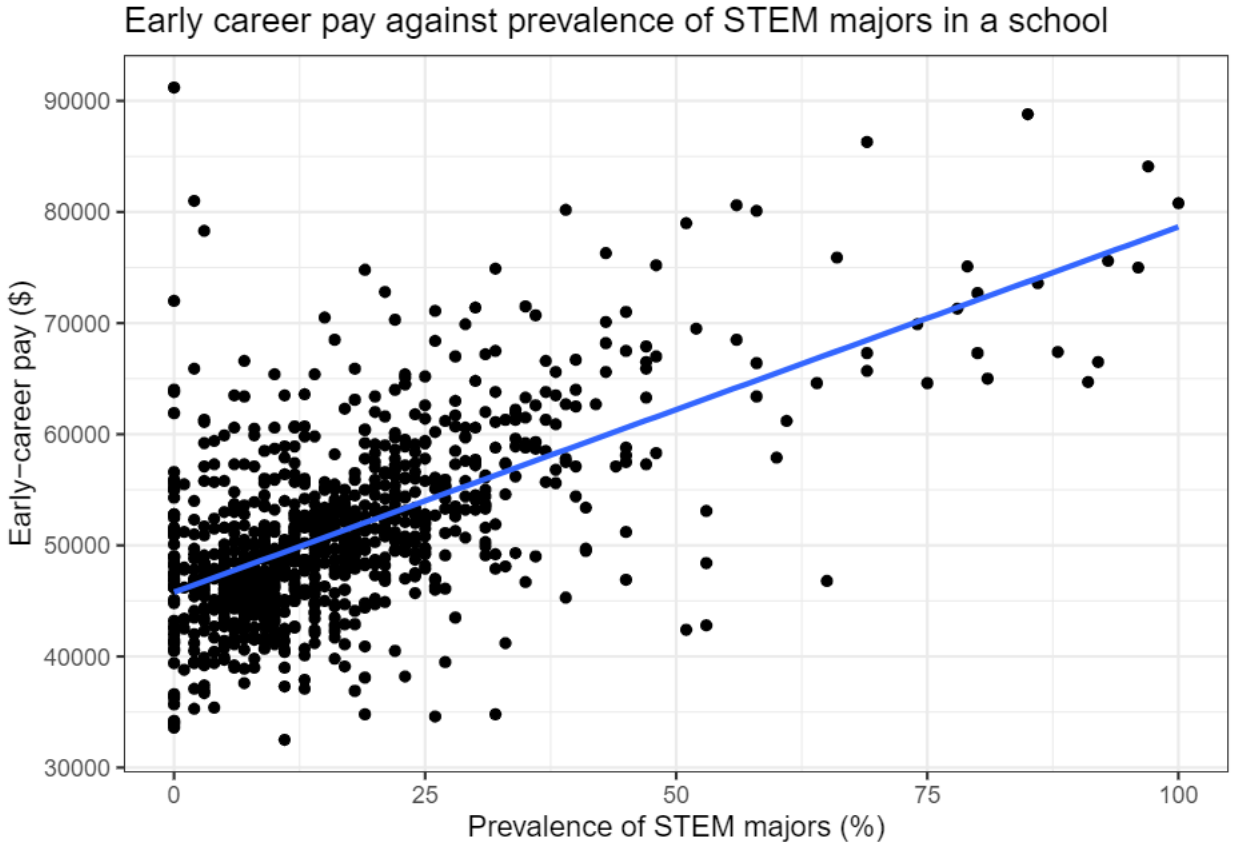
In order to undergo this exploration, we wanted to see the density plots.

Plotset 4


Distribution of cost−of−attendance by student residence
Density
3e−05
2e−05
1e−05
0e+00
0
20000
40000
60000
80000
Cost of attendance ($)
In−state
Out−of−state


Distribution of cost−of−attendance by college type
Density
4e−05
2e−05
0e+00
0
20000
40000
60000
80000
Cost of attendance ($)
For Profit
Private
Public

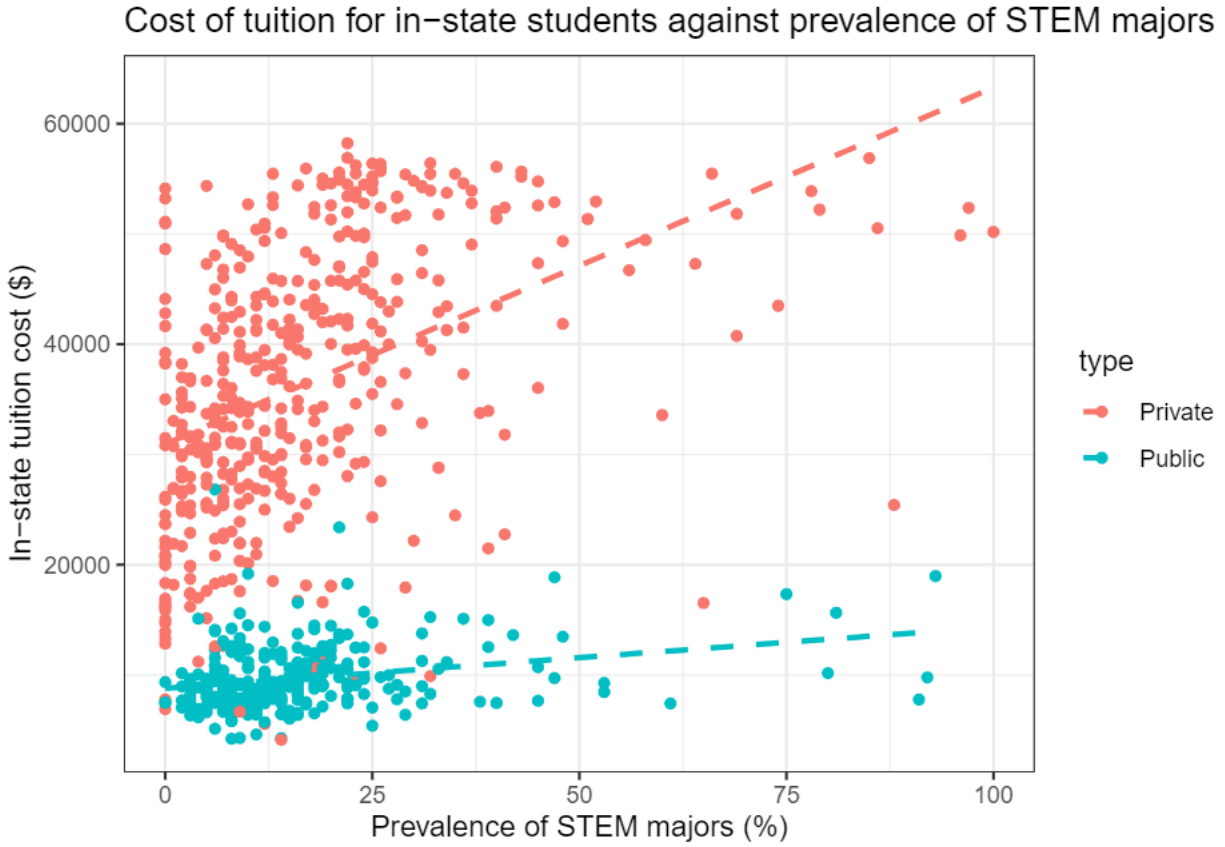
Plot 5



The second density plot is a clear explanation for the lack of significance we observed in our previous plot. Comparing the first density plot with the second shows that there is a distinct subgroup that can explain why certain schools tend to cost students more money to attend. While initially we may have thought that student residence was the most influential factor in determining how much a student would expect to pay to enroll, the plots show the rather obvious observation that private schools will tend to cost students more than public schools.

Next we thought about what insights we could draw from our `salary_potential` table. We hear a lot that STEM tends to be a lucrative fieldset, and we wanted to determine whether our dataset would support this hypothesis. Plot 5 shows a positive, moderately strong correlation between a school's STEM prevalence and graduates' expected early career pay with correlation coefficient $r^2 = 0.605$

Plot 6

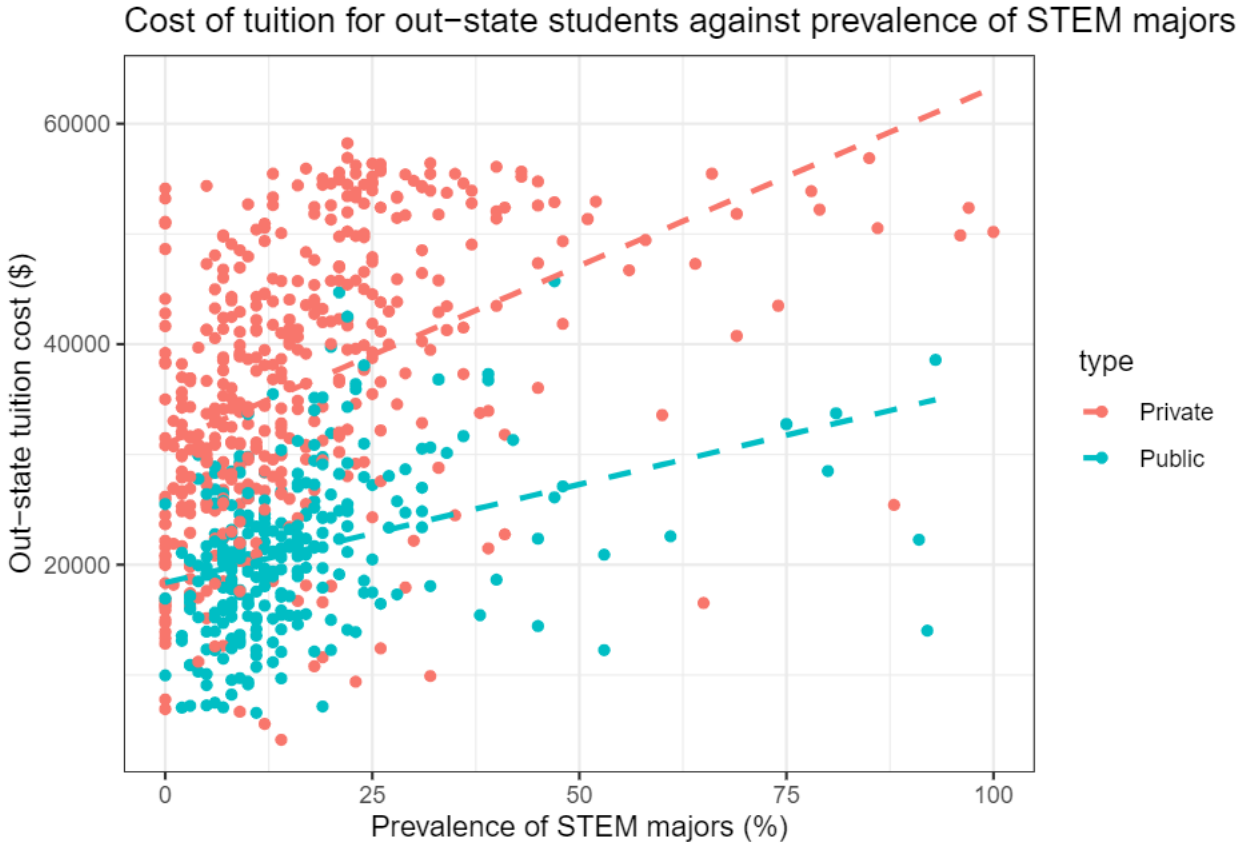


Our final question was a bit more adventurous: is there any correlation between how much students would expect to pay in tuition given a school's prevalence of STEM enrollment? our thinking was that if we assume STEM departments tend to cost more money due to the equipment and setups required, then students may ultimately bear the burden of the school's investment. Perhaps the more STEM majors there are at a school, the more the school will tend to spend on equipment for STEM.

As we can see from plot 6, there may exist some positive correlation in the dataset, but it is weak with correlation coefficient $r^2 = 0.433$ for private institutions and $r^2 = 0.265$ for public institutions. We might conclude from these insights that public schools are generally less volatile with tuition pricing than private schools, and that private schools may tend to charge their students more liberally, especially if the school has a high prevalence of STEM majors.

Further exploration will be required to discover other explanations for the spread seen in the points for the private schools subgroup.

Plot 7



Interestingly, after this final plot we derived the insight that out-of-state students generally tend to pay more volatile and less predictable prices for tuition. We are not sure exactly why this is the case, but it seems to support our initial hypothesis. The difference between this plot and plot 6 implies that we may have missed critical insight that could have been derived from a simple conditional distribution plot between school type and student residence.

Conclusion

Through this project and this data set we learned a lot this semester. We better understood the way in which college tuition differs based on residence, area, stem percentage and degree length. It was a wonderful exercise into handling our very own data set and our first foray into the world of a data scientist.

Acknowledgements

We would like to thank Professor Heather Kitada-Smalley for her assistance and guidance throughout this semester-long project, and our class for their continued and collaborative support.

Appendix

Table 1

State	Number	Proportion
California	258064	0.2105311
New York	195364	0.1593799
Pennsylvania	102400	0.0835389
Texas	88804	0.0724472
Ohio	64516	0.0526328
Illinois	62500	0.0509881
North Carolina	54756	0.0446705
Massachusetts	34596	0.0282238
Florida	30976	0.0252705
Georgia	24964	0.0203659
Virginia	24964	0.0203659
Michigan	24336	0.0198535
Missouri	21316	0.0173898
Minnesota	20164	0.0164500
Wisconsin	17956	0.0146487
Indiana	15376	0.0125439
Tennessee	15376	0.0125439
Washington	14400	0.0117477
South Carolina	12996	0.0106023
Alabama	11664	0.0095156
New Jersey	11664	0.0095156
Iowa	10816	0.0088238
Kansas	10816	0.0088238
NA	10816	0.0088238
Arkansas	8464	0.0069050
Maryland	8100	0.0066081
Kentucky	7744	0.0063176
Oklahoma	6400	0.0052212
Oregon	6400	0.0052212
Colorado	5776	0.0047121
Connecticut	5184	0.0042292
Arizona	4624	0.0037723
Louisiana	4624	0.0037723
Nebraska	4356	0.0035537
Mississippi	4096	0.0033416
West Virginia	3600	0.0029369
Maine	2916	0.0023789
New Mexico	2304	0.0018796
Montana	1936	0.0015794
New Hampshire	1764	0.0014391
Vermont	1444	0.0011780
North Dakota	1296	0.0010573
South Dakota	1296	0.0010573
Hawaii	784	0.0006396
Utah	784	0.0006396
Idaho	676	0.0005515
Rhode Island	484	0.0003949
Nevada	400	0.0003263
Delaware	324	0.0002643

State	Number	Proportion
Wyoming	256	0.0002088
Alaska	144	0.0001175

Table 2

Degree Factor	Number of Schools
2 Year	2240
4 Year	3704
Other	0

Codedump

```
tcFours = tcFactored %>%
  filter(degFactor=="4 Year")

tcTwos = tcFactored %>%
  filter(degFactor=="2 Year")
```

```
#4-year out-of-state
tc4YOOS_Summary = tcFours%>%
  summarise(count_4YOOS=n(),
            min=min(tcFours$out_of_state_total, na.rm=TRUE),
            Q1=quantile(tcFours$out_of_state_total, prob=0.25,na.rm=TRUE),
            med=median(tcFours$out_of_state_total, na.rm=TRUE),
            Q3=quantile(tcFours$out_of_state_total, prob=0.75,na.rm=TRUE),
            max=max(tcFours$out_of_state_total, na.rm=TRUE))

#4-year in-state
tc4YIS_Summary = tcFours%>%
  summarise(count_4YIS=n(),
            min=min(tcFours$in_state_total, na.rm=TRUE),
            Q1=quantile(tcFours$in_state_total, prob=0.25,na.rm=TRUE),
            med=median(tcFours$in_state_total, na.rm=TRUE),
            Q3=quantile(tcFours$in_state_total, prob=0.75,na.rm=TRUE),
            max=max(tcFours$in_state_total, na.rm=TRUE))

#2-year out-of-state
tc2YOOS_Summary = tcTwos%>%
  summarise(count_2YOOS=n(),
            min=min(tcTwos$out_of_state_total, na.rm=TRUE),
            Q1=quantile(tcTwos$out_of_state_total, prob=0.25,na.rm=TRUE),
            med=median(tcTwos$out_of_state_total, na.rm=TRUE),
            Q3=quantile(tcTwos$out_of_state_total, prob=0.75,na.rm=TRUE),
            max=max(tcTwos$out_of_state_total, na.rm=TRUE))

#2-year in-state
tc2YIS_Summary = tcTwos%>%
  summarise(count_2YIS=n(),
            min=min(tcTwos$in_state_total, na.rm=TRUE),
```

```
            Q1=quantile(tcTwos$in_state_total, prob=0.25,na.rm=TRUE),
            med=median(tcTwos$in_state_total, na.rm=TRUE),
            Q3=quantile(tcTwos$in_state_total, prob=0.75,na.rm=TRUE),
            max=max(tcTwos$in_state_total, na.rm=TRUE))
```

```
tc4YOOS_Summary
```

Boxplot 5-number Summaries

```
## # A tibble: 1 x 6
##   count_4YOOS   min    Q1   med    Q3   max
##         <int> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1        1852  1430 24951 34888 46670 75003
```

```
tc4YIS_Summary
```

```
## # A tibble: 1 x 6
##   count_4YIS   min    Q1   med     Q3   max
##        <int> <dbl> <dbl> <dbl>  <dbl> <dbl>
## 1       1852  1430 18199 28287 44846. 75003
```

```
tc2YOOS_Summary
```

```
## # A tibble: 1 x 6
##   count_2YOOS   min    Q1   med    Q3   max
##         <int> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1        1120  1376 8196. 10291 13598 68640
```

```
tc2YIS_Summary
```

```
## # A tibble: 1 x 6
##   count_2YIS   min    Q1   med    Q3   max
##        <int> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1       1120   962 3364. 4972.  8946 68640
```

Table and plot code

```
#Table 1
bystate = gatheredtc %>%
  group_by(state) %>%
  mutate(freq = n()) %>%
  summarize(numSchools = sum(freq)) %>%
  mutate(prop=numSchools/sum(numSchools)) %>%
  arrange(desc(prop))
knitr::kable(bystate,"pipe",col.names=c("State","Number","Proportion"))

#Table 2
knitr::kable(table(gatheredtc$degFactor),"pipe",col.names=c("Degree Factor","Number of Schools"))

#Plot 1
ggplot(gatheredtc, aes(x=totalCost,fill=in_out))+geom_histogram()+expand_limits(x=80000,y=430) +
  ggtitle("Total cost of attendance for students by residence")+ # for the main title
  xlab("Cost of attendance")+ # for the x axis label
  ylab("Count") + # for the y axis label
  theme_bw()+theme(
    legend.title = element_blank(),
```

```
    ) + scale_fill_discrete(name = "Student Residence", labels = c("In-state", "Out-of-state"))

#Plot 2
ggplot(gatheredtc, aes(x = degFactor, y = totalCost, fill=in_out)) +  # ggplot function
  geom_boxplot()+
  ggtitle("Breakdown of total cost of attendence by degree factor")+ # for the main title
  xlab("Degree Factor (2/4 year)")+ # for the x axis label
  ylab("Cost of Attendance")+ # for the y axis label
  theme_bw()+theme(
    legend.title = element_blank(),
    ) + scale_fill_discrete(name = "Student Residence", labels = c("In-state", "Out-of-state"))

#Plot 3
ggplot(gatheredtc, aes(x=totalCost,fill=in_out))+geom_histogram(bins=75)+expand_limits(x=80000) +
  ggtitle("Total cost of attendance for students by residence")+ # for the main title
  xlab("Cost of attendance")+ # for the x axis label
  ylab("Count") + # for the y axis label
  theme_bw()+theme(
    legend.title = element_blank(),
    ) + scale_fill_discrete(name = "Student Residence", labels = c("In-state", "Out-of-state"))

#Plotset 4
ggplot(gatheredtc,aes(x=totalCost, color=in_out)) +
  geom_density() +
  expand_limits(x=80000) +
  ggtitle("Distribution of cost-of-attendance by student residence")+
  xlab("Cost of attendance ($)") +
  ylab("Density") +
  theme_bw()+theme(
    legend.title = element_blank(),
    ) + scale_color_discrete(labels = c("In-state", "Out-of-state"))

ggplot(gatheredtc,aes(x=totalCost, color=type)) +
  geom_density() +
  expand_limits(x=80000) +
  ggtitle("Distribution of cost-of-attendance by college type")+
  xlab("Cost of attendance ($)") +
  ylab("Density") +
  theme_bw()+theme(
    legend.title = element_blank()) + scale_color_discrete(labels=c('For Profit','Private','Public'))

#Plot 5
ggplot(sp, aes(stem_percent,early_career_pay)) + geom_point() +
  geom_smooth(method="lm",se=FALSE)+
  ggtitle("Early career pay against prevalence of STEM majors in a school")+
  xlab("Prevalence of STEM majors (%)")+
  ylab("Early-career pay ($)")+
  theme_bw()

#Plot 6
ggplot(tcFacJoinSp, aes(x=stem_percent,y=in_state_tuition,color=type)) +
  geom_point() +
  geom_smooth(method="lm",se=FALSE,lty=2) +
  ggtitle("Cost of tuition for in-state students against prevalence of STEM majors")+
```

```
  xlab("Prevalence of STEM majors (%)")+
  ylab("In-state tuition cost ($)")+
  theme_bw()

#Plot 7
ggplot(tcFacJoinSp, aes(x=stem_percent,y=out_of_state_tuition,color=type)) +
  geom_point() +
  geom_smooth(method="lm",se=FALSE,lty=2) +
  ggtitle("Cost of tuition for out-state students against prevalence of STEM majors")+
  xlab("Prevalence of STEM majors (%)")+
  ylab("Out-state tuition cost ($)")+
  theme_bw()

#Miscellaneous relevant code
# IMPORTING DATASETS
tuition_cost <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/d
tc = tuition_cost

tuition_income <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master
ti = tuition_income

salary_potential <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/mast
sp = salary_potential

historical_tuition <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/ma
ht = historical_tuition

diversity_school <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/mast
ds = diversity_school

#Creating degree length factor
tcFactored = tc %>%
  mutate(degFactor = as.factor(degree_length))

#Gathering columns wrt total cost and in/out-of state status while filtering out extraneous school type
gatheredtc = tcFactored %>%
filter(type!='Other') %>%
gather(key="in_out", value="totalCost",c(in_state_total,out_of_state_total))

jointisp = ti %>%
left_join(sp) %>%
group_by(stem_percent)%>%
summarize(medianNet=median(net_cost))

tcFacJoinSp = tcFactored %>%
  filter(tcFactored$type!='For Profit') %>%
  inner_join(sp, by=c("name"="name"))
```