FINAL PROJECT

Louise Daly, Bee Garcia, Nicholas Peterson, Mingchen Wang

2022-11-16

Part 1 - Beginnings

1.1 - Motivating Question and Scope of Analysis:

Applying to college is a challenging task that millions of high school and college students take part in every year. Deciding on which college to choose is even more difficult, as there are many factors that play into your decision, such as the size of the school, the location, and the rigor. One of the most important factors is the cost of school, and whether or not they provide scholarships. Many colleges offer both merit and need based scholarships, but this generosity varies between schools. In this project, we will be assessing how large the difference is between what lower income students and higher income students have to pay among different schools, and if that difference is more pronounced in public institutions or private institutions.

1.2 - US Colleges Background:

- **Private vs. Public**: Colleges are typically split in whether they are public or private institutions. Public institutions are colleges that receive their primary funding from their state's government, where they may be owned by the state itself. Private institutions are colleges that typically function as educational nonprofits that garner the bulk of their funding from students, alumni, donors, or other means that are not reliant on state funds. Public colleges run less costly for students who reside within the state of the public college when compared to private colleges due to public universities receiving the bulk of funds from the state itself. Public colleges are also incentivized to take in in-state students since their funding is primarily from the state government they reside in.
- **Non-profit vs For-profit**: All public institutions are considered non-profit, which means they invest all the money they receive (through government, tuition, etc) back into the school. Private schools are mostly non-profit, but some are considered for-profit schools. For-profit schools are those whose main goal is to make money and therefore use the money they receive on aspects such as marketing and recruiting instead of putting it all back into the education system.
- **Fees paid by students**: All students attending a college are required to pay tuition which is the biggest fee they have to fork up. Each school has its own tuition and for public institutions it is often higher for out-of-state or international students. The other main fee students have to pay is their room and board fee. This includes their housing for the school year and often includes a meal plan as well. This fee is not required for commuting students, but those living on campus have to pay it.
- **Scholarships/Financial Aid**: Students can apply for scholarships or financial aid to reduce the fees they have to pay to their university. The amount of financial support they receive is often based on their family's income level as these institutions seek to help students who can't afford an undergraduate education and want to reduce the amount of student loans. Public institutions generally offer financial aid only to in-state students, while private institutions offer it to both in-state and out-of-state students. International students usually don't receive any aid but there are some merit-based and need-based exceptions.

- **2-Year Colleges (Community Colleges)**: For this project, 2-year colleges will be referenced as community colleges. Additionally, for this project a community college is an educational institution that provides credits towards a 4-year degree, an associate degree, or a technical/career certificate. Community colleges are institutions that can function as either a junior college or its own academic institution. These institutions are funded by a mixture of state-funds and student payments like public universities, but to a lesser degree. Community colleges are less expensive than 4-year colleges and provide roughly the same amount of credits for two years as a 4-year university does in its first two years. Community colleges may vary and include other costs such as housing for on campus students while some may not offer on-campus housing as an option.
- **In-state or Out-of-state**: Students can choose to attend any educational institution they were accepted to, but cost may vary depending on if the higher education institution resides within their state or outside of it. For public universities, tuition costs will be high for those students who are residents of another state. This difference in tuition is not as present in private schools who generally charge the same price regardless of a student's residency.
- **On-campus or Off-campus**: Students can decide to either live on the campus grounds of an institution or live off of the campus through personal living accommodations or school sponsored or owned apartments nearby. Pricing for on-campus students is higher since they must pay for housing and possible meal plan costs that off campus students do not necessarily have to account for.
- **Income levels**: A student's income level can be representative of how much they must spend to attend an institution. Typically, the higher income level a student or their family is in, the more a student must pay for their education. Income levels are not the sole factor of the price a student must pay. Outside factors like how many family members a student has could change the aid received by the students. There is potential for 2 students in different income levels to receive the same aid due to other outside factors. Income brackets serve as mainly a predictor and not a final say.

1.3 - Variables

Tuition_income dataset

Name (string) - This value contains the names of all the colleges that had reported data.

State (string) - Which state the institution is located in, given in a two-letter abbreviation.

Income level (string) - Students' family income range. There are 5 brackets: 0-30,000, 30,001 - 48,000, 48,001 - 75,000, 75,001 - 110,000, and over 110,000. The data is categorized by income level. Net cost and total price are likely to vary based on the students' income bracket.

Total price (Decimal) - The total or "sticker" price of the university including room and board, books, tuition, etc.

Net cost (Decimal) - The average amount of money that students actually pay to the institution after scholarships/awards.

Tuition_cost dataset

Name -(character) This value contains all the colleges that had reported data to Chronicle of Higher Education. There are 2973 unique universities that reported their data in the year 2018-2019.

Type (public/private) -(character) If the college classifies themselves as a public, private or for-profit institution.

Room and board -(numeric) The cost of the university's room and board per year for all students. This value is measured in US dollars.

In state tuition-(numeric) The cost of the university's tuition per year for in-state students. This value is measured in US dollars.

In state total -(numeric) The total cost of tuition per year for in-state students. This value is measured in US dollars.

Out of state tuition -(numeric) The cost of the university's tuition per year for out of state students. This value is measured in US dollars.

Out of state total -(numeric) The total cost of tuition per year for out of state students. This value is measured in US dollars.

Part 2 - Data Analysis

Reading Datasets

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.4.0     v purrr   0.3.5
## v tibble  3.1.8     v stringr 1.4.1
## v tidyr   1.2.1     v forcats 0.5.2
## v readr   2.1.3
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
## Warning: package 'purrr' was built under R version 4.2.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)

tuition_cost <-
read.csv("tuition_cost.csv")
tuition_income <-
read.csv("tuition_income.csv")

tuition_cost %>%
select(name,state,type)%>%
head(5)
```

```
##                                   name      state       type
## 1              Aaniiih Nakoda College    Montana     Public
## 2         Abilene Christian University      Texas    Private
## 3 Abraham Baldwin Agricultural College    Georgia     Public
## 4                      Academy College  Minnesota For Profit
## 5            Academy of Art University California For Profit
```

```
summary(tuition_cost)
```

```
##      name               state            state_code            type
##  Length:2973        Length:2973        Length:2973        Length:2973
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  degree_length      room_and_board  in_state_tuition in_state_total
##  Length:2973        Min.   :   30   Min.   :  480    Min.   :  962
##  Class :character   1st Qu.: 7935   1st Qu.: 4890    1st Qu.: 5802
##  Mode  :character   Median :10000   Median :10099    Median :17669
##                     Mean   :10095   Mean   :16491    Mean   :22872
##                     3rd Qu.:12424   3rd Qu.:27124    3rd Qu.:35960
##                     Max.   :21300   Max.   :59985    Max.   :75003
##                     NA's   :1094
##  out_of_state_tuition out_of_state_total
##  Min.   :  480        Min.   : 1376
##  1st Qu.: 9552        1st Qu.:11196
##  Median :17486        Median :23214
##  Mean   :20533        Mean   :26913
##  3rd Qu.:29208        3rd Qu.:39054
##  Max.   :59985        Max.   :75003
##
```

```
tuition_income %>%
select(name,state,total_price)%>%
head(5)
```

```
##                               name state total_price
## 1 Piedmont International University    NC       20174
## 2 Piedmont International University    NC       20174
```

```
## 3 Piedmont International University    NC       20174
## 4 Piedmont International University    NC       20174
## 5 Piedmont International University    NC       20514
```

```
summary(tuition_income)
```

```
##      name              state            total_price          year
##  Length:209012      Length:209012      Min.   :  4906   Min.   :2010
##  Class :character   Class :character   1st Qu.: 19186   1st Qu.:2012
##  Mode  :character   Mode  :character   Median : 26286   Median :2014
##                                        Mean   : 30102   Mean   :2014
##                                        3rd Qu.: 38831   3rd Qu.:2017
##                                        Max.   :114083   Max.   :2018
##     campus             net_cost       income_lvl
##  Length:209012      Min.   :-15101   Length:209012
##  Class :character   1st Qu.: 10149   Class :character
##  Mode  :character   Median : 16207   Mode  :character
##                     Mean   : 16785
##                     3rd Qu.: 22350
##                     Max.   : 95675
```

cleaning up tuition_income

```
cleaned_tuition_income<-
tuition_income %>%
filter(year == c(2018)) %>%
filter(net_cost != 0) %>%
##mention something about fixing names issue
mutate(name=str_replace(name
,"-", ": "))
cleaned_tuition_income %>% head()
```

```
##                                name state total_price year     campus net_cost
## 1 Piedmont International University    NC       20829 2018  On Campus 11847.56
## 2 Piedmont International University    NC       20829 2018  On Campus 11822.79
## 3 Piedmont International University    NC       20829 2018  On Campus 16755.92
## 4 Piedmont International University    NC       20829 2018  On Campus 16098.23
## 5 Piedmont International University    NC       26870 2018 Off Campus 11847.56
## 6 Piedmont International University    NC       26870 2018 Off Campus 11822.79
##          income_lvl
## 1       0 to 30,000
## 2  30,001 to 48,000
## 3  48_001 to 75,000
## 4 75,001 to 110,000
## 5       0 to 30,000
## 6  30,001 to 48,000
```

Joining our data

```
income_and_cost<-
tuition_cost%>%
inner_join(cleaned_tuition_income,by=c("name"="name"))
income_and_cost%>%
head()
```

```
##                           name state.x state_code    type degree_length
## 1       Aaniiih Nakoda College Montana         MT  Public        2 Year
## 2       Aaniiih Nakoda College Montana         MT  Public        2 Year
## 3       Aaniiih Nakoda College Montana         MT  Public        2 Year
## 4 Abilene Christian University   Texas         TX Private        4 Year
## 5 Abilene Christian University   Texas         TX Private        4 Year
## 6 Abilene Christian University   Texas         TX Private        4 Year
##   room_and_board in_state_tuition in_state_total out_of_state_tuition
## 1             NA             2380           2380                 2380
## 2             NA             2380           2380                 2380
## 3             NA             2380           2380                 2380
## 4          10350            34850          45200                34850
## 5          10350            34850          45200                34850
## 6          10350            34850          45200                34850
##   out_of_state_total state.y total_price year     campus net_cost
## 1               2380      MT       17030 2018 Off Campus  8541.00
## 2               2380      MT       17030 2018 Off Campus  7371.00
## 3               2380      MT       17030 2018 Off Campus 10492.00
## 4              45200      TX       49722 2018  On Campus 23867.21
## 5              45200      TX       49722 2018  On Campus 25567.14
## 6              45200      TX       49722 2018  On Campus 26921.04
##          income_lvl
## 1       0 to 30,000
## 2  30,001 to 48,000
## 3 75,001 to 110,000
## 4       0 to 30,000
## 5  30,001 to 48,000
## 6  48_001 to 75,000
```

Revaleuing the NAs

```
##Accounting for DC and no room and board and fixing aesthetic inconsistency
income_and_cost<-
income_and_cost %>%
replace_na(list(state.x='DC',room_and_board = 0)) %>%
mutate(income_lvl=str_replace(income_lvl,"_", ","))
```

remove unneeded columns

```
income_and_cost <-
income_and_cost %>%
select(name,state.x,type,degree_length,campus,
room_and_board,in_state_tuition,in_state_total,
```

```
out_of_state_tuition,out_of_state_total,
net_cost,total_price,income_lvl)

income_and_cost %>%
head()
```

```
##                           name state.x    type degree_length     campus
## 1       Aaniiih Nakoda College Montana  Public        2 Year Off Campus
## 2       Aaniiih Nakoda College Montana  Public        2 Year Off Campus
## 3       Aaniiih Nakoda College Montana  Public        2 Year Off Campus
## 4 Abilene Christian University   Texas Private        4 Year  On Campus
## 5 Abilene Christian University   Texas Private        4 Year  On Campus
## 6 Abilene Christian University   Texas Private        4 Year  On Campus
##   room_and_board in_state_tuition in_state_total out_of_state_tuition
## 1              0             2380           2380                 2380
## 2              0             2380           2380                 2380
## 3              0             2380           2380                 2380
## 4          10350            34850          45200                34850
## 5          10350            34850          45200                34850
## 6          10350            34850          45200                34850
##   out_of_state_total net_cost total_price        income_lvl
## 1               2380  8541.00       17030       0 to 30,000
## 2               2380  7371.00       17030  30,001 to 48,000
## 3               2380 10492.00       17030 75,001 to 110,000
## 4              45200 23867.21       49722       0 to 30,000
## 5              45200 25567.14       49722  30,001 to 48,000
## 6              45200 26921.04       49722  48,001 to 75,000
```

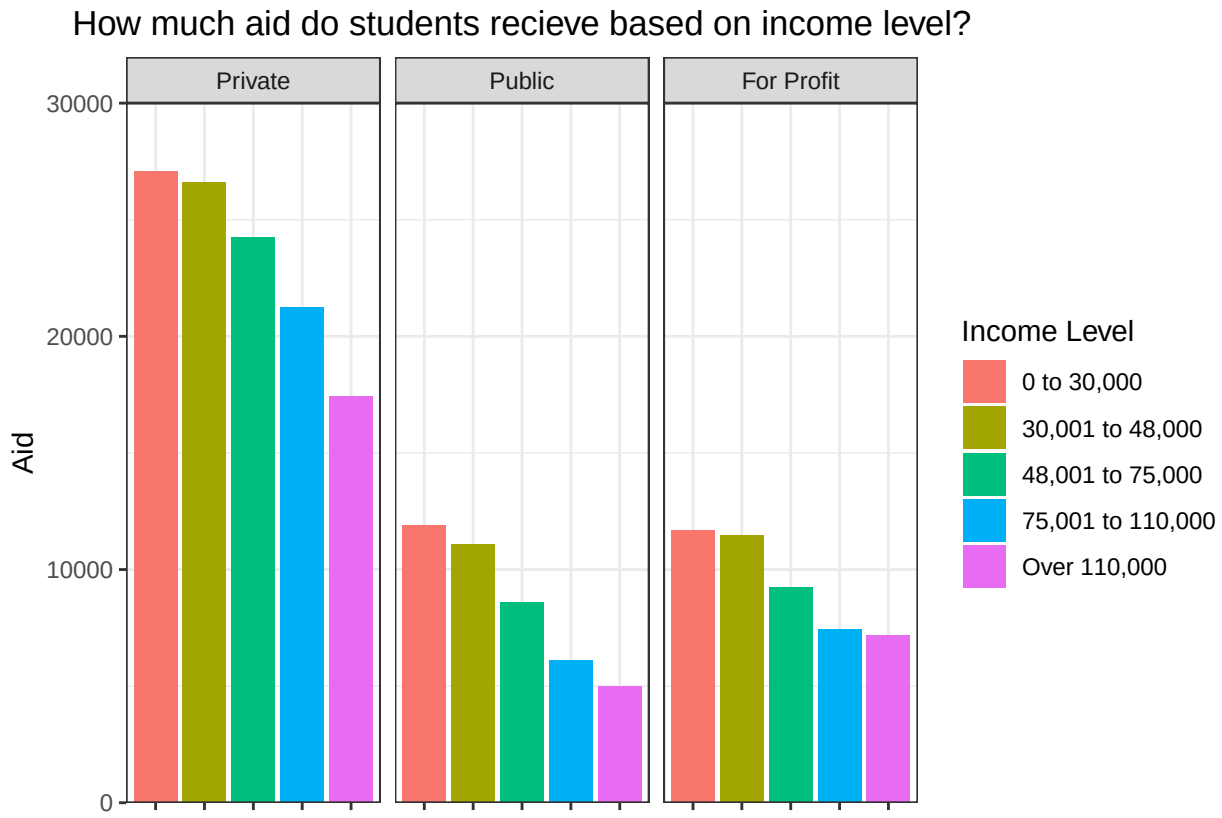
```
##negative Aid viewing
income_and_cost_aid<-
income_and_cost %>%
mutate(aid=total_price-net_cost)

average_income_and_cost_aid<-
income_and_cost_aid %>%
group_by(type, income_lvl) %>%
summarise(average_aid= mean(aid))
```

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

```
ggplot(average_income_and_cost_aid)+
geom_col(aes(x=income_lvl, y=average_aid, fill=income_lvl))+
ylab("Aid")+
xlab("")+
ggtitle("How much aid do students recieve based on income level?")+
facet_grid(~factor(type,levels=
c('Private','Public','For Profit'))) +
scale_x_discrete(limits=rev)+
theme_bw()+
theme(axis.text.x=element_blank())+
scale_x_discrete(guide = guide_axis(angle
```

```
= 90))+
theme(plot.title=element_text(hjust=0.5))+
guides(fill=guide_legend(title="Income Level"))+
scale_y_continuous(expand=c(0,0),limits=c(0,30000))
```

```
## Scale for x is already present.
## Adding another scale for x, which will replace the existing scale.
```



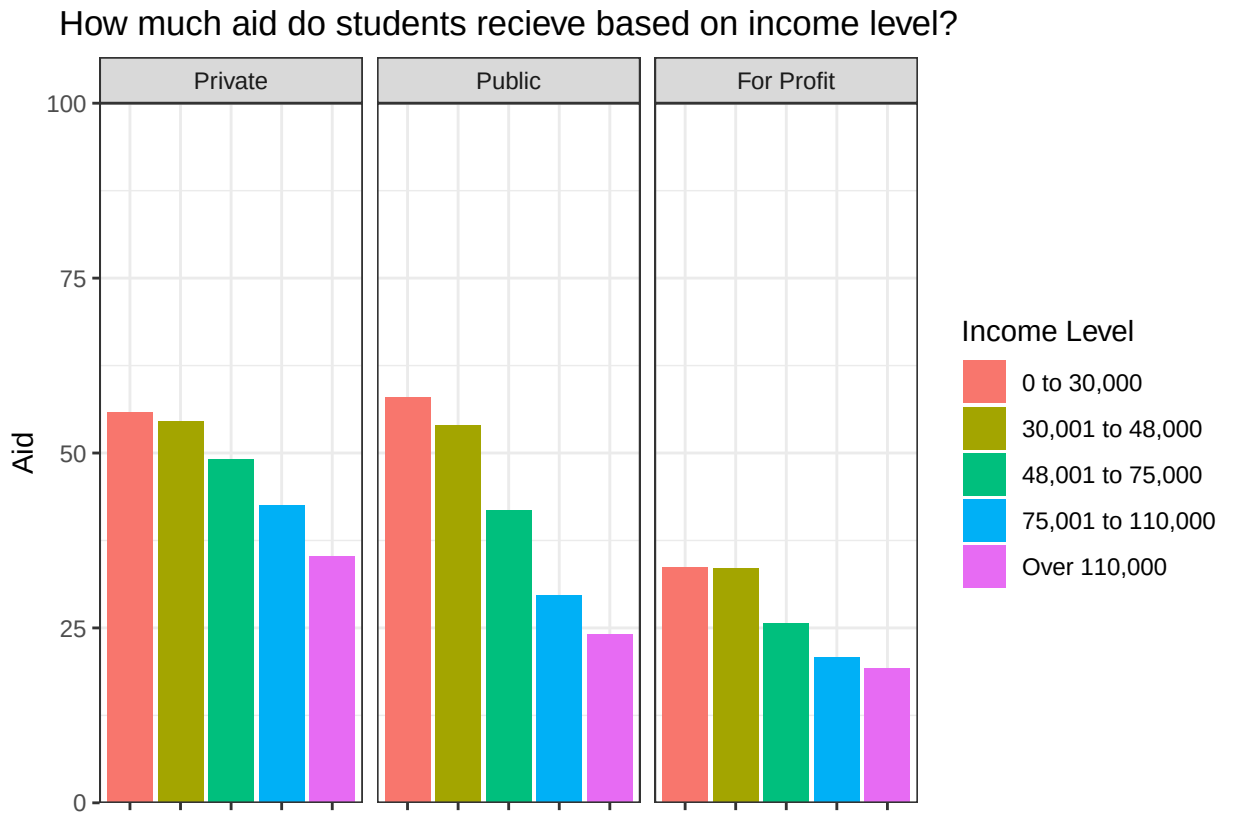
A common trend among all types of schools is that families with lower incomes receive the most aid, however, overall, private schools give out the most financial aid regardless of income level.

```
income_and_cost_aid<-
income_and_cost %>%
mutate(aid=total_price-net_cost)

income_and_cost_percent_aid<-
income_and_cost_aid %>%
mutate(percent_aid=(aid/total_price)*100)

average_income_and_cost_percent_aid<-
income_and_cost_percent_aid %>%
group_by(type, income_lvl) %>%
summarise(average_percent_aid= mean(percent_aid))
```

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

```
ggplot(average_income_and_cost_percent_aid)+
geom_col(aes(x=income_lvl, y=average_percent_aid, fill=income_lvl))+
ylab("Aid")+
xlab("")+
ggtitle("How much aid do students recieve based on income level?")+
facet_grid(~factor(type,levels=
c('Private','Public','For Profit'))) +
scale_x_discrete(limits=rev)+
theme_bw()+
theme(axis.text.x=element_blank())+
scale_x_discrete(guide = guide_axis(angle
= 90))+
theme(plot.title=element_text(hjust=0.5))+
guides(fill=guide_legend(title="Income Level"))+
scale_y_continuous(expand=c(0,0),limits=c(0,100))
```

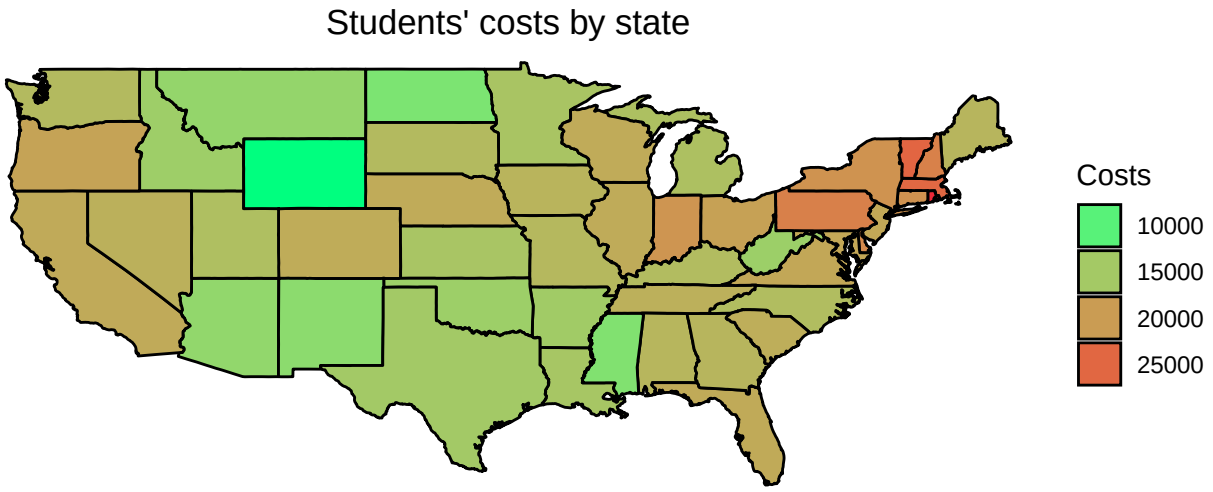
```
## Scale for x is already present.
## Adding another scale for x, which will replace the existing scale.
```



Although private schools give more aid as demonstrated in the previous bar graph, public schools cover a slightly higher percentage of tuition through financial aid for those in the lowest income level bracket, but a smaller percentage for those in higher income brackets.

Let's check how much tuition do student pay based on the state the college is in.

```
states_map<-
map_data("state")

average_cost_by_state<-
income_and_cost %>%
group_by(state.x) %>%
mutate(average_cost_after_aid = mean(net_cost))


ggplot(average_cost_by_state, aes(map_id = tolower(state.x)))+
geom_map(aes(fill=average_cost_after_aid), map = states_map, color = "black")+
expand_limits(x=states_map$long,y=states_map$lat)+
coord_equal()+
theme_void()+
ggtitle("Students' costs by state")+
theme(plot.title=element_text(hjust=0.5))+
guides(fill=guide_legend(title="Costs"))+
scale_fill_gradient(low="#00FF7F",high="#ED2938")
```



In the visual below, it is clear that Colorado schools have the lowest average cost after aid, while schools on the east coast tend to be more expensive, specifically in states like Massachusetts and Vermont.

Let's check how much tuition do student pay based on their family's income level.

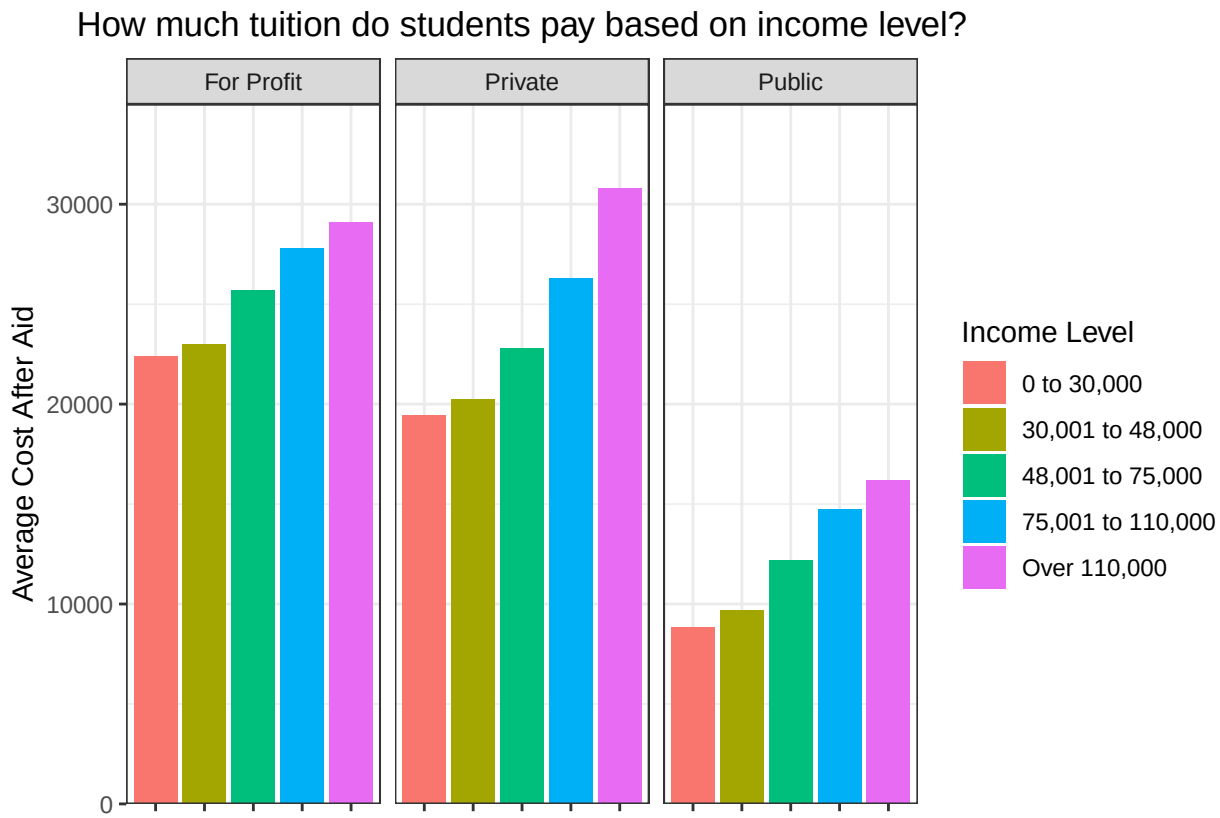
```
average_cost_by_income_lvl <-
income_and_cost%>%
```

```
group_by(type, income_lvl) %>%
summarise(average_cost_after_aid= mean(net_cost))
```

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

```
ggplot(average_cost_by_income_lvl)+
geom_col( aes(x = income_lvl, y=average_cost_after_aid, fill = income_lvl))+
ylab("Average Cost After Aid")+
xlab("")+
ggtitle("How much tuition do students pay based on income level?")+
facet_wrap(~type,scales = "free_x") +
scale_x_discrete(limits=rev)+
theme_bw()+
theme(axis.text.x=element_blank())+
scale_x_discrete(guide = guide_axis(angle
= 90))+
theme(plot.title=element_text(hjust=0.5))+
guides(fill=guide_legend(title="Income Level"))+
scale_y_continuous(expand=c(0,0),limits=c(0,35000))
```

```
## Scale for x is already present.
## Adding another scale for x, which will replace the existing scale.
```



```
## what happened with for profit school?
```

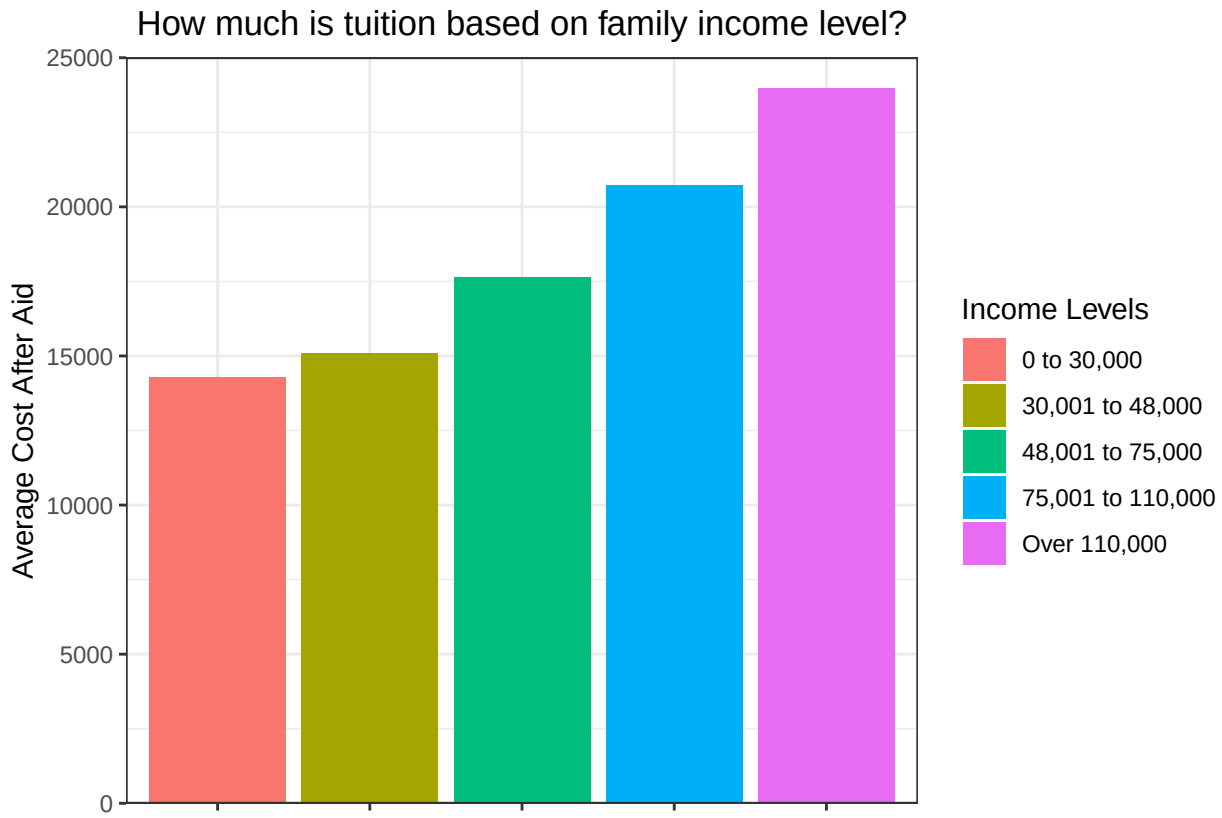
Overall, the trends among income levels for each type of school are similar; students with higher income levels pay more than those with lower income levels. However, this graph also indicates that public schools are cheaper overall than for profit and private schools, and additionally, the gap between what the highest income students pay versus the lowest income students is much wider in private schools.

Let's check how much tuition do student pay based on their family's income level

```
average_cost_by_income_lvl<-
income_and_cost %>%
mutate(income_lvl=str_replace(income_lvl, "_", ",")) %>%
group_by(income_lvl) %>%
summarise(average_cost_after_aid = mean(net_cost))

average_cost_by_income_lvl %>% head()
```

```
## # A tibble: 5 x 2
##   income_lvl        average_cost_after_aid
##   <chr>                              <dbl>
## 1 0 to 30,000                       14294.
## 2 30,001 to 48,000                  15093.
## 3 48,001 to 75,000                  17629.
## 4 75,001 to 110,000                 20708.
## 5 Over 110,000                      23970.
```

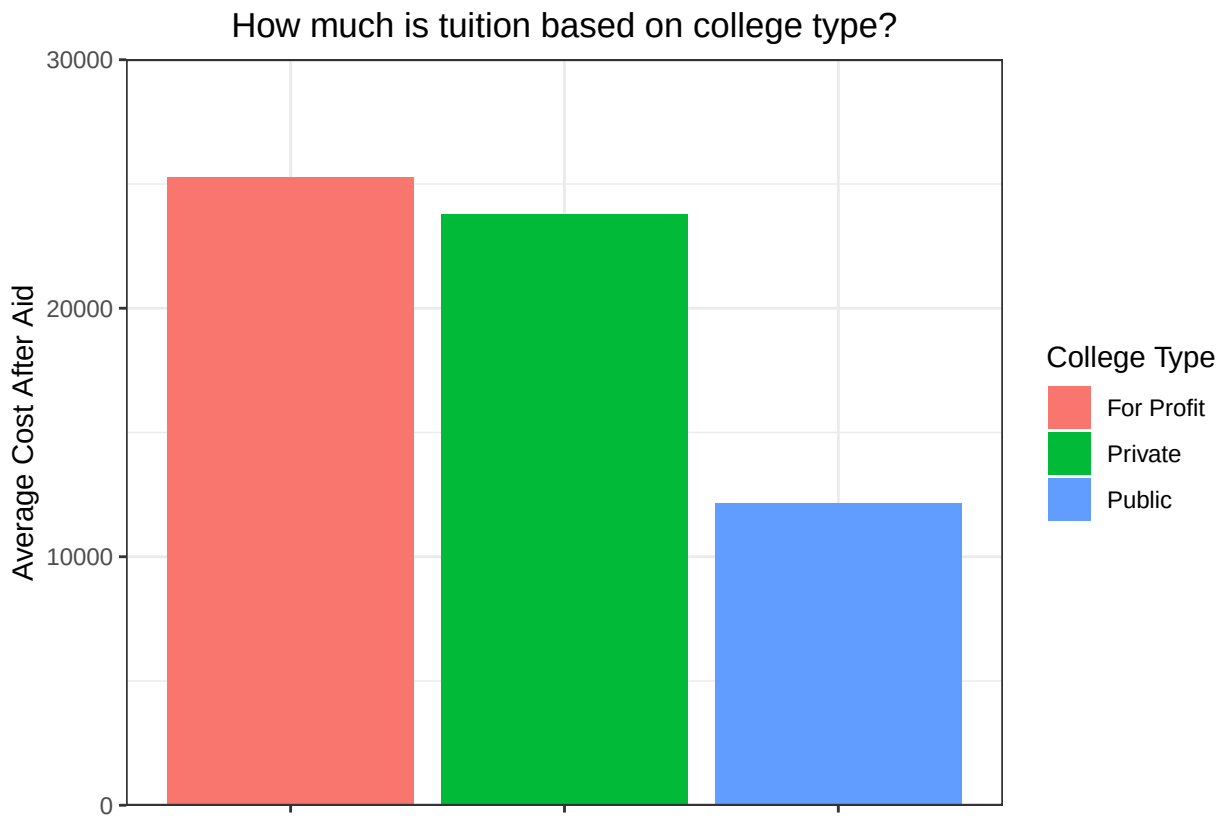
```
ggplot(average_cost_by_income_lvl)+
geom_col( aes(x=income_lvl,y=average_cost_after_aid,fill=income_lvl))+
ylab("Average Cost After Aid")+
xlab("")+
ggtitle("How much is tuition based on family income level?")+
theme_bw()+
theme(axis.text.x=element_blank())+
guides(fill=guide_legend(title="Income Levels"))+
theme(plot.title=element_text(hjust=0.5))+
scale_y_continuous(expand=c(0,0),limits=c(0,25000))
```



Similar to the last graph, this bar plot indicates that families with higher income levels pay more for college than families with lower income levels.

Let's check how much tuition do student pay based on the type of college

```
average_cost_by_type<-
income_and_cost %>%
group_by(type) %>%
summarise(average_cost_after_aid=mean(net_cost))


ggplot(average_cost_by_type)+
geom_col(aes(x=type,y=average_cost_after_aid,fill=type))+
ylab("Average Cost After Aid")+
xlab("")+
ggtitle("How much is tuition based on college type?")+ theme_bw()+
theme(axis.text.x=element_blank())+
guides(fill=guide_legend(title="College Type"))+
theme(plot.title=element_text(hjust=0.5))+
scale_y_continuous(expand=c(0,0),limits=c(0,30000))
```



For profit institutions, on average, cost the most while public institutions cost the least.

Public vs private vs for profit by state

```
states_map<-
map_data("state")

number_of_private_by_state<-
income_and_cost %>%
group_by(state.x)%>%
filter(type == "Private")%>%
select(type, name, state.x)%>%
distinct() %>%
summarize(count=n())%>%
rename(Number_of_Private_Schools=count)

number_of_private_by_state%>%
arrange(desc(Number_of_Private_Schools)) %>% head()
```

```
## # A tibble: 6 x 2
##   state.x       Number_of_Private_Schools
##   <chr>                             <int>
## 1 New York                             84
## 2 Pennsylvania                         84
## 3 California                           70
## 4 Massachusetts                        55
## 5 Ohio                                 55
```

```
## 6 Illinois                              45
```

```
number_of_public_by_state<-
income_and_cost %>%
group_by(state.x)%>%
filter(type == "Public")%>%
select(type, name, state.x)%>%
distinct()%>%
summarize(count=n())%>%
rename(Number_of_Public_Schools=count)

number_of_public_by_state%>%
arrange(desc(Number_of_Public_Schools)) %>% head()
```

```
## # A tibble: 6 x 2
##   state.x        Number_of_Public_Schools
##   <chr>                             <int>
## 1 California                          137
## 2 Texas                                67
## 3 North Carolina                       65
## 4 Illinois                             47
## 5 Georgia                              42
## 6 Michigan                             41
```

```
number_of_for_profit_by_state<-
income_and_cost%>%
group_by(state.x)%>%
filter(type=="For Profit")%>%
select(type,name,state.x)%>%
distinct()%>%
summarize(count=n())%>%
rename(Number_of_For_Profit_Schools=count)

number_of_for_profit_by_state%>%
arrange(desc(Number_of_For_Profit_Schools)) %>% head()
```

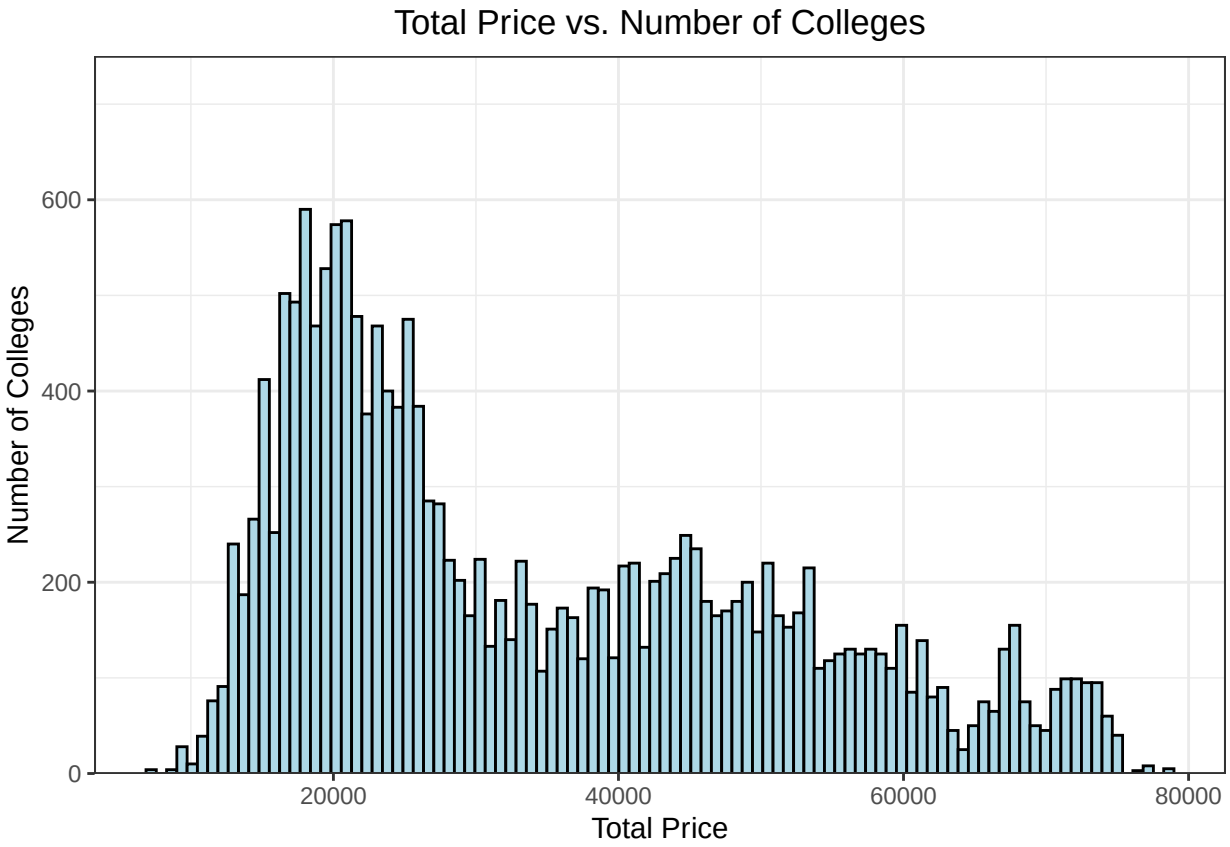
```
## # A tibble: 6 x 2
##   state.x    Number_of_For_Profit_Schools
##   <chr>                             <int>
## 1 New York                              9
## 2 California                            4
## 3 Florida                               3
## 4 Tennessee                             3
## 5 Arizona                               2
## 6 Illinois                              2
```

```
## must be in this order because only public by state has 50 rows
public_and_private<-
number_of_public_by_state%>%
left_join(number_of_private_by_state,"state.x")

number_of_type_by_state<-
```

```
public_and_private%>%
left_join(number_of_for_profit_by_state,by=c("state.x" = "state.x"))
```

Outliers total price

```
ggplot(income_and_cost,aes(x = total_price))+
geom_histogram(fill="lightblue",color="black",bins=100)+
theme_bw()+
scale_y_continuous(expand=c(0,0),limits=c(0,750))+
xlab("Total Price")+ylab("Number of Colleges")+
ggtitle("Total Price vs. Number of Colleges")+
theme(plot.title=element_text(hjust=0.5))
```

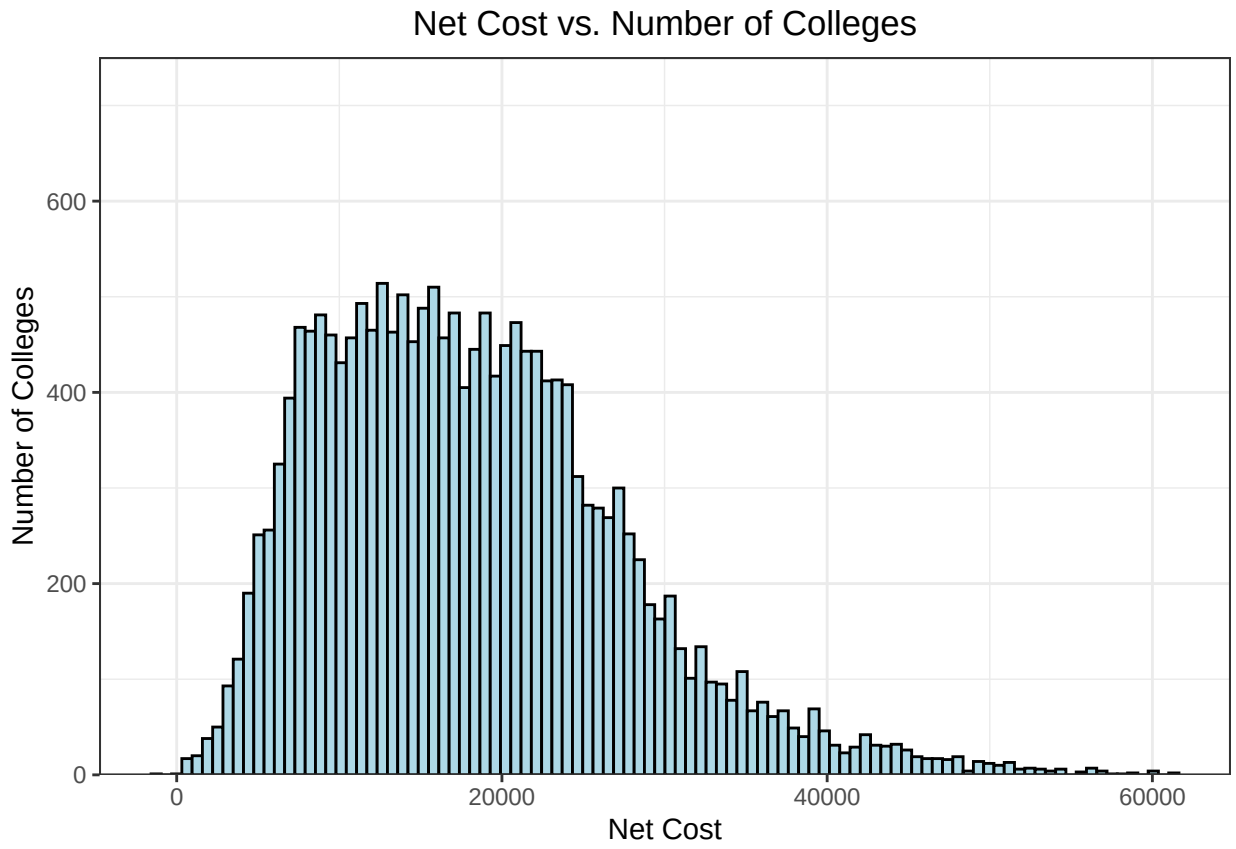


Although the graph is skewed slightly to the right, the overall distribution is pretty even with not obvious outliers that are dragging the average too far up/down.

Outliers - net cost

```
ggplot(income_and_cost,aes(x = net_cost))+
geom_histogram(fill="lightblue",color="black",bins=100)+
theme_bw()+
scale_y_continuous(expand=c(0,0),limits=c(0,750))+
```

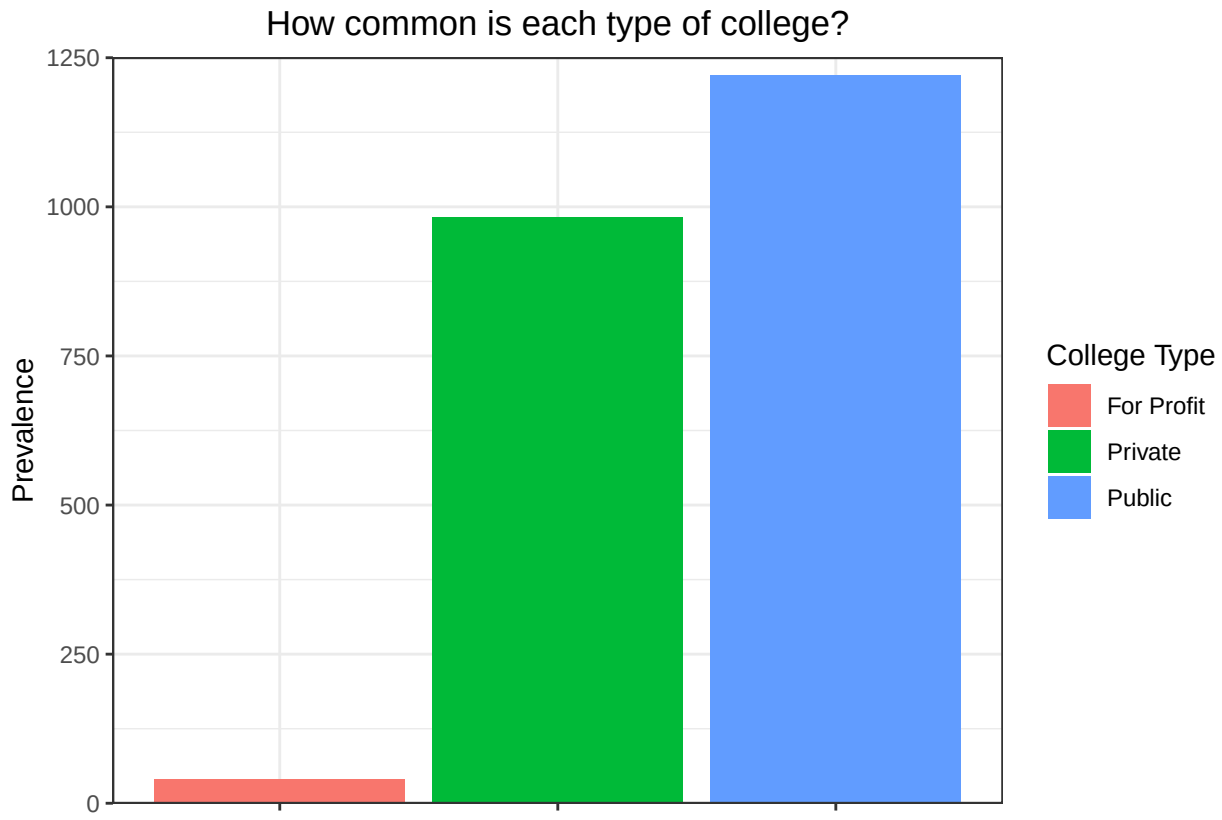
```
xlab("Net Cost")+ylab("Number of Colleges")+
ggtitle("Net Cost vs. Number of Colleges")+
theme(plot.title=element_text(hjust=0.5))
```



Similarly to total price, the graph is slightly skewed to the right, however there are no values that are drastically bringing up or down the mean.

```
number_of_type_by_state_refined<-
number_of_type_by_state %>%
rename("Public"="Number_of_Public_Schools",
"Private"="Number_of_Private_Schools",
"For Profit"="Number_of_For_Profit_Schools") %>%
pivot_longer(cols=c("Public":"For Profit"),
names_to='type',values_to='n')

ggplot(number_of_type_by_state_refined)+
geom_col(aes(x =type,y=n,fill=type))+
ylab("Prevalence")+xlab("")+
ggtitle("How common is each type of college?")+
theme_bw()+
theme(axis.text.x=element_blank())+
guides(fill=guide_legend(title="College Type"))+
theme(plot.title=element_text(hjust=0.5))+
scale_y_continuous(expand=c(0,0),limits=c(0,1250))
```

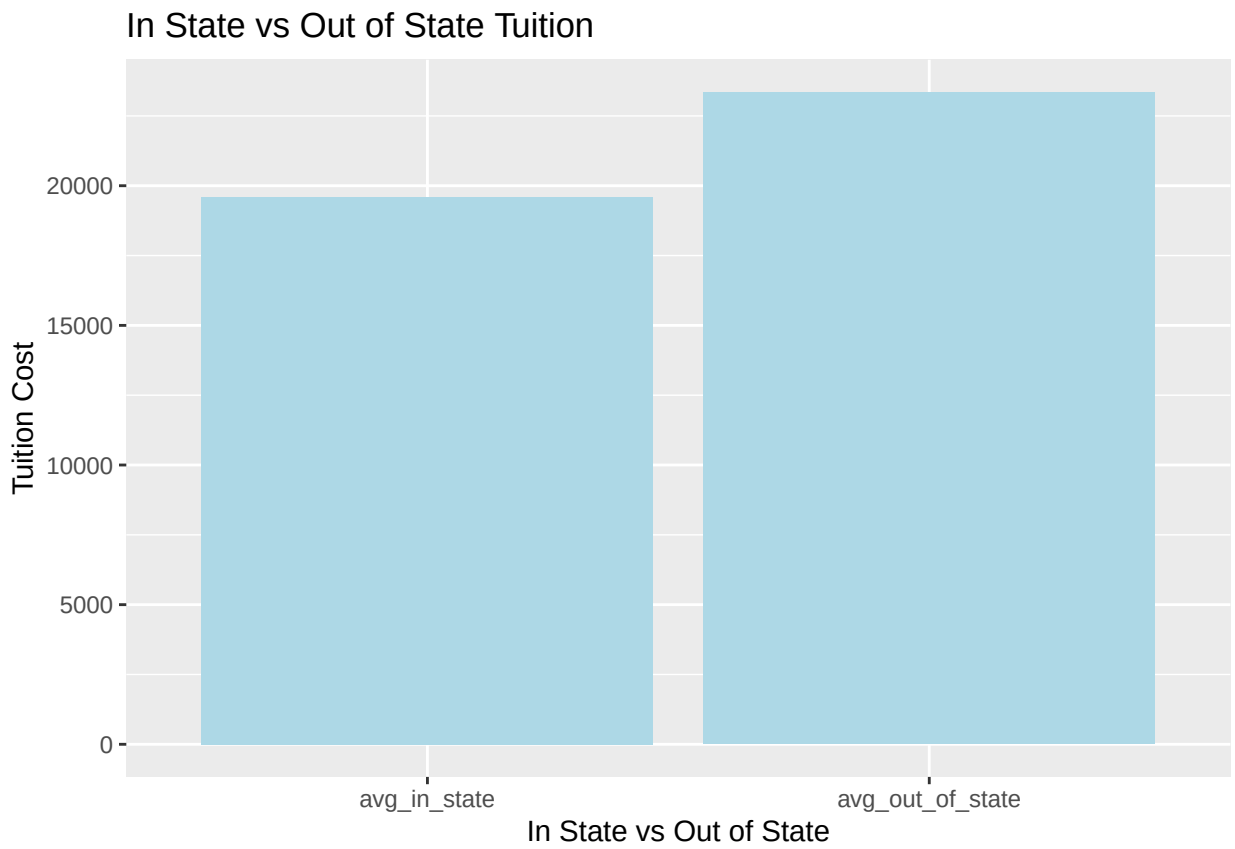
```
## Warning: Removed 32 rows containing missing values (‘position_stack()‘).
```



Public schools are the most common type of school in our dataset.

In state vs out of state tuition

```
in_vs_out_state<-
income_and_cost%>%
summarise(avg_out_of_state=mean(out_of_state_tuition),
avg_in_state=mean(in_state_tuition))

in_vs_out_state%>%
pivot_longer(cols=c("avg_out_of_state","avg_in_state"),
names_to='averages',values_to='tuition') %>%
ggplot(aes(x = averages, y = tuition))+
geom_bar(stat="identity",fill="lightblue")+
ylab("Tuition Cost")+
xlab("In State vs Out of State")+
ggtitle("In State vs Out of State Tuition")
```

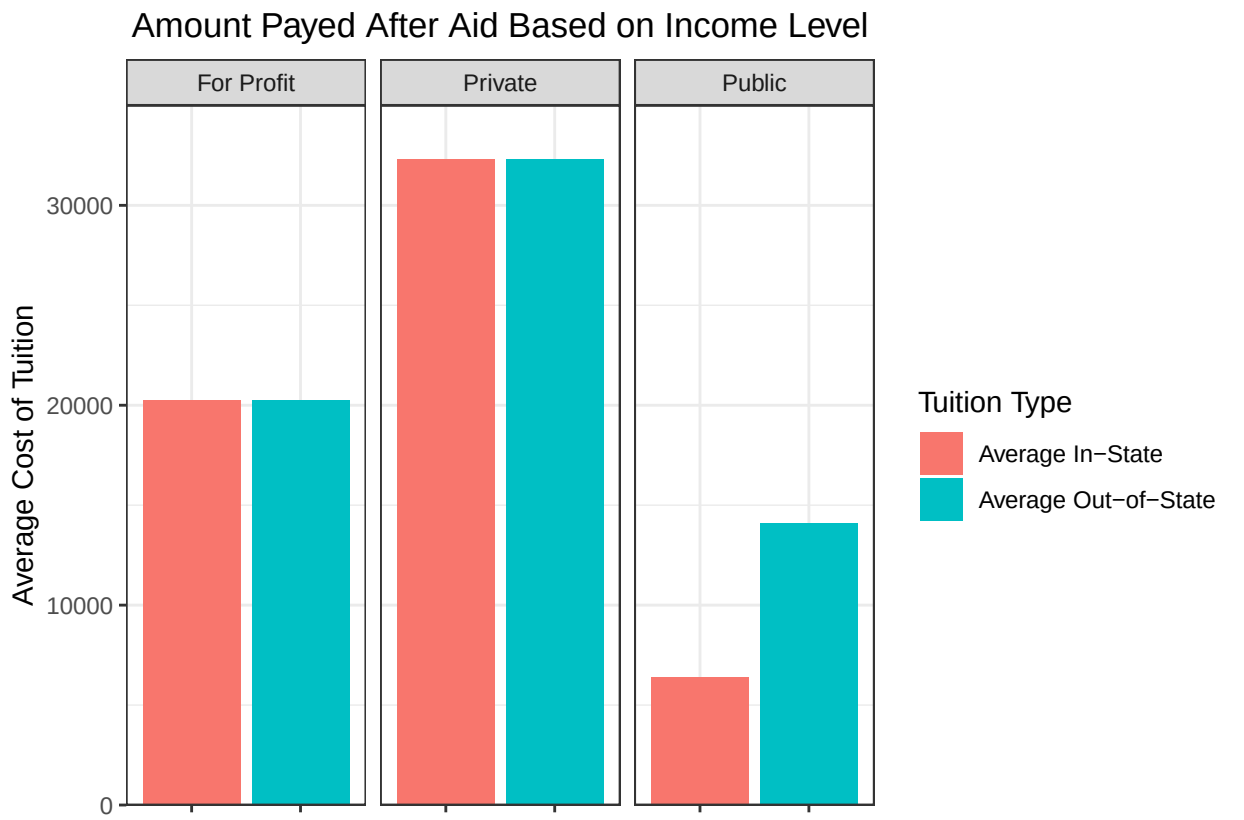


On average, out of state costs are much higher than the in state costs.

In state vs out of state tuition by type

```
in_out_by_type<-
income_and_cost%>%
group_by(type)%>%
summarise(avg_out_of_state=mean(out_of_state_tuition),
avg_in_state = mean(in_state_tuition))


in_out_by_type<-in_out_by_type%>%
pivot_longer(cols=c("avg_out_of_state","avg_in_state"),
names_to='averages',values_to='tuition') %>%
mutate(averages=str_replace(averages,
"avg_out_of_state","Average Out-of-State")) %>%
mutate(averages=str_replace(averages,
"avg_in_state","Average In-State"))

ggplot(in_out_by_type)+
geom_col(aes(x =averages,y=tuition,fill=averages))+
ylab("Average Cost of Tuition")+xlab("")+
ggtitle("Amount Payed After Aid Based on Income Level")+
facet_wrap(~type)+
theme_bw()+theme(axis.text.x=element_blank())+
guides(fill=guide_legend(title="Tuition Type"))+
theme(plot.title=element_text(hjust=0.5))+
```

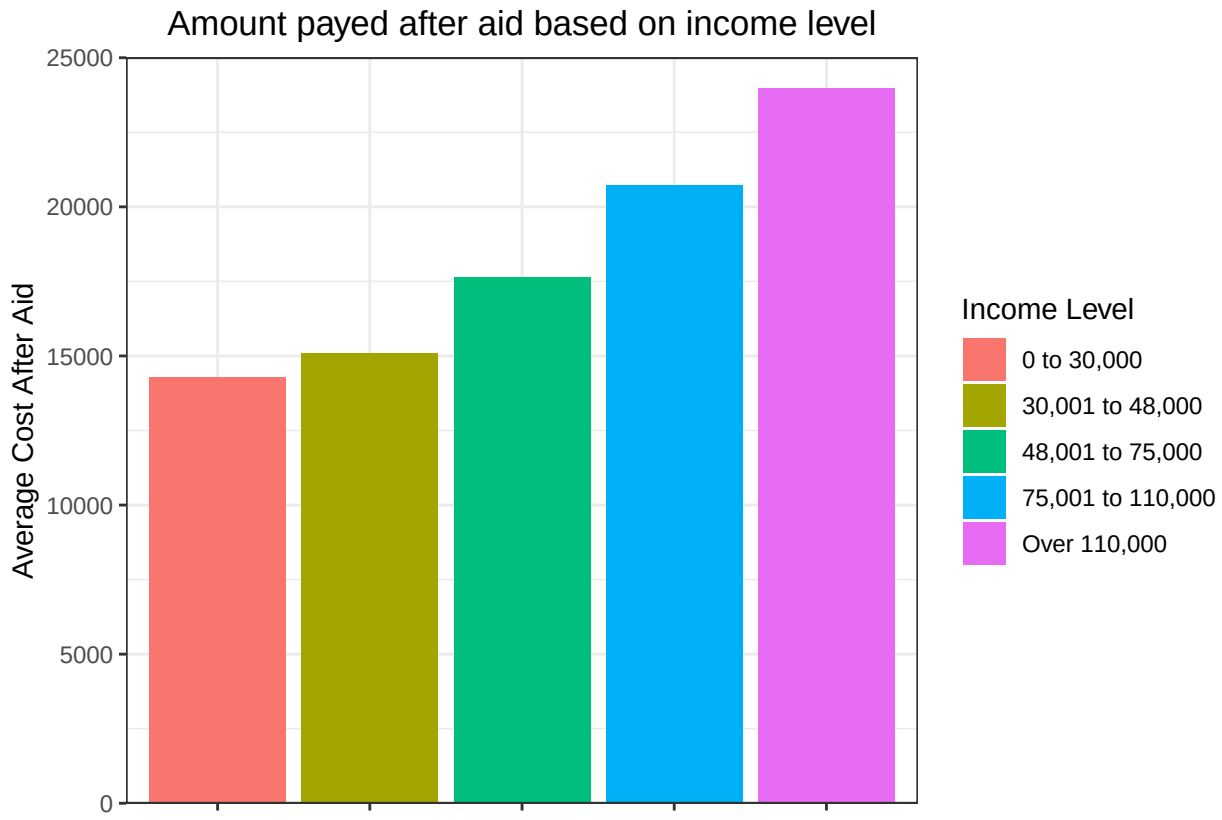
```
scale_y_continuous(expand=c(0,0),limits=c(0,35000))
```



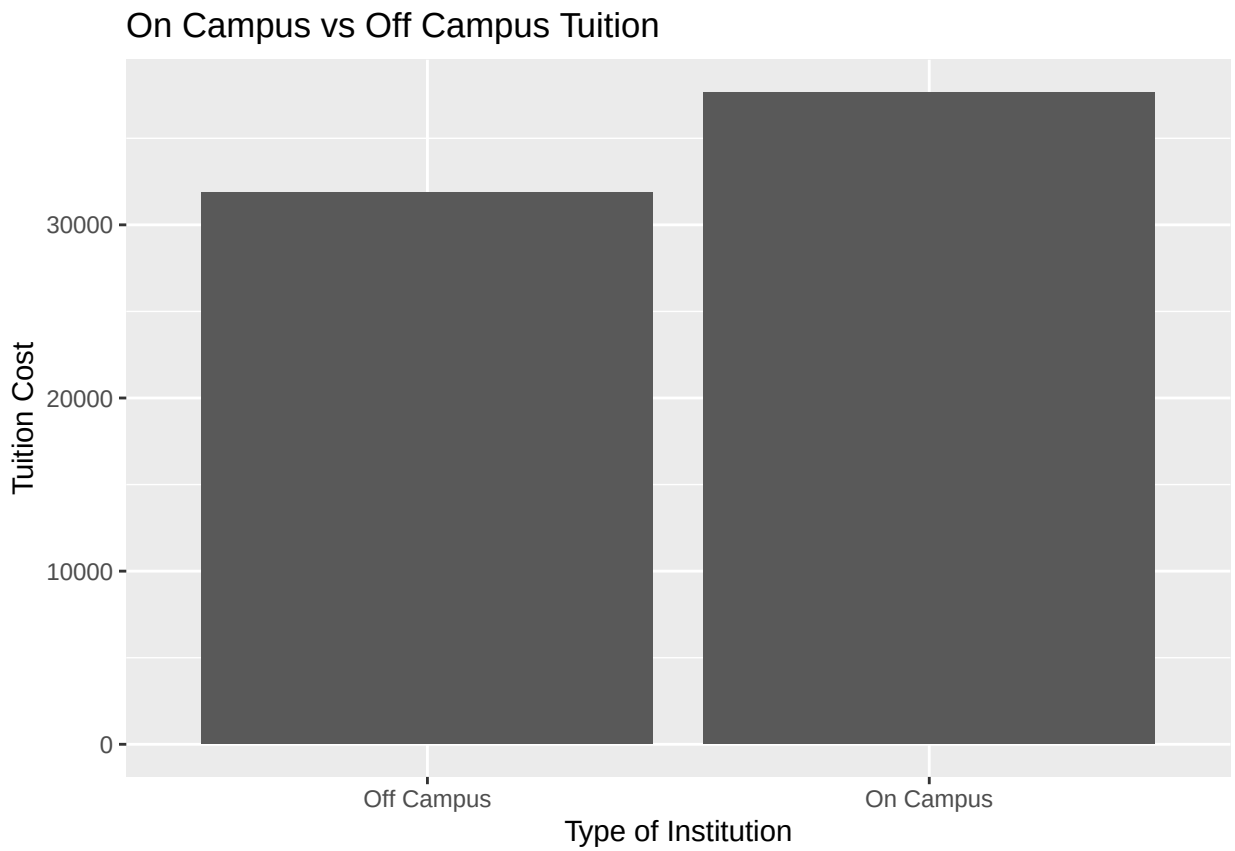
Not surprisingly, tuition for private and for profit schools are the same regardless of whether or not a student is from in state or out of state. On the other hand, in-state students on average pay almost $10,000 less than out-of-state students.

How much do student pay based on income level?

```
net_cost_by_income_lvl<-
income_and_cost%>%
group_by(income_lvl)%>%
summarise(net_price=mean(net_cost))

ggplot(net_cost_by_income_lvl)+
geom_col(aes(x =income_lvl,y=net_price,fill=income_lvl))+
ylab("Average Cost After Aid")+xlab("")+
ggtitle("Amount payed after aid based on income level")+
theme_bw()+theme(axis.text.x=element_blank())+
guides(fill=guide_legend(title="Income Level"))+
theme(plot.title=element_text(hjust=0.5))+
scale_y_continuous(expand=c(0,0),limits=c(0,25000))
```




```
on_off_campus<-
income_and_cost%>%
group_by(campus)%>%
summarise(avg_price=mean(total_price))

on_off_campus%>%
pivot_longer(cols=c("avg_price"),names_to='averages',values_to='tuition')%>%
ggplot(aes(x=campus,y=tuition))+
geom_bar(position="dodge", stat="identity")+
ylab("Tuition Cost")+
xlab("Type of Institution")+
ggtitle("On Campus vs Off Campus Tuition")
```



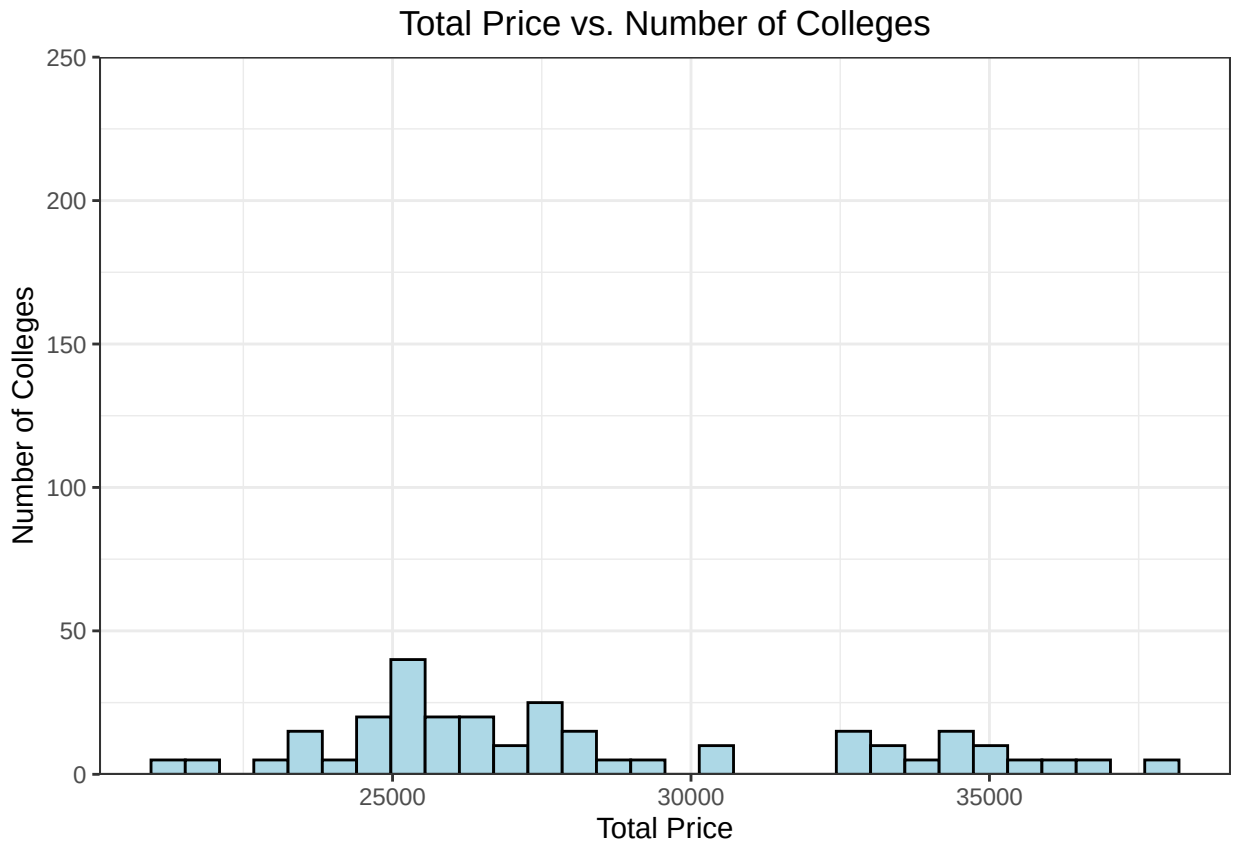
UC vs Cal State Deep Dive

After looking at the general price trends of colleges across the U.S. lets look at how California fairs looking at the pricings between it various public higher education institutions.

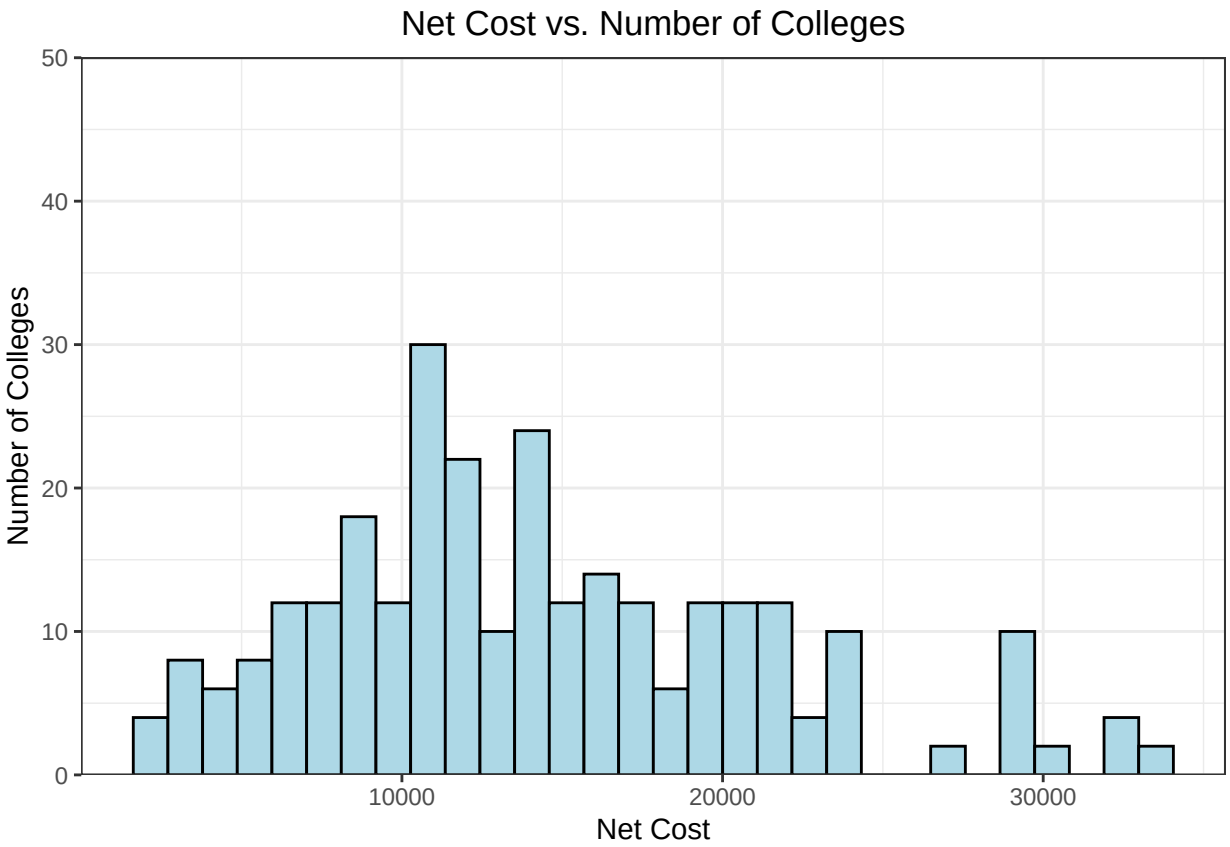
```
##First, we must create a new data frame to see specifically how each system

A<-income_and_cost %>%
filter(grepl('University of California:',name)) %>%
add_column(college_type="UC")


B<-income_and_cost %>%
filter(grepl('California State|California Polytechnic|San Diego State University
|San Francisco State University|San Jose State University',name)) %>%
add_column(college_type="Cal State")

california_colleges<-rbind(A,B) %>% select(-type)

view(california_colleges)
```

Since we are looking for how each college differs in pricing depending on what type of college it is in California, we create a new column called college_type classifying each college as either a UC, Cal State, or Community College.

```
## Here we are looing for any potential price outliers
ggplot(california_colleges,aes(x = total_price))+
geom_histogram(fill="lightblue",color="black",bins=30)+
theme_bw()+
scale_y_continuous(expand=c(0,0),limits=c(0,250))+
xlab("Total Price")+ylab("Number of Colleges")+
ggtitle("Total Price vs. Number of Colleges")+
theme(plot.title=element_text(hjust=0.5))
```




```
ggplot(california_colleges,aes(x = net_cost))+
geom_histogram(fill="lightblue",color="black",bins=30)+
theme_bw()+
scale_y_continuous(expand=c(0,0),limits=c(0,50))+
xlab("Net Cost")+ylab("Number of Colleges")+
ggtitle("Net Cost vs. Number of Colleges")+
theme(plot.title=element_text(hjust=0.5))
```



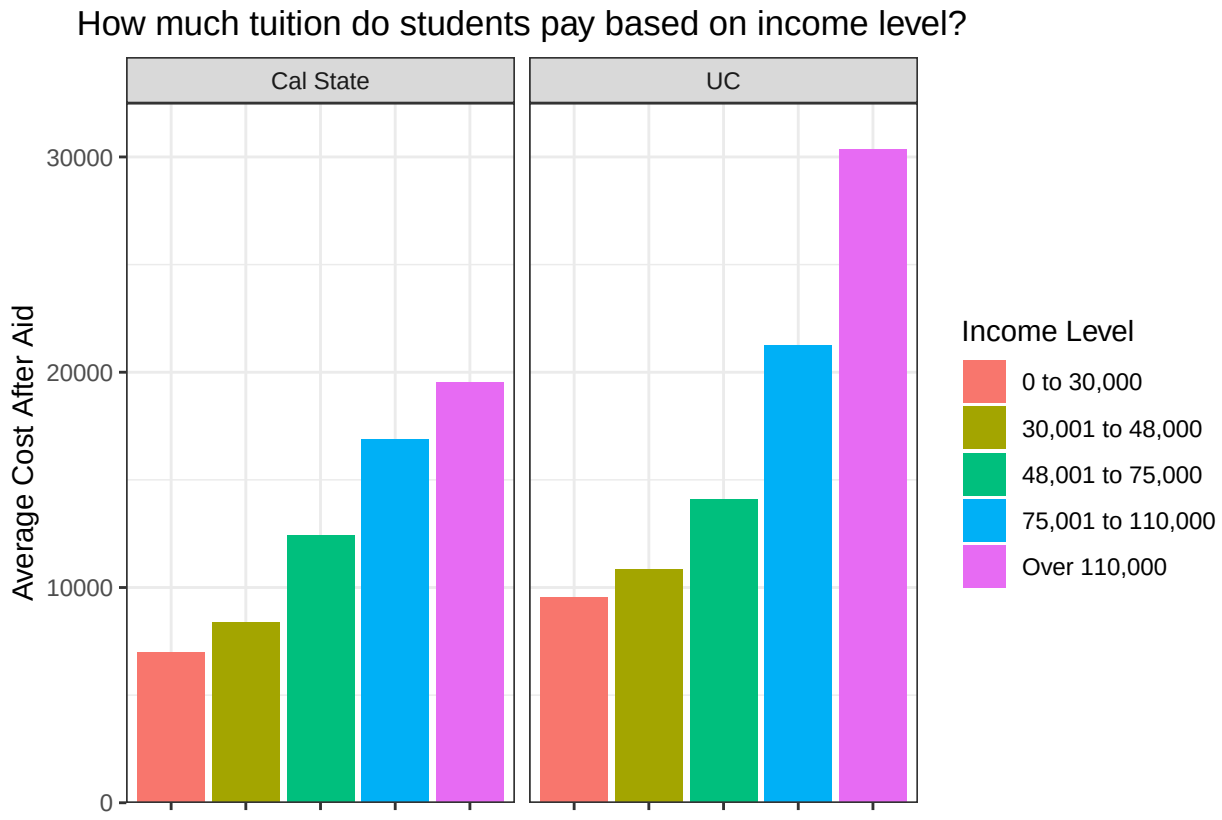
Both histograms make pretty even bell curves, with the outliers being not extreme enough to be excluded from the dataset.

```
#see how throughout California's Public College systems the prices compare
average_cost_by_income_lvl_california <-
california_colleges%>%
group_by(college_type, income_lvl) %>%
summarise(average_cost_after_aid_california= mean(net_cost))
```

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

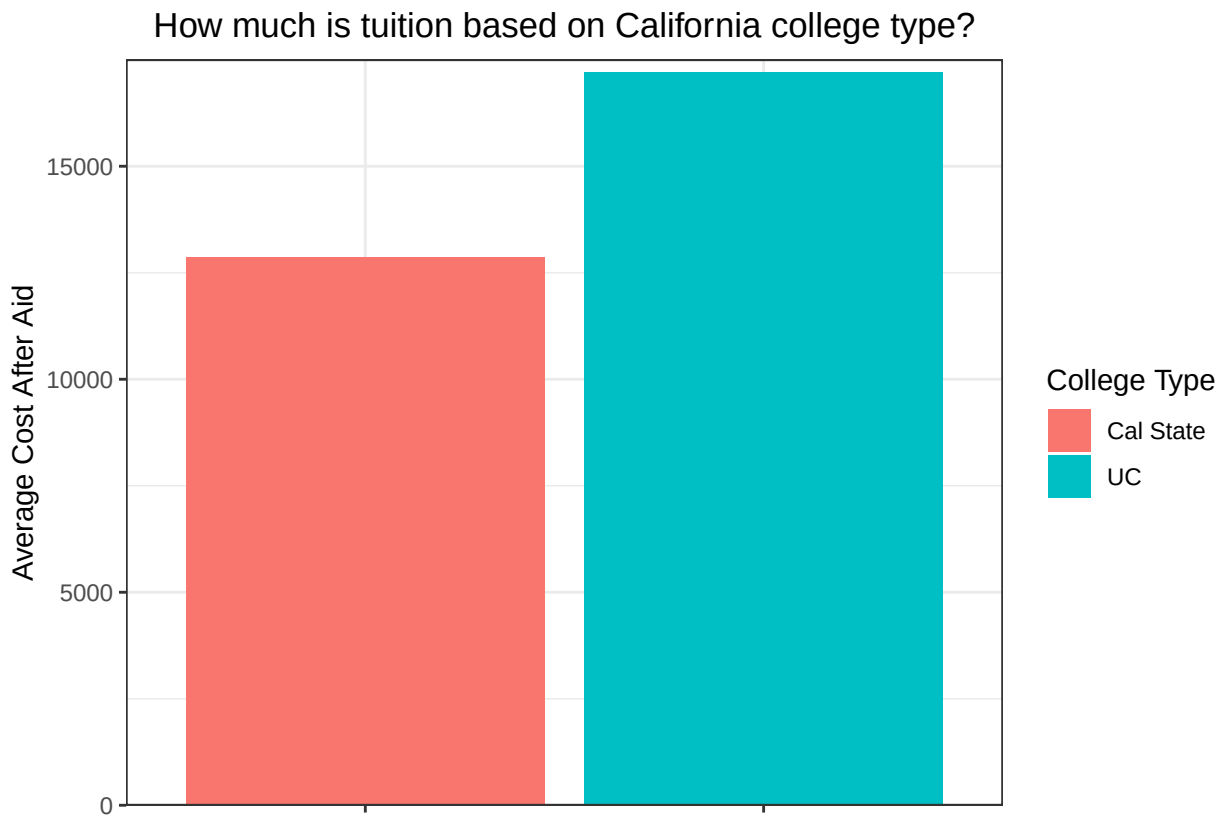
```
ggplot(average_cost_by_income_lvl_california)+
geom_col(aes(x=income_lvl,y=average_cost_after_aid_california,fill=income_lvl))+
ylab("Average Cost After Aid")+
xlab("")+
ggtitle("How much tuition do students pay based on income level?")+
facet_grid(~factor(college_type,levels=
c('Cal State','UC')))+
scale_x_discrete(limits=rev)+
theme_bw()+
theme(axis.text.x=element_blank())+
scale_x_discrete(guide = guide_axis(angle=90))+
theme(plot.title=element_text(hjust=0.5))+
guides(fill=guide_legend(title="Income Level"))+
scale_y_continuous(expand=c(0,0),limits=c(0,32500))
```

```
## Scale for x is already present.
## Adding another scale for x, which will replace the existing scale.
```



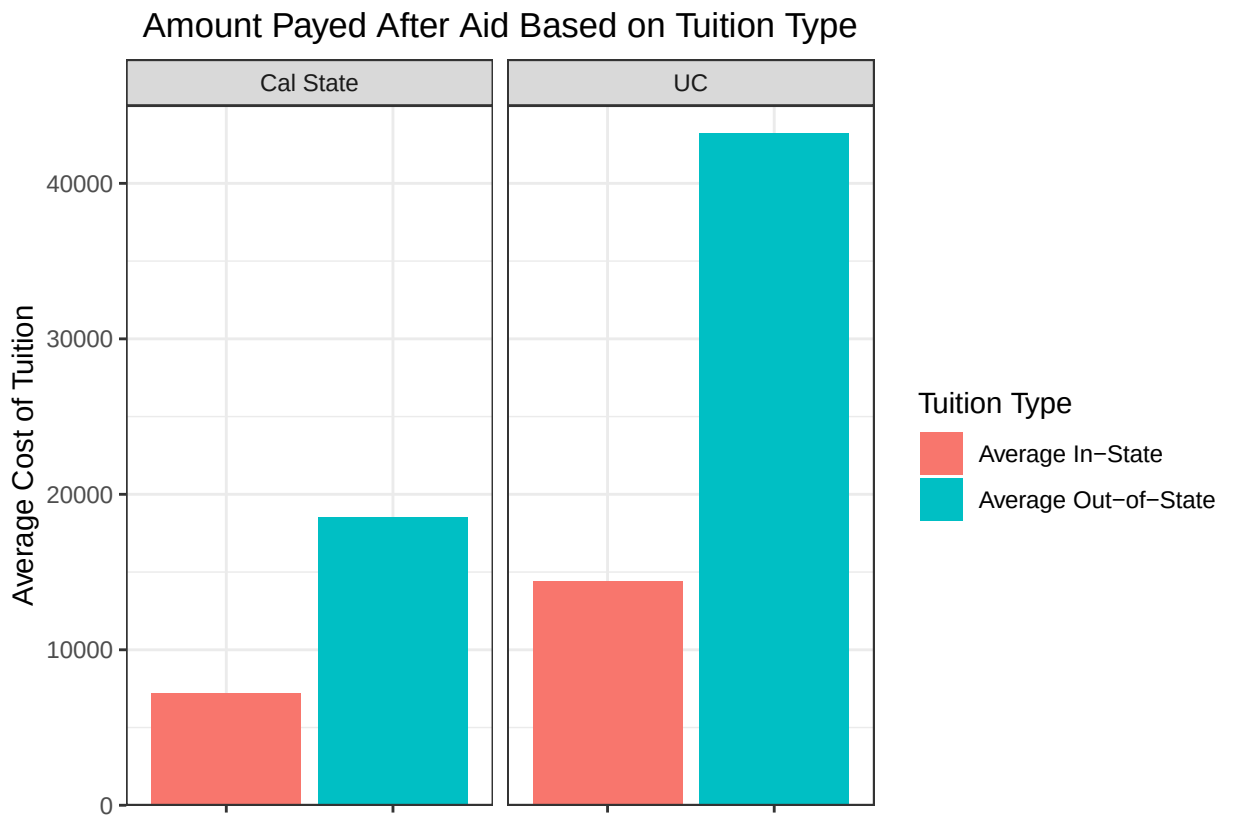
Again, the same pattern of lower income households paying less continues, however UC schools are more than $10,000 more expensive for individuals in the highest income bracket on average.

```
#See how much is paid after any aid based on California public higher level

average_cost_by_type_california<-
california_colleges %>%
group_by(college_type) %>%
summarise(average_cost_after_aid_california=mean(net_cost))

ggplot(average_cost_by_type_california)+
geom_col(aes(x=college_type,y=average_cost_after_aid_california,
fill=college_type))+
ylab("Average Cost After Aid")+
xlab("")+
ggtitle("How much is tuition based on California college type?")+ theme_bw()+
theme(axis.text.x=element_blank())+
guides(fill=guide_legend(title="College Type"))+
theme(plot.title=element_text(hjust=0.5))+
scale_y_continuous(expand=c(0,0),limits=c(0,17500))
```



Overall, UC schools are more significantly more expensive than Cal State schools.

```
##How do in state and out of state students fair in pricing
in_out_by_type_california<-
california_colleges%>%
group_by(college_type)%>%
summarise(avg_out_of_state=mean(out_of_state_tuition),
avg_in_state = mean(in_state_tuition))


in_out_by_type_california<-in_out_by_type_california%>%
pivot_longer(cols=c("avg_out_of_state","avg_in_state"),
names_to='averages',values_to='tuition') %>%
mutate(averages=str_replace(averages,
"avg_out_of_state","Average Out-of-State")) %>%
mutate(averages=str_replace(averages,
"avg_in_state","Average In-State"))

ggplot(in_out_by_type_california)+
geom_col(aes(x =averages,y=tuition,fill=averages))+
ylab("Average Cost of Tuition")+xlab("")+
ggtitle("Amount Payed After Aid Based on Tuition Type")+
facet_wrap(~college_type)+
theme_bw()+theme(axis.text.x=element_blank())+
guides(fill=guide_legend(title="Tuition Type"))+
theme(plot.title=element_text(hjust=0.5))+
scale_y_continuous(expand=c(0,0),limits=c(0,45000))
```



Similar to the patterns seen before, out-of-state students pay more than in-state students, however the amount paid is much higher for UC schools. Additionally the gap between in-state students tuition and out-of-state student tuition is far greater in the UC system.

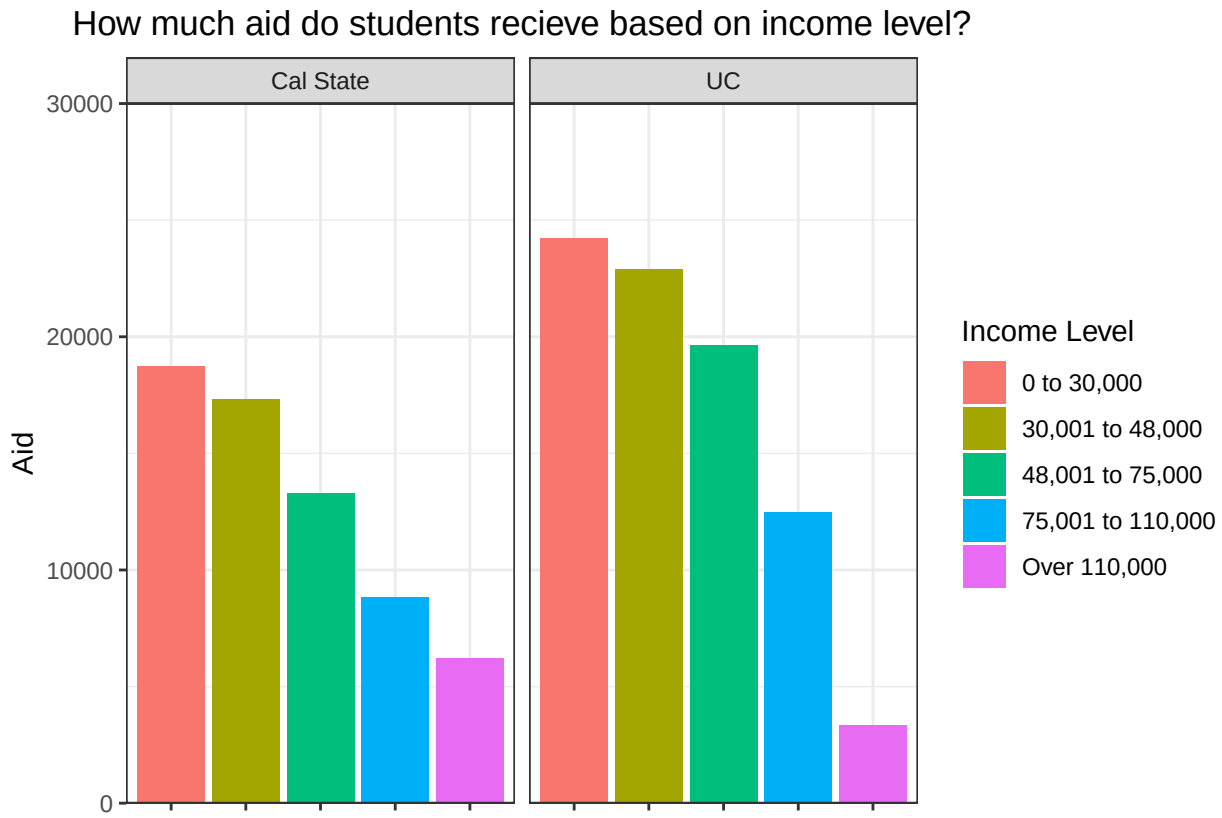
```
income_and_cost_aid_california<-
california_colleges %>%
mutate(aid=total_price-net_cost)

average_income_and_cost_aid_california<-
income_and_cost_aid_california %>%
group_by(college_type, income_lvl) %>%
summarise(average_aid= mean(aid))
```

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

```
ggplot(average_income_and_cost_aid_california)+
geom_col(aes(x=income_lvl, y=average_aid, fill=income_lvl))+
ylab("Aid")+
xlab("")+
ggtitle("How much aid do students recieve based on income level?")+
facet_grid(~factor(college_type,levels=
c('Cal State','UC'))) +
scale_x_discrete(limits=rev)+
theme_bw()+
theme(axis.text.x=element_blank())+
```

```
scale_x_discrete(guide = guide_axis(angle
= 90))+
theme(plot.title=element_text(hjust=0.5))+
guides(fill=guide_legend(title="Income Level"))+
scale_y_continuous(expand=c(0,0),limits=c(0,30000))
```

```
## Scale for x is already present.
## Adding another scale for x, which will replace the existing scale.
```



Although we have seen that the UC schools are more expensive, they offer a larger amount of financial aid to students in all income brackets besides the highest one compared to Cal State schools.

Part 3: Data Modeling and Predictions

3.1 t-test

T- tests will allow us to see if the means of two groups are similar to each other, Let's run a few t-tests to see if the price students in pay in different scenarios are similar.

First, let's see how much does financial aid play a role in a student's tuition.

```
t.test(income_and_cost$total_price,income_and_cost$net_cost)
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  income_and_cost$total_price and income_and_cost$net_cost
## t = 113.84, df = 28387, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  15835.18 16390.02
## sample estimates:
## mean of x mean of y
##  34297.53  18184.93
```

Here, we produced a t-value of 112.51, we can clearly see that there is a significant difference between cost before aid and cost after aid on average, this means that finical aid plays a crucial factor in student's tuitions.

Let's check if the price for public school is similar to those of private school:

```
public_price <-
  income_and_cost %>%
  filter (type == "Public") %>%
  select (net_cost)

private_price <-
  income_and_cost %>%
  filter (type == "Private") %>%
  select (net_cost)

t.test(private_price, public_price)
```

```
##
##  Welch Two Sample t-test
##
## data:  private_price and public_price
## t = 110.41, df = 15636, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  11412.02 11824.55
## sample estimates:
## mean of x mean of y
##  23789.00  12170.72
```

From the large t-value of 111.24 and sample estimates, we can clearly see that there is a difference between the average tuition in a private college and that of a public one.

With these information on hand, let's try to get closer our motivating question, and find out if the aid from public and private school offers are similar at all.

To do this, we need to create a few new variables.

```
public_aid <-
  income_and_cost %>%
  filter (type == "Public") %>%
  mutate (aid = total_price - net_cost) %>%
  select (aid)
```

```
private_aid <-
  income_and_cost %>%
  filter (type == "Private") %>%
  mutate (aid = total_price - net_cost) %>%
  select (aid)

t.test(private_aid, public_aid)
```

```
##
##  Welch Two Sample t-test
##
## data:  private_aid and public_aid
## t = 107.04, df = 11595, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  14469.67 15009.50
## sample estimates:
## mean of x mean of y
## 23422.046  8682.458
```

Though much cheaper than private colleges, we can see that public schools tends to offer students much less of finial aid from this test, and that the average finical aid given by these two types of insitutions are dramatically different from each other.

3.2 correlation tests, checking for signaifice.

Back to our motivating question, we want to know if lower income students and higher income students have to pay the same price for college. In other words, if there is a correlation between the student family's income level and the tuition they pay.

With a correlation test, we can check how significant are the correlation between income level and tuition paid.

```
income_lvl_factor <-
  factor(income_and_cost$income_lvl,
         levels = as.character(unique(income_and_cost$income_lvl)))

income_lvl_num <- as.numeric(income_lvl_factor)

cor(income_lvl_num,
    income_and_cost$net_cost,
    method = c("pearson", "kendall", "spearman"))
```

```
## [1] 0.3293511
```

The 0.32 correlation level shows that there is indeed a low positive correlation between the two variables. Though not a significant amount, it is true that the higher the student family's income level, the more they have to pay for the tuition.

Something else we want to know is that if this correlation is stronger in public, private or for profit institutions, let us filter to the specific group only and find out.

```
public_schools <- income_and_cost %>%
  filter (type == "Public")

public_income_lvl_factor <-
  factor(public_schools$income_lvl,
         levels = as.character(unique(public_schools$income_lvl)))

public_income_lvl_num <- as.numeric(public_income_lvl_factor)
cor(public_income_lvl_num,
    public_schools$net_cost,
    method = c("pearson", "kendall", "spearman"))
```

```
## [1] 0.4467293
```

```
private_schools <- income_and_cost %>%
  filter (type == "Private")

private_income_lvl_factor <-
  factor(private_schools$income_lvl,
         levels = as.character(unique(private_schools$income_lvl)))

private_income_lvl_num <-
  as.numeric(private_income_lvl_factor)

cor(private_income_lvl_num,
    private_schools$net_cost,
    method = c("pearson", "kendall", "spearman"))
```

```
## [1] 0.4764292
```

```
fpt_schools <- income_and_cost %>%
  filter (type == "For Profit")

fpt_income_lvl_factor <-
  factor(fpt_schools$income_lvl,
         levels = as.character(unique(fpt_schools$income_lvl)))

fpt_income_lvl_num <-
  as.numeric(fpt_income_lvl_factor)
  cor(fpt_income_lvl_num,
  fpt_schools$net_cost,
  method = c("pearson", "kendall", "spearman"))
```

```
## [1] 0.265049
```

From this analysis, we can see that there is a surprising moderate positive correlation between the two variables for public and private institutions. Meaning that the student's income level do play a good factor in terms of paying tuition in these type of colleges. On the other hand, we see that there is only a 0.25 correlation between the two variable for the for-profit institutions. This means that the students pay about the same amount of tuition no matter what income level they belong in. This makes sense since the for-private institutions are used to generate generate revenue and the students that attend those colleges do not receive state or federal aid.

3.3 - Linear Modeling, Predictions

The predict() function in R can be used to perdict values based on the input data. We've looked at the relationships between the variables, now let's try to predict how much will the college cost based on the student family's income level using linear modeling methods.

```
netcost_vs_incomelvl <- lm(net_cost~income_lvl, data = income_and_cost)
pricedata <- tibble(income_lvl = unique(income_and_cost$income_lvl))
predict_price <- pricedata %>%
mutate(price = predict(netcost_vs_incomelvl, pricedata))
predict_price
```

```
## # A tibble: 5 x 2
##   income_lvl         price
##   <chr>               <dbl>
## 1 0 to 30,000       14294.
## 2 30,001 to 48,000  15093.
## 3 75,001 to 110,000 20708.
## 4 48,001 to 75,000  17629.
## 5 Over 110,000      23970.
```

The model is able to see the pay gap between the income levels clearly. On average, we estimate a student to pay about 23948.54 USD for the college they are attending if their family's income level is over 110000, and only 14412.97 USD if their family's income level is about 0~30000.

Let's check out how much college will cost depending on the type of college

```
type_vs_cost <- lm(net_cost~type, data = income_and_cost)
type_data <- tibble(type = unique(income_and_cost$type))
type_predict <- type_data %>%
mutate(price = predict(type_vs_cost, type_data))
type_predict
```

```
## # A tibble: 3 x 2
##   type        price
##   <chr>       <dbl>
## 1 Public     12171.
## 2 Private    23789.
## 3 For Profit 25290.
```

We can see that there is quite a bit of difference between the tuition one have to pay for public vs private institutions. Let us combine these information we know and predict the pricate one have to pay in different income levels, specified by public and private institutions.

```
pb_netcost_vs_incomelvl <- lm(net_cost~income_lvl, data = public_schools)
pb_pricedata <- tibble(income_lvl = unique(public_schools$income_lvl))
pb_predict_price <- pb_pricedata %>%
mutate(price = predict(pb_netcost_vs_incomelvl, pb_pricedata))
pb_predict_price
```

```
## # A tibble: 5 x 2
##   income_lvl         price
```

```
##   <chr>              <dbl>
## 1 0 to 30,000        8811.
## 2 30,001 to 48,000   9671.
## 3 75,001 to 110,000 14757.
## 4 48,001 to 75,000  12191.
## 5 Over 110,000      16191.
```

```
pr_netcost_vs_incomelvl <- lm(net_cost~income_lvl, data = private_schools)
pr_pricedata <- tibble(income_lvl = unique(private_schools$income_lvl))
pr_predict_price <- pr_pricedata %>%
mutate(price = predict(pr_netcost_vs_incomelvl, pr_pricedata))
pr_predict_price
```

```
## # A tibble: 5 x 2
##   income_lvl         price
##   <chr>              <dbl>
## 1 0 to 30,000       19435.
## 2 30,001 to 48,000  20228.
## 3 48,001 to 75,000  22769.
## 4 75,001 to 110,000 26283.
## 5 Over 110,000      30785.
```

```
fpt_netcost_vs_incomelvl <- lm(net_cost~income_lvl, data = fpt_schools)
fpt_pricedata <- tibble(income_lvl = unique(fpt_schools$income_lvl))
fpt_predict_price <- fpt_pricedata %>%
mutate(price = predict(fpt_netcost_vs_incomelvl, fpt_pricedata))
fpt_predict_price
```

```
## # A tibble: 5 x 2
##   income_lvl         price
##   <chr>              <dbl>
## 1 0 to 30,000       22394.
## 2 30,001 to 48,000  23001.
## 3 48,001 to 75,000  25708.
## 4 75,001 to 110,000 27815.
## 5 Over 110,000      29076.
```

From these predictions, we see that the positive correlations we explored earlier had been recognized by the model, leading us to the predictions of the price one student should pay in different intuitions between different income levels.

3.4 Insights and Improvements

From our t-tests, we were able to conclude that financial aid plays a role in reducing tuition costs for students. This financial help is increasing depending on the family's financial income which supports the idea that these universities are helping those who need it the most. However, we also compared the relationship between these two variables using the comparison test which showed that although there is a correlation between income level and net costs, the correlation isn't super high. We then ran the same comparison test but filtering for each type of institution. These results showed that for public and private institutions there is a higher positive correlation between family income level and net cost than for non-profit institutions. In other words, your family's income plays a bigger role in reducing your tuition at private and public schools than non-profit ones. Between public and private, private institutions have a slightly higher correlation than public ones.

```
This is most likely attributed to the fact that private and public institutions offer more financial ai
```

Although we found an answer to our question that we feel is accurate due to our extensive data analysis, we feel that there are steps we would've taken if we were to re-do the project that could've provided even greater insight on the topic we were analyzing. For instance, the data set we used was compiled of data only from 2018. More updated data would've provided a more accurate conclusion to what the numbers would look like today, especially since big changes such as Covid could've influenced the collegiate system. Another thing we could do would be to research and compare even more variables related to our topic. For instance, comparing the percentage of tuition that financial aid covers for each type of institution. Although this doesn't give us the answer to our question, it could provide insight if we made our question more broad.

Part 5 - Conclusion

Through our linear modeling and predictions, we were able to see how the insights we derived are reflected in the different prices predicted. For each type of institution, prices went up in relation to the family's income level. When analyzing by type of institution, the difference in price from the highest income level to the lowest is largest in private, followed by public and then non-profit. This again shows how your income level plays a larger role in reducing tuition costs for each institution in that respective order. When predicting the average net costs for each type of school, public institutions are significantly cheaper than both private and non-profit ones. When breaking it down by family income level, our model shows how even if you are in the lowest income level category at a private or non-profit school, you will most likely have to pay more than a student in the highest income level at a public school, even if the student is out-of-state. This analysis allows us to conclude that if you are seeking the cheapest education possible, you should look for an in-state public school to attend. However, if you are looking for the largest "discount" from financial aid that you will receive based on your income level, a private institution would be favorable, followed by public and then non-profit. We hope that this information can be useful for students when it comes to finding the most affordable college to go to.