Final_University_Data

Myli Brown

2022-11-26

R Markdown

```
library(hexbin)
library(patchwork)
library(tinytex)
library(ggplot2)
library(ggExtra)
library(dplyr)

##
## 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(GGally)

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

library(ggpubr)
library(ggridges)
library(RColorBrewer)
library(wesanderson)
library(gridExtra)

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

library(cowplot)

##
## Attaching package: 'cowplot'
## The following object is masked from 'package:ggpubr':
##
##     get_legend
```

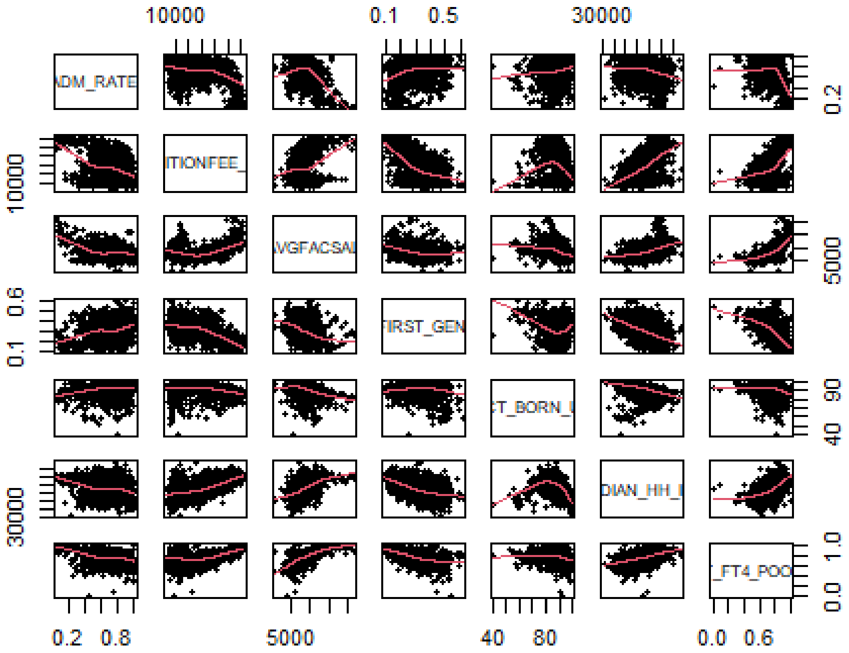
```
## The following object is masked from 'package:patchwork':
##
##     align_plots
```

```r
uni<-read.csv("D:/Fall 2022/Programming/Universities.csv",header=TRUE)
```

Introduction: The data that I am using in this project is the College Scorecard data, collected and reported by the United States Department of Education. This data was updated to its most recent level in September of 2022. This data set holds data from thousands of colleges throughout the United States. I decided to only use data from colleges in the Contiguous U.S. In this data set, I investigated the relationship between 4th year student retention rates and admission rate, in-state tuition, average faculty salary, percentage of first generation students, percentage of students born in the United States, and the median household income. I would like to observe the effect of each of these variables on 4th year retention rates. I also looked at how the region of the U.S. affected these same variables.

Exploratory Data Analysis: I started out by getting a big picture of the data in the dataset.

```r
pairs(~ADM_RATE + TUITIONFEE_IN + AVGFACSAL + FIRST_GEN + PCT_BORN_US +
MEDIAN_HH_INC + RET_FT4_POOLED, data = uni, panel = panel.smooth, pch = 20)
```



This figure shows correlation between all of the variables in the data. It allows for visualization of possible trends in the data set. The figure revealed strong relationships between four-year retention rate and all of the rest of the variables. It was still unclear how strong of correlation there was between the variables, so I created the correlation chart seen below.

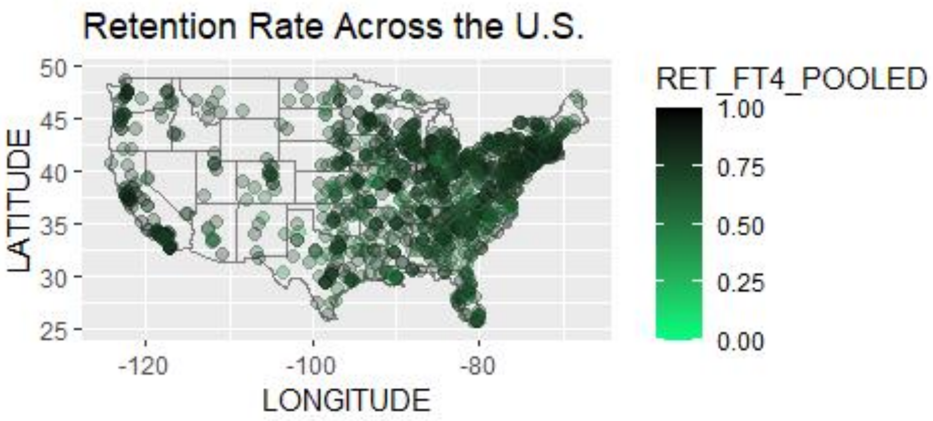
```
cor(uni[,unlist(lapply(uni, is.numeric))])

##                     UNITID    LATITUDE   LONGITUDE    ADM_RATE
TUITIONFEE_IN
## UNITID         1.00000000  0.10730936  0.10231779  0.12951931   -
0.06518964
## LATITUDE       0.10730936  1.00000000  0.11844396  0.07843752
0.19052463
## LONGITUDE      0.10231779  0.11844396  1.00000000 -0.07217803
0.13548455
## ADM_RATE       0.12951931  0.07843752 -0.07217803  1.00000000   -
0.35010211
## TUITIONFEE_IN -0.06518964  0.19052463  0.13548455 -0.35010211
1.00000000
## AVGFACSAL     -0.03782516  0.09087955  0.01756730 -0.37064391
0.23401239
## AGE_ENTRY      0.05176931 -0.13115846 -0.13758385  0.19470780   -
0.30308363
## FIRST_GEN      0.02033584 -0.23817786 -0.14983884  0.25094920   -
0.52392512
## PCT_BORN_US    0.12628576  0.14620053  0.10846876  0.21211282   -
0.12655410
## MEDIAN_HH_INC -0.03413248  0.23693367  0.13040773 -0.22899395
0.53362157
## RET_FT4_POOLED -0.07538439  0.15526983  0.04195836 -0.28206492
0.34575396
##                  AVGFACSAL    AGE_ENTRY   FIRST_GEN  PCT_BORN_US
MEDIAN_HH_INC
## UNITID        -0.03782516  0.051769313  0.02033584  0.126285761   -
0.03413248
## LATITUDE       0.09087955 -0.131158457 -0.23817786  0.146200525
0.23693367
## LONGITUDE      0.01756730 -0.137583846 -0.14983884  0.108468761
0.13040773
## ADM_RATE      -0.37064391  0.194707804  0.25094920  0.212112821   -
0.22899395
## TUITIONFEE_IN  0.23401239 -0.303083628 -0.52392512 -0.126554097
0.53362157
## AVGFACSAL      1.00000000 -0.320454748 -0.32467988 -0.414720424
0.51023024
## AGE_ENTRY     -0.32045475  1.000000000  0.63051181  0.001912073   -
0.30289310
## FIRST_GEN     -0.32467988  0.630511807  1.00000000 -0.148436594   -
0.47542982
## PCT_BORN_US   -0.41472042  0.001912073 -0.14843659  1.000000000   -
0.34810303
```

```
## MEDIAN_HH_INC   0.51023024 -0.302893103 -0.47542982 -0.348103030
1.00000000
## RET_FT4_POOLED  0.64581549 -0.441940378 -0.54925711 -0.206269575
0.49819524
##                RET_FT4_POOLED
## UNITID            -0.07538439
## LATITUDE           0.15526983
## LONGITUDE          0.04195836
## ADM_RATE          -0.28206492
## TUITIONFEE_IN      0.34575396
## AVGFACSAL          0.64581549
## AGE_ENTRY         -0.44194038
## FIRST_GEN         -0.54925711
## PCT_BORN_US       -0.20626958
## MEDIAN_HH_INC      0.49819524
## RET_FT4_POOLED     1.00000000
```

This figure shows the numerical correlation between variables in the data and confirms the high correlation that the variables had with retention rate.

I also wanted to look at the affect of location on fourth year retention to see whether there will be a difference in the data based on location. The scatter plot on the map below shows that there are some areas in the United States that have much higher retention than others. However, the points on the scatter plot are a bit too dense to get a real sense of trends in the data.

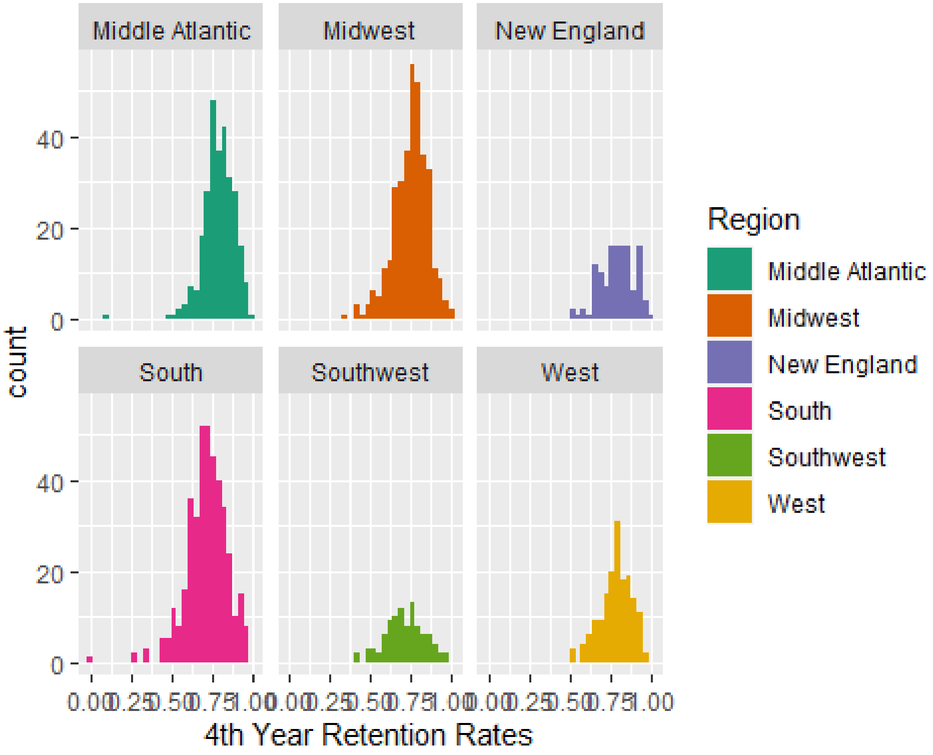
```
uni%>%
 ggplot(aes(LONGITUDE, LATITUDE, color =
RET_FT4_POOLED))+borders("state")+geom_point(size=2,
alpha=.3)+coord_quickmap()+ggtitle("Retention Rate Across the
U.S.")+scale_colour_gradient(low = "springgreen", high = "black")
```



In order to see how each region of the U.S. affects retention and the other variables in the data, I separated the colleges by region in the U.S.

The plot below is separated by U.S. regions and shows the retention rates in each region. the height of the histograms do not have as much importance as the spread of the histogram because the height shows the count of colleges whereas the spread shows the variability in retention rate. The height varies from region to region based on the number of colleges and also the area represented in each region. The New England area is relatively small thus it will have less colleges than the South which has much more land area. The spread of the data in this graph shows that for New England and West U.S., there is less variability in the data. It seems that in those two regions, the retention rate is more consistently higher than in the other regions. On the other hand, in the South, the data is spread out much further and the data seems slightly skewed to the left showing that colleges in the south tend to have lower retention rates.

```
ggplot(uni, aes(RET_FT4_POOLED, fill =
Region))+geom_histogram()+facet_wrap(~Region)+scale_fill_brewer(palette="Dark
2")+xlab("4th Year Retention Rates")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

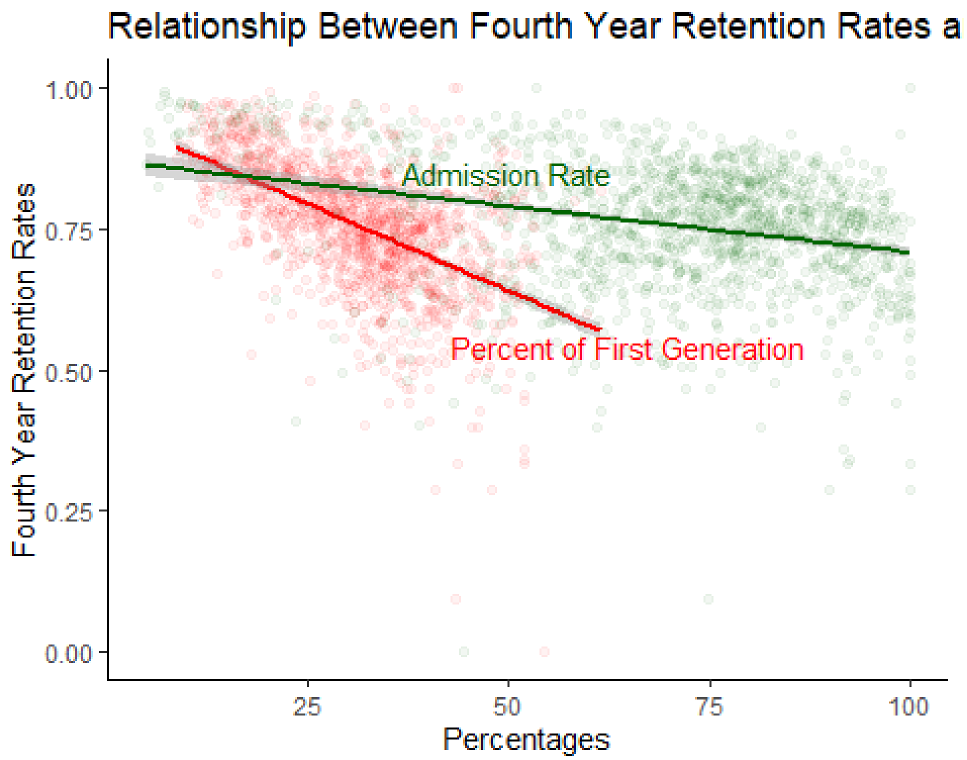


This graph shows differences between the regions that I will be looking into more deeply in the data analyses.

Data Analyses My overall purpose throughout the analyses will be to examine how each variable effects retention rates and how each variable is affected by geographical location.

The first variables that I am analyzing are the student admission rate and the percent of first generation students. I will be analyzing these variables as they relate to retention rate.

```
gg <- ggplot(uni, aes(y=RET_FT4_POOLED))
gg <- gg + geom_smooth(aes(x=FIRST_GEN*100), method="lm", colour="red")
gg <- gg + geom_point(aes(x=FIRST_GEN*100), alpha=.05, colour="red")
gg <- gg + geom_smooth(aes(x=ADM_RATE*100), method="lm", colour="darkgreen")
gg <- gg + geom_point(aes(x=ADM_RATE*100), alpha=.05, colour="darkgreen")
gg <- gg + annotate("text", x=65, y=.54, label="Percent of First Generation",
color="red")
gg <- gg + annotate("text", x=50, y=.85, label="Admission Rate",
color="darkgreen")
gg <- gg + ggtitle("Relationship Between Fourth Year Retention Rates and
Student Proportions")
gg <- gg + ylab("Fourth Year Retention Rates")
gg <- gg + xlab("Percentages")
gg <- gg + theme_classic()
gg

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



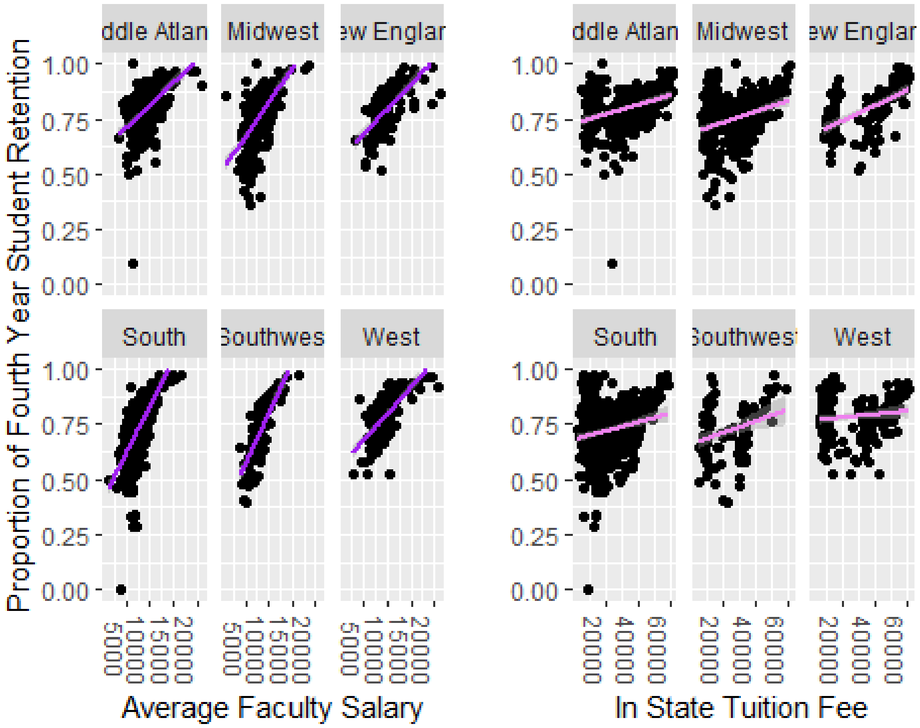
The graph shows that as student admission rate decreases, the retention rate of students increases. This means that as the college accepts less students, they retain more of their students.

The graph also shows that as the percent of first generation students decreases, the fourth year retention rates increase. This means that as the number of first generation students increases, the college loses more of its students. The line that represents percent of first generation students only goes to roughly 60 percent because there are not any colleges in the data that have more than sixty percent of first generation students.

The next variables that I will be looking at are average monthly faculty salary and the in-state tuition fee. I will also look at how these variables are related to retention rate.

```
f1 <- uni%>%
  ggplot(aes(x=AVGFACSAL,
y=RET_FT4_POOLED))+geom_point(color="black")+geom_smooth(color = "purple",
method = "lm")+xlab("Average Faculty Salary")+ylab("Proportion of Fourth Year
Student Retention")+facet_wrap(~Region)+theme(axis.text.x =
element_text(angle = 270))+ylim(0,1)

f2 <- uni%>%
  ggplot(aes(x=TUITIONFEE_IN,
y=RET_FT4_POOLED))+geom_point(color="black")+geom_smooth(color = "violet",
method = "lm")+xlab("In State Tuition
Fee")+ylab("")+facet_wrap(~Region)+theme(axis.text.x = element_text(angle =
270))

f3 <- ggplot(uni, aes(x=AVGFACSAL, y=TUITIONFEE_IN))+geom_smooth(method="lm")
```

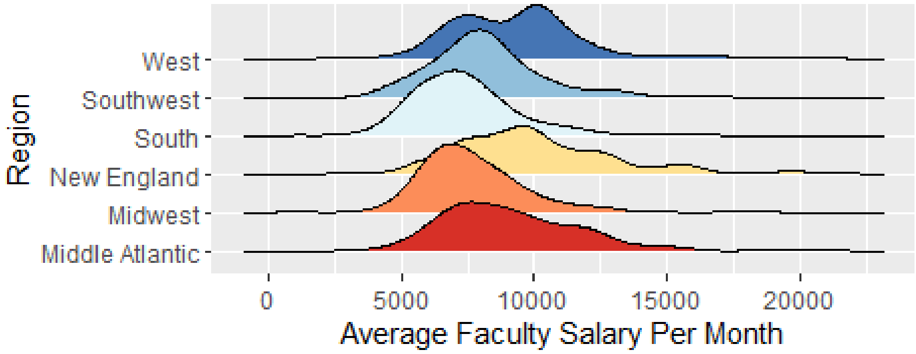
```
f1+f2

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

## Warning: Removed 67 rows containing missing values (`geom_smooth()`).

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

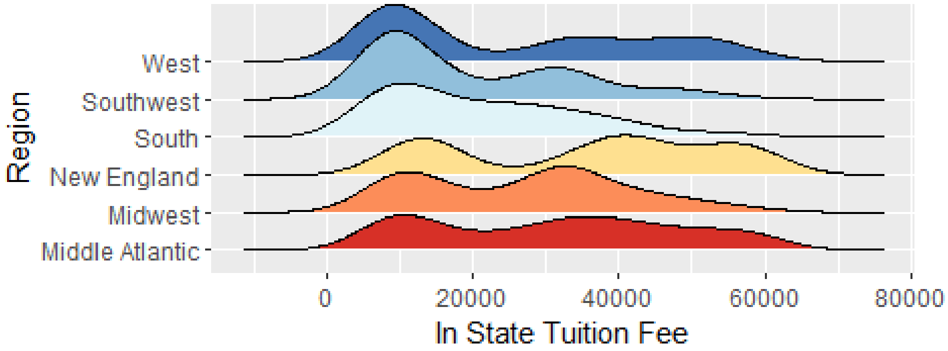


These graphs both show positive correlation with student retention rate. Faculty salary seems to be more heavily correlated than in-state tuition fees. This means that as teachers are paid more, the number of students that drop out of college is lower. Retention rate's relationship with in-state tuition fees shows slightly less correlation than with faculty salary but it still shows positive correlation. This means that as tuition fees increase, the amount of students dropping out of college decreases. It is interesting to note that the correlation for both variables in the Western region of the U.S. is less extreme than in the other regions. The West tended to have higher retention rates regardless of the other variables.

```
p1 <- ggplot(uni, aes(x=AVGFACSAL, y=Region))+
  geom_density_ridges(aes(fill = Region))
+scale_fill_brewer(palette="RdYlBu") +theme(legend.position =
"none")+xlab("Average Faculty Salary Per Month")
p2 <- ggplot(uni, aes(x=TUITIONFEE_IN, y=Region))+
  geom_density_ridges(aes(fill = Region))
+scale_fill_brewer(palette="RdYlBu")+xlab("In State Tuition
Fee")+theme(legend.position="none")
p1/p2
```

```
## Picking joint bandwidth of 655

## Picking joint bandwidth of 4840
```





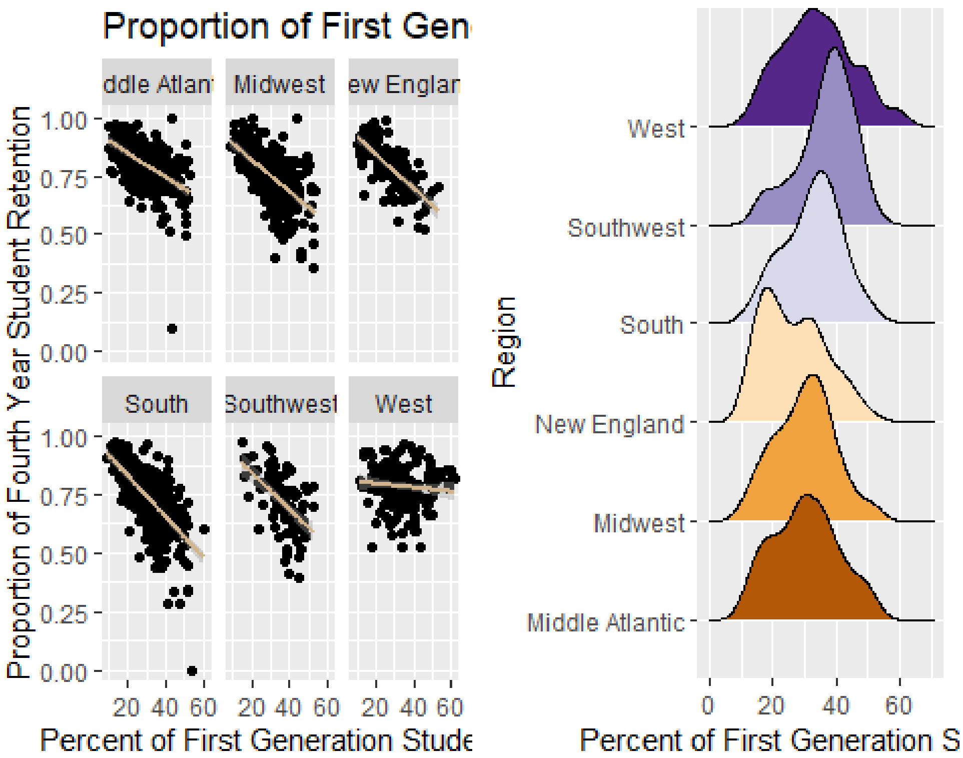
Looking a bit deeper into how the region affects tuition fees and faculty salaries, the graph shows that in New England and the West, in-state tuition fees can get very high, but with that, the faculty salary is also generally higher than in most other regions.

The graphs also show a typically bimodal relationship in tuition fees throughout the regions. This shows that there are a lot of schools that have higher tuition fees and a lot that have lower tuition fees instead of most of the schools having around the same tuition.

Starting to look a bit closer at individual relationships between variables and regions, the next graphs show the relationship between the proportion of first generation students and retention rates as well as showing the difference between each region.

```
fg<- uni%>%
  ggplot(aes(x=FIRST_GEN*100,
y=RET_FT4_POOLED))+geom_point()+geom_smooth(color="tan", method =
"lm")+xlab("Percent of First Generation Students")+ylab("Proportion of Fourth
Year Student Retention")+ggtitle("Proportion of First Generation Students
Relationship with Student Retention")+facet_wrap(~Region)

fg2<- ggplot(uni, aes(x=FIRST_GEN*100, y=Region))+
  geom_density_ridges(aes(fill = Region))
+scale_fill_brewer(palette="PuOr")+xlab("Percent of First Generation
Students")+theme(legend.position="none")
```

```
ggarrange(fg, fg2)

## `geom_smooth()` using formula = 'y ~ x'
## Picking joint bandwidth of 3.04
```



It is seen that the proportion of first generation students is negatively correlated with fourth year student retention rates. This shows that as more students are first generation college students, the school loses more students. It can also be seen that in the West region of the United States, student retention is less impacted than in the rest of the regions, with almost no correlation shown between the variables.

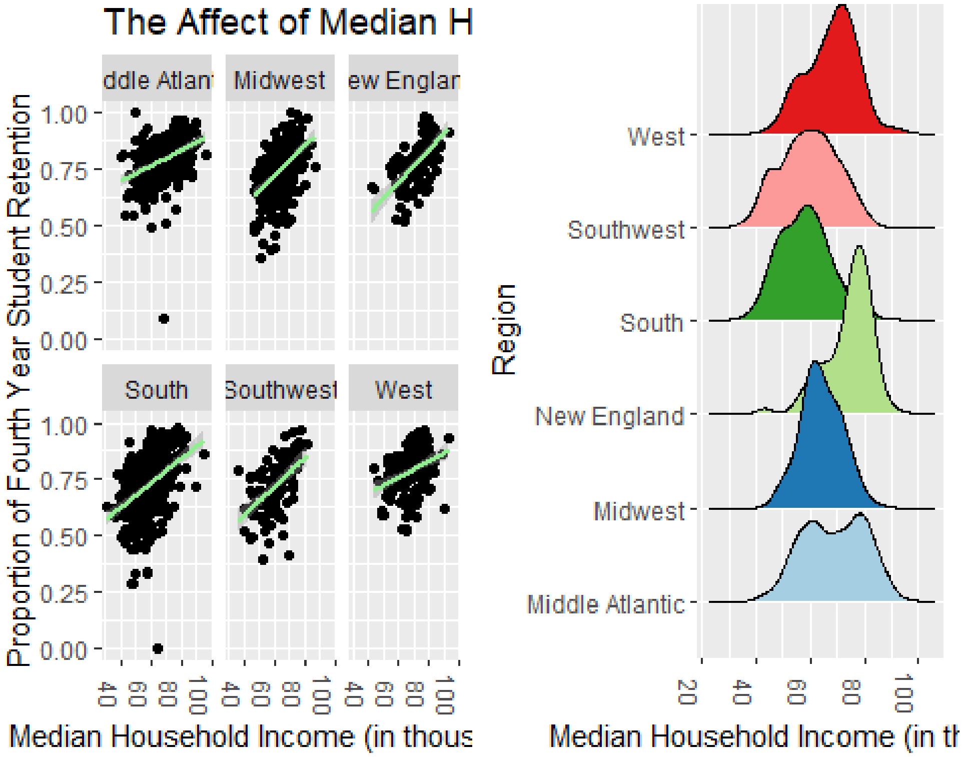
The graph of first generation students by region shows the South and Southwest regions as having a higher proportion of first generation students than the other regions. The New England region is bimodal and also shows the least proportion of first generation students.

The next variable I looked at was median household income as it relates to region and student retention rates.

```
hhi<- uni%>%
  ggplot(aes(x=MEDIAN_HH_INC/1000,
y=RET_FT4_POOLED))+geom_point()+geom_smooth(color = "lightgreen", method =
"lm")+xlab("Median Household Income (in thousands)")+ylab("Proportion of
Fourth Year Student Retention")+ggtitle("The Affect of Median Household
Income on Student Retention")+facet_wrap(~Region)+theme(axis.text.x =
element_text(angle = 270))
```

```
hhi2<- ggplot(uni, aes(x=MEDIAN_HH_INC/1000, y=Region))+
  geom_density_ridges(aes(fill = Region))
+scale_fill_brewer(palette="Paired")+theme(axis.text.x = element_text(angle =
270))+xlab("Median Household Income (in
thousands)")+theme(legend.position="none")

ggarrange(hhi, hhi2)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## Picking joint bandwidth of 2.93
```



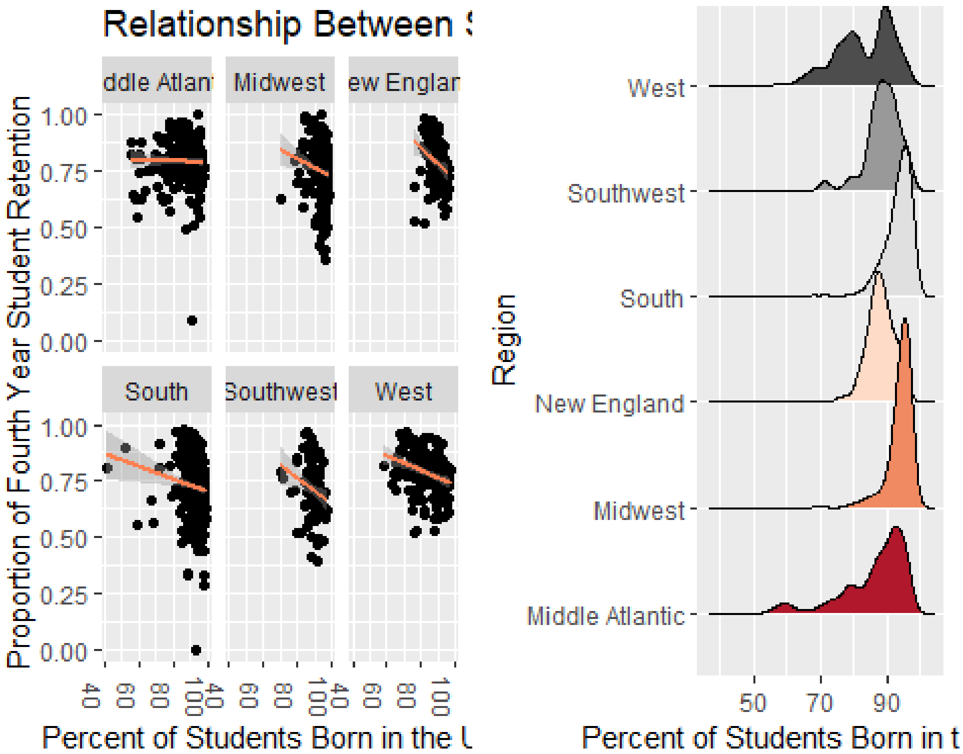
There is a strong positive correlation between household income and fourth year student retention rates. This shows that as a student's household income increases, they are less likely to drop out of college. The regions show almost the same correlation throughout.

It is also seen that household income is highest in the West, Middle Atlantic, and New England regions, however there is less variation in New England as there is in the other regions. There is also a bimodal distribution in the Middle Atlantic region showing that there are more areas with higher and lower household incomes as opposed to a middle income level.

The next graph shows the relationship between percent of students born in the United States and fourth year retention rates.

```
us<- uni%>%
  ggplot(aes(x=PCT_BORN_US,
y=RET_FT4_POOLED))+geom_point()+geom_smooth(color="coral", method =
"lm")+xlab("Percent of Students Born in the U.S.")+ylab("Proportion of Fourth
Year Student Retention")+ggtitle("Relationship Between Students Born in the
U.S. and Student Retention")+facet_wrap(~Region)+theme(axis.text.x =
element_text(angle = 270))

us2<- ggplot(uni, aes(x=PCT_BORN_US, y=Region))+
  geom_density_ridges(aes(fill =
Region))+scale_fill_brewer(palette="RdGy")+theme(legend.position="none")+xlab
("Percent of Students Born in the U.S.")

ggarrange(us, us2)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## Picking joint bandwidth of 1.6
```



There is a negative correlation between Percent of students born in the U.S. and fourth year retention rates except in the Middle Atlantic region. This shows that as the college has more students born in the United States, there are more students likely to drop out of college early.

It is also shown that in the West and Southwest, there are more students born outside of the U.S that in the other regions.

The data used in this study should be accurate and recent, but due to the confidentiality regarding college statistics, the study is limited to those colleges who gave unrestricted access to their information. This might introduce a slight bias in the data analyses because the types of schools likely to keep their information private are not represented in the analyses.

Conclusion:

There exists strong correlation between student retention rates and the other variables in this dataset. These correlations can give insight into the data showing what could have an effect on retention and possible solutions that would decrease the number of students dropping out.

First, the college's admission rate is negatively correlated with student retention rates. This is due to the fact that colleges who are more picky on admitting students will typically end up with students that treat their education more seriously, have better grades and test scores, and who are more driven to stay in college.

Next, as the percent of first generation students increases, the fourth year retention rate decreases. This could possibly be explained due to the fact that first generation students will likely have less college support than students whose parents went to college. With parents who did not go to college, a student may have less help navigating the transition to college life and might end up dropping out due to difficulties adjusting. Looking at the regions, it can be seen that the West region seems to have highest variability in first generation students than the rest of the regions. Similarly, the retention rates in the west seem less affected by percent of first generation students than in other regions.

Student retention also seems greatly affected by faculty salary. It seems the more money that faculty are paid, the higher the student retention rates are for the college. This makes sense because faculty that are paid well should be more likely to enjoy their job and therefore will put more effort into teaching. This will have a positive impact on their teaching, encouraging students to stay in college and not drop out. An increase in faculty salary can be seen in the New England and Western region of the United States which is understandable due to the typically high cost of living in those regions.

In-state tuition fee is positively correlated to retention rates as well. This also makes sense because if a student pays more towards their education, they are less likely to want to drop out and thus waste the money they have already spent. In-state tuition follows the normal trends expected by region. Regions with a higher cost of living typically have higher tuition rates. However,it is interesting to note that for each region, it seems that there are colleges with much higher tuition fees and much lower tuition fees, creating a gap in the data and the bimodal distribution seen in each region.

Median household income has an unsurprising relationship with retention rates. As the median household income increased, the fourth year retention rates increased. This could possibly be explained due to the fact that students with more money can better afford to go to college and are therefore less likely to drop out due to financial instability. The median household income by region followed the cost of living trend and thus did not reveal much about the data.

Finally, the percent of students born in the United States has a surprising correlation with student retention rates. The colleges with more students born in the United States had lower retention rates. This relationship can possibly be explained due to the fact that students born outside of the U.S. probably had to work a lot harder to be able to enroll in college than students who were born in the U.S. did. This is obviously not always the case but after having talked to many international college students, being able to come to the United States is a privilege that they worked extremely hard for. These students are very unlikely to drop out of college because of the amount of work involved to make it to college.

Overall, this study looked into college retention rates and the factors that could affect college retention rates. It was seen that some of the variables with the most impact on retention rates can't really be controlled such as household income or first generation student percentages. However, some factors could be experimented with to try and increase college retention rate such as paying the professors more or having a lower admission rate.