# Project 2

Load data

```
college <- read.csv('college.csv')
library(naniar)
```

```
## Warning: package 'naniar' was built under R version 3.6.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
##
## 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(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

## 1

```
nrow(college)
```

```
## [1] 1269
```

```
ncol(college)
```

```
## [1] 17
```

The data contains 1269 rows and 17 columns

## 2 Missing data

```
any_na <- sapply(college, function(cvg) sum(is.na(college)))
any_na
```

```
##                 id               name               city              state
##                  0                  0                  0                  0
##             region     highest_degree            control             gender
##                  0                  0                  0                  0
##     admission_rate            sat_avg         undergrads            tuition
##                  0                  0                  0                  0
## faculty_salary_avg  loan_default_rate        median_debt                lon
##                  0                  0                  0                  0
##                lat
##                  0
```

There is no missing data

## 3

```
summary(college)
```

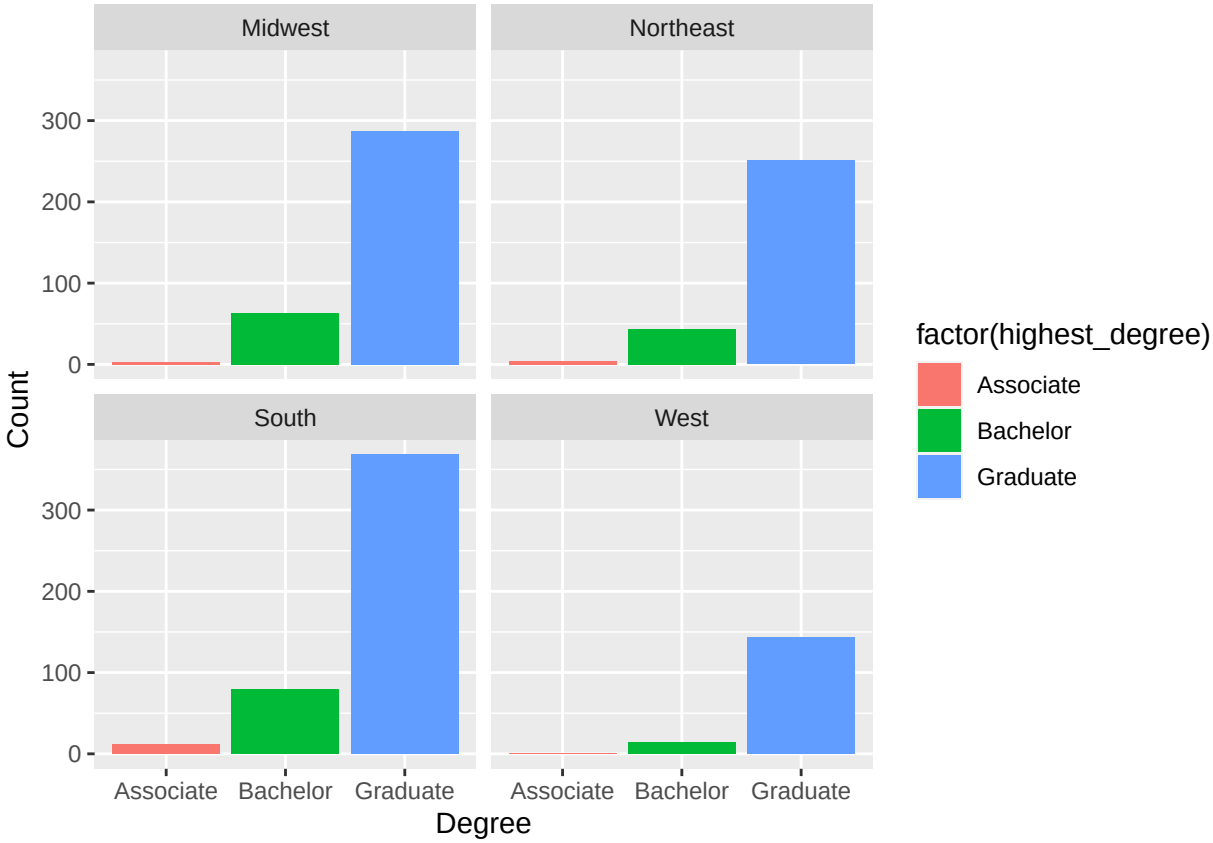
```
##        id                             name                city          state
##  Min.   :100654   Westminster College:   3   New York    :  15   PA     :101
##  1st Qu.:153250   Anderson University:   2   Boston      :  11   NY     : 84
##  Median :186283   Aquinas College    :   2   Chicago     :  10   CA     : 71
##  Mean   :186988   Bethany College    :   2   Philadelphia:   9   TX     : 63
##  3rd Qu.:215284   Bethel University  :   2   Cleveland   :   8   OH     : 52
##  Max.   :484905   Emmanuel College   :   2   Los Angeles :   8   IL     : 47
##                   (Other)            :1256   (Other)     :1208   (Other):851
##        region      highest_degree    control      gender     admission_rate
##  Midwest  :353   Associate:  20   Private:763   CoEd :1237   Min.   :0.0509
##  Northeast:299   Bachelor : 200   Public :506   Men  :   4   1st Qu.:0.5339
##  South    :459   Graduate :1049                 Women:  28   Median :0.6687
##  West     :158                                               Mean   :0.6501
##                                                              3rd Qu.:0.7859
##                                                              Max.   :1.0000
##
##     sat_avg       undergrads       tuition      faculty_salary_avg
##  Min.   : 720   Min.   :   47   Min.   : 2732   Min.   : 1451
##  1st Qu.: 973   1st Qu.: 1296   1st Qu.: 8970   1st Qu.: 6191
##  Median :1040   Median : 2556   Median :20000   Median : 7272
##  Mean   :1060   Mean   : 5629   Mean   :21025   Mean   : 7656
##  3rd Qu.:1120   3rd Qu.: 6715   3rd Qu.:30364   3rd Qu.: 8671
##  Max.   :1545   Max.   :52280   Max.   :51008   Max.   :20650
##
##  loan_default_rate  median_debt         lon               lat
##  0.057  :  32      Min.   : 6056   Min.   :-157.92   Min.   :19.71
##  0.04   :  23      1st Qu.:21250   1st Qu.: -94.17   1st Qu.:35.22
##  0.046  :  22      Median :24589   Median : -84.89   Median :39.74
##  0.027  :  21      Mean   :23483   Mean   : -88.29   Mean   :38.61
```

```
##  0.035  :  19      3rd Qu.:27000   3rd Qu.: -78.63   3rd Qu.:41.81
##  0.038  :  19      Max.   :41000   Max.   : -68.59   Max.   :61.22
##  (Other):1133
```

# 4

- Highest degree in each region

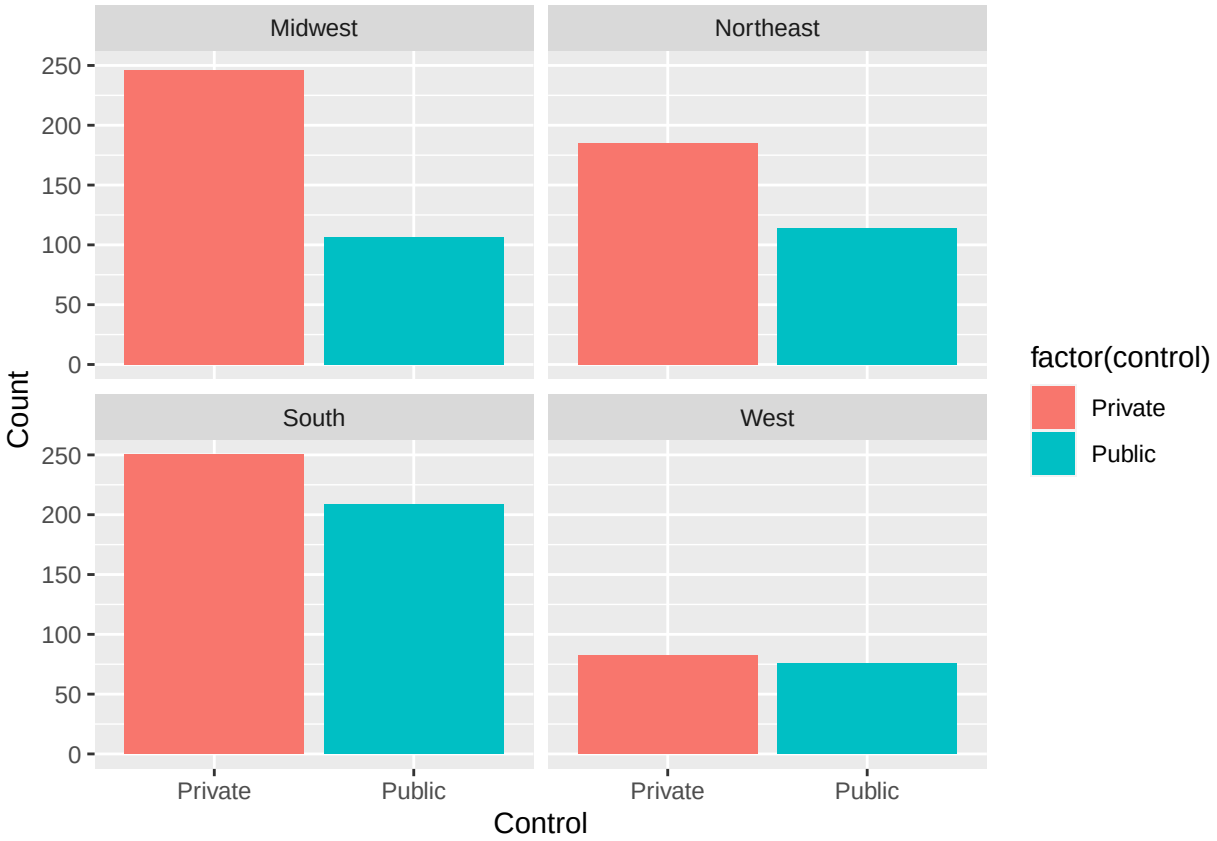
```
ggplot(college, aes(highest_degree, fill = factor(highest_degree))) + geom_bar(position = 'dodge') + fa
```




From the bar plots above, we can see Graduate degree partipates as the highest counts than Associate Degree & Bachelor Degree combine among four regions

- Control in each region

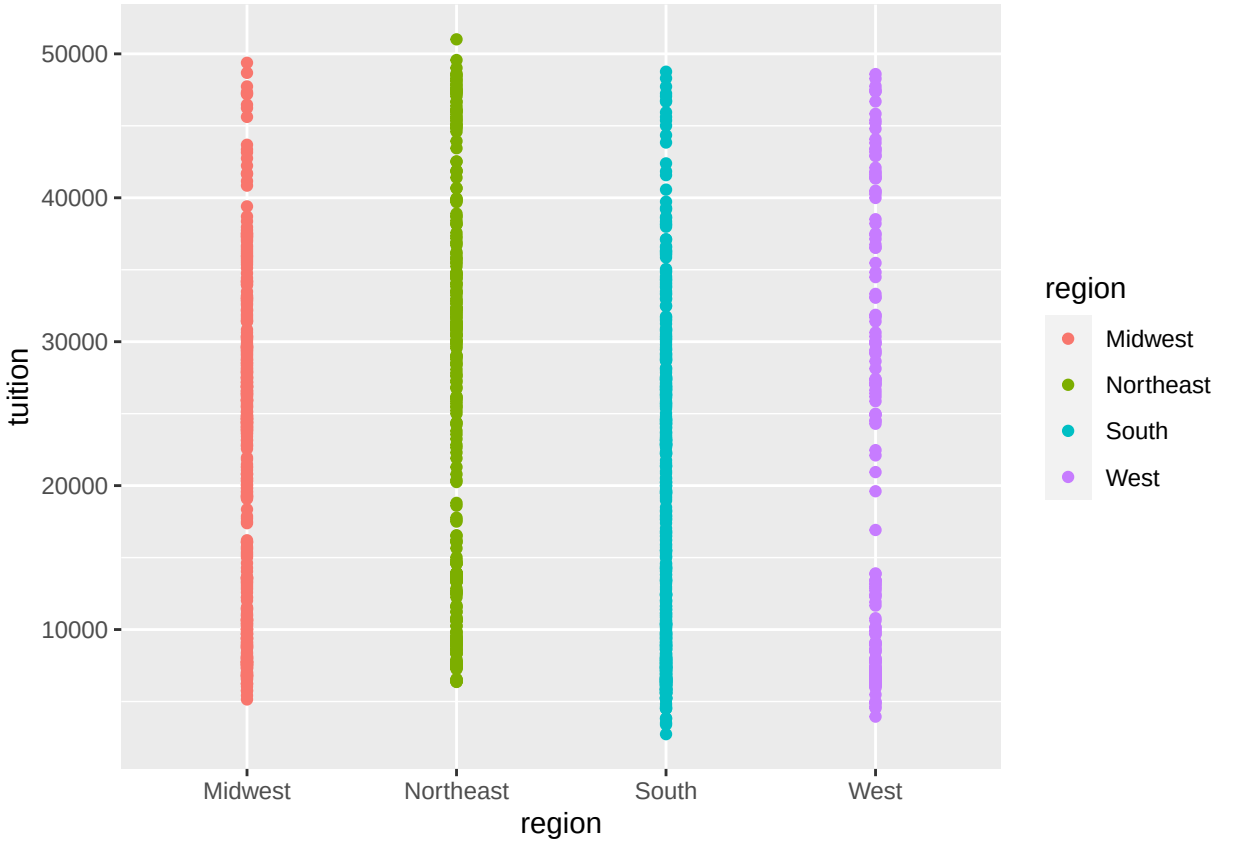
```
ggplot(college, aes(control, fill = factor(control))) + geom_bar(position = 'dodge') + facet_wrap(~regi
```




Private colleges make up a large amount than public ones in Midwest, Northeast, and South. Private colleges is still higher in the West but somewhat is accounted for the same amount of schools as public ones. To conclude, the West area has the lowest number of private schools compared to Midwest, Northeast, and South

- Tuition in each region

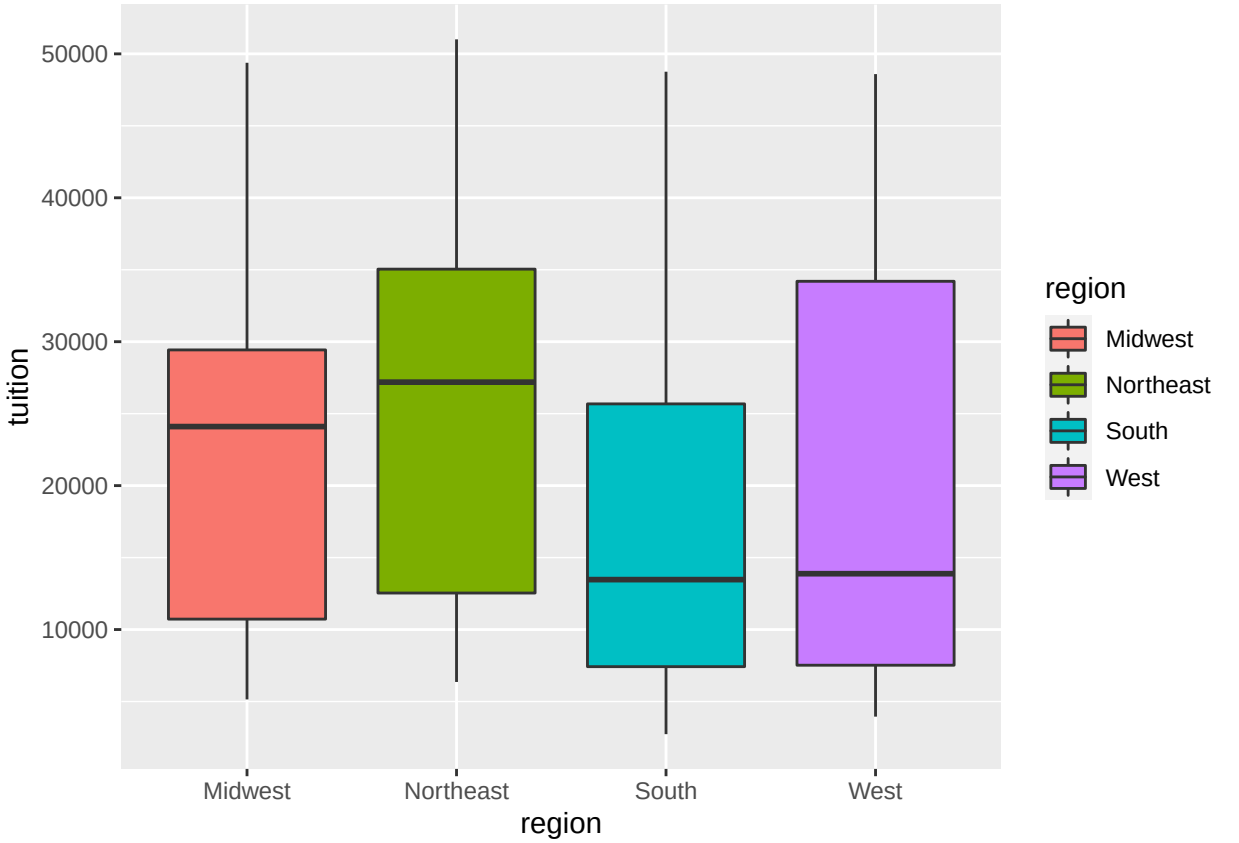
```
ggplot(college, aes(x = region, y = tuition, col = region)) + geom_point()
```




Tuition in each region all ranges from about $5000 to $50000. In the Midwest, majority of tuition is from $5000 - $40000. Tuition fee is more equally scattered within the range in Northest area. In the South, however, the tuition fee is hugely ranged from about $3000 to about $35000. Significantly, tuition in the West is majorly around $4000 to almost $15000. So, tuition fee in the West seems to be the lowest among other three regions

- Median tuition in each college

```
ggplot(college, aes(x = region, y = tuition, fill = region)) + geom_boxplot()
```

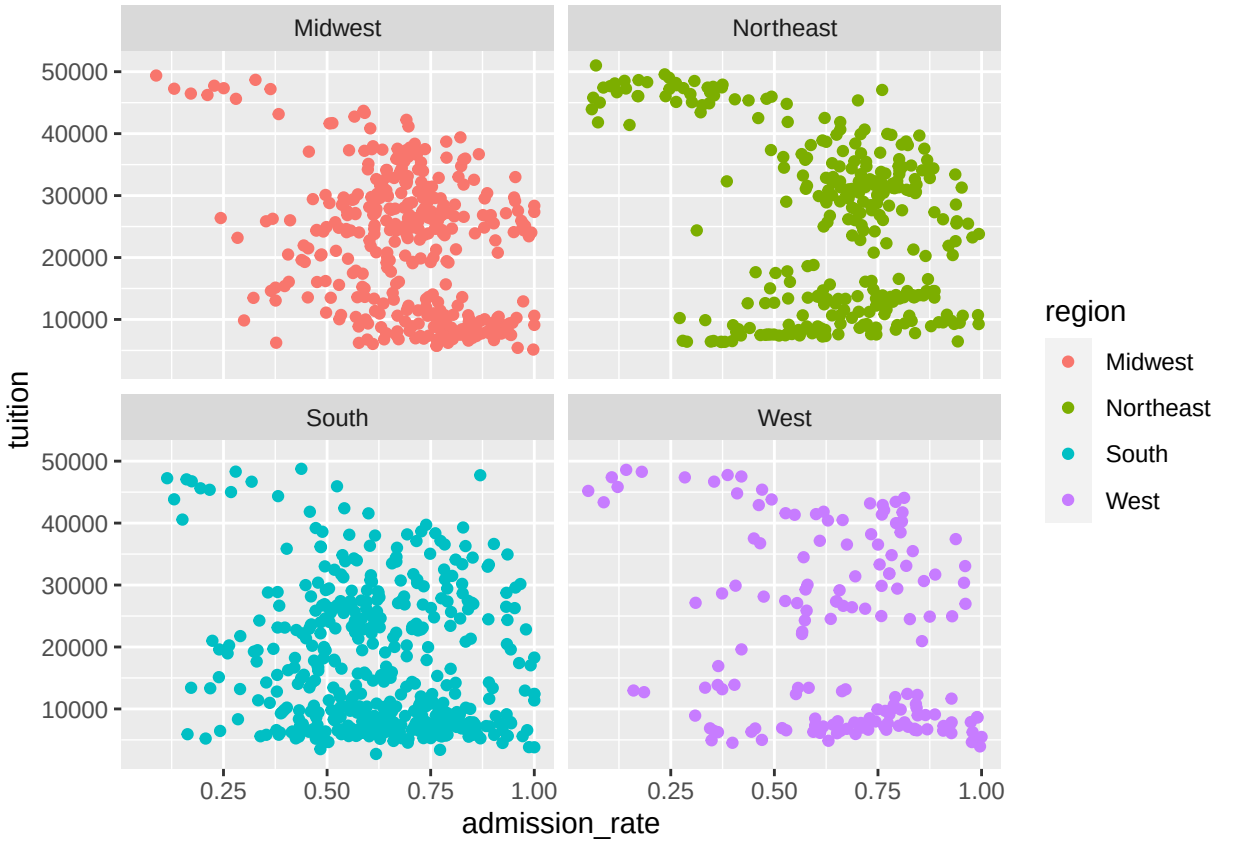



From the scatter plot statistic above, we dwell in the median tuition fee in each region. The Northeast has the highest median of tuition (about $28000) than Midwest, South, and West. The Midwest ranks second place with the median of about $24,500. If we look at the scatter plot above, it seems that the West area has the lowest tuition fee among four regions. However, in this box plot, median values for tuition fee of the South and West are almost similar, which is about $14000. We use median to compare because median is immuned to extreme data points than it is for mean

## 5

- Admission rate vs Tuition in each region

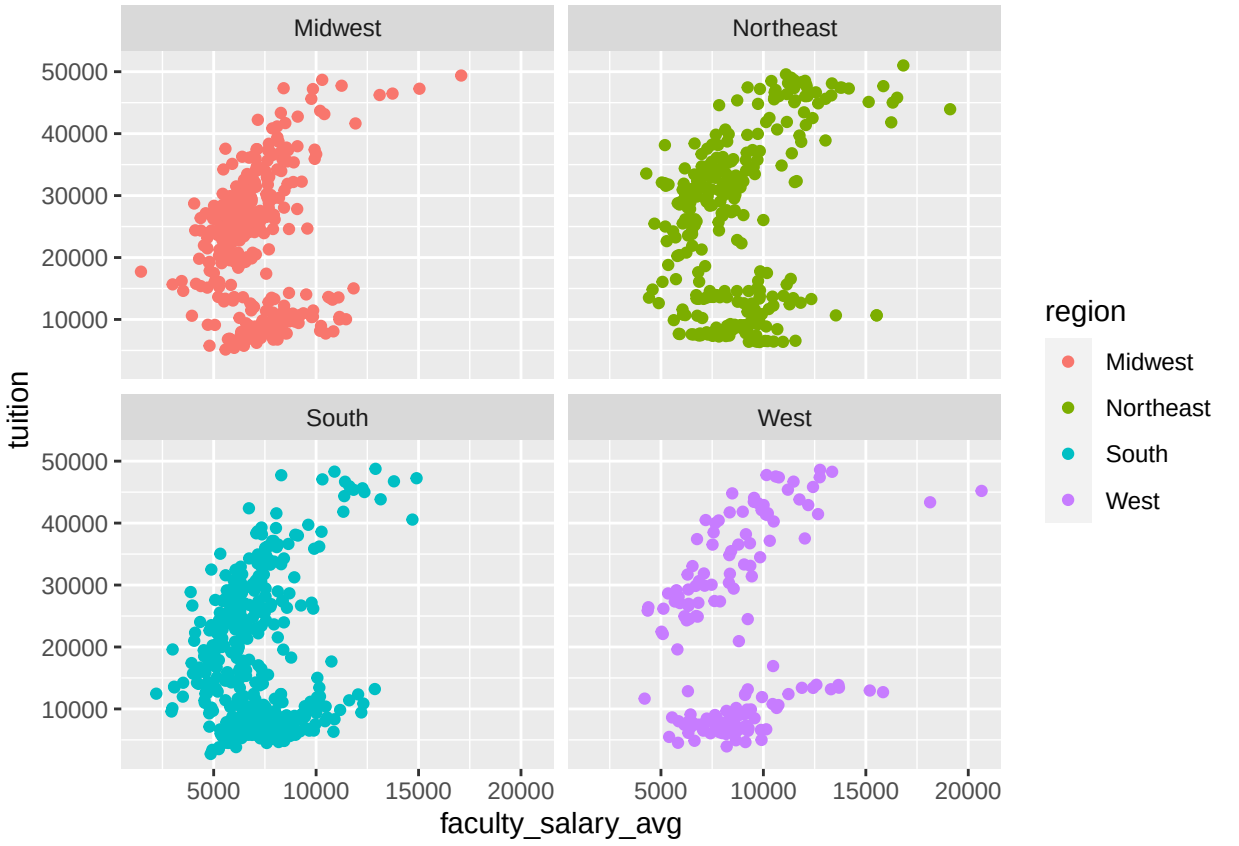
```
ggplot(college, aes(admission_rate, tuition, col = region)) + geom_point() + facet_wrap(~region)
```




Generally, the lower tuition fee, the higher admission rate, especially in Midwest, Northeast, and South. When the tuition is below $40000, there is 50% to 100% acceptance. In the West, admission_rate 0.75 to 1 when tuition fee is below $15000

- Faculty avg salary vs tuition in each region

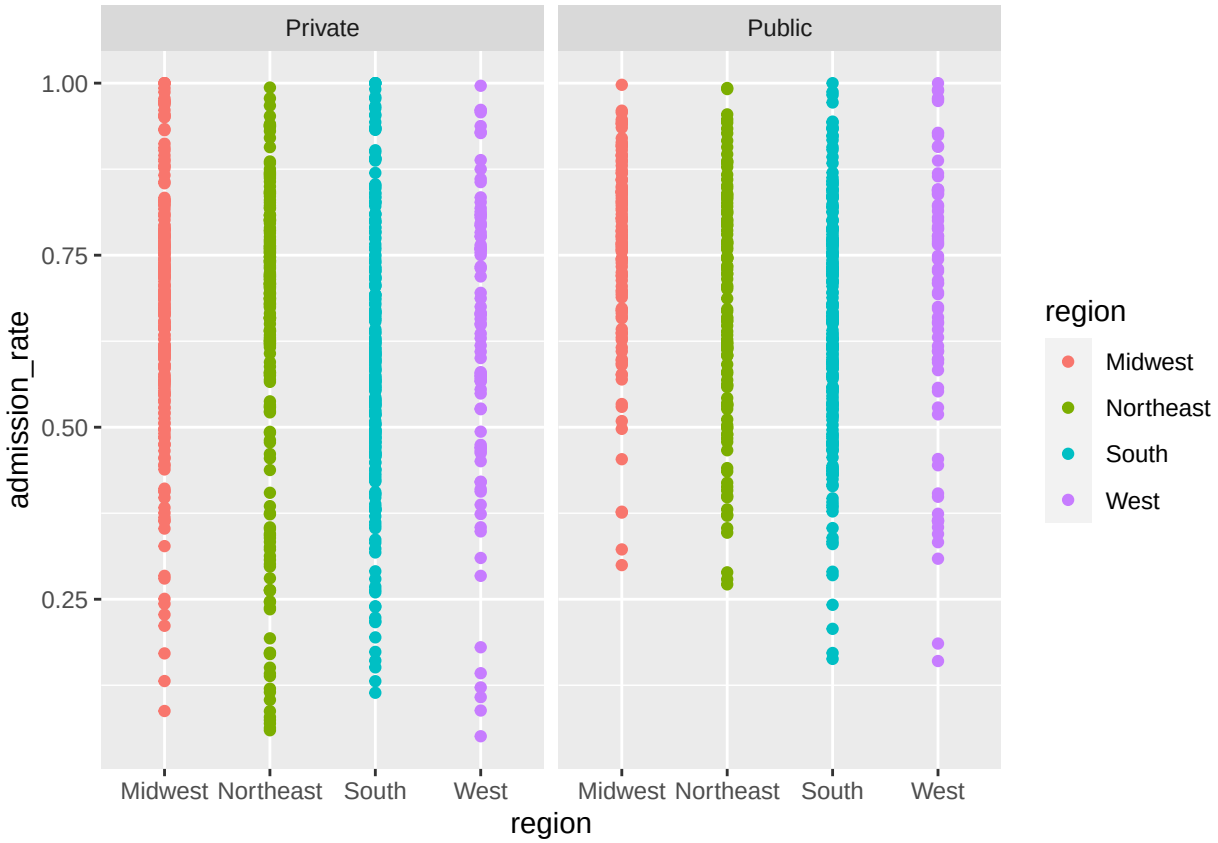
```
ggplot(college, aes(faculty_salary_avg, tuition, col = region)) + geom_point() + facet_wrap(~region)
```




As the tuition below $40000, average faculty salary is ranged from $5000 to $10000. In some rare case, as tuition fee is above $40000, average salary of faculty reaches $15000

- Control has high admission rate in which region

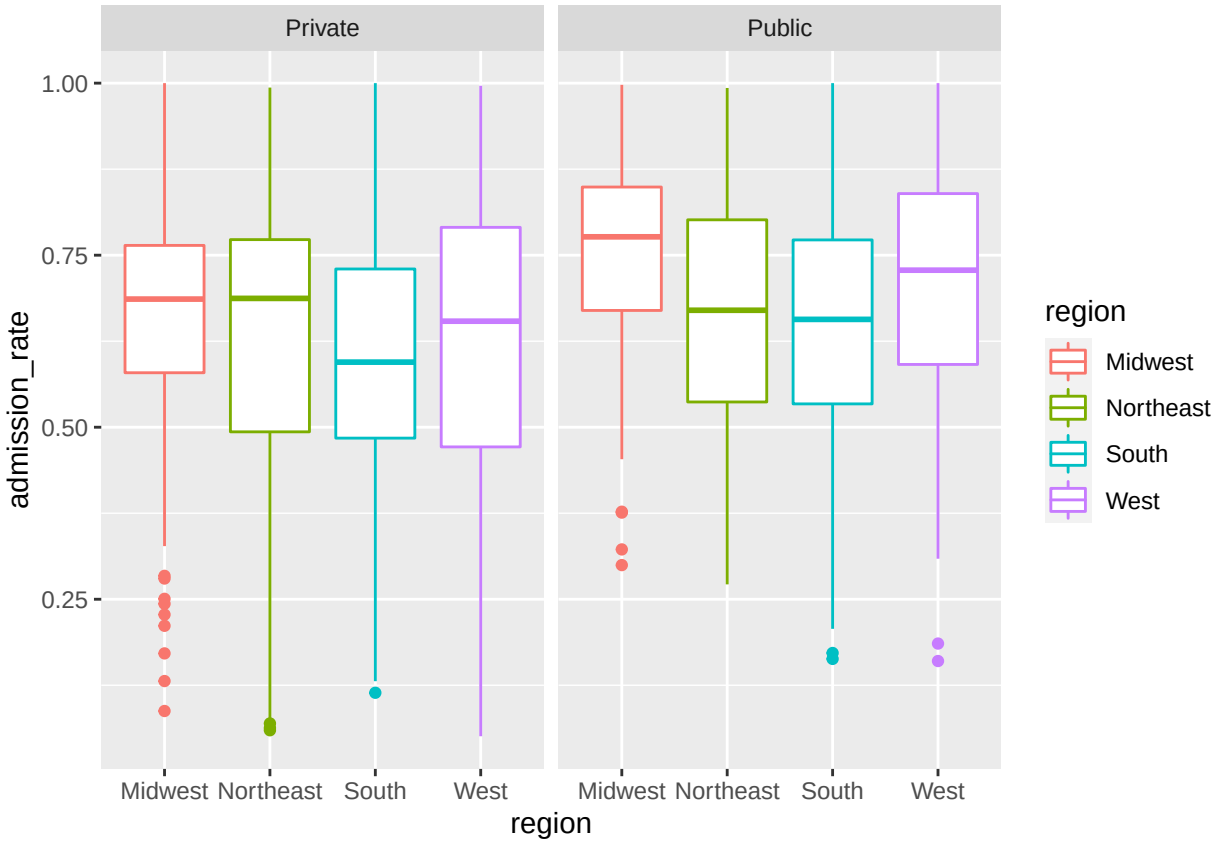
```
ggplot(college, aes(x = region, y = admission_rate, col = region)) + geom_point() + facet_wrap(~control)
```




Generally, majority of public school has the admission rate greater than 0.25 for all region than it does for private school. Significantly, both Private and Public schools in the South notably has the admission rate above 0.25

- Median admission_rate among region

```
ggplot(college, aes(x = region, y = admission_rate, col = region)) + geom_boxplot() + facet_wrap(~contr
```




The median of admission rate of public schools are generally higher than it is for private schools, accounted for all of the region. For Public school system, Midwest area comparts the highest median of admission rate (above 0.75) than other three regions, the West stands in second, and both Northeast and South are the lowest. In Private school system,Midwest and Northest have the highest median admission_rate, the West stands in second and the South accounts for the lowest median admission_rate. Private school in Midwest has many outliers for admission_rate(below 0.25) than for Public school (outliers above 0.25). Private and Public school in the South both have lowest median admission rate than other regions

## Proposed questions

1. Are there any correlations among average faculty salary, tuition, and students median debt

```
library(PerformanceAnalytics)
```

```
## Warning: package 'PerformanceAnalytics' was built under R version 3.6.3
```

```
## Loading required package: xts
```

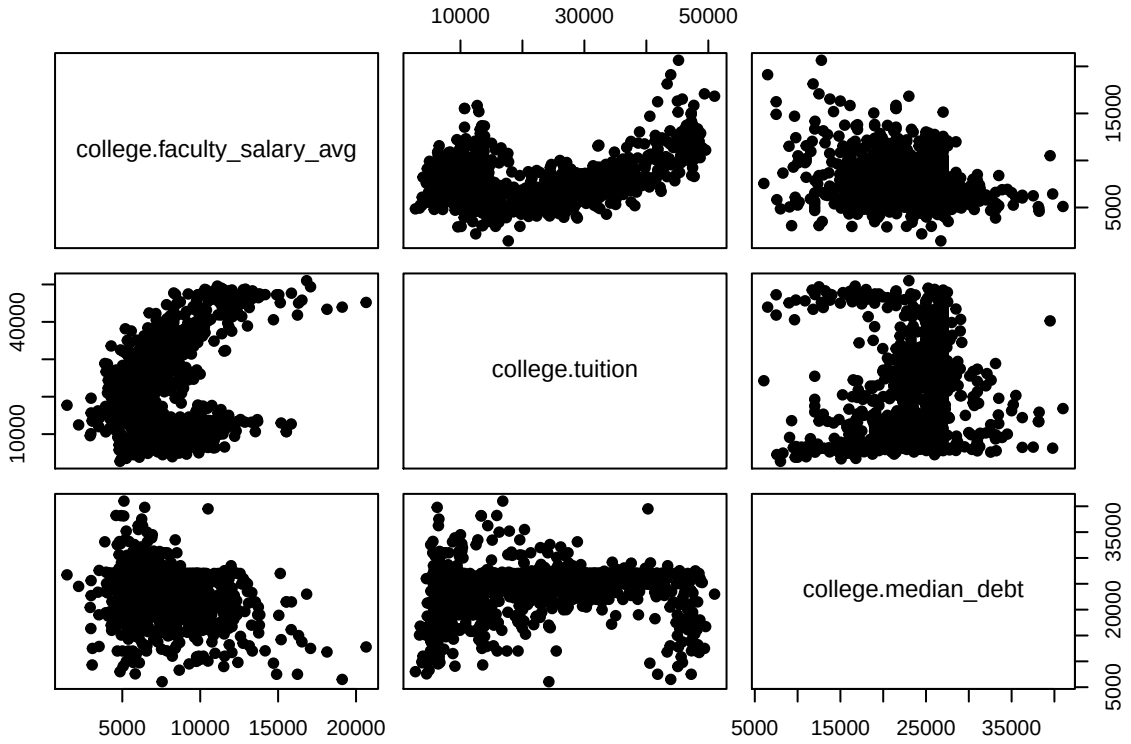
```
## Warning: package 'xts' was built under R version 3.6.3
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 3.6.3
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':
##
##     first, last

##
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':
##
##     legend
```

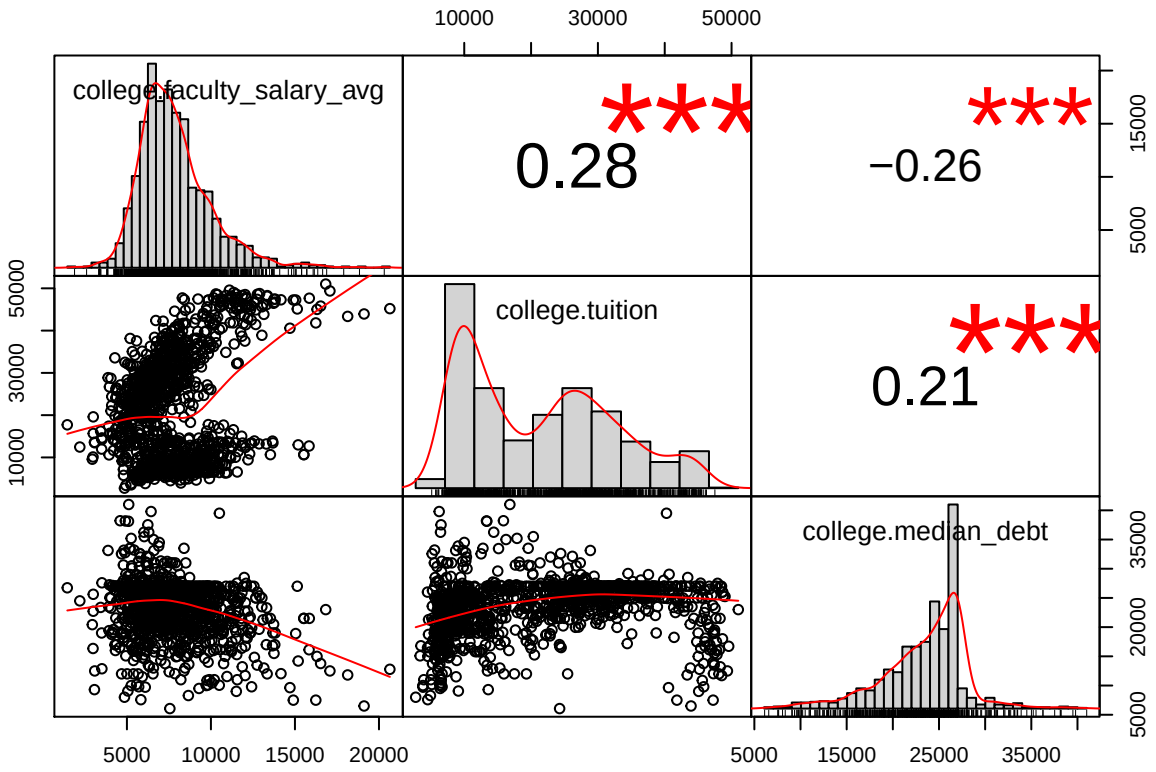
```r
salary_tuition <- data.frame(college$faculty_salary_avg, college$tuition, college$median_debt)
pairs(salary_tuition, pch = 19)
```



```
round(cor(salary_tuition), digits = 3)
```

```
##                              college.faculty_salary_avg college.tuition
## college.faculty_salary_avg                        1.000           0.281
## college.tuition                                   0.281           1.000
## college.median_debt                              -0.256           0.205
##                              college.median_debt
## college.faculty_salary_avg              -0.256
## college.tuition                          0.205
## college.median_debt                      1.000
```

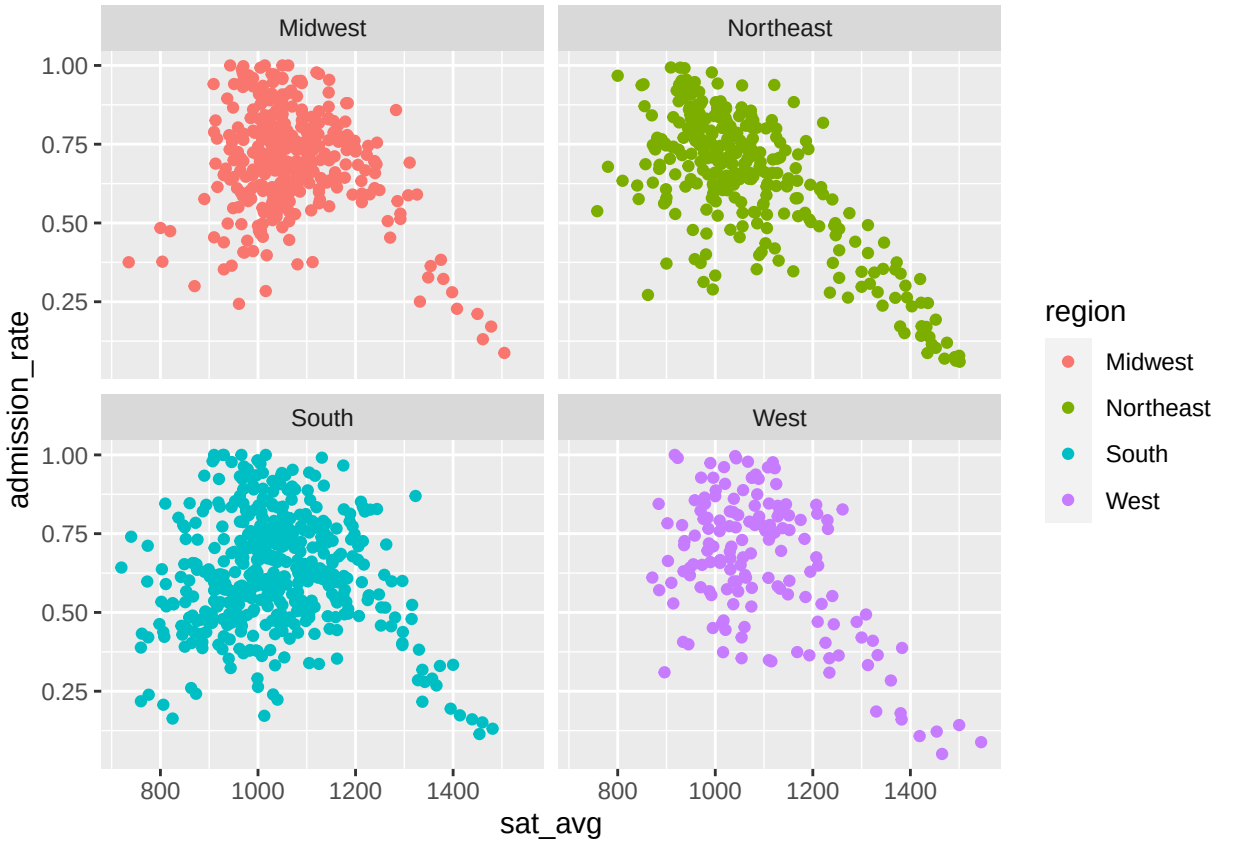
```
chart.Correlation(R = salary_tuition, histogram = TRUE, pch = 19)
```




There are not any strong correlations among the three chosen variables.

2. SAT average and admission rate among regions and if there is correlation

```
ggplot(college, aes(x = sat_avg, y = admission_rate, col = region)) + geom_point() + facet_wrap(~region
```

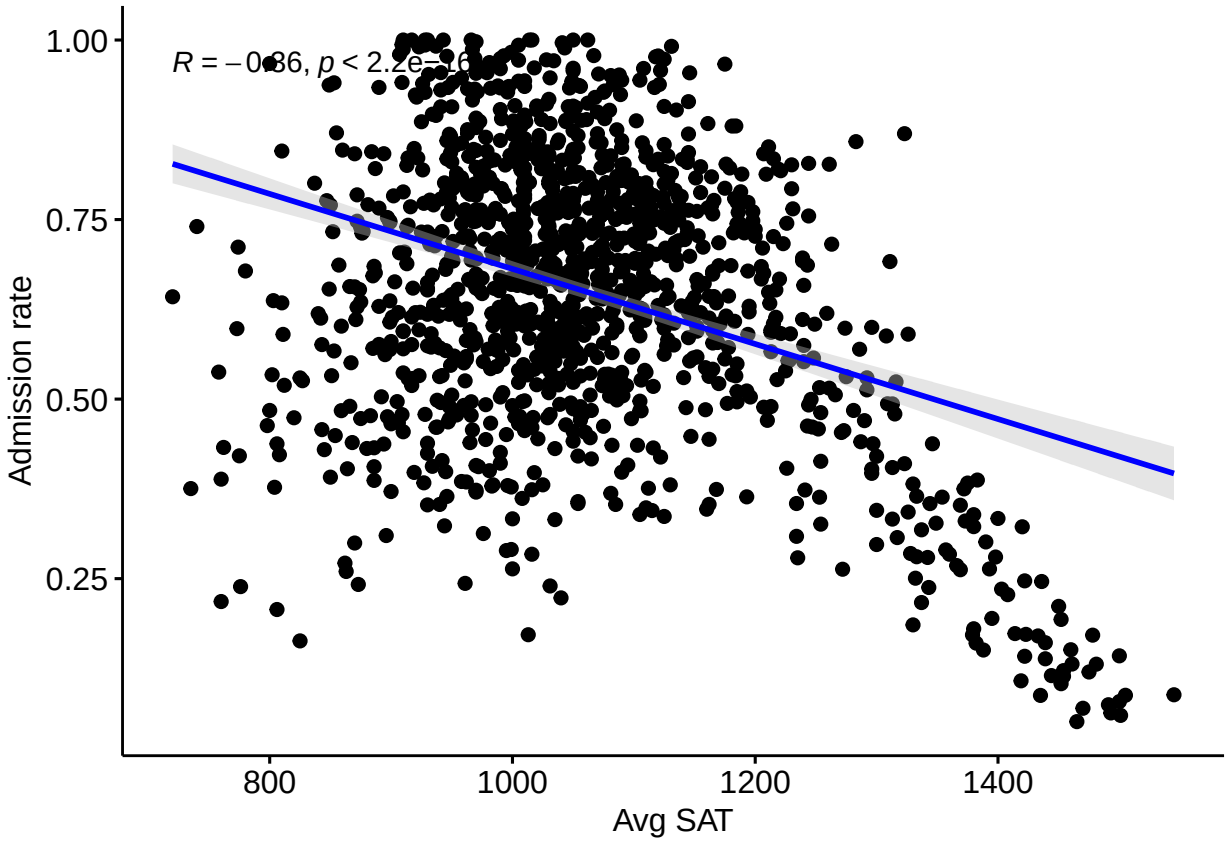



```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 3.6.3
```

```
ggscatter(college, x = "sat_avg", y = "admission_rate",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson", add.params = list(color = "blue", fill = "gray"),
          xlab = "Avg SAT", ylab = "Admission rate")
```

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




Overall, average SAT from 1000 to 1200 has higher admission_rate then average SAT above 1200, accounted for 4 regions

There is week correlation (R = -0.36) between average SAT and admission rate

3. Ivy League faculty average salary and student median debt in Northeast

```
library(dplyr)
college %>% filter(region == 'Northeast')%>% filter(name %in% c("Yale University", "Harvard University"
```

```
##                         name faculty_salary_avg median_debt
## 1            Yale University              16529       13774
## 2         Harvard University              19115        6500
## 3       Princeton University              16242        7500
## 4 University of Pennsylvania              15855       21500
```

Among four chosen Ivy League school in Northeast area, Havard has the higest faculty salary average (19115) and lowest student median debt(6500). Mean while, University of Pennsylvania has the lowest faculty salary average (15855) and highest student median debt (21500)

4. Top colleges faculty average salary and student median debt in West

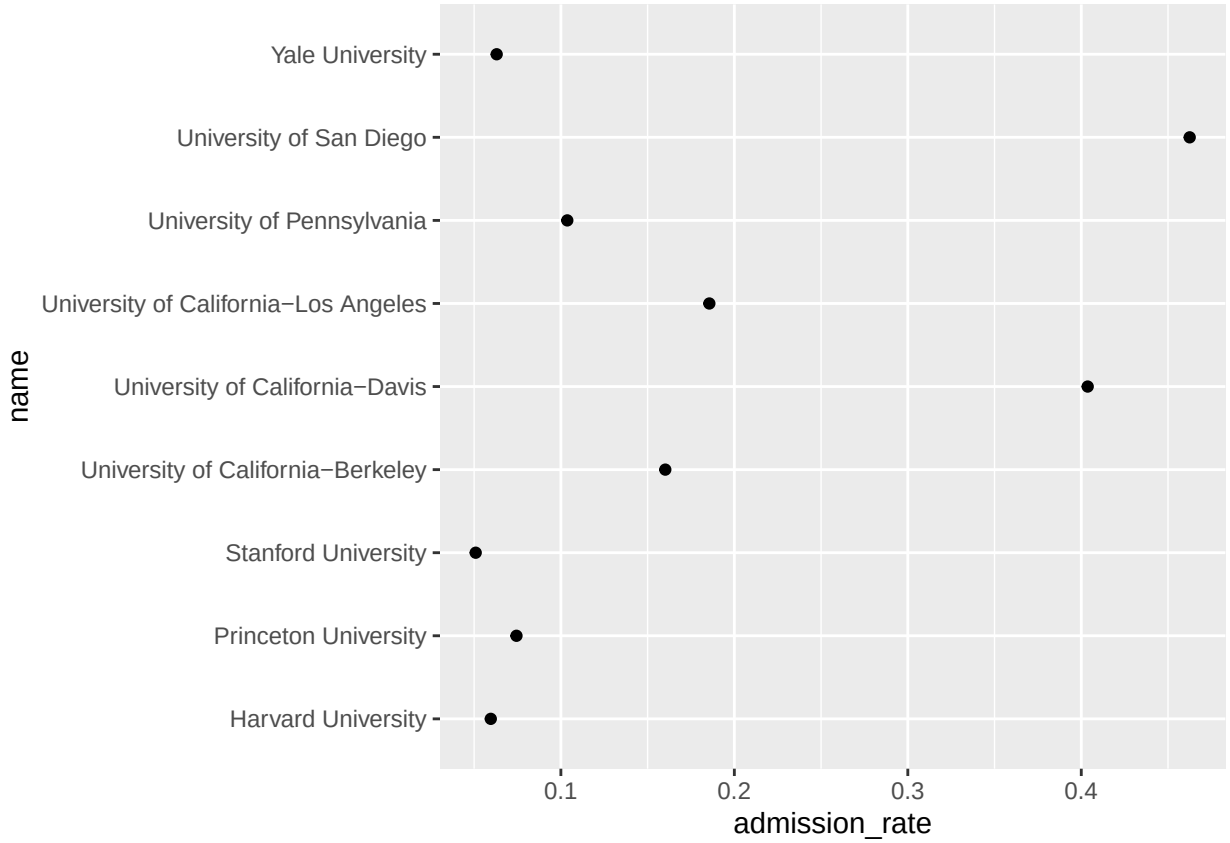
```
college %>% filter(region == 'West')%>% filter(name %in%  c("Stanford University", "University of Calif
```

```
##                                 name faculty_salary_avg median_debt
## 1                  Stanford University              20650       12782
## 2    University of California-Berkeley              15194       14200
## 3       University of California-Davis              12587       14833
## 4 University of California-Los Angeles              15841       16126
## 5              University of San Diego              12190       22750
```

Among five chosen top colleges West area, Stanford has the higest faculty salary average (20650) and lowest student median debt(12782). Mean while, University of San Diego has the lowest faculty salary average (12190) and highest student median debt (22750)

5. Admission rate among top colleges

```
n <- college %>% filter(region == 'Northeast')%>% filter(name %in% c("Yale University", "Harvard Univers
w <- college %>% filter(region == 'West')%>% filter(name %in%  c("Stanford University", "University of C
E <- rbind(n,w)
ggplot(E, aes(x = admission_rate, y = name)) + geom_point()
```




Majority of top schools have admission rate below 0.2 .Interestingly, University of California-Davis has admission rate about 0.4