# An analysis of University Admission Data

Rohan Taneja

2024-07-21

## Introduction

This project seeks to conduct Exploratory Data Analysis on the topic of University Admission Data.

The original data set has been sourced from the following link - https://www.kaggle.com/datasets/tarekmuhammed/university-students-data

## Aim of this experiment

We aim to answer some questions below with the help of this data. They are as follows:

1. How do the number of applications, acceptances, and enrollments vary across institutions? (I think public schools get more applications)
2. What is the acceptance rate and enrollment rate for each institution, and how do these rates vary between private and public institutions? ( I think public schools are more selective)
3. What is the relationship between the percentage of students in the top 10% and top 25% of their high school class and the acceptance rate? ( I think as more students who are in the prime of their class apply to a school , the acceptance rates should increase)
4. How does the student-to-faculty ratio impact the graduation rate? (One would think that more faculty = more student success)
5. Do private school alumni end up donating more to their alma matter than public school alumni? (Yeah they do)
6. How does the expenditure per student influence the graduation rate, and does this relationship vary between public and private institutions? ( Yes the more money you spend on a student, the more likely they are to graduate - we will use Linear Regression to answer how much it affect the grad rate.)

## Structure of the data in this data set

### Explanation of the Data

The data-set used in this analysis contains information about various aspects of colleges and universities. Below is a detailed description of each variable included in the dataset:

1. **Apps**: Number of applications received from prospective students.
2. **Accept**: The number of applications accepted.
3. **Enroll**: The number of new students enrolled.

4. **Top10perc**: The percentage of new students from the top 10% of their high school class.
5. **Top25perc**: The percentage of new students from the top 25% of their high school class.
6. **F.Undergrad**: The number of full-time undergraduates.
7. **P.Undergrad**: The number of part-time undergraduates.
8. **Outstate**: The out-of-state tuition fee.
9. **Room.Board**: The costs for room and board.
10. **Books**: The estimated costs for books.
11. **Personal**: The estimated personal spending.
12. **PhD**: The percentage of faculty with Ph.D. degrees.
13. **Terminal**: The percentage of faculty with terminal degrees.
14. **S.F.Ratio**: The student-to-faculty ratio.
15. **perc.alumni**: The percentage of alumni who donate.
16. **Expend**: The instructional expenditure per student.
17. **Grad.Rate**: The graduation rate.

In addition to the variables listed above, we have added three more columns to the data-set:

18. **AcceptanceRate**: The acceptance rate, calculated as the number of applications accepted divided by the number of applications received.
19. **EnrollmentRate**: The enrollment rate, calculated as the number of new students enrolled divided by the number of applications accepted.
20. **Type_of_institution**: A categorical variable indicating the type of institution (e.g., public, private).

## Loading libraries

```
# Load necessary libraries
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
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
```

```r
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.3.3
```

```r
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.3.3
```

```
##
## Attaching package: 'gridExtra'
```

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

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

```
## corrplot 0.92 loaded
```

```r
library(maps)
```

```
## Warning: package 'maps' was built under R version 4.3.3
```

```r
library(ggmap)
```

```
## Warning: package 'ggmap' was built under R version 4.3.3
```

```
## i Google's Terms of Service: <https://mapsplatform.google.com>
##   Stadia Maps' Terms of Service: <https://stadiamaps.com/terms-of-service/>
##   OpenStreetMap's Tile Usage Policy: <https://operations.osmfoundation.org/policies/tiles/>
## i Please cite ggmap if you use it! Use `citation("ggmap")` for details.
```

```r
library(zipcodeR)
```

```
## Warning: package 'zipcodeR' was built under R version 4.3.3
```

```r
library(dplyr)
library(purrr)
```

```
##
## Attaching package: 'purrr'
```

```
## The following object is masked from 'package:maps':
##
##     map
```

```r
library(zipcodeR)
library(ggplot2)
library(maps)
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

## Importing the dataset

```r
data <- read.csv('Kmeans_assignment_data.csv')
```

```r
# View the structure of the data-set
str(data)
```

```
## 'data.frame':    777 obs. of  19 variables:
##  $ X          : chr  "Abilene Christian University" "Adelphi University" "Adrian College" "Agnes Sco
##  $ Private    : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Apps       : int  1660 2186 1428 417 193 587 353 1899 1038 582 ...
##  $ Accept     : int  1232 1924 1097 349 146 479 340 1720 839 498 ...
##  $ Enroll     : int  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : int  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : int  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: int  2885 2683 1036 510 249 678 416 1594 973 799 ...
##  $ P.Undergrad: int  537 1227 99 63 869 41 230 32 306 78 ...
##  $ Outstate   : int  7440 12280 11250 12960 7560 13500 13290 13868 15595 10468 ...
##  $ Room.Board : int  3300 6450 3750 5450 4120 3335 5720 4826 4400 3380 ...
##  $ Books      : int  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : int  2200 1500 1165 875 1500 675 1500 850 500 1800 ...
##  $ PhD        : int  70 29 53 92 76 67 90 89 79 40 ...
##  $ Terminal   : int  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: int  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : int  7041 10527 8735 19016 10922 9727 8861 11487 11644 8991 ...
##  $ Grad.Rate  : int  60 56 54 59 15 55 63 73 80 52 ...
```

```r
summary(data)
```

```
##       X                Private              Apps           Accept
##  Length:777         Length:777         Min.   :   81   Min.   :   72
##  Class :character   Class :character   1st Qu.:  776   1st Qu.:  604
##  Mode  :character   Mode  :character   Median : 1558   Median : 1110
##                                        Mean   : 3002   Mean   : 2019
##                                        3rd Qu.: 3624   3rd Qu.: 2424
##                                        Max.   :48094   Max.   :26330
```

```
##      Enroll       Top10perc       Top25perc      F.Undergrad
##  Min.   :  35   Min.   : 1.00   Min.   :  9.0   Min.   :  139
##  1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992
##  Median : 434   Median :23.00   Median : 54.0   Median : 1707
##  Mean   : 780   Mean   :27.56   Mean   : 55.8   Mean   : 3700
##  3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005
##  Max.   :6392   Max.   :96.00   Max.   :100.0   Max.   :31643
##   P.Undergrad         Outstate       Room.Board       Books
##  Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0
##  1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0
##  Median :  353.0   Median : 9990   Median :4200   Median : 500.0
##  Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4
##  3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0
##  Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0
##     Personal         PhD            Terminal       S.F.Ratio
##  Min.   : 250   Min.   :  8.00   Min.   : 24.0   Min.   : 2.50
##  1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50
##  Median :1200   Median : 75.00   Median : 82.0   Median :13.60
##  Mean   :1341   Mean   : 72.66   Mean   : 79.7   Mean   :14.09
##  3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50
##  Max.   :6800   Max.   :103.00   Max.   :100.0   Max.   :39.80
##   perc.alumni        Expend        Grad.Rate
##  Min.   : 0.00   Min.   : 3186   Min.   : 10.00
##  1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
##  Median :21.00   Median : 8377   Median : 65.00
##  Mean   :22.74   Mean   : 9660   Mean   : 65.46
##  3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :64.00   Max.   :56233   Max.   :118.00
```

### Checking for NA values:

```
colSums(is.na(data))
```

```
##           X     Private        Apps      Accept      Enroll   Top10perc
##           0           0           0           0           0           0
##   Top25perc F.Undergrad P.Undergrad    Outstate  Room.Board       Books
##           0           0           0           0           0           0
##    Personal         PhD    Terminal   S.F.Ratio perc.alumni      Expend
##           0           0           0           0           0           0
##   Grad.Rate
##           0
```

No NA values :D

### Data Wrangling and creating some new variables:

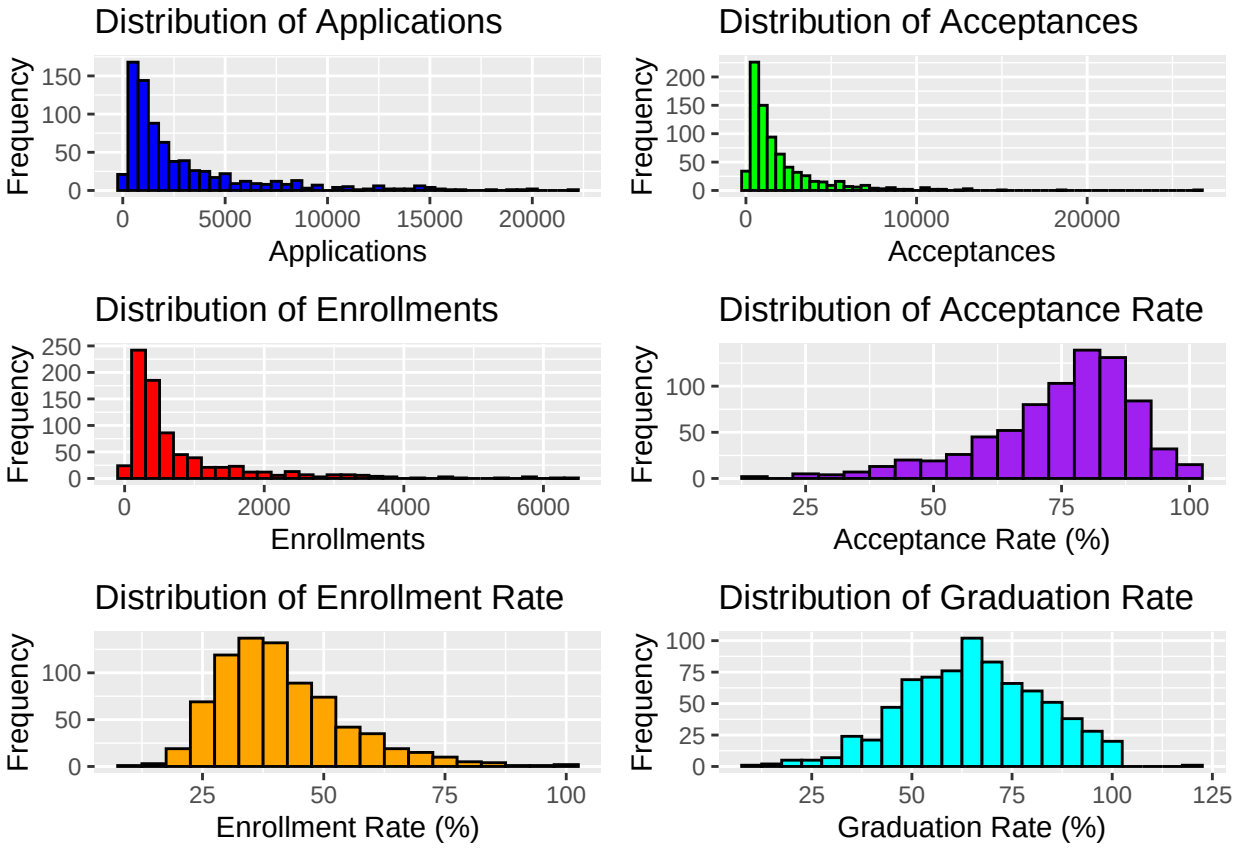
```
# Create new columns for acceptance rate and enrollment rate
data <- data %>%
  mutate(AcceptanceRate = Accept / Apps * 100,
```

```
        EnrollmentRate = Enroll / Accept * 100)
# Creating a new column for type of institution.
data <- data %>%
  mutate(Type_of_institution = ifelse(Private == "Yes", "Private", "Public"))

#One university has 48k applications which is way more than normal so we are excluding it
filtered_data <- data %>% filter(Apps <= 40000)
```

## Visual EDA

A brief look at the data distribution of the following variables : Applications, Accepted Applications, Enrolled , Acceptance Rate, Enrollment Rate and Grad Rate

```
p1 <- ggplot(filtered_data, aes(x = Apps)) +
  geom_histogram(binwidth = 500, fill = "blue", color = "black") +
  labs(title = "Distribution of Applications", x = "Applications", y = "Frequency")

p2 <- ggplot(data, aes(x = Accept)) +
  geom_histogram(binwidth = 500, fill = "green", color = "black") +
  labs(title = "Distribution of Acceptances", x = "Acceptances", y = "Frequency")

p3 <- ggplot(data, aes(x = Enroll)) +
  geom_histogram(binwidth = 200, fill = "red", color = "black") +
  labs(title = "Distribution of Enrollments", x = "Enrollments", y = "Frequency")

p4 <- ggplot(data, aes(x = AcceptanceRate)) +
  geom_histogram(binwidth = 5, fill = "purple", color = "black") +
  labs(title = "Distribution of Acceptance Rate", x = "Acceptance Rate (%)", y = "Frequency")

p5 <- ggplot(data, aes(x = EnrollmentRate)) +
  geom_histogram(binwidth = 5, fill = "orange", color = "black") +
  labs(title = "Distribution of Enrollment Rate", x = "Enrollment Rate (%)", y = "Frequency")

p6 <- ggplot(data, aes(x = Grad.Rate)) +
  geom_histogram(binwidth = 5, fill = "cyan", color = "black") +
  labs(title = "Distribution of Graduation Rate", x = "Graduation Rate (%)", y = "Frequency")
```




### A brief map of these universities

Now we will attempt to create a map of universities in the US using the zipcode. I obtained a scraper from the internet that would allow me to import postcode data from the US News Rankings API.

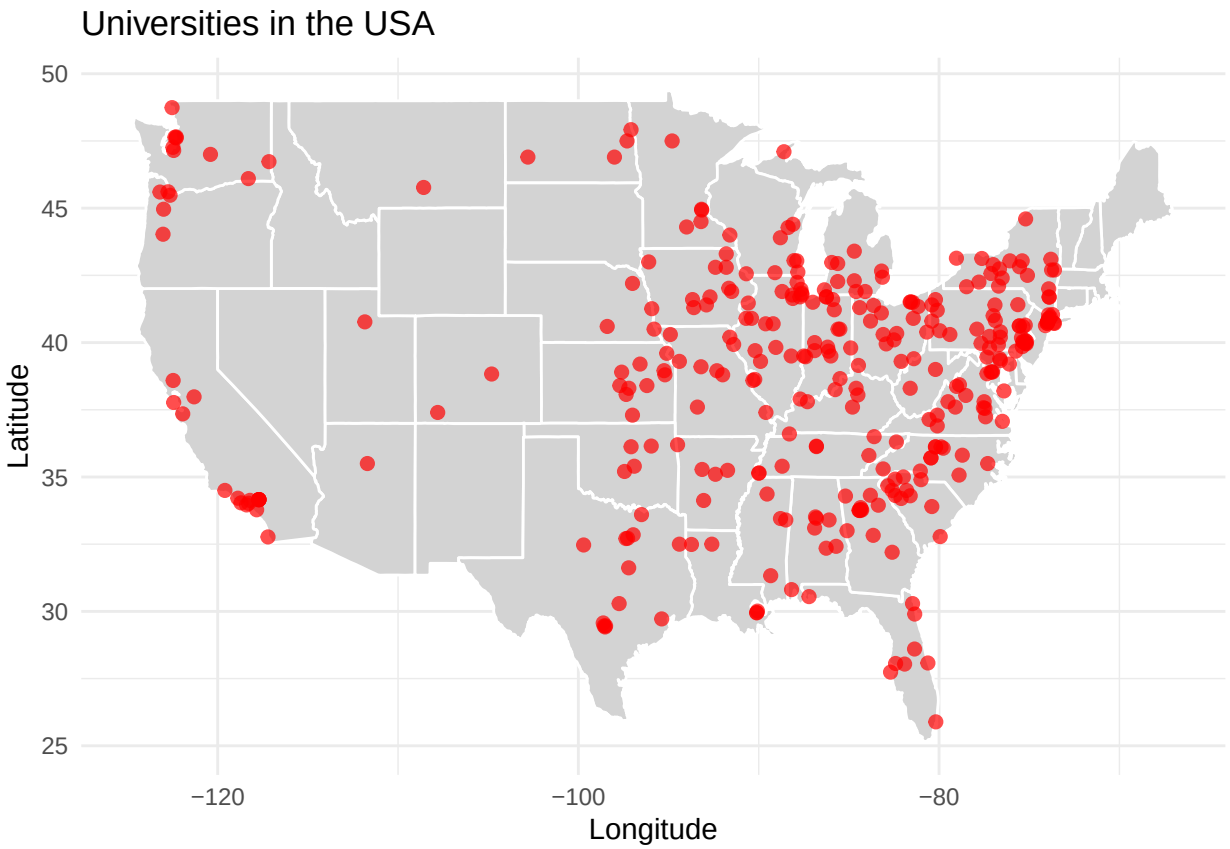
A link to it is here - https://github.com/kajchang/USNews-College-Scraper/tree/master

```
# Importing
postcode <- read.csv("usn/USNews-College-Scraper/data.csv")

# Select relevant columns from postcode data
postcodes1 <- postcode %>%
  select(institution.displayName, institution.zip)

# Merge data with postcode data
merged_data <- merge(data, postcodes1, by.x = "X", by.y = "institution.displayName", all.x = TRUE)

filtered_data <- merged_data %>%
  filter(!is.na(institution.zip))

# Define a function to check if a ZIP code is valid using tryCatch
is_valid_zip <- function(zip) {
  tryCatch({
    !is.null(zipcodeR::geocode_zip(zip))
  }, error = function(e) {
```

```
    FALSE
  })
}

# Filter out invalid ZIP codes
valid_data <- filtered_data %>%
  filter(sapply(institution.zip, is_valid_zip))

# Geocode the valid zip codes using zipcodeR
geocoded_data <- valid_data %>%
  mutate(geocode = map(institution.zip, ~zipcodeR::geocode_zip(.x))) %>%
  unnest(cols = c(geocode))

# Filter out invalid coordinates (latitude and longitude ranges for the contiguous USA)
geocoded_data <- geocoded_data %>%
  filter(lat >= 24 & lat <= 49, lng >= -125 & lng <= -66)

# View the geocoded data
head(geocoded_data)
```

```
## # A tibble: 6 x 26
##   X      Private  Apps Accept Enroll Top10perc Top25perc F.Undergrad P.Undergrad
##   <chr>  <chr>   <int>  <int>  <int>     <int>     <int>       <int>       <int>
## 1 Abile~ Yes      1660   1232    721        23        52        2885         537
## 2 Adelp~ Yes      2186   1924    512        16        29        2683        1227
## 3 Adria~ Yes      1428   1097    336        22        50        1036          99
## 4 Agnes~ Yes       417    349    137        60        89         510          63
## 5 Albio~ Yes      1899   1720    489        37        68        1594          32
## 6 Alfre~ Yes      1732   1425    472        37        75        1830         110
## # i 17 more variables: Outstate <int>, Room.Board <int>, Books <int>,
## #   Personal <int>, PhD <int>, Terminal <int>, S.F.Ratio <dbl>,
## #   perc.alumni <int>, Expend <int>, Grad.Rate <int>, AcceptanceRate <dbl>,
## #   EnrollmentRate <dbl>, Type_of_institution <chr>, institution.zip <int>,
## #   zipcode <chr>, lat <dbl>, lng <dbl>
```

```
# Get the map of the USA
usa_map <- map_data("state")

# Plot the universities on the map
ggplot() +
  geom_polygon(data = usa_map, aes(x = long, y = lat, group = group), fill = "lightgrey", color = "white
  geom_point(data = geocoded_data, aes(x = lng, y = lat), color = "red", size = 2, alpha = 0.7) +
  labs(title = "Universities in the USA", x = "Longitude", y = "Latitude") +
  theme_minimal()
```

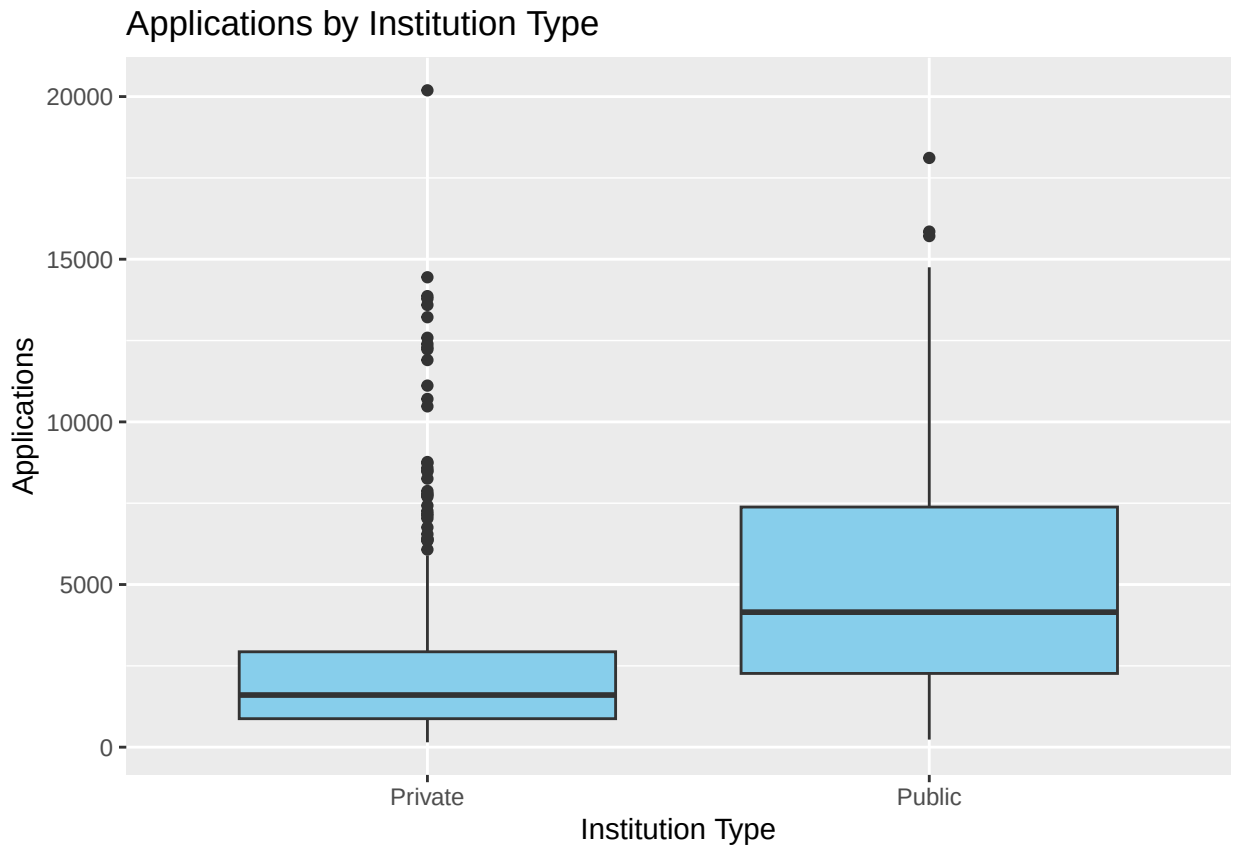



Now we will look to answer our first question:

1. How do the number of applications, acceptances, and enrollments vary across institutions?

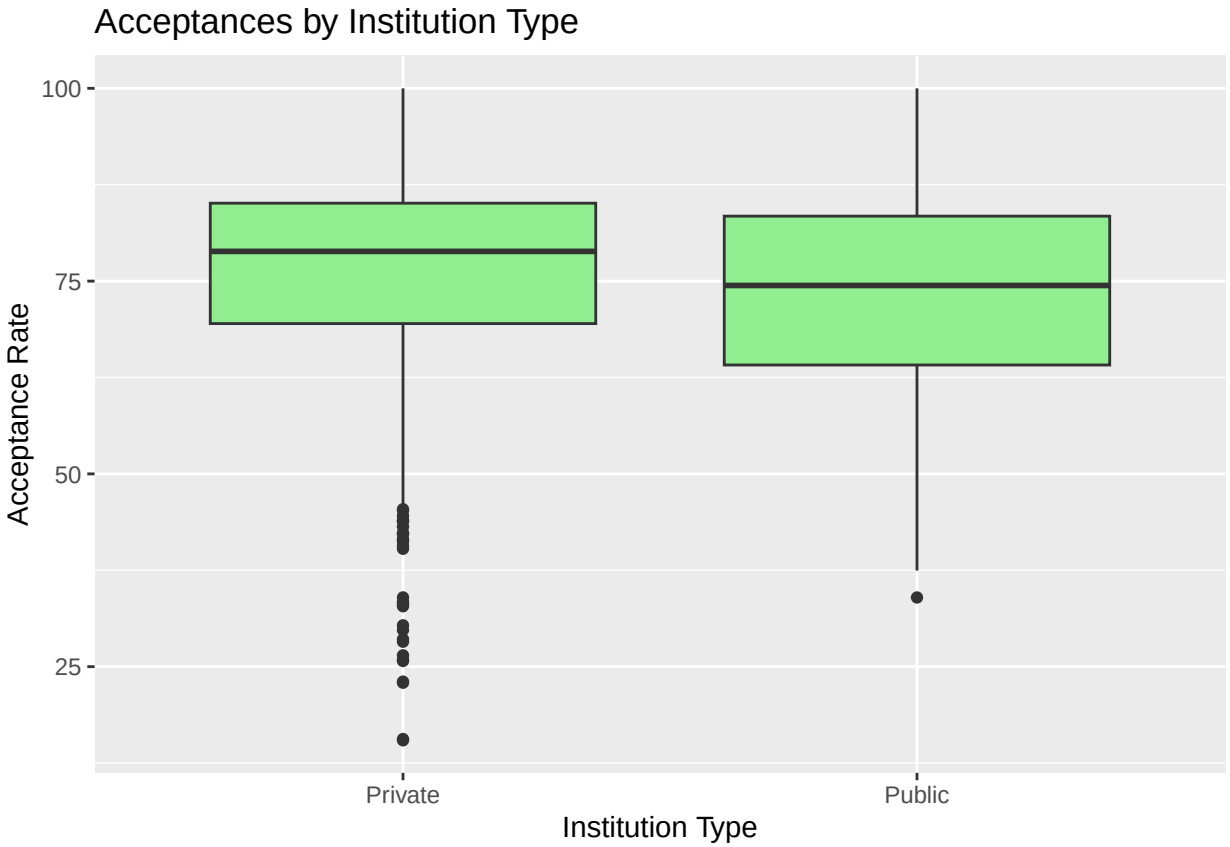
We know there are two types of institutions - private and public (hence we created our new variable earlier)

```
p7 <- ggplot(filtered_data, aes(x = Type_of_institution, y = Apps)) +
  geom_boxplot(fill = "skyblue") +
  labs(title = "Applications by Institution Type", x = "Institution Type", y = "Applications")
```



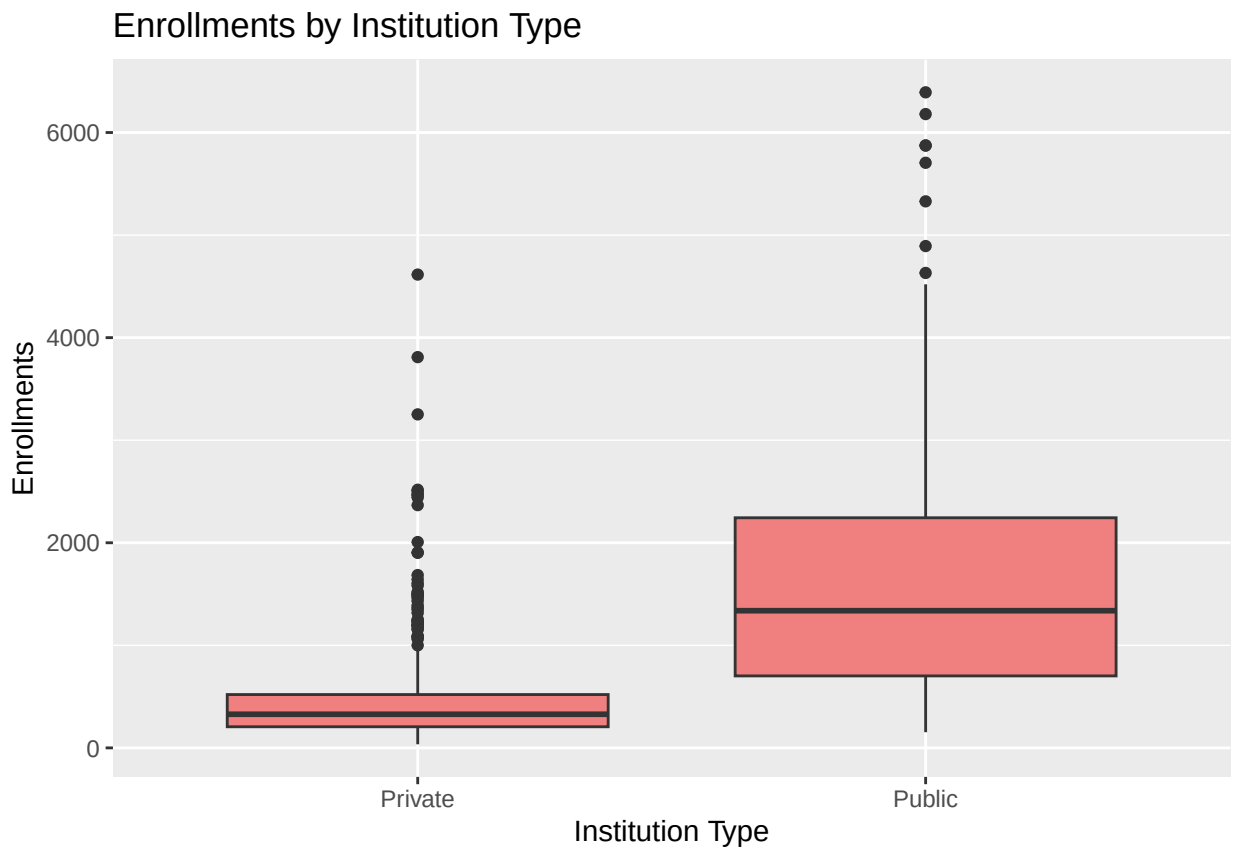

So from what we can see Private schools as a whole get fewer applications than public schools. The median applications number for public schools is a lot higher than private schools. The IQR of public schools also suggest that they have more variability in number of applications as compared to private schools. Public schools do have an outlier which received more than 40,000 applications. Private schools in general also have more outliers suggesting that certain private schools get more admission applications than other schools.

```
p8 <- ggplot(data, aes(x = Type_of_institution, y = AcceptanceRate)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Acceptances by Institution Type", x = "Institution Type", y = "Acceptance Rate")
```




Above we compare acceptance rates according to type of institution, we can see their median acceptance rate is about the same. The IQR for public schools is wider than private schools. This suggests that public schools have wider rates of acceptance compared to private schools. Private schools also have significantly more outliers than public schools which suggests private schools can be selective.

```
p9 <- ggplot(data, aes(x = Type_of_institution, y = Enroll)) +
  geom_boxplot(fill = "lightcoral") +
  labs(title = "Enrollments by Institution Type", x = "Institution Type", y = "Enrollments")
```

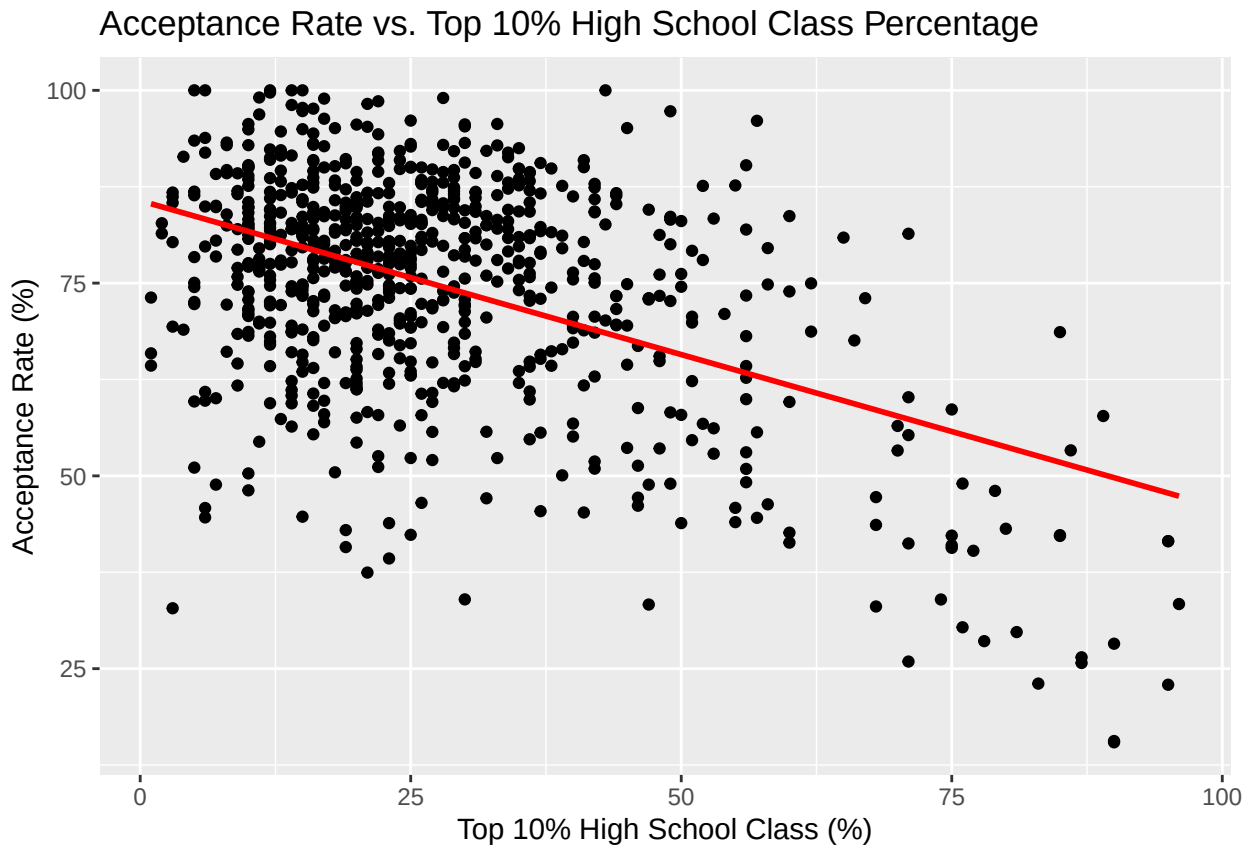



So from what we can see in the boxplot, public institutions generally have higher enrollments compared to private institutions. The median enrollment for public institutions is significantly higher, and they also show greater variability, with numerous outliers indicating some public schools have exceptionally high enrollments. Private institutions, while having lower median enrollments, display a higher number of outliers, suggesting variability in their enrollments too.

### Do students in the top 10% of their class have high acceptance rates?

```
p10 <- ggplot(data, aes(x = Top10perc, y = AcceptanceRate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Acceptance Rate vs. Top 10% High School Class Percentage", x = "Top 10% High School Cla

print(p10)
```

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




So from what we can see in the scatter plot, there is a noticeable negative relationship between the percentage of students in the top 10% of their high school class and the acceptance rate of institutions.

First, the trend line (red line) shows a clear downward slope, indicating that as the percentage of students in the top 10% of their high school class increases, the acceptance rate tends to decrease. This suggests that institutions with a higher proportion of top-performing students are generally more selective.

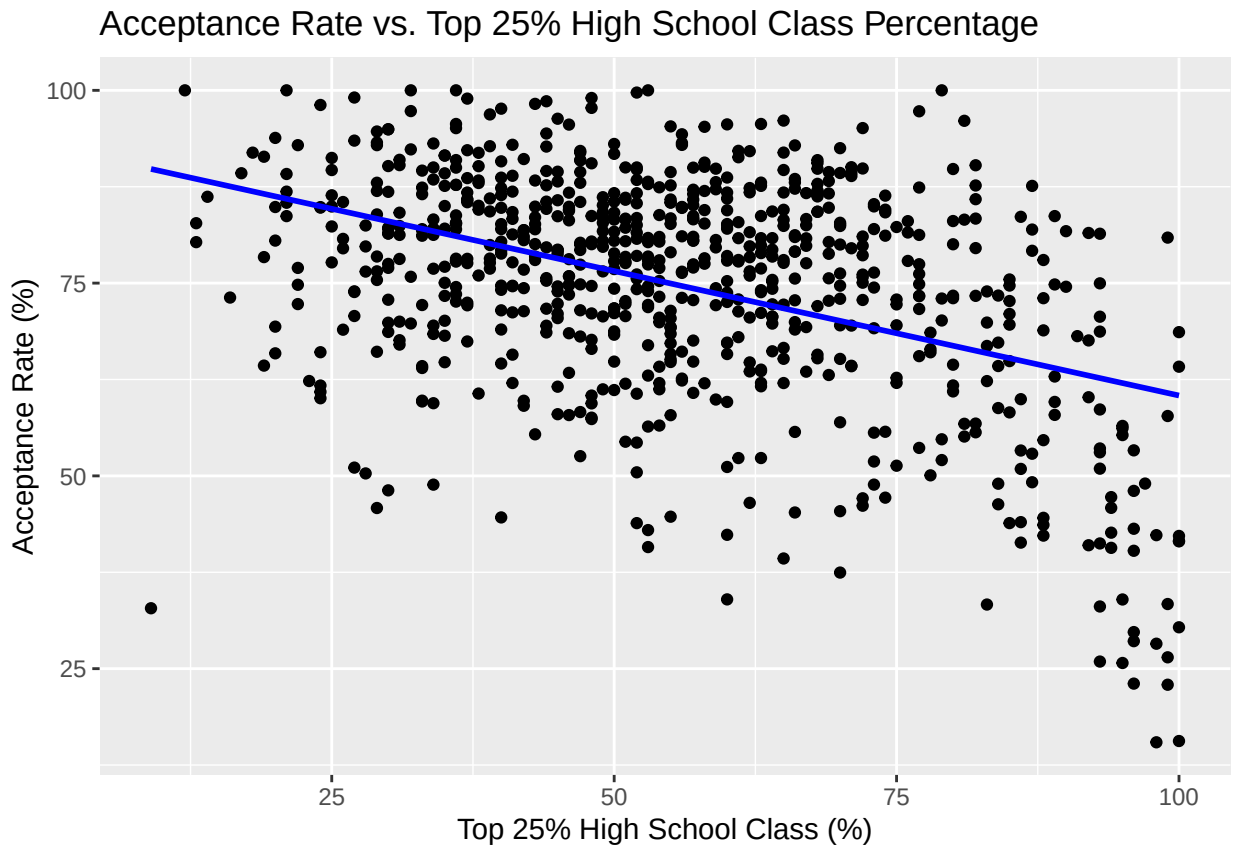
Next, we observe that institutions with a higher percentage of top 10% high school students tend to have acceptance rates clustered towards the lower end. This aligns with the expectation that more selective institutions attract academically elite students, resulting in lower acceptance rates.

Additionally, the data points are spread out more widely at lower percentages of top 10% students, showing higher variability in acceptance rates for these institutions. This could indicate that schools with fewer top 10% students have a broader range of selectivity, from very selective to less selective.

### Do students in the top 25% of their HS class do as well as the top 10%?

```
p11 <- ggplot(data, aes(x = Top25perc, y = AcceptanceRate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Acceptance Rate vs. Top 25% High School Class Percentage", x = "Top 25% High School Cla

print(p11)
```

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

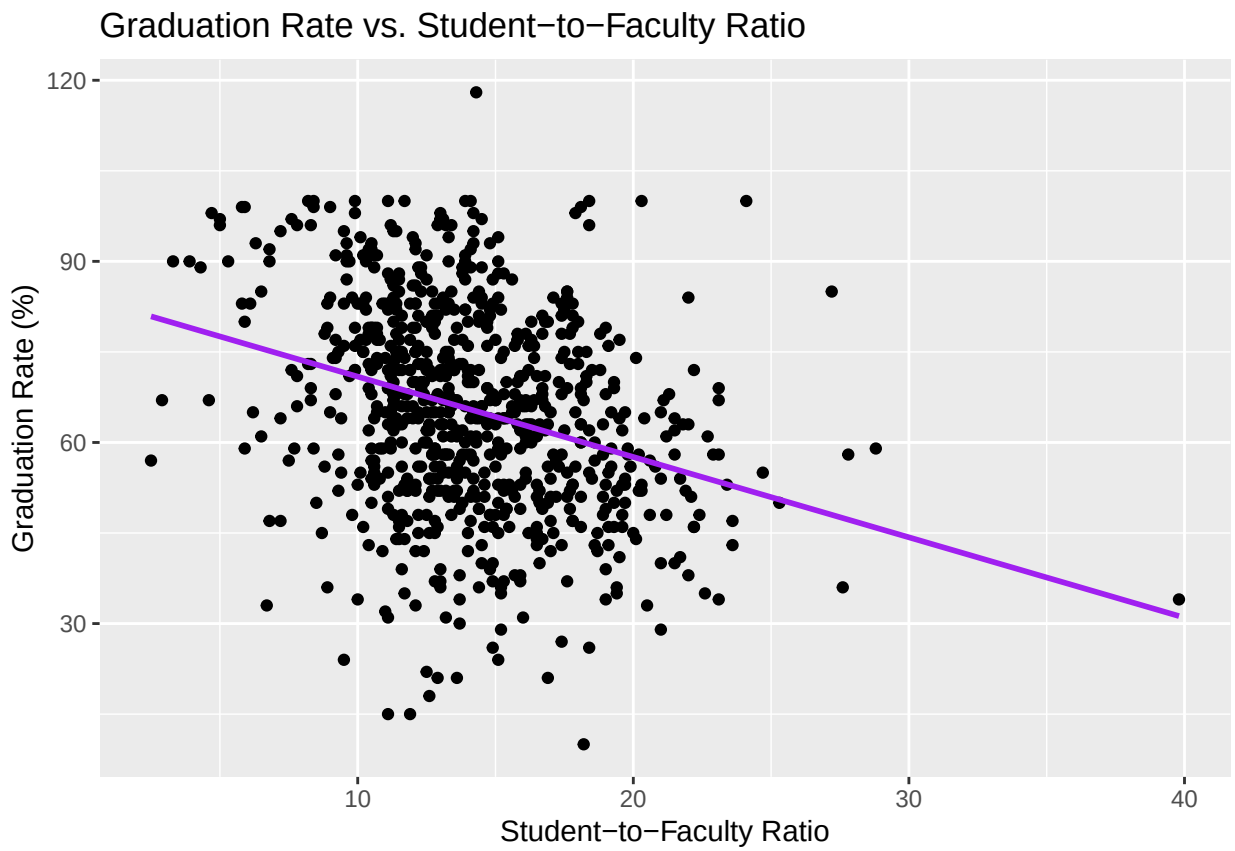



In short, it is the same story as their fellow 10% compatriots. There is a negative relationship between the percentage of students in the top 25% of their high school class and the acceptance rate. As the percentage of top-performing students increases, the acceptance rate tends to decrease. Turns out competition exists in every bracket of admissions.

### How does Student Faculty ratio affect the graduation rate?

```
p12 <- ggplot(data, aes(x = S.F.Ratio, y = Grad.Rate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "purple") +
  labs(title = "Graduation Rate vs. Student-to-Faculty Ratio", x = "Student-to-Faculty Ratio", y = "Grad

print(p12)
```

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

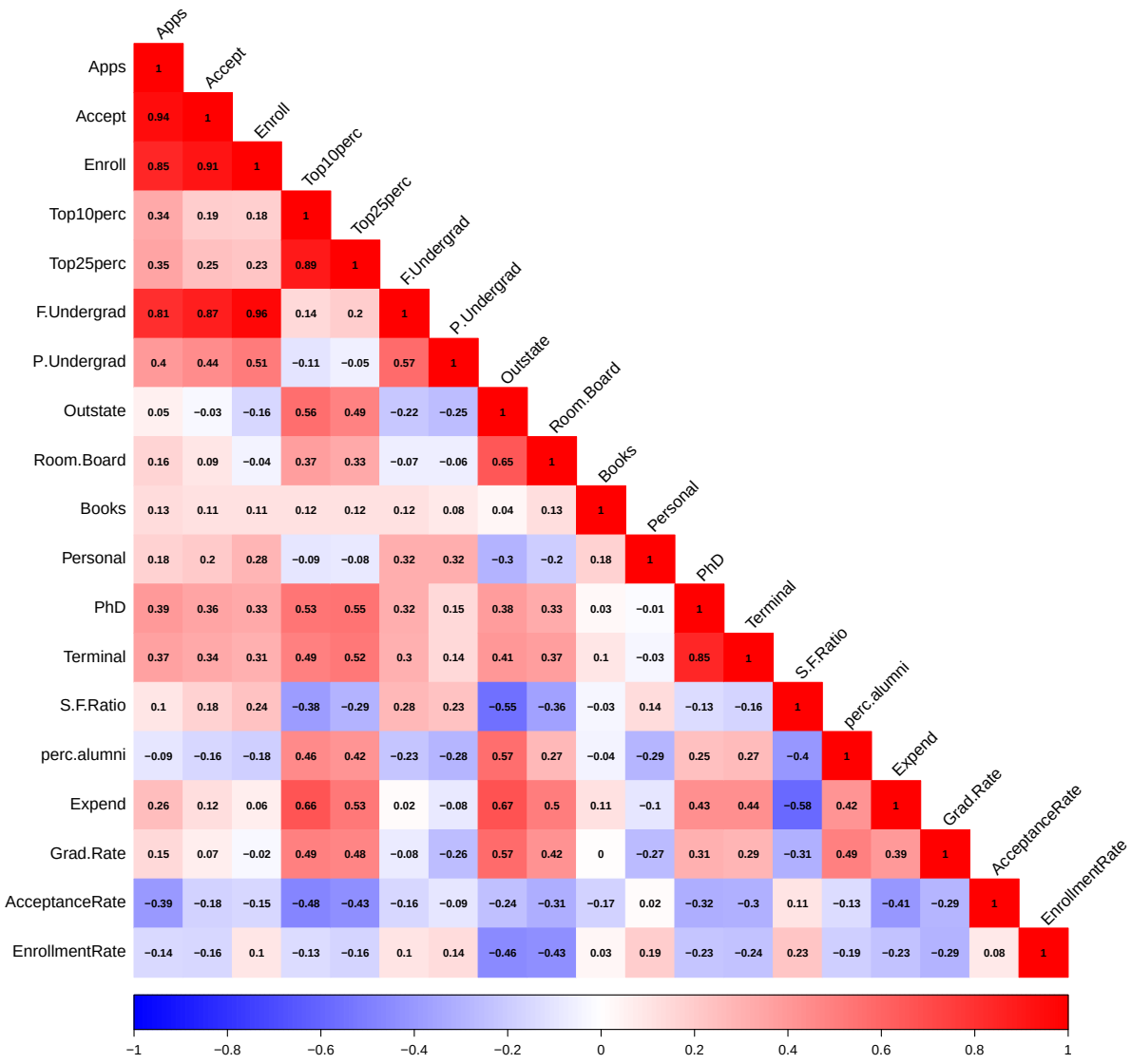



There is a clear negative relationship between the student-to-faculty ratio and the graduation rate. As the student-to-faculty ratio increases, the graduation rate tends to decrease. This suggests that institutions with smaller class sizes generally have higher graduation rates.

## Heatmap

We now look into a heat map

```
# Select only the numerical columns for the correlation matrix
numerical_data <- data %>%
  select(Apps, Accept, Enroll, Top10perc, Top25perc, F.Undergrad, P.Undergrad,
         Outstate, Room.Board, Books, Personal, PhD, Terminal, S.F.Ratio,
         perc.alumni, Expend, Grad.Rate, AcceptanceRate, EnrollmentRate)

# Calculate the correlation matrix
correlation_matrix <- cor(numerical_data)

# Create a heat map of the correlation matrix

corrplot(correlation_matrix, method = "color", type = "lower",
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black", number.cex = 0.7,
         col = colorRampPalette(c("blue", "white", "red"))(200))
```




So from what we can see in the heatmap, several key relationships between the variables in our dataset become evident.

First, we observe a strong positive correlation between the number of applications and the number of acceptances (0.94). This tells us that institutions receiving more applications tend to accept more students, which is expected. Similarly, there's a notable correlation between applications and enrollments (0.89), suggesting that higher application numbers generally translate to more students enrolling.

Next, there's a significant positive correlation between the number of full-time undergraduates and both the number of applications (0.87) and acceptances (0.87). This indicates that larger institutions tend to attract and accept more applicants. We also see a strong correlation between the number of full-time undergraduates and the number of enrollments (0.96), emphasizing that larger institutions enroll more students.

Interestingly, the percentage of alumni who donate has a moderate positive correlation with the percentage of students in the top 10% of their high school class (0.21) and with the graduation rate (0.31). This suggests that institutions with more academically elite students and higher graduation rates may also have higher

alumni donation rates.

The student-to-faculty ratio shows a moderate negative correlation with the graduation rate (-0.36) and a strong negative correlation with the percentage of alumni who donate (-0.40). This implies that institutions with lower student-to-faculty ratios (indicating smaller class sizes) tend to have higher graduation rates and more generous alumni.

Additionally, we see that expenditure per student is moderately positively correlated with the number of full-time undergraduates (0.66) and the graduation rate (0.39), suggesting that higher spending is associated with larger student bodies and better graduation outcomes.

Finally, acceptance rate has a moderate negative correlation with the number of applications (-0.39) and the number of full-time undergraduates (-0.36). This indicates that institutions with more applications and larger student bodies tend to have lower acceptance rates, likely due to higher selectivity.

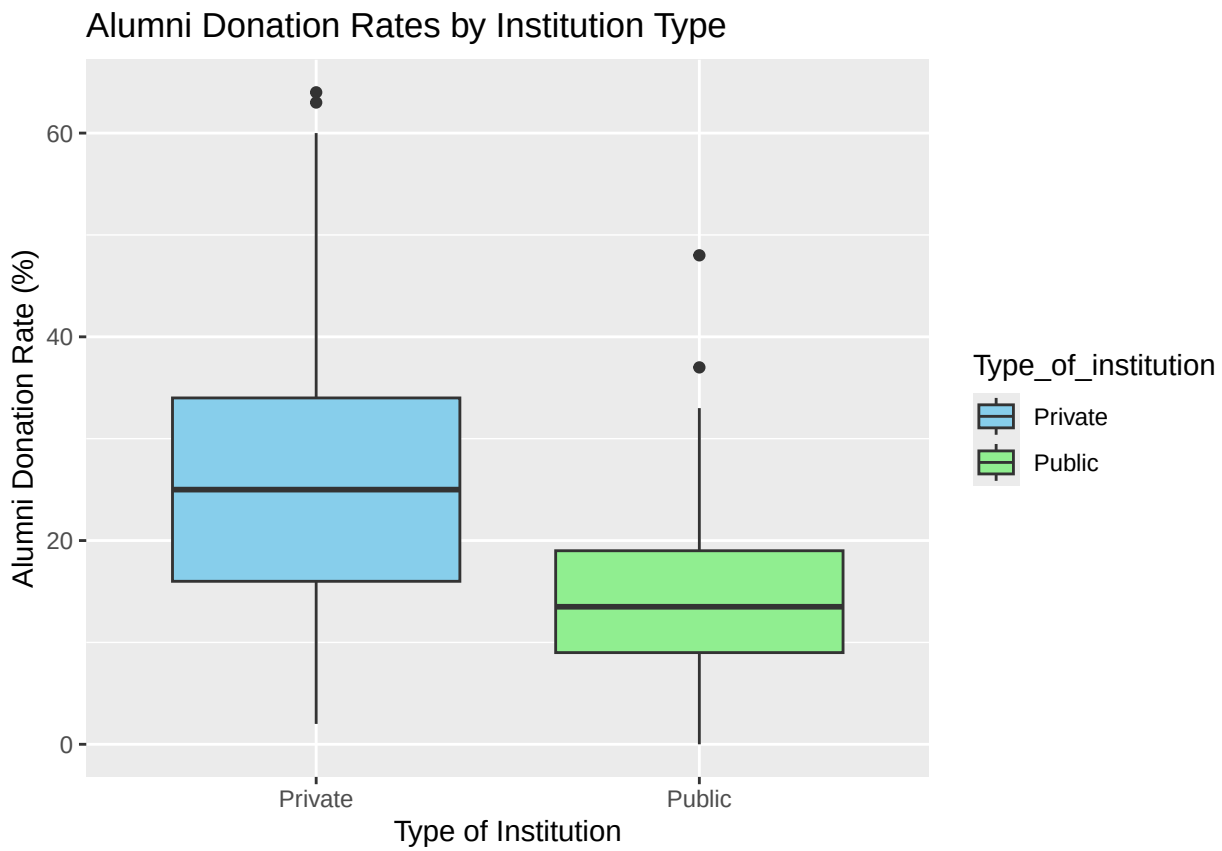
## Do private school alumni donate more?

Let's find out.

```
# Calculate summary statistics for alumni donation rates
summary_stats <- data %>%
  group_by(Type_of_institution) %>%
  summarise(
    mean_donation_rate = mean(perc.alumni, na.rm = TRUE),
    median_donation_rate = median(perc.alumni, na.rm = TRUE),
    sd_donation_rate = sd(perc.alumni, na.rm = TRUE)
  )

print(summary_stats)
```

```
## # A tibble: 2 x 4
##   Type_of_institution mean_donation_rate median_donation_rate sd_donation_rate
##   <chr>                            <dbl>                <dbl>            <dbl>
## 1 Private                           25.9                 25               12.4
## 2 Public                            14.4                 13.5              7.52
```

```
# Visualize the comparison of alumni donation rates between private and public institutions
p <- ggplot(data, aes(x = Type_of_institution, y = perc.alumni, fill = Type_of_institution)) +
  geom_boxplot() +
  labs(title = "Alumni Donation Rates by Institution Type",
       x = "Type of Institution",
       y = "Alumni Donation Rate (%)") +
  scale_fill_manual(values = c("Private" = "skyblue", "Public" = "lightgreen"))

print(p)
```




First, private institutions generally have higher alumni donation rates compared to public institutions. The median donation rate for private institutions is significantly higher than that for public institutions, indicating that alumni from private schools tend to donate more frequently.

Next, the interquartile range (IQR) for private institutions is wider than that for public institutions. This suggests greater variability in donation rates among private institutions. Some private schools have exceptionally high donation rates, while others have lower rates.

We also observe several outliers in both types of institutions. Notably, private institutions have outliers with alumni donation rates exceeding 60%, which indicates that some private schools receive substantial support from their alumni. Public institutions, on the other hand, have fewer and less extreme outliers.

Additionally, the overall range of alumni donation rates (whiskers) is broader for private institutions. This further emphasizes the variability and the potential for high alumni support in private schools.

In summary, private institutions tend to have higher and more variable alumni donation rates compared to public institutions. This suggests that alumni from private schools are generally more likely to donate and that there is a broader spectrum of donation behaviors among these institutions.

## Our model

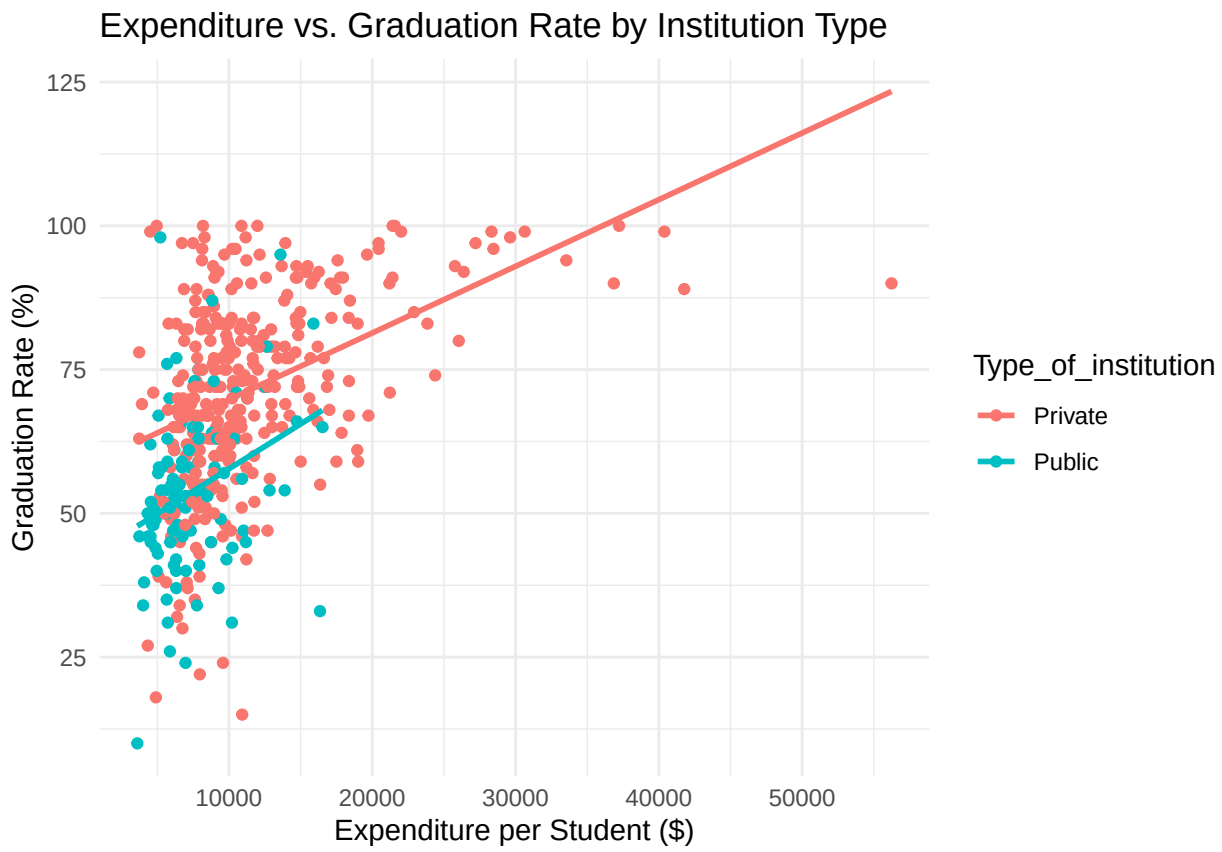
In the case of building a model to answer a question, we will do that here.

Our question in particular is:

How does the expenditure per student influence the graduation rate, and does this relationship vary between public and private institutions?

```r
p <- ggplot(filtered_data, aes(x = Expend, y = Grad.Rate, color = Type_of_institution)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Expenditure vs. Graduation Rate by Institution Type",
       x = "Expenditure per Student ($)",
       y = "Graduation Rate (%)") +
  theme_minimal()

print(p)
```

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




For both public (blue) and private (red) institutions, increased expenditure per student is associated with higher graduation rates. Private institutions show a steeper increase, indicating a stronger relationship between spending and graduation outcomes compared to public institutions. This visual evidence supports the regression analysis, highlighting that while both institution types benefit from higher expenditure, private institutions see a more pronounced impact on graduation rates.

```r
# Separate data for public and private institutions
public_data <- filtered_data %>% filter(Type_of_institution == "Public")
private_data <- filtered_data %>% filter(Type_of_institution == "Private")

# Linear regression model for public institutions
public_model <- lm(Grad.Rate ~ Expend, data = public_data)
summary(public_model)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Expend, data = public_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.771  -7.217  -0.490   7.682  47.723
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.215e+01  3.787e+00  11.129   <2e-16 ***
## Expend      1.560e-03  4.778e-04   3.264   0.0015 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.66 on 102 degrees of freedom
## Multiple R-squared:  0.09456,    Adjusted R-squared:  0.08569
## F-statistic: 10.65 on 1 and 102 DF,  p-value: 0.001496
```

```
# Linear regression model for private institutions
private_model <- lm(Grad.Rate ~ Expend, data = private_data)
summary(private_model)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Expend, data = private_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.852  -8.417   0.624   8.977  36.063
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.819e+01  1.671e+00  34.826   <2e-16 ***
## Expend      1.159e-03  1.284e-04   9.029   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.7 on 349 degrees of freedom
## Multiple R-squared:  0.1894, Adjusted R-squared:  0.187
## F-statistic: 81.53 on 1 and 349 DF,  p-value: < 2.2e-16
```

```
# Comparing the models
stargazer(public_model, private_model, type = "text",
          title = "Regression Results: Graduation Rate vs. Expenditure",
          column.labels = c("Public Institutions", "Private Institutions"),
          covariate.labels = c("Expenditure per Student ($)"),
          dep.var.labels = "Graduation Rate (%)",
          no.space = TRUE)
```

```
##
## Regression Results: Graduation Rate vs. Expenditure
## ================================================================================
```

```
##                                      Dependent variable:
##                              ---------------------------------------------
##                                      Graduation Rate (%)
##                                Public Institutions    Private Institutions
##                                        (1)                    (2)
## ------------------------------------------------------------------------
## Expenditure per Student ()         0.002***               0.001***
##                                    (0.0005)               (0.0001)
## Constant                           42.149***              58.190***
##                                     (3.787)                (1.671)
## ------------------------------------------------------------------------
## Observations                          104                    351
## R2                                   0.095                  0.189
## Adjusted R2                          0.086                  0.187
## Residual Std. Error            13.656 (df = 102)      14.699 (df = 349)
## F Statistic                 10.653*** (df = 1; 102) 81.531*** (df = 1; 349)
## ========================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

The regression analysis shows that increased expenditure per student significantly improves graduation rates for both public and private institutions. In public institutions, each additional dollar spent per student increases the graduation rate by 0.002 percentage points, while in private institutions, it increases by 0.001 percentage points. Both relationships are statistically significant. However, the models explain only 9.5% and 18.9% of the variance in graduation rates for public and private institutions, respectively, suggesting other factors also influence graduation rates. Overall, private institutions show a slightly stronger relationship between expenditure and graduation rates compared to public institutions.

## Conclusion

To conclude, this dataset has provided us with many insights into the world of university admissions. One might look at some of these observations and expect an obvious answer but some answers have been different from what we expected. Thanks for reading :)