STAT508_DAA1

Code

Ada Lazuli

2026-01-19

Front Matter

Hide

```
library(ggplot2)
library(dplyr)
library(reactable) # for knitting interactive tables into the report
college_df <- read.csv("../sample_data/L01_Colleges.csv")
# 875820569106c091e8c8e3b0e4e08c06bf51cce0  L01_Colleges.csv
```

Problem 1 - Enrollment

The research objective is to explore the *Colleges.csv* and identify patterns in the the number of students enrolled.

Data Overview

The data consists of 777 observations with 19 fields for each observation.

Hide

```
reactable(college_df, searchable = TRUE, compact = TRUE, striped = TRUE, defaultPag
        eSize = 5, showPageSizeOptions = TRUE)
```

Search

Name	Private	Apps	Accept	Enroll	Top10perc	Top25p
Abilene Christian	Yes	1660	1232	721	23	

University					
Adelphi University	Yes	2186	1924	512	16
Adrian College	Yes	1428	1097	336	22
Agnes Scott College	Yes	417	349	137	60
Alaska Pacific University	Yes	193	146	55	16

1–5 of 777 rows Show 10 Previous **1** 2 3 4 5 ... 156 Next

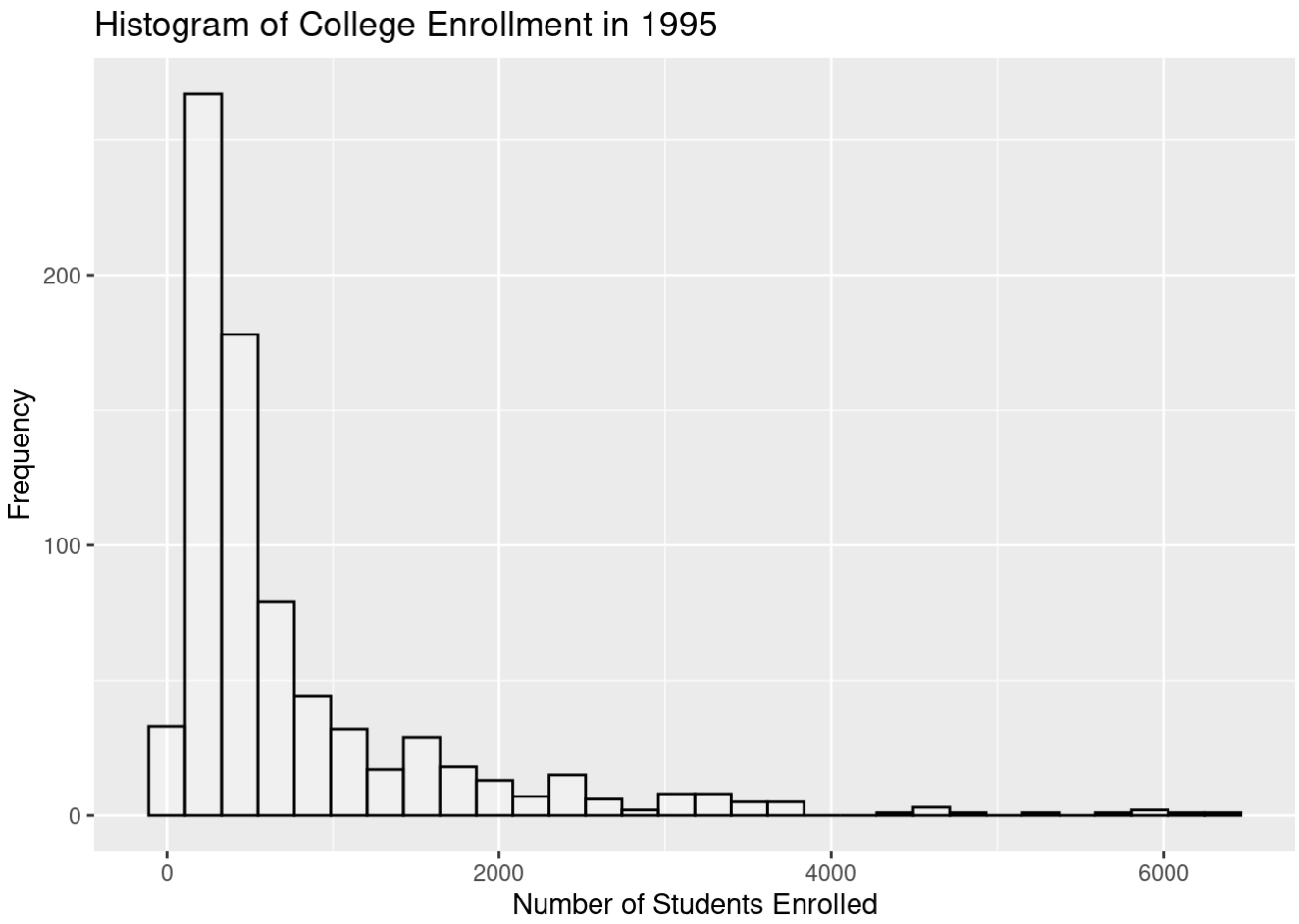
exception of the Name and whether the college is private, all of the fields are quantitative.

Enrollment Exploration

Starting with the exploration of the distribution of student enrollment.

Hide

```
ggplot(college_df, aes(Enroll)) +
  geom_histogram(alpha = .45, color = "black", fill = "white") +
  xlab("Number of Students Enrolled") +
  ylab("Frequency") +
  ggtitle("Histogram of College Enrollment in 1995")
```



The histogram shows a uni-modal distribution with an extreme right skew, with the vast majority of the colleges enrolling less than 2000 students.

Moreover, this data has the following numerical summaries:

1. **Mean**: 779.972973
2. **Median**: 434
3. **Standard Deviation**: 929.1761901
4. **IQR**: 660

The median number of students enrolled in colleges during the year 1995 was 434 student and it should be noted that this measure is affected by the extreme right skew observed earlier.

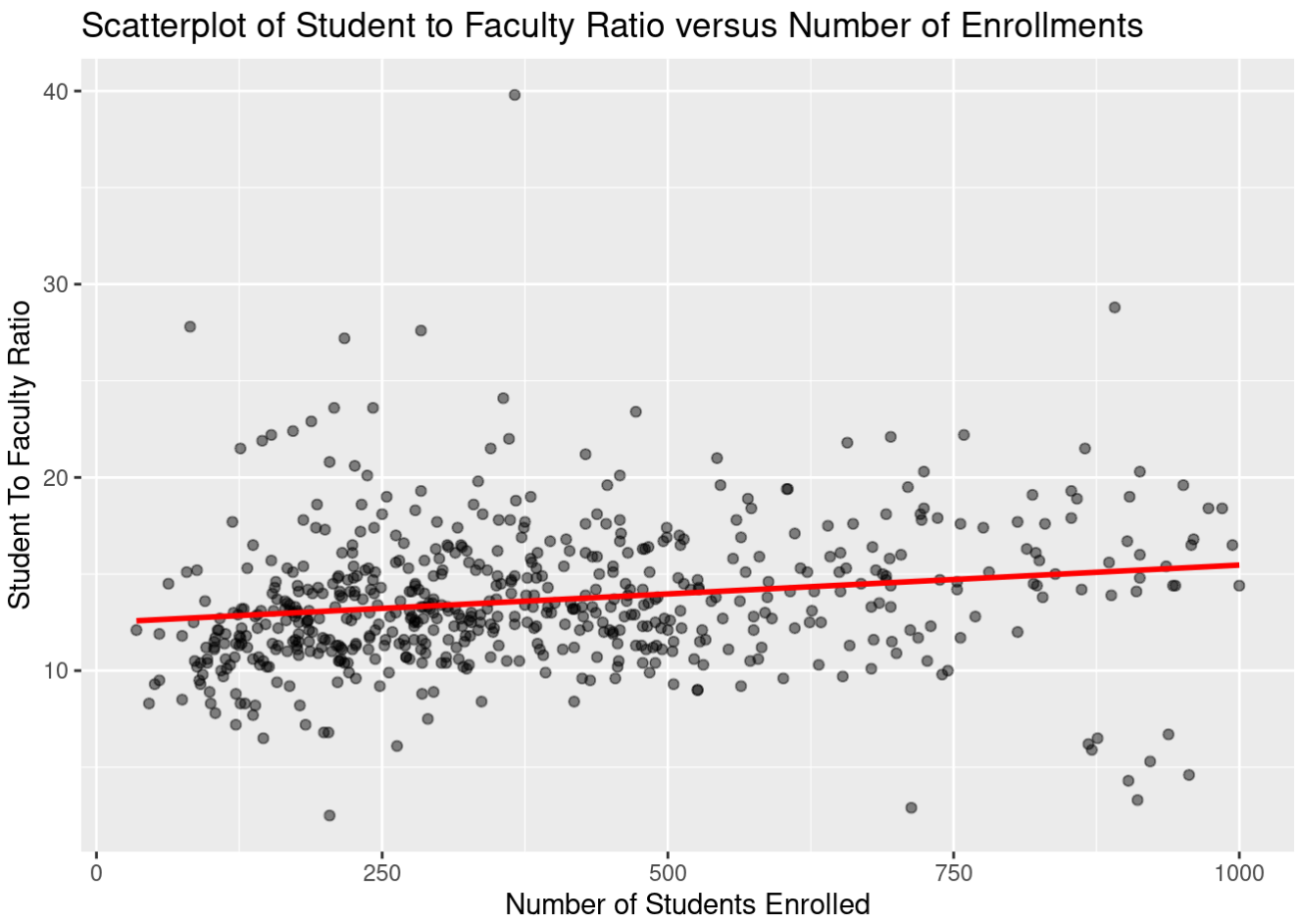
Problem 2 - Enrollment vs. Student to Faculty Ratio

The research question is whether there is a linear relationship between students enrolled and the faculty-to-student ratio for colleges with no more than 1000 students enrolled in 1995.

Hide

```
colleges_1000 <- college_df %>% filter(Enroll <= 1000)

ggplot(colleges_1000, aes(x = Enroll, y = S.F.Ratio)) +
  geom_point(alpha = .45) +
  geom_smooth(method = lm, se = FALSE, color = "RED") +
  xlab("Number of Students Enrolled") +
  ylab("Student To Faculty Ratio") +
  ggtitle("Scatterplot of Student to Faculty Ratio versus Number of Enrollments")
```



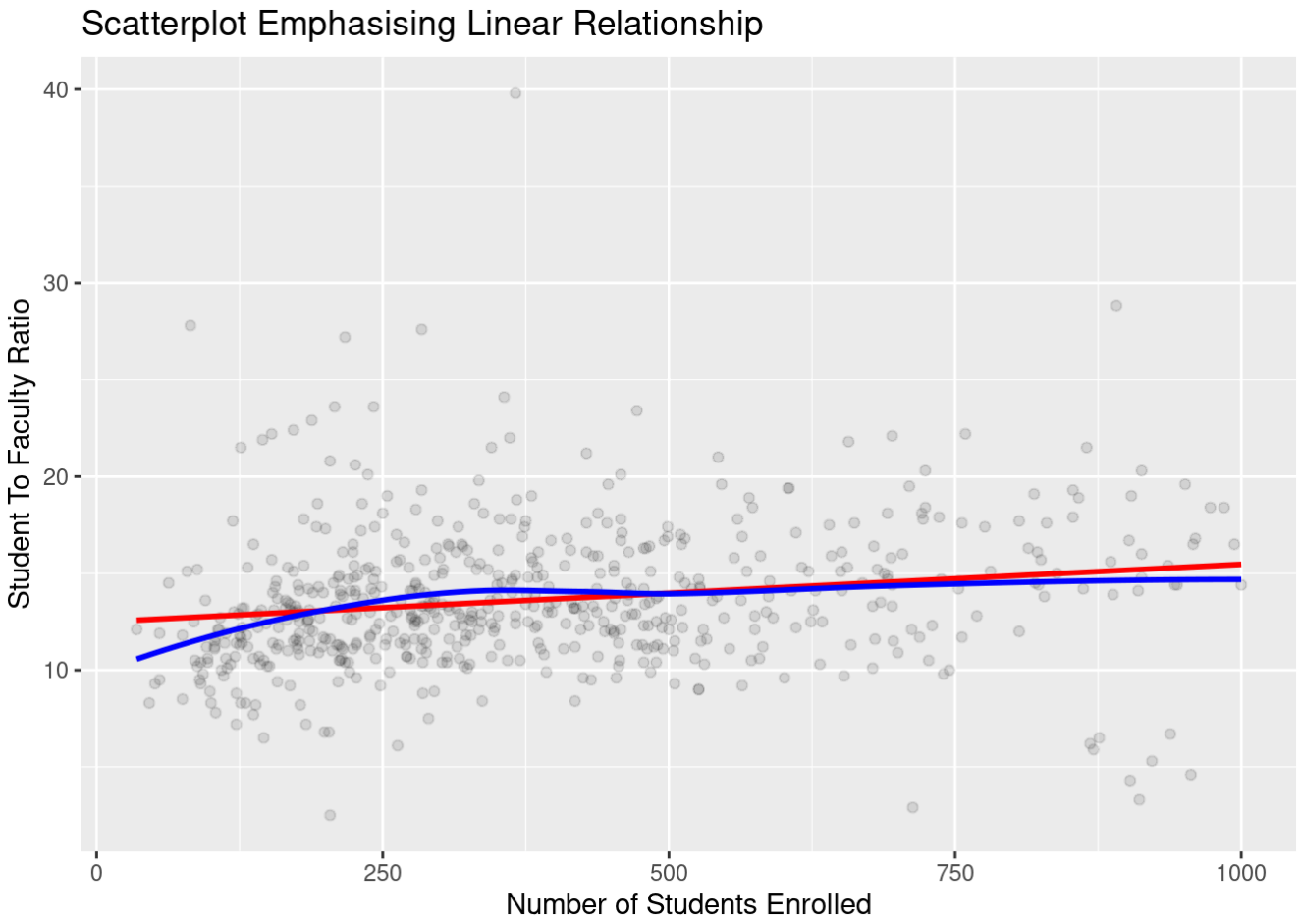
Based on the scatter plot, it appears that the student to faculty ratio slightly increases as the the number of students increases. However, this is not a strong linear relationship and more testing needs to be done.

The plot below with the added loess smoothing shows that the relationship could almost be logarithmic instead of linear.

Hide

```
colleges_1000 <- college_df %>% filter(Enroll <= 1000)

ggplot(colleges_1000, aes(x = Enroll, y = S.F.Ratio)) +
  geom_point(alpha = .1) +
  geom_smooth(method = lm, se = FALSE, color = "RED") +
  geom_smooth(method = loess, se = FALSE, color = "Blue") +
  xlab("Number of Students Enrolled") +
  ylab("Student To Faculty Ratio") +
  ggtitle("Scatterplot Emphasising Linear Relationship")
```



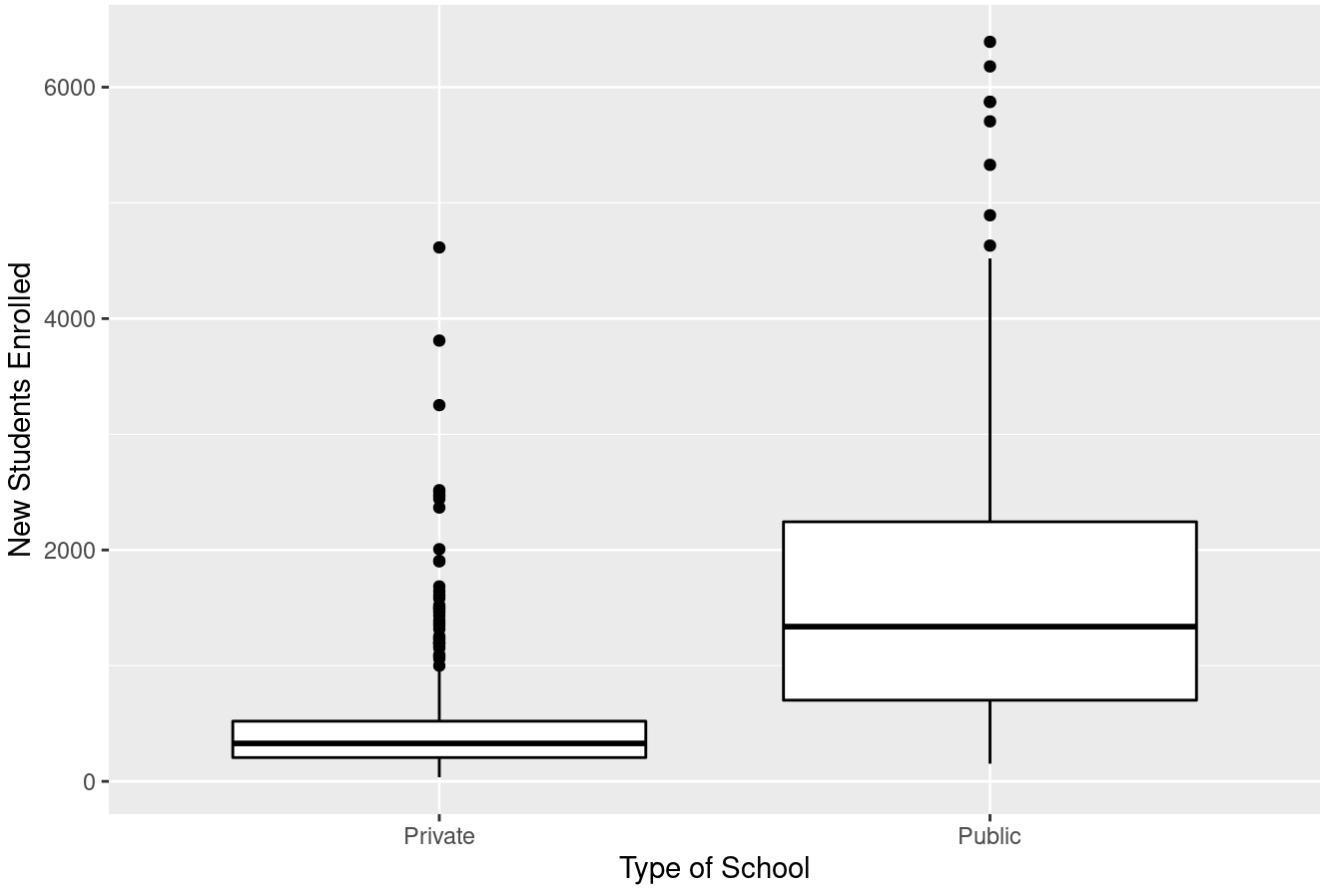
Problem 3 - Enrollment vs Type of School

The research questions is whether there is a relationship between the type of college and the enrollment.

Boxplots will be used for the initial investigation into the relationship.

Hide

```r
ggplot(
    college_df %>% mutate(Private = ifelse(Private == "No", "Public", "Private"))
    , aes(y = Enroll, x = Private)) +
  geom_boxplot(fill = "white", color = "black") +
  xlab("Type of School") +
  ylab("New Students Enrolled") +
  ggtitle("") # Title was absent in reference image.
```



Aside from a few outliers, it appears to be the case that private colleges enroll far fewer students than public colleges. However, there is some class imbalance here that could affect how representative this finding is. There are only **212 public colleges**, while there are **565 private colleges** represented in the sample.

Problem 4 - Outlier Investigation

While answering the previous research question, several outliers were observed regarding the enrollments of some private colleges.

Hide

```
top_7_private <- college_df %>%
  mutate(Type = ifelse(Private == "Yes", "Private", "Public")) %>%
  select(Name, Type, Enroll) %>%
  filter(Type == "Private") %>%
  arrange(desc(Enroll)) %>%
  top_n(7)

reactable(top_7_private, defaultPageSize = 7, compact = TRUE, striped = TRUE)
```

Name	Type	Enroll
Brigham Young University at Provo	Private	4615
Boston University	Private	3810
University of Delaware	Private	3252
Northeastern University	Private	2517
New York University	Private	2505
University of Southern California	Private	2477
University of Pennsylvania	Private	2464

It turns out that *Brigham Young University at Provo* is one of the largest outliers in regard to private college enrollments.

Problem 5 - Creating new variable and writing your own question

When the enrollment numbers of colleges are studied, questions around acceptance rates often arise. The sample does not provide the acceptance rate, but it could be calculated by:

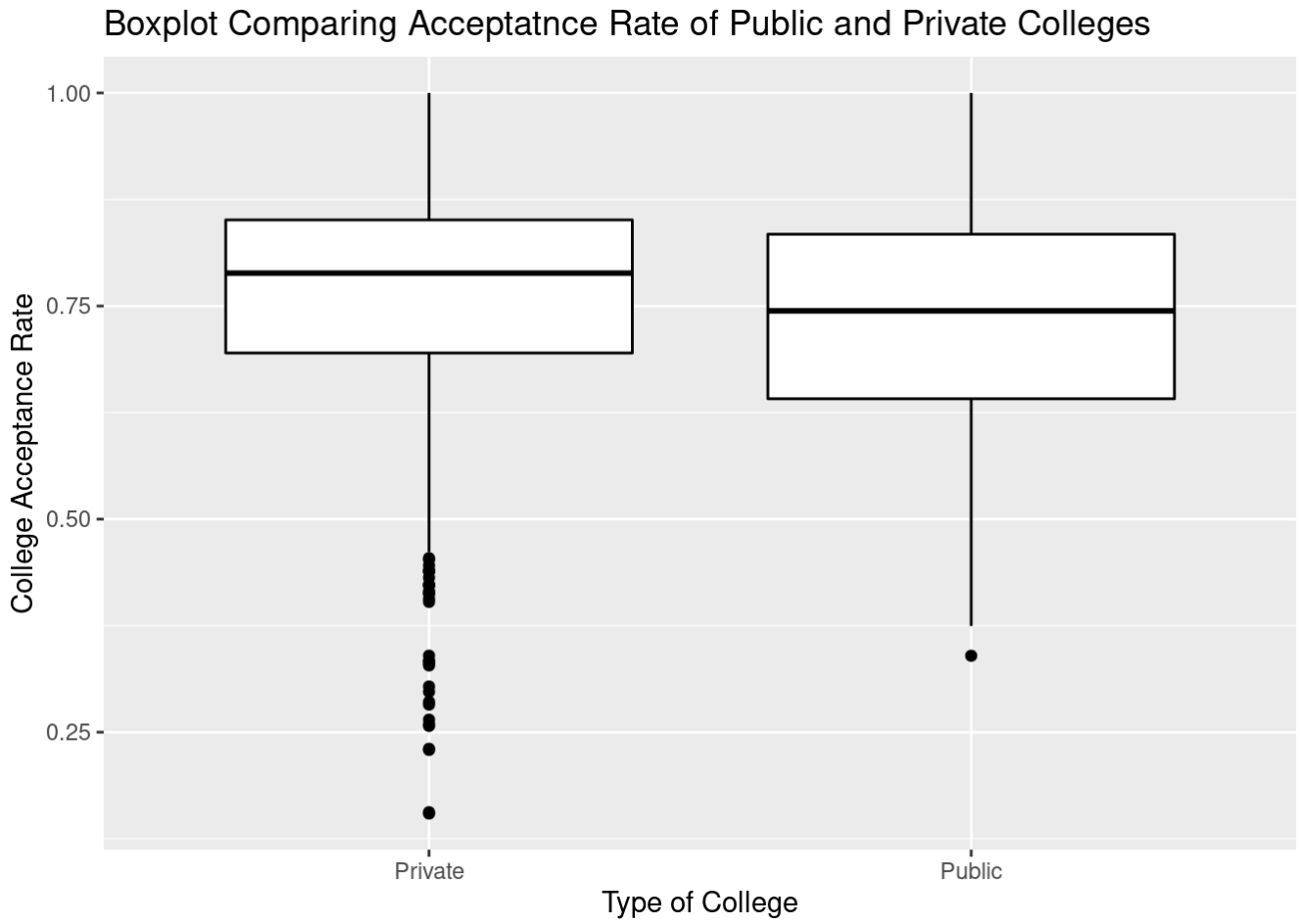
$$\text{Acceptance Rate} = \frac{\text{Accepted Applications}}{\text{Number of Applications}}$$

Since it was established earlier that private colleges enroll fewer students, it would be interesting to determine if it follows that the acceptance rate is also lower.

Hide

```
enriched_college_df <- college_df %>%
  mutate(AcceptanceRate = Accept/Apps) %>%
  mutate(Type = ifelse(Private == "Yes", "Private", "Public"))

ggplot(enriched_college_df, aes(x = Type, y = AcceptanceRate)) +
  geom_boxplot(fill = "white", color = "black") +
  xlab("Type of College") +
  ylab("College Acceptance Rate") +
  ggtitle("Boxplot Comparing Acceptatnce Rate of Public and Private Colleges")
```



Interestingly, the distribution of the acceptance rate seems to be very similar between public and private colleges.

Hide

```
enriched_summary <- enriched_college_df %>%
  group_by(Type) %>%
  summarize(
    Mean = mean(AcceptanceRate),
    Median = median(AcceptanceRate),
    "Standard Deviation" = sd(AcceptanceRate),
    "Interquartile Range" = quantile(AcceptanceRate, .75)[[1]] - quantile(Acceptanc
        eRate, .25)[[1]]
  )
reactable(enriched_summary, striped = TRUE)
```

Type	Mean	Median	Standard Deviation	Interquartile Range
Private	0.754581179824027	0.788565264293432	0.146734950352968	0.156121844884773
Public	0.726530480915445	0.744338718962189	0.146482254797395	0.193040848551747

As evidenced by the previous table, the numeric summary of the acceptance rate of private and public colleges is very similar.