EDAV Final Project Report

Qi Feng, Zhida Zhang, Zhiyuan Zhao

INTRODUCTION

With the fierce competition in the labor pool nowadays, more and more high schoolers choose to pursue a college education after graduation. College education not only equips students with knowledge and skills that are essential for their development, but also expands their networks which could advance them in future career. Yet, the search for an appropriate college can be overwhelming, given the vast range of factors one has to consider and the limitation of knowledge one has for colleges at hand.

Therefore, in our project, we would like to provide a way for students and their families to compare across different colleges based on the **cost-value** tradeoffs catering to their own needs, academic and career goals. We split cost-value tradeoffs into multiple parts, and define **cost** as integration of *admission requirement* and *affordability* of the college, and **value** as the combination of *education quality*, *diversity*, *completion* and *earnings after graduation*. Through our project, we will conduct exploratory data analysis on each part mentioned.

To guide the project, we propose the main research question: **how the values are related to the costs?** Specifically, do schools with higher admission requirement and higher tuition fees provide higher education quality, diversity, completion and earnings?

Our team members include Qi Feng, Zhida Zhang, and Zhiyuan Zhao. Qi Feng and Zhiyuan Zhao are responsible for data quality, and Zhida Zhang is in charge of presenting the results on the interactive platform of Shiny. Other than those, we split the remaining tasks evenly by each taking on two parts of the cost-value tradeoffs to perform analysis as well as writing the report. All of our work can be found here in the Github repository: https://github.com/ZhangZhida/EDAV_FinalProject.

DESCRIPTION OF DATA

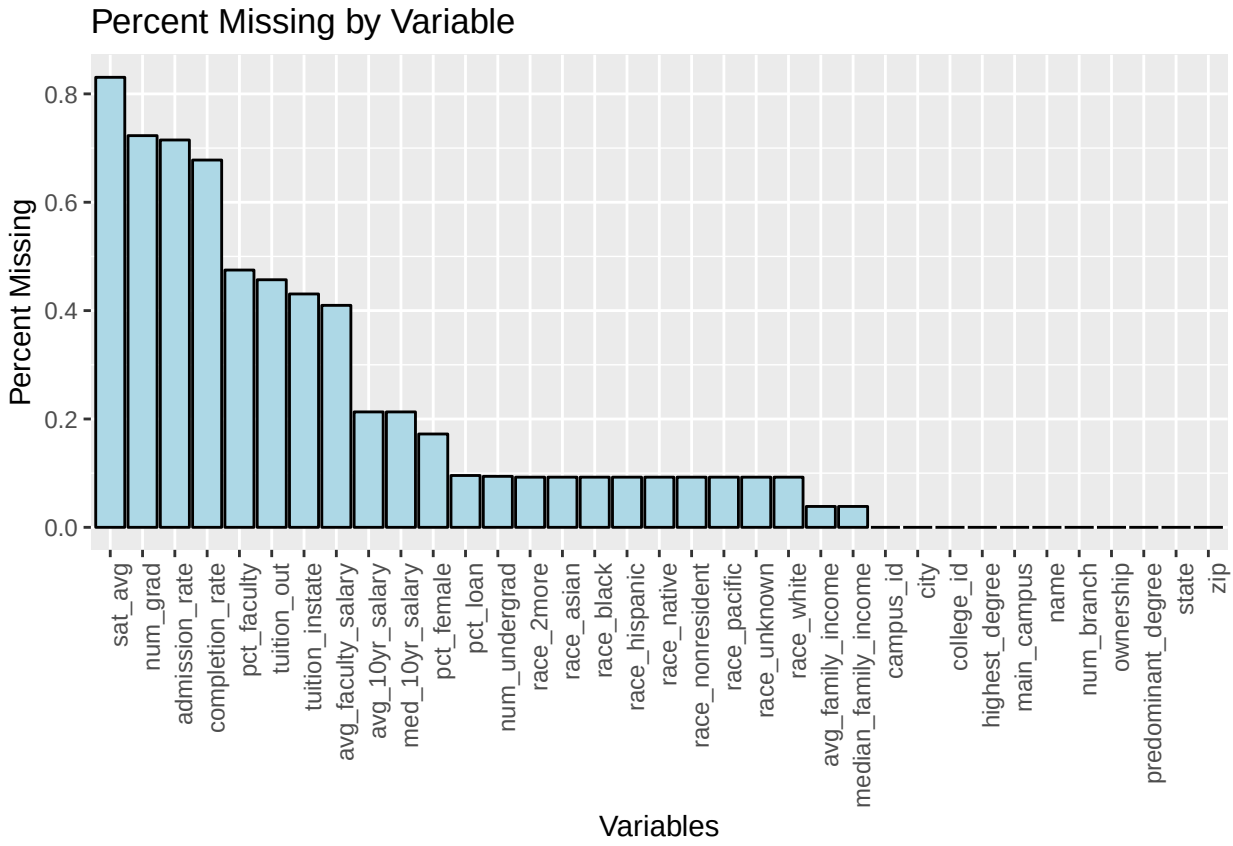
The dataset employed in this project was obtained from *The College Scorecard* (https://collegescorecard.ed.gov/data/), which integrates data from sources including IPEDS, NSLDS, and Department of Treasury. The data was collected from federal reporting of undergraduate degree-granting institutions of higher education. There were 1899 features in the raw dataset, covering topics including school, academics, admissions, student, cost, aid, repayment, completion and earnings. For our interest, we only focus on the data from 2014 to 2015, which is the most recent time span when earnings for graduates are available.

ANALYSIS OF DATA QUALITY

Since *The College Scorecard* project is supervised under the U.S. Department of Education, we are safe to presume the source of our data is valid. For our interest, we selected 35 features, of which 25 are continuous and 10 are categorical. A detailed data dictionary could be found at the GitHub repository. (https://github.com/ZhangZhida/EDAV_FinalProject/blob/master/Data%20Dictionary.pdf) Data cleaning code can also be found here https://github.com/ZhangZhida/EDAV_FinalProject/blob/master/Data%20cleaning.R.

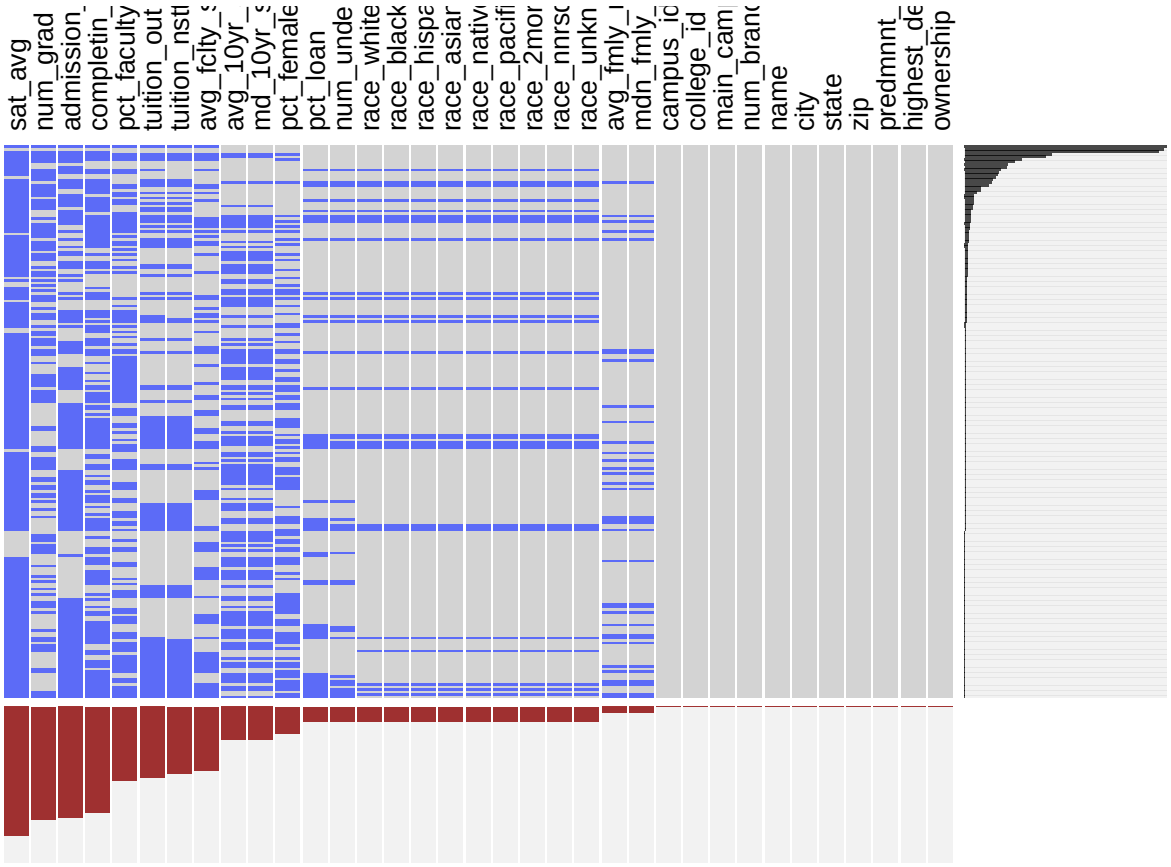
For missing value and patterns, we have the following analysis. According to the histogram below, for most schools, average SAT scores are missing. Number of graduate students, admission rate, and completion rate have missing value over 50%. Percent of full time faculty, in state tuition, out-of-state tuition, and average faculty salary have missing value a little under 50%. Other variables have lower than 12% missing value.

```
setwd(("/Users/Mavis1/Documents/ColumbiaUniv/Fall2018/EDAV/"))
library(tidyverse)
library(gridExtra)
college = read_rds("college.rds")
college[college=='NULL'] <- NA
a=data.frame(apply(college, 2, function(col)sum(is.na(col))/length(col)))
a$variables = rownames(a)
names(a) <- c("missing","variables")
rownames(a) <- NULL

ggplot(a, aes(x=reorder(factor(a$variables),-a$missing), y=a$missing))+
  geom_bar(stat="identity",fill = "lightblue", color = "black")+
  ggtitle("Percent Missing by Variable")+
  labs(x = "Variables", y = "Percent Missing") +
  theme(axis.text.x = element_text(size = rel(1),angle = 90, hjust = 1),
        axis.text.y = element_text(size = rel(1)),
        strip.text = element_text(size = rel(1)))
```



The most frequent missing pattern is missing average SAT, number of graduate, admission rate, completion rate, percent of full time faculty, in-state tuition, out-of-state tuition, and average faculty salary at the same time. The eight variables are also the ones with most missing values. For race, missing race percentages are the same among different race types, and the missing pattern for race percentage is the same across data, meaning when one race percent is missing, all are missing.

```
library(extracat)
visna(college,sort="b")
```


sat_avg
num_grad
admission_
completin_
pct_faculty
tuition_out
tuition_nstl
avg_fclty_s
avg_10yr_
md_10yr_s
pct_female
pct_loan
num_unde
race_white
race_black
race_hispa
race_asiar
race_nativ
race_pacifi
race_2mor
race_nnrst
race_unkn
avg_fmly_i
mdn_fmly_
campus_id
college_id
main_cam
num_branc
name
city
state
zip
predmnnt_
highest_de
ownership

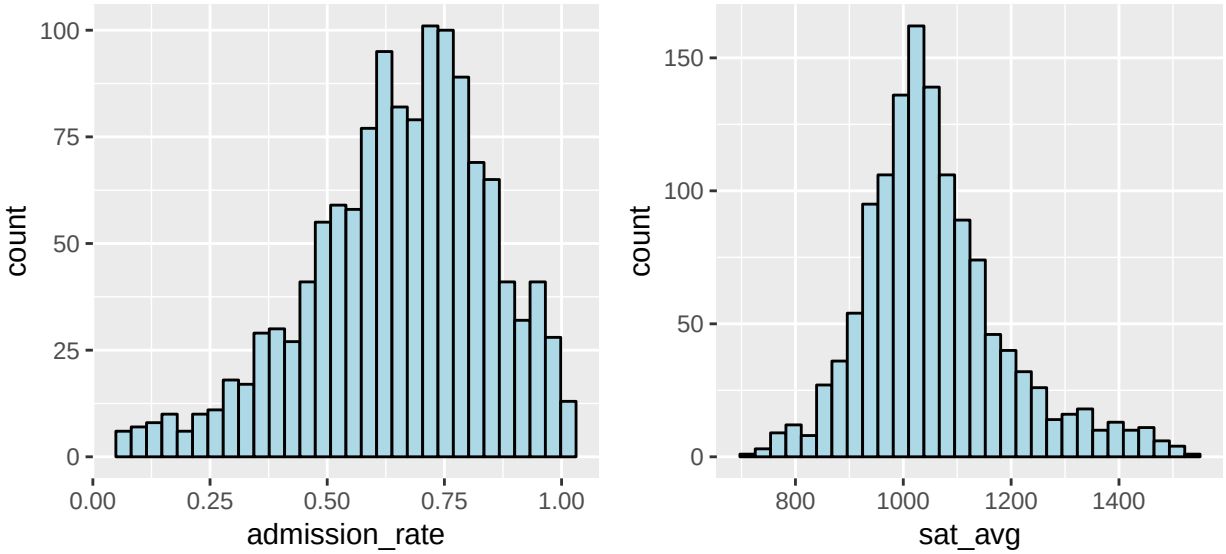
MAIN ANALYSIS

COST

Admission Requirement

To get a sense of how admission and SAT scores are distributed respectively, we plot the histogram on this two variables. From the histogram of admission rate, we could see that most schools have the admission rate range between 0.5 to 0.85. Only a small proportion of schools have admission rate below 0.25. From the SAT average histogram, we see that the majority of students enrolled into universities have scores in the range of 900 to 1200. The higher the score is, the less number of students there are.

```
library(ggplot2)
college <- readRDS('college.rds')
college_no_na <- college[with(college, (!is.na(sat_avg)) & (!is.na(admission_rate))), ]

plot_admi <- ggplot(college_no_na, aes(x = admission_rate)) +
  geom_histogram(fill = "lightblue", color = "black")

plot_sat <- ggplot(college_no_na, aes(x = sat_avg)) +
  geom_histogram(fill = "lightblue", color = "black")

grid.arrange(plot_admi, plot_sat, ncol = 2)
```

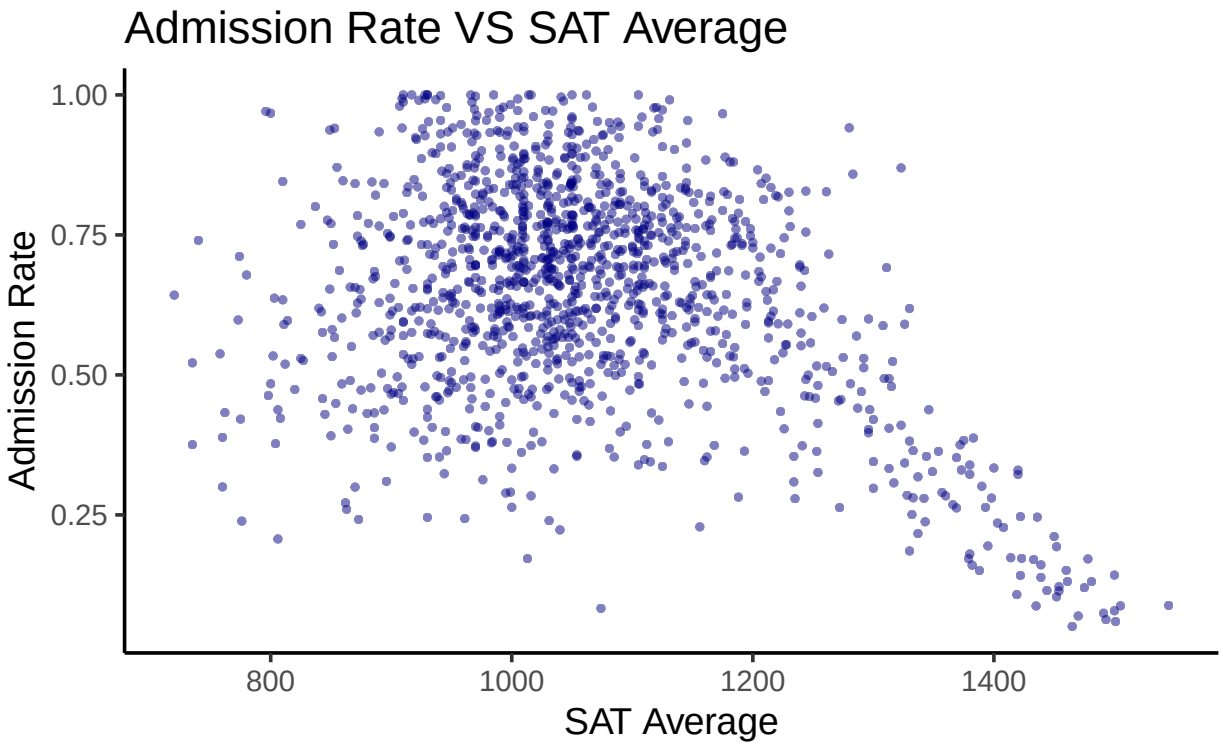


However, how do we know the relationship between admission rate and SAT average? Do we have any intuitive pattern like higher score and lower admission rate?

We try scatter plot to explore the relationship between them. As we can see below, there are mainly two patterns. For those schools with SAT average below 1200, the admission rate seems not tightly related with SAT average scores. That is, the SAT average is not the main factor when university enrolling students. On the other hand, for those universities with SAT average higher than 1200, we could clearly see a decreasing trend of admission rate with increasing SAT average. However, we cannot infer the internal reason for that. Based on our common sense, the reason for this pattern might be that when a school has very high SAT score requirement, it will lead to lower admission rate.

```
college_no_na <- college[with(college, (!is.na(sat_avg)) & (!is.na(admission_rate))), ]
sat_avg <- college_no_na$sat_avg
```

```r
admission_rate <- college_no_na$admission_rate
baseplt <- ggplot(college_no_na, aes(x = sat_avg, y = admission_rate))
baseplt + geom_point(alpha= 0.5, color = "navy", stroke = 0) +
  theme_classic(14) +
  labs(title = "Admission Rate VS SAT Average",
       x = "SAT Average",
       y = "Admission Rate")
```



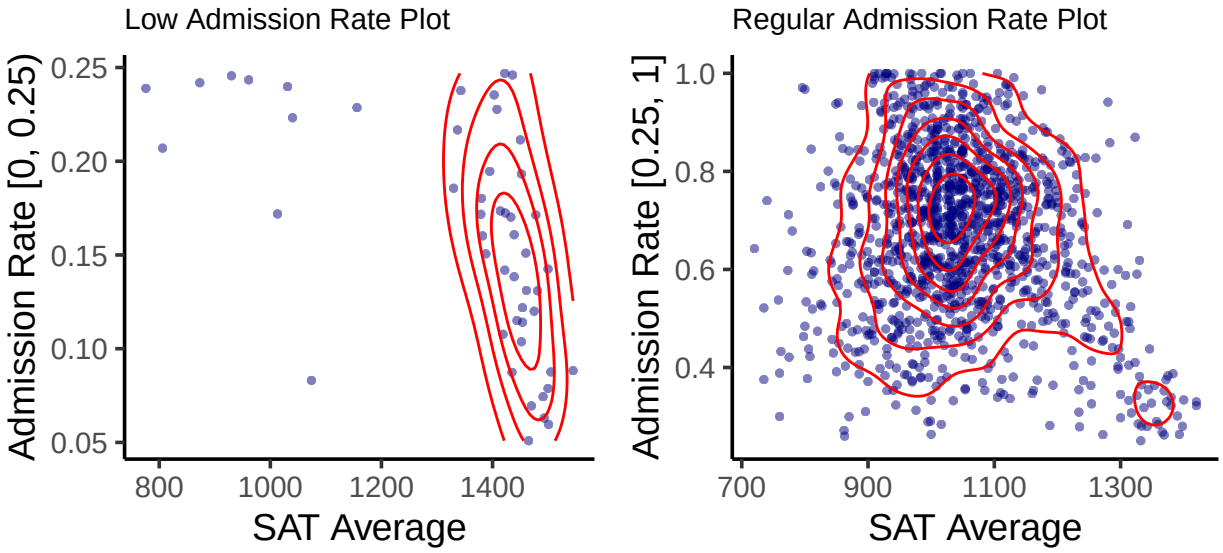
To follow on the previous patterns, we plot the density contour lines based on the scatter plots.

From the left-hand side graph below, we see there is a sharp decline pattern, that is, as SAR average become slightly higher, the admission rate immediately drops a lot. The reason for this may be that score higher than 1400 are very rare.

Meanwhile, we see some university with relatively lower SAT average still have very low admission rate. It is possibly based on the fact that SAT score is only one single requirement for universities to enroll students. These universities on the left part of the plot may have more emphasis on parts other than score, such as internship experience, sports talents, etc.

```r
college_no_na_low <- college[with(college, (!is.na(sat_avg)) &
                                  (!is.na(admission_rate)) & (admission_rate < 0.25)), ]
college_no_na_high <- college[with(college, (!is.na(sat_avg)) &
                                  (!is.na(admission_rate)) & (admission_rate >= 0.25)), ]
sat_avg <- college_no_na$sat_avg
admission_rate <- college_no_na$admission_rate

plot_low <- ggplot(college_no_na_low, aes(x = sat_avg, y = admission_rate))+
  geom_point(alpha= 0.5, color = "navy", stroke = 0) +
  theme_classic(14) + theme(plot.title = element_text(size = rel(0.8))) +
  geom_density_2d(color = "red", bins = 5) +
  labs(title = "Low Admission Rate Plot",
```

```
       x = "SAT Average",
       y = "Admission Rate [0, 0.25)")

plot_high <- ggplot(college_no_na_high, aes(x = sat_avg, y = admission_rate))+
  geom_point(alpha= 0.5, color = "navy", stroke = 0) +
  theme_classic(14) + theme(plot.title = element_text(size = rel(0.8))) +
  geom_density_2d(color = "red", bins = 8) +
  labs(title = "Regular Admission Rate Plot",
       x = "SAT Average",
       y = "Admission Rate [0.25, 1]")

grid.arrange(plot_low, plot_high, ncol = 2)
```

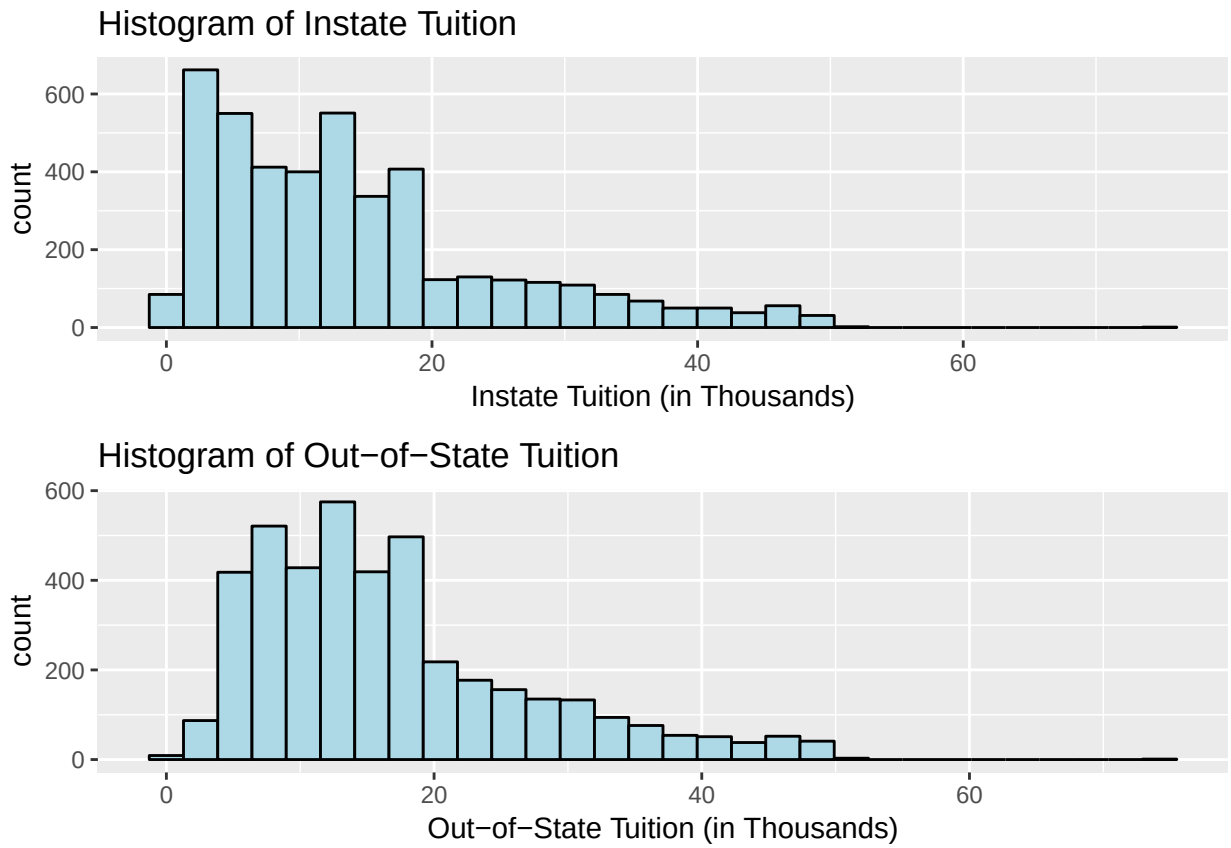


Affordability

Instate and out-of-state tuition are the tuition and fees for instate, and out-of-state students, respectively. From the histogram below, we can see that the distribution of the two tuition are similar, and they both skewed right. Out-of-state students have overall higher tuition fee than instate students.

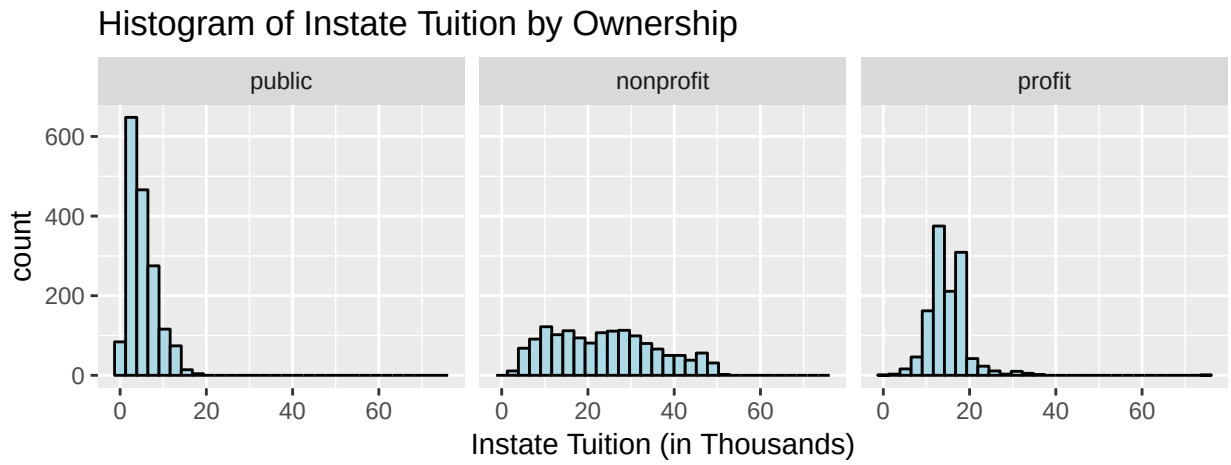
```
college = read_rds("college.rds")
college[college=='NULL'] <- NA
college1 =  college[!is.na(college$tuition_instate), ]
college2 =  college[!is.na(college$tuition_out), ]
college1$tuition_instate = college1$tuition_instate/1000
college2$tuition_out = college2$tuition_out/1000
recode <- c('1'='public','2'='nonprofit','3'='profit')
college1$ownership <- recode[college1$ownership]
college1$ownership<-fct_relevel(college1$ownership,'public','nonprofit','profit')
college2$ownership <- recode[college2$ownership]
college2$ownership<-fct_relevel(college2$ownership,'public','nonprofit','profit')
p3<-ggplot(college1, aes(x=tuition_instate))+
  geom_histogram(bins = 30,fill = "lightblue", color = "black")+
  labs(x="Instate Tuition (in Thousands)")+
  ggtitle("Histogram of Instate Tuition")
p4<-ggplot(college2, aes(x=tuition_out))+
```

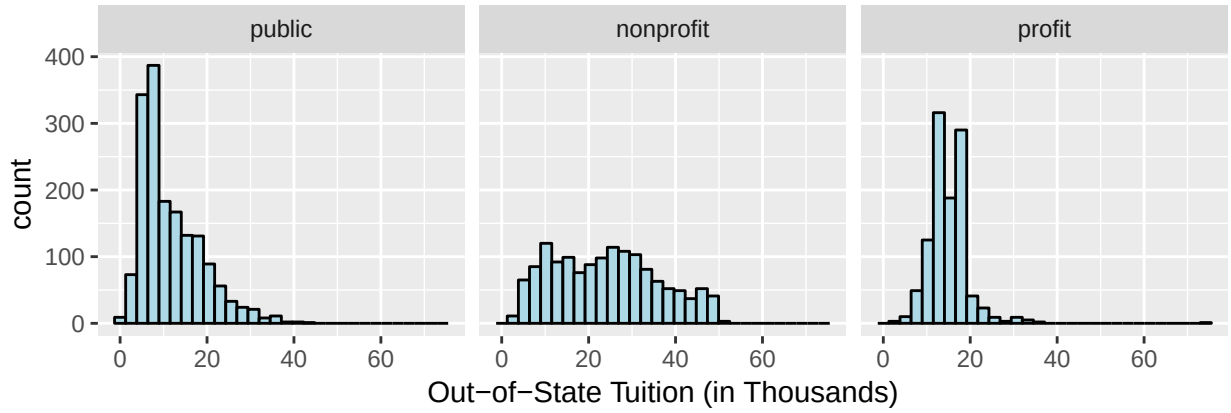
```
  geom_histogram(bins = 30,fill = "lightblue", color = "black")+
  labs(x="Out-of-State Tuition (in Thousands)")+
  ggtitle("Histogram of Out-of-State Tuition")
grid.arrange(p3,p4,ncol=1)
```



By faceting instate and out-of-state tuition into different ownership groups, we can see the different distributions for public, private nonprofit, and private for-profit schools, but the pattern for instate and out-of-state tuition is similar. For public schools, both instate and out-of-state tuition are lowest among all three ownership type, and the distribution is unimodal and skewed right. For private nonprofit schools, the tuition have larger range, and its distribution is bimodal. Therefore, the tuition differs for students from different ownership type schools. We would recommend public schools for students with limited tuition budget.

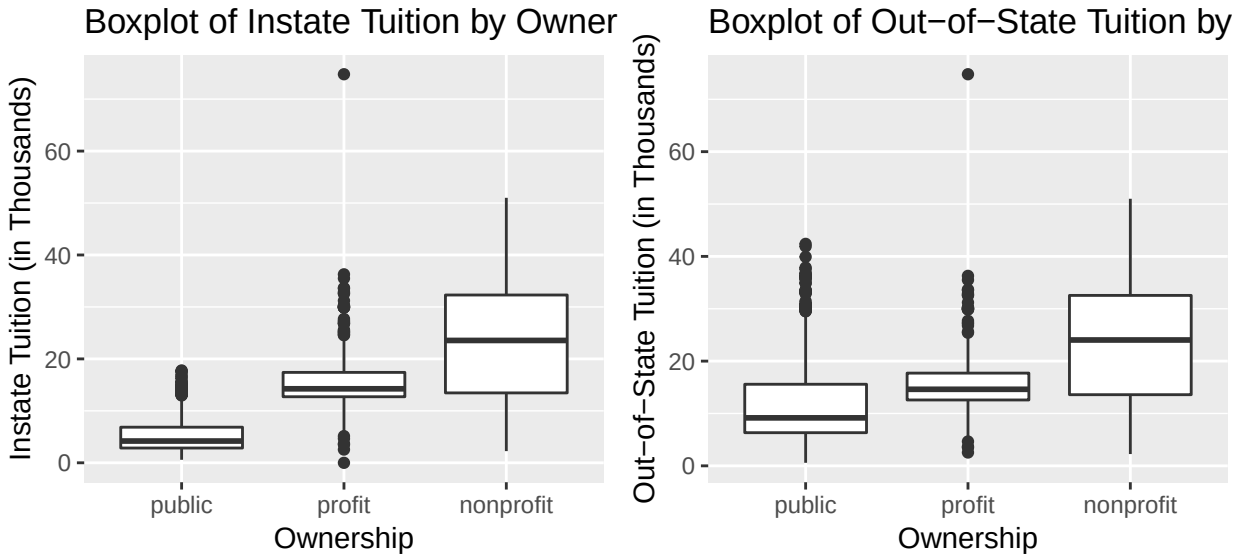
```
p5 <- ggplot(college1, aes(x=tuition_instate))+
  geom_histogram(bins = 30,fill = "lightblue", color = "black")+
  facet_grid(.~ownership)+
  labs(x="Instate Tuition (in Thousands)")+
  ggtitle("Histogram of Instate Tuition by Ownership")
p6 <- ggplot(college2, aes(x=tuition_out))+
  geom_histogram(bins = 30,fill = "lightblue", color = "black")+
  facet_grid(.~ownership)+
  labs(x="Out-of-State Tuition (in Thousands)")+
  ggtitle("Histogram of Out-of-State Tuition by Ownership")
grid.arrange(p5,p6,ncol=1)
```





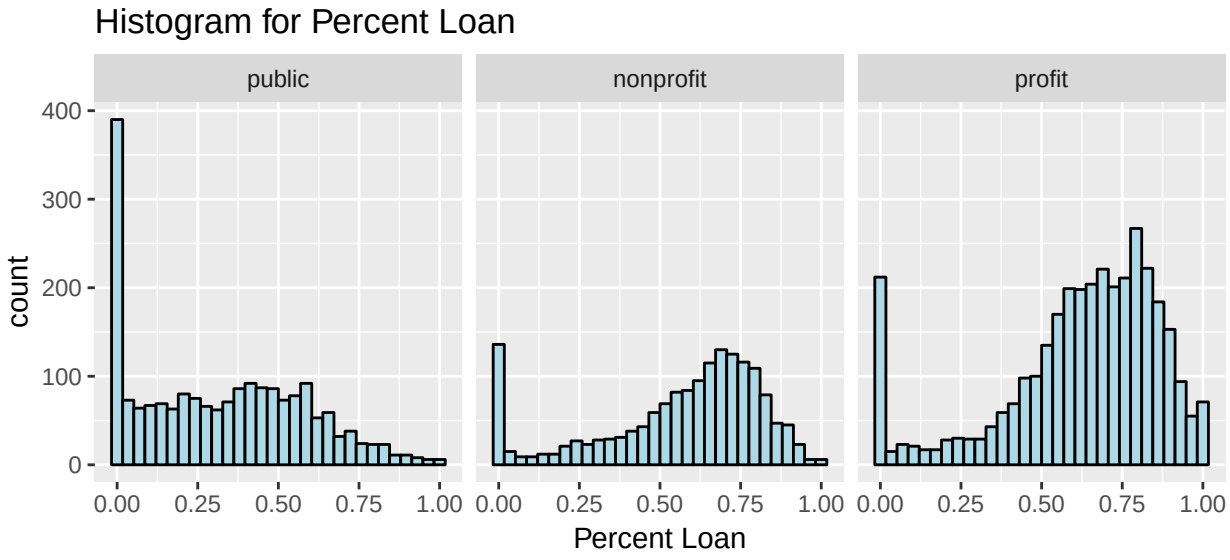
The boxplot for tuition by ownership below gives us a better understanding of the median, as well as outliers of the data. We can see that the median for private nonprofit schools is highest, while that of public is lowest. Moreover, we can clearly see the outliers. The one outlier for private for-private schools with value over 70,000 dollar seems like an unusal observation.

```
p7 <- ggplot(college1, aes(x=reorder(ownership,tuition_instate),y=tuition_instate))+
  geom_boxplot()+
  labs(y="Instate Tuition (in Thousands)",x="Ownership")+
  ggtitle("Boxplot of Instate Tuition by Ownership")
p8 <- ggplot(college2, aes(x=reorder(ownership,tuition_out),y=tuition_out))+
  geom_boxplot()+
  labs(y="Out-of-State Tuition (in Thousands)",x="Ownership")+
  ggtitle("Boxplot of Out-of-State Tuition by Ownership")
grid.arrange(p7,p8,ncol=2)
```



The percent loan variable represents percent of all undergraduate students receiving a federal student loan. By faceting into different ownership groups, the histogram shows that private for-profit schools have higher rate of students receiving a federal loan, while public schools have the lowest. For all ownership types, there are a lot of students receiving no federal loan.

```
college3 =  college[!is.na(college$pct_loan), ]
college3$ownership <- recode[college3$ownership]
college3$ownership<-fct_relevel(college3$ownership,'public','nonprofit','profit')
ggplot(college3, aes(x=pct_loan))+
  geom_histogram(fill = "lightblue", color = "black")+
  labs(x="Percent Loan")+
  ggtitle("Histogram for Percent Loan")+
  facet_grid(.~ownership)
```



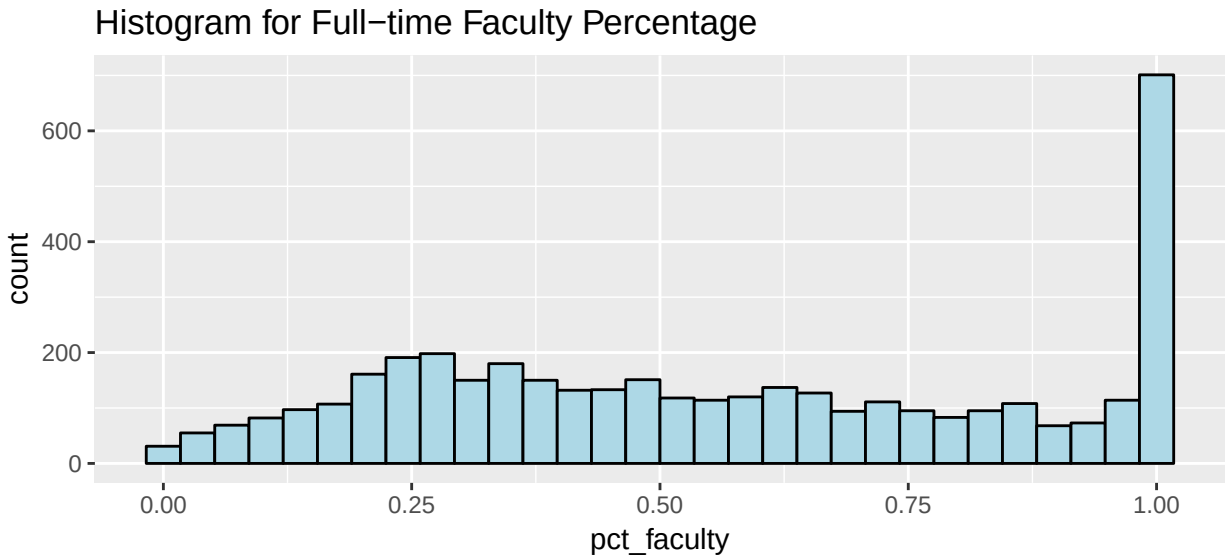
Values

Education Quality

More and more colleges are hiring part-time faculty as a means of cost-cutting. However, we should emphasize the importance of `full-time faculty` as it might be closely related to students' academic performance. For one thing, full-time faculty members are more stable in terms of curriculum teaching and planning. Because of high flexibility, they are also more likely to have higher levels of availability for students after class.

Using the feature `pct_faculty`, *percentage of full-time faculty*, we see that the distribution of percent of full-time faculty is slightly positively skewed with a mode of 1.00. Thus, most colleges have entirely full-time faculty members.

```
library(readr)
library(reshape)

college = read_rds("college.rds")
faculty = college %>% select(name, avg_faculty_salary, pct_faculty, num_undergrad, num_grad)

ggplot(faculty, aes(x = pct_faculty)) +
  geom_histogram(na.rm = TRUE, fill = "lightblue", color = "black")+
  labs(title = "Histogram for Full-time Faculty Percentage")
```

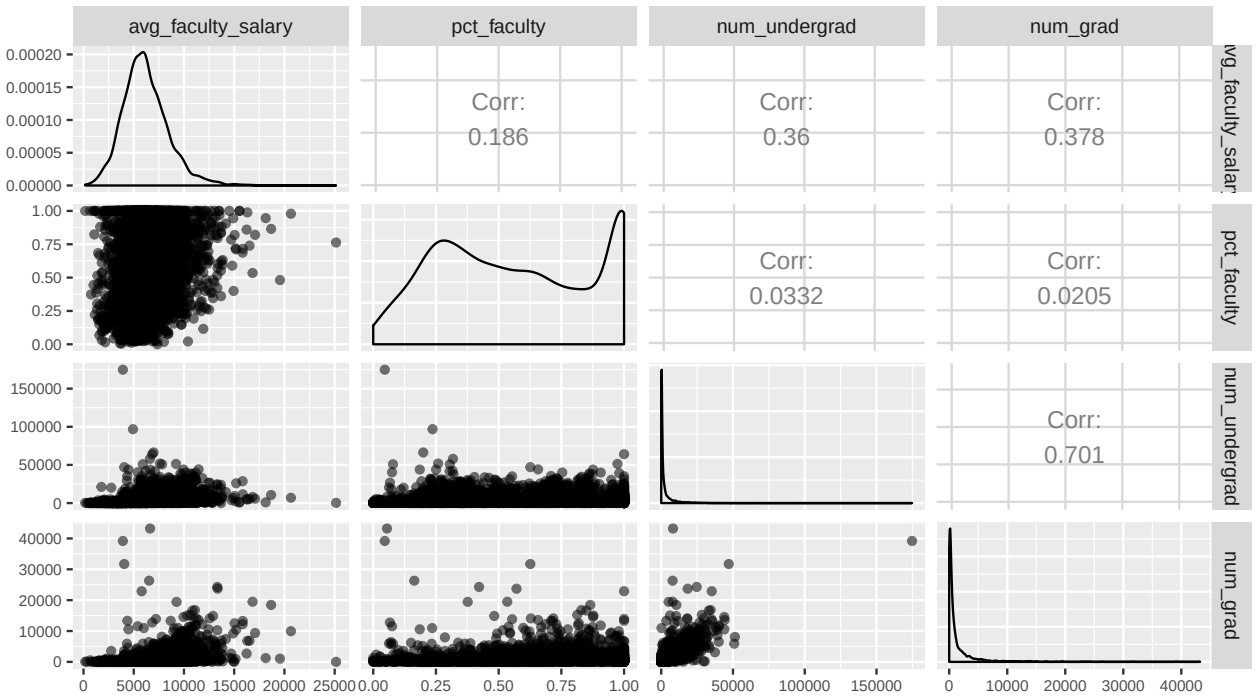


For further analysis of `full-time faculty`, we propose two research questions:

1. Is % of full-time faculty associated with the number of undergraduates/graduates?
2. Is % of full-time faculty associated with mean faculty salary?

To investigate the relationships, we obtained the following scatterplot matrix. We conclude that there is barely any relationship between percent of full-time faculty and other variables we proposed, as the correlations are negligible.

```
GGally::ggpairs(faculty, aes(alpha = 0.01),columns = colnames(faculty %>% select(-name))) +
  theme(axis.text.x = element_text(size = rel(0.8)),
        axis.text.y = element_text(size = rel(0.8)),
        strip.text = element_text(size = rel(0.8)))
```



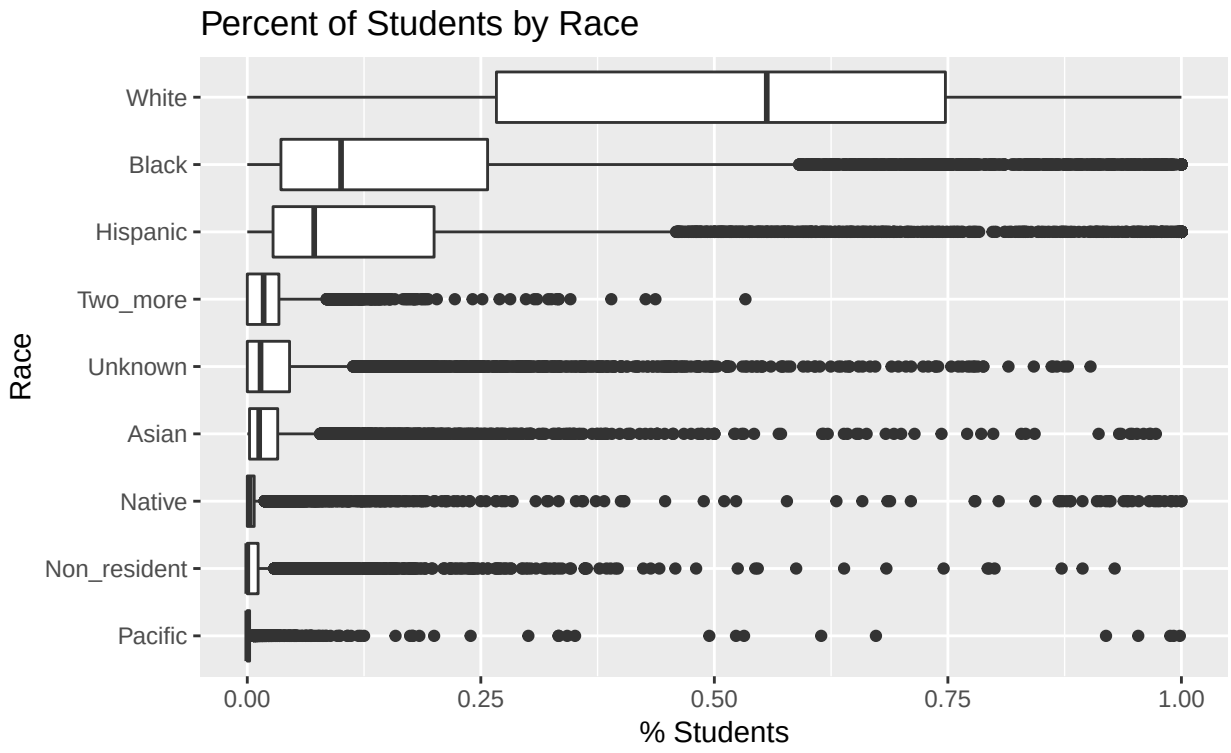
Diversity

1. Racial Diversity

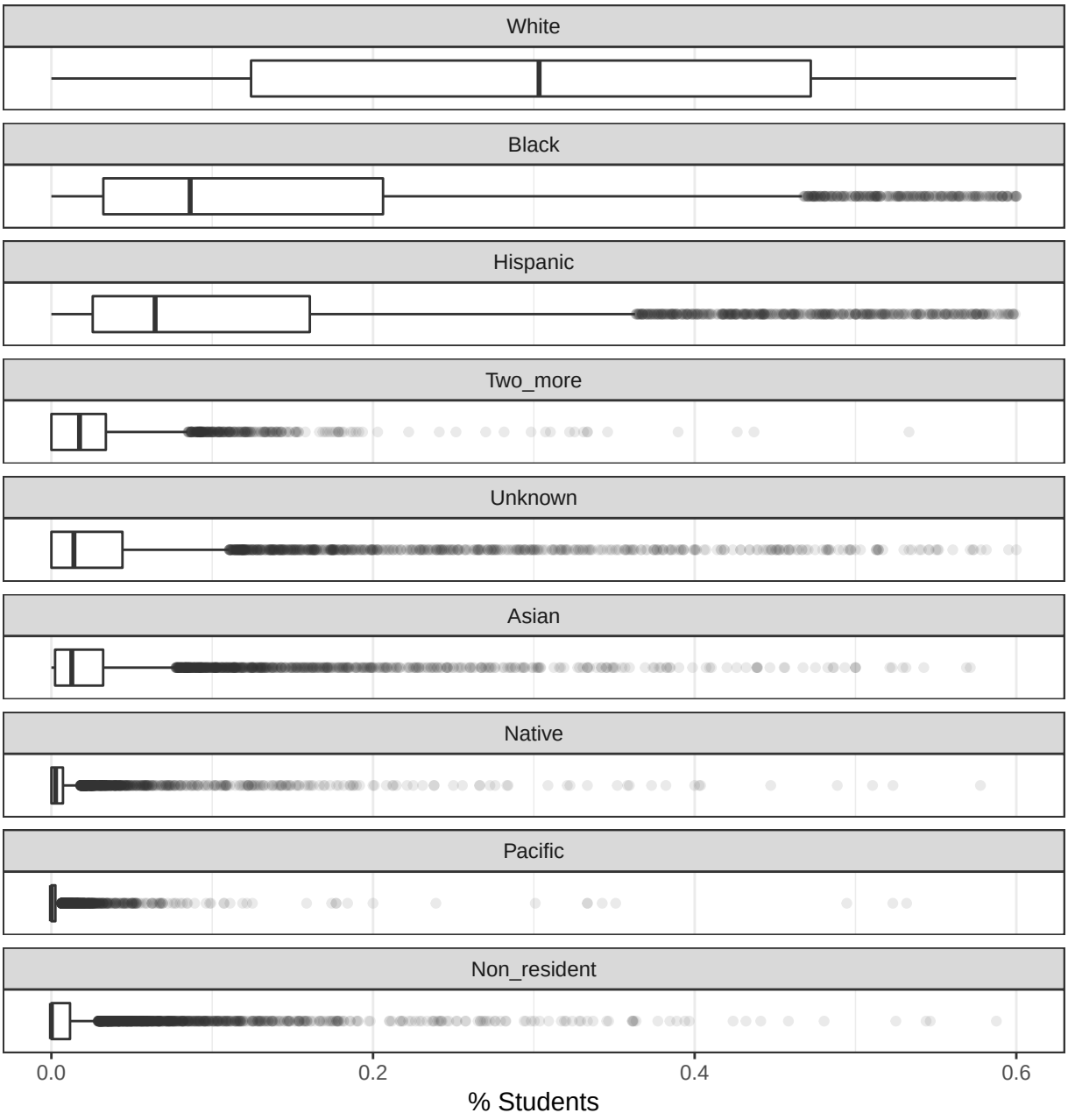
Racial segregation, with its root in the practice of slavery and segregation, is still present today in the United States. Racial separation impedes the establishment of a multicultural and equal society. In particular, it poses challenges for minority students to improve in educational settings, as discrimination, racial biases, and prejudices pervade. On the contrary, racial diversity can benefit students for it fosters collaboration, encourages creativity, and ultimately improves academic achievement. Therefore, we would like to provide an evaluation of racial diversity among colleges.

To begin the analysis, we first explore **how common each race is in colleges**.

```
college = read_rds("college.rds")
race_cat = c("race_white", "race_black", "race_hispanic",
             "race_asian", "race_native", "race_pacific",
             "race_2more", "race_nonresident", "race_unknown")
race = college %>% select ("name", race_cat)
colnames(race) = c("name", "White", "Black", "Hispanic", "Asian",
                   "Native", "Pacific", "Two_more", "Non_resident", "Unknown")
race_melted = melt(data.frame(race), id = "name")
g1 = ggplot(race_melted, aes(x = fct_reorder(`variable`, value, median, na.rm = TRUE),
                             y = value)) + geom_boxplot() + coord_flip() +
  labs(x = "Race", y = "% Students", title = "Percent of Students by Race")
g1
```




```
# Adjust for visualization
race_melted = race_melted %>%
  mutate(`variable` = fct_reorder(`variable`, value, median, na.rm = TRUE, .desc = TRUE))

# use "x = 1" to trick ggplot2 into making a single boxplot per facet
g2 = ggplot(race_melted, aes(x = 1, y = value)) +
  geom_boxplot(outlier.alpha = .1) +
  coord_flip() + facet_wrap(~`variable`, ncol = 1) +
  theme_bw() +
  scale_x_continuous(limits = c(0.4, 1.6)) +
  scale_y_continuous(limits = c(0, 0.6)) +
  # get rid of y-axis
  theme(axis.line.y = element_blank(),
        axis.ticks.y = element_blank(),
        axis.text.y = element_blank(),
        axis.title.y = element_blank(),
        panel.grid.major.y = element_blank(),
        panel.grid.minor.y = element_blank()) +
  labs(x = "Race", y = "% Students", title = "Percent of Students by Race (capped at 0.6)")
g2
```



Comparing the medians in the boxplots we find the between-group variation to be significant. Specifically, `White`, `Black` and `Hispanic` seem to be the most common race while `Pacific` and `Non_resident` are the least. Also, in-group variation is larger for `White`, `Black` and `Hispanic` as well, so that the proportions of these racial groups vary more from college to college. Note that there are a great many outliers in the dataset.

Racial Diversity Index (RDI)

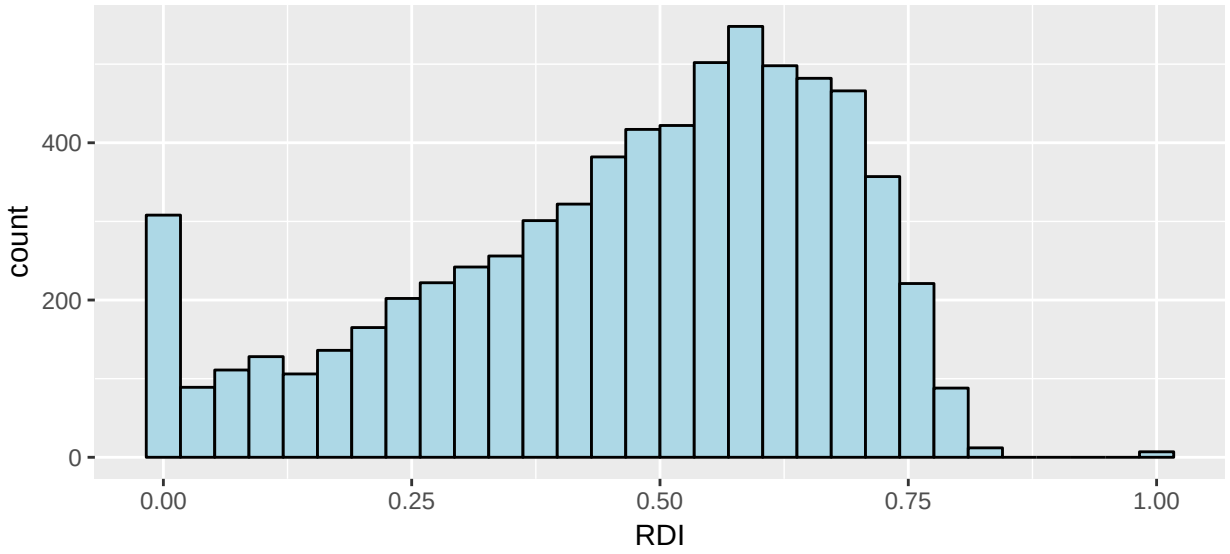
Next, to analyze racial diversity, we introduce *Racial Diversity Index (RDI)*, which is defined by the probability that any two students selected at random would have different races. This index was developed by Missouri

State University (https://diversity.missouristate.edu/DiversityIndex.htm) and is calculated by

$$RDI = 1 - Pr(\text{same race}) = 1 - \sum_i (\text{race i})^2,$$

where *race i* is the percentage of students of *race i*.

```
race <- race %>% mutate(RDI = 1- {(race$White)^2 + (race$Black)^2 +
    (race$Hispanic)^2 + (race$Asian)^2 + (race$Native)^2 + (race$Pacific)^2 +
    (race$Two_more)^2 + (race$Non_resident)^2 + (race$Unknown)^2})

g3 = ggplot(race, aes(x = RDI)) +
  geom_histogram(na.rm = TRUE, fill = "lightblue", color = "black")

g3
```



An RDI close to 1 implies high racial diversity while an RDI close to 0 implies low. From the histogram, we observe that the distribution of `RDI` is slightly negatively skewed, so that more colleges have moderate to high `RDI`'s (over 0.50). The mode is approximately 0.6 so that most colleges have moderately high racial diversity.

Note that there are a few colleges scoring 0.00 (*Table 1*), as all of their students are from one exclusive race. For instance, all students of *Thunderbird School of Global Management* are *Hispanic*. At the other tail, note the isolated bar for `RDI` = 1.00 , which is impractical as no college would have exactly even distribution of races. A closer look at the data reveals that the `race` features for these colleges are all 0, indicating that their `population` might not have been recorded properly (*Table 2*). Excluding these colleges, we find the most racially diverse colleges listed in *Table 3* below.

```
library(knitr)
library(kableExtra)
kable(head(race %>% filter(RDI == 0)), format = "latex",
             caption = "Colleges with RDI = 0") %>%
  kable_styling(latex_options="scale_down")
```

```
kable(head(race %>% filter(RDI == 1)), format = "latex",
             caption = "Colleges with RDI = 1") %>%
  kable_styling(latex_options="scale_down")
```

```
kable(head(race %>% filter(RDI != 1) %>% arrange(-RDI) %>% select(name, RDI)),
      format = "latex",
      caption = "Colleges with Highest Racial Diversity")
```

Table 1: Colleges with RDI = 0

name	White	Black	Hispanic	Asian	Native	Pacific	Two_more	Non_resident	Unknown	RDI
Thunderbird School of Global Management	0	0	1	0	0	0	0	0	0	0
Velvatex College of Beauty Culture	0	1	0	0	0	0	0	0	0	0
Universal College of Beauty Inc-Los Angeles 1	0	1	0	0	0	0	0	0	0	0
Carver Bible College	0	1	0	0	0	0	0	0	0	0
Mr Leon's School of Hair Design-Moscow	1	0	0	0	0	0	0	0	0	0
Hebrew Theological College	1	0	0	0	0	0	0	0	0	0

Table 2: Colleges with RDI = 1

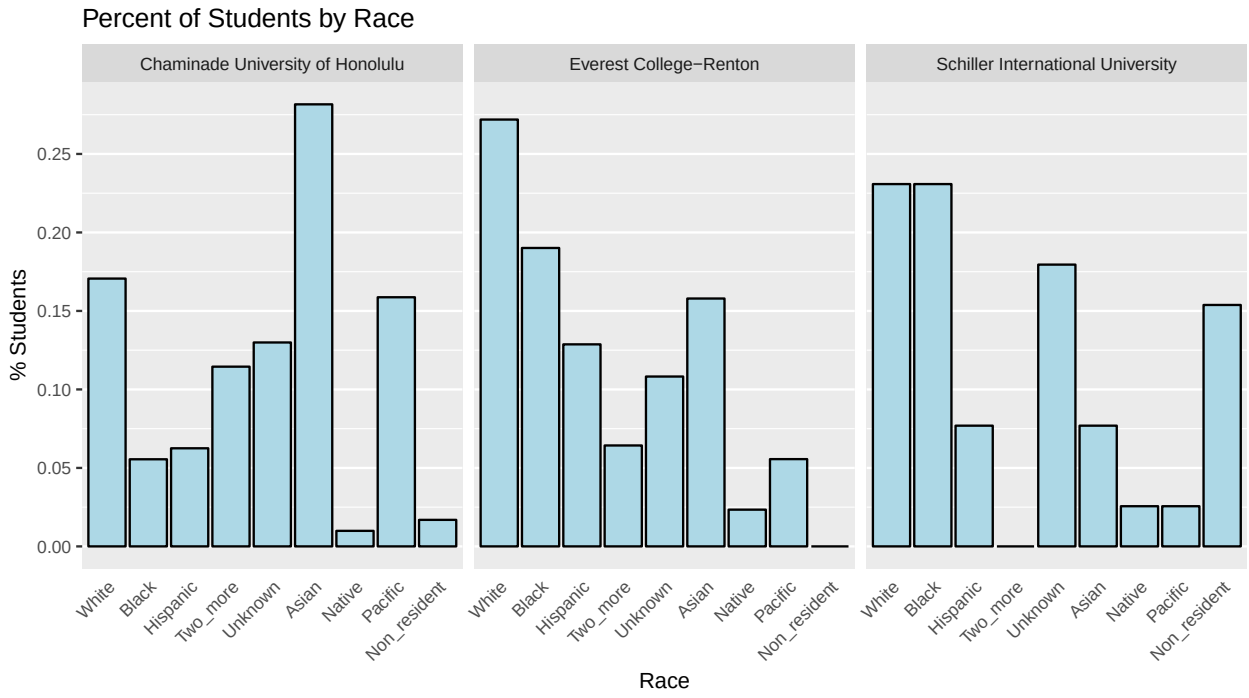
name	White	Black	Hispanic	Asian	Native	Pacific	Two_more	Non_resident	Unknown	RDI
American Conservatory Theater	0	0	0	0	0	0	0	0	0	1
Prince Institute-Rocky Mountains	0	0	0	0	0	0	0	0	0	1
Lyme Academy College of Fine Arts	0	0	0	0	0	0	0	0	0	1
Professional Business College	0	0	0	0	0	0	0	0	0	1
Spanish-American Institute	0	0	0	0	0	0	0	0	0	1
Taft University System	0	0	0	0	0	0	0	0	0	1

Finally, taking a look at the make-ups of the three most racially diverse colleges, we can see that the proportions of races can still vary a lot within each college. For instance, at *Chaminade University of Honolulu*, almost one-third of students are `Asian` but only few are `Native` and `Non-resident`.

```
top3 = c("Chaminade University of Honolulu","Everest College-Renton",
         "Schiller International University")

race_top3 = race_melted %>% filter(name %in% top3)

g4 = ggplot(race_top3, aes(x = variable, y = value)) +
  geom_bar(stat="identity", position = position_dodge(),
           fill = "lightblue", color = "black") + facet_grid(.~ name) +
  labs(title = "Percent of Students by Race", x = "Race", y = "% Students") +
  theme(axis.ticks.x = element_blank(),
        axis.text.x = element_text(angle = 45, hjust = 1),
        panel.grid.major.x = element_blank(),
        text = element_text(size = 10)) +
  scale_y_continuous(breaks = seq(0, 0.4, 0.05))
#Remark. Some questions for further exploration: is *racial diversity* correlated to
#other variables such as `ownership`, `predominant_Degree` and `highest_degree`?

g4
```

Table 3: Colleges with Highest Racial Diversity

name	RDI
Chaminade University of Honolulu	0.8290570
Everest College-Renton	0.8289556
Schiller International University	0.8244501
Pacific Rim Christian University	0.8215979
Golden Gate University-San Francisco	0.8205131
Charles A Jones Career and Education Center	0.8203344

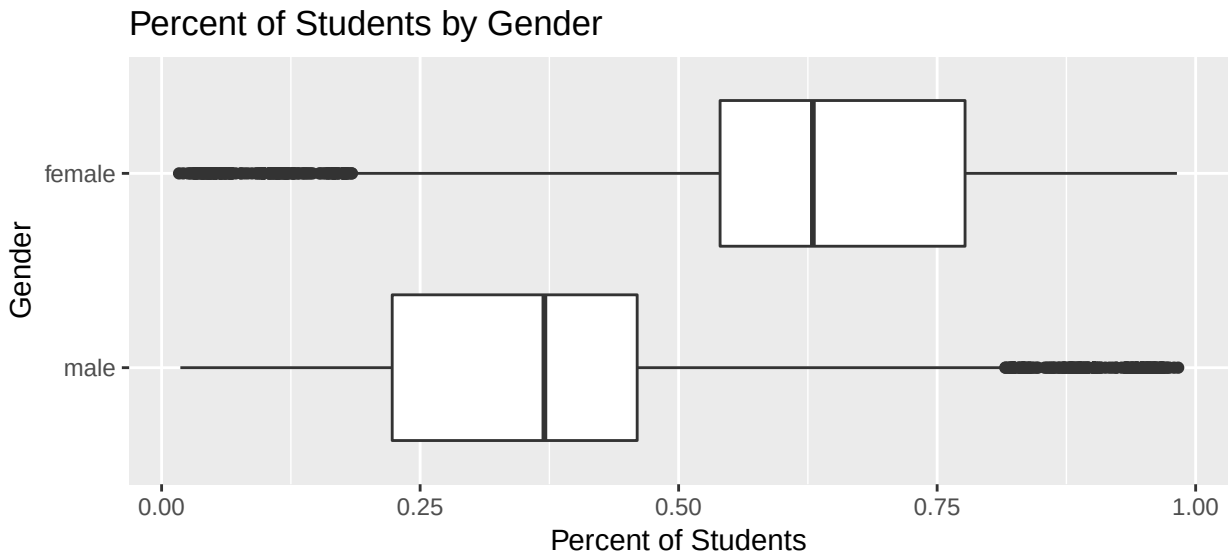


2. Gender Diversity

Gender equality has been an issue in the workplace, where women are known to have struggled for parity in pay and influence. To form awareness of gender equality and reduce challenges in workplace in the future, more and more colleges have been working on promoting gender diversity.

To investigate gender diversity, we first explore **how common each gender is in colleges**. Comparing the medians in the boxplot we see that most colleges seem to have more female students than male.

```
gender = college %>% select(name, pct_female) %>%
  mutate(pct_male = 1-pct_female)
colnames(gender) = c("name", "female", "male")
gender_melted = melt(data.frame(gender), id = "name")
g5 = ggplot(gender_melted, aes(x = fct_reorder(`variable`, value, median, na.rm = TRUE),
                               y = value)) + geom_boxplot() + coord_flip() +
  labs(x = "Gender", y = "Percent of Students", title = "Percent of Students by Gender")
g5
```



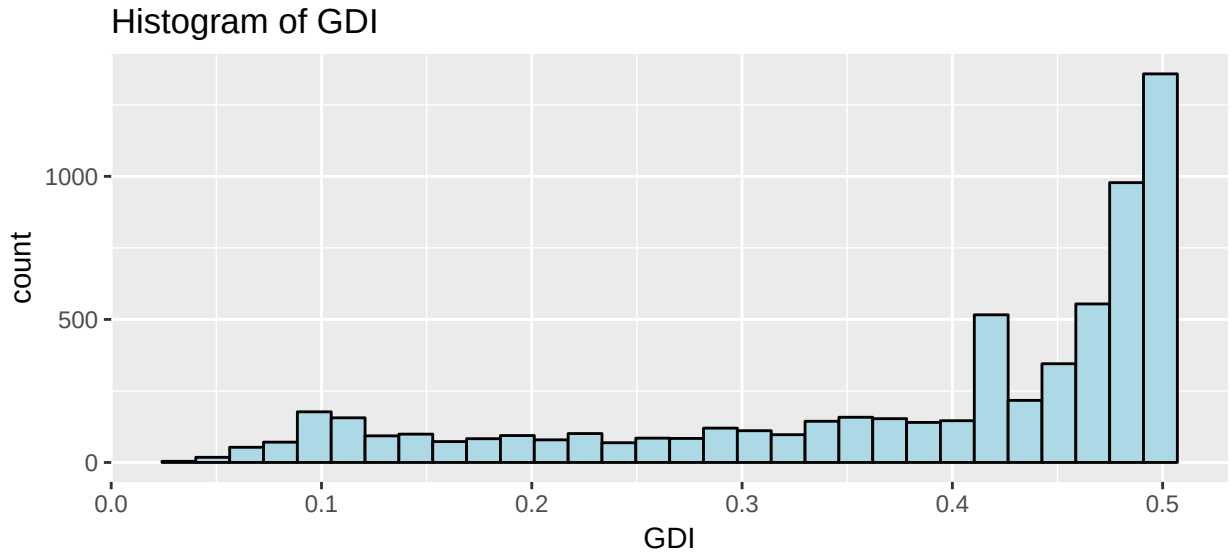
Gender Diversity Index (GDI)

To give a substantial measurement of gender diversity, we develop the *Gender Diversity Index (GDI)*. Like **racial diversity**, it is defined by the probability that any two students selected at random would have different genders:

$$GDI = 1 - Pr(\text{same gender}) = 1 - (\text{male})^2 - (\text{female})^2,$$

where *male* and *female* are percent of male and female students.

```
gender = gender %>% mutate(GDI = 1-female^2-male^2)
ggplot(gender, aes(x = GDI)) +
  geom_histogram(na.rm = TRUE, fill = "lightblue", color = "black")+
  labs(title = "Histogram of GDI")
```



A GDI close to 0.5 implies high racial diversity (i.e., the percentage of male students are close to percentage of female students) while a GDI close to 0 implies low racial diversity. From the histogram, we observe that the distribution of `GDI` is negatively skewed with a long left tail. Note that the mode is 0.5, so that most colleges have equal % of female and male students.

Table 4: Colleges with Highest Gender Diversity

name	female	male	GDI
VanderCook College of Music	0.5000000	0.5000000	0.5
South Carolina State University	0.5000000	0.5000000	0.5
American Institute of Beauty	0.5000000	0.5000000	0.5
SOLEX College	0.5000000	0.5000000	0.5
Texas State Technical College-Harlingen	0.5001274	0.4998726	0.5
University of Idaho	0.5001372	0.4998628	0.5

Finally, the colleges with the highest gender diversity can be found in Table 4.

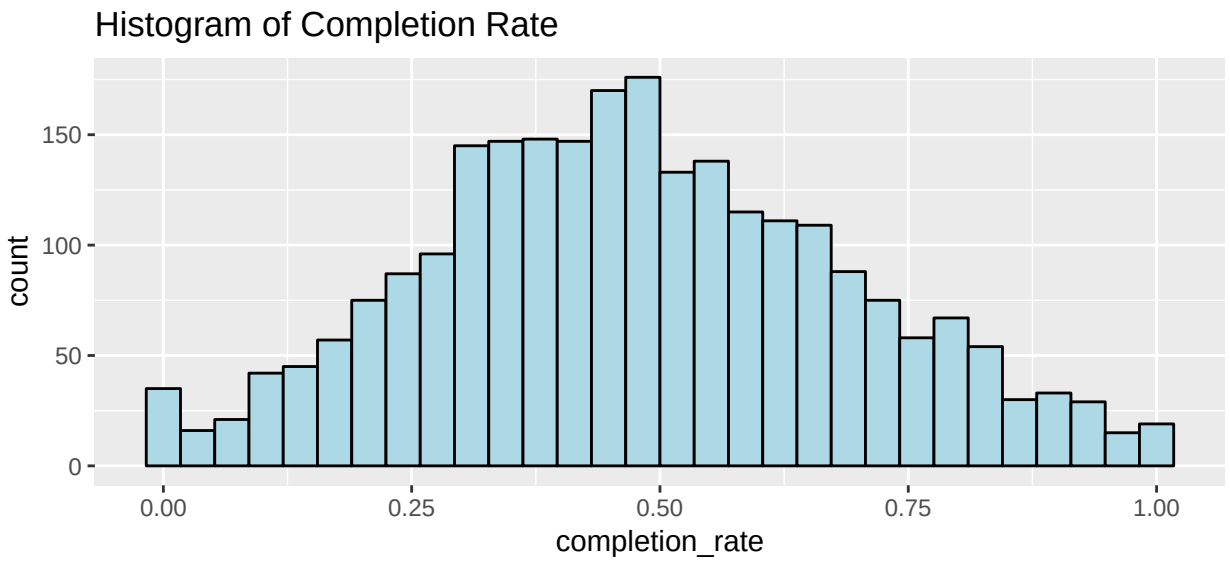
```
kable(head(na.omit(gender) %>% arrange(-GDI)),
      caption = "Colleges with Highest Gender Diversity")
```

Completion

1. Completion Rate

`Completion rate` is recorded as *percentage of students that graduated within 6 years* in our dataset. The overall distribution of completion rate is roughly symmetric, with mode around 0.47. Thus, for most colleges, just under half (47%) of students graduated within 6 years of starting college.

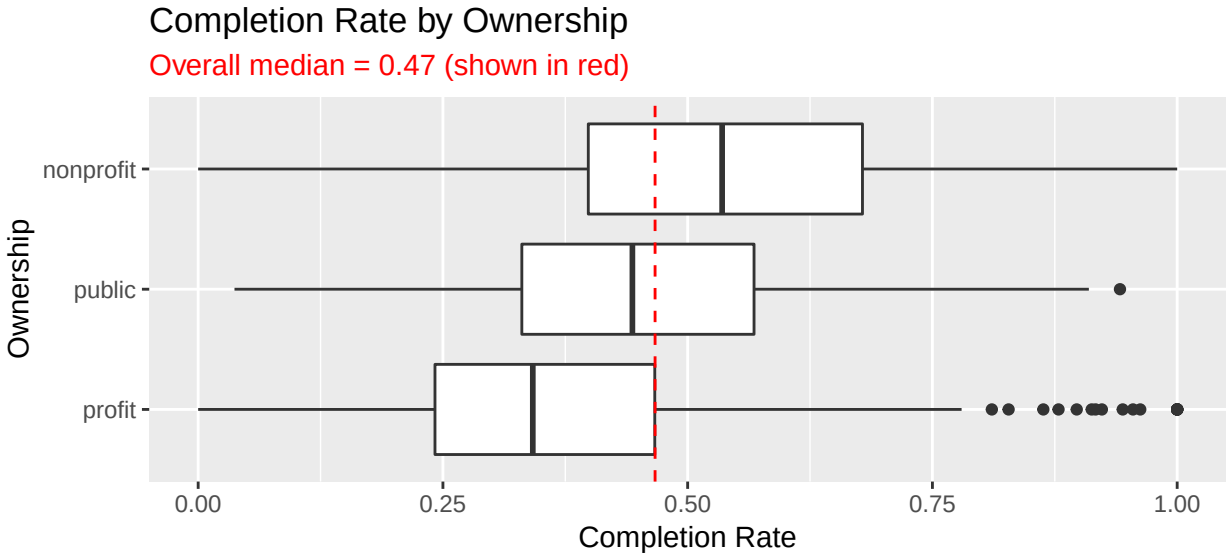
```
college = read_rds("college.rds")

ggplot(college, aes(x = completion_rate)) +
  geom_histogram(na.rm = TRUE, fill = "lightblue", color = "black")+
  labs(title = "Histogram of Completion Rate")
```



Completion Rate by Institution Type

Taking institution type into consideration, we can see that private *non-profit* schools have the best completion rates (highest median) while private *for-profit* schools have the worst. Only non-profit schools have higher median completion rate than the overall median. Note that unlike *non-profit* and *for-profit* schools, public schools have neither a completion rate of 0.00 or 1.00.

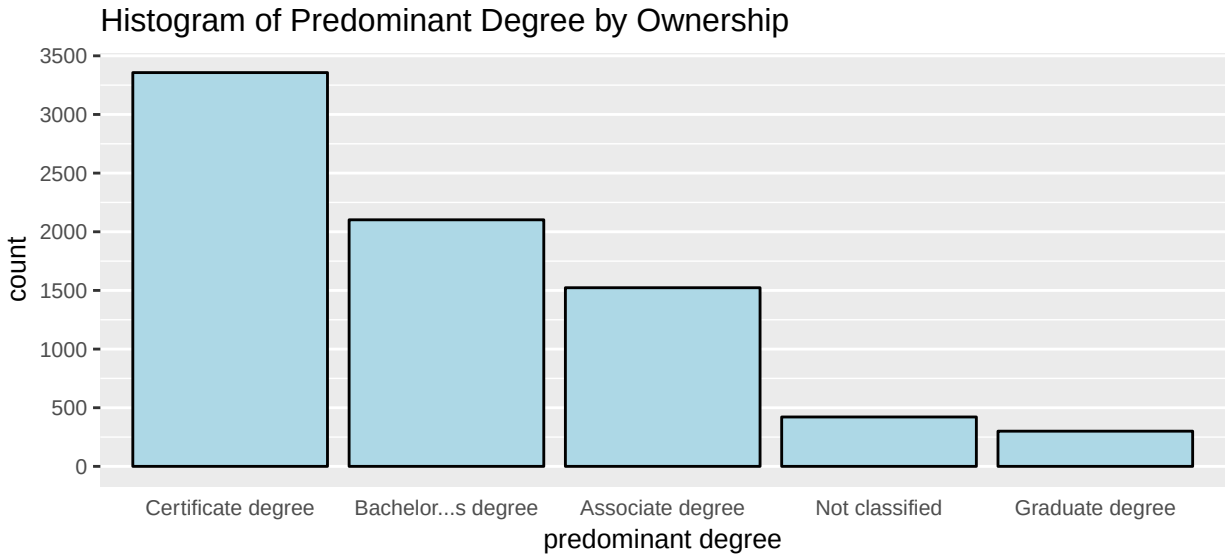
```r
comp_rate = college %>% select(name, ownership, completion_rate)
levels(comp_rate$ownership) = c("public","nonprofit","profit")

med = median(comp_rate$completion_rate, na.rm = TRUE)

ggplot(comp_rate, aes(x = fct_reorder(`ownership`, completion_rate, median, na.rm = TRUE),
                      y = completion_rate)) + geom_boxplot() + coord_flip() +
  geom_hline(yintercept = med, color = "red", linetype = 2) +
  labs(x = "Ownership", y = "Completion Rate", title = "Completion Rate by Ownership",
       subtitle = paste0("Overall median = ", round(med,2), " (shown in red)")) +
  theme(plot.subtitle = element_text(color = "red"))
```



2. Degree Awarded

Overall, *Certificate degree* is the most frequently awarded while *Graduate degree* is the least. Such a result is reasonable as certificate programs are less demanding than degree programs. Also, since graduate population is usually smaller than undergraduate population, we shall expect that fewer students receive a *Graduate degree*.

```r
degree = college %>% select(name, ownership, predominant_degree, highest_degree)
levels(degree$ownership) = c("public","nonprofit","profit")
levels(degree$predominant_degree) = c("Not classified", "Certificate degree", "Associate degree", "Bache
levels(degree$highest_degree) = c("Non-degree-granting", "Certificate degree", "Associate degree","Bache

gc3 = ggplot(degree, aes(x = fct_reorder(predominant_degree, predominant_degree, function(x){length(x)}
  geom_bar(fill = "lightblue", color = "black") +
  labs(x = "predominant degree") +
  theme(axis.ticks.x = element_blank(),
        panel.grid.major.x = element_blank(),
        text = element_text(size = 10)) +
  scale_y_continuous(breaks = seq(0, 3500, 500))+
  labs(title = "Histogram of Predominant Degree by Ownership")
gc3
```



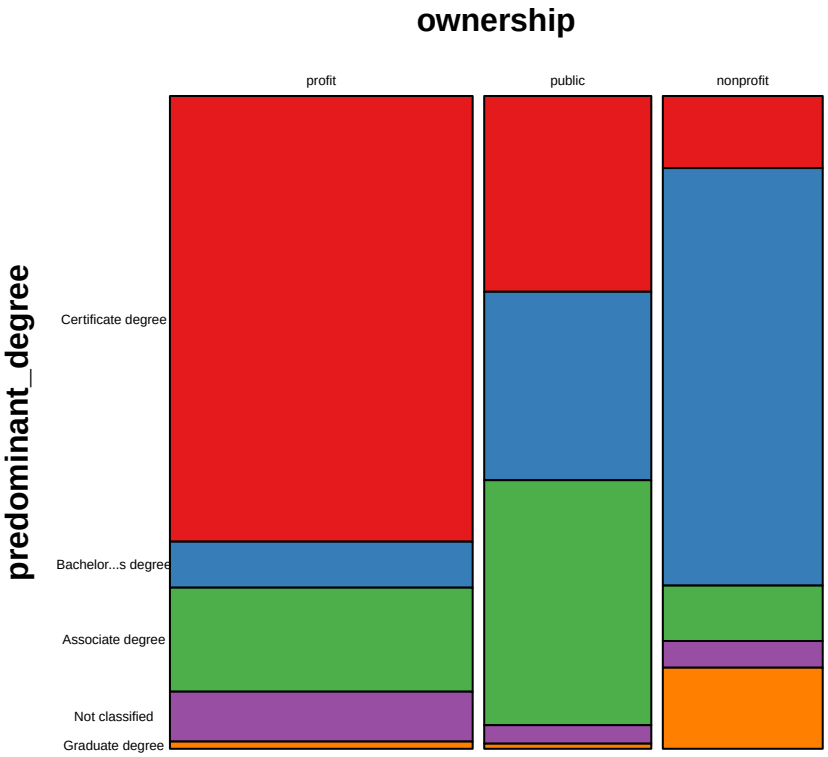
Predominant Degree by Ownership

Predominant degree seems to be associated with `ownership` as the boxes in the mosaic plot below are not proportional. Specifically, we can infer that the predominant degree awarded is mostly

- *Certificate degree* for private *for-profit* schools;
- *Associate degree* for *public* schools;
- *Bachelor's degree* for private *non-profit* schools.

```
library(grid)
library(vcd)
library(RColorBrewer)
fillcolors = brewer.pal(5, "Set1")

# Sort predominant_degree by frequency
count1 = degree %>% group_by(ownership, predominant_degree) %>% summarize(Freq = n())
pred_deg_order = count1 %>% group_by(predominant_degree) %>%
  summarize(sum = sum(Freq)) %>% arrange(-sum) %>%
  pull(predominant_degree)
count1$predominant_degree = factor(count1$predominant_degree, levels = pred_deg_order)

# Sort ownership by frequency
ownership_order = count1 %>% group_by(ownership) %>%
  summarize(sum = sum(Freq)) %>% arrange(-sum) %>%
  pull(ownership)
count1$ownership = factor(count1$ownership, levels = ownership_order)

gc4 = mosaic(predominant_degree~ownership, count1, direction = c("v", "h"),
             gp = gpar(fill = fillcolors),
             labeling = labeling_border(gp_labels = gpar(fontsize = 5),
                                        rot_labels = c(0, 0, 0, 0),
                                        offset_varnames = c(0,0,0,2),
                                        offset_labels = c(0,0,0,2.5)),
             main = "Predominant Degree vs. Ownership")
```

Predominant Degree vs. Ownership




```
gc4
```

```
##                    ownership profit public nonprofit
## predominant_degree
## Certificate degree              2528    613       216
## Bachelor's degree                261    590      1251
## Associate degree                 590    767       166
## Not classified                   283     58        80
## Graduate degree                   41     16       243
```

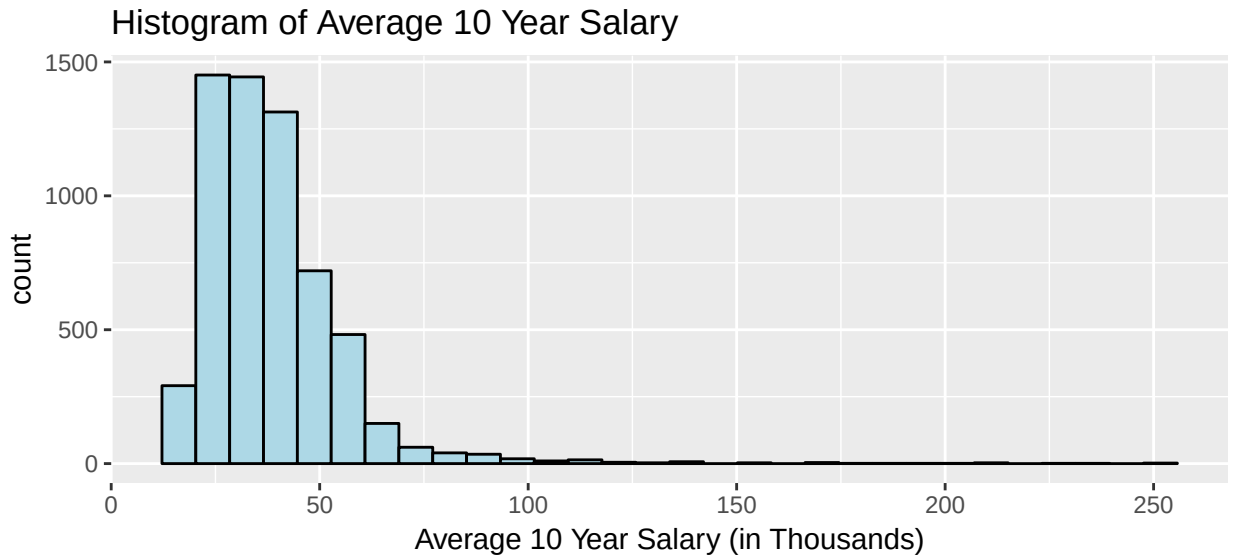
Earnings

Although earnings are not the only criteria to see how success a person is, it is a very straightforward indication of performance after graduation, so we also examined earnings after graduation as a part of value a college provides.

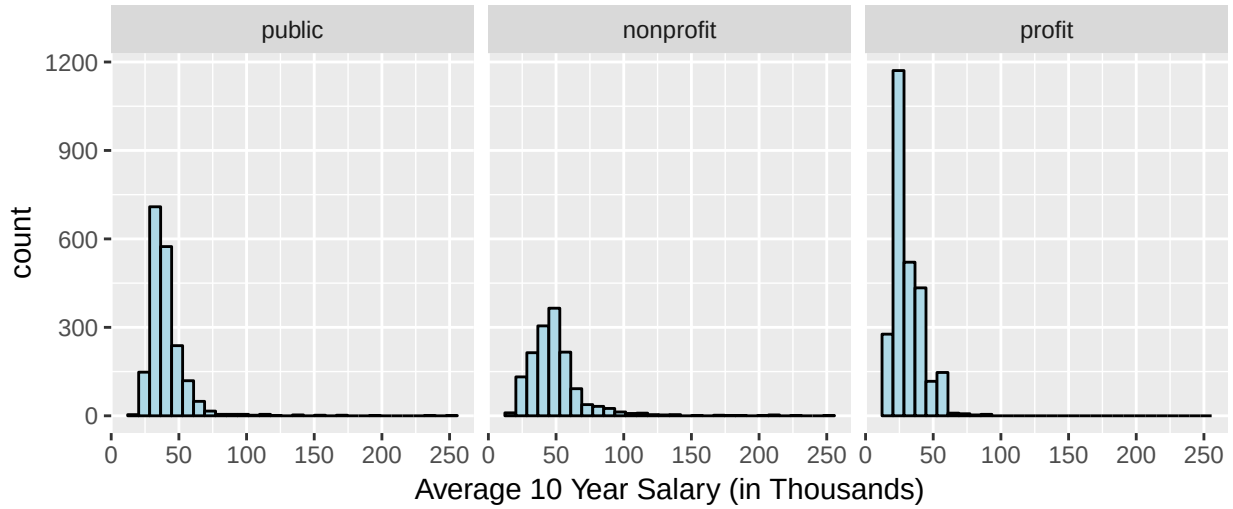
The `average 10 year salary` variable represents the mean earnings of students working and not enrolled at school 10 years after entry. The overall distribution of the average 10 year salary is skewed to the right. Most of our observations have an average 10 year salary of around 20,000 to 40,000 dollars. There are also a couple observations with higher average 10 year salary than 100,000 dollars.

```
college = read_rds("college.rds")
college[college=='NULL'] <- NA
#college = na.omit(college$avg_10yr_salary)
college =  college[!is.na(college$avg_10yr_salary), ]
college$avg_10yr_salary = college$avg_10yr_salary/1000
college$tuition_instate = college$tuition_instate/1000
college$tuition_out = college$tuition_out/1000
```

```
recode <- c('1'='public','2'='nonprofit','3'='profit')
college$ownership <- recode[college$ownership]
college$ownership<-fct_relevel(college$ownership,'public','nonprofit','profit')
ggplot(college, aes(x=avg_10yr_salary))+
  geom_histogram(bins = 30,fill = "lightblue", color = "black")+
  labs(x="Average 10 Year Salary (in Thousands)")+
  ggtitle("Histogram of Average 10 Year Salary")
```




```
a=count(college[ which(college$ownership=='public'),])
b=count(college[ which(college$ownership=='nonprofit'),])
c=count(college[ which(college$ownership=='profit'),])
ggplot(college, aes(x=avg_10yr_salary))+
  geom_histogram(bins = 30,fill = "lightblue", color = "black")+
  facet_grid(.~ownership)+
  labs(x="Average 10 Year Salary (in Thousands)")+
  ggtitle("Histogram of Average 10 Year Salary by Ownership")
```

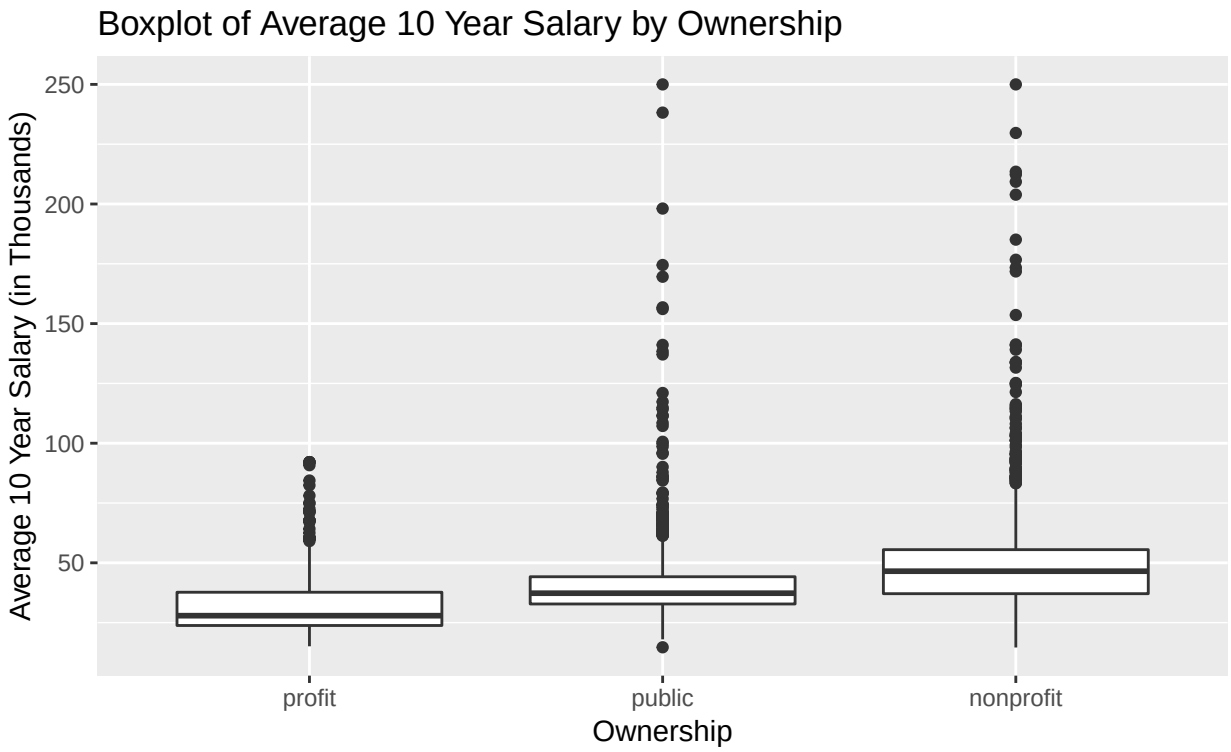


By faceting average 10 year salary into different ownership group, we can see the different distribution for

public, private nonprofit, and private for-profit schools. The total numbers of observations for the three ownership type are 1890, 1481, 2691. Among the three types, although private nonprofit schools have the least observations, they have the distribution most skewed right, with highest mean and mode. This indicates students from private nonprofit schools have highest average 10 year salary among the three school types. While public schools have the mode for average 10 year salary at around 30,000 dollars, private for-profit schools have the mode at around 20,000 dollars, and overall, students from public schools rank the second in average 10 year salary. Therefore, the average 10 year salary differs for students from different ownership type schools.

The boxplot below for average 10 year salary gives us a better understanding of the median, as well as outliers of the data. We can see that the median for private nonprofit schools is highest, while that of private for-profit ones is lowest. Moreover, we can clearly see the outliers. Comparing to other ownership types, private for-profit schools have the least number of outliers, and the value of outlier is lowest (no outliers with value over 100,000). The boxplot also indicates students from private nonprofit schools have higher average 10 year salary than those from public and private for-profit schools.

```
ggplot(college, aes(x=reorder(ownership,avg_10yr_salary),y=avg_10yr_salary))+
  geom_boxplot()+
  labs(y="Average 10 Year Salary (in Thousands)",x="Ownership")+
  ggtitle("Boxplot of Average 10 Year Salary by Ownership")
```



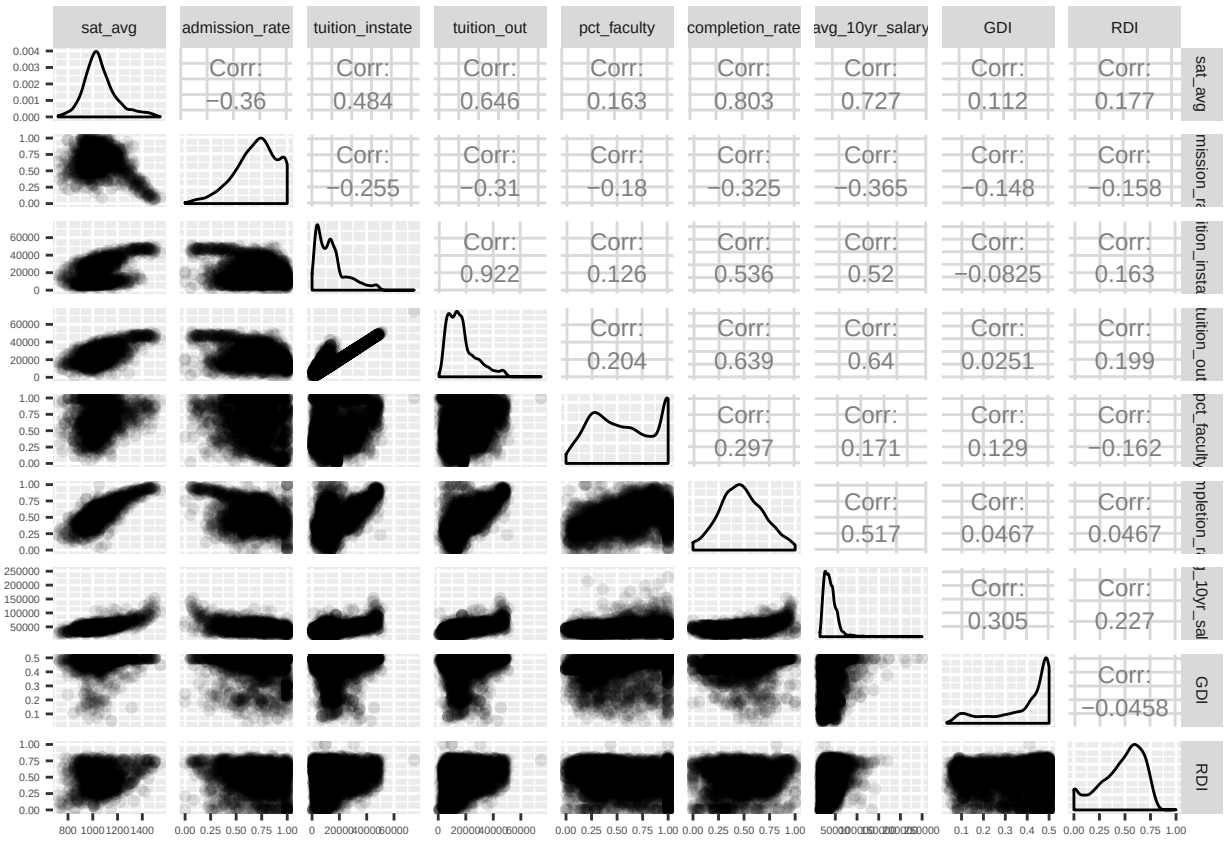
Cost-Value Tradeoffs

Eventually, to explore the cost-value trade-off, we obtained the following scatterplot matrix. We conclude the correlation to be *strong*, if $r \geq 0.7$; *weak*, if $r \leq 0.3$; *moderate* otherwise. Some key observations are:

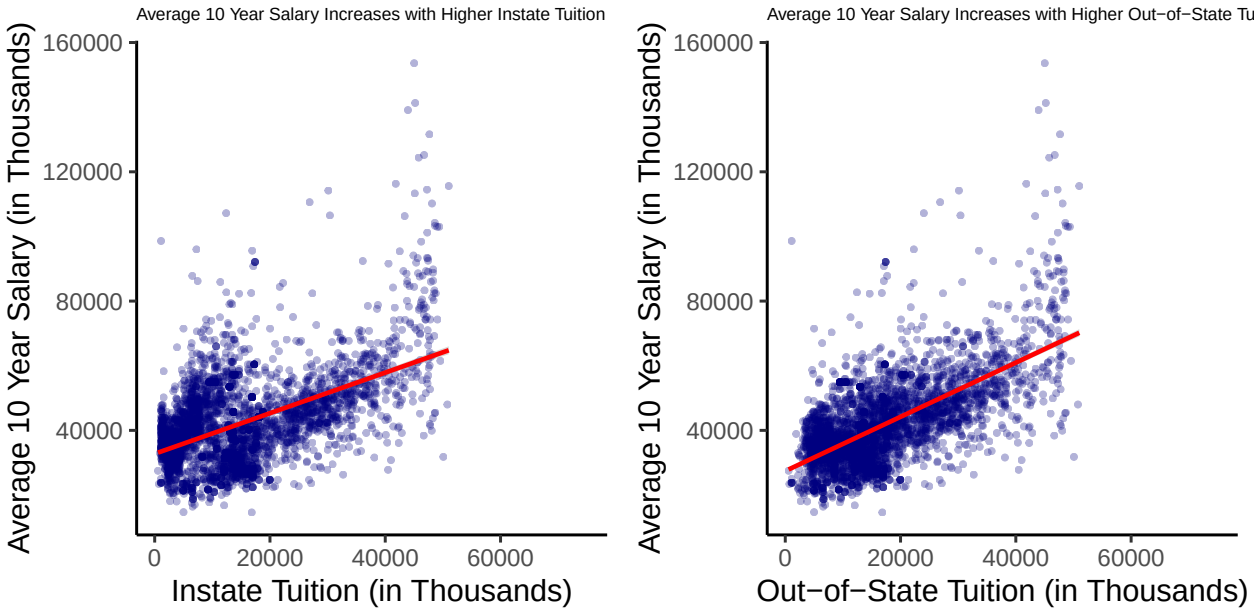
- **Multicollinearity** exists for both the cost and value variables such that
  - `sat_avg` is moderately positively associated with `tuition_instate`/`tuition_out`
  - `sat_avg` is moderately negatively associated with `admission_rate`
  - `tuition_instate` is strongly positively associated with `tuition_out`
  - `completion_rate` is moderately positively associated with `avg_10yr_salary`

- `sat_avg` is strongly positively associated with both `completion_rate` and `avg_10yr_salary`
- `admission rate` is moderately negatively associated with both `completion_rate` and `avg_10yr_salary`
- `tuition_instate` is moderately positively associated with both `completion_rate` and `avg_10yr_salary`
- `tuition_out` is moderately positively associated with both `completion_rate` and `avg_10yr_salary`

```
library(GGally)
college = read_rds("college.rds")

##### For individual file: duplicate code with Diversity.Rmd
race_cat = c("race_white", "race_black", "race_hispanic",
             "race_asian", "race_native", "race_pacific",
             "race_2more", "race_nonresident", "race_unknown")
race = college %>% select ("name", race_cat)
colnames(race) = c("name", "White", "Black", "Hispanic", "Asian",
                   "Native", "Pacific", "Two_more", "Non_resident", "Unknown")
race = race %>% mutate(RDI = 1- {(race$White)^2 + (race$Black)^2 +
    (race$Hispanic)^2 + (race$Asian)^2 + (race$Native)^2 + (race$Pacific)^2 +
    (race$Two_more)^2 + (race$Non_resident)^2 + (race$Unknown)^2})
gender = college %>% select(name, pct_female) %>%
  mutate(pct_male = 1-pct_female)
colnames(gender) = c("name", "female", "male")
gender = gender %>% mutate(GDI = 1-female^2-male^2)
#####

cost_value = college %>%
  select("sat_avg","admission_rate","tuition_instate",
         "tuition_out","pct_faculty", "completion_rate", "avg_10yr_salary") %>%
  mutate(GDI = gender$GDI) %>%
  mutate(RDI = race$RDI)

cost = c("sat_avg", "admission_rate", "tuition_instate", "tuition_out")
value = c("pct_faculty", "completion_rate", "avg_10yr_salary", "GDI", "RDI")

ggpairs(cost_value,
                upper = list(continuous = wrap("cor", size = 3, alignPercent = 1)),
                lower = list(continuous = wrap("points",alpha = 0.1))) +
  theme(axis.text.x = element_text(size = rel(0.5)),
        axis.text.y = element_text(size = rel(0.5)),
        strip.text = element_text(size = rel(0.5)))
```




```
college =  college[!is.na(college$tuition_out), ]
p1 <- ggplot(college, aes(x=tuition_instate,y=avg_10yr_salary))+
  geom_point(alpha = .3, color = "navy", stroke = 0)+
  labs(x="Instate Tuition (in Thousands)",y="Average 10 Year Salary (in Thousands)")+
  ggtitle("Average 10 Year Salary Increases with Higher Instate Tuition")+
  geom_smooth(method='lm',formula=y~x,color="red")+
  theme_classic(14)+
  theme(plot.title = element_text(size = rel(0.6)))
p2 <- ggplot(college, aes(x=tuition_out,y=avg_10yr_salary))+
  geom_point(alpha = .3, color = "navy", stroke = 0)+
  labs(x="Out-of-State Tuition (in Thousands)",y="Average 10 Year Salary (in Thousands)")+
  ggtitle("Average 10 Year Salary Increases with Higher Out-of-State Tuition")+
  geom_smooth(method='lm',formula=y~x,color="red")+
  theme_classic(14)+
  theme(plot.title = element_text(size = rel(0.6)))
r2_1 = format(summary(lm(college$avg_10yr_salary~college$tuition_instate))$r.squared*100, digits = 3)
r2_2 = format(summary(lm(college$avg_10yr_salary~college$tuition_out))$r.squared*100, digits = 3)
grid.arrange(p1,p2,ncol=2)
```



We can see from the scatter plot that there is some positive correlation between both instate, out-of-state tuition and average 10 year salary. The more expensive the tuition is, the higher the average 10 year salary is. 28, 40.9 percent of the variation in average 10 year salary is explained by the variation in predictor instate tuition, and out-of-state tuition, respectively.

EXECUTIVE SUMMARY

Our project provides analysis of different colleges to help high school students and their family to make informed decisions for college selection based on **cost-value** tradeoffs, where **cost** includes (1) *admission requirement* and (2) *affordability*, and **value** includes (1) *education quality*, (2) *diversity*, (3) *completion* and (4) *earnings*. The major findings for each factor and the relationship between costs and values will be discussed below.

Admission requirement. Admission requirement is assessed by **SAT score** and **admission rate**. Importantly, for schools with average SAT below 1200, admission rate doesn't seem to have strong relationship with average SAT scores. On the other hand, for those with average SAT higher than 1200, admission rate appears to decrease with higher SAT score. Such academically demanding schools may have an explicit requirement for SAT score for admission.

Affordability. Affordability is analyzed through **tuition** (out-of-state/instate) and **federal loan**. Generally, while out-of-state tuition is higher than instate tuition, schools with higher out-of-state tuition tend to have higher instate tuition. The distributions for out-of-state and instate tuition are similar across ownership groups. Public schools have the lowest tuition whereas private non-profit schools have the highest with larger variation. For federal loan, while private for-profit schools have the highest percentage of all undergraduate students receiving a federal loan, public schools have lowest. Therefore, we recommend that high schoolers with a limited budget to attend public schools where tuition is comparatively low, and private for-profit schools where they are more likely to be granted a federal loan.

Education quality. **Full-time faculty percentage** is selected as the indicator for education quality since it is closely related to students' academic performances. Most colleges have 100% full-time faculty while the rest have varying percentage from none to full.

Diversity. **Racial and gender diversity** are examined by Racial Diversity Index (RDI) and Gender Diversity Index (GDI), each of which represents the probability that any two students selected at random would have different races or genders. Colleges with the highest RDI's and GDI's are identified in the previous report.

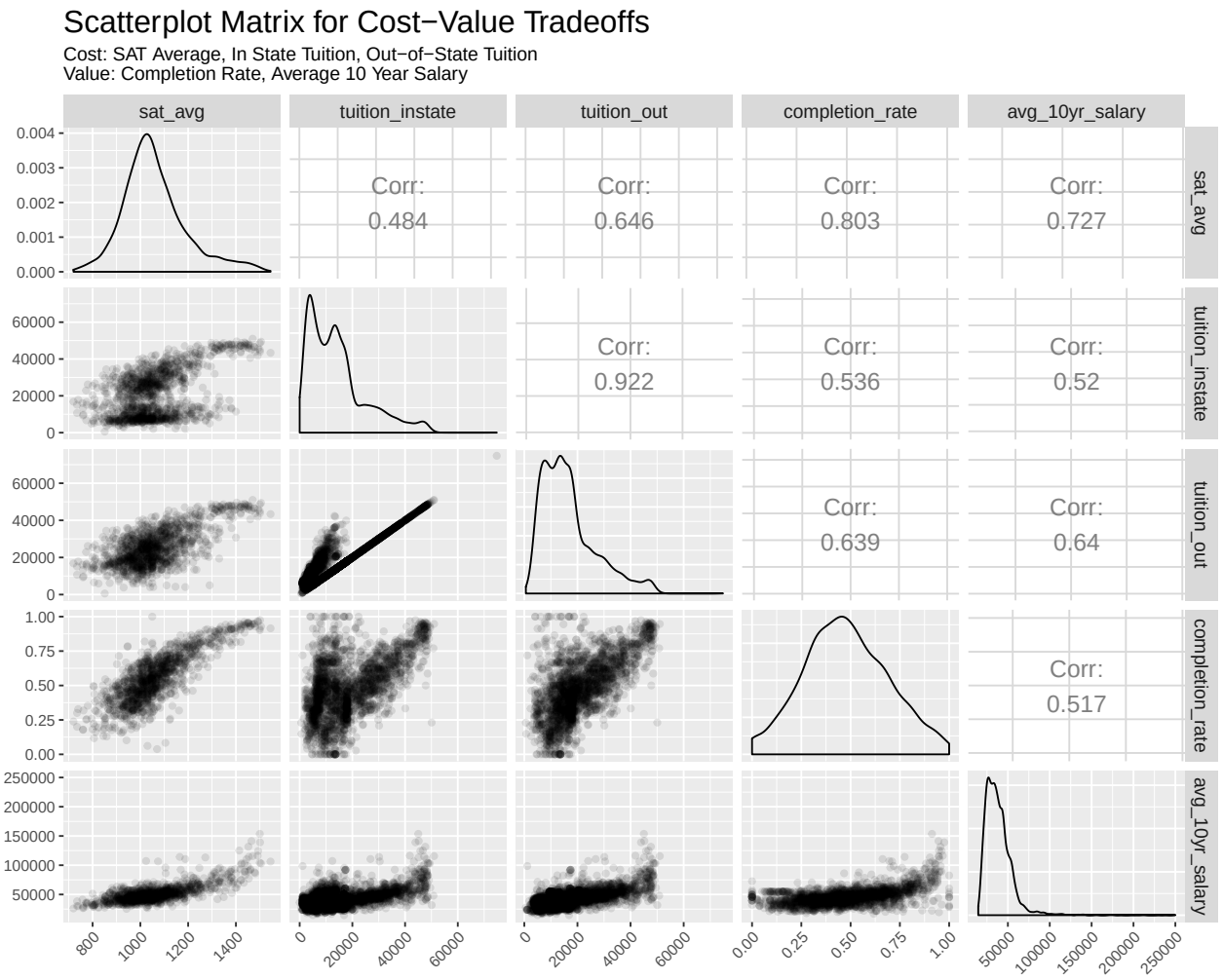
Completion. Completion is evaluated by **completion rate** and **predominant degree awarded**. For most colleges, completion rate is not satisfactory with just under half of the students graduating within 6 years of starting colleges. The most frequently awarded predominant degree is Certificate, while the least is Graduate. Additionally, ownership seems to affect completion. We find that private non-profit schools have the highest completion rate, while private for-profit ones have the lowest. The predominant degree awarded are mostly Certificate, Bachelor's and Associate degree, for private for-profit, private non-profit and public schools, respectively. Therefore, depending on the type of degree one wants to pursue, one can choose colleges of different ownership types accordingly.

Earnings. We investigate earnings by **average 10-year salary** after entry under different ownership types. Private non-profit schools have the highest median for average 10-year salary, while private for-profit schools have the lowest. Additionally, there seems to be a positive relationship between tuition and average 10-year salary, so that students who paid higher tuition at college had higher average 10-year salary.

Cost-value trade-off. Lastly, we inspect cost-value trade-off using the scatterplot matrix. Most importantly, higher **average SAT score** and **tuition** seem to imply higher **completion rate** and **average 10-year salary**. Thus, we have evidence to state that schools with **higher cost** to attend will lead to **higher value**.

```
college = read_rds("college.rds")

cost_value = college %>%
  select("sat_avg","tuition_instate","tuition_out", "completion_rate", "avg_10yr_salary")

ggpairs(cost_value,
                upper = list(continuous = wrap("cor", size = 5, alignPercent = 1)),
                lower = list(continuous = wrap("points",alpha = 0.1))) +
```

```
  theme(axis.text.x = element_text(size = rel(1),angle = 45, hjust = 1),
        axis.text.y = element_text(size = rel(1)),
        strip.text = element_text(size = rel(1)),
        plot.title = element_text(size=18))+
  labs(title = "Scatterplot Matrix for Cost-Value Tradeoffs",
       subtitle="Cost: SAT Average, In State Tuition, Out-of-State Tuition
Value: Completion Rate, Average 10 Year Salary")
```



INTERACTIVE COMPONENT

Our interactive dashboard (https://yourbestcollegechoice.shinyapps.io/final_proj/) aims to help users to filter the characteristics of a specific university they are interested in and to make comparisons across universities. For example, if a person is very interested in Columbia University, he can select Columbia University in the dashboard and see where Columbia University falls on each of the distribution plots of variables including education quality, diversity, completion rate and earnings. From the plots, users can easily compare Columbia University with other universities.

We choose Shiny as our interactive plot tool based on the following reasons. Firstly, Shiny is an effective high-level tool for the interactive plot. Shiny provides very convenient functions for creating a dashboard including base components like sidebar, main body, layout, and reactive plot from data change. Secondly, since Shiny is an R-based GUI programming tool, we could base our shiny app on the R codes we already have. Instead of spending much time on the grammar of a new language, we could focus on the logical part and user experience of this interactive dashboard.

Introduction to our Shiny Dashboard:

1. User cases
   Our research target is to find how schools with higher admission requirement and higher tuition fees correlate to higher education quality, diversity, completion rate, and earnings. We already have researched each variable and their relationships respectively. That means, now for a student or his/her parent, he has a sense of how these variables concerning with cost and value are distributed. However, he will ultimately focus on several universities that he can or wants to go to. Therefore, we provide an interactive tool for students and parents to narrow down the schools he will probably go to and see clearly the aspects of education quality, diversity, completion rate and earnings of that specific university.

2. Functions

(a) Search universities by SAT scores and admission rate SAT is a requirement on admission to most universities and admission rate is an essential aspect students would consider when applying to a university. Therefore, we let users select the universities in some SAT score and admission rate range on the scatterplot, so that they could narrow down the range of universities they are able to attend.

(b) Search a university by name
Users could type the name of the university in the university table to access the info of that university directly.

(c) Plot distribution on education quality, diversity, completion rate and earnings
When a selection of a specific university has been made, the distribution on education quality, diversity, completion rate and earnings will be shown in the form of histograms. Where that university falls in the histogram is also visualized for a quick search.

(d) Data exploring
We provide a separate page for the data we use, because we want to enable users to explore the original data themselves. We also provide the metadata information and data missing patterns for reference.

3. User instructions
   Draw a square range based on the SAT scores and admission rate on the scatterplot. A list of universities within that SAT range and admission rate range will be shown. Then click a university that you are interested in in the table. The distribution plots with the specified university highlighted will be shown below. Besides, you can go to the data page to explore the metadata information, the original data and the data missing patterns.

4. Code links
   https://github.com/ZhangZhida/EDAV_FinalProject/blob/master/shiny/Final_Proj/app.R

CONCLUSION

One limitation of our project is the completeness of the SAT scores. 6398 out of the 7703 (roughly 83%) observations have missing SAT scores, which could jeopardize the generality of our conclusion. Also, since we subsetted the data set to consider only the most recent time period (2014-2015) when average 10-year-salary is available, the observed patterns could have changed over the past few years. Additionally, since we did not cover all the features in the original data set, we might overlook some other important factors, which could include some confounding variables that intermediate the correlations we observed. Lastly, when analyzing the cost-value trade-off, since we only take continuous features into consideration, a caveat should be made regarding the categorical features such as predominant degree, which could affect the overall pattern.

Hence, future exploration could make use of larger span of the original dataset, with longer time span and more variables included. Furthermore, one could apply methods such as logistic regression to analyze the unexplored categorical features in cost-value trade-off.

REFERENCE

Shiny tutorial: https://shiny.rstudio.com/tutorial/

Shiny Interactive visualizations: https://shiny.rstudio.com/gallery/movie-explorer.html