Assignment 2

Jie Gu
SID: 913953707

1.In the data set of college scorecard, there are 14 features with no missing values. They are id, open8_id, open6_id, name, city, state, degrees awarded predominant, degrees awarded highest, ownership, main campus, branches, institutional characteristics. level, zip, academic year. Among these features, id is the unique identification number assigned to postsecondary institutions as surveyed through IPEDS; open8_id and open6_id are the identification numbers used by the U.S. Department of Education's Office of Postsecondary Education (OPE) and Federal Student Aid Office (FSA) (by data documentation). Other features are basic descriptive information about the institution, like name of colleges, so it is convenient for researchers to record these features.

There are 10 features with most missing values in this data set. They are minority_serving.historically_black, minority_serving.predominantly_black, minority_serving.annh, minority_serving.tribal, minority_serving.aanipi, minority_serving.hispanic, minority_serving.nant, men_only, women_only, and operating. Most of these features are about special mission: minority_serving.historically_black is about Historically Black Colleges and Universities; minority_serving.predominantly_black is about Predominantly Black Institutions; minority_serving.annh is about Alaska Native-/Native Hawaiian-serving Institutions; minority_serving.tribal is about Tribal Colleges and Universities; minority_serving.aanipi is about Asian American-/Native American-Pacific Islander-serving Institutions; minority_serving.hispanic is about Hispanic-serving Institutions; minority_serving.nant is about Native American Non-Tribal Institutions; men_only is about institutions for men only; women-only is about the institutions for women only. Since Data on special missions are provided only in the latest Scorecard data file, our data set do not have values for these features. As for operating, it is about whether institutions are currently operating or not. This data element is included only in the latest Scorecard data file, so our data set do not have values for this feature.

To look for the patterns of missing values, let us look at the tables below. These two tables are about the relation between missing values and years and about relations between missing values and ownerships.

Year	Missing values
2012	470009
2013	464046
2014	467017
2015	456149
2016	438477

Ownership	Missing Values
public	554795
Private nonprofit	586651
Private for-profit	1154252

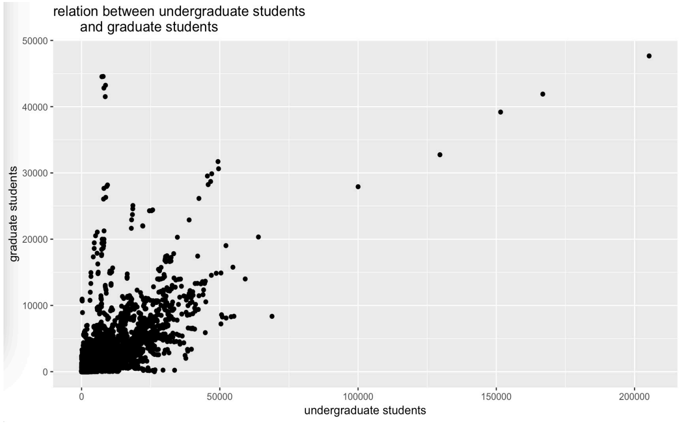
From these two tables, we find that the number of missing values is decreasing from 2012 to 2016. The reason for the pattern may be that the research on these data is improved with time passing by. Also, the reason may be the decreasing of private for-profit schools (By project1). In addition, the missing values in private for-profit schools are much more than that in other two kinds of schools, which also may be caused by the difference in numbers of three kinds of schools ---- the amount of private for-profit schools is much more than other two kinds of schools.

2.

ownership	Median undergraduate students	Mean undergraduate students	Median graduate students	Mean graduate students
Private for-profit	158.0000	483.9548	132.0000	991.8964
Private nonprofit	904.0000	1689.6713	303.0000	984.6068
Public	3181.0000	6011.5654	1281.5000	2456.0197

From the table above, we can easily know that the population for public school is much more than other two kinds of school, and the population for private for-profit school is the least one no matter for graduate students or undergraduate students.

Also, by analysis the mean and median of undergraduate students and graduate schools in relation to years, I find that the mean and median only change a little.

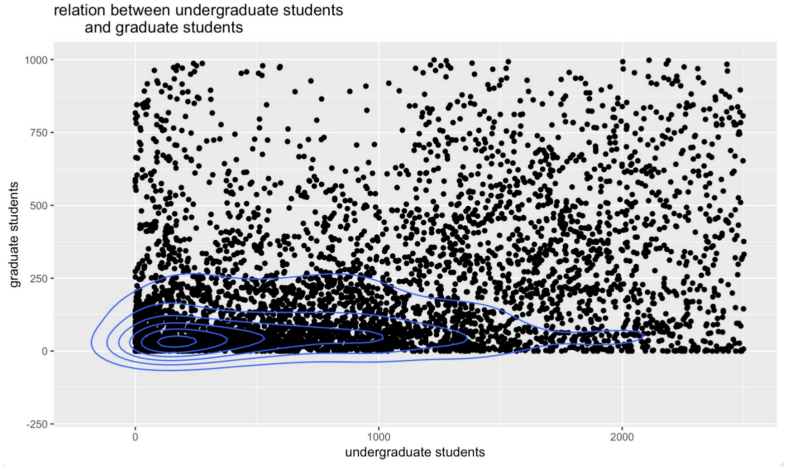


From the dot plot about the relation between the population of undergraduate students and graduate students, we see some outliers with extremely large population. Also, there are some schools with 0 undergraduate students or graduate students.

Walden University, University of Phoenix-Online Campus, University of Phoenix-Arizona are three schools with unusual large number of graduate students and graduate students. The common character shared by Walden University and University of Phoenix is that they all have online programs, which is the reason why they have so many students.

There are 119 observations with 0 graduate students, and 50 observations with 0 undergraduate students. Among schools with 0 graduate students, some of these schools, like California College San Diego, do not have graduate degree. Others, like Pacific Union College, are mainly for undergraduate students even though they have a graduate degree. Similarly, for schools with 0 undergraduate students, they are mainly for graduate students. In addition, we have two schools with 0 graduate students and 0 undergraduate students: Lyme Academy College of Fine Arts in 2014 and School of the Museum of Fine Arts at Tufts University in 2016.

To look for the relationship between undergraduate and graduate populations, we need to zoom in the dot plot and plot a density plot.



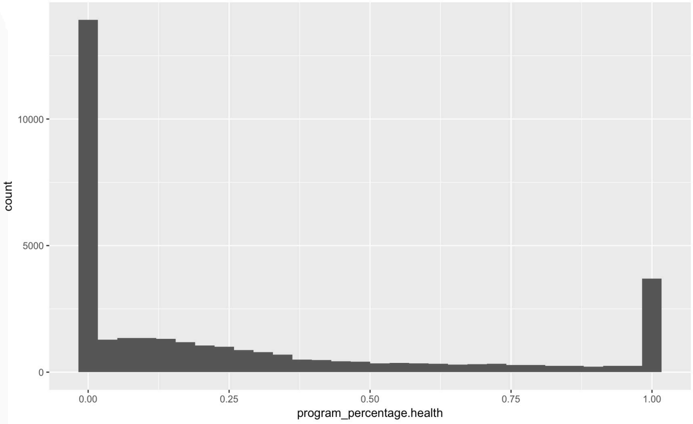
From this density plot, we find that most colleges have an undergraduate population smaller than 2000 and a graduate population smaller than 250. The colleges outside the density plot are all exceptions. Also, from the original dot plot, schools with a large graduate population always have a large undergraduate population. As for this character, Walden University is an exception since it has a small undergraduate population and an extremely large graduate population.

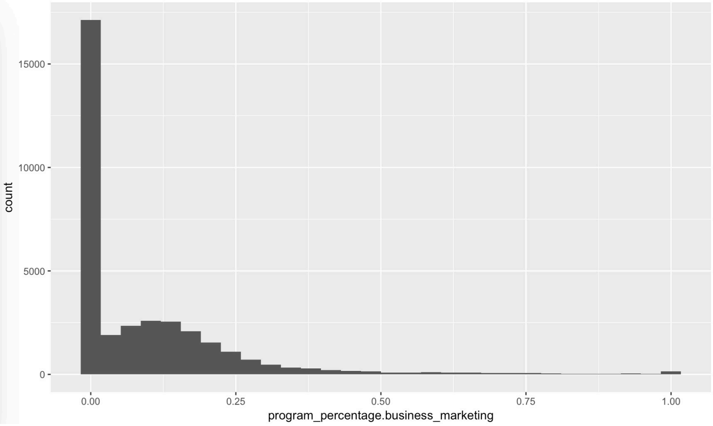
3.To study the popularity of a program, we can look at the mean of each program percentage in each year. The following table shows the programs with 2 highest mean and the programs with two lowest mean in each year.

year	2012		2013		2014		2015		2016	
Most popular	health	Personal culinary	health	Personal culinary	health	Personal culinary	health	Personal culinary	health	Personal culinary
Mean	0.273	0.208	0.270	0.211	0.268	0.218	0.267	0.216	0.268	0.207
Least popular	military	library	library	military	military	library	military	library	military	library
Mean	0.000062	0.000052	0.000067	0.000056	0.000100	0.000062	0.000136	0.000051	0.000132	0.000039

From the table, it is clear that health and personal culinary are the most popular programs; library and military are the least popular programs.

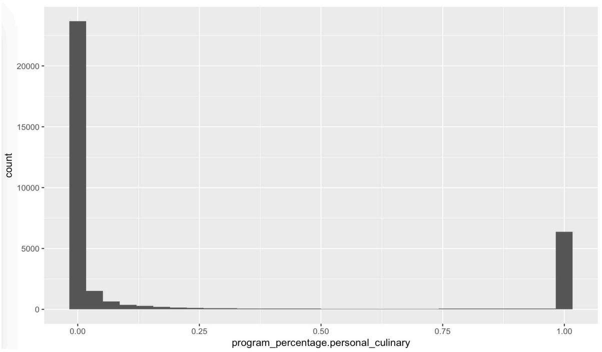
By looking at the mean and median of each program percentage, we find that the median for most program almost equals to 0. However, health and business marketing show a different pattern: their median are both much larger than 0 relative to other programs. So I use a histogram to show their patterns.





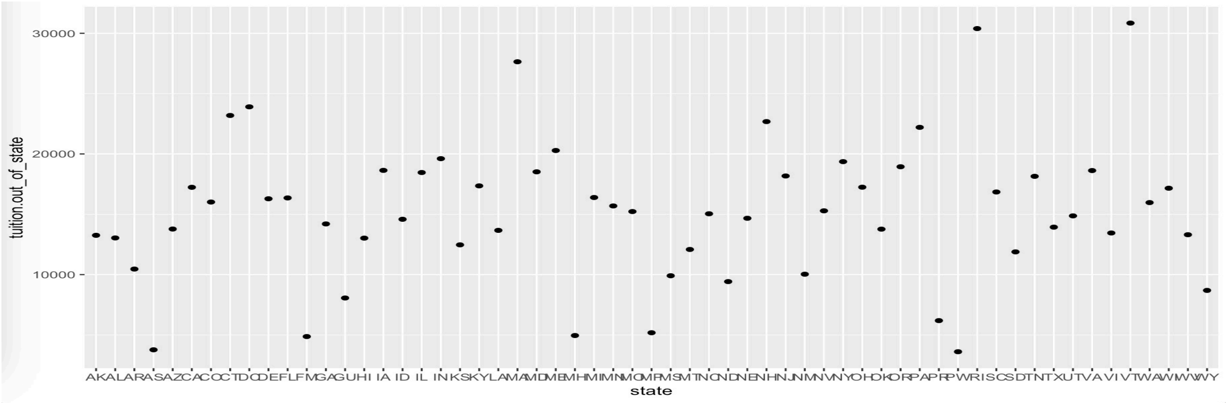
From the histogram of health percentage, although there are many observations around 0, there are also many observations between 0 to 0.50; also, there are many observations around 1. So, these make the median much larger than 0 relatively.
From the histogram of business marketing, there are many observations between 0 to 0.25, which makes the median larger than 0 relatively.

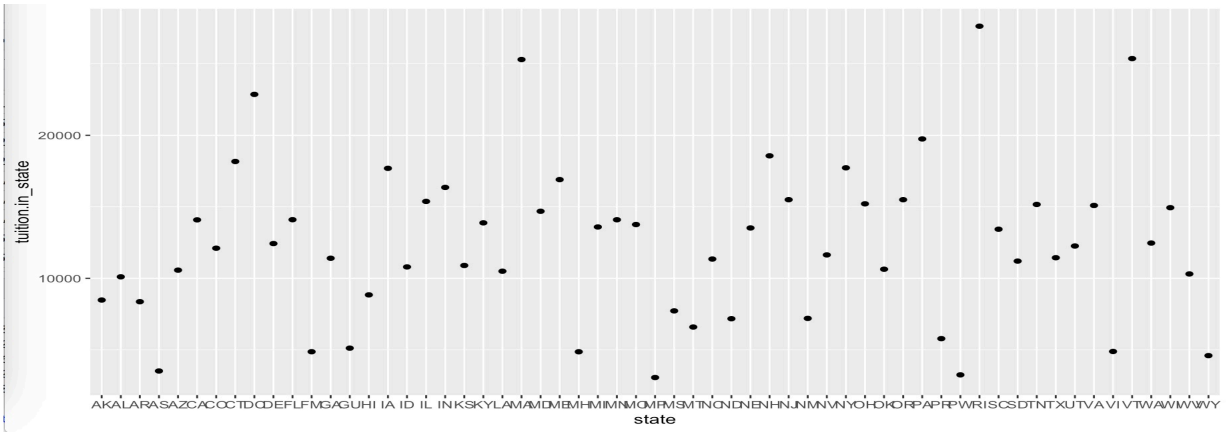
In addition, the percentage of personal culinary also shows a different pattern since its mean is really big. However, its median is still 0. Let us look at the following histogram of personal culinary.



In this histogram, there are many observations around 1, which makes the mean larger than others. However, the majority of observations are around 0, and there are not much observations in the middle, so its median is still close to 0.

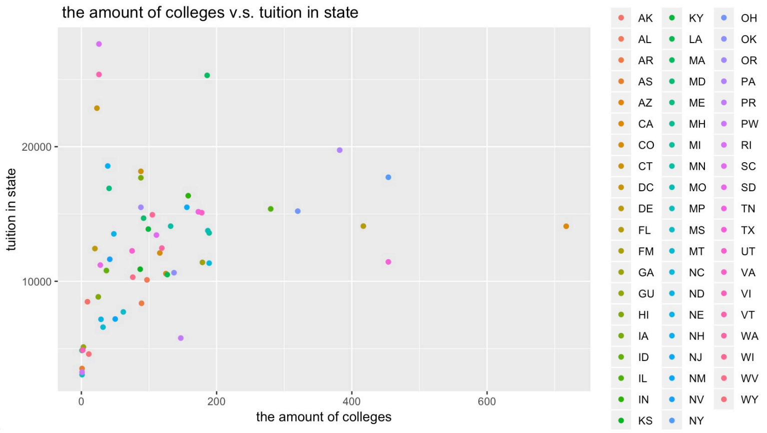
4.By calculating the standard deviation for the mean of both instate and outstate tuition in each states, I find that in each year, the standard deviation keeps the same since the mean keeps the same for each year. So we can use standard deviation of any years to represent general situation. The standard deviation for the mean of instate tuition is 5467.105; the standard deviation for the mean of outstate tuition is 5811.77. Since both standard deviation is large, it means that both instate tuition and outstate tuition vary greatly in different states. Additionally, since the standard deviation is similar, the degrees of variation for instate and outstate tuition are similar. We can see the variation more clearly by dot plots which reflect the mean of instate and outstate tuition in each states.



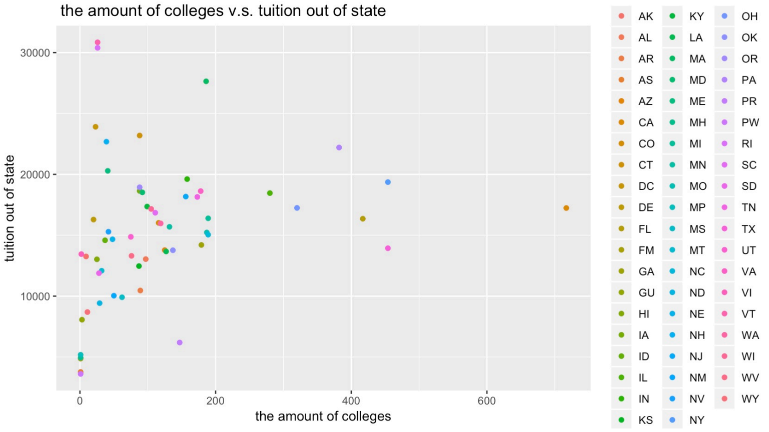


From the plot above, we can easily see that instate tuition and outstate tuition both vary greatly in different states, and the degrees of variation of instate tuition and outstate tuition are similar.

To look for a relationship between the number of universities in a state and tuition, we can look at the following dot plots. The following plots are about the relationship between amount of colleges and mean tuition in state and about the relationship between amount of colleges and mean tuition out of state



In this plot, with the increase in the amount of colleges, the degree of variation in mean tuition in state become smaller.



In this plot, with the increase in the amount of colleges, the degree of variation in mean tuition out of state become smaller.

Therefore, as for the relationship between the number of universities in a state and tuition, the patterns for tuition in state and tuition out of state almost keep the same.

5. To measure the "diversity", we should consider all "demographic" variables. More precisely, Colleges with the most diverse demographics means that the colleges should have all kinds of race and the proportion for each race should be near to each other. Also, the colleges should have veteran and first generation. Moreover, the ratio of the amount

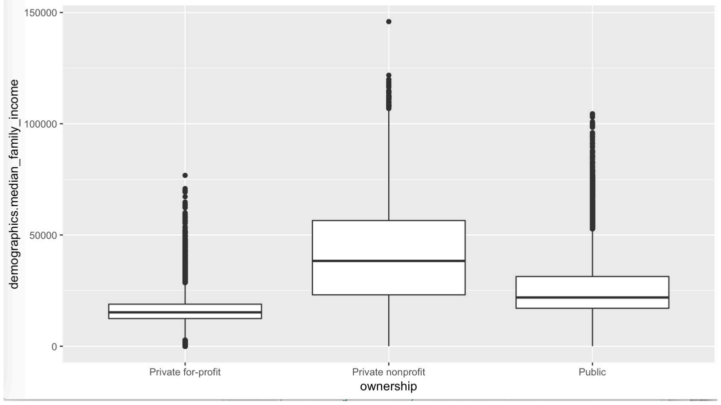
to women to amount of men should be close to 1, which means the colleges have similar amount of men and women.

I first select colleges with all ‘demographics’ variables larger than 0 and with the proportions of men and women both larger than 0.499 to guarantee that the colleges have similar amount of men and women. Then from these colleges, I calculate the standard deviation of race for each college. And finally I choose 6 school with lowest standard deviation, which is smaller than 0.15, from these schools. The 6 most diverse colleges are **Bergen Community College in 2012, Duke University in 2012, Irvine Valley College in 2013, Wayland Baptist University in 2013, Las Positas College in 2014, and University of Washington-Bothell Campus in 2014.**

6. **Q1: What is the situation of median family income for schools with different ownerships?**

To study the situation of median family income, we first look at the mean of it.
The mean of median family income of public school is 27085.61.
The mean of median family income of private for-profit school is 16108.14.
The mean of median family income of private nonprofit school is 41509.47.
Therefore, the mean for private for-profit school is lowest; the mean for private nonprofit school is much higher than others.

Also, we can use a boxplot to show the situation.

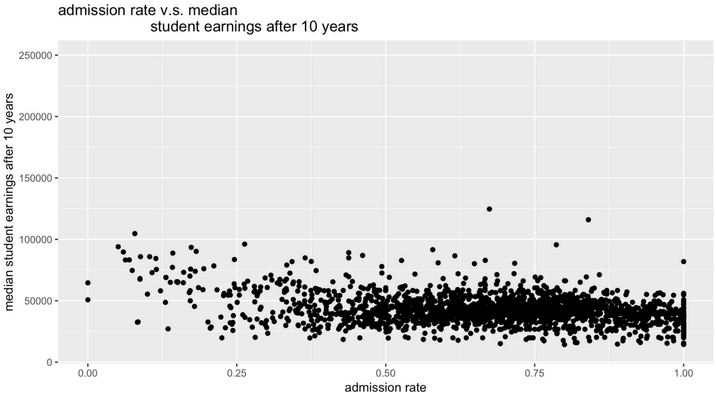


From this plot, we see the median of median family income for private for-profit school is lowest, the median of median family income for private nonprofit school is highest. Also, the standard deviation of median family income for private nonprofit school is much higher than others.

Therefore, we can conclude that the median family income of private nonprofit school is usually higher than others, and the median family income of private nonprofit school varies greatly. The median family income of private for-profit school is usually lower than others, and the degree of variation is small.

Q2: For colleges in 2014, what is the relation between admission rate and the median student earnings after 10 years?

We can use a scatter plot to solve this problem.



From this plot, we can see that before the admission rate smaller than 0.50, the median student earnings after 10 years decreases slowly with the increasing of admission rate; but after the admission rate larger than 0.50, the median student earnings after 10 years almost keep the same. There are also some outliers with both a high admission rate and high median student earnings after 10 years. MCPHS University and Albany College of Pharmacy and Health Sciences are two typical examples of outliers.

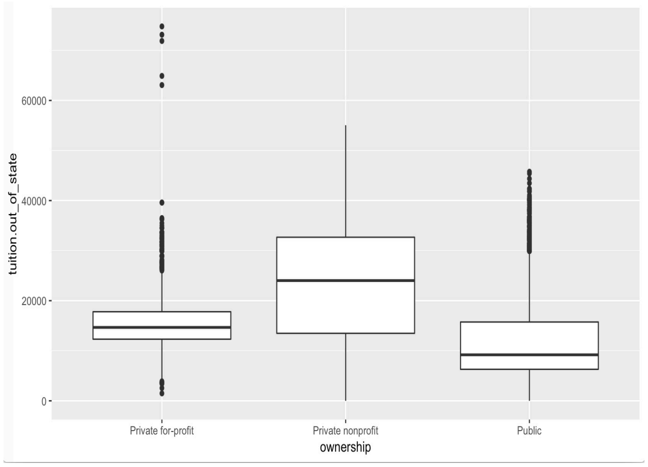
7. For Q1, it leads to an interesting conclusion that the median family income of private for-profit school is usually lower than median family income of public school. This conclusion is interesting since the tuition fee in private for-profit school is usually higher than that of public school. So it is really strange that the median family income of private for-profit school is the lowest one. Therefore, the question is raised that why people with a lower family income choose the private for-profit school even though the tuition fee is higher than that of public school. The question will make a result interesting since we can know the “unique charm” of private for-profit school by solving it.

For Q2, it also leads to an interesting conclusion that after the admission rate larger than 0.50, the median student earnings after 10 years almost keep the same. Usually, a college with a lower admission rate means that the college is hard to enter and is a good college. Also, students in better colleges are more likely to have a better future, which means higher income. So it is interesting that the median student earnings after 10 years almost keep the same instead of increasing after the admission larger than 0.50. For this problem, I have two questions: why the median student earnings after 10 years almost keep the same after the admission rate larger than 0.50; why some colleges with a high admission rate, like MCPHS University and Albany College of Pharmacy and Health Sciences, have a high “median student earnings after 10 years”? These questions will make the result more interesting since it may break down our impression that a high admission rate means the college is not good enough; or it may break down our impression that students in a college that is not good enough can hardly have a good development.

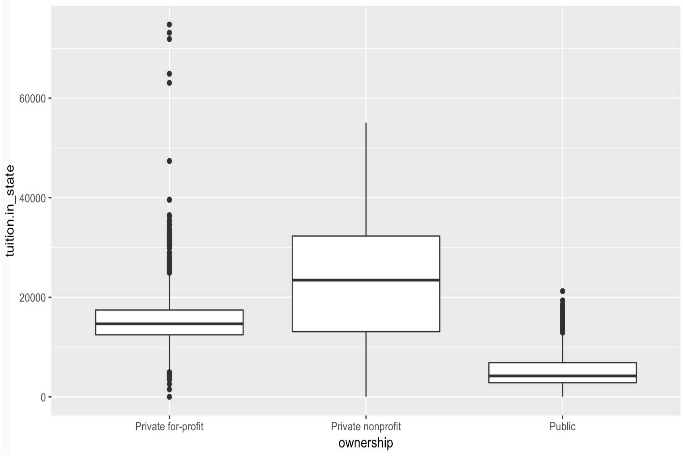
8. **Q1 How do tuition vary between public school, private for-profit school and private nonprofit school?**
This question is raised by problem 4 and 6Q1. By doing problem 4, we have known the relationship between tuition fee and state, so I am also interested in the relationship

between tuition and ownership. Also, since the median family income of private for-profit school is usually lower than that of public school, I want to know the study the tuition fee for private for-profit school and public school.

We can do a boxplot to show the variation and relationship in this question.



From this boxplot, the median of out of state tuition of private nonprofit schools is much higher than that of others, the median of out of state tuition of public school is lowest. The degree of variation for private nonprofit school is greatest; the degree of variation for private for-profit school is smallest. For private for-profit school, it has many outliers, some of which have an extremely high out of state tuition. The Aviator College of Aeronautical Science and Technology is a typical example



This situation for in state tuition for private school is almost the same as the situation for out of state tuition. However, the degree of variation for public school is smallest for in state tuition. Compared with out of state tuition, in state tuition become much cheaper for public school. As for the outliers for private for-profit school, The Aviator College of Aeronautical Science and Technology is also a typical example.

Q2 In question 2, why some schools have both 0 graduate students and 0 undergraduate students?

I think it is really strange for colleges have both 0 graduate students and 0 undergraduate students, so I want to look whether they share some patterns.

They are Lyme Academy College of Fine Arts in 2014 and School of the Museum of Fine Arts at Tufts University in 2016.

They are both school about fine art. The reason for the unusual number of population is that Lyme Academy College of Fine Arts become a part of University of New Haven in

2014 and School of the Museum of Fine Arts at Tufts University become a part of Tufts University in 2016.(By Wikipedia)

Appendix

```
colleges = readRDS('college_scorecard.rds')
library(ggplot2)

#classify colleges by years
colleges2012 <- subset(colleges, academic_year == '2012')
colleges2013 <- subset(colleges, academic_year == '2013')
colleges2014 <- subset(colleges, academic_year == '2014')
colleges2015 <- subset(colleges, academic_year == '2015')
colleges2016 <- subset(colleges, academic_year == '2016')

#classify colleges by ownership
public_colleges <- subset(colleges,ownership == 'Public')
private_forprofit_colleges <- subset(colleges,ownership ==
'Private for-profit')
private_nonprofit_colleges <- subset(colleges,ownership ==
'Private nonprofit')



#Find features with no missing values and most missing values
#----------------------
colleges_isna <- sapply(colleges,is.na)
colleges_isna_count <- apply(colleges_isna,2,sum)
colleges_nomissing = colleges[,colleges_isna_count == 0]
most_misssing = max(colleges_isna_count)
colleges_mostmissing = colleges[,colleges_isna_count ==
most_misssing]

names(colleges_nomissing)
names(colleges_mostmissing)

#by years
sum(is.na(colleges2012))
sum(is.na(colleges2013))
sum(is.na(colleges2014))
sum(is.na(colleges2015))
sum(is.na(colleges2016))

#by ownership
sum(is.na(public_colleges))
sum(is.na(private_forprofit_colleges))
sum(is.na(private_nonprofit_colleges))


#Explore student population------------------

#population for three types of school
aggregate(size ~ ownership, colleges,summary,na.rm = TRUE)
aggregate(grad_students ~ ownership, colleges,summary,na.rm =
TRUE)
aggregate(size ~ ownership, colleges,sd,na.rm = TRUE)
aggregate(grad_students ~ ownership,colleges,sd,na.rm = TRUE)


#poulation for school of different years
aggregate(size ~ academic_year, colleges,summary,na.rm = TRUE)
aggregate(grad_students ~ academic_year, colleges,summary,na.rm =
TRUE)
aggregate(size ~ academic_year, colleges,sd,na.rm = TRUE)
aggregate(grad_students ~ academic_year, colleges,sd,na.rm =
TRUE)

#unusuall population colleges

small_undergraduate = subset(colleges,size == 0)
small_graduate = subset(colleges,grad_students == 0)
unusual_small = subset(colleges,grad_students == 0 & size == 0)

large_exception = subset(colleges,size > 100000 | grad_students >
40000)



#relation between undergraduate and graduate population
ggplot(colleges,aes(x = size,y = grad_students)) +
  geom_point() +
  labs(title = 'relation between undergraduate students
       and graduate students', x = 'undergraduate students',
       y = 'graduate students')

ggplot(colleges,aes(x = size,y = grad_students)) +
  geom_point() + geom_density2d() + xlim(-200,2500) + ylim(-
200,1000) +
  labs(title = 'relation between undergraduate students
       and graduate students', x = 'undergraduate students',
       y = 'graduate students')
one_exception <- subset(colleges, grad_students > 40000 & size <
25000)



#Explore the program percentage-----------------


names(colleges)
program = colleges[47:84]
program2013 = colleges2013[47:84]
program2014 = colleges2014[47:84]
```

```
program2015 = colleges2015[47:84]
program2016 = colleges2016[47:84]





program2012_mean = sapply(program2012,mean,na.rm = TRUE)
sort(program2012_mean,decreasing = TRUE)
program2013_mean = sapply(program2013,mean,na.rm = TRUE)
sort(program2013_mean,decreasing = TRUE)
program2014_mean = sapply(program2014,mean,na.rm = TRUE)
sort(program2014_mean,decreasing = TRUE)
program2015_mean = sapply(program2015,mean,na.rm = TRUE)
sort(program2015_mean,decreasing = TRUE)
program2016_mean = sapply(program2016,mean,na.rm = TRUE)
sort(program2016_mean,decreasing = TRUE)

sapply(program,summary,na.rm = TRUE)

#mean equal to zero, but median is large
ggplot(colleges, aes(x = program_percentage.personal_culinary)) +
geom_histogram()
#mean not equal to 0
ggplot(colleges, aes(x = program_percentage.health)) +
geom_histogram()
ggplot(colleges, aes(x = program_percentage.business_marketing))
+ geom_histogram()



#Analysis tuition---------------
#How does tuition vary across different states?
tuition_list_instate<-
lapply(year_split,function(x)aggregate(tuition.in_state~state,col
leges,mean))
tuition_list_instate


sd(tuition_list_instate$`2012`[[2]])
sd(tuition_list_instate$`2013`[[2]])
sd(tuition_list_instate$`2014`[[2]])
sd(tuition_list_instate$`2015`[[2]])
sd(tuition_list_instate$`2016`[[2]])

tuition_list_outstate<-
lapply(year_split,function(x)aggregate(tuition.out_of_state~state
,colleges,mean))
tuition_list_outstate
```

```
sd(tuition_list_outstate$`2012`[[2]])
sd(tuition_list_outstate$`2013`[[2]])
sd(tuition_list_outstate$`2014`[[2]])
sd(tuition_list_outstate$`2015`[[2]])
sd(tuition_list_outstate$`2016`[[2]])

tution_mean_instate <-
aggregate(tuition.in_state~state,colleges,mean)
ggplot(tution_mean_instate,aes(x = state,y = tuition.in_state)) +
geom_point()

tution_mean_outstate <-
aggregate(tuition.out_of_state~state,colleges,mean)
ggplot(tution_mean_outstate,aes(x = state,y =
tuition.out_of_state)) + geom_point()

#relationship between the number of universities in a state and
tuition
school2016_number = as.double(table(colleges2016$state))
ggplot(tution_mean_instate,aes(x = school2016_number,y =
tuition.in_state,color = state)) +
  geom_point() +labs(title = ' the amount of colleges v.s.
tuition in state',
                 x = 'the amount of colleges',y = 'tuition in
state') +
  guides(color = guide_legend('State'))

ggplot(tution_mean_outstate,aes(x = school2016_number,y =
tuition.out_of_state,color = state)) +
  geom_point() +labs(title = ' the amount of colleges v.s.
tuition out of state',
                  x = 'the amount of colleges',y = 'tuition out
of state') +
  guides(color = guide_legend('State'))



#Demographics-------------------


all(colleges$demographics.age_entry != 0,na.rm = TRUE)
colleges_firststep <-
subset(colleges,demographics.race_ethnicity.white != 0 &
                          demographics.race_ethnicity.black != 0
&

demographics.race_ethnicity.hispanic != 0 &
                            demographics.race_ethnicity.asian !=
0 &
```

```
                            demographics.race_ethnicity.aian != 0
&
                            demographics.race_ethnicity.nhpi != 0
&

demographics.race_ethnicity.two_or_more != 0 &

demographics.race_ethnicity.non_resident_alien != 0 &

demographics.race_ethnicity.unknown != 0 &
                            demographics.veteran != 0 &
                            demographics.first_generation != 0 &
                            demographics.men >= 0.49 &
                            demographics.women >= 0.49)

names(colleges)
colleges_secondstep = colleges_firststep[87:95]
SD = apply(colleges_secondstep,1,sd,na.rm = TRUE)
colleges_thirdstep =
cbind(colleges_firststep[4],SD,colleges_fisrtstep[138:142])
colleges_most_diverse <- subset(colleges_thirdstep,SD < 0.15)

View(colleges_most_diverse)




#The average median family income for different types of schools-
------
mean(public_colleges$demographics.median_family_income,na.rm =
TRUE)
mean(private_forprofit_colleges$demographics.median_family_income
,na.rm = TRUE)
mean(private_nonprofit_colleges$demographics.median_family_income
,na.rm = TRUE)

ggplot(colleges,aes(x = ownership,y =
demographics.median_family_income)) + geom_boxplot()




#For colleges in 2014, what is the relation between
#admission rate and the median student earnings after 10 years
ggplot(colleges2014,aes(x = admission_rate.overall,
                    y = earn_10_yrs_after_entry.median)) +
  geom_point() + labs(title = 'admission rate v.s. median
                  student earnings after 10 years',
  x = 'admission rate',y = 'median student earnings after 10
years')
```

```
unusual_value <-subset(colleges2014,admission_rate.overall > 0.50
&
                     earn_10_yrs_after_entry.median >100000)



#how do tuition vary between three types of school
ggplot(colleges,aes(x = ownership,y = tuition.out_of_state)) +
geom_boxplot()
ggplot(colleges,aes(x = ownership,y = tuition.in_state)) +
geom_boxplot()

high_tuition <-
subset(private_forprofit_colleges,tuition.out_of_state > 60000 |
                     tuition.in_state > 60000)

#why some colleges have both 0 graduate students and 0
undergraduate students
View(unusual_small)
subset(colleges,name == 'Lyme Academy College of Fine Arts')
subset(colleges,name == 'School of the Museum of Fine Arts at
Tufts University')
#schools about fine art, only have
program_percentage.visual_performing.
```