Report

Ruochen Zhong 912888970

Question 1

This dataset has 3312 observations. After checking the variable "main_campus", it shows there are 2431 observations are main campus and 881 observations are not. This means there are 2431 colleges are recorded.

Question 2

When classifying all features, it shows the following table:

Character	Factor	Integer	Logical	numeric
4	**4**	**15**	**3**	**25**

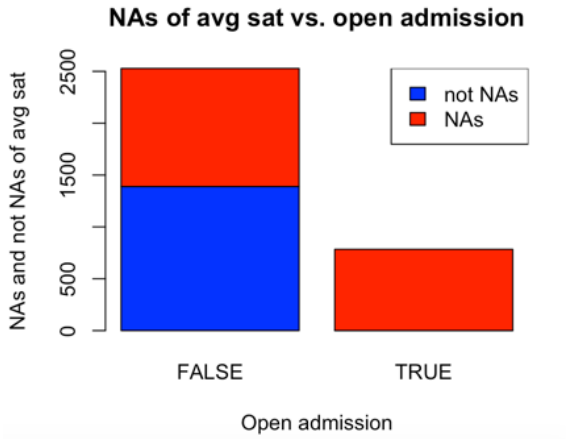
After looking through every feature, the feature "unit_id" should be a categorical feature, but because it is showed as a 6 digits number, it is recognized by R as integer.

Therefore, there are 51 features here, 12 of them are categorical, 14 of them are discrete. Except these two features, 25 of them are continuous.

Question 3

There are 23917 missing values in the dataset, the feature "avg_sat" has the most missing values.

The pattern is that none observations with open admissions have available average sat. The following graph shows the NAs distribution by classifying observations with open admissions and without open admissions:



It shows that if observations have open admissions, it will not have data about average sat. The reason of this pattern is that those schools with open admissions may not care about their applicants' sat scores in their applications, so they do not collect the data of sat scores.

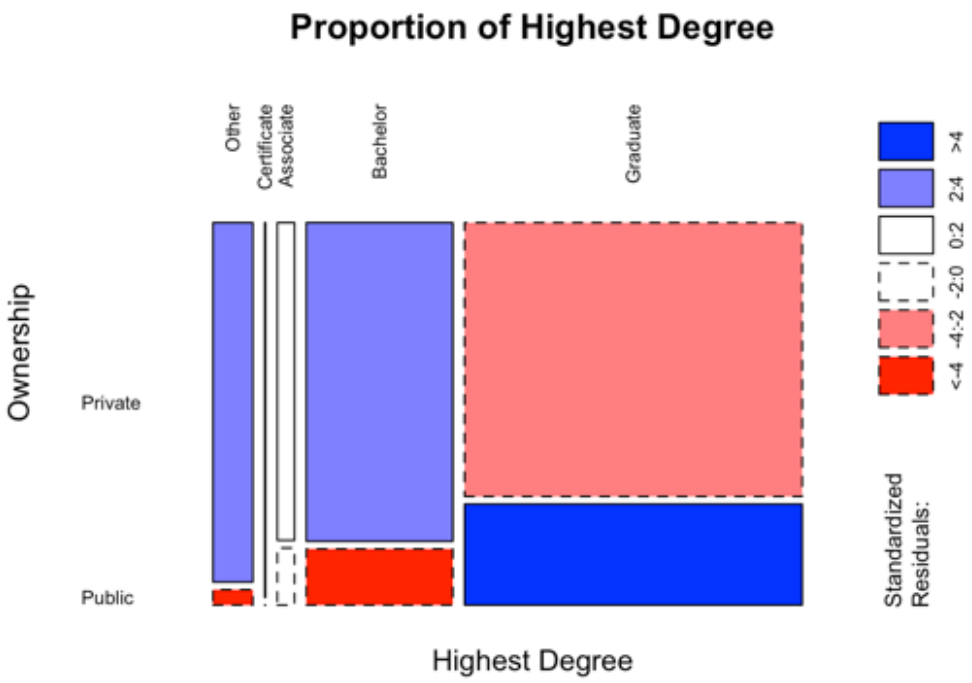
Question 4

There are more private colleges recorded. 716 are public colleges and 2596 are private colleges.

The proportions of highest degree awarded are show in the following table:

	Private	Public
Other	9.014%	1.397%
Certificate	0.193%	0%
Associate	3.428%	2.223%
Bachelor	29.314%	18.715%
Graduate	58.05%	77.654%
Total	100%	100%

Then, draw a mosaic plot:



From this plot, it is clear to find that graduate degree is most common in both private schools and public schools. Bachelor degree ranks at the second. There are only a few Certificate and Associate degree for both types of schools. For every kinds of highest degrees, the number of private schools is more than public schools.

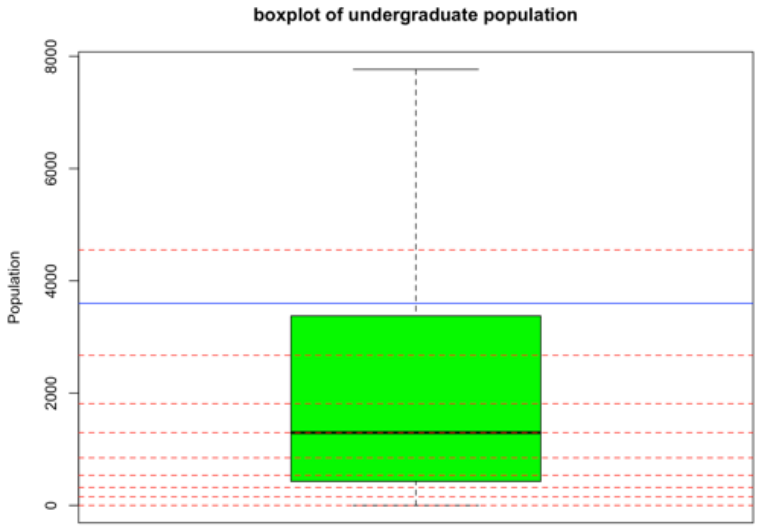
Question 5

The average undergraduate population is 3600, and the median is 1295.

The deciles are:

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0	153.0	319.2	536.0	847.6	1295.0	1811.8	2674.5	4550.8	9629.8	166816.0

Draw a boxplot of the undergraduate population:



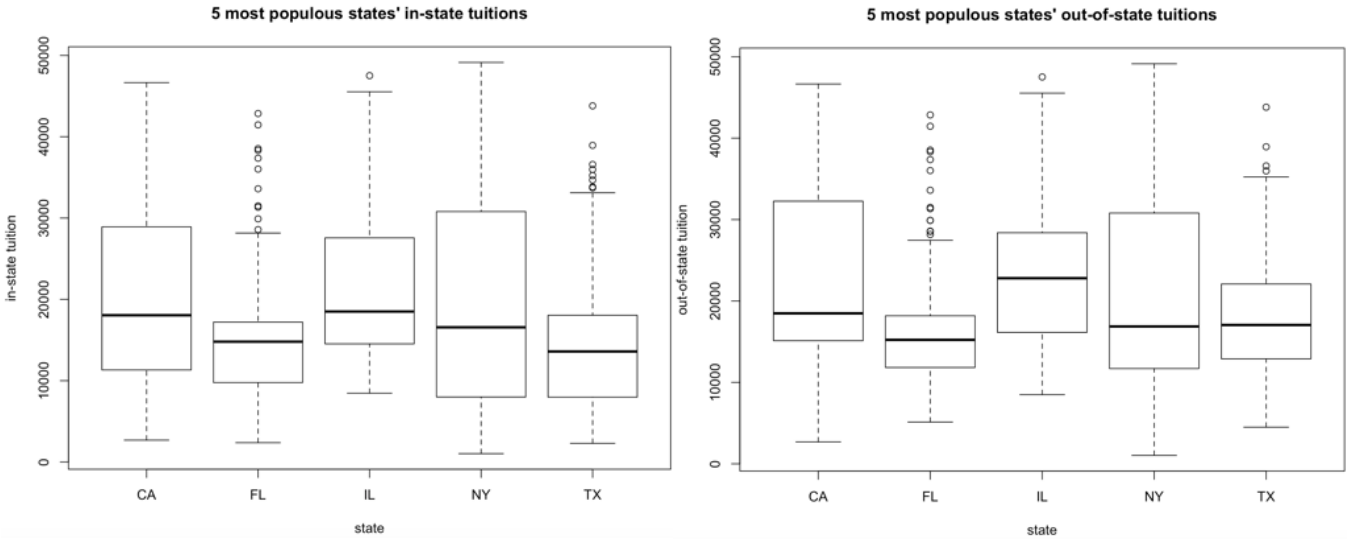
In this plot, the red line is deciles and the blue line is the mean line.

The unusual thing is that the mean line is between the 70% and 80% schools, which means it is far more larger than the median. The reason of this unusual thing is that there are many outliers in observations, more than 10% of them are outliers. Therefore, even if more than 70% observations are below the mean line, those huge outliers improve the mean a lot, and make it almost three times larger than the median.

Question 6

The 5 most populous states is CA, FL, IL, NY, and TX.

Draw the in-state tuitions boxplot and out-of-state tuition boxplot:



From in-state tuition to out-of-state tuition, the difference is that the distribution of CA and NY are skewed to the right more extremely, FL and TX move from skewing to the left to symmetric, and IL move from skewing to the left to skewing to the right. Except IL, there are no obvious changes of the median tuitions of other 4 out-of-state tuitions. These changes means in CA, NY, FL and TX, many schools with high tuitions will be more higher for out-of-state students, and for IL, many schools whose in-state tuition are below $20000 increase a lot for out-of-state tuitions. What's more, the tuition range for CA and NY are more widely, they includes both low tuition schools and high tuition schools. In comparison, In FL, the tuition range tend to be limited in the range between $0 to $30000. In FL and TX, there are several outliers whose tuition are far more higher than other schools in their states, the tuition of those outliers are almost $40000. Those outliers maybe some very famous schools in FL and TX. The median of IL's tuitions are highest for both in-state tuition and out-of-state tuition.

Question 7

Part A

California Institute of Technology has the largest value of average sat.

Part B

Yes, University of Phoenix-Online Campus has the largest amount of undergraduate population and it has open admissions.

Part C

The zip code of the public university with the smallest value of average family income is 11101.

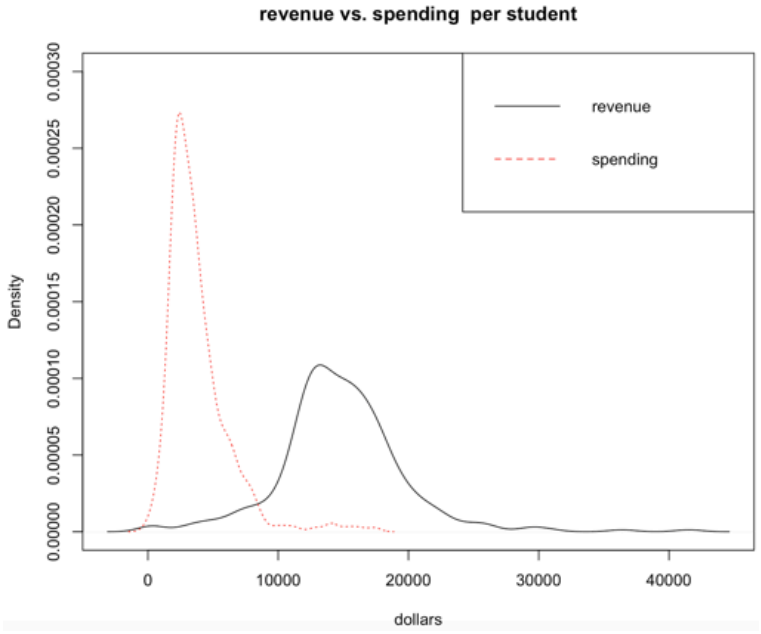
Part D

No, another school called Walden University has the largest amount of graduate population.

Question 8

Part A

Draw a density plot of revenue and spending per student for colleges:



Comparing the density line of revenue and spending, most spending per student concentrate around $4000 to $6000, and most revenue concentrate around $10000 to $20000. This means most schools whose ownership is for-profit and provide Bachelor degrees will make profits successfully.

Then find a linear relationship between revenue and spending per student and draw a regression line:

	Estimate	Std. Error	t - value
intercept	-60.350	382.7844	-0.158
slope	0.2614	0.02454	10.654
Adjusted R-squared	0.2701		

When fitting a linear regression model, it is clear to find several outliers whose spending per student is very high but the revenue per student is relatively low. Those outliers will influence the accuracy of our linear regression model. What's more, although the t-value (10.654) suggests the significance of the positive relationship between revenue per student and the spending per student, the relatively small adjusted R-squared (0.2701) suggests that revenue per student only explain a small amount of variation of spending per student. There are other reasons to influence the rest larger amount of variation of the spending per students.

Part B

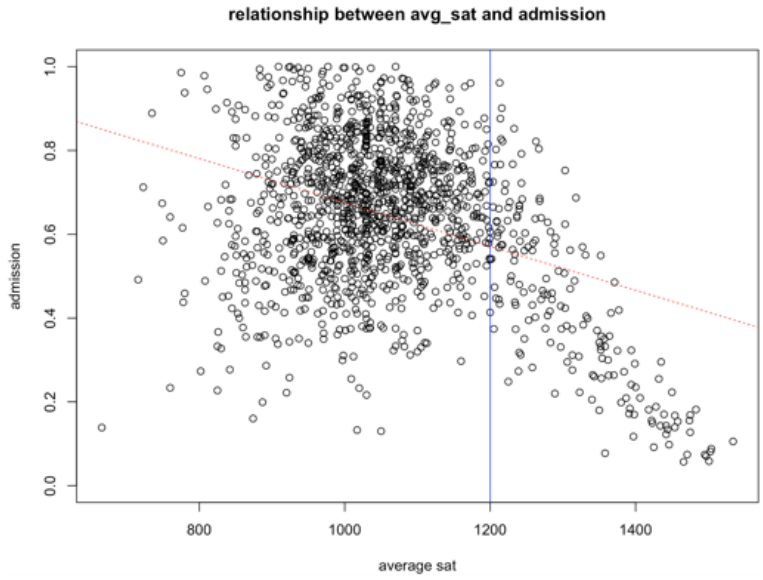
The top 5 earning schools are listed in the following:

School name	School's total net income
University of Phoenix-Online Campus	2632326.2
Ashford University	620193.2
Capella University	578458.9
Grand Canyon University	436293.8
Kaplan University-Davenport Campus	432927.6

Question 9

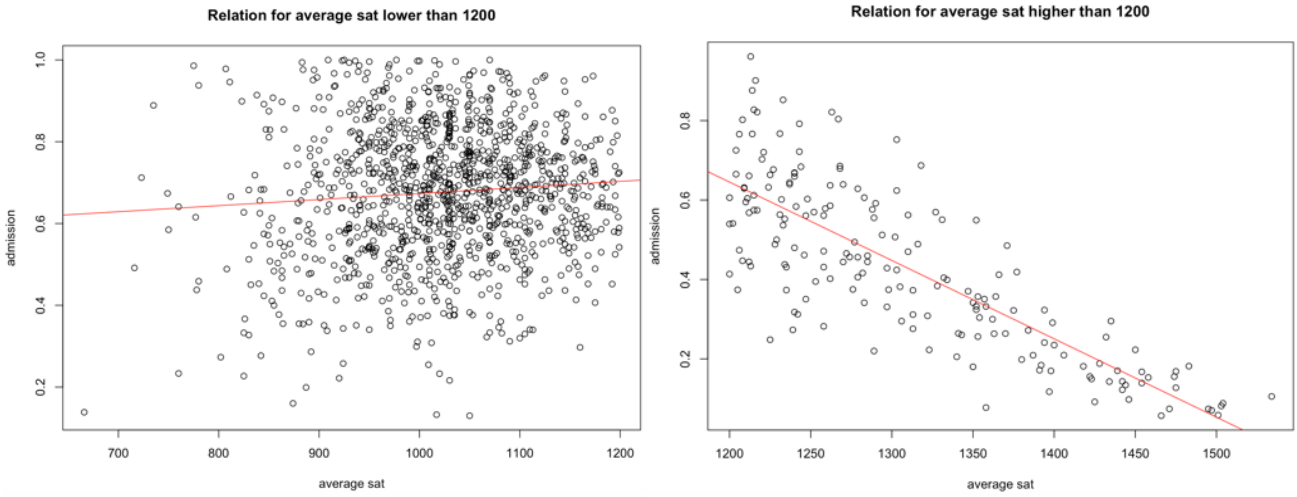
Part A

Plot the observations in a graph, the red line is the regression line of average sat and admission:



Although the regression line shows a negative relationship between average sat and admission of all observations, the distribution shows a reverse U shape trend, which means the admission increase as the average sat increase at first, and then decrease as the average sat increase. From my observation, those dots whose average sat exceeds 1200 shows an obvious decrease trend, so split data base on whether their average sat exceed 1200.

Then, draw two graphs to justify whether these two groups really have difference:



Comparing these two graphs, it is clear to see that for the group whose average sat is lower than 1200 has a weak positive relationship between average sat and admission, but for the group whose average sat is higher

than 1200, there is a clearly negative relationship between average sat and admission.

Part B (a)

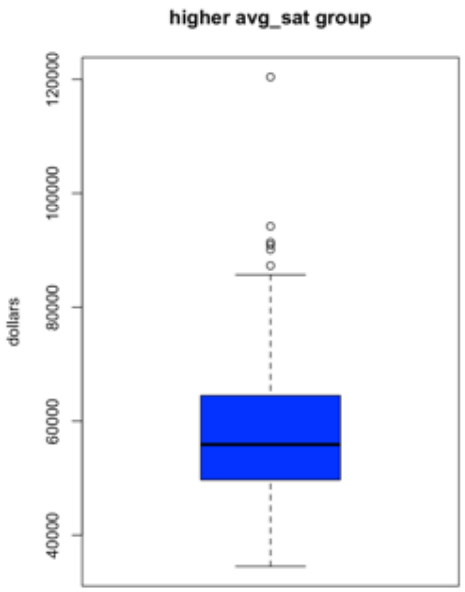
Summary for **higher average sat group's** median salary of students 10 years after starting college:

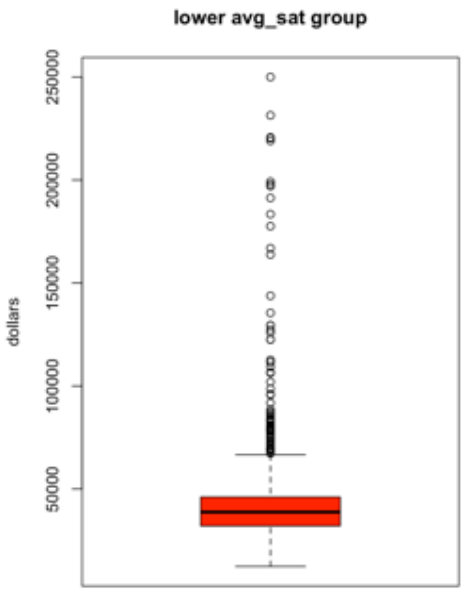
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
34500	49700	55900	58517	64500	120400

Summary for **lower average sat group's** median salary of students 10 years after starting college:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
12400	32100	38800	40205	46100	250000

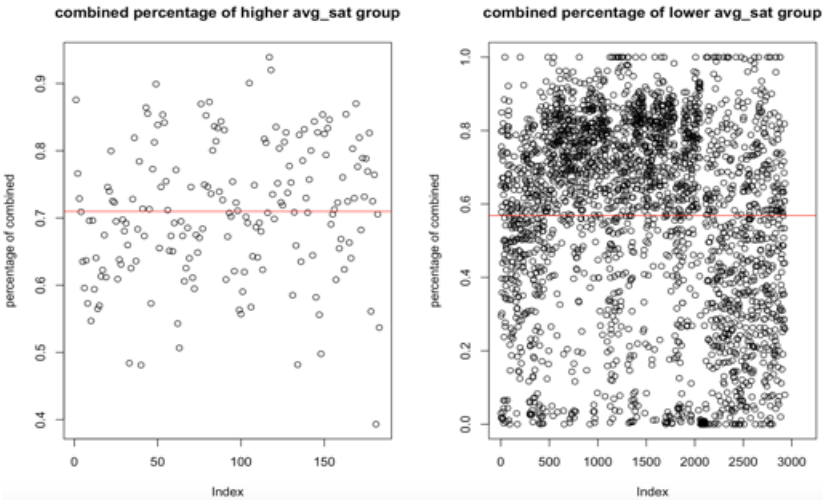
Boxplot for two group's median salary of students 10 years after starting college:





From the summary table and the boxplot, the median and mean of the higher average sat group are both higher than the lower sat group. The range of the higher average sat group is more widely and reach a high level such as $80000. In comparison, although lower average sat groups has some outliers whose median salary of students 10 years after starting college are very high, more than 75% of them are below $50000. In conclusion, joining in schools belong to higher average sat group is more likely to get a higher salary after starting college 10 years.

(b) draw a scatter plot for each group, red line is the mean line:



Summary for **higher avg_sat group's** combined percentage:

Min.	1st Qu	Median	Mean	3rd Qu	Max.
0.3932	0.6404	0.7084	0.7097	0.7846	0.9391

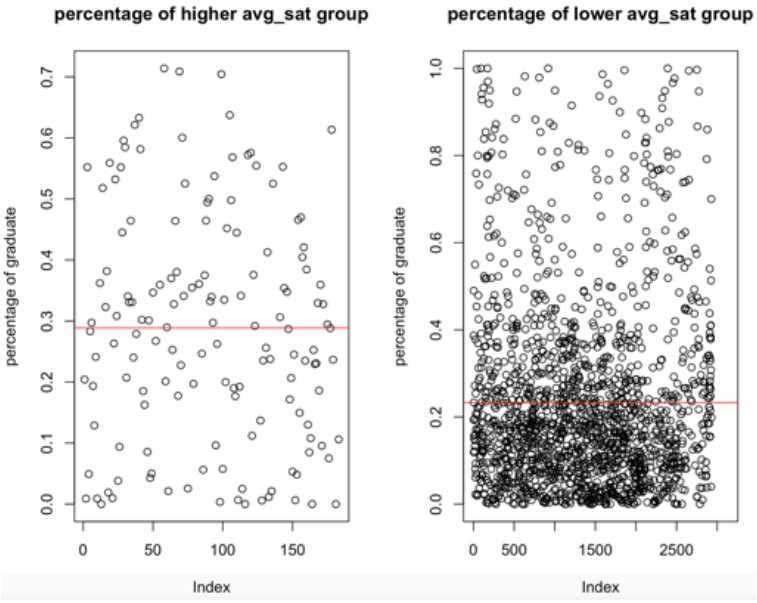
Summary for **lower avg_sat group's** combined percentage:

Min.	1st Qu	Median	Mean	3rd Qu	Max.
0.0	0.3913	0.6343	0.5685	0.7844	1.0

Comparing these two plots, it is clear to see that for higher average sat group, its percentage of Asian and White are almost within the range from 50% to 90%, but for lower average sat group, the percentage vary from 0% to 100%. What's more, higher average sat group has a higher average combined

percentage of White and Asian.

(c) draw a scatter plot for each group, red line is the mean line:



Summary for **higher avg_sat group's** graduate percentage:

Min.	1st Qu	Median	Mean	3rd Qu	Max.
0.0	0.1432	0.2883	0.2887	0.4086	0.7139

Summary for **lower avg_sat group's** combined percentage:

Min.	1st Qu	Median	Mean	3rd Qu	Max.
0.0	0.0862	0.1749	0.2329	0.3061	1.0

Comparing these two plots and tables, it is clear to see that the range of the lower average sat group's graduate percentage is more widely, it varies from 0% to 100% while the higher average sat group's range varies from 0% to 70%. Although the lower sat group's graduate percentage has a higher upper bound, 75% of them are below 30%. In comparison, the distribution of the higher average sat group is more balanced. What's more, the mean of the higher average sat group's graduate percentage is higher.

Part C a) **Group vs. Open admission**

Group \ Open admission	FALSE	TRUE
FALSE	2345	784
TRUE	183	0

From this table, it is clear to see that group and open admission are dependent with each other. The reason may be that schools in the group are very concerned about their applicants' SAT scores, so they will not adopt open admissions.

b) **Group vs. Main Campus**

Group \ Main Campus	FALSE	TRUE
FALSE	880	2249
TRUE	1	182

From this table, although there is 1 observation in group which is not main Campus, all others are main campus, so group and main campus are still dependent with each other. This is because main campus usually has the best resources, education, and reputation, so the main campus will require higher average sat of its applicants. Therefore, schools in group are almost all main campus.

c) **Group vs. Ownership**

Group \ Ownership	Public	Nonprofit	For Profit
FALSE	670	1574	885
TRUE	46	136	1

From this table, it seems that group and ownership are dependent with each other. Only one school in the group are "For Profit", and all others are "Public" and "Nonprofit". The reason is that Schools which are "For Profit" may not very care about their students' average sat score. Instead, they only care about whether they can get enough tuitions to make profits from their applicants, so they accept some students whose sat scores are not very good. Therefore, in the group, there are almost no "For Profit" schools, and this means group and ownerships are dependent with each other.

d) Group vs. Unique branch

Group \ Unique branch	FALSE	TRUE
FALSE	1076	2053
TRUE	14	169

From this table, group and unique branch seems dependent on each other. For schools not in the group, the ratio of unique branch or not is nearly 1:2, but for schools in the group, the ratio becomes almost 1: 12. This huge change of distribution means group are dependent on unique branch. The reason may be that most top Universities in US are concentrate all their resources in one campus, especially those top private schools such as "MIT", "CIT", "Harvard", or "Columbia". These schools usually require a higher average sat scores for their applicants, so the group are dependent with unique branch.

Question 10

Part A

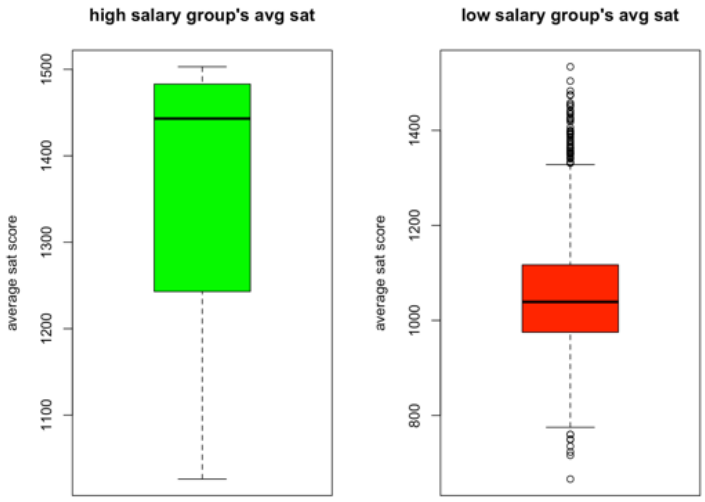
Draw a plot to show the relation between average family income and salary after starting college 10 years:

The red line is the regression line.

From this plot, although it shows a positive relationship between average family income and average salary after starting college 10 years, there are some observations which is very unusually because they have a relatively low family income but a very high salary. Therefore, I divide those observations whose average salary are higher than $100000 as a group because they seems to affect the accuracy of the regression line. In this group,

a higher family income doesn't show a positive relationship with salary. This contradicts the whole trends, so there may be some other important reasons to cause students in those schools get a high salary.

After investigation, there are some tables and graphs to show the difference between these 2 group:



Summary for **high salary group's** average sat:

Min.	1st Qu	Median	Mean	3rd Qu	Max.
1026	1250	1443	1368	1477	1503

Summary for **lower avg_sat group's** average sat:

Min.	1st Qu	Median	Mean	3rd Qu	Max.
666	975	1039	1057	1117	1534

From those data and distributions, it is clear to conclude that group with higher salary after starting college 10 years usually has higher average sat scores. 50% of high salary group's average sat are more than 1443, but for low salary group, 75% are below 1117. Both two group's mean and median of average sat have an obvious difference.

Part B

In real life, it is easy to understand the difference in Part A. Observations with obvious high average sat are usually those top schools in US. Those schools' reputation, resources, and net-working are very helpful for students to find a high salary job. Even though students in those schools don' t have enough support from their family, they can also get high salary. This explain why those points are unusually.

Based on this, **the categorical variable should be the average sat scores, and the level is 1300.** In other words, if average sat >1300, it is true, otherwise, it is false. To make it be able to add in the regression, when average sat > 1300, making this categorical variable equals to 1, otherwise it equals to 0.

For new regression:

	Estimate	Std. Error	T value
Intercept	30470	865.4	35.22
categorical	23870	1196	19.95
Avg_family_income	0.264	0.0131	20.16
Adjusted R-squared	0.4832		

For old regression:

	Estimate	Std. Error	T value
Intercept	33740	637.4	52.94
Avg_family_income	0.2469	0.01153	21.42
Adjusted R-squared	0.1405		

After adding the categorical variable into the regression line, the adjusted R-squared improves effectively. This justifies this categorical variable can improve the fit of the regression line.

Appendix

Cite: Discuss with Chloe Liu about question 9 and Question 10

Use some Professor's in-class codes

Discuss with Hao Luo about the method of question 7 a

```
#Ruochen Zhong 912888970

hw1data = readRDS("/Users/apple/Desktop/college_scorecard_2013.rds")

####question 1####
#to check the number of rows and columns of the dataset, and find the number of main campus
dim(hw1data)
summary(hw1data$main_campus)

####question 2####
#to see every variable's catagories
str(hw1data)
#to caculate the total numbers of every catagory
table(sapply(hw1data, class))

####question 3####
#find total NA numbers
sum(is.na(hw1data))
#check each column's NA numbers
colSums(is.na(hw1data))
#find the column which has largest NA numbers
which.max(colSums(is.na(hw1data)))

#try to find patterns
#get a table of two catagorical variables
pattern1 <- table(is.na(hw1data$avg_sat),hw1data$open_admissions)
#draw a barplot to show the relationship and also add legend to make it clearly
barplot(pattern1, col = c("blue","red"), main = "NAs of avg sat vs. open admission",
        ylab = "NAs and not NAs of avg sat", xlab = "Open admission")
legend("topright",legend=c("not NAs", "NAs"), fill = c("blue", "red"))
```

```
####question 4####
#check each catagory's number
summary(hw1data$ownership)

#create a new varaible who only contain "Public" and "Private" types of ownerships
hw1data$IsPublic<-factor(hw1data$ownership=="Public",labels=c("Private","Public"))

#create a new subset which only contain two variables
highest_degree<-hw1data[c("highest_degree", "IsPublic")]
#make a table of the constitution of the highest degree and transform it to proportion form
degree_table<-table(highest_degree)
prop.table(degree_table,margin = 2)
#draw a mosaic plot of that table
mosaicplot(degree_table, las=2,
           color = TRUE, shade = TRUE, xlab="Highest Degree",
           ylab="Ownership",main="Proportion of Highest Degree")

####question5####
#to get the median and mean
summary(hw1data$undergrad_pop)
#caculate the deciles of the data
quantile(hw1data$undergrad_pop, prob = seq(0, 1, length = 11), na.rm = TRUE)

#draw boxplot of undergraduate population and the line of deciles and mean, exclude NA and outliers
boxplot(hw1data$undergrad_pop, outline = FALSE, main = "boxplot of undergraduate population",
        ylab = "Population",col = "green")
abline(h = quantile(hw1data$undergrad_pop,seq(0, 1, 0.1),na.rm = TRUE), col = 'red',lty = 20)
abline(h = mean(hw1data$undergrad_pop, na.rm = TRUE),col='blue')

####question6####
#create a subset which only contain observations of those 5 populous states
five_states <- hw1data[hw1data$state%in%c("CA","TX","NY","IL","FL"),]
five_states <- droplevels(five_states)
#compare their in-state tuition cost by boxplot
boxplot(five_states$tuition ~ five_states$state, xlab = "state",
        ylab = "in-state tuition", main = "5 most populous states' in-state tuitions")
```

```
#compare their out-of-state tuition cost by boxplot
boxplot(five_states$tuition_nonresident ~ five_states$state, xlab = "state",
        ylab = "out-of-state tuition", main = "5 most populous states' out-of-state tuitions")

####question7#### partA
# find which row has largest sat and get that row's variable "name"
which.max(hw1data$avg_sat)
hw1data[105,]$name

####question7#### partB
#find which row has largest udnergraduate population and also check that row's name and open admissions
which.max(hw1data$undergrad_pop)
hw1data[2371,]$name
hw1data[2371,]$open_admissions

####7#### partC
#create a subset which only contain Public schools
public_university <- subset(hw1data, hw1data$ownership %in% c("Public"))
#find the observation with minimum average family income and its zipcode
which.min(public_university$avg_family_inc)
public_university[348,]$zip

####question7#### partD
#check whether the school with largest graduate population is same as school in partB
which.max(hw1data$grad_pop)
hw1data[248,]$name

####question8####parta
#create a subset
schools <- subset(hw1data, hw1data$ownership %in% c("For Profit") & hw1data$primary_degree %in%
c("Bachelor"))

#draw density line of two variables in one graph to see their distribution
plot(density(schools$revenue_per_student), main = "revenue vs. spending    per student"
     , xlab = "dollars", ylim = c(0, 0.00030))
lines(density(schools$spend_per_student), col = 'red', lty = 3)
```

```
legend("topright",c("revenue", "spending"),col=c(1,2), lty=c(1,2))

#fitting those 2 varaibles into a regression model and draw a scatter plot with regression line
reg1 <- lm (schools$spend_per_student ~ schools$revenue_per_student)
summary(reg1)
plot(schools$revenue_per_student, schools$spend_per_student, xlab = "revenue per student",
     ylab = "spend per student", main = "regression model of spending vs. revenue")
abline(reg1, col = 'red')

####question8#### partb
#set all NA in this subset equal to zero
schools[is.na(schools)] <- 0
#create variable "total_net_income", the unit is thousands
schools$total_net_income <- ((schools$revenue_per_student - schools$spend_per_student) *
(schools$undergrad_pop + schools$grad_pop)) / (1000)

#create a new data frame which only contains 2 variables, rank all observations by total_net_income
schools_income <- data.frame(schools$name, schools$total_net_income)
income_order <- order(schools_income$schools.total_net_income, decreasing = TRUE)
schools_income <- schools_income[income_order,]
#show top 5
head(schools_income, 5)

####question9#### parta
#fitting those 2 varaibles into a regression model and draw a scatter plot with regression line
reg2 <- lm(hw1data$admission ~ hw1data$avg_sat)
plot(hw1data$avg_sat, hw1data$admission, xlab = "average sat", ylab = "admission",
     main= "relationship between avg_sat and admission")
abline(v=1200, col="blue")
abline(reg2, col = "red", lty = 3)

#create a new variable to split data, and update it
hw1data$group <- (hw1data$avg_sat >= 1200)
hw1data$group.update <- ifelse(is.na(hw1data$group), FALSE, hw1data$group)

#construct two subsets
```

```
sat_lower <- subset(hw1data, hw1data$avg_sat < 1200)
sat_higher <- subset(hw1data, hw1data$avg_sat >= 1200)

#draw the plot and regression line for group whose sat lower than 1200
plot(sat_lower$avg_sat, sat_lower$admission, xlab = "average sat", ylab = "admission",
     main= "Relation for average sat lower than 1200")
reg3 <- lm(sat_lower$admission ~ sat_lower$avg_sat)
abline(reg3, col = "red")

#draw the plot and regression line for group whose sat higher than 1200
plot(sat_higher$avg_sat, sat_higher$admission, xlab = "average sat", ylab = "admission",
     main= "Relation for average sat higher than 1200")
reg4 <- lm(sat_higher$admission ~ sat_higher$avg_sat)
abline(reg4, col = "red")

####question9#### partB(A)
#view a rough distribution of two groups
summary(hw1data$med_10yr_salary[hw1data$group.update == "TRUE"])
summary(hw1data$med_10yr_salary[hw1data$group.update == "FALSE"])

#using boxplot to show their difference
par(mfrow=c(1,2))
boxplot(hw1data$med_10yr_salary[hw1data$group.update == "TRUE"],
        ylab = "dollars",main = "higher avg_sat group", col = "blue")
boxplot(hw1data$med_10yr_salary[hw1data$group.update == "FALSE"],
        ylab ="dollars",main = "lower avg_sat group", col = "red")

####question9#### partB(B)
#create a new variable
hw1data$race_combined <- hw1data$race_asian + hw1data$race_white

#view a rough distribution of two groups
summary(hw1data$race_combined[hw1data$group.update == "TRUE"])
summary(hw1data$race_combined[hw1data$group.update == "FALSE"])

#draw scatter plot to check their difference
```

```
par(mfrow=c(1,2))
plot(hw1data$race_combined[hw1data$group.update == "TRUE"],
     ylab = "percentage of combined", main = "combined percentage of higher avg_sat group" )
abline(h = 0.7097, col = 'red')
plot(hw1data$race_combined[hw1data$group.update == "FALSE"],
     ylab = "percentage of combined",main = "combined percentage of lower avg_sat group")
abline(h = 0.5685, col = 'red')

####question9#### partB(C)
#create a new variable
hw1data$grad_student_rate <- (hw1data$grad_pop)/(hw1data$undergrad_pop + hw1data$grad_pop)

#view a rough distribution of two groups
summary(hw1data$grad_student_rate[hw1data$group.update == "TRUE"])
summary(hw1data$grad_student_rate[hw1data$group.update == "FALSE"])

#draw scatter plot to check their difference
par(mfrow=c(1,2))
plot(hw1data$grad_student_rate[hw1data$group.update == "TRUE"],
     ylab = "percentage of graduate", main = "percentage of higher avg_sat group" )
abline(h = 0.2887, col = 'red')
plot(hw1data$grad_student_rate[hw1data$group.update == "FALSE"],
     ylab = "percentage of graduate",main = "percentage of lower avg_sat group")
abline(h = 0.2329, col = 'red')



####question9#### partC
#for each relationship, check their dependence by table

#a
relation1<-hw1data[c("group.update", "open_admissions")]
table1<-table(relation1)
table1

#b
relation2 <- hw1data[c("group.update", "main_campus")]
```

```
table2 <- table(relation2)
table2

#c
relation3 <- hw1data[c("group.update", "ownership")]
table3 <- table(relation3)
table3

#d
# create a new variable first, then create a table
hw1data$unique_branch <- (hw1data$branches == "1")
relation4 <- hw1data[c("group.update","unique_branch")]
table4 <- table(relation4)
table4



####question10#### part A
#create a regression model
reg5 <- lm(hw1data$avg_10yr_salary ~ hw1data$avg_family_inc)
summary((reg5))

# draw the regression line
par(mfrow=c(1,1))
plot(hw1data$avg_family_inc, hw1data$avg_10yr_salary, xlab = "average family income",
     ylab = "average salarys after starting college 10 years", main = "10 years salary vs. family income")
abline(reg5,col = 'red')

# split the dots by a line by observation
abline(h = 100000, col = 'blue')

#create a new variable to split data into two groups, update it
hw1data$group2 <- (hw1data$avg_10yr_salary > 100000)
hw1data$group2.update <- ifelse(is.na(hw1data$group2), FALSE, hw1data$group2)

#create two subsets
high_salary <- subset(hw1data, hw1data$group2.update == "TRUE")
```

```
lower_salary <- subset(hw1data, hw1data$group2.update == "FALSE")

#study some difference between those two subsets
summary(high_salary$avg_sat)
summary(lower_salary$avg_sat)
par(mfrow=c(1,2))
boxplot(high_salary$avg_sat, ylab = "average sat score",
        main = "high salary group's avg sat", col = 'green')
boxplot(lower_salary$avg_sat, ylab = "average sat score",
        main = "low salary group's avg sat", col = 'red')

####question10b####
#create a catagorical variable, if avg_sat >= 1300, it equals 1, if avg_sat < 1300, it equals 0
hw1data$catego <- ifelse((hw1data$avg_sat >= 1300), 1, 0)
#make it from numeric to catagorical for caculation
hw1data$catego <- as.factor(hw1data$catego)
summary(hw1data$catego)

#add this catagorical variable to the previous regression to create a new regression
reg_modified <- lm(avg_10yr_salary ~ catego + avg_family_inc, data = hw1data)
#check the difference between the new and old regression
summary(reg_modified)
summary(reg5)
```