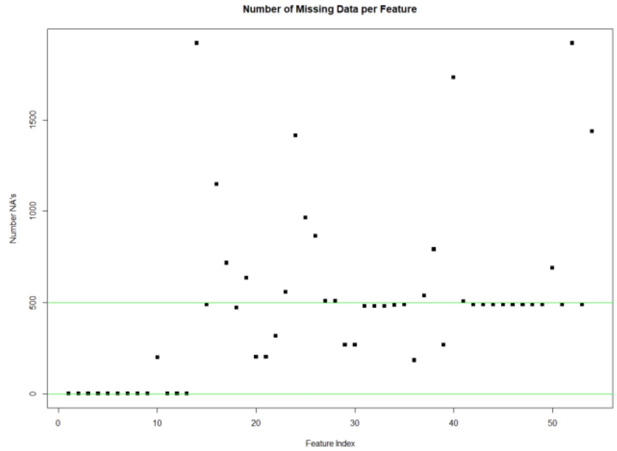
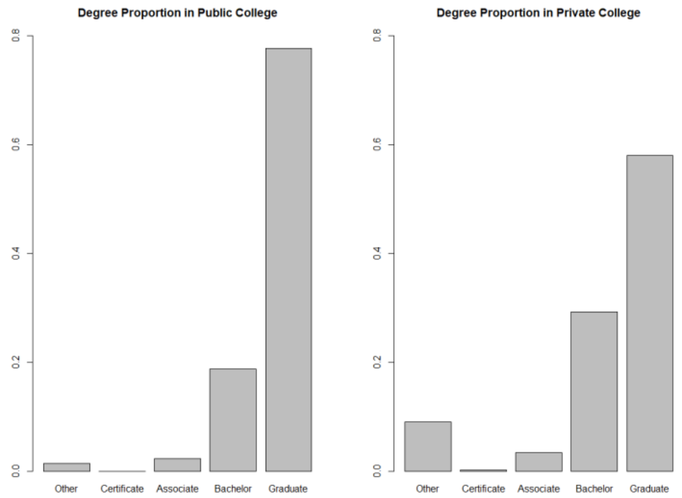
STA 141A HW 1
Jared Yu

1. There are a total of 3,312 observations within the data. For number of colleges there are 2,431.
2. In total there are 51 variables, so that makes for 51 features for this dataset. The categorical variables are: unit_id, ope_id, name, city, state, and zip. The numerical discrete variables are: branches, undergrad_pop, and grad_pop. Additionally, there are various numerical variables which are continuous: avg_sat, cost, tuition, tuition_nonresident, revenue_per_student, spend_per_student, avg_faculty_salary, ft_faculty, admission, retention, completion, fed_loan, pell_grant, avg_family_inc, med_family_inc, avg_10yr_salary, sd_10yr_salary, med_10yr_salary, med_debt, med_debt_withdraw, default_3yr_rate, repay_5yr_rate_withdraw, repay_5yr_rate, avg_entry_age, veteran, first_gen, male, female, race_white, race_black, race_hispanic, race_asian, race_native, race_pacific, net_cost, and race_other. Ordinal variables are: primary_degree and highest_degree. Logical variables (R only): main_campus, open_admissions, online_only.
3. There are 23,197 NA's in the dataset. The greatest number of NA's are in the variable avg_sat, with a total of 1,923 NA's. Revenue_per_student and spend_per_student both have 201 NA's. They both are related to the money used by each student, so it makes sense that information either exists for both or none of them. The variables fed_loan and pell_grant both have 510 NA's, both are related to the grant money provided by the government, so it makes sense that data exists for either both or neither. The variables avg_family_inc and med_family_inc both have 267 NA's, and it makes sense that data either exists for both the average and median family income or neither. The variables avg_10yr_salary, sd_10yr_salary, med_10yr_salary all have 480 NA's, and it makes sense that average, standard deviation, and median information exists for 10-year salaries or none exists at all. The variables med_debt and med_debt_withdraw are similar at 487 and 489 NA's. The definitions are similar, so the difference between the two is likely an anomaly.



When plotting the number of NA's for each feature index, it seems that most of the NA's are within the range of 0-500 for each feature that has NA's.

4. There are 715 public colleges, while there are 2,596 private colleges which means that there are many more private than public colleges.

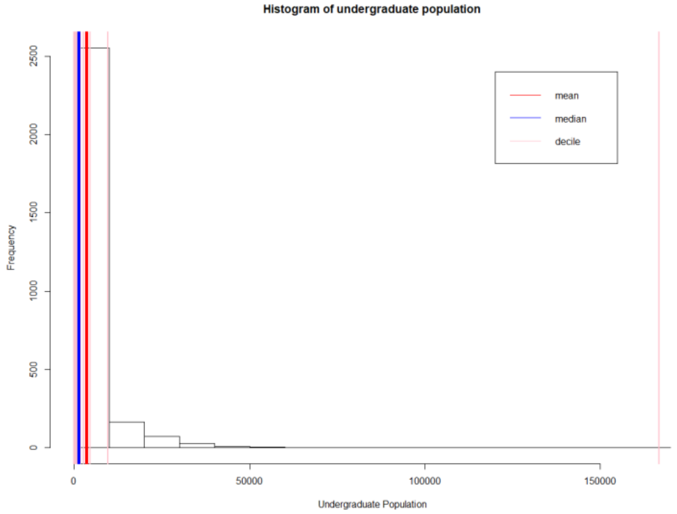


From the plot it seems that public colleges have a larger proportion of graduate in comparison to private colleges where the proportion between the graduate and bachelor students is more balanced. It seems that public universities in general offer the traditional track of a bachelor's degree and afterwards have some graduate programs within their possible education tracks. Private universities seem more balanced in that they offer students a variety of other degrees, possibly professional degrees which balance the number of bachelors and graduate students.

5. The average undergraduate population is 3599.502 when the mean of the category is taken. The median for this category is 1295.

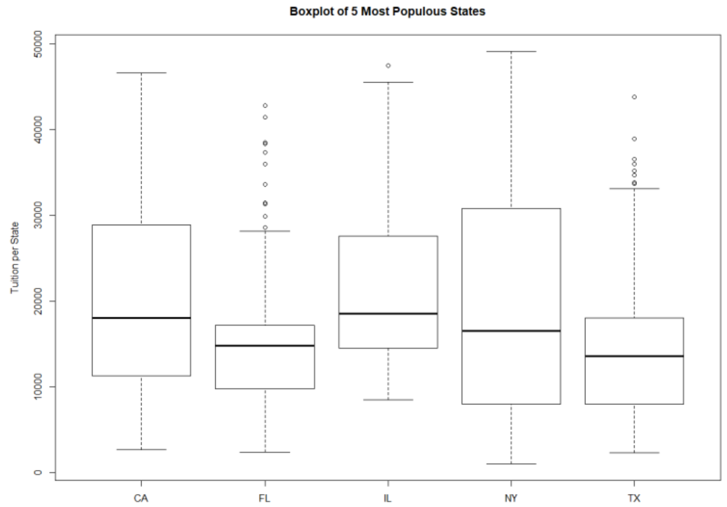
   The deciles are as follow:

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.0	152.7	319.0	536.0	847.2	1295.0	1812.1	2677.5	4551.0	9650.5	166816.0



Above is a histogram of the undergraduate population, along with the frequency for the populations. Most of the population seems to be on the far left side, with only a few exceptions towards the right. There is one extreme population to the right. Much of the skew seems attributable to the University of Phoenix, Arizona. This school has a population of over 150,000 students in their undergraduate population. Removing this piece of data and redoing the histogram plot seems to spread out the statistics on the data somewhat more evenly. However, the data still appears quite skew in appearance. (second plot not included)

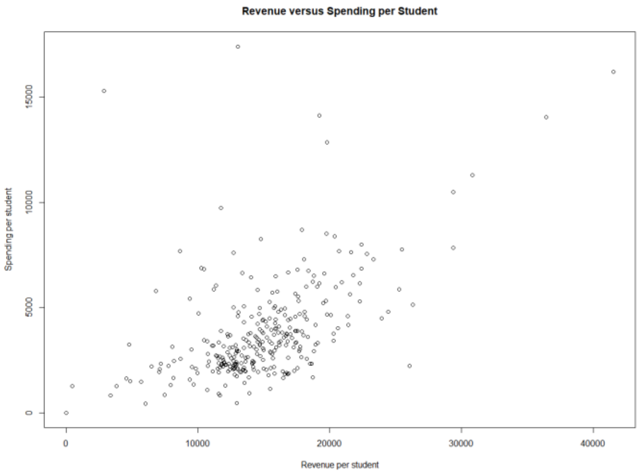
6. According to the TA, the 5 most populous states are determined to be California, Florida, Illinois, New York, and Texas.



From the boxplot it seems that all the states have a skew towards higher tuition. New York seems to have both the highest and lowest tuition. Florida and Texas tend to have a greater number of outliers than the other states. Texas and Florida also seem to have a lower tuition than others. California, the largest state seems to have the most balanced boxplot, with the median and the different quartiles looking slightly more balanced than the others. This makes more sense since it has almost twice the population of other states, so it has more data which can begin to appear more normal and balanced. The variance in New York is also the largest.

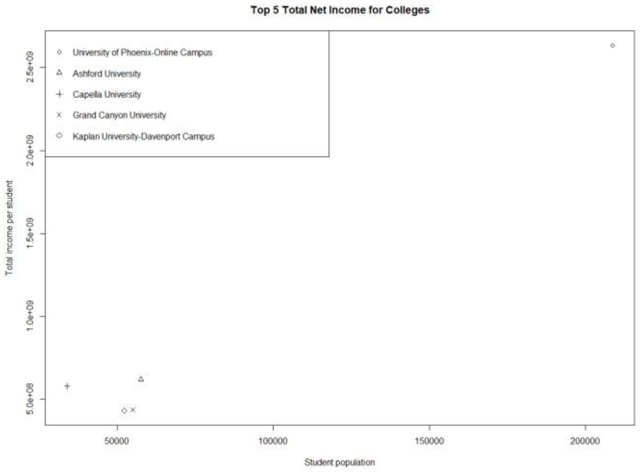
7. Part a) Using the following line in R: cs[which.max(cs$avg_sat), 'name'], it is possible to determine that the college with the largest 'avg_sat' is 'California Institute of Technology.'
Part b) Using the following line in R: cs[which.max(cs$undergrad_pop), c('name', 'open_admissions')], it is possible to determine that the largest university with open admissions, the 'University of Phoenix-Online Campus' does have open admissions set to 'TRUE', so no.
Part c) After using the subset function to create a variable which selects public universities, it is then possible to determine the index of the university with the smallest 'avg_family_inc.' The zip code of this school is '11101.'
Part d) The index of the school in part B was '2371.' After using this to check the 'grad_pop', the same school was determined to have '41,900' for 'grad_pop.' Then checking the dataset again to determine what school had the largest 'grad_pop' in the data, it was found out that there was a different larger number for the maximum 'grad_pop.' So the answer for this question is no.

8. part a) Below is a scatter plot of the two variables plotted against each other:



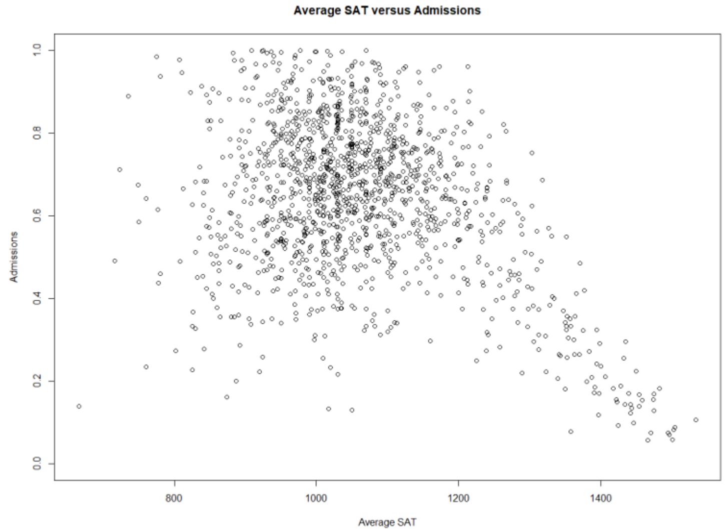
The relationship between the variables spend_per_student and revenue_per_student seems to have a somewhat linear relationship. The plot seems to imply that the greater the revenue that a college receives from a student, the greater the college will also be spending on each student. The majority of this happens around the $10,000 - $20,000 revenue per student and the $0 – 5,000 range for spending per student. This seems to imply that a college will receive up to twice however much it may spend on each student, which makes sense since they still need to make some sort of profit. There are some outliers, such as where certain colleges spend much more on their students than others, and some colleges make much more revenue than others per student. Some of the outliers to the top left of the plot may possibly violate certain assumptions of linearity.

Part b) Below is a visual of the top 5 earning schools:



The variable for total net income comes from subtracting the spending per student from the revenue per student for each college. This gives the total profit received from each student received by the college. After this calculation, the total profit per student is multiplied by the entire student population which consists of graduate students summed together with undergraduate students. These top 5 are then plotted which are: University of Phoenix-Online Campus, Ashford University, Capella University, Grand Canyon University, and Kaplan University-Davenport Campus. There is one outlier here which is the University of Phoenix-Online Campus which seems to make a great deal more revenue than the others which seem quite near to each other in terms of their total net income. The reason for this is likely that as an online school it can enroll many more students at one time for certain courses, giving them an ability to receive much more revenue from these online students, and the money can be focused on improving their online experience rather than building infrastructure or maintaining a campus. Much of the teaching for these classes can likely also be done in a large online classroom setting, giving more profit for enrolled students in comparison to fewer teachers being hired.

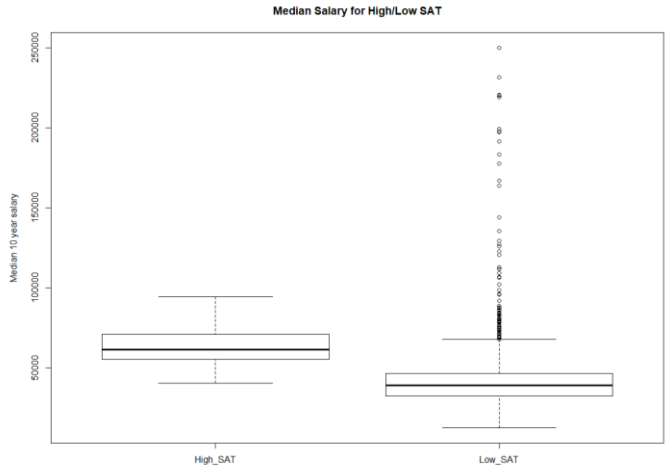
9. Part a) Below is a plot of average SAT versus admissions for colleges:



There seems to be a density around 800-1200 for average SAT scores which make up most of the colleges. An appropriate cutoff for the data from this plot seems to be around 1200 for SAT scores. For admissions it seems to be below 0.4 when it narrows. Therefore, the variable *group* can be created by distinguish low SAT scores (<1200) versus high SAT scores (>= 1200) and admissions which are (<=0.4) or (>0.4). As admissions become much narrower in the 1200+ range, there is a narrowing of colleges, and this seems to imply that the admissions rate for colleges become less varied, or in a sense are much more difficult to get accepted into depending on the SAT score of the students that enroll.

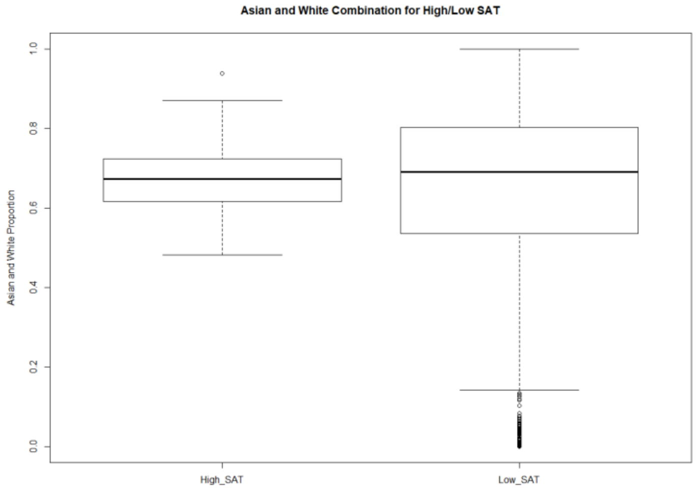
Part b) Below is a boxplot of the median 10 year salary with the 'group' variable:

(a)



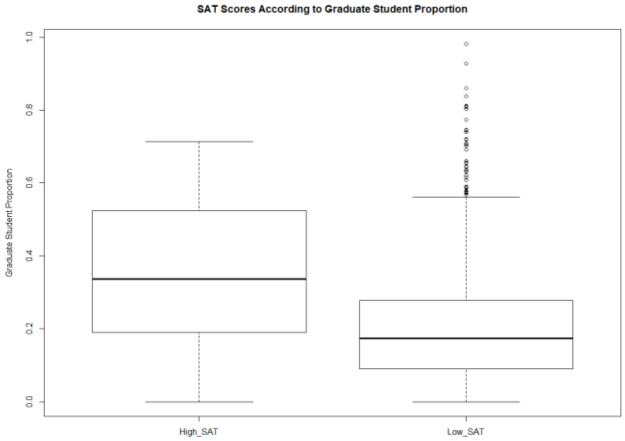
The 10 year median salary for students graduating from high SAT colleges tend to earn more than those with lower SAT's. There are still outliers in the low SAT colleges which make it so that their students can also make higher salaries. It is still more common however for high SAT college graduates to be earning more than their counterparts.

(b) Below is a boxplot of the percentage of Asians and Whites combined:



Colleges which have lower proportions of Asian and Whites can also have lower SAT scores. In high SAT average schools, this sort of pattern is not common and doesn't become apparent.

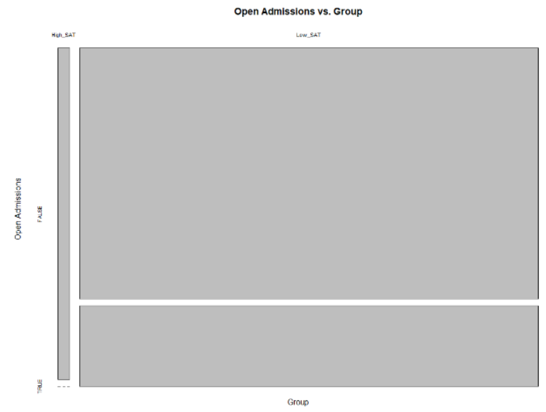
(c) Below is a boxplot of the percentage of graduate students enrolled:



It follows that colleges with higher SAT's will have a greater proportion of Graduate students than colleges that have a lower SAT average. However, there are some outliers in the lower SAT range. The exceptional outliers in the low SAT group seem to come from colleges which are focused on giving graduate level degrees, which explains how their student population can have both lower SAT's on average while still consisting mainly of graduate students.

Part c) (in the following mosaic plots the table() function was used to check whether a variable had 0 for a particular category, e.g., in the next plot Open Admissions which are High_SAT have 0 true values)

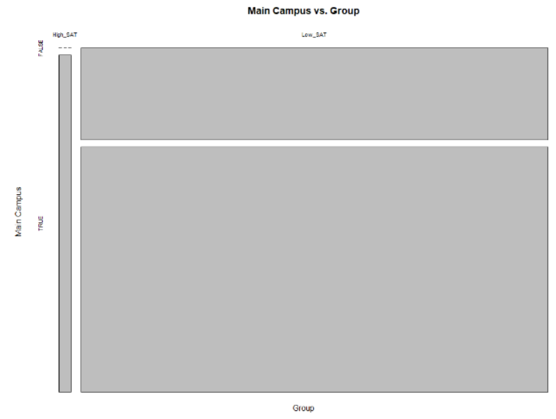
(a) Open admissions and group are plotted below using mosaic plot:



From the mosaic plot it becomes clear that if a school has a high SAT average, then it doesn't have open admissions. This makes sense, since those high SAT schools tend to be more exclusive and require high SAT's for acceptance. This wouldn't necessarily have to apply for schools with low SAT averages and accept most people who apply. From the table and plot it's clear that the two ratios are quite different.

	FALSE	TRUE
High_SAT	83	0
Low_SAT	2,445	784

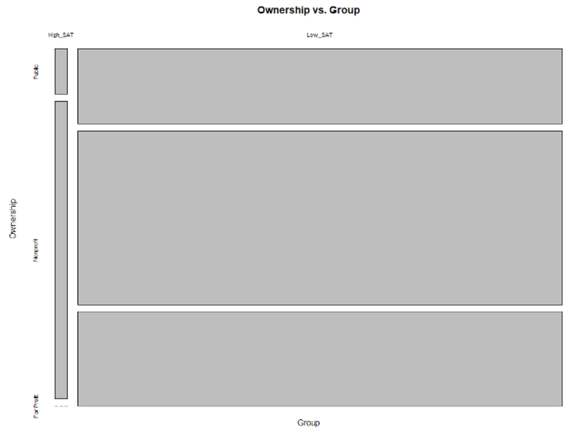
(b) Main campus and group are plotted using mosaic plot:



Here it's apparent that if a school is the main campus for a college, then it'll have a high SAT average which is an interesting observation. This seems to imply that schools which have high SAT averages are never off-campuses for a college. Although this may not have any special meaning, it is strange that no off-campuses have a high SAT average, despite which main campus it may come from. So clearly here, the main campus and group variable are dependent. From the table and plot it's clear that the two ratios are quite different.

	FALSE	TRUE
High_SAT	0	83
Low_SAT	881	2,348

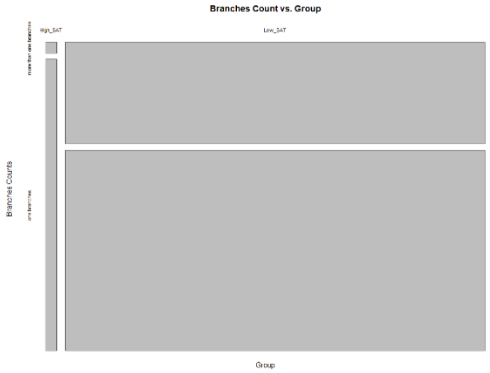
(c) Ownership is plotted below against Group:



Here it becomes apparent that high SAT average schools are either Nonprofit or Public universities. Most them which have high SAT averages are actually Nonprofit schools, while a much smaller proportion belong to the Public university system. For low SAT average schools, they can be either of the three, with about an equal amount being either Public or For Profit. Still the largest portion is the Nonprofit for low SAT schools. So, in this case again the two variables ownership and group are dependent on each other. From the table and plot it's clear that the three ratios are quite different.

	For Profit	Nonprofit	Public
High_SAT	0	72	11
Low_SAT	886	1,638	705

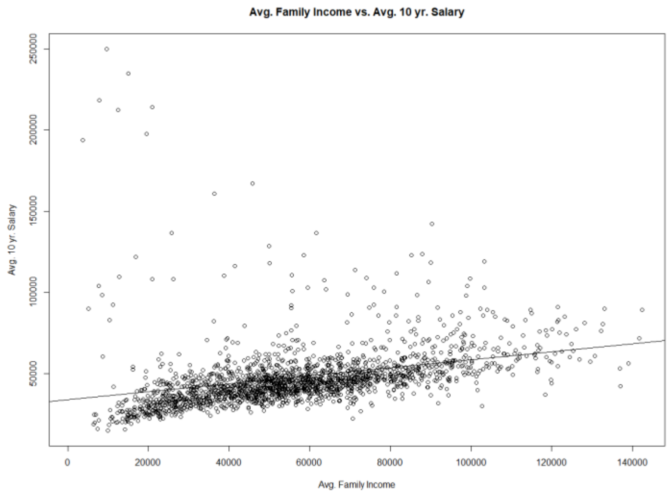
(d) Multiple branches are plotted against Group below:



According to the mosaic plot, it seems that there is bias towards high SAT schools having only one branch, while low SAT schools are more likely to have more than one branch. Therefore, the fact that a school has one or multiple branches is dependent on the school belonging to a high or low SAT group. From the table and plot it's clear that the two ratios are quite different.

	1+ branches	1 branch
High_SAT	3	80
Low_SAT	1,087	2,142

10. Part a) Below is a plot of Avg. Family Income against Avg. 10 Year Salary:



There is also a regression line through the data which is created by predicting the average 10 year salary according to average family income. After looking at the regression line, it seems that there are a few points in the top left region of the plot which may be affecting the usefulness of the regression line. A possible cutoff region to separate these pieces of data would be when the average 10 year salary is greater than 150,000. There are only 9 of these data points, after creating a variable for the outliers, and checking the number of observations. After looking at the 9 data points in that region it is clear from the names they all seem to be medical-related colleges.

Part b) Looking further into the data it seems that all of them have a Graduate degree as their highest available degree, and all except two of them have a Graduate program as their primary degree offered. This would help explain the high salary, since it is sensible to conclude that students graduating from medical schools can obtain high salary jobs despite their family's income. An interesting variable that could be used to identify a variable to help improve the fit is 'highest_degree' which has 5 factor levels: Other, Certificate, Associate, Bachelor, and Graduate.

References:

http://alumni.media.mit.edu/~tpminka/courses/36-350.2001/lectures/day12/

https://stackoverflow.com/questions/7706876/remove-na-values-from-a-vector

https://stackoverflow.com/questions/5577727/is-there-an-r-function-for-finding-the-index-of-an-element-in-a-vector

https://www.r-bloggers.com/quartiles-deciles-and-percentiles/

https://stackoverflow.com/questions/8161836/how-do-i-replace-na-values-with-zeros-in-an-r-dataframe

Patrick Vacek (Piazza, @33, @35, @37, @62, @63, @64, @73, @76, @77)

Shuva Gupta (lecture 4/5, 4/12. 4/17)

Code Appendix:

```
setwd("C:/Users/qizhe/Desktop/STA 141A/HW1") # set working directory

cs <- readRDS("college_scorecard_2013.rds") # give a name to the data

# 1 Total of 3312 observations, 2431 different colleges.
dim(cs) # 3312 observations
# number of colleges
sum(cs$main_campus) #2431 main campuses

# 2
# examine the number of different classes of variables
tab1 <- sapply(cs, class)
table(tab1)

# examine each variable type using str()
str(cs)

# 3
sum(is.na(cs)) # number of NA's 23197

# find number of NA's per variable, find the largest
count_na_by_features = sapply(cs, function(x) sum(is.na(x)))
t(t(count_na_by_features)) # avg_sat is largest with 1,923

# graphically display the NA's per feature
NA_counts<-colSums(is.na(cs)) # save NA's to variable
plot(NA_counts, main="Number of Missing Data per Feature",
     xlab = "Feature Index",
     ylab = "Number NA's", cex = 1, pch=15)
abline(h=c(500, 0), col="Green") # many around this region

# 4
# number of private versus public colleges
length(which(cs$ownership=="Public")) # public 716
length(which(cs$ownership!="Public")) # private aka For Profit and Nonprofit
2,596

# separate public and private colleges into variables
cs_public = cs[cs$ownership=="Public",]
cs_private = cs[cs$ownership!="Public",]

# graphically display proportions of degrees for private and public colleges
public_highest_deg_table = table(cs_public$highest_degree)/sum(table(cs_publi
c$highest_degree))
private_highest_deg_table = table(cs_private$highest_degree)/sum(table(cs_pri
vate$highest_degree))
par(mfrow=c(1,2))
barplot(public_highest_deg_table, main = "Degree Proportion in Public College
```

```r
", ylim=c(0,0.8))
barplot(private_highest_deg_table, main = "Degree Proportion in Private Colle
ge", ylim=c(0,0.8))
dev.off() # reset frames

# 5
# mean, median, and decile for undergraduate populations
ugmean <- mean(cs$undergrad_pop,na.rm = TRUE) #3599.502
ugmed <- median(cs$undergrad_pop,na.rm = TRUE) #1295
decile <- quantile(cs$undergrad_pop, prob = seq(0,1,length=11),type = 5,na.rm
 = TRUE)
hist(cs$undergrad_pop, main = "Histogram of undergraduate population", xlab =
 "Undergraduate Population")
abline(v = ugmean, col='red', lwd=5)
abline(v = decile, col='pink', lwd=2)
abline(v = ugmed, col='blue', lwd=5)
legend(x=120000, y=2400, c("mean", "median", "decile"), col = c("red", "blue
", "pink"), lty=c(1,1,1))
#There is an extreme outlier, the Universit of Phoenix, Arizona whose populat
ion of over
#150,000 students creates an extreme skew to the histogram.

# testing for without the previous outlier, little difference in overall skew
cs1 = cs[-which(cs$undergrad_pop==166816),] # removes outlier
ugmean <- mean(cs1$undergrad_pop,na.rm = TRUE) #3599.502
ugmed <- median(cs1$undergrad_pop,na.rm = TRUE) #1295
decile <- quantile(cs1$undergrad_pop, prob = seq(0,1,length=11),type = 5,na.r
m = TRUE)
hist(cs1$undergrad_pop, main = "Histogram of undergraduate population")
abline(v = ugmean, col='red', lwd=5)
abline(v = decile, col='pink', lwd=2)
abline(v = ugmed, col='blue', lwd=5)
legend(x=120000, y=2400, c("mean", "median", "decile"), col = c("red", "blue
", "pink"), lty=c(1,1,1))

# 6
# 5 most populous states: California, Texas, New York, Illinois, Florida
top = cs[cs$state %in% c("CA", "TX", "NY", "IL", "FL"),]
top$state <- droplevels(top$state)
boxplot(top$tuition~top$state, main = "Boxplot of 5 Most Populous States", yl
ab = "Tuition per State")

# 7
# part a
# name of university with largest avg_sat
cs[which.max(cs$avg_sat), 'name'] #California Institute of Technology

# part b
# largest university & open admissions
cs[which.max(cs$undergrad_pop), c('name', 'open_admissions')]
```

```r
# part c
# zip code of smallest avg_family_inc for public schools
publicUni = subset(cs,ownership=="Public") # subset by public
publicUni$zip[which.min(publicUni$avg_family_inc)] # 11101

# part d
# also largest grad_pop (referring to part b)
cs[which.max(cs$undergrad_pop), 'grad_pop'] == max(cs$grad_pop, na.rm=TRUE) # checks if part b is the max grad pop

# 8
# subset For Profit & Bachelor primary degree schools
profbach = subset(cs, ownership=="For Profit" & primary_degree=="Bachelor")
money_per_student <- profbach[,c('revenue_per_student', 'spend_per_student')]

# part a
plot(money_per_student, main = "Revenue versus Spending per Student", xlab = "Revenue per student"
     , ylab = "Spending per student")
#There are a few outliers which may violate the assumptions of linearity.

# part b
# create new variable to use
profbach[is.na(profbach)] <- 0 # set NA's to 0
net_income <- (profbach$revenue_per_student - profbach$spend_per_student) # net income for colleges
total_net_income <- (net_income*(profbach$undergrad_pop + profbach$grad_pop)) # total net income
# order profbach, and then choose top 5
profbach_top = profbach[order(total_net_income, decreasing=TRUE),]
profbach_top5 = profbach_top[1:5,] #University of Phoenix-Online Campus, Ashford University,
#Capella University, Grand Canyon University, Kaplan University-Davenport Campus

# plot net income and number of students
# create total student population, net income, and total net income categories for sub8_top5
profbach_top5$student_pop <- profbach_top5$grad_pop + profbach_top5$undergrad_pop
profbach_top5$net_income <- (profbach_top5$revenue_per_student - profbach_top5$spend_per_student)
profbach_top5$total_net_income <- (profbach_top5$net_income*(profbach_top5$undergrad_pop + profbach_top5$grad_pop))
plot(profbach_top5$student_pop, profbach_top5$total_net_income, xlab = "Student population",
     ylab = "Total income per student", main = "Top 5 Total Net Income for Colleges",
     pch=1:5)
```

```r
legend("topleft", profbach_top5$name, pch=1:5)

# 9
plot(cs[,c("avg_sat", "admission")], main = "Average SAT versus Admissions",
xlab = "Average SAT",
     ylab = "Admissions")
#Around avg_sat = 1200, the admission begins to narrow.

# split according to avg_sat >= 1200 and admission <= 0.4
group <- ifelse(cs$avg_sat >= 1200 & cs$admission <= 0.4, "High_SAT", "Low_SA
T")
group <- ifelse(is.na(group),"Low_SAT",group)
cs$group <- group
boxplot(cs$med_10yr_salary~cs$group, ylab = "Median 10 year salary", main = "
Median Salary for High/Low SAT")

cs$race_aw <- cs$race_asian + cs$race_white # create asian and white variable
boxplot(cs$race_aw~group, ylab = "Asian and White Proportion", main = "Asian
and White Combination for High/Low SAT")

cs$grad_proportion <- (cs$grad_pop/(cs$grad_pop + cs$undergrad_pop)) # create
 proportion variable
boxplot(cs$grad_proportion~group, ylab = "Graduate Student Proportion",
        main = "SAT Scores According to Graduate Student Proportion")

lsat=subset(cs, cs$group=="Low_SAT") #low SAT's
lsat$grad_proportion <- (lsat$grad_pop/(lsat$grad_pop + lsat$undergrad_pop))
lsat[order(lsat$grad_proportion, decreasing = TRUE),][1:5,]

# part C
mosaicplot(~  group + open_admissions, data = cs, main = "Open Admissions vs.
 Group",
           xlab = "Group", ylab = "Open Admissions")
table(as.character(cs$group), as.character(cs$open_admissions))

# examine table of 'group' and 'open_admissions, also view the distribution'
subgroup = cs$group
subgroup[is.na(subgroup)] = "NA"
suboa = cs$open_admissions
suboa[is.na(suboa)] = "NA"
table(suboa,subgroup)
subas = cs$avg_sat
subas[is.na(subas)] = "NA"
hist(cs$avg_sat)

# part C, part B
mosaicplot(~  group + main_campus, data = cs, main = "Main Campus vs. Group",
           xlab = "Group", ylab = "Main Campus")
table(as.character(cs$group), as.character(cs$main_campus)) # High_SAT FALSE
= 0
```

```r
# part C, part C
mosaicplot(~  group + ownership, data = cs, main = "Ownership vs. Group",
           xlab = "Group", ylab = "Ownership")
table(as.character(cs$group), as.character(cs$ownership)) # High_SAT For Profit = 0

# part C, part D
branches_counts = ifelse(cs$branches == 1, "one branches", "more than one branches")
mosaicplot(~ group + branches_counts, data=cs, main = "Branches Count vs. Group",
           xlab = "Group", ylab = "Branches Counts")
cs$branches_counts <- branches_counts
table(as.character(cs$group), as.character(cs$branches_counts)) # High_SAT For Profit = 0

# 10
# part A
plot(cs$avg_family_inc, cs$avg_10yr_salary, main = "Avg. Family Income vs. Avg. 10 yr. Salary",
     xlab = "Avg. Family Income", ylab = "Avg. 10 yr. Salary")
abline(lm(avg_10yr_salary~avg_family_inc, data=cs))

outliers = cs[cs$avg_10yr_salary > 150000 & !is.na(cs$avg_10yr_salary) & !is.na(cs$avg_family_inc),]
dim(outliers) # 9 observations
outliers$name

# part B
# check the number of factor levels
levels(cs$highest_degree)

# optional R/Stats practice, test the new categorical variable's factor levels to see
# if there's an improvement of fit
plot(cs$avg_family_inc, cs$avg_10yr_salary, main = "Avg. Family Income vs. Avg. 10 yr. Salary",
     xlab = "Avg. Family Income", ylab = "Avg. 10 yr. Salary")
fit1 = lm(avg_10yr_salary~avg_family_inc + highest_degree, data=cs)
fit1$coef
abline(a = fit1$coef[1], b = fit1$coef[2], col="orange")
abline(a = fit1$coef[1]+fit1$coef[3], b = fit1$coef[2], lty=2, col="pink")
abline(a = fit1$coef[1]+fit1$coef[4], b = fit1$coef[2], lty=3, col="green")
abline(a = fit1$coef[1]+fit1$coef[5], b = fit1$coef[2], lty=4, col="blue")
abline(a = fit1$coef[1]+fit1$coef[6], b = fit1$coef[2], lty=5, col="red")
abline(lm(avg_10yr_salary~avg_family_inc, data=cs))
legend("topright", c("other", "Certificate", "Associate", "Bachelor", "Graduate",
                     "Original regression line"),
       col = c("orange", "pink", "green", "blue", "red", "black"), lty = c(1,
2,3,4,5,1))
```