STA 141A Data Analysis Report

1.

Total number of observation	3312
Number of observed colleges	2431

By extracting data from the original dataset, we obtain the recorded observation number 3312 and the recorded colleges number 2431.

2.

Total feature number	Character	Factor	Integer	Logical	Numeric
51	4	4	15	3	25

By using the **table()** function in R, we see that R groups the 51 data features into 5 data types as the above table. Among them, 4 factor variables are considered categorical features, 15 integer variables are considered discrete features.

However, this grouping is from the R point of view. By using **str()** to see the internal structure of the data set and checking the variable descriptions, we can adjust some of the variable types as follows:

(1) *unit_ID* should be character feature just as *ope_ID* and *zip* because each ID doesn't really have that value in it, the numbers in an ID are used to represent identity.

(2) Though in the data set, *avg_sat*, *avg_faculty_salary*, *avg_10yr_salary*, *sd_10yr_salary*, and, *med_10yr_salary* are considered integer features, they should be considered numeric features just as *avg_entry_age*, *avg_family_inc*, *med_family_int*, *med_debt*, and, *med_debt_withdraw* because they are means, standard deviations or medians which usually contains floating points since it's hard to compute each of these as integers.

Therefore, we consider that among 51 features, the new feature numbers table should be:

Total feature number	Character	Factor	Integer	Logical	Numeric
51	5	4	9	3	30

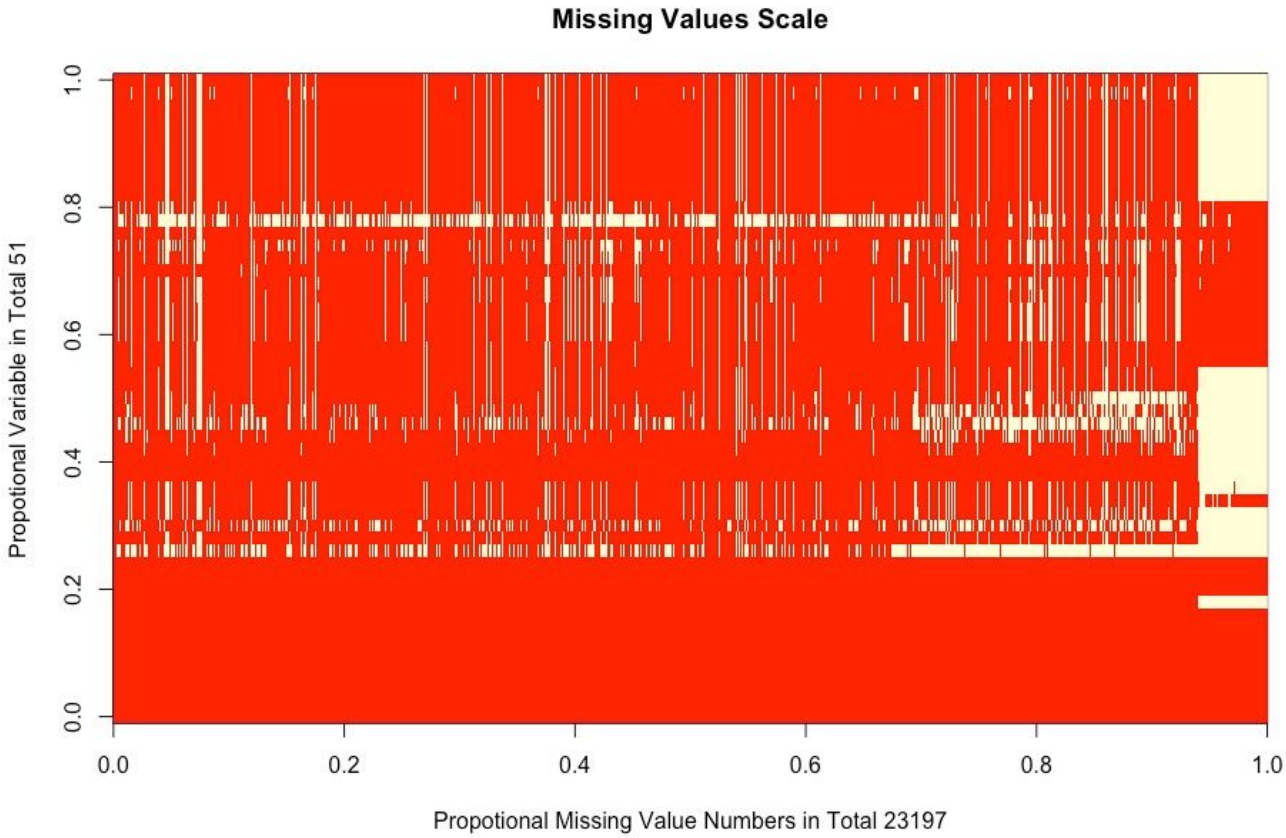
There are 4 categorical and 9 discrete features, the other features are character, logical and numeric.

3. By using **is.na()**, **which()**, **length()** functions in R, we find that there are 23197 missing values.

By extracting the variable name of which having the maximum missing value, we find that *avg_sat* has most missing values, 1923. And the first graph below shows that the most missing value is above 1500 and locates at feature index 15, which means *avg_Sat* as well. This graph also shows that mostly the features indexed less than 15 have no missing value (see red line) and almost half of the other features have around 500 missing values (see blue line), especially those indexed around 30 to 50.

The second graph below is the missing value scale. We can see the missing value pattern from it: Looking at the y-axis, at around column variable proportion(1 means the 51st variable) 0.25, 0.3, 0.45, 0.79, it seems that there are continuously missing values; especially at 0.25 proportion, values are missing starting from around where the 70th percentages of total observations until the 100th percentages.

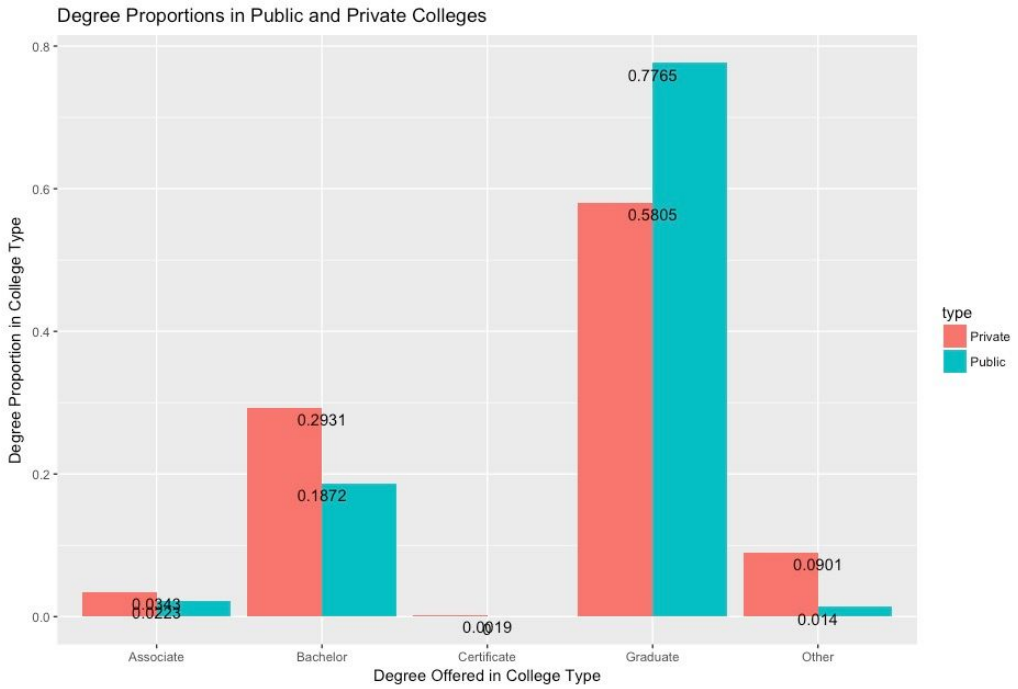
Looking at the x-axis, we find that the yellow solid areas are at around 0.95 to 1 in proportional variables 0.25 to 0.55 and 0.8 to 1, which means that in our dataset, approximately the last 5% of the observations in variables in the middle 30% columns and last 20% columns contain almost all missing values.



4.

Public	Nonprofit	For Profit	Private (Nonprofit + For Profit)
716	1710	886	2596

By using **table()** function in R, the results show as the above table that there are more private colleges recorded. Below is the plot showing different degrees offered in public and private colleges with its proportion in the certain college type. It shows that the proportions of highest degree awarded (graduate) in public college is 77.65% and in private college is 58.05%. And, we observe from the graph that public colleges don't offer certificate degree, and the proportions for associate and other degrees are similar; while private colleges offer five degrees with the proportion order from largest to smallest: graduate, bachelor, other, associate, certificate. We also notice that public colleges offer graduate degree in a larger proportion than private colleges, and private colleges offer other degree types in larger proportions than public colleges.



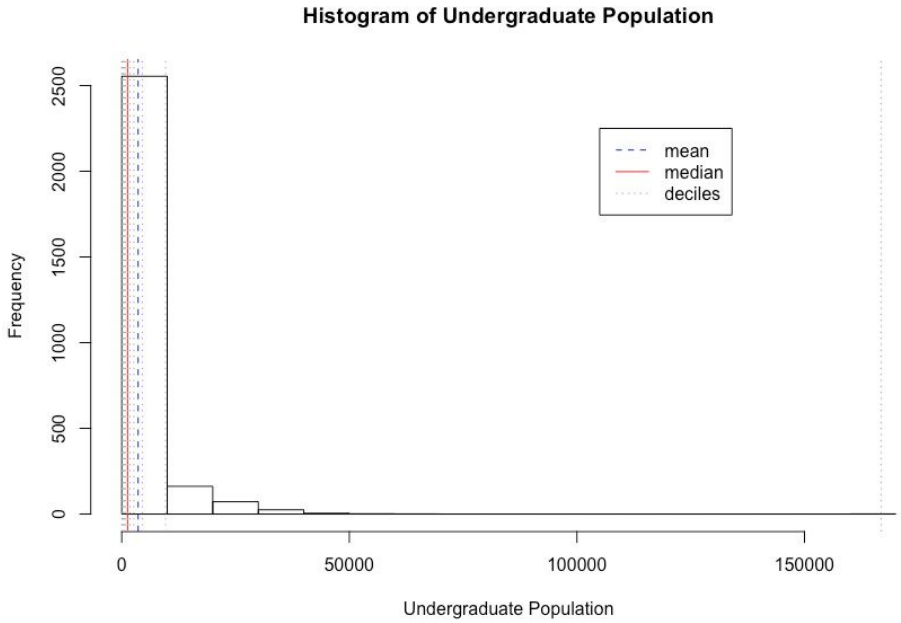
5.

By extracting results from the summary() output, we obtain the average undergraduate population is 3599.502, the median is 1295, and the deciles are:

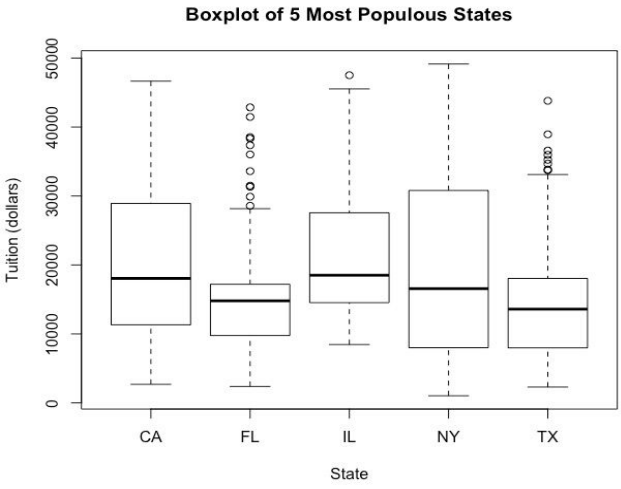
0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0.0	153.0	319.2	536.0	847.6	1295.0	1811.8	2674.5	4550.8	9629.8	166816.0

Below is the histogram of undergraduate population with labels at mean, median and the deciles. The histogram is right skewed with a long tail towards over 150000. Most undergraduate population values seem concentrate on the range from 0 to 50000. We look at the original dataset, among 2431

observations, we might think that the undergrad_pop = 166816 is probably an outlier affecting the scale of the histogram skewness and all the other informative values.



6.



Above is the boxplot of the 5 most populous states in tuition. As we can see from the plot, each box represents a state's tuition distribution. Since the boxes have different shapes, these 5 states have different tuition distributions. FL and TX seem to have a smaller box ranging around 10000 to 20000, meaning their tuitions are mainly at lower range than the other states. While CA, FL, NY and TX have approximate the same first quantile value (bottom short lines), IL has it at tuition around 10000. Also, 5 states all have approximate median at 20000, but CA and IL are closer, while FL, NY and TX are closer (middle short bold lines). FL has a lowest third quantile value at around 30000, and TX also has low third quantile value compared to the others. And we can see that these two states have some outliers whose tuition values are above their third quantiles, meaning that the distribution of FL and TX tuitions are right skewed. IL has one outlier shortly away from its third quantile. The third quantile values of CA and IL are similar around 47000, NY has an almost 50000 dollars tuition third quantile. CA and NY have large range between first and third quantiles, meaning that the tuition range vary a lot in these two states.

7.

(a). By using code, **data[which.max(data$avg_sat), "name"]**, where data is my dataset name, we find that the name of the university with the largest value of *avg_sat* is California Institute of Technology.

(b). By using code, **data[which.max(data$undergrad_pop), "open_admissions"]**, we know that the university with the largest amount of *undergrad_pop* have open admissions since this code returns TRUE.

(c). By using the following codes in order:

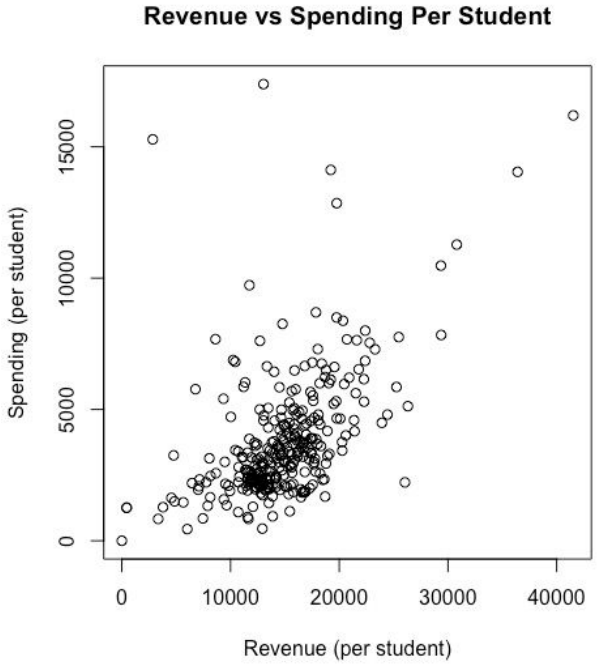
data.filtered <- data[data$ownership=="Public",]
min.index <- which.min(data.filtered$avg_family_inc)
data.filtered$zip[min.index]

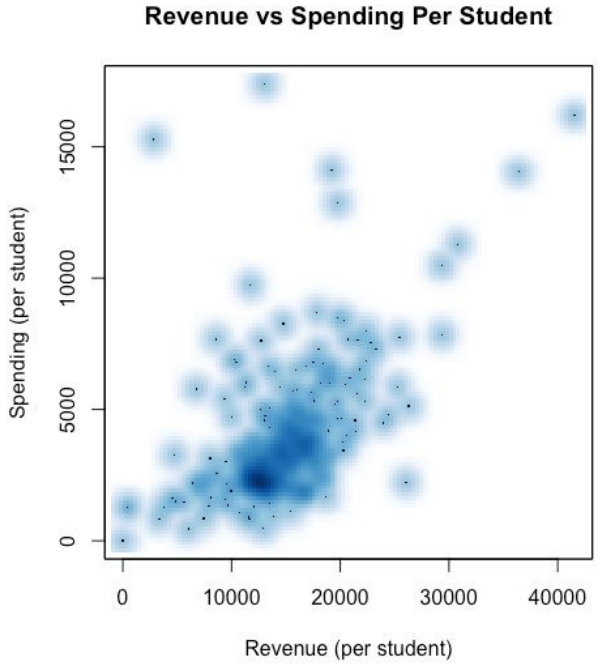
We find that the zip code of the public university with the smallest value of *avg_family_inc* is "11101".

(d). By using code, **data[which.max(data$avg_sat), "grad_pop"]==max(data$grad_pop, na.rm = TRUE)**, which returns FALSE, we find that the university we found in part b doesn't have the largest amount of *grad_pop*.

8.

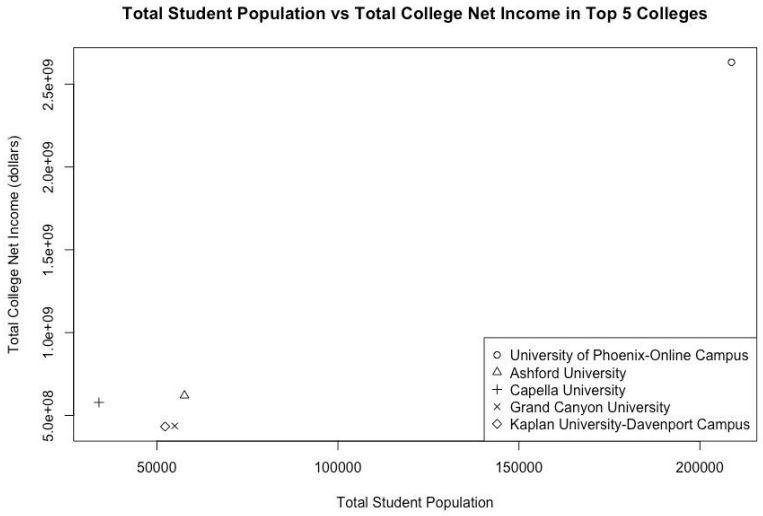
(a). Below is the data plot and smoothscatter plot for revenue_per_student and spend_per_student, as we can see from both plots, mostly as revenue per student increases, spending per student also increases and most data points concentrate on the lower left plot areas where revenue is between 10000 and 20000, spending is between 0 to 5000. We can approximately say that revenue and spending per student has a linear relationship based on these plots, but we can also see that there are a few points far away on the upper left, which are considered outliers and can violate the linearity assumption when fitting a linear regression model .





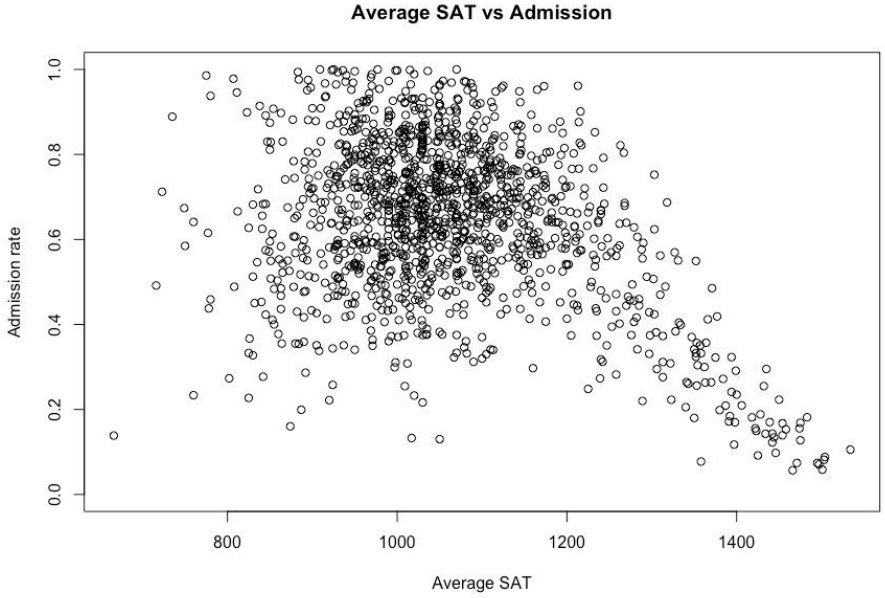
(b).

After creating the new variable called *total_net_income* and some subsetting, we find the top 5 earning schools by order: University of Phoenix-Online Campus, Ashford University, Capella University, Grand Canyon University, Kaplan University-Davenport Campus. On the right is the visualization of them and their total net income.



As we can see from the plot, University of Phoenix-Online Campus is on the top right where student population and total college net income are maximum among these schools. The other schools have similar student populations and college net income, them spreading on the left bottom in the plot.

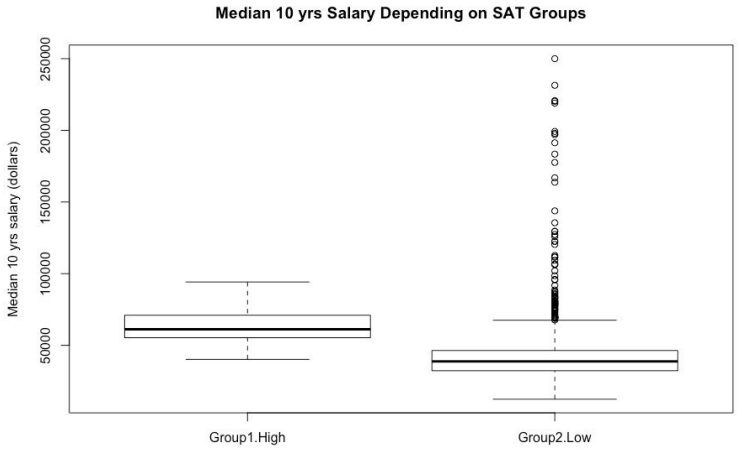
9.



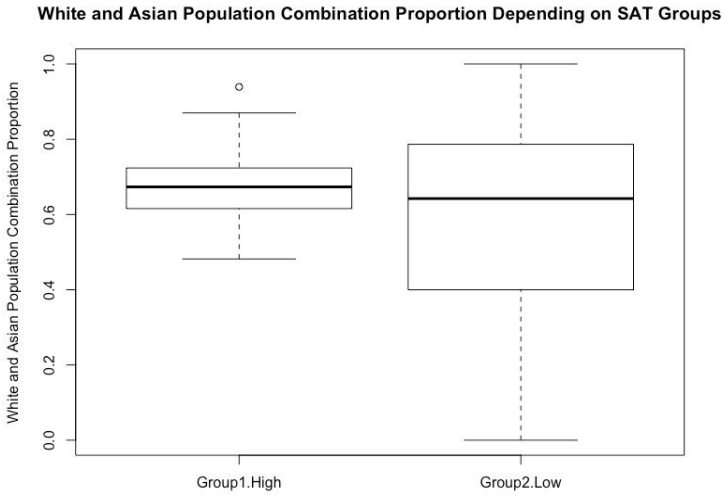
a. Above is the data plot for relationship between average SAT and admission. Here we can see that the data points seems spread evenly at range average SAT scores from 800 to 1200, but as average SAT increases from 1200, admission rate starts to drop down from around 0.4 and become narrow. So we split the data into two groups based on this: group1 to be high SAT group where SAT>=1200 and admission rate <= 0.4, group2 to be low SAT group where the others and NA go into.

b.

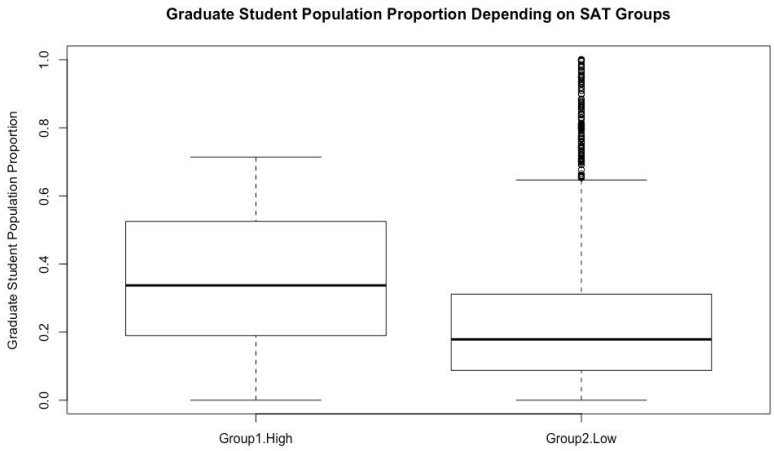
(a) On the right is the boxplot for median 10 years salary change depending on high SAT or low SAT group. We can see that in this dataset, two boxes have similar shape and range width. Yet, those schools which have higher mean SAT scores tend to have a higher median 10yrs salary because the box for high SAT is higher than the one for low SAT. The median of the median 10 yrs salary in Group1.High is above 50000, while in Group2.Low is below 50000. Also, there are some outliers in Group2.Low which means some of those whcih have low mean SAT might also earn high median 10yrs salary.



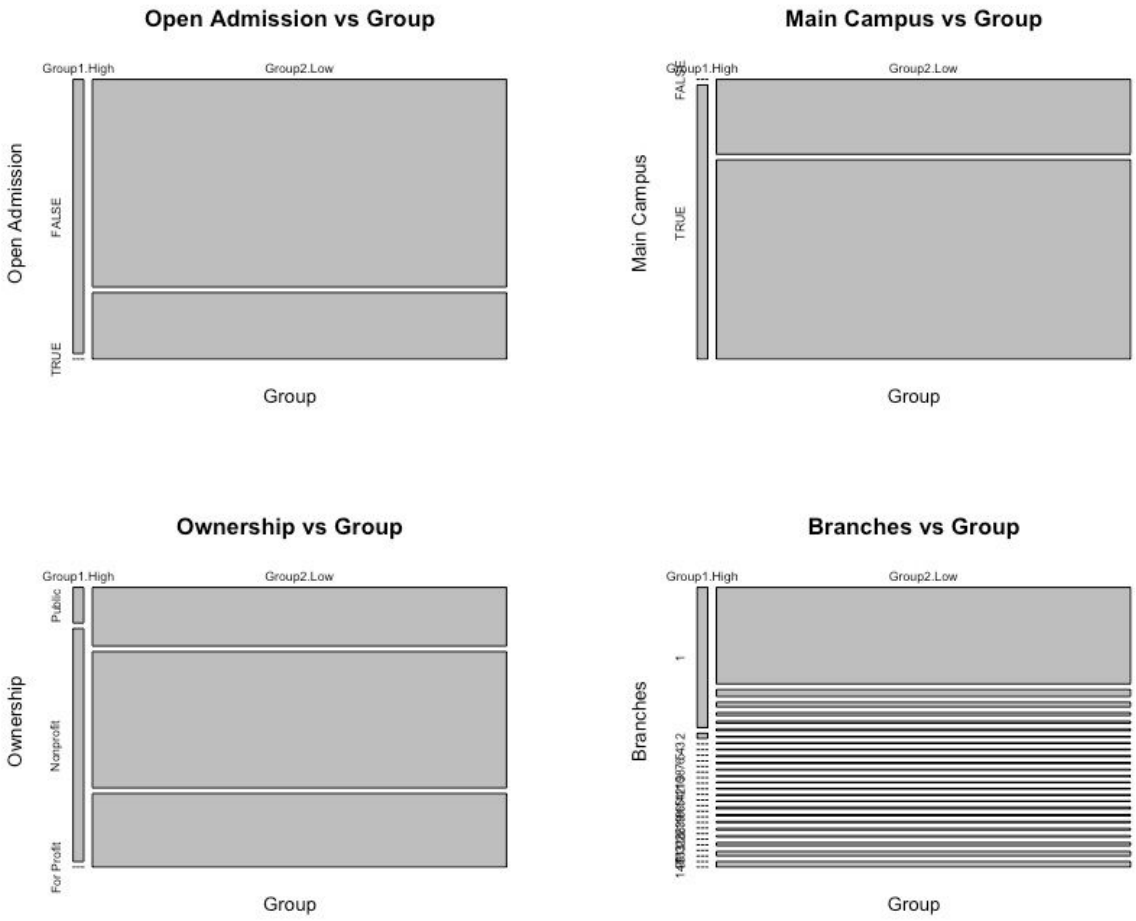
(b) On the right is the boxplot for White and Asian Population combination proportion depending on high SAT or low SAT group. We can see that the box for low SAT group is larger and the range from first quantile to third quantile is also wider. However, the median combination proportion for both groups are similar, around 0.65. There is one outlier in group1. Those schools whose students have low mean SAT don't really have a certain proportion in white and asian population since the range of Group2 is approximately from 0 to 1.



(c) On the right is the boxplot for graduate student population proportion depending on SAT groups. We can see that schools which mean SAT is high tend to have higher proportion of graduate students. And, there are some outliers in the top of the right box, meaning that schools that having a mean SAT as low might also have high proportion in graduate students.



(c)



(a) *open_admission* is independent of *group* because we can see from the mosaic plot that no matter SAT is high or low, open admission still has false as main response.

(b) *main_campus* is also independent of *group* because we can see from the mosaic plot that no matter SAT is high or low, main_campus still has true as main response.

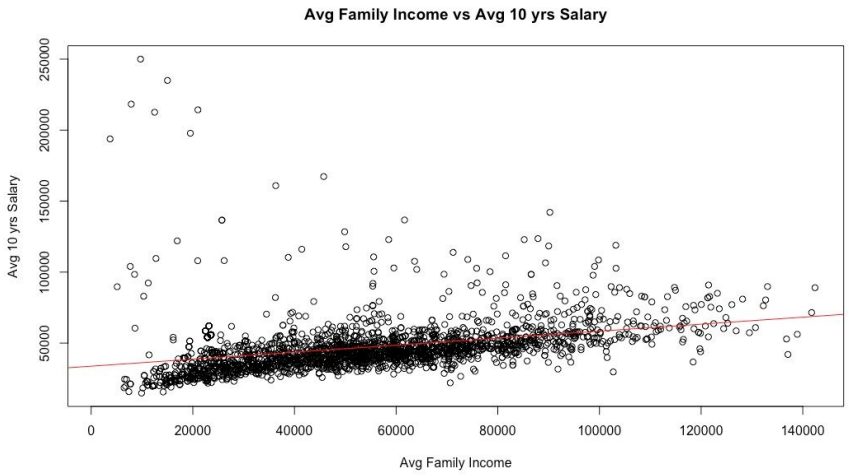
(c) *ownership* is not independent of *group* because we can see from the mosaic plot that in low SAT, the proportion of ownership is by order non profit to for profit to public, while in high SAT, the order is non profit to public to for profit.

(d) Whether the university has more than 1 branch or not is not independent of group because we can see from the mosaic plot that in low SAT, the proportion of only 1 branch is less than half, while in high SAT, the proportion of only 1 branch seems close to half or even more.

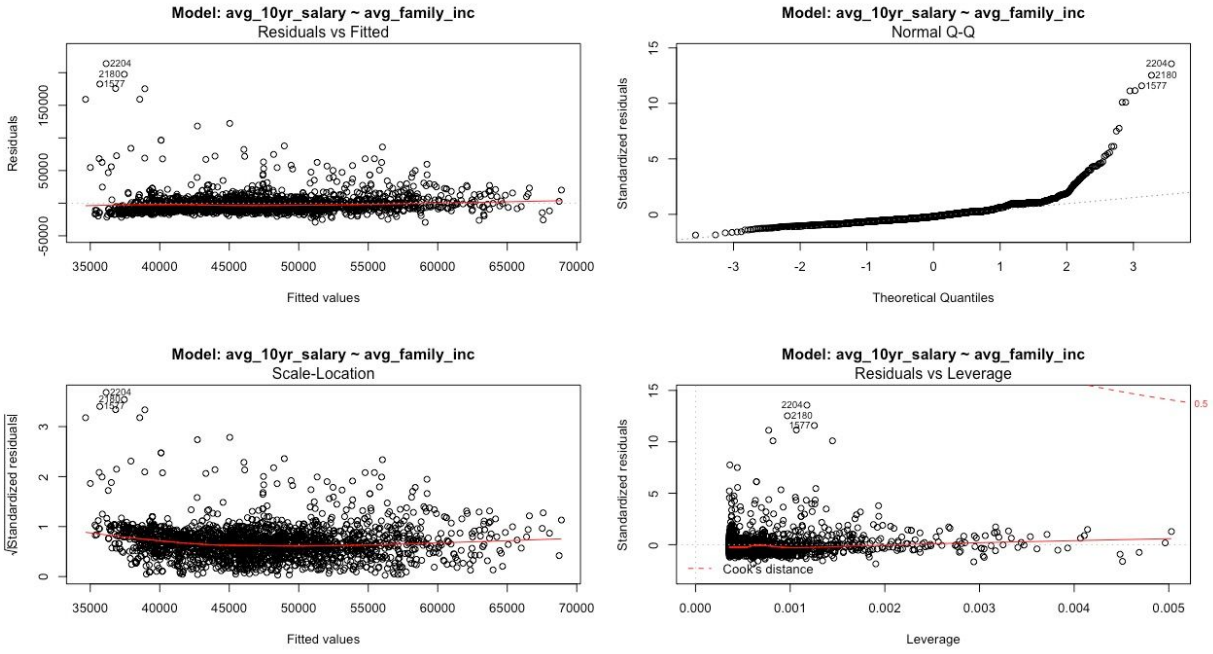
10.

(a) Below is the relationship plot for avg family income and avg 10yrs salary. It seems that the data are approximately linear from the lower left to the higher right. However, there are still lots of points are far away from the red regression line, meaning that the regression model might be violated by some data. We consider those that are furthest away from the main data points as outliers. There

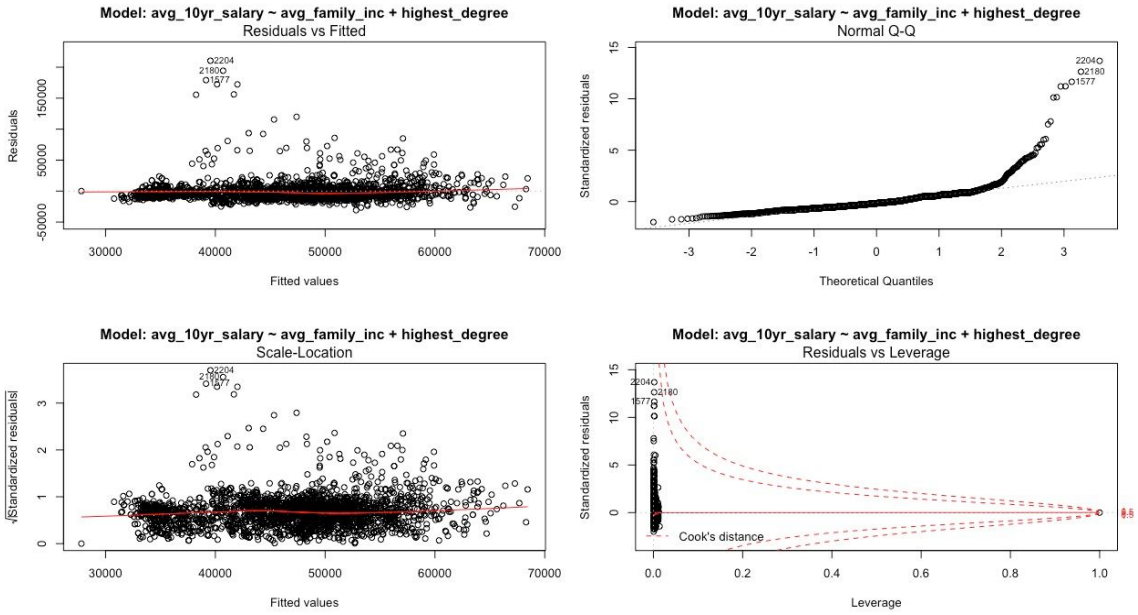
are 9 outliers if we set the cutoff as those who have avg 10 yrs salary higher or equal to 150000 dollars.



(b) There are in total 4 categorical variables in this dataset: *state*, *primary_degree*, *highest_degree*, and *ownership*. And if we look into the information of the outliers, we find that all these outliers schools offer Graduate degree as their highest degree. So we assume that the highest degree factor may affect the linear regression assumption. By adding *highest_degree* into our model fitting, we then compare two regression plots and see how the assumptions change or not.



Before adding *highest_degree* factor into the regression model, we get the above graphs. And after adding it into the regression model, we get the below graphs.



Comparing the two sets of plots, we find that: in the new residuals vs fitted plot, points are more concentrated and the red line is still flat with no pattern, meaning the linearity assumption still holds. The points in the new Normal Q-Q plot don't seem change that much, so adding this factor doesn't affect the normality assumption. In the new Scale-Location plot, the points are more evenly spread along with the red line and the red line is flatter, which means that the equal variance assumption hold better. And last, in the new residuals vs leverage plot, almost all points concentrate on the very left of the plot, and couple points are near the cook's distance lines but not exceed, meaning that there are no outliers within both models. Therefore, we conclude that the *highest_degree* factor does improve the linear regression fit. The regression line fit better as the following plot shows: