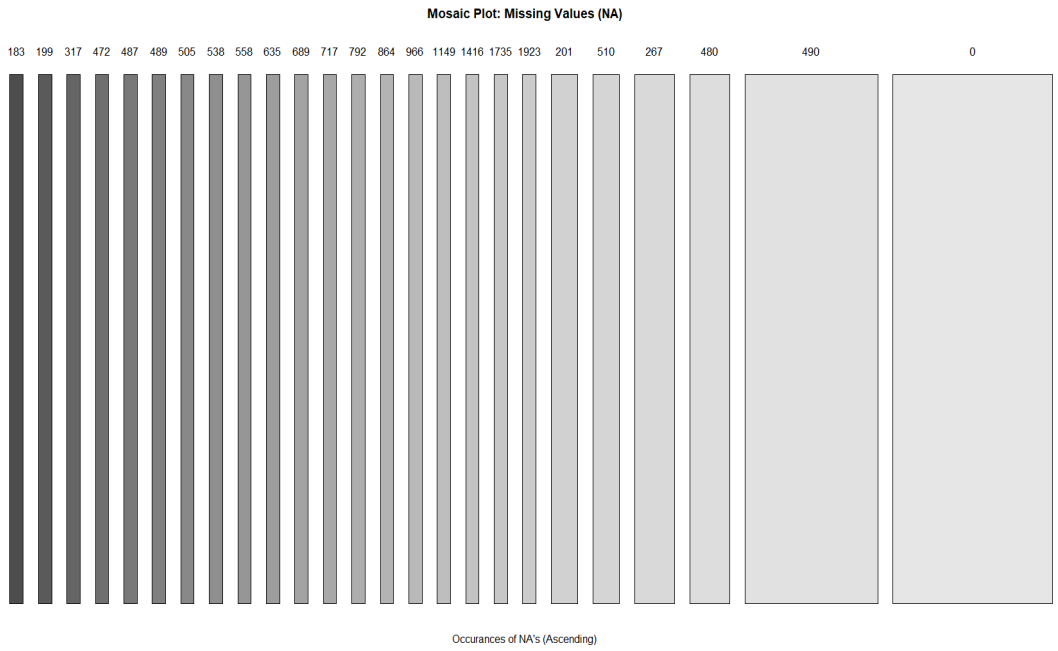
Mitchell Layton

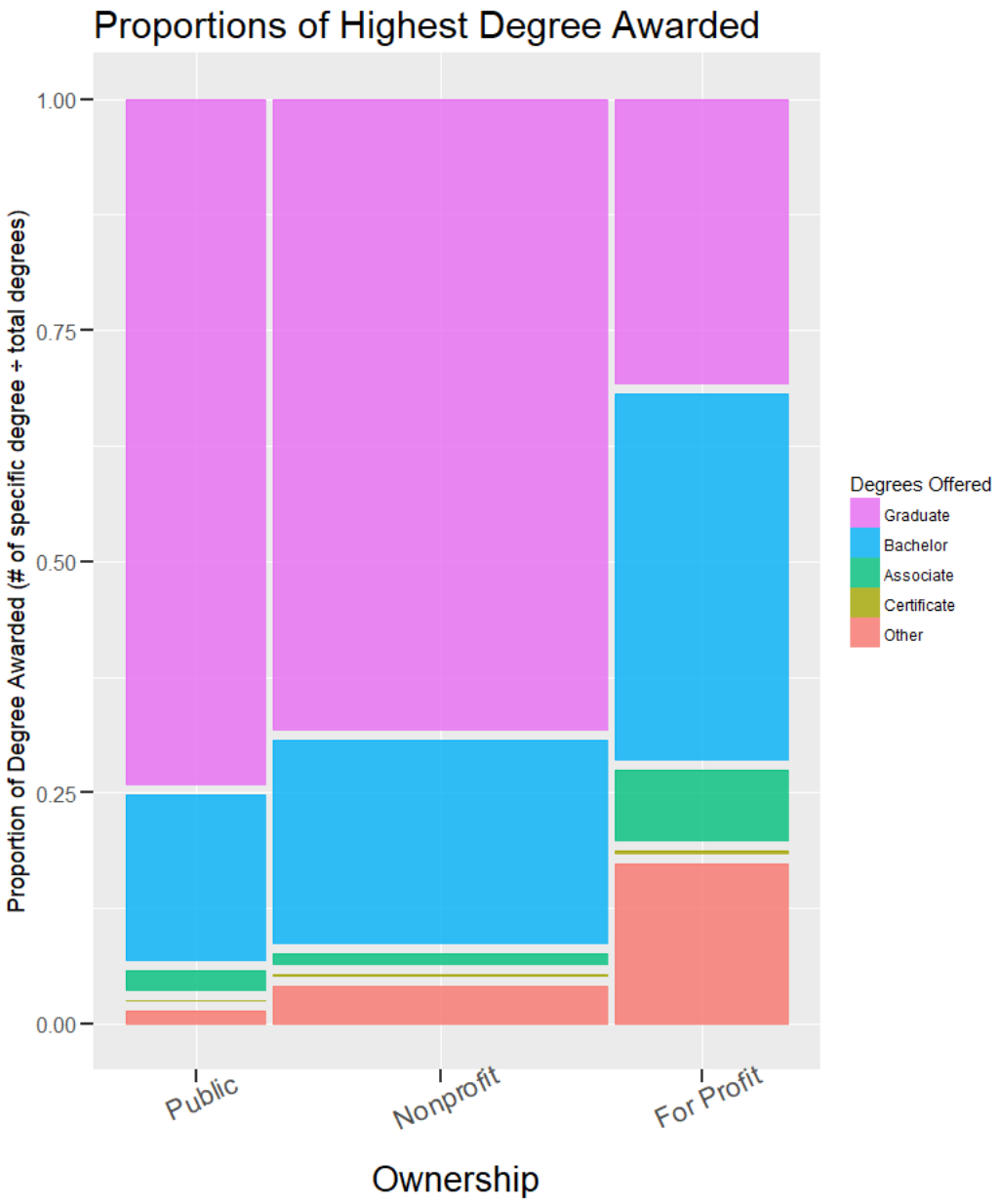
912307956

STA 141A – Homework Assignment 1

1. There are 3,312 rows in the recorded data set “college_scoredboard_2013”. This equates to 168,912 total number of observations in the dataset including any missing values (NA). The number of colleges recorded is 3,312.
2. There are 51 features in our dataset. Of them there are 11 categorical and only 1 discrete. Other notable features to the data set is that most of the data frame consists of numeric or integer vectors that are continuous variables which means the data is very reliant on precise numbers. There are 25 numeric and 15 integer class vectors in the data set.
3. There are 23,197 missing values (NA’s) in the data set and “avg_sat” has the most missing values with 1,923 NA. There are some straightforward trends in the NA’s in our dataset such as the high frequency of 490 NA values which are all prevalent in the male, female, and races columns. Furthermore, there are 4 total column vectors where they have over 1000 missing values that being mean SAT score for students, graduate student population, admission, and veterans.
   ##################FINISH



4. There are more for profit or private colleges (886) recorded compared to public colleges (716). Here is the mosaic plot which analyzes the 3 subgroups in our college data ownership column

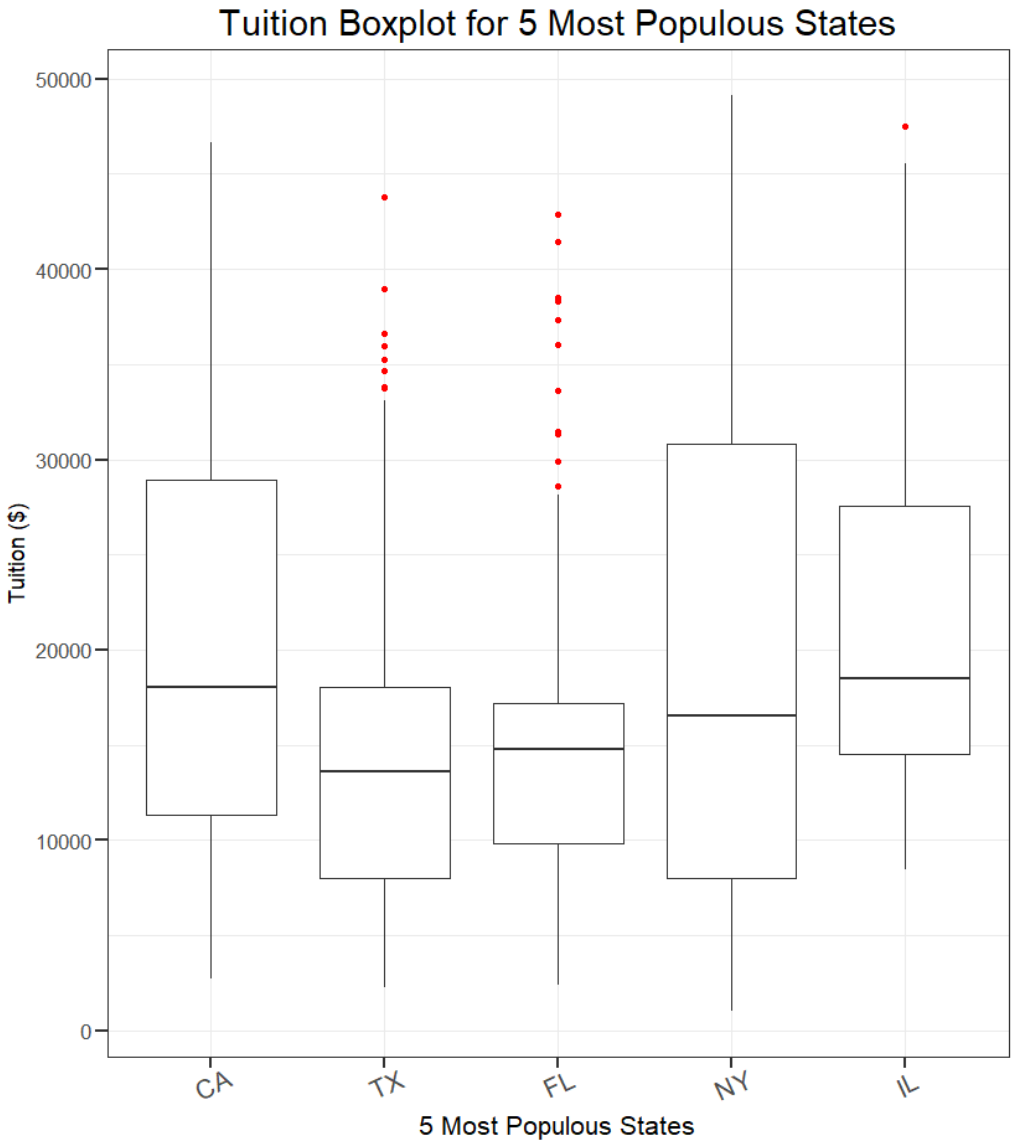


vector. Within each vertical subgroup, the mosaic plot displays a bar graph of constant height and varying width. The width is determined by the number of colleges for each respective subgroup. So with the most is Nonprofit, followed by For Profit and last Public. We also notice, within each column group are 5 subdivided categories representing the proportion of the type of degrees offered. A characteristic that stands out is that public colleges offer the largest proportion of the highest degree (Graduate) while the nonprofit colleges com in a close second, although with a larger amount of colleges. Interestingly, the private colleges offer the largest proportion of "Other" degrees. This could be because private institutions could offer their own certification or program specific to a more intricate or narrowed down field of study. Furthermore, the private colleges column visually seems to have the least amount of proportional variability between degrees offered yet it could be tested statistically to be certain.

5. The average undergraduate population is 3600 and the median undergraduate population is 1295. The deciles are:

10%	20%	30%	40%	50%	60%	70%	80%
153.0	319.2	536.0	847.6	1295.0	1811.8	2674.5	4550.8

90%	100%
9629.8	166816.0



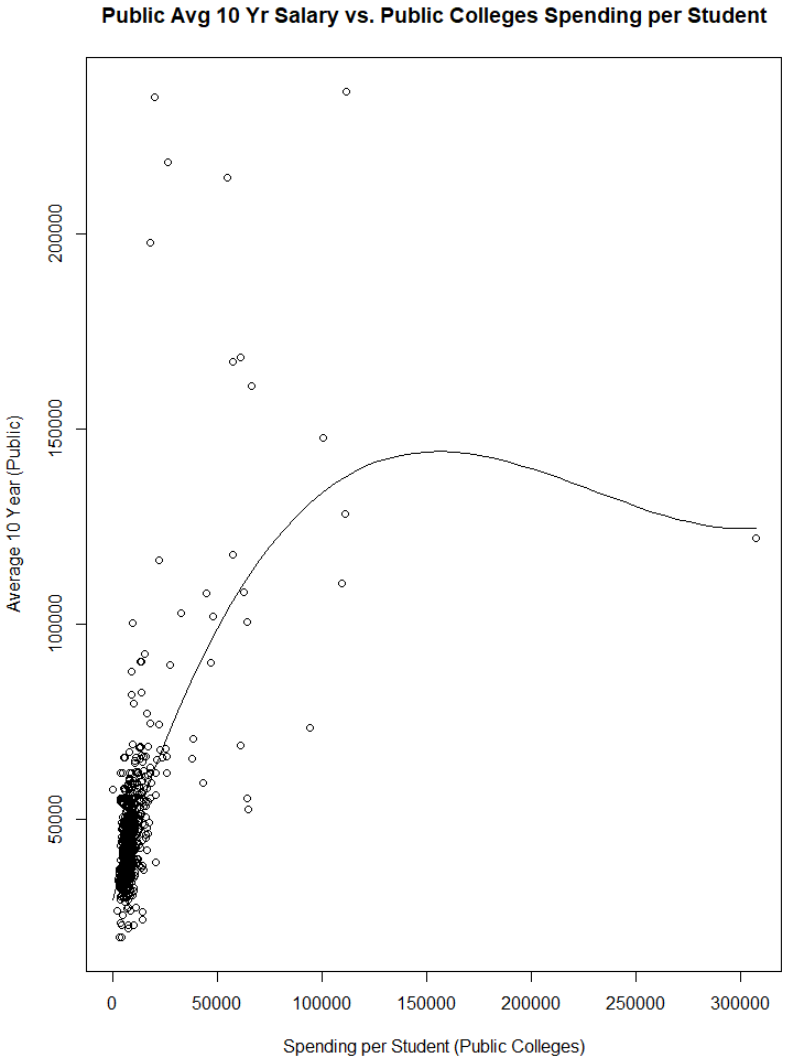
GRAPH ADD LATER

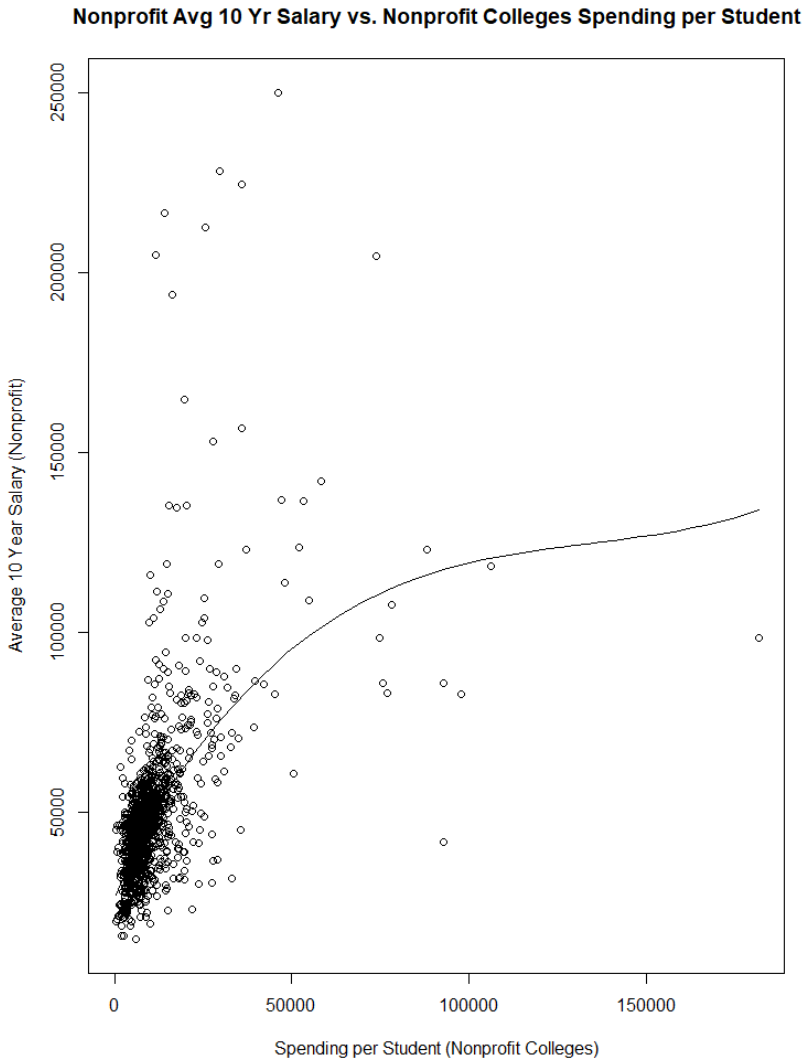
6. I created a boxplot showing the tuition for each of the 5 most populous states according to census.gov (link at bottom of answer #6). I ordered them from left to right from the most populous (CA) to the least populous (IL). The first interesting observation that comes to mind is that TX and FL have the most outliers in the data. I believe this could be due to many reasons. For one, consider that CA, NY, and IL have a good amount of established colleges in both their private and public entities where tuition is no joke to begin with. Consequently, that is why we see each of their average tuition being higher than TX and FL. This leads me to believe that the outliers are there because since they might not have as many established colleges, or networks of colleges such as our UC system, then any private institutions within TX and FL will charge higher tuition both in state and out of state. It becomes like a supply and demand dynamic where you will pay more in TX and FL for those private universities compared to a state like CA because they are much more prevalent and established in CA.

The next observation I look at are the interquartile ranges of each state. NY has the largest IQR and FL has the smallest IQR. We could interpret this as NY having a wider variety of colleges and programs that range from smaller and cheaper certification programs to large private institutions such as Columbia or Cornell which are highly regarded around the world. Comparatively, when you look at the colleges in FL, their highest regarded university is most likely University of Miami, or University of Tampa which do not carry as much weight than the ones in NY. This could be a logical explanation for the large discrepancies in the IQR. These

observations also lead to the high means of TX and FL which are both in the upper quartile. CA, NY, and IL all have means which are in the lower quartiles.

7. Here are the 4 graphical representations of colleges average 10-year student salaries versus spending per student respective to the ownership. There is public, nonprofit, private, and total in the graphs below respectively.

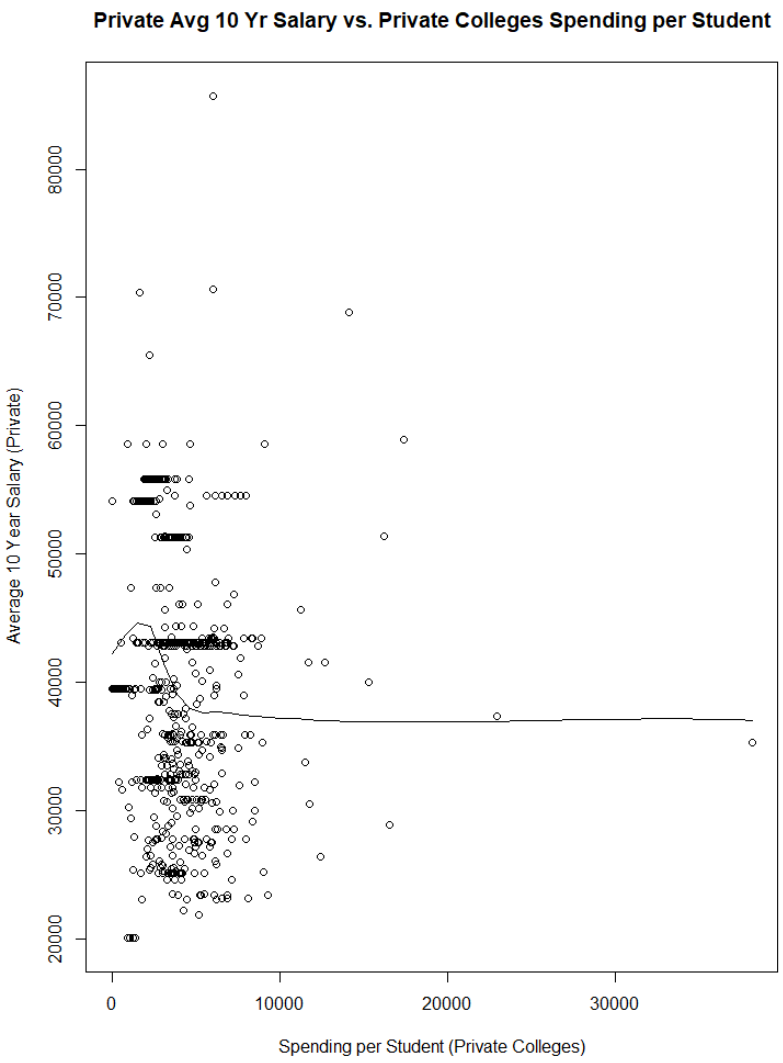


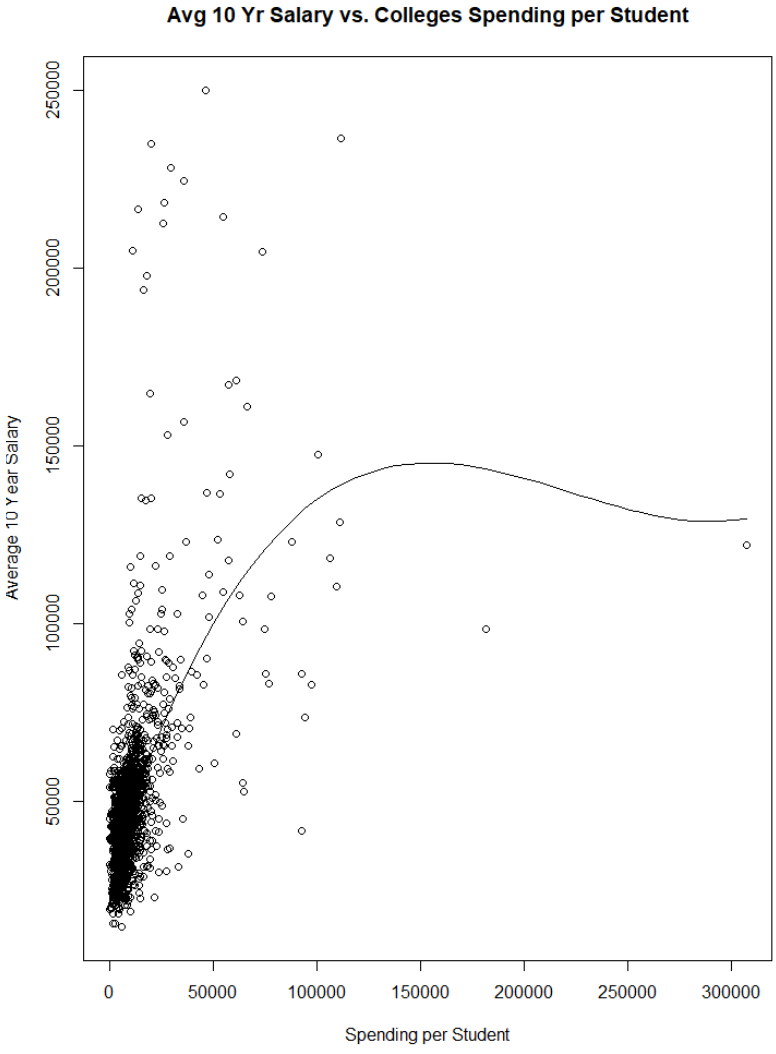


We notice that in general between both Public and Nonprofit that it requires relatively little amounts of spending per student to increase their average 10-year salary from around $30,000 to around $60,000. It's even less expensive for public colleges to have to spend on their students for a significant increase in 10-year salary past a certain point. From our data, in general the public colleges students' 10-year salaries are higher because of increased spending on students only up to its' plateau of around $110,000 spending per student. As it pertains to nonprofit colleges, there is a quicker leveling off positive linear trend. The nonprofits' clustering and variability seems to be a bit larger than the public colleges and the graph shows many more observations of higher earning 10-year salaries between $0-$50,000. Although, when looking closely at both of these graphs you may notice that it may be the very closely packed data of public 10-year salaries that are skewing the

higher sloping trend line. It's also important to note that the nonprofit colleges in general seem to spend more on students which is interesting because they are using funds from donations and such whereas one would think public colleges are stingy towards students in how to plan to spend their money.

Also, when comparing all 4 we notice that the public average 10-year salary versus public spending per student graph looks most like the last graph which represents total spending throughout all colleges. Lastly, when looking at the private graph (left), we notice how vastly different it looks than any others. There are what seems to be a few horizontal subgroups of close to definitively linear "Average 10-Year Salaries'. This could mean that a combination of the private institutions resources, networking, and specialty concentrations or degrees gives their students a direct path to a specific job which is slotted very specifically at those salary points. The graph also shows no trend for spending by the private colleges that increases or decreases the average 10-year salary which may be because students are paying so much for private institutions of which many are not as highly touted as other select private colleges.





8. Starting off with the answer of my findings, the college that gives the best earnings for the cost is Sitting Bull College in North Dakota, and according to Google their tuition is $3,910 as of 2015 and their total costs are $14,156. I had determined this by adding a column vector in the dataset provided which took the median_10yr_salary variable and divided it by the net_cost variable to give us a ratio of earnings over cost. I think a confounding variable in this procedure is not accounting for the huge discrepancy between the tuition costs of different universities while also accounting for the median 10-year salary. I believe that both the salary and tuition variables are not weighted equally. Furthermore, this method did not consider overall value and only tells us so much. It's understandable that the best earnings with respect to college cost is some unheard-of college in ND with a tuition of $3,910 (no offense to Sitting Bull students), but the value of the degree does not seem to match up in the long run compared to other colleges. Consider a graduate from UC Berkley where the total costs are $29,784, which is significantly more than Sitting Bull's total cost $14,156. Now there might be some solid accredited degrees that land someone a good job from Sitting Bull College, but I will always take UC Berkley's well-known reputation, degree, and networking to be of higher value, especially in the long run. I

9. I thought it was a good idea to show the top 5 most diverse colleges after seeing a classmate include that information. I had extreme difficulty with this problem initially because I was trying to use vector operations to give me a standard deviation of the race vectors. It had slipped my mind for the longest time, trying so many different things until I realized I can just use an apply function for each row of my data subset. So, I grabbed each column vector of race and put them into a data frame, then used an apply function to add a the standard deviation column vector to our main dataset, then I sorted the column vector and took care of the 0's and grabbed the top 5 observations in a variable. I then used that variable to filter the main dataset to grab all of the accompanying data by filtering standard deviation and using the %in% function to equate them to my variable with the 5 lowest standard deviations. Lastly, I selected the accompanying columns of race data and college and standard deviation.

   The reason we want the lowest standard deviation is because it means that the different race percentages will be closer together. There will be less variability among the race percentage data resulting in more diversity of students. The higher the standard deviation, the more variability there is in the data which accounts to certain colleges with race percentages of say .90.

   The top five colleges were: 1. Holy Names University 2. Pacific Rim University 3. Golden Gate University-San Francisco 4. California State University-East Bay 5. The University of Texas MD Anderson Cancer Center

10. For one, I started off by sub setting our main data set based on the UC's versus all other state schools just to get an idea for how we stack up as a system. For one, our retention rate is 91.11% compared to the state schools 70.77%. Retention is important because it means people value the school they are going to. It could mean the UC's have strong institutions that care about the students since they are doing a good job at keeping them there. Furthermore, the mean of the median 10 year salaries for UC's are $65,763 whereas state schools are $40,713. Now UC Davis specifically it's $55,000 which is still significantly higher than state colleges. The standard deviation column vector we created earlier shows that UC Davis has .156 where all other states schools average out to 0.253. So diversity is much more apparent by up to 10 percentage points. All in all, Davis crushes state school competitors in 3 very important aspects. Lastly, the

completion percentages are way up in the UC's as well as with UC Davis comparative to all the state schools. UC Davis completion percentage is %81.33 and states schools are %42.55