Big Data Project 2 by David Kim and Katherine Holotko

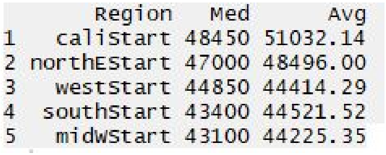
Introduction

In this project, teams were given an opportunity to display their knowledge and mastery over the statistical computing and graphics language, R. For this project, the datasets we used for college salaries, was from the Wall Street Journal, which was limited in scope to the top colleges of their choosing, and not representative of the entire United States. However, we felt that the dataset was representative of the schools that we wanted to look into because they were some of the top earning schools in the nation and some of the most recognized. The dataset for the entire college information was represented in colleges.csv and gave both educational data (SATs, admissions statistics, cost, etc.) as well as demographics that were involved with the college admissions/enrollment process.

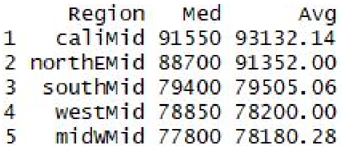
In this assignment we used data related to college location, salary, and admission data in order to analyze and answer multiple questions. R files for this project that were used have been consolidated into one file and added to IBM cloud at the location /home/2019/nyu/spring/6513/keh384/P2.

Q1: College Region and Starting/Mid career median salary?

A question we had as we were looking at the data was whether the region of the college mattered to starting or mid career median salary. In order to determine this sort of relationship, first we had to subset the data by the different regions and including only the median salaries of both starting and mid career salaries as well as the schools within the regions. We decided not to include the quartiles nor the 10th percentile of earning salary amounts as we wanted to capture an understanding of the normalized amounts, namely the median, because that would give us a standard from which we could base our assumptions. We didn't want any statistical outliers affecting our data. We then performed summary statistics of each of the regions for both starting and mid career median salaries to capture the median (of the medians) and average salary amounts for each of the regions for both starting and mid career salaries and only included those amounts with the regions. The following shows the results of the calculations with the left being starting median salaries and the right being mid career median salaries by region.

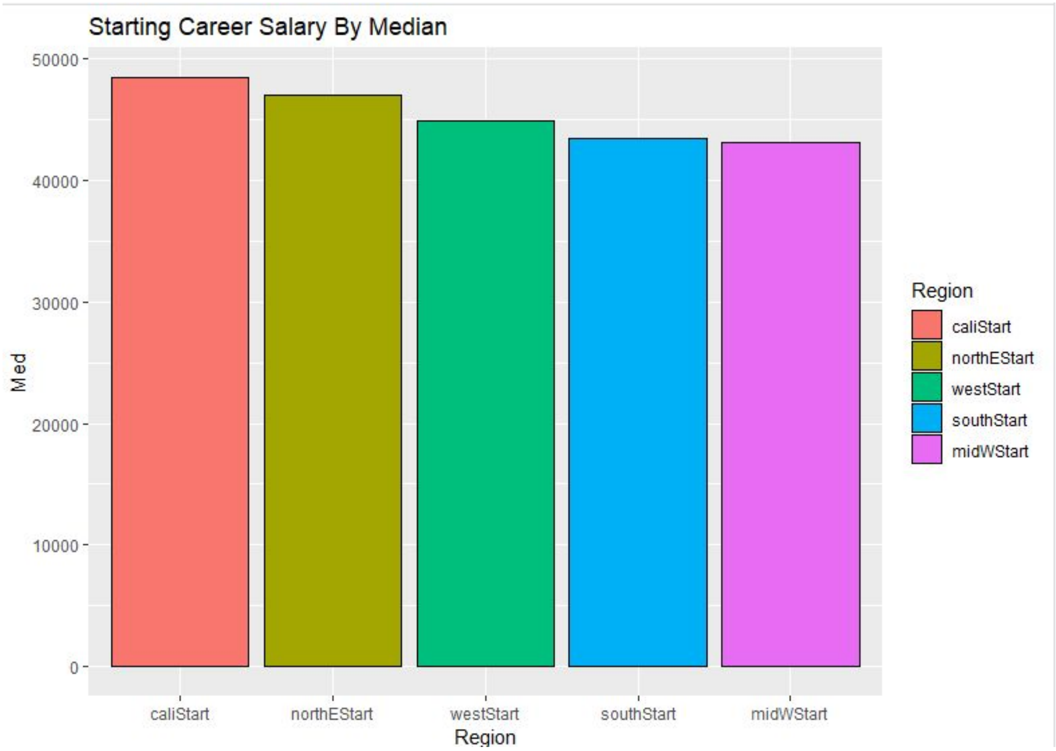


```
      Region   Med      Avg
1   caliStart 48450 51032.14
2 northEStart 47000 48496.00
3   westStart 44850 44414.29
4  southStart 43400 44521.52
5   midWStart 43100 44225.35
```

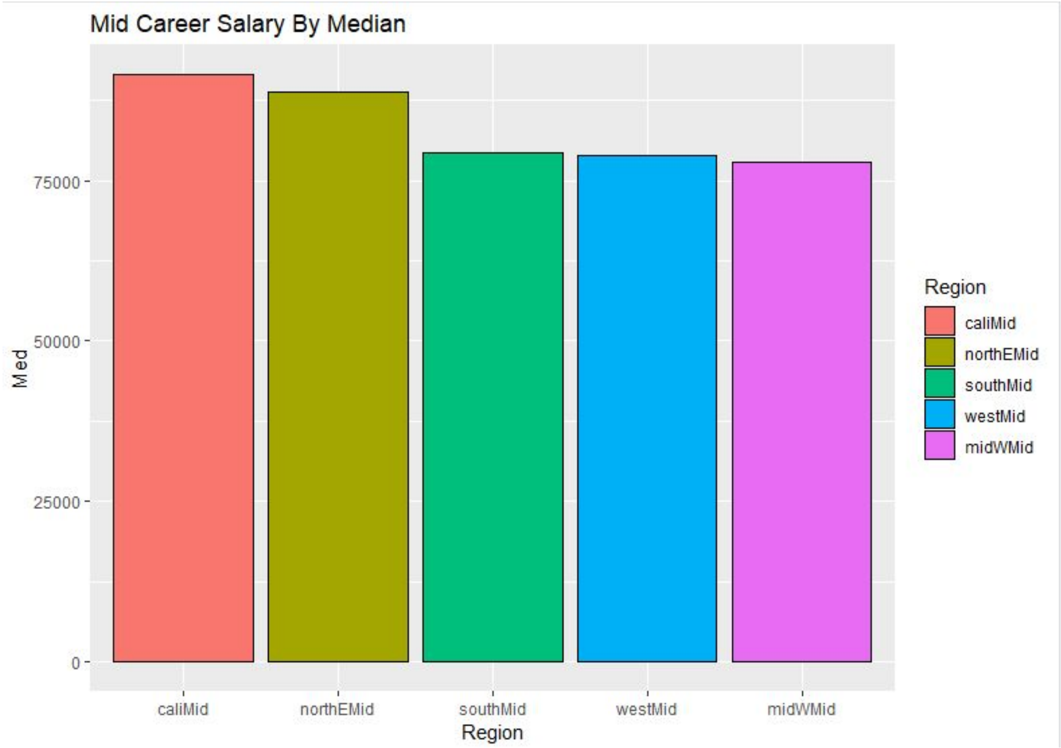


```
     Region   Med      Avg
1   caliMid 91550 93132.14
2 northEMid 88700 91352.00
3  southMid 79400 79505.06
4   westMid 78850 78200.00
5   midWMid 77800 78180.28
```

The graphs below visualize the data showing that for starting career salaries, the differences between the regions seemed to not be too different, with the ranges varying in the 40,000 salaries.



However, with the mid career median salaries, the differences were pretty drastic with the California and Northeastern regions jumping up 10K+.

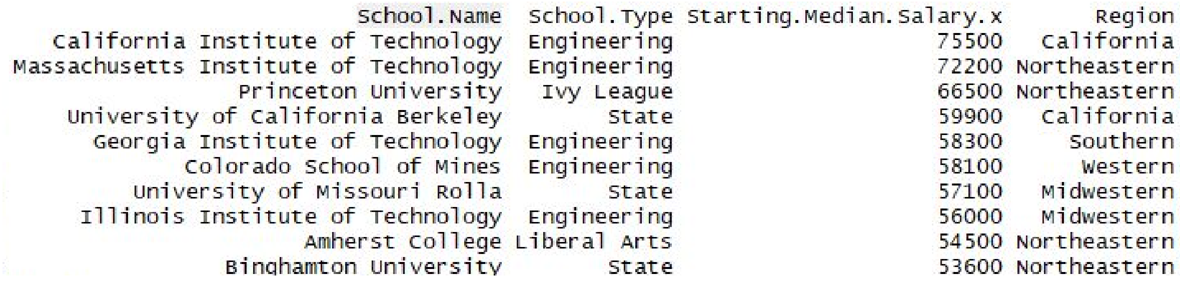


From the numbers, we were able to see that in either of the cases, coming from a california or even a northeastern region educational institute, it probably paid off to study there, with the south and western schools switching places from starting to mid career median salaries. It might pay off probably if a student bit the bullet and were to have attended school in a western region.

Q2: Regions, College Type, and starting/mid career earnings?

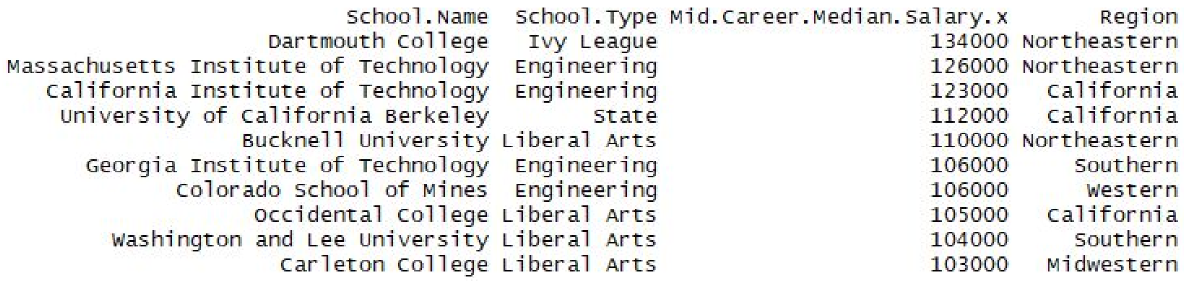
Another question we had was that within the regions, did certain colleges stand out in terms of college type for starting/mid career earnings? In order to begin exploring the question, we had to once again subset the colleges by the college type. The types were already predetermined by the dataset that came from Wall Street Journal, and it came in the flavors of: Engineering, Party, Liberal Arts, Ivy League, and State schools. After subsetting by the types, we used the already subset regional data and performed dplyr's left join function to filter by school name for each of the subtypes (i.e. engineering + california, engineering + MW, etc). This created objects for a combination of each of the regions and school types. We then placed all the 21 objects (20 from the combinations and 1 from Ivy league schools only coming from the northeastern part of the United States) in to a list to be able to extract the maximum salaries from them.

This portion of the code was where some of the bigger problems were met because when appending the maximum values that were extracted from a for loop that iterated through the objects, rbind wouldn't work with a list. Through much of the research we looked into, we realized it was a data typing issue and that the mismatch was the reason for rbind not being able to append the data. We converted the list into a vector of objects and iterated through the vector finding the maximum starting and mid career salary schools from each of the regions/school types, appending it to the empty dataframe. The lists of the top 10 schools by starting/mid career salaries by the type of school we came up with is the following:



```
                         School.Name  School.Type Starting.Median.Salary.x       Region
    California Institute of Technology  Engineering                    75500   California
 Massachusetts Institute of Technology  Engineering                    72200 Northeastern
                  Princeton University   Ivy League                    66500 Northeastern
       University of California Berkeley        State                    59900   California
     Georgia Institute of Technology  Engineering                    58300     Southern
              Colorado School of Mines  Engineering                    58100      Western
          University of Missouri Rolla        State                    57100   Midwestern
     Illinois Institute of Technology  Engineering                    56000   Midwestern
                       Amherst College Liberal Arts                    54500 Northeastern
                 Binghamton University        State                    53600 Northeastern
```

From the starting media salary data, we saw that amongst the top 10 schools of earners, 4 were Northeastern schools and 5 of the 10 were engineering type schools. This pointed to the evidence that again, Northeastern schools were probably a better region to attend, however, all 5 regions were represented with Engineering type schools as being the max starting salaries. This was a strong indicator that Engineering schools were probably the best bet to go for higher starting career salaries with schools such as students from Caltech or MIT being some of the top earners.



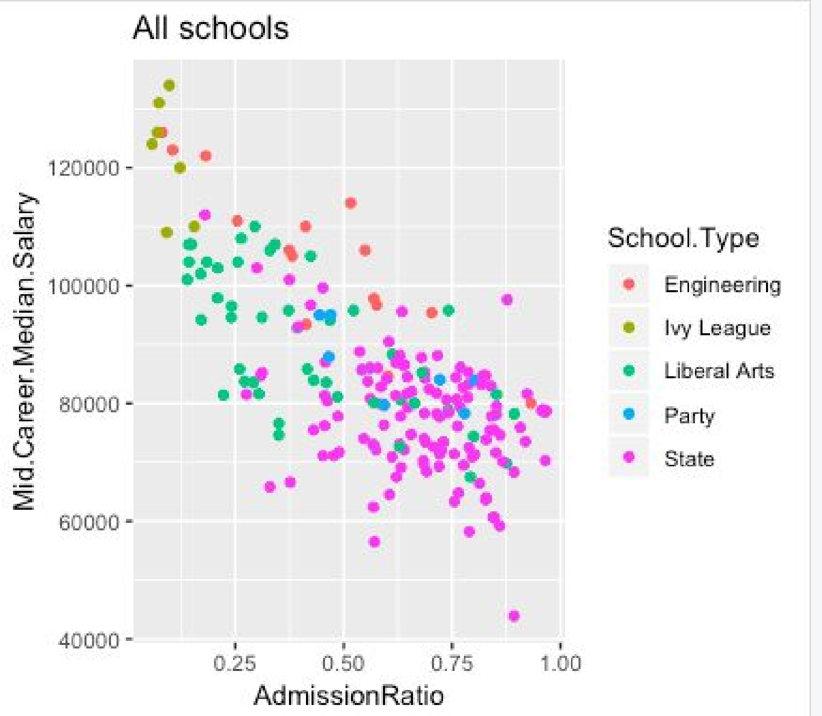
```
                          School.Name  School.Type Mid.Career.Median.Salary.x       Region
                    Dartmouth College   Ivy League                     134000 Northeastern
Massachusetts Institute of Technology  Engineering                     126000 Northeastern
   California Institute of Technology  Engineering                     123000   California
    University of California Berkeley        State                     112000   California
                  Bucknell University Liberal Arts                     110000 Northeastern
      Georgia Institute of Technology  Engineering                     106000     Southern
             Colorado School of Mines  Engineering                     106000      Western
                   Occidental College Liberal Arts                     105000   California
        Washington and Lee University Liberal Arts                     104000     Southern
                     Carleton College Liberal Arts                     103000   Midwestern
```

In terms of the mid career median salaries, we saw that the top 10 schools were all in the 6 digit salary amounts. Amongst them, California and Northeastern schools being the most represented with 6 schools of the top 10. Of the school types, Engineering and Liberal Arts type schools were the most represented, which meant that it is likely to have a higher mid career salary to be going to schools that are of the Liberal Arts or Engineering types, but for the most return on educational investment, Dartmouth would probably be the best bet.

Q3: College competitiveness versus Median Salary

One question we explored was whether the competitiveness of a school will ultimately correlate with a higher median salary. Here we defined the competitiveness of a school as the ratio of Admitted students to the total number of students who Applied. First, we added a column to the dataframe in order to store our competitiveness ratio. Then we merged the data of college salary and college type. Running a correlation over the Competitiveness and Median Mid Career Salary, we got a value of -0.70043. This indicates that as salary increases, the ratio of admitted students decreases. Since a correlation of .25 is considered a weak correlation and .75 is a strong correlation, this indicates a strongly connected link between our variables.

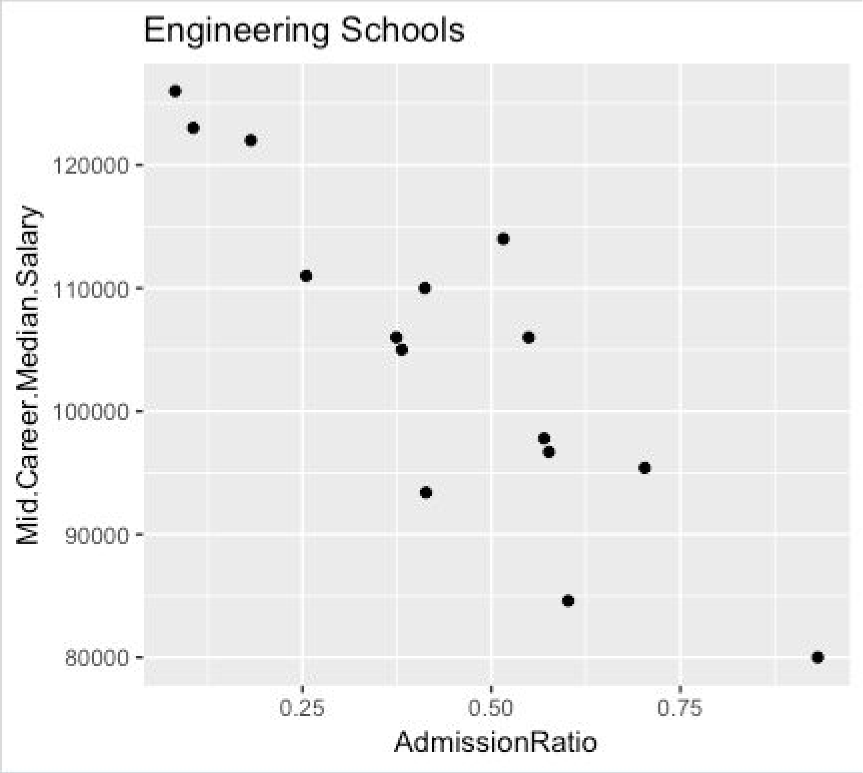
Below is a graph of all the Admissions ratios. You can see the relationship between the competitiveness and salary. There are 5 types of schools: Engineering, Liberal Arts, State, Party, and Ivy League:

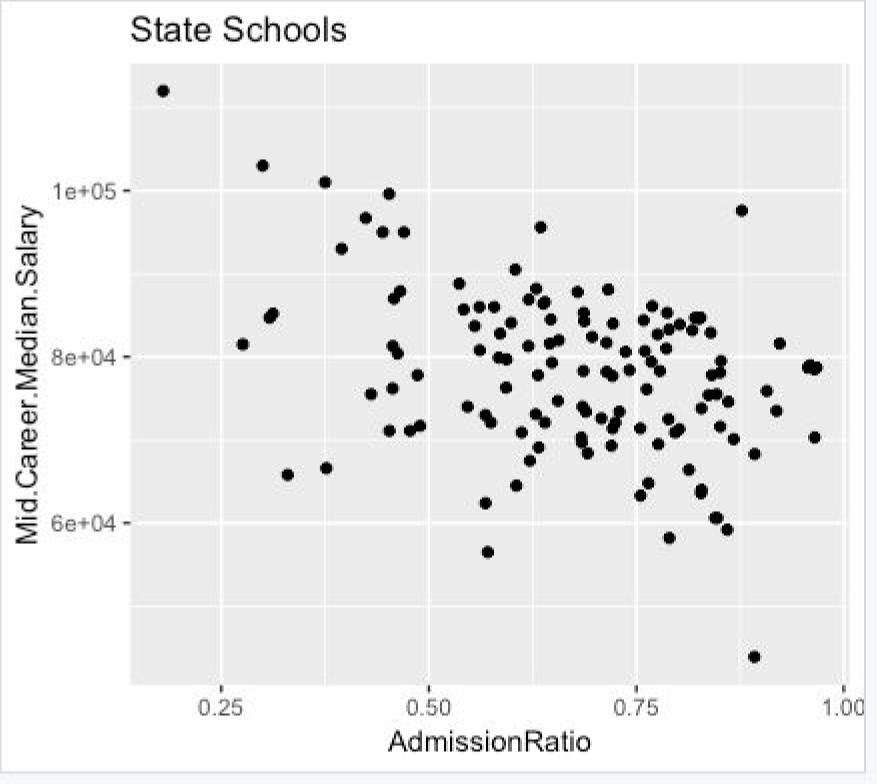


Once we had established a link between our variables we decided to investigate further within the different types of schools. We calculated the correlation coefficients to be:

Engineering	Liberal Arts	Party	Ivy League	State
-0.8758221	-0.6668737	-0.627974	-0.5447294	-0.383238

From left to right, left being the highest correlation and right being the lowest correlation, we can see each school type is somewhat correlated, but Engineering schools are very strongly correlated with competitiveness and salary whereas State schools are not as highly correlated. This can also be seen visually in the difference between the graphs for Engineering/State data points, where the Engineering graph has a more tight line to it, whereas the State graph is wider and less visually connected, even though it does have a certain trajectory to it:


Engineering Schools
Mid.Career.Median.Salary
120000
110000
100000
90000
80000
0.25
0.50
0.75
AdmissionRatio


State Schools
Mid.Career.Median.Salary
1e+05
8e+04
6e+04
0.25
0.50
0.75
1.00
AdmissionRatio

We grouped together the data and ran some analysis using dplyr library. We discovered the standard deviation of median salary was greatest in the Engineering schools and least in the Party schools. This was a bit surprising since a larger range in the salaries seems like it would have more opportunity for variation, and therefore a less-related correlation. It was also surprising that the State schools had a smaller standard deviation than Engineering schools, given that there were only 15 Engineering data points whereas there were 133 State data points.

Since there was such variation in the correlation coefficients from school to school, in order to ensure the validity of the data we also did a boxplot, which allowed us to eliminate some median salary data that would be considered outlier data. We discovered the salaries we should consider lay in the range: 43900 < Mid.Career.Median.Salary <122000. Then we recalculated the overall correlation coefficient to get -0.6293735, which is still between a moderate to strong correlation.

The data implies it does pay to work hard and get into a highly competitive school, no matter what type of school you will be attending. However if you plan on going to State school and didn't get into a competitive school, you have more of a chance to make the same amount of money as other students who graduate competitive State schools, whereas if you are an Engineering student you should definitely try your best to get into a competitive school in order to make a salary comparable to others. Even Liberal Arts, Party and Ivy League schools have a strong enough connection with competitiveness and salary that it would be in an applicants best interest to go with a more competitive school.

Q4: College cost versus Median Salary

A question that arose was whether or not college cost also correlates with median salary. This seemed like a given originally, since we think of expensive colleges as being more competitive, and we already established a connection between a school's competitiveness and salary. We calculated the correlation between the median salary and the cost of tuition and fees in 2013, to arrive at a correlation coefficient of 0.6147998. Yet again, there is a moderate to strong correlation in the data. However if we split out the data by school type we get much different results than the competitiveness:

Engineering	State	Liberal Arts	Party	Ivy League
0.6254379	0.4352455	0.3844867	0.2059881	-0.3392702

None of the results is as strongly related to salary as the competitiveness. There is only a weak to moderate correlation between cost and salary for the majority of school types. In fact, at an Ivy League school going to a more expensive school results in lower salary! Since the correlations are weak for the most part, the overall good news is that you can probably get away with going to a cheaper school, and end up getting paid similarly to students who paid more. However since there is a correlation, albeit weak, you still might have to overcome more obstacles and work harder to get that salary. The exception is Engineering schools, where cost of school seems to be more closely tied to salary. This is good news for NYU Tandon students!

Q5: Demographics comparison

For this section, we asked the question of how the student demographics differ between more competitive schools and less competitive schools. We took our college, college type, and salary data from Q3, as well as our Admission/Applicant ratio, and sorted the data by our competitiveness ratio. Then we picked the top 10 most competitive schools and stored their data in a new dataframe, as well as the top 10 least competitive schools and stored their data in a different data frame, and ran averages over each of the demographic categories. Our results were as follows.

% of Student population in demographic groups

	American Indian/Native Alaskan	Asian	Black/African American	Hispanic/Latinx
Most competitive	0.1	14.7	5.5	8.9
Least competitive	0.5	1.6	5.0	3.6

	White	Mixed	Women
Most competitive	42.7	4.1	46.4
Least competitive	77.7	2.4	49.8

The results show that more competitive colleges have a significantly more diverse community than the least competitive schools. For example, at least competitive schools we have 77.7% of the population is White, a whopping majority, whereas at most competitive schools this is only 42.7% of the population. Most competitive colleges also have a significantly higher Asian student population, a 13.1% difference from least competitive schools, as well as higher percentages across the board except for two outliers. One reason this could be is that many competitive schools tend to be in less rural areas, which attracts people of a more diverse background as well as more international students. Also, when colleges are more competitive and have their choice of who to admit, they can keep diversity in mind, whereas a college that doesn’t have as many applicants might not have that ability.

One outlier is American Indian/Native Alaskan populations, which have .1% population representation at most competitive schools and .5% representation at less competitive schools. This is not a large percentage of the population, so this could be part of the explanation as we only took top 10 most/least competitive schools into consideration. Another possible explanation is that many American Indian/Native Alaskan populations live outside of cities and on more rural locations, which might not have competitive colleges available. Another outlier demographic group was women, which had higher representation at least competitive schools than at most competitive schools. As we already established a relationship between competitive schools and mid career salary, this might be one piece of the gender pay gap puzzle where women are often paid less than men, which is another interesting question worth exploring.

Conclusion

During this project, we learned the following:

1. We learned how powerful R can be in computing statistics quickly and easily. It's very powerful and intuitive to learn, and is an all around great tool to analyze datasets.
2. Once again, designing and planning our analyses was very important. With questions in mind and hypotheses in store, analyses require very logical steps in order to achieve the sorts of results that we envision. Coming up with a plan, taking it step by step, and creating the algorithms with the end goal in mind is important.
3. Graphical tools like ggplot2 are very powerful tools to visualize the data and help the user/reader understand more concretely the data that has been analyzed.

Contributions

David Kim: Completed the analysis for questions 1 and 2. Contributed to a lot of the data cleaning. Contributed to the introduction and conclusion.

Katherine Holotko: Completed the analysis for questions 3, 4 and 5. Contributed to the Introduction.