Capstone Project 1

Client Report

Introduction

The cost of a higher education increased roughly 67% between 2002 and 2012. Since the cost is largely uncontrollable, someone may wish to consider what factors affect post-graduation salary to offset the expense. Therefore, the question which I wish to address is: How is post-graduation salary influenced by tuition, pre-admission placement score, institutional control (whether an institution is public or private), an institution's geographic region, and which programs are offered at the institution (does the institution offer an engineering degree or a computer science degree)?

Here, I will use linear regression to create the model. I use linear regression because it helps to explain the relationship between one dependent variable (salary) and one or more independent variables (tuition, etc.).

The tuition values used in this report are for in-state tuition and reflect four-year tuition paid. Likewise, the salary values show the mean salary received by a graduate two years after graduation. Also, the earnings included in files prior to 2011/2012 are inflation adjusted to 2014 dollars using the Consumer Price Index for all Urban Consumers (CPI-U). Beginning with the 2012/2013 data file, the remaining earnings are inflation adjusted to XXXX+3 dollars using the CPI-U (i.e. earnings included in the 2013/2014 data file are inflation adjusted to 2016 dollars).

The initial dataset contained roughly 140,000 records with nearly 8,000 columns. After reviewing, wrangling, and cleansing the data[1], I reduced the dataset to just over 10,000 records with 8 columns. Our dataset contains the following information:

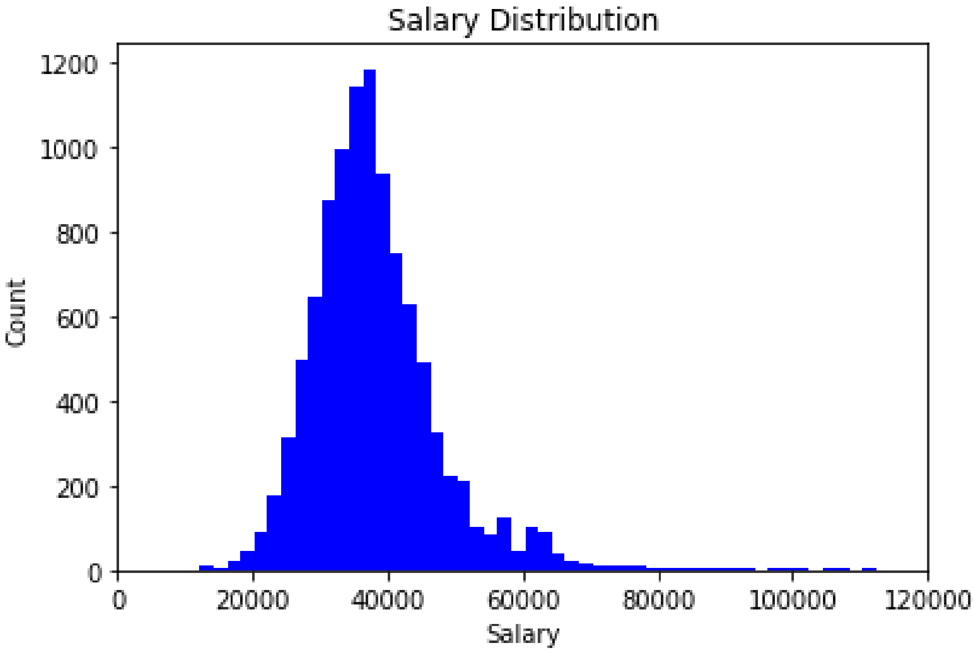
- The mean salary for the institution
- The tuition for the institution
- The salary-to-tuition ratio
- The geographic region name for the state where the institution resides (i.e. New England)
- An indicator which denotes if an institution is public or private
- An indicator which denotes if the institution offers a Computer Science degree
- An indicator which denotes if the institution offers an Engineering degree
- The state-wide mean SAT Score for the state where the institution resides[2]
- The state-wide mean ACT Score for the state where the institution resides

[1] For more information regarding the preparation of the data, please see Appendix 1.

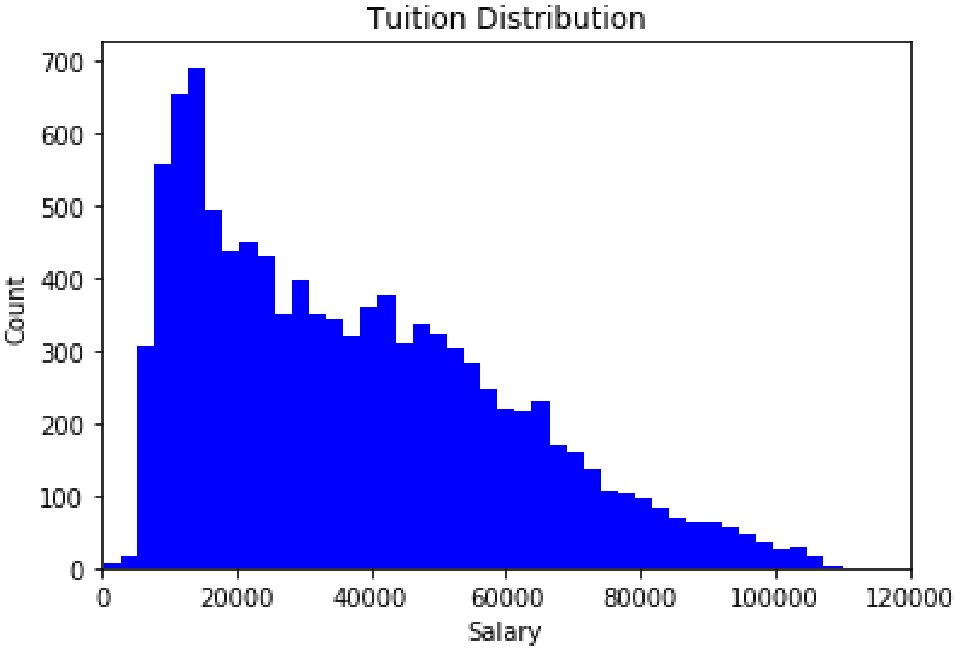
[2] Note that I calculated the mean SAT scores and the mean ACT scores by rolling the available institutional scores up to the state level. I then averaged these state-level scores for each statein the region to determine the mean placement score that is reported here.

Exploratory Data Analysis

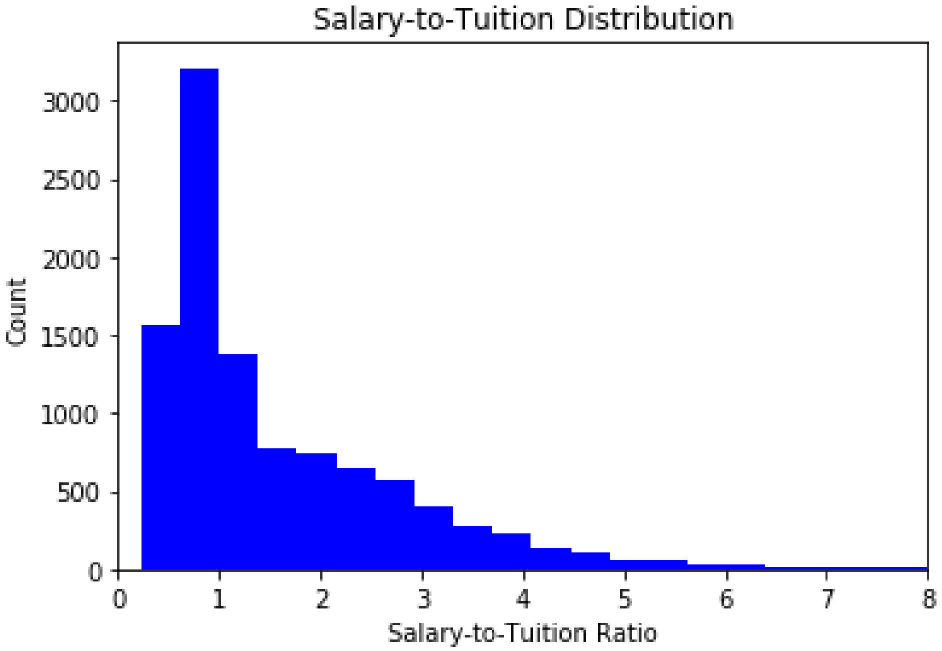
Here we present an overall exploration of the data set. We begin by plotting a histogram of the salary data which shows a slightly skewed-right distribution with a mean value of $38,100 and a median value of $36,800.



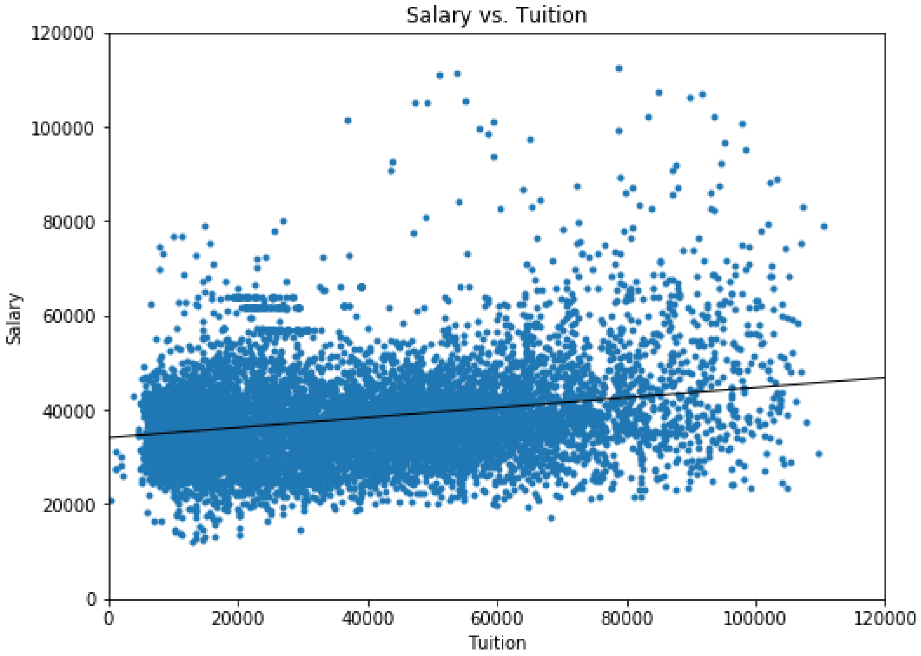
We also plot a histogram of the tuition data. Here, we find a heavily skewed-right distribution with a mean value of $37,300 and a median value of $33,300.



Then, we plot the data set's salary-to-tuition ratio distribution. Here, again, we find a heavily skewed-right distribution with a mean value of 1.61 and a median value of 1.07. We define a *lower cutoff threshold* as the value where an institution has a salary-to-tuition ratio equal to one (which means that the reported annual mean salary is equal to the calculated total tuition) and a *upper cutoff threshold* as the value where an institution has a salary-to-tuition ratio greater than four (which means that the reported annual mean salary is at least four times the calculated total tuition). Roughly 47 % of institutions fall beneath the lower cutoff threshold; roughly 48% fall between the lower cutoff threshold and the upper cutoff threshold; and, roughly 5% of the institutions fall above the upper cutoff threshold. We also note that institutions with upper cutoff thresholds reside largely in the Southeastern and Southwestern regions of the United States.



Next, let's plot salary vs. tuition.



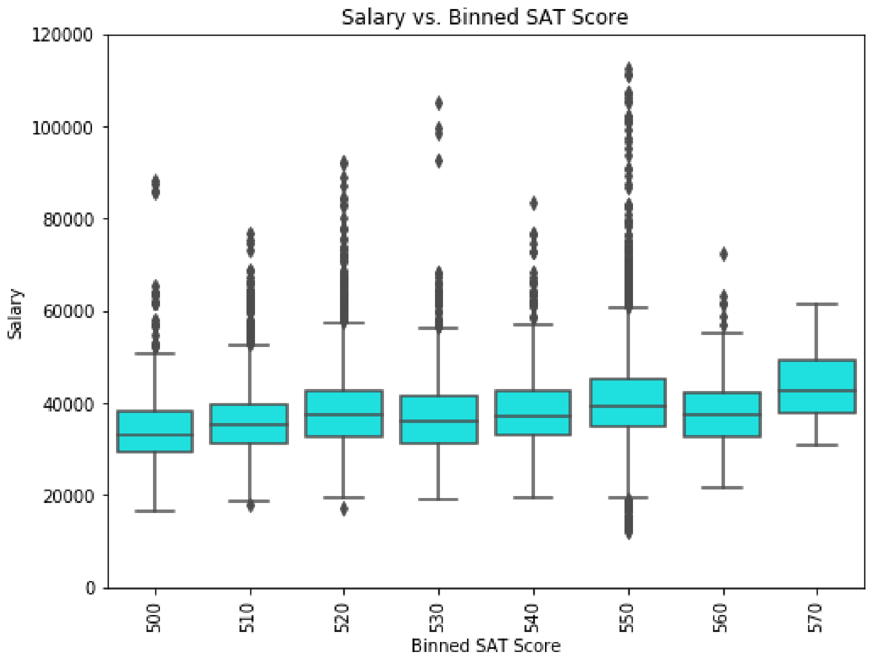
To quantify the resulting graph, I fit a linear regression line to the data; the designated slope, y intercept and R-squared[3] values are reported below. I also calculated the data's Pearson correlation coefficient[4], along with the coefficient's corresponding p-value. All results are summarized here:

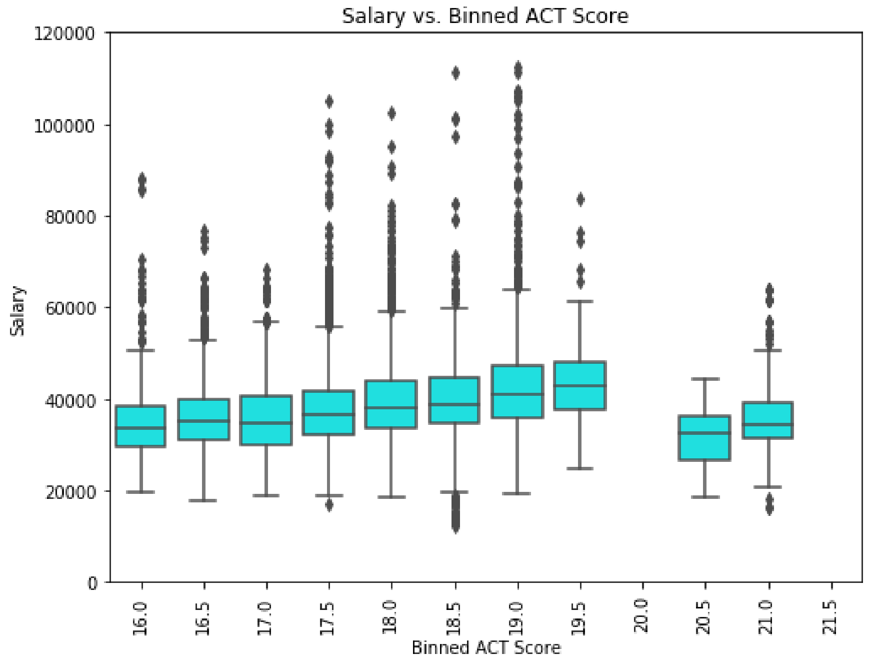
Slope	0.106
Intercept	34,157
R-squared	0.062
Pearson correlation coefficient	0.250
p-value	< 0.01

We see that, in the absence of other factors, an increase of $1,000 in tuition would result in an increase of $106 in salary. Also, the R-squared value is quite small. This indicates that, even though there is a correlation, there is a small variation in the salary due to variations in the tuition. This is also suggested by the magnitude of the Pearson correlation coefficient; here we note a slight positive relationship between the salary and tuition. This result corresponds with the calculated slope of the regression line. Also, the corresponding p-value indicates that the result is not due to randomness; there is definitely a correlation between the tuition and the salary.

[3] R-squared value indicates how the variation in salary changes with the variation in tuition. A low value indicates that there is not much variation in salary due to variations in tuition.
[4] See Appendix 2 for an explanation of the Pearson correlation coefficient.

Next, we examine salary vs. SAT score and salary vs. ACT score.

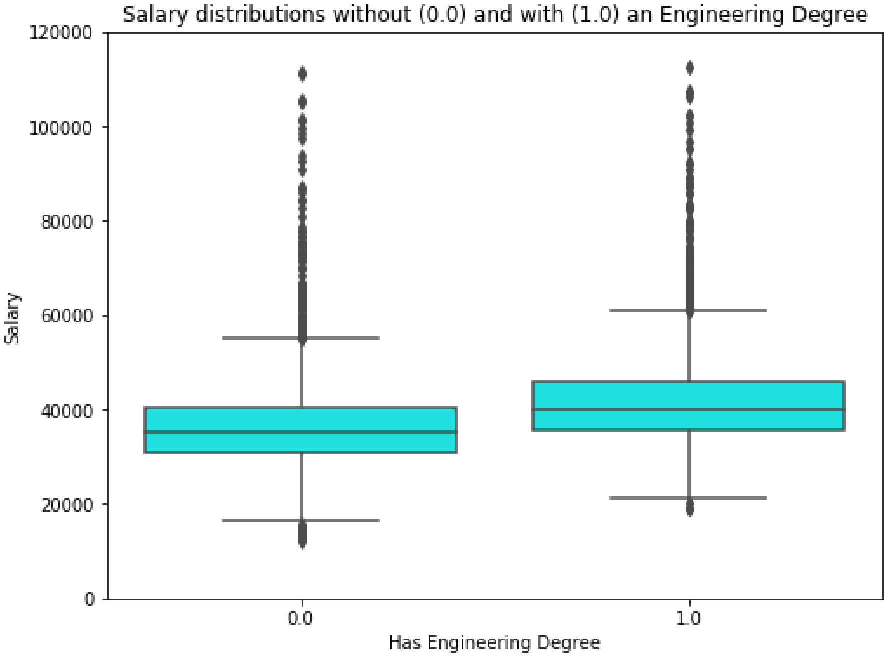




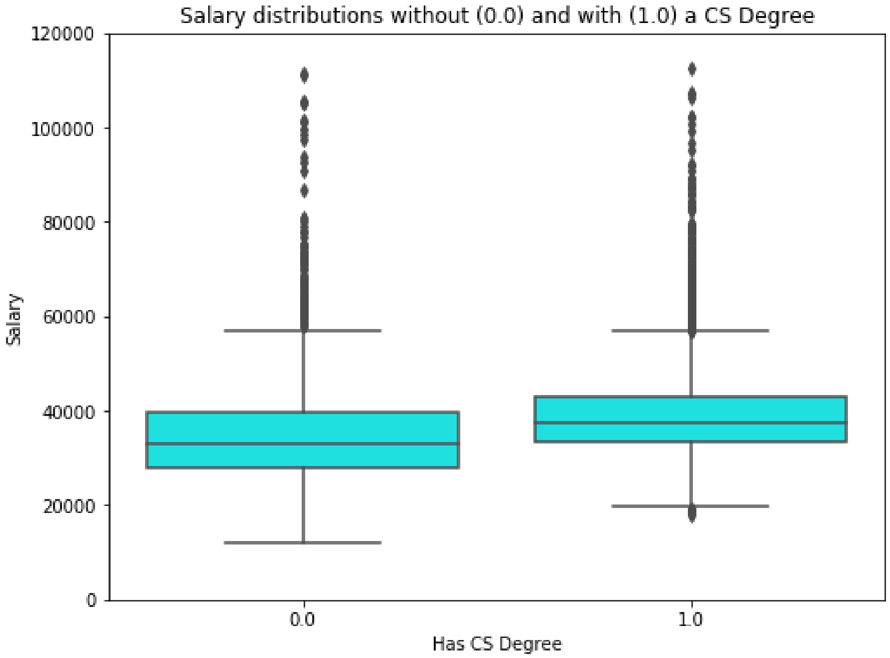
In each case, the plots show what appears to be an increase in salary with an increase in score. For each set of variables, we also calculate the Pearson correlation coefficient, along with the coefficient's corresponding p-value. Here we find:

Sample set	Pearson correlation coefficient	p-value
Salary vs. Mean SAT score	0.202	< 0.01
Salary vs. Mean ACT score	0.104	< 0.01

Finally, we explore how salary is influenced by the programs offered at an institution. Here, by plotting the salaries of institutions with engineering degrees and the salaries of institutions without engineering degrees, we see how an engineering program influences a graduate's salary.



Likewise, we can see how a computer science program influences a graduate's salary by plotting the salaries of institutions with computer science degrees and the salaries of institutions without computer science degrees.



In each of these plots, we want to verify that the differences seen are not simply do to randomness. To do this, we performed 10,000 permutation tests on each set of data and calculated the corresponding p-values for that test. In each case, we found that each p-value was less than 0.01. This indicates that the difference in means for the two distributions is not due to randomness. The results are summarized below.

Sample set	Difference in means	Margin of error	p-value
Engineering	5436	419	< 0.01
Computer Science	3777	487	< 0.01

Machine Learning

Now that we have a general idea of what the data looks like we can use linear regression to build a model that helps us understand what contributes to salary.

Here, we use ordinary least squares to fit the data to the following features: tuition, geographic region, mean state-level ACT score, mean state-level SAT score, public/private indicator, engineering offered indicator, and computer science offered indicator.

The best fit for this data has an R-squared value of 0.179. The feature coefficients are summarized below:

Feature	Coefficient	Coefficient type	p-value
Tuition	0.079	Multiplier	< 0.01
Mean ACT Score	216	Multiplier	0.011
Mean SAT Score	39	Multiplier	< 0.01
Private Institution	9633	Addend	< 0.01
Public Institution	8591	Addend	< 0.01
Region – New England	-1184	Addend	< 0.01
Region – Mid Atlantic	830	Addend	0.015
Region – Southeast	-3672	Addend	< 0.01
Region – Great Lakes	-3519	Addend	< 0.01
Region – Plains	-2685	Addend	< 0.01
Region – Rocky Mountains	-1989	Addend	< 0.01
Region – Southwest	-1333	Addend	< 0.01
Engineering program offered	5018	Addend	< 0.01
Computer Science program offered	2801	Addend	< 0.01

The tuition, mean SAT score, and mean ACT score coefficients may be understood as multipliers. To calculate the effect of the feature, the value is multiplied by the feature's coefficient. For example, to calculate the change in salary due to a $1000 change in tuition, someone simply multiplies $1000 by the tuition's coefficient to find: $1000 x 0.079 = a $79 increase in salary.

The remaining coefficients act as addends; they correspond to features which either apply or don't apply to the given calculation. If the feature is relevant, the coefficient is included in the calculation; if it is not, the coefficient is ignored. For example, a student won't attend both a public and a private institution; she will attend one or the other. So, we would add either $9,633 or $8,591 to account for the influence of attending either a private or a public institution in post-graduation salary.

Conclusion

We analyzed a dataset which contained admissions and placement data for colleges and universities around the United States. Using exploratory data analysis, we discovered that the salary and tuition distributions were right-skewed and that the salary-to-tuition ratio could help us differentiate the potential values of institutions. We also noted that there appeared to be a relationship between salary and pre-admission placement score. And, we saw that the programs offered at an institution have an impact on the salaries earned by the institution's graduates.

Finally, we used linear regression to model the relationship between post-graduation salary and tuition, institutional control, programs offered, geographic region, and pre-admission placement score. We reviewed the two different types of coefficients, and showed how each contributes to the overall salary.

Based on these results, I would recommend that a student consider attending a private school in the Mid-Atlantic region which offers both an engineering degree and a computer science degree. This would offer the best opportunity, above and beyond solid placement test scores, to attain a robust salary. Also, I would suggest not attending a public liberal arts college in the Southeastern region. Based on the data that we have available, salary expectations for such an institution are poor.

Also, we note that, where this model does have some explanatory power, it is still quite limited. There are, no doubt, many factors which can determine salary; some of these might include family of origin, gender, ethnicity, or even height.

This study requires more data and more analysis. For example, having access to a complete set of placement scores, as well as, salary and tuition data reported at the *program* level would be quite beneficial. Also, examining how the salary varies for public vs. private education may show some potentially interesting results. Finally, examining how the data changes with respect to time could also provide more insights.

Appendix 1- Data Review, Wrangling, and Cleansing

I selected a data set from data.gov which contains tuition and salary data for all colleges and universities in the United States. The data set includes data from academic years 1996/1997 through 2015/2016. In total the data set contains approximately 140,000 records.

The shape of the data did not require any modification; however, I did need to scrub the data to remove several string values which indicated that that the data was suppressed. To address this I replaced these values with nulls. Also, by comparing series' counts and series' sizes, I found that numerous null values spread throughout the tuition and salary data. In lieu of simply dropping the null values, I elected to determine where the nulls were located and found that the vast majority of them are located in the data for academic years that begin prior to the year 2000.

Data Set Manipulation and Cleansing

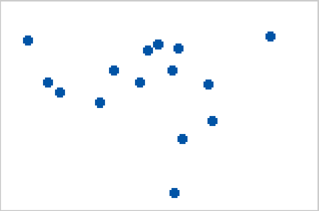
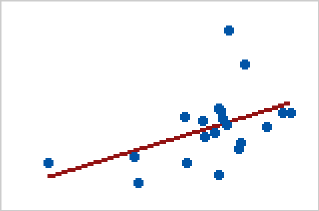
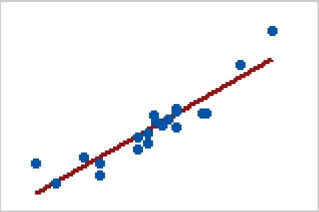
In order to process the data, I needed to address the presence of multiple null and text values in numerical fields. After an initial exploratory data analysis, I did the following to clean up the data:

- I filled null values in the fields which specified whether a certain major was offered with zeros. This aligned with the original data using a zero to designate that an institution did not offer a certain curriculum.
- I replaced invalid text values in the salary fields with null values and then removed records which contained nulls in both the tuition and salary fields.
- I rolled SAT and ACT scores up to the state level because not all institutions reported the values for each applicable year.
- I created fields for region names and a calculated salary-to-tuition ratio.
- Once the data cleanup was complete, I dropped all unnecessary fields.

Data sets which would make this analysis more robust would include annual institution-level SAT and ACT scores and more granular institution-level tuition and salary values.

Appendix 2 - The Pearson Correlation Coefficient

The Pearson correlation coefficient is a calculated value that reflects the strength of the relationship between two variables. Its magnitude is between +1 and -1, where +1 means that the data has a perfect linear relationship and zero indicates that there is no correlation between the variables. The sign of the coefficient indicates the direction of the relationship, where a positive value means that when one value increases the other value increases (i.e. the slope of a line through the data would slant upwards) and a negative value means that when one value increases the other value decreases (i.e. the slope of a line through the data would slant downwards). Below are some examples:

No relationship, Pearson r = 0 	*Moderate positive relationship, r = 0.4* 
Strong positive relationship, r = 0.9 	*Strong negative relationship, r = - 0.9*