Ryan M. Walker

College Scorecard

INTRODUCTION:

For most people seeking decent job prospects, a good quality-of-life, and an overall prosperous future, postsecondary education has become necessary to securing opportunities to achieve aspirations and advance careers in the 21st century. According to the National Center for Educational Statistics, as of 2013 there were 7,253 postsecondary Title IV institutions in the U.S., of which 3,026 were 4-year colleges.[1] Even if you reduce the list to 4-year colleges alone, that is still a substantial set of prospective schools for a student to investigate and potentially choose from. To assist prospective students in their search for schools of higher education, the U.S. Department of Education implemented College Scorecard. College Scorecard is a web accessible database with information on thousands of institutions in the U.S. and according to College Scorecard,

> The College Scorecard project is designed to increase transparency, putting the power in the hands of students and families to compare how well individual postsecondary institutions are preparing their students to be successful. This project provides data to help students and families compare college costs and outcomes as they weigh the tradeoffs of different colleges, accounting for their own needs and educational goals. (Data Documentation for College Scorecard, 2017, para. 1)[2]

As a college student, I find this database deeply intriguing. When I was first considering which schools I would be interested in attending, I know I would have spent numerous hours looking over this information. Also, as a first-generation college student, I feel there would have been great benefit in having the guidance and information available to me when blindly trying to discern what was in my best interest for higher education. Although I am now settled in my educational path, I still find this data to be of interest. However, now my interest extends to the raw metadata and the information that can be gathered from its vast entries. Specifically, my interest concerns the various variables in the dataset related to the postsecondary institutions and their association with first-generation college students in the U.S.

College Scorecard is an extremely large dataset provided by the U.S. Department of Education and published by the Office of Planning, Evaluation, and Policy Development. According to the Data Documentation, data elements in the set were provided predominantly by the Integrated Postsecondary Education Data Systems. The raw data available from data.gov is functionally an accessible database of all postsecondary institutions and info related to those institutions for use with web-based analysis searches at collegescorecard.ed.gov. Essentially, the data is a census of all postsecondary institutions that qualify under Title IV of the Higher Education Act to receive funding from the federal government to assist students with tuition. This contributes to the size of the dataset and means its information ranges from the likes of for-profit, cosmetology schools to highly-active research universities. With that in mind, limiting the parameters of the data is essential for effective analysis in our report.

For the purposes of our observational analysis and the scope of this project, we decided to limit the population to all four-year colleges/universities in the U.S. that predominantly award bachelor's degrees using five variables available in the original dataset: predominant degree awarded, highest degree awarded, accreditation, control of the institution, and Carnegie Basic Score. The institutions in the dataset needed to be predominantly bachelor's degree granting, four-year colleges and universities with the highest degrees awarded being either bachelor's or graduate. The schools needed to be

[1] https://nces.ed.gov/fastfacts/display.asp?id=84

[2] https://collegescorecard.ed.gov/assets/FullDataDocumentation.pdf

accredited by the six active, regional accreditation agencies specific to colleges and universities: Middle States Commission on Higher Education, New England Association of Schools and Colleges, North Central Association of Colleges and Schools, Northwest Commission on Colleges and Universities, Southern Association of Colleges and Schools, and the Western Association of Schools and Colleges, Senior Colleges and University Commission.[3] Further, the schools' control needed to be private and non-profit or public with a Carnegie Basic Score between 15-22, which ensures the data includes only schools ranging from Doctoral Universities with highest research activity to Baccalaureate Colleges with diverse fields.[4] We also removed schools in outlying regions (i.e. Guam and Puerto Rico). With these data constraints, the institutions are limited to a list of 1,448. Of the 1,448 institutions listed, we randomly sampled 400 using R for our analysis.

The variables of interest in our report are limited to two categorical: Research Type (Non-research/Research) and control (private/public); and eight quantitative: admissions rate, proportion of first generation student enrolled, median family income of all students enrolled, average age of entry for students, in-state tuition, out-of-state tuition, median debt of students at graduation, and graduation rate of all students within four years. Of the variables listed, one limitation evident is that only graduation rate is a dependent variable. We have numerous independent variables which we will have to address in the analysis. The focus of regression analysis will therefore be more towards graduation rate and its association with other quantitative variables of interest.

It is our aim through graphical representation and statistical summaries that we will be able to better visualize and interpret the relationship between each quantitative and categorical variable in our dataset, with interest towards first generation students. In our statistical analysis and inference we will be using a 95% confidence interval for our inference on single means and difference in means using the classical approach in addition to sample bootstrap distributions. Finally, in our correlation and regression models we will be testing the association between graduation rate and proportion of first-generation students enrolled, median family income, and average age of entry for all schools in our sample. With these processes in mind, we hope we can extrapolate meaningful data from the dataset.

[3] https://ope.ed.gov/accreditation/agencies.aspx

[4] http://carnegieclassifications.iu.edu/classification_descriptions/basic.php

SECTION II:

Quantitative Variables

Admissions Rate:

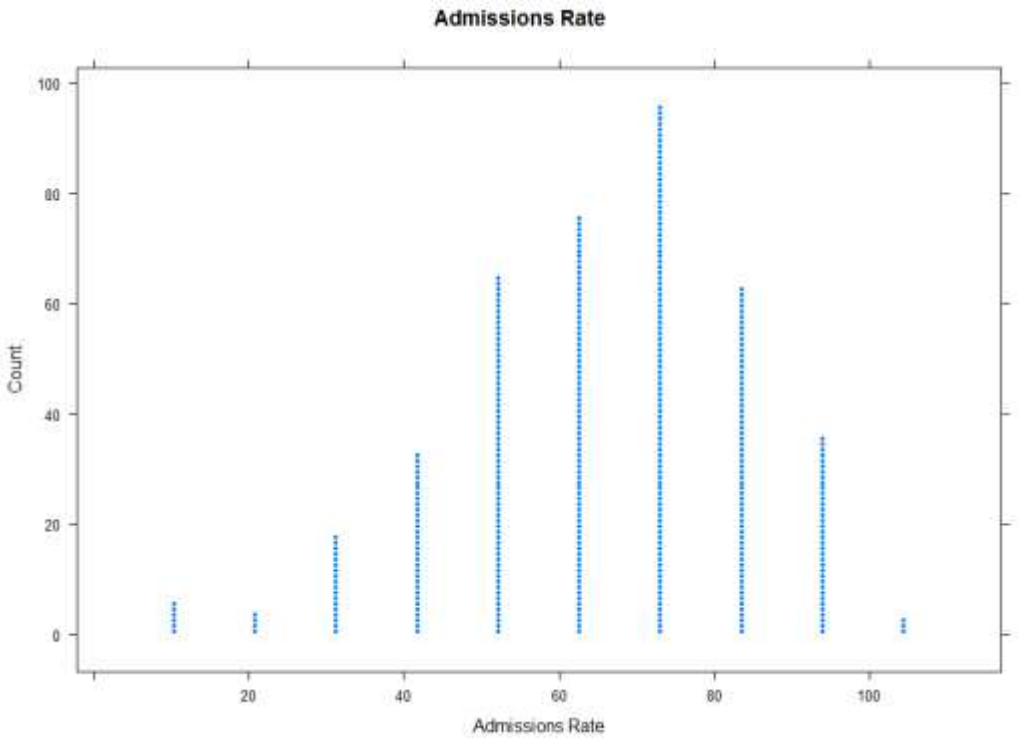

Figure 1. Dot plot showing relationship of admissions rate to count of institutions.

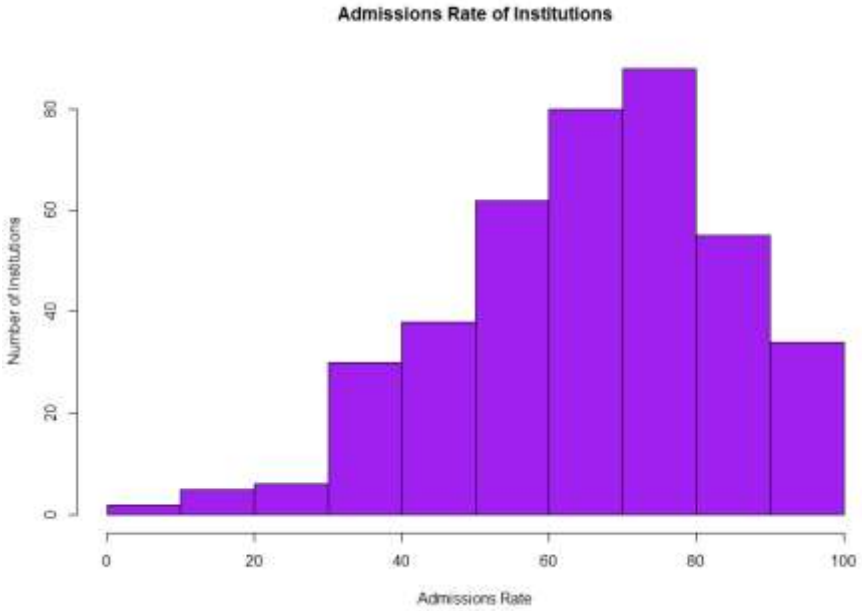

Figure 2. Bar graph showing relationship of admissions rate to count of institutions.

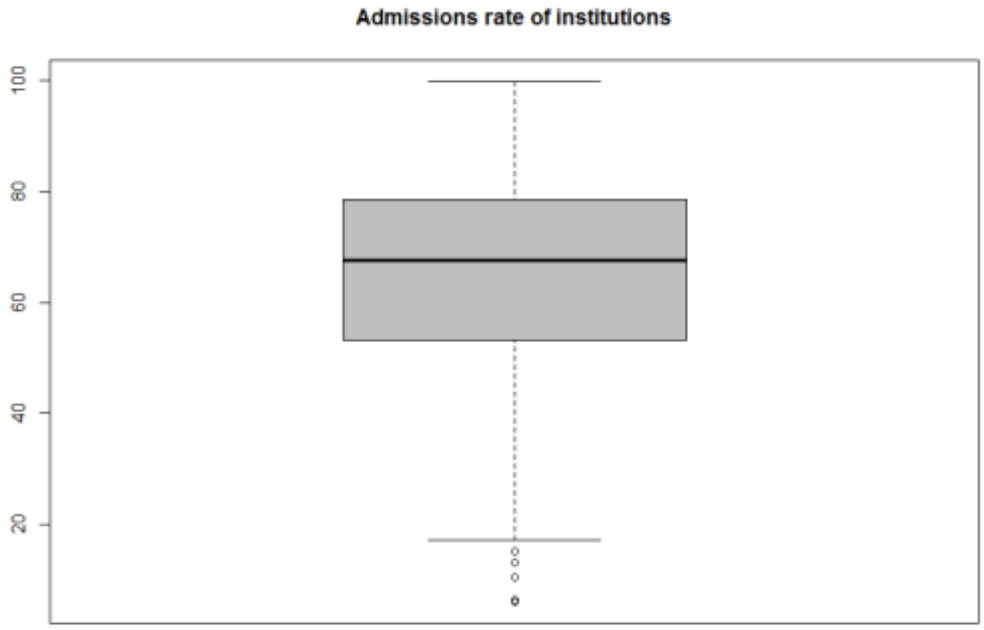

Figure 3. Box plot demonstrating five number summary of admissions rate of institutions.

Summary Statistics											
n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
400	0.00	65.35	18.41	-0.50	0.11	5.96	53.26	67.54	78.47	99.89	25.22

Table 1. Table showing the summary statistics including the five-number summary for the admissions rate at each institution.

Commentary:

Distribution is skewed left and sixe outliers exist in our data in the lower end.

In-State Tuition:

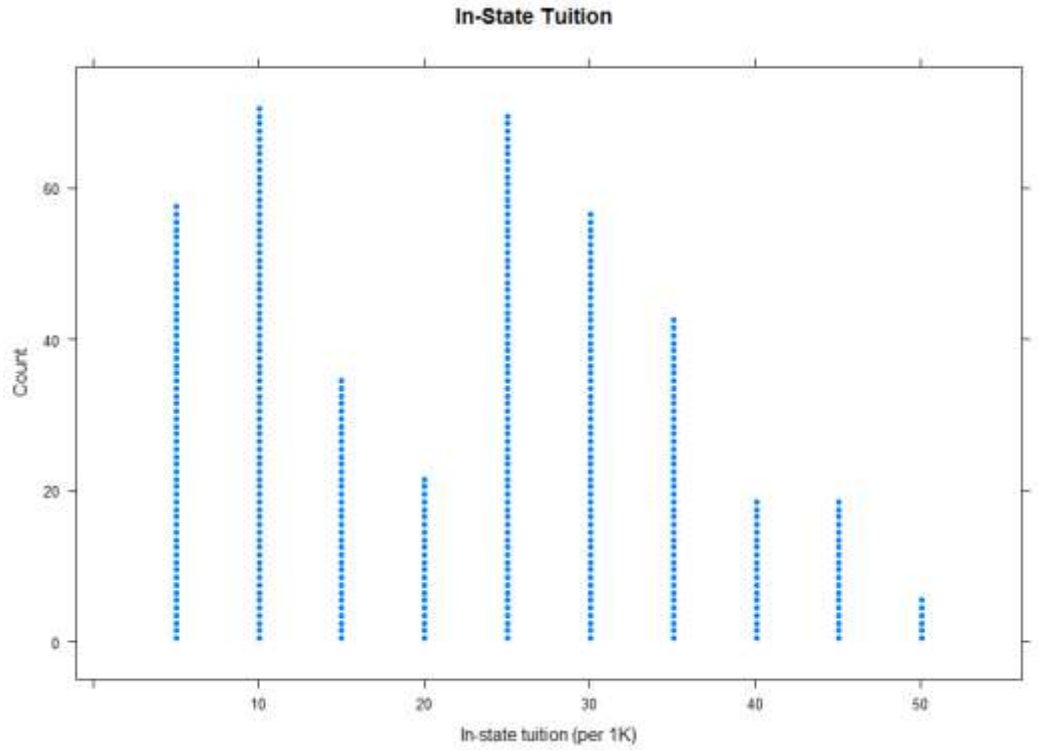

Figure 4. Dot plot showing relationship of in-state tuition (per 1K) to count of institutions.

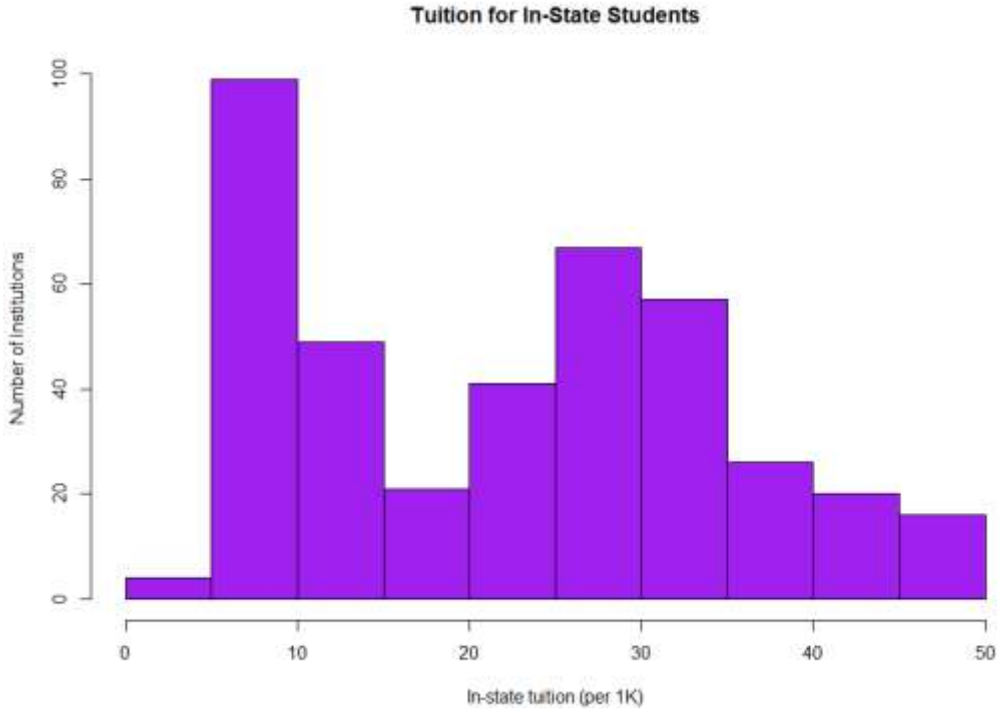

Figure 5. Bar graph showing relationship of in-state tuition (per 1K) to count of institutions.

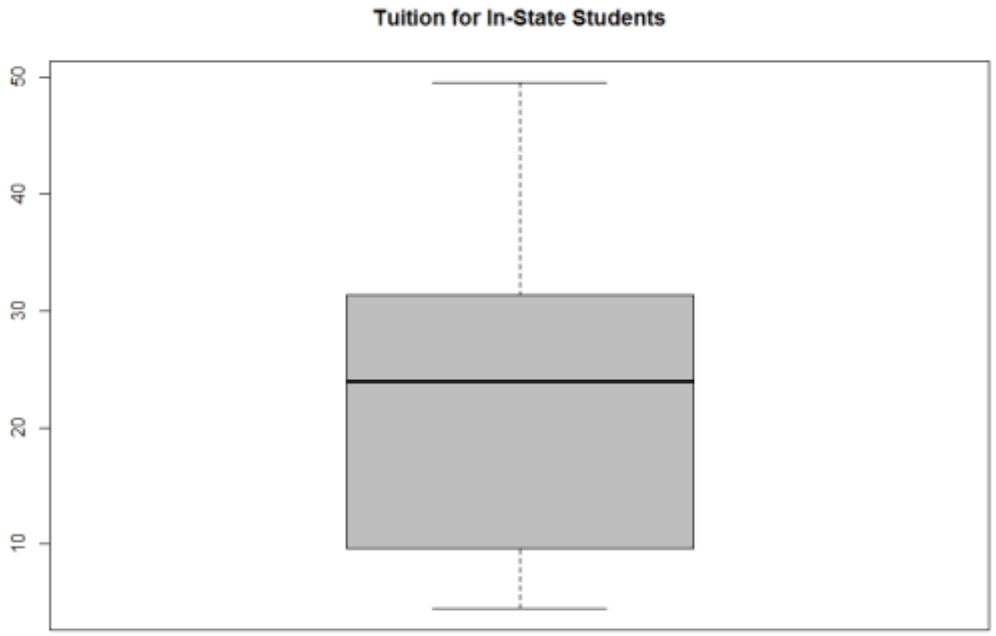

Figure 6. Box plot demonstrating five number summary of tuition (per 1K) for in-state students.

Summary Statistics											
n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
400	0.00	22.27	12.30	0.24	-1.06	4.50	9.69	23.97	31.37	49.57	21.68

Table 2. Table displaying the summary statistics including the five-number summary of in-state tuition (per 1K) at each institution.

Commentary:

Plots are non-symmetric and bimodal, potentially due to the proportion of private institutions in the study (and the U.S) and the fact that tuition for those institutions tend to be larger than in-state public tuition, and tuition is the same for in-state and out-of-state students at private institutions. There are no outliers. While the bimodal nature of this data is really interesting, it makes it difficult to use for meaningful analysis in our report.

Out-of-State Tuition:

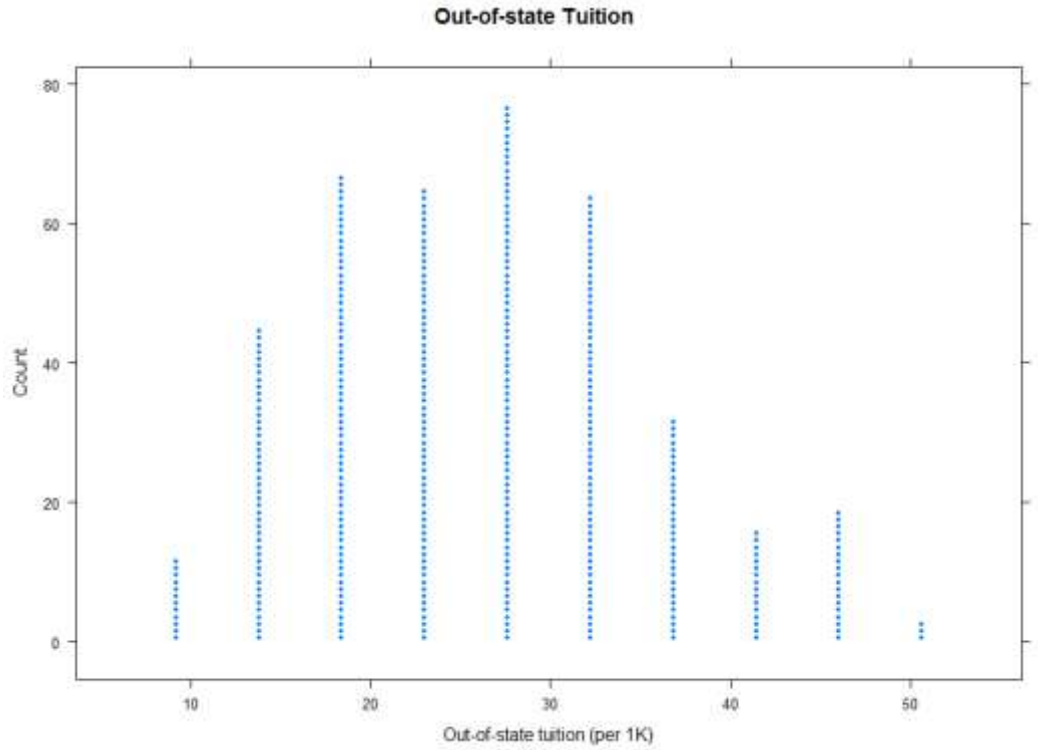

Figure 7. Dot plot showing the relationship of out-of-state tuition (per 1K) to count of institutions.

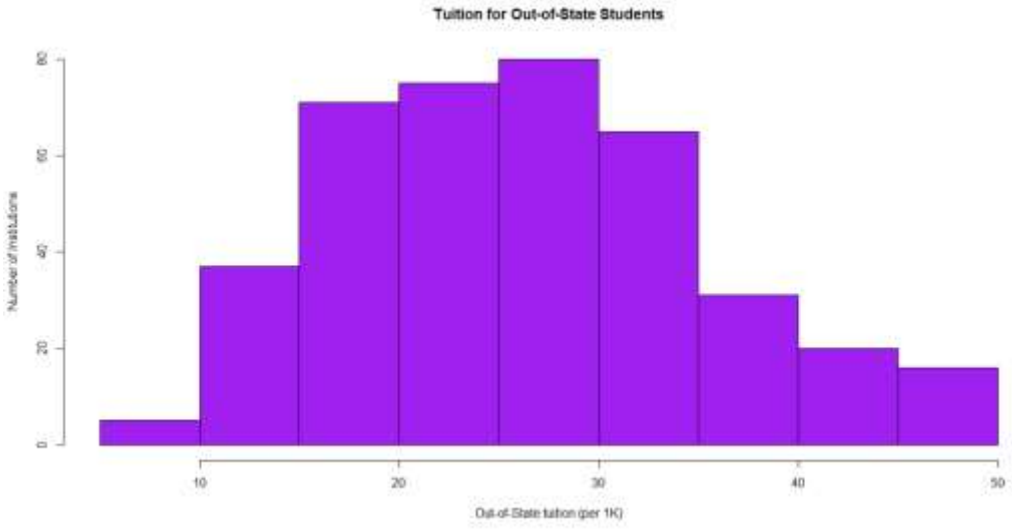

Figure 8. Bar graph showing relationship of out-of-state tuition (per 1K) to count of institutions.

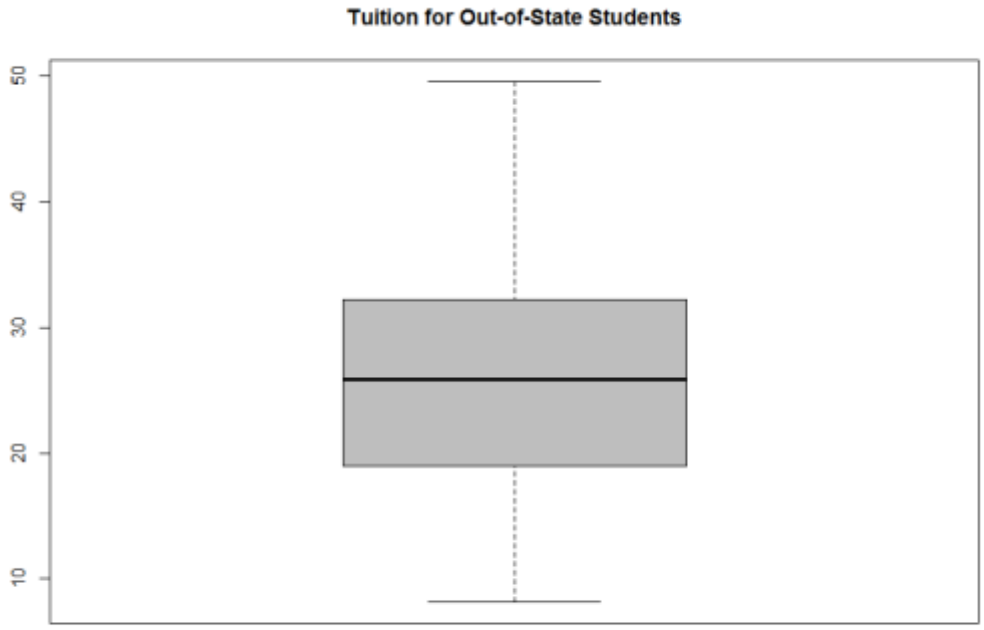

Figure 9. Box plot demonstrating five-number summary of tuition (per 1K) for out-of-state students.

Summary Statistics											
n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
400	0.00	26.24	9.19	0.40	-0.38	8.16	19.06	25.90	32.18	49.57	13.13

Table 3. Table displaying the summary statistics including the five-number summary of out-of-state tuition (per 1K) at each institution.

Commentary:

Plots have a normal distribution and there are no outliers.

Median Debt of Students:

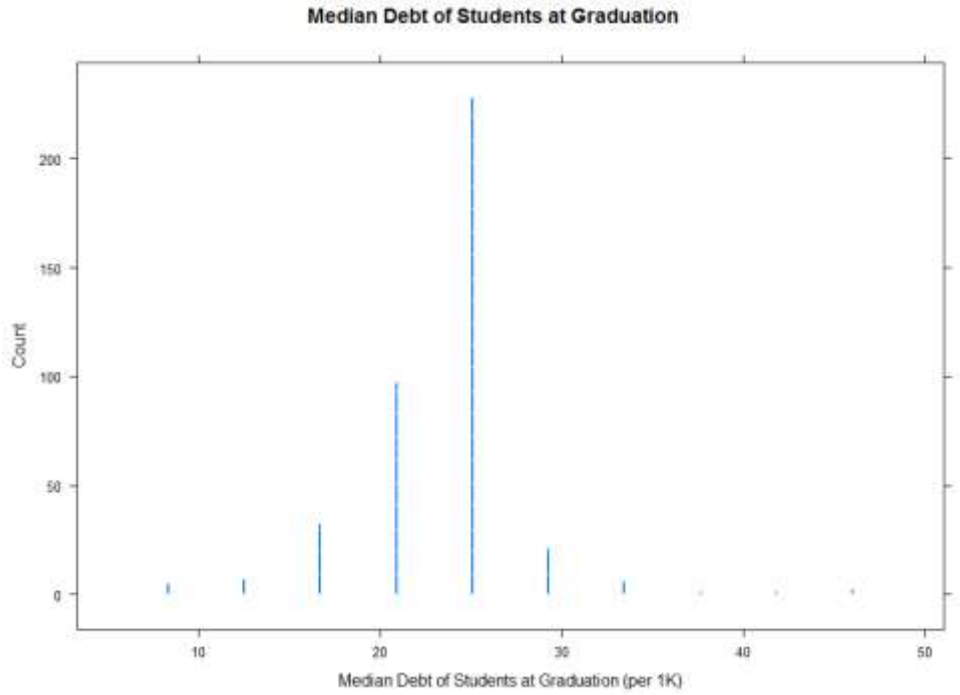

Figure 10. Dot plot showing relationship of median debt of students at graduation (per 1K) to count of institutions.

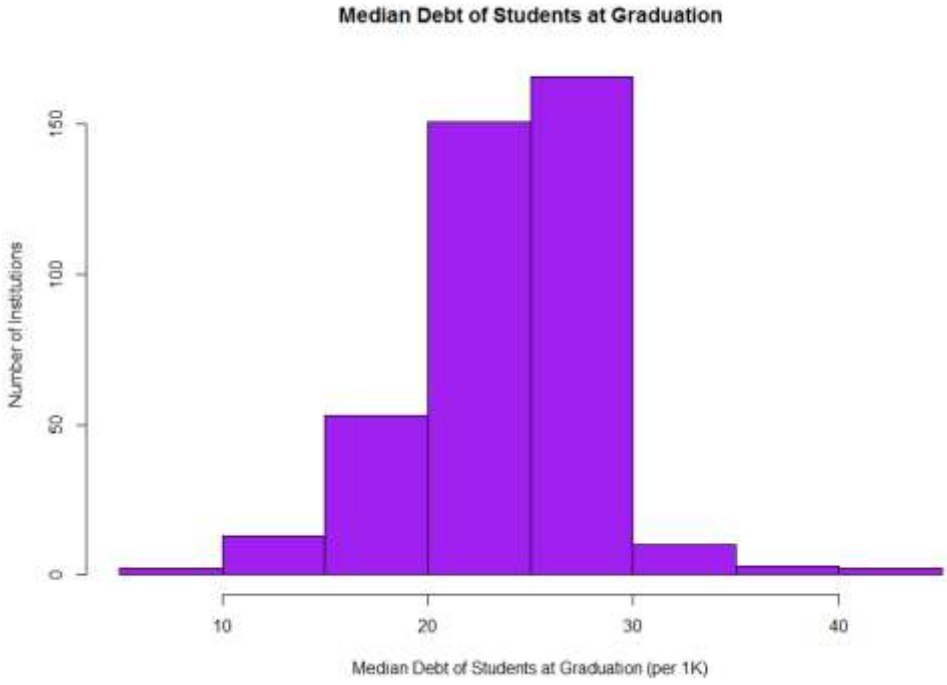

Figure 11. Bar graph showing relationship of median debt of students at graduation (per 1K) to count of institutions.

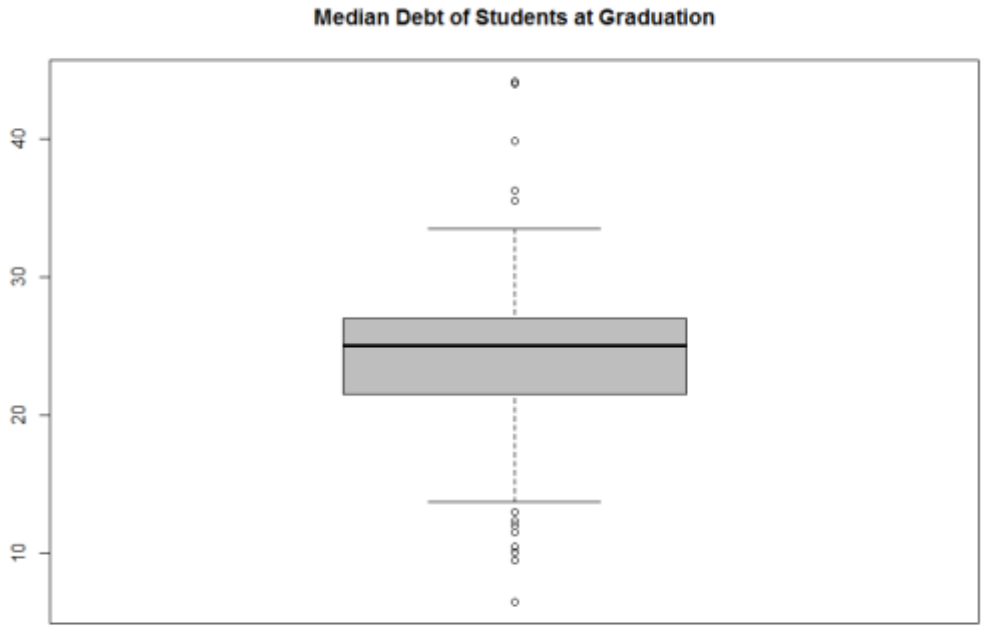

Figure 12. Box plot demonstrating five-number summary of median debt of students at graduation (per 1K).

Summary Statistics											
n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
400	0.00	23.94	4.48	-0.24	3.04	6.50	21.50	25.00	27.00	44.15	5.50

Table 4. Table displaying summary statistics including the five-number summary of median debt of students at graduation (per 1K) at each institution.

Commentary:

The data has 15 outliers with 10 on the lower end of the data and 5 on the higher end. Data appears close to a normal distribution, but has a steep drop-off after 30K.

Average Age of Entry:

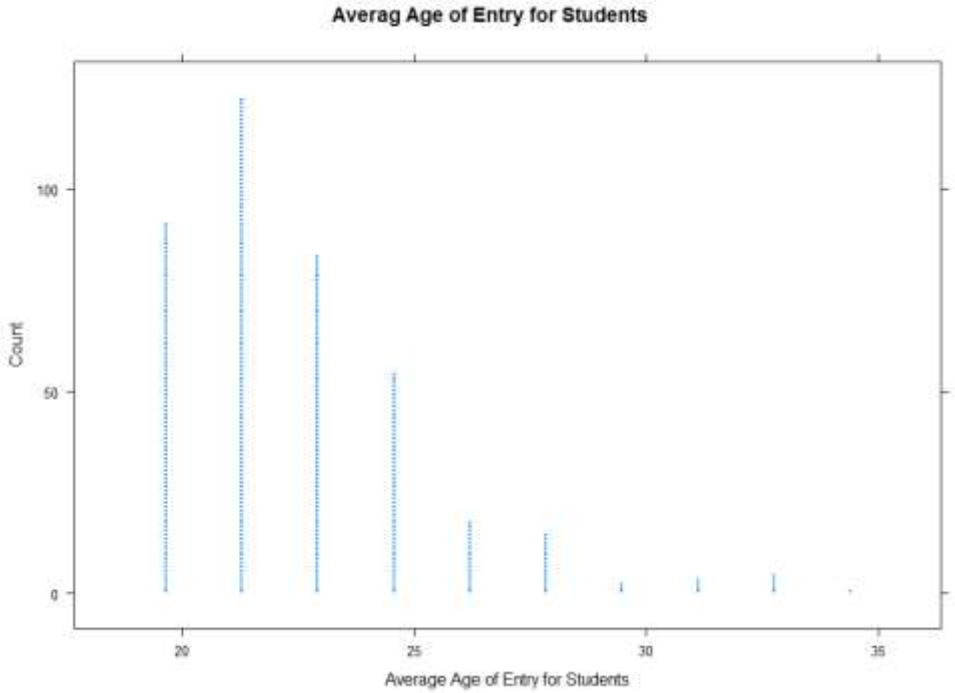

Figure 13. Dot plot showing relationship of average age of entry for students to count of institutions.

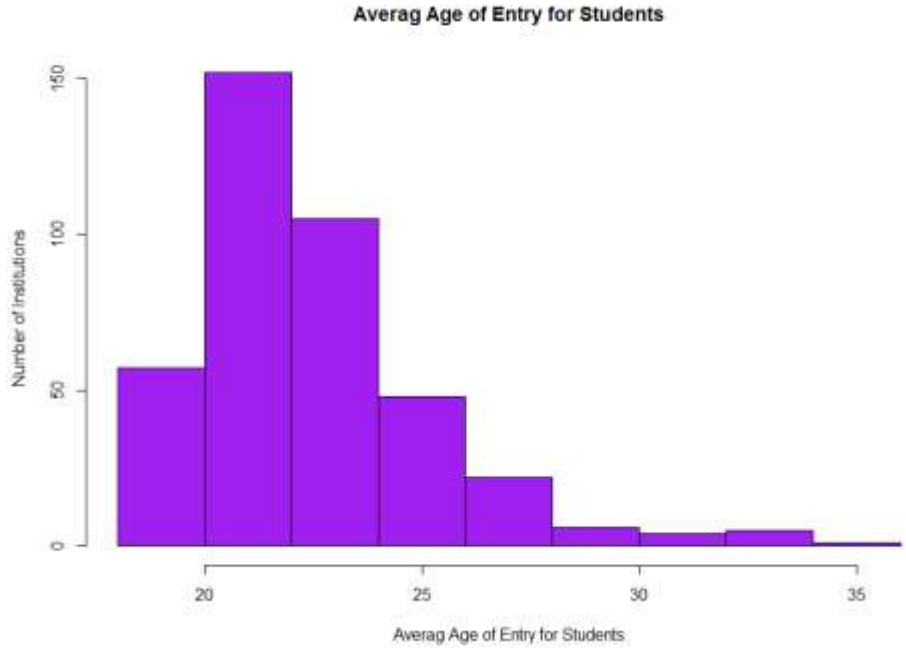

Figure 14. Bar graph showing relationship of average age of entry for students to count of institutions.

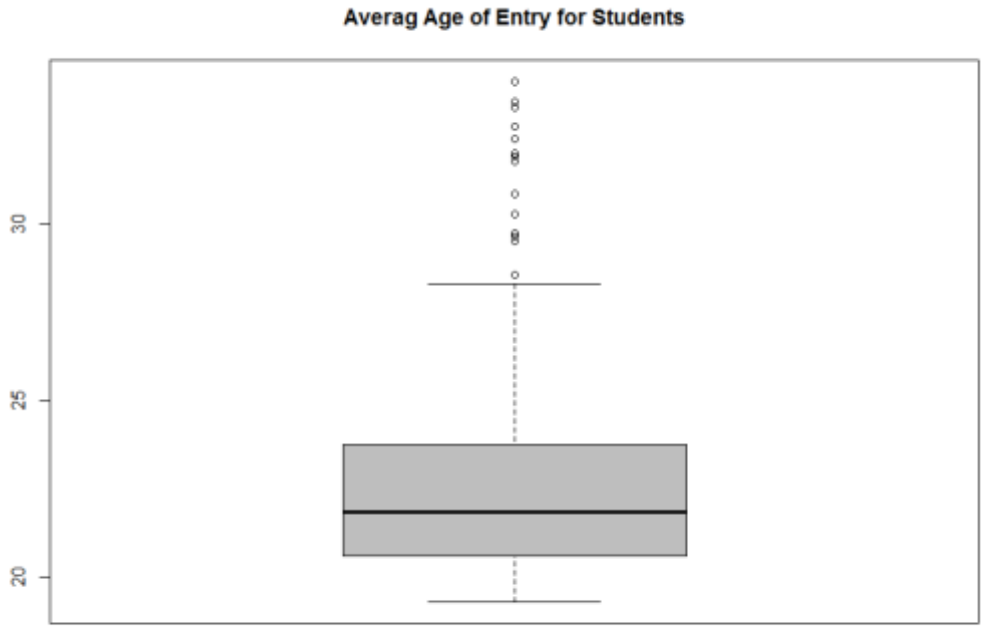

Figure 15. Box plot demonstrating five-number summary of average age of entry for students.

Summary Statistics											
n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
400	0.00	22.51	2.67	1.60	3.26	19.30	20.60	21.84	23.76	34.03	3.16

Table 5. Table displaying the summary statistics including the five-number summary of the age of entry for students for each institution.

Commentary:

The data has 14 outliers at the high end of the data and is heavily skewed right.

First Generation Students:

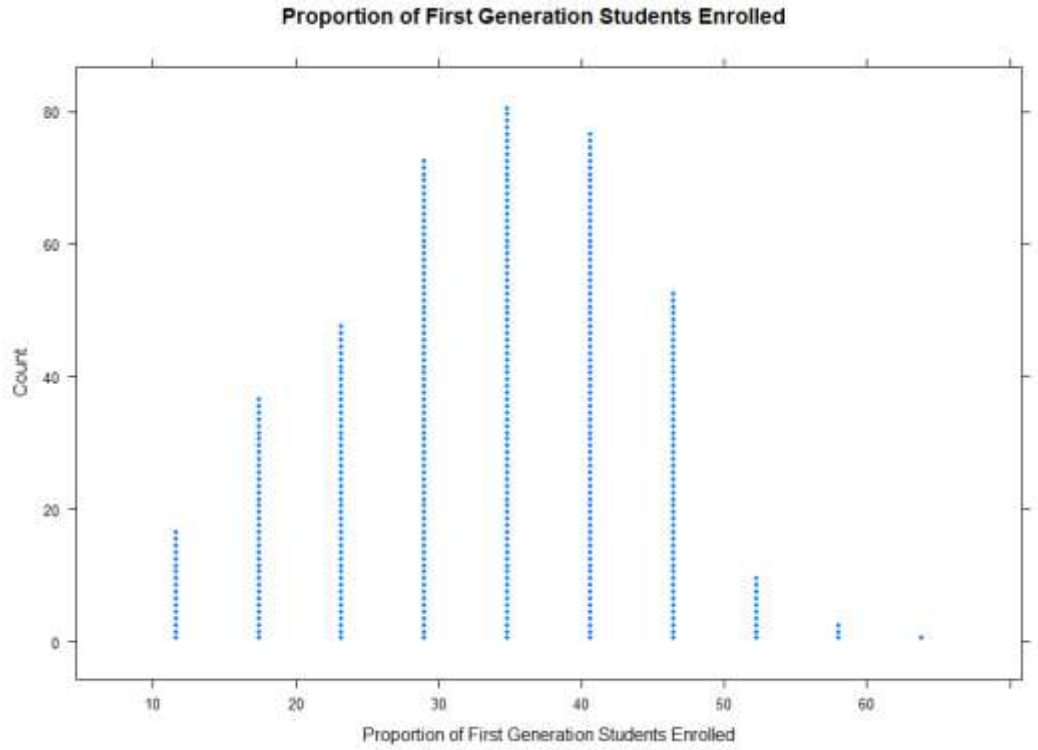

Figure 16. Dot plot showing relationship of proportion of first generation students enrolled to count of institutions.

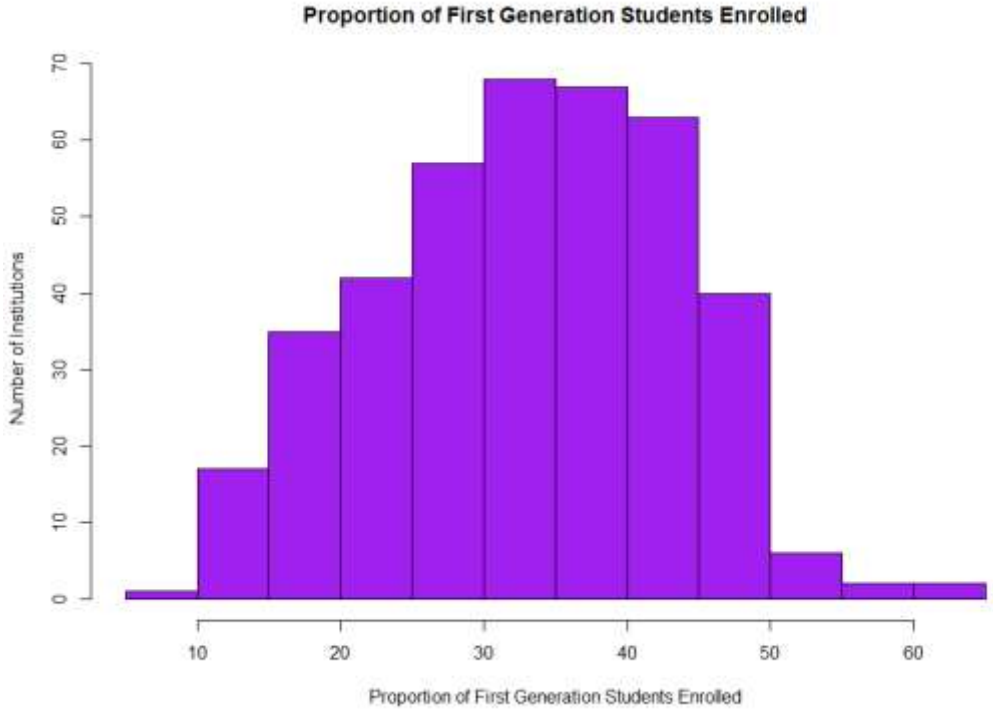

Figure 17. Bar graph showing relationship of proportion of first generation students enrolled to count of institutions.

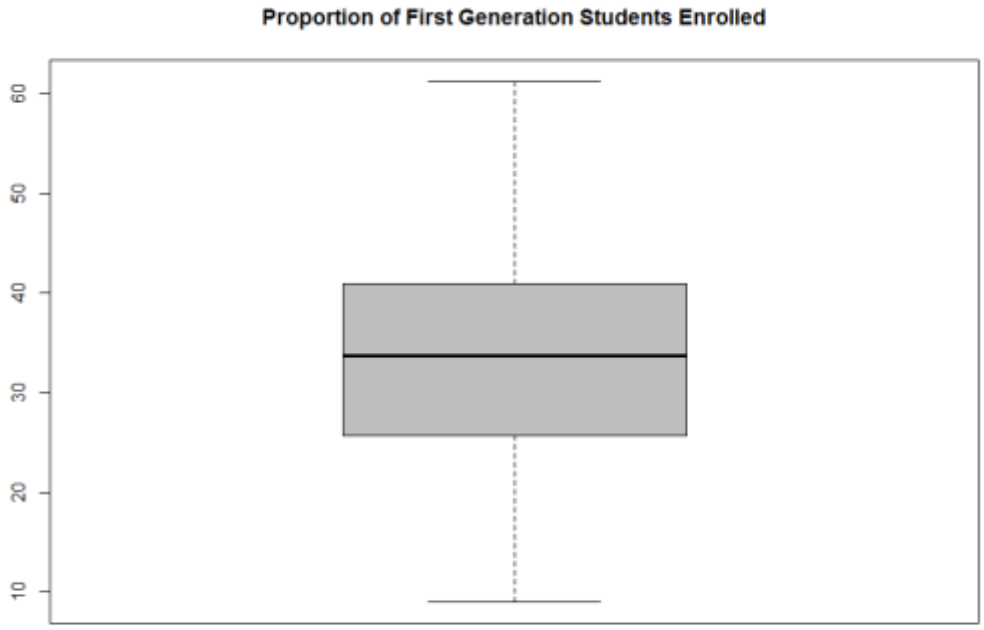

Figure 18. Box plot demonstrating five-number summary of proportion of first generation students enrolled at each institution.

Summary Statistics											
n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
400	0.00	33.05	10.28	-0.10	-0.62	9.03	25.85	33.67	40.88	61.27	15.03

Table 6. Table displaying summary statistics including the five-number summary of proportion of first generation students at each institution.

Commentary:

The data appears normally distributed with an abrupt drop-off at 50% and there are no outliers.

Graduation Rate:

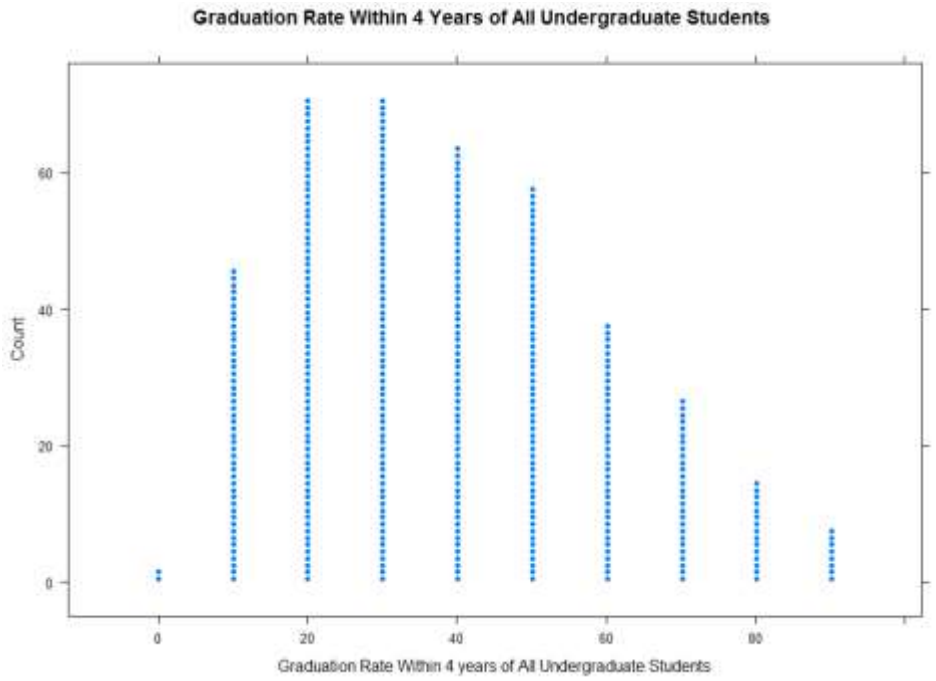

Figure 19. Dot plot showing relationship of graduation rate within 4 years of all undergraduate students to count of institutions.

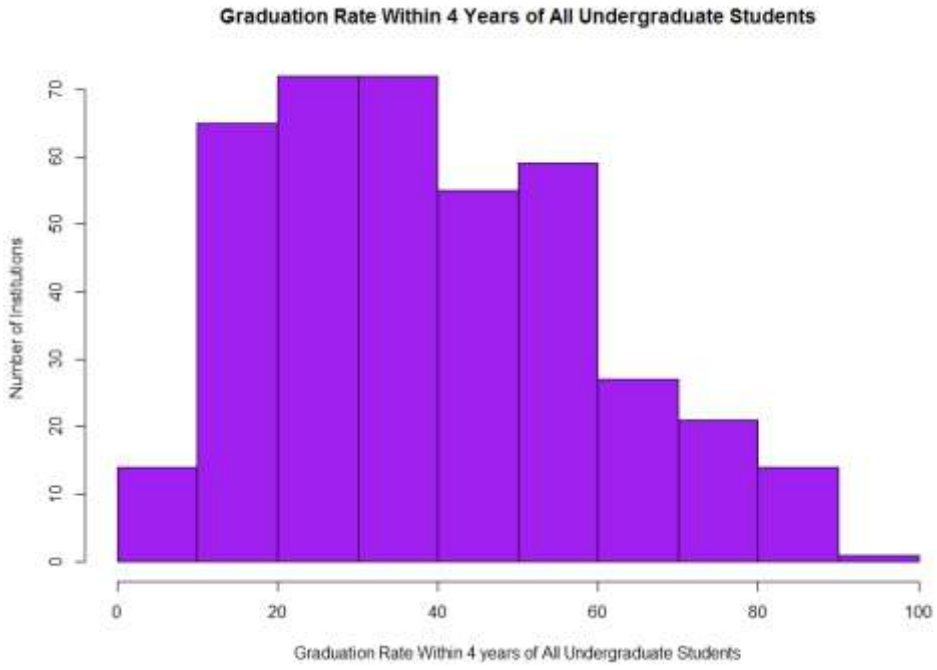

Figure 20. Bar graph showing relationship of graduation rate within 4 years of all undergraduate students to count of institutions.

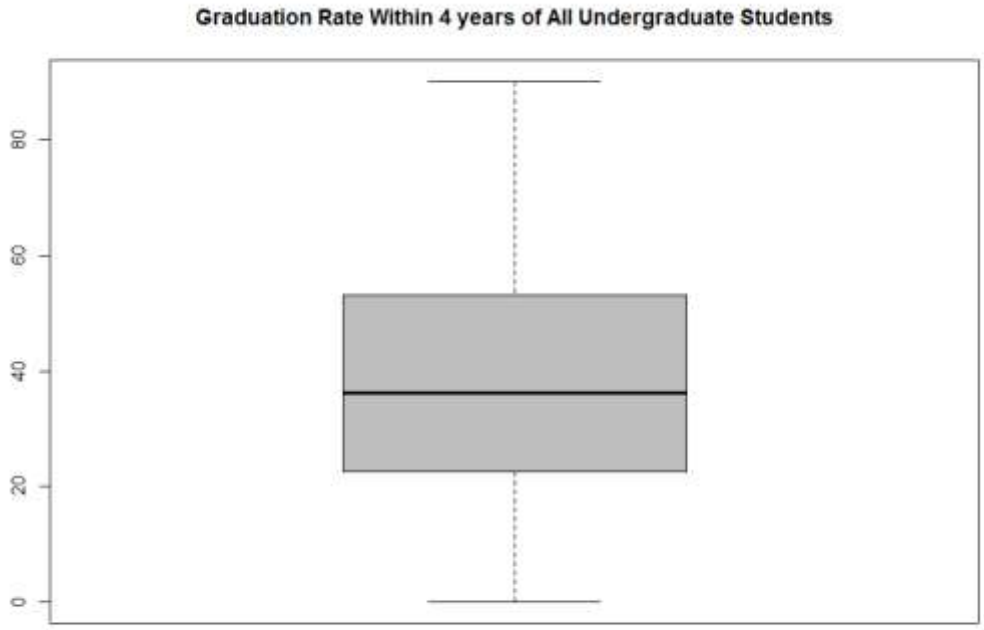

Figure 21. Box plot demonstrating five-number summary of graduation rate within 4 years of all undergraduate students.

Summary Statistics											
n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
400	0.00	39.08	20.18	0.43	-0.58	0.00	22.63	36.31	53.11	90.20	30.48

Table 7. Table displaying summary statistics including the five-number summary of graduation rate at each institution.

Commentary:

Data is very slightly skewed to the right, but almost normally distributed and there are no outliers.

Median family income:

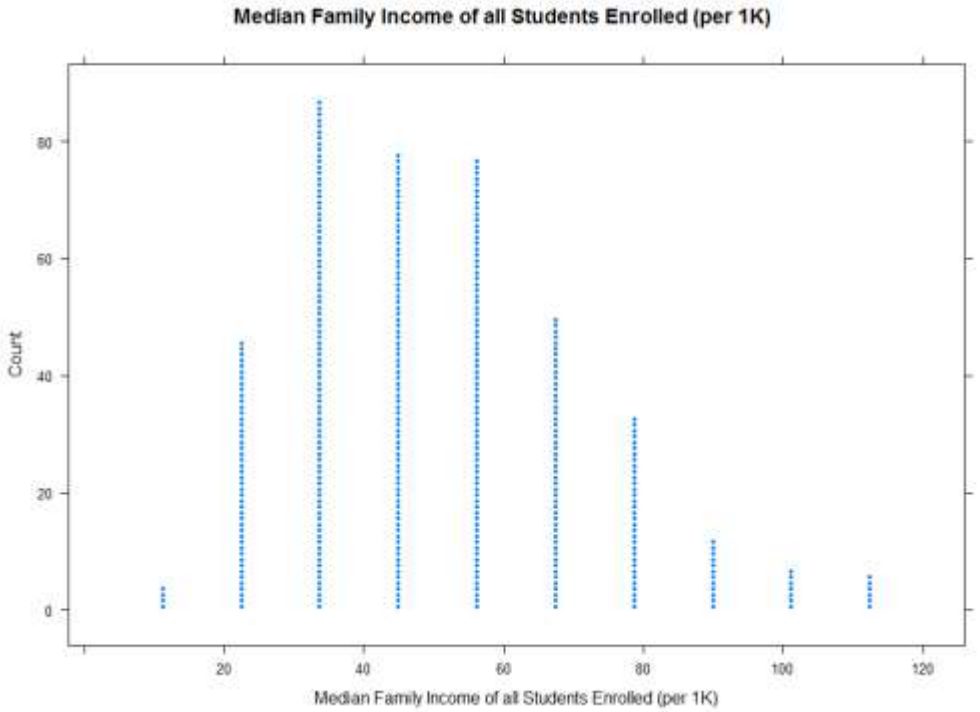

Figure 22. Dot plot showing relationship of median family income of all students enrolled (per 1K) to count of institutions.

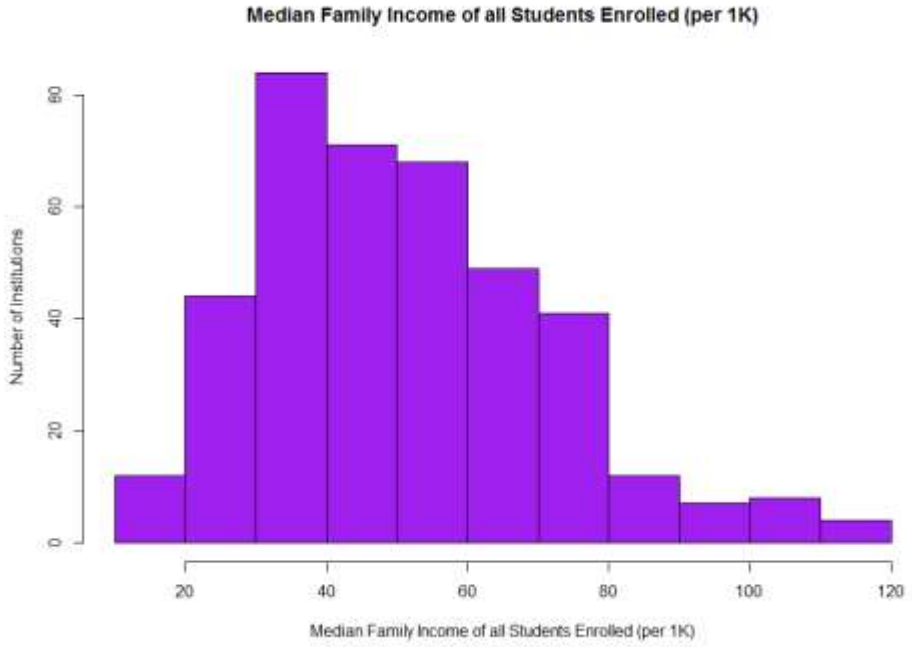

Figure 23. Bar graph showing relationship of median family income of all students enrolled (per 1K) to count of institutions.

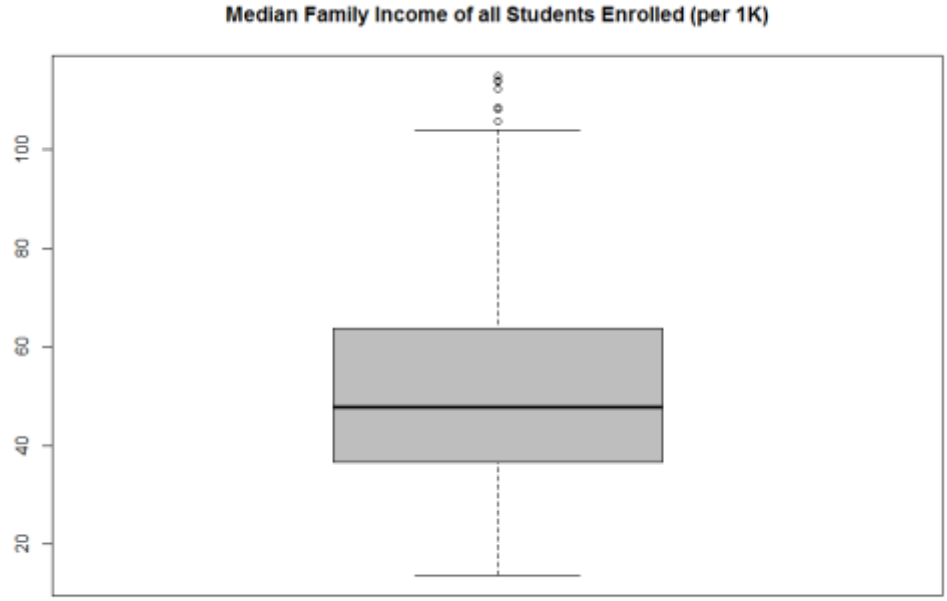

Figure 24. Box plot demonstrating five-number summary of median family income of all students enrolled (per 1K).

Summary Statistics											
n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
400	0.00	50.89	20.37	0.67	0.25	13.71	36.63	47.91	63.52	114.94	26.89

Table 8. Table displaying summary statistics including the five-number summary of median family income of all students enrolled (per 1K) at each institution.

Commentary:

Plots for median family income of all students represents data from all 400 institutions. Data is skewed right and there are 7 outliers at the higher end of the data.

Categorical Variables

Control of Institution (Private/Public):

Frequency Table			
	Private	Public	Total
Count:	255	145	400
Percent:	63.7	36.2	100

Table 9. Table showing frequency count and percentage for private and public institutions in sample.

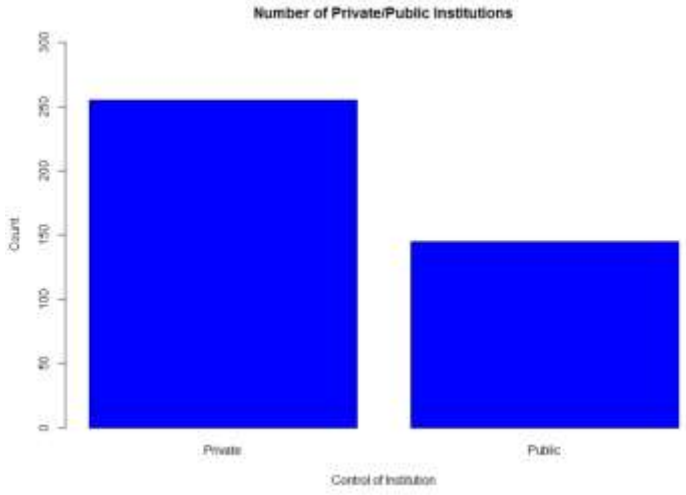

Figure 25. Table showing the number of private and public institutions represented in sample.

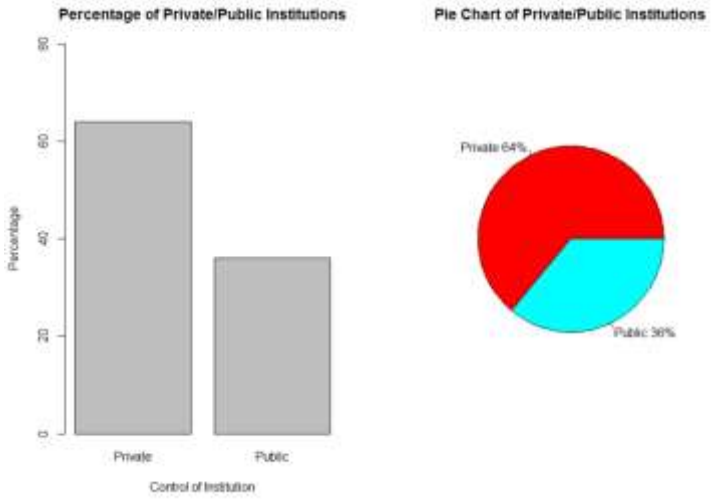

Figure 26. Bar graph and pie chart demonstrating the proportions of private and public institutions represented in sample.

Type of Institution (Non-Research/Research):

Frequency Table			
	Non-research	Research	Total
Count:	313	87	400
Percent:	78.2	21.7	100

Table 10. Table showing frequency count and percentage for non-research and research institutions in sample.

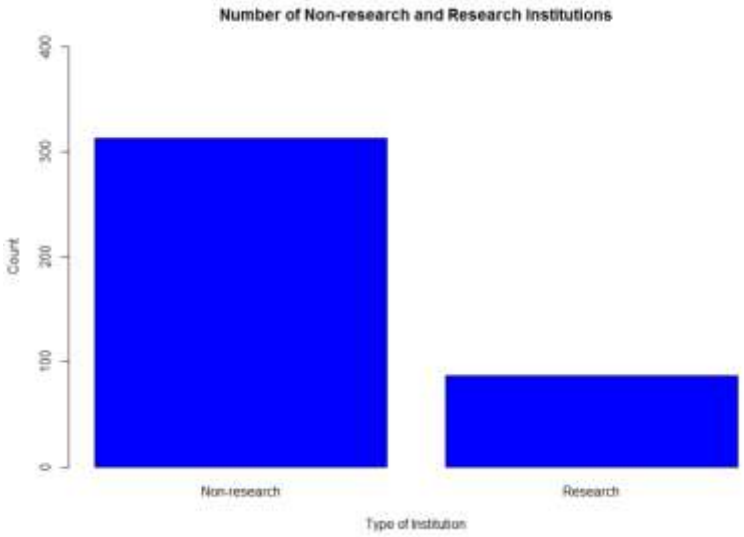

Figure 29. Bar graph showing count of non-research and research institutions represented in sample.

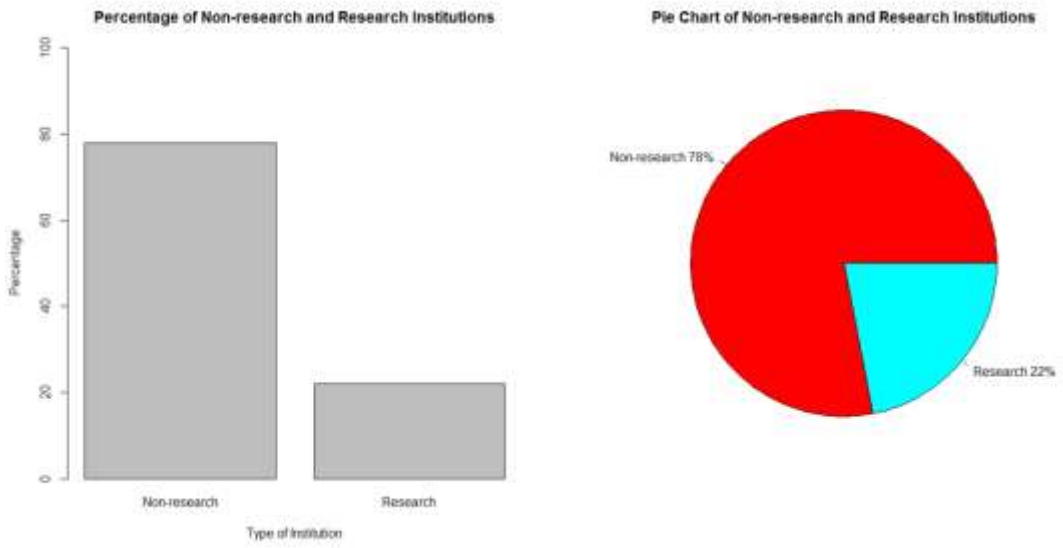

Figure 30. Bar graph and pie chart demonstrating the proportions of non-research and research institutions represented in sample.

First Generation Students by Research Type:

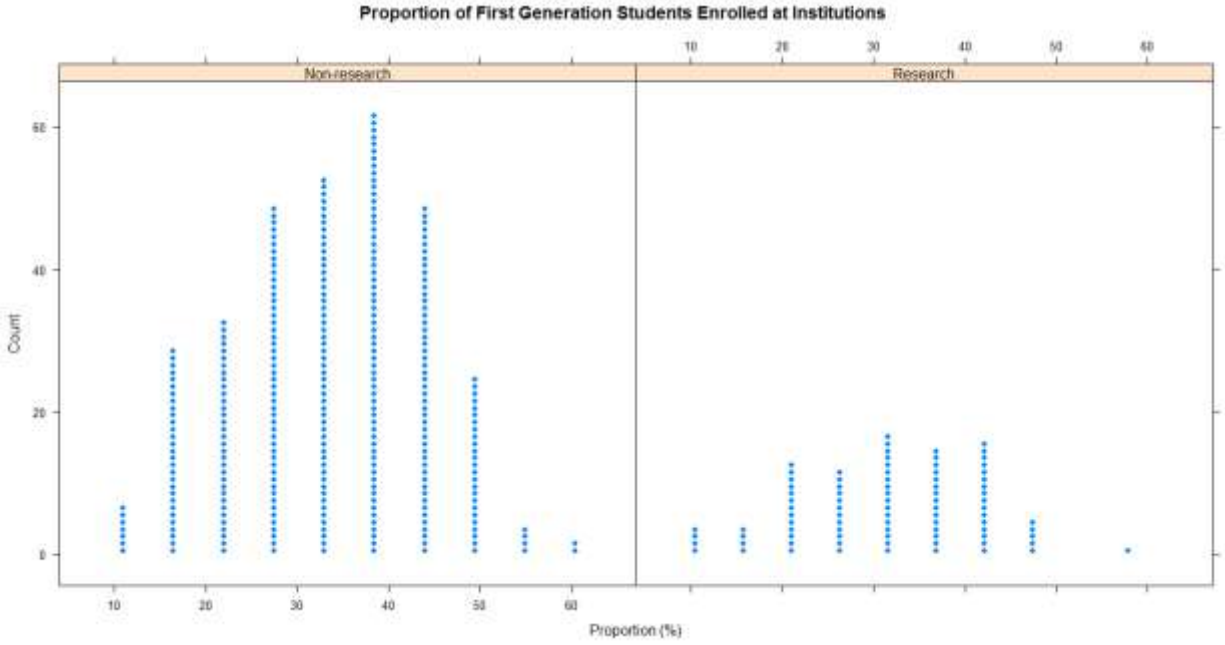

Figure 31. Dot plot showing relationship of proportion of first generation students enrolled to count of institutions by research type.

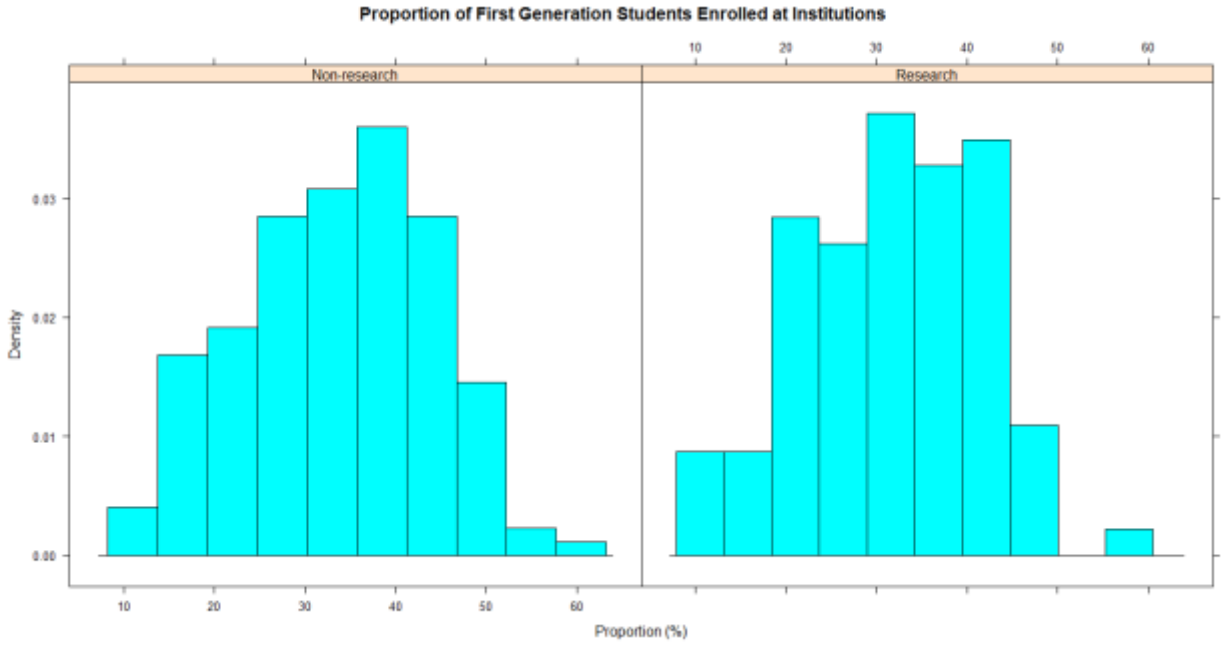

Figure 32. Bar graph showing relationship of proportion of first generation students enrolled to count of institutions by research type.

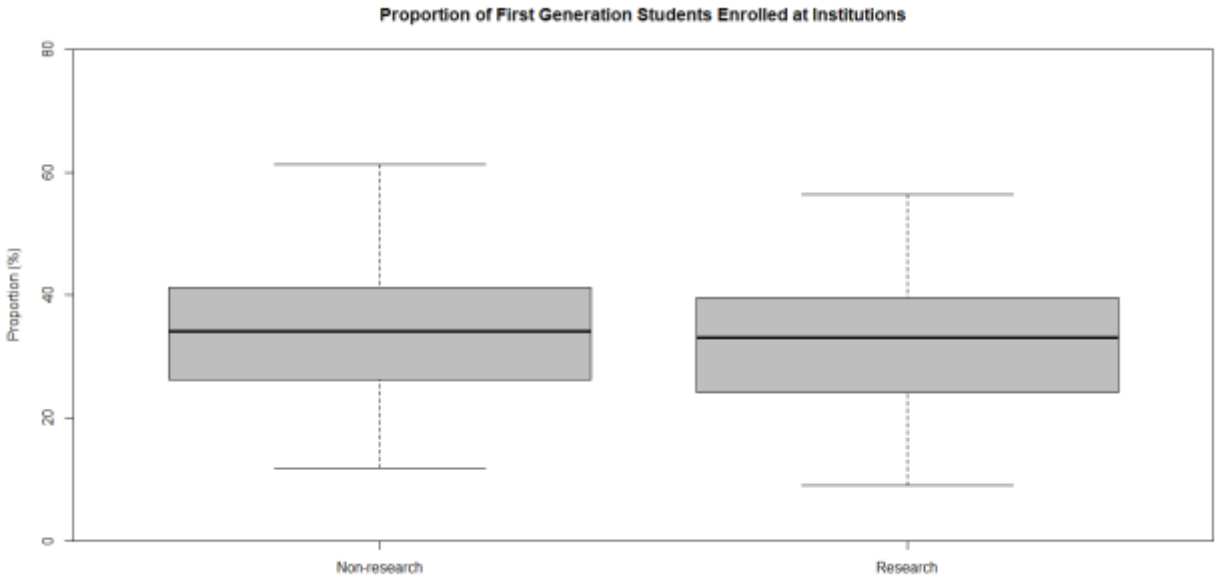

Figure 33. Box plot demonstrating five-number summary of proportion of first generation students enrolled at each institution by research type.

Summary Statistics												
	n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
Non-research	313.00	0.00	33.39	10.40	-0.09	-0.66	11.79	26.25	34.07	41.30	61.27	15.05
Research	87.00	0.00	31.81	9.79	-0.19	-0.50	9.03	24.22	33.04	39.57	56.41	15.36

Table 11. Table displaying summary statistics including the five-number summary of proportion of first generation students at each institution by research type.

Commentary:

The data contains a larger proportion of non-research institutions to research institutions, and non-research institutions have a much larger number of first generation students enrolled as opposed to research institutions. However, their mean proportions are close to one another, as can be seen in the box blots.

Graduation Rate by Research Type:

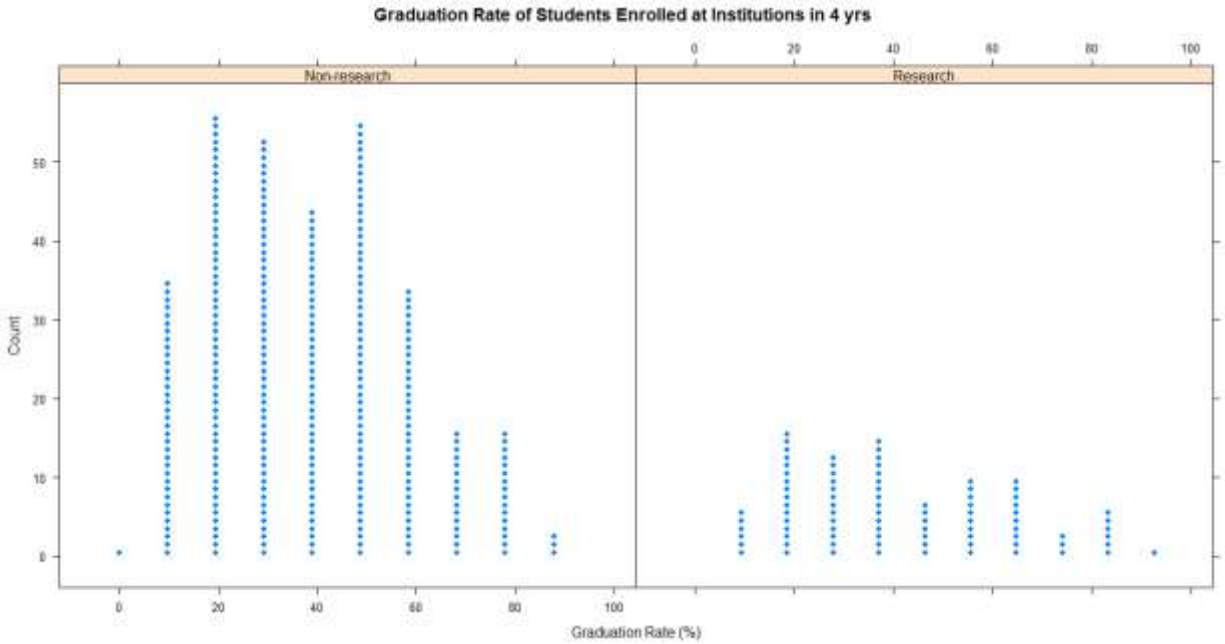

Figure 34. Dot plot showing relationship of graduation rate within 4 years of all undergraduate students to count of institutions by research type.

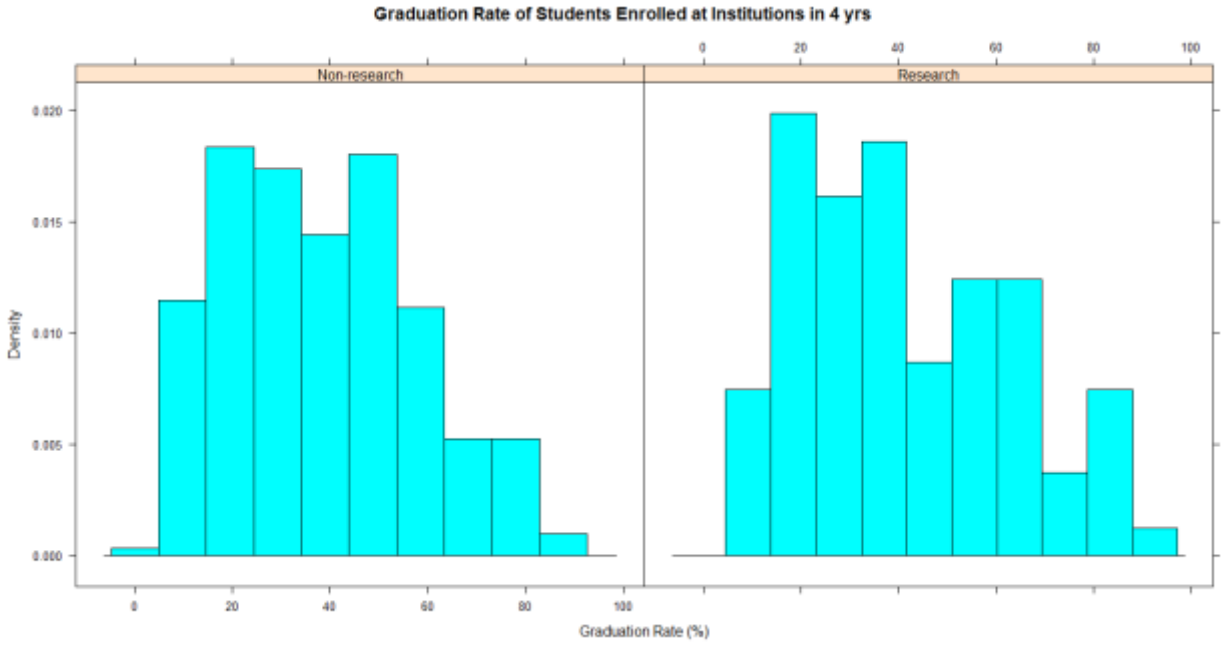

Figure 35. Bar graph showing relationship of graduation rate within 4 years of all undergraduate students to count of institutions by research type.

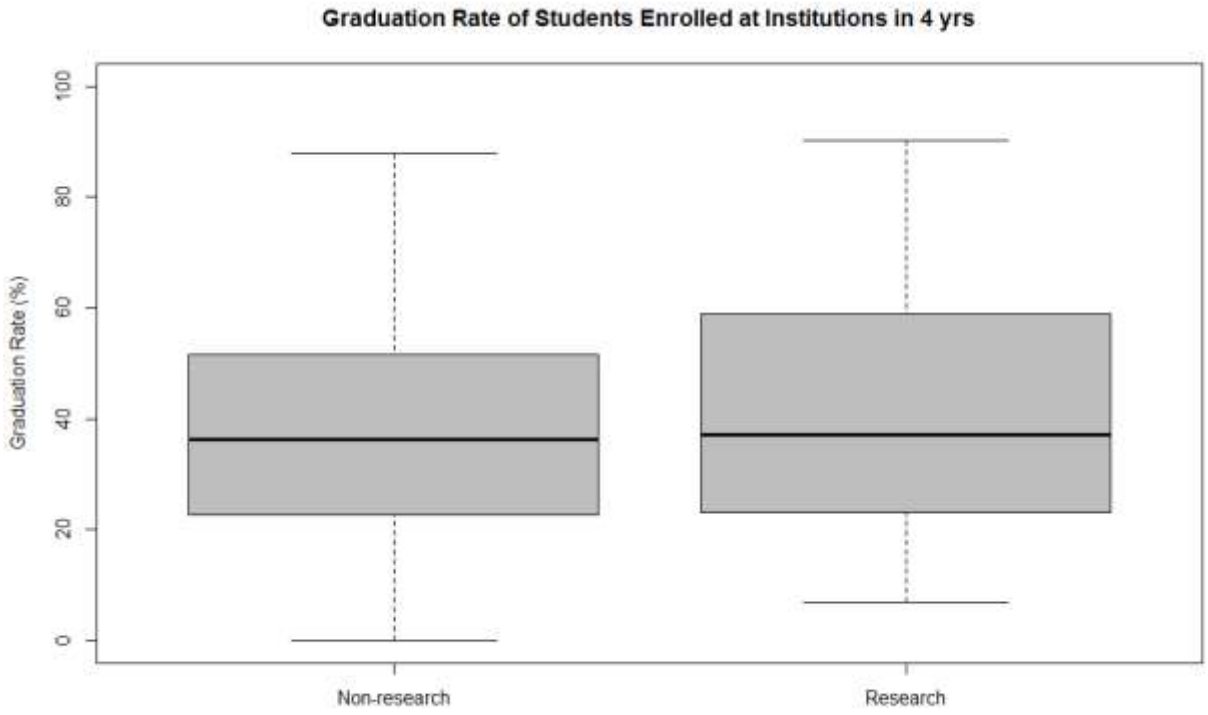

Figure 36. Box plot demonstrating five-number summary of graduation rate within 4 years of all undergraduate students by research type.

Summary Statistics												
	n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
Non-research	313.00	0.00	38.33	19.42	0.39	-0.58	0.00	22.58	36.25	51.63	87.77	29.05
Research	87.00	0.00	41.77	22.62	0.45	-0.81	6.86	23.08	37.18	59.02	90.20	35.95

Table 12. Table displaying summary statistics including the five-number summary of graduation rate at each institution by research type.

Commentary:

Again, the data contains a larger proportion of non-research institutions to research institutions, yet graphically their statistics look similar and the five number summary stats and boxplots show the mean graduation rates are close to one another.

Average Age of Entry by Research Type:

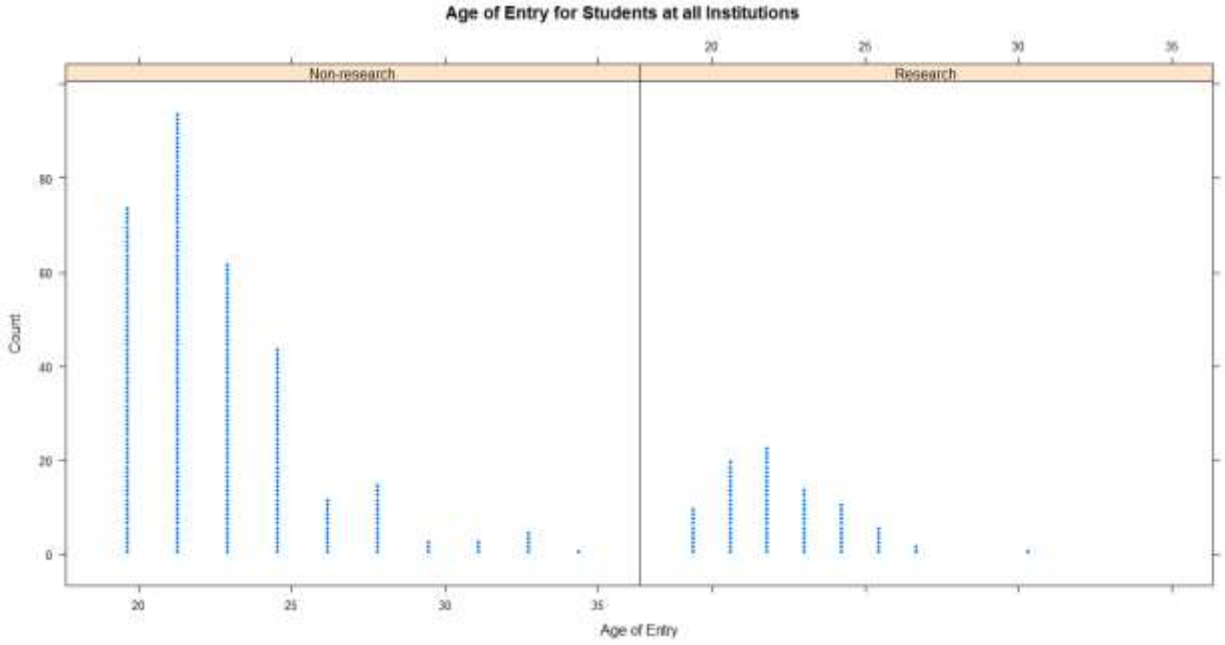

Figure 38. Dot plot showing relationship of average age of entry for students to count of institutions by research type.

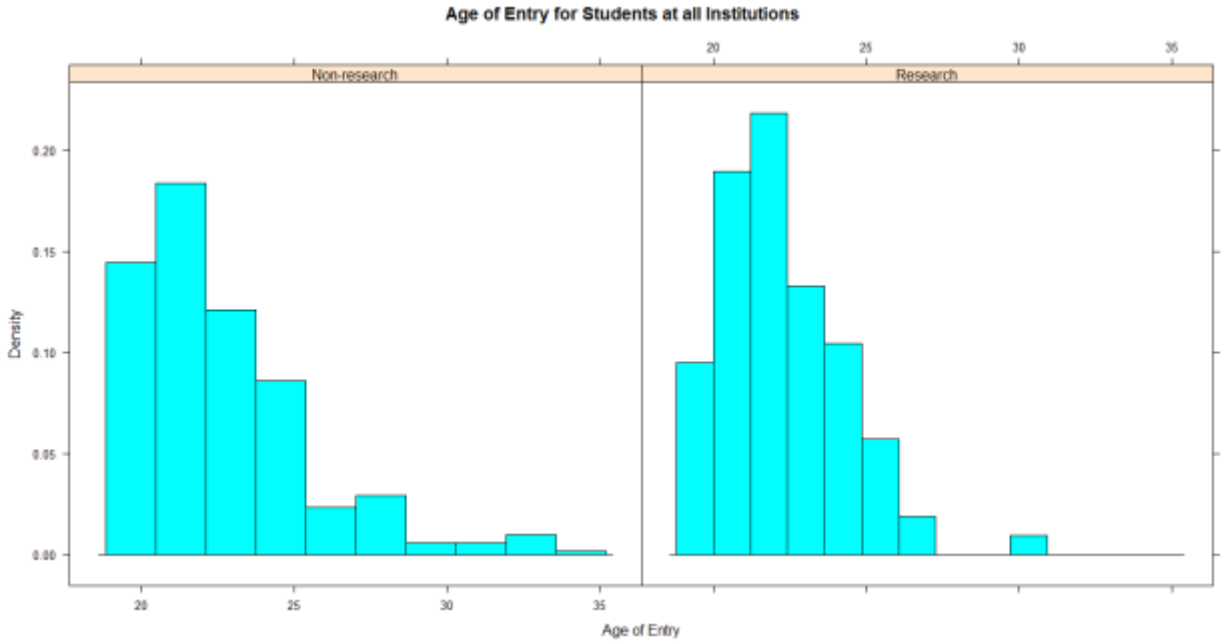

Figure 39. Bar graph showing relationship of average age of entry for students to count of institutions by research type.

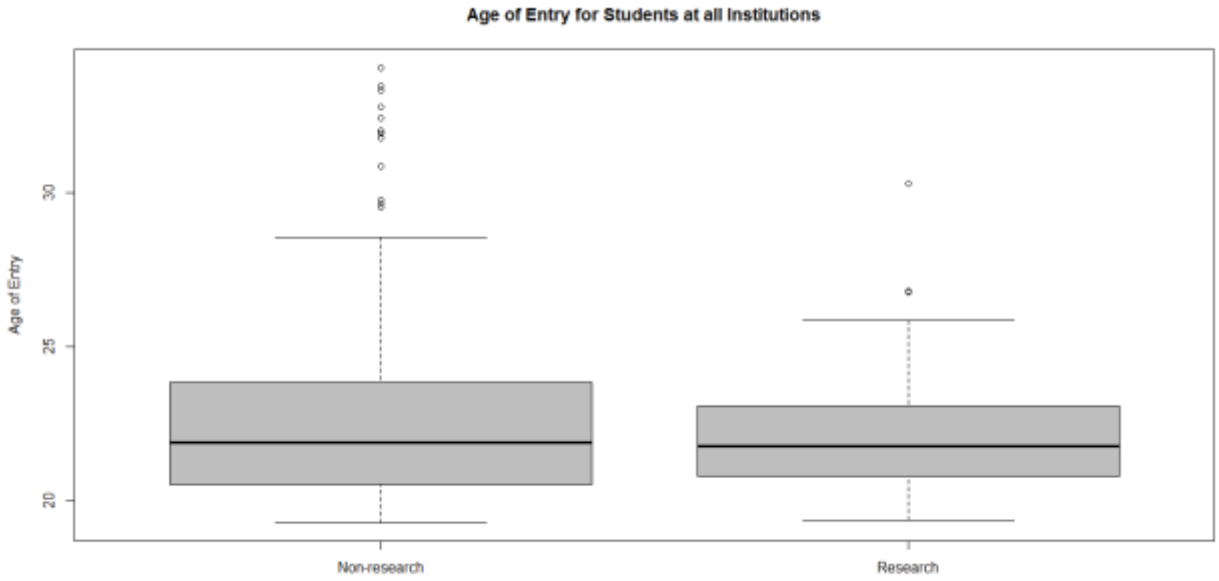

Figure 40. Box plot demonstrating five-number summary of average age of entry for students by research type.

Summary Statistics												
	n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
Non-research	313.00	0.00	22.60	2.83	1.57	2.90	19.30	20.54	21.89	23.85	34.03	3.31
Research	87.00	0.00	22.20	2.00	1.15	2.12	19.37	20.79	21.79	23.06	30.28	2.27

Table 13. Table displaying the summary statistics including the five-number summary of the age of entry for students for each institution by research type.

Commentary:

The data still shows a right skew with the breakdown of each research type with research institutions having a smaller spread versus non-research institutions. Mean ages of entry are similar, yet non-research institutions have a much larger number of outliers in addition to a larger spread concerning the age of entry.

Median Debt by Research Type:

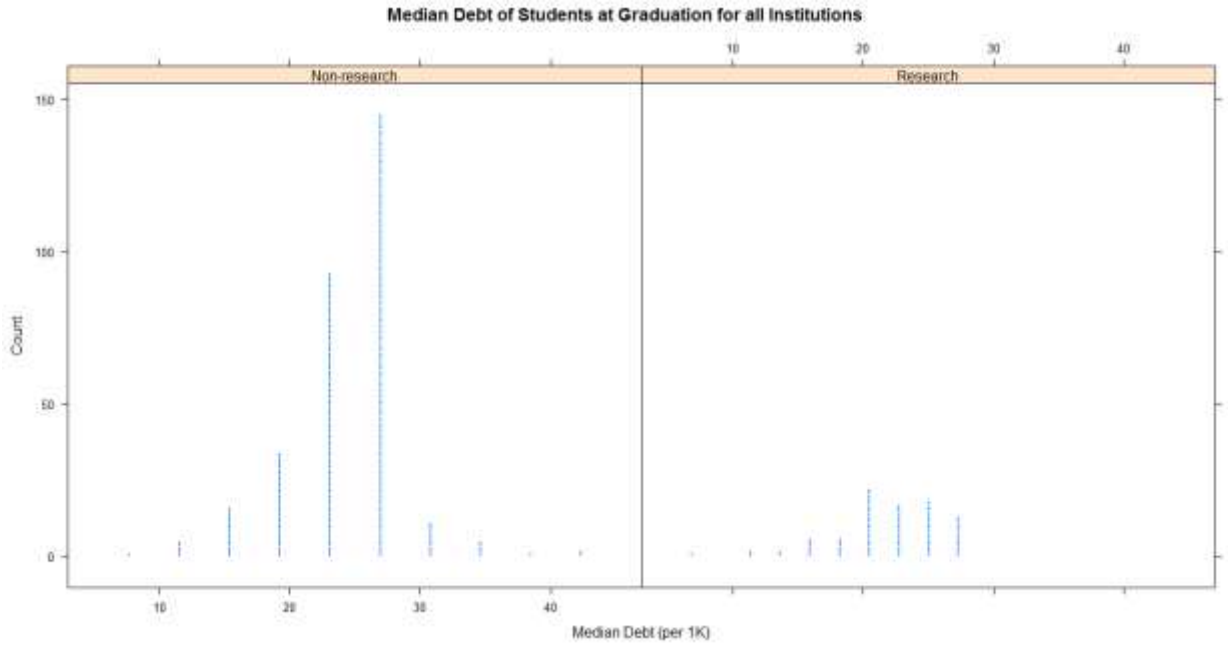

Figure 41. Dot plot showing relationship of median debt of students at graduation (per 1K) to count of institutions by research type.

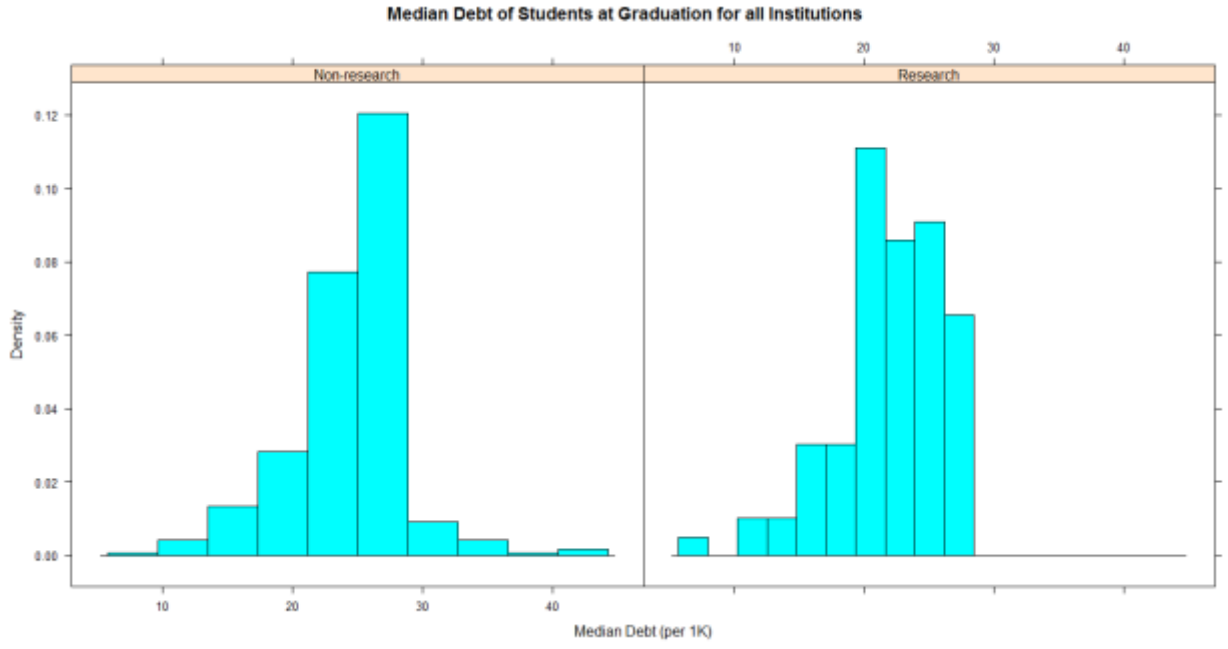

Figure 42. Bar graph showing relationship of median debt of students at graduation (per 1K) to count of institutions by research type.

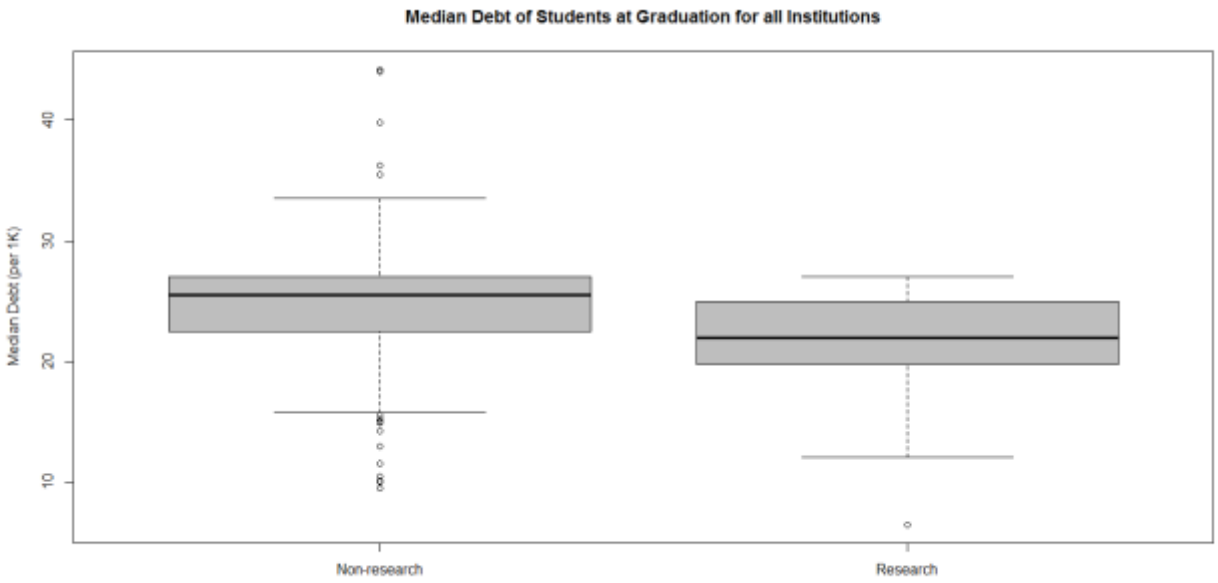

Figure 43. Table displaying summary statistics including the five-number summary of median debt of students at graduation (per 1K) at each institution by research type.

Summary Statistics												
	n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
Non-research	313.00	0.00	24.54	4.42	-0.15	3.57	9.50	22.50	25.50	27.00	44.15	4.50
Research	87.00	0.00	21.78	4.05	-1.01	1.53	6.50	19.75	21.94	25.00	27.00	5.25

Table 14. Table displaying summary statistics including the five-number summary of median debt of students at graduation (per 1K) at each institution by research type.

Commentary:

Data for each has an abrupt drop-off after 30K and the spread for research institutions is smaller than that of non-research. Non-research institutions have many more outliers on the low and high end of the data.

SECTION III:

Inference on Single Mean and Difference in Means (Classical):

CI for Single Mean:

Admissions Rate:
We are 95% confident that the average admissions rate for all institutions is between 63.2% and 67.2%.

In-State Tuition:
We are 95% confident that the average in-state tuition cost for all institutions is between 21.1K and 23.5K dollars.

Out-of-State Tution:
We are 95% confident that the average out-of-state tuition cost for all institutions, is between 25.3K and 27.1K dollars.

Median Debt:
We are 95% confident that the average median debt for students at graduation at all institutions lies between 23.5K and 24.4K dollars.

Average Age of Entry:
We are 95% confident that the average age of entry for students at all institution is between 22.3 years and 22.8 years.

First Generation Students:
We are 95% confident that the mean proportion of first generation students at all institutions lies between 32% and 34.1%.

Graduation Rate:
We are 95% confident that the average graduation rate for all institution is between 37.1% to 41.1%.

Median Family Income:
We are 95% confident that the average median family income for students at all institutions is between 48.9K and 52.9K dollars.

CI for Difference in Means:

First Generation Students by Research Type:
μ_1= true mean for the proportion of first generation students enrolled at all research institutions
μ_2= true mean for the proportion of first generation students enrolled at all non-research institutions
Null hypothesis: $\mu_1=\mu_2$
Alternative hypothesis: $\mu_1\neq\mu_2$
p-value= 0.05 < 0.1902, therefore we fail to reject the null hypothesis.

Based on the data, we do not have enough evidence to conclude that the true mean for the proportion of first generation students enrolled at all research institutions is not equal to the true mean for the proportion of first generation students enrolled at all non-research institutions.

Median Student Debt by Research Type:
μ_1= true mean of median student debt at graduation for students at all research insittutions
μ_2= true mean of median student debt at graduation for students at all non-research insitutions
Null hypothesis: $\mu_1=\mu_2$
Alternative hypothesis: $\mu_1\neq\mu_2$
p-value= 0.05 > 0.00, therefore we reject the null hypothesis.

We have enough evidence that the true mean of median student debt at graduation for students at all research insittutions is not equal true mean of median student debt at graduation for students at all non-research insitutions.

Average Age of Entry by Research Type:
μ_1= true mean for the average age of entry for students at all research institutions
μ_2= true mean for the average age of entry for students at all non-research institutions
Null hypothesis: $\mu_1=\mu_2$
Alternative hypothesis: $\mu_1\neq\mu_2$
p-value= 0.05 < 0.14, therefore we fail to reject the null hypothesis.

Based on the data, we do not have enough evidence to determine that the mean for the average age of entry for students at all research institutions is not equal to the mean for the average age of entry for students at all non-research institutions.

Graduation Rate by Research Type:
μ_1= true mean of graduation rate for all research institutions
μ_2= true mean of graduation rate for all non-research institutions
Null hypothesis: $\mu_1=\mu_2$
Alternative hypothesis: $\mu_1\neq\mu_2$
p-value= 0.05 < 0.20, therefore we fail to reject the null hypothesis.

We do not have enough evidence to conclude that the true mean of graduation rate for all research institutions is not equal to the true mean of graduation rate for all non-research institutions.

Inference on Single Mean and Difference in Means (Bootstrap Method):

Single Mean

Admissions Rate:

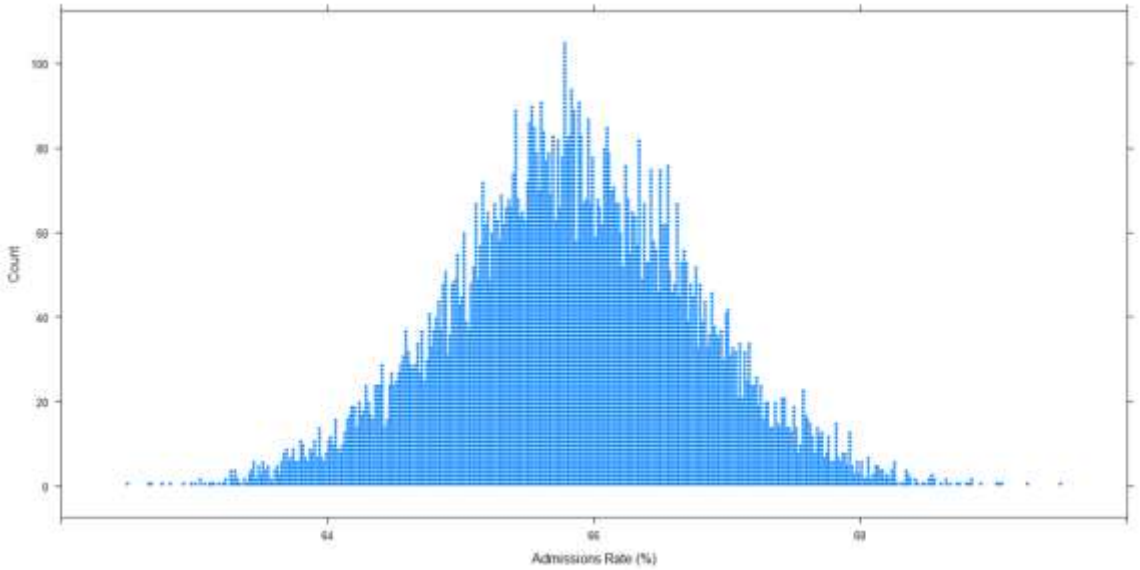

Figure 44. Dot plot showing the bootstrap distribution of a single mean for admissions rate.

We are 95% confident that the mean admissions rate of all institutions is between 64.003% and 67.681%.

In-State Tuition:

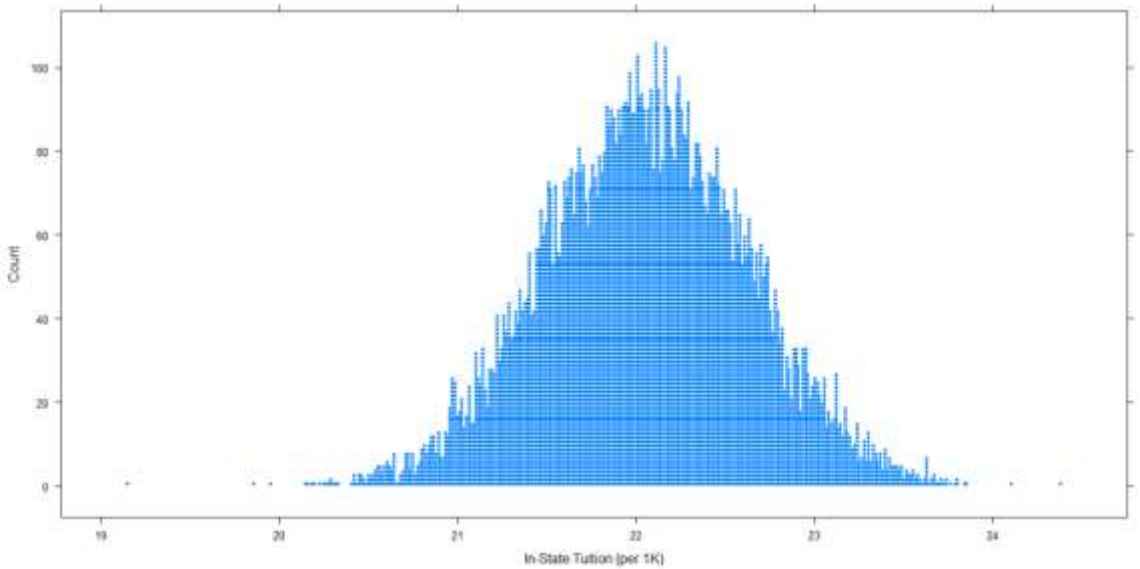

Figure 45. Dot plot showing the bootstrap distribution of a single mean for in-state tuition (per 1K).

We are 95% confident that the mean in-state tuition for all institutions is between 20.950K and 23.174K.

Out-of-State Tuition:

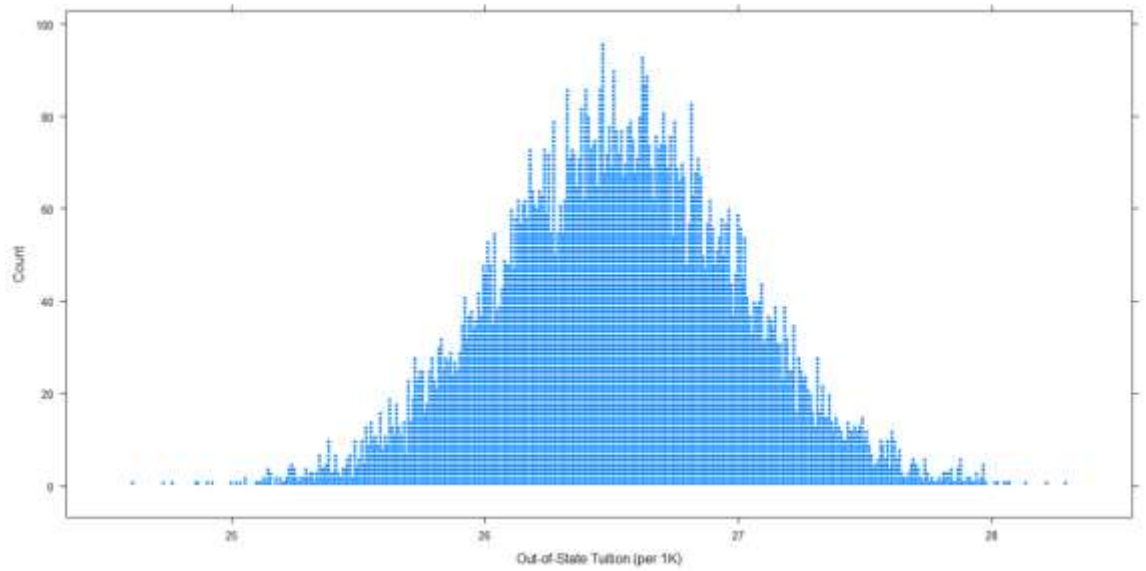

Figure 46. Dot plot showing the bootstrap distribution of a single mean for out-of-state tuition (per 1K).

We are 95% confident that the mean out-of-state tuition for all institutions is between 25.587K and 27.481K.

Median Debt:

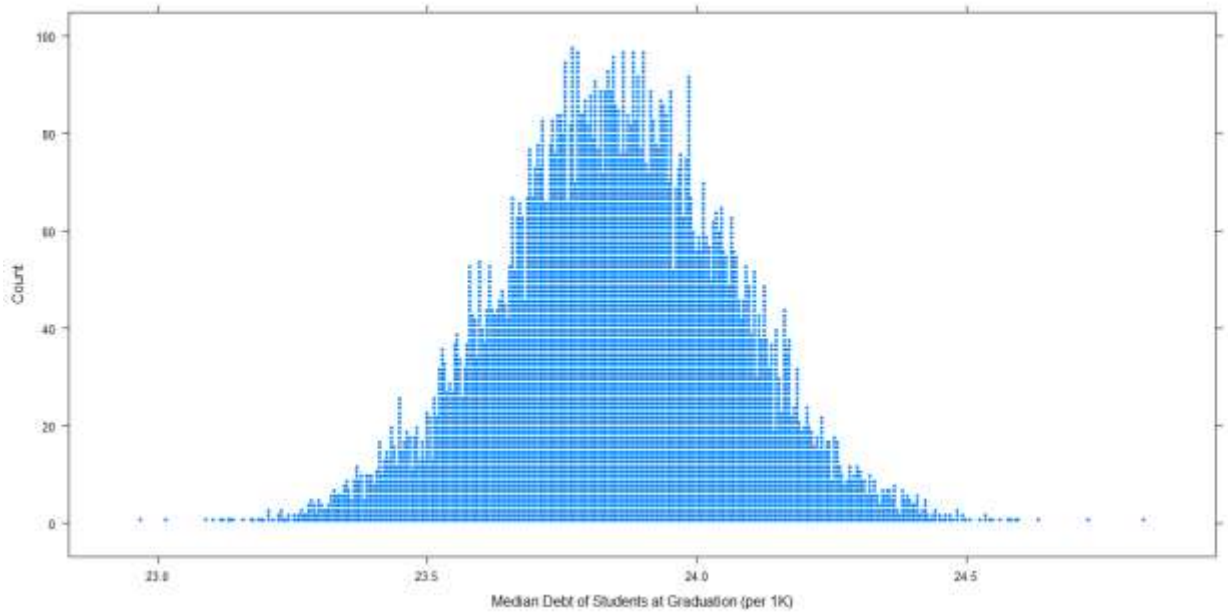

Figure 47. Dot plot showing the bootstrap distribution of a single mean for median debt of students at graduation (per 1K).

We are 95% confident that the mean of median debt for all students at all institutions is between 23.416K and 24.276K.

Average Age of Entry:

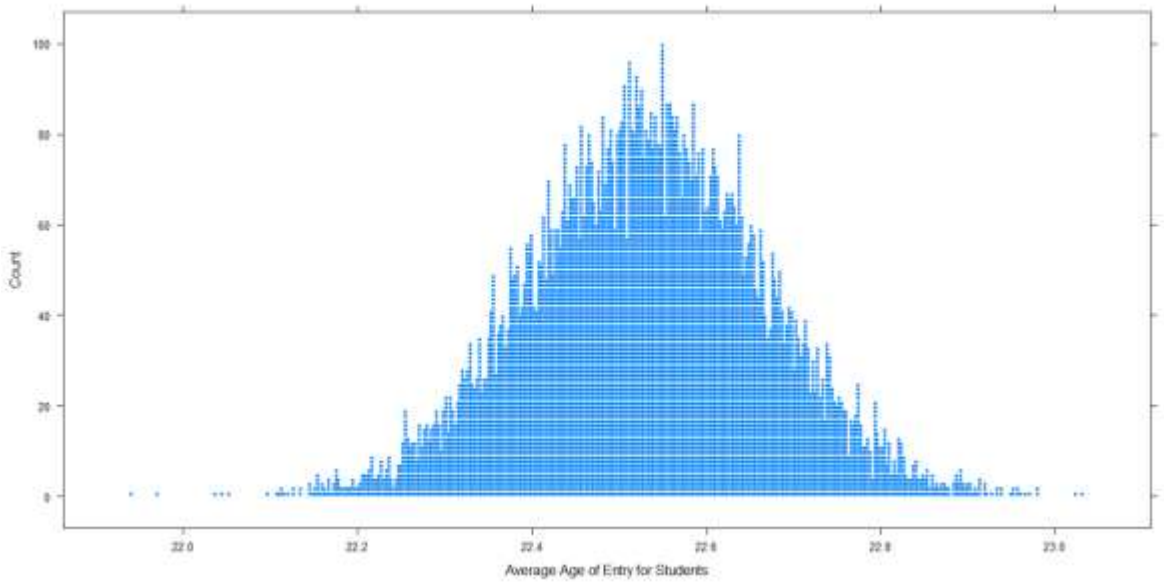

Figure 48. Dot plot showing the bootstrap distribution of a single mean for average age of entry for students.

We are 95% confident that the average age of entry for all students at all institutions is between 22.269 years and 22.795 years.

Proportion of First Generation Students:

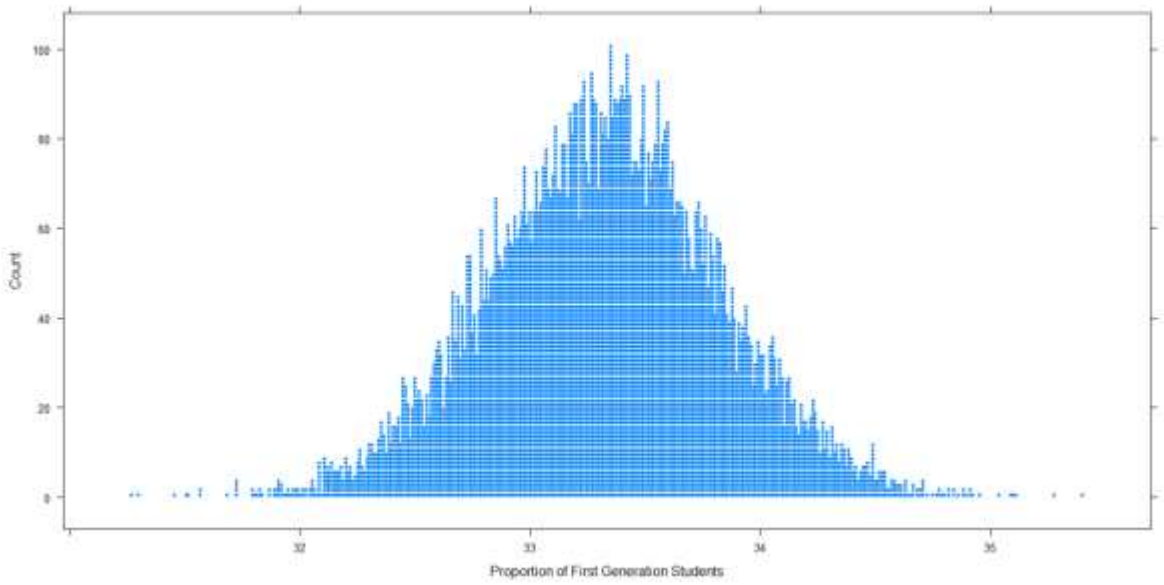

Figure 49. Dot plot showing the bootstrap distribution of a single mean for proportion of first generation students enrolled.

We are 95% confident that mean proportion of first generation students at all institutions is between 32.346% and 34.304%.

Graduation Rate:

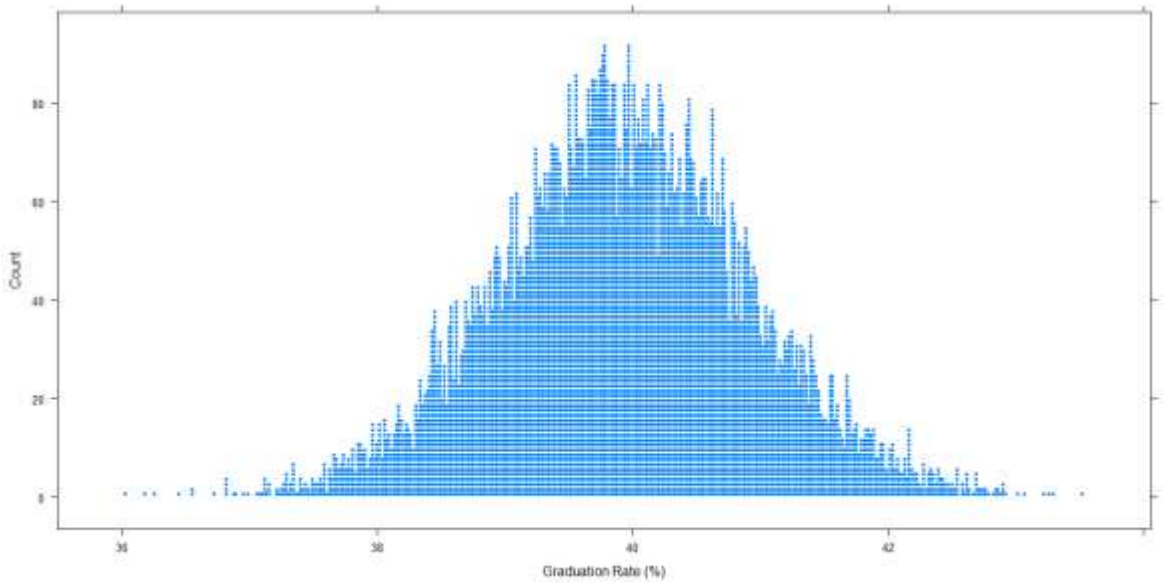

Figure 50. Dot plot showing the bootstrap distribution of a single mean for graduation rate (%).

We are 95% confident that the mean graduation rate at all institutions is between 38.130% and 41.902%.

Median Family Income:

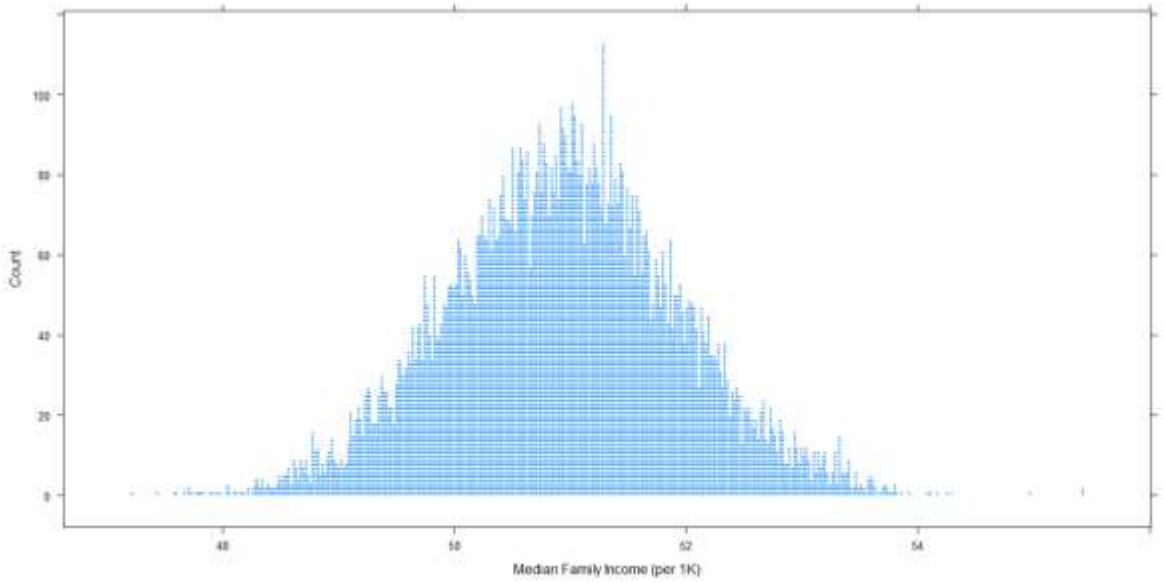

Figure 51. Dot plot showing the bootstrap distribution of a single mean for median family income of students (per 1K).

We are 95% confident that the mean median family income of all students at all institutions is between 49.004K and 52.940K.

Difference in Means

First Generation Students by Research Type:

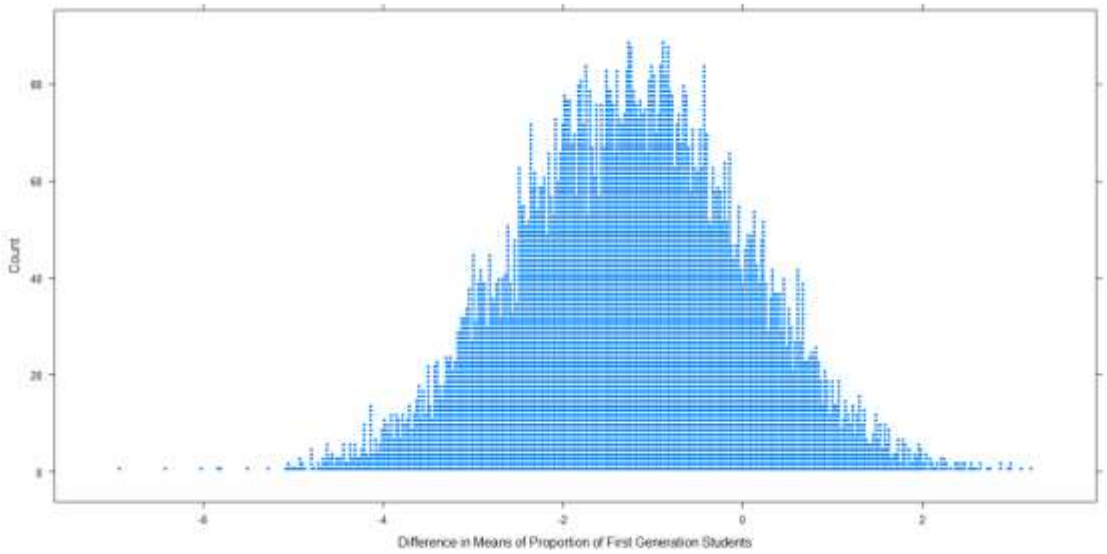

Figure 52. Dot plot showing the bootstrap distribution of difference in means between the mean proportion of first generation students enrolled at research institutions and the mean proportion of first generation students enrolled at non-research institutions.

Confidence Interval:

95% CI for difference in means: (-3.755, 1.234)

Graduation Rate by Research Type:

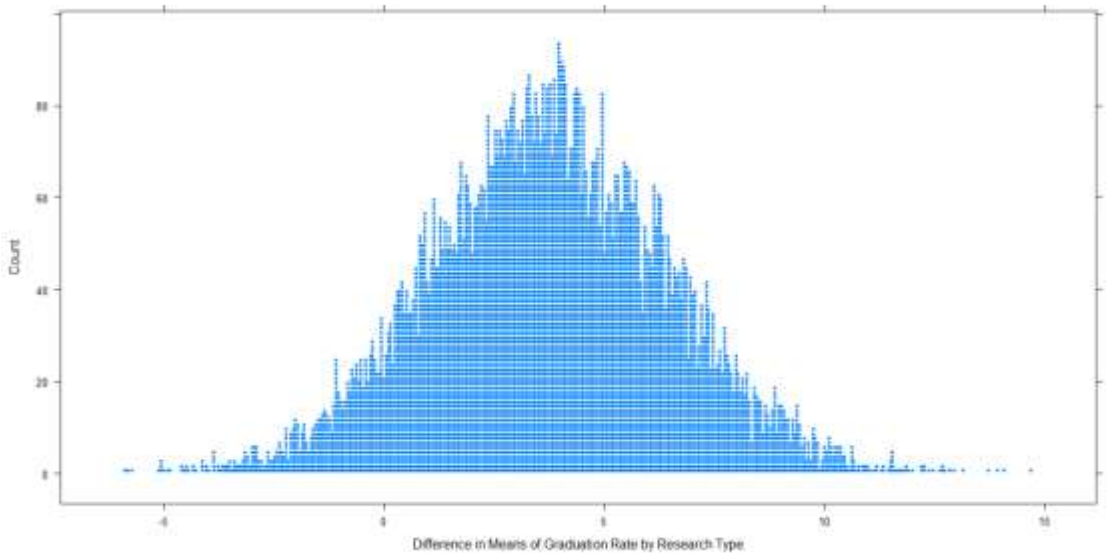

Figure 53. Dot plot showing the bootstrap distribution of difference in means between the mean graduation rate at research institutions and the mean graduation rate at non-research institutions.

Confidence Interval:

95% CI for difference in means: (-1.516, 9.113)

Average Age of Entry by Research Type:

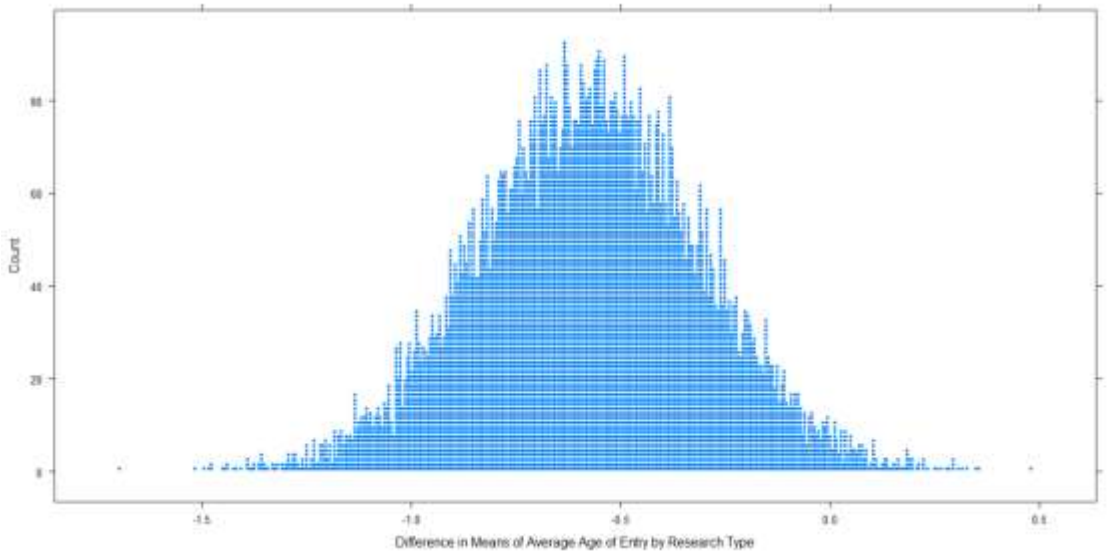

Figure 54. Dot plot showing the bootstrap distribution of difference in means between the mean average age of entry at research institutions and the mean average age of entry at non-research institutions.

Confidence Interval:

95% CI for difference in means: (-1.105, -0.048)

Median Debt by Research Type:

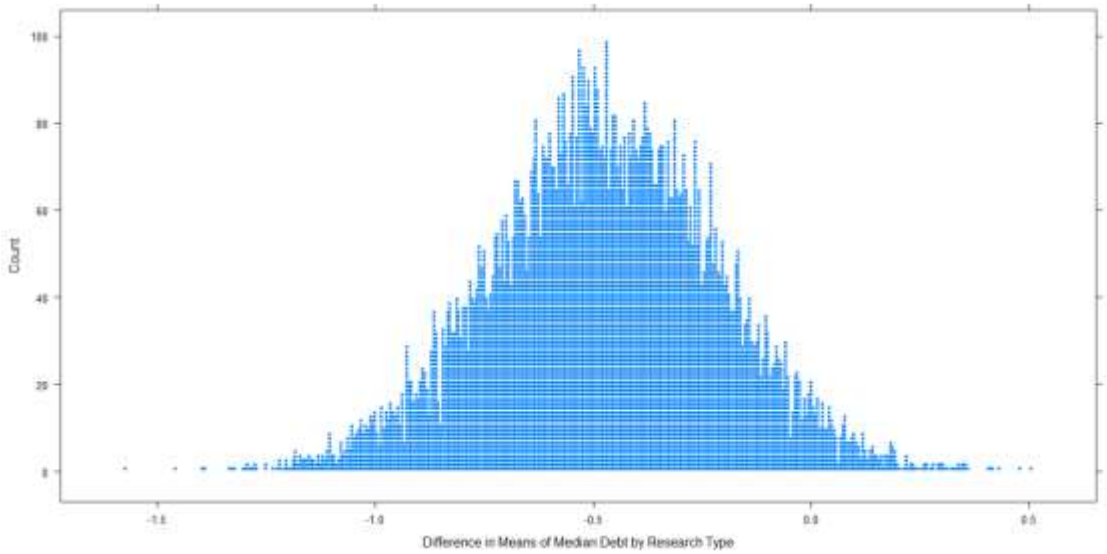

Figure 55. Dot plot showing the bootstrap distribution of difference in means between the mean median debt of student at graduation for research institutions and the mean median debt of student at graduation for non-research institutions.

Confidence Interval:

95% CI for difference in means: (-0.998, 0.035)

Commentary:

All data is normally distributed under bootstrap samples, which validates the legitimacy of the sample data and further allows us to analyze and extrapolate information.

SECTION IV:

<u>Graduation Rate by Proportion of First Generation Students</u>:

Summary Statistics											
n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
400	0.00	39.08	20.18	0.43	-0.58	0.00	22.63	36.31	53.11	90.20	30.48

Table 15. Table displaying summary statistics including the five-number summary of graduation rate at each institution.

Summary Statistics											
n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
400	0.00	33.05	10.28	-0.10	-0.62	9.03	25.85	33.67	40.88	61.27	15.03

Table 16. Table displaying summary statistics including the five-number summary of proportion of first generation students at each institution.

Test of Correlation:

> ***H_0*** **:** R = 0; There is not an association between graduation rate and proportion of first generation students.
>
> ***H_a*** **:** R ≠ 0 ; There is an association between graduation rate and proportion of first generation students.
>
> **Test:**
>
> **Confidence Interval:** We are 95% confident that R lies between -0.762 and -0.666.
>
> **T-value:** -20.532
>
> **P-value:** <2.2e-16
>
> **Level of Significance (α):** 0.05
>
> **Conclusion:**
>
> P-value < α, therefore, we reject the null hypothesis.
>
> We have enough evidence to determine that there is an association between graduation rate and proportion of first generation students.

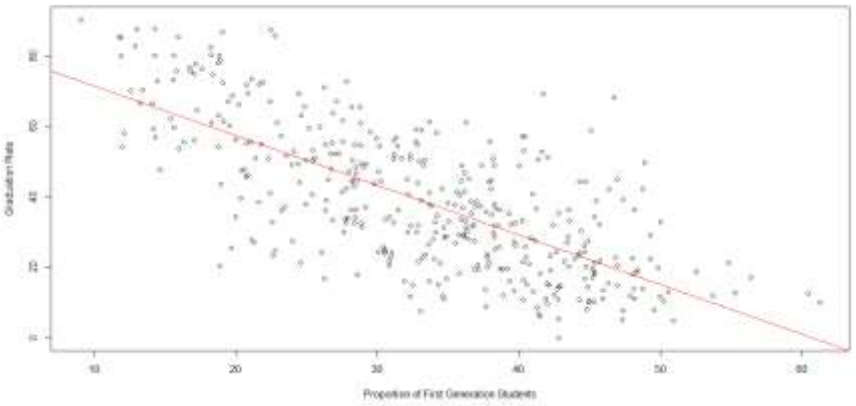

Figure 56. Plot showing correlation and line of regression for graduation rate by proportion of first generation students.

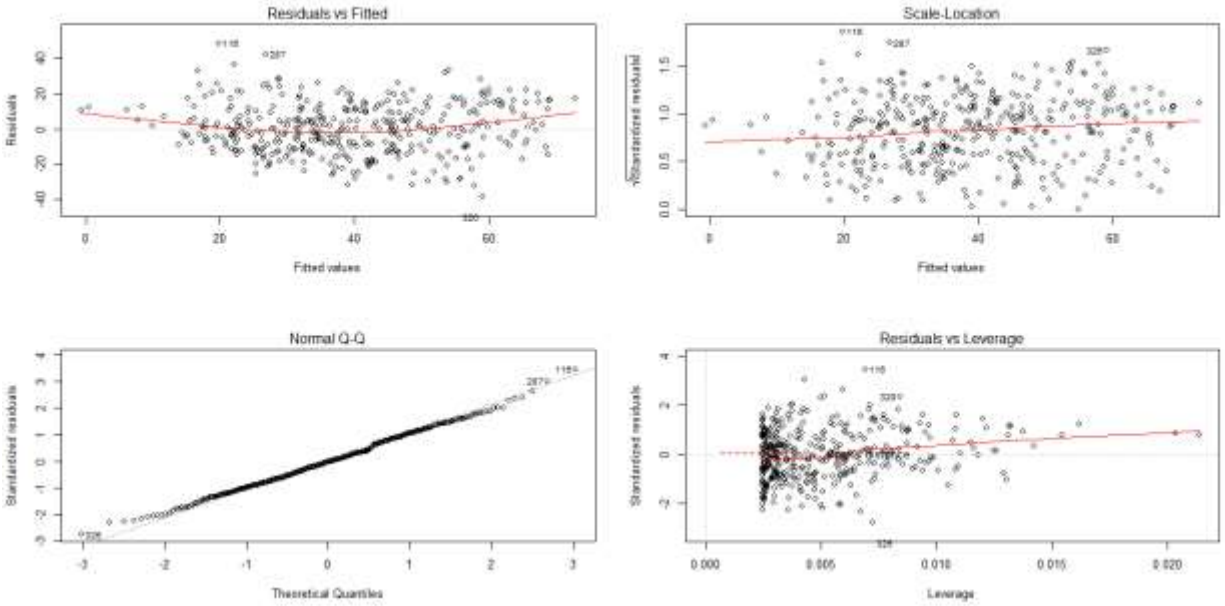

Figure 57. Representation of residual plots for grad rate by first generation students.

ANOVA (Grad Rate by First Gen)					
	Df	Sum Sq	Mean Sq	F-value	P-value
FirstGen	1	83581	83581	421.56	< 2.2e-16
Residuals	398	78909	198		

Table 17. Table displaying ANOVA regression test for grad rate by first generation students.

Least Square Regression Model:

R = -0.72

$$\widehat{GradRate} = 85.62 - 1.41(Prop\ First\ Gen)$$

Test of Least Square Regression Model:

Hypothesis: The regression line fits the model.

Intercept: 85.62

T-value: 36.07

P-value: <2e-16

Slope: -1.41($Prop\ First\ Gen$)

T-value: -20.53

P-value: <2e-16

Level of Significance (α): 0.05

Conclusion:

P-value < α. Therefore, the tested values are significant.

We have enough evidence to conclude that the model fits the regression line.

Commentary:

Regression analysis and statistical test confirm our assumptions about graduation rate by proportion of first generation students enrolled. There is a strong, negative association between our explanatory variable (proportion of first generation students) and our response variable (graduation rate). That is, as the proportion of first generation students enrolled at an institution increases by 1, the graduation rate decreases at a rate of -1.41.

Graduation Rate by Median Family Income:

Summary Statistics											
n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
400	0.00	39.08	20.18	0.43	-0.58	0.00	22.63	36.31	53.11	90.20	30.48

Table 18. Table displaying summary statistics including the five-number summary of graduation rate at each institution.

Summary Statistics											
n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
400	0.00	50.89	20.37	0.67	0.25	13.71	36.63	47.91	63.52	114.94	26.89

Table 19. Table displaying summary statistics including the five-number summary of median family income of all students enrolled (per 1K) at each institution.

Test of Correlation:

H_0 **:** R = 0; There is not an association between graduation rate and median family income.

H_a **:** R ≠ 0 ; There is an association between graduation rate and median family income.

Test:

Confidence Interval: We are 95% confident that R lies between 0.634 and 0.737.

T-value: 18.957

P-value: <2.2e-16

Level of Significance (α): 0.05

Conclusion:

P-value < α, therefore, we reject the null hypothesis.

We have enough evidence to determine that there is an association between graduation rate and median family income.

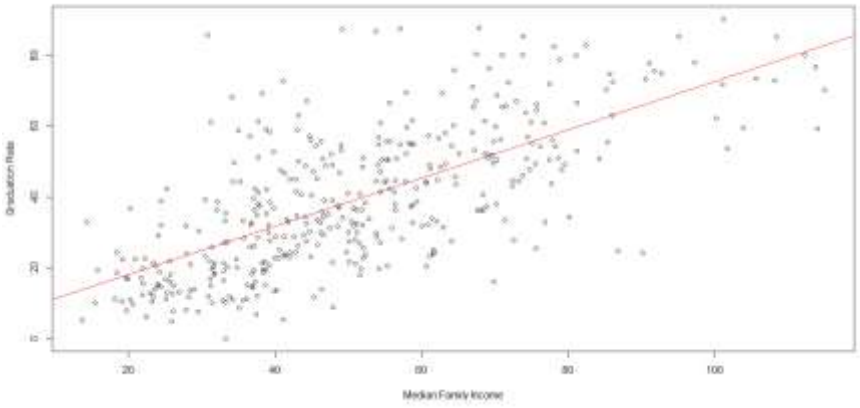

Figure 58. Plot showing correlation and line of regression for graduation rate by median family income of all students.

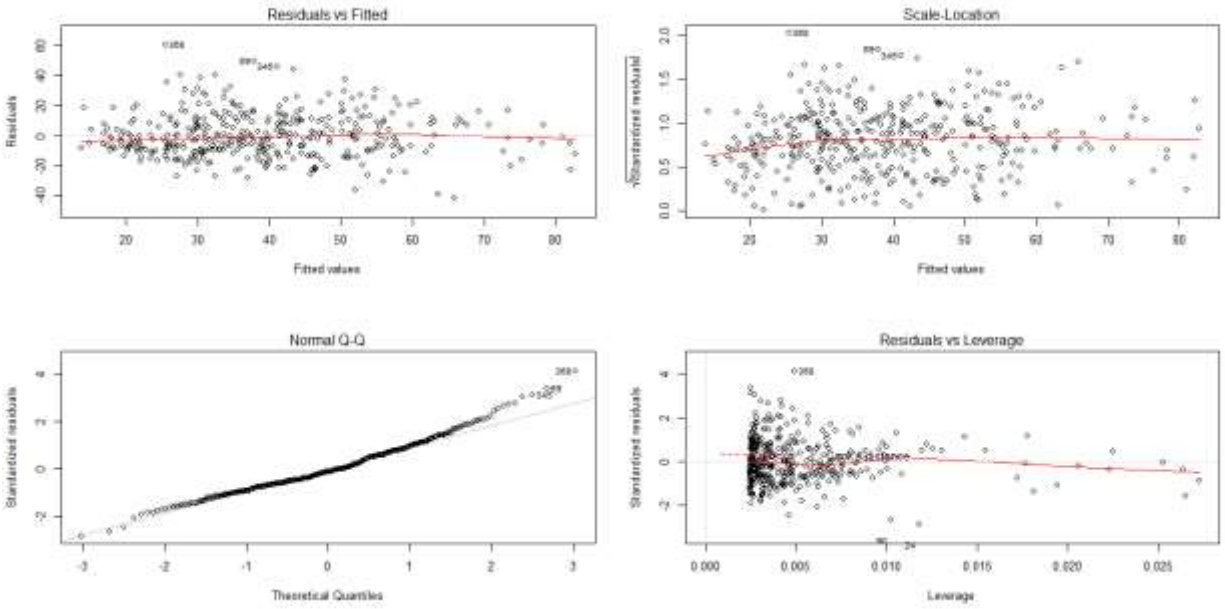

Figure 59. Representation of residual plots for grad rate by median family income.

ANOVA (Grad Rate by Family Income)					
	Df	SumSq	MeanSq	F-value	P-value
Family Income	1	77101	77101	359.37	< 2.2e-16
Residuals	398	85389	215		

Table 20. Table displaying ANOVA regression test for grad rate by median family income.

Least Square Regression Model:

R = 0.69

$$\widehat{GradRate} = 4.35 + 0.68(Median\ Family\ Income)$$

Test of Least Square Regression Model:

Hypothesis: The regression line fits the model.

Intercept: 4.35

T-value: 2.21

P-value: 0.028

Slope: 0.68 ($Median\ Family\ Income$)

T-value: 18.957

P-value: <2e-16

Level of Significance (α): 0.05

Conclusion:

P-value < α. Therefore, the tested values are significant.

We have enough evidence to conclude that the model fits the regression line.

Commentary:

Regression analysis and statistical test confirm our assumptions about graduation rate by median family income. There is a strong, positive association between our explanatory variable (median family income) and our response variable (graduation rate). That is, as the median family income increases by 1K, the graduation rate increases at a rate of 0.68.

Graduation Rate by Age of Entry:

Summary Statistics											
n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
400	0.00	39.08	20.18	0.43	-0.58	0.00	22.63	36.31	53.11	90.20	30.48

Table 21. Table displaying summary statistics including the five-number summary of graduation rate at each institution.

Summary Statistics											
n	miss	mean	sd	skew	krts	min	qrt1	mdn	qrt3	max	IQR
400	0.00	22.51	2.67	1.60	3.26	19.30	20.60	21.84	23.76	34.03	3.16

Table 22. Table displaying the summary statistics including the five-number summary of the age of entry for students for each institution.

Test for Correlation:

H_0 **:** R = 0; There is not an association between graduation rate and age of entry.

H_a **:** R ≠ 0 ; There is an association between graduation rate and age of entry.

Test:

Confidence Interval: We are 95% confident that R lies between -0.528 to -0.372.

T-value: -10.161

P-value: <2.2e-16

Level of Significance (α): 0.05

Conclusion:

P-value < α, therefore, we reject the null hypothesis.

We have enough evidence to determine that there is an association between graduation rate and age of entry.

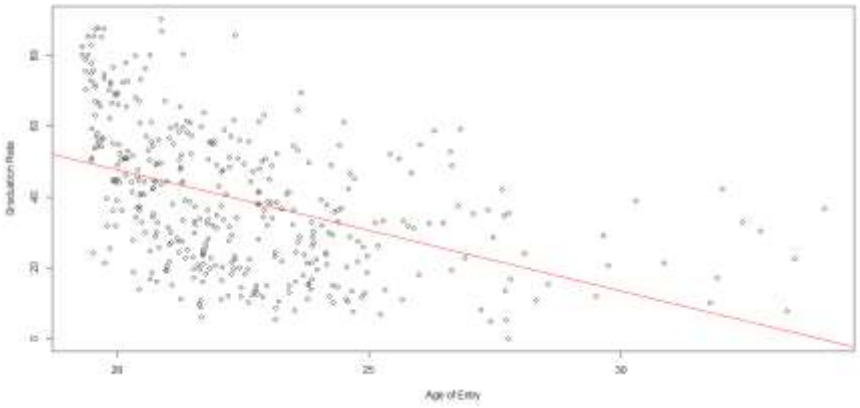

Figure 60. Plot showing correlation and line of regression for graduation rate by age of entry for all students.

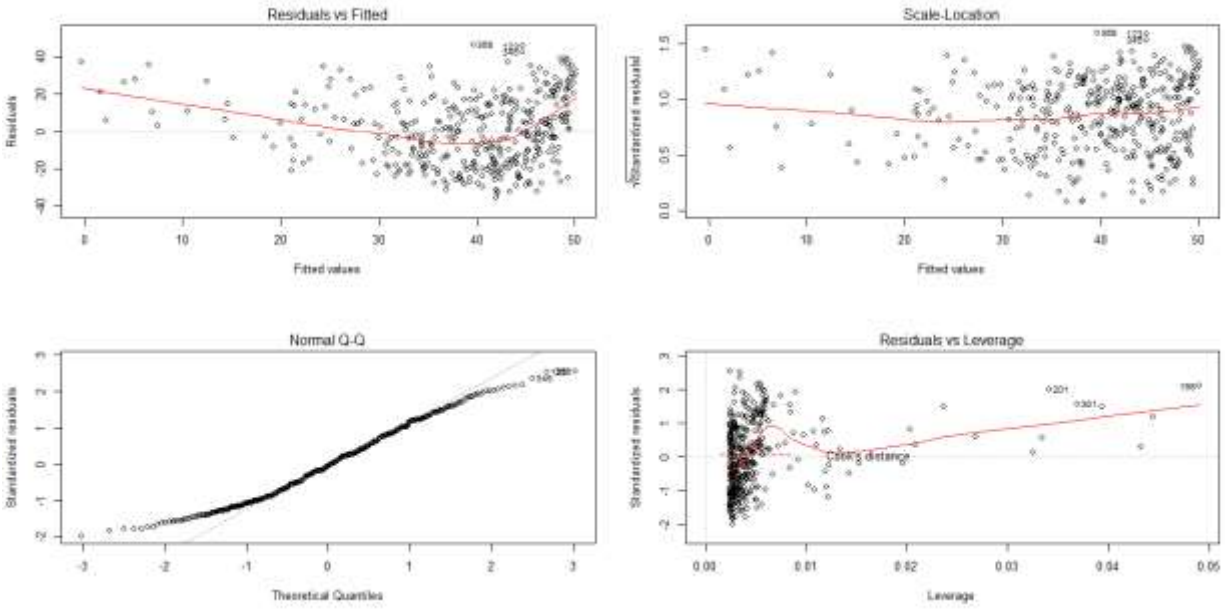

Figure 61. Representation of residual plots for grad rate by age of entry.

ANOVA (Grad Rate by Age of Entry)					
	Df	Sum Sq	Mean Sq	F-value	P-value
Age of Entry	1	33468	33468	103.24	< 2.2e-16
Residuals	398	129022	324		

Table 23. Table displaying ANOVA regression test for grad rate by age of entry.

Least Square Regression Model:

R = -0.45

$$\widehat{GradRate} = 116.18 - 3.42(Age\ of\ Entry)$$

Test of Least Square Regression Model:

Hypothesis: The regression line fits the model.

Intercept: 116.18

T-value: 15.20

P-value: <2e-16

Slope: -3.42$(Age\ of\ Entry)$

T-value: -10.16

P-value: <2e-16

Level of Significance (α): 0.05

Conclusion:

P-value < α. Therefore, the tested values are significant.

We have enough evidence to conclude that the model fits the regression line.

Commentary:

Regression analysis and statistical test confirm our assumptions about graduation rate by average age of entry for students. There is a moderate, negative association between our explanatory variable (average age of entry) and our response variable (graduation rate). That is, as the proportion of first generation students enrolled at an institution increases by 1 year, the graduation rate decreases at a rate of -3.42.

CONCLUSION:

Now that graphic summaries, statistic summaries and tests have been run, what practical information can be gleaned from what we have observed and who can use it, and for what purpose? We began this observational study with the intent of further understanding how the proportion of first generation students enrolled at a four-year, bachelor's degree granting, post-secondary institution interacts with other variables in the data set. So, did we succeed?

The mean proportion of first generation students enrolled at institutions in our study is 33.05%. Since the data is normally distributed under bootstrapping, we can establish the confidence interval for where the true proportion of first generation students fall at all similar institutions in the US. That is, we are 95% confident that mean proportion of first generation students at all institutions is between 32.35% and 34.30%. We then see that between non-research and research universities the mean proportions are 33.39% and 31.81% respectively. Under bootstrapping conditions the difference in those means is established: we are 95% confident that the difference between proportion of first generation students enrolled at non-research and research institutions is between -3.755 and 1.234 percent. Zero, lies in the range, so we can implicitly conclude that there is no statistical difference between the two means. This is confirmed using the classical approach with the t-test, where we see that we fail to reject the null hypothesis that the two means are equal. Therefore, there is not any statistical difference between the two means. So, even though there sometimes appears to be more first generation students enrolled at non-research institutions since the range is wider (see First generation student by research type in section II), we now know with more certainty that the mean proportion of first generation students at non-research vs. research institutions is not significantly different.

After establishing an interval for the true proportion of first generation students enrolled and testing the usefulness and significance of the data, we then compare it to other variables in the dataset. One large limitation of the dataset that we were aware of at the start was that we only have one dependent variable: graduation rate. While other info would have been of great interest, we worked within the parameters of the data and discovered that there is a strong association between the proportion of first generation students enrolled at an institution and graduation rate. Unfortunately, it is not data that inspires a lot of hope in me. We see that there is a strong negative association between the two. That is, as the proportion of first generation students enrolled at an institution increases the graduation rate within four years decreases. Using that info, we created a predictive model that can determine the graduation rate of first generation students for all similar institutions:

$$\widehat{GradRate} = 85.62 - 1.41(Prop\ First\ Gen)$$

Tests of the model show that the values are significant and useful within the scope of our study.

We know that the data is significant and useful, but what is its practical application and who can use it? The answer truly is anyone, but ideally those interested might be administrators of four-year, bachelor degree granting institutions since this is the population our data generalizes to. It is data that one might use to decide what programs and incentives that could be implemented at an institution to encourage the success of first generation students, or those who come from families with lower incomes or who enter college at an older age.

While median family income and age of entry were not our primary interest, we do observe similarly interesting info to that of proportion of first generation students enrolled. Since we were mostly interested in first generation students, I have outlined the data analysis process in this conclusion. However, the amazing thing about statistical analysis is that it is in fact a rinse and repeat process. Only the variables change and therefore the outcome regarding those variables. With median family income, we ultimately see that there is a positive association between income and graduation rate—i.e. the higher the median family income of students at an institution, the higher the graduation rate. And with average age of entry we see a moderately negative association with graduation rate, being the higher the average age of entry at an institution the lower the graduation rate becomes.

An administrative official can take this data into account and decide if and how they might want to address population of students that fall under these three variables. I would like to add, though, for me there is a disconcerting side to this data. All we can do is observe and analyze the information. How and if any issues that this data implicates are addressed are not up to us. This means that someone can interpret this info in many ways. For instance, one may look at the data regarding the proportion of first generation students enrolled at an institution and decide that programs should be implemented to ensure the success of those students, while another can look at that data and decide that perhaps the institution should admit less first generation students to improve the graduation rate statistics. For us, it's about the analysis of the data and not necessarily the decisions made from the data. Yet, I do think it's important to keep in mind the importance of thorough and responsible analysis of the data for this sake.

So, again I ask. Did we succeed in our goal of further understanding how the proportion of first generation students enrolled at a four-year, bachelor's degree granting, post-secondary institution interacts with other variables in the data set? We now have a more thorough understanding of the data in its context, but I would also emphasize that for further study it's just a starting point. For more information, longitudinal studies might be of interest or even more focused studies with specific institutions in mind. However, regarding our observational study, my short answer: yes.

Bibliography

Accreditation: Agency List. (n.d.). Retrieved May 06, 2017, from https://ope.ed.gov/accreditation/agencies.aspx

Carnegie Classification of Institutions of Higher Education ®. (n.d.). Retrieved May 06, 2017, from http://carnegieclassifications.iu.edu/classification_descriptions/basic.php

Data Documentation for College Scorecard. (n.d.). Retrieved May 6, 2017, from https://collegescorecard.ed.gov/assets/FullDataDocumentation.pdf

Fast Facts: Educational Institutions. (n.d.). Retrieved May 06, 2017, from https://nces.ed.gov/fastfacts/display.asp?id=84