Anish Gupta

anish.gupta@rutgers.edu

12/16/2023

1. Introduction/Problem Statement

The goal of my research is to try to determine the following - What are the most important predictors for 4-year and 6-year graduation rates in higher education institutions for fulltime degree seeking undergraduate students? My key variables of interest based on my college experience, literature review, conversations with other students are - percentage of undergraduates receiving a Pell Grant, whether a college is public or private, median SAT value, student retention rate, percentage of full-time students, completions per 100 full-time undergraduate students - national sector average, gender, and ethnicity.

2. Theoretical Framework/Literature Review

Why is this a topic of interest? Why is it important to have a college degree?

This topic is of great interest because individuals who obtain a college degree can achieve middle-class status. The college credential is needed to boost economic growth and competitiveness in our country. Currently, there is a growing concern that the United States is steadily losing their edge in global competitiveness when it comes to college completion. According to the U.S. Department of Labor, greater educational attainment leads to both higher salaries and lower unemployment rates. As per the U.S. Census Bureau, in 2009, only 27.4% of the adult population in the United States completed a college degree. However, there is a wide gap in degree attainment by race/ethnicity and gender.

In 1990, the United States Congress began to acknowledge that completing a college degree was very important in the country. As a result, the Student Right-to-Know and Campus Security Act was

passed. This meant that institutions were required to report their retention and graduation rates publicly. Four-year institutions need to accurately assess how they should design paths for students that are enrolled toward graduation.

In today's rapidly evolving job market, the value of a bachelor's degree is a topic of intense debate. According to the article “10 Benefits of Having a College Degree” from Northeastern University Bachelor’s Degree Completion, students are able to stand out from others in a competitive job market on obtaining a bachelor’s degree.

The economic argument for a bachelor's degree is an important factor to consider. With a bachelor’s degree, adults can earn 66% more income compared to those with a high school diploma. In addition, unemployment rates are significantly lower for degree holders, and they tend to enjoy higher job satisfaction. Return on investment raises the concern whether a college degree is worth it due the rising cost of tuition.

Moving beyond the financial aspect, college graduates develop a broad skillset and plenty more opportunities open for them in healthcare, engineering, or technology. Employers seek college graduates that have critical thinking, communication, and problem-solving skills.

Research in Measuring College Graduation Rates

The article “Completing College: Assessing Graduation Rates at Four-Year Institutions” highlighted several important points in their research. The main goal of the research was to evaluate the national average and institutional variation in graduation rates for first-time, full-time students who began studying in 2003. This research also focused on identifying and analyzing disparities in graduation rates based on factors such as race, ethnicity, socioeconomic background, and transfer student status. Based on the analysis, graduation rates explained the complex factors contributing to the observed variations. The results of this research also offered recommendations for

policymakers, institutions, and educators to address the issue of low graduation rates and promote academic success for all students.

The authors collected data from the 2004 CIRP Freshman Survey (TFS) and the National Student Clearinghouse (NCS). The two datasets were merged, and this permitted the authors to examine retention and degree attainment of 210,056 first-time, full-time, students at 356 four-year non-profit colleges. The multiple imputation method was used to handle missing values. Logistic regression was the main analytical method used to create a model and predict degree completion. The dependent variable was dichotomous, and it was coded as 1 for students that graduated within four, five, or six years and 0 if they did not graduate. Private universities had the highest four-year degree completion rate at 64%, while the public four-year colleges had the lowest graduation rate at 23.5%.

In another research on “Prediction of Graduation with Naïve Bayes Algorithm and Principal Component Analysis (PCA) on Time Series Data”, the authors aim to predict which students graduate on time by using data mining methods. One of the methods that was used in this research was Principal Component Analysis which has a straightforward structure and can simplify complex academic data. The results had been processed using Naïve Bayes classification. Beyond just course grades, numerous variables impact whether a student graduates. By leveraging data mining techniques on students' academic data, graduation rates can be predicted and students at risk can be proactively prevented from falling behind. Seeking to enhance both efficiency and accuracy, this research explored combining Naïve Bayes with PCA. The main takeaway of the result of this research is how to predict graduation accurately. The conclusion was that the combined algorithm improved the accuracy. The PCA-Naïve Bayes combination algorithm gave 6.04% better accuracy compared to an original Naive Bayes classification.

What is PCA?

According to the article “Machine Learning Applications in Graduation Prediction at the University of Nevada, Las Vegas”, the author talks about Principal Component Analysis (PCA) and it is a method used to reduce the number of attributes. PCA is a technique where dimensionality reduction can convert the data and make it easier to comprehend. Training models can be very helpful and make it more efficient by using many attributes. Data can be transformed into two or three-dimensional space by visualizing them into traditional plots. Clusters, outliers, and the distribution of the data can be detected. This method is applied once the data has been altered, scaled, and imputed.

Some concerns about using Predictive Analytics to boost college graduation rates

The article "Predictive Analytics Are Boosting College Graduation Rates, But Also Invade Privacy and Reinforce Racial Inequities” explores complex intersection of predictive analytics in higher education. While these algorithms hold promising results for boosting graduation rates, there is also the concern of students dropping out of college, privacy concern, bias, and potential reinforcement of existing inequities. The authors mention that 1,400 colleges and universities are turning to predictive analytics to help them find trends and patterns with vast amounts of historical data and using those patterns to forecast the future. Many of these institutions are utilizing data analytics to make sure to keep students enrolled and on path towards graduation. Some colleges have reported success, like Georgia State University saw an increase in graduation after implementing predictive analytics. Students on college campuses are not aware of being monitored. People argue that student data raises privacy concern. The potential for surveillance and profiling students based on their predicted behavior could erode trust and autonomy.

Another point that this article talks about is bias and reinforcing inequalities. Algorithms are only as good as the data they are fed. If historical data contains biases, this means the algorithms may perpetuate those biases. If minority students have lower graduation rates, an algorithm based on such data could wrongly flag them as high-risk, leading to negative consequences.

There should be careful consideration of ethical implications before fully implementing predictive analytics in higher education. Being transparent in data use should be an absolute priority.

3. Data and Measures

The dataset for this project contains information on college completion data from 3,800 degree-granting institutions in the United States. The dataset includes four csv files, and they are: cc_institution_details, cc_institution_grads, cc_state_sector_details, and cc_state_sector_grads. In my analysis I have used the first 2 csv files. The college completion datasets were pulled from the College Completion microsite and produced by The Chronicle of Higher Education with support from the Bill and Melinda Gates Foundation. This data was published in data.world by Jonathan Ortiz.

The data examines data and trends at 3,800 degree-granting institutions in the United States (excluding territories) that reported a first-time, full-time degree-seeking undergraduate cohort, had a total of at least 100 students at the undergraduate level in 2013, and awarded undergraduate degrees between 2008 and 2013.

The cc_institution_details table contains 2339 rows and 63 columns for 4-year colleges.

For this table the following data wrangling was performed:

1. I filtered for 4-year institutions.
2. Created a new binary variable from the variable “control” called “control_public_private” to indicate if an institution is public or private (1 = Public and 2 = Private for-profit and Private not-for-profit)
3. Created a subset of the table and dropped unnecessary columns
4. The dataset had the missing values listed in Table 1 for my variables of interest.

   As grad_100_value and grad_150_value are predictors, the NA value rows were dropped from the analysis. About 43% of the rows have missing values for the med_sat_value column. The

column endow value has about 28% of the rows with missing values. Both these columns were dropped from the analysis. For retain_percentile and ft_pct, I used listwise deletion.

The cc_institution_grads table has 1302102 rows and 10 columns

For this table the following data wrangling was performed:

1. Filtered for 4-year institutions and removed rows for all races and all gender (total)
2. For institutions that reported data for more than 1 year, filtered out the information for the most recent year so we eventually have 1 row per school.
3. I created a subset of the table and dropped unnecessary columns.
4. I created a new column “grad_cohort_percent” as a percent of the race gender combinations were needed for the analysis and not the actual numbers.

Next the two table were joined, and more data wrangling was performed:

1. The 2 datasets cc_institution_details and cc_institution_grads were left joined on the school id
2. Performed a pivot wider of the merged dataset so that each gender-race combination is a separate column and there is one row per institution in the final table.

The merger table has 1927 rows and 21 columns.

Following are the variables used in the merged dataframe for the analysis:

Dependent variables

- grad_100_value - percentage of first-time, full-time, degree-seeking undergraduates who complete a degree or certificate program within 100 percent of expected time (bachelor's-seeking group at 4-year institutions)

- grad_150_value - percentage of first-time, full-time, degree-seeking undergraduates who complete a degree or certificate program within 150 percent of expected time (bachelor's-seeking group at 4-year institutions)

Independent variables and reasoning for inclusion in analysis

- control_public_private - Control of institution (Public, Private not-for-profit, Private for-profit)
  There is a significant difference between public and private institutions in terms of graduation rates especially for 4-year graduation rates. Please refer to Tables 3 and 4 in the tables section.
- aid_value - The average amount of student aid going to undergraduate recipients.
  aid_value is correlated with both 4-year and 6-year graduation rates as shown in Graphs 1 and 2
- retain_percentile - Institution's percent rank for freshman retention percentage within sector
  retain_percentile - is correlated with both 4-year and 6-year graduation rates as shown in Graphs 3 and 4
- ft_pct - Percentage of undergraduates who attend full-time
  fct_pct has correlation with both 4-year and 6-year graduation rates as shown in Graph 5 and 6
- fte_percentile – Percentile of full-time equivalent undergraduates
  fte_percentile has a correlation with both 4-year and 6-year graduation rates as shown in Graph 7 and 8
- M_W - % of Male White Students
- F_W - % of Female White Students
- M_B - % of Male Black Students
- F_B - % of Female Black Students
- M_H - % of Male Hispanic Students
- F_H - % of Female Hispanic Students

- M_Ai - % of Male American Indian Students
- F_Ai - % of Female American Indian Students
- M_A - % of Male Asian Students
- F_A - % of Female Asian Students

  All above gender race combinations are initially included in the model to study the impact of gender and race on graduation rates.

Summary Table for descriptive statistics has been provided in Table 2 under Tables section.

4. Methods

The two primary statistical procedures used were Multiple Linear Regression and Principal Components Analysis. The rationale behind using Multiple Linear Regression - the goal was to build a model with good predictors to predict the 4-year and 6-year graduation rates individually with multiple independent variables. The rationale behind using Principal Components Analysis – is that there was a good amount of correlation between the independent variables as shown in Graph 9.

Multiple Linear Regression

I created five models. The first 3 models namely model_1, model_1.1, model_2 have been used to predict the 4-year graduation. The first model predicted the 4 year graduation rate with all variables included except for 1 which was F_A (female Asian). Due to the inclusion of this variable in the model, the vif of the model was generating an error - Error in vif.default(model_1) : there are aliased coefficients in the model. Therefore, this variable was dropped from the analysis at this stage.

For Multiple Linear Regression the following steps were performed, and the linear regression assumptions were handled –

1. The first assumption was that the variables are measured correctly. The readme file of the data documentation does not mention about any variables being incorrectly measured. Therefore, it can be assumed that the variables are measured correctly.
2. Next the first models for 4-year and 6-year graduation rates had all variables. However, the subsequent models had all the variables specified correctly. There were no missing variables and so there were no omitted variable bias. The dependent and independent variables were not reversed. There were also no extra variables in the model.
3. There were non-linear relationships in this dataset. I generated histograms of all the variables and found that there was quiet a few variables with a right skew. The variable aid_value was right skewed and not zero inflated. A log transformation helped the variable approach normality. The other right skewed independent variables F_A, F_Ai, F_B, M_A, M_Ai, M_B, F_H, M_H were zero inflated distributions, and a log transformation would not help here. The dependent grad_100_value was right skewed. However, a log transformation was making it left skewed. Therefore, it was not transformed.
4. The residuals of model 1 had a mean of zero.
5. The residuals suffered from heteroskedasticity. Upon plotting the residuals against the independent variables, it was found that the residuals had a cone shaped relationship with M_B, F_B, M_H, F_H, M_Ai, F_Ai, M_A (Graphs 10-16).
6. This model also suffered from multicollinearity as shown by the Variance Inflation Factor on model_1 and are shown in table variables and it can be see that the variables M_W, F_W, M_B, F_B, F_H, and M_A have VIFs > 4.
7. In the second model, model_1.1, the issues with nonlinear relationships had been addressed by using a quadratic model. However, no changes were observed.
8. The third model, model_2 addressed the multicollinearity as well as heteroskedasticity by the by exclusion of the variables ln_aid_value, awards_per_natl_value, F_B, M_B, M_Ai, F_Ai,

M_A, and F_H. While the multicollinearity was resolved (Table 6), the heteroskedasticity was not.

9. The fourth and fifth models predicted the 6-year graduation rate. The fourth model is the same as model 1 with all the assumption violations of model 1.
10. The fifth model is the same as model 3 except that it predicts the 6-year graduation rate where the issue of multicollinearity was addressed but heteroskedasticity was not fixed.

Principal Components Analysis

The dataset used for this analysis was fairly complex. It had lot of NA values that had to be handled. It contained a lot of variables and could use some help with compressing them into a simpler form and reducing them into something more meaningful. In addition, this dataset had high levels of multicollinearity. Therefore, was a good candidate for PCA. There were many variables with a correlation greater than 0.3 in this dataset.

The following steps were taken for performing the PCA:

1. Most of the variables of interest were passed to the PCA to see if it could generate meaningful PCs.
2. Based on the eigenvalues, combined % variance of the original variables, and the scree plot, the PCs were selected to perform regression analysis.
3. Regression analysis was performed both for the 4-year and 6-year graduation rate prediction.

5. Results

Results of Multiple Linear Regression

model_1 – Predict 4-year graduation rate

The linear regression model_1 included all variables except for F_A as explained in the methods section. The summary of the model is presented in Tables 7 and 8, the histogram of residuals is

presented in Graph 17 and the scatterplot of residual vs predicted values is presented in Graph 18. Looking at table 17 we see that all the independent variables except for fte_percentile are significant. The R2 / R2 adjusted are 0.644 / 0.641. The F-test is significant with a p-value: < 0.00000000000000022 shown in Table 8. The histogram of residuals shows that the residuals are normally distributed with mean of 0. The scatterplot of residuals has a cone formation and is not homoscedastic. The scatterplot of residuals has a cone formation with many of the independent variable M_B, F_B, M_H, F_H, M_Ai, F_Ai, M_A (Graphs 10-16)

model_1.1 – Predict 4-year graduation rate

As the independent variable display a correlation with the dependent variable, but not a linear relationship, a quadratic model was created as the second model. The summary of the model are presented in Tables 9 and 10, the histogram of residuals is presented in Graph 19 and the scatterplot of residual vs predicted values is presented in Graph 20. The results are exactly same as model 1 with the same R2 and adjusted R2, significance of variables and the F-test. Therefore, the quadratic model did not improve things.

model_2 - Predict 4-year graduation rate

Next, I ran a new model to address multicollinearity and heteroskedasticity. Here I dropped the variables with high VIF and cone shaped residual plots. This model was built by dropping ln_aid_value, awards_per_natl_value, F_B, M_B, M_Ai, F_Ai, M_A,F_H due to multicolinearity. The summary of the model is presented in Tables 11 and 12, the histogram of residuals is presented in graph 21 and the scatterplot of residual vs predicted values is presented in Graph 22. Looking at Table 11 we see that all the independent variables except for fte_percentile are significant. The R2 / R2 adjusted are 0.558 / 0.557. The F-test is significant with a p-value: < 0.00000000000000022 shown in Table 12. The histogram of residuals shows that the residuals are normally distributed with mean of 0. The scatterplot of residuals has a cone formation and is not homoscedastic. Although the R2/ Adjusted

R2 have fallen, the model no longer displays severe multicollinearity in comparison to model_1 as shown in table 6.

model_3 – Predict 6-year graduation rate

This model follows everything in model_1 except the dependent variable is grad_150_value. The model is very similar to model 1. The summary of the model is presented in Tables 13 and 14, the histogram of residuals is presented in Graph 23 and the scatterplot of residual vs predicted values is presented in Graph 24. Looking at Table 13 we see that all the independent variables are significant. The R2 / R2 adjusted are 0.674 / 0.671. The F-test is significant with a p-value: < 0.00000000000000022 shown in Table 14. The histogram of residuals shows that the residuals are normally distributed with mean of 0. The residual plot shows heteroskedasticity. Table 15 shows values for the VIF of the model and displays severe multicollinearity.

model_4 - Predict 6-year graduation rate

In model 4, I dropped the variables ln_aid_value, awards_per_natl_value, F_B, M_B, M_Ai, F_Ai, M_A,F_H due to multicollinearity. The summary of the model is presented in Tables 16 and 17, the histogram of residuals is presented in Graph 25 and the scatterplot of residual vs predicted values is presented in Graph 26. Looking at table 16 we see that all the independent variables except for control_public_private are significant. The R2 / R2 adjusted are 0.590 / 0.588. The F-test is significant with a p-value: < 0.00000000000000022 shown in Table 17. The histogram of residuals shows that the residuals are normally distributed with mean of 0. The scatterplot of residuals has a cone formation and is not homoscedastic. Although the R2/ Adjusted R2 have fallen, the model no longer displays severe multicollinearity in comparison to model_3 as shown in Table 18.

Overall, the linear regression models improved in terms of multicollinearity and the R2 and adjusted R2 are not very low and the models 2 and 4 are significant with most of the predictors significant at $\alpha = 0.00$. However, PCA might be a better fit.

Results of Principal Components Analysis

Here, only PC1, PC2, PC3, PC4, PC5, PC6, and PC7 have eigenvalues > 1. The first PC accounts for 20%, the second accounts for 14%, the third accounts for 11%, the fourth accounts for 8%, the fifth accounts for 8%, the sixth PC accounts for about 8%, and the seventh PC accounts for about 7% of the variance of the original variables. Among the PCs, PC 1, 2, 3, 4, 5, 6, and 7 together account for 75.64% of the variance of the original variables. Please refer to Table 19.

On analyzing the loading scores, I saw that for PC1 the pell_percentile and the F_B (female Black) have the strongest effect on PC1. The M_H (male Hispanic), F_H (female Hispanic), M_A (male Asian), and F_A (female Asian) have the strongest effect on PC2 as shown in Table 20.

From the scree plot on Graph 27, we can see that there is a break after PC4 but there is also a change after PC7. The scree plot bar graph shows the same pattern on Graph 28.

The biplot is shown on Graph 29.

I ran a linear regression model using the first 7 PCs to predict the 4-year graduation. This decision was based on the fact that they together account for 75.64% of the variance of the original variables and had eigenvalues > 1 and also the scree plot showed a break at the PC7 point. So, I decided to go with 7 PCs.

In Tables 21 and 22 show that PC1, 5, and 6 have a positive coefficient and therefore the variables in these PCs have an overall positive relationship with the 4-year graduation %. PC2, 3, 4, and 7 have a negative coefficient and therefore the variables in these PCs have an overall negative relationship with the 4-year graduation %. PC1, 2, 3,4, and 5 are statistically significant at $\alpha = 0.001$. PC6 is significant at 0.05 and PC7 is not significant. In this model, the R2 / R2 adjusted are 0.610 / 0.608. The F-test was significant at $\alpha = 0.0001$ with p-value: < 0.00000000000000022. With this R^2 61% of the variation in the dependent variable is explained by the model. With Adjusted R-squared 60% of the

variation in the dependent variable is explained by the model. Both R-squared and adjusted R-squared are good and the F-test is also significant, and this suggests that the model is a good fit. Graphs 29 and 30 show the residual distribution and scatterplot. The histogram is normally distributed with mean 0. However, the residual plot is heteroskedastic.

I ran a linear regression model using the first 7 PCs to predict the 6-year graduation and the justification for that was the same as was for the 4-year graduation prediction.

Tables 23 and 24 show that PC1, 4, 5, and 6 have a positive coefficient and therefore the variables in these PCs have an overall positive relationship with the 6-year graduation %. PC2, 3, and 7 have a negative coefficient and therefore the variables in these PCs have an overall negative relationship with the 6-year graduation %. All PC1, 2, 3, 6, and 7 are statistically significant at $\alpha = 0.0001$. PC4 and 5 are significant at 0.05. In this model, the R2 / R2 adjusted 0.616 / 0.615. The F-test significant at $\alpha = 0.0001$ with p-value: < 0.00000000000000022. With this R^2 61% of the variation in the dependent variable is explained by the model. With Adjusted R-squared 62% of the variation in the dependent variable is explained by the model. Both R-squared and adjusted R-squared are good and the F-test is also significant, and this suggests that the model is a good fit. Graphs 31 and 32 show the residual distribution and scatterplot. The histogram is normally distributed with mean 0. However, the residual plot is heteroskedastic.

Looking at tables 25 and 26 we can say that PCA is slightly better than Linear Regression Model for this analysis both in terms of R2 and significance level.

6. Discussion, Policy Implications, and Conclusion

My findings here suggest that, based on the PCA and also Linear Regression Models, the most important predictors that impact the 4-year and 6-year college graduation rates are - institution's percent rank for freshman retention percentage within sector, Pell percentile rank for an undergraduate

institution, if an institution is public vs private, the % of black female population , % of female white population, % of male Hispanic, % of female Hispanic, % male Asian, % female Asian, the average amount of student aid going to undergraduate recipients, total % of full time undergraduates.

One limitation of this research was that the residual plot was heteroskedastic. Further investigation is needed on this issue to get models with a better fit. The other limitation of this research was that too many important indicators for graduation rate prediction like median SAT value, endowment value of an institution, had a high percentage of NA values. This limited the research to fewer and less stronger predictors.

To conclude, this research provided valuable guidance as to what impacts graduation rates in colleges. With better quality data and more time, this research can be taken further to come to better conclusions about the predictors that will help predict the 4-year and 6-year graduation rates in colleges.

7. References List

Aslanian, S., & Barshay, J. (2021). Retrieved from https://hechingerreport.org/predictive-analytics-boosting-college-graduation-rates-also-invade-privacy-and-reinforce-racial-inequities/

DeAngelo, L., Franke, R., Hurtado, S., Pryor, J. H., & Tran, S. (2011). Retrieved from https://www.heri.ucla.edu/DARCU/CompletingCollege2011.pdf

Joubert, S. (2020). 10 Benefits of Having a College Degree. Retrieved from https://bachelors-completion.northeastern.edu/news/is-a-bachelors-degree-worth-it/

Herlambang, W. D., Laksitowening, K. A., & Asror, I. (2021). Retrieved from https://ieeexplore.ieee.org/document/9527443

Ploutz, E. C. (2018). Retrieved from

https://digitalscholarship.unlv.edu/cgi/viewcontent.cgi?article=4312&context=thesesdissertations

8. Tables/Figures

Variable Name	Number of Missing Values
grad_100_value	327
grad_150_value	327
med_sat_value	1024
retain_percentile	258
ft_pct	4
endow_value	664

Table 1

Statistic	N	Mean	St. Dev.	Min	Median	Max
unitid	1,927	204,981.700	83,647.770	100,654	191,649	466,921
grad_100_value	1,927	33.946	22.993	0.000	30.300	100.000
grad_150_value	1,927	48.964	21.572	0.000	48.800	100.000
control_public_private	1,927	1.703	0.457	1	2	2
aid_value	1,927	11,386.760	7,398.423	1,450	8,925	41,580
pell_percentile	1,927	48.074	28.767	0	47	100
retain_percentile	1,927	49.318	28.742	0	49	100
ft_pct	1,927	81.724	18.229	4.400	87.500	100.000
fte_percentile	1,927	52.402	28.742	0	53	100
awards_per_state_value	1,927	22.183	2.565	11.900	22.300	34.200
awards_per_natl_value	1,927	22.466	0.922	21.500	22.500	24.600
M_W	1,927	32.113	19.151	0.000	33.333	100.000
F_W	1,927	37.217	19.763	0.000	40.909	100.000
M_B	1,927	6.360	10.185	0.000	2.792	98.865
F_B	1,927	8.942	14.926	0.000	3.306	100.000
M_H	1,927	3.796	6.251	0.000	1.709	66.667
F_H	1,927	4.991	8.105	0.000	2.174	71.429
M_Ai	1,927	0.634	4.217	0.000	0.000	100.000
F_Ai	1,927	0.791	4.409	0.000	0.120	100.000
M_A	1,927	2.260	4.633	0.000	0.797	80.000
F_A	1,927	2.898	6.383	0.000	0.986	100.000

Table 2

median_grad_100_private <dbl>	median_grad_150_private <dbl>
35.7	50

Table 3

median_grad_100_public <dbl>	median_grad_150_public <dbl>
21.8	45.9

Table 4

VIF

control_public_private	ln_aid_value	pell_percentile	retain_percentile	ft_pct	fte_percentile	M_W	F_W	M_B	F_B	M_H	F_H	M_Ai	F_Ai	M_A
1.265	1.757	1.815	1.955	1.410	1.349	13.385	14.893	5.461	9.868	2.711	4.832	1.728	1.785	3.484

Table 5

VIF

control_public_private	pell_percentile	retain_percentile	ft_pct	fte_percentile	F_W	F_A
1.021	1.646	1.771	1.161	1.235	1.133	1.098

Table 6

	grad_100_value		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-86.47	-101.68 – -71.25	**<0.001**
control public private	3.09	1.58 – 4.61	**<0.001**
ln aid value	14.11	12.80 – 15.42	**<0.001**
pell percentile	-0.16	-0.18 – -0.13	**<0.001**
retain percentile	0.22	0.19 – 0.25	**<0.001**
ft pct	0.13	0.09 – 0.17	**<0.001**
fte percentile	0.01	-0.01 – 0.04	0.406
M W	-0.36	-0.48 – -0.25	**<0.001**
F W	-0.17	-0.29 – -0.05	**0.007**
M B	-0.36	-0.50 – -0.22	**<0.001**
F B	-0.35	-0.48 – -0.22	**<0.001**
M H	-0.24	-0.40 – -0.08	**0.003**
F H	-0.54	-0.70 – -0.37	**<0.001**
M Ai	-0.40	-0.60 – -0.21	**<0.001**
F Ai	-0.43	-0.61 – -0.24	**<0.001**
M A	-0.60	-0.84 – -0.35	**<0.001**
Observations	1927		
R^2 / R^2 adjusted	0.644 / 0.641		

Table 7

```
##
## Call:
## lm(formula = grad_100_value ~ control_public_private + ln_aid_value +
##     pell_percentile + retain_percentile + ft_pct + fte_percentile +
##     M_W + F_W + M_B + F_B + M_H + F_H + M_Ai + F_Ai + M_A, data = institution_grad_details_wider)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.502  -7.971  -0.180   7.345  73.792
##
## Coefficients:
##                          Estimate Std. Error t value             Pr(>|t|)
## (Intercept)            -86.46635    7.75868 -11.144 < 0.0000000000000002 ***
## control_public_private   3.09341    0.77180   4.008       0.000063571973 ***
## ln_aid_value            14.10899    0.66782  21.127 < 0.0000000000000002 ***
## pell_percentile         -0.15594    0.01469 -10.612 < 0.0000000000000002 ***
## retain_percentile        0.22463    0.01526  14.717 < 0.0000000000000002 ***
## ft_pct                   0.13202    0.02044   6.458       0.000000000134 ***
## fte_percentile           0.01055    0.01268   0.832              0.40557
## M_W                     -0.36371    0.05995  -6.067       0.000000001564 ***
## F_W                     -0.16584    0.06128  -2.706              0.00686 **
## M_B                     -0.36202    0.07200  -5.028       0.000000541636 ***
## F_B                     -0.34869    0.06604  -5.280       0.000000143935 ***
## M_H                     -0.24267    0.08265  -2.936              0.00336 **
## F_H                     -0.53750    0.08510  -6.316       0.000000000333 ***
## M_Ai                    -0.40356    0.09782  -4.125       0.000038595133 ***
## F_Ai                    -0.42786    0.09511  -4.499       0.000007243881 ***
## M_A                     -0.59607    0.12641  -4.715       0.000002588797 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.77 on 1911 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6413
## F-statistic: 230.5 on 15 and 1911 DF,  p-value: < 0.00000000000000022
```

Table 8

	grad_100_value		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-86.47	-101.68 – -71.25	**<0.001**
control public private	3.09	1.58 – 4.61	**<0.001**
ln aid value	14.11	12.80 – 15.42	**<0.001**
pell percentile	-0.16	-0.18 – -0.13	**<0.001**
retain percentile	0.22	0.19 – 0.25	**<0.001**
ft pct	0.13	0.09 – 0.17	**<0.001**
fte percentile	0.01	-0.01 – 0.04	0.406
M W	-0.36	-0.48 – -0.25	**<0.001**
F W	-0.17	-0.29 – -0.05	**0.007**
M B	-0.36	-0.50 – -0.22	**<0.001**
F B	-0.35	-0.48 – -0.22	**<0.001**
M H	-0.24	-0.40 – -0.08	**0.003**
F H	-0.54	-0.70 – -0.37	**<0.001**
M Ai	-0.40	-0.60 – -0.21	**<0.001**
F Ai	-0.43	-0.61 – -0.24	**<0.001**
M A	-0.60	-0.84 – -0.35	**<0.001**
Observations	1927		
R^2 / R^2 adjusted	0.644 / 0.641		

Table 9

```
##
## Call:
## lm(formula = grad_100_value ~ control_public_private + ln_aid_value +
##     pell_percentile + pell_percentile^2 + retain_percentile +
##     retain_percentile^2 + ft_pct + ft_pct^2 + fte_percentile +
##     fte_percentile^2 + M_W + F_W + M_B + F_B + M_H + F_H + M_Ai +
##     F_Ai + M_A, data = institution_grad_details_wider)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.502  -7.971  -0.180   7.345  73.792
##
## Coefficients:
##                          Estimate Std. Error t value             Pr(>|t|)
## (Intercept)            -86.46635    7.75868 -11.144 < 0.0000000000000002 ***
## control_public_private   3.09341    0.77180   4.008       0.000063571973 ***
## ln_aid_value            14.10899    0.66782  21.127 < 0.0000000000000002 ***
## pell_percentile         -0.15594    0.01469 -10.612 < 0.0000000000000002 ***
## retain_percentile        0.22463    0.01526  14.717 < 0.0000000000000002 ***
## ft_pct                   0.13202    0.02044   6.458       0.000000000134 ***
## fte_percentile           0.01055    0.01268   0.832              0.40557
## M_W                     -0.36371    0.05995  -6.067       0.000000001564 ***
## F_W                     -0.16584    0.06128  -2.706              0.00686 **
## M_B                     -0.36202    0.07200  -5.028       0.000000541636 ***
## F_B                     -0.34869    0.06604  -5.280       0.000000143935 ***
## M_H                     -0.24267    0.08265  -2.936              0.00336 **
## F_H                     -0.53750    0.08510  -6.316       0.000000000333 ***
## M_Ai                    -0.40356    0.09782  -4.125       0.000038595133 ***
## F_Ai                    -0.42786    0.09511  -4.499       0.000007243881 ***
## M_A                     -0.59607    0.12641  -4.715       0.000002588797 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.77 on 1911 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6413
## F-statistic: 230.5 on 15 and 1911 DF,  p-value: < 0.00000000000000022
```

Table 10

	grad_100_value		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-26.99	-32.05 – -21.93	**<0.001**
control public private	10.07	8.56 – 11.58	**<0.001**
pell percentile	-0.16	-0.19 – -0.13	**<0.001**
retain percentile	0.28	0.25 – 0.31	**<0.001**
ft pct	0.31	0.27 – 0.35	**<0.001**
fte percentile	0.02	-0.00 – 0.05	0.074
F W	0.28	0.24 – 0.32	**<0.001**
F A	0.32	0.21 – 0.43	**<0.001**
Observations	1927		
R^2 / R^2 adjusted	0.558 / 0.557		

Table 11

```
##
## Call:
## lm(formula = grad_100_value ~ control_public_private + pell_percentile +
##     retain_percentile + ft_pct + fte_percentile + F_W + F_A,
##     data = institution_grad_details_wider)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.722  -7.922   0.706   8.747  76.937
##
## Coefficients:
##                         Estimate Std. Error t value             Pr(>|t|)
## (Intercept)            -26.98956    2.58042 -10.459 < 0.0000000000000002 ***
## control_public_private  10.06979    0.77103  13.060 < 0.0000000000000002 ***
## pell_percentile         -0.15810    0.01556 -10.163 < 0.0000000000000002 ***
## retain_percentile        0.27672    0.01615  17.135 < 0.0000000000000002 ***
## ft_pct                   0.30693    0.02062  14.887 < 0.0000000000000002 ***
## fte_percentile           0.02409    0.01349   1.786               0.0743 .
## F_W                      0.28151    0.01878  14.987 < 0.0000000000000002 ***
## F_A                      0.31803    0.05727   5.553          0.000000032 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.31 on 1919 degrees of freedom
## Multiple R-squared:  0.5584, Adjusted R-squared:  0.5568
## F-statistic: 346.6 on 7 and 1919 DF,  p-value: < 0.00000000000000022
```

Table 12

	grad_150_value		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-59.60	-73.27 – -45.93	**<0.001**
control public private	-5.43	-6.79 – -4.07	**<0.001**
ln aid value	12.52	11.35 – 13.70	**<0.001**
pell percentile	-0.08	-0.11 – -0.05	**<0.001**
retain percentile	0.28	0.26 – 0.31	**<0.001**
ft pct	0.12	0.08 – 0.15	**<0.001**
fte percentile	0.05	0.02 – 0.07	**<0.001**
M W	-0.20	-0.30 – -0.09	**<0.001**
F W	-0.11	-0.22 – -0.00	**0.050**
M B	-0.32	-0.45 – -0.20	**<0.001**
F B	-0.27	-0.39 – -0.15	**<0.001**
M H	-0.30	-0.45 – -0.16	**<0.001**
F H	-0.31	-0.46 – -0.16	**<0.001**
M Ai	-0.46	-0.63 – -0.28	**<0.001**
F Ai	-0.49	-0.66 – -0.32	**<0.001**
M A	-0.34	-0.56 – -0.12	**0.003**
Observations	1927		
R^2 / R^2 adjusted	0.674 / 0.671		

Table 13

```
##
## Call:
## lm(formula = grad_150_value ~ control_public_private + ln_aid_value +
##     pell_percentile + retain_percentile + ft_pct + fte_percentile +
##     M_W + F_W + M_B + F_B + M_H + F_H + M_Ai + F_Ai + M_A, data = institution_grad_details_wider)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.368  -6.083  -0.093   5.681  82.511
##
## Coefficients:
##                          Estimate Std. Error t value             Pr(>|t|)
## (Intercept)            -59.60252    6.97099  -8.550 < 0.0000000000000002 ***
## control_public_private  -5.42670    0.69344  -7.826  0.00000000000000829 ***
## ln_aid_value            12.52318    0.60002  20.871 < 0.0000000000000002 ***
## pell_percentile         -0.07999    0.01320  -6.058  0.00000000165388858 ***
## retain_percentile        0.28499    0.01371  20.781 < 0.0000000000000002 ***
## ft_pct                   0.11657    0.01837   6.347  0.00000000027285391 ***
## fte_percentile           0.04697    0.01139   4.122  0.00003912429826997 ***
## M_W                     -0.19671    0.05386  -3.652             0.000267 ***
## F_W                     -0.10810    0.05506  -1.963             0.049740 *
## M_B                     -0.32324    0.06469  -4.997  0.00000063603696248 ***
## F_B                     -0.26956    0.05934  -4.543  0.00000589468921890 ***
## M_H                     -0.30088    0.07426  -4.052  0.00005289382813157 ***
## F_H                     -0.30618    0.07646  -4.004  0.00006458377470523 ***
## M_Ai                    -0.45733    0.08789  -5.203  0.00000021676915942 ***
## F_Ai                    -0.49216    0.08545  -5.760  0.00000000980332743 ***
## M_A                     -0.34193    0.11358  -3.010             0.002642 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.37 on 1911 degrees of freedom
## Multiple R-squared:  0.6736, Adjusted R-squared:  0.671
## F-statistic: 262.9 on 15 and 1911 DF,  p-value: < 0.00000000000000022
```

Table 14

VIF

control_public_private	ln_aid_value	pell_percentile	retain_percentile	ft_pct	fte_percentile	M_W	F_W	M_B	F_B	M_H	F_H	M_Ai	F_Ai	M_A
1.265	1.757	1.815	1.955	1.410	1.349	13.385	14.893	5.461	9.868	2.711	4.832	1.728	1.785	3.484

Table 15

	grad_150_value		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-13.61	-17.49 – -9.73	**<0.001**
control public private	0.91	-0.46 – 2.28	0.192
retain percentile	0.37	0.34 – 0.39	**<0.001**
ft pct	0.27	0.23 – 0.31	**<0.001**
fte percentile	0.09	0.06 – 0.11	**<0.001**
F W	0.29	0.26 – 0.32	**<0.001**
M W	0.14	0.11 – 0.18	**<0.001**
F A	0.37	0.27 – 0.48	**<0.001**
Observations	1927		
R^2 / R^2 adjusted	0.590 / 0.588		

Table 16

```
##
## Call:
## lm(formula = grad_150_value ~ control_public_private + retain_percentile +
##     ft_pct + fte_percentile + F_W + M_W + F_A, data = institution_grad_details_wider)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.060  -5.684   0.639   6.958  79.422
##
## Coefficients:
##                          Estimate Std. Error t value             Pr(>|t|)
## (Intercept)            -13.60962    1.97798  -6.881  0.00000000000804208 ***
## control_public_private   0.90997    0.69709   1.305                0.192
## retain_percentile        0.36748    0.01332  27.597 < 0.0000000000000002 ***
## ft_pct                   0.26896    0.01873  14.359 < 0.0000000000000002 ***
## fte_percentile           0.08513    0.01216   7.001  0.00000000000349116 ***
## F_W                      0.29087    0.01642  17.710 < 0.0000000000000002 ***
## M_W                      0.14179    0.01770   8.011  0.00000000000000195 ***
## F_A                      0.37295    0.05366   6.950  0.00000000000498408 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.85 on 1919 degrees of freedom
## Multiple R-squared:  0.5896, Adjusted R-squared:  0.5881
## F-statistic: 393.8 on 7 and 1919 DF,  p-value: < 0.00000000000000022
```

Table 17

VIF

control_public_private	retain_percentile	ft_pct	fte_percentile	F_W	M_W	F_A
1.021	1.472	1.172	1.227	1.059	1.154	1.179

Table 18

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.7859 1.5204 1.3144 1.1572 1.11890 1.10134 1.03433
## Proportion of Variance 0.1993 0.1445 0.1080 0.0837 0.07825 0.07581 0.06686
## Cumulative Proportion  0.1993 0.3438 0.4518 0.5355 0.61373 0.68954 0.75640
##                            PC8     PC9   PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.88762 0.84954 0.7076 0.70557 0.64660 0.59621 0.58312
## Proportion of Variance 0.04924 0.04511 0.0313 0.03111 0.02613 0.02222 0.02125
## Cumulative Proportion  0.80565 0.85075 0.8821 0.91316 0.93929 0.96151 0.98276
##                           PC15                  PC16
## Standard deviation     0.52516 0.000000000000001455
## Proportion of Variance 0.01724 0.000000000000000000
## Cumulative Proportion  1.00000 1.000000000000000000
```

Table 19

```
loading_scores <- my_pca$rotation
loading_scores
```

```
##                                PC1         PC2         PC3         PC4
## control_public_private  0.07185157 -0.01329234 -0.30556933 -0.38151989
## pell_percentile        -0.41021738  0.06325529 -0.01073192 -0.03996399
## retain_percentile       0.37978133 -0.24267325 -0.09219150  0.13827623
## ft_pct                  0.21793197 -0.06458738 -0.40027193  0.04487793
## fte_percentile          0.19646286 -0.26799158 -0.12647980 -0.11846941
## M_W                     0.26511437  0.29273903  0.09681984  0.15150241
## F_W                     0.28749480  0.30160184  0.08226081 -0.24945763
## M_B                    -0.38609665 -0.04338976 -0.39346701  0.12416548
## F_B                    -0.40085215 -0.08290491 -0.37687834  0.06780561
## M_H                    -0.08360324 -0.41714558  0.24590398 -0.34884041
## F_H                    -0.09439129 -0.42000646  0.24010367 -0.45010567
## M_AI                   -0.06977275  0.06280312  0.22594090  0.16627494
## F_AI                   -0.05756825  0.04348859  0.28194767  0.11812359
## M_A                     0.10591097 -0.41609628  0.05321756  0.41848105
## F_A                     0.08546413 -0.37676394  0.03558178  0.37911218
## ln_aid_value            0.29047887 -0.06489025 -0.39501788 -0.16677170
##                                PC5         PC6         PC7         PC8
## control_public_private  0.3340983400 -0.20890576  0.35791442 -0.538728878
## pell_percentile         0.0208417298 -0.17800359  0.27859658  0.268586516
## retain_percentile       0.0175710691  0.04974042 -0.28959657  0.024563617
## ft_pct                  0.0520647360 -0.34587849 -0.17746962  0.539493848
## fte_percentile          0.0127933866  0.41357386 -0.37832624 -0.257038779
## M_W                    -0.2895412082 -0.48929535 -0.12399242 -0.303257201
## F_W                     0.0587946620  0.45319522  0.24600924  0.302751478
## M_B                    -0.0263988697  0.01791458 -0.17548489 -0.105580253
## F_B                     0.0189189994  0.18346265 -0.14144233 -0.018189983
## M_H                    -0.1413663420 -0.28779588 -0.12985303  0.001898796
## F_H                     0.0003879021 -0.06889966 -0.02373745  0.176741064
## M_AI                    0.5798893281 -0.13752102 -0.22264488  0.151598159
## F_AI                    0.5887946705 -0.01831263 -0.19316389 -0.142828499
## M_A                    -0.0295975277 -0.08311547  0.26782812 -0.019553978
## F_A                     0.0793926601  0.13038378  0.46437605 -0.038283646
## ln_aid_value            0.2951604648 -0.15372476  0.13204240  0.132323784
##                                PC9        PC10        PC11        PC12
## control_public_private  0.077132203 -0.081340449  0.087077752 -0.25860844
## pell_percentile         0.006227841  0.399106005 -0.374513360  0.23014192
## retain_percentile      -0.163226085 -0.179232439  0.272479083  0.31599566
## ft_pct                 -0.109571868 -0.062113172 -0.207884354 -0.51322979
## fte_percentile          0.246296133  0.383381128 -0.510596393 -0.09844132
## M_W                     0.051680082  0.008548136 -0.265346032  0.14475424
## F_W                    -0.011233157  0.145174892  0.163771400 -0.16513386
## M_B                    -0.038898355  0.056245888  0.169291189 -0.07785930
## F_B                    -0.027794110 -0.167921193  0.072296311  0.13824999
## M_H                     0.023472318  0.168432859  0.235370717 -0.22534932
## F_H                    -0.058830063 -0.273912130 -0.167532108  0.23298874
## M_AI                    0.671831317 -0.081608368  0.100282785  0.01489390
## F_AI                   -0.657493089  0.138481603 -0.131704678 -0.08512255
## M_A                     0.029840437  0.516572955  0.310350326 -0.05872976
## F_A                     0.034347169 -0.406074953 -0.368419398 -0.08639651
## ln_aid_value           -0.020987673  0.194177468  0.001368186  0.55900880
##                               PC13        PC14        PC15
## control_public_private  0.29475672 -0.07725087  0.103767459
## pell_percentile         0.37916591 -0.37938348 -0.062669751
## retain_percentile       0.45831041 -0.48874417 -0.014812776
## ft_pct                  0.09083494  0.08327677  0.064865071
## fte_percentile          0.03170619 -0.01168835  0.005052680
## M_W                     0.02281960  0.02906297  0.028848484
## F_W                     0.02824457 -0.13371612  0.006646245
## M_B                    -0.42908469 -0.44924490  0.361384154
## F_B                     0.32502533  0.43262422 -0.338417516
## M_H                    -0.16847970 -0.17838896 -0.542353948
## F_H                     0.02431257  0.16341793  0.528959617
## M_AI                    0.02472268 -0.04144492  0.019022433
## F_AI                   -0.04411070  0.03278292 -0.034633704
## M_A                     0.13455918  0.28151546  0.276155592
## F_A                    -0.18854987 -0.20042158 -0.222011444
## ln_aid_value           -0.42001742  0.12533396 -0.183827740
##                                PC16
## control_public_private  0.000000000000000004944777
## pell_percentile        -0.000000000000000004333466
## retain_percentile      -0.000000000000000015927209
## ft_pct                 -0.000000000000000002363582
## fte_percentile          0.000000000000000043604042
## M_W                     0.533616271620035278541 25
## F_W                     0.550662039393537921057566
## M_B                     0.283782769080082345337664
## F_B                     0.415887863870766518762669
## M_H                     0.174188782950036856150 09
## F_H                     0.225830858147269886160 30
## M_AI                    0.117497293186172843681 98
## F_AI                    0.122839376693206325796 84
## M_A                     0.129097508917815581387 46
## F_A                     0.177847629391645473129 20
## ln_aid_value           -0.000000000000000009447479
```

Table 20

	grad_100_value		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	33.95	33.30 – 34.59	**<0.001**
PC1	8.82	8.46 – 9.18	**<0.001**
PC2	-1.69	-2.11 – -1.27	**<0.001**
PC3	-5.47	-5.96 – -4.98	**<0.001**
PC4	-1.08	-1.64 – -0.53	**<0.001**
PC5	3.26	2.69 – 3.84	**<0.001**
PC6	0.87	0.29 – 1.45	**0.004**
PC7	-0.25	-0.87 – 0.37	0.429
Observations	1927		
R^2 / R^2 adjusted	0.610 / 0.608		

Table 21

```
##
## Call:
## lm(formula = grad_100_value ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 +
##     PC7, data = scores_combined)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.782  -8.226  -0.272   8.516  72.074
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  33.9463     0.3278 103.554 < 0.0000000000000002 ***
## PC1           8.8222     0.1836  48.050 < 0.0000000000000002 ***
## PC2          -1.6896     0.2157  -7.834  0.00000000000000775 ***
## PC3          -5.4662     0.2495 -21.911 < 0.0000000000000002 ***
## PC4          -1.0828     0.2833  -3.821              0.000137 ***
## PC5           3.2626     0.2931  11.133 < 0.0000000000000002 ***
## PC6           0.8703     0.2977   2.923              0.003506 **
## PC7          -0.2506     0.3170  -0.790              0.429371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.39 on 1919 degrees of freedom
## Multiple R-squared:  0.6097, Adjusted R-squared:  0.6083
## F-statistic: 428.3 on 7 and 1919 DF,  p-value: < 0.00000000000000022
```

Table 22

	grad_150_value		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	48.96	48.37 – 49.56	**<0.001**
PC1	8.72	8.38 – 9.05	**<0.001**
PC2	-2.11	-2.50 – -1.71	**<0.001**
PC3	-3.99	-4.45 – -3.54	**<0.001**
PC4	0.62	0.10 – 1.13	**0.019**
PC5	0.61	0.07 – 1.14	**0.026**
PC6	1.17	0.63 – 1.71	**<0.001**
PC7	-1.94	-2.52 – -1.37	**<0.001**
Observations	1927		
R^2 / R^2 adjusted	0.616 / 0.615		

Table 23

```
##
## Call:
## lm(formula = grad_150_value ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 +
##     PC7, data = scores_combined)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.745  -5.805   0.550   6.849  81.815
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  48.9635     0.3050 160.561 < 0.0000000000000002 ***
## PC1           8.7165     0.1708  51.034 < 0.0000000000000002 ***
## PC2          -2.1056     0.2006 -10.495 < 0.0000000000000002 ***
## PC3          -3.9936     0.2321 -17.208 < 0.0000000000000002 ***
## PC4           0.6179     0.2636   2.344               0.0192 *
## PC5           0.6067     0.2726   2.225               0.0262 *
## PC6           1.1684     0.2770   4.219      0.0000257223139 ***
## PC7          -1.9438     0.2949  -6.591      0.0000000000562 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.39 on 1919 degrees of freedom
## Multiple R-squared:  0.6163, Adjusted R-squared:  0.6149
## F-statistic: 440.3 on 7 and 1919 DF,  p-value: < 0.00000000000000022
```

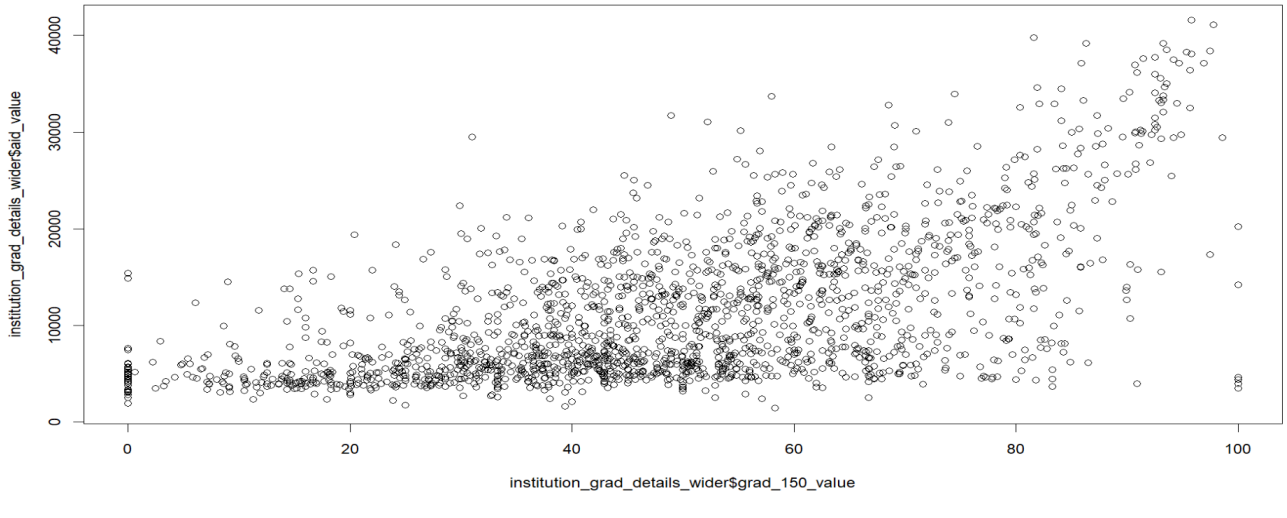
Table 24

	grad_100_value			grad_100_value		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	-26.99	-32.05 – -21.93	**<0.001**	33.95	33.30 – 34.59	**<0.001**
control public private	10.07	8.56 – 11.58	**<0.001**			
pell percentile	-0.16	-0.19 – -0.13	**<0.001**			
retain percentile	0.28	0.25 – 0.31	**<0.001**			
ft pct	0.31	0.27 – 0.35	**<0.001**			
fte percentile	0.02	-0.00 – 0.05	0.074			
F W	0.28	0.24 – 0.32	**<0.001**			
F A	0.32	0.21 – 0.43	**<0.001**			
PC1				8.82	8.46 – 9.18	**<0.001**
PC2				-1.69	-2.11 – -1.27	**<0.001**
PC3				-5.47	-5.96 – -4.98	**<0.001**
PC4				-1.08	-1.64 – -0.53	**<0.001**
PC5				3.26	2.69 – 3.84	**<0.001**
PC6				0.87	0.29 – 1.45	**0.004**
PC7				-0.25	-0.87 – 0.37	0.429
Observations	1927			1927		
R^2 / R^2 adjusted	0.558 / 0.557			0.610 / 0.608		

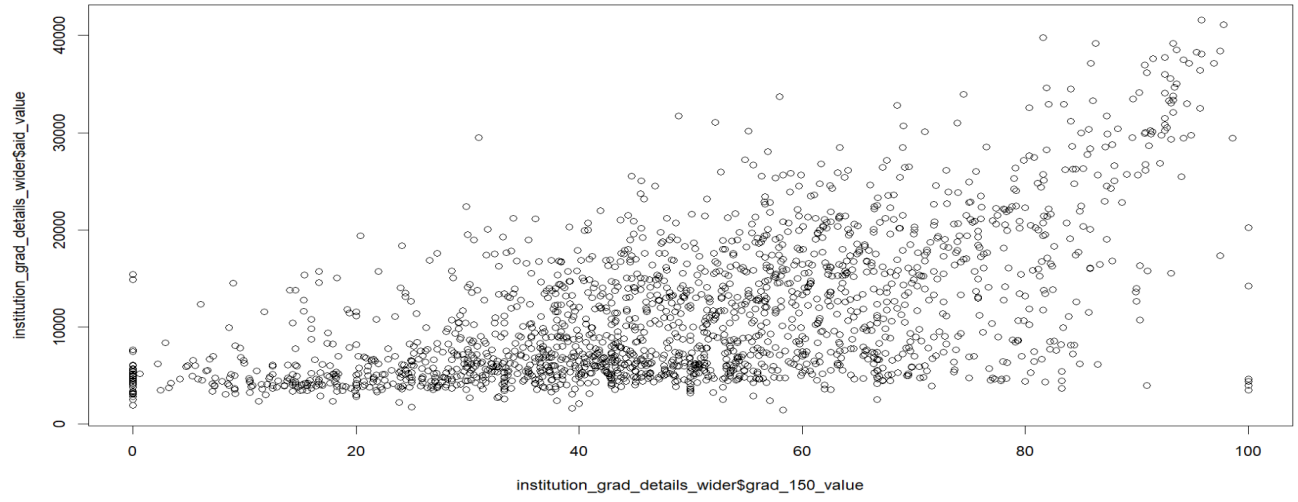
Table 25

	grad_150_value			grad_150_value		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	-13.61	-17.49 – -9.73	**<0.001**	48.96	48.37 – 49.56	**<0.001**
control public private	0.91	-0.46 – 2.28	0.192			
retain percentile	0.37	0.34 – 0.39	**<0.001**			
ft pct	0.27	0.23 – 0.31	**<0.001**			
fte percentile	0.09	0.06 – 0.11	**<0.001**			
F W	0.29	0.26 – 0.32	**<0.001**			
M W	0.14	0.11 – 0.18	**<0.001**			
F A	0.37	0.27 – 0.48	**<0.001**			
PC1				8.72	8.38 – 9.05	**<0.001**
PC2				-2.11	-2.50 – -1.71	**<0.001**
PC3				-3.99	-4.45 – -3.54	**<0.001**
PC4				0.62	0.10 – 1.13	**0.019**
PC5				0.61	0.07 – 1.14	**0.026**
PC6				1.17	0.63 – 1.71	**<0.001**
PC7				-1.94	-2.52 – -1.37	**<0.001**
Observations	1927			1927		
R^2 / R^2 adjusted	0.590 / 0.588			0.616 / 0.615		

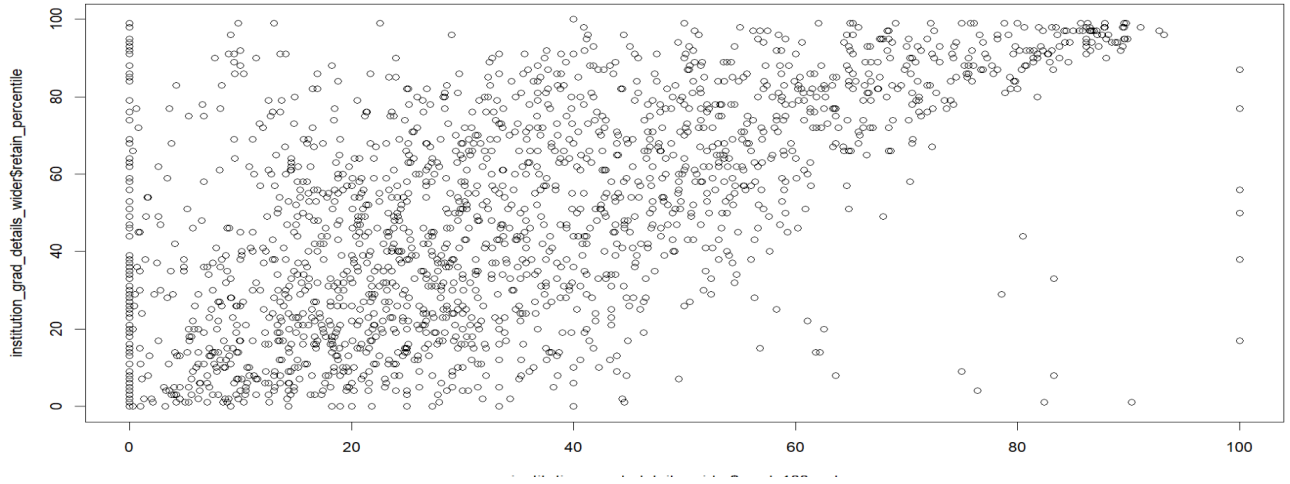
Table 26



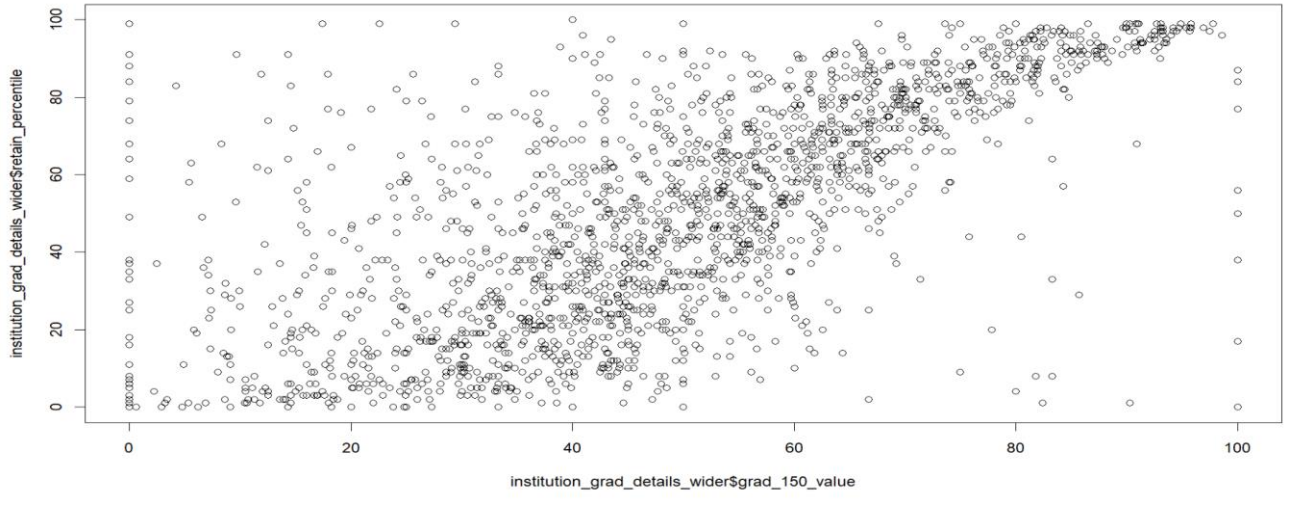
Graph 1



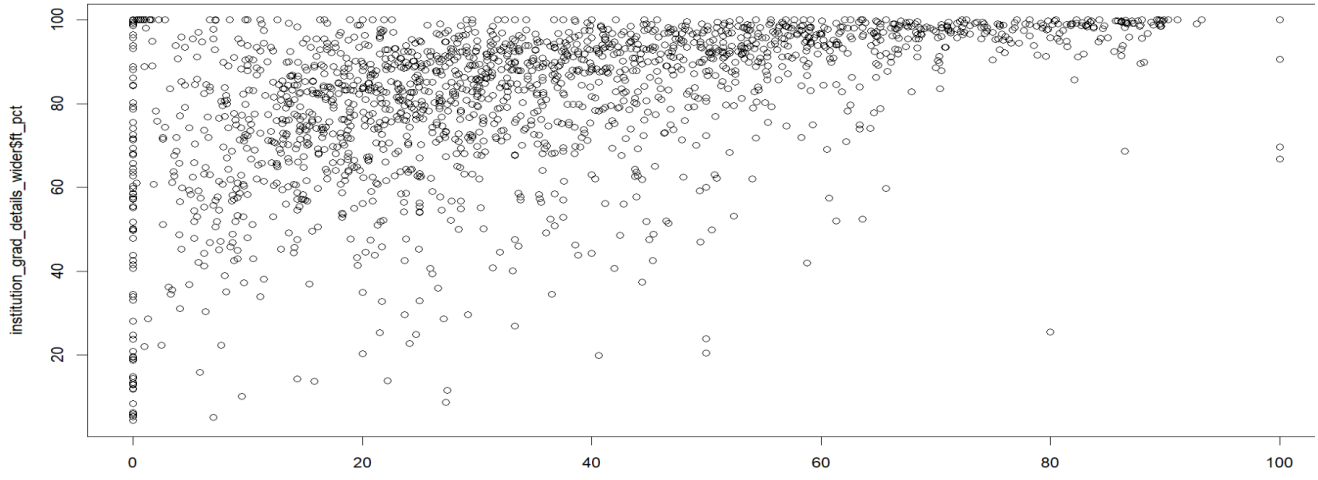
Graph 2



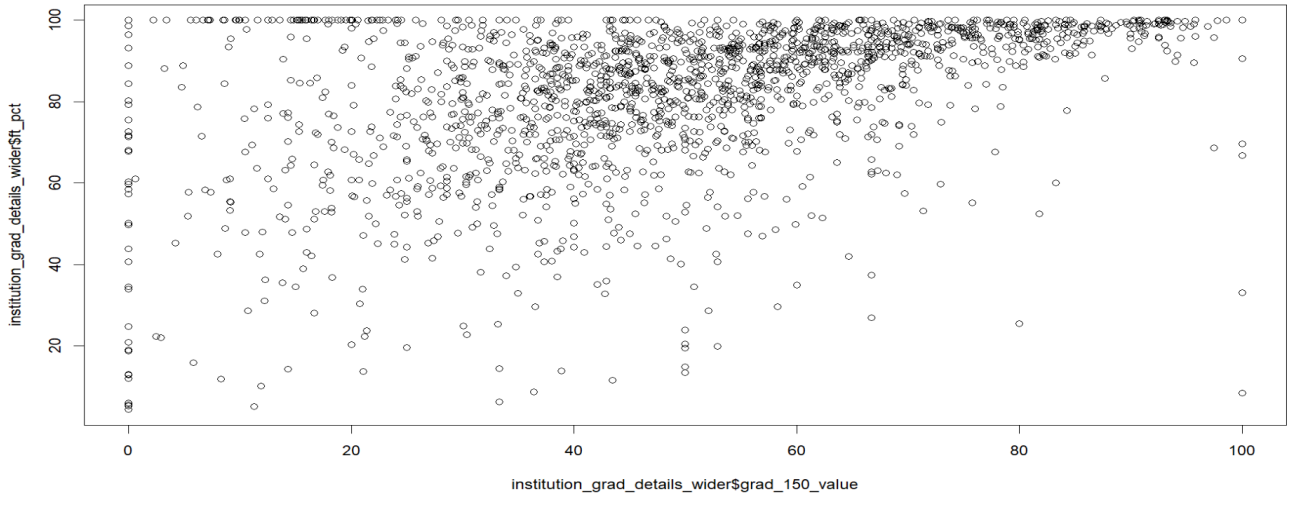
Graph 3



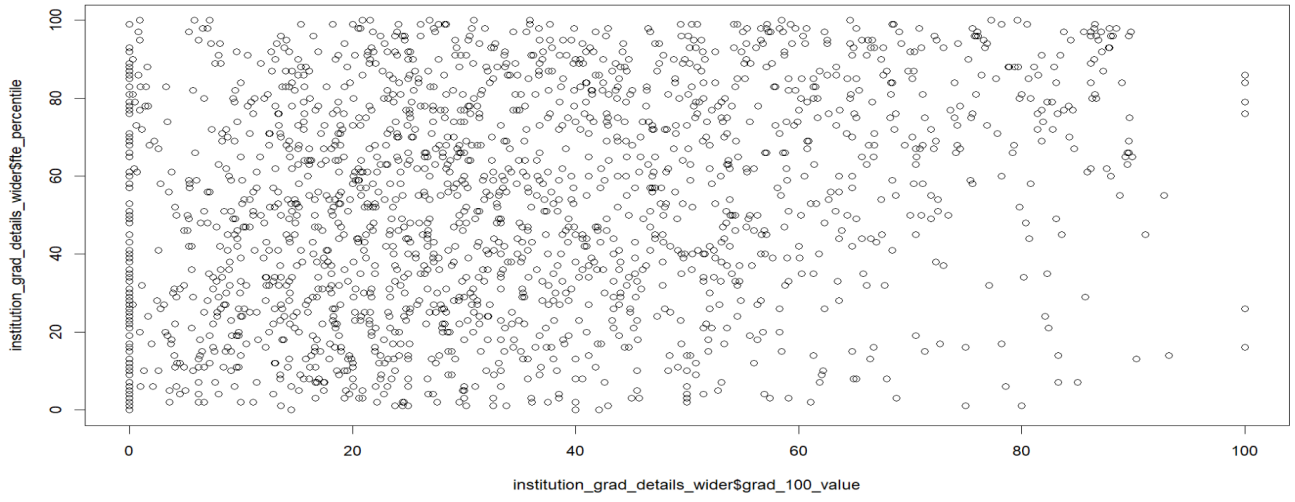
Graph 4



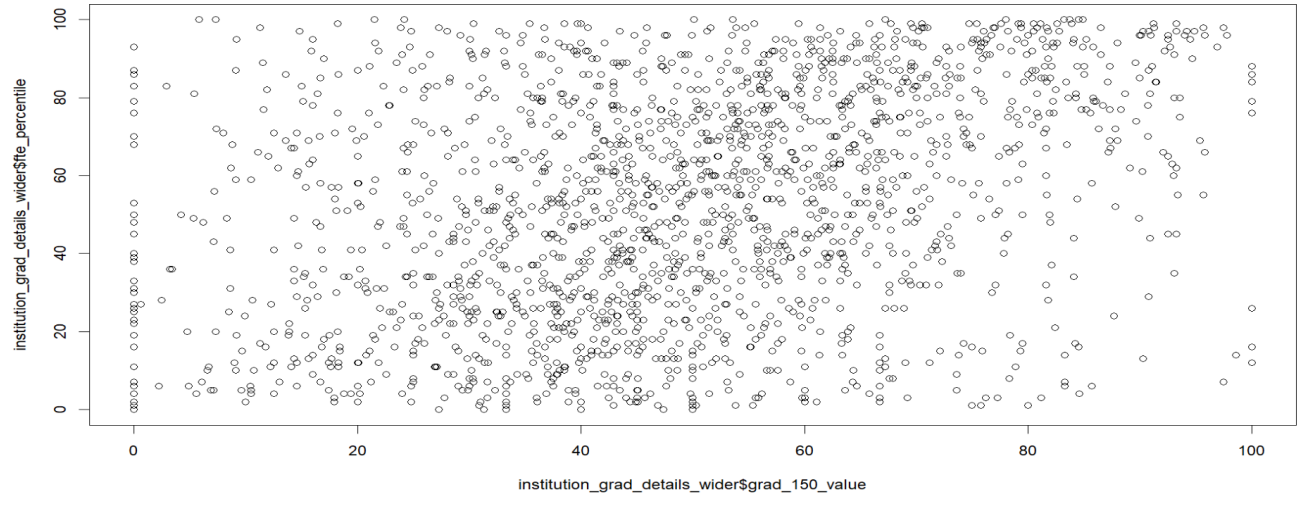
Graph 5



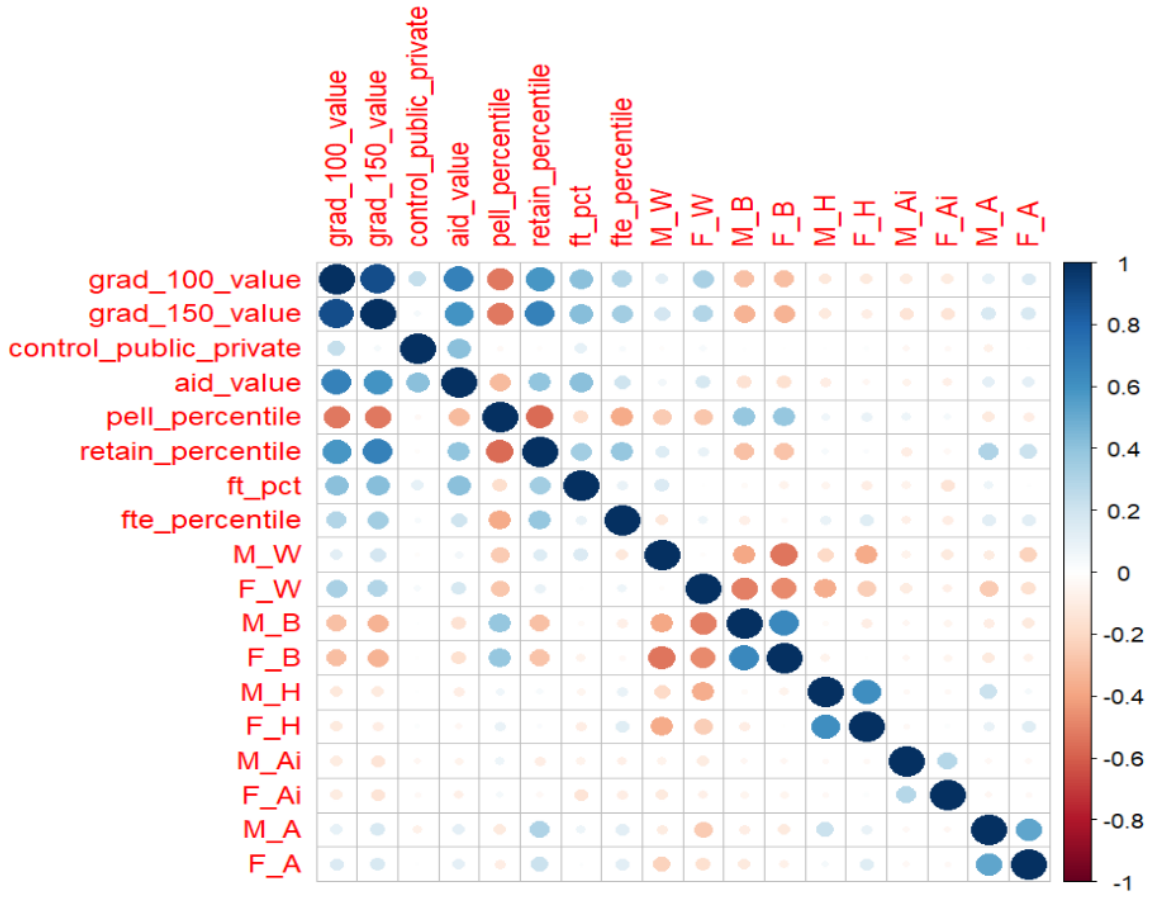
Graph 6



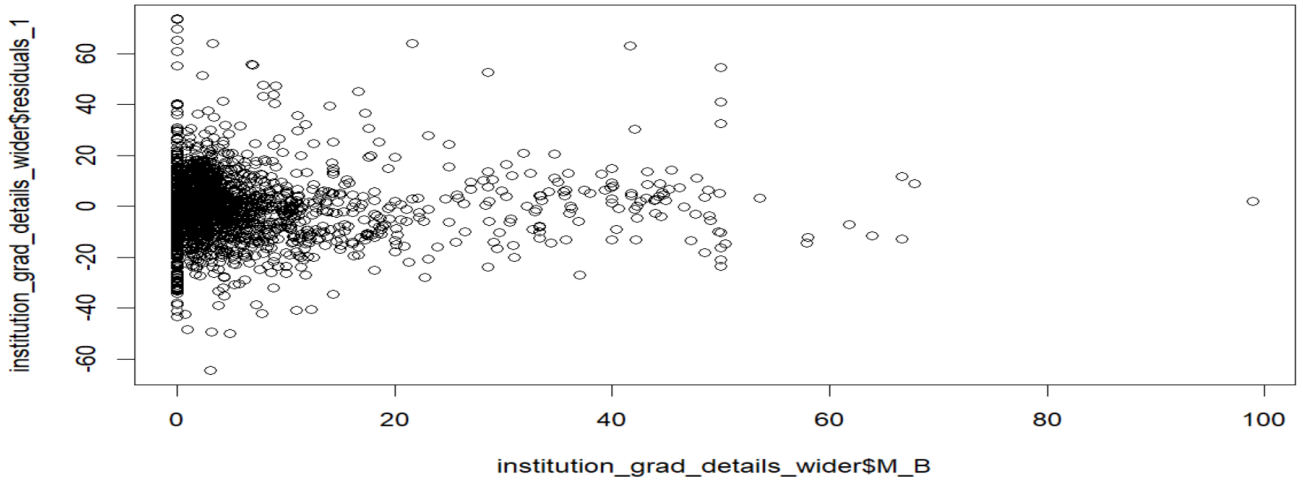
Graph 7



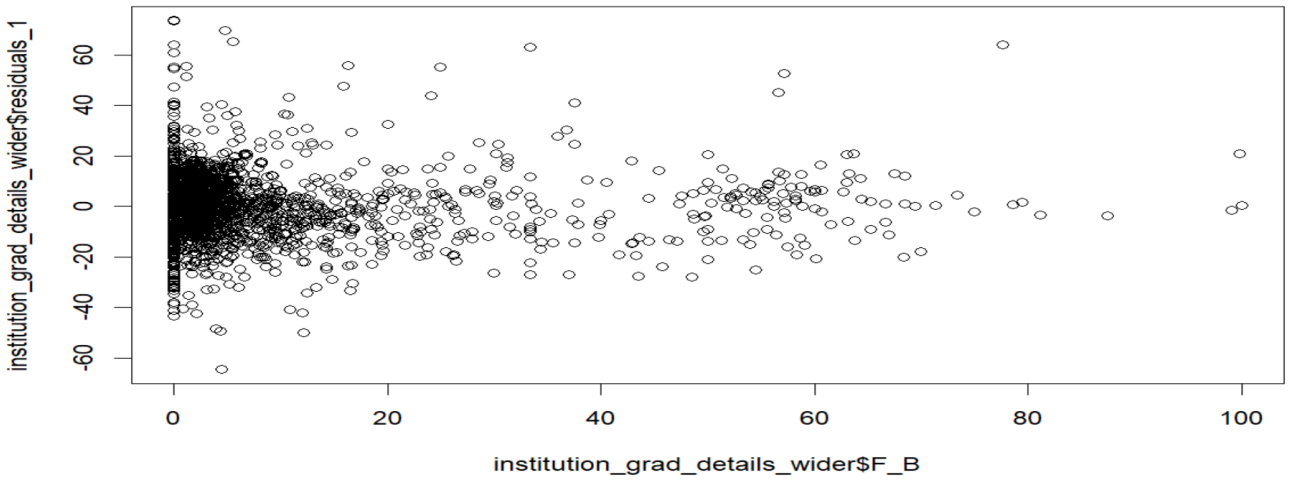
Graph 8



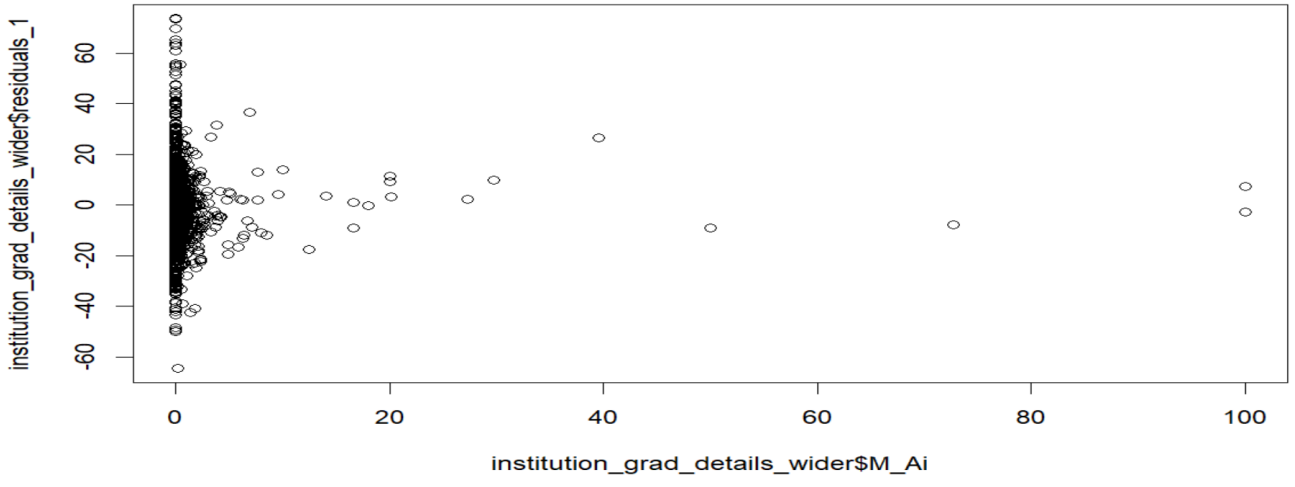
Graph 9



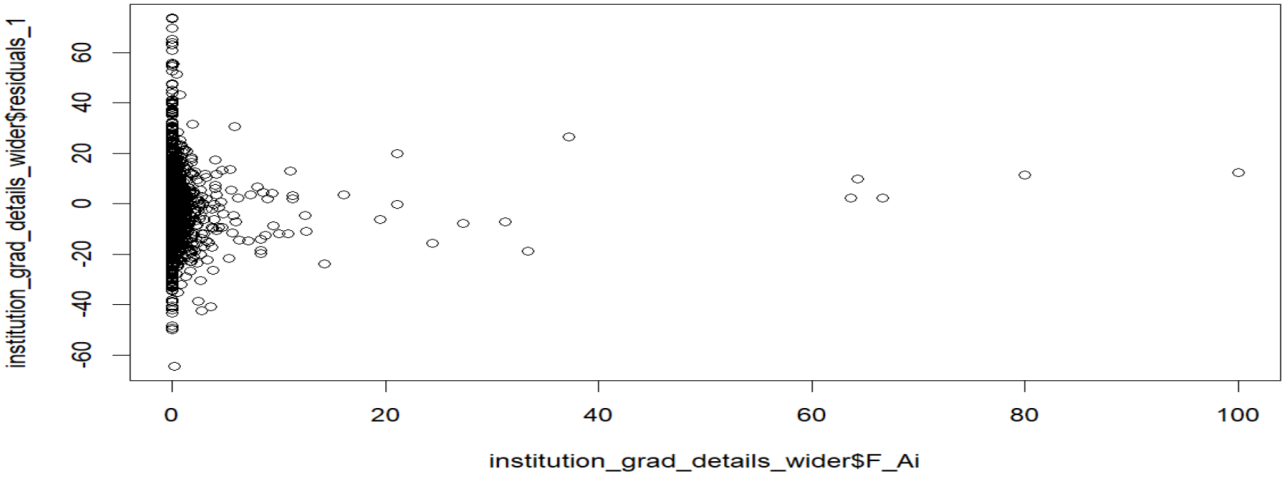
Graph 10



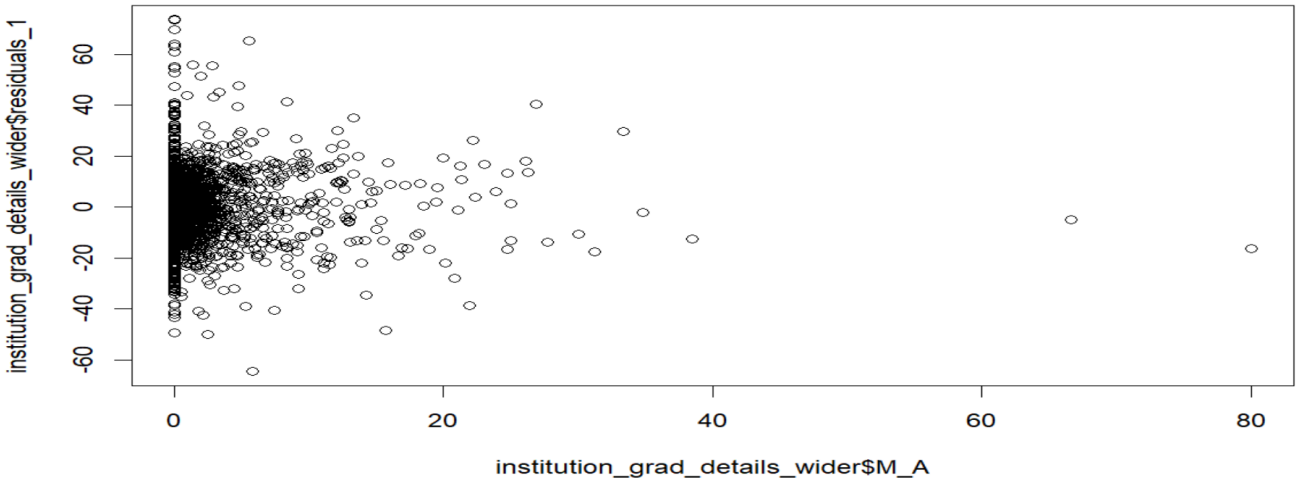
Graph 11



Graph 12

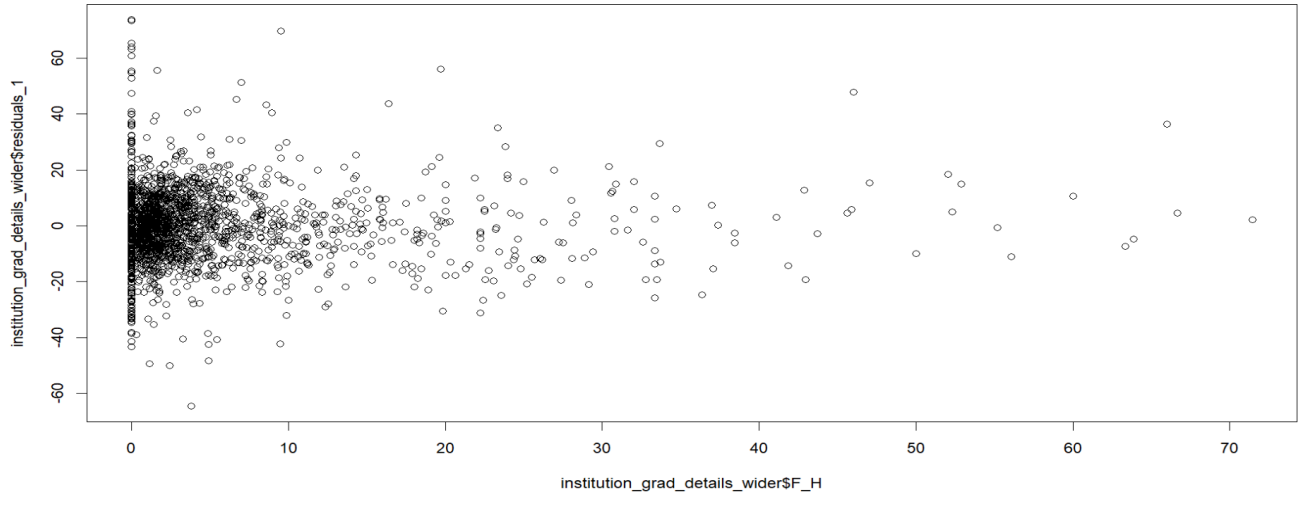


Graph 13



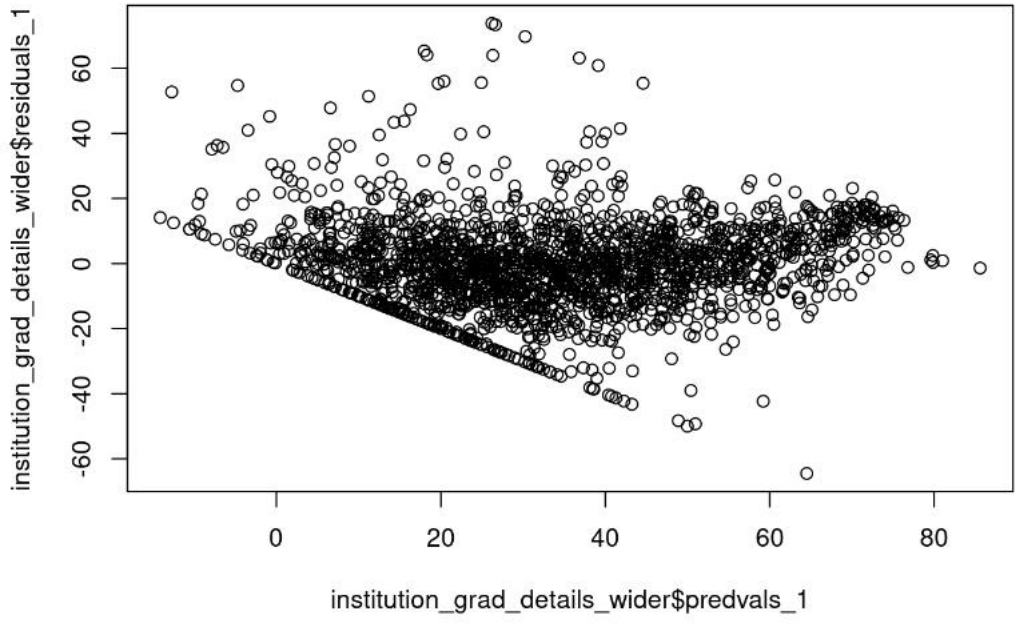
Graph 14

Graph 15



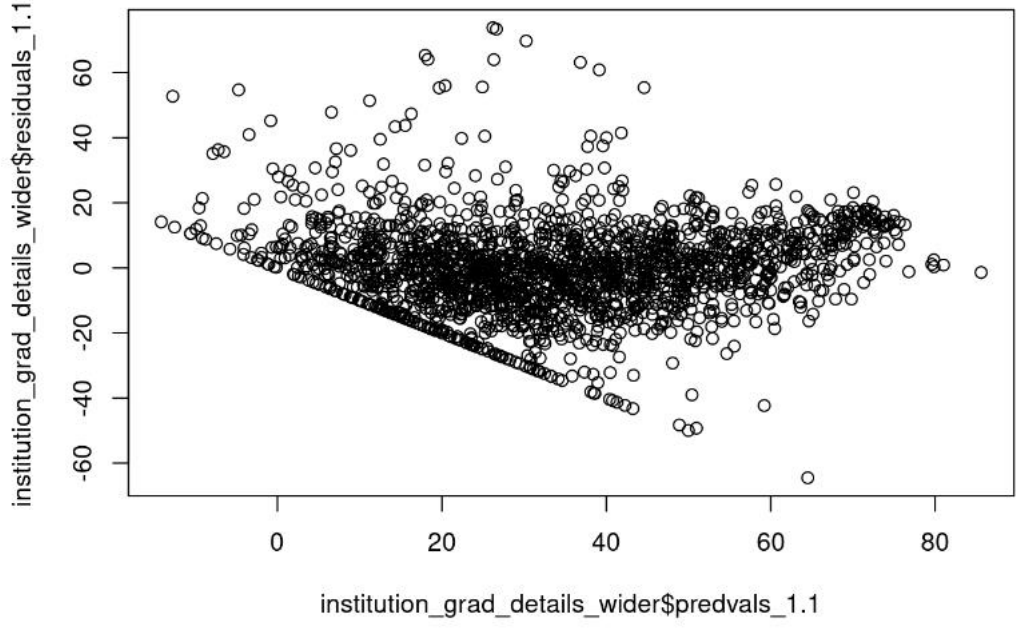
Graph 16

Graph 17

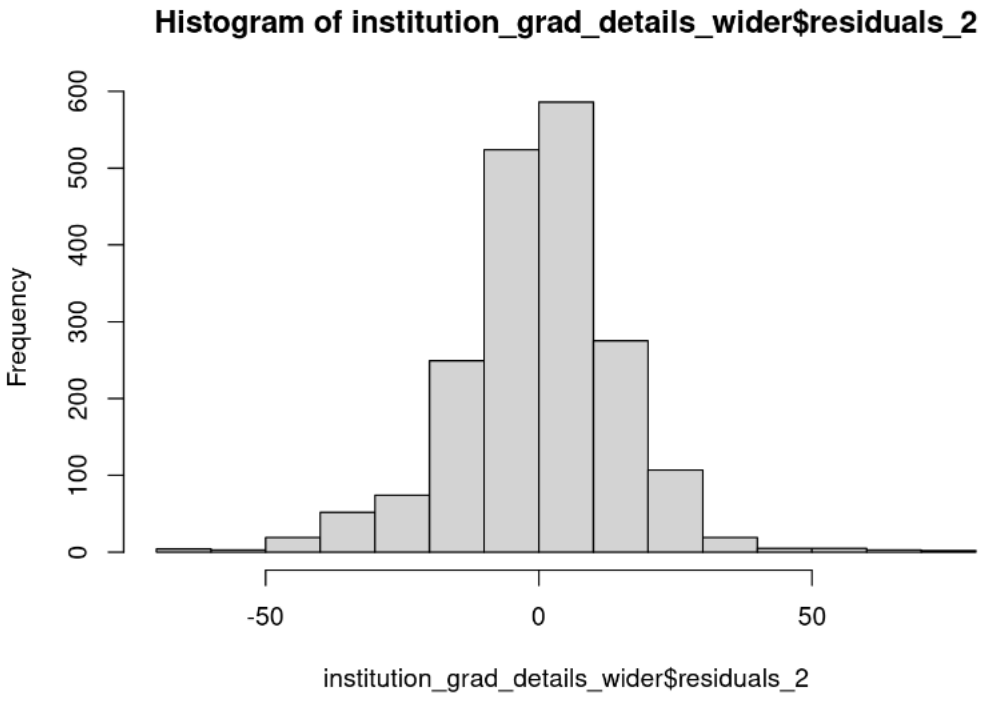


Graph 18

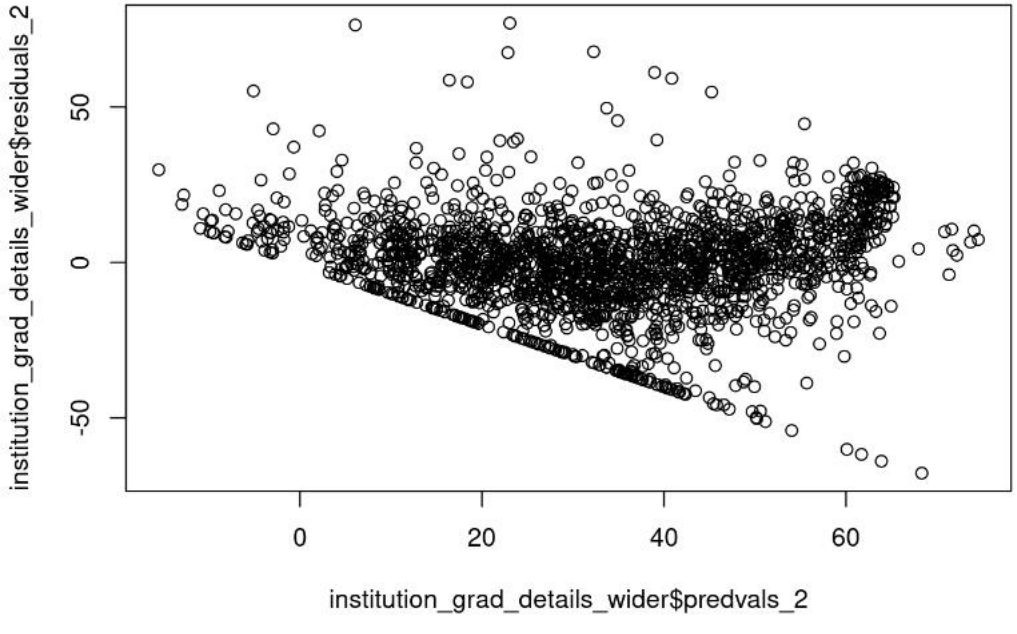
Graph 19



Graph 20

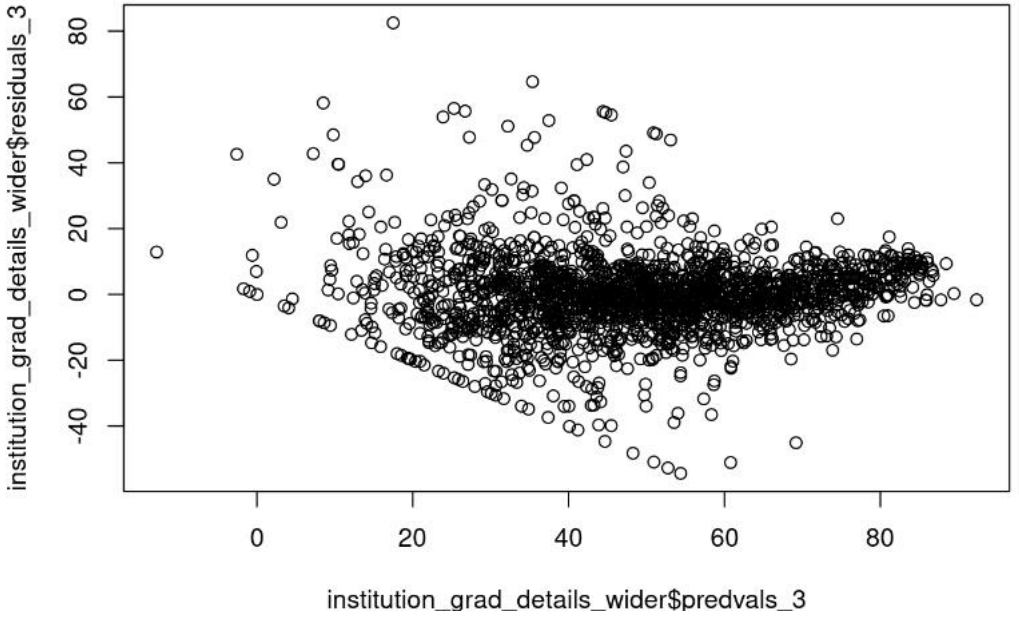


Graph 21

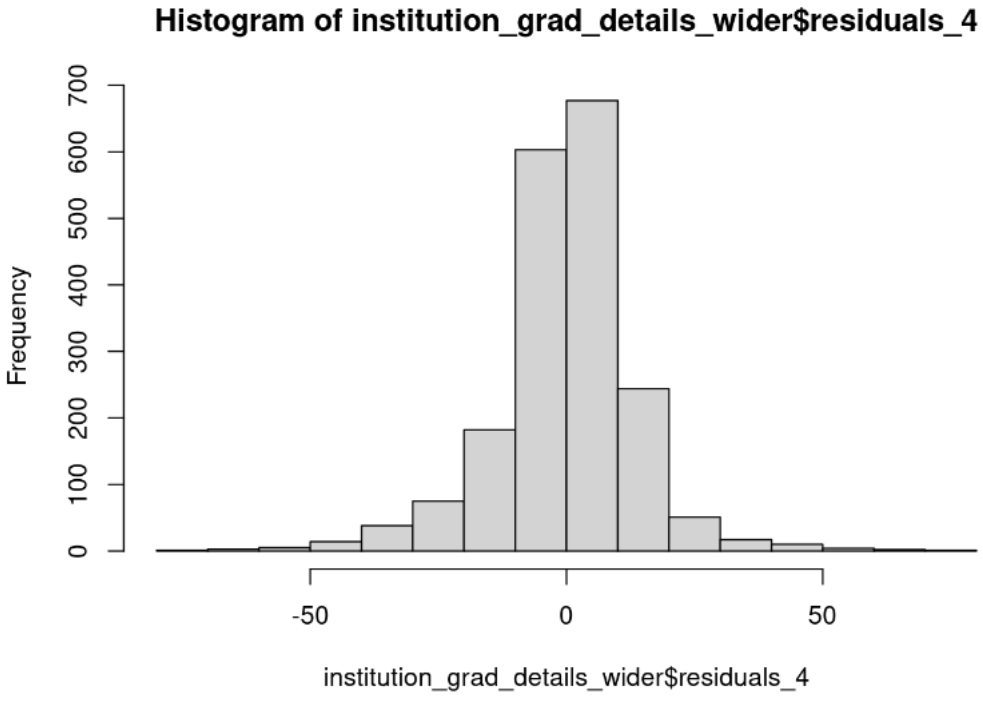


Graph 22

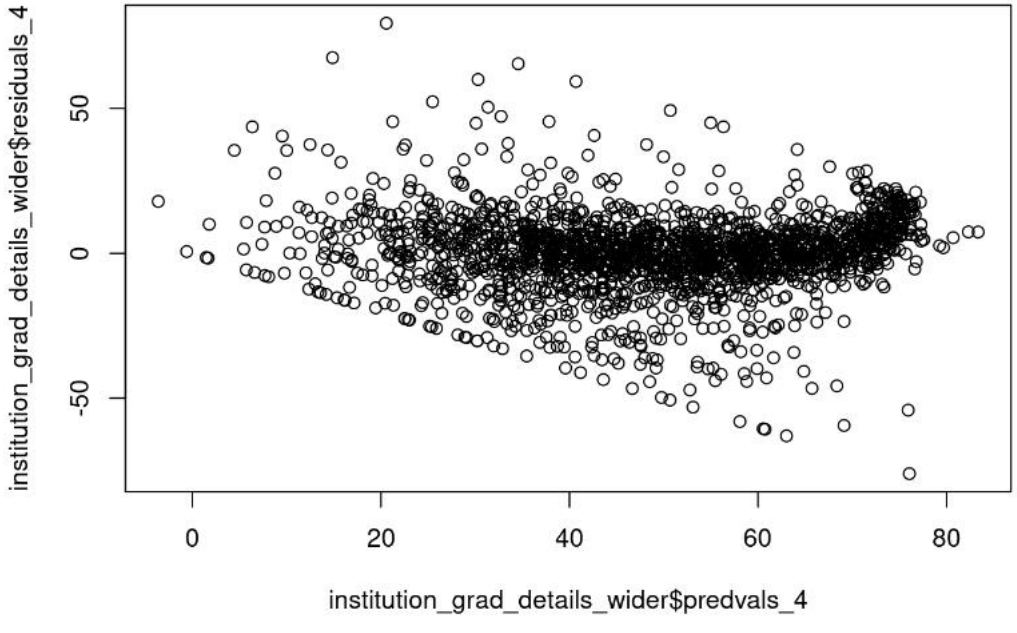
Graph 23



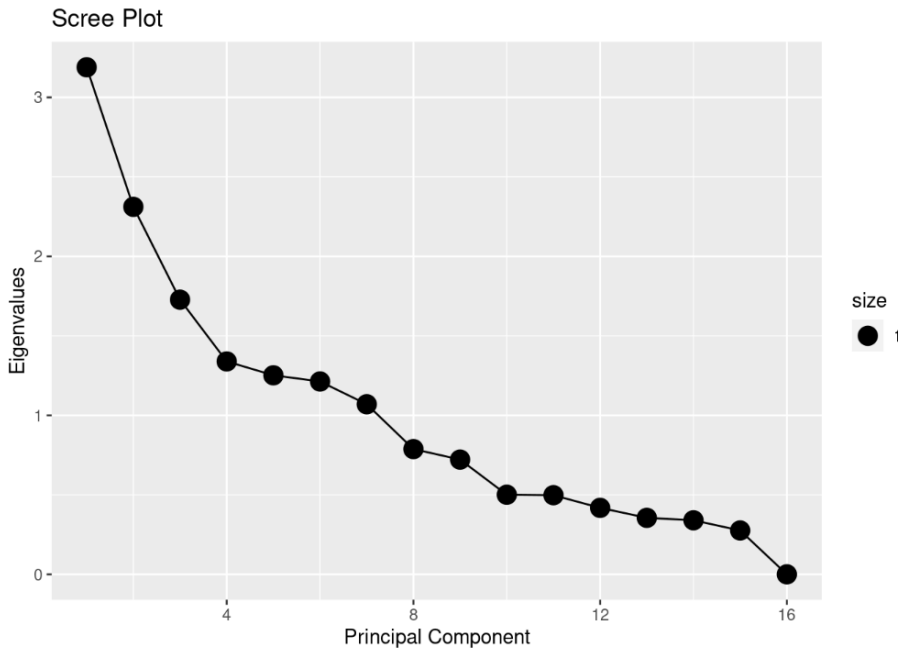
Graph 24



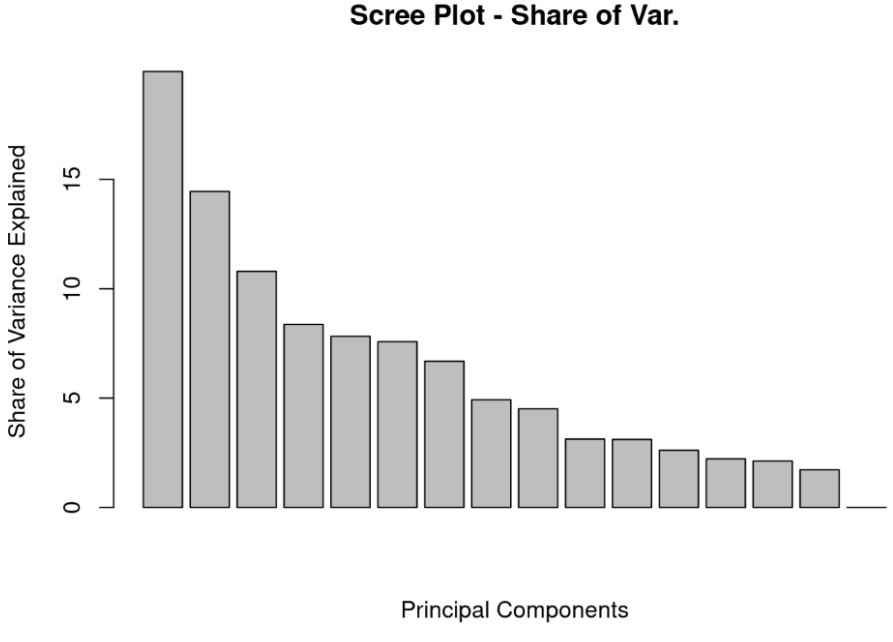
Graph 25



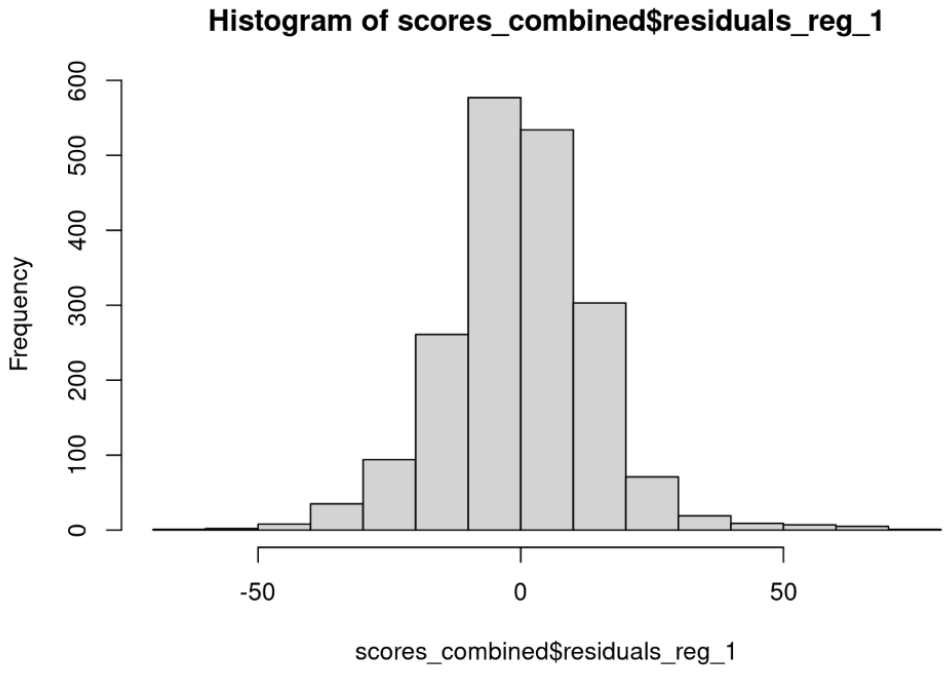
Graph 26



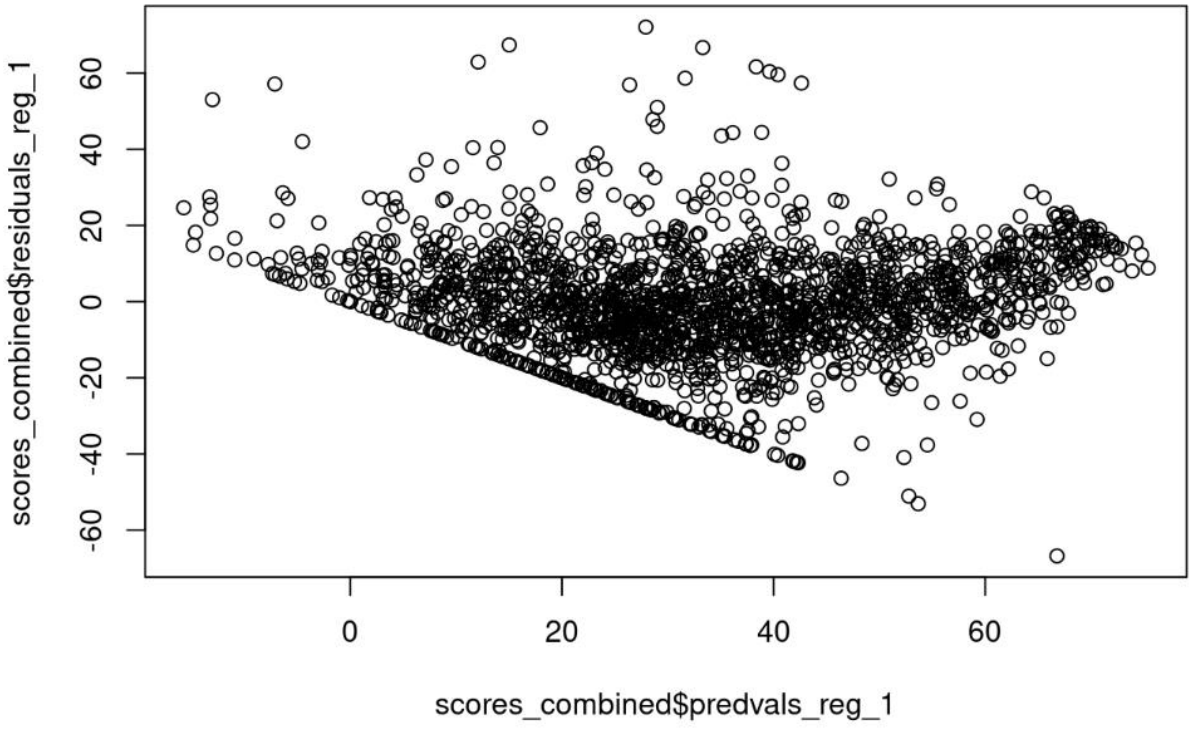
Graph 27



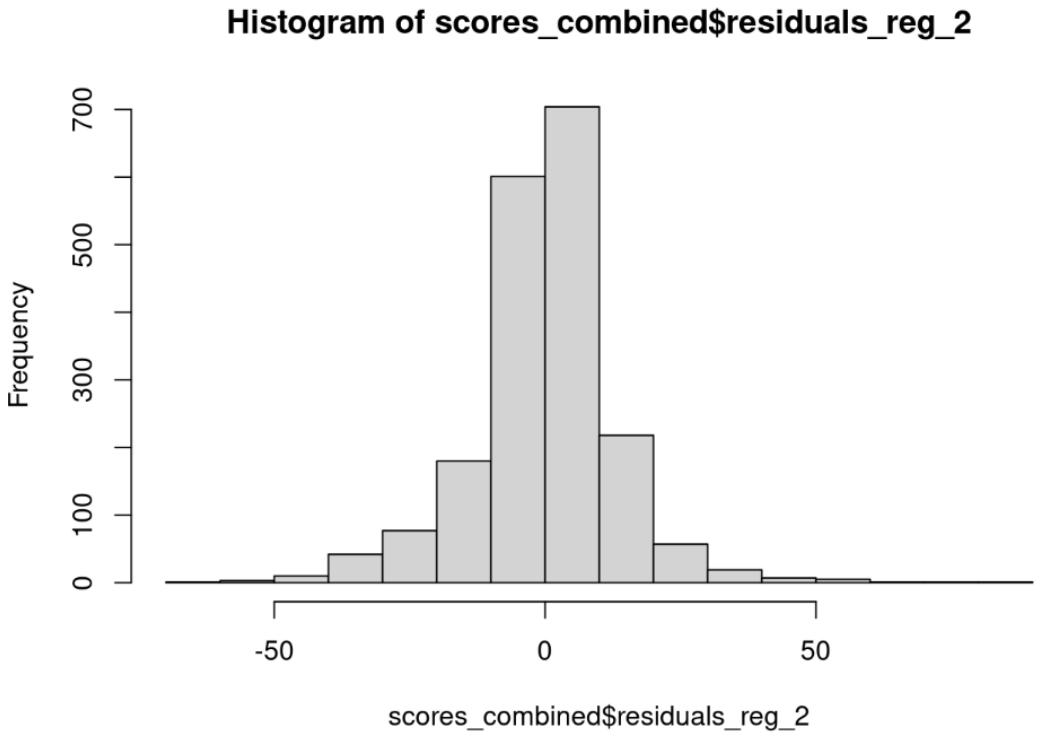
Graph 28



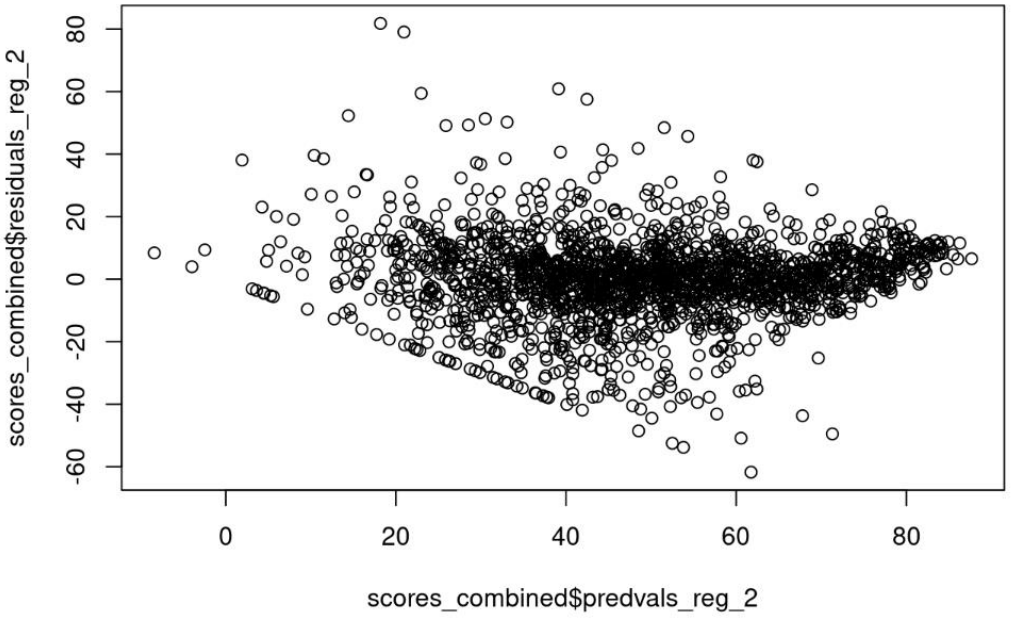
Graph 29



Graph 30


Histogram of scores_combined$residuals_reg_2
Frequency
0
100
300
500
700
-50
0
50
scores_combined$residuals_reg_2

Graph 31


scores_combined$residuals_reg_2
-60
-40
-20
0
20
40
60
80
0
20
40
60
80
scores_combined$predvals_reg_2

Graph 32

9. Acknowledgements

I would like to thank Professor Nolan for her guidance and support during my work on this project.

10. Attach your R Code & Output

Final Project

Anish Gupta

2023-11-07

Read College Completion Datasets

```
#setwd("C:\\Users\\anish\\Documents\\Applied Multivariate #Methods\\databeats-college-completion\\databeats-college-completion\\data")

institution_details <- read_csv("cc_institution_details.csv")
```

```
## Rows: 3798 Columns: 62
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (12): chronname, city, state, level, control, basic, hbcu, flagship, sit...
## dbl (50): unitid, long_x, lat_y, student_count, awards_per_value, awards_per...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
institution_grads <- read_csv("cc_institution_grads.csv")
```

```
## Rows: 1302102 Columns: 10
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (3): gender, race, cohort
## dbl (7): unitid, year, grad_cohort, grad_100, grad_150, grad_100_rate, grad_...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Data Wrangling for institution_details

```
#From the institution details dataframe, filter and keep 4-year institutions.
four_year_institution_details <- institution_details %>%
  filter(level == "4-year")

#Check to see if the control binary variable is Public, then assign a value of 1, else assign a value of 2.
four_year_institution_details$control_public_private = ifelse(four_year_institution_details$control == "Public", 1, 2)

dim(four_year_institution_details)
```

```
## [1] 2339   63
```

```
four_year_institution_details %>%
  summarise(sum_grad_100_value = sum(is.na(grad_100_value)),
            sum_grad_150_value = sum(is.na(grad_150_value)),
            sum_med_sat_value = sum(is.na(med_sat_value)),
            sum_aid_value = sum(is.na(aid_value)),
            sum_pell_percentile = sum(is.na(pell_percentile)),
            sum_retain_percentile = sum(is.na(retain_percentile)),
            sum_ft_pct = sum(is.na(ft_pct)),
            sum_endow_value = sum(is.na(endow_value)),
            sum_fte_percentile = sum(is.na(fte_percentile)),
            sum_awards_per_state_value = sum(is.na(awards_per_state_value)),
            sum_awards_per_natl_value = sum(is.na(awards_per_natl_value)),
            sum_control_public_private = sum(is.na(control_public_private)))
```

```
## # A tibble: 1 × 12
##   sum_grad_100_value sum_grad_150_value sum_med_sat_value sum_aid_value
##                <int>              <int>             <int>         <int>
## 1                327                327              1024             0
## # ℹ 8 more variables: sum_pell_percentile <int>, sum_retain_percentile <int>,
## #   sum_ft_pct <int>, sum_endow_value <int>, sum_fte_percentile <int>,
## #   sum_awards_per_state_value <int>, sum_awards_per_natl_value <int>,
## #   sum_control_public_private <int>
```

Create a subset of the necessary variables and drop med_sat_value, endow_value.

```
four_year_institution_details_subset <- four_year_institution_details %>%
  select(unitid, grad_100_value, grad_150_value, control_public_private, aid_value, pell_percentile, retain_percentile, ft_p
ct, fte_percentile, awards_per_state_value,awards_per_natl_value)

dim(four_year_institution_details_subset)
```

```
## [1] 2339   11
```

Drop remaining rows with NA values

```
#For this project, drop the NA values.
clean_four_year_institution_details <- four_year_institution_details_subset %>% drop_na()

dim(clean_four_year_institution_details)
```

```
## [1] 1930   11
```

Wrangle institution_grads

```
#table(institution_grads$cohort)

institution_grads %>%
  summarise(sum_institution_grads_year = sum(is.na(year)),
            sum_institution_grads_gender = sum(is.na(gender)),
            sum_institution_grads_race = sum(is.na(race))
            )
```

```
## # A tibble: 1 × 3
##   sum_institution_grads_year sum_institution_grads_gender sum_institution_grad…¹
##                        <int>                         <int>                  <int>
## 1                          0                             0                      0
## # ℹ abbreviated name: ¹sum_institution_grads_race
```

```
#Filter for Bachelor's/equivalent-seeking cohort at 4-year institutions
four_year_institution_grads <- institution_grads %>%
  filter(cohort == "4y bach",
         gender != "B",
         race != "X")
four_year_institution_grads <- four_year_institution_grads %>%
  group_by(unitid) %>%
  filter(year == max(year))

#Create a subset of the necessary variables.
four_year_institution_grads_subset <- four_year_institution_grads %>%
  select(unitid, gender, race, grad_cohort)

#For this project, drop the NA values.
clean_four_year_institution_grads <- four_year_institution_grads_subset %>% drop_na()

clean_four_year_institution_grads_percent <- clean_four_year_institution_grads %>%
  group_by(unitid) %>%
  mutate(grad_cohort_percent = (grad_cohort / sum(grad_cohort))*100)
```

```
institution_grad_details <- clean_four_year_institution_details %>% left_join( clean_four_year_institution_grads_percent,
        by=c('unitid'='unitid'))

summary(institution_grad_details)
```

```
##      unitid       grad_100_value   grad_150_value   control_public_private
##  Min.   :100654   Min.   :  0.00   Min.   :  0.00   Min.   :1.000
##  1st Qu.:154350   1st Qu.: 16.40   1st Qu.: 35.00   1st Qu.:1.000
##  Median :191662   Median : 30.35   Median : 48.70   Median :2.000
##  Mean   :205178   Mean   : 33.93   Mean   : 48.93   Mean   :1.703
##  3rd Qu.:219347   3rd Qu.: 49.70   3rd Qu.: 63.60   3rd Qu.:2.000
##  Max.   :466921   Max.   :100.00   Max.   :100.00   Max.   :2.000
##
##    aid_value     pell_percentile  retain_percentile     ft_pct
##  Min.   : 1450   Min.   :  0.00   Min.   :  0.00    Min.   :  4.40
##  1st Qu.: 5553   1st Qu.: 23.00   1st Qu.: 25.00    1st Qu.: 73.60
##  Median : 8922   Median : 47.00   Median : 49.00    Median : 87.50
##  Mean   :11379   Mean   : 48.04   Mean   : 49.33    Mean   : 81.73
##  3rd Qu.:15554   3rd Qu.: 73.00   3rd Qu.: 74.00    3rd Qu.: 95.50
##  Max.   :41580   Max.   :100.00   Max.   :100.00    Max.   :100.00
##
##  fte_percentile   awards_per_state_value awards_per_natl_value
##  Min.   :  0.00   Min.   :11.90          Min.   :21.50
##  1st Qu.: 28.00   1st Qu.:20.90          1st Qu.:21.50
##  Median : 53.00   Median :22.30          Median :22.50
##  Mean   : 52.32   Mean   :22.19          Mean   :22.47
##  3rd Qu.: 77.00   3rd Qu.:23.70          3rd Qu.:22.50
##  Max.   :100.00   Max.   :34.20          Max.   :24.60
##
##    gender              race            grad_cohort      grad_cohort_percent
##  Length:19300       Length:19300       Min.   :   0.00   Min.   :  0.0000
##  Class :character   Class :character   1st Qu.:   1.00   1st Qu.:  0.1842
##  Mode  :character   Mode  :character   Median :   5.00   Median :  1.7409
##                                        Mean   :  65.26   Mean   : 10.0000
##                                        3rd Qu.:  33.00   3rd Qu.:  9.7998
##                                        Max.   :3126.00   Max.   :100.0000
##                                                          NA's   :30
```

```
institution_grad_details <- institution_grad_details %>% drop_na()
summary(institution_grad_details)
```

```
##      unitid       grad_100_value   grad_150_value   control_public_private
##  Min.   :100654   Min.   :  0.00   Min.   :  0.00   Min.   :1.000
##  1st Qu.:154235   1st Qu.: 16.40   1st Qu.: 35.00   1st Qu.:1.000
##  Median :191649   Median : 30.30   Median : 48.80   Median :2.000
##  Mean   :204982   Mean   : 33.95   Mean   : 48.96   Mean   :1.703
##  3rd Qu.:219295   3rd Qu.: 49.80   3rd Qu.: 63.60   3rd Qu.:2.000
##  Max.   :466921   Max.   :100.00   Max.   :100.00   Max.   :2.000
##    aid_value     pell_percentile  retain_percentile     ft_pct
##  Min.   : 1450   Min.   :  0.00   Min.   :  0.00    Min.   :  4.40
##  1st Qu.: 5553   1st Qu.: 23.00   1st Qu.: 25.00    1st Qu.: 73.60
##  Median : 8925   Median : 47.00   Median : 49.00    Median : 87.50
##  Mean   :11387   Mean   : 48.07   Mean   : 49.32    Mean   : 81.72
##  3rd Qu.:15555   3rd Qu.: 73.00   3rd Qu.: 74.00    3rd Qu.: 95.50
##  Max.   :41580   Max.   :100.00   Max.   :100.00    Max.   :100.00
##  fte_percentile  awards_per_state_value awards_per_natl_value
##  Min.   :  0.0   Min.   :11.90          Min.   :21.50
##  1st Qu.: 28.0   1st Qu.:20.90          1st Qu.:21.50
##  Median : 53.0   Median :22.30          Median :22.50
##  Mean   : 52.4   Mean   :22.18          Mean   :22.47
##  3rd Qu.: 77.0   3rd Qu.:23.70          3rd Qu.:22.50
##  Max.   :100.0   Max.   :34.20          Max.   :24.60
##     gender              race           grad_cohort      grad_cohort_percent
##  Length:19270       Length:19270       Min.   :   0.00   Min.   :  0.0000
##  Class :character   Class :character   1st Qu.:   1.00   1st Qu.:  0.1842
##  Mode  :character   Mode  :character   Median :   5.00   Median :  1.7409
##                                        Mean   :  65.36   Mean   : 10.0000
##                                        3rd Qu.:  33.00   3rd Qu.:  9.7998
##                                        Max.   :3126.00   Max.   :100.0000
```

```
stargazer(as.data.frame(institution_grad_details),header=FALSE, type = 'text')
```

```
##
## ===============================================================================
## Statistic                N       Mean      St. Dev.    Min     Max
## -------------------------------------------------------------------------------
## unitid                 19,270 204,981.700 83,628.230 100,654 466,921
## grad_100_value         19,270   33.946      22.987    0.000  100.000
## grad_150_value         19,270   48.964      21.567    0.000  100.000
## control_public_private 19,270    1.703       0.457      1       2
## aid_value              19,270 11,386.760  7,396.695   1,450  41,580
## pell_percentile        19,270   48.074      28.761      0      100
## retain_percentile      19,270   49.318      28.735      0      100
## ft_pct                 19,270   81.724      18.225    4.400  100.000
## fte_percentile         19,270   52.402      28.735      0      100
## awards_per_state_value 19,270   22.183       2.564   11.900  34.200
## awards_per_natl_value  19,270   22.466       0.922   21.500  24.600
## grad_cohort            19,270   65.362     200.542      0     3,126
## grad_cohort_percent    19,270   10.000      16.965    0.000  100.000
## -------------------------------------------------------------------------------
```

Pivot the merged dataframe wider

```
institution_grad_details_wider <- institution_grad_details %>%
 select(-grad_cohort) %>%
  pivot_wider(
    names_from = c(gender, race),
    values_from = grad_cohort_percent
  )
```

Examine the data and run summary statistics

```
summary(institution_grad_details_wider)
```

```
##      unitid       grad_100_value   grad_150_value   control_public_private
##  Min.   :100654   Min.   :  0.00   Min.   :  0.00   Min.   :1.000
##  1st Qu.:154292   1st Qu.: 16.40   1st Qu.: 35.00   1st Qu.:1.000
##  Median :191649   Median : 30.30   Median : 48.80   Median :2.000
##  Mean   :204982   Mean   : 33.95   Mean   : 48.96   Mean   :1.703
##  3rd Qu.:219286   3rd Qu.: 49.75   3rd Qu.: 63.60   3rd Qu.:2.000
##  Max.   :466921   Max.   :100.00   Max.   :100.00   Max.   :2.000
##    aid_value     pell_percentile  retain_percentile     ft_pct
##  Min.   : 1450   Min.   :  0.00   Min.   :  0.00    Min.   :  4.40
##  1st Qu.: 5554   1st Qu.: 23.00   1st Qu.: 25.00    1st Qu.: 73.60
##  Median : 8925   Median : 47.00   Median : 49.00    Median : 87.50
##  Mean   :11387   Mean   : 48.07   Mean   : 49.32    Mean   : 81.72
##  3rd Qu.:15554   3rd Qu.: 73.00   3rd Qu.: 74.00    3rd Qu.: 95.50
##  Max.   :41580   Max.   :100.00   Max.   :100.00    Max.   :100.00
##  fte_percentile  awards_per_state_value awards_per_natl_value      M_W
##  Min.   :  0.0   Min.   :11.90          Min.   :21.50         Min.   :  0.00
##  1st Qu.: 28.0   1st Qu.:20.90          1st Qu.:21.50         1st Qu.: 21.86
##  Median : 53.0   Median :22.30          Median :22.50         Median : 33.33
##  Mean   : 52.4   Mean   :22.18          Mean   :22.47         Mean   : 32.11
##  3rd Qu.: 77.0   3rd Qu.:23.70          3rd Qu.:22.50         3rd Qu.: 40.91
##  Max.   :100.0   Max.   :34.20          Max.   :24.60         Max.   :100.00
##       F_W              M_B              F_B               M_H
##  Min.   :  0.00   Min.   : 0.000   Min.   :  0.000   Min.   : 0.0000
##  1st Qu.: 25.00   1st Qu.: 1.043   1st Qu.:  1.116   1st Qu.: 0.6123
##  Median : 40.91   Median : 2.792   Median :  3.306   Median : 1.7094
##  Mean   : 37.22   Mean   : 6.360   Mean   :  8.942   Mean   : 3.7963
##  3rd Qu.: 50.16   3rd Qu.: 6.642   3rd Qu.:  8.742   3rd Qu.: 4.0209
##  Max.   :100.00   Max.   :98.865   Max.   :100.000   Max.   :66.6667
##       F_H              M_Ai              F_Ai              M_A
##  Min.   : 0.0000   Min.   :  0.0000   Min.   :  0.0000   Min.   : 0.0000
##  1st Qu.: 0.7776   1st Qu.:  0.0000   1st Qu.:  0.0000   1st Qu.: 0.0000
##  Median : 2.1739   Median :  0.0000   Median :  0.1203   Median : 0.7968
##  Mean   : 4.9909   Mean   :  0.6338   Mean   :  0.7909   Mean   : 2.2596
##  3rd Qu.: 5.3184   3rd Qu.:  0.3636   3rd Qu.:  0.4711   3rd Qu.: 2.2982
##  Max.   :71.4286   Max.   :100.0000   Max.   :100.0000   Max.   :80.0000
##       F_A
##  Min.   :  0.000
##  1st Qu.:  0.000
##  Median :  0.986
##  Mean   :  2.898
##  3rd Qu.:  2.893
##  Max.   :100.000
```

```
stargazer(as.data.frame(institution_grad_details_wider), median = TRUE, type = 'text')
```

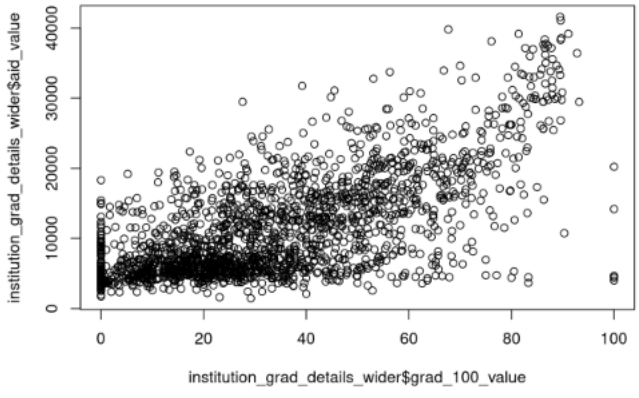
```
##
## ==========================================================================
## Statistic                 N      Mean      St. Dev.    Min   Median    Max
## --------------------------------------------------------------------------
## unitid                 1,927 204,981.700 83,647.770 100,654 191,649 466,921
## grad_100_value         1,927   33.946      22.993    0.000  30.300  100.000
## grad_150_value         1,927   48.964      21.572    0.000  48.800  100.000
## control_public_private 1,927    1.703      0.457       1       2       2
## aid_value              1,927 11,386.760  7,398.423   1,450   8,925  41,580
## pell_percentile        1,927   48.074      28.767      0      47      100
## retain_percentile      1,927   49.318      28.742      0      49      100
## ft_pct                 1,927   81.724      18.229    4.400  87.500  100.000
## fte_percentile         1,927   52.402      28.742      0      53      100
## awards_per_state_value 1,927   22.183      2.565     11.900  22.300  34.200
## awards_per_natl_value  1,927   22.466      0.922     21.500  22.500  24.600
## M_W                    1,927   32.113      19.151    0.000  33.333  100.000
## F_W                    1,927   37.217      19.763    0.000  40.909  100.000
## M_B                    1,927    6.360      10.185    0.000   2.792  98.865
## F_B                    1,927    8.942      14.926    0.000   3.306  100.000
## M_H                    1,927    3.796      6.251     0.000   1.709  66.667
## F_H                    1,927    4.991      8.105     0.000   2.174  71.429
## M_Ai                   1,927    0.634      4.217     0.000   0.000  100.000
## F_Ai                   1,927    0.791      4.489     0.000   0.120  100.000
## M_A                    1,927    2.260      4.633     0.000   0.797  80.000
## F_A                    1,927    2.898      6.383     0.000   0.986  100.000
## --------------------------------------------------------------------------
```

```
sum(is.na(institution_grad_details_wider))
```

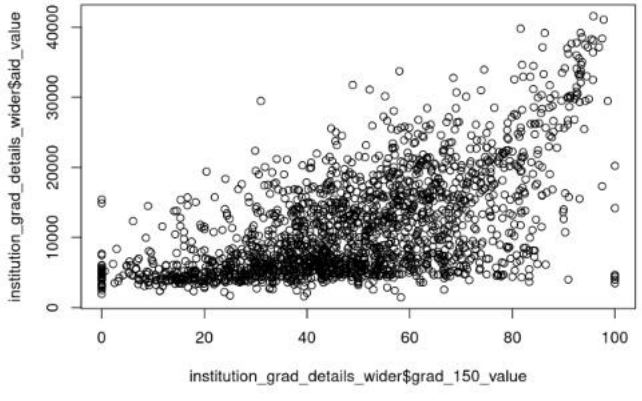
```
## [1] 0
```

Plot Residuals

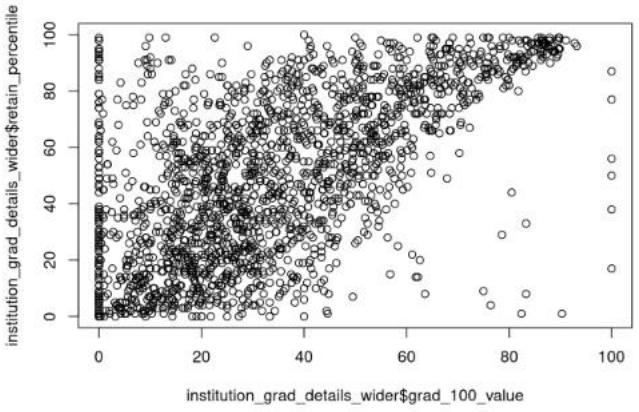
```
plot(institution_grad_details_wider$grad_100_value,institution_grad_details_wider$aid_value )
```



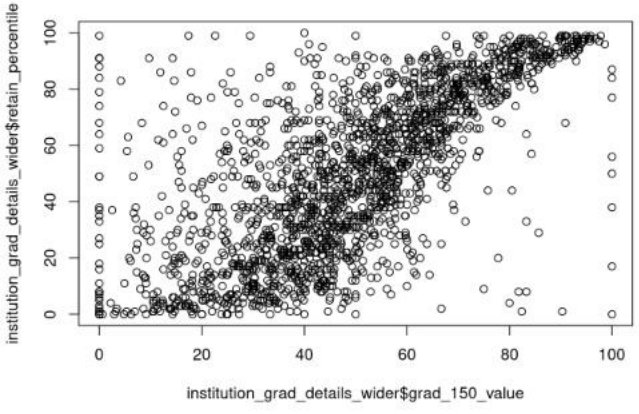

```
plot(institution_grad_details_wider$grad_150_value,institution_grad_details_wider$aid_value )
```



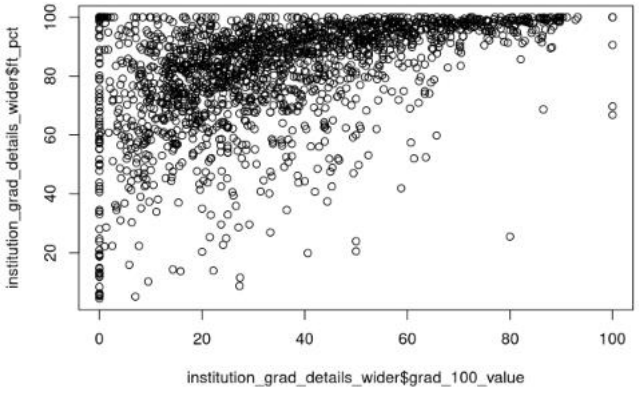
```
plot(institution_grad_details_wider$grad_100_value,institution_grad_details_wider$retain_percentile)
```



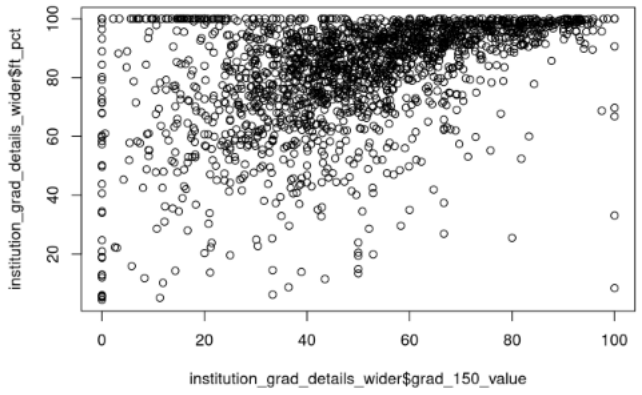

```
plot(institution_grad_details_wider$grad_150_value,institution_grad_details_wider$retain_percentile)
```



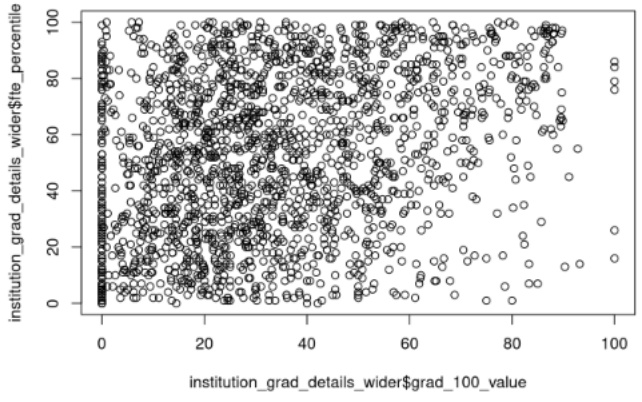

```
plot(institution_grad_details_wider$grad_100_value,institution_grad_details_wider$ft_pct)
```



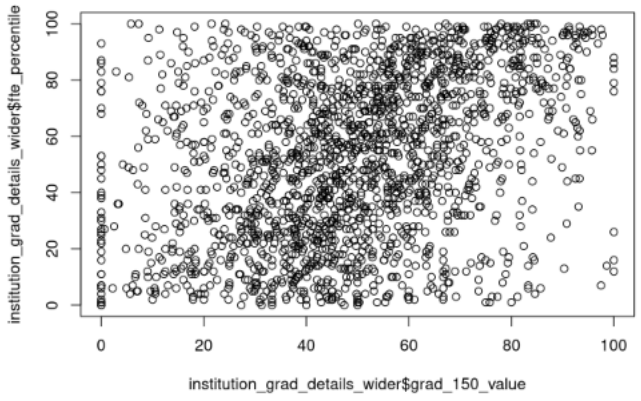
```
plot(institution_grad_details_wider$grad_150_value,institution_grad_details_wider$ft_pct)
```



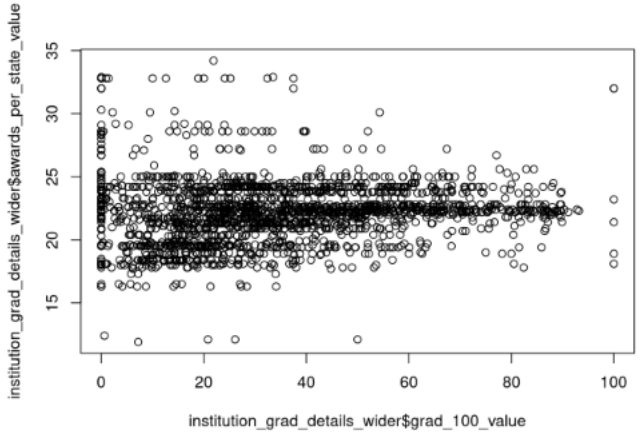

```
plot(institution_grad_details_wider$grad_100_value,institution_grad_details_wider$fte_percentile)
```



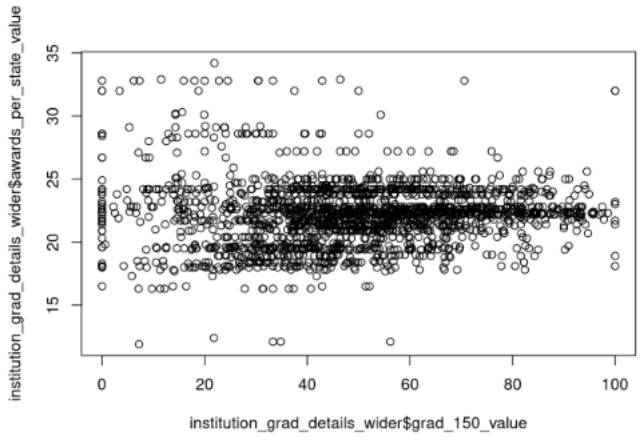
```
plot(institution_grad_details_wider$grad_150_value,institution_grad_details_wider$fte_percentile)
```



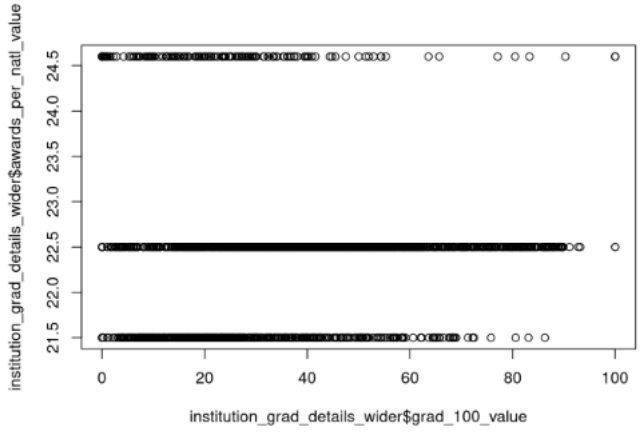

```
plot(institution_grad_details_wider$grad_100_value,institution_grad_details_wider$awards_per_state_value)
```



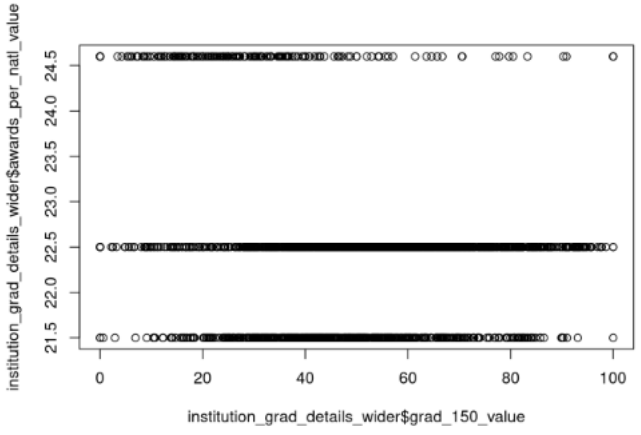
```
plot(institution_grad_details_wider$grad_150_value,institution_grad_details_wider$awards_per_state_value)
```




```
plot(institution_grad_details_wider$grad_100_value,institution_grad_details_wider$awards_per_natl_value)
```

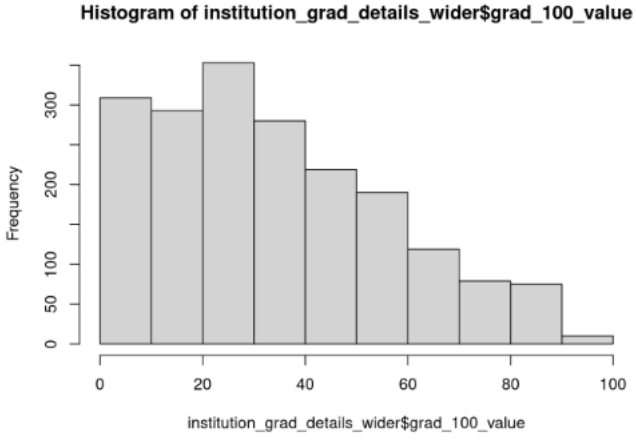


```
plot(institution_grad_details_wider$grad_150_value,institution_grad_details_wider$awards_per_natl_value)
```

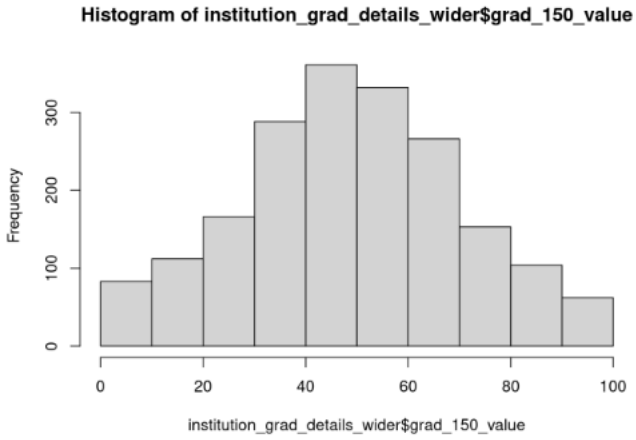


Create Histogram for grad_100_value and grad_150_value and all other variables

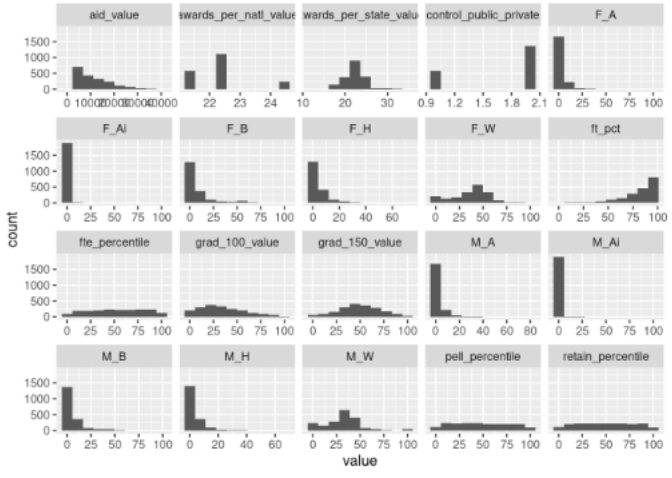
```
hist(institution_grad_details_wider$grad_100_value)
```



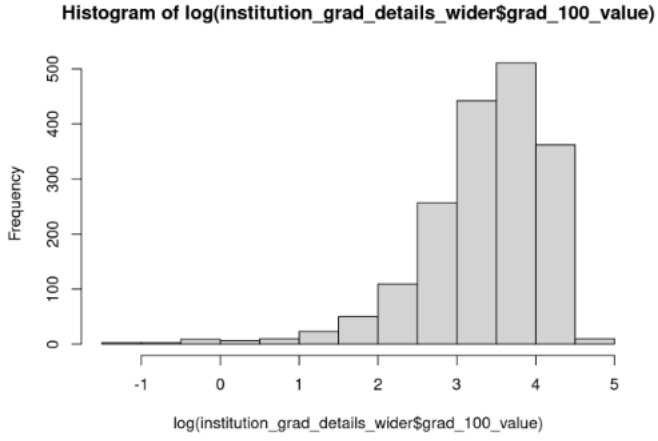
```
hist(institution_grad_details_wider$grad_150_value)
```



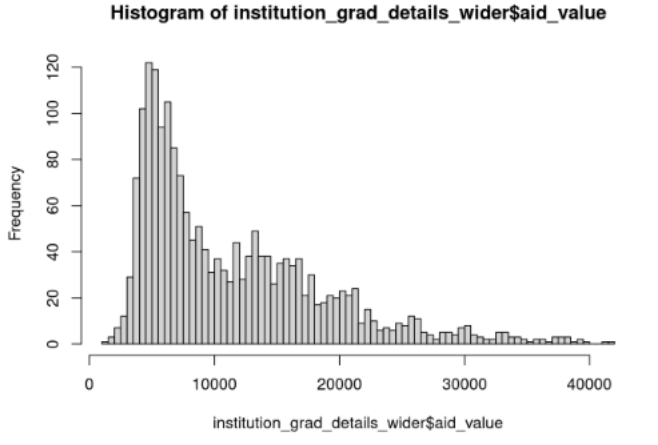

```
ggplot(gather(institution_grad_details_wider[,-1]), aes(value)) +
  geom_histogram(bins = 10) +
  facet_wrap(~key, scales = "free_x")
```



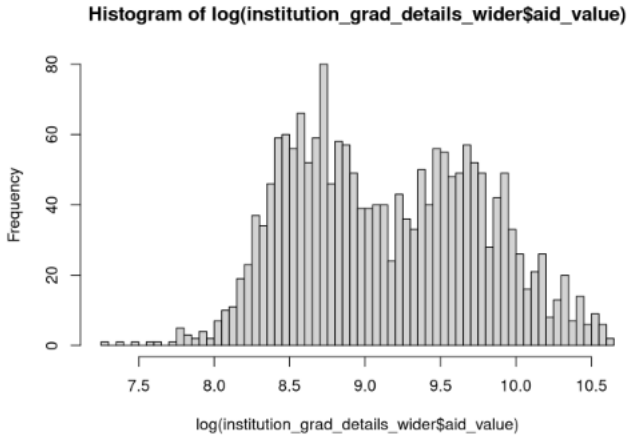
```
hist(log(institution_grad_details_wider$grad_100_value))
```



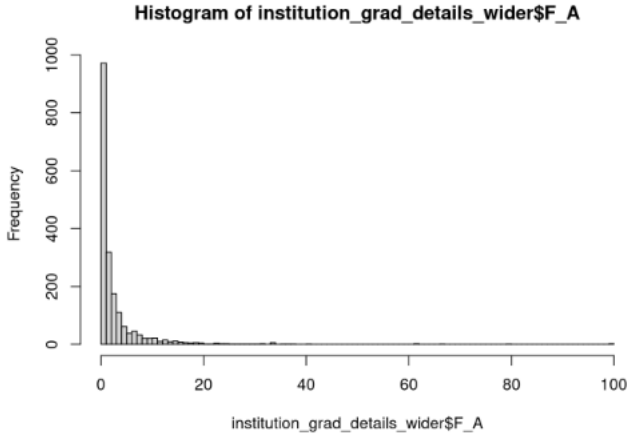

```
hist(institution_grad_details_wider$aid_value, breaks = 100)
```



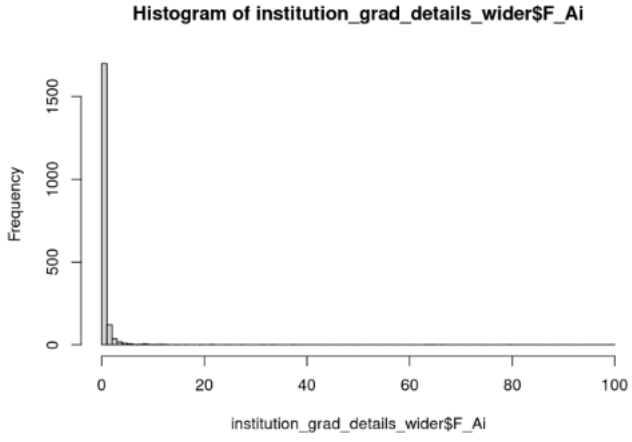
```
hist(log(institution_grad_details_wider$aid_value), breaks = 100)
```



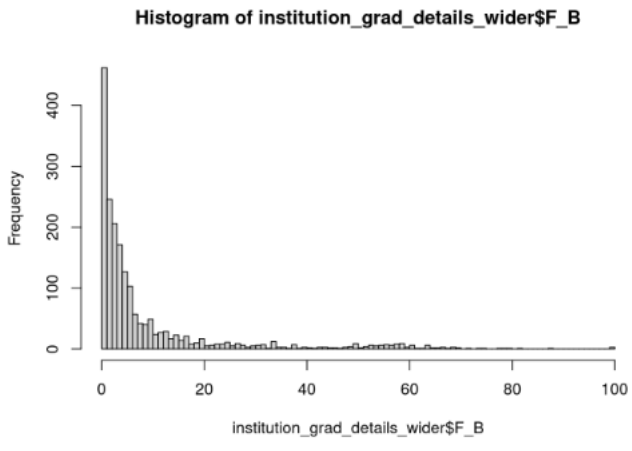

```
hist(institution_grad_details_wider$F_A, breaks = 100)
```



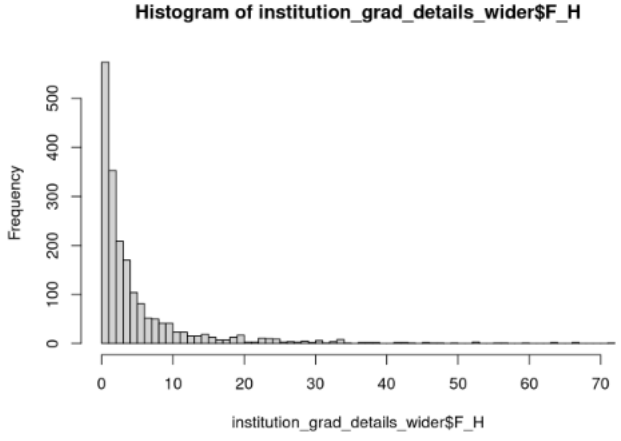
```
hist(institution_grad_details_wider$F_Ai, breaks = 100)
```



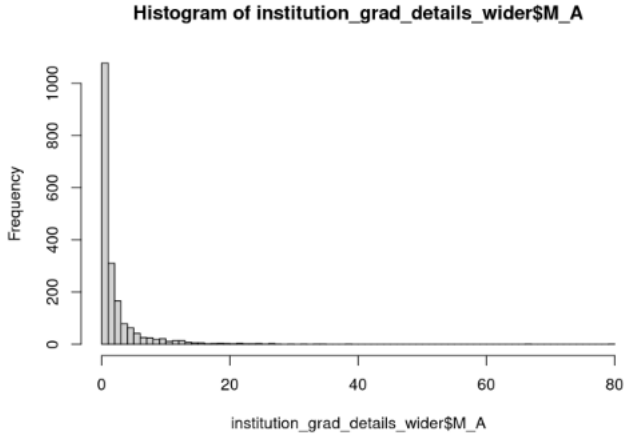

```
hist(institution_grad_details_wider$F_B, breaks = 100)
```



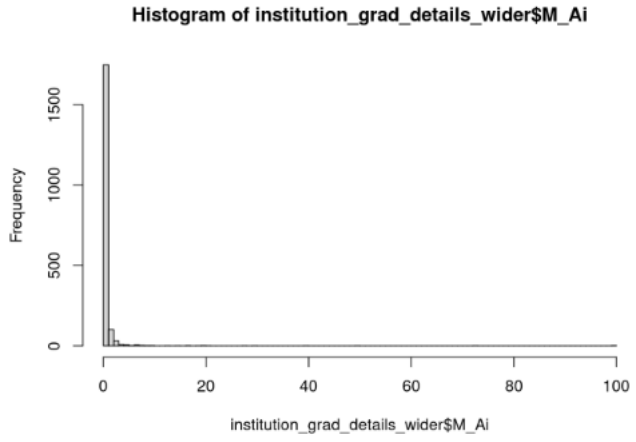
```
hist(institution_grad_details_wider$F_H, breaks = 100)
```



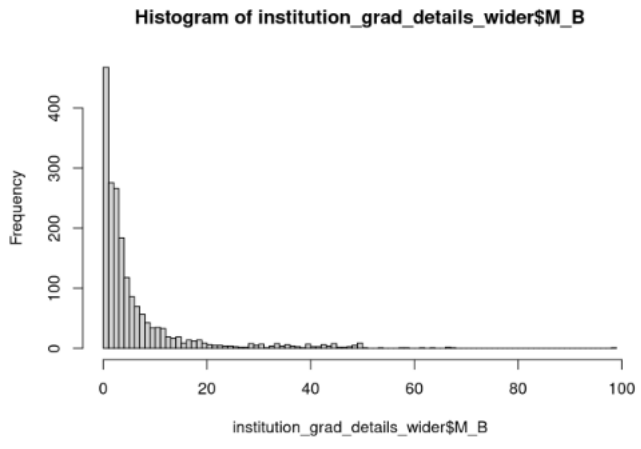

```
hist(institution_grad_details_wider$M_A, breaks = 100)
```



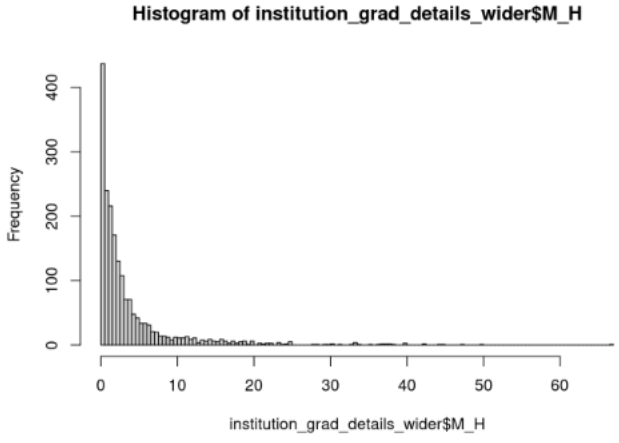
```
hist(institution_grad_details_wider$M_Ai, breaks = 100)
```




```
hist(institution_grad_details_wider$M_B, breaks = 100)
```



```
hist(institution_grad_details_wider$M_H, breaks = 100)
```



The distribution for the four year graduation is slightly right-skewed. It also has an outlier where a small number of schools the four year graduation rate is more than 80%. A log transformation here might be helpful. Upon log transformation we see that the variable becomes left skewed and therefore it will not be useful.

The distribution for the six year graduation is normal.

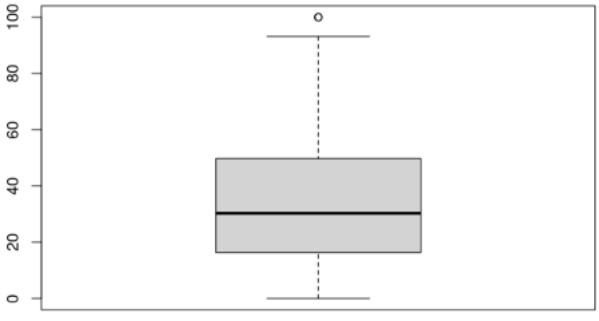
The F_A, F_Ai, F_B, M_A, M_Ai, M_B, F_H, M_H are zero inflated distributions and a log transformation will not help here.

aid_value is not zero inflated distribution and therefore may benefit from log transformation. Upon log transformation we see that the variable is now more normally distributed than before.

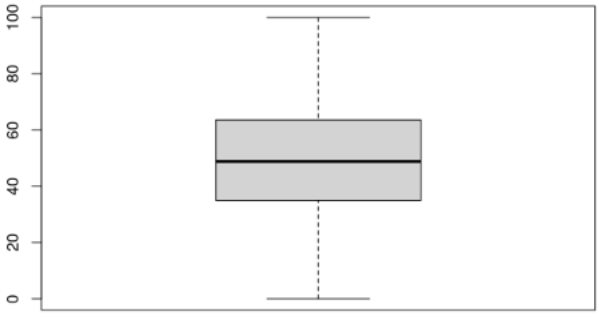
aid_value will be log transformed in the analysis

Boxplot for grad_100_value and grad_150_value

```
boxplot(institution_grad_details_wider$grad_100_value)
```




```
boxplot(institution_grad_details_wider$grad_150_value)
```



The boxplot for the four year graduation rate has a couple of outliers.
The boxplot for the six year graduation rate is normally distributed and contains no outliers.

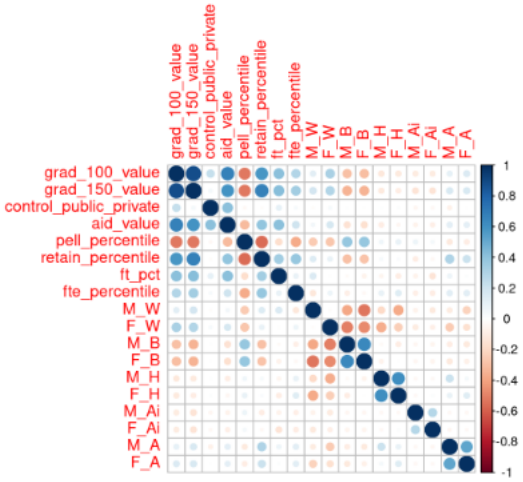
Correlation Matrix

```
#glimpse(institution_grad_details_wider)
cor_matrix <- cor(institution_grad_details_wider[,c(-1,-10,-11)])
cor_matrix
```

```
##                          grad_100_value grad_150_value control_public_private
## grad_100_value              1.0000000     0.88273624           0.2373047062
## grad_150_value              0.8827362     1.00000000           0.0476991331
## control_public_private      0.2373047     0.04769913           1.0000000000
## aid_value                   0.6746068     0.59211138           0.4106587532
## pell_percentile            -0.5275745    -0.52871464          -0.0307820151
## retain_percentile           0.5899294     0.67732477          -0.0176223443
## ft_pct                      0.4195159     0.42553433           0.1085011278
## fte_percentile              0.2955668     0.34127221           0.0341035403
## M_W                         0.1243815     0.18770028          -0.0167680499
## F_W                         0.3225823     0.29967543           0.0397406058
## M_B                        -0.2995270    -0.34297664           0.0074277001
## F_B                        -0.3047189    -0.34223093          -0.0027884306
## M_H                        -0.1220086    -0.11449203           0.0067321877
## F_H                        -0.1170274    -0.09506878           0.0158729202
## M_Ai                       -0.1073811    -0.14390078          -0.0456268416
## F_Ai                       -0.1067137    -0.14551433          -0.0345533803
## M_A                         0.1095542     0.16629499          -0.0704032028
## F_A                         0.1517371     0.16422527           0.0002985701
##                           aid_value pell_percentile retain_percentile
## grad_100_value          0.67460676     -0.52757450        0.58992938
## grad_150_value          0.59211138     -0.52871464        0.67732477
## control_public_private  0.41065875     -0.03078202       -0.01762234
## aid_value               1.00000000     -0.31605450        0.39045307
## pell_percentile        -0.31605450      1.00000000       -0.56133840
## retain_percentile       0.39045307     -0.56133840        1.00000000
## ft_pct                  0.41851782     -0.17557180        0.34100989
## fte_percentile          0.20110119     -0.37451476        0.38538925
## M_W                     0.05108207     -0.25281887        0.14124815
## F_W                     0.17133545     -0.27589405        0.09334587
## M_B                    -0.15508259      0.38623798       -0.29037054
## F_B                    -0.16635781      0.38979066       -0.28370903
## M_H                    -0.09534813      0.06160169        0.03875947
## F_H                    -0.04352602      0.09689378        0.02100170
## M_Ai                   -0.06369413      0.07564964       -0.09409682
## F_Ai                   -0.08004472      0.04623996       -0.03687624
## M_A                     0.11990256     -0.11302844        0.30016844
## F_A                     0.11169414     -0.09824533        0.21906349
```

```
##                              ft_pct fte_percentile         M_W         F_W
## grad_100_value          0.419515914     0.29556683  0.12438150  0.322582264
## grad_150_value          0.425534326     0.34127221  0.18770028  0.299675432
## control_public_private  0.108501128     0.03410354 -0.01676805  0.039740606
## aid_value               0.418517821     0.20110119  0.05108207  0.171335452
## pell_percentile        -0.175571805    -0.37451476 -0.25281887 -0.275894050
## retain_percentile       0.341009890     0.38538925  0.14124815  0.093345874
## ft_pct                  1.000000000     0.09062153  0.15729217 -0.009044494
## fte_percentile          0.090621527     1.00000000 -0.12641337  0.074079140
## M_W                     0.157292173    -0.12641337  1.00000000 -0.014485149
## F_W                    -0.009044494     0.07407914 -0.01448515  1.000000000
## M_B                    -0.023480227    -0.08497508 -0.38201798 -0.504481580
## F_B                    -0.069399182    -0.03439220 -0.53616794 -0.473187180
## M_H                    -0.052862351     0.09727191 -0.19429015 -0.364614513
## F_H                    -0.095053461     0.13358744 -0.37935255 -0.247828540
## M_Ai                   -0.066242543    -0.08048124 -0.06804416 -0.106802823
## F_Ai                   -0.140962121    -0.09621520 -0.11151329 -0.080077419
## M_A                     0.076589182     0.12826474 -0.10416026 -0.251237967
## F_A                     0.013818051     0.12756444 -0.22262202 -0.161264162
##                              M_B           F_B         M_H           F_H
## grad_100_value         -0.29952704 -0.3047188643 -0.122008621 -0.1170273570
## grad_150_value         -0.34297664 -0.3422309264 -0.114492028 -0.0950687804
## control_public_private  0.00742770 -0.0027884306  0.006732188  0.0158729202
## aid_value              -0.15508259 -0.1663578105 -0.095348131 -0.0435260150
## pell_percentile         0.38623798  0.3897906648  0.061601691  0.0968937777
## retain_percentile      -0.29037054 -0.2837090251  0.038759474  0.0210016953
## ft_pct                 -0.02348023 -0.0693991819 -0.052862351 -0.0950534614
## fte_percentile         -0.08497508 -0.0343922032  0.097271908  0.1335874432
## M_W                    -0.38201798 -0.5361679446 -0.194290147 -0.3793525502
## F_W                    -0.50448158 -0.4731871801 -0.364614513 -0.2478285404
## M_B                     1.00000000  0.6491459649 -0.028750737 -0.0953930540
## F_B                     0.64914596  1.0000000000 -0.065317801 -0.0001671512
## M_H                    -0.02875074 -0.0653178014  1.000000000  0.6111035890
## F_H                    -0.09539305 -0.0001671512  0.611103589  1.0000000000
## M_Ai                   -0.04411329 -0.0474529184 -0.033940371 -0.0329556499
## F_Ai                   -0.05638101 -0.0585048922 -0.031599837  0.0185469977
## M_A                    -0.09886869 -0.1166374525  0.217030852  0.0918988508
## F_A                    -0.11628831 -0.0798176671  0.041857250  0.1321102885
##                              M_Ai        F_Ai         M_A           F_A
## grad_100_value         -0.10738107 -0.10671371  0.10955420  0.1517370909
## grad_150_value         -0.14390078 -0.14551433  0.16629499  0.1642252695
## control_public_private -0.04562684 -0.03455338 -0.07040320  0.0002985701
## aid_value              -0.06369413 -0.08004472  0.11990256  0.1116941435
## pell_percentile         0.07564964  0.04623996 -0.11302844 -0.0982453269
## retain_percentile      -0.09409682 -0.03687624  0.30016844  0.2190634895
## ft_pct                 -0.06624254 -0.14096212  0.07658918  0.0138180512
## fte_percentile         -0.08048124 -0.09621520  0.12826474  0.1275644374
## M_W                    -0.06804416 -0.11151329 -0.10416026 -0.2226220212
## F_W                    -0.10680282 -0.08007742 -0.25123797 -0.1612641623
## M_B                    -0.04411329 -0.05638101 -0.09886869 -0.1162883089
## F_B                    -0.04745292 -0.05850489 -0.11663745 -0.0798176671
## M_H                    -0.03394037 -0.03159984  0.21703085  0.0418572503
## F_H                    -0.03295565  0.01854700  0.09189885  0.1321102885
## M_Ai                    1.00000000  0.28459639 -0.03661402 -0.0393627779
## F_Ai                    0.28459639  1.00000000 -0.04378228 -0.0302425352
## M_A                    -0.03661402 -0.04378228  1.00000000  0.5202239211
## F_A                    -0.03936278 -0.03024254  0.52022392  1.0000000000
```

```
corrplot(cor(institution_grad_details_wider[,c(-1,-10,-11)]))
```



Run the first multiple linear regression model for 4-year graduation with all variables

```
# Log transform the variables
institution_grad_details_wider$ln_aid_value <-log(institution_grad_details_wider$aid_value)

model_1 <- lm(grad_100_value~control_public_private+ln_aid_value+pell_percentile+retain_percentile+ft_pct+fte_percentile+M_W
+F_W+M_B+F_B+M_H+F_H+M_Ai+F_Ai+M_A, data = institution_grad_details_wider)

alias(model_1)
```

```
## Model :
## grad_100_value ~ control_public_private + ln_aid_value + pell_percentile +
##     retain_percentile + ft_pct + fte_percentile + M_W + F_W +
##     M_B + F_B + M_H + F_H + M_Ai + F_Ai + M_A
```

```
summary(model_1)
```

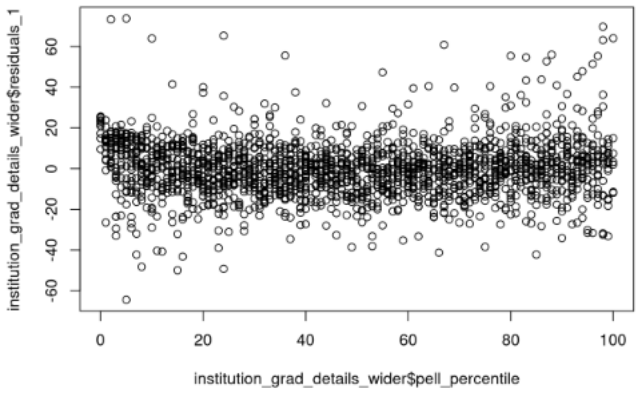
```
##
## Call:
## lm(formula = grad_100_value ~ control_public_private + ln_aid_value +
##     pell_percentile + retain_percentile + ft_pct + fte_percentile +
##     M_W + F_W + M_B + F_B + M_H + F_H + M_Ai + F_Ai + M_A, data = institution_grad_details_wider)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.502  -7.971  -0.180   7.345  73.792
##
## Coefficients:
##                         Estimate Std. Error t value             Pr(>|t|)
## (Intercept)           -86.46635    7.75868 -11.144 < 0.0000000000000002 ***
## control_public_private   3.09341    0.77180   4.008       0.000063571973 ***
## ln_aid_value            14.10899    0.66782  21.127 < 0.0000000000000002 ***
## pell_percentile         -0.15594    0.01469 -10.612 < 0.0000000000000002 ***
## retain_percentile        0.22463    0.01526  14.717 < 0.0000000000000002 ***
## ft_pct                   0.13202    0.02044   6.458       0.000000000134 ***
## fte_percentile           0.01055    0.01268   0.832              0.40557
## M_W                     -0.36371    0.05995  -6.067       0.000000001564 ***
## F_W                     -0.16584    0.06128  -2.706              0.00686 **
## M_B                     -0.36202    0.07200  -5.028       0.000000541636 ***
## F_B                     -0.34869    0.06604  -5.280       0.000000143935 ***
## M_H                     -0.24267    0.08265  -2.936              0.00336 **
## F_H                     -0.53750    0.08510  -6.316       0.000000000333 ***
## M_Ai                    -0.40356    0.09782  -4.125       0.000038595133 ***
## F_Ai                    -0.42786    0.09511  -4.499       0.000007243881 ***
## M_A                     -0.59607    0.12641  -4.715       0.000002588797 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.77 on 1911 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6413
## F-statistic: 230.5 on 15 and 1911 DF,  p-value: < 0.00000000000000022
```

```
#stargazer(model_1, title="Results", align=TRUE,type = "text")
tab_model(model_1)
```

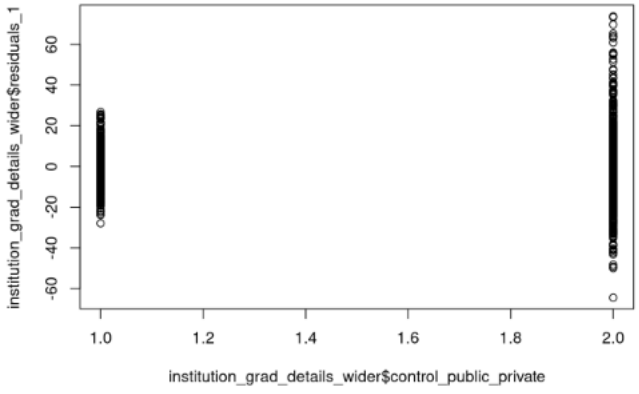
	grad_100_value		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-86.47	-101.68 – -71.25	**<0.001**
control public private	3.09	1.58 – 4.61	**<0.001**
ln aid value	14.11	12.80 – 15.42	**<0.001**
pell percentile	-0.16	-0.18 – -0.13	**<0.001**
retain percentile	0.22	0.19 – 0.25	**<0.001**
ft pct	0.13	0.09 – 0.17	**<0.001**
fte percentile	0.01	-0.01 – 0.04	0.406
M W	-0.36	-0.48 – -0.25	**<0.001**
F W	-0.17	-0.29 – -0.05	**0.007**
M B	-0.36	-0.50 – -0.22	**<0.001**
F B	-0.35	-0.48 – -0.22	**<0.001**
M H	-0.24	-0.40 – -0.08	**0.003**
F H	-0.54	-0.70 – -0.37	**<0.001**
M Ai	-0.40	-0.60 – -0.21	**<0.001**
F Ai	-0.43	-0.61 – -0.24	**<0.001**
M A	-0.60	-0.84 – -0.35	**<0.001**
Observations	1927		
R^2 / R^2 adjusted	0.644 / 0.641		

Checking the regression assumptions for Model 1 and create histogram

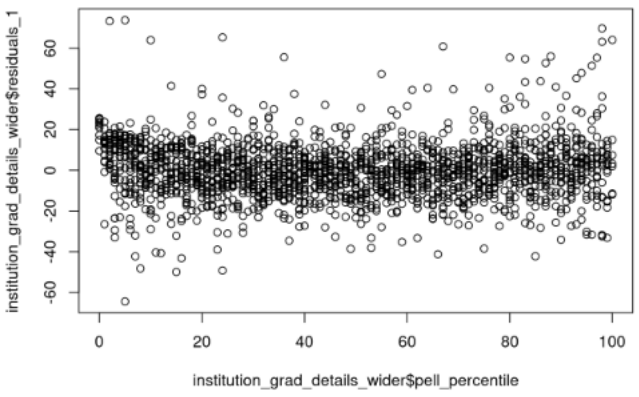
```
plot(institution_grad_details_wider$pell_percentile,institution_grad_details_wider$residuals_1)
```



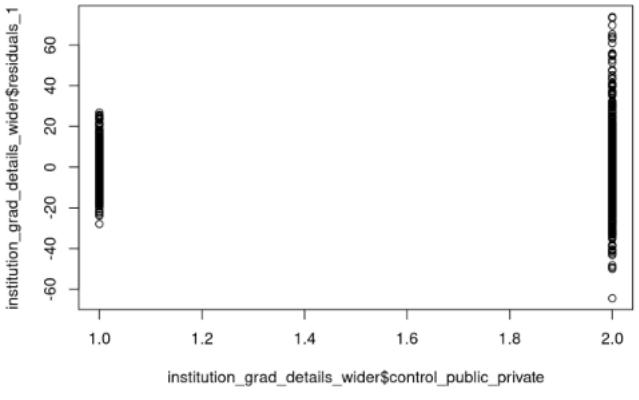

```
plot(institution_grad_details_wider$control_public_private,institution_grad_details_wider$residuals_1)
```



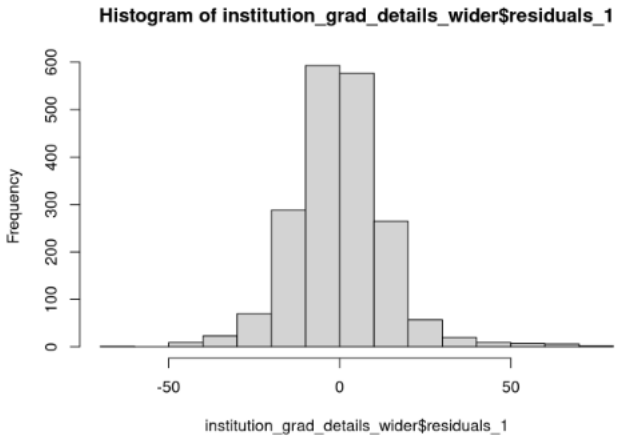
```
plot(institution_grad_details_wider$pell_percentile,institution_grad_details_wider$residuals_1)
```



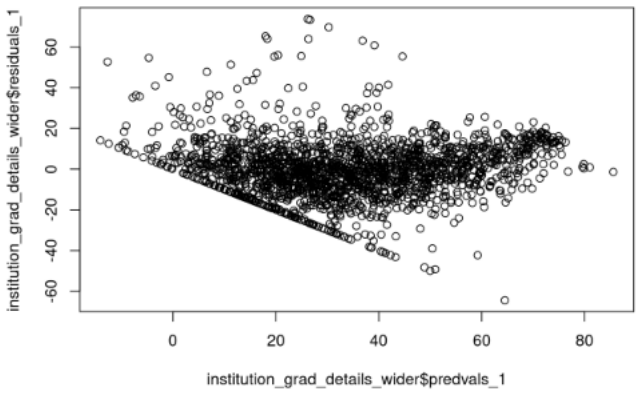

```
plot(institution_grad_details_wider$control_public_private,institution_grad_details_wider$residuals_1)
```



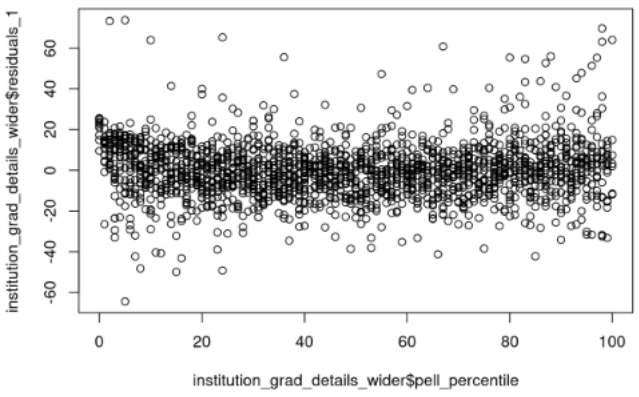
```
institution_grad_details_wider$residuals_1 <- residuals(model_1)
hist(institution_grad_details_wider$residuals_1)
```



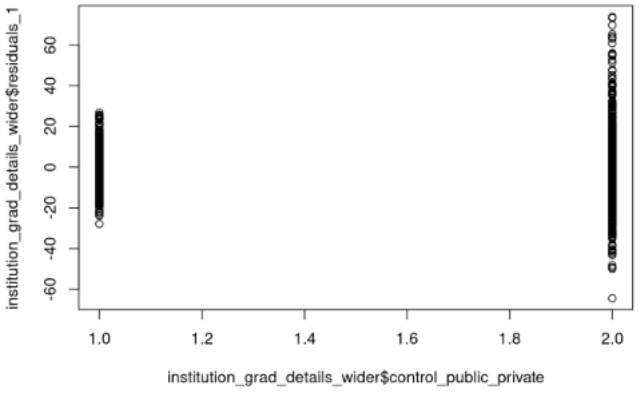

```
institution_grad_details_wider$predvals_1<-fitted(model_1)
plot(institution_grad_details_wider$predvals_1,institution_grad_details_wider$residuals_1)
```



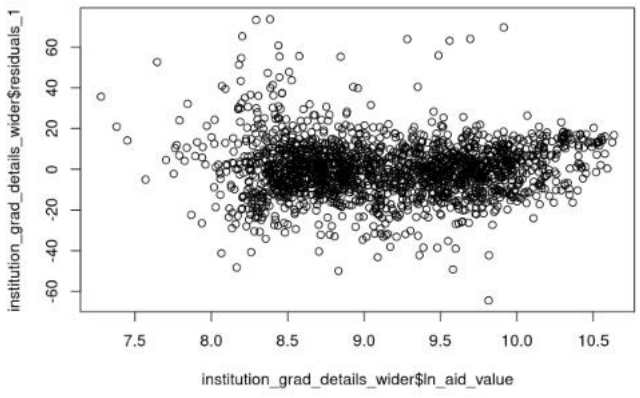
```
plot(institution_grad_details_wider$pell_percentile,institution_grad_details_wider$residuals_1)
```



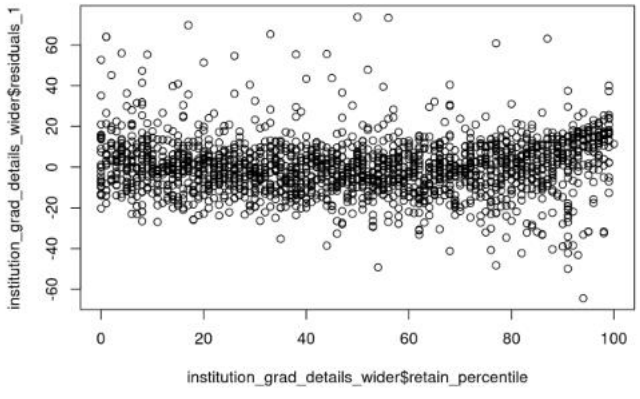

```
plot(institution_grad_details_wider$control_public_private,institution_grad_details_wider$residuals_1)
```



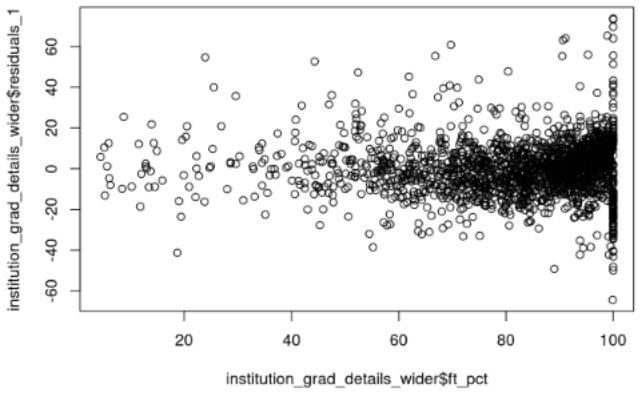
```
plot(institution_grad_details_wider$ln_aid_value,institution_grad_details_wider$residuals_1)
```



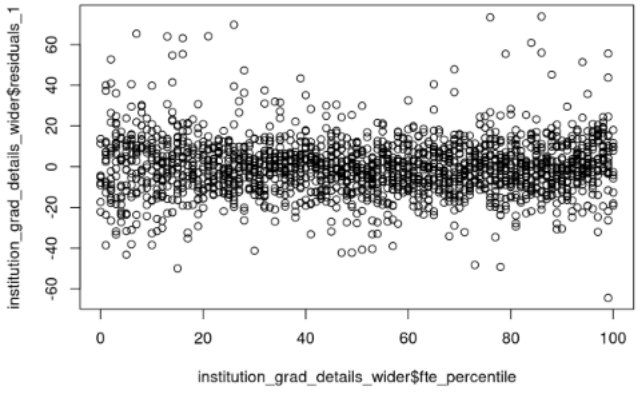

```
plot(institution_grad_details_wider$retain_percentile,institution_grad_details_wider$residuals_1)
```



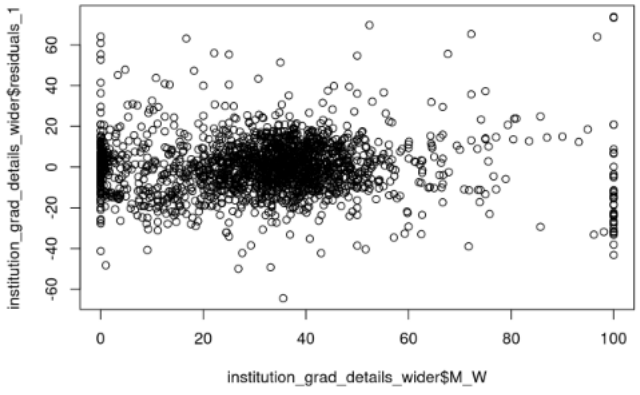
```
plot(institution_grad_details_wider$ft_pct,institution_grad_details_wider$residuals_1)
```



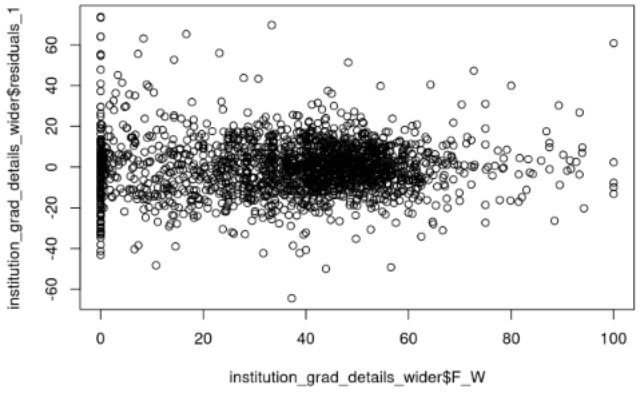

```
plot(institution_grad_details_wider$fte_percentile,institution_grad_details_wider$residuals_1)
```



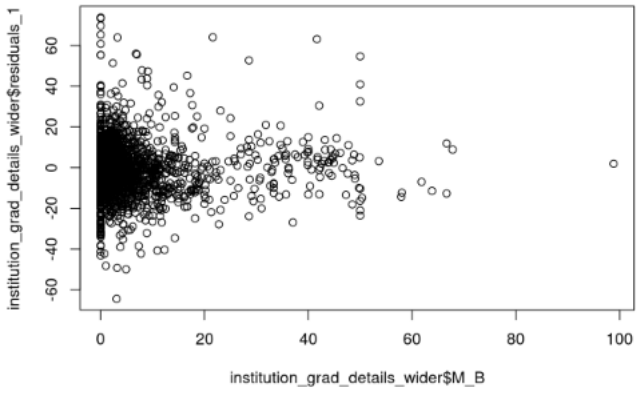
```
plot(institution_grad_details_wider$M_W,institution_grad_details_wider$residuals_1)
```



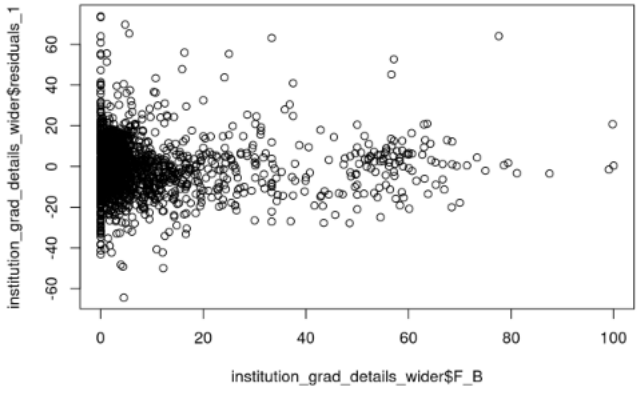

```
plot(institution_grad_details_wider$F_W,institution_grad_details_wider$residuals_1)
```



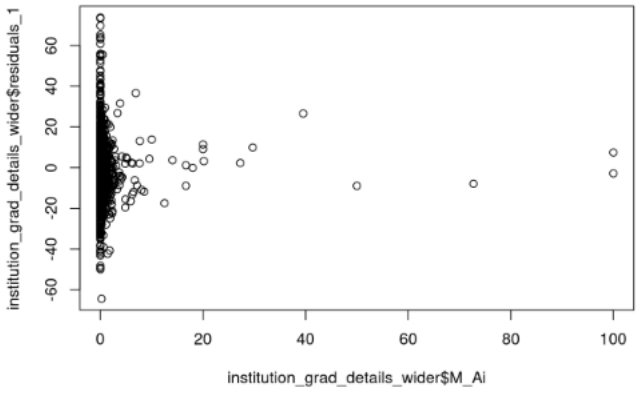
```
plot(institution_grad_details_wider$M_B,institution_grad_details_wider$residuals_1)
```



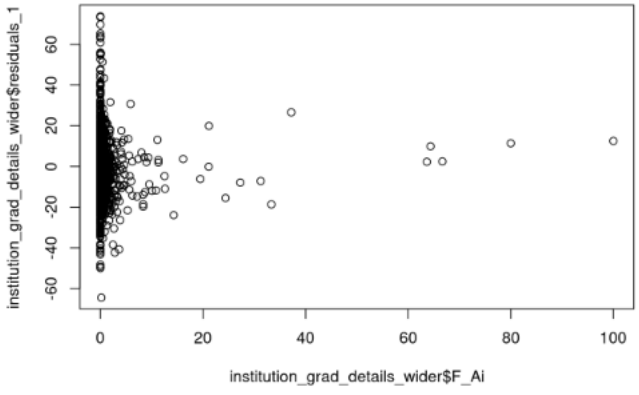

```
plot(institution_grad_details_wider$F_B,institution_grad_details_wider$residuals_1)
```



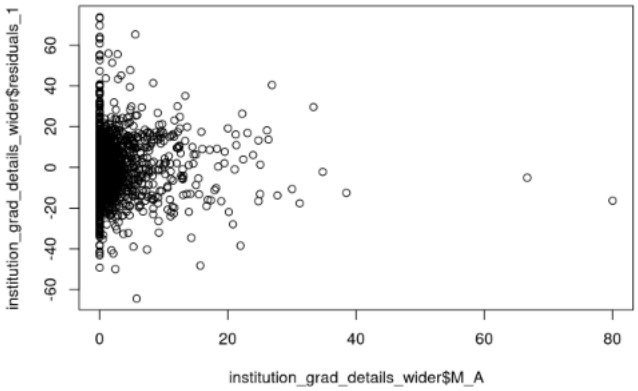
```
plot(institution_grad_details_wider$M_Ai,institution_grad_details_wider$residuals_1)
```



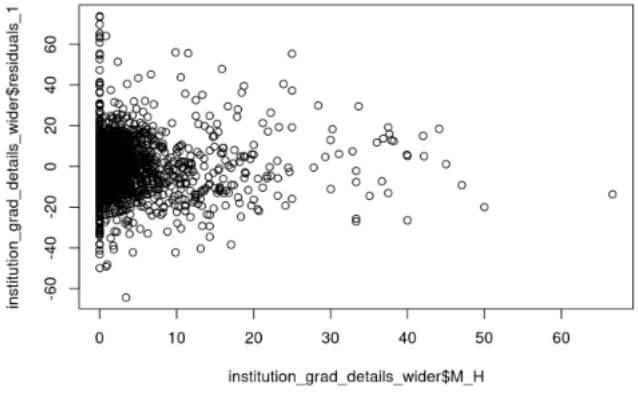

```
plot(institution_grad_details_wider$F_Ai,institution_grad_details_wider$residuals_1)
```



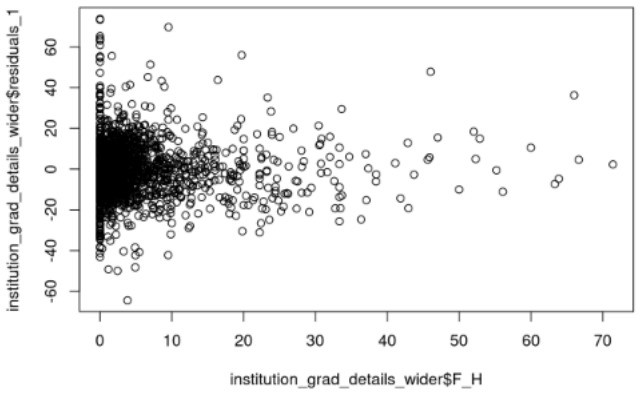
```
plot(institution_grad_details_wider$M_A,institution_grad_details_wider$residuals_1)
```




```
plot(institution_grad_details_wider$M_H,institution_grad_details_wider$residuals_1)
```



```
plot(institution_grad_details_wider$F_H,institution_grad_details_wider$residuals_1)
```




```
vif(model_1)
```

```
## control_public_private          ln_aid_value       pell_percentile
##               1.264601              1.756978              1.814847
##      retain_percentile                ft_pct        fte_percentile
##               1.954589              1.410043              1.349353
##                    M_W                   F_W                   M_B
##              13.385420             14.893361              5.460959
##                    F_B                   M_H                   F_H
##               9.867783              2.711186              4.831660
##                   M_Ai                  F_Ai                   M_A
##               1.728153              1.785355              3.483914
```

```
stargazer(vif(model_1), title="VIF", type = "text")
```

```
##
## VIF
## ============================================================================================================================================
========================
## control_public_private ln_aid_value pell_percentile retain_percentile ft_pct fte_percentile  M_W    F_W    M_B   F_B   M_
H   F_H  M_Ai  F_Ai   M_A
## --------------------------------------------------------------------------------------------------------------------------------------------
-------------------------
## 1.265                  1.757          1.815           1.955       1.410      1.349      13.385 14.893 5.461 9.868 2.7
11 4.832 1.728 1.785 3.484
## --------------------------------------------------------------------------------------------------------------------------------------------
-------------------------
```

Checking to see if a quadratic regression model will improve model fit and will fix heteroskedasticity

```
model_1.1 <- lm(grad_100_value~control_public_private+ln_aid_value+pell_percentile+pell_percentile^2+retain_percentile+retai
n_percentile^2+ft_pct+ft_pct^2+fte_percentile+fte_percentile^2+M_W+F_W+M_B+F_B+M_H+F_H+M_Ai+F_Ai+M_A, data = institution_gra
d_details_wider)

summary(model_1.1)
```

```
##
## Call:
## lm(formula = grad_100_value ~ control_public_private + ln_aid_value +
##     pell_percentile + pell_percentile^2 + retain_percentile +
##     retain_percentile^2 + ft_pct + ft_pct^2 + fte_percentile +
##     fte_percentile^2 + M_W + F_W + M_B + F_B + M_H + F_H + M_Ai +
##     F_Ai + M_A, data = institution_grad_details_wider)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.502  -7.971  -0.180   7.345  73.792
##
## Coefficients:
##                         Estimate Std. Error t value             Pr(>|t|)
## (Intercept)            -86.46635    7.75868 -11.144 < 0.0000000000000002 ***
## control_public_private   3.09341    0.77180   4.008       0.000063571973 ***
## ln_aid_value            14.10899    0.66782  21.127 < 0.0000000000000002 ***
## pell_percentile         -0.15594    0.01469 -10.612 < 0.0000000000000002 ***
## retain_percentile        0.22463    0.01526  14.717 < 0.0000000000000002 ***
## ft_pct                   0.13202    0.02044   6.458       0.000000000134 ***
## fte_percentile           0.01055    0.01268   0.832              0.40557
## M_W                     -0.36371    0.05995  -6.067       0.000000001564 ***
## F_W                     -0.16584    0.06128  -2.706              0.00686 **
## M_B                     -0.36202    0.07200  -5.028       0.000000541636 ***
## F_B                     -0.34869    0.06604  -5.280       0.000000143935 ***
## M_H                     -0.24267    0.08265  -2.936              0.00336 **
## F_H                     -0.53750    0.08510  -6.316       0.000000000333 ***
## M_Ai                    -0.40356    0.09782  -4.125       0.000038595133 ***
## F_Ai                    -0.42786    0.09511  -4.499       0.000007243881 ***
## M_A                     -0.59607    0.12641  -4.715       0.000002588797 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.77 on 1911 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6413
## F-statistic: 230.5 on 15 and 1911 DF,  p-value: < 0.00000000000000022
```

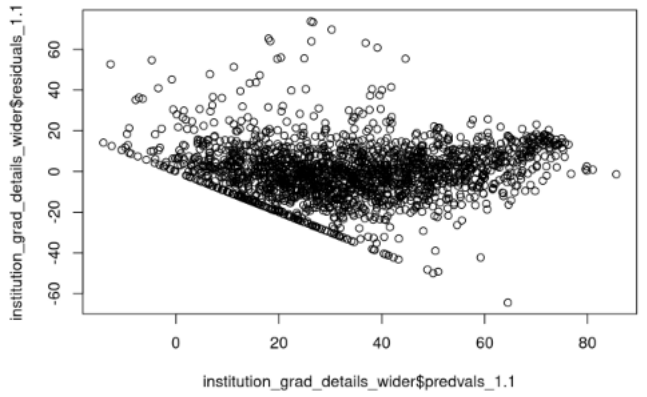
```
#stargazer(model_1.1, title="Results", align=TRUE,type = "text")
tab_model(model_1.1)
```

	grad_100_value		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-86.47	-101.68 – -71.25	**<0.001**
control public private	3.09	1.58 – 4.61	**<0.001**
ln aid value	14.11	12.80 – 15.42	**<0.001**
pell percentile	-0.16	-0.18 – -0.13	**<0.001**
retain percentile	0.22	0.19 – 0.25	**<0.001**
ft pct	0.13	0.09 – 0.17	**<0.001**
fte percentile	0.01	-0.01 – 0.04	0.406
M W	-0.36	-0.48 – -0.25	**<0.001**
F W	-0.17	-0.29 – -0.05	**0.007**
M B	-0.36	-0.50 – -0.22	**<0.001**
F B	-0.35	-0.48 – -0.22	**<0.001**
M H	-0.24	-0.40 – -0.08	**0.003**
F H	-0.54	-0.70 – -0.37	**<0.001**
M Ai	-0.40	-0.60 – -0.21	**<0.001**
F Ai	-0.43	-0.61 – -0.24	**<0.001**
M A	-0.60	-0.84 – -0.35	**<0.001**
Observations	1927		
R^2 / R^2 adjusted	0.644 / 0.641		

Checking the regression assumptions for Model 1.1 and create histogram

```
institution_grad_details_wider$residuals_1.1 <- residuals(model_1.1)
hist(institution_grad_details_wider$residuals_1.1)
```


```
institution_grad_details_wider$predvals_1.1<-fitted(model_1.1)
plot(institution_grad_details_wider$predvals_1.1,institution_grad_details_wider$residuals_1.1)
```




```
vif(model_1.1)
```

```
## control_public_private          ln_aid_value       pell_percentile
##               1.264601              1.756978              1.814847
##      retain_percentile                ft_pct        fte_percentile
##               1.954589              1.410043              1.349353
##                    M_W                   F_W                   M_B
##              13.385420             14.893361              5.460959
##                    F_B                   M_H                   F_H
##               9.867783              2.711186              4.831660
##                   M_Ai                  F_Ai                   M_A
##               1.728153              1.785355              3.483914
```

```
stargazer(vif(model_1.1), title="VIF", type = "text")
```

```
##
## VIF
## ============================================================================================================================
=========================
## control_public_private ln_aid_value pell_percentile retain_percentile ft_pct fte_percentile  M_W    F_W    M_B   F_B   M_
H   F_H  M_Ai  F_Ai   M_A
## ----------------------------------------------------------------------------------------------------------------------------
-------------------------
## 1.265                  1.757          1.815           1.955       1.410      1.349      13.385 14.893 5.461 9.868 2.7
11 4.832 1.728 1.785 3.484
## ----------------------------------------------------------------------------------------------------------------------------
-------------------------
```

Run the second multiple linear regression model by dropping ln_aid_value, awards_per_natl_value, F_B, M_B, M_Ai, F_Ai, M_A,F_H due to multicolinearity

```
model_2 <- lm(grad_100_value~control_public_private+pell_percentile+retain_percentile+ft_pct+fte_percentile+F_W +F_A, data =
institution_grad_details_wider)
summary(model_2)
```

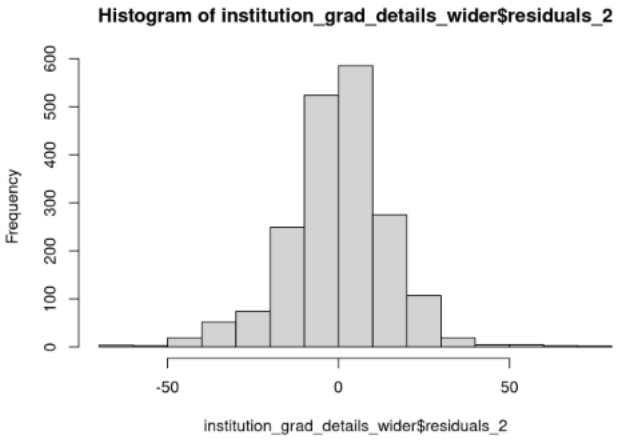
```
##
## Call:
## lm(formula = grad_100_value ~ control_public_private + pell_percentile +
##     retain_percentile + ft_pct + fte_percentile + F_W + F_A,
##     data = institution_grad_details_wider)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.722  -7.922   0.706   8.747  76.937
##
## Coefficients:
##                          Estimate Std. Error t value             Pr(>|t|)
## (Intercept)            -26.98956    2.58042 -10.459 < 0.0000000000000002 ***
## control_public_private  10.06979    0.77103  13.060 < 0.0000000000000002 ***
## pell_percentile         -0.15810    0.01556 -10.163 < 0.0000000000000002 ***
## retain_percentile        0.27672    0.01615  17.135 < 0.0000000000000002 ***
## ft_pct                   0.30693    0.02062  14.887 < 0.0000000000000002 ***
## fte_percentile           0.02409    0.01349   1.786               0.0743 .
## F_W                      0.28151    0.01878  14.987 < 0.0000000000000002 ***
## F_A                      0.31803    0.05727   5.553          0.000000032 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.31 on 1919 degrees of freedom
## Multiple R-squared:  0.5584, Adjusted R-squared:  0.5568
## F-statistic: 346.6 on 7 and 1919 DF,  p-value: < 0.00000000000000022
```

```
#stargazer(model_2, title="Results", align=TRUE,type = "text")
tab_model(model_2)
```

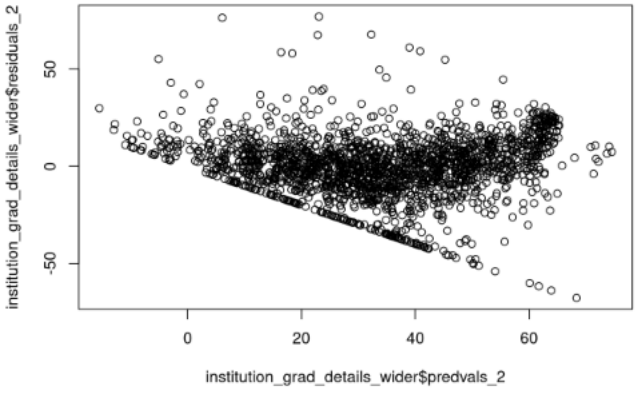
	grad_100_value		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-26.99	-32.05 – -21.93	**<0.001**
control public private	10.07	8.56 – 11.58	**<0.001**
pell percentile	-0.16	-0.19 – -0.13	**<0.001**
retain percentile	0.28	0.25 – 0.31	**<0.001**
ft pct	0.31	0.27 – 0.35	**<0.001**
fte percentile	0.02	-0.00 – 0.05	0.074
F W	0.28	0.24 – 0.32	**<0.001**
F A	0.32	0.21 – 0.43	**<0.001**
Observations	1927		
R^2 / R^2 adjusted	0.558 / 0.557		

Checking the regression assumptions for Model 2 and create histogram

```
institution_grad_details_wider$residuals_2 <- residuals(model_2)
hist(institution_grad_details_wider$residuals_2)
```




```
institution_grad_details_wider$predvals_2<-fitted(model_2)
plot(institution_grad_details_wider$predvals_2,institution_grad_details_wider$residuals_2)
```



```
vif(model_2)
```

```
## control_public_private        pell_percentile      retain_percentile
##               1.021468               1.646026               1.770759
##                 ft_pct         fte_percentile                    F_W
##               1.161043               1.234940               1.132752
##                    F_A
##               1.098460
```

```
stargazer(vif(model_2), title="VIF", type = "text")
```

```
##
## VIF
## ================================================================================
## control_public_private pell_percentile retain_percentile ft_pct fte_percentile  F_W   F_A
## --------------------------------------------------------------------------------
## 1.021                      1.646            1.771       1.161      1.235      1.133 1.098
## --------------------------------------------------------------------------------
```

Multiple Linear Regression for 6-year graduation with all variables

```
model_3 <- lm(grad_150_value~control_public_private+ln_aid_value+pell_percentile+retain_percentile+ft_pct+fte_percentile+M_W
+F_W+M_B+F_B+M_H+F_H+M_Ai+F_Ai+M_A, data = institution_grad_details_wider)
summary(model_3)
```

```
##
## Call:
## lm(formula = grad_150_value ~ control_public_private + ln_aid_value +
##     pell_percentile + retain_percentile + ft_pct + fte_percentile +
##     M_W + F_W + M_B + F_B + M_H + F_H + M_Ai + F_Ai + M_A, data = institution_grad_details_wider)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.368  -6.083  -0.093   5.681  82.511
##
## Coefficients:
##                          Estimate Std. Error t value             Pr(>|t|)
## (Intercept)            -59.60252    6.97099  -8.550 < 0.0000000000000002 ***
## control_public_private  -5.42670    0.69344  -7.826 0.00000000000000829 ***
## ln_aid_value            12.52318    0.60002  20.871 < 0.0000000000000002 ***
## pell_percentile         -0.07999    0.01320  -6.058 0.00000000165388858 ***
## retain_percentile        0.28499    0.01371  20.781 < 0.0000000000000002 ***
## ft_pct                   0.11657    0.01837   6.347 0.00000000027285391 ***
## fte_percentile           0.04697    0.01139   4.122 0.00003912429826997 ***
## M_W                     -0.19671    0.05386  -3.652             0.000267 ***
## F_W                     -0.10810    0.05506  -1.963             0.049740 *
## M_B                     -0.32324    0.06469  -4.997 0.00000063603696248 ***
## F_B                     -0.26956    0.05934  -4.543 0.00005894689218900 ***
## M_H                     -0.30088    0.07426  -4.052 0.00005289382813157 ***
## F_H                     -0.30618    0.07646  -4.004 0.00006458377470523 ***
## M_Ai                    -0.45733    0.08789  -5.203 0.00000021676915942 ***
## F_Ai                    -0.49216    0.08545  -5.760 0.00000000980332743 ***
## M_A                     -0.34193    0.11358  -3.010             0.002642 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.37 on 1911 degrees of freedom
## Multiple R-squared:  0.6736, Adjusted R-squared:  0.671
## F-statistic: 262.9 on 15 and 1911 DF,  p-value: < 0.0000000000000022
```

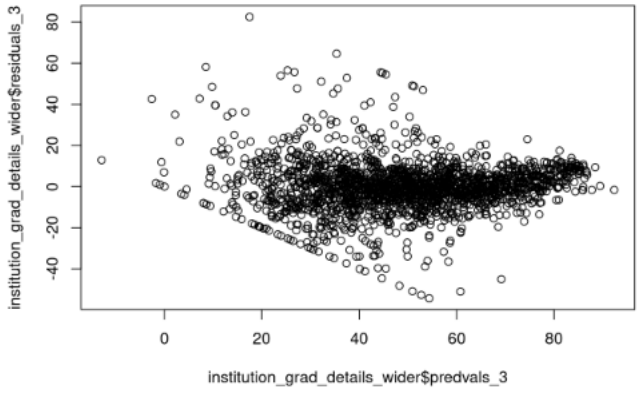
```
#stargazer(model_3, title="Results",type = "text")
tab_model(model_3)
```

	grad_150_value		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-59.60	-73.27 – -45.93	**<0.001**
control public private	-5.43	-6.79 – -4.07	**<0.001**
ln aid value	12.52	11.35 – 13.70	**<0.001**
pell percentile	-0.08	-0.11 – -0.05	**<0.001**
retain percentile	0.28	0.26 – 0.31	**<0.001**
ft pct	0.12	0.08 – 0.15	**<0.001**
fte percentile	0.05	0.02 – 0.07	**<0.001**
M W	-0.20	-0.30 – -0.09	**<0.001**
F W	-0.11	-0.22 – -0.00	**0.050**
M B	-0.32	-0.45 – -0.20	**<0.001**
F B	-0.27	-0.39 – -0.15	**<0.001**
M H	-0.30	-0.45 – -0.16	**<0.001**
F H	-0.31	-0.46 – -0.16	**<0.001**
M Ai	-0.46	-0.63 – -0.28	**<0.001**
F Ai	-0.49	-0.66 – -0.32	**<0.001**
M A	-0.34	-0.56 – -0.12	**0.003**
Observations	1927		
R^2 / R^2 adjusted	0.674 / 0.671		

Hist for model 3

```
institution_grad_details_wider$residuals_3 <- residuals(model_3)
hist(institution_grad_details_wider$residuals_3)
```


```
institution_grad_details_wider$predvals_3<-fitted(model_3)
plot(institution_grad_details_wider$predvals_3,institution_grad_details_wider$residuals_3)
```



```
vif(model_3)
```

```
## control_public_private           ln_aid_value        pell_percentile
##               1.264681               1.756978               1.814847
##      retain_percentile                 ft_pct         fte_percentile
##               1.954589               1.410843               1.349353
##                    M_W                    F_W                    M_B
##              13.385428              14.893361               5.460959
##                    F_B                    M_H                    F_H
##               9.867783               2.711186               4.831660
##                   M_AI                   F_AI                    M_A
##               1.728153               1.785355               3.483914
```

```
stargazer(vif(model_3), title="VIF", type = "text")
```

```
##
## VIF
## ============================================================================================================================================
## control_public_private ln_aid_value pell_percentile retain_percentile ft_pct fte_percentile  M_W    F_W    M_B   F_B   M_
H   F_H  M_AI  F_AI   M_A
## --------------------------------------------------------------------------------------------------------------------------------------------
## 1.265                  1.757          1.815           1.955          1.410      1.349       13.385 14.893 5.461 9.868 2.7
11 4.832 1.728 1.785 3.484
## --------------------------------------------------------------------------------------------------------------------------------------------
```

Run the fourth multiple linear regression model for 6 year graduation by dropping ln_aid_value, awards_per_natl_value, F_B, M_B, M_Ai, F_Ai, M_A,F_H due to multicolinearity

```
model_4 <- lm(grad_150_value~control_public_private+retain_percentile+ft_pct+fte_percentile+F_W+M_W+F_A, data = institution_
grad_details_wider)
summary(model_4)
```

```
##
## Call:
## lm(formula = grad_150_value ~ control_public_private + retain_percentile +
##     ft_pct + fte_percentile + F_W + M_W + F_A, data = institution_grad_details_wider)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.060  -5.684   0.639   6.958  79.422
##
## Coefficients:
##                        Estimate Std. Error t value             Pr(>|t|)
## (Intercept)          -13.60962    1.97798  -6.881 0.0000000000804208 ***
## control_public_private   0.98997    0.69709   1.385                0.192
## retain_percentile        0.36748    0.01332  27.597 < 0.0000000000000002 ***
## ft_pct                   0.26896    0.01873  14.359 < 0.0000000000000002 ***
## fte_percentile           0.08513    0.01216   7.001 0.0000000000340116 ***
## F_W                      0.29087    0.01642  17.710 < 0.0000000000000002 ***
## M_W                      0.14179    0.01770   8.011 0.0000000000000195 ***
## F_A                      0.37295    0.05366   6.950 0.0000000000498408 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.85 on 1919 degrees of freedom
## Multiple R-squared:  0.5896, Adjusted R-squared:  0.5881
## F-statistic: 393.8 on 7 and 1919 DF,  p-value: < 0.00000000000000022
```

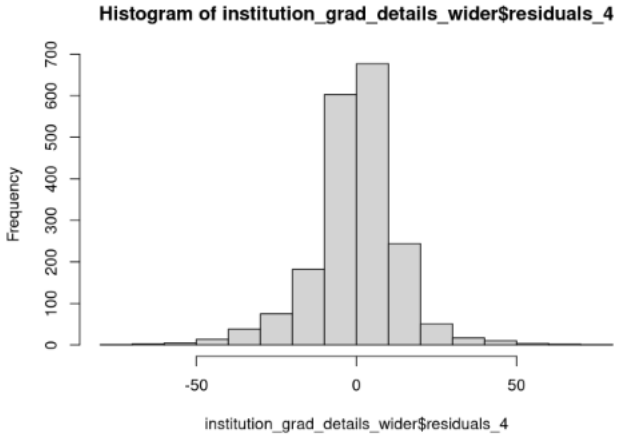
```
#stargazer(model_4, title="Results",type = "text")
tab_model(model_4)
```

```
#stargazer(model_4, title="Results",type = "text")
tab_model(model_4)
```

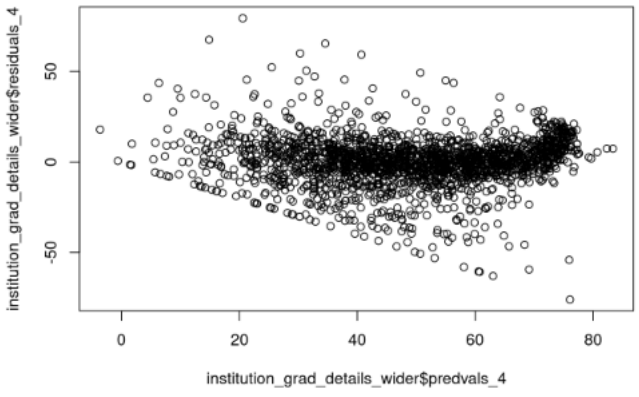
	grad_150_value		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-13.61	-17.49 – -9.73	**<0.001**
control public private	0.91	-0.46 – 2.28	0.192
retain percentile	0.37	0.34 – 0.39	**<0.001**
ft pct	0.27	0.23 – 0.31	**<0.001**
fte percentile	0.09	0.06 – 0.11	**<0.001**
F W	0.29	0.26 – 0.32	**<0.001**
M W	0.14	0.11 – 0.18	**<0.001**
F A	0.37	0.27 – 0.48	**<0.001**
Observations	1927		
R^2 / R^2 adjusted	0.590 / 0.588		

Histogram for model 4

```
institution_grad_details_wider$residuals_4 <- residuals(model_4)
hist(institution_grad_details_wider$residuals_4)
```



```
institution_grad_details_wider$predvals_4<-fitted(model_4)
plot(institution_grad_details_wider$predvals_4,institution_grad_details_wider$residuals_4)
```




```
vif(model_4)
```

```
## control_public_private      retain_percentile                 ft_pct
##               1.020674              1.471818              1.171512
##         fte_percentile                   F_W                   M_W
##               1.227026              1.058575              1.154408
##                   F_A
##               1.178724
```

```
stargazer(vif(model_4), title="VIF", type = "text")
```

```
##
## VIF
## ================================================================================
## control_public_private retain_percentile ft_pct fte_percentile  F_W   M_W   F_A
## --------------------------------------------------------------------------------
## 1.021                        1.472       1.172      1.227      1.059 1.154 1.179
## --------------------------------------------------------------------------------
```

PCA

```
#glimpse(institution_grad_details_wider)
my_pca <- prcomp(institution_grad_details_wider[,c(-1,-2,-3,-5,-10,-11,-23,-24,-25,-26,-27,-28,-29,-30,-31,-32)], scale = TR
UE)
summary(my_pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.7859 1.5204 1.3144 1.1572 1.11890 1.10134 1.03433
## Proportion of Variance 0.1993 0.1445 0.1080 0.0837 0.07825 0.07581 0.06686
## Cumulative Proportion  0.1993 0.3438 0.4518 0.5355 0.61373 0.68954 0.75640
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.88762 0.84954 0.7076 0.70557 0.64660 0.59621 0.58312
## Proportion of Variance 0.04924 0.04511 0.0313 0.03111 0.02613 0.02222 0.02125
## Cumulative Proportion  0.80565 0.85075 0.8821 0.91316 0.93929 0.96151 0.98276
##                           PC15                  PC16
## Standard deviation     0.52516 0.00000000000001455
## Proportion of Variance 0.01724 0.00000000000000000
## Cumulative Proportion  1.00000 1.00000000000000000
```

Here, only PC1, PC2, PC3, PC4, PC5, PC6, PC7 have an eigenvalue > 1.

Loading Scores

```
loading_scores <- my_pca$rotation
loading_scores
```

```
##                                PC1         PC2         PC3         PC4
## control_public_private  0.07185157 -0.01329234 -0.30556933 -0.38151989
## pell_percentile        -0.41021730  0.06325529 -0.01073192 -0.03996399
## retain_percentile       0.37978133 -0.24267325 -0.09219150  0.13827623
## ft_pct                  0.21793197 -0.06450738 -0.40027193  0.04407793
## fte_percentile          0.19646286 -0.26799158 -0.12647980 -0.11846941
## M_W                     0.26511437  0.29273903  0.09681984  0.15150241
## F_W                     0.28749400  0.30160184  0.08226081 -0.24945763
## M_B                    -0.38609665 -0.04330976 -0.39346701  0.12416548
## F_B                    -0.40085215 -0.08290491 -0.37687834  0.06780561
## M_H                    -0.08360324 -0.41714558  0.24590398 -0.34884041
## F_H                    -0.09439129 -0.42000646  0.24010367 -0.45010567
## M_Ai                   -0.06977275  0.06280312  0.22594990  0.16627494
## F_Ai                   -0.06756825  0.04340859  0.28194767  0.11812359
## M_A                     0.10591097 -0.41609628  0.05321756  0.41848105
## F_A                     0.08546413 -0.37676394  0.03558178  0.37911218
## ln_aid_value            0.29947887 -0.06489025 -0.39501788 -0.16677170
```

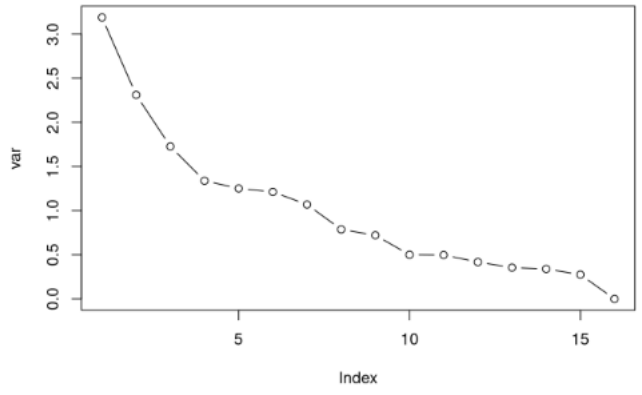
```
##                                PC5         PC6         PC7          PC8
## control_public_private  0.3340983400 -0.20890576  0.35791442 -0.530728878
## pell_percentile         0.0208417298 -0.17800359  0.27859658  0.268506516
## retain_percentile       0.0175710691  0.04974042 -0.28959657  0.024563617
## ft_pct                  0.0520647360 -0.34507849 -0.17746962  0.539493848
## fte_percentile          0.0127933866  0.41357386 -0.37832624 -0.257030779
## M_W                    -0.2896412002 -0.48929535 -0.12399242 -0.303257201
## F_W                     0.0507946620  0.45319522  0.24600924  0.302751478
## M_B                    -0.0263908697  0.01791458 -0.17540489 -0.105500253
## F_B                     0.0189189994  0.18346265 -0.14144233 -0.018189983
## M_H                    -0.1413663420 -0.28779588 -0.12985303  0.001890796
## F_H                     0.0003079021 -0.06089966 -0.02373745  0.176741064
## M_Ai                    0.5798893281 -0.13752102 -0.22264488  0.151590159
## F_Ai                    0.5887946705 -0.01831263 -0.19316389 -0.142828499
## M_A                    -0.0295975277 -0.08311547  0.26782812 -0.019553978
## F_A                     0.0793926601  0.13030378  0.46437695 -0.030203646
## ln_aid_value            0.2951604648 -0.15372476  0.13204240  0.132323784
##                                PC9         PC10         PC11         PC12
## control_public_private  0.077132203 -0.081349449  0.087077752 -0.25869844
## pell_percentile         0.006227841  0.399106905 -0.374513360  0.23014192
## retain_percentile      -0.163226085 -0.179232439  0.272479083  0.31599566
## ft_pct                 -0.109571868 -0.062113172 -0.207804354 -0.51322979
## fte_percentile          0.246296133  0.383301120 -0.510596393 -0.09844132
## M_W                     0.051680082  0.008548136 -0.265346032  0.14475424
## F_W                    -0.011233157  0.145174892  0.163771490 -0.16513306
## M_B                    -0.030898355  0.056245880  0.169291189 -0.07785930
## F_B                    -0.027794110 -0.167921193  0.072296311  0.13024999
## M_H                     0.023472318  0.168432859  0.235370717 -0.22534932
## F_H                    -0.058830063 -0.273912130 -0.167532108  0.23298074
## M_Ai                    0.671831317 -0.081608368  0.100202785  0.01409390
## F_Ai                   -0.657493089  0.138481603 -0.131704678 -0.08512255
## M_A                     0.029840437  0.516572955  0.310350326 -0.05872976
## F_A                     0.034347169 -0.406074953 -0.368419398 -0.08639651
## ln_aid_value           -0.020987673  0.194177468  0.001368186  0.55900880
##                               PC13        PC14         PC15
## control_public_private  0.29475672 -0.07725087  0.103767459
## pell_percentile         0.37916591 -0.37938348 -0.062669751
## retain_percentile       0.45831041 -0.48874417 -0.014812776
## ft_pct                  0.09083494  0.08327677  0.064865071
## fte_percentile          0.03170619 -0.01168035  0.005052689
## M_W                     0.02281960  0.02906297  0.028848484
## F_W                     0.02824457 -0.13371612  0.006646245
## M_B                    -0.42908469 -0.44924490  0.361384154
## F_B                     0.32502533  0.43262422 -0.338417516
## M_H                    -0.16047970 -0.17838896 -0.542353948
## F_H                     0.02431257  0.16341793  0.528959617
## M_Ai                    0.02472268 -0.04144492  0.019022433
## F_Ai                   -0.04411070  0.03278292 -0.034633704
## M_A                     0.13455918  0.28151546  0.276155592
## F_A                    -0.18854907 -0.20042158 -0.222011444
## ln_aid_value           -0.42001742  0.12533396 -0.183827749
##                                              PC16
## control_public_private  0.000000000000000004944777
## pell_percentile        -0.000000000000000094333466
## retain_percentile      -0.000000000000000015927209
## ft_pct                 -0.000000000000000002363582
## fte_percentile          0.000000000000000043604942
## M_W                     0.533616271620935278541 25
## F_W                     0.550662039935379210575 66
## M_B                     0.283782769800823453376 64
## F_B                     0.415887863079665187626 69
## M_H                     0.174100702959063065615009
```

```
## F_H                  0.22583085814726988616030
## M_Ai                 0.11749729318617284368198
## F_Ai                 0.12283937669320632579684
## M_A                  0.12909759891781558138746
## F_A                  0.17784762939164547312920
## ln_aid_value        -0.00000000000000009447479
```

Scree Plot

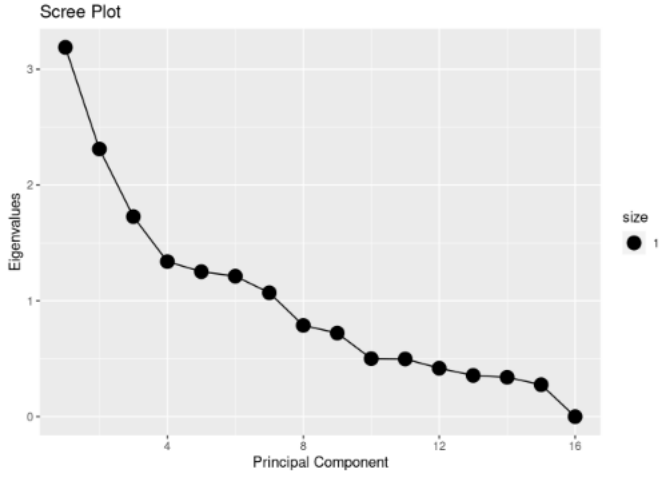
```
var <- my_pca$sdev^2

#in base R
plot(var, type = "b", lty = 1)
```




```
#using ggplot
qplot(c(1:16), var) +
  geom_line() +
  geom_point(aes(size=1))+
  xlab("Principal Component") +
  ylab("Eigenvalues") +
  ggtitle("Scree Plot") #+
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

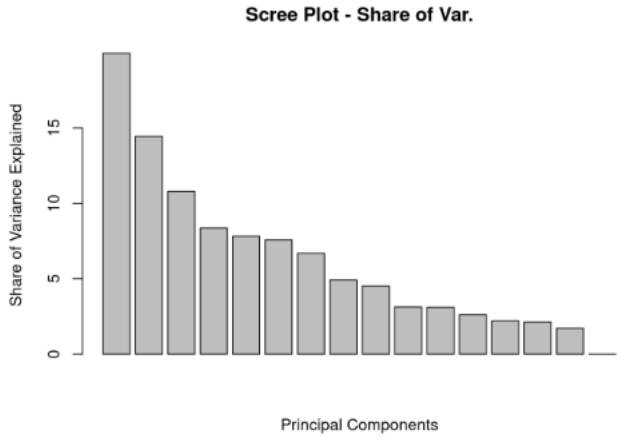



```
#ylim(0, 2.5)
```

Scree Plot by bar graph

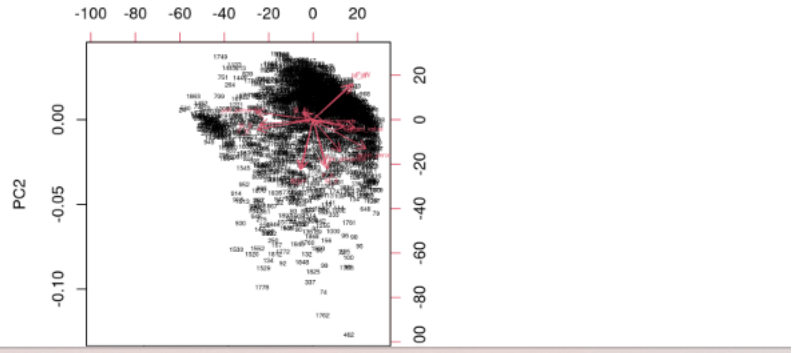
```
#CREATE A SCREE-PLOT OF SHARE OF VARIANCE
var_pct <- var/sum(var)*100

#barplot in base R
barplot(var_pct, main="Scree Plot - Share of Var.",
        xlab="Principal Components",
        ylab="Share of Variance Explained")
```



Biplot

```
#CREATE BIPLOT FOR PC1 AND PC2
biplot(my_pca, cex=.4)
```



Present a table of the scores for the first six observations in your original dataset.

```
scores <- my_pca$x
head(scores, n=6)
```

```
##             PC1         PC2        PC3        PC4        PC5        PC6
## [1,] -4.9632036 -0.05080501 -2.7716109  1.4725695 -0.3890633  0.2765541
## [2,] -0.2324147 -0.02753155  0.1566652  1.1699813 -0.7686946  1.2683107
## [3,] -2.2276914  0.24278975  2.3859729 -1.9091869 -0.6225395 -0.6564535
## [4,]  0.2746086  0.65540035  0.7379391  1.1683244 -0.7536568  0.2586581
## [5,] -4.9730701 -0.08513094 -2.7506879  1.4100161 -0.3794571  0.4799098
## [6,]  1.9337076  0.17684581 -0.1157167  0.5676344 -0.5874026  1.2756310
##             PC7        PC8            PC9         PC10        PC11       PC12
## [1,] -1.1982907 1.1177272 -0.42438945762  0.060249135  0.09589897 0.1480814
## [2,] -0.9951714 0.1562488 -0.07562669937 -0.136841157  0.07967809 0.3947869
## [3,]  1.4259369 0.1768367 -0.05138444091 -1.146239231 -0.92150720 0.2016009
## [4,] -0.9944236 0.2032518 -0.29429383691 -0.466336510  0.37065048 0.6143726
## [5,] -1.2052508 1.0148148 -0.31173922080  0.159186537 -0.09447783 0.2678268
## [6,] -1.9254713 0.3658772  0.00009958649  0.005160563 -0.20886673 0.1869447
##              PC13         PC14         PC15                        PC16
## [1,] -0.74145541 -0.24084720  0.11003637 -0.000000000000000272039699
## [2,]  0.09600449 -0.16878693 -0.16174965  0.000000000000000030836449
## [3,]  1.02009655  0.53826243  2.43521792 -0.000000000000000114804379
## [4,]  0.09424731  0.03094191 -0.20179223  0.000000000000000008687254
## [5,] -0.58987031 -0.04745349 -0.01427813 -0.000000000000000275055467
## [6,] -0.23821187 -0.03477901 -0.11207230  0.000000000000000129587355
```

Using these new PCA scores, regress the principal components you elected to retain on your dependent variable from your original dataset.

```
scores_combined <- cbind(institution_grad_details_wider,scores)

reg_1 <- lm(grad_100_value ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7, data=scores_combined)
summary(reg_1)
```

```
##
## Call:
## lm(formula = grad_100_value ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 +
##     PC7, data = scores_combined)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.782  -8.226  -0.272   8.516  72.074
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  33.9463     0.3278 103.554 < 0.0000000000000002 ***
## PC1           8.8222     0.1836  48.050 < 0.0000000000000002 ***
## PC2          -1.6896     0.2157  -7.834  0.0000000000000775 ***
## PC3          -5.4662     0.2495 -21.911 < 0.0000000000000002 ***
## PC4          -1.0828     0.2833  -3.821             0.000137 ***
## PC5           3.2626     0.2931  11.133 < 0.0000000000000002 ***
## PC6           0.8703     0.2977   2.923             0.003506 **
## PC7          -0.2506     0.3170  -0.790             0.429371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.39 on 1919 degrees of freedom
## Multiple R-squared:  0.6097, Adjusted R-squared:  0.6083
## F-statistic: 428.3 on 7 and 1919 DF,  p-value: < 0.00000000000000022
```

```
#stargazer(reg_1, title="Results",type = "text")
tab_model(reg_1)
```

	grad_100_value		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	33.95	33.30 – 34.59	**<0.001**
PC1	8.82	8.46 – 9.18	**<0.001**
PC2	-1.69	-2.11 – -1.27	**<0.001**
PC3	-5.47	-5.96 – -4.98	**<0.001**
PC4	-1.08	-1.64 – -0.53	**<0.001**
PC5	3.26	2.69 – 3.84	**<0.001**
PC6	0.87	0.29 – 1.45	**0.004**
PC7	-0.25	-0.87 – 0.37	0.429
Observations	1927		
R^2 / R^2 adjusted	0.610 / 0.608		

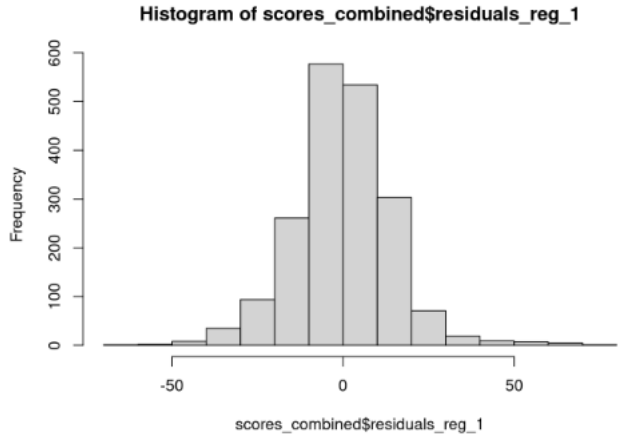
```
reg_2 <- lm(grad_150_value ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7, data=scores_combined)
summary(reg_2)
```

```
##
## Call:
## lm(formula = grad_150_value ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 +
##     PC7, data = scores_combined)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.745  -5.805   0.550   6.849  81.815
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  48.9635     0.3050 160.561 < 0.0000000000000002 ***
## PC1           8.7165     0.1708  51.034 < 0.0000000000000002 ***
## PC2          -2.1056     0.2006 -10.495 < 0.0000000000000002 ***
## PC3          -3.9936     0.2321 -17.208 < 0.0000000000000002 ***
## PC4           0.6179     0.2636   2.344               0.0192 *
## PC5           0.6067     0.2726   2.225               0.0262 *
## PC6           1.1684     0.2770   4.219      0.0000257223139 ***
## PC7          -1.9438     0.2949  -6.591      0.000000000562 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.39 on 1919 degrees of freedom
## Multiple R-squared:  0.6163, Adjusted R-squared:  0.6149
## F-statistic: 440.3 on 7 and 1919 DF,  p-value: < 0.00000000000000022
```

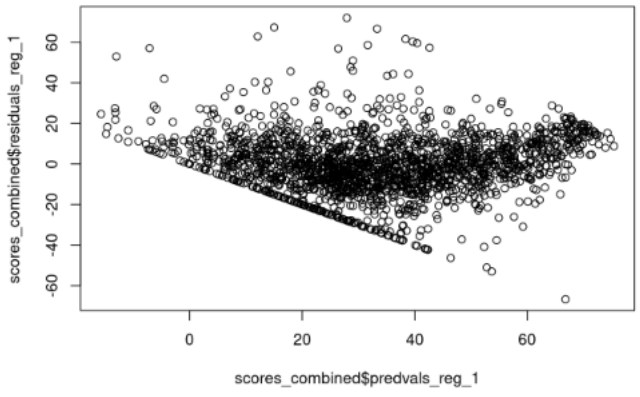
```
#stargazer(reg_2, title="Results",type = "text")
tab_model(reg_2)
```

	grad_150_value		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	48.96	48.37 – 49.56	**<0.001**
PC1	8.72	8.38 – 9.05	**<0.001**
PC2	-2.11	-2.50 – -1.71	**<0.001**
PC3	-3.99	-4.45 – -3.54	**<0.001**
PC4	0.62	0.10 – 1.13	**0.019**
PC5	0.61	0.07 – 1.14	**0.026**
PC6	1.17	0.63 – 1.71	**<0.001**
PC7	-1.94	-2.52 – -1.37	**<0.001**
Observations	1927		
R^2 / R^2 adjusted	0.616 / 0.615		

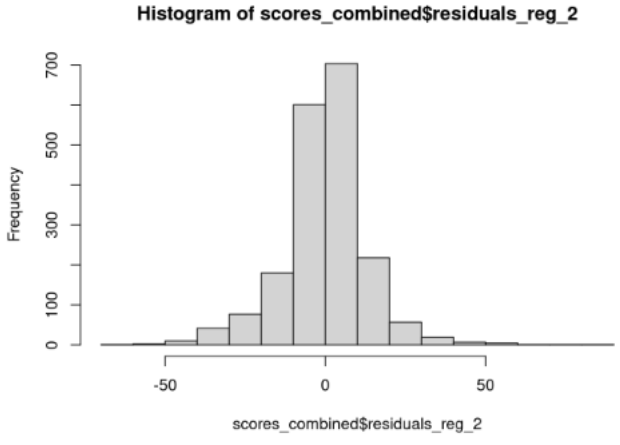
```
scores_combined$residuals_reg_1 <- residuals(reg_1)
hist(scores_combined$residuals_reg_1)
```



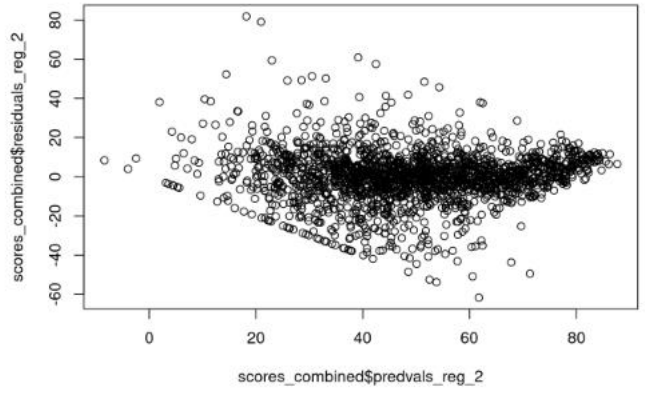
```
scores_combined$predvals_reg_1<-fitted(reg_1)
plot(scores_combined$predvals_reg_1,scores_combined$residuals_reg_1)
```




```
scores_combined$residuals_reg_2 <- residuals(reg_2)
hist(scores_combined$residuals_reg_2)
```




```
scores_combined$predvals_reg_2<-fitted(reg_2)
plot(scores_combined$predvals_reg_2,scores_combined$residuals_reg_2)
```



```
tab_model(model_2,reg_1)
```

	grad_100_value			grad_100_value		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	-26.99	-32.05 – -21.93	**<0.001**	33.95	33.30 – 34.59	**<0.001**
control public private	10.07	8.56 – 11.58	**<0.001**			
pell percentile	-0.16	-0.19 – -0.13	**<0.001**			
retain percentile	0.28	0.25 – 0.31	**<0.001**			
ft pct	0.31	0.27 – 0.35	**<0.001**			
fte percentile	0.02	-0.00 – 0.05	0.074			
F W	0.28	0.24 – 0.32	**<0.001**			
F A	0.32	0.21 – 0.43	**<0.001**			
PC1				8.82	8.46 – 9.18	**<0.001**
PC2				-1.69	-2.11 – -1.27	**<0.001**
PC3				-5.47	-5.96 – -4.98	**<0.001**
PC4				-1.08	-1.64 – -0.53	**<0.001**
PC5				3.26	2.69 – 3.84	**<0.001**
PC6				0.87	0.29 – 1.45	**0.004**
PC7				-0.25	-0.87 – 0.37	0.429
Observations	1927			1927		
R^2 / R^2 adjusted	0.558 / 0.557			0.610 / 0.608		

```
tab_model(model_4,reg_2)
```

	grad_150_value			grad_150_value		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	-13.61	-17.49 – -9.73	**<0.001**	48.96	48.37 – 49.56	**<0.001**
control public private	0.91	-0.46 – 2.28	0.192			
retain percentile	0.37	0.34 – 0.39	**<0.001**			
ft pct	0.27	0.23 – 0.31	**<0.001**			
fte percentile	0.09	0.06 – 0.11	**<0.001**			
F W	0.29	0.26 – 0.32	**<0.001**			
M W	0.14	0.11 – 0.18	**<0.001**			
F A	0.37	0.27 – 0.48	**<0.001**			
PC1				8.72	8.38 – 9.05	**<0.001**
PC2				-2.11	-2.50 – -1.71	**<0.001**
PC3				-3.99	-4.45 – -3.54	**<0.001**
PC4				0.62	0.10 – 1.13	**0.019**
PC5				0.61	0.07 – 1.14	**0.026**
PC6				1.17	0.63 – 1.71	**<0.001**
PC7				-1.94	-2.52 – -1.37	**<0.001**
Observations	1927			1927		
R^2 / R^2 adjusted	0.590 / 0.588			0.616 / 0.615		

```
#glimpse(scores_combined)
scores_combined %>%
  filter(control_public_private == 2) %>%
  summarise(median_grad_100_private = median(grad_100_value),
            median_grad_150_private = median(grad_150_value))
```

```
##   median_grad_100_private median_grad_150_private
## 1                   35.7                      50
```

```
scores_combined %>%
  filter(control_public_private == 1) %>%
  summarise(median_grad_100_public = median(grad_100_value),
            median_grad_150_public = median(grad_150_value))
```

```
##   median_grad_100_public median_grad_150_public
## 1                   21.8                    45.9
```