Executive Summary:

The purpose of this report is to examine the tuition fee of US colleges based on number of characteristics. The dataset consists of a sample data collected from 1121 records and includes the following variables (prefixed by their column name in the data file):

- tuition: College tuition ("out-of-state" rate for those with in-state discount).
- pcttop25: Percent of new students from the top 25% of high school class.
- sf_ratio: Student to faculty ratio.
- accrate: Fraction of applicants accepted for admission.
- graduat: Percent of students who graduate.
- pct_phd: Percent of faculty with Ph.D.'s.
- fulltime: Percent of undergraduates who are full time students.
- alumni: Percent of alumni who donate.
- num_enrl: Number of new students enrolled.
- public_private: Is the college a public or private institution? public=0, private=1
- fac_comp: Average faculty compensation.

Abstract:

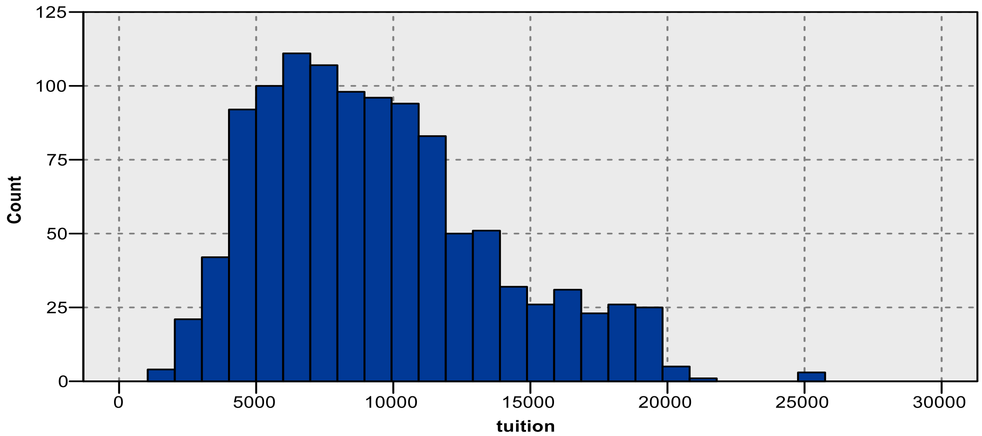
In this project, we will perform data mining processes on a dataset of tuition covering all the other factors of colleges in calculation and prediction. The objective of this project is to create a predictive model of college tuition based on a number of characteristics gathered from higher education institutions.

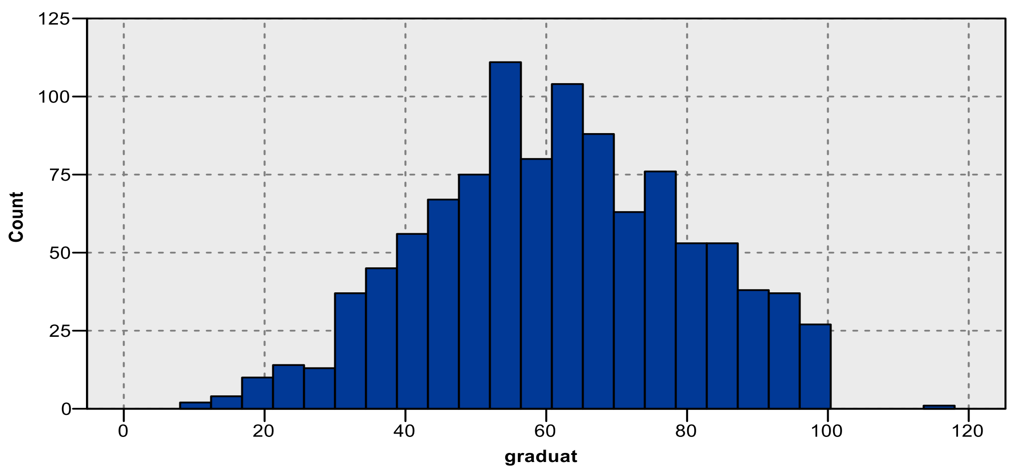
We will perform data mining process addressing the following points:

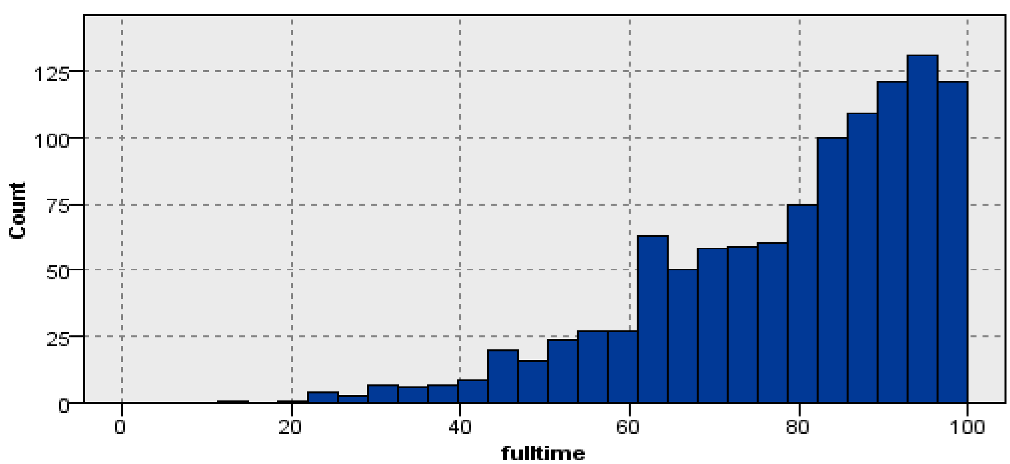
1. Explore the data to get some initial insights, if you think it is useful (your call):

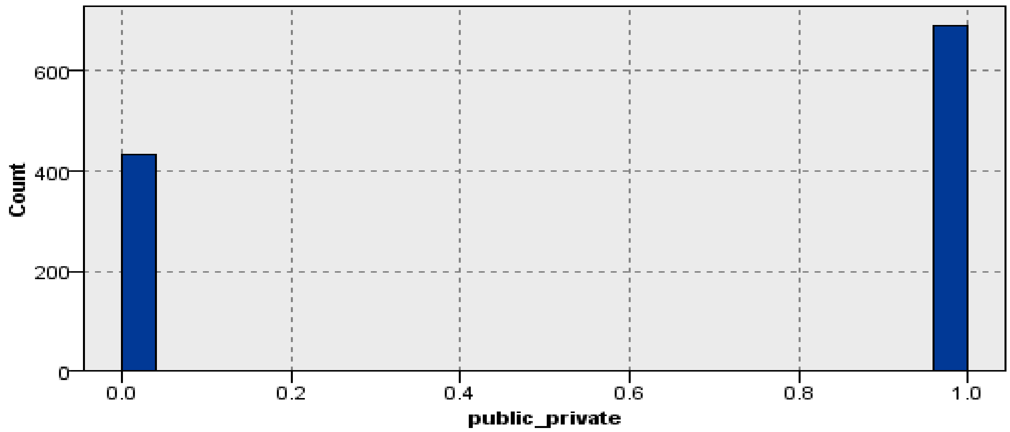
Our first step is to have a visual of some of the data, these will give us a rough idea about now the data is connected to each other, how the data is flowing from one variable to other. In our exploratory data analysis, we explored this phase and the data preparation phase simultaneously in order to utilize the new ideas of how to graphically explore the data every time the new areas of the data are uncovered.

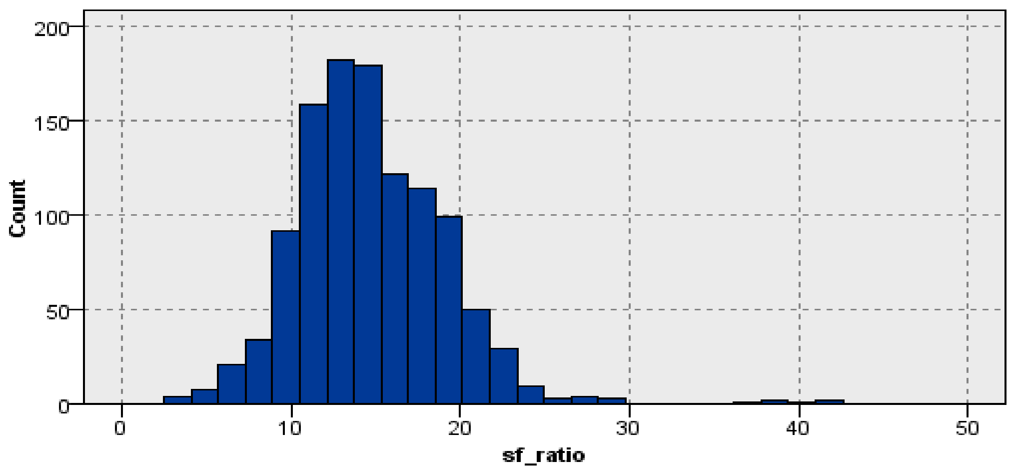
Firstly, we performed a data audit on our provided CVS file. We have found that the data set is not complete has lots of missing values across different fields. We will start by having visuals of some of fields on these data set, below is the attached snapshot of it:


125
100
75
50
25
0
Count
0
5000
10000
15000
20000
25000
30000
tuition


125
100
75
50
25
0
Count
0
20
40
60
80
100
120
graduat


125
100
75
50
25
0
Count
0
20
40
60
80
100
fulltime





2. Identify outliers and decide what to do with them:

Data Audit node report the outliners, defined here as values between 3 and 5 standard deviations from the mean, in the following fields: tuition, sf_ratio, accrate, graduat, pcr_phd, full time, alumni, num_enrl, fac_comp. There were values exceeding 5 standard deviations from the mean in the fields: Accrate, pct_phd, fulltime, num_enrl and fac_comp.

Data Audit of [11 fields] #2

File Edit Generate

Audit Quality Annotations

Complete fields (%): 27.27% Complete records (%): 71.72%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
tuition	Continuous	3	0	None	Never	Fixed	100	1121	0	0	0	0
pcttop25	Continuous	0	0	None	Never	Fixed	86.619	971	150	0	0	0
sf_ratio	Continuous	3	6	None	Never	Fixed	99.822	1119	2	0	0	0
accrate	Continuous	14	0	None	Never	Fixed	99.197	1112	9	0	0	0
graduat	Continuous	1	0	None	Never	Fixed	94.023	1054	67	0	0	0
pct_phd	Continuous	7	0	None	Never	Fixed	97.502	1093	28	0	0	0
fulltime	Continuous	9	0	None	Never	Fixed	98.037	1099	22	0	0	0
alumni	Continuous	5	0	None	Never	Fixed	85.37	957	164	0	0	0
num_enrl	Continuous	13	6	None	Never	Fixed	99.732	1118	3	0	0	0
public_private	Continuous	0	0	None	Never	Fixed	100	1121	0	0	0	0
fac_comp	Continuous	9	0	None	Never	Fixed	100	1121	0	0	0	0

Data Audit of [11 fields] #2

File Edit Generate

Audit Quality Annotations

Complete fields (%): 27.27% Complete records (%): 71.72%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
tuition	Continuous	3	0	None	Never	Fixed	100	1121	0	0	0	0
pcttop25	Continuous	0	0	None	Never	Fixed	86.619	971	150	0	0	0
sf_ratio	Continuous	3	6	None	Never	Fixed	99.822	1119	2	0	0	0
accrate	Continuous	14	0	None	Never	Fixed	99.197	1112	9	0	0	0
graduat	Continuous	1	0	None	Never	Fixed	94.023	1054	67	0	0	0
pct_phd	Continuous	7	0	None	Never	Fixed	97.502	1093	28	0	0	0
fulltime	Continuous	9	0	None	Never	Fixed	98.037	1099	22	0	0	0
alumni	Continuous	5	0	None	Never	Fixed	85.37	957	164	0	0	0
num_enrl	Continuous	13	6	None	Never	Fixed	99.732	1118	3	0	0	0
public_private	Continuous	0	0	None	Never	Fixed	100	1121	0	0	0	0
fac_comp	Continuous	9	0	None	Never	Fixed	100	1121	0	0	0	0

3. Missing data appears to be a problem with this data set. Prepare a copy of the dataset, where the missing values are each replaced with their field means. Report on how this substitution has affected the fields (summary stats, etc.), if at all. What do you think of this method of dealing with missing values?

First let's have a visual of missing dataset, the snapshot is attached below:

File Edit Generate

Audit Quality Annotations

Complete fields (%): 27.27% Complete records (%): 71.72%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete
tuition	Continuous	3	0	None	Never	Fixed	100
pcttop25	Continuous	0	0	None	Never	Fixed	86.619
sf_ratio	Continuous	3	6	None	Never	Fixed	99.822
accrate	Continuous	14	0	None	Never	Fixed	99.197
graduat	Continuous	1	0	None	Never	Fixed	94.023
pct_phd	Continuous	7	0	None	Never	Fixed	97.502
fulltime	Continuous	9	0	None	Never	Fixed	98.037
alumni	Continuous	5	0	None	Never	Fixed	85.37
num_enrl	Continuous	13	6	None	Never	Fixed	99.732
public_private	Continuous	0	0	None	Never	Fixed	100
fac_comp	Continuous	9	0	None	Never	Fixed	100

From the above we can see that the data is not complete, so now we will generate a missing node with addressing all the missing imputes by their mean

File Edit Generate

Audit | Quality | Annotations

Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
tuition	Continuous	3	0	None	Never	Fixed	100	1121	0	0	0	0
pcttop25	Continuous	0	0	None	Never	Fixed	100	1121	0	0	0	0
sf_ratio	Continuous	3	6	None	Never	Fixed	100	1121	0	0	0	0
accrate	Continuous	15	0	None	Never	Fixed	100	1121	0	0	0	0
graduat	Continuous	1	0	None	Never	Fixed	100	1121	0	0	0	0
pct_phd	Continuous	7	0	None	Never	Fixed	100	1121	0	0	0	0
fulltime	Continuous	10	0	None	Never	Fixed	100	1121	0	0	0	0
alumni	Continuous	9	0	None	Never	Fixed	100	1121	0	0	0	0
num_enrl	Continuous	13	6	None	Never	Fixed	100	1121	0	0	0	0
public_private	Continuous	0	0	None	Never	Fixed	100	1121	0	0	0	0
fac_comp	Continuous	9	0	None	Never	Fixed	100	1121	0	0	0	0

Below is complete summary stats of dataset after handling the missing values.

Statistics | Annotations

Collapse All | Expand All

tuition

Statistics

Count	1121
Mean	9446.558
Min	1044.000
Max	25750.000
Range	24706.000
Variance	17997213.074
Standard Deviation	4242.312
Standard Error of Mean	126.707

pcttop25

Statistics

Count	1121
Mean	53.493
Min	11.000
Max	100.000
Range	89.000
Variance	373.501
Standard Deviation	19.326
Standard Error of Mean	0.577

sf_ratio

Statistics

Count	1121
Mean	14.753
Min	2.500
Max	42.600
Range	40.100
Variance	19.705
Standard Deviation	4.439
Standard Error of Mean	0.133

accrate

Statistics

Count	1121
Mean	0.759
Min	0.154
Max	1.000
Range	0.846
Variance	0.023
Standard Deviation	0.151
Standard Error of Mean	0.005

Statistics | Annotations

Collapse All | Expand All

graduat

Statistics

Count	1121
Mean	61.421
Min	8.000
Max	118.000
Range	110.000
Variance	328.638
Standard Deviation	18.128
Standard Error of Mean	0.541

pct_phd

Statistics

Count	1121
Mean	70.202
Min	8.000
Max	103.000
Range	95.000
Variance	289.161
Standard Deviation	17.005
Standard Error of Mean	0.508

fulltime

Statistics

Count	1121
Mean	79.089
Min	11.430
Max	99.940
Range	88.510
Variance	264.248
Standard Deviation	16.256
Standard Error of Mean	0.486

alumni

Statistics

Count	1121
Mean	21.448
Min	0.000
Max	64.000
Range	64.000
Variance	136.742
Standard Deviation	11.694
Standard Error of Mean	0.349

num_enrl
Statistics

Count	1121
Mean	833.453
Min	21.000
Max	7425.000
Range	7404.000
Variance	851431.594
Standard Deviation	922.731
Standard Error of Mean	27.560

public_private
Statistics

Count	1121
Mean	0.615
Min	0.000
Max	1.000
Range	1.000
Variance	0.237
Standard Deviation	0.487
Standard Error of Mean	0.015

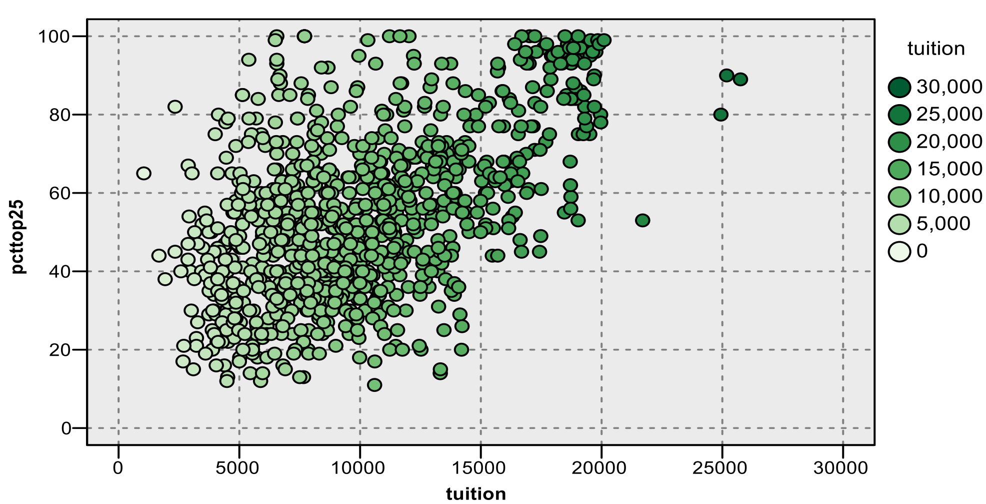
fac_comp
Statistics

Count	1121
Mean	52679.839
Min	26500.000
Max	107500.000
Range	81000.000
Variance	147960628.903
Standard Deviation	12163.907
Standard Error of Mean	363.304

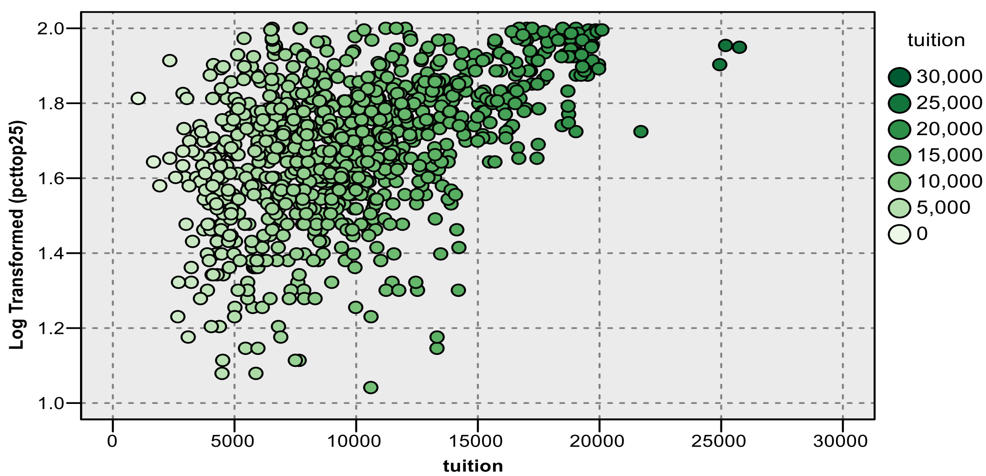
Using the mean of the field seems the best option to complete the missing values, if we have deal with complete data set, but mean field will only give us a projected value for that specific field, for this part of the project mean field seems to the best option.

4. **Provide a table describing the relationship of each explanatory variable with tuition (hint: use scatter plots). If the relationship is not linear, you can1 make it so by transforming the predictor variable.**

In order to describe the relationship of each variable we first start by having the visual of the data, below is the scatter plot of **tuition vs Pcttop25.** For this part of the project we have used the original data set with missing values to the graphical representation.



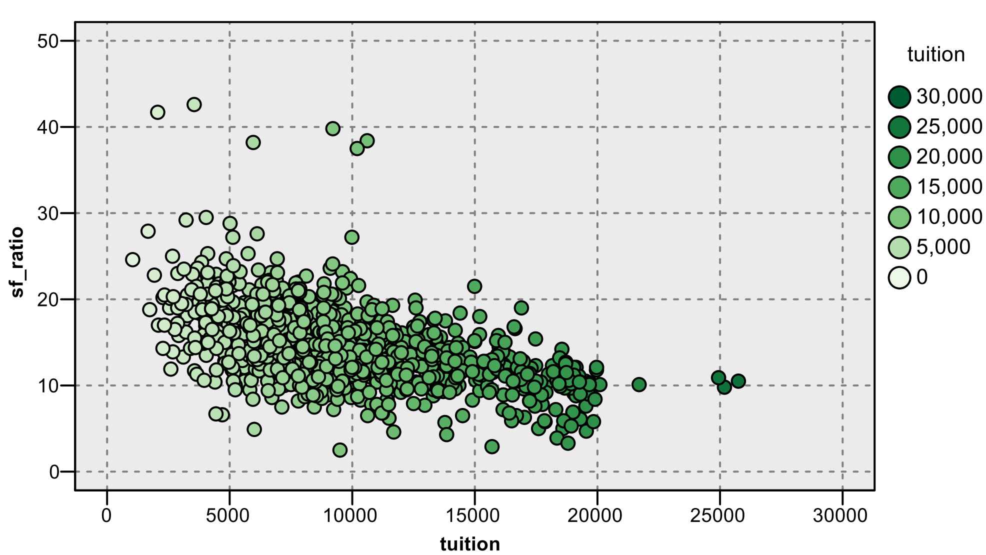
The above data for percent of new student from top 25% of high school class on the y axis and tuition on the x axis seems a bit skewed, but can have out idea that higher percentage of new students from top 25% of high school class has higher tuition fee. Let's have a log transformed y axis (percentage of new students from top 25% of high school class).



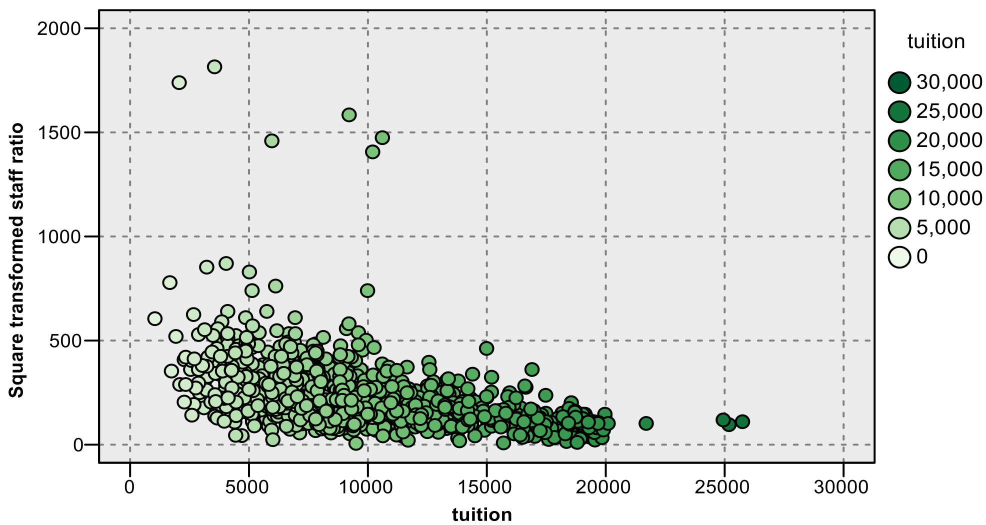
The above graph gives us a idea that with increase in percentage of new students from top 25% of high school class we have a increase in tuition fee.

Looking into the tuition vs staff ratio:

We first have a simple visual of data in graph.

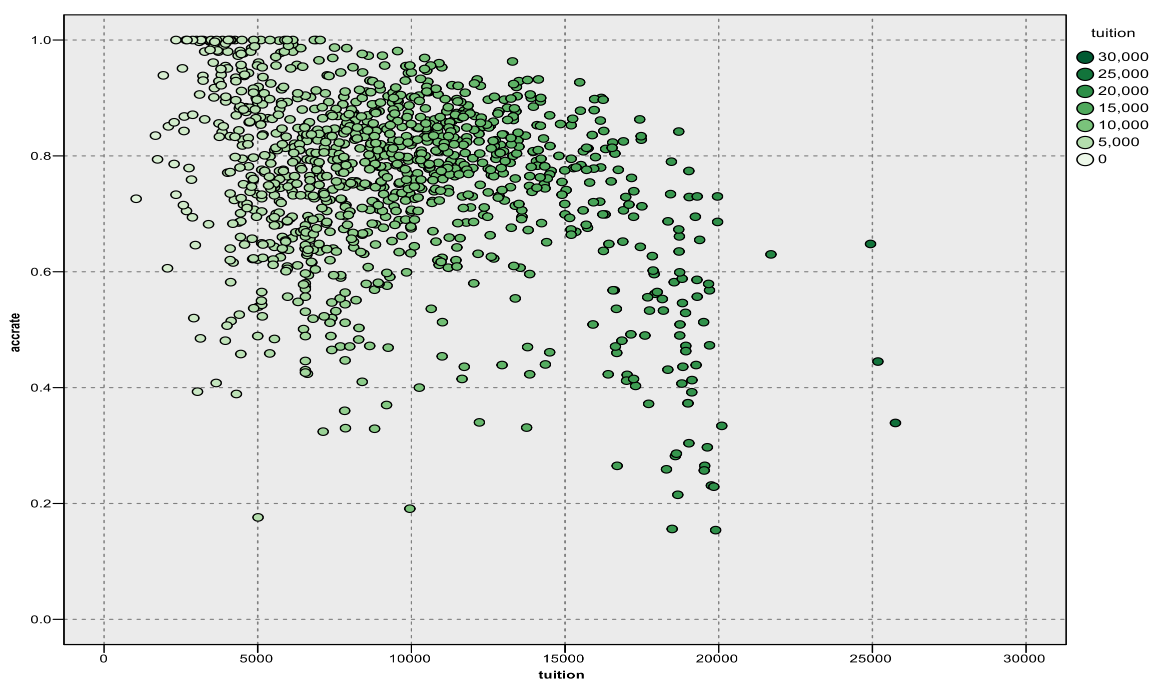


The above data set is seeming inversely proportional with staff ratio to tuition, let's try to transform the staff ratio axis if that makes the graph more logical and understandable.



In the above graph, we have used the square transformation for staff ratio, the graph is still bit skewed but have a visual understanding that with the increase in tuition fee we have decrease in staff ratio.

looking into tuition vs accurate:

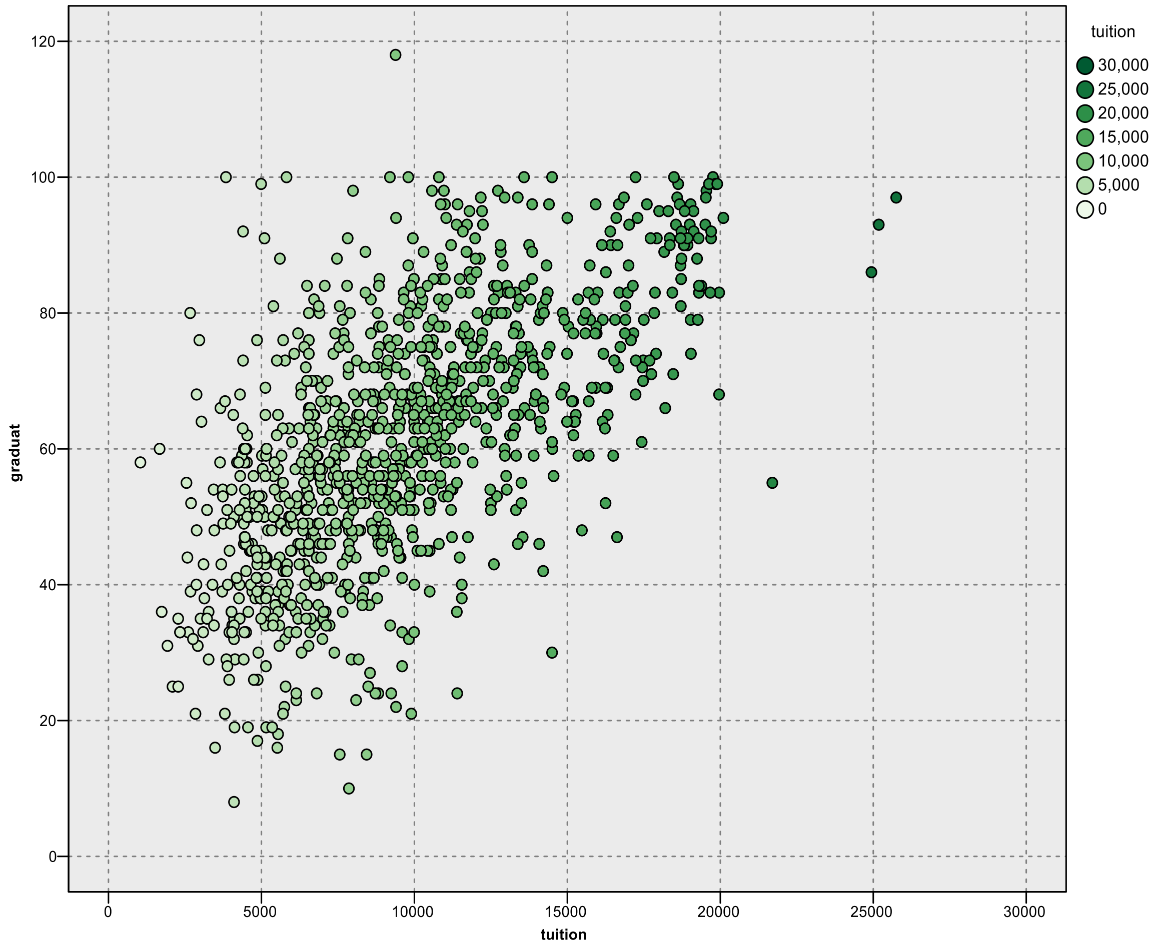


The above graph is skewed is hard to say anything with the visuals. Let’s try with different transformation representation of the graph.

After different hit and trial, squared transformation for fraction of applicants accepted for admission vs the tuition fee for US universities, we can say the graph is mostly skewed, but we have some density where we can say the fraction of applications accepted are with tuition fee around 5000 to 10000.

Tuition vs Graduation:

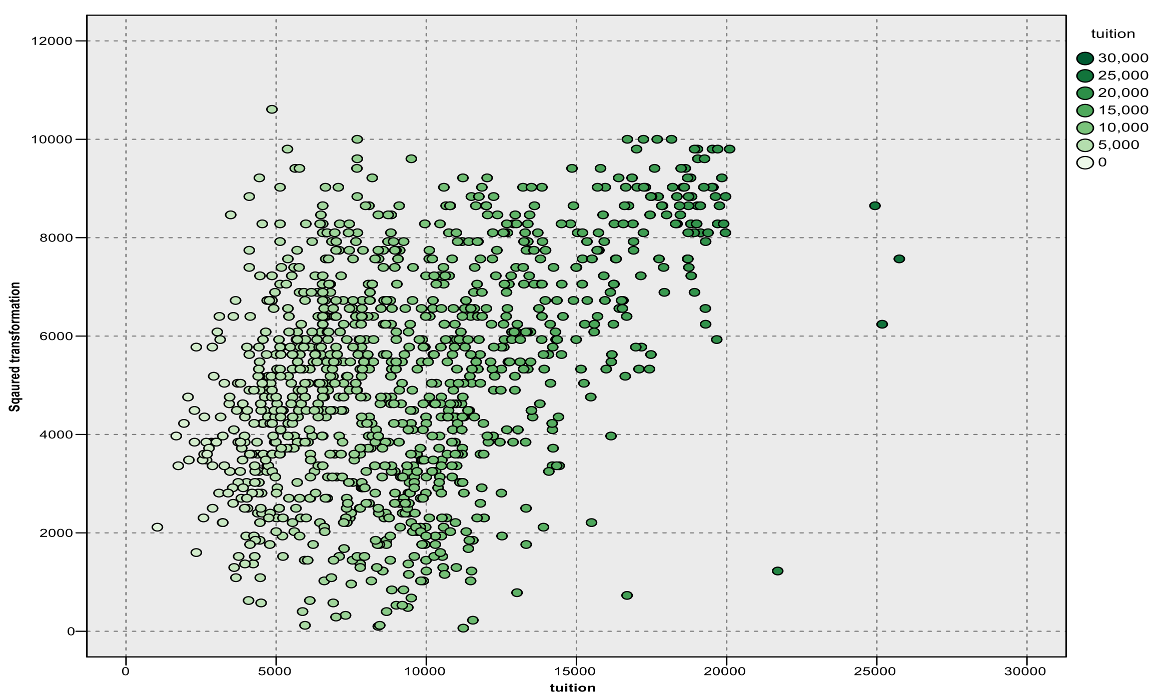
The theth



From the above graph we can say that the percentage of students who graduated increases with the increase in tuition fee.

Tuition vs PHD:

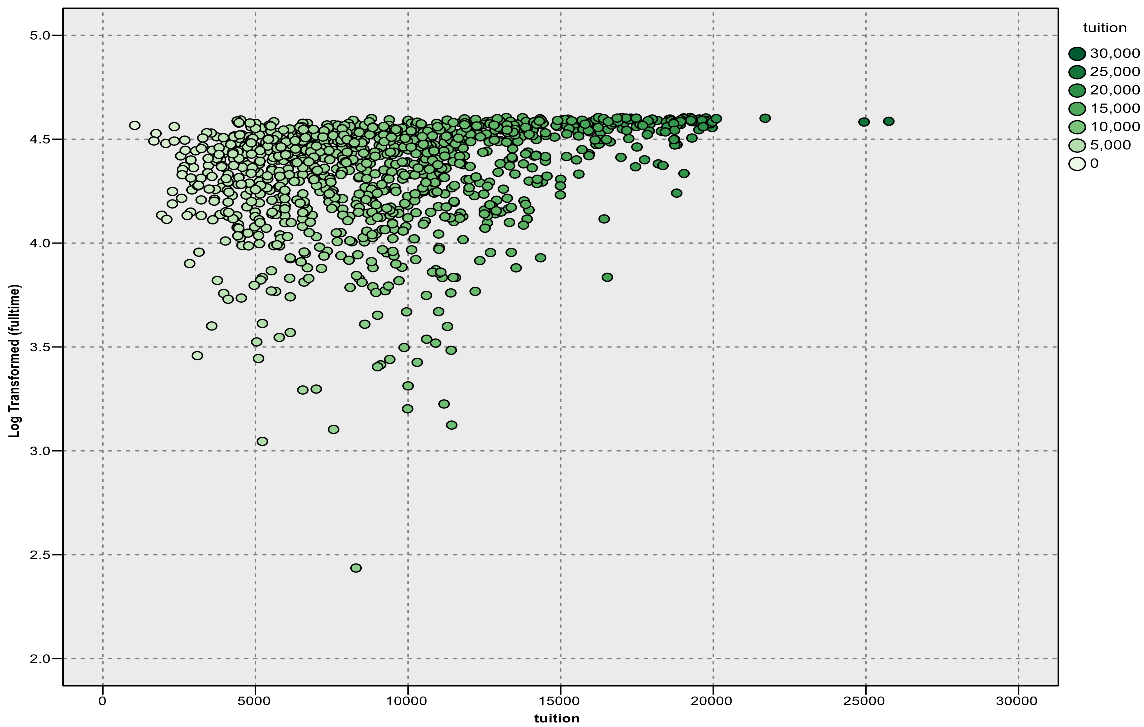
From the above graph we can say that with increase in tuition fee we have an increase in percentage of faculty who have a Ph.D.'s. Let's have transformed y axis if we have more clear visual.



The above graph is still skewed but we can say that with the increase in tuition fee we have more percentage of faculties with PHD degrees from the given data set.

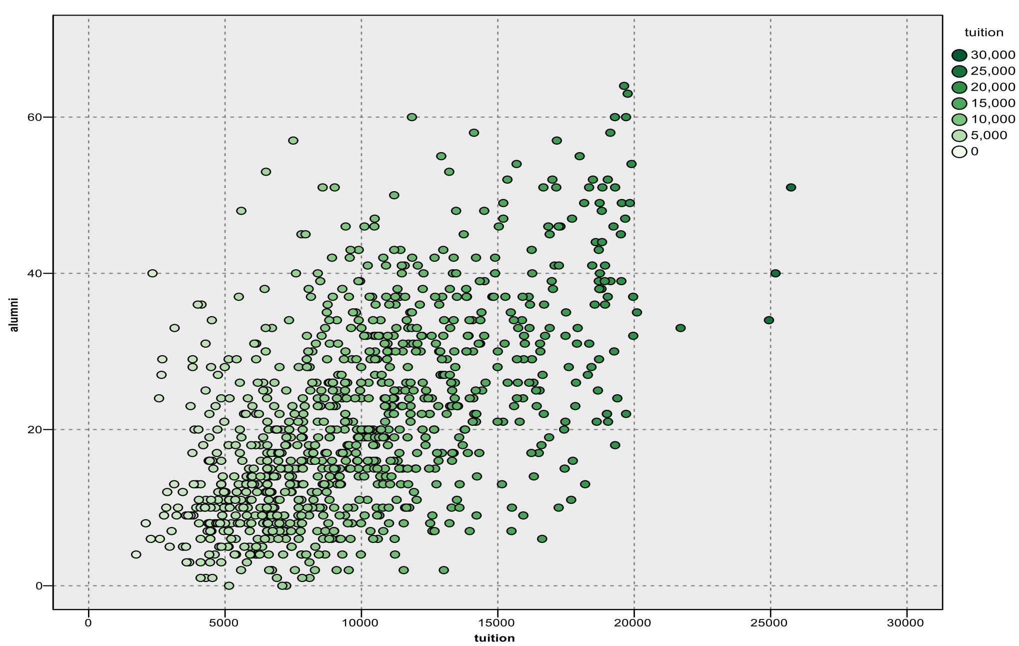
Tuition vs fulltime

The above graph is skewed but get have visual that with the increase in tuition we have a higher percentage of undergrads who are full time students.



From the we can see that we still have skewed graph but we from the visuals we can say that with the increase in tuition fee we have slightly more percentage of undergrads who are full time students.

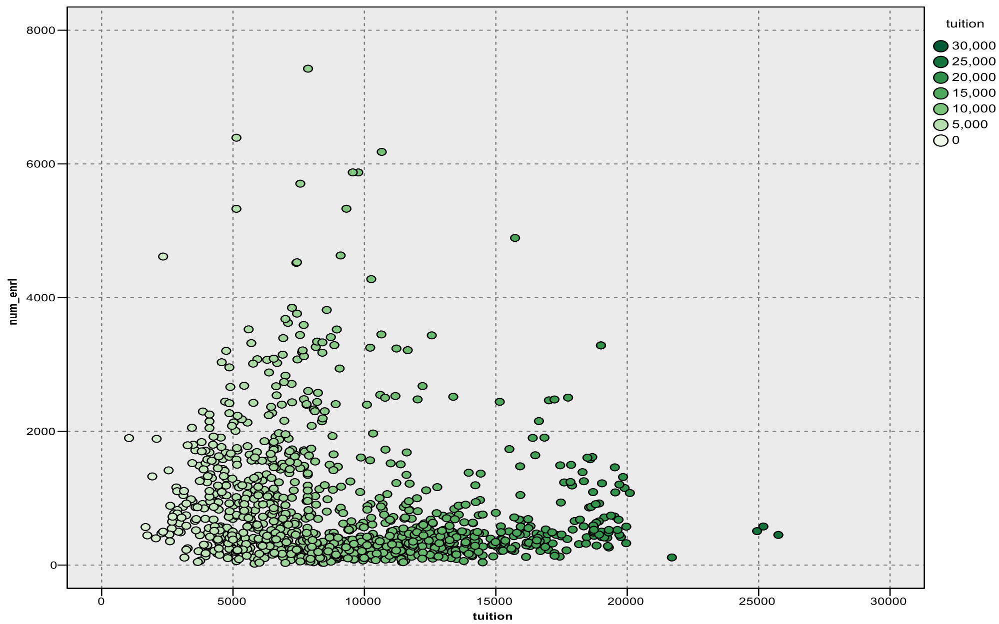
Tuition vs Alumni



The above is the normal graph between tuition on the x axis and alumni – percentage of alumni who donate. We can say that with the increase in percentage of donation we have increase in tuition fee as well. Let’s try a log transformation of y axis we have clearer visual.

From the above graph we can say that we the increase in percentage of donation we have increase in tuition fee as well.

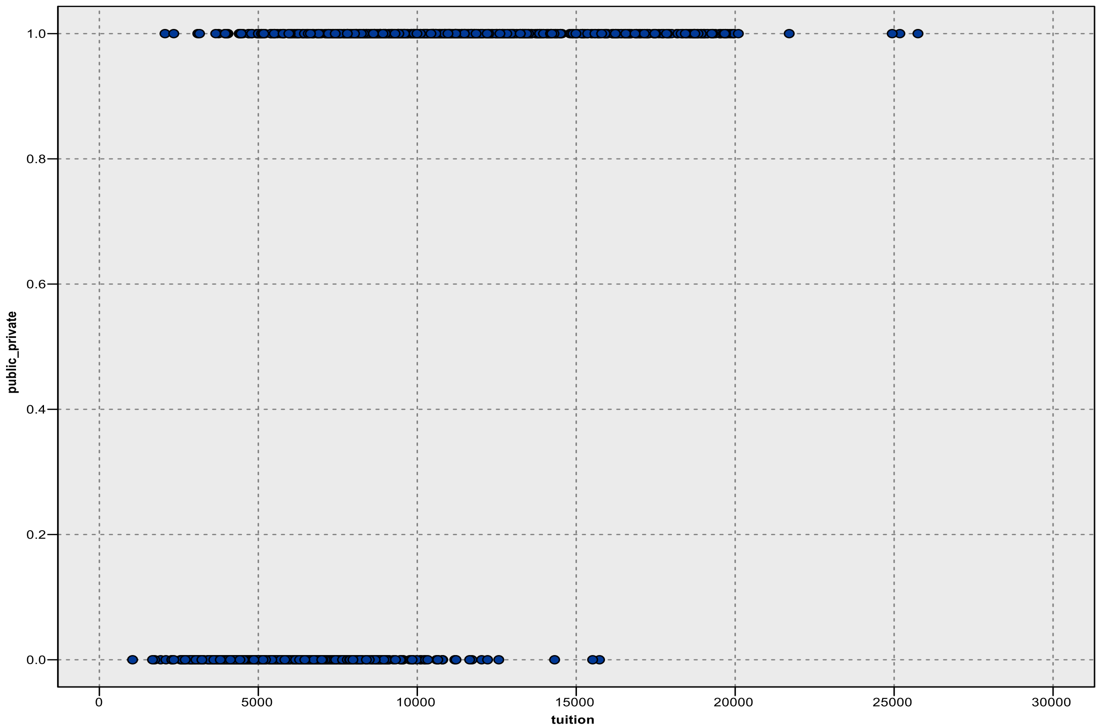
Tuition vs number of new students enrolled:



The above graph is very skewed is not possible to say anything from the visual, let try some transformations.

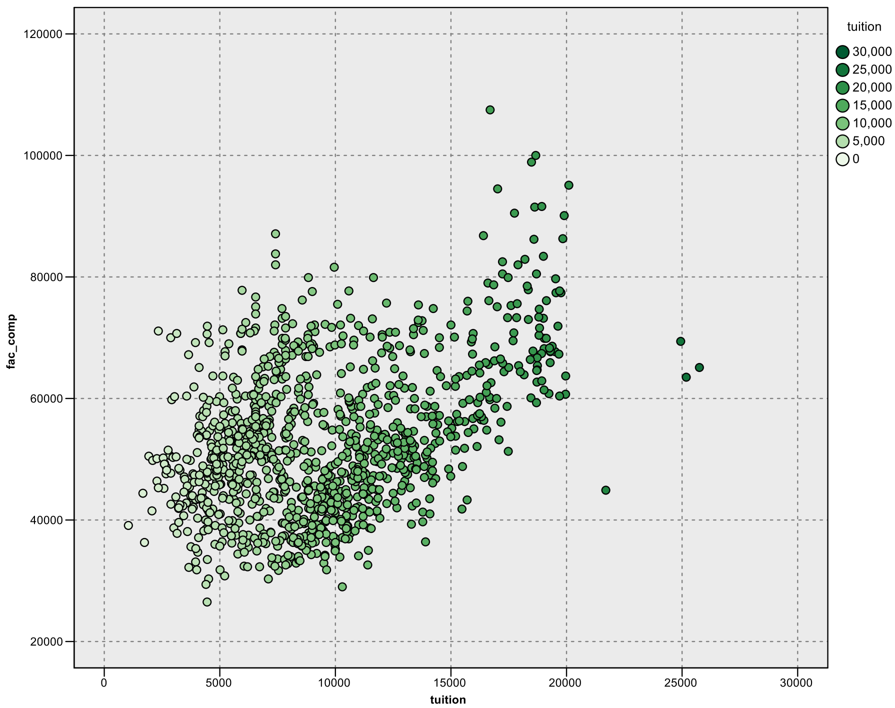
With the log transformation of number of new students enrolled we can say that we have a greater number of new students enrolled with tuition fee increasing mostly between 8000 – 15000.

Tuition vs Public Private:

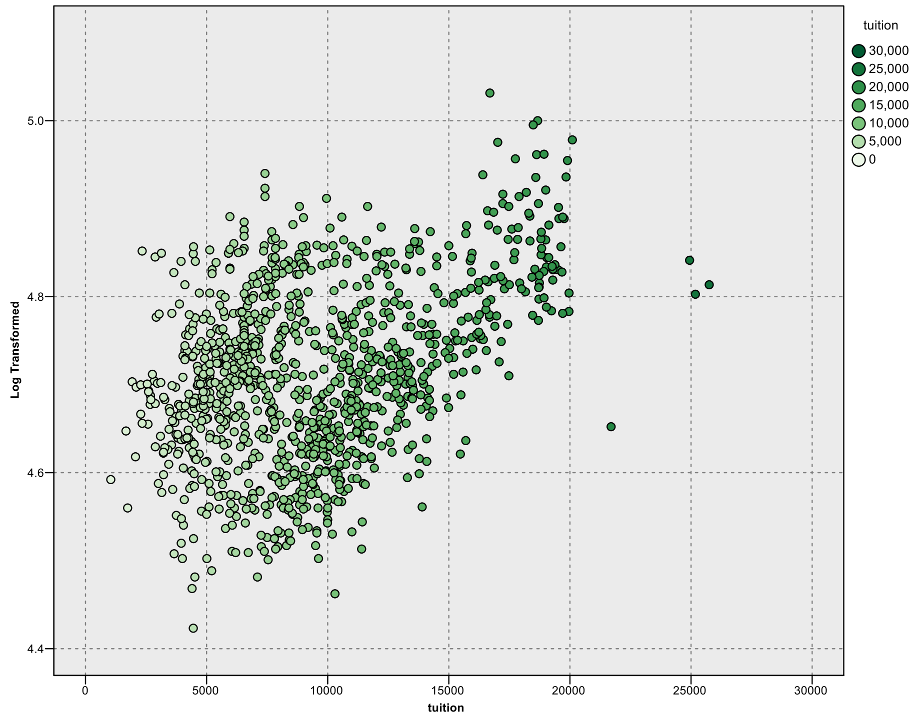


As we have a binary data for tuition vs public = 0 and private =1, we can say that private schools have more higher tuition fee as compared to the public schools.

Tuition vs average faculty Compensation:



From the above we can say that with the increase in tuition fee we have a increase in average faculty compensation, lets try some transformation if we can get more clear visuals.



The above graph is still bit skewed but we can say that with the increase in tuition fee we have increase in average compensation of faculty.

5. **Investigate the correlation among the predictor variables. Suggest a creative course of action (rather than simply omitting a variable) for dealing with any medium or strong correlations encountered (e.g. textbook, section 9.7; avoid any method linked to principal component analysis, as we have not covered it yet).**

The correlation among the predictor can be seen using statics node with given data set.

Statistics | Annotations

Collapse All | Expand All

tuition

Statistics

Count	1121
Mean	9446.558
Min	1044.000
Max	25750.000
Range	24706.000
Variance	17997213.074
Standard Deviation	4242.312
Standard Error of Mean	126.707
Median	8820.000
Mode	6550.000

Pearson Correlations

pcttop25	0.517	Strong
sf_ratio	-0.544	Strong
accrate	-0.323	Strong
graduat	0.635	Strong
pct_phd	0.386	Strong
fulltime	0.289	Strong
alumni	0.576	Strong
num_enrl	-0.166	Strong
public_private	0.609	Strong
fac_comp	0.415	Strong

pcttop25

Statistics

Count	971
Mean	53.493
Min	11.000
Max	100.000
Range	89.000
Variance	431.258
Standard Deviation	20.767
Standard Error of Mean	0.666
Median	51.000
Mode	40.000

Pearson Correlations

tuition	0.517	Strong
sf_ratio	-0.304	Strong
accrate	-0.451	Strong
graduat	0.495	Strong
pct_phd	0.549	Strong
fulltime	0.390	Strong
alumni	0.392	Strong
num_enrl	0.208	Strong
public_private	0.166	Strong
fac_comp	0.550	Strong

Statistics | Annotations

Collapse All | Expand All

sf_ratio

Statistics

Count	1119
Mean	14.753
Min	2.500
Max	42.600
Range	40.100
Variance	19.740
Standard Deviation	4.443
Standard Error of Mean	0.133
Median	14.300
Mode	12.100*

*Multiple modes exist. The smallest value is shown.

Pearson Correlations

tuition	-0.544	Strong
pcttop25	-0.304	Strong
accrate	0.183	Strong
graduat	-0.396	Strong
pct_phd	-0.110	Strong
fulltime	-0.083	Strong
alumni	-0.428	Strong
num_enrl	0.247	Strong
public_private	-0.485	Strong
fac_comp	-0.094	Strong

accrate

Statistics

Count	1112
Mean	0.759
Min	0.154
Max	1.000
Range	0.846
Variance	0.023
Standard Deviation	0.152
Standard Error of Mean	0.005
Median	0.784
Mode	1.000

Pearson Correlations

tuition	-0.323	Strong
pcttop25	-0.451	Strong
sf_ratio	0.183	Strong
graduat	-0.302	Strong
pct_phd	-0.347	Strong
fulltime	-0.147	Strong
alumni	-0.179	Strong
num_enrl	-0.123	Strong
public_private	-0.003	Weak
fac_comp	-0.500	Strong

Statistics | Annotations

Collapse All | Expand All

graduat

Statistics

Count	1054
Mean	61.421
Min	8.000
Max	118.000
Range	110.000
Variance	349.549
Standard Deviation	18.696
Standard Error of Mean	0.576
Median	61.000
Mode	63.000

Pearson Correlations

tuition	0.635	Strong
pcttop25	0.495	Strong
sf_ratio	-0.396	Strong
accrate	-0.302	Strong
pct_phd	0.289	Strong
fulltime	0.296	Strong
alumni	0.511	Strong
num_enrl	-0.075	Strong
public_private	0.465	Strong
fac_comp	0.317	Strong

pct_phd

Statistics

Count	1093
Mean	70.202
Min	8.000
Max	103.000
Range	95.000
Variance	296.575
Standard Deviation	17.221
Standard Error of Mean	0.521
Median	73.000
Mode	77.000

Pearson Correlations

tuition	0.386	Strong
pcttop25	0.549	Strong
sf_ratio	-0.110	Strong
accrate	-0.347	Strong
graduat	0.289	Strong
fulltime	0.276	Strong
alumni	0.242	Strong
num_enrl	0.322	Strong
public_private	-0.113	Strong
fac_comp	0.663	Strong

Statistics | Annotations

Collapse All | Expand All

fulltime

Statistics

Count	1099
Mean	79.089
Min	11.430
Max	99.940
Range	88.510
Variance	269.543
Standard Deviation	16.418
Standard Error of Mean	0.495
Median	83.560
Mode	84.570*

*Multiple modes exist. The smallest value is shown.

Pearson Correlations

tuition	0.289	Strong
pcttop25	0.390	Strong
sf_ratio	-0.083	Strong
accrate	-0.147	Strong
graduat	0.296	Strong
pct_phd	0.276	Strong
alumni	0.278	Strong
num_enrl	0.129	Strong
public_private	0.081	Strong
fac_comp	0.192	Strong

alumni

Statistics

Count	957
Mean	21.448
Min	0.000
Max	64.000
Range	64.000
Variance	160.199
Standard Deviation	12.657
Standard Error of Mean	0.409
Median	19.000
Mode	10.000

Pearson Correlations

tuition	0.576	Strong
pcttop25	0.392	Strong
sf_ratio	-0.428	Strong
accrate	-0.179	Strong
graduat	0.511	Strong
pct_phd	0.242	Strong
fulltime	0.278	Strong
num_enrl	-0.201	Strong
public_private	0.456	Strong
fac_comp	0.146	Strong

Statistics | Annotations

Collapse All | Expand All

num_enrl

Statistics

Count	1118
Mean	833.453
Min	21.000
Max	7425.000
Range	7404.000
Variance	853718.339
Standard Deviation	923.969
Standard Error of Mean	27.634
Median	478.500
Mode	169.000*

*Multiple modes exist. The smallest value is shown.

Pearson Correlations

tuition	-0.166	Strong
pcttop25	0.208	Strong
sf_ratio	0.247	Strong
accrate	-0.123	Strong
graduat	-0.075	Strong
pct_phd	0.322	Strong
fulltime	0.129	Strong
alumni	-0.201	Strong
public_private	-0.534	Strong
fac_comp	0.454	Strong

public_private

Statistics

Count	1121
Mean	0.615
Min	0.000
Max	1.000
Range	1.000
Variance	0.237
Standard Deviation	0.487
Standard Error of Mean	0.015
Median	1.000
Mode	1.000

Pearson Correlations

tuition	0.609	Strong
pcttop25	0.166	Strong
sf_ratio	-0.485	Strong
accrate	-0.003	Weak
graduat	0.465	Strong
pct_phd	-0.113	Strong
fulltime	0.081	Strong
alumni	0.456	Strong
num_enrl	-0.534	Strong
fac_comp	-0.195	Strong

Statistics | Annotations

Collapse All | Expand All

- tuition
- pcttop25
- sf_ratio
- accrate
- graduat
- pct_phd
- fulltime
- alumni
- num_enrl
- public_private
- fac_comp

Statistics

Count	1121
Mean	52679.839
Min	26500.000
Max	107500.000
Range	81000.000
Variance	147960628.903
Standard Deviation	12163.907
Standard Error of Mean	363.304
Median	50900.000
Mode	41800.000

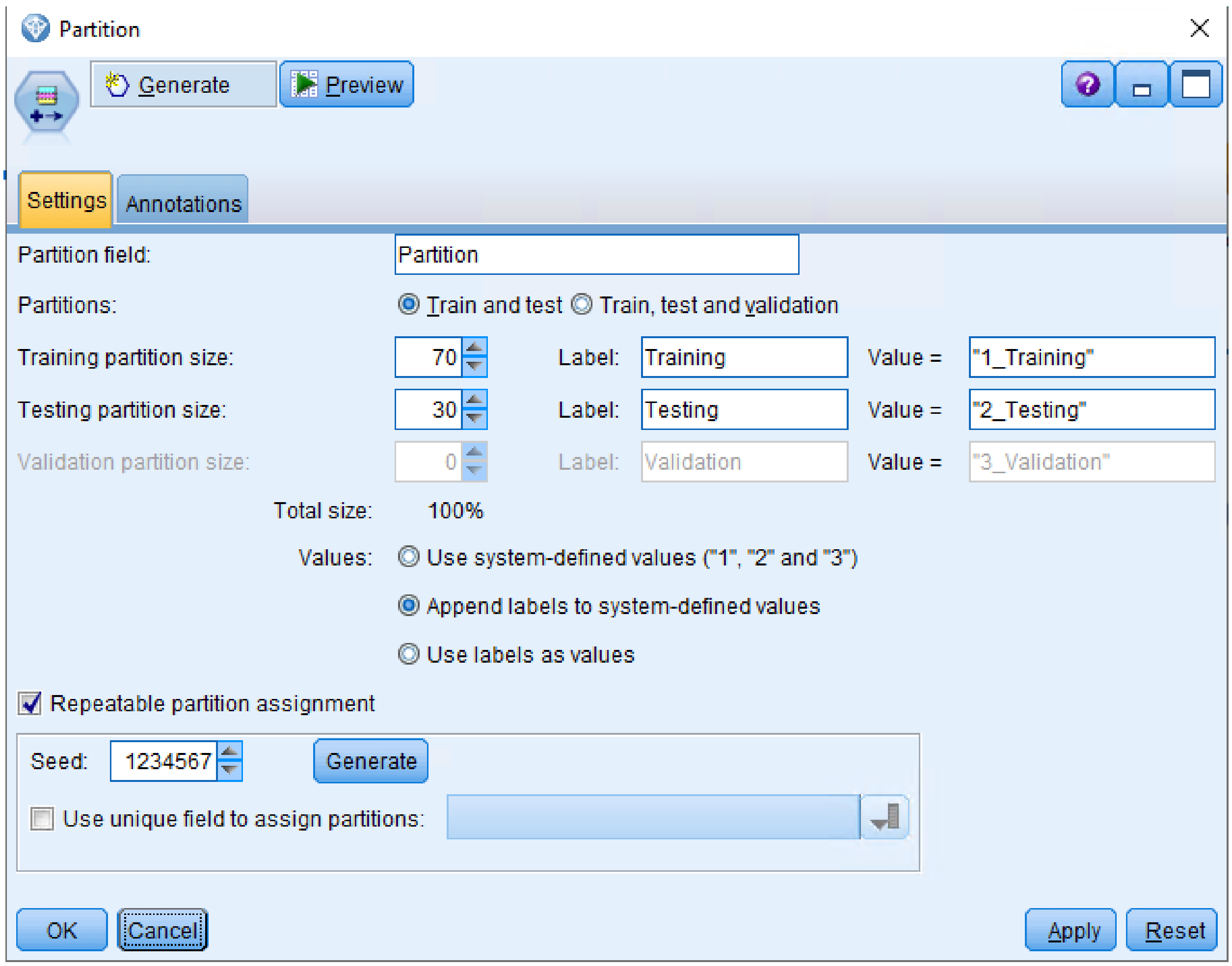
Pearson Correlations

tuition	0.415	Strong
pcttop25	0.550	Strong
sf_ratio	-0.094	Strong
accrate	-0.500	Strong
graduat	0.317	Strong
pct_phd	0.663	Strong
fulltime	0.192	Strong
alumni	0.146	Strong
num_enrl	0.454	Strong
public_private	-0.195	Strong

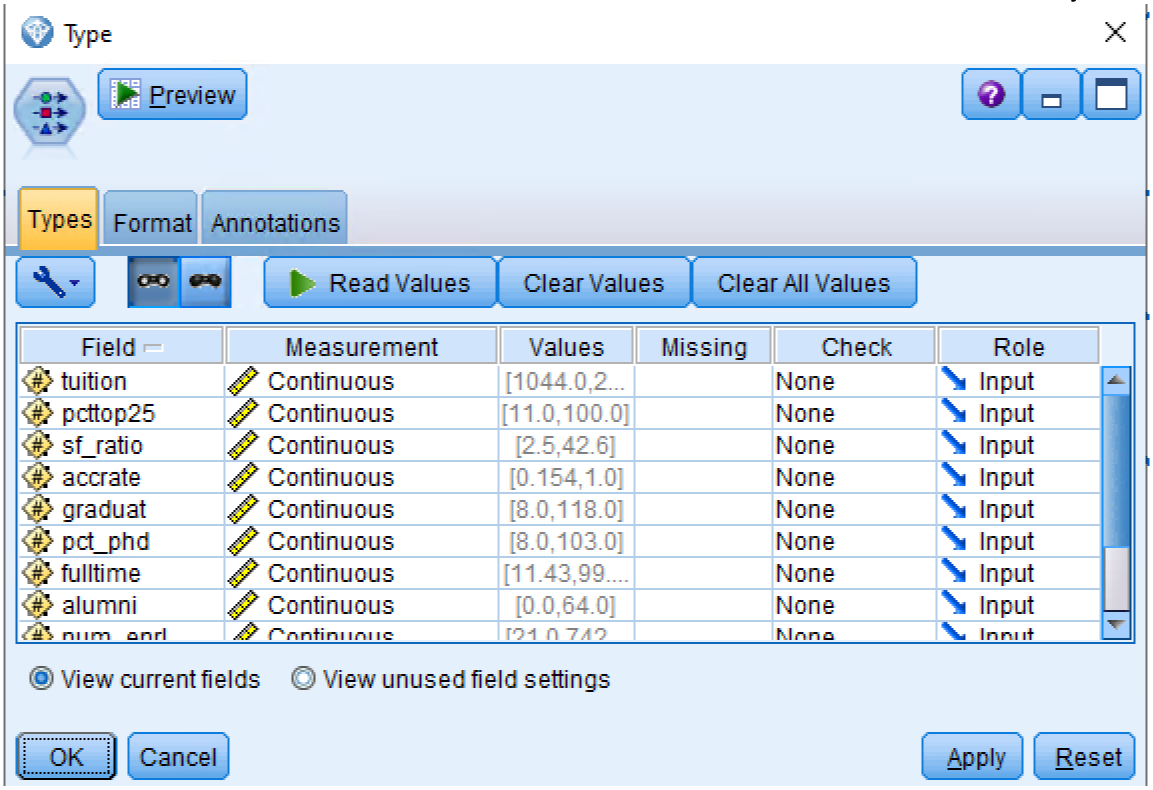
The above we have noticed that we have strong correlation in the fields for College tuition, percentage of new students from the top 25% of high school class, percentage of alumni who donate and in public and private college fields.

6. **Use SPSS Modeler linear regression tool to investigate whether a linear relationship exists between tuition and the other variables. Investigate the differences in the models, if any, among these methods: enter, stepwise, backwards. Construct a table showing method, variables included, statistical tests on regression coefficients, goodness of fit metric(s), predictive accuracy metric on training and test data. Discuss. Which model do you prefer and why?**

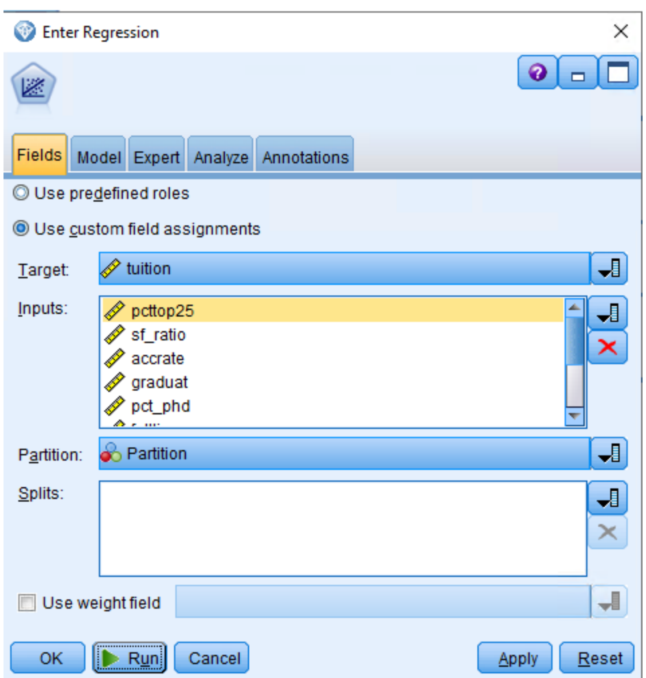
For this part of the question before modeling the data, we have first divided the given data set into 70: 30 ratios for Training and Testing, snapshot has been attached below:

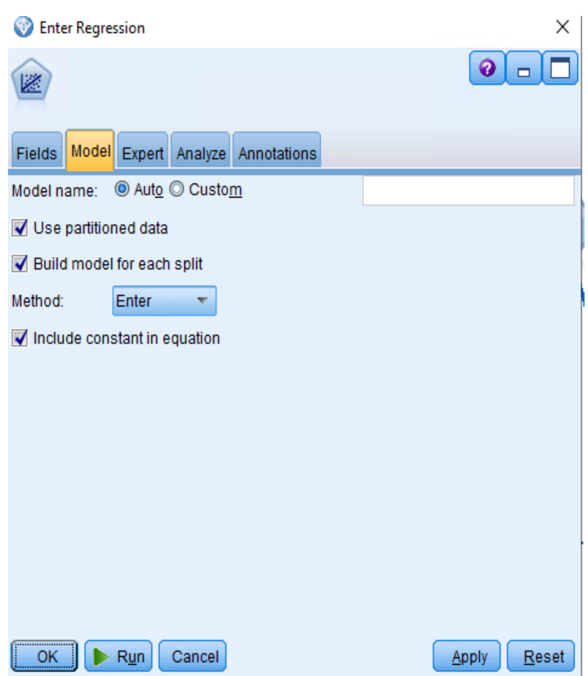


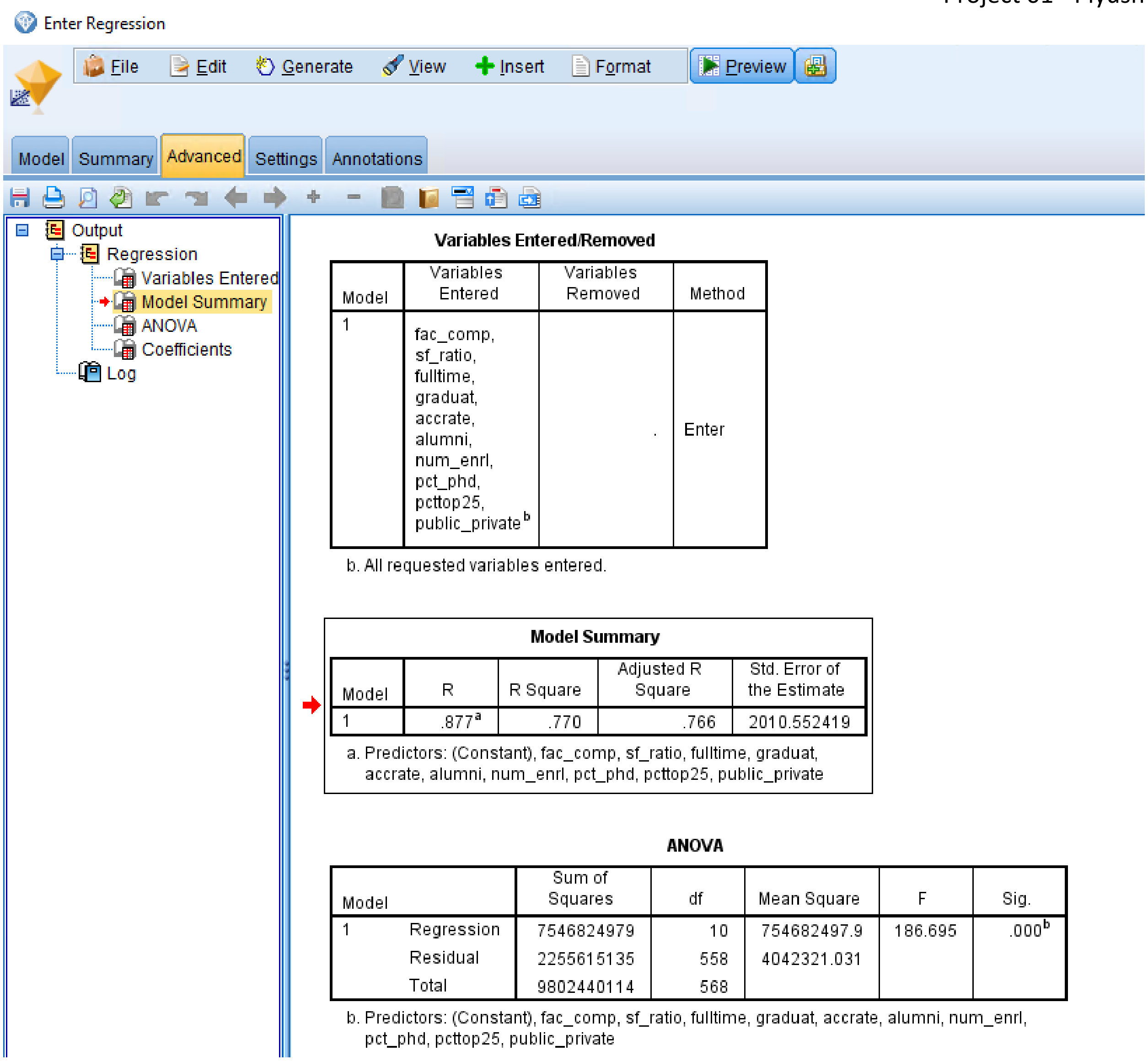
Then we have used the type node to read values of dataset, below is the attached snapshot:



The next step is to perform the Regression from SPSS, we have first have started with Enter Regression the choices for implementing the Enter regression is attached below:







Variables Entered/Removed

Model	Variables Entered	Variables Removed	Method
1	fac_comp, sf_ratio, fulltime, graduat, accrate, alumni, num_enrl, pct_phd, pcttop25, public_private [b]	.	Enter

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.877 [a]	.770	.766	2010.552419

a. Predictors: (Constant), fac_comp, sf_ratio, fulltime, graduat, accrate, alumni, num_enrl, pct_phd, pcttop25, public_private

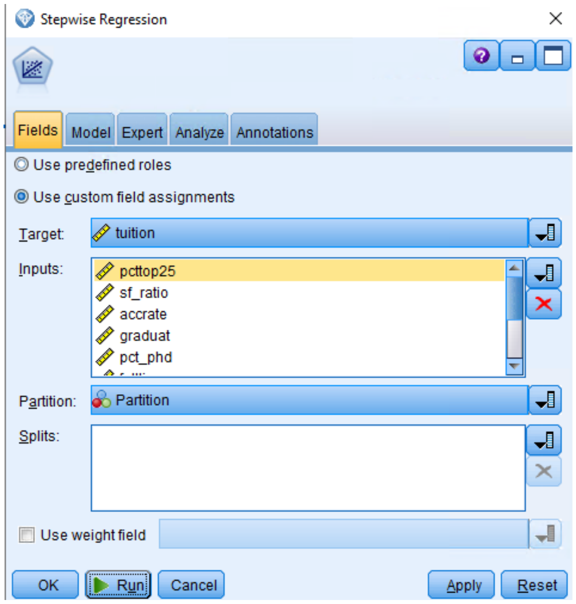
ANOVA

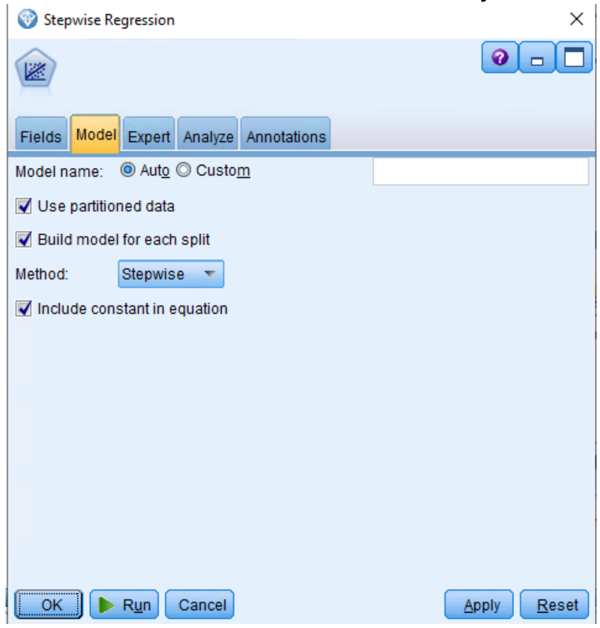
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7546824979	10	754682497.9	186.695	.000 [b]
	Residual	2255615135	558	4042321.031		
	Total	9802440114	568			

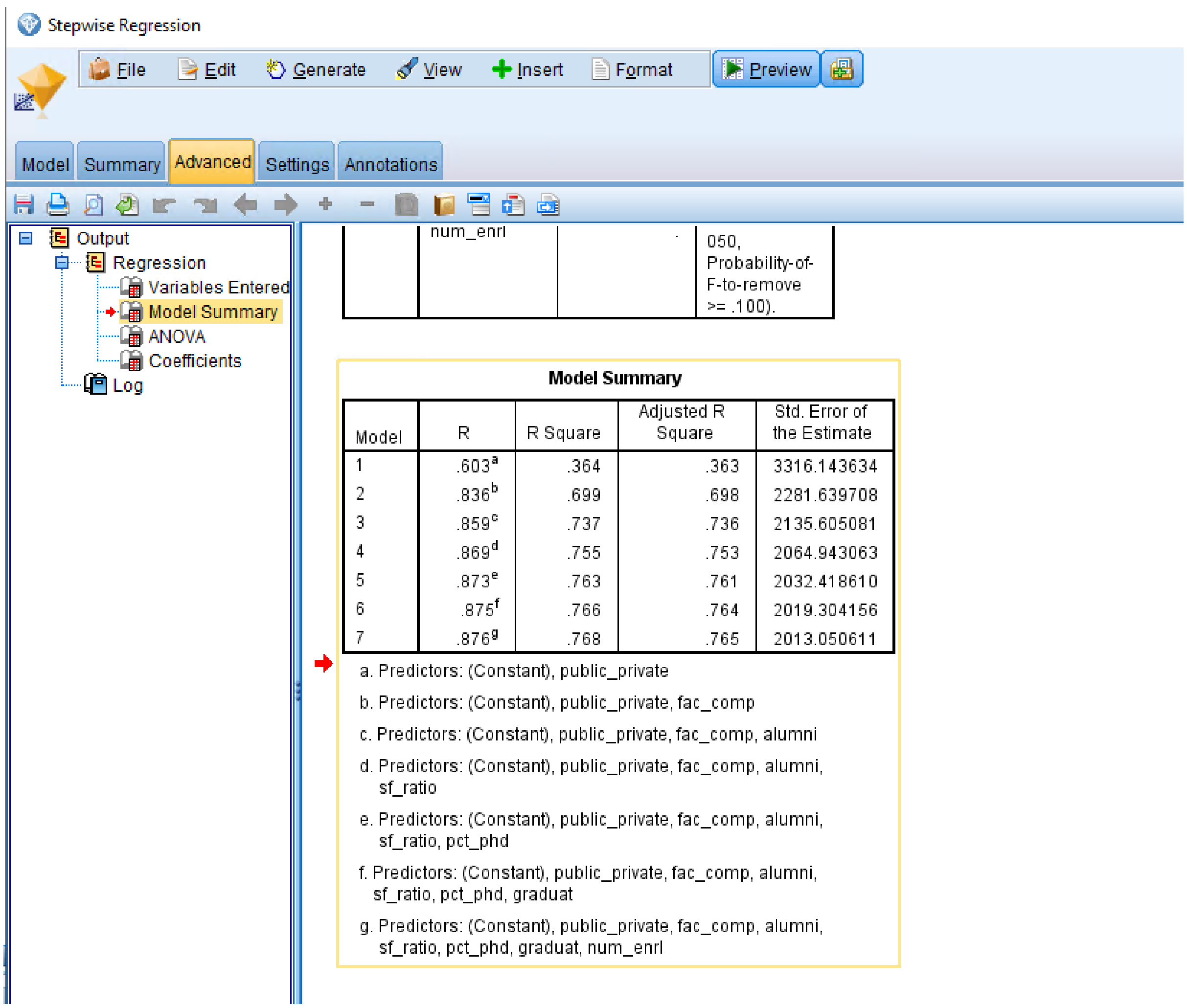
b. Predictors: (Constant), fac_comp, sf_ratio, fulltime, graduat, accrate, alumni, num_enrl, pct_phd, pcttop25, public_private

From the above model we have R value = 87.7, R squared value = 76.6 and F value is 186.695. From the above regression we can also say that the both of the regression coefficients are statically significant. We know that because the significant value obtain from the above regression is 0.000 which indicates that $p < 0.001$.

Now let's perform the Stepwise regression using the same parameter, we have the below Model summary:







Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.603[a]	.364	.363	3316.143634
2	.836[b]	.699	.698	2281.639708
3	.859[c]	.737	.736	2135.605081
4	.869[d]	.755	.753	2064.943063
5	.873[e]	.763	.761	2032.418610
6	.875[f]	.766	.764	2019.304156
7	.876[g]	.768	.765	2013.050611

a. Predictors: (Constant), public_private

b. Predictors: (Constant), public_private, fac_comp

c. Predictors: (Constant), public_private, fac_comp, alumni

d. Predictors: (Constant), public_private, fac_comp, alumni, sf_ratio

e. Predictors: (Constant), public_private, fac_comp, alumni, sf_ratio, pct_phd

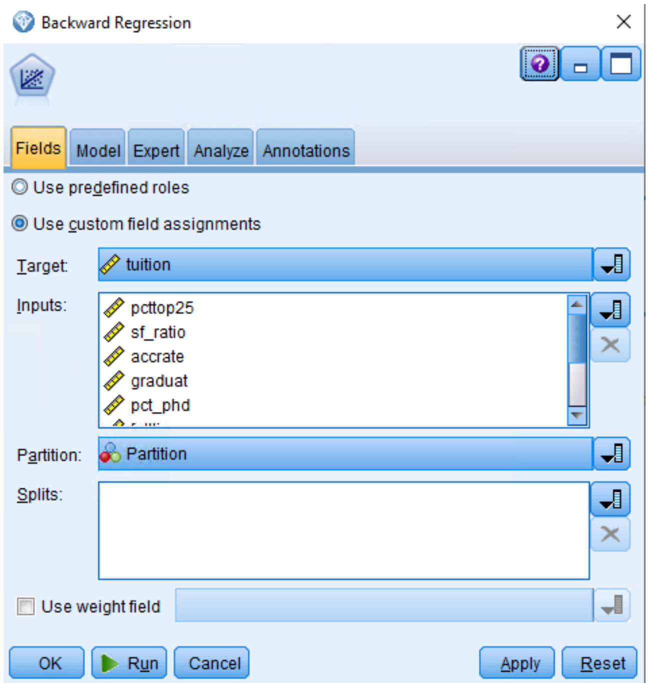
f. Predictors: (Constant), public_private, fac_comp, alumni, sf_ratio, pct_phd, graduat

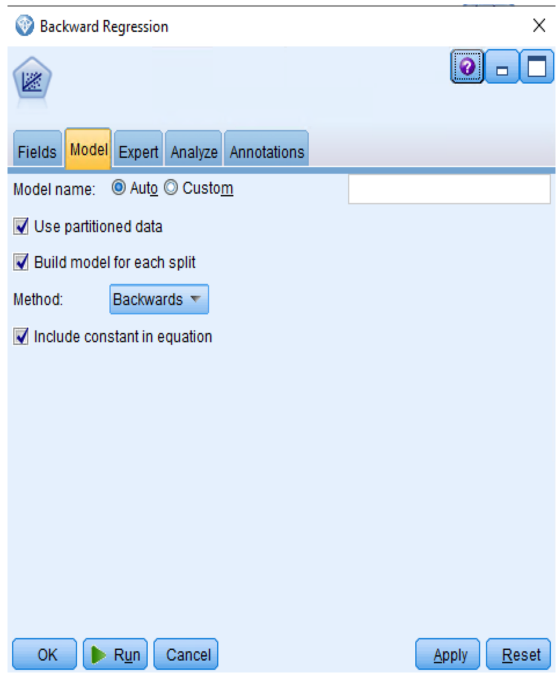
g. Predictors: (Constant), public_private, fac_comp, alumni, sf_ratio, pct_phd, graduat, num_enrl

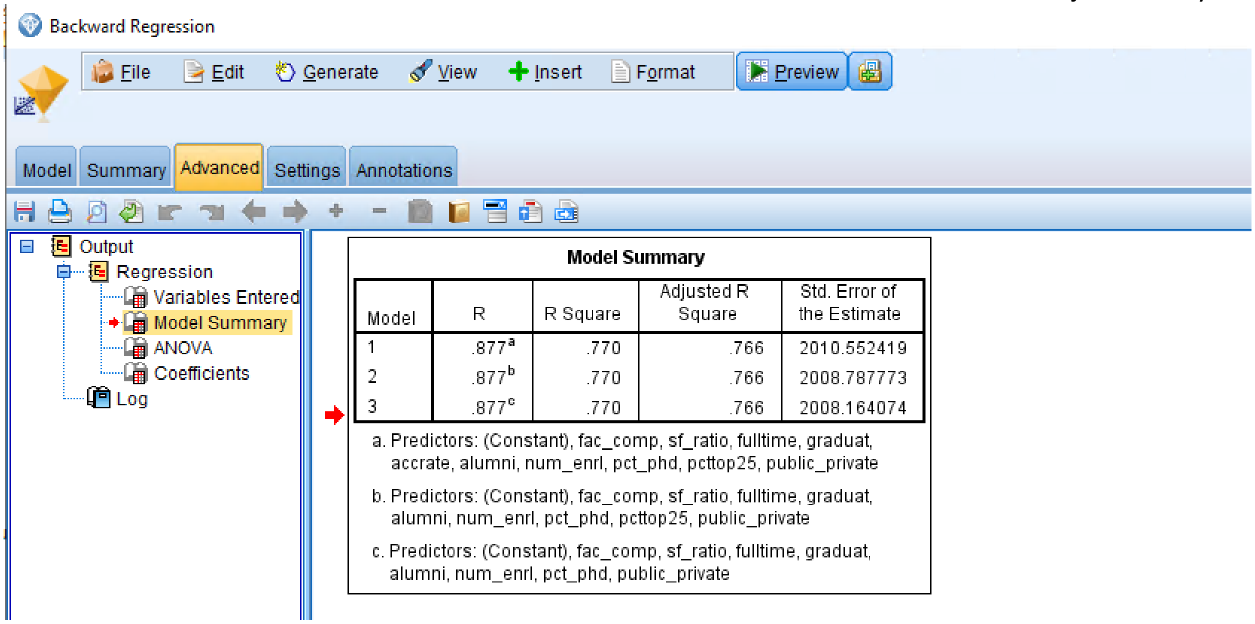
From the above model we have R value = 87.6, R squared value = 76.8 and F value is 186.695. From the above regression we can also say that the both of the regression coefficients are statically significant. We know that because the significant value obtain from the above regression is 0.000 which indicates that $p < 0.001$.

Now let’s perform the backward regression using the same parameter, we have the below Model summary:

Now let’s perform the backward Regression:







Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.877[a]	.770	.766	2010.552419
2	.877[b]	.770	.766	2008.787773
3	.877[c]	.770	.766	2008.164074

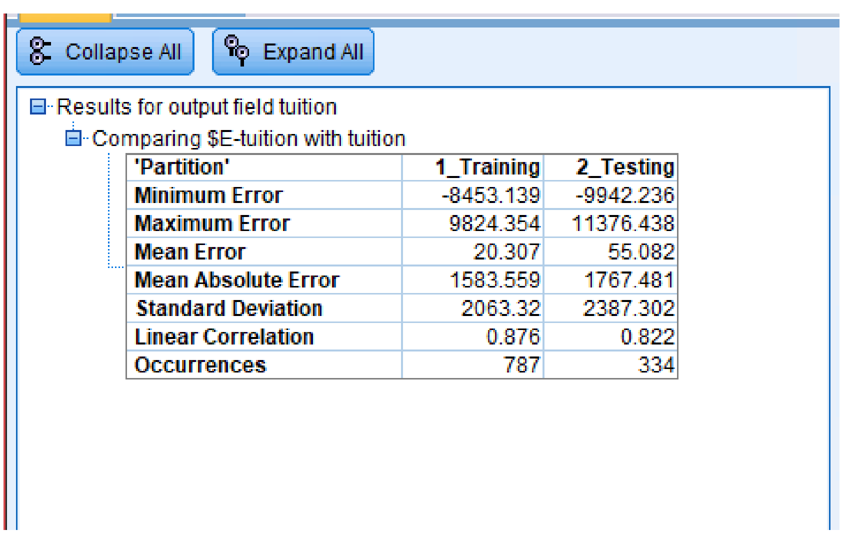
a. Predictors: (Constant), fac_comp, sf_ratio, fulltime, graduat, accrate, alumni, num_enrl, pct_phd, pcttop25, public_private

b. Predictors: (Constant), fac_comp, sf_ratio, fulltime, graduat, alumni, num_enrl, pct_phd, pcttop25, public_private

c. Predictors: (Constant), fac_comp, sf_ratio, fulltime, graduat, alumni, num_enrl, pct_phd, public_private

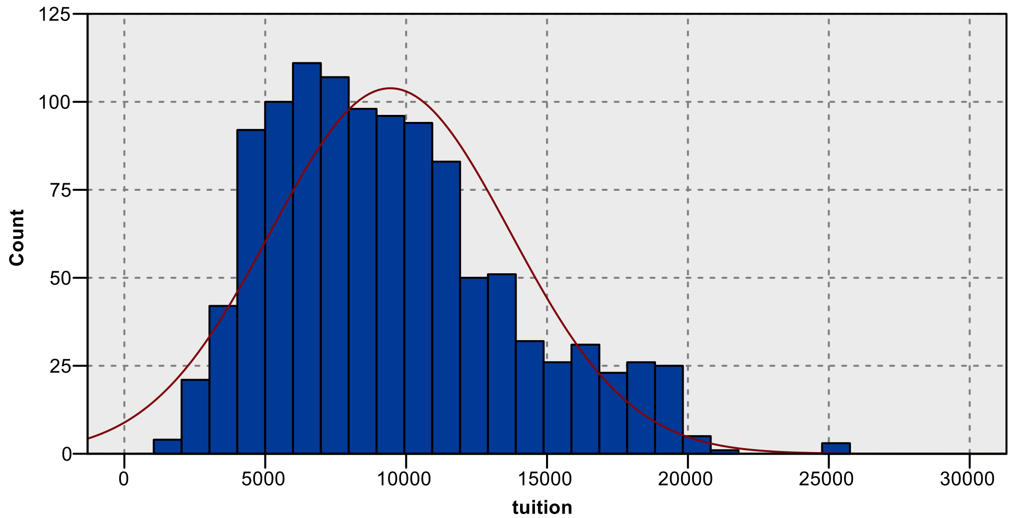
From the above model we have R value = 87.7, R squared value = 77 and F value is 185.95. From the above regression we can also say that the both of the regression coefficients are statically significant. We know that because the significant value obtain from the above regression is 0.000 which indicates that $p < 0.001$.

From the above three models we have similar values for R, R squared value, F and P values. If we have to choose a model we would move forward or select stepwise model, just because of its model processing, as step wise model is estimated on every step and we would select the stepwise model from the above 3 models. Below are the complete details stats for stepwise model.



'Partition'	1_Training	2_Testing
Minimum Error	-8453.139	-9942.236
Maximum Error	9824.354	11376.438
Mean Error	20.307	55.082
Mean Absolute Error	1583.559	1767.481
Standard Deviation	2063.32	2387.302
Linear Correlation	0.876	0.822
Occurrences	787	334

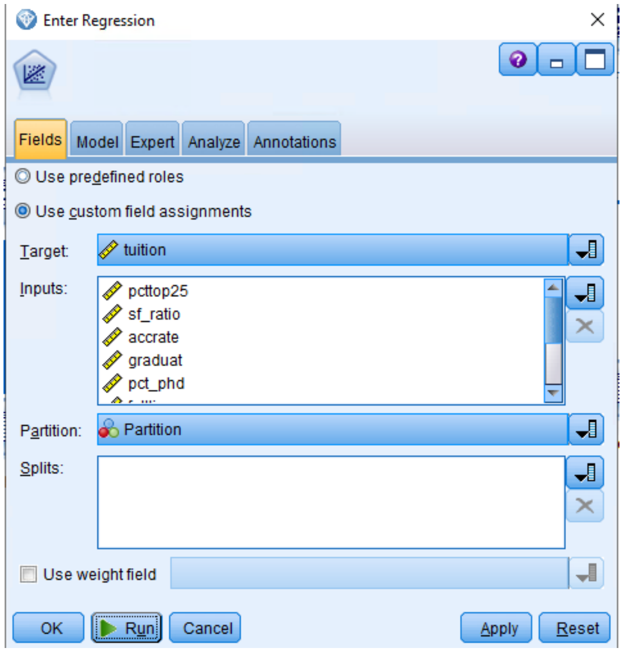
We will discuss more in details about the above stats and graph is below question number 9.

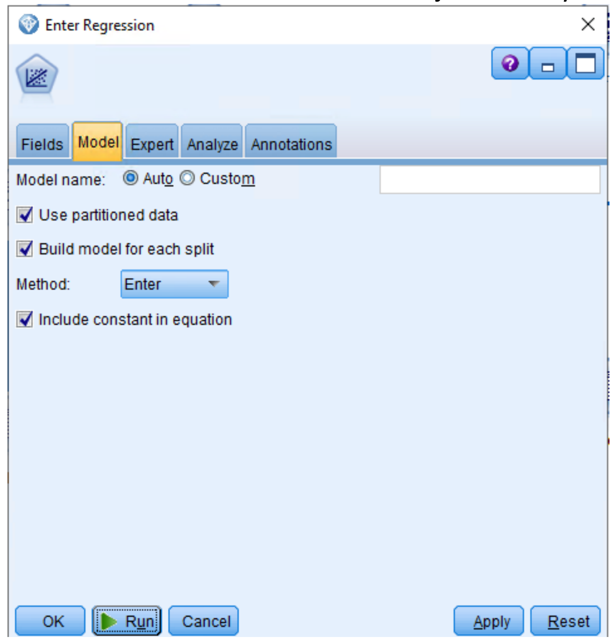


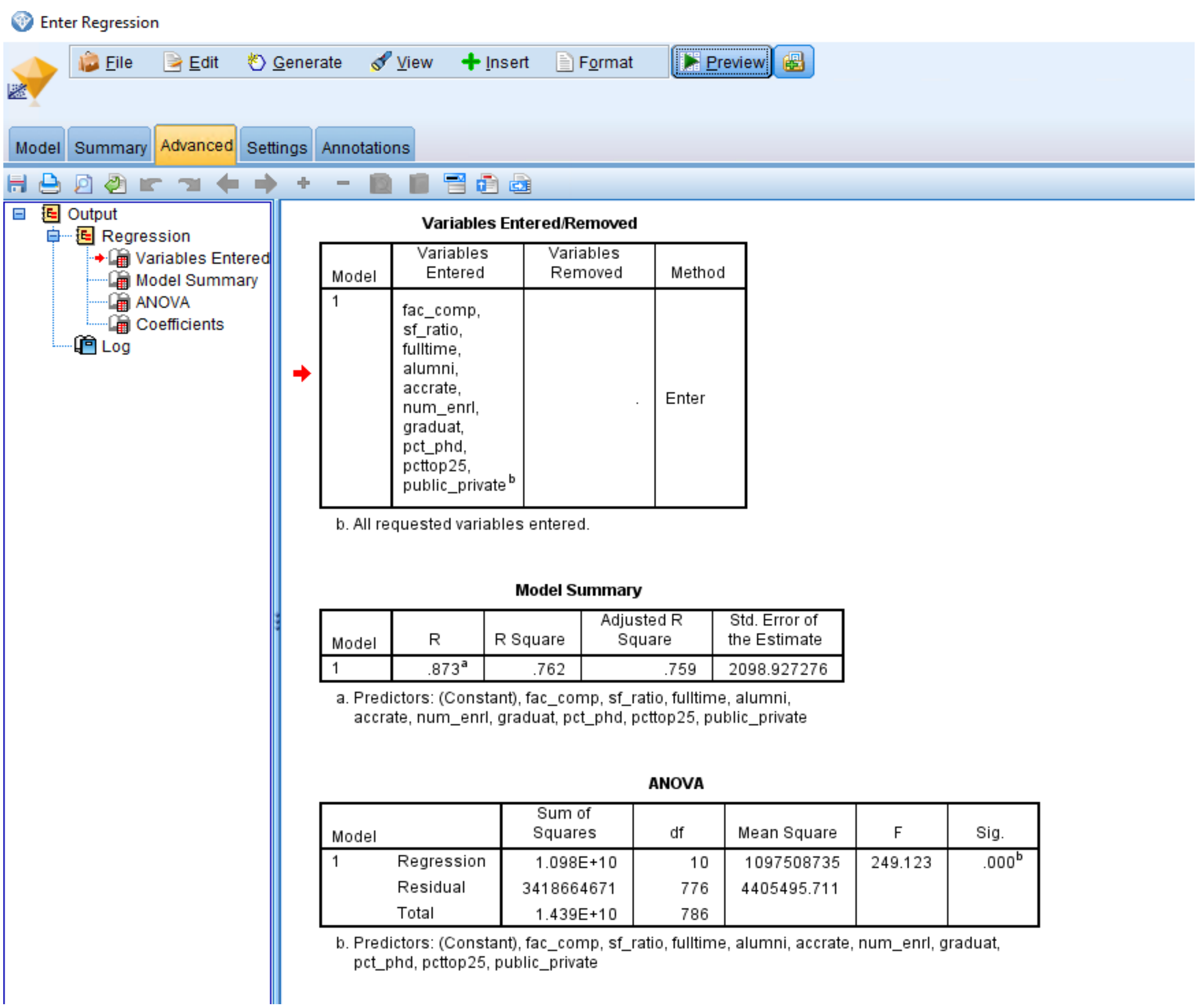
7. **Use SPSS Modeler linear regression tool on the data set where the missing values are each replaced with their field means. Investigate the differences in the models, if any, among these methods: enter, stepwise, backwards. Construct a table showing method, variables included, statistical tests on regression coefficients, goodness off its metric(s), predictive accuracy metric on training and test data. Discuss. Which model do you prefer and why?**

For this part of the question before modeling the data, we have first divided the given data set into 70: 30 ratios for Training and Testing, then will perform the enter, stepwise and backward regressions on our data set after handling the missing data. Below is the snapshot for each regression models.

Starting with Enter Regression:







Variables Entered/Removed

Model	Variables Entered	Variables Removed	Method
1	fac_comp, sf_ratio, fulltime, alumni, accrate, num_enrl, graduat, pct_phd, pcttop25, public_private[b]	.	Enter

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.873[a]	.762	.759	2098.927276

a. Predictors: (Constant), fac_comp, sf_ratio, fulltime, alumni, accrate, num_enrl, graduat, pct_phd, pcttop25, public_private

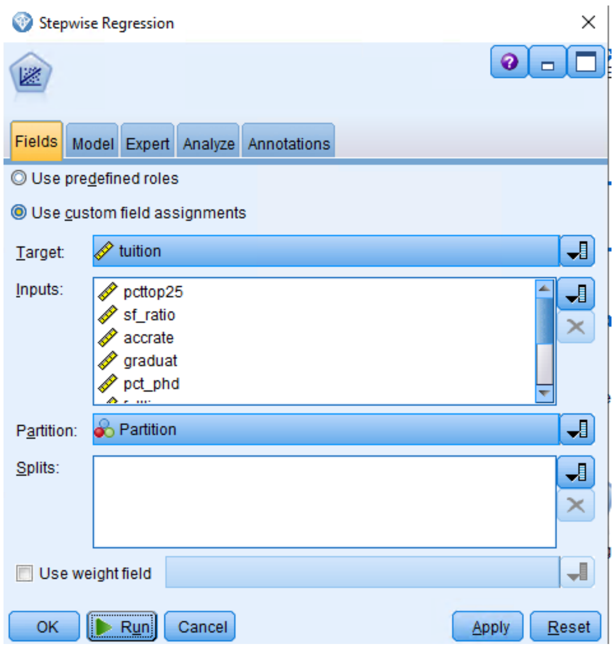
ANOVA

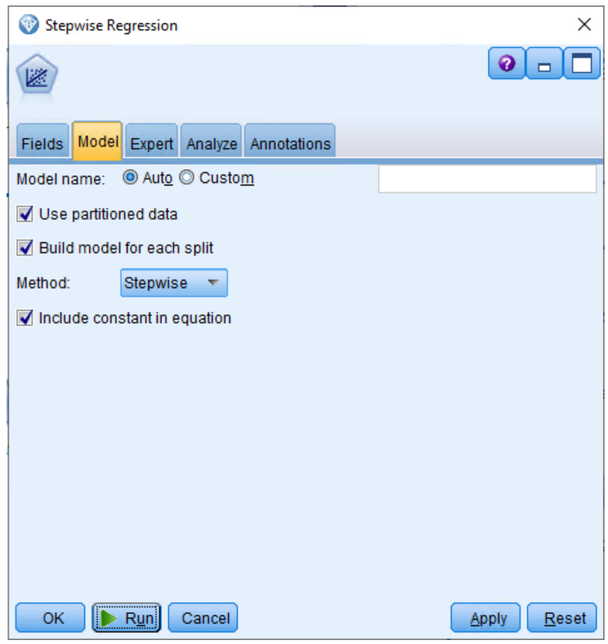
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.098E+10	10	1097508735	249.123	.000[b]
	Residual	3418664671	776	4405495.711		
	Total	1.439E+10	786			

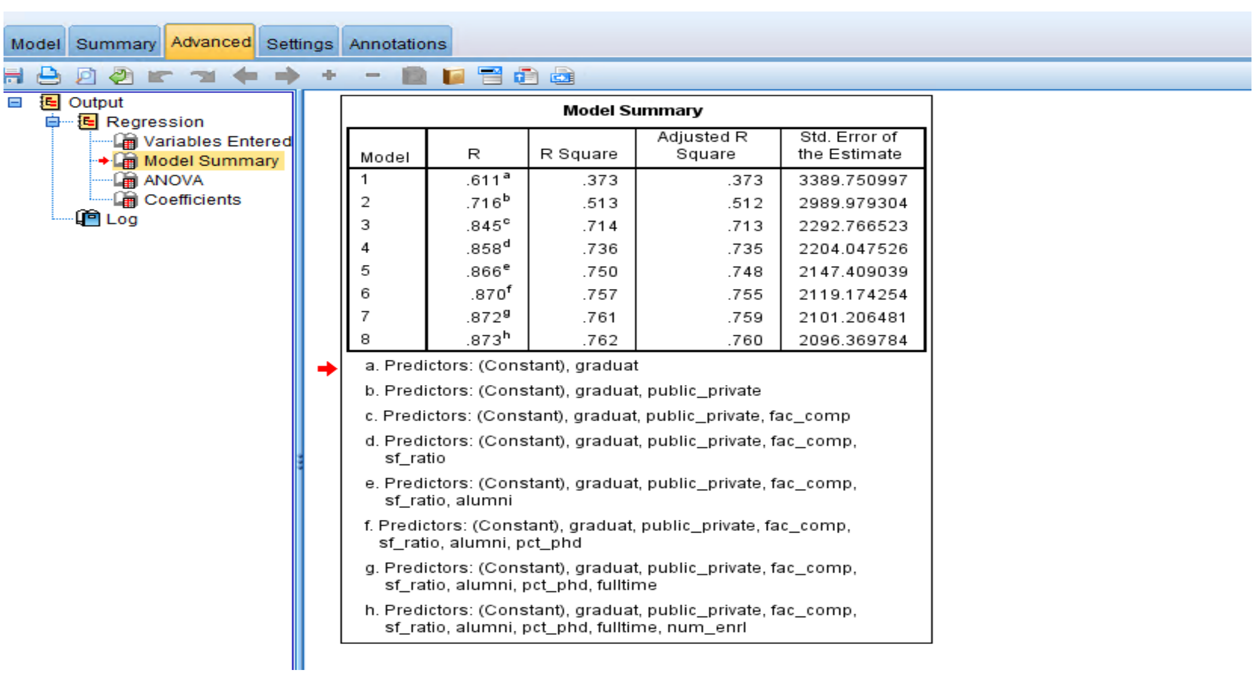
b. Predictors: (Constant), fac_comp, sf_ratio, fulltime, alumni, accrate, num_enrl, graduat, pct_phd, pcttop25, public_private

From the above model we have R value = 87.3, R squared value = 75.9 and F value is 249.123. From the above regression we can also say that the both of the regression coefficients are statically significant. We know that because the significant value obtain from the above regression is 0.000 which indicates that $p < 0.001$.

Stepwise Regression:







Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.611[a]	.373	.373	3389.750997
2	.716[b]	.513	.512	2989.979304
3	.845[c]	.714	.713	2292.766523
4	.858[d]	.736	.735	2204.047526
5	.866[e]	.750	.748	2147.409039
6	.870[f]	.757	.755	2119.174254
7	.872[g]	.761	.759	2101.206481
8	.873[h]	.762	.760	2096.369784

a. Predictors: (Constant), graduat

b. Predictors: (Constant), graduat, public_private

c. Predictors: (Constant), graduat, public_private, fac_comp

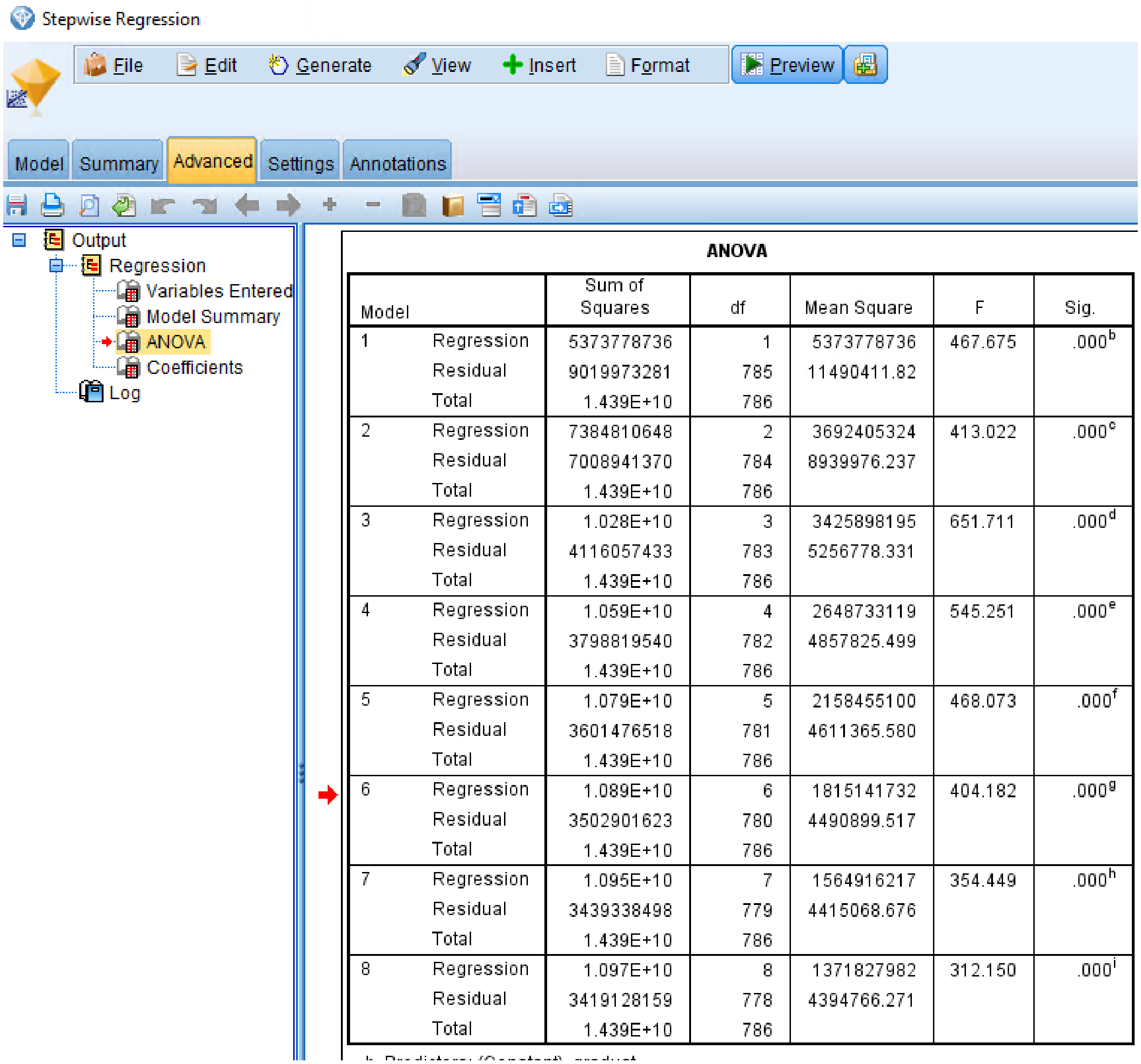
d. Predictors: (Constant), graduat, public_private, fac_comp, sf_ratio

e. Predictors: (Constant), graduat, public_private, fac_comp, sf_ratio, alumni

f. Predictors: (Constant), graduat, public_private, fac_comp, sf_ratio, alumni, pct_phd

g. Predictors: (Constant), graduat, public_private, fac_comp, sf_ratio, alumni, pct_phd, fulltime

h. Predictors: (Constant), graduat, public_private, fac_comp, sf_ratio, alumni, pct_phd, fulltime, num_enrl

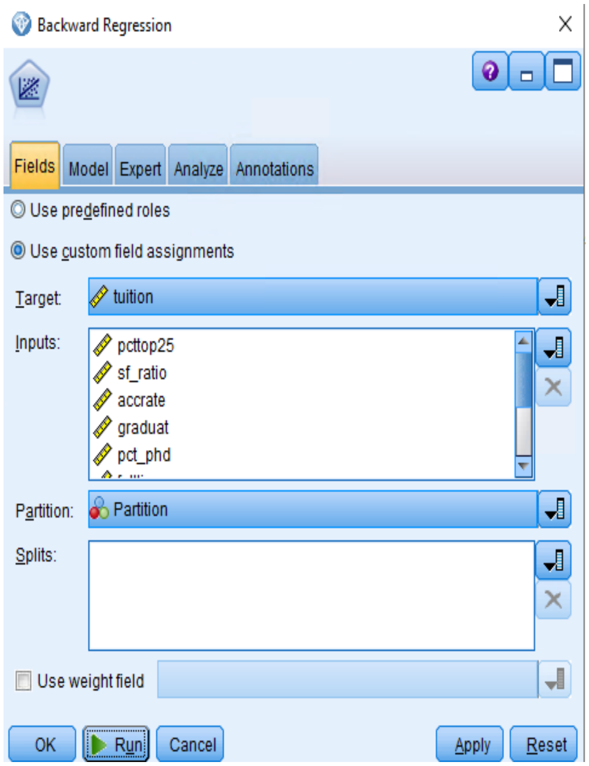


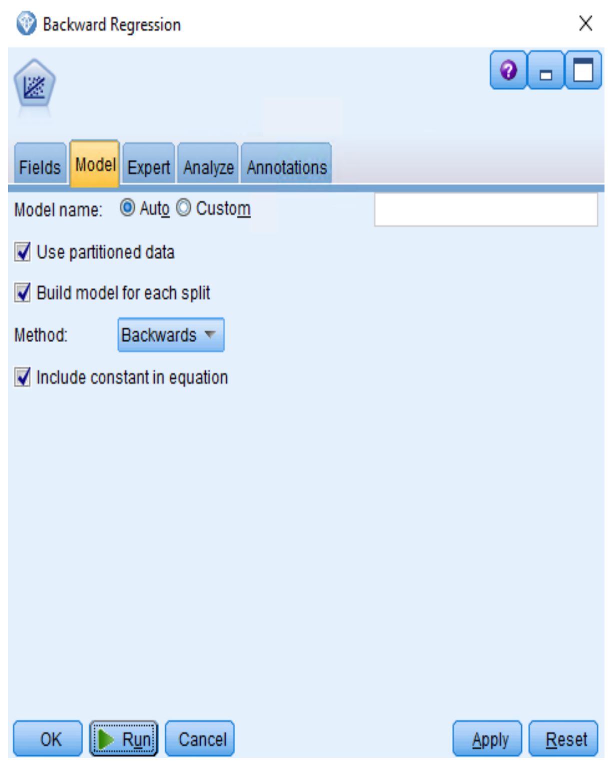
ANOVA

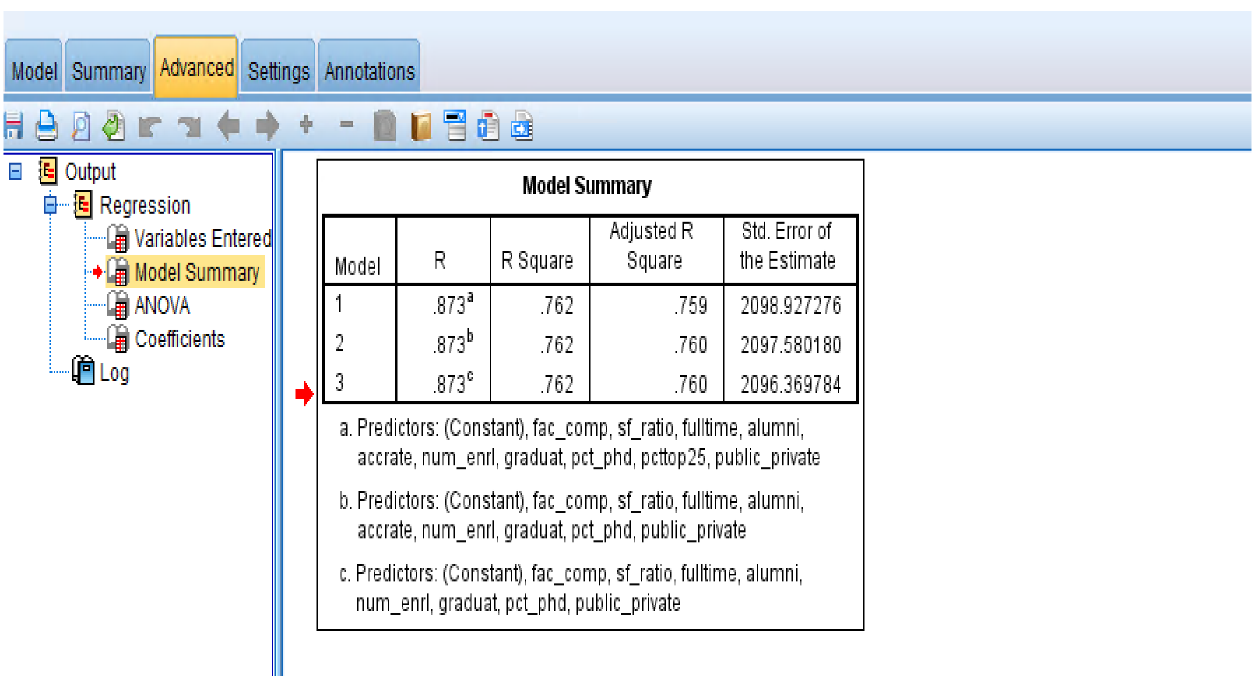
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5373778736	1	5373778736	467.675	.000[b]
	Residual	9019973281	785	11490411.82		
	Total	1.439E+10	786			
2	Regression	7384810648	2	3692405324	413.022	.000[c]
	Residual	7008941370	784	8939976.237		
	Total	1.439E+10	786			
3	Regression	1.028E+10	3	3425898195	651.711	.000[d]
	Residual	4116057433	783	5256778.331		
	Total	1.439E+10	786			
4	Regression	1.059E+10	4	2648733119	545.251	.000[e]
	Residual	3798819540	782	4857825.499		
	Total	1.439E+10	786			
5	Regression	1.079E+10	5	2158455100	468.073	.000[f]
	Residual	3601476518	781	4611365.580		
	Total	1.439E+10	786			
6	Regression	1.089E+10	6	1815141732	404.182	.000[g]
	Residual	3502901623	780	4490899.517		
	Total	1.439E+10	786			
7	Regression	1.095E+10	7	1564916217	354.449	.000[h]
	Residual	3439338498	779	4415068.676		
	Total	1.439E+10	786			
8	Regression	1.097E+10	8	1371827982	312.150	.000[i]
	Residual	3419128159	778	4394766.271		
	Total	1.439E+10	786			

From the above Stepwise Regression model, we have R value = 87.3, R squared value = 76.2 and F value is 312.150. From the above regression we can also say that the both of the regression coefficients are statically significant. We know that because the significant value obtain from the above regression is 0.000 which indicates that $p < 0.001$

Backward Regression:







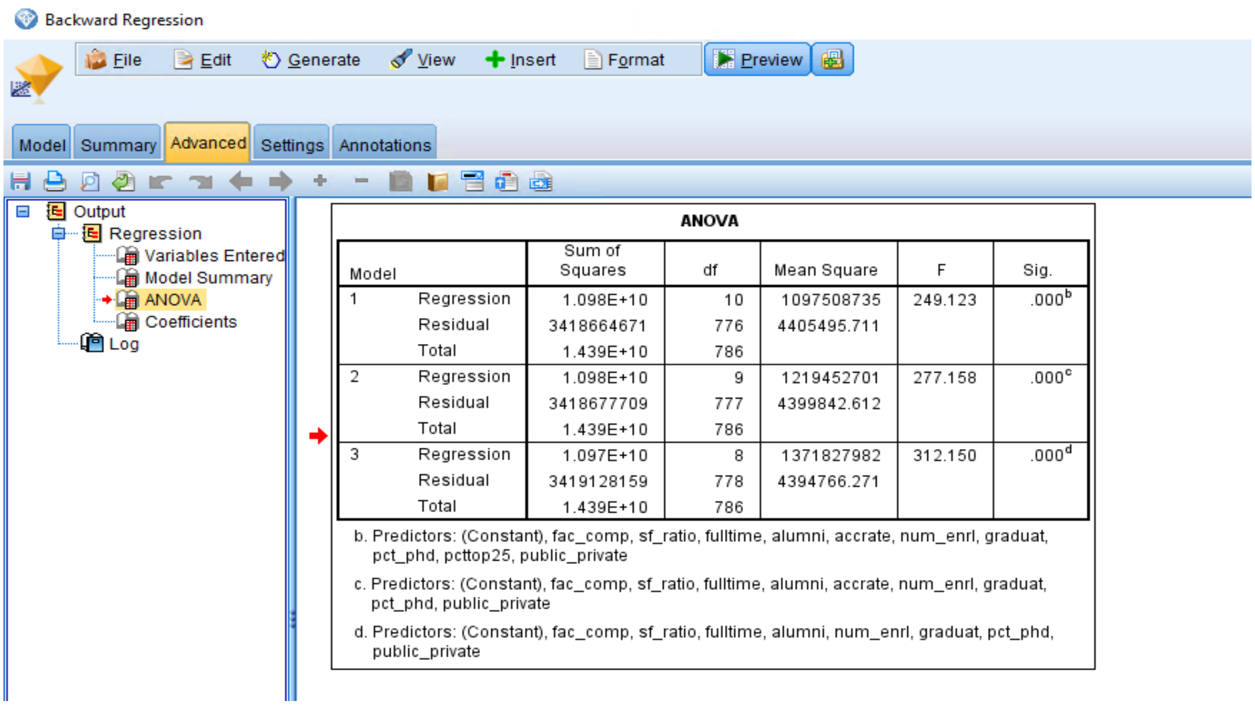
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.873[a]	.762	.759	2098.927276
2	.873[b]	.762	.760	2097.580180
3	.873[c]	.762	.760	2096.369784

a. Predictors: (Constant), fac_comp, sf_ratio, fulltime, alumni, accrate, num_enrl, graduat, pct_phd, pcttop25, public_private

b. Predictors: (Constant), fac_comp, sf_ratio, fulltime, alumni, accrate, num_enrl, graduat, pct_phd, public_private

c. Predictors: (Constant), fac_comp, sf_ratio, fulltime, alumni, num_enrl, graduat, pct_phd, public_private



ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.098E+10	10	1097508735	249.123	.000[b]
	Residual	3418664671	776	4405495.711		
	Total	1.439E+10	786			
2	Regression	1.098E+10	9	1219452701	277.158	.000[c]
	Residual	3418677709	777	4399842.612		
	Total	1.439E+10	786			
3	Regression	1.097E+10	8	1371827982	312.150	.000[d]
	Residual	3419128159	778	4394766.271		
	Total	1.439E+10	786			

b. Predictors: (Constant), fac_comp, sf_ratio, fulltime, alumni, accrate, num_enrl, graduat, pct_phd, pcttop25, public_private

c. Predictors: (Constant), fac_comp, sf_ratio, fulltime, alumni, accrate, num_enrl, graduat, pct_phd, public_private

d. Predictors: (Constant), fac_comp, sf_ratio, fulltime, alumni, num_enrl, graduat, pct_phd, public_private

From the above backward Regression model, we have R value = 87.7, R squared value = 76.2 and F value is 312.150. From the above regression we can also say that the both of the regression coefficients are statically significant. We know that because the significant value obtain from the above regression is 0.000 which indicates that $p < 0.001$.

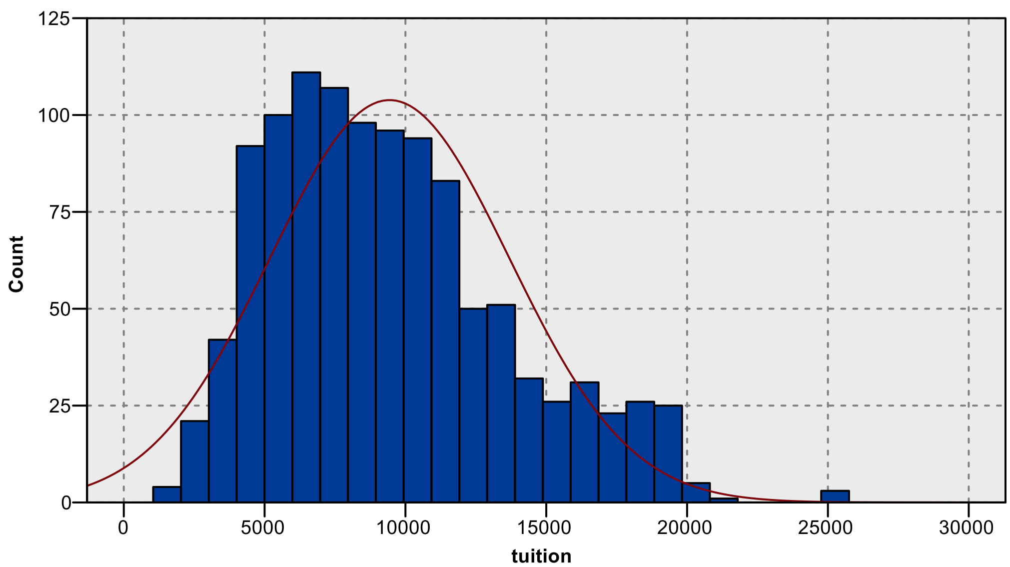
From the above three models we have similar values for R, R squared value, F and P values. If we have to choose a model we would move forward or select stepwise model, just because of its model processing, as step wise model is estimated on every step and we would select the stepwise model from the above 3 models. Below are the complete details stats for stepwise model.

Collapse All Expand All

Results for output field tuition

Comparing $E-tuition with tuition

'Partition'	1_Training	2_Testing
Minimum Error	-9124.261	-10284.888
Maximum Error	10129.296	10969.061
Mean Error	-0.0	7.388
Mean Absolute Error	1591.721	1746.892
Standard Deviation	2085.674	2307.435
Linear Correlation	0.873	0.831
Occurrences	787	334



8. Compare the best model in 6 and the best model in 7. Which model do you prefer and why?

After comparing two models which are stepwise model with missing values and stepwise regression model without missing values, if we have the choose a model we would choose the model with missing values just because taking a mean for the important features for predicting a model, and we had approximately 76 % complete data set, rest of them are with missing values. If the missing values feature we less than 5% then we might would have choose Regression model with after handling the missing data. But for the given conditions and data set, we would go with the Stepwise Regression model with missing values as it would be best the best model for predicting.

9. **For the final (chosen) model:**
   a) **Write out the estimated regression equation and explain the meaning of the coefficients**

Tuition = (-3774.1) +

sf_ratio * -154.5 +

Graduate* 18.4 +

Pct_phd * 25.46 +

Fulltime * 19.94 +

Alumni * 36.9 +

Num_enrl * - 0.2562 +

Public_private * 4416.5 +

Fac_comp * 0.1464 +

- An intercept of -3774.1 has no interpretation as it would be the tuition fee of university based on different factors included in the data set.
- A slope of -154.5 in sf-ratio means that with each increase in tuition fee the ratio of student to faculty reduces by 154.5.
- A slope of 18.4 in graduate means that with increase in tuition fee the percentage of students who graduated increases by 18.4%.
- A slope of 25.46 in pct_phd means that with increase in tuition fee the percentage of faculty with PHD's increases by 25.46%.
- A slope of 19.94 in fulltime means that with increase in tuition fee the percentage of undergraduates who are full time students increases by 19.94%.
- A slope of 36.9 in alumni means that with increase in tuition fee the percentage of alumni who donate increases by 36.9%.

- A slope of (-0.2562) in num_enrl means that with increase in tuition fee the number of new students who enrolled decreases by 0.2562.

- A slope of (4416.5) in public_private has no interpretation as it would be the tuition fee of university based on public = 0 and private = 1, from the graphical representation in question 3, we have seen that private universities have more tuition fee as compared with to public universities.

- A slope of 36.9 in alumni means that with increase in tuition fee the percentage of alumni who donate increases by 36.9%.

- A slope of 0.146 in fac_comp means that with the increase in tuition fee the average compensation of faculty increases by 0.146.

b) Provide a full report of the chosen regression model and report its metrics (goodness of fit, predictive performance) and statistics on training and test data Make sure you tweak your models to get the best performance. Use 70/30 partition in all cases

Below is the complete metrics report of the chosen model

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.611[a]	.373	.373	3389.750997
2	.716[b]	.513	.512	2989.979304
3	.845[c]	.714	.713	2292.766523
4	.858[d]	.736	.735	2204.047526
5	.866[e]	.750	.748	2147.409039
6	.870[f]	.757	.755	2119.174254
7	.872[g]	.761	.759	2101.206481
8	.873[h]	.762	.760	2096.369784

a. Predictors: (Constant), graduat

b. Predictors: (Constant), graduat, public_private

c. Predictors: (Constant), graduat, public_private, fac_comp

d. Predictors: (Constant), graduat, public_private, fac_comp, sf_ratio

e. Predictors: (Constant), graduat, public_private, fac_comp, sf_ratio, alumni

f. Predictors: (Constant), graduat, public_private, fac_comp, sf_ratio, alumni, pct_phd

g. Predictors: (Constant), graduat, public_private, fac_comp, sf_ratio, alumni, pct_phd, fulltime

h. Predictors: (Constant), graduat, public_private, fac_comp, sf_ratio, alumni, pct_phd, fulltime, num_enrl

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5373778736.398	1	5373778736.398	467.675	.000[b]
	Residual	9019973281.414	785	11490411.823		
	Total	14393752017.812	786			
2	Regression	7384810647.687	2	3692405323.844	413.022	.000[c]
	Residual	7008941370.125	784	8939976.237		
	Total	14393752017.812	786			
3	Regression	10277694584.560	3	3425898194.853	651.711	.000[d]
	Residual	4116057433.252	783	5256778.331		
	Total	14393752017.812	786			
4	Regression	10594932477.866	4	2648733119.467	545.251	.000[e]
	Residual	3798819539.946	782	4857825.499		
	Total	14393752017.812	786			
5	Regression	10792275500.092	5	2158455100.018	468.073	.000[f]
	Residual	3601476517.720	781	4611365.580		
	Total	14393752017.812	786			
6	Regression	10890850394.600	6	1815141732.433	404.182	.000[g]
	Residual	3502901623.212	780	4490899.517		
	Total	14393752017.812	786			
7	Regression	10954413519.574	7	1564916217.082	354.449	.000[h]
	Residual	3439338498.238	779	4415068.676		
	Total	14393752017.812	786			
8	Regression	10974623858.588	8	1371827982.324	312.150	.000[i]
	Residual	3419128159.224	778	4394766.271		
	Total	14393752017.812	786			

b. Predictors: (Constant), graduat

c. Predictors: (Constant), graduat, public_private

d. Predictors: (Constant), graduat, public_private, fac_comp

e. Predictors: (Constant), graduat, public_private, fac_comp, sf_ratio

f. Predictors: (Constant), graduat, public_private, fac_comp, sf_ratio, alumni

g. Predictors: (Constant), graduat, public_private, fac_comp, sf_ratio, alumni, pct_phd

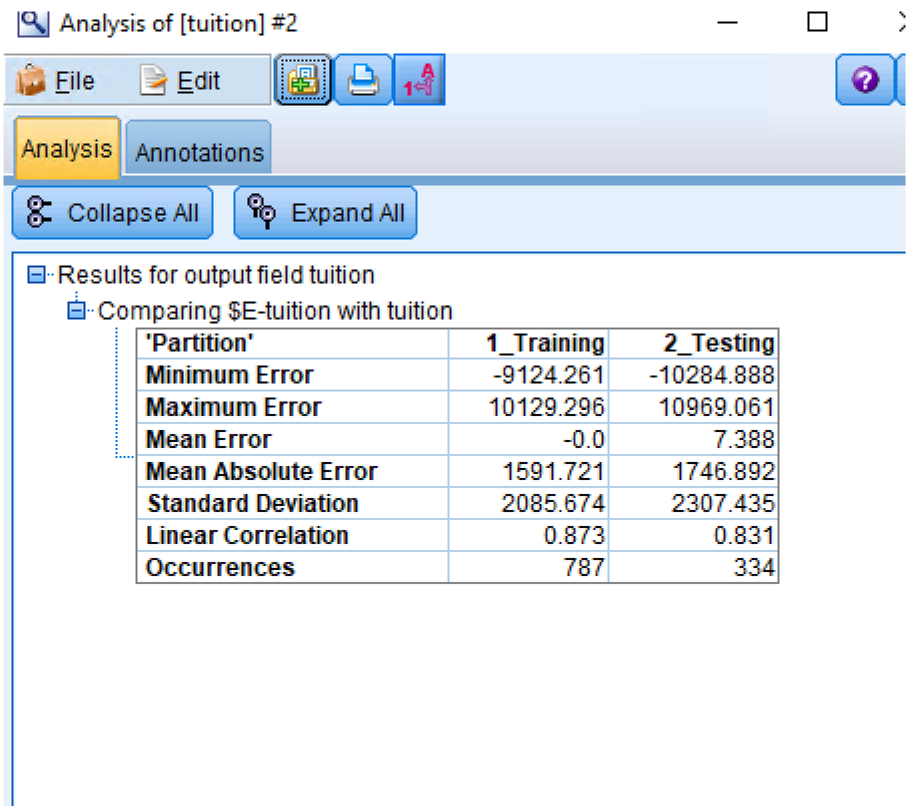
h. Predictors: (Constant), graduat, public_private, fac_comp, sf_ratio, alumni, pct_phd, fulltime

i. Predictors: (Constant), graduat, public_private, fac_comp, sf_ratio, alumni, pct_phd, fulltime, num_enrl

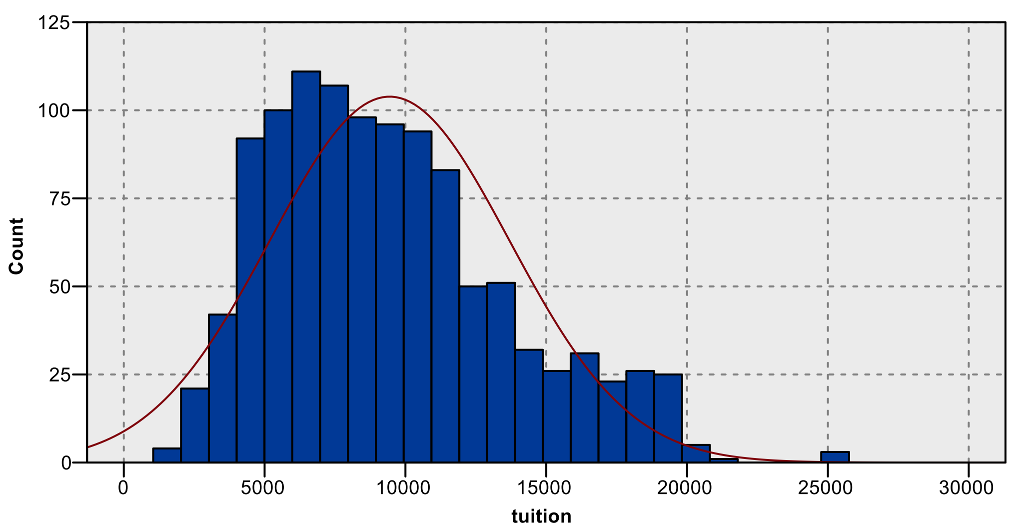
Statistical tests on Regression Coefficients for Stepwise Method:

Coefficients

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	674.471	427.046		1.579	.115
	graduat	143.565	6.639	.611	21.626	.000
2	(Constant)	1120.596	377.855		2.966	.003
	graduat	98.652	6.577	.420	15.000	.000
	public_private	3711.486	247.461	.420	14.998	.000
3	(Constant)	-5533.657	405.480		-13.647	.000
	graduat	36.074	5.705	.154	6.323	.000
	public_private	5557.363	205.425	.629	27.053	.000
	fac_comp	.178	.008	.515	23.459	.000
4	(Constant)	-1740.869	610.095		-2.853	.004
	graduat	31.939	5.508	.136	5.798	.000
	public_private	4797.264	218.732	.543	21.932	.000
	fac_comp	.168	.007	.485	22.668	.000
	sf_ratio	-172.209	21.310	-.177	-8.081	.000
5	(Constant)	-2218.834	598.890		-3.705	.000
	graduat	22.244	5.568	.095	3.995	.000
	public_private	4531.614	216.945	.513	20.888	.000
	fac_comp	.165	.007	.476	22.802	.000
	sf_ratio	-148.960	21.064	-.153	-7.072	.000
	alumni	49.137	7.511	.141	6.542	.000
6	(Constant)	-2745.366	601.606		-4.563	.000
	graduat	21.146	5.499	.090	3.845	.000
	public_private	4623.430	214.988	.523	21.506	.000
	fac_comp	.141	.009	.407	16.057	.000
	sf_ratio	-151.408	20.794	-.156	-7.281	.000
	alumni	42.552	7.545	.122	5.640	.000
	pct_phd	27.968	5.970	.113	4.685	.000
7	(Constant)	-3636.337	641.060		-5.672	.000
	graduat	17.561	5.534	.075	3.173	.002
	public_private	4636.095	213.191	.524	21.746	.000
	fac_comp	.141	.009	.407	16.183	.000
	sf_ratio	-156.528	20.662	-.161	-7.576	.000
	alumni	39.104	7.536	.112	5.189	.000
	pct_phd	24.776	5.978	.100	4.144	.000
	fulltime	18.722	4.934	.072	3.794	.000
8	(Constant)	-3774.122	642.803		-5.871	.000
	graduat	18.400	5.535	.078	3.324	.001
	public_private	4416.536	236.059	.500	18.709	.000
	fac_comp	.146	.009	.424	16.126	.000
	sf_ratio	-154.500	20.636	-.159	-7.487	.000
	alumni	36.902	7.588	.106	4.863	.000
	pct_phd	25.459	5.973	.103	4.262	.000
	fulltime	19.941	4.956	.076	4.024	.000
	num_enrl	-.256	.119	-.052	-2.144	.032

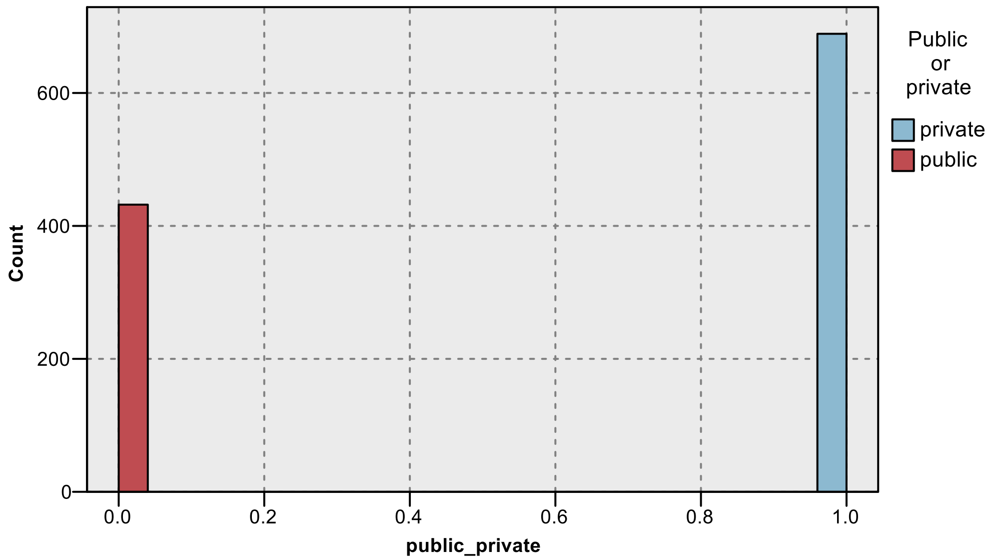


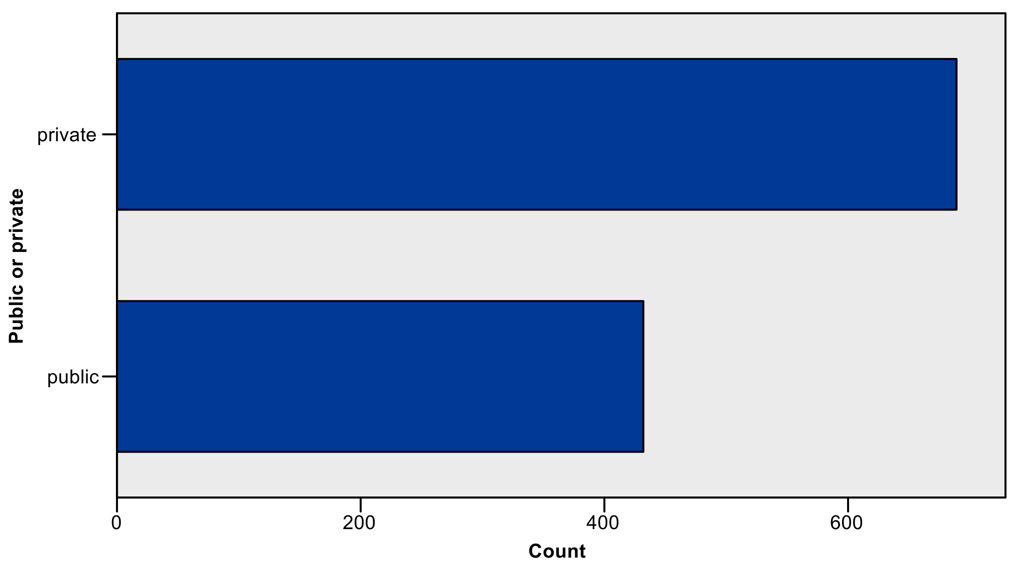
'Partition'	1_Training	2_Testing
Minimum Error	-9124.261	-10284.888
Maximum Error	10129.296	10969.061
Mean Error	-0.0	7.388
Mean Absolute Error	1591.721	1746.892
Standard Deviation	2085.674	2307.435
Linear Correlation	0.873	0.831
Occurrences	787	334



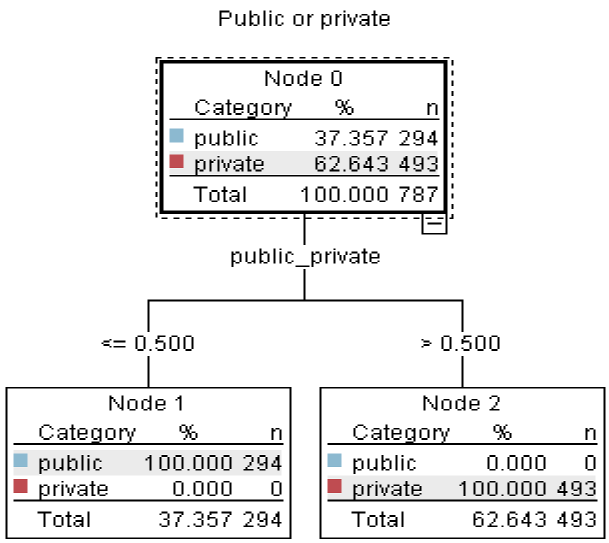
10. Decision tree classification: Using the public_private variable as categorical (flag), or deriving from it a flag variable, model the profile of a typical public and private college with a C5.0 decision tree algorithm, using all other variables as predictors (disregard tuition, given the typical difference in tuition between state and private institutions). Compute the confusion matrix and derive proper performance metrics.

The data set is unbalanced, as there are more private university 61.46 % which is 689 than compared to public university 38.54% which is 432.





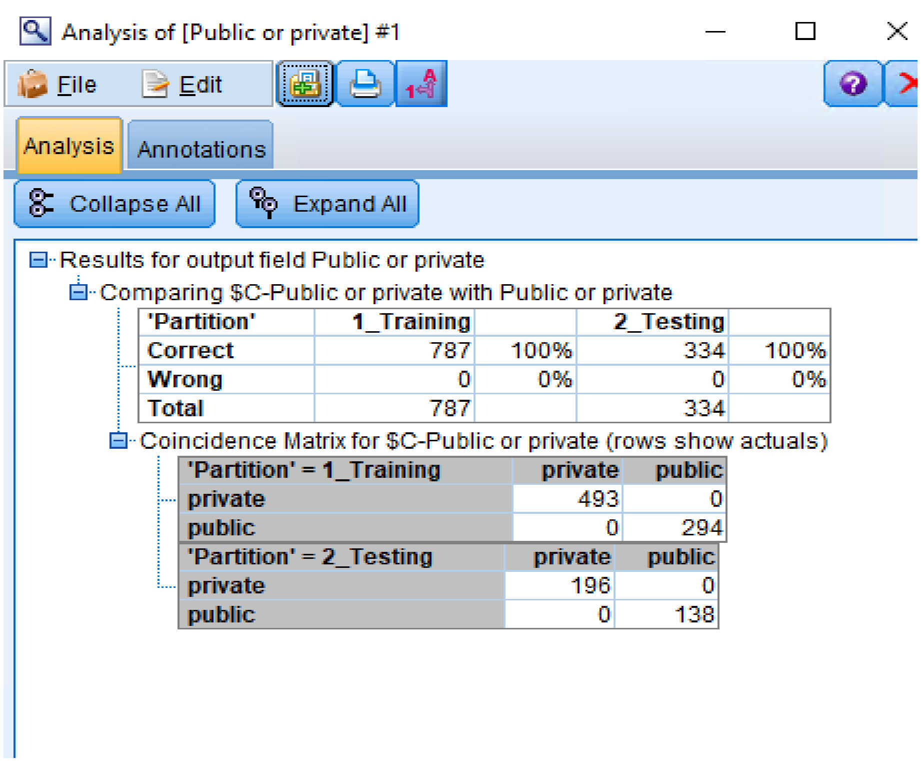
The decision tree can be built as below:



Decision Rules:

If	Consequence	Support	Confidence
If university = public	Public <= 0.500	294/ 787	294/ 787 = 37.357%
If university = private	Private < 500	493/ 787	493/ 787 = 62.643

Performance Evaluation (1)- Confusion Matrix and Derived Metrics can be given as


Analysis of [Public or private] #1

File Edit

Analysis Annotations

Collapse All Expand All

Results for output field Public or private

Comparing $C-Public or private with Public or private

'Partition'	1_Training		2_Testing	
Correct	787	100%	334	100%
Wrong	0	0%	0	0%
Total	787		334	

Coincidence Matrix for $C-Public or private (rows show actuals)

'Partition' = 1_Training	private	public
private	493	0
public	0	294

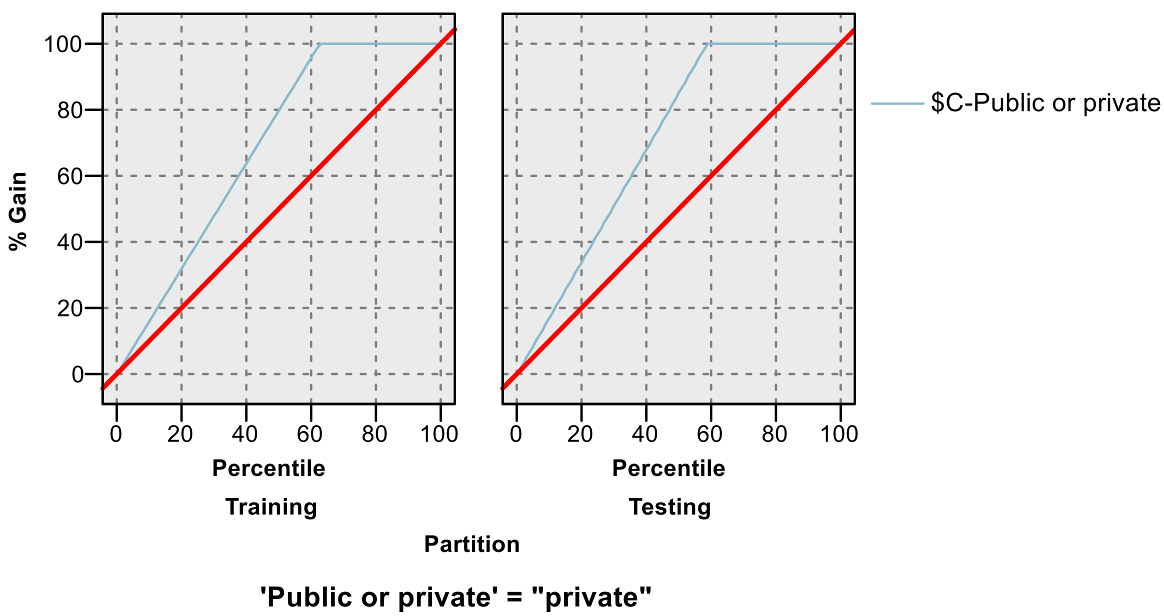
'Partition' = 2_Testing	private	public
private	196	0
public	0	138

For Testing set:

Since the data set is unbalanced, it is necessary to examine Recall and Precision in addition to the Accuracy:

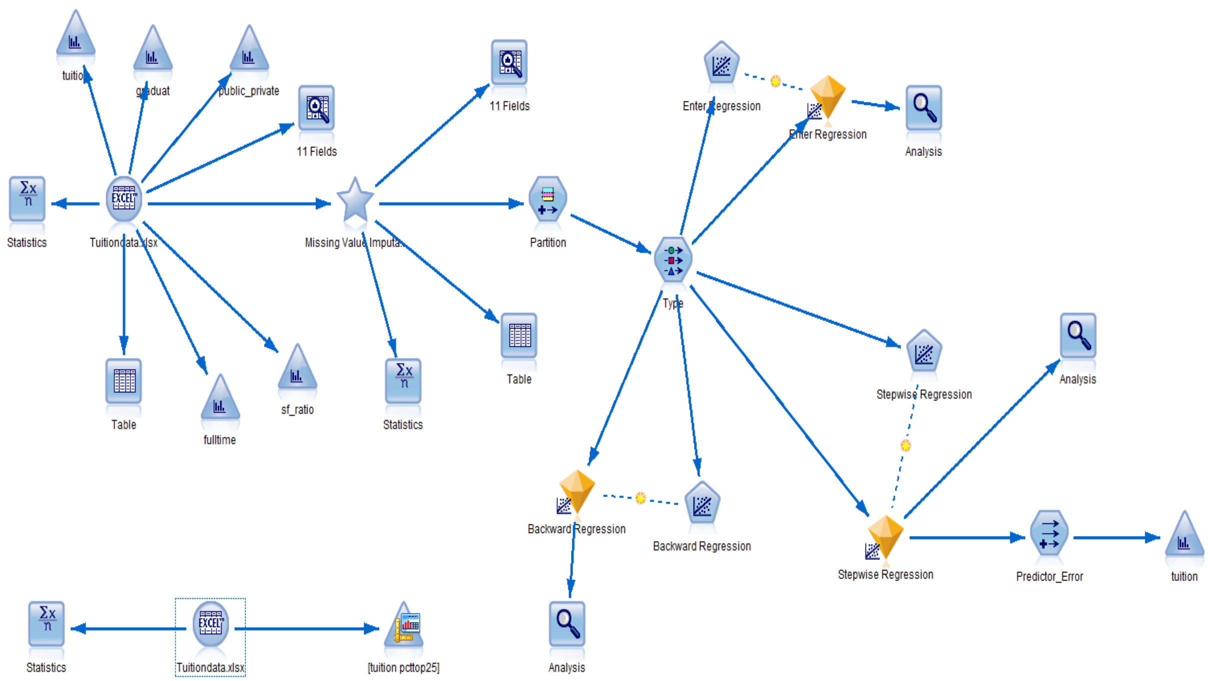
Accuracy	(TP+TN)/(TP+FP+TN+FN) = (196+138)/334	100%
Recall	TP/(TP+FN) = 196/ (196+0)	100%%
Precision	TP/(TP+FP) = 196/ (196+0)	100%
False Positive (1 - Specificity)	1-(TN/(FP+TN)) = 1 - (138/ (0+138))	0%

Performance Evaluation (2) – Gain Chart can be given as:

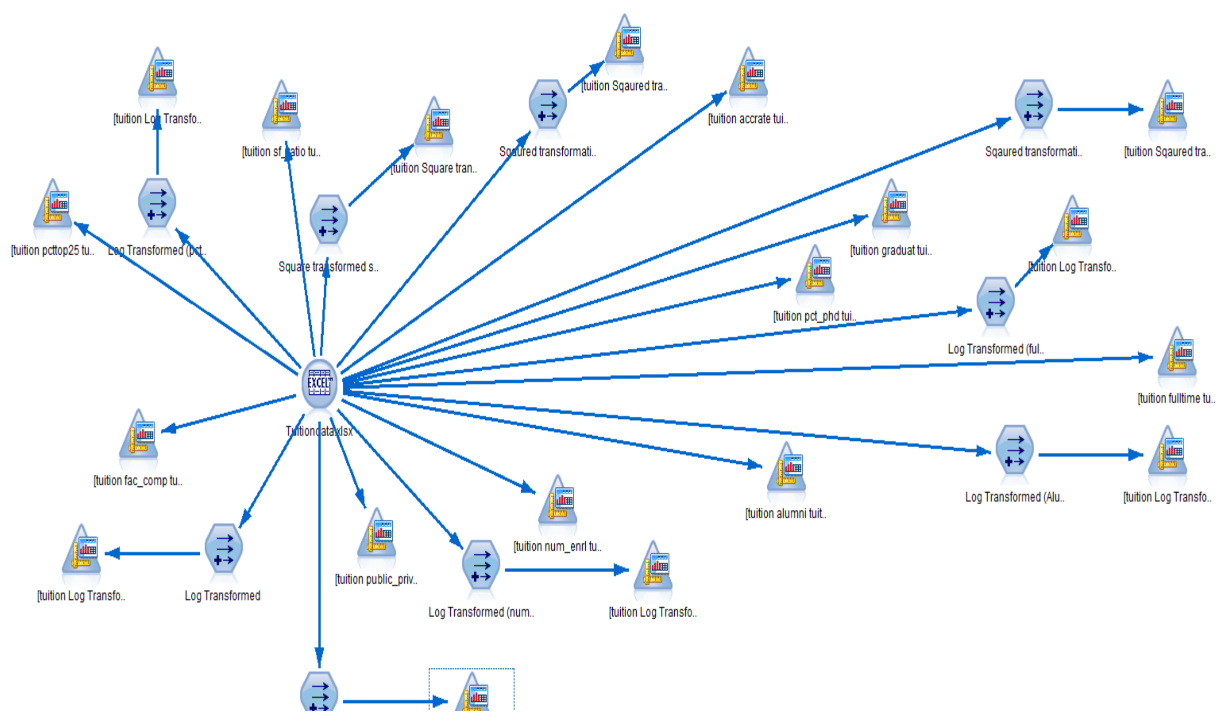


The above is Gain chart.

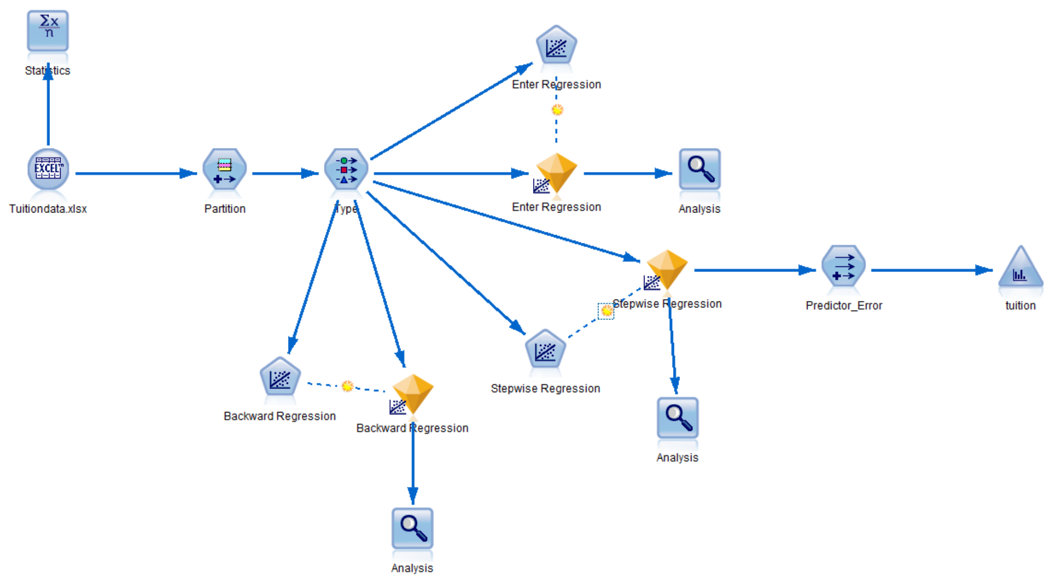
Stream for Q 1,2,3 and 7:



Stream for Question 4:



Stream for Q5 and Q6:



Stream for Question 10: