Data Mining and Predictive Analytics : Project 1

Dr. Eitel Lauria

October 21st, 2020

Shashank Kala

Sindhoori Kotapati

Table of Contents:

Executive Summary

The primary objective of this project is to provide an in depth analysis and a predictive model for the U.S. Department of Education. The predictive model will carefully analyze several key variables that were gathered from higher educational institutions as described below in correlation with tuition. The purpose of this analysis is to gather insight on the relationship between these variables with tuition in a graphical and statistical sense. The investigation of these relationships will help to provide clarity on the factors that affect tuition fees for any potential college candidate.

In this project we have utilized the SPSS modeler to explore the data, identifying the outliers and missing data, and subsequently how to treat these values. The data also contained few outliers which we chose to keep as they add to the data variation.The missing data values after being identified were replaced with their field means, and statistically compared to the original data set. The variables contained in the dataset without any alterations were used to create scatter plots that exemplified the relationship between the variables with tuition and whether a linear relationship exists. Furthermore, we have investigated the type of correlation among the predictor variables to further investigate the relationship amongst these component variables. We have also explored if a linear relationship exists between the variables in both the original data set, as well as the data set with the handled missing variables using 3 methods, *stepwise*, *backwards* and *enter*. The statistical tests were compared between the models in order to determine the overall best predictive model in the case.The final model which was selected which was created using stepwise method and contained missing values. Lastly, a decision tree classification was employed in order to model public vs private colleges.

Key Component Variables

- 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.

1. Exploratory Data Analysis

- To identify missing data connect the data file to a chart audit node.

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

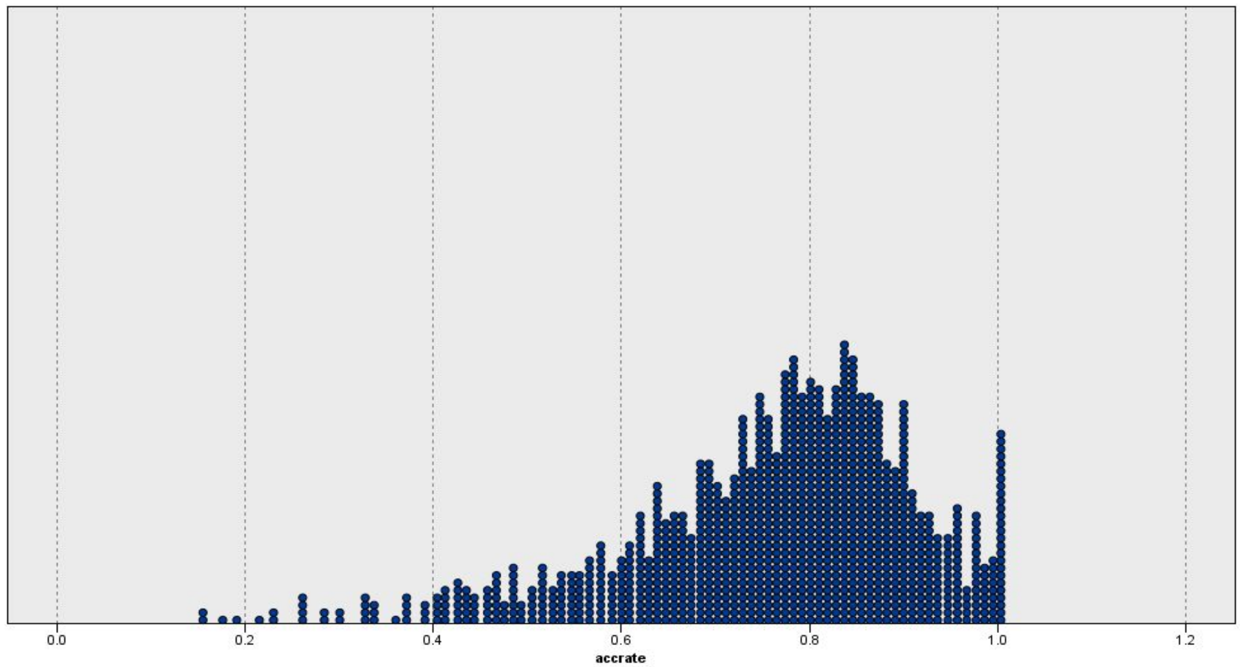
We read the field public_private as Flag:

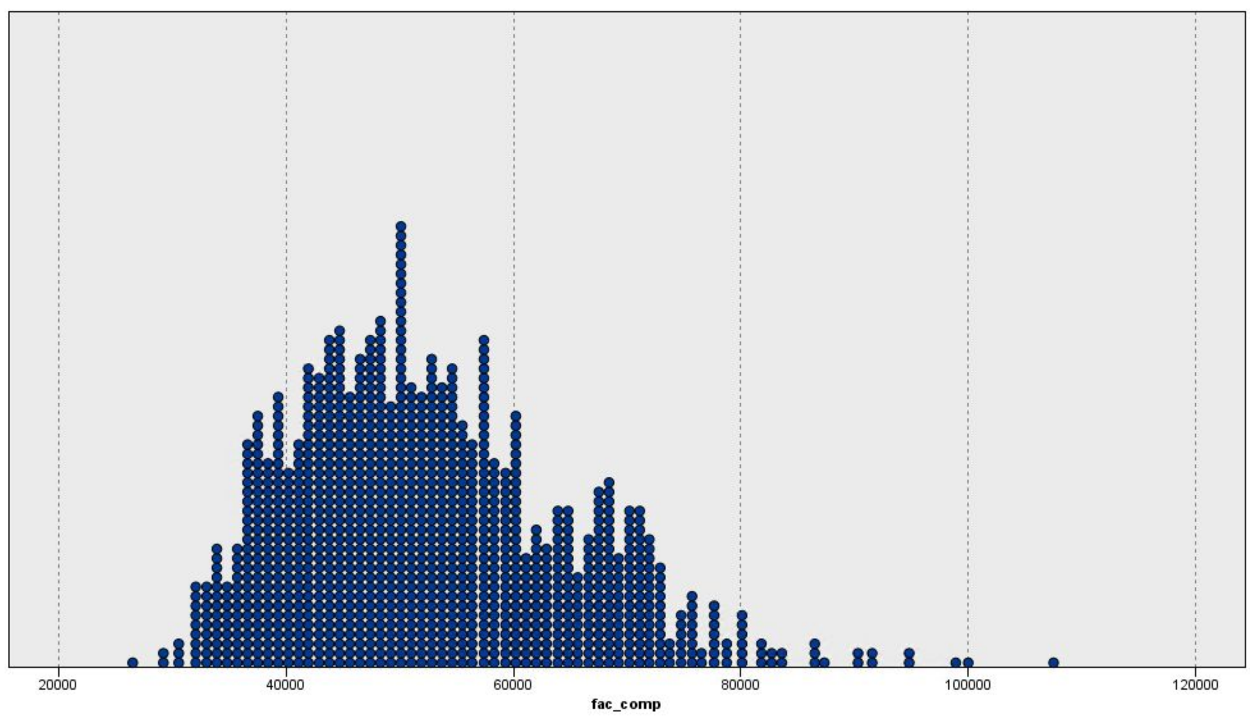
Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method
tuition	Continuous	3	0	None	Never	Fixed
pcttop25	Continuous	0	0	None	Never	Fixed
sf_ratio	Continuous	3	6	None	Never	Fixed
accrate	Continuous	14	0	None	Never	Fixed
graduat	Continuous	1	0	None	Never	Fixed
pct_phd	Continuous	7	0	None	Never	Fixed
fulltime	Continuous	9	0	None	Never	Fixed
alumni	Continuous	5	0	None	Never	Fixed
num_enrl	Continuous	13	6	None	Never	Fixed
public_private	Flag	--	--	--	Never	Fixed
fac_comp	Continuous	9	0	None	Never	Fixed

- There are 8 fields with missing data as shown above
- Pcttop25 (86.619% complete), sf_ratio (99.822% complete), accurate (99.197% complete), graduat (94.023% complete), pct_phd (97.502% complete), fulltime (98.037% complete), alumni (85.37% complete), num_enri (99.732% complete) .
- There are only two fields having about 15 percent of missing values,Pcttop25 (86.619% complete) and alumni (85.37% complete).
- The data set will require some cleaning in order to handle missing values and outliers.

2. Identification of outliers

The fields that have outliers include; tuition (3), sf_ratio(3), accurate(14), graduat(1), pct_phd(7), fulltime(9),alumni(5), num_c_enrl(13), fac_comp(9). The outliers can be visualized through dot plot as shown below for accrate & fac_comp :





We have decided to consider all the outliers , as they represent the variation in the data.

3. Handling missing values and it's analysis

The summary of data with missing values is as shown below:

accrate

Statistics

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

pcttop25

Statistics

Mean	53.493
Min	11
Max	100
Range	89
Variance	431.258
Standard Deviation	20.767
Standard Error of Mean	0.666
Median	51
Mode	40

sf_ratio

Statistics

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*

graduat

Statistics

Mean	61.421
Min	8
Max	118
Range	110
Variance	349.549
Standard Deviation	18.696
Standard Error of Mean	0.576
Median	61
Mode	63

pct_phd

Statistics

Mean	70.202
Min	8
Max	103
Range	95
Variance	296.575
Standard Deviation	17.221
Standard Error of Mean	0.521
Median	73
Mode	77

fulltime

Statistics

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*

alumni

Statistics

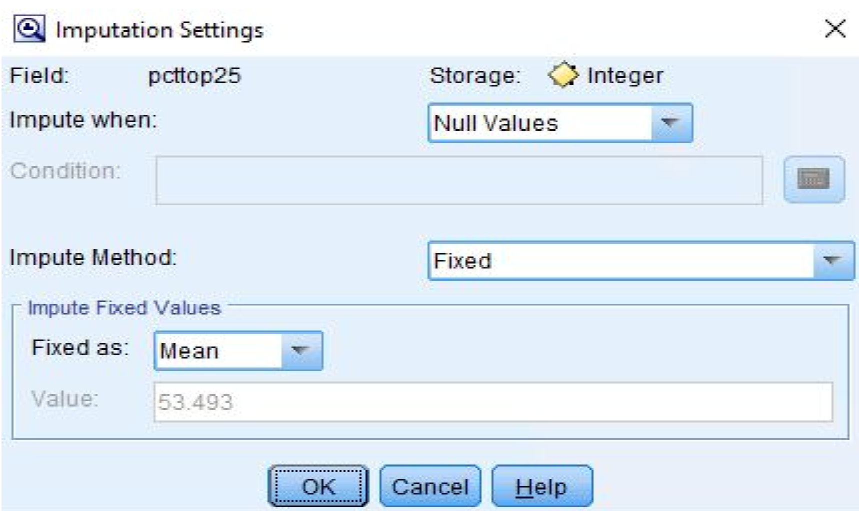
Mean	21.448
Min	0
Max	64
Range	64
Variance	160.199
Standard Deviation	12.657
Standard Error of Mean	0.409
Median	19
Mode	10

num_enrl

Statistics

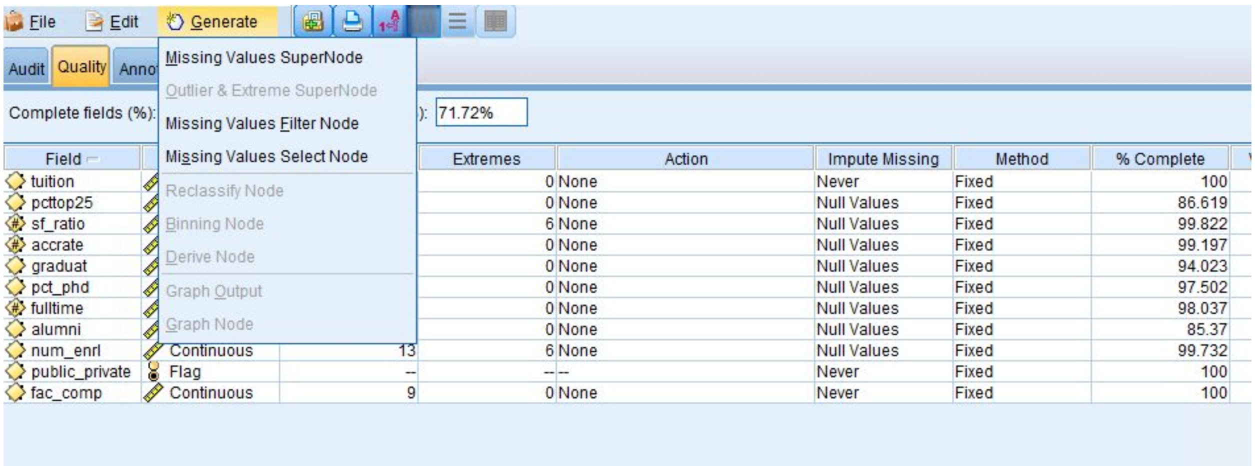
Mean	833.453
Min	21
Max	7425
Range	7404
Variance	853718.339
Standard Deviation	923.969
Standard Error of Mean	27.634
Median	478.500
Mode	169*

To handle the missing values specify impute missing as the null values and specify the method as mean shown below:



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

Generate missing value supernode and connect it to data audit node to see if the missing values are handled.



Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete
tuition	Continuous	3	0	None	Never	Fixed	100
pcttop25	Continuous	0	0	None	Never	Fixed	100
sf_ratio	Continuous	5	6	None	Never	Fixed	100
accrate	Continuous	16	0	None	Never	Fixed	100
graduat	Continuous	0	0	None	Never	Fixed	100
pct_phd	Continuous	28	0	None	Never	Fixed	100
fulltime	Continuous	23	0	None	Never	Fixed	100
alumni	Continuous	2	0	None	Never	Fixed	100
num_enrl	Continuous	13	6	None	Never	Fixed	100
public_private	Flag	--	--	--	Never	Fixed	100
fac_comp	Continuous	9	0	None	Never	Fixed	100

Let's check how the summary looks like after handling the missing data pcttop25

Statistics

Count	1121
Mean	46.335
Min	0
Max	100
Range	100
Variance	705.461
Standard Deviation	26.561
Standard Error of Mean	0.793

sf_ratio

Statistics

Count	1121
Mean	14.727
Min	0.000
Max	42.600
Range	42.600
Variance	20.093
Standard Deviation	4.482
Standard Error of Mean	0.134

accrate

Statistics

Count	1121
Mean	0.753
Min	0.000
Max	1.000
Range	1.000
Variance	0.028
Standard Deviation	0.166
Standard Error of Mean	0.005

graduat

Statistics

Count	1121
Mean	57.750
Min	0
Max	118
Range	118
Variance	540.830
Standard Deviation	23.256
Standard Error of Mean	0.695

pct_phd

Statistics

Count	1121
Mean	68.449
Min	0
Max	103
Range	103
Variance	409.292
Standard Deviation	20.231
Standard Error of Mean	0.604

fulltime

Statistics

Count	1121
Mean	77.536
Min	0.000
Max	99.940
Range	99.940
Variance	384.703
Standard Deviation	19.614
Standard Error of Mean	0.586

alumni

Statistics

Count	1121
Mean	18.310
Min	0
Max	64
Range	64
Variance	194.248
Standard Deviation	13.937
Standard Error of Mean	0.416

num_enrl

Statistics

Count	1121
Mean	831.222
Min	0
Max	7425
Range	7425
Variance	853287.266
Standard Deviation	923.735
Standard Error of Mean	27.590

Comparison of Summary values:

Pcttop 25 prior to handling the missing data had a Mean (53.493), SD (20.767), Standard error (0.666), Variance (431.258) .Pcttop25 after handling the missing data had a Mean (46.335), SD (26.561) and a Standard Error of Mean (0.793), Variance of (705.461). It is clear from these two summaries that there was a decrease in mean and variance, but an increase in SD, Standard Error of Mean.

Prior to handling missing data Sf_ratio had a Mean (14.753), SD (4.443), and Standard error (0.133), Variance (19.740). After handling the missing data, Sf_ratio had a Mean (14.727) , SD(4.482) Standard error of mean (0.134), Variance of (20.093). The values did not change dramatically from before and after.

Prior to handling missing data accrate had a Mean(0.759), SD (0.152) and Standard error(0.005), Variance (.023) . After handling the missing data, accrate had a mean (0.753) SD (0.166) Standard error (0.005), Variance of (0.028). The values here did not have a significant change from before to after.

Prior to handling the missing data, graduat had a mean (61.421), SD (18.696), Standard Error of Mean (0.576), Variance (540.830). After handling the missing data, graduat had a mean (57.750), SD (23.256), Standard Error of Mean (0.695), Variance (349.549). The mean has decreased, the SD and Standard Error of Mean increased after, and the Variance decreased after.

Prior to handling the missing data, pct_phd had a mean (70.202), SD (17.221), Standard Error of Mean (0.521), Variance (296.575). After handling the missing data, pct_phd had a mean (68.449), SD (20.231), Standard Error of Mean (0.604), Variance (409.292). It is clear that the mean has decreased slightly, the SD increased, the Standard Error of Mean increased slightly after, and the Variance increased.

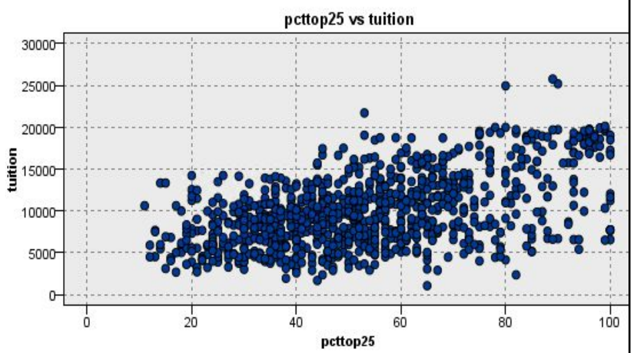
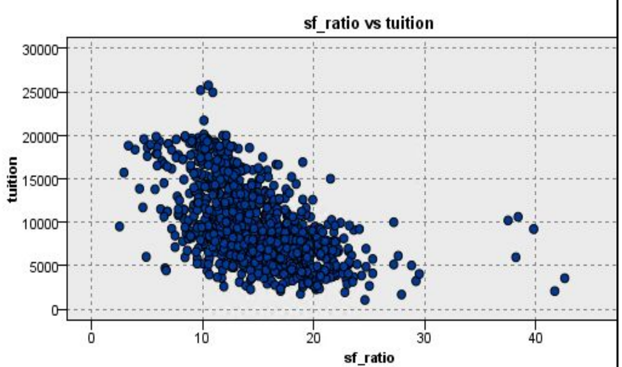
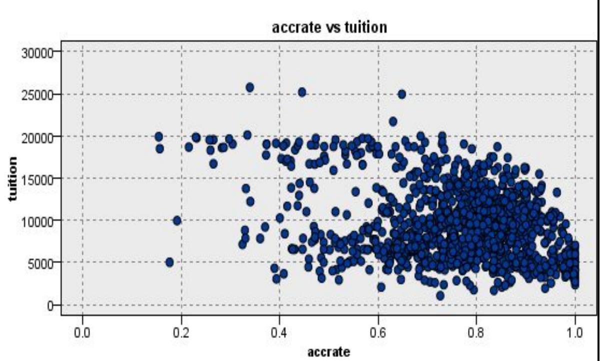
Prior to handling the missing data, fulltime had a mean (79.089), SD (16.418), Standard Error of Mean (0.495), Variance (269.543).After handling the missing data, fulltime had a mean (77.536), SD (19.614), Standard Error of Mean (0.586), Variance (384.703). From these two summaries, the mean has decreased, the SD and Standard Error of Mean increased, as well as the variance.

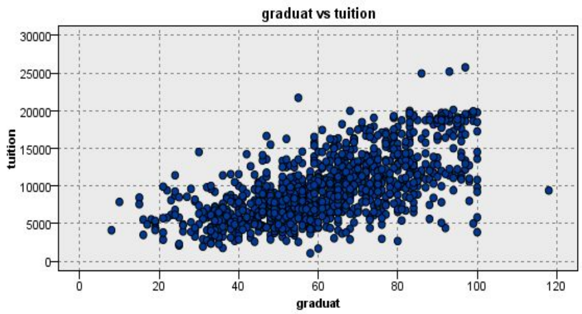
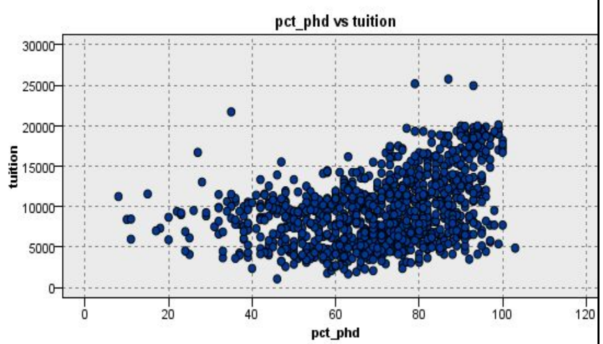
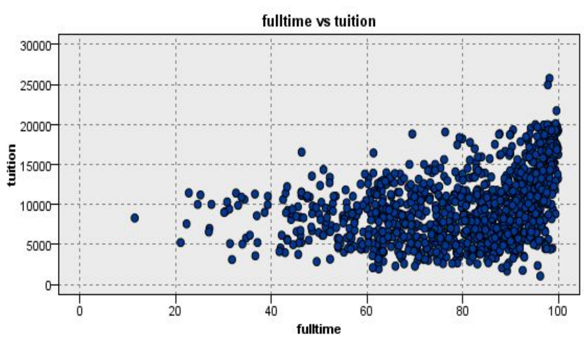
Prior to handling the missing data, alumni had a mean (21.448), SD (12.657), Standard Error of Mean (0.409), Variance (160.199).After handling the missing data, alumni had a mean (18.310), SD (13.937), Standard Error of Mean (0.416), Variance (194.248) .The mean has decreased, the SD and Standard Error of Mean increased, as well as the variance.

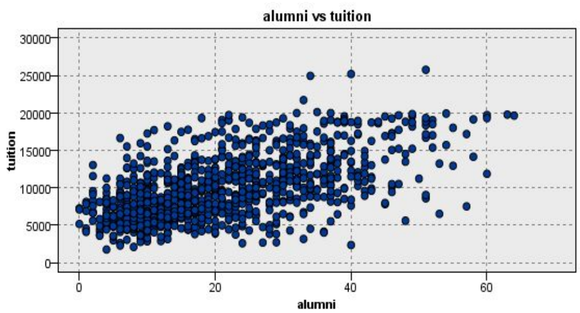
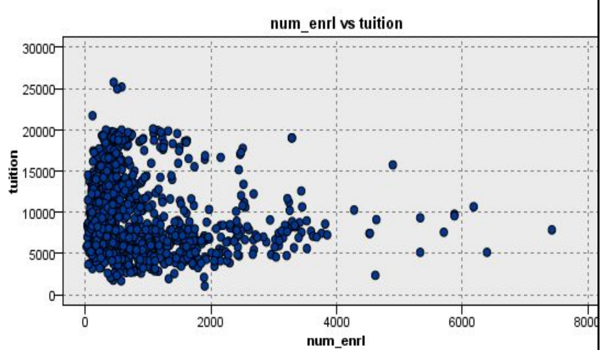
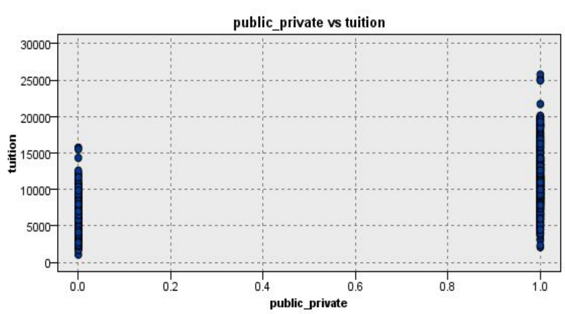
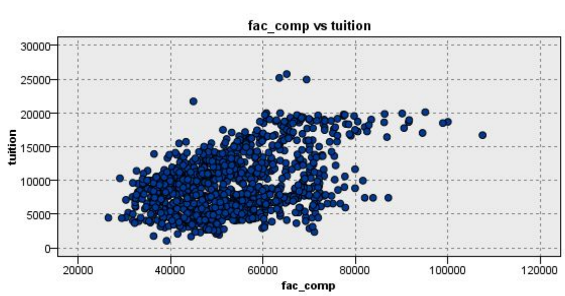
Prior to handling the missing data, num_enrl had a mean (833.453),SD (923.969), Standard Error of Mean (27.634), Variance (853718.339). After handling the missing data, num_enrl had a mean (831.222), SD (923.735),Standard Error of Mean (27.590), Variance (853287.266). The mean has decreased, but the change in the other values were not significant.

Replacing the missing values with the mean is not always a good idea, especially when the data set is limited such as what we are given, and there are too many missing values. Mean imputation may cause reduction in variance and thereby creating bias in the model. But in our case the variance is not getting reduced, hence it won't create any bias.

4. Relationship between Tuition and other variables

Variables	Plot	Description
Pcttop25 vs tuition		When we try to fit the line through a cluster of points, we observe that as pcttop25 increases the tuition variable also increases, hence we can say that it has a positive strong linear relationship.
Sf_ratio vs tuition		When we try to fit the line through a cluster of points, we observe that as sf_ratio increases the tuition variable decreases, hence we can say that it has a negative strong linear relationship as most of the points lie close to the line. There are also few outliers as shown in the graph.
Accrate vs tuition		In this case, it is clear that as the value of accrate increases, tuition tends to decrease indicating a negative linear relationship.

<table>
<tr>
<td>Graduat vs tuition</td>
<td>
</td>
<td>If you apply a straight line, we can see that as the graduat increases, the tuition also increases indicating a strong positive linear relationship.</td>
</tr>
<tr>
<td>Pct_phd vs tuition</td>
<td>
</td>
<td>If a straight line is applied to the cluster of points, we can visualize that there is a positive linear relationship between pct_phd and tuition.</td>
</tr>
<tr>
<td>Fulltime vs tuition</td>
<td>
</td>
<td>In this case a curve seems to be a better fit to cover the most of the points, hence both of them has non linear relationship</td>
</tr>
</table>

Alumni vs tuition		If you apply a straight line, we can see that as the alumni increases, the tuition also increases indicating a positive linear relationship.
Num_enrl vs tuition		In this case a curve seems to be a better fit to cover the most of the points and it indicates non linear relationship between Num_enrl and tuition
Public_private vs tuition		When we try to fit the line through a cluster of points, we observe that these are two separate groups and are not in linear relationship with tuition.
Fac_comp vs tuition		When we try to fit in a straight line between the cluster of points, we notice that as fac_comp increases tuition also increases, hence it has a strong positive linear relationship.

5. Correlation among Predictor Variables

We can see that most of the predictor variables are strongly correlated and this can result in multicollinearity,which may lead to incoherent results. Although it doesn't affect the prediction of the target variable, we should ensure that it is minimum. To avoid this we can use a user defined composite. We should take the mean of standardized values of variables and then peerform the regression.

tuition
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
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

sf_ratio
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
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

graduat
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
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

fulltime

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

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

num_enrl

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

Pearson Correlations

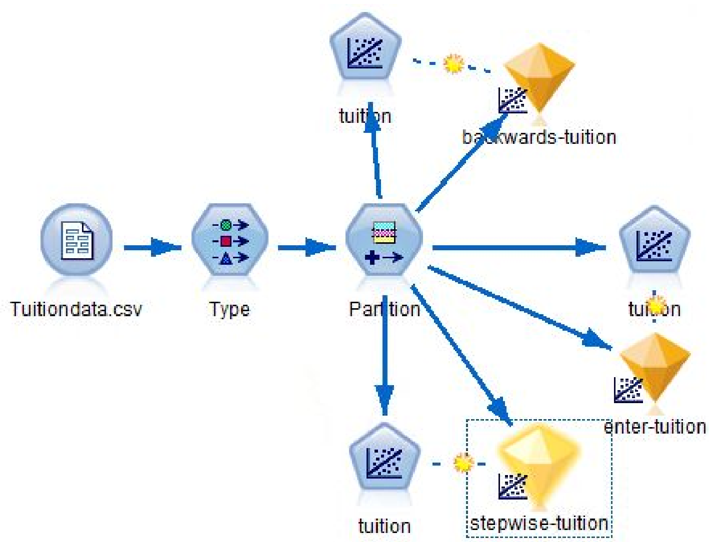
tuition	0.609	Strong
pcttop25	0.166	Strong
sf_ratio	-0.485	Strong
accrate	-0.003	Weak
graduat	0.465	Strong
pct_phd	-0.113	Strong
fulltime	0.081	Strong
alumni	0.456	Strong
num_enrl	-0.534	Strong
fac_comp	-0.195	Strong

fac_comp

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

6. Multiple Linear Regression



- The dataset was partitioned 70/30, in which the target variable selected was tuition and all other variables were input. Three methods were employed, *enter, stepwise,* and *backwards* to create the linear regression models.
- Below are the models containing regression equation, statistical tests, and analysis for each of the 3 methods (stepwise, backwards, enter)

Method: Enter

Variable	Metric slope	Std. error	t	p
pcttop25	-4.772	6.273	-0.761	0.447
sf_ratio	-170.9	26.851	-6.363	0.000
accrate	94.94	685.663	-0.138	0.890
graduat	18.54	6.420	2.887	0.004
pct_phd	31.74	7.735	4.104	0.000
fulltime	11.67	5.762	2.025	0.043
alumni	43.41	8.419	5.156	0.000
num_enrl	-0.2936	0.133	2.202	0.028

public_private	4309.8	289.558	14.884	0.000
fac_comp	0.1441	0.012	12.351	0.000

R=0.877, R square = 0.770, Adjusted R square = 0.766, Std Error = 2010.55

Method: Stepwise

Variable	Metric slope	Std. error	t	p
sf_ratio	-165.746	26.587	-6.234	0.000
graduat	19.207	6.249	3.073	0.002
pct_phd	32.808	7.352	4.462	0.000
alumni	44.763	8.168	5.481	0.000
num_enrl	-0.278	0.131	-2.121	0.034
public_private	4305.303	286.666	15.019	0.000
fac_comp	0.141	0.011	13.144	0.000

R=0.876, R square = 0.768, Adjusted R square = 0.765, Std Error = 2013.051

Method:Backwards

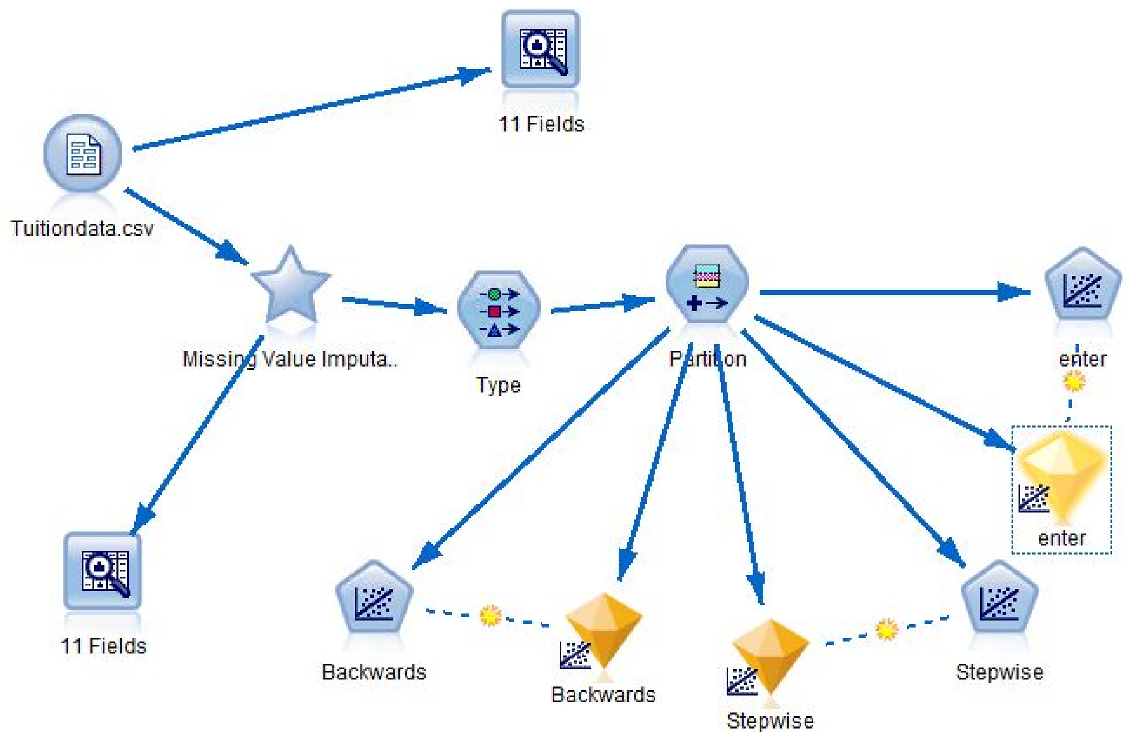
Variable	Metric slope	Std. error	t	p
pcttop25				
sf_ratio	-167.703	26.542	-6.318	0.000
accrate				
graduat	17.613	6.829	2.801	0.005
pct_phd	30.266	7.452	4.062	0.000
fulltime	10.924	5.654	1.932	0.054
alumni	42.123	8.261	5.099	0.000
num_enrl	-0.303	0.131	-2.305	0.022
public_private	4299.794	285.983	15.035	0.000

fac_comp	0.141	0.011	13.211	0.000

R=0.877, R square = 0.770, Adjusted R square = 0.766, S = 2008.164

As we can note from above that the best models are the one created using stepwise and backward regression. We can see that the R values are almost similar in both the cases which is 0.87, we would choose the stepwise regression as our predictor variables are strongly correlated.

7. Multiple Linear Regression(without missing data)



Method: Enter

Variable	Metric slope	std.Error	t	P
pcttop25	0.3136	5.516	-4.045	0.955
sf_ratio	-153.5	20.805	0.057	0.000
accrate	-171.7	575.698	-7.376	0.766
graduat	18.19	5.608	-0.298	0.001
pct_phd	25.26	6.139	3.244	0.000
fulltime	19.83	5.055	4.115	0.000
alumni	37.14	7.717	3.923	0.000
num_enrl	-.2539	0.121	4.813	0036

public_private	4413.2	238.824	-2.095	0.000
fac_comp	0.1452	0.01	18.479	0.000

R=0.873 , Rsquare = 0.763 , Adjusted R square = 0.76

Method:Stepwise

Variable	Metric slope	Std Error	t	P
sf_ratio	-154.1	20.635	-7.467	0.000
graduat	18.42	5.540	3.325	0.001
pct_phd	25.31	5.973	4.237	0.000
fulltime	19.88	4.954	4.013	0.000
alumni	37.32	7.609	4.905	0.000
num_enrl	-0.256	0.119	-2.144	0.032
public_private	4411.2	236.069	18.686	0.000
fac_comp	0.1463	0.009	16.131	0.000

R= 0.873 , R square = 0.763, Adjusted R square = 0.760 , Std.Error = 2095.535

Method: Backwards

Variable	Metric slope	Std Error	t	P
sf_ratio	-189.4	-18.859	-10.043	0.000
accrate	-1741.6	415.609	-4.191	0.000
graduat	14.4	5.567	2.586	0.010
pct_phd	22.994	6.017	3.822	0.000
fulltime	13.878	4.811	2.885	0.004
alumni	37.518	7.624	4.921	0.000

public_private	4402.123	206.312	21.337	0.000
fac_comp	0.123	0.008	14,755	0.000

R= 0.979 , R square = 0.959, Adjusted R square = 0.959 , Std.Error = 2119.439

In this case we have replaced all the missing values with their means and then performed regression. The best model in this case is the backward regression. The value of R is 0.979 which is better than the other two models. Also standard error of estimate is 2119,439 which lies in the same range as other models.

8. Comparison of multiple linear regression models

The best models selected are as follows:

a) Stepwise - When missing data was not imputed.

Comparing $E-tuition with tuition

'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

R=0.876, R square = 0.768, Adjusted R square = 0.765, Std Error = 2013.051

b) Backwards - When missing data was imputed

Comparing $E-tuition with tuition

'Partition'	1_Training	2_Testing
Minimum Error	-8991.88	-10702.142
Maximum Error	9841.528	10566.651
Mean Error	-24.277	-20.809
Mean Absolute Error	1603.913	1749.983
Standard Deviation	2109.841	2329.211
Linear Correlation	0.87	0.828
Occurrences	787	334

R= 0.979 , R square = 0.959, Adjusted R square = 0.959 , Std.Error = 2119.439

When we compare both the models, we notice that the R -value of backward (0.979) is greater than the stepwise(0.876) . The standard error estimate is better in case of stepwise hence we can proceed with it and can describe it as shown in next section.

9. Analysis of final model

For the final (chosen) model

Stepwise Regression without alterations to missing data

a. Write out the estimated regression equation and explain the meaning of the coefficients

The estimated regression for the stepwise model shown in Q6 is described below

tuition= sf_ratio*(-165.7) + graduat*(19.21)+ pct_phd*(32.81)+ alumni*(44.76) + num_enrl*(-0.2777)+public_private*(4305.3)+fac_comp*0.1411 -2292.6

The intercept is indicated by (-2292.6) . When all other variables are held constant, a slope of (-165.7) indicates that a unit decrease of student to faculty ratio will decrease the tuition. A slope of 19.21 indicates that when all the other variables are held constant, a unit increase of the percent of students who graduate will increase the tuition. The slope of 32.81 indicates that when you keep other variables constant, the unit increase of percent faculty with Ph. D.'s will increase the tuition subsequently. A slope of 44.76 indicates that when every other variable is constant, the unit increase of the percent of alumni who donate will increase the tuition by 44.76. A slope of num_enrl*(-0.2777) indicates that when every other variable is constant, the unit decrease of the number of students enrolled will decrease the tuition by .28. A slope of public_private*(4305.3) indicates that when every other variable is constant, the unit increase of the type of school will increase the tuition by 4305.3. A slope of fac_comp*0.1411 indicates that when every other variable is constant, the unit increase of the average faculty compensation will increase tuition by 0.14.

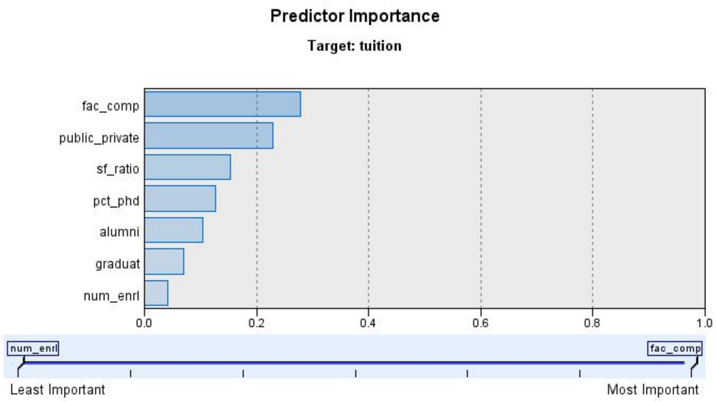
b. Provide a full report of the chosen regression model and report its metrics

(goodness of fit, predictive performance) and statistics on training and test data

Results for output field tuition

Comparing $E-tuition with tuition

'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



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.603[a]	.364	.363	3316.144
2	.836[b]	.699	.698	2281.640
3	.859[c]	.737	.736	2135.605
4	.869[d]	.755	.753	2064.943
5	.873[e]	.763	.761	2032.419
6	.875[f]	.766	.764	2019.304
7	.876[g]	.768	.765	2013.051

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

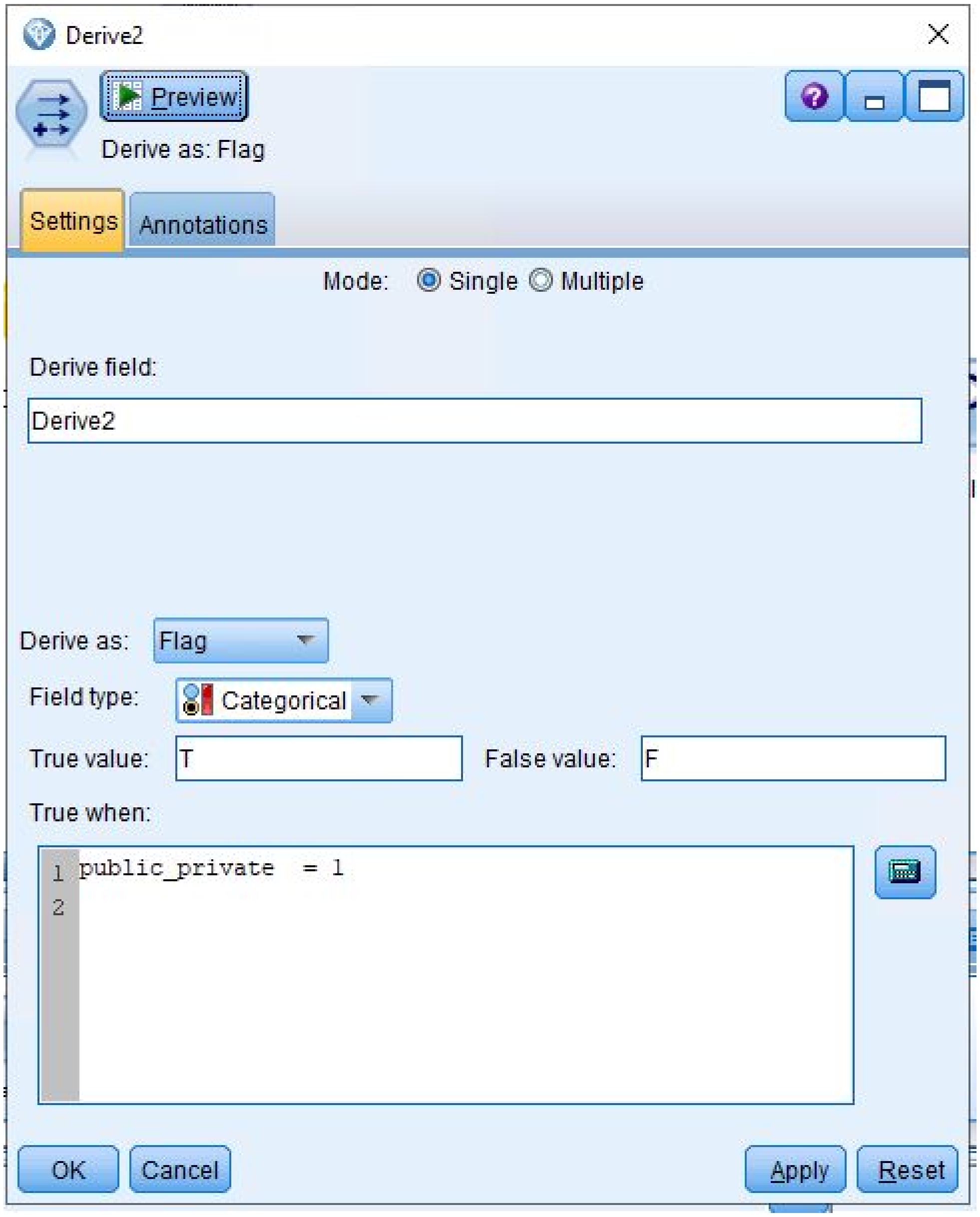
Variable	Metric slope	Std. error	t	p
sf_ratio	-165.746	26.587	-6.234	0.000
graduat	19.207	6.249	3.073	0.002
pct_phd	32.808	7.352	4.462	0.000
alumni	44.763	8.168	5.481	0.000
num_enrl	-0.278	0.131	-2.121	0.034
public_private	4305.303	286.666	15.019	0.000
fac_comp	0.141	0.011	13.144	0.000

R=0.876, R square = 0.768, Adjusted R square = 0.765, Std Error = 2013.051

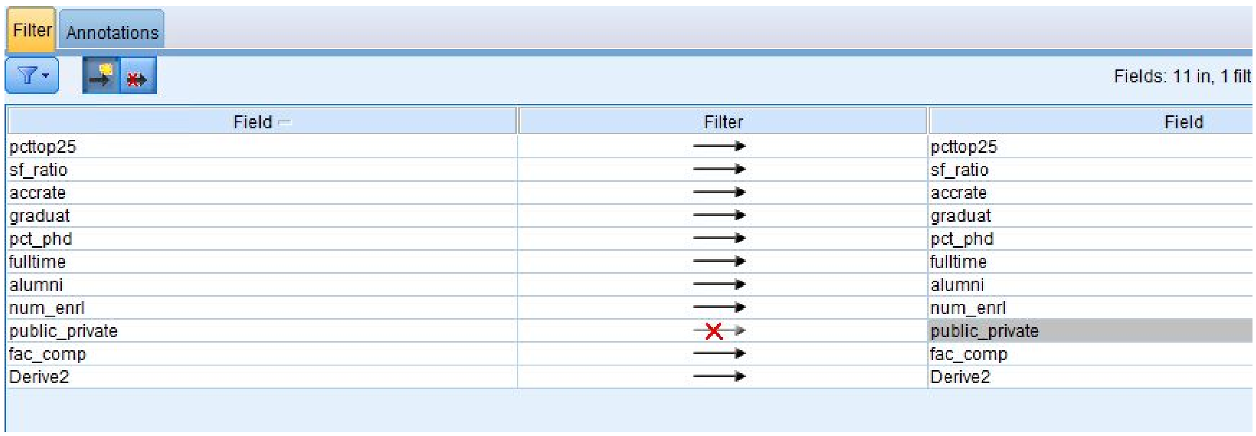
10. Decision tree classification

Steps:

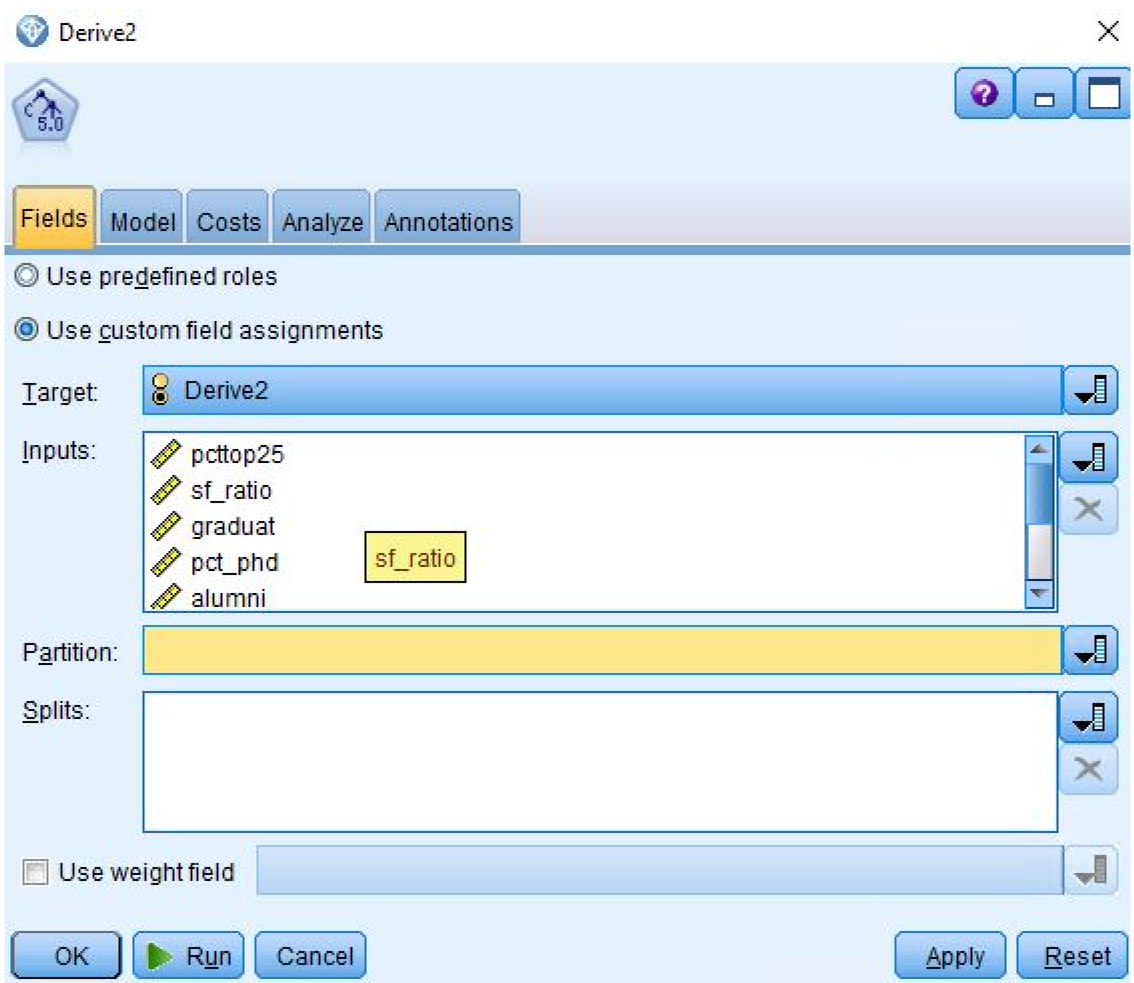
1. Convert the public_private into categorical variable as shown below using derive node



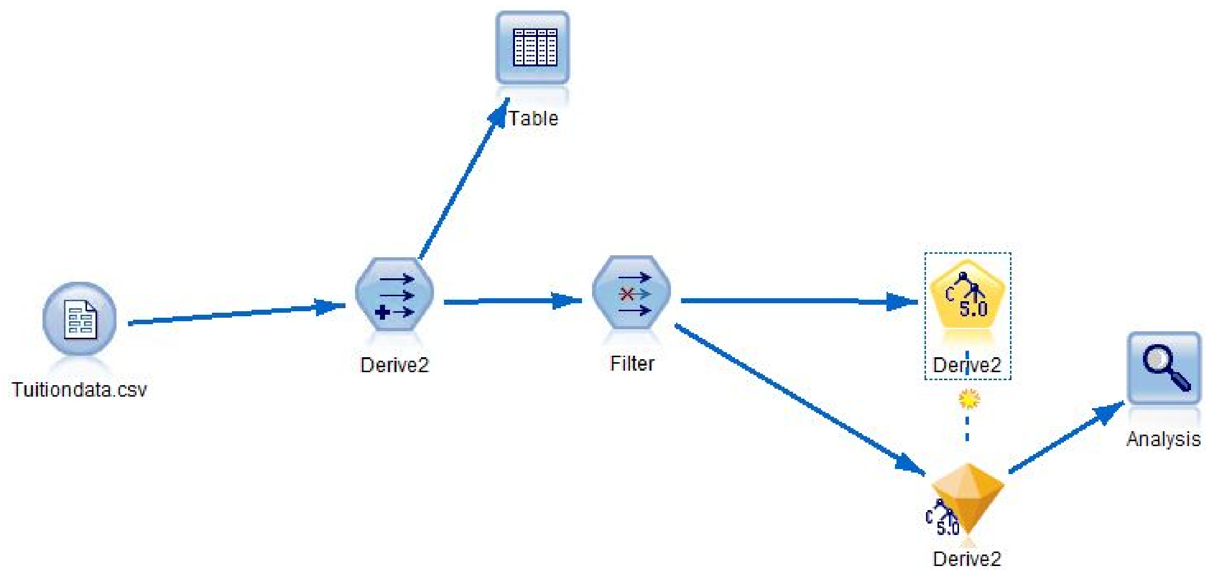
2. Now using filter node , remove the original public_private as we have a new categorical variable.

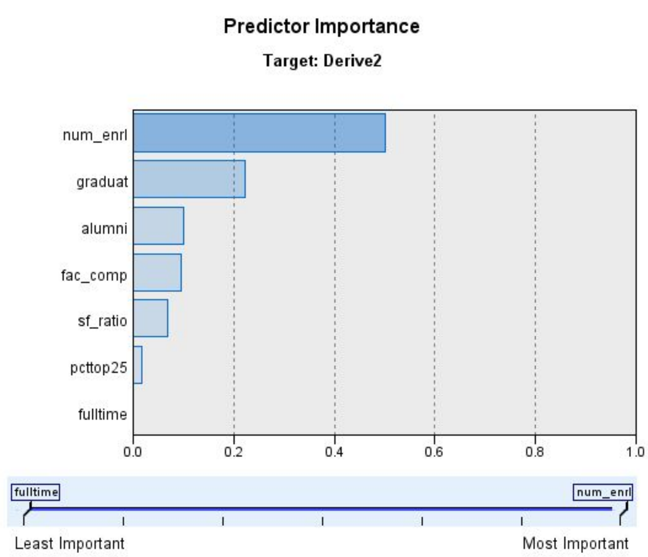


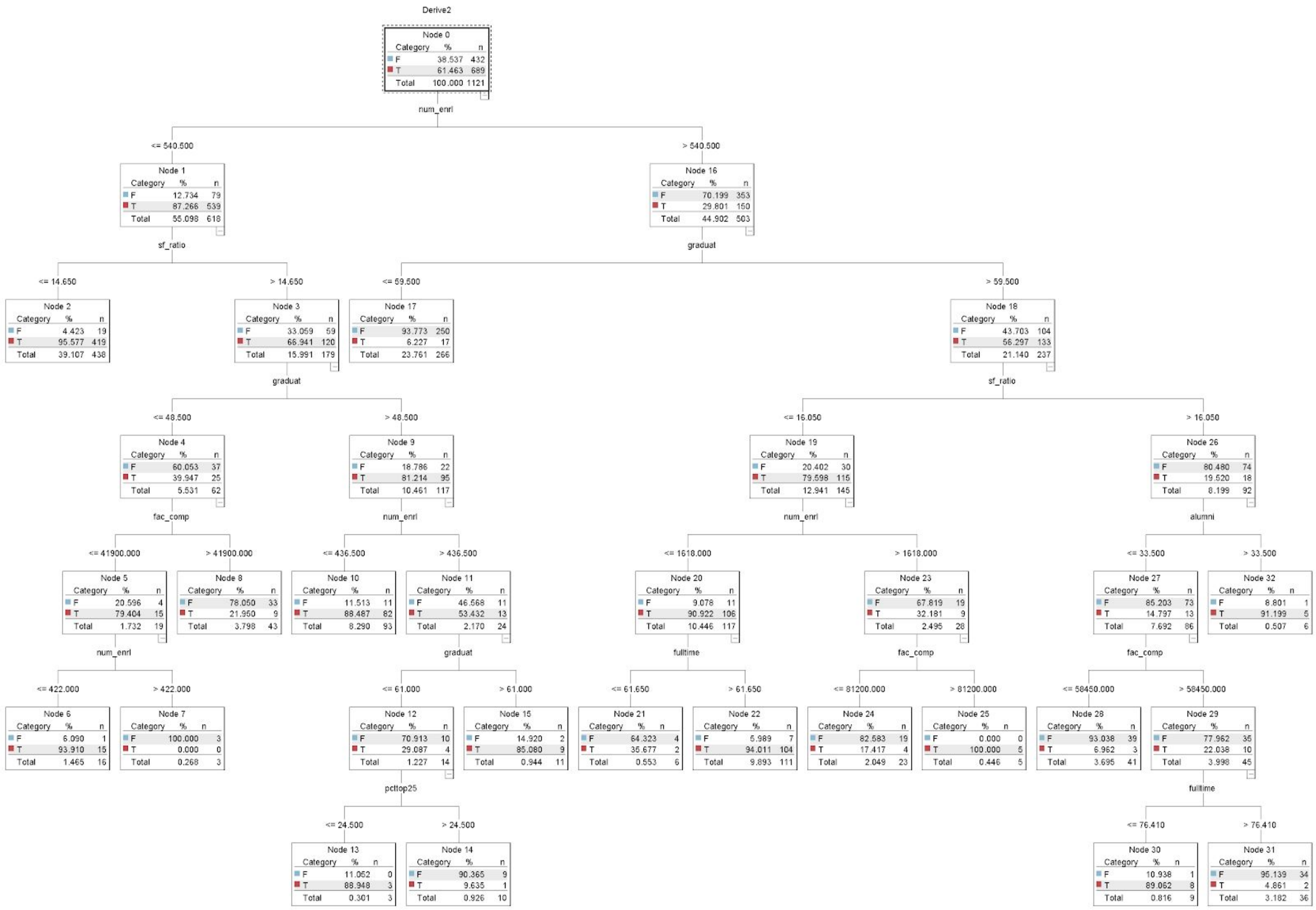
3. Connect it to the C5.0 model and select target variables , variable as shown below:



4. Run the model and connect it to analysis







The coincidence matrix is as shown below:

Results for output field Derive2

Individual Models

Comparing $C-Derive2 with Derive2

Correct	1,039	92.69%
Wrong	82	7.31%
Total	1,121	

Coincidence Matrix for $C-Derive2 (rows show actuals)

	F	T
F	389	43
T	39	650

Performance Evaluation

F	0.858
T	0.423

Evaluation Metrics

Model	AUC	Gini
$C-Derive2	0.944	0.887

References

1. Data Mining and Predictive Analytics, Daniel T. Larose ,Chantal D.Larose
2. https://statisticsbyjim.com/regression/interpret-r-squared-regression/#:~:text=R%2Dsquared%20evaluates%20the%20scatter,around%20the%20fitted%20regression%20line.&text=For%20the%20same%20data%20set,that%20a%20linear%20model%20explains.
3. https://www.sciencedirect.com/topics/mathematics/standard-error-of-estimate