Analysis and Prediction of Cost of Attendance using CRISP-DM

Sri Lakshmi Tammineni - 11482201

SriLakshmiTammineni@my.unt.edu

Anusha Putta - 11500342

AnushaPutta@my.unt.edu

University of North Texas-Denton

This report was done as a part of final project submission for CSCE 5310 - Empirical Analysis for Computer Sciences department of University of North Texas - Denton. We would like to thank Professor Levent Bulut for his suggestions and encouragement throughout the course, our project and especially in completing this report successfully.

Abstract

College education is expensive. College costs have increased over the last few decades and this trend is expected to continue. Typical college cost of attendance consists of several components and in this project, we study a dataset consisting of 1368 students' cost of attendance data and all the various components. We would like to understand the trends of individual components of the cost and the overall trend of cost of attendance. We would like to understand the various determinants for college costs. We hope to analyze the data and understand the trends of identified determinants and hopefully predict the future costs.

This analysis might help future students with valuable insights into how much one should save for education, planning for education funding or what components of the costs can be managed effectively to keep the overall cost of going to college for higher education. While these are our proposed aims, we are open to let the data mold our research and conclude our findings. Apart from the technical study of the data, we hope to conclude the project with a set of suggestions for any student who wishes to pursue an economical college education.

As part of this project, we apply our design and understanding of various data extraction, cleaning, analysis, and modeling techniques learnt during the length of our course and evaluate the dataset provided. The dataset consists of 1368 observations from 1368 institutions with 65 variables describing each observation. A detailed study of data is carried out in subsequent parts of the report.

1. Analysis and Prediction of Cost of Attendance using CRISP-DM

In this project we apply our design and understanding of various data extraction, cleaning, analysis, and modeling techniques learnt during the length of our course and evaluate the dataset provided. For the purposes of our project, we are analyzing the Cost of Attendance data set provided in our Data Sources section of the course project. The dataset consists of 1368 observations with 65 variables describing each observation. The data consists of observations both from undergraduate and graduate levels of education. We apply the open standard data mining model called CRISP-DM (Cross-industry Standard Process for Data Mining) to analyze this data. For the purposes of data cleaning, we used MS Excel, for analysis of determinant variables we used SAS tools and utilities and finally for linear regression purposes we used both SAS and Python programming in Jupyter Notebook.

2. Method and Brief Review

2.1. History of the model

Cross-industry standard process for data mining, known as CRISP-DM, is an open standard process model that describes common approaches used by data mining experts. It is the most widely used analytics model. A CRISP-DM project has 6 major phases and these phases as implemented in our project are discussed in below sub-sections.

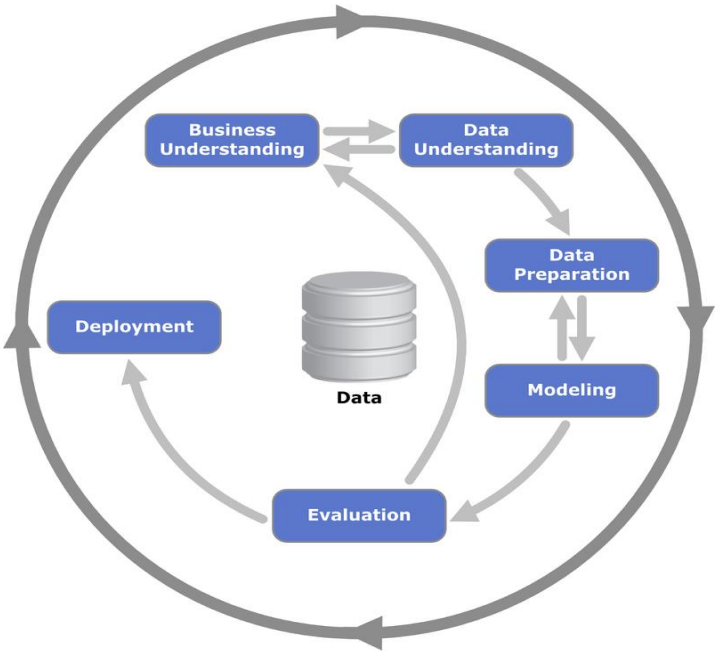


Fig 1: CRISP-DM Phases

2.1.1. Business Understanding

In this initial phase of the project, we studied the dataset to understand and propose a business problem. In our case, we would like to understand the determinant variables or factors that influence the overall cost of attendance. Once these variables are identified, we were able to put forward the following question and try to answer them based on the results of our analysis:

- Can we predict the future cost of education?
- Which components of college cost should students concentrate on to reduce expenses?
- Can our prediction help future students in choosing a college experience with reasonable costs?

2.1.2. Data Understanding

During the data understanding phase of the project, we acquire the dataset and try to identify the variables that are significant. In this process, we shortlist a set of 28 variables which we believe are significant for the purposes of analysis. Also, we choose to analyze only data about full-time students. Hence, any variables representing part-time students have been omitted. Below is the list of variables that we have shortlisted for exploratory data analysis (Years 2016-2019):

- University ID
- Published In-state and out-of-station Tuition and Fees
- Books and Supplies
- On and Off campus, room, and boarding
- On and Off campus other expenses

With a good understanding of data and an insight into what needs to be done to prepare and clean data, we move to the next step of the CRISP-DM process.

2.1.3. Data Cleansing and Preparation

For this phase of the project, we start with MS Excel as the primary tool to identify blank cells and combine multiple year specific columns for the same variable. For empty cells, we used a combination of techniques and SAS Studio Tools and Utilities for either approximate values or filter out these outliers. We need to construct the final dataset from the raw dataset by formatting the data as below:

- We choose to consider only full-time students for the purposes of this study.
- We choose to consider both in-state and out of state tuition students.
- We choose to consider only on-campus students and off campus students not living with family. These use cases we believe cover a significant number of students and these choices would narrow our problem to be analyzed effectively.
- We created a new categorical variable named ‘Year’ to highlight the academic year and re-ordered the data to combine columns of the same data into a single column with the year as identifier.
- Missing data values for single cells in the data are populated with linear estimates. Also, for institutes with only one type of housing we duplicated the housing cost data for housing types that are unavailable.
- Finally, for ‘Books and Supplies’ empty cells are populated with the mean of all other institutes’ data for that academic year.

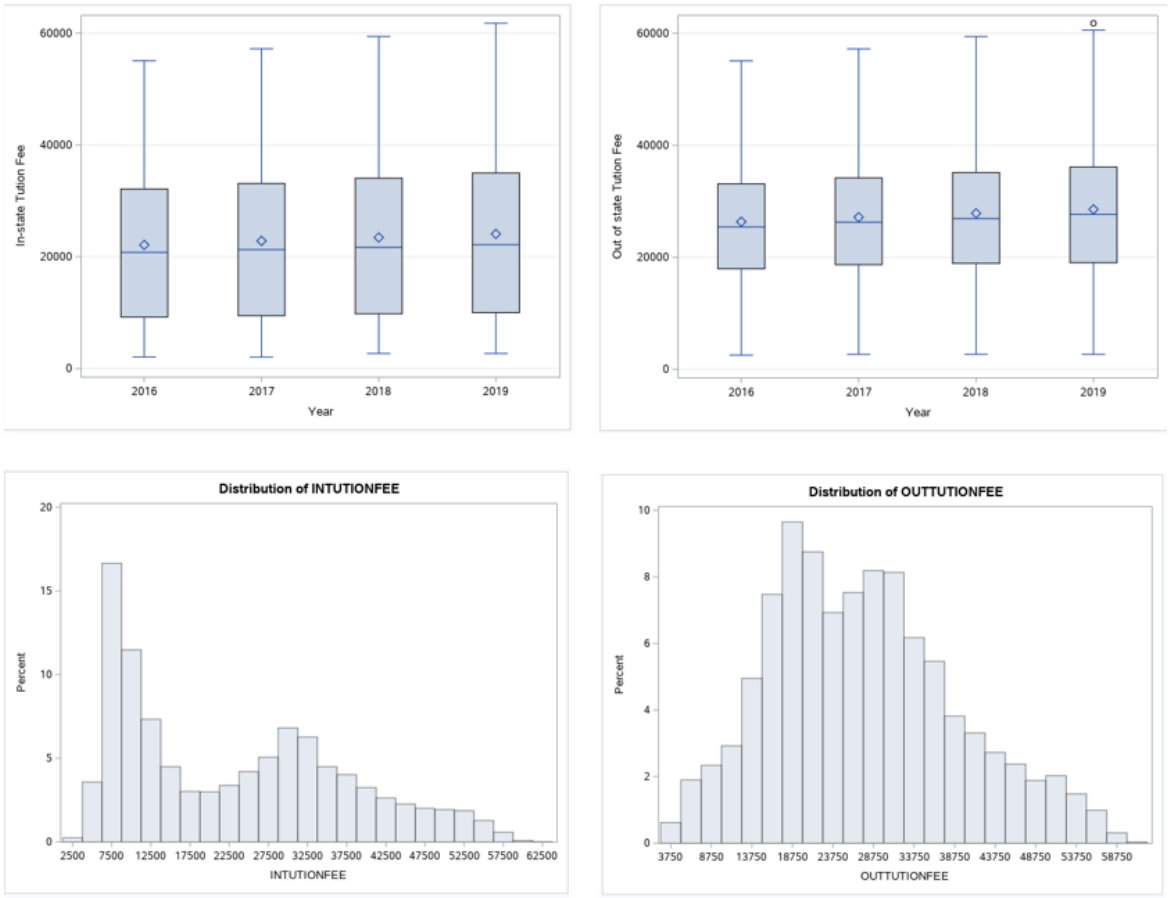


Fig 2: Bar charts and Histogram for In-state and out of state tuition

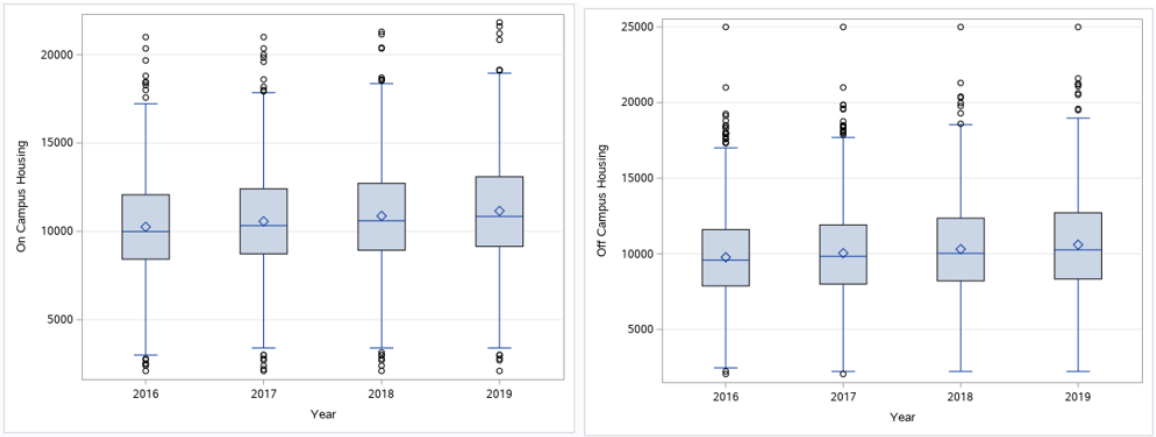


Fig 3:Box plots for On and Off campus Housing

The above box plots in Fig 2 are of in-state and out-of-state tuition fees. They are two of the most significant variables in the Cost of Attendance metric; both are on an upward trend. This upward trend is mimicked by the trend of rising housing costs of both on-

campus and off-campus housing options available to students. However, it should be noted that the rate of increase is higher in on-campus housing compared to off-campus options. The other variables like Books and Supplies and Other Expenses do not follow this pattern.

2.1.4. Modelling

For analyzing and prediction modelling Algorithms being considered for analysis:

ANOVA to identify the determinants for Cost of Attendance.

We perform correlation analysis and One-way ANOVA to identify the relation between the 7 variables deemed as significant and analyze our findings. Below are the charts and observations from Correlation analysis and ANOVA performed in SAS Studio.

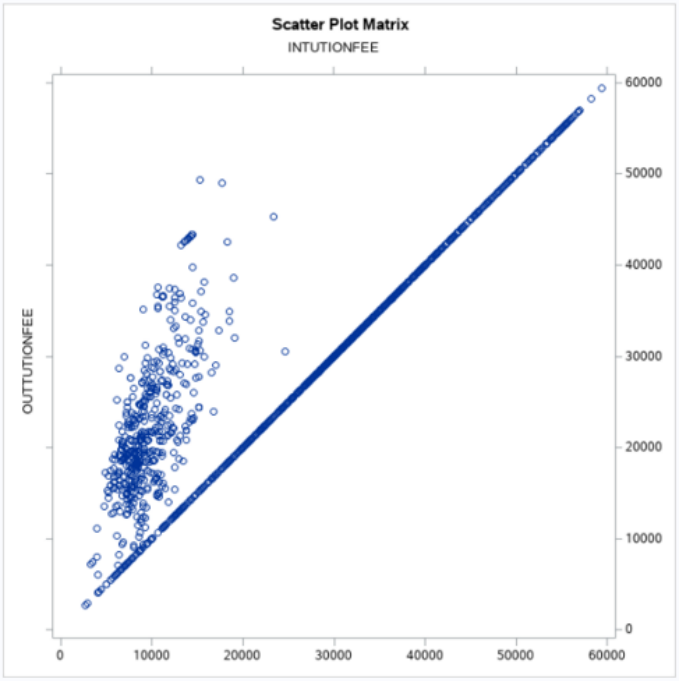


Fig 4: Correlation Analysis between In-State vs Out-of-State Tuition

Fig 4 shows the relation between In-state and out-of-state tuition fees. They have an almost linear relationship with out-of-state tuition having outliers with a great rise in costs compared to that of in-state tuition.

Fig 5 shows the correlation between tuition fees (in-state and out-of-state) and cost of Books and Supplies. The rate of increase in tuition costs do not impact this cost.

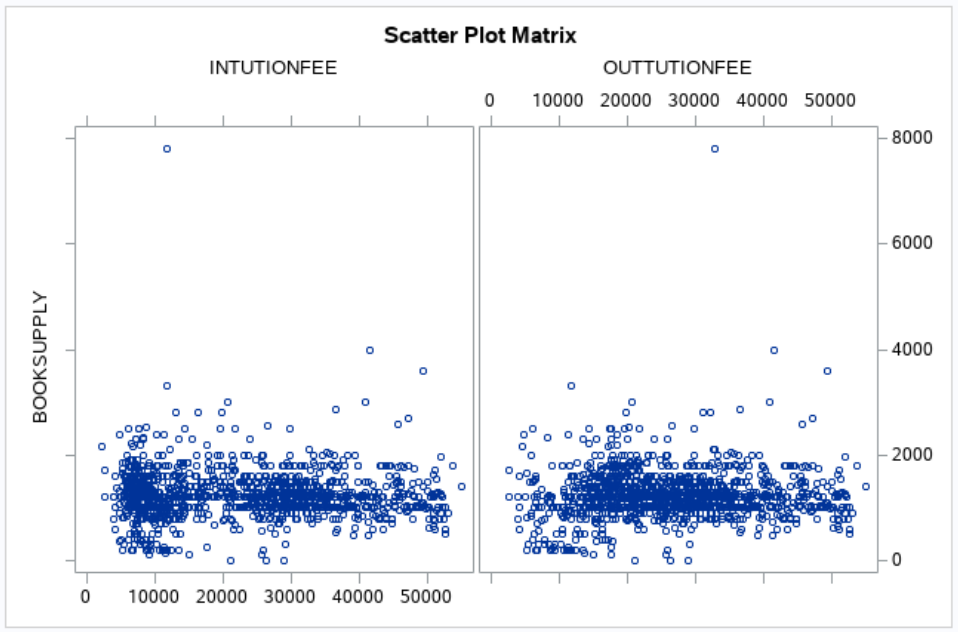


Fig 5: Correlation Analysis between Books and Supplies vs In-State & Out-of-State Tuition

Fig 6 shows the correlation between tuition fees and on-campus housing costs. As can be seen in plot, on-campus housing has a close to linear relation with tuition fees.

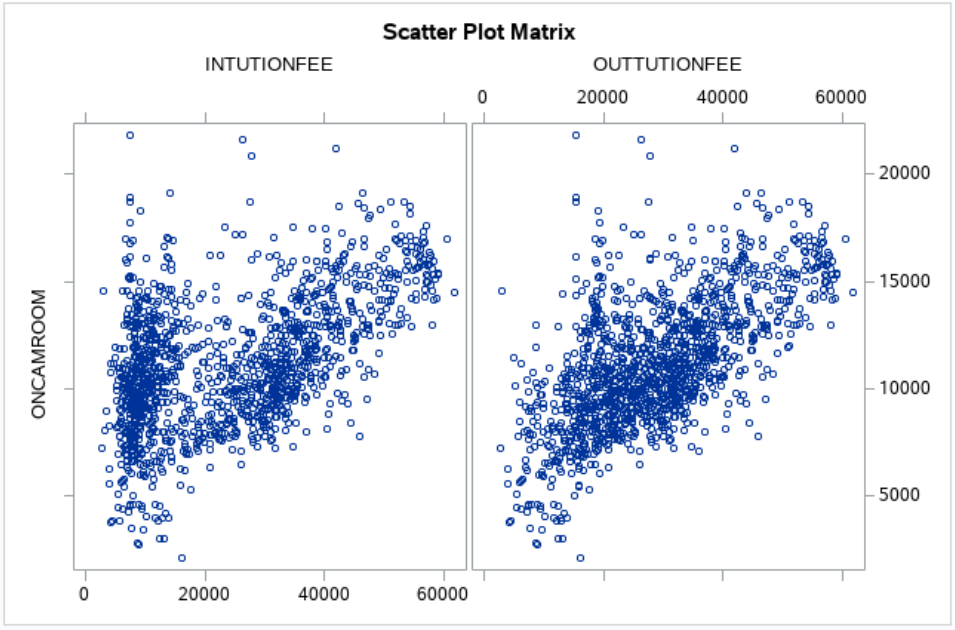


Fig 6: Correlation Analysis between On-Campus room vs In-State & Out-of-State Tuition

Fig 7 shows the correlation between tuition fees and on-campus housing costs. As can be seen in plot, off-campus housing has a relation with tuition fees, but they are not closely related.

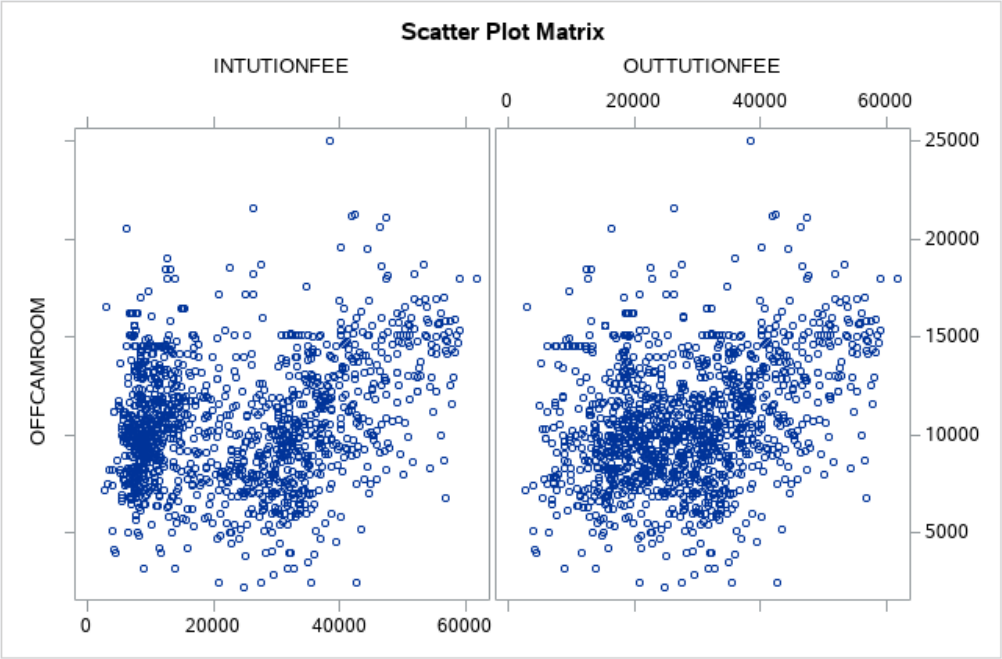


Fig 7: Correlation Analysis between Off-Campus room vs In-State & Out-of-State Tuition

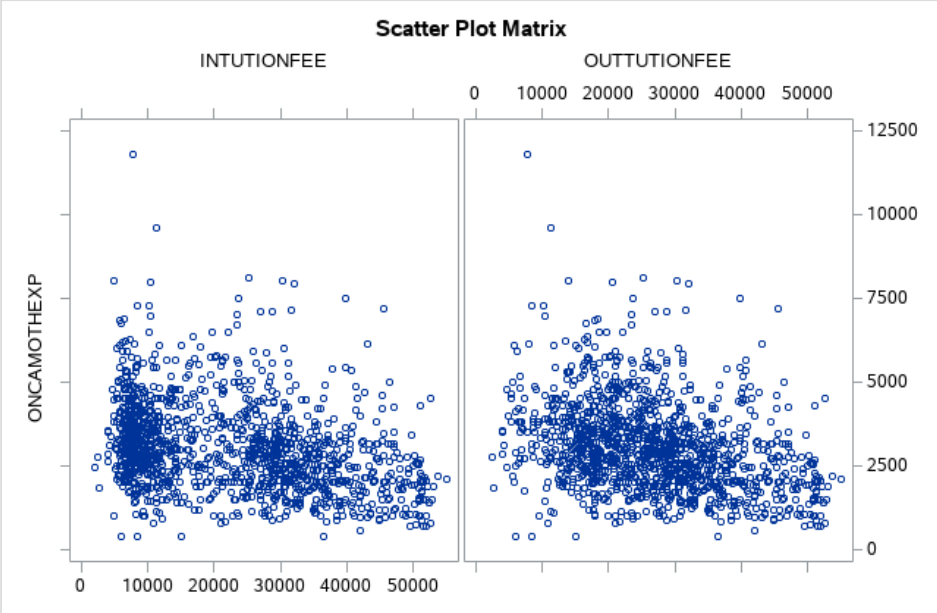


Fig 8: Correlation Analysis between On-Campus Expenses vs In-State & Out-of-State Tuition

Fig 8 and Fig 9 represent the correlation between other expenses (On and Off campus) and tuition fees. As can be seen in the box they do not show a strong relation between these variables.

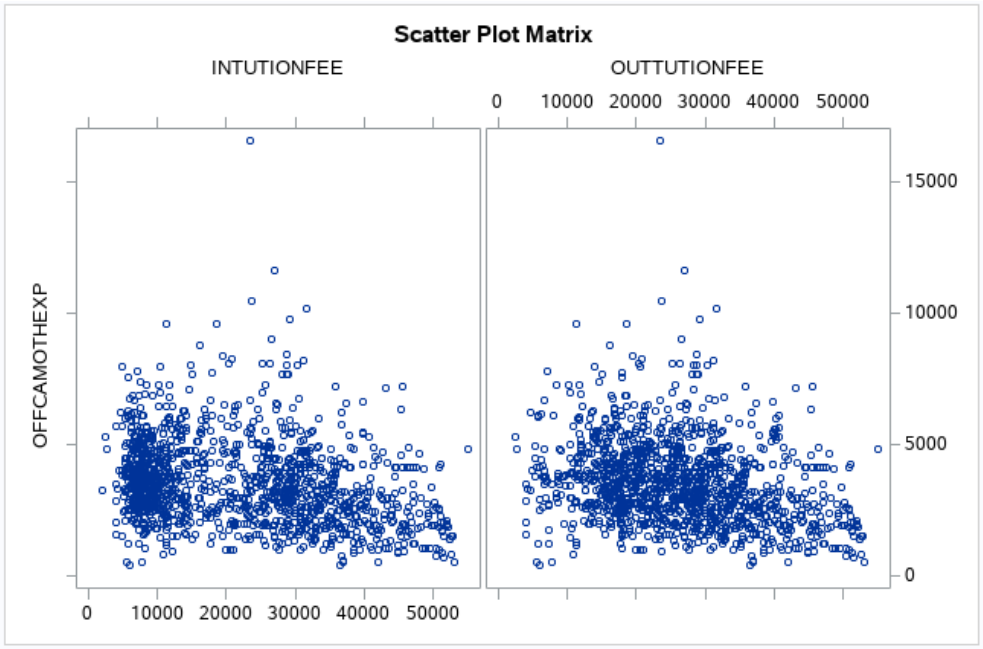


Fig 9: Correlation Analysis between Off-Campus Expenses vs In-State & Out-of-State Tuition

Class Level Information		
Class	Levels	Values
Year	4	2016 2017 2018 2019

Number of Observations Read	5472
Number of Observations Used	5471

Dependent Variable: INTUTIONFEE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2888727793.3	962909264.45	4.75	0.0026
Error	5467	1.1089348E12	202841559.34		
Corrected Total	5470	1.1118235E12			

R-Square	Coeff Var	Root MSE	INTUTIONFEE Mean
0.002598	61.60502	14242.25	23118.64

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Year	3	2888727793	962909264	4.75	0.0026

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Year	3	2888727793	962909264	4.75	0.0026

Class Level Information		
Class	Levels	Values
Year	4	2016 2017 2018 2019

Number of Observations Read	5472
Number of Observations Used	5471

Dependent Variable: OUTTUTIONFEE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3658397010.5	1219465670.2	9.11	<.0001
Error	5467	731978098816	133890268.67		
Corrected Total	5470	735636495826			

R-Square	Coeff Var	Root MSE	OUTTUTIONFEE Mean
0.004973	42.10778	11571.10	27479.71

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Year	3	3658397010	1219465670	9.11	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Year	3	3658397010	1219465670	9.11	<.0001

Fig 10: Result of ANOVA for In-state and out-of-state tuition

One way ANOVA was performed on all the seven variables and results of the analysis showed that means of in-state and out-of-state tuition have changed over the four-year period and in an upward trend. Results of the analysis are in above screenshot. Note the low P value < 0.05 significance. With a null hypothesis that each of the expenses has been the same across 4 years, the value of P proves or disproves the null hypothesis. A similar analysis of Books and Supplies show that this expense largely stayed the same. Both on-campus and off-campus housing changed and finally, other expenses have largely stayed the same.

Based on these observations we can conclude that the three major determinants for Cost of Attendance calculation are in-state tuition, out-of-state tuition, on-campus housing, and off-campus housing.

Multi Linear Regression to predict Cost of Attendance

We use a multi linear regression to create models that assume a linear relationship between a subset of our variables (X1, X2) and predict a set of variables (Y1, Y2, etc.). Using these predicted values, we can propose a list of predicted annual cost ranges for various components of Cost of Attendance with a significant confidence interval.

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n}$$

where:

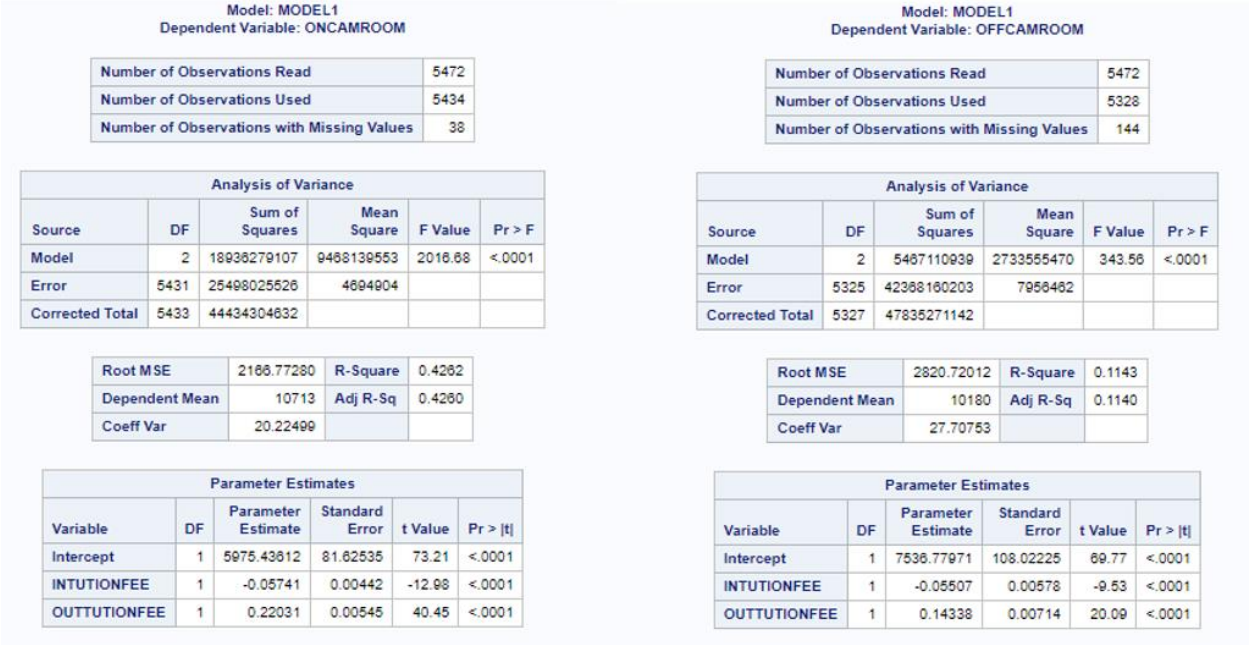
y is the predicted variable

x is the input variable(s)

β_0 is the intercept

$\beta_n$is the coefficient for x_n

For this linear regression model creation, we use Linear regression model defined in the sklearn module of Python. We create two subsets of the data for model creation and validation purposes. We consider the in-state tuition and out of state tuitions as the input variables and we predict the remaining components of the cost i.e., output variables. Below screenshot shows a similar regression results performed in SAS Studio. As the only variables that are correlated to two tuition components are the housing costs (on-campus and off-campus) we have these variables predicted using the regression model. Both books and supplies and other expenses costs can be estimated within 95% confidence using the range trend of these variables.



Model: MODEL1
Dependent Variable: ONCAMROOM

Number of Observations Read	5472
Number of Observations Used	5434
Number of Observations with Missing Values	38

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	18936279107	9468139553	2016.68	<.0001
Error	5431	25498025526	4694904		
Corrected Total	5433	44434304632			

Root MSE	2166.77280	R-Square	0.4262
Dependent Mean	10713	Adj R-Sq	0.4260
Coeff Var	20.22499		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5975.43612	81.62535	73.21	<.0001
INTUTIONFEE	1	-0.05741	0.00442	-12.98	<.0001
OUTTUTIONFEE	1	0.22031	0.00545	40.45	<.0001

Model: MODEL1
Dependent Variable: OFFCAMROOM

Number of Observations Read	5472
Number of Observations Used	5328
Number of Observations with Missing Values	144

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5467110939	2733555470	343.56	<.0001
Error	5325	42368160203	7956462		
Corrected Total	5327	47835271142			

Root MSE	2820.72012	R-Square	0.1143
Dependent Mean	10180	Adj R-Sq	0.1140
Coeff Var	27.70753		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7536.77971	108.02225	69.77	<.0001
INTUTIONFEE	1	-0.05507	0.00578	-9.53	<.0001
OUTTUTIONFEE	1	0.14338	0.00714	20.09	<.0001

Fig 11: Results of Multi-linear regression

2.1.5. Evaluation

To evaluate the overall fit and accuracy of our Linear model we use the train/test split method by splitting data into two subsections. We measure the root mean squared deviation and R-squared values for the model to evaluate the difference between values predicted by

our model and the values observed. The values of the evaluation parameters are displayed in the Fig 12. When interpreted in the context of our source data Root SME value are ~2100. Also, the R-squared value is between 0 and 0.6 showing our model ability to predict values with reasonable confidence.

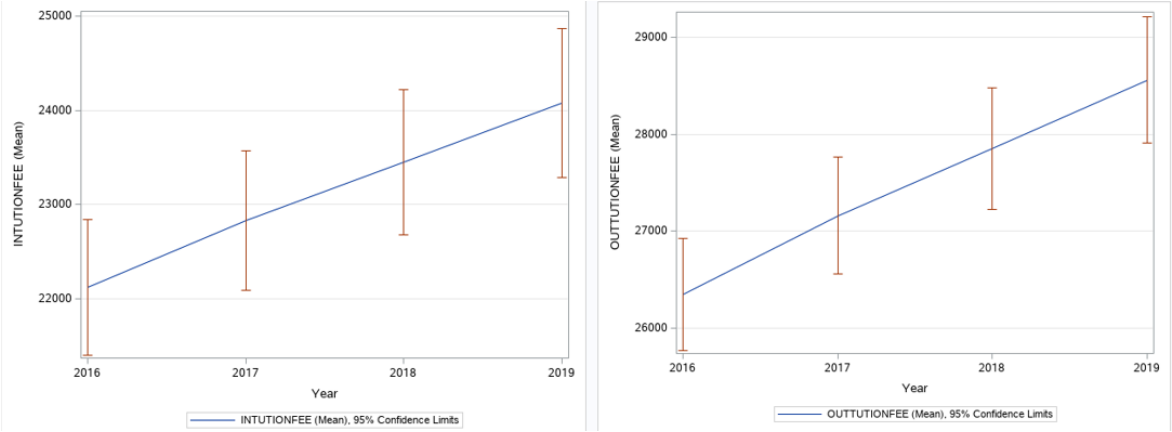


Fig 12: Cost patterns for tuition fees and 95% confidence interval for each

Based on the study of our data, we know that both in-state and out-of-state tuition costs published by institutes are the most significant variables in cost of attendance equation. Let us call them primary variables. The cost of housing variables directly depends on the on these primary variables. Using the trend of data for primary variables and extrapolating the confidence ranges for these variables we can get a range of possible primary variable values. Using this data along without prediction model we can predict the range of values for the secondary variables. The summation of the resulting primary and secondary variables should give us an estimate of total cost of attendance for college into the future. Based on our extrapolations we have an estimated in-state tuition range of [$23654, $25726] and estimated out-of-state tuition range of [$29410, $31630] with an approximate confidence range of 95%.

2.1.6. Deployment

Deployment is the process of using your new insights to make improvements within your organization. In general, the deployment phase of CRISP-DM includes two types of activities:

- Planning and monitoring the deployment of results
- Completing wrap-up tasks such as producing a final report and conducting a project review

In our case, as this is just a study of data and summarizing our observation and results, we do not have this component in our project scope.

3. Discussion of findings

In this section of the report, we will go back to the business question(s) we started with and let us try to answer the questions with the results of our analysis:

1. Can we predict the future cost of education?

   Based on our analysis and findings we see that the features in-state and out-of-state tuition have the highest significance among all considered features. Hence it is related to the cost of attendance calculation. The scatter plots from section 2.1.4 indicate the presence of absence of relationship between the various features considered. To conclude with the data given in the dataset and based on the assumptions we made, we can effectively predict the future cost of education.

2. Which components of college cost should students concentrate on to reduce expenses?

   Based on our analysis of data, the significant components of cost of education are tuition cost and housing costs. Students should try to concentrate on ways to decrease tuition component either by available scholarship or assistantships to decrease the

tuition costs. Housing costs can be controlled by choosing to stay further from the institute campus which might help them access to lower cost housing.

3. Can our prediction help future students in choosing a college experience with reasonable costs?

   Yes, we can offer couple of suggestions based on our data analysis.

   - Choosing a college within the state of residence gives you the highest savings as out-of-state residents end up paying significantly higher towards their tuition.
   - Choosing off-campus housing is another option to save. While on-campus housing offers significant benefits toward college experience, staying off campus and utilizing facilities offered with in the campus can help in offsetting the difference in college experience.
   - Books and supplies though not a huge portion of the overall cost of attendance add up to thousands of dollars through the course of four years of college, library facilities in college can be used to save.

References

1. IBM SPSS Modeler CRISP-DM Guide, IBM Corporation 2011.

   URL: https://inseaddataanalytics.github.io/INSEADAnalytics/CRISP_DM.pdf

2. Gregory Piatetsky**,** CRISP-DM, still the top methodology for analytics, data mining, or data science projects (2014).

   URL: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

3. CRISP-DM - Data Science Process Alliance

   URL: https://www.datascience-pm.com/crisp-dm-2/

4. Data Mining using CRISP-DM methodology

   URL: https://www.section.io/engineering-education/data-mining-using-crisp-dm-methodology/

5. Using CRISP-DM to Predict Car Prices

   URL: https://medium.com/@christophberns/using-crisp-dm-to-predict-car-prices-f15eb5b14025