Software Design Specification

for

Does Variations in Tuition Affect Incoming Student Enrollment at St. Thomas University?

Version 1.0

Prepared by Armando Zapata, Camilo Riviere, and Zimin Lu

Team Tuition

Creation: 10/7/18

Table of Contents

Table of Contents 2

Revision History 2

1. Introduction 1
   1. Goals 1
   2. Statement of scope 1
   3. Model context 2
   4. Major constraints 2
2. Data design 2
   1. Data sources 2
   2. Internal data structure 2

2.3Variable Description 3

1. Independent Variables 3

2. Sub-Independent Variables 3

3.Dependent Variables 3

2.4. Pre-design Analysis 4

2.5. Tidying procedure 5

2.6 Database description 5

1. Model Architecture 5
   1. Type of Model 5
   2. Training set 6
   3. Testing set 6
2. Approach 6
   1. Implementation Details 6
   2. System requirements 6
3. Testing Strategy 6
   1. Classes of tests 6
   2. Expected response 7
   3. Performance bounds 7
   4. External review and validation 7
4. References 7

Appendix A: Glossary 7

Appendix C: Issues List 7

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Date** | **Reason For Changes** | **Version** |
|  |  |  |  |
|  |  |  |  |

1. Introduction

Given the highly competitive landscape of college recruiting in the state of Florida, a report is being generated to shed light on a current trend that has been present at the institution for the past 10 years that can potentially provide insight on how to increase the school’s attractiveness. To help with this matter, a preliminary analysis will be performed to determine if there is a correlation between variations in tuition cost and the number of incoming students over the last 10 years at St. Thomas University. Further analysis will include sub-variables such as:

* Differences in domestic vs international students that conformed the student body for each year.
* Differences in ethnicities that made up the entire undergraduate population.
* Gender, proportion of male and female incoming undergraduate students for each year and spot variation within each year.

Part of the motivation behind this project is to provide insight into variations in the undergraduate student population over the last 10 years. By providing a report derived from a regression analysis, the individuals who ultimate have the final say in institution matters can effectively allocate resources and even consider implementing a new strategy that could potentially increase the number of incoming students in the years to come.

* 1. Goals
* Develop a predictive model to establish a relationship between variations in tuition cost and incoming undergraduate student enrollment.
* Model’s predictive power should provide predictions based on STU data only about tuition cost and incoming undergraduate student enrollment.
  1. Statement of scope

Given the limited timeframe of 5 weeks to fulfill the business sponsor’s desires, it has been established that 10 years-worth of data is broad enough to create and implement a model. To possibly increase the model’s accuracy, including incoming full-time graduate students for the different schools in St. Thomas University as part of the whole sample size was considered. However, the presence of different tuition rates for each respective school, as well as difficulty in accessing the preliminary data needed to assess this additional variable made it infeasible. Also, it could potentially create flaws to the model rather than increase its validity.

With the preliminary sample data containing approximately 23,000 of student records that was gathered during the discovery phase, the data was aggregated into 10 data points which each corresponded to a specific year in the 10-year scope. These 10 data points represent the independent and dependent variables that will be inputted into the model.

The 10-year scope assigned for this project is considered a good sample size given the degree of variations in tuition cost each year. A trend can be useful to provide insights to the university’s President about the presence of a relationship, and moving forward, subsequent models can be built upon to generate more valuable insights that could potential increase the institution’s attractiveness.

* 1. Model context

A critical part of the project consists of creating a model that can validate the results the it will generate. Considering the nature of the team’s 2-variable problem, the model that will be created will be based on a regression analysis. For simplicity purposes, the team will create and test the model by implementing the programming language R. Moving further and deep into the analytical process, the validity of the model will be tested by inputting the different sub-variables and increase the model’s accuracy.

* 1. Major constraints

For the purposes of this project, no major constraints were encountered that would limit the model. There were no clearly-defined guidelines or requirements as to what specific programming language needed to be used, or if this model was going to be integrated into the university’s website.

Still, a possible constraint during the model building phase is that the model may only accept school year figures. Therefore, the model will take years of data with larger variance in price differences to train a confident prediction model and increase its validity even more.

1. Data design

Aside from using the primary independent and dependent variables established, additional sub-variables will be incorporated, such as differences in number of domestic and international incoming undergraduate students, total number of different ethnicities that represent the sample population each year, and gender difference, meaning total number of male and female students each year for incoming undergraduate students in order to train the confident prediction model and improve its validity.

* 1. Data sources

The sample data was retrieved from a publicly available source using SQL queries. The information gathered in regard to undergraduate incoming student’s records was completely anonymous. The structured data gathered by the team was formatted in Excel, where some preliminary testing was performed to generate a graphic representation of the sample data in order to visualize the changes and subtleties that were present within the data.

* 1. Internal data structures

Input Data:

* *Incoming* - Data frame containing a set of 10 elements of type where each element in the set represents the total number of incoming students for each academic year for the previous 10 academic years.
  + Data Structure: Data frame
  + Data Type: Integer
* *Tuition* –Data frame containing a set of 10 elements of type integer (converted to numeric data type in R code) where each element in the set represents the total price for tuition for each academic year for the previous 10 academic years.
  + Data Structure: Data frame
  + Data Type: Integer
* *Year* – Data frame containing a set of 10 elements of type factor where each element in the set represents each academic year. Each element was exclusively used for labeling data points on the graphs produced by the linear model.
  + Data Structure: Data frame
  + Data Type: Factor
* *training\_set –* Data frame containing a set of 8 elements of type integer. *training\_set is a* randomly sampled subset of data structures *Incoming, Tuition,* and *Year. training\_set was* utilized to train the linear model.
  + Data Structure: Data frame
  + Data Type: Integer
* *test\_set* – Data frame containing a set of 2 elements of type integer. *test\_set* is a randomly sampled subset of data structures *Incoming, Tuition, and Year.* *test\_set* was utilized to compare the results of our trained linear model.
  + Data Structure: Data frame
  + Data Type: Integer

2.3 Variable Description

1. Independent Variables

1. Tuition price

2. Description: Price per year of tuition at St. Thomas University over the 10-year span.

3. Numeric

2. Sub-Independent Variables

1. Gender

1.1 Description: each incoming student’s gender

1.2 Textual

2. Ethnicity

2.1 Proportion of incoming undergraduate students among different ethnicities.

2.2 Numeric

3. International incoming students vs. domestic incoming students.

3.1 Number of incoming international students and domestic students each year over the 10-year scope.

3.2 Numeric

3. Dependent Variables

1. Incoming Student Enrollment

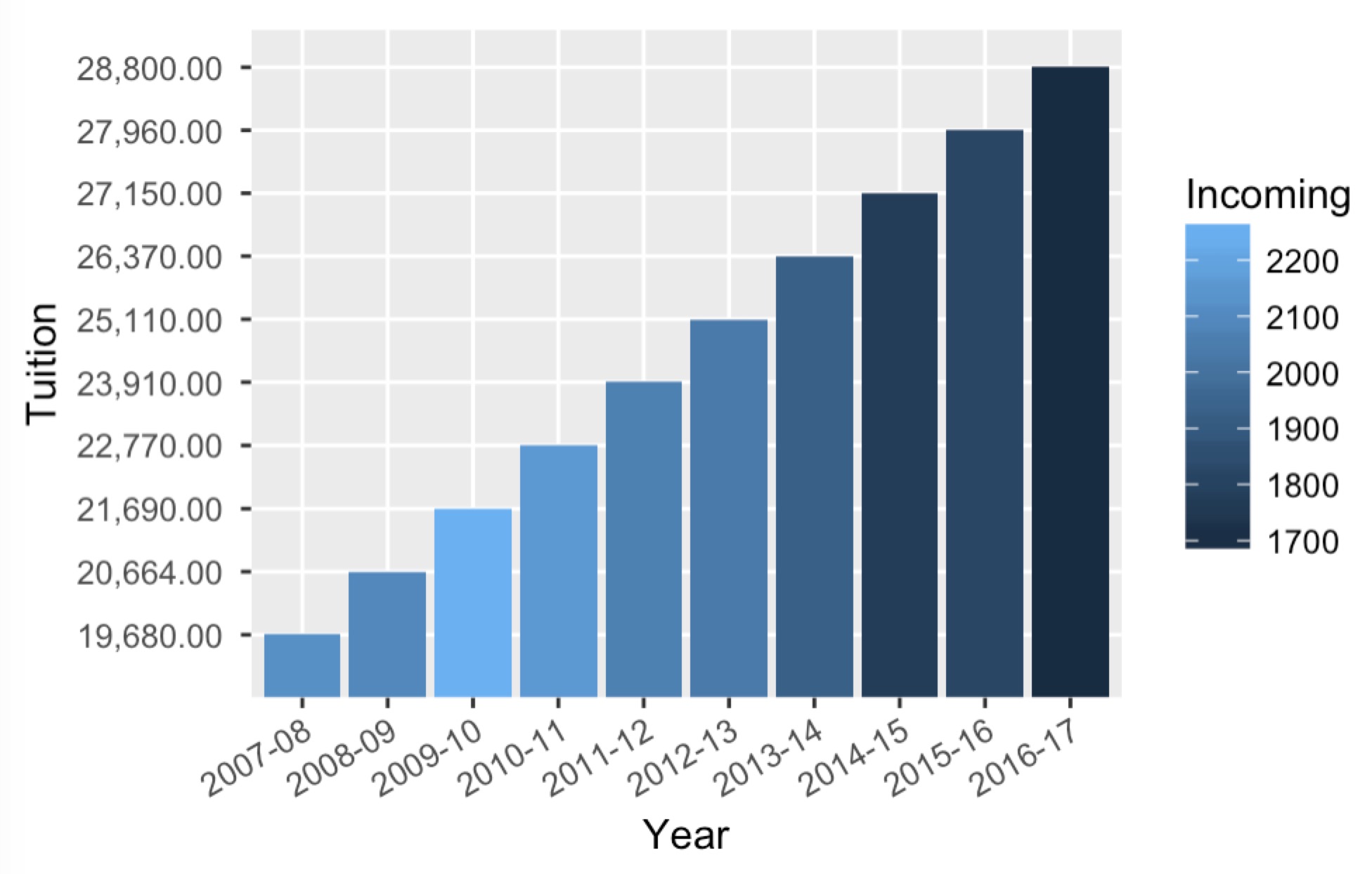
2. Description: Number of incoming students at STU who are enrolling in their first semester.

3. Numeric

2.4. Pre-design Analysis

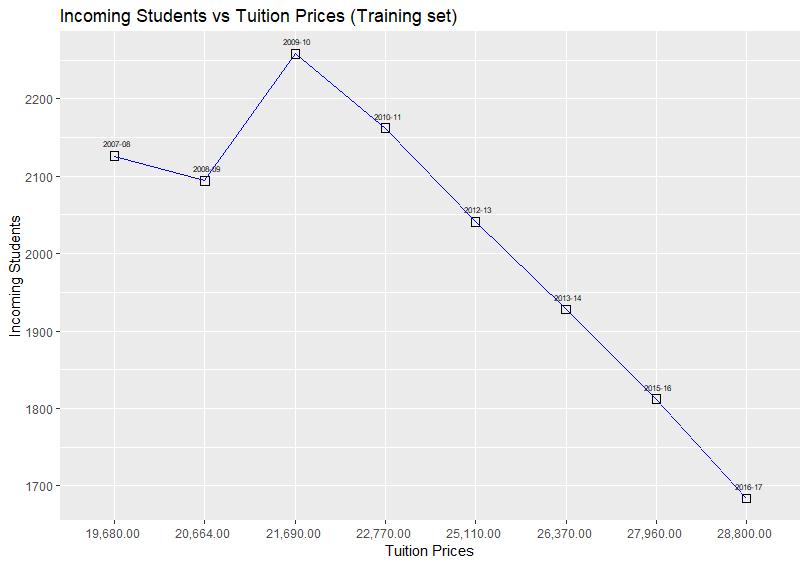
Before the model was designed, other variables were considered for observation such as proportion of total number of students per semester in a calendar year. During this sub phase, graphs were produced such as histograms and line charts to show the quantity and proportion each year. However, because the term division in each calendar year is not the same, it is hard to make an equal comparison.

It was also taken into consideration other factors such as how graduation rate could potentially affect incoming student enrollment. Still, due to the small proportion of influence, it was not included as part of the sample data set. At the same time, statistical calculations were performed on all full-time incoming undergraduate student data points and tuition rates, and it resulted in a negative correlation with a value of -0.7501.



Annual tuition changes

Moving forward as more in-depth analysis is conducted, a more accurate linear model will be developed by inputting the sub-variables in order to enhance the validity of the model.



Incoming students vs TUITION PRICES

2.5. Tidying procedure

In terms of normalization and classification processes, the total number of incoming undergraduate students was aggregated for each year as well as total tuition rate for the full-time incoming undergraduate students corresponding to that year, in order to input them in the initial rounds for the testing phase of the model. Within the spreadsheet in Excel, a SUM function was used to aggregate the total number of incoming undergraduate students each year. Total tuition for each year was directly inputted without using any function or procedure. This value is a predetermined fixed value with no cleaning or further tidying required.

2.6 Database description

The project’s spreadsheet contains anonymized data on incoming undergraduate student enrollment at St. Thomas University from the years 2007 to 2017. The sample data gathered, which was gathered in a CSV file, represents approximately 2,1980 students, with information such as term start time, reporting time, degree time, grade, gender, country of citizenship, race, student’s major, year, tuition per year, and total number of incoming undergraduate student. These variables were sorted and retrieved via SQL queries and formatted into an Exel file, and subsequently imported into Rstudio, where screening tests were subsequently performed.

1. Model Architecture

The initial model is based on a CSV file that was formatted in Excel. During the data preparation phase, a preview of the various independent variables contained in the data was done to determine the final independent variables used.

“Gender”, “race”, “International and domestic student” were used in order to develop a bar chart and that would indicate any changes and relationships between them. “Term” refers to the year a student is starting its college career, and it was used to generate a graphs in order to plot a trend performance for incoming undergraduate students in relation to each specific year.

* 1. Type of Model

Model used:

* Linear regression: Used to model the relationship between variations in tuition fees and the number of incoming students. Based on the fact that both variables are numeric ones, the model can use point-polt. It’s worth mentioning that regression models are inherently brittle to only the information that is given, so one can only make predictions on the information inputted to the model. These predictions doesn’t necessarily mean are going to be 100% guaranteed. A line chart was not used because the observation in each year is one, so it was deemed best to make a model with a dot plot.
  1. Training set

Training set constituted by “student id”, “term start time”, “term”, “reporting time”, “degree time”, “grade”, “gender”, “ccountry of ccitizenship”, “race”, “major”, “school”, “year”, “tuition per year”, “total number of incoming student".

Training set selected by enrollment students from 2007 to 2018 years.

Total data number is 21980.

* 1. Testing set

The testing set will constitute an initial test for the first 5 data points for the odd years of the project’s scope. These data points are based on the aggregated value of the independent variable corresponding to the tuition cost, as well as the dependent variable pertaining to the number of incoming students.

The total raw data extracted from the database was 21,980 cells.

There was a data set that was excluded from the model, and it referred to the different schools within St. Thomas University’s graduate program, degree date for each student, reporting time, major, and student ID, since the project requires complete anonymity and this set of data was deemed not relevant to the scope of the project.

1. Approach
   1. Implementation Details

Computer languages and versions: R version 3.4.3

Platform: x86\_64-apple-darwin15.6.0 (64-bit)

Running under: macOS High Sierra 10.13.6

libraries: tidyverse

* 1. System requirements

minimum hard drive space：16.7M

network bandwidth：Y

RMA amont：226.7M

CPU speed：0.8%/CPU

1. Testing Strategy
   1. Classes of tests

* White box testing - Unit testing

Base case:

* Input –
  + Data Frame Training Set (8 rows).
  + Data Frame Test Set (2 rows).
* Expected Output –
  + Linear graph with 8 data points from training set.
  + Linear graph with 2 data points from test set to compare with results of training set.

Edge case:

* Input –
  + Training Set (With 11 rows), and Test Set (With 3 rows)
* Expected Output –
  + Index out of bounds error.
* Black box testing – User acceptance testing.

Base case:

* Input –
  + csv file with 10 rows of data.
* Expected Output –
  + Linear graph with 8 data points from training set.
  + Linear graph with 2 data points from test set to compare with results of training set.

Edge case:

* Input –
  + Csv file with 11 rows of data.
* Expected Output –
  + Index out of bounds error
  1. Expected response

2 Linear regression graphs. One linear regression graph created with a training set and another linear regression graph created with the test set to compare with training set linear regression graph. with demonstrating negative correlation between student enrollment and tuition.

* 1. Performance bounds

Performance bounds are subject to the limitations of the hardware of our execution machine:

RAM- 8GB

Hard drive – 80GB

Processor – Intel i5

* 1. External review and validation

The validation process for the model was done by performing a series of testing sets. Among them was the 80-20 test, in which 80% of the observations collected were used for the training set and 20% of the observations collected were used for the test set. The test set was randomly sampled.

For this project, there is not an ideal way to externally test the model, especially since the purpose of the project is to identify a relationship between two variables.

1. References

Appendix A: Glossary

Correlation – A mutual relationship or connection between two or more things.

Linear Regression – In statistics, linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression.

Data model – A data model is an abstract model that organizes elements of data and standardizes how they relate to one another and to properties of the real world entities.

White-box testing – White box testing is a method of testing software that tests internal structures or workings of an application, as opposed to its functionality. In white-box testing an internal perspective of the system, as well as programming skills, are used to design test cases.

Black-box testing – Black-box testing is a method of software testing that examines the functionality of an application without peering into its internal structures or workings. This method of testing can be applied virtually to every level of software testing: unit, integration, system and acceptance.

Subset – A set of which all elements are contained in another set.

Appendix B: Issues List

* Result interpretation is a potential issue since this is the first time this type of analysis has been done for St. Thomas university.
* External data such as currency inflation was not considered in the analysis due to time constraints.
* Tidying procedures vary depending on the specific project and the model used, therefore selecting the right tidying method for the data will require a careful evaluation and may pose issues deeper into the analytical stages.