An Analysis of Retention Rates

****
****

Springboard Capstone Data Science Project
Presenter: Sandra Rucker
Date: August 3, 2019

The Problem

The viability of Higher Education Institutions relies on their ability to retain and graduate students. The revenues of both nonprofit and for profit educational institutions are affected by the number of students that attend the institutions. Hence retention rate is an important factor for institutions to understand and predict. The primary goal of this project is to gain a deeper understanding of the relationship between the retention rate and other selected variables at Higher Education Institutions, i.e., Post-Secondary Institutions in the state of Georgia. We decided to focus on data obtained in the year 2017. The data was obtained from a very credible source. The data was obtained from one of the United States governments' educational databases. Understanding important variables which affect retention rates at Institutions will permit administrators to make more informed decisions regarding the adequacy of their pricing schemes. So, we will explore several issues related to Retention rates

Our Initial Research Inquiries and the Variables

Our primary explanatory variable of interest is the Student Faculty ratio. This is a value which tells us how many students on average are associated with one professor in the classroom at a particular Institution. First the average values for the Retention rates and the Student Faculty ratios will be determined. Secondly, we will explore whether or how Student Faculty Ratios at Institutions affect Retention rates. Next we will examine whether other variables significantly affect Retention rates. Some of the other variables which will be evaluated include the prices for attending the Institutions, and the gender of students.

Initial Questions of Interest

1. What is the average retention rate for post secondary institutions in the state of Georgia?
2. What is the average student to faculty ratio in the state of Georgia?
3. What is the relationship between the Retention rate and the Student to Faculty ratio?
4. Is the relationship between Retention rate and Student to Faculty ratio statistically significant?
5. Are the relationships between Retention rate and the other variables in the dataset statistically significant?
6. What are the best explanatory variable(s) that could be used to predict the Retention rate?

The Datasets

The data for this project was obtained from the National Center for Education Statistics IPEDS Institutional Data for 2017. This data set includes information for all 171 Post Secondary Institutions in the State of Georgia in the United States of America. Retention rate is the dependent variable, i.e. what we want to predict. The data is available at http://www.data.gov. Specifically data from the Integrated Postsecondary Education System (IPEDS) was used for this project, i.e. https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx. The initial data set contained the following data; Institutions in the State of Georgia for 2017, where n = 171;

Variables

Full Variable Name	Abbreviation
Total price for in state students living on campus	Price_IO
Total price for out-of-state students living on campus	Price_OO
Total price for out-of-state students living off campus	Price_OF
Total price for in state students living off campus	Price_IF
Full Time Retention Rate	Reten
Student-to-faculty ratio	SF_Ratio
Total men undergraduate	Men
Total women undergraduate	Wom

Methods

Data wrangling, statistics, and regression analysis were used in analyzing possible relationships between the predictor variable and the explanatory variables. The best variables for predicting the retention rate will be identified.

Rationale for Removing Data

Some Institutions in the state of Georgia do not have dormitories or on campus housing for students and do not have price data for those variables. Two columns (variables) were removed based on the huge amount of missing data. The two columns removed were Total price for in state students living on campus (Price_IO), and Total price for of state students living on campus (Price_OO). Sixty-five percent, of the Institutions had missing data in these columns. The institutions either do not have on campus housing or did not submit data for variables as it relates to Total price for on campus housing in the year 2017.

Removing Missing Data

The variable we wish to predict is Retention Rate. There were 26 Institutions which listed no data regarding Retention Rate. The value of 26 was obtained in the following manner; there was 16 Institutions which provided no data for any of the 8 variables we are considering, and there were 10 Institutions which provided no data values for Retention in the year 2017.

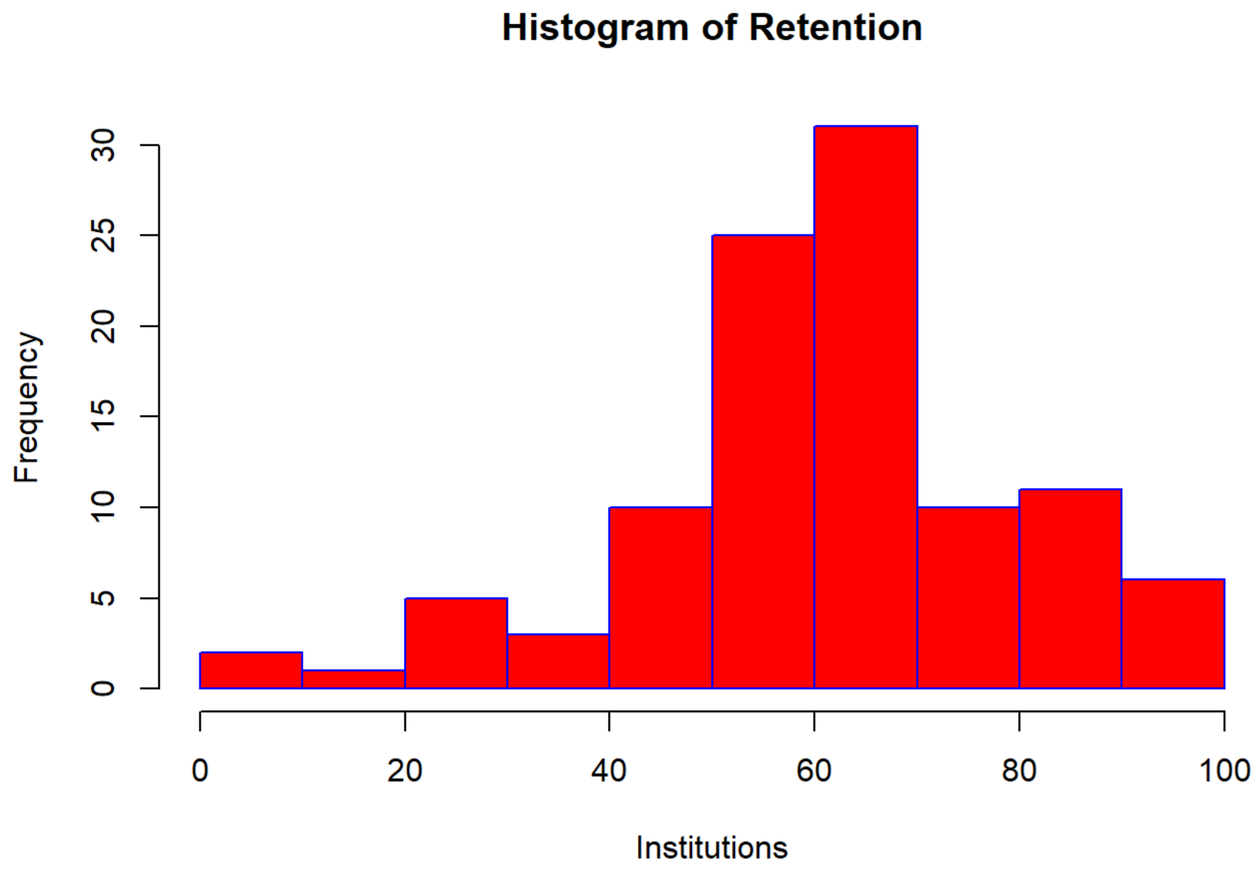
Considerations for the Statistical Analyses

The data was downloaded from the site https://www.data.gov in a certain format. We are interested in the Retention Rate and variables which may affect retention. The Retention data was not given at the student level, i.e. where 1 would specify that a student is retained, and 0 would specify that a student was not retained. Only summarized data is given, not 0’s and 1’s. Hence, the data given from the site was at the summary level. That is, the percentage of students who were retained, or returned was listed for each institution in the data set.

Average Retention Rates in the State of Georgia

The first issue addressed is the values of the retention rates at the Institutions in the study, and to find the average value of those retention rates. To obtain a visual picture of the retention rates, a histogram was created. Viewing the histogram gives us a feel for the average values of the “*mean*” and “*median*”. But to get the precise values, we followed by running the summary.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	53.00	63.00	61.61	71.00	100.00

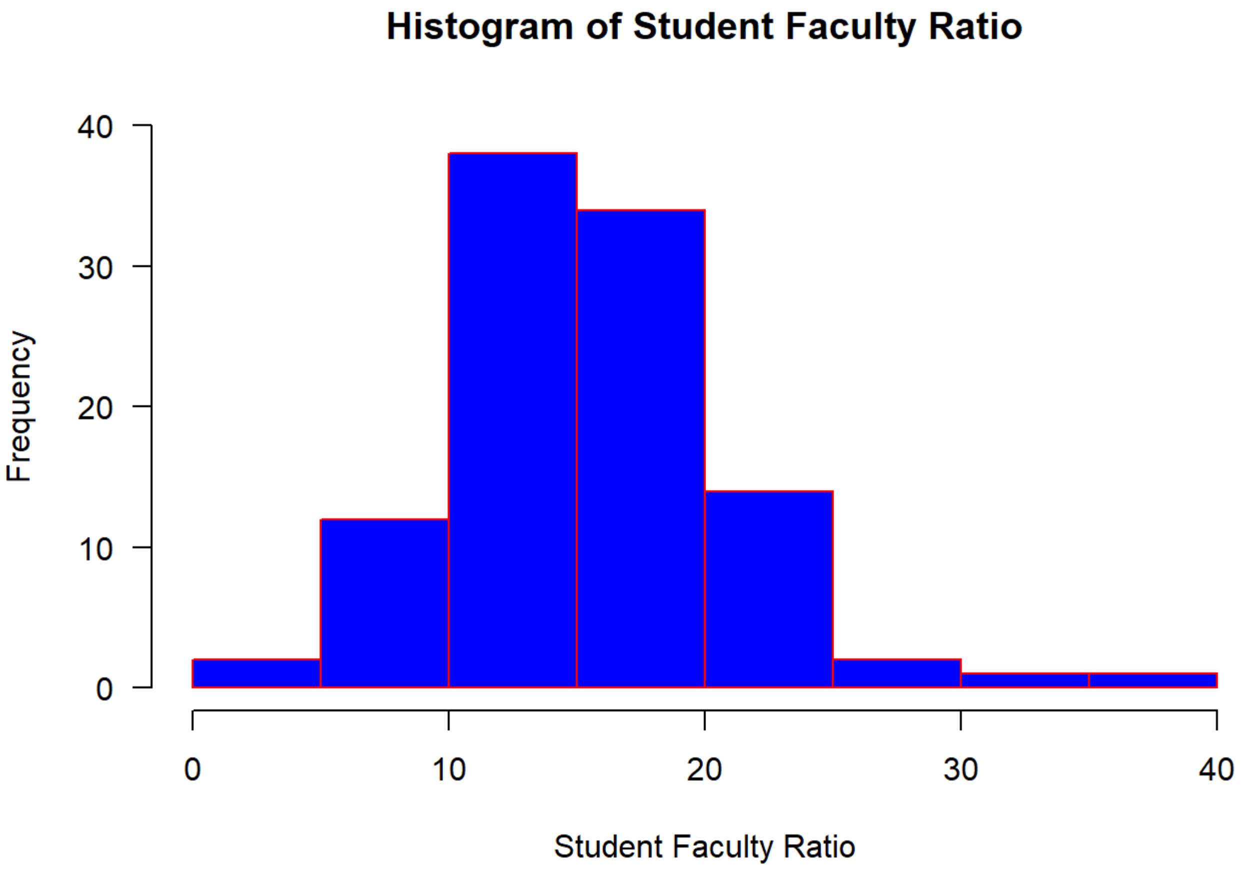


The summary shows that the mean Retention rate is 61.61.

Student Faculty Ratio

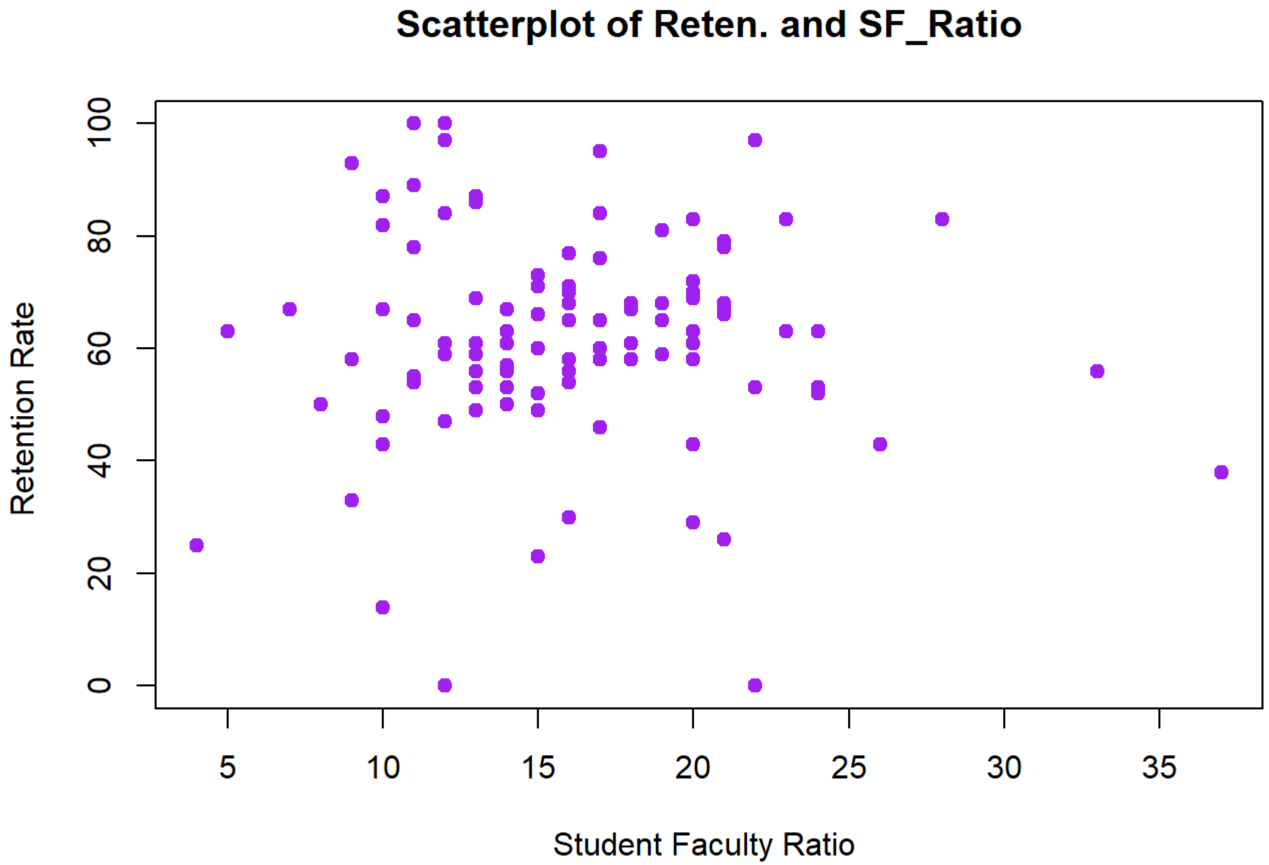
The second issue we explored was the Student Faculty ratio at Institutions. We created a histogram which shows the Student Faculty ratios at the institutions. The visual gives an indication of the average values, but more precise information regarding the average values for the mean and median are given in the summary.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.00	12.00	15.50	15.92	20.00	37.00



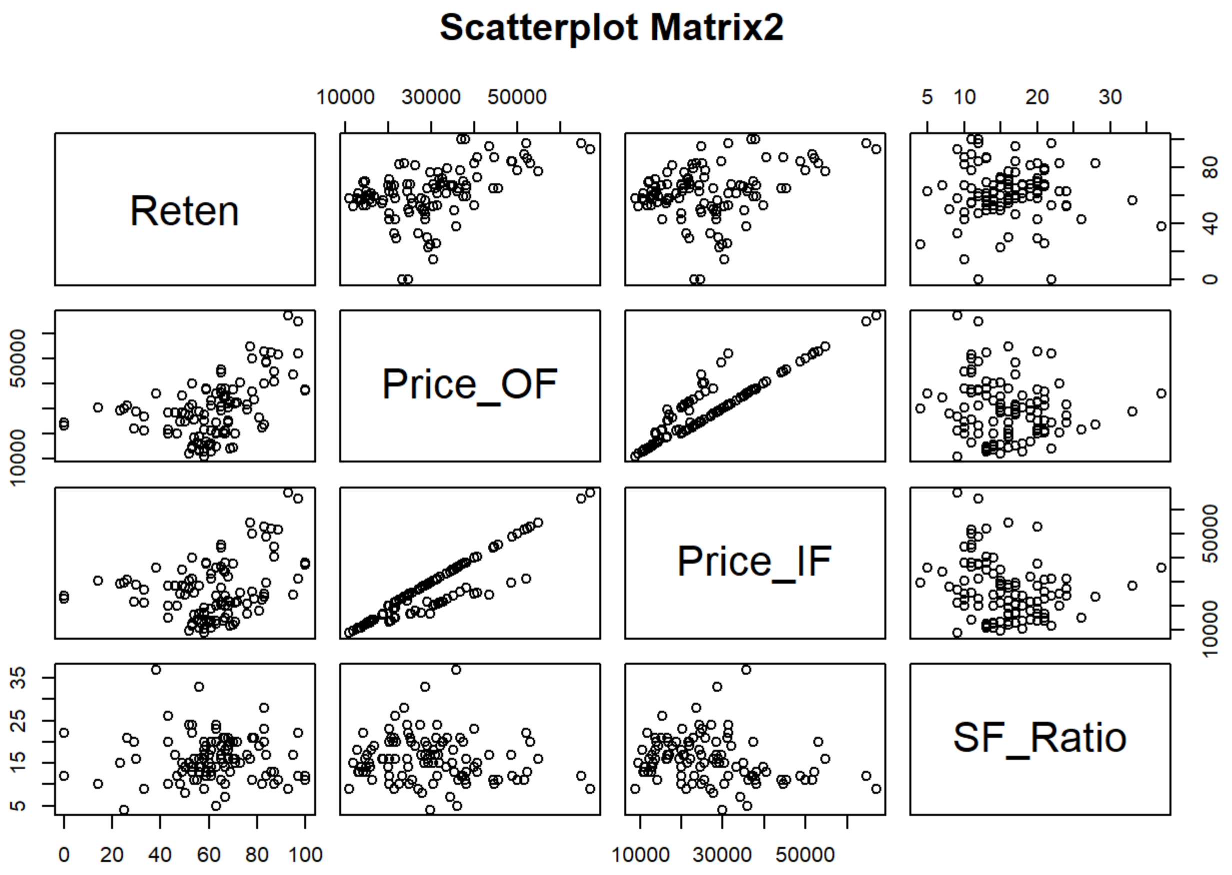
The mean value is 15.92, and the median is 15.50, which round up to 16. So we observe that on the average there are 16 students to 1 professor in the college classrooms in the state of Georgia.

Our next step is to examine the relationships of the variables in the data set. We first created a scatterplot, because scatterplots tell us if the variables seem to be related in a meaningful way.



The Scatter Matrix

We then developed several scatter matrices which show relationships between all variables in the dataset. Inspection of the scatter matrix below shows that there is a linear relationship between the price of students who live off campus and Retention. Possible correlations between the explanatory variables also appear. This was investigated.



The Model

Since the output variable is a continuous variable rounded to the nearest unit, a linear regression model was used. Multiple variable linear regression seemed to fit best because there is more than one independent variable used to predict a college’s retention rate. In order to explore the potential relationships between variables related to retention rates in colleges and universities in the state of Georgia.

Linear Regression Models

The linear regression model addresses the following research questions.

Research Question 1: Is the relationship between Retention rate and Student to Faculty ratio statistically significant?

Research Question 2: Are the relationships between Retention rate and the other variables in the dataset statistically significant?

Research Question 3: What are the best explanatory variable(s) that could be used to predict Retention rate?

Multiple Linear Regression Model

We begin with a multiple linear regression model which includes all the variables in our dataset following the data wrangling process. In multiple linear regression we can first consider all the explanatory variables, and after evaluating the summary report, remove variables which are not useful. That is, we will remove from the model variables which are not significantly significant. This method is called the backward elimination strategy. The backward elimination strategy begins with the full model which includes all possible explanatory variables and eliminates variables one at a time, until we are only left with variables which are statistically significant.

```
Model1 <-m(stats$Reten~stats$SF_Ratio+stats$Wom+stats$Price_OF+stats$Price_IF+stats$Men)
```

Model1 and p-values

The most important column in the summary table is the column which indicates the probability values or the p-values. The p-value indicates whether a particular explanatory variable is statistically significant in the given model. The summary data for the first multiple linear regression model, Model1, indicates that the price for out of state students who live off campus Price_OF has ** which means that this value has statistical significance in this model. The p-value for Price_OF is .00121. The variable Price_IF has **.** listed which means that it has significance, but not at or below the .05 level. We reject the Null hypothesis, and there is a high statistically significant relationship between the predictor variable, Retention, and the variable Price_OF. Note that the variable Men has the largest p-value and is not statistically significant in Model1. The adjusted R squared value is .3186.

Model2

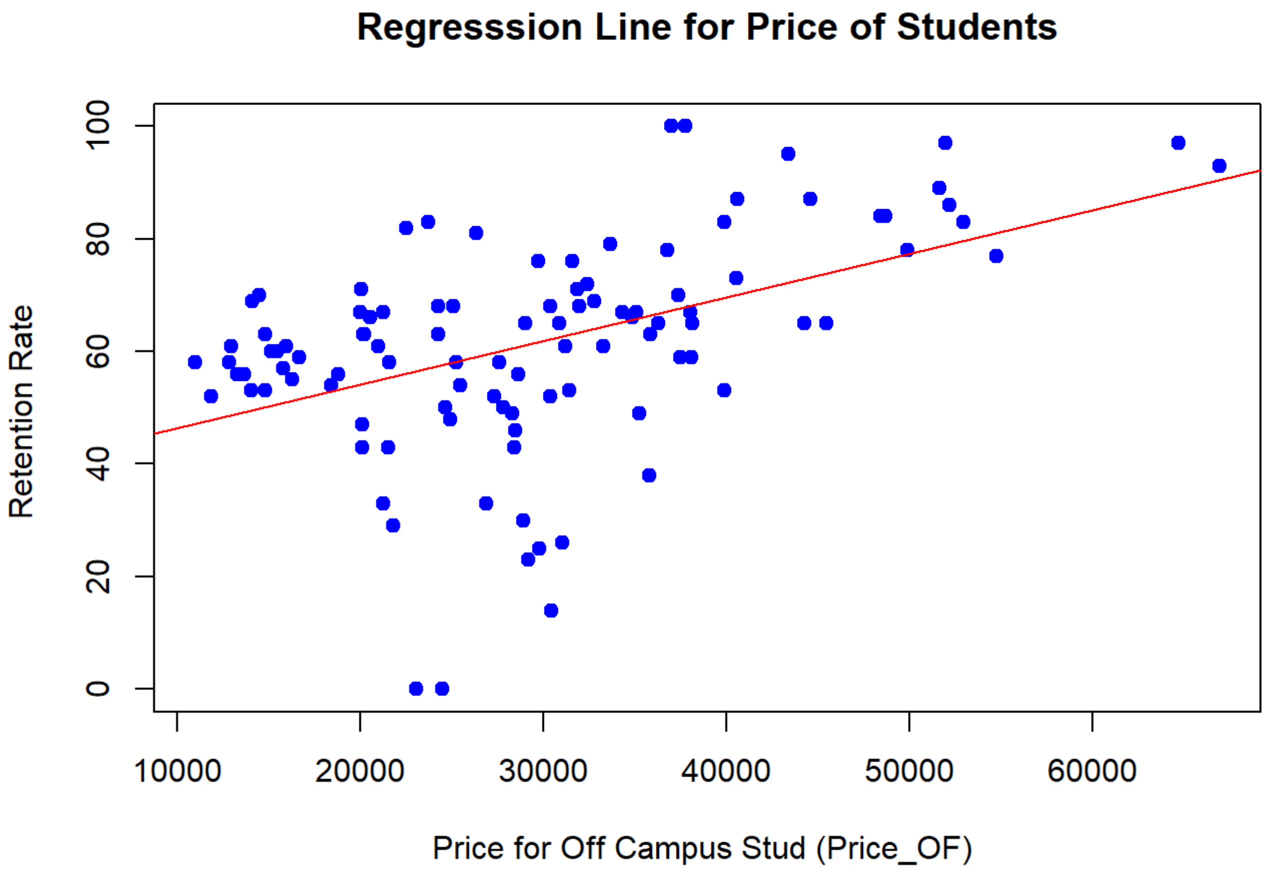
We ran the second linear multiple regression model, Model2, omitting the variable denoting the number of undergraduate men, Men. We omitted Men because it had the largest p-value in Model1. The p-value for Men was .754 in the previous model. We also note that the variable Men and Wom, are highly correlated, hence only 1 of these variables should be used in a linear multiple regression model. In the current model, Model2, the variable SF_Ratio has the largest p value, p = .31357 and is not statistically significant in the model. This variable will be removed in the next model. The variable Price_OF still has **, i.e. high statistical significance with p-value =.00117 which is less than its p-value in the previous model. The adjusted R squared value has increased in this model, Model2 to .3248.

Examination of all Models

We evaluated many models to determine which model contained variables which are most likely to best predict the Retention rate, and are highly statistically significant to the Retention rate. Six models were carefully studied using the backward elimination strategy. We will summarize the findings in the Conclusion.

A Regression Line

We ran many plots. We show the one below for the reason given. Price_OF has the lowest p-value, and hence highest significance level, a plot of the regression line for Price of out of state students who live off campus and Retention rate was completed.



Conclusions

We found that our original speculation was incorrect. We thought that the Student Faculty ratio might have a significant relationship with the Retention rate. The analyses show that Student Faculty Ratio (SF_Ratio) is not statistically significant for predicting the Retention rate (Reten). There were many issues to consider when determining the best model given the variables under consideration. After all considerations, we are left with two models for more discussion, Model4 includes one explanatory variable, Price_OF and Model6 includes two explanatory variables, Price_OF and Wom. Even though Model6 has the lowest p-value, i.e. p-value = .0000003 for Price_OF. The adjusted R squared value for Model6 is .2202. Model4 has p-value = .000000536 for the price of attendance for out of state students who live off campus (Price_OF) and .000239 for the number of undergraduate women (Wom). The adjusted R squared value for Model4 is .3115. We will select Model4 as the best model given our data set because the adjusted R squared value is higher and both variables have high statistical significance, i.e. p-values below .05, which corresponds to rejecting the Null hypothesis.

Conclusion

Hence the strongest predictors of Retention were the price of attendance for students from out of state who lived off campus, Price_OF and the number of undergraduate women students, Wom. It is noted that in the state of Georgia, the cost of tuition at state colleges and universities is usually more expensive for out of state students than the cost for in state students who are Georgia residents. It could be that their ability to continue at the institutions, i.e., their ability to be retained is cost dependent. Therefore, the Retention rates at these institutions seem to be statistically related to the price charged to out of state students who live off campus and the number of undergraduate women who attend the instituion. The other variables considered were not as strongly related to Retention.

Recommendations

The study gives evidence to support the recommendation that colleges and universities in the state of Georgia, consistently evaluate their prices if student retention is an issue of concern and importance. Secondly, we note that discussions regarding equalizing the cost of attendance for out of state and in state students at state colleges be considered if increasing the Retention Rate is a priority. Results indicate that the number of undergraduate women who attend these institutions should also enter into considerations. Given that the number of men and women are highly correlated, increasing the male population at these institutions also is advised.

Future Work

All post-secondary Institutions in the state of Georgia were included in the initial data set of this study. The state of Georgia is considered a Southern State. Future studies could replicate this study using all colleges and universities in either the Southern, Northern, Eastern, or Western States of the United States. A more comprehensive study would include all colleges and universities in the United States. Another worthy study would involve increasing the number and type of variables in the dataset. Hence, the model could possibly be improved by adding other variables to the dataset. Additional variables could lead to a more precise regression model. The most robust model might result from engaging in a time series analysis of Retention over multiple years at all colleges and universities in the United States.