Northeastern University
College of Professional Studies

ALY 6015 - INTERMEDIATE ANALYTICS

Prof. Roy Wada

ANALYTICAL REPORT ON COLLEGE SCORECARD DATA

08/20/2022

Pooja Kairamkonda

Priyanga Sreeram

Sakthi Pradeepa Muthukumar

INTRODUCTION

College Scorecard data helps prospective students to make informed decisions about their college education plans by providing essential information about the different school options available to them.

The College Scorecard dataset is adapted from the United States Department of Education which publishes aggregate data for each institution in the country. This dataset contains information about all the higher education institutions in the US between the years 1997 and 2015. The dataset contains information such as the institution name, location, classification (private or public), rate of admission, standard test scores, enrolment rate, financial aid and debt, graduate earnings post-completion, diversity rates, and cost of attendance. With this rich and varied information, useful insights can be gained on different aspects of the universities in the US and helps students to compare and choose schools that align with their educational goals and needs.

In this proposal, our group aims to identify the answer to the certain key questions from the College Scorecard dataset using various statistical methods and techniques using R programming.

We have chosen the questions based on some of the main factors that a student may contemplate while choosing an institution to frame these questions. In addition to this, as part of the analysis, we will also be exploring the factors that influence the debt a student may accumulate by the end of their education. As there is an increasing student debt crisis in the US during recent years, it becomes significant to understand the contributing elements to student debts.

DESCRIPTIVE STATISTICS

The dataset 'CollegeScorecard.csv' was imported into R studio to perform the descriptive statistics. The dataset contained 132402 observations and 66 features. The data cleaning was first performed before initiating the data analysis.

The following table Table 1 shows the calculated values for the mean and standard deviation of the entire dataset.

Table 1. Descriptive statistics for College Scorecard Dataset

Variables	mean	sd	median
name*	6158.23	3495.6	6183
city*	1919.87	1133.02	1948
state*	29.72	16.62	29
predominantdegree*	2.39	0.98	2

highestdegreegranted*	3.24	1.58	3
controlofinstitution*	2.79	0.92	3
carnegieclassification*	2.23	5.73	1
admissionrate	0.7	0.22	0.72
twoplusracesundergraduatepct	0.01	0.03	0
instatetuitionfees	10842.34	9064.6	8760
outstatetuitionfees	12902.25	8360.52	11130
pctpellgrant	0.5	0.24	0.48
pctfedstudentloan	0.53	0.3	0.59
mediandebt*	6925.16	5457.24	7118
agebegin*	41432.2	26474.05	43047.5
femalepct*	33746.14	20663.4	37361
meanearnings6yrs*	64.87	120.5	1
medianearnings6yrs*	57.4	109.17	1
stddvearnings6yrs*	35.88	71.92	1
meanearnings8yrs*	73.54	148.98	1
medianearnings8yrs*	62.51	129	1
stddvearnings8yrs*	44.17	93.35	1
year	2006.32	5.53	2007
undergraduateenrollment	2179.91	4773.4	493
whiteundergraduatepct	0.61	0.28	0.67
blackundergraduatepct	0.19	0.22	0.1
hispanicundergraduatepct	0.14	0.22	0.05
asianpacificislander	0.05	0.1	0.02
nativeamalundergraduatepct	0.02	0.09	0.01
nonresidentalienundergraduatepct	0.03	0.07	0.01
unknownraceundergraduatepct	0.08	0.14	0.04
parttimepct	0.32	0.23	0.28
averagenetprice	15818.64	8229.66	15369
titleivstudents	281.04	597.6	114
costofattendance	22010.5	10670.48	20154
completionrate	0.53	0.25	0.54
latitude	39.11	12.76	39.22
longitude	-93.53	34.55	-86.69

A subset of the dataset was then filtered out and represented in the following table Table 2, which displays the mean and standard deviation values for the different ethnic groups of undergraduate students enrolled in all the universities.

Table 2. Descriptive Statistics for ethnic groups in the dataset

Variables	**n**	**mean**	**sd**	**median**
whiteundergraduatepct	117567	0.61	0.28	0.67
blackundergraduatepct	110486	0.19	0.22	0.10
hispanicundergraduatepct	106917	0.14	0.22	0.05
asianpacificislander	89913	0.05	0.10	0.02
nativeamalundergraduatepct	76916	0.02	0.09	0.01

nonresidentalienundergraduatepct	55931	0.03	0.07	0.01
unknownraceundergraduatepct	72913	0.08	0.14	0.04

Missing Values

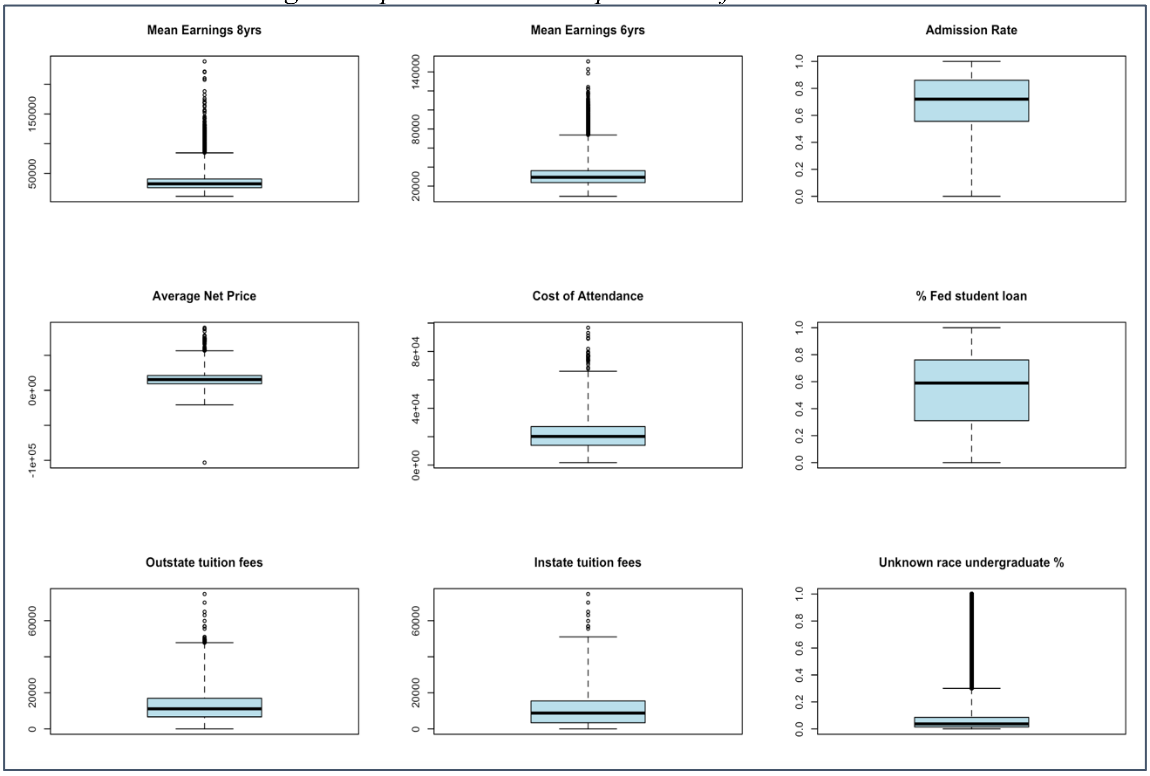
The dataset consisted of a large number of 'NA' or missing values, and it was necessary to eliminate such variables from the dataset for the sake of accuracy of results.

First, we checked for duplicate values and discovered that there were none. We then calculated the percentage of missing values for each variable. It was observed many variables had more than 50% of the data missing. We decided to remove only those variables with more than 80% of the data missing because removing 50% of the missing data variables would result in the removal of most of the important variables. As a result, we removed 28 variables that had more than 80% of the data missing and the remaining data was considered for the analysis

Detecting Outliers

We have defined outliers as observations that are detected 3 standard deviations away from the mean. Box plots are an efficient way to detect outliers. For our analysis, we have chosen those variables which have beyond 30% missing values to detect the outliers.

Fig 1. Boxplots to detect the presence of outliers



We can see from Fig 1 that only the variables “Admission rate” and “% Federal student Loan” have no outliers. All the other variables have outliers beyond 1.5 times the interquartile range along with the upper quartile Q3. It can also be seen that the variables “Average Net Price”, “Instate tuition fees”, “Mean Earnings 6yrs”, “Cost of attendance”, “Admission Rate”, “Mean Earnings 8yrs”, “Outstate tuition fees” are almost symmetrical while the variable “% Federal Student loan” is negatively skewed and the variable “Unknown race undergraduate %” is positively skewed. The missing data of the skewed variables are imputed with median data and the missing data of symmetrical variables are imputed with mean values.

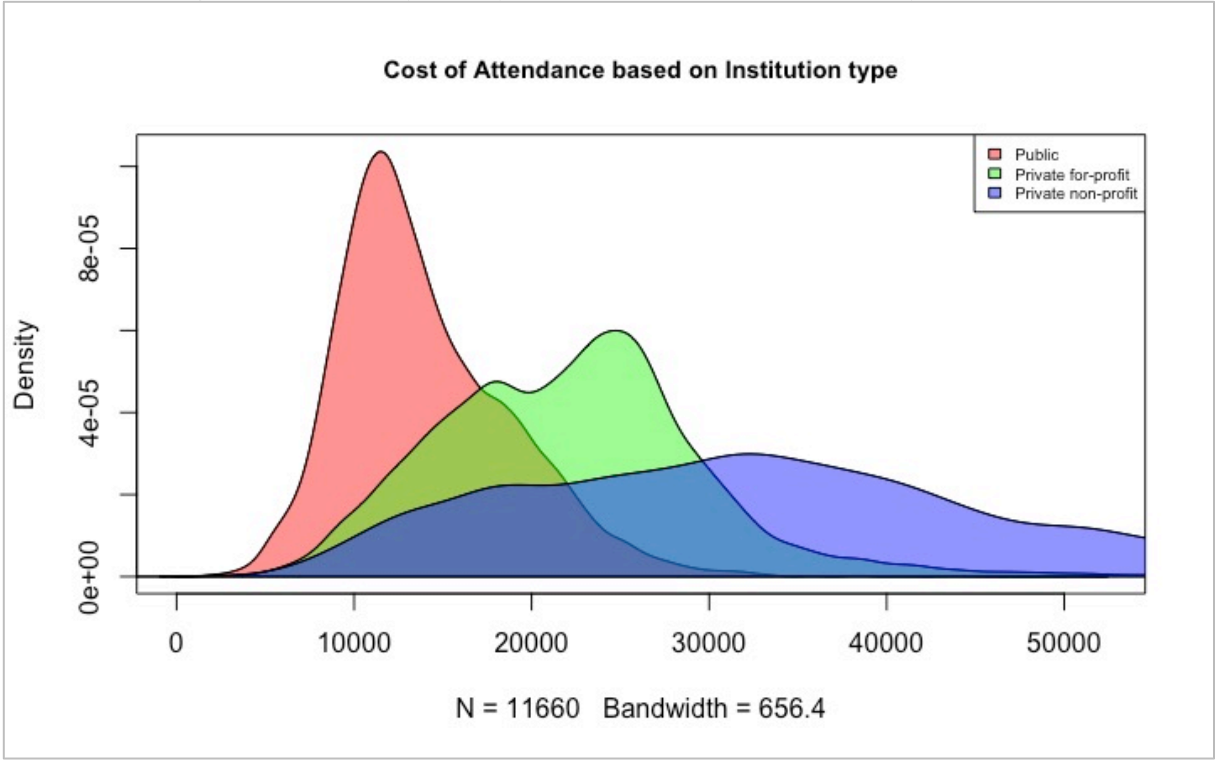
SUB-GROUP ANALYSIS

Sub-group analysis is typically performed to repeat the analysis on a specific sub-group within the dataset. Based on a shared characteristic, one or more subgroups can be identified from the dataset and comparisons can be done to ascertain and illustrate how the student debt analysis plays a role within the sub-groups.

1. Cost of Attendance based on ownership of institution

The following plot in Fig 2. represents how the Cost of Attendance is distribution based on the ownership type of the institution. It can be observed that the cost of attendance is on a higher scale for publicly owned institutions when compared to private institutions. For private non-profit institutions, the cost of attendance is relatively lower.

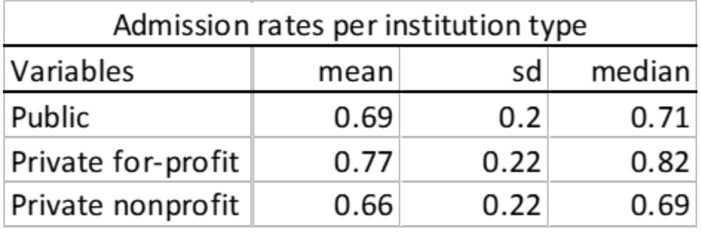
Fig 2. Distribution of Cost of Attendance based on ownership of institution



Another subset of the data was taken to explore the acceptance rate based on the ownership of the institution which is described in the below Table 3.

The average admission rate for private profit-making institutions is 77% which is slightly higher than the public and private non-profit-making institutions. Even though the public institutions offer less tuition fees, their admission rate is like that of private non-profit institutions having 69% & 66% rates respectively.

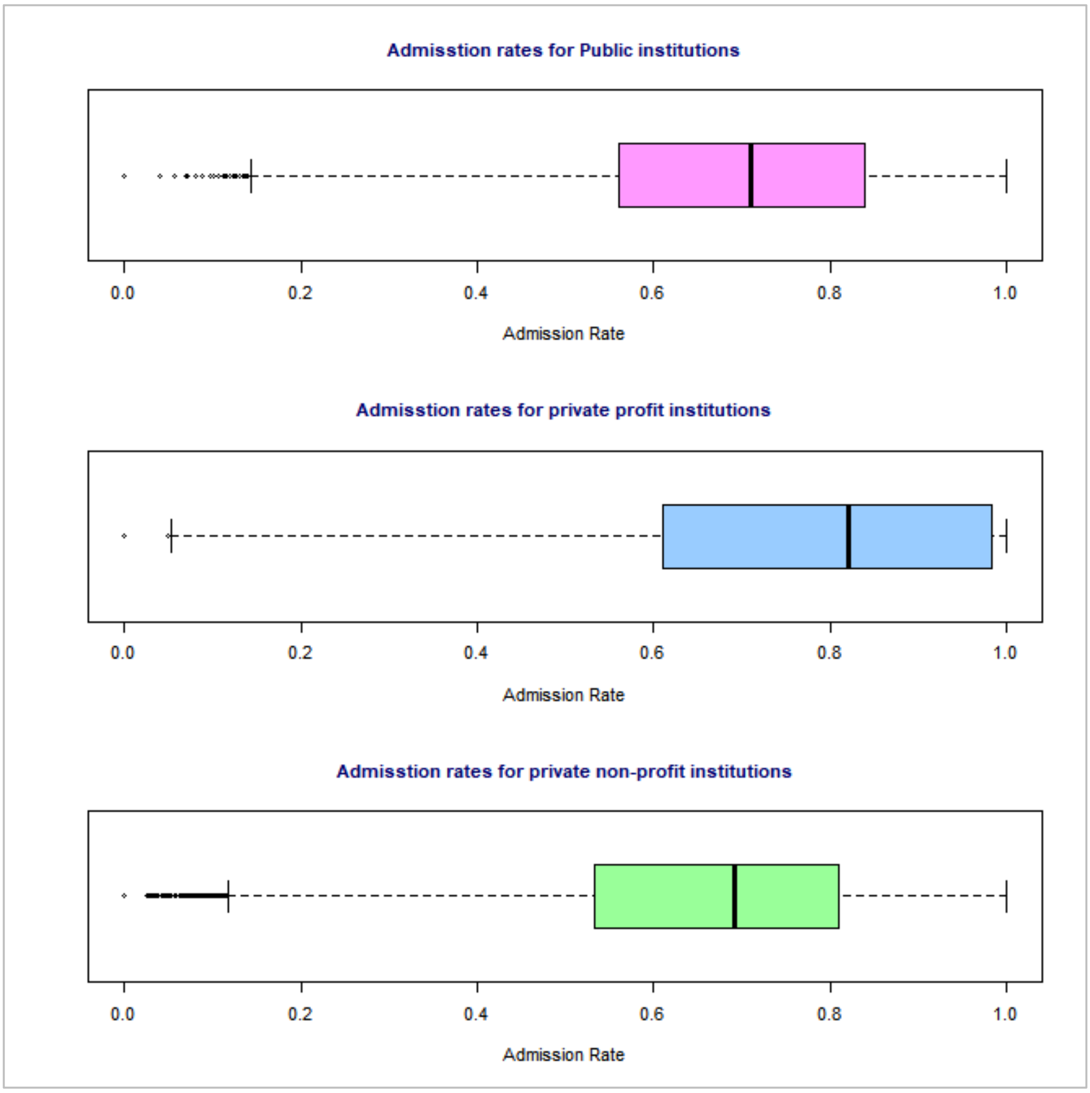
Table 3. Descriptive Statistics for admission rates per institution type



Admission rates per institution type

Variables	mean	sd	median
Public	0.69	0.2	0.71
Private for-profit	0.77	0.22	0.82
Private nonprofit	0.66	0.22	0.69

Fig 3. Admission rates per institution type



From the above boxplots in Fig 3, a few outliers are detected which can be removed from the dataset if the data is not significant. The median admission rate for private for-profit institutions is on a higher level when compared to the median admission rate for the public and private non-profit institutions which have a similar acceptance rate value.

2. Admission rates by institution type in a decade

The entire dataset is divided into the subgroup of 'control of institution' as – Public, Private for profit and Private for non-profit. The admission rates for these 3 categories over the decade are observed and the figures indicate how they differ throughout the years.

As compared to the year 2002, the average admission rate decreased from 72% to 66% for public universities. The universities with Private for-profit category had the highest average admission rate of 84% among the 3 categories, which eventually decreased to 76% in 2012. The average admission rate for the Private non-profit category was the lowest throughout the decade.

The table below indicates the admission rates over the decade starting in from 2002 to 2012 per institution type in the U.S.

Table 4. Descriptive Statistics for admission rates over the decade

	Year-2002			**Year-2012**		
	mean	**sd**	**median**	**mean**	**sd**	**median**
Public	0.72	0.19	0.75	0.66	0.2	0.68
Private for Profit	0.84	0.19	0.91	0.76	0.19	0.79
Private non-profit	0.7	0.2	0.75	0.64	0.21	0.65

Fig 4. Admission rate of students in the year 2002

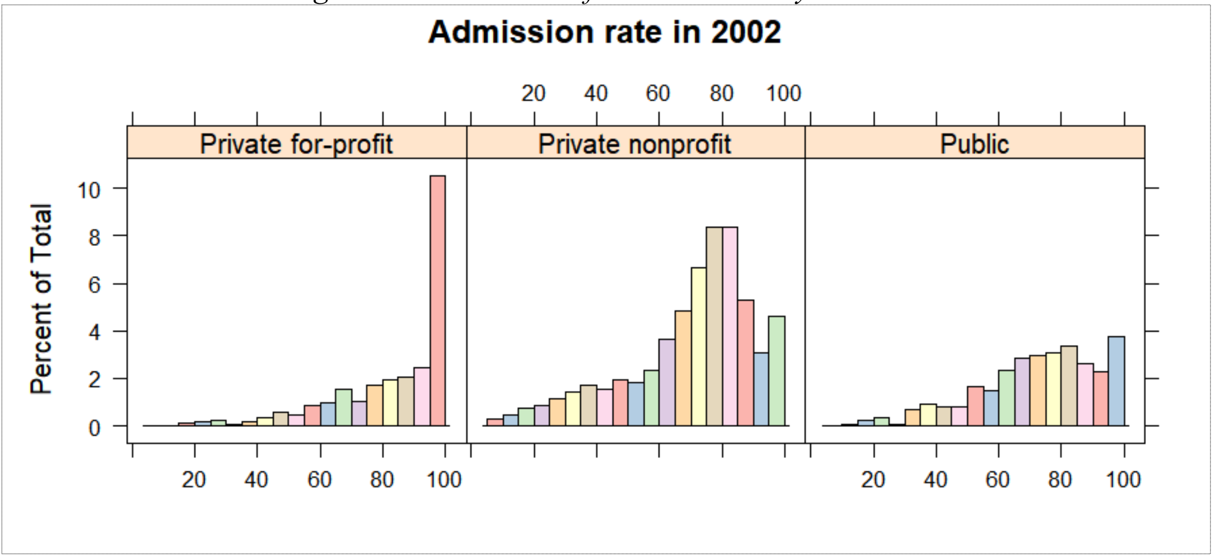
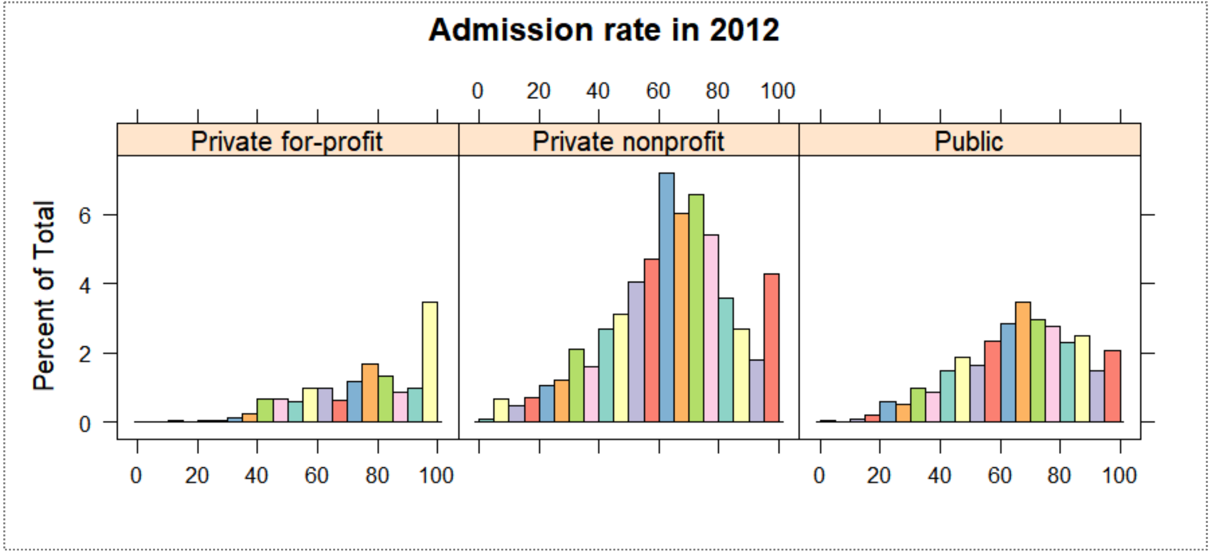


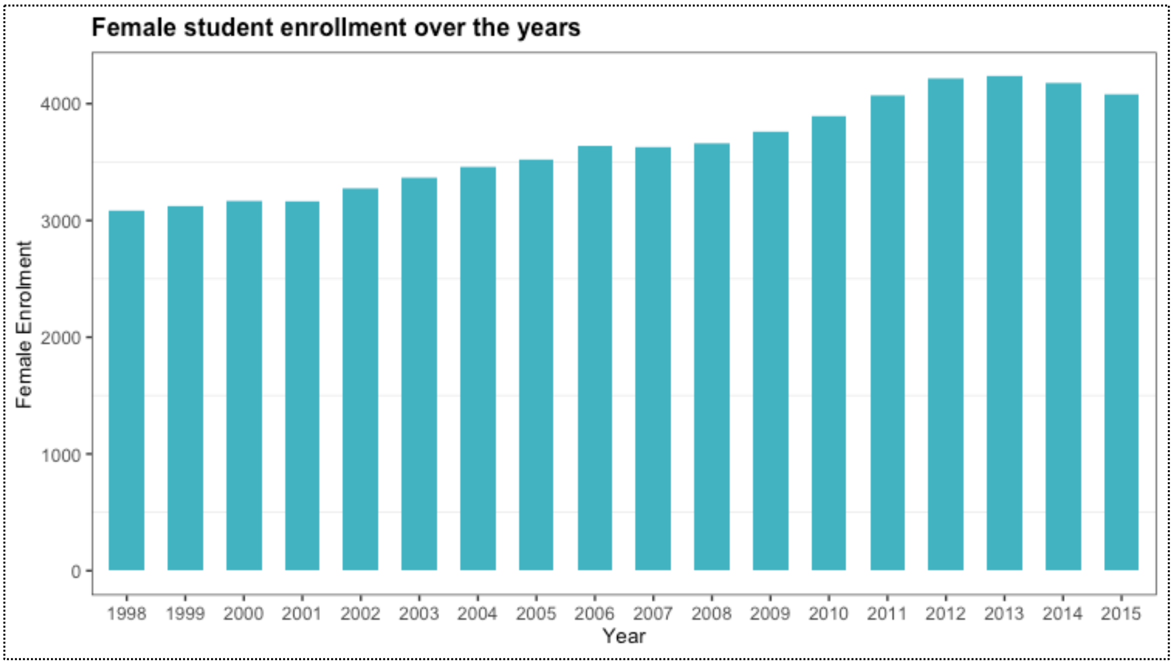
Fig 5. Admission rate of students in the year 2012



3. Female student enrollment by year

The following figure represents the female student enrolment rate between the years 1998 to 2015. Although there isn’t any significant increase seen during the 1990s, after the year 2000, the rate can be steadily seen to increase on a yearly basis.

Fig 6. Female enrolment rate throughout the years

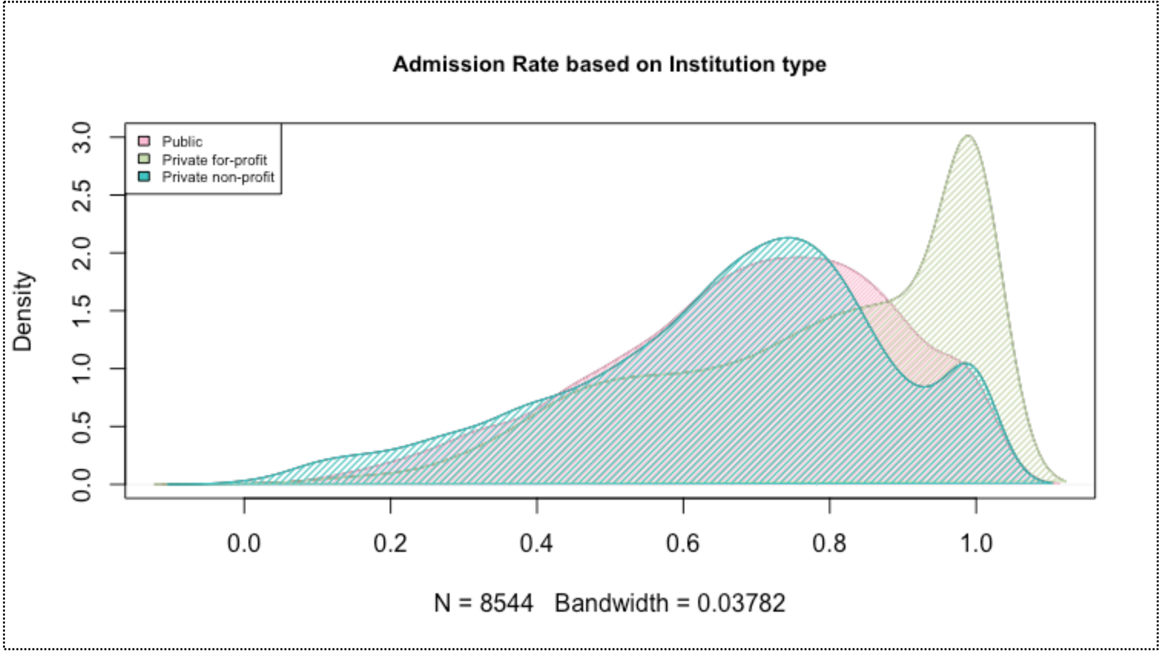


ANALYSIS

Question 1: Does the acceptance rate vary based on the ownership of the institution?

The following figure shows the distribution of admission rates data based on the ownership of the institution – public, private for profit and private non-profit. Based on the observation, institutions that are privately owned and for profit seem to have a higher rate of admission with publicly owned colleges having considerably lower admission rates.

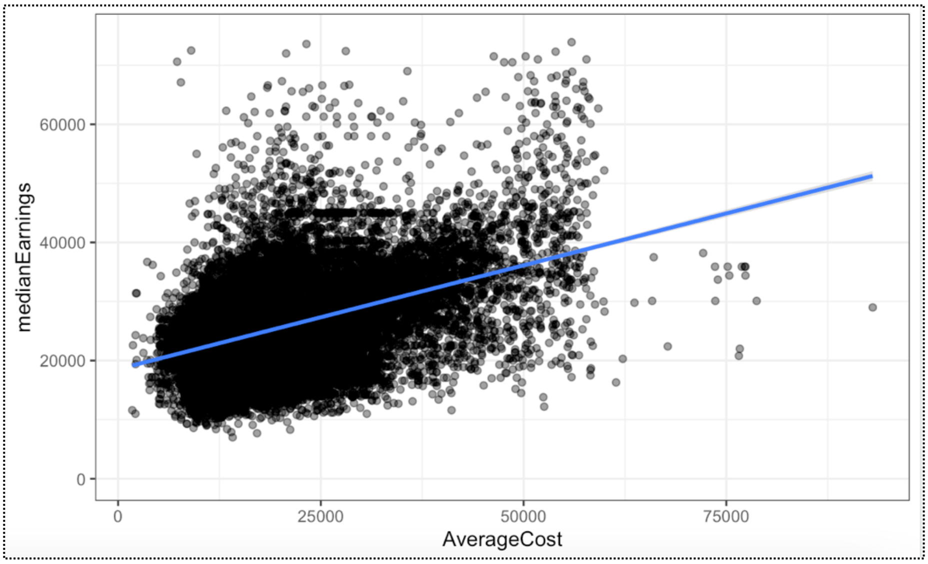
Fig 7. Admission rates per institution type



Question 2: Do high-cost institutions yield higher wages for students?

To understand if students who study in a highly priced institution have a better chance of earning higher wages, we consider two variables namely cost of attendance (Average Cost) and the median earning for analysis.

Fig 8. Relationship between average cost and median earnings

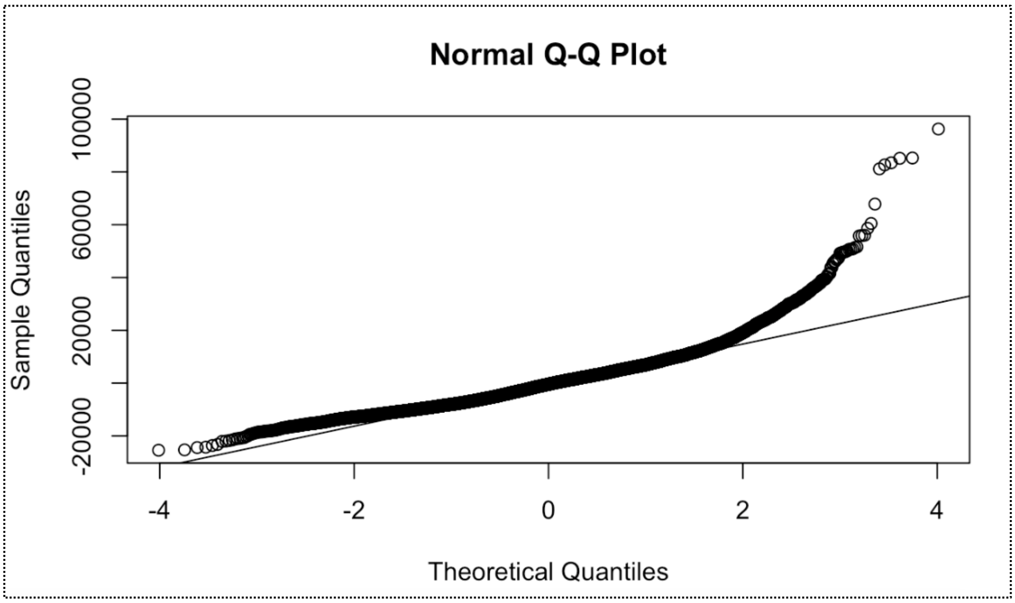


From the above figure, we can observe that there is a positive linear relationship between the cost of the institution and the earnings of the students. As the cost of the institution increases, the earnings of the student also increase though most of the students choose average-cost institutions and get an average salary. We can also see that there are many outliers and hence we can quantify this relationship by building a linear regression model.

FutureEarnings = lm(medianEarnings ~ AverageCost, data=medianearnings)

Once the model is fit, we test the linearity assumption of the residuals to see if the residuals are normally distributed or if the presence of outliers varies the data residuals vastly. Hence, we do a Q-Q plot to determine the normality assumption of the residuals.

Fig 9. Normal Q-Q plot of residuals



We can observe that the Q-Q plot is not linear indicating that the residuals are not linearly distributed. There is a strong right skew and hence we can perform a log transformation which will solve the assumption of linearity.

So, the model is fit once again after taking log transformation and we now observe the distribution of residuals.

The below histogram shows that the distribution of residuals is normal thus satisfying the assumption of linearity.

Fig 10. Distribution of residuals after logarithmic function of median earnings

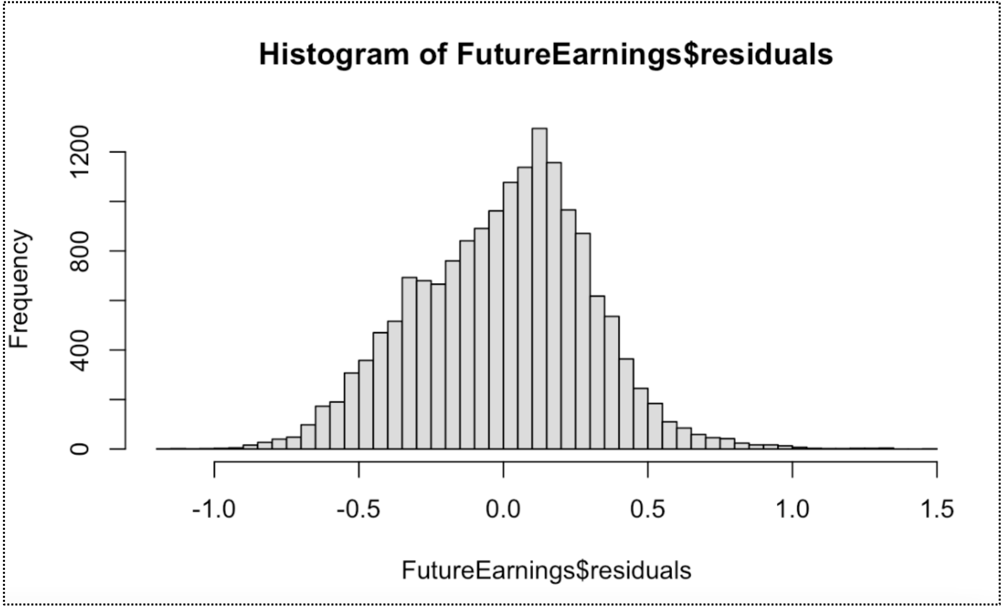
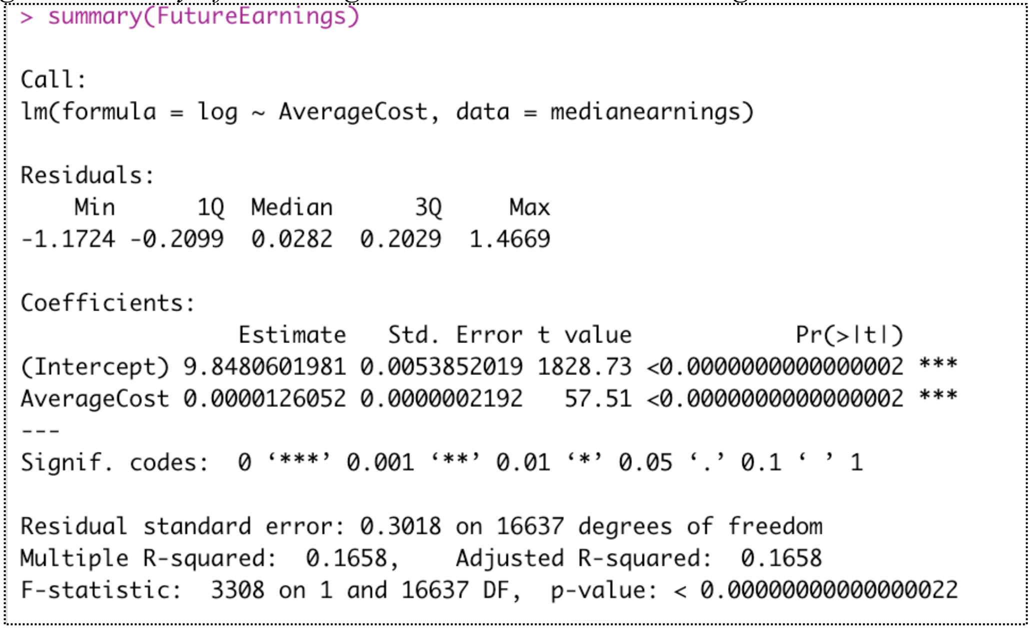


Fig 11. Summary of linear regression model between average cost and median earnings



```
> summary(FutureEarnings)

Call:
lm(formula = log ~ AverageCost, data = medianearnings)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.1724 -0.2099  0.0282  0.2029  1.4669 

Coefficients:
                Estimate   Std. Error t value            Pr(>|t|)    
(Intercept) 9.8480601981 0.0053852019 1828.73 <0.0000000000000002 ***
AverageCost 0.0000126052 0.0000002192   57.51 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3018 on 16637 degrees of freedom
Multiple R-squared:  0.1658,	Adjusted R-squared:  0.1658 
F-statistic:  3308 on 1 and 16637 DF,  p-value: < 0.00000000000000022
```

Hence from the above model, we observe that there is a statistically significant relationship between the log(medianEarnings) and AverageCost, which represents that high priced institutions yield students with much higher salaries. The R-squared value is 0.16 which indicates that there is 16% of the variance observed in the dependent variable (log(MediaEarnings)) is caused by the predictor variable (AverageCost).

Question 3: Do institutions with high granting degrees yield students with higher wages?

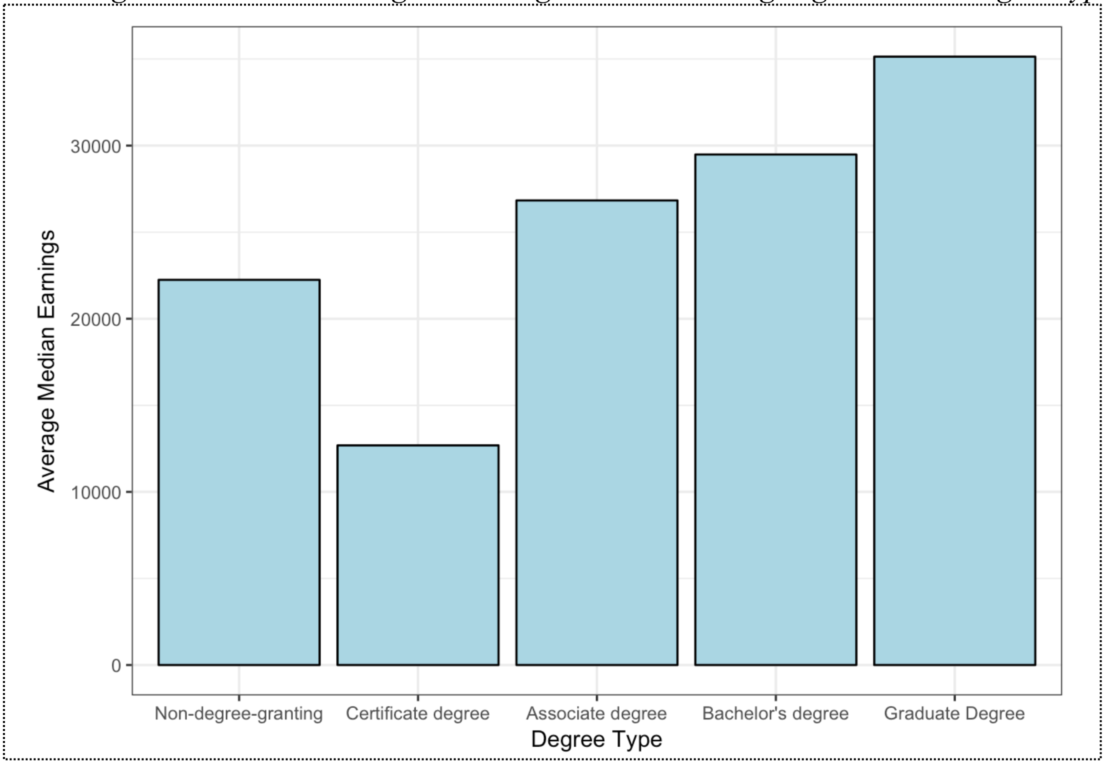
Next, to study if the institutions that provide high or advanced degree programs may yield students who may earn higher salaries is taken for analysis. This can be observed if it is true by analyzing the trend between the highest degree that are provided by various institutions and their median earnings. For this analysis, we have considered the variable "highestdegreegranted".

The median salary of all the institutions is grouped by the highest degree granted by the institutions. The variable "highestdegreegranted" is classified as below:

1. Non-degree granting
2. Certificate Degree
3. Associate Degree
4. Bachelor's degree
5. Graduate degree

We now plot the average of the median earnings against the highest degree granted by any institution.

Fig 12. Bar Plot showing the average median earnings against each degree type

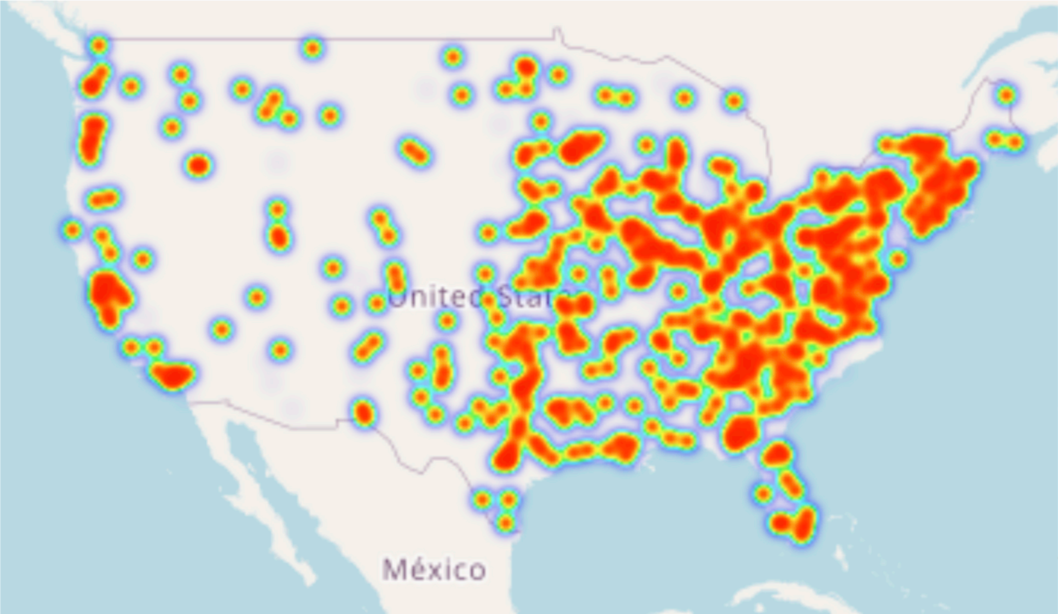


From the above graph, we can observe that there is a strong relationship between the highest degree granted and the median earnings. We can see that the students graduating from an institution that offers highest degree as a graduate degree yield students with more than three times wages when compared to the students coming from institutions that offer certificate degree as the highest degree.

Question 4: Does the location of the institution play a role in yielding high income for students?

The best approach to visualize clustered data is with a heat map. Data values are presented to us in the form of colored scales provided with hierarchical clustering. The below map represents those regions in the U.S. where the students' earnings is considerably high. From the map, it is clearly visible that the north & south eastern part of the country including popular states like Massachusetts, New York, North Carolina, South Carolina, Pennsylvania, West Virginia. The data was clustered using the variable 'medianearnings8yrs'.

Fig 13. U.S. Regions where the Students' earnings are high



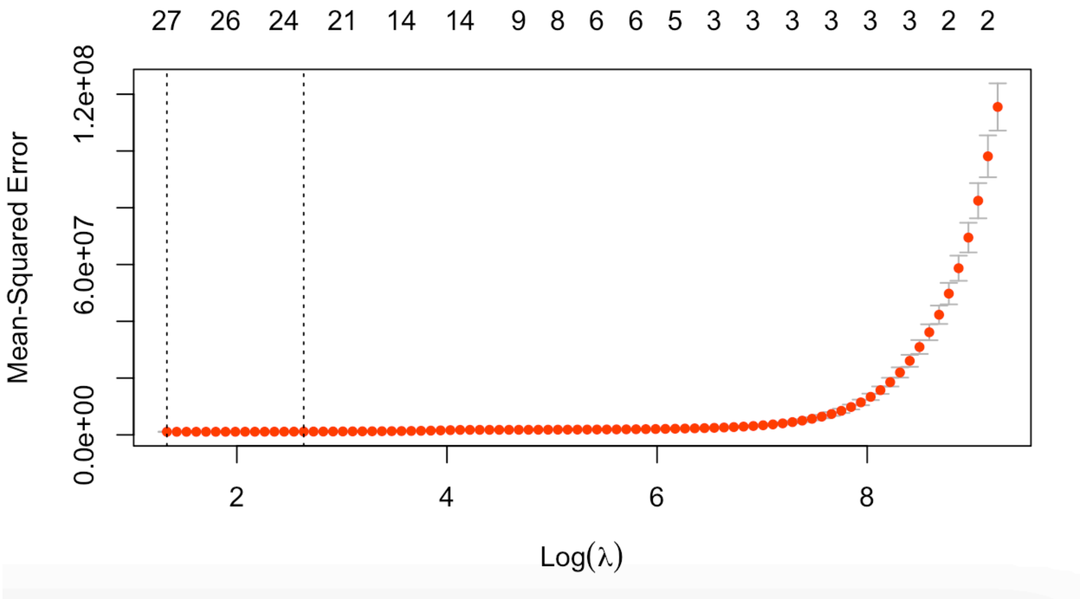
ANALYTICS METHODS AND TECHNIQUES

According to US News & World Report, the median weekly earnings in 2020 for a bachelor's degree was $1,305, for a master's degree was and for a doctoral degree was $1,885 as per the U.S. Bureau of Labor Statistics data. A prediction model was built to explore how the different variables influenced the median earnings of a student studying any particular degree in a US institution.

<u>LASSO Regression:</u>

By using the LASSO regression technique we have omitted the variables that have the correlation coefficients with value approximately equal to zero. The below code snippet displays the variables that are eliminated by using this method. First, we calculated the lambda min and lambda 1 standard error values.

Fig 13. Lasso model feature selection showing Lambda min and lambda1se values



From the above figure 13, we can see that the lambda min is 24 features and the lambda 1se has 27 features. We build both the models of lasso with both lambda min and lambda 1se values. On comparing we found that lambda 1se model has the better rmse output. Hence, we used the same for predicting both the train and test data.

From model lambda 1se value output we can identify that the variables twoplusracesundergraduetepct, mediandebt, undergratuateenrollment, nonresidentialundergratuatepct, unknownraceundergraduatepct are reduced to zero and only the remaining variables are used for prediction in the lambda 1se model.

Table 4: LASSO Model results

	RMSE
Train	1040.02
Test	1023.28

From the above table, we can see that the root mean square error value for train data is 1040.02 and that of test data is 1023.28. By dividing the values, we get the rmse output value as 1.01 which is very high and greater than 1. Hence this is not a good model for our prediction.

Logistic Regression model

A prediction model was built to explore how the different variables influenced the median earnings of a student studying for any particular degree in a US institution.

The logistic Regression technique was adopted for the analysis with the mean earnings 8 years after graduation (meanearning8yrs) as the outcome variable. Logistic regression is a machine learning algorithm that is used for the prediction of a dependent variable based on previous observations and this technique is used only when the response variable has a binary output. Since our response variable meanearning8yrs has multiple outcomes, we decided to convert it to a binary output. We found the mean of the meanearning8yrs to be 43200 and the median of the meanearning8yrs to be 41500. Since both the mean and median are almost equal, the data is found to be distributed approximately symmetrical. Hence, we decided to have any earnings less than $43000 be "low" earnings and above $43000 be "high" earnings thus converting our response variable as a binary output variable consisting of two factors "low" and "high".

Next, we created the data partitioning into train and test datasets with 70:30 ratio and used set.seed() function for generating a random sequence from the dataset.

The model is built using a **glm()** function. For the first model, we have used the variable "meanearning8yrs" as the response variable, and all the other 31 numeric variables in the dataset as the predictor variables. The AIC value of this model is calculated and the value is 403.385.

Next, we used stepAIC method for feature selection. Upon using this selection method, the number of features has been reduced to 17 and then the model is created using these 17 features as the predictor variables. The AIC of model 2 is calculated.

. *Table 5: Comparing the AIC values for Model 1 and Model 2*

Value	Model 1	Model 2
AIC	403.39	383.49

From the above table, we can find that the AIC of model 1 is 403.39 and the AIC of model 2 is 383.49. The lower the AIC, the better is the model. Hence, we choose model 2 to be the better model and we use the same for the prediction of mean earning at 8yrs. We use the test dataset for predicting the model. After prediction, we created a confusion matrix to know the accuracy of the model.

Table 4. Confusion Matrix

	High	**Low**
High	456	15
Low	17	644

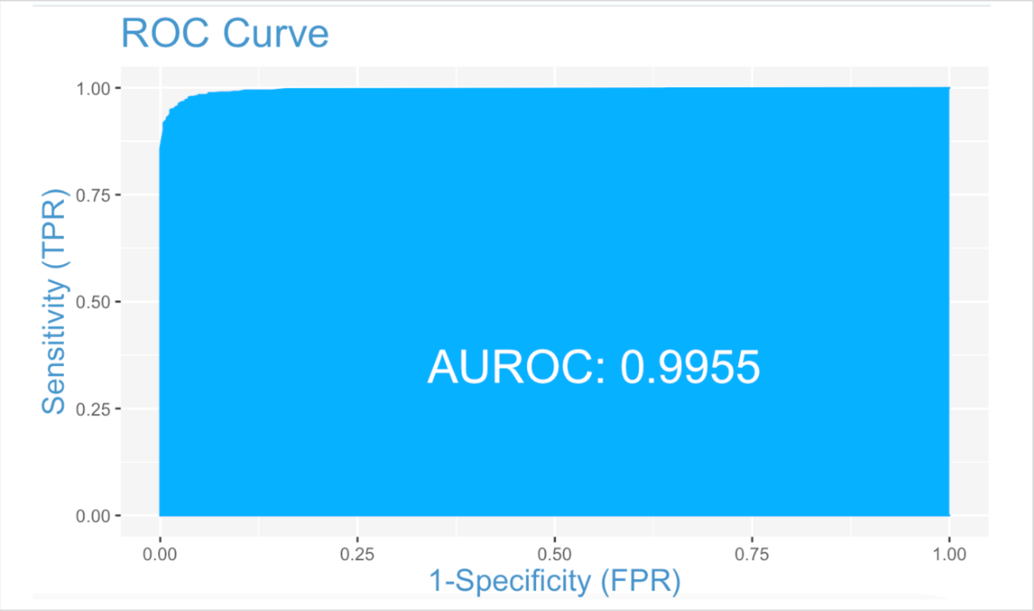
From table 4, we can see that the true positives predicting high earnings as high is 456 and the false positives is 17 where the high earnings are predicted to be low earnings. The true negative is predicting low as low earnings, and it is 644 and the false negatives are 15 where the low earnings are predicted as high earnings. The misclassification of low being predicted as high which is our type II error is more dangerous because the student might expect a high earning based on the prediction whereas in reality the student will end up in a low earning.

Table 5. Confusion Matrix Statistics

Metrics	**% Value**
Accuracy	97.17
Precision	96.4
Recall	96.81
Specificity	97.42

From table 5, we find that the accuracy of the confusion matrix is 97.17% which indicates that our model is good for prediction. The precision and recall metrics are the percentages of positives predicted as positives and both the values are 96.4% and 96.81% respectively. The specificity factor predicting the negatives as negatives is 97.42%. Overall all the metrics are above 95% indicating that our model is very good for prediction.

Fig 14. ROC Curve for the test set



We can see that the ROC curve is almost perfect to an ideal roc curve and there is only a minute curve deviation on the top left side of the curve. And also we can see that the Area Under the Curve value is 0.99 which is very close to 1. Hence, this proves that ours is a good model for prediction.

CONCLUSION

Overall, the College Scorecard data tell us about the factors that students consider while planning to apply to universities in the U.S. We have learned the importance of exploratory data analysis. Through the EDA, we were able to fetch the main variables from the dataset, by using which we could answer the research questions through our analysis. The Logistic Regression and StepAIC Feature Selection technique led us to build a model with 97% accuracy. Hence, we can say that the mean earning of 97% of future graduates will be predicted accurately. By the time a student graduates, the majority of students are focused on finding a dream career with a respectable salary. As per their goals, they'll prefer those universities which fall in big cities in the northeastern part of the United States where they can find ample career opportunities to get high earnings and also choose an institution that offers higher degrees like graduate degrees than certificate courses which will yield better earnings. The higher the earnings, the students will be able to complete their debts at the earliest as clearing the debts is one of the major hurdles for most of the students with lower earnings.

REFERENCES

[1] U.S Department of Education College Scorecard. Retrieved from - https://collegescorecard.ed.gov/data/

[2] Bluman G. Allan (2018). Elementary Statistics: A Step by Step Approach. Eighth Edition. New York, NY: McGraw-Hill Education.

[3] Yaohui Zeng, Patrick Breheny. (2018) The biglasso Package: A Memory- and Computation-Efficient Solver for Lasso Model Fitting with Big Data in R. Journal of Statistical Software. Retrieved from - https://arxiv.org/pdf/1701.05936.pdf

[4] Alex (March 2019). The QQ Plot in Linear Regression. BOOSTEDML. Retrived from - https://boostedml.com/

[5] Bassalat Sajjad (June 2017). LASSO regression in R exercises. Retrieved from - https://www.r-bloggers.com/2017/06/lasso-regression-in-r-exercises/

[6] Selva Prabhakaran. Logistic Regression. r-statistics.co. Retrieved from – http://r-statistics.co/Logistic-Regression-With-R.html

APPENDIX

We have enclosed the R script file named as “M6_Group2_FinalProject_RCodes.R” that we have used for all the above analyses.