Assignment 3: EDA and Logistic Regression for College dataset

Yash S

2024 – 2 – 2

Northeastern University: College of Professional Studies

Introduction

Introduction

The objective of this report is to explore and analyze the College dataset from the ISLR package and build a logistic regression model to predict whether a university is public or private. This analysis follows a structed and methodical approach, such as descriptive statistics, correlation analysis, feature selection, logistic regression model development and performance evaluation. The goal of this report is to provide with key insights and visualizations, with meaningful interpretations that align with the best practices of statistics.

Dataset Overview

The dataset comprises of 777 university with 18 variables which captures the various institutional characteristics of these universities. Our target variable here is the "Private" which contains information of whether the institution is a public or a private campus, this column has been converted to a binary format depicting 1 for private and 0 for public, here using one-hot encoding could make the dataset more complex so a manual conversion was done to achieve the binary format. We found that the data did not have missing values indicating that it is a complete dataset ready for analysis. The independent variables include numerical attributes that are related to enrolled students, qualifications of faculty, tuition fees and expenditure per student.

Key Variables:

1. Nominal (Categorical variable):-

a) **Private:** (Converted Binary, 1 = Private, 0 = Public) This variable informs us whether the institute is public or private.

2. Ordinal (Categorical variable):-

a) **PhD:** Percentage of Faculty with PhD in that particular institute.
b) **Terminal:** Percentage of Faculty with Terminal Degree in that particular institute.
c) **perc.alumni:** Percentage of alumni who donate.

3. Discrete:-

a) **Apps:** The total number of applications received by that particular institute.
b) **Accept:** The total number of applications accepted by that particular institute.
c) **Enroll:** The total number of students enrolled in that institute.
d) **F.Undergrad:** The total number of full-time undergraduates in the particular institute.

4. Continuous:-

a) **Outstate:** Cost for out of the state tuition cost for that institute.
b) **Expend:** Amount of expenditure per student for instructional purposes by that institute.
c) **S.F.Ratio:** Student to faculty ratio of that institute.

Data Analysis

Explanatory Descriptive Analysis:

1.Descriptive Statistics of Key Variables:

```
 Private      Apps           Accept          Enroll       F.Undergrad     Outstate
 0:212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   :  139   Min.   : 2340
 1:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:  992   1st Qu.: 7320
         Median : 1558   Median : 1110   Median : 434   Median : 1707   Median : 9990
         Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   : 3700   Mean   :10441
         3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.: 4005   3rd Qu.:12925
         Max.   :48094   Max.   :26330   Max.   :6392   Max.   :31643   Max.   :21700
      PhD            Terminal       S.F.Ratio      perc.alumni        Expend
 Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
 1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
 Median : 75.00   Median : 82.0   Median :13.60   Median :21.00   Median : 8377
 Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
 3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
 Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
```

```
Standard Deviation for Outstate: 4023.016
Standard Deviation for Expend: 5221.768
```

a. Private [Nominal Binary(0,1)]

There are around **212 public institutes** and **565 private institutes**.

b. Apps [Discrete (Count)]

The average number of applications received by the institutes are **3,002** with a median of **1,558** with lowest number of applications received of **81** and highest number of applications received is **48,094.**

c. Accept [Discrete (Count)]

The average number of accepted applications by the institutes are **2,019** with a median of **1,110** with lowest number of accepted applications by an institute at **72** and highest number of accepted applications by an institute is **26,330.**

d. Enroll [Discrete (count)]

The average number of students enrolled at an institute is **780** with lowest number of students enrolled at an institute at **35** and highest number students enrolled at an institute at **6,392**.

e. F.Undergrad [Discrete (Count)]

The number of full-time undergraduates has a range from **139 to 31,643** students with a mean of **3,700** and median of **1707**.

f. Outstate [Continuous ($)]

The average of out of state tuition is **$10,441** with a standard deviation of **$4,032**, with the cheapest tuition at **$2,340** and the most expensive tuition at **$21,700**.

g. PhD [Ordinal (%)]

The average of faculty percentage with PhD in an institute is **73%**, with a maximum value of **103%** indicating potential data inconsistency.

h. Terminal [Ordinal (%)]

The average of faculty percentage with a Terminal degree in an institute is **80%**, with a maximum value of **100%** indicating consistent data as compared to PhD faculty.

i. S.F.Ratio [Continuous (Ratio)]

The student-faculty ratio has a mean of **14.09** with a median of **13.60**.

j. perc.alumni [Ordinal (%)]

The average percentage of alumni who donate to their particular institutes are **22.74%** with a minimum of **0%** and maximum of **64%**.

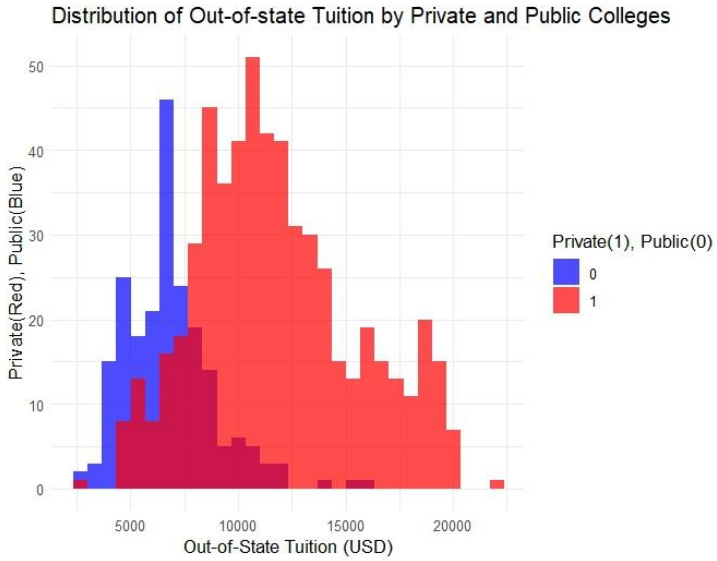
k. Expend [Continuous ($)]

Every institute have an average expenditure of **$9,660** and a standard deviation of **$5,222**, with a lowest spend of **$3,186** and highest spend of **$56,233**.

2. Visualizations:

a. Out of State Tuition Distribution:

This graph shows the distribution of the Out of State tuition cost for public (blue) and private (red) institutes. As we can see in the graph that private institutes usually have higher out of state tuition as compare to the public institutes.

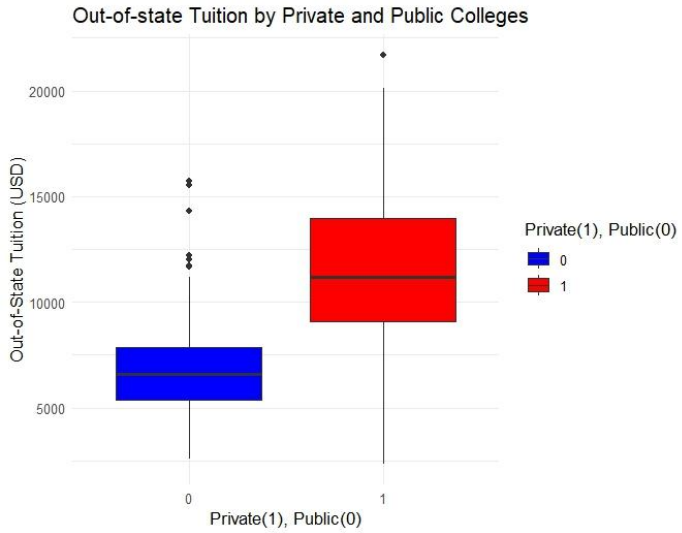


The chart shows comparative analysis of Out of State tuition cost: Public vs Private Colleges

b. Boxplots:

a. Out of State Tuition:

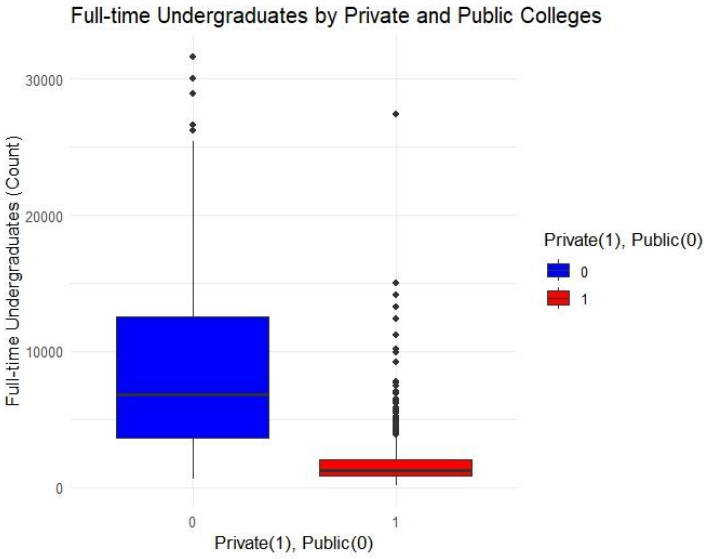
This graph shows that some of the public institues have potenial unusual observations, which are even above the average cost of private institutes. This warrants detection for these outliers during model development.



Boxplot of Out of State tuition cost: Public vs Private Colleges

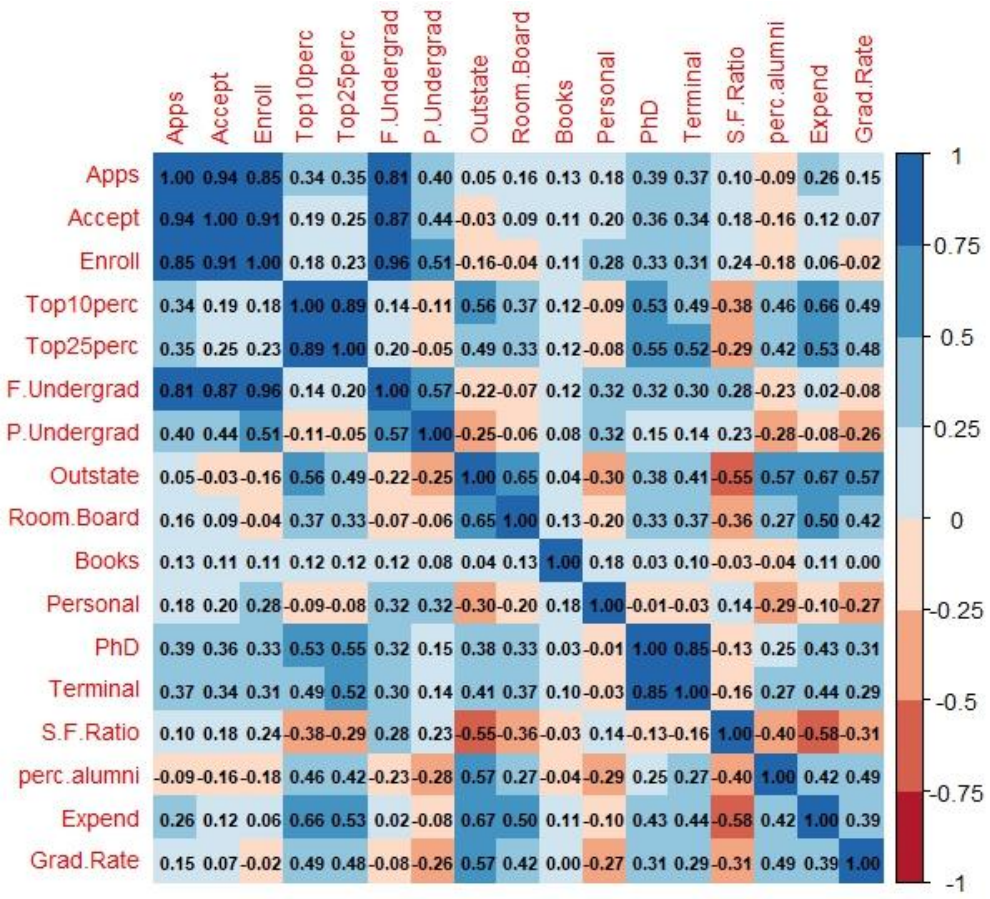
ii. Full-time Undergraduates:

This graph shows the number of full-time undergraduate students enrolled in public and private institute. The noticable trend here is that private colleges tend to have more full-time undergraduates making this an important variable for our prediction.



Full-time Undergraduates Enrolled: Private vs Public

Correlation Matrix:



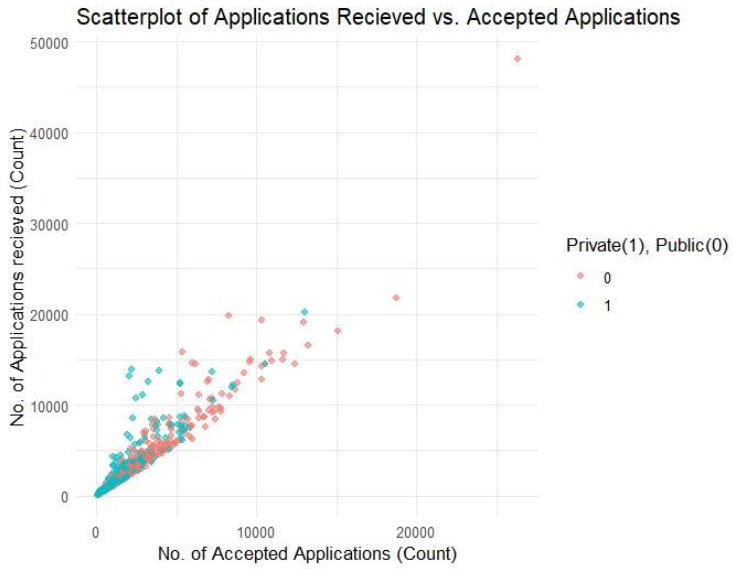
Correlation Matrix Plot

In this correlation matrix all the different independent variables are paired up in every different possible way and every combination is explored and each box is filled with colours. As we can see, the blue shades mean the variables move in the same direction, the darker the shade is the stronger the bond. Similarly, the red blocks have opposite relationship, meaning when one goes up the other goes down, the darker the shade is the weaker the bond. The perfect diagonal blocks always show perfect a positive correlation since it is compared with itself.

From the above correlation matrix following were the key inferences made:

i. Applications received (Apps) and Accepted Applications (Accept):

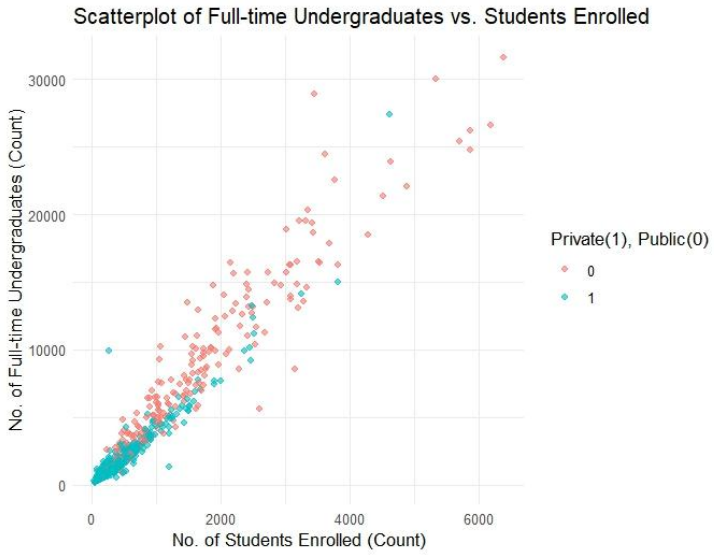
- High Correlation (0.94345057)
- Scatterplot showing a positive correlation between Applications received and Accepted applications, this simply means that acceptance rate improves when more applications are received.



The graph indicates a strong positive correlation between Applications Received and Accepted Applications.

ii. **Full-time Undergraduates (F.Undergrad) and Enrolled Students (Enroll):**

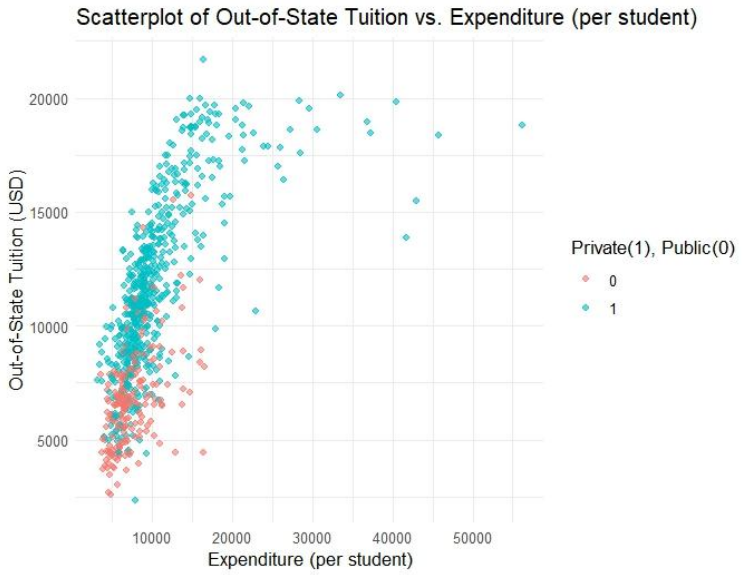
- High Correlation (0.96463965)
- Scatterplot showing a positive correlation between Number of Full-time Undergraduates and No of Students Enrolled, indicating that higher enrolment rate is directly associated with a large number of full-time undergraduates.



The graph indicates strong positive correlation between the number of Full-time Undergraduates and number of students enrolled.

iii. **Out of State Tuition (Outstate) and Expenditure (Expend):**

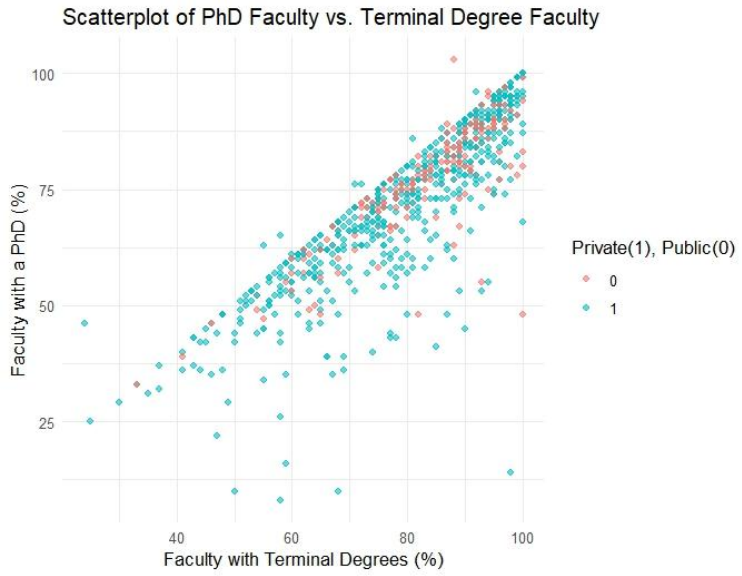
- High Correlation (0.67277862)
- Scatterplot showing a positive correlation between Out of State tuition and Expenditure per Students, suggesting that higher out of state tuition tend to have high expenditure per student.



The graph indicates strong positive correlation between out of state tuition and expenditure

iv. PhD holder faculty % (PhD) and Terminal degree faculty % (Terminal):

- High Correlation (0.84958703)
- Scatterplot showing a positive correlation between percentage of faculty with PhD and percentage of faculty with terminal degree, indicating that institutes with higher percentage of faculty with PhD also have higher percentage of faculty with terminal degree.



The graph indicates strong positive correlation between percentage of faculty with PhD and Terminal Degree.

Regression Analysis:

a) Data Partitioning

The dataset was split into 70% of training data and 30% of testing data, this was done using the “caret” package ensuring a balanced distribution of public and private colleges (Target Variable).

b) Step-wise Feature Selection

The method for feature selection was step-wise selection which uses AIC-based forward and backward selection to select the most influential predictors.

```
Call:
glm(formula = Private ~ Apps + Enroll + F.Undergrad + Outstate +
    PhD + perc.alumni + Expend + Grad.Rate, family = binomial,
    data = train_set)

Coefficients:
              Estimate Std. Error z value     Pr(>|z|)
(Intercept) -1.8963348  1.3926982  -1.362     0.173316
Apps        -0.0003739  0.0002027  -1.845     0.065102 .
Enroll       0.0019856  0.0010665   1.862     0.062636 .
F.Undergrad -0.0007323  0.0002338  -3.132     0.001738 **
Outstate     0.0007395  0.0001351   5.473 0.0000000444 ***
PhD         -0.0766800  0.0200659  -3.821     0.000133 ***
perc.alumni  0.0359692  0.0259347   1.387     0.165468
Expend       0.0002574  0.0001424   1.807     0.070735 .
Grad.Rate    0.0274096  0.0162281   1.689     0.091215 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 639.40  on 544  degrees of freedom
Residual deviance: 151.78  on 536  degrees of freedom
AIC: 169.78

Number of Fisher Scoring iterations: 8
```

The regression equation for the above model is as follows:

Private= −1.8963 −0.0003739×Apps +0.0019856×Enroll +0.0000234×F.Undergrad +0.0007395×Outstate −0.0766800×PhD +0.0359692×perc.alumni +0.0002574×Expend +0.0270462×Grad.Rate

We can interpret the logistic regression coefficients as follows:

i. Intercept (-1.8963):

This value represents the baseline log-odds of being a private institute when all the other predictors are zero. Although this is not meaningful for any interpretation, it is part of the regression equation.

ii. Apps (−0.0003739):

As the number of applications increases it slightly decreases the probability of that institute being a private one.

iii. Enroll (0.0019856):

A higher number of students enrolled increases the probability of it being a private institute.

iv. F.Undergrad (0.0000234):

An institute with a greater number of full-time undergraduate students has a slight probability of being private.

v. Outstate (0.0007395):

A higher out of state tuition fees highly increases the chance of it being a private institute, this has the strongest positive effect.

vi. PhD (-0.0766800):

A higher number of faculty with PhD decreases the likelihood of an institute being private.

vii. Perc.alumni (0.0359692):

A higher number of donations slightly increases the probability of it being a private institute. Private institutes often have a strong alumni network and a donor culture.

viii. Expend (0.0002574):

More expenditure on instructional purposes increases the probability of it being a private institute.

ix. Grad.Rate (0.0270462):

A higher graduation rate increases the probability of the institute being private.

The AIC value for the step-wise model is 169.78.

c) Logistic Regression Model Development

Performing a Logistic Regression in this case helps us to predict whether the institute is Public or Private (dependent variable) and we choose only 5 explanatory variables (Apps, F.Undergrad, Outstate, PhD, Expend) this helps uncover the factors that significantly influences the probability of an institute being either Public or Private. By using R, we constructed an initial model with five predictor variables with high correlation by referring to the correlation matrix and the VIF number of the step-wise model. The top five variables are:

1. Applications Received (Apps)
2. Full-time Graduates (F.Undergrad)
3. Out of State tuition (Outstate)
4. Percentage of faculty with PhD (PhD)
5. Expenditure per student (Expend)

```
Call:
glm(formula = Private ~ Apps + F.Undergrad + Outstate + PhD +
    Expend, family = binomial, data = train_set)

Coefficients:
             Estimate Std. Error z value        Pr(>|z|)
(Intercept) -0.0134915  1.1545492  -0.012        0.990677
Apps        -0.0001174  0.0001633  -0.719        0.472060
F.Undergrad -0.0005545  0.0001545  -3.589        0.000332 ***
Outstate     0.0008276  0.0001231   6.725 0.0000000000176 ***
PhD         -0.0750085  0.0197795  -3.792        0.000149 ***
Expend       0.0002051  0.0001341   1.530        0.125959
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 639.40  on 544  degrees of freedom
Residual deviance: 161.95  on 539  degrees of freedom
AIC: 173.95

Number of Fisher Scoring iterations: 8
```

The regression equation for the above model is as follows:
Private=−0.0134915−0.0001174×Apps−0.0005545×F.Undergrad+0.0008276×Outstate−0.0750085×PhD+0.0002051×Expend

The interpretation of the above model is as follows:

i. Intercept (−0.0134915):
   - This value represents the baseline log-odds of being a private institute when all the other predictors are zero. Although this is not meaningful for any interpretation, it is part of the regression equation.

ii. Applications Received (Apps) (−0.0001174):
   - As the number of applications increases it slightly decreases the probability of that institute being a private one.

iii. Full-time Undergraduate students (F.Undergrad) (−0.0005545):
   - An institute with a greater number of full-time undergraduate students has a slight probability of not being private.

iv. Out of State Tuition (Outstate) (0.0008276):
   - A higher out of state tuition fees highly increases the chance of it being a private institute, this has the strongest positive effect.

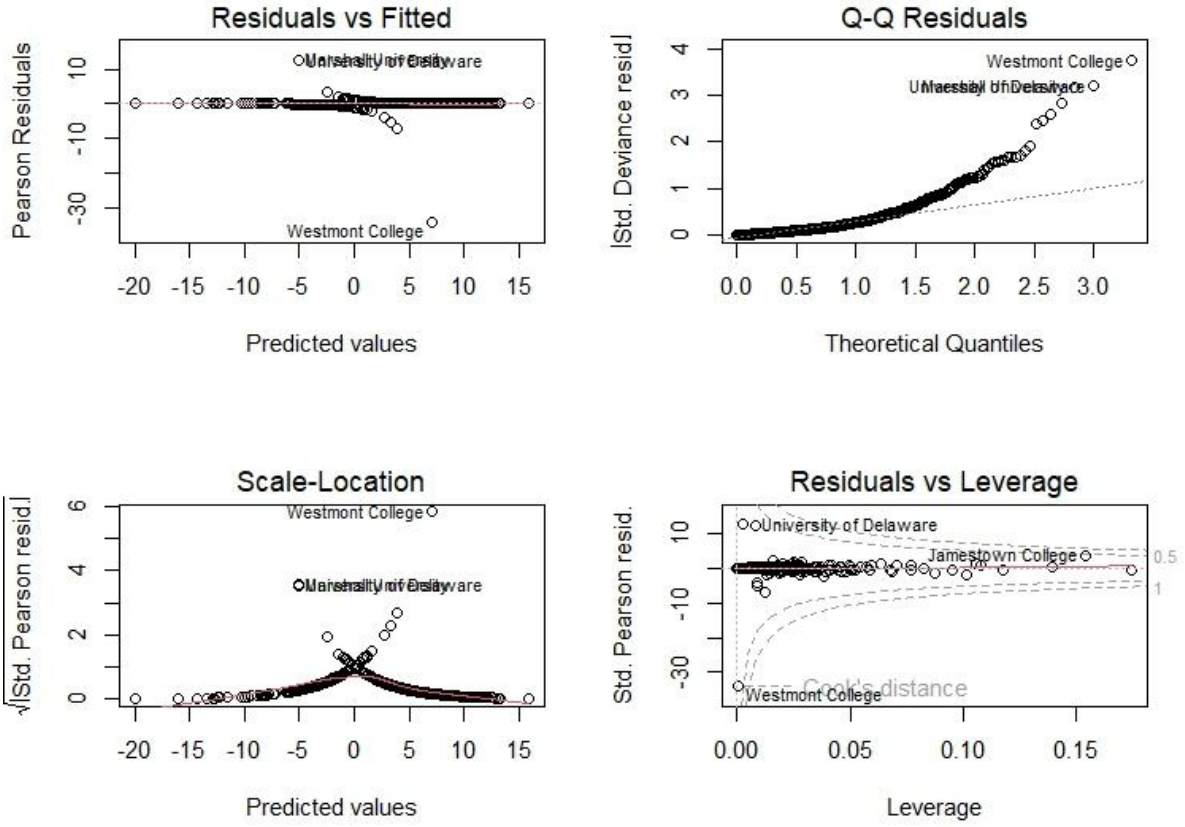
v. Percentage of faculty with PhD (PhD) (-0.0750085):
   - A higher number of faculty with PhD decreases the likelihood of an institute being private.

vi. Expenditure per student (Expend) (0.0002051):
   - More expenditure on instructional purposes increases the probability of it being a private institute.

AIC Value (173.95) is higher warranting further analysis of unusual observations.

d) Diagnostic Plots and Refinement

i. Residual vs Fitted: There is a visible trend of randomly scattered points and outliers suggesting potential issues with linearity and homoscedasticity.
ii. Q-Q Residuals: The Q-Q plot shows deviation from the line, indicating the residuals are not perfectly normally distributed.
iii. Scale Location: We can see the points are equally spread and not funnel shaped, but the pattern suggests outliers.
iv. Residuals vs Leverage: The points with high leverage are potential outliers as shown by Cook's Distance, these impact the regression model significantly.

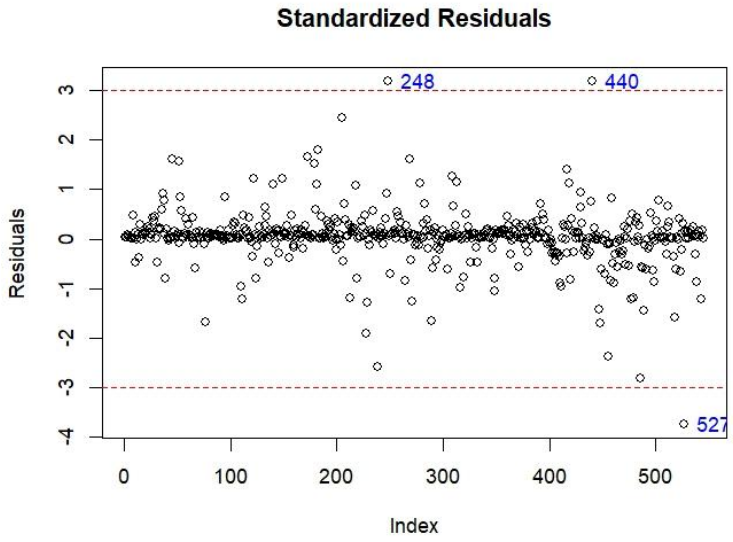


VIF for all predictors that was under 5, indicating there is some correlation between the predictors but they are not severe enough to cause any issues in the regression model.

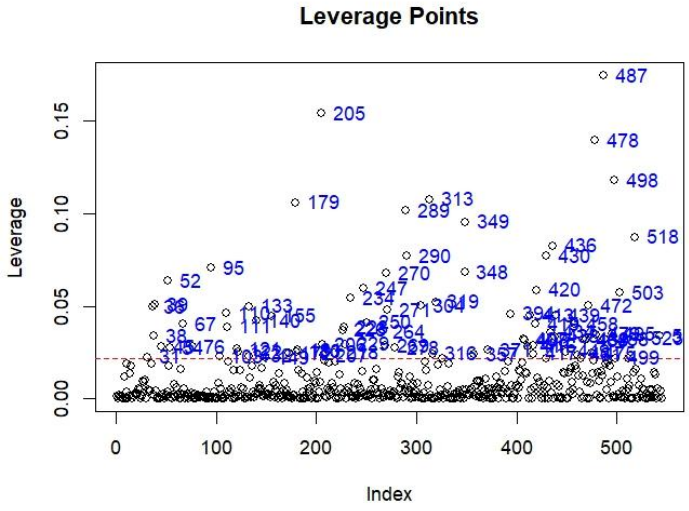
Apps	F.Undergrad	Outstate	PhD	Expend
3.567719	3.242469	1.829918	1.613285	2.019838

Handling the Unusual Observations for improved linear regression modelling:

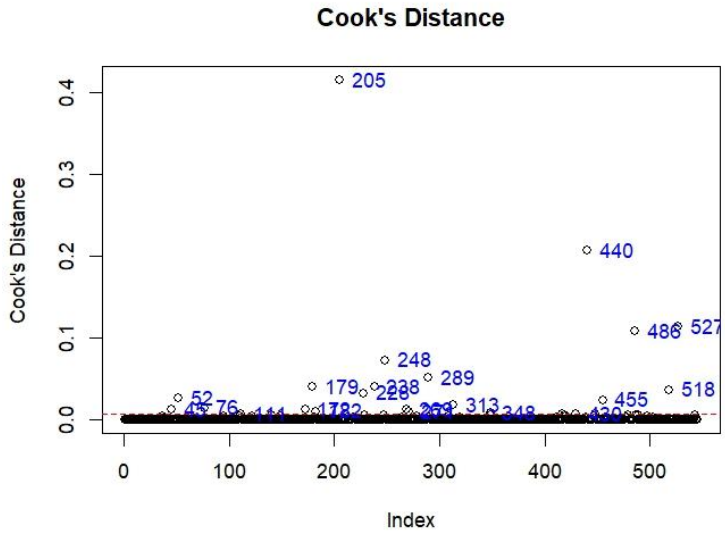
1. Outliers (3): Three values with high standardized residuals (z-score) were calculated from the initial Logistic Regression Model, these values were recognized and removed as they were above the threshold value of 3. These observations do not fit the genral pattern and prove to be a poor fit for these points, such points are represented by unusual financial structure, tuition, enrollment characteristics that are different from majority of the colleges.



2. <u>High-Leverage Points(81):</u> The Hat values were calculated and 81 observations were recognized to be significantly higher than the average leverage values, this is two times the average leverage values. This is calculated by measuring the influence of the 81 individual data-points on the fitted values.



3. <u>Influencial Points (22):</u> The Cooks Distance showed 22 data-points to be influencial points, this is calculated my measuring the influence of deleteing the given observations. These are identified by checking the values greater than 4 divided by the number of observations.



4. 88 points unusual observations warrants the removal of all the unsual observations and re-running the model once again.

After removing the outliers and re-running the model resulted in increased AIC value from 173.95 to 25.779 as we can see below.

```
Call:
glm(formula = Private ~ Apps + F.Undergrad + Outstate + PhD +
    Expend, family = binomial, data = train_set_cleaned)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.4529804  7.4492809   0.598   0.5500
Apps        -0.0006902  0.0023231  -0.297   0.7664
F.Undergrad -0.0013816  0.0019475  -0.709   0.4781
Outstate     0.0026050  0.0012957   2.010   0.0444 *
PhD         -0.3741138  0.2335330  -1.602   0.1092
Expend       0.0011037  0.0013484   0.819   0.4131
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 477.665  on 456  degrees of freedom
Residual deviance:  13.779  on 451  degrees of freedom
AIC: 25.779

Number of Fisher Scoring iterations: 12
```

e) All-subset Regression

After dealing with the outliers and refining the predictor values, an all-subset regression was used to finally choose the following 5 predictors:

i. Full-time Undergraduate Students (F.Undergrad)
ii. Out of State Tuition (Outstate)
iii. Percentage of Faculty with PhD (PhD)
iv. Student-Faculty Ratio (S.F.Ratio)
v. Expenditure per student (Expend)

```
Call:
glm(formula = Private ~ F.Undergrad + Outstate + PhD + S.F.Ratio +
    Expend, family = binomial, data = train_set_cleaned)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.4923140 11.8034789  -0.042   0.9667
F.Undergrad -0.0017754  0.0009463  -1.876   0.0606 .
Outstate     0.0024546  0.0011363   2.160   0.0308 *
PhD         -0.4057568  0.2470685  -1.642   0.1005
S.F.Ratio    0.2940679  0.5021670   0.586   0.5581
Expend       0.0016600  0.0017096   0.971   0.3316
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 477.665  on 456  degrees of freedom
Residual deviance:  13.515  on 451  degrees of freedom
AIC: 25.515

Number of Fisher Scoring iterations: 12
```

BICThe equation for the final model is
Private=−0.4923140−0.001754×F.Undergrad+0.0024546×Outstate−0.405768×PhD+0.2940679×S.F.Ratio+0.0016600×Expend

After comparing the second model with the best model (initial model doesn't have the same dataset size), we see that the **Akaike Information Criterion (AIC) value** decreased to **25.515**, with indication towards the best model.

The below image shows the best model variables have the least complexity among themselves.

```
> # Comparing all the models except the initial model since it is not of the same dataset
> anova(second_model, best_model)
Analysis of Deviance Table

Model 1: Private ~ Apps + F.Undergrad + Outstate + PhD + Expend
Model 2: Private ~ F.Undergrad + Outstate + PhD + S.F.Ratio + Expend
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       451     13.779
2       451     13.515  0  0.26435
> AIC(second_model, best_model)
             df      AIC
second_model  6 25.77919
best_model    6 25.51484
> BIC(second_model, best_model)
             df      BIC
second_model  6 50.52729
best_model    6 50.26294
```

1. **ANOVA test:** This test indicates that the best model has least residual deviance
2. **Akaike Information Criterion (AIC) value:** This indicates that the best model has the lowest AIC value.
3. **Bayesian Information Criterion (BIC) value:** This indicates that the best model has the lowest BIC value.

f) Model Evaluation and Performance

1. Confusion Matrix (Training Set)

```
          Actual
Predicted   0   1
        0  98   1
        1   1 357
```

Interpretation of the misclassification errors:

True Positive (TP): Correctly predicted private institutes **(357)**

True Negative (TN): Correctly predicted public institutes **(98)**

False Positive (TP): Falsely predicted private institutes **(1)**

False Negative (TN): Falsely predicted public institutes **(1)**

2. Performance Metrics (Training Set)

```
> train_accuracy
[1] 0.9956236
> train_precision
[1] 0.9972067
> train_recall
[1] 0.9972067
> train_specificity
[1] 0.989899
```

Accuracy is **99.5%,**

Precision is **99.7%,**

Recall is **99.7%** &

Specificity is **98.9%**

3. Confusion Matrix (Testing Set)

```
          Actual
Predicted   0   1
        0  57   7
        1   6 162
```

Interpretation of the misclassification errors:

True Positive (TP): Correctly predicted private institutes **(162)**

True Negative (TN): Correctly predicted public institutes **(57)**

False Positive (TP): Falsely predicted private institutes **(7)**

False Negative (TN): Falsely predicted public institutes **(6)**

4. Performance Metrics (Testing Set)

```
> test_accuracy
[1] 0.9439655
> test_precision
[1] 0.9585799
> test_recall
[1] 0.9642857
> test_specificity
[1] 0.890625
```

Accuracy is **94.3%,**

Precision is **95.8%,**

Recall is **96.4%** &

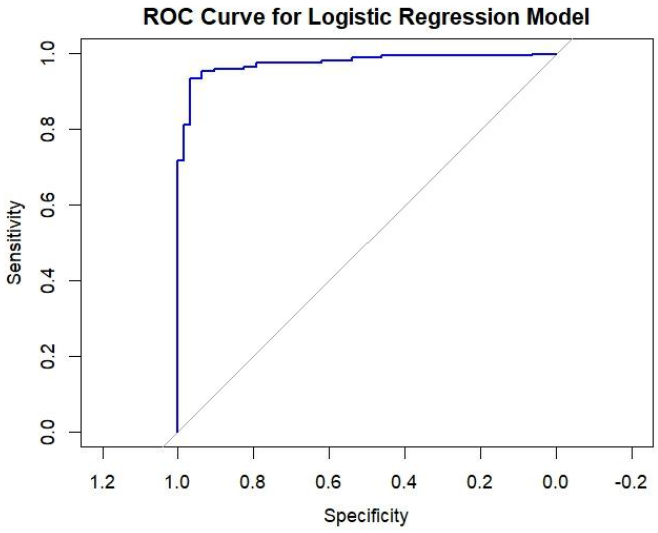
Specificity is **89%**

- The Accuracy of the testing data is slightly less than the training data suggesting there is a slight drop in accuracy with the model is introduced to a new data.

- The Precision of the testing data testing data is slightly less than the training data suggesting the model makes slightly more false positives when exposed to new data.
- The Recall of the testing data testing data is slightly less than the training data suggesting that the model missed a few true positives as compared to the training data.
- The Specificity of the testing data testing data is slightly less than the training data suggesting the ability of the model to detect true negatives is lower on the testing data.

The False Negatives are more damaging than False positive, as private colleges have a different tuition structure, funding, student expectations, etc. than public colleges.

5. Receiver Operator Characteristics (ROC) Curve and AUC (Area Under the Curve)



"AUC: 0.975673898750822"

- As we can understand from the ROC curve, the y-axis represents the True Positive Rate (TPR)/ Sensitivity, the x-axis represents the FalsePositive Rate (FPR)/ Specificity and the diagonal line represents a random classifer with no discrimination ability.
- The blue curve represents the performance of our logistic regression model, the closer the blue line is to the top left corner, the better is the ability of the model to distinguish between the positive and negative classes.
- We can see that our regression model is performing very well as it follows the left hand side border and then the top border.
- Area Under Curve (AUC) Value: 0.975673898750822
- AUC values typically range from 0 to 1, with a value closer to 1 meaning better model performance
- An AUC Value of 0.97 near to 1 indicates that the logistic regression model discriminates the positive and negative classes effectively, this means we can detect which institutes are private and which institutes are publics efficiently.

Conclusion

Conclusion:

In this analysis, we successfully developed a logistic regression model, and compared its performance metrics to figure out how accurately the model predicts our response variable (Public or Private) based on institutional features. The final model is the best model which properly discriminates which institutes are private and which institutes are publics.

Key Findings:

- Private colleges charge significantly more Out of State tuition and even have a high Expenditure per student.
- We found from the correlation matrix that **Outstate** and **Expend** are the strongest predictors of whether an institute is public or private.
- We noticed that removing the **unusual observation** helped improve the model's accuracy and stability significantly.
- We see that the final model achieves a **high classification accuracy (94.3%)** on the testing data, this indicates a strong generalization.

Works Cited

- Bluman, A. (2018). *Elementary statistics: A step-by-step approach* (10th ed.). McGraw Hill.
- Kabacoff, R. I. (2022). *R in action: Data analysis and graphics with R and tidyverse* (3rd ed.). Manning Publications.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R. Springer.*
- RDocumentation. (n.d.). ISLR::College dataset. Retrieved from *https://rdrr.io/cran/ISLR/man/College.html*

Appendix

R Code:

```
#Author: Yash s
#Created: 2025-01-25
#Edited: 2025-02-02
#Course: ALY6015

cat("\014") # clears console
rm(list = ls()) # clears global environment
try(dev.off(dev.list()["RStudioGD"]), silent = TRUE) # clears plots
try(p_unload(p_loaded(), character.only = TRUE), silent = TRUE) # clears packages
options(scipen = 100) # disables scientific notion for entire R session

library(pacman)
p_load(tidyverse, ISLR, caret, pROC, corrplot, RColorBrewer, car, leaps)

# Loading the college dataset and saving it as dataframe
data("College")
college <- as.data.frame(College)
# write.csv(college, "college_data.csv", row.names = TRUE)

# Converting the Private column to a binary column depicting as 1 as private and 0 as public
# This is done for applying logistic regression
college <- college %>%
  mutate(
    Private = if_else(Private == "Yes", 1, 0),
    Private = as.factor(Private)
  )
```

Report on Logistic Regression on College Dataset

```
# EDA on the college dataset
summary(college) # 0 NA
# Calculating the standard deviation for 'Outstate' and 'Expend'
sd_outstate <- sd(college$Outstate, na.rm = TRUE)
sd_expend <- sd(college$Expend, na.rm = TRUE)

cat("Standard Deviation for Outstate:", sd_outstate, "\n")
cat("Standard Deviation for Expend:", sd_expend, "\n")

# Visualizations
# Distribution of Out-of-state Tuition by private and public college
ggplot(college, aes(x = Outstate, fill = Private)) +
  geom_histogram(bins = 30, alpha = 0.7, position = "identity") +
  scale_fill_manual(values = c("blue", "red")) +
  theme_minimal() +
  labs(
    title = "Distribution of Out-of-state Tuition by Private and Public Colleges",
    x = "Out-of-State Tuition (USD)",
    y = "Private(Red), Public(Blue)",
    fill = "Private(1), Public(0)"
  )
# Box plot of Out-of-state Tuition by private and public college
ggplot(data = college, aes(x = Private, y = Outstate, fill = Private)) +
  geom_boxplot() +
  scale_fill_manual(values = c("blue", "red"))+
  theme_minimal() +
  labs(
    title = "Out-of-state Tuition by Private and Public Colleges",
    x = "Private(1), Public(0)",
```

```
    y = "Out-of-State Tuition (USD)",
    fill = "Private(1), Public(0)"
  )
# Box plot of Full-time Undergraduates by private and public college
ggplot(college, aes(x = Private, y = F.Undergrad, fill = Private)) +
  geom_boxplot() +
  scale_fill_manual(values = c("blue", "red")) +
  theme_minimal() +
  labs(title = "Full-time Undergraduates by Private and Public Colleges",
    x = "Private(1), Public(0)",
    y = "Full-time Undergraduates (Count)",
    fill = "Private(1), Public(0)")

# Correlation matrix
# Selecting only numeric columns
num_cols <- college %>% select_if(is.numeric)
# Computing and plotting the correlation matrix
corr_matrix <- cor(num_cols)
corrplot(corr_matrix, method = "color", col = brewer.pal(n = 8, name = "RdBu"),
         tl.cex = 0.8, cl.cex = 0.8, number.cex = 0.6, addCoef.col = "black")

# Checking correlation for all key predictors
# Apps
apps_corr  <- corr_matrix["Apps", ]
apps_corr <-sort(apps_corr, decreasing = TRUE)
apps_corr
high_corr_apps <- names(apps_corr[2])
high_corr_apps #Accept
# Scatter plot for highest correlated variable showing relationship between Accepted
applications and Applications received
```

```
ggplot(college, aes(x = Accept, y = Apps, color = Private)) +
  geom_point(alpha = 0.6) +
  theme_minimal()  +
  labs(title = "Scatterplot of Applications Recieved vs. Accepted Applications",
       x = "No. of Accepted Applications (Count)",
       y = "No. of Applications recieved (Count)",
       color = "Private(1), Public(0)")
# F.Undergad
fundergrad_corr  <- corr_matrix["F.Undergrad", ]
fundergrad_corr <- sort(fundergrad_corr, decreasing = TRUE)
fundergrad_corr
high_corr_fundg <- names(fundergrad_corr[2])
high_corr_fundg #Enroll
# Scatter plot for highest correlated variable showing relationship between Expenditure (per student) and Out-of-state tuition
ggplot(college, aes(x = Enroll, y = F.Undergrad, color = Private)) +
  geom_point(alpha = 0.6) +
  theme_minimal()  +
  labs(title = "Scatterplot of Full-time Undergraduates vs. Students Enrolled",
       x = "No. of Students Enrolled (Count)",
       y = "No. of Full-time Undergraduates (Count)",
       color = "Private(1), Public(0)")
# Outstate
outstate_corr  <- corr_matrix["Outstate", ]
outstate_corr <- sort(outstate_corr, decreasing = TRUE)
outstate_corr
high_corr_outstate <- names(outstate_corr[2])
high_corr_outstate #Expend
# Scatter plot for highest correlated variable showing relationship between Expenditure (per student) and Out-of-state tuition
```

```
ggplot(college, aes(x = Expend, y = Outstate, color = Private)) +
  geom_point(alpha = 0.6) +
  theme_minimal()  +
  labs(title = "Scatterplot of Out-of-State Tuition vs. Expenditure (per student)",
       x = "Expenditure (per student)",
       y = "Out-of-State Tuition (USD)",
       color = "Private(1), Public(0)")
# PhD
phd_corr  <- corr_matrix["PhD", ]
phd_corr <-sort(phd_corr, decreasing = TRUE)
phd_corr
high_corr_phd <- names(phd_corr[2])
high_corr_phd # Terminal
# Scatter plot for highest correlated variable showing relationship between Terminal Degree Faculty and PhD Faculty
ggplot(college, aes(x = Terminal, y = PhD, color = Private)) +
  geom_point(alpha = 0.6) +
  theme_minimal()  +
  labs(title = "Scatterplot of PhD Faculty vs. Terminal Degree Faculty",
       x = "Faculty with Terminal Degrees (%)",
       y = "Faculty with a PhD (%)",
       color = "Private(1), Public(0)")



# Creating the Training and Testing dataset- maintaining a % of event rate 70/30 split
set.seed(123)
trainIndex <- createDataPartition(college$Private, p = 0.7, list = FALSE, times = 1)
train_set <- college[trainIndex, ]
test_set <- college[-trainIndex, ]
```

```
# Feature Selection- Step-wise Selection
model_step <- step(glm(Private ~ ., data = train_set, family = binomial), direction = "both")
summary(model_step)

# Creating an initial classifier model from the above step with the 5 predictors
initial_model <- glm(formula = Private ~ Apps + F.Undergrad + Outstate +
    PhD + Expend, family = binomial,
  data = train_set)
summary(initial_model)

# Plot diagnostic graphs for the regression model
par(mfrow = c(2, 2))
plot(initial_model)
dev.off()

# Check for multi-collinearity using VIF
vif(initial_model)

# Handling unusual observations
# Identifying Outliers
standardized_residuals <- rstandard(initial_model)
outlier_threshold <- 3
outliers <- which(abs(standardized_residuals) > outlier_threshold)
print(outliers)
# Visualizing outliers
plot(standardized_residuals, main = "Standardized Residuals", ylab = "Residuals", xlab =
"Index")
abline(h = c(-outlier_threshold, outlier_threshold), col = "red", lty = 2)
text(outliers, standardized_residuals[outliers], labels = outliers, col = "blue", pos = 4)
```

```
# Identifying high-Leverage points
leverage <- hatvalues(initial_model)
leverage_threshold <- 2 * mean(leverage)
high_leverage <- which(leverage > leverage_threshold)
print(high_leverage)
# Visualizing high leverage points
plot(leverage, main = "Leverage Points", ylab = "Leverage", xlab = "Index")
abline(h = leverage_threshold, col = "red", lty = 2)
text(high_leverage, leverage[high_leverage], labels = high_leverage, col = "blue", pos = 4)

# Identifying influential observations
cooks <- cooks.distance(initial_model)
influential_threshold <- 4 / nrow(train_set)
influential_points <- which(cooks > influential_threshold)
print(influential_points)
# Visualizing influential points
plot(cooks, main = "Cook's Distance", ylab = "Cook's Distance", xlab = "Index")
abline(h = influential_threshold, col = "red", lty = 2)
text(influential_points, cooks[influential_points], labels = influential_points, col = "blue", pos
= 4)

# Combining all unusual observations into a single vector
unusual_points <- sort(unique(c(high_leverage, outliers, influential_points)))
print(unusual_points)

# Removing the unusual observations
train_set_cleaned <- train_set[-unusual_points, ]

second_model <- glm(formula = Private ~ Apps + F.Undergrad + Outstate +
                    PhD + Expend, family = binomial,
```

```
                data = train_set_cleaned)
summary(second_model)

# All subset regression
best_subset <- regsubsets(Private ~., data = train_set_cleaned, nvmax = 5)
reg_summary <- summary(best_subset)
reg_summary
# Best model by Mallow's Cp and BIC
which.min(reg_summary$cp)
which.min(reg_summary$bic)

# Let us make the best model using 5 predictors
best_model <- glm(formula = Private ~ F.Undergrad + Outstate +
    PhD + S.F.Ratio + Expend, family = binomial,
  data = train_set_cleaned)
summary(best_model)

# Comparing all the models except the initial model since it is not of the same dataset
anova(second_model, best_model)
AIC(second_model, best_model)
BIC(second_model, best_model)

# Prediction on training set
train_pred <- predict(best_model, newdata = train_set_cleaned, type = "response")

# Converting probabilities to binary classification (0.5 threshold)
train_pred_class <- ifelse(train_pred > 0.5, 1, 0)

# Confusion Matrix for Training Set
```

```
conf_matrix_train <- table(Predicted = train_pred_class, Actual = train_set_cleaned$Private)
print(conf_matrix_train)

# Computing performance metrics
train_accuracy <- sum(diag(conf_matrix_train)) / sum(conf_matrix_train)
train_precision <- conf_matrix_train[2, 2] / sum(conf_matrix_train[, 2])
train_recall <- conf_matrix_train[2, 2] / sum(conf_matrix_train[2, ])
train_specificity <- conf_matrix_train[1, 1] / sum(conf_matrix_train[1, ])

# Prediction on testing set
test_pred <- predict(best_model, newdata = test_set, type = "response")

# Converting probabilities to binary classification
test_pred_class <- ifelse(test_pred > 0.5, 1, 0)

# Confusion Matrix for Testing Set
conf_matrix_test <- table(Predicted = test_pred_class, Actual = test_set$Private)
print(conf_matrix_test)

# Computing performance metrics for testing set set
test_accuracy <- sum(diag(conf_matrix_test)) / sum(conf_matrix_test)
test_precision <- conf_matrix_test[2, 2] / sum(conf_matrix_test[, 2])
test_recall <- conf_matrix_test[2, 2] / sum(conf_matrix_test[2, ])
test_specificity <- conf_matrix_test[1, 1] / sum(conf_matrix_test[1, ])

# Computing ROC curve
roc_curve <- roc(test_set$Private, test_pred)

# Plotting ROC Curve
```

```
plot(roc_curve, col = "blue", main = "ROC Curve for Logistic Regression Model")


# Computing AUC value

auc_value <- auc(roc_curve)

print(paste("AUC:", auc_value))


##############################################################################
#####

# End of Assignment 3
```