MODULE 3 - R PRACTICE ASSIGNMENT REPORT

ANALYSIS OF ISLR COLLEGE DATASET

Ayush Anand

NORTHEASTERN UNIVERSITY
COURSE: ALY 6015: INTERMEDIATE ANALYTICS
SUBMITTED BY: AYUSH ANAND
PROFESSOR: XYZ
DATE: 20TH NOVEMBER 2024

Table of Contents

Ayush Anand

Introduction

About the Dataset

This report is about the "College" dataset from something called the ISLR package. It has information on 777 colleges and universities in the United States. There are 18 things we look at, like how many students get in, how many join, money stuff, and how good the teachers are. The main goal is to figure out if a college is private or public based on these things.

Objectives

The main goals of this analysis are:

1. To look at the differences between private and public universities.
2. To find the most important things that make private and public schools different.
3. To make and check a model that can guess if a school is private or public.

Questions of Interest

1. What are the biggest differences between private and public universities when it comes to the number of students, how much they cost, and how good they are at teaching?
2. Which things are the most important for deciding if a school is private or public?
3. How good can we be at guessing if a school is private or public based on its details?
4. How do out-of-state tuition costs and graduation rates relate to each other for private and public universities?

Data Preparation

We started the analysis by loading the College dataset and checking if anything was missing. Nothing was missing, so we had all the data we needed for our analysis.

Descriptive Statistics

We made a big summary table using the gtsummary package to compare important numbers between private and public universities. The table shows averages, how much the numbers vary, and p-values to compare the groups.

Summary Table by University Type

Variable	Overall (N = 777)1	No (N = 212)1	Yes (N = 565)1	p-value2
Accept	2,018.8 (2,451.1)	3,919.3 (3,477.3)	1,305.7 (1,369.5)	<0.001
Enroll	780.0 (929.2)	1,640.9 (1,261.6)	456.9 (457.5)	<0.001
Top10perc	27.6 (17.6)	22.8 (16.2)	29.3 (17.9)	<0.001
Top25perc	55.8 (19.8)	52.7 (20.1)	57.0 (19.6)	0.007
Outstate	10,440.7 (4,023.0)	6,813.4 (2,145.2)	11,801.7 (3,707.5)	<0.001
Room.Board	4,357.5 (1,096.7)	3,748.2 (858.1)	4,586.1 (1,089.7)	<0.001

Ayush Anand

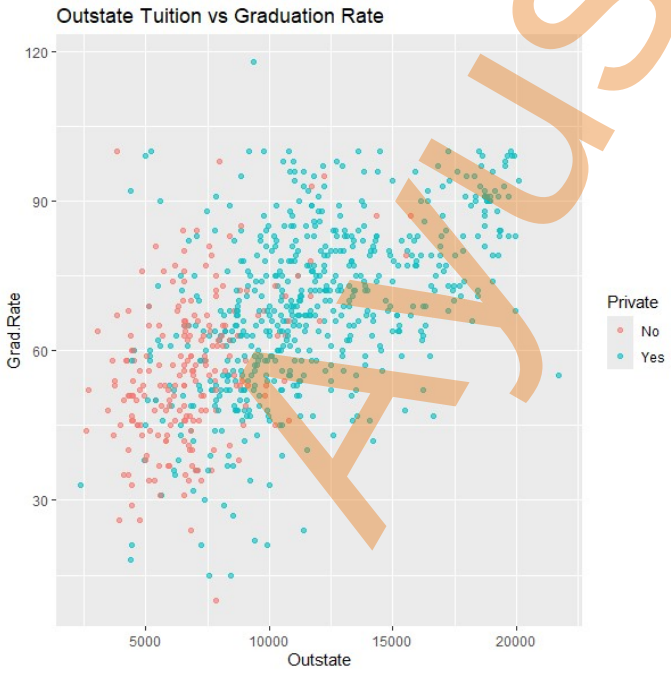
Books	549.4 (165.1)	554.4 (135.7)	547.5 (174.9)	0.002
Personal	1,340.6 (677.1)	1,677.0 (677.5)	1,214.4 (632.9)	<0.001
PhD	72.7 (16.3)	76.8 (12.3)	71.1 (17.4)	<0.001
Terminal	79.7 (14.7)	82.8 (12.1)	78.5 (15.5)	0.003
S.F.Ratio	14.1 (4.0)	17.1 (3.4)	12.9 (3.5)	<0.001
perc.alumni	22.7 (12.4)	14.4 (7.5)	25.9 (12.4)	<0.001
Expend	9,660.2 (5,221.8)	7,458.3 (2,695.5)	10,486.4 (5,682.6)	<0.001
Grad.Rate	65.5 (17.2)	56.0 (14.6)	69.0 (16.7)	<0.001

Key Insights

- **Admissions and Enrollment:** Private universities are pickier and usually get more students.
- **Student Profile:** There isn't much difference in how good the students are between private and public universities, looking at the Top10Perc and Top25Perc scores.
- **Costs:** Private universities cost more, including tuition, room, and food.
- **Faculty and Resources:** Private universities have more teachers with advanced degrees and smaller classes.
- **Outcomes:** Private universities have higher graduation rates, and their graduates stay more connected.

Data Visualizations

1. **Scatter Plot: Outstate Tuition vs Graduation Rate**



Key Insights:

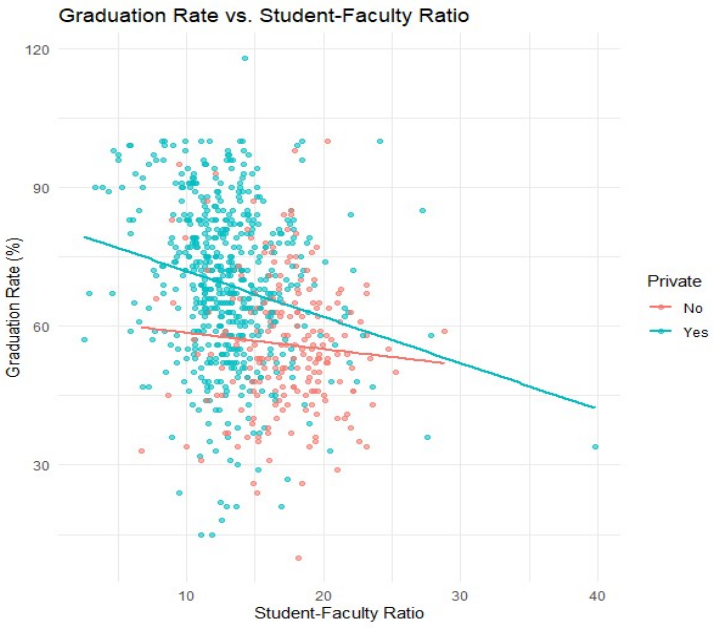
No Clear Connection: This means higher out-of-state tuition doesn't always mean better or worse graduation rates.

Private Universities: Private schools usually have higher out-of-state tuition compared to public ones.

Outliers: There are some schools, especially private ones, that have high out-of-state tuition but low graduation rates. These schools might have special things about them or certain problems.

Ayush Anand

2. **Scatter Plot: Graduation Rate vs Student-Faculty Ratio**



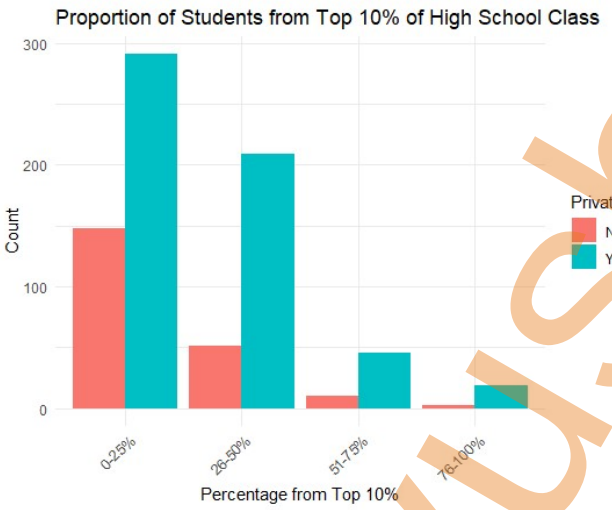
Key Insights:

Negative Connection: When there are more students per teacher, graduation rates usually go down. This means that smaller classes and more time with teachers might help students graduate more.

Private Universities: Private schools have fewer students per teacher and higher graduation rates compared to public schools. This shows that private schools might have more money to spend on teachers and small classes, which helps students do better.

Overall: The graph shows that the student-teacher ratio is important for graduation rates, especially in private schools. But we also need to think about other things that could affect graduation rates.

3. **Grouped Bar Plot: Proportion of Students from Top 10% of High School Classes**

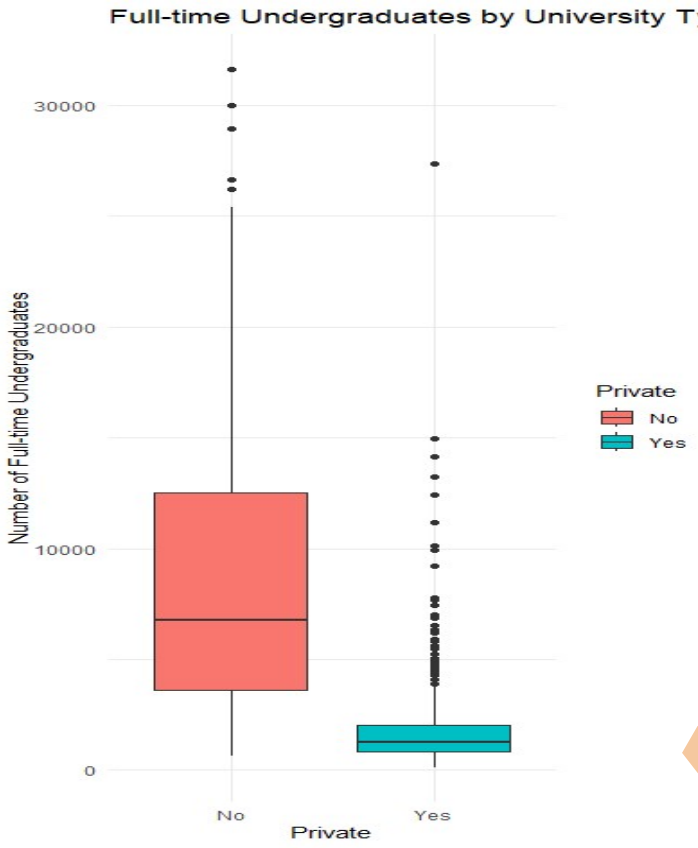


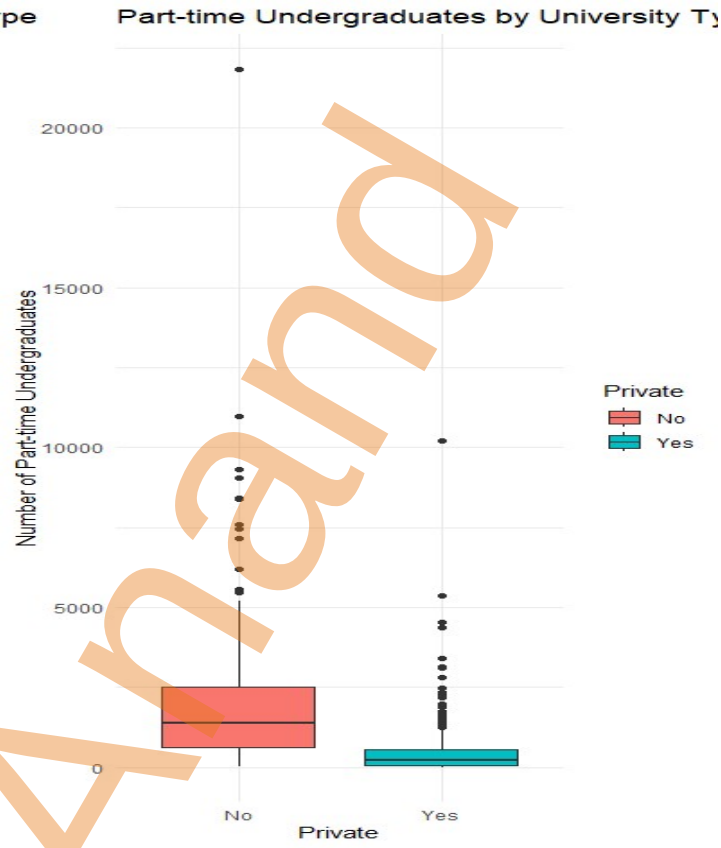
Key Insights:

The bar chart shows how universities are split based on how many students are from the top 10% of their high school class, and whether the school is private or not. Private universities usually get more students who did really well in high school. But the chart also shows that both private and public schools have students from all kinds of backgrounds.

Ayush Anand

4. **Boxplots: Full-Time and Part-Time Undergraduate Enrollment by University Type**



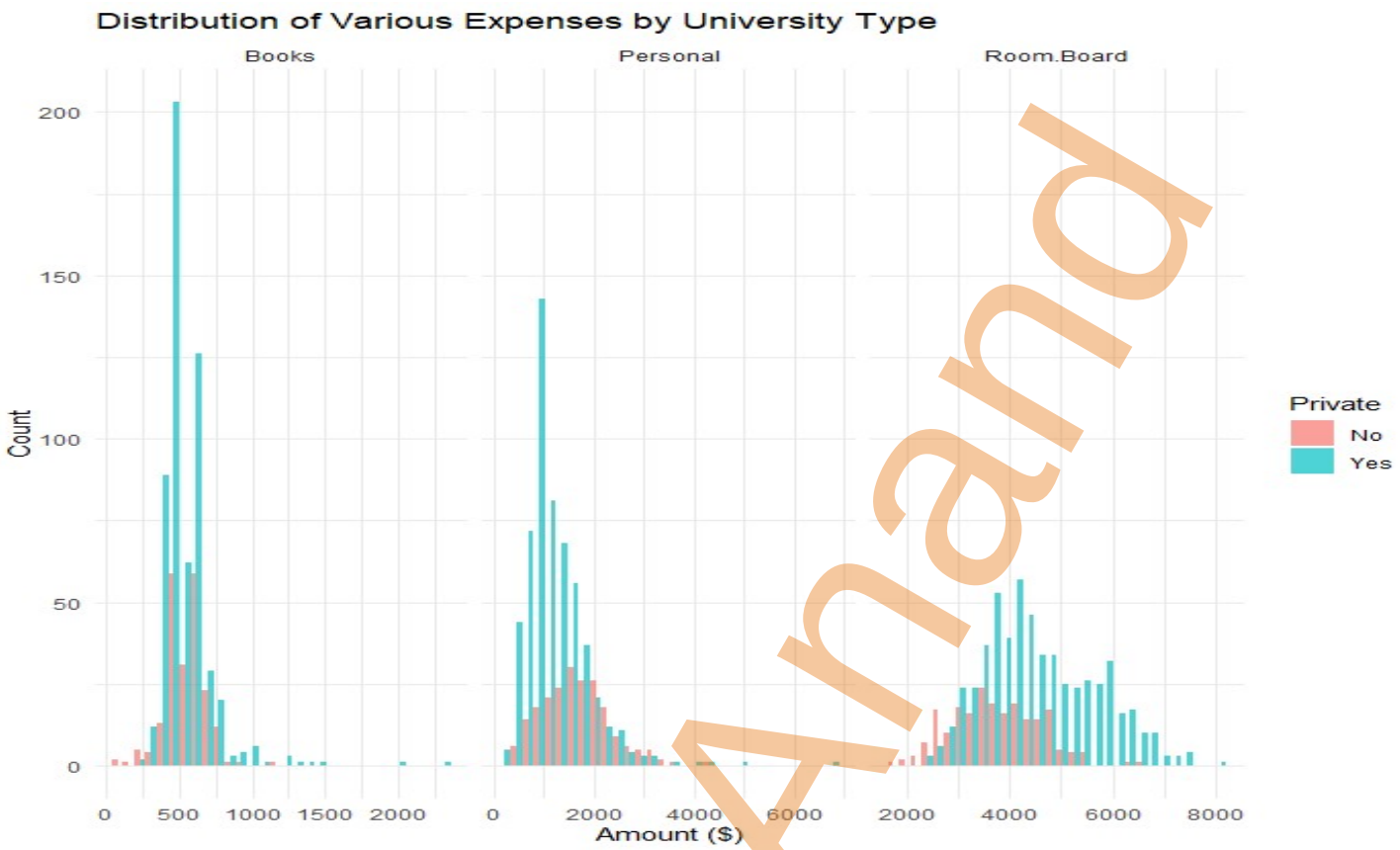


Key Insights:

- **Private Universities**: Private schools usually have more students, both full-time and part-time, compared to public schools.
- **Variation**: Private schools have a bigger difference in the number of students they have, which means they come in lots of different sizes and use different ways to get students.
- **Public Universities**: Public schools have a more even number of students, with not much difference in how many students they have.

Ayush Anand

5. **Faceted Histogram: Various Expenses by University Type**



Key Insights:

Books:

- **Public Universities:** The book costs are mostly around $500, with a few schools having higher costs.
- **Private Universities:** The book costs are mostly around $750, which means they are usually higher than public schools.

Personal Expenses:

- **Public Universities:** Most schools have personal costs around $2000, with a few that are higher.
- **Private Universities:** Personal costs are usually around $3000, which is higher compared to public schools.

Room and Board:

- **Public Universities**: Most schools have room and board costs around $4000, with a few that are higher.
- **Private Universities:** Room and board costs are mostly around $6000, which is much higher than public schools.

Ayush Anand

Logistic Regression Modeling

We made two logistic regression models:

1. **Model-1**: This one used all the information we had.
2. **Model-2**: This one used only some of the information: number of full-time students (F.Undergrad), out-of-state tuition (Outstate), percentage of teachers with PhDs (PhD), and graduation rate (Grad.Rate).

Model Comparison and Evaluation

Metric	Model-1	Model-2
Full-time Undergrads	-0.001**	-0.001***
	(0.000)	(0.000)
Out-of-state Tuition	0.001***	0.001***
	(0.000)	(0.000)
% Faculty with PhD	-0.064*	-0.069***
	(0.034)	(0.018)
Graduation Rate	0.029+	0.026+
	(0.018)	(0.015)
Num.Obs.	545	545
AIC	183.8	171.4
BIC	274.1	192.9

p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Key Insights:

Both models used the same information: number of full-time undergraduates, out-of-state tuition, percent of teachers with PhDs, and graduation rate. The numbers and their importance were similar in both models, but Model 2 showed more important numbers, with more stars.

- **Full-time Undergraduates**: The negative number means that as the number of full-time students goes up, the chance of the outcome goes down.
- **Out-of-State Tuition**: The positive number means that as out-of-state tuition goes up, the chance of the outcome goes up.
- **% Faculty with PhD**: The negative number means that as more teachers have PhDs, the chance of the outcome goes down.
- **Graduation Rate**: The positive number means that as graduation rate goes up, the chance of the outcome goes up.

Ayush Anand

Model Fit:

AIC (Akaike Information Criterion): Model 2 has a lower AIC value (171.4) than Model 1 (183.8). A lower AIC means the model fits better.

BIC (Bayesian Information Criterion): Model 2 also has a lower BIC value (192.9) compared to Model 1 (274.1). A lower BIC means it's a better model with fewer things to look at.

Why Model-2 is better:

Model 2 is better because it has lower AIC and BIC values. These numbers make models with too many details less good, and Model 2 seems to be a good balance between fitting well and not being too complicated. Also, Model 2 has more important numbers, which means it might show the real connections between things better.

Model Evaluation for Train Set

Confusion Matrix for Training Set

	Actual Values		
		No	**Yes**
Predicted Values	**No**	True Positive 133	False Positive 11
	Yes	False Negative 16	True Negative 385

Performance matrices for training sets:

- **Accuracy:** (TP + TN) / (TP + FP + FN + TN) = 95.05% The model correctly classifies 95.05% of all universities.
- **Precision:** TP / (FP + TP) = 92.36% Of universities predicted as public, 92.36% are actually public.
- **Specificity:** TP / (TP + FN) = 89.26% The model correctly identifies 89.26% of public universities
- **Sensitivity:** TN / (TN + FP) = 97.23% The model correctly identifies 97.23% of private universities

Key Insight:

Mistaking private universities for public ones (false negatives) is more harmful than the other way around. This is because it could mean not knowing the real costs and not planning resources properly.

Model Evaluation for Test Set

Confusion Matrix for Testing Set

	Actual Values		
		No	**Yes**
Predicted Values	**No**	**True Positive 55**	**False Positive 6**
	Yes	**False Negative 8**	**True Negative 163**

Ayush Anand

Performance matrices for testing sets:

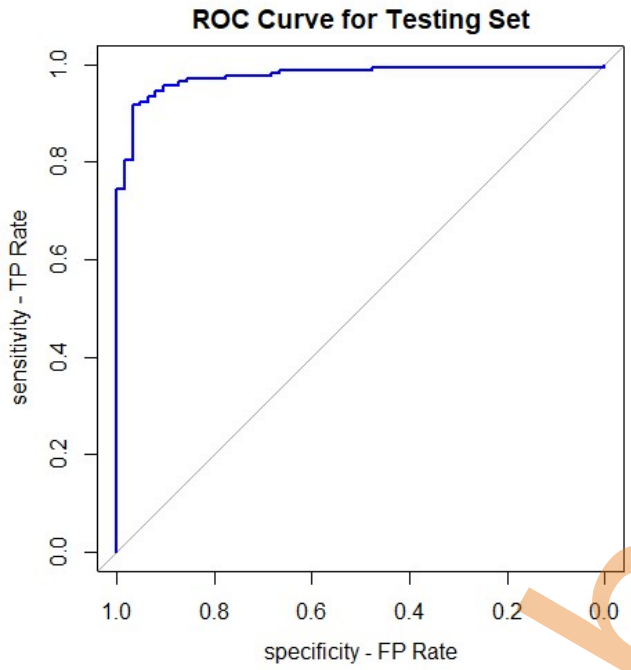
Accuracy: (TP + TN) / (TP + FP + FN + TN) = 93.97% The model correctly classifies 93.97% of all universities.

Precision: TP / (FP + TP) = 90.16% Of universities predicted as public, 90.16% are actually public.

Specificity: TP / (TP + FN) = 87.3% The model correctly identifies 87.3% of public universities

Sensitivity: TN / (TN + FP) = 96.45% The model correctly identifies 96.45% of private universities

ROC and AUC for Testing Set:



The value of Area Under the Curve (AUC) is 0.9766.

Model Performance: The AUC value of 0.9766 shows that the model works really well, since it's almost 1.0, which means perfect classification. This high AUC means the model is very good at telling the difference between private and public universities.

Key Insight: The ROC curve and the high AUC value tell us that the logistic regression model is really good at figuring out if a university is private or public using the chosen features.

Conclusion

This analysis looked at the College dataset from the ISLR library to see what makes private and public universities different in the United States. We used statistics, data analysis, and logistic regression to look at 777 schools.

Key Findings:

- Private universities have higher costs for out-of-state tuition, room and board, and education.
- They have fewer students per teacher, more teachers with advanced degrees, better alumni networks, and higher graduation rates.

Models: We made two models, and Model-2 (using F.Undergrad, Outstate, PhD, and Grad.Rate) was the best because it had lower AIC and BIC values.

Model Performance:

- Training set accuracy: 95.05%

Ayush Anand

- Testing set accuracy: 93.97%
- AUC value: 0.9766

The confusion matrix showed that getting private universities wrong (false negatives) was worse than getting public ones wrong (false positives). This could lead to problems with planning and using resources.

The ROC curve showed the model is great at telling private and public schools apart, with a high AUC value of 0.9766. This means the features we used do a good job of classifying schools, making the model a reliable tool for figuring out if a university is private or public.

References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). ***An introduction to statistical learning: With applications in R*****. Springer. Retrieved from** https://www.statlearning.com/

GeeksforGeeks. (n.d.). ***Confusion matrix in R*****. Retrieved from** https://www.geeksforgeeks.org/confusion-matrix-in-r/

R-Bloggers. (2015, September). ***How to perform a logistic regression in R*****. Retrieved from** https://www.r-bloggers.com/2015/09/how-to-perform-a-logistic-regression-in-r/

Appendix

Code Used

```
# Load required libraries
library(ISLR)
library(DataExplorer)
library(gtsummary)
library(ggplot2)
library(gridExtra)
library(dplyr)
library(caret)
library(pROC)

#-----------------------------------------------------------------------------------------------------------

# Load the College dataset
data(College)  # Load dataset
force(College)  # Ensure data is properly loaded


# Get column names
column_names <- names(College)
```

Ayush Anand

```
print(column_names)  # Print column names to check the variables


# Checking for missing values in each column
colSums(is.na(College))  # Summarize missing values per column


# View missing data patterns
plot_missing(College)  # Plot to visualize missing data patterns


#-------------------------------------------------------------------------------------------------------------------
# Summary statistics
# Create a summary table
college_summary <- College %>%
 select(-Apps, -F.Undergrad, -P.Undergrad) %>%  # Exclude some variables for brevity
 tbl_summary(
  by = Private,  # Group by Private/Public status
  statistic = list(
   all_continuous() ~ "{mean} ({sd})",  # Mean and standard deviation for continuous variables
   all_categorical() ~ "{n} ({p}%)"  # Count and percentage for categorical variables
  ),
  digits = all_continuous() ~ 1,  # Limit digits to 1 for continuous variables
  missing = "no"  # Don't show missing data
 ) %>%
 add_p() %>%  # Add p-values for group comparisons
 add_overall() %>%  # Add an overall column
 modify_header(label = "**Variable**") %>%  # Modify column headers
 modify_spanning_header(c("stat_1", "stat_2") ~ "**University Type**") %>%
 bold_labels()  # Make labels bold for better visibility


# Print the summary table
print(college_summary)


#-------------------------------------------------------------------------------------------------------------------
# Visualizations
# Visualize the distribution of public vs private universities
ggplot(College, aes(x = Private)) +
```

Ayush Anand

```
  geom_bar(fill = "steelblue") +
  ggtitle("Distribution of Public vs Private Universities")

# Boxplot of Outstate tuition by Private/Public status
ggplot(College, aes(x = Private, y = Outstate)) +
  geom_boxplot(fill = "orange") +
  ggtitle("Outstate Tuition by University Type")

# Scatter plot of Outstate tuition vs Graduation Rate
ggplot(College, aes(x = Outstate, y = Grad.Rate, color = Private)) +
  geom_point(alpha = 0.6) +
  ggtitle("Outstate Tuition vs Graduation Rate")

# Boxplot of Full-time and Part-time Undergraduates by university type
p1 <- ggplot(College, aes(x = Private, y = F.Undergrad)) +
  geom_boxplot(aes(fill = Private)) +
  labs(title = "Full-time Undergraduates by University Type", y = "Number of Full-time Undergraduates") +
  theme_minimal()

p2 <- ggplot(College, aes(x = Private, y = P.Undergrad)) +
  geom_boxplot(aes(fill = Private)) +
  labs(title = "Part-time Undergraduates by University Type", y = "Number of Part-time Undergraduates") +
  theme_minimal()

# Arrange the two plots side by side
grid.arrange(p1, p2, ncol = 2)

# Scatter plot of Graduation Rate vs. Student-Faculty Ratio
ggplot(College, aes(x = S.F.Ratio, y = Grad.Rate, color = Private)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Graduation Rate vs. Student-Faculty Ratio", x = "Student-Faculty Ratio", y = "Graduation Rate (%)") +
  theme_minimal()

# Grouped bar plot of Top 10% of High School Class
```

Ayush Anand

```
College$Top10perc_cat <- cut(College$Top10perc, breaks = c(0, 25, 50, 75, 100), labels = c("0-25%", "26-50%", "51-75%", "76-100%"))


# Grouped bar chart
ggplot(College, aes(x = Top10perc_cat, fill = Private)) +
 geom_bar(position = "dodge") +
 labs(title = "Proportion of Students from Top 10% of High School Class", x = "Percentage from Top 10%", y = "Count") +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))


# Various Expenses by University Type
College_long <- College %>%
 select(Private, Books, Personal, Room.Board) %>%
 tidyr::pivot_longer(cols = c(Books, Personal, Room.Board), names_to = "Expense_Type", values_to = "Amount")


# Histogram of expenses by university type
ggplot(College_long, aes(x = Amount, fill = Private)) +
 geom_histogram(position = "dodge", bins = 30, alpha = 0.7) +
 facet_wrap(~Expense_Type, scales = "free_x") +
 labs(title = "Distribution of Various Expenses by University Type", x = "Amount ($)", y = "Count") +
 theme_minimal()


#-------------------------------------------------------------------------------------------------------------------------
# Split the data into training and testing sets
set.seed(123)  # Set seed for reproducibility
train_index <- createDataPartition(College$Private, p = 0.7, list = FALSE)  # Split data into 70% train and 30% test
train_data <- College[train_index, ]
test_data <- College[-train_index, ]


# Fit logistic regression models
model1 <- glm(Private ~ ., data = train_data, family = "binomial")  # Full model using all predictors
summary(model1)


model2 <- glm(Private ~ F.Undergrad + Outstate + PhD + Grad.Rate, data = train_data, family = "binomial")  # Reduced model using selected predictors
summary(model2)
```

Ayush Anand

```
library(modelsummary)

# Create a side-by-side comparison table for both models
models_comparison <- modelsummary(
 list("Model-1" = model1, "Model-2" = model2),
 title = "Comparison of Logistic Regression Models",
 stars = TRUE,
 gof_map = c("nobs", "aic", "bic", "r.squared"),
 coef_map = c(
  "F.Undergrad" = "Full-time Undergrads",
  "Outstate" = "Out-of-state Tuition",
  "PhD" = "% Faculty with PhD",
  "Grad.Rate" = "Graduation Rate"
 )
)

# Print the comparison table
print(models_comparison)

#-----------------------------------------------------------------------------------------------------------------
# Create confusion matrices for training and testing sets
# Training set predictions
train_pred <- predict(model2, train_data, type = "response")
train_pred_class <- ifelse(train_pred > 0.5, "Yes", "No")
train_conf_matrix <- confusionMatrix(factor(train_pred_class), factor(train_data$Private), positive = 'Yes')
print(train_conf_matrix)

# Testing set predictions
test_pred <- predict(model2, test_data, type = "response")
test_pred_class <- ifelse(test_pred > 0.5, "Yes", "No")
test_conf_matrix <- confusionMatrix(factor(test_pred_class), factor(test_data$Private), positive = 'Yes')
print(test_conf_matrix)

#-----------------------------------------------------------------------------------------------------------------
```

Ayush Anand

```
# Plot ROC curve for training set
roc_train <- roc(train_data$Private, train_pred)
plot(roc_train, main = "ROC Curve for Training Set", col = "blue", ylab = "Sensitivity - TP Rate", xlab = "Specificity - FP Rate")

# Calculate and print the AUC for Training Set
auc_value_train <- auc(roc_train)
cat("AUC for Training Set:", auc_value_train, "\n")

# Plot ROC curve for testing set
roc_test <- roc(test_data$Private, test_pred)
plot(roc_test, main = "ROC Curve for Testing Set", col = "blue", ylab = "Sensitivity - TP Rate", xlab = "Specificity - FP Rate")

# Calculate and print the AUC for Testing Set
auc_value_test <- auc(roc_test)
cat("AUC for Testing Set:", auc_value_test, "\n")
```

Ayush Anand