MODULE THREE PROJECT

ANALYSIS OF COLLEGE USING GLM AND LOGISTIC REGRESSION

Submitted By: **HARSHIT GAUR**
MASTER OF PROFESSIONAL STUDIES IN ANALYTICS
ALY 6015 : INTERMEDIATE ANALYTICS
CRN : 21454
MARCH 12, 2022
WINTER 2022

Submitted To: **PROF. ROY WADA**

INTRODUCTION

Generalized Linear Models (GLM) represent a class of regression models that allows us to generalize the linear regression approach to accommodate many types of dependent variables.

In a Generalized Linear Model, the dependent variable does not need to be continuous or normally distributed.

Three components of a GLM:

- *Random component* – probability distribution of the response variable
- *Systematic component* - specifies the explanatory variables (X_1, X_2, ... X_k) in the model, more specifically their linear combination in creating the so called linear predictor
- *Link function* - specifies the link between random and systematic components

Logistic Regression is a type of statistical analysis (also known as logit model) which is often used for predictive analytics and modelling, and extends to applications in machine learning. In this analytics approach, the dependent variable is finite or categorical: either A or B (binary regression) or a range of finite options A, B, C or D (multinomial regression). It is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation.

This type of analysis can help predict the likelihood of an event happening or a choice being made. For example, we may want to know the likelihood of a visitor choosing an offer made on our website — or not (dependent variable). Our analysis can look at known characteristics of visitors, such as sites they came from, repeat visits to our site, behaviour on our site (independent variables). Logistic regression models help us determine a probability of what type of visitors are likely to accept the offer — or not. As a result, we can make better decisions about promoting our offer or make decisions about the offer itself.

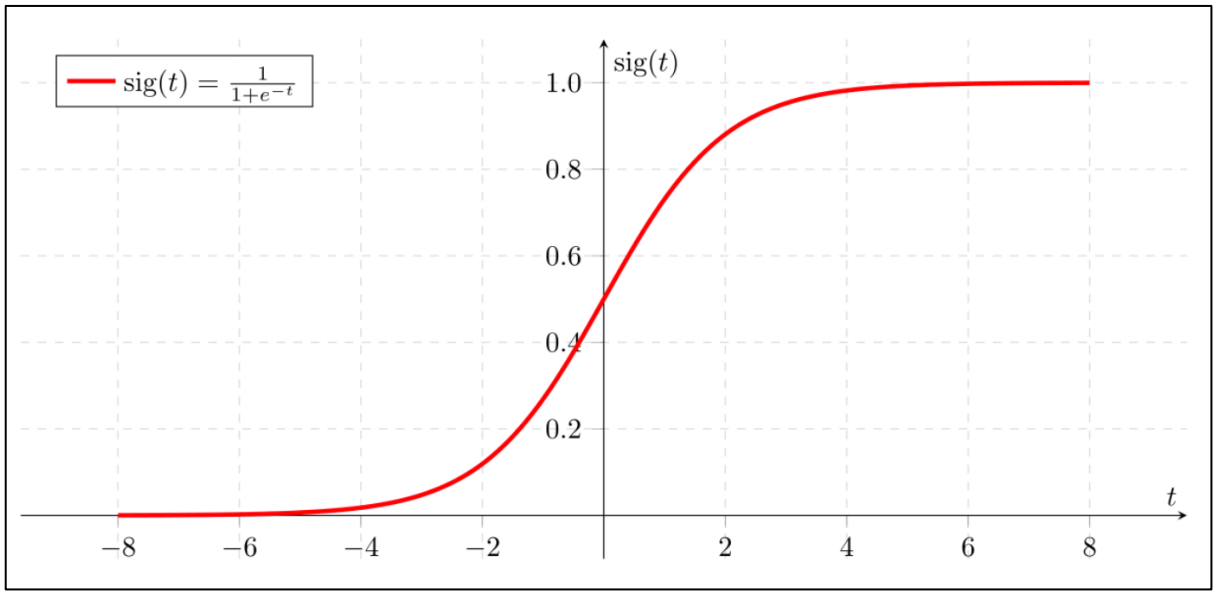

Figure 1.1: Sigmoid Function used for Logistic Regression

ANALYSIS

With the medium of this project, we will build a logistic regression model to predict whether a college is private or public.

COLLEGE DATA SET -

The first 6 data points of the college data set has been presented below for a glimpse of the data set and what all variables are included in the data set.

College Dataset (with 7 out 18 features)

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate
Abilene Christian University	Yes	1,660	1,232	721	23	52	2,885	537	7,440
Adelphi University	Yes	2,186	1,924	512	16	29	2,683	1,227	12,280
Adrian College	Yes	1,428	1,097	336	22	50	1,036	99	11,250
Agnes Scott College	Yes	417	349	137	60	89	510	63	12,960
Alaska Pacific University	Yes	193	146	55	16	44	249	869	7,560
Albertson College	Yes	587	479	158	38	62	678	41	13,500

Table 1.1: College Dataset

Exploratory Data Analysis -

We will analyse the data set to find out some insights and investigate the variables.

Descriptive Statistics of College Dataset

	n	mean	sd	median	min	max	range	skew	kurtosis
Private*	777	1.73	0.45	2.0	1.0	2.0	1.0	-1.02	-0.96
Apps	777	3,001.64	3,870.20	1,558.0	81.0	48,094.0	48,013.0	3.71	26.52
Accept	777	2,018.80	2,451.11	1,110.0	72.0	26,330.0	26,258.0	3.40	18.75
Enroll	777	779.97	929.18	434.0	35.0	6,392.0	6,357.0	2.68	8.74
Top10perc	777	27.56	17.64	23.0	1.0	96.0	95.0	1.41	2.17
Top25perc	777	55.80	19.80	54.0	9.0	100.0	91.0	0.26	-0.57
F.Undergrad	777	3,699.91	4,850.42	1,707.0	139.0	31,643.0	31,504.0	2.60	7.61
P.Undergrad	777	855.30	1,522.43	353.0	1.0	21,836.0	21,835.0	5.67	54.52
Outstate	777	10,440.67	4,023.02	9,990.0	2,340.0	21,700.0	19,360.0	0.51	-0.43
Room.Board	777	4,357.53	1,096.70	4,200.0	1,780.0	8,124.0	6,344.0	0.48	-0.20
Books	777	549.38	165.11	500.0	96.0	2,340.0	2,244.0	3.47	28.06
Personal	777	1,340.64	677.07	1,200.0	250.0	6,800.0	6,550.0	1.74	7.04
PhD	777	72.66	16.33	75.0	8.0	103.0	95.0	-0.77	0.54
Terminal	777	79.70	14.72	82.0	24.0	100.0	76.0	-0.81	0.22
S.F.Ratio	777	14.09	3.96	13.6	2.5	39.8	37.3	0.66	2.52
perc.alumni	777	22.74	12.39	21.0	0.0	64.0	64.0	0.60	-0.11
Expend	777	9,660.17	5,221.77	8,377.0	3,186.0	56,233.0	53,047.0	3.45	18.59
Grad.Rate	777	65.46	17.18	65.0	10.0	118.0	108.0	-0.11	-0.22

Table 1.2: College Dataset Descriptive Statistics Summary

1. The data set contains 777 data points with 18 features.
2. The mean of Top 10 Percentage students from High School (Top10Perc) at 27.56 which is almost close to the median of 23. This shows that the distribution of this feature might be a normal distribution. We can check the normality using Q-Q Plot and Shapiro Wilks test as well. But, from the test results, we can somewhat say that the feature is deviating from the normal distribution with outliers present.

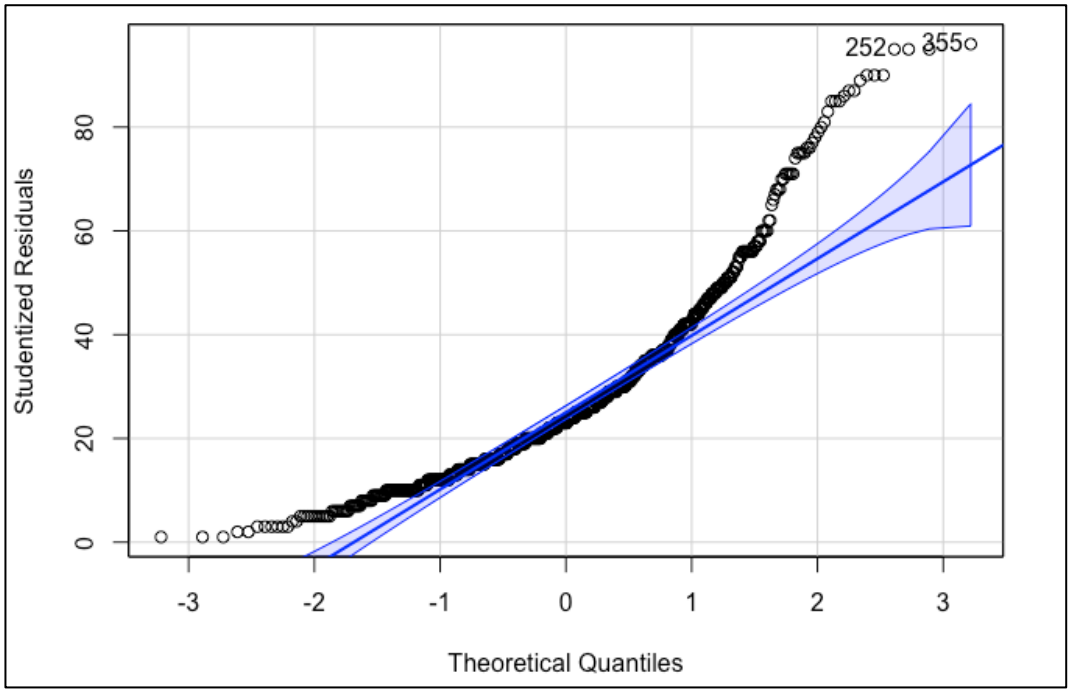

Figure 1.2: Quantile-Quantile Plot for Top10Perc feature

3. The mean of Top 25 Percentage students from High School (Top25Perc) at 55.80 which is almost close to the median of 54. This shows that the distribution of this feature might be a normal distribution. We can check the normality using Q-Q Plot and Shapiro Wilks test as well. From the test results, we can somewhat say that the feature is from the normal distribution with outliers present.
4. This feature is also the least skewed (right) amongst the features available in dataset.

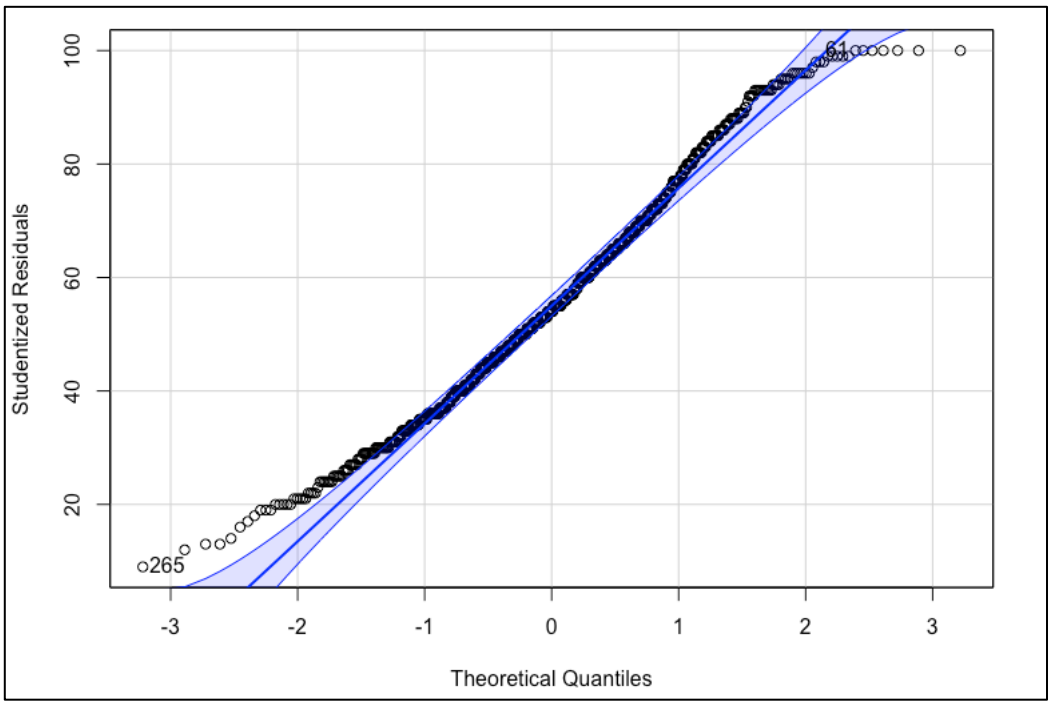

Figure 1.3: Quantile-Quantile Plot for Top25Perc feature

5. The maximum Out-of-state tuition cost from these colleges is $21,700.00.
6. The standard deviations of the feature are also significant which means the data is dispersed throughout the data set.
7. The wins distribution is slightly negatively skewed as suggested by the value of -0.18 of skewness.

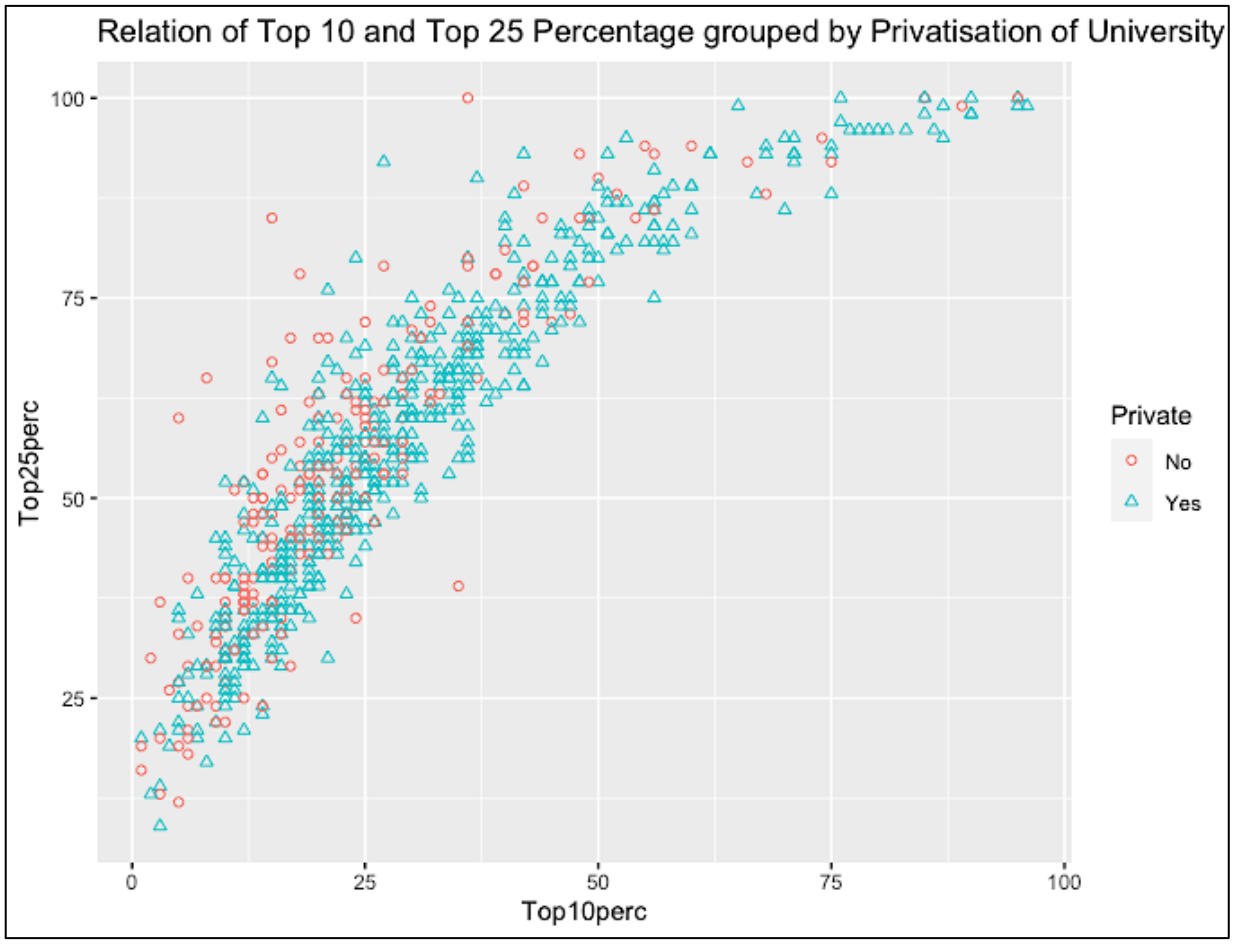

Figure 1.4: New Top 10 and Top 25 Percentage relation grouped by Private/Public University

1. The above plot shows that there is a linear relationship between the new Top 10 percentage of students from high school and new Top 25 percentage of students from high school.

2. The grouping by private/public university shows that there exists a linear relationship between these 2 groups to the relation of Top10 and Top25 features.

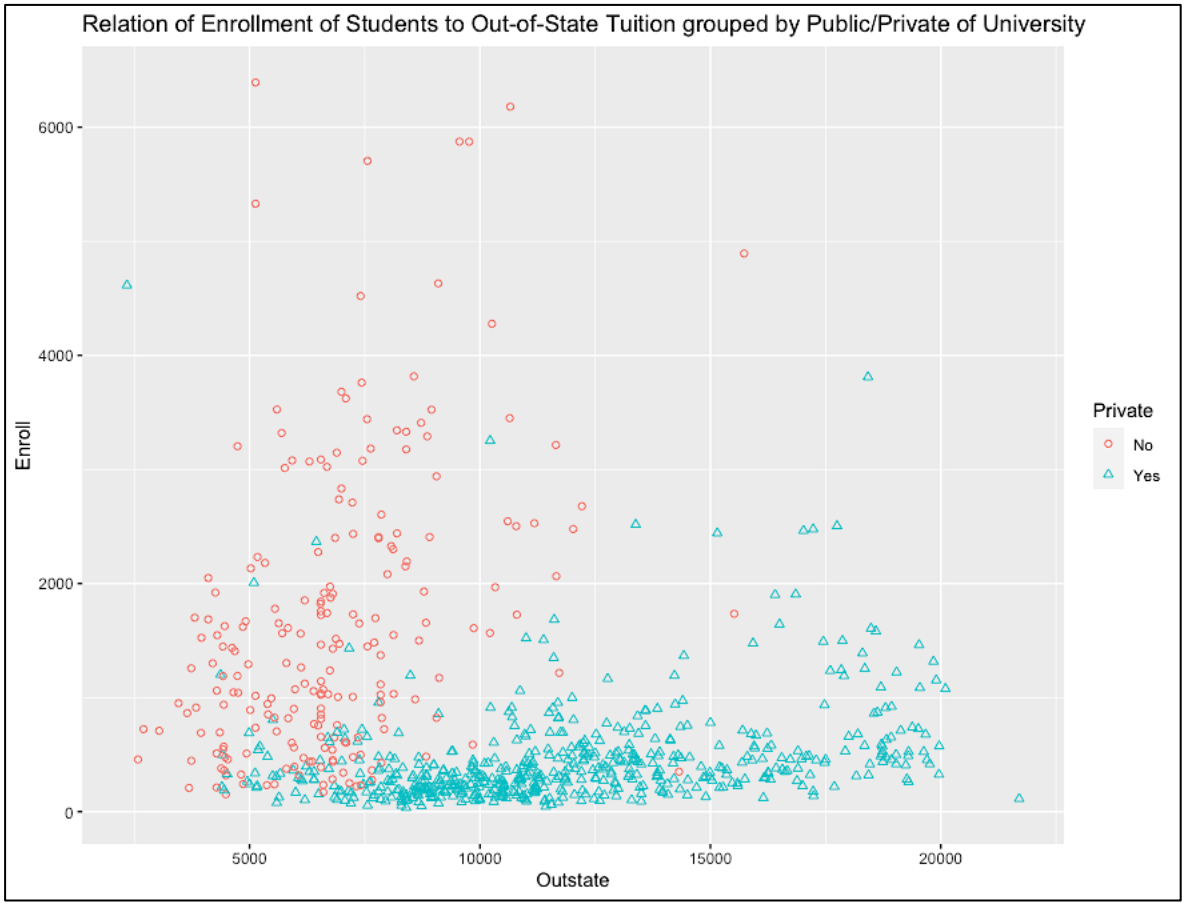

Figure 1.5: Enrolment of Students to Out-of-State Tuition relation grouped by Private/Public University

1. The enrolment of students in the Public University has grown very high as well in the range of 4000 to more than 6000 students per university when the Out-Of-State tuition rates are in the range of $5000 - $11,000.

2. The plot shows that the Private Universities have enrolment rates less in comparison to the enrolment rate in Public Universities when the Out-of-State tuition is less than $10,000.

3. The Out-of-State tuition fees are less for the Public Universities when compared to the Private Universities.

4. The Out-of-State tuition fee can go very high in the range of $20,000 and students enrol in these private universities as well.

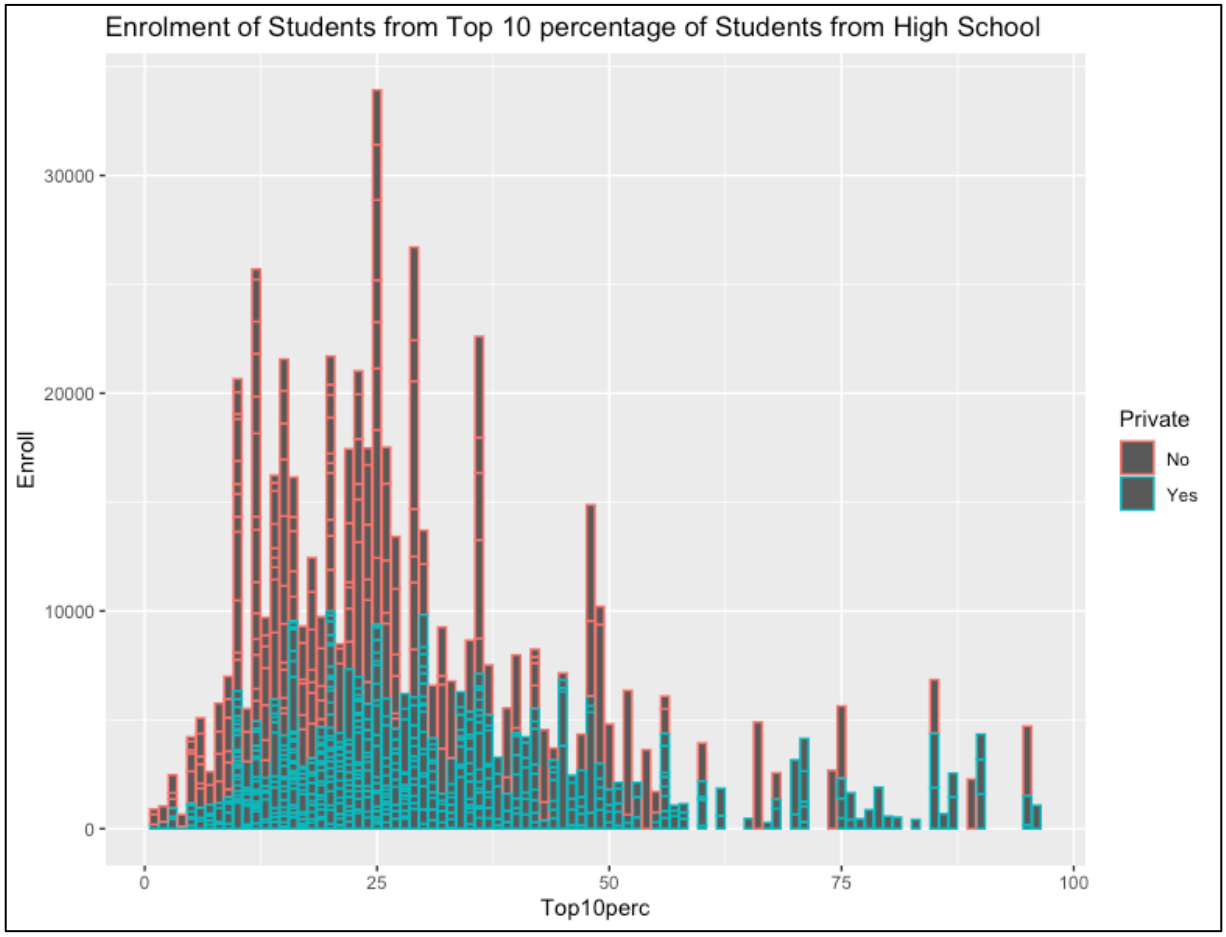

Figure 1.6: Enrolment of Students from Top 10 Percentage of Students from High School grouped by Private/Public University

1. Most of the students form the group of Top 10 percentage of students from high School enrol in Public Universities more significantly than in the Private Universities.

2. The capping of Enrolment of Top 10 percentage of students from high school in Private Universities is around 10,000.

Splitting the Data into Train and Test Sets -

We will create a partition of the data set randomly in the ratio of 70:30 where 70% of the random data will go into the training set and the rest 30% of the data will move into the testing set.

Dependent variable - *Private*

Logistic Regression Model using Generalized Linear Models (GLM) -

Logistic Regression model was fitted to the data set using all the features, except the dependent variable (*Private*), as predictor variables in the model.

Generalized Linear Models (GLM) was used to fit the logistic regression model to the training data set to in order to predict whether a university is private or public.

GLM Logistic Regression Test Coefficients Summary Table

	Estimate	Std..Error	z.value	Pr(>|z|)
(Intercept)	0.6056	2.2722	0.2665	0.7898
Apps	-0.0007	0.0003	-2.5536	0.0107
Accept	0.0001	0.0006	0.2579	0.7965
Enroll	0.0013	0.0010	1.3326	0.1827
Top10perc	0.0012	0.0312	0.0383	0.9695
Top25perc	0.0143	0.0210	0.6799	0.4966
F.Undergrad	-0.0002	0.0002	-1.5625	0.1182
P.Undergrad	-0.0001	0.0002	-0.3292	0.7420
Outstate	0.0006	0.0001	4.9151	0.0000
Room.Board	0.0002	0.0003	0.7597	0.4475
Books	0.0019	0.0017	1.1125	0.2659
Personal	-0.0006	0.0003	-1.8989	0.0576
PhD	-0.0723	0.0312	-2.3165	0.0205
Terminal	-0.0457	0.0297	-1.5386	0.1239
S.F.Ratio	-0.0634	0.0709	-0.8944	0.3711
perc.alumni	0.0410	0.0244	1.6827	0.0924
Expend	0.0003	0.0002	2.0857	0.0370
Grad.Rate	0.0266	0.0149	1.7877	0.0738

Table 1.4: GLM Logistic Regression Test Coefficients Summary (1st Model)

Null deviance: 639.40 on 544 degrees of freedom
Residual deviance: 177.18 on 527 degrees of freedom
AIC: 213.18

Observations -

From the Coefficients summary table, we can figure out that there are several features which are insignificant or almost insignificant to the training data set and model. Only few features deemed to be significant as their Probability-values (P-value) were less than or around 0.05.

These significant features are:

1. *Outstate* - The out of state tuition of the university
2. *Apps* - Number of applications received for the enrolment
3. *Personal* - Estimated Personal Spending
4. *PhD* - Percentage of Faculty with PhDs
5. *Expend* - Instructional expenditure per student in the university
6. *Grad.Rate* - Graduation Rate of students in the university
7. *perc.alumni* - Percentage of alumni who donate

Multi-Collinearity using Variance Inflation Factor (VIF) -

Variance Inflation Factor (VIF) function was used to check multi-collinearity of all the features in the data set.

Variance Inflation Factor Summary Table

	vif(model1)
Apps	18.346727
Accept	31.122754
Enroll	16.135155
Top10perc	4.727865
Top25perc	3.881612
F.Undergrad	7.951507
P.Undergrad	1.682087
Outstate	2.367777
Room.Board	1.905158
Books	1.196735
Personal	1.152363
PhD	4.175153
Terminal	3.802725
S.F.Ratio	1.654409
perc.alumni	1.202351
Expend	3.543402
Grad.Rate	1.484726

Table 1.5: Variance Inflation Factor (VIF) Summary (1st Model)

Observations -

From the table of Variance Inflation Factor (VIF) Summary, we can check that the following features have VIF value greater than or equal to 5 which means that these features are multi-collinear and need not be included in our model for predictions.

These features can be derived using other features and are not significant to the data set as they are considered as derivable variables.

The features with high Variance Inflation Factor (VIF) values are :

1. *Apps* - 18.346727
2. *Accept* - 31.122754
3. *Enroll* - 16.135155
4. *F.Undergrad* - 7.951507

But, the variable (*Apps*) is collinear with Accept, Enroll, F.Undergrad variables and its P-value is also very significant to the regression model (0.0107). Therefore, we can eliminate other multi-collinear features to fit our logistic regression model and instead use this variables (Apps) only to remodel it using GLM.

Another logistic regression model was fitted using the significant features only and after eliminating these derivable variables (those having large VIF value) for a better accuracy and prediction. *But, since the variables (Apps, F.Undergrad) have significant P-value and high variance inflation factor (VIF) value, we will keep only one variable amongst them.*

GLM Logistic Regression Test Summary Table

	Estimate	Std..Error	z.value	Pr(>|z|)
(Intercept)	-0.1806	1.3486	-0.1339	0.8935
Outstate	0.0007	0.0001	6.2509	0.0000
Personal	-0.0006	0.0003	-1.8142	0.0696
PhD	-0.1057	0.0202	-5.2376	0.0000
Expend	0.0004	0.0001	3.4215	0.0006
perc.alumni	0.0464	0.0230	2.0173	0.0437
Grad.Rate	0.0264	0.0136	1.9403	0.0523
Apps	-0.0006	0.0001	-7.2289	0.0000

Table 1.6: GLM Logistic Regression Test Coefficients Summary (2nd Model)

Null deviance: 639.40 on 544 degrees of freedom
Residual deviance: 186.08 on 537 degrees of freedom
AIC: 202.08

Observations -

From the new Coefficients summary table made with the significant features, all of the our features are significant or almost significant to the training data set and model as their Probability-values (P-value) are less than or equal to 0.05.

The ***Akaike Information Criterion (AIC)*** value of the second model - **202.08** (with the significant features only) is also smaller than the Akaike Information Criterion (AIC) value of the first model - **213.18** (with all the features). It suggests that we should use this second logistic regression model for our prediction.

Multi-Collinearity using Variance Inflation Factor (VIF) -

Variance Inflation Factor (VIF) function was used to check multi-collinearity of the chosen significant features in the data set for our second model.

Variance Inflation Factor Summary Table

	vif(model2)
Outstate	1.874432
Personal	1.035437
PhD	1.827159
Expend	2.294169
perc.alumni	1.105617
Grad.Rate	1.267659
Apps	1.616078

Table 1.7: Variance Inflation Factor (VIF) Summary (1st Model)

Observations -

From the table of Variance Inflation Factor (VIF) Summary, we can check that all the features (*significant chosen using P-value and VIF value considerations*) have low VIF value. So, they do not have multi-collinearity and can be used for modelling purposes.

Log-Odds and Odds Conversion

The coefficients of the features in summary of the model are in **Log Odds.**
We converted these coefficients from Log Odds to Odds.

Logistic Regression Model Odds Table

Row Names	Log Odds	Odds
(Intercept)	-0.1806247340	0.8347486
Apps	-0.0005788275	0.9994213
Expend	0.0003981471	1.0003982
Grad.Rate	0.0263569606	1.0267074
Outstate	0.0006664267	1.0006666
perc.alumni	0.0463721295	1.0474641
Personal	-0.0005537657	0.9994464
PhD	-0.1056579989	0.8997323

Table 1.8: Log-Odds and Odds (converted) Summary Table

Predicting the Training Data Set using Logistic Regression Model -

To check the accuracy/underfitting nature of the logistic regression model, we will apply the model to predict the training data set from which the model has been fitted for logistic regression. If the accuracy of the prediction is high, it means there is no underfitting of the model and it suggests that the model is ready to use for the testing data set. But, if the accuracy of the model is not high, it shows that the model is underfitting and not ready to use.

Confusion Matrix -

Training Data Confusion Matrix Table

	Actual (Reference)	
Prediction	No	Yes
No	134	14
Yes	15	382

Table 1.9: Training Data Confusion Matrix Table

Observations -

The Confusion Matrix for the training data set shows that

1. True Negative - **134**
2. False Negative - **14 (*Type II Error*)**
3. True Positive - **382**
4. False Positive - **15 (*Type I Error*)**

The True Negative value means that the predicted value were same with their actual values in case of "No". The True Positive value means that the predicted value were same with their actual values in case of "Yes".

The False Negative value means that the predicted value were not the same with their actual values in case of "No" and they are considered as *Type II Error*. The False Positive value means that the predicted value were not the same with their actual values in case of "Yes" and they are considered as *Type I Error*.

To talk more about the confusion matrix, we are a table of metrics involved in it.

Training Data Confusion Matrix Metrics Table

	Metric Value
Sensitivity	0.9646465
Specificity	0.8993289
Pos Pred Value	0.9622166
Neg Pred Value	0.9054054
Precision	0.9622166
Recall	0.9646465
F1	0.9634300
Prevalence	0.7266055
Detection Rate	0.7009174
Detection Prevalence	0.7284404
Balanced Accuracy	0.9319877

Table 1.10: Training Data Confusion Matrix Metrics Table

Observations -

The **Balanced Accuracy** of the logistic regression model on our training data set is around **93%** which seems to be good for the model. It shows that the model is not underfitting and can be used to predict the values in the testing data set.

Precision at around **96%** is also high in this case as it the number of correctly identified positive results divided by the number of all positive results, including those not identified correctly. Therefore, it is good so far for our model.

Recall at around **96.5%** is also high in our model as it is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive. This also adds significance to the credibility of our model.

Specificity at around **90%** is also high in our model as it is the number of correctly identified negative results divided by the number of all samples that should have been identified as negative. This also adds significance to the credibility of our model.

False Negative Or False Positive (Which One?) -

According to me, the damaging nature of Type I Error and Type II Error depends on the problem and its use case.

In our example, we are predicting whether the university is a Private University or a Public University.

If we consider the use case that public universities are government funded and they are less expensive than private universities, we can say that in our problem, miscalculations in Type II Error is more damaging to the model and its audience than Type I Error.

Predicting the Testing Data Set using Logistic Regression Model -

To check the robustness and overfitting nature of the logistic regression model, we will apply the model to predict the testing data set for logistic regression. If the accuracy of the prediction is high, it means there is no overfitting of the model and it suggests that the model's results are correct. But, if the accuracy of the model is not high, it shows that the model is overfitting and not showing correct results.

Confusion Matrix -

Testing Data Confusion Matrix Table

	Actual (Reference)	
Prediction	No	Yes
No	56	11
Yes	7	158

Table 1.11: Testing Data Confusion Matrix Table

Observations -

The Confusion Matrix for the testing data set shows that

1. True Negative - **56**
2. False Negative - **11 (*Type II Error*)**
3. True Positive - **158**
4. False Positive - **7 (*Type I Error*)**

The True Negative value means that the predicted value were same with their actual values in case of "No". The True Positive value means that the predicted value were same with their actual values in case of "Yes".

The False Negative value means that the predicted value were not the same with their actual values in case of "No" and they are considered as *Type II Error*. The False Positive value means that the predicted value were not the same with their actual values in case of "Yes" and they are considered as *Type I Error*.

To talk more about the confusion matrix, we are a table of metrics involved in it.

Testing Data Confusion Matrix Table

	Metric Value
Sensitivity	0.9349112
Specificity	0.8888889
Pos Pred Value	0.9575758
Neg Pred Value	0.8358209
Precision	0.9575758
Recall	0.9349112
F1	0.9461078
Prevalence	0.7284483
Detection Rate	0.6810345
Detection Prevalence	0.7112069
Balanced Accuracy	0.9119001

Table 1.12: Testing Data Confusion Matrix Metrics Table

Observations -

The **Balanced Accuracy** of the logistic regression model on our testing data set is around **91%** which seems to be good for the model. It shows that the model is not overfitting and showing correct results throughout the data.

Precision at around **96%** is also high in this case as it the number of correctly identified positive results divided by the number of all positive results, including those not identified correctly. Therefore, it is good so far for our model.

Recall at around **93.5%** is also high in our model as it is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive. This also adds significance to the credibility of our model.

Specificity at around **89%** is also high in our model as it is the number of correctly identified negative results divided by the number of all samples that should have been identified as negative. This also adds significance to the credibility of our model.

Receiver Operating Characteristic (ROC) Curve -

A ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR). The true positive rate is the proportion of observations that were correctly predicted to be positive out of all positive observations (TP/(TP + FN)). Similarly, the false positive rate is the proportion of observations that are incorrectly predicted to be positive out of all negative observations (FP/(TN + FP)).

The ROC curve shows the trade-off between **sensitivity** (or TPR) and **specificity** (1 – FPR). Classifiers that give curves closer to the top-left corner indicate a better performance. The plot of the ROC curve is presented below-

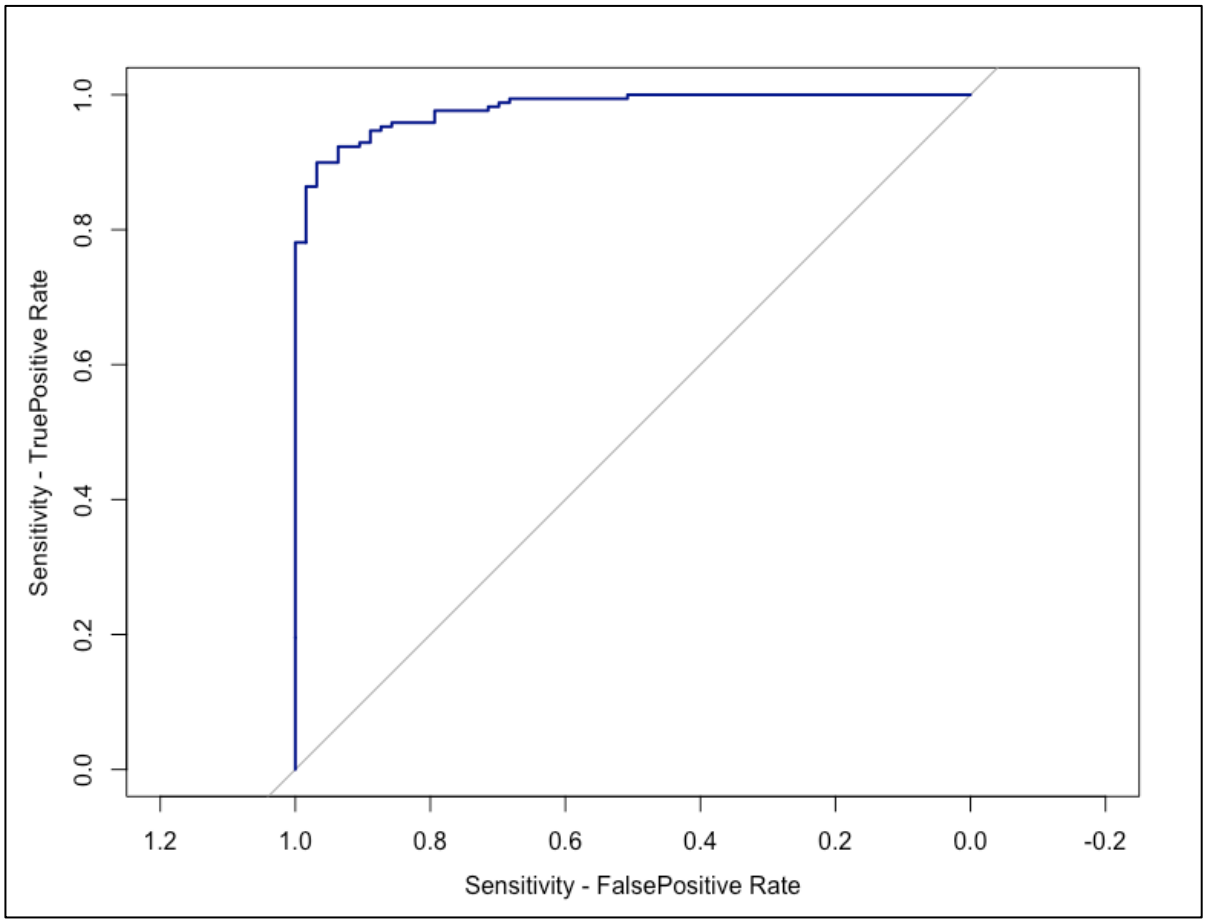

Figure 1.6: ROC Curve

Observations -

The **ROC Curve** is almost closer to the top-left corner and not to the 45-degree diagonal. This signifies that the model has shown better performance and the tests are almost accurate.

Area Under ROC Curve (AUC) -

Area under the ROC Curve - **0.98**

Higher the AUC value, the better the model is at predicting 'Yes' classes as 'Yes', and 'No' classes as 'No'. This value shows that the model has a good measure of separability.

CONCLUSION

We have conducted/performed the *Logistic Regression Modelling using Generalized Linear Model (GLM)* on the College data set to predict whether an university is a Private University or a Public University.

We split our College data set into a training and a testing data set in the ratio of 70-30. Randomly, 70% of data points from the data set were assigned to the training data set, and similarly, the rest 30% of the data points were assigned to the testing data set. We used this training data set to fit our logistic regression model created using Generalized Linear Model (GLM) and used this model on this training data set first to check its accuracy or the underfitting nature.

Upon predicting the training data set, we found an accuracy of about 93% with all the metrics (Precision, Recall, Specificity) in the range of 93-96% which shows that the model has not underfit and can be used to predict the testing data.

The testing data was predicted using this logistic regression model and provided an accuracy of about 91% along with other metrics high in value and significance (Precision, Recall, Specificity). This shows that the model has not overfit and is showing correct results.

The Type I Error and Type II Errors are also very low and in our case, it shows that a low Type II Error is significant to our problem since it would make a Public University be considered as a Private University.

We can conclude that the prediction of an university to be either a Private University or Public University has been conducted successfully and correctly and we were able to achieve an accuracy of about 91% in our testing data set. The Type I Error and Type II Error are also very low and our logistic regression model is also correct.

BIBLIOGRAPHY

1. *Home - RDocumentation*. (2021). Functions in R - Documentation. https://www.rdocumentation.org/

2. ALY 6015 - Prof Roy Wada - *Lesson 3-1 — Generalized Linear Models (GLM)* (2022, March), https://northeastern.instructure.com/courses/98028/pages/lesson-3-1-generalized-linear-models-glm?module_item_id=6646988

3. ALY 6015 - Prof Roy Wada - *Lesson 3-2 — Logistic Regression* (2022, March), https://northeastern.instructure.com/courses/98028/pages/lesson-3-2-logistic-regression?module_item_id=6646993

4. ALY 6015 - Prof Roy Wada - *GLM and Logistic Regression — Pre Assignment Lab* (2022, March), https://northeastern.instructure.com/courses/98028/assignments/1207975?module_item_id=6647007

5. *What is Logistic regression? | IBM*. (2021). IBM. https://www.ibm.com/topics/logistic-regression

6. Shivaprasad, P. (2021, December 16). *Understanding Confusion Matrix, Precision-Recall, and F1-Score.* Medium. https://towardsdatascience.com/understanding-confusion-matrix-precision-recall-and-f1-score-8061c9270011

APPENDIX

```
#-------- ALY6015_M3_GLM&LogisticRegression_HarshitGaur --------#

print("Author : Harshit Gaur")
print("ALY 6015 Week 3 Assignment - GLM and Logistic Regression")

# Declaring the names of packages to be imported
packageList <- c("tidyverse", "ISLR", "caret", "pROC", "car",
                 "RColorBrewer", "psych", "flextable")

for (package in packageList) {
  if (!package %in% rownames(installed.packages()))
  { install.packages(package) }

  # Import the package
  library(package, character.only = TRUE)
}


######################################################################
# College Dataset
######################################################################

# Import/Attach the data set
attach(College)

collegeDatasetHead <- College %>%
  select(Private, Apps, Accept, Enroll, Top10perc, Top25perc, F.Undergrad, P.Undergrad, Outstate)
save_as_docx('College Dataset' = flextable(data = cbind(rownames(head(collegeDatasetHead)),
head(collegeDatasetHead))),
          path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class
3/Assignment/Tables/College_Data_Table.docx')

# Get the glimpse of data set
glimpse(College)

describeCollegeFlex <- College %>%
  psych::describe(quant = c(.25, .75), IQR = TRUE) %>%
  select(n, mean, sd, median, min, max, range, skew, kurtosis)

describeCollegeFlex <- round(describeCollegeFlex, 2)
describeCollegeFlex <- cbind(e = rownames(describeCollegeFlex), describeCollegeFlex)

save_as_docx('Descriptive Statistics of College Dataset' = flextable(data = describeCollegeFlex),
          path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class
3/Assignment/Tables/College_Desc_Stats_Table_main.docx')

# Scatterplots
qplot(x = Top10perc, y = Top25perc, color = Private, shape = Private, geom = "point", main = "Relation of Top
10 and Top 25 Percentage grouped by Privatisation of University") + scale_shape(solid = FALSE)
qplot(x = Outstate, y = Enroll, color = Private, shape = Private, geom = "point", main = "Relation of Enrollment
of Students to Out-of-State Tuition grouped by Public/Private of University") + scale_shape(solid = FALSE)

College %>% ggplot(aes(y = Enroll, x = Top10perc, color = Private)) +
  geom_bar(stat = "identity") +
  labs(title = "Enrolment of Students from Top 10 percentage of Students from High School")

College %>% ggplot(aes(x = Private, y = Outstate, fill = Grad.Rate)) +
  geom_bar(stat = "identity")
```

```
# Normality Check for features using Q-Q Plot and Shapiro-Wilks Test.
qqPlot(College$Top10perc, ylab = "Studentized Residuals", xlab = "Theoretical Quantiles")
shapiro.test(College$Top10perc)

qqPlot(College$Top25perc, ylab = "Studentized Residuals", xlab = "Theoretical Quantiles")
shapiro.test(College$Top25perc)

#####################################################################
# Split data into train and test data
#####################################################################
set.seed(454)
trainIndex <- createDataPartition(College$Private, p = 0.70, list = FALSE)
train <- College[trainIndex,]
test <- College[-trainIndex,]


#####################################################################
# Fit a Logistic Regression Model
#####################################################################
model1 <- glm(Private ~ ., data = train, family = binomial(link = "logit"))
model.summary <- summary(model1)
model.summary

# Save 3-Line Table
df <- data.frame(unclass(round(model.summary$coefficients, 4)), stringsAsFactors = FALSE, check.rows =
TRUE)
save_as_docx('GLM Logistic Regression Test Summary Table' = flextable(data = cbind(trimws(rownames(df)),
df)),
        path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class
3/Assignment/Tables/GLM1_Summary.docx')

# Check for Multi-Collinearity
vif_model1 <- as.data.frame(vif(model1))
vif_model1
save_as_docx('Variance Inflation Factor Summary Table' = flextable(data = cbind(rownames(vif_model1),
vif_model1)),
        path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class
3/Assignment/Tables/GLM1_VIF_Summary.docx')


# Removing Insignificant Variables
model2 <- glm(Private ~ Outstate + Personal + PhD + Expend + perc.alumni +
          Grad.Rate + Apps, data = train, family = binomial(link = "logit"))
model.summary <- summary(model2)
model.summary

# Save 3-Line Table
df <- data.frame(unclass(round(model.summary$coefficients, 4)), stringsAsFactors = FALSE, check.rows =
TRUE)
save_as_docx('GLM Logistic Regression Test Summary Table' = flextable(data = cbind(trimws(rownames(df)),
df)),
        path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class
3/Assignment/Tables/GLM2_Summary.docx')

# Check for Multi-Collinearity
vif_model2 <- as.data.frame(vif(model2))

save_as_docx('Variance Inflation Factor Summary Table' = flextable(data = cbind(rownames(vif_model2),
vif_model2)),
        path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class
3/Assignment/Tables/GLM2_VIF_Summary.docx')
```

```
# Display Regression Coefficients (Log Odds)
coef(model2)

# Display Regression Coefficients (Odds)
exp(coef(model2))

# Create a data frame of Log-Odds and Odds
odds_df <- merge(as.data.frame(coef(model2)), as.data.frame(exp(coef(model2))), by = "row.names")
colnames(odds_df) <- c("Row Names", "Log Odds", "Odds")

# Save 3-Line Table
save_as_docx('Logistic Regression Model Odds Table' = flextable(data = odds_df),
        path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class
3/Assignment/Tables/GLM_Odds_Table.docx')



#####################################################################
# Train set predictions
#####################################################################
probabilities.train <- predict(model2, newdata = train, type = "response")
predicted.classes.train <- as.factor(ifelse(probabilities.train >= 0.5, "Yes", "No"))

# Confusion Matrix for Train Set
train_ConfusionMatrix <- confusionMatrix(predicted.classes.train, train$Private, mode = "everything", positive
= "Yes")

# Save 3-Line Table
save_as_docx('Training Data Confusion Matrix Table' = flextable(data =
as.data.frame.matrix(train_ConfusionMatrix$table)),
        path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class
3/Assignment/Tables/Train_CM_Table.docx')

train_CM_Metrics <- as.data.frame(train_ConfusionMatrix$byClass)
# Save 3-Line Table
save_as_docx('Training Data Confusion Matrix Table' = flextable(data = cbind(rownames(train_CM_Metrics),
train_CM_Metrics)),
        path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class
3/Assignment/Tables/Train_CM_Metrics_Table.docx')


#####################################################################
# Test set predictions
#####################################################################
probabilities.test <- predict(model2, newdata = test, type = "response")
predicted.classes.test <- as.factor(ifelse(probabilities.test >= 0.5, "Yes", "No"))

# Confusion Matrix for Test Set
test_ConfusionMatrix <- confusionMatrix(predicted.classes.test, test$Private, mode = "everything", positive =
"Yes")

# Save 3-Line Table
save_as_docx('Testing Data Confusion Matrix Table' = flextable(data =
as.data.frame.matrix(test_ConfusionMatrix$table)),
        path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class
3/Assignment/Tables/Test_CM_Table.docx')

test_CM_Metrics <- as.data.frame(test_ConfusionMatrix$byClass)
# Save 3-Line Table
save_as_docx('Testing Data Confusion Matrix Table' = flextable(data = cbind(rownames(test_CM_Metrics),
test_CM_Metrics)),
```

```
        path = 'Documents/Northeastern University/MPS Analytics/ALY 6015/Class
3/Assignment/Tables/Test_CM_Metrics_Table.docx')


######################################################################
# ROC and AUC Curve
######################################################################

# Plot the Receiver Operating Characteristic Curve
ROC <- roc(test$Private, probabilities.test)
plot(ROC, col = "DARKBLUE", ylab = "Sensitivity - TruePositive Rate", xlab = "Sensitivity - FalsePositive Rate")

# Plot the Area Under the ROC Curve
auc <- auc(ROC)
auc

#-------- END --------#
```