Northeastern

MPS Analytics

ALY6015 : Intermediate Analytics

Module 3 Assignment

GLM and Logistic Regression

Prepared By : Shyamala Venkatakrishnan

Date: 01/28/2023

Introduction

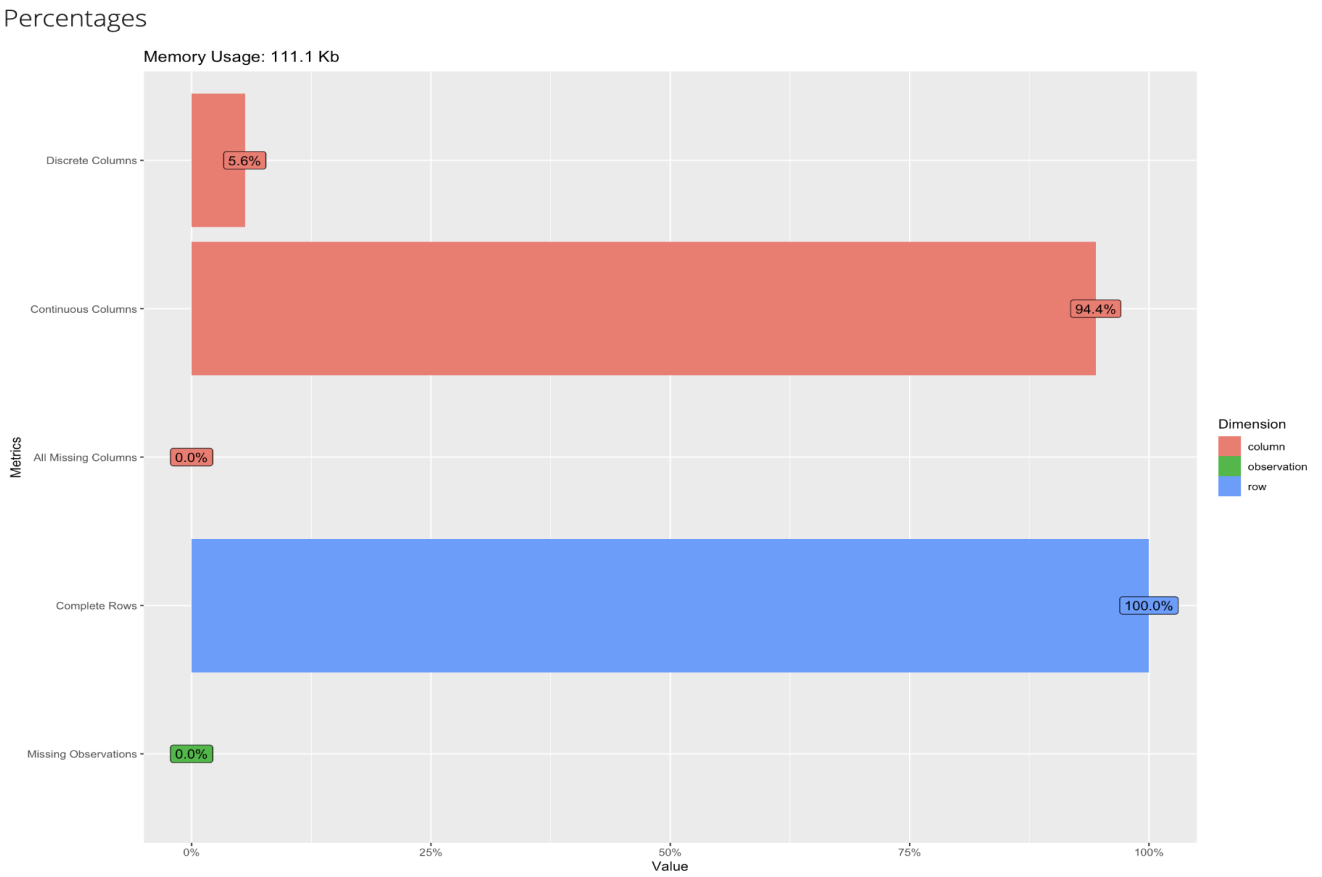
As part of this assignment, a dataset related to the details of the US Colleges from the 1995 issue of US News and World Report is shared and EDA and descriptive summary statistics are to be generated to find some useful insights about the college data. The selection of the college can be very crucial for the students based on many factors like private/public university, graduation rate, college expenses, faculties holding a PhD/Terminal degree. Both private and public universities have their own set of pros and cons and the analysis of this dataset can be useful in understanding about private/public university colleges from various perspectives. Generalized linear model (GLM) functions are to be used with logistic regression to build a model to predict whether a university is private or public. A confusion matrix is to be generated to study the accuracy of the model, and other metrics like recall, precision, specificity, sensitivity are to be interpreted from the results. An ROC curve is to be plotted and AUC should be calculated to evaluate the overall effectiveness of the model.

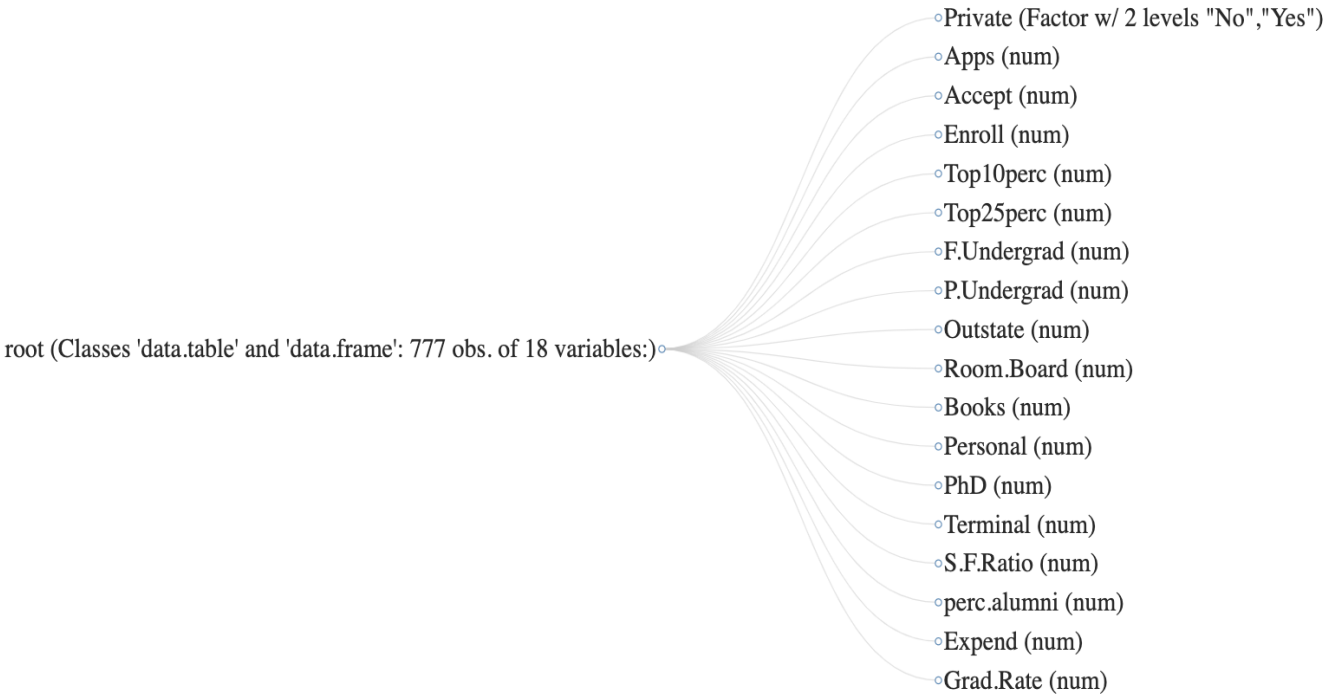
Understanding the dataset

- This dataset has a total of 777 observations (details about 777 colleges) and 18 attributes, each of which describes various variables related to colleges.
- It has 17 continuous columns and 1 discrete column.
- The meaning of these columns can be referred to in the data dictionary available in this link.

Raw Counts

Name	Value
Rows	777
Columns	18
Discrete columns	1
Continuous columns	17
All missing columns	0
Missing observations	0
Complete Rows	777
Total observations	13,986
Memory allocation	111.1 Kb





- The above graph displays the columns present in this dataset. Private variable is a categorical variable which has 2 factors ‘Yes’, ‘No’ and has a value of ‘Yes’ if a college belongs to Private university and a value of ‘No’ if a college belongs to Public university.

- There is no missing data or observations in this dataset.

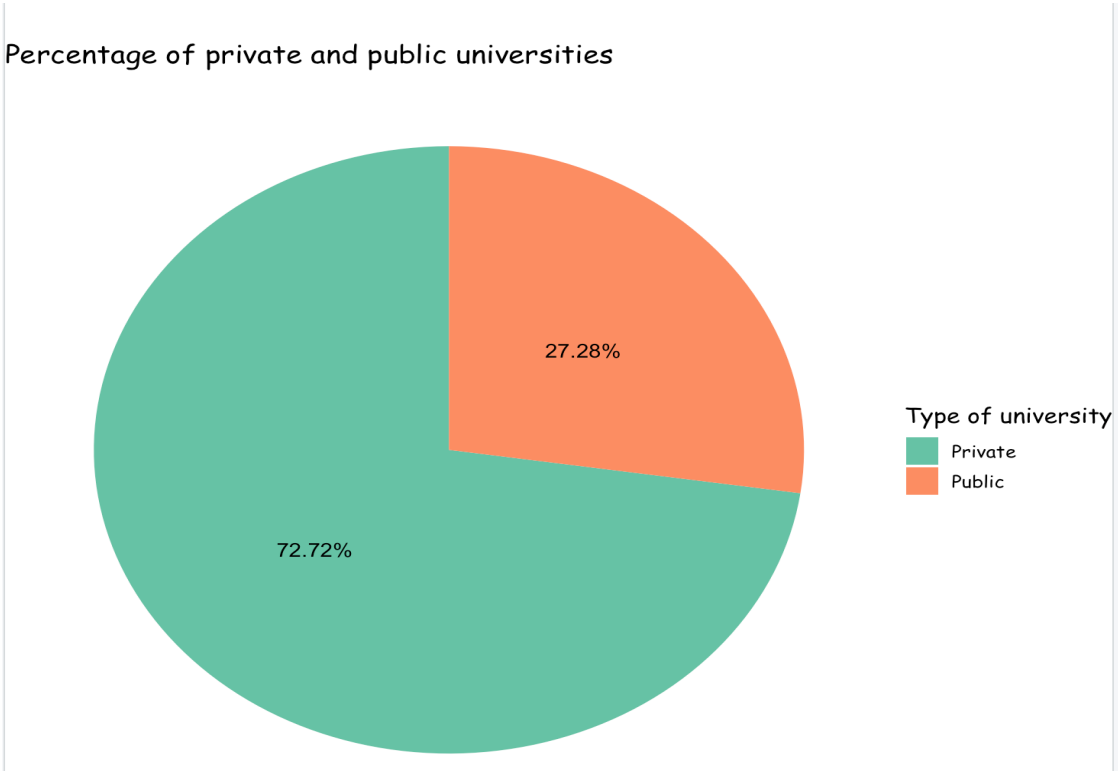
Descriptive summary statistics of the dataset:

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Private*	1	777	1.73	0.45	2.0	1.78	0.00	1.0	2.0	1.0	-1.02	-0.96	0.02
Apps	2	777	3001.64	3870.20	1558.0	2193.01	1463.33	81.0	48094.0	48013.0	3.71	26.52	138.84
Accept	3	777	2018.80	2451.11	1110.0	1510.29	1008.17	72.0	26330.0	26258.0	3.40	18.75	87.93
Enroll	4	777	779.97	929.18	434.0	575.95	354.34	35.0	6392.0	6357.0	2.68	8.74	33.33
Top10perc	5	777	27.56	17.64	23.0	25.13	13.34	1.0	96.0	95.0	1.41	2.17	0.63
Top25perc	6	777	55.80	19.80	54.0	55.12	20.76	9.0	100.0	91.0	0.26	-0.57	0.71
F.Undergrad	7	777	3699.91	4850.42	1707.0	2574.88	1441.09	139.0	31643.0	31504.0	2.60	7.61	174.01
P.Undergrad	8	777	855.30	1522.43	353.0	536.36	449.23	1.0	21836.0	21835.0	5.67	54.52	54.62
Outstate	9	777	10440.67	4023.02	9990.0	10181.66	4121.63	2340.0	21700.0	19360.0	0.51	-0.43	144.32
Room.Board	10	777	4357.53	1096.70	4200.0	4301.70	1005.20	1780.0	8124.0	6344.0	0.48	-0.20	39.34
Books	11	777	549.38	165.11	500.0	535.22	148.26	96.0	2340.0	2244.0	3.47	28.06	5.92
Personal	12	777	1340.64	677.07	1200.0	1268.35	593.04	250.0	6800.0	6550.0	1.74	7.04	24.29
PhD	13	777	72.66	16.33	75.0	73.92	17.79	8.0	103.0	95.0	-0.77	0.54	0.59
Terminal	14	777	79.70	14.72	82.0	81.10	14.83	24.0	100.0	76.0	-0.81	0.22	0.53
S.F.Ratio	15	777	14.09	3.96	13.6	13.94	3.41	2.5	39.8	37.3	0.66	2.52	0.14
perc.alumni	16	777	22.74	12.39	21.0	21.86	13.34	0.0	64.0	64.0	0.60	-0.11	0.44
Expend	17	777	9660.17	5221.77	8377.0	8823.70	2730.95	3186.0	56233.0	53047.0	3.45	18.59	187.33
Grad.Rate	18	777	65.46	17.18	65.0	65.60	17.79	10.0	118.0	108.0	-0.11	-0.22	0.62

- The average number of applications received for undergraduate courses in the listed 777 colleges is around 3000.
- The mean number of applications accepted by the universities is around 2018.
- The average enrollment rate in these universities in general is quite less than the number of applications received and around 779. This can be attributed to the fact that only the students who pass the admission criteria and those who demonstrate good knowledge through their high school projects,marks, and other achievements stand a chance in securing an admission in any college.
- The average graduation rate is around 65%. More than half of the total strength of the students have a good chance of successfully graduating.
- The average number of full time undergraduates is greater than the average number of part time undergraduates.
- All the universities have an average of 72 faculties holding a PhD degree and 80 faculties holding a terminal degree which marks the highest academic degree.
- The average student to faculty ratio is 14 which means that for every 14 students, 1 faculty is allotted in these universities.
- The average out of state tuition fee is around 10440 dollars in these universities. This value can generally be lesser in public universities than the private universities.

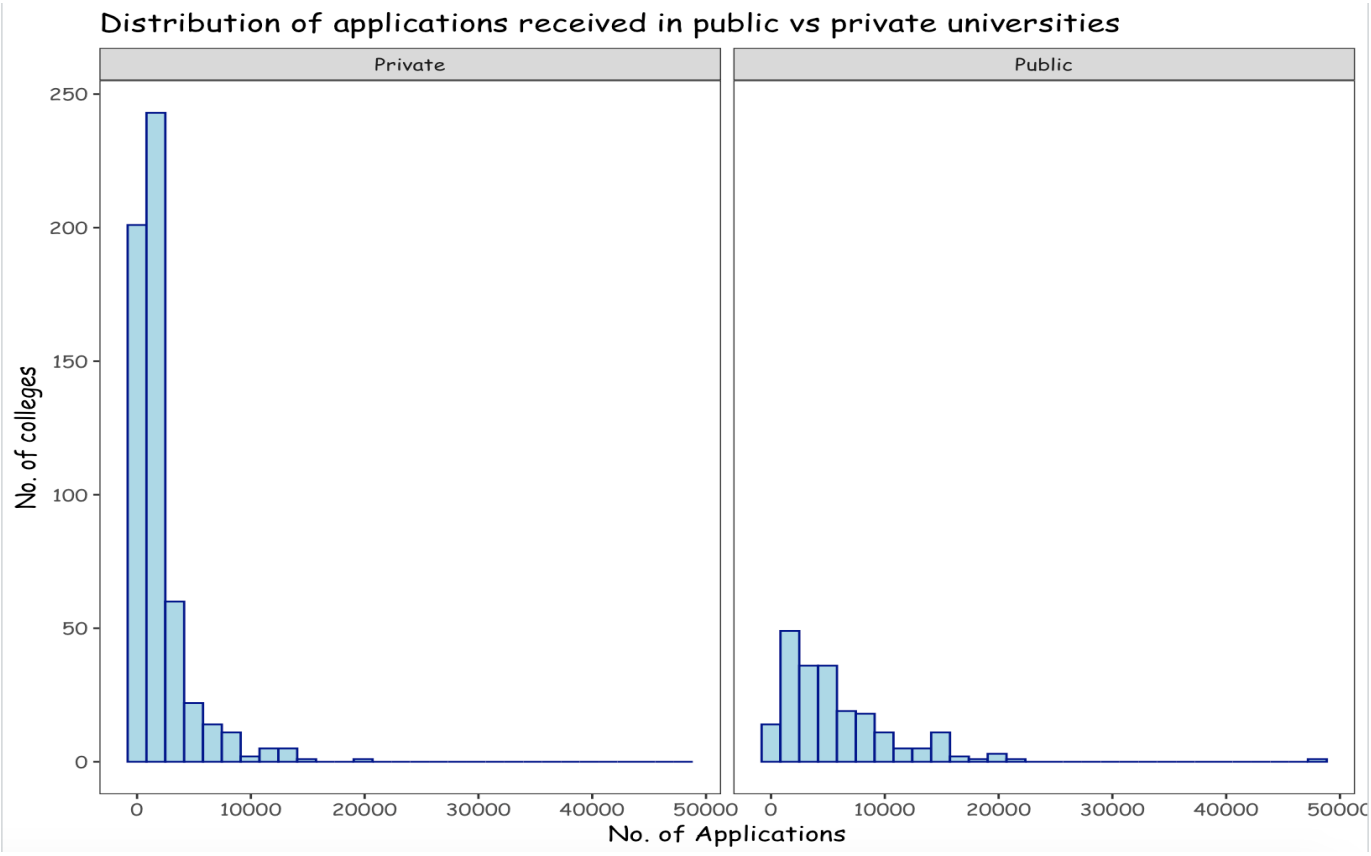
Data Exploration

1. What is the percentage of private and public universities in this dataset?



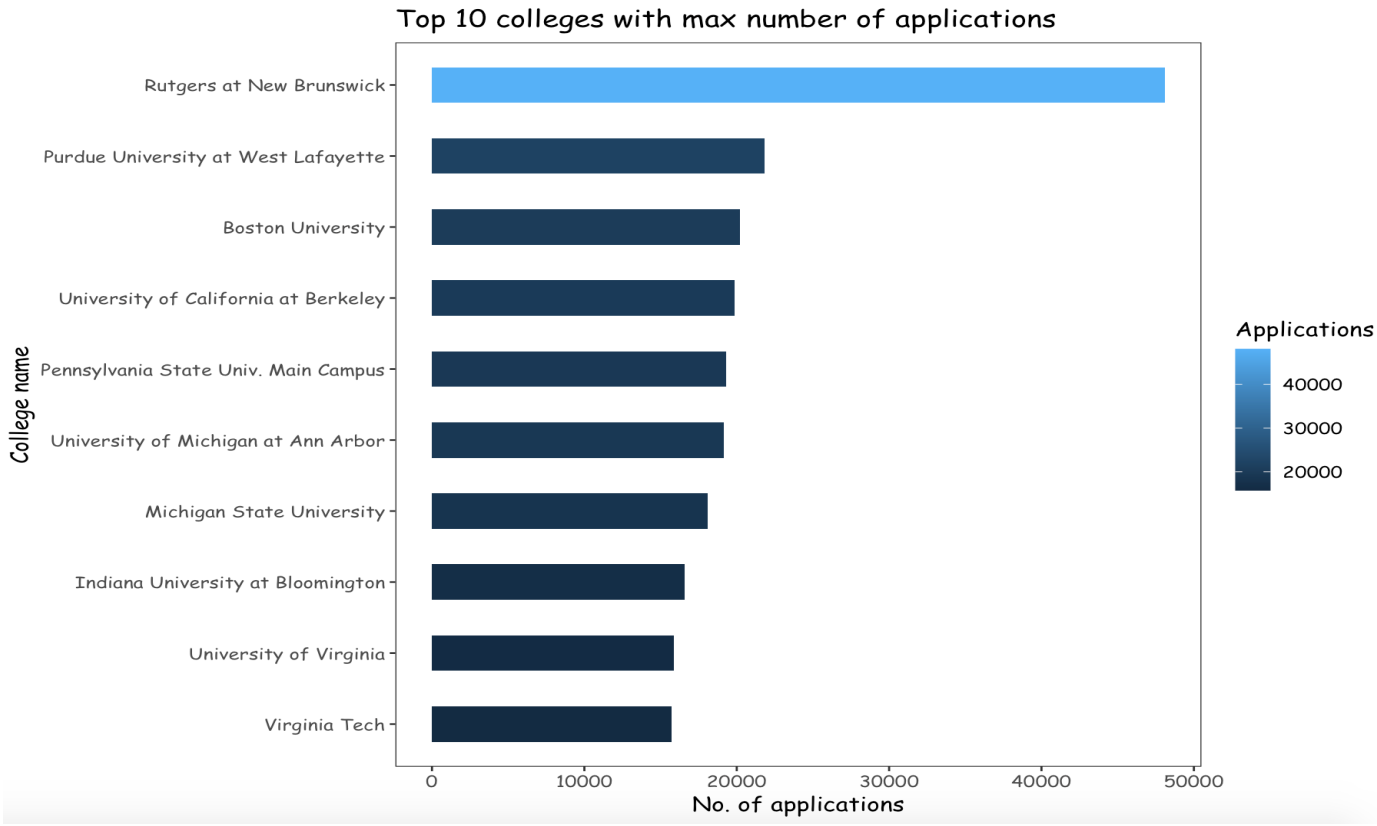
From the above chart, it can be observed that around 73% of the universities in this dataset are private and 27% of the universities are public.

2. How many applications are received in the public and private universities?



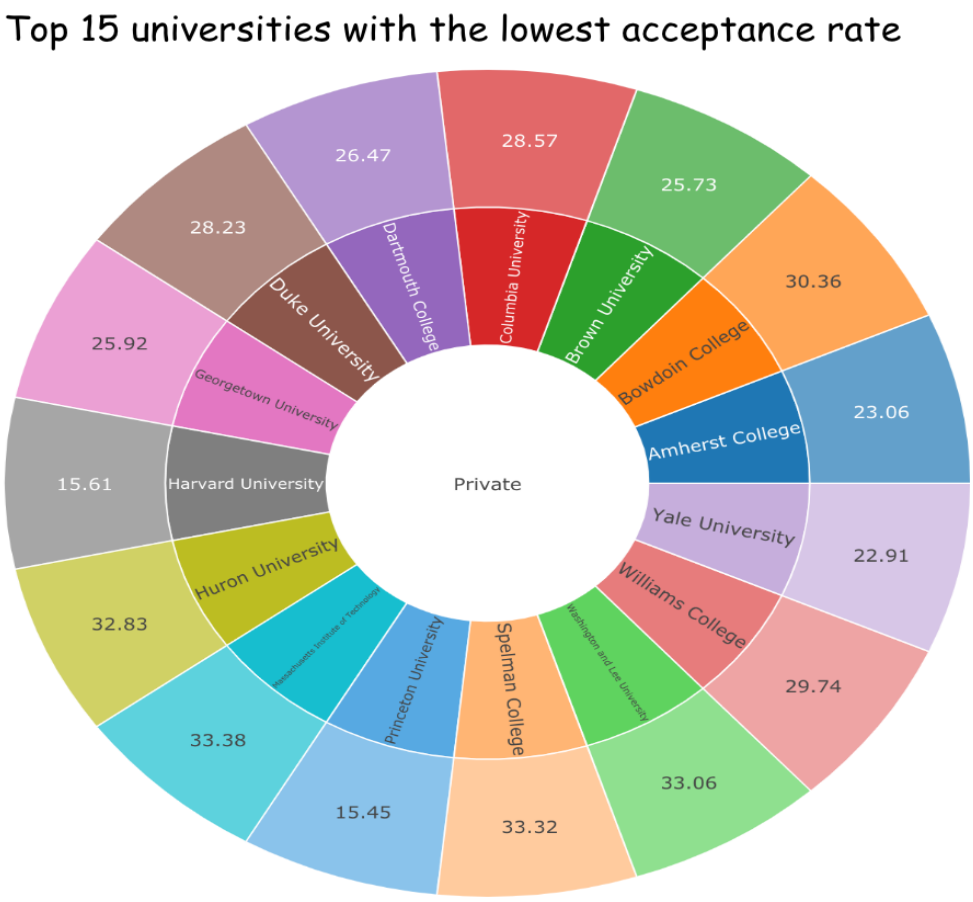
The distribution of the number of admissions in both public and private universities is right skewed. Both universities receive less than 10000 applications and more private universities receive applications than the public universities in the range of 0 - 10000. There is an outlier value in a public university receiving around 50000 applications.

3. What are the top 10 colleges with the maximum number of applications received?



These are the top 10 colleges which received the maximum number of applications, as per this dataset.

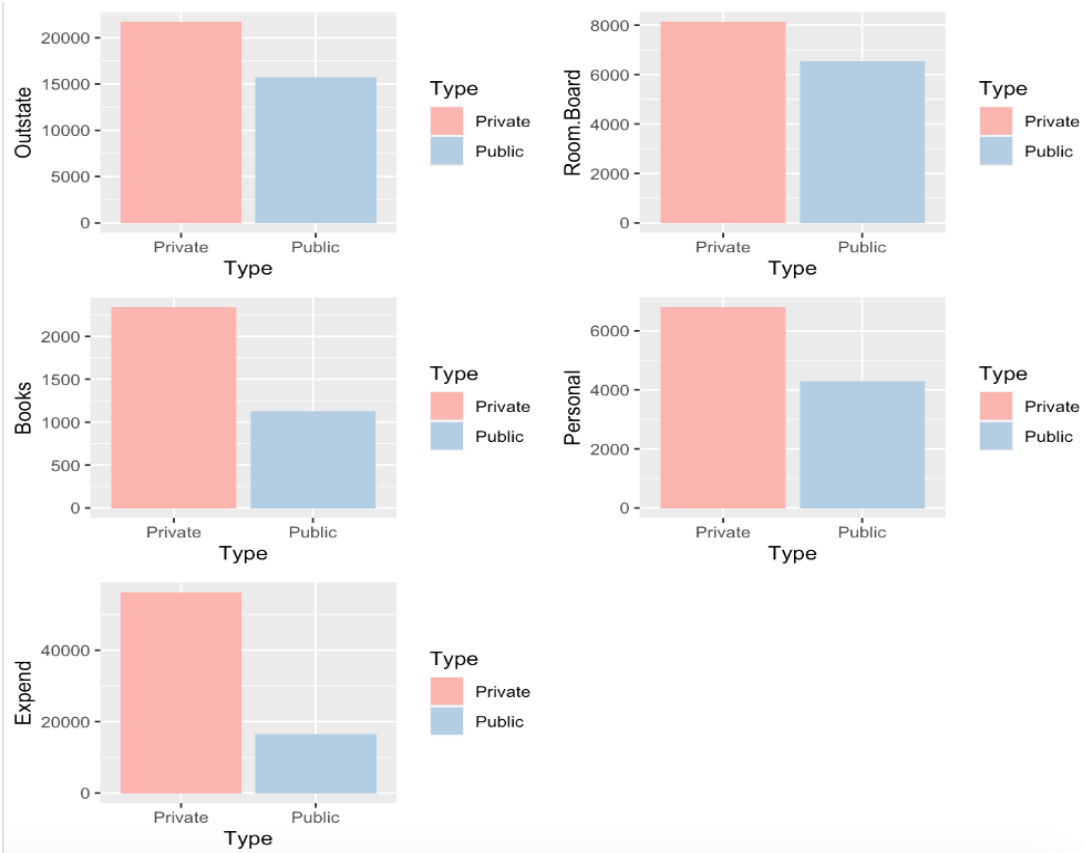
4. Which are the top 15 colleges with the least acceptance rate?



The above chart displays the top 15 universities which have the lowest acceptance rates and all of these universities are private. There are not any public universities listed in this chart which confirms the fact that it is quite difficult to get an admission in private universities than the public universities. Also there is a heavy competition between the students to secure an admission in private universities as top 10 percent and 25 percent of students with very good academic records compete every year.

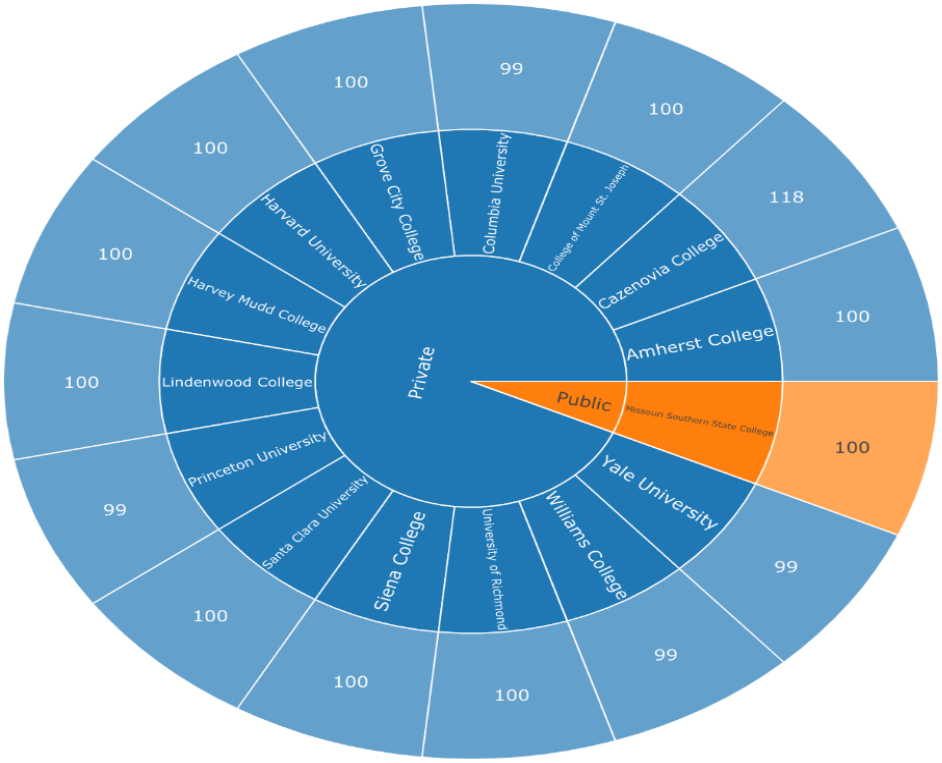
5. Which type of universities are considered costlier with respect to various charges - Public or Private?

Which type of universities are costlier with respect to various charges?



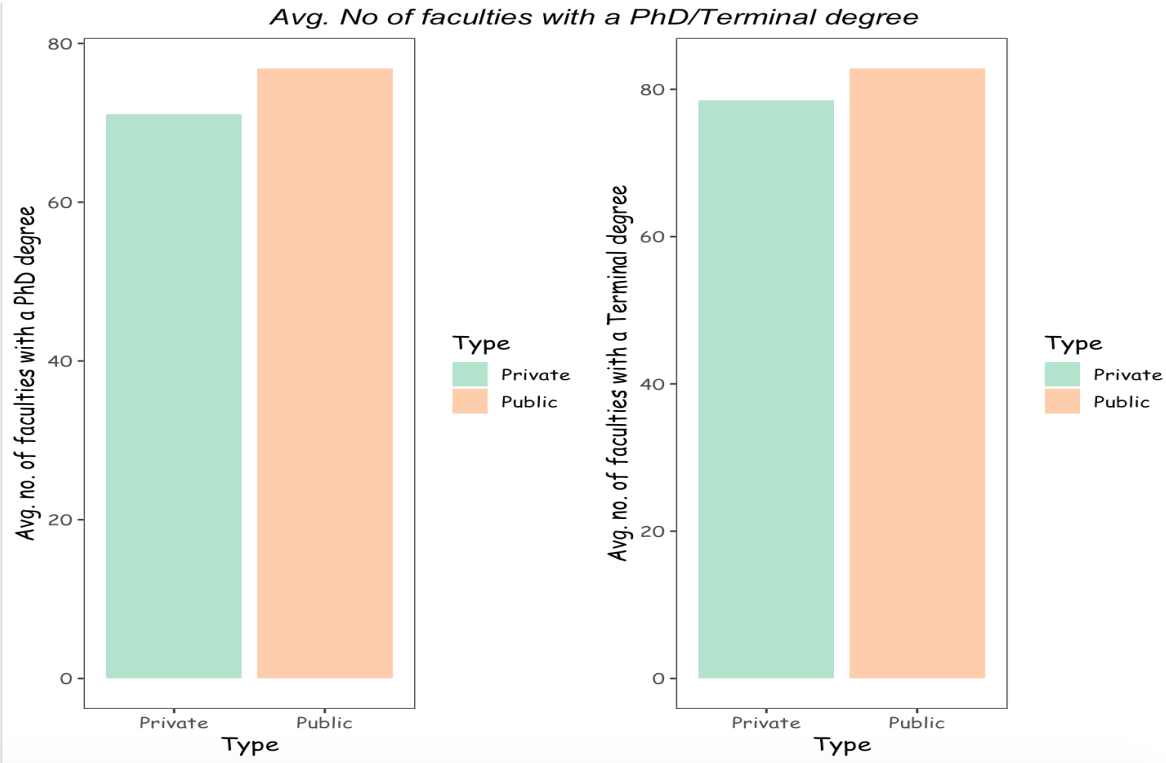
From the above graph, it can be inferred that private universities are quite expensive than the public universities with respect to all types of charges like Out of state tuition, Room and board costs, estimated book costs, overall expenses for personal spending. However they also offer good scholarships or financial aid to students with a better profile than the public universities. For example, Harvard university (Private university) is quite expensive as all the college related costs - from tuition fee to miscellaneous expenses is around 73800 USD, but they can also cover all the financial needs of the students who demonstrate great academic track record and highly talented students.

6. Which are the top 15 universities with the highest graduation rates?



Four-year graduation rate equals 4-year graduates divided by the adjusted cohort at the end of 4 years. Graduation rates are the calculated percentages of students who graduate or complete their program within a specified timeframe. A strong graduation rate typically highlights students' hard work, teachers' commitment to education and multiple support systems. From the above graph, it can be inferred that in general, private universities offer higher graduation rates than the public universities.

7. Which type of colleges have more qualified faculties with a PhD or a terminal degree?



From the above graph, we can observe that both the types of universities have almost an equal number of faculties holding a PhD degree or a terminal level degree. This indicates that high quality faculties are employed in both types of universities and public universities slightly tend to have more number of qualified professors compared to private universities.

8. Which are the top 15 colleges with a lower Student to Faculty Ratio?

Top 15 colleges with a lower Student to Faculty Ratio

Type of University	college_name	Student_Faculty_Ratio
Private	University of Charleston	2.5
Private	Case Western Reserve University	2.9
Private	Johns Hopkins University	3.3
Private	Washington University	3.9
Private	Wake Forest University	4.3
Private	Saint Louis University	4.6
Private	Dartmouth College	4.7
Private	Duke University	5.0
Private	Emory University	5.0
Private	University of Chicago	5.3
Private	Vanderbilt University	5.8
Private	Yale University	5.8
Private	Columbia University	5.9
Private	University of Miami	5.9
Private	University of Rochester	5.9

From the above table, we can observe that mostly private universities have a lesser S.F. ratio than the public universities. Ex: Student_Faculty_Ratio with a value of 5.0 indicates that 1 faculty is allotted for every 5 students. According to the National Center for Educational Statistics, 16:1 is recommended in the United States (1 faculty for every 16 students).

Training and Testing datasets:

A model is to be built with Glm function and logistic regression in order to predict whether a college university is private or public.

Before proceeding to build a model, it is always a good practice to split the input dataset into training and testing data.

Training dataset can be used to train the model with a best set of predictor variables to predict the target.
Testing dataset can be used to check the accuracy of the model and compare the performance with that of the training dataset. If the accuracy and precision results are the same during training and testing phase and if there are not many deviations found in the results, the model can be reliable and it can make accurate predictions.

70% data can be used for the training phase and remaining 30% data can be used for the testing phase.

createDataPartition() function can be used to split the dataset.

```
#Splitting the data into training and testing dataset:

set.seed(3456)
trainIndex <- createDataPartition(college_df$Private, p = 0.7, list = FALSE,
                                  times = 1)
caret_train <- college_df[ trainIndex,]
caret_test <- college_df[-trainIndex,]
```

Logistic regression model using GLM function

A **Generalized Linear Model** can be used to generalize a 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.

The target variable in this case is Private - to detect if the university is private or not ('Yes' / 'No'). The model family can be **Binomial** since the target variable has 2 possible values.

The link function used is "**logit**" which provides a link between the response and the predictor variables.

```
models_best <- regsubsets(Private~., data = caret_train,
                          nvmax = 5)

summary(models_best)

model2 <- glm(Private ~ Outstate + F.Undergrad, data = caret_train,
              family = binomial(link = "logit"))
summary(model2)
```

In order to find the best set of attributes which can predict the target variable well, the best subsets regression method is used. The variables like "Outstate" tuition fees and "F.Undergrad" - number of full time undergraduates can be used to predict the target.

```
Call:
glm(formula = Private ~ Outstate + F.Undergrad, family = binomial(link = "logit"),
    data = caret_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5368  -0.0193   0.1163   0.3031   5.9002

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.141e+00  6.075e-01  -5.171 2.33e-07 ***
Outstate     7.078e-04  8.034e-05   8.810  < 2e-16 ***
F.Undergrad -5.815e-04  7.612e-05  -7.640 2.17e-14 ***
---
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: 223.81  on 542  degrees of freedom
AIC: 229.81

Number of Fisher Scoring iterations: 7
```

From the above summary, it can be observed that both the independent variables are significant in predicting the response variable. Null deviance metric tells how well the response variable can be predicted by a model with only an intercept term. Residual deviance metric tells how well the response variable can be predicted by a model with the predictor variables. AIC (Akaike information criterion) metric gives the difference between Null deviance and residual deviance and the lower the AIC value, the better is the model fit. AIC can be used to compare various models and the model with the lowest AIC value can be selected.

Creating a Confusion Matrix - Training Dataset

A confusion matrix can be created to check for the accuracy of the model predictions. It can give the results of true positives, false positives, true negative, false negatives.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  131  15
       Yes  18 381

               Accuracy : 0.9394
                 95% CI : (0.916, 0.958)
    No Information Rate : 0.7266
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8466

 Mcnemar's Test P-Value : 0.7277

            Sensitivity : 0.9621
            Specificity : 0.8792
         Pos Pred Value : 0.9549
         Neg Pred Value : 0.8973
             Prevalence : 0.7266
         Detection Rate : 0.6991
   Detection Prevalence : 0.7321
      Balanced Accuracy : 0.9207

       'Positive' Class : Yes
```

From the above confusion matrix, the below points can be inferred:

True positive count - 381 - (Case when a university is private and the model predicted as private)
True Negative count - 131 - (Case when a university is not private and the model predicted as not private)
False positive count - 18 - (Case when a university is not private and the model predicted as private)
False Negative count - 15 - (Case when a university is private and the model predicted as not private)

Sensitivity value is 0.96 which means that 96% of the university records are correctly identified as private.
Specificity value is 0.87 which means that 87% of the university records are correctly identified as public, not private.

It can be observed that the **model accuracy is 93%**. This means that the model has correctly identified the true positives and true negatives for 93% of the university records. The **remaining 7%** can be considered as the **error rate of the model** - for false positives and false negatives.

Which misclassifications are more damaging for the analysis, False Positives or False Negatives?

Both False positives and False negatives can be equally damaging since both public and private universities have their own set of pros and cons. If an actual public university is wrongly predicted as private, then it means that it can charge more tuition fees, and have low acceptance rates, can provide high scholarships but in reality it is not the case. If an actual private university is wrongly predicted as public, then it means that it has many majors to select but actual private universities have a smaller number of majors to study. Any misclassification can give the students a wrong perception about private and public universities.

Accuracy, Precision, Recall, and Specificity

- **Accuracy** of this model is a metric which tells how well the model results are accurate and reliable.It is calculated using the below formula:

  Accuracy = (Tp+Tn)/(Tp+Tn+Fp+Fn)

  Substituting the values from the confusion matrix in the above formula:

  Accuracy = (381 + 131)/(131+15+18+381) = 0.93

  This means that the model can predict 93% of the results accurately.

- **Precision** is a metric which is defined as positive predicted values. When we have a class imbalance, accuracy can become an unreliable metric for measuring our performance. Therefore we need to look at class specific performance metrics too.

  Precision <- Tp / (Tp+Fp)

  Precision <- 381/(381+18) = 0.95

This value of 0.95 or 95% means that the model is able to predict 95% of the relevant positive cases.

- **Recall**, also called sensitivity ,is the proportion of actual positive cases which are correctly identified.

  Recall <- Tp/(Tp+Fn)

  Recall <- 381/(381+15) = 0.96

This value of 0.96 or 96% means that the model is able to accurately predict 96% of the universities which are Private.

- **Specificity** is the true negative rate = 0.87 , from the above correlation matrix results.

This value of 0.87 or 87% means that the model is able to accurately predict 87% of the universities which are not private (public).

Creating a Confusion Matrix - Testing Dataset

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No   55   9
       Yes   8 160

               Accuracy : 0.9267
                 95% CI : (0.8853, 0.9567)
    No Information Rate : 0.7284
    P-Value [Acc > NIR] : 1.967e-14

                  Kappa : 0.8157

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9467
            Specificity : 0.8730
         Pos Pred Value : 0.9524
         Neg Pred Value : 0.8594
             Prevalence : 0.7284
         Detection Rate : 0.6897
   Detection Prevalence : 0.7241
      Balanced Accuracy : 0.9099

       'Positive' Class : Yes
```

- The accuracy of the model with the testing dataset is 0.92 which means that the model can accurately predict 92% of the results.

- The accuracy of the model with both training and testing datasets is not deviating much and hence there is no problem of overfitting.

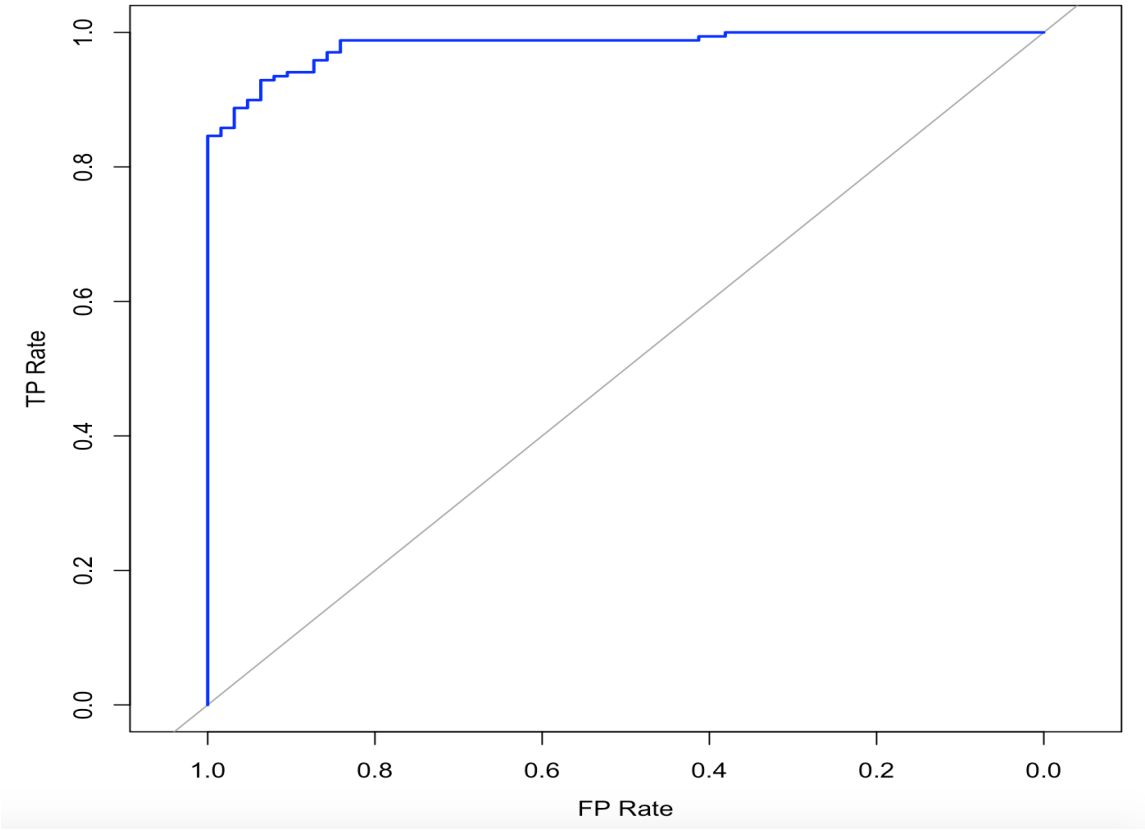
- Sensitivity metric has a value of 0.94, which means that the model can accurately predict 94%

of the universities which are private.

- Specificity metric has a value of 0.87, and the model can accurately predict 87% of the universities which are public (which are not private).

ROC curve and AUC

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. It is plotted between True positives and False positives.



If the ROC curve is closer to the diagonal, the model can give poor results. In this case, The ROC curve is closer to the top left corner, and hence the model can give better results.

AUC (Area Under Curve) value can be calculated using auc() function.

```
> auc <- auc(ROC1)
> auc
Area under the curve: 0.9815
```

The AUC of this model is close to the value of 1 (0.9815).

This indicates that the model does a very good job of predicting whether or not a university is public or private.

Conclusion:

The college dataset consisting of a large number of US Colleges from the 1995 issue of US News and World Report, was imported and analyzed. Descriptive summary statistics and some of the useful insights from the dataset were extracted and documented. A logic regression model was built to predict whether an university is a public or a private university. A Generalized linear method was used for this purpose for predicting a binomial outcome using link function as logit(). The dataset was splitted into training and testing dataset and the model was trained with 70% of the data. Remaining 30% data was used to evaluate the model performance. The results of the model were interpreted using a confusion matrix where the metrics like Accuracy, Precision,Recall, Specificity were studied. An ROC curve and area under curve were plotted and calculated to evaluate the overall performance of the model.

References:

- *Robert I. Kabacoff. (2015).* R in Action,second edition. Manning Publications Co. www.manning.com

- *Allan G.Bluman. (2018).* 558010983-Elementary-Statistics-a-Step-by-Step-Approach-10th-Edition. McGraw-Hill Education

Appendix:

.

```
#-------------------------------------------------------------------------------#
# Shyamala Venkatakrishnan                                        01/26/2023 #
#                                                                               #
#             ALY6015: Module 3 Assignment - GLM and Logistic Regression        #
#                                                                               #
#-------------------------------------------------------------------------------#

install.packages(c("dplyr","ggplot2","sqldf","tidyverse","RColorBrewer",
                   "plotly","gmodels","formattable","tidyr"))

loadlibrary <- c("dplyr","ggplot2","sqldf","tidyverse","RColorBrewer",
                 "plotly","gmodels","formattable","tidyr")

lapply(loadlibrary,require,character.only=TRUE)


install.packages("ISLR")
library(ISLR)

college_df <- College

head(college_df)
```

```r
summary(college_df)
View(college_df)
dim(college_df)

install.packages('DataExplorer')

library(DataExplorer)

create_report(college_df)
rownames(college_df)

college_df = college_df %>%
  mutate(
    college_name = rownames(college_df)
  )

college_df = college_df %>%
  mutate(
    Type = case_when(
      Private == "Yes" ~ "Private",
      Private == "No"  ~ "Public"
    )
  )

View(college_df)

#EDA:

install.packages("psych")
library(psych)

formattable(describe(college_df),
            caption = "Descriptive statistics summary of the baseball dataset")

describe(college_df)

#1. What is the percentage of private and public universities in this dataset?

private_table <- table(college_df$Type)

private_table_perc <- round((private_table / 777) * 100, digits = 2)

private_table_df <- data.frame(private_table_perc)

colnames(private_table_df)[1] <- "Type_of_university"
colnames(private_table_df)[2] <- "Percentage"

ggplot(private_table_df, aes(x = "", y = Percentage, fill = Type_of_university)) +
  geom_col() + geom_text(aes(label = paste(Percentage,"%", sep = ""))
                         ,position = position_stack(vjust = 0.5)) +
```

```
  guides(fill = guide_legend(title = "Type of university")) +
  coord_polar(theta = "y") +
  theme_void() +
  scale_fill_brewer(palette = "Set2")+
  ggtitle("Percentage of private and public universities")+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        text=element_text(size=12,  family="Comic Sans MS", color= "black"))

#Almost 73% of colleges are private universities and rest 27% of colleges are
#public universities.

#2. How many applications are received in public and private universities?

ggplot(college_df, aes(x=Apps))+
  geom_histogram(color="darkblue", fill="lightblue")+
  facet_grid(. ~ Type)+
  theme_bw()+xlab("No. of Applications")+ylab("No. of colleges")+
  ggtitle("Distribution of applications received in public vs private universities")+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        text=element_text(size=12,  family="Comic Sans MS", color= "black"))

#3. Top 10 colleges with max  number of applications received.

avg_applns_df <- sqldf("select college_name,Apps as Applications from
                       college_df group by college_name
                      order by Applications desc limit 10")


ggplot(avg_applns_df, aes(x = reorder(college_name, Applications), y = Applications)) +
  geom_col(width = 0.5,aes(fill  = Applications))+
  ggtitle("Top 10 colleges with max number of applications")+
  xlab("College name")+
  ylab("No. of applications")+theme_bw()+coord_flip()+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        text=element_text(size=12,  family="Comic Sans MS", color= "black"))

#5. Top 10 colleges with the least acceptance rate:


least_accept_df <- sqldf("select college_name,Type,round((Accept/Apps) * 100,2)
                      as Acceptance_rate from college_df
                      group by college_name
                      order by acceptance_rate limit 15")

ggplot(least_accept_df, aes(x = reorder(college_name, -acceptance_rate),
                            y = acceptance_rate)) +
  geom_col(width = 0.5,aes(fill  = acceptance_rate))+
  ggtitle("Top 15 colleges with the least acceptance rate")+
  xlab("College name")+
  ylab("Acceptance rate")+theme_bw()+coord_flip()+
```

```
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        text=element_text(size=12,  family="Comic Sans MS", color= "black"))

install.packages("devtools")
library(devtools)

devtools::install_github("yogevherz/plotme",force = TRUE)

library(plotme)

least_accept_df %>%
  count(Type, college_name,Acceptance_rate) %>%
  count_to_sunburst()

#6. Which type of colleges generally are costlier to study?

c1 <- ggplot(college_df, aes(fill=Type, y=Outstate, x=Type)) +
  geom_bar(position="dodge", stat="identity")+
  scale_fill_brewer(palette = "Pastel1")

c2 <- ggplot(college_df, aes(fill=Type, y=Room.Board, x=Type)) +
  geom_bar(position="dodge", stat="identity")+
  scale_fill_brewer(palette = "Pastel1")

c3 <-  ggplot(college_df, aes(fill=Type, y=Books, x=Type)) +
  geom_bar(position="dodge", stat="identity")+
  scale_fill_brewer(palette = "Pastel1")

c4 <-  ggplot(college_df, aes(fill=Type, y=Personal, x=Type)) +
  geom_bar(position="dodge", stat="identity")+
  scale_fill_brewer(palette = "Pastel1")


c5 <-  ggplot(college_df, aes(fill=Type, y=Expend, x=Type)) +
  geom_bar(position="dodge", stat="identity")+
  scale_fill_brewer(palette = "Pastel1")


install.packages("lattice")
library(lattice)
library(gridExtra)
library(grid)

grid.arrange(c1,c2,c3,c4,c5,ncol=2,nrow=3)

#7. Which type of colleges have good graduation rates? Name the top 15

top_grad_rate_df <- sqldf("select college_name,Type,`Grad.Rate`
                    as Graduation_Rate from college_df
                    order by Graduation_Rate desc limit 15")
```

```
top_grad_rate_df %>%
  count(Type, college_name,Graduation_Rate) %>%
  count_to_sunburst()

#8. Which type of colleges have more qualified faculties with a PhD or a terminal degree?

q1_phd <- sqldf("select Type, avg(PhD) as avg_PhD from college_df
                  group by Type")

q2_terminal <- sqldf("select Type, avg(Terminal) as avg_Terminal from college_df
                  group by Type")

q1 <- ggplot(q1_phd, aes(fill=Type, y=avg_PhD, x=Type)) +
  geom_bar(position="dodge", stat="identity")+
  scale_fill_brewer(palette = "Pastel2")+
  ylab("Avg. no. of faculties with a PhD degree")+theme_bw()+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        text=element_text(size=12,  family="Comic Sans MS", color= "black"))

q2 <- ggplot(q2_terminal, aes(fill=Type, y=avg_Terminal, x=Type)) +
  geom_bar(position="dodge", stat="identity")+
  scale_fill_brewer(palette = "Pastel2")+
  ylab("Avg. no. of faculties with a Terminal degree")+theme_bw()+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        text=element_text(size=12,  family="Comic Sans MS", color= "black"))

grid.arrange(q1,q2,ncol=2,nrow=1,
             top=textGrob("Avg. No of faculties with a PhD/Terminal
degree",gp=gpar(fontsize=14,font=3)))

#9. what are the top 15 colleges with a lower Student to Faculty Ratio?

sf_ratio <- sqldf("select Type as 'Type of University', college_name,`S.F.Ratio`
                       as Student_Faculty_Ratio from college_df
                      order by Student_Faculty_Ratio limit 15")

prince <- sqldf("select Type, college_name,`S.F.Ratio`
                       as Student_Faculty_Ratio from college_df
                      where college_name like '%Princeton%'")
formattable(sf_ratio,
            caption = "Top 15 colleges with a lower Student to Faculty Ratio",
            align = c("l","c","r"))

ggplot(neigh_recent, aes(x = reorder(Neighborhood, count_houses)
                         , y = count_houses)) +
  geom_segment(aes(x = reorder(Neighborhood, count_houses),
                   xend = reorder(Neighborhood, count_houses),
                   y = 0, yend = count_houses)) +
  geom_point(size = 4, pch = 21, bg = 4, col = 1)  +
```

```
  xlab("Neighborhood")+theme_bw()+
  ylab("No of houses")+
  ggtitle("Which neighborhood has the most number of houses built after 1980?")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  coord_flip()+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        text=element_text(size=12,  family="Comic Sans MS", color= "black"))

#Splitting the data into training and testTerminalt:

set.seed(3456)

q2trainIndex <- createDataPartition(college_df$Private, p = 0.7, list = FALSE,
                                    times = 1)
caret_train <- college_df[ trainIndex,]
caret_test <- college_df[-trainIndex,]

#Glm model

install.packages("leaps")
library(leaps)

models_best <- regsubsets(Private~., data = caret_train,
                          nvmax = 5)

summary(models_best)

model2 <- glm(Private ~ Outstate + F.Undergrad, data = caret_train,
              family = binomial(link = "logit"))
summary(model2)

#Trainset predictions:

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

#Model accuracy:

confusionMatrix(predicted.classes.min, caret_train$Private, positive = 'Yes')

#Model accuracy is high = 0.93
#sensitivity = 0.96 = Tue positive rate - the percentage of colleges
#the model correctly predicted to be private university.

#Specificity = 0.87 = True negative rate -the percentage of colleges
#the model correctly predicted to be public university.

#Error rate = 0.0606 = 6.06%

#Accuracy,precision,recall,specificity
```

```r
accuracy <- (tp+tn)/tp+tn+fp+fn

accuracy <- (381 + 131)/(131+15+18+381)
accuracy

#Precision is defined as positive predicted values
precision <- tp/tp+fp
precision <- 381/(381+18)

precision

#Recall -  it is the proportion of actual positive cases which are correctly identified.

recall <- tp/(tp+fn)
recall <- 381/(381+15)

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

#Model accuracy:

confusionMatrix(predicted.classes.min, caret_test$Private, positive = 'Yes')

#Accuracy is almost same with both  train and testing data.Hence there is no
#problem of overfitting

#Roc and Auc
library(pROC)

ROC1 <- roc(caret_test$Private,probabilities.test )

plot(ROC1, col = "blue", ylab = "TP Rate",
     xlab = "FP Rate")

auc <- auc(ROC1)

#AUC of this model is close to 1.
#this indicates that the model does a very good job of predicting
#whether or not a college is a private university or public univ.
```