CPS: Data Analytics & AI

Northeastern University

Name: Nam Ho Phan

Course: Intermediate Analytics (ALY6015.71882.202115)

Professor: Vladimir Shapiro

Major: Analytics

Title: GLM and Logistic Regression Assignment

Date: October 5th, 2020

Introduction:

This assignment will help us practice the generalized logistic model and regression. In real life, this analysis is very popular for classification because it can let us know the "yes/no" question accurately with more insights. For example, if the hospital wants to detect the disease of victim, they use the machine learning with a good generalized logistic model, and then this will let them find the potential disease effectively. Moreover, the generalized logistic model can help people to predict model with a non-normal distribution.

For this time, we have the data of universities in America to predict whether there is a private or non-private university. This is helpful in some cases such as the education organization want to find out the general standard of private university, or the business want to investigate money into underrated universities. In this data, we have totally 777 observations and 18 variables. This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the ASA Statistical Graphics Section's 1995 Data Analysis Exposition.

Before we find out how to find out the private university, it is also important for us to discover this set of data to get more helpful information.

Analysis:

Exploratory Data Analysis

To explore the new information from a set of given data I get from ILSR, I need to put some questions related to College data, which is the difference of private and non-private university. Here are some questions:

- Whether private university has more top 10% and 25% student from class?
- There is the similarity between top 10 highest out-of-state tuition in private and non-private university
- Do the universities having lower acceptance rate have more drop-out students?

```
library(ISLR)
college_data = ISLR::College

head(college_data)

##                              Private Apps Accept Enroll Top10perc T
op25perc
## Abilene Christian University     Yes 1660   1232    721        23
52
## Adelphi University               Yes 2186   1924    512        16
29
## Adrian College                   Yes 1428   1097    336        22
50
## Agnes Scott College              Yes  417    349    137        60
89
## Alaska Pacific University        Yes  193    146     55        16
44
## Albertson College                Yes  587    479    158        38
62
##                              F.Undergrad P.Undergrad Outstate Room.
Board Books
## Abilene Christian University        2885         537     7440
```

```
3300   450
## Adelphi University                  2683        1227    12280
6450   750
## Adrian College                      1036          99    11250
3750   400
## Agnes Scott College                  510          63    12960
5450   450
## Alaska Pacific University            249         869     7560
4120   800
## Albertson College                    678          41    13500
3335   500
##                              Personal PhD Terminal S.F.Ratio perc.a
lumni Expend
## Abilene Christian University     2200  70       78      18.1
12   7041
## Adelphi University               1500  29       30      12.2
16  10527
## Adrian College                   1165  53       66      12.9
30   8735
## Agnes Scott College               875  92       97       7.7
37  19016
## Alaska Pacific University        1500  76       72      11.9
2  10922
## Albertson College                 675  67       73       9.4
11   9727
##                              Grad.Rate
## Abilene Christian University        60
## Adelphi University                  56
## Adrian College                      54
## Agnes Scott College                 59
## Alaska Pacific University           15
## Albertson College                   55
```

```r
#The quality between private and non-private universities
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

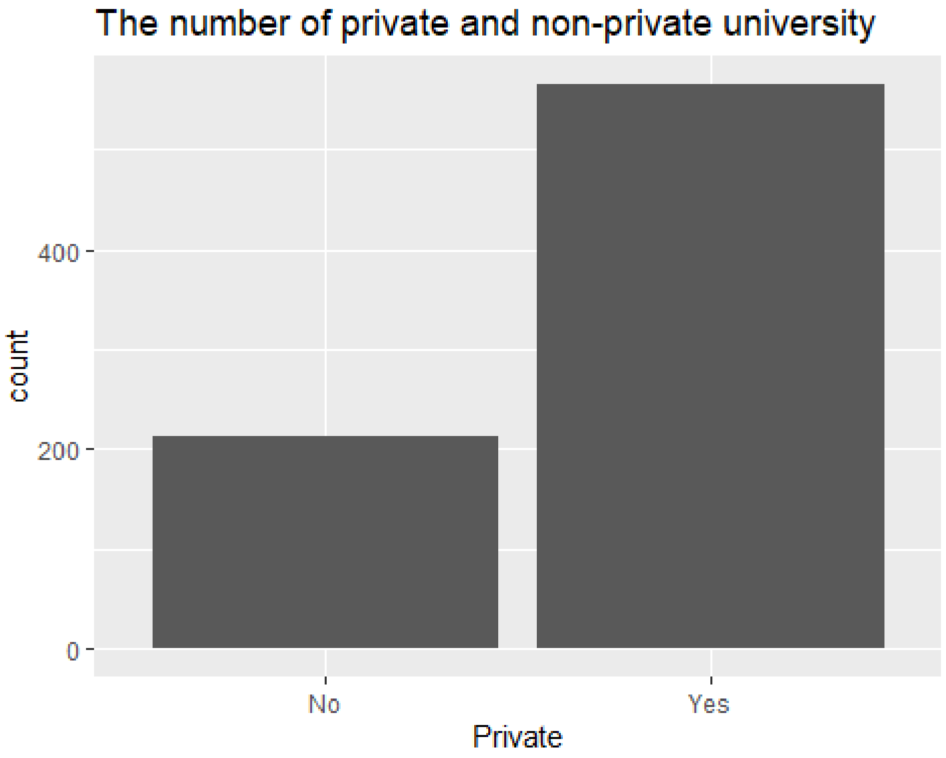
```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

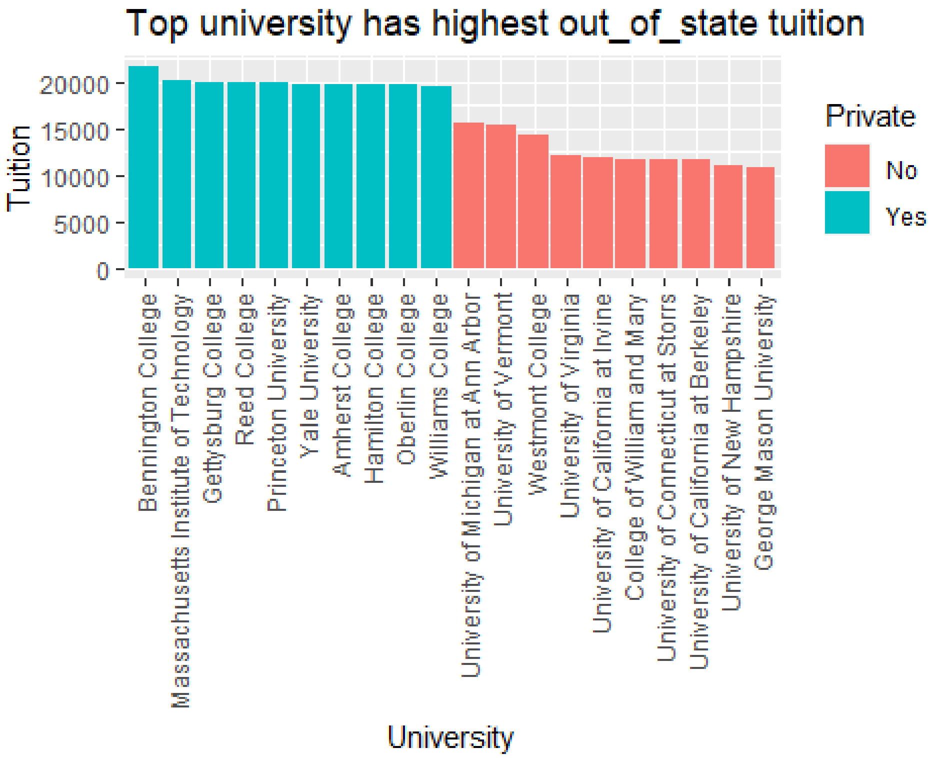
```
ggplot(college_data,aes(Private)) + geom_bar() + labs(title="The numbe
r of private and non-private university")
```



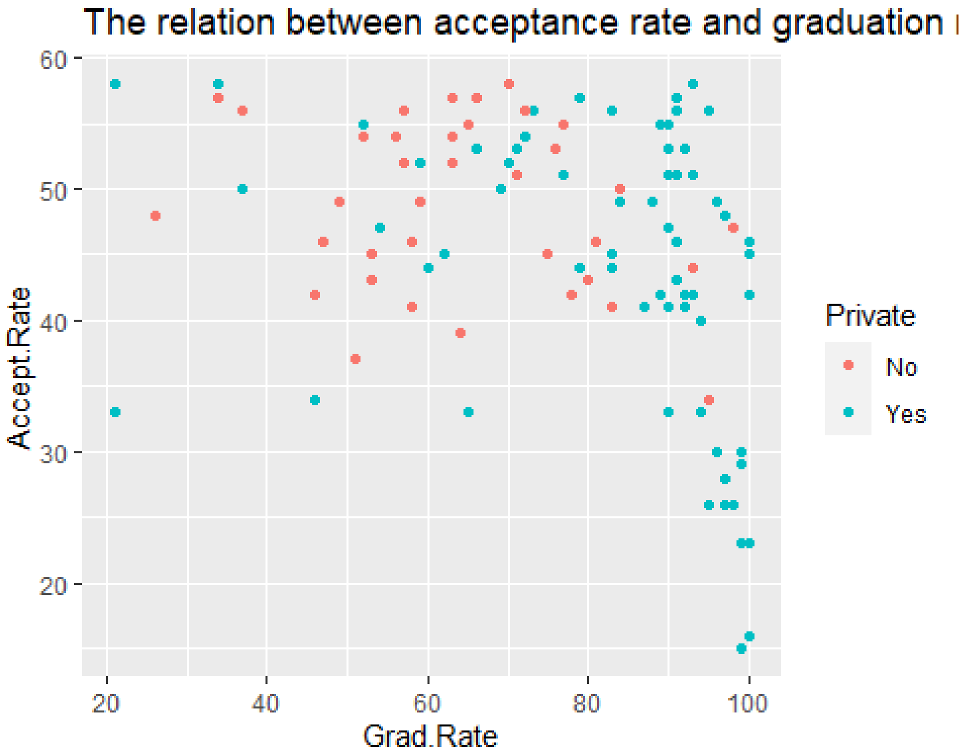
```
table(college_data['Private'])
```

```
##
##  No Yes
## 212 565
```

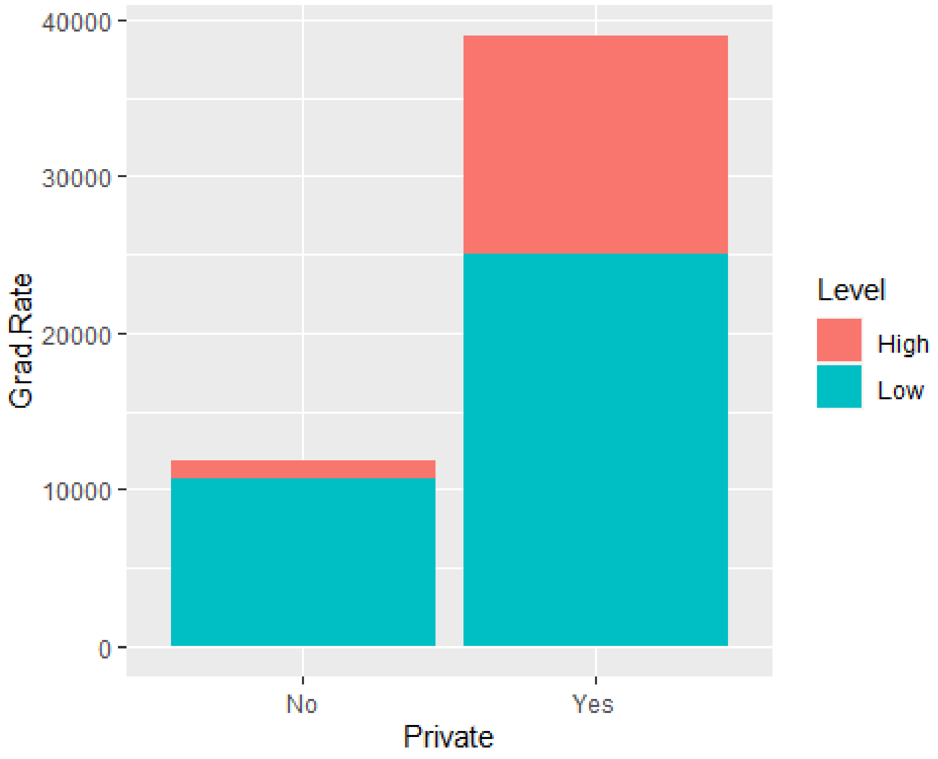
```
#Whether private university has more top 10% and 25% student from clas
s?
p_student <- college_data %>% subset(Private == "Yes") %>% select(Priv
ate,Top10perc,Top25perc)  %>% summarise(mean10 = mean(Top10perc),mean2
5=mean(Top25perc))

np_student <- college_data %>% subset(Private == "No") %>% select(Priv
ate,Top10perc,Top25perc)  %>% summarise(mean10 = mean(Top10perc),mean2
5= mean(Top25perc))

combine = rbind(p_student,np_student)
row.names(combine) = c("private","non-private")


#   Top 10 private university has highest out of state tuition?

top10_pri <- college_data %>% subset(Private == "Yes") %>% select(Priv
ate,Outstate,Accept) %>% arrange(desc(Outstate)) %>% head(n=10)

top10_nonpri <- college_data %>% subset(Private == "No") %>% select(Pr
ivate,Outstate,Accept) %>% arrange(desc(Outstate)) %>% head(n=10)
top_10 = rbind(top10_pri,top10_nonpri)

ggplot(top_10,aes(x=reorder(rownames(top_10),-Outstate),y=Outstate,fil
l=Private)) + geom_col() + theme(axis.text.x = element_text(angle = 90
, vjust = 0.5, hjust=1)) + labs(title="Top university has highest out_
of_state tuition",x="University",y="Tuition")
```



```
#The relationship between acceptance rate and the rate of drop off?
college_data$University = row.names(college_data)
accept_rate <- college_data %>%
mutate(Accept.Rate= round(Accept/Apps *100,0)) %>% arrange(Accept.Rate
) %>% select(Private,University,Accept.Rate,Grad.Rate) %>% head(100)

ggplot(accept_rate,aes(x=Grad.Rate,y=Accept.Rate,col=Private)) + geom_
point() +labs(title="The relation between acceptance rate and graduati
on rate")
```

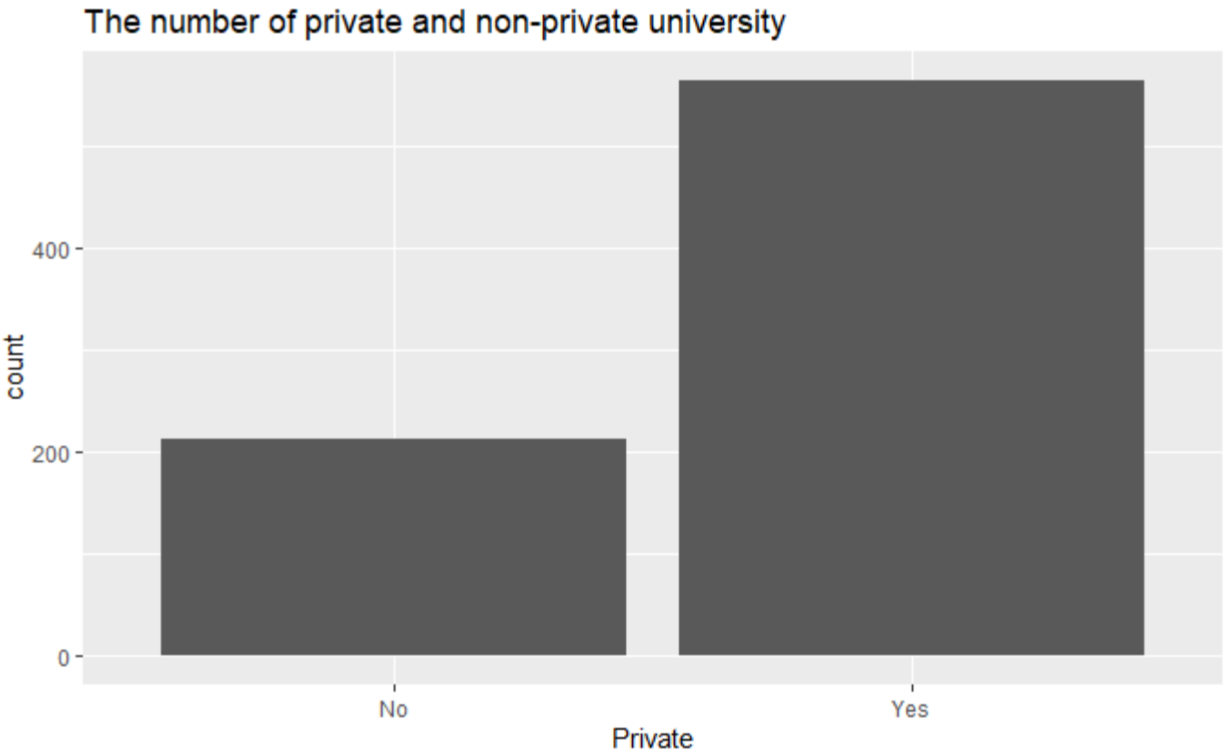


```
#The graduation rate in private university is better than non-private
university?
graduation_rate <- college_data  %>% arrange(desc(Grad.Rate)) %>% sele
ct(Private,Grad.Rate) %>% mutate(Level = ifelse(Grad.Rate < 80,"Low","
High"))
table(graduation_rate$Private,graduation_rate$Level)

##
##       High Low
##   No    13 199
##   Yes  158 407

ggplot(graduation_rate,aes(Private)) + geom_col(aes(y=Grad.Rate,fill=L
evel))
```



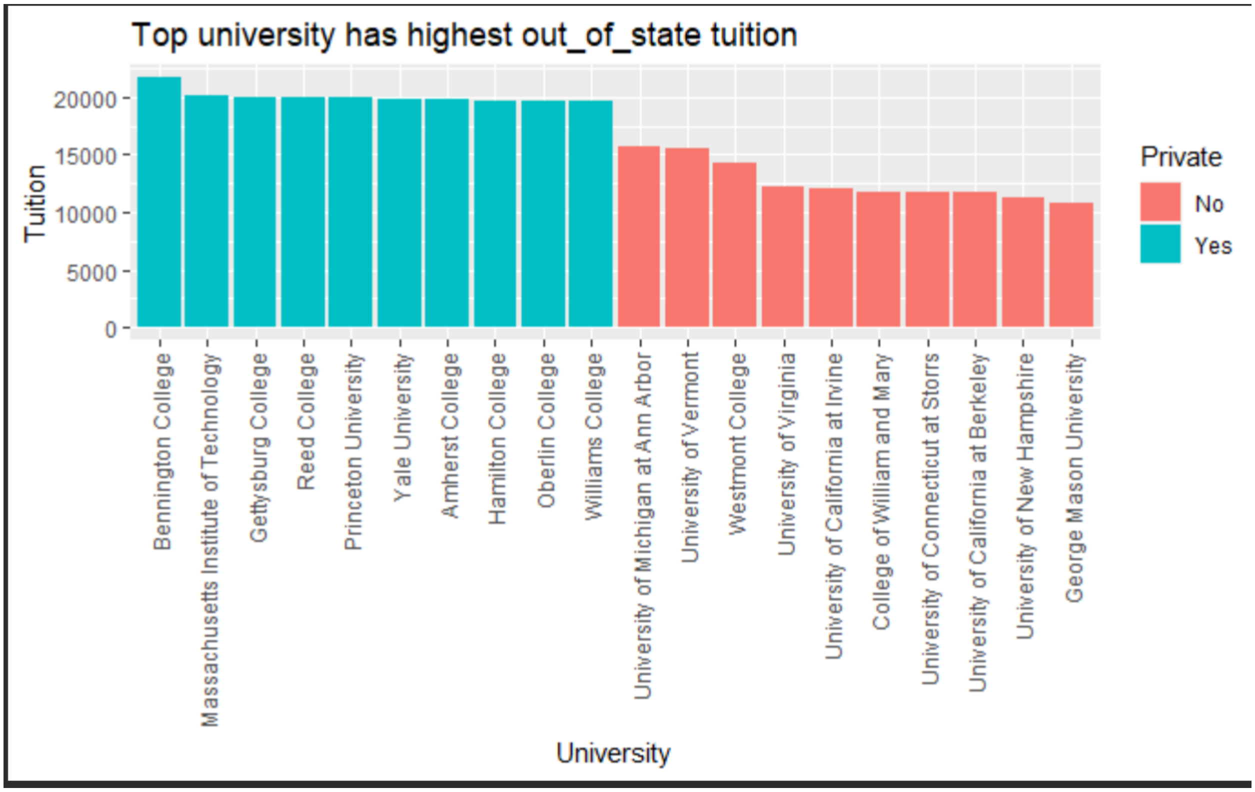
```
head(graduation_rate)

##   Private Grad.Rate Level
## 1     Yes       118  High
## 2     Yes       100  High
## 3     Yes       100  High
## 4     Yes       100  High
## 5     Yes       100  High
## 6     Yes       100  High
```

It is helpful to look at the number of private and non-private universities in the US


The number of private and non-private university

It seems like the number of private universities is significantly more than non-private university. The number of non-private universities is over 200 while the number of private universities is over 550. One of factors contributing to the reputation of universities is their attractiveness to top high school students. Therefore, it causes me to be curious about the university having these students, I decided to compare the mean between two types of university.

Both the average of top 10% and top 25 % students comes from private universities, which is 29.33 and 56.95, compared to 22.83 and 52.70 from non-private universities. Next, my question is whether the highest out-of-state tuition of two types of university is equal?



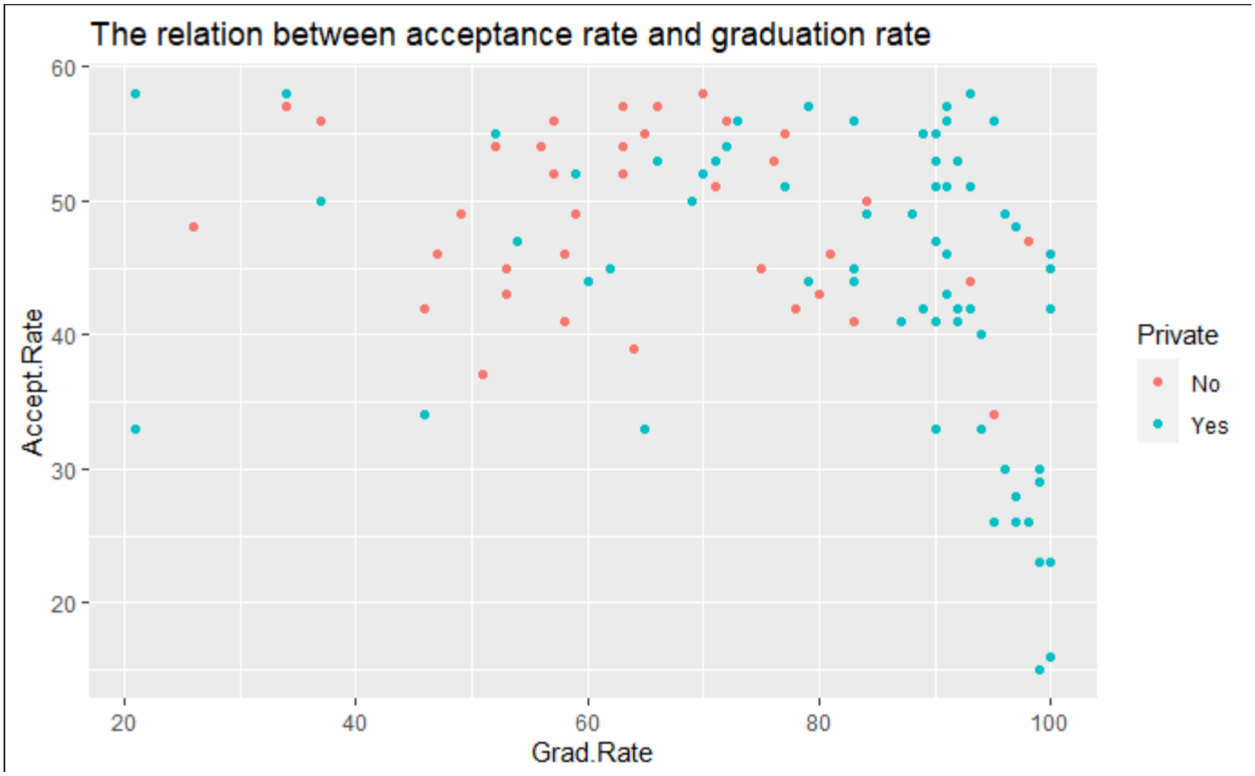
The graph shows me most of highest tuition is from private universities. Additionally, it is good to know more the tuition of private and non-private universities.

```
summary(top10_pri['Outstate'])

##     Outstate
##  Min.   :19629
##  1st Qu.:19715
##  Median :19870
##  Mean   :20022
##  3rd Qu.:19963
##  Max.   :21700

summary(top10_nonpri['Outstate'])
```

```
     ##     Outstate

##  Min.   :10800
##   1st Qu.:11650
##   Median :11872
##   Mean   :12681
```

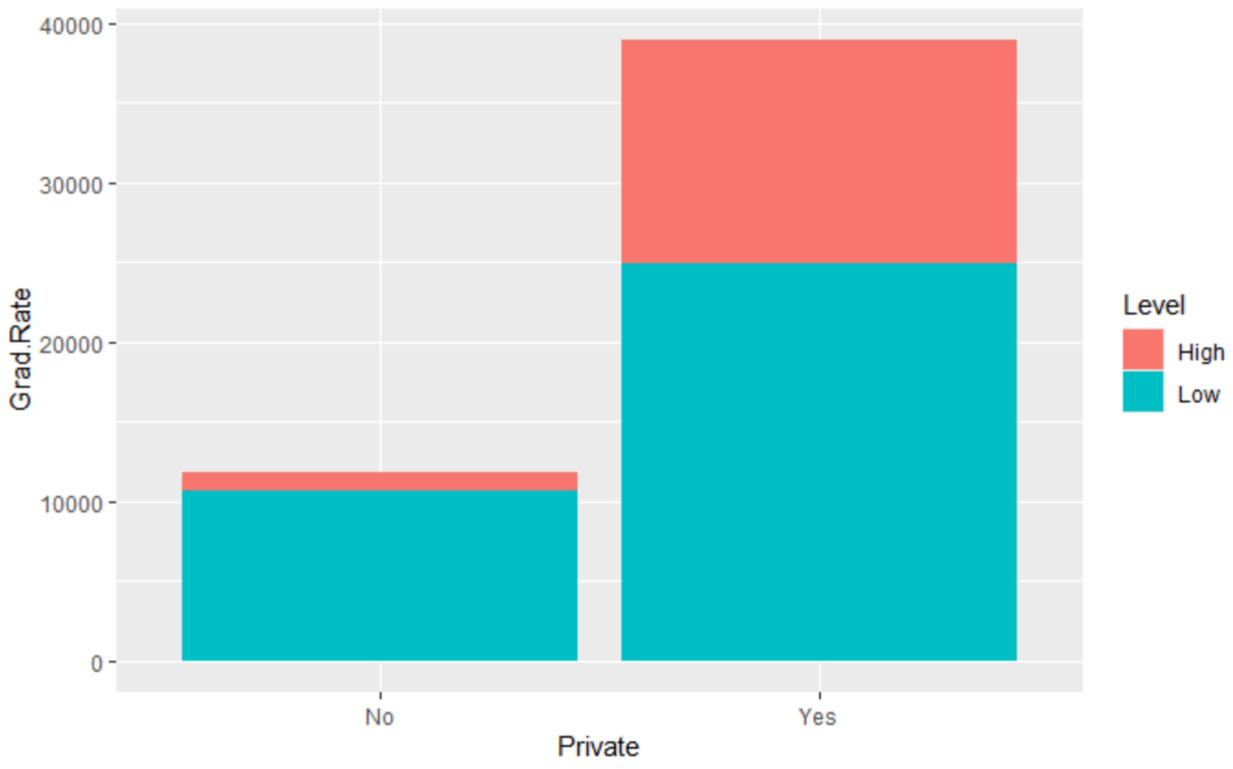
The summary shows most private universities have the tuition higher than 19,963 while the maximum tuition of non-private university is just 15,732

Next, it is good to know whether there is the relation between the acceptance rate and graduation rate among private and non-private universities. We can see that there is no relation between the acceptance rate and graduation rate despite few outliers.



Finally, I want to know if the graduation rate of private university is higher than non-private university. I suppose the graduation rate less than 80% is low and greater than 80% is high. The percentage of private university is 0.2, greater than 0.06 of non-private university.

```
      High Low
  No    13 199
  Yes  158 407
```



To create a model for prediction, it is necessary to split the data into training set and test set. Training data is used to fit the model and the test data is to predict outcome. The training set is implemented to build up the model, while the test set is to validate the model built.

#Split the data into a train and test set – refer to the pdf document for information on how to split a dataset.

```
college_data['University'] = NULL
# Create Train and Test set - maintain % of event rate (70/30 split)
N = nrow(college_data)
target = round(N*0.75)
vector = runif(N)

sample_rows <- sample(nrow(college_data), nrow(college_data)*75/100)
#training set
college_training = college_data[sample_rows,]
head(college_training)
```

```
##                                    Private  Apps Accept Enroll Top10
perc
## New York University                   Yes 13594   7244   2505
70
## Davidson College                      Yes  2373    956    452
77
## Columbia University                   Yes  6756   1930    871
78
## Longwood College                       No  2747   1870    724
12
## Brigham Young University at Provo     Yes  7365   5402   4615
48
## University of San Francisco           Yes  2306   1721    538
23
##                                    Top25perc F.Undergrad P.Undergrad
Outstate
## New York University                      86       12408        2814
17748
## Davidson College                         96        1601           6
17295
## Columbia University                      96        3376          55
18624
## Longwood College                         47        2874         118
7920
## Brigham Young University at Provo        82       27378        1253
2340
## University of San Francisco              48        4309         549
13226
##                                    Room.Board Books Personal PhD Ter
minal
## New York University                     7262   450     1000  87
98
## Davidson College                        5070   600     1011  95
97
## Columbia University                     6664   550      300  97
```

```
98
## Longwood College                        3962   550     2200  74
80
## Brigham Young University at Provo       3580   860     1220  76
76
## University of San Francisco             6452   750     2450  86
86
##                                   S.F.Ratio perc.alumni Expend Grad
.Rate
## New York University                     7.8          16  21227
71
## Davidson College                       12.0          46  17581
94
## Columbia University                     5.9          21  30639
99
## Longwood College                       18.4          23   5553
62
## Brigham Young University at Provo      20.5          40   7916
33
## University of San Francisco            13.6           8  10074
62
```

```
#test set
college_test= college_data[-sample_rows,]
head(college_test)
```

```
##                           Private Apps Accept Enroll Top10perc Top2
5perc
## Adrian College                Yes 1428   1097    336        22
50
## Albertson College             Yes  587    479    158        38
62
## Albertus Magnus College       Yes  353    340    103        17
45
## Albion College                Yes 1899   1720    489        37
68
## Alderson-Broaddus College     Yes  582    498    172        21
44
## Baker University              Yes  602    483    206        21
47
##                           F.Undergrad P.Undergrad Outstate Room.Boa
rd Books
## Adrian College                    1036          99    11250       37
50   400
## Albertson College                  678          41    13500       33
35   500
```

```
## Albertus Magnus College           416         230    13290       57
20   500
## Albion College                   1594          32    13868       48
26   450
## Alderson-Broaddus College         799          78    10468       33
80   660
## Baker University                  958         466     8620       41
00   400
##                           Personal PhD Terminal S.F.Ratio perc.alum
ni Expend
## Adrian College                1165  53       66      12.9
30   8735
## Albertson College              675  67       73       9.4
11   9727
## Albertus Magnus College       1500  90       93      11.5
26   8861
## Albion College                 850  89      100      13.7
37  11487
## Alderson-Broaddus College     1800  40       41      11.5
15   8991
## Baker University              2250  58       68      11.0
21   6136
##                           Grad.Rate
## Adrian College                   54
## Albertson College                55
## Albertus Magnus College          63
## Albion College                   73
## Alderson-Broaddus College        52
## Baker University                 65
```

```
#training_data
null_model = glm(as.numeric(Private)~1,data=college_training,family="g
aussian")
full_model = glm(as.numeric(Private)~.,data=college_training,family="g
aussian")
step_model = step(null_model,scope = list(lower=null_model,upper=full_
model),direction="forward")
```

```
## Start:  AIC=682.72
## as.numeric(Private) ~ 1
##
##               Df Deviance    AIC
## + F.Undergrad  1   68.014 408.23
## + Enroll       1   73.674 454.76
## + Outstate     1   77.942 487.53
## + Accept       1   85.379 540.57
## + S.F.Ratio    1   86.141 545.75
## + P.Undergrad  1   88.825 563.60
## + Apps         1   90.456 574.19
## + perc.alumni  1   92.776 588.93
## + Room.Board   1   97.076 615.30
## + Grad.Rate    1   98.025 620.96
## + Personal     1   99.402 629.08
## + Expend       1  102.357 646.13
## + PhD          1  106.643 670.00
## + Top10perc    1  106.652 670.05
## + Terminal     1  107.962 677.15
## + Top25perc    1  108.588 680.52
## <none>           109.375 682.72
## + Books        1  109.320 684.43
##
## Step:  AIC=408.23
## as.numeric(Private) ~ F.Undergrad
##
##               Df Deviance    AIC
## + Outstate     1   48.698 215.80
## + Room.Board   1   57.457 312.06
## + S.F.Ratio    1   58.501 322.54
## + Grad.Rate    1   59.477 332.18
## + Expend       1   60.062 337.87
## + Top10perc    1   60.404 341.17
## + perc.alumni  1   60.540 342.48
## + Top25perc    1   62.830 364.09
## + Apps         1   66.155 394.10
## + Accept       1   66.551 397.58
## + P.Undergrad  1   66.677 398.68
## + Personal     1   66.996 401.46
## + Terminal     1   67.362 404.63
## + Enroll       1   67.511 405.91
## <none>             68.014 408.23
## + PhD          1   67.797 408.37
## + Books        1   67.798 408.38
##
## Step:  AIC=215.8
## as.numeric(Private) ~ F.Undergrad + Outstate
##
##               Df Deviance    AIC
## + PhD          1   44.905 170.60
## + Terminal     1   45.482 178.03
## + S.F.Ratio    1   47.779 206.71
## + Grad.Rate    1   48.397 214.19
## + Room.Board   1   48.424 214.52
## + P.Undergrad  1   48.428 214.57
## + perc.alumni  1   48.468 215.04
## + Apps         1   48.507 215.51
## <none>             48.698 215.81
## + Books        1   48.599 216.62
## + Expend       1   48.601 216.64
## + Accept       1   48.641 217.12
## + Top25perc    1   48.665 217.40
## + Top10perc    1   48.695 217.77
## + Personal     1   48.697 217.79
## + Enroll       1   48.698 217.79
##
## Step:  AIC=170.6
## as.numeric(Private) ~ F.Undergrad + Outstate + PhD
##
##               Df Deviance    AIC
## + S.F.Ratio    1   43.990 160.62
## + Grad.Rate    1   44.369 165.62
## + perc.alumni  1   44.401 166.03
## + Top10perc    1   44.427 166.37
## + Room.Board   1   44.548 167.96
## + Top25perc    1   44.634 169.08
## + P.Undergrad  1   44.703 169.99
## + Terminal     1   44.744 170.52
## + Apps         1   44.745 170.53
## <none>             44.905 170.60
## + Accept       1   44.834 171.68
## + Books        1   44.885 172.34
## + Expend       1   44.899 172.53
## + Personal     1   44.901 172.55
## + Enroll       1   44.905 172.60
##
## Step:  AIC=160.62
## as.numeric(Private) ~ F.Undergrad + Outstate + PhD + S.F.Ratio
##
##               Df Deviance    AIC
## + Grad.Rate    1   43.454 155.49
## + perc.alumni  1   43.636 157.93
## + Room.Board   1   43.652 158.13
## + Top10perc    1   43.735 159.25
## + Apps         1   43.777 159.80
## + Terminal     1   43.798 160.08
## + P.Undergrad  1   43.799 160.09
## + Top25perc    1   43.821 160.39
## <none>             43.990 160.62
## + Expend       1   43.850 160.77
## + Accept       1   43.921 161.71
## + Books        1   43.985 162.56
## + Enroll       1   43.986 162.57
## + Personal     1   43.989 162.62
##
## Step:  AIC=155.49
## as.numeric(Private) ~ F.Undergrad + Outstate + PhD + S.F.Ratio +
##     Grad.Rate
##
##               Df Deviance    AIC
## + Apps         1   43.095 152.66
## + Room.Board   1   43.175 153.74
## + Terminal     1   43.239 154.61
## + perc.alumni  1   43.270 155.03
## <none>             43.454 155.49
## + Expend       1   43.333 155.87
## + Accept       1   43.339 155.95
## + Top10perc    1   43.342 155.98
## + P.Undergrad  1   43.368 156.34
## + Top25perc    1   43.398 156.74
## + Enroll       1   43.435 157.23
## + Books        1   43.443 157.35
## + Personal     1   43.447 157.39
##
## Step:  AIC=152.66
## as.numeric(Private) ~ F.Undergrad + Outstate + PhD + S.F.Ratio +
##     Grad.Rate + Apps
##
##               Df Deviance    AIC
## + Room.Board   1   42.725 149.64
## + Terminal     1   42.850 151.34
## + Top10perc    1   42.907 152.11
## <none>             43.095 152.66
## + perc.alumni  1   42.959 152.82
## + P.Undergrad  1   42.990 153.24
## + Top25perc    1   43.012 153.54
## + Accept       1   43.035 153.85
## + Expend       1   43.046 154.00
## + Books        1   43.079 154.45
## + Enroll       1   43.088 154.56
## + Personal     1   43.089 154.58
##
## Step:  AIC=149.64
## as.numeric(Private) ~ F.Undergrad + Outstate + PhD + S.F.Ratio +
##     Grad.Rate + Apps + Room.Board
##
```

```
##               Df Deviance    AIC
## + Terminal     1   42.427 147.57
## + Top10perc    1   42.466 148.11
## + perc.alumni  1   42.475 148.23
## + P.Undergrad  1   42.557 149.36
## <none>             42.725 149.64
## + Top25perc    1   42.607 150.03
## + Accept       1   42.666 150.84
## + Expend       1   42.674 150.95
## + Enroll       1   42.704 151.35
## + Personal     1   42.715 151.50
## + Books        1   42.722 151.61
##
## Step:  AIC=147.57
## as.numeric(Private) ~ F.Undergrad + Outstate + PhD + S.F.Ratio +
##     Grad.Rate + Apps + Room.Board + Terminal
##
##               Df Deviance    AIC
## + perc.alumni  1   42.159 145.89
## + Top10perc    1   42.180 146.17
## + Top25perc    1   42.276 147.49
## + P.Undergrad  1   42.280 147.55
## <none>             42.427 147.57
## + Accept       1   42.369 148.77
## + Expend       1   42.384 148.98
## + Books        1   42.406 149.28
## + Enroll       1   42.413 149.38
## + Personal     1   42.416 149.42
##
## Step:  AIC=145.89
## as.numeric(Private) ~ F.Undergrad + Outstate + PhD + S.F.Ratio +
##     Grad.Rate + Apps + Room.Board + Terminal + perc.alumni
##
##               Df Deviance    AIC
## + Top10perc    1   41.998 145.66
## <none>             42.159 145.89
## + P.Undergrad  1   42.030 146.10
## + Accept       1   42.064 146.57
## + Top25perc    1   42.069 146.64
## + Expend       1   42.098 147.03
## + Books        1   42.131 147.49
## + Personal     1   42.135 147.55
## + Enroll       1   42.152 147.78
##
## Step:  AIC=145.66
## as.numeric(Private) ~ F.Undergrad + Outstate + PhD + S.F.Ratio +
##     Grad.Rate + Apps + Room.Board + Terminal + perc.alumni +
##     Top10perc
##
##               Df Deviance    AIC
## + Accept       1   41.730 143.93
## + Expend       1   41.820 145.18
## <none>             41.998 145.66
## + P.Undergrad  1   41.901 146.32
## + Personal     1   41.979 147.39
## + Books        1   41.982 147.44
## + Enroll       1   41.991 147.56
## + Top25perc    1   41.998 147.65
##
## Step:  AIC=143.93
## as.numeric(Private) ~ F.Undergrad + Outstate + PhD + S.F.Ratio +
##     Grad.Rate + Apps + Room.Board + Terminal + perc.alumni +
##     Top10perc + Accept
##
##               Df Deviance    AIC
## <none>             41.730 143.93
## + Expend       1   41.630 144.53
## + P.Undergrad  1   41.662 144.98
## + Personal     1   41.703 145.56
## + Books        1   41.708 145.62
## + Enroll       1   41.713 145.69
## + Top25perc    1   41.716 145.74
```

```
summary(step_model)
```

```
##
## Call:
## glm(formula = as.numeric(Private) ~ F.Undergrad + Outstate +
##     PhD + S.F.Ratio + Grad.Rate + Apps + Room.Board + Terminal +
##     perc.alumni + Top10perc + Accept, family = "gaussian", data = c
ollege_training)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.84585  -0.14687   0.02711   0.16705   1.45565
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.843e+00  1.007e-01  18.289  < 2e-16 ***
## F.Undergrad -3.277e-05  5.414e-06  -6.053 2.58e-09 ***
## Outstate     3.913e-05  5.186e-06   7.546 1.79e-13 ***
## PhD         -4.836e-03  1.439e-03  -3.360 0.000832 ***
## S.F.Ratio   -1.119e-02  3.611e-03  -3.098 0.002045 **
## Grad.Rate    1.706e-03  8.468e-04   2.015 0.044380 *
## Apps        -3.052e-05  9.952e-06  -3.067 0.002264 **
## Room.Board   4.105e-05  1.395e-05   2.942 0.003396 **
## Terminal    -3.209e-03  1.589e-03  -2.019 0.043967 *
## perc.alumni  1.985e-03  1.194e-03   1.662 0.097086 .
## Top10perc    2.202e-03  1.031e-03   2.135 0.033202 *
## Accept       3.418e-05  1.785e-05   1.915 0.056056 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.07321078)
##
##     Null deviance: 109.37  on 581  degrees of freedom
## Residual deviance:  41.73  on 570  degrees of freedom
## AIC: 143.93
##
## Number of Fisher Scoring iterations: 2
```

```
library(pROC)

college_training$Private = as.numeric(college_training$Private)
formula = Private ~ F.Undergrad + Outstate + PhD + S.F.Ratio + Grad.Ra
te + Apps + Accept + Room.Board + Top10perc + Terminal + perc.alumni
# Make predictions on the test dataset
college_model = glm(formula = formula, family = "gaussian", data = col
lege_training)
summary(college_model)
```

```
##
## Call:
## glm(formula = formula, family = "gaussian", data = college_training
)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.84585  -0.14687   0.02711   0.16705   1.45565
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.843e+00  1.007e-01  18.289  < 2e-16 ***
## F.Undergrad -3.277e-05  5.414e-06  -6.053 2.58e-09 ***
## Outstate     3.913e-05  5.186e-06   7.546 1.79e-13 ***
## PhD         -4.836e-03  1.439e-03  -3.360 0.000832 ***
## S.F.Ratio   -1.119e-02  3.611e-03  -3.098 0.002045 **
## Grad.Rate    1.706e-03  8.468e-04   2.015 0.044380 *
## Apps        -3.052e-05  9.952e-06  -3.067 0.002264 **
## Accept       3.418e-05  1.785e-05   1.915 0.056056 .
## Room.Board   4.105e-05  1.395e-05   2.942 0.003396 **
## Top10perc    2.202e-03  1.031e-03   2.135 0.033202 *
## Terminal    -3.209e-03  1.589e-03  -2.019 0.043967 *
## perc.alumni  1.985e-03  1.194e-03   1.662 0.097086 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.07321078)
##
##     Null deviance: 109.37  on 581  degrees of freedom
## Residual deviance:  41.73  on 570  degrees of freedom
## AIC: 143.93
##
## Number of Fisher Scoring iterations: 2
```

```
college_training$prop = predict(college_model,type="response")

college_training$pred = ifelse(college_training$prop>= mean(college_tr
aining$prop),2,1)

# matrix

matrix = table(college_training$Private,college_training$pred)

#accuracy

accuracy = mean(college_training$Private == college_training$pred)
print(paste("accuracy", accuracy))
```

```
## [1] "accuracy 0.852233676975945"
```

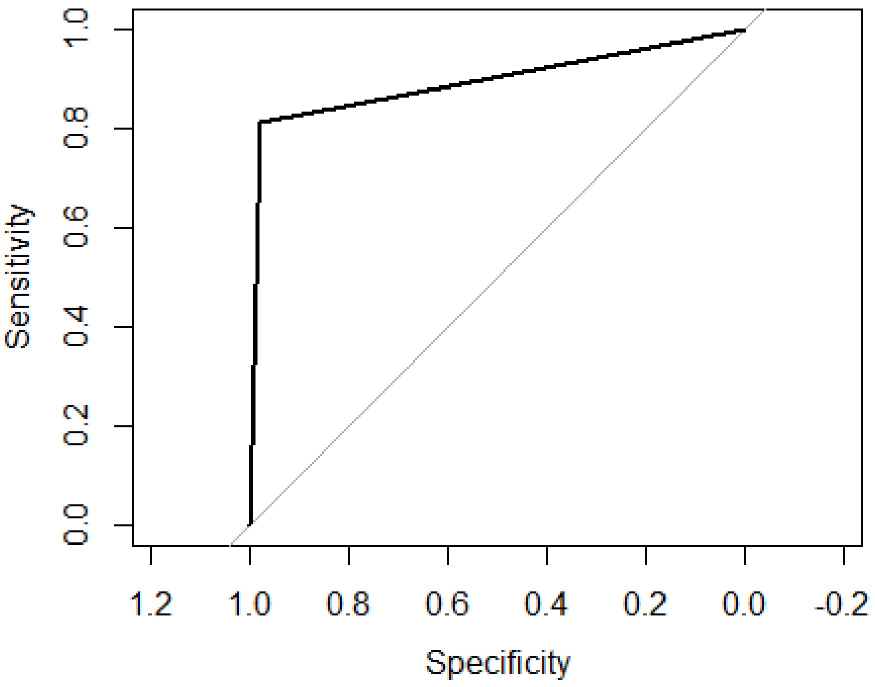
```
##precision - type I error rate - PPV

precision = matrix[1,1] / (matrix[1,2]+matrix[1,1])
print(paste("precision", precision))
```

```
## [1] "precision 0.979452054794521"
```

```
#specificity  - TNR

specificity= matrix[2,2] / (matrix[2,2] + matrix[1,2])

print(paste("specificity", specificity))
```

```
## [1] "specificity 0.991573033707865"
```

```
#recall  - type II error rate - TPR
recall =matrix[1,1] / (matrix[1,2]+matrix[1,1])
print(paste("recall", recall))

## [1] "recall 0.979452054794521"

#sensitivity

#ROC
Roc = roc(college_training$Private,college_training$pred)

## Setting levels: control = 1, case = 2

## Setting direction: controls < cases

plot(Roc,colors="red")
```

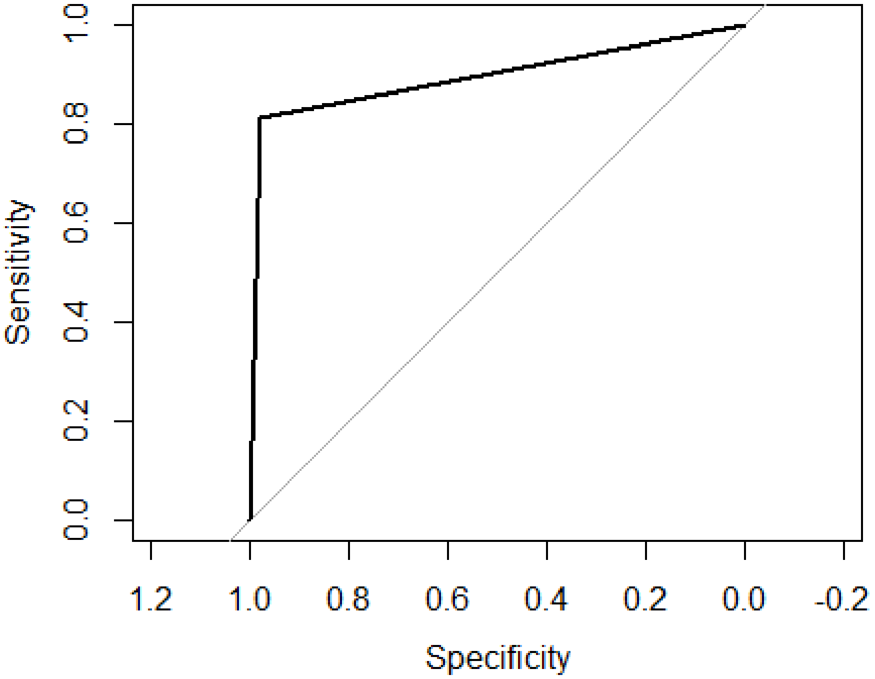


```
#AUR

auc(Roc)

## Area under the curve: 0.8945

matrix

##
##        1   2
##   1 143   3
##   2  83 353
```

```
library(pROC)

college_training$Private = as.numeric(college_training$Private)
formula = Private ~ F.Undergrad + Outstate + PhD + S.F.Ratio + Grad.Ra
te + Apps + Accept + Room.Board + Top10perc + Terminal + perc.alumni
# Make predictions on the test dataset
college_model = glm(formula = formula, family = "gaussian", data = col
lege_training)
summary(college_model)
```

```
##
## Call:
## glm(formula = formula, family = "gaussian", data = college_training
)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.84585  -0.14687   0.02711   0.16705   1.45565
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.843e+00  1.007e-01  18.289  < 2e-16 ***
## F.Undergrad -3.277e-05  5.414e-06  -6.053 2.58e-09 ***
## Outstate     3.913e-05  5.186e-06   7.546 1.79e-13 ***
## PhD         -4.836e-03  1.439e-03  -3.360 0.000832 ***
## S.F.Ratio   -1.119e-02  3.611e-03  -3.098 0.002045 **
## Grad.Rate    1.706e-03  8.468e-04   2.015 0.044380 *
## Apps        -3.052e-05  9.952e-06  -3.067 0.002264 **
## Accept       3.418e-05  1.785e-05   1.915 0.056056 .
## Room.Board   4.105e-05  1.395e-05   2.942 0.003396 **
## Top10perc    2.202e-03  1.031e-03   2.135 0.033202 *
## Terminal    -3.209e-03  1.589e-03  -2.019 0.043967 *
## perc.alumni  1.985e-03  1.194e-03   1.662 0.097086 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.07321078)
##
##     Null deviance: 109.37  on 581  degrees of freedom
## Residual deviance:  41.73  on 570  degrees of freedom
## AIC: 143.93
##
## Number of Fisher Scoring iterations: 2
```

```
college_training$prop = predict(college_model,type="response")

college_training$pred = ifelse(college_training$prop>= mean(college_tr
aining$prop),2,1)

# matrix

matrix = table(college_training$Private,college_training$pred)

#accuracy

accuracy = mean(college_training$Private == college_training$pred)
print(paste("accuracy", accuracy))

## [1] "accuracy 0.852233676975945"

##precision - type I error rate - PPV

precision = matrix[1,1] / (matrix[1,2]+matrix[1,1])
print(paste("precision", precision))

## [1] "precision 0.979452054794521"

#specificity  - TNR

specificity= matrix[2,2] / (matrix[2,2] + matrix[1,2])

print(paste("specificity", specificity))

## [1] "specificity 0.991573033707865"

#recall  - type II error rate - TPR
recall =matrix[1,1] / (matrix[1,2]+matrix[1,1])
print(paste("recall", recall))

## [1] "recall 0.979452054794521"

#sensitivity

#ROC
Roc = roc(college_training$Private,college_training$pred)

## Setting levels: control = 1, case = 2

## Setting direction: controls < cases

plot(Roc,colors="red")
```



```
#AUR

auc(Roc)

## Area under the curve: 0.8945

matrix

##
##         1   2
##    1 143   3
##    2  83 353

#test_data

null_model = glm(as.numeric(Private)~1,data=college_test,family="gauss
ian")
full_model = glm(as.numeric(Private)~.,data=college_test,family="gauss
ian")
step_model = step(null_model,scope =list(lower=null_model,upper=full_m
odel),direction="forward")

## Start:  AIC=265.56
## as.numeric(Private) ~ 1
##
##                  Df Deviance    AIC
## + F.Undergrad  1   27.186 175.18
## + Outstate     1   28.285 182.91
## + P.Undergrad  1   29.891 193.68
## + Enroll       1   30.064 194.80
## + S.F.Ratio    1   32.660 210.95
## + Accept       1   33.143 213.81
## + perc.alumni  1   33.285 214.65
## + Apps         1   33.334 214.94
## + Grad.Rate    1   37.162 236.13
## + Room.Board   1   38.119 241.09
## + Personal     1   39.565 248.35
## + Expend       1   40.388 252.37
## + Top10perc    1   42.154 260.71
## + Terminal     1   42.545 262.51
## + PhD          1   42.634 262.92
## + Top25perc    1   42.906 264.16
## <none>             43.662 265.56
## + Books        1   43.661 267.56
##
## Step:  AIC=175.18
## as.numeric(Private) ~ F.Undergrad
##
##               Df Deviance     AIC
## + Outstate     1   18.085  97.691
## + S.F.Ratio    1   21.539 131.777
## + perc.alumni  1   22.132 137.070
## + Grad.Rate    1   22.607 141.211
## + Expend       1   23.911 152.145
## + Room.Board   1   23.991 152.797
## + Top25perc    1   24.606 157.733
## + Top10perc    1   24.845 159.618
## + P.Undergrad  1   25.453 164.334
## + Enroll       1   25.616 165.582
## + Accept       1   25.729 166.439
## + Personal     1   26.064 168.963
## + Apps         1   26.854 174.783
## <none>             27.186 175.175
## + Books        1   26.923 175.284
## + PhD          1   27.139 176.840
## + Terminal     1   27.175 177.101
##
## Step:  AIC=97.69
## as.numeric(Private) ~ F.Undergrad + Outstate
##
##               Df Deviance    AIC
## + Terminal     1   16.092 76.924
## + PhD          1   16.145 77.572
## + S.F.Ratio    1   17.423 92.419
## + Apps         1   17.601 94.397
## + perc.alumni  1   17.743 95.966
## + P.Undergrad  1   17.805 96.644
## <none>             18.085 97.691
## + Grad.Rate    1   17.910 97.791
## + Expend       1   17.917 97.875
## + Room.Board   1   17.953 98.258
## + Top10perc    1   17.958 98.319
## + Enroll       1   18.009 98.869
## + Personal     1   18.041 99.212
## + Top25perc    1   18.067 99.496
## + Books        1   18.082 99.665
## + Accept       1   18.083 99.670
##
## Step:  AIC=76.92
## as.numeric(Private) ~ F.Undergrad + Outstate + Terminal
##
##               Df Deviance    AIC
## + perc.alumni  1   15.467 71.205
## + S.F.Ratio    1   15.559 72.355
## + PhD          1   15.826 75.673
## + P.Undergrad  1   15.861 76.110
## + Apps         1   15.869 76.198
## <none>             16.092 76.924
## + Personal     1   15.959 77.309
## + Grad.Rate    1   15.980 77.568
## + Expend       1   16.044 78.339
## + Top25perc    1   16.047 78.374
## + Enroll       1   16.055 78.480
## + Books        1   16.077 78.737
## + Accept       1   16.091 78.913
## + Room.Board   1   16.092 78.920
## + Top10perc    1   16.092 78.923
##
## Step:  AIC=71.2
## as.numeric(Private) ~ F.Undergrad + Outstate + Terminal + perc.alum
ni
##
##               Df Deviance    AIC
## + S.F.Ratio    1   14.994 67.138
## + PhD          1   15.192 69.701
## <none>             15.467 71.205
## + Apps         1   15.337 71.555
## + P.Undergrad  1   15.358 71.821
## + Personal     1   15.408 72.450
## + Expend       1   15.412 72.505
## + Enroll       1   15.434 72.780
## + Accept       1   15.446 72.931
## + Books        1   15.447 72.951
## + Top10perc    1   15.451 73.001
## + Grad.Rate    1   15.456 73.057
## + Room.Board   1   15.459 73.097
## + Top25perc    1   15.466 73.180
##
## Step:  AIC=67.14
## as.numeric(Private) ~ F.Undergrad + Outstate + Terminal + perc.alum
ni +
##     S.F.Ratio
##
##                Df Deviance    AIC
## + Expend       1   14.714 65.466
## + PhD          1   14.752 65.963
## <none>             14.994 67.138
## + P.Undergrad  1   14.861 67.398
## + Personal     1   14.876 67.599
## + Apps         1   14.878 67.628
## + Enroll       1   14.906 68.000
## + Accept       1   14.916 68.122
## + Top10perc    1   14.931 68.328
## + Books        1   14.956 68.646
## + Grad.Rate    1   14.975 68.896
## + Room.Board   1   14.991 69.100
## + Top25perc    1   14.992 69.110
##
## Step:  AIC=65.47
## as.numeric(Private) ~ F.Undergrad + Outstate + Terminal + perc.alum
ni +
##     S.F.Ratio + Expend
##
##                Df Deviance    AIC
## + PhD          1   14.502 64.643
## <none>             14.714 65.466
## + P.Undergrad  1   14.590 65.816
## + Personal     1   14.590 65.820
## + Accept       1   14.621 66.228
## + Enroll       1   14.641 66.494
## + Books        1   14.662 66.774
## + Apps         1   14.671 66.897
## + Room.Board   1   14.687 67.109
## + Grad.Rate    1   14.689 67.131
## + Top10perc    1   14.705 67.350
## + Top25perc    1   14.711 67.433
## 
## Step:  AIC=64.64
## as.numeric(Private) ~ F.Undergrad + Outstate + Terminal + perc.alum
ni + 
##     S.F.Ratio + Expend + PhD
## 
##               Df Deviance    AIC
## <none>            14.502 64.643
## + Personal     1   14.393 65.163
## + P.Undergrad  1   14.402 65.292
## + Accept       1   14.405 65.333
## + Grad.Rate    1   14.422 65.559
## + Enroll       1   14.440 65.801
## + Books        1   14.447 65.900
## + Apps         1   14.474 66.259
## + Top25perc    1   14.475 66.275
## + Room.Board   1   14.479 66.323
## + Top10perc    1   14.502 66.637
```

```
summary(step_model)
```

```
## 
## Call:
## glm(formula = as.numeric(Private) ~ F.Undergrad + Outstate + 
##     Terminal + perc.alumni + S.F.Ratio + Expend + PhD, family = "ga
ussian", 
##     data = college_test)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -0.95626  -0.16238   0.02058   0.16048   0.66703  
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  2.074e+00  1.607e-01  12.906  < 2e-16 ***
## F.Undergrad -2.976e-05  4.561e-06  -6.525 6.19e-10 ***
## Outstate     5.818e-05  8.287e-06   7.021 3.95e-11 ***
## Terminal    -4.881e-03  2.307e-03  -2.116  0.03566 *  
## perc.alumni  5.779e-03  2.122e-03   2.723  0.00708 ** 
## S.F.Ratio   -1.895e-02  6.577e-03  -2.881  0.00443 ** 
## Expend      -9.764e-06  5.448e-06  -1.792  0.07475 .  
## PhD         -3.461e-03  2.096e-03  -1.652  0.10031    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for gaussian family taken to be 0.07755302)
##
##     Null deviance: 43.662  on 194  degrees of freedom
## Residual deviance: 14.502  on 187  degrees of freedom
## AIC: 64.643
##
## Number of Fisher Scoring iterations: 2
```

The confusion matrix gives us more information about the classification of model. The meaning of false positive is that there is a non-private university, but it detects private university, the meaning of false negative is that there is the private university but it detects non-private university. The damage of misclassification to model depend on the purpose of the researcher. For example, we are mainly finding the private university, if the false negative increases, we are risky to lose private universities. However, in most cases, the cost of false negative is higher than the cost of false positive because I can say that I would be more so doubtful rather than so ignorant to get the huge consequence.

Next, we will find the good model for test set.

```
college_test$Private = as.numeric(college_test$Private)
formula = Private~F.Undergrad + Outstate + Terminal + P.Undergrad + Ap
ps
# Make predictions on the test dataset
college_model = glm(formula = formula, family = "gaussian", data = col
lege_test)
summary(college_model)
```

```
##
## Call:
## glm(formula = formula, family = "gaussian", data = college_test)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.11198  -0.17770   0.03734   0.18628   0.69516
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.732e+00  1.182e-01  14.653  < 2e-16 ***
## F.Undergrad -1.521e-05  9.823e-06  -1.549   0.1231    
## Outstate     6.820e-05  6.257e-06  10.898  < 2e-16 ***
## Terminal    -7.196e-03  1.610e-03  -4.470 1.35e-05 ***
## P.Undergrad -4.827e-05  2.529e-05  -1.908   0.0579 .  
## Apps        -2.404e-05  1.275e-05  -1.885   0.0609 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 0.08237391)
## 
##     Null deviance: 43.662  on 194  degrees of freedom
## Residual deviance: 15.569  on 189  degrees of freedom
## AIC: 74.477
## 
## Number of Fisher Scoring iterations: 2
```

```
college_test$prop = predict(college_model,type="response")

college_test$pred = ifelse(college_test$prop>= mean(college_test$prop)
,2,1)

# matrix

matrix = table(college_test$Private,college_test$pred)

##precision - type I error rate - PPV

precision = matrix[1,1] / (matrix[1,2]+matrix[1,1])
print(paste("precision", precision))
```

```
## [1] "precision 0.939393939393939"
```

```
#specificity  - TNR

specificity= matrix[2,2] / (matrix[2,2] + matrix[1,2])

print(paste("specificity", specificity))
```

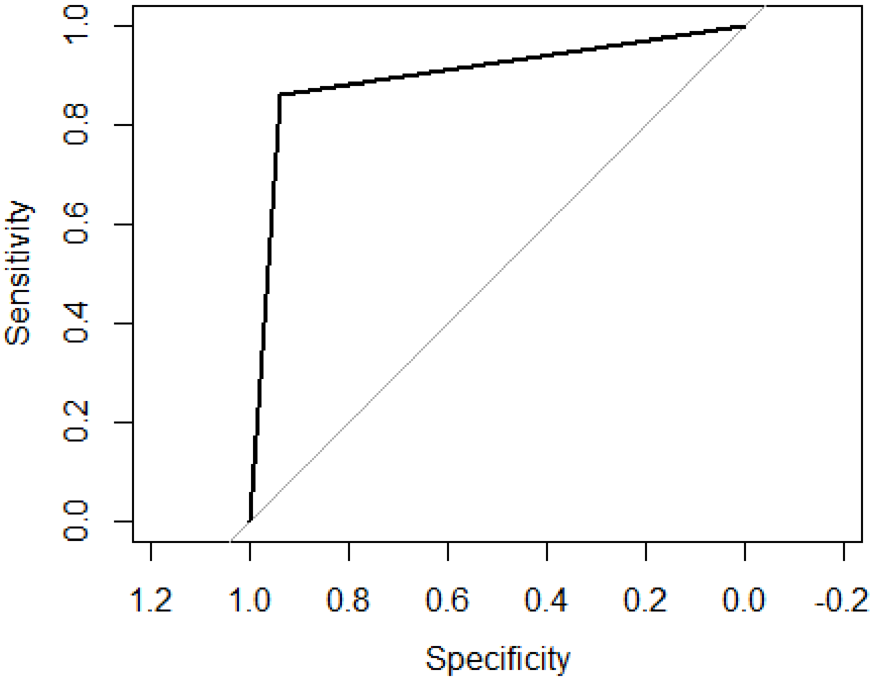
```
## [1] "specificity 0.965217391304348"
```

```
#recall  - type II error rate - TPR
recall =matrix[1,1] / (matrix[1,2]+matrix[1,1])
print(paste("recall", recall))
```

```
## [1] "recall 0.939393939393939"
```

```
#sensitivity
```

```
#accuracy

accuracy = mean(college_test$Private == college_test$pred)
print(paste("accuracy", accuracy))

## [1] "accuracy 0.887179487179487"

#ROC
Roc = roc(college_test$Private,college_test$pred)

## Setting levels: control = 1, case = 2

## Setting direction: controls < cases

plot(Roc,colors="red")
```



```
#AUR

auc(Roc)

## Area under the curve: 0.8999
```

Conclusion:

From what we explored from College data, we know that the graduation rate from private university is better than non-private university, while the low graduation rate of both types of university is significantly large compared to the general. I was wrong when I have thought the high acceptance rate will be associated with the good rate of graduation, but there is no relation between these two variables. For the classification, the model has given me the good model to distinguish the private and non-private university. For training set, I just need to take these variables to predict the outcome, with the lowest AIC 120.51, it means this model is the most accuracy and simplest one for using. The AUC for this training set is nearly 90%.

```
F.Undergrad + Outstate + PhD + S.F.Ratio + Grad.Rate + Apps +
Room.Board + Terminal + perc.alumni + Top10perc + Accept
```

For test set, we have these variables to predict the outcome, with the lowest AIC 46.53. The AUC for this test set is nearly 90% as well.

```
F.Undergrad + Outstate + Terminal + P.Undergrad + Apps
```

Reference:

Precision and recall. Retrieved from https://en.wikipedia.org/wiki/Precision_and_recall#:~:text=Recall%20in%20this%20context%20is%20also%20referred%20to%20as%20the,rate%20is%20also%20called%20specificity.

Retrieved from https://www.rdocumentation.org/packages/ROCR/versions/1.0-11/topics/prediction

Narkhede. S (2015). Understanding Confusion Matrix. Retrieved from https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62