title: “LAB 2” author: “Nehal Ur Rahman” date: “2023-01-24” output: word_document

#NAME: NEHAL UR RAHMAN #STUDENT ID: 991691259

##Introduction #In Lab 2 We will be analyzing a College dataset and performing various functions to understand the variables. We will also be creating a new variable and provide visualizations.

```
#Import
#Here we are going to load and read all variables with its values from the
college1 dataset.
college <- read.csv("college1.csv")
head(college)
```

```
##                              X Private Apps Accept Enroll Top10perc
Top25perc
## 1 Abilene Christian University     Yes 1660   1232    721        23
52
## 2           Adelphi University     Yes 2186   1924    512        16
29
## 3               Adrian College     Yes 1428   1097    336        22
50
## 4          Agnes Scott College     Yes  417    349    137        60
89
## 5    Alaska Pacific University     Yes  193    146     55        16
44
## 6            Albertson College     Yes  587    479    158        38
62
##   F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD Terminal
## 1        2885         537     7440       3300   450     2200  70       78
## 2        2683        1227    12280       6450   750     1500  29       30
## 3        1036          99    11250       3750   400     1165  53       66
## 4         510          63    12960       5450   450      875  92       97
## 5         249         869     7560       4120   800     1500  76       72
## 6         678          41    13500       3335   500      675  67       73
##   S.F.Ratio perc.alumni Expend Grad.Rate
## 1      18.1          12   7041        60
## 2      12.2          16  10527        56
## 3      12.9          30   8735        54
## 4       7.7          37  19016        59
## 5      11.9           2  10922        15
## 6       9.4          11   9727        55
```

```
#The fix function is used to fix the 1st column and not store it as data as
they are just labels
fix(college)
```

```
#Here we fix the dataset by adding a column called row.names which records
the name of all the universities
```

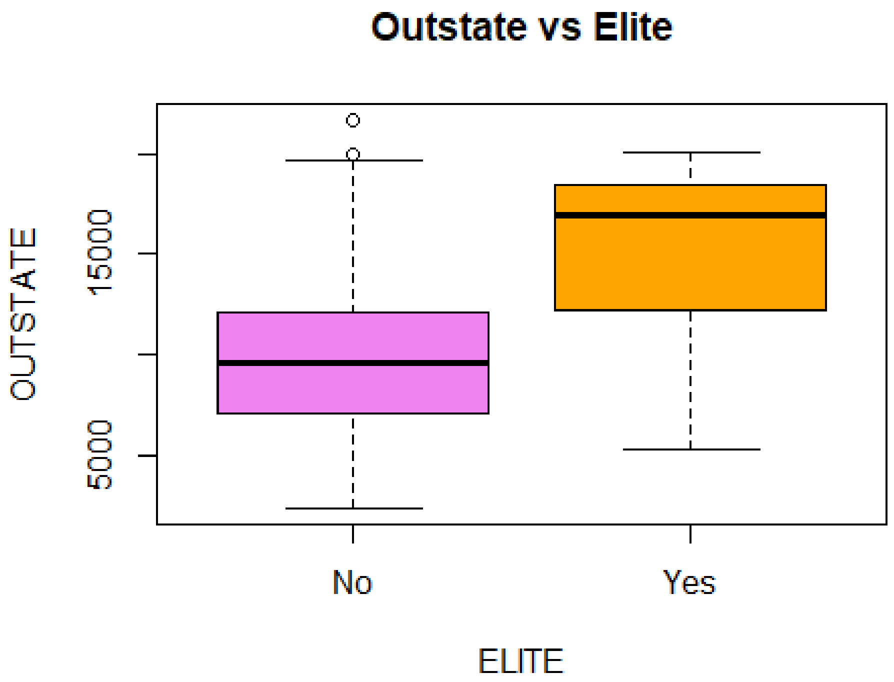
```
row.names(college) = college[,1]
fix(college)

#Using the function given below we delete the 1st column in the college
dataset as it is not required in our analysis.
college = college[,-1]
fix(college)

#The as.factor function converts the character variable(Private) to vector
with numerical values
college$Private<-as.factor(college$Private)
```

###Question 1

```
#First we create a variable by using the function rep() which replicates the
college with 777 number of rows with a value of "No"
Elite <- rep("No",nrow(college))
#Now we record the values as "Yes" in the Elite column with a condition that
the proportion of students coming from the top 10% of their high school
exceeds 50%.
Elite[college$Top10perc >50] <- "Yes"
#The as.factor function displays the variable(Elite) as vector with
levels(Yes & No)
Elite <- as.factor(Elite)
#Now we create a dataframe with college and elite
college <- data.frame(college , Elite)
TotalEliteSchools <- length(college$Elite[college$Elite=="Yes"])
#We then calculate the total number of Elite colleges and display the number.
message("The total number of Elite Schools are : ", TotalEliteSchools)

## The total number of Elite Schools are : 78
```

###Question 2

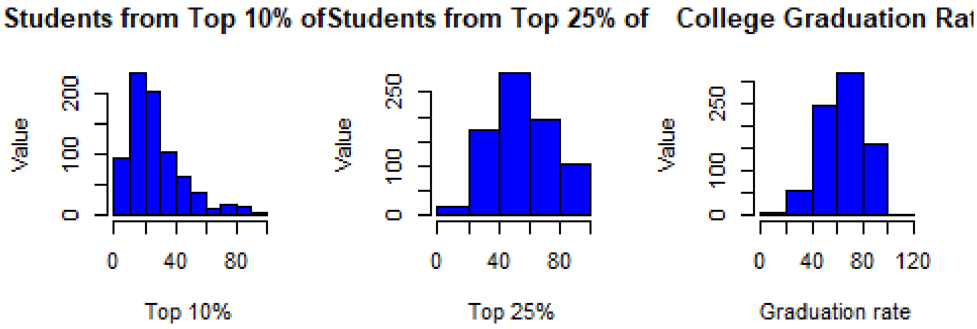
```
#The summary function is used here to get details of the Elite column
summary(college$Elite)

##  No Yes
## 699  78

#A side by side boxplot of Outstate Vs Elite is created using plot function
plot(college$Outstate ~ college$Elite, col = c("violet", "orange"),
xlab="ELITE",ylab="OUTSTATE", main = "Outstate vs Elite", border = "black")
```

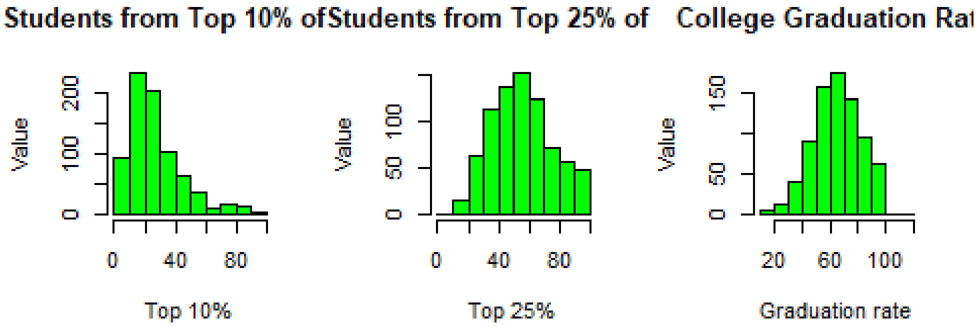


#Boxplot: From the boxplot we can see that the number of Elite colleges are more in the Outstate.

###Question 3

```
#The hist() function is used here to produce histograms with variable number of bins for 3 of the quantitative variables like Top10perc, Top25perc and Grad.Rate.
#breaks = 6 & 8 is assigned first which gives lesser number of bins
#breaks = 12 is assigned next to get more number of bins.

#The par() function divides the frame into the required number to display the histograms within one window.
par(mfcol=c(2,3))

hist(college$Top10perc, col = "blue",breaks=8, xlab = "Top 10%", ylab = "Value", main="Students from Top 10% of H.S")
hist(college$Top10perc, col = "green",breaks=12, xlab = "Top 10%", ylab = "Value", main="Students from Top 10% of H.S")

hist(college$Top25perc, col = "blue",breaks= 6, xlab = "Top 25%", ylab = "Value", main="Students from Top 25% of H.S")
hist(college$Top25perc, col = "green",breaks=12, xlab = "Top 25%", ylab = "Value", main="Students from Top 25% of H.S")

hist(college$Grad.Rate, col = "blue",breaks=6, xlab = "Graduation rate", ylab = "Value", main="College Graduation Rate")
hist(college$Grad.Rate, col = "green",breaks=12, xlab = "Graduation rate",
ylab = "Value", main="College Graduation Rate")
```





#The histogram for the 3 variables: Top10perc, Top25perc and Grad.Rate have been displayed with different number of bins.