Mini Project 2

Names of Group members: Sashidhar Donthiri and Sathya Pooja Ramireddy

Contributions:

1. a) >data=csv.read(“C:/users/shash/Downloads/college.csv”) #to read and store the csv document using Data editor

   b) >fix(data) #fix()-> used to read display the read file

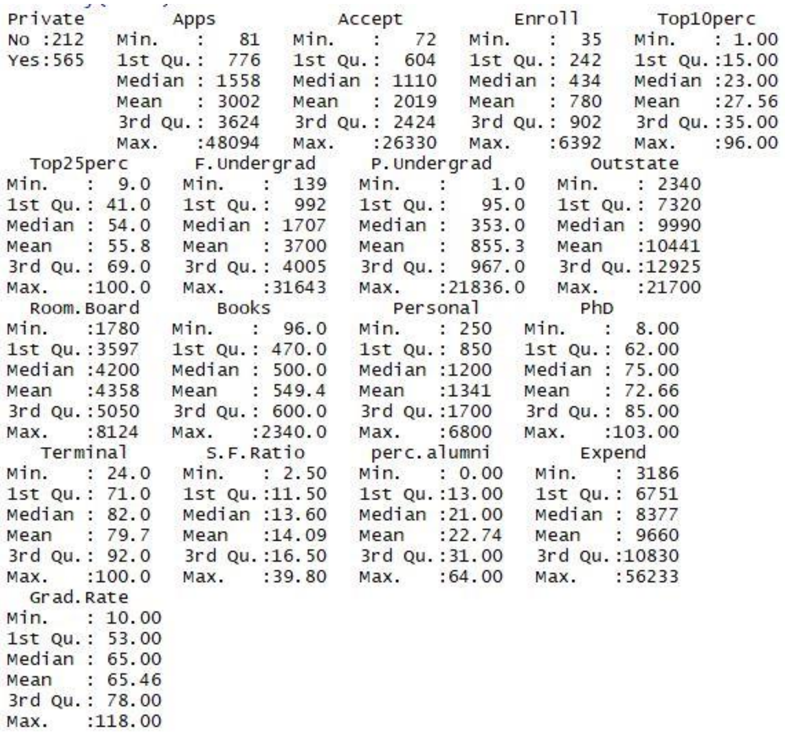
   >rownames(data)=data[,x] #x->column number #to rename any column as per #need

   >fix(data) #to display the file after renaming

   >data= data[,-1] #Since the first column just details about the name of the #university, its of no logical use for data analysis:

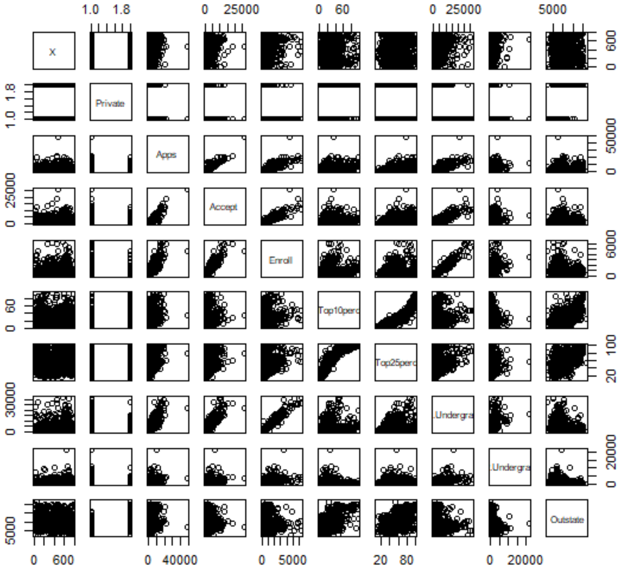
   >fix(data) #to display the edited document

   c) (i) >summary(data) #numerical summary of data(for each of the variables #in the data set; 5 point Summary of each variable in data set

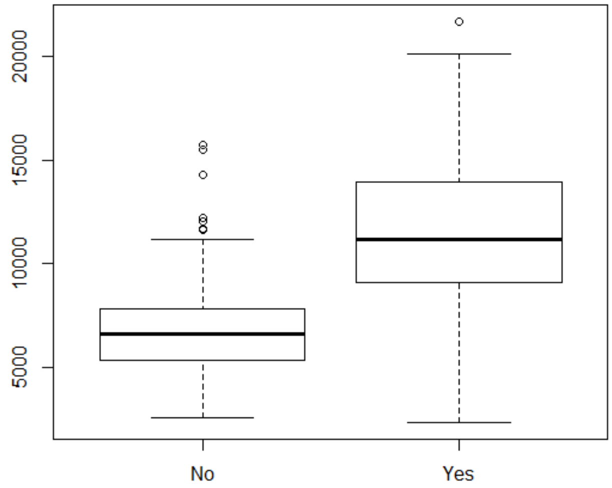


```
 Private        Apps           Accept          Enroll        Top10perc
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
           Median : 1558   Median : 1110   Median : 434   Median :23.00
           Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
           3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
           Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
   Top25perc      F.Undergrad     P.Undergrad         Outstate
 Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
 1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
 Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
 Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
 3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
 Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
   Room.Board       Books           Personal         PhD
 Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
 1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
 Median :4200   Median : 500.0   Median :1200   Median : 75.00
 Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
 3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
 Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
    Terminal       S.F.Ratio      perc.alumni        Expend
 Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
 1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
 Median : 82.0   Median :13.60   Median :21.00   Median : 8377
 Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
 3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
 Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
   Grad.Rate
 Min.   : 10.00
 1st Qu.: 53.00
 Median : 65.00
 Mean   : 65.46
 3rd Qu.: 78.00
 Max.   :118.00
```

(ii) >pairs(data[,a:b]) #a,b>0 the *ij*th scatterplot contains x[,i] plotted against x[,j]. The #scatterplot can be customised by setting panel functions to appear as something completely different.



(iii) >plot(data$Private,data$Outstate) #side by side box plot for comparison



The above boxplots are for Private and Outstate in the dataset.

Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution.

In the former case one wishes to discard them or use statistics that are robust to outliers, while in the latter case they indicate that the distribution has high skewness and that one should be very cautious in using tools or intuitions that assume a normal distribution.

In case of private schools/universities- NO, looking at the range between Q4(extreme max data point and box plot max datapoint), (box plot max point and median)Q3 , Q2(median and boxplot min point), Q1(box plot min point to min data point), the IQR between (Q1,Q4)and (Q2,Q3)pairs is almost symmetrical indicating the normal distribution. Here, student's fees follows approximately normal distribution with around mean=median~= 7000. The IQR is not large indicating less variance in the data.

On the other-hand, Private schools-YES in data just one outlier which could be an error in data, and when strictly observed median<mean indicating that the distribution is not normal and skewed i.e.; In a right skewed bell curve, most data points fall to the left of the middle, there are more exceptionally small than exceptionally large values on the right. Most data points fall in the Q2, there are more exceptionally small number of students with fees above 11000 than exceptionally large number of students with fees under 11000. Also, the whiskers are spread over a large range and IQR range is more indicating more variance.

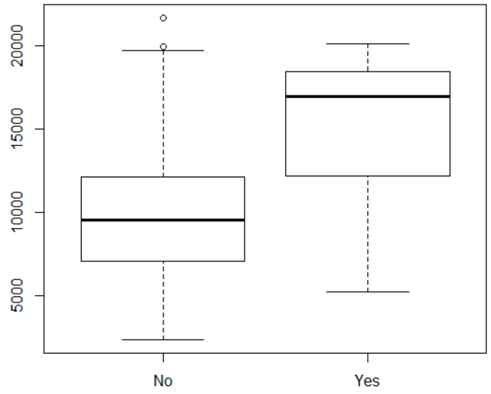
(iv) #divide universities into two groups based on whether or not the proportion of students coming from the top #10% of their high school classes exceeds 50% by binning Top10perc Variable in data and introducing a new qualitative variable.

```
>check=rep("No",nrow(data))
>check[data$Top10perc > 50 ]= "Yes"
>check=as.factor(check)
>check=data.frame(data,Elite)
>summary(data) 5 point Summary of each variable in data set
```

RESULTS:

```
check
No : 699
Yes: 78
```

```
>plot(data$check,data$outstate)
```

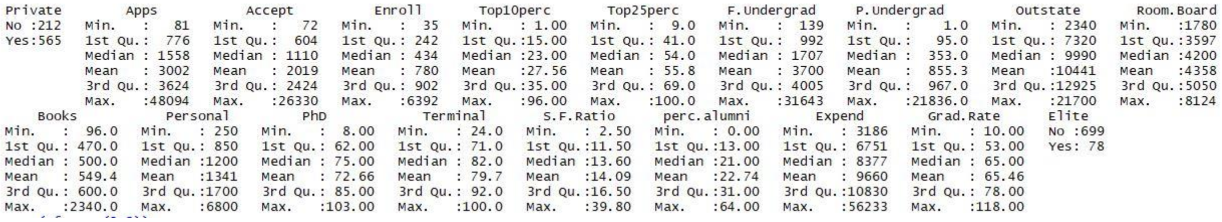


Check indicates whether or not the proportion of students coming from the top #10% of their high school classes exceeds 50% by binning Top10perc Variable in data. From above boxplots, NO indicates that though the IQR is large, the outliers are observed and median =mean which means the distribution of data points for NO is normal. The distribution is approximately normal, IQR range indicates comparably less and symmetric variance.

On the other hand, the YES is left skewed that means, most data points fall to the right of the middle, there are more exceptionally large than exceptionally small values on the left. More students pay fees whose value falls to the right of the middle, there are more exceptionally large than exceptionally small values. Proportion of students coming from the top #10% of their high school classes exceeds 50% by binning Top10perc Variable, pay more fees.

Also, the IQR in YES indicates that it has comparably more variance than the NO.

>summary(data)   #5 point Summary of each variable in data set



```
Private        Apps           Accept          Enroll        Top10perc       Top25perc      F.Undergrad     P.Undergrad        Outstate      Room.Board
No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780
Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597
          Median : 1558   Median : 1110   Median : 434   Median :23.00   Median : 54.0   Median : 1707   Median :  353.0   Median : 9990   Median :4200
          Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358
          3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050
          Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124
    Books           Personal         PhD            Terminal       S.F.Ratio      perc.alumni        Expend        Grad.Rate      Elite
Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00   No :699
1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00   Yes: 78
Median : 500.0   Median :1200   Median : 75.00   Median : 82.0   Median :13.60   Median :21.00   Median : 8377   Median : 65.00
Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
```

(v) >par(mfrow=c(2,2)) #to display multiple graphs In a single graph
>hist(data$Apps) #histogram of Apps
>hist(data$perc.alumni,col=2) #histogram of perc.alumni with color of bars chosen to be 2
>hist(data$S.F.Ratio) #histogram of S.F.Ratio
>hist(data$Apps,breaks=100) #histogram of Apps that are divided into 100 breaks over data

> hist(data$Accept,breaks=100)   #hist(data$variable_name,breaks=k,col=s)   k->divide the data set into k #separate datasets further for, s->determines the color over s

> hist(data$Enroll,breaks=100)

> hist(data$Top10perc,breaks=100)

> hist(data$Top25perc,breaks=100)

> hist(data$F.Undergrad,breaks=100)

> hist(data$P.Undergrad,breaks=100)

> hist(data$Outstate,breaks=100)

> hist(data$Room.Board,breaks=100)

> hist(data$Books,breaks=100)

> hist(data$Personal,breaks=100)

> hist(data$PhD,breaks=100)

> hist(data$Terminal,breaks=100)

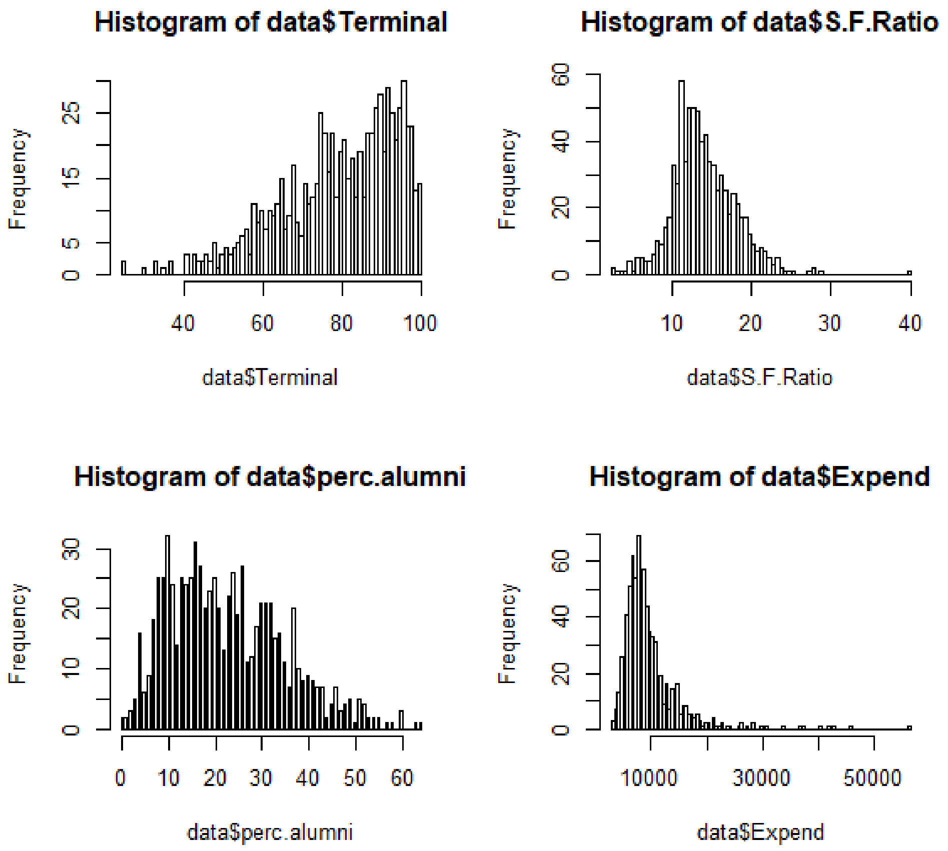
> hist(data$S.F.Ratio,breaks=100)

> hist(data$perc.alumni,breaks=100)

> hist(data$Expend,breaks=100)

> hist(data$Grad.Rate,breaks=100)

> hist(data$Elite,breaks=100)



Above graphs, histogram1 represents a graph to be left skewed and histogram2,3,4 represents a graph of distribution that is right skewed.

In a right skewed distribution, most data points fall to the left of the middle, there are more exceptionally small than exceptionally large values and Median<mean.

In a left skewed distribution, most data points fall to the right of the middle, there are exceptionally large than exceptionally small values and mean<median.

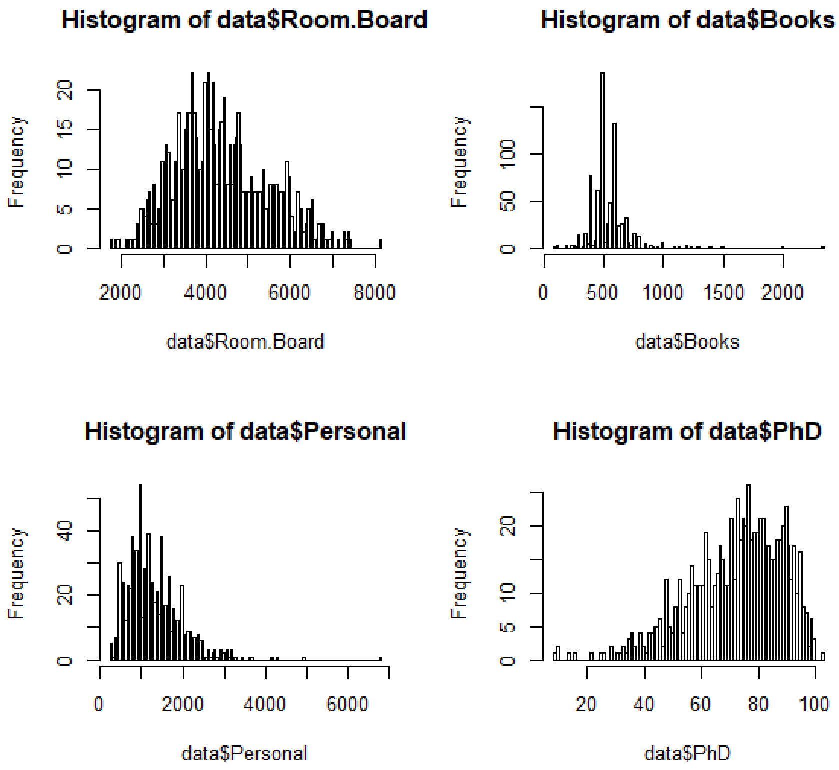
Observations:

Histogram 1: Terminal has more data points on left of median.

Histogram 2: More Student faculty ratio data points on the right of median than exceptionally large values.

Histogram 3: More percentage of alumni data points on right of median than exceptionally large values.

Histogram 4: More expends of datapoints on right of median than exceptionally large values.



In the above graph, histogram1,2,3 represents a distribution that is right symmetric and histogram 4 represents a distribution that is left symmetric.

In a right skewed distribution, most data points fall to the left of the middle, there are more exceptionally small than exceptionally large values and Median<mean.

In a left skewed distribution, most data points fall to the right of the middle, there are exceptionally large than exceptionally small values and mean<median.

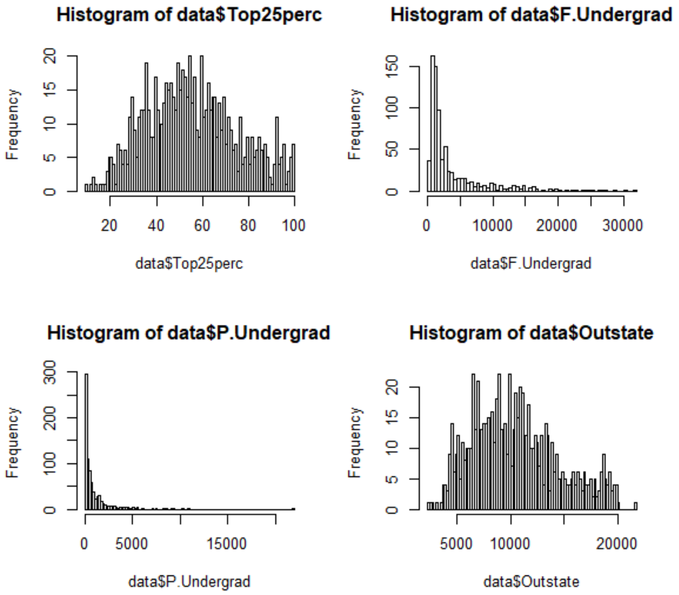
Observations:

Histogram 1: Roomboard is right skewed i.e.; it has more data points on the right of median than exceptionally large values.

Histogram 2:Books has more data points on the right of median than exceptionally large values.

Histogram 3: Personal has more data points on the right of median than exceptionally large values.

Histogram 4: PhD strength has more data points on left of median.



In the above graph, histogram 2,3 represent left skewed and careful observation of mean of histogram1 shows that it is left skewed, where as histogram 4 represents right skewed.

In a right skewed distribution, most data points fall to the left of the middle, there are more exceptionally small than exceptionally large values and Median<mean.

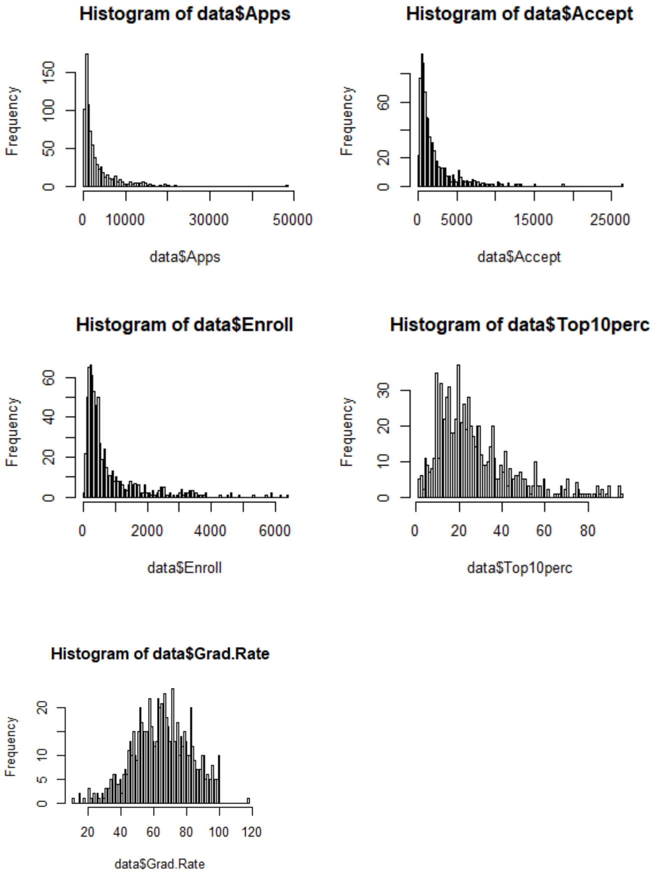
In a left skewed distribution, most data points fall to the right of the middle, there are exceptionally large than exceptionally small values and mean<median.

Histogram 1: Top25perc from high school has more data points on left of median.

Histogram 2: F.Undergrad has more data points on left of median.

Histogram 3: P.Undergrad has more data points on left of median.

Histogram 4: Outstate is right skewed, it has more data points on the right of median than exceptionally large values.



Observations:

Histogram 1:Number of applications has distribution Right skewed i.e.; it has more data points on the right of median than exceptionally large values.

Histogram 2: Acceptance number has distribution Right skewed i.e.; it has more data points on the right of median than exceptionally large values.

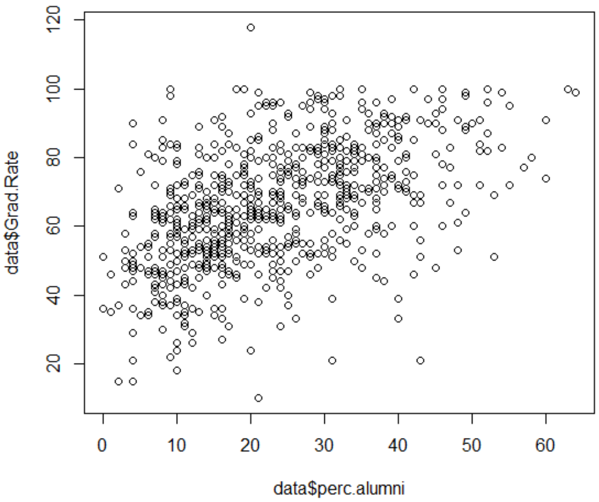
Histogram 3: Students Enrolled number has distribution Right skewed i.e.; it has more data points on the right of median than exceptionally large values.

Histogram 4: Top 10 perc has distribution Right skewed i.e.; it has more data points on the right of median than exceptionally large values.

Histogram 5: Grad rate has distribution Left skewed i.e.;. it has more data points on the right of median than exceptionally large values

(vi)

```
>Par(mfrow(1,1)) #used to display (nxn) graphs in a single  graph.
```

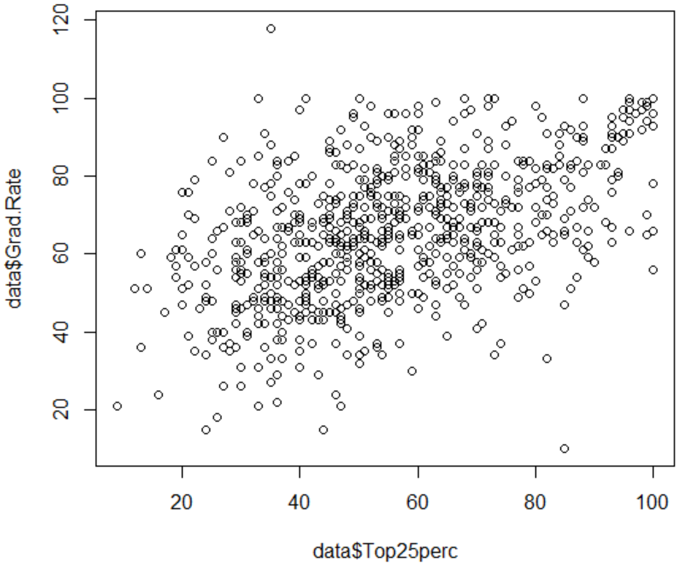


Observation:

```
>Plot(data$perc.alumni,data$Grad.Rate)
```

Grad rate is dependent on number of successive grads over total students enrolled.

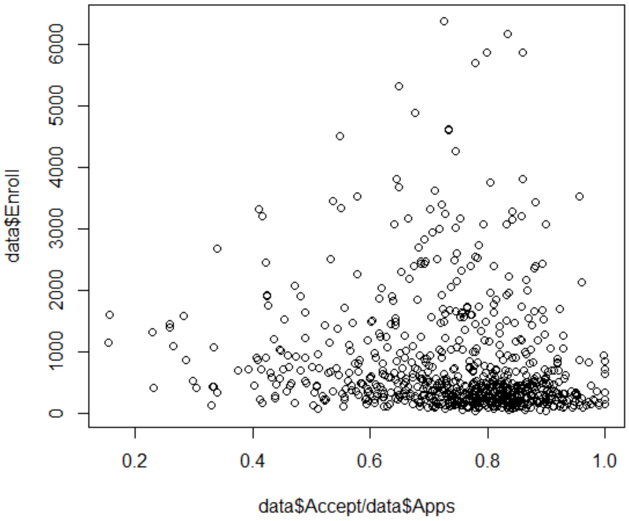
The scatter plot is drawn between Perc.alumni vs Grad.Rate, and shows that percentage of alumni doesn't necessarily have to do anything with highest graduation rate for the colleges.



Graph: Top25perc in high school vs GradRate.

```
>plot(data$Top25perc, data$Grad.Rate)
```

Observations: Grad Rate is dependent on the number of students who graduated over number of students who enrolled. From the graph, the Colleges with the most students from Top 25% don't necessarily have to highest graduation rate.

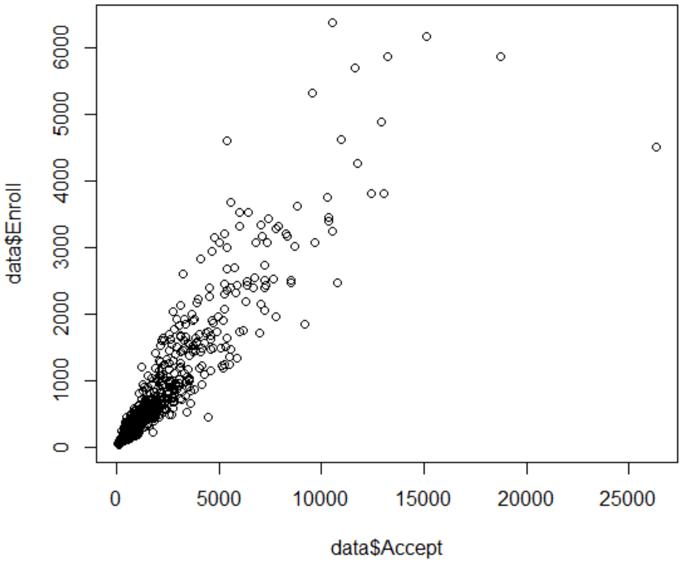


Graph: acceptance rate vs enroll

>plot(data$accept/data$Apps,data$Enroll)

Observation:

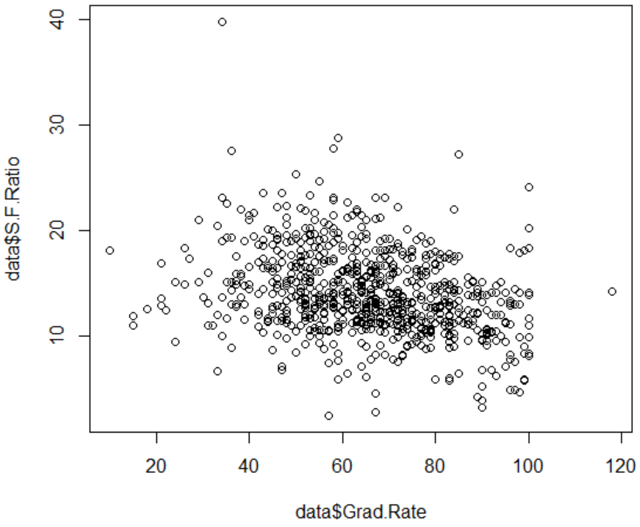
From the graph, it can be inferred that more students have enrolled once they have received their acceptance from colleges. The higher the acceptance rate, the lower the number of students enrolled.



Graph: Accepted vs Enrolled

>plot(data$Accept, data$Enroll)

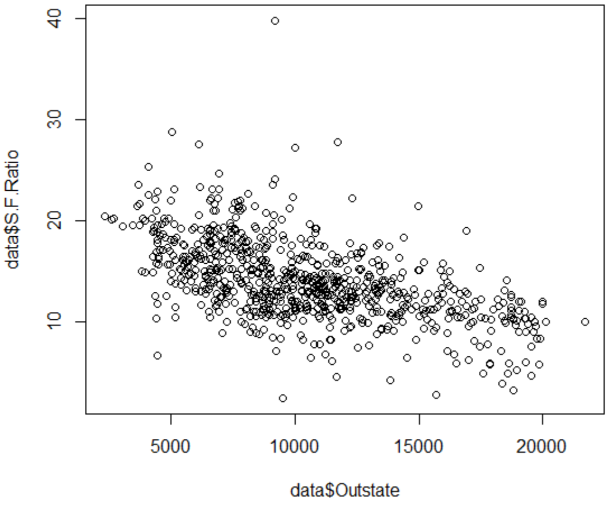
Observation: From the above graph, it can be inferred that in the colleges, as the number of applications accepted are less, eventually the number of students enrolled are less.



Graph: Grad.Rate vs S.F.Rate

>plot(data$S.F.Rate, data$Grad.Rate)

From the above graph, it can be inferred that there is not any correlation between the Student faculty ratio and the graduation rate. Grad.Rate does not have to do anything with the S.F.Ratio.



Graph: Outstate vs S.F.Rate

>plot(data$outstate, data$S.F.Ratio)

From the above graph, it can be inferred that there is not any correlation between the outstate and the S.F.Ratio. Outstate fees does not have to do anything with the S.F.Ratio.