**Mini Project 2**

**Name :**

**Dhwaniben Kaneria (drk170130)**

**Utkarsh Gandhi (usg170030)**

**Contribution of team members:**

Dhwani:

* Wrote R code for different plots
* Worked on all questions
* Wrote explanation for (a) and (b) part

Utkarsh:

* Wrote R code for different plots
* Worked on all questions
* Wrote explanation for (c)

**Ans 1:**

1. **# R - code**

file="College.csv"; # taking file name into a object

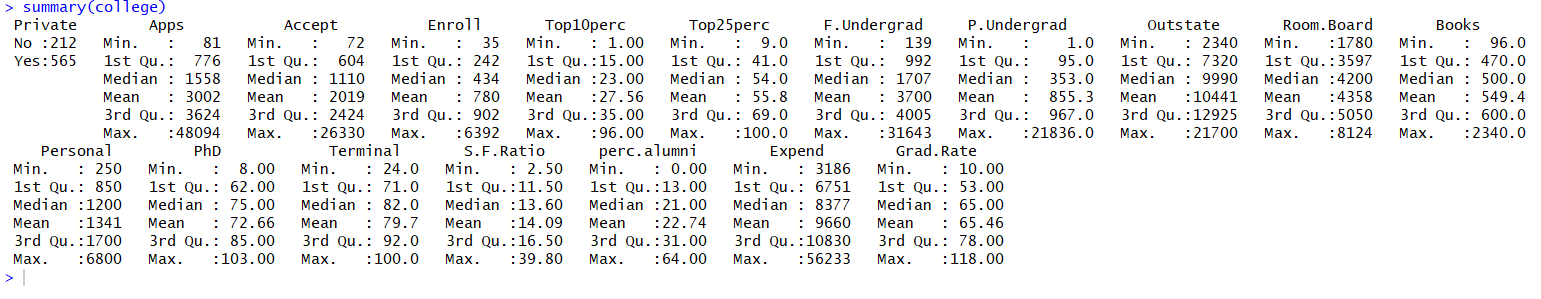
college=read.csv(file); # reading csv file into a dataframe

**(b) # R - code**

1. file="College.csv"; # taking file name into a object
2. college=read.csv(file); # reading csv file into dataframe
3. fix(college); # invoked edit on college dataframe
4. rownames(college)=college[,1]; #created a row.names column which contains the #name of each university
5. fix(college)
6. college=college[,-1]; # Removed the first column from the college dataframe
7. fix(college) # Now we will be able to see the updated the dataframe

**(c)**

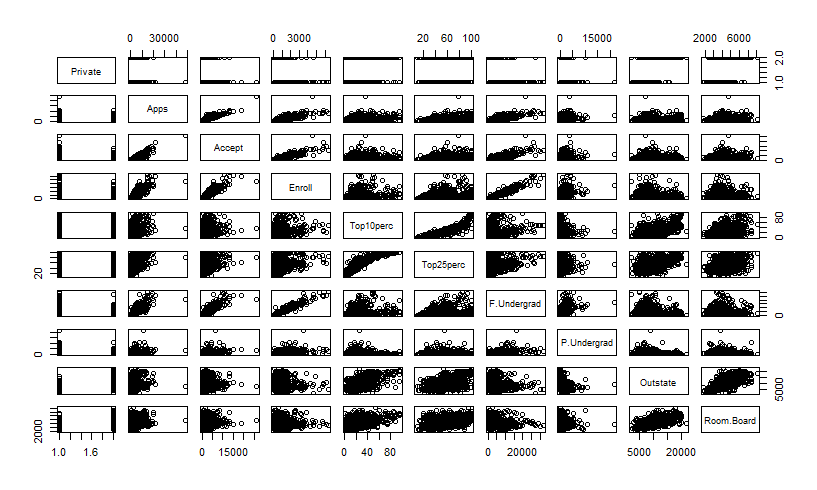
**i.** summary(college) #gives numerical summary of each attribute of the college data. If attribute is categorical, it will just give count for each category and for numerical attributes, it will output 5 points summary and the mean.



**ii**. pairs(college[ ,1:10 ])

**Explanation:**

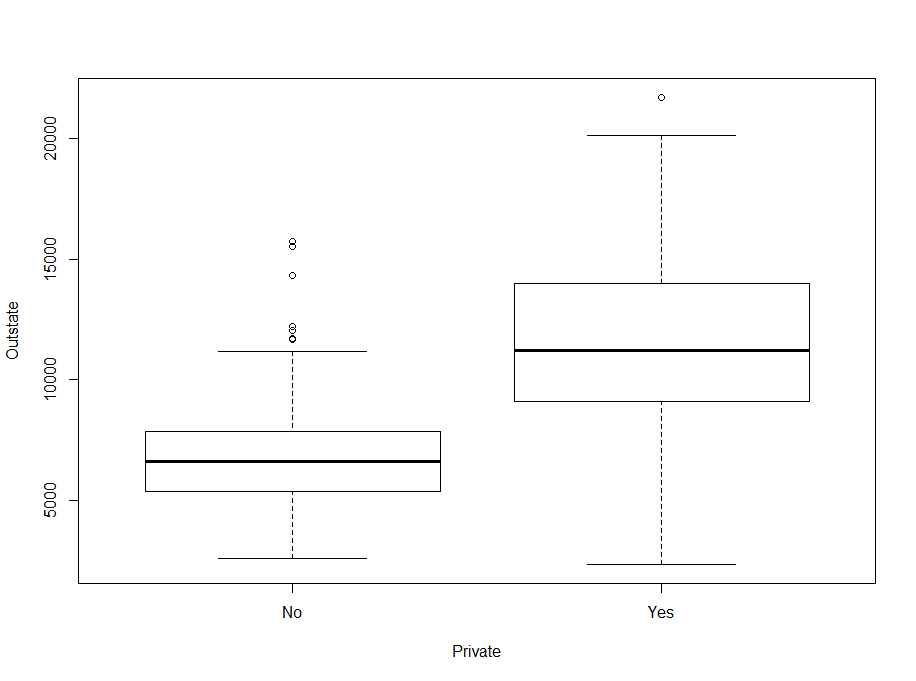
* scatter plot matrix of first ten columns. It will show correlation between multiple attributes.
* In our case there are 10 attributes so it will show correlation between every combination of any two attributes.
* If the plot between any two attributes follows a straight line if the slope of the line is positive that means the correlation is positive between the attributes as the correlation will be greater than 0. If the slope is negative and the plot follows a straight line then the correlation is negative. If the plot does not follows a straight line the correlation will be 0.
* For the college dataset, we can see that Enroll and Full Time undergrad are positively correlated. So when enrollment increase in particular college, number of full time undergraduates will be also increase.
* For attribute Top 25 percent of high school class and Room board cost, there is no correlation between them as the plot doesn’t follows a straight line.



**iii**. plot(college[ ,"Private"],college[ ,"Outstate"],xlab="Private",ylab="Outstate") # plotting boxplot side by side

**Explanation:**

* Below is the side-by-side boxplot of Private and Outstate attributes of the college dataset.
* From the boxplot, we can conclude that if the university/college is Private then the out-state tuition is high compared to universities/colleges that are not private.
* We can also see that universities which are private have more variation in outstate tuition.



**iv.**

**#R - code**

1. Elite=rep("No",nrow(college)) # Initialise Elite dataframe with “No” and having rows as college dataframe contains.
2. Elite[college$Top10perc>50]="Yes" # If classes exceeded 50 % with top 10% of students.
3. Elite=as.factor(Elite) # It will convert a vector into a factor.
4. college=data.frame(college,Elite) # It will append Elite factor to college dataframe.
5. summary(college$Elite) # It will give the summary of Elite column, which is categorical.

**Output:**

No Yes

699 78

**Explanation :**

* The output states that there are 78 universities which has more than 50% students from top 10% high school class and 699 universities which has less than 50% students from top 10% high school class.

**v.**

**#R - code**

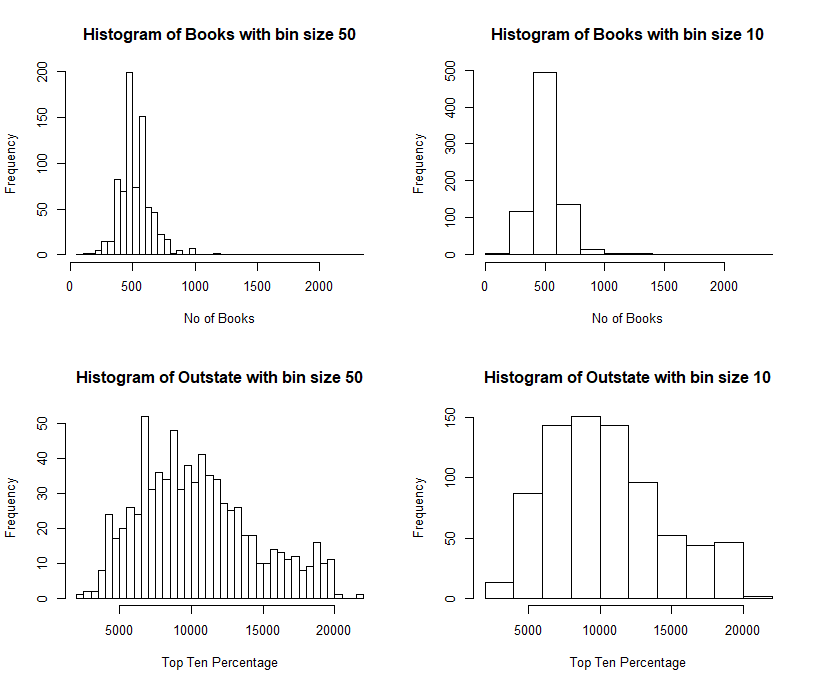
#histograms with differing numbers of bins for the quantitative variables (Books and Outstate)

1. par(mfrow=c(2,2)) # dividing print window into 4 regions

# plotting histogram for different values of break parameter and different attribute of college dataset

1. hist(college$Books,breaks = 50,main="Histogram of Books with bin size 50", xlab="No of Books")
2. hist(college$Books,breaks = 10,main="Histogram of Books with bin size 10", xlab="No of Books")
3. hist(college$Outstate,breaks = 50,main="Histogram of Outstate with bin size 50", xlab="Top Ten Percentage")
4. hist(college$Outstate,breaks = 10,main="Histogram of Outstate with bin size 10", xlab="Top Ten Percentage")

**Output:**

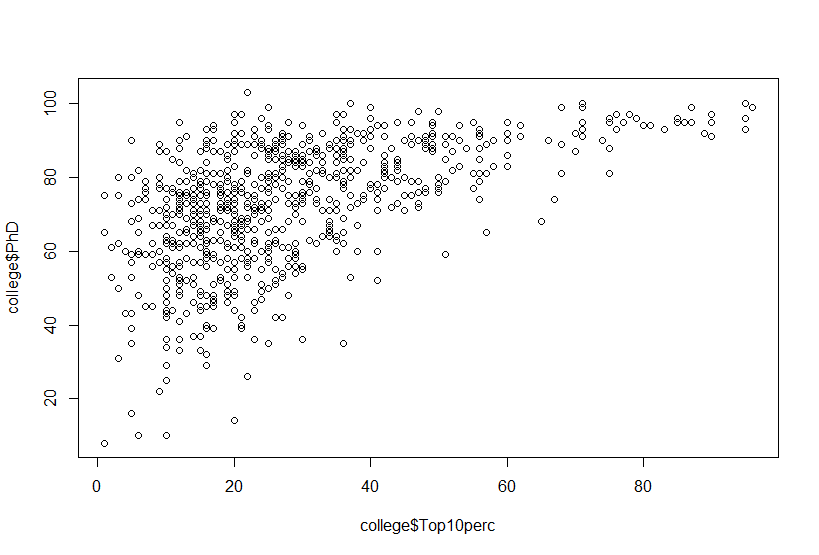


**vi.**

**Following are the conclusions by exploring the data more:**

1. plot(college$Top10perc,college$PhD)

* It will show the relation between top 10% of students in high school class and percentage of faculty with Ph.D.
* From scatter plot, we can conclude that the top 10% of high school students tends to attract more to the universities/colleges having more Ph.D. faculties.

****

**2) Plotting histogram and QQnorm for Grad Rate**

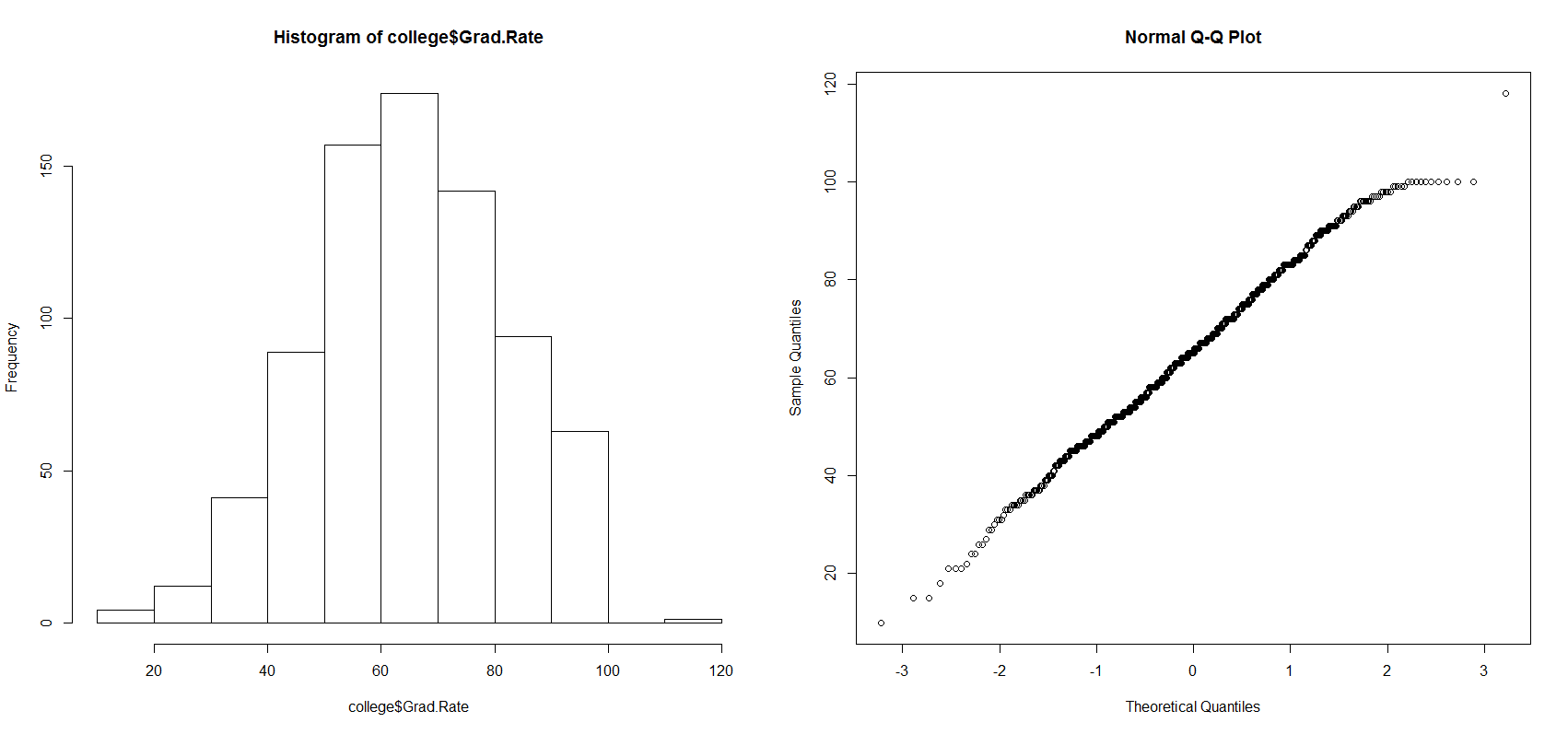
**# R-code**

par(mfrow=c(1,2))

hist(college$Grad.Rate)

qqnorm(college$Grad.Rate)

We can see from the histogram and QQ plot that the distribution is normal so we can say that Graduate rate is normal all over universities.

****

**3) Plot between S.F ratio and Outstate tuition**

par(mfrow=c(1,1))

plot(college$Outstate,college$S.F.Ratio)

**Explanation:**

As we can see from the scatter plot the attributes are negatively correlated , i.e. when Outstate tuition is higher then Student/Faculty ratio is low which concludes that faculties are more or students are less if the tuition is higher.

