Statistical Methods for Data Science

Mini Project 2

Members:

SHIVA RANGA CHAWALA sxc167630

KRISHNA SINDHU KOTA kxk171030

Contribution:

Shiva Ranga Chawala – Equally contributed

Krishna Sindhu Kota – Equally contributed

8. This exercise relates to the College data set, which can be found in the file College.csv. It contains a number of variables for 777 different universities and colleges in the US.

a) Use the read.csv() function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

College = read.csv("College.csv", header=TRUE, sep=",")
#this command is used to read the data from csv file and store it in College.

Read.csv is used to read the data from College.csv file and stores the loaded data into College. Attribute header = TRUE specifies that the data has header in the first row and sep=”,” specifies that data is separated by commas.

b) Look at the data using the fix() function.

rownames(College) = College[,1]
#A name is assigned to each row by using rownames command corresponding to the appropriate university names.
fix(College)
#fix invokes edit on college and assigns the new version of college

Data Editor

File Edit Help

	row.names	X	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate
1	Abilene Christian University	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440
2	Adelphi University	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280
3	Adrian College	Adrian College	Yes	1428	1097	336	22	50	1036	99	11250
4	Agnes Scott College	Agnes Scott College	Yes	417	349	137	60	89	510	63	12960
5	Alaska Pacific University	Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560
6	Albertson College	Albertson College	Yes	587	479	158	38	62	678	41	13500
7	Albertus Magnus College	Albertus Magnus College	Yes	353	340	103	17	45	416	230	13290
8	Albion College	Albion College	Yes	1899	1720	489	37	68	1594	32	13868
9	Albright College	Albright College	Yes	1038	839	227	30	63	973	306	15595
10	Alderson-Broaddus College	Alderson-Broaddus College	Yes	582	498	172	21	44	799	78	10468
11	Alfred University	Alfred University	Yes	1732	1425	472	37	75	1830	110	16548
12	Allegheny College	Allegheny College	Yes	2652	1900	484	44	77	1707	44	17080
13	Allentown Coll. of St. Francis de Sales	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64	1130	638	9690
14	Alma College	Alma College	Yes	1267	1080	385	44	73	1306	28	12572
15	Alverno College	Alverno College	Yes	494	313	157	23	46	1317	1235	8352
16	American International College	American International College	Yes	1420	1093	220	9	22	1018	287	8700
17	Amherst College	Amherst College	Yes	4302	992	418	83	96	1593	5	19760
18	Anderson University	Anderson University	Yes	1216	908	423	19	40	1819	281	10100
19	Andrews University	Andrews University	Yes	1130	704	322	14	23	1586	326	9996
20	Angelo State University	Angelo State University	No	3540	2001	1016	24	54	4190	1512	5130
21	Antioch University	Antioch University	Yes	713	661	252	25	44	712	23	15476
22	Appalachian State University	Appalachian State University	No	7313	4664	1910	20	63	9940	1035	6806
23	Aquinas College	Aquinas College	Yes	619	516	219	20	51	1251	767	11208
24	Arizona State University Main campus	Arizona State University Main campus	No	12809	10308	3761	24	49	22593	7585	7434
25	Arkansas College (Lyon College)	Arkansas College (Lyon College)	Yes	708	334	166	46	74	530	182	8644
26	Arkansas Tech University	Arkansas Tech University	No	1734	1729	951	12	52	3602	939	3460
27	Assumption College	Assumption College	Yes	2135	1700	491	23	59	1708	689	12000
28	Auburn University-Main Campus	Auburn University-Main Campus	No	7548	6791	3070	25	57	16262	1716	6300
29	Augsburg College	Augsburg College	Yes	662	513	257	12	30	2074	726	11902
30	Augustana College IL	Augustana College IL	Yes	1879	1658	497	36	69	1950	38	13353
31	Augustana College	Augustana College	Yes	761	725	306	21	58	1337	300	10990
32	Austin College	Austin College	Yes	948	798	295	42	74	1120	15	11280
33	Averett College	Averett College	Yes	627	556	172	16	40	777	538	9925
34	Baker University	Baker University	Yes	602	483	206	21	47	958	466	8620
35	Baldwin-Wallace College	Baldwin-Wallace College	Yes	1690	1366	662	30	61	2718	1460	10995
36	Barat College	Barat College	Yes	261	192	111	15	36	453	266	9690

College = College[,-1]
#This command ignores the first column of the data and copies remaining data to College matrix.
fix(College)

File Edit Help

	row.names	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board
1	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300
2	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450
3	Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750
4	Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450
5	Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120
6	Albertson College	Yes	587	479	158	38	62	678	41	13500	3335
7	Albertus Magnus College	Yes	353	340	103	17	45	416	230	13290	5720
8	Albion College	Yes	1899	1720	489	37	68	1594	32	13868	4826
9	Albright College	Yes	1038	839	227	30	63	973	306	15595	4400
10	Alderson-Broaddus College	Yes	582	498	172	21	44	799	78	10468	3380
11	Alfred University	Yes	1732	1425	472	37	75	1830	110	16548	5406
12	Allegheny College	Yes	2652	1900	484	44	77	1707	44	17080	4440
13	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64	1130	638	9690	4785
14	Alma College	Yes	1267	1080	385	44	73	1306	28	12572	4552
15	Alverno College	Yes	494	313	157	23	46	1317	1235	8352	3640
16	American International College	Yes	1420	1093	220	9	22	1018	287	8700	4780
17	Amherst College	Yes	4302	992	418	83	96	1593	5	19760	5300
18	Anderson University	Yes	1216	908	423	19	40	1819	281	10100	3520
19	Andrews University	Yes	1130	704	322	14	23	1586	326	9996	3090
20	Angelo State University	No	3540	2001	1016	24	54	4190	1512	5130	3592
21	Antioch University	Yes	713	661	252	25	44	712	23	15476	3336
22	Appalachian State University	No	7313	4664	1910	20	63	9940	1035	6806	2540
23	Aquinas College	Yes	619	516	219	20	51	1251	767	11208	4124
24	Arizona State University Main campus	No	12809	10308	3761	24	49	22593	7585	7434	4850
25	Arkansas College (Lyon College)	Yes	708	334	166	46	74	530	182	8644	3922
26	Arkansas Tech University	No	1734	1729	951	12	52	3602	939	3460	2650
27	Assumption College	Yes	2135	1700	491	23	59	1708	689	12000	5920
28	Auburn University-Main Campus	No	7548	6791	3070	25	57	16262	1716	6300	3933
29	Augsburg College	Yes	662	513	257	12	30	2074	726	11902	4372
30	Augustana College IL	Yes	1879	1658	497	36	69	1950	38	13353	4173
31	Augustana College	Yes	761	725	306	21	58	1337	300	10990	3244
32	Austin College	Yes	948	798	295	42	74	1120	15	11280	4342
33	Averett College	Yes	627	556	172	16	40	777	538	9925	4135
34	Baker University	Yes	602	483	206	21	47	958	466	8620	4100
35	Baldwin-Wallace College	Yes	1690	1366	662	30	61	2718	1460	10995	4410
36	Barat College	Yes	261	192	111	15	36	453	266	9690	4300

College[,1] fetches the first column of the college matrix for which rownames command assigns each row a name.
fix(college) invokes edit on college and assigns the new version of college.
College[,-1] ignores the first column of the data and copies remaining data to College matrix. As we can see from the above screen shot the first column of the data is Private now.

c) i. Use the summary() function to produce a numerical summary of the variables in the data set.

summary(College)
#provides a 5 point summary along with mean of each column

Summary command gives us 5 point summary(i.e Min, Q1, Median, Mean, Q3, Max) along with mean of each column present in College. Below are the screen shots.

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

```
     Expend        Grad.Rate      Elite
Min.   : 3186   Min.   : 10.00   No :699
1st Qu.: 6751   1st Qu.: 53.00   Yes: 78
Median : 8377   Median : 65.00
Mean   : 9660   Mean   : 65.46
3rd Qu.:10830   3rd Qu.: 78.00
Max.   :56233   Max.   :118.00
```

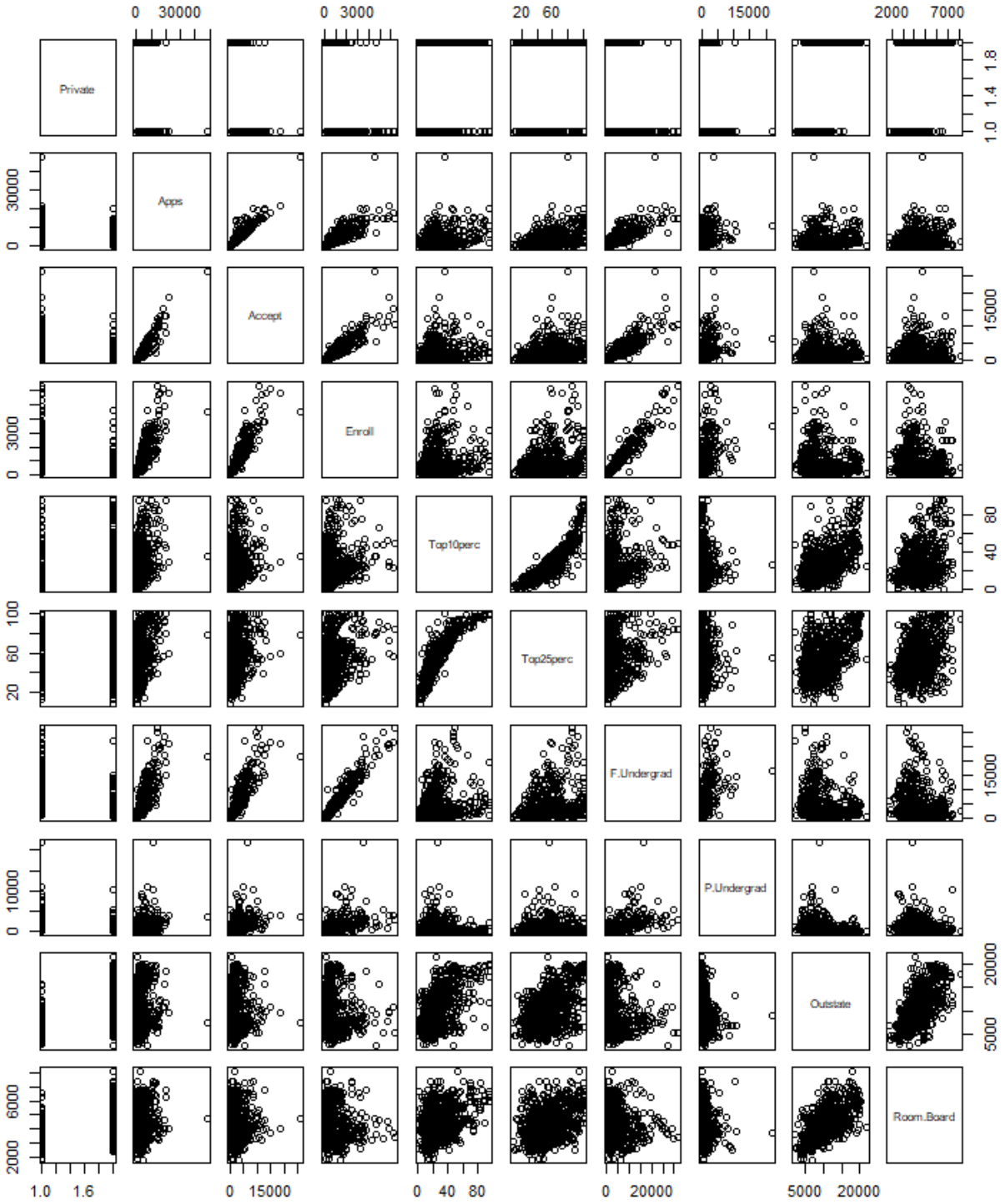
If we look at the summary of Apps (no. of applications received),

- we can interpret that 81 is minimum no. of applications received and this data corresponds to Christendom College

- maximum no. of applications is received by Rutgers at New Brunswick whose value is 48094
- Mean > Median so the data is right skewed.
- Inter Quartile Range (IQR) = (Q3-Q1) = 3624-776 = 2848
- Right whisker = min{max, Q3 + 1.5*IQR} = min{48094, 7896} = 7896
- Left whisker = max{min, Q1 – 1.5*IQR} = max{81, -3496} = 81
- As right whisker < max value we can say that there are outliers in the data.
- In the same way this 5 point summary gives results of Min, 1st Quartile, Median, Mean, 3rd Quartile, Max of all the columns present in the College matrix.

ii. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data.

pairs(College[,1:10])
#produces the scatter plot matrix of first 10 columns of the data.

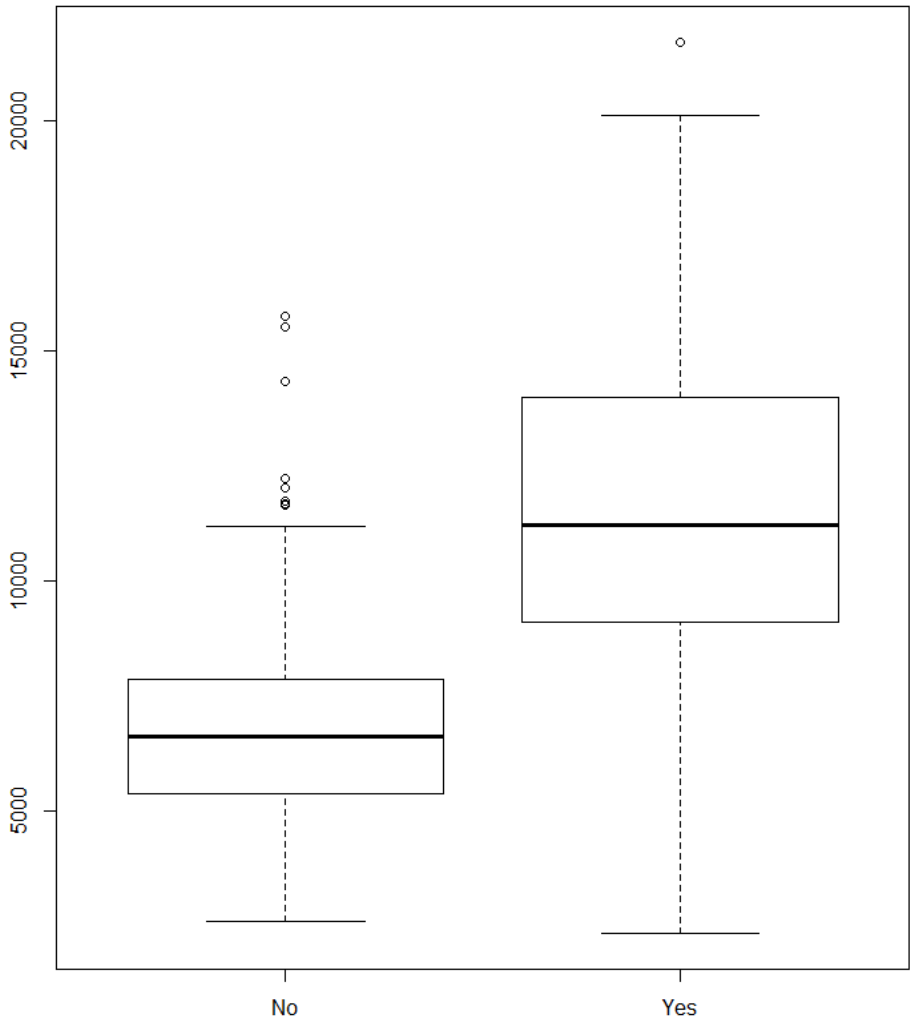


- Scatter plot represent how much one variable is dependent on another variable.
- In above scatterplot matrix the variables are written in a diagonal line from top left to bottom right. Then each variable is plotted against each other. For example, if we consider the box in the 3rd row 2nd column is an individual scatterplot of Apps and Accept, with Apps as the X-axis and Accept as the Y-axis. We can interpret that as the no. of applications received increases no. of applications accepted will also increase. This same plot is

replicated in 2nd row 3rd column with Accept on X-axis and Apps on Y-axis. In essence, the boxes on the upper right-hand side of the whole scatterplot are mirror images of the plots on the lower left hand.

iii. Use the plot () function to produce side-by-side boxplots of Outstate versus Private.

plot(college[,1], college[,9])
#generates a side-by-side boxplot of Outstate and Private columns.



The first box plot represents the 5 point summary/box plot of all the Out-of-state tuition data whose Private value is “NO” and second box plot represent the 5 point summary/box plot of all the Out-of-state tuition data whose Private value is “YES”.

Interpretation:

- Box plot 2 has more range than Box plot 1.
- Box plot 2 has more variance in data than Box plot 1.
- Box plot 1 is symmetric but Box plot 2 is slightly right skewed.
- Box plot 1 has more outlier than Box plot 2.
- IQR for Box plot 2 is higher than Box plot 1.
- Out of state tuition has more private indicators than public indicators.

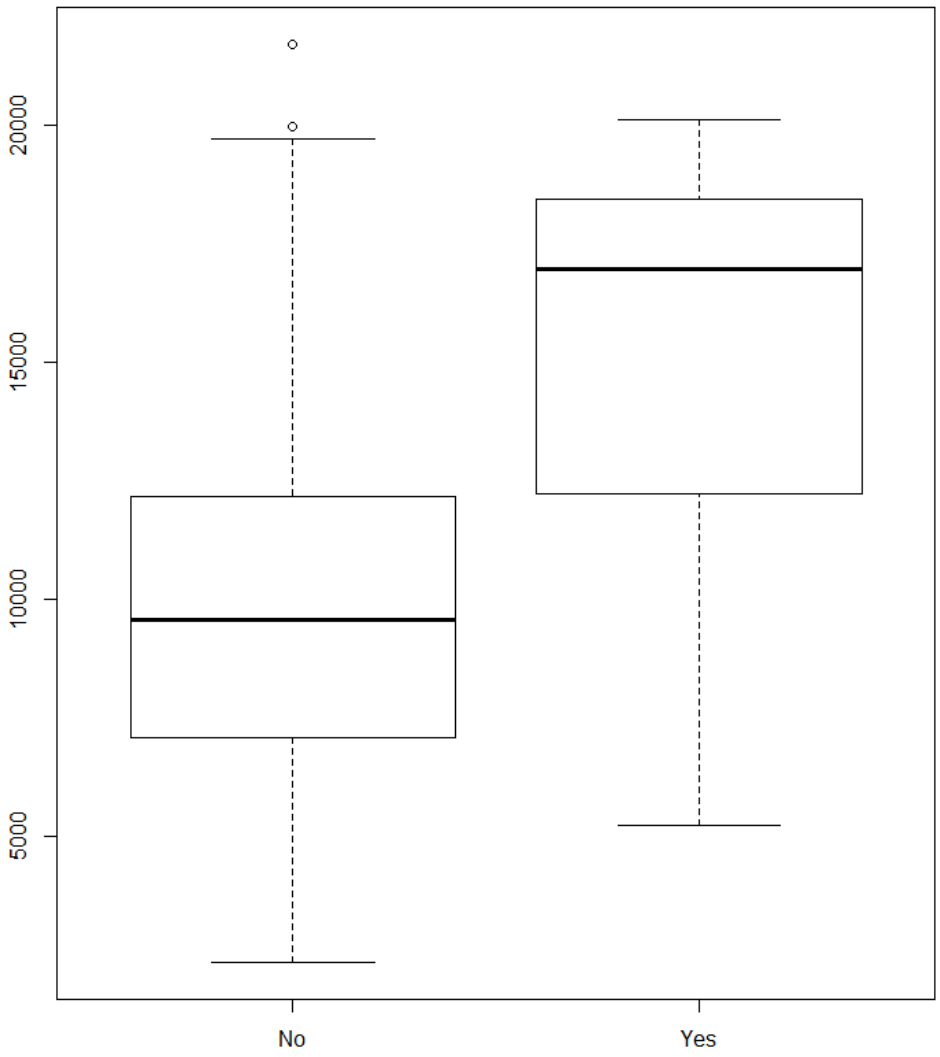
iv. Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

```
Elite = rep("No",nrow(College))
#replicate values "No" to the number of rows calculated by nrow command and assigns it to Elite variable
Elite[College$Top10perc >50] = "Yes"
 #assigns "Yes" to the rows that satisfies the condition of students coming from the top 10% of their classes exceeds 50%
Elite = as.factor(Elite)
#This is used to encode vector as a factor
College = data.frame(College ,Elite)
#It creates data frame with each column passed as a separate argument
summary(College)
plot(College$Elite, College$Outstate)
```

- Summary of college now produces output which includes elite column also.

```
   Grad.Rate      Elite
 Min.   : 10.00   No :699
 1st Qu.: 53.00   Yes: 78
 Median : 65.00
 Mean   : 65.46
 3rd Qu.: 78.00
 Max.   :118.00
> |
```

- From this data we can interpret that there are 78 values of Top10perc data whose percentage exceeds 50% and 699 values whose percentage is less than 50%.
- So, the proportion of students coming from the Top10perc of their high school classes exceeding 50% are less when compared to Top10perc of their high school classes not exceeding 50%.



The left box plot represents the 5 point summary of data whose proportion of students coming from the top 10% of their high school classes doesn't exceeds 50% and right one represent proportion of students coming from the top 10% of their high school classes exceeds 50%.

Interpretation:

- Box plot 1 has more range than Box plot 2.
- Box plot 1 is symmetric but Box plot 2 is left skewed.
- Box plot 1 has outliers but Box plot 2 doesn't have any.
- IQR for Box plot 2 is higher than Box plot 1.

v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables.

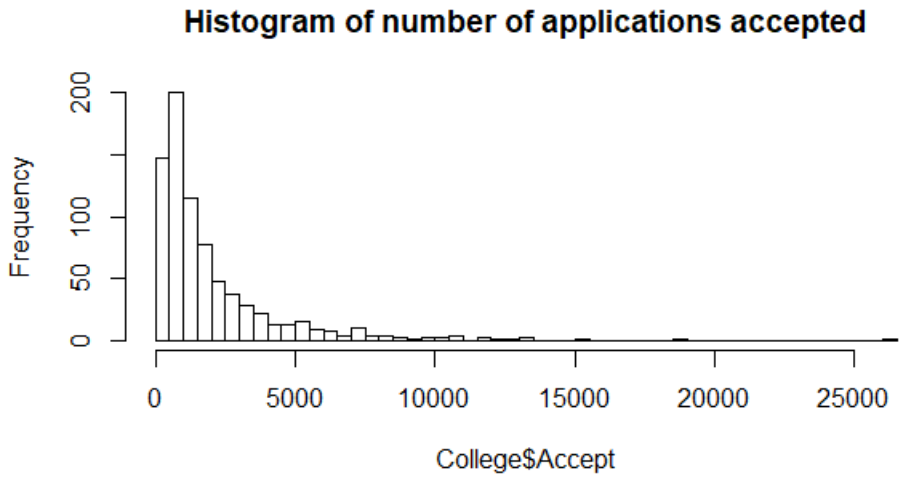
par(mfrow=c(2,1))
#divides the single plot frame into 2x2 frames

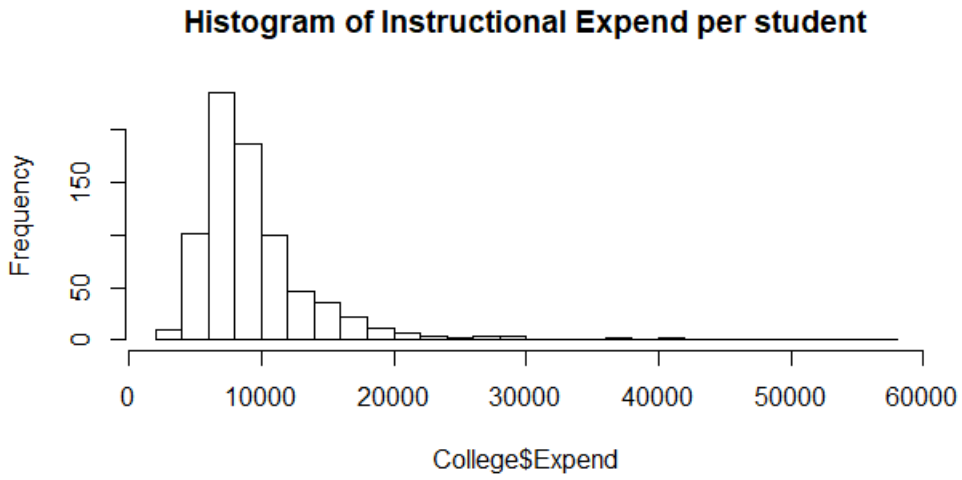
hist(College$Accept, main = paste("Histogram of number of applications accepted"), breaks=50)
#generates a histogram with 50 bins

hist(College$Expend, main = paste("Histogram of Instructional Expend per student"), breaks = 30)
#generates histogram with 30 bins

hist(College$Enroll, border="blue", main = paste("Histogram of Enroll"), breaks = 20)
#generates histogram with 20 bins with blue border

hist(College$PhD, main = paste("Histogram of faculty with PhD"), axes="TRUE", breaks = 10)
#generates histogram with 10 bins, if axes was FLASE then it generates histogram with no axes. By default the axes value is TRUE.



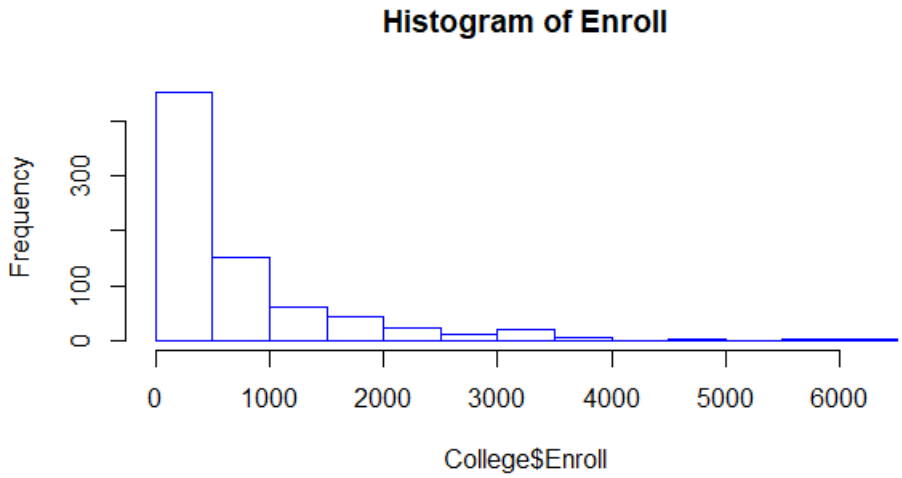


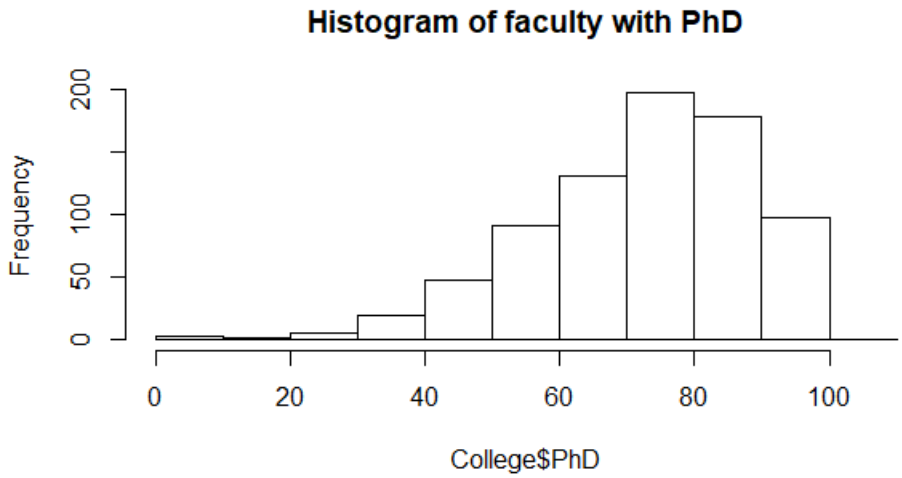
Histogram of no. of applications accepted:

- It has exponential nature with right skewed property.
- Around 500 to 1000 no. of applications are accepted by 200 universities which is the highest.
- We can see that there are few bins far away from actual graph, these can be interpreted as Outliers
- There are less number of universities which are accepting more number of applications.

Histogram of Instructional Expenditure per student:

- It has normal nature with right skewed property.
- Above 200 universities gave 5000 to 7500 instructional expenditure per student.
- As this histogram is right skewed we can say that Mode < Median < Mean.
- We can see that there are few bins far away from actual graph, these can be interpreted as Outliers





Histogram of number of new students enrolled:

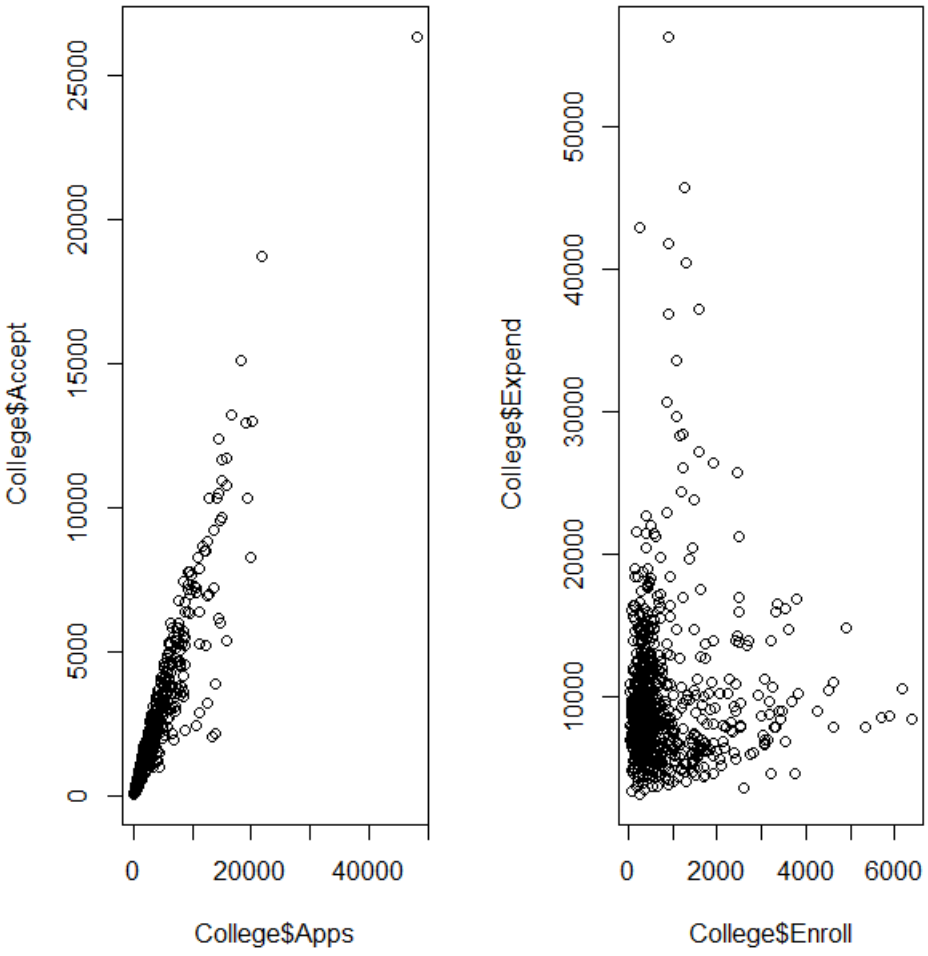
- It has exponential nature with right skewed property.
- Around 1 to 500 no. of new students enrolled in 450 universities which is the highest.
- We can see that there are few bins far away from actual graph, these can be interpreted as Outliers
- There are less number of universities which have more number of new students enrolled.

Histogram of Percent of faculty with Ph.D.'s:

- It has normal nature with left skewed property.
- About 200 universities have 70% - 80% of professors having PhD's which is the highest.
- As this histogram is left skewed we can say that Mode > Median > Mean.

vi. Exploring the data, and provide a brief summary of what you discover.

```
par(mfrow=c(2,2))
plot(College$Apps, College$Accept) #It produces the scatter plot for given attributes
plot(College$Enroll, College$Expend)
plot(College$Accept, College$Enroll)
plot(College$Top25perc, College$Books)
plot(College$Enroll, College$F.Undergrad)
plot(College$P.Undergrad, College$Expend)
plot(College$PhD, College$Terminal)
```

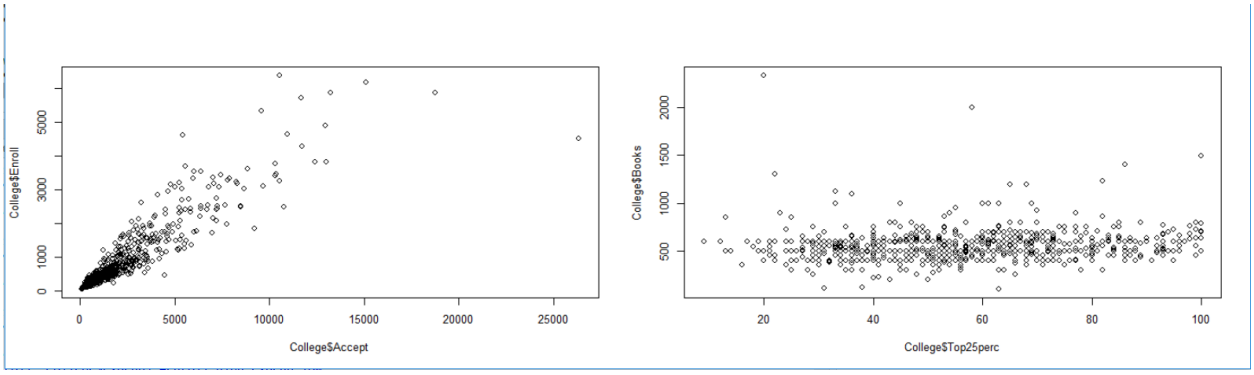


Plot between no. of applications received and no. of applications accepted:

- As the no. of applications received to the university increases, the no. of applications accepted increases.

Plot between no. of new students enrolled and Instructional Expenditure per student:

- As the no. of new students enrolled increases, the instructional expenditure per student decreases.

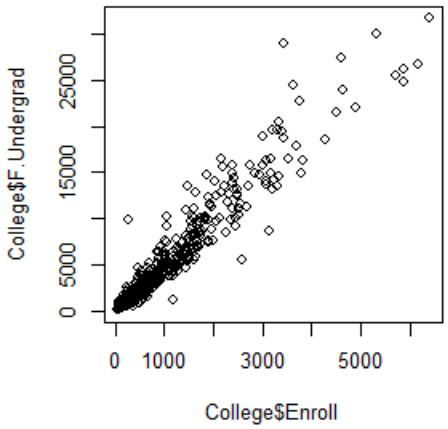


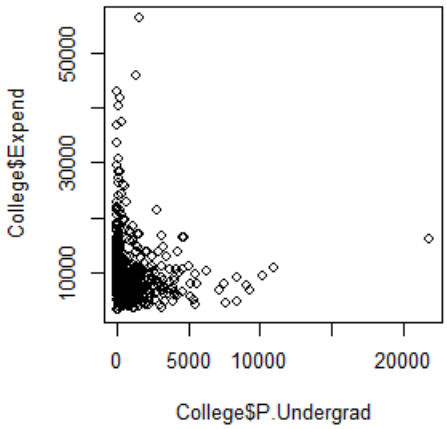
Plot between no. of applications accepted and no. of new students enrolled:

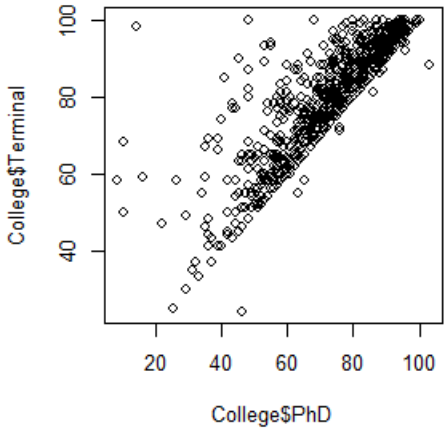
- As the no. of applications accepted increases, the no. new students enrolled will also increase.

Plots between New students from top 25% of high school class and Estimated book cost:

- Even though the new students from top 25% of high school class increases the estimated book cost remains same.







Plot between no. of new students enrolled and no. of full time under grads:

- As the no. of new students enrolled increases, the no. full time under grads also increase

Plot between no. of part time under grads and instructional expenditure per student:

- As the no. of part time under grads increases, instructional expenditure per student decreases.

Plot between percent of faculty with PhD's and percent of faculty with terminal degree:

- As the percent of faculty with PhD's increases, the percent of faculty with terminal degree also increase.