STATISTICAL METHODS FOR DATA SCIENCE
MINI PROJECT #2

SPRING 2018

NAMES OF GROUP MEMBERS:
ADRITA DUTTA(axd172930)
NEETHU ANTONY(nxa171330)

Contribution of team members:

ADRITA:
Learned R coding
Tried the R Codes
Documented the Codes and its Output
Tried out different plots and Boxplots
Explored the Data using plots

NEETHU:
Learned R coding
Tried the R Codes
Explained the Plots and Graphs
Wrote Code explanation for section-1
Explored the Data using plots

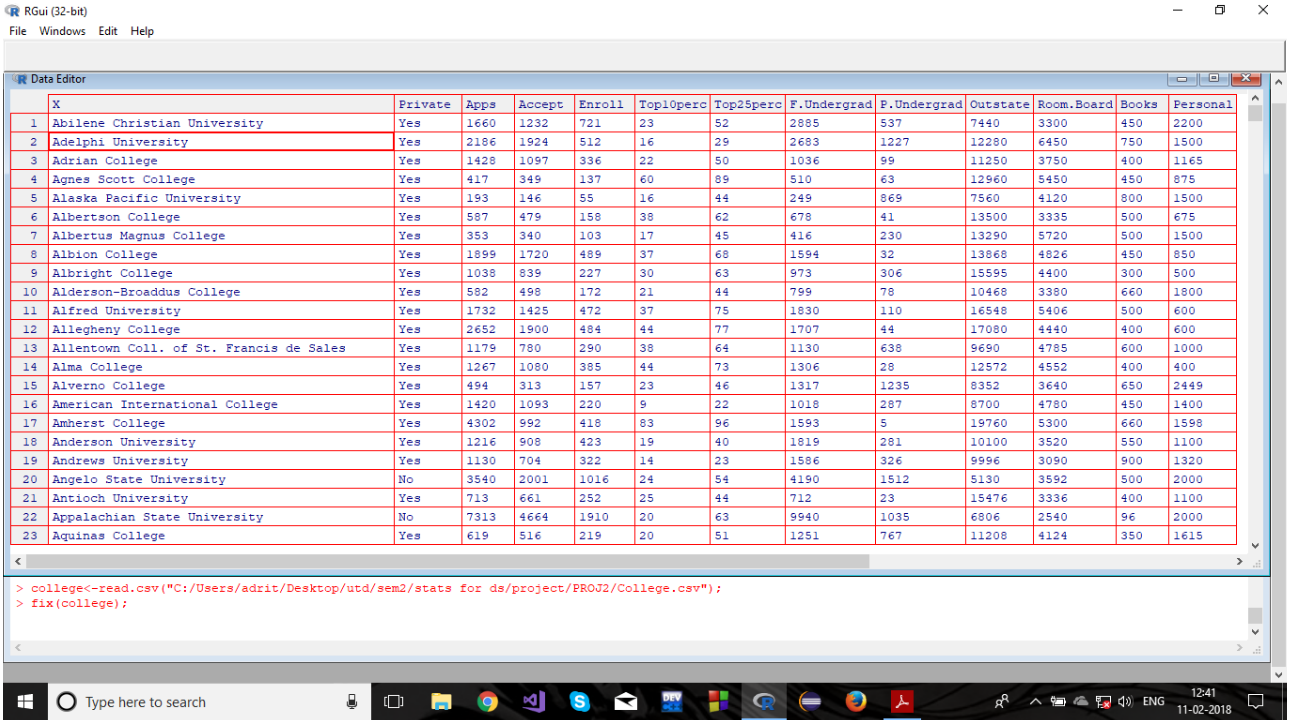
Section 1:

8.

a)read file college.csv

b)

->see the look of it

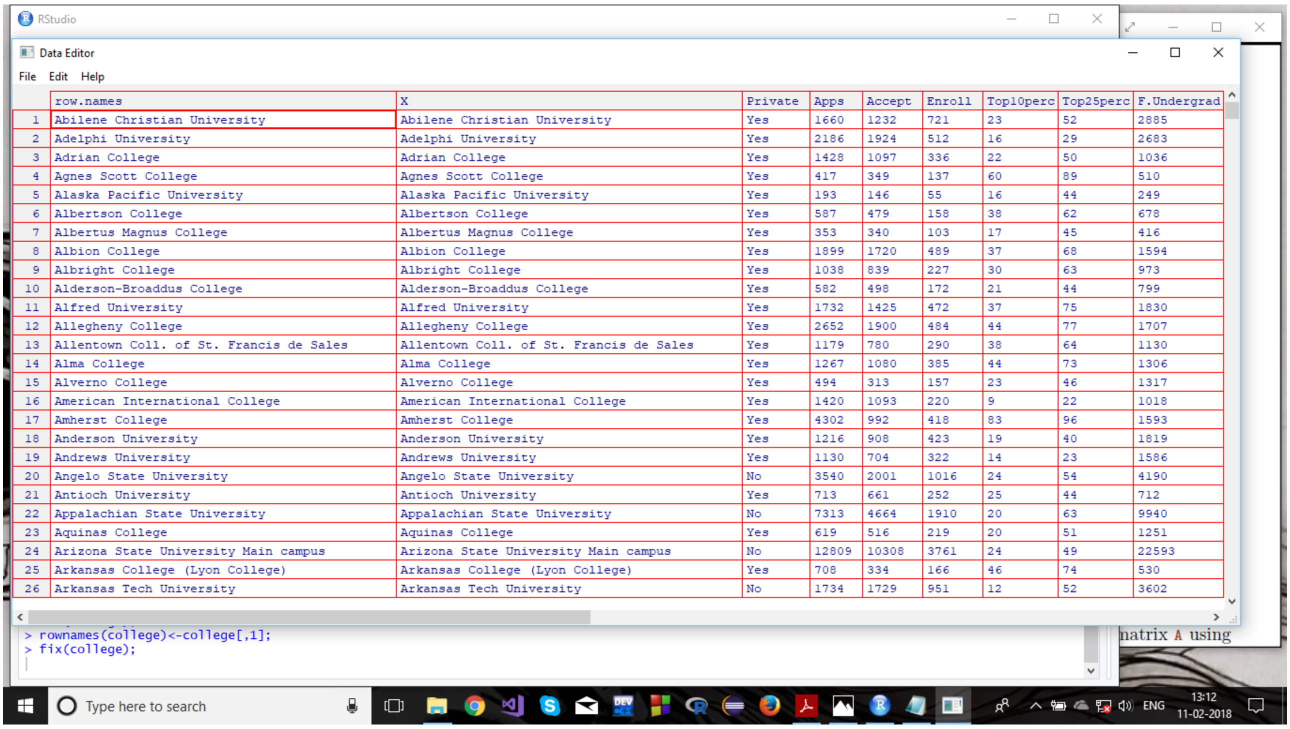


	X	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal
1	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200
2	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500
3	Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165
4	Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875
5	Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800	1500
6	Albertson College	Yes	587	479	158	38	62	678	41	13500	3335	500	675
7	Albertus Magnus College	Yes	353	340	103	17	45	416	230	13290	5720	500	1500
8	Albion College	Yes	1899	1720	489	37	68	1594	32	13868	4826	450	850
9	Albright College	Yes	1038	839	227	30	63	973	306	15595	4400	300	500
10	Alderson-Broaddus College	Yes	582	498	172	21	44	799	78	10468	3380	660	1800
11	Alfred University	Yes	1732	1425	472	37	75	1830	110	16548	5406	500	600
12	Allegheny College	Yes	2652	1900	484	44	77	1707	44	17080	4440	400	600
13	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64	1130	638	9690	4785	600	1000
14	Alma College	Yes	1267	1080	385	44	73	1306	28	12572	4552	400	400
15	Alverno College	Yes	494	313	157	23	46	1317	1235	8352	3640	650	2449
16	American International College	Yes	1420	1093	220	9	22	1018	287	8700	4780	450	1400
17	Amherst College	Yes	4302	992	418	83	96	1593	5	19760	5300	660	1598
18	Anderson University	Yes	1216	908	423	19	40	1819	281	10100	3520	550	1100
19	Andrews University	Yes	1130	704	322	14	23	1586	326	9996	3090	900	1320
20	Angelo State University	No	3540	2001	1016	24	54	4190	1512	5130	3592	500	2000
21	Antioch University	Yes	713	661	252	25	44	712	23	15476	3336	400	1100
22	Appalachian State University	No	7313	4664	1910	20	63	9940	1035	6806	2540	96	2000
23	Aquinas College	Yes	619	516	219	20	51	1251	767	11208	4124	350	1615

```
> college<-read.csv("C:/Users/adrit/Desktop/utd/sem2/stats for ds/project/PROJ2/College.csv");
> fix(college);
```

->add row names from the 1st column of the data set that stores the university names.

->look of the edited table, now each row has a name corresponding to the appropriate university marked by the column row.names.

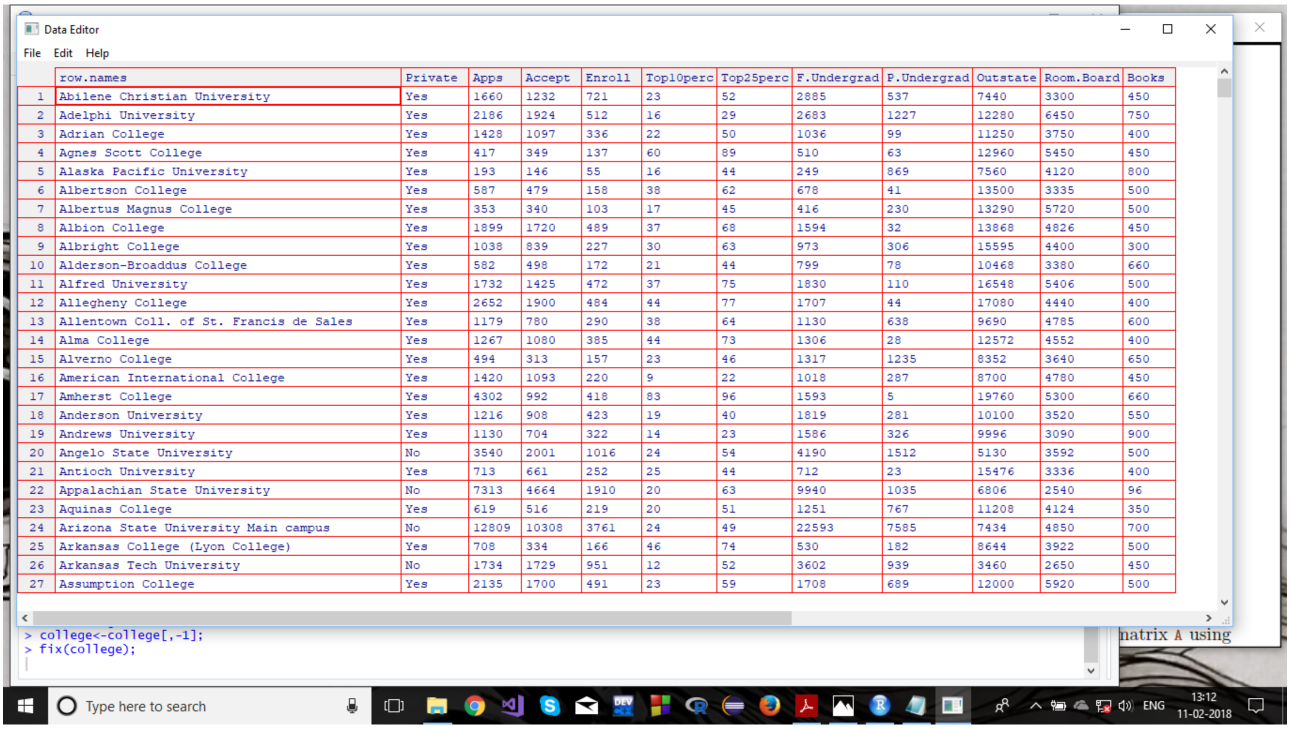


	row.names	X	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad
1	Abilene Christian University	Abilene Christian University	Yes	1660	1232	721	23	52	2885
2	Adelphi University	Adelphi University	Yes	2186	1924	512	16	29	2683
3	Adrian College	Adrian College	Yes	1428	1097	336	22	50	1036
4	Agnes Scott College	Agnes Scott College	Yes	417	349	137	60	89	510
5	Alaska Pacific University	Alaska Pacific University	Yes	193	146	55	16	44	249
6	Albertson College	Albertson College	Yes	587	479	158	38	62	678
7	Albertus Magnus College	Albertus Magnus College	Yes	353	340	103	17	45	416
8	Albion College	Albion College	Yes	1899	1720	489	37	68	1594
9	Albright College	Albright College	Yes	1038	839	227	30	63	973
10	Alderson-Broaddus College	Alderson-Broaddus College	Yes	582	498	172	21	44	799
11	Alfred University	Alfred University	Yes	1732	1425	472	37	75	1830
12	Allegheny College	Allegheny College	Yes	2652	1900	484	44	77	1707
13	Allentown Coll. of St. Francis de Sales	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64	1130
14	Alma College	Alma College	Yes	1267	1080	385	44	73	1306
15	Alverno College	Alverno College	Yes	494	313	157	23	46	1317
16	American International College	American International College	Yes	1420	1093	220	9	22	1018
17	Amherst College	Amherst College	Yes	4302	992	418	83	96	1593
18	Anderson University	Anderson University	Yes	1216	908	423	19	40	1819
19	Andrews University	Andrews University	Yes	1130	704	322	14	23	1586
20	Angelo State University	Angelo State University	No	3540	2001	1016	24	54	4190
21	Antioch University	Antioch University	Yes	713	661	252	25	44	712
22	Appalachian State University	Appalachian State University	No	7313	4664	1910	20	63	9940
23	Aquinas College	Aquinas College	Yes	619	516	219	20	51	1251
24	Arizona State University Main campus	Arizona State University Main campus	No	12809	10308	3761	24	49	22593
25	Arkansas College (Lyon College)	Arkansas College (Lyon College)	Yes	708	334	166	46	74	530
26	Arkansas Tech University	Arkansas Tech University	No	1734	1729	951	12	52	3602

```
> rownames(college)<-college[,1];
> fix(college);
```

->delete the first column that has college names in it from the data set, since this is treated as part of the data. We have already stored this data as row.names.

->look at edited table

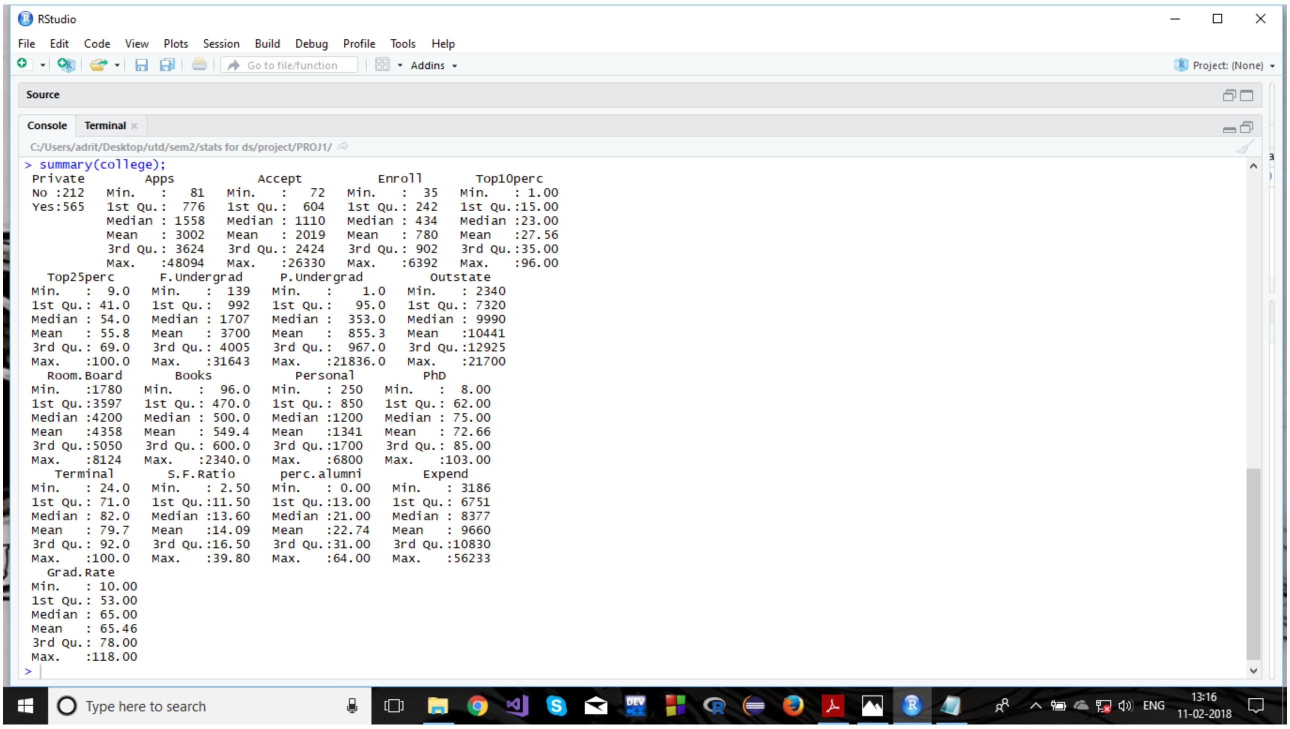


	row.names	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books
1	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450
2	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750
3	Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400
4	Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450
5	Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800
6	Albertson College	Yes	587	479	158	38	62	678	41	13500	3335	500
7	Albertus Magnus College	Yes	353	340	103	17	45	416	230	13290	5720	500
8	Albion College	Yes	1899	1720	489	37	68	1594	32	13868	4826	450
9	Albright College	Yes	1038	839	227	30	63	973	306	15595	4400	300
10	Alderson-Broaddus College	Yes	582	498	172	21	44	799	78	10468	3380	660
11	Alfred University	Yes	1732	1425	472	37	75	1830	110	16548	5406	500
12	Allegheny College	Yes	2652	1900	484	44	77	1707	44	17080	4440	400
13	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64	1130	638	9690	4785	600
14	Alma College	Yes	1267	1080	385	44	73	1306	28	12572	4552	400
15	Alverno College	Yes	494	313	157	23	46	1317	1235	8352	3640	650
16	American International College	Yes	1420	1093	220	9	22	1018	287	8700	4780	450
17	Amherst College	Yes	4302	992	418	83	96	1593	5	19760	5300	660
18	Anderson University	Yes	1216	908	423	19	40	1819	281	10100	3520	550
19	Andrews University	Yes	1130	704	322	14	23	1586	326	9996	3090	900
20	Angelo State University	No	3540	2001	1016	24	54	4190	1512	5130	3592	500
21	Antioch University	Yes	713	661	252	25	44	712	23	15476	3336	400
22	Appalachian State University	No	7313	4664	1910	20	63	9940	1035	6806	2540	96
23	Aquinas College	Yes	619	516	219	20	51	1251	767	11208	4124	350
24	Arizona State University Main campus	No	12809	10308	3761	24	49	22593	7585	7434	4850	700
25	Arkansas College (Lyon College)	Yes	708	334	166	46	74	530	182	8644	3922	500
26	Arkansas Tech University	No	1734	1729	951	12	52	3602	939	3460	2650	450
27	Assumption College	Yes	2135	1700	491	23	59	1708	689	12000	5920	500

```
> college<-college[,-1];
> fix(college);
```

c.i) summary of all data sets:-

Here, since we asked for the summary of the entire data set, summary of each column is returned in the output. The column private has only 2 values- Yes or No, hence the summary will be the number of data having 'Yes' value for column Private and number of data having 'No' value for column Private. For all other columns, we have the 5 number summary which includes Minimum, 1st Quartile, Median(2nd Quartile), 3rd Quartile and Maximum.



```
> summary(college);
 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) scatterplot matrix of first 10 columns:

Considering the first 10 columns we see that it forms a matrix that is symmetric along its diagonal.
Each scatterplot represents the correlations between the 2 elements.
On consideration of each scatterplot separately we can observe a few different kinds of relations among the variables:-

1)Horizontal lines:no correlation. The vertical axis is a value axis and the horizontal axis is a category axis. Instead of displaying values, a category axis shows evenly spaced groupings (categories) of data. Because my data has only values and no categories.

2)Vertical lines:no correlation.The horizontal axis is a value axis and the vertical axis is a category axis. Instead of displaying values, a category axis shows evenly spaced groupings (categories) of data. Because my data has only values and no categories.

3)diagonal line(/):a line with a positive slope indicates that the 2 variables have a positive correlation.
A line with a negative slope would represent a negative correlation, but there is no such relation in this matrix.

4)approximately diagonal line: These show low positive correlation. The plot is almost diagonal but is not exactly diagonal.

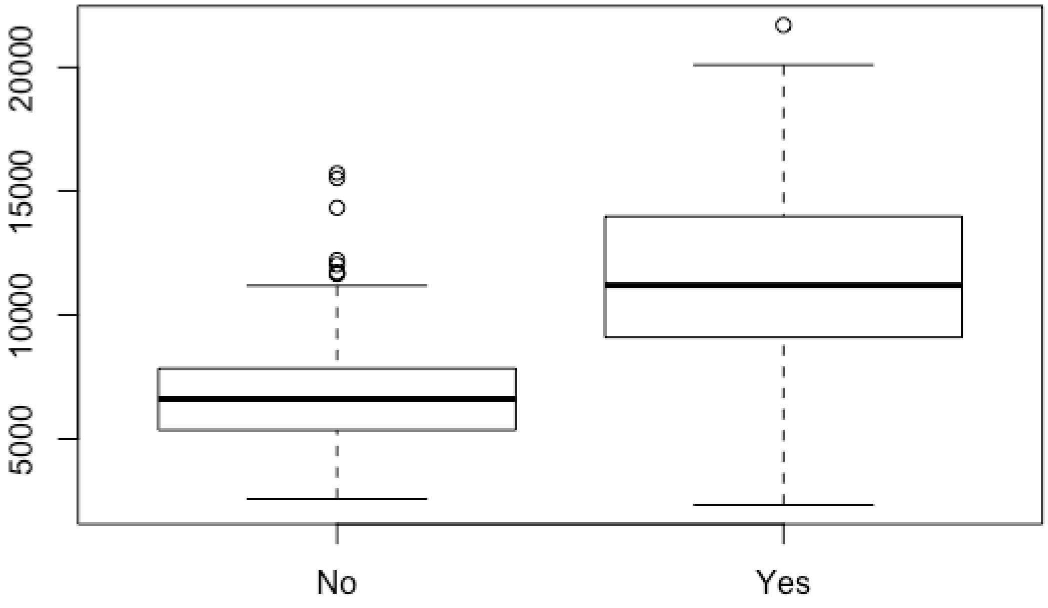
iii) side-by-side boxplots of Outstate vs Private:

Boxplots are only able to deal with one quantitative variable. However, side-by-side box plots can be applied to data sets with one quantitative and one categorical variable, which makes them especially useful for many real-world statistical problems.

Here Private is a categorical variable in the data set and Outstate is a quantitative variable. Private variable can take up two values or Categories(Yes or No). Hence there will be 2 boxplots drawn side-by-side as shown below. Side-by-side boxplots present all of the information that box plots do for each instance of a categorical variable. Box plots corresponding to each instance of the categorical variable (in this case two instances namely, Yes and No) summarize the data in five different numbers:-

- The Median
- The Lower Quartile
- The Upper Quartile
- The Minimum
- The Maximum

From the below side-by-side boxplot, we can infer that the Private Universities has a higher Out-of-state tuition compared to the public universities.



iv) create Elite, by binning the Top10perc variable

->see how many Elite universities are there using summary: This gives the number of universities that are marked as Yes in Elite column and number of universities that are marked as No in the Elite Column.

NO	YES
699	78

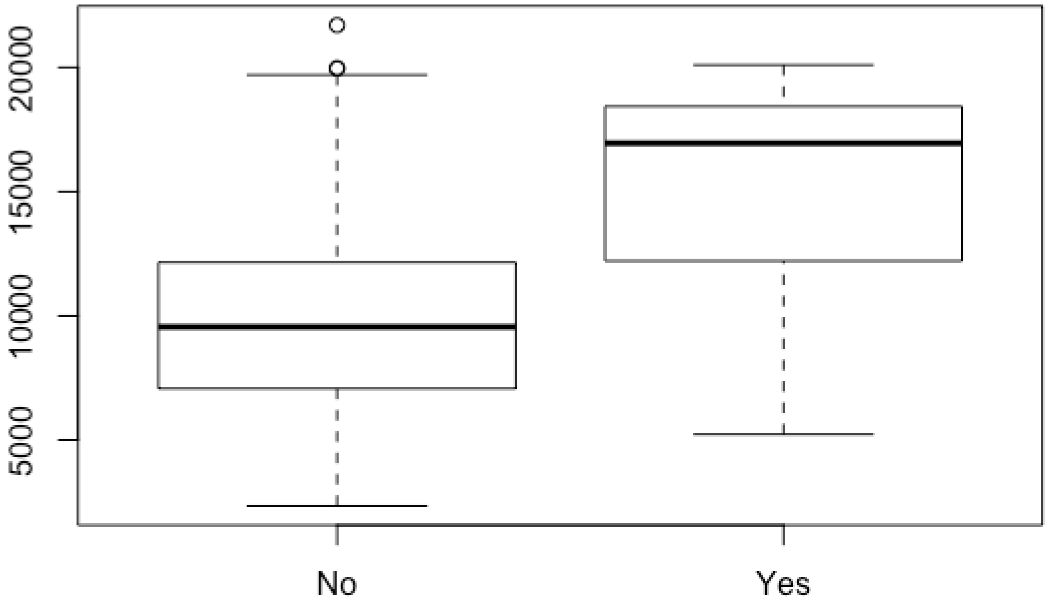
->side-by-side boxplots of Outstate vs Elite

Boxplots are only able to deal with one quantitative variable. However, side-by-side box plots can be applied to data sets with one quantitative and one categorical variable, which makes them especially useful for many real-world statistical problems.

Here Private is a categorical variable in the data set and Outstate is a quantitative variable. Private variable can take up two values or Categories(Yes or No). Hence there will be 2 boxplots drawn side-by-side as shown below. Side-by-side boxplots present all of the information that box plots do for each instance of a categorical variable. Box plots corresponding to each instance of the categorical variable (in this case two instances namely, Yes and No) summarize the data in five different numbers:-

- The Median
- The Lower Quartile
- The Upper Quartile
- The Minimum
- The Maximum

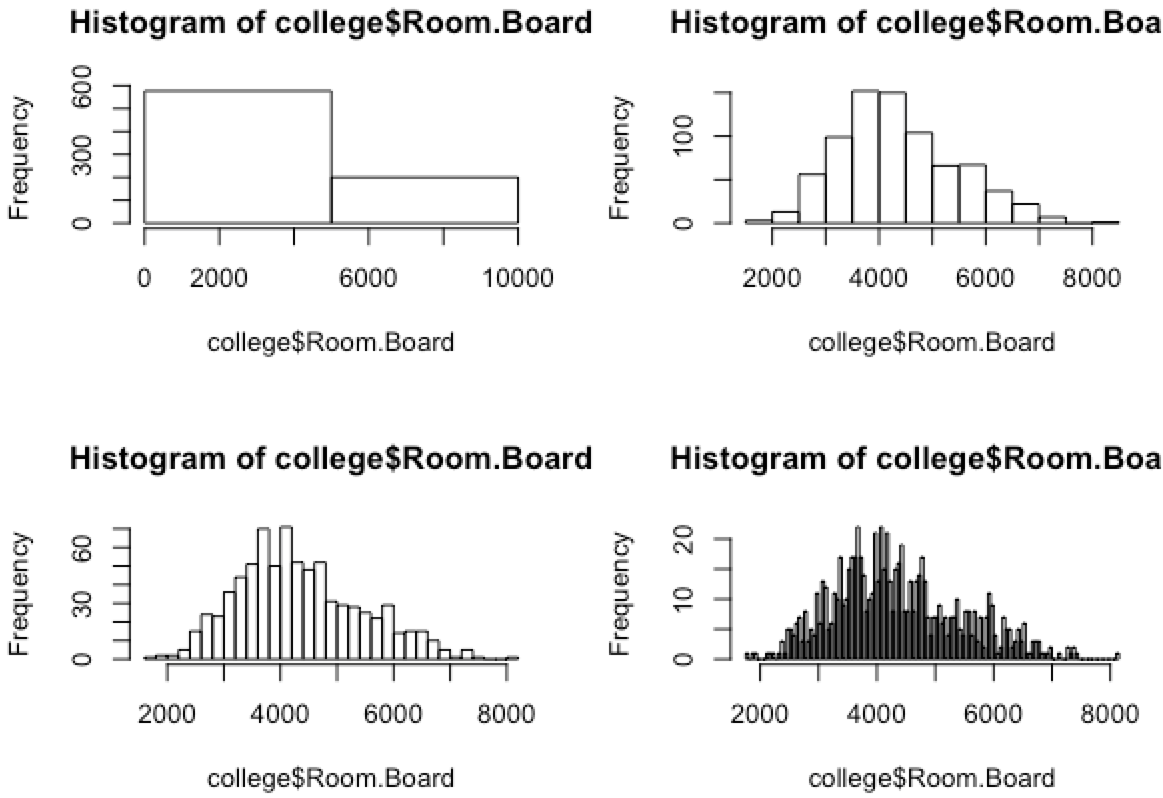
From the below side-by-side boxplot, we can infer that the Elite Universities has a higher Out-of-state tuition compared to the other universities. Also, the box plot corresponding to Elite universities is left skewed, since the median is closer to 3rd Quartile, this again presses the fact that most of the Elite universities has a higher out-of-state tuition.



v) histograms of a few variables with different number of bins.

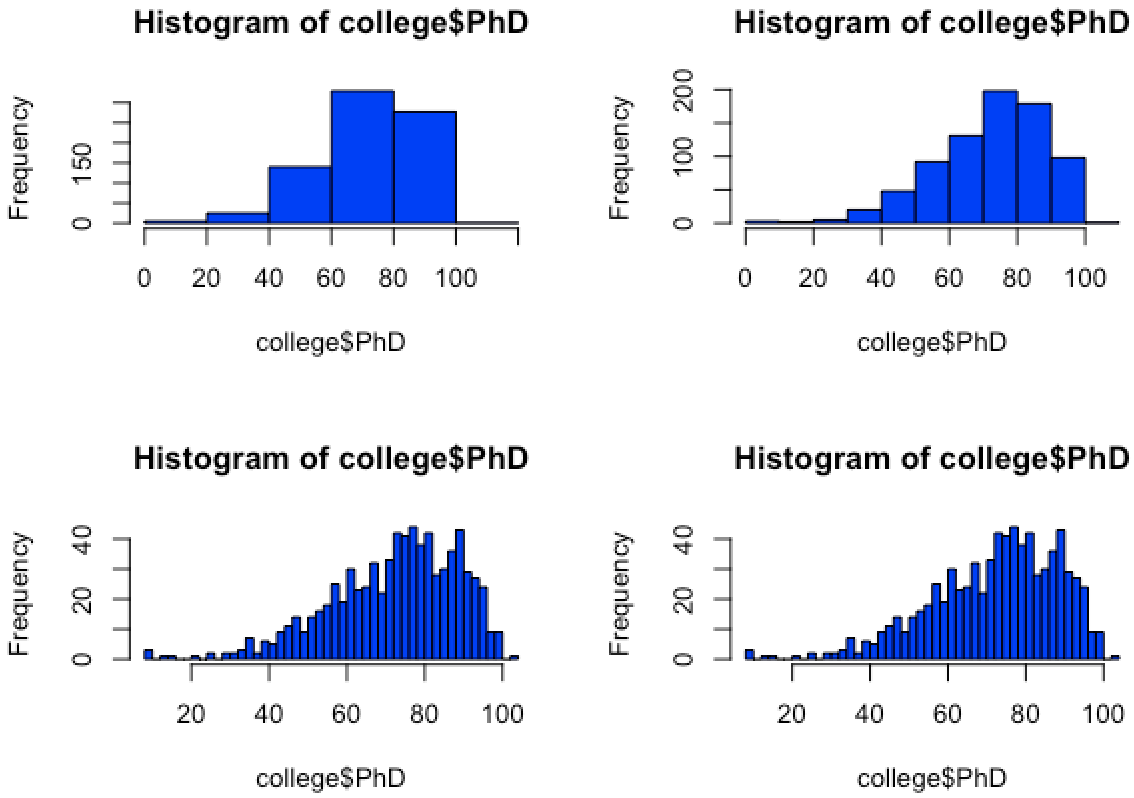
(a) Histogram of the quantitative variable Room.Board with differing number of bins.

From the below plots, it is clear that the distribution is right skewed. We can see that in most of the universities the Room and Board costs lies between 2000 and 7000. There are a few universities to the right that have a Room and Board costs higher than 7000 and these are the bars that make the data have a shape that is skewed right.



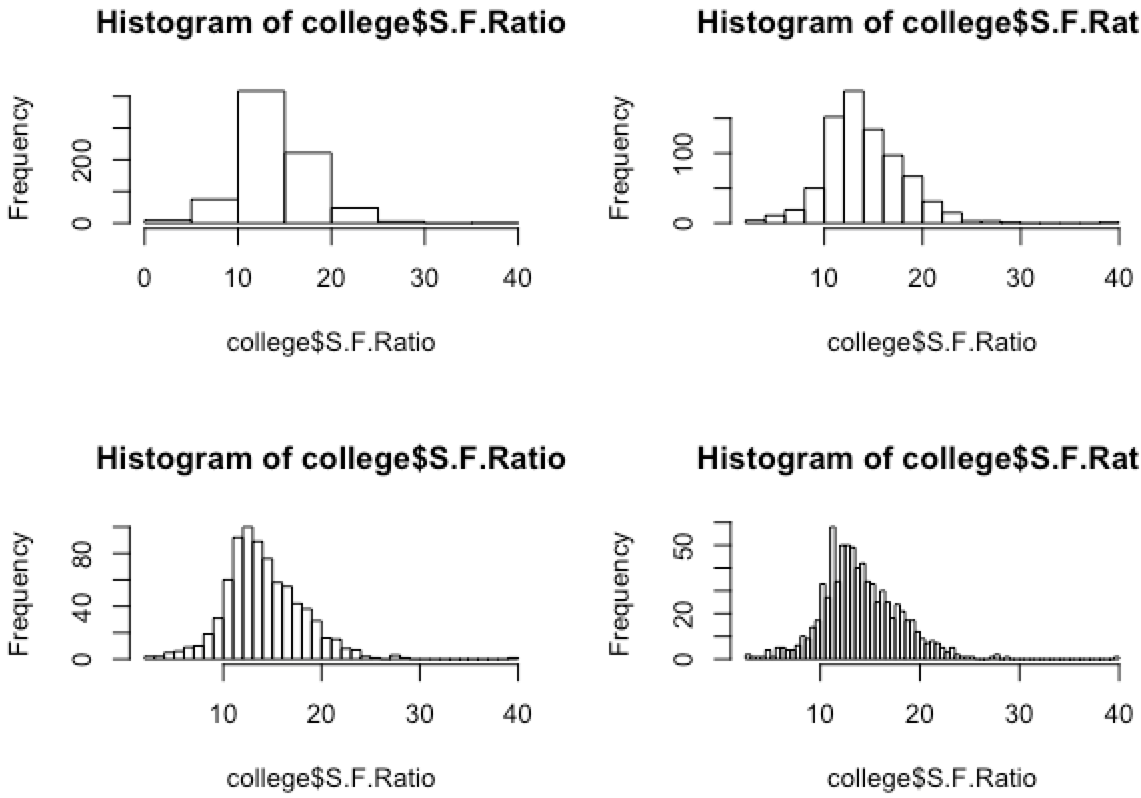
(b) Histogram of the quantitative variable PhD with differing number of bins.

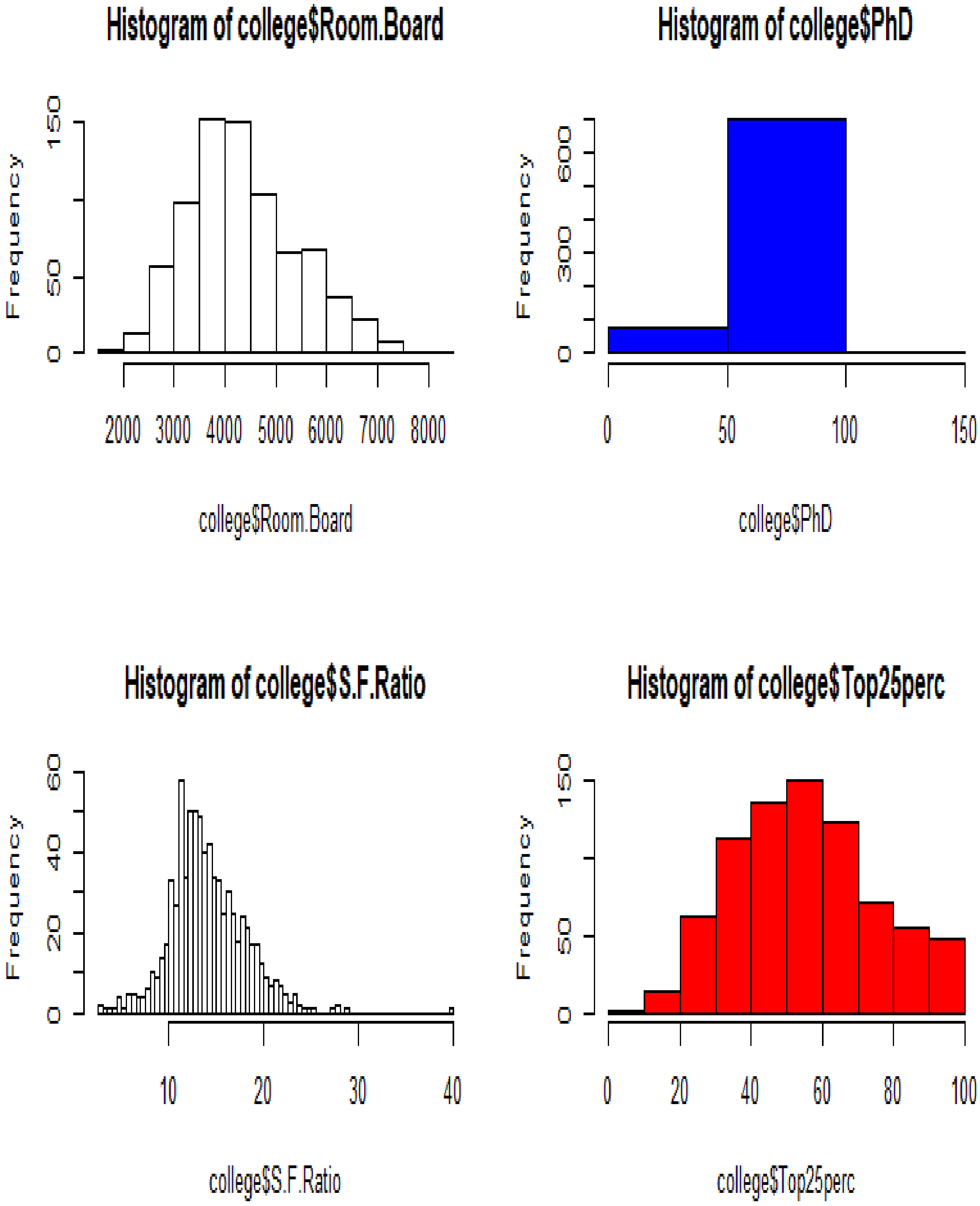
From the below plots, it is clear that the distribution is left skewed. We can see that in most of the universities the percent of faculties with PhD's is between 40 and 100. There are a few universities to the left that have a percent of faculties with PhD's less than 40 and these are the bars that make the data have a shape that is skewed left.



(c) Histogram of the quantitative variable S.F.Ratio with differing number of bins.

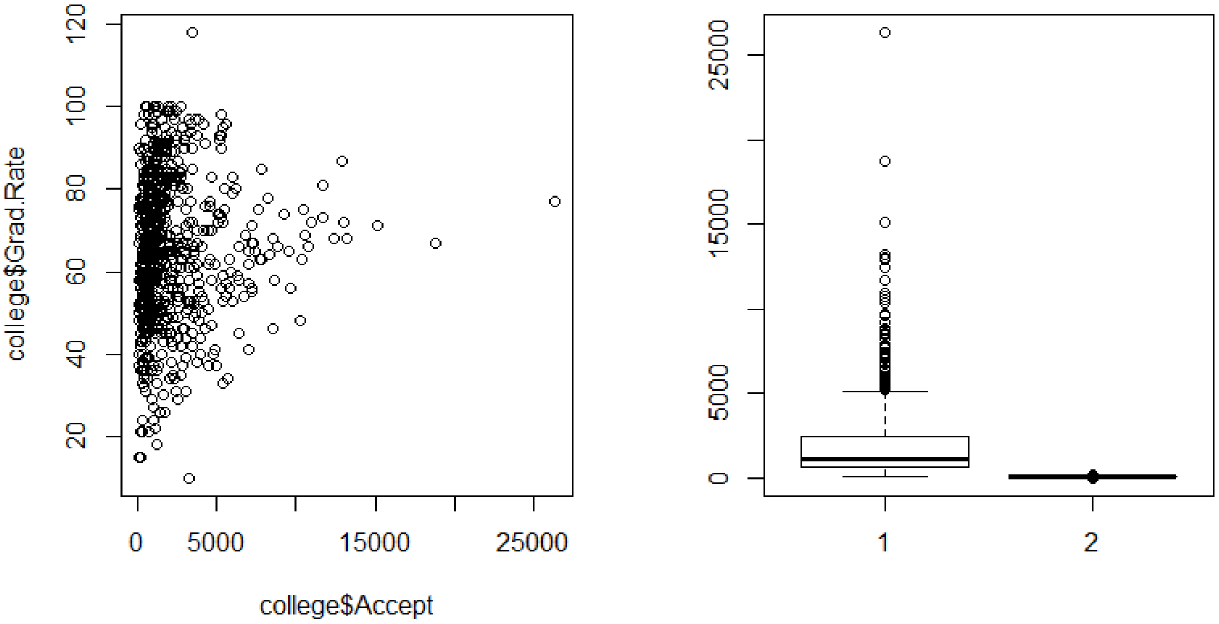
From the below plots, it is clear that the distribution is left skewed. We can see that in most of the universities the percent of faculties with PhD's is between 40 and 100. There are a few universities to the left that have a percent of faculties with PhD's less than 40 and these are the bars that make the data have a shape that is skewed left.


Histogram of college$S.F.Ratio
Frequency
0
200
0
10
20
30
40
college$S.F.Ratio
Histogram of college$S.F.Ratio
Frequency
0
100
10
20
30
40
college$S.F.Ratio
Histogram of college$S.F.Ratio
Frequency
0
40
80
10
20
30
40
college$S.F.Ratio
Histogram of college$S.F.Ratio
Frequency
0
20
50
10
20
30
40
college$S.F.Ratio

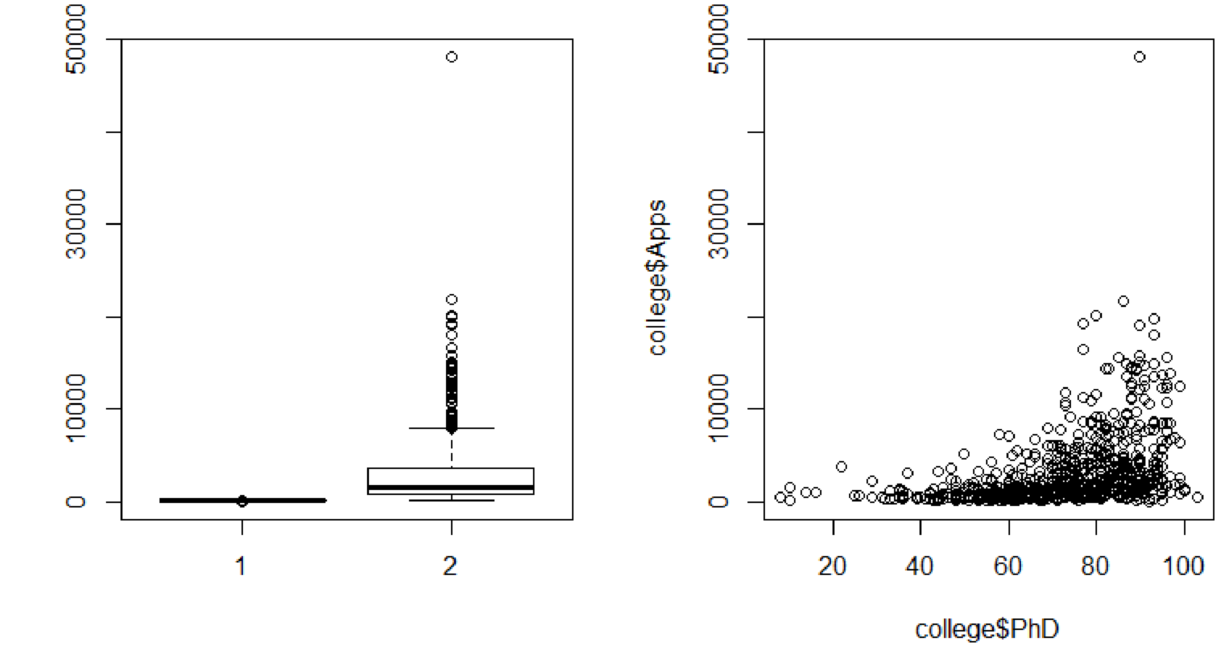

Histogram of college$Room.Board
Frequency
0
50
150
2000
3000
4000
5000
6000
7000
8000
college$Room.Board
Histogram of college$PhD
Frequency
0
300
600
0
50
100
150
college$PhD
Histogram of college$S.F.Ratio
Frequency
0
20
40
60
10
20
30
40
college$S.F.Ratio
Histogram of college$Top25perc
Frequency
0
50
150
0
20
40
60
80
100
college$Top25perc

vi) On further experimenting with the given data it is observed that:

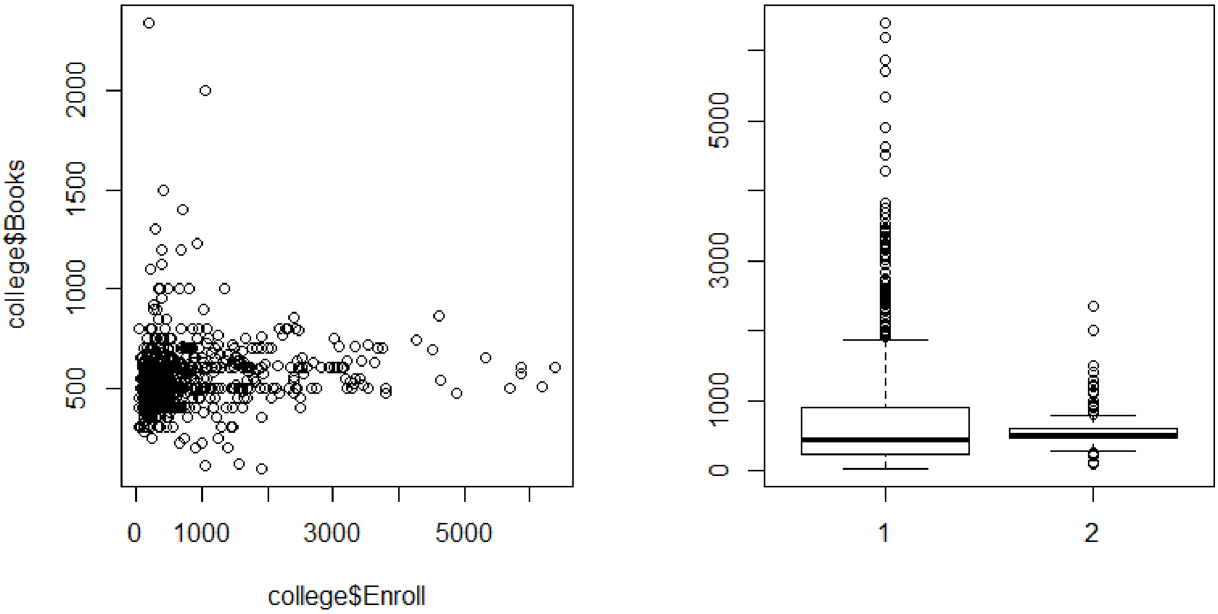
->the higher the acceptance, graduation rate has lesser variance and is thus lower in average



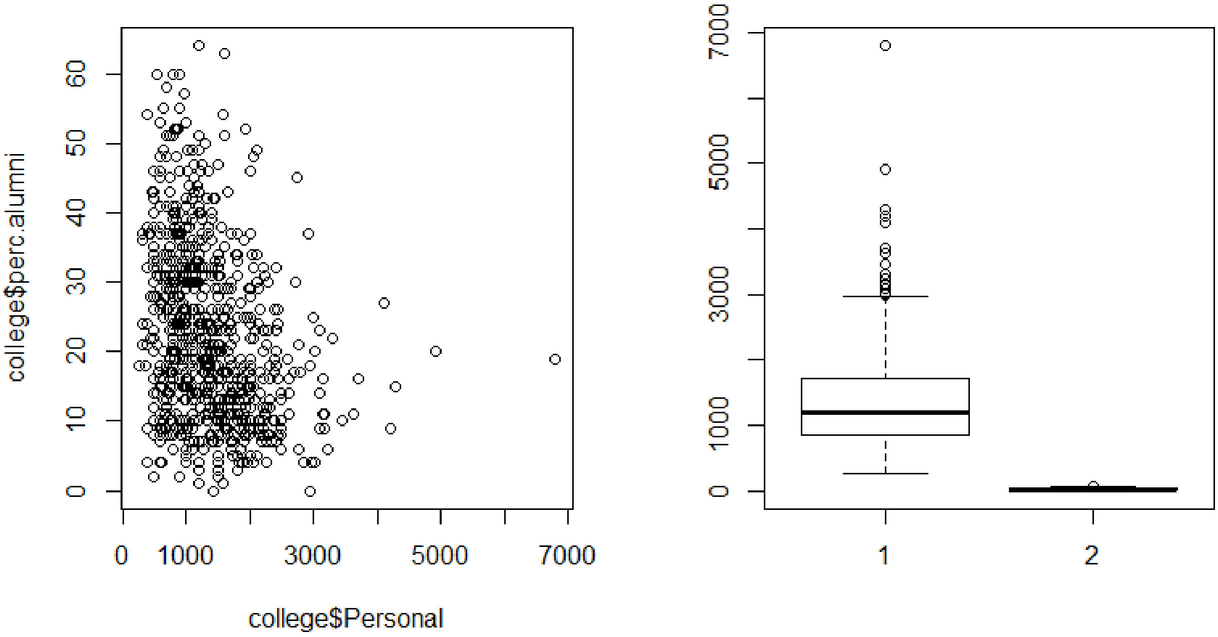
->more applicants if more number of faculty have phd



->more variance in costs of books when lesser students enrolled



->as personal expenses increase donations have lower variance and thus lower average values



Section 2:

R-Code:

```
#a)read file college.csv
college<-read.csv("C:/Users/adrit/Desktop/utd/sem2/stats for
ds/project/PROJ2/College.csv");

#b)
#fix file, see the look of it
fix(college);

#now we set names to each row(R will not perform any operations on row
names:)
#extra column for row name added for each row
rownames(college)<-college[,1];

#see the look of the table again(this time with an extra column for names)
fix(college);

#now that we have 2 columns with uni name, delete 1
#delete the original 1 as calculations can be done on that.
college<-college[,-1];

#see table again:
fix(college);

#c)
#i)summary of data set(MIN,Q1,Q2,MEAN,Q3,MAX)
summary(college);


#ii)scatterplot matrix of first 10 columns:
pairs(college[,1:10]);
```

```
#iii)side by side boxplots of Outstate vs Private
plot(college$Private, college$Outstate)

#iv)creating Elite(top 10 percentile)
Elite=rep("No",nrow(college))
Elite[college$Top10perc>50]="Yes"
Elite=as.factor(Elite)
college=data.frame(college,Elite)


#summary of Elite universities
summary(college$Elite)

#boxplots of outstate vs Elite
plot(college$Elite, college$Outstate)

#v)histograms of a few variables with different bin size
#dividing display page
par(mfrow=c(2,2))

#histogram of room and board costs
hist(college$Room.Board)

#histogram of PhD students
hist(college$PhD, col=4, breaks = 2)

#histogram of student faculty ratio
hist(college$S.F.Ratio, breaks =10)
hist(college$S.F.Ratio, breaks =25)
hist(college$S.F.Ratio, breaks =50)
hist(college$S.F.Ratio, breaks =100)


#histogram of top 25 percentile
hist(college$Top25perc, col=2)
```

```
#vi)doing other experiments on data:
#boxplot of no of acceptances and graduation rate
#we see here the higher the acceptance, graduation rate has lesser
variance)
par(mfrow=c(1,2))
plot(college$Accept, college$Grad.Rate)
boxplot(college$Accept, college$Grad.Rate)


#boxplot of applicant no vs percentage of faculty with PHD
#more applicants if more number of faculty have phd
plot(college$PhD, college$Apps)
boxplot(college$PhD, college$Apps)


#boxplot of Enroll and Books
#more variance in costs of books when lesser students enrolled
plot(college$Enroll, college$Books)
boxplot(college$Enroll, college$Books)


#personal exp vs percentage of alumni who donate
#as personal expenses increase donations have lower variance and thus
lower average values.
#if they spent less in college they are more likely to donate
plot(college$Personal, college$perc.alumni)
```