STATISTICAL METHODS FOR DATA SCIENCE
MINI PROJECT #2

NAME: ADRITA DUTTA
NET-ID: axd172930

SPRING 2018

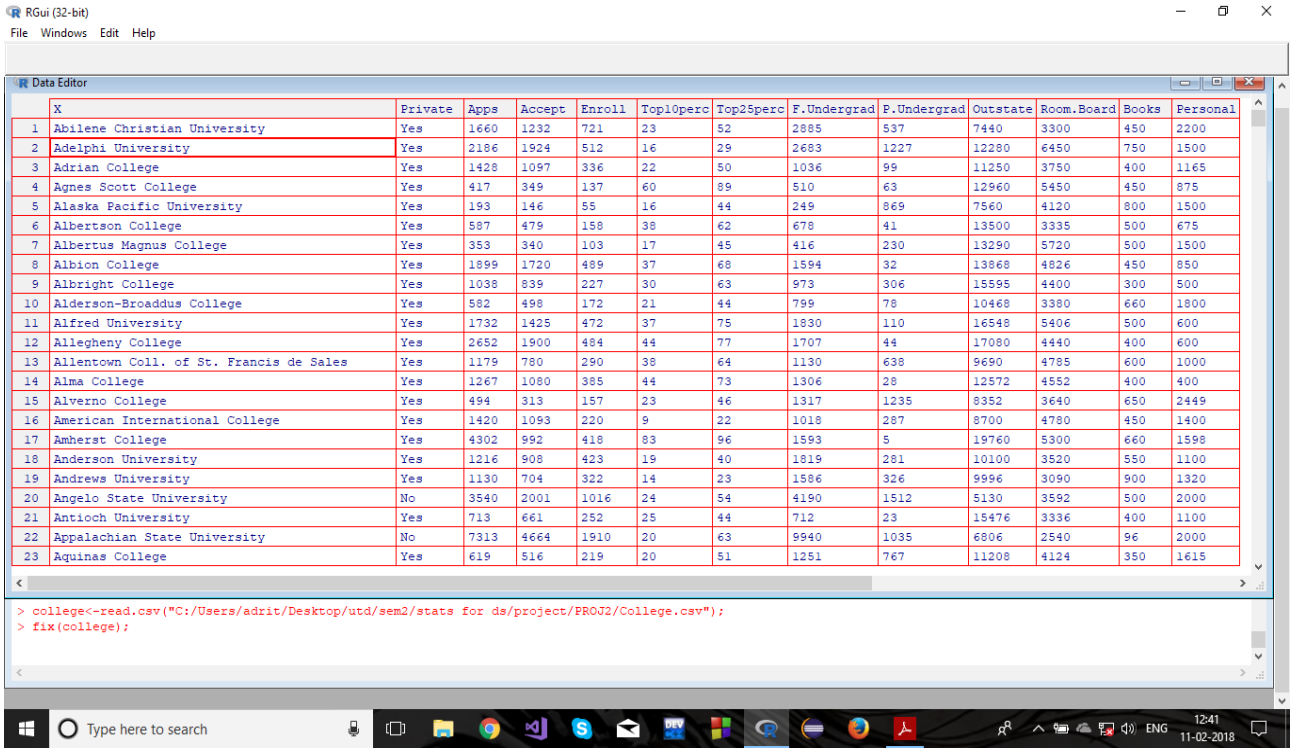
Section 1:

8.

a)read file college.csv

b)

->see the look of it



RGui (32-bit)

File Windows Edit Help

Data Editor

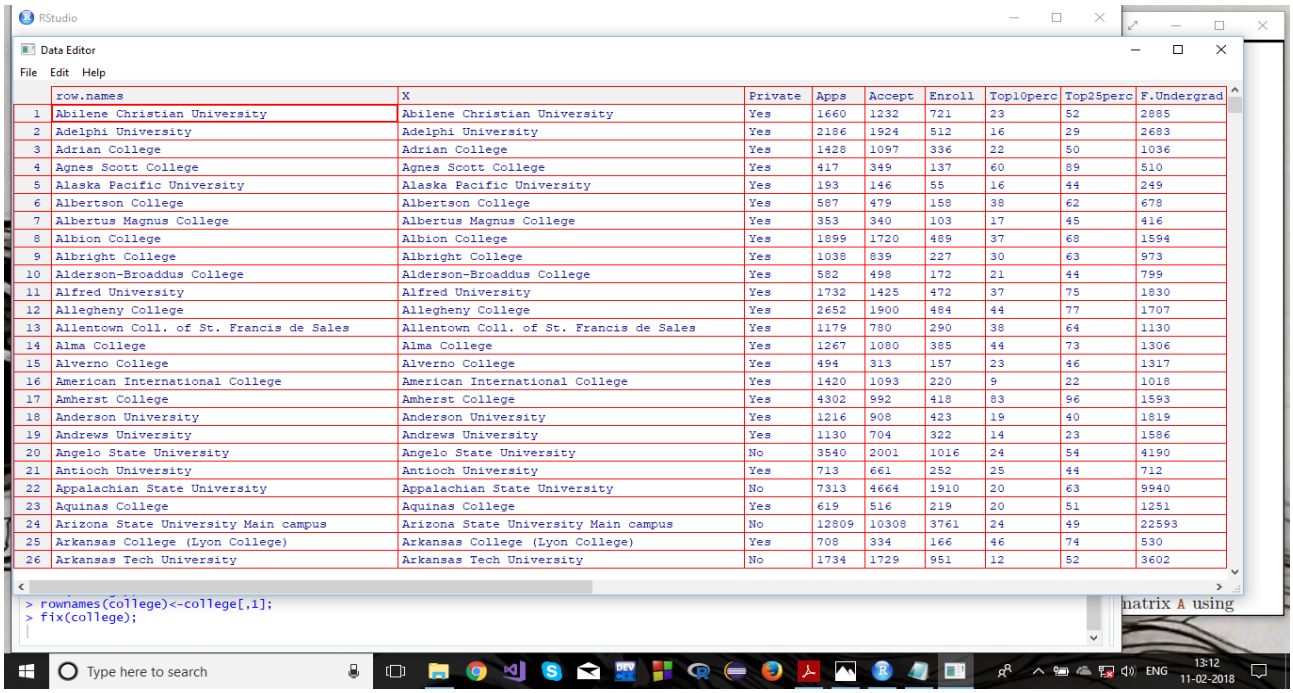
	X	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal
1	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200
2	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500
3	Adrian College	Yes	1428	1097	336	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);
```

Type here to search ENG 12:41 11-02-2018

->add row names

->look of the edited table



RStudio

Data Editor

File Edit Help

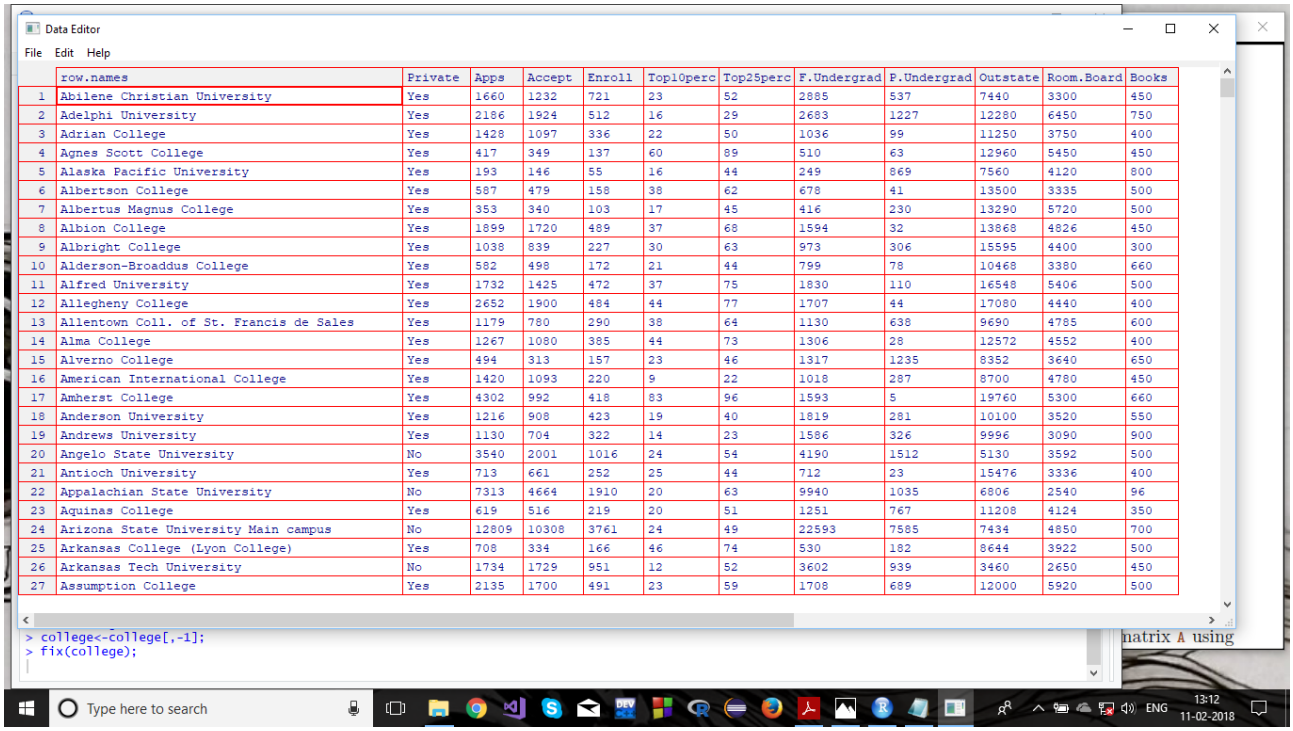
	row.names	X	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad
1	Abilene Christian University	Abilene Christian University	Yes	1660	1232	721	23	52	2885
2	Adelphi University	Adelphi University	Yes	2186	1924	512	16	29	2683
3	Adrian College	Adrian College	Yes	1428	1097	336	22	50	1036
4	Agnes Scott 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);
```

natrix A using

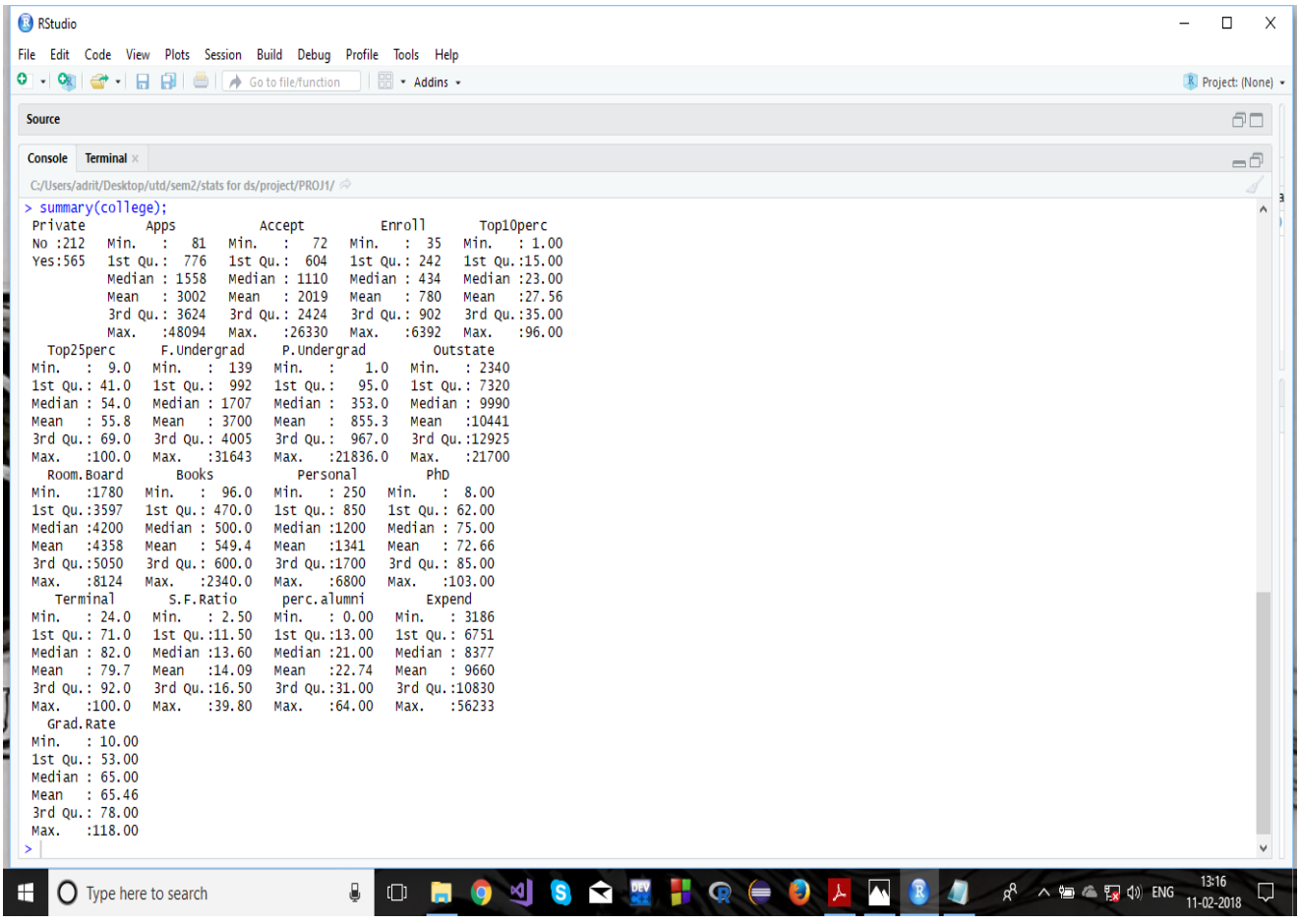
Type here to search ENG 13:12 11-02-2018

->delete college name column that can have calulations on them
->look at edited table

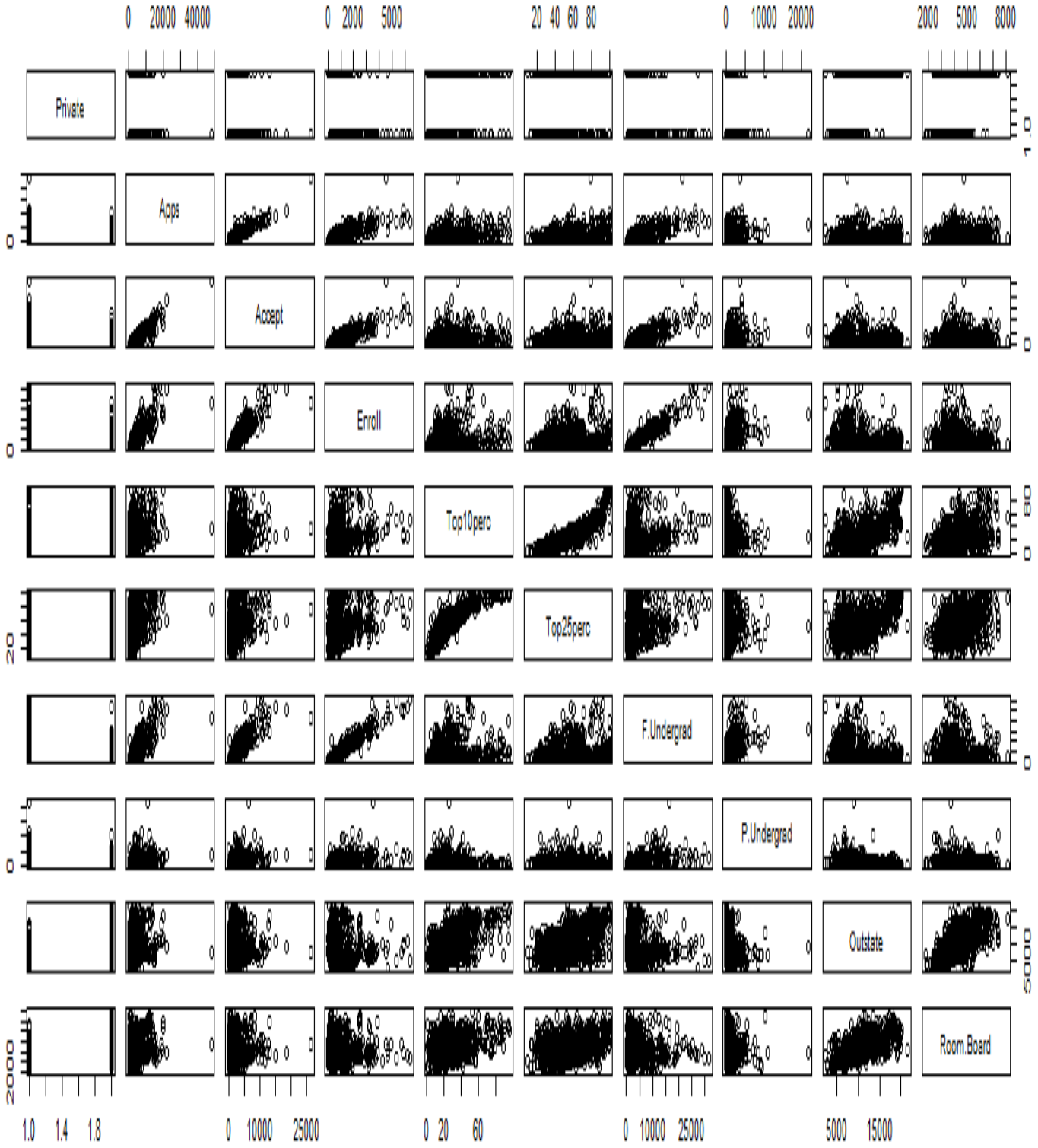


	row.names	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books
1	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450
2	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750
3	Adrian College	Yes	1428	1097	336	22	50	1036	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

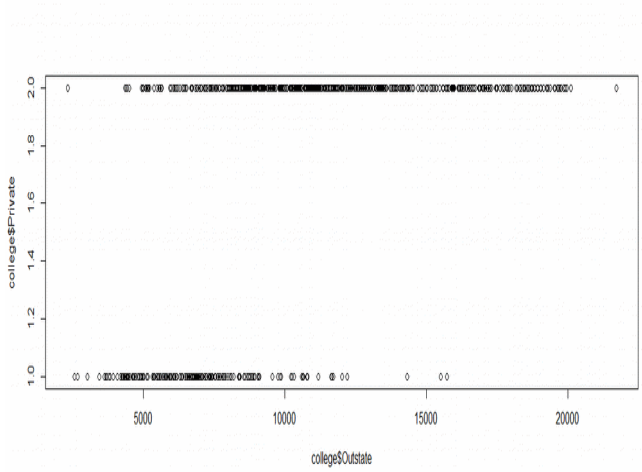
c.i)summary of all data sets

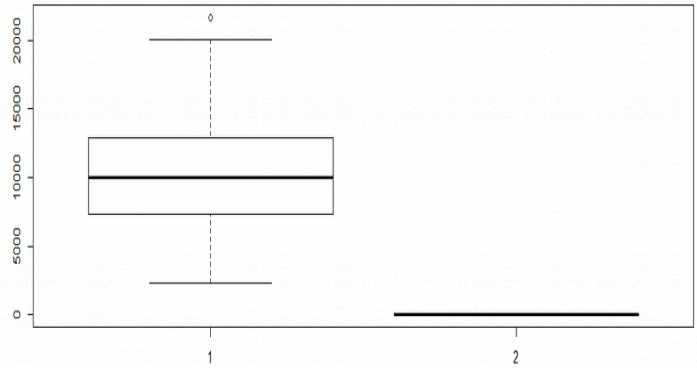


ii)scatterplot matrix of first 10 columns:



iii)side-by-side plot and boxplot of outstate vs private:



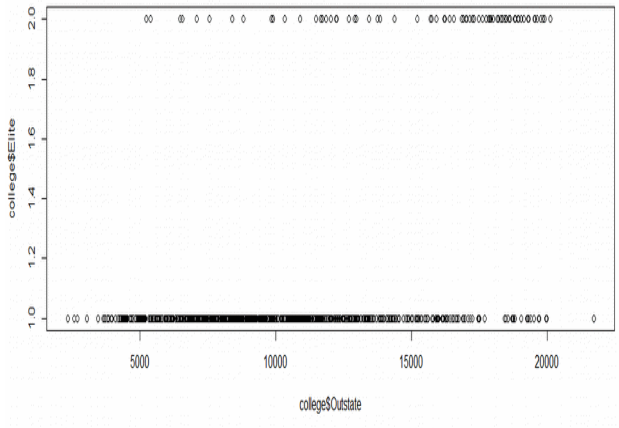


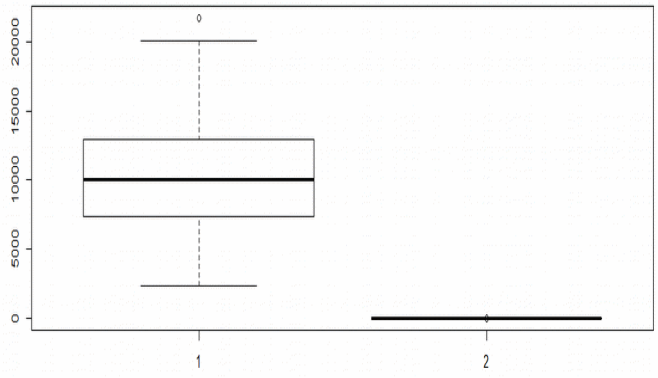
iv)create Elite, by binning the Top10perc variable

->see how many Elites using summary:

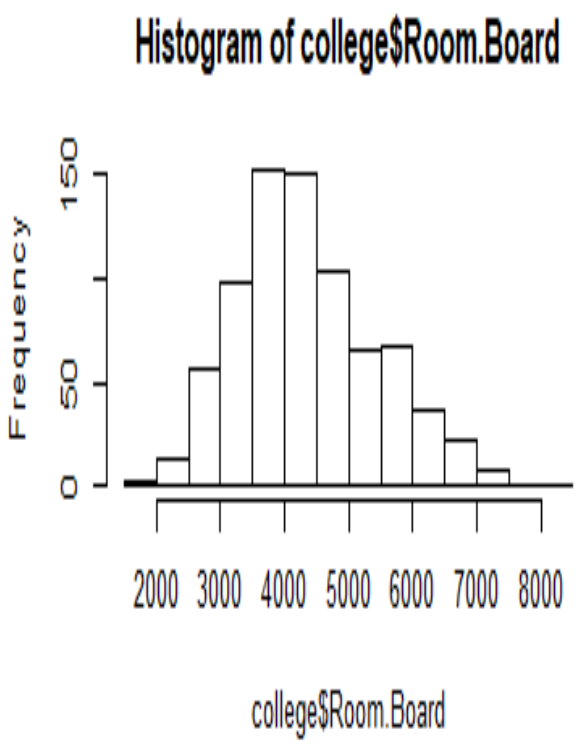
NO	YES
699	78

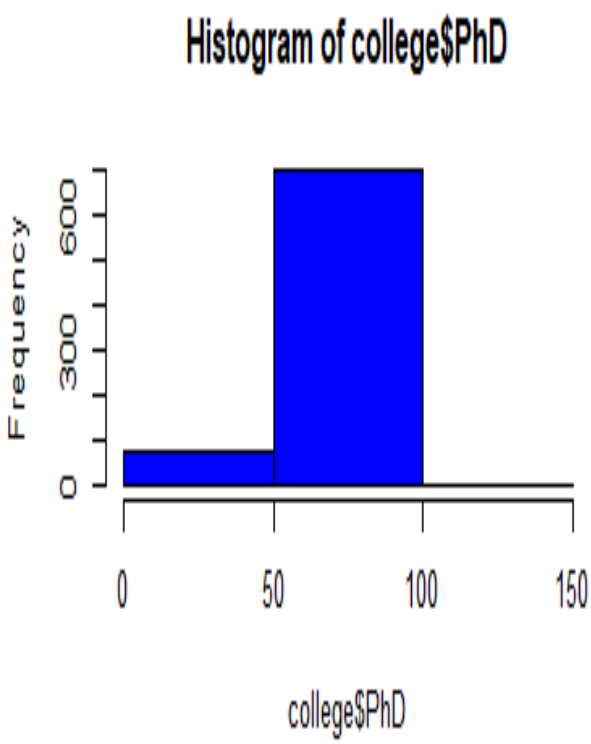
->side-by-side plot and boxplot of outstate vs Elite

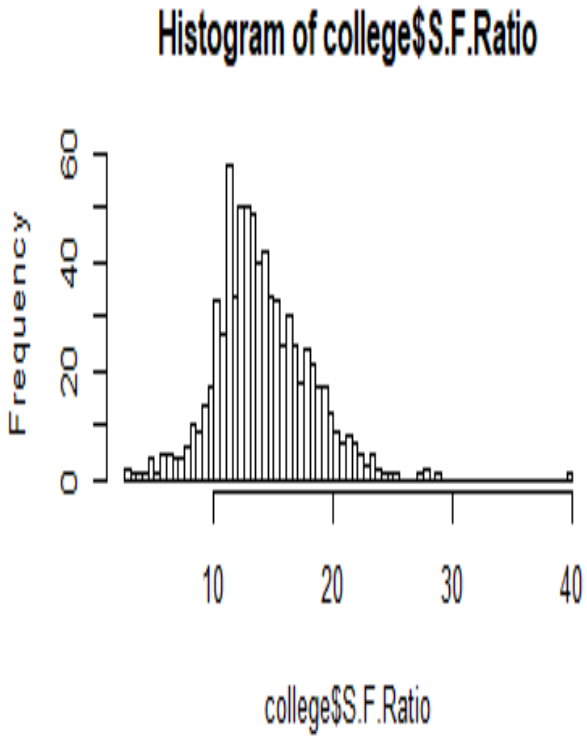


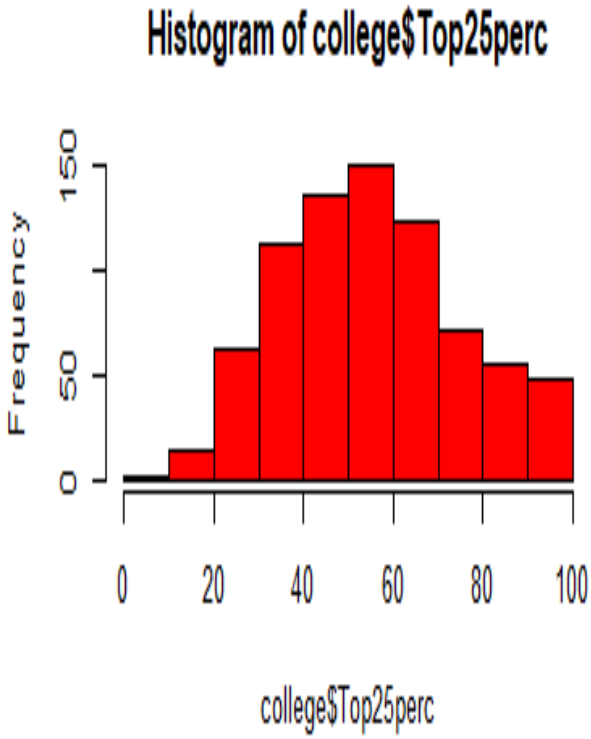


v)histograms of a few variables with different bin size



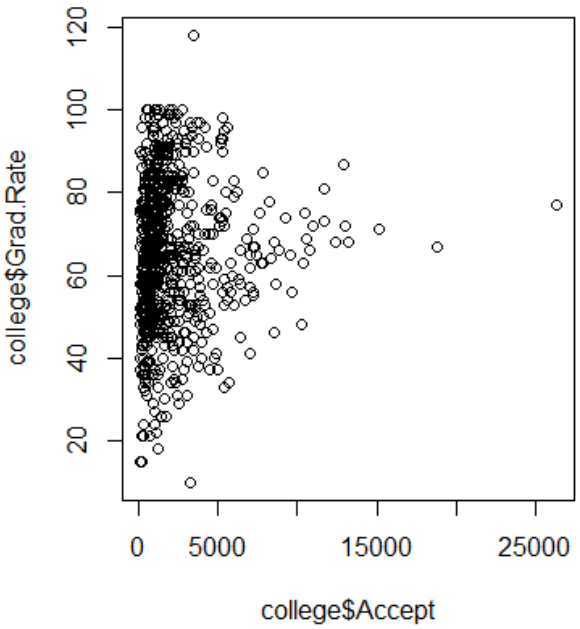


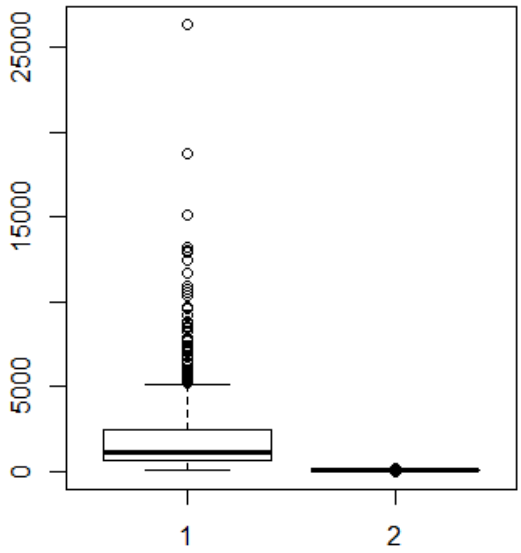




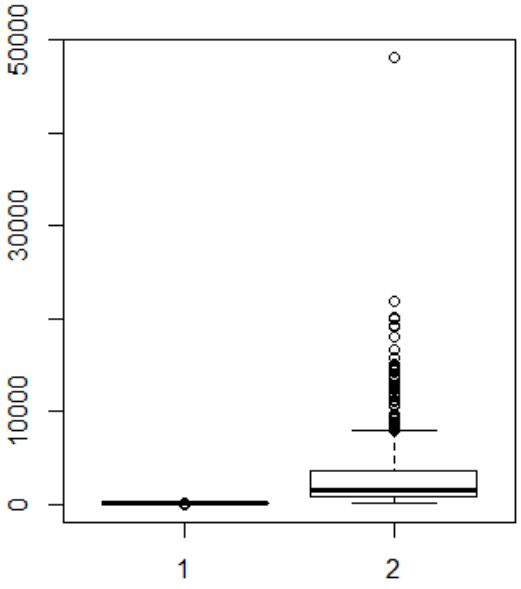
vi) On further experimenting with the given data it is observed that:

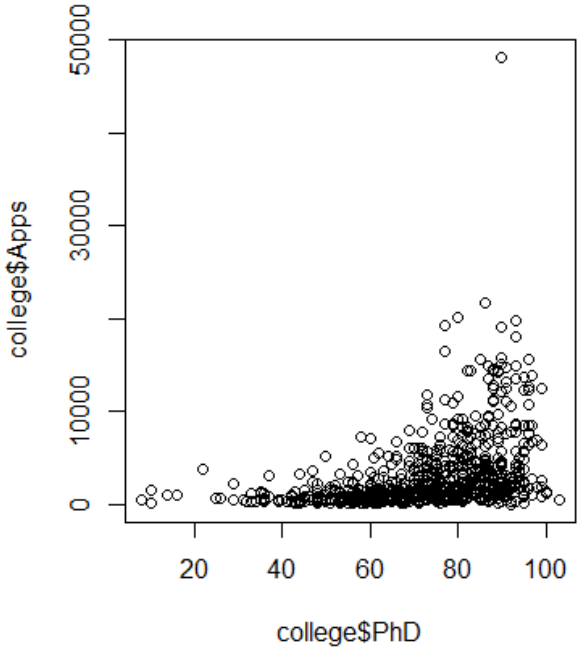
->the higher the acceptance, graduation rate has lesser variance and is thus lower in average



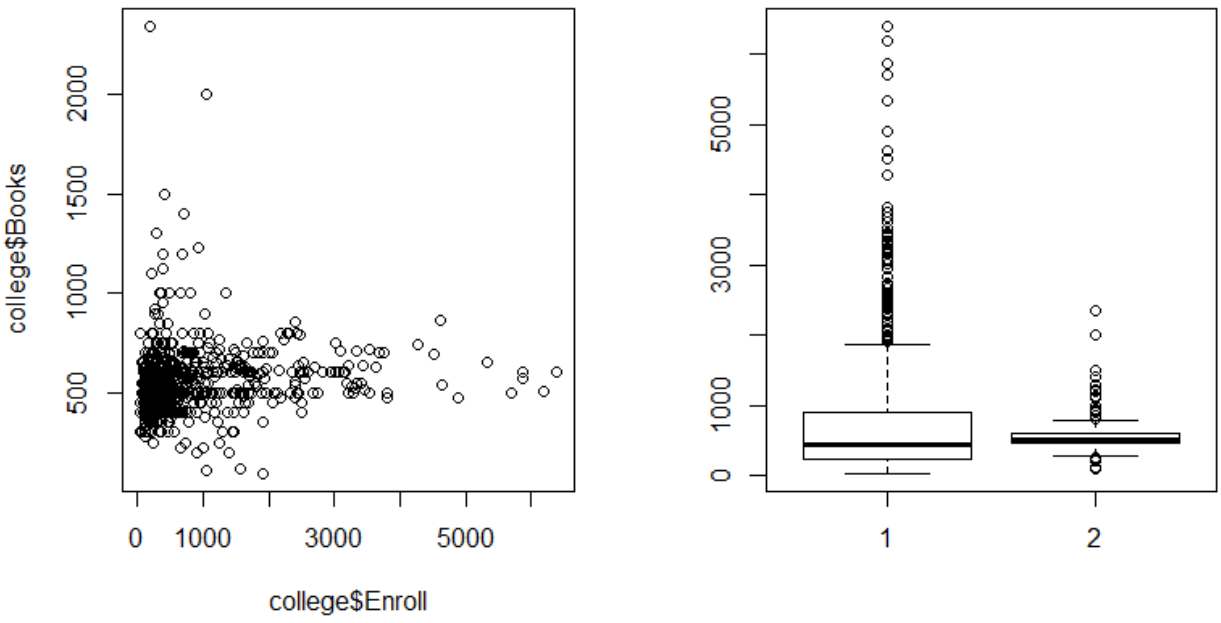


->more applicants if more number of faculty have phd

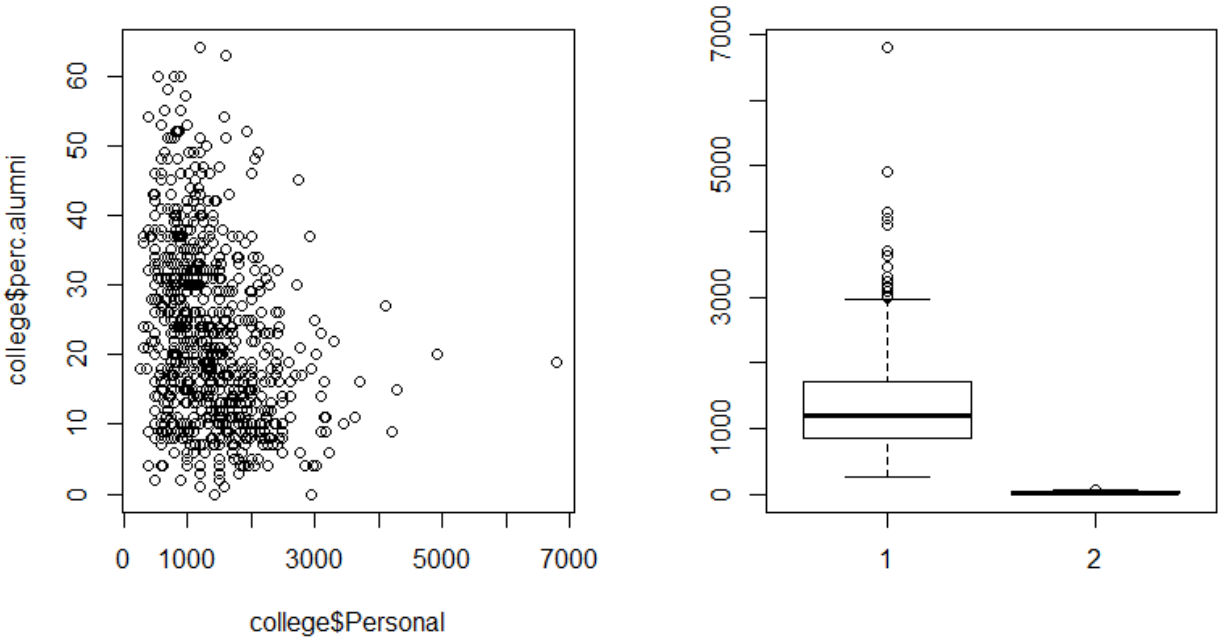




->more variance in costs of books when lesser students enrolled



->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$Outstate,college$Private)
boxplot(college$Outstate,college$Private)
```

```
#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$Outstate,college$Elite)
boxplot(college$Outstate,college$Elite)


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

#histogram of romm 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 =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)
```