

Introduction to Statistical Learning- Lab1

Name: Sandeep Reddy Salkuti

Student id: 16296868

Email: sswf7@umsystem.edu

1)

- 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.

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains an R script with the following code:

```
1 setwd("E:/Fall2020/ISL-1/Lab/Lab1")
2 print(getwd())
3 college=read.csv(file="college.csv", header=TRUE, sep=",")
4 print(college)
5
6
```
- Console:** Displays the output of the script, showing the current working directory and a table of college data with 19 variables and 777 observations.
- Environment Pane:** Shows the loaded data object `college` with 777 observations and 19 variables.
- Files Pane:** Shows the file `college.csv` in the current directory.

The console output for the `print(college)` command is as follows:

		X	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F. Undergrad	P. Undergrad
1	Abilene Christian university	Yes	1660	1232	721	23	52	2885	537	
2	Adelphi university	Yes	2186	1924	512	16	29	2683	1227	
3	Adrian college	Yes	1428	1097	336	22	50	1036	99	
4	Agnes Scott college	Yes	417	349	137	60	89	510	63	
5	Alaska Pacific university	Yes	193	146	55	16	44	249	869	
6	Albertson college	Yes	587	479	158	38	62	678	41	
7	Albertus Magnus college	Yes	353	340	103	17	45	416	230	
8	Albion college	Yes	1899	1720	489	37	68	1594	32	
9	Albright college	Yes	1038	839	227	30	63	973	306	
10	Alderson-Broaddus college	Yes	582	498	172	21	44	799	78	
11	Alfred university	Yes	1732	1425	472	37	75	1830	110	
12	Allegheny college	Yes	2652	1900	484	44	77	1707	44	
13	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64	1130	638	
14	Alma college	Yes	1267	1080	385	44	73	1306	28	
15	Alverno college	Yes	494	313	157	23	46	1317	1235	
16	American International college	Yes	1420	1093	220	9	22	1018	287	
17	Amherst college	Yes	4302	992	418	83	96	1593	5	
18	Anderson university	Yes	1216	908	423	19	40	1819	281	
19	Andrews university	Yes	1130	704	322	14	23	1586	326	
20	Angelo State university	No	3540	2001	1016	24	54	4190	1512	
21	Antioch university	Yes	713	661	252	25	44	712	23	
22	Appalachian State university	No	7313	4664	1910	20	63	9940	1035	
23	Aquinas college	Yes	619	516	219	20	51	1251	767	
24	Arizona State university Main campus	No	12809	10308	3761	24	49	22593	7585	
25	Arkansas college (Lyon college)	Yes	708	334	166	46	74	530	182	
26	Arkansas Tech university	No	1734	1729	951	12	52	3602	939	
27	Assumption college	Yes	2135	1700	491	23	59	1708	689	

```

1 setwd("E:/Fall2020/ISL-1/Lab/Lab1")
2 print(getwd())
3 college=read.csv(file="College.csv", header=TRUE, sep=",")
4 print(college)
5
6

```

	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.	Ratio	perc.alumni	Expend	Grad.Rate
1	7440	3300	450	2200	70	78	18.1		12	7041	60
2	12280	6450	750	1500	29	30	12.2		16	10527	56
3	11250	3750	400	1165	53	66	12.9		30	8735	54
4	12960	5450	450	875	92	97	7.7		37	19016	59
5	7560	4120	800	1500	76	72	11.9		2	10922	15
6	13500	3335	500	675	67	73	9.4		11	9727	55
7	13290	5720	500	1500	90	93	11.5		26	8861	63
8	13868	4826	450	850	89	100	13.7		37	11487	73
9	15595	4400	300	500	79	84	11.3		23	11644	80
10	10468	3380	660	1800	40	41	11.5		15	8991	52
11	16548	5406	500	600	82	88	11.3		31	10932	73
12	17080	4440	400	600	73	91	9.9		41	11711	76
13	9690	4785	600	1000	60	84	13.3		21	7940	74
14	12572	4552	400	400	79	87	15.3		32	9305	68
15	8352	3640	650	2449	36	69	11.1		26	8127	55
16	8700	4780	450	1400	78	84	14.7		19	7355	69
17	19760	5300	660	1598	93	98	8.4		63	21424	100
18	10100	3520	550	1100	48	61	12.1		14	7994	59
19	9996	3090	900	1320	62	66	11.5		18	10908	46
20	5130	3592	500	2000	60	62	23.1		5	4010	34
21	15476	3336	400	1100	69	82	11.3		35	42926	48
22	6806	2540	96	2000	83	96	18.3		14	5854	70
23	11208	4124	350	1615	55	65	12.7		25	6584	65
24	7434	4850	700	2100	88	93	18.9		5	4602	48
25	8644	3922	500	800	79	88	12.6		24	14579	54
26	3460	2650	450	1000	57	60	19.6		5	4739	48
27	12000	5920	500	500	92	92	13.8		30	7100	88

```

1 setwd("E:/Fall2020/ISL-1/Lab/Lab1")
2 print(getwd())
3 college=read.csv(file="College.csv", header=TRUE, sep=",")
4 print(college)
5
6

```

	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.	Ratio	perc.alumni	Expend	Grad.Rate
28	6300	3933	600	1908	85	91	16.7		18	6642	69
29	11902	4372	540	950	65	65	12.8		31	7836	58
30	13353	4173	540	821	78	83	12.7		40	9220	71
31	10990	3244	600	1021	66	70	10.4		30	6871	69
32	11280	4342	400	1150	81	95	13.0		33	11361	71
33	9925	4135	750	1350	59	67	22.4		11	6523	48
34	8620	4100	400	2250	58	68	11.0		21	6136	65
35	10995	4410	1000	1000	68	74	17.6		20	8086	85
36	9690	4300	500	500	57	77	9.7		35	9337	71
37	19264	6206	750	750	98	98	10.4		30	13894	79
38	17926	8124	600	850	83	93	10.3		33	12580	91
39	11290	5360	600	1800	76	78	12.6		11	9084	72
40	6450	3920	600	1346	71	76	18.5		38	7503	72
41	12850	5400	400	800	78	89	12.2		30	8954	73
42	8840	2950	750	1290	74	82	13.1		31	6668	84
43	9000	4850	300	2480	78	85	13.2		10	7550	52
44	7800	3664	650	900	61	61	11.1		19	7614	49
45	16304	3616	355	715	87	95	11.1		26	12957	69
46	4425	2700	660	1800	57	62	19.6		16	3752	46
47	9550	3850	350	250	64	84	14.1		18	5922	58
48	21700	4100	600	500	35	59	10.1		33	16364	55
49	13800	5510	630	850	87	87	17.5		20	10941	82
50	8050	3940	350	2375	80	80	16.3		17	10511	63
51	8740	3363	550	1700	62	68	11.6		29	7718	48
52	8540	3580	500	1400	61	80	8.8		32	8324	56

b) Look at the data using the `fix()` function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands: `> rownames(college)=college[,1]` `> fix(college)` You should see that there is now a `row.names` column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try `> college=college[, -1]` `> fix(college)` Now you should see that

the first data column is Private. Note that another column labeled row.names now appears before the Private column. However, this is not a data column but rather the name that R is giving to each row.

The screenshot shows the RStudio interface. The top pane displays the following R code:

```
1 setwd("E:/Fall2020/ISL-1/Lab/Lab1")
2 print(getwd())
3 #1. a)
4 College=read.csv(file="College.csv", header=TRUE, sep=",")
5 print(College)
6 #1. b)
7 rownames(College)=College[,1]
8 fix(College)
9
```

The bottom pane shows the Data Editor window with a table of college data. The table has 12 columns: row.names, Private, Apps, Accept, Enroll, Top10perc, Top25perc, F.Undergrad, F.Undergrad, and Outstate. The first column, row.names, contains the names of 32 colleges. The other columns contain numerical data for each college.

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

- To eliminate the first column in the data where the names are stored below is the command typed and screenshot attached.

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Untitled1* x

Source on Save Run Source

```

1 setwd("E:/Fall2020/ISL-1/Lab/Lab1")
2 print(getwd())
3 #1. a)
4 College=read.csv(file="college.csv", header=TRUE, sep=",")
5 print(College)
6 #1. b)
7 rownames(College)=College[,1]
8 fix(College)
9 #Eliminate first column in data
10 College=College[,-1]
11 fix(College)
12
13

```

10:1 (Top Level) R Script

Data Editor

File Edit Help

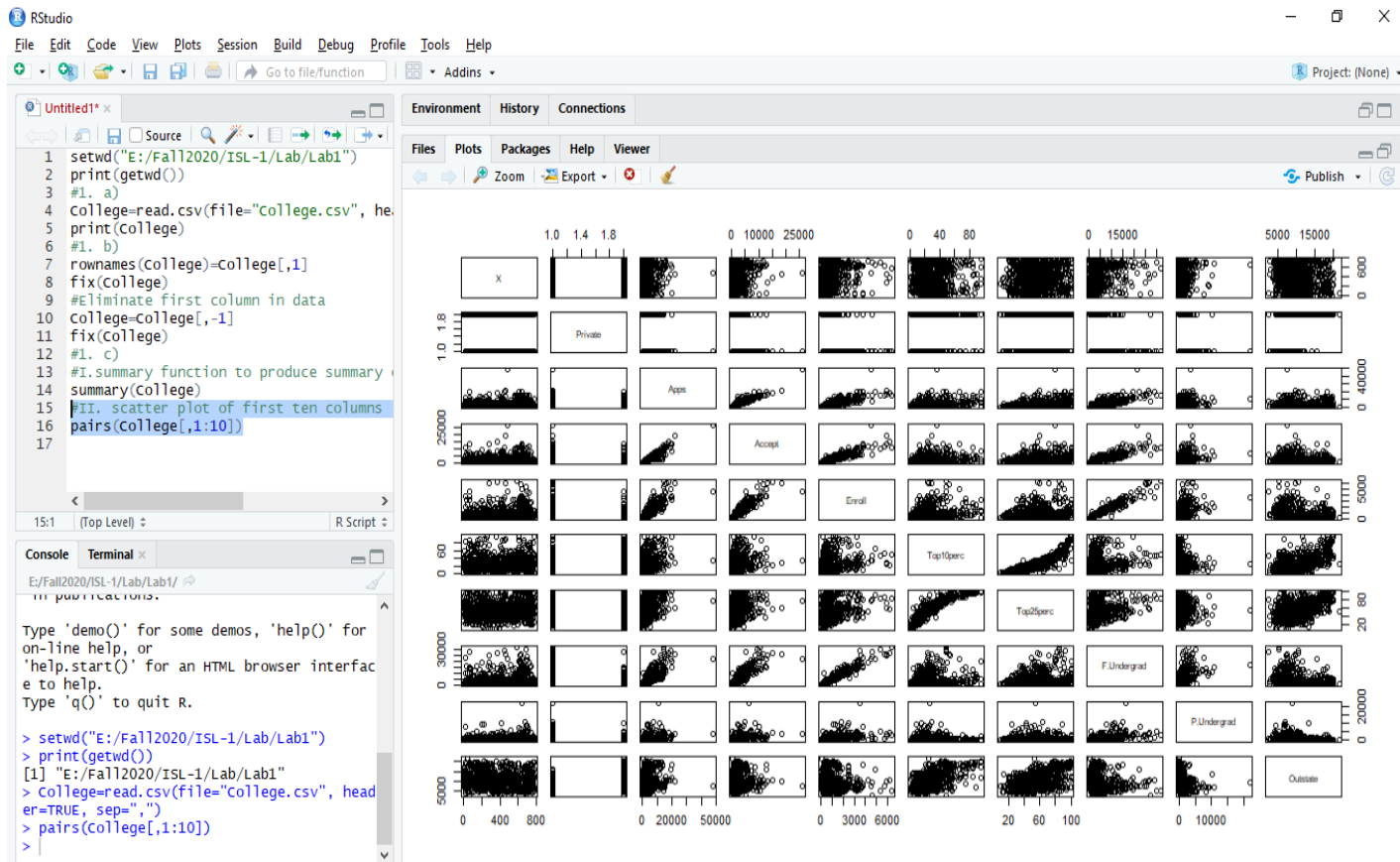
	row.names	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal
1	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200
2	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500
3	Adrian College	Yes	1428	1097	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-Broadus 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
24	Arizona State University Main campus	No	12809	10308	3761	24	49	22593	7585	7434	4850	700	2100
25	Arkansas College (Lyon College)	Yes	708	334	166	46	74	530	182	8644	3922	500	800
26	Arkansas Tech University	No	1734	1729	951	12	52	3602	939	3460	2650	450	1000
27	Assumption College	Yes	2135	1700	491	23	59	1708	689	12000	5920	500	500
28	Auburn University-Main Campus	No	7548	6791	3070	25	57	16262	1716	6300	3933	600	1908
29	Augsburg College	Yes	662	513	257	12	30	2074	726	11902	4372	540	950
30	Augustana College IL	Yes	1879	1658	497	36	69	1950	38	13353	4173	540	821
31	Augustana College	Yes	761	725	306	21	58	1337	300	10990	3244	600	1021
32	Austin College	Yes	948	798	295	42	74	1120	15	11280	4342	400	1150

c)

I) Use the summary() function to produce a numerical summary of the variables in the data set

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Untitled1*
11 #I. c)
12 #I. summary function to produce summary of a variables in dataset
13 summary(College)
14
15
16
16:1 (Top Level) R Script
Console Terminal
E:/Fall2020/ISL-1/Lab/Lab1/
> 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 Room.Board
Min. : 9.0 Min. : 139 Min. : 1.0 Min. : 2340 Min. :1780
1st Qu.: 41.0 1st Qu.: 992 1st Qu.: 95.0 1st Qu.: 7320 1st Qu.:3597
Median : 54.0 Median : 1707 Median : 353.0 Median : 9990 Median :4200
Mean : 55.8 Mean : 3700 Mean : 855.3 Mean :10441 Mean :4358
3rd Qu.: 69.0 3rd Qu.: 4005 3rd Qu.: 967.0 3rd Qu.:12925 3rd Qu.:5050
Max. :100.0 Max. :31643 Max. :21836.0 Max. :21700 Max. :8124
Books Personal PhD Terminal S.F.Ratio
Min. : 96.0 Min. : 250 Min. : 8.00 Min. : 24.0 Min. : 2.50
1st Qu.: 470.0 1st Qu.: 850 1st Qu.: 62.00 1st Qu.: 71.0 1st Qu.:11.50
Median : 500.0 Median :1200 Median : 75.00 Median : 82.0 Median :13.60
Mean : 549.4 Mean :1341 Mean : 72.66 Mean : 79.7 Mean :14.09
3rd Qu.: 600.0 3rd Qu.:1700 3rd Qu.: 85.00 3rd Qu.: 92.0 3rd Qu.:16.50
Max. :2340.0 Max. :6800 Max. :103.00 Max. :100.0 Max. :39.80
perc.alumni Expend Grad.Rate
Min. : 0.00 Min. : 3186 Min. : 10.00
1st Qu.:13.00 1st Qu.: 6751 1st Qu.: 53.00
Median :21.00 Median : 8377 Median : 65.00
Mean :22.74 Mean : 9660 Mean : 65.46
3rd Qu.:31.00 3rd Qu.:10830 3rd Qu.: 78.00
Max. :64.00 Max. :56233 Max. :118.00
>
```

II) Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

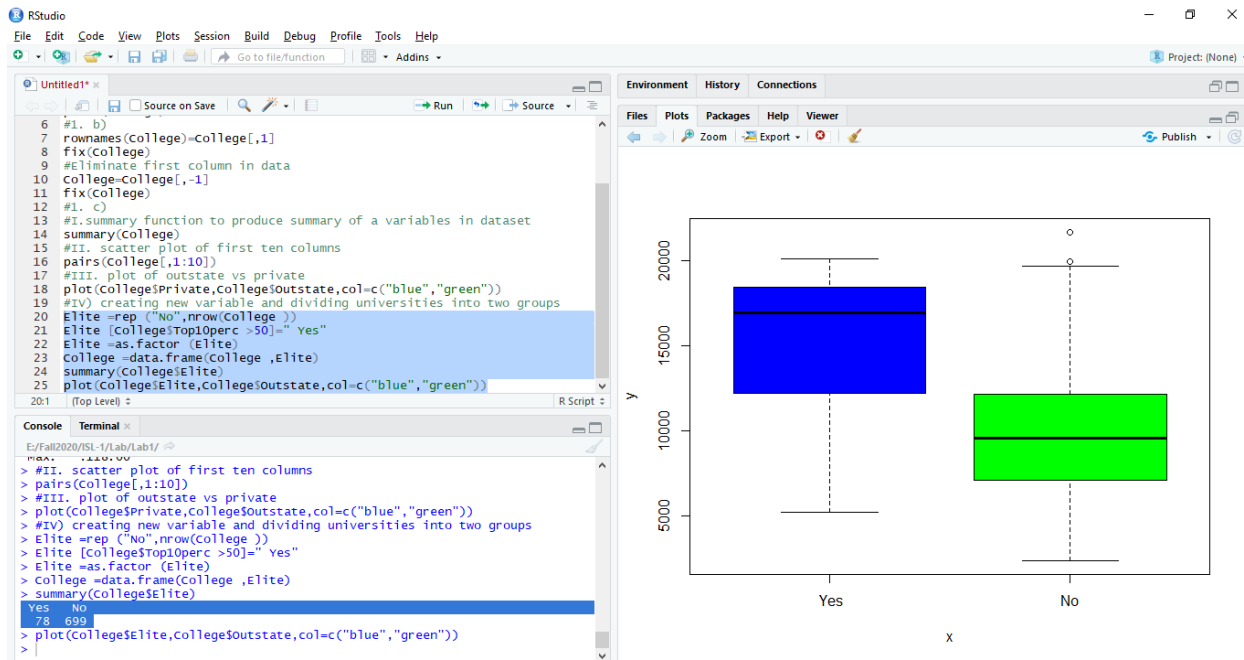


III) Use the plot() function to produce side-by-side boxplots of Outstate versus Private.

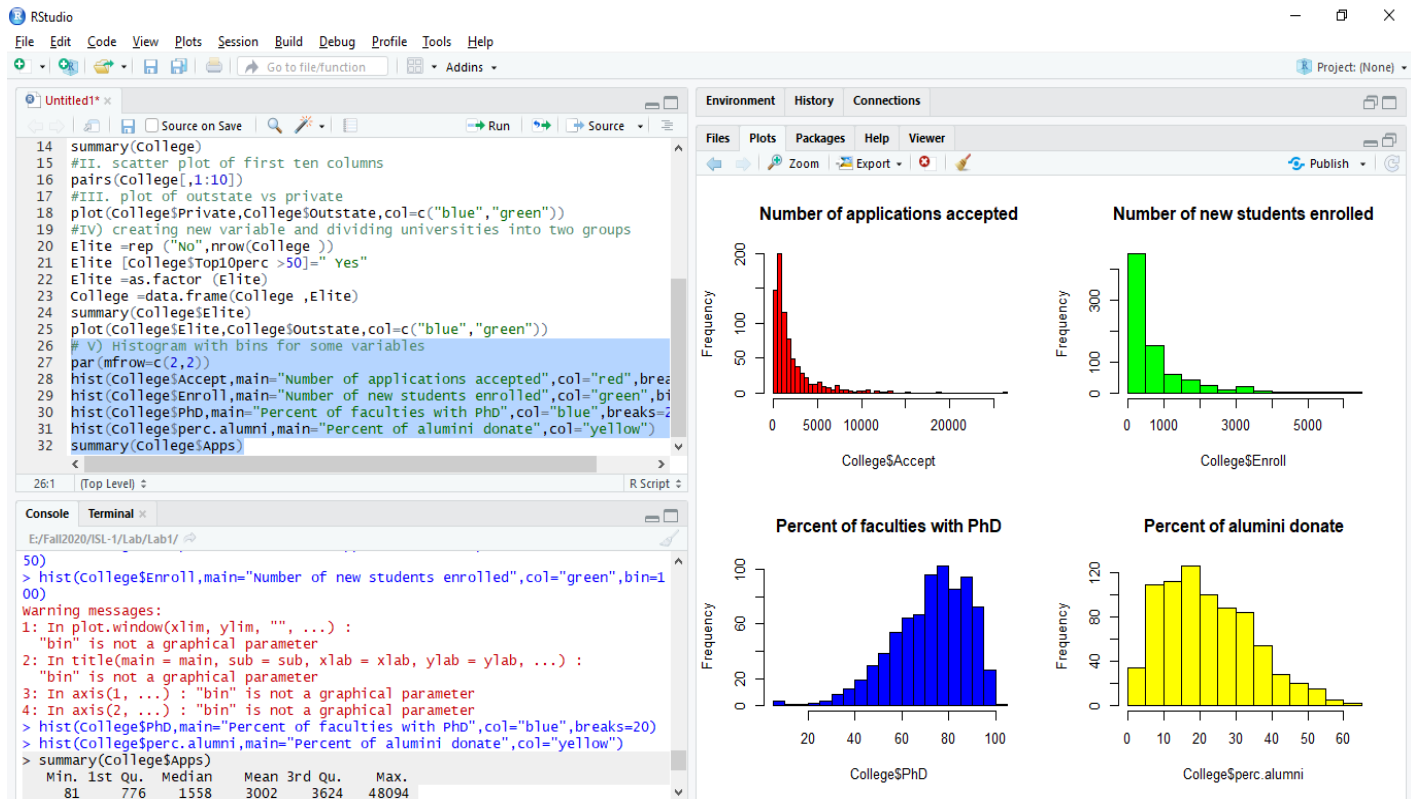


IV) Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from

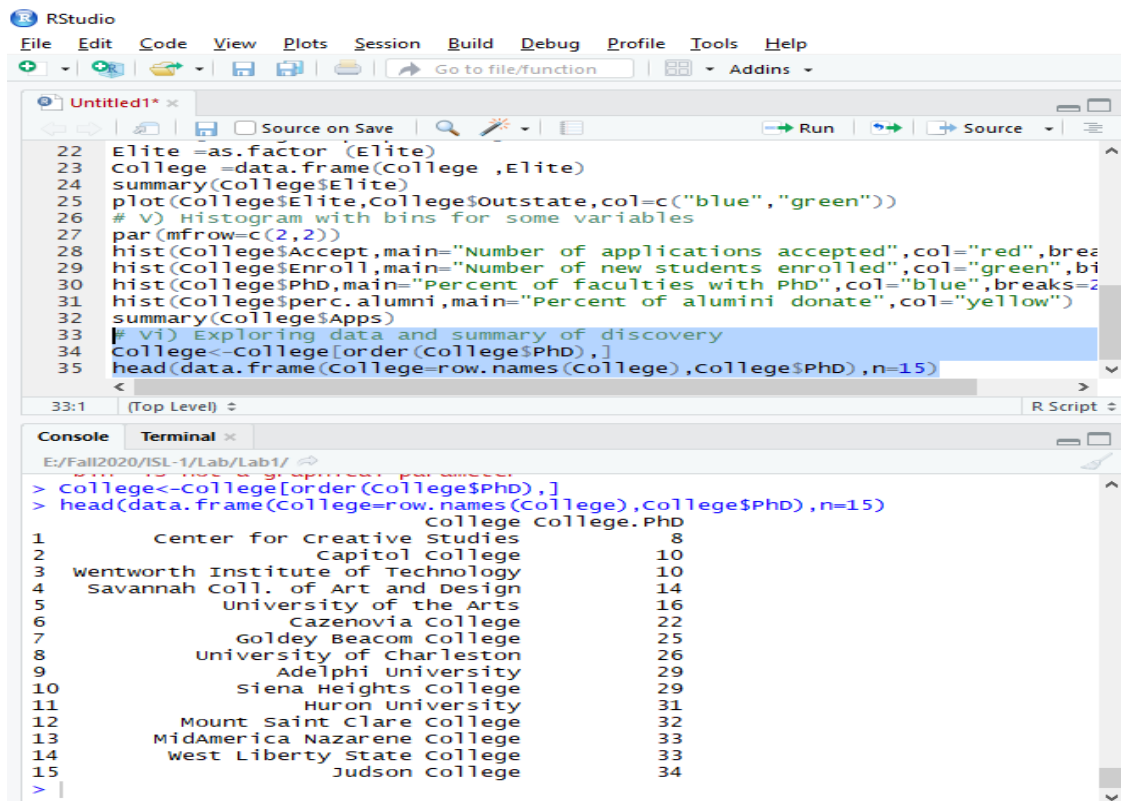
the top 10% of their high school classes exceeds 50%. > Elite =rep ("No",nrow(college)) > Elite [college\$Top10perc >50]=" Yes" > Elite =as.factor (Elite) > college =data.frame(college ,Elite) Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite. v. Use the hist() function to produce some histograms with differing



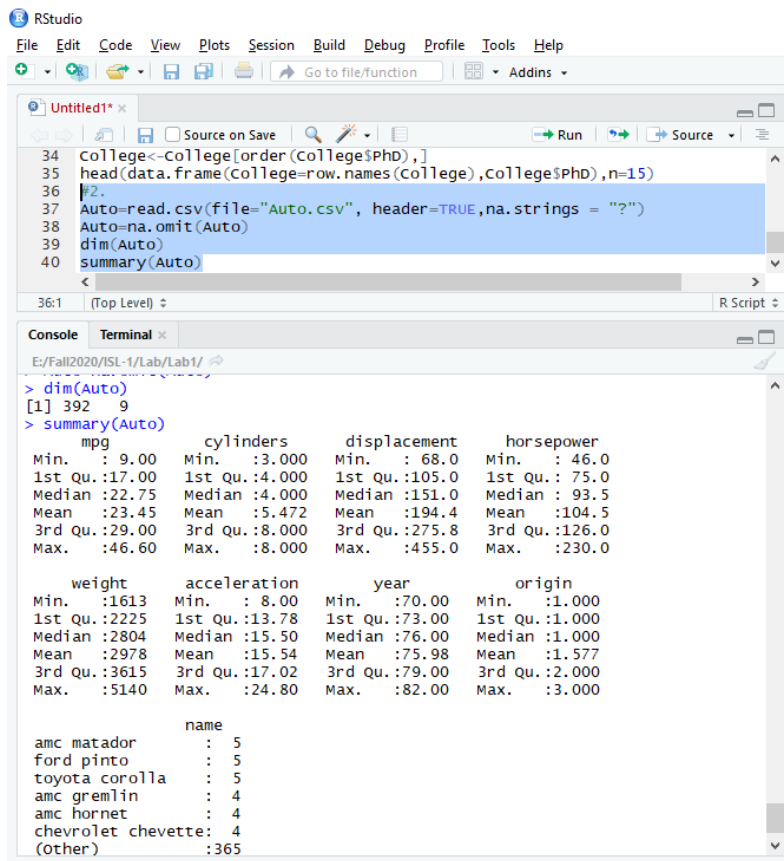
V) Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.



VI) Continue exploring the data, and provide a brief summary of what you discover



2)



The screenshot shows the RStudio interface. The script editor contains the following R code:

```
34 college<-College[order(College$PhD),]
35 head(data.frame(College=row.names(College),college$PhD),n=15)
36 #2.
37 Auto=read.csv(file="Auto.csv", header=TRUE,na.strings = "?")
38 Auto=na.omit(Auto)
39 dim(Auto)
40 summary(Auto)
```

The console shows the output of the commands:

```
> dim(Auto)
[1] 392 9
> summary(Auto)
```

mpg				cylinders	displacement	horsepower	
Min.	: 9.00	Min.	:3.000	Min.	: 68.0	Min.	: 46.0
1st Qu.	:17.00	1st Qu.	:4.000	1st Qu.	:105.0	1st Qu.	: 75.0
Median	:22.75	Median	:4.000	Median	:151.0	Median	: 93.5
Mean	:23.45	Mean	:5.472	Mean	:194.4	Mean	:104.5
3rd Qu.	:29.00	3rd Qu.	:8.000	3rd Qu.	:275.8	3rd Qu.	:126.0
Max.	:46.60	Max.	:8.000	Max.	:455.0	Max.	:230.0

weight		acceleration	year	origin	
Min.	:1613	Min.	: 8.00	Min.	:1.000
1st Qu.	:2225	1st Qu.	:13.78	1st Qu.	:1.000
Median	:2804	Median	:15.50	Median	:1.000
Mean	:2978	Mean	:15.54	Mean	:1.577
3rd Qu.	:3615	3rd Qu.	:17.02	3rd Qu.	:2.000
Max.	:5140	Max.	:24.80	Max.	:3.000

name	
amc matador	: 5
ford pinto	: 5
toyota corolla	: 5
amc gremlin	: 4
amc hornet	: 4
chevrolet chevette	: 4
(other)	:365

a) Which of the predictors are quantitative, and which are qualitative?

quantitative variables: mpg, cylinders, displacement, horsepower, weight, acceleration

qualitative variables: year, origin, name

```
40 summary(Auto)
41 #a.
42 range(Auto$mpg)
43 range(Auto$weight)
44 summary(Auto[,c(7:9),])
45 summary(Auto[,c(1:6),])
```

```
41:1 (Top Level) R Script
```

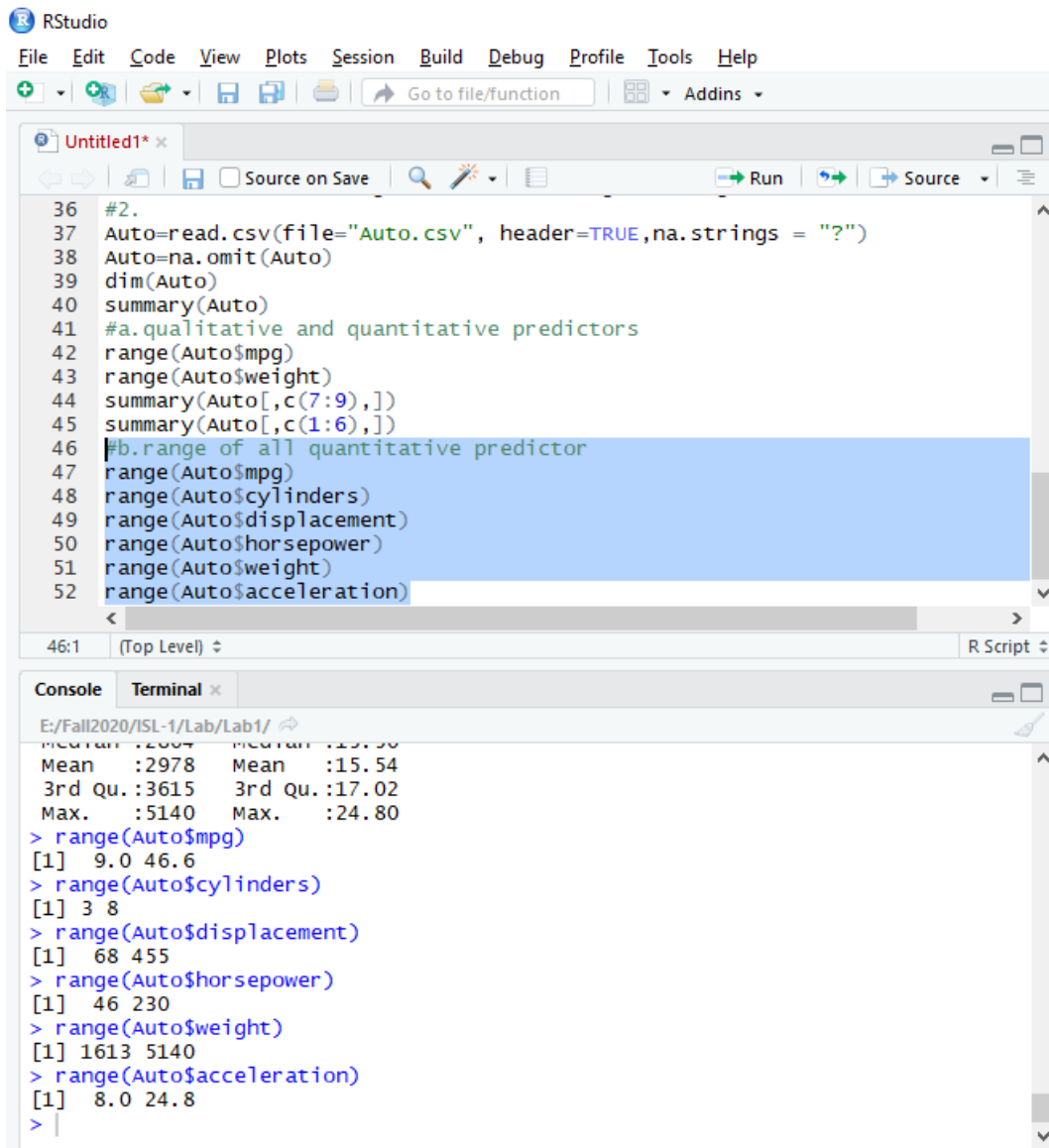
Console Terminal

E:/Fall2020/ISL-1/Lab/Lab1/

```
> range(Auto$mpg)
[1] 9.0 46.6
> range(Auto$weight)
[1] 1613 5140
> summary(Auto[,c(7:9),])
  year      origin      name
Min.  :70.00  Min.  :1.000  amc matador      : 5
1st Qu.:73.00 1st Qu.:1.000  ford pinto       : 5
Median :76.00 Median :1.000  toyota corolla   : 5
Mean   :75.98 Mean   :1.577  amc gremlin      : 4
3rd Qu.:79.00 3rd Qu.:2.000  amc hornet       : 4
Max.   :82.00 Max.   :3.000  chevrolet chevette: 4
              (other) :365

> summary(Auto[,c(1:6),])
  mpg      cylinders      displacement      horsepower
Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0
1st Qu.:17.00 1st Qu.:4.000   1st Qu.:105.0 1st Qu.: 75.0
Median :22.75 Median :4.000   Median :151.0 Median : 93.5
Mean   :23.45 Mean   :5.472   Mean   :194.4 Mean  :104.5
3rd Qu.:29.00 3rd Qu.:8.000   3rd Qu.:275.8 3rd Qu.:126.0
Max.   :46.60 Max.   :8.000   Max.   :455.0 Max.   :230.0
  weight      acceleration
Min.   :1613   Min.   : 8.00
1st Qu.:2225 1st Qu.:13.78
Median :2804 Median :15.50
Mean   :2978 Mean   :15.54
3rd Qu.:3615 3rd Qu.:17.02
Max.   :5140 Max.   :24.80
```

b) What is the range of each quantitative predictor? You can answer this using the `range()` function.



```
#2.
Auto=read.csv(file="Auto.csv", header=TRUE, na.strings = "?")
Auto=na.omit(Auto)
dim(Auto)
summary(Auto)
#a.qualitative and quantitative predictors
range(Auto$mpg)
range(Auto$weight)
summary(Auto[,c(7:9),])
summary(Auto[,c(1:6),])
#b.range of all quantitative predictor
range(Auto$mpg)
range(Auto$cylinders)
range(Auto$displacement)
range(Auto$horsepower)
range(Auto$weight)
range(Auto$acceleration)
```

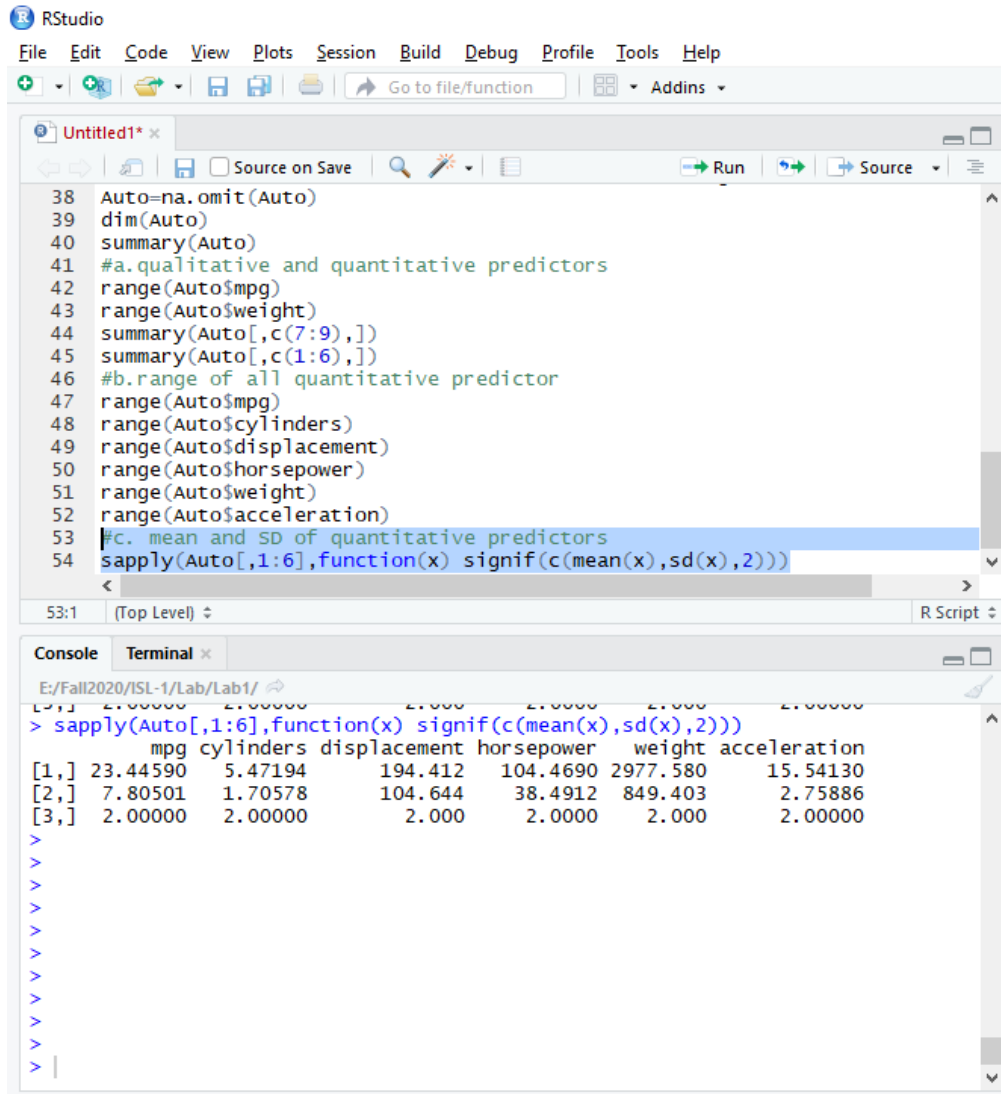
46:1 (Top Level) R Script

Console Terminal x

E:/Fall2020/ISL-1/Lab/Lab1/

```
Median :2804      Median :15.50
Mean   :2978      Mean   :15.54
3rd Qu.:3615      3rd Qu.:17.02
Max.   :5140      Max.   :24.80
> range(Auto$mpg)
[1] 9.0 46.6
> range(Auto$cylinders)
[1] 3 8
> range(Auto$displacement)
[1] 68 455
> range(Auto$horsepower)
[1] 46 230
> range(Auto$weight)
[1] 1613 5140
> range(Auto$acceleration)
[1] 8.0 24.8
>
```

c) What is the mean and standard deviation of each quantitative predictor?



The screenshot shows the RStudio interface. The source editor contains the following R code:

```
38 Auto=na.omit(Auto)
39 dim(Auto)
40 summary(Auto)
41 #a.qualitative and quantitative predictors
42 range(Auto$mpg)
43 range(Auto$weight)
44 summary(Auto[,c(7:9),])
45 summary(Auto[,c(1:6),])
46 #b.range of all quantitative predictor
47 range(Auto$mpg)
48 range(Auto$cylinders)
49 range(Auto$displacement)
50 range(Auto$horsepower)
51 range(Auto$weight)
52 range(Auto$acceleration)
53 #c. mean and SD of quantitative predictors
54 sapply(Auto[,1:6],function(x) signif(c(mean(x),sd(x),2)))
```

The console shows the output of the last command, displaying the mean and standard deviation for the first six predictors (mpg, cylinders, displacement, horsepower, weight, acceleration) for the first three rows of the data:

```
E:/Fall2020/ISL-1/Lab/Lab1/
> sapply(Auto[,1:6],function(x) signif(c(mean(x),sd(x),2)))
      mpg cylinders displacement horsepower  weight acceleration
[1,] 23.44590    5.47194    194.412    104.4690 2977.580    15.54130
[2,]  7.80501    1.70578    104.644     38.4912  849.403     2.75886
[3,]  2.00000    2.00000      2.000      2.0000   2.000     2.00000
>
>
>
>
>
>
>
>
>
```

d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

The screenshot shows the RStudio interface. The script editor contains the following R code:

```

43 range(Auto$weight)
44 summary(Auto[,c(7:9),])
45 summary(Auto[,c(1:6),])
46 #b.range of all quantitative predictor
47 range(Auto$mpg)
48 range(Auto$cylinders)
49 range(Auto$displacement)
50 range(Auto$horsepower)
51 range(Auto$weight)
52 range(Auto$acceleration)
53 #c. mean and sd of quantitative predictors
54 sapply(Auto[,1:6],function(x) signif(c(mean(x),sd(x),2)))
55 #d. removing 10-85th observations and range,mean,sd of remaining predictors
56 new.auto=subset(Auto[-c(10:85),])
57 sapply(new.auto[, -c(9)],range)
58 sapply(new.auto[, -c(9)],mean)
59 sapply(new.auto[, -c(9)],sd)

```

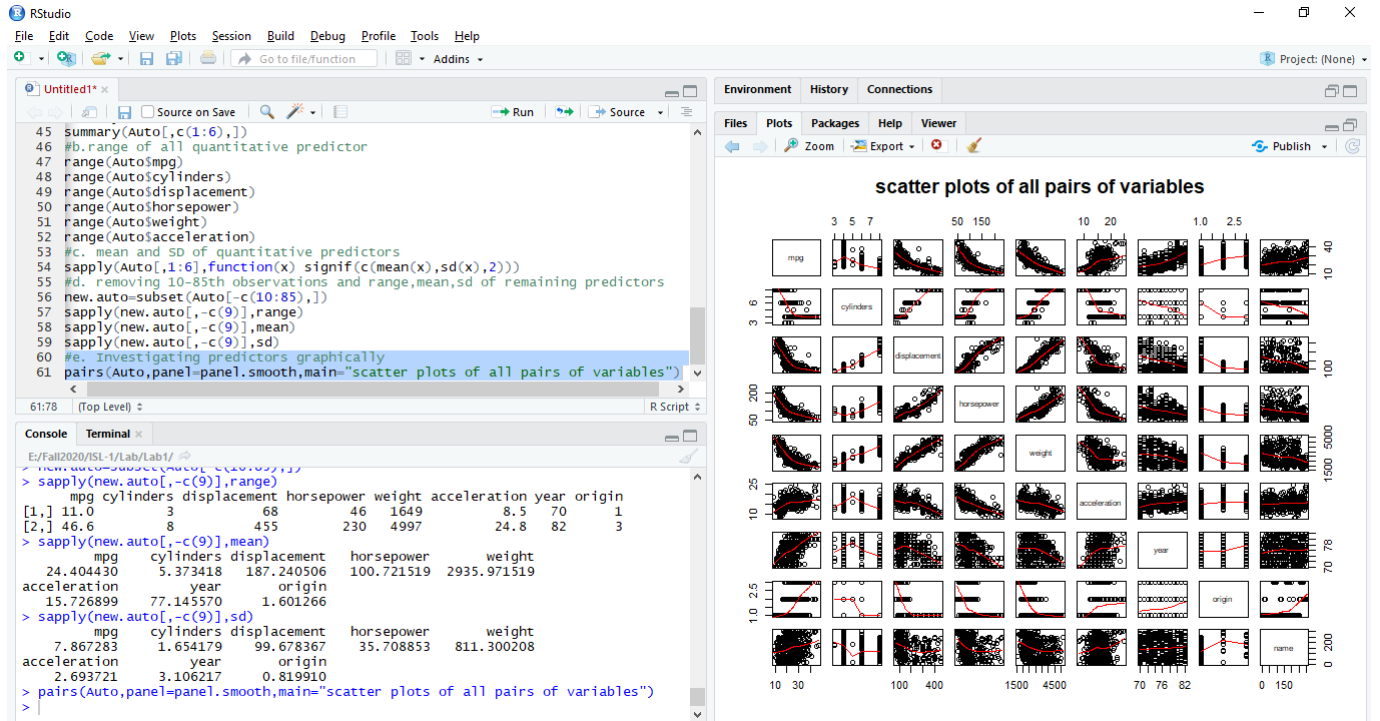
The console shows the output of the executed commands:

```

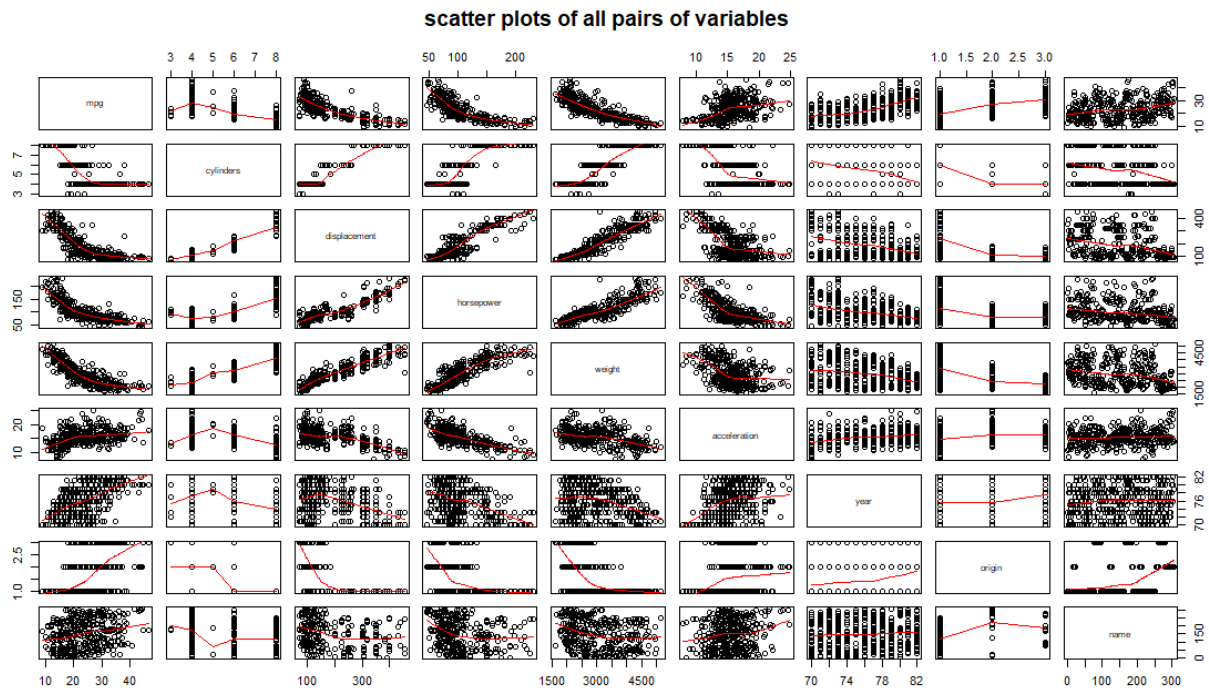
> new.auto=subset(Auto[-c(10:85),])
> sapply(new.auto[, -c(9)],range)
      mpg cylinders displacement horsepower weight acceleration year origin
[1,] 11.0         3          68         46   1649          8.5    70      1
[2,] 46.6         8         455        230   4997         24.8    82      3
> sapply(new.auto[, -c(9)],mean)
      mpg cylinders displacement horsepower weight
24.404430  5.373418  187.240506  100.721519 2935.971519
acceleration year      origin
15.726899  77.145570  1.601266
> sapply(new.auto[, -c(9)],sd)
      mpg cylinders displacement horsepower weight
7.867283  1.654179  99.678367  35.708853  811.300208
acceleration year      origin
2.693721  3.106217  0.819910
>

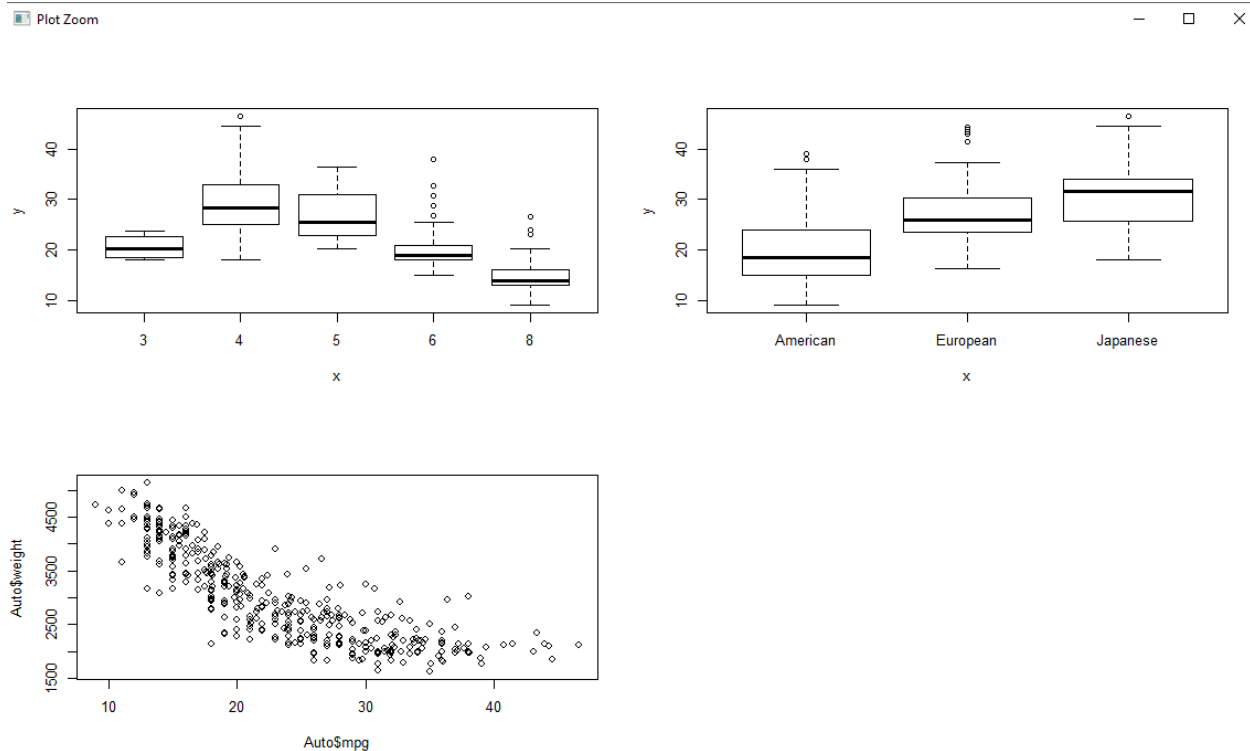
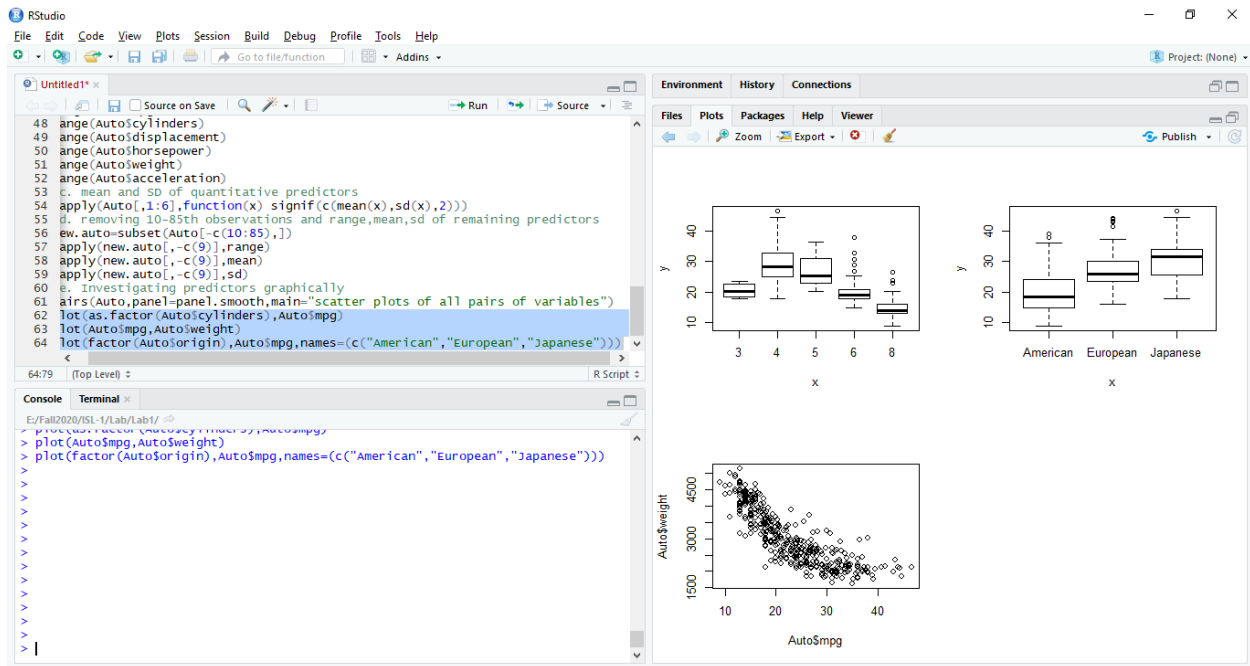
```

e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.



Plot Zoom



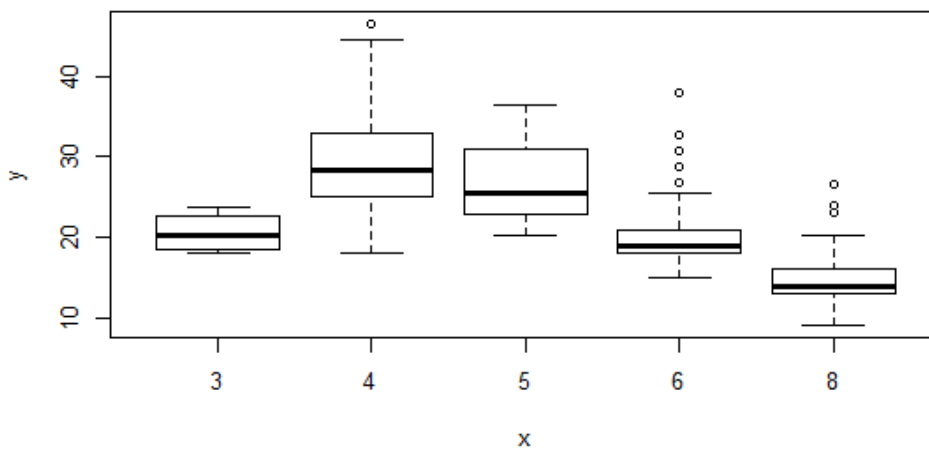


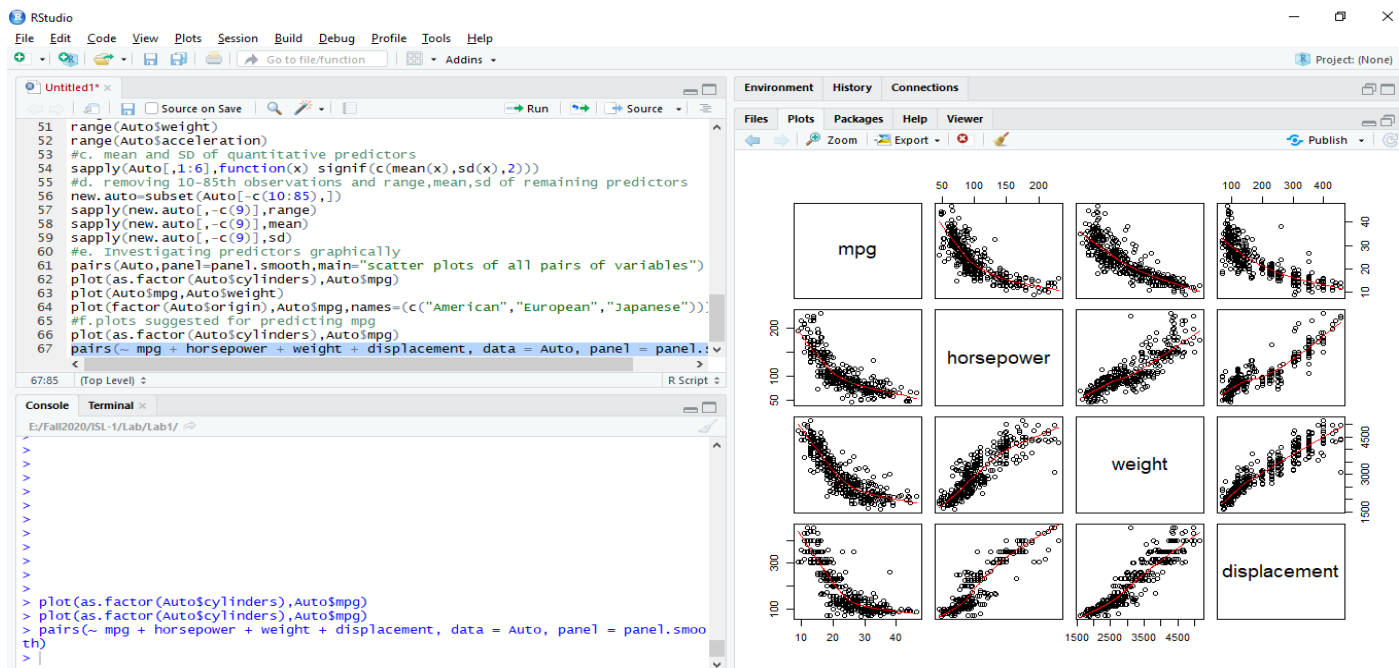
By looking at above graphs Weight, displacement and horse power seems to have an inverse effect with mpg. While displacement with horse power are directly proportional.

f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.


```
51 range(Auto$weight)
52 range(Auto$acceleration)
53 #c. mean and SD of quantitative predictors
54 sapply(Auto[,1:6],function(x) signif(c(mean(x),sd(x),2)))
55 #d. removing 10-85th observations and range,mean,sd of remaining predictors
56 new.auto=subset(Auto[-c(10:85),])
57 sapply(new.auto[, -c(9)],range)
58 sapply(new.auto[, -c(9)],mean)
59 sapply(new.auto[, -c(9)],sd)
60 #e. Investigating predictors graphically
61 pairs(Auto,panel=panel.smooth,main="scatter plots of all pairs of variables")
62 plot(as.factor(Auto$cylinders),Auto$mpg)
63 plot(Auto$mpg,Auto$weight)
64 plot(factor(Auto$origin),Auto$mpg,names=c("American","European","Japanese"))
65 #f.plots suggested for predicting mpg
66 plot(as.factor(Auto$cylinders),Auto$mpg)
```

Plot Zoom

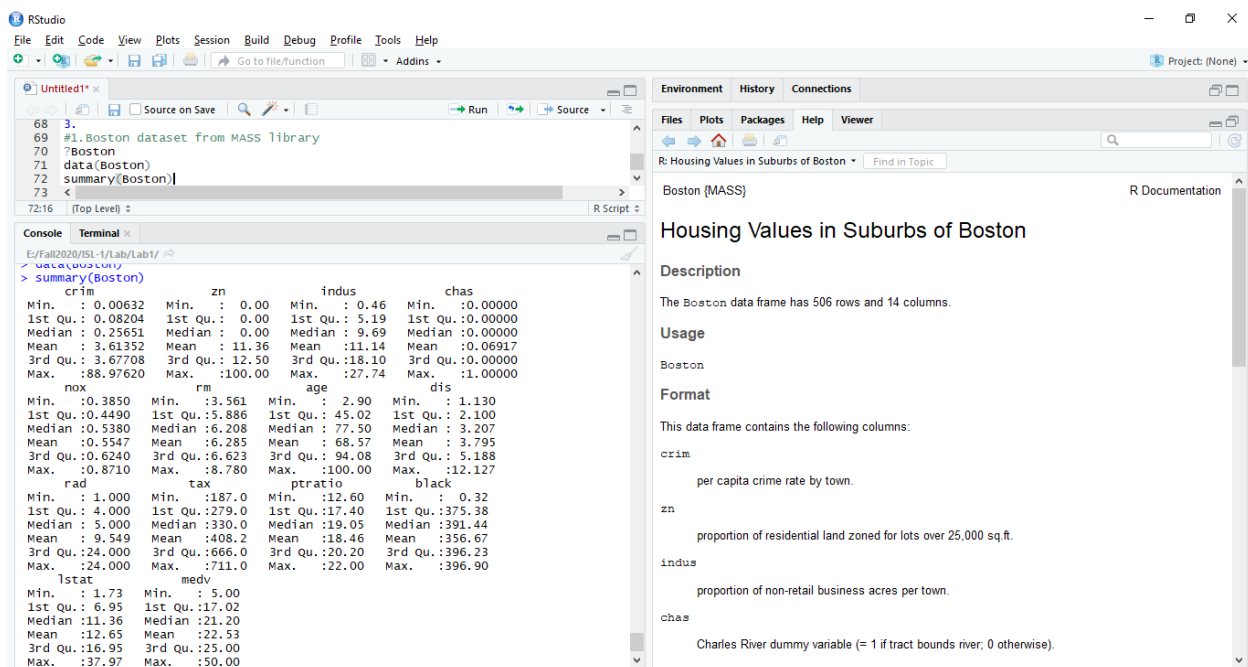




3)

a) To begin, load in the Boston data set. The Boston data set is part of the MASS library in R. `> library(MASS)` Now the data set is contained in the object Boston. `> Boston` Read about the data set: `> ?Boston` How many rows are in this data set? How many columns? What do the rows and columns represent?

For downloading Boston dataset as it is part of MASS library in R lets install package where dataset exists.



The screenshot shows the RStudio environment. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. Below the menu is a toolbar with icons for saving, running, and other functions. The main editor window, titled 'Untitled1*', contains the following R code:

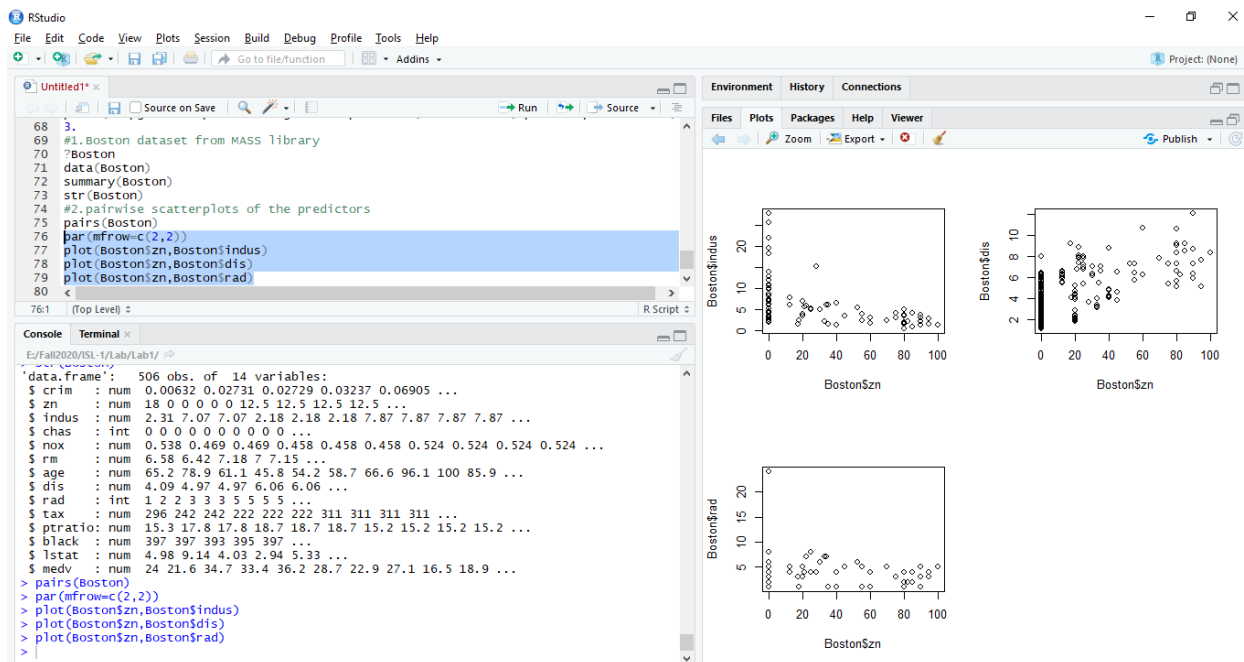
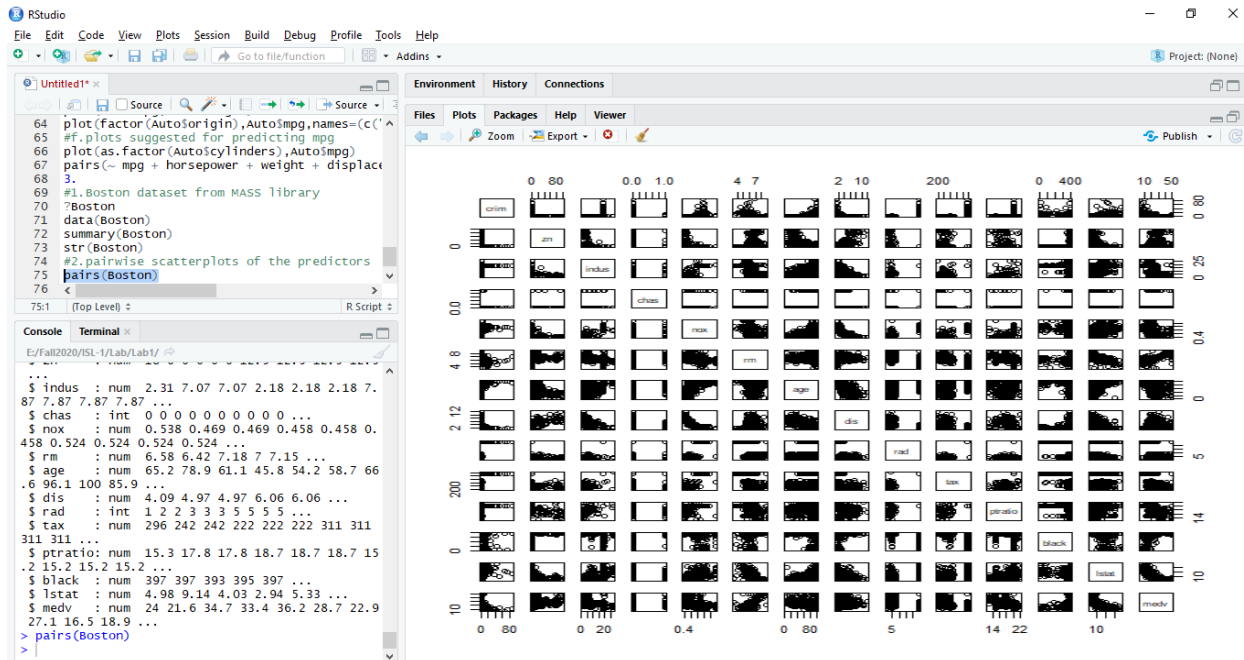
```
63 plot(Auto$mpg,Auto$weight)
64 plot(factor(Auto$origin),Auto$mpg,names=c("American","European","Japanese"))
65 #f.plots suggested for predicting mpg
66 plot(as.factor(Auto$cylinders),Auto$mpg)
67 pairs(~ mpg + horsepower + weight + displacement, data = Auto, panel = panel.s
68 3.
69 #1.Boston dataset from MASS library
70 ?Boston
71 data(Boston)
72 summary(Boston)
73 str(Boston)
74
```

Below the editor is a console window showing the output of the commands. It displays summary statistics for the 'Auto' dataset and the structure of the 'Boston' dataset.

```
E:/Fall2020/ISL-1/Lab/Lab1/
Median :11.36   Median :21.20
Mean   :12.65   Mean   :22.53
3rd Qu.:16.95   3rd Qu.:25.00
Max.   :37.97   Max.   :50.00
> str(Boston)
'data.frame':  506 obs. of  14 variables:
 $ crim    : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ zn      : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ indus   : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ chas    : int   0 0 0 0 0 0 0 0 0 0 ...
 $ nox     : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ rm      : num  6.58 6.42 7.18 7 7.15 ...
 $ age     : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ dis     : num  4.09 4.97 4.97 6.06 6.06 ...
 $ rad     : int   1 2 2 3 3 3 5 5 5 5 ...
 $ tax     : num  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio : num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ black   : num  397 397 393 395 397 ...
 $ lstat   : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv    : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
>
```

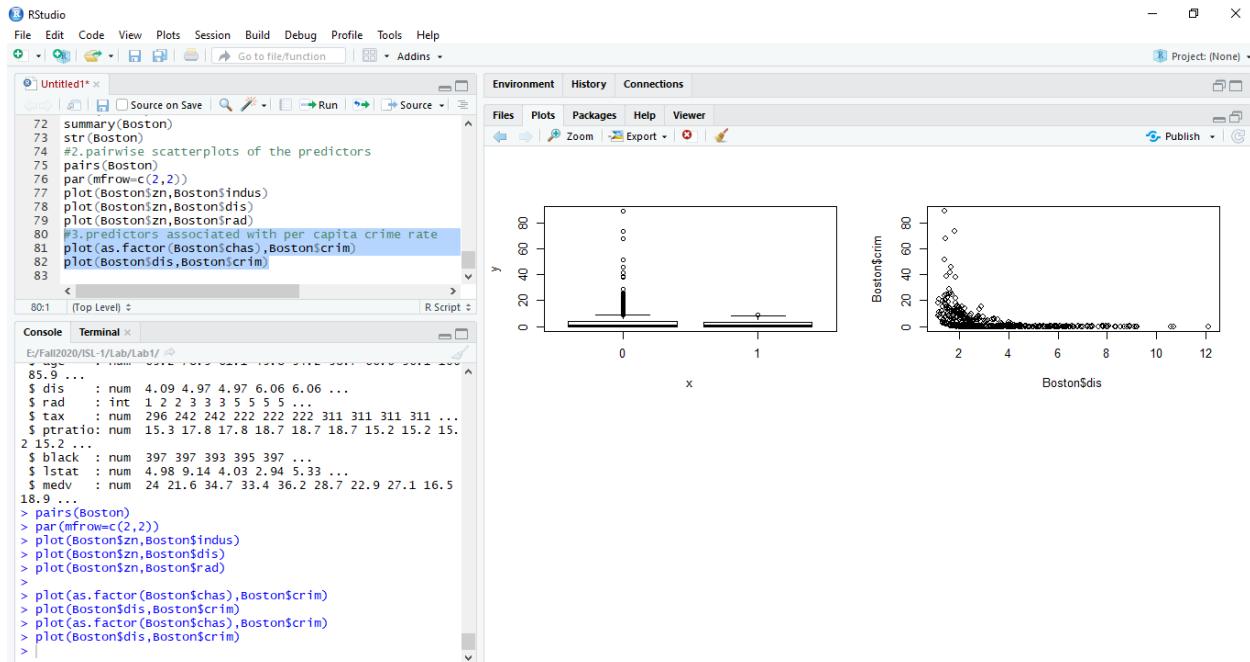
From above there 506 rows and 14 columns

b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

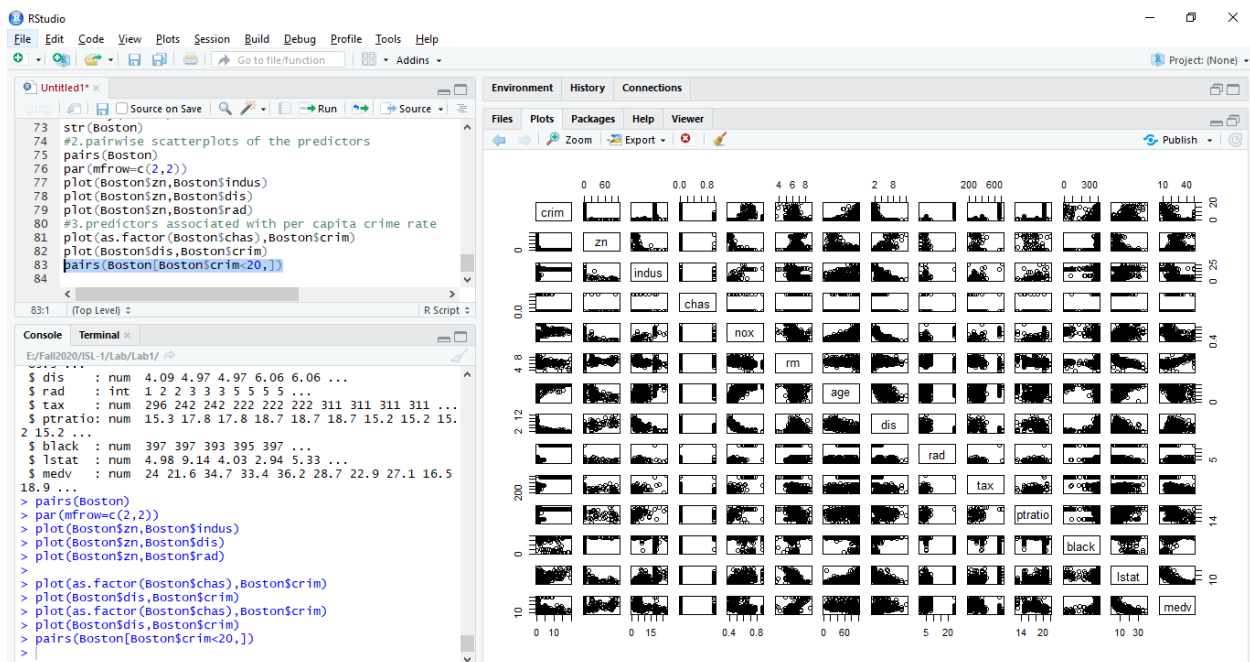


There will be number of scatter plots like above it will become difficult to read all, so maybe a heatmap will be easier to read. Data cleaning is hard.

c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

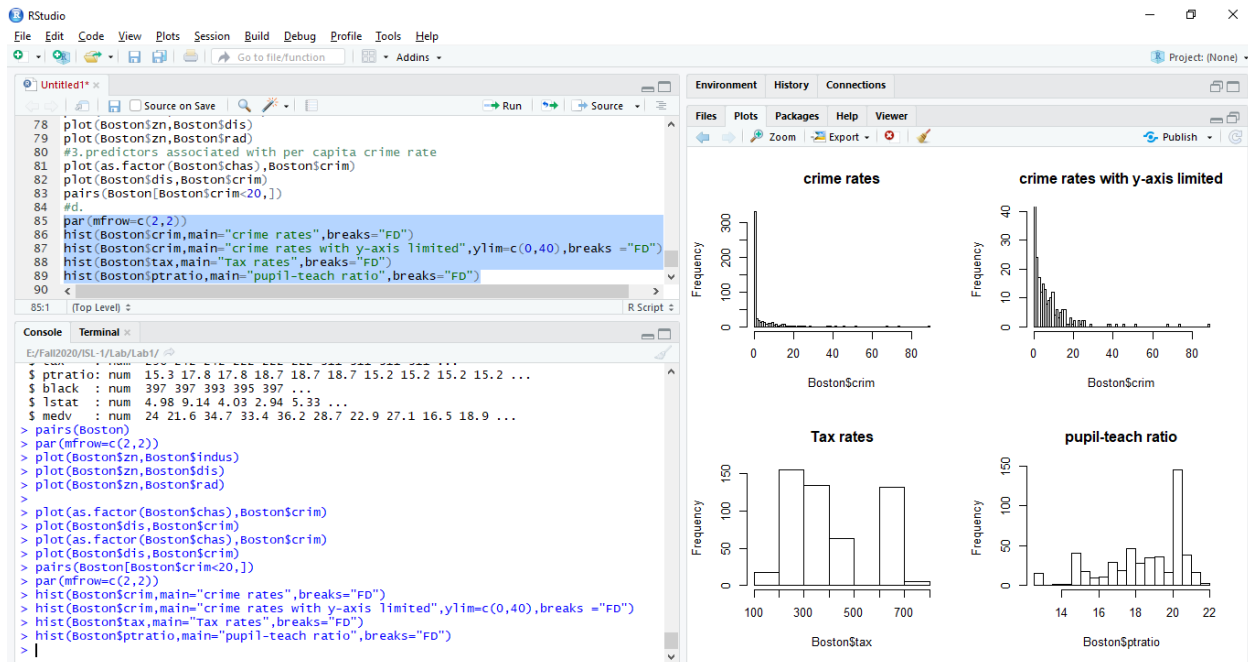


From above diagram crime is large close to five Boston employment centers



From above Lower status populated areas have the more crime rate.

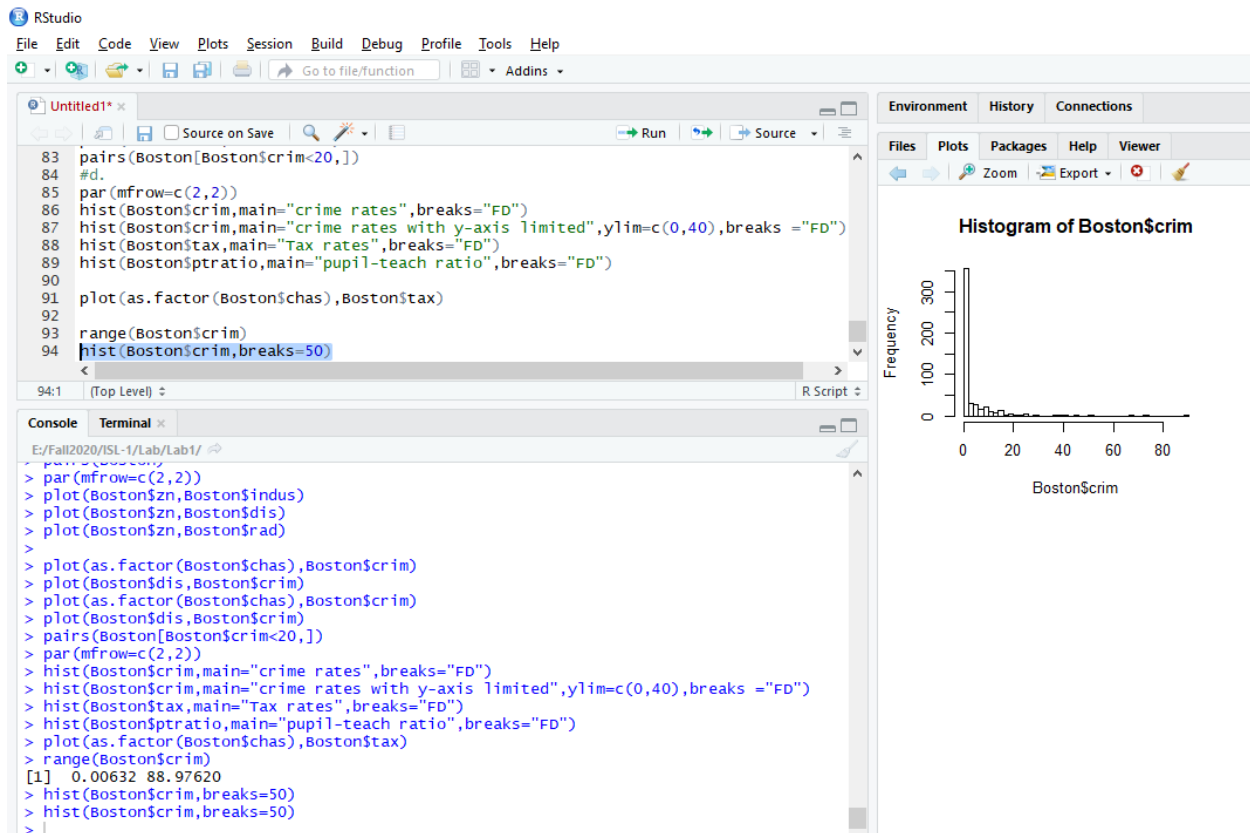
d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.



Most of the suburbs do not have any crime rate.



From above surprisingly tax is less near river area



e) How many of the suburbs in this data set bound the Charles river

```
> table(Boston$chas)
```

```

  0    1
471  35

```

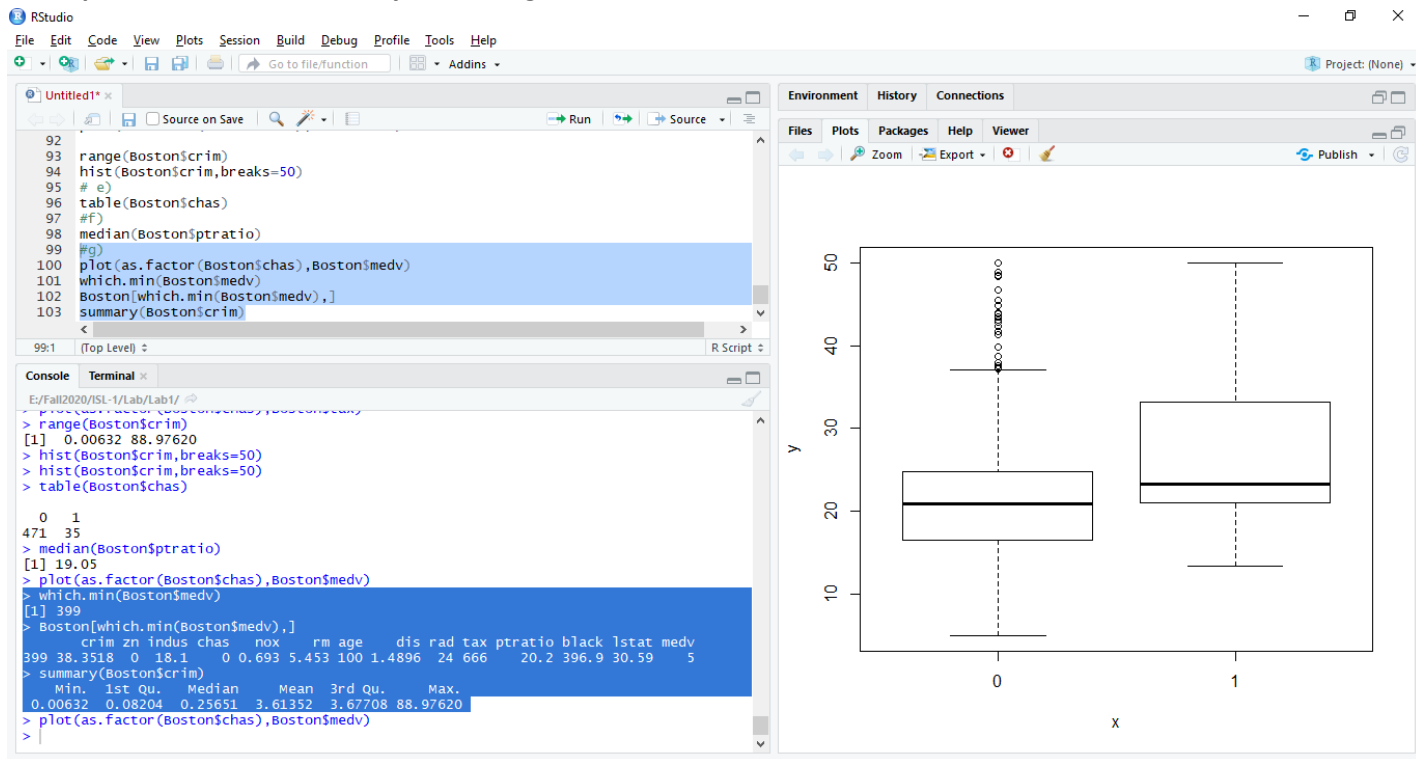
f) What is the median pupil-teacher ratio among the towns in this data set?

```
> median(Boston$ptratio)
```

```
[1] 19.05
```

g) Which suburb of Boston has lowest median value of owner occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for

those predictors? Comment on your findings



Median 0.25 , Maximum is 88.7 and the crime in the median value of owner occupied homes is 38.3518 and we can see that the crime is larger in this area.. It is far from radial highways and Charles river area.

h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

```

> summary(Boston$rm)
      Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
 3.561    5.886    6.208    6.285    6.623    8.780

```

Average is around 6.285 rooms

```

> table(Boston$rm > 7)

```

```

FALSE  TRUE
  442    64

```

From above more than 7 rooms is 64 houses

```

> table(Boston$rm >8)

```

```

FALSE  TRUE
  493    13

```

More than 8 rooms – 13 houses and these have the lesser crime rate.

```
> rooms8 = Boston[Boston$rm > 8, ]  
> summary(rooms8)
```

crim	zn	indus	chas		
Min. :0.02009	Min. : 0.00	Min. : 2.680	Min. :0.0000		
1st Qu.:0.33147	1st Qu.: 0.00	1st Qu.: 3.970	1st Qu.:0.0000		
Median :0.52014	Median : 0.00	Median : 6.200	Median :0.0000		
Mean :0.71879	Mean :13.62	Mean : 7.078	Mean :0.1538		
3rd Qu.:0.57834	3rd Qu.:20.00	3rd Qu.: 6.200	3rd Qu.:0.0000		
Max. :3.47428	Max. :95.00	Max. :19.580	Max. :1.0000		
nox	rm	age	dis	rad	
Min. :0.4161	Min. :8.034	Min. : 8.40	Min. :1.801	Min. : 2.	
000					
1st Qu.:0.5040	1st Qu.:8.247	1st Qu.:70.40	1st Qu.:2.288	1st Qu.: 5.	
000					
Median :0.5070	Median :8.297	Median :78.30	Median :2.894	Median : 7.	
000					
Mean :0.5392	Mean :8.349	Mean :71.54	Mean :3.430	Mean : 7.	
462					
3rd Qu.:0.6050	3rd Qu.:8.398	3rd Qu.:86.50	3rd Qu.:3.652	3rd Qu.: 8.	
000					
Max. :0.7180	Max. :8.780	Max. :93.90	Max. :8.907	Max. :24.	
000					
tax	ptratio	black	lstat	medv	
Min. :224.0	Min. :13.00	Min. :354.6	Min. :2.47	Min. :21.9	
1st Qu.:264.0	1st Qu.:14.70	1st Qu.:384.5	1st Qu.:3.32	1st Qu.:41.7	
Median :307.0	Median :17.40	Median :386.9	Median :4.14	Median :48.3	
Mean :325.1	Mean :16.36	Mean :385.2	Mean :4.31	Mean :44.2	
3rd Qu.:307.0	3rd Qu.:17.40	3rd Qu.:389.7	3rd Qu.:5.12	3rd Qu.:50.0	
Max. :666.0	Max. :20.20	Max. :396.9	Max. :7.44	Max. :50.0	

```
> table(rooms8$chas)
```

```
0 1  
11 2
```

Crime seems to be less in the houses which have 8 rooms

```
> summary(rooms8$black)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
354.6	384.5	386.9	385.2	389.7	396.9

11 of the houses with 8 rooms are not near Charles river (only 2 are near Charles river)

```
> summary(Boston$black)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.32	375.38	391.44	356.67	396.23	396.90

All the rooms8 houses blacks population

