```r
#### Exercise 1.1 ####

## Question a: reading csv file
setwd("/Users/tenzintashi/Downloads/CSc 460 - DS/Assignment")

# file <- file.choose()
# college <- read.csv(file)
college <- read.csv('Dataset/College.csv')
```

```r
## Question b: View data set
View(college)
```

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate
Abilene Christian University	2	1660	1232	721	23	52	2885	537	7
Adelphi University	2	2186	1924	512	16	29	2683	1227	12
Adrian College	2	1428	1097	336	22	50	1036	99	11
Agnes Scott College	2	417	349	137	60	89	510	63	12
Alaska Pacific University	2	193	146	55	16	44	249	869	7
Albertson College	2	587	479	158	38	62	678	41	13
Albertus Magnus College	2	353	340	103	17	45	416	230	13
Albion College	2	1899	1720	489	37	68	1594	32	13
Albright College	2	1038	839	227	30	63	973	306	15
Alderson-Broaddus College	2	582	498	172	21	44	799	78	10
Alfred University	2	1732	1425	472	37	75	1830	110	16
Allegheny College	2	2652	1900	484	44	77	1707	44	17
Allentown Coll. of St. Francis de Sales	2	1179	780	290	38	64	1130	638	9
Alma College	2	1267	1080	385	44	73	1306	28	12
Alverno College	2	494	313	157	23	46	1317	1235	8

```r
rownames(college) <- college[, 1]
View(college)
```

	X	Private	Apps	Accept	Enroll	Top10perc	Top25perc
Abilene Christian University	Abilene Christian University	Yes	1660	1232	721	23	52
Adelphi University	Adelphi University	Yes	2186	1924	512	16	29
Adrian College	Adrian College	Yes	1428	1097	336	22	50
Agnes Scott College	Agnes Scott College	Yes	417	349	137	60	89
Alaska Pacific University	Alaska Pacific University	Yes	193	146	55	16	44
Albertson College	Albertson College	Yes	587	479	158	38	62
Albertus Magnus College	Albertus Magnus College	Yes	353	340	103	17	45
Albion College	Albion College	Yes	1899	1720	489	37	68
Albright College	Albright College	Yes	1038	839	227	30	63
Alderson-Broaddus College	Alderson-Broaddus College	Yes	582	498	172	21	44
Alfred University	Alfred University	Yes	1732	1425	472	37	75

```r
# removing the fist column in the data where the name are stored
college <- college[,-1]
View(college)
```

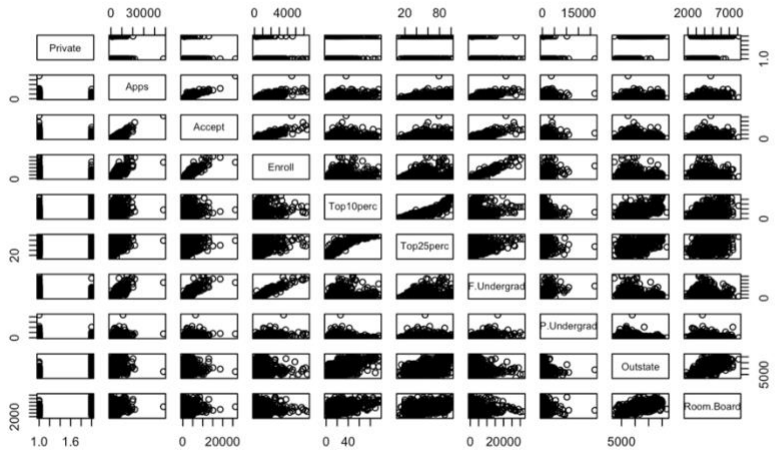
	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate
Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7
Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12
Adrian College	Yes	1428	1097	336	22	50	1036	99	11
Agnes Scott College	Yes	417	349	137	60	89	510	63	12
Alaska Pacific University	Yes	193	146	55	16	44	249	869	7
Albertson College	Yes	587	479	158	38	62	678	41	13
Albertus Magnus College	Yes	353	340	103	17	45	416	230	13
Albion College	Yes	1899	1720	489	37	68	1594	32	13
Albright College	Yes	1038	839	227	30	63	973	306	15
Alderson-Broaddus College	Yes	582	498	172	21	44	799	78	10
Alfred University	Yes	1732	1425	472	37	75	1830	110	16
Allegheny College	Yes	2652	1900	484	44	77	1707	44	17
Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64	1130	638	9
Alma College	Yes	1267	1080	385	44	73	1306	28	12

```
## Question c:
```{r}
# getting the summary
summary(college)
```
```

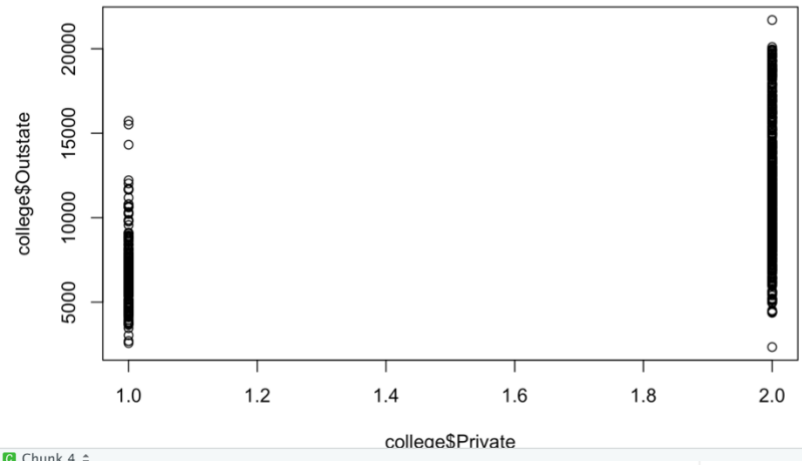
```
  Private             Apps           Accept        Enroll       Top10perc      Top25perc
 Length:777         Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0
 Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0
 Mode  :character   Median : 1558   Median : 1110   Median : 434   Median :23.00   Median : 54.0
                    Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8
                    3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0
                    Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0
  F.Undergrad     P.Undergrad       Outstate      Room.Board       Books           Personal
 Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250
 1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850
 Median : 1707   Median :  353.0   Median : 9990   Median :4200   Median : 500.0   Median :1200
 Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341
 3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700
 Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800
       PhD            Terminal       S.F.Ratio      perc.alumni        Expend        Grad.Rate
 Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
 1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
 Median : 75.00   Median : 82.0   Median :13.60   Median :21.00   Median : 8377   Median : 65.00
 Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
 3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
 Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
```

```
```

```r
# Use pairs() to produce a scatterplot matrix
college[,1] = as.numeric(factor(college[,1]))
pairs(college[, 1:10])
```



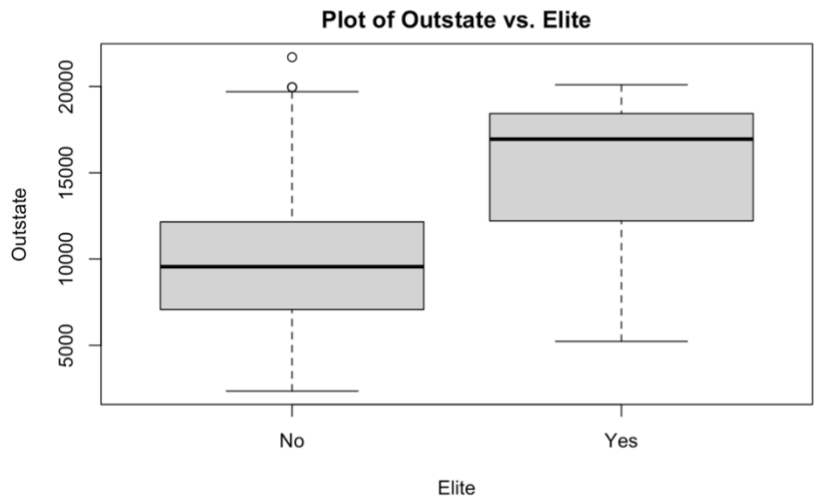

```r
# plot side by side boxplots of Outstate vs Private
plot(college$Private, college$Outstate)
```



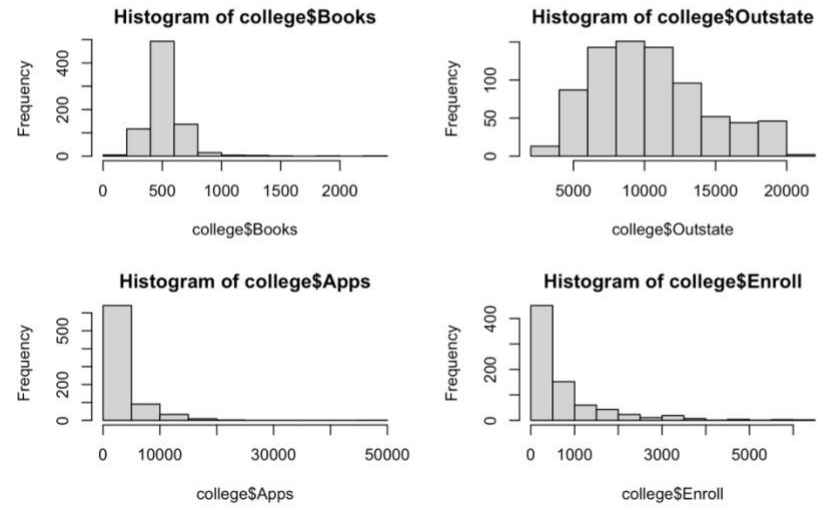

```r
# create a new qualitative variable
Elite <- rep("No", nrow(college))
Elite[college$Top10perc > 50] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college, Elite)

summary(college$Elite)
```

```
 No Yes
699  78
```

```r
# plot side by side boxplots of Outstate vs Elite
plot(college$Elite, college$Outstate, main = "Plot of Outstate vs. Elite", xlab = "Elite", ylab = "Outstate")
```




```r
par(mfrow = c(2,2))
hist(college$Books)
hist(college$Outstate)
hist(college$Apps)
hist(college$Enroll)
```



Exercise 1.2

```{r}
Auto <- read.csv('Dataset/Auto.csv', header = T, na.strings = "?")
Auto <- na.omit(Auto)
```

```{r}
summary(Auto)
```

```
      mpg          cylinders      displacement     horsepower        weight      acceleration
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613   Min.   : 8.00
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225   1st Qu.:13.78
 Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804   Median :15.50
 Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978   Mean   :15.54
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615   3rd Qu.:17.02
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140   Max.   :24.80
      year           origin          name
 Min.   :70.00   Min.   :1.000   Length:392
 1st Qu.:73.00   1st Qu.:1.000   Class :character
 Median :76.00   Median :1.000   Mode  :character
 Mean   :75.98   Mean   :1.577
 3rd Qu.:79.00   3rd Qu.:2.000
 Max.   :82.00   Max.   :3.000
```

```{r}
head(Auto)
```

	m... <dbl>	cylinders <int>	displacement <dbl>	horsepower <int>	weight <int>	acceleration <dbl>	year <int>	origin <int>	name <chr>
1	18	8	307	130	3504	12.0	70	1	chevrolet chevelle malibu
2	15	8	350	165	3693	11.5	70	1	buick skylark 320
3	18	8	318	150	3436	11.0	70	1	plymouth satellite
4	16	8	304	150	3433	12.0	70	1	amc rebel sst
5	17	8	302	140	3449	10.5	70	1	ford torino
6	15	8	429	198	4341	10.0	70	1	ford galaxie 500

6 rows

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

```{r}
# The quantitative variables are cylinders, origin, and name
# and the rest are the qualitative.
qualitative_columns <- c(2, 8, 9)
```

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

```{r}
sapply(Auto[, -qualitative_columns], range)
```

```
      mpg displacement horsepower weight acceleration year
[1,]  9.0           68         46   1613          8.0   70
[2,] 46.6          455        230   5140         24.8   82
```

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

```{r}
sapply(Auto[, -qualitative_columns], mean)
sapply(Auto[, -qualitative_columns], sd)
```

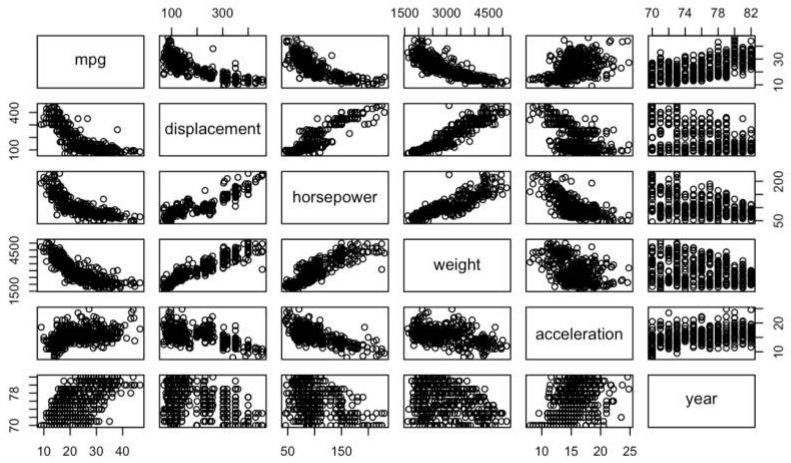
```
         mpg displacement   horsepower       weight acceleration         year
    23.44592    194.41199    104.46939   2977.58418     15.54133     75.97959
         mpg displacement   horsepower       weight acceleration         year
    7.805007   104.644004    38.491160   849.402560     2.758864     3.683737
```

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?

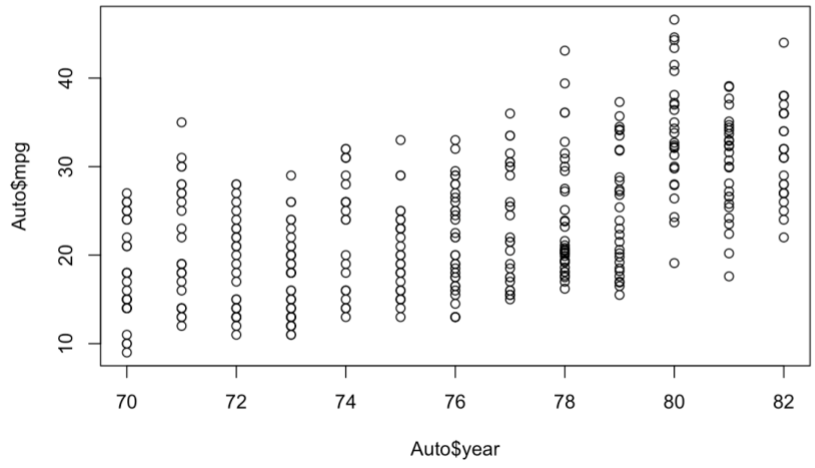
```{r}
sapply(Auto[-seq(10, 85), -qualitative_columns], mean)
```

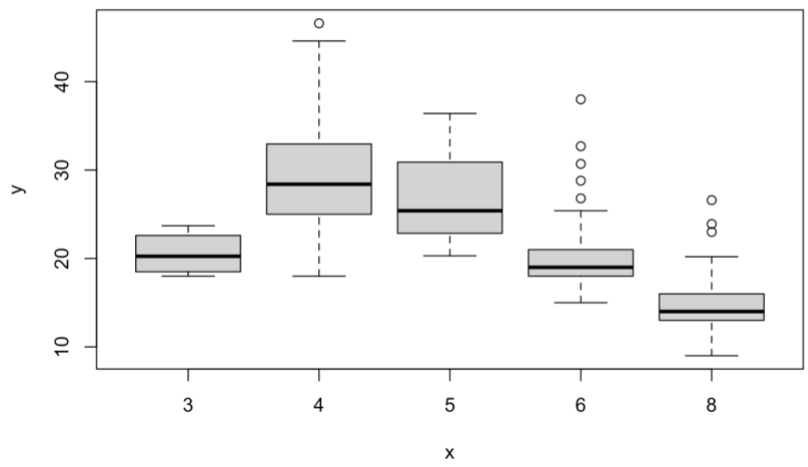
```
         mpg displacement   horsepower       weight acceleration         year
    24.40443    187.24051    100.72152   2935.97152     15.72690     77.14557
```

```
## 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.
```{r}
pairs(Auto[, -qualitative_columns])
```
```




```{r}
plot(Auto$year, Auto$mpg)
```



```
## 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.
```{r}
# Lets plot some mpg vs. some of our qualitative features:
plot(as.factor(Auto$cylinders), Auto$mpg)
```
```




```
```{r}

plot(as.factor(Auto$origin), Auto$mpg)
```
```