Assignment 1

Group 2

Shubhra Rajadhyaksha

Bhaumik Ichhaporia

1.

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. The variables are

- Private : Public/private indicator
- Apps : Number of applications received
- Accept : Number of applicants accepted
- Enroll : Number of new students enrolled
- Top10perc : New students from top 10 % of high school class
- Top25perc : New students from top 25 % of high school class
- F.Undergrad : Number of full-time undergraduates
- P.Undergrad : Number of part-time undergraduates
- Outstate : Out-of-state tuition
- Room.Board : Room and board costs
- Books : Estimated book costs
- Personal : Estimated personal spending
- PhD : Percent of faculty with Ph.D.'s
- Terminal : Percent of faculty with terminal degree
- S.F.Ratio : Student/faculty ratio
- perc.alumni : Percent of alumni who donate
- Expend : Instructional expenditure per student
- Grad.Rate : Graduation rate

Before reading the data into R, it can be viewed in Excel or a text editor.

(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.

(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:

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 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.

```
> 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.

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

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%.

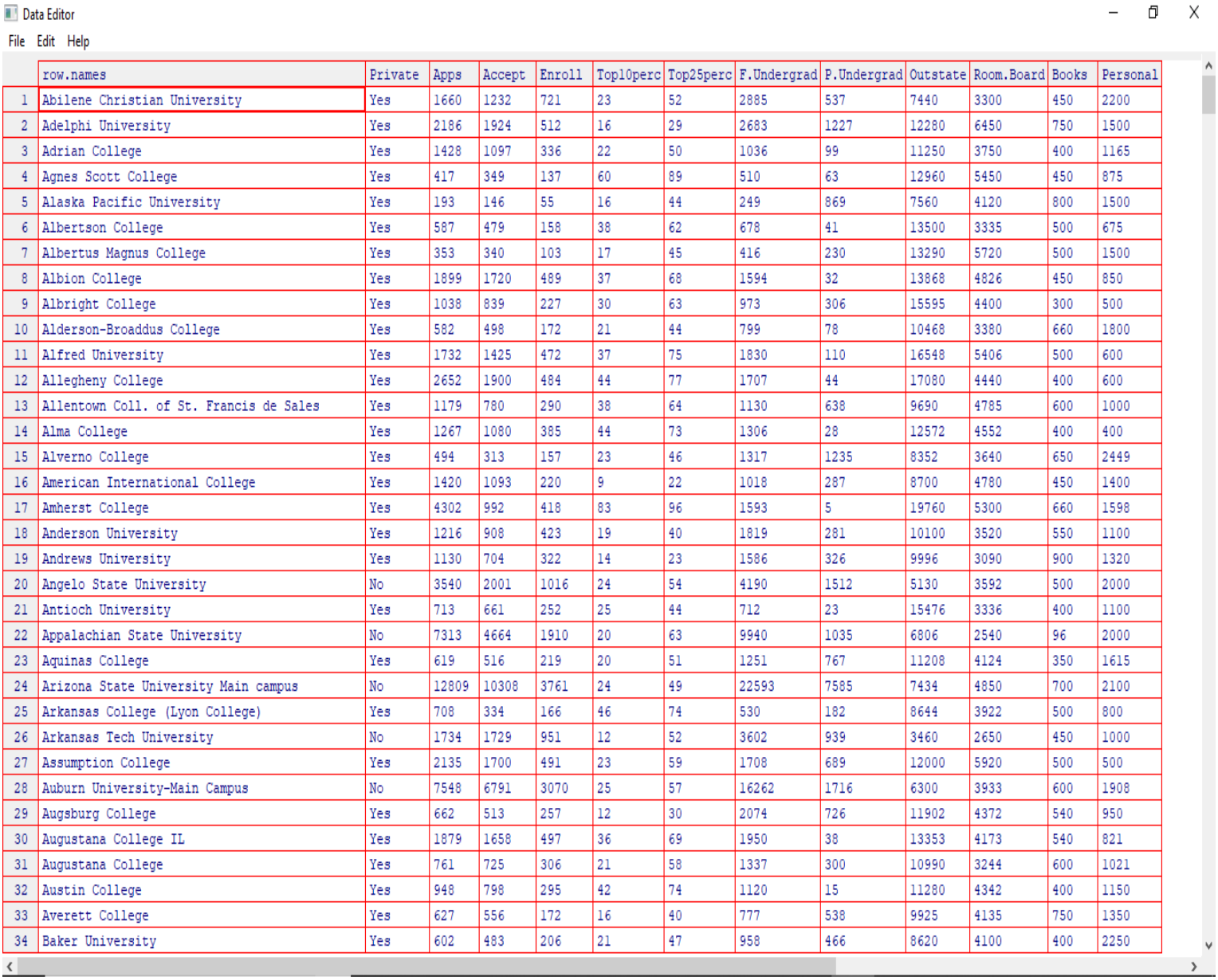
```
> Elite=rep("No",nrow(college))
> Elite[college$Top10perc >50]="Yes"
> Elite=as.factor(Elite)
> college=data.frame(college ,Elite)
```

Show all your R code and output. For part(c)
provide a brief summary of your answers to:
What is the university with the most students in the top 10% of class?
What university has the smallest acceptance rate?

8)a) 1. Reading the csv file in the variable College and calling the loaded data.

```
> college<- read.csv(file="hw1/College.csv", header=TRUE, sep=",")
>
> str(college)
'data.frame':   777 obs. of  19 variables:
 $ X          : Factor w/ 777 levels "Abilene Christian University",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ Apps       : int  1660 2186 1428 417 193 587 353 1899 1038 582 ...
 $ Accept     : int  1232 1924 1097 349 146 479 340 1720 839 498 ...
 $ Enroll     : int  721 512 336 137 55 158 103 489 227 172 ...
 $ Top10perc  : int  23 16 22 60 16 38 17 37 30 21 ...
 $ Top25perc  : int  52 29 50 89 44 62 45 68 63 44 ...
 $ F.Undergrad: int  2885 2683 1036 510 249 678 416 1594 973 799 ...
 $ P.Undergrad: int  537 1227 99 63 869 41 230 32 306 78 ...
 $ Outstate   : int  7440 12280 11250 12960 7560 13500 13290 13868 15595 10468 ...
 $ Room.Board : int  3300 6450 3750 5450 4120 3335 5720 4826 4400 3380 ...
 $ Books      : int  450 750 400 450 800 500 500 450 300 660 ...
 $ Personal   : int  2200 1500 1165 875 1500 675 1500 850 500 1800 ...
 $ PhD        : int  70 29 53 92 76 67 90 89 79 40 ...
 $ Terminal   : int  78 30 66 97 72 73 93 100 84 41 ...
 $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
 $ perc.alumni: int  12 16 30 37 2 11 26 37 23 15 ...
 $ Expend     : int  7041 10527 8735 19016 10922 9727 8861 11487 11644 8991 ...
 $ Grad.Rate  : int  60 56 54 59 15 55 63 73 80 52 ...
```

b) Fix function to see the data like excel sheet.
->fix(College)


Data Editor

File Edit Help

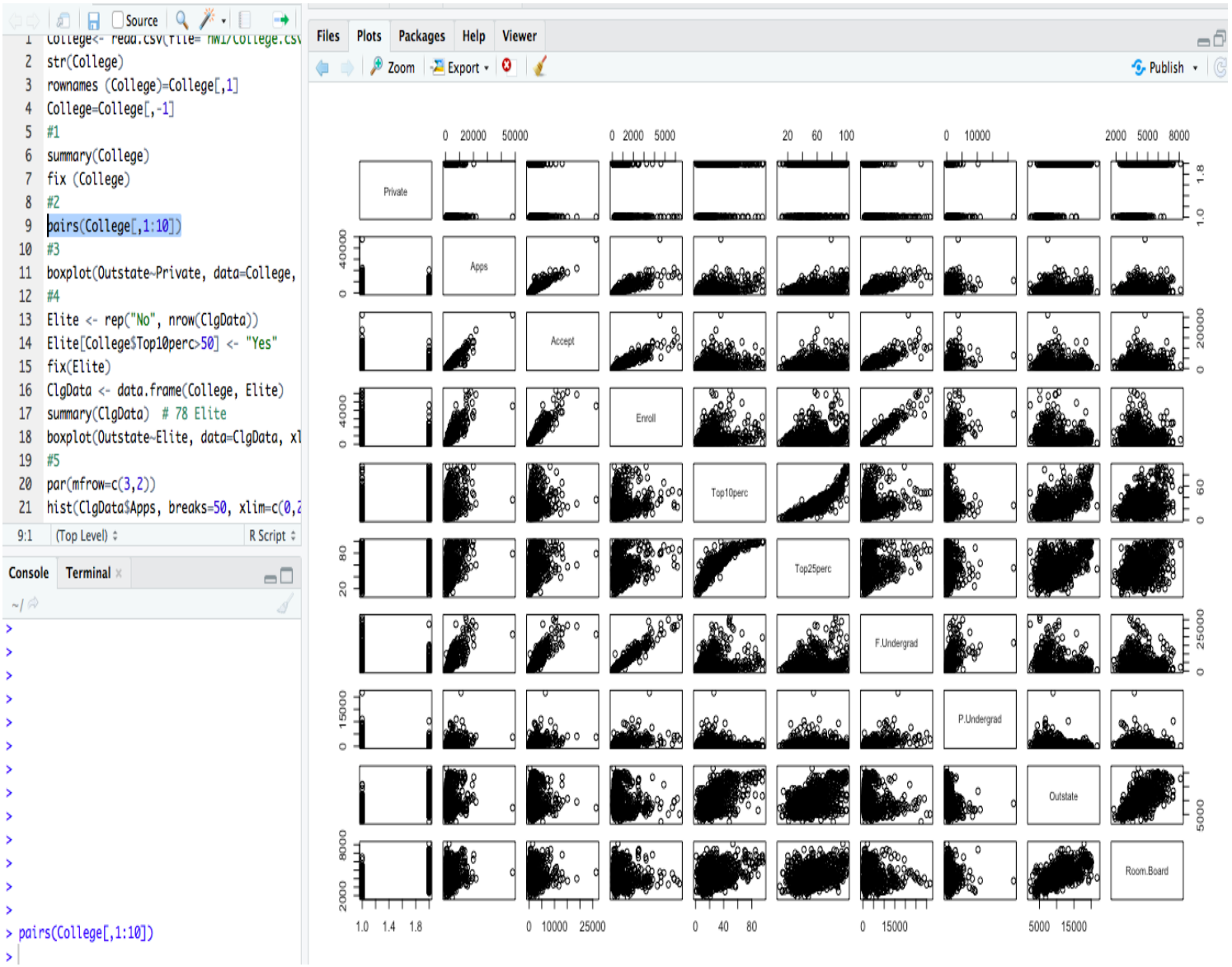
	row.names	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal
1	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200
2	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500
3	Adrian College	Yes	1428	1097	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-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
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
33	Averett College	Yes	627	556	172	16	40	777	538	9925	4135	750	1350
34	Baker University	Yes	602	483	206	21	47	958	466	8620	4100	400	2250

C) i.summary(College)

Summary function to produce a numerical summary of the variables in the data set.

```
> college<- read.csv(file="hw1/College.csv", header=TRUE, sep=",")
> summary(college)
                          X       Private        Apps           Accept         Enroll       Top10perc      Top25perc
 Abilene Christian University:  1   No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0
 Adelphi University          :  1   Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0
 Adrian College              :  1             Median : 1558   Median : 1110   Median : 434   Median :23.00   Median : 54.0
 Agnes Scott College         :  1             Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8
 Alaska Pacific University   :  1             3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0
 Albertson College           :  1             Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0
 (Other)                     :771
  F.Undergrad     P.Undergrad         Outstate       Room.Board      Books           Personal         PhD            Terminal
 Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   :  250   Min.   :  8.00   Min.   : 24.0
 1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.:  850   1st Qu.: 62.00   1st Qu.: 71.0
 Median : 1707   Median :  353.0   Median : 9990   Median :4200   Median : 500.0   Median : 1200   Median : 75.00   Median : 82.0
 Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   : 1341   Mean   : 72.66   Mean   : 79.7
 3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.: 1700   3rd Qu.: 85.00   3rd Qu.: 92.0
 Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   : 6800   Max.   :103.00   Max.   :100.0

   S.F.Ratio      perc.alumni        Expend        Grad.Rate
 Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
 1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
 Median :13.60   Median :21.00   Median : 8377   Median : 65.00
 Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
 3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
 Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
```

C.ii) pairs() function to produce a numerical summary of the variables in data set.



3.iii) Plot function to produce side by side box plots.

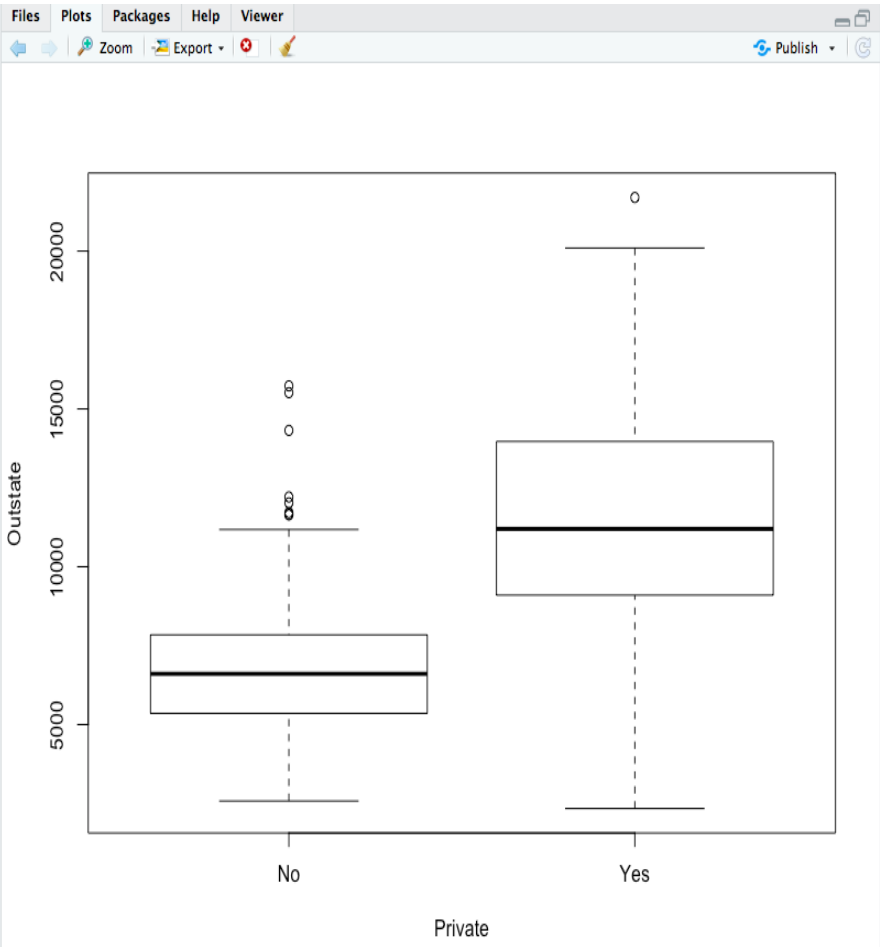
boxplot(Outstate~Elite, data=ClgData, xlab="Elite", ylab="Outstate")

```
College<- read.csv("file= hw1/college.csv", header=TRUE, sep=",")
str(College)
rownames (College)=College[,1]
College=College[,-1]
#1
summary(College)
fix (College)
#2
pairs(College[,1:10])
#3
boxplot(Outstate~Private, data=College, xlab="Private", ylab="Outstate")
#4
Elite <- rep("No", nrow(ClgData))
Elite[College$Top10perc>50] <- "Yes"
fix(Elite)
ClgData <- data.frame(College, Elite)
summary(ClgData)  # 78 Elite
boxplot(Outstate~Elite, data=ClgData, xlab="Elite", ylab="Outstate")
#5
par(mfrow=c(3,2))
hist(ClgData$Apps, breaks=50, xlim=c(0,25000), main="Apps")
```

11:1 (Top Level) R Script

Console Terminal

```
> pairs(College[,1:10])
> boxplot(Outstate~Private, data=College, xlab="Private", ylab="Outstate")
>
```



c) iv. 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"
fix(Elite)
College <- data.frame(College, Elite)
summary(College)  # 78 Elite
boxplot(Outstate~Elite, data=College, xlab="Elite", ylab="Outstate")
```

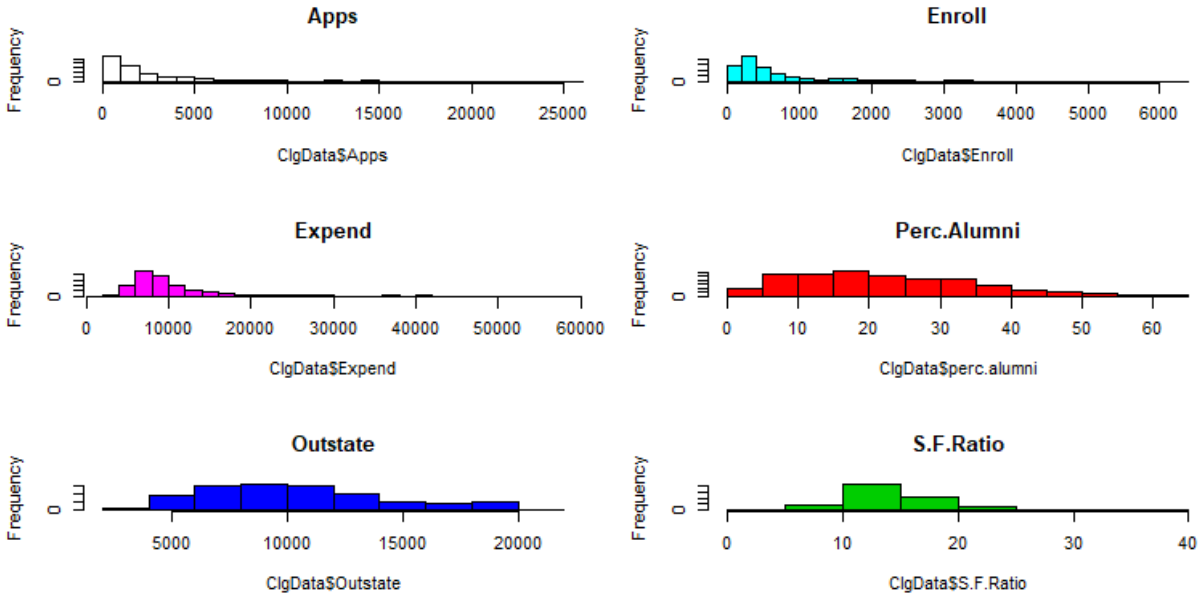
summary(College)

```
> summary(ClgData)  # 78 Elite
                              X       Private        Apps           Accept
 Abilene Christian University:  1   No :212   Min.   :   81   Min.   :   72
 Adelphi University           :  1   Yes:565   1st Qu.:  776   1st Qu.:  604
 Adrian College               :  1             Median : 1558   Median : 1110
 Agnes Scott College          :  1             Mean   : 3002   Mean   : 2019
 Alaska Pacific University    :  1             3rd Qu.: 3624   3rd Qu.: 2424
 Albertson College            :  1             Max.   :48094   Max.   :26330
 (Other)                      :771
     Enroll       Top10perc       Top25perc      F.Undergrad     P.Undergrad
 Min.   :  35   Min.   : 1.00   Min.   :  9.0   Min.   :  139   Min.   :    1.0
 1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0
 Median : 434   Median :23.00   Median : 54.0   Median : 1707   Median :  353.0
 Mean   : 780   Mean   :27.56   Mean   : 55.8   Mean   : 3700   Mean   :  855.3
 3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0
 Max.   :6392   Max.   :96.00   Max.   :100.0   Max.   :31643   Max.   :21836.0

    Outstate       Room.Board       Books           Personal         PhD
 Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
 1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
 Median : 9990   Median :4200   Median : 500.0   Median :1200   Median : 75.00
 Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
 3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
 Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00

    Terminal       S.F.Ratio      perc.alumni        Expend       Grad.Rate
 Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
 1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
 Median : 82.0   Median :13.60   Median :21.00   Median : 8377   Median : 65.00
 Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
 3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
 Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
```

Boxplot()

c)v. Use hist() function to plot various histograms.

```
#5
par(mfrow=c(3,2))
hist(College$Apps, breaks=50, xlim=c(0,25000), main="Apps")
hist(College$Enroll,col=5, breaks=25, main="Enroll")
hist(College$Expend, col=6,breaks=25, main="Expend")
hist(College$perc.alumni,col=2,main="Perc.Alumni")
hist(College$Outstate,col=4, main="Outstate")
hist(College$S.F.Ratio,col=3,main="S.F.Ratio")
```



c) vi. provide a brief summary of your answers to

-What is the university with the most students in the top 10% of class?

-What university has the smallest acceptance rate?

```
x=College[which.max(College$Top10perc),]
print(x[,1])

y=College[which.min(College$Accept/College$Apps),]
print(y[,1])
```

```
> y=College[which.min(College$Accept/College$Apps),]
> fix(y)
> print(y[,1])
[1] Princeton University
777 Levels: Abilene Christian University Adelphi University ... York College of Pennsylvani
a
> x=College[which.max(College$Top10perc),]
> print(x[,1])
[1] Massachusetts Institute of Technology
777 Levels: Abilene Christian University Adelphi University ... York College of Pennsylvani
a
```

2. The dataset “housetype.data” represents an extract from a commercial marketing database created from questionnaires filled out by shopping mall customers in the San Francisco Bay area. Report the commands you use to:
A. Read the data into a data frame. Be sure to keep the row and column names around. Show the dimensions of the data matrix and the upper 5x5 submatrix.

#Read the data into a data frame. Be sure to keep the row and column names around.
housetype_data <-read.table("hw1/housetype_data.txt",header=TRUE,sep = ",")

#Show the dimensions of the data matrix
dim(housetype_data)

```
> dim(housetype_data)
[1] 9013   14
```

Show the upper 5x5 submatrix
housetype_data[1:5,1:5]

```
> housetype_data[1:5,1:5]
  ht sex ms age edu
1  1   2  4   7   4
2  1   2  1   5   4
3  1   1  1   5   5
4  3   2  1   3   5
5  1   2  5   1   2
```

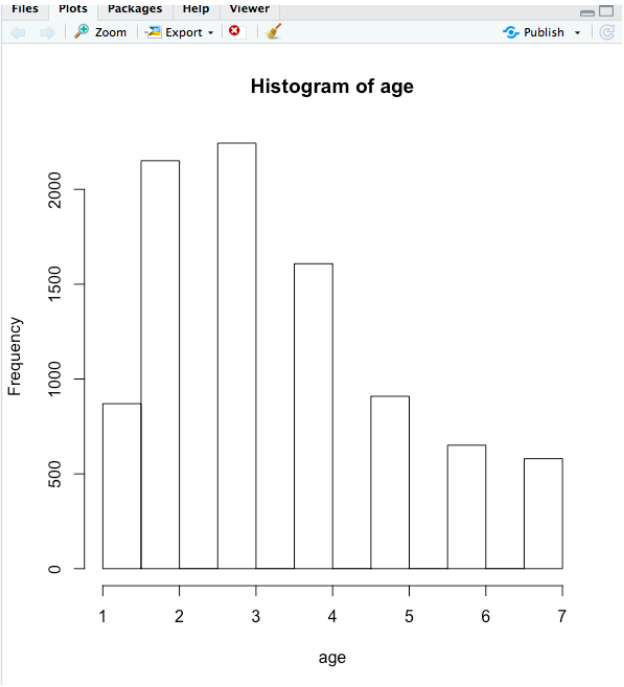
B. Write a function attributeHist that takes the name of an attribute, such as “age", finds the corresponding column in the table, and produces the histogram. By default, it should put the long attribute name into the title and on the horizontal axis. If the attribute contains “missing values” (represented as NA in the data), a message with the missing count should be printed. Specifically,
a. attributeHist(“age") should produce a histogram of the values for this attribute.
b. attributeHist(“hello") should print a message pointing out that here is no such attribute.
c. attributeHist(“eth”) should display the histogram and print a message “61 missing values.” Turn in the source code for your function as well as the output on the three calls above.

```
attributeHist <- function(attribute_name) {
  housetype_data <-read.table("hw1/housetype_data.txt",header=TRUE,sep = ",")

  if (attribute_name %in% colnames(housetype_data)){

  column<-housetype_data[attribute_name]

  missing<-length(column[is.na(column)])

  missing
```

```
  hist(column[!is.na(column)],freq = TRUE,main =paste("Histogram of" , attribute_name),
       xlab = toString(attribute_name))

  if (missing > 0) {
    print(paste(missing," missing values"))
  }
  }
  else {
  msg <- paste(attribute_name , " : No such attribute exists")
  print(msg)
}

}
```

a.

```
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> attributeHist <- function(attribute_name) {
+     housetype_data <-read.table("hw1/housetype_data.txt",header=TRUE,sep = ","
)
+
+     if (attribute_name %in% colnames(housetype_data)){
+
+         column<-housetype_data[attribute_name]
+
+         missing<-length(column[is.na(column)])
+
+         missing
+
+         hist(column[!is.na(column)],freq = TRUE,main =paste("Histogram of" , a
ttribute_name),
+              xlab = toString(attribute_name))
+
+         if (missing > 0) {
+             print(paste(missing," missing values"))
+         }
+     }
+     else {
+         msg <- paste(attribute_name , " : No such attribute exists")
+         print(msg)
+     }
+
+ }
>
> attributeHist("age")
>
```

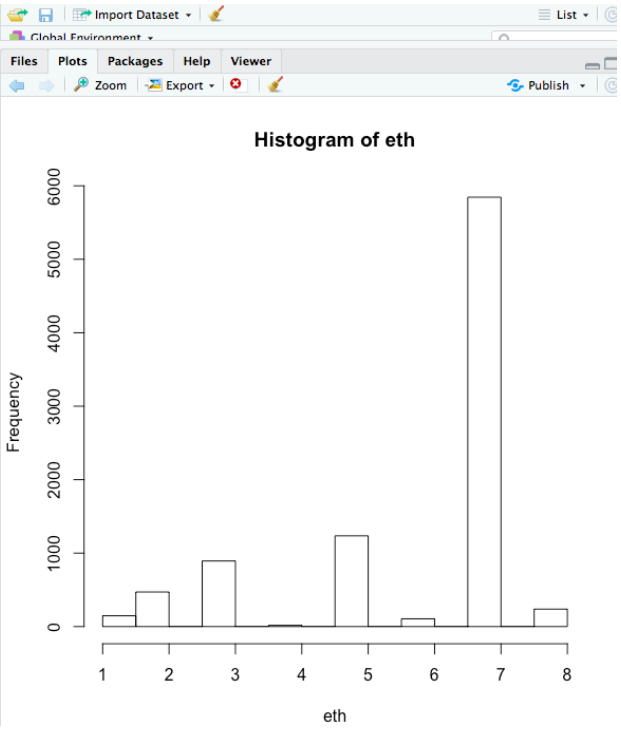


b.

```
> attributeHist("hello")
[1] "hello  : No such attribute exists"
```

c.

```
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> attributeHist <- function(attribute_name) {
+     housetype_data <-read.table("hw1/housetype_data.txt",header=TRUE,sep = ","
)
+
+     if (attribute_name %in% colnames(housetype_data)){
+
+         column<-housetype_data[attribute_name]
+
+         missing<-length(column[is.na(column)])
+
+         missing
+
+         hist(column[!is.na(column)],freq = TRUE,main =paste("Histogram of" , a
ttribute_name),
+              xlab = toString(attribute_name))
+
+         if (missing > 0) {
+             print(paste(missing," missing values"))
+         }
+     }
+     else {
+         msg <- paste(attribute_name , " : No such attribute exists")
+         print(msg)
+     }
+
+ }
>
> attributeHist("age")
> attributeHist("hello")
[1] "hello  : No such attribute exists"
> attributeHist("eth")
[1] "61  missing values"
```



3. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small.

Flexible statistical learning method will work better than inflexible method. The number of observations in the training sample is quite large. As there is large amount of data, there are lower chances of overfitting.

(b) The number of predictors p is extremely large, and the number

of observations n is small.
Flexible method will be worse as compared to inflexible method. As the number of observations is too small, flexible method might lead to overfitting of data.

(c) The relationship between the predictors and response is highly Non-linear.
Flexible method will work better in this case , as it is non linear.

(d) The variance of the error terms, i.e. σ2 = Var(), is extremely High.
Flexible method will be worse as the flexible model will also fit the error/noise.

4. Compound Bayesian Decision Theory. Suppose we have three categories with P(w1) = 1/2, P(w2)=P(w3)= ¼ and the following distributions

- $p(x|\omega_1) \sim N(0, 1)$
- $p(x|\omega_2) \sim N(0.5, 1)$
- $p(x|\omega_3) \sim N(1, 1)$

Using the following, which assumes the independence of x_i and $\omega(i)$:

$$p(\mathbf{X}|\boldsymbol{\omega}) = \prod_{i=1}^{4} p(x_i|\omega(i)) \quad ; \quad P(\boldsymbol{\omega}) = \prod_{i=1}^{4} P(\omega(i))$$

and using the *dnorm*() R function, calculate explicitly the probability that the sequence $\mathbf{X} = <0.6, 0.1, 0.9, 1.1>$ came from $\boldsymbol{\omega} = <\omega_1, \omega_3, \omega_3, \omega_2>$.

Note:
P(w1)=0.5
P(w2)=0.25
P(w3)=0.25
ω=<ω1,ω3,ω3,ω2>
Prior= P(w)= P(w1)*P(w3)*P(w3)*P(w2)

X <- c(0.6,0.1,0.9,1.1)
PW <- c(0.5,0.25,0.25,0.25)
mean <- c(0,0.5,1)
Sd <- c(1,1,1)

Prior probability of w is : P(w)= P(w1)*P(w3)*P(w3)*P(w2)

Note for self : The function dnorm returns the value of the probability density function for the normal distribution given parameters for x, μ, and σ.

Likelihood= P(X1|w1)*P(X2|w3)*P(X3|w3)*P(X4|w2)
All terms in likelihood function calculated using dnorm() function in R

Posterior probability= Likelihood*Prior / estimate
Estimate p(x)=

$$p(\mathbf{x}) = \sum_j p(\mathbf{x}|\omega_j)P(\omega_j)$$

R code :

```
# Data from problem
X <- c(0.6,0.1,0.9,1.1)
# Priors  for w1, w3,w3,w2
PW <- c(0.5,0.25,0.25,0.25)
# Mean for class w1, w2, w3
mean <- c(0,0.5,1)
# Standard deviations
Sd <- c(1,1,1)

# calculation of dnorm P(X|W)
prob_dens<-function(x,w) {
ret_val <- dnorm(x,mean[w],Sd[w],log=FALSE)
}

 # to find the estimate p(x): balancing term
px=0
for(i in 1:3)
{
   for(j in 1:3)
  {
      for(k in 1:3)
     {
         for(l in 1:3)
         {
            px <-  px + (prob_dens(X[1],i) * PW[i]* prob_dens(X[2],j) * PW[j] *prob_dens(X[3],k) *
PW[k] *prob_dens(X[4],l) *PW[l])

         }
      }
   }
}
# prior= multiplication of individual priors
# Prior probability of w is : P(w)= P(w1)*P(w3)*P(w3)*P(w2)
prior= prod(PW)

# likelihood calculation
likelihood<-prob_dens(X[1],1)*prob_dens(X[2],3)*prob_dens(X[3],3)*prob_dens(X[4],2)

# Using Bayes theorem
# Posterior= (prior*likelihood)/px
Numerator<-(likelihood * prior)
Posterior<-Numerator/px
```

```
> # Data from problem
> X <- c(0.6,0.1,0.9,1.1)
> # Priors  for w1, w3,w3,w2
> PW <- c(0.5,0.25,0.25,0.25)
> # Mean for class w1, w2, w3
> mean <- c(0,0.5,1)
> # Standard deviations
> Sd <- c(1,1,1)
>
> # calculation of dnorm P(X|W)
> prob_dens<-function(x,w) {
+     ret_val <- dnorm(x,mean[w],Sd[w],log=FALSE)
+ }
>
> # to find the estimate p(x): balancing term
> px=0
> for(i in 1:3)
+ {
+     for(j in 1:3)
+     {
+         for(k in 1:3)
+         {
+             for(l in 1:3)
+             {
+                 px <-  px + (prob_dens(X[1],i) * PW[i]* prob_dens(X[2],j) * PW[j] *prob_dens(X[3],k) *         PW[k] *prob_dens(X[4],l) *PW[l])
+
+             }
+         }
+     }
+ }
> # prior= multiplication of individual priors
> # Prior probability of w is : P(w)= P(w1)*P(w3)*P(w3)*P(w2)
> prior= prod(PW)
>
> # likelihood calculation
> likelihood<-prob_dens(X[1],1)*prob_dens(X[2],3)*prob_dens(X[3],3)*prob_dens(X[4],2)
>
> # Using Bayes theorem
> # Posterior= (prior*likelihood)/px
> Numerator<-(likelihood * prior)
> Posterior<-Numerator/px
>
> Posterior
[1] 0.007583795
```

So the probability is found to be 0.007583795

5. Discriminant Functions and Maximum-Likelihood (ML) Estimation. Consider the following data sets: D1 = {<3, 4>, <4, 6>, <2, 6>, <3, 8>} D2 = {<3, 0>, <1, -2>, <5, -2>, <3,

-4>} a. Using ML estimates for μ1, μ2 and S1, S2, write expressions for the discriminant functions g1(x) and g2(x) b. Assuming equal priors, find expression for decision boundary by setting g1(x) = g2(x) c. Draw, by hand or otherwise, decision boundary, means, and data points.

D1 = {<3, 4>, <4, 6>, <2, 6>, <3, 8>}
D2 = {<3, 0>, <1, -2>, <5, -2>, <3, -4>}

```
> # ques no 5
> #mean 1
> u1=matrix(c(3,6),2,1)
> #mean2
> u2=matrix(c(3,-2),2,1)
> u1
     [,1]
[1,]    3
[2,]    6
> u2
     [,1]
[1,]    3
[2,]   -2
> #Sigma1
> e1=matrix(c(0.5,0,0,2),2,2)
> e1
     [,1] [,2]
[1,]  0.5    0
[2,]  0.0    2
> e2=matrix(c(2,0,0,2),2,2)
> e2
     [,1] [,2]
[1,]    2    0
[2,]    0    2
> #e2 is sigma2
> #e1i is inverse of e1
> #e2i is inverse of e2
> e1i=solve(e1)
> e2i=solve(e2)
```

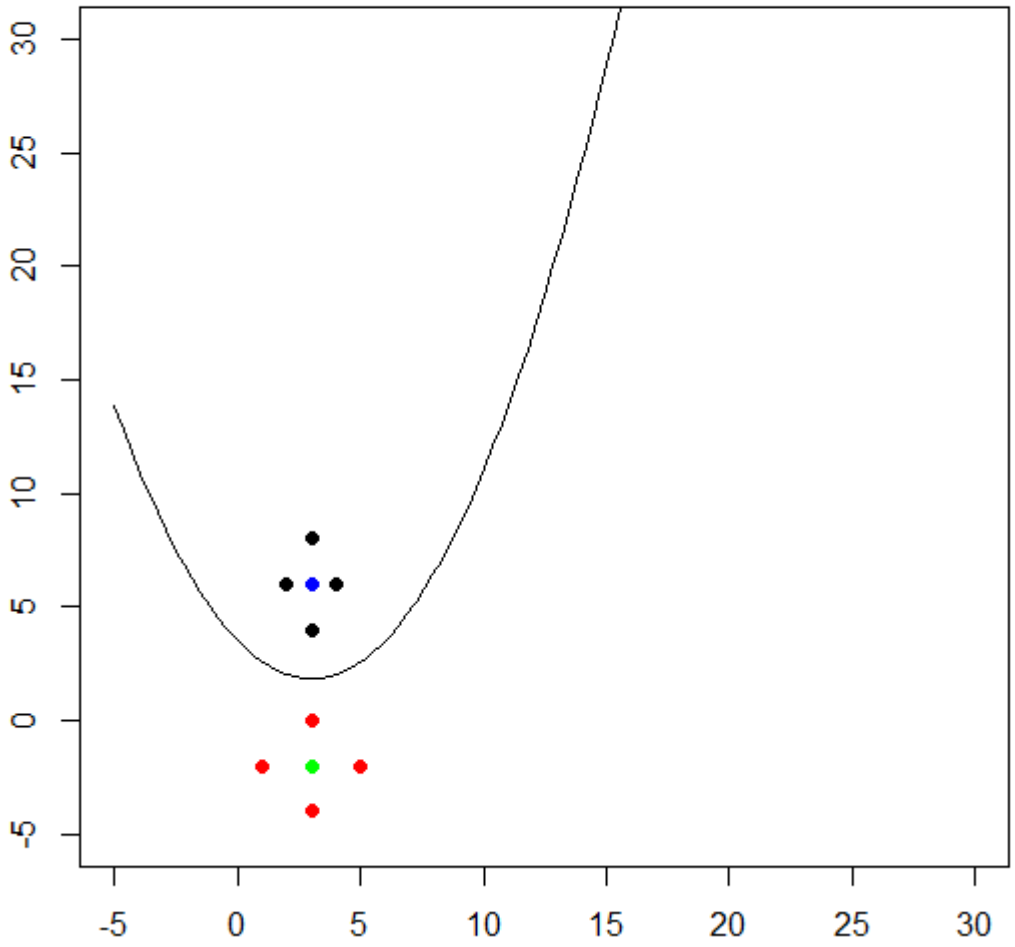
```
> #Find W1
> W1=-0.5*e1i
> W1
     [,1]  [,2]
[1,]   -1  0.00
[2,]    0 -0.25
> w1=e1i%*%u1
> w1
     [,1]
[1,]    6
[2,]    3
> #To find w10
> #u1 transpose
> #u1t is u1 transpose
> u1t=t(u1)
> u1t
     [,1] [,2]
[1,]    3    6
> #Intermediate steps in calculation of w10
> a=u1t%*%e1i
> b=a%*%u1
> first_term=-0.5*b
> first_term
     [,1]
[1,]  -18
> #determinant of e1
> de1=det(e1)
> de1
[1] 1
> #These values are substituted in equation for w10
```

```
> # Find Transpose of u2 as u2t
> u2t=t(u2)
> u2t
     [,1] [,2]
[1,]    3   -2
> #Find inverse of e2 as e2i
> e2i=solve(e2)
> e2i
     [,1] [,2]
[1,]  0.5  0.0
[2,]  0.0  0.5
> #To find W2
> W2=-0.5*e2i
> W2
      [,1]  [,2]
[1,] -0.25  0.00
[2,]  0.00 -0.25
> w2=e2i%*%u2
> w2
     [,1]
[1,]  1.5
[2,] -1.0
> #To find w20
> #Intermediate steps
> a=u2t%*%e2i
> b=a%*%u2
> f_t=-0.5*b
> f_t
      [,1]
[1,] -3.25
> # Find determinant of e2
> de2=det(e2)
> de2
[1] 4
> #These values are substituted in equation for w20
```

Solved on handwritten sheet. R used in intermediate steps and for drawing graphs.

R code :

```
->f <- function(x) 0.1875 * x^2 - 1.125 * x +3.514
->plot(f, xlim=c(-5,30),ylim=c(-5,30), lty=1)
->x= c(3,4,2,3)
->y=c(4,6,6,8)
->points(x,y,pch=19)
->points(3,6,pch=19,col="blue") #mean 1
->points(3,-2,pch=19,col="green")#mean
->x1=c(3,1,5,3)
->y1=c(0,-2,-2,-4)
->points(x1,y1,pch=19,col="red")
```



6. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs	X1	X2	X3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.

Euclidean distance=distance((x,y,z),(a,b,c))

$= \text{sqrt}((x-a)^2 + (y-b)^2 + (z-c)^2)$

Point	Calculation	Euclidean distance
1	sqrt(0+9+0)	3
2	sqrt(4+0+0)	2
3	sqrt(0+1+9)	3.162
4	sqrt(0+1+4)	2.236
5	sqrt(1+0+1)	1.414
6	sqrt(1+1+1)	1.732

(b) What is our prediction with K = 1? Why?
With K=1 :
Closest point is point 5, which is green
So our prediction is Green

(c) What is our prediction with K = 3? Why?
With K=3 :
3 closest points are :
Point 5: Green
Point 6: Red
Point 2: Red

As most of the points are red, our prediction is Red.

(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why?
If we increase the value of K, the decision boundary becomes linear. So if the Bayes decision boundary is highly non linear, we would expect the value of K to be small.