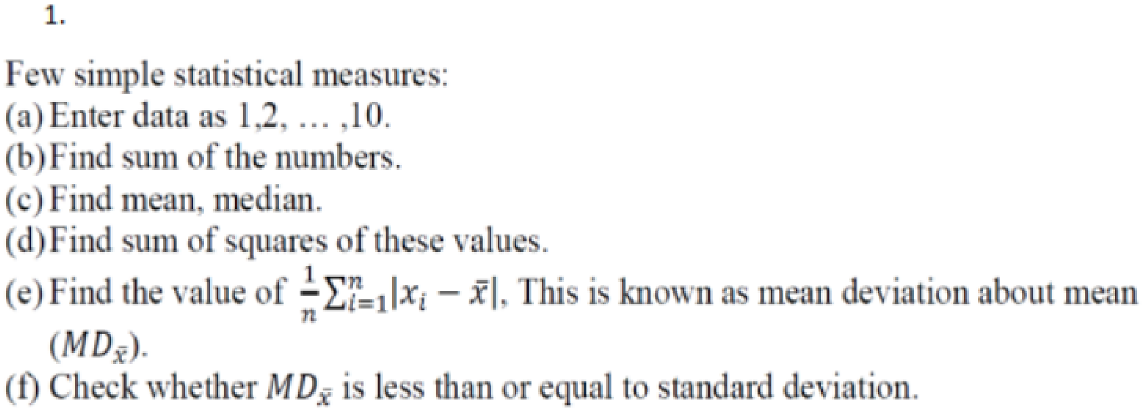
21BCE7371
RADHA KRISHNA GARG

APPLIED STATS

FINAL LAB ASSIGNMENT



1.

Few simple statistical measures:

(a) Enter data as 1,2, … ,10.

(b) Find sum of the numbers.

(c) Find mean, median.

(d) Find sum of squares of these values.

(e) Find the value of $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$, This is known as mean deviation about mean ($MD_{\bar{x}}$).

(f) Check whether $MD_{\bar{x}}$ is less than or equal to standard deviation.

(a)

```
> a= c(1:10)
 [1]  1  2  3  4  5  6  7  8  9 10
```

(b)

```
> sum(a)
[1] 55
```

(c)

```
> mean(a)
[1] 5.5
> median(a)
[1] 5.5
```

(d)

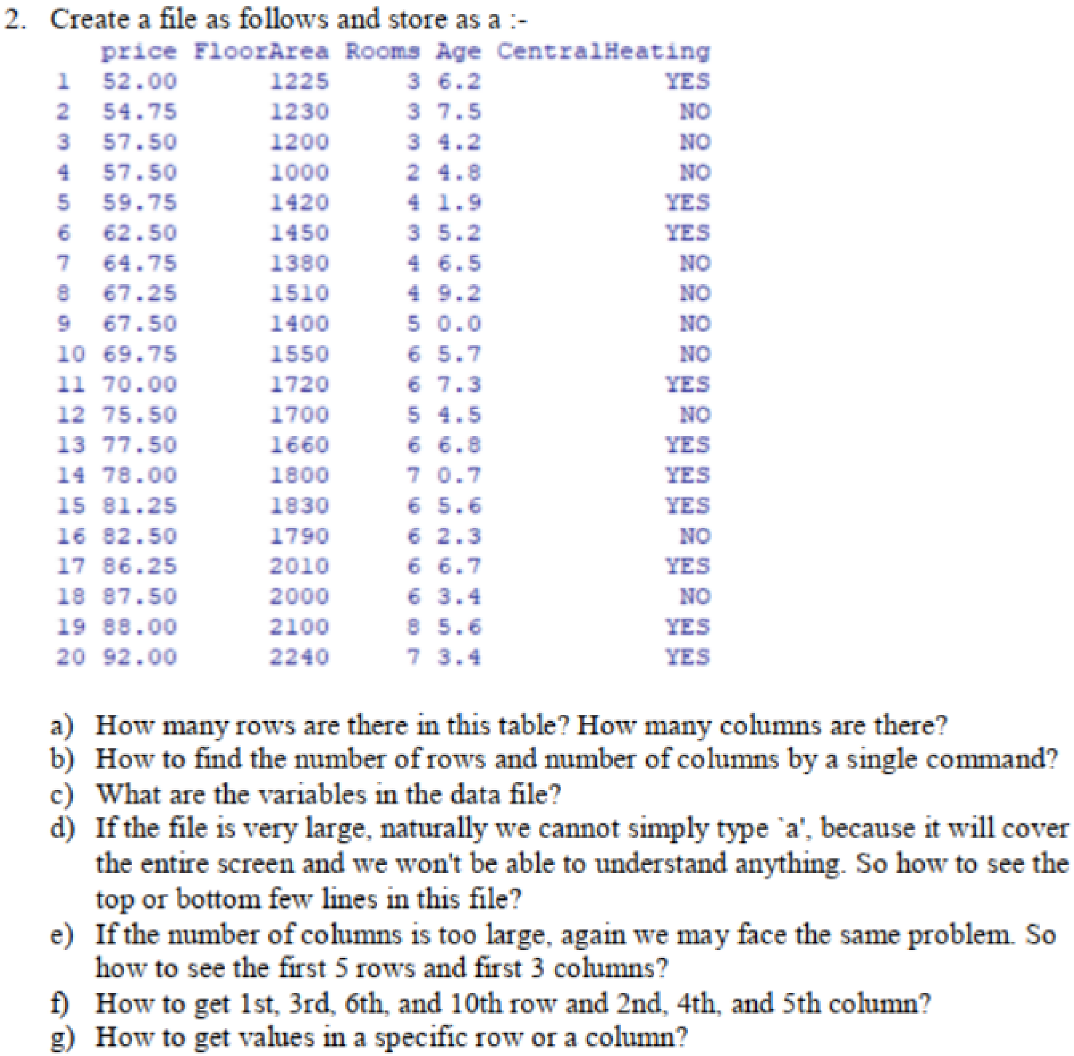
```
> sum(a^2)
[1] 385
```

(e)

```
> mean(-1*(a-mean(a)))
[1] 0
> sd(a)
[1] 3.02765
```

(f)

```
> mean(abs(a - mean(a))) <= sd(a)
[1] TRUE
```



2. Create a file as follows and store as a :-

```
   price FloorArea Rooms Age CentralHeating
1  52.00      1225     3 6.2            YES
2  54.75      1230     3 7.5             NO
3  57.50      1200     3 4.2             NO
4  57.50      1000     2 4.8             NO
5  59.75      1420     4 1.9            YES
6  62.50      1450     3 5.2            YES
7  64.75      1380     4 6.5             NO
8  67.25      1510     4 9.2             NO
9  67.50      1400     5 0.0             NO
10 69.75      1550     6 5.7             NO
11 70.00      1720     6 7.3            YES
12 75.50      1700     5 4.5             NO
13 77.50      1660     6 6.8            YES
14 78.00      1800     7 0.7            YES
15 81.25      1830     6 5.6            YES
16 82.50      1790     6 2.3             NO
17 86.25      2010     6 6.7            YES
18 87.50      2000     6 3.4             NO
19 88.00      2100     8 5.6            YES
20 92.00      2240     7 3.4            YES
```

a) How many rows are there in this table? How many columns are there?
b) How to find the number of rows and number of columns by a single command?
c) What are the variables in the data file?
d) If the file is very large, naturally we cannot simply type 'a', because it will cover the entire screen and we won't be able to understand anything. So how to see the top or bottom few lines in this file?
e) If the number of columns is too large, again we may face the same problem. So how to see the first 5 rows and first 3 columns?
f) How to get 1st, 3rd, 6th, and 10th row and 2nd, 4th, and 5th column?
g) How to get values in a specific row or a column?

```
> price=
c(52.00,54.75,57.50,57.50,59.75,62.50,64.75,67.25,67.50,69.75,70.00,75.50,77.50,78.00,81.25,82.50,86.25,87.50,88.00,92.00)
> FloorArea=
c(1125,1230,1200,1000,1420,1450,1380,1510,1400,1550,1720,1700,1660,1800,1830,1790,2010,2000,2100,2240)
> Rooms= c(3,3,3,2,4,3,4,4,5,6,6,5,6,7,6,6,6,6,8,7)
> Age= c(6.2,7.5,4.2,4.8,1.9,5.2,6.5,9.2,0.0,5.7,7.3,4.5,6.8,0.7,5.6,2.3,6.7,3.4,5.6,3.4)
> CentralHeating=
c('YES','NO','NO','NO','YES','YES','NO','NO','NO','NO','YES','NO','YES','YES','YES','NO','YES
','NO','YES','YES')
> data1= data.frame(price,FloorArea,Rooms,Age,CentralHeating)
> data1
   price FloorArea Rooms Age CentralHeating
1  52.00      1125     3 6.2            YES
2  54.75      1230     3 7.5             NO
3  57.50      1200     3 4.2             NO
4  57.50      1000     2 4.8             NO
5  59.75      1420     4 1.9            YES
6  62.50      1450     3 5.2            YES
7  64.75      1380     4 6.5             NO
8  67.25      1510     4 9.2             NO
9  67.50      1400     5 0.0             NO
10 69.75      1550     6 5.7             NO
11 70.00      1720     6 7.3            YES
12 75.50      1700     5 4.5             NO
13 77.50      1660     6 6.8            YES
14 78.00      1800     7 0.7            YES
15 81.25      1830     6 5.6            YES
16 82.50      1790     6 2.3             NO
17 86.25      2010     6 6.7            YES
18 87.50      2000     6 3.4             NO
19 88.00      2100     8 5.6            YES
20 92.00      2240     7 3.4            YES
>
```

(a)

```
> dim(data1)
[1] 20  5
```

(c)

```
> #Variables in the file are: price, FloorArea, Rooms, Age, CentralHeating
```

(d)

```
> head(data1)
  price FloorArea Rooms Age CentralHeating
1 52.00      1125     3 6.2            YES
2 54.75      1230     3 7.5             NO
```

```
3 57.50      1200     3 4.2            NO
4 57.50      1000     2 4.8            NO
5 59.75      1420     4 1.9           YES
6 62.50      1450     3 5.2           YES
> tail(data1)
   price FloorArea Rooms Age CentralHeating
15 81.25      1830     6 5.6            YES
16 82.50      1790     6 2.3             NO
17 86.25      2010     6 6.7            YES
18 87.50      2000     6 3.4             NO
19 88.00      2100     8 5.6            YES
20 92.00      2240     7 3.4            YES
```

(e)

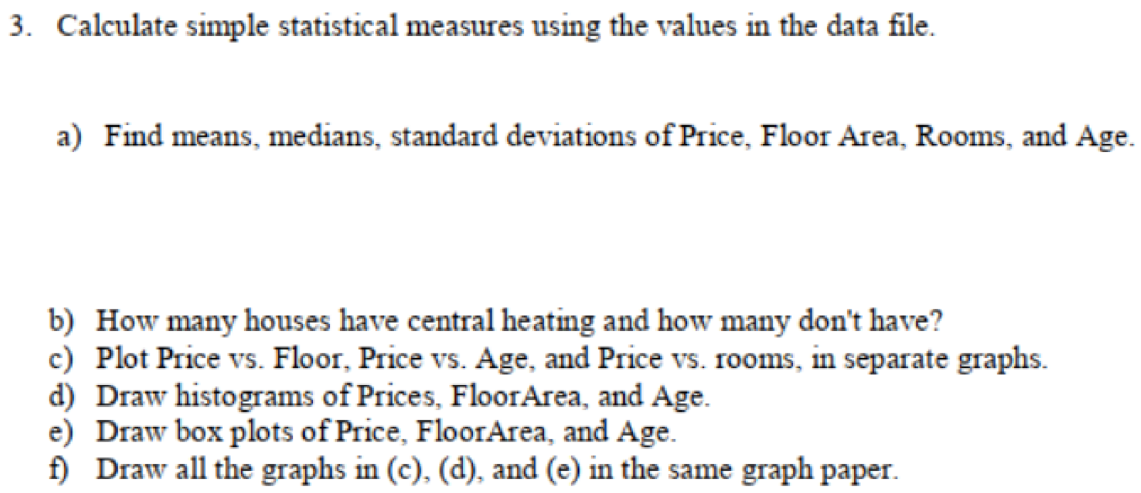
```
> data1[1:5,1:3]
  price FloorArea Rooms
1 52.00      1125     3
2 54.75      1230     3
3 57.50      1200     3
4 57.50      1000     2
5 59.75      1420     4
```

(f)

```
> data1[c(1,3,6,10),c(2,4,5)]
   FloorArea Age CentralHeating
1       1125 6.2            YES
3       1200 4.2             NO
6       1450 5.2            YES
10      1550 5.7             NO
```

(g)

```
> data1[10,]
   price FloorArea Rooms Age CentralHeating
10 69.75      1550     6 5.7             NO
> data1[,4]
 [1] 6.2 7.5 4.2 4.8 1.9 5.2 6.5 9.2 0.0 5.7 7.3 4.5 6.8
[14] 0.7 5.6 2.3 6.7 3.4 5.6 3.4
```



3. Calculate simple statistical measures using the values in the data file.

   a) Find means, medians, standard deviations of Price, Floor Area, Rooms, and Age.

   b) How many houses have central heating and how many don't have?
   c) Plot Price vs. Floor, Price vs. Age, and Price vs. rooms, in separate graphs.
   d) Draw histograms of Prices, FloorArea, and Age.
   e) Draw box plots of Price, FloorArea, and Age.
   f) Draw all the graphs in (c), (d), and (e) in the same graph paper.

(a)

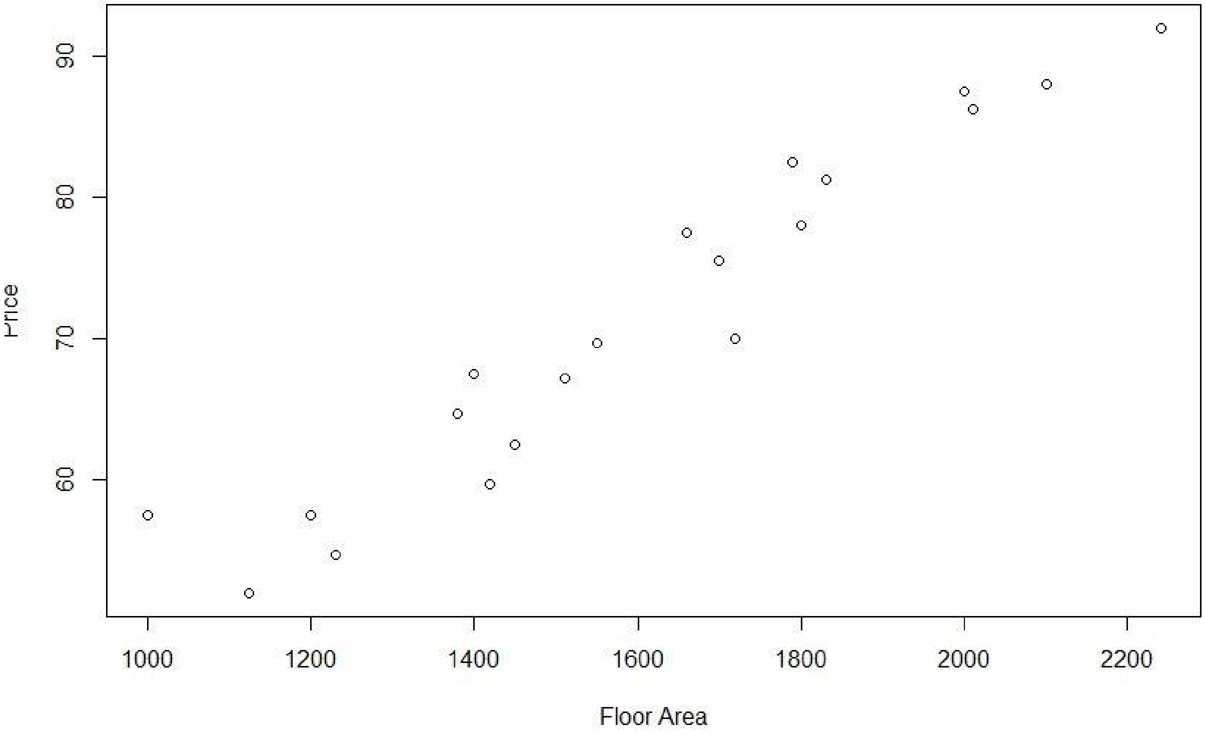
```
> mean(data1$price)
[1] 71.5875
> mean(data1$FloorArea)
[1] 1605.75
> mean(data1$Rooms)
[1] 5
> mean(data1$Age)
[1] 4.875
> median(data1$price)
[1] 69.875
> median(data1$FloorArea)
[1] 1605
> median(data1$Rooms)
[1] 5.5
> median(data1$Age)
[1] 5.4
>
> sd(data1$price)
[1] 12.21094
> sd(data1$FloorArea)
[1] 338.7643
> sd(data1$Rooms)
[1] 1.65434
> sd(data1$Age)
[1] 2.366182
```

(b)

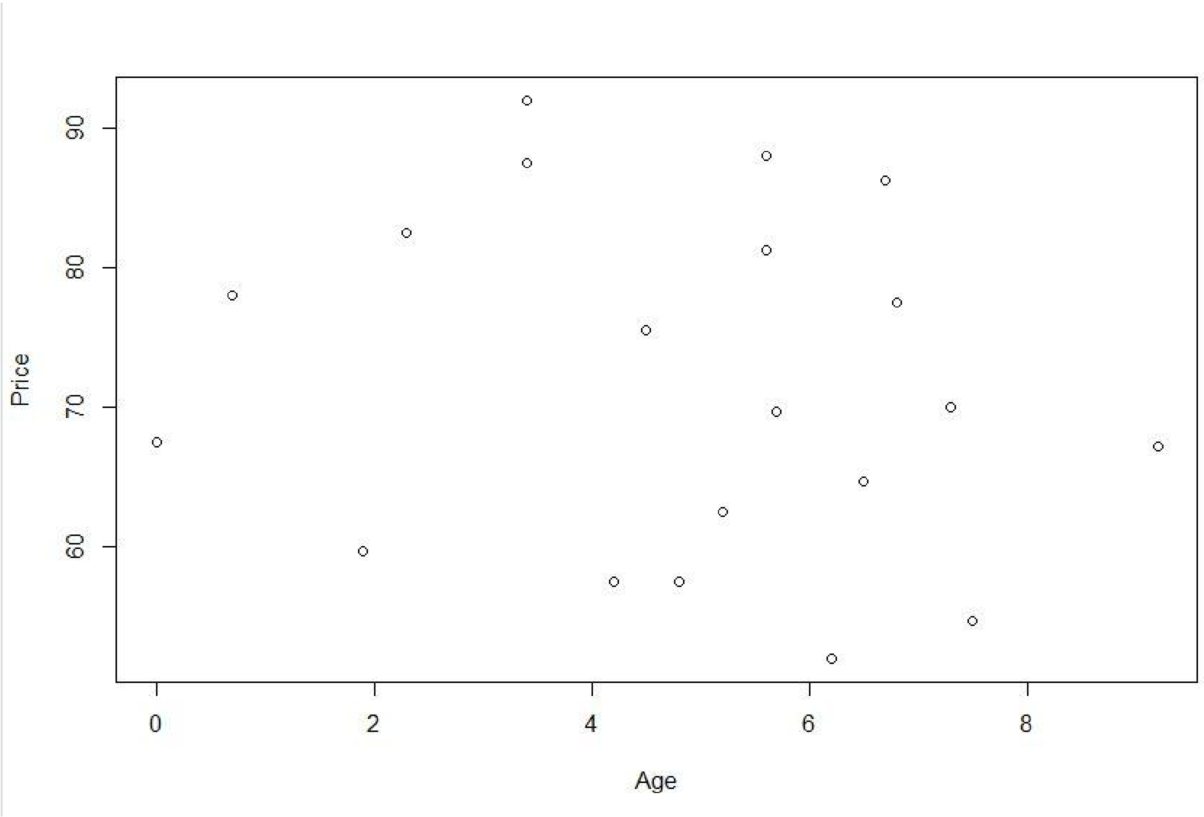
```
> sum(data1$CentralHeating== "YES")
[1] 10
> sum(data1$CentralHeating== "NO")
[1] 10
```

(c)

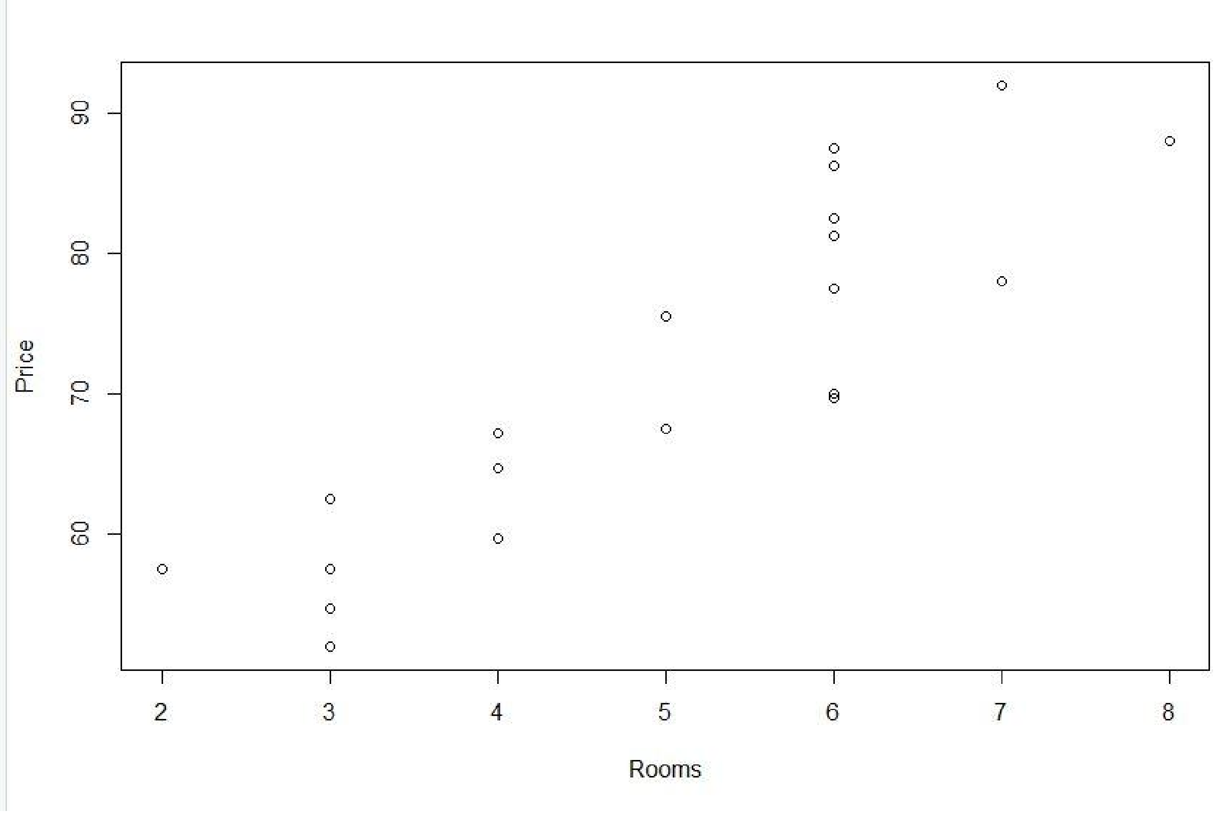
```
> plot(data1$FloorArea, data1$price, xlab = "Floor Area", ylab = "Price")
```



```
> plot(data1$Age, data1$price, xlab = "Age", ylab = "Price")
```

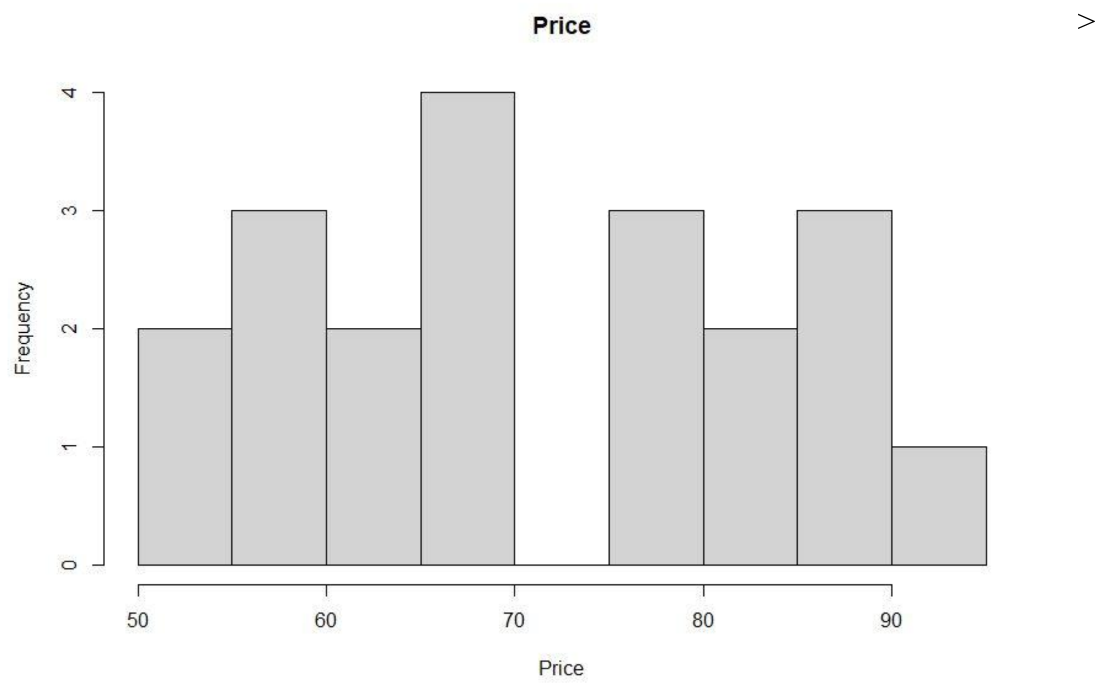


```
> plot(data1$Rooms, data1$price, xlab = "Rooms", ylab = "Price")
```

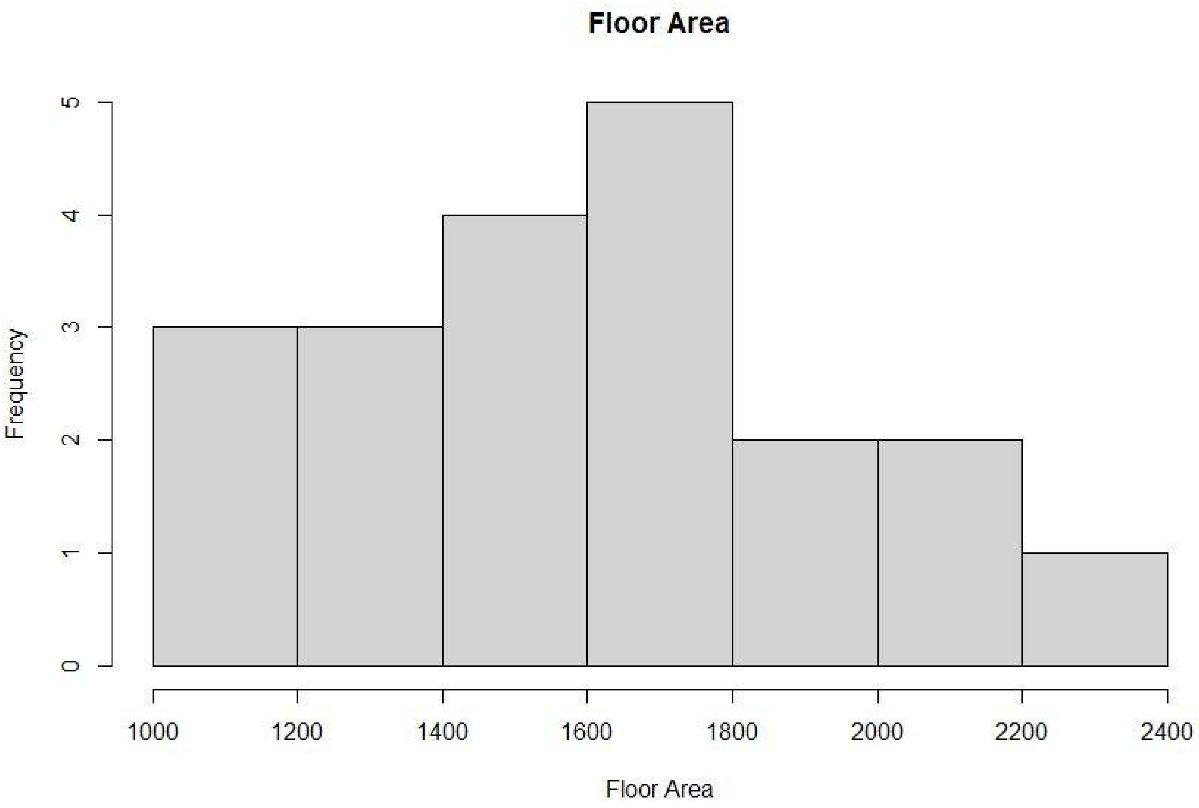


(d)

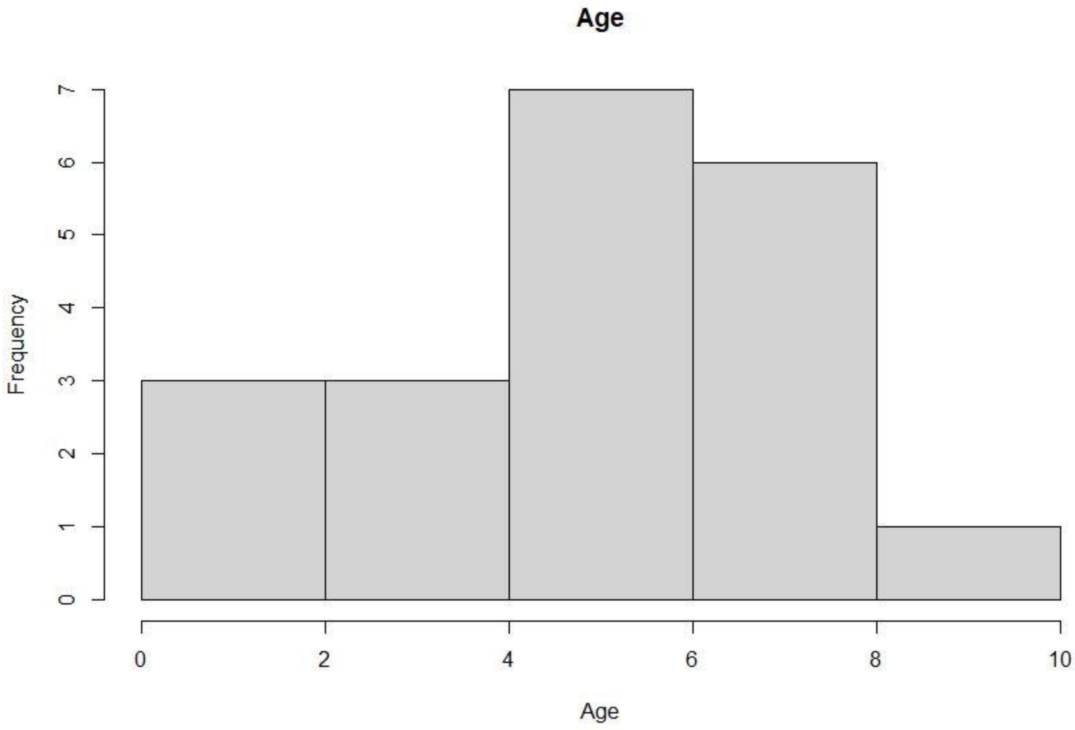
```
> hist(data1$price, main = "Price", xlab = "Price")
```



```
hist(data1$FloorArea, main = "Floor Area", xlab = "Floor Area")
```

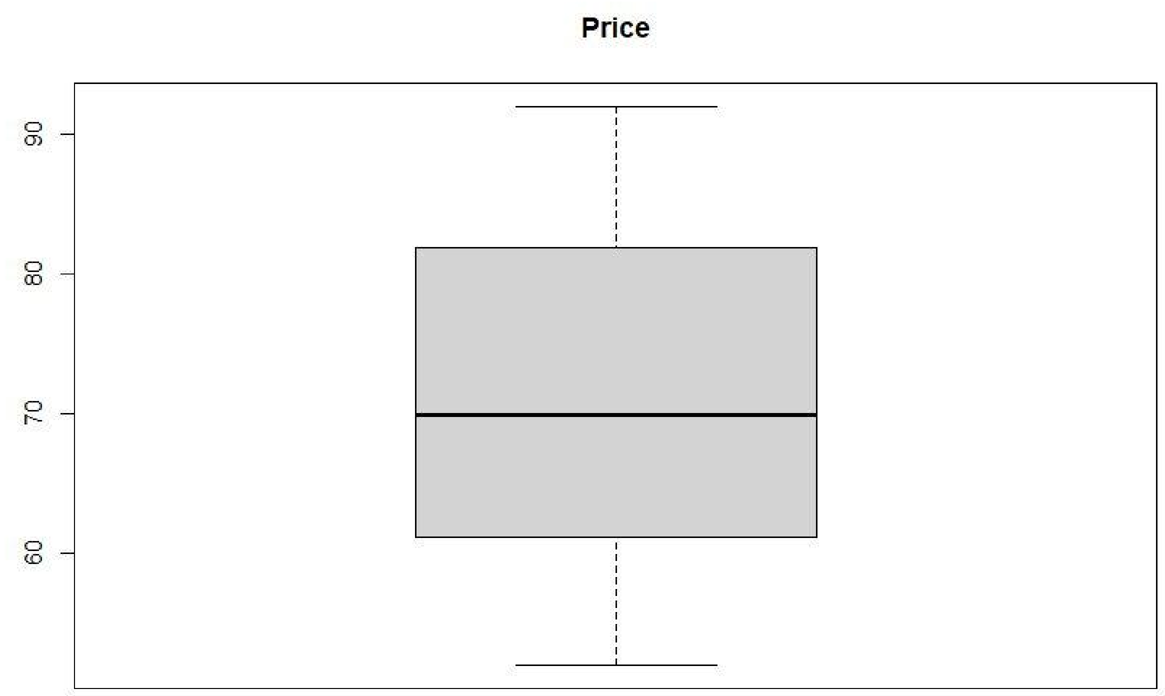


```
> hist(data1$Age, main = "Age", xlab = "Age")
```

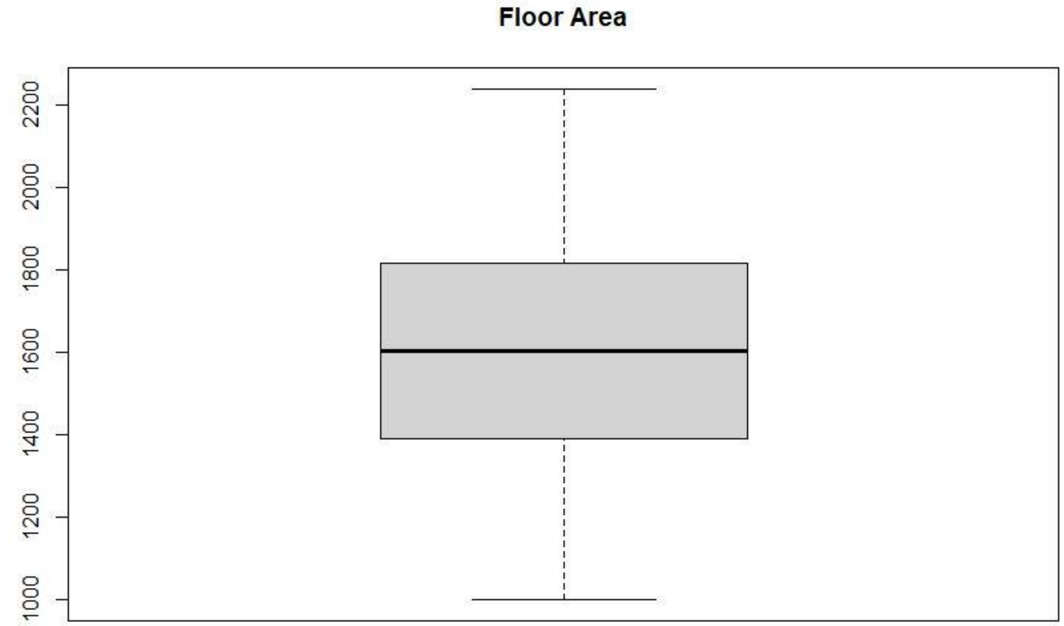


(e)

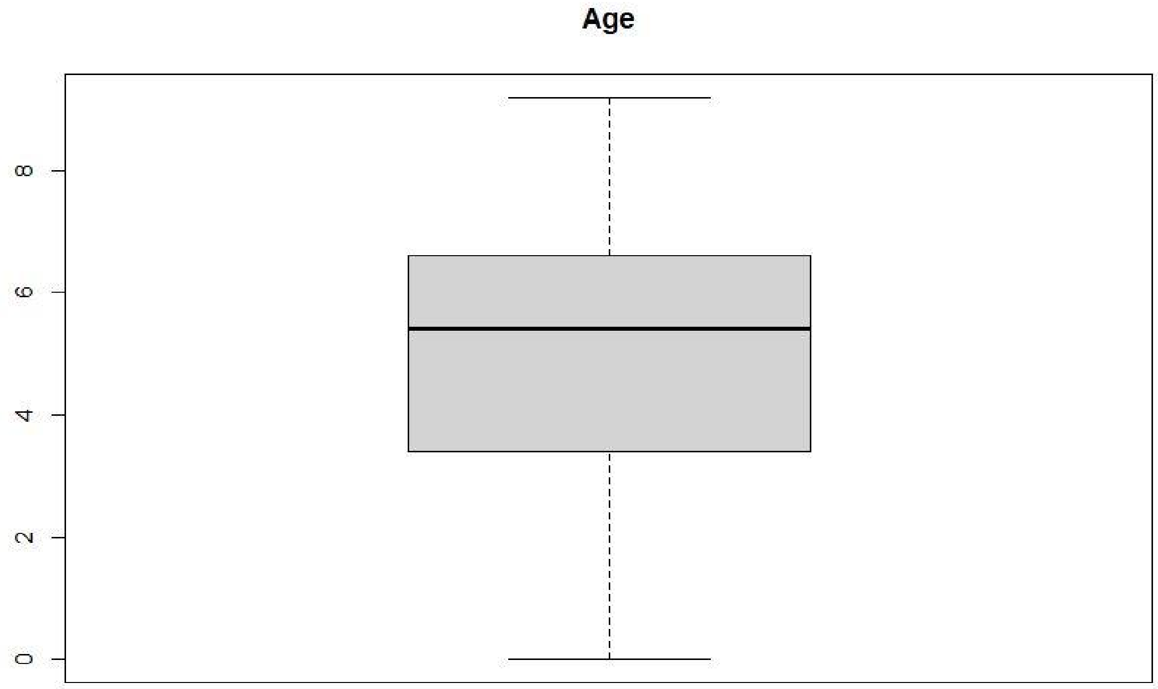
```
> boxplot(data1$price, main = "Price")
```

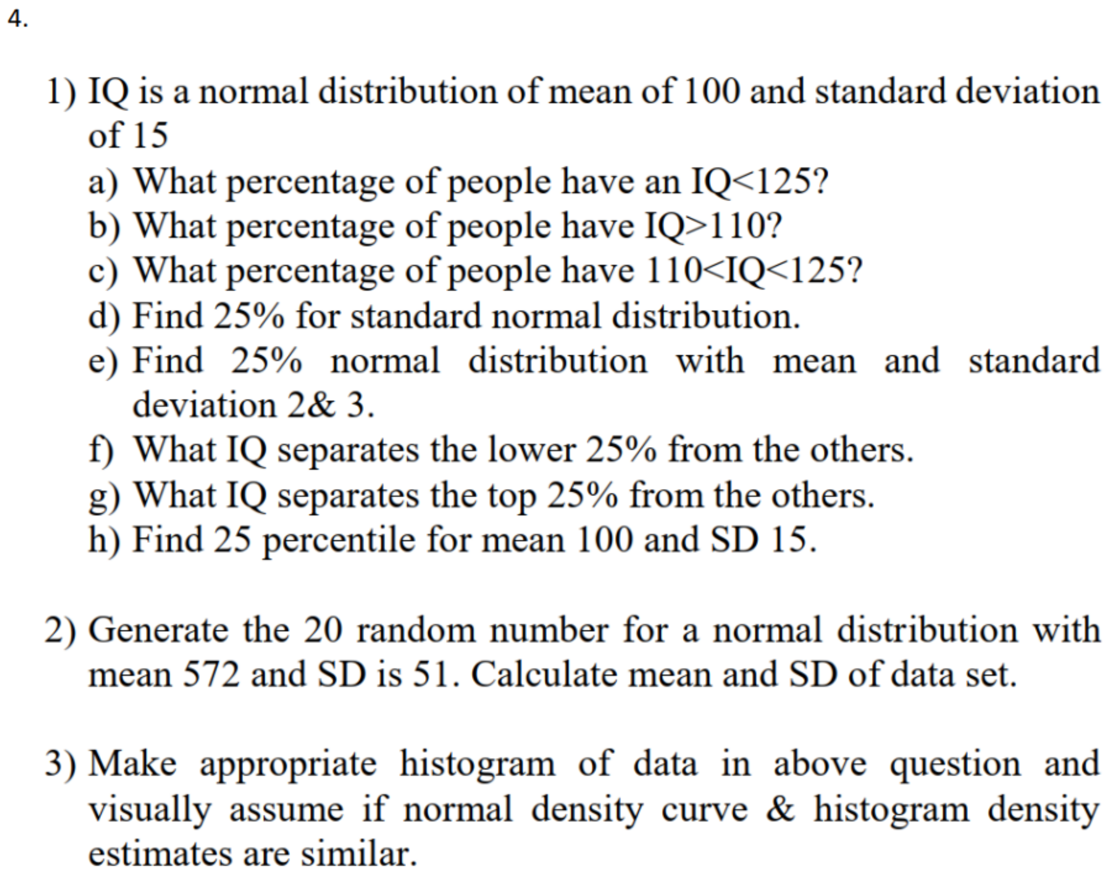


```
> boxplot(data1$FloorArea, main = "Floor Area")
```



```
> boxplot(data1$Age, main = "Age")
```





4.

1) IQ is a normal distribution of mean of 100 and standard deviation of 15
   a) What percentage of people have an IQ<125?
   b) What percentage of people have IQ>110?
   c) What percentage of people have 110<IQ<125?
   d) Find 25% for standard normal distribution.
   e) Find 25% normal distribution with mean and standard deviation 2& 3.
   f) What IQ separates the lower 25% from the others.
   g) What IQ separates the top 25% from the others.
   h) Find 25 percentile for mean 100 and SD 15.

2) Generate the 20 random number for a normal distribution with mean 572 and SD is 51. Calculate mean and SD of data set.

3) Make appropriate histogram of data in above question and visually assume if normal density curve & histogram density estimates are similar.

1)

(a)

```
> pnorm(125, mean = 100, sd = 15)
[1] 0.9522096
```

(b)

```
> pnorm(125, mean = 100, sd = 15) * 100
[1] 95.22096
```

(c)

```
> 1 - pnorm(110, mean = 100, sd = 15)
[1] 0.2524925
>
> (1 - pnorm(110, mean = 100, sd = 15)) * 100
[1] 25.24925
>
> pnorm(125, mean = 100, sd = 15) - pnorm(110, mean = 100, sd = 15)
```

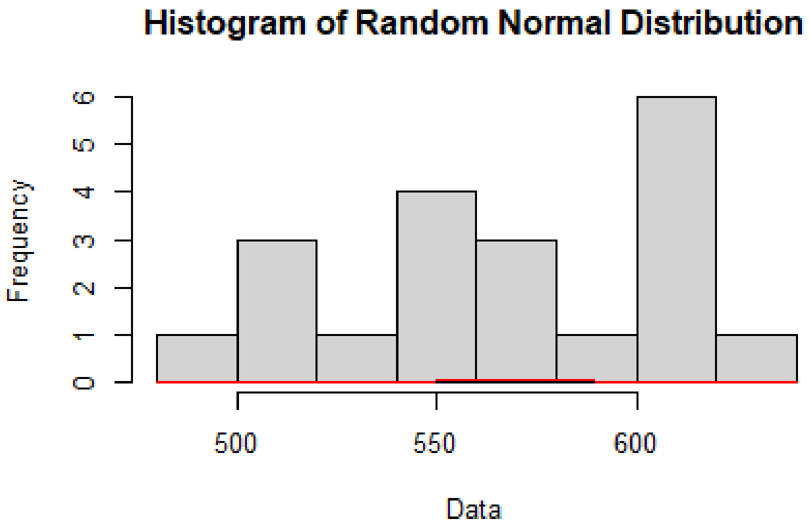
```
[1] 0.2047022
>
> (pnorm(125, mean = 100, sd = 15) - pnorm(110, mean = 100, sd = 15)) * 100
[1] 20.47022
>
> qnorm(0.25)
[1] -0.6744898
>
> qnorm(0.25, mean = 2, sd = 3)
[1] -0.02346925
>
> qnorm(0.25, mean = 100, sd = 15)
[1] 89.88265
>
> qnorm(0.75, mean = 100, sd = 15)
[1] 110.1173
>
> qnorm(0.25, mean = 100, sd = 15)
[1] 89.88265
>
> # Generating random numbers from normal distribution
> data <- rnorm(20, mean = 572, sd = 51)
```

2)

```
> # Calculating mean and standard deviation of data set
> mean(data)
[1] 579.2228
> sd(data)
[1] 49.60593
>
data <- rnorm(20, mean = 572, sd = 51)
data"
> data <- rnorm(20, mean = 572, sd = 51)
```

3)

```
> # Making a histogram of the data set
> hist(data, main = "Histogram of Random Normal Distribution", xlab = "Data")
>
> # Adding a density curve to the histogram
> curve(dnorm(x, mean = mean(data), sd = sd(data)), col = "red", add = TRUE)
```


Histogram of Random Normal Distribution
Frequency
Data