SAVEETHA SCHOOL OF ENGINEERING
AVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES

COMPUTER SCIENCE AND ENGINEERING PROGRAMME

Day 1

List of Programs:
1The intervals and corresponding frequencies are as follows. age frequency
1-5. 200
5-15 450
15-20 300
20-50 1500
50-80 700
80-110 44
Compute an approximate median value for the data

Input:

```
#age, frequency
age<-c(5,15,20,50,80,110)
frequency<-c(200,450,300,1500,700,44)
median(age)
median(frequency)
```

output:

```
> #age, frequency
> age<-c(5,15,20,50,80,110)
> frequency<-c(200,450,300,1500,700,44)
> median(age)
[1] 35
> median(frequency)
[1] 375
```

2.Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
(a) What is the mean of the data? What is the median?
(b) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
(c) What is the midrange of the data?
(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

Input:

```
#mean,median,mode,quatile
age<-c(13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70)
```

mean(age)
median(age)
mode_age<-names(table(age))[table(age)==max(table(age))]
mode_age
range(age)
quantile(age,.25)
quantile(age,.75)

output:

```
> #mean,median,mode,quatile
> age<-c(13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70)
> mean(age)
[1] 29.96296
> median(age)
[1] 25
> mode_age<-names(table(age))[table(age)==max(table(age))]
> mode_age
[1] "25" "35"
> range(age)
[1] 13 70
> quantile(age,.25)
 25%
20.5
> quantile(age,.75)
75%
 35
>
```

3.Data Preprocessing :Reduction and Transformation

Use the two methods below to normalize the following group of data: 200, 300, 400, 600, 1000 (a) min-max normalization by setting min = 0 and max = 1 (b) z-score normalization

4.Data:11,13,13,15,15,16,19,20,20,20,21,21,22,23,24,30,40,45,45,45,71, 72,73,75

a) Smoothing by bin mean
b) Smoothing by bin median
c) Smoothing by bin boundaries

input:

data <- c(11,13,13,15,15,16,19,20,20,20,21,21,22,23,24,30,40,45,45,45,71,72,73,75)
bins <- 5
bin_indices <- cut(data, bins)
mean_smooth <- tapply(data, bin_indices, mean)
print(mean_smooth)
median_smooth <- tapply(data, bin_indices, median)
median_smooth
min_max_smooth <- tapply(data, bin_indices, function(x) c(min(x), max(x)))
print(min_max_smooth)

output:

```
> median_smooth <- tapply(data, bin_indices, median)
> median_smooth
(10.9,23.8] (23.8,36.6] (36.6,49.4] (49.4,62.2] (62.2,75.1]
       19.5        27.0        45.0          NA        72.5
> data <- c(11,13,13,15,15,16,19,20,20,20,21,21,22,23,24,30,40,45,45,45,71,72,73,75)
> bins <- 5
> bin_indices <- cut(data, bins)
> mean_smooth <- tapply(data, bin_indices, mean)
> print(mean_smooth)
(10.9,23.8] (23.8,36.6] (36.6,49.4] (49.4,62.2] (62.2,75.1]
   17.78571    27.00000    43.75000          NA    72.75000
> median_smooth <- tapply(data, bin_indices, median)
> median_smooth
(10.9,23.8] (23.8,36.6] (36.6,49.4] (49.4,62.2] (62.2,75.1]
       19.5        27.0        45.0          NA        72.5
> min_max_smooth <- tapply(data, bin_indices, function(x) c(min(x), max(x)))
> print(min_max_smooth)
$`(10.9,23.8]`
[1] 11 23

$`(23.8,36.6]`
[1] 24 30

$`(36.6,49.4]`
[1] 40 45

$`(49.4,62.2]`
NULL

$`(62.2,75.1]`
[1] 71 75
```

5. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(a) Calculate the mean, median, and standard deviation of age and %fat.

(b) Draw the boxplots for age and %fat.

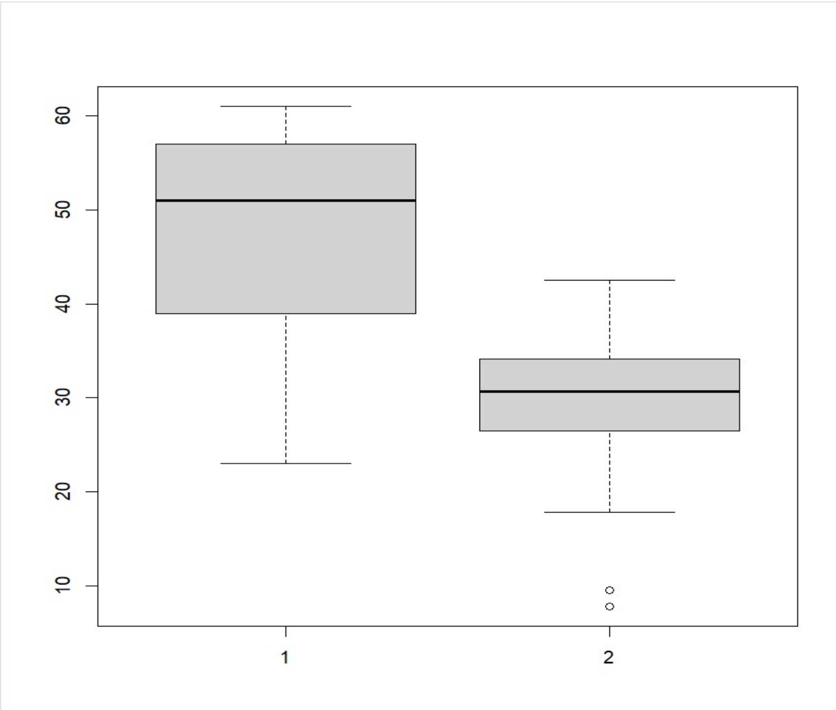
(c) Draw a scatter plot and a q-q plot based on these two variables.

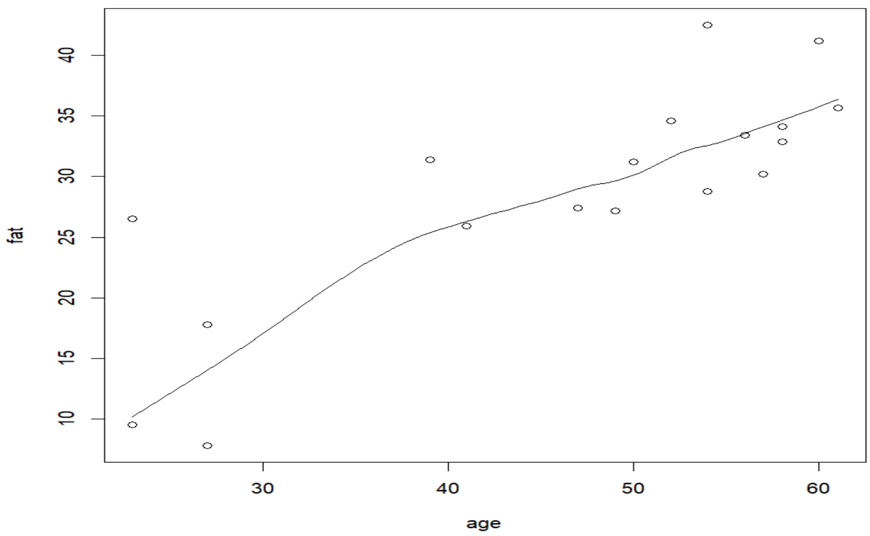
Input:

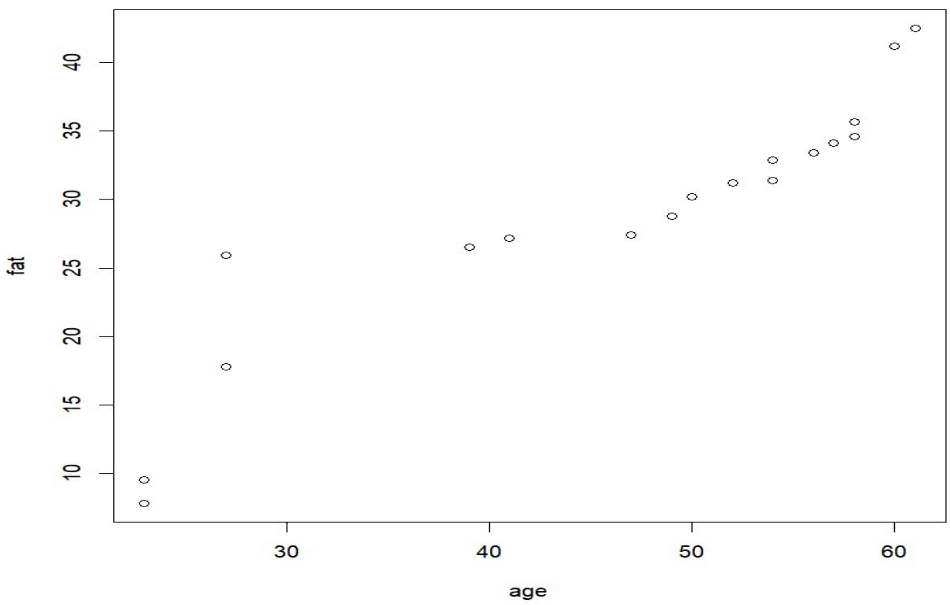
```
age<-c(23,23,27,27,39,41,47,49,50,52,54,54,56,57,58,58,60,61)
fat<-
c(9.5,26.5,7.8,17.8,31.4,25.9,27.4,27.2,31.2,34.6,42.5,28.8,33.4,30.2,34.1,32.9,41.2,35.7)
mean(age)
median(age)
```

```
sd(age)
mean(fat)
median(fat)
sd(fat)
#boxplot
boxplot(age,fat)
#scatter plot
scatter.smooth(age,fat)
#qplot
qqplot(age,fat)
```

output:







6.Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

(i) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0].
(ii) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.

(iii) Use normalization by decimal scaling to transform the value 35 for age. Perform the above functions using R – tool

Input:

```
v<-c(23,23,27,27,39,41,47,49,50,52,54,54,56,57,58,58,60,61)
min<-0
max<-1
#min_max
min_max=((35-min(v))/(max(v)-min(v)))
print(min_max)
#z-score
m=mean(v)
s<-12.94
z_score=(35-m)/s
print(z_score)
#decimal scaling
m<-35
j=max(m)<1
decimal_scaling=m/10^j

print(decimal_scaling)
```

output:

```
print(min_max)
] 0.3157895
#z-score
m=mean(v)
s<-12.94
z_score=(35-m)/s
print(z_score)
] -0.8844238
#decimal scaling
m<-35
j=max(m)<1
decimal_scaling=m/10^j
print(decimal_scaling)
] 35
```

7.The following values are the number of pencils available in the different boxes. Create a vector and find out the mean, median and mode values of set of pencils in the given data.

Box1	Box2	Box3	Box4	Box5	Box6	Box7	Box8	Box9	Box 10
9	25	23	12	11	6	7	8	9	10

Input:

```
pencils<-c(9,25,23,12,11,6,7,8,9,10)
mean(pencils)
median(pencils)
mode=names(table(pencils))[table(pencils)==max(table(pencils))]
mode
```

output:

```
pencils<-c(9,25,23,12,11,6,7,8,9,10)
mean(pencils)
] 12
median(pencils)
] 9.5
mode=names(table(pencils))[table(pencils)==max(table(pencils))]
mode
] "9"
```

8. the following table would be plotted as (x,y) points, with the first column being the x values as number of mobile phones sold and the second column being the y values as money. To use the scatter plot for how many mobile phones sold.

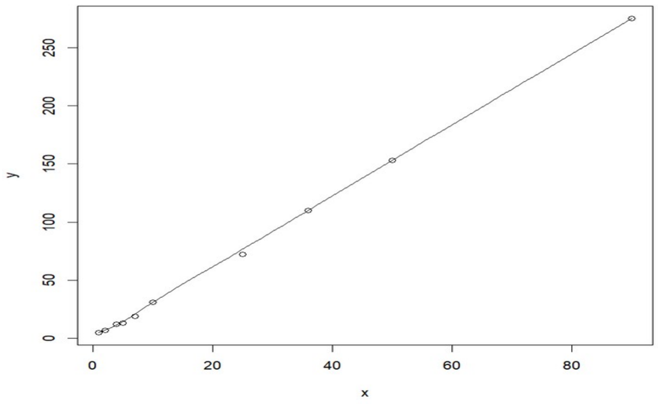
x :4 1 5 7 10 2 50 25 90 36

y :12 5 13 19 31 7 153 72 275 110

input:

```
#scatterplot
x<-c(4,1,5,7,10,2,50,25,90,36)
y<-c(12,5,13,19,31,7,153,72,275,110)
scatter.smooth(x,y)
```

output:



9. Implement of the R script using marks scored by a student in his model exam has been sorted as follows: 55, 60, 71, 63, 55, 65, 50, 55,58,59,61,63,65,67,71,72,75. Partition them into three bins by each of the following methods. Plot the data points using histogram.

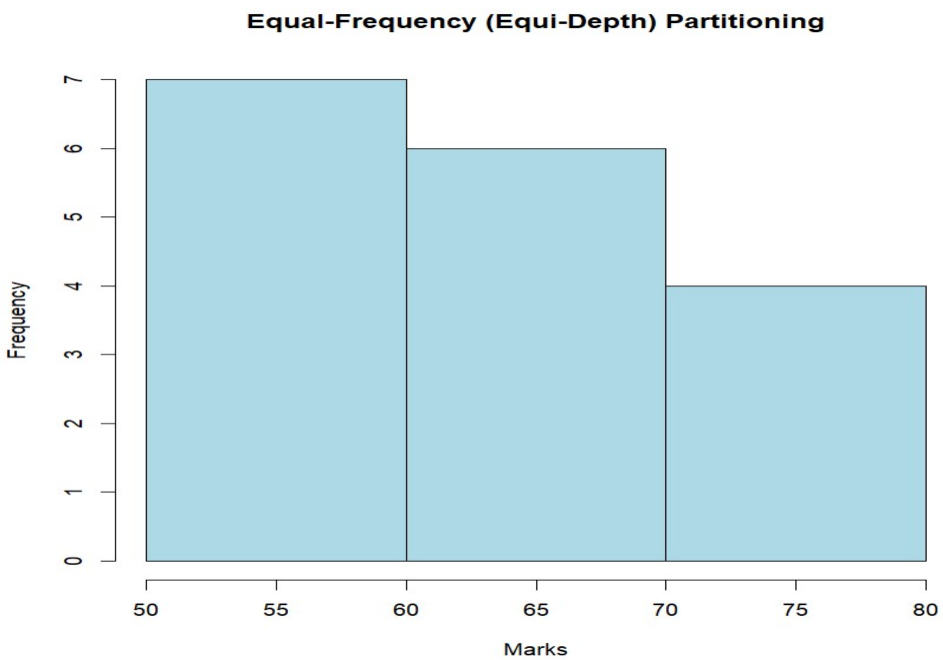
(a) equal-frequency (equi-depth) partitioning (b) equal-width partitioning

Input:

```
marks <- c(55, 60, 71, 63, 55, 65, 50, 55, 58, 59, 61, 63, 65, 67, 71, 72, 75)
num_bins <- 3
bins_eq_frequency <- cut(marks, breaks = num_bins, labels = FALSE)
hist(marks, breaks = num_bins, col = "lightblue", xlab = "Marks", main = "Equal-Frequency
(Equi-Depth) Partitioning")
marks <- c(55, 60, 71, 63, 55, 65, 50, 55, 58, 59, 61, 63, 65, 67, 71, 72, 75)
bin_mean <- tapply(data, cut(data, num_bins), mean)
smoothed_data_by_mean <- unname(bin_mean[as.character(cut(data, num_bins))])
bin_median <- tapply(data, cut(data, num_bins), median)
smoothed_data_by_median <- unname(bin_median[as.character(cut(data, num_bins))])
bin_boundaries <- tapply(data, cut(data, num_bins), function(x) c(min(x), max(x)))
smoothed_data_by_boundaries <- unlist(bin_boundaries[as.character(cut(data,
num_bins))])
print("Original data:")
print(data)
print("Smoothed data by bin mean:")
print(smoothed_data_by_mean)
print("Smoothed data by bin median:")
print(smoothed_data_by_median)
print("Smoothed data by bin boundaries:")
print(smoothed_data_by_boundaries)
```

output:

```
> print(smoothed_data_by_mean)
 [1] 18.9375 18.9375 18.9375 18.9375 18.9375 18.9375 18.9375 18.9375
 [9] 18.9375 18.9375 18.9375 18.9375 18.9375 18.9375 18.9375 18.9375
[17] 43.7500 43.7500 43.7500 43.7500 72.7500 72.7500 72.7500 72.7500
> print("Smoothed data by bin median:")
[1] "Smoothed data by bin median:"
> print(smoothed_data_by_median)
 [1] 20.0 20.0 20.0 20.0 20.0 20.0 20.0 20.0 20.0 20.0 20.0 20.0 20.0
[14] 20.0 20.0 20.0 45.0 45.0 45.0 45.0 72.5 72.5 72.5 72.5
> print("Smoothed data by bin boundaries:")
[1] "Smoothed data by bin boundaries:"
> print(smoothed_data_by_boundaries)
(10.9,32.3]1 (10.9,32.3]2 (10.9,32.3]1 (10.9,32.3]2 (10.9,32.3]1
          11           30           11           30           11
(10.9,32.3]2 (10.9,32.3]1 (10.9,32.3]2 (10.9,32.3]1 (10.9,32.3]2
          30           11           30           11           30
(10.9,32.3]1 (10.9,32.3]2 (10.9,32.3]1 (10.9,32.3]2 (10.9,32.3]1
          11           30           11           30           11
(10.9,32.3]2 (10.9,32.3]1 (10.9,32.3]2 (10.9,32.3]1 (10.9,32.3]2
          30           11           30           11           30
```



10. Suppose that the speed car is mentioned in different driving style.

Regular 78.3 81.8 82 74.2 83.4 84.5 82.9 77.5 80.9 70.6 Speed

Calculate the Inter quantile and standard deviation of the given data.

Input:

```
#IQR, SD
v<-c(78.3,81.8,82,74.2,83.4,84.5,82.9,77.5,80.9,70.6)
```

IQR(v)
sd(v)

output:

```
v<-c(78.3,81.8,82,74.2,83.4,84.5,82.9,77.5,80.9,70.6)
IQR(v)
.] 4.975
sd(v)
.] 4.445835
```

11.Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

Input:

#Q1, Q2
age<-c(13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70)
quantile(age,.25)
quantile(age,.75)

output:

```
> #Q1, Q2
> age<-c(13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70)
> quantile(age,.25)
 25%
20.5
> quantile(age,.75)
75%
 35
```