

NAME:-SYED.MANSOOR BASHA

REG.NO:-192111385

SUB.CODE:-CSA1658

SUB.NAME:-DATA WAREHOUSING AND DATA MINING

FACULTY NAME:-DR.ARUMUGAM

DAY-1

1.The 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

PROGRAM:-

#age, frequency

age<-c(5,15,20,50,80,110)

frequency<-c(200,450,300,1500,700,44)

median(age)

median(frequency)

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?

PROGRAM:-

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)

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

PROGRAM:-

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)

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



1. 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.

PROGRAM:-

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)

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

PROGRAM:-

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)

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.

PROGRAM:-

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

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

PROGRAM:-

#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)

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

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)

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.

PROGRAM:-

#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)

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?

PROGRAM:-

#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)