Vidyavardhini's College of Engineering and Technology

Department of Artificial Intelligence & Data Science

AY: 2024-25

Class:	TE	Semester:	V
Course Code:	CSC504	Course Name:	D.W.M

Name of Student:	Sainath Khot
Roll No. :	20
Assignment No.:	2
Title of Assignment:	Intro to Data Mining
Date of Submission:	
Date of Correction:	

Evaluation

Performance Indicator	Max. Marks	Marks Obtained
Completeness	5	4
Demonstrated Knowledge	3	3
Legibility	2	2
Total	10	9

Performance Indicator	Exceed Expectations (EE)	Meet Expectations (ME)	Below Expectations (BE)
Completeness	5	3-4	1-2
Demonstrated Knowledge Legibility	3	2	1
Legibility	2	1	0

Checked by

Name of Faculty :

Signature :

Date :

Q. Suppose that the data for analysis include 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) Mean of the data : [13 + 15 + 16 + 16 + 19 + 20 + 20 + 21 + 22 + 22 + 25 + 25 + 25 + 25 + 30 + 33 + 35 + 33 + 35 + 35 + 35 + 35 + 36 + 40 + 45 + 46 + 52 + 70] / 27

= 29.962

Median of data = 25

b) Mode: This dataset has 2 modes, viz 25 & 35
Thus the dataset is bimodal

c) Mid range = (lowest value + highest value) / 2
= (13 + 70) / 2
= 41.5

d) Quartile (m) = (n+1) × m/4
∴ First Quartile (Q_1) = (27+1) × 1/4
= 7th element = 20

∴ Third Quartile = (27 + 1) × 3/4
= 21st element = 35

e) Mean of the given data is 29.962
Median of the given data is 25
Mode of given data is 25 & 35

Midrange of the data is 41.5

Quartile of the data are:

$Q_1 = 20$, $Q_2 = 25$, $Q_3 = 35$, $Q_4 = 70$

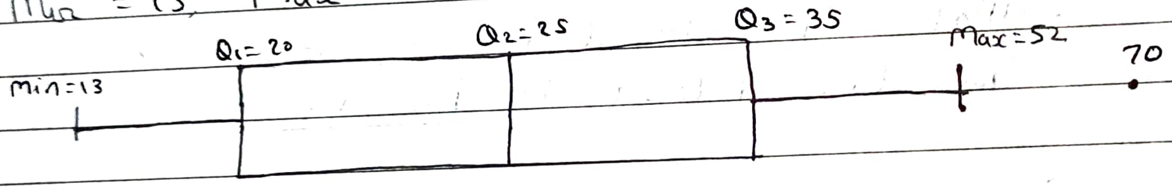
(f) Boxplot:

Inter Quartile Range (IQR) = $Q_3 = Q_1 = 35 \cdot 20 = 15$

Lower limit = $Q_1 - 1.5 \times IQR = 20 - [1.5 \times 15] = -2.5$

Upper limit = $Q_3 + 1.5 \times IQR = 35 + [1.5 \times 15] = 57.5$

Min = 13, Max = 52 Outlier = 70



Q2]

Age	frequency	Cumilative frequency
1-5	200	200
6-15	450	650
16-20	300	950
21-50	1500	2450
51-80	700	3150
81-110	44	3194

n = 3194

n/2 = 1597

This observation lies bet^n the class interval 21-50 which is the median class.

Lower class limit = 21

Class size (h) = 30

frequency of median class (f) : 1500

Cumulative freq of class preceding the median class (cf) : 950

$$\boxed{\text{Median} = l + \left(\frac{\frac{n}{2} - cf}{f} \right) \times h}$$

$$= 21 + \frac{(1597 - 950)}{1500} \times 30$$

$\therefore$ Median = 33.94

$P_1 (0,2)$, $P_2 (2,0)$, $P_3 (3,1)$ $P_4 (5,1)$

Euclidean distance = $[(x_2 - x_1)^2 + (y_2 - y_1)^2]^{1/2}$

$d(P_1 P_2) = [(2-0)^2 + (0-2)^2]^{1/2} = 2.828$

$d(P_1 P_3) = [(3-0)^2 + (1-2)^2]^{1/2} = 3.162$

$d(P_1 P_4) = [(5-0)^2 + ((1-2)^2]^{1/2} = 5.099$

$d(P_2 P_3) = [(3-2)^2 + (1-0)^2]^{1/2} = 1.414$

$d(P_2 P_4) = [(5-2)^2 + (1-0)^2]^{1/2} = 3.162$

$d(P_3 P_4) = [(5-3)^2 + (1-1)^2]^{1/2} = 2$

P_1	0	2.828	3.162	5.099
P_2	2.818	0	1.414	2
P_3	3.162	1.414	0	3.162
P_4	5.099	2	3.162	0
	P_1	P_2	P_3	P_4

Qs Data : 2, 10, 18, 18, 19, 20, 22, 25, 28
Bin size : 3

Soln: As data is already sorted in increasing order, divide the data into bins of size 3.

Bin 1: 2, 10, 18
Bin 2: 18, 19, 20
Bin 3: 22, 25, 28

- Smoothing by bin mean

Mean (Bin1) = (2+10+18)/3 = 10
Mean (Bin2) = (18+19+20)/3 = 19
Mean (Bin3) = (22+25+28)/3 = 25

- Replacing each value in the bin with its mean.

Bin1: 10, 10, 10
Bin2: 19, 19, 19
Bin3: 25, 25, 25

- Smoothening by bin median [Replacing each value in the bin with its median]

Bin1: 10, 10, 10
Bin2: 19, 19, 19
Bin 3: 25, 25, 25

- Smoothening by bin boundaries [Replacing each element by value it is closer to (1st or the last)]

Bin1: 2, 2, 18 ; Bin2: 18, 18, 20
Bin 3: 22, 22, 28