Vidyavardhini's College of Engineering and Technology Department of Artificial Intelligence & Data Science

Name:	Shreeya Sunil Hudekar
Roll No:	13
Class/Sem:	TE/V
Experiment No.:	3
Title:	Tutorial a) Data Exploration b) Data
Date of Performance:	02/08/24
Date of Submission:	09/08/24
Marks:	
Sign of Faculty:	

M NAROHINA M NAROHINA

Vidyavardhini's College of Engineering and Technology Department of Artificial Intelligence & Data Science

Aim: To solve problems in Data Exploration and Data Pre-processing.

Objective: To enable students to effectively identify sources of data and process it for data mining.

- 1. 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, 36, 40, 45, 46, 52, 70.
- a. What is the mean of the data? What is the median?
- a. What is the mode of the data? Comment on the data's modality (i.e., unimodal, bimodal, trimodal, etc.).
- a. What is the midrange of the data?
- a. Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
- a. Give the five-number summary of the data.
- a. Show a boxplot of the data.
- 2. Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows:

age	frequency
1–5	200
6–15	450
16–20	300
21–50	1500
51-80	700
81–110	44

Compute an approximate median value for the data.

- 3. Consider the data given below and compute the Euclidean distance between each point. P1 (0,2), P2(2,0), P3(3,1) and P4(5,1).
- 4. Suppose that the minimum and maximum values for the attribute income are \$12,000 and \$98,000 respectively. Normalize income value \$73,600 to the range [0.0, 1.0] using min-max normalization method.
- 5. Partition the given data into bins of size 3 using equi-depth binning method and perform smoothing by bin mean, bin median and bin boundaries. Consider the data: 2, 10, 18, 18, 19, 20, 22, 25, 28.

	The cone
4.	suppose that he data for analysis includes the attribute age. The age
α,	values for the data tuples are (in intreasing order).
	values for the data tuples are (in intreasing order. 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33,
	13, 15, 16, 16, 19, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20
	35, 35, 35, 36, 40, 45, 46, 52, 70
	i what is the mean of the data?
	Mean - Exi , n = no. of dataported
	n xi where i = 0, 1, 2, 3,
-	: £xi = 13+15+16+16+19+20+20+21+22+22+25+25+25+25+25+25+25+25+26+40+45+46+52+76
	30+33+33+35+35+35+36+40+45+46+52+70
	60. 808
	: £xi = 808
	n ∴ n = 27
	: Mean = 809 = 29.96 = 30
	27
	ii. What is the median?
	Median is the middle value in a set of data.
	· Median = 25
•	: Mearan - 23
	iii. What is me made of the data? comment on data's modality.
	Two values are reneased four times.
	Mode = 25,35
	Modality = Bimodal.
	iv what is the midsange of the data?
	Midsange = minimum value + maximum value = 13+70
	2
	Midsange = 41.5
Sundaram	FOR EDUCATIONAL USE

V. Q1 & Q3 of the data. Median ie. Q2 is 25. for finding Q1, dataset will be: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25 Q1 = (Lower of Lower bound + higher of Lower bound) Q1 = 20 Q1 = 19 : Q1 = 13 + 25 For finding Q2, dataset will be: 19 30, 23, 33, 25, 35, 35, 35, 36, 40, 45, 46, 52, 70 az = (bourer of higher bound + higher of upper bound) Q3 = 35 02 = 50 . Q2 = 30+70 vi. Give the fire-number surmary of data: a. Minimum value = 13 b. Q1 = 1 20 c. Q2 = 25 d. Q3 = 50 35 e. Maximum value = 70 Vii. Boxplot. Q1 Q2 03 04 20 30 40 50 60 70 80 FOR EDUCATIONAL USE (Sundaram)

R(0,2)					ance bt
P2 (2,0)		5	+		
II and the state of the state o		4	1		
			0,		
		2			P 3 ⊕
Formula: d(x,y) = 2 (xi - yi)2				P1	3 4
iel			0 1	7	3 4
	1100		x	-	
				_	
d (P, Ph) = J(0-5)2+ (2-1)2 = 5.09				1	
1 1 12 1 12				1	
		* 4			
		P.	P ₂	Р3	P4
a (12,14) - 3(2 3) + (021) - 3.11	Pı	0		3.16	5.09
$a(e_2, e_3) = ((3-e)^2 + (1-2)^2 = 3.16$		2.82	0	1,41	3.16
		3.16	1-41	0	2
	P4	5.09	3.16	2_	0
$d(\rho_{1}, \rho_{1}) = \int (5-0)^{2} + (1-2)^{2} = 5.09$					
	$\begin{aligned} & \rho_{2}(2,0) \\ & \rho_{3}(3,1) \\ & \rho_{41}(5,1) \end{aligned}$ Formula: $d(x,y) = \int_{1=1}^{\infty} (\alpha_{1} - y_{1})^{2} \\ & d(\rho_{1},\rho_{2}) = \int_{1}^{\infty} (0-2)^{2} + (2-0)^{2} = 2.82 \\ & d(\rho_{1},\rho_{3}) = \int_{1}^{\infty} (0-3)^{2} + (2-1)^{2} = 3.16 \\ & d(\rho_{1},\rho_{1}) = \int_{1}^{\infty} (0-5)^{2} + (0-1)^{2} = 5.09 \end{aligned}$ $d(\rho_{2},\rho_{1}) = \int_{1}^{\infty} (2-3)^{2} + (0-1)^{2} = 1.41 \\ d(\rho_{2},\rho_{3}) = \int_{1}^{\infty} (2-3)^{2} + (0-1)^{2} = 3.16 \\ d(\rho_{3},\rho_{1}) = \int_{1}^{\infty} (3-0)^{2} + (1-2)^{2} = 3.16 \\ d(\rho_{3},\rho_{2}) = \int_{1}^{\infty} (3-2)^{2} + (1-0)^{2} = 1.41 \\ d(\rho_{3},\rho_{4}) = \int_{1}^{\infty} (3-5)^{2} + (1-1)^{2} = 2 \end{aligned}$ $d(\rho_{4},\rho_{1}) = \int_{1}^{\infty} (5-0)^{2} + (1-2)^{2} = 3.16 \\ d(\rho_{3},\rho_{4}) = \int_{1}^{\infty} (3-5)^{2} + (1-1)^{2} = 2 \end{aligned}$	$P_{21}(S, I)$ $P_{22}(S, I)$ $P_{23}(S, I)$ $P_{24}(P_{21}, P_{22}) = \int (0-2)^{2} + (2-0)^{2} = 2.82$ $P_{24}(P_{21}, P_{21}) = \int (0-5)^{2} + (0-2)^{2} = 3.16$ $P_{24}(P_{21}, P_{21}) = \int (2-3)^{2} + (0-1)^{2} = 1.41$ $P_{24}(P_{21}, P_{21}) = \int (2-0)^{2} + (1-0)^{2} = 3.16$ $P_{24}(P_{21}, P_{22}) = \int (3-2)^{2} + (1-0)^{2} = 1.41$ $P_{24}(P_{21}, P_{22}) = \int (3-2)^{2} + (1-0)^{2} = 1.41$ $P_{25}(P_{21}, P_{22}) = \int (3-5)^{2} + (1-1)^{2} = 2$ $P_{25}(P_{21}, P_{22}) = \int (3-5)^{2} + (1-1)^{2} = 3.16$ $P_{25}(P_{21}, P_{22}) = \int (3-5)^{2} + (1-1)^{2} = 3.16$	$P_{3}(3,1)$ $P_{41}(5,1)$ 2 Formula: $d(x,y) = \int_{1-1}^{1} (\alpha_{1} - y_{1})^{2}$ $d(P_{1}, P_{2}) = \int_{1}^{2} (0 - 2)^{2} + (2 - 0)^{2} = 2.82$ $d(P_{1}, P_{3}) = \int_{1}^{2} (0 - 3)^{2} + (2 - 1)^{2} = 3.16$ $d(P_{1}, P_{1}) = \int_{1}^{2} (0 - 5)^{2} + (2 - 1)^{2} = 5.09$ P_{2} $d(P_{2}, P_{1}) = \int_{1}^{2} (2 - 9)^{2} + (0 - 2)^{2} = 2.82$ $d(P_{2}, P_{3}) = \int_{1}^{2} (2 - 3)^{2} + (0 - 1)^{2} = 1.41$ $d(P_{2}, P_{1}) = \int_{1}^{2} (2 - 5)^{2} + (0 - 1)^{2} = 3.16$ $P_{1} = \int_{1}^{2} (2 - 6)^{2} + (1 - 6)^{2} = 1.41$ $d(P_{3}, P_{2}) = \int_{1}^{2} (3 - 2)^{2} + (1 - 6)^{2} = 1.41$ $d(P_{3}, P_{4}) = \int_{1}^{2} (3 - 5)^{2} + (1 - 1)^{2} = 2$ $d(P_{3}, P_{4}) = \int_{1}^{2} (3 - 5)^{2} + (1 - 1)^{2} = 2$ $d(P_{4}, P_{4}) = \int_{1}^{2} (5 - 6)^{2} + (1 - 2)^{2} = 3.16$	$\begin{array}{cccccccccccccccccccccccccccccccccccc$	$\begin{array}{cccccccccccccccccccccccccccccccccccc$

Q 2.	Suy	Suppose that the values for a given see of data are grouped into intervals.						
	Age Gegrency cummulative frequency							
	a.	1-5	200	200				
	ь.	6-15	450	650				
	C .	16-20	300	950				
	d.	21-50	1500	2450				
	e.	51-80	700	3150				
	9.	81-110	44	3194				
- (cu	mulation	d Bearing	401:	2			
,	a.	0+200=	200	b = 200+450 = 650				
	_ c	= 650+ 2	00 = 950	d= 950+ 1500 = 2450				
	e	= 2450+	700 = 3150	; e = 3150 + 44 = 319) /.			
			5130	, e = 3154 +44 = 319	1 4			
	1		70 Pari 2026					
		The formula to calculate the median downing the median day:						
		formule	A to calcular	te the median dothin the m	edian Uay:			
		formule	x to calcular Median = 1	te the median disthin the m $L + (\frac{N/2 - CF}{2}) \times w$	edian Uay;			
			Median = 1	1 + (N/2 - CF) x w	edian Uay:			
	L =	= loves b	Median = 1 Oundary	1 - 21 - 21	edian Uass:			
	L:	= lower be	Median = 1 Oundary Oservation	$\frac{L + (N/2 - CF) \times w}{8}$ = 21 = 3194	edian Uay:			
	L :	= lower by = total of = cumula	Median = 1 Oundary bservation ative frequer	$\frac{L + (N/2 - CF) + w}{8}$ = 21 = 3194 = 950	edian Uay:			
	L :	= lower by = total of = cumula = frequence	Median = 1 Oundary Eservation ative frequer	$\frac{L + (N/2 - CF) + w}{8}$ = 21 = 3194 (cy = 950 = 1500	edian Uay:			
	L :	= lower by = total of = cumula = frequence = width	Median = 1 Oundary bservation ative prequer y of median c	$\frac{L + (N/2 - CF) + w}{8}$ = 21 = 3194 = 950 = 1500 Lass = 29	edian Uay:			
	L :	= lower by = total of = cumula = frequence = width	Median = 1 Oundary bservation ative prequer y of median c	$\frac{L + (N/2 - CF) + w}{8}$ = 21 = 3194 = 950 = 1500 Lass = 29				
	L :	= lower by = total of = cumula = frequence = width	Median = 1 Oundary bservation ative prequer y of median c	$\frac{L + (N/2 - CF) + w}{8}$ = 21 = 3194 (cy = 950 = 1500 (ass = 29	redian Way:			
	L :	= lower by = total of = cumula = frequence = width	Median = 1 Sundary bservation ative frequer y of median c Median =	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$				
	L :	= lower by = total of = cumula = frequence = width	Median = 1 Sundary bservation ative frequer y of median c Median =	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$				
	L :	= lower by = total of = cumula = frequence = width	Median = 1 Sundary bservation ative frequer y of median c Median =	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$				
	L :	= lower by = total of = cumula = frequence = width	Median = 1 Sundary bservation ative frequer y of median c Median =	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$				
	L :	= lower by = total of = cumula = frequence = width	Median = 1 Sundary bservation ative frequer y of median c Median =	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$				
	L :	= lower by = total of = cumula = frequence = width	Median = 1 Sundary bservation ative frequer y of median c Median =	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$				
/anj	L :	= lower by = total of = cumula = frequence = width	Median = 1 Sundary bservation ative frequer y of median c Median =	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$				

04	Minimum Value: \$ 12000 = mins						
	Maximum value: \$ 78000 = maxa						
	Range = [0.0, 1.0]						
	V = \$ 73600						
	Min-Max normalization formulae						
	-V'=V-minA						
	max - min (newmax, - newmina) + newmina						
	· V' = 73600 -12000 = 0.7163						
	98000-12000(1-0)+1						
•	. Income \$ 73600 is transpormed to 0.7165						
05	Data: 2,10, 18,18,19,20,22,25,28						
	n = 3						
	Three equi-depth buis of size 3						
	Bin1= 2,10,58						
	Bin 2: 18, 19,20						
	Bin 3 = 22, 25, 20						
	smoothing by bin means:						
•	Bin1: 10,10,10						
	Binz: 19,59,19						
	Bin 3 : 25, 25, 25						
	smoothing by bin boundaries:						
	Bin 1: 2,18,18						
	Bin 2: 18, 20, 20						
	Bin 3 : 28, 28, 28						
	smoothing by bin median:						
	Bin 1: 10,10,10						
	Bin 2: 19,19,19						
	Bin 3: 25, 25, 25						
Sundaram	FOR EDUCATIONAL USE						