Name:	Swarup Sat	ish Kal	cade				
Roll No:	17						
Class/Sem:	TE/V						
Experiment No.:	3						
Title:	Tutorial	on:	a)	Data	Exploration	b)	Data
	pre-proce	ssing					
<b>Date of Performance:</b>							
<b>Date of Submission:</b>							
Marks:							
Sign of Faculty:							



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?
- b. What is the mode of the data? Comment on the data's modality (i.e., unimodal, 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?
- e. Give the five-number summary of the data.
- f. Show a boxplot of the data.

#### **Solution:**

	Aim: To selve Roblems in Dato Exploration & Dato Pre-pocessing
	to I Trace
Ć	Suppose that the data for analysis includes the attribute age. The age
	values for the data typies are (in mores in acces) 1,15, 15, 15, 15, 15, 15, 15, 15, 15, 1
	Separse that the data for analysis includes the attribute again 13, 16, 16, 19, 20, values for the data tiples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 12, 25, 25, 25, 25, 25, 26, 33, 33, 35, 35, 35, 36, 44, 45, 46, 52, 25
	gurates the mean of data ) who his the median ?
(	dictaris the moon of data to the
	Hear, N=27 Mean Zz: = 13+15+6+18+19+20+2+12+22+25+25+25+25+25-25-25-25-25-25-25-25-25-25-25-25-25-2
	Mean 21 = 53+33+35+35+35+35+36+40+45+46+52+10
	/07
	= 809 - 29.96.
	27
	Me dian -> Middle value
	= 25/
	What is the mode of the data Comment on the data's modality?
$\mathcal{G}$	Mode = roost occured values
3	In the above sequence 15 & 35 are most occurred values
	Mode 25, 35 is bimodal
C	what is the midsange of the data
4	Midrange= Ittin vale + Max value /2
	,
	· 13H0/2
	Midrange 41.5
Cx	6 (6 ( 1 ) H & h - 16 (0) K 1) 1) 1 - 1 - 1 HA
(d)	Can you find (roughly) the first quartite (gi) the third quartitle)
Sundaram	- FOR EDUCATIONAL USE



## Vidyavardhini's College of Engineering and Technology

#### Department of Artificial Intelligence & Data Science

	the dato
=>	A : Median = 25
	Q = first Quartile = Middle value of 13, 15, 16, 16, 19, 20, 20, 21, 22, 27, 25, 25
	Q <sub>1</sub> : 20 ,,
	V .
	Q2 = Third Quartile= Hiddle value of 30, 33, 35, 35, 35, 35, 35, 36, 45, 45, 15, 25, 70
	\$ 35,,
,	C: B. Francisco man [ H. ] I-
<u> </u>	Give the 5 number Summary of the data -
	Frot Quartile=20
	Median value 25
	Third Qualik: 35 Haximum value: 70
$\mathcal{B}$	Show a bexplot of the data
	)
	- 9E
	13 10 25 33
	0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80
Sundaram	FOR EDUCATIONAL USE

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.

**Solution:** 



	n=3194
	0/2=1597
	The state of the partition of the partit
	This observation he between the class intowal
	21-50 which is the median klass.
	lower dase limit = 21 = (1)
	class size (h) = 30
	1 4 11 mearan was (f) = 1500
	almulative terequency of class preceding the
	facquercy of the median class (f)=1500 Kumulative facquercy of class preceding the median class (cf)=950
./	mudian= 1+(1/2-cf) xh= 21+ (1597-950) x30 = 21+1294 = 33.94
	f 1500
,	11.10. 00 1
4 ,	Median = 33.94.
- 1	
ram	FOR EDUCATIONAL USE
The second secon	

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).

#### **Solution:**

	Consider of Guelidean of	, P2(s	2,0), F	3(3,1)	& P41	5, D.	
78	SOLO :- Asials	2	y				
	PI	0	2				
	P2	2	0				
	РЗ	3	- 1				
	P4	5	1				

	d(x,y)=(2)	z:- y;)	2)/2 =	216	x:-yi)2		
	الله على الله الله الله الله الله الله الله ال	10	12.1	1=17	2 - 5/1	$(2-0)^2$	
	Pd(P, P2)	=1(x'-	x2)+(	y,-y2)	- 411		
	d(p,p2)	= 14+	4 = ~	8 =	2.888	8. 200 COLUMN	
	-2 [	~3./	32-	10-20	3+12-12	= 19+1 = 110=	3.16
	d(RB)= 1(x)	-X3)+(C	1,-ya) -	710 3			
:	d (P, P4) - V(x	1-X4)2+1	(41-74)	= 1(0-5	)2+(2-1)2	= \[ \sqrt{25+1} = \sqrt{26} =	5.0
	1400) [/	- >2//	1 1172	10-2	1-10-D2=	1+1= (2=1-41	4.
:.	d(BB)=1(x)	-23)+(	12-ya) =	1(2 )			
	d(P2P4)=16x	2-74/2+	(y2-y4	)== [(	2-5)2+(0	-1)2 = 19+1= Tu=	3.16
						12= 22= 24	
	d(BP4)= VI	25-24)	+ (43-Yu	) -40	4)		
		PI	P2	P3	рч		
	PI PI	0	2.828	3-16	5.09		
	P2	2-808	0	1-414	3.16		
	ρ3	3.16	3.16	2	0		
	] pu	5.09	3.16				
	5+						
	4 -						
	3 + 2 + P	(0.2)					
	2+		6/3	1) ,6	4(5,1)		
		1 4	B(2,0)	1 1	,	100	
	ō	1 2	P2(2,0)	4 5	×		
am			1	1	ATIONAL U	SE	

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.

Soln:- Let A be attribute income  Given:- mina = \$12,000	
$\frac{m_{0XA} = $98,000}{V = $73600}$ $V = $73600$ $n_{0W-mnA} = 0.0$ $n_{0W-mnA} = 0.0$	
V' = V-minA maxA-minA (new-maxA-new minA) + new-minA	
= 73,600 - 12,000 98,000 - 12,000 (1.0-0.0) + 0.0 = 61600 86,000	
= 0.7163 : Income \$73,600 is toonsformed to 0.7163	
in the	
o lomer i	
FOR EDUCATIONAL USE	
L USE	

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.



Solution:

0	1º the way of the gots being at size 3 us for any - door
0	boutition the given data into bins of size 3 using equi-dep
	binning nuthor & periform elmoothing by bin mean, bin me and bin boundaries. Consider the data: 2,10,16,18,19,20,22,25,
	Oata!-2,10,18,18,19,20,22,25,28.
	BEN SIZE = 3
	of data is already souted in increasing order of wide the
	data Porto bins of size 3.
	Bin 2:- 18, 19, 20
DESCRIPTION OF PERSONS ASSESSMENT	8103!-22,25,28
	7.0
+	smoothing by bin mean.
Ť	8°n 1 :-10, 10, 10
	30,2:-19,19
	8in 3:- 85 25,25.
+ )	smoothing by bin median
	Bin 1:-10, 10, 10
	Bin 2:- 19, 19, 19
	3°n 3:- 25,25,25
	(1) -15 - 37 mil 200 300 100 100
+ 0	moothing by bin boundances.
F	18h 1:- 2,2,18
	8in 21-18,18,820
	Pn 3: 22, 22, 28
	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

