Name:	Pratik Sanja	ay Avha	ad				
Roll No:	01						
Class/Sem:	TE/V						
Experiment No.:	3						
Title:	Tutorial	on:	a)	Data	Exploration	b)	Data
	pre-proces	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, 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:**

	Air: To salve Roblems in Odo Exploration & Data Re-paressing
	Aim: To solve hoblems in Lord expension
F	Suppose that the data for analysis includes the attribute age. The ag
U	supples that the data typies are (in increasing order) 13,15, 16,16,19,20
	Soprese that the data for analysis including the attitude age. It was to have for the data toples out (in increasing and 1), 15, 15, 16, 19, 20, 21, 12, 12, 25, 25, 25, 25, 25, 33, 33, 35, 35, 35, 34, 45, 46, 52, 26, 21, 12, 12, 125, 15, 25, 25, 25, 26, 33, 33, 35, 35, 35, 34, 45, 46, 52, 26, 21, 12, 12, 12, 12, 12, 12, 12, 12, 12
	Property of the second
(	gurates the mean of data what is the median ?
	Here, N=77
$\rightarrow$	N 53+30+35+35+35+55+00+00+
	1 19 9C
	= 809 = 29.96
	^
	Me dian-> Middle value
	= 25,
	What is the mode of the data? Comment on the data's modality?
$\Theta$	What is the mode of the dard Commer
=	Mode=most occured values  In the above sequence 25 & 35 are most occured values
7	In the above Sequence 23 4 55 and more
	Made 25, 35 is bimodal -
6	what is the midsange of the data !
-	Midage Thin rake + Hax value /
	12
	- 13470/2
	,
	Midrange 41,5
	Can you find (roughly) the first quartile (QI) he the third quartile



## Vidyavardhini's College of Engineering and Technology

#### Department of Artificial Intelligence & Data Science

	the doto
=)	a : Median - 25
	J.
	9, : First Quartile = Middle volv of 13, 15, K, 16, 19, 20, 20, 21, 22, 27
	Λ
	g; 20 ,,
	9, = April Quartile= Hiddle value of 30,33,35,35,35,35,35,35,35,36,405,15
	0 1 - 36
. 6	Give the 5 number summary of the data -
Ę,	Missimum value = 13
	First Quatile=20
	Medfan Value 25
	Third Quartile = 35
	Horimum vol 122 70
A	Show a barplet of the dota
0	Jivo
	1- 25 35
	13 70
1	0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80

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
	1/2=1597
	The state of the s
	This observation le between the class intowal 21-50 which is the median class.
	21-50 which is the median class.
	lower dass limit = 21 = (1)
	class size (h) = 30
	facquercy of the median class (+)=1500
	Minulative frequency of class preceding the median class (cf) = 950.
2	median= 1+(1/6-cf) xh = 21+ (1597-950) X30 = 21+1294 = 33.91
	4edian = 33.94
ram <sup>®</sup>	FOR EDUCATIONAL USE

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:**

	Cuclidean of	, P2(s	2,0), F	3(3,1) &	P4(5, 1	).	
7	adol 1: - Asinh		y				
	PI	0	2				
	P2	2	0				
	Р3	3	- 1				
	P4	5	1				

d(z,	y)=(2)	(x;-y;)	2)/2 =	艺(	(1-yi)2	
	(0 00)	= T6x	x 3+1	y,-y2)	$2 = \sqrt{(0-2)^2 + (0-2)^2}$	2-0)2
					2.838	
110	0) [	11 72 LL.	11172-	(0-3)	+ (2-1)2 = 59+1	= 10=
i. de	13)= V(X	-X3) T(C	11-9a/	0.00		
:. d (P)	Ry)= J(2	1-x42+	(41-yus	= 1(0-5	$1^{2}+(2-1)^{2}=\sqrt{25+1}$	= 126 = 5
		.2.1	-2	[40.0]	7-(0-02= 11+1=	JZ=1-414
, 210	0.7-16	(x, )2+	· (u uu	3== 10	2-5)2+(0-1)2=10	1+1= [4=3
:. d(B	P4)= J	(x3-x4)2.	+(43-44	)2 = 1(3	$-5)^{2}+(1-1)^{2}=\sqrt{2^{2}}$	= 3/1.
		PI	P2	P3	рч	
	PI	0	2.828	3.16	5.09	
		-	0	1-414	3.16	
	Po	0-808				
		3.16	1-24/24	0	. 2	
	P2			0	0	
	P2 P3 P4	3.16	1-24124			
	P2 P3	3.16	1-24124			
	P2 P3 P4	3.16	1-24124			
	P2 P3 P4 Y A 5 + 4 + 7	3.16	1-24124			
	P2 P3 P4 Y A 5 + 4 + 7	3.16	3-16		0	
	P2 P3 P4 Y A 5 + 4 + 7	3.16	3-16	2	0	

4. Suppose that the minimum and maximum values for the attribute income are \$12,000 and

# A STANAROUM

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

\$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.

#### Solution:

6	troutition the given data into bins of size 3 using aqui-depth binning nuthed & postorn smoothing by bin mean, bin med
	and her boundance. Consider the data: 2,10,18,18,19,20,22,25,2
$\rightarrow$	Oata !- 2,10,18,18,19,20,22,25,28.
	Bin size = 3 As data is already sorted in increasing order divide the
	data into bins of size 3.
	Bin 1:- 2,10,18
	8°n 2:- 18,19,20 8°n 3:- 22,25,28
	9.0
4	Smoothing by bin mean.
	8°n 1 :-10, 10, 10 8°n 2 :-19, 19, 19
	Bin 3:- 26,25,26.
+	&moothing by bin median
	Bin 1:-10, 10, 10
	8in 2:- 19, 19, 19
	8h 3:- 85, 25, 25
+	dmoothing by bin boundaries.
	Bh 1:- 2,2,18
	Bin 2!-18,18,1820
	8 m 3 - 22, 22, 28
	The second secon
1000	

