Name:	Priyanka N	Bhand	ari				
Roll No:	02						
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:							

# 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, 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 solve Pooblems in Onta Exploration and "
Oata pour-pourcessing.
O scuppose that the data for analysis includes the attribute age. The age values for the data tuples
Lave ( & Shewashy order) 13, 15, 16, 16, 19, 20, 20, 21,
22, 22, 25, 25, 25, 25, 30, 33, 35, 33, 35, 35, 35,
36,40,45,46, 52,70.
What & the man of the date 2 will be it in 10 2
a) What is the mean of the data? What is the median? -
Mean = 2 = 13+15+16+16+19+20+20+21+22+22+25+25+25+
N 25+30+33+35+35+35+35+36+40+45+46+52+70_
27
.: Mean = 809 = 2 9.96
and the state of t
modern = most Reported Locaround value.
=> 25-, 35 = Bimedal.
M 10 ) = 110
Median → The middle value of the stequence.
median => a5
(b) what is the mode of the dale ?
(b) what is the mode of the data? comment on the data's modality (i.e. unimodal, bimedal, teimodal, de)
-> Rol n:-
mode = most recurred value.
In the & above sequence 25 & 35 are mostly
Occurred.
FOR EDUCATIONAL USE



(e)	ague the free-number summary of the data.
	Sol = 13
	Ffrst Quaestile = 20 Median value = 25
	Tried Quaetile = 35 Maximum Value = 70
(e)	show a bearplot of the data.
	$min = 13$ , $q_1 = 20$ , $q_2 = 35$ . $max = 70$ , $q_2 = 35$
	20 25 35
_	0 5 10 15 20 25 30 35 40 45 50 35 60 65 70 75

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
	n/2=1597
	This observation le between the class interval
	21-60 which is the median clay.
	louiser doss limit = 21 = (1)
	class size (h) = 30
	facquercy of the median class (+) = 1500
	midian slaw (cs)-050
./	median= 1+(1/2 -(F) xh = 21+ (1597-950) X30 - 21+1294= 33.9
9,	Median = 33.94
ram)	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:**

3	Cuclidean	the 10.1	data gr	ien belout	& compu	te the	
	Euclidean of	usta	nce betw	3(3, 1) & PI	1 point.		
->	2001 ° - Airle		4	3(3,1) 4 F	1(0, 1).		
	PI	0	2				
	P2	2	0				
	Р3	3	- 1				
	P4	5	1				
(merchan)	1 1 1		FOR	R EDUCATIONAL U	SF		

				-		
100	x,y)=(2(	x 4:)	2) 1/2 =	£ [6	(;-yi)2	
ac	(,9)-(2)	200	/	9=17	5 [	22 /2 D2
. 0	d(P, P2):	$=$ $\int (x, -$	x2)+(	y,-y2)	= 1 (0	$(2-0)^2$
	d(p,p2)					
i.d	(P, P3) = \((x,	-x3)2+(c	1,-ya)2=	10-3	+ (2-1)2	= 59+1 = 510= 3.
			. 0	100	12/12/12	- Josti = Ja6 = 5.
i.d	(P, P4)- V(x	1-X4)2+1	4-34	= 1(0.5	1 +(2-1) =	= \[ \sqrt{25+1} = \[ \sqrt{26} = 5. \]
- 1/	007-112	-x-)2+(c	10-407	1/2-3	7-(0-02=	11+1= 12=1-414.
·al	1213/-100	737	32 30			- F 2
i.dl	BPW= 16x	2-24/2+	(y2-y4)	$)^{2} = -1($	2-5)2+(0-	1)2 = \q+1=\u=3.
						2= \ 22 = 2/.
r.d(	BRUE VI	3-74)	H 43-44	) - 410	2) 10 /	
		PI	P2	P3	рч	
	PI	0	2.828	3.16	5.09	
	P2	2-828	0	1-414	3.16	
	Р3	3.16	1-414	0	0	
	P4	5.09	3.16	2	0	
	× 4					
	4 -					
	3 +					
	3+ 2+P1	(0,2)	6/3	1) 6	.(< 1)	
	1 -					Mary page 1
		1 2	B(2,0)	1 1 5	×	
3				OR EDUC	ATIONAL US	E

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.

#### Solution:



