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

Name:	Sunit Sunil Khaire
Roll No:	19
Class/Sem:	TE/V
Experiment No.:	3
Title:	Tutorial on: a) Data Exploration b) Data pre-
	processing
Date of	
Performance:	
Date of Submission:	
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, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
- What is the mean of the data? What is the median?
- What is the mode of the data? Comment on the data's modality (i.e., unimodal, bimodal, trimodal, etc.).
- What is the midrange of the data?
- Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
- Give the five-number summary of the data.
- 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.



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

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.

	Aim: To solve problems in Data Explosiation and Data Pre-processing.
	and Data Pre-mocessing.
\rightarrow	Deluppose that the deta for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20,
	The attribute age. The age value for the data
	tuples are (in invecsing 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.
	the state of the contract of the state of th
	a What is the man of the date I what is the
100 1	Made of the same o
. T	Amo: Mian (4) = $5x = 809$ = 2008
, a , a	= 29.96 ≈ 30
* **** * * * * * * * * * * * * * * * *	Median Criddle value of the ordered set, as the number of values in the set is odd) of the
M. A	number of values in the set is odd) of the
	data = 25.
	(b) What is the mode of the data? Comment on the data's modality (i.e., unimodal, bimodal, trimadal,
	data's modality (i.e., unimodal, bimodal, trumadal
	Ans: This data set has two values that occur
	with the same height friguency and is therefore.
	with the same height forequency and is, therefore, bimodal. The mody (values occurring with the
	greatest frequency) of the data and 25 and 35.
Sundaram)	FOR PRIVATE AND ADDRESS OF THE PRIVATE ADDRESS OF THE PRIVATE AND ADDRESS OF THE PRIVATE ADDRESS OF THE
	FOR EDUCATIONAL USE

	(E) What is the mideange of the data?
	Any: The midrange Cauviage of the largest and
	Ans: The midrange (aureage of the largest and smallyt values in the data set) of the data is:
- F-72	(70+13) = 41.5
S	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
	The state of the s
- ;	(d) Can you find (grouphly) the livest quantile (Q1) and
	d) Can you find (roughly) the first quantile (Q1) and the othered quantile (Q3) of the data? Ans: The 1st quartile curresponding to the 25th percentile of the data is: 200. The 3rd
	Anni The 1st accountile Conservamoding to the
	25th sourcestile of the data is: 200. The 3rd
	quartile (covies ponding to the 75th percentile) of
	the data is: 35.
	The aud is . O)
	E) live the five-number dummary of the data.
	Ans: The five number summary of a distribution
	consists of the minimum value, first quartile,
	median value, there quartile, and maximum value.
	Minimum: 13
	Firest Quartile (Q1): 20
	Median (Q2): 25
	Third Quartile (Q3): 35
	Maximum: 70
	1- warran , /
· · · · · · · · · · · ·	FOR EDUCATIONAL USE

4	(F) Show o	e boxinle	of the data	. 1198-1	
	Ansi-			: N	
January -	it was in	25 12 12	ni illi	the union in ?	
	- in		a With	TO THE STATE OF TH	
				The many sand	
2 Z	0 10 20 30 40 50 60 70 80 90				
2 ž					
_	Min: 13	, Q1:20	, Q2:25	, Q3: 35 , Max: 70	
		2			
				and the second	
>	2) Duppon	that v	alives for a	given set of data are	
	groupid	into int	wirds. The	intervals and	
	Lovespon	ding f	requercies à	given set of data are intervals and re as follows:	
	age fouquency				
		1-5	200	nihit	
		6-15	300		
				A. M.	
		21-30	1000		
		01-11	0 44		
	Anu-		01 74		
		age	I SO ALLA SA	1.12.1	
		1-5	- frequency	lumidative prequency	
		6-15	450	200	
		16-20	300	650	
		21 - 50	1500	950	
		51-80	700	2450	
		81-110	44	3 \ 5 O 3 \ 9 \	
<u>Sundaram</u> ®			FOR EDUCATIONA	L USE	
				그 그는 그는 이 이렇게 하는 어머니는 사람이 하나 이 그릇이다.	

This observation lie between the class interval 21-50 which is the median days. ouve das limit = 21 beguency of median class (f)=1500 lamulative frequency of class preceding the median class (cf)=1950 Median = 33.94 Consider the data given below and compute -Euclidean 0)2+(1-2)2]112 2+ (1-02-T1)2= -12= 1.414=d $(5-2)^2+(1-0)^2)^{1/2}=\sqrt{10}=3.162=d(P_4)$ FOR EDUCATIONAL USE Sundaram

			1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1			
	P	0	2.828	3.162	5.099	
	P2	2.828	0	/	2	
	Pa	3:162	1.414		3.162	
, -	Py	5.099	2	3.162	. 70	
=		P,	P	Pz	Py	
0:5)	Partition	the gir	un date	a into	bins of	81213
	using eg	jui- de	pth bis	nning r	nethod	and perform median data:
	- Como	othing.	by bir	med	in, bin	median'
	and bin	- Obc	sundari	y · Consi	der tri	data:
	2,10,18,	18, 19,	20,22,	25, 28.		·
010		0 1				
% n:-	Os data is abready souted in increasing orders,					
	divide the data Pinto bins of size 03.					
	Bin 1: 2, 10, 18 Bin 2: 18, 19, 20					
	15in 2.	3 3 3 3 3 3 3	, 20	· · · · · · · · · · · · · · · · · · ·	ilian China	
	Bin 3.	22, 25	, 28			
	· Smooth	٦ ا	- William			
			34 bin	mean:	0 110 5	- 12
	Muan Obinio = $(2+10+18)/3=10$					
	Man (Bin3) = $(22+25+28)$ 3 = 25					
	1901 (5110) - (21725428) 13-25					
	Riplacine	each	value	in the b	in with	010
	Riv 2: 19 19 19					
	Rin 3: 25 25 25					
			1			
<u>Sundaram</u> ®			FOR EDI	UCATIONAL USI	Ξ	

	· Amoothing by hin median I Replacing each
12 (a) 14 2 (a) 14 3 (a) 14	· Smoothing by bin median I Replacing each value Oin other bin with its median
	Bin 1: 10, 10, 10
	Bin 2: 19, 19, 19
	Bin 1: 10, 10, 10 Bin 2: 19, 19, 19 Bin 3: 25, 25, 25
- '	
	· Olmoothing by bin boundaries I Replacing each element by value it is closur to Cator.
. <u>Si</u>	element Obe Value it is closer to (stor:
	the last of 3 Bin 1: 2,2,18
	Bin 1 = 2,2,18
•	Bin 3: 22,22,28
	Bin 3: 22,22,28
(10	110. 110 + ++ ·1 + · · · · · · · · · · · · · · ·
<u> (V.4) .</u>	Sol": Let A be attribute income
	ljven: - min = \$ \$ 12,000
	$max_{8} = $98,000$ V = \$73600
	$nu_{min_{A}} = 0.0$, $nu_{max_{A}} = 1.0$
	VI = V - mina
	Many - ming (new maxy-new ming) + new ming
	=73,600-12000
	98000 - 12000 (1.0-0.0) + 0.0
	= 61600
	86000
• (1)	=0.1163
	:- Income \$13600 is transferred to
	0.7163.
8	FOR EDUCATIONAL USE
Sundaram	TOR EDUCATIONAL USE