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:	

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?
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.

Q1. 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

i. What is the mean of the data?

Mean = $\frac{\Sigma x_i}{n}$, n = no. of datapoints

x_i where $i = 0, 1, 2, 3, \ldots$

$\therefore \Sigma x_i = 13+15+16+16+19+20+20+21+22+22+25+25+25+25+30+33+33+35+35+35+35+36+40+45+46+52+70.$

$\therefore \Sigma x_i = 809$

$\therefore n = 27$

$\therefore$ Mean = $\frac{809}{27} = 29.96 \cong 30$

ii. What is the median?

Median is the middle value in a set of data.

$\therefore$ Median = 25

iii. What is the mode of the data? comment on data's modality.

Two values are repeated four times.

Mode = 25, 35

Modality = Bimodal.

iv. What is the midrange of the data?

Midrange = $\frac{\text{minimum value} + \text{maximum value}}{2} = \frac{13+70}{2}$

$\therefore$ Midrange = 41.5

v. Q_1 & Q_3 of the data.

Median ie. Q_2 is 25.

For finding Q_1, dataset will be:

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25

Q_1 = (lower of lower bound + higher of lower bound) / 2

$\therefore Q_1 = \frac{13 + 25}{2}$ $\therefore Q_1 = 19$ $Q_1 = 20$

For finding Q_2, dataset will be:

30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

Q_2 = (lower of higher bound + higher of upper bound) / 2

$\therefore Q_2 = \frac{30+70}{2}$ $\therefore Q_2 = 50$ $Q_3 = 35$

vi. Give the five-number summary of data:

a. Minimum value = 13

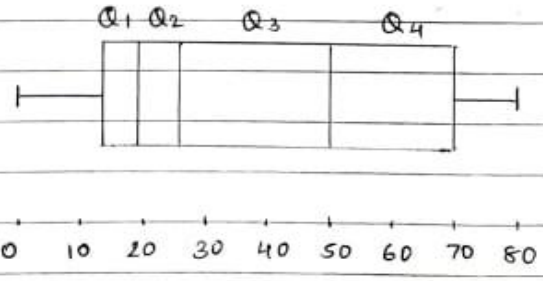
b. Q_1 = 20

c. Q_2 = 25

d. Q_3 = 35

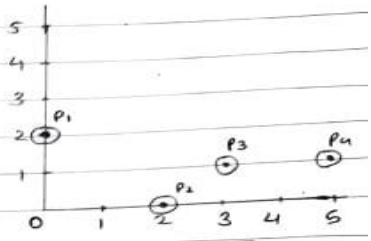
e. Maximum value = 70

vii. Boxplot.



Q3. Consider the data given below & compute the Euclidean distance btw.

$P_1(0, 2)$
$P_2(2, 0)$
$P_3(3, 1)$
$P_4(5, 1)$



Formula: $d(x, y) = \sum_{i=1}^{n} \sqrt{(x_i - y_i)^2}$

$d(P_1, P_2) = \sqrt{(0-2)^2 + (2-0)^2} = 2.82$
$d(P_1, P_3) = \sqrt{(0-3)^2 + (2-1)^2} = 3.16$
$d(P_1, P_4) = \sqrt{(0-5)^2 + (2-1)^2} = 5.09$

$d(P_2, P_1) = \sqrt{(2-0)^2 + (0-2)^2} = 2.82$
$d(P_2, P_3) = \sqrt{(2-3)^2 + (0-1)^2} = 1.41$
$d(P_2, P_4) = \sqrt{(2-5)^2 + (0-1)^2} = 3.16$

$d(P_3, P_1) = \sqrt{(3-0)^2 + (1-2)^2} = 3.16$
$d(P_3, P_2) = \sqrt{(3-2)^2 + (1-0)^2} = 1.41$
$d(P_3, P_4) = \sqrt{(3-5)^2 + (1-1)^2} = 2$

$d(P_4, P_1) = \sqrt{(5-0)^2 + (1-2)^2} = 5.09$
$d(P_4, P_2) = \sqrt{(5-2)^2 + (1-0)^2} = 3.16$
$d(P_4, P_3) = \sqrt{(5-3)^2 + (1-1)^2} = 2$

DP	x	y
P_1	0	2
P_2	2	0
P_3	3	1
P_4	5	1

	P_1	P_2	P_3	P_4
P_1	0	2.82	3.16	5.09
P_2	2.82	0	1.41	3.16
P_3	3.16	1.41	0	2
P_4	5.09	3.16	2	0

Q2. Suppose that the values for a given set of data are grouped into intervals.

	Age	Frequency	cummulative frequency
a.	1-5	200	200
b.	6-15	450	650
c.	16-20	300	950
d.	21-50	1500	2450
e.	51-80	700	3150
f.	81-110	44	3194

cummulative frequency for:

a. 0+200 = 200 ; b = 200+450 = 650

c = 650+300 = 950 ; d = 950+1500 = 2450

e = 2450+700 = 3150 ; e = 3150+44 = 3194

The formula to calculate the median within the median class:

$$\text{Median} = L + \left(\frac{N/2 - CF}{f}\right) \times w$$

L = lower boundary = 21

N = total observation = 3194

CF = cumulative frequency = 950

f = frequency = 1500

w = width of median class = 29

$$\therefore \text{Median} = 21 + \left(\frac{3194/2 - 950}{1500}\right) \times 29$$

$$\therefore \text{Median} = 33.50$$

Q4 Minimum value : $ 12000 = $\min_A$

Maximum value : $ 98000 = $\max_A$

Range = [0.0 , 1.0]

V = $ 73600

Min-Max normalization formulae

$$V' = \frac{V - \min_A}{\max_A - \min_A (new\max_A - new\min_A) + new\min_A}$$

$$\therefore V' = \frac{73600 - 12000}{98000 - 12000(1-0)+1} = 0.7163$$

∴ Income $ 73600 is transformed to 0.7165

Q5 Data: 2, 10, 18, 18, 19, 20, 22, 25, 28

n = 3

Three equi-depth bins of size 3

Bin1 = 2, 10, 18

Bin 2 : 18, 19, 20

Bin 3 = 22, 25, 28

Smoothing by bin means:

Bin 1 : 10, 10, 10

Bin 2 : 19, 19, 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