Name:	Vinith Shetty
Roll No:	52
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:	

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.

Aaryan C. Gole
TE - AI/DS (10)

DWM

Experiment - 3

Aim: To solve problems in Data Exploration and Data Pre-processing.

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?

→ $\text{Mean}(\mu) = \frac{\Sigma x}{n} = \frac{[\text{Sum of values}]}{\text{Total no. of values}} = \frac{809}{27} = 29.96$

Median = 25 [∵ 27 is odd so the centre-most value]

b. Mode of the data? Comment on the data's modality (i.e., unimodal, bimodal, trimodal, etc.)

→ Mode = 25 and 35 (repeated 4 times)
The data is of bimodal modality.

c. What is the mid-range of the data?

→ $\text{Mid-range} = \frac{\text{Minimum Value} + \text{Maximum Value}}{2}$

$= \frac{13+70}{2} = \frac{83}{2} = 41.5$ //.

d. Can you find (roughly) the first quartile (Q_1) and the third quartile (Q_3) of the data?

→ Since there are 27 values in the data
$\therefore \frac{27}{4} \approx 7$ (approx.) and from median 25;

Hence Q_1 (roughly) = 7th element = 20
and Q_3 (roughly) = 21st element = 35
(median of 13-25 range; median of 25-70 range)

e. Give the five-number summary of the data.

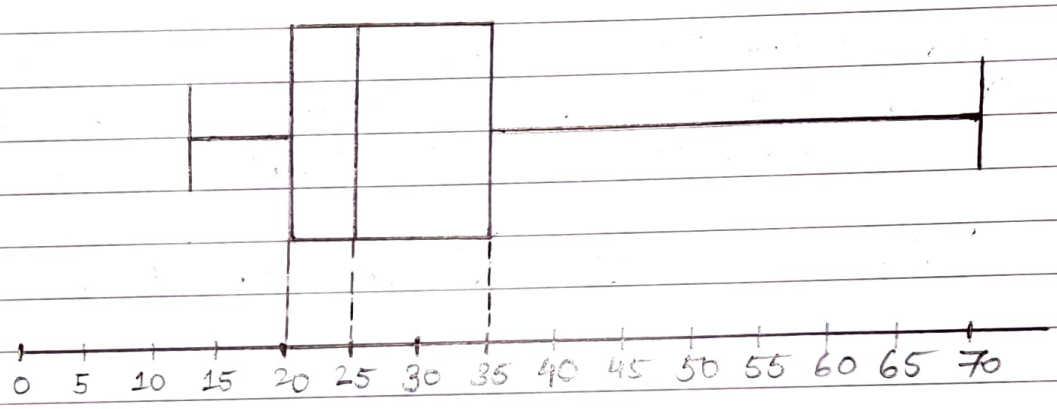
→ The five-number summary of a distribution consists of the :-
(i) min. value ; (ii) first quartile ; (iii) median; (iv) third quartile ; (v) maximum value.

∴ min. value = 13
first quartile = 20
median = 25
third quartile = 35
max. value = 70

Hence, five-number summary of the data is 13, 20, 25, 35, 70.

f. Show a boxplot of the data.

→ BOX-PLOT



Min: 13, Q_1: 20, Q_2: 25; Q_3: 35; max: 70

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

→ The frequency of class intervals between 1 to 5 is 200.
And now for 6–15 is 450 + 200 = 650
Likewise for 16–20 is 300 + 650 = 950
for 21–50 = 950 + 1500 = 2450
for 51–80 = 2450 + 700 = 3150
for 81–110 = 3150 + 44 = 3194

∴ cumulative frequency table

age	c.f.
1–5	200
6–15	650
16–20	950
21–50	2450
51–80	3150
81–110	3194

$N/2 = \frac{3194}{2} = 1597$
which lies in range 21–50

N = 3194
L = 21 [lower limit of median class]
F = cf of prev class of median class
= 950
f = freq. of median class = 1500
h = class width = 50 − 21 + 1 = 30

$\therefore$ Median $= L + \left(\frac{\frac{N}{2} - F}{f}\right) \times h$

$\therefore$ $\frac{3194}{21} + \left(\frac{1597 - 950}{50}\right) \times 20$

$\therefore$ $\frac{3194}{21} + \frac{647}{50}$

$\therefore$ Median $= \frac{3194}{21} + 12.94$

$= 3206.94$ $\quad 33.94$ //.

4. Minimum value $= \$12000 = min_A$

Maximum value $= \$98000 = max_A$

Range $= [0.0, 1.0]$

$v = \$73600$

MIN-MAX NORMALIZATION FORMULA

$\Rightarrow v' = \frac{v - min_A}{max_A - min_A (newmax_A - newmin_A) + newmin_A}$

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

$\therefore$ Income \$73600 is transformed to 0.7163

5. Data: 2, 10, 18, 18, 19, 20, 22, 25, 28.
$n = 3$

Three equi-depth bins of size 3 [Data elements = 9] are :-

Bin 1 : 2, 10, 18
Bin 2 : 18, 19, 20
Bin 3 : 22, 25, 28

- Smoothing by bin means is :-

Bin 1 : 10, 10, 10
Bin 2 : 19, 19, 19
Bin 3 : 25, 25, 25

- Smoothing by bin boundaries is :-

Bin 1 : 2, 18, 18
Bin 2 : 18, 20, 20
Bin 3 : 22, 28, 28

- Smoothing by bin median is :-

Bin 1 : 10, 10, 10
Bin 2 : 19, 19, 19
Bin 3 : 25, 25, 25

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

By computing the Euclidean distance between each point is :-

Euclidean distance $= [(x_2-x_1)^2 + (y_2-y_1)^2]^{1/2}$

$d(P_1P_2) = [(2-0)^2+(0-2)^2]^{1/2} = \sqrt{8} = 2.828 = d(P_2P_1)$

$d(P_1P_3) = [(3-0)^2+(1-2)^2]^{1/2} = \sqrt{10} = 3.162 = d(P_3P_1)$

$d(P_1P_4) = [(5-0)^2+(1-2)^2]^{1/2} = \sqrt{26} = 5.099 = d(P_4P_1)$

$d(P_2P_3) = [(3-2)^2+(1-0)^2]^{1/2} = \sqrt{2} = 1.414 = d(P_3P_2)$

$d(P_2P_4) = [(5-2)^2+(1-0)^2]^{1/2} = \sqrt{10} = 3.162 = d(P_4P_2)$

$d(P_3P_4) = [(5-3)^2+(1-1)^2]^{1/2} = 2 = d(P_4P_3)$

P_1	0	2.828	3.162	5.099
P_2	2.828	0	1.414	2
P_3	3.162	1.414	0	3.162
P_4	5.099	2	3.162	0
	P_1	P_2	P_3	P_4