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:	

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.

- 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.

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 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?

Ans:- Mean $(\mu) = \frac{\Sigma x}{N} = \frac{809}{27}$ ~~= 29.96~~

$= 29.96 \approx 30$

Median (middle value of the ordered set, as the 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, trimodal, etc.).

Ans:- This data set has two values that occur with the same height frequency and is, therefore, bimodal. The modes (values occuring with the greatest frequency) of the data are 25 and 35.

(c) What is the midrange of the data?

Ans:- The midrange (average of the largest and smallest values in the data set) of the data is:

$\frac{(70+13)}{2} = 41.5.$

(d) Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

Ans:- The 1st quartile (corresponding to the 25th percentile) of the data is : 20. The 3rd quartile (corresponding to the 75th percentile) of the data is : 35.

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

Ans:- The five number summary of a distribution consists of the minimum value, first quartile, median value, third quartile, and maximum value.

Minimum: 13

First Quartile (Q_1): 20

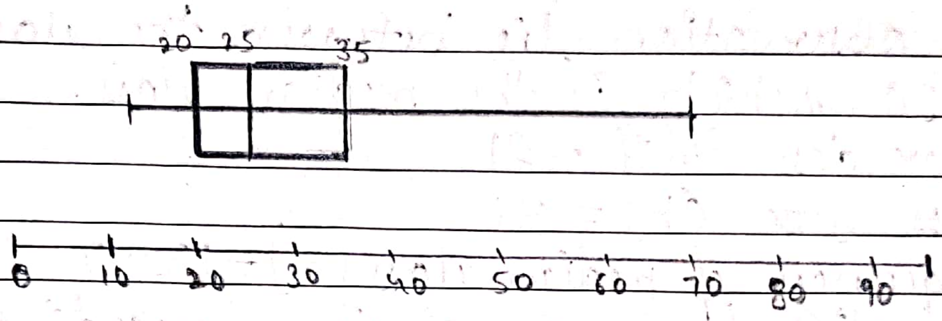
Median (Q_2) : 25

Third Quartile (Q_3) : 35

Maximum : 70

(f) Show a boxplot of the data.

Ans:-



Min: 13, Q_1: 20, Q_2: 25, Q_3: 35, Max: 70

(2) Suppose that 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

Ans:-

age	frequency	cumulative frequency
1-5	200	200
6-15	450	650
16-20	300	950
21-50	1500	2450
51-80	700	3150
81-110	44	3194

$n = 3194$

$n/2 = 1597$

This observation lie between the class interval 21-50 which is the median class.

lower class limit = 21

class size (h) = 30

frequency of median class (f) = 1500

Cumulative frequency of class preceding the median class (c.f) = 950

$$\text{Median} = l + \frac{(n/2 - c.f)}{f} \times h$$

$$= 21 + \left(\frac{1597 - 950}{1500}\right) \times 30$$

$$= 21 + 12.94$$

$\therefore$ Median $= 33.94$

Q.3) Consider the data given below and compute the Euclidean distance between each point.

$P1\ (0,2)$, $P2\ (2,0)$, $P3\ (3,1)$ and $P4\ (5,1)$.

Soln:- $d(P_1, P_2) = [(2-0)^2 + (0-2)^2]^{1/2} = \sqrt{8} = 2.828 = d(P_2 P_1)$

$d(P_1 P_3) = [(3-0)^2 + (1-2)^2]^{1/2} = \sqrt{10} = 3.162 = d(P_3 P_1)$

$d(P_1 P_4) = [(5-0)^2 + (1-2)^2]^{1/2} = \sqrt{26} = 5.099 = d(P_4 P_1)$

$d(P_2 P_3) = [(3-2)^2 + (1-0)^2]^{1/2} = \sqrt{2} = 1.414 = d(P_3 P_2)$

$d(P_2 P_4) = [(5-2)^2 + (1-0)^2]^{1/2} = \sqrt{10} = 3.162 = d(P_4 P_2)$

$d(P_3 P_4) = [(5-3)^2 + (1-1)^2]^{1/2} = 2 = d(P_4 P_3)$

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

Q.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.

Soln:- As data is already sorted in increasing orders, divide the data into bins of size 3.

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

- Smoothing by bin mean:

Mean (Bin1) = (2 + 10 + 18)/3 = 10
Mean (Bin2) = (18 + 19 + 20)/3 = 19
Mean (Bin3) = (22 + 25 + 28)/3 = 25

Replacing each value in the bin with its mean:
Bin 1: 10, 10, 10
Bin 2: 19, 19, 19
Bin 3: 25, 25, 25

- Smoothing by bin median [Replacing each value in the bin with its median]

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

- Smoothing by bin boundaries [Replacing each element by value it is closer to (1st or the last)]

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

Q.4) Solⁿ:- Let A be attribute income

given:- $min_A = \$ 12,000$

$max_B = \$ 98,000$

$V = \$73600$

$new\_min_A = 0.0$, $new\_max_A = 1.0$

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

$$= \frac{73,600 - 12000}{98000 - 12000 (1.0 - 0.0) + 0.0}$$

$$= \frac{61600}{86000}$$

$$= 0.7163$$

∴ Income ~~$73600~~ $13600 is transferred to 0.7163.