Aim: To solve problems on 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.
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.

Experiment-3

Aim: To solve problems on Data Exploration and Data Pre-processing

Theory:

1) Suppose that the data for analysis includes the attributes 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) mean. $(\mu) = \frac{\Sigma x}{n}$

$$= \frac{13+15+16+16+19+20+20+21+22+22+25+25+25+25+30+33+33+35+35+35+35+36+40+45+46+52+70}{27}$$

$$= \frac{784}{27}$$

$$= 29.037$$

median = [redacted]

b) mode = 25, 35

∴ It is bimodal.

c) midrange = $\frac{13+70}{2} = 41.5$

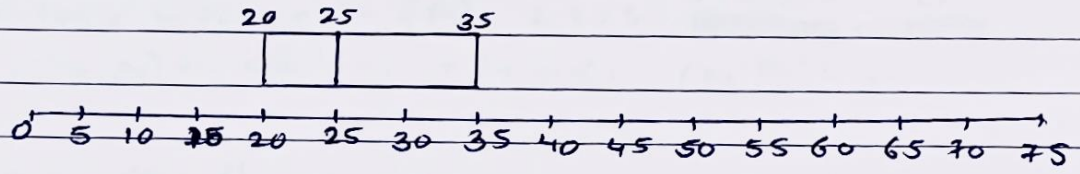
d) Q_1 = 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25

$Q_1 = 20$

Q_2 = 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

$Q_2 = 35$

e) Box Plot



2)

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

$\frac{N}{2} = \frac{3194}{2} = 1597$, $l_1 = 21$, $l_2 = 50$

$$\text{median} = l_1 + \frac{(l_2 - l_1)\left(\frac{N}{2} - C.F\right)}{f} = 21 + \frac{(50-21)(1597-950)}{1500}$$

median = 33.508

3) consider the given below and compute euclidean distance between each point: $P_1(0,2)$, $P_2(2,0)$, $P_3(3,1)$ & $P_4(5,1)$.

→

$$d = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$

$d(P_1,P_2) = \sqrt{(2-0)^2+(0-2)^2} = \sqrt{8} = 2.828$

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

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

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

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

$d(P_3,P_4) = \sqrt{(5-3)^2+(1-1)^2} = \sqrt{4} = 2$

∴ Distance between $(P_1 \& P_2) = 2.828$, $(P_1 \& P_3) = 3.162$, $(P_1 \& P_4) = 5.0$

$(P_2,P_3) = 1.414$, $(P_2,P_4) = 3.162$, $(P_3,P_4) = 2$

4) Suppose that the maximum and minimum value for the attribut income are \$12000 and \$95000 respectively. Normalize income value \$73600 to range (0.0, 1.0) using min-max normalization method.

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

where,

$v = \$73600$, $min_A = \$12000$, $max_A = \$95000$

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

$$\therefore v' = \frac{73600 - 12000}{95000 - 12000}(1.0 - 0.0) + 1.0$$

$$v' = 0.716$$

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.

→

2, 10, 18, 18, 19, 20, 22, 25, 28
bin size = 3
bin = 3

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

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

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

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