Name:	Pranita Kumbhar
Roll No:	70
Class/Sem:	TE/V
Experiment No.:	3
Title:	Tutorial on: a) Data Exploration b) Data pre-processing
Date of Performance:	07/8/25
Date of Submission:	14/8/25
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.

Solution:

1) Descriptive Statistics for Age Data

Data (27 values, sorted): 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

Calculation for Mean:

Sum of all values = 809
Number of values = 27
Mean = Sum / Number = 809 / 27 = 29.9630

Calculation for Median:

Since there are 27 values (odd count), median is the middle value.
Position = (n+1)/2 = (27+1)/2 = 14th value in sorted list = 25.0

Calculation for Mode:

Count frequency of each value. The most frequent values are modes.
Frequencies show [25, 35] occur the most times → Modality: bimodal

Calculation for Midrange:

Midrange = (Min + Max) / 2 = (13 + 70) / 2 = 41.5

Calculation for Quartiles:

Q1 ≈ 20.5, Q3 ≈ 35.0 using percentile position method.

Five-number summary: min=13,

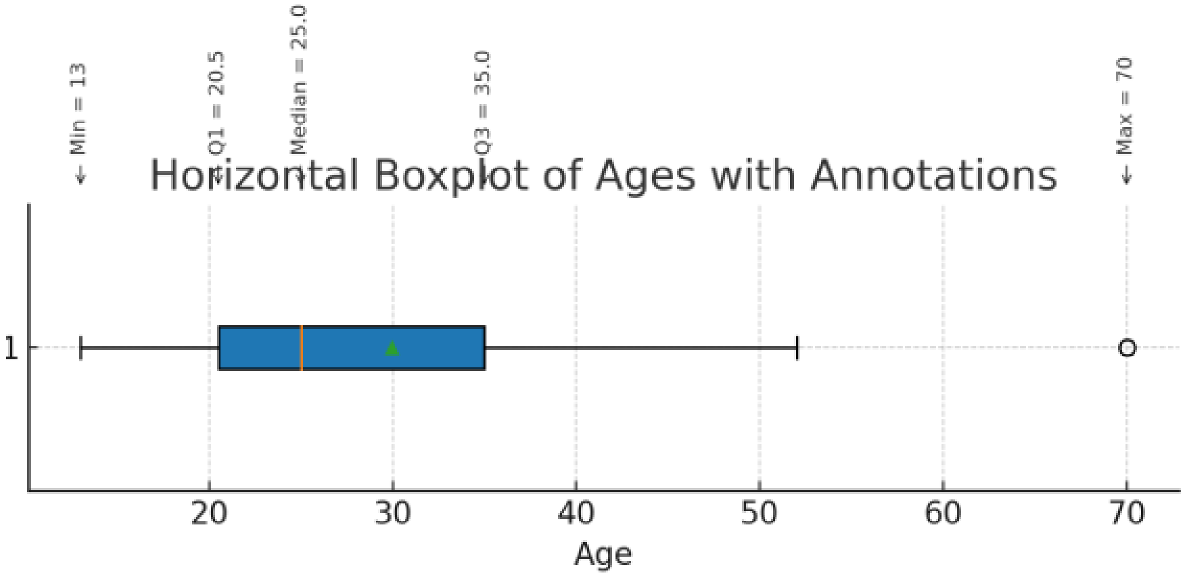
Q1=20.5000,

median=25.0,

Q3=35.0000,

max=70

Horizontal boxplot with annotated values:



2) Approximate Median from Grouped Data

Age Interval	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

Total N = 3194, median position = N/2 = 1597.0.

Median class = 81–110 (where cumulative frequency first exceeds N/2).

Using the grouped-data median formula: Median ≈ L + ((N/2 − cf) / f) × w, with continuous lower boundary L = class lower − 0.5, cumulative frequency before class cf, class frequency f, and width w.

Computed approximate median ≈ 33.4400

3) Euclidean Distances Between Points

Points: P1(0,2), P2(2,0), P3(3,1), P4(5,1)

Pair	Distance formula	Value
P1–P2	√((x1−x2)^2 + (y1−y2)^2)	2.8284
P1–P3	√((x1−x2)^2 + (y1−y2)^2)	3.1623
P1–P4	√((x1−x2)^2 + (y1−y2)^2)	5.0990
P2–P3	√((x1−x2)^2 + (y1−y2)^2)	1.4142
P2–P4	√((x1−x2)^2 + (y1−y2)^2)	3.1623
P3–P4	√((x1−x2)^2 + (y1−y2)^2)	2.0000

4) Min–Max Normalization

Min = 12000, Max = 98000, Value = 73600

Normalized value = (Value − Min) / (Max − Min)

Result: 0.716279

5) Equi-Depth Binning (bin size = 3) & Smoothing

Original data (sorted): [2, 10, 18, 18, 19, 20, 22, 25, 28]

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

Smoothing by bin mean:

10.0000, 10.0000, 10.0000, 19.0000, 19.0000, 19.0000, 25.0000, 25.0000, 25.0000

Smoothing by bin median:

10.0000, 10.0000, 10.0000, 19.0000, 19.0000, 19.0000, 25.0000, 25.0000, 25.0000

Smoothing by bin boundaries (replace each value by nearest boundary within its bin):

2, 2, 18, 18, 18, 20, 22, 22, 28