Vidyavardhini's College of Engineering and Technology

Department of Artificial Intelligence & Data Science

AY: 2025-26

Class:	TE	**Semester:**	V
Course Code:	CSC504	**Course Name:**	Data WareHousing & mining

Name of Student:	Pranita kumbhar
Roll No. :	70
Assignment No.:	02
Title of Assignment:	Introduction to Data mining, Data Exploration & Data Preprocessing.
Date of Submission:	
Date of Correction:	

Evaluation

Performance Indicator	Max. Marks	Marks Obtained
Completeness	5	5
Demonstrated Knowledge	3	3
Legibility	2	2
Total	10	10

Performance Indicator	Exceed Expectations (EE)	Meet Expectations (ME)	Below Expectations (BE)
Completeness	5	3-4	1-2
Demonstrated Knowledge Legibility	3	2	1
Legibility	2	1	0

Checked by

Name of Faculty : Ms. Neha Raut .

Signature :

Date :

Assignment 2.

Q.1] If a dataset is normally distributed, why are Mean, Median, and Mode approximately equal? What does this imply in data analysis?

→

A normal distribution is a symmetric, bell-shaped curve with one clear peak (unimodal). Because the left and right sides mirror each other, its 'center' is uniquely defined. In such symmetric unimodal data:

- Mode lies at the peak.
- Median splits the area.
- Mean is the balance point where positive & negative deviations cancel out.

Hence Mean ≈ Median ≈ Mode.

* Its implication in data analysis:

1] If mean ≈ median ≈ mode:

The data is symmetric

- There is no strong skewness.
- Outliers are not significantly influencing dataset.
- In this case, mean is reliable measure of central tendency.

2] If Mean ≠ Median ≠ Mode

The data is skewed.

- The mean is distorted by extreme values.
- The median is better choice to represent 'typical' value of dataset.

- The mode may be useful if the analysis is about the most frequent observation.

3] Practical implication :

- When analyzing data, always compare mean, median & mode.
- Their equality suggests normal distribution and allows you to apply many statistical models.
- If they differ widely, the dataset is skewed or contains outliers, so we should either transform the data, use robust statistics.

Q.2] A retail data warehouse stores the daily sales amount (in ₹) of 12 transactions as :
1500, 1800, 1700, 1600, 2000, 1550, 4000, 1700, 1800, 1900, 1700, 1600.

i] Calculate mean, median, mode & midrange.
ii] Draw boxplot.

→

i] a] mean : $\frac{\sum x}{n}$

$$\frac{1500 + 1800 + 1700 + 1600 + 2000 + 1550 + 4000 + 1700 + 1800 + 1900 + 1700 + 1600}{12}$$

$$\text{mean} = \frac{22850}{12} = 1904.17.$$

b] Median :

Sorted : 1500, 1550, 1600, 1600, 1700, 1700, 1700,
1800, 1800, 1900, 2000, 4000.

n = 12.

median = avg of 6th & 7th value.

$$= \frac{1700 + 1700}{2}$$

median = 1700

c] Mode.

Most frequent value = 1700 (appears 3 times)
so it is unimodal.

d] Midrange :

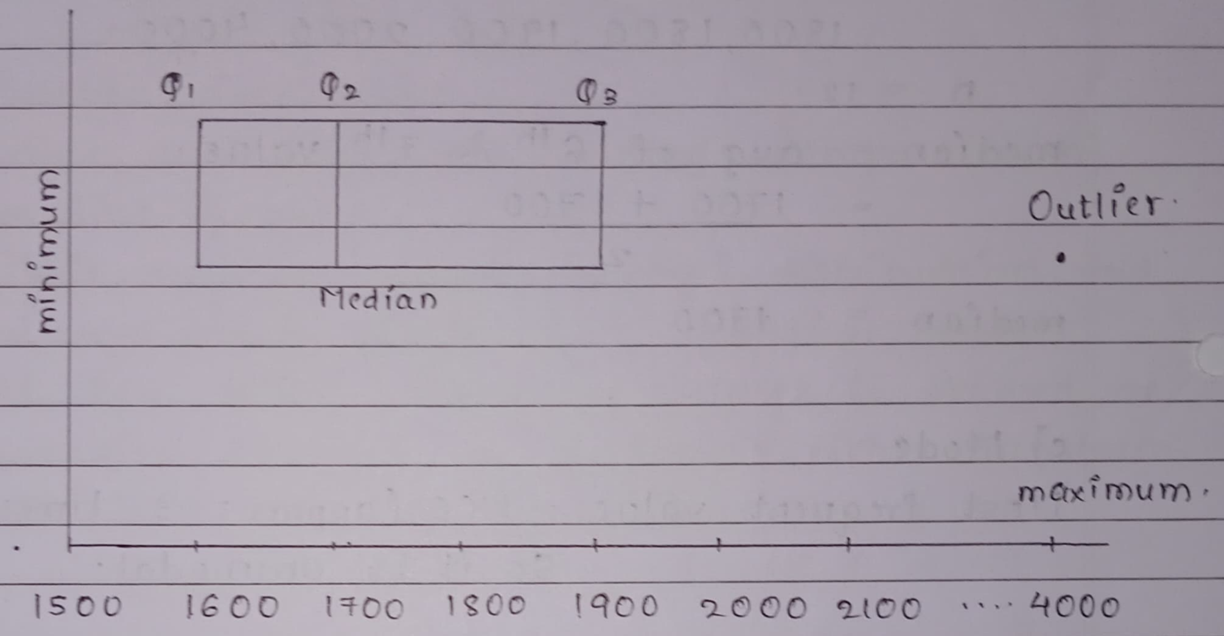
$$= \frac{\text{min} + \text{max}}{2}$$

$$= \frac{1500 + 4000}{2}$$

midrange = 2750.

Q1 = 1600
Q2 = 1700
Q3 = 1900.

ii] Boxplot :



Q1
Q2
Q3
Outlier.
minimum
Median
maximum.
1500
1600
1700
1800
1900
2000
2100
.... 4000