generally used for attribute selection.

Dimensionality Reduction

- The number of input features, variables, or columns present in a given dataset is known as dimensionality, and the process to reduce these features is called dimensionality reduction.

- It is a way of converting the higher dimensions dataset into lower dimension dataset ensuring that it provides similar information.

Problems

Measures of Central Tendency

- Mean
- Median
- Mode
- Mid Range
- Range

→ Mean :- $(\bar{x})$

$$\text{Mean} = \frac{\text{Sum of Observations}}{\text{No of observations}}$$

→ Median :-

It means to identify the middle value from a set of observations

i When n is odd :-

$\frac{(n+1)}{2}$ th position

ii When n is even :-

~~$\left(\frac{n}{2}+1\right)$ th~~

$$\frac{\left(\frac{n}{2}\right)\text{th observation} + \left(\frac{n}{2}+1\right)\text{th observation}}{2}$$

→ Mode :-

Mode is the most frequently occurring value in the dataset

a Find the mean of the first 10 odd integers

Obs → 1, 3, 5, 7, 9, 11, 13, 15, 17, 19

$$\text{Mean} = \frac{\text{Sum}}{10} = \frac{100}{10} = 10$$

→ Mid Range :-

It is the sum of maximum and minimum obs / 2 in the data

i.e Mid Range = $\frac{max + min}{2}$

→ Range :-

It is the difference between the maximum and minimum value of observation in the data

i.e Range : $\frac{max - min}{}$

→ Variance : (σ^2)

$$\sigma^2 = \sum_{i=0}^{n} \frac{(x_i - \bar{x})^2}{n}$$

It is the measure of distance of observed values from the mean

→ Standard Deviation : (σ)

$$\sigma = \sqrt{\frac{\sum_0^n (x_i - \bar{x})^2}{n}}$$

Q What is the median of the following data set

32, 6, 21, 10, 8, 11, 12, 36, 17, 16, 15, 18, 40, 24, 21, 23, 24, 24, 29, 16, 32, 31, 10, 30, 35, 32, 18, 39, 12,

Find the mid range and range too

Steps: Sort the data set (ascending)

After Sorting:-

6, 8, 10, 10, 11, 12, 12, 15, 16, 16, 17, 18, 18, 20, 21, 21, 23, 24, 24, 24, 29, 30, 31, 32, 32, 32, 35, 36, 39, 40

Median: $\frac{\left(\frac{n}{2}\text{th} + \frac{n}{2}+1\right)}{2}$

$\frac{15\text{th} + 16\text{th}}{2} = \frac{21+21}{2} = 21$

Mid Range: $\frac{40+6}{2} = 23$ $\left(\frac{\max + \min}{2}\right)$

Range: $\frac{40-6}{2} = 34$ $\left(\frac{\max - \min}{2}\right)$

a Identify mode

21, 19, 62, 21, 66, 28, 66, 48, 79, 59, 28, 62, 63, 63, 48, 66, 59, 66, 94, 79, 19, 94

Mode = 66 (repeated 4 times)
Modality - multimodal (66, 94) → multiple values are repeating.

Q A dataset for analysis included only 1 attribute x

x: {7, 12, 5, 8, 5, 9, 13, 12, 19, 7, 12, 12, 13, 3, 4, 5, 13, 8, 7, 6}

a Mean?
b Median?
c Standard Deviation

Sort the data

Xr {3, 4, 5, 5, 5, 6, 7, 7, 7, 8, 8, 9, 12, 12, 12, 12, 13, 13, 13, 19}

$$\text{Mean} = \frac{\text{Sum}}{\text{No of obs}} = \frac{180}{20} = 9$$

b Median

$$\frac{\left(\frac{n}{2}\text{th}\right) + \left(\frac{n}{2}+1\right)\text{th}}{2}$$

$$\frac{10\text{th} + 11\text{th}}{2} = \frac{5+5}{2} = 8$$

c Standard Deviation:-

$\sigma^2 = 3$

Modality of Data:-

→ Datasets with only one mode is known as unimodal
→ Datasets with two mode are known as bimodal
→ Datasets with > 2 mode are known as multimodal

Eg:- A: {1, 2, 3, 3, 4, 4, 5, 6}
Mode |A|: {3, 4}

A has two mode ∴ it is bimodal

a Calculate range, variance and standard deviation
-4, -2, 0, -2, 6, 4, 6, 0, -6, 4

Sorting

$-6, -4, -2, -2, 0, 0, 4, 4, 6, 6$

a) Range: $\frac{\max - \min}{}$

$= \frac{6 - (-6)}{} = 12$

b) Variance $= \frac{\sum_{i=0}^{n} (x_i - \bar{x})^2}{n}$

Mean $= \frac{\text{Sum}}{\text{No of obs}}$

$= \frac{-6-4-2-2+4+4+6+6}{10}$

$\frac{6}{10} = 0.6$

$\sigma^2 = (-6-0.6)^2 + (-4-0.6)^2 + (-2-0.6)^2 + (-2-0.6)^2 + (0-0.6)^2 + (0-0.6)^2 + (4-0.6)^2 + (4-0.6)^2 + (6-0.6)^2 + (6-0.6)^2$

10

$= 17.82$

c) $\sigma = \sqrt{\sigma^2} = \sqrt{17.82} = 4.22$

a Calculate range, variance and standard deviation

-3 -3 -3 -3 0 3 3 3 3

Range = $\frac{\max - \min}{2}$ ∴ $\frac{3-(-3)}{2}$: $\frac{6}{2}$: 3

Variance : $\sum_{i=0}^{n} \left(\frac{x_i - \bar{x}}{n}\right)^2$

mean = $\frac{\text{sum of obs}}{\text{no of obs}}$ = $\frac{0}{9}$ = 0

$\sigma^2 = \frac{(-3-0)^2 + (-3-0)^2 + (-3-0)^2 + (-3-0^2) + (0-0^2) + (3-0)^2 + (3-0)^2 + (3-0)^2 + (3-0)^2}{9}$

$= \frac{72}{9} = 8$

$\sigma = \sqrt{\sigma}$

$= \sqrt{8}$

Quartiles

Quartiles are the values from the dataset which divide the dataset into four equal parts where each part of the dataset contains an equal no of observations

aka. 1st quartile

Lower Quartile (Q_1) = $\frac{(n+1)}{4}$ th term

Median (Q_2) = $\frac{2(n+1)}{4}$ th term

Upper Quartile (Q_3) = $\frac{3(n+1)}{4}$ th term

aka 3rd quartile.

Inter-quartile range :- (IQR)

It is the difference between Q_3 and Q_1

$$IQR = Q_3 - Q_1$$

a Find the quartiles of the following age:

23, 13, 37, 16, 26, 35, 26, 35

$Q_1 = \left(\frac{n+1}{4}\right)$ th term

$n = 8$

$Q_1 = \frac{9}{4} = 2.25$ th term

Step 1 After sorting

13 16 23 26 26 35 35 37

We can rewrite Q_1 as

Q_1 = 2nd term + 0.25 (3rd term − 2nd term)

= 16 + 0.25 (23 − 16)

= 16 + 0.25 × 7

= 16 + 1.75

= 17.75

$Q_2 = \left(\frac{2(n+1)}{4}\right)$th term

$= \frac{2 \times 9}{4}$

$= \frac{18}{4} = 4.5$

Rewriting

Q_2 = 4th term + 0.5 (5th term − 4th term)

= 26 + 0.5 (26 − 26)

= 26

$Q_3 = \left(\frac{3(n+1)}{4}\right)$th term

$= \frac{3 \times 9}{4} = \frac{27}{4} = 6.75$

We can rewrite Q_1 as

Q_1 = 2nd term + 0.25 (3rd term - 2nd term)
= 16 + 0.25 (23 - 16)
= 16 + 0.25 × 7
= 16 + 1.75
= 17.75

$Q_2 = \left(\frac{2(n+1)}{4}\right)$th term

$= \frac{2 \times 9}{4}$

$= \frac{18}{4} = 4.5$

Rewriting

Q_2 = 4th term + 0.5 (5th term - 4th term)
= 26 + 0.5 (26 - 26)
= 26

$Q_3 = \left(\frac{3(n+1)}{4}\right)$th tm

$= \frac{3 \times 9}{4} = \frac{27}{4} = 6.75$

Rewriting

Q_3 = 6th term + 0.75 (7th term - 6th term)
= 35 + 0.75 (35 - 35)
= 35

a Find Quartile of the given data
10, 30, 5, 12, 20, 40, 25, 15, 18

After sorting

5, 10, 12, 15, 18, 20, 25, 30, 40
n = 9

$Q_1 = \left(\frac{(n+1)}{4}\right)$th term

$= \frac{10}{4} = 2.5$

Rewriting:

Q_1 = 2nd term + 0.5 (3rd term - 2nd term)
= 10 + 0.5 (12 - 10)
= 10 + 0.5 × 2
= 11

$Q_2 = \left(\frac{2(n+1)}{4}\right)$th term

$\frac{2(10)}{4} = 5 = 18$

$Q_3 = \left(\frac{3(n+1)}{4}\right)$-th term

$= \frac{30}{4}$ 7.5

Rounding

Q_3: 7th term + 0.5 (8th term − 7th term)

$= 25 + 0.5\,(30-25)$ $\quad \frac{0.5}{\times\ 5}\ \frac{}{5}$

$= 25 + 0.25$

$= 25.25$

HW Q Find 1st, and, 3rd quartile of

8, 5, 15, 20, 18, 30, 40, 25

Find IQR

Sort: 5, 8, 15, 18, 20, 25, 30, 40

Q_1: $[(n+1)/4]$-th term

$= \frac{9}{4} = 2.25$th term

= 2nd term + 0.25 (3rd term − 2nd term)

= 8 + 0.25 (15 − 8) = 9.75

$Q_2 = [2(n+1)/4]$ th term

$= \frac{18}{4} = 4.5$ th item

$\therefore$ 4th term + 0.5 (5th term − 4th term)

$= 18 + 0.5(20 - 18)$

$= 19$

$Q_3 = [3(n+1)/4]$ th term

$= 27/4 = 6.75$

$\therefore$ 6th term + 0.75 (7th term − 6th term)

$= 25 + 0.75(30 - 25)$

$= 28.75$

IQR $= Q_3 - Q_1$

$= 28.75 - 9.75$

$= 19$

Box Plot:-

It is a graphical method to visualize data distributions for gaining insights and making informed decisions.

A box plot is also known as Whisker plot and is created to display the summary of the set of data values

Elements of Box Plot:-

i. Minimum :- It is the minimum value in the dataset

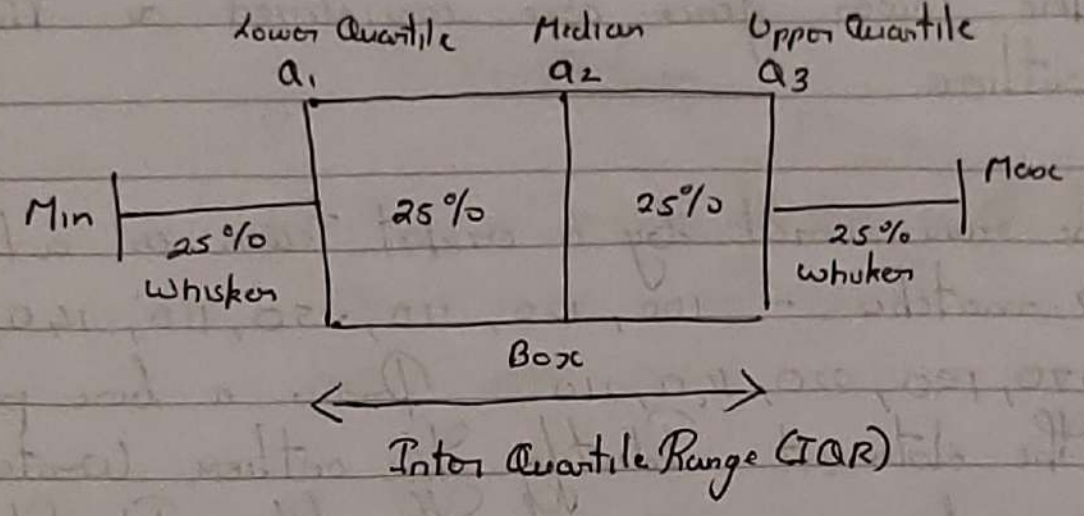
ii. Q_1 :- 25% of data lies below the First (lower) Quartile.

iii. Median (Q_2) :- Mid-point of the data set

iv. Third Quartile (Q_3) :- 75% data lies below the 3rd quartile (Upper quartile)

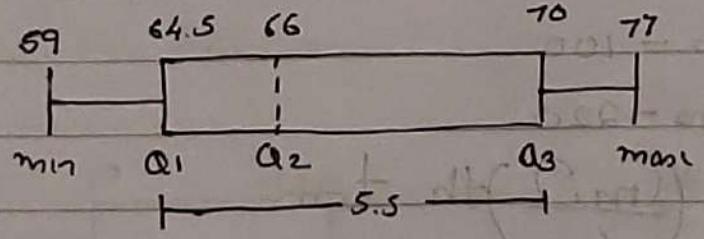
v. Maximum :- Maximum value in the dataset

This is the 5-Number Summary Data.



a) Heights of 40 students having the 5-number summary as follows. Construct a box plot for the dataset

min = 59 max = 77 Q1 = 64.5 Q2 = 66 Q3 = 70



Outliers :- (Finding out outliers)

Upper Fence = $Q_3 + (1.5 * IQR)$

Lower Fence = $Q_1 - (1.5 * IQR)$

The values above the upper fence and below the lower fence are considered as the outliers.

Q The runs scored by a cricket team in a league of 12 matches - 100, 120, 110, 150, 110, 140, 130, 170, 120, 220, 140, 110. Draw a box plot for the data set. Identify the outliers. Write the 5 number summary of the data. Identify the outliers

Sort → 100, 110, 110, 110, 120, 120, 130, 140, 140, 150, 170, 220

Minimum - 100

Maximum - 220

$Q_1 = \left(\frac{(n+1)}{4}\right)$th term

$\frac{12+1}{4}$th term

: 3.25 th term

3rd term + 0.25 (4th term - 3rd term)

110 + 0.25 (110 - 110)

110

$Q_3 = \frac{3(n+1)}{4}$ th term

$\therefore \frac{39}{4} = 9.75$ th term

$= 9$th term $+ 0.75$ (10th term $-$ 9th term)

$= 140 + 0.75\,(150 - 140)$

$= 140 + 147.5$

$Q_2 = \frac{2(n+1)}{4}$

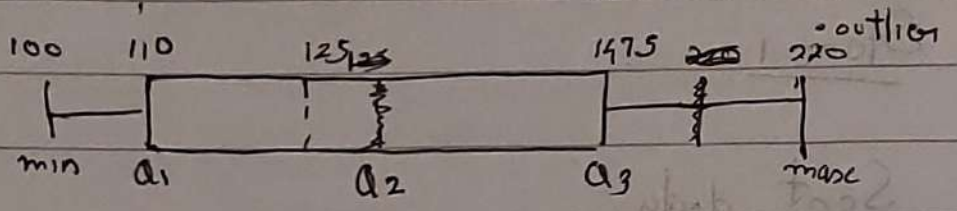
$\therefore \frac{26}{4} = 6.5$

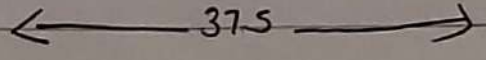
$=$ 6th term $+ 0.5$ (7th term $-$ 6th term)

$= 120 + 0.5\,(130 - 120)$

$= 125$

$IQR = Q_3 - Q_1 = 37.5$





Upper Fence $= Q_3 + (1.5 \times IQR)$

$= 147.5 + (1.5 \times 37.5) = 203.75$

Lower Fence $= Q_1 - (1.5 \times IQR) = 53.75$

Ostlier is 220

0 An IT company has two stores that sell computers. The company record the Number of sale each store made each month. In the past 12 months we have the following no of sold computers

Store 1 350, 460, 20, 160, 580, 250, ~~120~~ 210, 120, 200, 510, 290, 380

Store 2 520, 180, 260, ~~380~~ 80, 500, 630, 420, 210, 70, 440, 140.

Give the 5 no summary and draw box plot.

Store 1

Sort data

20, 120, 160, 200, 210, 250, 290, 350, 380, 460, 510, 580

Minimum = 20
Maximum = 580

Median (Q_2) = $\frac{2(n+1)}{4}$ th term

$= \frac{2 \times 13}{4}$ th term

$= \frac{26}{4}$ th term

$= 6.5$

6th term + 0.5 (7th term - 6th term)

250 + 0.5 (290 - 250)

250 + 0.5 × 40

250 + 20

270

$Q_1 = \frac{(n+1)}{4}$ th term

$= \frac{13}{4} = 3.25$

: 3rd term + 0.25 (4th term - 3rd term)

: 160 + 0.25 (200 - 160)

: 160 + $\frac{25}{100} \times 40$

: 170

$Q_3 = \frac{3(n+1)}{4}$ th term

$= \frac{3 \times 13}{4}$

$= \frac{39}{4}$ th term

$= 9.75$ th term

9th term + 0.75 (10th term − 9th term)

380 + 0.75 (460 − 380)

380 + 0.75 × 80

380 + 60

440

IQR. $Q_3 - Q_1$

$= 440 - 170$

$= 270$

5-number summary

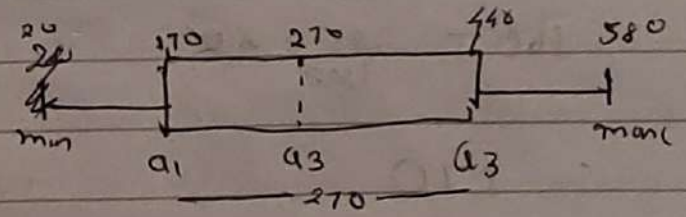
Minimum: 20

Q_1: ~~180~~ 170

Median: 270

Q_3: 440

Maximum: 580



Stone 2

Sort the data

70, 80, 140, 180, 210, 260, ~~300~~, 380, 420, 440, 500, 520, 630

Minimum: 70 Maximum: 630

$Q_1 = \left(\frac{n+1}{4}\right)$th term

$\frac{13}{4}$th term

3.25th term

= 3rd term + 0.25 (4th term - 3rd term)

= 140 + 0.25 (180 - 140)

= ~~150~~ 150

$Q_2 = 2\left(\frac{n+1}{4}\right)$ th term

$\frac{2 \times 13}{4} = 6.5$ th term

= 6th term + 0.5 (7th term - 6th term)

= 260 + 0.5 (380 - 260)

= 260 + 0.5 (60) = ~~240~~ 320

Q_3 : $3\frac{(n+1)}{4}$ th term

$\frac{3 \times 13}{4}$ th term

9.75 th term

9th term + 0.75 (10th term - 9th term)

420 + 0.75 (500 - 420)

420 + 0.75 × 20

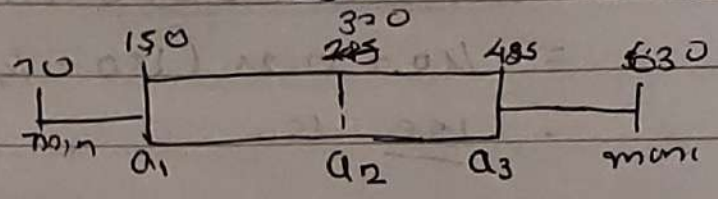
420 + 15

485

IQR. $Q_3 - Q_1$

: 435 - 150

285



5-number summary

Minimum: 70

Maximum: 630

Q_1: 150

Q_2: 285

Q_3: 435

2 Consider the data set: 180, 156, 9, 176, 163, 1827, 166, 171
Draw box plot and identify outliers

After sorting:

9, 156, 163, 166, 171, 176, 180, 1827

Q1. $\frac{(n+1)}{4}$ th term

$\therefore \left(\frac{9}{4}\right)$ th term = 2.25 th term

$\therefore$ 2nd term + 0.25 (3rd term − 2nd term)

$= 156 + 0.25\,(163 - 156)$

$= 156 + 0.25 \times 7$

$= 156.75$

$=$

Q2. $\frac{2(n+1)}{4}$ th term

$\therefore \frac{9}{2}$ th term

$\therefore$ 4.5 th term

$\therefore$ 4th term + 0.5 (5th term − 4th term)

$166 + 0.5\,(171 - 166)$

$166 + 0.5 \times 5 \quad : \quad 168.5$

$=$

$Q_3 = \frac{3(n+1)}{4}$

$= \frac{3 \times 9}{4} = \frac{27}{4}$ th term

6.75 th term

6th term + 0.75 (7th term - 6th term)

176 + 0.75 (180 - 176)

176 + 0.75 × 4

179

IQR = $Q_3 - Q_1$

= 179 - 157.75

= 21.25

Outlier

Upper Fence = $Q_3 + (1.5 \times IQR)$

= 179 + (1.5 × 21.25)

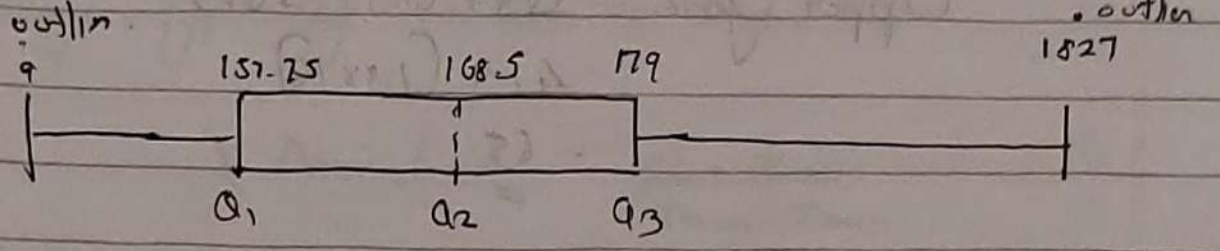
= 179 + 31.875

= 210.875

Lower Fence = $Q_1 - (1.5 \times IQR)$

= 157.75 - 81.875

= 75.875



Outlier :- 1827, 9

3 22, 24, 25, 28, 29, 31, 35, 37, 41, 53, 64

min - 22

Q_1 = ((n+1)/4)th term

= 12/4 = 3 → 25

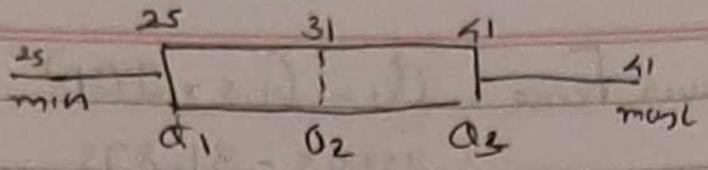
Q_2 = ((n+1)/4)th term

= 24/4 = 6 → 31

Q_3 = (3(n+1)/4)th term

= 36/4 = 9 → 41

IQR = $Q_3 - Q_1$ = 41 - 25 = 16



Upper fence = $Q_3 + (1.5 \times IQR)$
$= 41 + (1.5 \times 16)$
$= 65$

Lower fence = $Q_1 - (1.5 \times IQR)$
$= 25 - 24$
$= 1$

No outlier

Normalization :-

- Preprocessing technique
- Scaling technique in machine learning, applied during data preparation to change the values of numeric column in the dataset to use a column scale

Method of Normalization

-> Min - Max Normalization
-> Z-Score Normalization
-> Decimal Scaling

Min-Max Normalization

→ Min-max scaling transforms features to a specified range, typically between 0 and 1.

$$X \text{ Normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

a Use the min-max method to normalize the data
200, 300, 400, 600, 1000

min = 200

max = 1000

x	$\frac{x - x_{min}}{x_{max} - x_{min}}$
200	$\frac{200 - 200}{1000 - 200} = 0$
300	$\frac{300 - 200}{1000 - 200} = 0.125$
400	$\frac{400 - 200}{6000 - 200} = 0.25$
600	$\frac{600 - 200}{1000 - 200}$

$$1000 = \frac{1000 - 200}{1000 - 200} = 1$$

2 Z-Score Normalization

- Also known as standardization
- Transforms the data so that the mean is 0 and standard deviation is 1
- Process adjust the data values based on how far they deviate from the mean measured in units of standard deviation

$$Z = \frac{x - \mu}{\sigma} \qquad \sigma - S.D \qquad \mu - mean$$

a Use z-Score to normalize

200, 300, 400, 600, 1000

$$\mu = \frac{200 + 300 + 200 + 600 + 1000}{5}$$

$$= \frac{2500}{5} = 500$$

$$\text{Variance} = \frac{(200-500)^2 + (300-500)^2 + (400-500)^2 + (600-500)^2 + (1000-500)^2}{5}$$

S.D. $= \sqrt{\frac{90000 + 40000 + 10000 + 1000 + 250000}{5}}$

$= 209.79 \quad (282.8)$

Decimal Scaling Normalization.

→ In this technique, we move the decimal point of value of the attribute

→ The movement of the decimal points totally depends on the maximum value among all values in the attribute.

Normalized Value of Attribute $= \frac{v}{10^j}$

Step 1: find maximum absolute value.

Step 2: Find the value of j (if max value is 100 then j is, i.e the no of digits)

Step 3: $\frac{v}{10^j}$

Q Normalize using decimal scaling
200, 300, 400, 600, 1000

$$\frac{v}{10^j}$$

max = 1000 $\therefore j = 4$
(max absolute value)

x	x - normalized (v)
200	$\frac{200}{10^4} = \frac{200}{10000} = 0.02$
300	$\frac{300}{10^4} = \frac{300}{10000} = 0.03$
400	$\frac{400}{10^4} = \frac{400}{10000} = 0.04$
600	$\frac{600}{10^4} = \frac{600}{10000} = 0.06$
1000	$\frac{1000}{10^4} = \frac{1000}{10000} = 0.1$

Data Similarity and Dissimilarity

Data Similarity is the measure of how alike two data objects are.

→ Data dissimilarity is a measure of how different the data objects are.

→ The similarity measure is usually expressed as a numerical value

→ The value is high when the objects are more alike

* Proximity Refers to Similarity or Dissimilarity

Methods for Measuring Similarity / Dissimilarity.

→ Euclidean Distance :-
→ Manhattan Distance
→ Minkowski Distance
→ Supreme Distance

Euclidean Distance :

Let the two points i and j, the distance b/w the points is given by

$$d(i,j) = \sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \dots + (x_{in}-x_{jn})^2}$$

e Compute the Euclidean Distance b/w the two objects (22, 1, 42, 10) and (20, 0, 36, 8)

$$d_{i,j} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{in} - x_{jn})^2}$$

$$= \sqrt{(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2}$$

$$= \sqrt{2^2 + 1^2 + 6^2 + 2^2}$$

$$= \sqrt{45}$$

$$= 6.71$$

Manhattan Distance:-

$$d_{i,j} = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{in} - x_{jn}|$$

a (22, 1, 42, 10) and (20, 0, 36, 8)

$$d_{i,j} = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{in} - x_{jn}|$$

$$= |22-20| + |1-0| + |42-36| + |10-8|$$

$$= 2 + 1 + 6 + 2$$

$$= 11$$

Minkowski Distance:

$$d_{ij} = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \ldots + |x_{in} - x_{jn}|^p)^{1/p}$$

Q Compute the Minkowski Distance
(22, 1, 42, 10) and (20, 0, 36, 8) $p=3$

$$d(i,j) = (|x_{i1}-x_{j1}|^p + |x_{i2}-x_{j2}|^p + \dots + |x_{in}-x_{jn}|^p)^{1/p}$$

$$(|22-20|^3 + |1-0|^3 + |42-36|^3 + |10-8|^3)^{1/3}$$

$$= (8+1+6^3+2^3)^{1/3}$$

$$(8+1+6^3+8)^{1/3}$$

$$\sqrt[3]{233}$$

$$6.15$$

Supreme Distance :-

$$d(i,j) = \max(|x_{i1}-x_{j1}|, |x_{i2}-x_{j2}|, \dots |x_{in}-x_{jn}|)$$

Q Find the supreme distance between 2 objects
(22, 1, 42, 10) and (20, 0, 36, 8)

$$d(i,j) = \max(|x_{i1}-x_{j1}|, |x_{i2}-x_{j2}| \dots |x_{in}-x_{jn}|)$$

$$= \max(|22-20|, |1-0|, |42-36|, |10-8|)$$

$$= \max(2, 1, 6, 2)$$

$$= 6$$

Module 2

Database:-

→ Stores data in tables
→ Deals with operational or transactional data (current data).
→ Can store MBs to GBs of data.
→ Used for OLTP (Online Transaction Processing)