29|12|20 Lec-03

→ Major Issues in data mining (End of chp 01)

Chp 02 : Getting to know your data

→ data objects and attributes types (attributes : nominal, binary, ordinal, numeric, discrete vs. continuous)

→ Statistic description

i> central tendency (mean, median, mode)

ii> data dispersion (range, quartile, variance, std dev, IQR)

iii> Graphic displays

→ data objects / samples / examples / instances / data points / data tuples (rows)

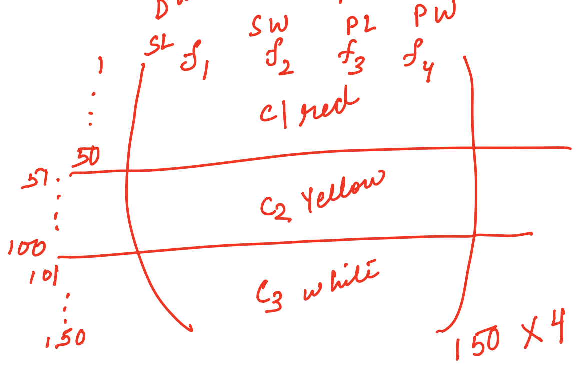
→ attributes / dimension / feature / variable (columns)

attributes → DM and data base

dimension → commonly used in DWH

feature → mostly used by ML experts

variable → mostly used in statistics

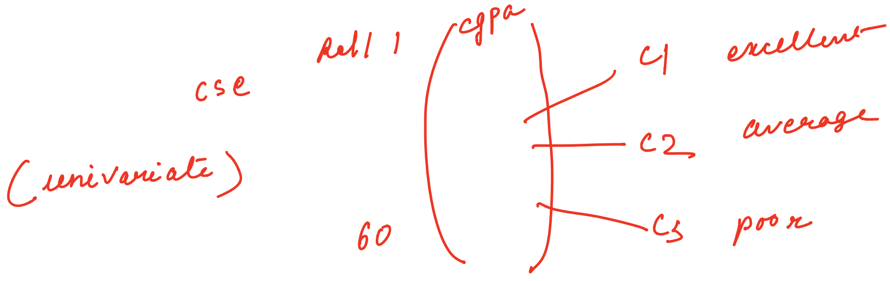


Iris is a flower

Row4 $\langle f_1, f_2\ f_3\ f_4 \rangle$

feature vector

<UCI - ML repository>



(bivariate)

$$\begin{array}{c} \\ 1 \\ \vdots \\ 60 \end{array}\left(\begin{array}{cc} f_1 & f_2 \\ \text{cgpa} & \text{Age} \\ & \\ & \end{array}\right)$$

Types of attributes :

1) **Nominal** : (some kind of category) / categorical attributes:

Eg: hair color : black, brown, blonde, red, auburn
grey white

marital status: single, married, divorced, widowed

2) **Binary** : (categorical → 0/1)

Eg: out come of medical test → 0 (normal, -ve test)
→ 1 (patient, +ve test)

symmetric
Eg gender
M F

asymmetric
HIV Test — pos (rare)
neg. (majority)

3) **Ordinal attribute** : (ranking)

Eg: drink size available at fast food restaurant → small, medium, large

grade : A, A+, A-, B+

customer satisfaction: 0 → very disatisfied
1 → somewhat satisfied
2 → neutral
3 → satisfied
4 → very satisfied

* The central tendency of an ordinal attribute can be represented by mode and median, but mean cannot be

defined .

* Nominal, binary and ordinal attributes are qualitative

4) <u>numeric attributes</u> (quantitative):

→ rep. either as integers or real nos:

for Eg: Age <0 .... 100>

cgpa <3.5 .... 9.5>

rainfall <30 40 . . . 110>
in cm

5) <u>discrete vs. continuous</u>:

Eg: of discrete : age (0... 100)

Binary attribute (0 or 1)

Infinite (cus- ID, PIN code etc)

Eg continuous: 30°C ... 30.59999°C

<u>Central tendency</u> (mean, median, mode)

average value — middle value — most common value

→ Gives an idea of "middle" or "center" of data distribn.

1) $\boxed{\text{Mean} = \frac{\sum x}{n}}$ or $\boxed{\text{mean} = \frac{\sum fx}{\sum f}}$

C1

Age :	19	20	21
f :	1	3	1

$$M = \frac{19 + 20\times3 + 1\times21}{5} = \frac{100}{5} = 20$$

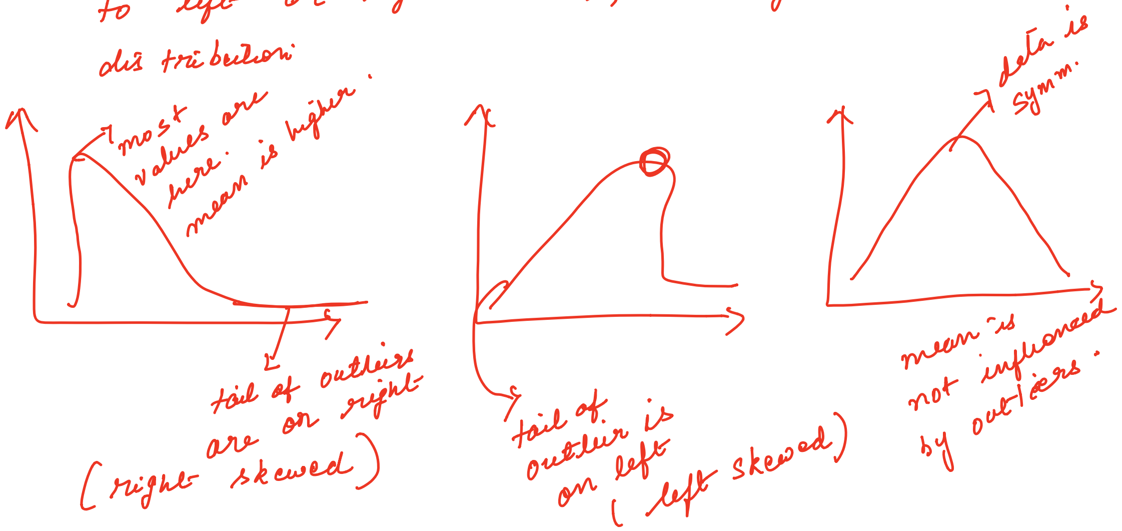
x x → outliers

C2 :

Age:	19	20	21	$\overset{\uparrow}{145}$	$\overset{\uparrow}{147}$
f:	3	6	3	1	1

$\mu = 38$ (presence of outlier has pulled the mean higher)

→ Presence of outlier pulls the data mean either to left or right. Thus, making the data skewed distribution.


most values are here. mean is higher.
tail of outliers are on right
(right skewed)
tail of outlier is on left
(left skewed)
data is symm.
mean is not influenced by outliers.

2) Finding the median :

STEPS :

1. Line up your numbers from smallest to largest.
2. If you have odd no. of values, the median is the one in the middle. If 'n' nos., then median is at posn. $\left(\frac{n+1}{2}\right)$.
3. If you have even no. of values, median is obtaind by adding two middle nos. together and dividing by 2.

Eg 1: Ages 19 19 20 20 $\boxed{20}$ 21 22 100 102

$n = 9$ $\therefore$ median pos $= \frac{n+1}{2} = \frac{9+1}{2} = 5^{th}$

value = 20.

Eg 2 :

Ages : 19 20 20 20 | 21 21 100 102

(4.5)

posn : $\frac{1+8}{2} = 4.5$

value : $\frac{20+21}{2} = 20.5$

Q Find the mean, median and check whether the data is skewed or not. Check whether mean is higher / lower than median.

i)

Values	1	2	3	4	5	6	7	8
f	4	6	4	4	3	2	1	1

ii)

Values :	1	4	6	8	9	10	11	12
f :	1	1	2	3	4	4	5	5