Data Mining

⇒ DM is the process of Extracting knowledge or insights from large amount of data using various statistical, & Computational techniques.

⇒ The data can be structured, semi-structured or Unstructured, and can be stored in various forms such as databases, data warehouses, and data lakes.

⇒ The primary goal of data mining is to discover hidden patterns and relationships in the data that can be used to make informed decisions or predictions.

⇒ This involves Exploring the data using various techniques such as clustering, Classification, Regression analysis, association rule mining, and anomaly detection.

⇒ Data mining has a wide range of applications across various industries, including marketing, finance, healthcare and telecommunications.

Eg:- In marketing data mining can be used to identify customer segmentation and target marketing campaigns.

In healthcare it can be used to identify risk factors for diseases and develop personalized treatment plans.

⇒ Database Data:- A db system also called database management system (DBMS), consists of a collection of interrelated data, known as database. and set of s/w programs to manage and access the data.

⇒ A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (record or row).

⇒ Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values.

⇒ A semantic data model such as entity-relational (ER) data model, is often constructed for relational database.

⇒ A ER model represents the database as a set of entities and their relationships.

eg:- "Show me a list of all items that were sold in the last quarter"

Here Relational languages also use aggregate functions such as sum, avg, count, max, min,

* Data Warehouse :- A DW is a repository of information collected from multiple sources.

⇒ DW are constructed via a process of data cleaning, data integration, data transformation, data loading, periodic of data refreshing.

⇒ A data warehouse is usually modeled by a multi dimensional data structure called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema.

of information ...

Transactional data :- In database captures a transaction, such as a customer's purchase, a flight booking, or a user's clicks on a web page.
A transactional database may have additional tables, which contain other information related to the transaction such as item description, information about the salesperson or the branch, and so on.
Eg:- "which items sold well together" This kind of market basket data analysis.

Trans-ID	List of Item-IDs
T100	I_1, I_3, I_8, I_{16}
T200	I_2, I_8
...	...

* Descriptive :- Descriptive mining tasks characterize properties of the data in a target dataset.
* Predictive :- predictive mining tasks perform induction on the current data in order to make predictions

1 → Basic statistical Description of data :-

The basic statistical description of data can be used to :
the properties of data.

* Measuring the Central Tendency :- (Mean, Median, Mode)

a) Mean :- Mean is most common & effective numeric measure which is used to measure the center of data.

Mean = $\bar{x}$

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Mean of sales

$$= \frac{x_1 + x_2 + \cdots x_N}{N}$$

center of the whole data.

if weight are associated

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3 \cdots w_N x_N}{w_1 + w_2 + w_3 \cdots w_N}$$

This is called weighted arithmetic mean or the weighted average.

b) Median : Median is the best measure for finding the center of the data.

Ex :- 40, 45, 47, 50, 52, 52, 57, 61, 64, 71, 110.

center $\frac{52+56}{2} = \frac{108}{2} = 54$ center position / middle position

c) Mode :- another measure for central tendency is mode.

→ Mean − Mode ≈ 3 × (mean − median)

d) Mid Range :- To measure the central tendency of numeric data set average of largest & smallest

$$= \frac{30 + 110}{2} = \frac{140}{2} = 70.$$

in which attribute or a set of attributes.

looking for a

Measuring the Dispersion of data :- (3)

Range :- The difference b/w largest & smallest.

Quantile :- Split the data distribution equal size consecutive sets

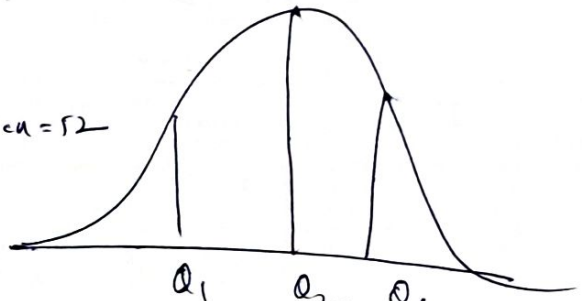
Quartile :- one fourth of data distribution

percentile :- 100 quantities are more commonly referred to as percentile

Interquartile range (IQR)

The distance between the first & third quartiles is a single measure of spread that gives the range covered by the middle half of the data. This distance is called the IQR

$$IQR = Q_3 - Q_1$$



$Q_1 = 47$ $Q_3 = 65$ & Median = 52

$\Rightarrow$ $IQR = 65 - 47$

$= 18$

25th %. Median 75-1.

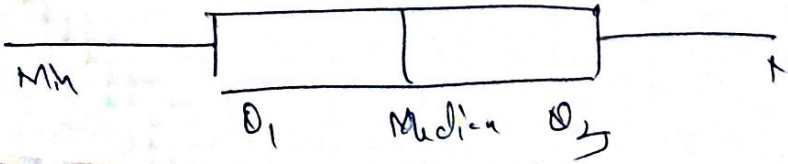
* Five Number Summary :-

To identify the middle and what next value will come.

It is also called boxplots.

* Five Number summary of a distribution consists of the median (Q_2) the quartiles Q_1 and Q_3, and the smallest and largest individual observations, written in the order of

Minimum, Q_1, Median, Q_3, Maximum



Variance :- The variance of N observations, $x_1, x_2 \ldots x_N$ for a numeric attribute X is

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$\bar{x}$ is the Mean value

$$= \left(\frac{1}{N} \sum_{i=1}^{N} x_i^2 \right) - \bar{x}^2$$

standard deviation, σ, of the observations is the square root of the variance, σ^2

$$\sigma \approx \sqrt{\sigma^2}$$

⇒ Data Mining Functionalities :-

* 5 functionalities

1→ Concept / class (definations) / descriptions :-

Data is always associated with class / concepts

⇒ class / concepts can be done in '2' ways

1) Data characterisation :-

2) Data descrimination :-

1) Data characterisation :-

Refers to the summary of the class / concept about which are studying.

o/p → General overview.

2) Data Descrimination :-

Compares the common features of the classes

Here we have '2' classes / concept's then it will compare the common features which are in between two classes or those two concepts

If we have any changes in the two classes or concept it will observe.

o/p :- Barchart, curves, graph etc.

2→ Mining frequent patterns, associations & correlations :-

Mining :- Data Mining is the process of sorting through large sets to identify patterns and relationships that can help solve the business problems through data analysis.

Anomaly :- Something different, abnormal, peculiar, or not easily clarified.

Frequent patterns :- Things which are found most commonly in

There are many kind of frequent patterns are there.

1. "frequent itemsets", 2) "frequent sub sequences", & 3) "frequent substructures".

[data item's / Object's]

1. frequent Itemsets :- Refers to a set of items that often appear together in a transactional data sets.

Eg:- Milk & bread. [which are frequently bought together in grocery stores by many customers.

2. Frequent subsequences :-

frequently occurring subsequences, such as pattern that customers, tend to purchase first laptop, followed by a digital camera, and then memory card is a frequent sequential pattern.

3. Frequent substructure :-

F.S can refer to different structure forms that may combined (Eg:- Graphs, trees, or lattices) with itemsets or subsequences.

"Association Analysis" :- → [relationship]

* It is a way of identifying the relationship between various item.

Eg:- Used to determine sales of items that are "frequently purchased together" ① dry fruit + ...

② How the data items are related to each other and what is association between the data items (x → y)

x → dry fruit Y → kho

⇒ Dry fruits are frequently purchased together items, next purchase together what are the items that are purchased together

Dry fruits

↳ frequently

↳ one data item

one set

Dry fruits ⇆ Chocolates

together

↳ Another data item

another set

Correlation Analysis :-

→ It is a Mathematical techniques.

→ shows how strongly pair of attributes are related together.

Eg:- Tall people tend to have more weight

(1) (2)

This is one attribute

This is another attribute

↳ two attributes are strongly related to each other; & how they are related to each other.

3. Classification and regression for predictive analysis :-

↳ Predicting the data → (Assuming the data)

if data is missing we have to fill it or assume it.

Classification is process of finding a model that distinguishes data items

↳ 'Using colour'

Eg:- decision one group

Chocolate ⇆ Dairy Milk

Kit Kat

Two different types of Chocolates will be classifying them. → second group

$S = \frac{\text{Item}}{\text{Total}}$

→ Nominal

class

Mass

Killed | Alive

↑ T

⇒ Here data items are classifying in to the name(co... based on the color of the wrapper.

⇒ Here what is the model? Wrapper color is the model your finding some model in order to distinguish the data items that you are having. this is the classification.

* "Decision trees" :- is a flowchart-like tree structure. based on decision tree we can easily classifying the data.

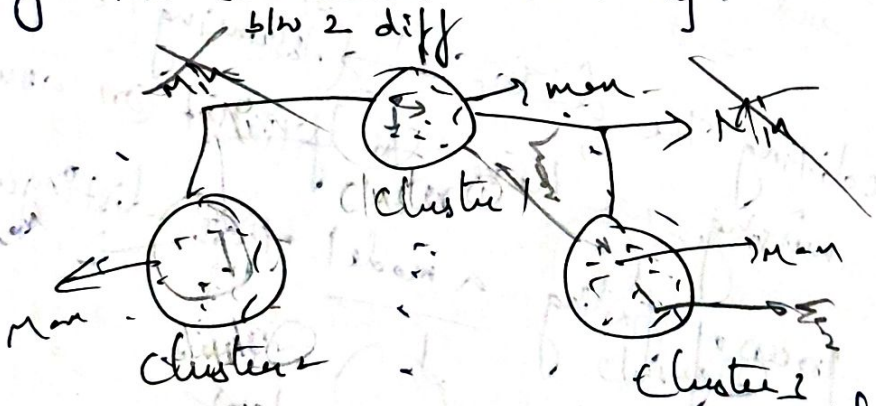
Regression :-

Statistical methodology that is used for numeric prediction of missing data. [done based on previous data]

Eg 2,3,4,6,7,8,10,11,12 ⟷ 15,16,17

↳ predicting the data.

4. Cluster Analysis :- The data items are clustered based on the principle of maximising the intraclass (with in the same) similarity and minimising the interclass (b/w 2 diff.) similarity:



→ Dots represents the data so these data items are classified in to three clusters based on the similarity between those things.

→ So what ever data sets are inside this cluster one whatever the data items are there all those data

items are similar to each. (6)

~~Item Inside the~~ Item which are inside this cluster is maximum, and the similarity between the items which are inside this cluster are also maximum.

⇒ But the similarity between the items which are inside this cluster should be minimum this what minimizing the inter class between two classes or between two clusters.

⇒ So within the cluster the similarity between the data objects should be maximum.

⇒ And between two different clusters the similarity between the data objects should be minimum

⇒ (Analysis ~~of that's all~~ clusters is called is of

⇒ " Analysis of these clusters is called as cluster Analysis "

Associat → Market Basket
outlie → Credit card, fraud detection.
clus →

5- Outlier Analysis :- (anomaly mining)

⇒ Among the data items in a db, there may be some items which do not follow the general behaviour of data.

These data items → Outlier [noise | Exceptions]

⇒ Anomaly :- Which do not obey the rules & which are not according to the rule which we have defined but some abnormalities we can say like that

Eg { 2, 4, 6, (7), 8, 10, 12, ... }

- 2, 4, 6, 8, 10, 12 → Even
- 7 → odd element → Outlier

One whatever. me

→ Proximity Measures for Binary Attributes :-

* Let i, j are two objects and 'p' total number of attributes. (0, 1 either F, T)

* Symmetric Binary attribute :- [N objects]

Eg:- M / F → Equal importance.

* Asymmetric Binary Attribute :- "Unequal importance"

Height important for one / other one it not much important.

(M / F)

$$d = \frac{q+r}{q+r+s+t}$$

dissimilarity calculation

← N / P
↓ +ve
-ve

q - 1 for both objects.

r - 1 for i, 0 for j } same.

s - 0 for i, 1 for j } different

t - 0 for both i and j

q, r, s, t

Obj j / Obj i	1	0	Sum
1	q	r	q+r
0	s	t	s+t
Sum	q+s	r+t	p → Total No of objects.

(Contingency Matrix)

Dissimilarity for symmetric Binary attribute

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

Dissimilarity for Asymmetric Binary attribute (all values are not important)

$$d(i,j) = \frac{r+s}{q+r+s}$$

Asymmetric Binary attribute → P + highest imp, -N lowest imp

't' both are negative importance [N] → t[0, 0] ↑ ignored.

⇒ Then Asymmetric Binary similarity

$$Sim(i,j) = 1 - dis(i,j)$$

$$= 1 - \frac{r+s}{q+r+s}$$

$$= \frac{q+r+s-r-s}{q+r+s}$$

$$Sim_{jaccard}(i,j) = \frac{q}{q+r+s}$$

Eg: let the values Y & P to be set → 1, N & Negative let set to 0

$$d(i,j) = \frac{r+s}{q+r+s}$$ dissimilarity for Asymmetric B.A

self patient table

O → Y, N
1, 0

S → 0, 1
N Y

q → 1 1
Y Y

$$dis(jack, jim) = \frac{1+1}{1+1+1}$$

$$= \frac{2}{3} = 0.67$$ 0.23

$D3(jim, mary) = \frac{\text{no of mismatch}}{\text{no of attributes}} = \frac{1+1+1}{1+1+1+1} = \frac{3}{4} = 0.75$

$D2(jack, mary) = \frac{0+1}{1+1+0+1} = \frac{1}{3} = 0.33$

$= \begin{bmatrix} 0.23 \\ 0.67 \\ 0.75 \end{bmatrix}$

→ Jack & Mary → same disease → similarity is more because distance is less

out of three patients, Jack & Mary are the most likely to have a similar disease.

→ Proximity Analysis for ordinal variables :-

here we use ranking.

Name	Rank	r_{if}	$z_{if} = \frac{r_{if}-1}{M_f-1}$	$r_{if}\ \frac{(4-1)}{3}$
Jack	Excellent	4	$(4-1)/3 = 1$	
Mary	Better	3	$(3-1)/3 = 0.67$	
Jim	Good	2	$(2-1)/3 = 0.33$	
pat	average	1	$(1-1)/3 = 0$	

Total ordinal states M_f = 4 no of states (ranks)

$z_{if} = r_{if}-1(M_f-1) \rightarrow (r_{if}-1)/3$

$= 4-1 = 3$

$d(i,j) = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + \cdots |x_{ip}-x_{jp}|$

→ Manhattan distance

dissimilarity Matrix

	Jack	Mary	Jim	pat
Jack	0			
Mary	0.33	0		
Jim	0.67	0.34	0	
pat	1	0.67	0.23	0

similarity Matrix

	1			
	0.67	1		
	0.33	0.66	1	
	0	0.33	0.67	1

$= 4 - 0.33$

$= \frac{4-1}{3}$

$= \frac{3-1}{3} = \frac{2}{3} = 0.67$

Dissimilarity for Attributes of Mixed types:-

No	Test I Nominal	Test II Ordinal	Test III Numeric
1	code A	excellent	45
2	code B	Fair	22
3	code C	Good	64
4	code D	excellent	28

[Nominal vedio]

⇒ Nominal $\frac{p-m}{p}$ =

Dissimilarity measure $= d(i,j) = (p-m)/p$

m - c no. of matches

p no. of variables.

p = 1

CA	0			
CB	1	0		
CC	1	1	0	
CD	1	1	1	0
	CA	CB	CC	CD

⇒ Ordinal attribute

	Nominal	ordinal	Numeric
1	CA	E	45
2	CB	F	22
3	CC	G	64
4	CD	E	28

$M_f = 3$ {ex, F, Good}

$z_{if} = (r_{if} - 1)/(M_f - 1)$

TII	Rank	r_{if}	z_{if}
1	ex	3	$(3-1)/2 = 1$
2	Fair	1	$(1-1)/2 = 0$
3	Good	2	$(2-1)/2 = 0.5$
4	Ex	3	$(3-1)/2 = 1$

Euclidean distance

code A(1)	0			
code B(2)	1	0		
code C(3)	0.5	0.5	0	
code A(1)	0	1	0.5	0
	code A(1)	code B(2)	code C(3)	code A(4)

	Name	Ordno	Num
1	LA	Σ	47
2	CB	F	22
3	CC	Go	64
4	CD	Σ	25

Min = 0

Max = 42

Divide by 42 for Normalization

Min Max

Code A	0			
Code B	23	0		
Code C	19	42	0	
Code D	17	6	26	0
	CA	CB	CC	CA

(9)

Data Mining

⇒ Similarity and dissimilarity :-

* More important - used in clustering, some classification, anomaly detection.

* **Similarity** :-
  1) Numerical measures of how alike two data objects are.
  2) Is higher when objects are more alike.
  3) often falls in the range [0,1]

* **Dissimilarity** :-
  1) Numerical measure of how different are two data objects.
  2) Lower when objects are more alike.
  3) Minimum dissimilarity often 0.
  4) Upper limit varies.

→ Proximity refers to a similarity or dissimilarity.

Sim | dis for Objects with single Attribute.

p, q are the attribute values for two data objects.

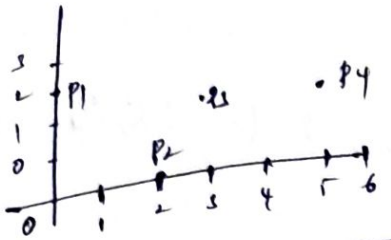
Attribute type	Dissimilarity	similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
ordinal	$d = \frac{\lvert p-q \rvert}{(n-1)}$	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p-q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ (or) $s = 1 - \frac{d - \min d}{\max d - \min d}$

and are

Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

point	x	y
P_1	0	2
P_2	2	0
P_3	3	1
P_4	5	1



$(P_1, P_2) = \sqrt{(x_1-x_2)^2 + (y_1-y_2)^2}$

$= \sqrt{(0-2)^2 + (2-0)^2} = \sqrt{(2)^2 + (2)^2} = \sqrt{4+4} = \sqrt{8}$

$= 2.828$

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

$= \sqrt{(0-3)^2 + (2-1)^2} = \sqrt{9+1} = \sqrt{10} = 3.162$

	P_1	P_2	P_3	P_4
P_1	0	2.828	3.162	5.099
P_2	2.828	0	1.414	3.162
P_3	3.162	1.414	0	2
P_4	5.099	3.162	2	0

{ distance joint to itself = zero / 0

Distance Matrix

$(P_1, P_4) = \sqrt{(0-5)^2 + (2-1)^2} = \sqrt{(5)^2 + (1)^2} = \sqrt{25+1} = \sqrt{26} = 5.099$

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

$(P_2, P_4) = \sqrt{(2-5)^2 + (0-1)^2} = \sqrt{(3)^2 + (1)^2} = \sqrt{9+1} = \sqrt{10} = 3.162$

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

Minkowski Distance

* Minkowski distance is a generalization of Euclidean Distance

Given two objects p & q

$$dist = \left(\sum_{k=1}^{n} |p_k - q_k|^r \right)^{1/r}$$

Where r is a parameter, n is the number of dimensions and p_k & q_k are, respectively, the kth attributes of data objects p and q.

→ r=1 city block (Manhattan distance.

→ r=2 :- Euclidean distance.

→ r=∞ :- "Supremum" (L_{max} norm, L_∞ norm) distance.

$$d(x,y) = \lim_{r \to \infty} \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Point	x	y
p_1	0	2
p_2	2	0
p_3	3	1
p_4	5	1

Manhattan distance formula

$\sqrt{1}$ | power of 1

L_1	p_1	p_2	p_3	p_4
p_1	0	4	4	6
p_2	4	0	2	4
p_3	4	2	0	2
p_4	6	4	2	0

(p_2-p_1) & same (p_1-p_2)

L_2	p_1	p_2	p_3	p_4
p_1	0	2.828	3.162	5.099
p_2	2.828	0	1.414	3.162
p_3	3.162	1.414	0	2
p_4	5.099	3.162	2	0

← Euclidean Distance formula $\sqrt{2}$ / power of 2

and so on the ...

Similarity and dissimilarity measures (Sneha Bagalkot in Delicious)

Manhattan Distance :

$P_1 = (0,2) = (x_1, y_1)$ $\quad P_2 = (2,0) = (x_2, y_2)$

Distance $= |x_1 - x_2| + |y_1 - y_2|$

$= |0-2| + |2-0|$

$= 2 + 2 = 4$

Distance from P_1 to P_2 is 4

And distance from P_2 to P_1 is 4.

→ Similarly for the other points.

Common properties of a Distance

1) $d(p,q) \geq 0$ for all p & q and $d(p,q) = 0$ only if $p = q$.
(positive definiteness)

2) $d(p,q) = d(q,p)$ for all p & q (symmetry)

3) $d(p,r) \leq d(p,q) + d(q,r)$ for all points p, q & r.
(Triangle Inequality).

where $d(p,q)$ is the distance (dissimilarity) between joints (data objects) p & q.

[Similarity Matching and coefficients]

SMC versus Jaccard : eg :-

→

$P = 1000000000$

0000001001

$M_{01} = 2$ (the no's of attributes where p was 0 and q was 1)

$M_{10} = 1$ (the no's of " where p was 1 & q was 0)

$M_{00} = 7$ (the no's of attributes where p was 0 & q was 0)

$M_{11} = 0$ (the number of attributes where p was 1 & q was 1)

$SMC = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$ (12)

$\Rightarrow (0+7) / (2+1+0+7)$

$= 7/10$

$= 0.7$

$J = (M_{11}) / (M_{01} + M_{10} + M_{11})$

$= 0 / (2+1+0) = 0$

SMC = No of Matches / no of attributes.

$= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

J = number of 11 matches / no of not-both-zero attribute values.

$= (M_{11}) / (M_{01} + M_{10} + M_{11})$

Cosine similarity

If d_1 & d_2 are two document vectors, then

$\cos(d_1, d_2) = (d_1 \cdot d_2) / \|d_1\| \|d_2\|$

where . indicates vector dot product and $\|d\|$ is the length of vector d.

Eg: $d_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$

$d_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$

$d_1 \cdot d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$\|d_1\| = (3*3 + 2*2 + 0*0 + 5*5 + 0*0 \ldots\ldots) = (42)^{0.5} = 6.481$

$\|d_2\| = (1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)^{0.5} = (6)^{0.5} = 2.245$

$\cos(d_1, d_2) = .3150$

$$\cos(d_1, d_2) = \frac{(d_1 \cdot d_2)}{\|d_1\| \|d_2\|}$$

$$\cos(x, y) = \frac{5}{(6.481)(2.245)}$$

$$= .3150$$

$\cos(x, y) = 0$ indicate both are dissimilar

$\cos(x, y) = 1$ indicate both are similar

Extended Jaccard Coefficient (Tanimoto)

$$T(p, q) = \frac{p \cdot q}{\|p\|^2 + \|q\|^2 - p \cdot q}$$

→ Problem :- "Euclidean distance" :-

	A_1	A_2	
x_1	1.5	1.7	0.1414
x_2	2	1.9	0.6708
x_3	1.6	1.8	0.2828
x_4	1.5	1.0	0.22236
			0.6081

(1.5, 1.7) → x_1, y_1

(1.4, 1.6) → x_2, y_2

* let $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$

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

$x = (1.4, 1.6)$ = $\sqrt{(1.4-1.5)^2 + (1.6-1.7)^2} = 0.1414 \Rightarrow 0.1414$

⇒ Manhatten Distance :-

	A_1	A_2	Manhatten Distance
1	1.5	1.7	$\lvert 1.4 - 1.5 \rvert + \lvert 1.6 - 1.7 \rvert = 0.2$
2	2	1.9	$= 0.9$
3	1.6	1.8	0.4
4	1.2	1.5	$= 0.3$
5	1.5	1.0	$= 0.7$

$x = (1.4, 1.6)$

let $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$

$$MD = \lvert x_2 - x_1 \rvert + \lvert y_2 - y_1 \rvert$$

and acces in

	A_1	A_2
1	1.5	1.7
2	2	1.9
3	1.6	1.8
4	1.2	1.5
5	1.5	1.0

$x = (1.4, 1.6)$

$h = 2$

Minkowski distance =

let $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$

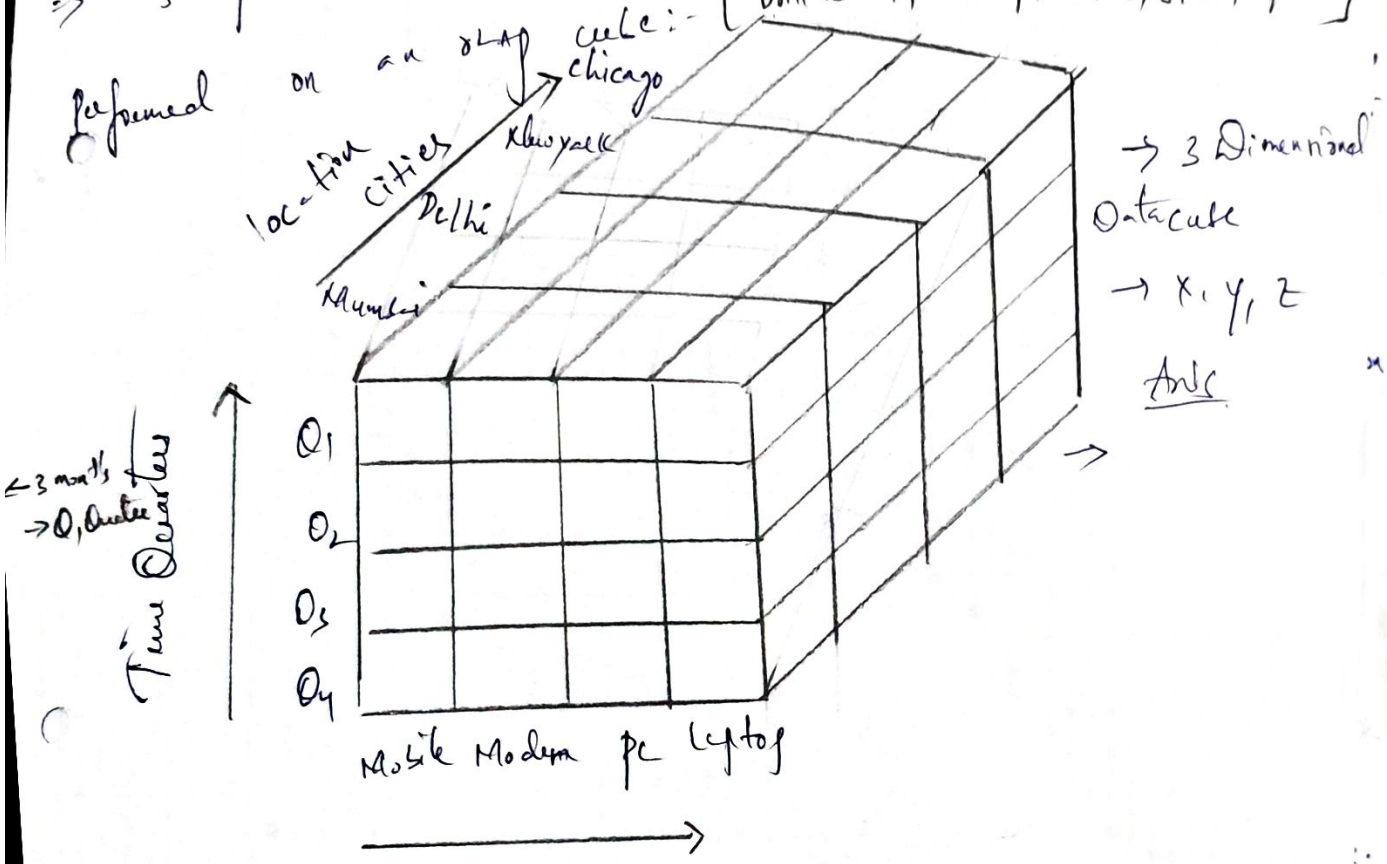
$$\sqrt[h]{(x_2 - x_1)^h + (y_2 - y_1)^h}$$

"OLAP operations" :-

=> They are based on multidimensional data model and allows the user to query on multidimensional data.

=> OLAP cub is a structure that is optimized for proper data analysis

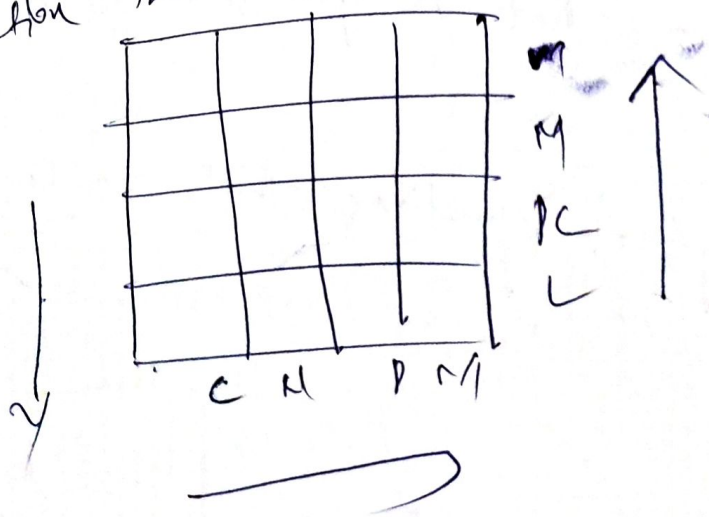
=> 5 operations basic analytical operations that can be performed on an OLAP cube:- [Drill Down, Rollup, Dice, slice, pivot]



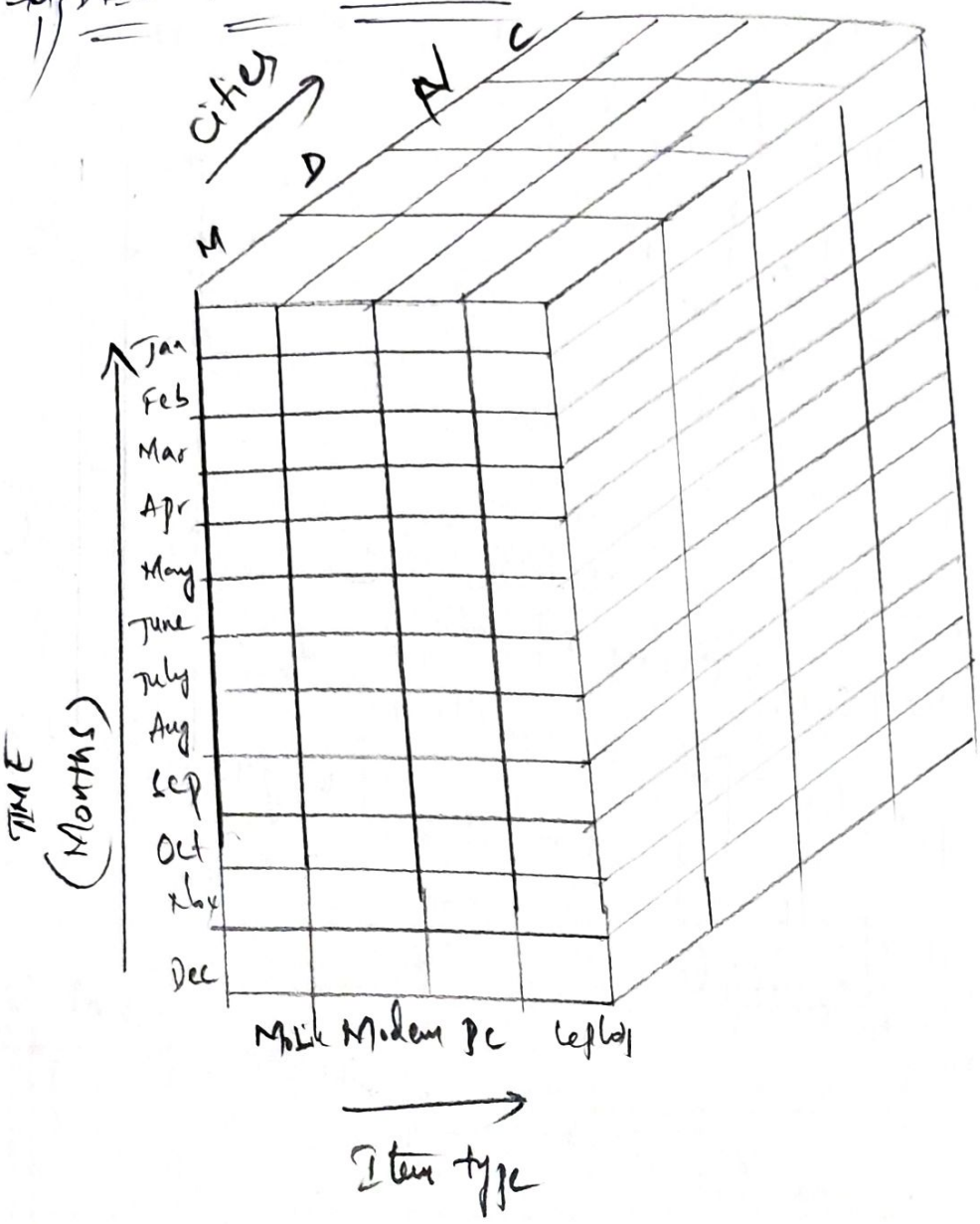
Drill Out:- We want detail's information or more information from the down side cube

⇒ Pivot :- [Rotating any given data]

& The positions of x, y, & are changes. & rotating of the information then it will atomatically reflected to the x, y, & cols.

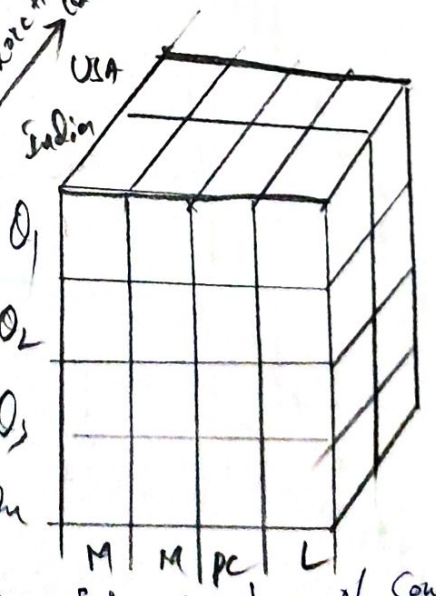


i) DRILL DOWN ON TIME :-



ii) Rollup :- Exact opposite side (or) Top of the cube operations we will perform.

Rollup on locations :



cities are four → Chicago } USA
→ New York }
→
→ Delhi } India
→ Mumbai }

* We want generalization information. [here in terms of country]
and access the data

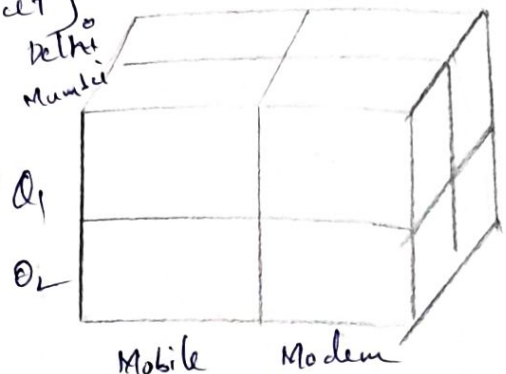
3. Dice :- Dice, Eg :- ludo

→ In Games we have

→ we called some part of portion of cube we call it has

* Dice. [some simple part]

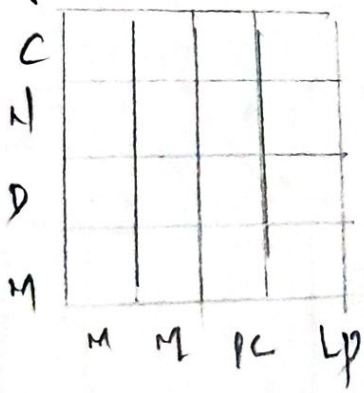
Dice



→ Here only we count the information for the (Q_1, Q_2) Quarters information as well as Mumbai & Delhi of particular product Mobile, Modem.

→ Some part of cube information we talking here that information by the overall cuboide.

→ 4. Slice :- Slice is cutting of the cube what ever information you need.



→ 'Q_1' Quarter information of the factories & product we count. in that just one slice of cube compact of information needed.