DMPA Chapter-1

117-248
18th Week

Thursday 27

## Data :-

It is a raw fact, information or statistics which can be in various forms like numbers, texts or any other form.

## Why Data Mining?

→ The explosive growth of data means production of data is too much in various areas. Due to increase in size of database, increase in computerized growth in the society, automatic analysis overcome on manual analysis which shows need of data mining.

1960 → Data collection and Database creation

1970 → Database Management System (SQL)

1980 → ~~RDBMS~~ RDBMS i.e advancement of DBMS

1990 - 2000 → Data Mining

## What is Data Mining?

→ It refers to extracting or mining knowledge from large datasets

→ It is also known as knowledge discovery from data.

KDD of data involves 7 steps :-

Essential Job to do Phone No.

Creation → Cleaning → Integration → Storing in warehouse → Selection → Transformation → Mining → Output as patterns

Saturday 29

Data Cleaning :- To remove the Noise and inconsistent data

Data Integration :- Where multiple data sources may be combined and stored in data warehouse.

Data Selection :- Where data relevant to the analysis are retrieved from database.

Data Transformation :- Where data are transformed into appropriate forms for mining by performing scaling operations.

Data Mining :- An essential process where intelligent methods are applied in order to extract data patterns.

Pattern Evaluation :- To identify interesting patterns, representing knowledge based on measures

30 Sunday

Knowledge Presentation :- Where visualisation and knowledge representation techniques are used.
Eg :- Pie Chart, Bar Chart, Area under Curve, etc.

## Architecture of Data Mining

(i) Data collected goes through cleaning, integration and selection. (Preprocessing)

Essential | Job to do | Phone No.

(ii) The preprocessed data is stored in warehouse server.

01 Monday

121-244
19th Week

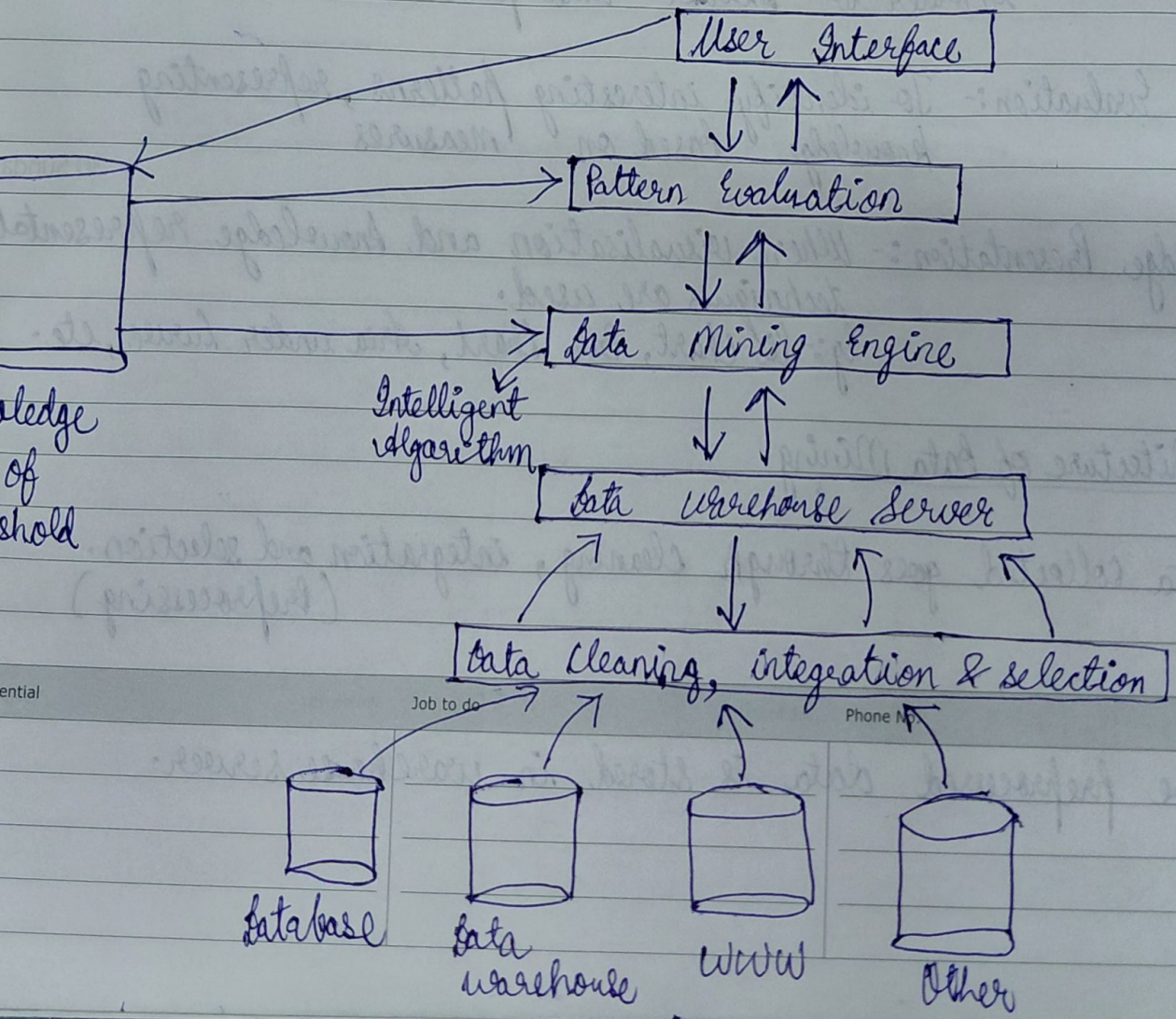
(iii) The intelligent algorithms are applied through data mining.

(iv) The mined patterns are evaluated, output is measured.

For finding out whether the output is accepted by user we require background knowledge.

The threshold in the extra knowledge, if greater than output, the output is acceptable by the user.
If threshold < output, output is not accepted.

Only the user can change the threshold value



122-243
19th Week

Tuesday 02

→ Database, Data Warehouse, WWW, other repositories

↳ These are set of database, datawarehouse, spreadsheet or other kind of information repositories.

→ The database or Data warehouse server is responsible for fetching the relevant data based on user requests.

→ Knowledge Base

↳ This is the background knowledge used to guide the search ~~engine~~ or evaluate the resulting patterns.

→ Data Mining Engine

↳ This consists of a set of modules for task prediction and evaluation analysis.

→ Pattern Evaluation

↳ This componet interact with data mining modules based on the search.

→ User Interface

↳ This module communicates between user and the data mining system. It allowes user to interact with system by specifying data mining query.

Essential | Job to do | Phone No.

03 Wednesday

123-242
19th Week

## Predictive Analysis

→ It is the process of extracting information from large dataset inorder to make prediction and estimates about future outcomes.

Eg:- Loan Default Prediction

Data Mining Knowledge – Analysing past records, the banks finds customers with high credit utilization and irregular ~~prediction~~ payment history are likely to default.

When a new customer applies for a loan the bank uses this pattern to credit, the likelihood of default.

Eg:- Health Care – predicting a patient prone to heart disease.

### What kind of Data?

→ Number of data repositories on which mining can be performed.

Repositories are:-

- Relational Database
- Data Warehouse
- Transactional Database
- Advanced Database System
  - → Object Oriented / Object Relational Database
  - → Spatial Database
  - → Temporal Database, Sequential Database
  - → Text Database, Multimedia Database
  - → Heterogeneous Database and Legacy Database
  - → Time Series Database
- World Wide Web

Essential Job to do Phone No.

124-241
19th Week

Thursday 04

## Relational Database

→ DBMS consists of database and set of software programs

→ The database consists of a collection of interrelated data

→ Software programs are used to manage and access the data, showed in database

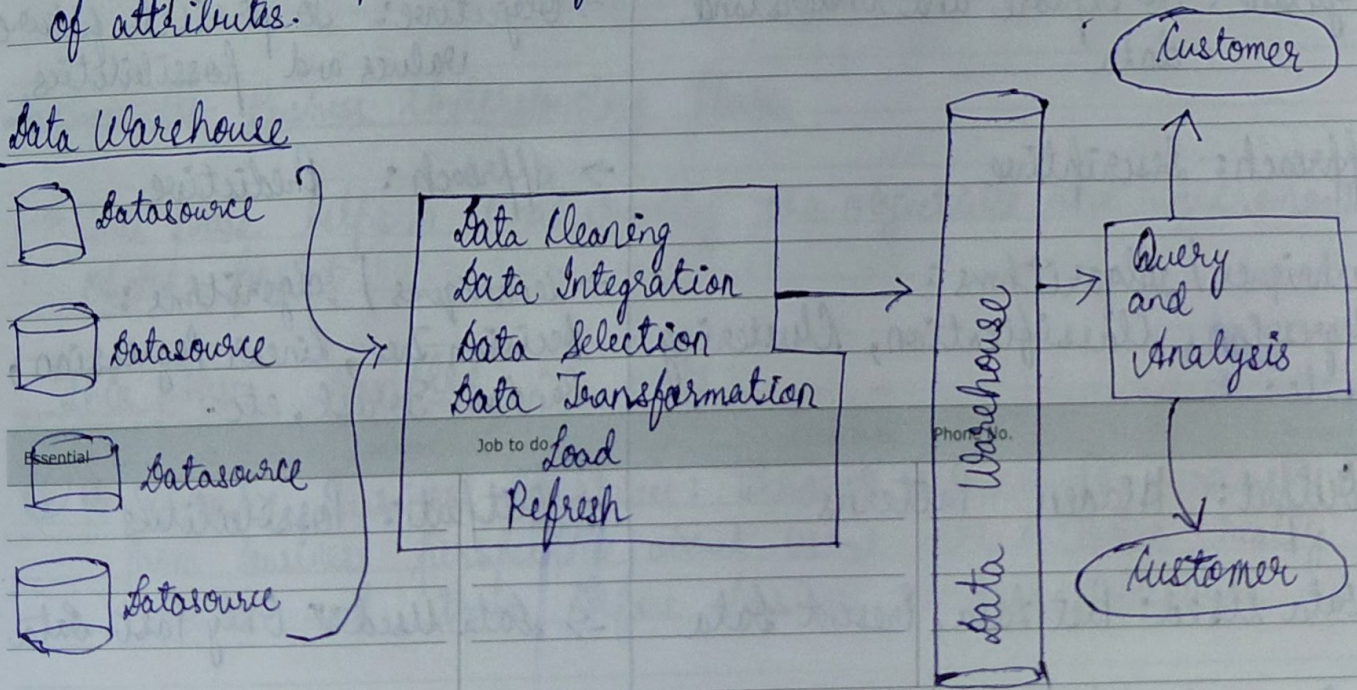
→ Software programs consists of database structure, data storing, distributing data access and for ensuring consistency and security of the information stored.

→ The Relational Database is a collection of tables, each table is assigned an unit name.

→ Each table, consists of set of attributes and large set of tuples.

→ Each tuple is represented by a unique key and described by a set of attributes.

## Data Warehouse



05 Friday

125-240
19th Week

→ A data warehouse is a repository of information collected from multiple sources stored under unified scheme and usually recides at a single point.

→ Data warehouse are constructed through a process of data cleaning, data integration, data transformation, data loading and periodic data refreshment refreshing.

→ In computing, data warehouse is also known as enterprised data warehouse (EDW). It is a system used for reporting and data analysis.

* Differentiate between Data Mining and Predictive Analysis

| Data Mining | Predictive Analysis |
|---|---|
| → It is a process of extracting knowledge from large database | → Applying the data mining knowledge or patterns to predict future outcome. |
| → Objective: To explore and understand data | → Objective: To predict future values and possibilities |
| → Approach: Descriptive | → Approach: Predictive |
| → Techniques / Algorithms: Regression, Classification, Clustering, etc. | → Techniques / Algorithms: Decision Tree, Linear Regression, Random Forest, etc. |
| → Output: Hidden Patterns | → Output: Possibilities |
| → Data Used: Past Data, Present Data | → Data Used: Only Past Data |
| → It is independent | → It is dependent |

Essential Job to do Phone No.

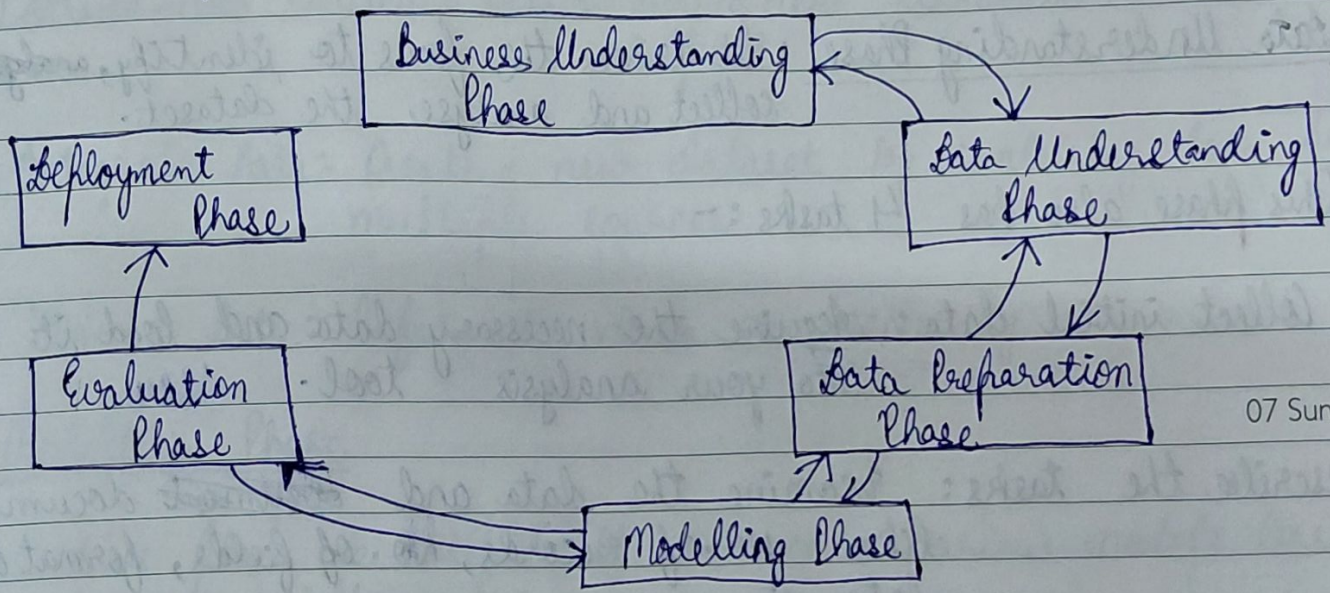
Saturday 06

## CRISP – DM :-

"Cross Industry Standard Process for Data Mining"

CRISP provides standardised data mining process across industries for fitting data mining into general problem solving strategy of a business or research unit.

It has 6 phases :-




07 Sunday

### ~~Business~~ Business Understanding Phase

→ This phase helps in understanding the objectives and requirements of the project.

This phase handles 4 tasks :-

Essential | Job to do | Phone No.

(i) Determine Business Objectives : ~~Through~~ Thoroughly understand from business perspective about what the customer really wants and then define business success criteria.

08 Monday



(ii) Access Situation: Determine Resources availaleility, project requirement, access contingency and conduct a cost benefit analysis.

(iii) Determine Data Mining Goals: Define technical Data Mining Perspective

(iv) Product Project Plan: Select technologies and tools and define detailed plan for each project phase.

Data Understanding Phase: It drives the focus to identify, ~~analyse~~ collect and analyse the dataset.

This phase also has 4 tasks:-

(i) Collect initial data: Acquire the necessary data and load it into your analysis tool.

(ii) Describe the tasks: Examine the data and ~~Document~~ document, like No. of records, No. of fields, format of data, etc.

(iii) Explore Data: Visualise and Identify relationship among data

(iv) Verify Data: Clear or first Data must have to be verified.

| Essential | Job to do | Phone No. |
| --- | --- | --- |
| | | |

129-236
20th Week

# Data Preparation Phase

Tuesday 09

It prepares the final dataset for modelling

Tasks:-

(i) Select Data: Choose the dataset that will be used

(ii) Clean Data: To correct or remove erronious data.

(iii) ~~Derive New~~ Construct Data: Derive new attribute that will be helpful

(iv) Integrate Data: Create a new dataset by combining data from multiple sources.

(v) Format Data: Reformat data as necessary

## Modelling Phase

→ In this phase we will build and access various models based on several modelling techniques.

Tasks:-

(i) Select Modelling Technique: Determine which algorithm you want to try

| Essential | Job to do | Phone No. |
|---|---|---|
| (ii) Generate Test Designs: | Needs to Split the data into training, testing and validation. | |

(iii) ~~Execute~~ Build Model: Execute few lines of code

(iv) Access Model: Multiple models are competing against each other to interpret model results based on domain knowledge.

## Evaluation Phase

It focuses on technical model assessment.

### Tasks:-

(i) Evaluate Results: Which one we approve for the business.

(ii) Review Process: Check the model is properly executed or not, summarise it and correct anything if needed.

(iii) Determine next Phase: Based on previous tasks, determine ~~what~~ whether to ~~deploy~~ proceed for deployment or not.

## Deployment Phase

A model is not particularly useful unless the customer can access ~~its~~ its result.

| Essential | Job to do | Phone No. |
|---|---|---|
| | | |
| | | |
| | | |

131-234
20th Week

Thursday

Tasks:-

(i) Plan Deployment : Develop and document a plan for deploying the model

(ii) Plan Monitoring and Maintenance - to avoid issues during the operation phase of model.

(iii) Produce final report - Document a summary of the project

(iv) Review Project - What went well, what would have been better, How to improve

Fallacies of Data Mining (Misconceptions)

(i) Data Mining tools are automated tools that can be deployed on data repositories to find and search our problems.

Reality: There are no automatic data Mining tools which will automatically solve your problems.

(ii) Data Mining process is autonomous requiring no human intervention

Reality: Without experts, blind use of data mining software will only provide wrong answers to wrong questions.

| Essential | Job to do | Phone No. |
|---|---|---|

(iii) Data Mining pays for itself quickly

Reality: Benefits take time. Costs include startup costs, data collection, data warehouse preparation costs, etc

12 Friday



(iv) Data Mining software packages are easy to use.

Reality : No ease of use varies it may require statistical and mathematical knowledge of particular application domain.

(v) Data Mining will identify the causes of other business problems or research problems.

Reality : It will help to show hidden patterns but human judgement and domain expertise is needed.

(vi) Data Mining will automatically clean up messy database

Reality : Still needs to handle missing values, outliers and inconsistency ~~manually~~ manually

(vii) Data Mining always provides positive results

Reality : Not Guarenteed. Poor data or assumption can lead to misleading outcomes.

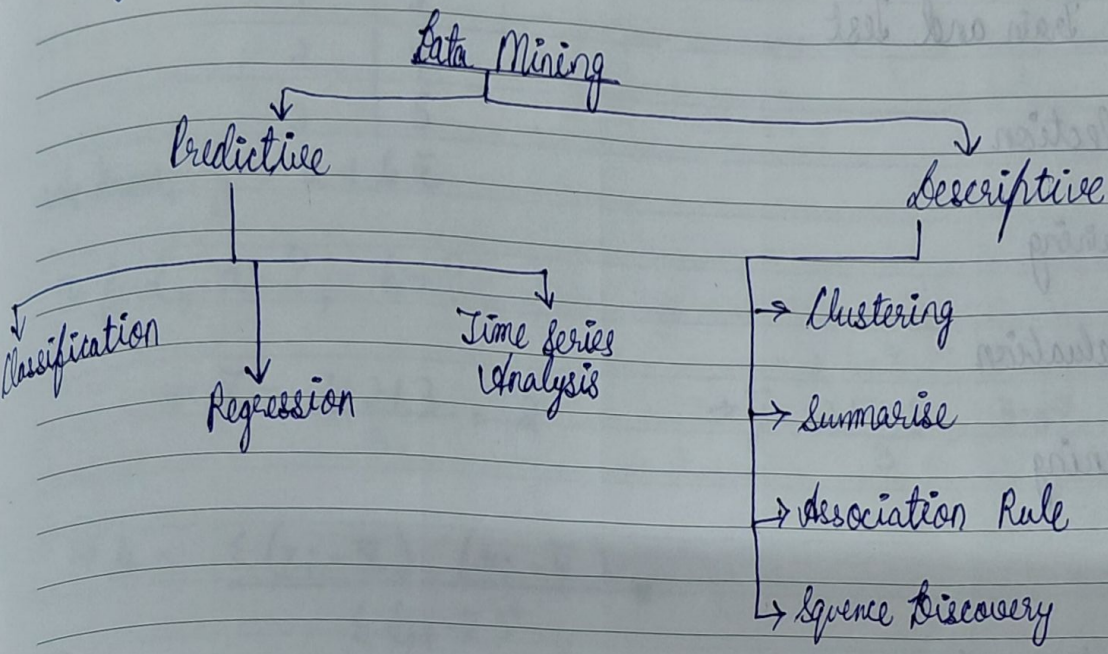
| Essential | Job to do | Phone No. |
| --- | --- | --- |
| | | |

133-232
20th Week

# Tasks of Data Mining

Saturday 13

Data Mining tasks involve finding patterns and useful information from large dataset.




## Classification

14 Sunday

→ It is a ~~pre defined~~ supervised learning technique in data mining that involves categorising or classifying data in predefined classes.

→ Categorising all groups based on features or attributes

→ In supervised learning, labelled data can be used to build a model that can predict the class of unseen data.

→ It is of two types Binary Classification and Multiclassification

Essential | Things to do | Phone No

15 Monday

135-230
21st Week

Steps to build classification:

(i) Data Preparation
(ii) ~~Model~~ ~~Feture~~ Feature Selection
(iii) Split for Train and Test
(iv) Model Selection
(v) Model Training
(vi) Model Evaluation
(vii) Model Tuning
(viii) Learning
(ix) Model Deployment

## Regression (Estimation)

→ It is a supervised learning technique used to predict a continuous numerical value by analysing past data.

Types:-

(i) Linear Regression (ii) Logistic Regression (iii) Polynomial Regression

Essential | Job to do | Phone No.

(iv) Lasso Regression (v) Ridge Regression

136-229
21st Week

Tuesday 16

Q) We have a dataset Hours studied (x) to predict Exams(y)
We need to find y if x = 4

| x | y |
|---|---|
| 1 | 2 |
| 2 | 4 |
| 3 | 5 |

We know, $\hat{y} = a + b\bar{x}$

To find, $a = ?$, $b = ?$

$\rightarrow \bar{x} = \frac{1+2+3}{3} = 2$ $\quad\quad$ $\rightarrow \bar{y} = \frac{2+4+5}{3} = 3.67$

$$\rightarrow b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Now,

| x | y | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})(y_i - \bar{y})$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|
| 1 | 2 | -1 | -1.67 | 1.67 | 1 |
| 2 | 4 | 0 | 1.67 | 0 | 0 |
| 3 | 5 | 1 | 2.67 | 2.67 | 1 |

$\therefore b = \frac{1.67 + 2.67}{1+1} = \frac{3.67}{2} \rightarrow b = 1.5$

$a = \bar{y} - b\bar{x} = 0.67$ $\quad\quad$ Test Data $\Rightarrow \bar{y} = 0.67 + 1.5x$

| Essential | Job to do | Phone No. |
|---|---|---|
| | So, For $x = 4$, | $y = 6.67$ |

17 Wednesday



## Prediction

→ This is similar to classification and estimation except for prediction, the result lie in future.

Eg:- ~~Data~~ Predicting prices of stock 3 months into the future.

## Time series analysis

→ It is a way of analysing sequence of data points connected over an interval of time.

It is composed of 4 main elements:-

(i) Trend: Long term movement of data

(ii) Seasonality: A predictable recurring patterns of repeatition over a fixed regular interval.

(iii) Cycle: Long term Movement or fluctuation

(iv) Irregularities: Random or unpredictable movement in data

Clustering: It means grouping of records, observations into a similar object.

This is mainly used ~~for~~ in unsupervised learning. The data points having less distance belongs to same cluster and having more distance belongs to different clusters.

138-227
21st Week

Clustering segments the whole dataset into relatively homogeneous subgroups.

Thursday 18

**Summarisation** : It is the process of reducing large dataset into shorter, more understandable format that highlight key patterns, trends and relationships.

**Types :-**

↳ Descriptive : Uses statistical measures to describe main features of numerical data.

Eg :- Mean, Median, Mode

↳ Aggregation : Combining data into simpler summarised form like sum, average, etc.

↳ Sampling : Select a representative subset of the data.

Association : Association for data mining is the job of finding which attributes go together.

Eg :- Market basket analysis

This rule shows how frequently an item set occurs in a transaction.

Sequential Discovery : It is a data mining technique used to identify frequently occuring patterns in sequential data such as purchase history.

MAY

19 Friday

2023

APPENDIX : Data Summarisation and Visualisation

| Applicant | Martial Status | Mortgage | Income | Rank | Year | Risk |
|---|---|---|---|---|---|---|
| 1 | Single | Y | 38K | 2 | 2009 | Good |
| 2 | Married | Y | 32K | 7 | 2010 | Good |
| 3 | Other | N | 25K | 3 | 2011 | Good |
| 4 | Other | N | 36K | 3 | 2009 | Good |
| 5 | Other | Y | 33K | 4 | 2010 | Good |
| 6 | Other | N | 24K | 10 | 2007 | Bad |
| 7 | Married | Y | 25K | 8 | 2009 | Good |
| 8 | Married | Y | 48K | 1 | 2010 | Good |
| 9 | Married | Y | 32K | 6 | 2011 | Bad |
| 10 | Married | Y | 32K | 5 | 2009 | Good |

Varialeles

ualitative (Categorical)

minal

Ordinal

Quantitative (Numerical)

Interval

Ratio

Discrete Value

Continous Value

Essential

Job to do

Phone No.

140-225
21st Week

20 Saturday

Qualitative :- Describes qualities or categories of data. It is also called categorical variables.

Eg :- Marital Status, Mortgage, Rank, Risk.

Quantitative :- Takes numerical values and allows arithmetic operation. It is also called numerical variables.

Eg :- Year, Income.

Nominal :- Used for names, labels or categories without order

Eg :- Marital Status, Mortgage, Risk.
(Blood Groups)

Ordinal :- Maintains a particular order Eg :- Rank

Interval :- It is a quantitative data defined on an interval without a natural zero (0). 21 Sunda

Eg :- Year

Ratio :- Quantitative data for which mathematical operations can be performed (can contain 0)

Eg :- Income

| Essential | Job to do | Phone No. |
| --- | --- | --- |
| | | |
| | | |
| | | |

22 Monday

142-223 22nd Week

Discrete Values :- It is a quantitative variable that can take finite or countable number inclusing zero.

→ It cannot take values in between two numbers
→ Gaps between values

→ Graphical Representation is Bar Chart.

Eg :- Year

Continuous Values :- It is a quantitative variable that can take any value within a given range.

→ No Gaps between the points. → Eg :- Income

→ Graphical Representation - Histogram

Predicter Variable : It is a variable whose value is used to predict the value of response variable

The predicter variable in table are all variable except Risk.

Response Variable : It is also called dependent variable or output variable or target variable which try to predict, explain or measure the effect.
↳ In table Example : Risk

Eg :- Effect of 'x' hours studies on exam score 'Y'.

Predicter = X
Response = Y.

Essential | Job to do | Phone No.

143-222
22nd Week

# Measurement of center, variability and position

Tuesday 23

→ Measures of Central tendency - It measures the location of middle or center of a data distribution. Eg:- Mean, Median, Mode and Midrange

**Mean -** It is an arithmetic average. The sum of all data values divided by the number of values.

Eg:- Dataset → 2, 4, 6, 8, 10

$$\text{Mean} = \frac{2+4+6+8+10}{5} = 6$$

Advantage :- Easy to compute including all data points

Disadvantages :- Sensitive to outliers

**Median -** The middle value when the data is sorted.

Eg:- 3, 5, 7, 9, 11

Median = 7

3, 5, 7, 9

$$\text{Median} = \frac{5+7}{2} = 6$$

Advantage: Not affected by outliers

**Mode** - The value that appears most frequently in the dataset.

It is of 3 types: Unimodal, Bimodal and Multimodal

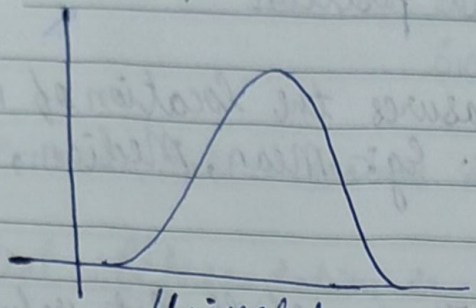
No Mode - If all values occur with same frequency

Eg:- 2, 3, 3, 4, 4, 5, 5, 8, 8, 8, 10, 12

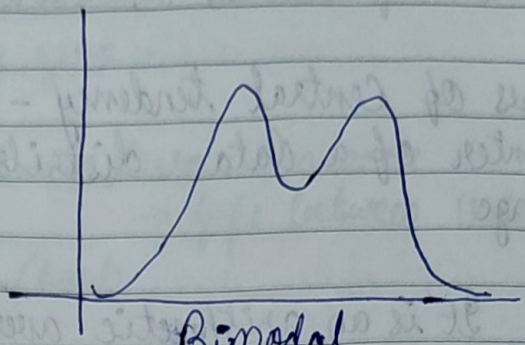
Mode = 8

24 Wednesday

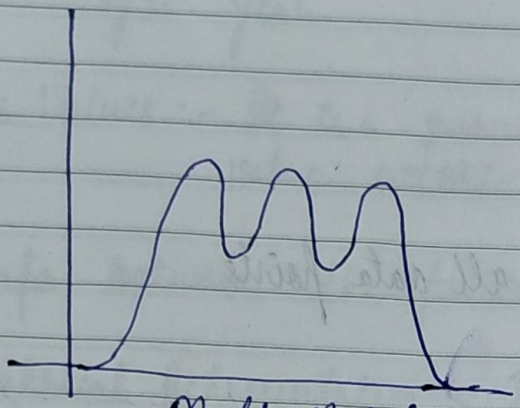




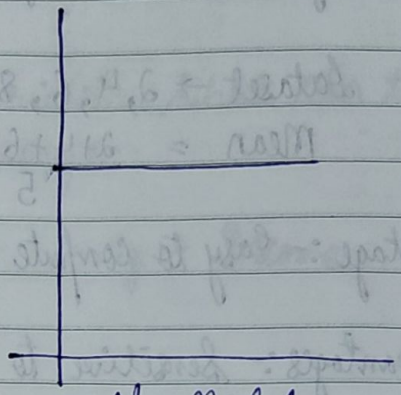
Unimodal



Bimodal



MultiModal



No Modal

Advantage :- It is used mostly for categorical data

Midrange - It is the average of largest and smallest value in dataset. It is applicable for numerical data.

| Measure | Best Used | Sensitive to outliers | Datatypes |
|---|---|---|---|
| Mean | Numerical data and Symmetric data | Yes | Interval / Ratio |
| Median | Skewed data or weak outliers | No | Ordinal and Interval / Ratio |
| ~~Mode~~ Mode ~~Mode~~ | Most frequent item needed | No | Nominal, ordinal, Interval / ~~Ratio~~ Ratio |

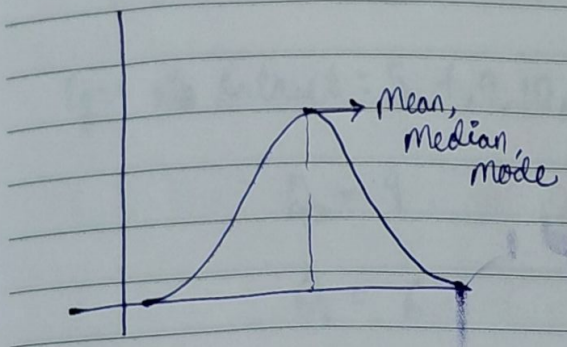
145-220
22nd Week

Thursday 25

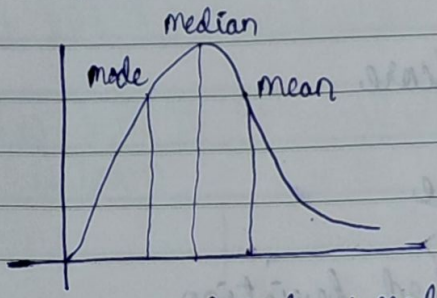
## Measuring Dispersion of data

In unimodal, frequency curve with perfect symmetric data distribution - mean, median and mode are all at same center value.

But, in reality, data is not symmetric.

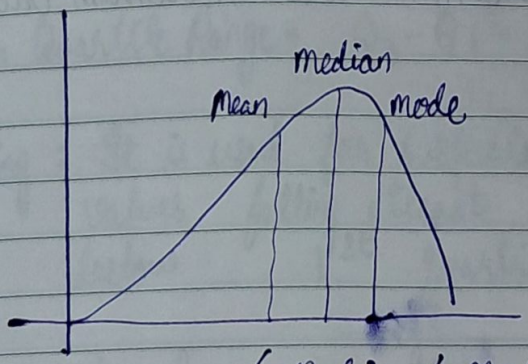



Positive Skewed:




(mode < median)

Negative Skewed:




(Median < Mode)

| Essential | Job to do | Phone No. |
|---|---|---|
| | | |
| | | |
| | | |

26 Friday



## Measurement of ~~variability~~ dispersion (variability)

(i) Range
(ii) Quartiles ($Q_1, Q_2, Q_3$)
(iii) Inter quartile range (IQR) → $Q_3 - Q_1$
(iv) 5 Number Summary ($Q_1, Q_2, Q_3$, min, max)
(v) Box Plot
(vi) Z Score
(vii) Variance
(viii) Standard Deviation

Range : It is the difference between maximum and minimum value in dataset.

Eg:- Dataset : 5, 7, 9, 10, 12

Range = 12 - 5 = 7

| Essential | Job to do | Phone No. |
| --- | --- | --- |
| | | |

147-218
22nd Week

Saturday 27

Quartile :- It divides ordered data into 4 equal parts

Q1 → First quartile, which represent 25% ile of data below it

Q2 → Second quartile, known as median, 50% ile of data below it

Q3 → Third quartile, which present 75% ile of data below it

Eg :- Dataset : 5, 7, 9, 10, 12

$Q_2 = 9$

$Q_1 = 6$

$Q_3 = 11$

28 Sunday

Inter Quartile Range = $Q_3 - Q_1 = 5$

Capping : It is used for suspected outliers, common rule is single out values falling atleast 1.5 X IQR (capping) above 3rd quartile or below 1st quartile.

Five Number Summary : It is a concise statistical summary of data consisting of min, $Q_1$, $Q_2$, $Q_3$, max

Dataset : 5, 7, 8, 12, 13, 14, 18, 21, 23, 25

Essential | Job to do | Phone No.

Min = 5 | Max = 25

$Q_2 = 13.5$ | $Q_1 = 8$ | $Q_3 = 21$

# Box Plot (Whisker Plot)

→ It is a graphical representation of 5 Number summary of a dataset.

→ Features:

(i) Box: Extends from $Q_1$ to $Q_3$

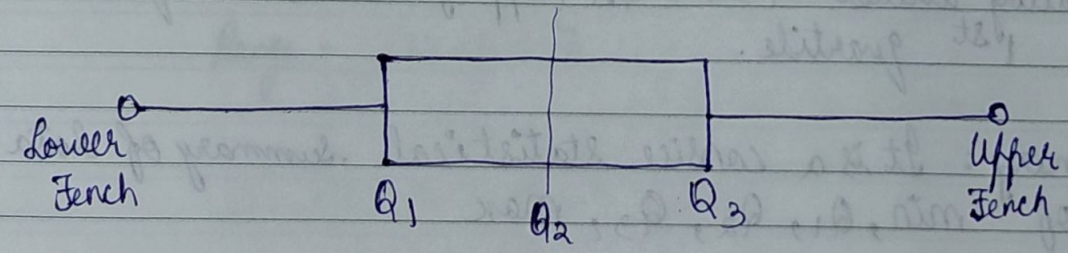
(ii) Median Line: A line inside the box shows median

(iii) Whisker: A line extending from $Q_1$ to minimum and $Q_3$ to maximum.

## Outliers

→ Data points that lie beyond lower fench and upper fench

Lower fench = $Q_1 - 1.5 \times IQR$
Upper fench = $Q_3 + 1.5 \times IQR$




Box Plot

150-215
23rd Week

30 Tuesday

Q) In dataset 11, 7, 35, 55, 64, 86, 88, 90, 95, 97

Sorted = 7, 11, 35, 55, 64, 86, 88, 90, 95, 97

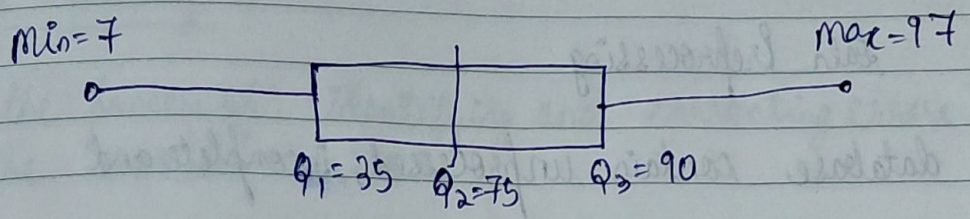
min = 7
max = 97
$Q_2$ = 75
$Q_1$ = 35
$Q_3$ = 90

IQR = 55

Lower = −47.5
Upper = 172.5

Capping = 1.5 × 55 = 82.5

So, No outliers




## Z Score

→ Z score measures how many standard deviations a data point is from the mean. It standadises data making values from different distribution comparable.

$$Z = \frac{x - \mu}{\sigma}$$

If Z = 0 → Value is exactly at mean

If Z < 0 → Value is ~~above~~ below mean

If Z > 0 → Value is above mean

→ If datapoint with |Z| > 3 are considered outliers

## Variance

→ It measures the average square deviation from the mean

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

## Standard Deviation

→ It is the square root of variance $\sqrt{\sigma^2}$ = SD

## Chapter-2 Data Preprocessing

→ Raw Data in database contain unprocessed, incomplete and noisy data.

→ Data Preprocessing is the process of preparing raw data for analysis by cleaning and transforming it into a usable format.

→ Real dataset often have missing or incomplete entries. Data may contain typo error, duplicates or outliers.

→ Different features may have different scale not suitable for data mining models.

→ Some features are irrelevant, redundant or highly corelated.



→ Values are not consistent with the policy.

152-213
23rd Week

Thursday 01

## Steps of data processing:-

(i) Data Cleaning

(ii) Data Integration

(iii) Data Transformation

(iv) Data Reduction

## Data ~~Clean~~ Cleaning

→ It is the process of identifying and correcting errors and inconsistencies in the ~~world~~ dataset.

→ It involves handling missing values, removing duplicates and correcting incorrect or outlier data to ensure the dataset is accurate.

→ Clean data is ~~of~~ essential for effective analysis as it improves the quality of results and enhance the performance of data models.

## Missing values

→ When a data is absent in a dataset is called missing values.

## Handling of Missing Values

| Essential | Job to do | Phone No. |
|---|---|---|
| a) Ignore the tuples or | records : This is usually done when class label is missing. | |

This method is not very effective unless the tuple contains several attributes with missing values.

b) Fill in the missing value manually

This method is time consuming may not be feasible for large dataset.

c) Use a global constant to fill in the missing values

d) Use the attribute mean to fill in the missing value.

e) Use the attribute mean for all the samples belonging to the same class as the given tuple.

f) Use the most prefferable, probable value to fill in the missing value.

Eg:- Desicion Tree, Regression ..

## Noisy Data

→ Noise is a random error or variance in a measured variable.

→ Incorrect attribute values may be due to faulty data, data entry problem, data ~~tran~~ transmission problem, technology limitations, inconsistency in naming convention.

→ Other problems include duplicate record, incomplete data

154-211
23rd WEEK

# Binning Method

Saturday 03

Sort the data and the sorted values are distributed into a number of buckets or bins. Then these bins can be smoothen up by mean, median and boundaries.

## Bin Mean

↳ Each value in bin is replaced by mean value of the bin

## Bin Median

↳ Each bin value is replaced by the bin median

## Bin boundary

↳ Maximum and Minimum values in a given bin are identified as bin boundaries.

↳ Each bin value is then replaced by the closest boundary value.

04 Sunday

## Equal Width Partitioning

→ It divides the range into N intervals of equal size.

→ If A and B are lowest and highest value of the attribute the width of the interval will be $\frac{B-A}{n} = W$

| Essential | Job to do | Phone No. |
|---|---|---|
| This is → the most straight presentation. | forward method but | outliers may dominate |

05 Monday

## Equal Depth Partitioning (Equal Frequency Partitioning)

→ It divides the range into N intervals each containing approximately same number of samples.

## Good Data Scaling

↳ Managing Categorial Attributes

Q) Sorted data for price is 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

## Smoothing By Bin Means

Mean = 9 22.75 29.25

Bin-1 = 9, 9, 9, 9

Bin-2 = 22.75, 22.75, 22.75, 22.75

Bin-3 = 29.25, 29.25, 29.25, 29.25

## Smoothing By Medians

Bin-1) 8.5 Bin-2) 22.5 Bin-3) 29.25

## Smoothing By Boundaries

Bin-1) 4 Bin-2) 21 Bin-3) 34

| Essential | Job to do | Phone No. |
|---|---|---|
| | | |

157-208
24th Week

Tuesday 06

## Equal Width Partitioning

Step-1
Min = 4
Max = 34
Range = 34 - 4 = 30

Step-2
Width = Interval = $\frac{\text{Range}}{\text{No. of Bins (default 3)}} = \frac{30}{3} = 10$

Step-3
Bin-1 [4, 4+10] = [4, 14] = 4, 8, 9

Bin-2 [14, 14+10] = [14, 24] = 15, 21, 21, 24

Bin-3 [24, 34] = 25, 26, 28, 29, 34

## Equal Depth / Frequency partitioning

Step-1 Data must be sorted

Step-2 Take 3 Bins. Total value in dataset is 12. So, each bin should have 4 values

Step-3
Bin-1 > 4, 8, 9, 15
Bin-2 > 21, 21, 24, 25
Bin-3 > 26, 28, 29, 34

07 Wednesday

158-207
24th Week

## Regression

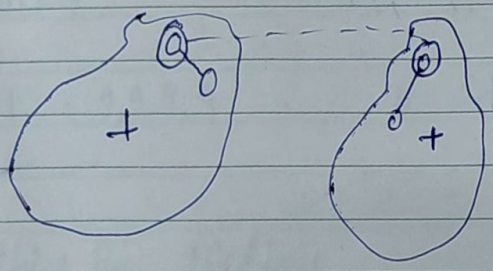
-> In Regression, data can be smooth by fitting the data to a function.

-> Two types :-

(i) Linear Regression

(ii) Multiple Linear Regression

## Clustering



## Miss classification

-> It means a model predict the wrong class level compared to the actual or true class level in your dataset

Dataset

| | Brand | Frequency |
|---|---|---|
| US X | USA | 1 |
| Europe | France | 1 |
| | US | 156 |
| | Europe | 46 |
| | Japan | 51 |

Essential | Job to do | Phone No.

Thursday 08

→ In frequency distribution dataset, having 5 classes however two of the class USA and France have count only one each which is clearly happening that two of the records have been inconsistently classified (misClassification) with respect to original data.

→ To maintain the consistency, record have been labelled US and Europe instead of USA and France.

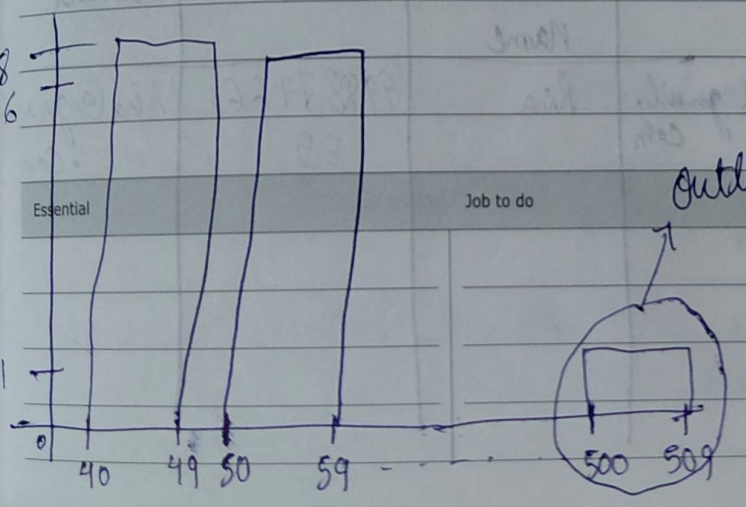
## Graphical Methods for identifying outliers

Outliers are extreme values that go against the trend of remaining data.

Graphical Methods for identifying outliers in numerical variable are Histogram and Scatterplot.

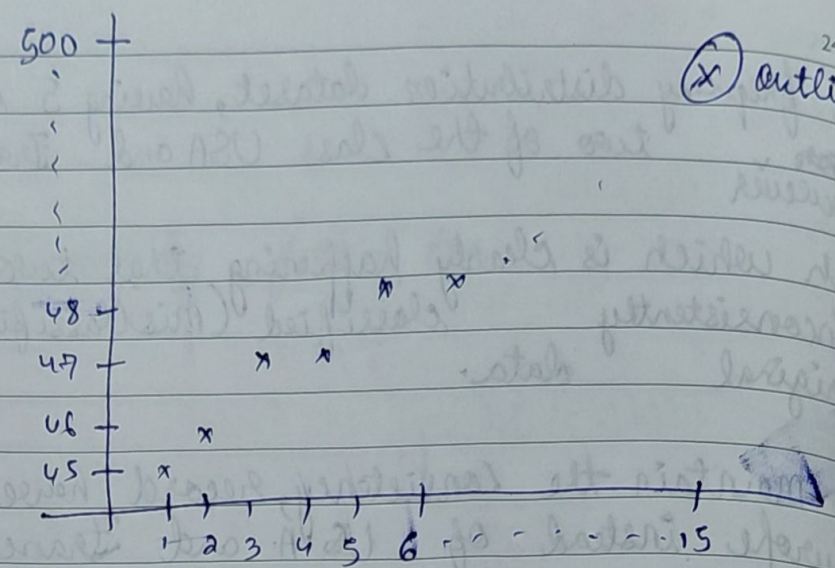
Dataset: 45, 50, 52, 48, 47, 49, 51, 46, 500, 50, 48, 49, 52, 47, 51

Sorted: ~~45, 46, 47, 48, 49, 50, 50, 51~~ 45, 46, 47, 47, 48, 48, 49, 49, 50, 50, 51, 51, 52, 52, 500



09 Friday

1-45
2-46
3-47
4-47
5-48
6-48
:
:
:
15-500




(x) outlier

## Data Integration

→ It is the merging of data from multiple data sources.

It helps to reduce and avoid redundancy and inconsistency in the resulting dataset which improve the accuracy and speed of the ~~subsequent~~ subsequent data mining process.

## Entity Identification Problem

Eg:- Employee Details

| Employee name | Mobile No | Email ID |
|---|---|---|
| Rina | 9988776655 | rina@gmail.com |

Customer Details

| Customer Name | Mobile No | Email |
|---|---|---|
| Rina | 9988776655 | rina@gmail.com |

Essential | Job to do | Phone No.

161-204
24th Week

Saturday 10

→ It occurs when we try to determine whether two or more records from different data subsets refer to the same real world entity.

Solution

→ Rule based Matching

→ Probabilistic Matching

→ Machine Learning Algorithms

Data Transformation

→ It is the process of converting data from one format or structure to another format or structure to ensure compatibility, consistency and better quality of analysis.

11 Sunday

Types

↳ Smoothing (Binning: Mean, Median, Boundary), ~~Regressing~~ Regression Clustering): It is used to remove the noise from data

↳ Aggregation - It summarises the data. Eg:- Average, Mean, Sum

↳ Generalization - Replacing detailed data with higher level concept

Essential | Job to do | Phone No.

Eg:- Age of Student → 23 → Youth Adult

↳ Normalization - Rescaling ~~the~~ the values to a standard range i.e from 0-1.

12 Monday

↳ Attribute Construction: Create a new attribute from the existing data. Eg:- Use DOB to calculate age.

## Min - Max Normalization (Scaling) Standard Range - (0-1)

→ It is a linear transformation on the original data

$$X_{mm} / X^* = \frac{X - \text{Min value of } X}{\text{Range (Max - Min) of } X}$$

where X is the original data, Min of X is minimum value of the attribute, Max of X is maximum value of attribute, $X_{mm} / X^*$ is normalised value

Q) You have a dataset 50, 60, 70, 80, 90, 100. Min Max Normalisation for 80.

$$X^* = \frac{80 - 50}{100 - 50} = 0.6$$

Q) Minimum and Maximum value for the attribute Income are 12000 and 98000 respectively. We like to map income to the range 0-1 by min max Normalisation of a value of 73600. Calculate transformed value

$$X^* = \frac{73600 - 12000}{98000 - 12000} = \frac{61600}{86000}$$

$$= 0.716$$

$$X^* = \frac{61600}{50000} = 1$$

Mid Range = $\frac{\text{Max value} + \text{Min value}}{2}$

Normalised value of Mid Range

$$X^* = \frac{50000 - 12000}{98000 - 12000} = \frac{38000}{86000} = 0.5$$

## Z score Standardisation

$$Z\text{-Score} = \frac{X - \text{Mean}(X)}{SD(X)}$$

## Decimal Scaling

Standard Range (-1 to 1)

→ It is a Normalised technique where Normalized value lies between -1 to +1.

$$X_{decimal} = \frac{X}{10^d}$$

where d = No. of digits in the data value with the largest absolute value

Q) -987, -120, 56, 300, 850 : Dataset

$$X_{decimal} = \frac{-987}{10^3} = -0.987$$

→ -0.987, -0.120, 0.56, 0.300, 0.850

| Essential | Job to do | Phone No. |
|---|---|---|
| | | |

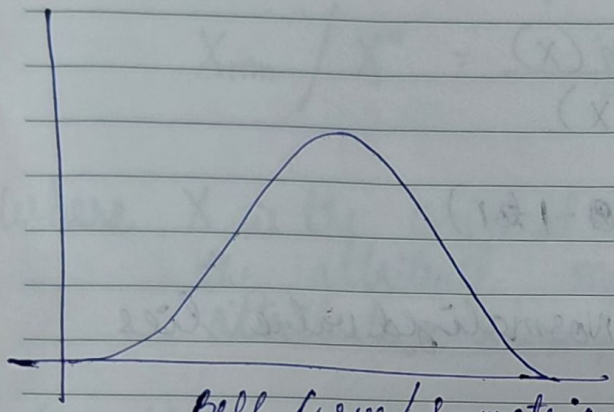
14 Wednesday

## Transformation to achieve Normality

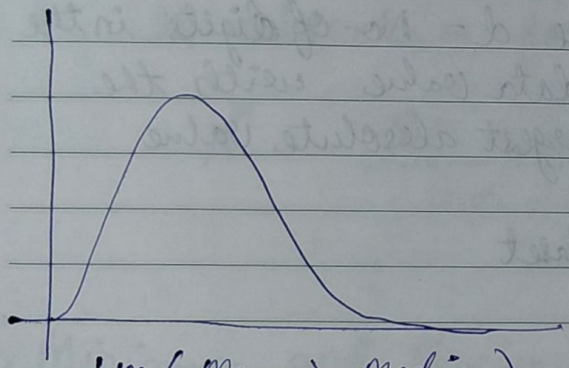
→ Normal distribution is a continuous probability distribution commonly known as Bell curve which is a symmetric curve.

Condition of Bell curve - SD = 1, Mean = 0

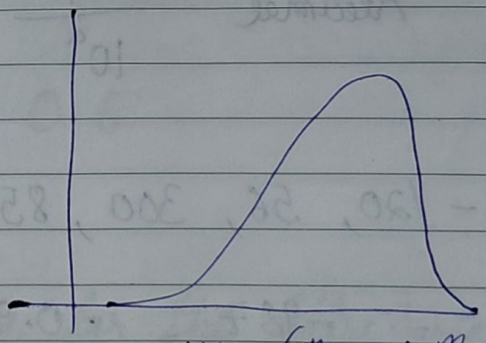
Symmetric curve - Mean, Median, Mode are peak of curve



Bell curve / Symmetric curve



+ve (Mean > Median)



-ve (Mean < Median)

$$\text{Skewness} = \frac{3(\text{Mean} - \text{Median})}{SD}$$

| Essential | Job to do | Phone No. |
|---|---|---|
| | | |

166-199
29th Week

Thursday 15

## Methods to remove Skewness

1. Log Transformation → $\bar{x} = \log(x+1)$
2. Square Root Transformation → $\bar{x} = \sqrt{x}$
3. Inverse Square Root Transformation → $\bar{x} = \frac{1}{\sqrt{x}}$

## Flag Variables

→ These are dummy / indicator variables which converts categorical variables into only two values (0 or 1).

→ Eg:-

Categorical Variable: Gender (Female, Male)

If Gender =: Female then Gender_Flag = 0

If Gender =: Male then Gender_Flag = 1

Flag =: If region = North then North_flag = 1 otherwise North_flag = 0

South_flag =: If region = South then South_flag = 1 otherwise South_flag = 0

| Essential | Job to do | Phone No. |
|---|---|---|

East_flag =: If region = East then East_flag = 1 otherwise East_flag = 0

Otherwise west_flag = 0

16 Friday



Categorical values must be converted into numerical format for machine learning algorithms :-

① Label Encoding

Eg:- Size — Large - 0
— Medium - 1
— Small - 2

② One Hot Encoding

Eg:- Colors — Red
— Green
— Blue

| Red | Green | Blue |
|---|---|---|
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |

③ Binary Encoding

Eg:- Binary Encoding — Bhubaneswar @ 1 → 0001
— Cuttack @ 2 → 0010
— Rourkela @ 3 → 0011

| Essential | Job to do | Phone No. |
|---|---|---|
| | | |

168-197
25th Week

# DMPA Questions

Saturday 17

Q) Suppose that a hospital tested the age and body fat data for some randomly selected adults.

| age | 23 | 23 | 27 | 27 | 39 | 41 | 47 | 49 | 50 |
|---|---|---|---|---|---|---|---|---|---|
| % fat | 9.5 | 26.5 | 7.8 | 17.8 | 31.4 | 25.9 | 27.4 | 27.2 | 31.2 |

a) Calculate Mean, Median, Mode and SD of age and % fat

b) Transform above numeric attribute using Z-score Normalization.

## age

Mean = 36.22

Mode = 23, 27 (Bimodal)

Median = 39

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

$$= \left(\frac{1}{9} \times 1019.5556\right)^{1/2}$$

$$= \sqrt{113.2839} = 10.64$$

## % fat

Mean = 22.74

Mode = NoModal

Median = 26.5

18 Sunday

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

$$= \frac{1}{9} \times 635.2024$$

$$= \sqrt{70.5780} = 8.401$$

| Essential | Job to do | Phone No. |
|---|---|---|
| | | |

19 Monday

170-195
26th Week

$$z\text{-score} = \frac{x - \mu}{\sigma}$$

z-score

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| age | -1.242 | -1.242 | -0.866 | -0.866 | 0.261 | 0.449 | 1.010 |
| % fat | -1.576 | 0.447 | -1.778 | -0.588 | 1.030 | 0.376 | 0.554 |

| | | |
|---|---|---|
| | 1.201 | 1.295 |
| | 0.530 | 1.007 |

(c) Calculate correlation coefficient $r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}}$

$= 0.71$

Q) BINNING

Sales Records : T = {50, 55, 65, 72, 11, 13, 15, 35, 92, 108, 150, 5, 8, 10, 187, 204, 210, 215}

Partition them into 4 bins and apply smoothing by bin boundaries

a) Equal width b) Equal depth

Essential | Job to do | Phone No.

Tuesday 20

Sorting: 5, 8, 10, 11, 13, 15, 35, 50, 55, 65, 72, 92, 108, 150, 187, 204, 210, 215

Min = 5, Max = 215, Range = ~~200~~ 210

Width = $\frac{210}{4}$ = 52.5

Bin 1 [5, 57.5] = [5, 8, 10, 11, 13, 15, 35, 50, 55, ~~65, 72,~~]

Bin 2 [57.5, 105] = [62, 72, 92]

Bin 3 [105, 155] = [108, 150]

Bin 4 [155, 205] = [187, 204, 210, 215]

Equal depth } Take 4, 4, 5, 5

## Transforming Categorical variable into Numerical Variables

Eg:- Region (East, West, North, South)

| Categorical Variable | Numeric Value |
|---|---|
| East | 1 |
| West | 2 |
| North | 3 |
| South | 4 |

Relation

1 < 2 < 3 < 4

East < West < North < South

21 Wednesday



→ This simply transforms categorical variable region into a single numerical variable rather than using several flag variables.

## Reclassifying Categorical Variable

→ Reclassifying Categorical Variables means combining regrouping or transforming categorical variables into fewer and more meaningful groups to improve data analysis or model performance.

→ When a categorical variable has too many categories or irrelevant categories, we merge or rename them to make analysis easier.

Reason - (i) Too many categories exist. Some categories have very few observations.

(ii)

(iii) Several Categories have similar meaning

(iv) To create more useful or predictive grouping for data mining algorithms.

Eg:- 50 states can be reclassified on the basis of economic-level as Richest, Middle Range and Poor

| Essential | Job to do | Phone No. |
| --- | --- | --- |
| | | |

173-192
26th Week

Thursday 22

## Adding an Index Field

→ An Index field is a new variable / column added to the dataset to uniquely identify each record.

→ It is especially useful when the dataset doesnot contain unique key or id.

Why required?

→ To uniquely identify each record, data cracking, debugging, joining dataset

Eg:-

| Index | Name | Age | Salary |
|---|---|---|---|
| 1 | Rina | 32 | 20K |
| 2 | Aditya | 55 | 50K |
| 3 | Raj | 15 | 70K |
| 4 | Malala | 22 | 55K |

## Removing variables that are not useful

→ Some variables in the dataset may not contain useful information for data mining task, these variables should be identified and removed to simplify the dataset and improve performance of the model.

Reason

(i) To reduce the noise (irrelevant data)

(ii) To improve computation

(iii) Simplify interpretation

(iv) Avoid redundancy

23 Friday



Eg:- Unary Variable: It is a variable / column that has only one value for all the records in a dataset.

Eg:- Table

| Index | Name | Age | Salary | Country |
|---|---|---|---|---|
| | | | | India |
| | | | | India |
| | | | | India |
| | | | | India |

Unary Variable

Nearly Unary Variable

| Country |
|---|
| India |
| India |
| India |
| US |

Nearly Unary Variable

Variables that should probably not be removed

Some variable that appear unimportant at first but later turn out to be quite useful

Types of variables

→ Categorical identifiers that group records meaningful

| Essential | Job to do | Phone No. |
|---|---|---|
| → Variables with | potential interaction | effect |
| → Variables with | missing / rare values | |

# Chapter-3 Exploratory Data Analysis (EDA)

Saturday 24

→ There are two distinct approaches to data analysis :-

① Hypothesis Testing :

~~Testing~~ test a confirmatory formal procedure that a pre specified idea or assumption.

② EDA (Exploratory Data Analysis) - It is an open ended discovery oriented process where the role is to learn what the data suggest without a fixed hypothesis.

Both approaches play a complementary role in data mining, 25 Sunday statistics and machine learning.

## HYPOTHESIS TESTING

→ It is a statistical method used to make decision / draw conclusion about a population based on sample data.

→ It helps to determine whether there is enough evidence to support or reject a particular belief (Hypothesis).

| Essential | Job to do | Phone No. |
|---|---|---|
| | | |

26 Monday

Key features :-

(i) It starts with a prior hypothesis or assumption (before examining the data in detail)

(ii) It involves two ~~complimenting st~~ competing statements

$H_0$ = Null Hypothesis

$H_1$ = Alternative Hypothesis
or $H_A$

Example :- A mobile phone operator takes a hypothesis :

$H_0$ = Market share has not decreased after fee hike.

$H_1$ = Market share has decreased after fee hike.

## Exploratory Data Analysis

→ It is an approach to analyse dataset that emphasises visual exploration and descriptive statistics to uncover patterns, anomalies and relationship without predetermining assumptions.

Essential

## Objectives

Job to do

Phone No.

(i) Understanding the structures of data - Eg :- Variables, Range

(ii) ~~Variables~~ Distribution of Categorical variables

Tuesday 27

(iii) Look at the histograms of numeric variables to understand their spread and shape.

(iv) Explore the relationship among sets of variables (Predictor and Target Variables)

(v) Detect Outliers, missing values and data quality issues.

(vi) Develop initial hypothesis and guide subsequent modelling.

Common EDA techniques

-> Graphical - Histogram, Scatter Plot, Box Plot, ...

-> Numerical - Summary Statistics (Mean, Median, Mode, Variance, Skewness), correlation coefficient

-> Subset / Group Analysis -

Identifying clusters, trends or interesting subsets.

* EDA acts as foundation of data analysis shaping the direction of further investigation and hypothesis testing.

| Essential | Job to do | Phone No. |
| --- | --- | --- |
| | | |

28 Wednesday

## Complementary Roles

→ EDA comes first (discovery stage)

(→ Helps analyst understand the dataset, distribution and uncover important relationships and patterns that could indicate important areas for further investigation.

→ Hypothesis testing follows (Confirmation Stage)

(→ Validates the patterns ~~or suspicions~~ suggested by EDA with statistical reasons i.e testing assumptions with formal procedures.

→ Together they form a powerful cycle of discovery and confirmation in data Mining and statistical analysis.

| | Hypothesis Testing | EDA |
|---|---|---|
| Purpose → | To perform or reject a pre specified | → To discover patterns, understand distributions and generate new ideas. |
| Approach → | Deductive, Confirmatory | → Inductive, Discovery oriented |
| When used → | When clearer theory driven questions exist | → When data are unfamiliar large or complex |
| Focuses → | Formal Decision making | → Investigation of variables, distributions and relationships |

| | | |
|---|---|---|
| Tools → | Statistical tests (ANOVA, Chi Square) | Graphical (Histogram, Scatter Plot, Box Plot) |
| Outcome → | Binary Decision | → Insights, hypothesis |
| Flexibility → | Rigid, Structured | → Flexible, iterative |

## EDA on the Churn Dataset (A case study)

→ The churn dataset (UCI M/L repository) is used to demostrate EDA methods applied in a real world business scenario.

→ EDA helps in:
- ↳ Detecting anomalies or missing data
- ↳ Identifying patterns and relationships among variab
- ↳ Suggesting potential predictions for target variable
- ↳ Gaining domain insights through visualisation

## Overview of Dataset

Essential → No. of observations = 3333
No. of Inputs/Predictors/features = 20
Target (Churn) - indicates whether a custo has left the company.

30 Friday



→ The dataset contains a mix of categorical, integer value and continuous features describing customer demographics, account information, service usage, etc.

## Variables in Churn Dataset

a) Customer Identification

→ State : Categorical

→ Account Length : Integer

→ Area Code : Categorical

→ Phone Number : Unique Identifier

b) Service Plans

→ International Plans : Dichotomous Categorical (yes/no)

→ VoiceMail Plans : Dichotomous Categorical (yes/No)

→ No. of voice mail messages : Integer

c) Usage Matrics

→ Total minutes, calls, charges, ... (Continuous Integer)

Essential | Job to do | Phone No.

d) Customer Service Interaction

→ No. of calls to customer service (Integer)

182-183
27th Week

Saturday 01

e) Target Variable

-> Churn - Boolean (True / False)
indicates whether a customer left the company

02 Su

| Essential | Job to do | Phone No. |
| --- | --- | --- |
| | | |
| | | |
| | | |
| | | |

03 Monday

| Type | Variables |
|---|---|
| Categorical | State, Area Code |
| Identification | Phone Number |
| Flag Variables | International Plan, Voice Mail Plan |
| Numerical (Continuous / Integer) | Account length, No. of voice mail messages |
| Target | Churn (True / False) |

## Univariate vs Multivariate Analysis

→ Univariate analysis explores a single variable in isolation to understand its distribution, central tendency, spread and shape.

→ It doesnot relationship or dependencies

→ Purpose:

↳ Understand data range, outliers and overall pattern

↳ Identify missing or extreme values

↳ Decide on data transformations (Normalisation and log based transformation)

↳ Check assumptions for future modelling

Tuesday 04

| Type of Variable | Common Techniques | Visualisation |
|---|---|---|
| Categorical | Frequency counts, mode, ... | ~~Box Plot~~ Barchart, ... |
| Numerical | mean, median, ... | Box Plot, Scatter Plot |

→ Multivariate analysis investigates two or more variables simultaneously to detect patterns, relationships, correlations and interactions between them.

→ Purpose:

↳ Find dependencies and interactions between variables.

↳ Identify predictors (inputs) for target variable.

↳ Support future selection and hypothesis formulation

| Relationship Type | Typical Analysis | Visualisation |
|---|---|---|
| Two categorical | Contingency table, Chi Square Test | Clustered Bar Chart |
| 1 categorical + 1 numeric | Group means, box plots | Side by Side Box Plots |
| Two numeric | Correlation, Regression line | ~~Plot~~ Scatter Plots |
| Many numeric | Heatmap | Matrix Plots |

Essential

Job to do

Phone No.

05 Wednesday

# Contingency Table

→ It is also known as Cross Table

→ It is a type of table using statistics to show the frequency distribution of the variables i.e. how two or more categorical variables are related to each other.

Job to do | Phone No.