Data Exploration and Data Preprocessing

- **Data Mining**
- In general terms, “Mining” is the process of extraction of some valuable material from the earth e.g. coal mining, diamond mining, etc.
- **In the context of computer science, “Data Mining” can be referred to as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.**
- **It is basically the process carried out for the extraction of useful information from a bulk of data or data warehouses.**
- **We can think of Data Mining as a step in the process of Knowledge Discovery or Knowledge Extraction.**
- We can say it is a process of extracting interesting knowledge from large amounts of data.
- That is stored in many data sources. Such as file systems, databases, data warehouses. Also, knowledge used to contributes a lot of benefits to business and individual.

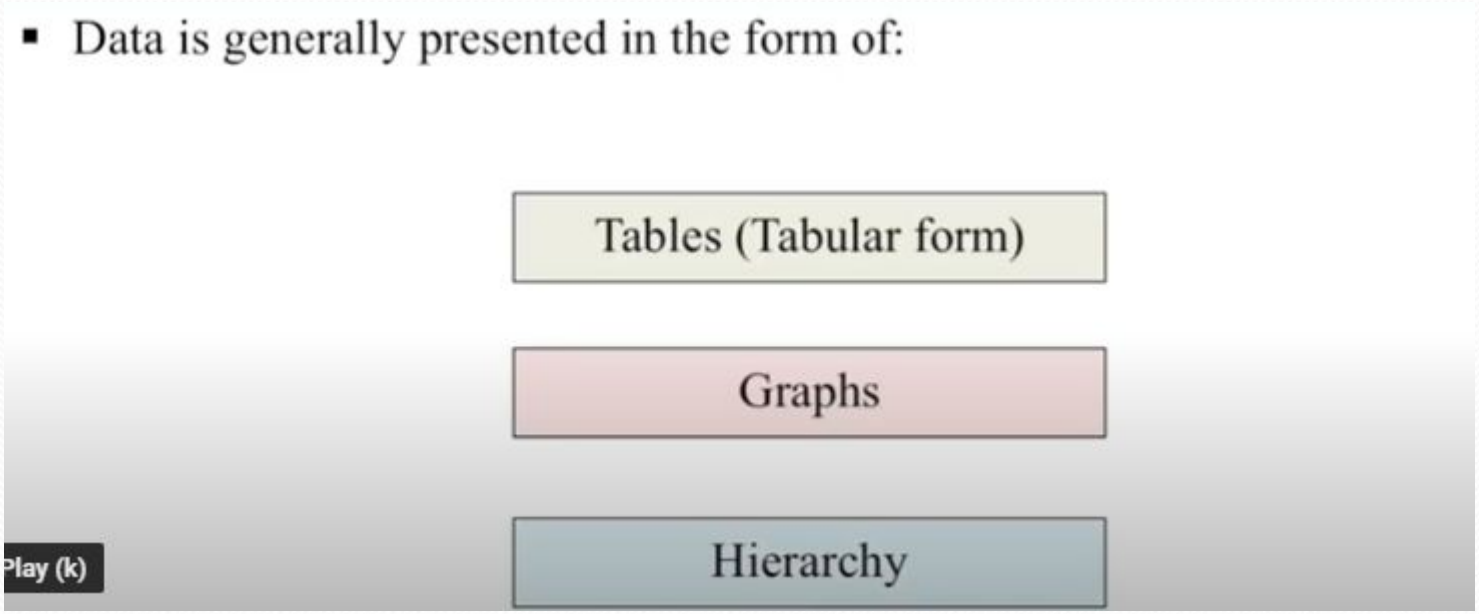
- Nowadays, data mining is used in almost all places where a large amount of data is stored and processed.
- For example, banks typically use ‘data mining’ to find out their prospective customers who could be interested in credit cards, personal loans, or insurance as well.
- Since banks have the transaction details and detailed profiles of their customers, they analyze all this data and try to find out patterns that help them predict that certain customers could be interested in personal loans, etc.

- ***Data->Information->knowlege***

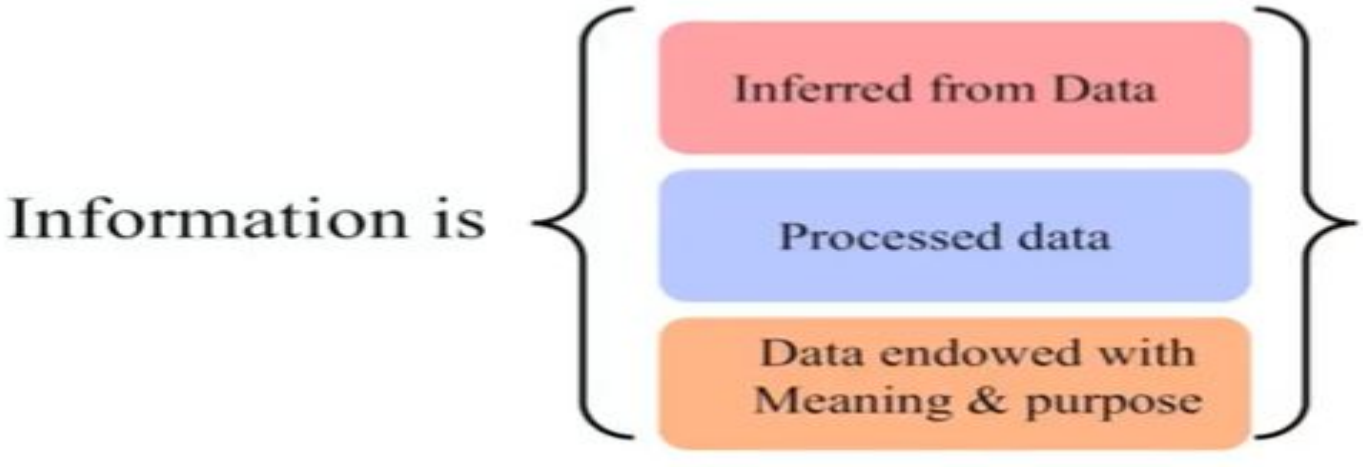
- Data is defined "as being discrete, objective facts or observations, which are unorganized and unprocessed and therefore have no meaning or value because of lack of context and interpretation"

Known facts about entity , Event , transaction etc.
Data is unorganized and unprocessed facts

- Data is generally presented in the form of:



Information is "organized or structured data, which has been processed in such a way that the information now has relevance for a specific purpose or context, and is therefore meaningful, valuable, useful and relevant"



Knowledge:

- Definitions of knowledge refer to information having been processed, organized or structured in some way, or else as being applied or put into action

- One of the most frequently quoted definitions of knowledge captures some of the various ways in which it has been defined by others : Knowledge is a fluid mix of

Framed experience

Values

Contextual information

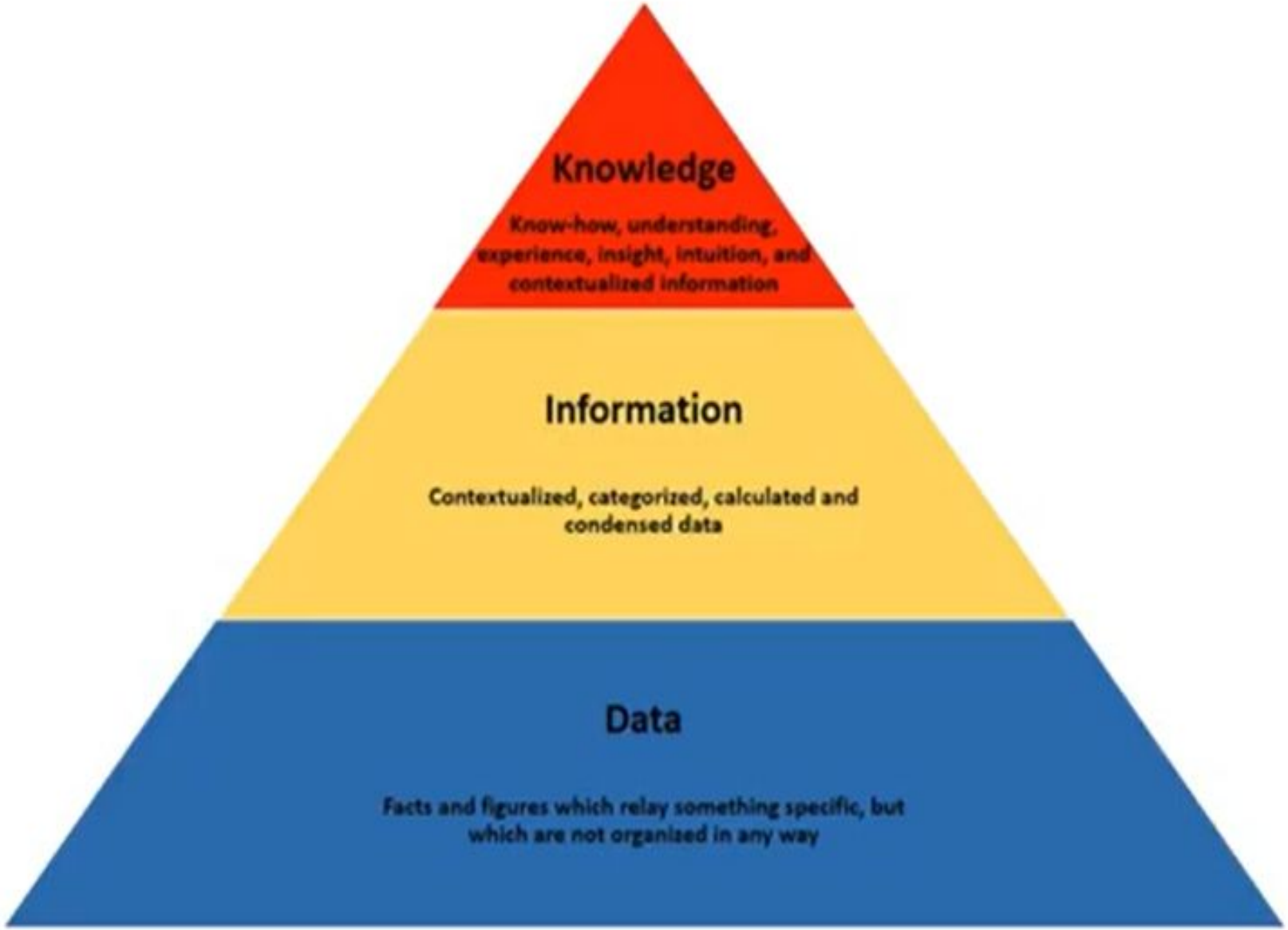
Expert insight

Grounded intuition

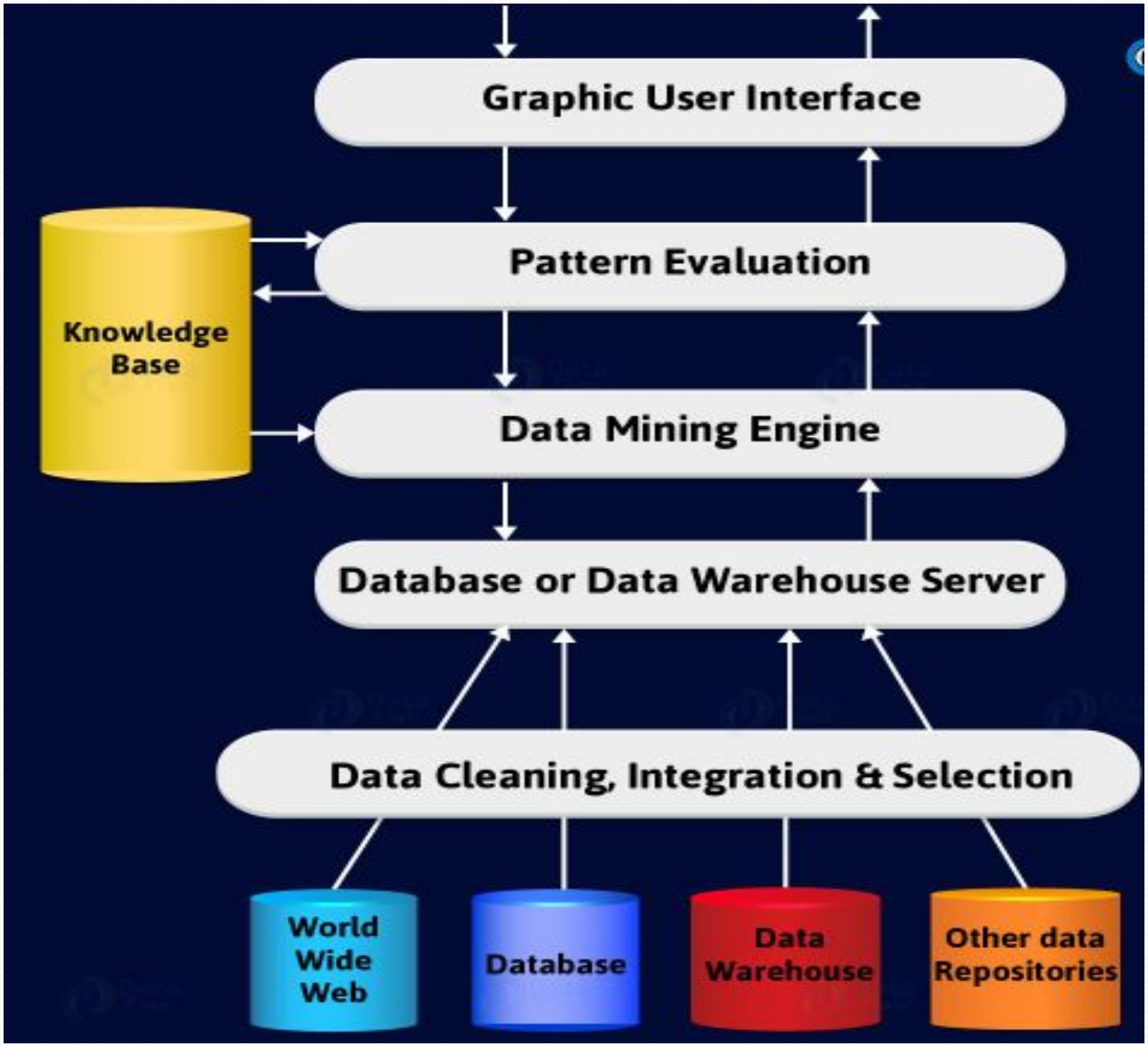
- That provides an environment and framework for evaluating and incorporating new experiences and information.

- ***Heart Disease Detection: Problem Area***
- ***Patients (weight,height,sugar,bp,BMI,HeartDisease)***
- ***80 5 120 100 5.5 yes*** ⟵ ***Information***
- ***70 5.1 100 80 5.0 No***

- ***weight,height,sugar,bp,BMI,HeartDisease***
- ***60 4.8 100 95 5.0 ?*** ⟵ ***Knowledge***


Knowledge
Know-how, understanding, experience, insight, intuition, and contextualized information
Information
Contextualized, categorized, calculated and condensed data
Data
Facts and figures which relay something specific, but which are not organized in any way

Data Mining Architecture



- ***Data mining Architecture system contains many components.***
- ***That is a data source, data warehouse server, data mining engine, and knowledge base.***

a. Data Sources

- *There are so many documents present.*
- *That is a database, data warehouse, World Wide Web (WWW).*
- *That are the actual sources of data. Sometimes, data may reside even in plain text files or spreadsheets.*
- *World Wide Web or the Internet is another big source of data.*

b. Database or Data Warehouse Server

- *The database server contains the actual data that is ready to be processed.*
- *Hence, the server handles retrieving the relevant data. That is based on the data mining request of the user.*

c. Data Mining Engine

- *In data mining system data mining engine is the core component.*
- *As It consists a number of modules.*
- *That we used to perform data mining tasks. That includes association, classification, characterization, clustering, prediction, etc.*

d. Pattern Evaluation Modules

- *This module is mainly responsible for the measure of interestingness of the pattern. For this, we use a threshold value.*
- *Also, it interacts with the data mining engine. That's main focus is to search towards interesting patterns.*

e. Graphical User Interface

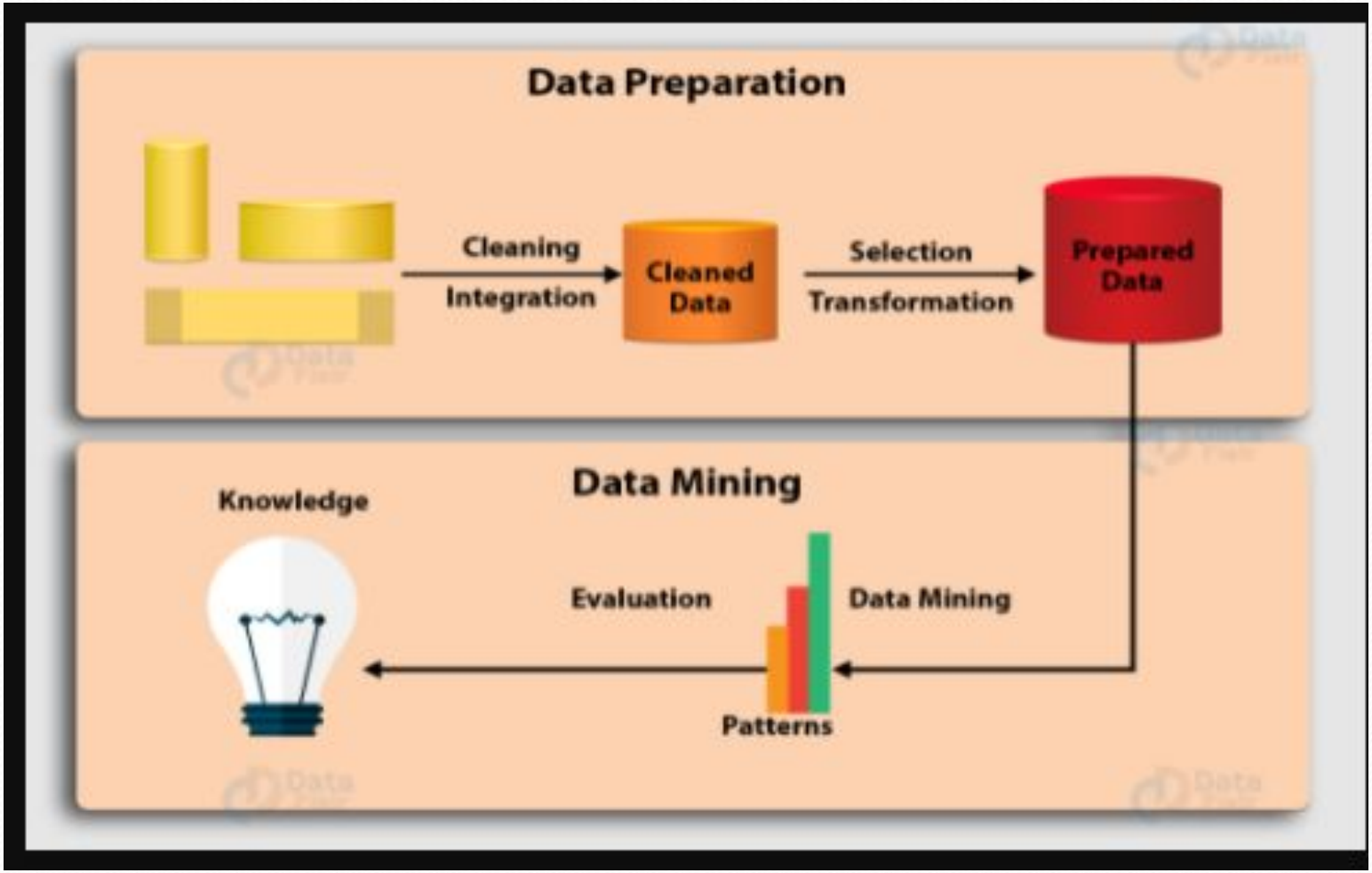
- *We use this interface to communicate between the user and the data mining system.*
- *Also, this module helps the user use the system easily and efficiently. They don't know the real complexity of the process.*
- *When the user specifies a query, this module interacts with the data mining system.*
- *Thus, displays the result in an easily understandable manner.*

- ***What is Data Mining?***
- ***It is a process of discovering hidden valuable knowledge by analyzing a large amount of data. Also, we have to store that data in different databases.***
- ***As data mining is a very important process.***
- ***It becomes an advantage for various industries. Such as manufacturing, marketing, etc. to increase their business efficiency. Therefore, the needs for a standard data mining process increased dramatically.***

- ***Stages of Data Mining Process***
- ***Data Mining Process is classified into two stages: Data preparation or data preprocessing and data mining***

Stages of Data Mining Process

Data Mining Process is classified into two stages: Data preparation or data preprocessing and data mining



- ***a. Data Cleaning***
- In the phase of data mining process, data gets cleaned.
- As we know data in the real world is noisy, inconsistent and incomplete.
- It includes a number of techniques. Such as filling in the missing values, combined compute.
- The output of the data cleaning process is adequately cleaned data.
- **b. Data Integration**
- In this phase of Data Mining process data in integrated from different data sources into one. As data lies in different formats in a different location.
- We can store data in a database, text files, spreadsheets, documents, data cubes, and so on.
- Although, we can say data integration is so complex, tricky and difficult task. That is because normally data doesn't match the different sources.
- We use metadata to reduce errors in the data integration process. Another issue faced is data redundancy.
- In this case, the same data might be available in different tables in the same database.
- Data integration tries to reduce redundancy to the maximum possible level. As without affecting the reliability of data

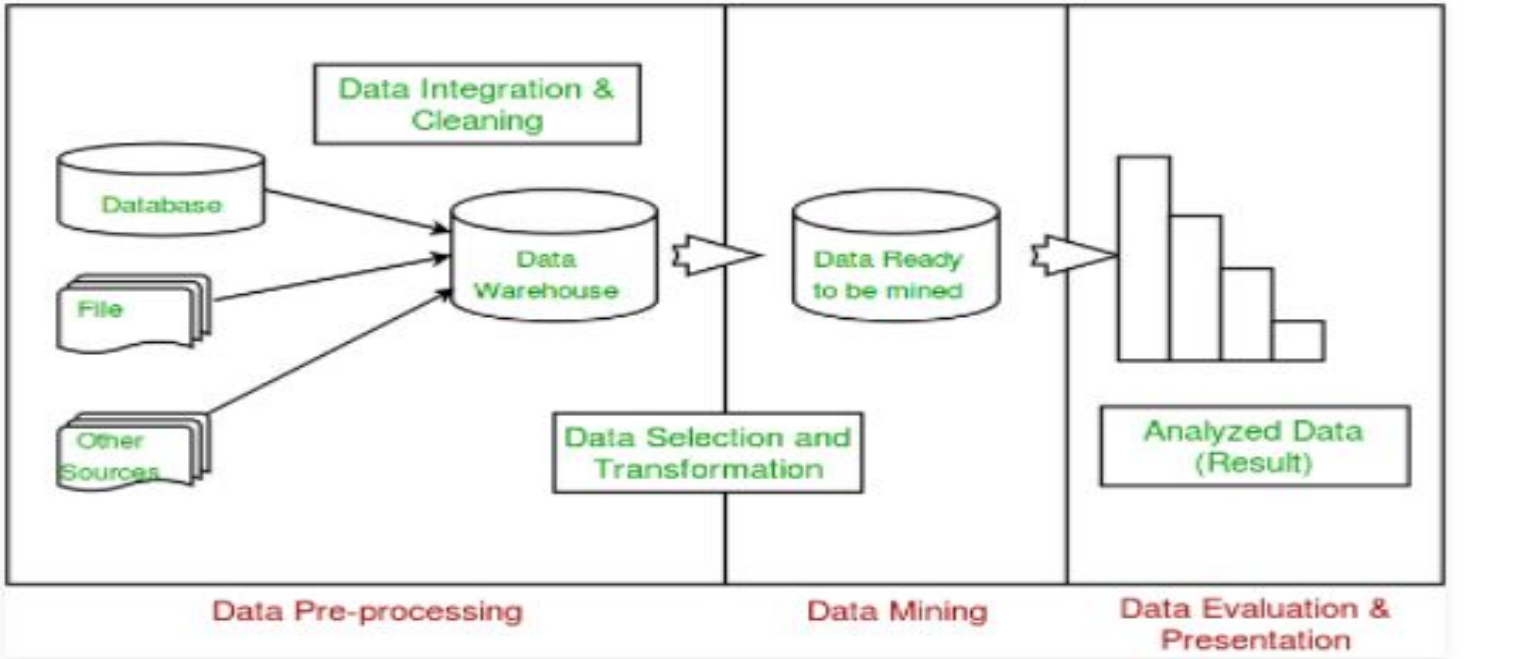
- ***c. Data Selection***
- This is the process by which data relevant to the analysis is retrieved from the database.
- As this process requires large volumes of historical data for analysis.
- So, usually, the data repository with integrated data contains much more data than actually required.
- From the available data, data of interest needs to be selected and stored.
- ***d. Data Transformation***
- In this process, we have to transform and consolidate the data into different forms. That must be suitable for mining.
- Normally this process includes normalization, aggregation, generalization etc.
- For example, a data set available as "-5, 37, 100, 89, 78" can be transformed as "-0.05, 0.37, 1.00, 0.89, 0.78". Here data becomes more suitable for data mining. After data integration, the available data is ready for data mining.
- ***e. Data Mining***
- In this phase of Data Mining process, we have applied methods to extract patterns from the data.
- As these methods are complex and intelligent. Also, this mining includes several tasks. Such as classification, prediction, clustering, time series analysis and so on.

f. Pattern Evaluation

- The pattern evaluation identifies the truly interesting patterns.
- That is representing knowledge based on different types of interesting measures.
- A pattern is considered to be interesting if it is potentially useful.
- Also, easily understandable by humans.

- ***g. Knowledge Representation***
- In the phase of Data Mining process, we have to represent data to the user in an appealing way.
- Also, that information is mined from the data. To generate output different techniques are need to be applied.

- ***Data Mining phases:***
- ***The whole process of Data Mining consists of three main phases:***
- ***Data Pre-processing – Data cleaning, integration, selection, and transformation takes place***
- ***Data Extraction – Occurrence of exact data mining***
- ***Data Evaluation and Presentation – Analyzing and presenting results***



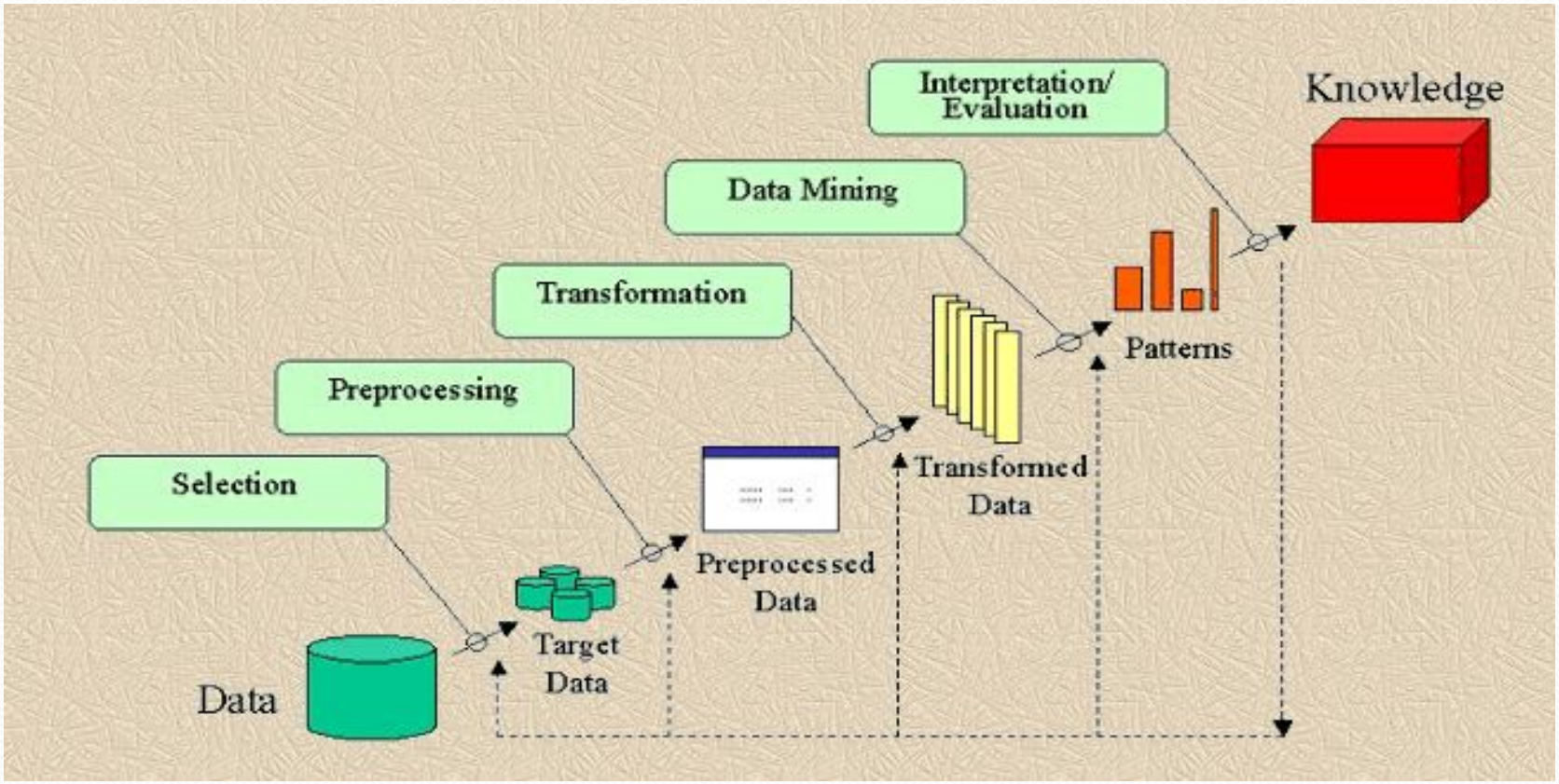
- ***What is the KDD Process?***

- ***The term Knowledge Discovery in Databases, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods.***
- ***It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.***

- ***KDD process is to extract knowledge from data in the context of large databases.***

- ***It does this by using data mining methods (algorithms) to extract (identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required preprocessing, subsampling, and transformations of that database.***

KDD Process:


Interpretation/
Evaluation
Knowledge
Data Mining
Transformation
Patterns
Preprocessing
Selection
Transformed
Data
Preprocessed
Data
Target
Data
Data

- *The overall process of finding and interpreting patterns from data involves the repeated application of the following steps:*
- ***Developing an understanding of***
- *the application domain*
- *the relevant prior knowledge*
- *the goals of the end-user*
- ***Creating a target data set:***
- *selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.*
- ***Data cleaning and preprocessing.***
- *Removal of noise or outliers.*
- *Collecting necessary information to model or account for noise.*
- *Strategies for handling missing data fields.*
- *Accounting for time sequence information and known changes.*
- ***Data reduction and projection.***
- *Finding useful features to represent the data depending on the goal of the task.*

- Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.
- ***Choosing the data mining task.***
- *Deciding whether the goal of the KDD process is classification, regression, clustering, etc.*
- ***Choosing the data mining algorithm(s).***
- *Selecting method(s) to be used for searching for patterns in the data.*
- *Deciding which models and parameters may be appropriate.*
- *Matching a particular data mining method with the overall criteria of the KDD process.*
- ***Data mining.***
- *Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.*
- ***Interpreting mined patterns.***
- ***Consolidating discovered knowledge.***

- **KDD** refers to the overall process of discovering useful knowledge from data.
- It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge.
- It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.
  Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the **KDD process.**
- ***Definitions Related to the KDD Process***
- ***Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.***

Data	A set of facts, F.
Pattern	An expression E in a language L describing facts in a subset F_E of F.
Process	KDD is a *multi-step process* involving data preparation, pattern searching, knowledge evaluation, and refinement with iteration after modification.
Valid	Discovered patterns should be true on new data with some degree of certainty. Generalize to the future (other data).
Novel	Patterns must be novel (should not be previously known).
Useful	Actionable; patterns should potentially lead to some useful actions.
Understandable	The process should lead to human insight. Patterns must be made understandable in order to facilitate a better understanding of the underlying data.

- ***Applications of Data Mining***

- ***Financial Analysis***
- ***Biological Analysis***
- ***Scientific Analysis***
- ***Intrusion Detection***
- ***Fraud Detection***
- ***Research Analysis***

- ***Data Mining Applications & Use Cases***
- ***Following are the applications of data mining in various sectors:***
- ***a. Data Mining in Finance***
- *We have to Increase customer loyalty by collecting and analyzing customer behavior data. Also, one needs to help banks that predict customer behavior and launch relevant services and products.*
- *Helps in Discovering hidden correlations between various financial indicators that need to detect suspicious activities with a high potential risk.*
- *Generally, it identifies fraudulent or non-fraudulent actions. As it done by collecting historical data. And then turning it into valid and useful information.*
- ***b. Data Mining in Healthcare***
- *Basically, it provides government, regulatory and competitor information that can fuel competitive advantage. Although, it supports the R&D process. And then go-to-market strategy with rapid access to information at every phase.*

- *Generally, it discovers the relationships between diseases and the effectiveness of treatments.*
- *That is to identify new drugs or to ensure that patients receive appropriate, timely care.*
- *Also, it supports healthcare insurers in detecting fraud and abuse.*

c. Data Mining for Intelligence

- *Generally, it reveals hidden data related to money laundering, narcotics trafficking, etc.*
- *Also, helps in Improving intrusion detection with a high focus on anomaly detection. And identify suspicious activity from a day one.*
- *Basically, convert text-based crime reports into word processing files.*
- *That can be used to support the crime-matching process.*

- ***d. Data Mining in Telecommunication***
- *In this, data mining gains a competitive advantage and reduce customer churn by understanding demographic characteristics and predicting customer behavior.*
- *Increases customer loyalty and improve profitability by providing customized services.*
- *As it supports customer strategy by developing appropriate marketing campaigns and pricing strategies.*

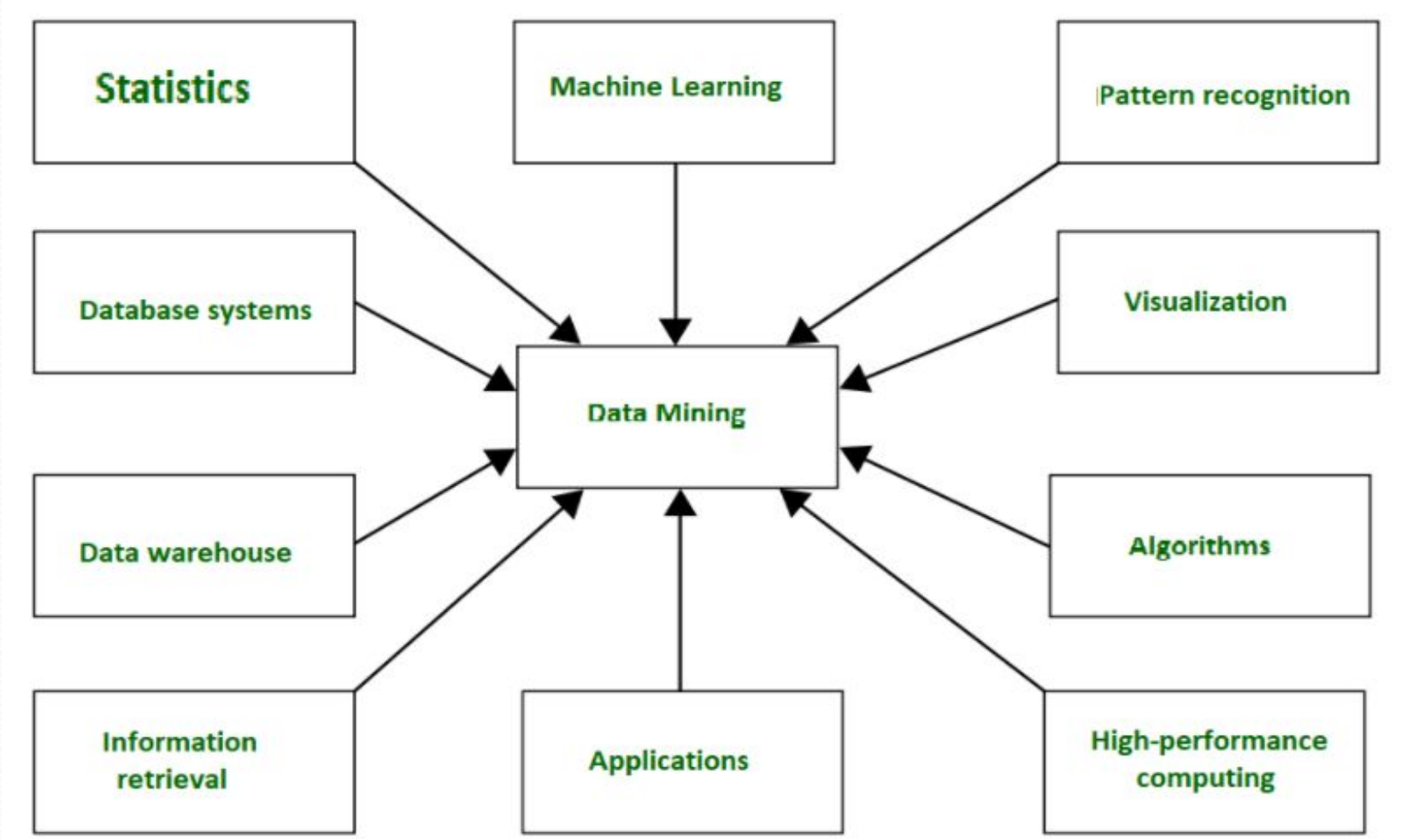
- ***f. Data Mining in Marketing and Sales***
- *Basically, it enables businesses to understand the hidden patterns inside historical purchasing transaction data. Thus helping in planning and launching new marketing campaigns.*
- *Generally, the following illustrates several data mining applications in sale and marketing.*
- *We use it for market basket analysis. That is to provide information on what product combinations have to purchased together. This information helps businesses promote their most profitable products and maximize the profit. In addition, it encourages customers to purchase related products.*
- *Retail companies use data mining to identify customer's behavior buying patterns.*
- ***g. Data Mining in E-commerce***
- *Many E-commerce companies are using data mining business Intelligence to offer cross-sells through their websites.*
- *One of the most famous of these is, of course, Amazon. They use sophisticated mining techniques to drive their 'People who viewed that product. Also liked this' functionality.*

- ***h. Data Mining in Education***
- There is a newly emerging field, called Educational Data Mining. As it concerns with developing methods. That discover knowledge from data originating from educational Environments.

- **The goals of EDM are identified as predicting students' future learning behavior, studying. We use data mining by an institution to take accurate decisions. And also to predict the results of the student.**

- With the results, the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured. And used to develop techniques to teach them.
- https://data-flair.training/blogs/data-mining-applications/

Main Purpose of Data Mining



- ***Basically, Data mining has been integrated with many other techniques from other domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, etc. to gather more information about the data and to helps predict hidden patterns, future trends, and behaviors and allows businesses to make decisions.***
- Technically, data mining is the computational process of analyzing data from different perspectives, dimensions, angles and categorizing/summarizing it into meaningful information.

Getting to Know Your Data

- ***Data Objects and Attribute Types***
- ***Basic Statistical Descriptions of Data***
- ***Measuring Data Similarity and Dissimilar***

- ***What is Data?***
- *Data sets are made up of data objects.*
- *A data object represents an entity.*
- *Also called sample, example, instance, data point, object, tuple.*
- *Data objects are described by attributes.*
- *An attribute is a property or characteristic of a data object. – Examples: eye color of a person, temperature, etc.*
- *Attribute is also known as variable, field, characteristic, or feature*
- *A collection of attributes describe an object.*
- *Attribute values are numbers or symbols assigned to an attribute.*

Data Objects

- *Data sets are made up of data objects.*
- *A **data object** represents an entity.*
- *Examples:*
  - ***sales database: customers, store items, sales***
  - ***medical database: patients, treatments***
  - ***university database: students, professors, courses***
- *Also called samples , examples, instances, data points, objects, tuples.*
- *Data objects are described by **attributes**.*
- *Database rows -> data objects; columns ->attributes.*

Types of Data Sets

- ***Record***
  - *Relational records*
  - *Data matrix, e.g., numerical matrix, crosstabs*
  - *Document data: text documents: term-frequency vector*
  - *Transaction data*
- ***Graph and network***
  - *World Wide Web*
  - *Social or information networks*
  - *Molecular Structures*
- ***Ordered***
  - *Video data: sequence of images*
  - *Temporal data: time-series*
  - *Sequential Data: transaction sequences*
  - *Genetic sequence data*
- ***Spatial, image and multimedia:***
  - *Spatial data: maps*
  - *Image data:*
  - *Video data:*

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

A Data Object

Attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Objects

database rows data objects
database columns attributes

Attributes:

- Attribute (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.
- **E.g., customer _ID, name, address**
- Attribute values are numbers or symbols assigned to an attribute.
- The attribute can be defined as a field for storing the data that represents the characteristics of a data object.
- The attribute is the property of the object.
- The attribute represents different features of the object.
- For example, hair color is the attribute of a lady.
- Similarly, rollno, and marks are attributes of a student.
- An attribute vector is commonly known as a set of attributes that are used to describe a given object.
  Type of attributes:
  We need to differentiate between different types of attributes during Data-preprocessing.
- There are different types of attributes. some of these attributes are mentioned below.

- So firstly, we need to differentiate between qualitative and quantitative attributes.
  1. **Qualitative Attributes** such as Nominal, Ordinal, and Binary Attributes.
  2. **Quantitative Attributes** such as Discrete and Continuous Attributes.
- Attributes can store all kinds of different descriptive information, however, depending on the descriptive information stored in each attribute, some properties are implied, so operations are permitted, and other operations and properties are not allowed and not applied respectively.
- Descriptive information does not imply any order, size, or any other quantitative information.

- ***Attributes can store all kinds of different descriptive information.***

.

- We can break attributes down into four different categories: nominal, ordinal, interval, and ratio.
- The first attribute category is the nominal attribute category.
- What's important here is that this That means that you cannot state that one attribute is greater than or less than another attribute or you cannot multiply attributes together, so for instance, it does not make sense to multiply the color blue by the color red.
- The only comparisons you can do with nominal attributes, are to check whether to attributes are equal or not equal.

Attributes

Attribute:

It can be seen as a data field that represents the characteristics or features of a data object.

For a customer, object attributes can be customer Id, address, etc. We can say that a set of attributes used to describe a given object are known as attribute vector or feature vector.

Attribute (*or* ***dimensions, features, variables*** *): a data field, representing a characteristic or feature of a data object.*

E.g., customer _ID, name, address

Types:

Nominal

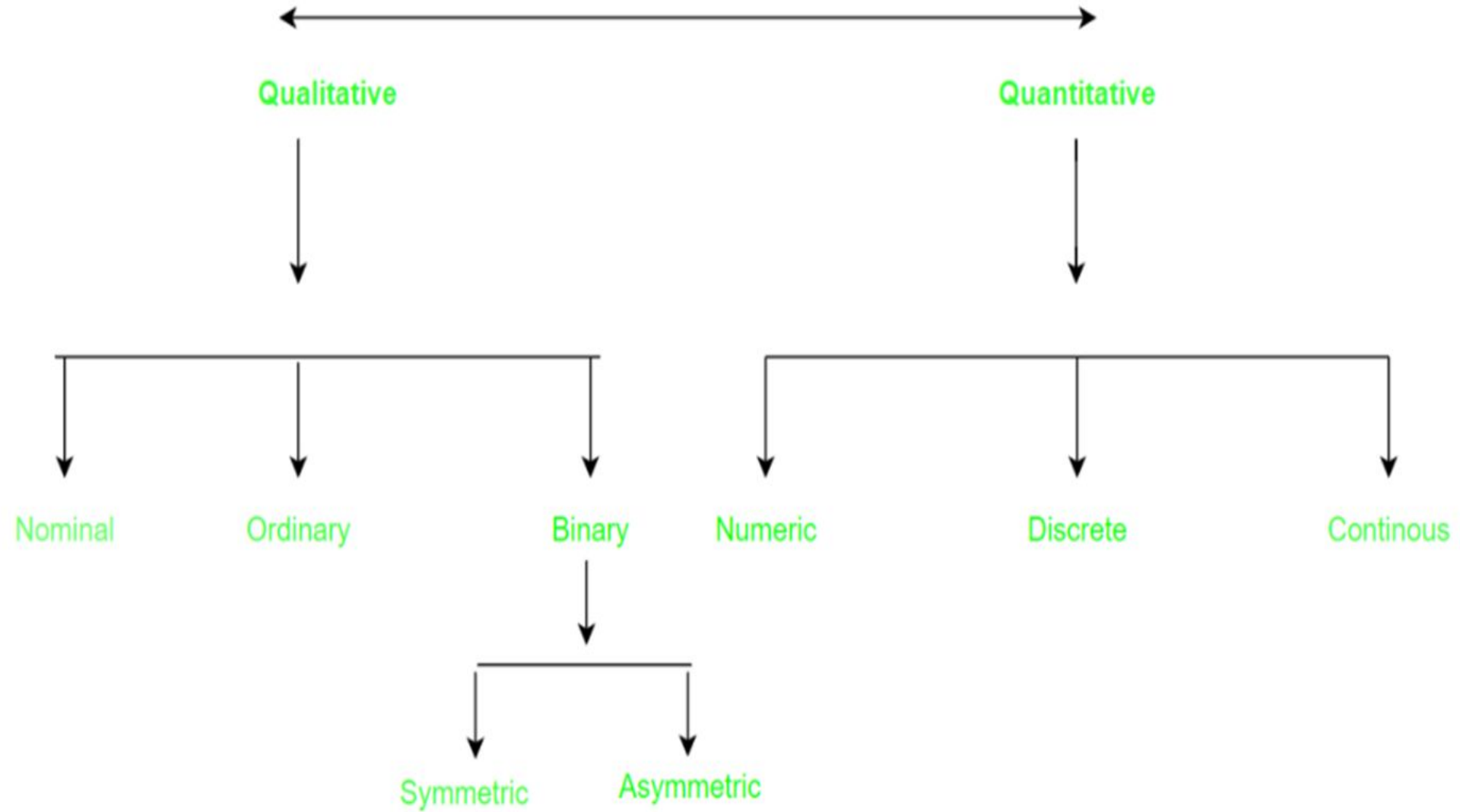
Ordinal

Interval-scaled

Ratio-scaled

- ***1. Nominal scale***
- Nominal scale deals with the non-numeric data that is with the categorical data
- It is a system of assigning number to the variable to label them only for identification and to distinguish them from each other. Example: Car-1, Buses-2
- It is a measure that simply divides objects or events into categories
- Here, categories are designated with names or numerals but ordering of categories is meaningless i.e. there is no order
- Examples: Gender, race, color preference, etc.
- Researchers may add the number or represent by number male=1 and female =2 but that number has no numerical value.
- These are nominal because they are numerical in name only
- The only mathematical operation that can be performed is count or say frequency.

Types of Attributes



Qualitative Attributes:

1.Nominal Attributes –

- A nominal attribute provides descriptive information about the object such as the color of the object, the name of an object so for instance a city name, or the type of an object.

related to names: The values of a Nominal attribute are names of things, some kind of symbols.

Nominal scales: Measurements where a value is used to represent something or someone.

Nominal values are typically coded, or converted to numeric values for later statistical analysis.

Values of Nominal attributes represents some category or state and that's why nominal attribute also referred as categorical attributes and there is no order (rank, position) among values of the nominal attribute.

Attribute	Values
Colours	Black, Brown, White
Categorical Data	Lecturer, Professor, Assistant Professor

- ***Nominal Attributes:***
- Nominal means “relating to names”.
- The values of a nominal attribute are symbols or names of things.
- **Each value represents some kind of category, code, or state,**
- Nominal attributes are also referred to as categorical attributes.
- The values of nominal attributes do not have any meaningful order.
- **Example: The attribute marital_status can take on the values single, married hair color,Occupation etc.**
- Because nominal attribute values do not have any meaningful order about them and they are not quantitative.
- **It makes no sense to find the mean (average) value or median (middle) value for such an attribute.**
- **Mode is one of the measures of Central tendency.**

- **Properties of nominal scale are:**
- Mutually exclusive
- Categories are distinct and homogeneous
- They cannot be measured or ordered but can be counted
- Data can reflect that they are different from each other but cannot be ordered as smaller or greater.
- The only mathematical operation that can be performed is count or say frequency.

- Nominal Scale, also called the categorical variable scale, is defined as a scale used for labeling variables into distinct classifications and doesn't involve a quantitative value or order.
- This scale is the simplest of the four variable measurement scales.
- Nominal scale is often used in research surveys and questionnaires where only variable labels hold significance.
- For instance, a customer survey asking "Which brand of smartphones do you prefer?" Options : "Apple"- 1 , "Samsung"-2, "OnePlus"-3.
- Nominal Scale Data and Analysis
- There are two primary ways in which nominal scale data can be collected:
- By asking an open-ended question, the answers of which can be coded to a respective number of label decided by the researcher.
- The other alternative to collect nominal data is to include a multiple choice question in which the answers will be labeled.

Nominal Scale Examples

- Gender
- Political preferences
- Place of residence

What is your Gender?	What is your Political preference?	Where do you live?
• M- Male • F- Female	• 1- Independent • 2- Democrat • 3- Republican	• 1- Suburbs • 2- City • 3- Town

- **Nominal Attributes: Binary attribute**
- A binary attribute is a special nominal attribute with only two states: 0 or 1.
- A binary attribute is symmetric if both of its states are equally valuable and carry the same weight.
- **Example: the attribute gender having the states male and female.**
- A binary attribute is asymmetric if the outcomes of the states are not equally important.
- **Example: Positive and negative outcomes of a medical test.**
- By convention, we code the most important outcome, which is usually the rarest one, by 1 (e.g. positive) and the other by 0 (e.g. negative).

2. Binary Attributes:

Binary data has only 2 values/states.

For Example yes or no, affected or unaffected, true or false.

Symmetric: Both values are equally important (Gender).

Asymmetric: Both values are not equally important (Result).

Attribute	Values
Cancer detected	Yes, No
result	Pass , Fail

3. Ordinal Attributes:

•An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known.

•**Example:** An ordinal attributes include grade (eg. A+,A,B+,B etc).

•This attribute has three possible values: small, medium, and large.

•The values have a meaningful sequence (which corresponds to increasing)

•Professional Rank etc.

•The central tendency of an ordinal attribute can be represented by its mode and its median (middle value in an ordered sequence), but the mean cannot be defined

3. Ordinal Attributes :

The Ordinal Attributes contains values that have a meaningful sequence or ranking(order) between them, but the magnitude between values is not actually known, the order of values that shows what is important but don't indicate how important it is.

Attribute	Value
Grade	A,B,C,D,E,F
Basic pay scale	16,17,18

Ordinal Scale Properties:

- It has unequal units
- It displays from highest to lowest by different measurement points
- It has no zero point i.e. it is arbitrary or absolute
- Interval size is unequal and unknown
- **M**utually exclusive
- Categories are distinct and homogeneous
- They cannot be measured but can be counted and ordered/ranked
- Data can express that one is different from another and one is greater or smaller than the other.
- However, data cannot say that one is ‘X’ units or ‘X’times greater or smaller than the other.

Examples: Academic performance: School-1, College-2, Bachelor-3, Masters-4

Disease severity: mild-1, moderate-2, severe-3

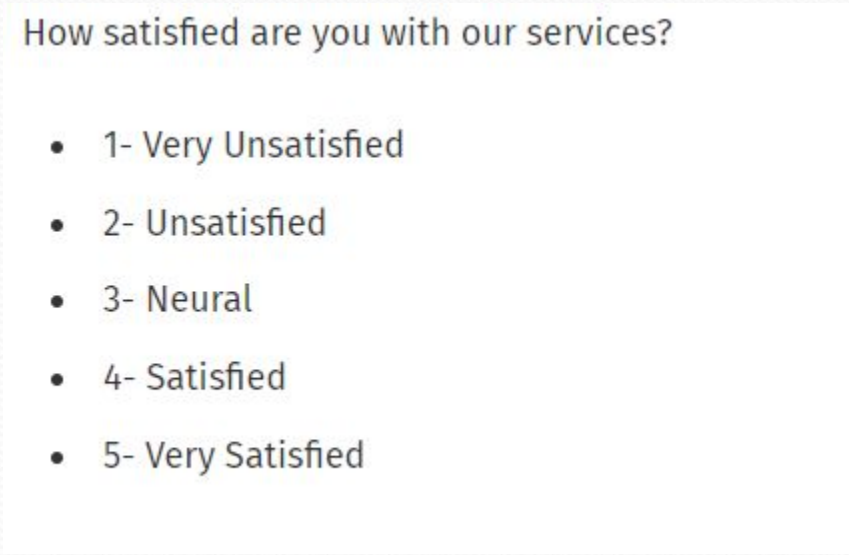
3. Ordinal Attributes :

Characteristics of Ordinal Variable

- It is an extension of nominal data.
- It has no standardized interval scale.
- It establishes a relative rank.
- It measures qualitative traits.
- The median and mode can be analyzed.
- It has a rank or order.

Interval size is unequal and unknown
Mutually exclusive
Categories are distinct and homogeneous
They cannot be measured but can be counted and ordered/ranked

- Ordinal Scale is defined as a variable measurement scale used to simply depict the order of variables and not the difference between each of the variables.
- These scales are generally used to depict non-mathematical ideas such as frequency, satisfaction, happiness, a degree of pain, etc.
- Origin of this scale is absent due to which there is no fixed start or “true zero”.
- example



Quantitative Attributes:

1. Numeric: A numeric attribute is quantitative because, it is a measurable quantity, represented in integer or real values. Numerical attributes are of 2 types, **interval**, and **ratio**.

- An **interval-scaled** attribute has values, whose differences are interpretable, but the numerical attributes do not have the correct reference point, or we can call zero points.
- Data can be added and subtracted at an interval scale but can not be multiplied or divided.
- **Consider an example of temperature in degrees Centigrade.**
- **If a day's temperature of one day is twice of the other day we cannot say that one day is twice as hot as another day.**

Numeric Attributes:

- ***A Numeric Attribute is quantitative , it is measurable quantity , represented in integer or real values. Numeric attributes can be interval-scaled or ratio-scaled.***
- ***Interval-Scaled Attributes:***
- Interval-Scaled Attributes are measured on a scale of equal-size units.
- The values of interval-scaled attributes have order and can be positive,0,negative.
- Thus, in addition to providing a ranking of values , such attributes allow and quantify the difference between values.

Eg. temprature attribute is interval scaled.

Quantitative Attributes:

- A **ratio-scaled** attribute is a numeric attribute with a fix zero-point.
- If a measurement is ratio-scaled, we can say of a value as being a multiple (or ratio) of another value.
- The values are ordered, and we can also compute the difference between values, and the mean, median, mode, Quantile-range, and Five number summary can be given.

- **Interval scale** is a system of assigning number to the variable to label them for identification and ranking based on a scale having equal interval size with arbitrary zero.
- Here, zero is arbitrary that allows measurement on either side of zero.
- Here, a variable is categorized in different subgroups in ascending or descending order and intervals between the successive categories are equal and constant.
- We can know bigger value and also how much bigger they are.
- Zero is not the lowest value there are point on scale which are below than it
- Examples include Celsius, Fahrenheit Temperature, IQ (intelligence scale), Here, 0 degree Celsius does not mean no temperature or no heat
- Likewise, 0 IQ does not mean no IQ
- Likewise, 64 degrees Fahrenheit is 32 units more than 32 degrees Fahrenheit but not twice as warm as 32 degrees Fahrenheit
- IQ 100 is 50 times more than IQ 50 but individual with IQ 100 is not two times intelligent than individual with IQ 50.

Characteristics of interval scale:

- It has equal units
- It has arbitrary (absolute) zero which is just a reference point
- Interval size is known, equal and constant
- Measurement is taken on both sides of zero
- Mutually exclusive.
- Categories are distinct and homogeneous.
- Can be measured and ranked/ordered.
- Data can show that one is different from other, one is greater or smaller than other, one is 'X' unit greater or smaller than other but cannot show that one is 'X' times greater or smaller than the other.
- Data can be added or subtracted but cannot be divided or multiplied.

. Ratio scale

•Ratio scale is a system of assigning number to the variable to label them for identification and ranking based on a scale having equal interval size with absolute zero that allows measurement on only one side of zero.

•Here, a variable is categorized in different subgroups in ascending or descending order and intervals between the successive categories are equal and constant.

•Same as the interval scale except zero has the true value i.e. zero represent the absolute value

•All the mathematical operations are applicable in this scale.

•Examples include: weight, height, sales figures, ruler measurements, number of children.

•Here, weight 0 means no weight and weight 60 kgs mean it is 30 units more than 30 kgs and two times of 30 kgs.

Characteristics of ratio scale:

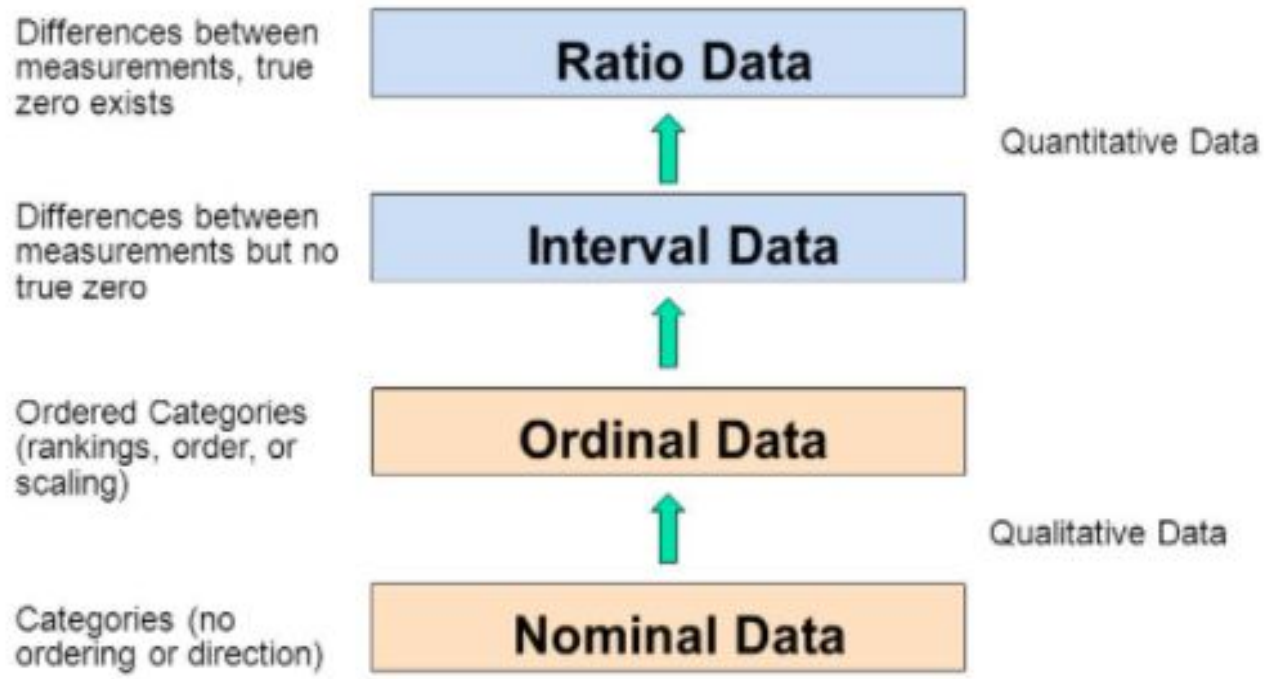
- It has equal units
- Has absolute zero
- Interval size is known, equal and constant.
- Measures on only one side of zero
- Mutually exclusive.
- Categories are distinct and homogeneous.
- Can be measured and ranked/ordered.
- Data can show that one is different from other, one is greater or smaller than other, one is 'X' unit greater or smaller than other, one is 'X' times greater or smaller than the other.
- Data can be added, subtracted, multiplied or divided.
- Ratio of two numbers can be meaningfully calculated and interpreted

- **Comparisons of Four Scales of Measurement:**

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Scales of Measurement	Arithmetic aspects that can be performed	Features/characteristics	Examples
Nominal scale	· Counting	· Unordered · Can say one is different from other	Religion, sex etc.
Ordinal scale	· Counting · Ranking	· Ordered category with unequal and unknown interval size	Economic status academic performance etc.
Interval scale	· Counting · Ranking · Measurement · Addition · Subtraction	· Ordered category with equal and known interval size and arbitrary zero · Can say one is different from the other · Can say one is greater or smaller than the other · Can say one is 'X' units greater or smaller than the other	Temperature, IQ score etc.

Ratio Scale	· Counting · Ranking · Measurement · Addition · Subtraction · Multiplication · Division	· Ordered category with equal and known interval size and absolute zero · Can say one is different from the other · Can say one is greater or smaller than the other · Can say one is ‘X’ units greater or smaller than the other · Can say one is ‘X’ times greater or smaller than the other	Body weight, height etc.


Differences between measurements, true zero exists
Ratio Data
Quantitative Data
Differences between measurements but no true zero
Interval Data
Ordered Categories (rankings, order, or scaling)
Ordinal Data
Qualitative Data
Categories (no ordering or direction)
Nominal Data

- **Nominal**
- A nominal scale describes a variable with categories that do not have a natural order or ranking. You can code nominal variables with numbers if you want, but the order is arbitrary and any calculations, such as computing a mean, median, or standard deviation, would be meaningless.

- Examples of nominal variables include:

- genotype, blood type, zip code, gender, race, eye color, political party

- **Ordinal**
- An ordinal scale is one where the order matters but not the difference between values.
- Examples of ordinal variables include:
- socio economic status (“low income”,”middle income”,”high income”), education level (“high school”,”BS”,”MS”,”PhD”), income level (“less than 50K”, “50K-100K”, “over 100K”), satisfaction rating (“extremely dislike”, “dislike”, “neutral”, “like”, “extremely like”).

- Note the differences between adjacent categories do not necessarily have the same meaning.
- For example, the difference between the two income levels “less than 50K” and “50K-100K” does not have the same meaning as the difference between the two income levels “50K-100K” and “over 100K”.

- ***Interval***
- An interval scale is one where there is order and the difference between two values is meaningful.
- Examples of interval variables include:
- temperature (Farenheit), temperature (Celcius), pH, SAT score (200-800), credit score (300-850).

- ***Ratio***
- A ratio variable, has all the properties of an interval variable, and also has a clear definition of 0.0. When the variable equals 0.0, there is none of that variable.
- Examples of ratio variables include:
- enzyme activity, dose amount, reaction rate, flow rate, concentration, pulse, weight, length, temperature in Kelvin (0.0 Kelvin really does mean “no heat”), survival time.
- When working with ratio variables the ratio of two measurements has a meaningful interpretation.
- For example, because weight is a ratio variable, a weight of 4 grams is twice as heavy as a weight of 2 grams.
- However, a temperature of 10 degrees C should not be considered twice as hot as 5 degrees C.
- If it were, a conflict would be created because 10 degrees C is 50 degrees F and 5 degrees C is 41 degrees F. Clearly, 50 degrees is not twice 41 degrees.
- Another example, a pH of 3 is not twice as acidic as a pH of 6, because pH is not a ratio variable.

Quantitative Attributes:

2. Discrete : Discrete data have finite values it can be numerical and can also be in categorical form. These attributes has finite or countably infinite set of values.

Example:

Attribute	Value
Profession	Teacher, Business man, Peon
ZIP Code	301701, 110040

Quantitative Attributes:

3.. **Continuous**: Continuous data have an infinite no of states. Continuous data is of float type. There can be many values between 2 and 3.

Example :

Attribute	Value
Height	5.4, 6.2 ...etc
weight	50.33 .....etc

- ***Attribute Summary***

Attribute Types – Categorical/Qualitative

Nominal: *categories, states, or "names of things"*

- ***Hair_color = {auburn, black, blond, brown, grey, red, white}***
- ***marital status, occupation, ID numbers, zip codes***

Binary

- ***Nominal attribute with only 2 states (0 and 1)***
- ***Symmetric binary: both outcomes equally important***
  - e.g., gender
- ***Asymmetric binary: outcomes not equally important.***
  - e.g., medical test (positive vs. negative)
  - Convention: assign 1 to most important outcome (e.g., positive/Negative)

Ordinal

- ***Values have a meaningful order (ranking) but magnitude between successive values is not known.***
- ***Size = {small, medium, large}, grades, army rankings***

Numeric Attribute Types

Quantity (integer or real-valued)

Interval

Measured on a scale of **equal-sized units**

Values have order

E.g., *temperature in C˚or F˚, calendar dates*

No true zero-point

Ratio

Inherent **zero-point**

We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).

e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

Discrete Attribute

- ***Has only a finite or countably infinite set of values***
  - E.g., zip codes, profession, or the set of words in a collection of documents
- ***Sometimes, represented as integer variables***
- ***Note: Binary attributes are a special case of discrete attributes***

Continuous Attribute

- ***Has real numbers as attribute values***
  - E.g., temperature, height, or weight
- ***Practically, real values can only be measured and represented using a finite number of digits***
- ***Continuous attributes are typically represented as floating-point variables***

- **Example of attribute**
- In this example, RollNo, Name, and Result are attributes of the object named as a student.

Rollo	Name	Result
1	Ali	Pass
2	Akram	Fail

- ***Nominal Attributes***
- Nominal data is in alphabetical form and not in an integer.
- Nominal Attributes are Qualitative Attributes.
- Examples of Nominal attributes

Attribute	Value
Categorical data	Lecturer, Assistant Professor , Professor
States	New, Pending, Working, Complete, Finish
Colors	Black, Brown, White, Red

- ***Binary Attributes***
- Binary data have only two values/states.
- For example, here Disease detected can be only Yes or No.
- Binary Attributes are Qualitative Attributes.
- Examples of Binary Attributes

Attribute	**Value**
Disease detected	Yes, No
Result	Pass, Fail

- **The binary attribute is of two types;**
- **Symmetric binary**
- **Asymmetric binary**
- Examples of Symmetric data
- Both values are equally important. For example, if we have open admission to our university, then it does not matter, whether you are a male or a female.
- Example:

Attribute	Value
Gender	Male, Female

- ***Examples of Asymmetric data***
- Both values are not equally important.
- For example, Disease detected is more important than Disease not detected.
- If a patient is with Disease and we ignore him, then it can lead to death but if a person is not Disease detected and we ignore it, then there is no special issue or risk.
- **Example:**

Attribute	Value
Disease detected	Yes, No
Result	Pass, Fail

- **Ordinal Attributes**
- All Values have a meaningful order.
- For example, Grade-A means highest marks, B means marks are less than A, C means marks are less than grades A and B, and so on. Ordinal Attributes are Quantitative Attributes.
- **Examples of Ordinal Attributes**

Attribute	Value
Grade	A, B, C, D, F
BPS- Basic pay scale	16, 17, 18

- **Discrete Attributes**
- Discrete data have a finite value. It can be in numerical form and can also be in a categorical form.
- Discrete Attributes are Quantitative Attributes.
- **Examples of Discrete Data**

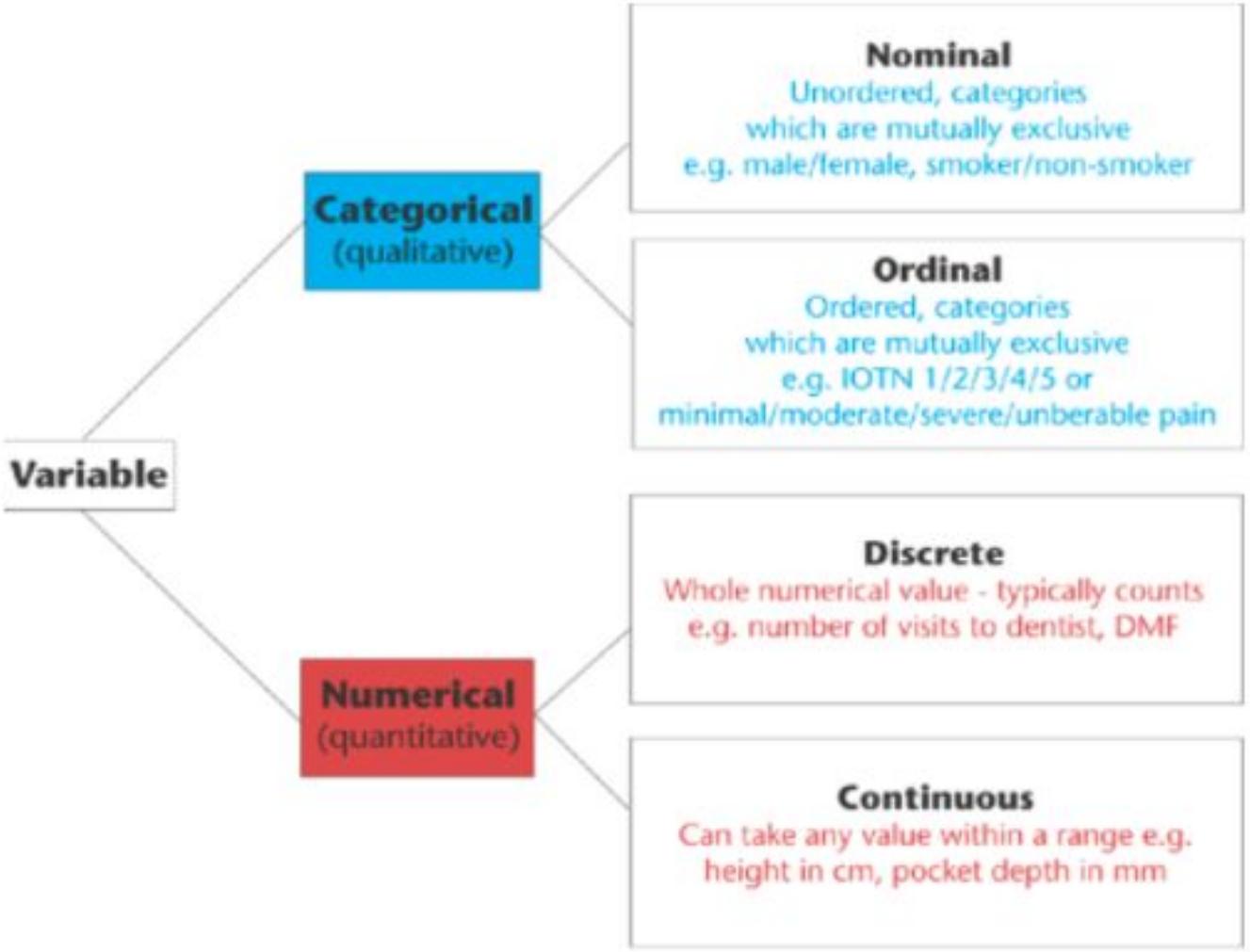
Attribute	Value
Profession	Teacher, Bussiness Man, Peon etc
Postal Code	42200, 42300 etc

Example of Continuous Attribute

- Continuous data technically have an infinite number of steps.
- Continuous data is in float type.
- There can be many numbers in between 1 and 2.
- These attributes are Quantitative Attributes.
- **Example of Continuous Attribute**

Attribute	Value
Height	5.4…, 6.5….. etc
Weight	50.09…. etc

- ***Quantitative (Numerical) vs Qualitative (Categorical)***
- There are other ways of classifying variables that are common in statistics. One is qualitative vs. quantitative.
- Qualitative variables are descriptive/categorical.
- Many statistics, such as mean and standard deviation, do not make sense to compute with qualitative variables.
- Quantitative variables have numeric meaning, so statistics like means and standard deviations make sense.


Variable
Categorical
(qualitative)
Nominal
Unordered, categories
which are mutually exclusive
e.g. male/female, smoker/non-smoker
Ordinal
Ordered, categories
which are mutually exclusive
e.g. IOTN 1/2/3/4/5 or
minimal/moderate/severe/unberable pain
Numerical
(quantitative)
Discrete
Whole numerical value - typically counts
e.g. number of visits to dentist, DMF
Continuous
Can take any value within a range e.g.
height in cm, pocket depth in mm

Attribute Types

Four main types of attributes

- ***Nominal: Categorical (Qualitative)***

categories, states, or "names of things"

Hair color, marital status, occupation, ID numbers, zip codes

An important nominal attribute: Binary

Nominal attribute with only 2 states (0 and 1)

- ***Ordinal: Categorical (Qualitative)***

Values have a meaningful order (ranking) but magnitude between successive values is not known.

Size = {small, medium, large}, grades, army rankings

- ***Interval: Numeric (Quantitative)***

  Measured on a scale of equal-sized units – Values have order:

  - **temperature in C° or F°, calendar dates**

  No true zero-point: ratios are not meaningful

- ***Ratio: Numeric (Quantitative)***

  Inherent zero-point: ratios are meaningful

  - **temperature in Kelvin, length, counts, monetary quantities**

- ***Four main types of attributes:***
- ***Ordinal Attributes***
- ***An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known.***
- ***Example: An ordinal attribute size corresponds to the size of available at a fast-food restaurant.***
- ***– This attribute has three possible values: small, medium, and large.***
- ***– The values have a meaningful sequence (which corresponds to increasing size); however, we cannot tell from the values how much bigger, say, a medium is than a large.***
- ***• The central tendency of an ordinal attribute can be represented by its mode and its median (middle value in an ordered sequence), but the mean cannot be defined***

- ***Interval Attributes***
- *Interval attributes are measured on a scale of equal-size units.*
- *We can compare and quantify the difference between values of interval attributes.*
- ***Example:*** *A temperature attribute is an interval attribute.*
- *We can quantify the difference between values.*
- *For example, a temperature of 20oC is five degrees higher than a temperature of 15oC.*
- *Temperatures in Celsius do not have a true zero-point, that is, 0oC does not indicate "no temperature."*
- *Although we can compute the difference between temperature values, we cannot talk of one temperature value as being a multiple of another.*
- *Without a true zero, we cannot say, for instance, that 10oC is twice as warm as 5oC .*
- *That is, we cannot speak of the values in terms of ratios.*
- *The central tendency of an interval attribute can be represented by its mode, its median (middle value in an ordered sequence), and its mean.*

- ***Ratio Attributes***
- *A ratio attribute is a numeric attribute with an inherent zero-point.*
- ***Example: A number_of_words attribute is a ratio attribute.***
- If a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value.
- The central tendency of an ratio attribute can be represented by its mode, its median(middle value in an ordered sequence), and its mean.

- ***Properties of Attribute Values***
- ***The type of an attribute depends on which of the following properties it possesses:***
- *Distinctness:* = □
- *Order:* < >
- *Addition:* + -
- *Multiplication:* * /
- ***Nominal attribute:*** ***Distinctness***
- ***Ordinal attribute:*** ***distinctness & order***
- ***Interval attribute:*** ***distinctness, order & addition***
- ***Ratio attribute:*** ***all 4 properties***

Properties of Attribute Values

Attribute Type	Description	Examples
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, □)	zip codes, employee ID numbers, eye color, sex: {male, female}
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {good, better, best}, grades, street numbers
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length,

- ***Attribute Types***
- *Categorical (Qualitative) and Numeric (Quantitative)*
- *Nominal and Ordinal attributes are collectively referred to as categorical or qualitative attributes.*
- *qualitative attributes, such as employee ID, lack most of the properties of numbers.*
- *Even if they are represented by numbers, i.e. , integers, they should be treated more like symbols .*
- *Mean of values does not have any meaning.*
- *Interval and Ratio are collectively referred to as quantitative or numeric attributes.*
- *Quantitative attributes are represented by numbers and have most of the properties of numbers .*
- *Note that quantitative attributes can be integer-valued or continuous.*
- *Numeric operations such as mean, standard deviation are meaningful*

- ***Discrete vs. Continuous Attributes***
- ***Discrete Attribute***
- Has only a finite or countably infinite set of values
- zip codes, profession, or the set of words in a collection of documents
- Sometimes, represented as integer variables
- Note: Binary attributes are a special case of discrete attributes
- Binary attributes where only non-zero values are important are called asymmetric binary attributes.
- ***Continuous Attribute***
- Has real numbers as attribute values
- temperature, height, or weight
- Practically, real values can only be measured and represented using a finite number of digits
- Continuous attributes are typically represented as floating-point variables

- **Example of attribute**
- In this example, RollNo, Name, and Result are attributes of the object named as a student.

Rollo	Name	Result
1	Ali	Pass
2	Akram	Fail

- ***Nominal Attributes***
- Nominal data is in alphabetical form and not in an integer.
- Nominal Attributes are Qualitative Attributes.
- Examples of Nominal attributes

Attribute	Value
Categorical data	Lecturer, Assistant Professor , Professor
States	New, Pending, Working, Complete, Finish
Colors	Black, Brown, White, Red

- ***Binary Attributes***
- Binary data have only two values/states.
- For example, here Disease detected can be only Yes or No.
- Binary Attributes are Qualitative Attributes.
- Examples of Binary Attributes

Attribute	**Value**
Disease detected	Yes, No
Result	Pass, Fail

- **The binary attribute is of two types;**
- **Symmetric binary**
- **Asymmetric binary**
- Examples of Symmetric data
- Both values are equally important. For example, if we have open admission to our university, then it does not matter, whether you are a male or a female.
- Example:

Attribute	Value
Gender	Male, Female

- ***Examples of Asymmetric data***
- Both values are not equally important.
- For example, Disease detected is more important than Disease not detected.
- If a patient is with Disease and we ignore him, then it can lead to death but if a person is not Disease detected and we ignore it, then there is no special issue or risk.
- **Example:**

Attribute	Value
Disease detected	Yes, No
Result	Pass, Fail

- **Ordinal Attributes**
- All Values have a meaningful order.
- For example, Grade-A means highest marks, B means marks are less than A, C means marks are less than grades A and B, and so on. Ordinal Attributes are Quantitative Attributes.
- **Examples of Ordinal Attributes**

Attribute	Value
Grade	A, B, C, D, F
BPS- Basic pay scale	16, 17, 18

- **Discrete Attributes**
- Discrete data have a finite value. It can be in numerical form and can also be in a categorical form.
- Discrete Attributes are Quantitative Attributes.
- **Examples of Discrete Data**

Attribute	Value
Profession	Teacher, Bussiness Man, Peon etc
Postal Code	42200, 42300 etc

Example of Continuous Attribute

- Continuous data technically have an infinite number of steps.
- Continuous data is in float type.
- There can be many numbers in between 1 and 2.
- These attributes are Quantitative Attributes.
- **Example of Continuous Attribute**

Attribute	Value
Height	5.4…, 6.5….. etc
Weight	50.09…. etc

- **Basic Statistical Descriptions of Data**
- ***Statistical description can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.***
- Basic statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.
- These descriptive statistics are of great help in understanding the distribution of the data.
- ***Three areas of basic statistical descriptions:***
- ***Measure of Central Tendency : Which measure the location of the middle or center of a data distribution.***
- Measures of central tendency include mean, median, mode, and midrange.
- **Measures of data dispersion** spread of the data which include quartiles, interquartile range (IQR), and variance.
- **Graphic Display of Basic Statistical Description of Data:**
- Graphical data representation of basic statistical descriptions(Quantile Plot, Histogram, Scatter Plot etc)

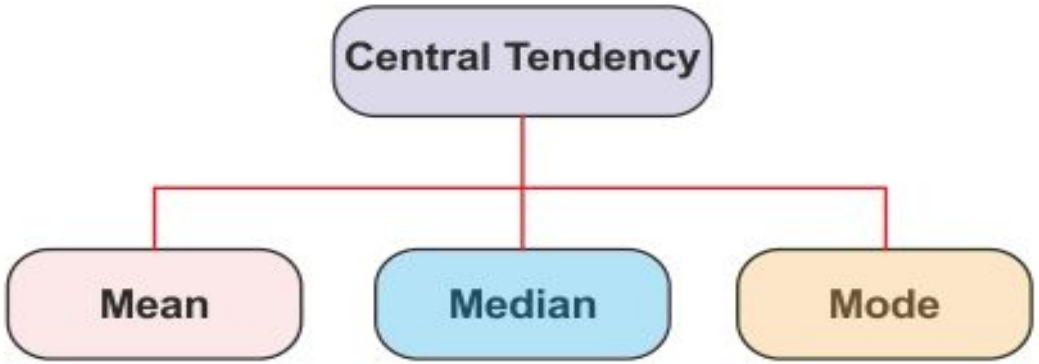
- ***Measures of Central Tendency & Dispersion***
- Measures that indicate the approximate center of a distribution are called measures of central tendency.
- Measures that describe the spread of the data are measures of dispersion. These measures include the mean, median, mode, range, upper and lower quartiles, variance, and standard deviation.

Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

Measure of Central Tendency

- A measure of central tendency is an important aspect of quantitative data.
- It is an estimate of a "typical" value.
- Three of the many ways to measure central tendency are the **mean**, **median** and **mode**.
- A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data.
- Measures of central tendency are sometimes called measures of central location.
- The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.
- It is also called as categorized as summary statistics.

Measures of central tendency are a key approach to address and communicate with graphs In real-world applications, you can practice tables and graphs of numerous types to present information and to obtain information from data that helps in the analyses and predictions.



The central tendency is said to be the statistical model that represents the single value of the entire distribution or database and aims to implement an exact description of the entire data in the distribution. There are three main measures of central tendency – Mean, Median and Mode.

- ***Measures of Central Tendency***
- Measures that indicate the approximate center of a distribution are called measures of central tendency.

A. Finding the Mean

- The mean of a set of data is the sum of all values in a data set divided by the number of values in the set. It is also often referred to as an arithmetic average. The Greek letter (“mu”) is used as the symbol for population mean and the symbol is used to represent the mean of a sample. To determine the mean of a data set:

1. Add together all of the data values.
2. Divide the sum from Step 1 by the number of data values in the set.

Formula:

$$mean = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Example:

Consider the data set: 17, 10, 9, 14, 13, 17, 12, 20, 14

$$mean = \frac{\sum x_i}{n} = \frac{17 + 10 + 9 + 14 + 13 + 17 + 12 + 20 + 14}{9} = \frac{126}{9} = 14$$

The mean of this data set is **14**.

- ***Measuring Central Tendency: Mean***

The most common and most effective numerical measure of the "center" of a set of data is the arithmetic mean.

Arithmetic Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sometimes, each value x_i in a set may be associated with a weight w_i.

- The weights reflect the significance and importance attached to their respective values.

Weighted Arithmetic Mean: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

Arithmetic Mean of Ungrouped Data

$If x_1,\ x_2,\ x_3, \ldots\ldots,\ x_n$ are n values of a variable X, then the arithmetic mean of these values is presented by the following methods:

Direct Method: If a variable X carries value $x_1,\ x_2,\ x_3, \ldots\ldots,\ x_n$ including corresponding frequencies $f_1,\ f_2,\ f_3, \ldots\ldots,\ f_n$ respectively then the arithmetic mean of these values is presented by:

$$\overline{x} = \frac{\sum x}{n}$$

Mean = (Sum of all the observations/Total number of observations)

Example: What is the mean of 2, 8, 10, 6 and 14?

$Step\ 1:\ First\ add\ all\ the\ numbers.$

$2 + 8 + 10 + 6 + 14 = 40$

$Step\ 2:\ Now\ divide\ by\ 5\ (here\ 5\ is\ the\ total\ number\ of\ observations).$

$Mean = \frac{40}{5} = 8$

Arithmetic Mean of Grouped Data

- For computing, arithmetic means in a continuous frequency distribution, first, we need to find the midpoint of class intervals (x), and then we need to multiply the mid-points with the corresponding frequency of the interval (fx).
- The sum of this product is obtained and finally, by dividing the sum of this product by the sum of frequencies we will obtain the arithmetic mean of the continuous frequency distribution.

Question: Find the mean of the below distribution.

x	6	4	15	9	10
f	10	5	8	10	7

Solution:

Calculation table for arithmetic mean:

x_i	f_i	x_if_i
6	10	60
4	5	20
15	8	120
9	10	90
10	7	70
–	$\sum f_i = 40$	$\sum x_i f_i = 360$

$$Mean = \bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{360}{40} = 9$$

$$Therefore\ Mean = 9$$

- ***Measuring Central Tendency: Mean***
- Although the mean is the single most useful quantity for describing a data set, it is not always the best way of measuring the center of the data.
- A major problem with the mean is its sensitivity to extreme (outlier) values.
- Even a small number of extreme values can corrupt the mean.
- To offset the effect caused by a small number of extreme values, we can instead use the trimmed mean,
- Trimmed mean can be obtained after chopping off values at the high and low extremes.

- ***Measuring Central Tendency: Median***
- Another measure of the center of data is the median.
- Suppose that a given data set of N distinct values is sorted in numerical order.
- If N is odd, the median is the middle value of the ordered set;
- If N is even, the median is the average of the middle two values.
- In probability and statistics, the median generally applies to numeric data; however, we may extend the concept to ordinal data.
- Suppose that a given data set of N values for an attribute X is sorted in increasing order.
- If N is odd, then the median is the middle value of the ordered set.
- If N is even, then the median may not be not unique.
- In this case, the median is the two middlemost values and any value in between.

- ***What is Median of Grouped Data?***
- Median of a grouped data is data that is arranged in ascending order and is written in a continuous manner.
- The data is in the form of a frequency distribution table that divides the higher level of data from the lower level of data.
- One of the simplest methods of finding the median of grouped data is by using the formula.
- As finding the middle value or median of a grouped data might be tough.
- Therefore, to find the median for grouped data we can use the following steps and formula:
- Step 1: Find the total number of observations.
- Step 2: Define the class size, and divide the data into different classes.
- Step 3: Calculate the cumulative frequency of each class.
- Step 4: Identify the class in which the median falls. (Median Class is the class where n/2 lies.)
- Step 5: Find the lower limit of the median class(l), and the cumulative frequency of the median class (c).

Median for Grouped Data

$$\text{Median} = l + \left[\frac{\frac{n}{2} - c}{f} \right] \times h$$

Where,

- l = lower limit of median class
- n = total number of observations
- c = cumulative frequency of the preceding class
- f = frequency of each class
- h = class size

- ***For example: Let's consider the data: 48, 20, 50, 69, 73. What is the median?***
- ***Solution:***
- Arranging in ascending order, we get: 20, 48, 50, 69, 73. Here, n (no.of observations) = 5
- So, to find the median of odd data we use the formula:
- [(n+1)/2] = (5 + 1)/2 = 6/2 = 3
- Therefore, Median = 3rd observation
- Median = 50.

- ***Steps to Find Median of Grouped Data***
- Median of grouped data is in the form of a frequency distribution arranged in ascending order and is continuous.
- Find the median of any given data is simple since the median is the middlemost value of the data. Since the data is grouped, it is divided into class intervals.
- Let us learn the steps to finding the median of grouped data.
- **Step 1: Construct the frequency distribution table with class intervals and frequencies.**
- **Step 2: Calculate the cumulative frequency of the data by adding the preceding value of the frequency with the current value.**
- **Step 3: Find the value of n by adding the values in frequency.**
- **Step 4: Find the median class. If n is odd, the median is the (n+1/2). And if n is even, then the median will be the average of the n/2th and the (n/2 +1)th observation.**
- **Step 5: Find the lower limit of the class interval and the cumulative frequency.**
- **Step 6: Apply the formula for median for grouped data: Median = l + [(n/2−c)/f] × h**

For Example: Calculate the median for the following data:

Marks	0 - 10	10 - 30	30 - 60	60 - 80	80 - 90
Number of students	6	20	37	10	7

Solution:

We need to calculate the cumulative frequencies to find the median.

Marks	**Number of students**	**Cumulative frequency**	
0 - 20	6	0 + 6	6
20 - 40	20	6 + 20	26
40 - 60	37	26 + 37	63
60 - 80	10	63 + 10	73
80 - 100	7	73 + 7	80

- N = sum of cf = 80, N/2 = 80/2 = 40
- Since n is even, we will find the average of the $n/2^{th}$ and the $(n/2 +1)^{th}$ observation i.e. the cumulative frequency greater than 40 is 63 and the class is 40 - 60. Hence, the median class is 40 - 60.
- l = 40, f = 37, c = 26, h = 20
- Using Median formula:
- Median = $l + [(n/2 - c)/f] \times h$
- $= 40 + [(37 - 26)/40] \times 20$
- $= 40 + (11/40) \times 20$
- $= 40 + (220/40)$
- $= 40 + 5.5$
- $= 45.5$
- Therefore, the median is 45.5.

Median

The data of the middlemost observation that is achieved after modifying the data in ascending order is termed the median of the data. The advantage of applying the median as a central tendency is that it is less influenced by outliers and skewed data.

Median for Ungrouped Data

- First, the data is arranged in either ascending or decreasing order.
- To determine the median, we need to recognize if n is even or odd.
- If the data set holds an odd number of values i.e n=odd, then the median is given by the formula.

$$Median = \frac{(n+1)^{th}}{2}\ observation.$$

- If the dataset holds an even number of data values i.e n = even, the median is computed by the formula.

$$Median = \frac{\left\{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}\ observation\right\}}{2}$$

- ***Measuring Central Tendency: Mode***
- Another measure of central tendency is the mode.
- The mode for a set of data is the value that occurs most frequently in the set.
- It is possible for the greatest frequency to correspond to several different values, which results in more than one mode.
- Data sets with one, two, or three modes: called unimodal, bimodal, and trimodal. – At the other extreme, if each data value occurs only once, then there is no mode.
- Central Tendency Measures for Numerical Attributes: Mean, Median, Mode
- Central Tendency Measures for Categorical Attributes: Mode (Median?)
- Central Tendency Measures for Nominal Attributes: Mode
- Central Tendency Measures for Ordinal Attributes: Mode, Median

Mode

The median is the middle character in a data set when the numbers are presented in ascending or descending order. Whereas the mode is the value that happens to appear most often in a data set and the range is the difference between the highest and lowest values in a data set. The types of modes are as follows.

Unimodal Mode – A collection of data/numbers with one mode is recognised as a unimodal mode.

For example, the mode of data set B = { 20, 14, 16, 17, 14, 18, 14, 19} is 14 as there is simply one value replicating itself. Therefore, it is a unimodal data set.

Bimodal Mode – A set of data including two modes is identified as a bimodal model. This indicates that there are two data values that possess the highest frequencies.

For example, the mode of data set B = { 8, 12, 12, 14, 15, 19, 17, 19} is 12 and 19 as both 12 and 19 are repeating twice in the given set. Therefore, the given set is a bimodal data set.

Trimodal Mode – A collection of data including three modes is identified as a trimodal mode. This implies that there are three data values that are holding the highest frequencies.

For example, the mode of data set B = {2, 2, 2, 3, 7, 7, 5, 6, 5,4, 7, 5, 8} is 2, 7, and 5 because all the three values are recurring thrice in the given set. Therefore, it is a trimodal data set.

Multimodal Mode – A set of data including four or more than four modes is recognised as a multimodal model.

For example, The mode of data set B = {101, 82, 82, 95, 95, 100, 90, 90,101 ,96 } is 82, 90, 95, and 101 because all the four values are recurring twice in the given set.

Now moving towards the mode formula; for **ungrouped data**, we only need to identify the observation which occurs at maximum times.

Mode = Observation with maximum frequency

For example in the data set: 7, 8, 9, 2, 4, 7, 7, 6, 3 the value 7 appears the most number of times.

Thus, the mode is equal to 6 for the set.

For **grouped data** or when the data is continuous, the mode can be determined using the following rules:

Step 1: Find the modal class i.e. the class with the highest frequency.

Step 2: Find mode applying the following formula:

$$Mode = l + \left\{ \frac{f_m - f_1}{2f_m - f_1 - f_2} \right\} \times h$$

$Where,$

$l = \ lower\ limit\ of\ modal\ class,$

$f_m = frequency\ of\ modal\ class,$

$f_1 = frequency\ of\ class\ preceding\ modal\ class,$

$f_2 = frequency\ of\ class\ succeeding\ modal\ class\ and\ h = class\ width.\ h$

Example: Determine the modal value for the given set of data.

12	14	20	18	22	16	20	20	11	12	16	14	20	22

Solution:

Arranging the above set of data in ascending order.

11	12	12	14	14	16	16	18	20	20	20	20	22	22

Here, we get 20 four times, 12, 14, 16, 22 twice each, and other terms i.e 11 and 18 only once.

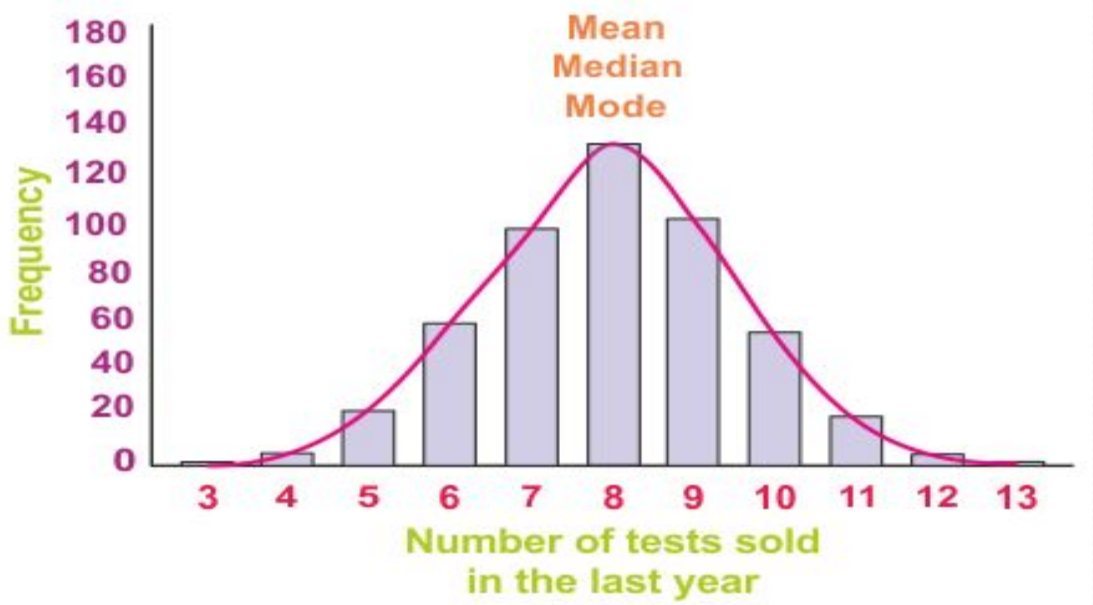
Therefore, the mode for a given set of data is 20.

Distributions and Central Tendency

A data set in central tendency is a distribution of *n* number of records or values. Distributions are basically of two types:

Normal distribution

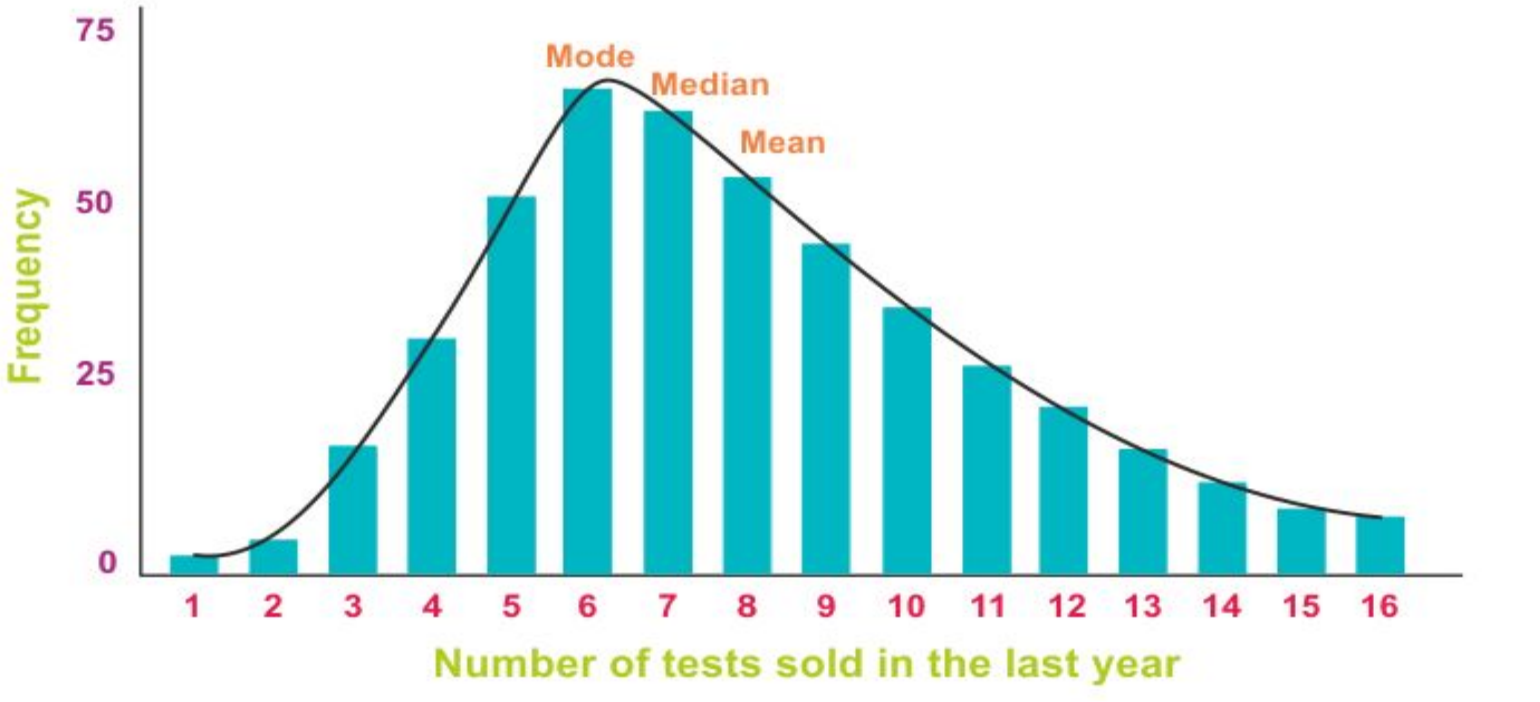
In a normal distribution(symmetric distribution), information or data is symmetrically distributed with no skew. Most of the values cluster around a central area, with values decreasing off as they travel away from the center. The mean, mode and median are the same in a normal distribution.



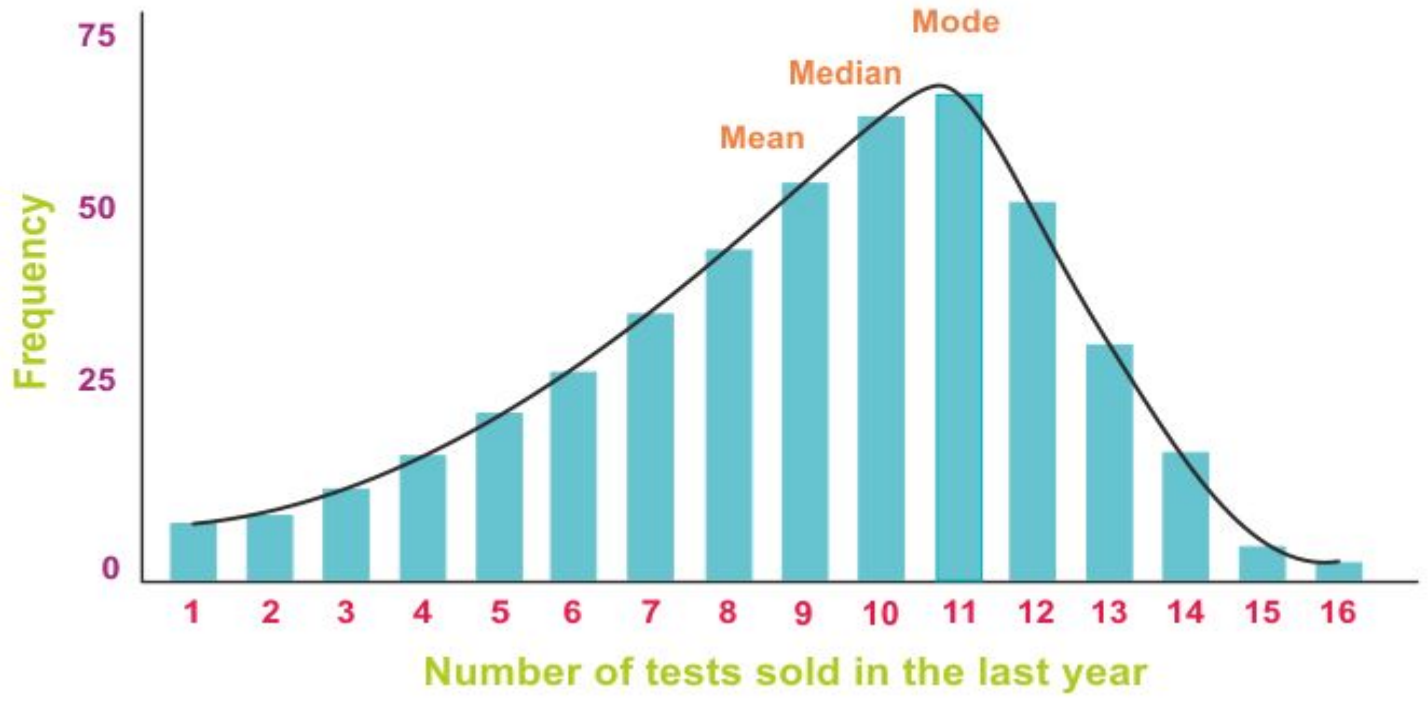
Skewed distributions

In skewed distributions, most of the values lie on one side of the center rather than the other, and the mean, median and mode all vary from one another. The direction of this tail shows us the side of the skew.

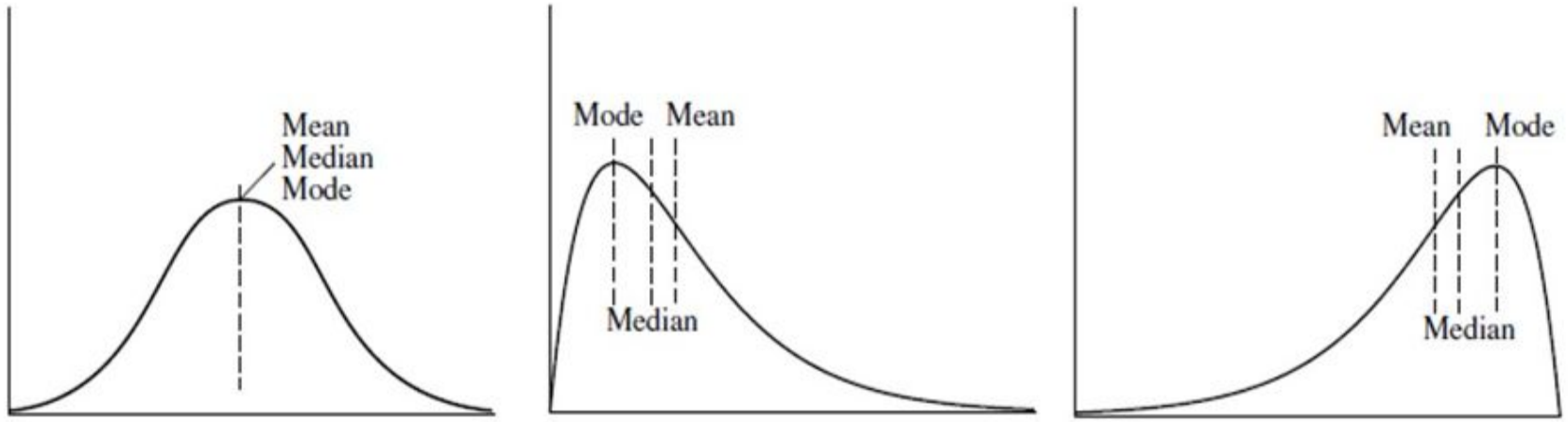
- If Right-skewed or positively skewed, the tail reaches out to the right.
- Mean > median > mode



- Left-skewed or negatively skewed, the tail reaches out to the left.
- Mode > median > mean



- ***Measuring Central Tendency Mean, Median, Mode***
- ***Median, mean and mode of symmetric, positively and negatively skewed data***



symmetric data *positively skewed data* *negatively skewed data*

The shape of the data helps us to determine the most appropriate measure of central tendency. The three most significant descriptions of shape are symmetric, left-skewed, and right-skewed. Skewness is a measure of the degree of asymmetry of the distribution.

•*Measuring Central Tendency: Example*

- *What are central tendency measures (mean, median, mode)for the following attributes?*
- *attr1 = {2,4,4,6,8,24}*
- *attr2 = {2,4,7,10,12}*
- *attr3 = {xs,s,s,s,m,m,l}*

- *Measuring Central Tendency: Example*
- *What are central tendency measures (mean, median, mode)for the following attributes? attr1 = {2,4,4,6,8,24}*
- *mean = (2+4+4+6+8+24)/6 = 8 average of all values*
- *median = (4+6)/2 = 5 avg. of two middle values*
- *mode = 4 attr2 = {2,4,7,10,12} most frequent item*
- *mean = (2+4+7+10+12)/5 = 7 average of all values*
- *median = 7 middle value*
- *mode = any of them (no mode) all of them has same freq.*
- *attr3 = {xs,s,s,s,m,m,l} mean is meaningless for categorical attributes.*
- *median = s middle value mode = s most frequent item*

Measuring Dispersion of Data

- **The degree to which numerical data tend to spread is called the dispersion, or variance of the data.**

The most common ***measures of data dispersion:***

- **Range:** Difference between the largest and smallest values.
- **Interquartile Range (IQR):** range of middle 50%
  - **quartiles**: Q1 (25th percentile), Q3 (75th percentile) IQR=Q3-Q1
  - **five number summary**: Minimum, Q1, Median,Q3, Maximum
- **Variance and Standard Deviation:** (*sample: s, population: σ)*
  - **variance** of N observations:

$$\sigma^2 = \frac{1}{n}\sum_{1}^{n}(x_i - \mu)^2 \qquad s^2 = \frac{1}{n-1}\sum_{1}^{n}(x_i - \mu)^2$$

where is the mean value of the observations

standard deviation σ *(s)* is the square root of varianceσ^2 *(s^2)*

Variance and Standard Deviation

The extent of statistical data is estimated by the standard deviation. Standard Deviation is the square root of variance. It is a measure of the extent to which data deviates from the mean. It is indicated by the symbol, 'σ'. The variance of the given data set is the average square distance between the mean value and specific data value.

For ungrouped data

Use the formulae given below to find out the variance for ungrouped data:

$$\sigma^2 = \sum \left[(x - \overline{x})^2 \right] / n$$

x = Article/objects given in the data

$\overline{x}$ = Mean of the data

n = Total number of articles

The standard deviation for grouped data is determined as the square root of variance obtained.

The Population variance is given by the formula:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

The Population Standard Deviation is given by the formula:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$$

- The three measures of central values i.e. mean, median and mode are correlated by the following relations (called an empirical relationship):
- 2 Mean + Mode = Median
- Mean is the chosen measure of central tendency when information is normally distributed.
- Median is the most beneficial measure of central tendency when data is skewed.
- While working with nominal variables, the mode is the most helpful measure of central tendency.
- Mean and median can not be zero unless all information values are zero. However, there may be no mode in the dataset.

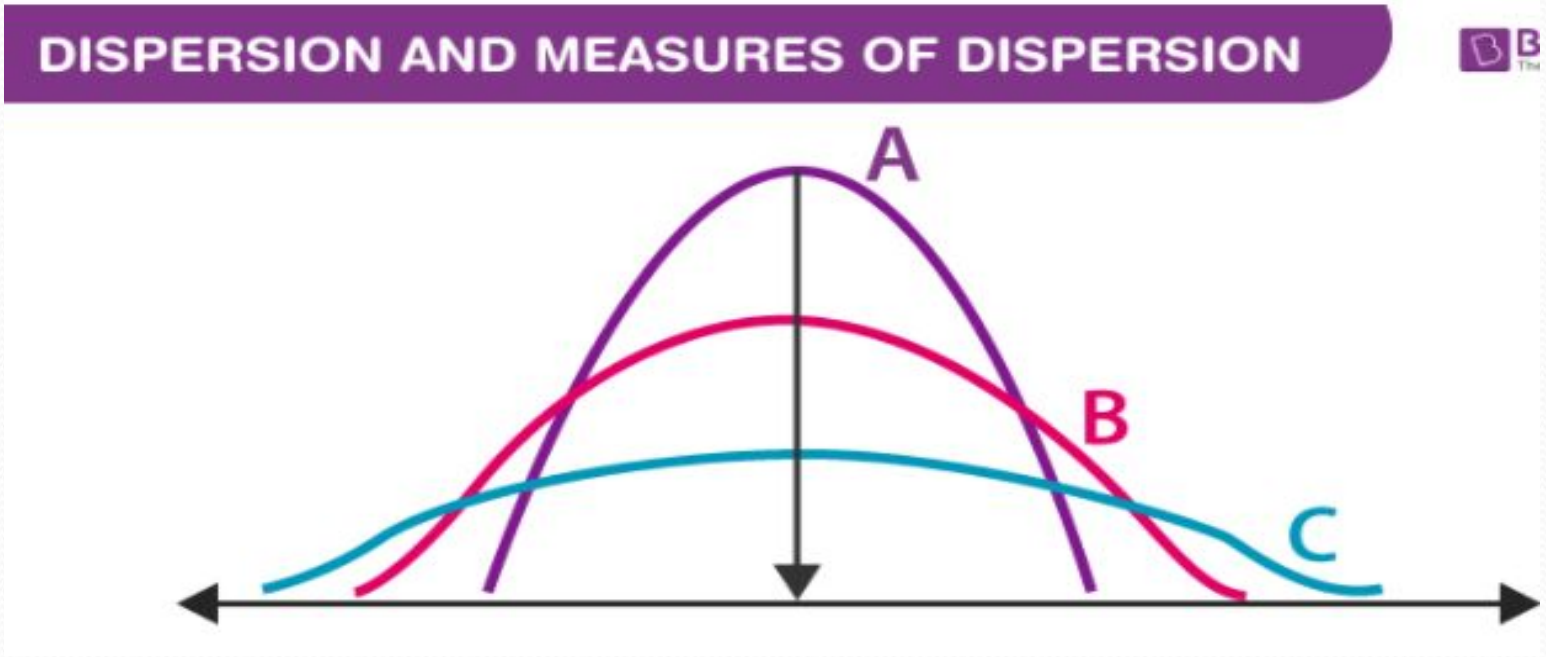
- **Standard Deviation and Variance:**
- Standard deviation and variance are two key measures commonly used in the financial sector.
- Standard deviation is the spread of a group of numbers from the mean.
- The variance measures the average degree to which each point differs from the mean.
- While standard deviation is the square root of the variance, variance is the average of all data points within a group.
- If the points are further from the mean, there is a higher deviation within the data but if they are closer to the mean, there is a lower deviation.
- So the more spread out the group of numbers are, the higher the standard deviation.
- The standard deviation is a statistical measure that people can use to determine how spread out numbers are in a data set.
- Variance, on the other hand, gives an actual value to how much the numbers in a data set vary from the mean.

- **Standard deviation and variance:**
- Suppose you have a series of numbers and you want to figure out the standard deviation for the group.
- The numbers are 4, 34, 11, 12, 2, and 26. We need to determine the mean or the average of the numbers.
- In this case, we determine the mean by adding the numbers up and dividing it by the total count in the group:
- $(4 + 34 + 18 + 12 + 2 + 26) \div 6 = 16$
- So the mean is 16. Now subtract the mean from each number then square the result:
- $(4 - 16)2 = 144$
- $(34 - 16)2 = 324$
- $(18 - 16)2 = 4$
- $(12 - 16)2 = 16$
- $(2 - 16)2 = 196$
- $(26 - 16)2 = 100$

- Variance= (144 + 324 + 4 + 16 + 196 + 100) ÷ 6 = 130.67
- This means we end up with a variance of 130.67.
- square root of the variance, which is 11.43

- The simple definition of the term variance is the spread between numbers in a data set.
- Variance is a statistical measurement used to determine how far each number is from the mean and from every other number in the set.
- You can calculate the variance by taking the difference between each point and the mean. Then square and average the results.
- Standard deviation measures how data is dispersed relative to its mean and is calculated as the square root of its variance.
- The further the data points are, the higher the deviation.
- Closer data points mean a lower deviation.
- In finance, standard deviation calculates risk so riskier assets have a higher deviation while safer bets come with a lower standard deviation.

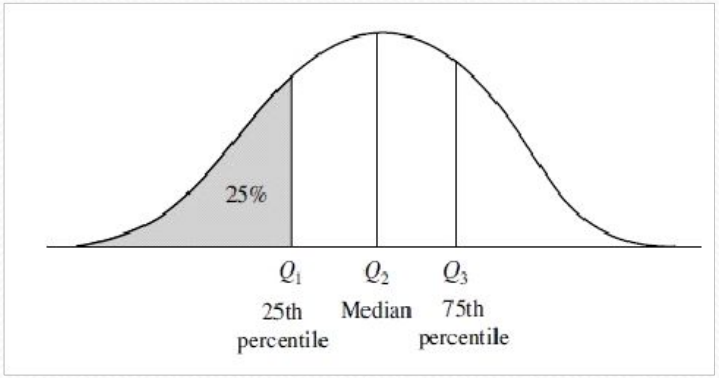
- ***What is Dispersion in Statistics?***
- Dispersion is the state of getting dispersed or spread. Statistical dispersion means the extent to which a numerical data is likely to vary about an average value.
- In other words, dispersion helps to understand the distribution of the data.



- Measures of Dispersion
- In statistics, the measures of dispersion help to interpret the variability of data i.e. to know how much homogenous or heterogeneous the data is. In simple terms, it shows how squeezed or scattered the variable is.
- ***The types of absolute measures of dispersion are:***
- **Range:** It is simply the difference between the maximum value and the minimum value given in a data set. Example: 1, 3,5, 6, 7 => Range = 7 -1= 6
- **Variance:** Deduct the mean from each data in the set then squaring each of them and adding each square and finally dividing them by the total no of values in the data set is the variance. Variance $(\sigma 2)=\sum(X-\mu)2/N$
- **Standard Deviation:** The square root of the variance is known as the standard deviation i.e. S.D. = $\sqrt{\sigma}$.
- **Quartiles and Quartile Deviation:** The quartiles are values that divide a list of numbers into quarters. The quartile deviation is half of the distance between the third and the first quartile.
- **Mean and Mean Deviation:** The average of numbers is known as the mean and the arithmetic mean of the absolute deviations of the observations from a measure of central tendency is known as the mean deviation (also called mean absolute deviation).

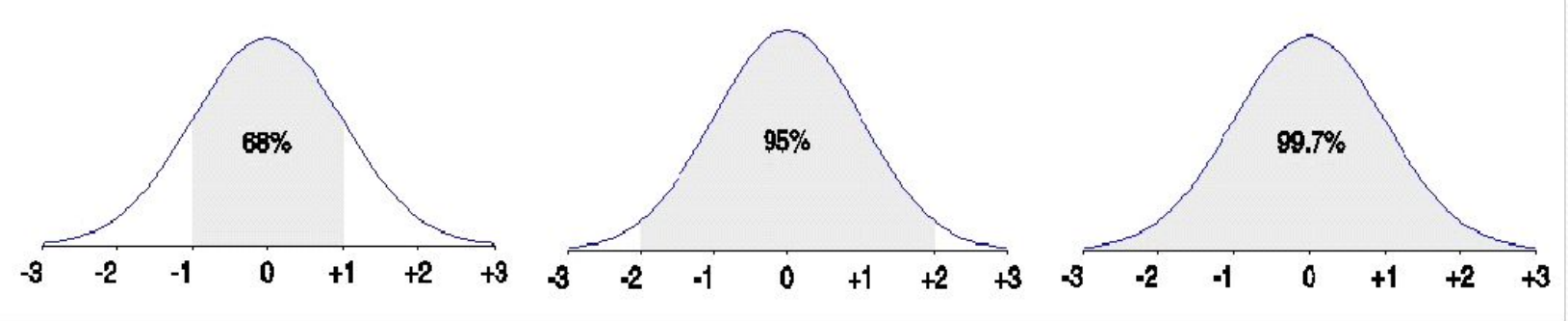
Measuring Dispersion of Data: Quartiles

- Suppose that set of observations for numeric attribute X is sorted in increasing order.
- **Quantiles** are points taken at regular intervals of a data distribution, dividing it into essentially equal size consecutive sets.
  - The **k^{th} q-quantile** for a given data distribution is the value x such that at most **k/q** of the data values are less than x and at most **(q-k)/q** of the data values are more than x, where k is an integer such that $0<k<q$. There are q-1 q-quantiles.
  - The 100-quantiles are more commonly referred to as **percentiles**; they divide the data distribution into 100 equal-sized consecutive sets.
- **Quartiles**: The 4-quantiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution.



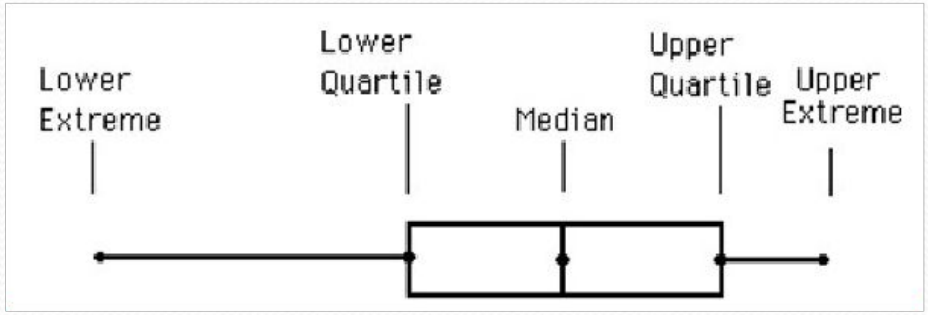
Measuring Dispersion of Data: Outliers

- **Outliers** can be identified by the help of ***interquartile range*** or ***standard deviation*** measures.
  - Suspected outliers are values falling at least 1.5xIQR above the third quartile or below the first quartile.
  - Suspected outliers are values that fall outside of the range of $\mu - N\sigma$ and $\mu + N\sigma$ where μ is mean and σ is standard deviation. N can be chosen as 2.5.
- The normal distribution curve: (μ: mean, σ: standard deviation)
  - From $\mu - \sigma$ to $\mu + \sigma$: contains about 68% of the measurements
  - From $\mu - 2\sigma$ to $\mu + 2\sigma$: contains about 95% of it
  - From $\mu - 3\sigma$ to $\mu + 3\sigma$: contains about 99.7% of it



Measuring Dispersion of Data: Boxplot Analysis

- **Five-number summary** of a distribution: Minimum, Q1, Median, Q3, Maximum
- **Boxplots** are a popular way of visualizing a distribution and a boxplot incorporates ***five-number summary***:
  - The ends of the box are at the **quartiles Q1 and Q3**, so that the box length is the interquartile range, IQR.
  - The **median** is marked by a line within the box. (**median** of values in IQR)
  - Two lines outside the box extend to the **smallest and largest observations** (outliers are excluded). Outliers are marked separately.
- If there are no outliers, lower extreme line is the smallest observation (Minimum) and upper extreme line is the largest observation (Maximum).

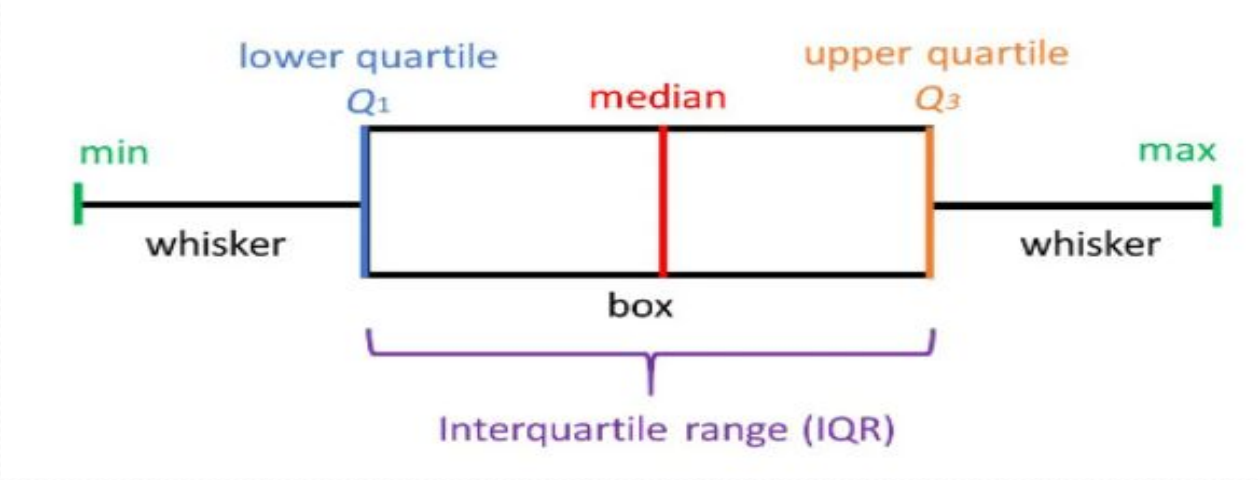


outliers outliers

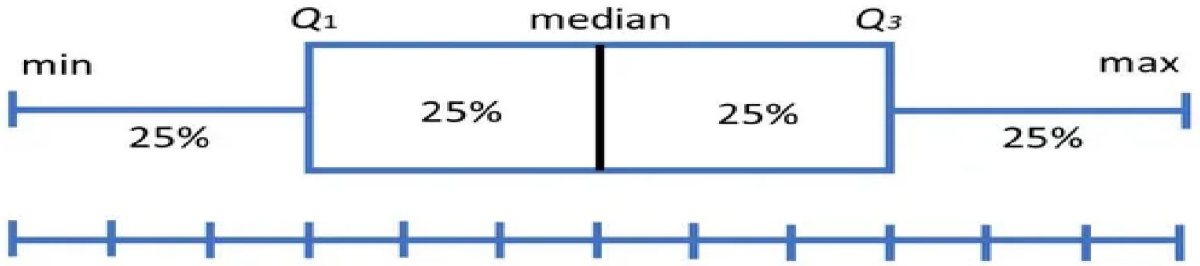
- ***Measuring Dispersion of Data:***
- Example
- Consider following two attribute values:
- attr1: {2,3,4,5,6,7,8,9}
- attr2: {1,5,9,10,11,12,18,30}
- Give interquartile ranges of attribute values?
- Are there any outliers (wrt IQR) in these datasets?
- Give a 4 element dataset whose standard deviation is zero?

- ***Measuring Dispersion of Data: Example***
- Consider following two attribute values:
- attr1: {2,3,4,5,6,7,8,9}
- attr2: {1,5,9,10,11,12,18,30}
- Give interquartile ranges of attribute values?
- attr1:
- Median=5+6/2=5.5
- Q1: (3+4)/2=3.5 Q3:(7+8)/2=7.5 IQR:3.5-7.5 = 4
- attr2: Q1: (5+9)/2=7 Q3:(12+18)/2=15 IQR:7-15 = 8
- Outlier>q3+1.5*IQR outlier<q1-1.5*IQR
- 7.5+1.5*4
- Are there any outliers (wrt IQR) in these datasets?
- Yes.
- attr2: Q1: (5+9)/2=7 Q3:(12+18)/2=15 IQR:7-15 = 8
- 30 in attr2. 30 > 15+1.5*IQR
- Give a 4 element dataset whose standard deviation is zero?
- {1,1,1,1}

- ***What is a box plot?***
- In descriptive statistics, a box plot or boxplot (also known as box and whisker plot) is a type of chart often used in explanatory data analysis.
- Box plots visually show the distribution of numerical data and skewness through displaying the data quartiles (or percentiles) and averages.
- Box plots show the five-number summary of a set of data: including the minimum score, first (lower) quartile, median, third (upper) quartile, and maximum score.



Box plots divide the data into sections that each contain approximately 25% of the data in that set.



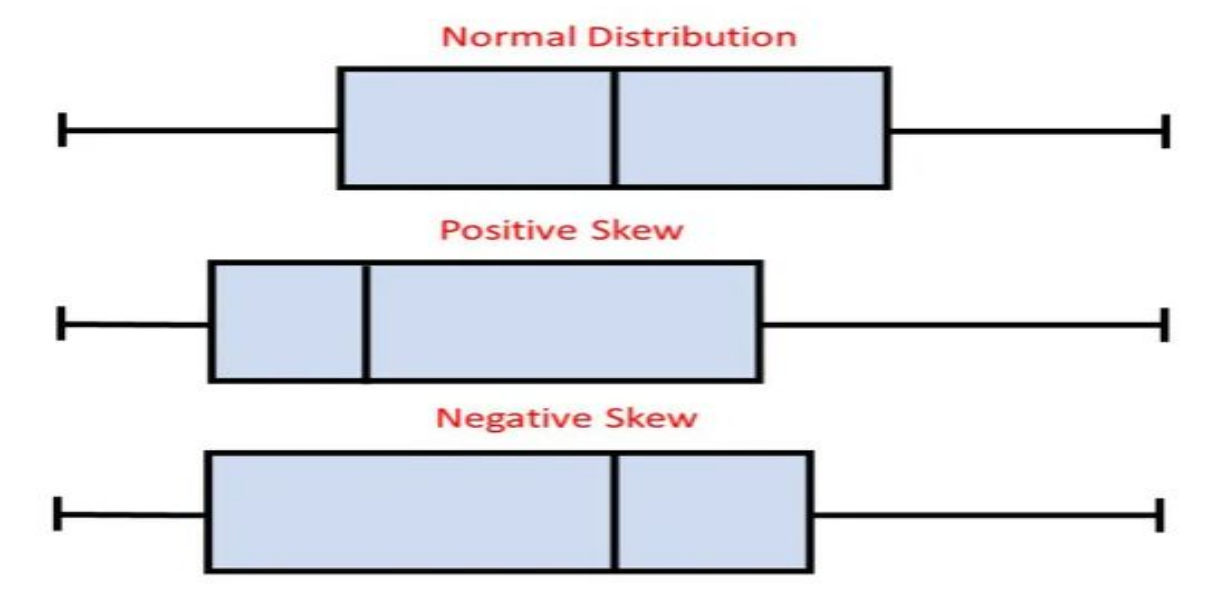
- ***Definitions***
- ***Minimum Score***
- ***The lowest score, excluding outliers (shown at the end of the left whisker).***
- ***Lower Quartile***
- ***Twenty-five percent of scores fall below the lower quartile value (also known as the first quartile).***
- ***Median***
- ***The median marks the mid-point of the data and is shown by the line that divides the box into two parts (sometimes known as the second quartile). Half the scores are greater than or equal to this value and half are less.***
- ***Upper Quartile***
- ***Seventy-five percent of the scores fall below the upper quartile value (also known as the third quartile). Thus, 25% of data are above this value.***
- ***Maximum Score***
- ***The highest score, excluding outliers (shown at the end of the right whisker).***

- ***Whiskers***
- ***The upper and lower whiskers represent scores outside the middle 50% (i.e. the lower 25% of scores and the upper 25% of scores).***
- ***The Interquartile Range (or IQR)***
- ***This is the box plot showing the middle 50% of scores (i.e., the range between the 25th and 75th percentile).***

- *Box plots are useful as they show the average score of a data set.*
- *The median is the average value from a set of data and is shown by the line that divides the box into two parts.*
- *Half the scores are greater than or equal to this value and half are less.*

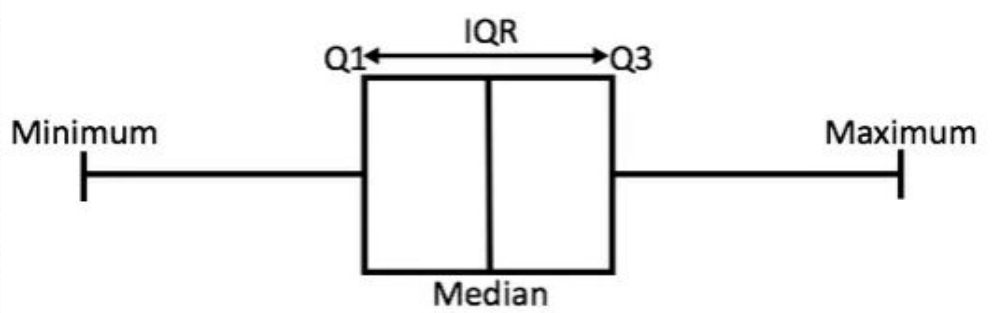
 Box plots are useful as they show the skewness of a data set

- *The box plot shape will show if a statistical data set is normally distributed or skewed.*



- ***When the median is in the middle of the box, and the whiskers are about the same on both sides of the box, then the distribution is symmetric.***
- ***When the median is closer to the bottom of the box, and if the whisker is shorter on the lower end of the box, then the distribution is positively skewed (skewed right).***
- ***When the median is closer to the top of the box, and if the whisker is shorter on the upper end of the box, then the distribution is negatively skewed (skewed left).***

- **In statistics, dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed.**
- The smallest value and largest value are found at the end of the ‘whiskers’ and are useful for providing a visual indicator regarding the spread of scores (e.g. the range).
- The interquartile range (IQR) is the box plot showing the middle 50% of scores and can be calculated by subtracting the lower quartile from the upper quartile (e.g. Q3−Q1).



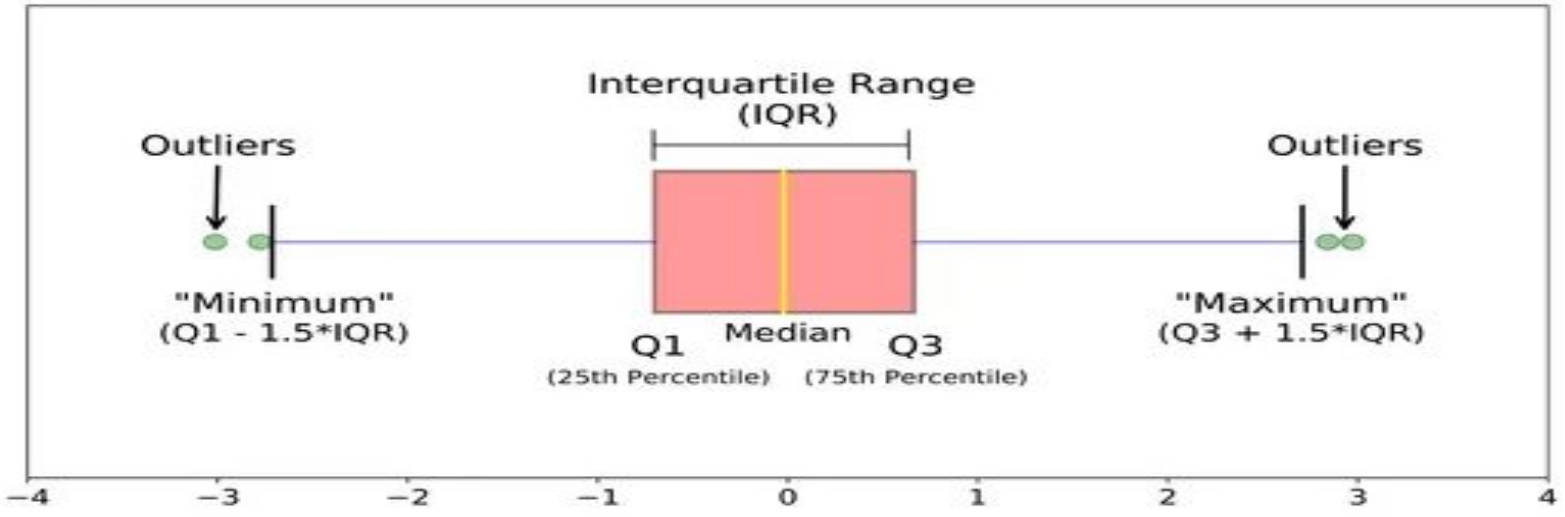
Box plots are useful as they show outliers within a data set.

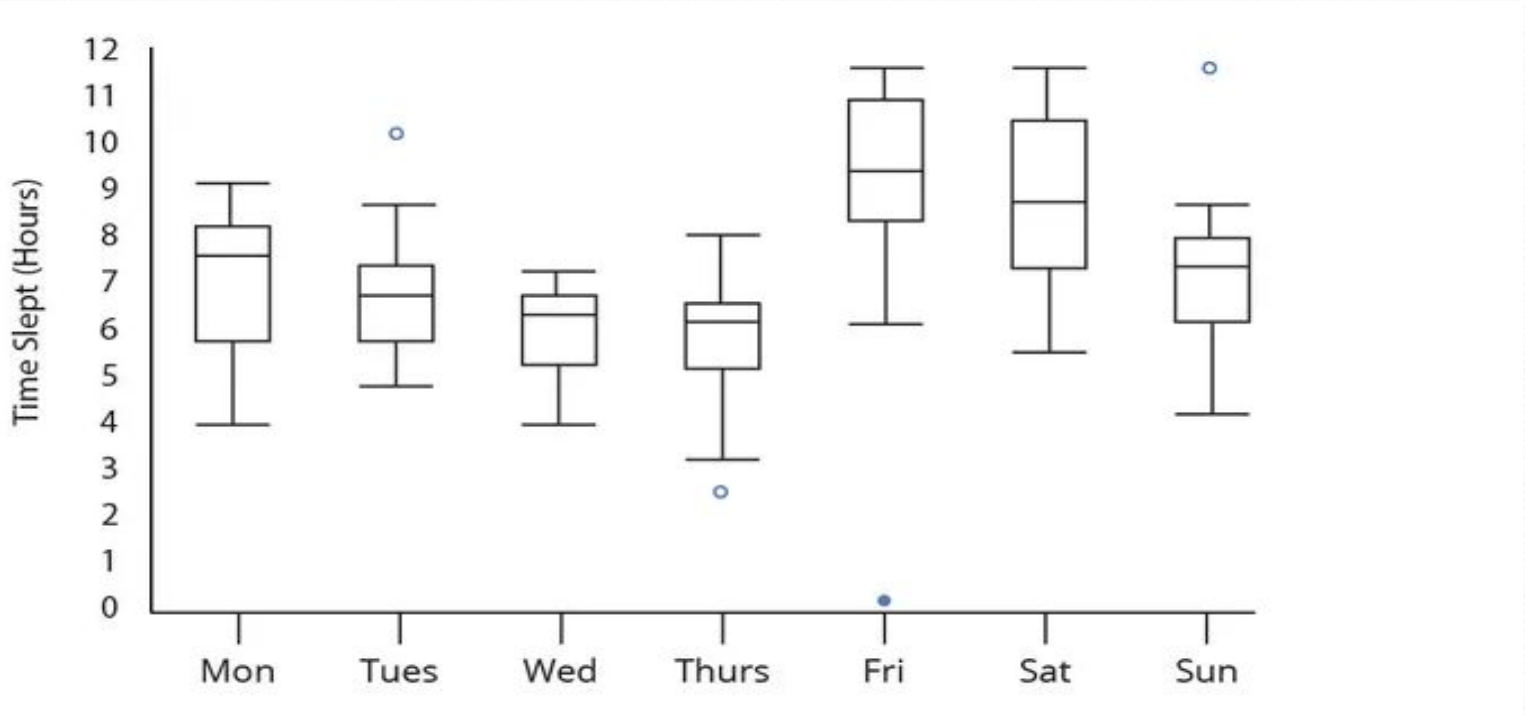
•An outlier is an observation that is numerically distant from the rest of the data.

•When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot.

•An outlier is an observation that is numerically distant from the rest of the data.

•When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot.



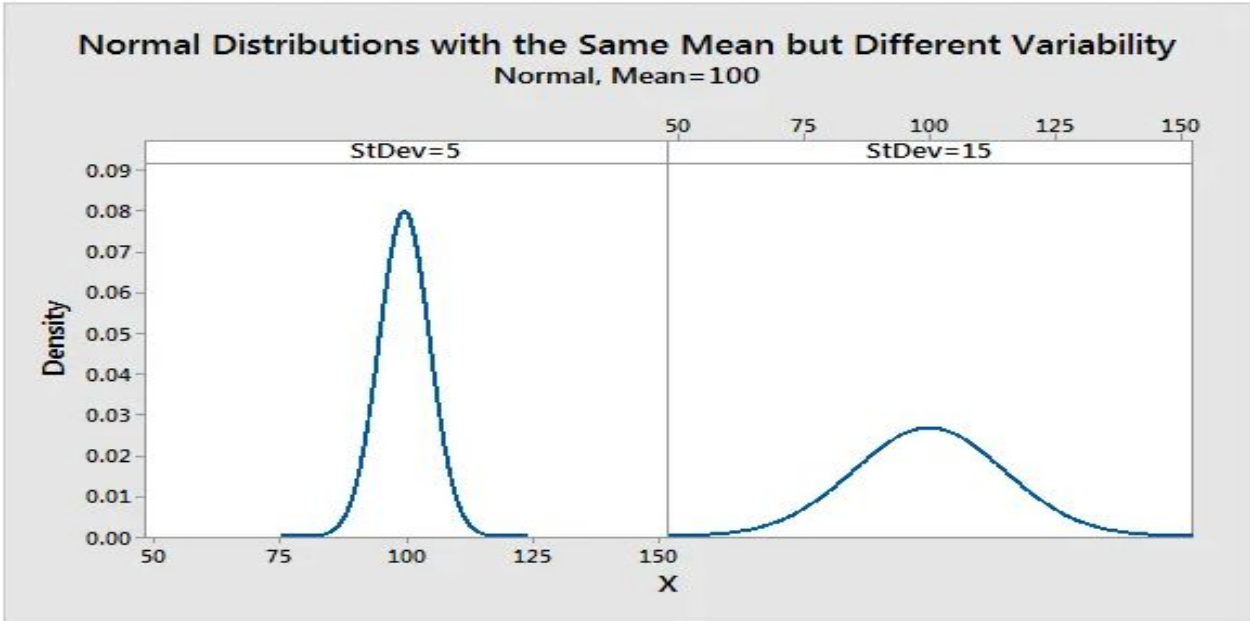

Time Slept (Hours)
12
11
10
9
8
7
6
5
4
3
2
1
0
Mon
Tues
Wed
Thurs
Fri
Sat
Sun

- *Detecting Outliers*
- *There are two simple ways you can detect outlier problem :*

1. Box Plot Method

- *If a value is higher than the 1.5*IQR above the upper quartile (Q3), the value will be considered as outlier.*
- *Similarly, if a value is lower than the 1.5*IQR below the lower quartile (Q1), the value will be considered as outlier.*
- *QR is interquartile range. It measures dispersion or variation. IQR = Q3 -Q1.*
- *Lower limit of acceptable range = Q1 - 1.5* (Q3-Q1)*
- *Upper limit of acceptable range = Q3 + 1.5* (Q3-Q1)*

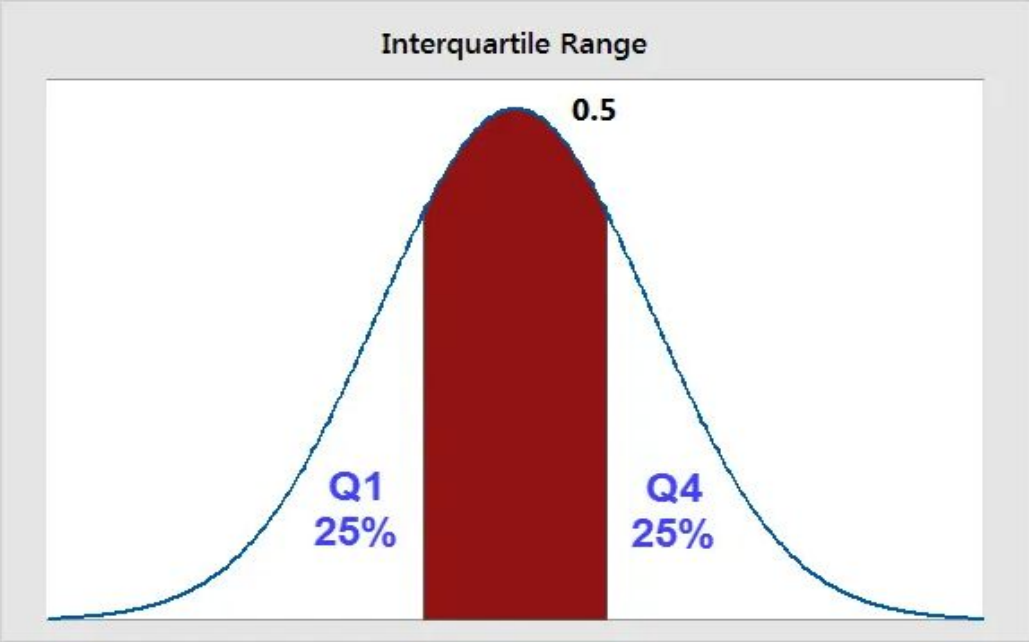
- ***Measures of Variability:***
- ***Range, Interquartile Range, Variance, and Standard Deviation***
- ***A measure of variability is a summary statistic that represents the amount of dispersion in a dataset.***
- ***How spread out are the values?***
- ***While a measure of central tendency describes the typical value, measures of variability define how far away the data points tend to fall from the center.***
- ***We talk about variability in the context of a distribution of values.***
- ***A low dispersion indicates that the data points tend to be clustered tightly around the center.***
- ***High dispersion signifies that they tend to fall further away.***


Normal Distributions with the Same Mean but Different Variability
Normal, Mean=100
StDev=5
StDev=15
Density
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
50
75
100
125
150
X

- **Range**
- Let's start with the range because it is the most straightforward measure of variability to calculate and the simplest to understand.
- The range of a dataset is the difference between the largest and smallest values in that dataset.
- For example, in the two datasets below, dataset 1 has a range of 20 – 38 = 18 while dataset 2 has a range of 11 – 52 = 41.
- Dataset 2 has a broader range and, hence, more variability than dataset 1.

Dataset 1	Dataset 2
20	11
21	16
22	19
25	23
26	25
29	32
33	39
34	46
38	52

- **The Interquartile Range (IQR) . . . and other Percentiles**
- The interquartile range is the middle half of the data. To visualize it, think about the median value that splits the dataset in half.
- Similarly, you can divide the data into quarters. Statisticians refer to these quarters as quartiles and denote them from low to high as Q1, Q2, and Q3.
- The lowest quartile (Q1) contains the quarter of the dataset with the smallest values.
- The upper quartile (Q4) contains the quarter of the dataset with the highest values.
- The interquartile range is the middle half of the data that is in between the upper and lower quartiles.
- In other words, the interquartile range includes the 50% of data points that fall between Q1 and Q3. The IQR is the red area in the graph below.


Interquartile Range
0.5
Q1
25%
Q4
25%

1	2	3	4
Data Point	Mean	Difference	Squared Difference
11	32	-21	441
16	32	-16	256
19	32	-13	169
20	32	-12	144
21	32	-11	121
22	32	-10	100
25	32	-7	49
26	32	-6	36
29	32	-3	9
33	32	1	1
34	32	2	4
38	32	6	36
39	32	7	49
46	32	14	196
52	32	20	400
55	32	23	529
58	32	26	676

Sum	3216
Divide by 16	201
	Variance

In the variance section, we calculated a variance of 201 in the table.

$$\sqrt{201} = 14.177$$

Therefore, the standard deviation for that dataset is 14.177.

Graphic Displays of Basic Statistical Descriptions

- ***Boxplot:*** *graphic display of five-number summary*
- ***Histogram****: x-axis are values, y-axis repres. frequencies*
- ***Quantile plot****: each value x_i is paired with f_i indicating that approximately 100 f_i% of data are $\leq x_i$*
- ***Quantile-quantile (q-q) plot****: graphs the quantiles of one univariant distribution against the corresponding quantiles of another*
- ***Scatter plot****: each pair of values is a pair of coordinates and plotted as points in the plane*

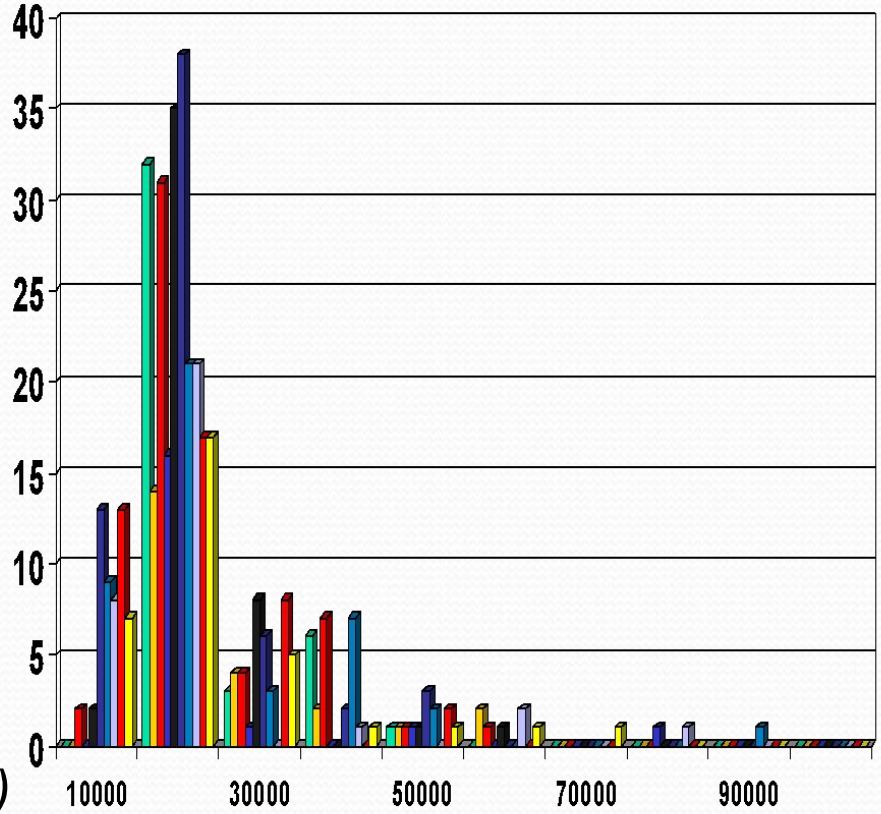
Histogram Analysis

Histogram: *Graph display of tabulated frequencies, shown as bars*

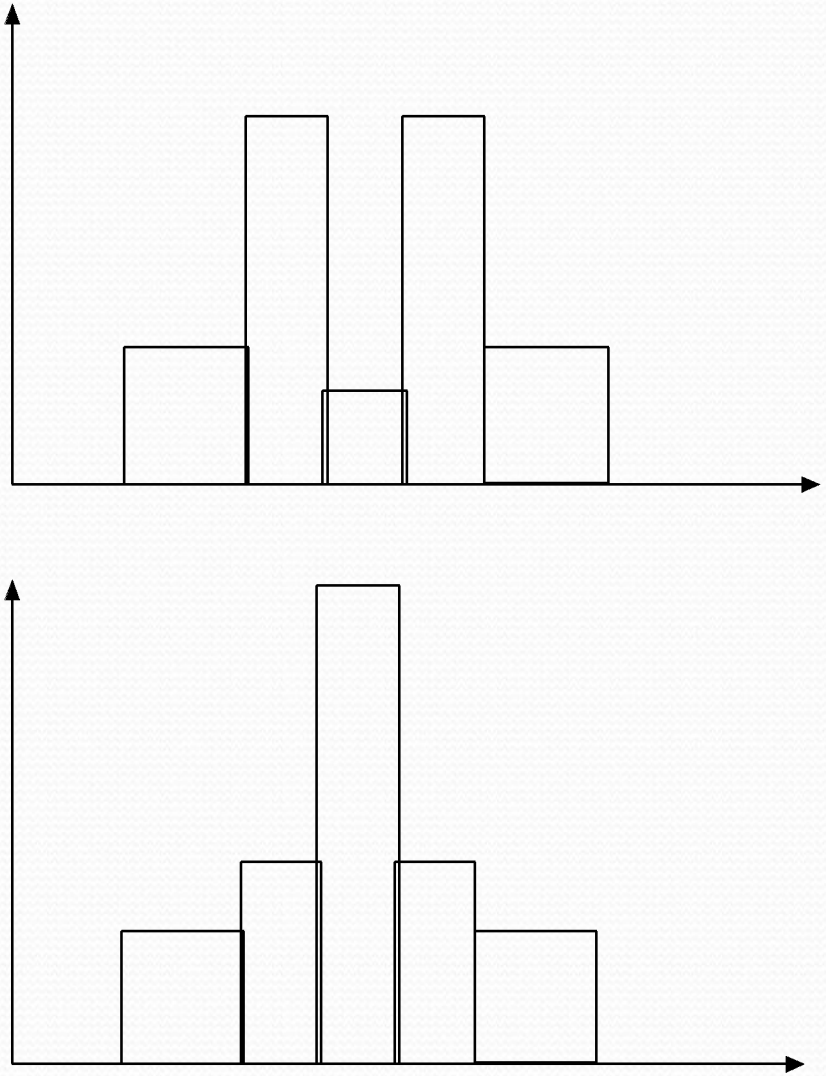
It shows what proportion of cases fall into each of several categories

Differs from a bar chart in that it is the area of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width

The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



Histograms Often Tell More than Boxplots



The two histograms shown in the left may have the same boxplot representation

- The same values for: min, Q1, median, Q3, max

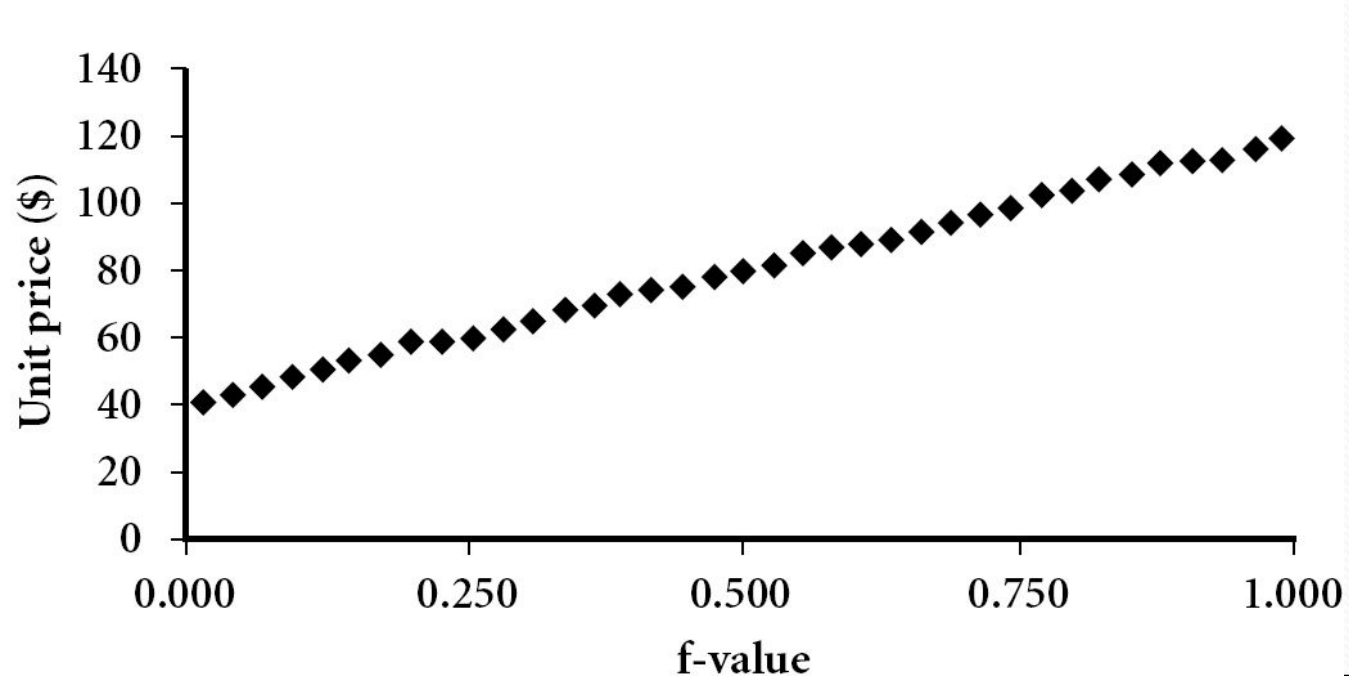
But they have rather different data distributions

Quantile Plot

Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)

*Plots **quantile** information*

For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



- The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not.
- Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution.
- They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

Quantiles And Percentiles

Quantiles are points in a dataset that divide the data into intervals containing equal probabilities or proportions of the total distribution. They are often used to describe the spread or distribution of a dataset. The most common quantiles are:

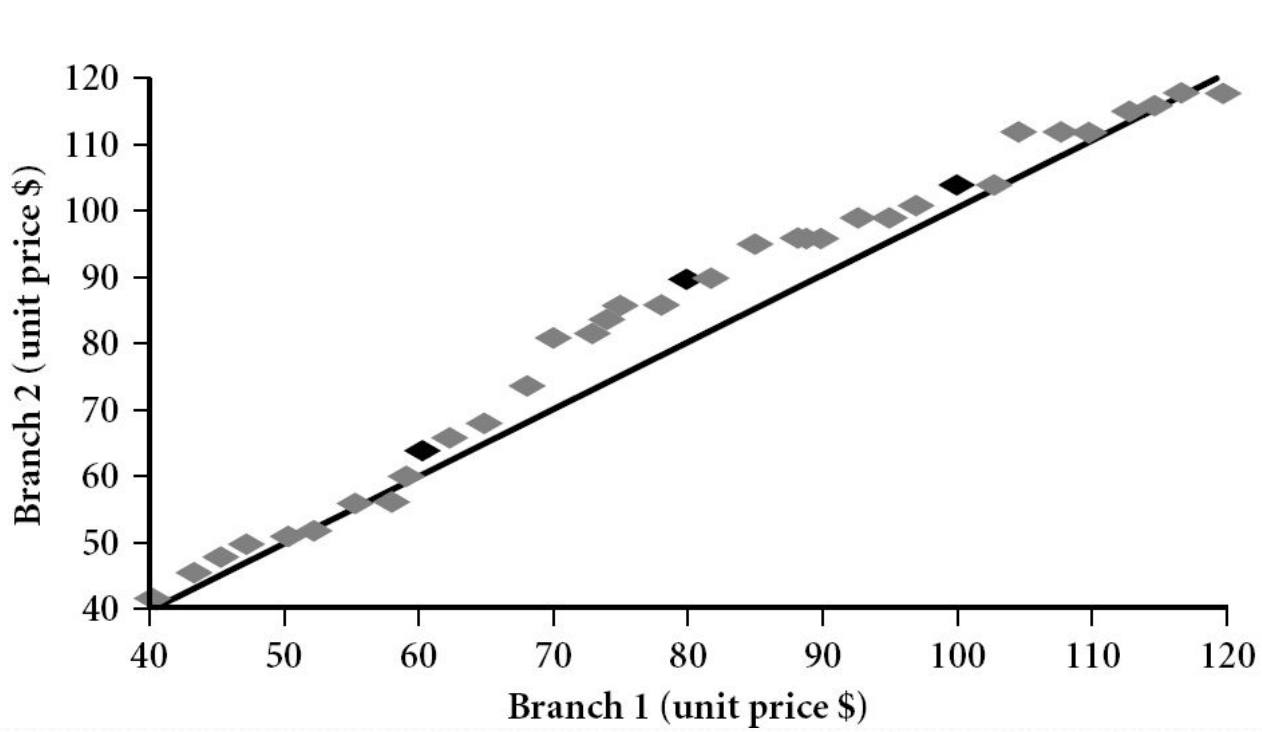
1. Median **(50th percentile)**: The median is the middle value of a dataset when it is ordered from smallest to largest. It divides the dataset into two equal halves.
2. Quartiles **(25th, 50th, and 75th percentiles)**: Quartiles divide the dataset into four equal parts. The first quartile (Q1) is the value below which 25% of the data falls, the second quartile (Q2) is the median, and the third quartile (Q3) is the value below which 75% of the data falls.
3. Percentiles: Percentiles are similar to quartiles but divide the dataset into 100 equal parts. For example, the 90th percentile is the value below which 90% of the data falls.

Quantile-Quantile (Q-Q) Plot

Graphs the quantiles of one univariate distribution against the corresponding quantiles of another

View: Is there is a shift in going from one distribution to another?

Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.

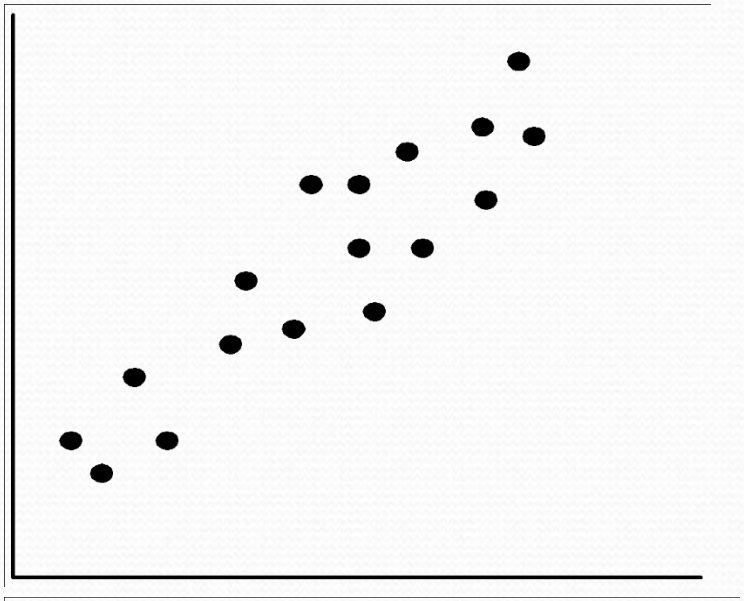


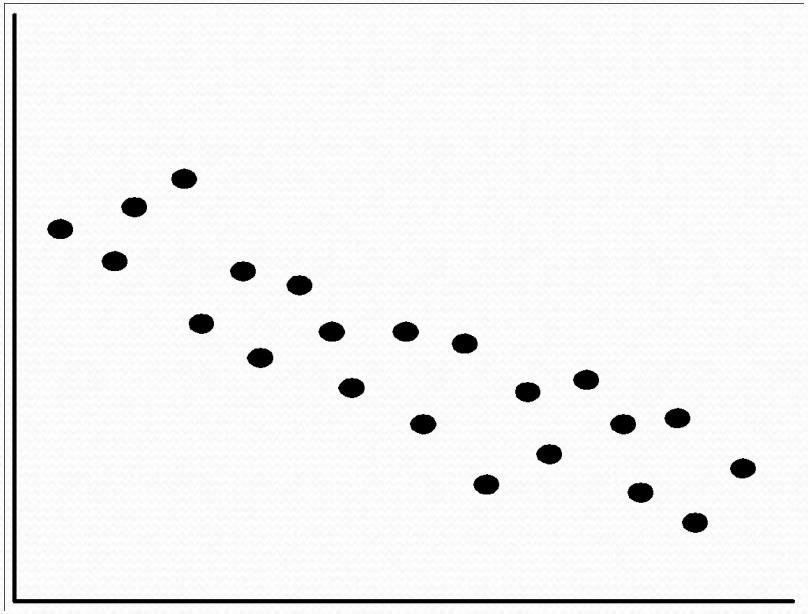
Scatter plot

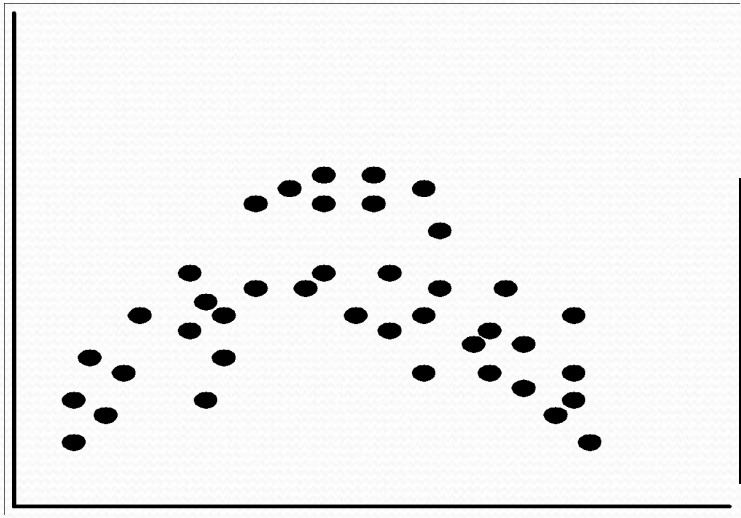
Provides a first look at bivariate data to see clusters of points, outliers, etc

Each pair of values is treated as a pair of coordinates and plotted as points in the plane

Positively and Negatively Correlated Data



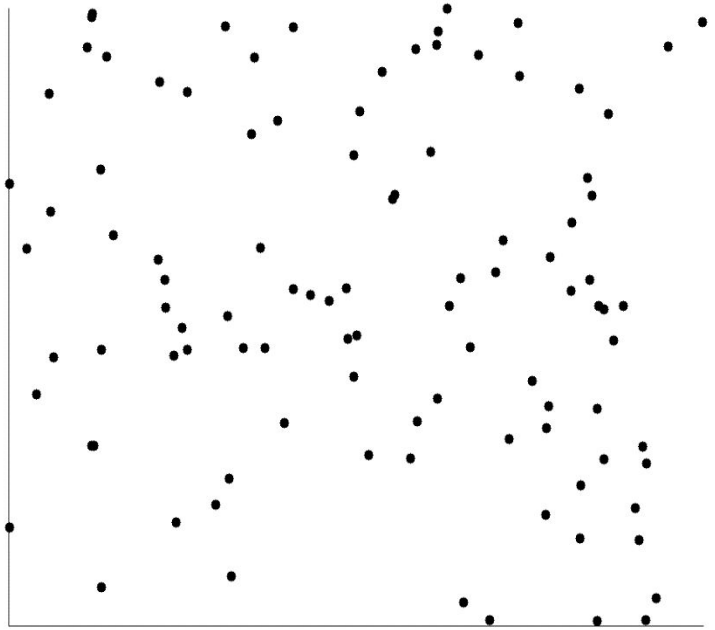


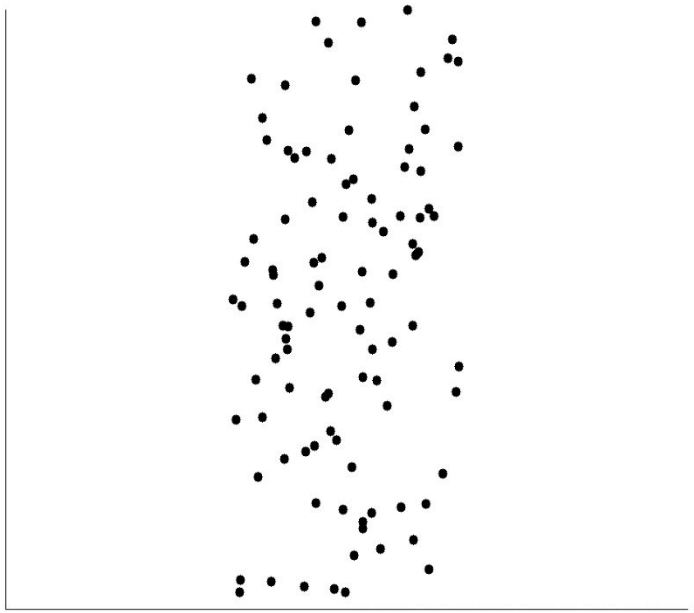


The left half fragment is positively correlated

The right half is negative correlated

Uncorrelated Data





- **Quantile Quantile plots**
- The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not.
- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
- By a quantile, we mean the fraction (or percent) of points below the given value.
- For the reference purpose, a 45% line is also plotted, if the samples are from the same population then the points are along this line.
- Normal Distribution:
- The normal distribution (aka Gaussian Distribution/ Bell curve) is a continuous probability distribution representing distribution obtained from the randomly generated real values.

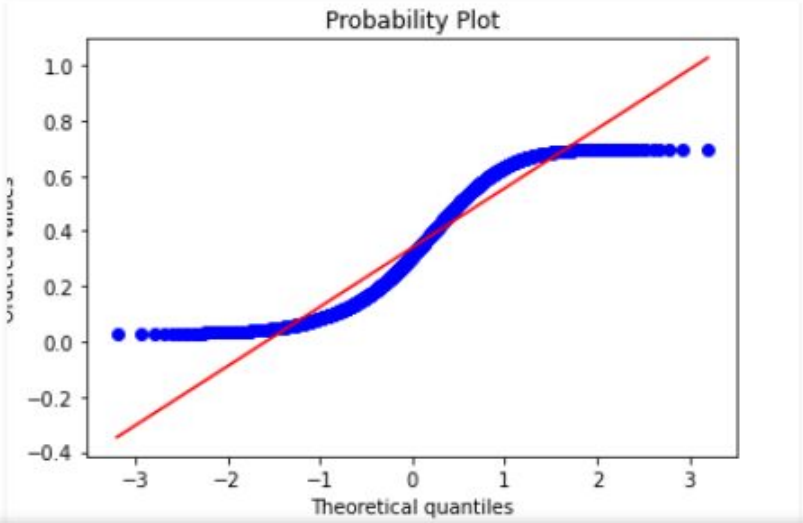
- **Usage:**
- The Quantile-Quantile plot is used for the following purpose:
- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behavior.

- **How to Draw Q-Q plot**
- Collect the data for plotting the quantile-quantile plot.
- Sort the data in ascending or descending order.
- Draw a normal distribution curve.
- Find the z-value (cut-off point) for each segment.
- Plot the dataset values against the normalizing cut-off points.

- ***Advantages of Q-Q plot***
- ***Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.***
- ***Since we need to normalize the dataset, so we don't need to care about the dimensions of values.***

Types of Q-Q plots

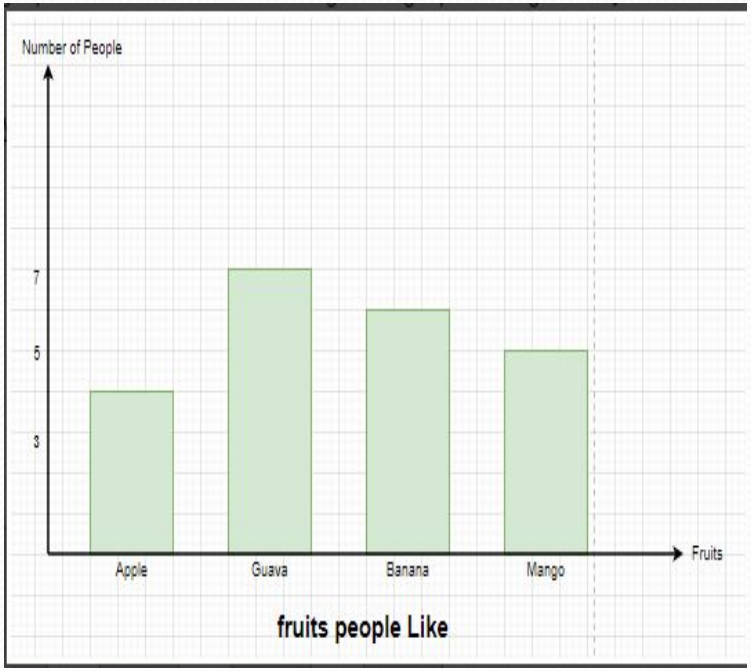
- For Left-tailed distribution: Below is the



Bar Graph:

- A bar graph is the graphical representation of categorical data using rectangular bars where the length of each bar is proportional to the value they represent.
- A histogram is the graphical representation of data where data is grouped into continuous number ranges and each range corresponds to a vertical bar.

- ***Bar Graph***
- The pictorial representation of data in groups, either in horizontal or vertical bars where the length of the bar represents the value of the data present on axis.
- They (bar graphs) are usually used to display or impart the information belonging to ‘categorical data’ i.e; data that fit in some category.
- Reading a Bar Graph and comparing two sets of data
- In order to read a Bar graph, we need to ask questions to ourselves looking at the displayed graph.
- Let’s understand reading a Bar graph through a very basic example,



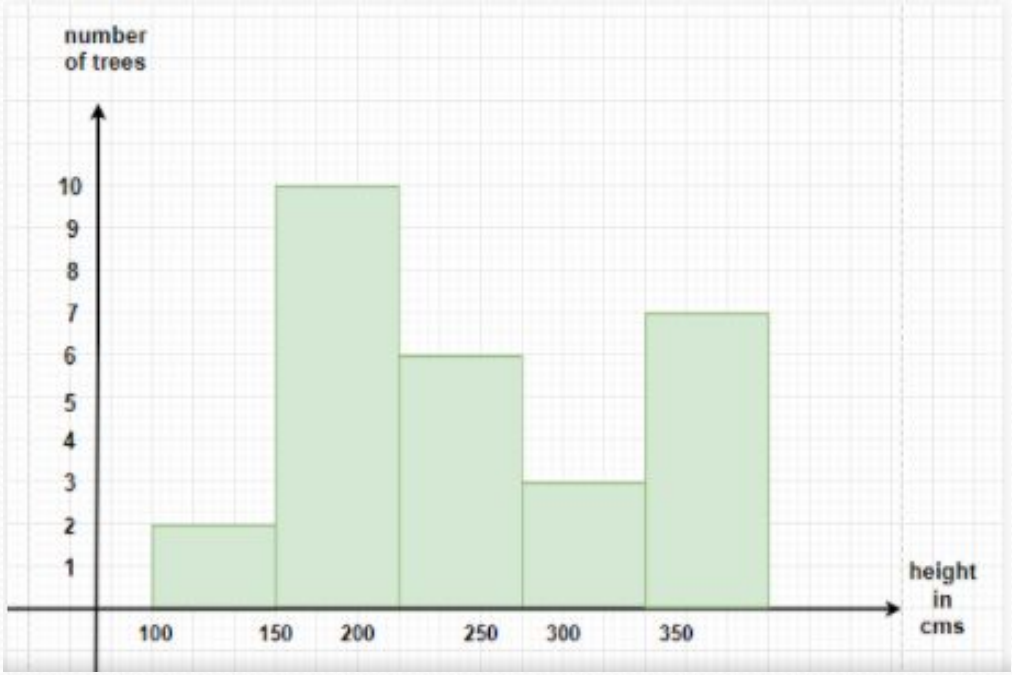
- **Properties of Bar Graph**
- All Bars have a common base.
- The length of each bar corresponds to its respective data mentioned on the axis (Y-axis for Vertical Graph, X-axis for Horizontal Graph).
- Each bar displayed has the same width.
- The distance between consecutive bars is the same
- **Significance of a Bar Graph**
- It is always easier and more comfortable to visually understand something than to look at the large table of Numerical data.
- Bar graphs are extensively used in presentations and reports.
- It is very prominently used as it summarizes data and displays it in a frequency distribution.

- **Histograms**
- Histograms are the graphical display of data with the help of bars, the heights of the bars may vary due to different data.
- The similarity of a Histogram resembles a bar chart, but histogram groups numbers into different ranges.
- The length of each bar tells how many fall into each range, and it can be decided which ranges are to be used.
- **Construction and interpretation of Histogram**
- First, mark the class intervals on X-axis and frequencies on Y-axis.
- Make sure that the scale of both axes is the same.
- The class Intervals shall always be exclusive.
- Create bars with class intervals on the x-axis and corresponding frequencies on the y-axis.
- The length of each bar reflects the Frequency when intervals are equal.
- The area of each bar is the same as its respective frequency when intervals are unequal.

- ***Difference Between Histograms and Bar graph***
- Bar graph is a one-dimensional figure while Histogram is a two-dimensional figure.
- In Bar graphs, the length of the bars shows the frequency, but the width has no special significance but in Histograms, the frequency is shown by the area of the bar.
- In bar graphs, the bars are separated from each other with equal spaces, while in Histograms, the bars are always touching each other.

Question 1: In a Park, there are 28 trees of different heights, the heights can be measured in centimeters and the range of the trees lie between 100-350 cms. Draw the Histogram for the following data,

Range of Height of Tree (in cm)	100-150	150-200	200-250	250-300	300-350
Number of Trees	2	10	6	3	7



- **What is a Scatter Plot?**
- A scatter plot is a chart type that is normally used to observe and visually display the relationship between variables.
- The values of the variables are represented by dots.
- The positioning of the dots on the vertical and horizontal axis will inform the value of the respective data point.
- Hence, scatter plots make use of Cartesian coordinates to display the values of the variables in a data set.
- Scatter plots are also known as scattergrams, scatter graphs, or scatter charts.

- A scatter plot is a chart type that is normally used to observe and visually display the relationship between variables.
- It is also known as a scattergram, scatter graph, or scatter chart.
- The data points or dots, which appear on a scatter plot, represent the individual values of each of the data points and also allow pattern identification when looking at the data holistically.
- The most common use of the scatter plot is to display the relationship between two variables and observe the nature of such a relationship.
- The relationships observed can either be positive or negative, non-linear or linear, and/or, strong or weak.
-

- **Scatter Plot Applications and Uses:**

1. Demonstration of the relationship between two variables

- The most common use of the scatter plot is to display the relationship between two variables and observe the nature of the relationship.
- The relationships observed can either be positive or negative, non-linear or linear, and/or, strong or weak.

2. Identification of correlational relationships

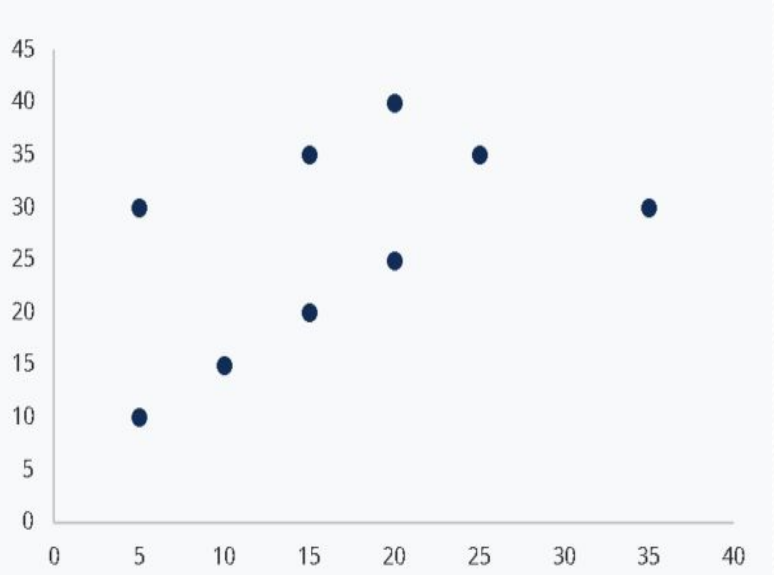
- Another common use of scatter plots is that they enable the identification of correlational relationships.
- Scatter plots tend to have independent variables on the horizontal axis and dependent variables on the vertical axis.
- It allows the observer to know or get an idea of what the possible vertical value may be, provided there is information on the horizontal value.

3. Identification of data patterns

- Data pattern identification is also possible with scatter plots. Data points can be grouped together based on how close their values are, and this also makes it easy to identify any outlier points when there are data gaps.

- **Creating a Scatter Plot Diagram**

Month	Series 1	Series 2
Sep	30	5
Aug	35	15
Jul	40	20
Jun	35	25
May	30	35
Apr	25	20
Mar	20	15
Feb	15	10
Jan	10	5



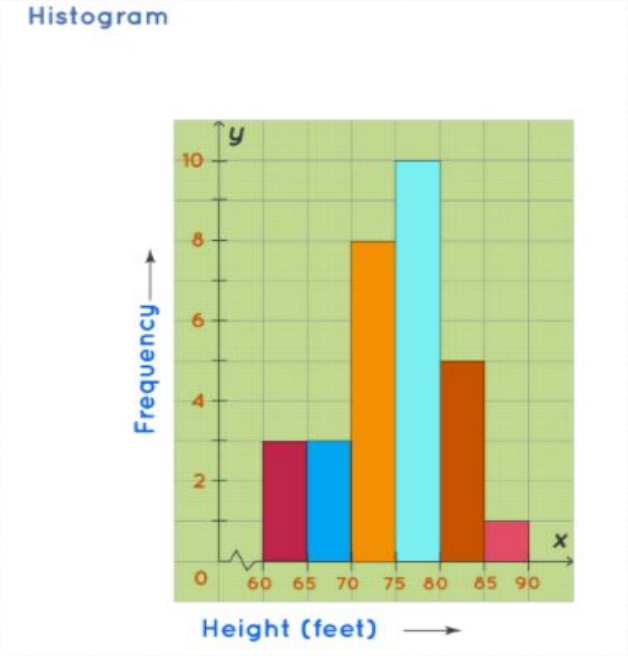
- A frequency plot is **a graphical data analysis technique for summarizing the distributional information of a variable**.
- The response variable is divided into equal sized intervals (or bins).
- The number of occurrences of the response variable is calculated for each bin

- ***Histogram***
- A histogram can be defined as a set of rectangles with bases along with the intervals between class boundaries.
- Each rectangle bar depicts some sort of data and all the rectangles are adjacent.
- The heights of rectangles are proportional to corresponding frequencies of similar as well as for different classes. Let's learn about histograms more in detail.
- ***What is Histogram?***
- A histogram is the graphical representation of data where data is grouped into continuous number ranges and each range corresponds to a vertical bar.
- The horizontal axis displays the number range.
- The vertical axis (frequency) represents the amount of data that is present in each range.
- The number ranges depend upon the data that is being used.

- ***Histogram Graph***
- A histogram graph is a bar graph representation of data.
- It is a representation of a range of outcomes into columns formation along the x-axis. in the same histogram, the number count or multiple occurrences in the data for each column is represented by the y-axis.
- It is the easiest manner that can be used to visualize data distributions.
- Let us understand the histogram graph by plotting one for the given below example.

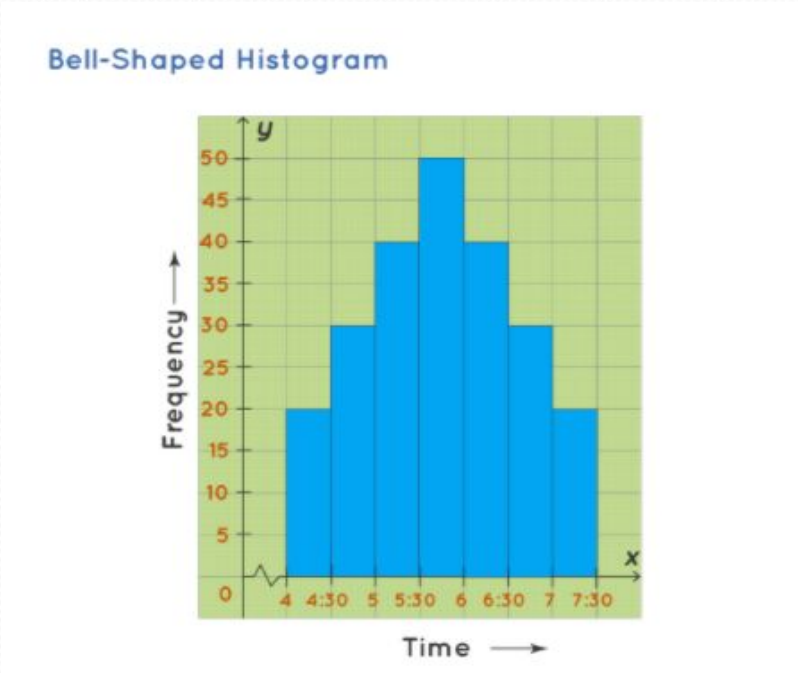
- ***Each tree is of a different height. The height of the trees (in inches): 61, 63, 64, 66, 68, 69, 71, 71.5, 72, 72.5, 73, 73.5, 74, 74.5, 76, 76.2, 76.5, 77, 77.5, 78, 78.5, 79, 79.2, 80, 81, 82, 83, 84, 85, 87.***
- ***We can group the data as follows in a frequency distribution table by setting a range:***

Height Range (ft)	Number of Trees (Frequency)
60 - 75	3
66 - 70	3
71 - 75	8
76 - 80	10
81 - 85	5
86 - 90	1

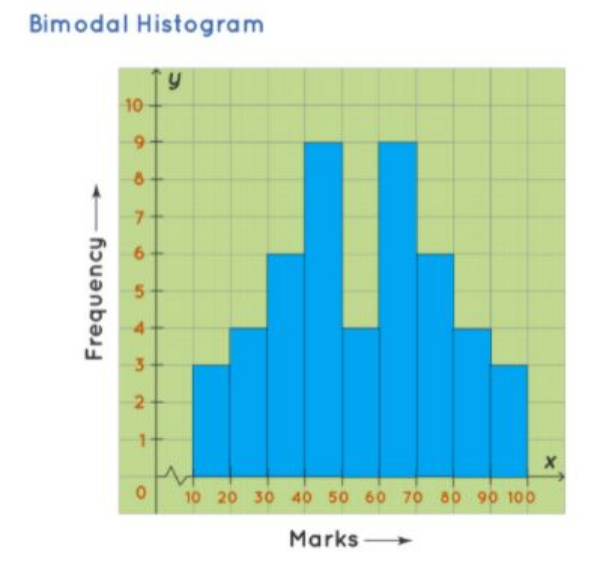


- ***Histogram Shapes***
- The histogram can be classified into different types based on the frequency distribution of the data.
- There are different types of distributions, such as normal distribution, skewed distribution, bimodal distribution, multimodal distribution, comb distribution, edge peak distribution, dog food distribution, heart cut distribution, and so on.
- The histogram can be used to represent these different types of distributions.
- We have mainly 5 types of histogram shapes.
- They are listed below:
- Bell Shaped Histogram
- Bimodal Histogram
- Skewed Right Histogram
- Skewed Left Histogram
- Uniform Histogram

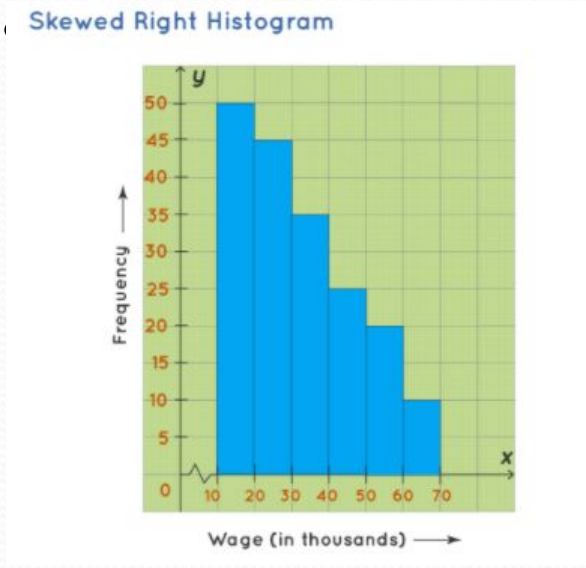
- **Bell-Shaped Histogram**
- A bell-shaped histogram has a single peak. The histogram has just one peak at this time interval and hence it is a **bell-shaped histogram**.
- For example**,** the following histogram shows the number of children visiting a park at different time intervals. This histogram has only one peak. The maximum number of children who visit the park is between 5.30 PM to 6 PM.
-



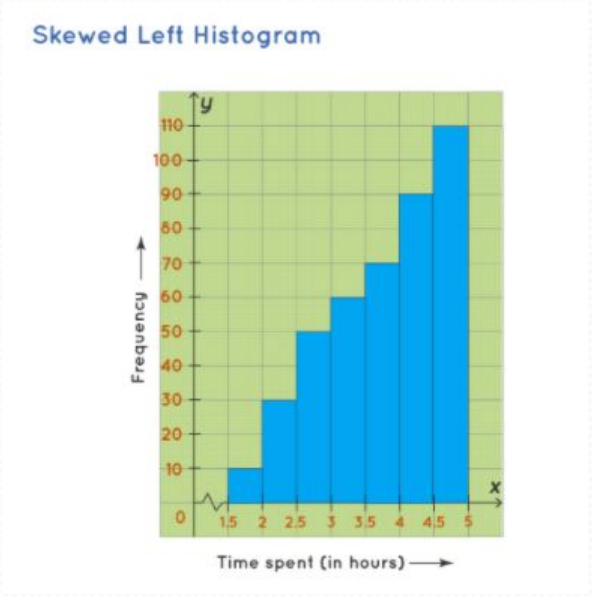
- **Bimodal Histogram**
- A bimodal histogram has two peaks and it looks like the graph given below.
- For example, the following histogram shows the marks obtained by the 48 students.
- The maximum number of students have scored either between 40 to 50 marks OR between 60 to 70 marks. This histogram has two peaks (between 40 to 50 and between 60 to 70) and hence it is a **bimodal histogram**.



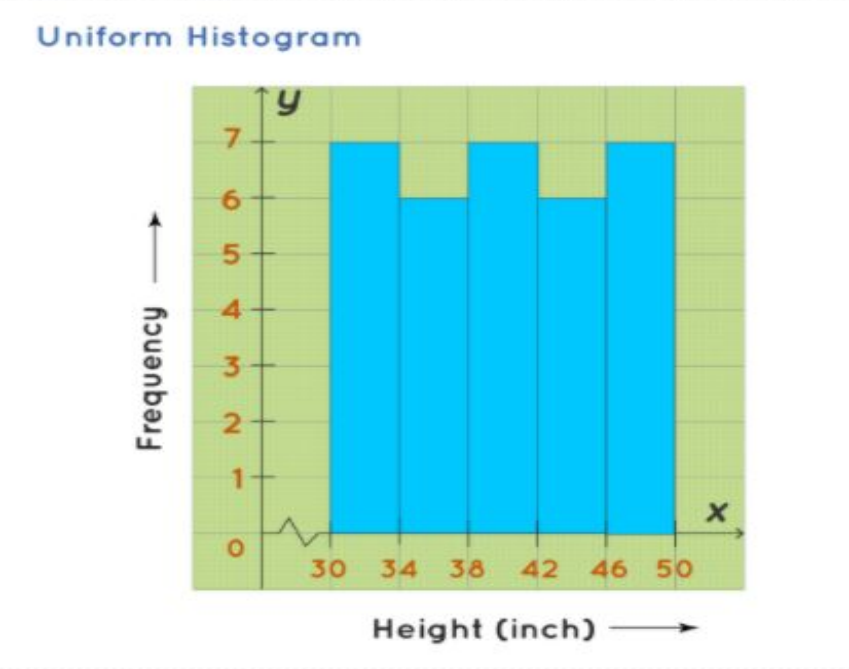
- **Skewed Right Histogram**



- **Skewed Left Histogram**

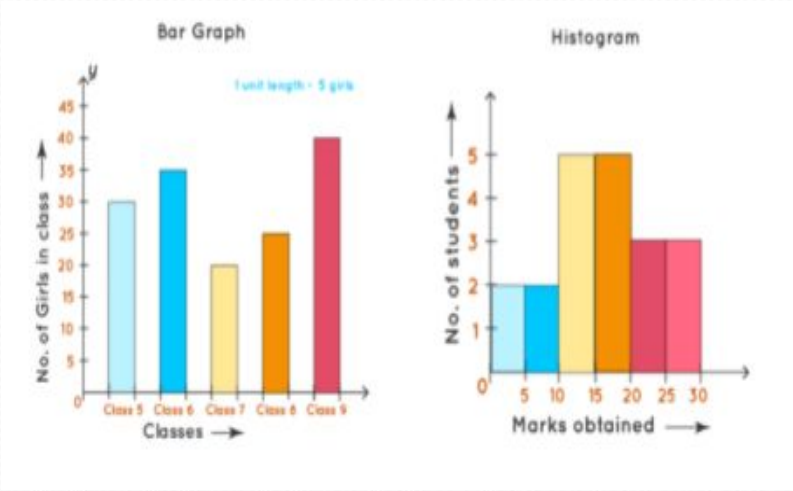


- **Uniform Histogram**
- A uniform histogram is a histogram where all the bars are more or less of the same height.
- In this histogram, the lengths of all the bars are more or less the same.
- Hence, it is a uniform histogram.



- **Difference Between a Bar Chart and a Histogram**
- The fundamental difference between histograms and bar graphs from a visual aspect is that bars in a bar graph are not adjacent to each other.
- A bar graph is the graphical representation of categorical data using rectangular bars where the length of each bar is proportional to the value they represent.
- A **histogram** is the graphical representation of data where data is grouped into continuous number ranges and each range corresponds to a vertical bar.
- The main differences between a bar chart and a histogram are as follows:

Bar Graph	Histogram
Equal space between every two consecutive bars.	No space between two consecutive bars. They should be attached to each other.
X-axis can represent anything.	X-axis should represent only continuous data that is in terms of numbers.



	Bar graph	Histogram
Data type	Compares discrete or categorical data	Visualizes continuous data
Bars	Has gaps between bars	Has no gaps between bars
Purpose	Compares different groups of data	Shows distribution of data
Example	Sales by region, survey results	Heights of individuals

Similarity and Dissimilarity

Similarity

- ***Numerical measure of how alike two data objects are***
- ***Value is higher when objects are more alike***
- ***Often falls in the range [0,1]***

Dissimilarity *(e.g., distance)*

- ***Numerical measure of how different two data objects are***
- ***Lower when objects are more alike***
- ***Minimum dissimilarity is often 0***
- ***Upper limit varies***

Proximity *refers to a similarity or dissimilarity*

Data Matrix and Dissimilarity Matrix

Data matrix

- *n data points with p dimensions*
- *Two modes*

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Dissimilarity matrix

- *n data points, but registers only the distance*
- *A triangular matrix*
- *Single mode*

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

- d(i,j) is the measured dissimilarity or the between object i and j.
- d(i,j) is a non-negative number that is close to 0 when objects i and j are highly similar or “near” each other, and becomes larger the more they differ.
- Measure of dissimilarity:
- Sim(i,j)=1-d(i,j)

Proximity measures are different for different types of attributes.

Similarity measure:

–Numerical measure of how alike two data objects are.

– Is higher when objects are more alike.

– Often falls in the range [0,1].

Dissimilarity measure:

–Numerical measure of how different two data objects are.

– Lower when objects are more alike.

– Minimum dissimilarity is often 0.

– Upper limit varies.

Dissimilarity Matrix:

Dissimilarity matrix is a matrix of pairwise dissimilarity among the data points. It is often desirable to keep only lower triangle or upper triangle of a dissimilarity matrix to reduce the space and time complexity.

1. It's square and symmetric(A^T = A for a square matrix A, where A^T represents its transpose).

2. The diagonals members are zero, meaning that zero is the measure of dissimilarity between an element and itself.

Proximity Measure for Nominal Attributes

Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)

Method 1: Simple matching

m: # of matches, p: total # of variables

$$d(i,j) = \frac{p - m}{p}$$

Method 2: Use a large number of binary attributes

creating a new binary attribute for each of the M nominal states

- **Proximity measures are mainly mathematical techniques that calculate the similarity/dissimilarity of data points. Usually, proximity is measured in terms of similarity or dissimilarity i.e., how alike objects are to one another.**

Proximity Measure for Nominal Attributes

object identifier	*test-1 (nominal)*	*test-2 (ordinal)*	*test-3 (numeric)*
1	code-A	excellent	45
2	code-B	fair	22
3	code-C	good	64
4	code-A	excellent	28

Take only test-1, which is nominal. p = 1

Dissimilarity matrix

$$= \begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

$$sim(i,j) = 1 - d(i,j) = \frac{m}{p}.$$

- ***m:- Number of matches.***
- ***P:-total number of attributes.***

- ***Object Identifier*** ***test-1*** ***test-2*** ***test-3***

	Nominal	Ordinal	Numeri
1	Code A	Excellent	45
2	Code B	Fair	22
3	Code C	good	64
4	Code A	Excellent	28

Compute the dissimilarity between objects of mixed attribute types given in Table 1.

Table 1: A sample data table containing attributes of mixed type.

Object identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code-A	excellent	45
2	code-B	fair	22
3	code-C	good	64
4	code-A	excellent	28

Proximity measures for ordinal attributes

An ordinal attribute is an attribute whose possible values have a meaningful order or ranking among them, but the magnitude between successive values is not known.

However, to do so, it is important to convert the states to numbers where each state of an ordinal attribute is assigned a number corresponding to the order of attribute values.

Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}.

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

where M is a maximum number assigned to states and r is the rank(numeric value) of a particular object.

The similarity can be calculated as:

$$s(i, j) = 1 - d(i, j)$$

Object ID	Attribute
1	High
2	Low
3	Medium
4	High

In this example, we have four objects having ID from 1 to 4.

Here for encoding our attribute column, we consider *High=1, Medium=2, and Low=3*. And, the value of Mf=3(since there are three states available)

Now, we normalize the ranking in the range of 0 to 1 using the above formula.

So, *High=(1-1)/(3-1)=0, Medium=(2-1)/(3-1)=0.5, Low=(3-1)/(3-1)=1.*

Finally,

so, *High=(1-1)/(3-1)=0, Medium=(2-1)/(3-1)=0.5, Low=(3-1)/(3-1)=1.*

Finally, we are able to calculate the dissimilarity based on difference in normalized values corresponding to that attribute.

– d(1,1)= 0-0 = 0

– d(2,1)= 1-0= 1

– d(3,1)= 0.5-0 = 0.5

– d(4,1)= 0-0 =0

– d(4,3)= 0.5-0=0

– d(2,2)= 3-3 = 0

– d(3,2)= 0.5-0 = 0.5

– d(4,2)= 1-0 = 1

– d(3,3)= 0.5-0.5 = 0

– d(4,4)= 0-0 = 0

T

Proximity Measures for Ordinal Variables

- Consider the data in the adjacent table:
- Here, the attribute Test has three states: fair, good and excellent, so M_f=3
- For step 1, the four attribute values are assigned the ranks 3,1,2 and 3 respectively.
- Step 2 normalizes the ranking by mapping rank 1 to 0.0, rank 2 to 0.5 and rank 3 to 1.0
- For step 3, using Euclidean distance, a dissimilarity matrix is obtained as shown
- Therefore, students 1 and 2 are most dissimilar, as are students 2 and 4

Student	Test
1	Excellent
2	Fair
3	Good
4	Excellent

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Proximity Measure for Binary Attributes

A contingency table for binary data

Object *i* \ Object *j*	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

Distance measure for symmetric binary variables:

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

Distance measure for asymmetric binary variables:

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

Jaccard coefficient (similarity measure for asymmetric binary variables):

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

Note: Jaccard coefficient is the same as "coherence":

$$coherence(i,j) = \frac{sup(i,j)}{sup(i)+sup(j)-sup(i,j)} = \frac{q}{(q+r)+(q+s)-q}$$

- q:- number of attributes that are equal to 1 for both objects I and j.
- r:- number of attributes that are equal to 1 for object I and 0 for object J.
- s:- number of attributes that are equal to 0 for object I and 1 for object J
- t:- number of attributes that are equal to 0 for both objects.
- The total number of attributes is p , where p=q+r+s+t

Dissimilarity between Binary Variables

Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

Gender is a symmetric attribute

The remaining attributes are asymmetric binary

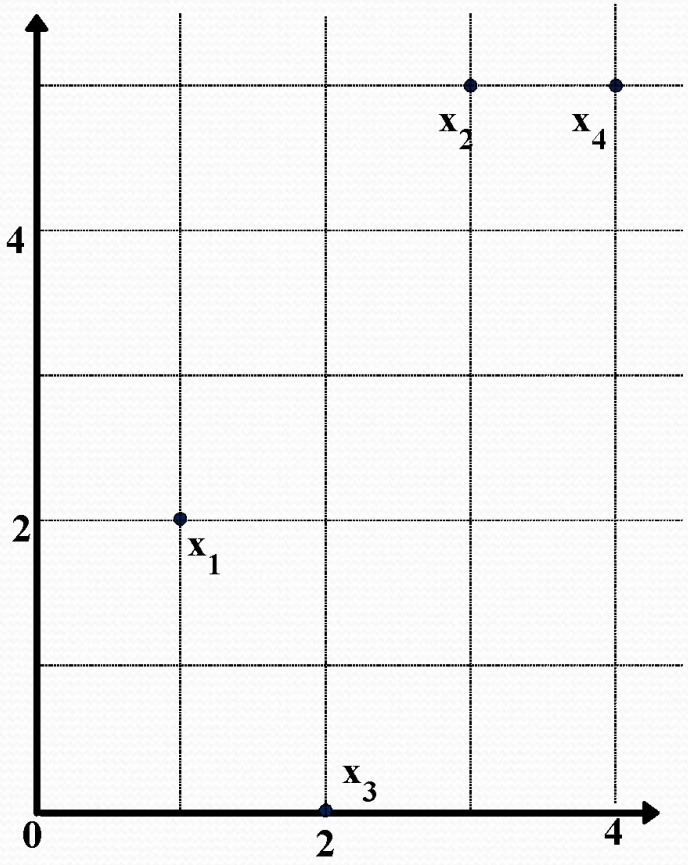
Let the values Y and P be 1, and the value N 0

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Example:
Data Matrix and Dissimilarity Matrix



$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2)}$$

Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrix

(with Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	5.1	5.1	0	
x4	4.24	1	5.39	0

- ***Euclidean distance:***
- ***Let i=(xi1, xi2, xi3, xi4,… xip)***
- ***j=(xj1, xj2, xj3, xj4,… xjp)***
- ***d(i,j)=sqrt((xi1- xj1)^2 + (xi2- xj2)^2 + (xi3- xj3)^2….. (xip- xjp)^2)***

Distance on Numeric Data: Minkowski Distance

Minkowski distance: A popular distance measure

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

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and h is the order (the distance so defined is also called L-h norm)

Properties

$d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positive definiteness)

$d(i, j) = d(j, i)$ (Symmetry)

$d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality)

A distance that satisfies these properties is a metric

Special Cases of Minkowski Distance

h = 1: Manhattan (city block, L_1 norm) distance

E.g., the Hamming distance: the number of bits that are different between two binary vectors

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

h = 2: (L_2 norm) Euclidean distance

$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2)}$$

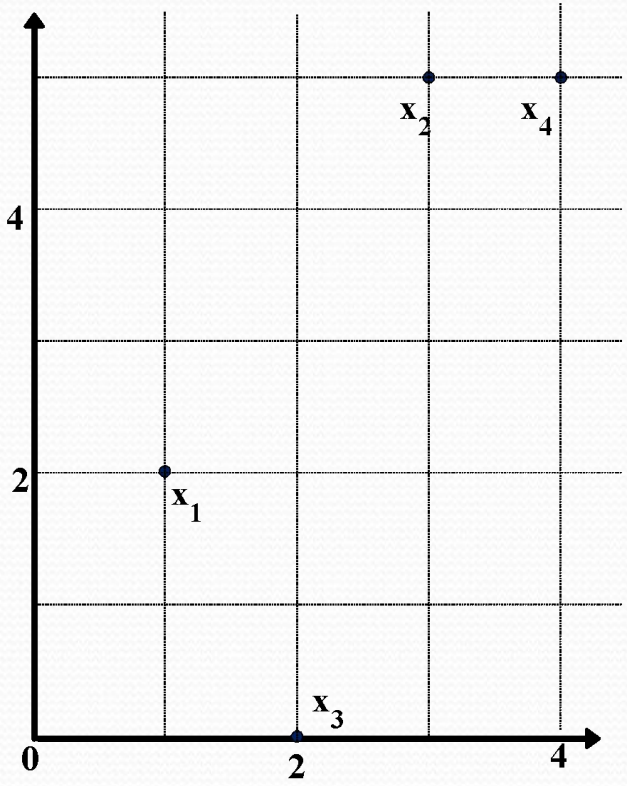
h → ∞. "supremum" (L_{max} norm, L_∞ norm) distance.

This is the maximum difference between any component (attribute) of the vectors

$$d(i,j)=\lim_{h\to\infty}\left(\sum_{f=1}^{p}|x_{if}-x_{jf}|^h\right)^{\frac{1}{h}}=\max_{f}^{p}|x_{if}-x_{jf}|$$

Example: Minkowski Distance

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Dissimilarity Matrices

Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Ordinal Variables

An ordinal variable can be discrete or continuous

Order is important, e.g., rank

Can be treated like interval-scaled

- ***replace x_{if} by their rank*** $r_{if} \in \{1, \ldots, M_f\}$
- ***map the range of each variable onto [0, 1] by replacing i-th object in the f-th variable by***

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- ***compute the dissimilarity using methods for interval-scaled variables***

Attributes of Mixed Type

A database may contain all attribute types

Nominal, symmetric binary, asymmetric binary, numeric, ordinal

One may use a weighted formula to combine their effects

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

f is binary or nominal:

$d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise

f is numeric: use the normalized distance

f is ordinal

Compute ranks r_{if} and

Treat z_{if} as interval-scaled

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Cosine Similarity

*A **document** can be represented by thousands of attributes, each recording the frequency of a particular word (such as keywords) or phrase in the document.*

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

Other vector objects: gene features in micro-arrays, …

Applications: information retrieval, biologic taxonomy, gene feature mapping, ...

Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

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

where • indicates vector dot product, ||d||: the length of vector d

Example: Cosine Similarity

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

where • indicates vector dot product, ||d|: the length of vector d

*Ex: Find the **similarity** between documents 1 and 2.*

$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$

$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$

$d_1 \cdot d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$

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

$||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12$

$cos(d_1, d_2) = 0.94$

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

It is often used to measure document **similarity in text analysis**.

Cosine similarity between two term-frequency vectors

Suppose that $\boldsymbol{x}$ and $\boldsymbol{y}$ are the first two term-frequency vectors in Table 2.5. That is, $\boldsymbol{x} = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$ and $\boldsymbol{y} = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$. How similar are $\boldsymbol{x}$ and $\boldsymbol{y}$? Using Eq. (2.23) to compute the cosine similarity between the two vectors, we get:

$$\begin{aligned} \boldsymbol{x}^t \cdot \boldsymbol{y} &= 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 \\ &\quad + 0 \times 0 + 0 \times 1 = 25 \\ ||\boldsymbol{x}|| &= \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2} \\ ||\boldsymbol{y}|| &= \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2} \\ sim(\boldsymbol{x}, \boldsymbol{y}) &= 0.94 \end{aligned}$$

```
Cos(x, y) = x . y / ||x|| * ||y||
```

where,

- **x . y** = product (dot) of the vectors 'x' and 'y'.
- **||x||** and **||y||** = length of the two vectors 'x' and 'y'.
- **||x|| * ||y||** = cross product of the two vectors 'x' and 'y'.