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:

Tables (Tabular form)

Graphs

Hierarchy

Play (k)

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

Objects

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

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.

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 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

How satisfied are you with our services?

- 1- Very Unsatisfied
- 2- Unsatisfied
- 3- Neural
- 4- Satisfied
- 5- Very Satisfied

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.

This type of data represents quantitative data with equal intervals between consecutive values.

Interval data has no absolute zero point, and therefore, ratios cannot be computed.

Examples of interval data include temperature, IQ scores, and time.

Interval data is used in data mining for clustering and prediction tasks.

- ***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).

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.
- This type of data is similar to interval data, but with an absolute zero point. In ratio data, it is possible to compute ratios of two values, and this makes it possible to make meaningful comparisons. Examples of ratio data include height, weight, and income.
- Ratio data is used in data mining for prediction and association rule mining tasks.

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

- ***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.
- For example, because weight is a ratio variable, a weight of 4 grams is twice as heavy as a weight of 2 grams.
-

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

- ***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

Discrete data represents **discrete variables**, which you can count in a finite amount of time. The key feature here is that these variables are countable instead of measurable.

- **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.***
- 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.
- 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.

Measuring the Central Tendency

Mean (algebraic measure) (sample vs. population):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

Note: n is sample size and N is population size.

Weighted arithmetic mean:

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

Trimmed mean: chopping extreme values

Median:

Middle value if odd number of values, or average of the middle two values otherwise

Estimated by interpolation (for grouped data):

$$median = L_1 + \left(\frac{n/2 - (\sum freq)l}{freq_{median}}\right) width$$

Mode

Value that occurs most frequently in the data

Unimodal, bimodal, trimodal

- ***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
–	$\sum f_i = 40$	$\sum x_i\, f_i = 360$

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

$$Therefore\ Mean = 9$$

- ***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.**

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.

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

- ***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.

- ***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

- ***Finding the Median of Grouped Data***
- ***The marks obtained in English test by 17 students were recorded. What is the median marks of the students?***

Marks out of 50	Frequency	Cumulative frequency
0-10	2	2
10-20	4	6
20-30	5	11
30-40	4	15
40-50	2	17

$$\text{Median} = l + \frac{\frac{N}{2} - F}{f} \times h$$

where l = lower limit of the median class

f = frequency of the median class

F = cumulative frequency of the class preceding the median class

N = total number of observations

h = width of the median class

From the table above, we have $l = 20,\ N = 17,\ F = 6,\ f = 5$ and $h = 10$.

$$\begin{aligned} \therefore \quad \text{Median} &= 20 + \frac{\frac{17}{2} - 6}{5} \times 10 \\ &= 20 + \frac{5}{10} \times 10 = 20 + 5 \\ &= 25 \end{aligned}$$

Therefore, the median marks is 25.

Examples

1. Calculate Mean, Median, Mode from the following grouped data

Class	Frequency
2 - 4	3
4 - 6	4
6 - 8	2
8 - 10	1

Solution:

Class (1)	Frequency (f) (2)	Mid value (x) (3)	$f \cdot x$ (4) = (2) × (3)	cf (6)
2-4	3	3	9	3
4-6	4	5	20	7
6-8	2	7	14	9
8-10	1	9	9	10
---	---	---	---	---
--	$n = 10$	--	$\sum f \cdot x = 52$	--

$$\text{Mean } \bar{x} = \frac{\sum fx}{n}$$

$$= \frac{52}{10}$$

$$= 5.2$$

2. Median $M = L + \dfrac{\frac{n}{2} - cf}{f} \cdot c$

To find Median Class

= value of $\left(\frac{n}{2}\right)^{th}$ observation

= value of $\left(\frac{10}{2}\right)^{th}$ observation

= value of 5th observation

From the column of cumulative frequency cf, we find that the 5th observation lies in the class 4 - 6.

$\therefore$ The median class is 4 - 6.

Now,

$\therefore L$ = lower boundary point of median class = 4

$\therefore n$ = Total frequency = 10

$\therefore cf$ = Cumulative frequency of the class preceding the median class = 3

$\therefore f$ = Frequency of the median class = 4

$\therefore c$ = class length of median class = 2

Median $M = L + \dfrac{\frac{n}{2} - cf}{f} \cdot c$

$= 4 + \dfrac{5 - 3}{4} \cdot 2$

$= 4 + \dfrac{2}{4} \cdot 2$

$= 4 + 1$

$= 5$

- ***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
- 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.

Examples

1. Calculate Mean, Median, Mode from the following grouped data

Class	Frequency
2 - 4	3
4 - 6	4
6 - 8	2
8 - 10	1

Solution:

Class (1)	Frequency (f) (2)	Mid value (x) (3)	$f \cdot x$ (4) = (2) × (3)	cf (6)
2-4	3	3	9	3
4-6	4	5	20	7
6-8	2	7	14	9
8-10	1	9	9	10
---	---	---	---	---
--	$n = 10$	--	$\sum f \cdot x = 52$	--

$$\text{Mean } \bar{x} = \frac{\sum fx}{n}$$

$$= \frac{52}{10}$$

$$= 5.2$$

To find Mode Class
Here, maximum frequency is 4.

$\therefore$ The mode class is 4 - 6.

$\therefore L$ = lower boundary point of mode class = 4

$\therefore f_1$ = frequency of the mode class = 4

$\therefore f_0$ = frequency of the preceding class = 3

$\therefore f_2$ = frequency of the succedding class = 2

$\therefore c$ = class length of mode class = 2

$$Z = L + \left(\frac{f_1 - f_0}{2 \cdot f_1 - f_0 - f_2}\right) \cdot c$$

$$= 4 + \left(\frac{4 - 3}{2 \cdot 4 - 3 - 2}\right) \cdot 2$$

$$= 4 + \left(\frac{1}{3}\right) \cdot 2$$

$$= 4 + 0.6667$$

$$= 4.6667$$

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***

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:

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

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

where μ is the mean value of the observations

standard deviation *σ (s)* is the square root of variance σ^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:

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

- In the case of a moderately skewed distribution, i.e. in general, the difference between mean and mode is equal to three times the difference between the mean and median. Thus, in this case, the empirical relationship is expressed as, Mean – Mode = 3 (Mean – Median).
- In the case of a frequency distribution that has a symmetrical frequency curve, the empirical relation states that mean = median = mode.
- In the case of a positively skewed frequency distribution curve, mean > median > mode.
- In the case of negatively skewed frequency distribution, mean < median < mode.

mean < median < mode
mean = median = mode
mean > median > mode
Frequency
Mode
Median
Mean
x
Mean
Median
Mode
x
Mode
Median
Mean
x
Negatively Skewed
Symmetrical Distribution
Positively Skewed

- ***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.

Measuring the Dispersion of Data

Quartiles, outliers and boxplots

Range, Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)

Inter-quartile range: IQR = $Q_3 - Q_1$

Five number summary: min, Q_1, median, Q_3, max

Boxplot: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually

Outlier: usually, a value higher/lower than 1.5 x IQR

Variance and standard deviation (sample: s, population: σ)

Variance: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

Standard deviation s (or σ) is the square root of variance s^2 (or σ^2)

- ***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***

- 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.

- ***Distance between first and third quartile is simple measure of spread of Data that gives the range covered by the middle half of the data.***
- ***This distance is called IQR.***
- ***IQR=Q3-Q1***

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

Boxplot Analysis

Five-number summary *of a distribution*

Minimum, Q1, Median, Q3, Maximum

Boxplot

Data is represented with a box

The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR

The median is marked by a line within the box

Whiskers: two lines outside the box extended to Minimum and Maximum

Outliers: points beyond a specified outlier threshold, plotted individually

- ***With boxplots, this is done using something called "fences". The idea is that anything outside the fences is a potential outlier and shouldn't be included in the main group that we graph.***

Visualization of Data Dispersion: 3-D Boxplots

Properties of Normal Distribution Curve

The normal (distribution) curve

From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)

From μ–2σ to μ+2σ: contains about 95% of it

From μ–3σ to μ+3σ: contains about 99.7% of it

Problems

Given the set:

10, 13, 4, 7, 8, 11 10, 16, 18, 12, 3, 6, 9, 9, 4, 13, 20, 7, 5, 10, 17, 10, 16, 14, 8, 18

Find their mean, mode, quartiles and variance.

Problems

A pediatrician has obtained the following table which represents the number of children who begin to walk for the first time at different ages:

Months	Children
9	1
10	4
11	9
12	16
13	11
14	8
15	1

1. Draw the respective frequency plot.

2. Calculate the mode, median, mean and variance.

Problems

The result of throwing two dice 120 times is represented by the table

Sums	2	3	4	5	6	7	8	9	10	11	12
No. of Times	3	8	9	11	20	19	16	13	11	6	4

Calculate the mean and standard deviation.

draw the box plot for the above data

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 ≤ 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

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

<u>The left half fragment is positively correlated</u>

<u>The right half is negative correlated</u>

Uncorrelated Data

Chapter 2: Getting to Know Your Data

Data Objects and Attribute Types

Basic Statistical Descriptions of Data

Data Visualization

Measuring Data Similarity and Dissimilarity

Summary

Chapter 2: Getting to Know Your Data

Data Objects and Attribute Types

Basic Statistical Descriptions of Data

Data Visualization

Measuring Data Similarity and Dissimilarity

Summary

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***

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

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 Measure for Binary Attributes

A contingency table for binary data

		Object *j*		
		1	0	sum
Object *i*	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":

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

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

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

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

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \ldots + |x_{i_p} - x_{j_p}|$$

h = 2: (L_2 norm) Euclidean distance

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

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.*

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

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

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

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

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

$\cos(d_1, d_2) = 0.94$

Chapter 2: Getting to Know Your Data

Data Objects and Attribute Types

Basic Statistical Descriptions of Data

Data Visualization

Measuring Data Similarity and Dissimilarity

Summary

Summary

Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled

Many types of data sets, e.g., numerical, text, graph, Web, image.

Gain insight into the data by:

- ***Basic statistical data description: central tendency, dispersion, graphical displays***
- ***Data visualization: map data onto graphical primitives***
- ***Measure data similarity***

Above steps are the beginning of data preprocessing.

Many methods have been developed but still an active area of research.

- ***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}
- Which attribute has biggest standard deviation? Do not compute standard deviations.
- 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:
- 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
- Are there any outliers (wrt IQR) in these datasets?
- Yes.
- 30 in attr2. 30 > 15+1.5*IQR
- Give a 4 element dataset whose standard deviation is zero?
- {1,1,1,1}

- **Graphic Displays of Basic Statistical Descriptions**
- Boxplot: graphic display of five-number summary
- Bar Chart: compare data across different categories.
- Histogram: x-axis are values, y-axis represent frequencies
- Quantile plot: each value xi is paired with fi indicating that approximately 100 fi % of data are $< xi$
- 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.

Boxplot: graphic display of five-number summary

```
t3 <- data.frame(vals=c(1,5,9,10,11,12,18,30),vtypes=c(1,1,1,1,1,1,1,1))
boxplot(vals ~ vtypes, data = t3, col = "red", xlab="vtpes",
        ylab="vals",main="BoxPlot1")
```

```
fivenum(c(1,5,9,10,11,12,18,30))
```

```
1.0  7.0 10.5 15.0 30.0
```

```
t7 <- data.frame(vals=c(1,2,3,4,5,6,7),vtypes=c(1,1,1,1,1,1,1))
boxplot(vals~vtypes,data=t7,col="red",xlab="vtypes",ylab="vals",main="BP1")
fivenum(c(1,2,3,4,5,6,7))
1.0 2.5 4.0 5.5 7.0
```

```r
input <- mtcars[,c('mpg','cyl')]
boxplot(mpg ~ cyl, data = mtcars, xlab = "Number of Cylinders",
        ylab = "Miles Per Gallon", col="red", main = "Mileage Data")
```

```
> input
                    mpg cyl
Mazda RX4          21.0   6
Mazda RX4 Wag      21.0   6
Datsun 710         22.8   4
Hornet 4 Drive     21.4   6
Hornet Sportabout  18.7   8
Valiant            18.1   6
Duster 360         14.3   8
Merc 240D          24.4   4
Merc 230           22.8   4
Merc 280           19.2   6
Merc 280C          17.8   6
Merc 450SE         16.4   8
Merc 450SL         17.3   8
Merc 450SLC        15.2   8
Cadillac Fleetwood 10.4   8
Lincoln Continental 10.4  8
Chrysler Imperial  14.7   8
Fiat 128           32.4   4
Honda Civic        30.4   4
Toyota Corolla     33.9   4
Toyota Corona      21.5   4
Dodge Challenger   15.5   8
AMC Javelin        15.2   8
Camaro Z28         13.3   8
Pontiac Firebird   19.2   8
Fiat X1-9          27.3   4
Porsche 914-2      26.0   4
Lotus Europa       30.4   4
Ford Pantera L     15.8   8
Ferrari Dino       19.7   6
Maserati Bora      15.0   8
Volvo 142E         21.4   4
```

Bar Chart: Example in R

```
H <- c(7,12,28,3,41)
M <- c("Mar","Apr","May","Jun","Jul")
barplot(H,names.arg = M,xlab = "Month",ylab = "Revenue",col = "blue",
        main = "Revenue chart",border = "red")
```

- A **bar chart** represents data in rectangular bars with length of the bar proportional to the value of the variable.

Histogram

```
v <-  c(9,13,21,8,36,22,12,41,31,33,19)
hist(v,xlab = "Weight",col = "yellow",border = "blue")
```

A **histogram** represents the frequencies of values of a variable bucketed into ranges.

- Histogram is similar to bar chat but the difference is it groups the values into continuous ranges.

A straight line that represents the case of when, for each given quantile, the unit price at each branch is the same.

Scatter Plot

- A **scatter plot** is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes.
  - To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane.

- The **scatter plot** is a useful method for providing a first look at bivariate data to see clusters of points and outliers, or to explore the possibility of correlation relationships.

- Two attributes, X, and Y, are **correlated** if one attribute implies the other.
- **Correlations** can be **positive**, **negative**, or null (**uncorrelated**).

```
t2 <- data.frame(price=c(10,15,20,22,25,30,45,50,60,70,80,85,90,92,95,100),
   sales=c(100,110,112,120,125,130,135,150,140,132,125,120,119,110,108,98))

plot(x=t2$price,y=t2$sales,xlab="price",ylab="sales",main="Price vs Sales")
```

```
input <- mtcars[,c('wt','mpg')]
plot(x = input$wt,y = input$mpg,
     xlab = "Weight",
     ylab = "Milage",
     xlim = c(2.5,5),
     ylim = c(15,30),
     main = "Weight vs Milage"
)
```

- Scatterplots show many points plotted in the Cartesian plane. Each point represents the values of two variables.
- One variable is chosen in the horizontal axis and another in the vertical axis.

Scatter Plot:

Positively and Negatively Correlated Data

Uncorrelated

Uncorrelated

Scatter Plot:
Positively and Negatively Correlated Data

The left half fragment is positively correlated

The right half is negative correlated

- Data visualization provides an important suite of tools for identifying a qualitative understanding. This can be helpful when we try to explore the dataset and extract some information to know about a dataset and can help with identifying patterns, corrupt data, outliers, and much more.
- **What is Data Visualization?**
- Data visualization is defined as a graphical representation that contains the information and the data.

- By using visual elements like charts, graphs, and maps, data visualization techniques provide an accessible way to see and understand trends, outliers, and patterns in data.
- It is used in many areas such as:
- To model **complex events**.
- Visualize phenomenons that cannot be observed directly, such as **weather patterns**, **medical conditions**, or **mathematical relationships**.

- **The basic uses of the Data Visualization technique are as follows:**
- It is a powerful technique to explore the data with **presentable** and **interpretable** results.
- In the **data mining process**, it acts as a primary step in the pre-processing portion.
- It supports the **data cleaning process** by finding incorrect data and corrupted or missing values.
- It also helps to **construct and select variables**, which means we have to determine which variable to include and discard in the analysis.
- In the process of **Data Reduction**, it also plays a crucial role while combining the categories.
-

- ***Different Types of Analysis for Data Visualization***
- ***Mainly, there are three different types of analysis for Data Visualization:***

- ***Univariate Analysis: In the univariate analysis, we will be using a single feature to analyze almost all of its properties.***

- ***Bivariate Analysis: When we compare the data between exactly 2 features then it is known as bivariate analysis.***

- ***Multivariate Analysis: In the multivariate analysis, we will be comparing more than 2 variables.***

- ***Univariate Analysis Techniques for Data Visualization***
- ***1. Distribution Plot***
- It is one of the best univariate plots to know about the distribution of data.
- When we want to analyze the impact on the target variable(output) with respect to an independent variable(input), we use distribution plots a lot.
- This plot gives us a combination of both probability density functions(pdf) and histogram in a single plot.
- Implementation:
- The distribution plot is present in the Seaborn package.

- ***The code snippet is as follows:***
- sns.FacetGrid(hb,hue='SurvStat',size=5).map(sns.distplot,'age').add_legend()
- data visualization techniques | distribution plot
- Some conclusions inferred from the above distribution plot:
- From the above distribution plot we can conclude the following observations:
- We have observed that we created a distribution plot on the feature ‘Age’(input variable) and we used different colors for the Survival status(output variable) as it is the class to be predicted.
- There is a huge overlapping area between the PDFs for different combinations.
- In this plot, the sharp block-like structures are called histograms, and the smoothed curve is known as the Probability density function(PDF).

- The Probability density function(PDF) of a curve can help us to capture the underlying distribution of that feature which is one major takeaway from Data visualization or Exploratory Data Analysis(EDA).
-

- **2. Box and Whisker Plot**
- This plot can be used to obtain more **statistical details** about the data.
- The straight lines at the maximum and minimum are also called **whiskers**.
- Points that lie outside the whiskers will be considered as an outlier.
- The box plot also gives us a description of the **25th, 50th,75th quartiles**.
- With the help of a box plot, we can also determine the **Interquartile range(IQR)** where maximum details of the data will be present. Therefore, it can also give us a clear idea about the outliers in the dataset.

Interquartile Range
Minimum
1st Quartile
Median
3rd Quartile
Maximum

50
40
30
20
10
0
axil_nodes
1
2
SurvStat

- **Some conclusions inferred from the above box plot:**
- From the above box and whisker plot we can conclude the following observations:
- How much data is present in the 1st quartile and how many points are outliers etc.
- For class 1, we can see that it is very little or no data is present between the median and the 1st quartile.
- There are more outliers for class 1 in the feature named **axil_nodes**.
- **NOTE:**
- We can get details about outliers that will help us to well prepare the data before feeding it to a model since outliers influence a lot of Machine learning models.
-

- **Bivariate Analysis Techniques for Data Visualization**
- **1. Line Plot**
- This is the plot that you can see in the nook and corners of any sort of analysis between 2 variables.
- The line plots are nothing but the values on a series of data points will be connected with straight lines.
- The plot may seem very simple but it has more applications not only in machine learning but in many other areas.
-