

**LECTURE NOTES
ON
DATA SCIENCE AND ANALYTICS
(6THSEM,CSE)**

**PREPARED BY
SUNITA MAHAPATRA
SR.LECT. IN DEPT.OF CSE (PKAIET, BGH)**

Data Science

- Data Science, also known as Data-driven Science, is an interdisciplinary field of scientific methods, processes, and systems to extract knowledge or insights from data in various forms - either structured or unstructured, similar to data mining.
- Convergence of various knowledge domains
- Mathematics: Statistics, Linear algebra, Optimization, Time series etc.
- Machine Learning, Data Structures, Parallel algorithms etc.
- Storage and computing platforms, Statistical tools.
- Text, Finance, Images, Economics etc.
- Visualization, Infographics
- Best practices and Heuristics: Handle mixed values in data, transform and represent data.
- As such Data Science is one of the recent fields combining big data, unstructured data and combination of statistics and analytics and business intelligence.
- Data Science correlates between Structured and unstructured data.
- It is the discipline of using quantitative methods from Statistics and Mathematics along with technology to develop algorithms to discover patterns, predict outcomes and find optimal solutions to complex problems.

→ Now-a-days, Data Scientists are in great demand as they can transform unstructured data into actionable insights, helpful for business.

Terminology related with Data Science

- Big Data: Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low-latency.
- Big data usually includes data sets with sizes beyond the ability of commonly used R/W tools to capture, curate, manage and process data within a tolerable elapsed time.
- 10^3 KB (1 kilobit)
- 10^6 byte (1 MB)
- 10^9 byte (1 GB)
- 10^{12} byte (1 terabyte) (TB)
- 10^{15} byte (1 petabyte) (PB)
- 10^{18} byte (1 exabyte) (EB)
- 10^{21} byte (1 zettabyte) (ZB)
- 10^{24} byte (1 yottabyte) (YB)

* Because big data is simply larger than life itself, it can offer a detail of the user. The amount of data produced by users has surpassed the Petabytes (PB) levels.

Business Intelligence : (BI)

- BI is the technology which uses the transformed and loaded historical data to get or create the reports.
- It is a set of methodologies, process theories that transform raw data into useful information to help companies to make better decisions.

→ BI is a process of analyzing data and presenting actionable information to help executives, managers and other corporate end users make informed business decisions and thus help in decision making.

→ common functions of business intelligence technologies include reporting, online analytical processing, analytics, data mining, process mining, performance management etc.

Data Analytics: Data analytics used to describe the field as a comprehensive collection of associated methods.

→ Data analysts collect, process and perform statistical analysis of data. Their skills may not be that advanced as data scientists but their goals are the same → to discover how data can be used

Difference between big data and business intelligence

→ The difference between big data and BI is synonymous to fishing in the sea and fishing in the lake.

→ Big data collectively refers to the act of generally capturing and usually processing enormous amounts of data on a continuing basis.

→ BI collectively refers to b/w all systems that import data streams of any size and use them to generate informational displays that point towards specific decisions.

Data Wrangling:

→ The process of conversion of data, often through the use of scripting languages, to make it easier to work with. It is known as data wrangling or data munging.

Algorithm

A series of repeatable steps for carrying out a certain type of task with data. Specific data structures often play a role in how certain algorithms get implemented.

Web Analytics:

Statistical or machine learning methods applied to web data such as page views, hits, clicks, and conversions (sales), generally with a view to learning what web interactions are most effective in achieving the organizational goals (usually sales).

→ Key challenges in web analytics are the volume and constant flow of data.

Methods of Data Repository

Data repository is a term used to refer to a destination dedicated for data storage.

→ Data repository may assume several different shapes like:

1. Data Lakes
2. Data Marts
3. Data warehousing
4. Big data and Hadoop and similar frameworks.

Data Lake

A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed and needed elsewhere.

→ Data lake shares a data environment that comprises multiple repositories and capitalizes on big data technologies. It provides data to an organization for a variety of analytics processes.

→ A data lake is a storage repository that holds an enormous amount of raw or refined data

in native format until it is accessed. The term data lake is usually associated with Hadoop - oriented object storage in which an organization's data is loaded into the hadoop platform and then business analytics and data-mining tools are applied to the data where it resides on the hadoop cluster.

→ data flows from the streams to the lake. users have access to the lake to examine, take samples or dive in.

Characteristics of Data Lake

1. All data is loaded from source systems. No data is turned away.
2. Data is stored at the least level in an untransformed or nearly untransformed look.
3. Data is transformed and schema is applied to fulfill the needs of analysts.

DATA WAREHOUSE

- Data warehouse is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or adhoc queries and decision making.
- Data warehouse store current and historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons.
- The term "Data warehouse" was first coined by Bill Inmon in 1990. According to him, a data warehouse is a subject oriented, integrated, time-variant and non-volatile collection of data. This data helps analysts to take informed decisions in an organization.

- * A data warehouse is a database, which is kept separate from the organization's operational database.
- * There is no frequent updating done in a data warehouse.
- * It processes consolidated historical data, which helps the organization to analyze its business.
- * A data warehouse helps executives to understand and use their data to take strategic decisions.
- * A data warehouse system helps in consolidated historical data analysis.

Data Warehouse Models

1. Virtual warehouse
2. Data Mart
3. Enterprise warehouse

Virtual warehouse

The view over an operational data warehouse is known as a virtual warehouse. It is easy to build a virtual warehouse. Building a virtual warehouse requires less capacity on operational database servers.

Data Mart

Data Mart contains a subset of organization-wide data. This subset of data is valuable to specific groups of an organization.

- * Windows-based or Unix/Linux-based servers are used to implement data marts. They are implemented on low-cost servers.
- * The implementation data mart cycles is measured in short periods of time. i.e. in weeks rather than months or years.
- * The life cycle of a data mart may be complex in long run, if its planning and design are not organization-wide.

- * Data marts are small in size.
- * Data marts are customized by department.
- * The source of a data mart is departmentally structured data warehouse.
- * Data marts are flexible.

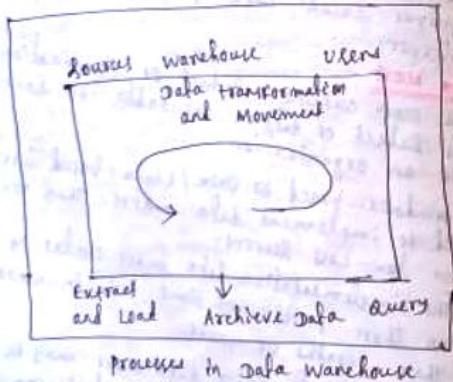
Enterprise warehouse:

- An enterprise warehouse collects all the information and the subject spanning an entire organization.
- It provides us enterprise-wide data integration.
- The data is integrated from operational systems and external information providers.
- The transformation can vary from a few gigabytes to hundreds of gigabytes, terabytes or beyond.

Process flow in data warehouse:

There are four major processes that contribute to a data warehouse.

1. Extract and load the data.
2. Cleaning and transforming the data.
3. Backup and archive the data.
4. Managing queries and directing them to the appropriate data sources.



Functions of data warehouse tools and utilities:

1. **Data extraction:** involves gathering data from multiple sources.
 2. **Data cleaning:** involves finding and correcting the errors in data.
 3. **Data transformation:** involves converting the data from legacy format to warehouse format.
 4. **Data loading:** involves sorting, summarizing, checking integrity etc.
 5. **Refreshing:** involves updating from data sources to warehouse.
- a) **Extract and Load process:** Data extraction takes data from the source systems. Data load takes the extracted data and loads it into the data warehouse.
 - b) **controlling the process:** controlling the process involves determining when to start data extraction and the consistency check on data.
 - c) **When to initiate extract:** Data needs to be in a consistent state when it is extracted, i.e. the data warehouse should represent a single, consistent version of the information to the user.
 - d) **Loading the data:** After extracting the data, it is loaded into a temporary data store where it is cleaned up and made consistent.
 - e) **Clean and transform process:** Once the data is extracted and loaded into the temporary data store, it is time to perform cleaning and transforming.
 - * clean and transform the loaded data into a structure.
 - * partition the data
 - * Aggregation
 - f) **Clean and transform the loaded data into a structure:** Cleaning and transforming the loaded data helps speed up the queries.

It can be done by making the data consistent:

1. within itself.
2. with other data within the same data source.
3. with the data in other source systems.
4. with the existing data present in the warehouse.

i) Provide the data: It will optimize the h/w performance and simplify the management of data warehouse.

ii) Aggregation: Aggregation is required to speed up common queries. Aggregation relies on the fact that most common queries will analyze a subset or an aggregation of the detailed data.

iii) Backup and Restore the data: In order to recover the data in the event of data loss, h/w failure, or h/w failure, it is necessary to keep regular back up.

iv) Query management process:

1. Manages the queries.
2. Helps speed up the execution time of queries.
3. Directs the queries to their most effective data sources.
4. Monitors actual query profiles.

Person involved with data scientist:

1. DATA Scientist

A data scientist is someone who is better at statistics than any software engineer and better at S/W engineering than any statistician.

- Data scientist implies the ability to work with large volumes of data generated not by studies but by ongoing organizational processes.
- Data scientist conduct unstructured research and frame open-ended industry questions.
- Extract huge volumes of data from multiple internal and external sources.

- Thoroughly clean and prune data to discard irrelevant information.
- Explore and examine data from a variety of angles to determine hidden weaknesses, trends.

(b) Technical skills required:

- * Math
- * Statistics
- * Machine learning tools and techniques
- * Software engg. skills
- * Data cleaning and merging
- * Data visualization
- * Unstructured data techniques
- * R and/or SAS language
- * Python, C, C++, Java, Perl
- * Big data platform like Hadoop
- * Cloud tools like Amazon

(c) Business skills required:

- * Analytic problem-solving
- * Effective communication
- * Industry knowledge

DATA Analyst:

Data analysts collect, process and perform statistical analyses of data. Their skills may not be as advanced as data scientists, but their goals are the same — to discover how data can be used to answer questions and solve problems.

→ Work with IT teams, management and/or data scientists to determine organizational goals.

→ Mine data from primary and secondary sources.

→ Clean and prune data to discard irrelevant information.

- Analyze and interpret results using standard statistical tools and techniques.
- Identify new opportunities for process improvement.
- Design, create and maintain relational databases and data systems.
- Data analysts are sometimes called junior data scientists or data scientists in training.

Technical Skills of Data Analysts:

- * Statistical methods and packages (SPSS)
- * R and SAS language
- * Data warehousing and database querying language
- * Programming (e.g. XML, JavaScript)
- * Database design
- * Data mining
- * Data cleaning and munging
- * Machine learning techniques

Business Skills:

- * Analytic problem-solving
- * Effective communication
- * Creative thinking
- * Industry knowledge

Data Science vs Data Analysis:

Data Science

- Providing strategic actionable insights into the world
- Mathematical, technical and strategic knowledge are mandatory
- Deal with big data

Data Analysis

- Providing operational observations info values
- Data analysis and visualization skills required.
- Not necessarily deal with big data

DATA Engineer :

- A specialist in data wrangling. Data engineers are the ones that take the messy data and build the infrastructure for real, tangible analysts.
- Design, construct, install, test and maintain highly scalable data management platforms.
 - Ensure systems meet business requirements and industry practices.

- Build high performance algorithms, prototypes, predictive models and POC of concepts.
- Research opportunities for data acquisition and new uses for existing data.
- Develop data set processes for data modeling, mining and production.
- Employ a variety of languages and tools (e.g. scripting languages)
- Install and update disaster recovery procedures.

Desired Technical Skills:

- A B.Tech in CSE, applied math, physics, statistics etc.
- Statistical analysis and modeling
- Database architecture
- Hadoop-based technologies (Hive, Pig)
- SQL-based technologies (MySQL)
- Python, C/C++, Java, Perl
- Matlab, SAS, R
- Data warehousing solutions
- Machine learning
- Data mining
- Unix, Linux, MS-Windows

Expected Business Skills:

- Creative problem-solving
- Effective collaboration
- Intellectual curiosity
- Industry knowledge

DATA Architect

- Data architects create blueprints for data management systems. After assessing a company's potential data sources, architects design a plan to integrate, centralize, protect and maintain them.
- This allows employees to access critical information in the right place, at the right time.

Data architect responsibilities

- collaborate with IT teams and management to define a data strategy that addresses industry requirements.
- build an inventory of data needed to implement the architecture.
- research new opportunities for data acquisition.
- identify and evaluate current data management technologies.
- develop data models for database structures.
- implement measures to ensure data accuracy and accessibility.
- build new systems with existing warehouse structures.
- produce and enforce database development standards.

Technical skills required

- Application server software (e.g. Oracle)
- DBMS S/w (e.g. Microsoft SQL Server)
- User interface and query S/w (e.g. IBM DB2)
- Enterprise application integration S/w (e.g. XML)
- Development environment S/w
- Backup / archival S/w
- Data mining
- UML
- Python, C, C++, Java, Perl
- UNIX, Linux, MS Windows
- Hadoop and NoSQL databases.

Expected Bucket Skills

1. Analytical problem-solving
2. Effective communication
3. Expert management
4. Industry knowledge
5. Data Science

6. Data scientists are big data wranglers. They take an enormous mass of messy data points (unstructured and structured) and use their skills in math, statistics and programming to clean, manage and organize them.

Types of DATA

This Data and Big Data includes huge volume, high velocity and extensive variety of data. The data in it will be of three types.

1. Unstructured data : Word, PDF, Text, Media Log.
2. Semi Structured data : XML data
3. Meta data : Data about data
4. Structured data : Relational data

Unstructured big data

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it.

→ Typical example of unstructured data is, a heterogeneous data source containing a simple text files, images, videos etc.

→ Any data that can be stored, accessed and processed in the form of fixed format is termed as a structured data.

→ Now a day organizations have wealth of data available with them but unfortunately they don't know how to derive value out of it since this data is in its raw form or unstructured format.

Semi-structured data

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in XML file.

personal data stored in a XML file

```
<rec><name>Amitabh Singh </name></rec>
    male </sex><age>45</age></rec>
<rec><name>Anurag Jain </name></rec> male
    </sex><age>39</age></rec>
```

Web pages are generated in the form of HTML which is also an example of semi-structured data.

Meta data

Metadata is defined as the data providing information about one or more aspects of the data. It is used to summarize basic information about data which can make tracking and working with specific data easier.

→ There are three main types of metadata

* **descriptive metadata**: It describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keyword.

* **structural metadata**: It indicates how compound objects are put together, for example, how pages are ordered to form chapters.

* **administrative metadata**: It provides information to help manage a resource, such as when and how it was created, file type and who can access it.

Metadata repository: Meta data repository is an integral part of a data warehouse system. It contains the following metadata.

- business metadata
- operational metadata
- data for mapping from operational environment to data warehouse
- the algorithms for summarization.

Structured: Any data that can be stored, accessed and processed in the form of fixed format is termed as a structured data.

Example

EMP-ID	EMP-Name	Gender	Dept	Salary
A001	XX	M	Finance	600000
B001	YY	F	Admin	700000
A002	XZ	F	Admin	800000

The data science process

DSP is an agile, iterative data science methodology to deliver predictive analytics solutions and intelligent applications efficiently.

- DSP helps improve team collaboration and learning.
- we provide a generic description of the process here that can be implemented with a variety of tools.
- the process may involve 7 clear cut steps for data analysis.

- Step-1 : frame or define the (business) problem.
- Step-2 : collect the raw data needed for your problem.
- Step-3 : data preparation for process the data for analysis.
- Step-4 : Explore the data.
- Step-5 : perform in-depth analysis (Modelling).
- Step-6 : Evaluation
- Step-7 : visualization and communication of results of the analysis.

Step 1: The first thing you have to do before you solve a problem is to define exactly what it is. You need to be able to translate data questions into something achievable.

You should ask questions like the following:

1. Who are the customers?
2. Why are they buying our product?
3. How do we predict if a customer is going to buy our product?

→ It's important that at the end of this stage, you have all the information and context you need to solve this problem.

Step 2: Once you have defined the problem, you will need data to give you the insights needed to turn the problem around with a solution.

→ This part of the process involves thinking through what data you will need and finding ways to get that data, whether it's querying internal databases, or purchasing external datasets.

→ In case of big data, you have to adopt machine learning process.

Step 3: Now that you have all the raw data, you will need to trawl it before it before you can do anything. Effectiveness, data can be quite messy, especially if it hasn't been well-maintained.

→ You will see errors that will corrupt your analysis. We have to check for the following errors:

1. Missing values, perhaps customers without an initial contact date.
2. Corrupted values, such as invalid entries.
3. Time zone differences, perhaps your database doesn't take into account the different timezones of users.
4. Date range errors, perhaps you will have dates

that makes no sense, such as data registered from before sales started.

Step 4: When your data is clean, you'll start playing with it. The difficulty here is not coming up with ideas to test, but coming up with ideas that are likely to turn data insights. Here to this we have to trace the patterns and analyze it deeply.

Step 5: This step of the process is where you're going to have to apply your statistical, mathematical and technological knowledge and leverage all of the data science tools at your disposal to couch the data and find the insights. → Here you might have to create a predictive model.

Step 6: You can now combine all of those qualitative insights with data from your quantitative analysis to craft a story that moves people to action.

Step 7: Proper communication will mean the difference between action and inaction on your proposals.

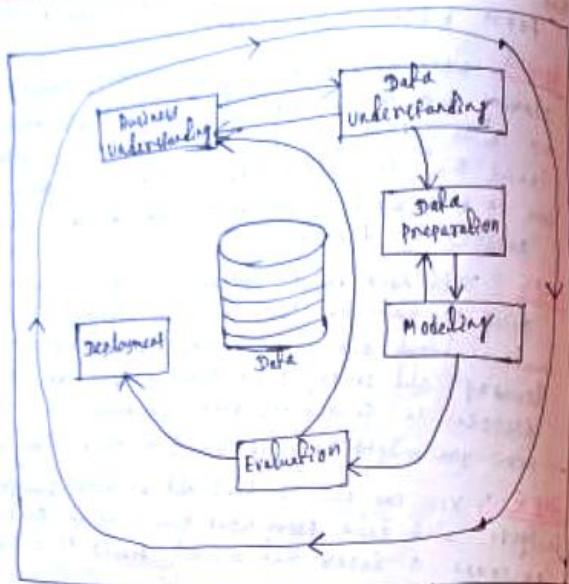
→ You need to craft a compelling story here that ties your data with their knowledge.

→ Throughout the data science process, your days-to-day vary significantly depending on where you are and you will definitely receive tasks that fall outside of this standard process.

Data Science project's lifecycle

→ The team data science process (TDSF) provides a lifecycle to structure the development of the data science project.

→ The lifecycle outlines the steps, from start to finish, that projects usually follow when they are executed.



(The 6 high-level phases of CRISP-DM suggested for the data science project)

- This life cycle has been designed for data science projects that ship as part of intelligent applications.
- These applications deploy machine learning or artificial intelligence models for predictive analysis.
- Exploratory data science projects or ad-hoc analytics projects can also benefit from using this process.
- The 6 high-level phases of CRISP-DM are still a good description for the analytics process, but the details and specifics need to be updated.
- CRISP-DM remains the top methodology for data mining projects. CRISP-DM was conceived around 1996.
- The lifecycle outlined the major stages that project typically execute, often iteratively.

* Business Understanding

* Data Acquisition and Understanding

* Modeling

* Deployment

* customer acceptance

popular data science tools

→ Tools are an important element of the data science field. The open source community has been contributing to the data science toolkit for years which has led to major advancements to the field.

1. R programming language

→ R is built by data scientists for data scientist. R is a programming language used for data manipulation and graphics.

→ R is one of the easier languages to learn as there are numerous packages and guides available for users.

→ R has a steep learning curve and is generally built for stand alone systems.

→ RStudio is the IDE for R for beginners.

2. Python: Python is another widely used language among data scientists, created by Dutch programmer Guido van Rossum.

→ It's a general-purpose programming language focusing on readability and simplicity.

→ we can do all sorts of tasks such as sentiment analysis or time series analysis with Python, a very versatile general-purpose programming language.

3. KNIME: KNIME is a SW company with headquarters in major tech hubs around the world. The company offers an open source analytics platform written in Java, used for data reporting, mining and predictive analysis.

4. SQL: Structured query language or SQL is a special-purpose programming language for data stored in relational database.

- SQL is used for more basic data querying and can perform tasks such as organizing and manipulating data or retrieving data from a database.
- 2. Apache Hadoop and other Big data tools:

 - a) Hadoop: It is a framework, written in Java, for processing large and complex datasets.
 - b) Machine Learning: It is an environment to build reliable machine learning algorithms. The algorithms are written on top of Hadoop. It is used for collaborative filtering, clustering and categorization.
 - c) Spark: It is a cluster-computing framework for data analysis. It is often preferred due to its speed.
 - d) Impala: It is the massive parallel processing (MPP) database for Apache Hadoop.
 - e) Apache Flink: It is a computational platform for real-time analytics. It is simple and easy to use.

- 3. DB (Data Science Tools):
 - D3: It is a JavaScript library for building interactive data visualizations within your browser. It allows data scientists to create rich visualizations with a high level of customizability.
 - f) Tensor Flow: Tensor Flow is the product of Google's brain team coming together for the purpose of advancing machine learning.
 - It is a Python library for numerical computation and built for everyone from students and researchers to hackers and innovators.
 - It allows programmers to access the power of deep learning.
- 4. RStudio: RStudio integrates with R as an IDE to provide further functionality. RStudio combines a source code editor, build automation tools and a debugger.

FAMILIARITY WITH EXAMPLE APPLICATIONS:

1. Airline Route Planning:

- Using data science, the airline companies can:
1. predict flight delay.
 2. decide which class of airplanes to buy.
 3. whether to directly land at the destination, or take a halt in between.
 4. Effective drive customer loyalty programs.

2. Fraud and Risk Detection:

One of the first applications of data science originated from finance discipline. Companies were fed up of bad debts and losses every year. They decided to bring in data science practices in order to rescue them out of losses.

3. Delivery Logistics:

Logistics companies like DHL, UPS have used data science to improve their operational efficiency.

→ Using data science, these companies have discovered the best routes to ship, the best delivery time to deliver, the best mode to transport to choose thus leading to cost efficiency, and many more.

4. uber's Taxi Service:

Uber is a smartphone-app based taxi booking service which connects users who need to get somewhere with drivers willing to give them a ride.

→ The business is rooted firmly in big data and leveraging this data in a more effective way than traditional taxi firms have managed has played a huge part in its success.

→ Uber's entire business model is based on the very big data potential of crowd sourcing. Fares are calculated automatically based on GPS. These algorithms monitor traffic conditions and journey times in real-time, meaning prices can be adjusted as demand for rides changes and traffic conditions mean journeys are likely to take longer.

5. people analytics

This application of analytics helps companies manage human resources.

- The aim is to find out which employees to hire, which to reward to promote, what responsibilities to assign, and similar human resource problems.

6. portfolio analytics

- A common application of business analytics is portfolio analysis. In that, a bank or lending agency has a collection of accounts of varying value and risk.
- The accounts may differ by the social status (wealthy, middle-class, poor, etc.) of the holder, the geographical location, its net value and many other factors.

7. Risk Analytics

- Predictive models in the banking industry are developed to bring certainty across the risk scores for individual customers.
- Furthermore, risk analyses are carried out in the financial world and the insurance industry.

8. digital analytics

- Digital analytics is a set of business and technical activities that define, create, collect, verify or transform digital data into reporting, research, analyses, optimizations, predictions and automation.
- Even banner ads and clicks come under digital analytics.

CHAPTER-2 DATA MANAGEMENT USING IBM SPSS.

DATA MANAGEMENT PLANNING:

- A data management plan is an integral part of the research plan. The data plan can be reviewed and expanded during research but main principles and procedures should be determined before the research starts, at the latest before data collection begins.
- Data collection and management aims to maximize efficiency of staff and resources, ensure the collection of accurate and reliable data, and focus on careful management of data once they have been collected.
- The collection, management and analysis of data, is often a neglected aspect of programme planning and implementation.

DATA ANALYSIS:

- Data analysis is at the heart of any scientific investigation.
- The module explores how scientists collect and record data, find patterns in data, explain those patterns and share their research with the larger scientific community.
- Data collection is the systematic recording of information. Data analysis involves working to uncover patterns and trends in datasets, data interpretation explaining those patterns and trends.

Data quality:

- describe procedures for ensuring data quality during the project. Technical and content decisions made at data entry stage influence the quality of data.
- Solutions chosen for post-collection processing also have an impact on data quality. Following are the measures to be adopted:
 - * Systematic and consistent naming of data files facilitates data management during research as well as data archiving and reuse.

- * Always test the technical instruments and equipment before collecting data.
- * When recording variables, use statistical software and if possible, consider recording the variables by using binary code instead of determining values for 'making' type data and 'can't say' type responses.

The Data

- * What kind of data are collected / generated?
- * In what way are data collected / generated?
- * What kind of data are collected is mainly determined by the research question. Research data are typically questionnaire surveys, interviews, focus group discussions, written material, visit or meeting recordings, official documents, websites or registers or media data.

Ethical issues - copy rights

- * Who owns the copy right, intellectual property rights and management right to data?
- * Who has the right to grant access to data?
- * What procedures are used to inform research participants?
- Copyright laws may be relevant for research data even though most empirical research data are outside the scope of standard copyright act.
- Data ^{not} requires permission from the copyright owner.
- Research teams should always make an agreement of data ownership and usage rights.
- Usage rights should be determined both for the research project and for usage after the project has been completed.
- Data & archives → specify access rights to archived data.

Confidentiality and data security

- Confidentiality in the research environment basically means planned and careful processing of personal data.
- Personal data should only be collected and processed to the degree necessary for the research and unauthorized access to the data must be prevented.

- When data contain personal data, the various personal data act requires the researcher to complete a description of the file.
- The data file description adds to the transparency of personal data processing.
- If data are collected through the internet, it is also possible to use privacy notice form.
- * Data security means keeping personal information collected, as well as computer systems, data files and transfers of data safe.
- Making back-ups of data files and preventing unauthorized access to them are thus integral parts of data security.

DATA MANAGEMENT PLAN

- A data management plan describes how research data are collected or created, how data are used and stored during research and how made accessible for others after the research has been completed.
- * It describes briefly what kind of data will be collected and how they will be collected.
- * Outline the types of data (Survey, interview, observation, photographs)
- * It also describes any editing data you will receive.

Date file format

File format is a primary factor in accessing and reading your data in the future. The file format in which we have the data may be in txt, MP3, jpg etc.

Preferring qualitative data

- Separately preserve all the documents that pertain to or have affected the data. These may include interview questions, writing instructions, transcription methods used and information sheets given to research participant.
- Documentation if each data unit is necessary to enable archiving and reuse of the data.

DATA COLLECTION AND MANAGEMENT

- * Data collection is the process of gathering and measuring information on targeted variables in an established systematic fashion, which then enables one to answer relevant questions and evaluate outcomes.
- Data collection is a component of research in all fields of study including physical and social sciences, humanities and business.
- But collecting data is only one step in a scientific investigation and scientific knowledge is much more than a simple compilation of data points.
- The world is full of observations that can be made, but not every observation constitutes a useful piece of data.
- Regardless of the field of study or preference for designing data accurate data collection is essential to maintaining the integrity of research.
- Many computer systems implement data entry forms, but data collection systems tend to be more complex, with possibly many related forms containing detailed user input fields, data validations and navigation links among the forms.
- Data collection methods are determined by the type of data collected.
- Quantitative data can be collected through interviews, postal and online questionnaires, by using existing source materials.
- Qualitative data are often collected by recording individual interviews, group interviews, lectures or meetings as audio or video files.

Data collection methods

The choice of data collection methods depends on the research problem under study, the research design and the information gathered about the variable.

1. Primary data collection methods

- The primary data are the first-hand data, collected by the researcher for the first time and is original in nature.
- The researcher collects the fresh data when the research problem is unique and no related research work is done by any other person.

2. Secondary data collection methods

- When the data is collected by someone else for his research work and has already passed through the statistical analysis is called the secondary data.
- Thus, the secondary data is the second-hand data which is readily available from the other sources.

Documentation on Data processing and controls

- Technical and context decisions made at data entry stage influence the quality of data.
- Decisions to be made include, for example, whether to enter information into a matrix or the technical solution chosen for audio or video recording.
- Solutions chosen for post-collection processing also have an impact on data quality.
- Sufficient data documentation in different logics of data collection and processing is a crucial factor for quality.
- Data documentation is also important for long-term preservation and usability.

Processing Quantitative data files

- A quantitative dataset is typically a data matrix consisting of rows and columns where one row corresponds to one observation and one column corresponds to one variable.
- To analyse quantitative data, it is required for statistical analysis as well as at least basic knowledge of statistics and quantitative methods.

i. Recording a data matrix

- Data collection method and instrument affect how the data are stored in digital form. In online surveys

- the responses are saved the moment they are submitted
- all recording methods have the potential to cause mistakes in the data.
- 1. check for and correct values out of range
- 2. check the entered data against a few randomly selected questionnaires.
- 3. check the length of rows and the number of variables.
- 4. do not record variables while entering the data.
- 5. check the accuracy of frequencies.
- 6. create documentation of all changes made to the dataset.

3. variable name and label

be consistent when naming variables favour short names that correspond to the numbering of the instrument used in data collection. Examples:

1. variables relating to actual survey questions: A good name for the variable that contains the responses to the first question in the questionnaire is q1. If it has several subquestions, the following form can be used: q1-1, q1-2, q1-3, --- and so on.
2. variables relating to questions about background information:

The background variables should be named in a consistent manner, for instance bV1, bV2, bV3, ---

3. other variables: The dataset can contain information that is not directly related to the research instrument. (e.g. observation id, date of response and time spent in responding).
- Data collected through online questionnaire often contain technical information like browser information, time of response, respondent's IP address. The variables related to this sort of information should also be named in a logical manner, for

example t1, t2, --- if there are only a few variables of this kind in the dataset, descriptive names, such as 'ID', 'date', 'time', 'IP' can also be used.

3. A variable label

A variable label refers to the description of the contents of a variable. Different statistical packages and file formats limit the length of variable labels. (For example, SPSS portable has a limit of 255 characters.)

- it is advisable to code the values of a variable to correspond to the numbering used in the research instrument, for instance:

Strongly disagree - 1

Disagree to some extent - 2

Neutral: neither agree nor disagree - 3

Agree to some extent - 4

Strongly agree - 5

- for missing data and 'categorical' types of classes, negative values and zero can be used.

4. Recoding variables

When analyzing the data, variables sometimes need to be recoded or new variables based on them need to be formed. For example, the respondent's years of birth are often required in the questionnaire, but the results are reported as age groups.

Missing data

- in almost all datasets, there are variables with missing data for some cases. For instance, a respondent may have decided not to answer a question in the questionnaire, or there may have been a failure in collecting the response.

- if the cases that have missing data are removed from the analysis, the total number of cases decreases and the accuracy of the results may suffer. The missing data should be coded so that they can be clearly distinguished from the actual values of a variable.

Weight variables

- If there are systematic errors in the dataset, it may be useful to weight the observations.
- with weight variables, potential bias in age, gender and region distributions resulting from the sampling can be corrected.

Using syntax

- Most statistical packages allow users to process and analyse data with the help of a programming language or syntax.
- Often the most effective features of statistical packages are available only through syntax commands, even though basic analyses can be performed through menus.
- The commands given in syntax can be saved to a separate file (script file).
- It is advisable to always use syntax rather than menus when editing data.
- Syntax allows users to see what changes have been made to the data and how.
- This makes it easy to perform quality control, search for potential mistakes, and make corrections and adjustments.

Processing qualitative data files

Qualitative research data may consist of many different types of research material. These may include transcribed interviews, ~~and~~ audio recordings, still images and various types of written texts.

1. Transcription

- The most common formats of qualitative data are written texts, interview data and focus group discussion data.
- In most cases, interview and discussion data are first digitally recorded and then transcribed.
 - Representing audiovisual data into written form is the most typical way of processing interview and discussion data into an ~~an~~ analyzable format.
 - The level of transcription is always decided by the original researcher or research team and is dependent on the objectives set for the data.
 - Transcription level decisions are often influenced by the resources available.

(2) Levels of Transcription

Different levels of transcription can be classified in a following manner, for instance:

1. Omitted / Summary transcription

Interview recordings are represented into written form only roughly, by listing or summarizing main points/topics. → direct quotations or parts of speech are only rarely written down.

→ Interpretation plays a big role in this kind of transcription because it is the transcriber who decides which parts are worth transcription.

2. Basic level transcription

Will produce an exact transcription of utterances but leaves out repeats, cut-offs of words and sentences, filler (*"you know"*) and non-lexical sounds (*"uh", "ah"*). → Utterances clearly nor in context can also be left out.

3. Exact transcription

All speech is transcribed, nothing is left out. Transcription is an exact, word-for-word replication of the verbal data, using the most common standardized notation symbols.

4. Conversation analysis transcription

Full verbal transcription using standardized notation symbols, with careful reproduction of speech patterns. → Transcription includes all words, timed pauses (in seconds), cut-offs of a word, volume, word stress, as well as non-lexical action (coughs, breaths, sighs, facial expressions) etc.

→ Both for the sake of one's own research and for the sake of data reuse, it is always better for the transcription to be too detailed than vice-versa.

→ If interview records have been represented into text only in a summary format, this may become a problem, even for the original researchers at the analysis stage.

→ The notation symbols used in transcription should be described in interview guidelines and consequent data documentation.

Organizing Data Files

- The data collected are entered into data files which are then stored in a data folder.
- If the research involves several independent data collections, it is advisable to create a separate data folder for each collection.
- We should include detailed information on the data collection and data processing procedures.
- Examples of relevant material:
 1. interview guidelines
 2. writing invitation
 3. observation instructions
 4. transcription guidelines
 5. writing instructions
- Depending on the amount of data, one data file can include one or more data units.
- If the textual data as a whole are not very large, it might be more useful to store all units in one data file.

Naming Data Files

Systematic and consistent naming of data files facilitates data management during research as well as later archiving and reuse.

Example: An example on how to document data file conventions used

Data files, names are formed in the following manner:

<date><type><ID1><gender><age><municipality>

<datatype><ID2>,

where

<date> is the date on which the data were collected.

<type> specifies the type of event/data material.

<ID1> is the ID of the collection event.

<gender> is the gender of the interviewee.

<age> is the age of interviewee.

<municipality> is the municipality of residence

of the interviewee.

<datatype> specifies the type of data the file contains.

For instance, "trans" means transcription.

"audio" means audio recording, and "image" means photograph.

<ID2> is the ID numbers used to separate the images connected to the collection event.

2. Documenting Background Information

- As mentioned above, from the point of view of data reuse and sharing, it is not a good practice to document background information in the file names alone.
- What background information is entered for each unit varies from data to data and is a decision of the original researcher.
- Background information may include information on the research subject and the data collection event, and notes of the researcher.

Life cycle

What happens to data after research has been finished?

- Subsequent use value of research data is largely dependent on data management measures carried out during the research.
- Effective data management before and during data collection and processing is an essential requirement for generating data that can be used afterwards for new research, learning, or teaching of methodologies.
- It is not worthwhile to preserve all research data permanently.
- Still, destroying a dataset must always be a conscious decision and not the result of an inadequate or careless data management.

Application programming interface (API)

An API is a set of ~~set~~ subroutine definitions, protocols, and tools for building application software, in general terms. It is a set of clearly defined methods of communication between various software components.

- A good API makes it easier to develop a computer program by providing all the building blocks, which are then put together by the programmers.

- An API may be for a web-based system, operating system, database, computer hardware or software library.
- An API specification can take many forms, but often includes specifications for routines, data structures, object classes, variables or remote calls.
- Test at a (GUI) makes it easier for people to use programs. Application programming interface (API) makes it easier for developers to use certain technologies in building applications.
- By abstracting the underlying implementation and only exposing objects or actions the developer needs, an API simplifies programming.

An Example of an API

When we use an application on our mobile phone, the application connects to the internet and sends data to a server. The server then retrieves that data, interprets it, performing the necessary actions and sends it back to your phone. The application then interprets that data and presents you with the information you wanted in a readable way. This is what an API is - all of this happens via API.

The Modern API

- 1. Modern APIs adhere to standards (typically HTTP and REST), that are developer-friendly, easily accessible and understood broadly.
- 2. They are treated more like products than code.
- 3. Because they are much more standardized, they have a much stronger discipline for security and governance.
- 4. Modern APIs have their own (SDLC) or design, testing, building, managing etc.

The API consists of:

1. Bring in the API module.
2. Obtain database connection

- 3. Issue SQL statements and then stored procedures
- 4. or close the connection.

Data Acquisition

- There are many ways to get a dataset like configuring an API, internet, database etc.
- To convert binary data into a useful data, we need to perform certain tasks which includes - decompress file, querying relational database etc.

Pre-processing

It is the most important part of data science, when data is incomplete or some values are missing, we need to fill some value into it and process that data to avoid any error.

Scrubbing data

The data that is obtained has inconsistencies, errors, weird characters, missing values or different problems. For that reason data has to be scrubbed or cleaned before using it.

The scrubbing techniques include:

1. Filter lines
2. Extract certain columns or words
3. Replace values
4. Handle missing values
5. Convert data from one format to another

Filtering lines: The first scrubbing operation is to filter lines. It means that from the input data every line will be calculated to determine whether it may be passed on as output. The operation may be based on certain aspects given below.

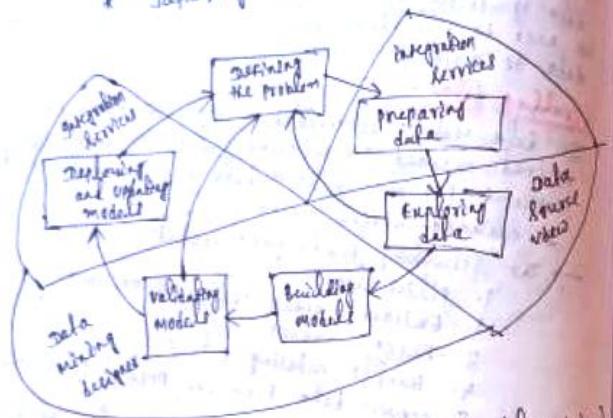
1. Based on location

2. Based on patterns

3. Based on randomness

EXPLORING DATA

- Actual to be the third step in the data mining process. It is explore the prepared data.
- It includes the following five steps:
 - * defining the problem
 - * preparing data
 - * exploring data
 - * building models
 - * exploring and validating models
 - * deploying and updating models



(Exploring data is third stage in data mining)

- The process is cylindrical, meaning that creating a data mining model is a dynamic and iterative process.
- It means we have to build several models and again if that is not adequate then we have to redefine the problem.
- Exploration techniques include calculating the minimum and maximum values, calculating mean and standard deviations and looking at the distribution of the data.

- Data that strongly deviates from a standard distribution might be biased, or might represent an accurate picture of a real-life problem, but make it difficult to fit a model to the data.
- By exploring data we can get deeper knowledge about the problem.

Building Models

case study • data collection for the CSH program
(Controlled industrial Hygiene)

- Step 1: prepare know your population and develop survey sampling frame.
- Step 2: finalize survey modules.
 - Translate and adapt modules using a translation protocol.
 - conduct pilot test for all modules.
- Step 3: collect data
 - create and expand survey packages
 - Ensure survey modules are labelled and create quality assurance measures.
- Step 4: prepare and create data entry strategies and environment.
 - create SPSS data entry templates
 - set up quality assurance measures
 - train individuals for data entry.
- Step 5: enter and clean data
 - perform ongoing quality assurance checks of data entered.
 - develop and run an error check program.
 - provide error check program and output to evaluation team.
- Step 6: **Manage and analyse data**
 - Track data as it comes in.
 - Analyse primary variables.
 - Discuss preliminary data outcomes.
 - Create standard secondary variables for all sites to use for analysis - Review and analyse secondary variables.

Labelling Survey

1. survey are labelled in the following round
data collection time - point - when follow up.
surveys are planned (TIME).
2. module type (MODULE)
 3. participant identification numbers (ID)
 4. region (Adult Module) or setting (workplace, clinical practice, youth modules) code (REGION / SETTING).
- participant identification numbers are embedded on each survey instrument or interview, including forms to be completed with biometric measures and participant consent forms.

Collecting data Survey

- An important initial step in collecting survey data is to establish a point-person for each setting.
- Once this person has been identified then a communication strategy should be developed between any evaluators and the point-person.
- Depending on time, cost and feasibility, interviewing may be administered via paper and pencil, laptop or PDA.

Data collection method

① paper and pencil

a) Type of administration (Self administered)

- * Advantage: can collect data in shorter period of time.

→ Low cost in collecting data from large sample
→ Required less people for implementation of data collection.

* Disadvantage:

- There might be delay when entering the data
- There is lag time between data collection and data entry.

b) Interview based

- * Advantage: Have higher response rate and increased rate of survey completion.
- can ensure better quality of data collected.

Disadvantage:

- will take longer time for data collection process
- might get biased response on sensitive questions.

② Observational data collection method

a) Type of administration (Data administrator)

- * Advantage: can review the data collected immediately after data is collected.
- improved quality data.

- * Disadvantage: requires certain level infrastructure in order to use this method (e.g. Internet, electricity, PDA, etc.)

- May have more logistic challenges if letting for data collection does not have equipment for data collection (e.g. computers).

b) Interview based

- * Advantage: can review the data collected immediately after data is collected.
- improved quality of data

- * Have higher response rate and increased rate of survey completion.
- ensure better data quality.

Disadvantage:

- * Requires certain level of infrastructure.
- Will take longer time for data collection process.

- * might get biased response on sensitive questions.

Storage Management

- The term Storage management encompasses the techniques and processes organizations use to maximize or improve the performance of their data storage resources.
- It is a broad category that includes virtualization, replication, mirroring, security, compression, threat automation, traffic analysis.

Storage management benefits

- Many storage management technologies, like Replication, virtualization, de-duplication and compression allow companies to better utilize their existing storage.
- The benefits of these approaches include lower costs both the one-time capital expenses associated with storage devices and the ongoing operational costs for maintaining those devices.
- Most storage management techniques also simplify the management of storage networks and devices.
- This can allow companies to save time and money by reducing the number of IT workers needed to maintain their storage systems, which in turn also reduces overall storage operating costs.
- Storage management can also help improve a data center's performance. For example, compression and deduplication can enable faster I/O, and automatic storage provisioning can speed the process of assigning storage resources to various applications.

Storage Resource Management (SRM)

- SRM often refers particularly to software used to manage storage networks and devices.
- By contrast, the term "storage management" can refer to devices and processes, as well as actual software.

→ SRM may include asset management, charge back, capacity management, configuration management, data and media migration, event management etc.

Mass Storage Management

- Mass storage refers to various techniques and devices for storing large amounts of data.
- Mass storage is distinct from memory, which refers to temporary storage areas within the computer.
- Unlike RAM, mass storage devices retain data even when the computer is turned off.

Examples of Mass Storage devices (MSD)

1. Solid State Drives (SSD)

2. Hard drives
3. External hard drives
4. Optical drives
5. Tape drives
6. RAID storage
7. USB storage
8. Flash memory cards

→ Storage management is also closely associated with networked storage solutions such as storage area network (SAN) and network-attached storage devices (NAS) devices.

→ Using SAN and NAS is more complicated for organizations than using direct-attached storage (DAS).

1. Network-attached Storage (NAS)

- Before storage networking computing environments used disks directly attached to servers to store data, while clients connected to servers over an Ethernet network using TCP/IP.
- Then SUN Microsystems developed network file system (NFS) that allowed data to be shared over an Ethernet network.

→ NAS - is a dedicated disk storage subsystem residing on a local-area network (LAN) that provides access to native files using the CIFS (Common Internet File System) (Windows) or NFS (UNIX) protocols.

2. Storage-area network (SAN):

- A SAN addresses essentially the same challenge as a network-attached storage
- The critical difference between SAN and NAS is that SAN storage delivers data in blocks rather than whole files, which makes SAN - especially Fibre Channel (FC) SAN - well suited to deliver large amounts of transactional data at very high levels of I/O performance, such as for databases
- SAN and NAS are distinguishable by the way they delivered data
- A key advantage of FC SANs is that the storage network is entirely separate from LAN. This can bring benefits to data storage and back-up.
- 3. Unified storage.
- Unified Storage - also known as multiprotocol storage - is a single data storage system that supports both file and block access
- The subsystem can support NAS, as well as Fibre Channel (block-based access) storage protocols simultaneously.
- There are two ways of achieving unified storage. You can buy arrays that are built as dedicated multiprotocol storage arrays with a NAS server and storage-area network controller combined under a single management interface.

Storage Management Technology.

- The primary organizations involved in establishing storage management standards is the Storage Networking Industry Association (SNIA).

→ It has put forth several important storage specifications including the Storage Management Interface Specification (SMI-S) and the Cloud Data Management Interface (CDMI).

→ The work of the SNIA builds on previous work done by the Distributed Management Task Force (DMTF), which has also been involved in establishing storage management standards.

Importing Data:

We shall be using IBM SPSS Statistics software for learning how to import data.

IBM SPSS Statistics:

IBM SPSS Statistics is a comprehensive system for analyzing data. SPSS Statistics can take data from almost any type of file and use them to generate tabulated reports, charts and plots of distributions and trends, descriptive statistics, and complex statistical analyses.

→ SPSS Statistics makes statistical analysis more accessible for the beginner and more convenient for the experienced user.

IBM SPSS Statistics 21 Student version:

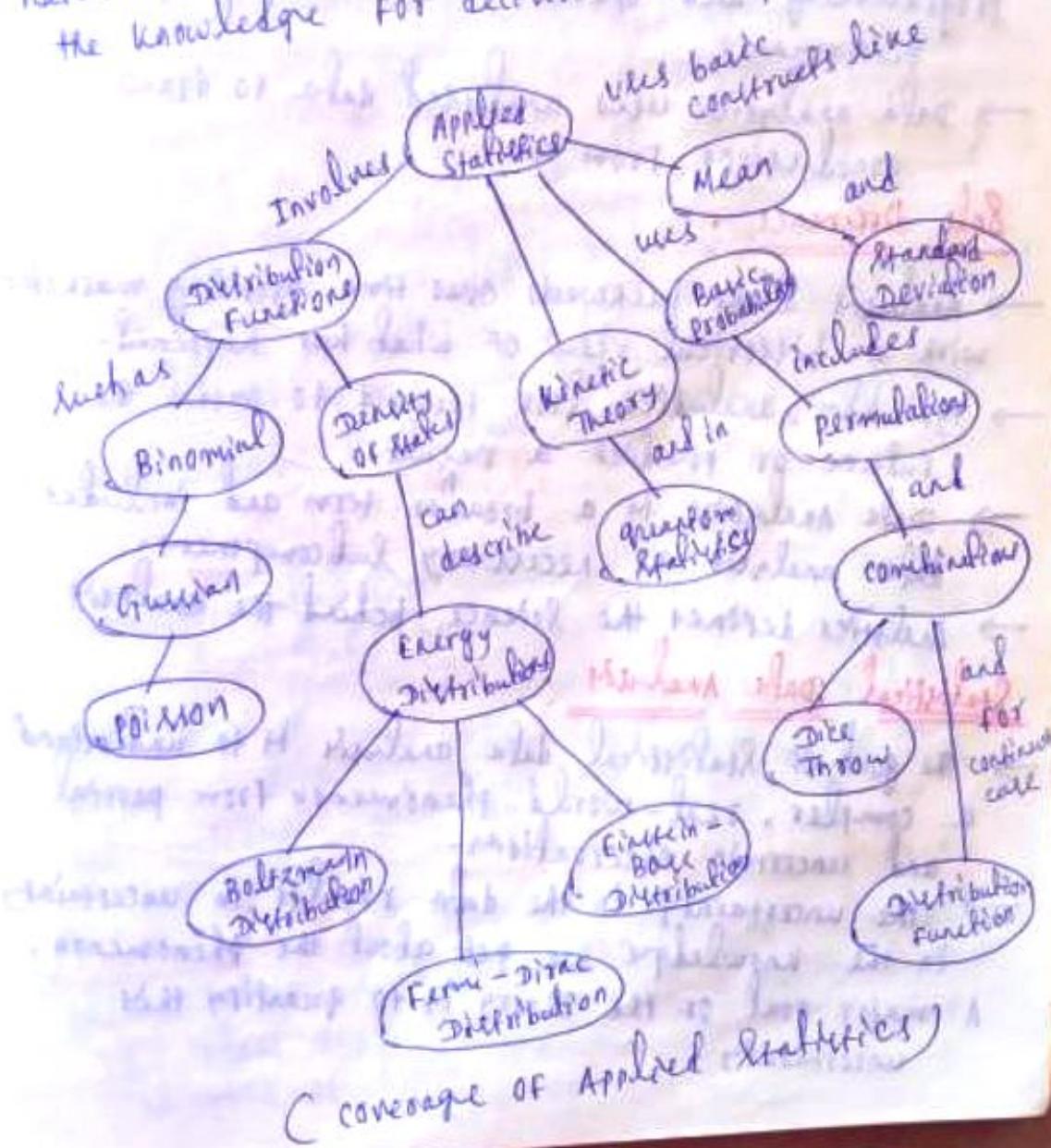
capability: The student version contains many of the important data analysis tools SPSS Statistics including:

1. spreadsheet-like data editor for entering, modifying and viewing data files.
2. Statistical procedures, including t tests, analysis of variance and correlations.
3. generative graphics that allow you to change or add chart elements.

limitations: Student version is limited to used by students and instructors for educational purpose only.

Introduction to applied Statistical techniques :

- Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation and organization of data.
- In applying Statistics to e.g., a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model process to be studied.
- Applied Statistics is a branch which covers natural processes and phenomena and provides us the knowledge for decision making.



Data analysis

- Analysis is the process of breaking a complex system into smaller parts in order to gain a better understanding of it.
- Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users.

Data analytics

- Analytics is the discovery, interpretation, and communication of meaningful patterns in data.
- Essentially valuable in areas such as statistical inference, analysis relies on a simultaneous application of statistics, computer programming and operations research to quantity performance.
- Data analysts uses analyzed data to draw conclusions from it.

Data science

- Studies looks forward over time providing motivation with a historical view of what has happened.
- Typically analysts look toward to model the future or predict a result.
- Data analytics is a broader term and includes data analysis as ancillary subcomponents.
- Analysts defend the science behind the analysis.

Statistical data analysis

- The goal of statistical data analysis is to understand a complex, real-world phenomena from partial and uncertain observations.
- The uncertainty in the data results in uncertainty in the knowledge we get about the phenomena. A major goal of the theory is to quantify this uncertainty.

→ Mathematical theory can be precisely rigorous. Surprisingly, mathematicians were able to build an exact mathematical framework to deal with uncertainty.

→ Nevertheless, there is a subjective part in the way statistical analysis yields actual turned decisions.

Univariate and Multivariate methods

In most cases, we can consider two dimensions in our data:

1. observations (or samples for machine learning tasks)
2. variables (or features)

→ Typically, observations are independent realizations of the same random process.

→ Each observation is made of one or several variables. Most of the time, variables are either numbers, or elements belonging to a finite set (that is, taking a finite number of values).

→ The first step in an analysis is to understand what your observations and variables are.

→ Our problem is univariate if we have one variable. It is bivariate if we have two variables and multivariate if we have at least two variables.

Types of Statistical Data

1. Numerical (Discrete and continuous)

1.1. Categorical, set

1.2. Ordinal

→ discrete data are called digital or binary data and continuous data types are called analog data in information technology.

1. Numerical Data: These data have meaning after a measurement, such as person's height, weight, blood pressure, or they're a count, such as number of black shares a person owns etc.

- Numerical data can be further broken into two types: discrete and continuous.
- discrete data: measurable items that can be counted. They take on possible values that can be listed. All the list of possible values may be fixed or called finite. It may go from 0, 1, 2, ... up to infinity (making it countable infinite).
- continuous data: represent measurements. These possible values cannot be counted and can only be described using intervals of the real numbers line. For example, the exact amount of gas purchased at the pump for cars with 20-gallon tanks would be continuous data from 0 gallons to 20 gallons, represented by the interval $[0, 20]$. Likewise, you might pump 8.45 liters or 9.41, or 8.414863 liters or any possible number from 0 to 20.

2. Categorical data or Qualitative data:

- Categorical data represent characteristics such as a person's gender, marital status, hometown or the type of movies they like.
- This type of data can take numerical value (such as 1 for male & 2 for female or yes/no type value).

3. Ordinal data:

- Ordinal data mixes numerical and categorical data.
- The data fall into categories, but the numbers placed to the categories have meaning. For e.g., rating a resource on a scale from 0 (lowest) to 4 (highest).

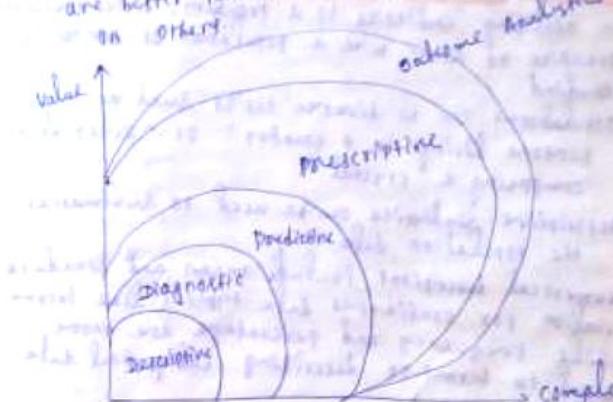
Constant and variable data:

- Variable conforming onto nominal or ordinal measurements cannot be reasonably measured numerically, sometimes they are grouped together as categorical variables, whereas ratio and interval measurements are grouped together as quantitative

variables which can be either discrete or continuous due to their numerical nature.

Type of Big Data Analytics:

- It is useful to distinguish between different kinds of analytics. Some types of analytics are better performed on some platforms than on others.



(Evaluation of Analytics within Data Science)

1. descriptive: What is Happening?

- It is the first stage of business analytics, which still accounts for the majority of all business analytics today.
- Descriptive analytics looks at past performance and understands that performance by mining historical data to look for the reasons behind past success or failure.
- Descriptive analytics, such as reporting / DASH dashboards / Scoreboards, and sales visualisation - have been widely used for sometime, and are the core applications of traditional BI (Business Intelligence).

- Descriptive analytics provides insights into what has happened historically and will provide you with tools to dig into in more detail. It gains insight from historical data with reporting, scorecards, clustering etc.
- Descriptive Statistics Application
- In applying statistics to a problem, it is common practice to start with a population or process to be studied.
- Population can be diverse topics such as "all persons living in a country" or "every atom comprising a crystal".
- Descriptive statistics can be used to summarize the population data.
- Numerical descriptors include mean and standard deviation for continuous data types (like income), while frequency and percentage are more useful to terms of describing categorical data (like race).
- Examples of descriptive analytics also include summaries, statistics, clustering and association rules used in market basket analysis.

Key points

- * Backward looking
- * pattern detection and description
- * focused on descriptions and comparisons

Diagnostic Analytics: why is it happening?

- This is the next step up in complexity in data analytics after descriptive analytics.
- An assessment of the descriptive data, diagnostic analytical tools will empower an analyst to drill down and in so doing isolate the root cause of a problem.
- Data scientists turn to this technique when trying to determine why something happened

Key points

- * Backward looking
- * focused on causal relationships and sequences
- * target / dependent variable with independent variables / dimensions

Predictive Analytics: what is likely to happen?

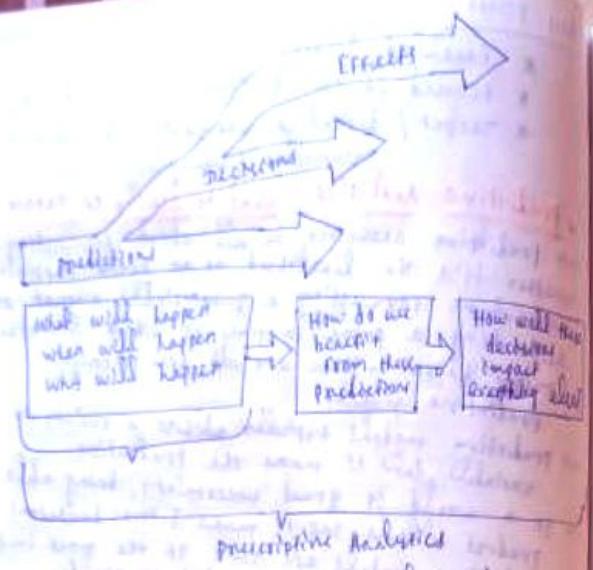
- Predictive analytics is all about forecasting whether it's the likelihood of an event happening in future - forecasting a questionable amount of estimating a point in time at which something might happen - these are all done through predictive models.

- Predictive models typically utilize a variety of variable data to make the prediction.
- In a world of great uncertainty, being able to predict allows one to make better decisions.
- Predictive models are some of the most important utilized across a number of fields.
- Examples of predictive analytics include next best offer, churn risk and medical risk analysis.

Prescriptive Analytics: what do I need to do?

- The next step up in terms of value and complexity is the prescriptive model.
- The prescriptive model utilizes an understanding of what has happened, why it has happened and a variety of "what-might-happen" scenarios to help the user determine the best course of action to take.

- Prescriptive analysis is typically not just with one individual action, but is in fact a host of other actions.
- * Forward looking
- * Focused on optimal decisions for future situations.
- * Simple rules to complex models that are centered on an automated or programmatic basis.



prescriptive Analytics

(example of prescriptive Analytics)

- A good example of this is a traffic application helping you choose the best route home and taking into account the distance of each route, the speed at which one can travel on each road, and crucially, the current traffic constraints.
- Another example might be predicting an exam time-table such that no students have clashing schedules.

Outcome Analytics

- Also referred to as consumption analytics. This technique provides insight into customer behaviour that drives specific outcomes.
- This analysis is meant to help you know your customers better and learn how they are interacting with your products and services.
- These outcomes can be varied - optimized marketing efforts, increased cash flow, accurate prediction

of customer behaviour and improved customer engagement

- * Backward looking, real-time and forward looking
- * useful application
- * definition of usage thresholds
- * focused on consumption patterns

Collecting Data for Sampling and Distribution

- Accurate data collection is essential to many business processes, to the enforcement of many government regulations, and to maintain the integrity of scientific research.
- The problem with collecting data is that you do not generally know what distribution the data follows. So you have a sample, but no distribution to help figure it out.
- We have to find out something workable distribution as per central limit theorem.

Probability

- Probability is the measure of the likelihood that an event will occur. Probability is quantified as a number between 0 and 1, where 0 indicates impossibility and 1 indicates certainty.
- The higher the probability of an event, the more likely it is that the event will occur.
- The simple example is the tossing of a fair coin. Since the coin is fair, the two outcomes (heads and tails) are both equally probable, the probability of either "heads" or "tails" is $\frac{1}{2}$, which could be written as 0.5 or 50%.

Tossing dice: when a single die is thrown, there are six possible outcomes: 1, 2, 3, 4, 5, 6. The probability of any one of them is $\frac{1}{6}$.

probability of an event happening = $\frac{\text{Number of ways it can happen}}{\text{Total Number of outcomes}}$

Frequentist: The chance of rolling a "4" with a die
 i) Number of ways it can happen : 1
 Given it only 1 face with a "4"
 Total number of outcomes : 6
 So the probability = $\frac{1}{6}$

Probability to get a quide: exactly what will probability does not tell us exactly what will happen. It's just a guide.
 → For example if we toss a coin 100 times,
 how many heads will come up?
 A probability says that we will have a $\frac{1}{2}$ chance.
 So we can expect 50 heads. But when we actually toss it we might get 48 heads, or 53 heads or anything, but in most cases it will be a number near 50.

Frequency distribution:
 A frequency distribution is a table that displays the frequency of various outcomes in a sample. It may be a graph or table.
 → Each entry in the table contains the frequency or count of the occurrences of values.
 → Frequency shows how often an event happens in any phenomena.

Frequency distribution Table:

Example : Goals
 Shyam's team has scored the following numbers of goals in recent games:
 2, 3, 1, 2, 1, 3, 3, 3, 4, 5, 4, 2, 2, 3
 Shyam put the numbers in order, then added up:
 * how often 1 occurs (2 times),
 * how often 2 occurs (5 times), and so on

- (1) Univariate distributions
- (2) Bivariate distributions
- (3) Multivariate distributions

Score:	
Score	Frequency
1	3
2	5
3	9
4	2
5	1

→ This table is also called as contingency table.
Bivariate (two way)

Bivariate (two way) frequency distributions are often represented as (two-way) contingency table

	Dance	Sports	TV	Total
Men	2	10	8	30
Women	16	6	8	30
Total	18	16	16	50

Population and parameters:
 In statistics, a population is a set of similar items or events which is of interest for some question or experiment.
 → A statistical population can be group of actually existing objects (e.g. the set of all stars within the milky way galaxy).

→ We have seen that descriptive statistics provide information about our immediate group of data. For example, we could calculate the mean and standard deviation or the mean marks for the 100 students and this could provide valuable information about this group of 100 students.

→ Any group of data like this, which includes all the data you are interested in, is called a population.

→ Statistical sampling is used quite often in statistics. A statistical sample of size n involves a single group of n individuals or subjects that have been randomly chosen from the population.

Central Tendency or Central Value:

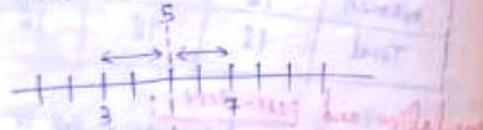
→ In statistics, a central tendency (or measure of central tendency) is a central or typical value for a probability distribution.

→ It may also be called a centre or location of the distribution.

→ The most common measures of central tendency are the arithmetic mean, the median and the mode.

Central Value:

Example : What is the central value for 3 and 7?
Answer : Half way between, which is 5.



$$3+7/2 = 10/2 = 5$$

Example : central value ~~now~~ or 3, 7 and 8?

$$\text{Ans} \quad 3+7+8/3 = 18/3 = 6$$

Measures of Central Tendency:

→ They are a combination of two words i.e. 'measure' and 'central tendency'. Measure means methods and central tendency means average value of any collected data.

→ There are three main measures of central tendency:

1. Mean
2. Median
3. Mode

→ Each of these measures describes a different indication of the typical or central value to the distribution. All are valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

The Mean:

We have been calculating the mean (or the Average):
Mean : Add up the numbers and divide by how many numbers are there. Sometimes mean are not useful.

→ If we have n values in a data set and they have values x_1, x_2, \dots, x_n , the sample mean,

$$\text{usually denoted by } \bar{x} = \frac{(x_1 + x_2 + \dots + x_n)}{n}$$

$$\text{or } \bar{x} = \frac{\sum x}{n} \quad \sum x : \text{sum of values} \\ \text{or } \bar{x} : \text{mean}$$

$$\mu = \frac{\sum x}{n} \quad (\mu : \text{mu})$$

→ The mean is essentially a model of your data set. It is the value that is most common.

→ You will notice, however, that the mean is not often one of the actual values that you have observed in your data set.

When not to use the mean

- The mean has one main disadvantage: it is particularly susceptible to the influence of outliers.
- There are values that are unusual compared to the rest of the data set by being especially small or large in numerical value.

For example:

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15K	11K	10K	14K	15K	12K	13K	17K	10K	95K

- Here the mean salary for these ten staff is 30.7K. However, inspecting the raw data suggests that this mean value might not be the best way to accurately predict the typical salary of a worker, as most workers have salaries to the 12K to 17K range.

- The mean is skewed by the two large salaries, so they need to have better central tendency like median.

- Skewness is a measure of symmetry, or more precisely, the lack of symmetry.
- A distribution of data set is symmetric if it looks the same to the left and right of the centre point.
- We usually prefer the median over the mean (or mode) to when our data is skewed.
- If we consider the normal distribution - as this is the most frequent distribution - when the data is perfectly normal, the mean, median and mode are identical. However, the median best retains this position and is not as strongly influenced by the skewed values.

Median

- The median is the middle score for a set of data that has been arranged in order of magnitude.
- The median is less affected by outliers and skewed data. In order to calculate the median, suppose we have the data below.

65	55	89	56	35	14	50	55	87	45	92
----	----	----	----	----	----	----	----	----	----	----

We first need to rearrange test data into order of magnitude (smallest first):

14	25	45	55	55	56	56	65	87	89	92
----	----	----	----	----	----	----	----	----	----	----

→ Our median mark is the middle mark - in this case, 56.

- This works fine for odd numbers, but if we have seven number data set then we have two take two middle values and find out the average of them to get a median.

→ Suppose we have 10 numbers are there for the above set.

Then Median = $\frac{5\text{th score} + 6\text{th score}}{2}$

$$= \frac{55+56}{2} = \frac{111}{2} = 55.5$$

Mode

- The mode is the most frequent score in our data set. On a histogram it represents the highest bar in a bar chart or histogram.
- We can, therefore, sometimes consider the mode as being the most popular option.

Example:

The next 8 He value that occurs most often
is 12, 13, 15, 19, 23 & 18

Example: What is the mode of 3, 4, 4, 5, 6, 6, 7?
Mode can be sometime tricky, there can be
more than one mode sometimes.
4 and 6 are modes.

- Here both y and G are nodes.



note of a populated area)

- we can see above that the most common form of transport, in this particular data set, is the bus.
 - one of the problems with mode is that it is not unique. in such cases where more than one mode is there then it's difficult to find out the best mode which describes the best way to describe the central tendency.
 - It is more problematic in case of continuous data.
 - when there are two modes to be called "bimodal", whereas when there are three or more modes we call it "multimodal".

Outline

- Outlier

 - An outlier is an observation point that is distant from other observations.
 - An outlier may be due to variability in the measurement or it may indicate experimental error. Outliers can occur by chance in any distribution and as such are more than often excluded from the data set. Otherwise an outlier can cause serious problems to statistical analysis.

There can have many anomalous cases

Unsettled anomalies

Causes: Outliers can have many anomalous causes.
A physical apparatus for taking measurements
may have suffered a transient malfunction.
There may have been an error in data
transmission or transcription.

Different Types of Electrical Meas.

- concrete types

 1. The arithmetic mean
 2. geometric mean,
 3. median and
 4. Mode

N. Mode

Arithmetic Mean: The arithmetic mean is perhaps the most commonly used statistical mean to measure the central tendency of data. It is also called the "average". It is used in most scientific experiments.

Mathematically, the arithmetic mean is given by:

= arithmetic mean

Specifically we write

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n}(x_1 + x_2 + \dots + x_n)$$

The Mean vs. the Median

- All measures of central tendency, the mean and the median each have advantages and disadvantages.
- The median may be a better indicator of the most typical value if a lot of scores has an outlier. An outlier is an extreme value that deviates greatly from other values.
 - However, when the sample size is large and does not include outliers, the mean score usually provides a better measure of central tendency.

Geometric Mean

- The geometric mean is relevant on certain sets of data, and is different from the arithmetic mean.
- Mathematically, the geometric mean is the n th root of the product of n numbers.
 - They can be written as:

$$\text{Geometric Mean} = (\alpha_1 \times \alpha_2 \times \dots \times \alpha_n)^{\frac{1}{N}}$$

where:

N = Number of datapoints

α = Score of a datapoint

or

$$(\prod_{i=1}^n \alpha_i)^{\frac{1}{N}} = \sqrt[N]{\alpha_1 \alpha_2 \dots \alpha_n}$$

- The geometric mean is relevant on those sets of data that are products or exponential in nature.
- This includes a variety of branches of natural sciences and social sciences.
- In social sciences, we frequently encounter this in a number of ways. For example, the human population growth is expressed as a percentage, and thus when population growth needs to be averaged, it is the geometric mean that is most relevant.

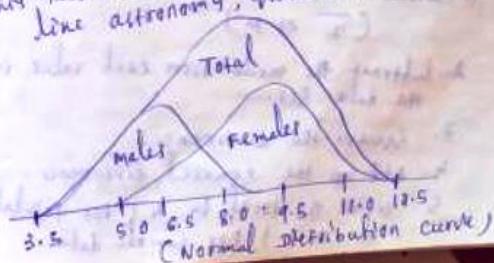
PROBLEMS OF Estimation: population or sample

A common problem in statistics is to obtain information about the mean, μ of a ~~population~~ population.

- For example: we might want to know
 1. the mean age of people in the child labour force,
 2. the mean cost of a wedding.
- If the population is small, we can ordinarily determine μ but if the population is large, however, it is generally impractical, extremely expensive or impossible.
- In such case we can take samples from the population.

Normal distribution curve

- In statistics, the theoretical curve that shows how often an experiment will produce a particular result. The curve is symmetrical and bell shaped, showing that totals will usually give a result near the average, but will occasionally deviate by large amounts.
- Normal distribution occurs very frequently in statistics, economics and natural sciences.
- This theory also finds use in advanced sciences like astronomy, quantum mechanics etc.



→ In case of perfectly symmetrical distribution
the mean, median and mode are equal.

Type of variable best measure of central tendency

Nominal	Mode
Interval	Median
Interval/Ratio (Raw)	Mean
Interval/Ratio (Ranked)	Median

Characteristic of Normal Distribution curve

→ There are four characteristics of a normal distribution. Normal distributions are:

1. Symmetric
 2. Unimodal
 3. Asymptotic
- Mean, median and mode are all equal.

→ A normal distribution is perfectly symmetrical around its centre. That is, the right side of the centre is a mirror image of the left side.

Calculation of Standard Deviation

1. Calculate the mean of the data set
(\bar{x} or M)

2. Subtract the mean from each value in the data set.

3. Square the difference.

4. Add up the squared difference.

5. Divide the total by N (for population data)
or $(n-1)$ for sample data.

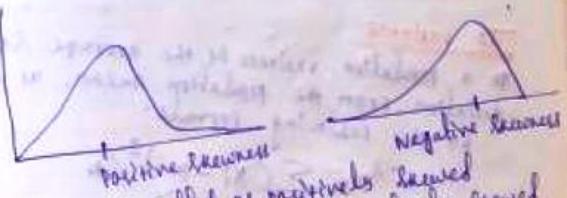
6. Take the square root of result to get the standard deviation.

SKEWNESS

→ Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution is said to be symmetric if it looks the same to the left and right of the centre point.

→ Skewness tells us about the direction of variation of the data set.

Types of Skewness: Positive Skewness Negative Skewness



The first one is called as positively skewed and the second one is known as negatively skewed curve.

KURTOSIS

→ Kurtosis is a parameter that describes the shape of a random variable's probability distribution.

→ It is the measure of the shape distribution.

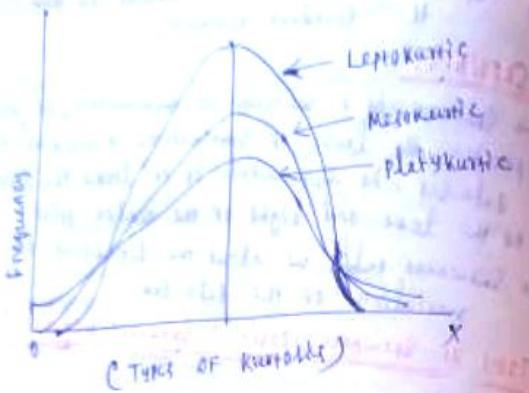
→ Kurtosis is a measure of the relative peakedness of its frequency curve.

→ Various frequency curves can be divided into three categories depending upon the shape of their peak. These shapes are formed as

* Leptokurtic

* Mesokurtic

* Platykurtic



The variance

In a population, variance is the average squared deviation from the population mean, as defined by the following formula:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Population variance

Population mean

$x_i \rightarrow$ ith element from the population

$N \rightarrow$ number of elements in the population

Sample variance (s^2)

$$= \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Sample mean

$x_i \rightarrow$ ith element from the sample

$n \rightarrow$ number of elements in the sample

The Standard Deviation.

The standard deviation is the square root of the variance.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

(Population Standard deviation)

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

↓
Sample standard deviation

Example Calculate the standard deviation for the following sample: 2, 4, 7, 6, 10, 5, 12.

Solution Actual mean method.

$$\bar{x} = \frac{2+4+8+6+10+12}{6} = \frac{42}{6} = 7$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$\sum (x_i - \bar{x})^2 = (\bar{x} - \bar{x})^2 = 0$$

$$(2-7)^2 = (-5)^2 = 25$$

$$(4-7)^2 = (-3)^2 = 9$$

$$(7-7)^2 = 0$$

$$(6-7)^2 = 1$$

$$(10-7)^2 = 9$$

$$(12-7)^2 = 25$$

$$\sum (x_i - \bar{x})^2 = 70$$

$$s^2 = \frac{70}{6-1} = \frac{70}{5} = 14$$

$$s = \sqrt{14} = 3.74$$

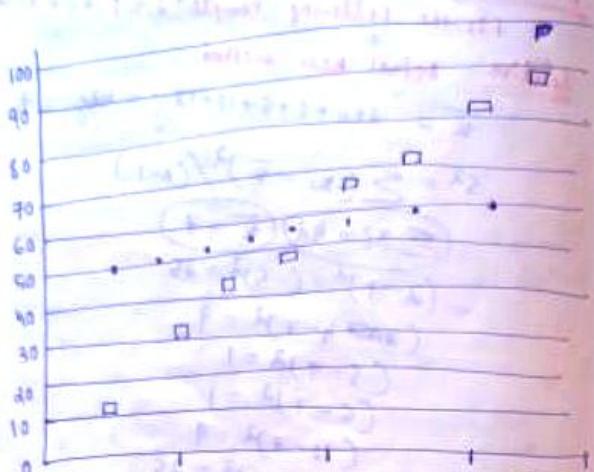
$$\text{For } \sigma^2 = \frac{70}{6} = \frac{35}{3}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{35}{3}} = 3.49$$

Through the standard deviation, we can measure the distribution of data about the mean.

Example, for example, the data points 50, 51, 52, 55, 56, 57, 58 and 60 have a mean at 55 (rectangle).

Another data set of 12, 34, 43, 48, 64, 71, 83 and 87 and it had 100-mean of 55 (rectangle).



(Two different Kinds of data collection)

- Here we can see the first data set is much more closely packed than the second one so less deviation is from the mean.
- Standard deviation provides a way to check the results.
- very large values of standard deviation can mean the experiment is fault - either there is too much noise from outside or there could be a fault in the measuring instrument.

Variability:
variability refers to how spread out a group of data is. In other words, variability measures

how much your scores differ from each other. Variability is also referred to as dispersion or spread.

→ Data sets with similar values are said to have little variability, while data sets that have values are spread out have high variability.

Student	Midterm	Final
1	71	80
2	90	99
3	100	85
4	75	95
5	55	72
6	52	78
7	100	100
8	90	92
9	85	100
10	81	87
11	83	69
12	99	88
13	89	97

- Above the table represents students marks in midterm and final. Only one student has obtained same grade in both exams.
- We want to know if the students score on each exam are similar to each other. Or if the scores are spread out. This is called variability.

Measuring variability in statistics:

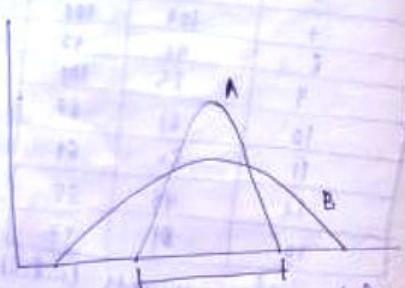
- The most common measures of variability are the range, the Interquartile range (IQR), variance, and standard deviation.

The Range:

- The range is the difference between the largest and the smallest values in a set of values.
- For example, consider the following numbers: 1, 2, 4, 5, 6, 6, 7, 11. For this set of numbers, the range would be $(11 - 1) = 10$.

Measures of variability:

- The range is the simplest measure of variability that tells us about the spread of our data.
- The range is sensitive to outliers, or values that are significantly higher or lower than the rest of data set, and should not be used when outliers are present.



Measures of variability

- Here data set B is wider and more spread out than data set A. This indicates that data set B has more variability.

- Let's calculate the range for midterm exam grades.

$$\text{Range} = \text{Highest midterm grade} - \text{lowest midterm grade}$$

$$\text{Range} = 100 - 52 = 48$$

The Interquartile Range (IQR):

The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles.

- Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second and third quartiles, and they are denoted by Q_1 , Q_2 and Q_3 , respectively.

Q_1 is the middle value in the first half of the rank-ordered data set.

Q_2 is the median value in the set.

Q_3 is the middle value in the second half of the rank-ordered data set.

- The IQR is equal to Q_3 minus Q_1 .

Example: consider the following numbers:

$$1, 3, 4, 5, 5, 6, 7, 11$$

$$Q_1 = \text{middle number in 1st half } (1, 3, 4, 5) \\ = 3 + 4 / 2 = 7 / 2 = 3.5$$

$$Q_2 = 5 + 5 / 2 = 10 / 2 = 5$$

$$Q_3 = \text{middle number in 2nd half } (5, 6, 7, 11)$$

$$= 6 + 7 / 2 = 13 / 2 = 6.5$$

$$\text{So } \text{IQR} = Q_3 - Q_1 = 6.5 - 3.5 = 3$$

R Programming Language

- R is a programming language and software environment for statistical analysis, graphics representation and reporting. The R authors are Robert Gentleman and Ross Ihaka.
- R is a well-developed, simple and effective programming language (called 'S') which includes conditionals, loops, user defined recursive functions and input and output facilities.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R is world's most widely used statistical programming language. It's the #1 choice of data scientists.

Calculating various Statistical Measures using R

- * mean
- * median
- * range
- * quartile
- * percentile
- * box plot
- * central tendency
- * standard deviation
- * correlation coefficient
- * t-test
- * interquartile range
- * variance
- * covariance
- * kurtosis

→ All these measures are based on the previously discussed built-in data set faithful.

→ old faithful Geyser data

waiting time between eruptions and the duration of the eruption for the old faithful geyser in yellowstone national park, Wyoming, USA.

Usage `faithful`

Format A data frame with 272 observations on 2 variables.

[1] 1 eruptions numeric Eruptions time in (1½ min to 5 min) minutes.

[2] 2 waiting numeric Waiting time to next eruption (in minutes) (45 min to 2 hours)

Mean Find the mean eruption duration and mean ~~as~~ waiting in the data set faithful

Solution

```
> duration = faithful$erupt
> mean(duration)
[1] 3.4878
```

```
> waiting = faithful$waiting
> mean(waiting)
[1] 70.89706
```

Here mean function is applied and the mean duration is 3.4878 minutes.

Median The median ^{is} an observations variable is the value at the middle when the data is sorted in ascending order

Example: Find the median of the eruption duration in the data set faithful.

Ans: we apply the median function to the data set faithful

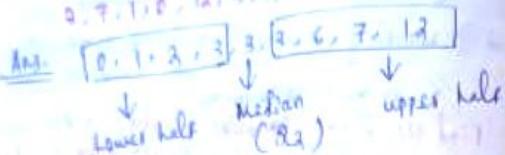
```
> duration = faithful$erupt
> median(duration)
[1] 4
```

```
> waiting = faithful$waiting
> median(waiting)
[1] 76
```

The median of the eruption duration is 4 minutes.

Quantile find the IQR for the following data set.

2, 7, 7, 8, 12, 2, 2, 3, 6



$$Q_1 = \frac{1+2}{2} = \frac{3}{2} = 1.5$$

$$Q_3 = \frac{6+7}{2} = \frac{13}{2} = 6.5$$

$$IQR = Q_3 - Q_1 = 6.5 - 1.5 = 5$$

There are general quartiles of an observation variable. The first quartile, or lower quartile, is the value that cuts off the first 25% of the data when it is sorted in ascending order. The second quartile is the median which cuts off the first 50%. The third quartile, or upper quartile, is the value that cuts off the first 75%.

Example

① duration = faithful \$ eruptions

> quantile(duration)

0%	25%	50%	75%	100%
1.60000	4.16275	4.80000	4.95425	5.10000

② quantile(EXAMPLE \$ AGE)

③ quantile(EXAMPLE \$ HEIGHT)

Range

Range = Largest value - Smallest value

max(EXAMPLE \$ AGE) - min(EXAMPLE \$ AGE)

$$(5.1) - (1.6) = 3.5$$

> duration = faithful \$ eruptions

> max(duration) - min(duration)

$$[1] 3.5 \quad (5.1) - (1.6) = 3.5$$

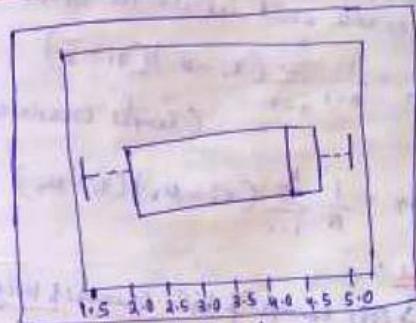
The range of the eruption duration is 3.5 minutes.

Box plot:

The box plot of an observation variable is a graphical representation based on its quartiles, as well as its smallest and largest values. It attempts to provide a visual shape of the data distribution.

> boxplot(duration, horizontal = TRUE)

Ans:



(Box plot)

Variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ (sample variance)

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

> var(duration)
[1] 1.3027 > var(EXAMPLE \$ AGE)

Standard deviation

The standard deviation of an observation variable is the square root of its variance.

Example: Find the standard deviation of the eruption duration in the data set faithful.

> duration = faithful[, "eruption"]

> sd(duration)

[1] 1.1414

The standard deviation of the eruption duration is 1.1414.

Covariance

The covariance of two variables x and y in a data set measures how the two are linearly related. A positive covariance would indicate a positive linear relationship between the variables x and y , a negative covariance would indicate the opposite.

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

(sample covariance)

$$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)$$

Example:

① > cov(eruptions ~ AGE, eruptions ~ Height)

②

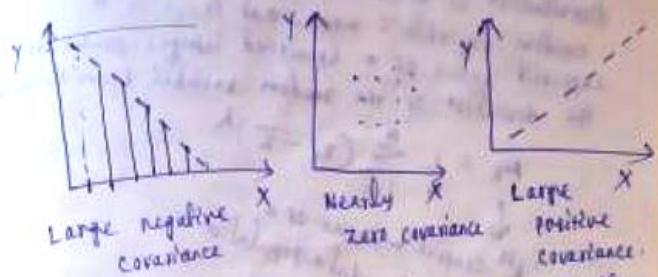
> duration = faithful[, "eruption"]

> waiting = faithful[, "waiting"]

> cov(duration, waiting)

[1] 13.978

The covariance of eruption duration and waiting time is about 13.978, indicates a positive linear relationship between the two variables.



Covariance indicates the relationship of two variables whenever one variable changes. If an increase in one variable results in an increase in the other variable, both variables are said to have positive covariance, otherwise negative covariance.

Correlation coefficient

The correlation co-efficient of two variables in a data set equals to their covariance divided by the product of their individual standard deviations. It is a normalized measurement of how the two are linearly related.

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \rightarrow \text{sample covariance}$$

(sample correlation co-efficient)

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \rightarrow \text{population covariance}$$

(population correlation co-efficient)

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \rightarrow \text{population std. deviations of } x \text{ & } y$$

(population correlation co-efficient)

Example: cov(duration, waiting)

[1] 0.9081

The co-rel. co-efficient is here 0.9081. Since it is rather close to 1, we can conclude that the two variables are positively linearly related.

Central moment

The central moment is a moment of a probability distribution of a random variable about the random variable's mean, that is, it is the expected value of a specified integer power of the deviation of the random variable from the mean.

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

↓
kth central moment of a data population

$$\mu_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$$

↓
kth central moment of a data sample

Example

> library (E1071)

> duration = faithful[, eruptions]

> moment (duration, order = 3, center = TRUE)

[1] -0.6149

The third central moment of duration is -0.6149.

Skewness

$$V_1 = \mu_3 / \mu_2^{3/2}$$

$V_1 \rightarrow$ Skewness

$\mu_2 \rightarrow$ 2nd central moments

$\mu_3 \rightarrow$ 3rd central moment

- negative skewness indicates that the mean of the data values is less than the median, and the data distribution is left-skewed while if mean is larger than median then it is right-skewed.

Example

> library (E1071)

> duration = faithful[, eruptions]

> skewness (duration)

[1] -0.41955

it indicates that the eruption duration distribution is skewed toward left

Kurtosis

$$V_2 = \mu_4 / \mu_2^2 - 3$$

Kurtosis describes the tail shape of the data distribution. The normal distribution has zero kurtosis and thus the standard tail shape. It is said to be mesokurtic. Negative kurtosis would indicate a thin-tailed data distribution, and it is said to be platykurtic. Positive kurtosis would indicate a fat-tailed distribution, and it is said to be leptokurtic.

- Find the kurtosis of eruption duration in the data set faithful.

> library (E1071)

> duration = faithful[, eruptions]

> kurtosis (duration)

[1] -1.5116 (Kurtosis is negative so platykurtic)

help (Kurtosis)

Examining the distribution of a set of data. We can examine (univariate) a set of data in large number of ways. The simplest is to examine the numbers.

> attach (faithful)

> summary (eruptions)

summary will be displayed on console

> mean(eruptions)
 $[1] 1.0800, 2.1085, 4.0000, 4.4585, 5.1000]$

> sdm(eruptions)
 → following will be displayed

> attach(faithful)

> summary(eruptions)

min	1st d.v.	median	mean	3rd d.v.	Max
1.0000	2.1085	4.0000	2.485	4.4585	5.1000

> sdm(eruptions)

attach() function in R language is used to access the variables present in the data framework without calling the data frame.

Summary(.) is a generic function used to produce result summaries of various model fitting functions.

sdm() : gt produces a stem-and-leaf plot of the values in x. gt extracts the numeric data and splits them into two parts namely, the stem and leaf.

> hist(eruptions)

> lines(density(eraptions, bw = 0.1))

→ Here bw means bandwidth defines how close are the distance between two points must be to influence the estimation of the density.

→ The density is the area of the bar that tells us the frequency in a histogram, not its height. Instead of plotting Frequency on the Y axis, we plot the Frequency density.

> sgf(eruptions) # Show the actual data points.
Quantitative data.

Quantitative data, also known as continuous data, consists of numeric data that support arithmetic operations. This is in contrast with qualitative data, whose values belong to pre-defined classes with no arithmetic operation allowed.

> head(faithful)

This head function will show the preview or faithful datframe.

(gt will show the first 6 rows of the data frame)

> dim()

Shows the dimensions means the number of rows present and the no. of columns present.

① > str(c) → name of dataset

Shows the structure of data frame

② summary(c) : provides summary statistics on the columns of the data frame.

③ colnames(c) : Shows the name of each column in the data frame.

To install new packages like (E1071) in R
 Go to Tools then install packages
 then select required packages and
 install.

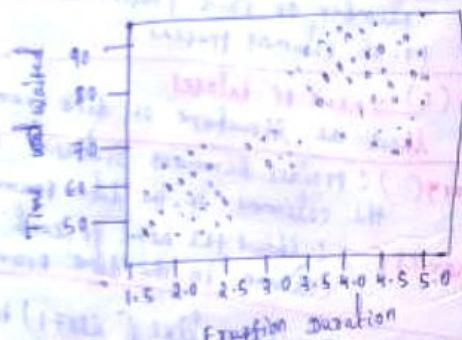
multiple labels are available
 multiple scrollable windows are used which
 allows us to view different windows at
 same time.

Scatter plot

A scatter plot pairs up values of two quantitative variables in a data set and displays them as points in a cartesian diagram.

Example: In the data set Eruption, we pair up the eruption and waiting values in the same observations at (x, y) coordinates. Then we plot the points in the cartesian plane.

- > duration = faithful\$eruptions
- > waiting = faithful\$waiting
- > plot (duration, waiting)
- + xlab = "Eruption duration"
- + ylab = "Time waited"



(Scatter data plot of eruption of geyser in the USA yellowstone national park)

Cartesian applications in the Social Sciences

In data science there are mainly three algorithms are used:

1. Data preparation, merging and process algorithms

2. Optimization algorithms

3. Machine learning algorithms

Machine Learning Algorithms

Machine learning is used to predict, categorize, classify, finding polarity, etc from the given datasets and concerned with minimizing the error.

	Unsupervised	Supervised
Clustering	<ul style="list-style-type: none"> - SVD - PCA - K-means 	<ul style="list-style-type: none"> • Regression <ul style="list-style-type: none"> * Linear * polynomial • Decision Trees • Random Forests
Categorization	<ul style="list-style-type: none"> • Association analysis • Bayesian • FP-Growth • Hidden markov model 	<ul style="list-style-type: none"> • Classification <ul style="list-style-type: none"> * KNN * Trees * Logistic regression * Naive-Bayes * SVM

Variety of machine learning algorithms

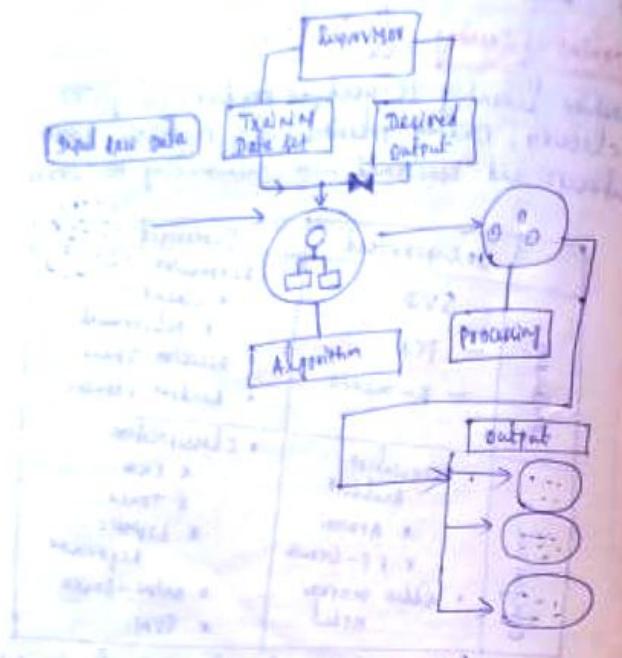
→ Machine learning is a term closely associated with data science. It uses training data for artificial intelligence.

Supervised learning

It is used for structured dataset. It analyzes the training data and generates function which will be used for the datasets.

→ It is machine learning for making predictions.

→ Core concept is to use tagged data to train predictive models. Tagged data means observations where ground truth is already known.



(Supervised Learning process)

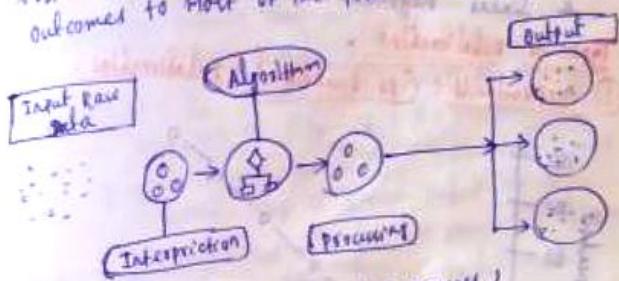
We can understand supervised learning in an even better way by looking at it through two types of problems:

1. Classification • classification problems categorize all the variables that form the output. Examples of these categories formed through classification would include demographic data such as marital status, sex, or age.

2. Regression • problems that can be classified as regression problems include types where the output variables are set at a real number. The format for this problem often follows a linear format.

Unsupervised Learning

- It is used for raw datasets that make task to convert raw data to structured data.
- In today's world there is a huge amount of raw data in every field. Therefore it is the most important part of machine learning.
- Unsupervised learning is commonly used for finding meaningful patterns and grouping inherent in data.
- Extracting generative features for analysis purpose.
- The concept of unsupervised learning is not as well defined and frequently used as supervised learning.
- During the process of unsupervised learning, the system does not have concrete data sets and the outcomes to most of the problems are largely unclear.



(Unsupervised Learning process)

Reinforcement Learning

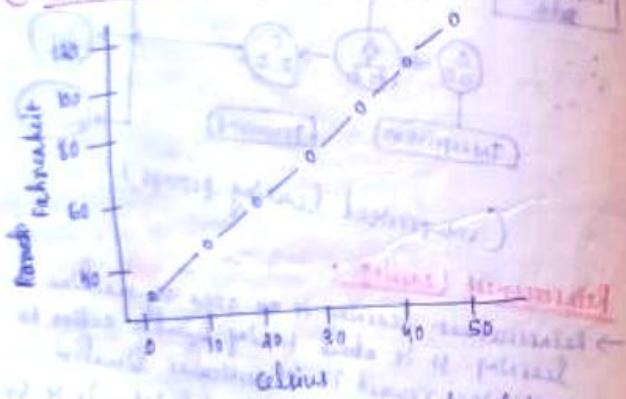
- Reinforcement learning is an area of machine learning that is about taking suitable action to maximize reward in a particular situation.
- In the absence of a training dataset, it is bound to learn from its experience.
- Some of the autonomous driving tasks where reinforcement learning could be applied include trajectory optimization, motion planning, dynamic pathing. For example, parking can be achieved by learning alternative parking policies.

Linear Regression

- Definition: Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables.
- It attempts to model the relationships between two variables by fitting a linear equation to observed data.
- One variable is considered to be an explanatory variable, and others are considered to be dependent variables.
- For example, a modeler might want to relate the weight of individuals to their height using a linear regression model.

Type of Relationship

① Deterministic (or functional) relationship



(Linear regression between Fahrenheit vs Celsius plot)

$$Fahr = 1.8 \text{ cels} + 32$$

That is, if we know the temperature in degree Celsius, we can use this equation to determine the temperature in degree Fahrenheit.

Uses of Linear Regression

- Three major uses for regression analysis are:
1. Determining the strength of prediction
 2. Forecasting an aspect, and
 3. Trend forecasting
- The regression might be used to identify the strength of the effect that the independent variable(s) has on a dependent variable.
 - It can be used to forecast effects or impact of changes.
 - Third, regression analysis predicts trends and future values.

Multiple regression analysis

Simple Linear Regression Model

$$E(y) = (\beta_0 + \beta_1 x)$$

where, β_0 is the y intercept of the regression line. And β_1 is the slope.

$E(y)$ is the mean or expected value of y for a given value of x .

→ A regression line can show a positive linear relationship, a negative linear relationship, or no relationship.

→ There are several types of linear regression analyses available to the researcher.

- * Simple linear regression
- * Multiple linear regression
- * Logistic regression
- * Ordinal regression
- * Multinomial regression
- * Discriminant analysis

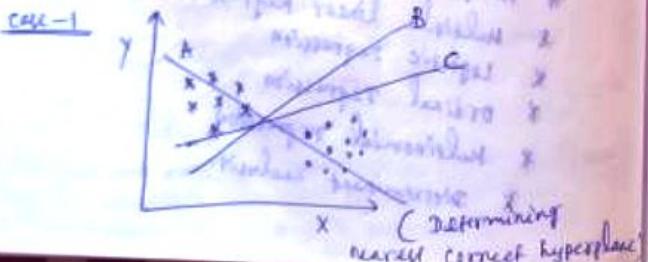
Support vector machine (SVM)

- SVM is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems.
- In this algorithm, we plot each data item as a point in n-dimensional space (where n is number features you have) with the value of each feature being the value of a particular coordinate.
- Then, we perform classification by finding the hyper-plane (in 2D space it is a classifier line, mostly straight) that differentiates the two classes (group of data) very well.



- Support vectors →ჩჩხენ ჩა წრთხის რიგი, რომელიც უკეთს განუსაზღვრს გარე კლასებს (hyper-plane/line).

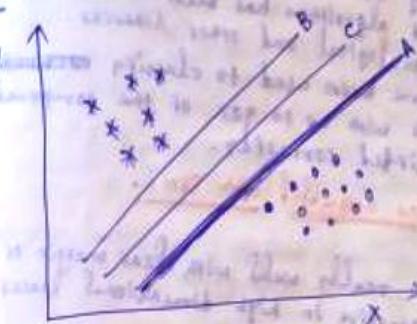
determining the right hyper-plane:



Here we have three hyper-planes (A, B and C) segregating two distinct data groups to classification stars and circle.

- Select the hyper-plane which segregates the two classes better. In this scenario, hyper-plane "C" is best option among the three.

case 3



(determining best among better Hyperplanes)

Here, C is the right hyper-plane.

case 3 Example with outliers



Hence the programmer may be unable to segregate the two classes using a straight line, as one of stars lies in the territory of other (circle) class as an outlier. One star at other end is also an outlier for star class. SVM ignores outliers and find the maximum margin.

Applications:

- SVMs can be used to solve various real world problems
- classification of images can also be performed using SVMs.
- Hand-written characters can be recognized using SVM.
- The SVM algorithm has been widely applied to the biological and other sciences.
- They have been used to classify ~~compounds~~ proteins with up to 50% of the compounds classified correctly.
- pros and cons associated with SVM.

Pros:

1. It works really well with clear margin of separation.
2. It is effective in high dimensional spaces.

Cons:

1. It doesn't perform well when we have large data set because the required training time is higher.
2. It also doesn't perform very well, when the data set has more noise i.e. target classes are overlapping.

Drawbacks with SVM:

Potential drawbacks of the SVM include the following aspects:

1. Requires full labelling of input data.
2. Uncalibrated class membership probabilities.
3. The SVM is only directly applicable to two-class tasks.
4. Parameters of a learned model are hard to interpret.

Naive Bayes:

- Naive Bayes is a simple technique for constructing classifiers that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.
- Naive Bayes is a family of algorithms not a single algorithm.
- All naive bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.
- For example, a fruit may be considered to be an apple if it is red, round and about 10cm in diameter.
- A classifier sorting fruits into apples and oranges would know that apples are red, round and are a certain size, but would not observe all these things at once. Oranges are round too, after all.
- An advantage of naive Bayes is that it only requires small numbers of training data to estimate the parameters necessary for classification.
- Naive Bayes classifiers algorithms use Bayes theorem.
- * The most popular application of naive Bayes is spam filters which looks at email messages for certain keywords and puts them in a spam folder if they match.
- Naive Bayes model is easy to build and particularly useful for very large data sets.
- Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Naive Bayes

Naive Bayes provides a way of calculating marginal probabilities $P(C|X)$ from $P(C)$, $P(X)$ and $P(C|X)$.

$$P(C|A) = \frac{P(CAA)}{P(CA)} \rightarrow \text{probability of } C \text{ given } A$$

$$P(C|A) = \frac{P(ABA)}{P(CA)} \rightarrow \text{probability of } A \text{ given } C$$

$$\Rightarrow P(C|AB) = P(C|A) \cdot P(A)$$

Ex: what are probabilities of 2 girls given at least one girl

$$\begin{aligned} &= P(2G | \text{at least } 1G) \\ &= \frac{P(1G|2G) \cdot P(2G)}{P(1G)} \quad \text{prob. of } G = 100\% \\ &= \frac{1 \cdot \frac{2}{3}}{\frac{3}{4}} = \frac{2}{3} \quad \text{GG, GB, BG, BB} \\ &= \frac{1}{2} \times \frac{2}{3} = \frac{1}{3} \end{aligned}$$

Pros and Cons of Naive Bayes

Pros

1. It is easy and fast to predict class of test data (i.e., it also performs well in multi-class prediction).
2. When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.

(cont.)

1. If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 probability and will be unable to make a prediction. That is called often as "zero frequency".

2. On the other side Naive Bayes is also known as a bad estimator, so the probability outputs are not to be taken too seriously.

Applications of Naive Bayes Algorithms

Frequentist and Bayesian methods

a) Frequentists interpret a probability as a statistical average across many independent realizations (law of large numbers).

→ Bayesian interpret it as a degree of belief (no need for many realizations). The Bayesian interpretation is very useful when only a single trial is considered.

Real time prediction

b) Naive Bayes is an easier learning and it is fast. Thus, it could be used for making predictions in real time.

c) Multi class prediction: Here we can predict the probabilities of multiple classes of target variable.

d) Text classification / Spam filtering / Sentiment analysis: Naive Bayes classifiers mostly used in text classification have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering and sentiment analysis (to identify positive and negative customer sentiments).

e) Recommendation System: Here recommendation system uses machine learning and data mining techniques to filter user's information and predict whether a user would like a given resource or not.

CHAPTER - 4 DATA VISUALISATION

Data Visualisation

- Data visualisation is a general term that describes any effort to help people understand the significance of data by placing it in a visual context.
- Patterns, trends and correlations that might go undetected in text-based data can be exposed and recognized easier with data visualization software.
- Data visualisation is both art and science that is viewed as a branch of descriptive statistics by some, but also as a grounded theory development tool by others.

Data attributes

When it comes to data attributes, there are two categories: quantitative data and qualitative data.

Quantitative data

Quantitative data can take shape of:

1. Ratio (age 10 yrs old, 20 yrs old)
2. Data we can perform arithmetic operations on (add, divide etc.)
3. Intervals (temperature -5°, 10°, 25° or time 1 am, 5 pm)
4. Data with a set value that you cannot perform all arithmetic operations on.
E.g. we can't calculate the sum of temp. during a week but we can calculate the average temp. per day and the high/low for each day.

Other data types for visualisation

1. Time-series: A single variable is captured over a period of time, such as the unemployment rate.

② Ranking: categorical subdivisions are ranked in ascending or descending order, such as a ranking of sales performance by sales person during a single period.

③ Deviation: A bar chart can show comparison of the actual versus the reference amount.

④ Frequency distribution: Shows the number of observations of a particular variable for given interval, such as the number of years in which the stock market return is between intervals such as 0-10%, 11-20%. A histogram, a type of bar chart, may be used for this analysis.

Correlation

⑤ Geographic or geospatial: Comparison of a variable across a map or layout, such as the unemployment rate by state or the number of persons on the various floors of a building. A cartogram is a typical graphic used.

Qualitative data

Ordinal (size small, medium, large or position 1st place, 2nd place)

E.g.: A large elephant in India is very different from a large elephant in Africa.
(data with a fixed ranking with intermediate distance between the values)

Nominal: (computers laptop vs. desktop)
Data where you can distinguish between values but not order them.

→ Based on these classification, the methods for aggregation and visualisation of the data needs to adjust accordingly.

Importance of data visualization:

- Makes it easier to visualize data to communicate information clearly and concisely via statistical graphics, plots and information graphics.
- Numerical data may be encoded using dots, lines and bars.
- Effective visualization helps users analyze and reason about data and evidence.
- It makes complex data more accessible, understandable and usable.
- Because of the way the human brain processes information, using charts or graphs to visualize large amounts of complex data is easier than poring over spreadsheets or reports.

Data visualization can also:

- * Identify areas that need attention or improvement.
- * clarify which factors influence customer behaviour.
- * Help you understand which products to place where.
- * predict sales volumes.

Graphical displays should:

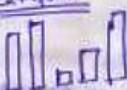
1. Show the data.
2. Directly induce the viewer to think about the substance rather than about methodology.
3. avoid distorting what the data has to say.
4. present many numbers in a small space.
5. make large data sets coherent.
6. Encourage the eye to compare different pieces of data.
7. Serve a reasonably clear purpose: description, exploration, tabulation or decoration

Conventional Data visualization methods

- Many conventional data visualization methods are often used. They are: table, histogram, scatter plot, line chart, bar chart, pie chart, flow chart, bubble chart, tree diagram, data flow diagram, E-R diagram etc.
- The additional methods are parallel coordinates, treemap and semantic network etc.
 - Parallel coordinates are used to plot individual data elements across many dimensions. Parallel coordinate is very useful when to display multidimensional data.
 - Treemap is an effective method for visualizing hierarchies.

Visual perception and data visualization

- A human can distinguish differences in like length, shape, orientation, and colour readily without significant processing effort. These are referred to as "pre-attentive attributes".
- For example, it may require significant time and effort to identify the number of times the digit "5" appears in a series of numbers called "attentive attributes".

<u>pattern</u>	<u>Example</u>
High, low and in between	
Going up, down and remaining flat	
Sleep and gradual	
Steady and fluctuating	

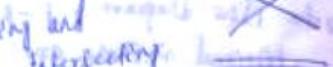
* Linear and Repetitive



* Straight and curved



* Non-interacting and interacting



* Isometrical at viewed



* wide and narrow



* clusters and gaps



* Tightly and loosely distributed



* Normal and abnormal

→ visualizations are not only static, they can be interactive. The steps of interactive visualization are as follows:

* Selecting

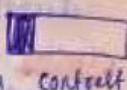
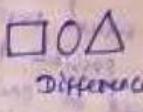
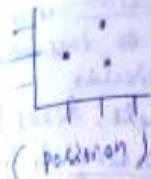
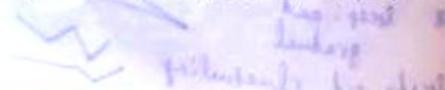
* Linking

* Filtering

* Rearranging or Remapping

Mapping of Data visualization:

For all data to be mapped to a visualization, there are some basic options of display:



(Data mapping: Basic options of display)

Retinal variables:

Seven variables for visualization:

1. two planar variables

2. FIVE retinal variables

* size

* color value There are two planar variables (size and color)

* color hue

* shape

* orientation

(the x and y position on the map plane)

Types of visual variables:

A visual variable can be:

1. Selective (e.g. color hue)

2. Allocative (e.g. shape)

3. Ordered

→ A visual variable is selective (e.g. colour hue) and therefore fundamental for symbolization of data, if all symbols can be easily isolated to form a group of similar symbols based on this variable.

Applications of DATA Science Technologies

For visualization and Bokeh (Python)

Applications of Data Science technologies for visualization:

- data visualization has become the standard for modern business intelligence (BI).
- virtually all BI software has strong data visualization functionality.
- data visualization tools have ~~become~~ been important in data analytics and making data-driven insights available to workers throughout an organization.
- data visualization software also plays an important role in big data and advanced analytics projects.
- data visualization tools can be used in a variety of ways. Many business departments implement data visualization S/w to track their own initiatives. For example, a marketing team might implement the software to monitor the performance of an email campaign, tracking metrics like open rate.

Introduction to python:

- Firstly, python is a general purpose programming language and it's not only for data science.
- It is a high-level language. It was made to be simple, "user-friendly" and easy-to-interpret.
- Python handles different data structures very well.
- It has very powerful statistical and data visualization libraries.

- It has many packages that assist solving complex analytics projects just as much as advanced data science projects.
- When it comes to learn data coding - one has to focus on three main languages:
 - Python 2 & Python 3
 - R
 - Batch
- Python 2.7 is around since 2008 - and 95% of the data science related features and libraries have been migrated already.
- On the other hand Python 2 won't be supported after 2020. So learning Python 2 is useful in some cases but the future is for Python 3.
- Choosing a development environment:
- Once Python has installed, there are various options for choosing an environment. Here are the 3 most common options:
 - Terminal / shell based
 - IDLE (default environment)
 - Ipython notebook
- When we run Python 3 in the command line, it means we are telling the interpreter to start translating.
- Python 2 and Python 3 are the same language but they rely on slightly different interpreters to translate.
- For e.g. In Python 2, we print "VK Jain", but in Python 3 we do print ("VK Jain"). The meaning is the same but the way we write it is different.
- The syntax for Python 2 is not understood by the Python 3 interpreter, and vice versa.

Starting with Python:

Starting Python: 

Start → All programs → Python 2.7
→ IDLE (Python GUI):

```
>>> 101 * 101
10201
```

Basic Numeric operations

To do numeric calculations in Python, you can write expressions that look more or less like algebraic expressions in many other common languages.

The "+" operator is addition, "-" is subtraction, "*" to multiply, and use "/" to divide.

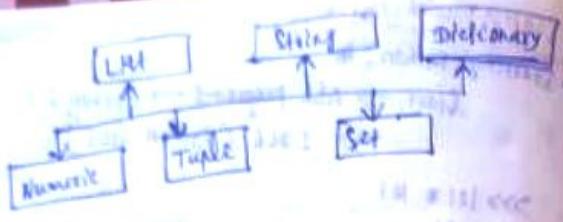
Here are some examples:

```
>>> 100 - 99
1
>>> 99 * 100
9900
>>> 79 * 5.2 - 62.2
395.0
```

Data types in Python:

Python supports various data types. These data types defines the operations possible on the variables and the storage methods.

All of standard data types available in Python are discussed below:



Numeric

- Just as expected numeric data type is None.
- numeric values.
- they are immutable data types, that means that you cannot change its value.
- python supports three different Numeric data type.

Integer type: It holds all the integer values i.e. all the positive and negative whole numbers.

Floating type: It holds the real numbers and are represented by decimal and sometimes each scientific notations with E or e indicating the power of 10 ($2.5 \times 10^3 = 2500$), example -

10.25

Complex type: There are of the form $a+bi$, where a and b are floats and i represents the square root of -1 (which is an imaginary number). Example $-10+6i$.

1. $A = 10$

2. # convert it into float type

3. $B = float(A)$

4. print(B)

>>> A = 10

convert it into float type

>>> B = float(A)

>> print(B)

>>> print(B)

10.0

1. $A = 10.76$

2. # convert it into int type

3. $B = int(A)$

4. print(B)

Following will be output on screen:

>>> A = 10.76

convert it into int type

B = int(A)

>> print(B)

10.0

>>>

List Data Type

you can consider the List as Array in C, but in List you can store elements of different types, but in array all the elements should be of the same type.

→ List is the most versatile data type available in python which can be written as a list of comma separated values (Items) between square brackets.

1. Subjects = ['physics', 'chemistry', 2]
2. print(Subjects)

Following will be the output on screen

>>> Subjects = ['physics', 'chemistry', 2]

>>> print(Subjects)

['physics', 'chemistry', 2]

LINt Operations

Syntax Result

Subjects[0] physics → This will give the index 0 value from the Subjects List.

Subjects[0:2] physics, chemistry → They will not include 2 in the List.

`subject[2] = "maths"` \rightarrow `[physics, chemistry,
"maths"]`

This will update the list and add
"maths" at index 3 and remove the value

`del subject[2]` \rightarrow `[physics, chemistry]`
This will delete the index value 2
from subject list

`len(subject)` \rightarrow `[physics, chemistry,
"maths", 2, 1, 2, 3]`

This will return the length
of the list.

`subject * 2` \rightarrow `[physics, chemistry,
"maths", 2]`
`[physics, chemistry,
"maths", 2]`

This will repeat the subjects
list twice.

String Data Type

Strings are amongst the most popular types
in python. We can create them simply
by enclosing characters in quotes.

→ Python treats single and double quotes in
exactly the same fashion. Consider the
example below:

`1 A = "welcome To python Tutorial"`

`2 B = " That Is great "`

Following shell be output on screen:

`>>> A = " Welcome To python Tutorial"`

`>>> print(A) [A:0]`

Welcome To python Tutorial

`>>> B = " That Is great "`

`>>> print(B)`

`>>> That Is great`

`>>> sys.exit()`

`print(len(String-Name))` String Length

`print(String-name.index("char"))` Locates a
character in string

`print(String-Name.count("char"))`

Count the number of times
a character is repeated in a string

`print(String-Name[start:stop])` Slicing

`print(String-Name[: -1])` Reverse a
string

`print(String-Name.upper())` converts the
letters in a
string to upper-
case

`print(String-Name.lower())` converts the letters
in a string to
lower-case

Examples

`>>> A = "GOD"`

`>>> print(A[-1: -1])`

`GOD`

`>>>`

`>>> A = "upper case"`

`>>> print(A.upper())`

`UPPER CASE`

`>>> A = "UPPER CASE"`

`>>> print(A.lower())`

`lower case`

`>>> print(A.lower())`

`lower case`

- Set Data Type
- A set is an unordered collection of items.
 - Every element is unique.
 - A set is created by placing all the items (elements) inside curly braces {}, separated by comma.
 - Consider the example below:
 - In set every element has to be unique.
 - Try running the below code:

```
1. set-1 = {1, 2, 3}
2. set-2 = {1, 2, 3, 2}
```

Here 2 is repeated twice but it will print it only once.

 - a) Union: Union of A and B is a set of all elements from both sets. Union is performed using operator. Consider the below example:

```
A = {1, 2, 3, 4}
B = {3, 4, 5, 6}
```

instead, print(A | B)

Output shall be:

```
set([1, 2, 3, 4, 5, 6])
```

 - b) Intersection: intersection of A and B is a set of elements that are common in both sets. Intersection is performed using & operator.
 - Consider the example below:
 1. A = {1, 2, 3, 4}
 2. B = {3, 4, 5, 6}
 3. print(A & B)

Output shall be:

```
set([3, 4])
```

Dictionary Data Type

- Let us take example of Aadhar card. Which is a unique ID which has been given to all Indian citizens. So for every Aadhar number there is a name and few other details attached.
- Now you can consider the Aadhar number as a key and the person's detail as the value attached to that key.
- In Python dictionary like type contains these key value pairs enclosed within curly braces and keys and values are separated with ':'. Type following at the interpreters command prompt

```
>>>
1. dict = {'Name': 'Jyotikha', 'Age': 44}
```

We will go through various dictionary operations.

Dictionary Operations:

- a) Access elements from a dictionary:
 1. dict = {'Name': 'Jyotikha', 'Age': 44}
 2. print(dict['Name'])
- b) changing elements in a dictionary:

Type following at the interpreters command prompt >>>

```
1. dict = {'Name': 'Jyotikha', 'Age': 44}
>>> print(dict['Name'])
```

Jyotikha

```
>>> dict['Age'] = 32
>>> dict['Address'] = 'Bengaluru'
>>> print(dict['Name'], dict['Age'], dict['Address'])
```

```
>>> print(dict['Age'])
32
>>> print(dict['Address'])
>>> print(dict['Address'])
>>> Bengaluru
```

Loop in Python

In python, there are three types of loops:
• while • for • Nested

- a) while loop: Here first the condition is checked and if it's true, control will move inside the loop and execute the statements inside the loop until the condition becomes false.
→ Type following at the interpreters command prompt >>>

```
count = 0
while (count < 9):
    print "The count is:", count
    count = count + 1

print "Good bye!"
```

The count is: 0
The count is: 1
The count is: 2
The count is: 3
The count is: 4
The count is: 5
The count is: 6
The count is: 7
The count is: 8
Good bye!

- b) for loop: Like the while loop, the for loop also allows a code block to be repeated certain number of times.
→ The difference is, in for loop we know the amount of iterations required unlike while loop, where iterations depend on the condition.

```
Fruits = ['Mango', 'apple', 'Grapes']
for index in range(len(Fruits)):
    print (Fruits[index])
```

Mango
Apple
Grapes

- c) Nested loops: It basically means a loop inside a loop. It can be a for loop can be inside a for loop or a while loop inside a while loop. It can be also a for loop inside a while loop and vice-versa.

```
1. count = 1
2. for i in range(8):
3.     print (str(i)*i)
4. for j in range(0,i):
5.     count = count + 1
```

→ The output shall be:

```
1
2 2
3 3 3
4 4 4 4
5 5 5 5 5
6 6 6 6 6 6
7 7 7 7 7 7 7
```

Modules: A module allows you to logically organize your python code. Grouping related code into a module makes the code easier to understand and use.
→ A module is a python object with arbitrarily named attributes that you can bind and reference.
→ Simply, a module is a file containing python code.
A module can define functions, classes and variables.
A module can also include reusable code.

```
#!/usr/bin/python
import calendar
cal = calendar.month(2018, 1)
print ("Here is the calendar:")
print (cal)
```

They would produce the following result:

Here is the calendar						
January 2019						
Mo	Tu	We	Th	Fr	Sa	Su
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

The Import Statement:

You can use any Python source file as a module by executing an import statement in some other Python source file. The import has the following syntax -

```
import module1[, module2[, ... moduleN]]
```

→ When the interpreter encounters an import statement, it → imports the module if the module is present in the search path.

→ For example, to import the module support.py, you need to put the following command :

```
# !usr/bin/python
# import module support
import support
```

Now you can call defined function from module as follows

```
support.print_func("zara")
```

```
0%! Hello : zara
```

A module is loaded only once, regardless of the number of times it is imported. This prevents the module execution from happening over and over again if multiple imports occur.

The From... Import Statement:

→ Python's from statement lets you import specific attributes from a module into the current namespace. The from...import has the following syntax -

```
from modname import name1[, name2[,... nameN]]
```

```
from fib import Fibonacci
```

→ This statement does not import the entire module fib into the current namespace, it just introduces the item Fibonacci from the module fib into the global symbol table.

The From... Import * Statement:

→ It is also possible to import all names from a module into the current namespace by using the following import statement -

```
from modname import *
```

→ This provides an easy way to import all the items from a module into the current namespace; however, this statement should be used sparingly.

Locating Modules:

When you import a module, the Python interpreter searches for the module in the following sequences -

The current directory.

→ If the module isn't found, Python then searches each directory in the shell variable `pythonpath`.

→ If all else fails, Python checks the default path. On UNIX, this default path is normally `/usr/local/lib/python/`.

→ The module search path is stored in the built-in module list as the `sys.path` variable.

The PYTHONPATH Variable

- The PYTHONPATH is an environment variable consisting of a list of directories. The syntax of PYTHONPATH is the same as that of the shell variable PATH.
- Here is a typical PYTHONPATH from a Windows system -
Set PYTHONPATH = C:\python20\lib;
- And here is a typical PYTHONPATH from a UNIX system -
Set PYTHONPATH = /user/local/lib/python

Namepaces and Scoping

- variables are named (identifiers) that map to objects. A namespace is a dictionary of variable names (keys) and their corresponding objects (values).
- A python statement can access variables in a local namespace and in the global namespace.
- If a local and a global variable have the same name, the local variable shadows the global variable.
- Each function has its own local namespace.
- class methods follow the same scoping rule as ordinary functions. Python makes educated guesses on whether variables are local or global. It assumes that any variable assigned a value in a function is local.
- Therefore, in order to assign a value to a global variable within a function, you must first use global statement.

Example : Money = 2000

```
def AddMoney():
    # global Money
    Money = Money + 1
```

print(Money)

AddMoney()

print(Money)

Here, we define Money to global namespace. Within the function Money, we assign Money a value. Therefore Python assumed Money as local variable. However, we accessed the value of the local variable Money before letting it, so an UnboundLocalError is the result.

The dir() Function

- The dir() built-in function returns a sorted list of strings containing the names defined by a module.
- The list contains the names of all the modules, variables and functions that are defined in a module.

Example : #!/user/bin/python
import built-in module math
import math
content = dir(math)
print(content)

O/P :

```
[ '__doc__', '__loader__',  
 '__name__', '__package__', '__spec__',  
 'acos', 'acoth', 'acsin', 'atanh', 'atan2',  
 'atanh', 'ceil', 'comb', 'copysign', 'cos',  
 'cosh', 'degrees', 'e', 'erf', 'erfc', 'exp',  
 'expm1', 'fabs', 'factorial', 'floor', 'fmod',  
 'frexp', 'fsum', 'gamma', 'hypot', 'isclose',  
 'isnan', 'isqrt', 'ldexp', 'lgamma', 'log',  
 'log10', 'log1p', 'log2', 'modf', 'pi', 'perm',  
 'pi', 'pow', 'rad2deg', 'rad2pi', 'rsh', 'sqrt',  
 'tan', 'tanh', 'true']
```

Library

A Python library is a reusable chunk of code that you may want to include in your program.

→ Here, a "library" loosely describes a collection of core modules.

① Numpy: It is used for numerical python. The most powerful feature of Numpy is n-dimensional arrays. This library also contains basic linear algebra functions, Fourier transforms, advanced random number capabilities and tools for integration with other low-level languages like Fortran, C, C++ etc.

② Scipy: It is used for scientific python. Scipy is built on Numpy. It is very useful library for science & engineering modules like Fourier transforms, linear algebra, optimization & Sparse matrices.

③ matplotlib: It is used for plotting vast variety of graphs, starting from histograms - line plots, bar charts, pie charts etc.

④ pandas: It is used for structured data operations and manipulations, data munging and wrangling. pandas is a s/w library written for the python programming language for data manipulation and analysis.

⑤ dash: It is used for creating interactive plots, dashboards and data applications on modern web-browsers. It can be used over very large or streaming datasets.

⑥ seaborn: It is used for statistical data visualization. It is based on matplotlib.

→ Seaborn is a library for making attractive and informative statistical graphics in python.

⑦ Scikit: It is used for machine learning. It is built on Numpy, Scipy and matplotlib, that contains a lot of efficient tools for machine learning and statistical modeling including regression, classification, clustering etc.

Matplotlib (matplotlib) (Python matplotlib):

Matplotlib is a plotting library used for 2D graphics in python language. It can be used in python scripts, shell, web application servers and other GUI toolkits. There are several toolkits available that extend python matplotlib functionality. Some examples are: Batmips, cartopy, Excel tools, Matplotlib etc.

Example: From matplotlib import pyplot as plt
plt.plot([1, 2, 3], [3, 4, 5])
plt.show()

Bar charts

Use the `bar()` function to make bar charts, which includes customizations such as error bars, bar(x, height, width=0.8, bottom=None, align='center', data=None, **kwargs)

Example: (IDLE)

```
import matplotlib.pyplot as plt
n = ["Science", "Coms", "Art"]
h = [200, 300, 500]
plt.bar(n, h)
plt.xlabel("Courses")
plt.ylabel("Students enrolled")
plt.title("Students enrolled for diff courses")
plt.show()
```