



Table Of Contents

[Chapter 1 Linear Algebra 1-1](#_Toc147741360)

[Learning concepts 1-1](#_Toc147741361)

[Data Science Categories 1-1](#_Toc147741362)

[Execution Environments 1-1](#_Toc147741363)

[Data Asset Management 1-1](#_Toc147741364)

[Code Asset Management 1-4](#_Toc147741365)

[Development Environments 1-4](#_Toc147741366)

[Open-Source Tools for Data Science 1-5](#_Toc147741367)

[List the open-source data management tools. 1-5](#_Toc147741368)

[List the open-source data integration and transformation tools. 1-5](#_Toc147741369)

[List the data visualization tools. 1-5](#_Toc147741370)

[List the model tools for building, deployment, monitoring, and assessment. 1-5](#_Toc147741371)

[List tools for code and data asset management. 1-5](#_Toc147741372)

[Algorithm Analysis 1-5](#_Toc147741373)

[Algorithm Analysis 1-5](#_Toc147741374)

[Data Examples 1-5](#_Toc147741375)

[Algorithm Analysis 1-6](#_Toc147741376)

[Chapter 2 Data 2-7](#_Toc147741377)

[What Is Data? 2-7](#_Toc147741378)

[Types of Data 2-7](#_Toc147741379)

[Quantitative data: 2-7](#_Toc147741380)

[Qualitative data: 2-7](#_Toc147741381)

[Data Examples 2-8](#_Toc147741382)

[Types of Data 2-8](#_Toc147741383)

Top of Form

# 

# **Linear Algebra**

## ***Learning concepts***

Overview of Tools

Language

Libraries, APIs, Datasets, and Models

## ***Data Science Categories***

Before raw data can be useful, raw data must pass through different data science categories such as Data Management, Data Integration and Transformation, Data visualization, Model Building, Model deployment, Model Monitoring and Assessment.

***Data Asset Management***

***Execution Environment***

***Development Environments***

***Code Asset Management***

***Execution Environments***

An execution environment has libraries to compile the source code and system resources that execute and verify the code.

Cloud based 🡪 IBM Watson studio, AWS, Google Cloud, Azure.

* + 1. ***Data Asset Management***
* Data asset management, also called digital asset management (DAM).
* Platform for organizing and managing of important data collected from different sources.
* It allows version controlling and collaboration. And also supports replication, backup and access right management for stored data.

1. Data Management

* Collecting, persisting, and retrieving data securely, efficiently and cost-effectively.
* Data is collected from many sources like,
  + Flipkart
  + Twitter
  + Media
  + Internet
  + Sensors
* Store-collected data is available so can be collected.

1. Data Integration and Transformation

* Extract, Transform and Load (ETL).
* **Extract** the data and save it in a central repository.
* Some of the data is distributed in multiple formats.
  + Database
  + Data cube
  + Flat files
* The data is extracted from these files and saved and stored in central repositories like data warehouses.
* **Data Transformation** is the process of transforming data into values, structure, and proper useful format.
* Transformed data is **loade**d back to the data warehouse.

1. Data visualization

* Graphical representation of data and information.
* Visual representation can be in the form of charts, plots, maps and animations.
* Data visualization conveys data more effectively for making decision .
* It is more crucial to data representation and the data science process.
* Charts
  + Bar charts 🡪 compares the size of each component
  + Tree map 🡪 displays hierarchy data,
  + Line chart 🡪 plots a series of data points over time
  + Map chart 🡪 displays data by location. Map charts can also be applied to other locations like websites.

1. Model Building

* Train the data and analyze patterns with machine learning algorithms.
* The system **learns** how to provide predictions or decisions by itself.
* Later, by using this model to make predictions on new, unseen data.

1. Model Deployment

* The process of integrating a developed model into a production environment.
* Business users can access and interact with the data through these Machine learning model that is available to third-party applications via APIs. This helps users to make data-based decisions.
* ***Example:*** SPSS Collaboration and Deployment Services can be used to deploy any type of asset created by the SPSS software tools suite.

1. Model Monitoring and Assessment

* Model monitoring 🡪 Tracks deployed models.
* Model assessment 🡪 Checks for Accuracy, robustness and fairness monitoring
* Model monitoring and assessment run continuous quality checks to ensure the model’s accuracy, fairness, and robustness.
* Model monitoring uses tools like Fiddler to track the performance of deployed models in a production environment.
* Model assessment uses evaluation metrics like
  + F1 score,
  + true positive rate, or
  + the sum of squared error

to understand a model's performance.

* **IBM Watson Open Scale** is a popular Model Monitoring and Assessment tool.
  + 1. ***Code Asset Management***
* Unified view where one can manage an inventory of assets.
* When one wants to develop a model, need to update it, fix bugs, or improve code features incrementally. All of this requires version control.
* Developers use versioning to track and manage changes to a software project's code.
* When working on a model, teams need a centralized repository where everyone can upload, edit, and manage the code files simultaneously.
* Collaboration allows diverse people to share and update the same project together.
* Example is GitHub.

***Development Environments***

* Development Environments are also called as Integrated development environment (IDEs).
* IDEs provide a workspace and tools to develop, implement, execute, test, and deploy source code.

## ***Open-Source Tools for Data Science***

* + 1. ***List the open-source data management tools.***
    2. ***List the open-source data integration and transformation tools.***
    3. ***List the data visualization tools.***
    4. ***List the model tools for building, deployment, monitoring, and assessment.***
    5. ***List tools for code and data asset management.***
* L

Scalar

## ***Algorithm Analysis***

## ***Algorithm Analysis***

* + 1. ***Data Examples***

## ***Algorithm Analysis***

# *****Data*****

## ***What Is Data?***

Data is defined as a collection of individual facts or statistics.

(While “datum” is technically the singular form of “data,” it’s not commonly used in everyday language.)

Data can be in the form of text, figures, images, numbers, graphs, or symbols. Data is gathered by observations, measurements, research, or analysis.

Data is a raw form of knowledge and is simple. It may even seem useless until it is analyzed, organized, and interpreted.

### ***Types of Data***

There are two main types of data:

* **Quantitative data** is provided in numerical form that can answer questions like “How many”, “How much”, and “How often”. For example the weight, volume, or cost of an item.
* **Qualitative data** is descriptive, but non-numerical, like the name, sex, or eye color of a person.

### *Quantitative data:*

Data are easily amenable (responsive)

### *Qualitative data:*

Qualitative data, also known as categorical data, describes the data that fits into the categories. Qualitative data are not numerical. Categorical measures are defined in terms of natural language specifications, but not in terms of numbers. One of the examples is grouped data. More precisely, categorical data could be derived from countable qualitative data analysis, or from quantitative data analysis grouped within given intervals.

### ***Data Examples***

* The number of visitors to a website in one month
* Inventory levels in a warehouse on a specific date
* Individual satisfaction scores on a customer service survey
* The price of a competitor’s product.Top of Form

### ***Types of Data***

* **Structured Data:**

Data that is organized in a tabular format, often found in databases. Each row corresponds to a record, and each column corresponds to a field.

* **Unstructured Data:**

Data that does not have a predefined structure, such as text documents, images, and videos.

* **Semi-Structured Data:**

Data that is not strictly organized like structured data but still has some level of structure, such as JSON or XML files.

Data can be categorized into various types based on its nature, structure, and characteristics. The main types of data include:

**\*\*1. Numerical Data:**

* Numerical data consists of numbers and is used for quantitative measurements.
* Two subtypes:
  + **Continuous Data**: Data that can take any value within a specific range. For example, height, weight, temperature.
  + **Discrete Data**: Data that can only take specific, distinct values. For example, the number of cars in a parking lot, and the count of items sold.

**2. Categorical Data:**

* Categorical data represents categories or labels and is used for qualitative analysis.
* Two subtypes:
  + Nominal Data: Categories without any inherent order or ranking. For example, colours, and types of animals.
  + Ordinal Data: Categories with a meaningful order or ranking. For example, educational levels (high school, bachelor's, master's).

**3. Text Data:**

* Text data consists of unstructured textual content, such as documents, articles, emails, and social media posts.
* Often used in natural language processing (NLP) tasks like sentiment analysis and text classification.

**4. Time Series Data:**

* Time series data is collected over successive time intervals. It's used for analyzing trends, patterns, and seasonality.
* Common in fields like finance, economics, and environmental monitoring.

**5. Image Data:**

* Image data represents visual information in the form of pixels.
* Used in computer vision tasks like image recognition, object detection, and image generation.

**6. Audio Data:**

* Audio data represents sound waves and is used in applications like speech recognition and music analysis.

**7. Geospatial Data:**

* Geospatial data includes information related to geographical locations, coordinates, and maps.
* Used in geographic information systems (GIS) and location-based services.

**8. Binary Data:**

* Binary data consists of only two possible values, often represented as 0s and 1s.
* Used in computer systems, cryptography, and digital communication.

**9. Mixed Data:**

* Mixed data involves a combination of different data types within a dataset. For instance, a dataset about customers might include numerical attributes (age), categorical attributes (gender), and text attributes (address).

Understanding the types of data is crucial for determining appropriate analysis methods, data preprocessing techniques, and modelling approaches. Different types of data require different methods for handling, visualizing, and extracting insights.

**Importance of Data**

Data serves as the foundation for making informed decisions, deriving insights, identifying trends, and developing models in various fields such as business, science, healthcare, and more.

**Data Collection**

The process of gathering data from various sources, which can include surveys, sensors, websites, databases, and more.

**Attributes or Features**

These are the individual characteristics or variables that make up a dataset. For example, in a dataset about customers, attributes could include name, age, and location.

SQL is a powerful language used for managing and manipulating relational databases. Below is a basic SQL course outline to get you started:

**1. Introduction to Databases:**

* Understand what databases are and why they are used.
* Learn about the differences between relational and non-relational databases.

**2. Basics of SQL:**

* Learn about SQL syntax, statements, and clauses.
* Explore SELECT statements to retrieve data from a database.

**3. Filtering and Sorting Data:**

* Use the WHERE clause to filter data based on conditions.
* Learn to use the ORDER BY clause to sort query results.

**4. Working with Functions:**

* Understand various SQL functions like COUNT, SUM, AVG, MAX, MIN, etc.
* Explore how to perform calculations on data.

**5. Joins and Relationships:**

* Learn about different types of joins (INNER, LEFT, RIGHT, FULL) to combine data from multiple tables.
* Understand primary keys and foreign keys for establishing relationships between tables.

**6. Grouping and Aggregating Data:**

* Use the GROUP BY clause to group data based on specific columns.
* Learn about the HAVING clause to filter grouped results.

**7. Subqueries and Derived Tables:**

* Explore subqueries to perform operations within queries.
* Understand derived tables (also known as subquery factoring or Common Table Expressions).

**8. Data Modification:**

* Learn how to INSERT, UPDATE, and DELETE data from tables.
* Understand the importance of data integrity and transaction management.

**9. Advanced SQL Topics:**

* Study concepts like indexes, views, stored procedures, and triggers.
* Explore techniques for optimizing SQL queries.

**10. Practical Projects:** - Apply SQL concepts to real-world scenarios by working on data manipulation, analysis, and reporting projects.