Essentials on

DATA SCIENCE

Prepared by:

Leonard Andrew Mesiera
IT Prof.

VALUE OF DATA

“DATA IS THE NEW GOLD”

Data

Top five companies like Apple, Alphabet, MIcrosoft, Amazon and Facebook – are all data riched firm.

Many companies are seizing this opportunity to improve their data.

Explosive Growth in Data

- **Fact:** The amount of data generated globally is growing exponentially.
- Sources: Social media, IoT devices, e-commerce, healthcare systems, etc.
- **Reason:** Organizations realize that data is a valuable asset, often referred to as "the new oil."

Advancement in Technology

- **Cheaper Storage:** Cloud platforms (e.g., AWS, Google Cloud) make storing large datasets affordable.
- **Faster Processing:** Advances in hardware (GPUs, TPUs) enable rapid computation of massive datasets.
- **AI & ML:** Better algorithms have made it easier to derive insights from data.

Increased Demand in Data-Driven Decisions

Companies are shifting from intuition-based to **data-driven decision-making**.

Industries like finance, marketing, and healthcare rely on predictive analytics for competitive advantage.

Example: Netflix uses data science to recommend content, boosting user engagement.

What is DATA SCIENCE ?

What is Data Science

What is Data Science?

- **Definition:**
  Data Science is a field that combines **math, statistics, programming, and domain knowledge** to analyze and interpret complex data for decision-making.
- **Goal:**
  To extract valuable **insights** from raw data.
- **Example:**
  Predicting customer behavior, optimizing delivery routes, detecting fraud.

Data Science Life Cycle

1. **Collect Data:** Gather raw data from various sources (e.g., surveys, sensors, logs).
2. **Clean Data:** Remove noise and inconsistencies in the data.
3. **Analyze Data:** Use statistical and computational methods.
4. **Visualize & Communicate:** Present insights using charts, dashboards, or reports.
5. **Apply Insights:** Help make better decisions or automate processes.

How Does Data Science Relate to Data Mining?

Data Mining:

- A subset of data science.
- Focuses on discovering patterns in **large datasets** using algorithms.
- Example: Finding groups of customers with similar buying habits.

Relationship:

Data mining provides raw patterns and structures that data science uses and interprets for actionable insights.

How Does Data Science Relate to Data Analytics?

- **Data Analytics:**
  - Another subset of data science.
  - Focuses on applying statistical techniques to **interpret data** and answer specific questions.
  - Example: Evaluating a marketing campaign's success.
- **Relationship:**
  Data analytics is like a tool or method that helps achieve the broader goals of data science.

Visual Comparison

Aspect	Data Science	Data Mining	Data Analytics
Scope	Broad (end-to-end process)	Narrow (pattern discovery)	Medium (interpretation-focused)
Purpose	Insights and predictions	Patterns and structures	Answer specific questions
Tools	Python, R, ML models, databases	SQL, clustering, classification	Excel, Tableau, Power BI

Data Science in 2025

Data science's rise is fueled by the **intersection of data availability, technological innovation, and business needs**. Its ability to solve real-world problems and drive innovation has made it indispensable in today's world.

WHAT IS DATA MINING ?

Data Mining

Data mining is the process of extracting useful and previously unknown information or patterns from large volumes of data. It involves using computational techniques to analyze and discover meaningful insights, trends, correlations, and relationships within datasets.

Importance of Data today ?

Thanks to Data Science, AI has ...

Computer Systems able to perform tasks that normally require human Intelligence, data science was able to bridge the gap between computer And human, by using well defined algorithms.

Data Analytics

Data analytics is a critical process for businesses and organizations looking to leverage the vast amounts of data that are available today to drive informed decision-making, improve operations, and stay competitive in the marketplace.

Data Analytics vs Data Science

- Both involve working with data.
- Data Analytics: Focuses on insights for decision-making from existing data.
- Data Science: Uses scientific methods for broader knowledge extraction from data.

What are the *different* types of *Data Analytics?*

Data Analytics

In a technical sense, data analytics can be described as the process of using data to answer questions, identify trends, and extract insights. There are multiple types of analytics that can generate information to drive innovation, improve efficiency, and mitigate risk.

Data Analytics

There are four key types of data analytics, and each answers a different type of question:

1. Descriptive analytics asks, “What happened?”
2. Diagnostic analytics asks, “Why did this happen?”
3. Predictive analytics asks, “What might happen in the future?”
4. Prescriptive analytics asks, “What should be done next?”

Descriptive Analytics

Descriptive analytics primarily uses observed data to identify key characteristics of a data set. It relies solely on historical data to provide reports on past events. This type of analysis is also used to generate ad hoc (as needed) reports that summarize large amounts of data to answer simple questions like “how much?” or “how many?” It can also be used to ask deeper questions about a specific problem. Descriptive analytics is not used to draw inferences or predictions from its findings; it is just a starting point used to inform decisions or to prepare data for further analysis.

Descriptive Analytics

The descriptive analytics process is as follows:

1. Ask a historical question that needs an answer, such as “How much of product X did we sell last year?”
2. Identify required data to answer the question
3. Collect and prepare data
4. Analyze data
5. Present results

Examples of descriptive analytics include:

- Summarizing historical events such as sales, inventory, or operations data
- Understanding engagement data such as likes and dislikes or volume of page views over time
- Reporting general trends like revenue growth or employee injuries
- Collating survey results

Diagnostic Analytics

Diagnostic analytics enhances the descriptive analytics process by digging in deeper and attempting to discover the cause(s).

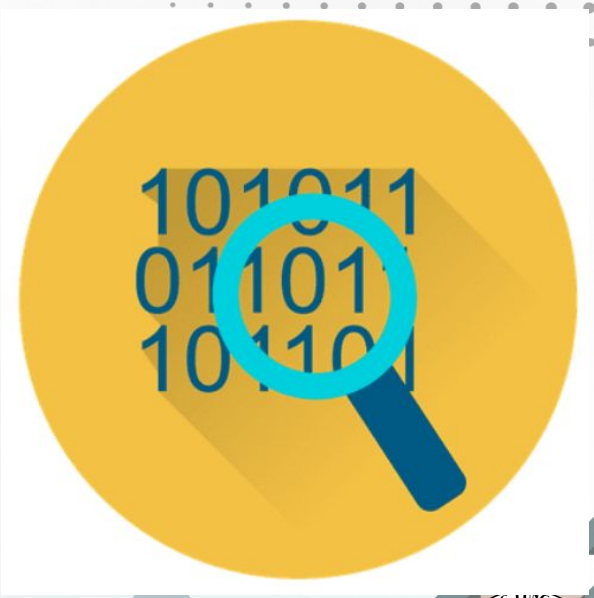
The diagnostic analytics process is as follows:

1. Identify anomalies (inconsistencies) in data sets
2. Collect data related to the anomalies
3. Use statistical techniques to uncover relationships and trends that could explain the anomalies
4. Present possible causes

An example of diagnostic analytics is using subscription cancellations, correlated with customer comments and ratings, to determine the most common reasons why users cancel subscriptions. Another example would be determining whether there is a correlation between the demographics of consumers and their purchasing patterns at specific times of year.

Predictive Analytics

Predictive analytics utilizes real-time and/or past data to make predictions based on probabilities. It can also be used to infer missing data or establish a predicted future trend. Predictive analytics uses simulation models and forecasting to suggest what could happen going forward, which can guide realistic goal setting, effective planning, management of performance expectations, and avoiding risks. This information can empower executives and managers to take a proactive and fact-based approach to strategy and decision making.



Predictive Analytics

The predictive analytics process is as follows:

1. Ask a forward-thinking question, such as “Can we predict how much product X we will sell next year?”
2. Collect and prepare data
3. Develop predictive analytics models
4. Apply models to the prepared data
5. Review models and present results

Examples of predictive analytics include:

- Forecasting customer behavior, purchasing patterns, and identifying sales trends
- Predicting customer preferences and recommending products to customers based on past purchases and search history
- Predicting the likelihood that a given customer will purchase another product or leave the store
- Identifying possible security breaches that require further investigation
- Predicting staffing and resourcing needs

Prescriptive Analytics

Leveraging insights to optimize outcomes.

Prescriptive analytics builds on descriptive and predictive analysis by recommending the most beneficial courses of action for your organization. In simpler terms, it tells you **what to do** in a given situation.

Empowering informed decisions at all levels.

This empowers **executives, managers, and employees** to make **data-driven decisions** that drive success.

Real-world example: Navigating with intelligence.

*Imagine using a GPS app that not only shows different routes, but also recommends the **best route** based on **real-time traffic, road conditions, and your preferences** (shortest distance or fastest time).*

Prescriptive analytics takes data analysis to the next level, guiding you towards optimal outcomes.

Objective

Topic Title	Topic Objective
Tools for Data Understanding	Describe common software tools used in data analytics.
Basic Excel Concepts and Features	Use basic Excel functions to gather and examine data.
Data Organization	Explain how variables and values are used in data analysis.

Data Analytics Process

Process

There are many different models of data analytics, and in this course, we will use the six-step Data Analysis Lifecycle shown in the figure below. Think of this model as iterative, meaning that some steps could be repeated multiple times before decision makers are confident enough to move forward to the next step.

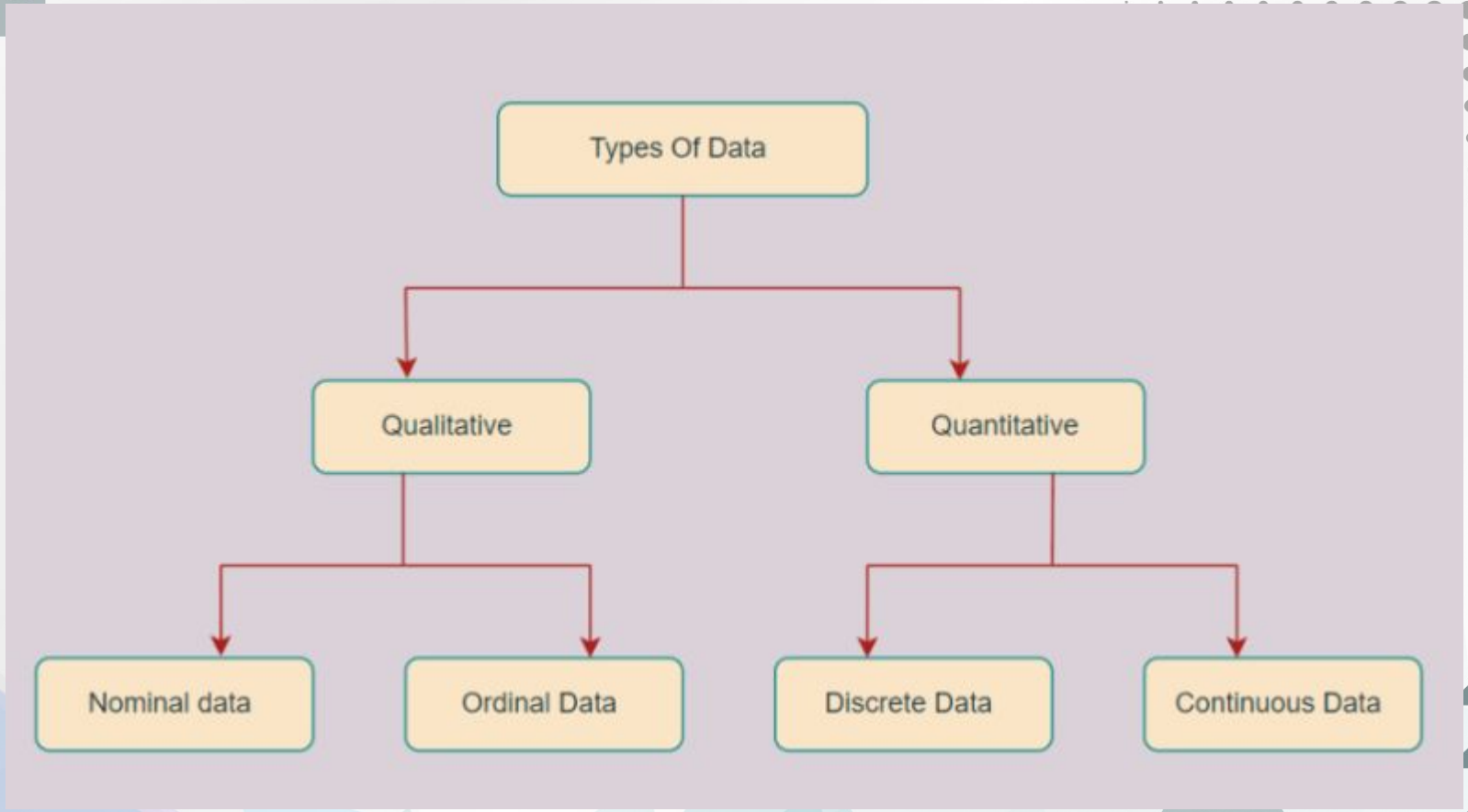
Types of Data

Types of Data

Today data is everywhere in every field. Whether you are a data scientist, marketer, businessman, data analyst, researcher, or you are in any other profession, you need to play or experiment with raw or structured data. This data is so important for us that it becomes important to handle and store it properly, without any error. While working on these data, it is important to know the types of data to process them and get the right results. **There are two types of data: Qualitative and Quantitative data** , which are further classified into**:**

The data is classified into four categories:

- Nominal data.
- Ordinal data.
- Discrete data.
- Continuous data.


Types Of Data
Qualitative
Quantitative
Nominal data
Ordinal Data
Discrete Data
Continuous Data

Qualitative Data

Qualitative or Categorical Data is data that can't be measured or counted in the form of numbers. These types of data are sorted by category, not by number. That's why it is also known as Categorical Data. These data consist of audio, images, symbols, or text. The gender of a person, i.e., male, female, or others, is qualitative data.

Qualitative data tells about the perception of people. This data helps market researchers understand the customers' tastes and then design their ideas and strategies accordingly.

The other examples of qualitative data are :

- What language do you speak
- Favorite holiday destination
- Opinion on something (agree, disagree, or neutral)
- Colors

Qualitative Data

Nominal Data

Nominal Data is used to label variables without any order or quantitative value. The color of hair can be considered nominal data, as one color can't be compared with another color.

The name "nominal" comes from the Latin name "nomen," which means "name." With the help of nominal data, we can't do any numerical tasks or can't give any order to sort the data. These data don't have any meaningful order; their values are distributed into distinct categories.

Examples of Nominal Data :

- Colour of hair (Blonde, red, Brown, Black, etc.)
- Marital status (Single, Widowed, Married)
- Nationality (Indian, German, American)
- Gender (Male, Female, Others)
- Eye Color (Black, Brown, etc.)

Qualitative Data

Ordinal Data

Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can't do any arithmetical tasks on them.

Ordinal data is qualitative data for which their values have some kind of relative position. These kinds of data can be considered "in-between" qualitative and quantitative data. The ordinal data only shows the sequences and cannot use for statistical analysis. Compared to nominal data, ordinal data have some kind of order that is not present in nominal data.

Examples of Ordinal Data :

- When companies ask for feedback, experience, or satisfaction on a scale of 1 to 10
- Letter grades in the exam (A, B, C, D, etc.)
- Ranking of people in a competition (First, Second, Third, etc.)
- Economic Status (High, Medium, and Low)
- Education Level (Higher, Secondary, Primary)

Quantitative Data

Quantitative data can be expressed in numerical values, making it countable and including statistical data analysis. These kinds of data are also known as Numerical data. It answers the questions like “how much,” “how many,” and “how often.” For example, the price of a phone, the computer’s ram, the height or weight of a person, etc., falls under quantitative data.

Quantitative data can be used for statistical manipulation. These data can be represented on a wide variety of graphs and charts, such as bar graphs, histograms, scatter plots, boxplots, pie charts, line graphs, etc.

Examples of Quantitative Data :

- Height or weight of a person or object
- Room Temperature
- Scores and Marks (Ex: 59, 80, 60, etc.)
- Time

Quantitative Data

Discrete Data

The term discrete means distinct or separate. The discrete data contain the values that fall under integers or whole numbers. The total number of students in a class is an example of discrete data. These data can't be broken into decimal or fraction values.

The discrete data are countable and have finite values; their subdivision is not possible. These data are represented mainly by a bar graph, number line, or frequency table.

Examples of Discrete Data :

- Total numbers of students present in a class
- Cost of a cell phone
- Numbers of employees in a company
- The total number of players who participated in a competition
- Days in a week

Quantitative Data

Continuous Data

Continuous data are in the form of fractional numbers. It can be the version of an android phone, the height of a person, the length of an object, etc. Continuous data represents information that can be divided into smaller levels. The continuous variable can take any value within a range.

The key difference between discrete and continuous data is that discrete data contains the integer or whole number. Still, continuous data stores the fractional numbers to record different types of data such as temperature, height, width, time, speed, etc.

Examples of Continuous Data :

- Height of a person
- Speed of a vehicle
- “Time-taken” to finish the work
- Wi-Fi Frequency
- Market share price

Python Libraries for DA

Libraries of the Python Ecosystem

- **NumPy**: Short for “Numerical Python”. Handles multidimensional arrays and matrices for efficient mathematical computations.
- **SciPy**: A scientific computing library for scientific, mathematical, and engineering operations.
- **Pandas**: Offers tabular data structures (like DataFrames) for data exploration and manipulation.
- **Scikit-Learn**: A machine learning library offering a variety of algorithms for tasks like regression, classification, and cluster analysis.

Libraries of the Python Ecosystem

- **Matplotlib**: Core library for data visualization in Python. Offers 2D and 3D plots, graphs, charts, and figures.
- **Seaborn**: Based on Matplotlib. Provides high-level, interactive, and organized plots.
- **Plotly**: Offers high-quality, interactive graphs like scatter charts, line charts, bar charts, histograms, boxplots, heatmaps, and subplots.

Data Sources

Data Source

A data source refers to the origin of a specific set of information. As businesses increasingly generate data year over year, data analysts rely on different data sources to measure business success and offer strategic recommendations. Having data literacy means you're capable of identifying, understanding, and interpreting crucial data and its results.

Data sources play a key role by bundling information into accessible formats, which enables seamless integrations between different types of systems. This ensures that relevant information about a data set is readily available while remaining hidden, allowing analysts to focus on data interpretation and analysis.

- **Internal Data** : Created by organizational processes, including marketing, customer profiles, and online activity.
- **External Data** : Derived from outside sources like social media, demographic data, and websites.
- **Third-Party Analytics** : Provided through analytics platforms like Google Analytics.
- **Open Data** : Free, publicly accessible data, like government and health and science data.
- **Machine Data** : Generated by users, stored in the input machine, and not easily shareable.
- **File Data** : Reside within single, shareable files, allowing multiple users to access and edit the data from different locations

Local Data Source (Open Data)

- Open Data Philippines : An online repository of open data from different government agencies in the Philippines.
- **Data Engineering Pilipinas** : Provides a list of government and public data, including the Philippine Standard Geographic Code (PSGC) and COVID-19 data from the Department of Health.
- **Philippine Statistics Authority (PSA)** : Offers various statistical datasets available for public use.
- PSA OpenSTAT : Provides access to various datasets on topics like population, migration, labor and employment.

Data Sets

Datasets are collections of raw data gathered during the research process usually in the form of numerical data. Many organizations, e.g. government agencies, universities or research institutions make the data they have collected freely available on the web for other researchers to use.

Data sets FILE FORMAT

Plain Text Formats

Plain Text Formats

a. CSV (Comma-Separated Values)

- **Description:** Stores tabular data in plain text, with values separated by commas.
- **Usage:** Common for sharing and analyzing data.
- **Advantages:** Easy to read and widely supported.
- **Limitations:** No support for metadata or complex structures.

b. TSV (Tab-Separated Values)

- **Description:** Similar to CSV, but values are separated by tabs.
- **Usage:** Preferred when data contains commas.

c. TXT (Plain Text)

- **Description:** Unstructured data stored as text.
- **Usage:** Logs, documentation, or datasets with no specific structure.

Spreadsheet Formats

.XLS/XLSX (Microsoft Excel)

- **Description:** Spreadsheet format supporting multiple sheets, formulas, and metadata.
- **Usage:** Used for data analysis, financial data, and business reports.
- **Advantages:** Easy to use and supports advanced formatting.
- **Limitations:** May not handle very large datasets efficiently.

b. ODS (OpenDocument Spreadsheet)

- **Description:** Open standard spreadsheet format.
- **Usage:** Alternative to XLS/XLSX for open-source applications.

Markup and Structured Data Formats

a. JSON (JavaScript Object Notation)

- **Description:** Lightweight, text-based format for structured data.
- **Usage:** APIs, web applications, and hierarchical datasets.
- **Advantages:** Human-readable and easy to parse.

b. XML (eXtensible Markup Language)

- **Description:** Text format for hierarchical data with custom tags.
- **Usage:** Data exchange in applications like SOAP-based web services.
- **Advantages:** Extensible and supports metadata.
- **Limitations:** Verbose compared to JSON.

Database and Binary Formats

SQL (Structured Query Language)

- **Description:** Relational database format storing structured data in tables.
- **Usage:** Large-scale databases managed by systems like MySQL, PostgreSQL.

b. SQLite (.sqlite/.db)

- **Description:** Self-contained database format for small-scale applications.
- **Usage:** Mobile apps, embedded systems, and prototyping.

Tools for Analytics


Tableau
SQL
SQL
Datapine
Apache Spark
Microsoft Power BI
Airtable
ClicData
SAS
Qlik
Thoughtspot

Advanced Data Analytic Tools

Current and future tools

https://docs.google.com/document/d/1kwiluGQ1u-hgkgbjRom6Fsi2-X4QVifjHQGmZUfkDvw/edit?usp=sharing

DATA FRAMES

What is a DataFrame ?

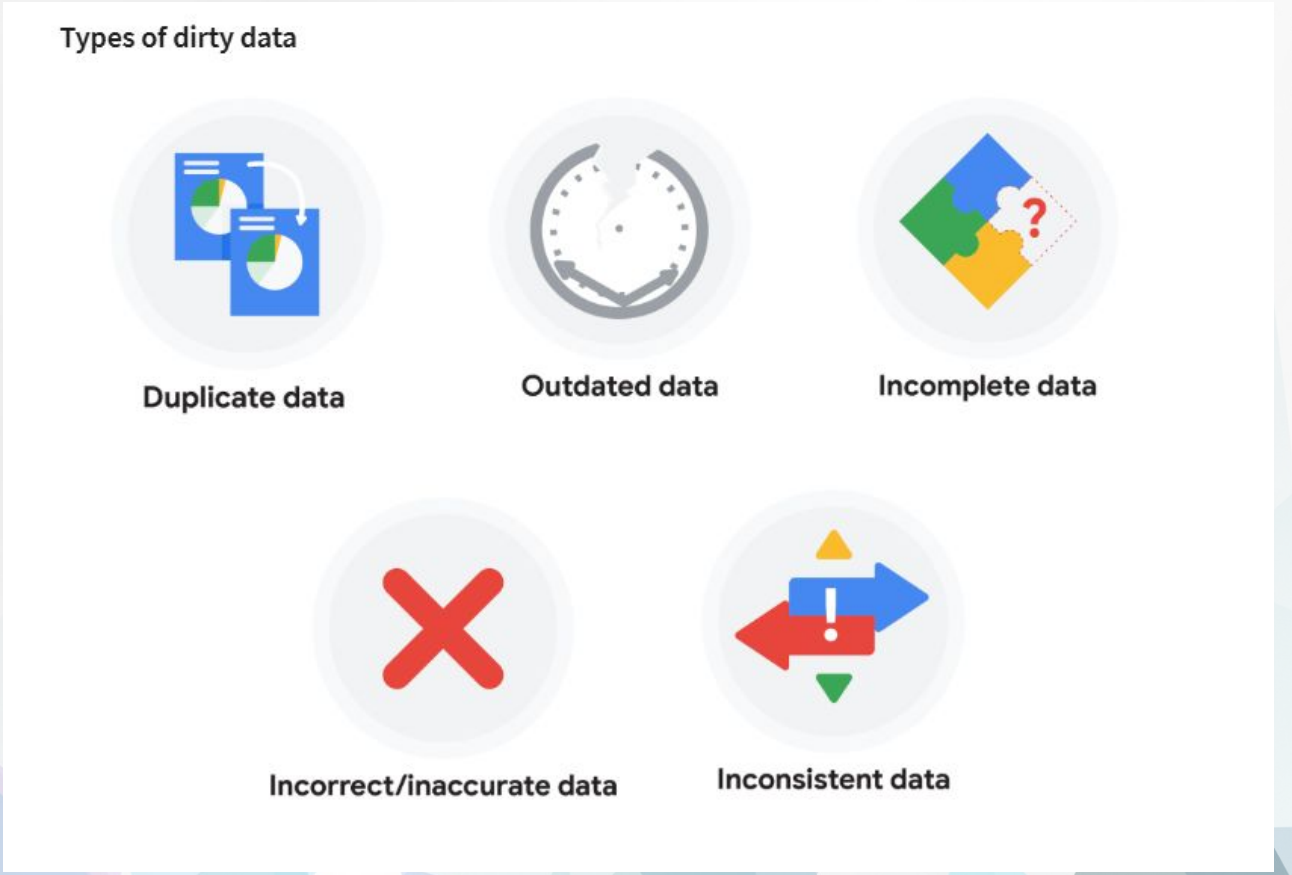
A **DataFrame** is a two-dimensional, tabular data structure commonly used in data analysis and manipulation. It is a core component of data analysis libraries such as **Pandas** in Python and **R's data.frame**. A DataFrame is organized into rows and columns, similar to a spreadsheet or SQL table.

Key Features of a DataFrame:

- **Rows and Columns:** Data is stored in rows (observations) and columns (variables).
- **Labeled Axes:** Rows and columns have labels (indices), making it easier to access and manipulate data.
- **Heterogeneous Data:** Columns can store different data types (e.g., integers, floats, strings).
- **In-Memory:** DataFrames typically operate in-memory, making them fast for interactive data analysis.

Cleaning a Data

Types of Dirty Data



Duplicate Data

Description	Possible causes	Potential harm to businesses
Any data record that shows up more than once	Manual data entry, batch data imports, or data migration	Skewed metrics or analyses, inflated or inaccurate counts or predictions, or confusion during data retrieval

Outdated Data

Description	Possible causes	Potential harm to businesses
Any data that is old which should be replaced with newer and more accurate information	People changing roles or companies, or software and systems becoming obsolete	Inaccurate insights, decision-making, and analytics

Incomplete Data

Description	Possible causes	Potential harm to businesses
Any data that is missing important fields	Improper data collection or incorrect data entry	Decreased productivity, inaccurate insights, or inability to complete essential services

Inaccurate / Incorrect Data

Description	Possible causes	Potential harm to businesses
Any data that is complete but inaccurate	Human error inserted during data input, fake information, or mock data	Inaccurate insights or decision-making based on bad information resulting in revenue loss

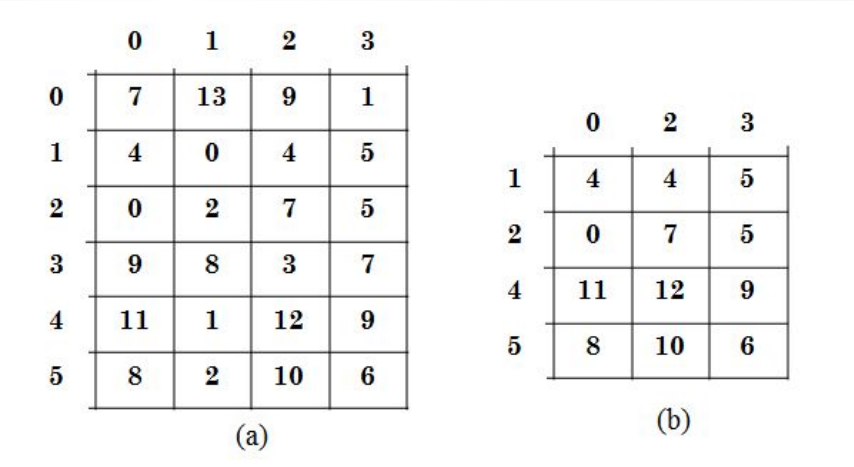
Inconsistent Data

Description	Possible causes	Potential harm to businesses
Any data that uses different formats to represent the same thing	Data stored incorrectly or errors inserted during data transfer	Contradictory data points leading to confusion or inability to classify or segment customers

Fundamentals of DATA ANALYTICS

Matrix

In data analytics, a ***matrix*** refers to a two-dimensional data structure with rows and columns. It is often used to represent a dataset or a specific subset of data for analysis. Matrices are fundamental in various analytical techniques and operations, such as linear algebra, matrix operations, and statistical calculations.



	0	1	2	3
0	7	13	9	1
1	4	0	4	5
2	0	2	7	5
3	9	8	3	7
4	11	1	12	9
5	8	2	10	6

(a)

	0	2	3
1	4	4	5
2	0	7	5
4	11	12	9
5	8	10	6

(b)

Matrices

Converting data into matrices is a necessary step in data mining as it enables

- efficient computation,
- compatibility with algorithms and libraries,
- application of feature engineering techniques,
- execution of mathematical operations,
- optimization of performance, and
- simplified representation and visualization.

Matrices provide a standardized format that is widely supported and optimized for various data mining tasks, making them a valuable tool in the data mining process.

Libraries for Data Analytics

Introduction to NUMPY

Objectives

- Understand Numpy Arrays and List
- Numpy array numerical data types
- Manipulating array shapes
- The stacking of Numpy arrays
- Partitioning NumPy arrays
- Changing the data type of Numpy arrays
- Creating Numpy views and copies
- Slicing Numpy arrays
- Boolean and fancy indexing
- Broadcasting arrays

Numpy

Short for Numerical Python.

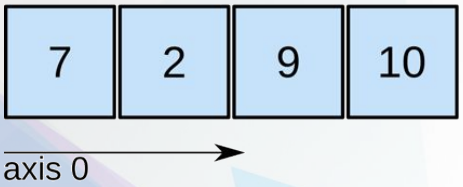
Numpy is the foundational Library for scientific computing in Python since it provides data structures and high-performing functions that the basic package in Python Cannot provide.

Features .:

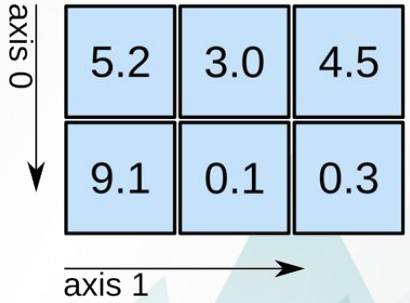
- Ndarray
- Reading Data Sets
- Element Wise Computation
- Integration with Other languages

Numpy Array

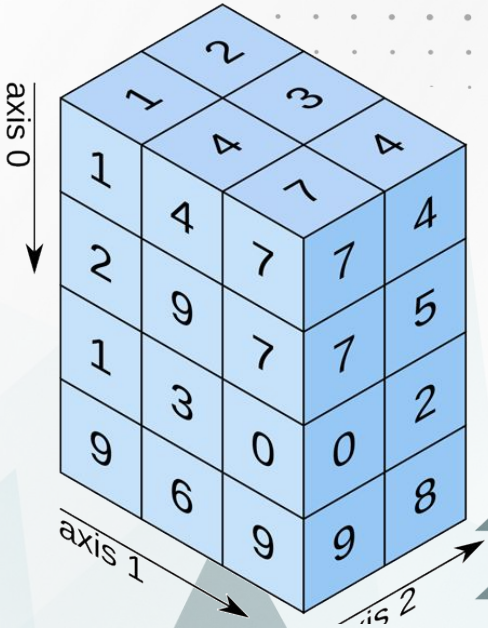
1D array



2D array



3D array



Basic Numpy Functions for Matrices

1. ***Reshape:*** The reshape function allows you to change the shape or dimensions of a matrix without altering its data. For example, you can convert a 1D array into a 2D matrix or vice versa by specifying the desired shape.
2. ***Transpose:*** The transpose function or the .T attribute transposes a matrix, interchanging its rows and columns. This operation flips the matrix along its main diagonal.
3. ***Concatenate:*** The concatenate function allows you to concatenate or join multiple matrices along a specified axis. It can be used to combine matrices vertically (axis 0) or horizontally (axis 1).
4. ***Split:*** The split function splits a matrix into multiple smaller matrices along a specified axis. It divides the matrix into equal-sized sections or based on predefined indices.
5. ***Resize:*** The resize function resizes a matrix to a specified shape. It can add or remove elements from the matrix to match the desired shape.

Basic Numpy Functions for Matrices

6. ***Flatten*** : The flatten function returns a 1D array by collapsing a matrix into a single row, concatenating its rows one after another.
7. ***Diagonal*** : The diagonal function extracts the diagonal elements of a matrix and returns them as a 1D array. It can also be used to extract or modify elements along any other specified diagonal.
8. ***Fill***: The fill function fills a matrix with a specified value, replacing all the existing elements with the given value.
9. ***Delete***: The delete function removes a specified row or column from a matrix, returning a new matrix with the selected row or column removed.
10. ***Insert***: The insert function inserts a specified value, row, or column into a matrix at a specified position, returning a new matrix with the new elements added.

DASHBOARDS

The beauty of dashboards

The following table summarizes the benefits of using a dashboard for both data analysts and their stakeholders.

Benefits	For data analysts	For stakeholders
Centralization	Share a single source of data with all stakeholders	Work with a comprehensive view of data, initiatives, objectives, projects, processes, and more
Visualization	Show and update live, incoming data in real time*	Spot changing trends and patterns more quickly
Insightfulness	Pull relevant information from different datasets	Understand the story behind the numbers to keep track of goals and make data-driven decisions
Customization	Create custom views dedicated to a specific person, project, or presentation of the data	Drill down to more specific areas of specialized interest or concern

Types of dashboards

For a refresher, consider the different types of dashboards a business may use. Often, businesses will tailor a dashboard for a specific purpose. The three most common categories are:

- **Strategic**: focuses on long term goals and strategies at the highest level of metrics
- **Operational:** short-term performance tracking and intermediate goals
- **Analytical:** consists of the datasets and the mathematics used in these sets