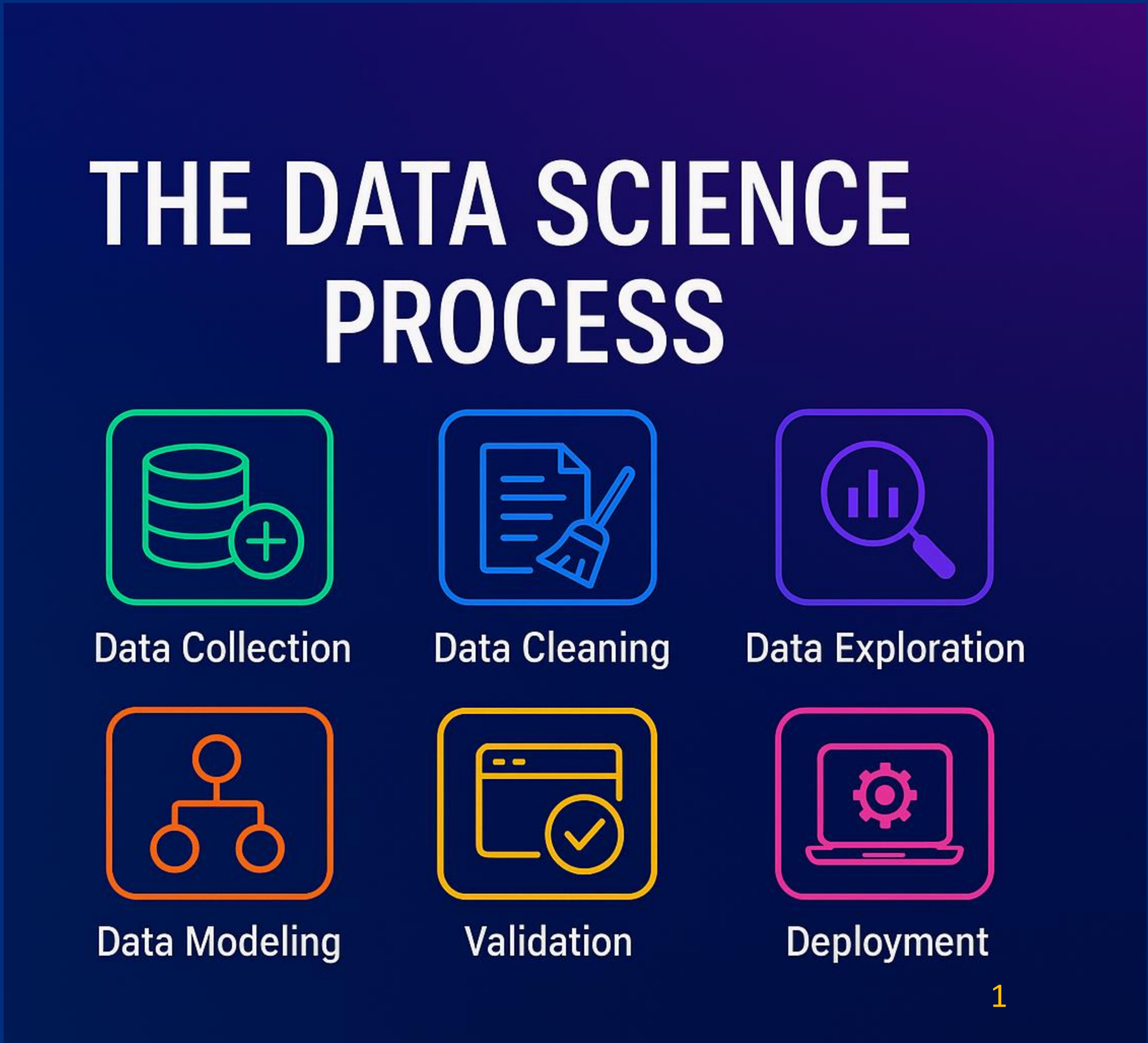
Chapter 2

The Data Science Process

Prepared by :
Dr. Bilal Dendani

جامعة باجي مختار - عنابة
BADJI MOKHTAR - ANNABA UNIVERSITY



Chapter 2. The Data Science Process

- Roles and responsibilities in a Data Science project
- Overview of the Data Science project life cycle
- **Step 1:** Define research objectives and create a project charter
- **Step 2:** Data collection
- **Step 3:** Data cleaning, integration, and transformation
- **Step 4:** Exploratory Data Analysis (EDA)
- **Step 5:** Model building
- **Step 6:** Presenting results and developing applications on top of them

2.1 Data science process

- The data science process consists of a series of systematic steps, starting from the definition of objectives to the presentation of results.

- This structured sequence of stages guides us in transforming raw data into actionable insights.

- A well-organized process ensures the quality, consistency, and reliability of the results obtained.

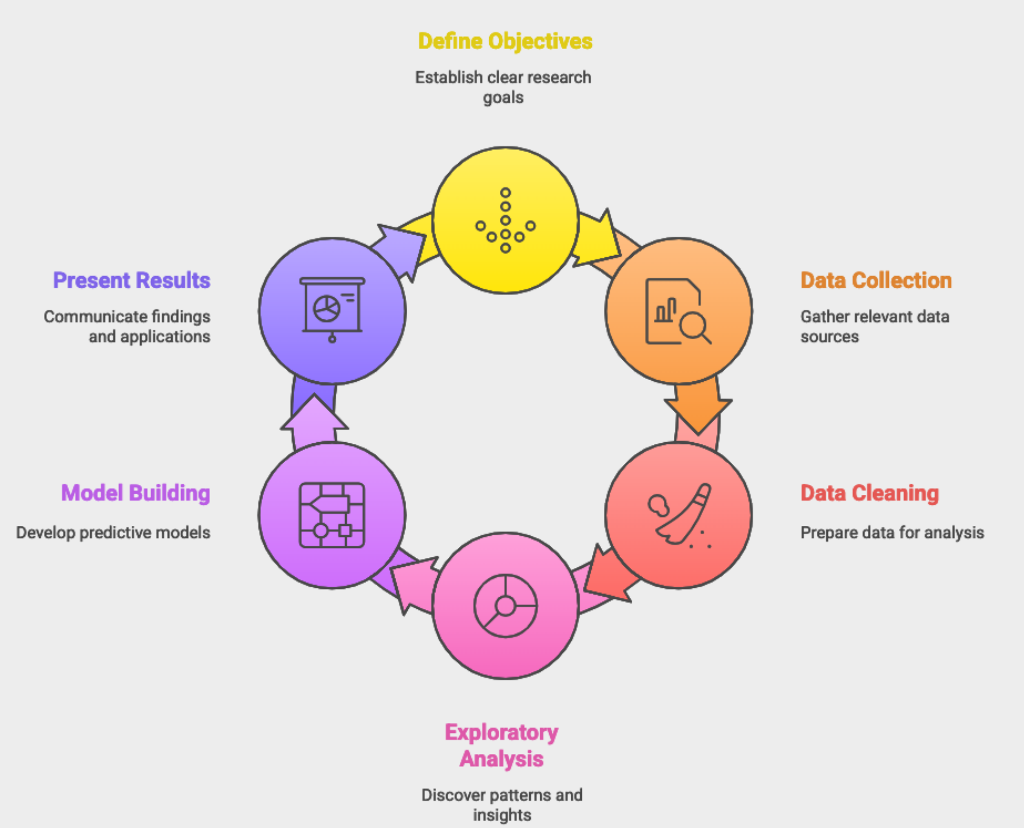

Define Objectives
Establish clear research goals
Data Collection
Gather relevant data sources
Data Cleaning
Prepare data for analysis
Exploratory Analysis
Discover patterns and insights
Model Building
Develop predictive models
Present Results
Communicate findings and applications

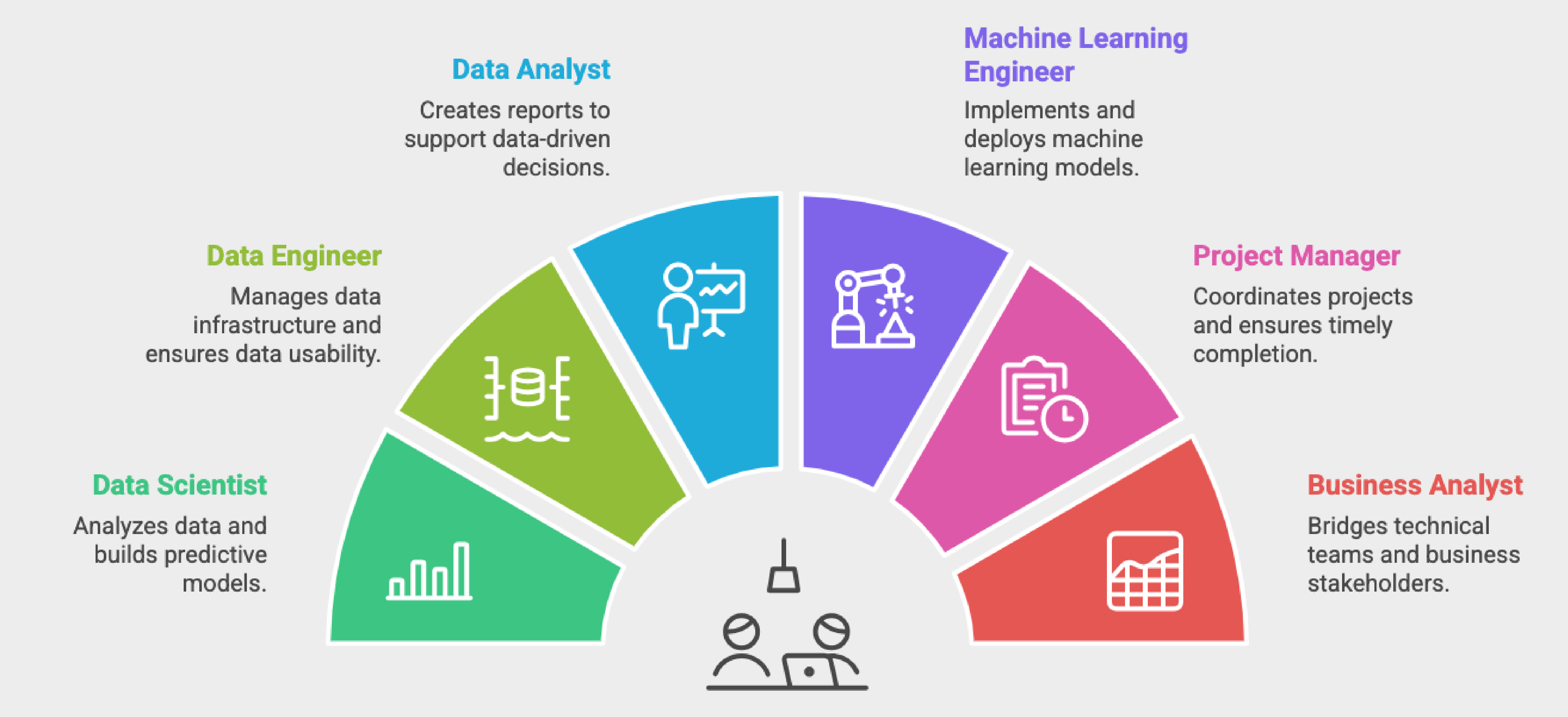
Figure 25. Data science process steps

2.2 Roles and Responsibilities in a Data Science Project

- Data Scientist: Responsible for data analysis, building predictive models, and extracting insights from data.
- Data Engineer: Manages the data infrastructure, collects, cleans, and transforms data to make it usable.
- Data Analyst: Analyzes data and creates reports to help stakeholders make data-driven decisions.
- Machine Learning Engineer: Implements and deploys models into production environments.
- Business Analyst: Acts as the bridge between the technical data science teams and business stakeholders. Their role is to understand business needs, translate them into analytical questions, and ensure that data analysis results align with the organization's objectives.
- Project Manager: Ensures project coordination and management, sets objectives, and monitors timelines and resources.

365
DATA SCIENCE JOBS
EXPLAINED

Data Science Roles



Data Scientist

Analyzes data and builds predictive models.

Data Engineer

Manages data infrastructure and ensures data usability.

Data Analyst

Creates reports to support data-driven decisions.

Machine Learning Engineer

Implements and deploys machine learning models.

Project Manager

Coordinates projects and ensures timely completion.

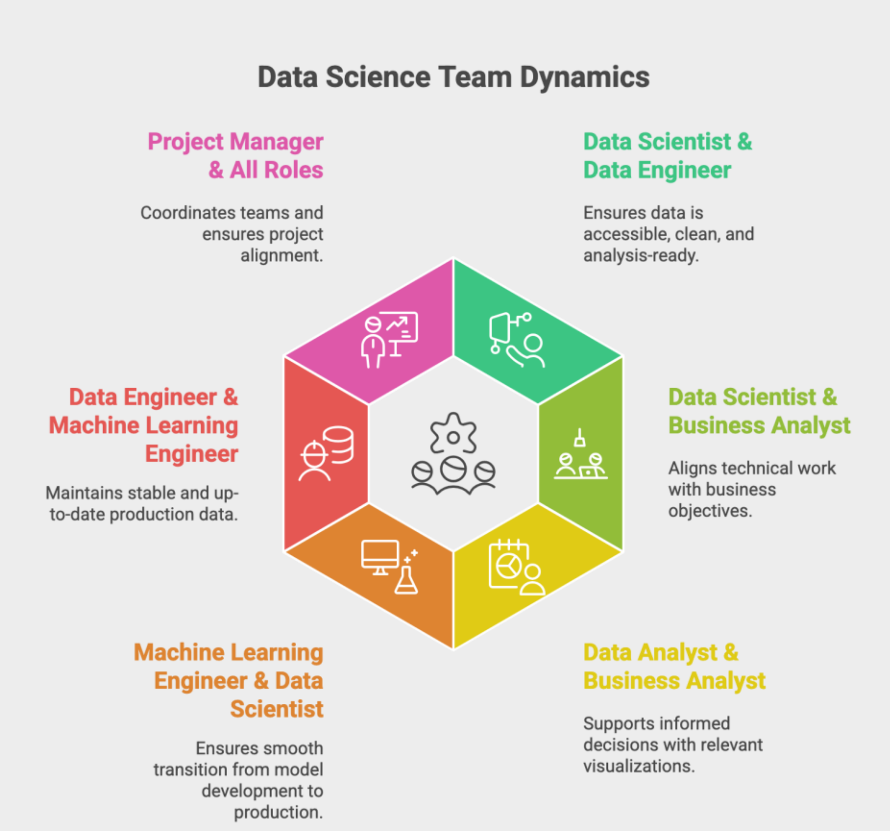
Business Analyst

Bridges technical teams and business stakeholders.

2.3 Collaboration Between Roles in a Data Science Project

In a data science project, each role has specific responsibilities, and close collaboration among them is essential to ensure project success. Below is an overview of typical interactions and why collaboration is so important:

1. **Data Scientist & Data Engineer**
2. **Data Scientist & Business Analyst**
3. **Data Analyst & Business Analyst**
4. **Machine Learning Engineer & Data Scientist**
5. **Data Engineer & Machine Learning Engineer**
6. **Project Manager & All Roles**



Data Science Team Dynamics

Project Manager & All Roles

Coordinates teams and ensures project alignment.

Data Scientist & Data Engineer

Ensures data is accessible, clean, and analysis-ready.

Data Engineer & Machine Learning Engineer

Maintains stable and up-to-date production data.

Data Scientist & Business Analyst

Aligns technical work with business objectives.

Machine Learning Engineer & Data Scientist

Ensures smooth transition from model development to production.

Data Analyst & Business Analyst

Supports informed decisions with relevant visualizations.

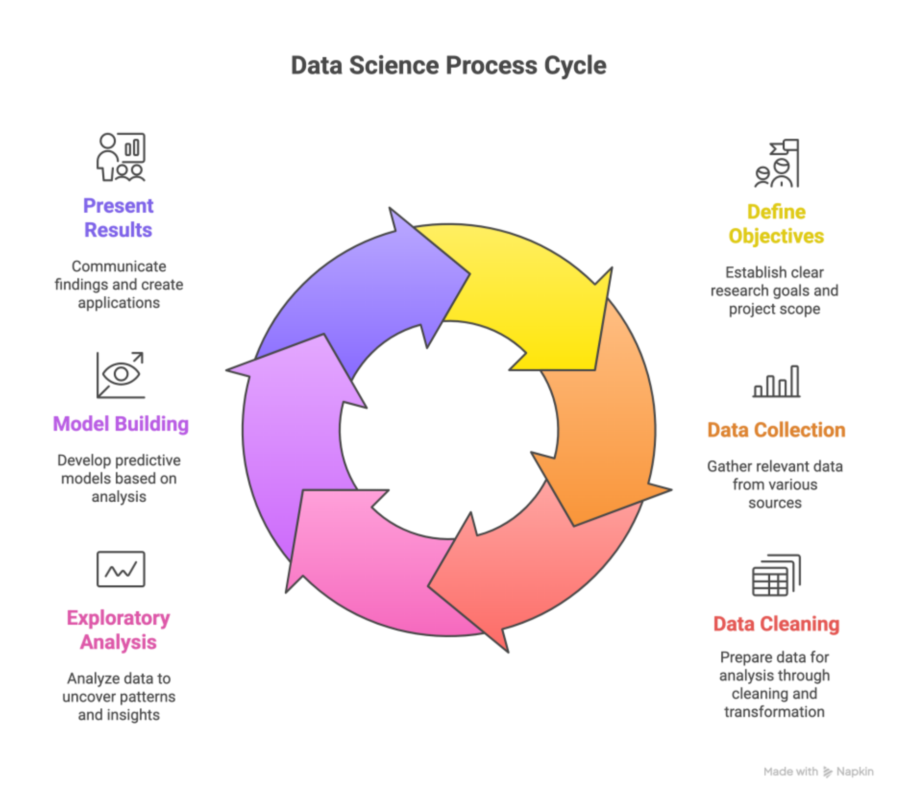
2.4 Lifecycle of a Data Science Project

The Data Science Project Life Cycle is a structured sequence of stages that guide the process of turning raw data into valuable insights and actionable solutions.

- It helps ensure clarity, consistency, and quality throughout the project, from problem definition to deployment.

Main Stages:

1. Define research objectives and create a project charter
2. Data collection
3. Data cleaning, integration, and transformation
4. Exploratory Data Analysis (EDA)
5. Model building
6. Presenting results and developing applications on top of them


Data Science Process Cycle
Present Results
Communicate findings and create applications
Define Objectives
Establish clear research goals and project scope
Model Building
Develop predictive models based on analysis
Data Collection
Gather relevant data from various sources
Exploratory Analysis
Analyze data to uncover patterns and insights
Data Cleaning
Prepare data for analysis through cleaning and transformation
Made with Napkin

2.4 Key Stages of the Data Science Project Life Cycle

1. Define Objectives and Create a Project Charter:
   Identify the problem to be solved and clearly define the project goals.
2. Data Collection:
   Identify and gather relevant data from both internal and external sources.
3. Data Cleaning and Transformation:
   Prepare the data by cleaning, integrating, and transforming it into a usable format.
4. Exploratory Data Analysis (EDA):
   Understand data distributions and relationships to guide future analysis.
5. Modeling and Model Building:
   Apply algorithms to build predictive or analytical models.
6. Presenting Results:
   Communicate insights and propose concrete actions based on model outcomes.

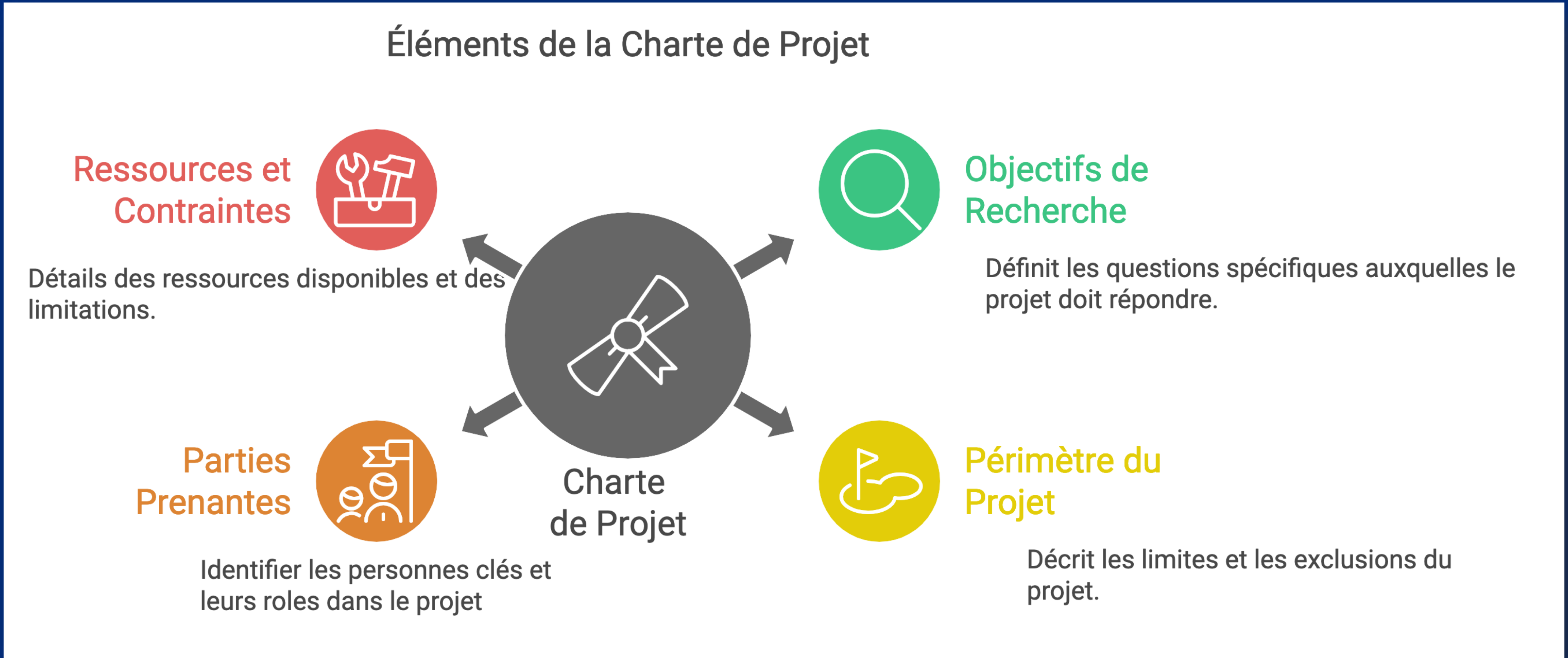
2.4.1 Step 1: Define Research Objectives and Create a Project Charter

- 1. Purpose of the Step

    - Define the Project Direction: Clearly identify what the team aims to achieve.
    - Stakeholder Alignment: Ensure that everyone involved shares a common understanding of the project’s objectives.

- 2. Developing the Project Charter
    - Establish the project’s scope, goals, timeline, and success criteria.
    - Identify key stakeholders, team roles, and responsibilities.
    - Outline the resources, risks, and expected deliverables.

Elements of the project charter



Éléments de la Charte de Projet

Ressources et Contraintes

Détails des ressources disponibles et des limitations.

Objectifs de Recherche

Définit les questions spécifiques auxquelles le projet doit répondre.

Charte de Projet

Parties Prenantes

Identifier les personnes clés et leurs roles dans le projet

Périmètre du Projet

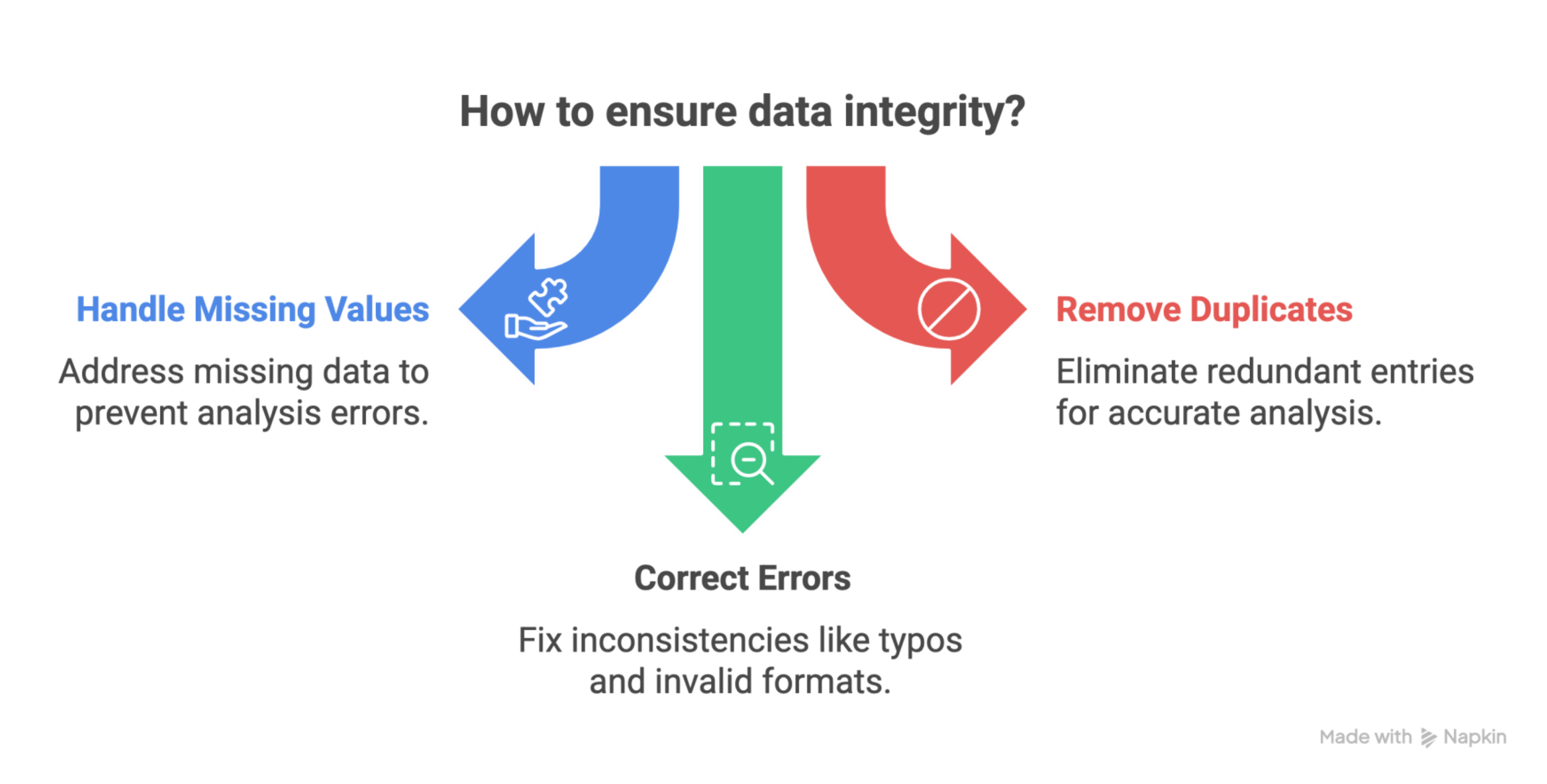
Décrit les limites et les exclusions du projet.

2.4.2 Step 2 : Data collection

- Collect the data required to achieve the objectives defined in the project charter. Data can come from different types and sources.
- **Data Sources**
  - Internal Data: Company databases (e.g., customer history, sales records).
  - External Data: Public datasets, social media platforms, third-party APIs.
  - IoT Devices: Data generated by sensors and connected devices (e.g., logistics tracking, environmental data).
- **Types of Collected Data**
  - Structured Data: Tables, relational databases.
  - Semi-Structured Data: JSON or XML files.
  - Unstructured Data: Images, videos, textual documents.

2.4.3 Step 3: Data Cleaning, Integration, and Transformation

The goal of this step is to ensure that the data is **accurate**, **complete**, **and ready for analysis**. As data scientists, we can only build reliable models if our data foundation is solid, so this phase is crucial in the data science process.



How to ensure data integrity?

Handle Missing Values

Address missing data to prevent analysis errors.

Correct Errors

Fix inconsistencies like typos and invalid formats.

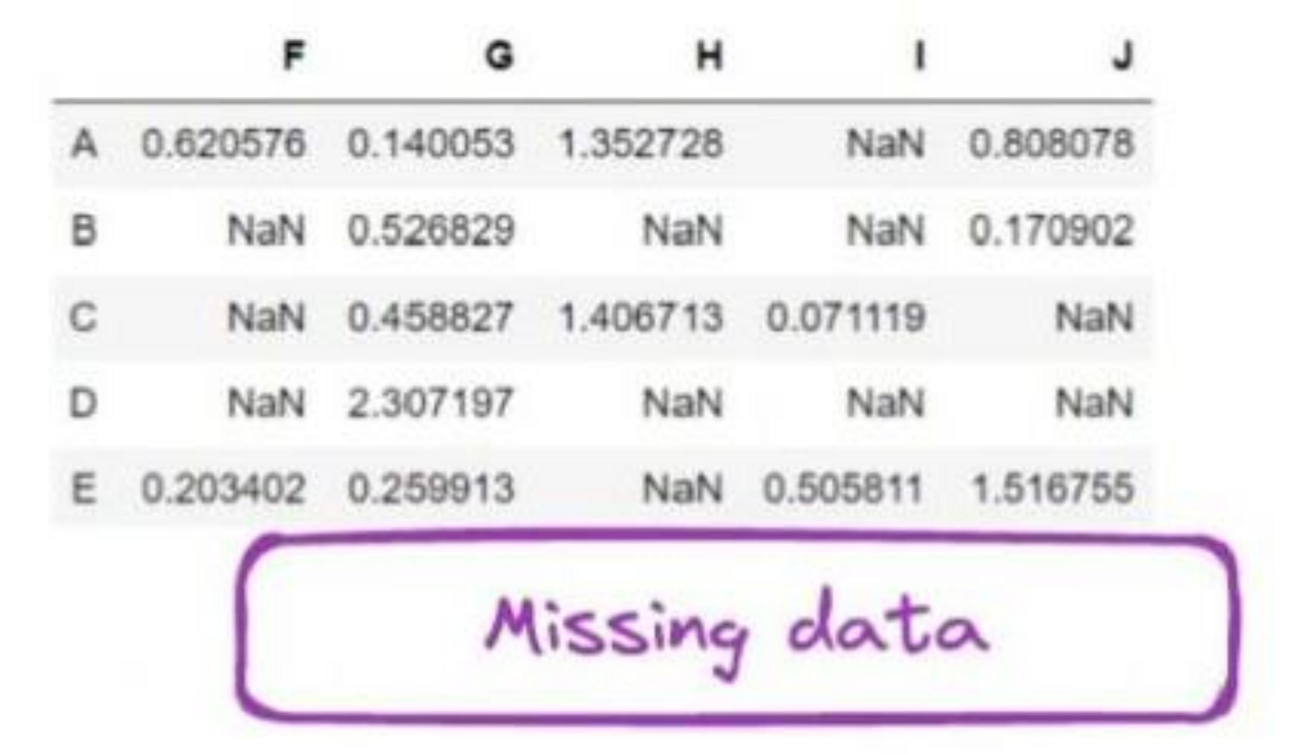
Remove Duplicates

Eliminate redundant entries for accurate analysis.

2.4.3 Step3: Data Cleaning, Integration, and Transformation

- Data Cleaning

- Identify and handle **missing values** and **remove duplicates**.
- Correct **errors and inconsistencies** such as typos, incorrect formats, or invalid entries.
- Ensure data integrity across all fields before moving forward.



	F	G	H	I	J
A	0.620576	0.140053	1.352728	NaN	0.808078
B	NaN	0.526829	NaN	NaN	0.170902
C	NaN	0.458827	1.406713	0.071119	NaN
D	NaN	2.307197	NaN	NaN	NaN
E	0.203402	0.259913	NaN	0.505811	1.516755

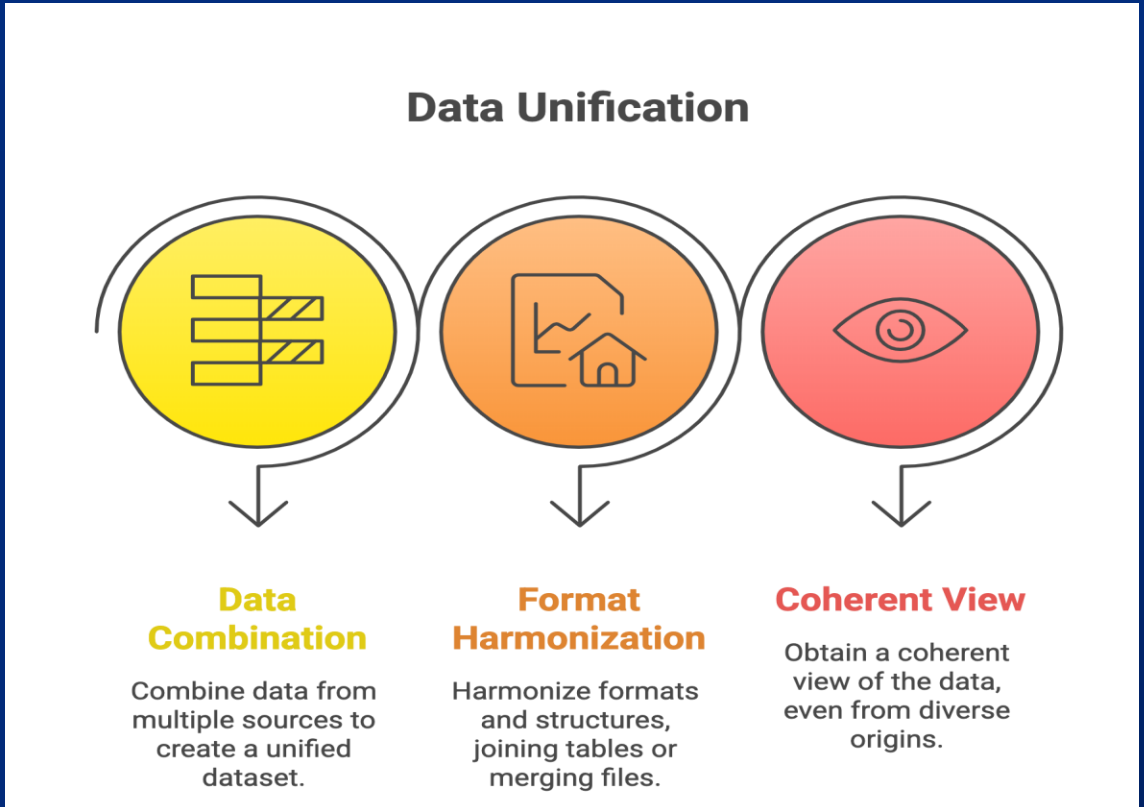
Missing data

2.4.3 Step3: Data Cleaning, Integration, and Transformation

The goal of this step is to ensure that the data is **accurate, complete, and ready for analysis**. As data scientists, we can only build reliable models if our data foundation is solid so this phase is crucial in the data science process.

Data Integration

- Combine data from multiple sources to create a unified dataset.
- Harmonize formats and structures for example, joining tables or merging files from different systems.
- The goal is to obtain a coherent view of the data, even when it comes from diverse origins.



Data Unification

Data Combination

Combine data from multiple sources to create a unified dataset.

Format Harmonization

Harmonize formats and structures, joining tables or merging files.

Coherent View

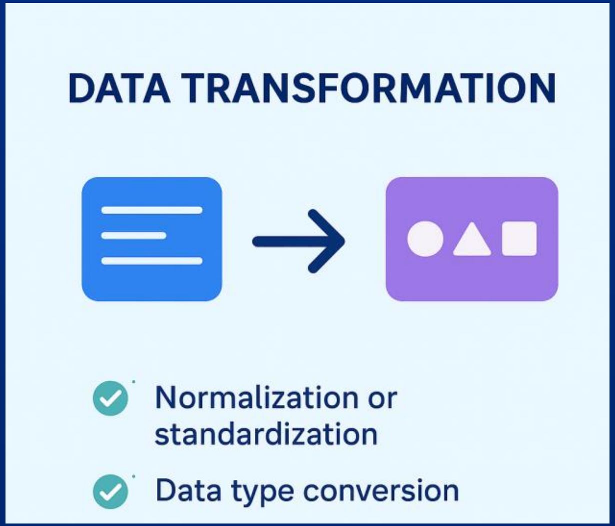
Obtain a coherent view of the data, even from diverse origins.

2.4.3 Step3: Data Cleaning, Integration, and Transformation

The goal of this step is to ensure that the data is **accurate, complete, and ready for analysis**. As data scientists, we can only build reliable models if our data foundation is solid so this phase is crucial in the data science process.

Data Transformation

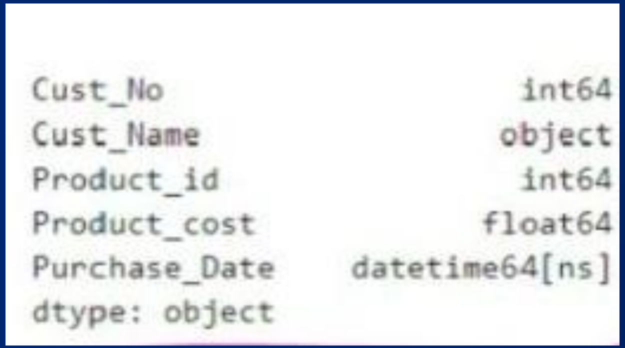
- Normalize or standardize numeric values to bring them onto comparable scales.
- Convert data types when necessary (e.g., turning string dates into proper date formats).
- This step prepares your dataset for statistical analysis and machine learning modeling.



DATA TRANSFORMATION

✓ Normalization or standardization

✓ Data type conversion



```
Cust_No                    int64
Cust_Name                 object
Product_id                 int64
Product_cost             float64
Purchase_Date     datetime64[ns]
dtype: object
```

2.4.4 Step4: Exploratory Data Analysis (EDA)

EDA is a fundamental step in the data science process, used to understand the structure and characteristics (features) of the data and to identify patterns or anomalies before applying modeling techniques.

Main Objectives of EDA

🔍 **Understand the Data**: Identify the main features, structure, and distribution of variables.

📈 **Detect Patterns and Trends**: Reveal hidden relationships, groupings, or time-based trends.

⚠️ **Identify Outliers and Anomalies**: Detect data points that deviate from expected patterns.

🔗 **Evaluate Data Quality**: Spot missing values, inconsistencies, or potential data entry errors.

🎯 **Guide Further Analysis**: Provide insights that inform model selection and feature engineering.

2.4.4 Step4: EDA Techniques and Methods

1. Descriptive Statistics

- Descriptive statistics are tools we use to **summarize and describe** the main characteristics (features) of a dataset. Instead of looking at thousands of rows of raw data, we use statistics to quickly understand what the data looks like.

a. Measures of Central Tendency

These tell us where the “center” of the data lies.

Mean (Average): The sum of all values divided by their number.

Median: The middle value when data are sorted.

Mode: The most frequent value.

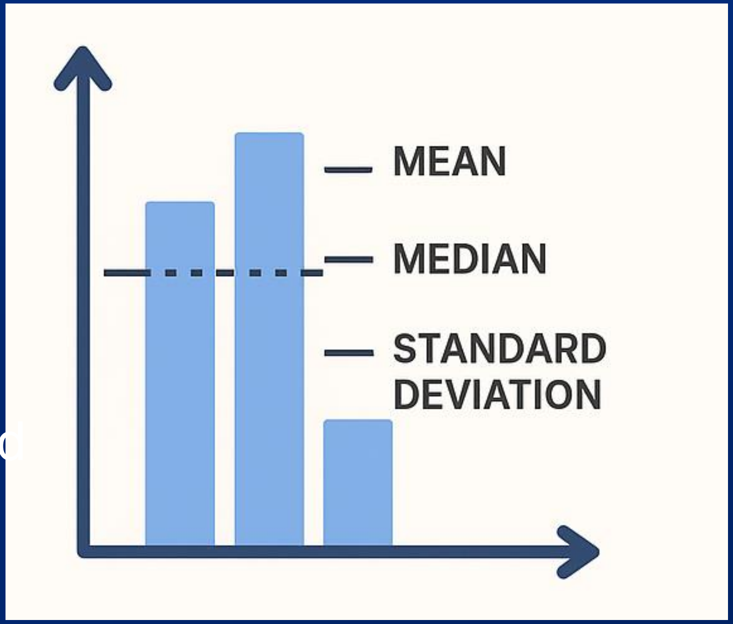
b. Measures of Dispersion (Spread)

These show how much the data varies.

Range: Difference between maximum and minimum.

Variance and Standard Deviation: Indicate how far values are spread from the mean.

Interquartile Range (IQR): Spread of the middle 50% of data.



2.4.4 Step4: EDA Techniques and Methods

1. Descriptive Statistics

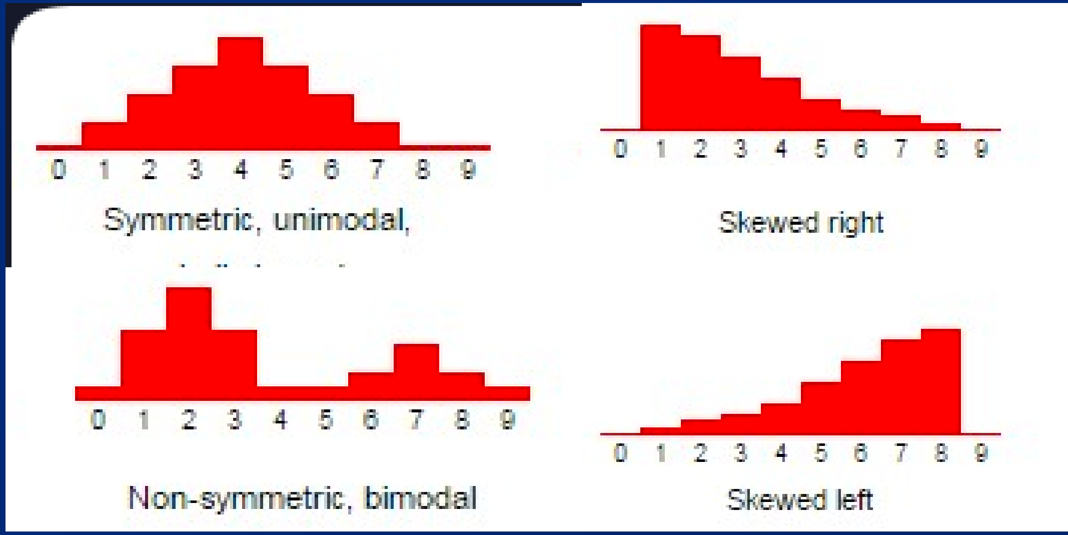
- **c. Shape of the Distribution**

Helps understand how the data are distributed.

Skewness: Is the data symmetric or shifted to one side?

Kurtosis: Are there many outliers or is flat/peaked?

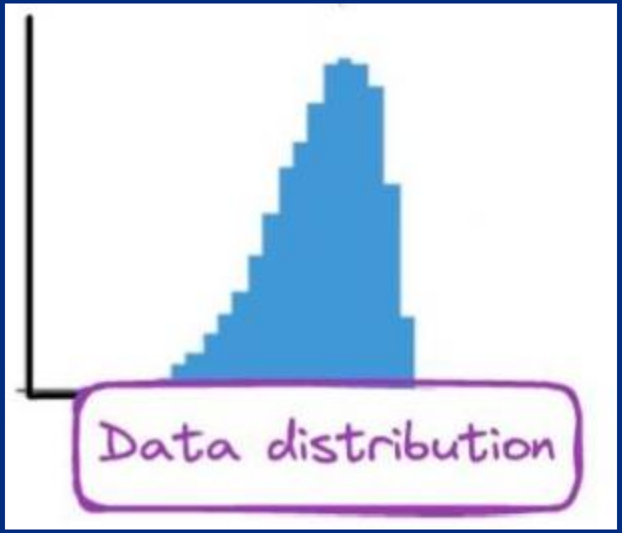
Visualize this with **histograms** or **box plots**.

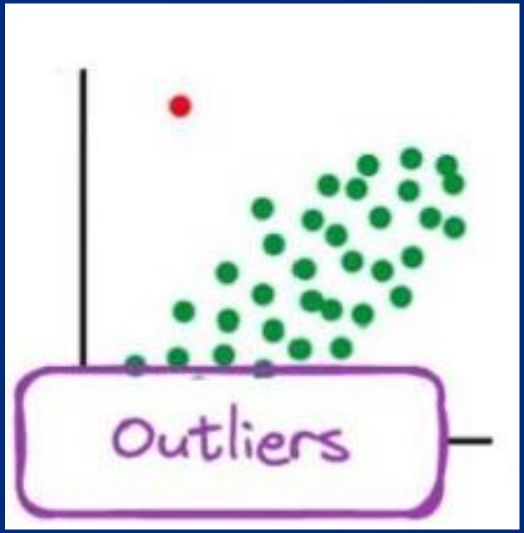


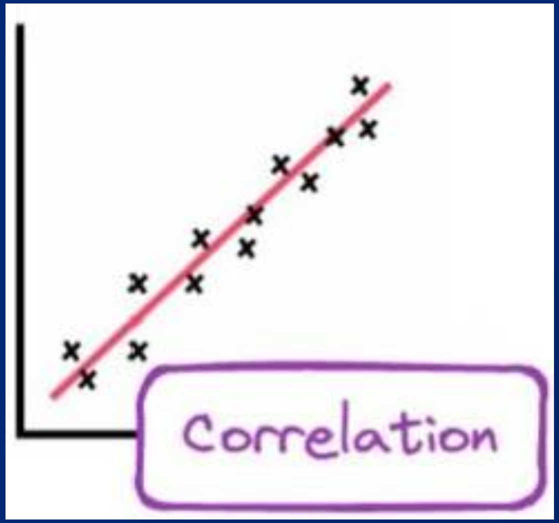
2.4.4 Step4: EDA Techniques and Methods

2. Visualization Techniques:

- **Histogram:** To understand data distribution
- **Box Plot:** To detect outliers and variability
- **Scatter Plot:** To study relationships between two variables
- **Pair Plot / Heatmap:** To observe global relationships







2.4.4 Step4: EDA Techniques and Methods

- 3. Correlation and Relationships:
  - Compute the **correlation matrix** to examine linear associations
  - Discuss examples where variables are strongly correlated (e.g., height vs weight)

4. Tools Commonly Used:

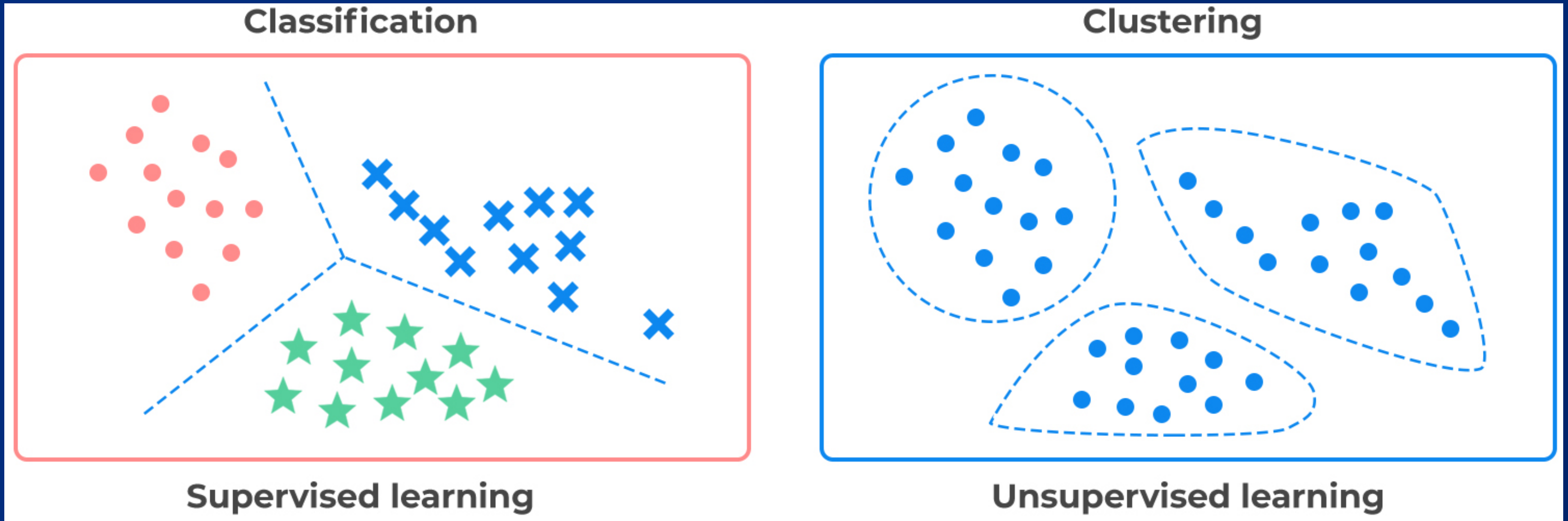
- Python libraries: pandas, matplotlib, seaborn, plotly
  Notebook environments: Jupyter, Google Colab

2.4.5 Step 5: Model Building

- This step involves using data to create models that help us answer the questions defined at the beginning of the project.

5.1 Types of Models:

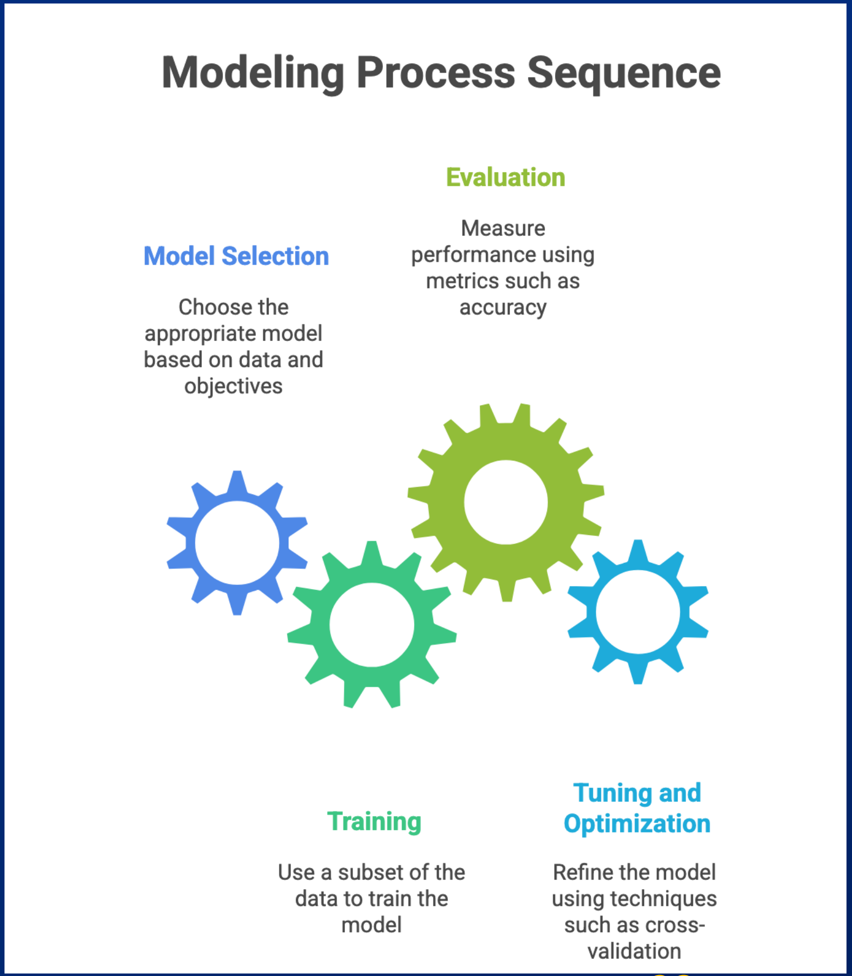
- **Supervised Models:** Regression, Classification (e.g., predicting a house price or classifying a customer).
- **Unsupervised Models:** Clustering, Dimensionality Reduction (e.g., segmenting customers by purchasing behavior).



2.4.6 Step5: Building the model: modelling process

5.2 Modeling Process:

- **Model Selection**: Choose the appropriate model based on the data and objectives (e.g., regression to predict continuous values, classification for descret values).
- **Training**: Use a subset of the data to train the model.
- **Evaluation**: Measure performance using metrics such as accuracy, recall, or RMSE.
- **Tuning and Optimization:** Refine the model using techniques such as cross-validation or hyperparameter tuning.



Modeling Process Sequence

Model Selection

Choose the appropriate model based on data and objectives

Evaluation

Measure performance using metrics such as accuracy

Training

Use a subset of the data to train the model

Tuning and Optimization

Refine the model using techniques such as cross-validation

2.4.6 Step 6 : Presenting Results and Building Applications

- This is the **final step** of the data science process. At this stage, data scientists must communicate their insights effectively and sometimes **turn those insights into practical applications** that create real value for the organization
- 🧭 **Goal**
  - To present the insights obtained from the analysis in a **clear, understandable, and actionable** way for all stakeholders — even for those without technical backgrounds
- 🛠️📊 **Data Visualization**
  - We use **charts**, **tables**, **and** **dashboards** to illustrate patterns, trends, and key findings.
  - Visual storytelling is a crucial skill that helps transform complex data into a message that is both **meaningful and persuasive**.

2.4.6 Step 6 : Presenting Results and Building Applications

- **Common Tools**
  - **Tableau** and **Power BI** for interactive dashboards and business reporting
  - **Matplotlib** and **Seaborn** in Python for customizable data visualizations
  - **Plotly** or **Dash** for creating web-based interactive applications