



# Training Outline

**Course:** Foundations of Data Analysis

**Week 1 Theme:** *From Raw Data to First Insights*

**Target Audience:** Beginners / Mixed-background  
learners

**Delivery Mode:** Lecture + Live Demo + Hands-on Lab

# Learning Outcomes

**By the end of Week 1, learners will be able to:**

1. Explain what data analysis is and why it matters
2. Distinguish between different types and formats of data
3. Identify common data quality issues
4. Use Jupyter / Colab to load a CSV dataset
5. Perform basic data inspection and descriptive statistics using Python
6. Extract simple, meaningful insights from a real dataset

# Concept Covered

- What is Data Analysis?
- Why It Matters in Decision-Making
- Types of Data: Structured vs. Unstructured
- Data Quality & Common Issues
- Introduction to Python & Jupyter
- Descriptive Statistics
- Hands-on Lab: Exploring Student Performance Data

# What Is Data Analysis?

*Data analysis is the process of inspecting, cleaning, transforming, and modeling data to discover useful information, draw conclusions, and support decision-making.*

- **Main goal:** Convert raw data into meaningful insights
- It goes beyond reporting - it involves asking *why* and *what if*

## The Data-Driven Decision Cycle

1. **Ask** – Clearly define the problem or question
2. **Collect** – Gather relevant data
3. **Clean** – Address errors and missing values
4. **Analyze** – Summarize, visualize, and model the data
5. **Act** – Make informed decisions and communicate results

*Example:* Universities analyze student data to reduce dropout rates.

# Real-World Impact of Data Analysis

- .  **Healthcare:** Predicting disease outbreaks
- .  **Finance:** Real-time fraud detection
- .  **Agriculture:** Crop yield forecasting using satellite data
- .  **Education:** Early identification of at-risk students

In Kenya, data literacy is a **national priority** under the Digital Economy Blueprint.

# Data Types, Formats and Structures

There are three different types of data structures

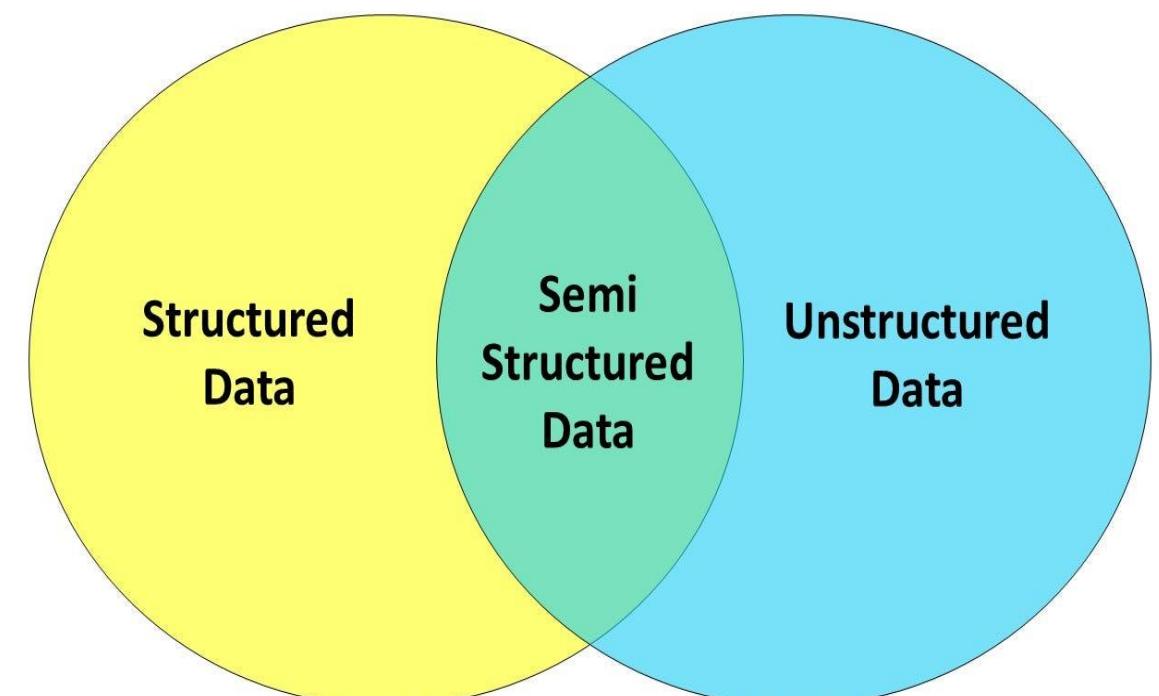
Data Type	Description	Example
Structured	Organized in rows and columns	Excel files, SQL tables
Unstructured	No predefined structure	Emails, images, social media
Semi-structured	Partially organized	JSON, XML

## Common Data Formats

- **CSV** (.csv) – Lightweight and widely supported
- **Excel** (.xlsx) – Multiple sheets and formulas
- **JSON** – Common in web APIs
- **Database tables** – MySQL, PostgreSQL

We begin with **CSV files** due to their simplicity and versatility.

This course focuses primarily on **structured data**, such as CSV files.



# What Is “Good” Data?

Good-quality data is:

- Accurate
- Complete
- Consistent
- Relevant

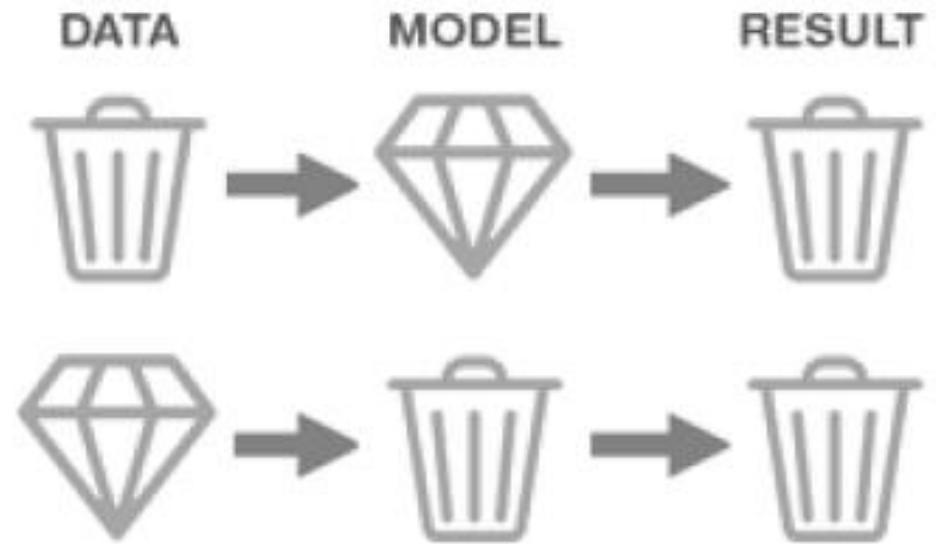
Poor-quality data leads to poor decisions

— *garbage in, garbage out.*



## Common Data Problems

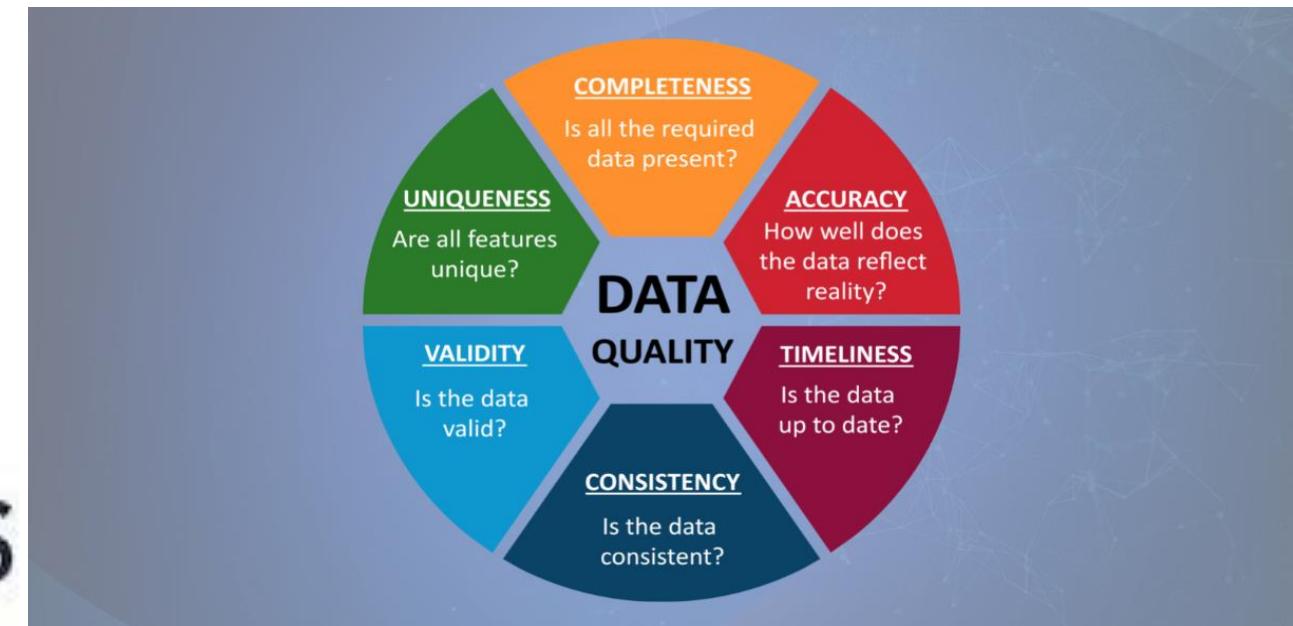
- Missing values (`NaN`)
- Duplicate records
- Outliers (e.g., unrealistic ages)
- Inconsistent formatting (e.g., “Male”, “M”, “m”)
- Incorrect data types



# Quality of the Data

**Good Quality Data + Good Quality Model = Good Results**

- ❖ High-quality data, defined by accuracy, completeness, and consistency, is the critical foundation for data analysis.
- ❖ This helps in ensuring models learn accurate patterns to produce reliable, trustworthy, and actionable results.
- ❖ Combining this with a well-designed, appropriate model prevents, garbage-in, garbage-out scenarios, leading to superior performance and informed decision-making.



# Introduction to Pivot Data Analysis

Turning Raw Student Data into Actionable Insights

## Key Points:

-  **750+ million users worldwide** — most common business tool
-  Used daily in **banks, NGOs, universities, and government** in Kenya
-  No coding required — ideal for beginners
-  **PivotTables = instant summaries** without formulas
-  Foundation for Power BI & Python later

*You'll be job-ready faster by mastering Excel first.*

# Our Dataset – Student Performance (1,500 Records)

What's in the file?

- ❖ Student ID, Gender, Age, Program, Year of Study
- ❖ Academic Scores: Math, English, Data Skills Pre-Test
- ❖ Behavioral Metrics: Attendance %, Study Hours/Week
- ❖ Background: Family Income, Scholarship Status
- ❖ Outcomes: Previous CGPA, Final Grade, At Risk of Retaking the Course or Not (Yes/No)

# What Is a PivotTable?

## Definition:

A PivotTable is an interactive Excel tool that **summarizes large datasets** by grouping, filtering, and calculating values-**without writing formulas**.

## Why it's powerful:

- . Drag-and-drop interface
- . Instantly answer questions like:
  - “What’s the average Final Grade by Program?”
  - “How many at-risk students are in Business?”
- . Updates automatically when data changes

# Creating Your First PivotTable

## Instructions:

1. Open student performance.csv in **Excel**
2. Select **any cell** in the data
3. Go to **Insert → PivotTable → OK**
4. In the PivotTable Fields pane:
  - Drag **Program** to **Rows**
  - Drag **Final Grade** to **Values** (it will default to **Average**)



You now see **average final grade per program!**

*Tip: Right-click any value → “Value Field Settings” to change from Average to Count, Sum, etc.*

# Common PivotTable Applications

PivotTables can answer common analytical questions quickly:

Analytical Question	PivotTable Configuration
Number of students per program	Rows: Program; Values: Student ID
Average attendance by gender	Rows: Gender; Values: Attendance Pct (Average)
Percentage of at-risk students by income level	Rows: Family Income Level; Values: At Risk (Average × 100)
Average study hours by scholarship status	Rows: Scholarship Status; Values: Study Hours Per Week

# Using Filters and Slicers

PivotTables become more powerful when interactivity is added.

- Use **Report Filters** to filter data by variables such as Year of Study
- Insert **Slicers** to visually filter by Gender, Program, or At-Risk status
- Combine multiple PivotTables and slicers to create interactive dashboards

*This approach mirrors professional Excel dashboards used by decision-makers.*

# Identifying At-Risk Students

To analyze student risk patterns, create the following PivotTable:

- ❖ Rows: Program
- ❖ Columns: Family Income Level
- ❖ Values: At\_Risk (Average)
- ❖ Format results as percentages

## Key insight:

Students from **low-income backgrounds**, particularly in **Education and Agriculture programs**, exhibit higher risk levels and may require targeted support.

*This demonstrates data-driven decision-making.*

# **Best Practices and Common Mistakes**

## **Best Practices:**

- . Ensure source data has no empty rows or columns
- . Use consistent category names (e.g., “Male/Female”)
- . Refresh PivotTables whenever the dataset changes

## **Common Mistakes to Avoid:**

- . Editing values inside the PivotTable
- . Saving PivotTables in CSV format (use .xlsx instead)

# Next Steps and Homework

## Assignment

1. Open `student_performance.csv` in Excel
2. Create **PivotTables** to answer the following:
  - Which program has the highest average final grade?
  - How does attendance relate to at-risk status?
  - How does scholarship status influence study hours?
3. Take a screenshot of your best PivotTable and share it in the class WhatsApp group

# Introduction to Python for Data Analysis

Python is preferred because it is:

- ❖ Easy to learn
- ❖ Supported by powerful libraries (pandas, numpy, matplotlib)
- ❖ Widely used by organizations such as Google, NASA, WHO, and Safaricom
- ❖ Free and open-source

*No prior programming experience is required.*

**Jupyter Notebook – Your Data Lab**

- ❖ Interactive environment for running code
- ❖ Combines code, text, and visualizations
- ❖ Excellent for learning and collaboration
- ❖ Available via **Anaconda** or **Google Colab**

**Tip:** *Google Colab is ideal for low-resource computers.*

# Core Python Libraries Used in This Course

## Top 10 Python Libraries



### Pandas

Data analysis and manipulation



### NumPy

Mathematical functions



### Matplotlib

Data visualisations



### Seaborn

Data visualisations



### Tensorflow

Machine Learning



### Keras

Deep Learning



### SciPy

Scientific computing



### PyTorch

Machine Learning



### Scrapy

Web crawling



### SQLModel

Interact with SQL databases

# What Is a Data Frame?

- . A DataFrame is a table of data
- . Rows represent observations
- . Columns represent variables
- . Similar to an Excel worksheet, but programmable

```
import pandas as pd  
df =  
pd.read_csv('student_performance_week1.csv  
' )
```

# Descriptive Statistics – The Basics

Key summary measures include:

- Mean
- Median
- Standard deviation
- Minimum and maximum
- Count

These statistics help describe the distribution of the data.

## Hands-On Demo – Loading and Inspecting Data

```
df =  
pd.read_csv('student_performance_week1.csv')  
df.head()  
df.info()  
df.describe()
```

These steps will be repeated during the lab session.

## Understanding the Dataset

### Dataset:

student\_performance\_week1.csv

- **Observations:** 150 students
- **Variables:**
  - Student\_ID, Name, Gender, Age, Program
  - Math\_Score, English\_Score, Data\_Skills\_PreTest
  - Attendance\_Pct, Final\_Outcome

The objective is to identify patterns and relationships.

# Variables Definition

Variable	Type	Description
`Student ID`	ID	Unique identifier
`Gender`	Categorical	Male/Female
`Age`	Integer	17–25
`Program`	Categorical	Computer Science, Education, Business, Nursing, Agriculture
`Year of Study`	Integer	1–4
`Math Score`	Float (0–100)	Baseline math ability
`English Score`	Float (0–100)	Baseline language ability
`Data Skill Pre-Test`	Float (0–100)	Prior exposure to data tools
`Attendance Pct`	Float (0–100)	Class attendance %
`Study Hours Pe Week`	Integer	Self-reported study time
`Parental Education`	Categorical	None, Secondary, Diploma, Degree
`Family Income Level`	Categorical	Low, Medium, High
`Has Internet at Home`	Binary	Yes/No
`Uses Learning Apps`	Binary	Yes/No
`Participates in Group Work`	Binary	Yes/No
`Previous CGPA`	Float (0–5.0)	Cumulative GPA from prior semesters
`Distance from Campus km`	Float	Commute distance
`Scholarship Status`	Binary	Yes/No
`Final Grade`	Float (0–100)	Target (continuous) – calculated from inputs + noise
`At Risk`	Binary	Target (classification)– Final Grade < 50