CERTIFIED DATA SCIENTIST

PEOPLE INFORMATION TECHNOLOGY PROGRAM
Lecture 02

Muhammad Nabeel Ibrahim Khan
NED University of Engineering and Technology

Agenda

- Importance of Projects in the Data Science Journey
- Introduction to Datasets

Why Projects Matter in Your Data Science

- Theoretical knowledge is important, but real learning happens when you apply that knowledge to projects.
- Projects provide tangible proof of skills for potential employers or clients.
- Projects expose you to real-world data problems, helping you build critical thinking and problem-solving skills.
- Projects allow you to experience the complete data science pipeline, from data collection to model deployment.

Types of Projects to Work On

- **Beginner Projects**: Start with simple datasets (e.g., Iris dataset, Titanic dataset).
- **Intermediate Projects**: Work on real-world datasets involving significant data cleaning and feature engineering (e.g., Kaggle competitions).
- **Advanced Projects**: Build end-to-end projects that include deployment, automation, and monitoring (e.g., fraud detection system, recommendation engine).

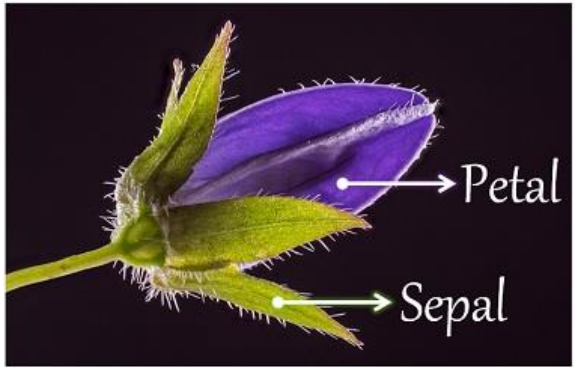
What is a Dataset?

- A dataset is a collection of data. It can contain rows and columns, where:
  - Rows represent individual observations or records (e.g., a customer, a transaction, a plant species).
  - Columns represent attributes or features of those observations (e.g., name, age, height, species).
- Types of Datasets:
  - Structured Data: Organized in rows and columns, such as spreadsheets or CSV files (e.g., Iris dataset).
  - Unstructured Data: Unorganized in a predefined way, such as images, videos, and text (e.g., raw social media posts).

Example Dataset

Iris Setosa

- Iris Dataset: Contains data about Iris flowers species.
- Data for each flower includes columns like: "Sepal Length", "Sepal Width", "Petal Length", “Petal Width” and "**Species**"



Iris Virginica

Example Dataset (Cont'd)

- Iris dataset is considered beginner-friendly for several reasons:
  - The dataset contains only 150 instances, making it easy to **load** and **manipulate**.
  - There are only four features (sepal length, sepal width, petal length, and petal width), simplifying analysis and visualization.
  - The three species are relatively well-separated, making it easier to build accurate models.
  - The dataset is widely used and well-documented, with plenty of resources available online.
  - There are no missing values, eliminating the need for data cleaning or imputation.

Iris Virginica

Example Dataset (Cont'd)

- Sample data in Iris dataset

S. No.	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
0	6.1	3.0	4.9	1.8	2
1	4.7	3.2	1.3	0.2	0
2	6.0	2.9	4.5	1.5	1
3	5.1	3.8	1.6	0.2	0
4	4.6	3.4	1.4	0.3	0

Group Activity – Brainstorming Project Ideas

Think of a simple data science project.

- What problem does the model solve?
- How will the **user interact** with it?
- How will you **deploy** the model?