Day 1: Introduction to Data Science & Basic Python

Theory

1.What is Data Science?

Definition:

- Data Science is the field of extracting insights from data using techniques from **statistics, programming, and machine learning**.
- It combines **data collection, cleaning, analysis, visualization, and modeling** to solve real-world problems.

Lifecycle of Data Science Project:

1. **Problem Understanding**
2. **Data Collection**
3. **Data Cleaning & Preprocessing**
4. **Exploratory Data Analysis (EDA)**
5. **Feature Engineering**
6. **Model Building**
7. **Model Evaluation**
8. **Deployment**
9. **Reporting & Insights**

Applications:

- Predicting sales
- Fraud detection
- Recommendation systems
- Healthcare predictions
- Customer segmentation

Roles in Data Science

- **Data Analyst** – Analyzes data, creates dashboards, insights.
- **Data Scientist** – Builds predictive models, applies ML/AI.
- **ML Engineer** – Deploys ML models into production.
- **Data Engineer** – Builds pipelines, manages databases.

Essential Skills

- **Programming:** Python (NumPy, Pandas, Matplotlib, Scikit-learn)
- **Math & Stats:** Probability, Hypothesis Testing, Linear Algebra
- **Databases:** SQL
- **Machine Learning & AI:** Regression, Classification, Deep Learning
- **Visualization Tools:** Power BI, Tableau, Matplotlib

- **Version Control:** Git & GitHub

2. Setup (Professional Environment)

Install Tools:

- Anaconda (Python + Jupyter Notebook)
- Git
- Create **GitHub Account**

3. First Jupyter Notebook – Basic Python for Data Science

Basic Python

```
[1]: print("Welcome to Data Science!")
```

```
Welcome to Data Science!
```

```
[2]: # Simple calculation
a = 10
b = 5
print("Sum:", a+b)
```

```
Sum: 15
```

```
[3]: # Python list
data = [1, 2, 3, 4, 5]
print("Average:", sum(data)/len(data))
```

```
Average: 3.0
```

Using Libraries (NumPy & Pandas)

Using Libraries (NumPy & Pandas)

```
import numpy as np
import pandas as pd
```

```
# NumPy array
arr = np.array([1, 2, 3, 4, 5])
print("Mean using NumPy:", np.mean(arr))
```

```
Mean using NumPy: 3.0
```

```
# Pandas DataFrame
data = {'Name': ['Gouthami', 'Masrath', 'Chandana'],
        'Score': [85, 90, 78]}
df = pd.DataFrame(data)
print(df)
```

```
       Name  Score
0  Gouthami     85
1   Masrath     90
2  Chandana     78
```

4. Mini Project

Objective: Analyze a simple excel file (e.g., sales data)

Steps:

1. Download sample dataset: Sample Superstore
2. Load dataset in Pandas
3. Show first 5 rows
4. Calculate:
    - Total Sales
    - Average Profit
    - Unique number of Customers
5. Print summary

Mini Project

```
import pandas as pd

# Load dataset
df = pd.read_excel("Sample - Superstore.xlsx")

# Basic EDA
print("First 5 rows:")
print(df.head())

print("Total Sales:", df['Sales'].sum())
print("Average Profit:", df['Profit'].mean())
print("Unique Customers:", df['Customer Name'].nunique())
```