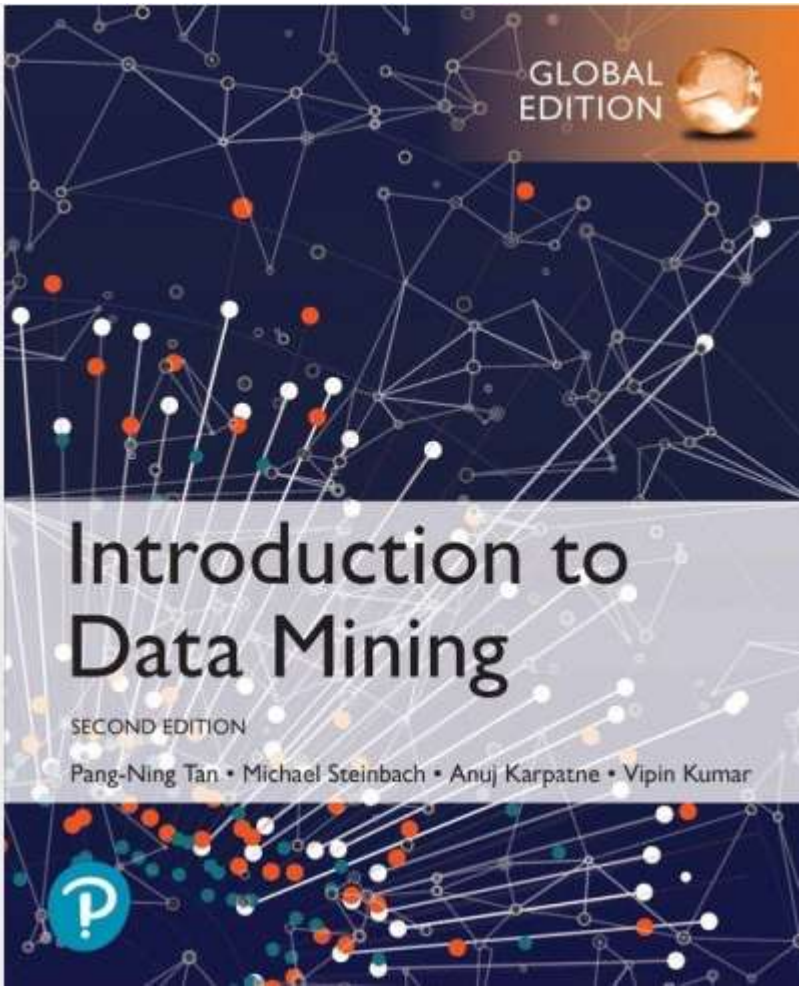
RIPHAH INTERNATIONAL UNIVERSITY



Instructor: Tajamul Shahzad

Lab 1: Environment Setup & EDA Kick-off

Objectives:

1. Set up the Python data science environment (Anaconda, Jupyter).
2. Learn to load datasets into Pandas DataFrames.
3. Perform initial data inspection and identify basic data quality issues.

Tools: Python, Pandas, NumPy, Matplotlib

Dataset: titanic.csv

Supporting Content:

- **Jupyter Notebook:** An interactive, web-based environment for writing and running code.
- **Pandas DataFrame:** A primary data structure for storing and manipulating tabular data.
- **Data Understanding:** The first step in any data mining process, crucial for defining further preprocessing.

Installation Process of Python via Anaconda:

:: 1) Remove defaults channel and force conda-forge

conda config --remove-key channels

conda config --add channels conda-forge

conda config --set channel_priority strict

:: 2) Create a fresh environment called "eda"

conda create -n eda python=3.10 -y

:: 3) Activate the environment

conda activate eda

:: 4) Install JupyterLab + core EDA libraries

conda install -y jupyterlab pandas numpy matplotlib seaborn scikit-learn

:: 5) Launch JupyterLab

jupyter lab

Practical Example 1:

```
# This is a code cell in a Jupyter Notebook
import pandas as pd

# Load the dataset
df = pd.read_csv('titanic.csv')

# Get a quick overview
print("First 5 rows:")
print(df.head())
print("\nDataset info:")
print(df.info())
print("\nSummary statistics:")
print(df.describe())
```

Lab Tasks:

1. Install Anaconda Navigator and launch Jupyter Lab.
2. Create a new notebook named Lab1_Data_Familiarization.ipynb.
3. Import the necessary libraries: pandas, numpy, and matplotlib.pyplot.
4. Load the titanic.csv dataset.
5. Use .head(), .info(), .describe(), and .shape to inspect the data.
6. Check for missing values in each column using .isnull().sum().
7. Plot a histogram of the 'Age' column using df['Age'].hist().

Exercise code Solutions

```
# Step 1: Import necessary libraries
import pandas as pd       # For data handling
import numpy as np        # For numerical computations
import matplotlib.pyplot as plt  # For plotting graphs

# Step 2: Load the Titanic dataset
df = pd.read_csv("titanic.csv")

# Step 3: Inspect the dataset
print("First 5 rows of the dataset:")
print(df.head())          # Shows the first 5 rows

print("\nDataset info:")
print(df.info())          # Gives column names, data types, and missing
values

print("\nSummary statistics:")
print(df.describe())      # Summary statistics for numeric columns

print("\nShape of the dataset:")
```

```
print(df.shape)           # Shows (rows, columns)

# Step 4: Check for missing values
print("\nMissing values in each column:")
print(df.isnull().sum())  # Count of missing values per column

# Step 5: Plot histogram of Age column
plt.figure(figsize=(8,5))
df['Age'].hist(bins=30, edgecolor='black')
plt.title("Age Distribution of Titanic Passengers")
plt.xlabel("Age")
plt.ylabel("Count")
plt.show()
```

Exercise:

8. How many passengers and features are in the dataset?
9. What is the data type of the 'Fare' column?
10. Calculate the percentage of missing values for the 'Cabin' column.
11. Plot a histogram for the 'Fare' column and describe its distribution.

2 – Data Analysis & Grouping in Titanic Dataset

Practical Example

```
# Import required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Load dataset
df = pd.read_csv("titanic.csv")

# Preview dataset
print(df.head())

# Example: Calculate survival rate by gender
survival_by_gender = df.groupby("Sex")["Survived"].mean()
print("\nSurvival Rate by Gender:")
print(survival_by_gender)

# Example: Plot survival by passenger class
df.groupby("Pclass")["Survived"].mean().plot(kind="bar")
plt.title("Survival Rate by Passenger Class")
plt.xlabel("Passenger Class")
plt.ylabel("Survival Rate")
plt.show()
```

Lab Tasks

12. Create a new notebook named Lab3_Titanic_Grouping.ipynb.
13. Import **pandas, numpy, matplotlib.pyplot**.
14. Load the titanic.csv dataset.
15. Display the first 10 rows of the dataset.
16. Filter the dataset to show only passengers who were younger than 18 (Age < 18).
17. Find the average age of passengers in each **passenger class (Pclass)**.
18. Calculate the **survival rate by gender** using .groupby().
19. Calculate the **survival rate by passenger class** using .groupby().
20. Plot a **bar chart** showing survival rate by gender.
21. Plot a **bar chart** showing survival rate by passenger class.

Exercise

1. How many passengers were children (Age < 18)?
2. Which passenger class had the **highest average age**?
3. Which gender had a higher survival rate? Provide the rate in percentages.
4. Compare survival rates across passenger classes. Which class had the **highest survival probability**?
5. Create a **stacked bar chart** showing the number of survivors vs non-survivors by gender.