Laboratory Exercise: Basic Data Cleaning and Visualization of Titanic Dataset

Objective

By the end of this laboratory exercise, students will:

- Understand and apply fundamental data cleaning techniques using Python and

Pandas. - Learn how to handle missing values, remove duplicates, convert data types,

and more. - Use the cleaned dataset for basic data analysis

Dataset

Download the Titanic dataset from Kaggle: https://www.kaggle.com/c/titanic/data

Use the file `train.csv` as your working dataset.

Lab Instructions

Step 1: Load the Data

Use Pandas to load the dataset into a DataFrame.

```
import pandas as pd

df = pd.read_csv("train.csv")
df.head()
```

Step 2: Understand the Data

Use basic functions to understand the structure of the dataset.

```
df.info()
df.describe()
df.columns
```

�� Tip: Look at column data types and note any unexpected ones.

Step 3: Check for Missing Values

```
df.isnull().sum()
```

�� Task: Identify which columns have missing values and how many.

Step 4: Handle Missing Values

Common options: remove rows/columns or fill in values.

```
# Fill Age with median
df["Age"].fillna(df["Age"].median(), inplace=True)

# Drop Cabin column (too many missing values)
df.drop(columns=["Cabin"], inplace=True)
```

Step 5: Remove Duplicates

```
# Check for duplicates
```

```
df.duplicated().sum()
# Remove duplicates
df.drop_duplicates(inplace=True)
```

Step 6: Fix Data Types

```
# Convert Survived and Pclass to categorical
df["Survived"] = df["Survived"].astype("category")
df["Pclass"] = df["Pclass"].astype("category")
```

Step 7: Standardize Column Names

```
# Convert all column names to lowercase
df.columns = df.columns.str.lower()
```

Step 8: Save the Cleaned Dataset

```
df.to_csv("titanic_cleaned.csv", index=False)
```

Basic Data Visualizations

As part of the exploratory data analysis (EDA), you are required to analyze the data to identify relations, correlations, and insights from the data. This is important to undergo this phase because this gives deeper understanding of the data and how it will behave throughout our conduct of computational thinking.

You need to import `matplotlib` into the notebook to access capability to create basic data visualizations.

To import the module, use the following code below: (It is best practice put all the module imports at the top of the cells)

```
import matplotlib.pyplot as plt
```

1. Bar Plot of Survival Count

```
df["survived"].value_counts().plot(kind="bar")
plt.title("Survival Count")
plt.xlabel("Survived")
plt.ylabel("Number of Passengers")
plt.show()
```

2. Histogram of Age Distribution

```
df["age"].plot(kind="hist", bins=20, edgecolor="black")
plt.title("Age Distribution")
plt.xlabel("Age")
plt.show()
```

3. Survival by Gender

```
df.groupby("sex")["survived"].mean().plot(kind="bar")
plt.title("Survival Rate by Gender")
```

```
plt.ylabel("Survival Rate")
plt.show()
```