IISAT UNIVERSITY

**Name Zohaib Danish**

**Roll No 311**

**Section ADP CS**

For this assignment, we'll follow the steps outlined and use a sample dataset from Kaggle.

**Data Set Selection**

For this exercise, we'll use the "Titanic - Machine Learning from Disaster" dataset available on Kaggle.

### Data Loading

We'll start by loading the dataset into a Pandas DataFrame.

**Code :**

import pandas as pd

# Load the dataset

df = pd.read\_csv('/path/to/titanic.csv')

# Display the first few rows of the dataframe

df.head()

### Data Exploration

Let's explore the dataset to understand its structure, features, and get a statistical summary.

**Code :**

# Check the structure of the dataset

df.info()

# Get the statistical summary

df.describe()

# Display the columns

df.columns

### Data Cleaning

We need to handle missing values, duplicates, and perform necessary data transformations.

**Code :**

# Check for missing values

df.isnull().sum()

# Drop rows where 'Embarked' is missing

df.dropna(subset=['Embarked'], inplace=True)

# Fill missing values in 'Age' with the median age

df['Age'].fillna(df['Age'].median(), inplace=True)

# Drop the 'Cabin' column due to a large number of missing values

df.drop(columns=['Cabin'], inplace=True)

# Check for duplicates

df.duplicated().sum()

# Drop duplicates if any

df.drop\_duplicates(inplace=True)

### Data Visualization

We'll use Matplotlib and Seaborn to create various visualizations and explain the purpose and insights derived from each.

#### Visualization 1: Distribution of Age

**Code :**

import matplotlib.pyplot as plt

import seaborn as sns

plt.figure(figsize=(10, 6))

sns.histplot(df['Age'], bins=30, kde=True)

plt.title('Age Distribution of Passengers')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.show()

**Purpose:** This histogram shows the distribution of passengers' ages.

**Insights:** Most passengers are between 20 and 40 years old.

#### Visualization 2: Survival Rate by Gender

**Code :**

plt.figure(figsize=(8, 6))

sns.countplot(x='Survived', hue='Sex', data=df)

plt.title('Survival Rate by Gender')

plt.xlabel('Survived')

plt.ylabel('Count')

plt.legend(title='Gender')

plt.show()

**Purpose:** This count plot shows the survival rate segmented by gender.

**Insights:** Females had a higher survival rate compared to males.

#### Visualization 3: Passenger Class Distribution

**Code :**

plt.figure(figsize=(8, 6))

sns.countplot(x='Pclass', data=df)

plt.title('Passenger Class Distribution')

plt.xlabel('Passenger Class')

plt.ylabel('Count')

plt.show()

**Purpose:** This count plot shows the distribution of passengers across different classes.

**Insights:** Most passengers were in the third class.

#### Visualization 4: Survival Rate by Passenger Class

**Code :**

plt.figure(figsize=(8, 6))

sns.countplot(x='Survived', hue='Pclass', data=df)

plt.title('Survival Rate by Passenger Class')

plt.xlabel('Survived')

plt.ylabel('Count')

plt.legend(title='Passenger Class')

plt.show()

**Purpose:** This count plot shows the survival rate segmented by passenger class.

**Insights:** Passengers in the first class had a higher survival rate compared to those in the second and third classes.

Analysis and Insights

After each visualization, we provided the analysis and insights derived from the visualizations.