**By**

**NAME: Zainab Imran**

**ROLL NO #: 1220100382**

**BS.SE (4th)**

**Fall 2024**

**COURSE TITLE: Artificial Intelligence**

**SUBMITTED TO: Sir Zubair**

**ASSIGMENT NO.4**

**Data Set Selection**:

We'll choose a dataset from a publicly available source.

**Data Loading**:

Load the dataset into a Pandas DataFrame.

**Data Exploration**:

Perform a detailed exploration of the dataset.

**Data Cleaning**:

Clean the data by handling missing values, duplicates, and performing necessary data transformations.

**Data Visualization**:

Create various graphs and charts using Pandas, Matplotlib, and Seaborn, and explain the purpose and insights derived from each visualization.

**Analysis and Insights**:

Provide analysis and insights after each visualization.

**Step 1: Data Set Selection**

For this assignment, let's use the "Titanic" dataset from Kaggle, which contains information about the passengers on the Titanic, including their survival status.

**Step 2: Data Loading**

First, we load the dataset into a Pandas DataFrame.

python

Copy code

import pandas as pd

# Load the dataset

url = 'https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv'

df = pd.read\_csv(url)

# Display the first few rows of the dataframe

df.head()

**Step 3: Data Exploration**

Let's explore the dataset to understand its structure and features.

python

Copy code

# Display basic information about the dataframe

df.info()

# Display summary statistics of the dataframe

df.describe()

# Display the first few rows to get a sense of the data

df.head()

**Step 4: Data Cleaning**

We'll clean the data by handling missing values, duplicates, and performing necessary data transformations.

python

Copy code

# Check for missing values

df.isnull().sum()

# Handle missing values: Fill 'Age' with the median, 'Embarked' with the mode, and drop rows where 'Cabin' is missing

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

df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)

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

# Check for duplicates and drop if any

df.drop\_duplicates(inplace=True)

# Verify the cleaning process

df.isnull().sum()

**Step 5: Data Visualization**

Now, let's create various visualizations to gain insights from the data.

**Visualization 1: Distribution of Age**

python

Copy code

import matplotlib.pyplot as plt

import seaborn as sns

# Plot the distribution of Age

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

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

plt.title('Distribution of Age')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.show()

**Insight**: The distribution of age among the passengers can help us understand the age demographics of the passengers on the Titanic.

**Visualization 2: Survival Rate by Sex**

python

Copy code

# Plot the survival rate by sex

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

sns.barplot(x='Sex', y='Survived', data=df)

plt.title('Survival Rate by Sex')

plt.xlabel('Sex')

plt.ylabel('Survival Rate')

plt.show()

**Insight**: This visualization helps us see if there was a significant difference in survival rates between male and female passengers.

**Visualization 3: Survival Rate by Passenger Class**

python

Copy code

# Plot the survival rate by passenger class

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

sns.barplot(x='Pclass', y='Survived', data=df)

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

plt.xlabel('Passenger Class')

plt.ylabel('Survival Rate')

plt.show()

**Insight**: This visualization helps us understand if the passenger class had an impact on the likelihood of survival.

**Visualization 4: Age Distribution by Survival Status**

python

Copy code

# Plot the age distribution by survival status

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

sns.histplot(data=df, x='Age', hue='Survived', multiple='stack', bins=30)

plt.title('Age Distribution by Survival Status')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.show()

**Insight**: This visualization helps us see if there were age groups that had higher survival rates compared to others.

**Step 6: Analysis and Insights**

1. **Distribution of Age**:
   * The age distribution is right-skewed with a peak around 20-30 years.
   * This suggests that most passengers were young adults.
2. **Survival Rate by Sex**:
   * The survival rate is significantly higher for females compared to males.
   * This indicates that women had a better chance of survival, possibly due to the "women and children first" protocol.
3. **Survival Rate by Passenger Class**:
   * Passengers in higher classes (1st class) had a higher survival rate compared to those in lower classes.
   * This suggests that socio-economic status played a role in survival likelihood.
4. **Age Distribution by Survival Status**:
   * The distribution shows that children had higher survival rates.
   * This further supports the notion that "women and children first" was a protocol followed during the evacuation.

By performing this detailed data analysis and visualization, we gained several insights into the demographics and survival factors of the Titanic passengers. This process demonstrates how data analysis and visualization can uncover important patterns and trends in a dataset.