EXP. NO: 2

21/8/25

DATA ANALYTICS LIFECYCLE: DATA DISCOVERY AND PREPARATION.

AIM:

To perform data discovery and exploratory analysis on a real-world dataset.

PROGRAM CODE:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer

df = pd.read_csv('titanic.csv')
print(df.head())
print(df.info())
print(df.isnull().sum())
print(df.describe())
sns.heatmap(df.isnull(), cbar=False, cmap='viridis')
plt.title('Missing Data Visualization')
plt.show()
numerical_cols = ['Age', 'Fare']
imputer = SimpleImputer(strategy='mean')
df[numerical_cols] = imputer.fit_transform(df[
                        numerical_cols])
df['Embarked'].fillna(df['Embarked'].mode()[0],
                        inplace=True)
print(df.isnull().sum())
```

OUTPUT:

PassengerId	Survived	Pclass
1	0	3
2	1	1
3	1	3
4	1	1
5	0	3

PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	0
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	687
Embarked	0

dtype: int64

Training Data Shape: (712, 8)

Testing Data Shape : (179, 8)

EXP. NO: 2

2/8/25

AIM:

To pe
analysis

PROGRAM

import

import

impor

impo

from

from

from

df

Pri

pri

p

p

s

```
x = df ['Pclass', 'Age', 'Fare', 'SibSp', 'Parch',
        'Embarked', 'Sex']]
y = df ['Survived']
x = pd.get_dummies (x, drop_first = True)
x_train, x_test, y_train, y_test = train_test_split
                                   (x, y, test_size = 0.2,
                                   random_state = 42)
Print (f"Training Data Shape: {x_train.shape}")
Print (f"Testing Data Shape: {x_test.shape}")
```

RESULT:

Thus data discovery and exploratory analysis has been performed successfully.