EX.NO02
2/8/25

Data Analytics Lifecycle
[Data discovery and preparation]

Aim

To perform data directory and exploratory analysis on real world dataset

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.

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 ()

numericals_cols = ['age', 'fare']

df [numericals_cols] = imputer.it_transform

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

Output :

	PassengerId	survived	pclass
0	1	0	3
1	2	1	1
2	3	1	3
3	4	1	1
4	5	0	3

	Name	Sex	Age	Sibsp
0	Braund, Mr. owen Harris	male	22.0	1
1	Cumings, Mrs. John Bradley	female	38.0	1
2	Allen, Mr. William Henry	male	35.0	0

#	Column	non-null count	Dtype
0	PassengerId	891 non-null	int64
1	survived	891 non-null	int64
2	pclass	891 non-null	int64
3	Name	891 non-null	object

memory usage : 83.7+ KB

passegerId : 0
survived : 0
pclass : 0
Name : 0
sex : 0
Age : 177
sibsp : 0

Training Data shape : (712, 8)
Testing Data shape : (179, 8)

```
x = df ['pclass', 'Age', 'Fare', 'sibsp', 'parch']]
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)

print (f" Training Data Shape : {x_train.shape}")
```

Result

Thus the data discovery and exploratory analysis on titanic dataset has been executed sucessfully.