Ex 02

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 import SimpleImputer.
df = pd.read_csv('titanic.csv')
print( df.head())
print(df.isnull().sum())
print( df.describe())
sns.heatmap(df.isnull(), char=False,
plt.show()
numerical_cols = ['Age', 'Fare']
df[numerical_cols] = imputer.fit_transform(df[
                                numerical_cols])
df['Embarded'].fillna(df['Embeded'].mode()[
print[df.isnull().sum())
```

output:

Passenger=d	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	177
SibSp	0
Parch	0
Ticket	0
Fare	0

training Data shape: (712, 8)

Testing Data shape: (179, 8)

```
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, 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.