Experiment No. 11
Program to perform exploratory data analysis using Numpy and Pandas
Date of Performance:
Date of Submission:

Experiment No. 11

Title: Program to demonstrate data frame creation and Manipulation using Pandas

Aim: To study and implement data frame creation and Manipulation using Pandas

Objective: To introduce Pandas package for python

Theory:

Pandas is an open-source library that is built on top of NumPy library. It is a Python package that offers various data structures and operations for manipulating numerical data and time series. It is mainly popular for importing and analyzing data much easier. Pandas is fast and it has high-performance & productivity for users.

code:

```
import numpy as np
import pandas as pd

data = pd.read_csv('dataset.csv')
print("First few rows of the dataset:")
print(data.head())
print("\nSummary statistics:")
print(data.describe())
print("\nMissing values:")
print(data.isnull().sum())

print("\nData types:")
print(data.dtypes)
print("\nUnique values in categorical columns:")
categorical_columns = data.select_dtypes(include=['object']).columns
for col in categorical_columns:
    print(f"{col}: {data[col].unique()}")
print("\nCorrelation matrix:")
```

```
print(data.corr())
```

Output:

First few rows of the dataset:

```
   ID   Name  Age  Salary
0   1   John   25   50000
1   2   Emma   30   60000
2   3  David   35   55000
3   4  Sarah   28   58000
4   5   Adam   32   62000
```

Summary statistics:

```
             ID        Age        Salary
count  5.000000   5.000000      5.000000
mean   3.000000  30.000000  57000.000000
std    1.581139   3.162278   4596.194221
min    1.000000  25.000000  50000.000000
25%    2.000000  28.000000  55000.000000
50%    3.000000  30.000000  58000.000000
75%    4.000000  32.000000  60000.000000
max    5.000000  35.000000  62000.000000
```

Missing values:

```
ID        0
Name      0
Age       0
Salary    0
dtype: int64
```

Data types:

ID int64

Name object

Age int64

Salary int64

dtype: object

Unique values in categorical columns:

Name: ['John' 'Emma' 'David' 'Sarah' 'Adam']

Correlation matrix:

	ID	Age	Salary
ID	1.000000	0.119523	-0.260254
Age	0.119523	1.000000	0.750453
Salary	-0.260254	0.750453	1.000000

Conclusion:

The program efficiently conducts exploratory data analysis (EDA) using NumPy and Pandas. It begins by loading the dataset and displaying its initial rows. Then, it provides summary statistics, identifies missing values, inspects data types and unique values in categorical columns, and computes the correlation matrix.