

Experiment No : 1

Load and Explore the Student Dataset

1.1 Objective

Load a CSV dataset using Pandas and understand its records, dimensions, column names, data types, statistical summary.

1.2 Learning outcomes

After completing this experiment, you should be able to:

- 1.import the Pandas library;
- 2.load a CSV file into a DataFrame
- 3.inspect the beginning and end of a dataset
- 4.identify the dataset's size, columns and data types
- 5.generate an initial statistical summary

1.3 Theory

NumPy and Pandas are two important Python libraries used extensively in data analysis and exploratory data analysis (EDA).

Python lists can store data, but performing data-analysis tasks using ordinary lists may require explicit loops, type checking, and additional bookkeeping. NumPy provides efficient numerical arrays, while Pandas provides labeled data structures such as Series and DataFrames.

1.3.1 NumPy

NumPy (Numerical Python) is a Python library designed for efficient numerical and scientific computing. A NumPy array can contain data in one or more dimensions. Normally, the elements of an array have a common data type, which allows numerical operations to be performed efficiently.

1.3.2 Pandas

Pandas is a Python library used for data manipulation, data cleaning, and data analysis.

Unlike a NumPy array, Pandas provides labels for rows and columns. This makes Pandas particularly suitable for working with tabular datasets. Pandas has two important data structures:

i) Series

A Series is a one-dimensional labeled data structure, similar to a single column of a table.

ii) DataFrame

A DataFrame is a two-dimensional tabular data structure consisting of rows and columns.

1.4 Program

```
# Import pandas library to work with tables of data
import pandas as pd

# Load the CSV file and stores it as a Pandas DataFrame
df = pd.read_csv('Student Academic Performance Record.csv')

# Display the first five records
df.head()
```

	Student_ID	Student_Name	Gender	Branch	PU_Marks_Pct	Attendance_Pct	Engineering_Marks_Pct	Internships_Completed
0	CS039	Vikram Desai	Male	ISE	84.6	72.0	74.0	0.0
1	CS034	Sneha Reddy	Female	ECE	86.3	86.2	71.7	1.0
2	CS011	Dhanush Bhat	Male	ECE	78.1	86.4	77.2	2.0
3	CS019	Lakshya Jain	Male	ISE	73.6	82.6	82.1	2.0
4	CS018	Keerthana Iyer	Female	ECE	74.3	78.7	70.1	2.0

Display the last five records

df.tail()

	Student_ID	Student_Name	Gender	Branch	PU_Marks_Pct	Attendance_Pct	Engineering_Marks_Pct	Internships_Completed
943	CS143	Aditya Menon	Female	ECE	79.7	76.0	72.6	1.0
944	CS249	Lakshya Jain	Female	ISE	79.9	87.4	70.4	2.0
945	CS547	Manoj Kapoor	Female	CSE	72.5	81.0	68.7	1.0
946	CS911	Gauri Shetty	Male	ECE	99.0	98.0	64.1	3.0
947	CS709	Chirag Kumar	Female	CSE	74.5	84.0	71.4	1.0

Find the dimensions of the dataset - stores the number of rows and columns in two variables

rows, columns = df.shape

print("Number of rows:", rows)

print("Number of columns:", columns)

Number of rows: 948

Number of columns: 15

Display the column names

df.columns

```
Index(['Student_ID', 'Student_Name', 'Gender', 'Branch', 'PU_Marks_Pct',  
      'Attendance_Pct', 'Engineering_Marks_Pct', 'Internships_Completed',  
      'Hackathons_Participated', 'Projects_Completed', 'Aptitude_Test_Pct',  
      'Hired', 'Placement_Package_LPA'],  
      dtype='object')
```

Display the data type of each column.

df.dtypes

Student_ID	object
Student_Name	object
Gender	object
Branch	object
PU_Marks_Pct	float64
Attendance_Pct	float64
Engineering_Marks_Pct	float64
Internships_Completed	float64
Hackathons_Participated	float64
Projects_Completed	int64
Aptitude_Test_Pct	float64
Hired	object
Placement_Package_LPA	float64
Unnamed: 13	float64
Unnamed: 14	float64
dtype:	object

View the DataFrame structure

df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 948 entries, 0 to 947
Data columns (total 15 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Student_ID                            948 non-null    object
1   Student_Name                          948 non-null    object
2   Gender                                948 non-null    object
3   Branch                                948 non-null    object
4   PU_Marks_Pct                          929 non-null    float64
5   Attendance_Pct                        911 non-null    float64
6   Engineering_Marks_Pct                 929 non-null    float64
7   Internships_Completed                 929 non-null    float64
8   Hackathons_Participated               929 non-null    float64
9   Projects_Completed                    948 non-null    int64
10  Aptitude_Test_Pct                     929 non-null    float64
11  Hired                                  948 non-null    object
12  Placement_Package_LPA                  464 non-null    float64
13  Unnamed: 13                            0 non-null      float64
14  Unnamed: 14                            0 non-null      float64
dtypes: float64(9), int64(1), object(5)
memory usage: 111.2+ KB
```

Generate statistical parameters

`df.describe(include='all').T`

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Student_ID	948	929	CS191	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Student_Name	948	674	Ishita Murthy	4	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gender	948	2	Male	496	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Branch	948	3	CSE	355	NaN	NaN	NaN	NaN	NaN	NaN	NaN
PU_Marks_Pct	929.0	NaN	NaN	NaN	77.67944	9.00747	55.0	71.6	77.3	83.1	101.0
Attendance_Pct	911.0	NaN	NaN	NaN	83.005488	6.612051	65.0	78.7	82.8	86.85	105.0
Engineering_Marks_Pct	929.0	NaN	NaN	NaN	74.884715	8.930758	51.8	69.3	74.1	79.0	112.0
Internships_Completed	929.0	NaN	NaN	NaN	1.610334	1.474986	0.0	1.0	1.0	2.0	8.0
Hackathons_Participated	929.0	NaN	NaN	NaN	2.027987	2.04983	0.0	1.0	2.0	3.0	12.0
Projects_Completed	948.0	NaN	NaN	NaN	1.699367	1.286482	0.0	1.0	2.0	2.0	5.0
Aptitude_Test_Pct	929.0	NaN	NaN	NaN	68.624543	7.601836	47.2	63.4	68.4	73.7	92.8
Hired	948	2	Yes	497	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Placement_Package_LPA	464.0	NaN	NaN	NaN	5.388578	0.766735	3.5	4.9	5.3	5.8	8.0
Unnamed: 13	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Unnamed: 14	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Inspect categorical labels

i) Count the values present in the Gender column.

`df["Gender"].value_counts(dropna=False)`

```
Gender
Male      496
Female    452
Name: count, dtype: int64
```

ii) Count the students belonging to each branch.

`df["Branch"].value_counts(dropna=False)`

```
Branch
CSE      355
ECE      301
ISE      292
Name: count, dtype: int64
```

iii) *Count the values present in the Hired column.*

```
df["Hired"].value_counts(dropna=False)
```

```
Hired
```

```
Yes    497
```

```
No     451
```

```
Name: count, dtype: int64
```