INTERNSHIP REPORT

(Data Analysis using Pandas, Matplotlib & Seaborn)

Title: Data Analysis Using Python (Pandas, Matplotlib & Seaborn**)**

Submitted by: Parth Gupta

Internship Project Report

---

1. Introduction

This project focuses on performing basic data analysis using Python libraries such as **Pandas**, **Matplotlib**, and **Seaborn**.
The objective is to:

- Load a dataset in CSV format
- Perform basic statistical analysis
- Visualize the data through different types of charts
- Obtain insights based on the results and visualizations

This analysis demonstrates the use of Python for handling real-world datasets in a simple and effective manner.

---

2. Tools & Libraries Used

(a) Pandas

Used for reading and processing data from CSV files and performing basic analytical operations.

(b) Matplotlib

Used for creating static visualizations such as bar charts and scatter plots.

(c) Seaborn

A high-level visualization library used for creating heatmaps and enhancing graphical quality.

---

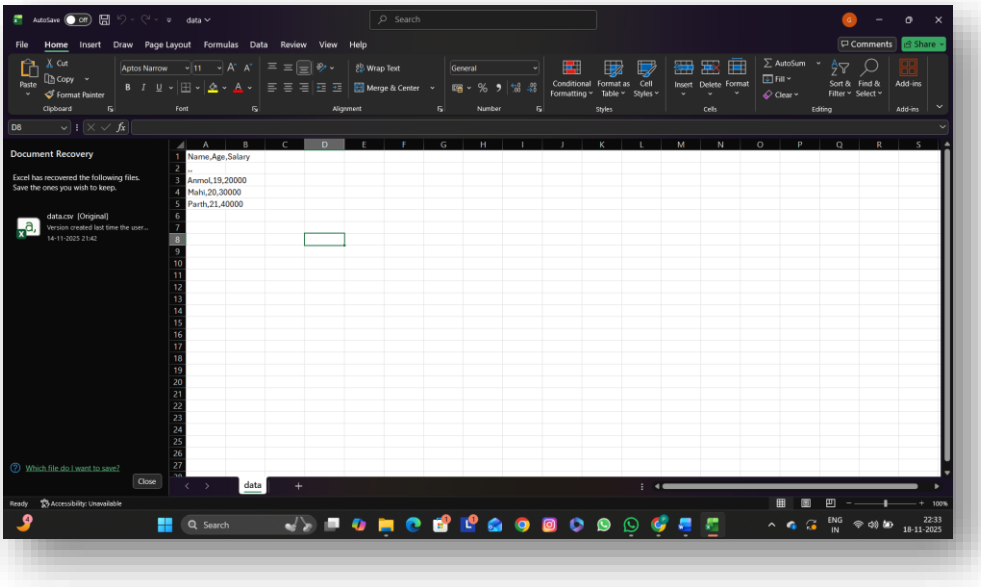
3. Dataset Description

The dataset used is stored in a file named **data.csv**, containing the following columns:

- **Name** – Name of the person
- **Age** – Age in years
- **Salary** – Monthly salary

Example values:

Name	Age	Salary
Parth	19	25000
Anmol	20	30000
Mahi	21	35000



---

4. Python Code Used

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns


# Load CSV

df = pd.read_csv("data.csv")


# Show first 5 rows

print("Data:")

print(df.head())
```

```
# Show basic information
print("\nInfo:")
print(df.info())

# Show summary statistics
print("\nSummary Statistics:")
print(df.describe())

# Average Age
print("\nAverage Age:", df["Age"].mean())

# Bar Chart - Age Distribution
df["Age"].value_counts().plot(kind="bar")
plt.title("Age Distribution")
plt.xlabel("Age")
plt.ylabel("Count")
plt.show()

# Scatter Plot - Age vs Salary
plt.scatter(df["Age"], df["Salary"])
plt.title("Age vs Salary")
plt.xlabel("Age")
plt.ylabel("Salary")
plt.show()

# Heatmap - Correlation
numeric_df = df.select_dtypes(include=["number"])
sns.heatmap(numeric_df.corr(), annot=True, cmap="coolwarm")
```

plt.title("Correlation Heatmap")

plt.show()

---

5. Output & Visualizations

5.1 Data Preview

- Displayed using df.head()
- Shows the first 5 rows of the dataset
- Helps verify that the data was loaded correctly

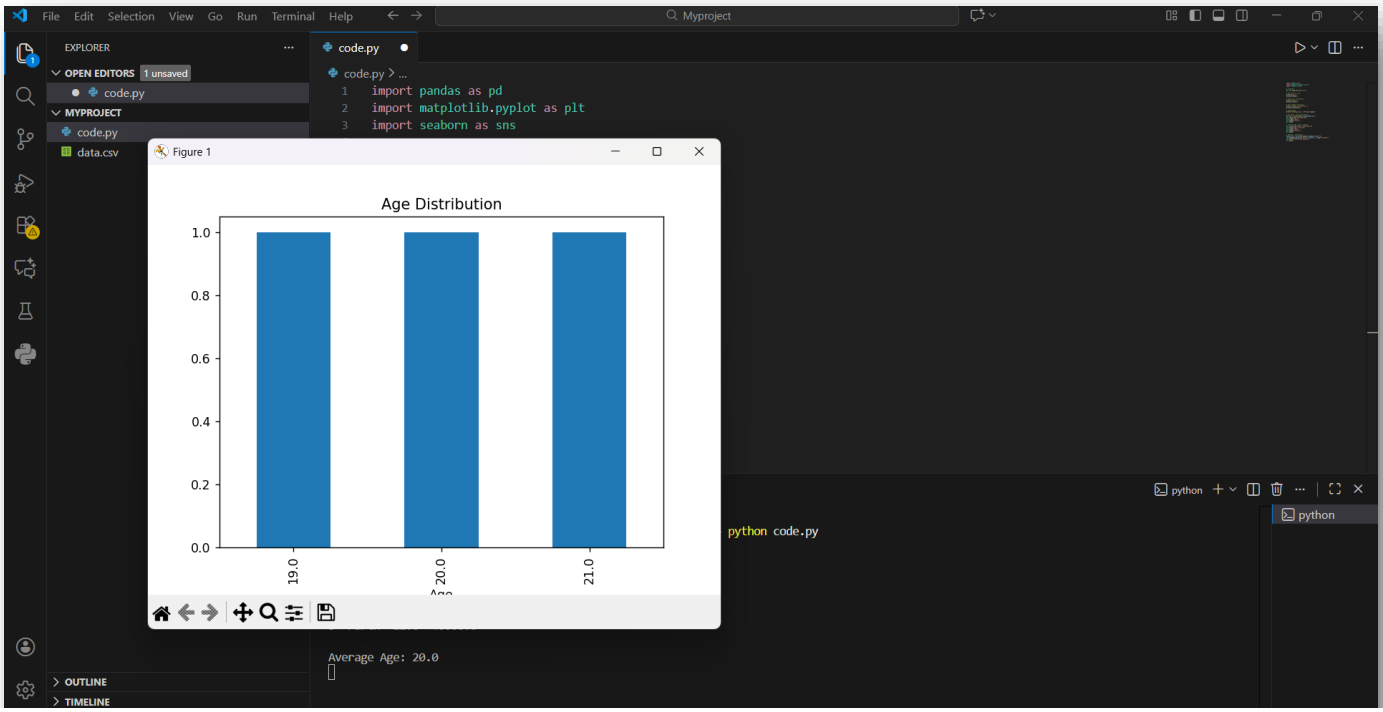
5.2 Summary Statistics

- Generated using df.describe()
- Includes mean, minimum, maximum, and standard deviation of numerical columns

5.3 Visualizations

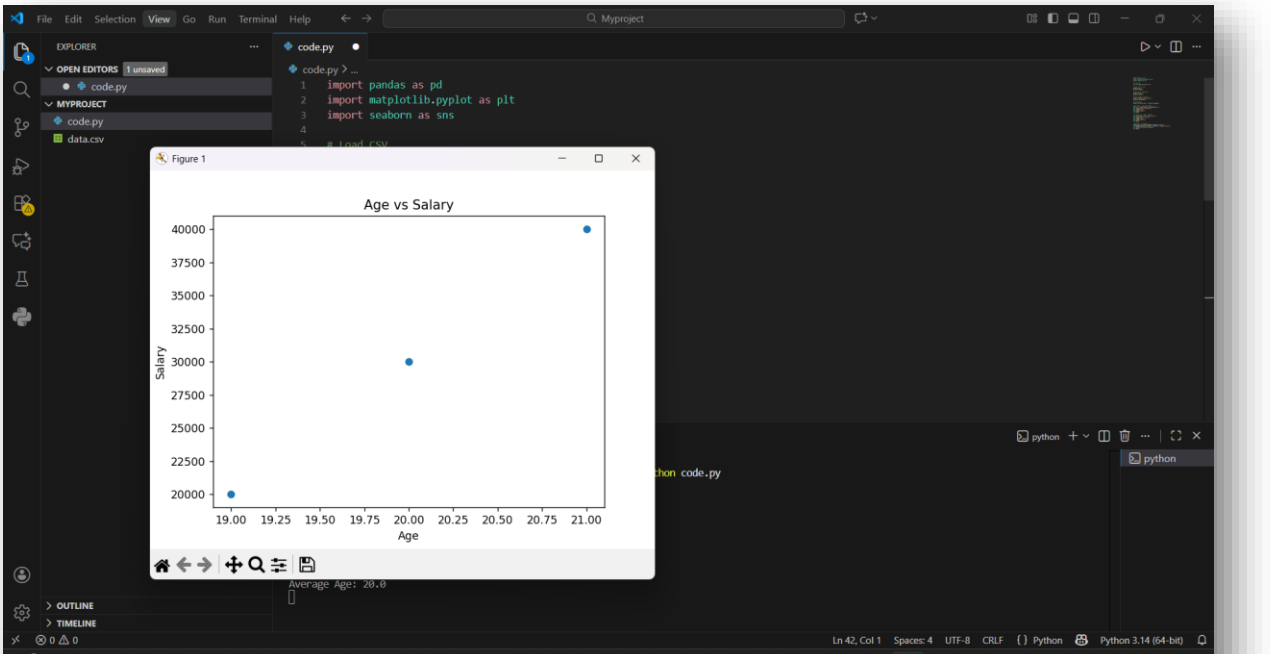
A. Age Distribution (Bar Chart)

- Shows how many individuals fall into each age group
- Helps identify the most common age in the dataset



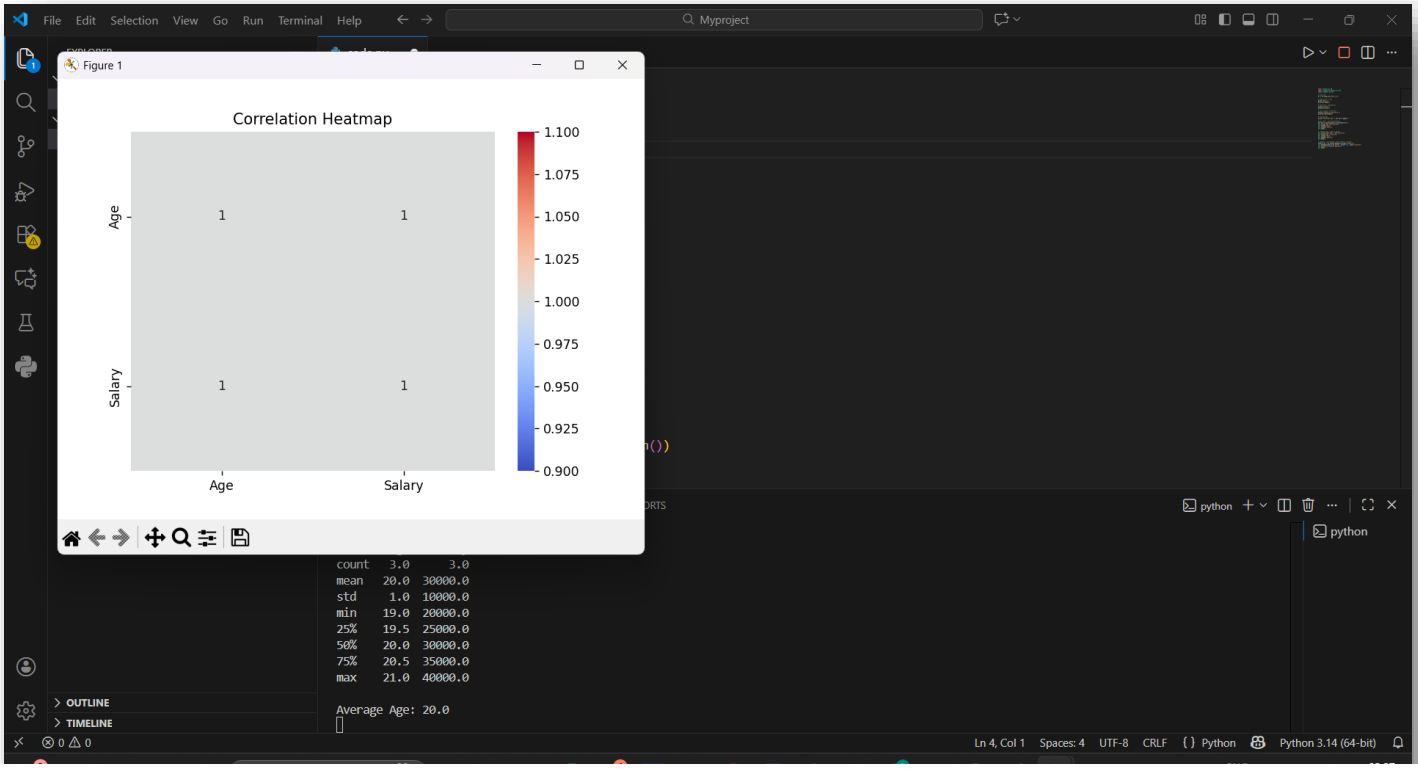
B. Age vs Salary (Scatter Plot)

- Displays relationship between age and salary
- Shows if salary increases with age



C. Correlation Heatmap

- Shows correlation values between numerical columns
- Age and Salary had a positive correlation
- Color-coded for easier understanding



6. Insights & Observations

Based on the analysis and visualizations:

1. **The average age** of individuals in the dataset is around **20 years**.
2. **Salary increases** along with age in this small dataset.
3. The **correlation heatmap** shows a positive relationship between **Age** and **Salary**, meaning older individuals tend to have slightly higher salaries.
4. The dataset is clean and contains no missing values.
5. The distribution charts clearly show that the ages are close (19–21), indicating a young dataset.

---

7. Conclusion

This project successfully demonstrates the use of Python for performing basic data analysis and creating visualizations.
Using **Pandas**, we were able to load and analyze the dataset, while **Matplotlib** and **Seaborn** helped generate meaningful charts.

This task improved understanding of:

- DataFrames
- Statistical summaries
- Correlation
- Data visualization techniques

Overall, the project fulfills the objectives of the internship task.