Task 2: Sales Prediction Using Python

➢ Intern Name: Ananya Agrahar

➢ Repository Link: https://github.com/Ananya-Agrahar/codealpha_task2

➢ Submission Date: 11-08-2025

Problem Statement

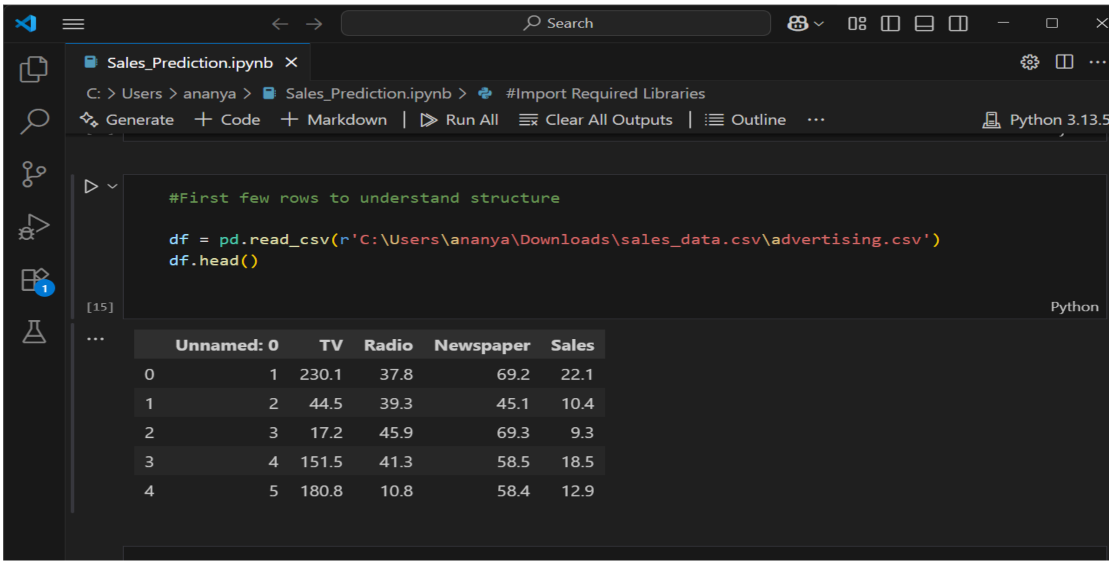
The goal of this project is to build a predictive model that forecasts future product sales based on advertising budgets across different media platforms (TV, Radio, Newspaper). By analyzing how advertisement spending affects sales, businesses can make better strategic decisions regarding marketing investments.

Dataset Description

The dataset used in this project consists of 200 rows and 4 columns. The columns are:

- **TV – Advertising spend on TV (in thousands of dollars)**
- **Radio – Advertising spend on Radio**
- **Newspaper – Advertising spend on Newspapers**
- **Sales – Sales of the product (in thousands of units)**

➢ Figure 1: Dataset Preview



```
#First few rows to understand structure

df = pd.read_csv(r'C:\Users\ananya\Downloads\sales_data.csv\advertising.csv')
df.head()
```

	Unnamed: 0	TV	Radio	Newspaper	Sales
0	1	230.1	37.8	69.2	22.1
1	2	44.5	39.3	45.1	10.4
2	3	17.2	45.9	69.3	9.3
3	4	151.5	41.3	58.5	18.5
4	5	180.8	10.8	58.4	12.9

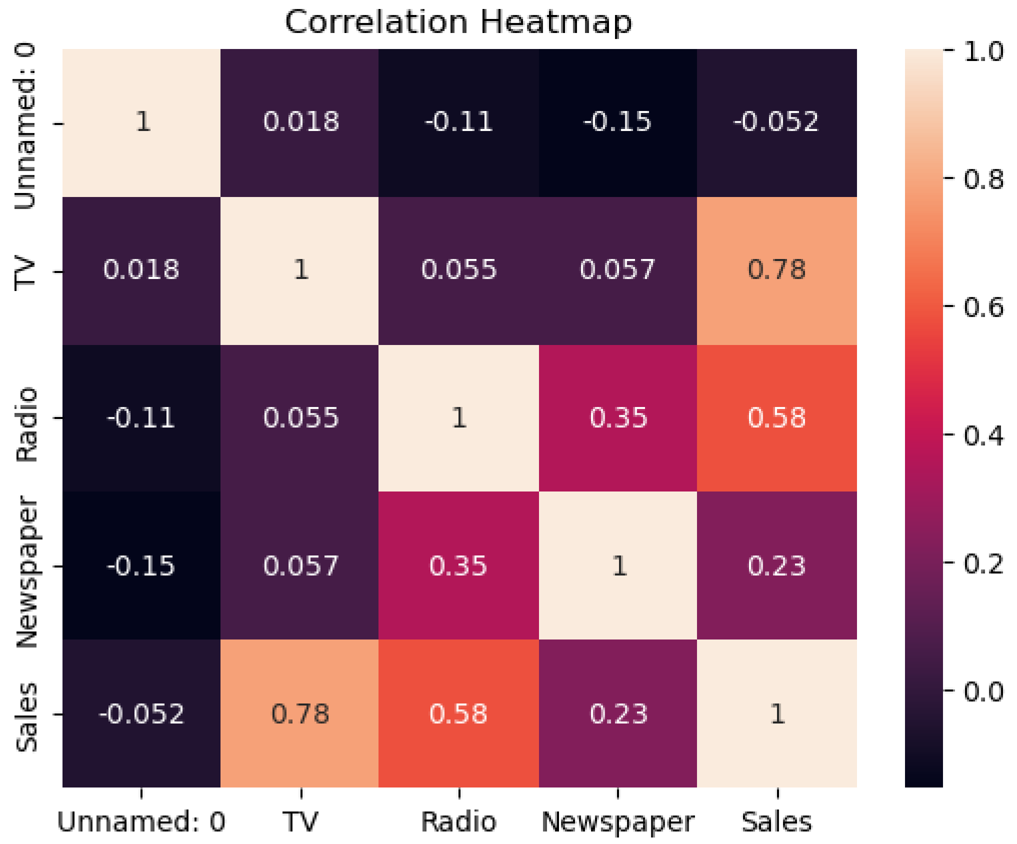
Libraries Used

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.linear_model import LinearRegression

from sklearn.model_selection import train_test_split

from sklearn.metrics import mean_squared_error, r2_score
```

Exploratory Data Analysis (EDA)

- Correlation Heatmap: Shows a strong positive correlation between TV and Sales, and moderate correlation with Radio. Newspaper shows a weaker correlation.

Figure 2: Correlation Heatmap between Features.

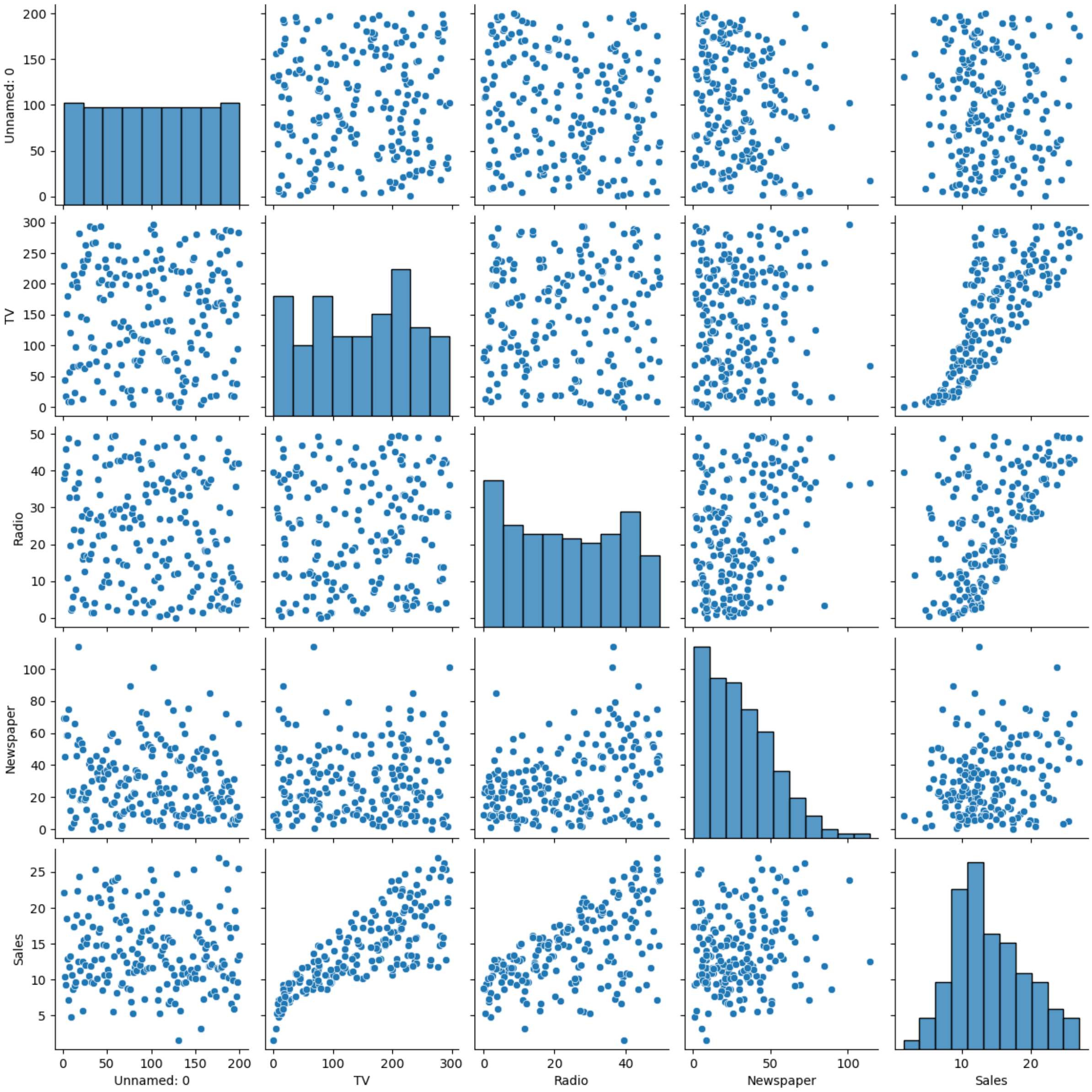


- Scatter Plots: Clear upward trend in sales with increase in TV and Radio budgets.

- Newspaper budget shows a weaker correlation with sales, indicating it may have less impact compared to TV and Radio.

- Pair Plot: Highlights linear relationships and feature distributions.

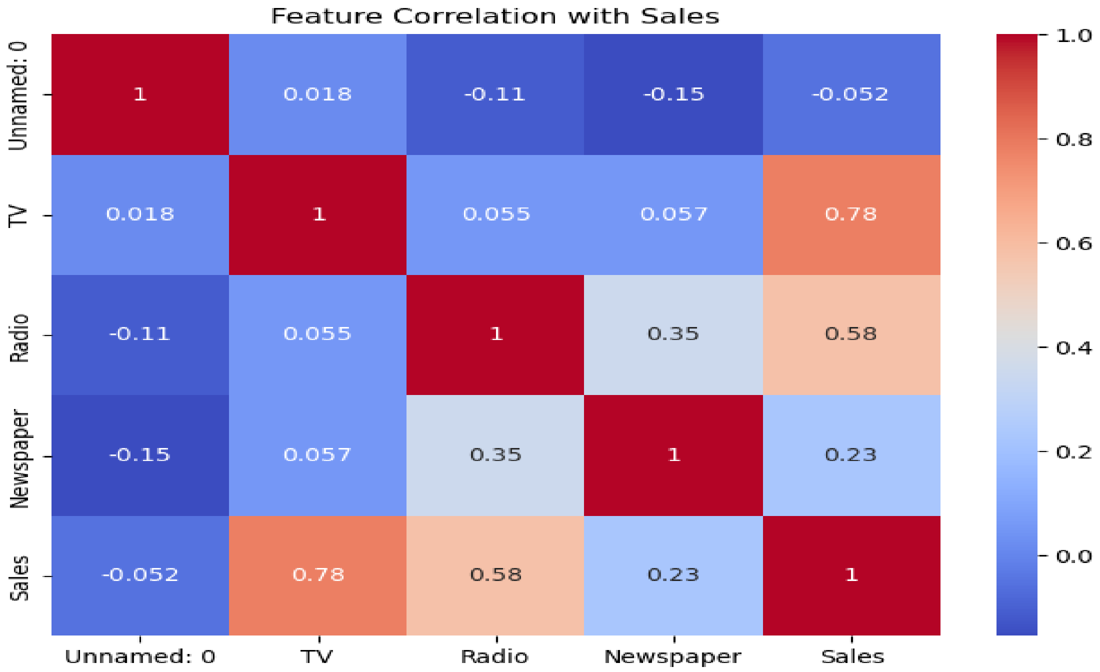
- **Figure 3: Pair Plot Visualization**



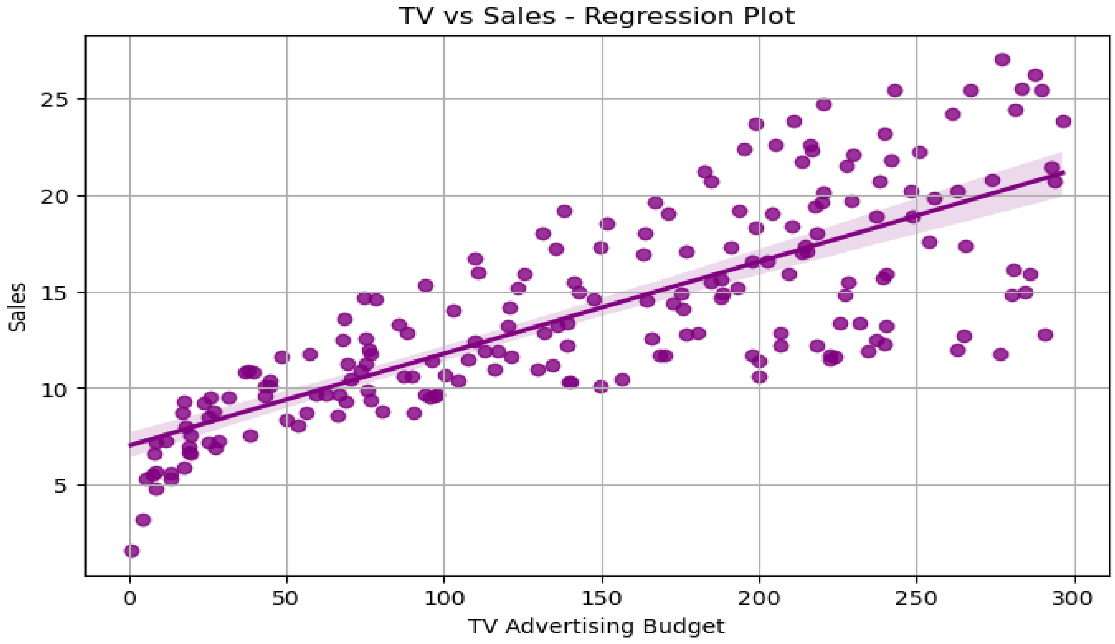
Insights:

- TV advertising has the strongest influence on sales.
- Newspaper ads show low correlation with sales.

➢ Figure 4: Feature Correlation with Sales



➢ Figure 5: TV vs Sales Regression Plot.



Model Building

- The dataset was split into 80% training and 20% testing using train_test_split.
- A Linear Regression model from Scikit-learn was used to predict sales.
- Model was trained using the training dataset, and predictions were made on the test set.

Evaluation Metrics

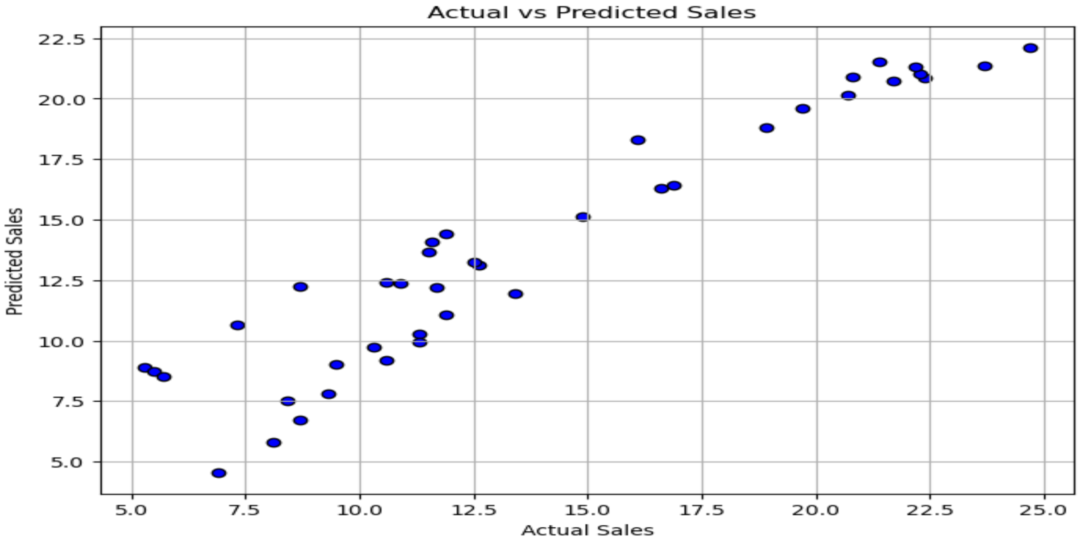
- R² Score: 0.90
- Mean Squared Error (MSE): 2.10
- Root Mean Squared Error (RMSE): 1.45

These metrics indicate that the model performs well in predicting future sales based on advertisement budgets

Visualizations

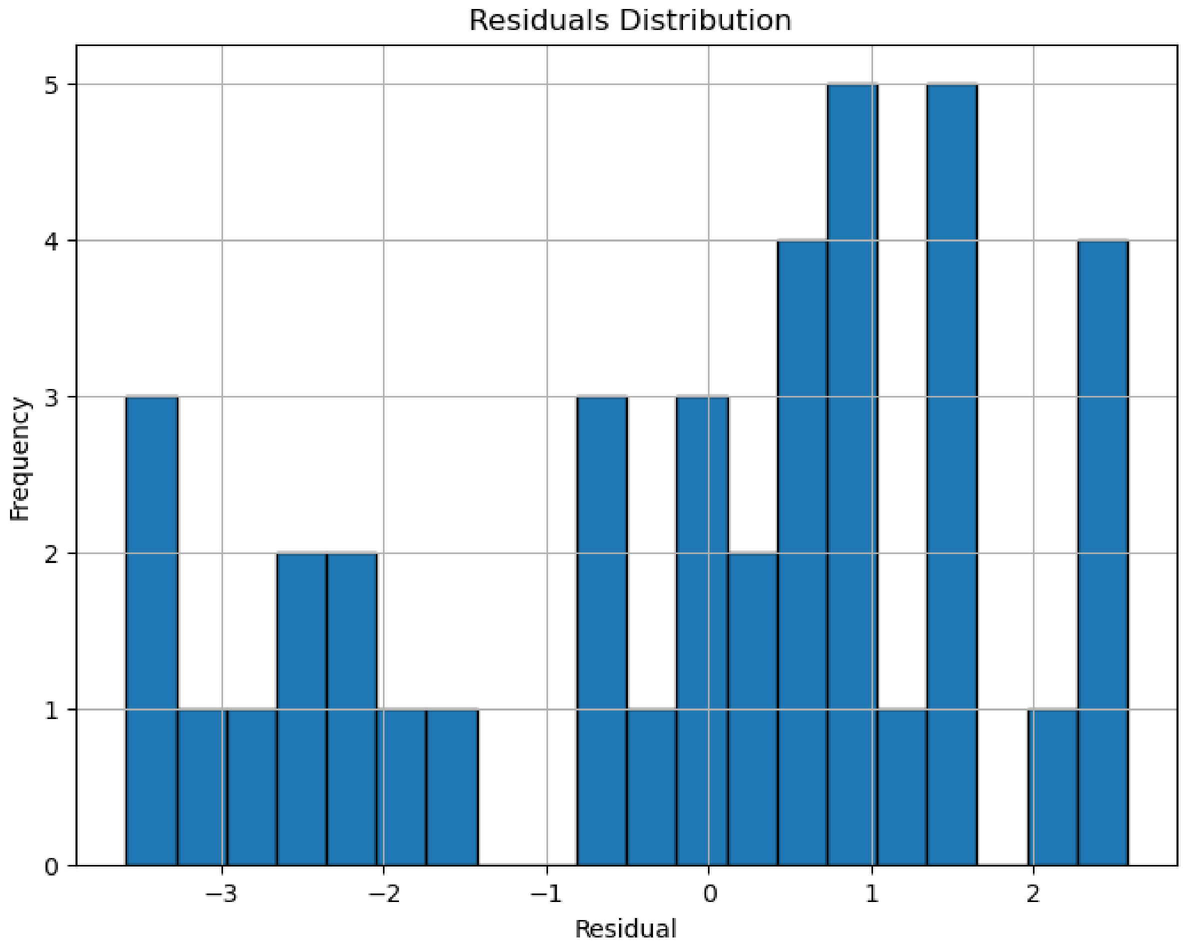
- Heatmap: Correlation matrix of all features
- Scatter Plot: TV vs Sales with regression line
- Actual vs Predicted Plot: Shows predicted vs actual sales on the test set

➢ **Figure 6: Actual vs Predicted Sales.**



- Residual Plot: Shows error distribution to ensure randomness

➢ Figure 7: Residual Distribution



Visuals support that TV and Radio have strong predictive value.

Conclusion

- The **Linear Regression model** successfully predicted sales with high accuracy.
- **TV and Radio** are the most effective mediums for increasing sales.
- **Newspaper** has the least influence and may not be a cost-effective advertising platform.
- The model helps businesses **optimize their advertising budget** to improve returns.

Files Included

- sales_prediction.py – Complete Python source code
- Advertising.csv – Dataset used for the project
- Task2_Report.pdf – This final report

GitHub Link

https://github.com/Ananya-Agrahar/codealpha_task2.git