Sales prediction using python

Abstract :

This report details the development of a sales prediction model using Python, based on advertising expenditure across TV, Radio, and Newspaper platforms. The analysis involved thorough data loading, initial exploration, and extensive visualization to understand feature distributions and relationships with sales. Key steps included data cleaning, correlation analysis, and the application of a Linear Regression model. The model's performance was evaluated using metrics such as Mean Squared Error, R2 Score, and Cross-Validation R2. The report concludes with a demonstration of sales prediction for new advertising inputs, highlighting the model's predictive capability.

Introduction :

Sales prediction is a crucial aspect of business strategy, enabling companies to forecast future revenue, optimize resource allocation, and make informed decisions. This project focuses on building a predictive model for sales based on advertising spending across different media channels: TV, Radio, and Newspaper. Utilizing Python's powerful data science libraries, we aim to understand the impact of each advertising medium on sales and predict future sales figures.

Data Loading and Initial Exploration :

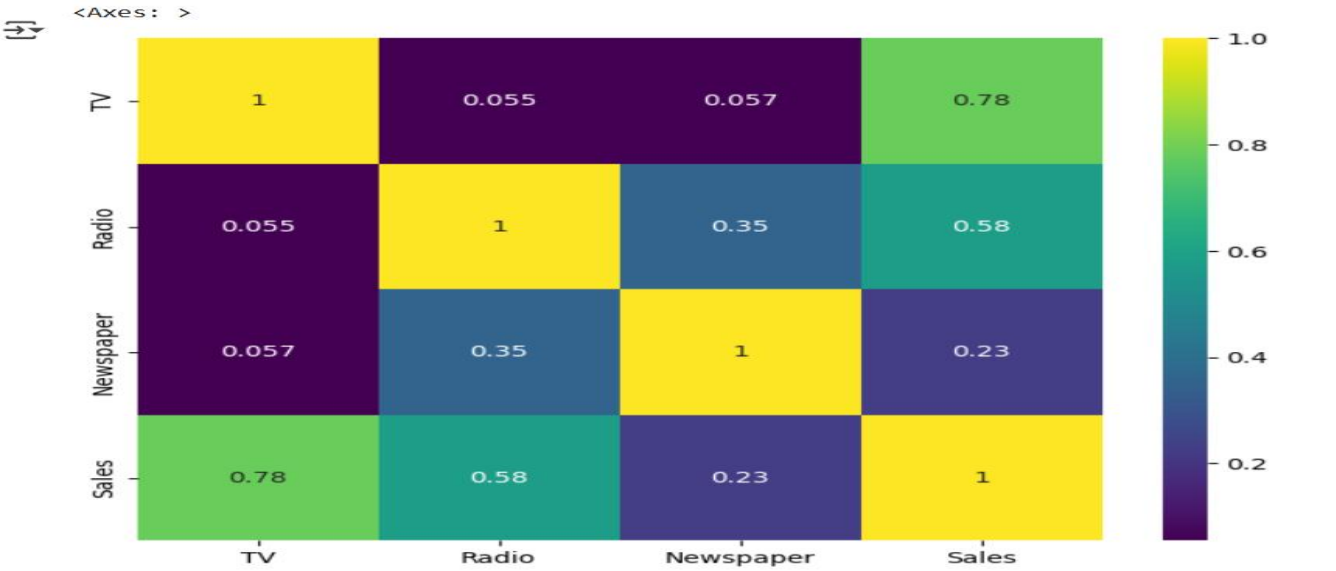
- Loading the Dataset
- Dropping Redundant Columns
- Dataset Overview: This step involved a series of checks to understand the dataset's characteristics:

    - ✓ dfads.head() was used to inspect the first few rows of the cleaned dataset.
    - ✓ dfads.shape provided the dimensions (number of rows and columns) of the dataset.
    - ✓ dfads.isnull().sum() confirmed the absence of missing values, ensuring data quality.
    - ✓ dfads.describe() generated basic statistical summaries (e.g., mean, standard deviation, min, max) for all numerical features, offering initial insights into data distribution.

Exploratory Data Analysis (EDA) and Data Visualization :

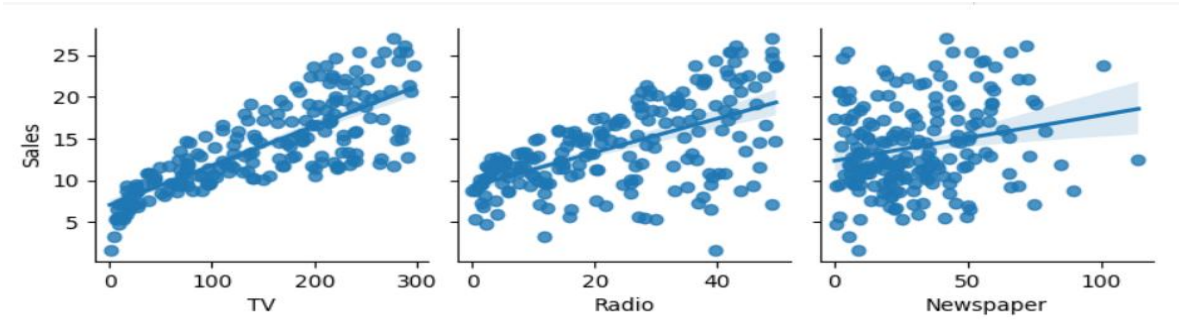
Extensive EDA and data visualization were performed to understand advertising expenditure and sales relationships:

- Correlation Analysis: A heatmap of the correlation matrix quantified linear relationships between advertising channels and sales, notably highlighting a strong positive correlation between TV advertising and Sales.

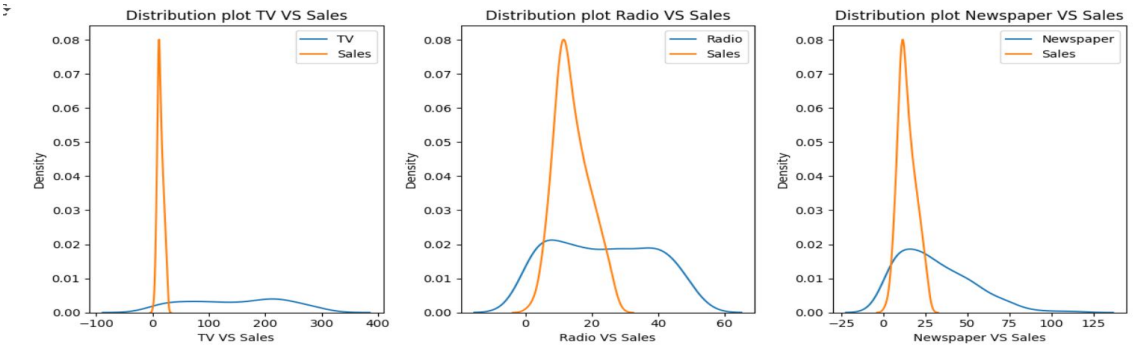


- Linearity and Relationship Checks:

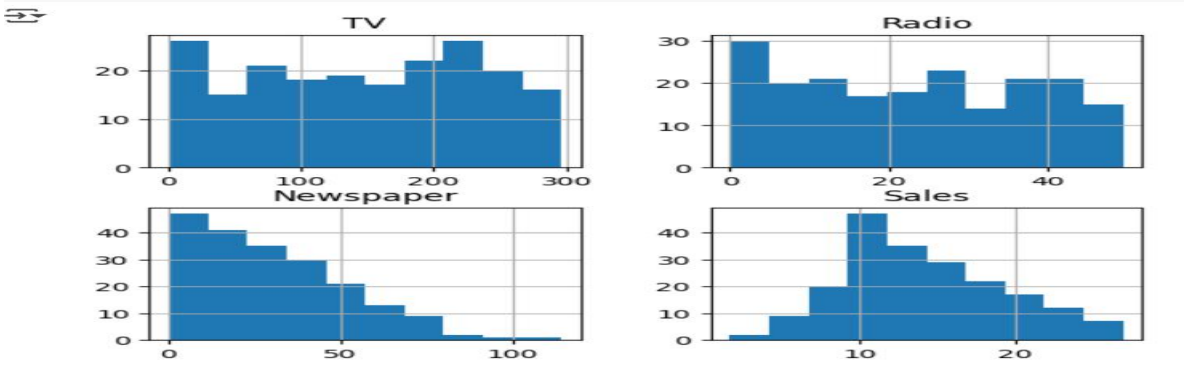
  - ✓ pairplot and individual scatter plots with regression lines were used to visualize relationships between each advertising medium and sales, assessing linearity.



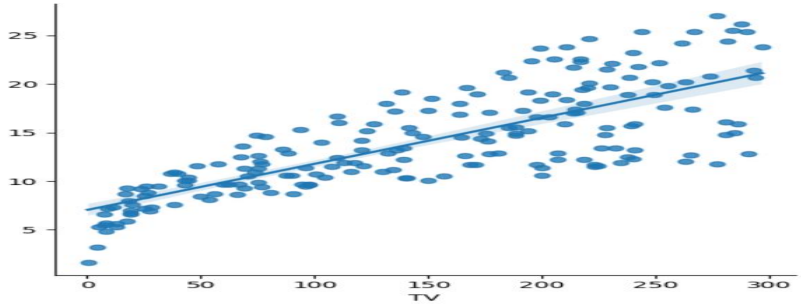
  - ✓ Distribution plots compared the spend distribution of each advertising feature against sales.

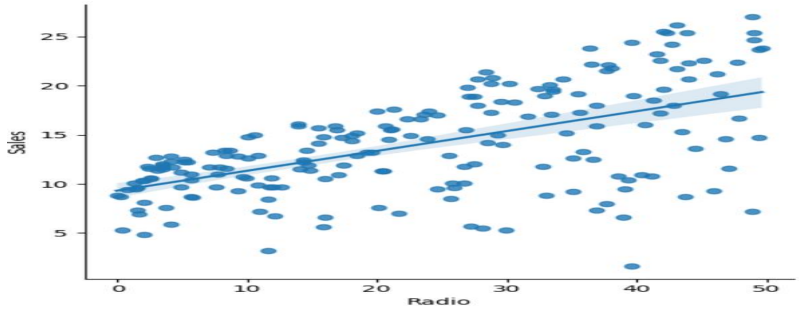


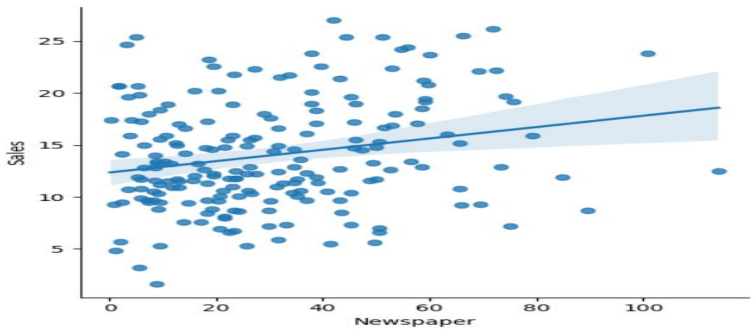
- Feature Distributions: Histograms were generated for all numerical columns to show the frequency distribution and spread of data.



- Regression Trends: lmplot was utilized to visualize the best-fit linear relationships between TV, Radio, Newspaper advertising, and Sales.







Data Preprocessing and Model Preparation :

Before training the linear regression model, the dataset underwent necessary preprocessing:

- Feature and Target Separation: The dataset was split into features (x) and the target variable
- Train-Test Split: Data was divided into 80% training and 20% testing sets for robust model evaluation.
- Target Variable Transformation (Note): LabelEncoder was present in the code for the target y, though typically unnecessary for continuous regression targets like Sales.
- Feature Scaling: StandardScaler was applied to features (x) to standardize data by removing the mean and scaling to unit variance, which can aid optimization.

```
      TV  Radio  Newspaper
0  230.1   37.8       69.2
1   44.5   39.3       45.1
2   17.2   45.9       69.3
3  151.5   41.3       58.5
4  180.8   10.8       58.4
..   ...    ...        ...
195 38.2    3.7       13.8
196 94.2    4.9        8.1
197 177.0   9.3        6.4
198 283.6  42.0       66.2
199 232.1   8.6        8.7

[200 rows x 3 columns]
```

Model Training (Linear Regression) :

A Linear Regression model was chosen to predict sales, given its suitability for identifying linear relationships between variables.

- Model Initialization: A LinearRegression model was initialized.
- Model Training: The model was trained on the preprocessed training data (x_train, y_train). The linear regression algorithm learns the optimal coefficients for each advertising medium to best predict sales.

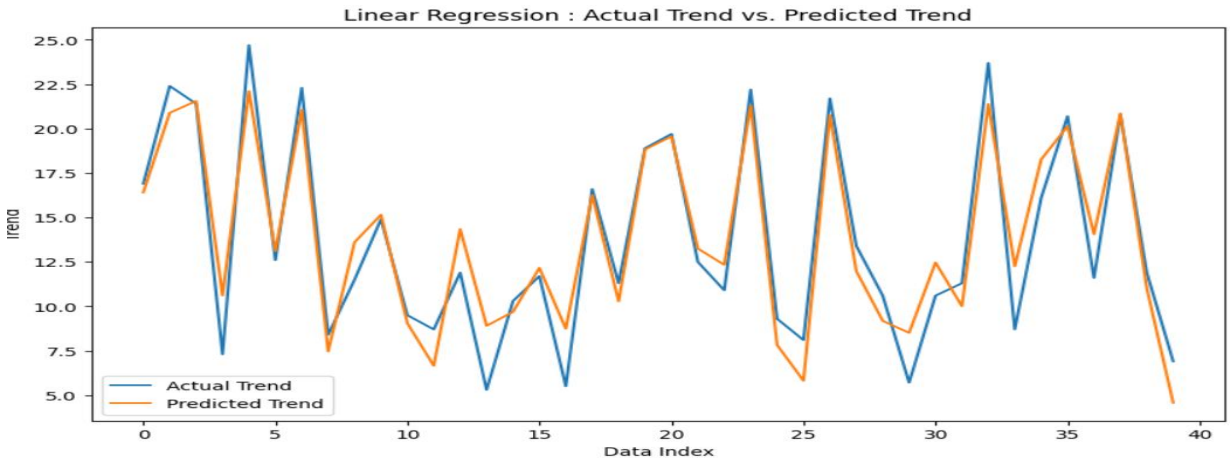
Prediction and Evaluation :

The trained Linear Regression model was evaluated using standard regression metrics and its predictive capability demonstrated:

- Model Performance: Evaluated using R2 Score, Mean Squared Error (MSE), and Cross-Validation R2 to assess accuracy and robustness.
- Sales Prediction: The model generated predictions on the test set and for new advertising inputs .

```
Predicted Sales: [16.96699408]
```

- Trend Visualization: An "Actual Trend vs. Predicted Trend" line plot visually compared the model's predictions against true sales values.



- Confusion Matrix Note: The Confusion Matrix is not applicable for evaluating regression models, as it is used for classification tasks.

Conclusion:

This sales prediction project successfully developed and evaluated a Linear Regression model to forecast sales based on advertising investments. The comprehensive data exploration and visualization steps provided deep insights into the correlations between advertising channels and sales. The model demonstrated a strong predictive capability, as evidenced by its R2 score and the visual comparison of actual versus predicted trends. The use of robust evaluation metrics, including MSE and cross-validation, ensures the reliability of the model's performance. This analysis serves as a valuable framework for businesses to optimize their advertising strategies and forecast sales more effectively.