Dataset Overview:

The dataset used in this analysis is the Advertising Dataset, which contains information about advertising expenditures across different media channels and their impact on sales. This dataset is commonly used for regression analysis and predictive modeling tasks.

Dataset Structure:

Rows: 200 (each row represents a unique observation).
Columns: 4 (3 independent variables and 1 target variable).

Dataset Features:

- TV:
  Represents the amount spent on TV advertising.
  Continuous numerical feature.
- Radio:
  Represents the amount spent on Radio advertising.
  Continuous numerical feature.
- Newspaper:
  Represents the amount spent on Newspaper advertising.
  Continuous numerical feature.
- Sales:
  Represents the sales figures corresponding to the advertising expenditures.
  Continuous numerical feature (target variable).

Methodology:

The methodology for this sales prediction project consists of several key steps:

1. **Data Collection & Preprocessing**
   - The dataset consists of advertising budgets across different media platforms (TV, `Radio`, `Newspapers`) and corresponding `Sales`.
   - Initial data cleaning involved removing unnecessary columns, handling missing values, and detecting duplicates.
   - Outliers were identified using the Interquartile Range (IQR) method and removed to improve model performance.
2. **Exploratory Data Analysis (EDA)**
   - Various statistical properties of the dataset were analyzed, including skewness, kurtosis, and correlations among features.
   - Data distributions were visualized using histograms, boxplots, and scatter plots to understand the relationship between independent variables and sales.
   - Feature engineering was performed to create new attributes, including:
     - Total Advertising = Sum of TV, Radio, and Newspaper budgets
     - TV-Radio Interaction = Multiplication of TV and Radio budgets
3. **Model Selection & Training**
   - Several regression models were trained and evaluated, including:
     - Linear Regression
     - Ridge and Lasso Regression
     - Decision Tree Regressor
     - Random Forest Regressor

        - Gradient Boosting Regressor
        - Support Vector Regression (SVR)
        - k-Nearest Neighbors (KNN)
        - XGBoost Regressor
    - **GridSearchCV** was used to optimize hyperparameters and select the best-performing model.
4. **Model Evaluation**
    - The trained models were evaluated using multiple performance metrics:
        - Mean Absolute Error (MAE)
        - Mean Squared Error (MSE)
        - Root Mean Squared Error (RMSE)
        - R^2 Score
    - The models were ranked based on RMSE, where lower values indicate better performance.
5. **Prediction on New Data**
    - The best-performing model, XGBoost Regressor, was used to predict sales based on new advertising budget values.
    - Also Linear Regression model was used to predict sales based on same advertising budget values for comparison with the best model.
    - Additional feature transformations ensured consistency between training and prediction inputs.

Observation:

- TV Advertising vs Sales
    - Strong Positive Relationship: The scatter points are closely aligned with the regression line, indicating a strong linear relationship.
- Radio Advertising vs Sales
    - Moderate Positive Relationship: The scatter points are more spread out compared to TV, but a positive trend is still visible.
- Newspaper Advertising vs Sales
    - Weak Positive Relationship: The scatter points are widely dispersed, and the regression line has a slight positive slope.

TV advertising has the highest correlation with Sales (approximately 0.78).

Radio advertising also shows a significant correlation with Sales (approximately 0.58).

Newspaper advertising has the lowest correlation with Sales (approximately 0.23).

Conclusion:

This project successfully demonstrated a machine learning-based approach to predicting sales from advertising budgets. The key findings are:

- TV and Radio advertising have a strong positive correlation with sales, whereas Newspaper advertising has a weaker impact.

- Feature engineering (e.g., Total Advertising and TV-Radio Interaction) improved model performance.
- Among the tested models, XGBoost Regressor provided the best predictive performance with an RMSE of 0.16, outperforming traditional linear regression models.
- The final trained model allows for accurate sales predictions given new advertising budget allocations, enabling businesses to make data-driven decisions regarding marketing expenditures.