Home Assignment: Sales Forecasting using Polynomial Regression

Learning Objective

The objective of this assignment is to apply **Polynomial Regression** to model a **non-linear relationship** between sales drivers and sales output.
Students will compare **Linear vs Polynomial Regression**, evaluate model performance, and justify model selection based on metrics and visualization.

Expected Completion Time

- **Best Case:** 60 minutes
- **Average Case:** 90 minutes

Problem Statement

You are provided with a **Sales dataset** that captures the relationship between **Advertising Spend** and **Sales Revenue**.
Initial analysis shows that sales growth does not increase linearly with advertising spend.

Your task is to:

- Build a **Polynomial Regression model**
- Compare it against a **Linear Regression baseline**
- Decide which model better represents the data

Dataset Description

Dataset file: `sales_data.csv`

Column Name	Description
Advertising_Spend	Amount spent on advertising
Sales	Revenue generated

Assignment Requirements

Step 1: Data Exploration

- Load the dataset using Pandas
- Display summary statistics
- Plot a **scatter plot** between Advertising Spend and Sales
- Comment on whether the relationship appears linear or non-linear

---

Step 2: Baseline Model – Simple Linear Regression

- Build a **Linear Regression** model using:
    - X = Advertising_Spend
    - y = Sales
- Train and predict on the same dataset
- Evaluate the model using:
    - Mean Squared Error (MSE)
    - R^2 Score
- Plot the regression line over the scatter plot

Step 3: Polynomial Feature Transformation

- Use `PolynomialFeatures` to transform the independent variable
- Train **Polynomial Regression models** with:
    - Degree 2
    - Degree 3
- Clearly mention which degree you are using for final evaluation

---

Step 4: Polynomial Regression Model Training

- Train a Linear Regression model on the transformed polynomial features
- Predict sales values using the polynomial model

Step 5: Model Evaluation and Comparison

Evaluate **both Linear and Polynomial models** using:

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 Score

Create a small comparison table showing:

- Model Type
- MSE
- R² Score

Step 6: Visualization

- Plot:
    - Actual sales data (scatter)
    - Linear regression line
    - Polynomial regression curve (smooth curve)
- Ensure:
    - Clear labels
    - Legend
    - Proper title

Step 7: User Input Prediction

- Ask the user to enter a new **Advertising Spend** value
- Predict sales using:
    - Linear Regression model
    - Polynomial Regression model
- Print both predictions and briefly comment on the difference

---

Step 8: Model Interpretation (Important)

Answer the following questions in comments or markdown:

1. Why does Polynomial Regression perform better or worse than Linear Regression?
2. What risks are associated with choosing a higher polynomial degree?
3. In a real business scenario, which model would you choose and why?

Hints for Learners

- Polynomial Regression is still **Linear Regression under the hood**
- Sort X values before plotting polynomial curves for smooth visuals
- Use:
- `np.argsort(X.flatten())`
- Higher R² alone does **not** guarantee a better model
- Watch out for **overfitting** when increasing degree