VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF ADVANCED SCIENCES

Winter Semester 2024-2025

Experiment 1

Objective

To build a simple linear regression model that can predict the price of a house based on its area in square feet. Specifically, we aim to estimate the prices of houses with areas of 3300 sq. feet and 5000 sq. feet using the given dataset.

Methodology

Step 1: Import Required Libraries

Necessary libraries for data handling (pandas), visualization (matplotlib), and modeling (scikit-learn) are imported. These tools allow us to efficiently manage the dataset, create visualizations, and implement the regression model.

Step 2: Load the Provided Data

The given dataset, consisting of the area and price of houses, is loaded into a pandas DataFrame. This structured format enables easy manipulation and analysis of the data.

Step 3: Understand the Data

The dataset contains two variables:

- **Area (X):** The independent variable representing the size of the house in square feet.
- **Price (Y):** The dependent variable representing the price of the house in USD.

We aim to model the relationship between area (X) and price (Y), meaning we want to predict the price of a house based on its area.

Step 4: Visualize the Data

A scatter plot is created to visually inspect the relationship between area and price. The scatter plot reveals a positive correlation: as the area of a house increases, its price also tends to increase.

Step 5: Prepare the Data for Modeling

The data is prepared by separating the independent variable (X) and dependent variable (Y). In simple linear regression, we aim to model the dependent variable (Y) as a linear function of the independent variable (X).

Step 6: Split the Data into Training and Testing Sets

In this case, we use all available data for training the model. In practice, larger datasets are typically split into a training set (used to build the model) and a testing set (used to evaluate the model's performance).

Step 7: Train the Simple Linear Regression Model

The simple linear regression model is based on the following equation:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where:

- Y is the predicted value (price of the house),
- X is the input variable (area of the house),
- β_0 is the intercept (the predicted price when the area is 0),
- β_1 is the slope (the rate of change in price for each unit change in area).

- ϵ is the **error term** or **residual**, which accounts for the variability in Y that cannot be explained by the linear relationship with X. This term represents the difference between the actual and predicted values of Y. It's assumed to follow a some distribution with a mean of 0 and constant variance.

In this step, the model learns the best-fitting line by minimizing the difference between the actual and predicted values, using a technique called **least squares estimation**.

Step 8: Make Predictions

Once the model is trained, it can be used to predict house prices for new areas. The model calculates the predicted price by plugging the area values (3300 sq. ft and 5000 sq. ft) into the equation derived from the model.

Step 9: Evaluate the Model

To assess the model's performance, we calculate two important metrics:

- **Mean Squared Error (MSE):** This measures the average squared difference between the actual values and the predicted values. A lower MSE indicates better accuracy.
- **R-squared (R^2):** This tells us how well the model explains the variation in the dependent variable (price). A value closer to 1 means a better fit of the model to the data.

Step 10: Visualize the Regression Line

The regression line (the best-fit line) is visualized on the scatter plot, which helps us see how well the model represents the relationship between area and price. The line should ideally pass close to most of the data points, indicating a good fit.

Results

Model Equation

The linear regression model has the following equation:

$$\text{Price} = 211,542 + 128.27 \times \text{Area}$$

Predicted Prices

- For an area of 3300 sq. feet, the predicted price is approximately **$669,411**.
- For an area of 5000 sq. feet, the predicted price is approximately **$854,375**.

Model Evaluation

- **Mean Squared Error (MSE):** The error is 983,144.82, which indicates that the model's predictions are not very accurate.
- **R-squared (R^2):** The R^2 value indicates that the model explains a reasonable portion of the variance in the data.

Discussion

Interpretation of Model Coefficients

- **Intercept:** The predicted starting price for a house with 0 square feet is about $211,542.
- **Slope:** For every additional square foot of house area, the price increases by $128.27.

Model Evaluation

- **MSE:** The relatively high MSE suggests that the model is not highly accurate in its predictions. With a small dataset, the model's predictions may be less reliable.
- **R-squared:** The R^2 value indicates that the model captures some of the variability in house prices, but more data would be needed to improve its fit.

Limitations

The dataset used in this experiment is quite small, which limits the generalizability of the model. A larger dataset would allow for better evaluation of the model?s performance and lead to more accurate predictions.

Conclusion

In this experiment, a simple linear regression model was successfully built to predict house prices based on the area of the house. We visualized the data, trained the model, made predictions, and evaluated its performance using the Mean Squared Error and R-squared metrics. The model showed a reasonable relationship between area and price, but its accuracy is limited by the small size of the dataset. A larger dataset would improve the model's performance and reliability.