Exploratory Data Analysis with Sales Data

DEBAJIT PRAMANICK,

B.Tech in Information Technology,

Government College of Engineering and Ceramic Technology

Period of Internship: 21ST January 2026 – 17TH February 2026

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

1. Abstract

This project focuses on performing Exploratory Data Analysis (EDA) and predictive modelling on a house price dataset from India. The objective was to understand the factors influencing house prices and develop a machine learning model to predict prices based on key features. Various preprocessing steps such as data cleaning, feature selection, and correlation analysis were performed. Strongly correlated features were selected to build a Linear Regression Model. The dataset was split into training and testing sets to evaluate the model performance. The R-squared metric was used to measure the prediction accuracy. The final model achieves an R^2 score of 0.56, indicating moderate predictive capability. The study highlights how property characteristics such as living area, number of bathrooms, and house grade impacts pricing. The project demonstrates practical application of Data Analysis and Machine Learning in Real Estate Analytics.

2. Introduction

House price prediction is one of the most practical applications of data analytics and machine learning. Real estate pricing depends on multiple factors such as area, number of rooms, conditions, locations and amenities. Understanding these factors help buyers, sellers, and real estate companies make data driven decisions.

This project was developed as a part of the 2026 Spring Internship Program to apply data science concepts learned during training. The project involves Exploratory Data Analysis (EDA), feature selection, and regression modelling using Python.

Technologies Used:

- Python
- Pandas
- NumPy
- Matplotlib
- Scikit-learn

Background and Literature Context

Linear Regression is a Supervised Learning Algorithm widely used for predicting continuous values. Correlation analysis helps identifying relationships between variables. R-squared is used to measure how well a regression model explains variability in the data.

Procedure Overview:

- Data Cleaning.
- Exploratory Data Analysis.
- Correlation Analysis.
- Feature Selection.

- Model Building.
- Model Evaluation.

Topics Covered During First Two Weeks of Internship:

- Introduction to Python
- Data Types and Control Structures
- Class, Functions and OOPS
- NumPy and Pandas
- Machine Learning Overview
- Regression in Machine Learning
- Classification in Machine Learning
- LLM Fundamentals
- Communication Skills

3. Project Objective

The main objectives of this project were:

- To perform exploratory data analysis on the house price dataset.
- To identify the most influential features affecting house prices.
- To analyse correlation between numerical variables and price.
- To build and evaluate a Linear Regression Model for price prediction.
- To interpret model performance using R-squared Metric.

No hypothesis or survey was conducted in this project.

4. Methodology

Step 1: Data Collection

The dataset was provided as a part of the internship training. It contained various features of houses such as:

- Number of bedrooms
- Number of bathrooms
- Living area
- Grade of the house
- Waterfront presence
- Basement Area
- Price (target variable)

Step 2: Data Cleaning

i) **Removed irrelevant columns**

- Date
- Longitude
- Renovation Year
- Postal Code
- Latitude
- living_area_renov
- lot_area_renov

ii) Checked for duplicates

iii) Checked for missing values

Step 3: Exploratory Data Analysis

- Generated descriptive statistics using .**describe( )**
- Grouped houses by number of bedrooms
- Compared average prices based on waterfront presence
- Computed correlation matrix

Step4: Feature Selection

- Selected features with correlation > 0.5
- Excluded 'Price' to avoid data leakage

Final Selected Features

- Number of bathrooms
- Living area
- Grade of the house
- Area excluding basement

Step5: Model Development

Model Used

Linear Regression

Train-Test Split

- 70% training
- 30% testing
- random_state = 123

Evaluation Metric

- R-squared (R^2)

Step 6: Model Validation

- Model trained on training data
- Predictions made on test data
- R^2 score calculated

Final R^2 score: **0.56**

Flowchart of Project Process

Data Collection

Data Cleaning

Exploratory Data Analysis

Correlation Analysis

Feature Selection

Train Test Split

Linear Regression Model

Model Evaluation

5. Data Analysis and Results

Descriptive Analysis

- Larger Living Areas correspond to higher prices.
- Houses with waterfront views have significantly higher average prices.
- Grade of the house strongly influences price.

Inferential Analysis

Correlation Analysis revealed:

- Living area strongly correlated with price.

- Bathrooms moderately correlated.
- Grade positively correlated.

Machine Learning Results

Model Used: Linear Regression

R^2 Score: 0.56

Interpretation

The model explains 56% of the price variation using selected features. This indicates moderate predictive performance.

6. Conclusion

The project successfully demonstrated how data analysis and machine learning can be applied to real estate price prediction. Exploratory data analysis helped identify key influencing factors such as living area, bathrooms, and grade. The Linear Regression model achieved an R^2 score of 0.56, indicating moderate accuracy. Although the model performs reasonably well, there is scope for improvement by including more features or using advanced machine learning algorithms such as Random Forest or Gradient Boosting. The project strengthened understanding of data preprocessing, feature selection, and model evaluation techniques. Future work can involve nonlinear models to improve prediction accuracy.

7. APPENDICES

Appendix A: References

- Scikit-learn Documentation
- Pandas Official Documentation
- Python for Data Analysis – Wes McKinney