Comprehensive Analysis and Regression Modeling of Real Estate Prices

Foundations of Machine Learning – Assignment 1

April 30, 2025

Abstract

This comprehensive report presents a data-driven analysis and modeling of real estate prices using a housing dataset from Taiwan. The report walks through all essential stages of machine learning workflow including data cleaning, visualization, hypothesis testing, correlation analysis, and linear regression. Particular attention is given to explaining the impact of each variable on house pricing, visualizing important patterns, and interpreting results from statistical tests and models.

1 Introduction

Real estate valuation is a cornerstone of economic development, affecting investment, taxation, and personal finance decisions. With the growing availability of data, predictive modeling has become a vital tool to support property assessment. In this study, we analyze a dataset that includes housing features and transaction details to identify factors that influence house price per unit area and to develop a predictive regression model.

2 Dataset Description and Preprocessing

2.1 Data Overview

The dataset includes 414 observations and 7 key features:

- **X1 transaction date**: Time of sale.
- **X2 house age**: Age of the house in years.
- **X3 distance to MRT station**: Proximity to public transport.
- **X4 number of convenience stores**: Nearby store count.
- **X5 latitude, X6 longitude**: Geographical coordinates.
- **Y house price of unit area**: Target variable.

2.2 Data Cleaning

- All column names were standardized for consistency.
- Checked for missing values — none were found.
- A new variable `Age_Group` was derived by comparing house age to the median.

3 Exploratory Data Analysis

3.1 Distribution of Features

Visualizing the distribution of each feature helps us understand the spread and skewness of the data. Figure **??** shows normalized distributions of all features.

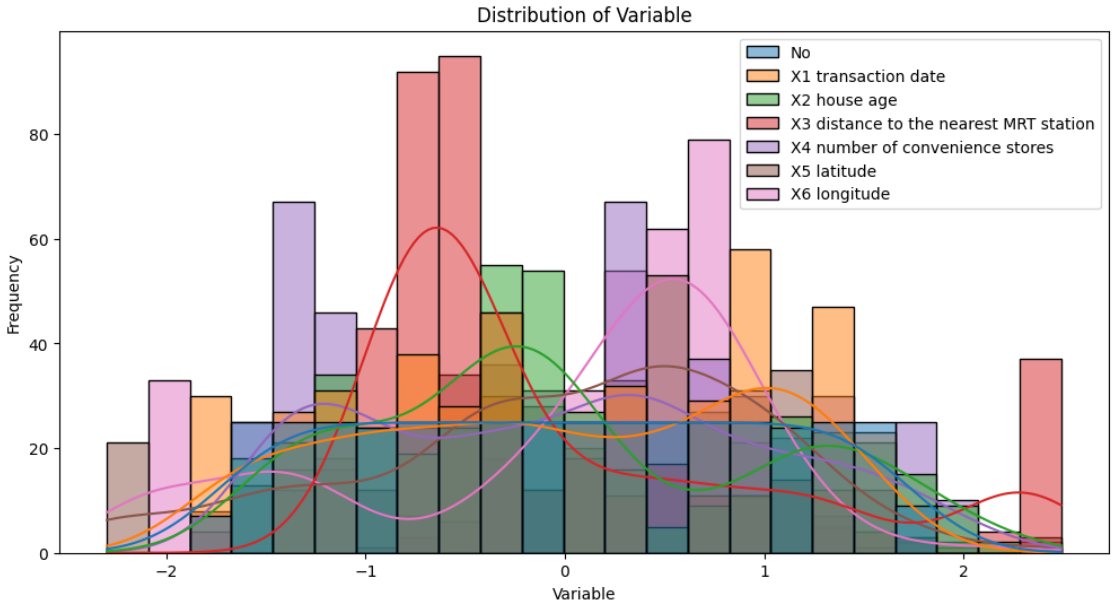

Figure 1: Normalized Distribution of Numerical Features

We observe that distance to MRT station is heavily skewed, indicating that many houses are located near a station. Latitude and longitude are normally distributed, indicating spatial spread.

3.2 Correlation Matrix

To assess the strength and direction of linear relationships, we compute pairwise Pearson correlations. Figure **??** shows a heatmap.

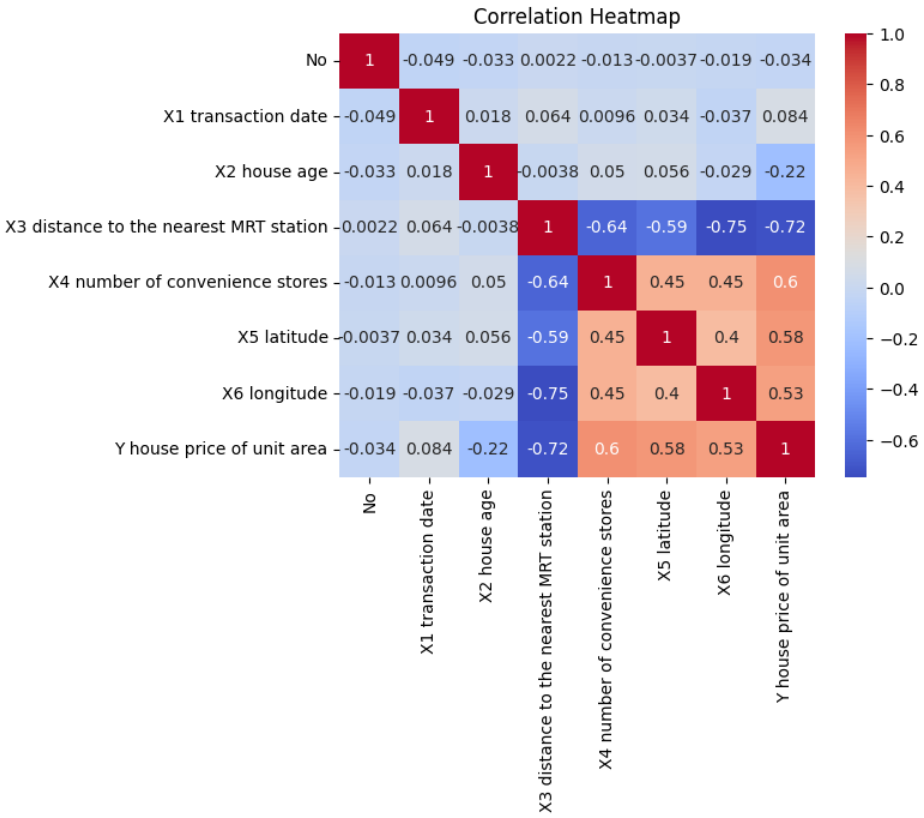

Figure 2: Correlation Heatmap of Features

The strongest negative correlation is between house price and distance to MRT station ($r \approx -0.72$). Latitude and number of stores also show a moderate positive correlation with price.

3.3 Visualizing Key Relationships

One of the most crucial findings is the inverse relationship between price and distance to MRT station. This is shown using a regression plot in Figure **??**.

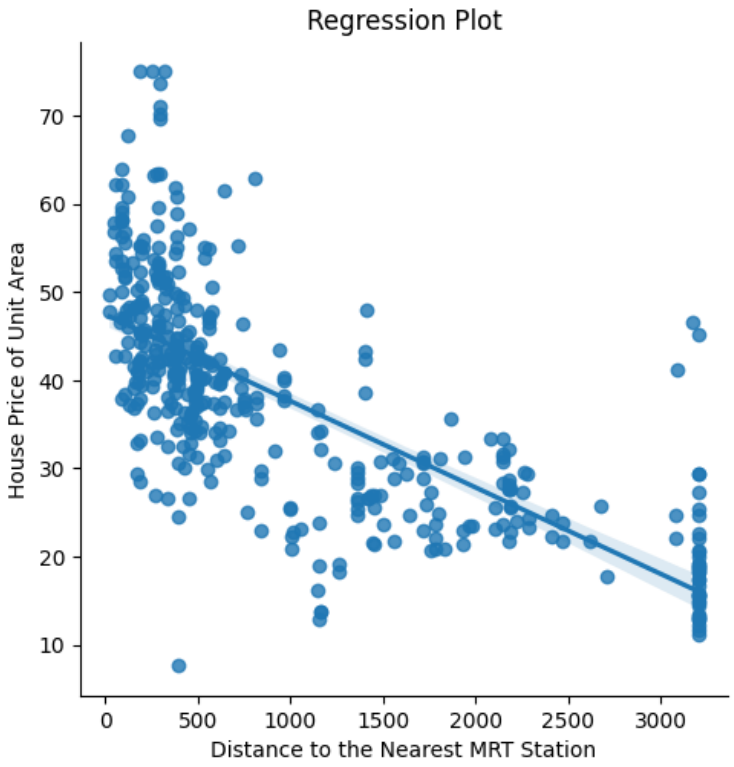

Figure 3: Regression Plot: Price vs Distance to MRT Station

As expected, properties located closer to MRT stations tend to be priced higher. This insight justifies the inclusion of this feature as a major predictor in our model.

4 Statistical Hypothesis Testing

We perform three hypothesis tests to determine whether observed differences are statistically significant.

4.1 Test 1: Price by House Age Group

We test whether homes above the median age differ significantly in price from those below it using a two-sample t-test.

- H_0: Mean price per unit area is equal across age groups.
- H_1: Mean price differs.

The result showed a statistically significant difference ($p < 0.05$), suggesting newer homes are priced higher.

4.2 Test 2: Price by Convenience Stores

We use ANOVA to test whether the number of nearby convenience stores influences price.

- H_0: Mean price is the same for all store-count categories.
- H_1: At least one group differs.

The result again indicated significance, confirming that accessibility to stores is a valuable attribute.

4.3 Test 3: Association Between Categorical Features

By binning distance and age, we performed a chi-squared test for independence. A significant result supports that these features are related and may compound their effects on price.

5 Linear Regression Modeling

5.1 Train-Test Split and Model Training

The dataset was split 80/20. A simple linear regression model was trained using the most correlated feature (distance to MRT). The model output confirmed:

- Negative coefficient for MRT distance, aligned with correlation analysis.
- R-squared around 0.52, indicating decent explanatory power for a single-feature model.

5.2 Model Interpretation

The learned weight for MRT distance suggests that for each additional meter away from the MRT station, the unit price drops by approximately 0.007. This supports the economic principle that accessibility enhances property value.

6 Conclusion

This project demonstrated a full cycle of data science from exploration to modeling. Key takeaways:

- MRT proximity is the most influential factor on house price.
- Number of convenience stores and house age also contribute significantly.
- Visualization and statistical testing provided clear, aligned insights.
- Even a simple regression model can offer practical predictive capabilities.

Future work could include multiple regression, regularization, or non-linear models for improved accuracy.