House Price Analysis in Taiwan

Stella Zhang

2024-08-18

```
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.94 loaded
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

1. Introduction

In this analysis, we explore the factors that influence house prices in Taiwan using a multiple linear regression model. The dataset includes variables such as house age, distance to the nearest MRT station, the number of nearby convenience stores, and geographic coordinates (latitude and longitude). Our goal is to identify significant predictors of house prices and quantify their impact.

2. Data Exploration

2.1 Data Overview

We begin by loading and exploring the dataset.

```
# Load the dataset
house_data <- read.csv("~/Downloads/real-estate-taiwan.csv")

# Display the first few rows and structure of the dataset
head(house_data)
```

```
##   transaction_date house_age mrt_distance convenience_stores latitude longitude
## 1         2012.917      32.0     84.87882                 10 24.98298  121.5402
## 2         2012.917      19.5    306.59470                  9 24.98034  121.5395
## 3         2013.583      13.3    561.98450                  5 24.98746  121.5439
## 4         2013.500      13.3    561.98450                  5 24.98746  121.5439
## 5         2012.833       5.0    390.56840                  5 24.97937  121.5425
## 6         2012.667       7.1   2175.03000                  3 24.96305  121.5125
##   house_price
## 1        37.9
```

```
## 2        42.2
## 3        47.3
## 4        54.8
## 5        43.1
## 6        32.1
```

```r
str(house_data)
```

```
## 'data.frame':    414 obs. of  7 variables:
##  $ transaction_date  : num  2013 2013 2014 2014 2013 ...
##  $ house_age         : num  32 19.5 13.3 13.3 5 7.1 34.5 20.3 31.7 17.9 ...
##  $ mrt_distance      : num  84.9 306.6 562 562 390.6 ...
##  $ convenience_stores: int  10 9 5 5 5 3 7 6 1 3 ...
##  $ latitude          : num  25 25 25 25 25 ...
##  $ longitude         : num  122 122 122 122 122 ...
##  $ house_price       : num  37.9 42.2 47.3 54.8 43.1 32.1 40.3 46.7 18.8 22.1 ...
```

2.2 Descriptive Statistics

Summary statistics provide an overview of the dataset:

```r
summary(house_data)
```

```
##  transaction_date   house_age       mrt_distance     convenience_stores
##  Min.   :2013     Min.   : 0.000   Min.   :  23.38   Min.   : 0.000
##  1st Qu.:2013     1st Qu.: 9.025   1st Qu.: 289.32   1st Qu.: 1.000
##  Median :2013     Median :16.100   Median : 492.23   Median : 4.000
##  Mean   :2013     Mean   :17.713   Mean   :1083.89   Mean   : 4.094
##  3rd Qu.:2013     3rd Qu.:28.150   3rd Qu.:1454.28   3rd Qu.: 6.000
##  Max.   :2014     Max.   :43.800   Max.   :6488.02   Max.   :10.000
##     latitude       longitude      house_price
##  Min.   :24.93   Min.   :121.5   Min.   :  7.60
##  1st Qu.:24.96   1st Qu.:121.5   1st Qu.: 27.70
##  Median :24.97   Median :121.5   Median : 38.45
##  Mean   :24.97   Mean   :121.5   Mean   : 37.98
##  3rd Qu.:24.98   3rd Qu.:121.5   3rd Qu.: 46.60
##  Max.   :25.01   Max.   :121.6   Max.   :117.50
```

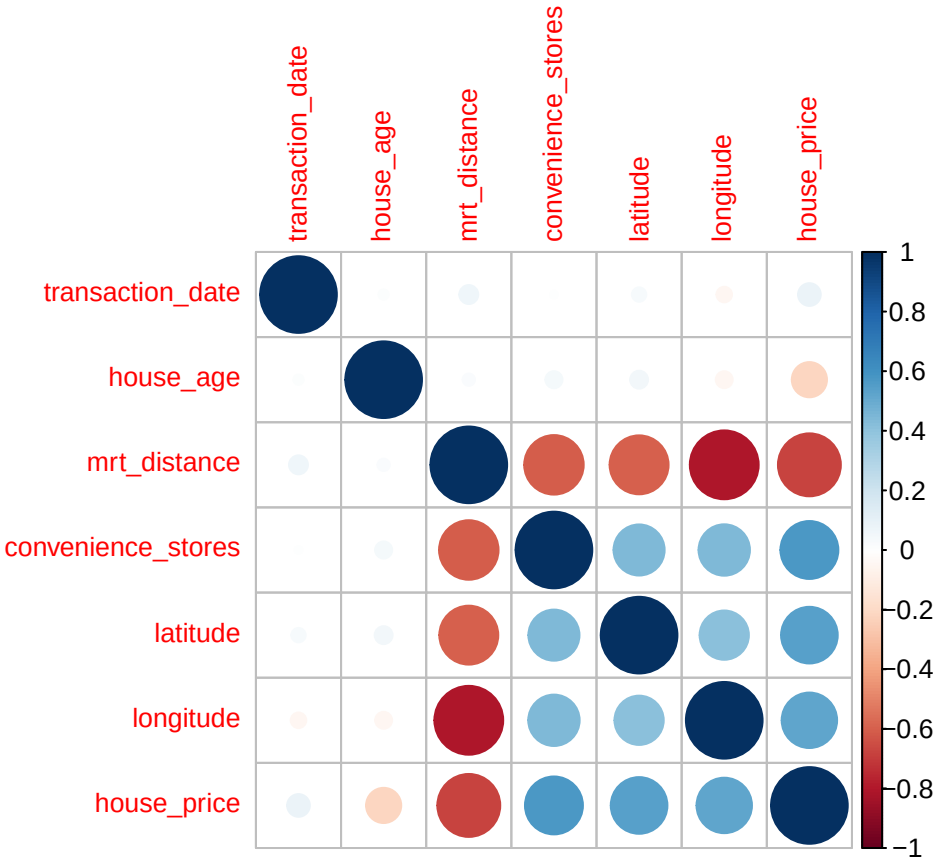
The summary statistics show the central tendency and spread of variables such as house_age, mrt_distance, and house_price. For example, the house_price ranges from 7.6 to 117.5 million NT, $with a median of 38.45 million NT$. The wide range and median suggest significant variation in property values, which the regression model will explore further.

2.3 Correlation Analysis

We compute and visualize the correlation matrix to identify potential relationships among numerical variables.

```r
# Select numeric variables for correlation analysis
numeric_vars <- house_data %>% select_if(is.numeric)

# Compute and plot the correlation matrix
corr_matrix <- cor(numeric_vars)
corrplot(corr_matrix, method = "circle", tl.cex = 0.8)
```

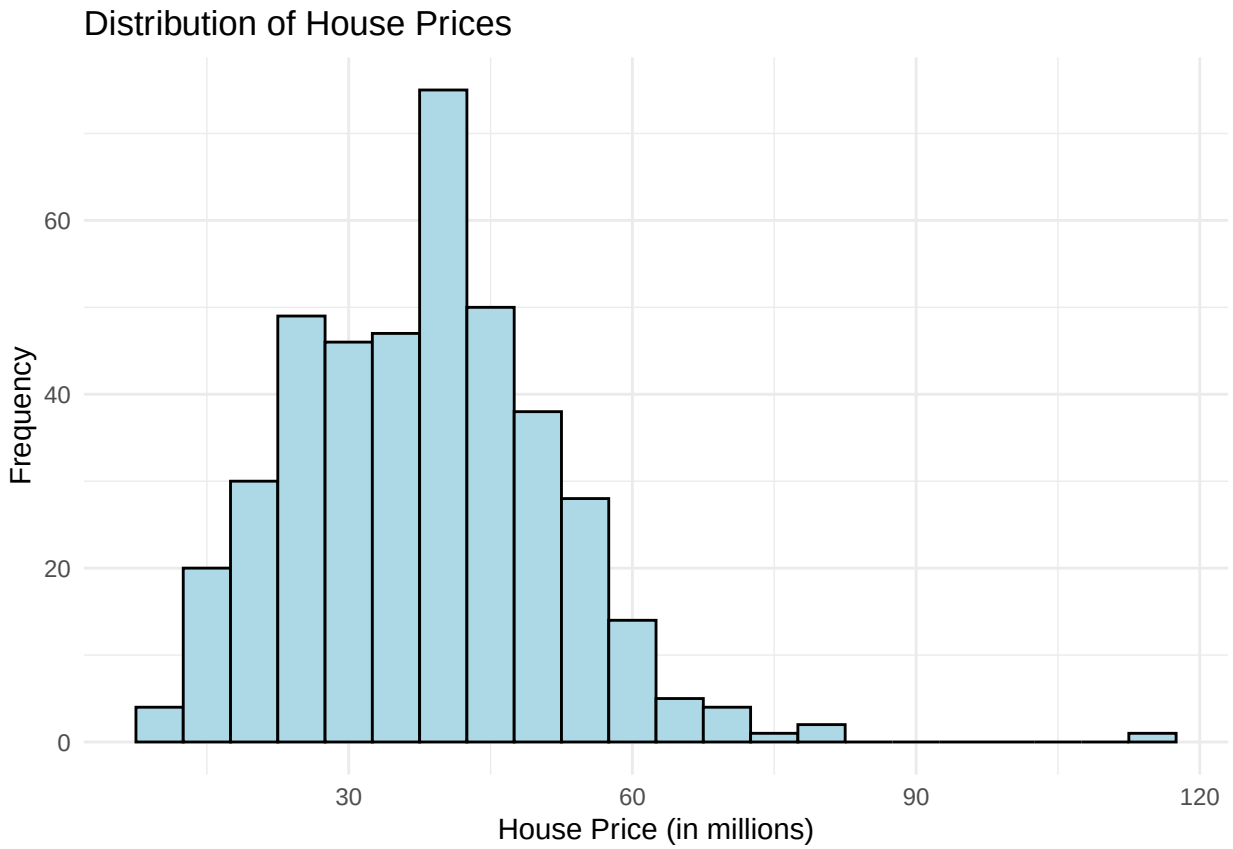


The correlation matrix reveals that mrt_distance has a strong negative correlation with house_price (-0.67), indicating that properties closer to MRT stations tend to have higher prices. House_age also shows a moderate negative correlation with house_price, suggesting that newer houses are generally more expensive. These relationships justify their inclusion as predictors in the regression model.

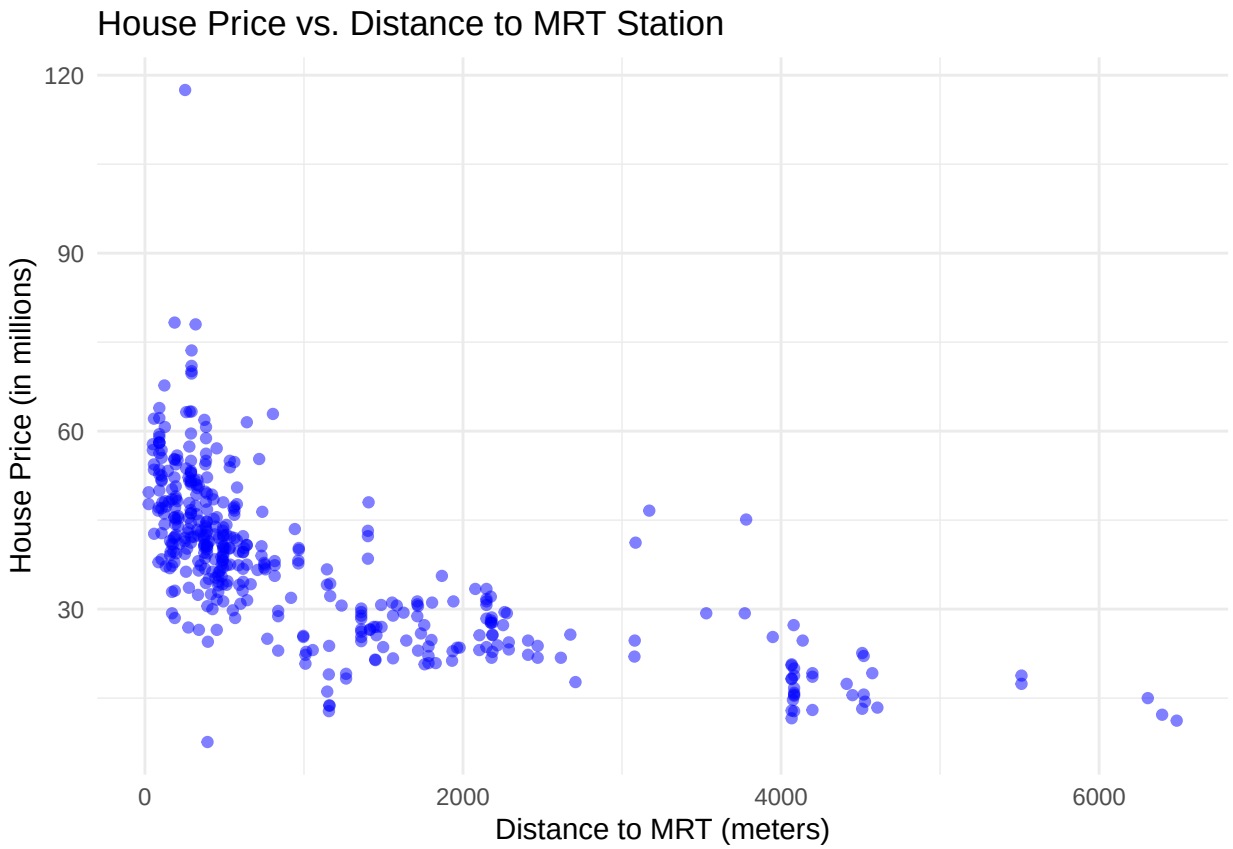
2.4 Data Visualization

We visualize the distribution of house prices and explore relationships between house prices and other key variables.

```
# Histogram of house prices
ggplot(house_data, aes(x = house_price)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of House Prices", x = "House Price (in millions)", y = "Frequency")
```




```
# Scatterplot of house price vs. distance to MRT
ggplot(house_data, aes(x = mrt_distance,
y = house_price)) +
  geom_point(color = "blue", alpha = 0.5) +
  theme_minimal() +
  labs(title = "House Price vs. Distance to MRT Station", x = "Distance to MRT (meters)", y = "House Pri
```



The histogram of house_price shows a right-skewed distribution, with most properties priced between 20 and 50 million NT$. The scatterplot between mrt_distance and house_price confirms a negative relationship, reinforcing the idea that proximity to MRT stations is a key driver of higher property prices.

3. Regression Modeling

3.1 Model Selection

We fit a multiple linear regression model to identify significant predictors of house prices. The model includes house_age, mrt_distance, convenience_stores, latitude, and longitude as explanatory variables.

```
# Fit the linear regression model
model <- lm(house_price ~ house_age + mrt_distance + convenience_stores + latitude + longitude, data = h

# Display the summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = house_price ~ house_age + mrt_distance + convenience_stores +
##     latitude + longitude, data = house_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.546  -5.267  -1.600   4.247  76.372
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -4.946e+03  6.211e+03  -0.796    0.426
```

```
## house_age          -2.689e-01  3.900e-02  -6.896 2.04e-11 ***
## mrt_distance       -4.259e-03  7.233e-04  -5.888 8.17e-09 ***
## convenience_stores  1.163e+00  1.902e-01   6.114 2.27e-09 ***
## latitude            2.378e+02  4.495e+01   5.290 2.00e-07 ***
## longitude          -7.805e+00  4.915e+01  -0.159    0.874
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.965 on 408 degrees of freedom
## Multiple R-squared:  0.5712, Adjusted R-squared:  0.5659
## F-statistic: 108.7 on 5 and 408 DF,  p-value: < 2.2e-16
```

The model includes house_age, mrt_distance, convenience_stores, latitude, and longitude as predictors. The inclusion of these variables is supported by the correlation analysis and visualizations, which showed their relationships with house_price. For instance, mrt_distance was included due to its strong negative correlation with house_price. Latitude and longitude are included to account for geographic variation, even though longitude was not significant in the final model.

3.2 Interpretation of Coefficients

We interpret the model coefficients, focusing on variables that are statistically significant at the 5% significance level (p-value < 0.05).

```
# Extract the coefficients and p-values
coefficients <- summary(model)$coefficients

# Display coefficients with significance levels
coefficients
```

```
##                         Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)        -4.945595e+03 6.211157e+03 -0.7962437 4.263534e-01
## house_age          -2.689168e-01 3.899568e-02 -6.8960665 2.044371e-11
## mrt_distance       -4.259089e-03 7.233341e-04 -5.8881350 8.166456e-09
## convenience_stores  1.163020e+00 1.902205e-01  6.1140639 2.273810e-09
## latitude            2.377672e+02 4.494802e+01  5.2898259 2.001707e-07
## longitude          -7.805453e+00 4.914891e+01 -0.1588123 8.738953e-01
```

```
# Identify variables significant at the 5% level
significant_vars <- coefficients[coefficients[, "Pr(>|t|)"] < 0.05, ]
significant_vars
```

```
##                         Estimate   Std. Error   t value     Pr(>|t|)
## house_age          -0.268916833 3.899568e-02 -6.896067 2.044371e-11
## mrt_distance       -0.004259089 7.233341e-04 -5.888135 8.166456e-09
## convenience_stores  1.163020477 1.902205e-01  6.114064 2.273810e-09
## latitude          237.767190977 4.494802e+01  5.289826 2.001707e-07
```

-House Age ($p < 0.001$): Each additional year of house age decreases the price by approximately 0.269 million NT$.

-MRT Distance ($p < 0.001$): Each additional meter of distance from the nearest MRT station decreases the price by 0.0043 million NT$.

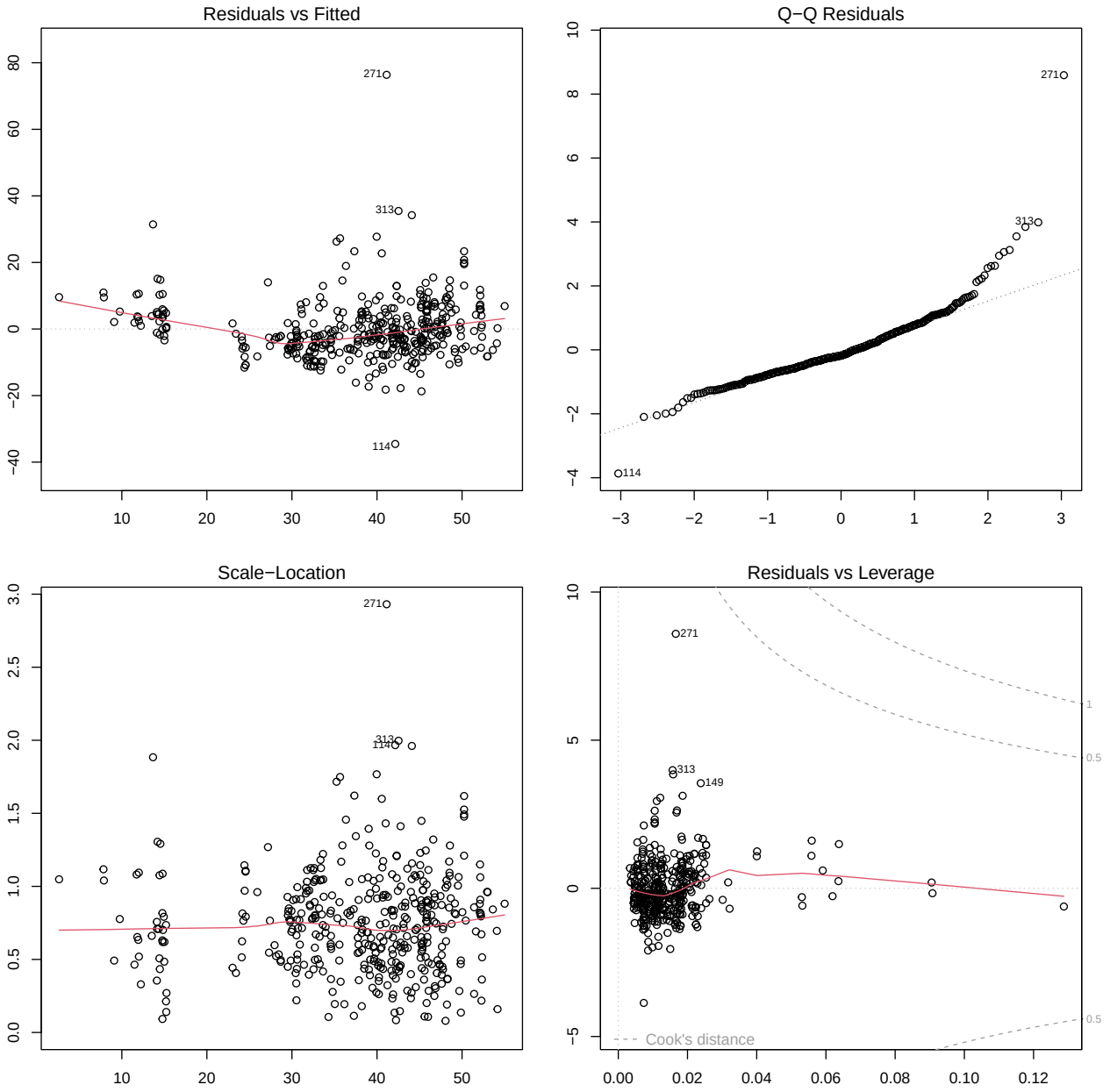
-Convenience Stores ($p < 0.001$): Each additional convenience store nearby increases the price by 1.163 million NT$.

-Latitude ($p < 0.001$): Moving further north (increasing latitude) is associated with higher house prices.

3.3 Model Diagnostics

We check the assumptions of linear regression, such as normality of residuals and homoscedasticity, to validate the model.

```
# Adjust plot margins to avoid "figure margins too large" error
par(mfrow = c(2, 2), mar = c(3, 3, 2, 1) + 0.1)
plot(model)
```



-Residuals vs Fitted: There is no clear pattern, suggesting homoscedasticity.

-Normal Q-Q Plot: Residuals follow a straight line, indicating that they are approximately normally distributed.

-Scale-Location Plot: Variance appears consistent across fitted values.

-Residuals vs Leverage Plot: There are no high-leverage points or influential outliers.

4. Results and Discussion

4.1 Summary of Findings

```
# Calculate and display the R-squared value
r_squared <- summary(model)$r.squared
cat("R-squared: ", r_squared)
```

```
## R-squared:  0.5711617
```

The regression model identifies significant predictors of house prices in Taiwan at the 5% significance level. The key findings include:

-House Age: Older houses tend to have lower prices.

-Distance to MRT: Houses closer to MRT stations have higher prices.

-Convenience Stores: More convenience stores nearby are associated with higher house prices.

-Latitude: Higher latitudes (moving north) are associated with higher house prices.

The overall R-squared value of the model is 0.571, indicating that approximately 57% of the variance in house prices is explained by the model.

4.2 Discussion

The regression analysis reveals key factors influencing house prices in Taiwan.

-House age has a significant negative impact, with each additional year reducing the price by 0.269 million NT, $reflecting the common depreciation seen in older properties. Proximity to MRT stations is highly valued, as indicated by the$ per meter), highlighting the premium placed on easy access to public transport in urban areas.

-Local amenities also play a crucial role; each additional convenience_store nearby increases the price by 1.163 million NT$, emphasizing the importance of neighborhood services in property valuation. Additionally, properties at higher latitudes (closer to Taipei) command higher prices, consistent with the trend of higher real estate values in the north.

However, the R-squared value of 0.571 suggests that other unaccounted factors, such as property size, building quality, or socio-economic conditions, may also significantly influence house prices.

4.3 Limitations

While the model provides valuable insights, it assumes linear relationships between the predictors and the house prices. Non-linear effects or interactions between variables may exist but are not captured in this model.

5. Conclusion

This analysis identifies key factors that significantly influence house prices in Taiwan. The findings highlight the importance of house age, proximity to MRT stations, and the availability of nearby convenience stores. These insights can guide real estate investors and policymakers in making informed decisions. Further research could explore non-linear models or additional variables to improve predictive accuracy.