Real Estate Valuation in Taiwan

DSC 241: Statistical Models - Project

AUTHOR
Harley Clifton & Tristan Cooper

PUBLISHED
May 5, 2025

Introduction

Problem Statement

Predicting housing prices is a valuable and practical application of data analysis. In this project, we use a historical real estate dataset collected from Sindian District in New Taipei City, Taiwan. The dataset includes variables such as transaction dates, the age of the house, distance to the nearest MRT station, the number of nearby convenience stores, as well as latitude and longitude coordinates. Our goal is to explore whether a combination of these features can be used to build a model that accurately predicts housing prices.

Research Question

Can a combination of property features accurately predict the unit price of real estate properties?

Variables of Interest

Response Variable: House price per unit area (Y)

Predictor Variables: Transaction date (X1), House age (X2), Distance to the nearest MRT station (X3), Number of convenience stores (X4), Latitude (X5), and Longitude (X6)

Prediction Goal

Develop a predictive model to estimate property prices based on available features and evaluate its accuracy

Data Overview

Variable Descriptions

The dataset was sourced from the UC Irvine Machine Learning Repository, which hosts a wide range of curated datasets for research and educational purposes. A detailed summary of the dataset variables, including their units and data types, is provided in the table below.

Project Variable Metadata

Varibles	Descriptions	Type	Units
Transaction Date	Date in the formate yyyy.### where ### indicate the number of the month divided by 12; for example, 2013.250=2013 March, 2013.500=2013 June, etc.	real number	
House Age	Age of the house in years, rounded to nearest tenth.	real number	year
Distance to nearest MRT Station	Distance in meters	real number	meter
Number of Convience Stores	Number of convience stores within 500 meters of the house	nonnegative integer	
Latitiude	geographic coordinate, latitude	real number	degree
Longitude	geographic coordinate, longitude	real number	degree
House Price (per unit area)	10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared	real number	10000 New Taiwan Dollar/Ping

Data Wrangling

We first observed that the transaction date variable is formatted in an unusual way. It appears as `YEAR.###`, where the decimal portion is not immediately intuitive. Upon reviewing the codebook, we found that the digits following the decimal represent the month as a fraction of the year (i.e., the month number divided by 12). For example, January is represented as 1/12 = 0.0833, so January 2021 appears as `2021.083`. Similarly, December would be 12/12 = 1.0, making December 2021 appear as `2022.000`.

To address this, we converted these values into standard date objects in the format `YYYY-MM-DD`, defaulting to the first day of the month (`DD` = 01) since day-level information is not available.

Exploratory Data Analysis

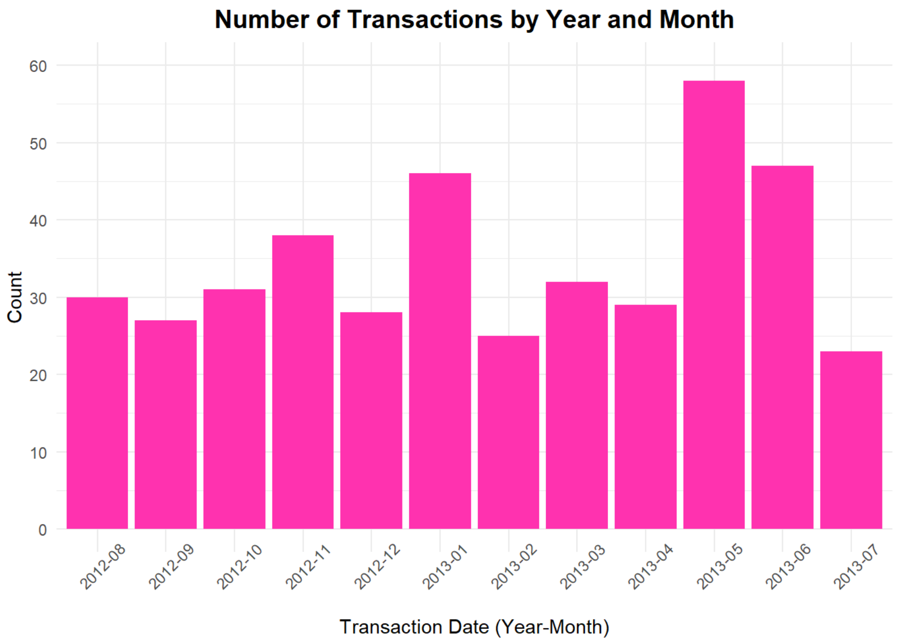
Exploratory data analysis (EDA) was carried out to examine the underlying structure of the dataset, supported by informative tables and figures.

Missing Data

We began our exploratory data analysis by visualizing missing values to assess data completeness and identify any potential issues in the dataset. Interestingly, we found that there are no missing values-an unexpected but welcome result (see Appendix for the missingness visualization). Since there is no missingness to address, we proceed by creating univariate visualizations to explore the distribution of individual variables.

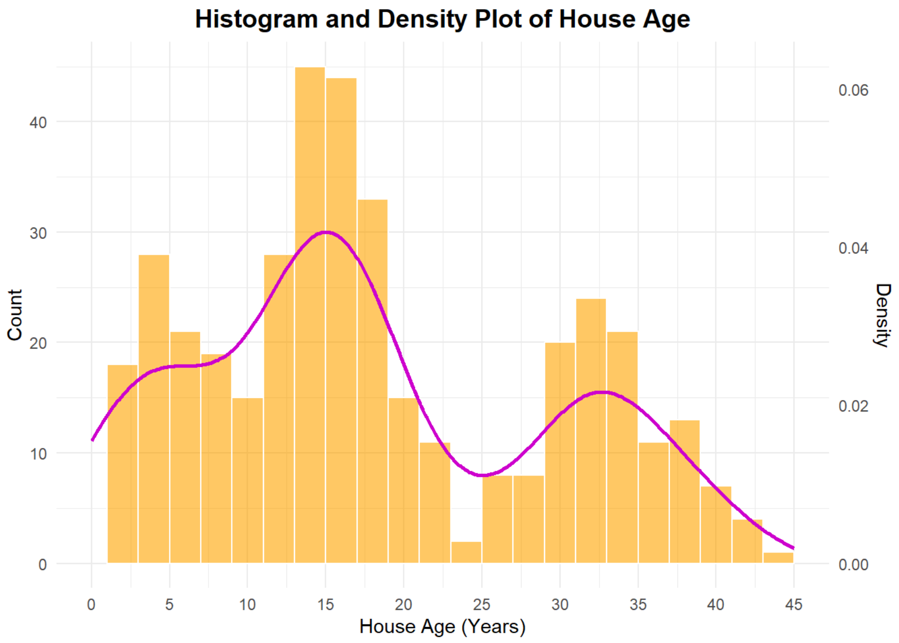
Univariate Graphs

We begin by visualizing the predictor variables using appropriate plots to better understand their distributions. The first variable of interest is **transaction date**, for which we created a bar plot to display the frequency of transactions over time.



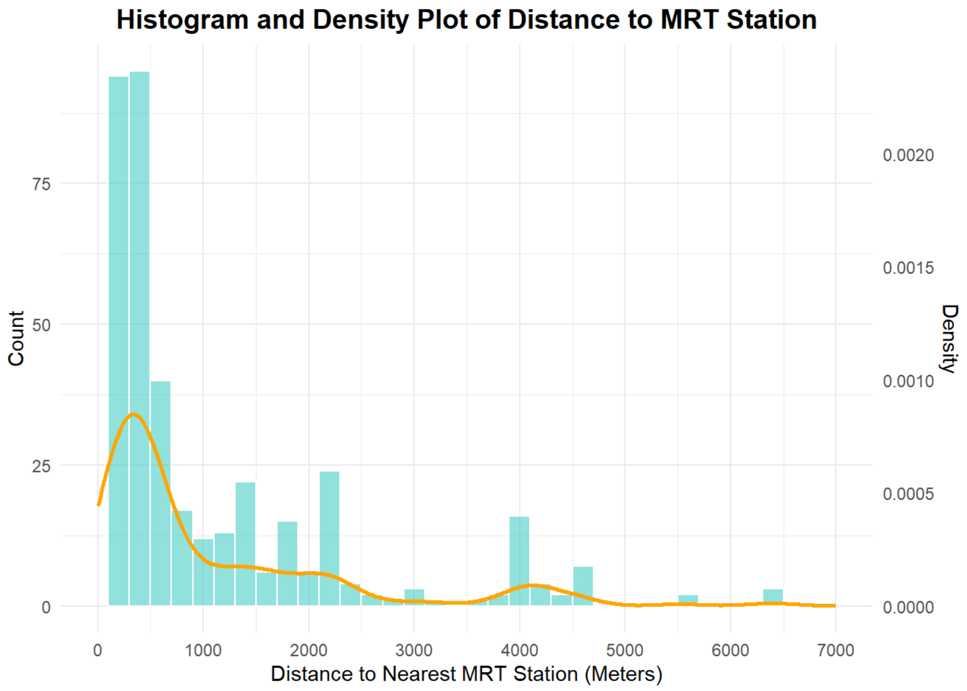
Transaction dates range from August 2012 to July 2013, with at least 20 observations for each month. There are no suspected outliers. The monthly number of housing transactions fluctuates, indicating some seasonality or irregular patterns. May 2013 had the highest number of transactions, with January 2013 and June 2013 also fairly close. Fewer transactions occurred in February 2013 and July 2013, possibly due to seasonal factors or holidays.

The next predictor of interest is **house age**, measured in years. To explore its distribution, we visualize this variable using a combined histogram and density plot, which allows us to examine both the frequency of observations and the overall shape of the distribution.



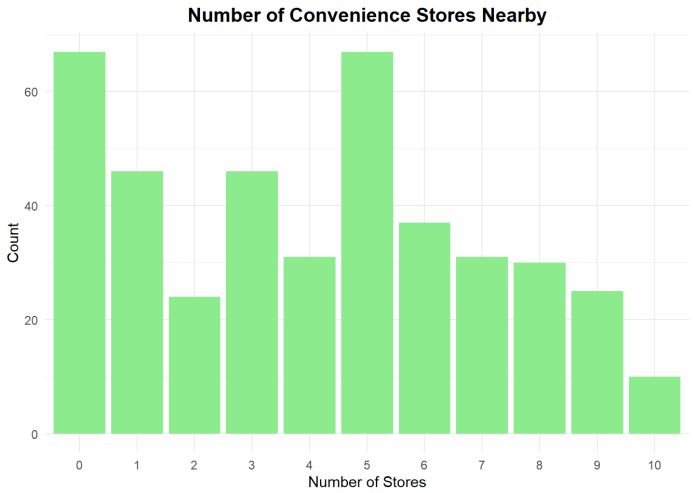
The distribution of House Age appears to be very irregular, but there is little suspicion of outliers. The density curve suggests a multimodal distribution, possibly indicating multiple waves of housing developments. The largest concentration of homes appears to be around 15 to 20 years old. There are relatively fewer houses under 5 years old or older than 40 years, suggesting limited recent construction and fewer retained very old properties. Given its spread and variability, house age may have a non-linear relationship with housing price, and might benefit from transformation or binning in predictive modeling.

Next, we examine the variable **distance to the nearest MRT station**. To explore its distribution, we use a combined histogram and density plot, which allows us to visualize both the counts and distribution.



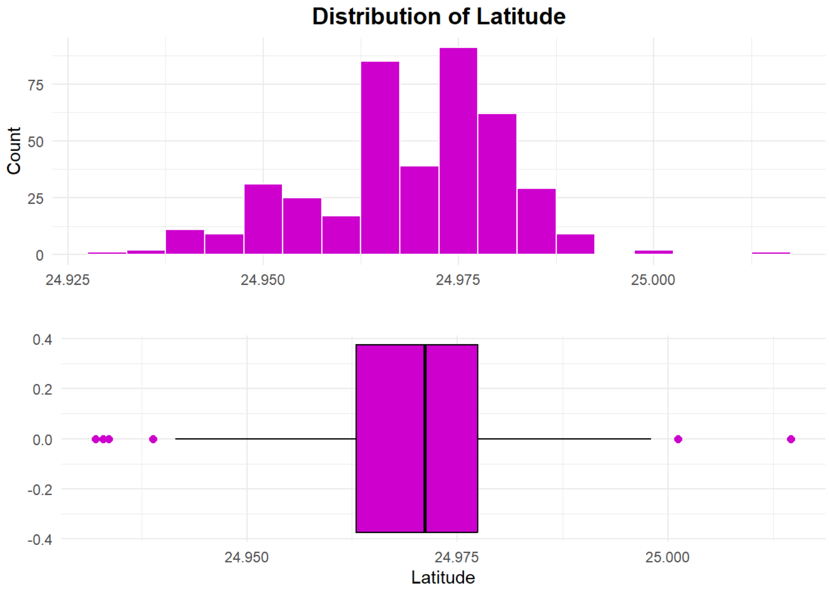
The distribution is heavily right-skewed, indicating that most houses are located close to MRT stations, with a long tail extending toward greater distances. A large concentration of properties (over half) are located within 500-1000 meters of an MRT station, suggesting strong demand or planning around transit accessibility. It is less common for homes to be located much farther away (up to 7000 meters). Due to its skewness, this variable may benefit from log transformation or binning when used in predictive models to reduce the impact of extreme values.

Next, we visualize the **number of convenience stores** within 500 meters of each house. Since this is a discrete count variable, we use a bar plot to display the frequency distribution across different store counts.



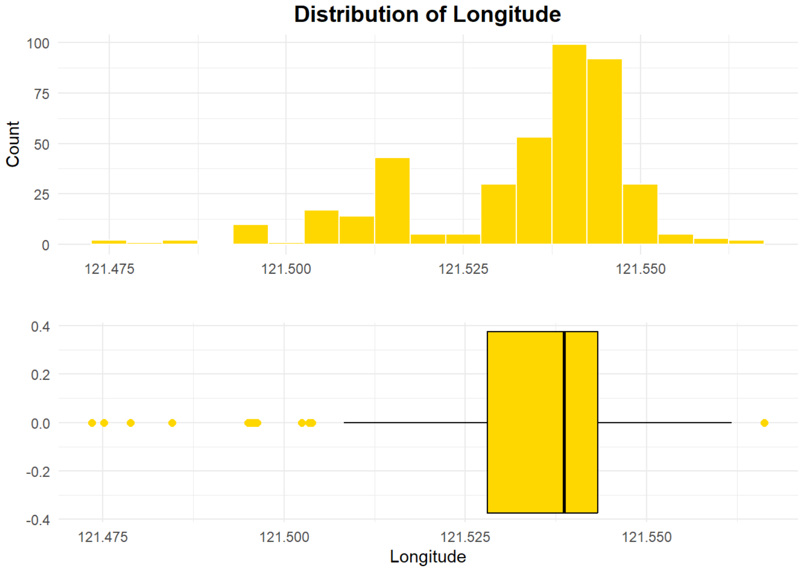
The distribution appears bimodal, with notable peaks at 0 and 5 convenience stores. This suggests two common housing contexts: (1) homes in less commercialized areas, and (2) homes in highly accessible urban zones. The number of nearby stores ranges from 0 to 10. The most frequent values are 0 and 5 stores, each with over 60 observations. Fewer homes are located near 2 or 10 stores, which may be edge cases in suburban or dense commercial zones, respectively.

The next variable we examine is **latitude**, which represents the north-south geographic position of each property. We visualize its distribution using a histogram and horizontal boxplot to observe how the properties are spatially distributed along this dimension.



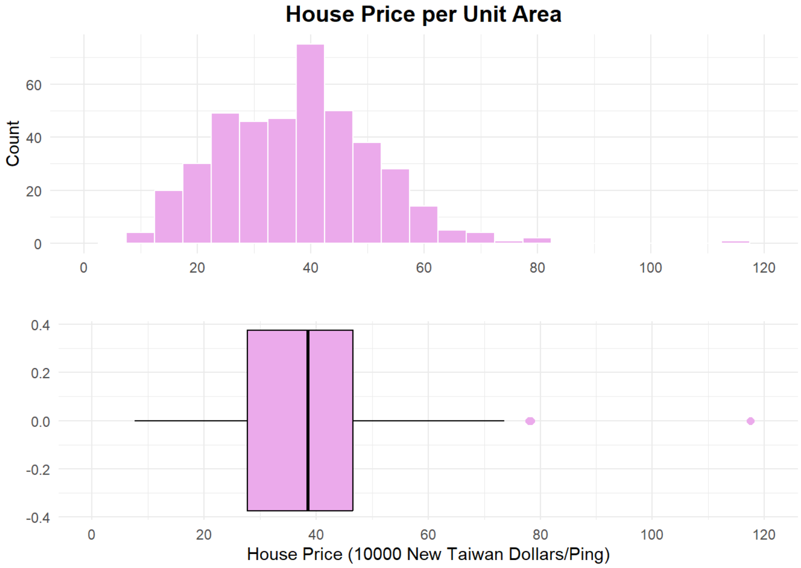
Most properties fall within a very narrow latitude range, indicating a relatively compact geographic area in terms of north-south spread. The histogram is skewed right, with the highest density of homes located around 24.97 degrees latitude. The boxplot indicates several outliers on both ends of the latitude range.

Next, we examine the **longitude** variable, which represents the east-west geographic position of each property. We visualize its distribution using a combined histogram and horizontal boxplot.



Most properties are located within a narrow longitude band, indicating limited east-west spread and suggesting a fairly localized study area. The histogram shows a moderate right skew and the boxplot reveals multiple outliers, especially on the lower end, meaning a few properties are located farther west than the main cluster. There's a noticeable spike in the number of properties at the center of the distribution, reinforcing that much of the housing data is concentrated in a specific area.

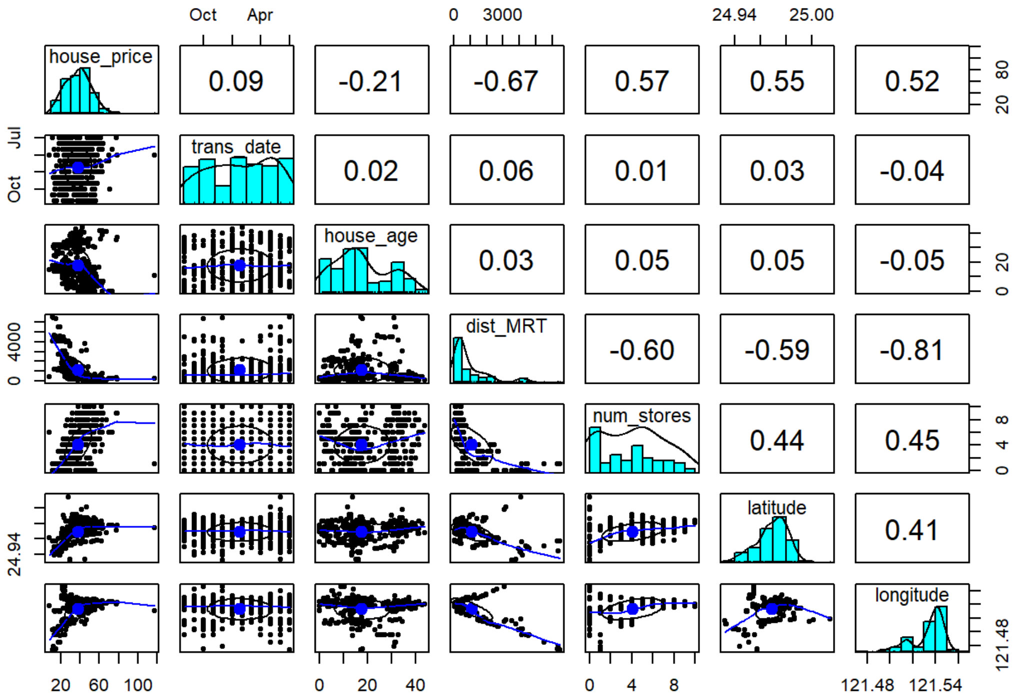
Next, we visualize our response variable, which is **house price per unit area**, measured in units of *10,000 New Taiwan Dollars (NTD) per Ping* where 1 Ping is equivalent to 3.3 square meters. Again, we use a combined histogram and boxplot.



The histogram shows a nearly bell shaped distribution centered around 40-45 (10,000 NTD/Ping), with a slight right skew, indicating some higher-priced properties pulling the tail to the right. The majority of homes are priced between 30-50 (10,000 NTD/Ping), suggesting this is the typical market rate in the sampled region. The boxplot highlights a few notable high outliers, suggesting luxury or premium real estate. While the distribution is fairly compact, the outliers may affect modeling.

Pairsplot & Correlation Matrix

To conclude our exploratory data analysis, we will generate a pairs plot and a correlation matrix to examine the relationships between all variables in the dataset. These visualizations will help identify potential multicollinearity, detect linear associations, and inform future feature selection for modeling.



Key takeaways from the Pairsplot & Correlation Matrix:

1. Distance to Nearest MRT Station is negatively correlated with house price (r = -0.67). This suggests that closer proximity to MRT stations is strongly associated with higher house prices-a potentially key predictor.
2. The number of nearby convenience stores is positively correlated with house price (r = 0.57), possibly reflecting neighborhood accessibility and desirability.
3. Latitude and longitude show moderate positive correlations with house price (r = 0.55 and r = 0.52, respectively). This indicates a spatial pattern-certain geographic locations (likely more central or desirable) are associated with higher property values.
4. *Multicollinearity Warning*: Latitude and longitude are moderately correlated with each other (r = 0.41-0.45).

Data Analysis

Modeling

Since our primary goal is prediction, as stated in our research question, we will split the data into training and testing sets. The training set will contain 70% of the data and will be used to build and tune the model, while the testing set will include the remaining 30% and will be used to evaluate the model's performance on unseen data. This approach helps ensure a more accurate and generalizable assessment of predictive accuracy.

The dataset was randomly split into training and testing subsets. The training set contains 289 observations (approximately 70%), which will be used to build the predictive model, while the testing set includes 125 observations (approximately 30%), which will be used to evaluate model performance on unseen data.

Fitting a Simple Linear Regression

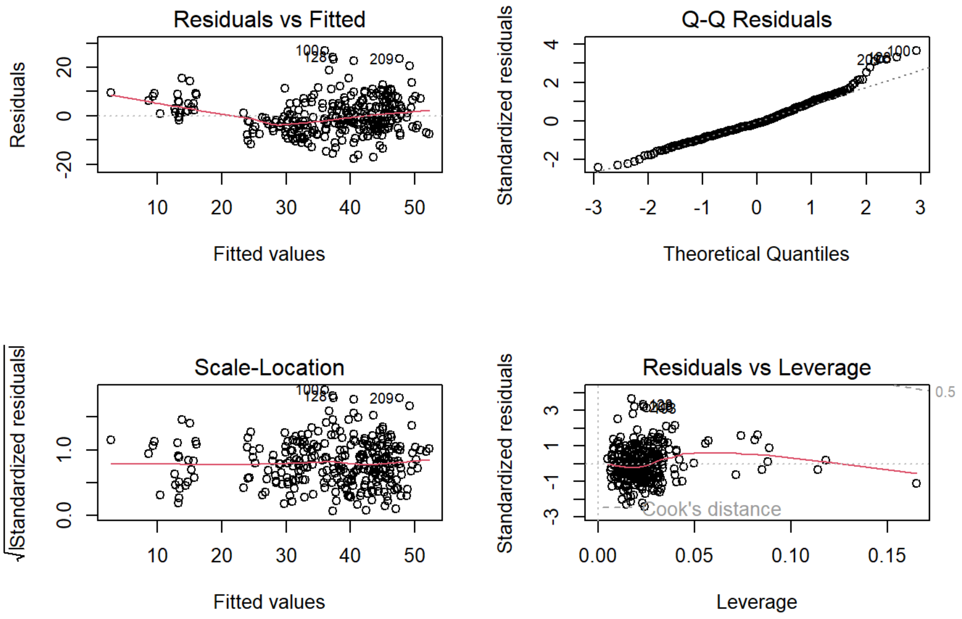
Next, we fit a simple linear regression model on the training data, using house price per unit area as the response variable. All available predictor variables were included additively, allowing us to assess the individual contribution of each feature to the predicted housing price. The model summary is displayed below.

Linear Regression Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	−9,956.392	5,792.356	−1.720	0.087
trans_date	0.009	0.004	2.080	0.038
house_age	−0.214	0.039	−5.470	0.000
dist_MRT	−0.004	0.001	−5.490	0.000
num_stores	1.043	0.192	5.430	0.000
latitude	252.379	43.093	5.860	0.000
longitude	29.246	45.876	0.640	0.524

Initial Model Diagnostics

To assess the validity of the model's underlying assumptions, we generated the standard suite of diagnostic plots, including those for residual normality, homoscedasticity, linearity, and influential observations.



Linearity: In the Residual vs. Fitted plot, the red line deviates from the zero line, indicating the presence of missed curvature and clear evidence against the linearity assumption.

Homoskedacity: In the Residual vs. Fitted plot, there is an increasing fanning pattern present. This indicates a clear violation of the Equal Variance assumption. Further, since there is a clear violation in that plot, we are unable to further assess the Scale-Location plot for this assumption.

Normality: In the Normal Q-Q Plot, there is a very heavy right tail - which is very problematic. This suggests that our residuals are skewed right, which means a clear violation of the Normality Assumption.

Outliers: In the Residuals vs. Leverage Plot, there are no obvious points that are beyond the Cook's Distance Boundary of 0.5, indicating no obvious violations or evidence of outliers.

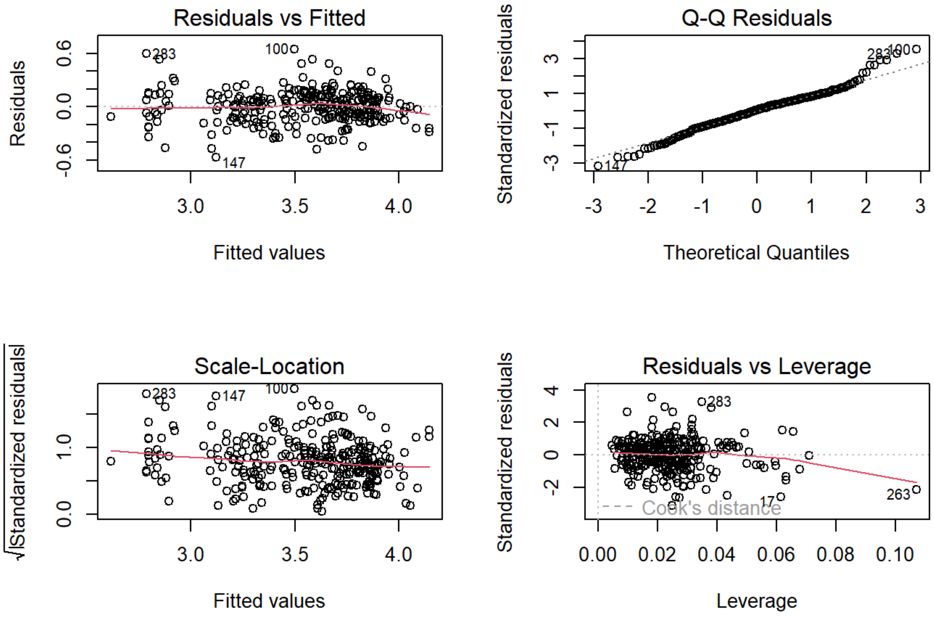
Since the diagnostic plots reveal clear violations of the assumptions of linearity, homoscedasticity (equal variance), and normality of residuals, we will explore transformations of the variables in an effort to improve model fit and better satisfy these assumptions.

Transforming Variables

We explored three common transformations: (a) log, (b) reciprocal, and (c) square root, each applied solely to the response variable (house price per unit area). However, none of these transformations yielded meaningful improvements in the model diagnostics (see Appendix for details).

Based on earlier exploratory analysis and visualizations, we observed that the distance to the nearest MRT station was heavily right-skewed, suggesting it may benefit from a log transformation to reduce the influence of extreme values. Therefore, we proceeded to fit a new model that applied a log transformation to both the response variable and the distance to the nearest MRT station.

Compare Dianostic Improvements



The transformation resulted in a notable improvement in linearity, as the red line now more closely follows the zero line.

When assessing equal variance, a slight diamond-shaped pattern remains (characterized by increasing and then decreasing spread in the residuals), but this still represents a modest improvement compared to the untransformed model.

Regarding normality, the heavy tails in the residual distribution have diminished in severity and are now more symmetrically distributed, rather than being concentrated on one side.

Compare R-Squared & Variance of Coefficients

To supplement our diagnostic analysis, we compared the original model with the transformed model in terms of R^2, coefficient variance, and interpretation of the coefficients. This comparison provides additional insight into model performance and stability, and supports our evaluation of whether the transformation meaningfully improved the model.

Comparison of Original and Transformed Linear Models

Model Type	R^2	Avg. Coef. Variance	Coefficient Interpretation
Original Model	0.635	4793622.030	Estimated coefficients represent additive change in raw price per unit area
Transformed Model	0.758	1652.004	Estimated coefficients represent multiplicative effects on log(price), interpreted as percent changes

R-Squared: The original model has an adjusted R-squared value of 0.635, and the transformed model has an adjusted R-squared of 0.758; this is a big improvement in variability explained by the predictors in the model.

Variance of Coefficients: The variance for the estimated coefficients are all smaller for the transformed model with the exception of the distance to the nearest MRT station.

Interpreting Coefficients: Since we took a log transformation of the response, the coefficients now represent the change in the *log of the house price* per unit increase for each predictor.

Since we took the log of the distance to nearest MRT station, that coefficient represents the increase in the log house price per unit increase in the log of the distance to nearest MRT station.

Overall, we consider this transformation to be a modest but meaningful improvement over the original model, as it resulted in better diagnostic performance, a higher R^2, and more stable standard error estimates for the coefficients.

Model Selection

We then applied feature selection to the transformed model using the `dredge()` function, which generates models with all possible combinations of predictors and ranks them based on their Akaike Information Criterion (AIC). Although AIC values are not meaningful in isolation, they are useful for comparing models; with lower AIC values indicating a better trade-off between model fit and complexity. This approach helps identify the most parsimonious (simple) model that still explains the data well.

Final Model

The best-performing model, with an AIC of -149.2, uses log-transformed house price as the response variable and includes all predictors from the transformed model. Based on both diagnostic improvements and model selection criteria, this model is considered the most appropriate for capturing the relationships in the data. The final estimated model is written below:

$$\begin{gathered}\log(\text{HousePrice}) = -674.3 + 0.0003771 \cdot \text{TransactionDate} - 0.005122 \cdot \text{HouseAge} \\ -0.1618 \cdot \log(\text{DistanceToMRT}) + 0.01067 \cdot \text{ConvenienceStores} \\ +10.08 \cdot \text{Latitude} + 3.468 \cdot \text{Longitude}\end{gathered}$$

Interpreting Model Coefficients

Intercept: When the distance to the nearest MRT station is 1 and all other predictors are 0, the log house price per unit area is -674.3.

Transaction Date: A one-day increase in transaction date is associated with a 0.0003771 increase in log house price per unit area.

House Age: A one-year increase in house age is associated with a 0.005122 decrease in log house price per unit area.

Log distance to nearest MRT station: A one-unit increase in the log of the distance to the nearest MRT station is associated with a 0.1618 decrease in log house price per unit area.

Number of nearby convenience stores: For every additional nearby convenience store, there is an associated 0.01067 increase in log house price per unit area.

Latitude: For each one unit increase in latitude, the log house price increases by 10.08 times the latitude increase.

Longitude: For each one unit increase in longitude, the log house price increases by 3.468 times the latitude increase.

Evaluating Predictive Ability

While interpreting model coefficients provides insight into the direction and relative importance of each predictor, the primary goal of our study is not interpretation but prediction - to assess how well the model can forecast housing prices for new data. To evaluate the predictive ability of our final, transformed model, we shift focus to performance metrics that reflect how accurately the model generalizes to unseen observations.

Specifically, we will assess the model using three standard metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2). RMSE gives greater weight to larger errors and indicates the typical magnitude of prediction error. MAE provides the average absolute deviation between predicted and actual prices, offering a more intuitive interpretation of error magnitude. R^2, on the other hand, reflects the proportion of variance in housing prices explained by the model, giving us a general sense of its predictive strength. The values of these metrics for our final transformed model are displayed in the table below.

Predictive Performance on Test Set

Metric	Value
Root Mean Squared Error (RMSE)	10.720
Mean Absolute Error (MAE)	6.100
R-squared	0.582

The results of our evaluation show that the final model achieves an **RMSE of 10.720**, meaning that, on average, the predicted house price per unit area deviates by about 10.72 units (in 10,000 NTD per Ping) from the actual value.

The **MAE of 6.100** indicates that the average absolute prediction error is relatively modest, providing further support for the model's reliability.

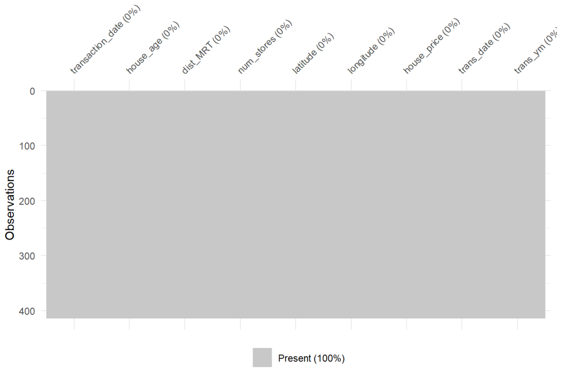
The model's R^2 **of 0.582** indicates that approximately 58.2% of the variance in housing prices is explained by the predictors in the model. While there is room for improvement, these results demonstrate that the model captures meaningful patterns in the data and provides a reasonable level of predictive accuracy.

Conclusion

In this project, we set out to answer the question: Can a combination of property features accurately predict the unit price of real estate properties? Through careful data wrangling, exploratory analysis, model fitting, and diagnostic evaluation, we developed a log-transformed linear regression model that demonstrated reasonable predictive accuracy. By incorporating features such as distance to the nearest MRT station, house age, number of nearby convenience stores, and geographic location, the final model explained approximately 58% of the variance in housing prices and produced relatively low error metrics on unseen data. While there is room for improvement (particularly in capturing nonlinear effects or incorporating additional contextual variables) our findings suggest that property features can indeed serve as a meaningful basis for predicting unit price. This supports the use of interpretable statistical models in real estate valuation and lays the groundwork for more advanced modeling approaches in future work.

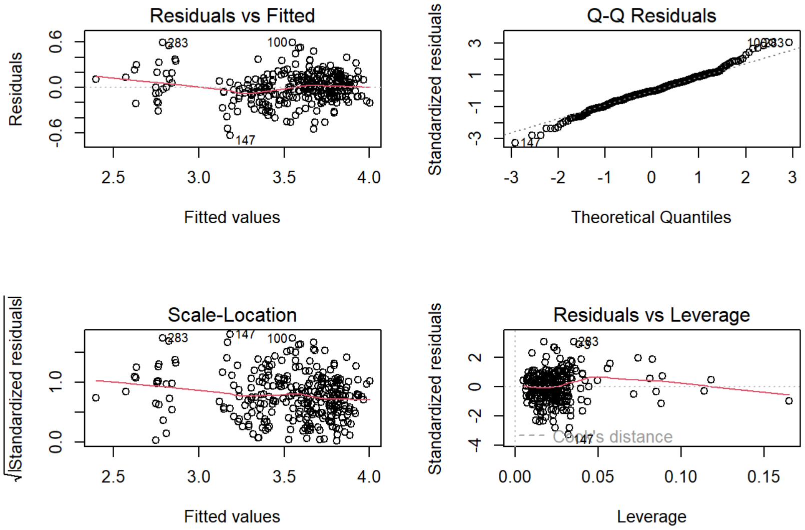
Appendix

Missing Data Visualization



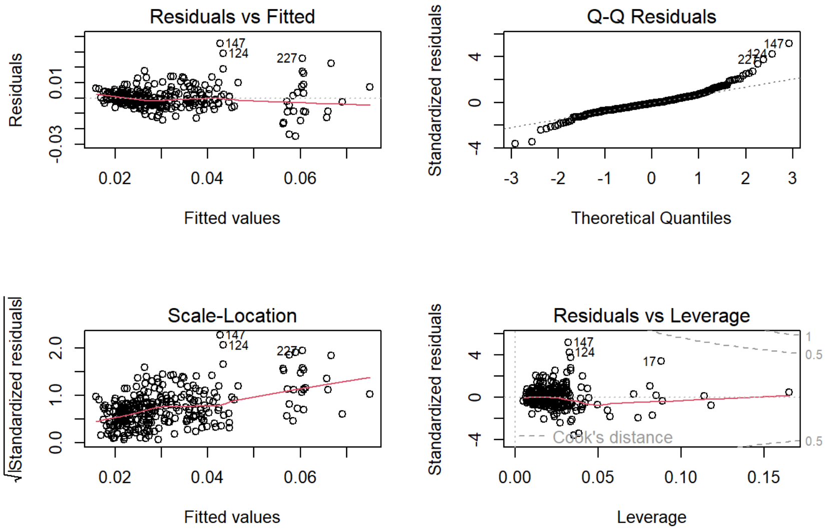
Other Varaible Transformations We Explored

Log Transformation of the Response:



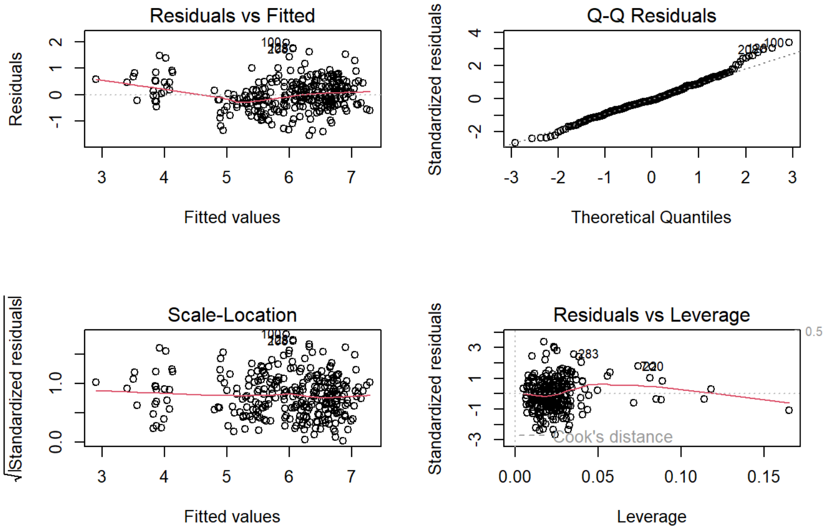
Linearity was slightly improved, still issues with Equal Variance and Heavy tails.

Reciprocal Transformation of the Response:



Weird stuff going on with Equal Variance, no noticeably change in Normal Q-Q plot.

Square Root Transformation of the Response:



Not much improvement with Linearity, Equal Variance, or in Normal Q-Q plot.

Contribution Statement

Tristan Cooper:

- Located Info for Project Dataset and Variable Descriptions
- Fit Simple Linear Regression Model
- Variable Transformations (in report and appendix)
- Compared Training Models: Variance of Coefficients
- Compared Training Models: Interpretations of Coefficients
- Interpretations of all Selected Model's Coefficients

Harley Clifton:

- Organized Project Variable Descriptions in Table Format
- Data Wrangling
- Missing Data Visualization (and comments)
- Univariate Graphs (and comments)
- Pairs Plot & Correlation Matrix (and comments)
- Split Data into Training and Testing Sets
- Initial Model Diagnostics
- Transformed Model Diagnostics
- Compared Training Models: R-Squared Values
- Model Selection with `dredge` function (and comments)
- Estimated Model in LaTeX
- Evaluated Predictive Ability
- Conclusion