# Real Estate Evaluation

## I. Introduction

In the dynamic landscape of the real estate market, especially with the current economic climate worldwide, accurate property valuation has become a crucial factor for buyers, sellers, investors, and policy makers alike.

Our aim is to provide a comprehensive understanding of the intricate factors influencing property prices. As such, our research aims to answer the question: **What factors significantly influence house prices in the study area, and to what extent do these variables contribute to the observed pricing patterns?**

Our analysis seeks to offer valuable insights into the housing market dynamics and aid stakeholders in making informed decisions. By identifying the key determinants of house prices, we aim to enhance transparency in real estate transactions. Understanding these factors can also help policymakers formulate effective housing policies that are tailored to the needs of the community.

Our data was taken from the UC Irvine Machine Learning Repository. (https://archive.ics.uci.edu/dataset/477/real+estate+valuation+data+set)
The data was originally collected from the public database of the Ministry of the Interior from two districts in Taipei City and two districts in New Taipei City, resulting in four datasets consisting of 414 records of real estate sales spanning from September 2012 to August 2013.

The data includes various factors influencing residential housing prices in Taipei City and New Taipei City.

Each observation in the provided data set has 6 continuous explanatory variables, which were also referred to as appraisal factors, and one continuous response variable.

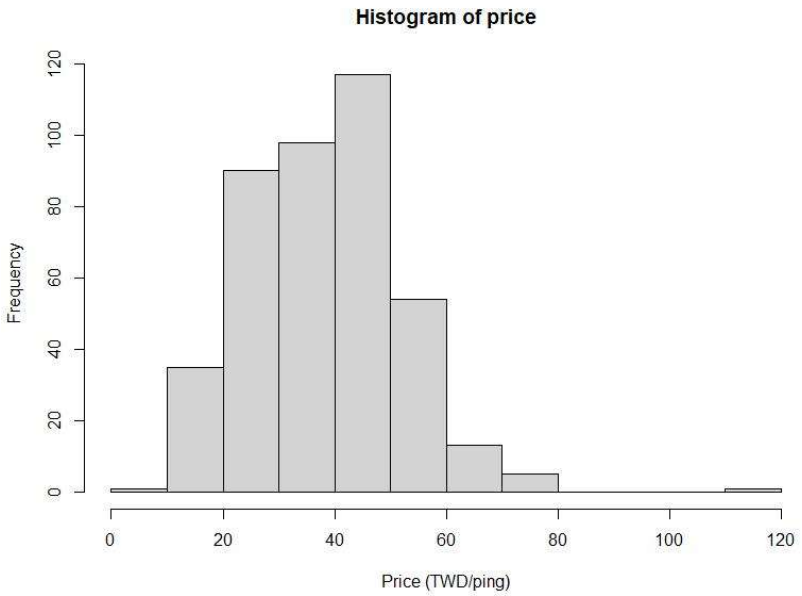
1. **Transaction Date** (date that the house was sold, for example, 2013.250 = 2013 March, 2013.500 = 2013 June, etc.): This variable is represented as a real number and accounts for the effect of market conditions on house prices.
2. **House Age** (unit: years): The age of the house at the time of transaction, measured in years, which affects both depreciation and living quality.

3. **Distance to the Nearest MRT Station** (unit: meter): This variable is calculated by determining the location coordinates of the house and the MRT stations, and then finding the nearest MRT station through minimization operations.
4. **Number of Convenience Stores in the living circle on foot** (unit: meter): This variable is determined by counting the number of convenience stores within a 500-meter radius of the house, calculated based on location coordinates.
5. **Geographical Coordinate - Latitude and Longitude** (unit: degree) These two variables represent the location coordinates of the house, which influence factors such as the distance to downtown and the associated time and cost implications.

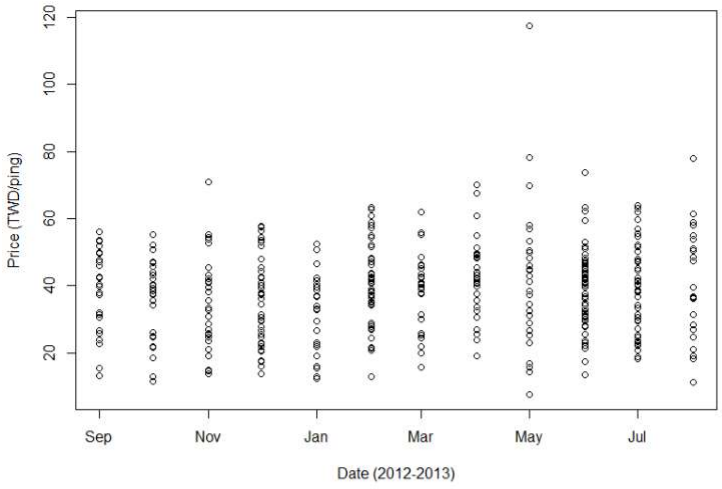
Our response variable is **house price per unit area** (New Taiwan Dollars per Ping): the ping (坪) is a local unit of area equivalent to approximately 3.3 square meters or 36 square feet.

## AI. Analysis

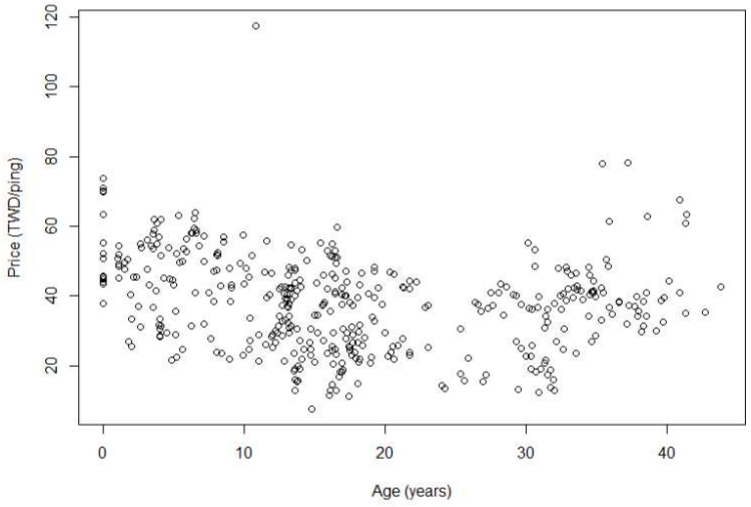
The prices per unit area of properties in our data range from 7.6 to 117.5 TWD/ping, with a mean of 38.0. It appears that the relationships between price and building age, and between price and distance to the nearest MRT station, are not linear. The observations with price 7.6 and price 117.5 may in fact be outliers, as they are quite far outside the typical range of prices for properties with similar values of the explanatory variables. Following are plots exploring the distribution of price and its possible relationships with the explanatory variables.



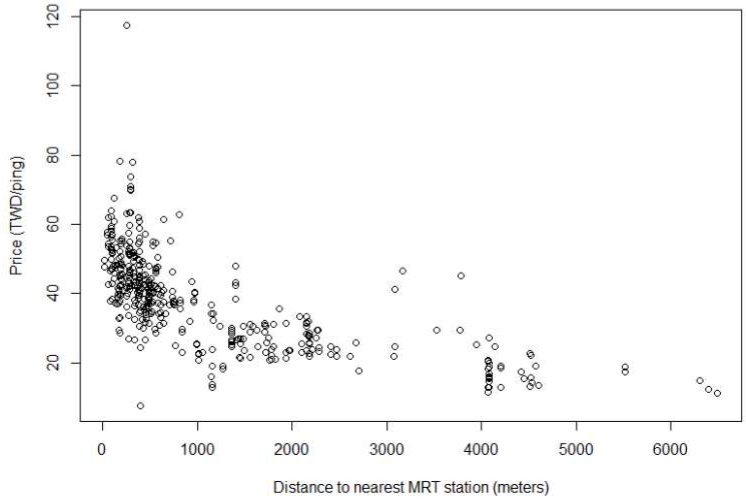
**Price vs. date**

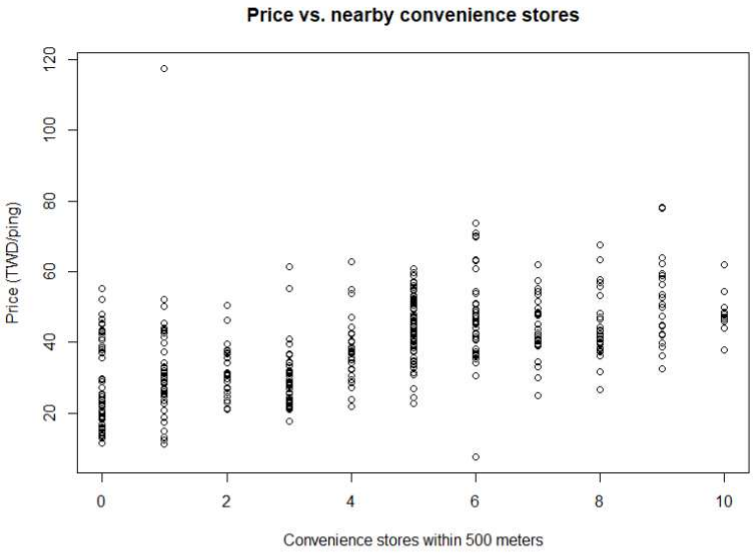


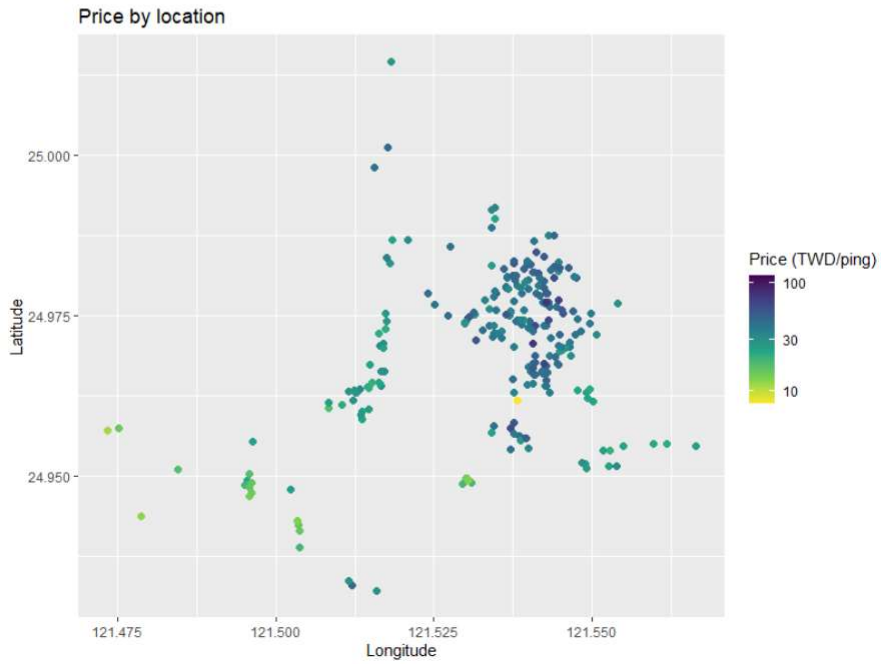

**Price vs. age**



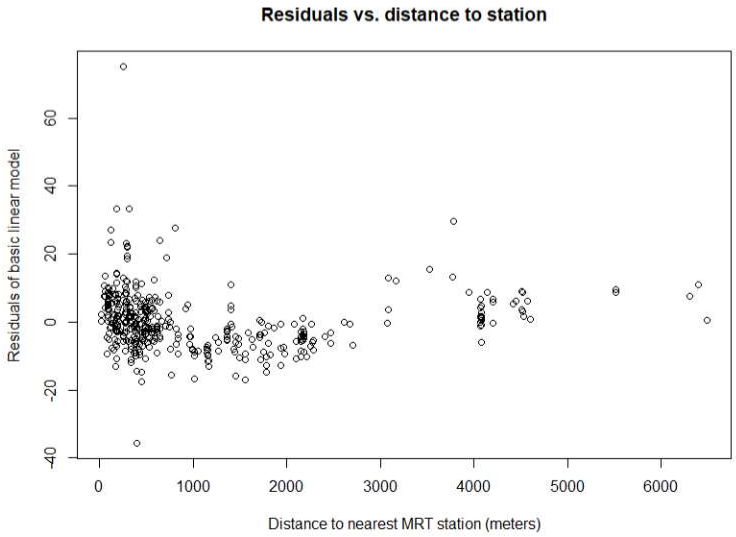

**Price vs. station distance**



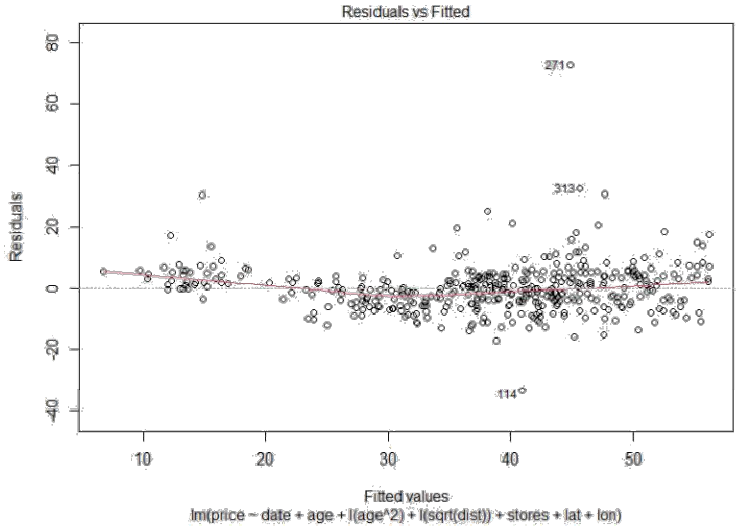


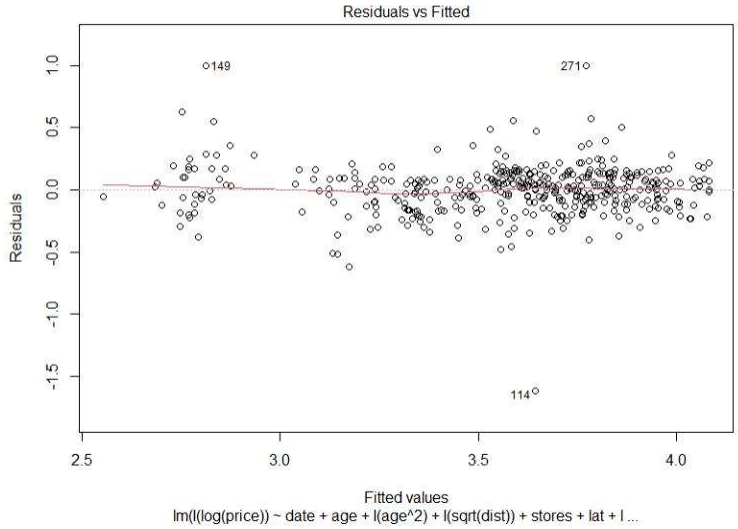

A basic linear model including all explanatory variables was fitted to the data, but we deemed this model to be inappropriate since the values of the response variable appear to depend non-linearly on both age and the distance to the nearest MRT station, even when controlling for the other explanatory variables.


After applying a square root transformation to the distance to the nearest transit station and fitting a model allowing a quadratic relationship between price and age, it became clear that not only was there another source of non-linearity in the data, the assumption of homoscedasticity was not satisfied.




Transforming price to the logarithm of price led to a better-fitting model that appeared to better satisfy the assumption of constant error variance.

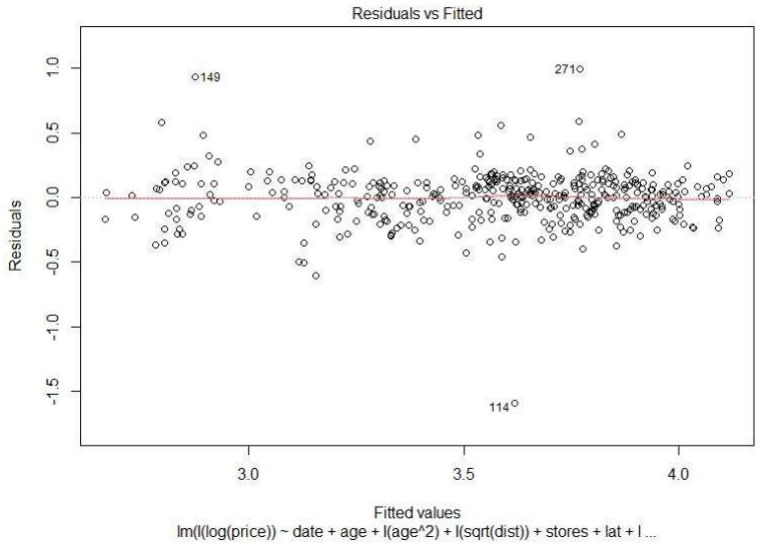



It appeared that there was still some source of non-linearity in the data, so a backward model search was performed, starting with a full model allowing for quadratic terms on all of the variables and interaction terms between most pairs of variables. Removing at each step the term with its respective t-test having the highest p-value until all p-values were less than 0.05 led to a model with terms for intercept, date, age, the square of age, the square root of distance to the nearest MRT station, the number of nearby convenience stores, the latitude, and an interaction term between latitude and longitude. Since the last term to be removed before arriving at this reduced model was the longitude with a marginal P-value of 0.0545, two models were considered: the one with the terms listed above, and the one also including a term for longitude. Since the model including a term for longitude as well as the above terms had slightly higher adjusted

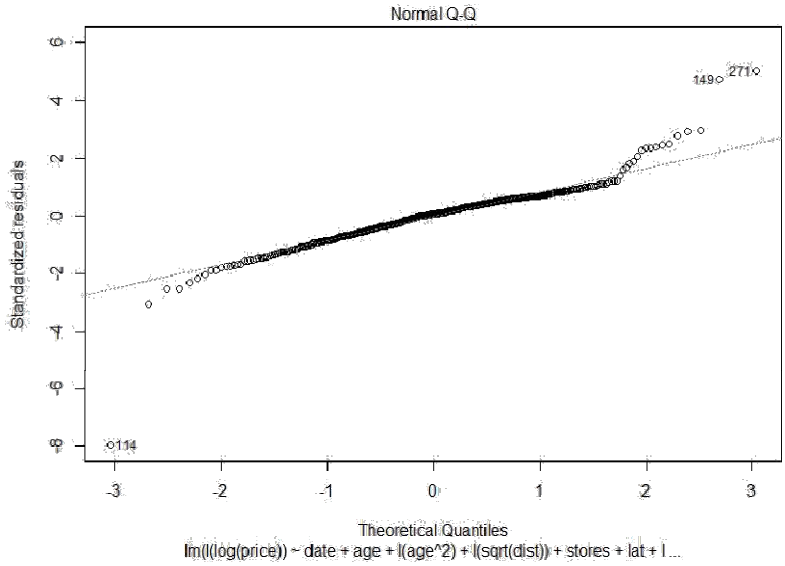
R-squared, lower Akaike's Information Criterion, and lower Mallow's $C_p$, it appears to be a more suitable model.

```
Call:
lm(formula = I(log(price)) ~ date + age + I(age^2) + I(sqrt(dist))
    + stores + lat + lon + I(lat - mean(lat)):I(lon - mean(lon)),
    data = data)

Residuals:
     Min       1Q    Median       3Q      Max
-1.59004 -0.11169   0.01305  0.11080  0.99605

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                           -5.117e+02  1.490e+02  -3.435 0.000654 ***
date                                   4.511e-04  9.767e-05   4.618 5.20e-06 ***
age                                   -2.437e-02  3.452e-03  -7.060 7.30e-12 ***
I(age^2)                               4.389e-04  8.354e-05   5.253 2.42e-07 ***
I(sqrt(dist))                         -1.229e-02  1.228e-03 -10.008  < 2e-16 ***
stores                                 1.108e-02  4.771e-03   2.321 0.020758 *
lat                                    9.349e+00  1.020e+00   9.164  < 2e-16 ***
lon                                    2.265e+00  1.175e+00   1.928 0.054544 .
I(lat - mean(lat)):I(lon - mean(lon)) 1.912e+02  5.814e+01   3.288 0.001098 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2013 on 405 degrees of freedom
Multiple R-squared:  0.7421,     Adjusted R-squared:  0.737
F-statistic: 145.7 on 8 and 405 DF,  p-value: < 2.2e-16
```

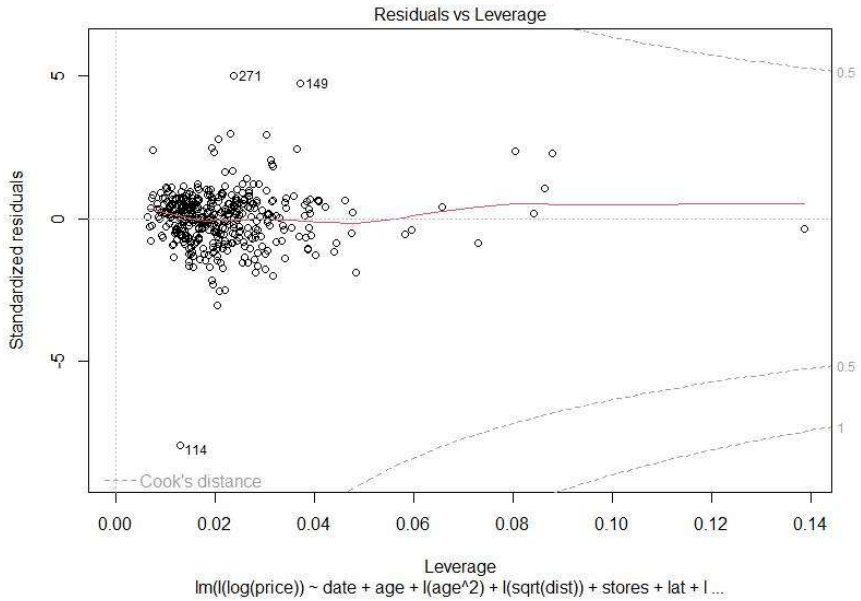
The model appears to be a reasonable fit, with 74% of the variability in log unit price accounted for by the explanatory variables.



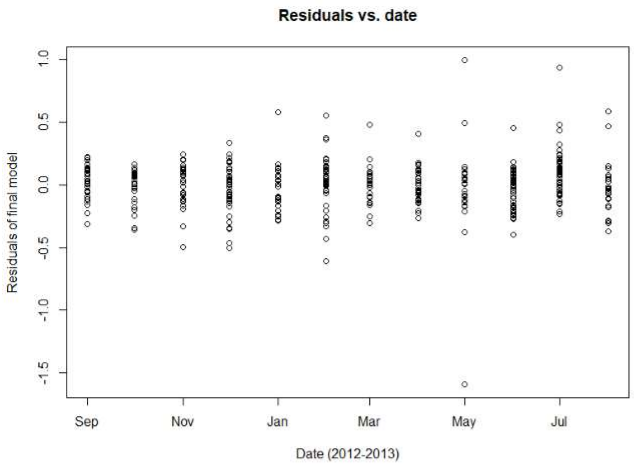

The assumptions of linearity and homoscedasticity appear to be reasonably satisfied.

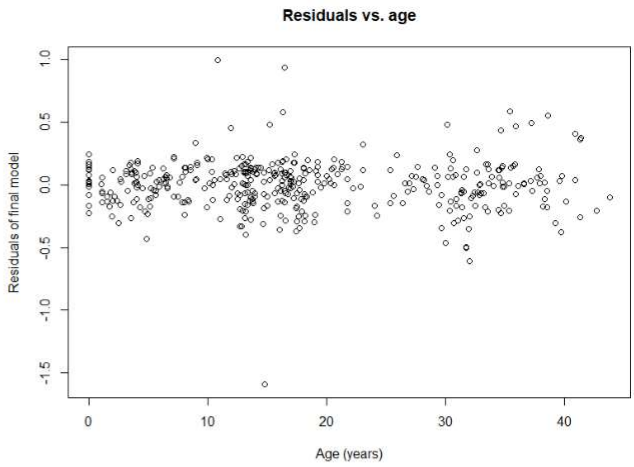


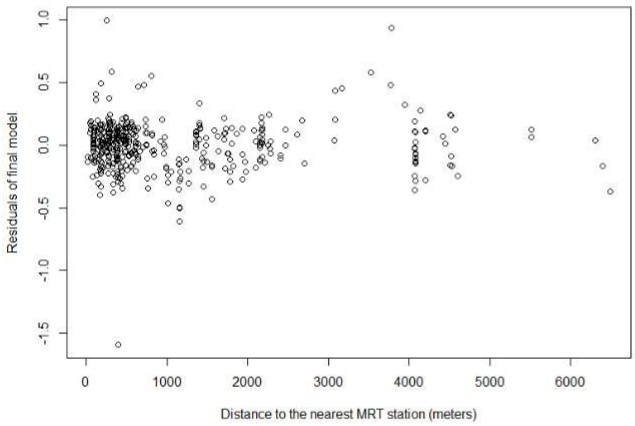

The assumption of Normally distributed errors also seems reasonable, despite the residuals appearing to follow a slightly skewed and heavy-tailed distribution.

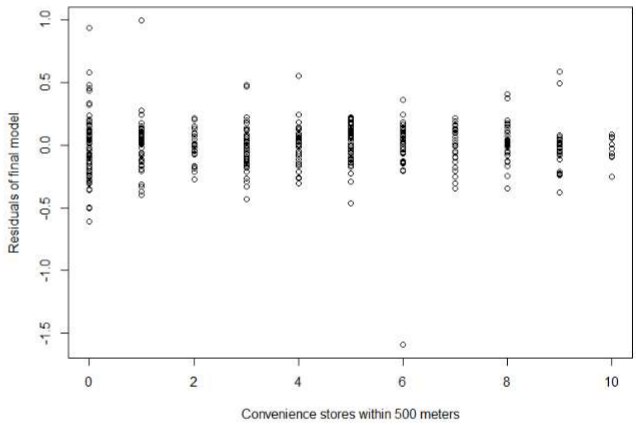


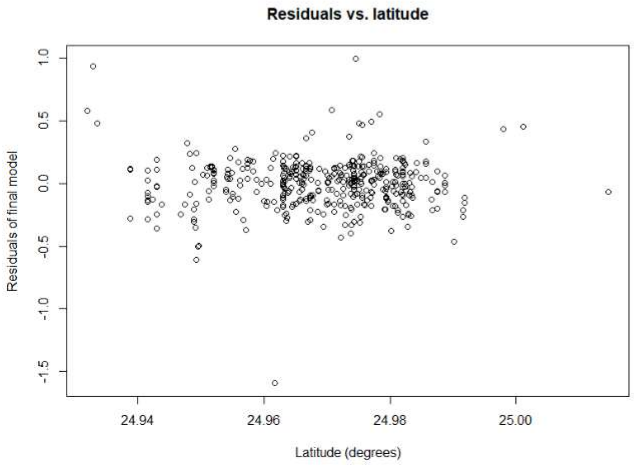

The potential outliers identified earlier do not have high leverage, and all of the points are well below a Cook's distance of 0.5, so there do not appear to be any problematic influential points. Plots of the residuals of the final model against each of the explanatory variables follow.

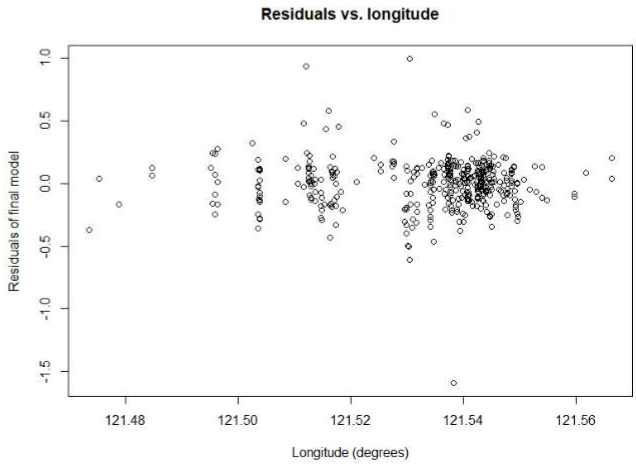


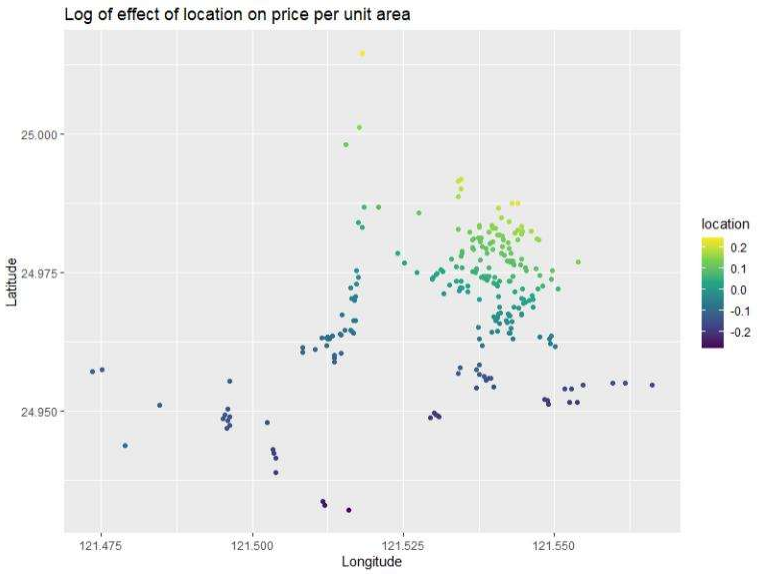




It appears, then, that house prices per unit area were on a rising trend in the Sindian district between September 2012 and August 2013, increasing by a factor of 1.18 over the course of the year when controlling for the other explanatory variables. Again

controlling for the other explanatory variables, buildings of age 27.8 years were the least preferred, with both newer and older buildings having higher prices per unit area. Values were also higher for properties closer to MRT stations, as well as for those with more convenience stores nearby. An additional convenience store within 500 meters corresponded to a 1.1% increase in price per unit area. Within the area covered by the data, a slight increase in value was associated with properties to the northeast of the central area, and a slight decrease to the northwest.




## BI. Conclusion

By removing some layers of uncertainty in land and building appraisal with a better understanding of what variables affect property prices, we can reduce the friction that exists around both housing policies and sales. Through the examination of various appraisal factors and their relationships with property prices, several key findings have emerged that highlight the critical role of accessibility, amenities, and infrastructure in determining pricing patterns.

Specifcally, our analysis revealed that properties situated in areas with closer proximity to amenities and transportation, as indicated by shorter distances to convenience stores and MRT stations, tend to command higher prices. However, while proximity to amenities and transportation positively impacts property values, there are diminishing returns beyond certain thresholds. Additionally, areas with higher population density tend to command higher property prices. It is important to keep in mind however that this correlation might also be driven by factors such as convenience, connectivity, and proximity to population centers, which contribute to increased demand and desirability in those densely populated areas.

Moreover, there appears to be a noticeable price trend (both positive and negative) for certain neighborhoods based on the temporal variation. This temporal variation aligns with the current conventional method of property valuation, given that one of the primary methods of determining property value is to start the valuation based on the asking prices and recent sales of the properties in close proximity to the property being appraised. This temporal insight can be particularly useful when one is trying to ascertain trends in certain locations or neighborhoods that are experiencing sudden fuctuations to property value.

All in all, the implications of our fndings promote a practical understanding of the underlying causes of property valuation in a dense ecosystem occupied by buyers, sellers, investors, and policymakers alike. These insights can be leveraged by all parties to help make more accurate and transparent decisions, and reduce uncertainty regarding whatever action they want to take in the real estate market. As this sector continues to grow and evolve, our fndings serve as a foundation for a guided and practical approach to real estate, which ultimately helps promote a more sound and efcient market.