Cover Page

Identifying Factors that have Impact on Real Estate Prices With Regression

Simon Fraser University
Course: STAT350
Professor Name: Derek Bingham

Group-29 Member:

- Liu Huilin
- Wang Muzi
- Wang Tianhang

Abstract

Background: The aim of the project is to identify factors that may have impact on the housing price using linear regression. **Methods:** 414 records of property sales were collected from real estate market in year 2012 to 2013. Variables included purchase year, house age, distance to nearest MRT station, number of convenience stores, latitude and longitude. We performed stepwise variable selection from baseline model with main effects to the full model using all interaction terms in linear regression. Additionally, we performed K-means clustering based on latitude and longitude. **Results**: The location is the most important variable, not only in the exact location defined by latitude and longitude, but also defined by the distance to the nearest MRT station and number of convenient stores. Specifically, the housing price would decrease as the house age or distance to the nearest MRT station increases, and as the number of decreases.

Introduction

Assessing the market value of real estate is of increasing interest to researchers who conduct economics research. The real estate market is exposed to many factors such as the location, the house age and how convenient it is to commute. It is challenging to collect data to predict the real estate price, due to the existing correlations with many variables. The aim of this project is to identify factors that may have impact on the housing price (per unit area) using statistical modeling techniques taught in class.

Data Description

Historical data were collected by Prof. I-Cheng Yeh based on real estate market from year 2012 to 2013, Sindian District, New Taipei City were used for the project (UCI machine learning repository https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set#). The dataset contains 414 records of property sales, in which each row represents a transaction record of an real estate property, and each column corresponds to a feature describing that record.

New additional data point: A simple random sample is implemented to randomly selected for each of the variables, where each of the 414 observations has an equal probability of being chosen. For the new additional data point with house price of unit area is 28.6, the year is 2012, the house age is 17, the distance to the nearest MRT station is 380, the number of *convenience stores is 4, the latitude is 24.98 and the longitude is 121.55.*

Response variable

- *House price of unit area (Y), defined in 10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared.*

Explanatory variables

- *Year of purchase (X1), Year 2012 or Year 2013 based on date of purchase.*
- *House age in years (X2), continuous variable ranging from 0 to 45 years.*

- *Distance to nearest MRT station (X3), continuous variable in meters.*
- *Number of convenience stores (X4), continuous variable.*
- *Location, defined by latitude (X5) and longitude (X6), in degrees.*

Methods

Multiple linear was implemented to identify factors (X1 to X6) that may have impact on the housing price. The full model was fitted for the explanatory variables using all variables including year of purchase, house age, distance to the nearest MRT station, number of convenience stores, latitude and longitude. The second model performed stepwise selection to determine the best predictor subset from all variables. The third model further investigated potential feature engineering to create new cluster indicators based on K-means clustering using house location (latitude and longitude). Models were then compared using Bayesian's Information Criterion (BIC). that leverages the trade-off between the goodness of fit and simplicity of the model. Linear assumptions (residuals) and multicollinearity (variance inflation factor) were checked for the final model. Statistical analyses were performed using RStudio (Version 1.3.1073). P-values less than 0.05 were considered statistically significant.

Results

Data Visualization

The histogram of house price of unit area was shown in Figure 1A, and the square root of original scale was presented in Figure 1B. In the following data analyses, we would focus on the square root of original scale, as it is more normally distributed.

The relationship between each of the variables and the response variable was visualized in Figure 2. Figure 2A plotted house price over house age, which did not show an obvious trend. This may imply that the house age is not so relevant to the house price. Figure 2B plotted house price over distance to the nearest MRT station, which indicated a clear decreasing trend as the distance increased. Houses with the highest sale price are within 1000 meters to an MRT station, which suggested that the distance to the nearest MRT station is an important predictor

for the house price. Figure 2C presented the house price over the number of convenience stores, which indicated a slowly increasing trend between these two variables. From Figure 2, there was a potential outlier with extremely high house price of unit area of 117.5 in Year 2013, whose house age of 10.8, distance to MRT of 252.6, number of convenience stores of 1. Moreover, Figure 2D indicates that the house which locate at latitude 24.97 and longitude 121.54 are the place people intend to purchase. In addition, the houses which away from this area are consider to be outliers in later analysis.

Figure 3 displayed the relationship between variables by matrix scatter plot, which indicated that both the house age and number of convenience stores had some degree of relationship with distance to the nearest MRT station. In specific, buyers tend to purchase the property with shorter distance to the nearest MRT station regardless of the age of the house. Moreover, the number of convenience stores within walking distance is higher if the distance to the nearest MRT station is short. Thus, we can conclude that distance to the nearest MRT station is an important predictor.

Model 0: Baseline model

The baseline model was fitted for the explanatory variables using all variables including year of purchase, house age, distance to the nearest MRT station, number of convenience stores, latitude and longitude. The baseline linear model with all main effects is defined as follows and the detailed output was presented in Appendix 1. The baseline model had $R^2 = 0.65$. By checking the variance inflation factor (VIF), we discovered that the X3 (distance to the nearest MRT station) had highest VIF of 4.30 and X6 (longitude) had the second highest VIF of 2.93.

$$sqrt(Y) = \alpha + \beta_1 * X1 + \beta_2 * X2 + \beta_3 * X3 + \beta_4 * X4 + \beta_5 * X5 + \beta_6 * X6 + \varepsilon.$$

Model 1: Full model

The full model was fitted for the explanatory variables using all variables including year of purchase, house age, distance to the nearest MRT station, number of convenience stores,

latitude and longitude, as well as all the pairwise interaction terms except for latitude and longitude. The full linear model is the most complex model a and the detailed output was presented in Appendix 1. The full model had higher $R^2 = 0.72$ and adjusted $R^2 = 0.71$. However, we noticed adding all the interaction terms led to main effects being insignificant, which might make the interpretation challenging.

$$sqrt(Y) = \alpha + \beta_1 * X1 + \beta_2 * X2 + \beta_3 * X3 + \beta_4 * X4 + \beta_5 * X5 + \beta_6 * X6 + \alpha_{12} * (X1 * X2) + \alpha_{13} * (X1 * X3) + \alpha_{14} * (X1 * X4) + \alpha_{15} * (X1 * X5) + \alpha_{16} * (X1 * X6) + \alpha_{23} * (X2 * X3) + \alpha_{24} * (X2 * X4) + \alpha_{25} * (X2 * X5) + \alpha_{26} * (X2 * X6) + \alpha_{34} * (X3 * X4) + \alpha_{35} * (X3 * X5) + \alpha_{36} * (X3 * X6) + \alpha_{45} * (X4 * X5) + \alpha_{46} * (X4 * X6) + \varepsilon.$$

Model 2: Stepwise variable selection

To identify the best predictor subset from all variables and interaction terms, we started from the simplest model and ended with the full model by performing the stepwise variable selection based on Bayesian's Information Criterion (BIC). The stepwise model included main effect X1 to X5 all being significant except for purchase year, and the interaction terms as defined below. Although the stepwise model had much less interaction terms, it had comparable $R^2 = 0.72$ and adjusted $R^2 = 0.72$.

$$sqrt(Y) = \alpha + \beta_1 * X1 + \beta_2 * X2 + \beta_3 * X3 + \beta_4 * X4 + \beta_5 * X5 + \alpha_{12} * (X1 * X2) + \alpha_{34} * (X3 * X4) + \alpha_{35} * (X3 * X5) + \alpha_{45} * (X4 * X5) + \varepsilon.$$

Model 3: K-means clustering

In this model, we employed feature engineering to create new cluster indicators based on K-means clustering using house location (latitude and longitude). Figure 4 identified three clusters, where the first and third cluster is closer to the standardized mean of the latitude and longitude thus could potentially be the residential area. We then used the cluster as a new variable (Z) which had three levels that represented the cluster 1, 2, 3 that contained information of both latitude and longitude. Similarly in the stepwise variable selection based on BIC, we started from the simple model with main effects X1-X4, Z, and ended with the full model with main effects X1-X4, Z, and all potential interaction terms. The stepwise model

included main effect X1-X4, Z all being significant except for purchase year, and the interaction terms as defined below. Although the stepwise model had much less interaction terms, it had comparable $R^2 = 0.73$ and adjusted $R^2 = 0.72$. The detailed output was shown in Appendix 2.

$$sqrt(Y) = \alpha + \beta_1 * X1 + \beta_2 * X2 + \beta_3 * X3 + \beta_4 * X4 + \beta_5 * Z + \alpha_{12} * (X1 * X2) + \alpha_{23} * (X2 * X3) + \alpha_{34} * (X3 * X4) + \gamma * (X4 * Z) + \varepsilon.$$

Model Selection

From previous three models using original variables X1-X6, the stepwise model selected based on BIC is the best, as it is less complicated as the full model but had similar performance with $R^2 = 0.72$ and adjusted $R^2 = 0.72$. However, the stepwise model did not select the longitude information. To further investigate on the dataset and make the best use of location information of both latitude and longitude information using K-means clustering. By including the new variable Z that represents the location information, we repeated the stepwise variable selection using BIC. The new stepwise model had slightly better performance with $R^2 = 0.73$ and adjusted $R^2 = 0.72$. Therefore, we chose the new stepwise model with the clustering membership as our final model which could explain 74% of the variability of the data.

Residuals Check

The Diagnostic plot of final model was shown in Figure 5.

- Residuals vs Fitted: The relationship between fitted values and residuals are flat.
- Normal Q-Q: The residuals do not fall close to the line (towards end of the right tail) and there are some deviations from normality, so the assumptions of residuals are not normally distributed and this assumption is violated.
- Scale-location: The red line in this plot is flat and the variances in the square root of the standardized residuals are consistently across fitted values. Therefore, this is a sign of homoscedasticity and the assumption is not violated.
- Residuals vs. Leverage: There is no values that fall in the upper and lower right hand side of the plot beyond the red bands, therefore there is no evidence of influential cases.

Final Model Interpretation

The results of the final stepwise model with the clustering membership are shown in Table 1. The factors that impacted house price of unit area are house age, distance to the nearest MRT station, number of convenience stores, and location clusters.

- As house age increased by 10 year while keeping all the other variables consistent, the square root of house price of unit area would decrease by 0.27 (95% CI: 0.21 to 0.34).
- As distance to the nearest MRT station increased by 1000 meters while keeping all the other variables consistent, the square root of house price of unit area would decrease by 2.9 (95% CI: 1.1 to 4.6).
- As number of convenient stores increased by 1 while keeping all the other variables consistent, the square root of house price of unit area would increase by 0.27 (95% CI: 0.19 to 0.36).
- The location cluster membership is challenging to interpret as we need more information to characterize the cluster defined by latitude and longitude. Compared to the houses in cluster 3, cluster 1 had higher house price of unit area while cluster 2 had lower house price of unit area.
- Additionally, the house price of unit area also related to the interaction terms of:
    - purchase year and house age;
    - house age and distance to nearest MRT station;
    - distance to nearest MRT station and number of convenience stores;
    - number of convenience stores and clustering membership.

Conclusion

Our findings from the final model are consistent with previous literature. We have identified factors such as house age, distance to the nearest MRT station, number of

convenience stores, and location clusters are key variables in interpreting the housing price per unit area. In particular, the location is the most important variable, not only in the exact location defined by latitude and longitude, but also defined by the distance to the nearest MRT station and number of convenient stores. In summary, the housing price per unit area would decrease as the house age or distance to the nearest MRT station increases, and as the number of convenient store decreases.

Future studies could further investigate on the interaction terms between location variables. In specific, we could link the zip code to the latitude and longitude, which could help with the interpretation. Other county-level variables such as the race and education should be considered in predicting the housing price per unit area. Furthermore, we could attempt to collapse the continuous variables into appropriate categories to help with the interpretation, especially for the interaction terms.

Reference

Yeh, I. C., & Hsu, T. K. (2018). Building real estate valuation models with comparative approach through case-based reasoning. Applied Soft Computing, 65, 260-271.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Table 1: Linear regression results of final model

	estimate	2.50%	97.50%	p-value
Year1	-0.02664	-0.13975	0.08647	0.64
X2.house.age	-0.02741	-0.03400	-0.02081	<0.0001
X3.distance.to.MRT	-0.00029	-0.00046	-0.00011	0.001
X4.number.of.stores	0.27360	0.18540	0.36180	<0.0001
cluster1	0.94701	0.71102	1.18301	<0.0001
cluster2	-0.52040	-0.84639	-0.19440	0.002
X4:cluster1	-0.16554	-0.23855	-0.09254	<0.0001
X4:cluster2	0.18235	0.05414	0.31057	0.005
X3:X4	-0.00016	-0.00022	-0.00010	<0.0001
X2:X3	0.00001	0.00000	0.00001	0.03
Year1:X2	-0.00548	-0.01096	0.00000	0.05

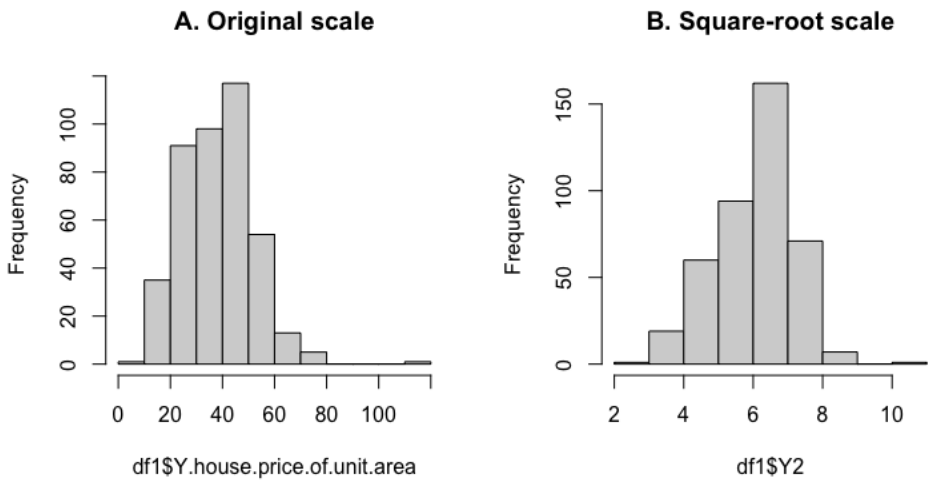

Figure 1. Histogram of House price of unit area, defined in 10000 New Taiwan Dollar/Ping.

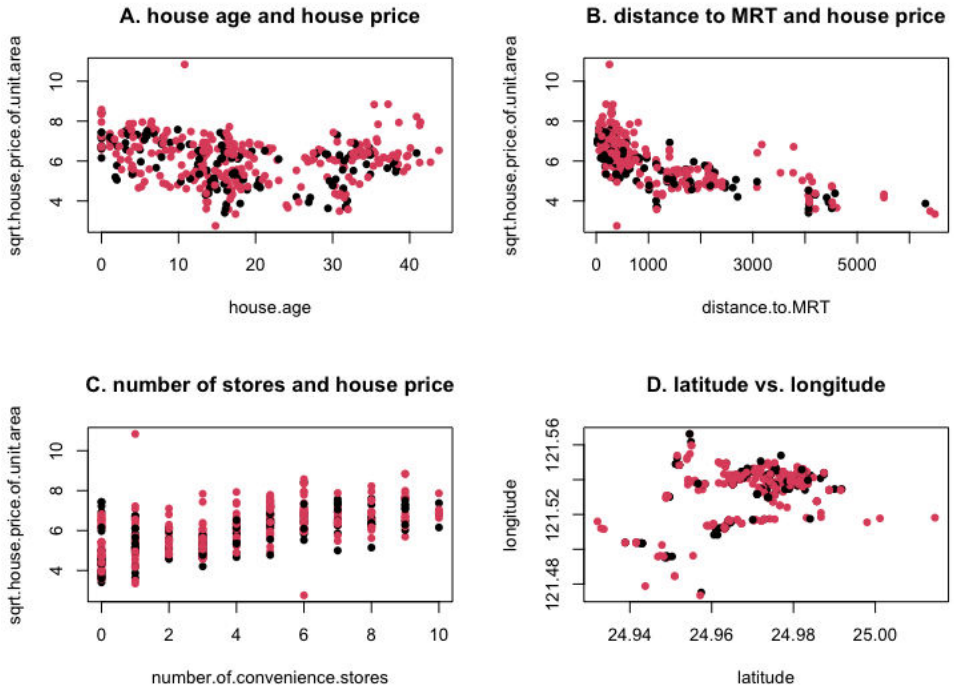

Figure 2. relationship between each of the variables and the response variable (black represents Year 2012 and red represents Year 2013).

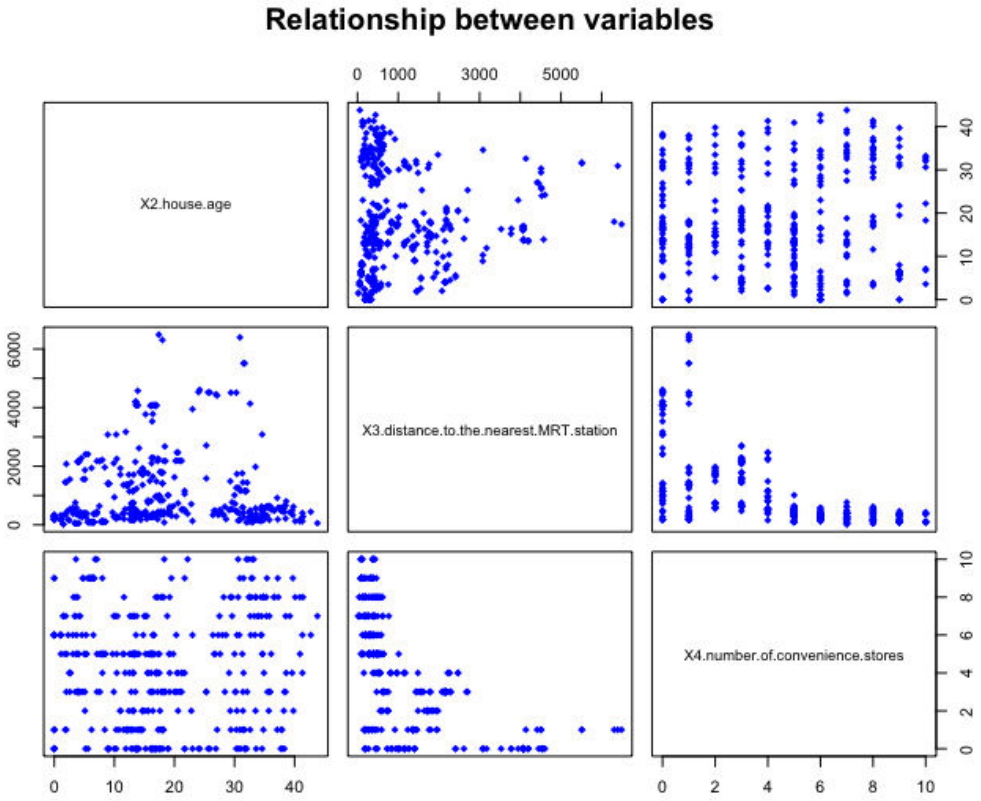

Figure 3. relationship between variables (X3-house age, X4- distance to MRT, X4- number of stores).

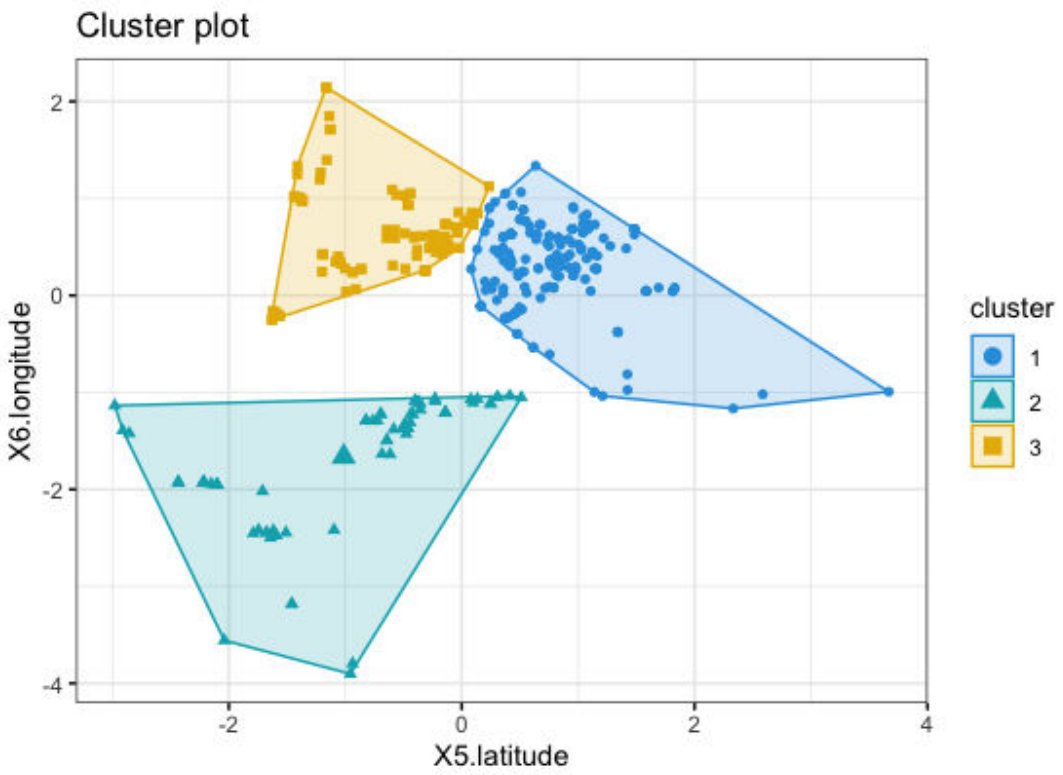

Figure 4. K-means clustering of latitude and longitude (k = 3).

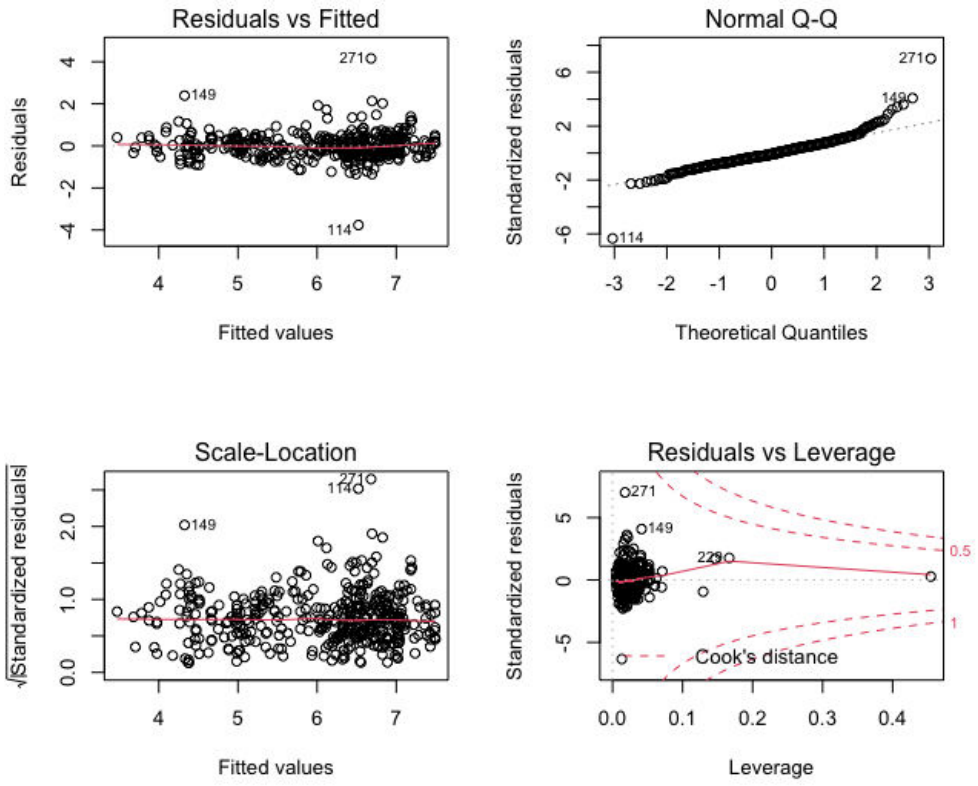

Figure 5. Diagnostic plot of final model.

Appendix 1

Baseline model output

```
Call:
lm(formula = Y2 ~ ., data = df1[, -c(1, 8)])

Residuals:
    Min      1Q  Median      3Q     Max
-3.7233 -0.4009 -0.0593  0.3415  4.3993

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                            -4.665e+02  4.664e+02  -1.000 0.317807
Year1                                  -1.206e-01  3.593e-02  -3.357 0.000862 ***
X2.house.age                           -2.179e-02  2.931e-03  -7.433 6.29e-13 ***
X3.distance.to.the.nearest.MRT.station -3.890e-04  5.435e-05  -7.156 3.89e-12 ***
X4.number.of.convenience.stores         9.184e-02  1.427e-02   6.434 3.49e-10 ***
X5.latitude                             2.128e+01  3.370e+00   6.313 7.16e-10 ***
X6.longitude                           -4.798e-01  3.691e+00  -0.130 0.896656
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.673 on 408 degrees of freedom
Multiple R-squared:  0.6452,    Adjusted R-squared:   0.64
F-statistic: 123.7 on 6 and 408 DF,  p-value: < 2.2e-16
```

Full model output

```
Call:
lm(formula = Y2 ~ (.)^2, data = df1[, -c(1, 8)])

Residuals:
    Min      1Q  Median      3Q     Max
-3.7674 -0.3246 -0.0498  0.2796  4.1371

Coefficients: (1 not defined because of singularities)
                                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                           -2.552e+03  1.188e+03  -2.149 0.032259 *
Year1                                                                  2.571e+02  4.482e+02   0.574 0.566525
X2.house.age                                                           4.077e+01  5.419e+01   0.753 0.452198
X3.distance.to.the.nearest.MRT.station                                 3.894e-01  2.402e-01   1.621 0.105769
X4.number.of.convenience.stores                                        2.694e+02  2.993e+02   0.900 0.368663
X5.latitude                                                            7.785e+01  1.245e+01   6.254 1.04e-09 ***
X6.longitude                                                           5.058e+00  9.109e+00   0.555 0.579042
Year1:X2.house.age                                                    -6.937e-03  2.901e-03  -2.391 0.017248 *
Year1:X3.distance.to.the.nearest.MRT.station                           9.337e-07  5.307e-05   0.018 0.985973
Year1:X4.number.of.convenience.stores                                  4.377e-03  1.391e-02   0.315 0.753206
Year1:X5.latitude                                                     -1.091e+00  3.721e+00  -0.293 0.769600
Year1:X6.longitude                                                    -1.892e+00  3.515e+00  -0.538 0.590792
X2.house.age:X3.distance.to.the.nearest.MRT.station                    3.911e-07  6.008e-06   0.065 0.948132
X2.house.age:X4.number.of.convenience.stores                           5.525e-04  1.065e-03   0.519 0.604252
X2.house.age:X5.latitude                                               1.573e-03  3.350e-01   0.005 0.996257
X2.house.age:X6.longitude                                             -3.360e-01  4.400e-01  -0.764 0.445451
X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores -1.141e-04  2.961e-05  -3.854 0.000136 ***
X3.distance.to.the.nearest.MRT.station:X5.latitude                     -1.744e-02  3.370e-03  -5.173 3.68e-07 ***
X3.distance.to.the.nearest.MRT.station:X6.longitude                     3.752e-04  2.073e-03   0.181 0.856448
X4.number.of.convenience.stores:X5.latitude                           -8.764e+00  1.649e+00  -5.315 1.79e-07 ***
X4.number.of.convenience.stores:X6.longitude                          -4.149e-01  2.380e+00  -0.174 0.861705
X5.latitude:X6.longitude                                                      NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6042 on 394 degrees of freedom
Multiple R-squared:  0.7238,    Adjusted R-squared:  0.7098
F-statistic: 51.62 on 20 and 394 DF,  p-value: < 2.2e-16
```

Stepwise selection output

```
Call:
lm(formula = Y2 ~ X3.distance.to.the.nearest.MRT.station + X4.number.of.convenience.stores +
    X2.house.age + X5.latitude + Year + X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores +
    X2.house.age:Year + X3.distance.to.the.nearest.MRT.station:X5.latitude +
    X4.number.of.convenience.stores:X5.latitude, data = df1[,
    -c(1, 8)])

Residuals:
    Min      1Q  Median      3Q     Max
-3.7651 -0.3264 -0.0508  0.2851  4.1582

Coefficients:
                                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                     -1.926e+03  1.965e+02  -9.801  < 2e-16 ***
X3.distance.to.the.nearest.MRT.station                           4.237e-01  6.587e-02   6.432 3.56e-10 ***
X4.number.of.convenience.stores                                  2.157e+02  3.423e+01   6.301 7.72e-10 ***
X2.house.age                                                    -2.366e-02  2.839e-03  -8.334 1.23e-15 ***
X5.latitude                                                      7.740e+01  7.870e+00   9.834  < 2e-16 ***
Year1                                                            6.759e-03  5.802e-02   0.117    0.907
X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores -1.114e-04  1.823e-05  -6.111 2.32e-09 ***
X2.house.age:Year1                                              -7.021e-03  2.813e-03  -2.496    0.013 *
X3.distance.to.the.nearest.MRT.station:X5.latitude              -1.698e-02  2.639e-03  -6.435 3.49e-10 ***
X4.number.of.convenience.stores:X5.latitude                     -8.633e+00  1.371e+00  -6.298 7.87e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5981 on 405 degrees of freedom
Multiple R-squared:  0.7218,    Adjusted R-squared:  0.7157
F-statistic: 116.8 on 9 and 405 DF,  p-value: < 2.2e-16
```

Appendix 2

Appendix 2 presented output based on K-means clustering membership as the new variable to replace latitude and longitude variables.

Baseline model output

```
lm(formula = Y2 ~ ., data = df3[, c(2:5, 9:10)])

Residuals:
    Min      1Q  Median      3Q     Max
-3.5898 -0.3796 -0.0341  0.2872  4.2499

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                             6.219e+00  1.112e-01  55.917  < 2e-16 ***
Year1                                  -1.176e-01  3.406e-02  -3.453 0.000613 ***
X2.house.age                           -2.325e-02  2.795e-03  -8.320 1.33e-15 ***
X3.distance.to.the.nearest.MRT.station -3.304e-04  4.693e-05  -7.039 8.25e-12 ***
X4.number.of.convenience.stores         1.072e-01  1.358e-02   7.894 2.73e-14 ***
cluster1                                4.803e-01  5.560e-02   8.639  < 2e-16 ***
cluster2                               -3.210e-01  8.360e-02  -3.839 0.000143 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6385 on 408 degrees of freedom
Multiple R-squared:  0.6806,    Adjusted R-squared:  0.6759
F-statistic: 144.9 on 6 and 408 DF,  p-value: < 2.2e-16
```

Stepwise selection output

```
Call:
lm(formula = Y2 ~ Year + X2.house.age + X3.distance.to.the.nearest.MRT.station +
    X4.number.of.convenience.stores + cluster + X4.number.of.convenience.stores:cluster +
    X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores +
    X2.house.age:X3.distance.to.the.nearest.MRT.station + Year:X2.house.age,
    data = df3[, c(2:5, 9:10)])

Residuals:
    Min      1Q  Median      3Q     Max
-3.6374 -0.3397 -0.0325  0.2727  4.0549

Coefficients:
                                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                     6.036e+00  1.621e-01  37.249  < 2e-16 ***
Year1                                                          -2.664e-02  5.754e-02  -0.463  0.64366
X2.house.age                                                   -2.741e-02  3.355e-03  -8.169 4.04e-15 ***
X3.distance.to.the.nearest.MRT.station                         -2.863e-04  8.856e-05  -3.234  0.00132 **
X4.number.of.convenience.stores                                 2.736e-01  4.487e-02   6.098 2.52e-09 ***
cluster1                                                        9.470e-01  1.200e-01   7.889 2.90e-14 ***
cluster2                                                       -5.204e-01  1.658e-01  -3.138  0.00183 **
X4.number.of.convenience.stores:cluster1                       -1.655e-01  3.714e-02  -4.458 1.08e-05 ***
X4.number.of.convenience.stores:cluster2                        1.824e-01  6.522e-02   2.796  0.00542 **
X3.distance.to.the.nearest.MRT.station:X4.number.of.convenience.stores -1.618e-04  3.209e-05  -5.041 7.02e-07 ***
X2.house.age:X3.distance.to.the.nearest.MRT.station             6.173e-06  2.838e-06   2.175  0.03019 *
Year1:X2.house.age                                             -5.483e-03  2.787e-03  -1.968  0.04980 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5905 on 403 degrees of freedom
Multiple R-squared:  0.7301,    Adjusted R-squared:  0.7228
F-statistic: 99.13 on 11 and 403 DF,  p-value: < 2.2e-16
```