Housing Price in Sindian District

Emmy Su

2024-03-16

Introduction

The purpose of this project is to generate the best model to predict price per unit area of real estate based on several factors (predictors). In this study, we investigate the relationship between the cost per unit of housing and predictor factors such as transaction date, house age, distance to public transit or store. The analysis contains 414 instances and is based on data collected from Sindian District, New Taipei City, Taiwan.

Date	House_age	MRT_distance	Store_distance	Latitude	Longitude	Price
2012.917	32.0	84.87882	10	24.98298	121.5402	37.9
2012.917	19.5	306.59470	9	24.98034	121.5395	42.2
2013.583	13.3	561.98450	5	24.98746	121.5439	47.3
2013.500	13.3	561.98450	5	24.98746	121.5439	54.8
2012.833	5.0	390.56840	5	24.97937	121.5425	43.1
2012.667	7.1	2175.03000	3	24.96305	121.5125	32.1

Multiple Linear Regression

We want to produce a regression model to predict the cost per unit for housing in Sindian District. The following includes 6 predictors notes as X.

- Y(price): price per unit
- X1(date): transaction date
- X2(house_age): house age
- X3(MRT_distance): Distance in meters to the nearest mass rapid transit(MRT)
- X4(store_distance): Distance to the convenience store
- X5(Latitude): Latitude
- X6(Longitude): Longitude

Regression model that directly predicts the price per unit using all of the 6 potential predictor variables listed above.

```
##
## Call:
## lm(formula = Price ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.667  -5.412  -0.967   4.217  75.190
```

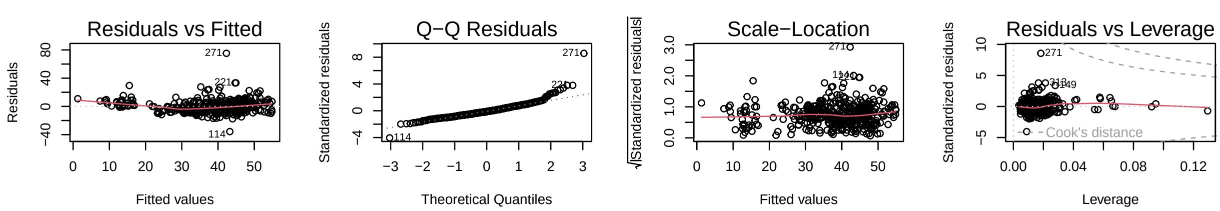
```
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.444e+04  6.775e+03  -2.132  0.03364 *
## Date            5.149e+00  1.557e+00   3.307  0.00103 **
## House_age      -2.697e-01  3.853e-02  -7.000 1.06e-11 ***
## MRT_distance   -4.488e-03  7.180e-04  -6.250 1.04e-09 ***
## Store_distance  1.133e+00  1.882e-01   6.023 3.83e-09 ***
## Latitude        2.255e+02  4.457e+01   5.059 6.38e-07 ***
## Longitude      -1.243e+01  4.858e+01  -0.256  0.79820
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.858 on 407 degrees of freedom
## Multiple R-squared:  0.5824, Adjusted R-squared:  0.5762
## F-statistic:  94.6 on 6 and 407 DF,  p-value: < 2.2e-16
```

Multiple R-squared is the proportion of variance in the dependent variable explained by the independent variables. In this case, about 58.24% of the variance in Price are explained by the predictor variables. The F-statistics test the overall significance of the regression model by comparing the fit of model with no predictors. With low p-value, it indicates that the overall model is statistically significant.

This model has a low p-value, indicating the significance of the individual predictor variables. The variables that are statistically significant in predicting the price of real estate are: Data, House_age, MRT_distance, Store_distance, and Latitude. The predictor variable that is most statistically significant is House_age because it has the lowest p-value. Longitude is not a significant predictor because the p-value is higher than 0.05. Hence, overall the predictor variables collectively have a significant impact on predicting real estate prices.

```
## Analysis of Variance Table
##
## Response: Price
##                 Df Sum Sq Mean Sq  F value    Pr(>F)
## Date             1    586     586   7.4666  0.006559 **
## House_age        1   3441    3441  43.8575 1.119e-10 ***
## MRT_distance     1  34857   34857 444.2919 < 2.2e-16 ***
## Store_distance   1   3576    3576  45.5812 5.064e-11 ***
## Latitude         1   2065    2065  26.3192 4.488e-07 ***
## Longitude        1      5       5   0.0655  0.798203
## Residuals      407  31931      78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the Analysis of Variance Table, the fitted values seem to appear statistically significant except for the Longitude. Among the predictors, MRT_distance has the highest F-value which can indicate the largest proportion of explained variance in Price.



The diagnosis plots show that the model assumption are NOT violated. The residual vs. fitted plot is used to assess the goodness of fit of a regression model. If the average residuals are zero, indicating a horizontal line, then it is a good fit. The normal quantile-quantile visualization calculates the normal quantiles of all values in a column. It is use to examine whether the residuals are normal distributed. It is a good fit if the residuals follow a straight dashed line. The conclusion drawn from the plot is that it has an approximately normal distribution. The scale-location plot uses the square root of residuals which makes it different from residual vs fit plot. The purpose is to check for homogeneity of variance of the residuals. For a good model, the values should be randomly distributed, following a horizontal line. In this case, our plot does follow a horizontal line. Residuals vs Leverages is use to identify influential observations in a regression analysis. The graph seems pretty normal but may have a few outliers that can be removed later. Ergo, the four diagnostic plots show that the model is valid.

Choosing Predictors

The predictor variable Longitude is not statistically significant because the p-value is greater than 0.05. Therefore, it is time to consider if our model significantly improves its fit compared to when it is reduced without a specific predictor.

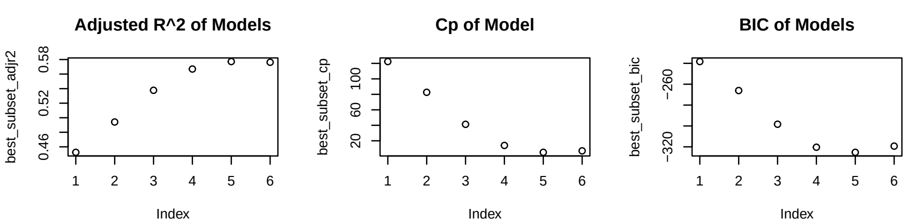
```
## Analysis of Variance Table
##
## Model 1: Price ~ Date + House_age + MRT_distance + Store_distance + Latitude
## Model 2: Price ~ Date + House_age + MRT_distance + Store_distance + Latitude +
##     Longitude
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    408 31937
## 2    407 31931  1    5.1353 0.0655 0.7982
```

By using the partial F test, the p-value is greater than 0.05, so we fail to reject null hypothesis. There is no sufficient evidence against the reduced model in factor of the full model. Therefore, the high p-value suggest that the reduced model is a better fit to the data.

All Possible Subsets

By using the method of all possible subsets, it suggests that using 5 predictor variables is best for my data.

```
##   Adj.R2 CP BIC
## 1      5  5   5
```



Forward Stepwise Regression

```
## Start:  AIC=1813.03
## Price ~ Date + House_age + MRT_distance + Store_distance + Latitude +
##     Longitude
```

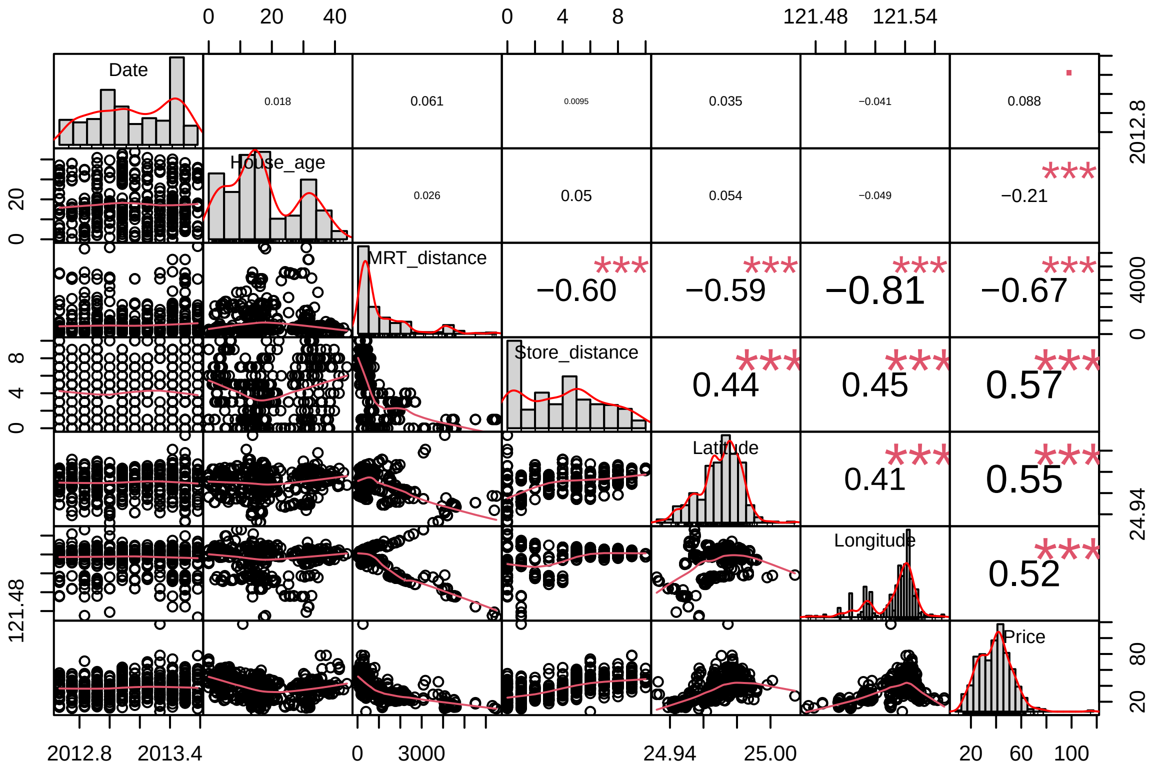
Backward Stepwise Regression

```
## Start:  AIC=1813.03
## Price ~ Date + House_age + MRT_distance + Store_distance + Latitude +
##     Longitude
##
##                  Df Sum of Sq   RSS    AIC
## - Longitude       1       5.1 31937 1811.1
## <none>                        31931 1813.0
## - Date            1     858.2 32790 1822.0
## - Latitude        1    2008.2 33940 1836.3
## - Store_distance  1    2846.3 34778 1846.4
## - MRT_distance    1    3064.6 34996 1849.0
## - House_age       1    3843.9 35775 1858.1
##
## Step:  AIC=1811.1
## Price ~ Date + House_age + MRT_distance + Store_distance + Latitude
##
##                  Df Sum of Sq   RSS    AIC
## <none>                        31937 1811.1
## - Date            1     855.0 32792 1820.0
## - Latitude        1    2064.9 34001 1835.0
## - Store_distance  1    2870.9 34807 1844.7
## - House_age       1    3838.9 35775 1856.1
## - MRT_distance    1    6181.9 38118 1882.3
```

After looking at both backward and forward stepwise regression, backward is preferred. The backward regression removes variables from model to improve the model's fit which is indicated by AIC. The AIC is lower for the backward stepwise regression after removing Longitude. The AIC went from 1813.03 to 1811.1 which suggest removing the variable results in a better-fitting model.

Checking for Pairwise Correlation

We can also check for pairwise correlation and remove a predictor variable with the highest correlation. The result below also supports that we should remove Longitude from our data.



Building the Model

The final model will have 5 predictors: Date, House_age, MRT_distance, Store_distance, Latitude.

Date	House_age	MRT_distance	Store_distance	Latitude	Price
2012.917	32.0	84.87882	10	24.98298	37.9
2012.917	19.5	306.59470	9	24.98034	42.2
2013.583	13.3	561.98450	5	24.98746	47.3
2013.500	13.3	561.98450	5	24.98746	54.8
2012.833	5.0	390.56840	5	24.97937	43.1
2012.667	7.1	2175.03000	3	24.96305	32.1

```
##
## Call:
## lm(formula = Price ~ Date + House_age + MRT_distance + Store_distance +
##     Latitude, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.625  -5.373  -1.020   4.243  75.343
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    -1.596e+04  3.233e+03  -4.938 1.15e-06 ***
## Date            5.138e+00  1.554e+00   3.305  0.00103 **
## House_age      -2.694e-01  3.847e-02  -7.003 1.04e-11 ***
## MRT_distance   -4.353e-03  4.899e-04  -8.887  < 2e-16 ***
## Store_distance  1.136e+00  1.876e-01   6.056 3.17e-09 ***
## Latitude        2.269e+02  4.417e+01   5.136 4.35e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.847 on 408 degrees of freedom
## Multiple R-squared:  0.5823, Adjusted R-squared:  0.5772
## F-statistic: 113.8 on 5 and 408 DF,  p-value: < 2.2e-16
```

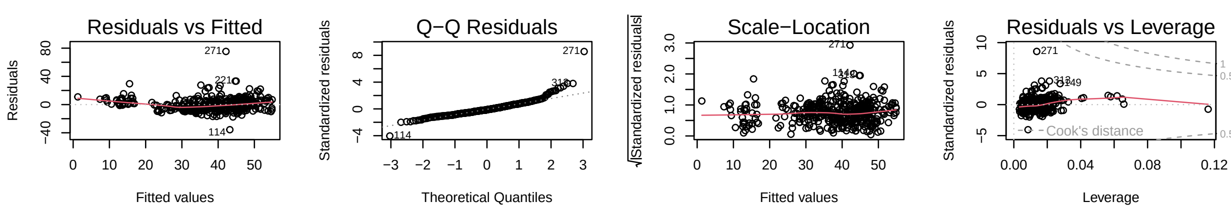
Looking at the summary of the best model, all of the predictor variables pass the statistics significance test where the p-value is lower than 0.05. Hence, making our overall predictor variables more reliable for predicting the response variable.

Model Diagnostic

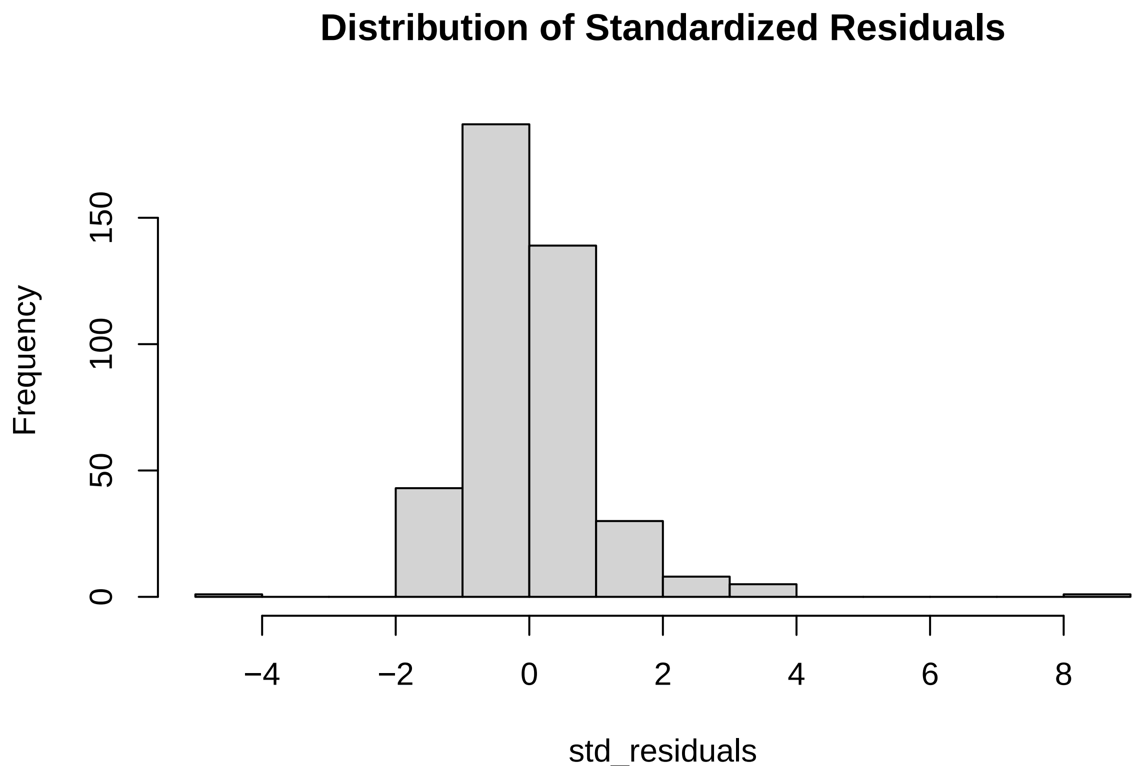
After choosing our best model, it is important to assess the quality and appropriateness of the statistical model. It can help identify the model's assumptions, evaluate the model's performance, and help detect any unusual points.

The plots of the reduced predictor variable below suggest that the data exhibits a relatively normal distribution, although there appear to be a few unusual points.

The residual vs. fitted plot seems to have a red line approximately horizontal at zero which indicates a good fit. The normal quantile-quantile visualization also looks like a good fit because it follows a straight dashed line. The scale-location plot seems like a good model since the values are randomly distributed, following a horizontal line. Residuals vs Leverages seems pretty normal but may have potential outliers.



We can also look at the Distribution of Standardized Residuals to further check the assumption of normality in the residuals of a statistical model. Looking at the distribution of standardized residual below, it is a normal distribution with potential outliers that can be further removed from the data.



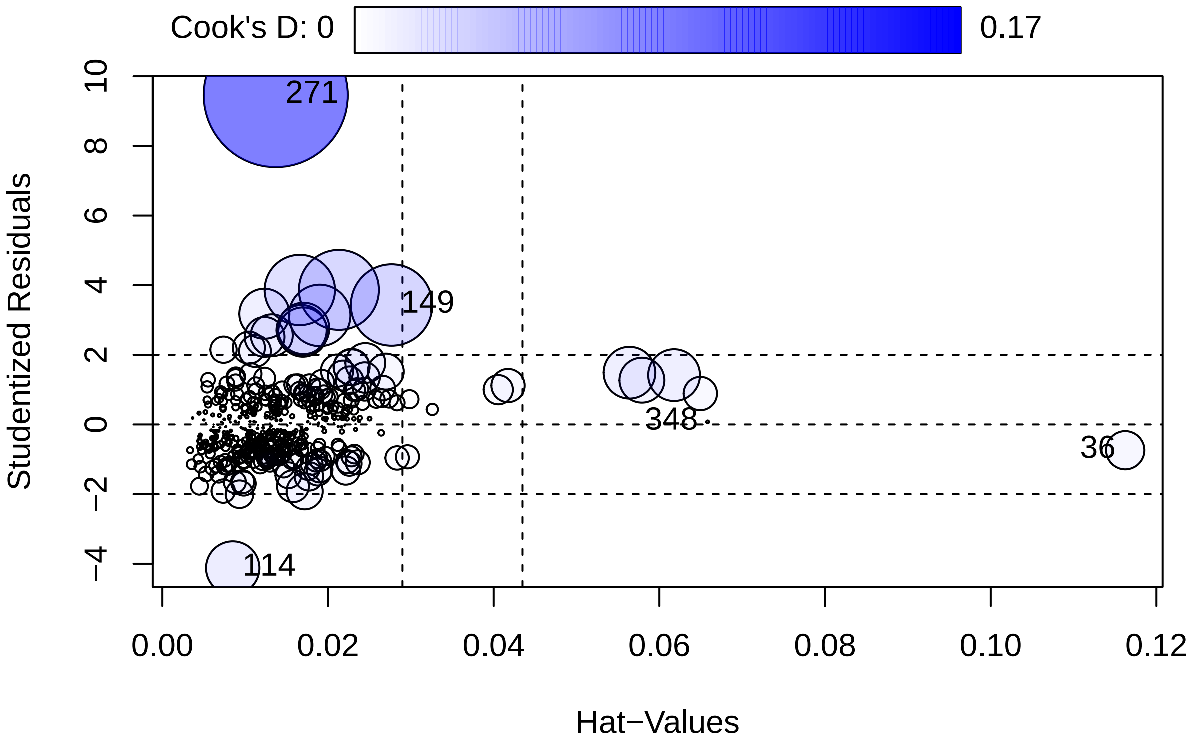
Checking if there is Multi-collinearity

I assessed multi-collinearity within my best model and found low indication of multi-collinearity. The VIF values close to 1 indicates low multi-collinearity which means the predictor variables does not have strong linear relationships with other predictors. A VIF exceeding 5 is an indicator of muli-collinearity but the variable predictors below have VIF values are from 1-2, which indicates multi-collinearity is not a significant concern in this model.

```
##           Date      House_age   MRT_distance Store_distance       Latitude
##       1.013815       1.013243       2.016820       1.611282       1.585625
```

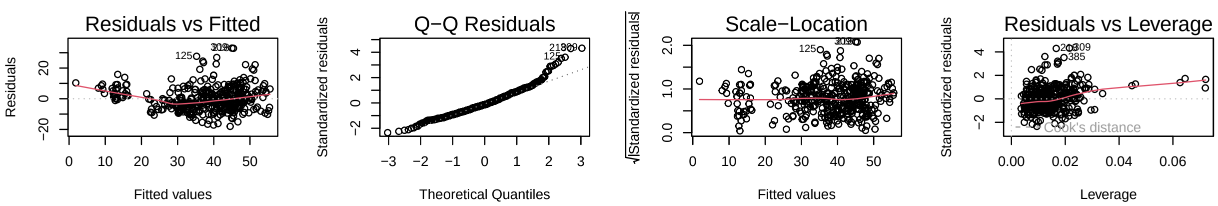
Identify Unusual Values

As we have identified potential outliers from the plots above, we can delve into it and use the influencePlot function to identify them. Based on the results below, there are five unusual values identified along with their corresponding statistics. We can then use this information to remove it from our reduced data and see if it makes a difference.




```
##         StudRes         Hat       CookD
## 36  -0.73961223 0.116237700 1.200469e-02
## 114 -4.12239104 0.008498624 2.336161e-02
## 149  3.43096568 0.027676344 5.440795e-02
## 271  9.45893827 0.013693328 1.701361e-01
## 348  0.07380196 0.065826840 6.412392e-05
```

After removing the unusual points, the Residual vs Fitted Plot is more spread out, which depicts how the spread is more significant. This indicates that the model's performance has been positively influenced by the removal of unusual points.



Inference

After removing the Longitude predictor and 5 values from the data, we have concluded our final model.

The summary of our new model below shows improvement of our data. The difference reflects how our model is more reliable compared to the full model. The residual standard error decreased from 8.858 to 7.716, the Adjusted R-squared increased from 0.5762 to 0.6445. In addition, the F-statistic increased from 94.6 to 148.9. Hence, the difference is an improvement to our previous summary.

```
##
## Call:
## lm(formula = Price ~ ., data = without_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.020  -5.183  -1.053   4.106  32.985
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.501e+04  2.863e+03  -5.242 2.57e-07 ***
## Date            4.548e+00  1.363e+00   3.336 0.000928 ***
## House_age      -2.642e-01  3.361e-02  -7.861 3.53e-14 ***
## MRT_distance   -4.111e-03  4.540e-04  -9.056  < 2e-16 ***
## Store_distance  1.294e+00  1.651e-01   7.837 4.14e-14 ***
## Latitude        2.360e+02  4.109e+01   5.744 1.82e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.716 on 403 degrees of freedom
## Multiple R-squared:  0.6488, Adjusted R-squared:  0.6445
## F-statistic: 148.9 on 5 and 403 DF,  p-value: < 2.2e-16
```

The F-test below indicates that the model is reliable and compared to the previous F-test with the full model, the one below has higher F-values. Since the F-values measures how well the model fits the data, the higher F-values indicates the teh reduced model without Longitude and the 5 points are a better fit for our data. In addition, they all have small p-values, indicating that they are statistically significant predictors for the response variable (Price).

```
## Analysis of Variance Table
##
## Response: Price
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## Date             1    628     628  10.541  0.001265 **
## House_age        1   3245    3245  54.505 8.965e-13 ***
## MRT_distance     1  34066   34066 572.129 < 2.2e-16 ***
## Store_distance   1   4433    4433  74.454 < 2.2e-16 ***
## Latitude         1   1965    1965  32.997 1.821e-08 ***
## Residuals      403  23996      60
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the tests conducted above, we can conclude that the 5 predictor variables is statistically significant in predicting the price per unit for housing in Sindian District, New Taipei City, Taiwan. The date when the house was purchased, age of the house, MRT distance, store distance, and latitude are all significant variables in predicting the price per unit of housing within the district. Originally, the dataset included a sixth variable to predict house prices. However, this variable did not show sufficient statistical evidence to justify its impact on housing prices. Thus, it was non-significant and removed from the final model.

The predictor variable that is most statistically significant is MRT_distance because it has the lowest p-value. This suggest it has a substantial impact on the price. The demand to buy a house near public transportation may be higher because it gives them more accessibility. This can significantly affect the price by increasing the cost because of high demand. The second most influential predictor that determines the price is house age. People may not want to buy a house that is old because it may require more maintenance or reconstructing, thus influencing the price. The third most influential is the store distance, how far the house is from a convenience store. Having stores nearby can also increase the price because the demand would be higher. The convenience store offers easy access to everyday necessities which increases its attractiveness in the housing market. Next influential factor would be the age of the house and latitude.

The result is evidence of how we are able to predict the price per unit of housing within the district based on the 5 predictor variables.

Citation

Yeh,I-Cheng. (2018). Real Estate Valuation. UCI Machine Learning Repository. https://doi.org/10.24432/C5J30W.