141A project

Table of Contents

[A. Introduction 1](#_Toc88603555)

[B. Data background and Questions of Interest 2](#_Toc88603556)

[C. Exploratory Analysis 3](#_Toc88603557)

[D. Data transformation and possible models 8](#_Toc88603558)

[E. Model fitting 8](#_Toc88603559)

[Linear regression model 8](#_Toc88603560)

[polynomial regression model 10](#_Toc88603561)

[F. Model diagnostic 11](#_Toc88603562)

[Model 1 12](#_Toc88603563)

[Model 2 14](#_Toc88603564)

[G. Statistical analysis of the questions 15](#_Toc88603565)

[Does the predictor MRT station distance, convenience store numbers and location have significant effects on house price? 15](#_Toc88603566)

[Is there a relationship between MRT station distance and convenience store numbers and location? 16](#_Toc88603567)

[Are there any houses seem to be overestimated or underestimated? (Function part is located here) 17](#_Toc88603568)

[F. Conclusion 18](#_Toc88603569)

[G. References 18](#_Toc88603570)

[H. R code 19](#_Toc88603571)

# A. Introduction

In this project, we will analyze New Taipei City Sindian District house price per unit area data, and propose suitable model for house price prediction. The dataset collects the historical data from real estate market from 2012 to 2013, in Sindian District, New Taipei City. Analyzing this dataset can help people know what are the significant predictors for the house price, and give people insights about determining whether the house price is above market or below market. The model is also helpful for government and broker agencies, so we want to analyze the New Taipei house dataset. To achieve our goal, we will first make exploratory analysis, finding meaningful predictors, checking data structure. We will perform statistical tests to select and diagnose the regression model, and use the final model to predict the average price and determine whether the price is significant off the market value.

# B. Data background and Questions of Interest

Real estate is the one of the most important purchase in most people’s life, so it’s crucial to everyone to solve the mystery of the pricing problem for real estate. Lots of factors can influence the house price: The community the house is in, the traffic convenience, the education and supermarkets nearby, and the house price will also change over time. There are lots of data and too many for consideration that people can’t make easy decisions from them. In this project, we have several predictors about transaction date, age, MRT station distance, convenience store numbers, latitude and longitude.

We obtain the New Taipei real estate price dataset from 2012 to 2013 in Sindian District, New Taipei City from UCI machine learning repository, website: [https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set#](https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set).

This dataset is from the paper Building real estate valuation models with comparative approach through case-based reasoning. Applied Soft Computing, 65, 260-271 by professor I-Cheng Yeh from TamKang University. In this dataset we have 414 real estate price with 6 predictors:

X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.) X2=the house age (unit: year) X3=the distance to the nearest MRT station (unit: meter) X4=the number of convenience stores in the living circle on foot (integer) X5=the geographic coordinate, latitude. (unit: degree) X6=the geographic coordinate, longitude. (unit: degree)

The response variable is: Y = house price of unit area, (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared).

We want to build a model using them to predict house price. We will deliver an understandable regression model to analysis how these variables play an important role on pricing property. The question of interest are:

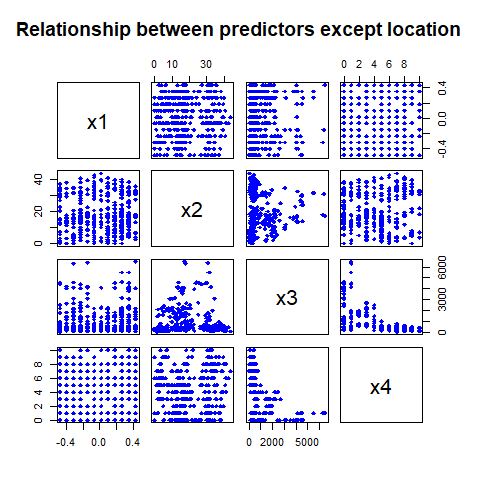
1. Does the predictor MRT station distance, convenience store numbers and location have significant effects on house price?
2. Is there a relationship between MRT station distance and convenience store numbers and location?
3. Are there any houses seem to be overestimated or underestimated?

# C. Exploratory Analysis

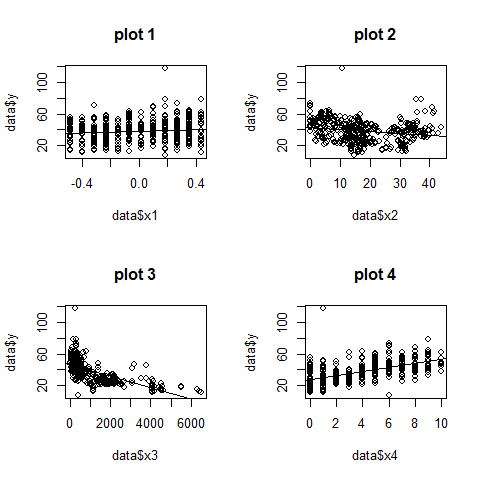
First we read in the data and inspect the first 6 rows of our dataset.

## # A tibble: 6 x 7  
## x1 x2 x3 x4 x5 x6 y  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2013. 32 84.9 10 25.0 122. 37.9  
## 2 2013. 19.5 307. 9 25.0 122. 42.2  
## 3 2014. 13.3 562. 5 25.0 122. 47.3  
## 4 2014. 13.3 562. 5 25.0 122. 54.8  
## 5 2013. 5 391. 5 25.0 122. 43.1  
## 6 2013. 7.1 2175. 3 25.0 122. 32.1

So we total have 6 predictors, while the lattitude and longtitude should have complex relationship between price and other predictors. We need to make a transformation of x1 to make it centered at 0, because otherwise they are too large to observe the small difference. Then first we can make a pairwise scatterplot with the other 4 predictors:

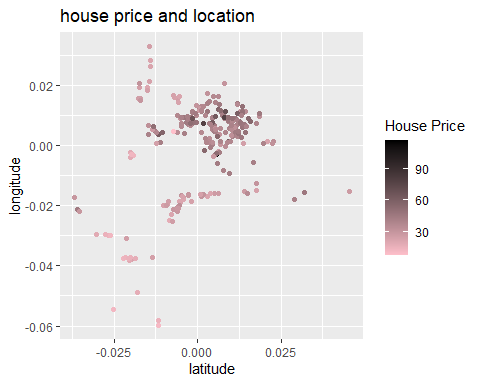


From this scatter plot, we can observe the relationship between x1(Transaction date) , x2 (house age), x3(the distance to nearest MRT station in meters) and x4 (convenience stores). As the scatter plots seems to have no linear relationship, the transaction date and age seems to have no relationship with any other predictors. It only seems significant that the distance to MRT is negative related with the convenience store numbers. Then we obtain the scatter plot for those predictors and the house price.



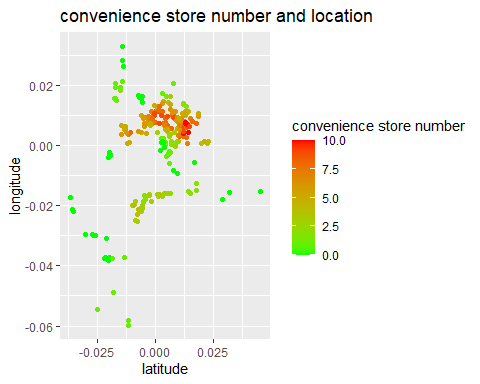
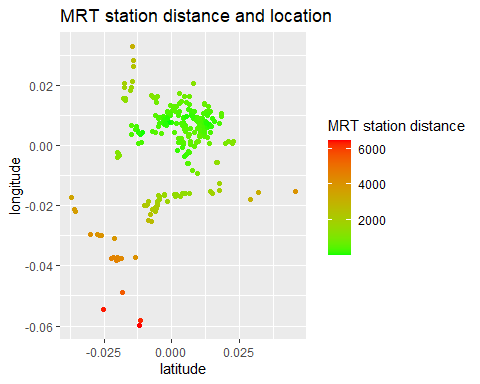
Plot 1 shows that the when transaction day go up, the house price has a small rise. Plot 2 shows the house price decrease when age increase, but some old house can be very expensive. Plot 3 shows a clear decreasing trend between the house price and the distance to the nearest MRT station. Houses within 1000 meters to an MRT station are much expensive than the houses far away from it. Plot 4 shows the relationship between the house price and the number of convenience stores. There is a slowly increasing trend between these two variables.

Then we draw the plot of the house price against the location (latitude and longitude)



From this plot we can clearly observe there is no linear relationship with the price and longtitude and latitude itself, but the trend is more related to the geology location. When the house is close to the center of the district, it tends to have higher price. (darker)

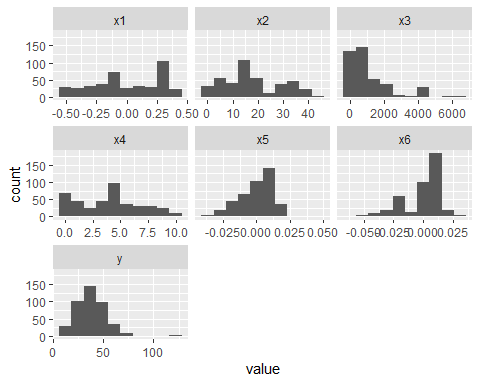
Then we inspect the correlation with location and the MRT station distance and the convenience store number:



These plots suggest perfect negative correlation of these two pairs. In the center, the distance is short, and the convenience store number is high. Reverse is for the suburbs.

# D. Data transformation and possible models

Here we use the histogram to check if we need some transformation in the dataset.



From this result, only x3 (distance near MRT station) is right skewed. The respond variable y is mainly normally distributed, not skewed. But I think it is not a good idea to transform because the houses close to the MRT station will not be influenced by the distance much, but after transformation, they will have a big difference. So we decide to not perform any transformation in the data except the location predictor.

From previous section we find there is no linear relationship between latitude, longitude and the price, but there can be a linear relationship of the price and a new variable: distance from the district center. Here we define the distance from the district center as the distance of a house from the population median of the latitude and longitude.

Another possible solution is that we can fit quadratic or qubic regression on the latitude and longitude. We will check them and compare their AIC in the next section.

# E. Model fitting

## Linear regression model

First we fit the model with x1, x2, x3, x4 and distance to the center x7:

In the previous section we find that there is positive correlation between price and mrt station, negative correlation between convenience store, so we want to check the collinearity of the predictors. We use the VIF to check the collinearity. VIF > 10 means that the variables may suffer from collinearity.

| Variables | Tolerance | VIF |
| --- | --- | --- |
| x1 | 0.9925043 | 1.007552 |
| x2 | 0.9271441 | 1.078581 |
| x3 | 0.0615488 | 16.247280 |
| x4 | 0.6300603 | 1.587150 |
| x7 | 0.0645707 | 15.486902 |

As the x3 (distance to MRT station) and x7 (distance to the center) both have >15 VIF, so we consider they have collinearity, thus we should remove the x3 in the final model.

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x4 + x7, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -38.397 -5.683 -1.304 4.254 76.776   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 46.29739 1.53270 30.206 < 2e-16 \*\*\*  
## x1 5.59878 1.59239 3.516 0.000487 \*\*\*  
## x2 -0.29660 0.03936 -7.535 3.17e-13 \*\*\*  
## x4 1.34586 0.18684 7.203 2.85e-12 \*\*\*  
## x7 -534.19679 43.09602 -12.396 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.1 on 409 degrees of freedom  
## Multiple R-squared: 0.5571, Adjusted R-squared: 0.5527   
## F-statistic: 128.6 on 4 and 409 DF, p-value: < 2.2e-16

Then we apply backward selection using AIC from the model, to inspect any predictors can be eliminated from the full model. From the AIC result, the full model has AIC 3010.27, which is smaller than the model with removal of any predictors. Thus the linear regression model should contain variables x1, x2, x4 and x7. The R square and adjusted R square of the model 1 is 0.5571 and 0.5527 respectively.

The summary of the linear regression model is below:

| Characteristic | Beta | 95% CI1 | p-value |
| --- | --- | --- | --- |
| (Intercept) | 46 | 43, 49 | <0.001 |
| transaction date | 5.6 | 2.5, 8.7 | <0.001 |
| house age | -0.30 | -0.37, -0.22 | <0.001 |
| convenience stores | 1.3 | 1.0, 1.7 | <0.001 |
| distance to center | -534 | -619, -449 | <0.001 |
| 1CI = Confidence Interval | | | |

## polynomial regression model

Then we also fit the model 2 with polynomial terms for the location predictor. The 2nd polynomial model has AIC 2953 and the 3rd polynomial model has AIC 2938, thus we should use the 3rd polynomial model as the polynomial model by choosing the model with smallest AIC. Also we test the collinearity and the resulting VIF table shows the variable x3 is still should be excluded. So the model 2 contains the variable x1, x2, x4 and 3rd polynomial term of x5 and x6.

| Variables | Tolerance | VIF |
| --- | --- | --- |
| x1 | 0.9725944 | 1.028178 |
| x2 | 0.9140732 | 1.094004 |
| x3 | 0.0204259 | 48.957467 |
| x4 | 0.5186326 | 1.928147 |
| poly(x5, 3, raw = TRUE)1 | 0.3145854 | 3.178787 |
| poly(x5, 3, raw = TRUE)2 | 0.2032087 | 4.921049 |
| poly(x5, 3, raw = TRUE)3 | 0.3626650 | 2.757366 |
| poly(x6, 3, raw = TRUE)1 | 0.2944368 | 3.396314 |
| poly(x6, 3, raw = TRUE)2 | 0.0093480 | 106.974771 |
| poly(x6, 3, raw = TRUE)3 | 0.0316382 | 31.607347 |

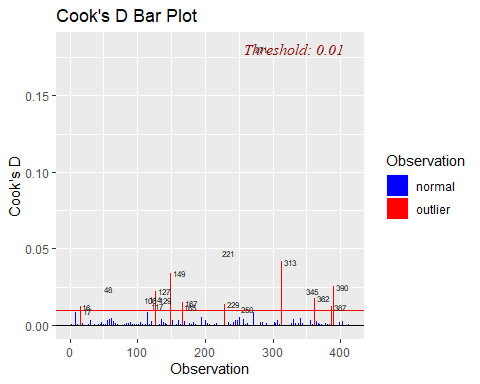
We summary those model 2 parameter below:

| Characteristic | Beta | 95% CI1 | p-value |
| --- | --- | --- | --- |
| (Intercept) | 43 | 40, 45 | <0.001 |
| transaction date | 4.7 | 1.7, 7.8 | 0.002 |
| house age | -0.31 | -0.39, -0.23 | <0.001 |
| convenience stores | 1.2 | 0.79, 1.5 | <0.001 |
| x5, 3 |  |  |  |
| x5, 3 | 429 | 307, 551 | <0.001 |
| x5, 3² | -3,644 | -8,006, 717 | 0.10 |
| x5, 3³ | -358,854 | -516,315, -201,394 | <0.001 |
| x6, 3 |  |  |  |
| x6, 3 | 93 | -7.9, 195 | 0.071 |
| x6, 3² | -17,830 | -24,731, -10,930 | <0.001 |
| x6, 3³ | -261,057 | -369,978, -152,136 | <0.001 |
| 1CI = Confidence Interval | | | |

Where the x5 and x6 in the table indicates latitude and longitude respectively. The AIC of the model 2 is 2983.785, R square is 0.5944 and Adjusted R-squared: 0.5854. From the first glimplse the model 2 is better than model 1 in AIC, R square and adj R square.

# F. Model diagnostic

First we detect the potential outliers by cook’s distance:

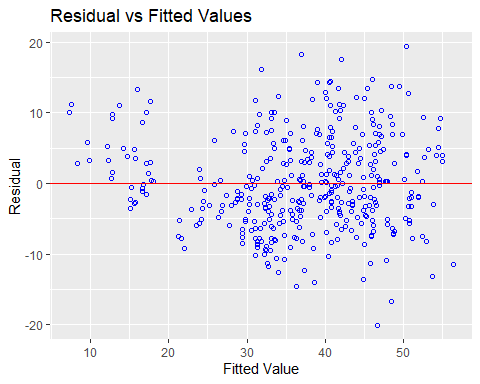


From the cook’s distance plot we can observe multiple outliers. We remove them and fit the model 1 and model 2 again:

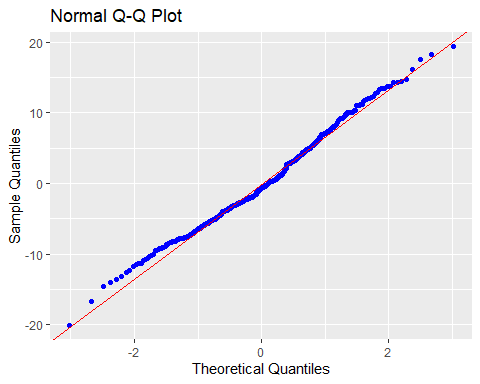
Now we check the model assumptions with the model fit on the dataset removing all outliers.

## Model 1

Residual fit plot: From the residual vs fitted plot, the we can find there is no non-linearity patterns in the residual, variance seems equal, and no obvious outlier.



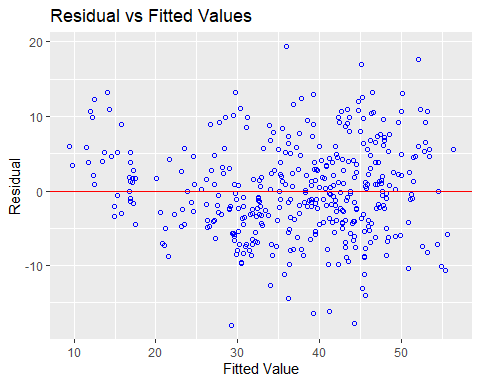
Residual qq plot: As the dots are approximately on the qqline, so the residual satisfy the normality assumption.

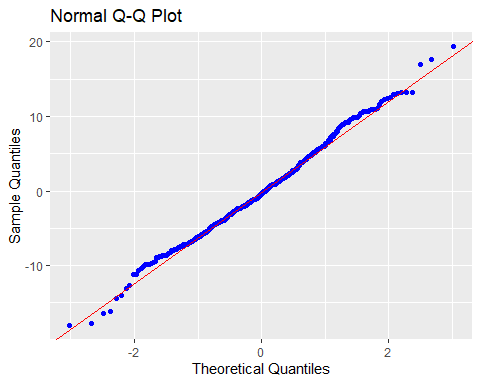


Breusch Pagan Test for Heteroskedasticity: We applied the Breusch Pagan Test, Ho: the variance is constant against Ha: the variance is not constant. The test statistic is 2.703788 with p value 0.100 > 0.05, so we fail to reject the null so there is no evidence of heteroskedasticity.

As the predictor contains 2 time variable, we want to test whether there is autocorrelation in the regression residual. We use the Durbin-Watson test with alternative hypothesis true autocorrelation is greater than 0. From the R result, the test statistic is 1.883 with p value 0.1236 > 0.05, so we fail to reject the H0 thus we have not enough evidence to state there is significant autocorrelation between the residuals.

## Model 2





The residual vs fitted plot and normal QQ plot for the model 2 is similar to the model 1. The Breusch Pagan Test for Heteroskedasticity return a test statistic 3.575 with p value 0.058 > 0.05, thus the variance is approximately equal. The Durbin-Watson test statistic is 1.9414 with p value 0.2827 also suggest no autocorrelation. So both models passed those diagnostics thus satisfy the regression assumption.

So we summary those 2 models here: Table

Description automatically generated

For the model selection conclusion, the AIC, R square and adjusted R square of model 2 is all better than model 1. So for statistic model 2 is the better model. We will use the model 2 for the following analysis. Still, the model 1 helps us to show that when a house is closer to the center, it tends to have higher price.

# G. Statistical analysis of the questions

## Does the predictor MRT station distance, convenience store numbers and location have significant effects on house price?

We use the model 2 result to analysis the predictor effects. From the model 2 result:

| Characteristic | Beta | 95% CI1 | p-value |
| --- | --- | --- | --- |
| (Intercept) | 42 | 40, 44 | <0.001 |
| transaction date | 3.0 | 0.73, 5.2 | 0.009 |
| house age | -0.36 | -0.42, -0.30 | <0.001 |
| convenience stores | 1.3 | 1.0, 1.5 | <0.001 |
| x5, 3 |  |  |  |
| x5, 3 | 359 | 269, 448 | <0.001 |
| x5, 3² | -4,246 | -7,555, -936 | 0.012 |
| x5, 3³ | -295,829 | -415,012, -176,645 | <0.001 |
| x6, 3 |  |  |  |
| x6, 3 | 112 | 38, 187 | 0.003 |
| x6, 3² | -17,034 | -22,085, -11,983 | <0.001 |
| x6, 3³ | -273,895 | -360,562, -187,228 | <0.001 |
| 1CI = Confidence Interval | | | |

The one sample t test of the convenience stores predictor has: null hypothesis: the slope of the convenience stores is 0. Alternative hypothesis: the slope of the convenience stores is not 0. As the p value is smaller than 0.001, we reject the null and conclude that the convenience store numbers have significant effect on house prices. In the context of the problem, on average when other conditions fixed, one increase in the convenience store numbers will increase the average price by 13000 New Taiwan Dollar/Ping. It’s a relatively high increase!

For the MRT station distance predictor, as it shows multicolinearity in both linear and polynomial model with the location variable, so we only need to analysis the location variable, then the effect of MRT station distance predictor can be obtained from the location variable, as when the location is closer to the center, distance from center is smaller, MRT station distance will decrease.

For the location variable, the model 2 (polynomial model) states that the location terms are significant. Though model 2 is better than model 1, but model 1 can help us identify how the location variable influence the price.

From the model 1, the slope of the distance variable x7 is significant and the slope is -511, which means 1 unit increase of the square root of 1 unit longitude and latitude will decrease the price of the house by 511 \* 10000 New Taiwan Dollar/Ping. The p value of smaller than 0.001 shows that effect is very significant.

| Characteristic | Beta | 95% CI1 | p-value |
| --- | --- | --- | --- |
| (Intercept) | 45 | 43, 48 | <0.001 |
| transaction date | 3.8 | 1.5, 6.2 | 0.001 |
| house age | -0.35 | -0.41, -0.29 | <0.001 |
| convenience stores | 1.5 | 1.2, 1.7 | <0.001 |
| x7 | -511 | -577, -445 | <0.001 |
| 1CI = Confidence Interval | | | |

## Is there a relationship between MRT station distance and convenience store numbers and location?

For the MRT station distance and location, there is collinearity between them, so there is very strong positive correlation between MRT station distance and distance to the district center. I guess probably only 1 MRT station in the middle of the district.

For the convenience store number, from the plot of convenience store number and the map location, we can observe that the closer to the center, the more the number.

We can also perform correlation test for their relationship: The Pearson’s product-moment correlation for x3 (MRT station distance) and x4 (convenience store number) have p value of smaller than 0.0001. The sample correlation is -0.6025 so they have strong negative correlation. The Pearson’s product-moment correlation for x7 (distance from center) and x4 (convenience store number) have p value of smaller than 0.0001. The sample correlation is -0.579 so they have strong negative correlation. The Pearson’s product-moment correlation for x7 (distance from center) and x3 (MRT station distance) have p value of smaller than 0.0001. The sample correlation is 0.9648 so they are probably the same variable.

## Are there any houses seem to be overestimated or underestimated? (Function part is located here)

Using the model 2, we can obtain a 95% prediction interval for the mean price of a given parameter set, so we can write a function to determine whether a given house is overestimated or underestimated.

# function to determine price  
# input x is the index in the data set  
price\_func = function(x){  
 input = data[x,]  
 pred = predict(model2, newdata = input, interval = "prediction")  
 if(input$y>pred[3])  
 {return(paste0("Overestimated, price is "  
 ,input$y\*10000,   
 " New Tai dollars/Ping, but the Confidence interval is ",  
 round(pred[1]\*10000), " New Tai dollars/Ping to " , round(pred[3]\*10000), " New Tai dollars/Ping" ))  
 }  
 else if(input$y>=pred[1])  
 {return(paste0("Fair price, price is "  
 ,input$y\*10000,   
 " New Tai dollars/Ping, and the Confidence interval is ",  
 round(pred[1]\*10000), " New Tai dollars/Ping to " , round(pred[3]\*10000), " New Tai dollars/Ping" ))  
 }   
 else if(input$y<pred[1])  
 {return(paste0("Underestimated, price is "  
 ,input$y\*10000,   
 " New Tai dollars/Ping, but the Confidence interval is ",  
 round(pred[1]\*10000), " New Tai dollars/Ping to " , round(pred[3]\*10000), " New Tai dollars/Ping" ))  
 }   
}  
# The number 14  
price\_func(14)

## [1] "Underestimated, price is 238000 New Tai dollars/Ping, but the Confidence interval is 238987 New Tai dollars/Ping to 363915 New Tai dollars/Ping"

# The number 36  
price\_func(36)

## [1] "Fair price, price is 273000 New Tai dollars/Ping, and the Confidence interval is 140869 New Tai dollars/Ping to 307272 New Tai dollars/Ping"

# The number 16  
price\_func(16)

## [1] "Overestimated, price is 505000 New Tai dollars/Ping, but the Confidence interval is 356962 New Tai dollars/Ping to 482016 New Tai dollars/Ping"

# F. Conclusion

In this project, after removing those outliers, we find the best model is the polynomial model 2, with predictors transaction date, house age, number of convenience stores in the living circle on foot (integer) and up to 3 polynomial terms for the latitude and longitude. The house price per unit area in Sindian District of NewTaipei City (10000 New Taiwan Dollar/Ping is the unit) will decrease 0.36 for every increased year in the house age, will increase 3.0 for every increased transaction data in year, increase 1.3 for every 1 more convenience stores nearby.

The number of convenience stores and the distance to MRT station has positive interaction with the distance from center, which means when the house is far from the center, the increase of convenience store will tend to increase the price more because of everyday convenience is more important in suburb area. But we remove the MRT station because it has multicolinearity with the location variable. Our model is reasonable and statistically significant. It also satisfies all of the assumptions quite well.

Future works can be done use more complex models. Because distance from the city center is only a crude predictor of the location for the house, so more complex models should be considered. Models should allow the city has multiple centers, and those houses in those centers would be significantly expensive than other houses. Polynomial model can interpret part of those influences but lack interpretation power.

# G. References

1. 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.
2. UCI machine learning repository, [https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set#](https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set)

# H. R code

knitr::opts\_chunk$set(  
 echo = FALSE,  
 error = FALSE,  
 message = FALSE,  
 warning = FALSE  
)  
library(car)  
library(ggplot2)  
library(MASS)  
library(readxl)  
library(olsrr)  
data <- read\_excel("Real estate valuation data set.xlsx", range = "B1:H415")  
colnames(data)[colnames(data)  
 == "X1 transaction date"] <- "x1"  
colnames(data)[colnames(data)   
 == "X2 house age"] <- "x2"  
colnames(data)[colnames(data)   
 == "X3 distance to the nearest MRT station"] <- "x3"  
colnames(data)[colnames(data)   
 == "X4 number of convenience stores"] <- "x4"  
colnames(data)[colnames(data)   
 == "X5 latitude"] <- "x5"  
colnames(data)[colnames(data)   
 == "X6 longitude"] <- "x6"  
colnames(data)[colnames(data)   
 == "Y house price of unit area"] <- "y"  
head(data)   
data$x1 = data$x1 - mean(data$x1)  
data$x5 = data$x5 - mean(data$x5)  
data$x6 = data$x6 - mean(data$x6)  
pairs(data[,1:4], col= "blue", pch=18, main= "Relationship between predictors except location")  
par(mfrow=c(2,2))  
plot(data$x1,data$y)  
abline(lm(y~x1,data))  
title("plot 1")  
plot(data$x2,data$y)  
abline(lm(y~x2,data))  
title("plot 2")  
plot(data$x3,data$y)  
abline(lm(y~x3,data))  
title("plot 3")  
plot(data$x4,data$y)  
abline(lm(y~x4,data))  
title("plot 4")  
ggplot(data,aes(x=x5,y=x6,color=y))+geom\_point()+  
 labs(  
 title = "house price and location",colour = "House Price",  
 x = "latitude", y = "longitude"  
) + scale\_color\_gradient(low="pink", high="black")  
ggplot(data,aes(x=x5,y=x6,color=x3))+geom\_point()+  
 labs(  
 title = "MRT station distance and location",colour = "MRT station distance",  
 x = "latitude", y = "longitude"  
) + scale\_color\_gradient(low="green", high="red")  
  
ggplot(data,aes(x=x5,y=x6,color=x4))+geom\_point()+  
 labs(  
 title = "convenience store number and location",colour = "convenience store number",  
 x = "latitude", y = "longitude"  
) + scale\_color\_gradient(low="green", high="red")  
library(tidyr)  
ggplot(gather(data), aes(value)) +   
 geom\_histogram(bins = 10) +   
 facet\_wrap(~key, scales = 'free\_x')  
data$x7 = sqrt((data$x5-median(data$x5))^2 +(data$x6-median(data$x6))^2)  
fit1 = lm(y~x1+x2+x3+x4+x7,data)  
#summary(fit1)  
l1 = ols\_coll\_diag(fit1)  
knitr::kable(l1$vif\_t)  
fit1 = lm(y~x1+x2+x4+x7,data)  
summary(fit1)  
stepAIC(fit1)  
#install.packages("gtsummary")  
library(gtsummary)  
tbl\_regression(fit1,  
 label=list(x1 ~ "transaction date",  
 x2 ~ "house age",  
 x4 ~ "convenience stores",  
 x7 ~ "distance to center"),  
 intercept=TRUE)  
data$x5 = data$x5 - median(data$x5)  
data$x6 = data$x6 - median(data$x6)  
fit2 = lm(y~x1+x2+x3+x4+  
 poly(x5,2,raw=TRUE)+  
 poly(x6,2,raw=TRUE),data)  
summary(fit2)  
AIC(fit2)  
  
fit3 = lm(y~x1+x2+x3+x4+  
 poly(x5,3,raw=TRUE)+  
 poly(x6,3,raw=TRUE),data)  
summary(fit3)  
AIC(fit3)  
  
l1 = ols\_coll\_diag(fit3)  
knitr::kable(l1$vif\_t)  
fit3 = lm(y~x1+x2+x4+  
 poly(x5,3,raw=TRUE)+  
 poly(x6,3,raw=TRUE),data)  
tbl\_regression(fit3,  
 label=list(x1 ~ "transaction date",  
 x2 ~ "house age",  
 x4 ~ "convenience stores"  
 ),  
 intercept=TRUE)  
l1 = ols\_plot\_cooksd\_bar(fit1)  
index = l1$data$color == "normal"  
data1 = data[index,]  
model1 = lm(y~x1+x2+x4+x7,data1)  
model2 = lm(y~x1+x2+x4+  
 poly(x5,3,raw=TRUE)+  
 poly(x6,3,raw=TRUE),data1)  
ols\_plot\_resid\_fit(model1)  
ols\_plot\_resid\_qq(model1)  
ols\_test\_breusch\_pagan(model1)  
library(lmtest)  
dwtest(model1)  
ols\_plot\_resid\_fit(model2)  
ols\_plot\_resid\_qq(model2)  
ols\_test\_breusch\_pagan(model2)  
dwtest(model2)  
library(stargazer)  
stargazer(model1,model2,out = "1.html")  
tbl\_regression(model2,  
 label=list(x1 ~ "transaction date",  
 x2 ~ "house age",  
 x4 ~ "convenience stores"  
 ),  
 intercept=TRUE)  
tbl\_regression(model1,  
 label=list(x1 ~ "transaction date",  
 x2 ~ "house age",  
 x4 ~ "convenience stores"  
 ),  
 intercept=TRUE)  
cor.test(data$x7, data$x3, method=c("pearson"))  
cor.test(data$x4, data$x3, method=c("pearson"))  
cor.test(data$x7, data$x4, method=c("pearson"))  
# function to determine price  
# input x is the index in the data set  
price\_func = function(x){  
 input = data[x,]  
 pred = predict(model2, newdata = input, interval = "prediction")  
 if(input$y>pred[3])  
 {return(paste0("Overestimated, price is "  
 ,input$y\*10000,   
 " New Tai dollars/Ping, but the Confidence interval is ",  
 round(pred[1]\*10000), " New Tai dollars/Ping to " , round(pred[3]\*10000), " New Tai dollars/Ping" ))  
 }  
 else if(input$y>=pred[1])  
 {return(paste0("Fair price, price is "  
 ,input$y\*10000,   
 " New Tai dollars/Ping, and the Confidence interval is ",  
 round(pred[1]\*10000), " New Tai dollars/Ping to " , round(pred[3]\*10000), " New Tai dollars/Ping" ))  
 }   
 else if(input$y<pred[1])  
 {return(paste0("Underestimated, price is "  
 ,input$y\*10000,   
 " New Tai dollars/Ping, but the Confidence interval is ",  
 round(pred[1]\*10000), " New Tai dollars/Ping to " , round(pred[3]\*10000), " New Tai dollars/Ping" ))  
 }   
}  
# The number 14  
price\_func(14)  
# The number 36  
price\_func(36)  
# The number 16  
price\_func(16)