**The Price Prediction Using Multiple Regression Model**

**Introduction**

These days, housing price in Asia varied upon geographical locations. According to the statement from ‘GlobalPropertyGuide’, that the house prices in TaiPei rose by 1.5% during the year to Q1 2019 [1]. To some extent, it is useful for buyers to know about a price estimate before making a transaction for it can reduce buyers’ traveling cost by comparing and deciding a desire price to relocate simply online. This project focuses on building such an model for prediction purpose by using multiple regression model.

**Procedure**

The procedure of the model involved in various steps listed below:

1. Prepare and mutate attributes in data
2. Visualize data
3. Construct model
4. Train and test the model
5. Evaluate model and make analysis and conclusion

The dimension (414x7) of dataset suggests the usage of multiple regression model for

more than one variable are stored in the dataset. The ‘Date’ variable is separated into its year and month portions. The year portion is coded as dummy variable (0:2012, 1:2013). The ‘month’ portion is transformed into its proper float value (stored as month\_new= month\*0.012). Three new variables ‘circle’, ‘avgst’ and ‘year\_x’ are added to the dataset and their metadata is listed below:

1. Circle: Area within the 1000 meter radius; center with the house unit
2. Avgst: Average numbers of stores in the circle area of the house unit
3. year\_x: The birth year of the house

**Visualization**

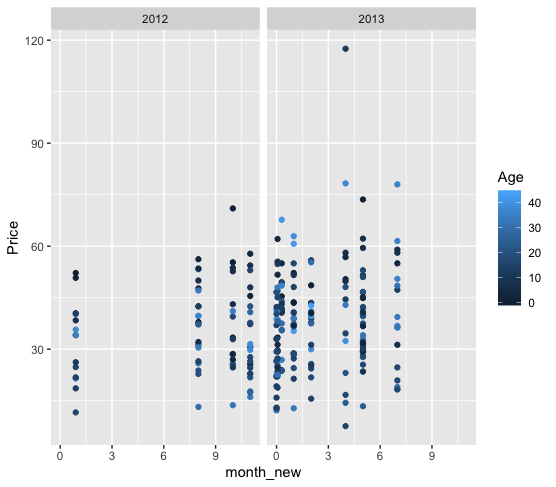
The data visualization began with a scatter plot of month and housing price. According to *No.1*, it is evident that the price of most houses are mostly sold beneath 60 NT/ft^2 and the neighborhood has a balanced set of young and old houses.

The bar plot *No.2* indicates that the highest number of stores nearby many housing units are range from 0 to 5 stores.

The histogram *No.3* of distance suggests that most housing units are less than 1900 meters away from the metro station. According to this piece of information, it is applicable to group housing units based on the ‘Distance’ variable for closer distance to metro indicates less traveling time. The result is shown in *No.4,* where housing units that are less than 1900 meters away from the metro stations are cluttered at coordinate (121.550,24.975) and they are a set of young and old houses.

The welch’s t test aims to find whether or not the average distance to metro station between 2012 and 2013 housing units differ as well as the average number of stores are distributed less in units for houses that are sold in 2012 comparing with houses that are sold in 2013. As a result, there is no evidence to support that the distance to metro station of houses in 2012 and 2013 are different (logged-p-value>0.05, 95% CI: -217.4656 to217.4656). Since the log-transformed mean and median are approximately the same (not shown in this report). It is safe to reverse transform the result. This leads to that the average distance to the metro station between two groups is equal, approximately 642.3915 meters.

In terms of average stores, there it is evident that houses that are sold in 2012 have a higher number of stores nearby comparing with houses that are sold in 2013 (p-value=0.05, 95%CI: -inf to 0.4469).

A close up of a device

Description automatically generated

*No.1 No.2*

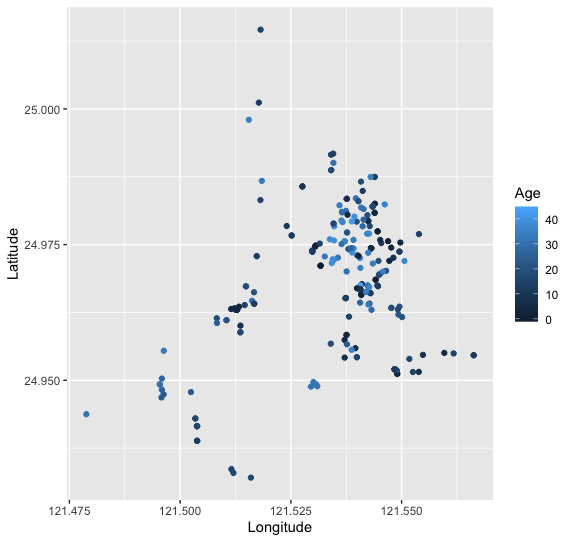
A close up of a logo

Description automatically generatedA screenshot of a cell phone

Description automatically generated

*No.3 No.4*

A close up of a white wall

Description automatically generated**

*No.5 No.6 Reference Plot*



**Regression Model**

One way to enhance the model performance in this project is to answer whether the month of the year can impact the price of the transaction. In terms of model, various steps involved in achieving a better model performance are listed below:

1. Apply mathematical functions to certain variables
2. Obtain regression results and its RMSE value
3. Perform Anova to test the significance of ‘month\_new’ variable
4. Use cook distance and residual plot to drop influential records
5. Find the model that explains the maximum amount of test dataset
6. Initial model

According to Anova, the absence of ‘month\_new’ variable increases the RSS slightly. In terms of hypothesis testing, there is no evidence to support that the full model (Regression\_full) is better than the reduced model (Regression\_reduced); (p-value>0.05). As a result, the reduced model is carried on for further improvement.



*Regression\_reduced Regression\_full*



1. Ultimate model

According to *group\_plot*, records such as 129 and 72 are dropped along with couple other influential points. The right set of *group\_plot* showed an adjusted overall pattern while not decreasing the predictive power of the model by much.

After several influential points are dropped accordingly from the visualization of *group\_plot*. The ultimate model explains ~73% of price variation from all explanatory variables. Comparing with the initial model, there is an 19.26% increase in adj-R2 value, a 26.38% decrease in train RMSE value, and a 0.3869% decrease in test RMSE value in the ultimate model. The ultimate model reduces the gap between train and test RMSE by 0.066 points.



A close up of a map

Description automatically generatedA close up of a map

Description automatically generated

*Group\_plot. Left; presence of influential points. Right; absence of influential points.*

**Analysis**

Along with many other details. This section discusses the detailed analysis mainly on:

1. Estimation of the geographical location of metro station
2. Model performance Evaluation

The pattern from No.1 suggests that the transaction price was not impacted by certain years or months when a certain transaction being made. This piece of information explained visually the non-significant impact of ‘month\_new’ variable in the regression model as well as this community housing price has not been fluctuating since 2012.

Patterns in No.4 implies that the west side of longitude 121.525 (No.6) obtains most 2013 houses with distance more than 1900 meters from the metro station. However, according to the pattern in No.5, there are numbers of 2013 houses mixed in with 2012 houses. The clutter provides the reason of equal mean distance from t-test are not reliable as the true distance from 2013 houses to the metro station was impacted by those that are cluttered with the 2012 houses. Thus, it is evident to say that the location of metro station must have satisfy a long range distance from most 2013 houses and a short distance from most 2012 houses. This leads to the conclusion that the metro station must have located on the south east or east direction from longitude 121.525 (No.6). The google map improved that the prediction of the location is correct.



*[2]: Green dot on the right indicates metro station*

*https://www.google.com/maps/place/Taipei,+Taiwan/@24.9750405,121.5407062,16z/data=!4m5!3m4!1s0x3442ac72bce20a99:0x3f6a35cedd0ac2e0!8m2!3d25.0329694!4d121.5654177.*

On behalf of regression model, all coefficients are significant in the final model (p-value<0.05). It is evident that housing price is strongly related to the average number of stores located in the circular area that centered from each house within the radius of 1000 meters after controlling all rest of independent variables (one unit increase in ‘avgst’ results in 8.846e4 increase in ‘log(price)’). The age and the distance to the metro station of average houses in Taipei impact the housing average price critically also. The increase in square root of both variables result in decrease of ‘log(price)’ . The RMSE value indicates the model is not over fitted.

**Conclusion**

The model explains test data fairly well. However, it includes mathematics functions that increase the complicity for further interpretation. As an future improvement, an increase in numbers of variables along with the size of dataset may reduce needs of mathematic functions in the model. As a result. It is evident that housing price in this particular neighborhood of Taipei are heavily influenced by its own housing characteristics as well as the access to transportations.

P.S. R-code template is attached to this report.

df<-read.csv('final\_training.csv')

dft<-read.csv('final\_test.csv')

df

dft

sum(is.na(df))

sum(is.na(dft))

library(dplyr)

library(tidyverse)

library(ggplot2)

library(broom)

library(psych)

library(fastDummies)

library(outliers)

##preparing train dataset

pairs(df3)

ggplot(df,aes(y=Latitude,x=Longitude, color=Age))+geom\_point()

ggplot(df,aes(y=Latitude,x=Longitude, color=Age))+geom\_point()+facet\_grid(~Distance<1900)

df1<-df%>%

separate(Date,into = c('year','month'))

df1[is.na(df1)]<-0

df1$year<-as.numeric(df1$year)

df1$month<-as.numeric(df1$month)

ggplot(df1,aes(y=Latitude,x=Longitude, color=Age))+geom\_point()+facet\_grid(~year)

df2<-df1%>%

mutate(month\_new=month\*0.012)

df2<-df2%>%

select(-c(Latitude,Longitude,No, month))%>%

mutate(circle=(1000\*\*2)\*pi)%>%

mutate(avgst=Stores/circle)%>%

mutate(year\_x=year-Age)

df3<-df2%>%

dummy\_columns('year', remove\_first\_dummy = T)%>%

select(-c(year))

#visualization

ggplot(df2,aes(Distance))+geom\_histogram(bins=30,binwidth = 200)

ggplot(df2,aes(x=month\_new,y=Price,color=Age))+geom\_point()+facet\_grid(~year)

ggplot(df2,aes(Stores))+geom\_histogram(bins=30)

#t test

year\_2012<-df2%>%

group\_by(year=2012)

year\_2013<-df2%>%

group\_by(year=2013)

t.test(log(year\_2012$Distance),log(year\_2013$Distance),alternative = 'two.sided',equal.variance=F,confidence.interval=0.95)

t.test(year\_2012$Stores,year\_2013$Stores,alternative = 'l',equal.variance=F)

describe(log(year\_2012$Distance))

describe(log(year\_2013$Distance))

exp(6.465198)

#preparing test dataset for analysis

dft1<-dft%>%

separate(Date,into = c('year','month'))

dft1[is.na(dft1)]<-0

dft1$year<-as.numeric(dft1$year)

dft1$month<-as.numeric(dft1$month)

dft2<-dft1%>%

mutate(month\_new=month\*0.012)%>%

mutate(circle=(1000\*\*2)\*pi)%>%

mutate(avgst=Stores/circle)%>%

mutate(year\_x=year-Age)

dft2<-dft2%>%

select(-c(Latitude,Longitude,No, month))

dft2

dft3<-dft2%>%

dummy\_columns('year', remove\_first\_dummy = F)%>%

select(-c(year,year\_2012))

dft3

#detecting outliers

lapply(df3,outlier)

df3[df3$Age==43.8,]

df3[df3$Distance==6396.283,]

df3[df3$Stores==10,]

df3[df3$Price==117.5,]

logg=log(df3$Price)

exp(logg[logg<2.5])

boxplot(log(df3$Price))

df3[df3$Price==11.6,]

df3[df3$Price==7.6,]

#critical influential points dropped

df3<-df3[-c(161,71,72,29,129,164,91,62,166,209,105,137),,]

#regressors, ANOVA

regressor=lm(log(Price)~avgst+sqrt(Age)+sqrt(Distance)+year\_x+year\_2013, data=df3)

regressor1=lm(log(Price)~avgst+sqrt(Age)+sqrt(Distance)+year\_x+year\_2013+month\_new, data=df3)

summary(regressor)

summary(regressor1)

anova(regressor,regressor1)

prediction\_train=predict(regressor,df3)

prediction\_test=predict(regressor,dft3)

regressor%>%

augment()%>%

mutate(res=.fitted-log.Price.)%>%

mutate(rsqr=.resid\*\*2)%>%

summarize(mse=mean(rsqr))%>%

mutate(rmse=sqrt(mse))

sqrt(mean((log(dft3$Price)-prediction\_test)\*\*2))

##plot section

par(mfrow=c(2,2))

plot(regressor,1:2)

plot(regressor,4)

length(df3$Price)

lm1hat = hatvalues(regressor) # hatvalues

length(lm1hat)

id.lm1hat = which(lm1hat > 2\*length(lm$coefficients)/length(df1$Price))

lm1hat[id.lm1hat]

plot(df3$Price,lm1hat)

abline(h=2\*length(lm1$coefficients)/length(vec))

#95%CI

confint(regressor)

**Reference**

[1]: Global Property Guide. “Taiwan's Housing Market - Overvalued, but Rising.” *Global Property Guide*, 8 Aug. 2019, https://www.globalpropertyguide.com/Asia/Taiwan/Price-History.