Regression Modeling for Predicting Optimal Real Estate Prices in Chile

MSBA310 – Applied Statistical Analysis

Dr. Imad Bou Hamad

Fall 2023

By:

Lara Baltaji

Hadi Knaiber

Shadi Youssef

Abstract

Real estate industry in Chile has been thriving in the past few years. However, we do not exactly know what factors determine the price of residential properties in Chile as predictive characteristics differ between one country and another. For that, we used data from a Real Estate company in Chile in order to determine significant predictors and applied statistical analysis and predictive modeling. The first aim of this study is to check which variables (floor number, typology, view orientation, size, and region) are significantly correlated with the sale price of apartments and can be used for predicting the prices. The second aim is to employ a predictive model using a multiple linear regression technique which uses significant predictors for estimating and predicting sale price of residential properties. The results of the study show that all the variables are significantly correlated with the price per meter squared except for the region (Norte and Santiago). The results also reveal that there exists a negative correlation between floor number and price per meter squared. This is likely due to the fact that Chile is subject to earthquakes, and that the demand for lower floor apartments is higher as they provide better safety from earthquakes. We employed two linear regression models, one including all the variables and the other all except the region. Despite the fact that region is not significantly correlated with price, the first model performed better in terms of prediction and region turned out to be a significant predictor after all. Future research can use additional variables for prediction models such as the age and specific location of the apartment. They can also use other predictive techniques such as Random Forest, Support Vector Machine and Neural Networks which can provide better options for prediction.

1 Introduction

The Real Estate industry has become one of the most popular industries in the world, and investing in real estates is considered one of the best ways to secure money. The Chilean Real Estate Market has been considered a market in demand for a very long time now. People choose Chile for Real Estate not only for its spectacular scenery and fascinating culture, but mainly for its thriving economy. The Republic of Chile is situated in the Western part of South America between the Andes to the east and the Pacific Ocean to the west as seen in **Figure 1**. The map of Chile shows a very unique shape of the Chilean land which is long and narrow.

Due to the effects of the coronavirus pandemic, the Chilean economy shrank by 5.8% in 2020 which eventually affected the Real Estate Market according to the World Bank. However, Chile's economy was able to recover after strict policy support by the Chilean government. Later, inflation has risen after the Russian invasion on Ukraine which as a result affected the prices in the Real Estate industry.

This paper uses a multiple linear regression model to analyze the real estate property market in Santiago and Norte, Chile, for the aim of predicting list price based on certain predictors. Here, it is important to clarify the difference between the list price and the sale price. The list price is the initial price amount of a property offered by a market. The sale price, on the other hand, is the transaction price amount which is the amount the consumer has actually paid for the property. Usually, sale prices are lower than list prices as they are affected by negotiation techniques. The objective of this study is to determine the estimated list prices of properties using a dataset collected by a company called Disca. Choosing the appropriate listing prices of properties is crucial not only for the buyer and the seller, but also for the loan provider or payer involved in the transaction.

Disca is a leading Real Estate company in Chile that has several constructing projects. They construct the apartments from A to Z including all structural calculations, land preparations for construction, and the construction process until the apartments become ready for sale. In this project, we will use statistical analysis and predictive modelling tools in order to check what the significant predictors of the price of real estate apartments are in Chile.

The remaining sections of the paper are organized as follows. Section 2 reviews existing research on the Chilean Real Estate market; Section 3 describes the present problem including traditional prediction models for Real Estate prices in Chile; Section 4 describes the data used; Section 5 discusses the results and limitations of the method used; Section 6 presents our conclusions as well as future recommendation and research.

2 Literature Review

Existing studies have explored various aspects of Chile's Real Estate sector for different purposes. Some focused on studying the residential property market in Santiago as being the capital and economic focus of Chile such as Masias et.al (2016) and Vergara-Perucich and Aguirre-Núñez (2019). Others focused on studying Real Estate in Chile as a whole such Perucich (2021).

The Real Estate sector is considered a significant component in Chile's economy due to its vital role in balancing household spending and financial systems according to Parrado and Cox (2008). As in many other countries in the world, housing is a key factor of household's wealth and it incorporates the household's major contribution to loans by the financial systems (Parrado & Cox, 2008). According to the National Socioeconomic Characterization survey, almost 70% of Chilean households own the property that they live in, while only 18% are renters. In addition, 73% of the people who own their houses have fully paid the costs, whereas 27% are still paying a mortgage loan (Parrado & Cox, 2008). All the above percentages serve as proof that the Chilean economy not only depends on the purchasing power of consumers in the Real Estate sectors, but also on the dependency on loans by banking and financial sectors in order to cover the expenses of bought residential properties.

Chile is located in the Western part of South America, a location that is extremely exposed to earthquakes. Natural disasters have long been a limitation to Real Estate development, especially in Chile. For example, the 1985 earthquake that hit Santiago caused multiple damages in the housing sector. The reconstruction of building took years because of the delicate economic situation at the time (Vergara-Perucich & Aguirre-Núñez, 2019). The typologies of building changed over the years. After the earthquake, building construction was maximum up to 5 floors since high floors are critically dangerous in the case of earthquakes.

Later, buildings started to increase in heights and after 1995, buildings with less than 3 floors were rarely built (Vergara-Perucich & Aguirre-Núñez, 2019).

3 Problem Description

The major factor which people consider when buying their household properties is the cost of living. For that, we can notice in major cities all over the world, the suburbs are congested with people more than the city itself. That is because the cost of living in suburbs is much cheaper than that of the city, especially the city center. The costs also include the pricing of real estates.

The prices of residential properties depend on many factors including but not limited to age of the property, size, view, floor number, typology, and location. In order to determine the main factors that affect the price of apartments in Chile we decided to go over a dataset by Disca Real Estate Company and apply statistical tools which reveal correlations between the variables. Then, we applied two multiple linear regression models based on the variables of the dataset and computed the accuracy of prediction in order to choose the one which offers the least error.

Existing studies have explored different prediction models for estimating housing prices in Chile. According to Masias et al. (2016), most Santiago housing market studies use econometric methods to estimate market price of residential properties. Some use linear regression models and others use semi-log and log-log models in their estimation (Masias, et al., 2016). Mosias et al. (2016) used different hedonic models such as Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN) and Multiple Linear Regression (MLR) and compared the predictive performance. Their results reveal that the Random Forest (RF) produced the best predictions for Santiago housing prices. The results, however, were close in accuracy; meaning that all three models can be used for prediction.

4 Data Description

Our data set consists of several variables, some are quantitative such as floor number, size, and price, and others are qualitative such as project name and region.

- Project name refers to the project under which the apartment comes from
- Floor represents the floor number of the apartment
- Status refers to whether the apartment is sold, available, reserved or promised.
- Product_id refers to the unique ID number that differentiates every apartment

- Bedrooms refers to the number of bedrooms per apartment
- Bathrooms refers to the number of bathrooms per apartment
- The view orientation refers to the compass direction of the view of each apartment
- Size per m^2 presents the size of each apartment in m^2
- Initial price posted is the listed price of the apartment in UF Chileno.
- Initial price per m^2 is the listed price of the apartment in m^2
- Sold price is the sale price of the sold apartment in UF Chileno.
- Sold price per m squared is the sale price of the apartment in m^2
- Region is the region were the apartment is located (Norte or Santiago)

The dataset was taken from Disca, a Real Estate Company in Chile which is responsible for numerous Real Estate projects in the north and in Santiago.

5 Results and Discussion

After testing and exploring the data, we came up with several insights and generated multiple results. At the beginning, we ran a quantile-quantile plot test to see if the prices are normally distributed among the two studied regions. The results of the plot (**Figure 2**) reveal that many points are on the line, however some points deviated. Therefore, the distribution is close to being normal and is not skewed, but is not perfectly normal.

First, we inspected the descriptive statistical information of the **sold price per meter squared** and found the following: Prices of sold apartments by this real estate company were 48.41 UF Chileno (1,921$) per meter squared on average, with the least selling price of an apartment being 28.82 UF Chileno which is equivalent to approximately 1,135 USD. We noticed that 25% of the apartments were sold at a price that is less than or equal to 41.56 UF Chileno (1,635$) and almost half the apartments were sold at a price that is less than or equal to 48.57 UF Chileno (1,911$) and almost 75% of the apartments had a selling price that is less than or equal to 54.72 UF Chileno (2,153$). Looking at the summary of the results, we saw that the highest sold price was 73.33 UF Chileno (2,886$). The difference of the sold prices between apartments was 8.63 UF Chileno (339$) on average and assuming the data was normally distributed, 68% of the apartments have a maintenance cost that is between 39.78 (1565$) and 57.04 (2245$).

The same procedure and analysis was done on the initial **list prices** of apartments. Surprisingly, the results were exactly the same as the former ones. We deduce here that the prices of the apartments are fixed and therefore, non-negotiable. So, it does not matter which price variable we use because the listed prices are the same as that of the sold prices.
The previous can be visualized graphically with the boxplot and histogram we made (**Figure 3**).

After that, we created a histogram for the sale price per meter squared (**Figure 4**) to inspect the distribution of the prices per meter squared. We saw that the highest frequency of sold apartments was at almost 50 UF Chileno per meter squared. The data visualization represented a slightly skewed distribution, so it is close to being normal. The lowest number of apartments sold were at a price of almost below 30 UFC and above 70 UFC. In addition to that, there were no outliers in the figures presented, which also confirms that the data is close to being normal.

In order to see if there is a difference in selling price of an apartment in the two studied regions, which are Norte and Santiago, we computed the mean price of an apartment across the two regions. The results were almost equal with a very slight increase in Norte (46.58) compared to Santiago (48.36). The stability of the currency displayed (UF Chileno) plays a role in this equality. UF Chileno is a unit of account that is used in Chile. UF Chileno is a non-circulating currency in which the rate of exchange between the UF Chileno and the original Chilean currency (Peso) is adjusted for the value of the UF to remain constant for inflation. Usually, it is used in bank loans, financing, purchases and investments. It also became the predominant currency to use in the real estate industry and in valuing houses and properties, whether private or belonging to the Chilean government. The bar graph visualization represents the mean price per meter squared in both regions, Norte and Santiago (**Figure 5**).

Afterwards, we delved deeper into the data and studied relationships and correlations between the variables in hand. We checked if there is a correlation between the floor number and price. After analyzing the data and applying a scatter plot (**Figure 6**), we saw that as floor number increases, the price decreases. Therefore, we can conclude that there is a negative correlation between floor number and price per area. To validate this, we conducted a correlation test between the selling price and the floor number, the results verify that there is a significant negative correlation between price and floor number. This means that as the floor increases, price decreases and vice versa. After contacting the company in Chile to ask about the reason, they

informed us that earthquakes are the main reason why floor number affects price negatively. Chile is one of the few locations on Earth where three major earthquake plates (Nazca, South American Tectonic plates and Antarctic Plate) meet; in other words, Chile is a "triple junction". Therefore, earthquakes occur in Chile on a regular basis with a frequency of almost 18 major earthquakes and 2 million minor earthquakes yearly. Generally, it is safer to be on lower floors when an earthquake happens since there are less wave vibrations on lower floors and this in turn means more stability and less destruction. That is the main reason why as floor number increases, price decreases and vice versa.

After that, we wanted to check if there is correlation between the number of bedrooms and price per m^2. Since price per $m^2$is numerical while number of bedrooms is categorical, we used boxplot in order to check for any correlation. After analyzing the boxplot generated (**Figure 7**) we can notice that the medians and quartiles differ significantly (for instance the median of one bedroom is higher than that of two bedrooms) for that we expect a strong correlation between the number of bedrooms and price per m^2.

In order to validate our assumption that there is correlation between the number of bedrooms and the price per m^2 we performed an ANOVA test since it is the test used in comparing a quantitative and qualitative variable. The results showed that p-value is less than alpha (0.05). So, we reject H_0. Therefore there exists a significant correlation between number of bedrooms and the price.

We expect a correlation between the number of bathrooms and the price per m^2 Similarly, we decided to dig deeper and see if there is a correlation between the number of bathrooms and the price per m^2. For that, and similar to what we did with the number of bedrooms and the price per m^2 we used boxplot to see whether there is a relationship between number of bathrooms and the price after analyzing the boxplot (**Figure 8**), We can see a significant difference in the medians and quartiles of the boxplot, so we expect a significant correlation between the number of bathrooms and the price

In order to validate our assumption, we performed ANOVA test (similar to the case of bedrooms) and we noticed that p-value is less than 0.05 which is alpha. Thus we reject H_0.So, we are 95% confident that there is a significant correlation between number of bathrooms and the price.

Moreover, we checked the potential relationship between the two quantitative variables size and price per area by plotting a scatter plot. The results show a negative decreasing trend line passing through the points (**Figure 9**) which indicates that there seems to be a negative correlation between the two variables. Therefore, we deduct that the bigger the size, the less price per meter squared. To verify the correlation between the 2 variables, we conduct a correlation test. The result shows a P-value that is way less than alpha (2.2e-16 < 0.05), so we confirm that there is a negative correlation between the sold price and size with a correlation result of a negative value. Therefore, as size increases, price decreases and vice versa. This is because costly utilities such as bathrooms and kitchens for example that require infrastructure will be the same in the big apartments as that of the small, therefore, the price per area will be higher in smaller apartments compared to the big apartments.

Adding up on the previous variables, we decided to go further by checking if there is a relationship between the view orientation and the price. Theoretically, there must be a correlation since view is one of the most factors that enter in marketing a real estate. For that, we used a box plot since our variables are categorical (view orientation) and numerical (price). By analyzing the box plot (**Figure 10**), we can tell that the medians and quartiles are significantly different, thus we assume that there is a correlation between the view orientation and the price.
In order to validate our assumption, we performed ANOVA test since it best fits the comparison. The results showed that the p-value is less than alpha. So, we are 95% confident that there exists a significant correlation between the price and the view orientation.

Reviewing the correlation between price per area and region, we constructed a side-by-side boxplot between the 2 variables. We notice that the median in Norte region was a bit higher than that of Santiago and conclude that there is no significant correlation between price and region. This goes back to the stability of the currency being used, which is UF Chileno. Following that, we performed an ANOVA test to confirm the correlation. The results show that the P-value is greater than alpha (0.726 > 0.05); thus, we conclude that there exists no significant correlation between price and region.

To study the significant predictors that might affect price, we built a multiple linear regression model (MLR) that studied all the variables we have and how they might affect price. After running the model, there were several significant results. All the significant results had a P-value that is less than alpha which show that the variable can be considered as a predictor and the

opposite is true. To start, we investigated the number of bedrooms and bathrooms. Number of bedrooms was considered as a significant predictor and as the apartment increases by one bedroom, the price per area increases by 11.2 UF Chileno. Concerning bathrooms, it was also considered as a significant predictor yet, as number of bathrooms increases by 1 unit, price decreases by 5 UF Chileno. The size was also a significant predictor in which as size increases, the price decreases. When the area increases by 1 unit, the price decreases by 0.3 UF Chileno.

Regarding the orientation, all the directions were studied and the significant ones were the apartments that had views to the north, the ones that had views to the north, east and west, and the ones that had west, south and east views. These orientations caused an increase in the price with 2.89, 13.33 and 11.5 UF Chileno respectively. All other orientations were considered as insignificant and did not qualify to be potential predictors.

The floor number was regarded as a significant predictor in which as the apartment goes up by a floor, its value per area decreases approximately 0.08 units. The Santiago region was also a significant predictor where prices in Santiago were less by almost 6.67 units.

After calculating the Root Mean Square Error (RMSE) of this model, its result was 6.9 and comparing this number to two times the mean of the sold price per area which is 96.8, the RMSE is much less and hence we can infer that the model is significant and can be used.

6 Conclusion and Recommendations

In conclusion, there are multiple factors that affect the real estate business and several factors that predict the price of an apartment. The main ones are the size, number of bedrooms, number of bathrooms, view orientation, and floor. Unlike the common trend, in Chile, as the floor gets higher, the price decreases. That is because of its geographic location that coincides with three earthquake tectonic plates. Despite the fact that region is not significantly correlated with price, it is a significant predictor. Including the variable region in our model resulted in a lower RMSE and this means that the model has a better predictive accuracy. This may be due to the multicollinearity of the region variable with some other predictors.

There are limitations in our study. First, the listed prices were exactly the same as the sold prices, suggesting that prices are non-negotiable and fixed. Usually, this is not what happens since in most cases as negotiations take place regularly, even if on a minimal scale. Another

limitation would be that the data includes only two regions in Chile. The studied analysis might or might not be the same in other regions. Moreover, the age of each property is not put in the list and it is known world-wide that the age of the apartment plays a significant role in the pricing.

Future research can use additional variables for prediction models such as the age and specific location of the apartment. Moreover, our study used the multiple linear regression model only, yet, trying other predictive techniques may provide better estimation and predictive accuracy and thus may support this study.

The real estate industry is growing with time because of the increasing demand for housing worldwide due to the general overgrowth of populations. Moreover, the rush of investors to invest in real estate has become a popular trend and this is due to the economic decline worldwide, especially in devaluation of currencies and the spread of inflations in many countries. More studies and analyses are required in the field of real estate as the trends that are shaping the future of this industry appear to be booming.

7 References

Garreton, M. (2017). City profile: Actually existing neoliberalism in Greater Santiago. ScienceDirect, 32-50.

Liu, G. (2022). Research on Prediction and Analysis of Real Estate Market. Hindawi Scientific Programming, 1-8.

Masias, V. H., Valle, M. A., Crespo, F., Crespo, R., Vargas, A., & Laengle, S. (2016). Property Valuation using Machine Learning Algorithms: A Study in a Metropolitan-Area in Chile. AMSE Conference Santiago/Chile (pp. 97-105). Santiago: ResearchGate.

Parrado, E., & Cox, P. (2008). Evolution of Housing Prices in Chile. Economia Chilena.

Perucich, F. V. (2021). Urban Determining Factors of Housing Prices in Chile: A Statistical Exploration. Revista Urbano.

Rosales, I., & Hernández, C. (2021). Building Models to Predict Real Estate List Prices using Ensemble Machine Learning Algorithms. International Conference on Industrial Engineering and Operations Management. Rome: IEOM Society International.

Vergara-Perucich, F., & Aguirre-Núñez, C. (2019). Housing Prices in Unregulated Markets: Study On Verticalised Dwellings in Santiago De Chile. Preprints .

Vergara-Perucich, J.-F. (2021). Typological Study of Financialized Housing in Chile: Verticalization in Estación Central. Civil Engineering and Architecture 9(3), 611-624.

8 Appendix

Figures:

Figure 1: Map of Chile

Figure 2: QQ-plot of Sale Price

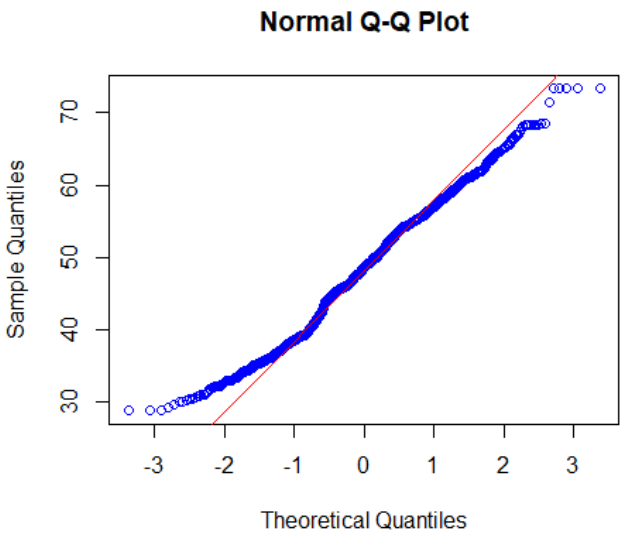


Figure 3:

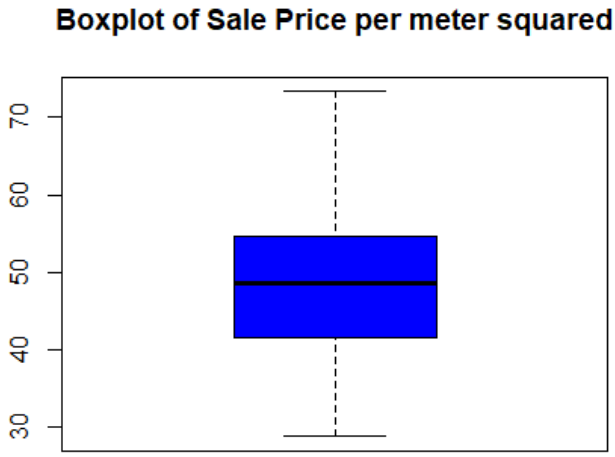


Figure 4:

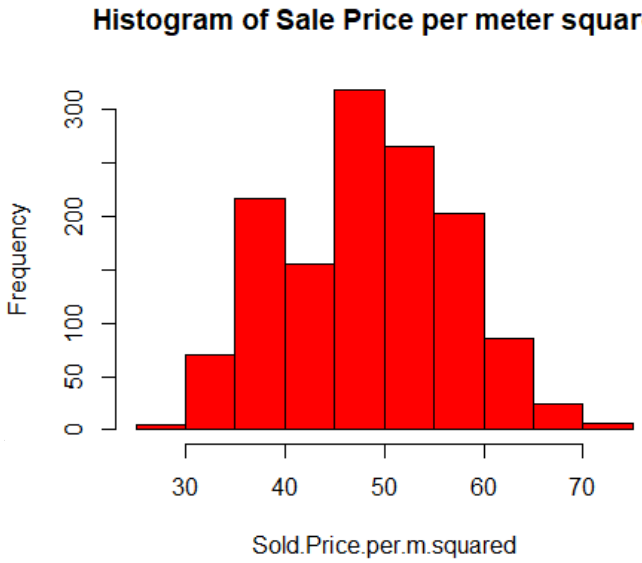


Figure 5:

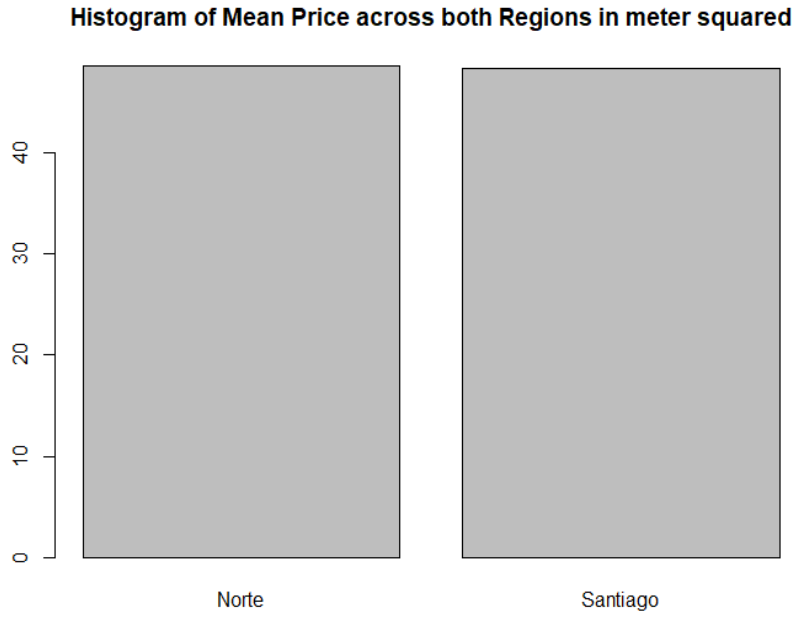


Figure 6:

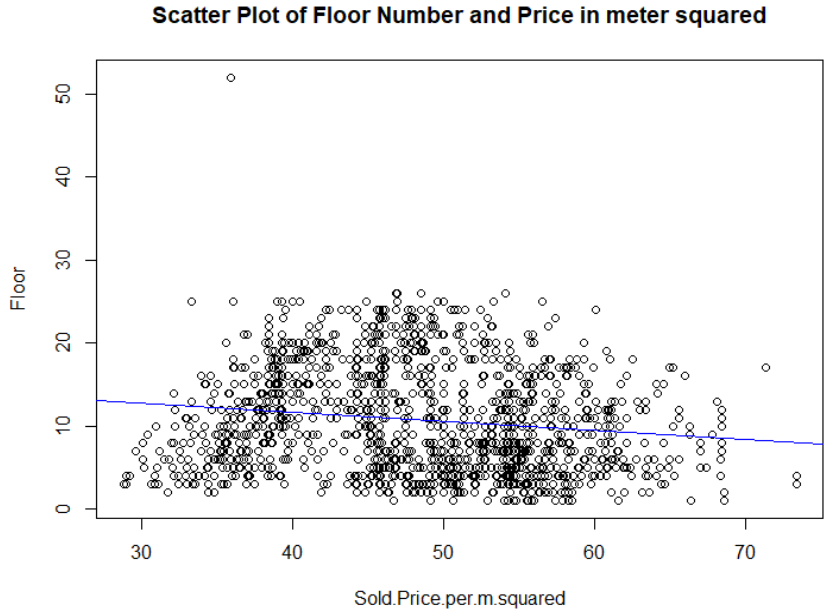


Figure 7:

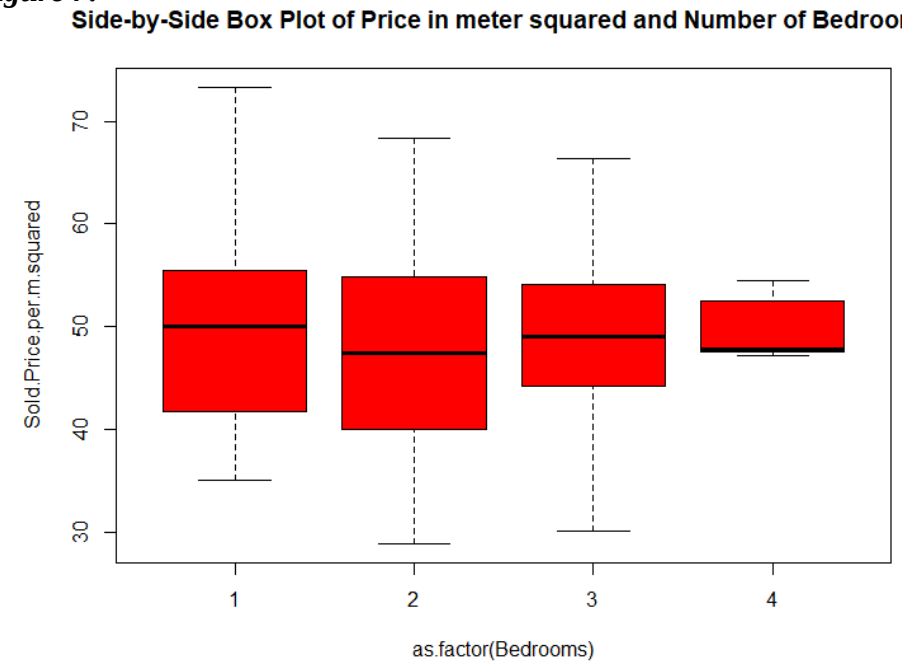


Figure 8:

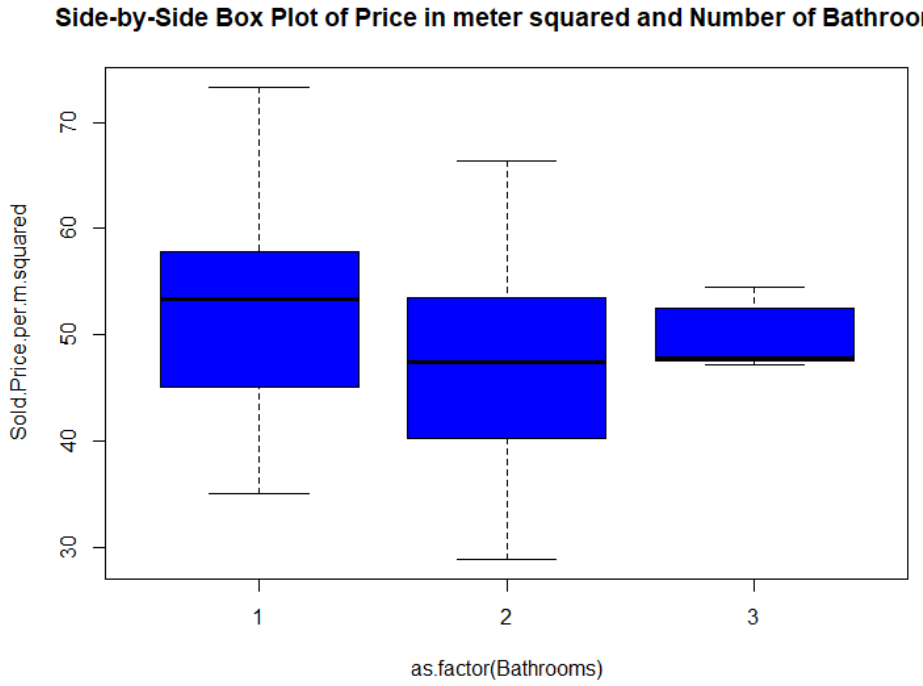


Figure 9:

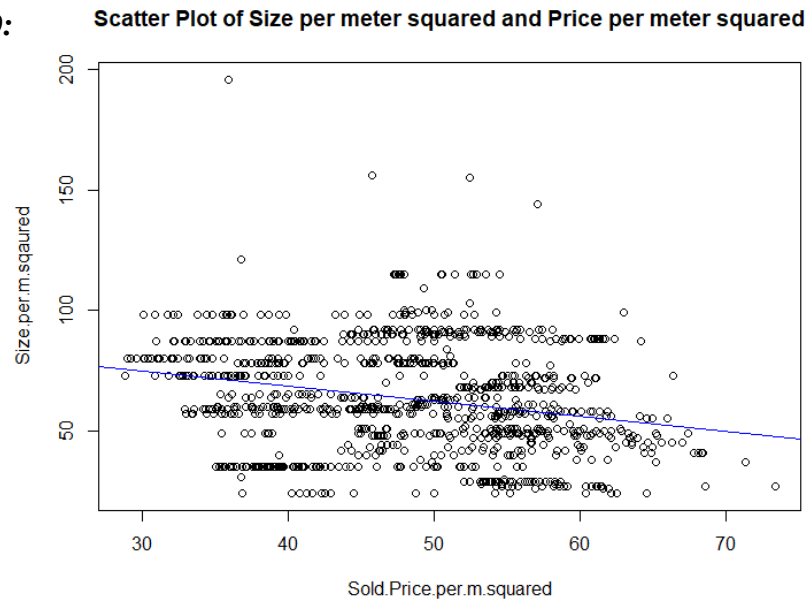


Figure 10:

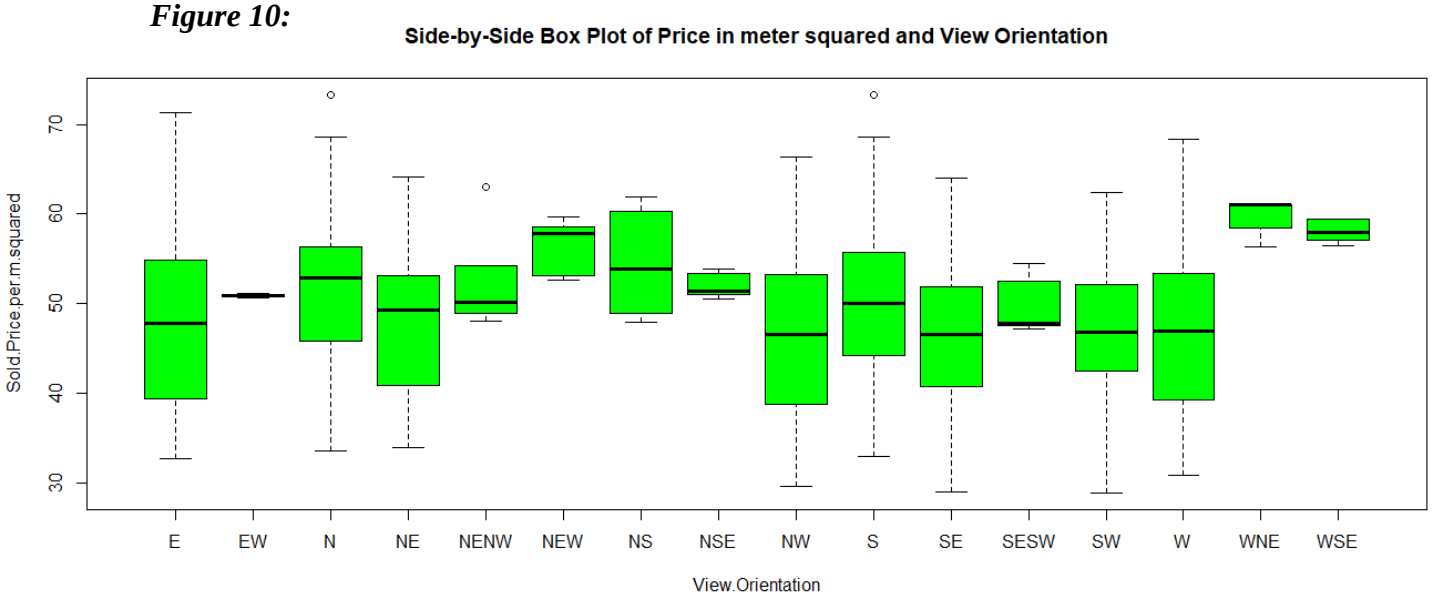
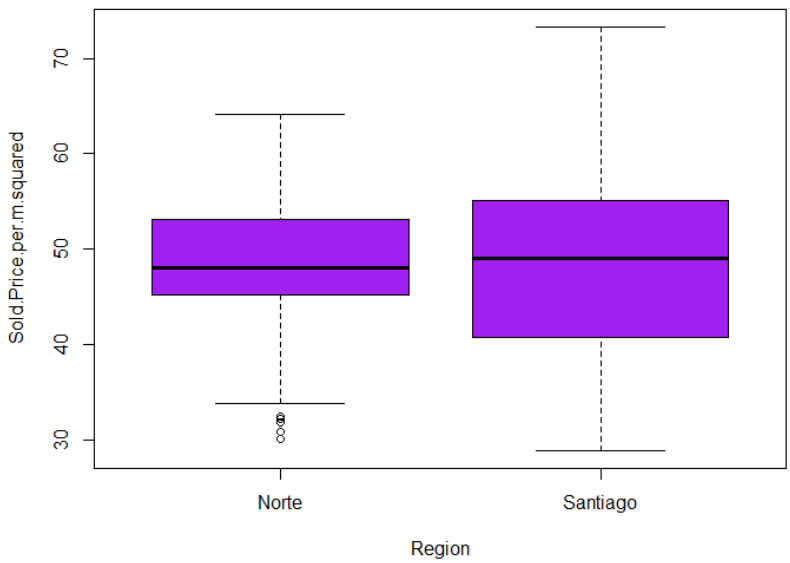


Figure 11:



R code:

```
> data <- read.csv("C:\\Users\\Lenovo\\Desktop\\MSBA\\Applied Statistics
  MSBA310\\Project\\Real Estate Chile Cleaned Data.csv")
> ## How many rows does the dataset include?
> nrow(data2)
[1] 1350
> nrow(data)
[1] 2264
> #1350 apartments are sold out of 2264 apartments
> ## Filter the data for the sold apartments only.
> data2 <- data[data$Status=="Sold",]
> summary(Sold.Price.in.UF.Chileno)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    884    2340    2898    3020    3664    8218
> summary(Initial.Price..posted..in.UF.Chileno)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    884    2340    2898    3020    3664    8218
> #We can notice that the list prices and the sale prices are exactly the same
  which means that prices are fixed in this company (non-negotiable)
> ## Are there any missing values?
> attach(data2)
> data2[!complete.cases(data2)]
data frame with 0 columns and 1350 rows
> #No missing values
> ## Report the mean, standard deviation, and quartiles for the sold price per
  m2. Interpret
> summary(Sold.Price.per.m.squared)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  28.82   41.56   48.57   48.41   54.72   73.33
> #The least sold price of an apartment is  28.82 UF Chileno
> #25% of apartments have an sold price less than or equal to 41.56 UN Chileno
> #50% of apartments have an sold price less than or equal to 48.57 UF Chileno
> #75% of apartments have an sold price less than or equal to 54.72 UF Chileno
> #The highest sold price is 73.33 UF Chileno
> #The Average price of an apartment by this Real Estate Company is 48.41 UF
  Chileno
> sd(Sold.Price.per.m.squared)
```

```
[1] 8.628078
> # The standard deviation is 8.63 which means that the difference of initial
  prices between apartments is 8.63 on average
> # Assuming that the data is normally distributed, 68% of apartments have a
  maintenance cost between (48.41 - 8.63) = 39.78 and (48.41 + 8.63) = 57.04
> ## Draw histogram and boxplot for the sold price per m2. Comment on the shape
  of distribution
> hist(Sold.Price.per.m.squared, col = "Red")
> boxplot(Sold.Price.per.m.squared, col = "blue")
> #No Outliers. The data seems to be normally distributed
> ## Draw a qqplot to determine if the Price is normally distributed
> qqnorm(Sold.Price.per.m.squared, col = "blue")
> qqline(Sold.Price.per.m.squared, col = "red")
> # The dataset is not perfectly normally distributed, but it is close to being
  normal as it is not skewed.
> ##    Compute the mean price of an apartment across the two regions. Comment
> table <- aggregate(Sold.Price.per.m.squared, list(Region), FUN=mean)
> barplot(table$x, names.arg = table$Group.1)
> #The mean price seems to be similar between both regions
> ## Draw a scatter plot between floor number and price per m2. Comment
> plot(Floor ~ Sold.Price.per.m.squared)
> abline(lm(Floor ~ Sold.Price.per.m.squared), col= "blue")
> #There is a negative relationship between price per m2 and the floor number
> ## Is there a correlation between the price per m2 and the floor number?
> cor.test(Sold.Price.per.m.squared, Floor)


  Pearson's product-moment correlation


data:  Sold.Price.per.m.squared and Floor
t = -5.3882, df = 1348, p-value = 8.389e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.19702801 -0.09256721
sample estimates:
       cor
-0.1452022
> #there seems to be significant correlation as the p-value is less than alpha
> ## Draw a boxplot that shows the price per m2 with the number of bedrooms?
```

```
> boxplot(Sold.Price.per.m.squared ~ as.factor(Bedrooms), col ="red")
> #We expect a correlation between the number of bedrooms and the price per m2
> ## Is there a correlation between the number of bedrooms and the price?
> effect = aov(Sold.Price.per.m.squared~ as.factor(Bedrooms))
> summary(effect)
                      Df Sum Sq Mean Sq F value Pr(>F)
as.factor(Bedrooms)    3    654  217.86   2.939 0.0322 *
Residuals           1346  99771   74.12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #p-value is less than 0.05. So there is a significant correlation between
  number of bedrooms and the price
> ## Draw a boxplot that shows the price per m2 with the number of bathrooms?
> boxplot(Sold.Price.per.m.squared ~ as.factor(Bathrooms), col ="blue")
> #We expect a correlation between the number of bathrooms and the price per m2


> ## Is there a correlation between the number of bathrooms and the price?
> effect2 = aov(Sold.Price.per.m.squared~ as.factor(Bathrooms))
> summary(effect2)
                       Df Sum Sq Mean Sq F value Pr(>F)
as.factor(Bathrooms)    2   6154    3077   43.97 <2e-16 ***
Residuals            1347  94271      70
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #p-value is less than 0.05. So there is a significant correlation between
  number of bedrooms and the price


> ## Draw a scatter plot between  size per m2 and price per m2. Comment
> plot(Size.per.m.sqaured ~ Sold.Price.per.m.squared)
> abline(lm(Size.per.m.sqaured~Sold.Price.per.m.squared), col="blue")
> #Negatively Correlated, the bigger the size, the smaller the price per m2


> ## Is there a correlation between the price per m2 and the size per m2?
> cor.test(Sold.Price.per.m.squared, Size.per.m.sqaured)

  Pearson's product-moment correlation
```

```
data:  Sold.Price.per.m.squared and Size.per.m.sqaured
t = -9.4853, df = 1348, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2994917 -0.1994459
sample estimates:
       cor
-0.2501364
> #p-value is less than alpha, so there exists a significant correlation
  between size per m2 and size


> ## Draw a boxplot that shows the price per m2 with the View Orientation.
> boxplot(Sold.Price.per.m.squared ~ View.Orientation, col = "green")
> #There seems to be a difference in the prices according to the view
  orientation


> ## Is there a correlation between the view orientation and the price?
> effect3 = aov(Sold.Price.per.m.squared~ View.Orientation)
> summary(effect3)
                   Df Sum Sq Mean Sq F value   Pr(>F)
View.Orientation   15   6657   443.8   6.313 4.32e-13 ***
Residuals        1334  93768    70.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #p-value is less than alpha. So there exists a significant correlation
  between the price and the view orientation


> ## Draw a boxplot that shows the price per m2 with the Region.
> boxplot(Sold.Price.per.m.squared ~ Region, col = "purple")
> #There seems to be some outliers in Norte. We expect a small correlation
  between the two


> ## Is there a correlation between the region and the price?
> effect4 = aov(Sold.Price.per.m.squared~ Region)
> summary(effect4)
              Df Sum Sq Mean Sq F value Pr(>F)
Region         1      9    9.16   0.123  0.726
Residuals   1348 100415   74.49
```

```
> #p-value is 0.726 greater than alpha. So there exist no significant
  correlation between the two.
> #We can try to remove the outliers in Norte and check the correlation again


> acc_error<- function(actual,pred){
+   mape <- mean(abs((actual - pred)/actual))*100
+   mae=mean(abs(actual-pred))
+   RMSE= sqrt(mean((actual-pred)^2))
+
+   vec=c(mape,mae, RMSE)
+
+   names(vec)= c("MAPE", "MAE", "RMSE")
+   return(vec)
+ }
> ## Split the data into 70% for training and 30% for validation
> set.seed(100)
> split=sample(1:2, nrow(data2), replace = TRUE, prob=c(0.7, 0.3))
> train=data2[split==1, ]
> val=data2[split==2, ]
> ## Apply a linear regression model using all variables as predictors.
> model1 <- lm(Sold.Price.per.m.squared ~ Floor + Bedrooms + Bathrooms +
  View.Orientation + Size.per.m.sqaured + Region, data = train)
> summary(model1)


Call:
lm(formula = Sold.Price.per.m.squared ~ Floor + Bedrooms + Bathrooms +
    View.Orientation + Size.per.m.sqaured + Region, data = train)


Residuals:
    Min      1Q  Median      3Q     Max
-17.583  -4.124   0.311   3.832  34.308


Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         63.64393    1.43677  44.297  < 2e-16 ***
Floor               -0.08148    0.03782  -2.154 0.031475 *
Bedrooms            11.28730    0.77747  14.518  < 2e-16 ***
```

```
Bathrooms             -5.19105    0.90135  -5.759 1.16e-08 ***
View.OrientationEW    -0.43174    4.26597  -0.101 0.919409
View.OrientationN      2.89790    0.86714   3.342 0.000866 ***
View.OrientationNE     0.06082    1.22760   0.050 0.960497
View.OrientationNENW   4.48755    3.11260   1.442 0.149721
View.OrientationNEW   13.33966    4.13882   3.223 0.001314 **
View.OrientationNS     6.46754    4.29997   1.504 0.132906
View.OrientationNSE    6.42644    3.37313   1.905 0.057072 .
View.OrientationNW    -1.56570    1.20291  -1.302 0.193386
View.OrientationS      0.73446    0.90346   0.813 0.416468
View.OrientationSE    -1.41980    1.21675  -1.167 0.243568
View.OrientationSESW   2.18901    2.61499   0.837 0.402758
View.OrientationSW    -2.24715    1.20636  -1.863 0.062821 .
View.OrientationW      0.24764    0.92842   0.267 0.789734
View.OrientationWNE    7.18978    3.69073   1.948 0.051715 .
View.OrientationWSE   11.59312    3.35294   3.458 0.000570 ***
Size.per.m.sqaured    -0.38024    0.02804 -13.562  < 2e-16 ***
RegionSantiago        -6.67914    0.81010  -8.245 5.77e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 7.038 on 907 degrees of freedom
Multiple R-squared:  0.3333,    Adjusted R-squared:  0.3186
F-statistic: 22.67 on 20 and 907 DF,  p-value: < 2.2e-16


> ## Apply a linear regression model using only the significant predictors
  which we determined previously using correlation tests.
> model2 <- lm(Sold.Price.per.m.squared ~ Floor + Bedrooms + Bathrooms +
  View.Orientation + Size.per.m.sqaured, data = train)
> summary(model2)
Call:
lm(formula = Sold.Price.per.m.squared ~ Floor + Bedrooms + Bathrooms +
    View.Orientation + Size.per.m.sqaured, data = train)


Residuals:
     Min       1Q   Median       3Q      Max
-16.2886  -4.4420   0.4247   4.8203  24.6356
```

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          55.70772    1.10531  50.400  < 2e-16 ***
Floor                -0.08424    0.03919  -2.150  0.03185 *
Bedrooms             11.05474    0.80511  13.731  < 2e-16 ***
Bathrooms            -7.17948    0.89994  -7.978 4.48e-15 ***
View.OrientationEW   -3.93869    4.39848  -0.895  0.37077
View.OrientationN     2.22889    0.89461   2.491  0.01290 *
View.OrientationNE   -1.50916    1.25668  -1.201  0.23010
View.OrientationNENW  6.16977    3.21842   1.917  0.05555 .
View.OrientationNEW  11.77034    4.28422   2.747  0.00613 **
View.OrientationNS    1.52020    4.41214   0.345  0.73051
View.OrientationNSE   2.47468    3.45985   0.715  0.47464
View.OrientationNW   -2.03066    1.24512  -1.631  0.10326
View.OrientationS     0.34303    0.93490   0.367  0.71377
View.OrientationSE   -2.62648    1.25168  -2.098  0.03615 *
View.OrientationSESW  4.54129    2.69355   1.686  0.09214 .
View.OrientationSW   -2.76754    1.24835  -2.217  0.02687 *
View.OrientationW     0.81368    0.95941   0.848  0.39661
View.OrientationWNE   5.14068    3.81575   1.347  0.17824
View.OrientationWSE   8.14552    3.44727   2.363  0.01834 *
Size.per.m.sqaured   -0.27136    0.02563 -10.588  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 7.293 on 908 degrees of freedom
Multiple R-squared:  0.2833,    Adjusted R-squared:  0.2683
F-statistic: 18.89 on 19 and 908 DF,  p-value: < 2.2e-16
> ## Study the predictive performances of models 1 and 2.
> pred1=predict(model1,val)
> perform1=acc_error(val$Sold.Price.per.m.squared, pred1)
> perform1
     MAPE       MAE      RMSE
11.877046  5.464962  6.901039
> pred2=predict(model2,val)
> perform2=acc_error(val$Sold.Price.per.m.squared, pred2)
> perform2
```

```
     MAPE       MAE      RMSE
12.860932  5.900873  7.364444
> #RMSE of model 2 is bigger than that of model 1. Therefore model 1 has a
  better predictive performance even though it includes Regions which is
  relatively has an insignificant correlation with price.
>##Compare both RMSE's with double of the Price's mean. Conclude.
> 2*mean(Sold.Price.per.m.squared)
   [1] 96.81139
>The RMSE of the chosen model is almost 7 which is considered acceptable
  compared with two times the mean (7%)
```