Melbourne Housing

Final Project

Kajal Talele

Applied Regression

Guided by

Professor Margo Bergman

Table of Contents

Abstract

The housing market plays an important role in the global economy. Housing prices continue to change from day to day and are sometimes raised rather than based on calculations. Research has shown that fluctuations in housing prices often affect homeowners and the housing market. The goal of our project is to predict housing prices in Melbourne using several statistical/machine learning prediction models. In total, 5 statistical models are built, one of them being the categorical variable of linear regression models. First, data is cleaned and explained using the principles of Exploratory Data Analysis. The target variable from the given dataset is Price. The best-performing model, among all models used for the project, is the Linear regression model. This report will also comprehensively validate multiple techniques in model implementation on regression and provide an optimistic result for housing price prediction. The predicted price is compared with the actual price to obtain the score for the model (coefficient of determination R2). The best-performing model has an R2 of 0.50. The findings of this analysis confirmed the use of Linear Regression as the most efficient model.

Introduction

Melbourne, Australia is facing rapid growth population growth but is facing concerns with the affordability of housing. As property prices in Melbourn**e** have become so expensive, it has become difficult for low and middle-income households to afford to house and fewer and fewer people are purchasing houses. With prices constantly changing due to different factors regarding the facilities, the price will be our dependent variable. One of the main aspects buyers, sellers, and investors are all interested in is how many bedrooms and bathrooms there are; quite simply because the more there are, the more expensive the property will be. In addition to this, things like garages, pools, backyards, air-conditioning, patios, and balconies, the list is endless, and all factor into a property's value (What, 2022). Facilities are also important when considering location: a city apartment will add value if it has a parking space, while a suburban home with a large backyard is perfect for families, and again adds value (What, 2022). Since housing price is determined by various factors, our scope of the project is to find out the key factors affecting house prices in Melbourne. In the dataset, our independent variables include house sizes, number of bathrooms, number of bedrooms, number of car spaces**, etc.** We will be using these various features to build a regression model to predict house prices in Melbourne.

Data Set

This dataset was discovered on Kaggle and is created through historic data available for house prices in an urban environment -Melbourne Housing in Australia.

Kaggle Dataset: Housing Dataset | Kaggle

I. Variable List:

Variable Name	Number of Observations	Variable Type	Description
Suburb	13580	String	Residential area on the outskirts
Address (Dropped)	13580	String	The physical address of the house
Rooms	13580	Integer	Number of rooms in the house
Type (Dropped)	13580	String	Structural type of the house
Price	13580	Numeric	Sale price of the house
Method (Dropped)	13580	String	The method the house was sold
SellerG (Dropped)	13580	String	Real Estate Agent
Date	13580	String (Timestamp)	The selling date of the house
Distance	13580	Numeric	The distance of the house
Postcode (Dropped)	13580	Numeric	The Postal Code of the house
Bedroom2	13580	Integer	The number of bedroom of the house
Bathroom	13580	Integer	The number of bathroom of the house

Car	13580	Integer	Number of car spots
Landsize	13580	Numeric	The total land size of the house
BuildingArea (Dropped)	7130	Numeric	The building area of the house
YearBuilt (Dropped)	8205	Numeric	The year that the house was built
CouncilArea (Dropped)	12211	String	Governing council for the area
Latitude (Dropped)	13580	Float	The North-South position of the house on the earth
Longitude (Dropped)	13580	Float	The East-West position of the house on the earth
Regionname	13580	String	A particular area that are broadly divided by physical characteristics
Propertycount	13580	Integer	Number of property objects that are in the properties collection

Number of variables: 10

Our team performed the primary analysis on all variables, we had 21 columns, but we removed one by one variable which was the least important to this project which is explained in our previous part of the Dataset. There were such columns that could have given a higher correlation between two variables, so our team removed them one by one variable to prepare our final data. We explored and refined our research question based on the available data. We decided to drop 10 variables which are as follows with reasons:

1. Type: Type just provides information on housing type, which has the least importance for this project.

2. Method: We don't want to take the status of selling a house into consideration, it would not have much impact on our research question.
3. SellerG: The name of the seller is of least importance and doesn't add value for this project to predict the housing price.
4. BuildingArea: Building area was dropped because there are only 7130 numbers of observations, which means that almost half of the data has null values.
5. YearBuilt: Year Built was dropped because there are only 8205 numbers of observations, which means that almost half of the data has null values.
6. CouncilArea, Latitude, Longitude, Address, PostCode: These variables are highly correlated to "Suburb" and "Regionname" categorical variables in our dataset so that they would not add additional analysis values to our project.

Categorical variable:

Region name

```
[1] "Northern Metropolitan"      "Western Metropolitan"
[3] "Southern Metropolitan"      "Eastern Metropolitan"
[5] "South-Eastern Metropolitan" "Eastern Victoria"
[7] "Northern Victoria"          "Western Victoria"
```

For Regionname, we think it does not require grouping because they are 8 in total numbers, and we will be utilizing them as categorical variables in our project

II. Data Cleaning

Below are Data Cleaning steps we've taken in R:

1. Exclude Type, Method, SellerG, BuildingArea, CouncilArea, Latitude, Longitude, Address, PostCode, and Year Built columns by sub-setting the columns.

2. Drop any missing values by using [complete.cases] function.

4. We extracted the categories by breaking them into subcategories using unique(CatLotData2$Suburb) code for Suburbs

5. We extracted the categories by breaking them into subcategories using unique(CatLotData2$Regionname) code for Regionname

6. Downloaded the CSV file with write_csv function

Data Analysis Plan

Our team performed a step-by-step analysis of the data. When we received the data from our IT department, we started exploring the data and their variables. We also took several steps to conclude our final data analysis techniques and plan. Following are the steps that our team took to answer our research questions.

I.Data Preparation in R studio

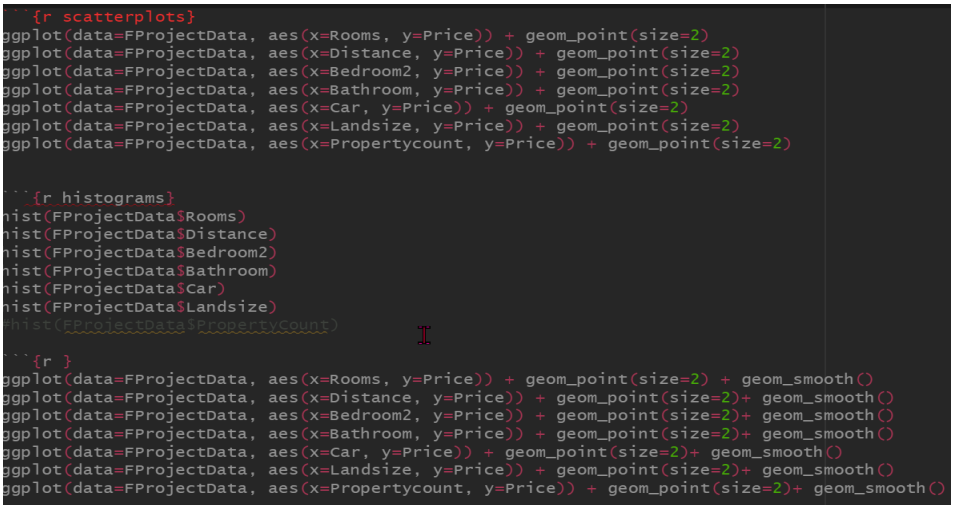
Data preparation is the most critical step in the analysis before regression testing and analyzing the results.

Step 1

Our team performed a Univariate analysis (Descriptive analysis) of all the numerical independent variables such as car, distance, land size, bedrooms, bathrooms, etc. to analyze summary statistics using the functions; summary(FProjectData) and View(FProjectData).

Step 2

In the next step, we created Scatter plots and Residual plots (Bivariate Analysis) between the label 'Price' and independent variables to check if there is any non-linearity in the data. Moreover, our team also created Histograms of each variable to observe skewness issues.



```
```{r scatterplots}
ggplot(data=FProjectData, aes(x=Rooms, y=Price)) + geom_point(size=2)
ggplot(data=FProjectData, aes(x=Distance, y=Price)) + geom_point(size=2)
ggplot(data=FProjectData, aes(x=Bedroom2, y=Price)) + geom_point(size=2)
ggplot(data=FProjectData, aes(x=Bathroom, y=Price)) + geom_point(size=2)
ggplot(data=FProjectData, aes(x=Car, y=Price)) + geom_point(size=2)
ggplot(data=FProjectData, aes(x=Landsize, y=Price)) + geom_point(size=2)
ggplot(data=FProjectData, aes(x=Propertycount, y=Price)) + geom_point(size=2)


```{r histograms}
hist(FProjectData$Rooms)
hist(FProjectData$Distance)
hist(FProjectData$Bedroom2)
hist(FProjectData$Bathroom)
hist(FProjectData$Car)
hist(FProjectData$Landsize)
#hist(FProjectData$Propertycount)

```{r }
ggplot(data=FProjectData, aes(x=Rooms, y=Price)) + geom_point(size=2) + geom_smooth()
ggplot(data=FProjectData, aes(x=Distance, y=Price)) + geom_point(size=2)+ geom_smooth()
ggplot(data=FProjectData, aes(x=Bedroom2, y=Price)) + geom_point(size=2)+ geom_smooth()
ggplot(data=FProjectData, aes(x=Bathroom, y=Price)) + geom_point(size=2)+ geom_smooth()
ggplot(data=FProjectData, aes(x=Car, y=Price)) + geom_point(size=2)+ geom_smooth()
ggplot(data=FProjectData, aes(x=Landsize, y=Price)) + geom_point(size=2)+ geom_smooth()
ggplot(data=FProjectData, aes(x=Propertycount, y=Price)) + geom_point(size=2)+ geom_smooth()
```

Step 3

In this step, our team selected a subset of the data and deleted missing data to make the final data more accurate. Later, we also got a new descriptive summary on our variables.
```

```
Select a subset of data

```{r select subset}
FProject2 <- FProjectData %>% select(Price, Rooms, Bedroom2, Bathroom, Car, Distance, Landsize,
Propertycount, Regionname)

```

Delete missing data

```{r delete Missing}
summary(FProject2)
FProject3 <-  FProject2[complete.cases(FProject2), ]
```
```

```
     Price             Rooms          Bedroom2        Bathroom          Car           Distance
 Min.   :  85000   Min.   : 1.00   Min.   : 0.00   Min.   :0.00   Min.   : 0.00   Min.   : 0.0
 1st Qu.: 650000   1st Qu.: 2.00   1st Qu.: 2.00   1st Qu.:1.00   1st Qu.: 1.00   1st Qu.: 6.2
 Median : 901000   Median : 3.00   Median : 3.00   Median :1.00   Median : 2.00   Median : 9.2
 Mean   :1074796   Mean   : 2.94   Mean   : 2.92   Mean   :1.54   Mean   : 1.61   Mean   :10.2
 3rd Qu.:1328000   3rd Qu.: 3.00   3rd Qu.: 3.00   3rd Qu.:2.00   3rd Qu.: 2.00   3rd Qu.:13.0
 Max.   :9000000   Max.   :10.00   Max.   :20.00   Max.   :8.00   Max.   :10.00   Max.   :48.1
    Landsize      Propertycount    Regionname
 Min.   :     0   Min.   :  249   Length:13518
 1st Qu.:   178   1st Qu.: 4380   Class :character
 Median :   442   Median : 6567   Mode  :character
 Mean   :   558   Mean   : 7456
 3rd Qu.:   651   3rd Qu.:10331
 Max.   :433014   Max.   :21650
```

II. Selection of Regression techniques and model

After data preparation and preliminary analysis, our team selected appropriate regression techniques and models by using the following steps. We also checked if there are any significant interaction variables that are affecting the model fit. Lastly, our team determined which variables are not significant by eliminating the variables one by one from the model by analyzing the P values, F-Stat, and their significance.

Step 1 - As we know that our data has numerical and Categorical variables, our team performed Multiple Linear Regressions to test the significant variables using p-value and R2 values. Our team took all the numerical variables to test the fit of the model. **The R2 value we got here is 0.395** with 'Propertcount' less significant among all.

Regression 1

(HouseFit1<lm(Price~Rooms+Distance+Bedroom2+Bathroom+Car+Landsize+Propertycount, data=FProject3)

summary(HouseFit1)

Step 2 - In this step, in order to improve the R2 value, we dropped one variable (Propertycount) as it was the least significant variable and again ran the model**. Our results showed exactly the same R2 Value.**

Regression 2

HouseFit2<-lm(Price~Rooms+Distance+Bedroom2+Bathroom+Car+Landsize, data=FProject3)

summary(HouseFit2)

Step 3 - In this step, we included a categorical variable (Regionmame) with previous variables to see our model fit. **Our R2 value increased from 0.395 to 0.502.**

Regression 3

HouseFit3<-lm(Price~Rooms+Distance+Bedroom2+Bathroom+Car+Landsize+Regionname, data=FProject3)

summary(HouseFit3)

Step 4 - In this step, we performed transformations on our variables and created histograms using the log function on our variables, and we found that the log of those variables corrects the skewness problem (explained in the results). As we have many zeros in our data, the log function worked on only the rooms variable and it increased our model. **This model increased our R2 value from 0.502 to 0.51.** All variables were significant except Regionname Western Victoria.

Regression 4

HouseFit4<lm(Price~log(Rooms)+Distance+Bathroom+Car+Regionname+Landsize+Bedroom2, data=FProject3)

summary(HouseFit4)

Step 5 - In this step, we saw all variables are significant except categorical one variable Regionname Western Victoria so we used the following functions to separate it from other dummy variables and ran the model again. Our R2 value changed from 0.51 to 0.506 with all the variables significant. And this is our final model.

Regression 5

```
print("Before Encoding")
print(names(FProjectData))
print(dim(FProjectData))

FProjectData = one_hot_encoding(FProjectData,"Regionname")
FProjectData = as.data.frame(subset(FProjectData, select = -c(Regionname)))

print("After Encoding")
print(names(FProjectData))
print(dim(FProjectData))
```

HouseFit5<lm(Price~log(Rooms)+Distance+Bedroom2+Bathroom+Car+Landsize+Regionname Eastern.Victoria+RegionnameNorthern.Metropolitan+RegionnameNorthern.Victoria+RegionnameSouthEastern.
Metropolitan+RegionnameSouthern.Metropolitan+RegionnameWestern.Metropolitan, data=train))

summary(model5)

Step 6 - In this step, we checked about other regressions such as logistic regression or Arima Model, but we found out that they are not the right choice of regressions testing as per the nature of our dataset as we do not seek any binary(1/0) output in our label 'Price'.

Results

In this section, our team interpreted the findings of our results as per the above-mentioned steps

I. Descriptive Statistics of all variables

The following is our summary of the descriptive statistic table for each of the variables in our dataset. Each statistic number is reasonable except there might be outliers for the land size.

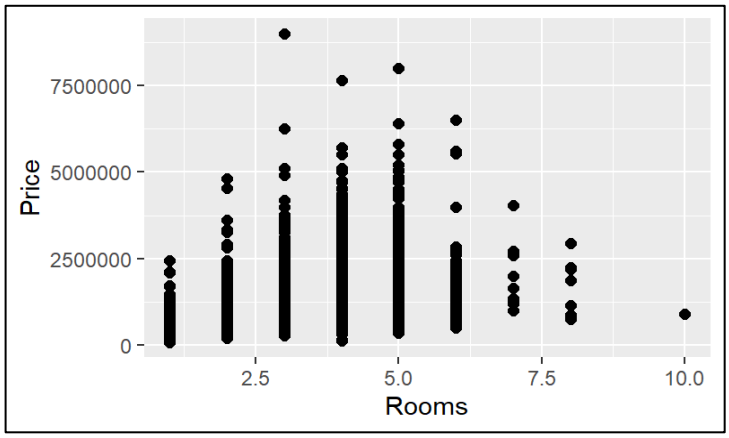
We calculated all Min, Max, Median, first and third quantiles of each variable.

```
   Suburb              Rooms           Price             Date
Length:13518       Min.   : 1.00   Min.   :  85000   Length:13518
Class :character   1st Qu.: 2.00   1st Qu.: 650000   Class :character
Mode  :character   Median : 3.00   Median : 901000   Mode  :character
                   Mean   : 2.94   Mean   :1074796
                   3rd Qu.: 3.00   3rd Qu.:1328000
                   Max.   :10.00   Max.   :9000000
   Distance       Bedroom2        Bathroom         Car           Landsize
Min.   : 0.0   Min.   : 0.00   Min.   :0.00   Min.   : 0.00   Min.   :     0
1st Qu.: 6.2   1st Qu.: 2.00   1st Qu.:1.00   1st Qu.: 1.00   1st Qu.:   178
Median : 9.2   Median : 3.00   Median :1.00   Median : 2.00   Median :   442
Mean   :10.2   Mean   : 2.92   Mean   :1.54   Mean   : 1.61   Mean   :   558
3rd Qu.:13.0   3rd Qu.: 3.00   3rd Qu.:2.00   3rd Qu.: 2.00   3rd Qu.:   651
Max.   :48.1   Max.   :20.00   Max.   :8.00   Max.   :10.00   Max.   :433014
 Regionname        Propertycount
Length:13518       Min.   :  249
Class :character   1st Qu.: 4380
Mode  :character   Median : 6567
                   Mean   : 7456
                   3rd Qu.:10331
                   Max.   :21650
```

II. Scatterplots

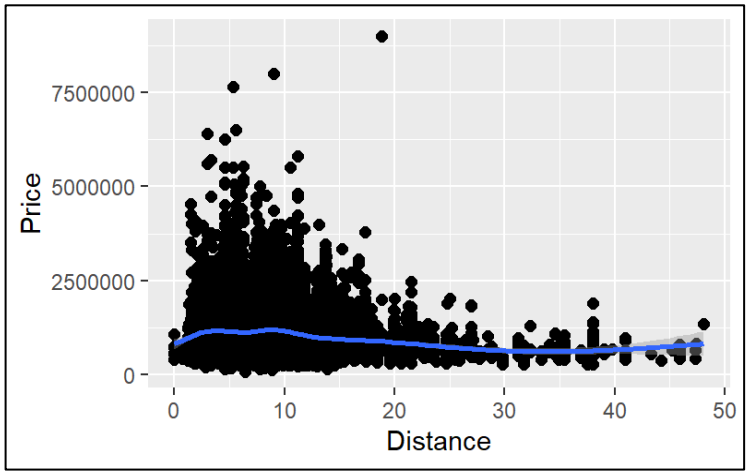
Our team created scatterplots to analyze each variable in detail.

1. **Rooms**



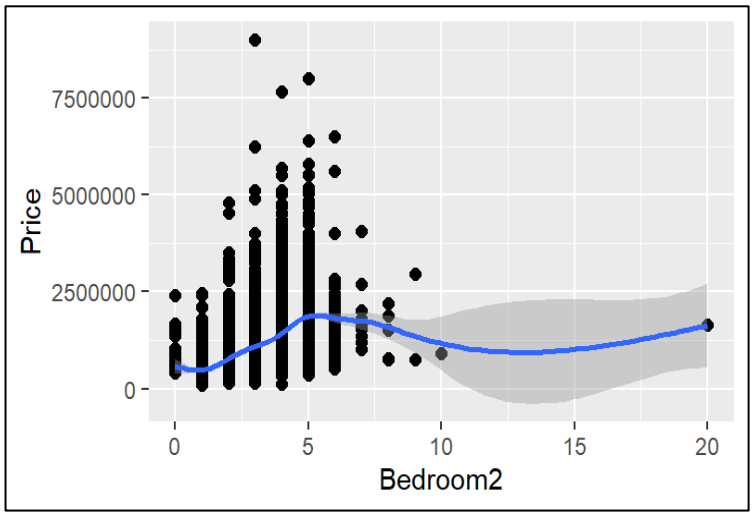
Analysis: We can see a trend from the scatterplot that as the number of rooms increases, the price goes up comparatively. As the number of rooms increases (more than 5), the price relatively drops but the starting point of the price is higher than the price for houses with rooms less than 5. The scatter plot seems to be normally distributed with some outliers. We can conclude that there is a relationship between the price and the number of rooms.

2. Distance



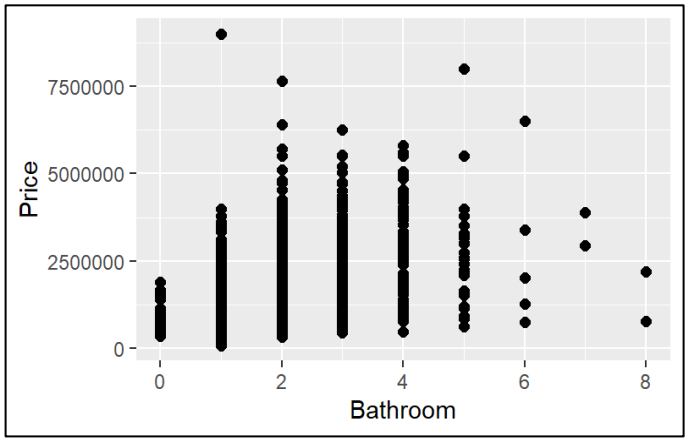
Analysis: We can see the trend that as the distance of the house goes up, the price tends to go down. The price for houses with less than 20 appears to be more expensive. More houses are distributed, and the scatter plot is denser from 0 to 20 distance.

3. Bedroom2



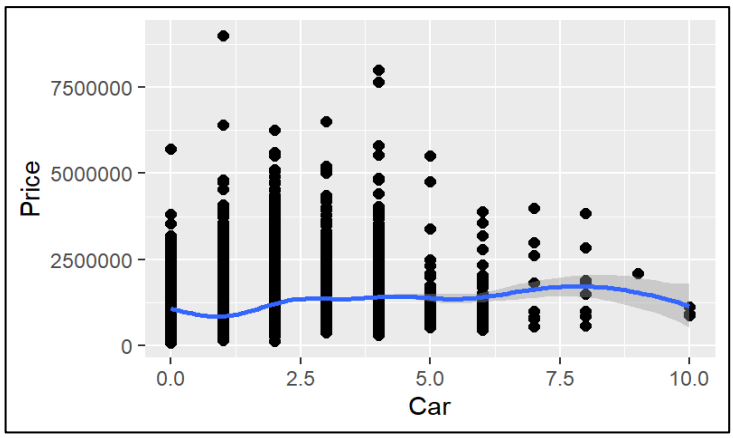
Analysis: From the scatterplot, we can conclude that as the number of bedrooms goes up, the price of the house goes up. For houses with bedrooms greater than 5, the price starts to decrease as the bedrooms increase.

4. Bathroom



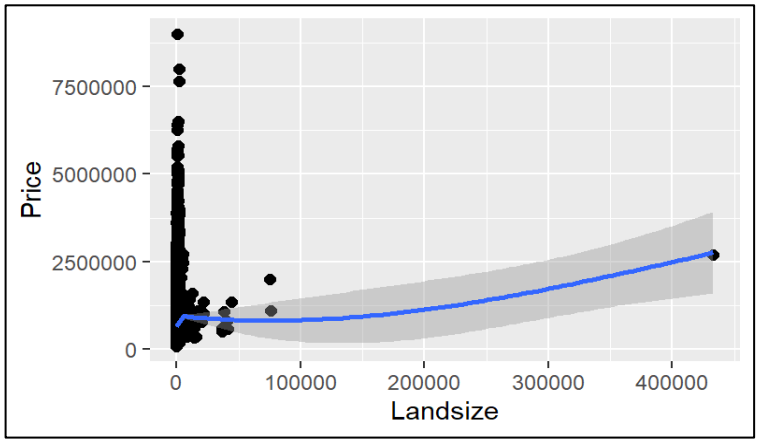
Analysis: We do not see an obvious trend in the scatterplots for Price and Bathroom. When there are more bathrooms, the starting price of the house goes up. The house price is comparatively high when there are 2 to 4 bathrooms. The house price would not be high when there is only 0-1 bathroom. Houses with 5 or more bathrooms are less in number and are also having less price with some exceptions.

5. Car



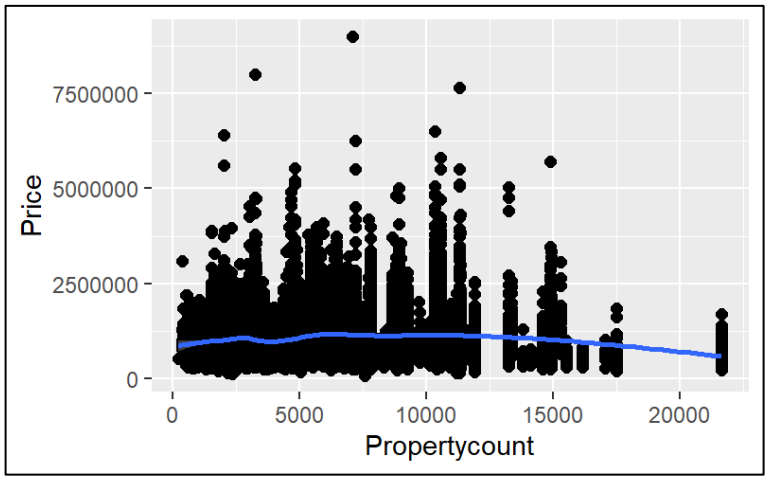
Analysis: We do not see a noticeably clear relationship between Price and the number of car spots. When there are around 2 to 5 car spots, the house tends to have higher prices.

6. Landsize



Analysis: There is one outlier within our data where the house has a super large land size. From the blue line, we can conclude that as the land size increases, the price tends to go up.

7. PropertyCount

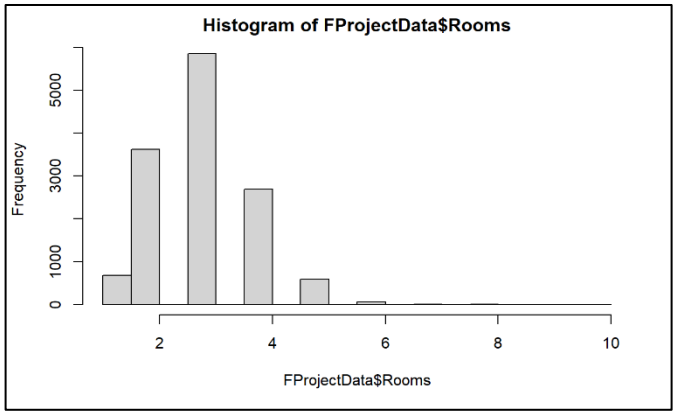


Analysis: We do not see a clear relationship between Price and Property Count.

III. Histograms

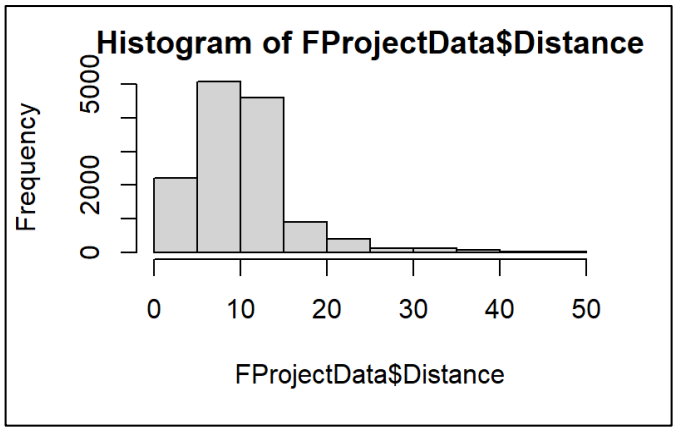
Our team also created histograms to analyze the distribution for each variable and check if there is any skewness problem.

1. Rooms



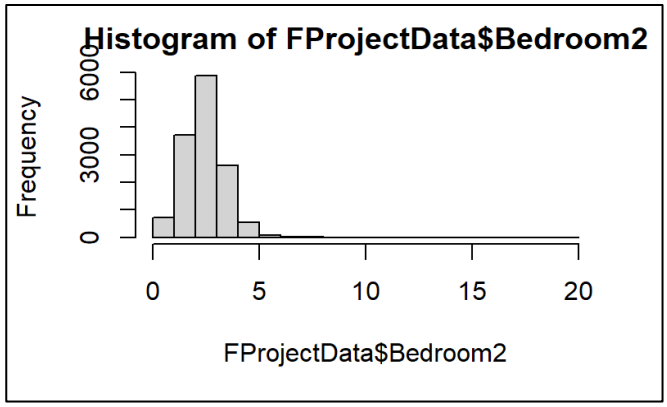
As shown in the histogram, most of the houses have 2-4 rooms in total. The distribution is a right-skewed (positive) distribution.

2. Distance



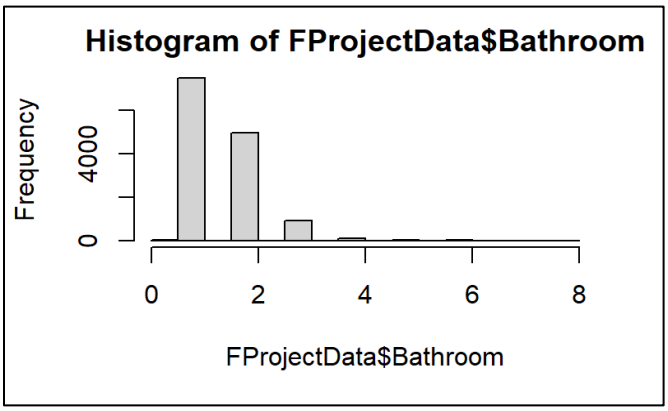
As shown in the histogram, most of the houses are 5 to 15 distances from the center area. The distribution is a right-skewed (positive) distribution.

3. Bedroom 2



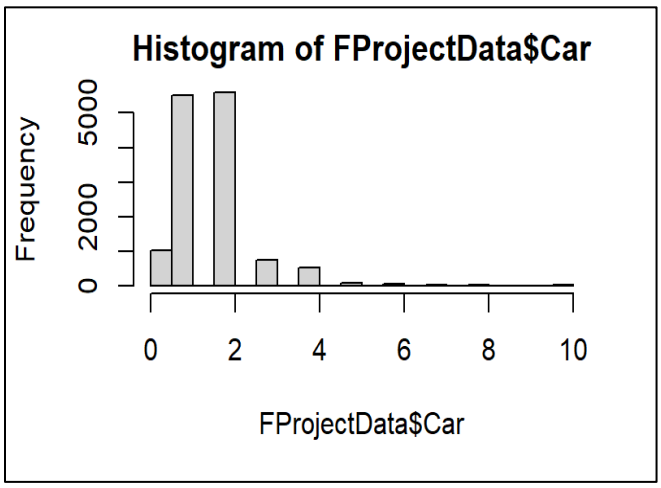
As shown in the histogram, most of the houses have 1-4 bedrooms. The distribution is a right-skewed (positive) distribution.

4. **Bathroom**



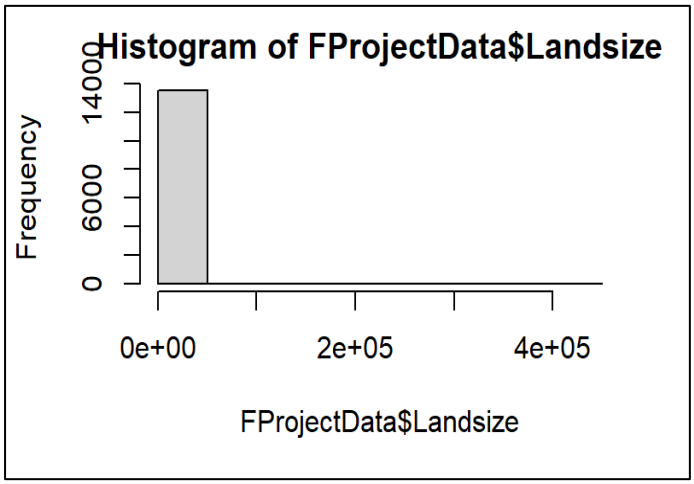
As shown in the histogram, most of the houses have 1-2 bathrooms. The distribution is a right-skewed (positive) distribution.

5. **Car**.



As shown in the histogram, most of the houses have 1 or 2 car spots. The distribution is a right-skewed (positive) distribution

6. Land size

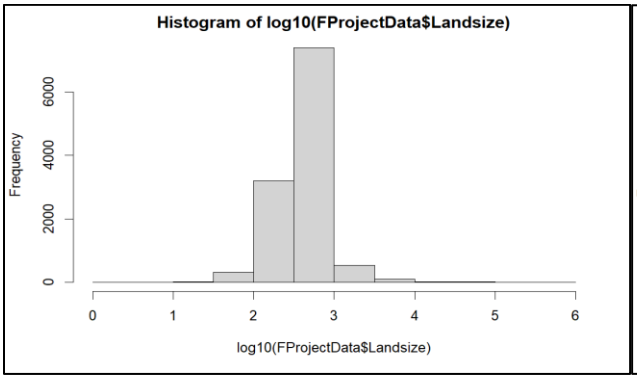


As shown in the histogram, most of the houses in our data have their land sizes within 1000. The reason that the frequency is super right-skewed is there is an outlier in the data. We may consider removing the outlier or transforming "landsize" using an algorithm.

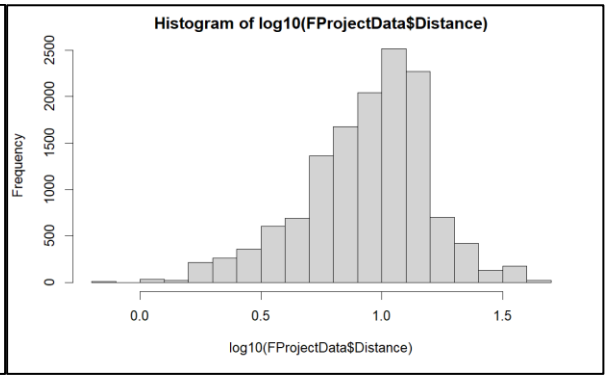
IV. Transformation on Skewed Histograms using Log Function:

To correct the skewness problem, our team applied the Log function on each variable to correct the skewness problem. All the following variables have huge improvements from highly right-skewed to normal distribution.

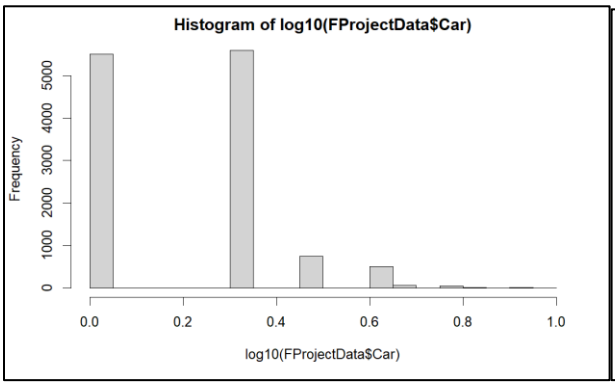
1. Land size



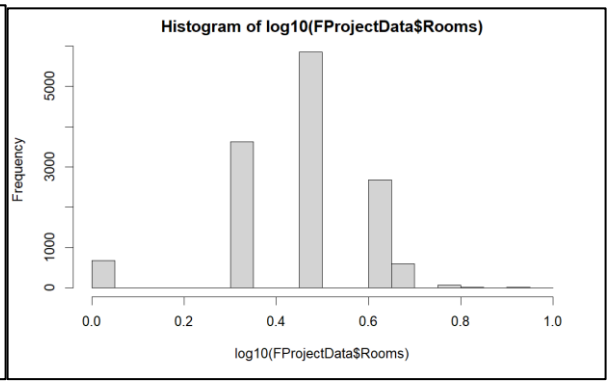
2. Distance



3. Car



4. Rooms

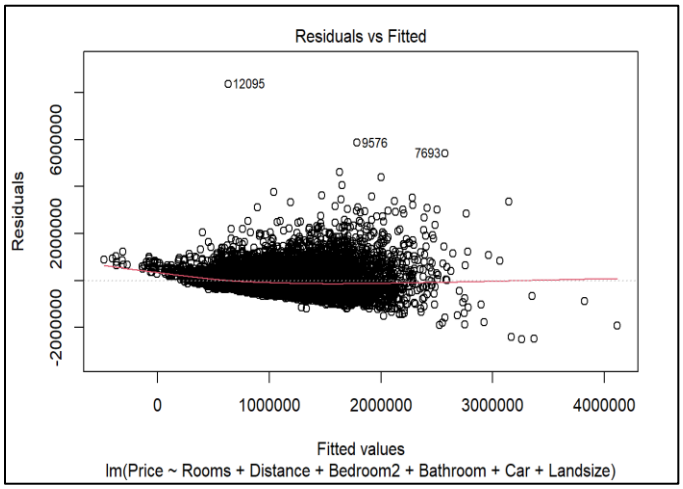


V. Residual Plots

Our team created the residual plots to analyze in detail.

1. **Residuals vs Fitted**

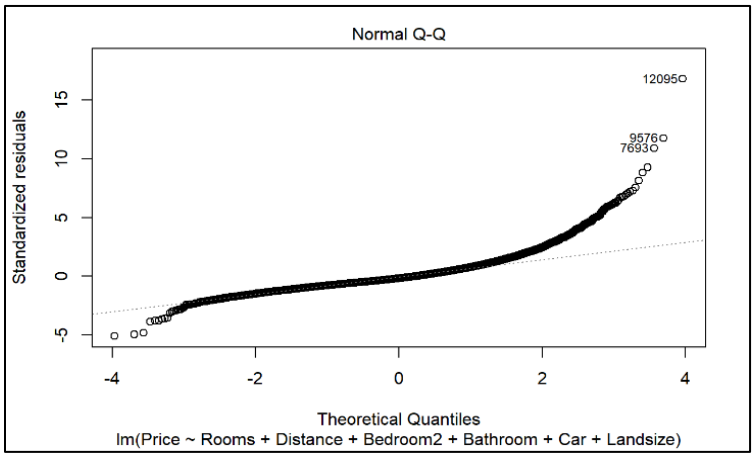
Here we see the linearity relationship. There are several outliers, with residuals close to 300000.



2. **Normal Q-Q**

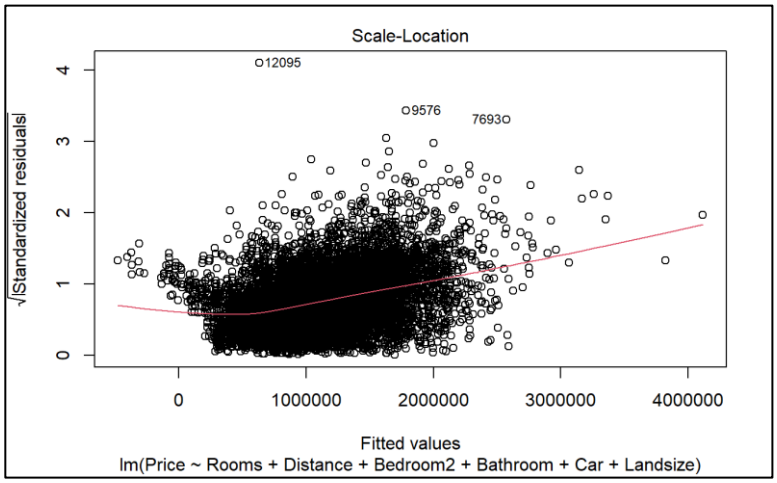
Normal Q-Q residual plot shows to evaluate the model fit and the skew of the data. The theoretical quantiles are shown on the x-axis and the standardized residuals are shown on the y-axis of this graph. Data that aligns closely to the dotted line indicates a normal distribution. If the points are

skewed drastically from the line, you could consider adjusting your model by adding or removing other variables in the regression model.



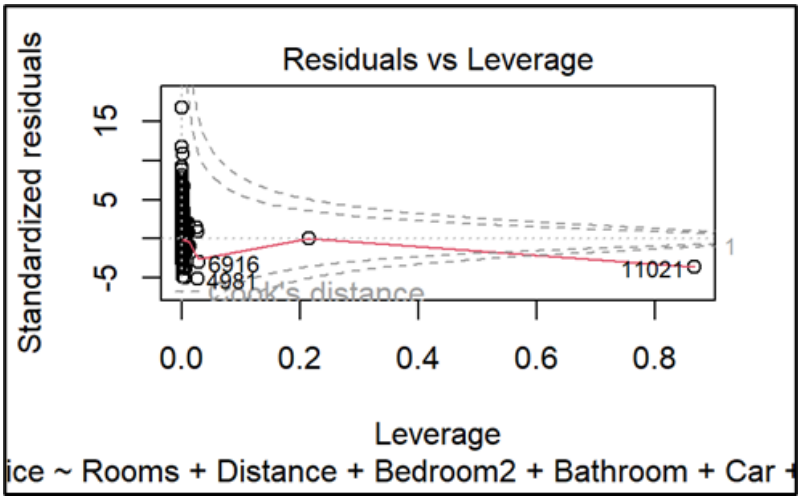
3. Scale-Location

The Scale Location plot suggests some non-linearity here, but what we can also see is that the spread of magnitudes is the lowest in the fitted values close to 0, highest in the fitted values around 1000000, and medium around 3000000. This suggests heteroskedasticity.



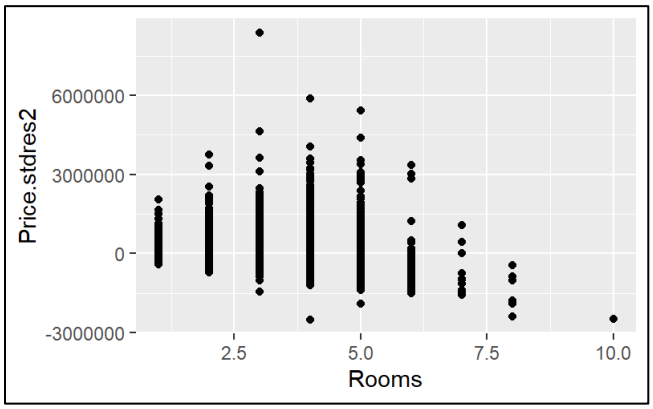
4. Residuals vs Leverage

We can see that observations lie closest to the border of Cook's distance, but it doesn't fall outside of the dashed line. This means there are not any influential points in our regression model.



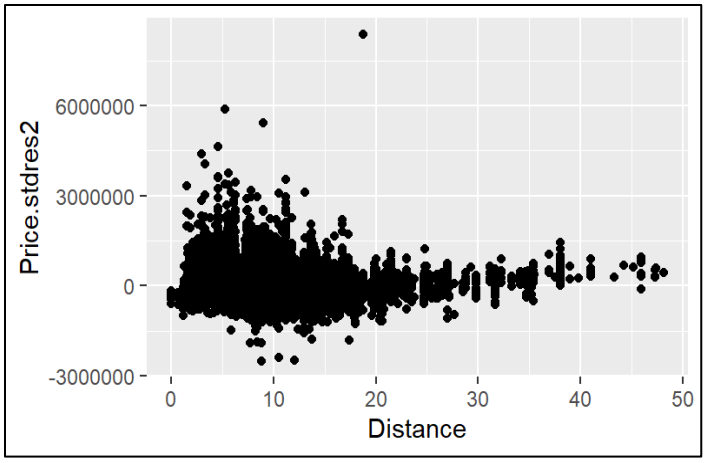
5. Residual Plot -Rooms

Due to a lack of data for rooms with 1, 6, and 6.5, the residuals for such houses are skewed towards one side or the other and all over the place. But for 3 and 5-room houses, it is somewhat normally distributed and centered around 0.



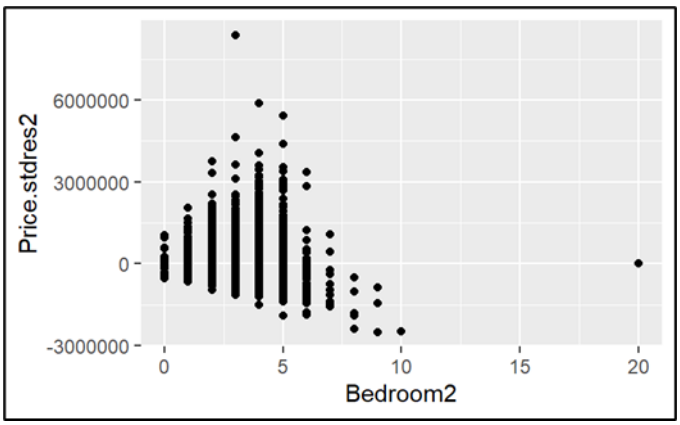
6. Residual Plot - Distance

In this plot, the residuals fluctuate randomly around 0 (at least there is no apparent trend in one direction or the other). However, it might be a problem that so many of our data points are clustered around the center.



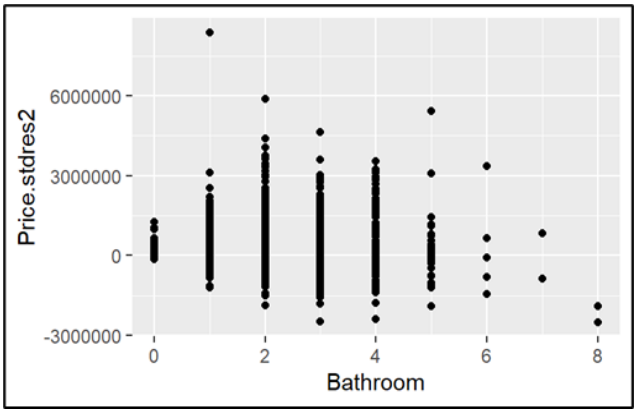
7. Residual Plot - Bedroom2

Like rooms, bedroom 2 is also evenly distributed between positive and negative values for 2 to 5.5. But it is all over the place for other values due to a lack of data.



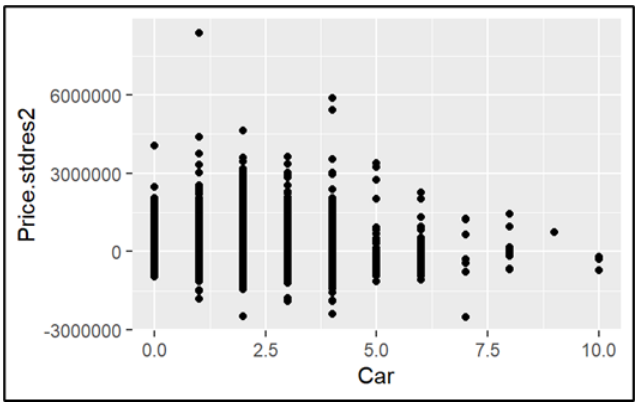
8. Residual Plot - Bathroom

The bathroom is also evenly distributed between positive and negative values for 1 to 4 bathrooms. But it is all over the place for other values due to lack of data.



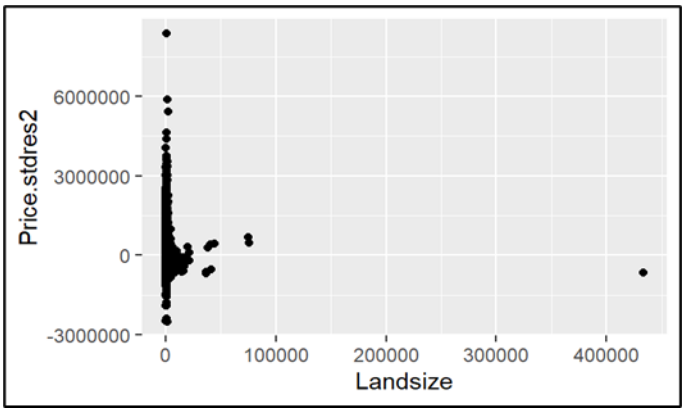
9. Residual Plot - Car

Due to lack of data for car it is all over the place from 3.5 to 10 on the residual plot.



10. Residual Plot - Landsize

This plot doesn't show much due to the presence of few higher landsize. The residuals are all clustered towards lower landsize. Trimming the landsize beyond certain values would have helped get some meaningful plots.



VI. Regression Results

Following are the results of our models and their analysis.

Model 1

```
Call:
lm(formula = Price ~ Rooms + Distance + Bedroom2 + Bathroom +
    Car + Landsize + Propertycount, data = FProject3)

Residuals:
     Min       1Q   Median       3Q      Max
-2485870  -293356   -88115   203870  8369378

Coefficients:
               Estimate Std. Error t value             Pr(>|t|)
(Intercept)   206825.657  16907.306   12.23 < 0.0000000000000002 ***
Rooms         252107.445  13783.877   18.29 < 0.0000000000000002 ***
Distance      -36919.009    779.141  -47.38 < 0.0000000000000002 ***
Bedroom2       35914.390  13531.552    2.65              0.00796 **
Bathroom      211534.965   7789.322   27.16 < 0.0000000000000002 ***
Car            51390.197   4983.634   10.31 < 0.0000000000000002 ***
Landsize           3.892      1.072    3.63              0.00028 ***
Propertycount     -1.689      0.981   -1.72              0.08532 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 498000 on 13510 degrees of freedom
Multiple R-squared:  0.395,     Adjusted R-squared:  0.394
F-statistic: 1.26e+03 on 7 and 13510 DF,  p-value: <0.0000000000000002
```

Analysis: For the first multiple regression model, we included all the numeric variables to check the significance level. The regression result is shown above. Since the p-value for Rooms, Distance, Bedroom, Bathroom, Car, and Land size are much lower than our selected significant level at 0.05, we can reject the null hypotheses and conclude that there are certain relationships between Melbourne house price and these variables. However, the p-value for Property Count is 0.08532 which is higher than 0.05, we can conclude that there is no significant relationship between the number of properties and house prices in Melbourne. The Multiple R-squared is 0.395 which means that only 39.5% of the data can be explained by our model. Our first model does not perform well in predicting house prices.

Model 2

```
Call:
lm(formula = Price ~ Rooms + Distance + Bedroom2 + Bathroom +
    Car + Landsize, data = FProject3)

Residuals:
     Min       1Q   Median       3Q      Max
-2499634  -293289   -87907   203443  8369468

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 192153.00   14600.71   13.16 < 0.0000000000000002 ***
Rooms       252413.06   13783.73   18.31 < 0.0000000000000002 ***
Distance    -36871.65     778.71  -47.35 < 0.0000000000000002 ***
Bedroom2     36183.66   13531.63    2.67              0.00750 **
Bathroom    211630.78    7789.69   27.17 < 0.0000000000000002 ***
Car          51244.43    4983.28   10.28 < 0.0000000000000002 ***
Landsize         3.90       1.07    3.64              0.00028 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 498000 on 13511 degrees of freedom
Multiple R-squared:  0.395,     Adjusted R-squared:  0.394
F-statistic: 1.47e+03 on 6 and 13511 DF,  p-value: <0.0000000000000002
```

Analysis: For the second multiple regression model, we decided to drop the ‘Propertycount’ since it is not significant. The above multiple regression results indicate that all our variables are very significant which is better than our first model. However, the Multiple R-squared stays the same (R2 value 0.394). Therefore, we need to add the categorical variables in our model to see if it helps to explain the data better.

Model 3

HouseFit3<-lm(Price~Rooms+Distance+Bedroom2+Bathroom+Car+Landsize+Regionname, data=FProject3)

```
                                          Pr(>|t|)
(Intercept)                        < 0.0000000000000002 ***
RegionnameEastern Victoria                0.0000000013 ***
RegionnameNorthern Metropolitan    < 0.0000000000000002 ***
RegionnameNorthern Victoria                  0.00249 **
RegionnameSouth-Eastern Metropolitan < 0.0000000000000002 ***
RegionnameSouthern Metropolitan    < 0.0000000000000002 ***
RegionnameWestern Metropolitan     < 0.0000000000000002 ***
RegionnameWestern Victoria                   0.82510
Rooms                              < 0.0000000000000002 ***
Distance                           < 0.0000000000000002 ***
Bedroom2                                     0.00021 ***
Bathroom                           < 0.0000000000000002 ***
Car                                < 0.0000000000000002 ***
Landsize                                  0.0000533097 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 452000 on 13504 degrees of freedom
Multiple R-squared:  0.502,     Adjusted R-squared:  0.502
F-statistic: 1.05e+03 on 13 and 13504 DF,  p-value: <0.0000000000000002
```

Analysis: For the third model, we include the categorical variable "Regionname" to see if it explains the data better. The p-values are all very significant except for Western Victoria. There is a significant improvement in the **R-squared from 0.395 to 0.502.**

Model 4

```
Call:
lm(formula = Price ~ Regionname + log(Rooms) + Distance + Bedroom2 +
    Bathroom + Car + Landsize, data = FProject3)

Residuals:
     Min       1Q   Median       3Q      Max
-2161367  -267647   -55174   192451  8101859

Coefficients:
                                  Estimate  Std. Error t value
(Intercept)                     325924.652   20150.414   16.17
RegionnameEastern Victoria      413904.517   65301.524    6.34
RegionnameNorthern Metropolitan -212285.458   14695.309  -14.45
RegionnameNorthern Victoria     254608.252   74075.153    3.44
RegionnameSouth-Eastern Metropolitan 232117.487   25714.518    9.03
RegionnameSouthern Metropolitan 223679.080   14163.292   15.79
RegionnameWestern Metropolitan -279232.331   14741.358  -18.94
RegionnameWestern Victoria       34455.641   81482.364    0.42
log(Rooms)                      716846.990   26758.784   26.79
Distance                        -44106.017     913.039  -48.31
Bedroom2                         51913.645   10299.070    5.04
Bathroom                        168056.799    7034.065   23.89
Car                              55082.344    4498.387   12.24
Landsize                             3.865       0.966    4.00
                                    Pr(>|t|)
(Intercept)                     < 0.0000000000000002 ***
RegionnameEastern Victoria               0.00000000024 ***
RegionnameNorthern Metropolitan < 0.0000000000000002 ***
RegionnameNorthern Victoria              0.00059 ***
RegionnameSouth-Eastern Metropolitan < 0.0000000000000002 ***
RegionnameSouthern Metropolitan < 0.0000000000000002 ***
RegionnameWestern Metropolitan  < 0.0000000000000002 ***
RegionnameWestern Victoria               0.67240
log(Rooms)                      < 0.0000000000000002 ***
Distance                        < 0.0000000000000002 ***
Bedroom2                                 0.00000047005 ***
Bathroom                        < 0.0000000000000002 ***
Car                             < 0.0000000000000002 ***
Landsize                                 0.00006341769 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 448000 on 13504 degrees of freedom
Multiple R-squared:  0.51,      Adjusted R-squared:  0.51
F-statistic: 1.08e+03 on 13 and 13504 DF,  p-value: <0.0000000000000002
```

Analysis: Our next step model was trying to use the transformed logarithm variables to see if the model is better when the skewness problems are solved. However, we could not transform the "Bedroom", "Bathroom", and "Car" because there are 0 values for the three variables in the dataset. We ended up only transforming the "Rooms" to "log (Rooms). As shown above, we can see that there is a slight improvement in the R-squared from 0.502 to 0.51.

Model 5 – Final Model

HouseFit5<lm(Price~log(Rooms)+Distance+Bedroom2+Bathroom+Car+Landsize+Regionname Eastern.Victoria+RegionnameNorthern.Metropolitan+RegionnameNorthern.Victoria+RegionnameSouthEastern.
Metropolitan+RegionnameSouthern.Metropolitan+RegionnameWestern.Metropolitan, data=train))

```
Coefficients:
                                  Estimate Std. Error t value            Pr(>|t|)
(Intercept)                       332492.0    24771.2   13.42 < 0.0000000000000002 ***
log(Rooms)                        724388.0    31470.9   23.02 < 0.0000000000000002 ***
Distance                          -44994.6     1112.4  -40.45 < 0.0000000000000002 ***
Bedroom2                           50427.2    11971.1    4.21        0.0000255127 ***
Bathroom                          160698.7     8610.7   18.66 < 0.0000000000000002 ***
Car                                57965.3     5532.0   10.48 < 0.0000000000000002 ***
Landsize                              11.7        3.7    3.18               0.0015 **
RegionnameEastern.Victoria        455475.1    77719.2    5.86         0.0000000048 ***
RegionnameNorthern.Metropolitan  -207676.8    18033.3  -11.52 < 0.0000000000000002 ***
RegionnameNorthern.Victoria       295927.5   101465.7    2.92               0.0035 **
RegionnameSouth.Eastern.Metropolitan  265439.3    31549.3    8.41 < 0.0000000000000002 ***
RegionnameSouthern.Metropolitan   222945.7    17351.5   12.85 < 0.0000000000000002 ***
RegionnameWestern.Metropolitan   -284281.0    18050.5  -15.75 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 449000 on 8999 degrees of freedom
Multiple R-squared:  0.507,     Adjusted R-squared:  0.506
F-statistic:  771 on 12 and 8999 DF,  p-value: <0.0000000000000002
```

Analysis: In our final model, we can observe that the coefficients are positive for Rooms, Bedroom2, Bathroom, Car, Landsize, RegionnameEastern.victoria, RegionnameNorthern.Victoria, RegionnameSouth.Eastern. Metropolitan, RegionnameSouthern.Metropolitan and these independent variables have a directly proportional relationship with price. Distance, RegionnameNorthern.Metropolitan, RegionnameWestern.Metropolitan, has a negative coefficient, and it is inversely proportional to price.

Parsimonious Model:

In this 5th model, we removed Regionname Western Victoria since it is not significant. Although the adjusted R-squared dropped from 0.51 to 0.506, we still would like to conclude that this is our parsimonious model because all the variables are significant.

Prediction based on the final model 5:

```
> price.predict = predict(model5,test)
> str(price.predict)
 Named num [1:4506] 1445864 778115 281420 994653 836244 ...
 - attr(*, "names")= chr [1:4506] "5" "6" "9" "11" ...
```

Explanation:

Price = 332,492 + 724,338*log (Rooms) - 44,944.6*Distance + 50,427.2*Bedroom2 + 160,698.7*Bathroom + 57,965.3*Car + 11.7*Landsize + 455,475.1*Eastern.Victoria - 207,676.8*Northern.Metropolitan + 295,927.5*Northern.Victoria + 265,439.3* Esatern.Metropolitan + 222,945.7* Southern.Metropolitan - 284,281* Western.Metropolitan

*Note: The input would be 1 for the region name of the house, and 0 for all other region names.

We conducted the prediction on our current data set to compare the prediction values and actual values. There is 50.6% of the predicted prices match the actual prices.

For future usage, we suggest creating an excel document and inputting the numbers for each of the variables of the houses in Melbourne, then using our model to predict the prices.

Limitations of Study

After going through the study, the best-performing model had 12 variables with an R2 of 0.50 which is not so great. This study focused only on selecting the appropriate variables and building a parsimonious model using only multiple linear regression only. This could be improved in either

1. By building other models like a tree-based or a neural network.

2. By collecting more data relevant to recent purchases and so on.

Housing prices also depend on the market and the region, and the other possible solution is to use the most recent purchases in the local neighborhood and use to build the model. We feel that will be more appropriate for this problem.

Conclusion

All in all, we can see that prices are affected by factors such as Rooms, Bathrooms, Bedrooms, Distance, Car, and Landsize with results of p-value less than the significant level of 0.05. On the other hand, Property Count had a p-value higher than the significant level of 0.05, so we can conclude that it is not a factor affecting price. The best performing model (Model 5) for predicting house price included variables with Rooms, Bedroom2, Bathroom, Car, Landsize, RegionnameEastern.victoria, RegionnameNorthern.Victoria, RegionnameSouth.Eastern. Metropolitan, RegionnameSouthern.Metropolitan Although this is the best model in predicting Melbourne housing in our study, we suggest that picking out a specific zip code will have a better result in predicting housing in Melbourne.

Moreover, we were able to build a prediction calculator to see the predictions on prices using our significant variables.

References

What affects the market value of a property? loans.com.au. (2022, October 24). Retrieved December 9, 2022, from https://www.loans.com.au/home-loans/property-reports/what-factors-affect-property-value