Curtin University

Business Quantitative Analysis

Unit Code: ECOM5002

GROUP 12

SYED MUHAMMAD AHMED ZAIDI - 20972008

MOHAMMAD ADIL JAN MALIK – 21320394

UMAIR SATTAR - 21361924

Table of Contents

INTRODUCTION

Quantitative analysis in business is a structured way of learning from and acting upon numerical information. Mathematical and statistical methods are applied to business data for analysis and interpretation. In this assignment, we took a deep dive into the housing market for Perth particularly focusing on three Suburbs which include Inglewood, Dunsborough and Bull Creek. The report thoroughly analyses and discusses their most latest sold house listings to find insights that would help develop a prediction model allowing us to forecast the Prices of currently advertised housing for sale. This model will particularly be helpful for objectively evaluating performance metrics, forming well-informed judgments, and discovering patterns and trends that can influence strategic planning, especially for homebuyers and sellers. This would pave the way towards investment planning as Real Estate investors can use the model to diversify their portfolio and also take help for risk management. Forecasted house pricing could also turn out to be a good economic indicator providing insights about the health of this industry ultimately aiding strategic management for both consumers as well as policy makers. For the report, we have the following methodology and structure.

1.1 Gathering Information

First and foremost was the gathering of information about the three assigned suburbs which was collected through RealEstate.com.au. By applying multiple filters of the most latest housing being sold, we extracted a total of 150 house listings that provided all the information that would be required to reduce as much missing data as possible. Each Suburb was allocated 50 listings that showed the latest houses recently sold in the market. The extracted information about each house includes its Price, Land Size (SQM), Number of Bedrooms, Number of Bathroom, Car Space, Distance to Primary School (KM), Distance to Secondary School (KM), Age of House and Address. All this information can be found in the datafile.csv attached with the report.

1.2 Preparing and Cleaning the Data:

As preparing and cleaning data is one of the most important steps for data analysis it requires fixing missing data, eliminating anomalies, and adjusting unit sizes (El-Nasr et al., 2021). For our report, we from the start only picked those housing properties that would provide us the complete information so that issues with missing data could be reduced later on. Properties that were too far away from the suburbs were avoided as to improve the prediction power of the model highly saturated areas within the suburb.

1.3 Statistical Analysis:

The properties of the data set are summarised and described using descriptive statistics such as mean, median, mode, variance, standard deviation, and more. It also consists of Inferential statistics which is a robust analytical technique utilized in research and decision-making across diverse disciplines including the Real Estate Market. Using samples enables us to draw informed inferences and forecasts on populations, even when confronted with insufficient data (Peren, 2021). As the whole population would be unrealistic to consider for this assignment, we have taken a sample of 50 properties from each allocated suburb which is used as the train data to create the prediction model.

1.4 Data Visualization:

Throughout the assignment we have made use of multiple visual aids such as charts and graphs to communicate findings and patterns better. This includes Bar Graphs, scatterplots, boxplots, and correlation plots to find interesting insights based on which we could apply our regression model which is a process that encompasses a range of techniques for modelling, predicting, and extracting significant insights from temporal data sets (Soltane et al., 2022).

1.5 Model Creation:

Once the data was explored, we dived into the creation of a prediction model based on the regression analysis. We made use of all the dimensions extracted for the suburbs to try building a predictor that could help us forecast the price of any house currently advertised in the market. We made use of functions like Correlation Coefficients, confidence intervals, and also ANOVA to find the strengths of relationships between different inputs and choose the ones that are impacting the model at most. This reduced regression model is then created that would serve the purpose of prediction which is immensely important in several domains, facilitating data-driven insights and forecasts to enhance decision-making (Pandey, 2020).

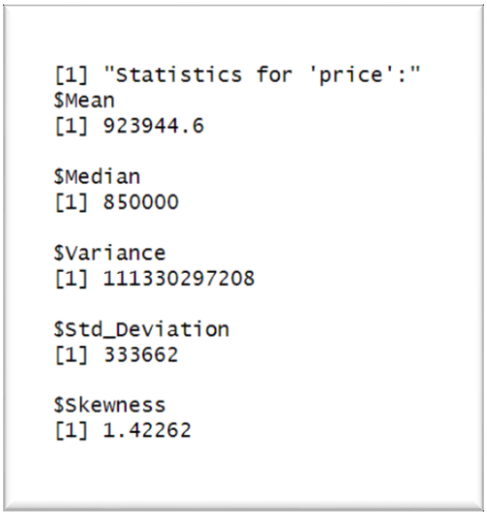
1.6 Forecasting:

Forecasting involves strategic predictions or estimations of future events, results, or trends. It informs planning and decision-making using historical data, patterns, and pertinent information. Once we had the model fully created, we went forward by selecting a currently advertised property and tried predicting the price of it. This was then compared with the authentic ANZ report for that particular house to make comparisons on how well the model has worked. However, projections are unpredictable and liable to change with new information and unforeseen events, making ongoing monitoring and adjustment necessary (Young et al., 2004). Hence, we found reasons for the results not being completely accurate and how we can make use of this analysis to improve our prediction model further.

ANALYSIS

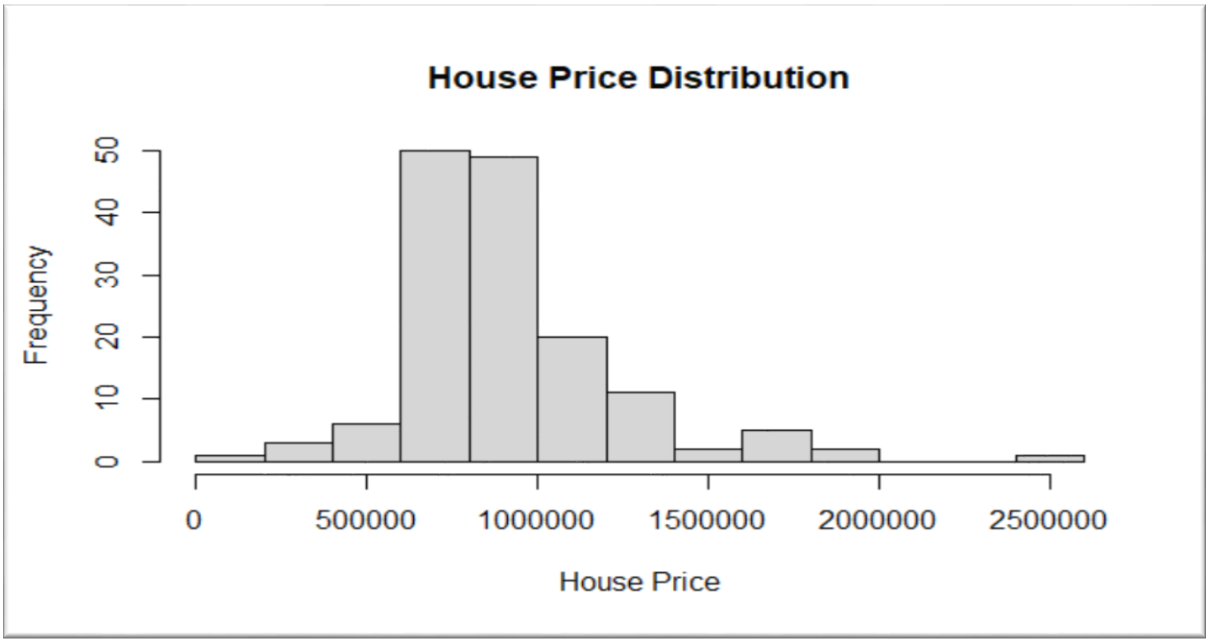
2.1 Summary Statistics And Central Tendency

The concept of central tendency plays a crucial role in statistical analysis. For this purpose, we have made use of multiple statistical concepts to deeply evaluate the current situation of the housing market in Perth. Within the code, we were able to generate statistical results for all the columns however our predictor which is price has the following results.



```
[1] "Statistics for 'price':"
$Mean
[1] 923944.6

$Median
[1] 850000

$Variance
[1] 111330297208

$Std_Deviation
[1] 333662

$Skewness
[1] 1.42262
```

The mean shows that the average house prices for the three allocated suburbs. This comes out to be $ 923,944.6 which is calculated by adding up all the prices of 150 properties and then dividing them against the count. The mean is very near to the price of an average house being sold in the market in all three suburbs reflecting the credibility of data collection. However, within the data their were some outliers as some of the house prices did reach to around $2500000. outliers can affect the mean which is why we can use median (Bird, 2018). This method is when the data set is sorted by ascending order, this is the value in the middle. This measure of central tendency suggests that 50% of the houses in Perth suburbs are priced below $850,000, and 50% are priced above. Moving forward, the variance indicate how house prices may vary from one to the other. As each of them had different rooms, bathrooms, garages etc. they vary significantly as the results suggest that $111,330,297,208 spread or variability in house prices in allocated suburbs is a lot. The standard diviation uses this variance to see how the prices of the houses are spread out from one another and according to the results, $333,662 is the average amount by which individual house prices deviate from the mean. Lastly, the results for skewness of 1.42262 suggests that the distribution of house prices is right-skewed. This reflects our previous discussion of outliers as there are some properties that are highly values which is why they are stretching the mean towards the higher side.



2.2 Histograms and Distribution Shape

Figure 1 House Price Distribution

The analysis of the dispersion of residential property prices within a specific geographical region. The horizontal axis represents the price of houses, spanning from $0 to $2,500,000, while the vertical axis represents the frequency, denoting the quantity of dwellings within each price range.

The histogram illustrates a right-skewed distribution of house prices, indicating a greater concentration of residences in the lower price categories compared to the higher price ranges. The approximate value of the median house price, which is the point at which 50% of houses are priced higher and 50% are priced lower, is $163,000.

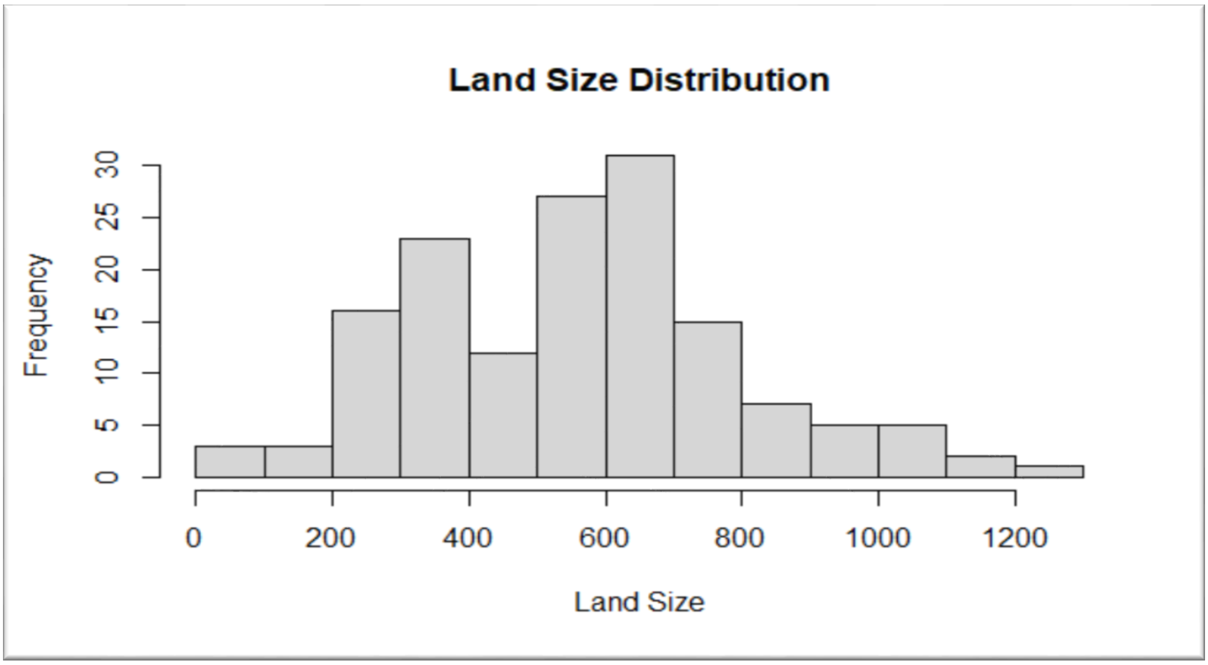


Figure 2 Land Size Distribution

There exist several potential rationales for the configuration of this distribution. One potential explanation pertains to a restricted inventory of luxury residences within the vicinity, resulting in housing costs escalating. An alternative hypothesis could be that a significant need exists for cost-effective housing within the locality.

The histogram is a valuable tool for addressing many inquiries about the local property market. As an illustration, the prevailing range for housing prices is $129,975 to $214,000. There exists a need for more residential properties valued at over $2 million.

Plot size is represented on the x-axis, with values ranging from 0 to 1200 square meters, and frequency is displayed along the y-axis.

According to the histogram, there is an even distribution of land sizes. This means there are around the same number of small, medium, and large plots of land. Approximately 600 square meters is the median land size, at which half of the land plots are larger, and half are smaller.

The distribution's form can be attributed to several factors. One possible explanation is that the area's land is very level and homogenous, making it ideal for various purposes. The site may also be zoned for many uses, including residential, commercial, and industrial development.

The histogram can be used to investigate various issues concerning the local real estate market. For instance, we may see that the typical lot size is between 450 and 750 square meters. It is also clear that lots larger than 1000 square meters are quite rare.

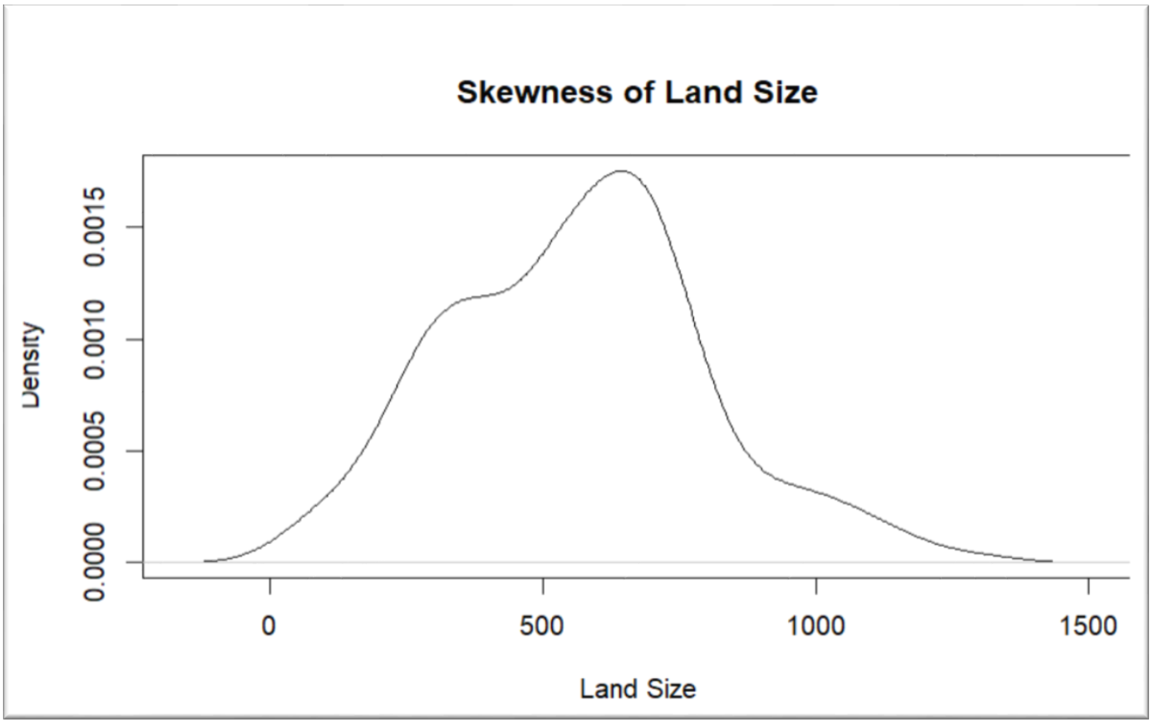


Figure 3 Skewness of Land Size

By looking at the skewness graph of Land Size it can depicted that it negatively skewed as majority of the points are on the left side of the graph. Negatively skewed generally means that the majority of the points are concentrated towards the right side of mean and left tail is stretched out. At this condition normally mean is less than median.

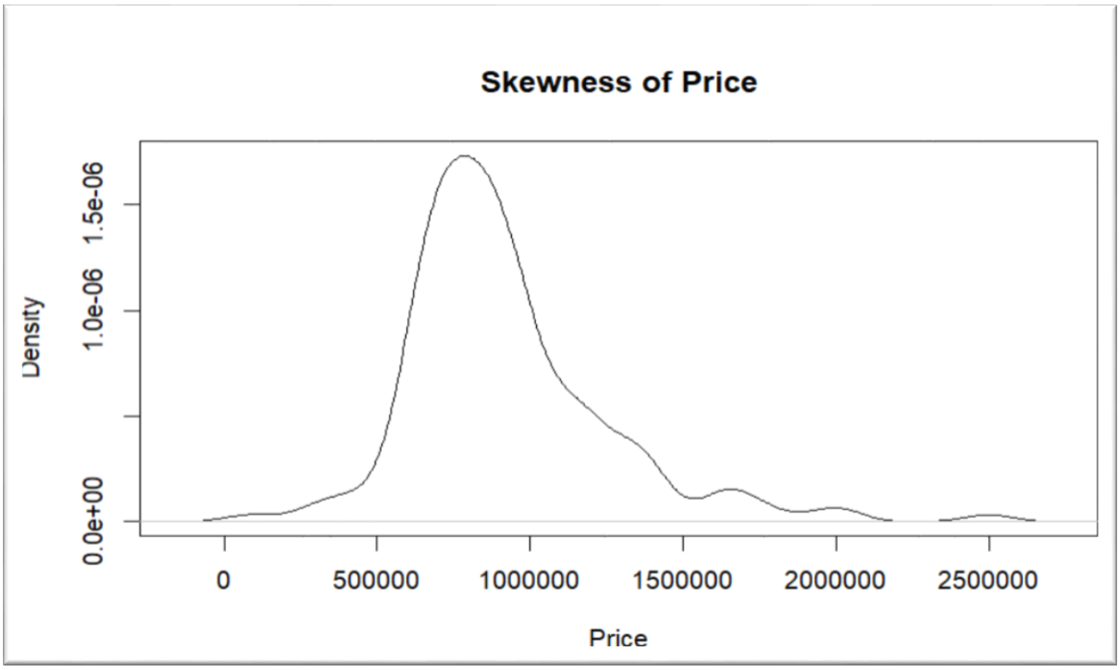
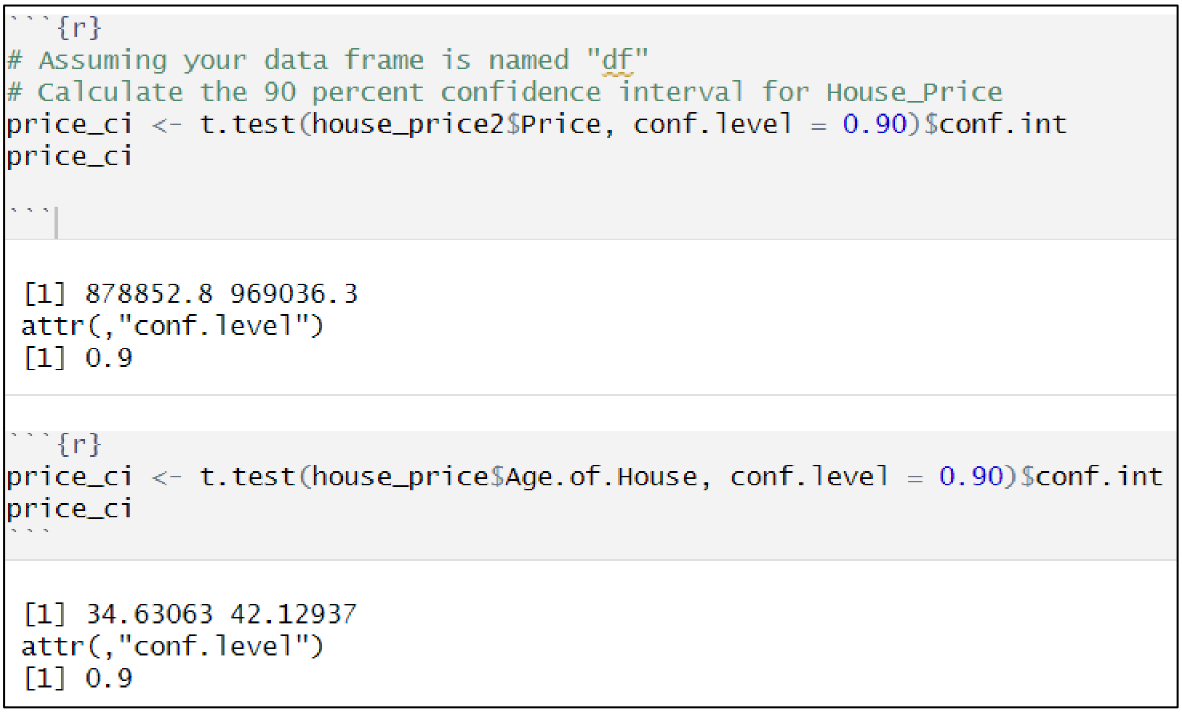


Figure 4 Skewness of Price

By looking at the skewness graph of Price it similar to Land Size so it can depicted that it is negatively skewed as majority of the points are on the left side of the graph. Negatively skewed generally means that the majority of the points are concentrated towards the right side of mean and left tail is stretched out. At this condition normally mean is less than median.

2.3 Confidence Intervals

Confidence Interval gives the range of possible values with certainty within which it can estimate the statistical measure of the data. Commonly used percentages for confidence intervals are 90%, 95%, and 99%, which shows that the narrower the margins, the higher accuracy can be obtained for the interval (Simundic, 2008). At this point of our analysis, we have just found the confidence interval for the Age of the house and the Price of the house. The following results are obtained:



~~~
```{r}
# Assuming your data frame is named "df"
# Calculate the 90 percent confidence interval for House_Price
price_ci <- t.test(house_price2$Price, conf.level = 0.90)$conf.int
price_ci

```
~~~

```
 [1] 878852.8 969036.3
 attr(,"conf.level")
 [1] 0.9
```

~~~
```{r}
price_ci <- t.test(house_price$Age.of.House, conf.level = 0.90)$conf.int
price_ci
```
~~~

```
 [1] 34.63063 42.12937
 attr(,"conf.level")
 [1] 0.9
```

Figure 5 CI's for House Price and Age

For Price, we are 90% confident that the values lie between 878,852.8 and 969,036.3, and for the Age of the House, the value lies between 34.63063 and 42.12937. Further, other variables are also considered to build the analysis of the model at different confidence intervals.

2.4 Scatterplots and Relationships

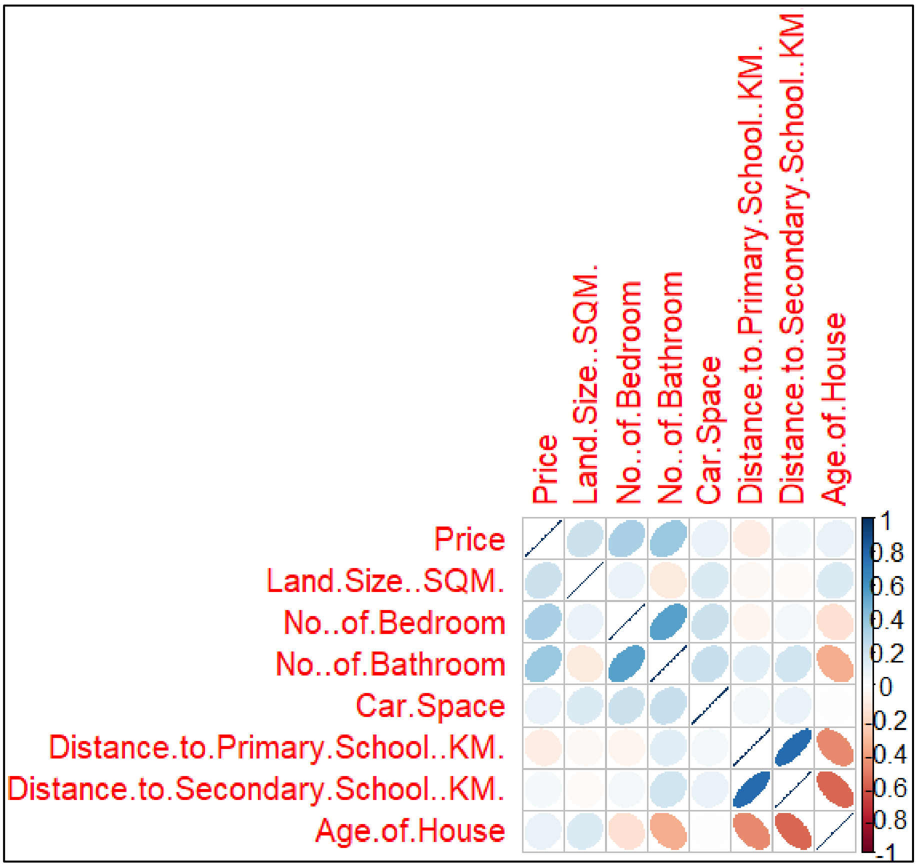


Figure 6 Correlation Matrix

This matrix indicates and explains the relationship of variables with each other. We can conclude from the figure that Distance from Primary and Secondary School are strongly correlated. Also, No. of bathroom and No. of Bathroom are also about 0.6 correlated which shows a good correlation. On the contrary, Age and Distance to Secondary School along with Primary are highly negatively correlated and showing opposite characteristics. However, other variables show little no correlation between the variables. Highly correlation means that both the variables show the same characteristics and behave similarly whereas, for negatively correlation it is vice versa.

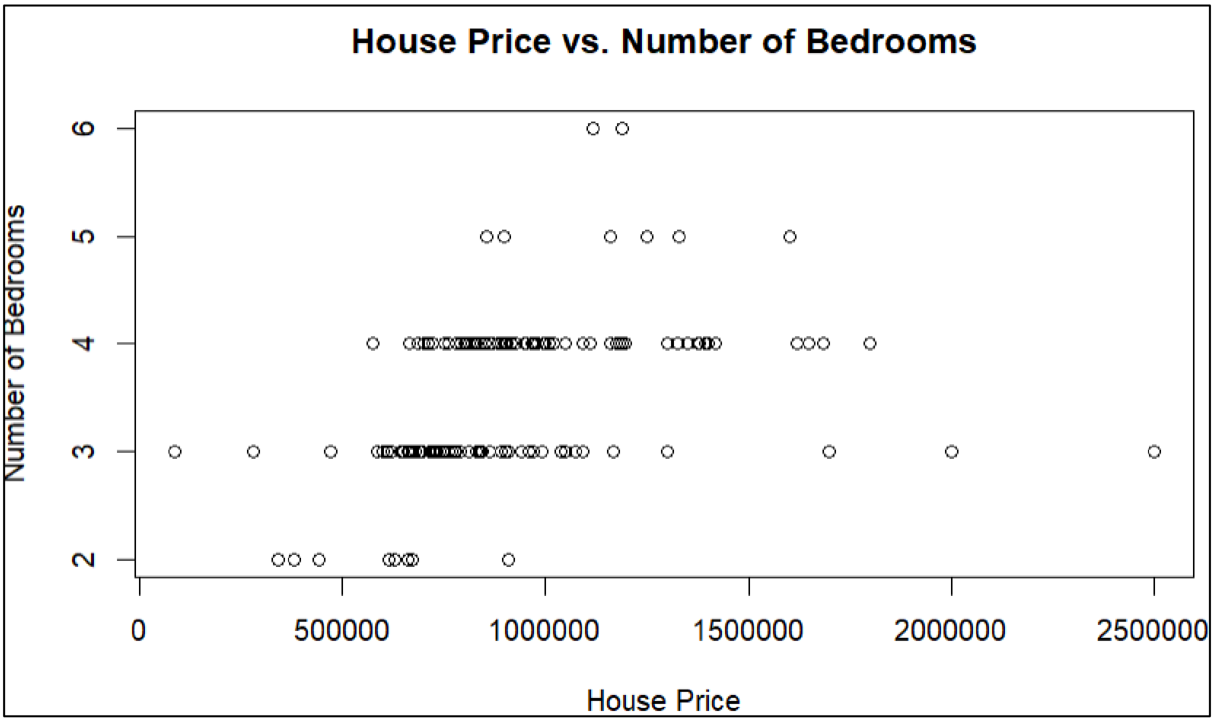


Figure 7 House Price Vs No. of Bedrooms

The scatter plot above illustrates the correlation between home prices and bedroom counts in a given area. The x-axis depicts the price of the home from zero to two million five hundred thousand dollars, and the y-axis displays the number of bedrooms from zero to six.

The scatter plot indicates a positive relationship between home price and bedroom count, with larger homes commanding higher prices on average. This is probably because purchasers value space and versatility, which are increased in homes with additional bedrooms. Houses with the same number of bedrooms exhibit much price fluctuation in the scatter plot. This is probably due to a combination of variables, including the house's location, condition, and proximity to local conveniences.

The scatter plot above depicts the correlation between house price and age in a certain location. House prices (shown along the x-axis) range from zero to two million and fifty thousand dollars, while house ages (shown along the y-axis) go from zero to one hundred.

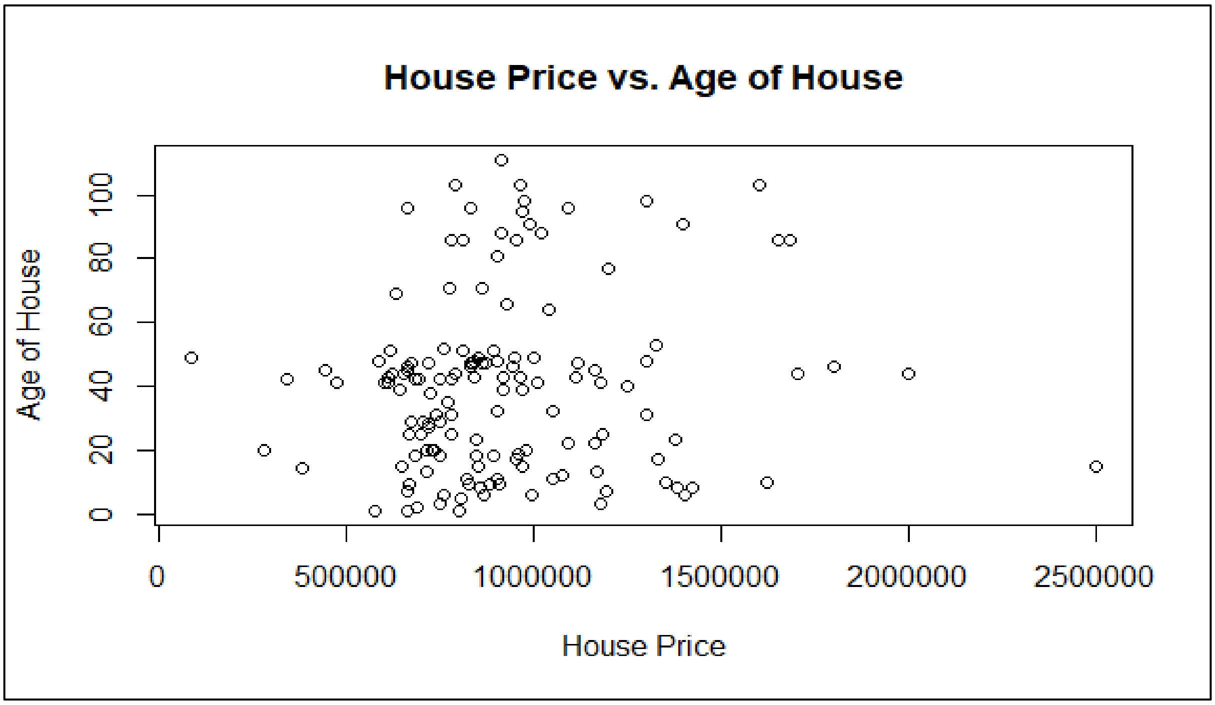


Figure 8 Scatterplot Price Vs Age

The scatter plot demonstrates a negative relationship between home price and residence age, suggesting that older residences are more affordable. This is probably because older homes have a higher potential for breakdown and repair costs and lack some of the conveniences seen in contemporary constructions. However, there is a wide range in home prices, even among homes of the same age. Several factors, including the house's location, condition, and proximity to conveniences, are likely responsible for this.

The scatter plot between the home price and the neighbourhood property age is informative. Keep in mind that the scatter plot only reveals correlation and not causation. Both home price and age could be influenced by external forces, such as the property's location or the quality of the local school district.

However, there is a wide range in home prices, even among homes of the same age. Several factors, including the house's location, condition, and proximity to conveniences, are likely responsible for this. The scatter plot between the home price and the neighbourhood property age is informative. Keep in mind that the scatter plot only reveals correlation and not causation. Both home price and age could be influenced by external forces, such as the property's location or the quality of the local school district.

The scatter plot above illustrates how land area and home price relate to a region. The x-axis depicts a land area from zero to 1200 square meters, while the y-axis depicts the price of a home from zero to two hundred thousand dollars.

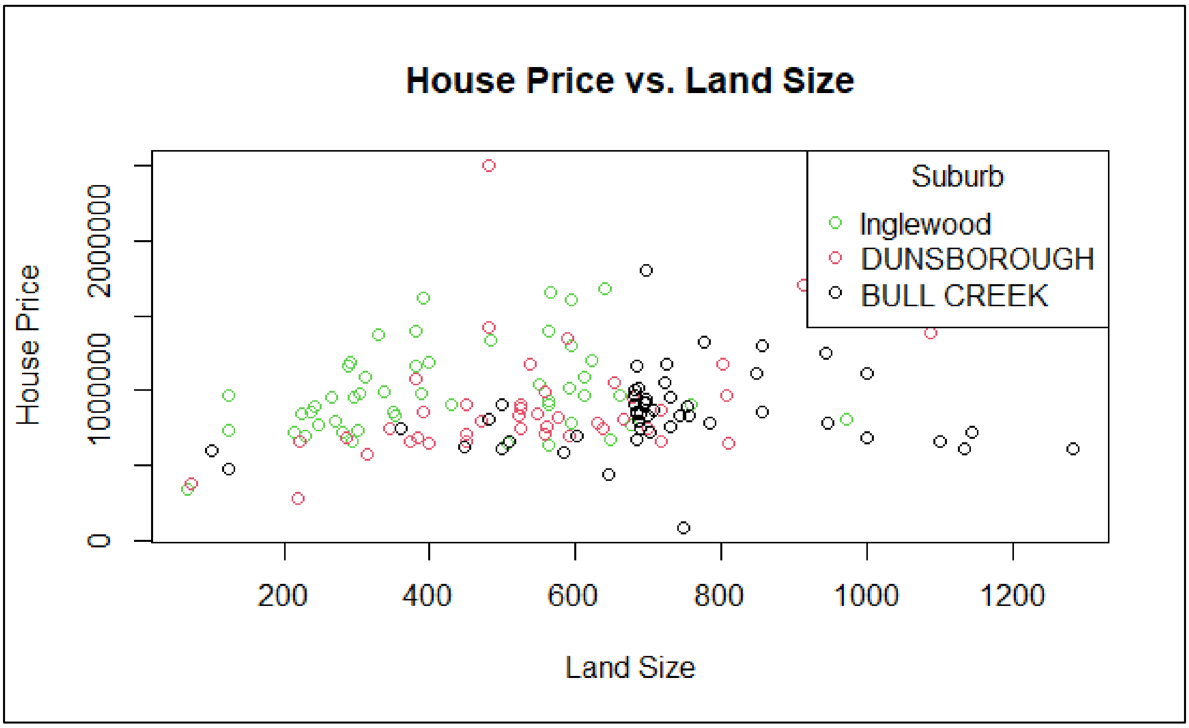


Figure 9 Scatterplot Price Vs Land Size

Houses situated on larger plots of land are reflected as being more expensive in the scatter plot of house prices versus land area. This is probably because bigger lots provide more room, seclusion, and building options.

However, there is a wide range in home prices, even among homes on similarly sized lots. Several factors, including the house's location, condition, and proximity to conveniences, are likely responsible for this.

The scatter plot shows the general trend between the home price and lot size in the neighbourhood. The scatter plot demonstrates merely correlation, not causation. The house's location and the quality of the surrounding school district may play a role in determining its asking price and lot size.

Houses on larger lots tend to be more expensive, as shown by the scatter plot you gave, which indicates a positive link between house price and land size. This is probably because bigger lots provide more room, seclusion, and building options.

House prices for lots of the same size exhibit much scatter in the scatter plot. Several factors, including the house's location, condition, and proximity to conveniences, are likely responsible for this. A lot of 600 square meters in a desirable area with excellent schools is likely to cost more than an equivalent lot in a less desirable area with subpar educational opportunities. Equally, a house on a 600-square-foot lot that has been well-maintained would command a higher price than a same-sized home that needs work. The scatter plot shows the general trend between the home price and lot size in the neighbourhood. Nonetheless, more aspects should be considered before a home purchase or sale.

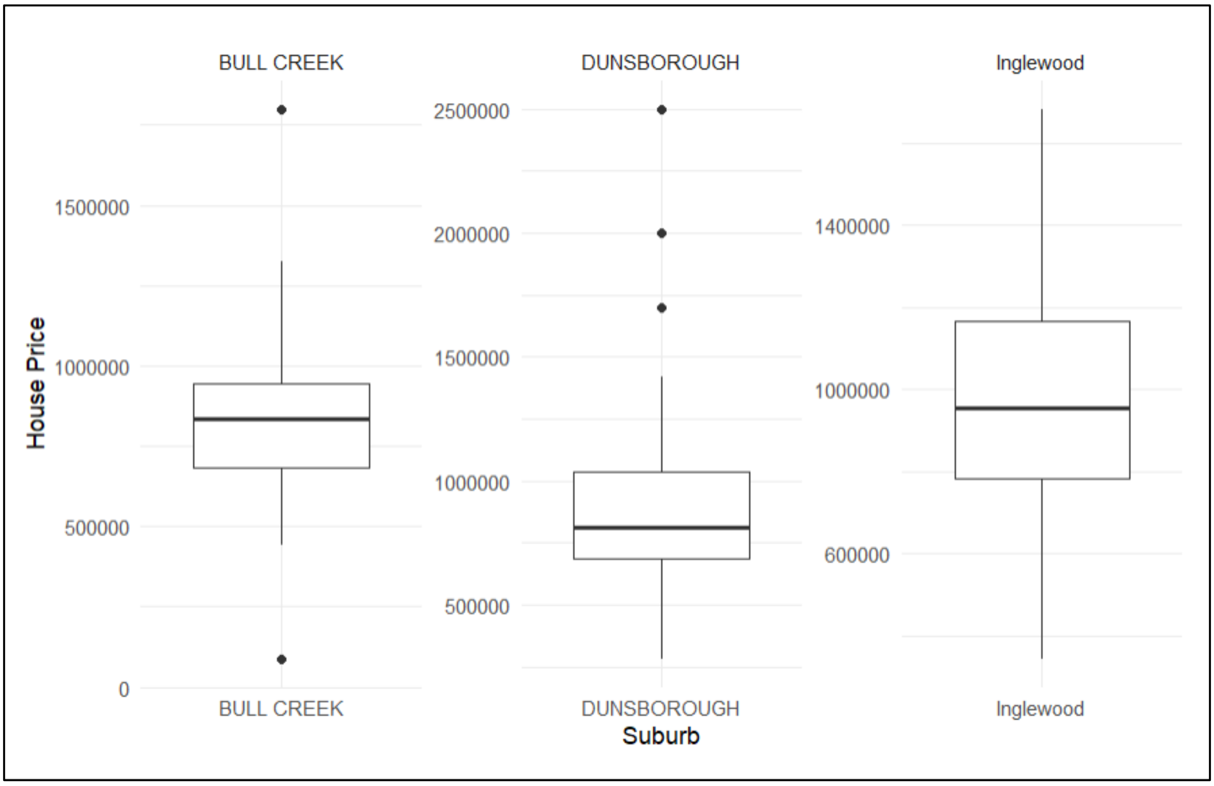


Figure 10 Boxplot

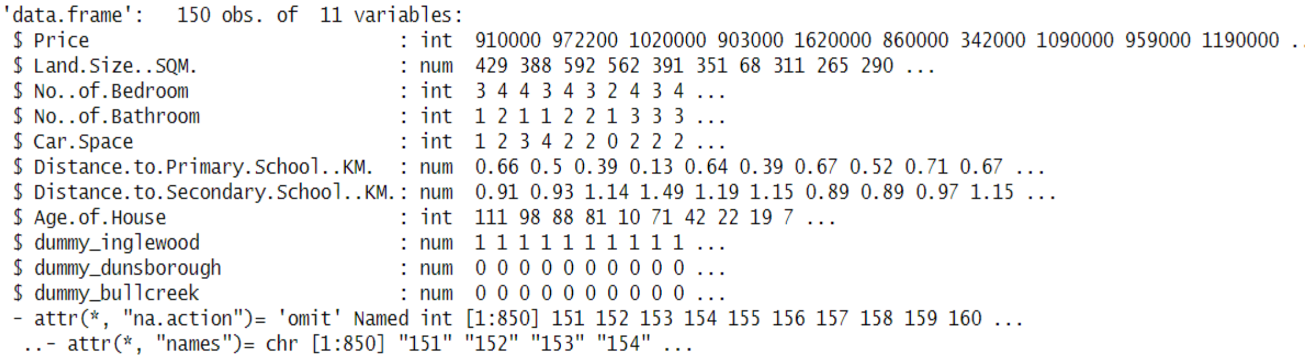
The boxplot uses the interquartile ranges to indicate where the value lies. Also, it shows the outliers for each of the suburb. The centre box shows the 50% of the data.

The upper part above is called Q3 , the part below the line is Q1 and there difference (Q3-Q1) is the interquartile range and the solid indicates the median of the values, For Bull Creek, we have two outliers one with the lowest value and one with the highest value. Whereas, Dunsborough has three outliers above Q3 with highest value in the suburb. Lastly, Inglewood does not have any outliers. Mainly, boxplot is used to indicate any outliers to check the value and frequency of outliers present in the dataset.

2.5 Multiple Regression Model

Multiple regression is a method to model the quantitative point on a graph to find a relationship between the points. It is the most common method for empirical research (Kelley & Bolin, 2013). To use the regression model, first, we created a dummy variable to represent them as a categorical variable for all our suburbs: Inglewood, Dunsborough, and Bull Creek. Furthermore, the "Address" and "Suburb" columns are excluded because these are not directly used in the analysis, and Dummy variables are already created for the suburbs; thus, the names of the suburbs are not required.

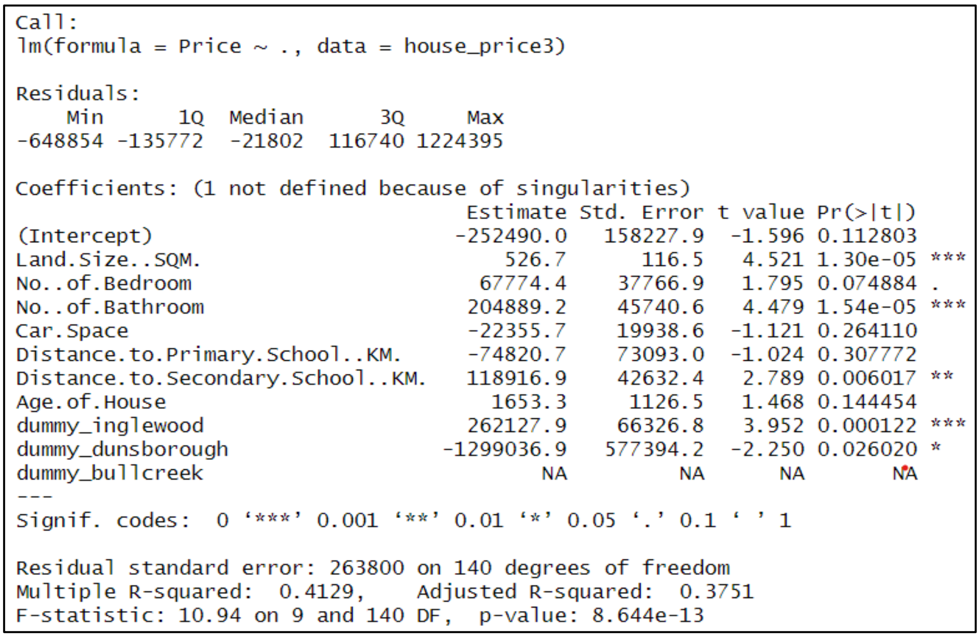
Overall, this part of the analysis is mainly inclined towards finding a relationship between the various property-related features (Land Size, No. of Bedroom, No. of Bathroom, Car Ports, Distance to primary school, Distance to secondary school, Age of house, dummy variable of suburbs) and price of the residence.



```
'data.frame':   150 obs. of  11 variables:
 $ Price                           : int  910000 972200 1020000 903000 1620000 860000 342000 1090000 959000 1190000 .
 $ Land.Size..SQM.                 : num  429 388 592 562 391 351 68 311 265 290 ...
 $ No..of.Bedroom                  : int  3 4 4 3 4 3 2 4 3 4 ...
 $ No..of.Bathroom                 : int  1 2 1 1 2 2 1 3 3 3 ...
 $ Car.Space                       : int  1 2 3 4 2 2 0 2 2 2 ...
 $ Distance.to.Primary.School..KM. : num  0.66 0.5 0.39 0.13 0.64 0.39 0.67 0.52 0.71 0.67 ...
 $ Distance.to.Secondary.School..KM.: num  0.91 0.93 1.14 1.49 1.19 1.15 0.89 0.89 0.97 1.15 ...
 $ Age.of.House                    : int  111 98 88 81 10 71 42 22 19 7 ...
 $ dummy_inglewood                 : num  1 1 1 1 1 1 1 1 1 1 ...
 $ dummy_dunsborough               : num  0 0 0 0 0 0 0 0 0 0 ...
 $ dummy_bullcreek                 : num  0 0 0 0 0 0 0 0 0 0 ...
 - attr(*, "na.action")= 'omit' Named int [1:850] 151 152 153 154 155 156 157 158 159 160 ...
  ..- attr(*, "names")= chr [1:850] "151" "152" "153" "154" ...
```

Figure 11 Data Frame of Variables

The results obtained with the Model are given below:



```
Call:
lm(formula = Price ~ ., data = house_price3)

Residuals:
    Min      1Q  Median      3Q     Max 
-648854 -135772  -21802  116740 1224395 

Coefficients: (1 not defined because of singularities)
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     -252490.0   158227.9  -1.596 0.112803    
Land.Size..SQM.                     526.7      116.5   4.521 1.30e-05 ***
No..of.Bedroom                    67774.4    37766.9   1.795 0.074884 .  
No..of.Bathroom                  204889.2    45740.6   4.479 1.54e-05 ***
Car.Space                        -22355.7    19938.6  -1.121 0.264110    
Distance.to.Primary.School..KM.  -74820.7    73093.0  -1.024 0.307772    
Distance.to.Secondary.School..KM. 118916.9    42632.4   2.789 0.006017 ** 
Age.of.House                       1653.3     1126.5   1.468 0.144454    
dummy_inglewood                  262127.9    66326.8   3.952 0.000122 ***
dummy_dunsborough               -1299036.9   577394.2  -2.250 0.026020 *  
dummy_bullcreek                        NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 263800 on 140 degrees of freedom
Multiple R-squared:  0.4129,	Adjusted R-squared:  0.3751 
F-statistic: 10.94 on 9 and 140 DF,  p-value: 8.644e-13
```

Figure 12 Summary of the Model

In this output, the points to notice are the significance of the p-value, t-value, and the Value of R-squared. Looking at ‘Pr(>|t|),’ we can tell about the significant and insignificant.

- Land Size is statistically significant
- No. of Bedroom are not statistically significant
- No. of Bathroom are statistically significant
- Car space is not statistically significant
- Distance to Primary. is not statistically significant
- Distance to Secondary. is statistically significant
- Age of House is not statistically significant
- Dummy variables Inglewood and Dunsborough represent categorical and statistically significant.

Adjusted R-squared is the goodness of fit and shows a percentage of 37.51%, which is a fairly good number for our model. Concerning the R-squared, the closer the value is to 1 or 100%, the better the model has a better goodness of fit. But, concerning our model, the percentage is expected to be low as the relationship is low between the variables.

Lastly, F-statistics show whether the whole model is significant or not. We obtained a 10.94, and the associated p-value is less than 0.05, so we can conclude that the model is statistically significant. Additionally, our model will tend to give us better accuracy for prediction, and the ANOVA table gives us a similar analysis of the model as above.

<u>MULTIPLE REGRESSION EQUATION:</u>

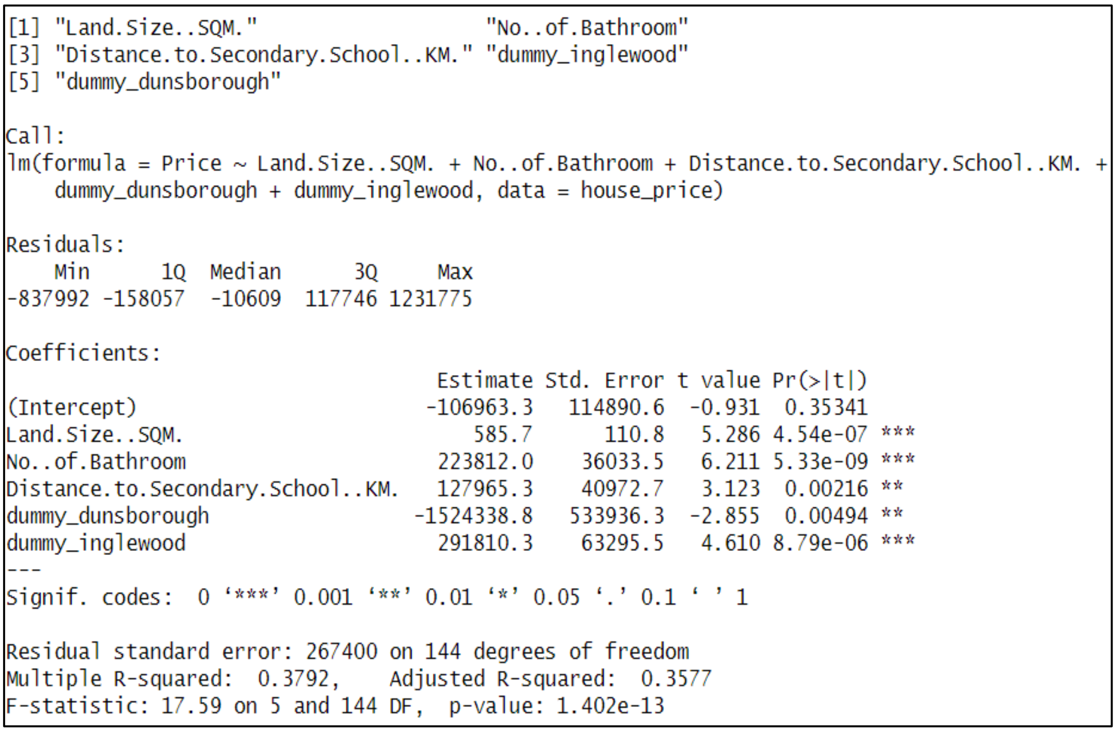
The relationship between the Criterion Variable (Dependent Variable) and Predictor Variable (Independent Variable) is examined through statistical techniques, which are based on the correlations (Jeon, 2015). With the use of coefficient calculation, an equation is managed to form which is:

```
Multiple Regression Equation:

Price =-252490+526.69 * Land.Size.SQM. +67774.36 * No..of.Bedroom
+204889.2 * No..of.Bathroom +-22355.74 * Car.Space +-74820.65 *
Distance. To.Primary.School..KM. +118916.9 * Distance.
To.Secondary.School..KM. +1653.3 * Age. Of.House +262127.9 *
dummy_inglewood +-1299037 * dummy_dunsborough +NA * dummy_bullcreek
```

2.6 Reduced Form Regression Model

Our dependent variable is “Price,” based on this, we predict other variables. The output includes the summary giving the information on the model's fit. Residuals are the difference between predicted and actual prices as reflected in Min, 1Q, Median, 3Q, and Max. Coefficients represent a change in Price (Dependent Variable) concerning a one-unit change in the other variable (Independent Variable).

In the next analysis step, the Regression Model is cleaned by removing all the insignificant variables. With this cleaning, only the following variables were left:



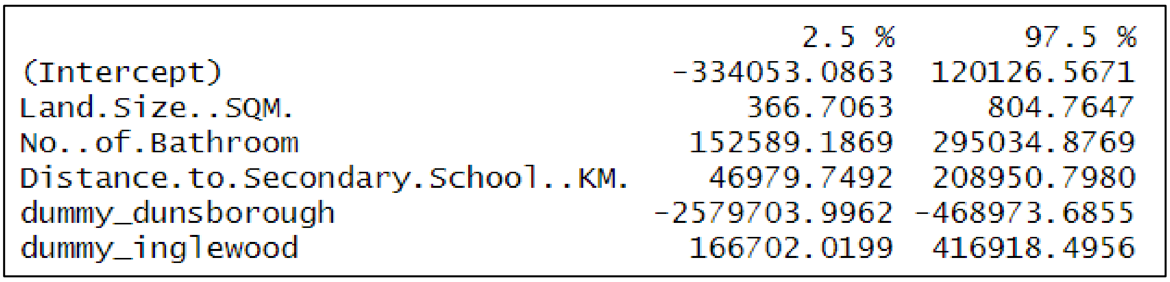
```
[1] "Land.Size..SQM."                  "No..of.Bathroom"
[3] "Distance.to.Secondary.School..KM." "dummy_inglewood"
[5] "dummy_dunsborough"

Call:
lm(formula = Price ~ Land.Size..SQM. + No..of.Bathroom + Distance.to.Secondary.School..KM. +
    dummy_dunsborough + dummy_inglewood, data = house_price)

Residuals:
    Min      1Q  Median      3Q     Max
-837992 -158057  -10609  117746 1231775

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        -106963.3   114890.6  -0.931  0.35341
Land.Size..SQM.                        585.7      110.8   5.286 4.54e-07 ***
No..of.Bathroom                     223812.0    36033.5   6.211 5.33e-09 ***
Distance.to.Secondary.School..KM.   127965.3    40972.7   3.123  0.00216 **
dummy_dunsborough                 -1524338.8   533936.3  -2.855  0.00494 **
dummy_inglewood                     291810.3    63295.5   4.610 8.79e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 267400 on 144 degrees of freedom
Multiple R-squared:  0.3792,    Adjusted R-squared:  0.3577
F-statistic: 17.59 on 5 and 144 DF,  p-value: 1.402e-13
```

Figure 13 Reduced Regression Model

In the reduced form, it is noted that the Adjusted R-squared is 35.77%. Similarly, the F-statistics is 17.59, and the associated P-value is still less than 0.05, so it can be concluded that the model is still statistically significant.

2.7 Confidence Intervals For All Slope Coefficients

Furthermore, the confidence interval is checked at 95%, and the results are shown below:



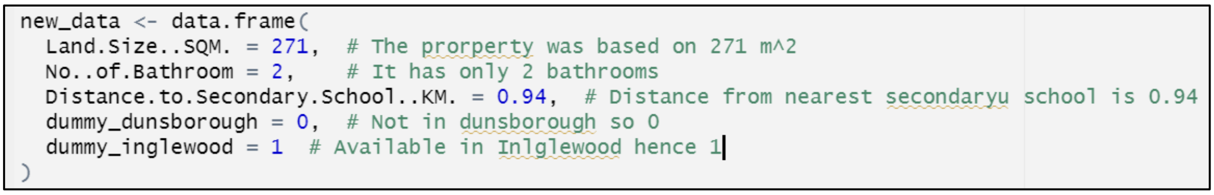
```
                                        2.5 %      97.5 %
(Intercept)                       -334053.0863 120126.5671
Land.Size..SQM.                       366.7063    804.7647
No..of.Bathroom                    152589.1869 295034.8769
Distance.to.Secondary.School..KM.   46979.7492 208950.7980
dummy_dunsborough                -2579703.9962 -468973.6855
dummy_inglewood                    166702.0199 416918.4956
```

Figure 14 Confidence Intervals

The 2.5% column shows the lower bound, which shows the range in which we can be 95% confident that the true population parameter lies. The analysis for this table is that we are 95% confident that the values for Land Size, No. of Bathroom, Distance to Secondary School, and dummy variables of Dunsborough and Inglewood fall within the ranges given in table 2.5% to 97.5%. It provides the uncertainty around estimated coefficients, which assists in understanding possible values that can lie within the range.

2.8 House Price Prediction

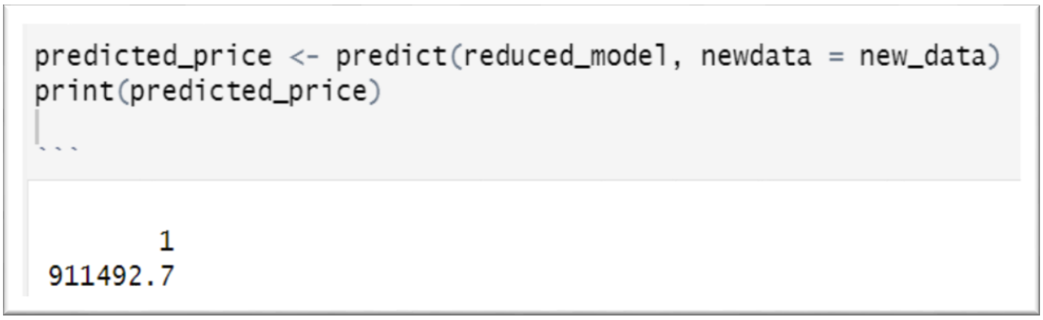
After creating a reduced model by choosing only the significant variables found, we tested the model by predicting the price of a house currently advertised in the market. We had to do thorough research as we wanted to find a house address available on realestate.com.au, and it also has an ANZ House report available. Due to this, we had to go back and forth by searching properties that fall on this criteria so we could use it for prediction. As the assignment requirement was only to test it on one property of any suburb, we randomly chose Inglewood. We searched for properties using filters that it is a house currently advertised for sale. One property we came across was **13A Walter Road West, Inglewood, WA 6052**. This address was available on both platforms and had all the information we needed to use our prediction model. As we reduced our model, we were left with only the significant variables, which include land size, no of bathrooms, distance to secondary schools, and suburbs. Therefore, we picked up those pieces of information from realestae.com.au and created a new data frame on R by including these relevant pieces of information. A screenshot of our prediction data frame can be found below.

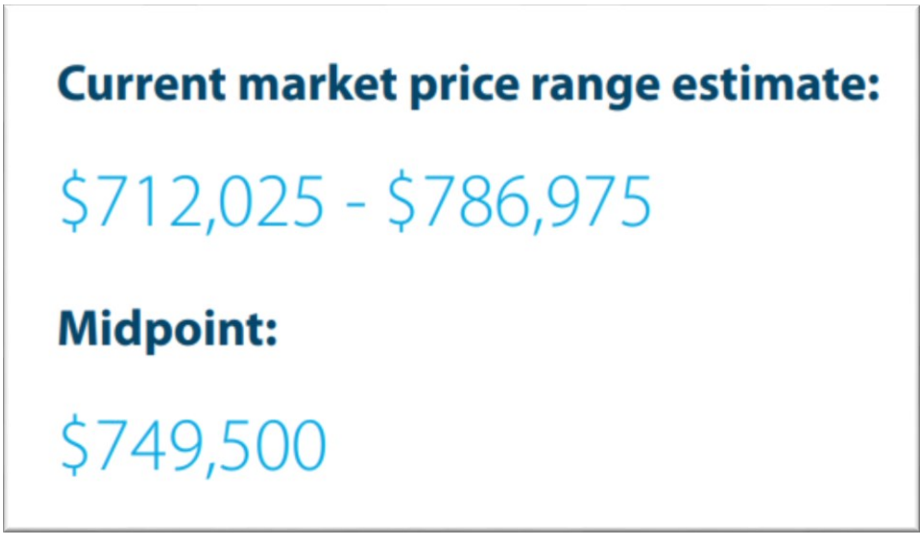


```
new_data <- data.frame(
  Land.Size..SQM. = 271,  # The prorperty was based on 271 m^2
  No..of.Bathroom = 2,    # It has only 2 bathrooms
  Distance.to.Secondary.School..KM. = 0.94,  # Distance from nearest secondaryu school is 0.94
  dummy_dunsborough = 0,  # Not in dunsborough so 0
  dummy_inglewood = 1  # Available in Inlglewood hence 1
)
```

Figure 15 Prediction Model

As we required only certain significant variables, the screenshot shows that the size of the land on which the house is based is 271m². It has only two bathrooms available. The nearest secondary school is 0.94 km away, and the house is in Inglewood's suburb. Once set, we used the predict function of R while using our reduced_model, alongside the values we have given above, to predict the price of this house. The results show that the house should be valued at a total price of $911492 based on such characteristics. This will further be compared with the results of the ANZ report to see how accurate the results are. A screenshot of the result and the code are shown below.



```
predicted_price <- predict(reduced_model, newdata = new_data)
print(predicted_price)
```
```

```
       1
911492.7
```

Figure 16 Predicted Value

2.9 Comparison With ANZ Forecast

The same address was found in the ANZ report for houses on sale, which allowed us to compare our model's accuracy in predicting the prices of currently advertised houses. Based on the report, we counter-checked all data we input in the model, and it shows that the house is approximately 271 sqm and has two bathrooms. It needs information about the nearest secondary schools, so we trusted realestate.com.au to provide the correct information. The report gives the estimated price range of the house to be between $712025 and $786975, while the midpoint is $749500.



Current market price range estimate:

$712,025 - $786,975

Midpoint:

$749,500

Figure 17 Market prices

As our model predicted the price to be $911,492, it shows that the prediction is 80.47% accurate as the absolute difference between the prediction and midpoint is 161,992, telling us that it is only 19.53% off the real price in the market. This is a fairly good result, considering that our model only had an R^2 of 0.357, so we
```

expected the results to be very different from the actual price. However, this difference between prediction and market value can be due to several reasons. A Few of them are mentioned below.

- First, in our model, the overall data set was only 150 rows. This was also reduced to 50 as each suburb has only 50 houses. Considering only 50 houses to predict the price of a currently listed house has a very high chance of proving inaccurate results as we have not fed the model enough data which it could use to learn and give us the results.
- Also, we have considered this model to be linear. There may be other dimensions that could add more depth to this model, making it nonlinear and improving the results overall. However, according to the assignment, we have only created a multi-regression model that shows a relationship between predictors and target variables, which is linear.
- As seen in the boxplots, while collecting data, we have encountered many outliers that can skew the results in one direction, causing inaccuracy. As we did not particularly account for dealing with outliers, these values may disproportionately influence our model.
- There is also a high chance the created model may be overfitting or Underfitting: The model may be overfitting the training data, capturing noise rather than the true underlying patterns, or it may be too simple (underfit), not capturing all the important relationships.
- The sample of 150 houses we used in the data is spread over a long period. Some were even dated a year earlier. This also opens up doors towards temporal effects as there has been a consistent change in economic activity and seasonality, which may be causing the results to be different now compared to what it is predicting.
- Similar to temporal, geographic factors may also come into play. Inglewood itself is a big suburb, so there may be factors apart from what we've chosen in the model that may be affecting the prices of houses. For example, nearness to a bus station or the mall may also play an important role not considered in this model.
- We only collected data from one website, realestate.com.au, which can be a question of the data quality. Even though the website is renowned for providing authentic data, we should still countercheck it with other websites and not completely trust the results from one, as discrepancies may also be one reason why the model prediction may be slightly off from the midpoint from the ANZ House report.

CONCLUSION

The assignment was a comprehensive journey that allowed us to start from scratch and go through the complete data analysis process. It started with data collection, which allowed us to understand the insights into real estate for the assigned suburbs, including Dunsborough, Inglewood, and Bullcreek. The next step of data collection allowed us to learn different methods of analyzing the data with statistical knowledge, which helped us get a deeper grasp of what exactly our data represents. Functions like a summary in R studio gave a complete overview of all the variables within the data set, allowing us to move towards the next step, model building.

With a keen understanding of the dataset, we constructed a multi-regression model that consisted of multiple variables, with the Price of the house being the predictor. We further used statistical knowledge discussed in the class to find which variables had more impact on the model. Through this, we selected only the significant ones and left the rest to reduce noise and improve the model's efficiency. This reduced model was then further used to predict the price of a selected house, which was later compared with an authentic ANZ report, allowing us to understand why the values may differ from our predicted price.

According to the figures obtained, the model has successfully obtained an approximately 80% accurate result. This shows that it can be considered a good model to predict house pricing; however, to make it even better, we must train this a lot more data points so that it can further improve the results. The model could also be tested further by adding more variables to the datasets, as there may be a high chance that other significant variables may exist that can improve the prediction power of the model. Therefore, our concluding remark would be that even though the model has returned a good result in terms of prediction, it can always be improved further by doing some tests and trials on the inputs until we receive an even higher accuracy.

REFERENCES

Bird, J., Mean, Median, Mode and Standard Deviation, *Basic Engineering Mathematics*, pp. 355–61, accessed 20 October 2023, September 2, 2018. DOI: 10.4324/9781315561776-49/MEAN-MEDIAN-MODE-STANDARD-DEVIATION-JOHN-BIRD

El-Nasr, M. S., Dinh, T. H. N., Canossa, A. and Drachen, A., Data Preprocessing, *Game Data Science*, pp. 33–58, accessed October 20, 2023, from https://academic.oup.com/book/39142/chapter/338586655, October 14, 2021. DOI: 10.1093/OSO/9780192897879.003.0002

Maiti, M., Risk Analysis, *Applied Financial Econometrics*, pp. 153–201, accessed October 20, 2023, from https://link.springer.com/chapter/10.1007/978-981-16-4063-6_6, 2021. DOI: 10.1007/978-981-16-4063-6_6

Pandey, S., Principles of Correlation and Regression Analysis, *Journal of the Practice of Cardiovascular Sciences*, accessed October 20, 2023, 2020. DOI: 10.4103/JPCS.JPCS_2_20

Peren, F. W., Inferential Statistics, *Statistics for Business and Economics*, pp. 77–109, accessed October 20, 2023, from https://link.springer.com/chapter/10.1007/978-3-662-64276-4_3, 2021. DOI: 10.1007/978-3-662-64276-4_3

Soltane, H. Ben, Naoui, K. and Alshammari, A., Systematic Illiquidity, Characteristic Illiquidity, and Stock Returns: Timeseries Analysis, *International Journal of Advances in Applied Sciences*, vol. **9**, no. 2, pp. 72–80, accessed October 20, 2023, February 1, 2022. DOI: 10.21833/IJAAS.2022.02.008

Simundic, A.-M. (2008). Confidence interval. *Biochemia Medica*, *18*(2), 154-161.

Young, P. and Shellswell, S., Time Series Analysis, Forecasting and Control, *IEEE Transactions on Automatic Control*, vol. **17**, no. 2, pp. 281–83, accessed October 20, 2023, April 27, 2004. DOI: 10.1109/TAC.1972.1099963

Jeon, E. H. (2015). Multiple regression. In *Advancing quantitative methods in second language research* (pp. 131-158). Routledge.

Kelley, K., & Bolin, J. H. (2013). Multiple regression. In *Handbook of quantitative methods for educational research* (pp. 69-101). Brill.