

**Melbourne Housing Market**

**Submitted By: Aniruddha Bose**

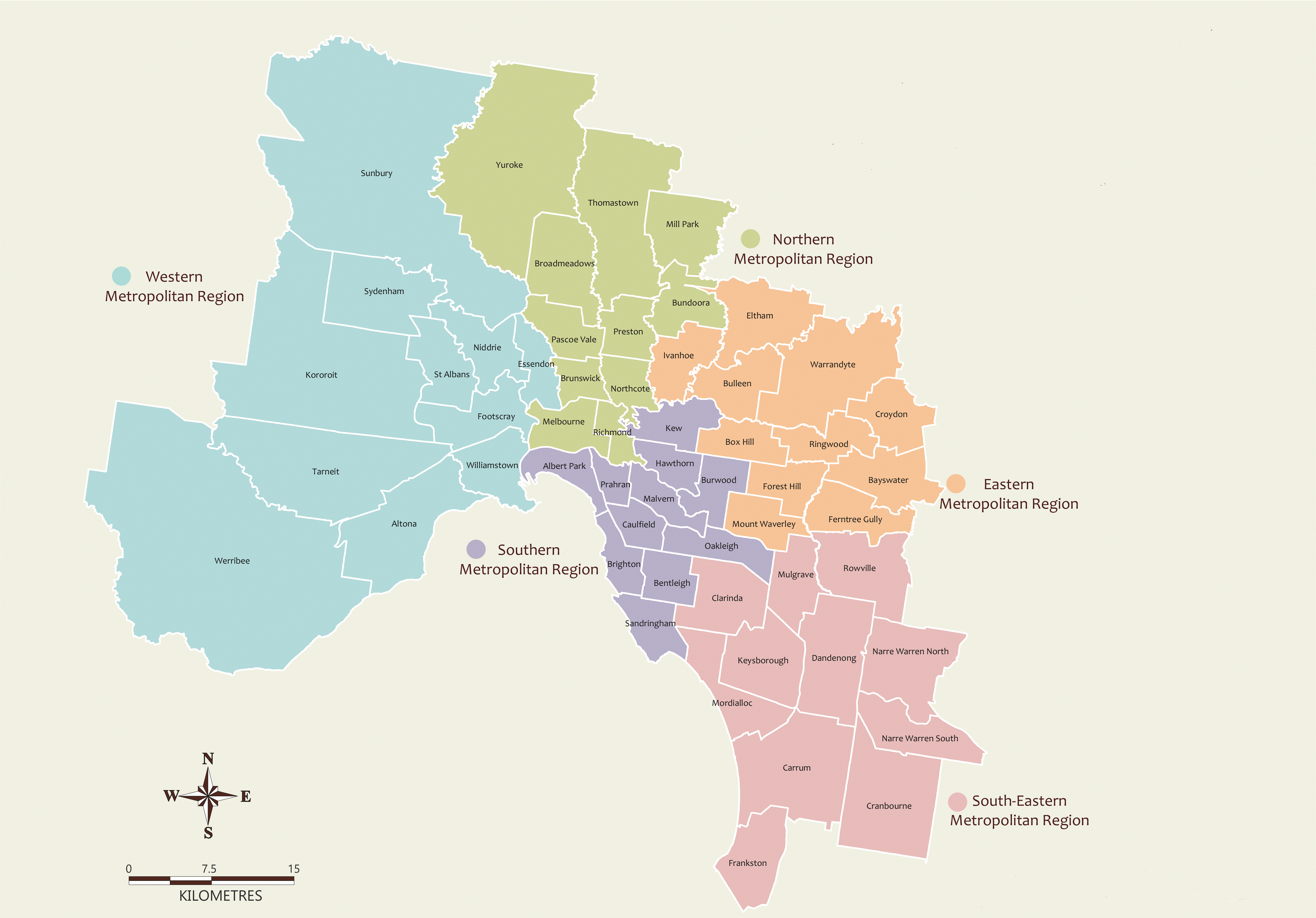
**Mentor: Ryan Rosario**

**Housing Prices**

Analysis and Model Building

Background

Melbourne, Australia is currently experiencing a housing bubble (some experts say it may burst soon). People living in Melbourne are interested to know the trend of prices of the properties located in Melbourne and its suburbs. Which suburbs/regions are the best to buy in? Which ones are value for money? Where is the expensive side of town? And more importantly where should an individual buy a 2-bedroom unit which is not that far Central Business District (CBD) of Melbourne?



Map prepared *by the Victoria Electoral Commission (VEC)*

*Copyright© 2014 – State Government of Victoria*

Task

After determining which factors relate to the selling prices of homes located in and around the Melbourne city center, the task was to develop a model to predict the housing prices.

Data

Data is readily available on a number of measures, including size of the land, location, age of the property, number of rooms, number of bathrooms, number of cars that can fit into the garage and distance from CBD. The dataset contains information of 12K+ residential properties sold during a recent 18-month period. The variables in the dataset are:

|  |  |
| --- | --- |
| **Variables** | **Description** |
| Price | Selling price of the property in AUD (Australian Dollar) |
| Suburb | A suburb is a [residential area](https://en.wikipedia.org/wiki/Residential_area) or a mixed use area, either existing as part of a city or urban area or as a separate residential community within [commuting](https://en.wikipedia.org/wiki/Commuting) distance of a city. |
| Address | Address of the property |
| Rooms | Number of rooms in the house |
| Type | Type of the Property.  h - house, cottage, villa, semi, terrace;  u - unit, duplex;  t - townhouse; dev site - development site;  o res - other residential |
| Method | The process is which the property has been sold  S - property sold;  SP - property sold prior;  PI - property passed in;  PN - sold prior not disclosed;  SN - sold not disclosed;  NB - no bid;  VB - vendor bid;  W - withdrawn prior to auction;  SA - sold after auction;  SS - sold after auction price not disclosed. N/A - price or highest bid not available. |
| Seller | Real Estate Agent |
| Date | Property Sold Date |
| Distance | Distance from CBD (Central Business District) |
| Postcode | Zip code of the Property where it is located |
| Bedroom | Number of bedrooms in the house |
| Bathroom | Number of bathrooms in the house |
| Car | Number of cars that can fit into the garage |
| Landsize | The size of the property plot |
| BuildingArea | Area of the building in Sqft |
| YearBuilt | The year when this house was built |
| CouncilArea | Council is the area defined by Local Government |
| Latitude | Latitude and longitude are angles that uniquely define points on a sphere. Together, the angles comprise a coordinate scheme that can locate or identify geographic positions on the surfaces of planets such as the earth. |
| Longitude | Latitude and longitude are angles that uniquely define points on a sphere. Together, the angles comprise a coordinate scheme that can locate or identify geographic positions on the surfaces of planets such as the earth. |

Data Management

Data management is the process of cleaning and unifying messy and complex datasets for easy access and analysis.

1. The junk Suburbs’ records are removed by dropping meaningless Suburb named as ‘RE’.
2. The Properties having missing price have been removed from analysis.
3. Missing 'Landsize' of any Property has been replaced with ‘Median’ Landsize of the Properties in a Suburb. ‘Mean’ Landsize of the properties in a Suburb is not considered for the presence of extreme values which are nothing but outliers.
4. The date field (Property Sold Date) has been formatted from Australian standard (dd/mm/yyyy) to US standard (mm/dd/yyyy).
5. There are around 70 suburbs. If we consider all the suburbs for analysis, apparently the more useful the model would be however the opposite is actually true – it greatly weakens the model’s predictive analysis. Hence we decided to categorize the suburbs in 5 regions (East, West, North, South & Central). Prepared a dataset having Suburb name, City name and Region and finally merged with the original dataset.

Data Visualization

Figure 1: **Number of houses by Region**

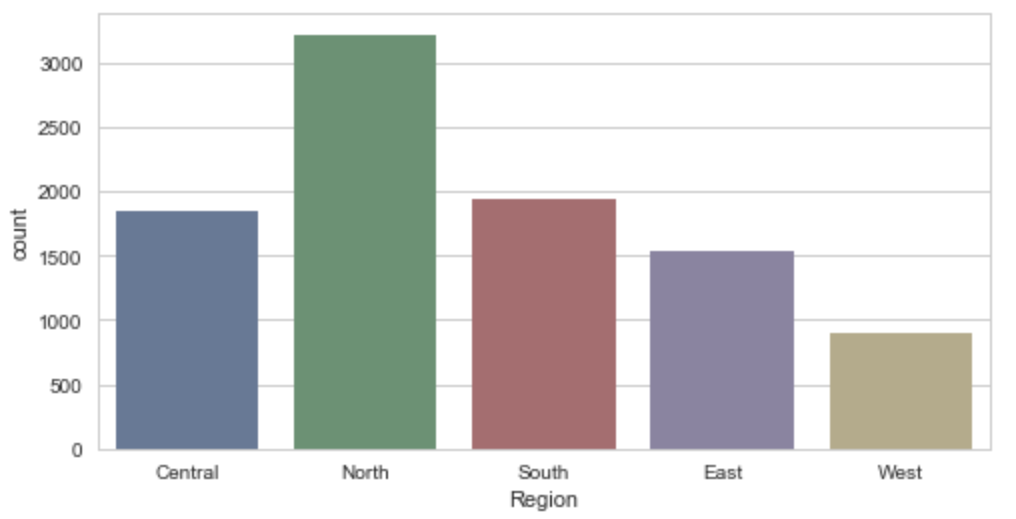
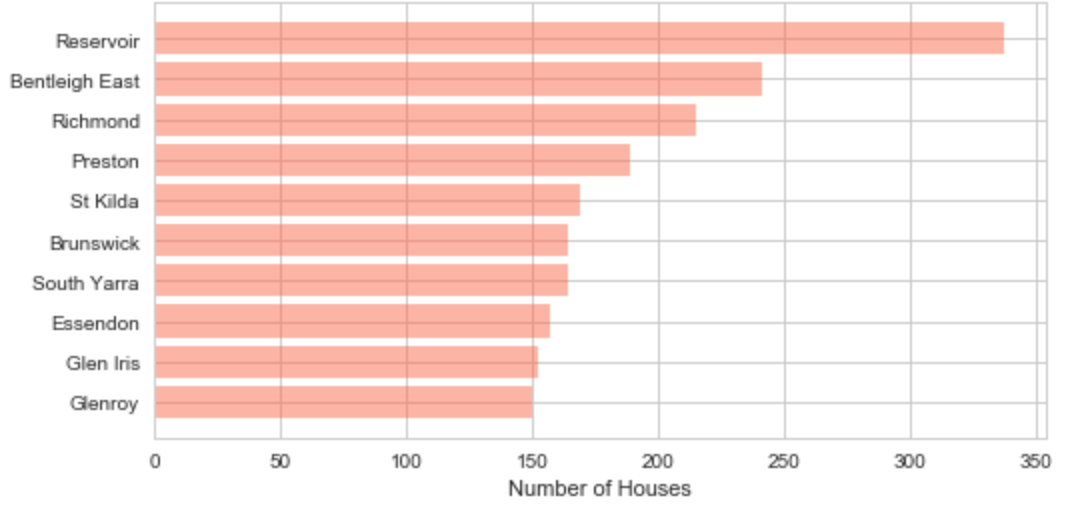


Figure 2: **Top 10 Suburbs by the Number of houses**



Detailed visualization and its Analysis

This analysis involves one response variable, the selling price of the home, and various potential predictors of selling price. From our initial analysis, we think seven (7) of the predictor variables measure, directly or indirectly: the ‘Size of the Land’, ‘Number of Rooms’, ‘Garage’, ‘Distance from CBD’, ‘Property Type’, ‘Suburb Region’ and ‘Property Sold Date’ could influence the price of the property. Among these 7 Independent variables ‘Property Type’ and ‘Suburb Region’ are considered to be categorical or dummy variables. There are other Independent Variables which can be considered. But adding more predictors to a multiple regression procedure does not mean that regression would be "better" or offer better predictions. Infact it can make things worse. This phenomenon is known as 'Overfitting'.

Figure 3:Scatterplot/Boxplot/Line-chart Matrix

|  |  |
| --- | --- |
| **'Property Price' vs 'Landsize'** | **'Property Price' vs 'CBD'** |
|  |  |
| **'Property Price' vs 'CBD' with ‘Region’** | **'Property Price' vs 'CBD' at ‘Central Region’** |
|  |  |
| **'Property Price' vs 'CBD' at ‘Northern Region’** | **'Property Price' vs 'CBD' at 'Southern Region’** |
|  |  |
| **'Property Price' vs 'CBD' at ‘Eastern Region’** | **'Property Price' vs 'CBD' at ‘Western Region’** |
|  |  |
| **'Property Price' vs 'Property Sold Date'** |  |
|  |  |
|  |  |
| **'Property Price' vs 'Number of Rooms'** | **'Property Price' vs 'Number of Rooms' with various Property Types - ‘h’, ‘u’ & ‘t’** |
|  |  |
| **'Property Price' vs 'Number of Rooms' with Property Type ‘h’ -** house, cottage, villa | **'Property Price' vs 'Number of Rooms' with Property Type ‘u’ - unit, duplex** |
|  |  |
| **'Property Price' vs 'Number of Rooms' with Property Type ‘t’ - townhome** | **'Property Price' vs 'Number of Cars fit into a Garage'** |
|  |  |

**From the data distributions, we can conclude that among the considered predictors except 'Landsize' all others ('Property Type', 'Property Sold Date', 'Distance from CBD (Central Business District)', 'Suburb Region', 'Number of Rooms' and ‘Number of cars fit into the Garage’) have some influence over the 'Property Price'.**

Regression Analysis

Now we will try to find potential relationship between each predictor and the ‘Sales Price’ using mathematical equations. In multiple regression, each coefficient is interpreted as the estimated change in ‘Sales Price’ corresponding to a unit change in a predictor, when all other predictors are held constant.

**Multicollinearity** is also an important aspect. This occurs when some/all predictors are correlated with each other. If we find high multicollinearity between 2 predictors both the variables cannot be considered in the regression equation; they are redundant.

Table 1:Simple regression models for each of these predictors

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Adj R^2** | **Model p-value** | **Coefficient p-value** |
| Price = 1192481 - 13622 \* [CBD] | 0.005 | 0.000 | 0.000 |
| Price = -133365691 + 183 \* [Property Sold Date] | 0.001 | 0.00121 | 0.001 |
| Price = - 25767 + 385203 \* [Number of Rooms] | |  |  | | --- | --- | |  | 0.315 | | 0.000 | 0.000 |
| Price = - 762741 + 208655 \* [Number of Cars that can fit into the Garage] | 0.086 | 0.000 | 0.000 |

The above table indicates that, based on the single-predictor models, the house price increases on average by:

* *Model#1*: AU$ 13,622 for each KM closer towards the CBD
* *Model#2*: AU$ 183 for each Day newer
* *Model#3*: AU$ 385,203 for each room addition
* *Model#4*: AU$ 208,655 for each addition of car into the Garage

But at the same time as the R^2 value is very low for all the models we can conclude that single predictor is **not** significantly influencing the Property price. From the above analysis we can also infer that the Independent variable 'Property Sold Date' is grossly impacting the R^2 value; almost ‘Zero’ R^2 value. Moreover, with huge negative intercept value (-133365691) the 2nd model suggests that there are homes with ‘zero’ room which is meaningless. **Hence it has been decided not to consider 'Property Sold Date' as predictor in the regression analysis.**

Table 2:Multiple regression models using 2 predictor variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Adj R^2** | **Model p-value** | **Coefficient p-value** |
| Price = 245716 + 440041\*[Number of Rooms]  – 48963 \*[CBD] | 0.379 | 0.000 | Distance: 0.000  Rooms: 0.000 |
| Price = 970831 + 243924\*[Number of Cars that can fit into the Garage] – 30630\*[CBD] | 0.112 | 0.000 | Distance: 0.000  Cars: 0.000 |
| Price = -40450 + 365837\*[ Number of Rooms]  + 54584\*[Number of Cars that can fit into the Garage] | 0.326 | 0.000 | Rooms: 0.000  Cars: 0.000 |

The above table indicates that, based on the two predictor models, the house price increases on average by:

* *Model#5*
  + AU$ 440,041 for each addition of room if CBD is held constant
  + AU$ 48,963 for each KM closer towards the CBD if number of rooms are held constant
* *Model#6*
  + AU$ 243,924 for each addition of car fits into the Garage if CBD is held constant
  + AU$ 30,630 for each KM closer towards the CBD if number of cars fit into the Garage are held constant
* *Model#7*
  + AU$ 365,837 for each car addition of room if number of cars that can fit into the Garage are held constant
  + AU$ 54,584 for each addition of car fits into the Garage if the number of rooms are held constant

From the above analysis we can also infer that the 6th model has much less impact on the housing Price. **Hence, we discard the 6th model of regression analysis.**

Checking of Multicollinearity among the predictor variables

Table 3: **Pairwise Correlation Matrix of the Predictor variables**

|  |  |  |  |
| --- | --- | --- | --- |
| **r = Corr(X, Y)** | Distance from CBD | Number of Rooms | Cars in the Garage |
| Distance from CBD |  |  |  |
| Number of Rooms | 0.299916 |  |  |
| Cars in the Garage | 0.295145 | 0.405438 |  |
| Property Sold Date | -0.006659 | -0.011493 | 0.016554 |

From the above correlation matrices we can suggest that none of the pairwise correlations among the predictor variables are particularly strong (*r* < 0.50 in each case). Only the ‘Number of Rooms’ and ‘Number of Cars fit into the Garage’ are moderately correlated (r = 0.41). We will do further analysis on the impact that this multicollinearity when we try to find Price of the property using both ‘Number of Rooms’ and ‘Number of Cars fit into the Garage’ as predictor variables in the regression analysis.

Table 4:Multiple regression models using 3 predictor variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Adj R^2** | **Model p-value** | **Coefficient p-value** |
| Price = 219232  + 405185\*[Number of Rooms]  + 96667\*[Number of Cars that can fit into the Garage]  – 50940\*[CBD] | 0.395 | 0.000 | Rooms: 0.000  Car: 0.000  Distance: 0.000 |

The *Model#8* indicates that, based on the triple-predictor variables, the house price increases on average by:

* AU$ 405,185 for each addition of room where CBD and Number of cars that can fit into the Garage are held constant
* AU$ 96,667 for each addition of car fits into the Garage where CBD and Number of Rooms are held constant
* AU$ 50,940 for each KM closer towards the CBD where Number of Rooms and Number of cars that can fit into the Garage are held constant

**Now we calculate Variance Inflation Factor (VIF) for the predictor “Number of Cars that can fit into the Garage”. Regressing the predictor “Number of Cars that fit into the Garage” on the remaining 2 predictors**

The calculated R^2 is 0.199 then

**VIFCars** = 1 / (1 – R^2) = 1 / (1 – 0.199) = 1.25

This variance inflation factor tells us that the variance of the “**Number of Cars that can fit into the Garage**” coefficient is inflated by a factor of 1.25. This is a very satisfactory number and it appears there is hardly any variance inflation that we have otherwise anticipated from the correlation analysis between the “Number of Rooms” and the “Number of Cars that can fit into the Garage”. In general, if an individual looks for a house with more number of rooms mean his family is big and most of the family members have their own car; hence it requires a bigger garage which can accommodate more cars. But in reality it may not be true; most of the family members might be young or children or the family members stay very near to the city’s main center may not require car for day-to-day commute – small garage works for them.

Adding Dummy/Categorical variable/s to the analysis

Table 5: **Adding just ‘Region’ (East, West, North, South and Central) in the analysis**

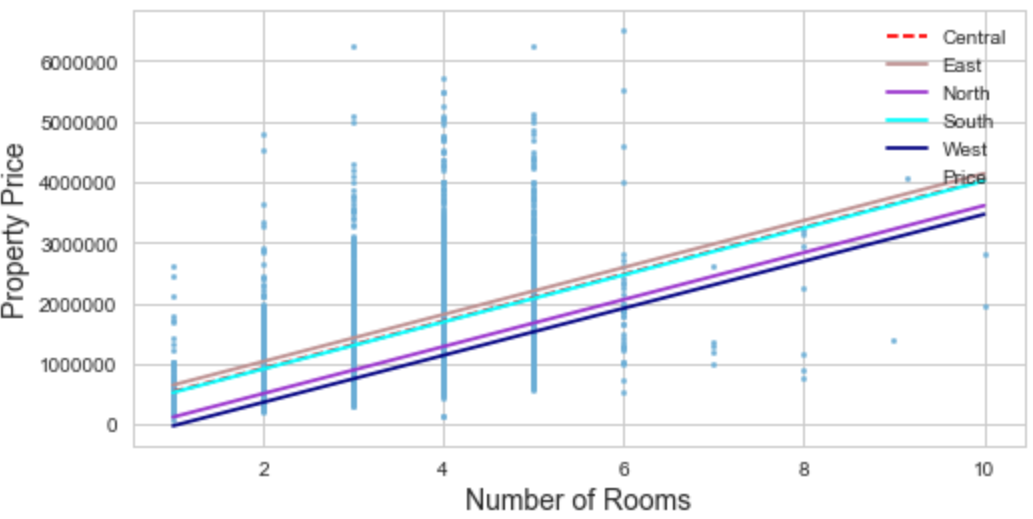
|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Adj R^2** | **Model p-value** | **Coefficient p-value** |
| Price = – 59939\*[CBD]  + 394412\*[Number of Rooms]  + 92806\*[Number of Cars fit into the Garage]  + 640490\*[East]  + 33212\*[West]  + 137069\*[North]  + 625488\*[South]  + 278890\*[Central] | 0.519 | 0.000 | Distance: 0.000  Rooms: 0.000  Car: 0.000  East: 0.000  West: 0.262  North: 0.000  South: 0.000  Central: 0.000 |

The *Model#9* indicates that, based on the triple-predictor variables along with ‘Region’ as Categorical variable, the house price increases on average by:

* AU$ 394,412 for each addition of room where CBD and Number of cars that can fit into the Garage are held constant
* AU$ 92,806 for each addition of car fits into the Garage where CBD and Number of Rooms are held constant
* AU$ 59,939 for each KM closer towards the CBD where Number of Rooms and Number of cars that can fit into the Garage are held constant

The impact of categorical variables can be best described graphically.

Figure 4: Influence of categorical variable 'Region' on the relationship between the 'Property Price' and the 'Number of Rooms'.



**We noticed that the lines are parallel. This is because the categorical variable affects only the intercept not the slope which is a function of ‘Number of Rooms’. From the graph, we can say that the properties in the East region are the most expensive as compare to other region properties.**

The p-value for West (0.262) is greater than the common alpha level of 0.05, which indicates that the relationship is not statistically significant. A larger (insignificant) p-value means the coefficient of a variable is almost Zero, that variable is not helping to predict the dependent variable. In other words, changes in the predictor are not associated with changes in the response. In our next analysis when we include Property Type would like to see if the Region ‘West’ still remains statistically insignificant.

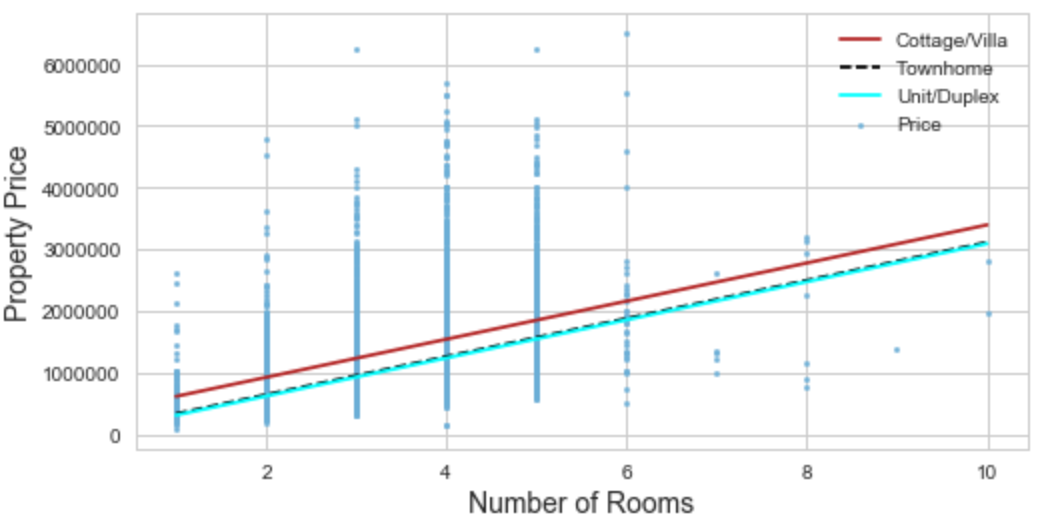
Table 6: **Excluding Region and adding ‘Property Type’ in the analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Adj R^2** | **Model p-value** | **Coefficient p-value** |
| Price = – 52646\*[CBD]  + 318964\*[Number of Rooms]  + 90892\*[Number of Cars fit into the Garage]  + 599091\*[PropTypeH]  + 360359\*[PropTypeT]  + 246945\*[PropTypeU] | 0.435 | 0.000 | Distance: 0.000  Rooms: 0.000  Car: 0.000  PropTypeH:0.000  PropTypeT:0.000  PropTypeU:0.000 |

The *Model#10* indicates that, based on the triple-predictor variables along with ‘Property Type’ as Categorical variable, the house price increases on average by:

* AU$ 318,964 for each addition of room where CBD and Number of cars that can fit into the Garage are held constant
* AU$ 90,892 for each addition of car fits into the Garage where CBD and Number of Rooms are held constant
* AU$ 52,646 for each KM closer towards the CBD where Number of Rooms and Number of cars that can fit into the Garage are held constant

Figure 5: Influence of the categorical variable 'Property Type' on the relationship between the 'Property Price' and the 'Number of Rooms'.



**It is clear from the above graph that the Cottage/Villa are the costliest as compare to other type of properties.**

Table 7: **Now adding both ‘Region’ and ‘Property Type’ in the analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Adj R^2** | **Model p-value** | **Coefficient p-value** |
| Price = – 63171\*[CBD]  + 267220\*[Number of Rooms]  + 89556\*[Number of Cars fit into the Garage]  + 871541\*[CentPrTyH]  + 449481\*[CentPrTyT]  + 300059\*[CentPrTyU]  + 1310795\*[EastPrTyH]  + 801356\*[EastPrTyT]  + 496622\*[EastPrTyU]  + 476584\*[WestPrTyH]  + 351009\*[WestPrTyT]  + 314062\*[WestPrTyU]  + 589390\*[NrthPrTyH]  + 437979\*[NrthPrTyT]  + 399726\*[NrthPrTyU]  + 1252049\*[SuthPrTyH]  + 907885\*[SuthPrTyT]  + 627385\*[SuthPrTyU] | 0.610 | 0.000 | Distance: 0.000  Rooms: 0.000  Car: 0.000  CentPrTyH:0.000  CentPrTyT:0.000  CentPrTyU:0.000  EastPrTyH:0.000  EastPrTyT:0.000  EastPrTyU:0.000  WestPrTyH:0.000  WestPrTyT:0.000  WestPrTyU:0.000  NrthPrTyH:0.000  NrthPrTyT:0.000  NrthPrTyU:0.000  SuthPrTyH:0.000  SuthPrTyT:0.000  SuthPrTyU:0.000 |

The *Model#11* indicates that, based on the triple-predictor variables along with ‘Region’ and ‘Property Type’ as Categorical variables, the house price increases on average by:

* AU$ 267,220 for each addition of room where CBD and Number of cars that can fit into the Garage are held constant
* AU$ 89,556 for each addition of car fits into the Garage where CBD and Number of Rooms are held constant
* AU$ 63,171 for each KM closer towards the CBD where Number of Rooms and Number of cars that can fit into the Garage are held constant

Figure 6: **Graphical Representation of the Model#11 – the influence of Categorical variables on the relationship between ‘Property Price’ and ‘Distance from CBD’**

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

***From the* Figure# 6 *we can say that the price of a property decreases steadily as the distance from CBD increases. It is true across all regions. And another interesting observation that across all regions, among the properties Cottages/regular houses are the costliest and on the other-hand duplex properties are less expensive, rather affordable. The properties in the 'Eastern' region for all property types are the most expensive whereas the price of the properties in the 'Western' region are the lowest.***

Model Summary

After carefully analyzing the results of all the Regression models (11 altogether) created so-far we can conclude that the **Model #11** is the most justified as far as the available data and its important/impacting independent variables are concerned because it has the highest Adjusted R^2 values. There is no multicollinearity among the considered Independent variables. The Adjusted R^2 is also same as the R^2 value. The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low/tend to zero p-value (< 0.05) indicates that we can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to the model because changes in the predictor's value are related to changes in the response variable. In this detailed analysis ‘Stepwise Regression’ has been followed. It provides an automated approach for identifying important variables and simplifying (reducing) the model.

And most importantly the relationships between the predictors and the ‘Sales Price’ established in the model are meaningful because as the number of rooms go up the price of the property may go up; similarly, if the garage can fit more cars the house will become costlier and again as the property moves further from CBD the price of the property decreases. The region and type of property have impact on the price where cottage is expensive than a duplex/small unit. The addition of categorical variables improves the R^2 value.

***Model #11*** *is the best model which fits to the dataset accurately. This model is not crowded with non-significant predictors; hence overfitting has been avoided.*

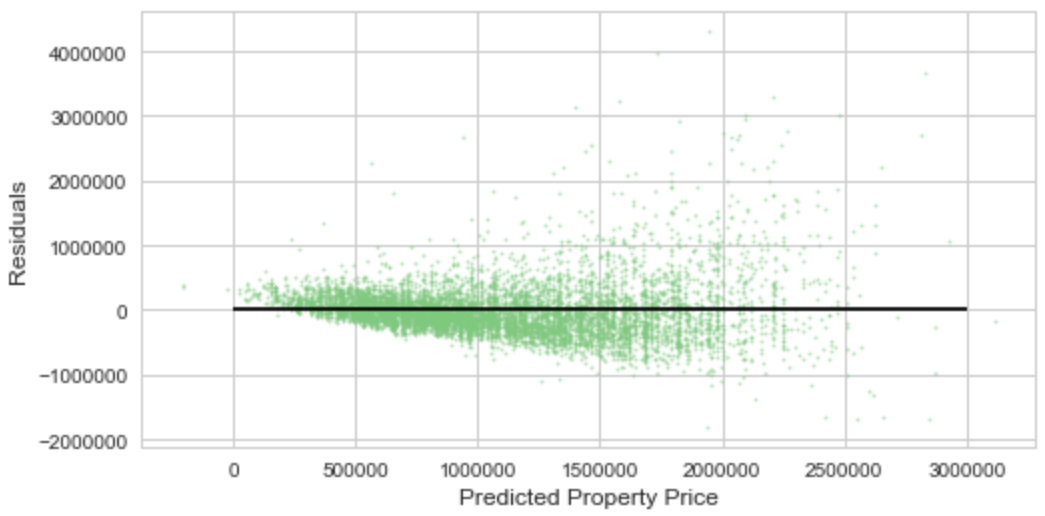
The model will help an individual to estimate/predict the property price if he/she has the information of the rooms present in the property, how big is the garage, how far is the property from CBD, suburb region and type of property. It is very convenient for an outsider to estimate the price of a house with some basic information. With this estimated number, the individual would be in a better position to negotiate with the sales agent/home owner/seller before zeroed in his/her choice of house.

Model Diagnostics

Residual Plots

Residuals represent the predictor errors. Residual plots are an effective way to visualize the errors in the model. If we can derive the right regression model then the residuals should be randomly scattered around the line zero.

Figure 7:



Per **Model #11**, the residuals against the Predicted Property Price have not been scattered randomly around ‘zero’. Randomness can be observed in patches. We are getting funnel-like shape. This is known as ‘**Heteroskedasticity**’ – the standard deviation is small to the left of the plot and large to the right. It occurs when different observations have different error variance.

OLS estimators are unbiased and consistent in the presence of heteroskedasticity, but they are not efficient and the estimated standard errors are inconsistent even though it is correctly measured, usual t statistics or F statistics or LM statistics for drawing inferences become no longer valid.

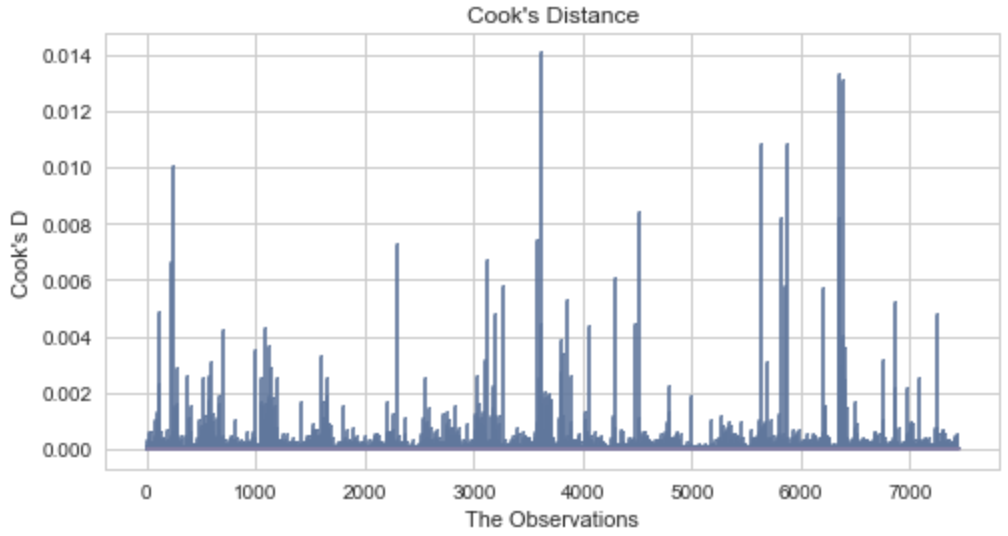
**Dealing with** **Heteroskedasticity:**

The easiest and most common solution is to use inefficient OLS estimator but use “robust” standard errors that allow for the presence of heteroskedasticity. Otherwise, weighted least squares (WLS) can be used to calculate efficient estimators, conditional on correct knowledge of the pattern of heteroskedasticity. This is a better solution if we know the pattern, which we usually don’t. The most common way to improve a model is to transform one or more variables, usually using a “log” transform. In our analysis, the size of the house (building area) which is a very important predictor is not considered due to insufficient data. The goal of OLS regression is to fit a straight line to the data at hand. If the functional form of the relationship between the dependent and independent variables is not linear, to avoid unreliable result it is advisable to use a better fit non-linear model, than the linear regression model.

Cook’s D

Cook’s D is a measure of the influence an individual point has on the model. Observations with Cook’s D values >1 are generally considered to be influential, high-leverage points.

Figure 8:

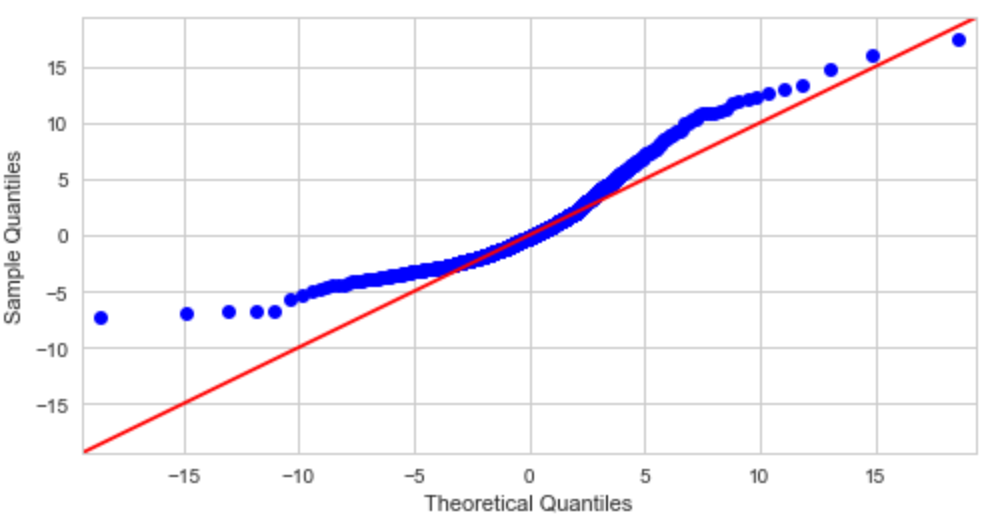


Per Figure #8, the Cook’s D value of all the observations is less than 1. The result indicates that there are very few influential points from the considered observations of the dataset.

Quantile Plot of Residuals

The quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.

Exhibit 15: **Left-skewed data**



The circles in this quantile plot start out on one side of the line, then are almost exclusively on the other side for a long stretch, then move to the other side of the line again. This behavior indicates some degree of skewing. The fact that the circles start consistently above the line, then stay consistently below it, then rise above it indicates *left* skewing.

Limitations and Future Work

As a future scope, the analysis could include other factors which may influence the decision of the buyer. For example, the property price of Eastern and Southern regions is more expensive than the property price of Central region which is surrounding the CBD. The other factors become out of scope of current analysis due to non-availability of relevant data.

**Scope of Improvement**

* *Look for a dataset which has very important predictor ‘size of the house’ (building area) of all the properties in Melbourne and its suburbs*
* *Need to relook on the outliers to avoid biasedness*
* *When we did model diagnostics we found the Residual plots is ‘****Heteroskedasticity’.*** *Heteroskedasticity diminishes the advantages of OLS model. Need more research to come up with alternative solution so that Heteroskedasticity can be minimized.*
* *Adjusted R^2 which penalizes model complexity to control for overfitting, generally under penalizes complexity. The****best****approach to feature selection is actually Cross validation. It provides a more reliable estimate of out-of-sample error, and thus is a better way to choose which of the models will best****generalize****to out-of-sample data.*

The statistical model developed for this analysis can be used as a crude way to identify which houses are statistically over- or undervalued. Houses with large studentized residuals (>3) may be overpriced relative to the statistical model. Likewise, houses with large negative studentized residuals (<3) may be underpriced relative to the model – they might end up being great bargains!

Conclusion

**Now the important question**: **If an individual wants to get a unit with two rooms, have it within cycling distance (approx. 2 Kms) of CBD (most likely the work place), with 1 car for other purpose how much the individual might need to pay??**

Per Model #11, the linear equation is

Price = – 63171\*[CBD] + 267220\*[Number of Rooms] + 89556\*[Number of Cars fit into the Garage]

+ (300059\*[CentPrTyU] OR 496622\*[EastPrTyU] OR 314062\*[WestPrTyU] OR 399726\*[NrthPrTyU] OR

627385\*[SuthPrTyU])

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Region** | **Property Type** | **# of Rooms** | **# of Car** | **Distance from CBD (KM)** | **Indicative Price to Pay** |
| Central | Unit | 2 | 1 | 2 | AU$ 797713 |
| East | AU$ 994276 |
| West | AU$ 811716 |
| North | AU$ 807824 |
| South | AU$ 1125039 |

**Clearly, the most affordable places to buy house (unit/duplex) are Northern and Central region.**

Acknowledgements

I would like to thank **Ryan Rosario** for his advice on this capstone project, especially for suggesting various techniques and patiently listening concern/issue. He is truly very knowledgeable and a great mentor.