Housing Price Prediction

by
Aniket Patole (20906426), Peyman Mohseni Kiasari (20919482)

Course: MSCI 623 - Big Data Analytics
University of Waterloo, Ontario, Canada, 2021

Abstract

In this project, we apply machine learning concepts on data collected for housing prices in the region of Melbourne, Australia to predict the selling price of a newly built home. With the help of information available of houses in the same vicinity such as suburb, number of rooms, property type, neighborhood, and various other variables, we develop a regression model to predict the optimal price, for which the newly built house can be sold for. The listings data is available on Kaggle which been scraped from Domain website for analysis, training and validating the model. Furthermore, we mine for patterns in the data with unsupervised machine learning approaches such as K-mean Clustering and supervised approaches in the form of Regression Analysis.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Business Problem

In the past few years there has been a steady rise in the real estate market in Australia. Along with most cities, Melbourne has several people who have decided to start building new houses in order to sell them.

The business problem that we are trying to solve benefits a person in a way such that, if they wish to build a house, it can estimate the cost of it for them. Along with that, it will also help in estimating what variables would be an important factor to include and what variable would be insignificant.

Assume that a person wants to build a new house in a specific suburb in Melbourne. As the property owner, they would want to know what would be the most optimal layout in the region in order to sell it at a better price. This model will highlight the level of importance of each variable such a the Distance from the Central Business District (CBD) or Number of Bedrooms required to name a few, which will aid them in creating the perfect house as per their budget and set a competitive price against multiple other landlords to sell their house.

This would be very beneficial to upcoming landlords because, looking for several characteristics simultaneously, such as, Suburb, Number of Rooms, Type of House, Cost of House, and Metropolitan to name a few, manually by looking for similar postings in their neighborhood and come up with a price and layout, would not only be an intricate task but also a time consuming one.

1.2 Motivation

There are multiple websites which contain enormous amounts of data about listings and their prices. This data could be used to develop a pricing model that would assist property-owners to set an optimal price. The input information required by the pricing model should

be easily available and understood by landlords. We hypothesize that there are two major factors that affect the price:

1. The characteristics of the house are a key factor in affecting the cost of it. This can range from the type of the house to the number of bedrooms in the house. To come up with the right feature of the house and the exact quantity of it to be used is extremely crucial.

2. The neighborhood of the listing. If the house is located in a prime vicinity with most facilities within a walk-able distance, then it is likely to be priced higher. Hence, to decide the distance of the house from important areas, such as CBD in our case is necessary.

To acquire neighborhood information, we can determine it with Google Maps, since the address of the listing is known. This also means that such a pricing model based on the above factors would be easy to use for seller. Moreover, this information for the existing listings can be obtained by scraping the web which would facilitate the model building process.

With the ever-growing rise in the real estate market due to the rise in demand of properties, It means that the number of home owners will increase with time. Our solution would benefit the upcoming property landlords to set the right price in the right location with the right features, in order to to remain competitive in the real estate market and being profitable at the same time.

1.3 Related Work

- The prediction of housing prices using linear regression dates back to as early as 1974 when Richardson et al. (1974) claimed that location and distance based variables play an important role along with house-specific characteristics in determining the price of a house [1].

- Jorge used multi-variate spatial methods kriging and cokriging to predict location price through continuous maps providing appraisers an overall view [2].

- Jakob proposed a model using economy-wide national income, mortgage repayment and nominal lending rates [3]. This model is used to predict house prices in OECD countries.

- Mitra and Suman proposed a model using 14 independent variables to predict house rent using artificial neural networks [4]. When compared with hedonic price models, the proposed model gave better results.

- Lynn and Erik provide a method to predict the housing market trends [5]. They even demonstrated how Google predicts future business activities.

Our data set and variables we use are different and challenging, this is where our project displays originality.

Chapter 2

Data

2.1 Data Collection

The data for this project has been obtained the source as described below:

Web Scraping

A listing posted by sellers on Domain includes information about the listing like price, location of the house, number of rooms, number of bathrooms, property type, parking space, size of the house, name of the seller and agent, and property features.

Using the location of the house we are able to obtain the latitudinal and longitudinal coordinates of the house. This can be done so using Google Maps.

A screenshot of the websites have been posted in the following page.

Kaggle Data Set

In order to acquire a solid data set for our analysis, we picked up the Melbourne Housing Market data set from Kaggle. The data that we downloaded and used for our model has been scraped from the Domain website.

Domain Group provides a listing of all the houses uploaded. This file had information about the listing (price, number of rooms, property type, Land size, latitude, longitude, description, etc.) and the seller . In total the dataset contains 34857 rows and 21 columns.

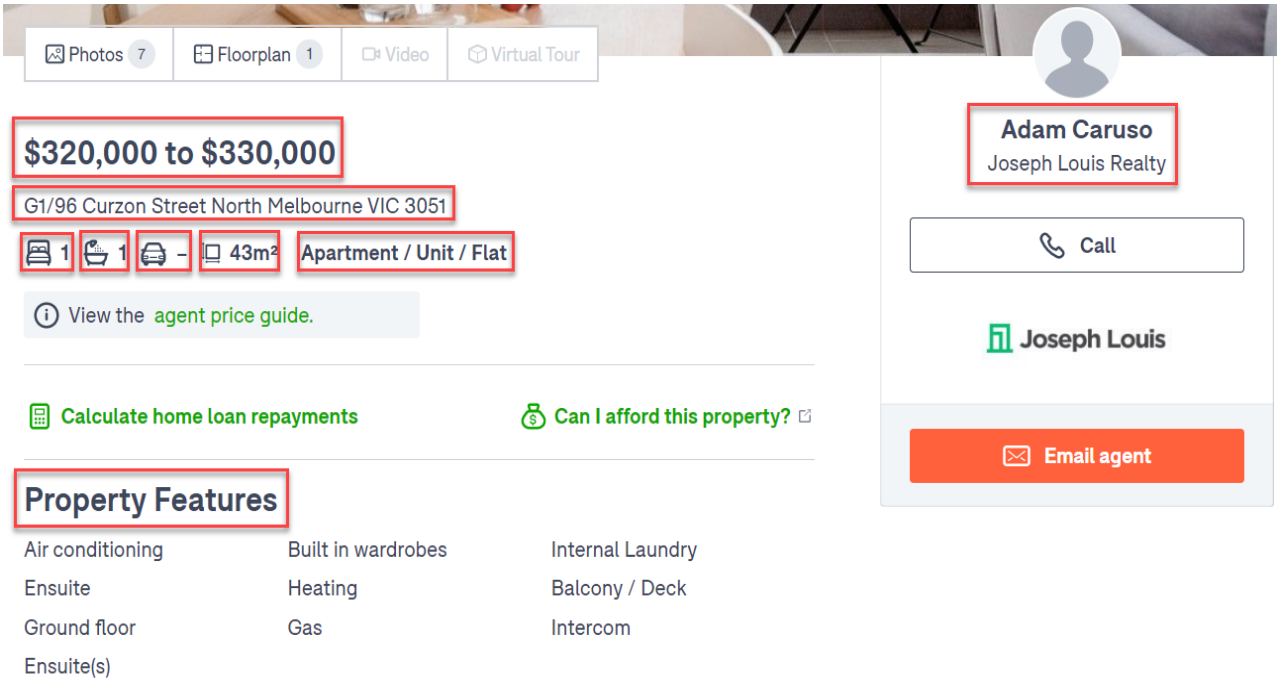


Figure 2.1: Features of the Listing

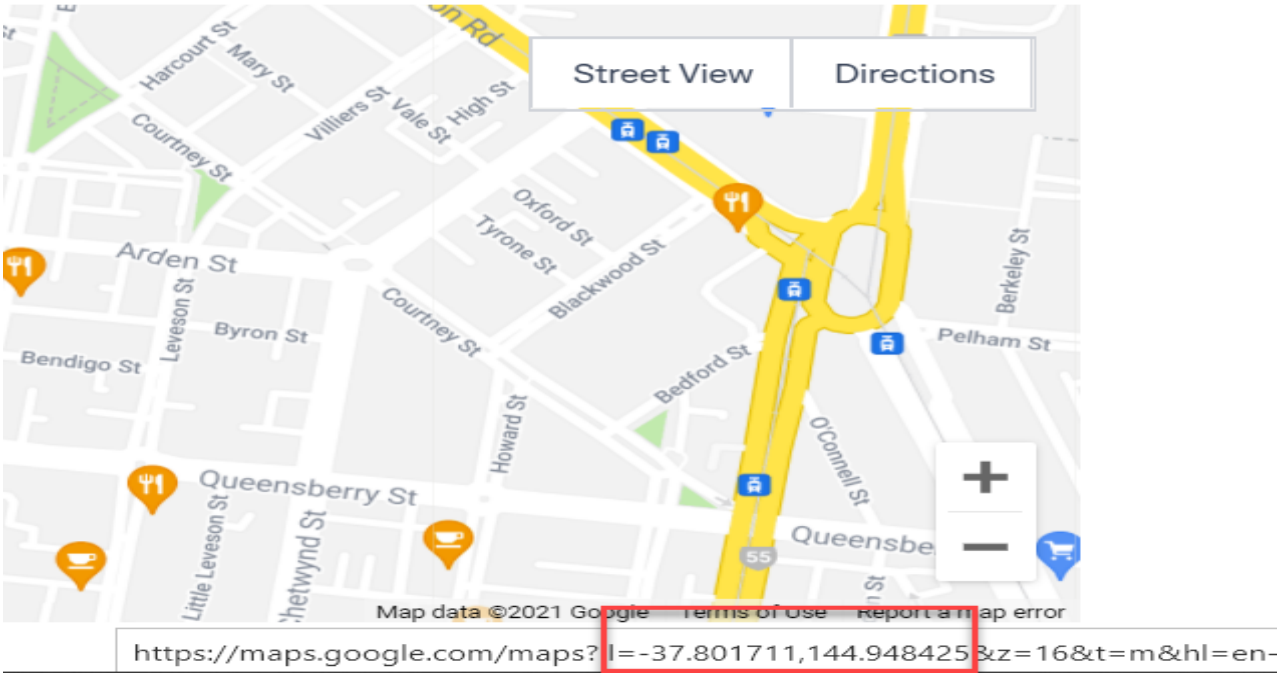


Figure 2.2: Location of the listing

2.2 Data Description

The data obtained from the sources discussed in the previous section can be grouped in 3 categories of explanatory variables. This section describes the variables that come under each category along with information about the data type of the variable (whether numeric or categorical when it was scraped. Some of these numeric variables were later converted to categorical for specific tasks, which would be discussed in the respective sections). Note that price is the target variable which is numeric for the regression analysis and is converted to binary for the classification task.

Listing Characteristics

1. property type (Categorical): House, Apartment, Condominium, Tent, etc.
2. bathrooms (Numeric): Number of bathrooms
3. bedrooms (Numeric): Number of bedrooms
4. parking (Numeric): Number of parking spots
5. land size (Numeric): size of land in Meters
6. Seller (Categorical): Real Estate Agent
7. Distance (Numeric): Distance from CBD in Kilometers
8. Regionname (Categorical): General Region (West, North West, North, North east . . .etc.)
9. Propertycount (Numeric): Number of properties that exist in the suburb.
10. Landsize (Numeric): Land Size in Meters
11. BuildingArea (Numeric): Building Size in Meters
12. Latitude (Numeric): Self explanatory
13. Longitude (Numeric): Self explanatory

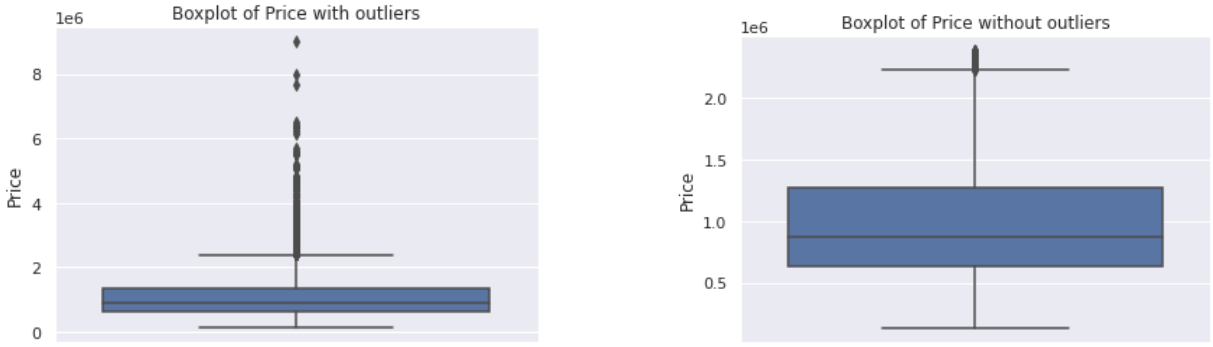
NOTE:

- The originally acquired dataset contains 34857 rows and 21 columns.
- The following columns were deleted because they were of no use in this analysis:
- **Postcode, Address, SellerG** (seller name)
- Columns **YearBuilt and BuildingArea** were both deleted, because they contained more than 60% Nans.
- After removing any row with a NaN, 17679 rows $\times$ 16 columns remained.

Chapter 3

Exploratory Data Analysis

Exploratory Data Analysis is an important first step which must be carried out before modeling the data. It can lead to some patterns, trends, or insights which we might be able to uncover either by simple descriptive statistics or visualizing the data in the form of different charts.



(a) Before Outlier Removal

(b) With Refined Data

Figure 3.1: Figure (a) and (b): Box Plot of Price

3.1 Dependent Variable

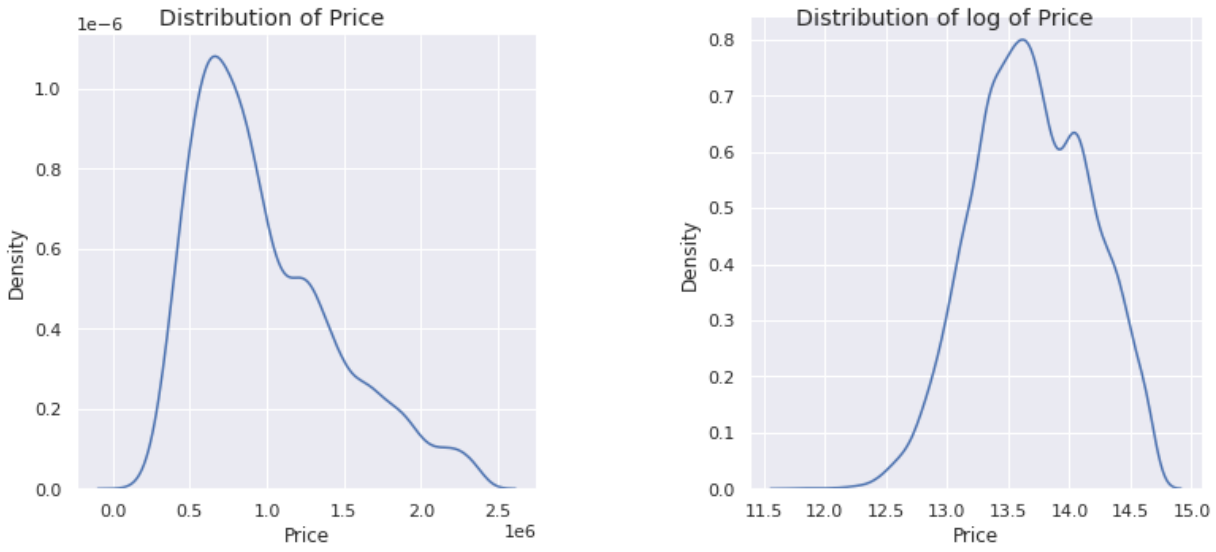
First, we focus on the dependent variable, which in our case is the price of the house. Using a box plot, we can detect the presence of large number of outliers. The maximum price was surprisingly around $9 Million for one listing, which was indeed a very luxurious home. Due to the presence of such data points, the value of mean would shift up by a considerable amount and affect the model fitting process. Hence, data had to be cleaned before proceeding further.

Data Cleaning

We implemented the common technique of outlier removal based on inter quartile range. This is done by first calculating the inter-quartile range which is the difference between the 75^{th} quantile and the 25^{th} quantile.

$$IQR = q_{0.75} - q_{0.25} \tag{3.1}$$

Any data point that has a value greater than 75^{th} quartile + 1.5 IQR or quartile 1.5 IQR. Based on this criterion, the refined dataset had 8779 records remaining with the price variable now having the maximum price as \$2.68 Million and the average as \$985 Thousand. The box plot of price in the refined data is illustrated in Figure.



(a) Distribution of Price

(b) Distribution of Log of Price

Figure 3.2: Figure (a) and (b): Price Distribution

Additionally, the probability density plot (on Fig. 3.2) shows normal with positive skew 0.87.

And log of price (on Fig. 3.2) will be normal with negative skew -0.054.

The following graph (on Fig. 3.3) shows that, from 2016 to late 2018, the cost of the properties did not change significantly during these two years. This is important because we do not need to delete any of the older data and will be able to utilize data available for all the years listed.

Figure 3.3: Mean Price per year

3.2 Explanatory Variables

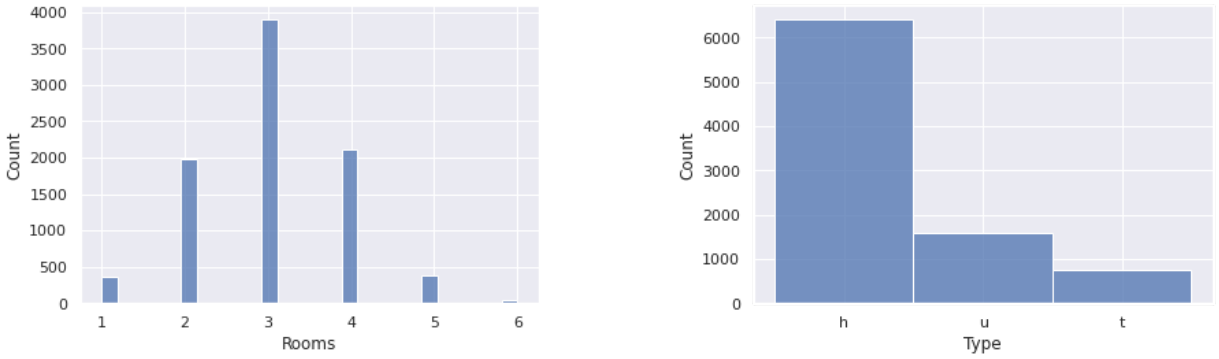
This section is categorized by the type of plots that were graphed for the different explanatory variables. Only the useful plots have been discussed in the report.

Bar Plots

The analysis of the bar plots in Figure 3.4 can be described as below:

(a) There are various amounts of rooms available in the listings posted, 3-bedroom houses are the most uploaded houses.

(b) Off all the types of listings which are: h - house, cottage, villa, semi, terrace; u - unit, duplex; t - townhouse; Type - h are the most common.

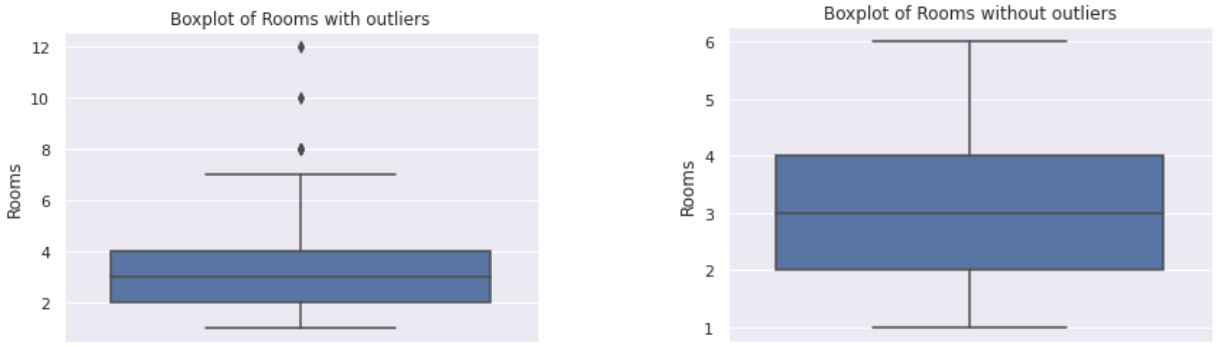


(a) Number of Rooms against Count

(b) Type of House against Count

Figure 3.4: Figure (a) and (b): Box Plots of Variables against Count

Box Plots



(a) Price by number of rooms w outliers

(b) Price by number of rooms w/o outliers

Figure 3.5: Box Plots

The two box plots in Figure 3.5 can be interpreted as follows:

There exists a few number of houses with more than 7 rooms, Since this number existed in very few of the houses, they would disrupt the prediction and hence were considered as outliers. There were only 11 outliers and consequently deleted.

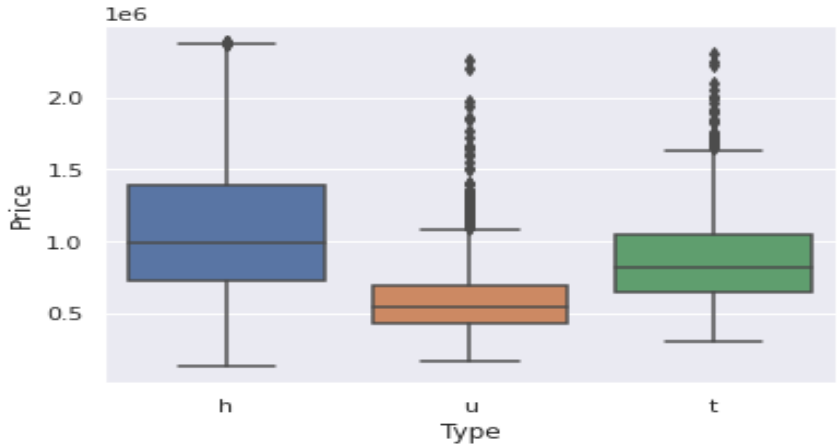


Figure 3.6: Box Plot of types of houses against the Price

As per, figure 3.6:

We have compared the price of each type of house available against each other. We infer that, the type - h which includes house, cottage, villa, semi, terrace, has the highest range of price. The type - u, which are unit, and duplex, have the lowest range.

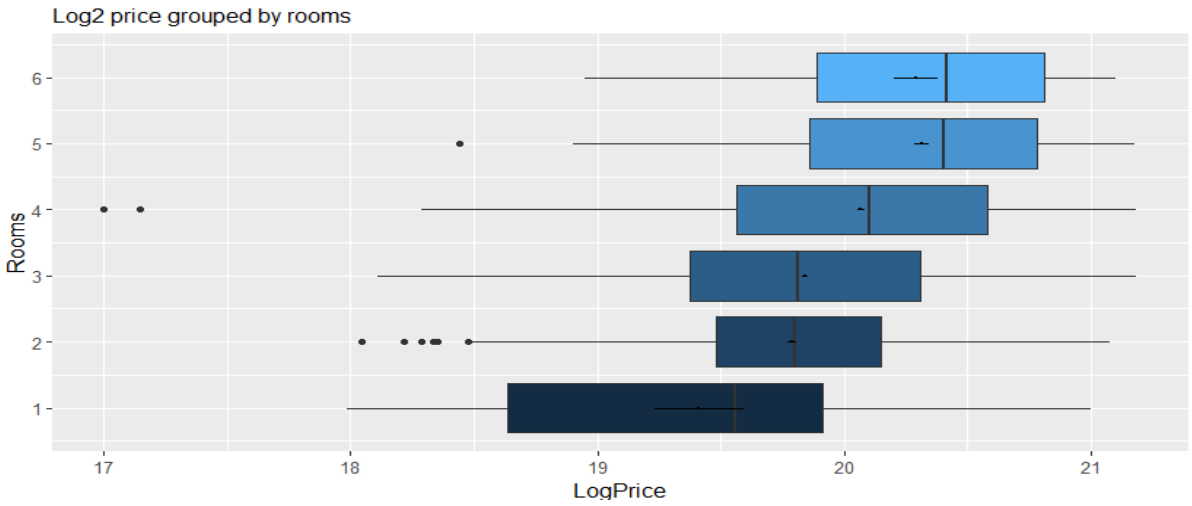


Figure 3.7: Box Plot of number of rooms against the Price

Figure 3.7, Demonstrates:

As the number of rooms vary the cost of the house also varies. We can see that as we increase the number of rooms in a house, the price range of the house also surges simultaneously.

Heat Map

Heatmap of prices over the density of each region as shown in figure 3.8. It illustrates that, price get higher as we approach the center of the town. This also exhibits that each region differs from others. Of all the regions present in Melbourne, we distinctly notice that Northern Victoria is very sparse.

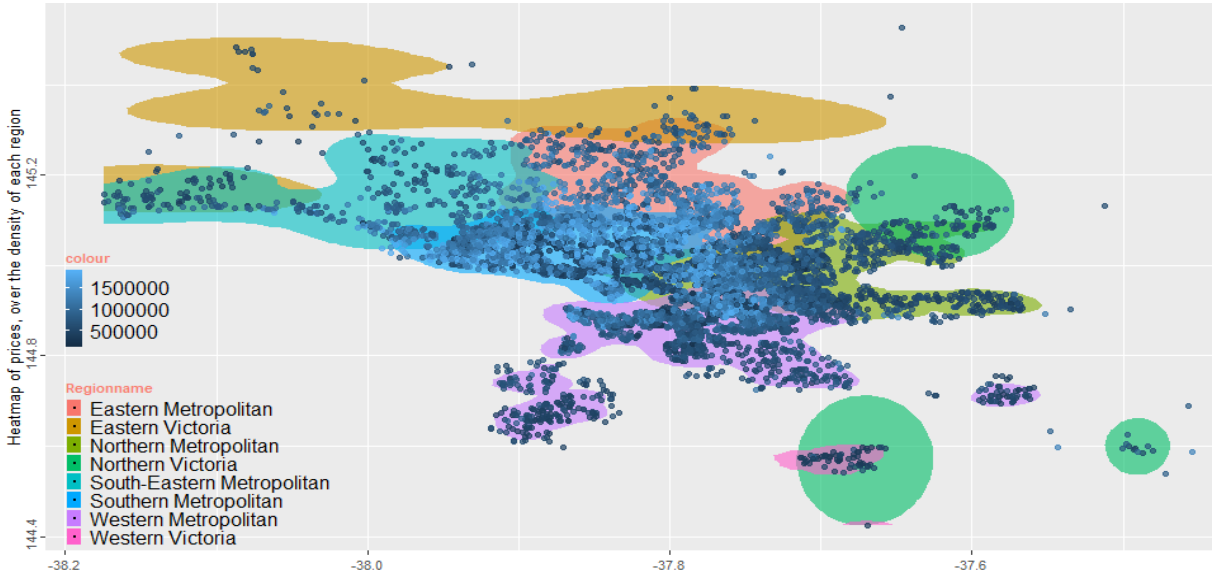


Figure 3.8: Heat Map of Prices over the density of each region

Violin Plot

This plot shown in figure 3.9 exemplifies that each region has a versatile price range. This means that the cost of housing in each area varies.

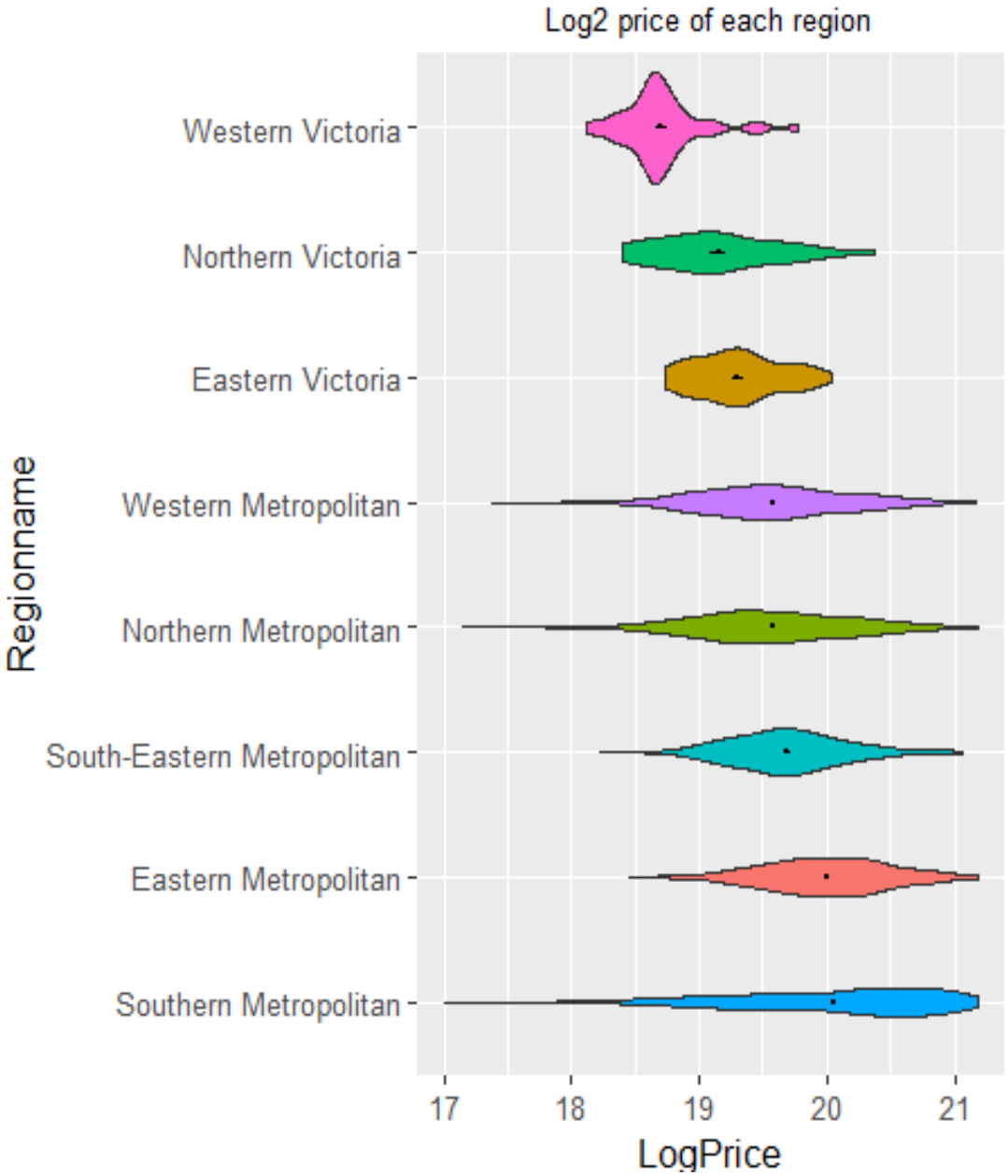


Figure 3.9: Violin Plot of Region over the log price

Hex Plot

The following Hex Plot displayed in Figure 3.10, represents how the cost of the housing increases as we approach the center of Melbourne also known as CBD.

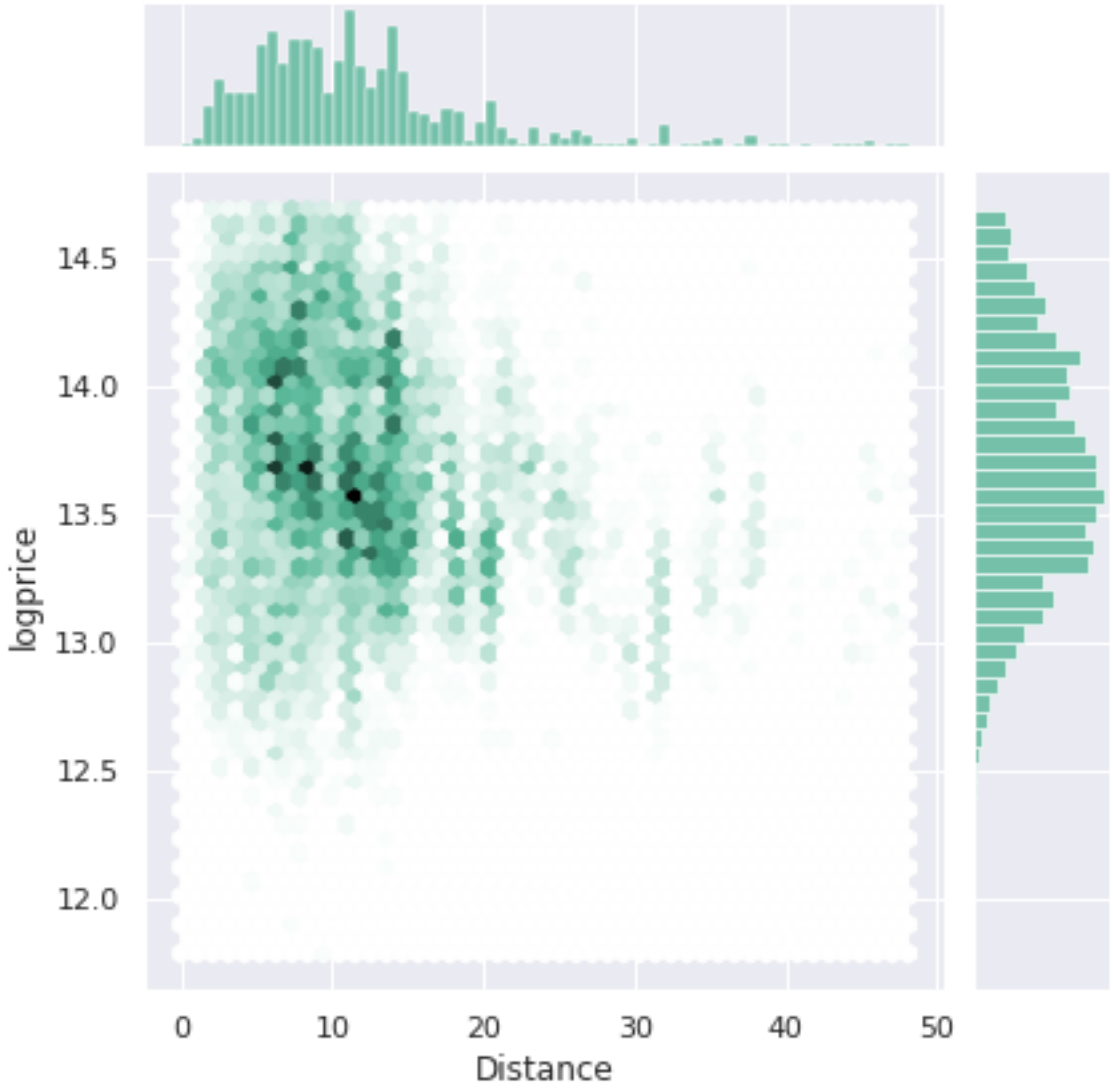


Figure 3.10: Hex Plot of Price against Distance from CBD

Chapter 4

Regression Analysis

This section describes the process of fitting a linear regression model to predict the listing price using the different groups of explanatory variables discussed in the previous section.

4.1 Data Pre-processing

Some of the variables in the original data had to be removed before being used for model fitting.

- Only type "h" housing values were considered for the model as it had the highest number of listings, this was done so because the housing of type "u" and "t" were very low in number and had to be removed as it would be in efficient.
- We also excluded the "Victorian" regions which includes the Western Victoria, Northern Victoria and Eastern Victoria Regions. This is because they are suburbs with very low population and again would lead to inefficiency of the model.
- Post cleaning of data we had 6231 listings remain.

We use **LogPrice** beacause, as shown in figure 3.2 when we compare the distribution of price against the distribution of log of price. It can be inferred that Distribution of Log of Price is Normally Distributed.

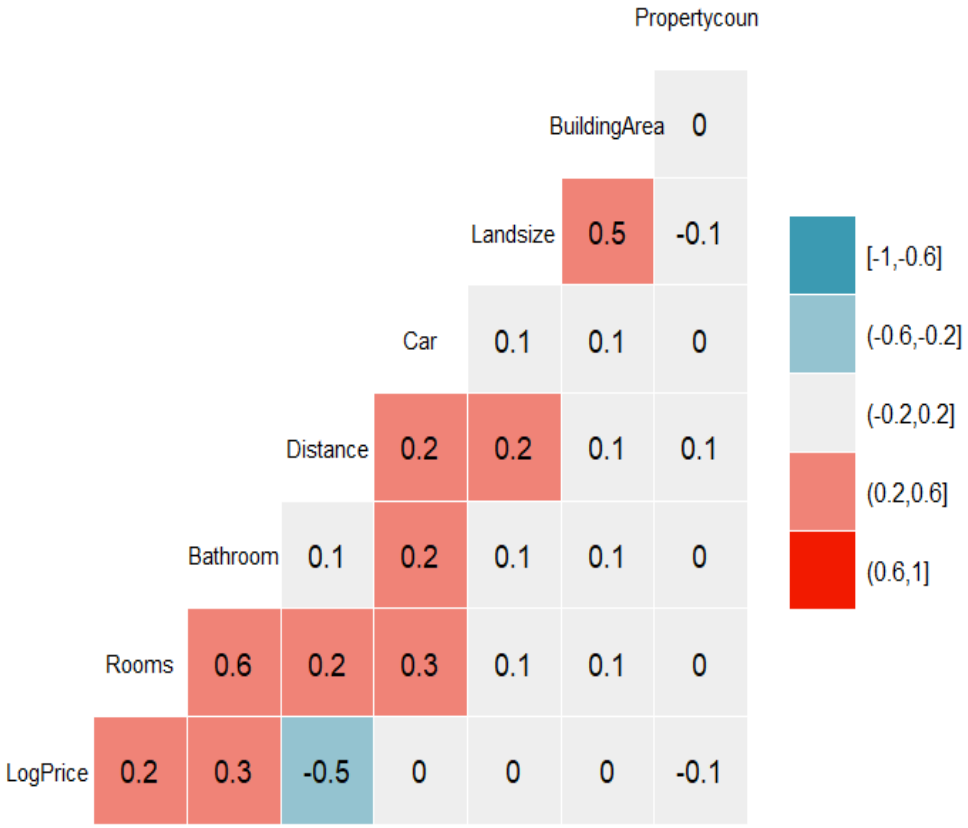


Figure 4.1: Correlation Chart of LogPrice

We plot a Correlation chart with **logPrice** as shown in fig. 4.1 to understand what variables must be selected in the model. As per the chart, the following variables should be selected: rooms, bathrooms, and distance alongside the 4 different regions.

Because room and bathroom are highly correlated to each other they both cannot be selected in one model. In order to check which of the two is a better parameter for prediction we must build 2 models to compare the same.

NOTE: None of these required variables need to be pre-processed.

4.2 Model Fitting

For fitting linear models, we make use of the Python library **statsmodels** instead of **sklearn**. This is because statsmodels outputs the model and co-efficient significance along with other evaluation metrics. Model fitting was done using ordinary least squares.

4.3 Models

T acquire the best model for this business problem, it needs to include the most significant features.

As seen in Figure 4.1, due to the high correlation between the variables rooms and bathrooms. Only either one can be used in a model at a time. Therefore, in order to find the most significant factor out of the two, we created 2 models: Model1 and Model2 which are shown in the Appendix A.

Model 1:

Model 1 is "logprice ∼ Rooms + Distance + Regionname" as shown in figures 8.1 and 8.2.

On Performing regression analysis, we infer that, Durbin-Watson test is slightly more than 1.5, which means that slight positive auto-correlation exists. R-squared value is 0.61 which means the model is explaining 61% of the data. All P_values are small, which means that choice of variables is correct. Coefficient of distance is low but when we remove it, we face a huge downfall in accuracy, therefore it is an important factor.

Model 2:

Model2 is "logprice ∼ Bathroom + Distance + Regionname" as shown in the figure 8.3.

This model is not better than model 1 because R-squared value is 0.58. which means that it explains lesser data as compared to model1.

For testing if our model is optimal, we built a regression model based on all of our variables which is "logprice ∼ Rooms + Distance + Regionname + Car + Landsize + BuildingArea + Propertycount + Latitude + Longitude" as seen in the following figure 8.5. The R_s is 0.65 and higher than our model1. This made us notice that we were missing something in our model. We realised that, the reason for this value being higher is that all coefficients are very low except Latitude and Longitude. Hence we infer that, we were right in overseeing those variables but made a mistake in not considering Latitude and Longitude, this is because we thought distance would be enough. In order to rectify this, we built a new model which includes latitude and longitude called Model3.

Model 3:

Model 3 is formulated as "logprice ∼ Rooms + Distance + Regionname + Latitude + Longitude" displayed in the figure 8.4. R_squared is 0.64 [which is close to 0.65 which we

found in the test model having all the variables, proving that other coefficients were not important].

From the above model, we realised that the P-value for 2 regions: South-Eastern Metropolitan and Western Metropolitan is greater than 0.1, which is not acceptable and hence have to be removed.

But since our primary goal is to build a model which to predict for all regions. We tackle this challenge by creating 2 models. One model contains the values of the Northern Metropolitan and Southern Metropolitan while the other contains the value of the regions which had a higher P-value in Model3 as seen in figures 4.2 and 8.6 in the Figure 4.2.

Model 4:

Model 4 Inference

As per the data analysis we make the following inferences:

- All the P-Values of this model are 0, it means that the choice of the variables are correct.
- The R-squared value is 0.659, which means that this model is the most accurate compared to the rest. It describes most of the 65% of the data making it the most optimal model we have achieved.
- The Skewness Value is -0.670, which falls below -0.5, we conclude that the model is moderately skewed, unlike the other models which are symmetrical.
- The Durbin - Watson value is 1.616 which is below 2, which indicates positive auto-correlation.
- The F-statistic is 1348, which is larger than the F-stat at 5% aplha. This means that the overall model is significant.

Due to the optimal features of this model, we consider it as our primary model. From the figure 4.2 we formulate our model4 as follows:

logprice = - 156.54 + 0.180 * (Rooms) - 0.0425 * (Distance) - 0.9469 * (Latitude) + 0.9277 * (Longitude) - 0.1171 * (Regionname: T. Northern Metropolitan) + 0.1475 * (Regionname: T. Southern Metropolitan)

```
Intercept                            -156.542590
Regionname[T.Northern Metropolitan]    -0.117138
Regionname[T.Southern Metropolitan]     0.147528
Rooms                                   0.180533
Distance                               -0.042560
Lattitude                              -0.946969
Longtitude                              0.927760
dtype: float64
                            OLS Regression Results
==============================================================================
Dep. Variable:               logprice   R-squared:                       0.659
Model:                            OLS   Adj. R-squared:                  0.659
Method:                 Least Squares   F-statistic:                     1348.
Date:                Wed, 04 Aug 2021   Prob (F-statistic):               0.00
Time:                        03:58:11   Log-Likelihood:                -6.0197
No. Observations:                4190   AIC:                             26.04
Df Residuals:                    4183   BIC:                             70.42
Df Model:                           6
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                      -156.5426     11.042    -14.177      0.000    -178.191    -134.894
Regionname[T.Northern Metropolitan]    -0.1171      0.016     -7.482      0.000      -0.148      -0.086
Regionname[T.Southern Metropolitan]     0.1475      0.015     10.085      0.000       0.119       0.176
Rooms                                   0.1805      0.005     37.240      0.000       0.171       0.190
Distance                               -0.0426      0.001    -46.321      0.000      -0.044      -0.041
Lattitude                              -0.9470      0.079    -11.973      0.000      -1.102      -0.792
Longtitude                              0.9278      0.081     11.497      0.000       0.770       1.086
==============================================================================
Omnibus:                      720.971   Durbin-Watson:                   1.616
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4503.731
Skew:                          -0.670   Prob(JB):                         0.00
Kurtosis:                       7.899   Cond. No.                     4.43e+05
==============================================================================
```

Figure 4.2: Model4: With Northern Metropolitan and Southern Metropolitan

Cross Validation

We test using cross-validation, in this we remove each variable from the model one at a time and test it as a new model. We notice that, each time the R_squared will decrease significantly (especially with the first 3 variables). Hence this illustrates that, our model is the best model and tested. The R_squared of removing each is listed below in figure 4.3:

Variable Removed	Model R-Squared Value
Rooms	**0.53**
Distance	**0.40**
Latitude	**0.59**
Longitude	**0.63**
Regionname (All regions)	**0.61**

Figure 4.3: R-Squared Values

Hence we can infer that the presence of distance is very important in an optimal model as the removal of it from our model provides the least R-Squared Value.

4.4 Residual Plot

To ensure that the linear regression assumptions of constant variance and normally distributed residuals hold, we analyze the residual plot.

The residual histogram as seen in the figure 4.4 is approximately normal distributed. The graph and the the mean value which is approximately equal to zero can be used to conclude that there is no violation.

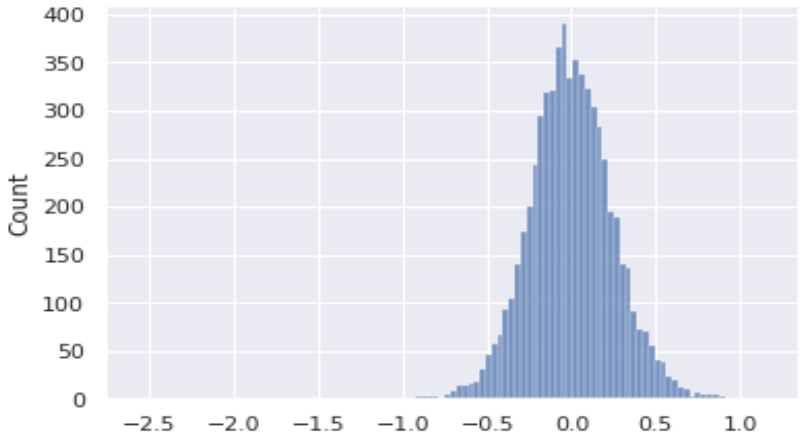


Figure 4.4: Residual Histogram of Model 3

Chapter 5

Clustering Analysis

In clustering analysis, the input we chose are those variables which were the most effective variables we discovered in regression analysis. The variables are as follows: 'Price', 'Rooms', 'Distance', 'Regionname'. The reason we did not include 'Latitude' and Longitude' for clustering is because the r-squared value with them in the model was 0.64 and without them was 0.61, which is not that significant and hence its use lacks effectiveness.

Our challenge is that, we have three kinds of variables. These are as follows: Range - price, ordered - rooms and categorical - region. It has been stated that "The k-Means algorithm is not applicable to categorical data, as categorical variables are discrete and do not have any natural origin. So computing euclidean distance for such as space is not meaningful." as mentioned in Clustering Algorithm for Data. [6]

We tackle this issue by using k-Prototype for categorical variables, it is an extension of K-means algorithm introduced in the paper "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values" [7]

In case of rooms which is an ordered variable, we should scale all of our number variables to the same range which is said in the paper "A conceptual version of the K-means algorithm" [8]

5.1 Data Preprocessing

The preprocessing required before clustering is data normalization. This is necessary because different numeric variables have different ranges. We use the Euclidean distance metric and therefore by normalizing we ensure that a unit difference in variable A has the same effect as that of a unit difference in another variable. For this purpose, we apply min-max normalization for each column (X) in the data, as depicted by Equation below where, Xmin is the minimum value in the column and Xmax is the maximum value in the column.

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

For our model we used Min-Max Normalization to normalize 'Price', 'Rooms', 'Distance' variables. Post normalizing the 3 variables we cluster all of them using k-Prototype.

5.2 k-Prototype Clustering

Figure 5.1 shows that steepest slope is from k = 1 to k = 2. we infer from the plot that, K = 2 or K = 3 are good for clustering.

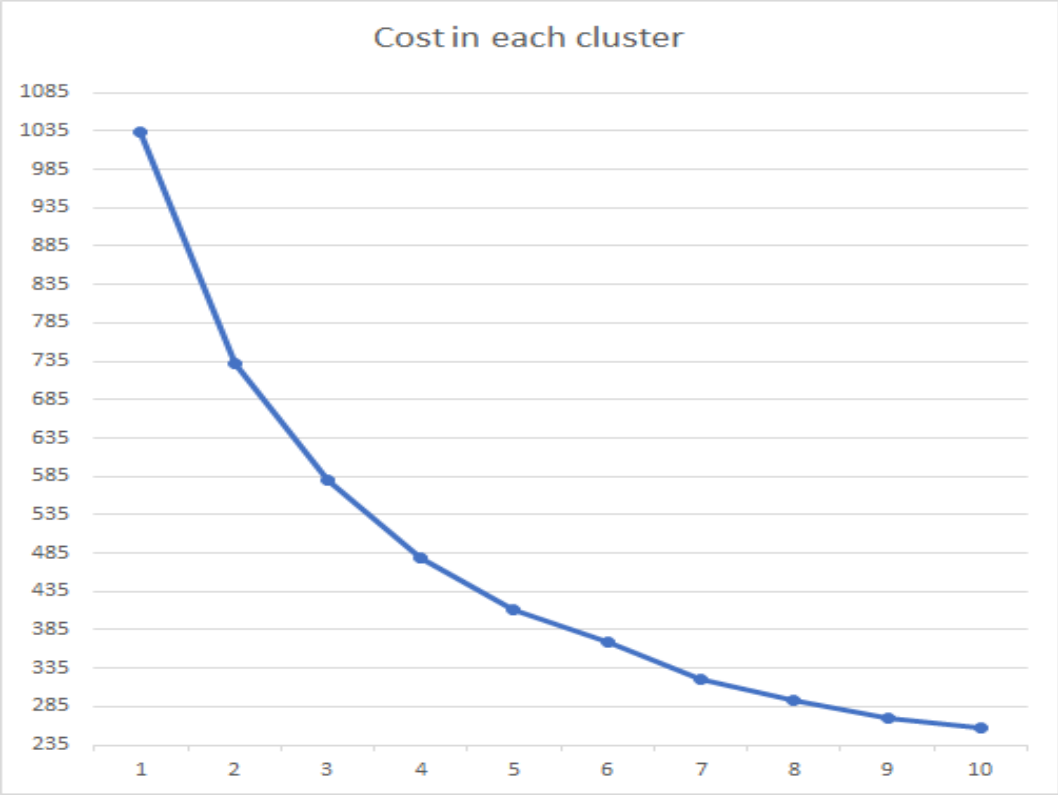


Figure 5.1: Slope of Clustering

5.3 Insights

In our analysis we wished to perform the Principal Component Analysis to chart the data, but due to the presence of categorical variables, we could not apply it. But in place of that we used maps of Latitude and Longitude and created clustering using values of K = 2 and K = 3.

Clustering with value of K = 2 and K = 3 5.2 are depicted. are in line with our original hypothesis that being close to center is the most important factor.

Looking at the figures 5.2, we infer that it proves our hypothesis right. It confirms that as the distance gets closer to the Central Business District (CBD) the cost of the house increases.

(a) Clustering with K = 2

(b) Clustering with K = 3

Figure 5.2: Figure (a) and (b): Clustering with K = 2 and K = 3

The cluster centers of each variables are shown in the figure below:

Variable	Cluster 1 (K = 2)	Cluster 2 (K = 3)
Price	0.3072771690414418	0.646319502262362
Rooms	0.4340143003064206	0.5114470842332589
Distance	0.33593960917690996	0.21435255206831322
Regionname	Northern Metropolitan	Southern Metropolitan

Figure 5.3: Cluster Centers for Cluster 1 and Cluster 2

Chapter 6

Conclusion

The main objective of this study was to develop a price prediction model for new house builders to assist them in setting a price for their houses based on the amenities and facilities available. The model was build using linear regression. We selected the features with analysis and recursive feature elimination. K-Prototype clustering analysis, which is an extension to the K-means algorithm, was carried out to identify the presence of groups of similar records in the data. The result was the discovery of the presence of 2 distinct clusters in the dataset - one corresponding to houses close to the central business district and the other far away.

Our analysis strongly shows that the most significant factor for house price is its distance from the central business district. And in the amenities of a house, the number of rooms is the most important. If someone wants to build a house, more rooms will benefit him/her more than the building area or parking spots. Also, if a person wants to upgrade their house, the best option will be to add a new room.

For the price prediction model, this could be developed into an application. It can be used by users who want to build a house and want to predict the price. They would just be required to input information about amenities and the location. This part of the project would be considered for future work.

For future works, we also suggest gathering more explanatory variables. More variables can build a more robust model and improve the analytic power of the model.

Chapter 7

References

1. Harry W Richardson, Joan Vipond, and Robert A Furbey. Determinants of urban house prices. Urban Studies, 11(2):189–199, 1974.

2. Jorge Chica-Olmo. Prediction of housing location price by a multivariate spatial method: cokriging. Journal of Real Estate Research, 2009.

3. Jakob B Madsen. A behavioral model of house prices. Journal of Economic Behavior Organization, 82(1):21–38, 2012.

4. Suman Kumar Mitra. Applicability of artificial neural network in predicting house rent, 2008.

5. Lynn Wu and Erik Brynjolfsson. The future of prediction: How google searches foreshadow housing prices and sales. In Economic analysis of the digital economy, pages 89–118. University of Chicago Press, 2014.

6. Satyam Kumar. Clustering Algorithm for data with mixed Categorical and Numerical features. 2021.

7. Huang, Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2, 283–304 (1998). https://doi.org/10.1023/A:1009769707641

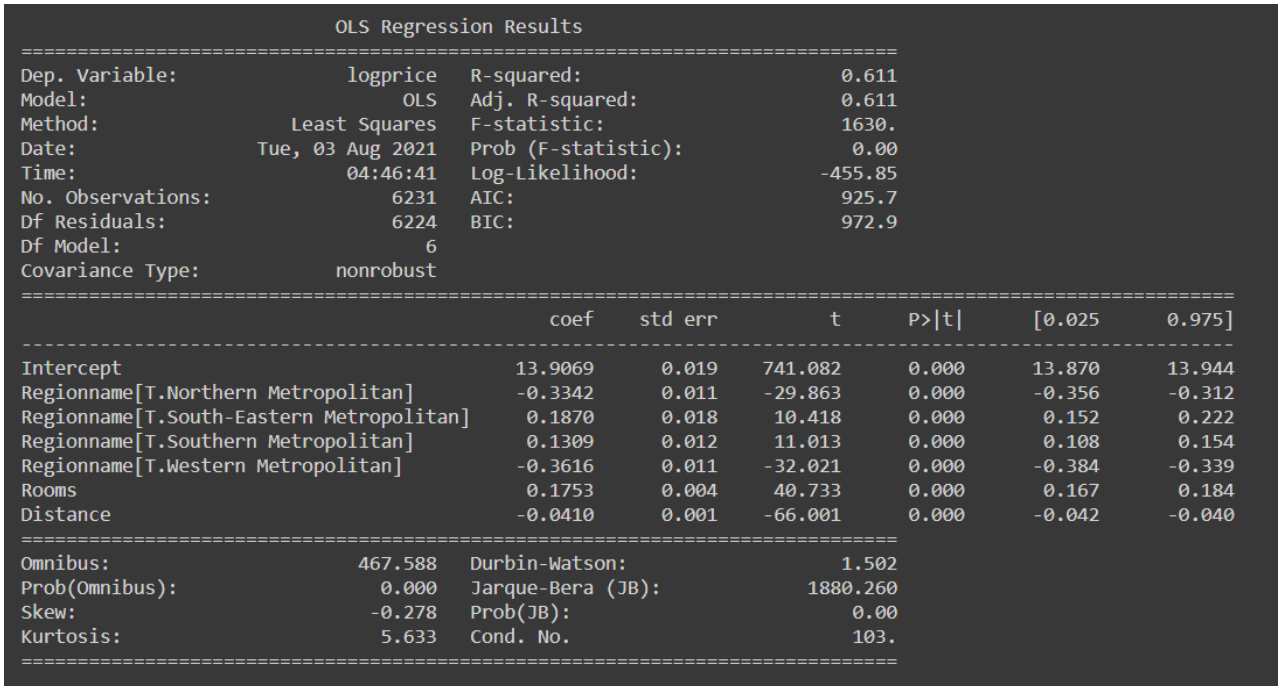
8. H. Ralambondrainy, A conceptual version of the K-means algorithm, Pattern Recognition Letters, Volume 16, Issue 11, 1995, Pages 1147-1157, ISSN 0167-8655.

Chapter 8

Appendix

Intercept	13.906861
Regionname[T.Northern Metropolitan]	-0.334174
Regionname[T.South-Eastern Metropolitan]	0.186979
Regionname[T.Southern Metropolitan]	0.130886
Regionname[T.Western Metropolitan]	-0.361584
Rooms	0.175271
Distance	-0.041049

Figure 8.1: Model1



OLS Regression Results

Dep. Variable:	logprice	R-squared:	0.611
Model:	OLS	Adj. R-squared:	0.611
Method:	Least Squares	F-statistic:	1630.
Date:	Tue, 03 Aug 2021	Prob (F-statistic):	0.00
Time:	04:46:41	Log-Likelihood:	-455.85
No. Observations:	6231	AIC:	925.7
Df Residuals:	6224	BIC:	972.9
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	13.9069	0.019	741.082	0.000	13.870	13.944
Regionname[T.Northern Metropolitan]	-0.3342	0.011	-29.863	0.000	-0.356	-0.312
Regionname[T.South-Eastern Metropolitan]	0.1870	0.018	10.418	0.000	0.152	0.222
Regionname[T.Southern Metropolitan]	0.1309	0.012	11.013	0.000	0.108	0.154
Regionname[T.Western Metropolitan]	-0.3616	0.011	-32.021	0.000	-0.384	-0.339
Rooms	0.1753	0.004	40.733	0.000	0.167	0.184
Distance	-0.0410	0.001	-66.001	0.000	-0.042	-0.040

Omnibus:	467.588	Durbin-Watson:	1.502
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1880.260
Skew:	-0.278	Prob(JB):	0.00
Kurtosis:	5.633	Cond. No.	103.

Figure 8.2: Model1

```
Intercept                                14.168119
Regionname[T.Northern Metropolitan]      -0.338969
Regionname[T.South-Eastern Metropolitan]  0.176224
Regionname[T.Southern Metropolitan]       0.112232
Regionname[T.Western Metropolitan]       -0.364671
Bathroom                                  0.180720
Distance                                 -0.038810
dtype: float64
                            OLS Regression Results
==============================================================================
Dep. Variable:               logprice   R-squared:                       0.584
Model:                            OLS   Adj. R-squared:                  0.584
Method:                 Least Squares   F-statistic:                     1459.
Date:                Tue, 03 Aug 2021   Prob (F-statistic):               0.00
Time:                        05:12:02   Log-Likelihood:                -662.37
No. Observations:                6231   AIC:                             1339.
Df Residuals:                    6224   BIC:                             1386.
Df Model:                           6
Covariance Type:            nonrobust
===========================================================================================================
                                          coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------
Intercept                                14.1681      0.016    891.752      0.000      14.137      14.199
Regionname[T.Northern Metropolitan]      -0.3390      0.012    -29.277      0.000      -0.362      -0.316
Regionname[T.South-Eastern Metropolitan]  0.1762      0.019      9.501      0.000       0.140       0.213
Regionname[T.Southern Metropolitan]       0.1122      0.012      9.146      0.000       0.088       0.136
Regionname[T.Western Metropolitan]       -0.3647      0.012    -31.233      0.000      -0.388      -0.342
Bathroom                                  0.1807      0.005     33.965      0.000       0.170       0.191
Distance                                 -0.0388      0.001    -61.075      0.000      -0.040      -0.038
==============================================================================
Omnibus:                      351.307   Durbin-Watson:                   1.536
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              986.108
Skew:                          -0.287   Prob(JB):                    7.40e-215
Kurtosis:                       4.862   Cond. No.                         96.3
==============================================================================
```

Figure 8.3: Model2

```
Intercept                                -154.609162
Regionname[T.Northern Metropolitan]        -0.116050
Regionname[T.South-Eastern Metropolitan]    0.030404
Regionname[T.Southern Metropolitan]         0.196700
Regionname[T.Western Metropolitan]         -0.041213
Rooms                                       0.174776
Distance                                   -0.039153
Lattitude                                  -0.628065
Longtitude                                  0.997266
dtype: float64
                            OLS Regression Results
==============================================================================
Dep. Variable:               logprice   R-squared:                       0.640
Model:                            OLS   Adj. R-squared:                  0.639
Method:                 Least Squares   F-statistic:                     1381.
Date:                Tue, 03 Aug 2021   Prob (F-statistic):               0.00
Time:                        05:31:04   Log-Likelihood:                -217.55
No. Observations:                6231   AIC:                             453.1
Df Residuals:                    6222   BIC:                             513.7
Df Model:                           8
Covariance Type:            nonrobust
================================================================================================
                                         coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                           -154.6092      7.900    -19.572      0.000    -170.095    -139.123
Regionname[T.Northern Metropolitan]   -0.1160      0.015     -7.961      0.000      -0.145      -0.087
Regionname[T.South-Eastern Metropolitan]    0.0304      0.022      1.393      0.164      -0.012       0.073
Regionname[T.Southern Metropolitan]    0.1967      0.013     14.835      0.000       0.171       0.223
Regionname[T.Western Metropolitan]    -0.0412      0.020     -2.110      0.035      -0.079      -0.003
Rooms                                  0.1748      0.004     42.073      0.000       0.167       0.183
Distance                              -0.0392      0.001    -63.622      0.000      -0.040      -0.038
Lattitude                             -0.6281      0.056    -11.313      0.000      -0.737      -0.519
Longtitude                             0.9973      0.053     18.892      0.000       0.894       1.101
==============================================================================
Omnibus:                      490.283   Durbin-Watson:                   1.569
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2393.713
Skew:                          -0.212   Prob(JB):                         0.00
Kurtosis:                       6.007   Cond. No.                     3.74e+05
==============================================================================
```

Figure 8.4: Model3

```
Intercept                                -153.552027
Regionname[T.Northern Metropolitan]        -0.098625
Regionname[T.South-Eastern Metropolitan]    0.046774
Regionname[T.Southern Metropolitan]         0.197939
Regionname[T.Western Metropolitan]         -0.032439
Rooms                                       0.143708
Distance                                   -0.041277
Car                                         0.015434
Landsize                                    0.000107
BuildingArea                                0.000352
Propertycount                               0.000002
Lattitude                                  -0.672763
Longtitude                                  0.978099
dtype: float64
                            OLS Regression Results
==============================================================================
Dep. Variable:               logprice   R-squared:                       0.658
Model:                            OLS   Adj. R-squared:                  0.657
Method:                 Least Squares   F-statistic:                     995.3
Date:                Tue, 03 Aug 2021   Prob (F-statistic):               0.00
Time:                        05:22:11   Log-Likelihood:                -58.792
No. Observations:                6231   AIC:                             143.6
Df Residuals:                    6218   BIC:                             231.2
Df Model:                          12
Covariance Type:            nonrobust
================================================================================================
                                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                         -153.5520      7.717    -19.899      0.000    -168.679    -138.425
Regionname[T.Northern Metropolitan]   -0.0986      0.015     -6.793      0.000      -0.127      -0.070
Regionname[T.South-Eastern Metropolitan]    0.0468      0.021      2.186      0.029       0.005       0.089
Regionname[T.Southern Metropolitan]    0.1979      0.013     15.285      0.000       0.173       0.223
Regionname[T.Western Metropolitan]   -0.0324      0.019     -1.698      0.089      -0.070       0.005
Rooms                                 0.1437      0.004     32.406      0.000       0.135       0.152
Distance                             -0.0413      0.001    -65.401      0.000      -0.043      -0.040
Car                                   0.0154      0.003      4.663      0.000       0.009       0.022
Landsize                              0.0001   1.03e-05     10.455      0.000    8.71e-05       0.000
BuildingArea                          0.0004   2.89e-05     12.207      0.000       0.000       0.000
Propertycount                      1.56e-06   7.57e-07      2.061      0.039    7.64e-08    3.04e-06
Lattitude                            -0.6728      0.055    -12.265      0.000      -0.780      -0.565
Longtitude                            0.9781      0.052     18.899      0.000       0.877       1.080
==============================================================================
Omnibus:                      812.062   Durbin-Watson:                   1.536
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5471.681
Skew:                          -0.429   Prob(JB):                         0.00
Kurtosis:                       7.510   Cond. No.                     2.13e+07
==============================================================================
```

Figure 8.5: Test Model

```
Intercept                          -245.997295
Regionname[T.Western Metropolitan]    0.240813
Rooms                                 0.176104
Distance                             -0.026406
Lattitude                            -0.190025
Longtitude                            1.739517
dtype: float64
                            OLS Regression Results
==============================================================================
Dep. Variable:               logprice   R-squared:                       0.530
Model:                            OLS   Adj. R-squared:                  0.529
Method:                 Least Squares   F-statistic:                     458.8
Date:                Wed, 04 Aug 2021   Prob (F-statistic):               0.00
Time:                        04:05:16   Log-Likelihood:                -94.326
No. Observations:                2041   AIC:                             200.7
Df Residuals:                    2035   BIC:                             234.4
Df Model:                           5
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                      -245.9973     15.302    -16.076      0.000    -276.007    -215.987
Regionname[T.Western Metropolitan]    0.2408      0.052      4.594      0.000       0.138       0.344
Rooms                          0.1761      0.008     23.188      0.000       0.161       0.191
Distance                      -0.0264      0.001    -22.057      0.000      -0.029      -0.024
Lattitude                     -0.1900      0.085     -2.234      0.026      -0.357      -0.023
Longtitude                     1.7395      0.099     17.567      0.000       1.545       1.934
==============================================================================
Omnibus:                       38.240   Durbin-Watson:                   1.536
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               44.636
Skew:                           0.274   Prob(JB):                     2.03e-10
Kurtosis:                       3.474   Cond. No.                     4.10e+05
==============================================================================
```

Figure 8.6: Model with Western Metropolitan