Melbourne House Price Prediction

Contents

1 Introduction

The focus of this data science project is to predict house prices using the Melbourne housing data set from Domain.com.au. As prospective buyers often find it challenging to determine the actual house price, a model that predicts the price based on a set of house features would be highly beneficial. This model could assist buyers in making informed decisions regarding whether a particular house is within their budget, whether it is worth the asking price, or what price to offer during negotiations.

The Melbourne housing data set includes information on houses sold in 2016 and 2017, and three different methods - k-nearest neighbors (knn), Linear Regression and Random Forest method - are utilized to predict house prices. The training data set is employed in each model to determine the Root Mean Squared Error (RMSE), with the optimal model selected based on the smallest cross-validation RMSE. Finally, the optimal model is used to estimate the RMSE on the validation data set.

This report is structured as follows: Section 1 outlines the analytical problem, Section 2 provides details on the data description and cleaning process, while Section 3 discusses data exploration and visualization. The methodologies and results are presented in Section 4, and the report concludes with a discussion of the limitations and potential for further analysis in Section 5.

2 Data Preparation

2.1 Data Description

Firstly, an overview of the data set is provided, which contains 13,580 observations and 21 variables, with one of the variables serving as the response variable, namely "Price". A comprehensive description of each variable is presented in Table 1.

Table 1: Data description

Variables	Description
Suburb	Name of houses' suburb
Address	Address of houses
Rooms	Number of rooms
Type	Houses' types: h - house, cottage, villa, semi, terrace; u - unit, duplex; t - townhouse
Price	Price in Dollars
Method	Methods used to sell houses
SellerG	Real Estate Agent
Date	Date sold
Distance	Distance from CBD
Postcode	Postcode address number
Bedroom2	Number of bedrooms (from different source)
Bathroom	Number of bathrooms
Car	Number of car spots
Landsize	Land size
BuildingArea	Building size
YearBuilt	Year built
CouncilArea	Governing council for the area
Regionname	General Region (West, North West, North, North east, etc)
Propertycount	Number of properties that exist in the suburb

2.2 Data Cleaning

The data set comprises 8 character variables and 13 numeric variables. Table 2 displays a summary of the character variables, while Table 3 presents the summary of numeric variables.

Table 2: Summary of character variables

No	Character_variables	Number_of_categories
1	Suburb	314
2	Address	13378
3	Type	3
4	Method	5
5	SellerG	268
6	Date	58
7	CouncilArea	34
8	Regionname	8

Table 3: Summary of numeric variables

No	Numeric_variables	Min	Median	Mean	Max	Number_of_NA
1	Rooms	1.00	3.0	2.94	10.00	0
2	Price	85000.00	903000.0	1075684.08	9000000.00	0
3	Distance	0.00	9.2	10.14	48.10	0
4	Postcode	3000.00	3084.0	3105.30	3977.00	0
5	Bedroom2	0.00	3.0	2.91	20.00	0
6	Bathroom	0.00	1.0	1.53	8.00	0
7	Car	0.00	2.0	1.61	10.00	62
8	Landsize	0.00	440.0	558.42	433014.00	0
9	BuildingArea	0.00	126.0	151.97	44515.00	6450
10	YearBuilt	1196.00	1970.0	1964.68	2018.00	5375
11	Lattitude	-38.18	-37.8	-37.81	-37.41	0
12	Longtitude	144.43	145.0	145.00	145.53	0
13	Propertycount	249.00	6555.0	7454.42	21650.00	0

The summary tables reveal that three variables - Car, BuildingArea, and Landsize - contain missing data. The N/A values for the Car variable are negligible (about 0.46%), and therefore, it is retained in the data set. However, the BuildingArea and Landsize variables have almost 50% missing values, and hence, they are dropped from the data set.

The missing values in the Car variable are dealt with using the median imputation method. This involves replacing the N/A values with the median value of the Car data.

Subsequently, we proceed by creating boxplots for several numeric variables to detect any outliers. Initially, we examine the boxplots for Price and Landsize variables, as displayed in Figure 1.

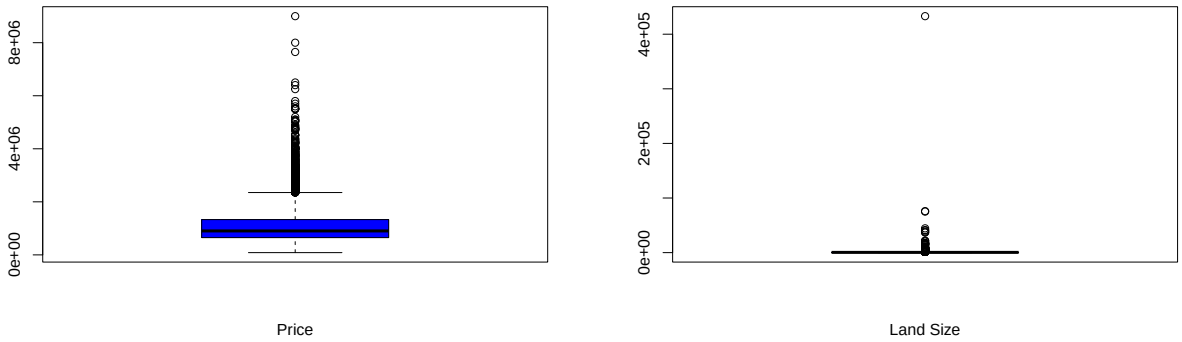

Figure 1: Boxplots of Price and Landsize

Based on the boxplot shown in Figure 1, we observe that there could be potential outliers in the highest values of Price variable, however, these values are not removed as they could still be reasonable. Conversely, around 14.3% of the data for Landsize variable equals to 0, which may indicate that this data was not provided. Therefore, we decide to exclude this variable from the prediction model.

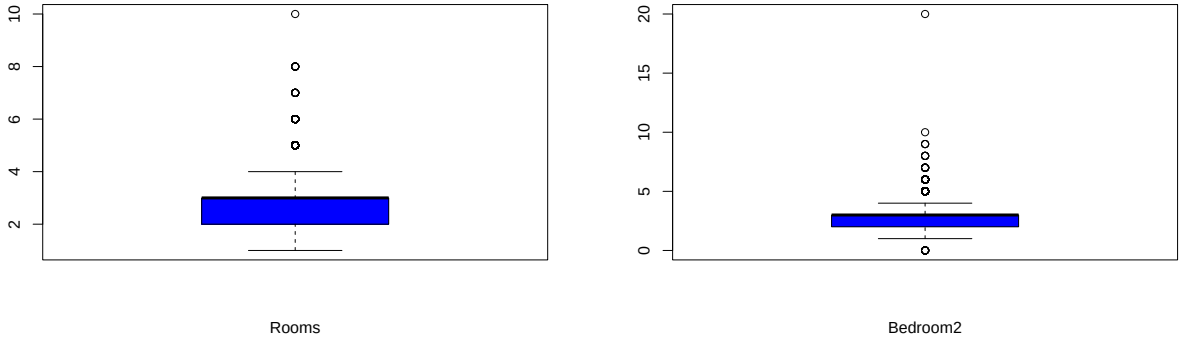

Figure 2: Boxplots of Rooms and Bedroom2

In Figure 2, we can see the boxplots of Rooms and Bedroom2 variables. It appears that there may be some outliers in the top end of both variables' data. To investigate further, we examine the highest values of these variables and present them in Table 4 and Table 5.

Table 4: Explore highest values of Rooms

Rooms	Type	Bedroom2	Bathroom	Car	Landsize
8	h	9	7	4	1472
8	u	4	2	4	983
8	h	8	4	4	638
8	h	6	2	4	663
8	h	6	4	3	668
8	h	8	3	3	614
8	h	8	8	4	650
10	h	10	3	2	313
8	h	8	3	1	1063

Table 5: Explore highest values of Bedroom2

Rooms	Type	Bedroom2	Bathroom	Car	Landsize
5	h	8	2	2	693
8	h	9	7	4	1472
8	h	8	4	4	638
4	h	9	8	7	1254
3	h	9	6	2	592
3	h	20	1	2	875
8	h	8	3	3	614
8	h	8	8	4	650
10	h	10	3	2	313
8	h	8	3	1	1063

Upon examining the highest values of Rooms and Bedroom2 variables, as shown in Table 4 and Table 5, we can see that some data points appear to be unreasonable, as the number of bedrooms exceeds the number of rooms. To address this issue, we remove all 203 data points with such discrepancies from the data set.

3 Data Exploration and Visualization

3.1 House Price Exploration

In this section, we delve deeper into the dataset and aim to visualize it wherever possible. As indicated in Table 2, the dataset consists of 13,378 unique houses across 314 suburbs. To initiate the exploration, we focus on the top 10 highest-priced houses in the dataset and their characteristics as outlined in Table 6.

Table 6: The top 10 highest price houses

Price	Suburb	Distance	Rooms	Type	Bedroom2	Bathroom	Car	Landsize
9000000	Mulgrave	18.8	3	h	3	1	1	744
8000000	Canterbury	9.0	5	h	5	5	4	2079
7650000	Hawthorn	5.3	4	h	4	2	4	1690
6500000	Kew	5.6	6	h	6	6	3	1334
6400000	Middle Park	3.0	5	h	5	2	1	553
6250000	Toorak	4.6	3	h	3	3	2	564
5800000	Brighton	11.2	5	h	5	4	4	1276
5700000	South Yarra	3.3	4	h	4	2	0	292
5600000	Middle Park	3.0	6	h	6	4	2	472
5525000	Armadale	6.3	6	h	5	3	4	1491

It is evident that the top 10 highest-priced houses possess distinct characteristics, as most of them have moderate values for each variable (such as land size, number of rooms, and proximity to the Central Business District). Additionally, it is noteworthy that all of these properties fall under the "house" category.

The lowest price houses are listed in Table 7. These 10 houses share similar characteristics, including having only 1 bedroom and 1 bathroom, small area, and no garage space. Additionally, eight of these houses are units.

Table 7: The top 10 lowest price houses

Price	Suburb	Distance	Rooms	Type	Bedroom2	Bathroom	Car	Landsize
200000	Kingsville	7.8	1	u	1	1	1	0
200000	Albion	13.9	1	u	1	1	1	1175
190000	Albion	13.9	2	u	2	1	1	0
185000	Albion	13.9	1	u	1	1	1	2347
185000	West Footscray	8.2	1	u	1	1	1	0
170000	Footscray	5.1	1	u	1	1	0	30
170000	Brunswick	5.2	1	u	1	1	0	1250
160000	Hawthorn	4.6	1	u	1	1	0	322
145000	Coburg	7.8	4	h	3	1	1	536
131000	Caulfield	8.9	4	h	4	1	2	499
85000	Footscray	6.4	1	u	1	1	0	0

Furthermore, we also analyze the house prices through visualization. The histogram of the house prices is presented in Figure 3.

From the histogram in Figure 3, we can observe that the right tail is longer than the left one, indicating that the distribution of house price is skewed right. This means that the mean value of Price is higher than the median value, likely due to the presence of some extremely high data points in the Price data. To address this issue, we transform the Price data into logarithmic form and create a new variable called Price_log. We then plot the histogram of Price_log, which is shown in Figure 4.

The histogram of the logarithm of the Price variable, as shown in Figure 4, displays a much more normal distribution than the original Price histogram. This suggests that the Price_log variable is more suitable for use in the prediction model than the original Price variable.

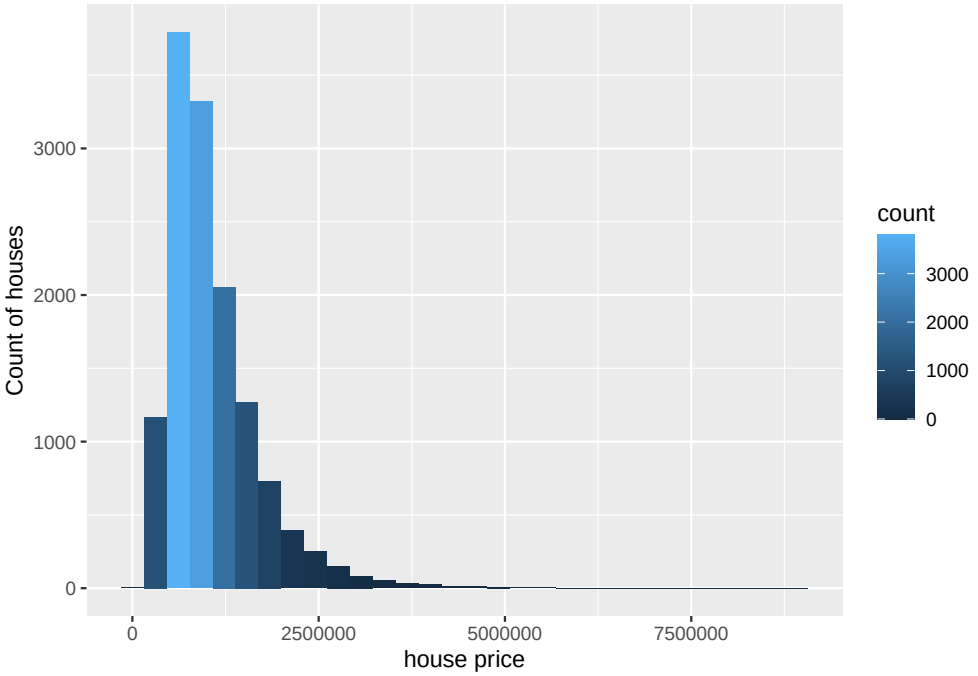

Figure 3: Histogram of Price

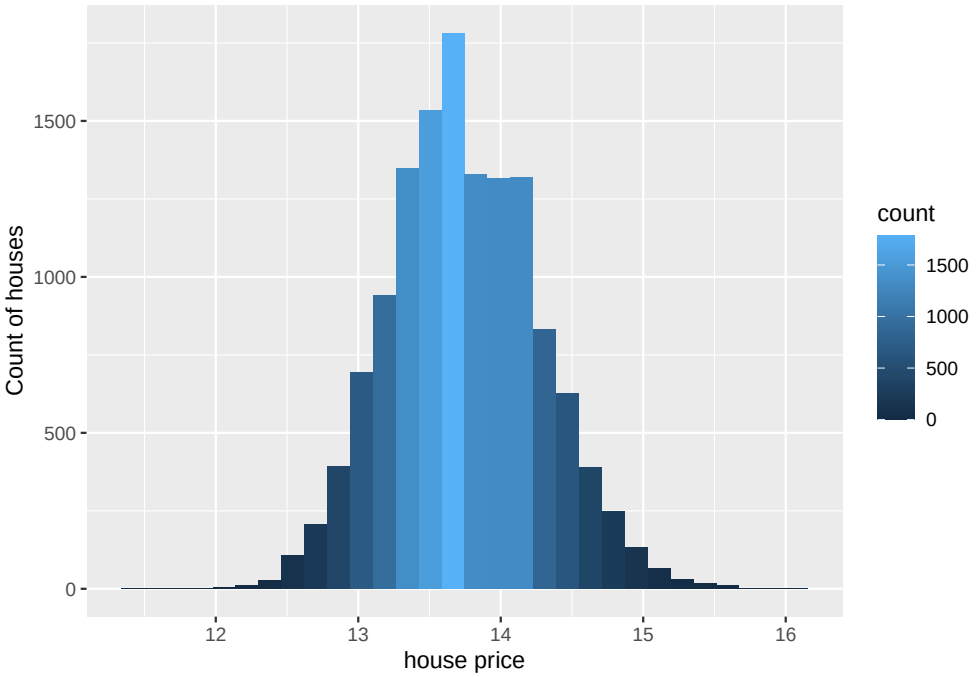

Figure 4: Histogram of log Price

3.2 House Price Exploration By Type

In this section, we explore the relationship between house price and house type. Firstly, we generate Table 8, which shows the total number of houses, average price, minimum and maximum price for each house type.

Table 8: House price and type

Type	Total	Max_Price	Min_Price	Average_Price
h	9301	9000000	131000	1240572.2
t	1105	3475000	300000	935035.2
u	2971	2460000	85000	603738.9

As presented in Table 8, the majority of houses in the dataset are categorized as "house", comprising around 66.77% of the total. Meanwhile, "unit" and "townhouse" represent the second and third most frequent house types, with 22.83% and 7.52% of the total, respectively. It is worth noting that the maximum price of "house" type is the highest among all other types, while "unit" type has the lowest minimum price.

In addition to the tabular representation, we can also visualize the average price of each house type using Figure 5.

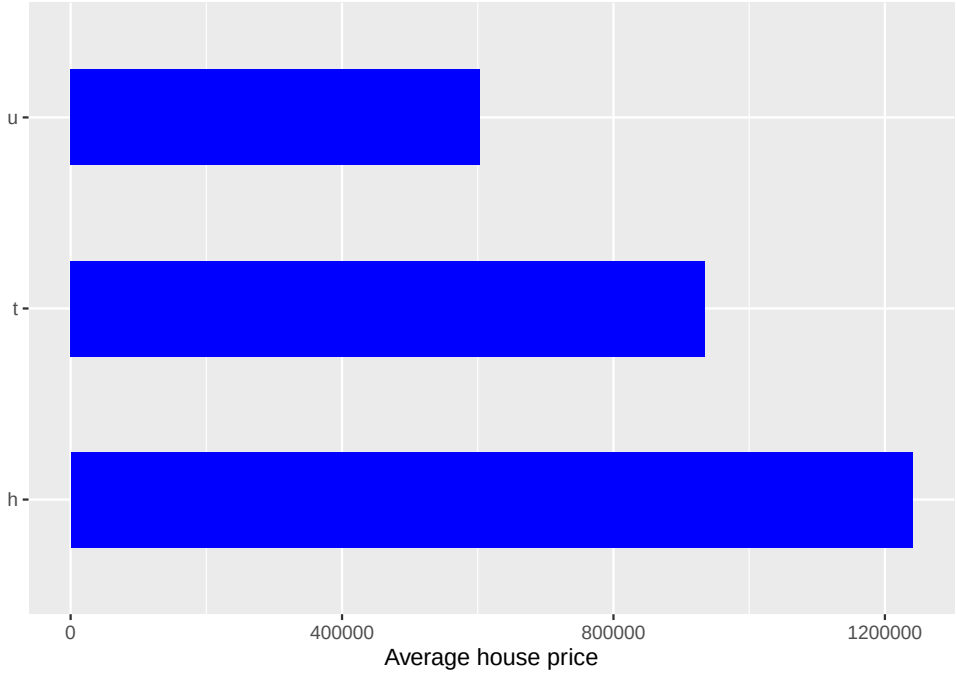

Figure 5: Average house price by type

The average price of the "house" type is significantly higher than the other two types, as shown in Figure 5. It is more than double the average price of the "unit" type.

3.3 House Price Exploration By Suburb

In the same manner, a table is created to show the relationship between house price and suburb by displaying the total number, average price, minimum price and maximum price of houses in each suburb. However, due to the large number of suburbs, only the top 10 suburbs with the highest and lowest house prices are presented. The top 10 suburbs with the highest house prices are illustrated in Figure 6.

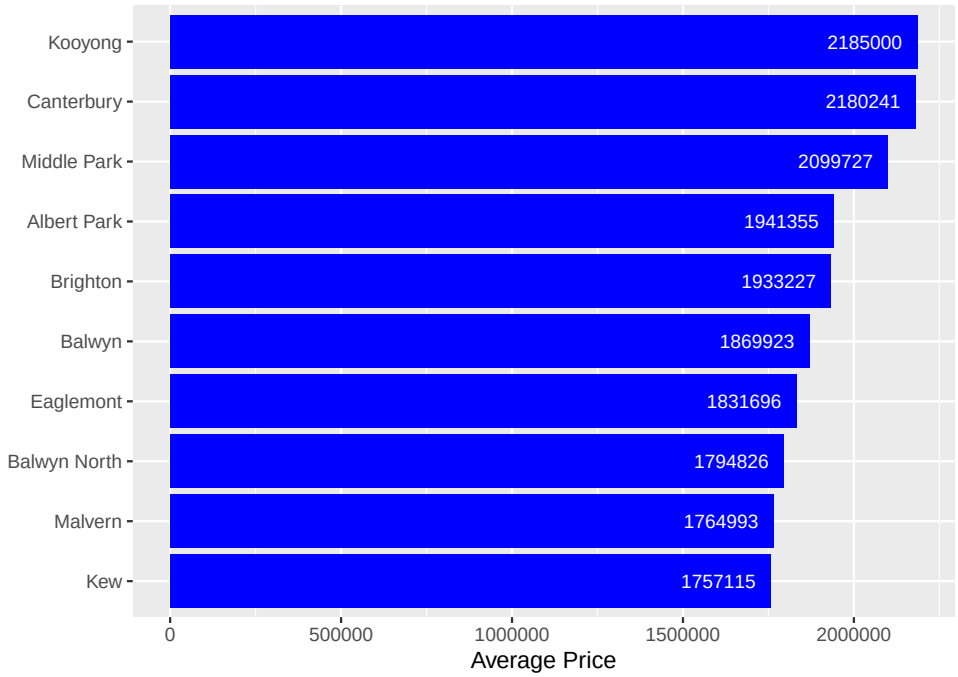

Figure 6: Top 10 highest average house price by suburb

Looking at the top 10 lowest house price suburbs presented in Figure 7, we can observe that most of the suburbs are located in the west and north-west areas of Melbourne. The lowest average house price is in Melton South, with less than 300,000 dollars, which is less than one-third of the average house price of the total data set.

The comparison between the most expensive and cheapest suburbs continues in Figure 7. Here, we see a contrasting picture with the top expensive suburbs. The highest average price in these cheapest suburbs is even less than half of the mean value of the Price data. It is interesting to note that in terms of the average price, the suburb of Kooyong (the most expensive suburb) has a number almost eight times higher than that of Bacchus Marsh (the cheapest suburb). Therefore, we can conclude that there is a significant variation in house prices among different suburbs.

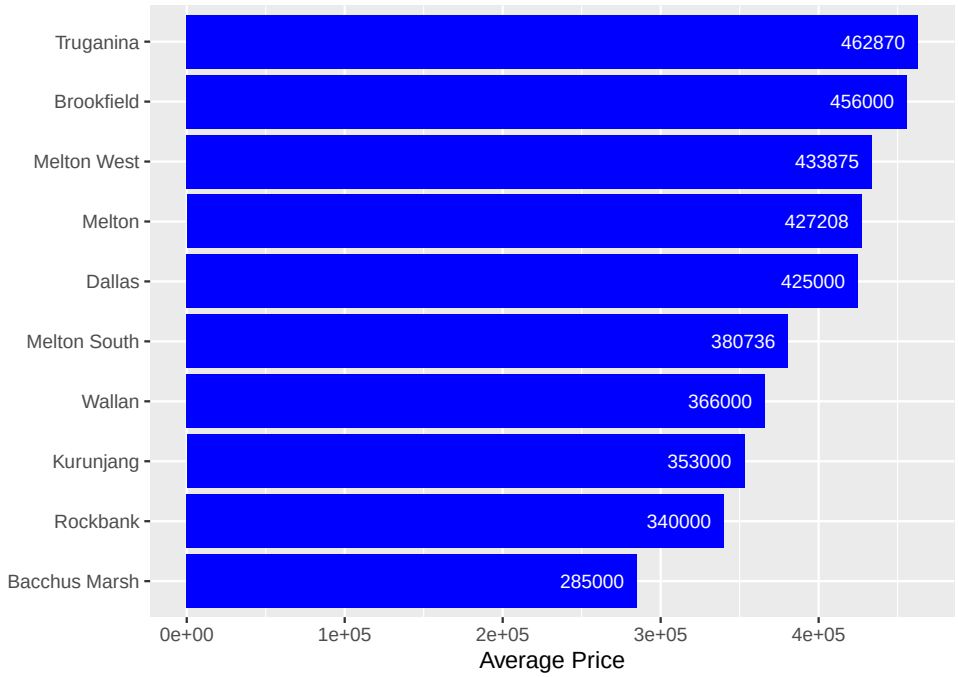

Figure 7: Top 10 lowest average house price by suburb

3.4 House Price Exploration By Other Variables

In this section, we explore the house price and other variables, including: Rooms, Bedroom2, Bathroom and Car, which is presented in Figure 8.

We can observe from Figure 8 that the highest number of rooms, bedrooms, bathrooms, and car spots does not necessarily come with the highest average price. However, the houses with the highest mean price are usually larger than normal ones (with more rooms, bedrooms, bathrooms, and car spots than average - around 7 to 9). On the other hand, the houses with the lowest average house price are usually the small ones (with 1 room, 1 bathroom, 1 bathroom, and 1 car spot).

Moving on to the next part, we explore the correlation between house price and distance. It is expected that houses closer to the CBD will be more expensive than the ones further away. We use a scatter plot and trend line to depict this relationship, which is shown in Figure 9.

As we observe the large variability in both house prices and distances, we can still identify a negative correlation between them. This implies that when the distance increases, the house price tends to decrease, and vice versa, which aligns with our initial expectations.

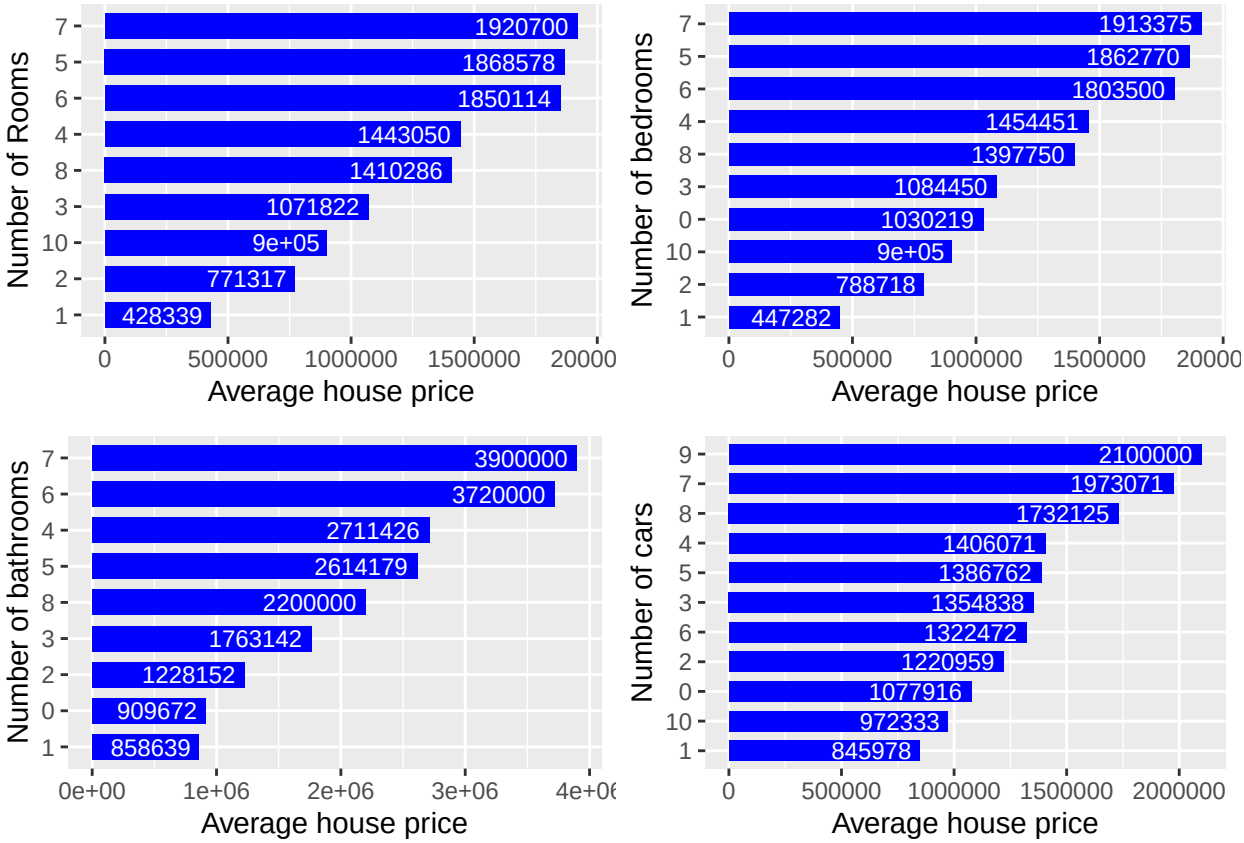

Figure 8: House price and Rooms, Bedroom2, Bathroom and Car

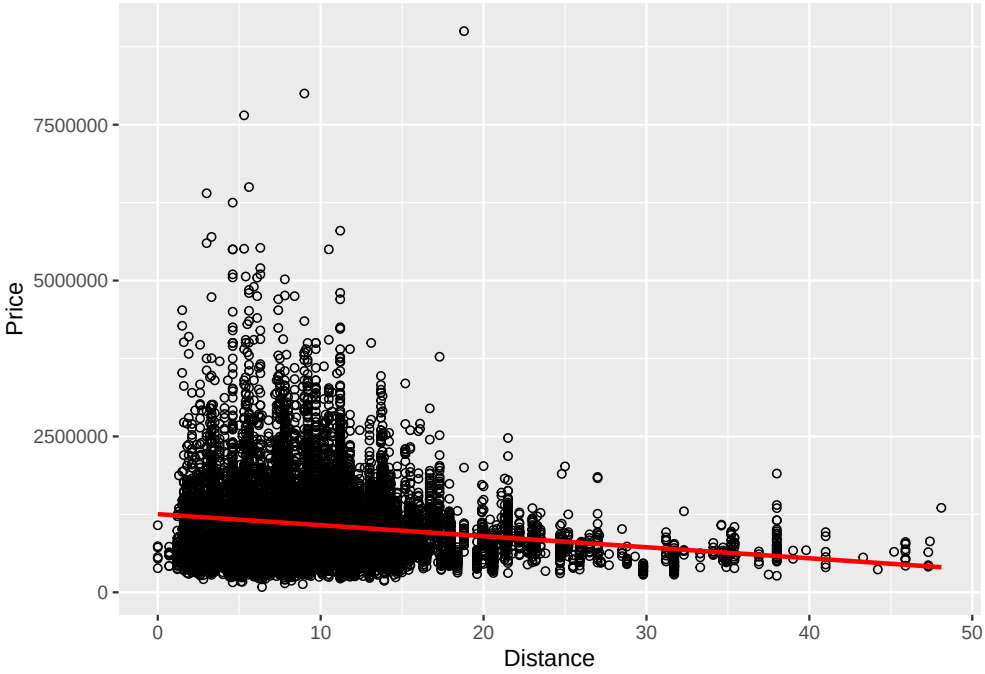

Figure 9: Scatter plot of house price and distance

4 Methodology and Result

As mentioned earlier, we utilized three methods, namely k-nearest neighbor (knn), linear regression, and random forest, to construct a house price prediction model. To select the best model with the smallest Root Mean Square Error (RMSE), we used a training data set, and then validated the chosen model with a separate validation data set. However, to reduce the number of variables, we only selected meaningful predictors out of the 20 available in the dataset, which were Rooms, Type, Distance, Bedroom2, Bathroom, and Car. We also used Price_log (the log form of Price) instead of the original Price.

We did not include the Suburb variable in the prediction model because it is a factor variable with 314 classes, making it difficult to handle within the models. Additionally, many of the suburbs only appear once in the data set (21 suburbs). Furthermore, since both Suburb and Distance variables are related to location, we deemed it sufficient to include only the Distance variable in the prediction model.

To train and develop the algorithm, we divided the data set into two parts: the training data set (which comprised 80% of the total data set), and the validation data set (which comprised 20% of the total data set). Dividing the data set in this way allows us to mimic the final evaluation process. The known outcome data set (i.e., the training data set) is used to develop and train the algorithm, and the validation set (or test set) is used to evaluate the algorithm's performance.

Typically, the proportion of the validation or test set is set between 10% and 30%. We chose to split the data set into two parts, with 80% for the training set and 20% for the validation set. This proportion allowed us to train a better prediction model while also testing how well the optimal model generalizes to unseen data. The validation set was only used to test the best model at the final part of this section.

4.1 K-Nearest Neighbors Method

To begin with, we utilized the knn method to construct the house price prediction model. The knn algorithm is a non-parametric method that works by calculating the distances between a query and all examples in the data set. It then selects the specified number of examples (k) that are closest to the query and either averages the labels or votes for the most frequent label in the case of regression or classification, respectively. Since we are dealing with continuous data, the knn algorithm produces the average value of the k closest examples.

To determine the optimal k with the smallest Root Mean Square Error (RMSE), we conducted 10-fold cross validation on the training set and experimented with different values of k, ranging from 2 to 50.

The optimal value of k with the smallest RMSE can be easily identified from Figure 10. As shown in Table 9, the final result for the optimal k is 22, with an RMSE of approximately 0.32. In the upcoming section, I will discuss the linear regression method.

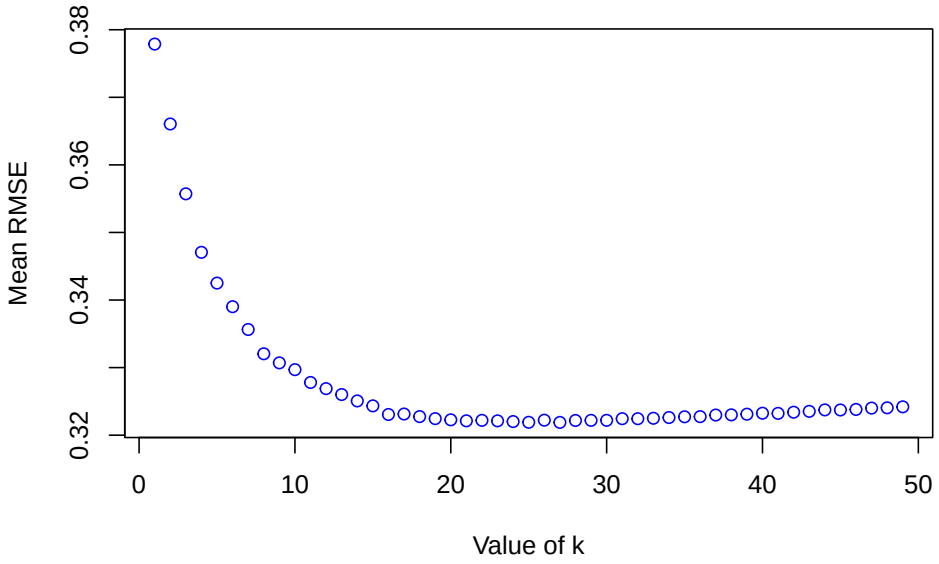

Figure 10: RMSE of knn Model

4.2 Linear Regression Method

We applied the linear regression method as the second approach to construct the house price prediction model. This is a simple parametric method that is appropriate for numerical data. Similar to the previous method, we also performed 10-fold cross-validation on the training set to determine the optimal model with the lowest Root Mean Square Error (RMSE). The final RMSE result for the linear regression model was approximately 0.35, as shown in Table 9.

4.3 Random Forest Method

Finally, we employed the random forest method to estimate house prices. The general idea behind the random forest algorithm is to create many predictors using regression or classification trees, and then compute a final prediction based on the average prediction of all these trees. To ensure that the individual trees are not identical, we used the bootstrap method to introduce randomness. These two features combined explain the name: the bootstrap makes the individual trees randomly different, and the combination of trees forms a forest. Upon running this algorithm, we obtained an RMSE result of approximately 0.31, as indicated in Table 9.

Three different types of prediction models have been constructed. To find the optimal model, we can compare the results in Table 9.

Table 9: Combined result of three methods

Method	RMSE
Knn Method	0.3218920
Linear Regression Method	0.3527677
Random Forest Method	0.3182078

According to the combined result table, the Random Forest model has the lowest RMSE, followed by the knn

and linear regression models, repectively. Therefore, we will select the Random Forest model and test it on the validation set in the next part.

4.4 Final Result

After selecting the optimal model, we proceed to evaluate its performance on the validation set. Table 10 displays the results obtained.

Table 10: Result of chosen model on validation set

Method	RMSE
Random Forest Method - Final Result	0.3323298

The validation set has an RMSE of approximately 0.33, which is slightly higher than the training set result. Additionally, we can examine the QQ-plot to evaluate the accuracy of the predicted values.

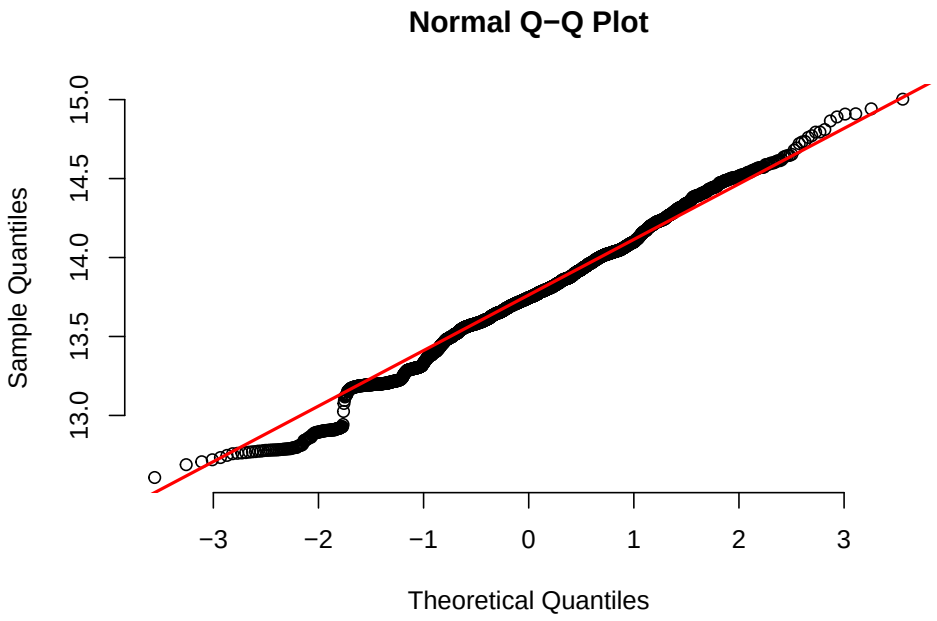

Figure 11: QQ-Plot

The variables importance index for the random forest model is presented in Figure 11 below. The QQ-plot indicates that the prediction model performs well, as the predicted outcome is nearly normally distributed.

The interpretability of the random forest method is limited, but one approach that can help is to examine the variable importance. Figure 12 presents the variable importance index for the random forest model. The graph on the left shows the Mean Decrease Accuracy, which measures how much the model's accuracy decreases if a variable is dropped. The graph on the right shows the Mean Decrease Gini, which measures the variable importance based on the Gini impurity index used for calculating splits in trees. Both graphs indicate that Distance is the most important variable, followed by Type, Bathroom, and Rooms. Conversely, Car and Bedroom2 are the least important variables. These results are sensible and reasonable.

Figure 12: Variable Importance Plot

5 Conclusion

n our project, we utilized the Melbourne housing dataset to develop a house price prediction model. By visualizing and exploring the data, we gained a more profound understanding of the house price situation in Melbourne. We experimented with three different methods, namely knn, linear regression, and random forest, in order to identify the optimal approach. After considering the RMSE results, we selected the random forest algorithm to build the final house price prediction model. The model's RMSE on the validation set was only slightly higher than that of the training set. Based on the combined QQ-plot result, we concluded that the selected model performed admirably. The most important variable in the model was Distance, followed by Type, Bathroom, and Rooms, whereas Car and Bedroom2 were the least important.

However, our project has a few limitations. Firstly, the data set only covers a two-year period (2016 and 2017) and contains limited variables. It was not possible to explore the house price trend over time due to the short time frame. Additionally, we did not employ any variable selection method to select the significant attributes. Instead, we only chose relevant attributes based on our subjective judgment. Expanding the house price dataset to include more years and attributes would be beneficial for future analyses. Additionally, a variable selection method should be applied to identify significant variables before feeding them into various prediction models, such as stepwise, LASSO, or elastic net methods. Finally, other algorithms, such as regression tree or gradient boosting machine, can be used to identify the optimal model.