House Price Project

Thi Hau Nguyen

Contents

1 Introduction

In this project, I use housing data set in Melbourne, Australia to predict house price. When people search for a house, they usually find on real estate websites. They can check information given for a specific house, but the house price is not always shown. The sellers can have some choices about house price, for instance: set an offer price, biding, open for negotiation, etc. However, the buyers need a guide for house price for some considerations like: Can they afford for that house? Whether that house is worth or not? What price they should set for negotiation? Or whether the offer price of the sellers is reasonable or not? Therefore, the objective of this project is to build a model to predict house price based on the features of a house to make recommendations about house price for buyers.

The data for this project is the Melbourne housing data set collected from Domain.com.au website with information of sold houses in 2016 and 2017. Three different methods are applied to predict house price, including k-nearest neighbors (knn), Linear Regression and Random Forest method. The training data set is used in each model to calculate the Root Mean Squared Error (RMSE), then find the optimal model with smallest cross-validation RMSE. Finally, the optimal model is chosen to estimate the RMSE on validation data set.

The structure of the report is as follows: Section 1 introduces the analytic problem, Section 2 presents data description and data cleaning, data exploration and visualization are included in Section 3, Section 4 discusses methodologies and results, finally a conclusion with limitations and further analysis is presented in Section 5.

2 Data preparation

2.1 Data description

To begin with, I describe the information of all variables in the data set. There are 21 variables, including 1 response (Price) and 13,580 observations in this data set. The details of each variable are shown 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

There are 8 character variables and 13 numeric variables in the data set. The summary of character variables and numeric variables are presented in Table 2 and Table 3 respectively.

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

We can see from the summary tables that there are 3 variables with missing data, including: Car, BuildingArea and Landsize. Because the number of N/A values of Car variable is quite small (about 0.46%), I still keep this variable. However, BuildingArea and Landsize variables have nearly a half of missing values, then I decide to remove them from the data set.

The missing values of Car variable is handled by median imputation method, which means that N/A values will be replaced by median value of Car data.

Next, I draw boxplots of some numeric variables and discover whether they have outliers or not. Firstly, we explore Price and Landsize variables boxplots in Figure 1.

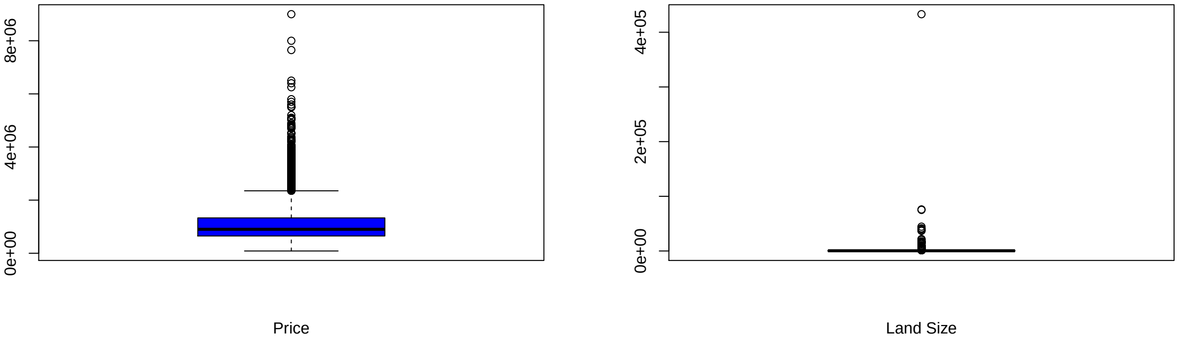

Figure 1: Boxplots of Price and Landsize

We see that there may be some outliers in Price (highest values), but I do not remove them as they may still reasonable. On the other hand, most data of Landsize equal 0 (about 14.3%), may be because this data was not given. Hence, I do not include this variable in prediction model.

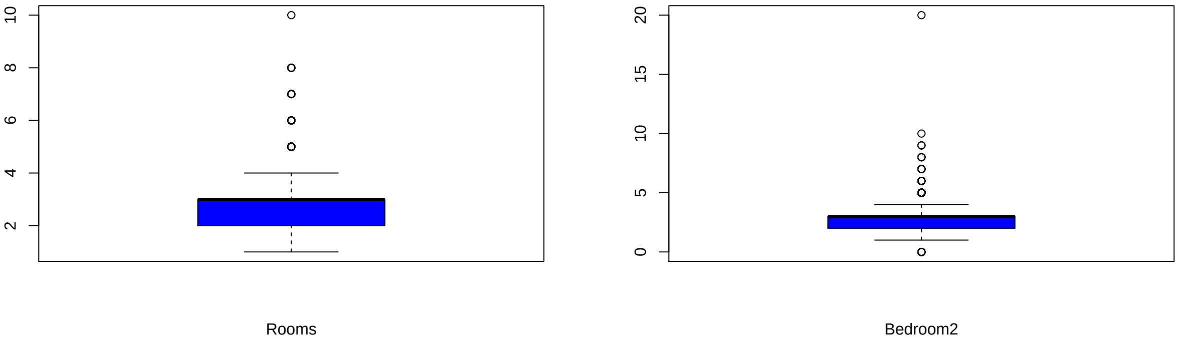

Figure 2: Boxplots of Rooms and Bedroom2

Figure 2 presents the boxplots of Rooms and Bedroom2 variables. There also may be some outliers in Rooms and Bedroom2 data on the top, so I check the highest values of these variables to find out more 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

The details of these highest values of Rooms and Bedroom2 show that it seems to be unreasonable when the number of bedrooms is bigger than the number of rooms. Therefore, I investigate and remove all of the data points like that in the data set (203 data points).

3 Data exploration and visualization

3.1 House price exploration

In this part, I explore the data set in more details and also visualize if possible. From Table 2, we can see that there are 13,378 different houses in 314 suburbs in the data set. Let's start with top 10 highest price houses and discover their features 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

We can see that the most expensive houses have different features, which most of them in the middle range of each variable (not too big in term of land size, not too many rooms, not so close to CDB, etc). Moreover, all of these houses are "house" type.

Table 7 presents the top 10 lowest price houses. The 10 cheapest houses have some similar characteristics: most of them have only 1 bedroom, 1 bathroom, small area and even no garage space. Besides that, eight of them 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

In addition, I also explore house price by visualization. The histogram of house price is shown in Figure 3.

As we can see from the histogram in Figure 3 that the right tail is longer than the left one, hence the distribution of house price is skewed right. It means that the mean value of Price is higher than the median value due to some extreme high data points in the Price data. Therefore, I transform Price data into the log form (create new variable called Price_log) and draw the histogram of log Price in Figure 4.

The histogram of log Price is much better than original Price, which shows that the distribution of Price_log is nearly normally distributed. Hence, I will use Price_log variable in prediction model instead of Price.

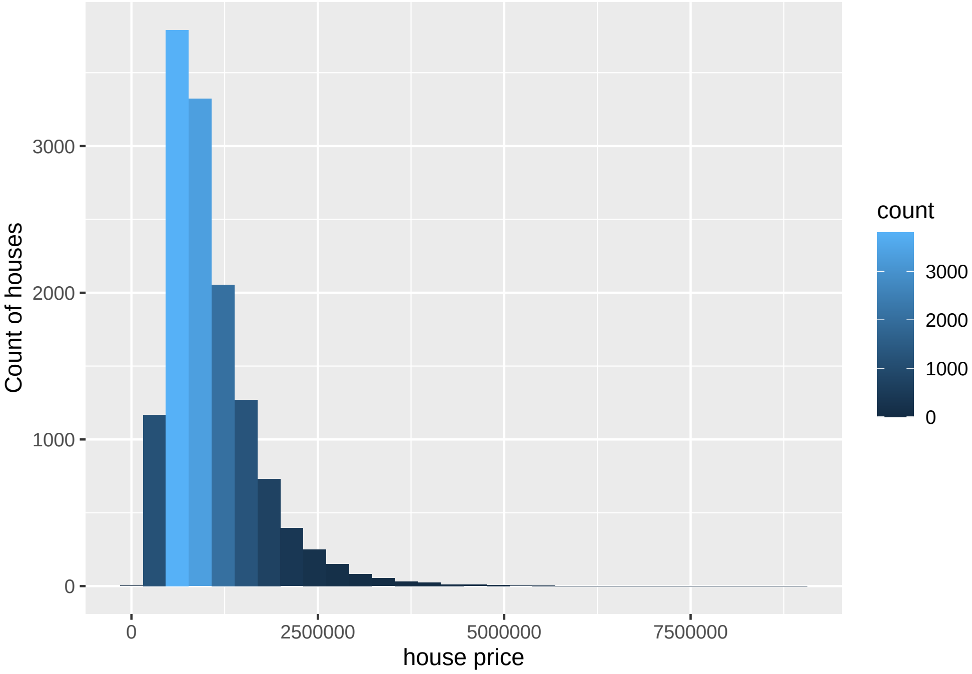

Figure 3: Histogram of Price

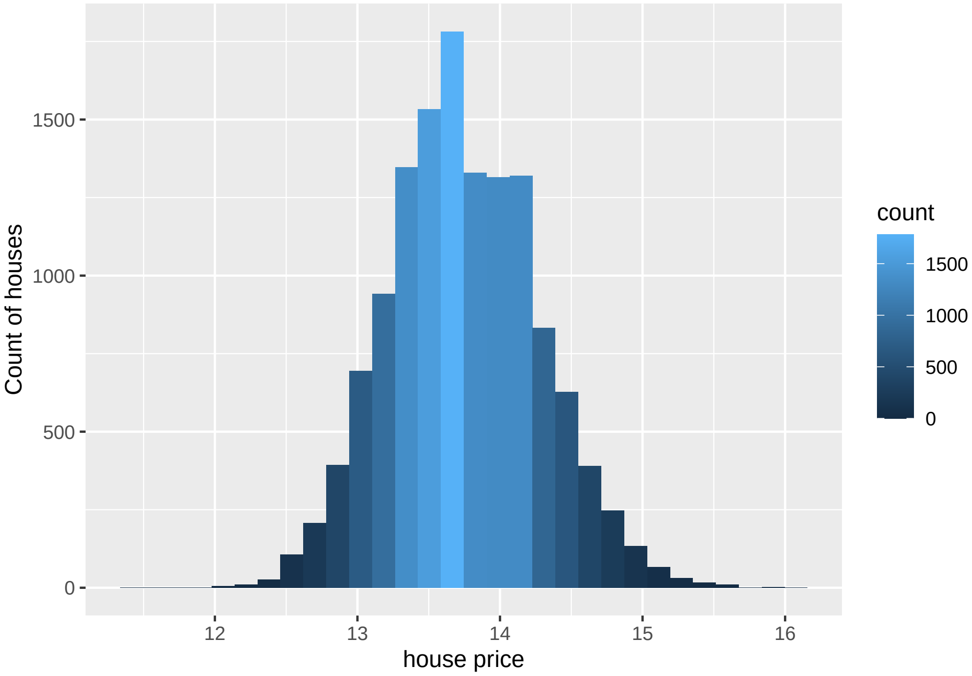

Figure 4: Histogram of log Price

3.2 House price exploration by type

In this section, I discover the house price and houses' types. Firstly, a table of Type variable with the total of houses, average price, minimum and maximum price in each house type is produced in Table 8.

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

Table 8 shows that the most popular house type is "house", which accounts for about two third of the total, followed by "unit" and "town house". "House" type also has the highest maximum price, on the other hand, the lowest minimum price is "unit" type. We also can explore the average price of each house type by visualization in Figure 5.

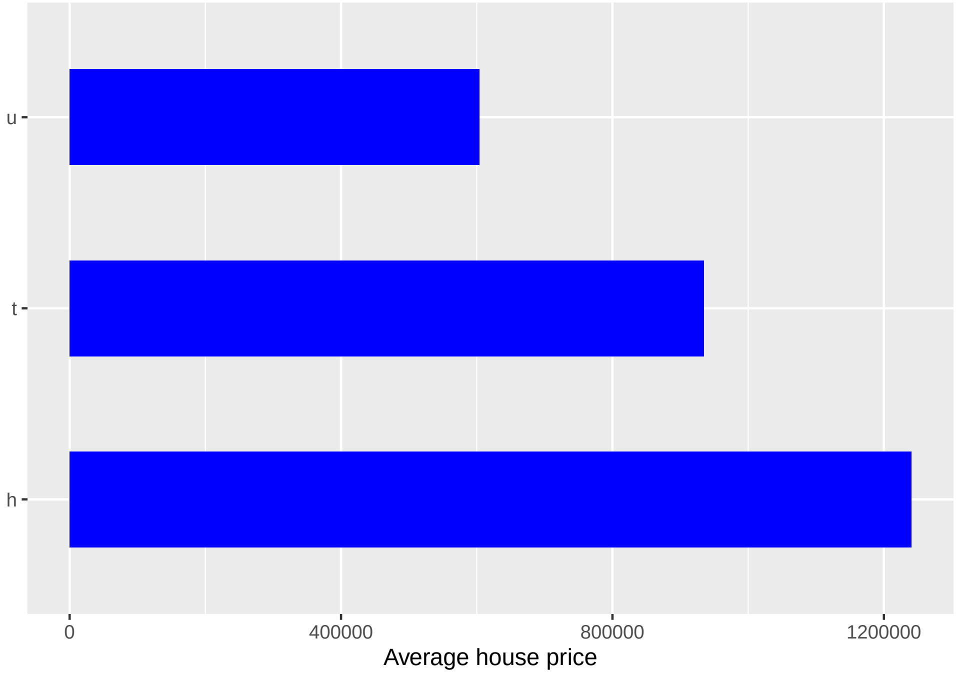

Figure 5: Average house price by type

It can be seen easily from the Figure 5 that the average price of "house" type is also highest among three types, and more than double the figure of "unit" type.

3.3 House price exploration by suburb

Similarly, a table of house price and suburb with the total number, average price, minimum price and maximum price in each suburb is created. However, I do not show all due to large number of suburbs. Instead, we discover the top 10 suburbs with highest and lowest house price. Let's start with the top 10 highest first which is shown in Figure 6.

Among of the top 10 highest, houses in Kooyong and Canterbury seem to be the most expensive ones, with average price higher than 2 million dollars (almost double the average house price of the total data set - at

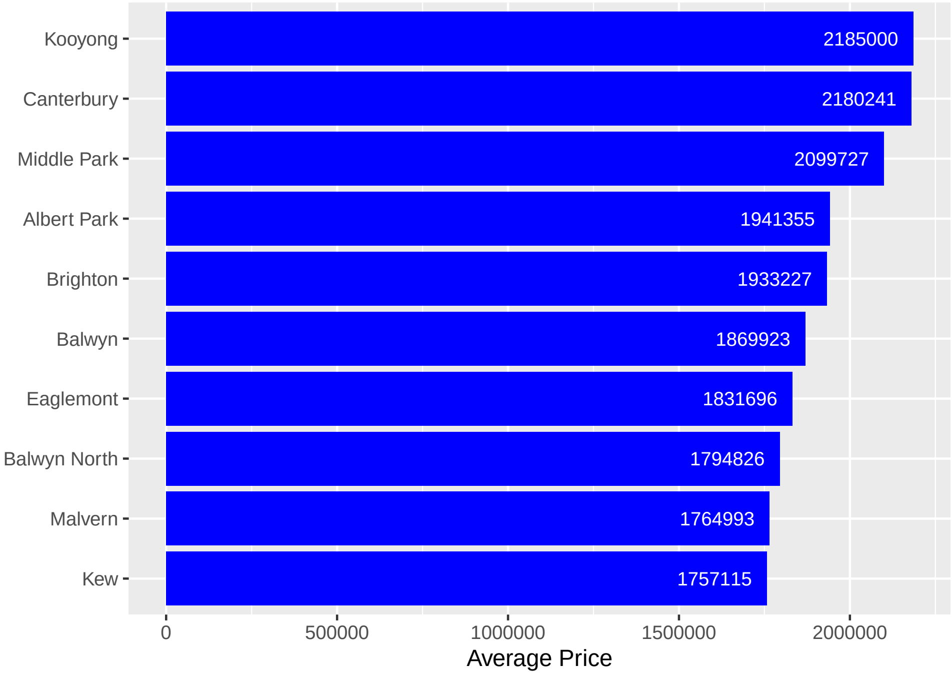

Figure 6: Top 10 highest average house price by suburb

around 1 million dollars from summary of data set in Table 3). Next, we explore the top lowest house price suburbs in Figure 7.

Figure 7 shows a different story compared with the top expensive suburbs. The highest average price of these cheapest suburbs is even less than a half of mean value of Price data. Moreover, in term of average price, Kooyong (the most expensive suburb) has the number nearly 8 times compared with Bacchus Marsh (the cheapest suburb). We can conclude that there is a big difference in house price among the suburbs.

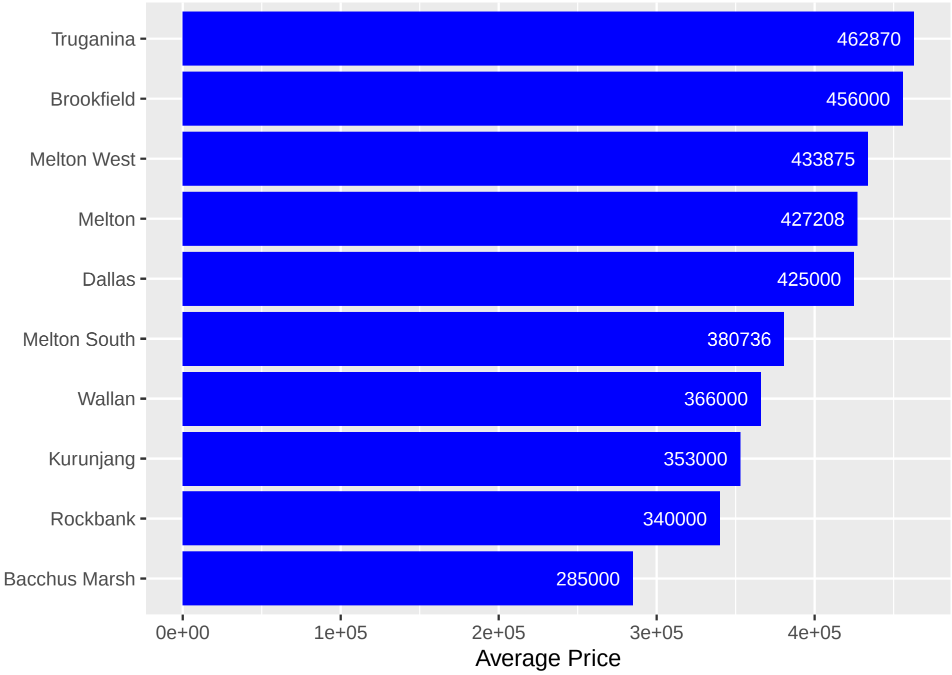

Figure 7: Top 10 lowest average house price by suburb

3.4 House price exploration by other variables

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

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

In the last part, I explore the correlation between house price and distance, we expect that houses closer to CBD will more expensive than the further ones. I use scatter plot and trend line to indicate this relationship, which is presented in Figure 9.

Although there are large variability in houses' price and distance, we still can see the negative correlation between them. It means that house price will decrease when distance increases or vice versa as we expected.

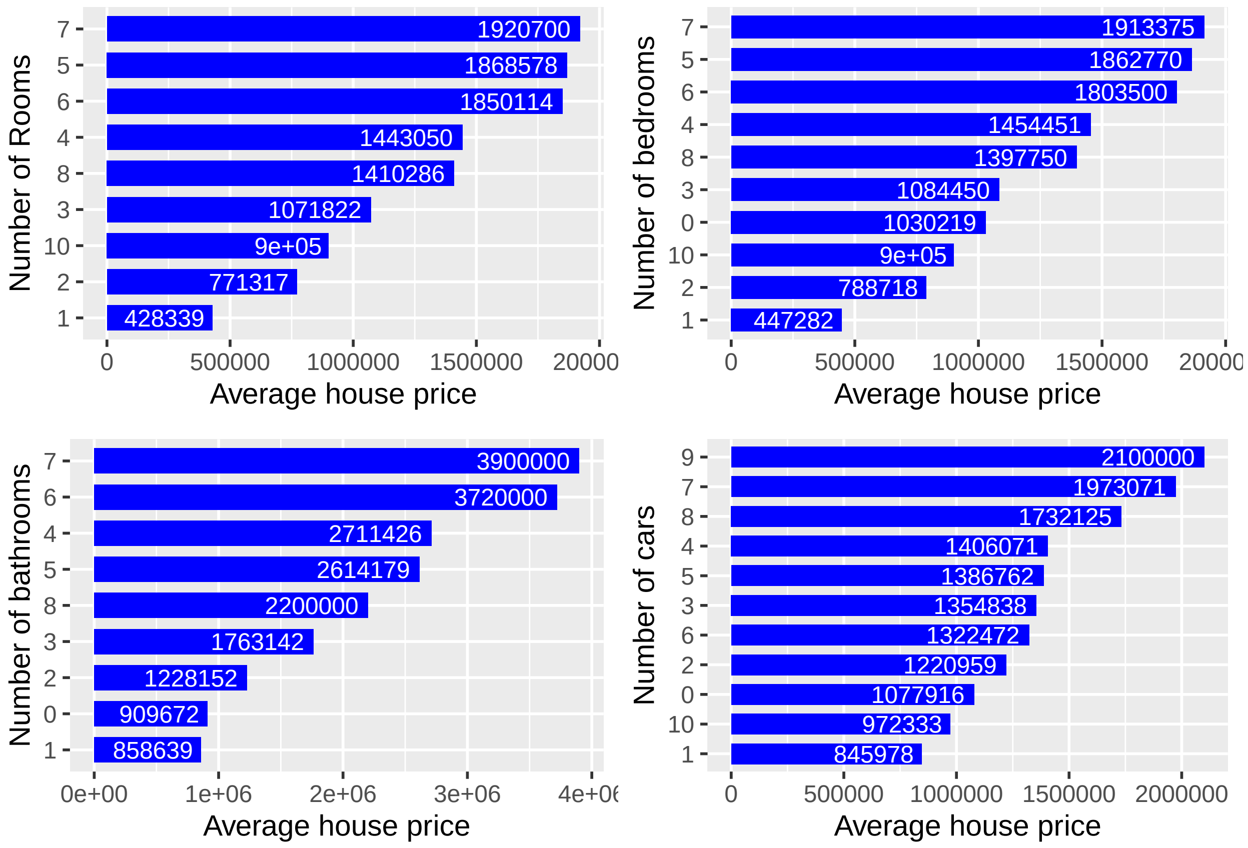

Figure 8: House price and Rooms, Bedroom2, Bathroom and Car

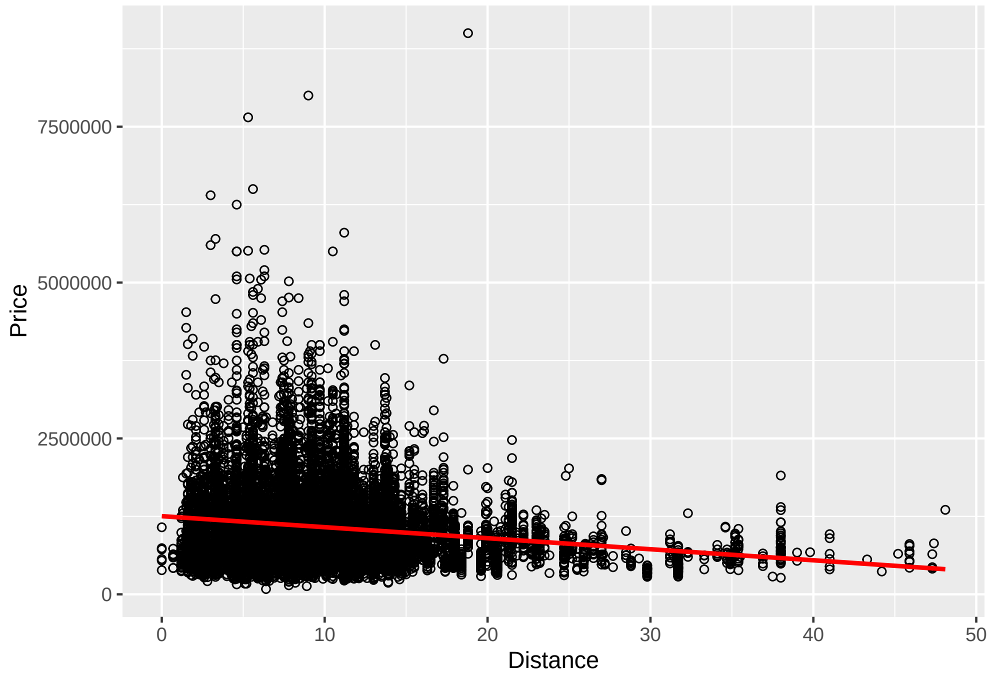

Figure 9: Scatter plot of house price and distance

4 Methodology and result

As mention in the first part, I use three different methods to build house price prediction model, including k-nearest neighbor (knn), linear regression and random forest method. A training data set is used to choose the best model with smallest Root Mean Square Error (RMSE), then a validation data set is put into the chosen model to get the final RMSE result. I will discuss more details each model in the following sections, but we need to select suitable variables first.

There are 20 predictors and 1 response in the data set, and I use Price_log (the log form of Price mentioned in Section 3.1) instead of original Price. Because there are so many variables and not all of them are really meaningful to put in the model. Therefore, I choose some suitable variables, including Rooms, Type, Distance, Bedroom2, Bathroom and Car.

Please note that I do not include Suburb variable in prediction model since this variable is factor data with 314 classes (so many to handle inside models) and many of them just appear 1 time only (21 suburbs). Moreover, both Suburb and Distance variables are refer to location, then including Distance in prediction model is enough.

The data set used in prediction models will be divided into two parts: training data set (accounts for 80% of the total) and validation data set (accounts for 20% of the total). Let's recall some knowledge here to explain why we have to split the data set into two different parts and why we set the proportion for each set like that. When developing an algorithm, we usually have a data set for which we already know the outcomes. Therefore, to mimic the ultimate evaluation process, usually we divide a data set into two parts and deal with one part as unknown outcomes. We will use the known outcome data set (training data set) to train and develop the algorithm, then after constructing it, the validation set (or test set) will be used to test the algorithm.

The proportion of validation set or test set is typically set around 10% to 30%, then I divide into two parts with 80% for training set and 20% for validation set. More training data set will help to "train" a better prediction model and 20% data set for validation set is enough to "test" how well the optimal model generalizes to unseen data. The validation set is only used to test the best model at the final part of this section.

4.1 k-Nearest neighbors method

First of all, I construct house price prediction model with knn method. The knn algorithm is a non-parametric method, it works by finding the distances between a query and all the examples in the data, selecting the specified number examples (k) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression). In this project, because the outcome is continuous data, knn algorithm produces the average value of k closest examples.

I do 10-fold cross validation on the training set and try different values of k from 2 to 50 to find the optimal k with smallest RMSE.

Figure 10 shows RMSE with different values of k, we can easily find out the optimal k from that. The combined result table of three methods Table 9 also shows the final result for optimal k is 22 with RMSE around 0.32. Linear regression method is presented in the upcoming part.

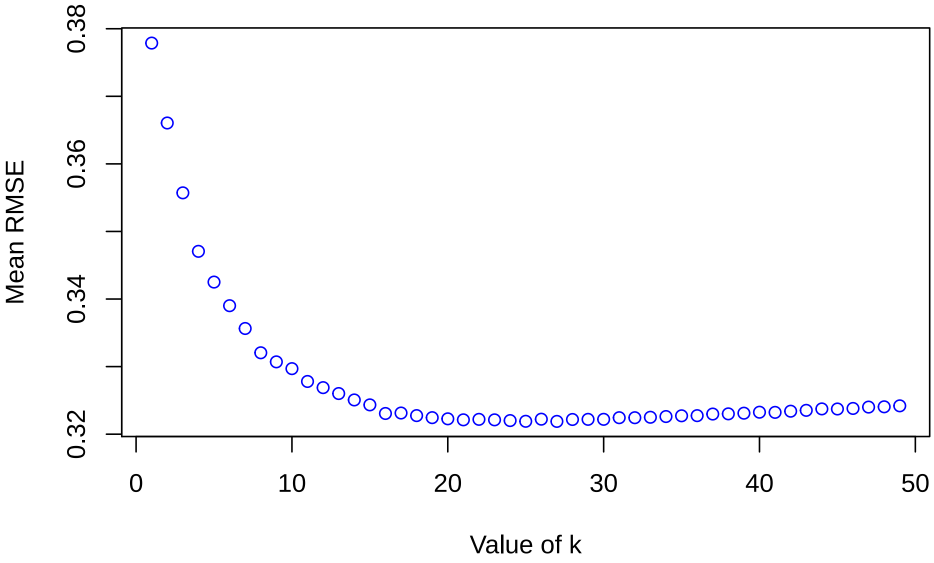

Figure 10: RMSE of knn Model

4.2 Linear regression method

The second method applied to build house price prediction model is linear regression. It's the simplest parametric method, but suitable with numeric data. I also do 10-fold cross validation with linear regression model, and the final RMSE result is around 0.35 (Table 9).

4.3 Random forest method

Finally, I use random forest method to estimate house price. The general idea of random forest algorithm is to generate many predictors, each using regression or classification trees, and then forming a final prediction based on the average prediction of all these trees. To assure that the individual trees are not the same, we use the bootstrap to induce randomness. These two features combined explain the name: the bootstrap makes the individual trees randomly different, and the combination of trees is the forest. After running this algorithm, I get RMSE result about 0.31 as shown in Table 9.

We already built three different kinds of prediction model, now we construct a combined result table, then compare and find the optimal one 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

The combined result table shows that Random Forest model has the lowest RMSE, followed by knn and linear regression model. Therefore, I will pick Random Forest model and test it on the validation set in the next part.

4.4 Final result on validation data set

After choosing optimal model, in this part we check how well this model works on the validation set. The result is shown in Table 10.

Table 10: Result of chosen model on validation set

Method	RMSE
Random Forest Method - Final Result	0.3323298

The RMSE of validation set is about 0.33, which is a little bit larger than the result of training set. We also can check the QQ-plot to check how well the predicted values is.

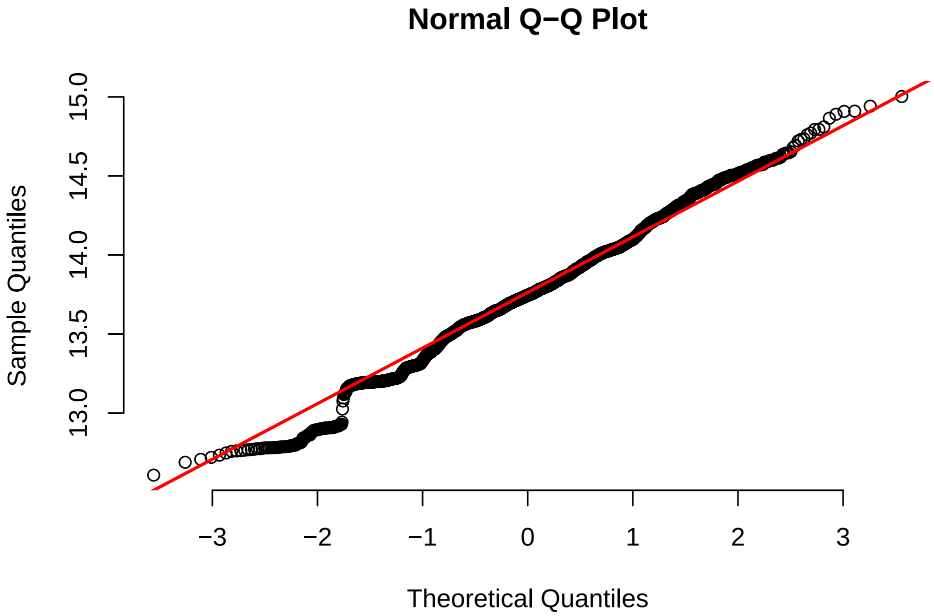

Figure 11: QQ-Plot

The QQ-plot also shows that the prediction model performs quite well since the predicted outcome is nearly normally distributed. Further more, we explore the results of random forest model by variables importance index, which is presented in Figure 12 below.

The main drawback of random forest method is that it's hard to interpret, but an approach that helps with interpretability is to examine variable importance. The left graph in Figure 12 shows Mean Decrease Accuracy (How much the model accuracy decreases if we drop that variable) and the right one shows Mean Decrease Gini (Measure of variable importance based on the Gini impurity index used for the calculation of splits in trees). Both graphs point out that the most important variable is Distance, followed by Type, Bathroom or Rooms. On the other hand, Car and Bedroom2 are the least important variables. The results make sense and reasonable.

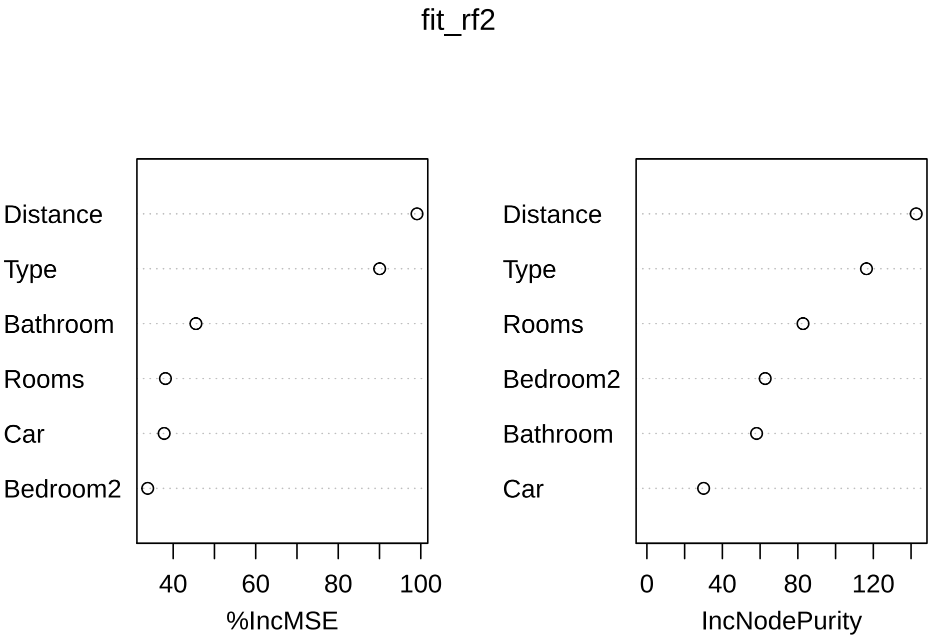

Figure 12: Variable Importance Plot

5 Conclusion

In this project, I use Melbourne housing data set to build a house price prediction model. By exploration and visualization, we can have a deeper insight and find out some more interesting information about the house price situation in Melbourne. I try three different methods to find the optimal one, including knn, linear regression and random forest algorithm. Based on the RMSE results, random forest algorithm is chosen to construct the final house price prediction model. We get the RMSE result on the validation set is about 0.33, which is not much bigger than RMSE result on the training set. Combined QQ-plot result, we can conclude that this chosen model performs quite well. Inside prediction model, the most important variable is Distance, on the other hand, Car and Bedroom2 are the least important variables.

However, this project also has some limitations. Firstly, the data set was collected for only two years (2016 and 2017) with limited variables. I also want to explore the house price change over time, but due to short time data set, then I could not explore it. Moreover, I did not use any variable selection method to choose significant variables. I just choose suitable attributes in the data set based on my opinion only, and it's not the optimal way. For further analysis, the house price data set can be expanded to larger data with more years and more attributes. Besides that, a variable selection method should be done to choose significant variables before putting into different prediction models (for example stepwise, LASSO or elastic net method). Finally, more algorithms can be applied to find the optimal model, like regression tree or gradient boosting machine.