Analysis of House Prices in Melbourne

Zizheng Wang

Rutgers University

May 7, 2021

Introduction

This data set is a snapshot of Melbourne's housing prices, including the variables such as:
Rooms: the number of rooms.
Price: price in dollars.
Regionname: the name of regions.
Propertycount: the number of properties in the suburb.
Distance: the distance to the CBD.
We want to use this data set to find a trend and give a prediction for the price of house.

Introduction

In this project, I try to:

1. clean and process the data.
2 do some descriptive statistical analysis through tables and plots.
3. build a linear regression model to predict the price.
4. the data is from https://www.kaggle.com/anthonypino/melbourne-housing-market.

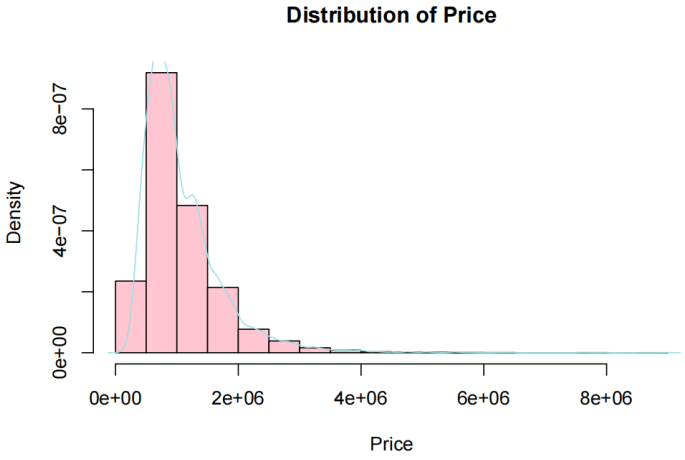
Data processing

I delete the observations, which has missing values in these five variables and extract these five variables for the following analysis.

```
# delete the missing values
index = !is.na(data$Price) & !is.na(data$Rooms) &
  !is.na(data$Regionname) & !is.na(data$Propertycount) &
  !is.na(data$Distance)
# extract these five variables
data = data[index,
            c("Price","Rooms","Regionname","Propertycount","Distance")]
```

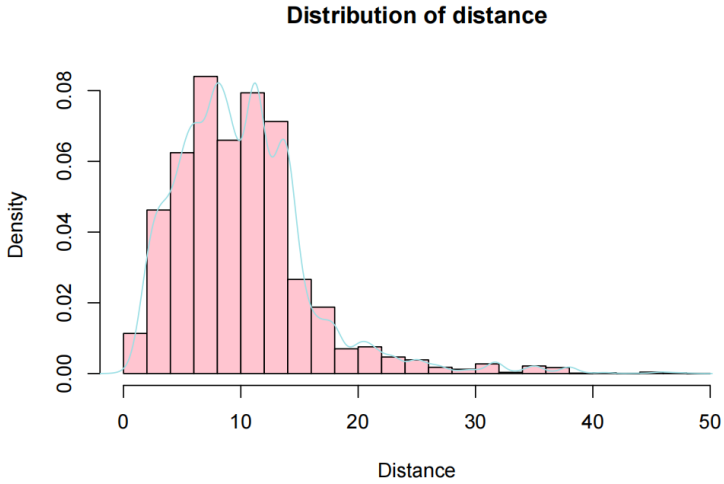
Analysis

Firstly, I make a histogram for the Price to see the distribution:



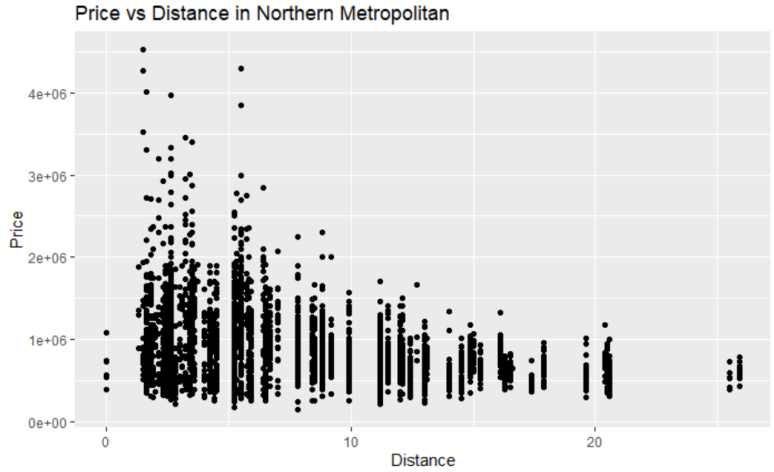
Analysis

Secondly, I make a histogram for the Distance to see the distribution.



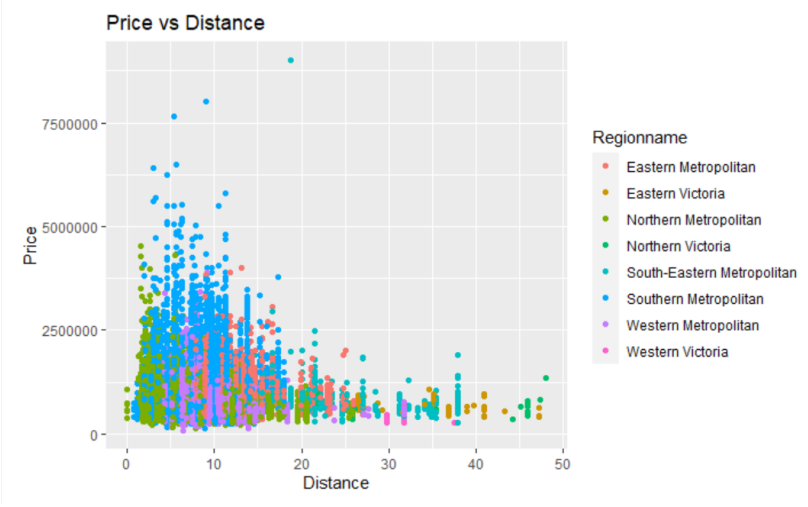
Analysis

Since the distance in the database is not a continuous variable, I build a scatter plot to see the distance-price relation in the Northern Metropolitan.



Analysis

Then I build another scatter plot to see the distance-price relation in the whole city of Melbourne.



Analysis

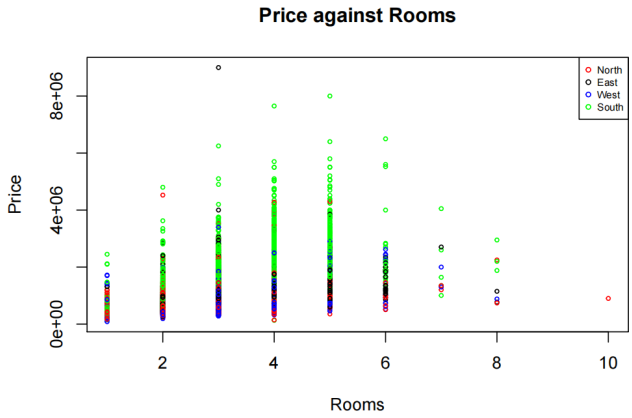
Next, I make a scatter plot for ‘Price‘ and ‘Rooms‘ with ‘Price‘ at the vertical axis and color the data by ‘Regionname‘. From above figures, we can see all regions with same "direction" share the same trend on the price.(For example, all regions whose name starting with "south" tend to have higher prices.) So we combine the regions into 4 classes, i.e., north, south, west and east, and make a more general analysis.

Analysis

```
# combine the Region into 4 classes
data[data$Regionname=="Northern Metropolitan" |
       data$Regionname=="Northern Victoria","Regionname"] = "North"
data[data$Regionname=="Eastern Metropolitan" |
       data$Regionname=="Eastern Victoria" |
       data$Regionname=="South-Eastern Metropolitan",
     "Regionname"] = "East"
data[data$Regionname=="Western Metropolitan" |
       data$Regionname=="Western Victoria","Regionname"] = "West"
data[data$Regionname=="Southern Metropolitan" |
       data$Regionname=="Southern Victoria","Regionname"] = "South"
```

Analysis

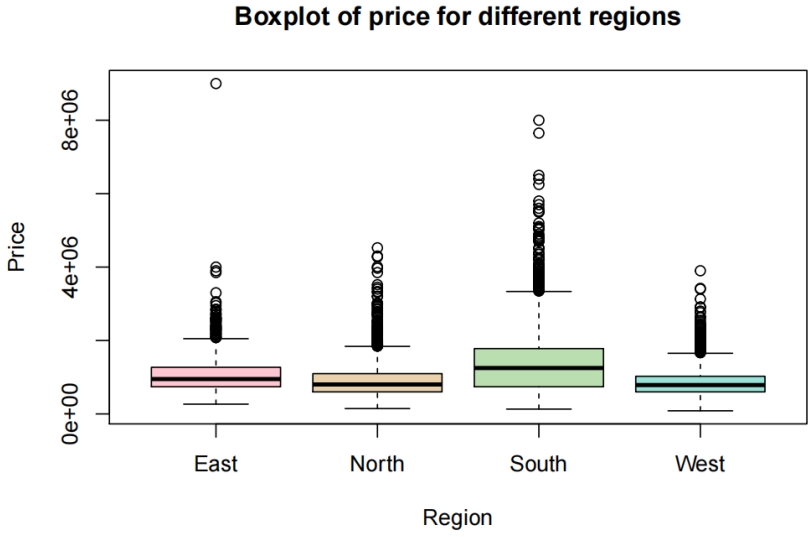
Now I make a scatter plot for Price and Rooms with Price at the vertical axis and color the data by Regionname.



Analysis

From the scatter plot, we can see that as the number of rooms changes from 1 to 5, the price increases gradually.
Furthermore, we can see that the price of house in southern area is the highest among the four regions while the prices of house in northern and western area are lower.
We can also obtain the same conclusion through a boxplot for the Price grouped by Region.

Analysis


Boxplot of price for different regions
Price
0e+00
4e+06
8e+06
East
North
South
West
Region

Linear Regression

Finally, I build a linear regression model using Price as dependent variable and other four variables as predictors.

Table 1: Coefficients estimates of linear regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.552257e+05	2.096700e+04	16.942132	0.0000000
Rooms	3.983395e+05	4.385979e+03	90.821102	0.0000000
RegionnameNorth	-2.318262e+05	1.485028e+04	-15.610896	0.0000000
RegionnameSouth	2.186858e+05	1.404608e+04	15.569167	0.0000000
RegionnameWest	-3.058005e+05	1.436864e+04	-21.282501	0.0000000
Propertycount	-2.205875e+00	9.722249e-01	-2.268893	0.0232904
Distance	-3.697188e+04	7.981737e+02	-46.320588	0.0000000

Conclusion

From the results, we can see that the number of Rooms has a positive impact on the house price.
The house in southern area is highest and then the eastern area, the northern area, and western area.
The distance has a negative impact on the house price. The more close to the CBD the higher price will be.

The End