midterm

2024-03-11

Project Introduction

1. Data background The dataset is from https://www.kaggle.com/datasets/shree1992/housedata/data The real estate markets, like those in Sydney and Melbourne, present an interesting opportunity for data analysts to analyze and predict where property prices are moving towards. Prediction of property prices is becoming increasingly important and beneficial. Property prices are a good indicator of both the overall market condition and the economic health of a country.
2. Problem we want to solve House price is a classic problem in machine learning and statistics. And house price is always a problem that everyone cares about. So we want to figure out what factors would affect the housing price, that is, whether the number of bedrooms, the number of bathrooms, the square footage of the living space, the lot, the places above the ground, the age of the house, the number of waterfront, the score of view and condition would affect the house price.

Methods

I download the data from kaggle, https://www.kaggle.com/datasets/shree1992/housedata/data And then I use read.csv to import the data into R. From dim(), we can see the data has 4600 rows, 18 columns Using the str() function, I checked the type of the variables.

```
data <- read.csv("data.csv", header=TRUE)
```

I use na.omit() to delete the rows with missing data, and the result shows that the data has no missing values. For variable "waterfront", "view", "condition", I turn them from int into type factor. Since the difference between 0,1 and 1,2 doesn't represent the same amount of difference.

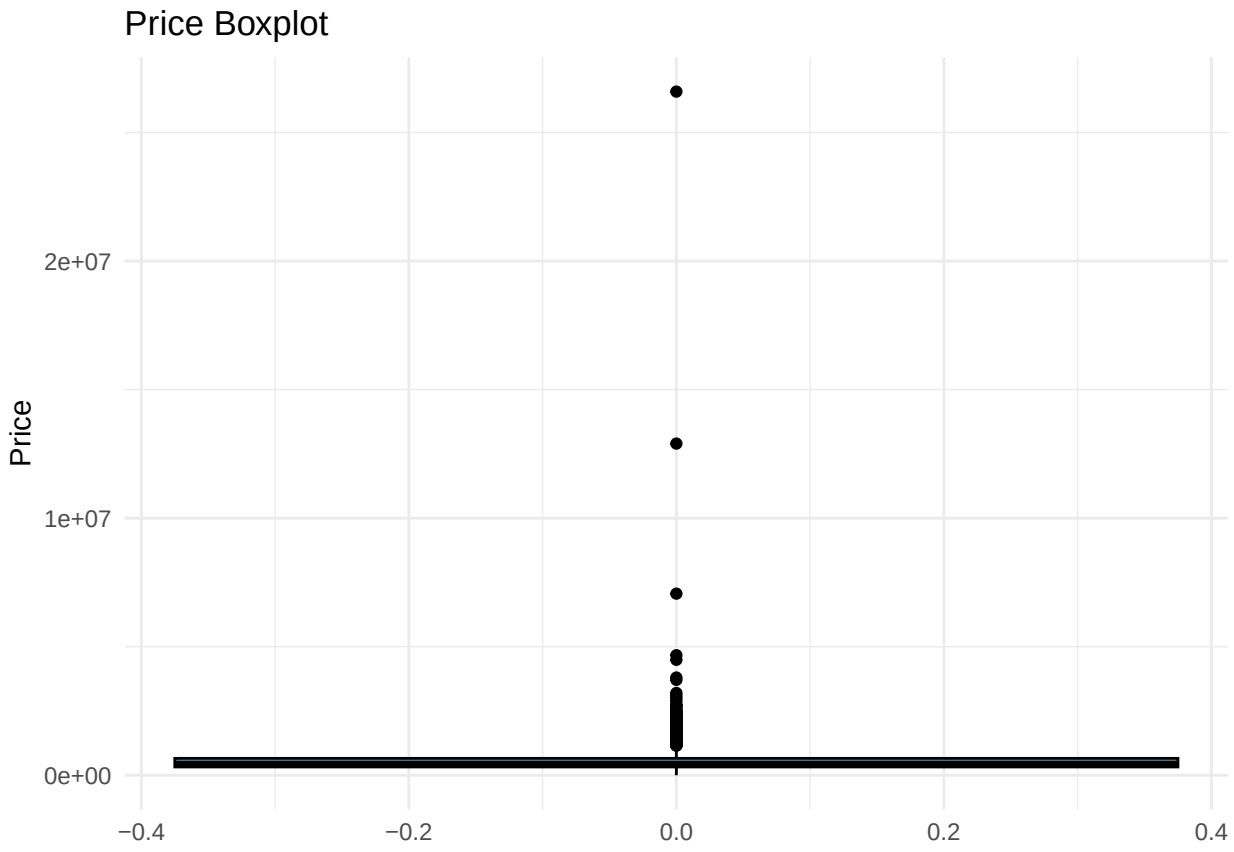
```
data<- na.omit(data)
data$waterfront <- factor(data$waterfront, levels = c(0, 1), labels = c("No", "Yes"))

data$view <- factor(data$view, levels = c(0, 1, 2, 3, 4), ordered = TRUE)

data$condition <- factor(data$condition, levels = c(1, 2, 3, 4, 5), ordered = TRUE)
```

Then I plot the distribution of price using box plot using ggplot. Then I found there are some extremely high prices form the boxplot. So I decide to clean those outliers since the those extreme value may affect the other data heavily because of their extremely high prices. After that, 4360 rows is left.

```
ggplot(data, aes(y = price)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Price Boxplot", y = "Price") +
  theme_minimal()
```




```r
Q1 <- quantile(data$price, 0.25, na.rm = TRUE)
Q3 <- quantile(data$price, 0.75, na.rm = TRUE)


IQR <- Q3 - Q1


lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# Filter the data to remove outliers
data <- data[data$price >= lower_bound & data$price <= upper_bound, ]
```

Also, I use the 2014-yr_built(the data was collected in 2014) to build a new variable age, since I think age may also be an important factor that affects the housing price. And the age is more directly than the yr_built variable.

```r
data$age <- 2014 - data$yr_built
```

Finally, I kept the variables that we really care in our problems. There are 11 variables left including price.

```r
data <- data[, c("price", "bedrooms", "bathrooms", "sqft_living", "sqft_lot",
                 "sqft_above", "floors", "waterfront",
                 "view", "condition", "age")]
```

Preliminary Results

I use a basic linear regression model to check if the variable are truly important factor to housing price. In the summary, here is some points that we need to pay attention to:

1. The coefficient for bedrooms is statistically significant (p-value < 0.05), but with a negative estimate. This implies that, all else being equal, an increase in the number of bedrooms is associated with a decrease in price, which could be counterintuitive and we need further investigation.
2. The variable view and condition, when broken down into its components, shows that the quality of view has a complex, non-linear relationship with price, which also needs further investigation. Maybe a more complex model to explain these two factor.
3. For variable sqft_lot, sqft_above, waterfront, these variables are not statistically significant at the 5% level, indicated by their p-values being greater than 0.05. This suggests that within the context of this model, they do not have a significant impact on the price.

```
model <- lm(price ~ bedrooms + bathrooms + sqft_living + sqft_lot + sqft_above +
            floors + waterfront + view + condition + age, data = data)
```

```
summary_lm <- summary(model)

coefficients_table <- summary_lm$coefficients

kable(coefficients_table, caption = "Linear Regression Analysis Summary")
```

Table 1: Linear Regression Analysis Summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.007638e+04	2.433589e+04	1.2358855	0.2165680
bedrooms	-2.984437e+04	3.773091e+03	-7.9097930	0.0000000
bathrooms	3.001026e+04	6.216072e+03	4.8278493	0.0000014
sqft_living	1.529601e+02	8.255115e+00	18.5291299	0.0000000
sqft_lot	-1.030693e-01	7.439290e-02	-1.3854722	0.1659793
sqft_above	8.298969e+00	7.956843e+00	1.0429976	0.2970074
floors	6.521921e+04	6.646098e+03	9.8131576	0.0000000
waterfrontYes	-2.105009e+04	4.409204e+04	-0.4774125	0.6330925
view.L	-1.566844e+04	2.213202e+04	-0.7079535	0.4790121
view.Q	-7.148922e+04	1.917447e+04	-3.7283544	0.0001952
view.C	3.055469e+04	2.113050e+04	1.4459990	0.1482496
view^4	-7.696650e+04	1.747373e+04	-4.4046968	0.0000108
condition.L	1.166286e+05	4.556583e+04	2.5595636	0.0105139
condition.Q	1.547632e+04	3.859147e+04	0.4010295	0.6884181
condition.C	-4.541529e+04	2.967240e+04	-1.5305566	0.1259518
condition^4	5.406898e+04	1.735041e+04	3.1162946	0.0018433
age	1.734222e+03	1.188679e+02	14.5894931	0.0000000

Summary so far

After some basic model and analysis, we can conclude that the number of bedrooms, the number of bathrooms, the square footage of the living space the age of the house would affect the housing price. But we need to remember that the coefficient of the number of bedrooms is negative, which is counterintuitive. Also for the variable condition and view, we need more complex model to explain their influence. In the future, we can build complex model such as Generalized linear model, poisson regerssion. We can also use machine learning model in the future to explain the issues left.