Housing Price Prediction Model for Modern Homes in New Taipei City

Bin Han

2025-07-09

Table of contents

1 Project description

This project investigates the factors influencing house prices in the Sindian district of New Taipei City, Taiwan, using housing data from 2012 and 2013. The dataset includes variables such as transaction date, house age, distance to the nearest Mass Rapid Transit (MRT) station, number of nearby convenience stores, and geographic coordinates (latitude and longitude). A linear regression model was developed using only houses newer than 20 years of age to ensure relevance to modern housing conditions.

The final model identifies significant predictors of house price: more recent transaction dates, higher latitude, shorter MRT distance, and a greater number of nearby convenience stores are associated with higher prices, while older house age correlates with lower prices. **The model demonstrates moderate explanatory power with an adjusted R-squared of 0.625 and a residual standard error of 8.589. Performance evaluation shows a Root Mean Squared Error (RMSE) of 8.50 and a Mean Absolute Percentage Error (MAPE) of 17.7%, indicating that predictions are,** on average, within 17.7% of the actual house prices.**

This model can be practically applied to estimate house prices by inputting key variables, using average values for the others, supporting future decision-making in urban planning, investment, or real estate assessment.

2 Load Data and R library

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
```

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

```
library(car)
```

```
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode
```

```
The following object is masked from 'package:purrr':

    some
```

```r
data<-read.csv("real-estate-taiwan.csv")
data<-data%>%
  filter(house_age>=0 & house_age<=20)
```

3 Exploratory data analysis

3.1 Data overview

```r
glimpse(data)
```

```
Rows: 277
Columns: 7
$ transaction_date   <dbl> 2012.917, 2013.583, 2013.500, 2012.833, 2012.667, 2~
$ house_age          <dbl> 19.5, 13.3, 13.3, 5.0, 7.1, 17.9, 6.3, 13.0, 13.2, ~
$ mrt_distance       <dbl> 306.59470, 561.98450, 561.98450, 390.56840, 2175.03~
$ convenience_stores <int> 9, 5, 5, 5, 3, 3, 9, 5, 4, 6, 1, 8, 7, 3, 7, 1, 7, ~
$ latitude           <dbl> 24.98034, 24.98746, 24.98746, 24.97937, 24.96305, 2~
$ longitude          <dbl> 121.5395, 121.5439, 121.5439, 121.5425, 121.5125, 1~
$ house_price        <dbl> 42.2, 47.3, 54.8, 43.1, 32.1, 22.1, 58.1, 39.3, 34.~
```

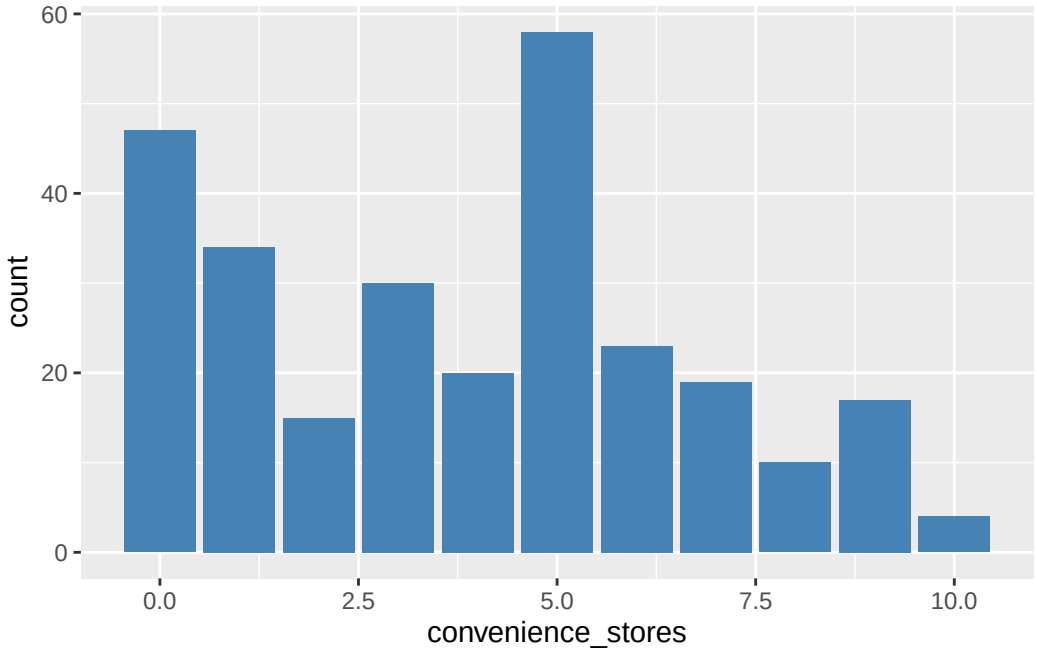
```r
summary(data)
```

```
 transaction_date   house_age      mrt_distance     convenience_stores
 Min.   :2013     Min.   : 0.00   Min.   :  23.38   Min.   : 0.000
 1st Qu.:2013     1st Qu.: 5.40   1st Qu.: 289.32   1st Qu.: 1.000
 Median :2013     Median :12.80   Median : 492.23   Median : 4.000
 Mean   :2013     Mean   :10.83   Mean   :1099.61   Mean   : 3.856
 3rd Qu.:2013     3rd Qu.:16.10   3rd Qu.:1643.50   3rd Qu.: 6.000
 Max.   :2014     Max.   :20.00   Max.   :6488.02   Max.   :10.000
    latitude       longitude      house_price
 Min.   :24.93   Min.   :121.5   Min.   :  7.60
 1st Qu.:24.96   1st Qu.:121.5   1st Qu.: 28.40
 Median :24.97   Median :121.5   Median : 40.10
 Mean   :24.97   Mean   :121.5   Mean   : 39.09
 3rd Qu.:24.98   3rd Qu.:121.5   3rd Qu.: 49.00
 Max.   :25.01   Max.   :121.6   Max.   :117.50
```

```
data|>
  count(convenience_stores, name="count")
```

```
   convenience_stores count
1                   0    47
2                   1    34
3                   2    15
4                   3    30
5                   4    20
6                   5    58
7                   6    23
8                   7    19
9                   8    10
10                  9    17
11                 10     4
```

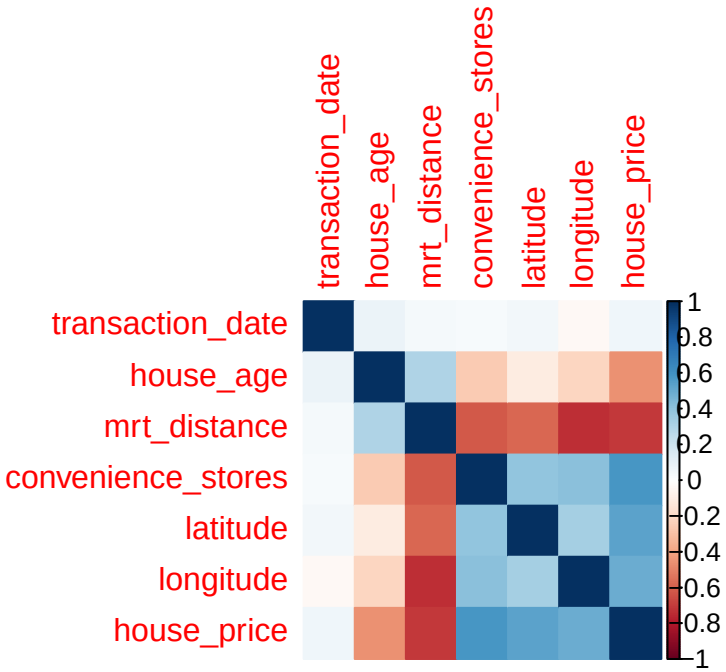
```
data%>%
  ggplot(aes(x=convenience_stores))+
  geom_bar(fill="steelblue")
```



In our data set, most houses have 5 convenience stores and 0 convenience stores.

3.2 Compute correlations

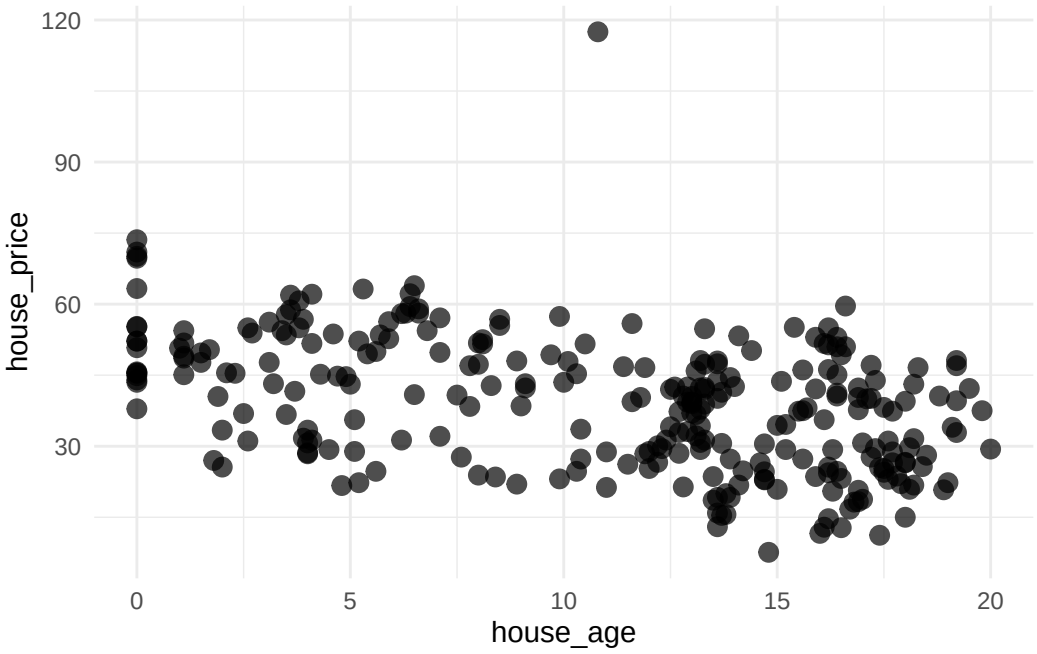
```
data%>%
  select(where(is.numeric))%>%
  cor(use="pairwise.complete.obs")%>%
  corrplot(method="color")
```



From the correlation plot, house price is proportional to longitude, latitude, and the number of convenience stores, and inversely proportional to Mass Rapid Transit (MRT) distance. Since the transaction data ranges from 2012.667 to 2013.583, house prices increase slightly over time. Additionally, house prices decrease slightly as the age of the house increases.

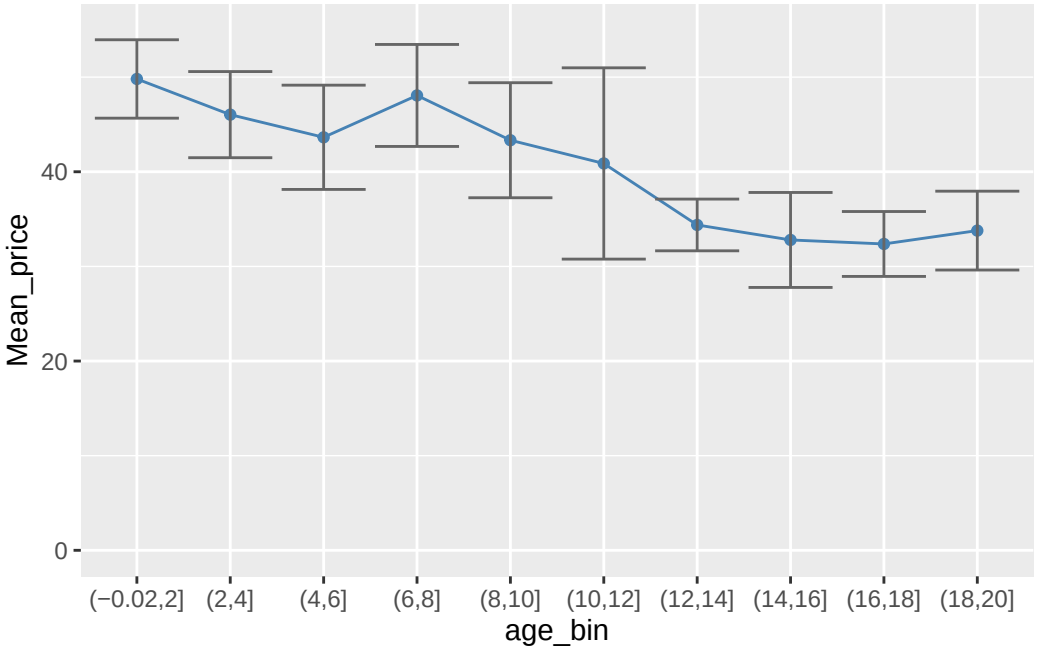
3.3 Scatter plot and line plot

```
ggplot(data, aes(x=house_age, y=house_price))+
  geom_point(size=3, alpha=0.7)+
  #scale_color_viridis_d(name="Convenience Stores")+
  theme_minimal()
```



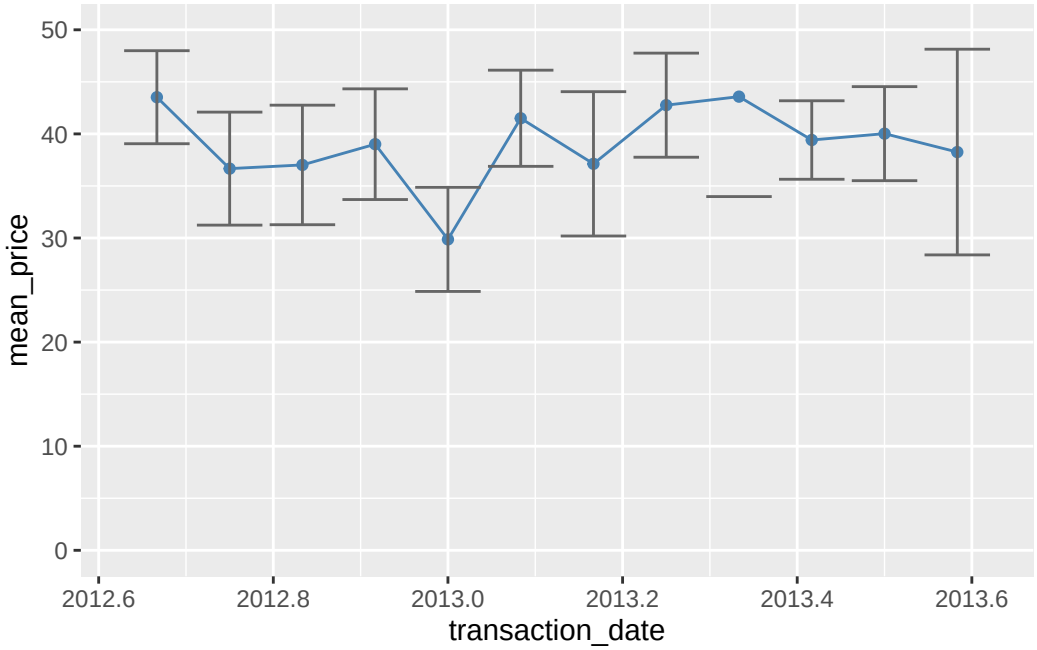
From the scatter plot, the house price decreases slightly with the increase of the house ages.

```
data%>%
  mutate(age_bin=cut(house_age, breaks=10))%>%
  group_by(age_bin)%>%
  summarise(Mean_price=mean(house_price),
            sd=sd(house_price),
            n=n(),
            se=sd/sqrt(n),
            lower=Mean_price-1.96*se,
            upper=Mean_price+1.96*se) %>%
  ggplot(aes(x=age_bin, y=Mean_price, group=1))+
  geom_point(color="steelblue")+
  geom_line(color="steelblue")+
  ylim(0,55)+
  geom_errorbar(aes(ymin=lower, ymax=upper), color="gray40")
```



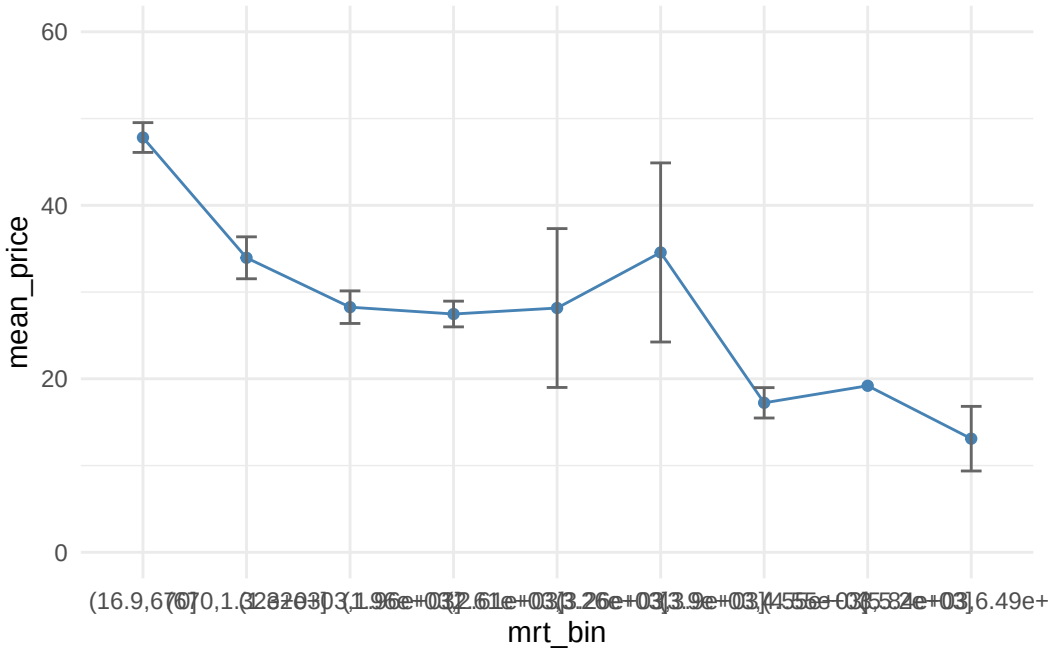
From the line plot, the house price decreases slightly with the increase of the house ages.

```
data%>%
  group_by(transaction_date)%>%
  summarise(mean_price=mean(house_price, na.rm=TRUE),
            sd=sd(house_price,na.rm=TRUE),
            n=n(),
            se=sd/sqrt(n),
            lower=mean_price-1.96*se,
            upper=mean_price+1.96*se
            )%>%
  ggplot(aes(x=transaction_date, y=mean_price))+
  geom_point(color="steelblue")+
  geom_line(color="steelblue")+
  ylim(0,50)+
  geom_errorbar(aes(ymin=lower, ymax=upper), color="gray40")
```



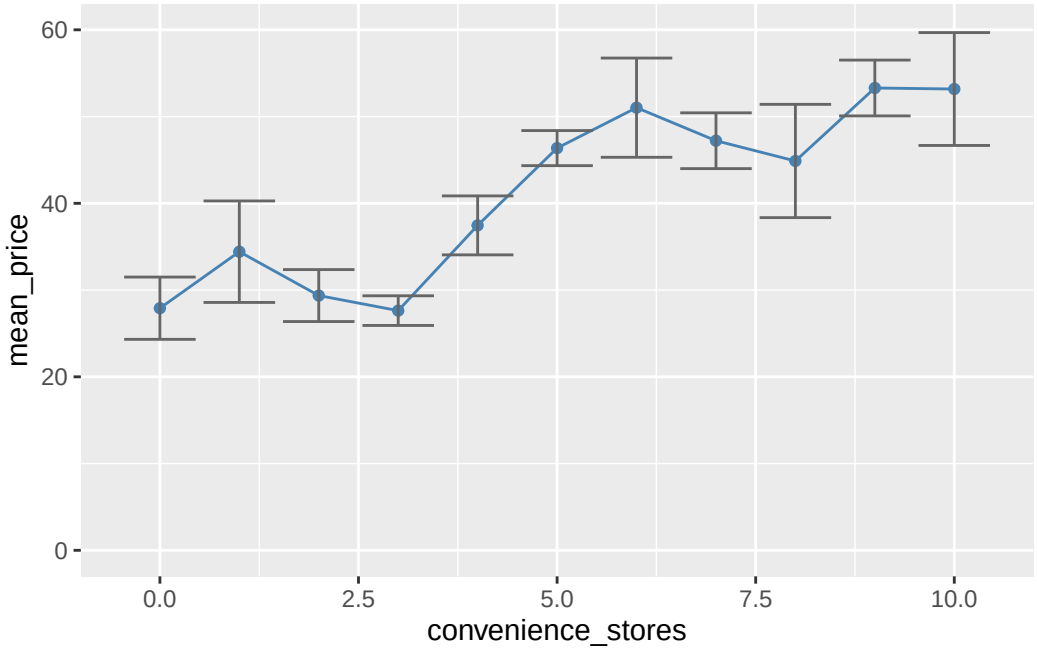
From the line plot,we see the house price increase slightly by time.

```
data %>%
  mutate(mrt_bin = cut(mrt_distance, breaks = 10)) %>%
  group_by(mrt_bin) %>%
  summarise(
    mean_price = mean(house_price, na.rm = TRUE),
    sd = sd(house_price, na.rm = TRUE),
    n = n(),
    se = sd / sqrt(n),
    lower = mean_price - 1.96 * se,
    upper = mean_price + 1.96 * se
  ) %>%
  ggplot(aes(x = mrt_bin, y = mean_price, group = 1)) +  # group = 1 is needed for lines o
  geom_line(color = "steelblue") +
  geom_point(color = "steelblue") +
  ylim(0,60)+
  geom_errorbar(aes(ymin = lower, ymax = upper), color = "gray40", width = 0.2) +
  theme_minimal()
```



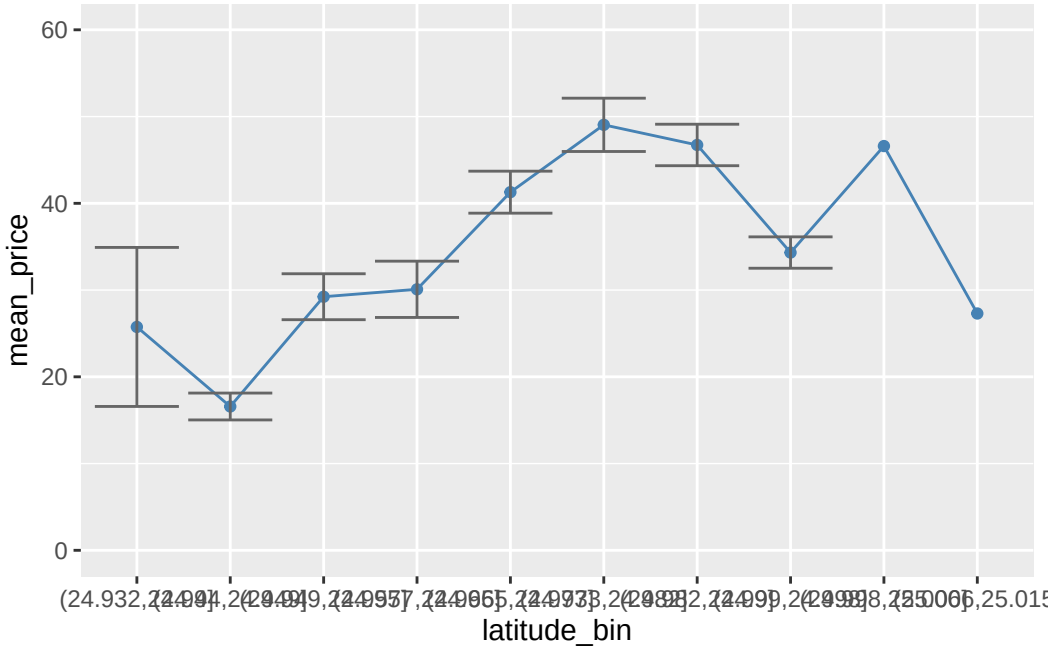
From the line plot, we can see that the mean price decreases as the distance to the nearest Mass Rapid Transit (MRT) station increases.

```
data%>%
  group_by(convenience_stores)%>%
  summarise(mean_price=mean(house_price),
            sd=sd(house_price),
            n=n(),
            se=sd/sqrt(n),
            lower=mean_price-1.96*se,
            upper=mean_price+1.96*se)%>%
  ggplot(aes(x=convenience_stores, y=mean_price))+
  geom_point(color="steelblue")+
  geom_line(color="steelblue")+
  ylim(0,60)+
  geom_errorbar(aes(ymin=lower, ymax=upper), color="gray40")
```



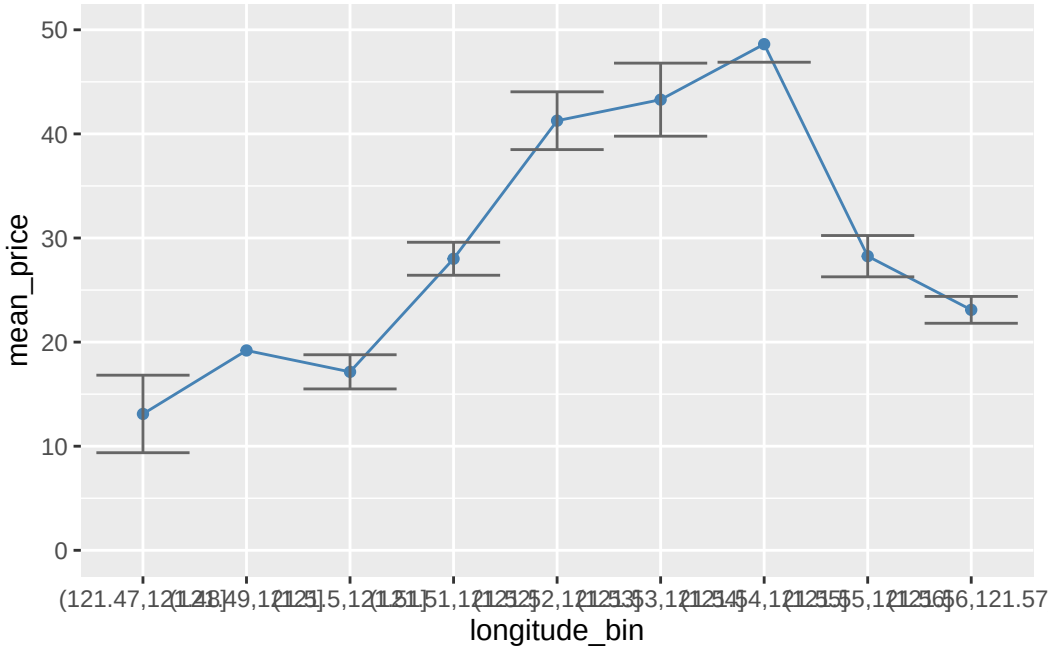
From the line plot, the more convenience stores close to the house, the more expensive the house is.

```
data%>%
  mutate(latitude_bin=cut(latitude, breaks=10))%>%
  group_by(latitude_bin) %>%
  summarise(mean_price=mean(house_price),
            sd=sd(house_price),
            n=n(),
            se=sd/sqrt(n),
            lower=mean_price-1.96*se,
            upper=mean_price+1.96*se,
            )%>%
  ggplot(aes(x=latitude_bin, y=mean_price,group=1))+
  geom_point(color="steelblue")+
  geom_line(color="steelblue")+
  ylim(0,60)+
  geom_errorbar(aes(ymin=lower, ymax=upper), color="gray40")
```



From the line plot, we can see that the house price increases with the increase in latitude.

```
data%>%
  mutate(longitude_bin=cut(longitude, breaks=10))%>%
  group_by(longitude_bin)%>%
  summarise(mean_price=mean(house_price),
            sd=sd(house_price),
            n=n(),
            se=sd/sqrt(n),
            lower=mean_price-1.96*se,
            upper=mean_price+1.96*se)%>%
 ggplot(aes(x=longitude_bin, y=mean_price, group=1))+
  geom_point(color="steelblue")+
  geom_line(color="steelblue")+
  geom_errorbar(aes(ymin=lower, ymax=upper), color="gray40")+
  ylim(0,50)
```



From the line plot, the house price slightly increases as the longitude increases.

4 Linear regression

4.1 Initial model

```
model1<-lm(house_price~., data=data)
summary(model1)
```

```
Call:
lm(formula = house_price ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-33.425  -5.105  -0.497   4.470  75.123

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.382e+04  7.351e+03  -1.880   0.0612 .
transaction_date   4.387e+00  1.874e+00   2.341   0.0200 *
house_age         -6.577e-01  9.343e-02  -7.040 1.59e-11 ***
mrt_distance      -4.539e-03  8.178e-04  -5.551 6.79e-08 ***
```

```
convenience_stores  9.562e-01  2.366e-01   4.041 6.94e-05 ***
latitude            2.201e+02  4.967e+01   4.432 1.36e-05 ***
longitude          -3.781e+00  5.077e+01  -0.074   0.9407
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.604 on 270 degrees of freedom
Multiple R-squared:  0.6321,    Adjusted R-squared:  0.6239
F-statistic: 77.31 on 6 and 270 DF,  p-value: < 2.2e-16
```

```
vif(model1)
```

```
  transaction_date          house_age       mrt_distance convenience_stores
          1.025332           1.127482           3.722308           1.657647
          latitude          longitude
          1.544453           2.226925
```

4.2 Model improvement

```
model2<-step(model1, direction=c("both"), trace=1)
```

```
Start:  AIC=1199.27
house_price ~ transaction_date + house_age + mrt_distance + convenience_stores +
    latitude + longitude

                     Df Sum of Sq   RSS    AIC
- longitude           1       0.4 19990 1197.3
<none>                            19990 1199.3
- transaction_date    1     405.6 20395 1202.8
- convenience_stores  1    1209.2 21199 1213.5
- latitude            1    1454.2 21444 1216.7
- mrt_distance        1    2281.1 22271 1227.2
- house_age           1    3668.8 23658 1243.9

Step:  AIC=1197.27
house_price ~ transaction_date + house_age + mrt_distance + convenience_stores +
    latitude

                     Df Sum of Sq   RSS    AIC
<none>                            19990 1197.3
+ longitude           1       0.4 19990 1199.3
- transaction_date    1     405.2 20395 1200.8
- convenience_stores  1    1217.4 21207 1211.7
```

```
- latitude            1    1490.5 21480 1215.2
- house_age           1    3670.8 23661 1242.0
- mrt_distance        1    3880.8 23871 1244.4
```

```
summary(model2)
```

```
Call:
lm(formula = house_price ~ transaction_date + house_age + mrt_distance +
    convenience_stores + latitude, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-33.411  -5.183  -0.458   4.452  75.169

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1.429e+04  3.850e+03  -3.711 0.000251 ***
transaction_date    4.384e+00  1.870e+00   2.344 0.019810 *
house_age          -6.578e-01  9.324e-02  -7.054 1.45e-11 ***
mrt_distance       -4.500e-03  6.204e-04  -7.253 4.27e-12 ***
convenience_stores  9.574e-01  2.357e-01   4.063 6.36e-05 ***
latitude            2.207e+02  4.909e+01   4.495 1.03e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.589 on 271 degrees of freedom
Multiple R-squared:  0.6321,    Adjusted R-squared:  0.6253
F-statistic: 93.12 on 5 and 271 DF,  p-value: < 2.2e-16
```

```
vif(model2)
```

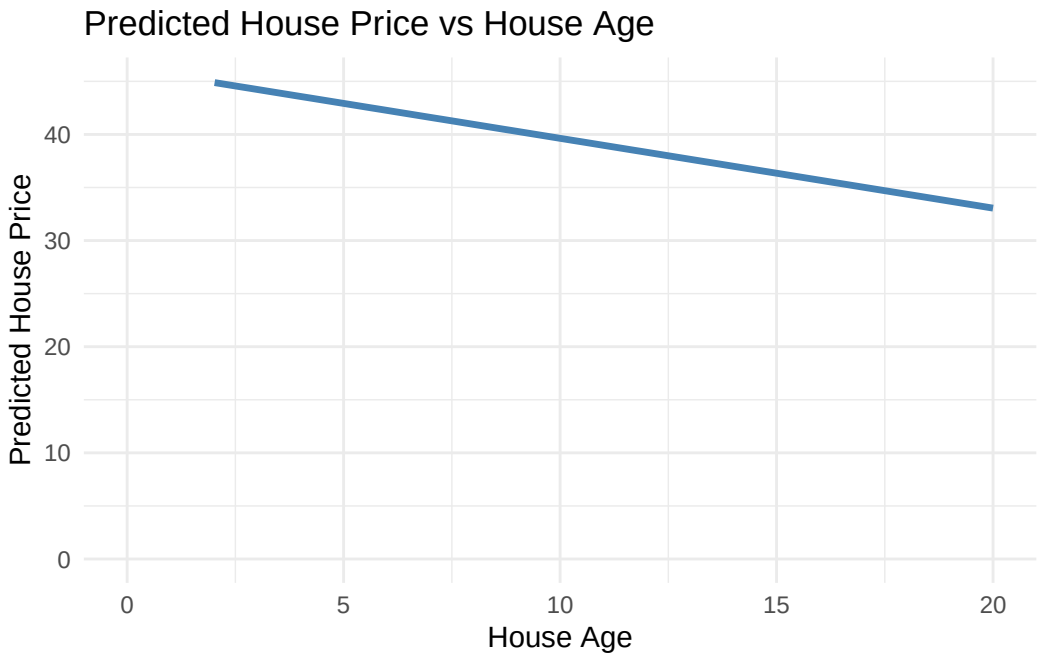
```
  transaction_date          house_age       mrt_distance convenience_stores
          1.024785           1.127227           2.150002           1.650441
          latitude
          1.514002
```

House prices increase with a later transaction date, a higher number of nearby convenience stores within walking distance, and higher latitude. However, house prices decrease with increasing house age and greater distance to the nearest MRT station.

4.3 Prediction plots

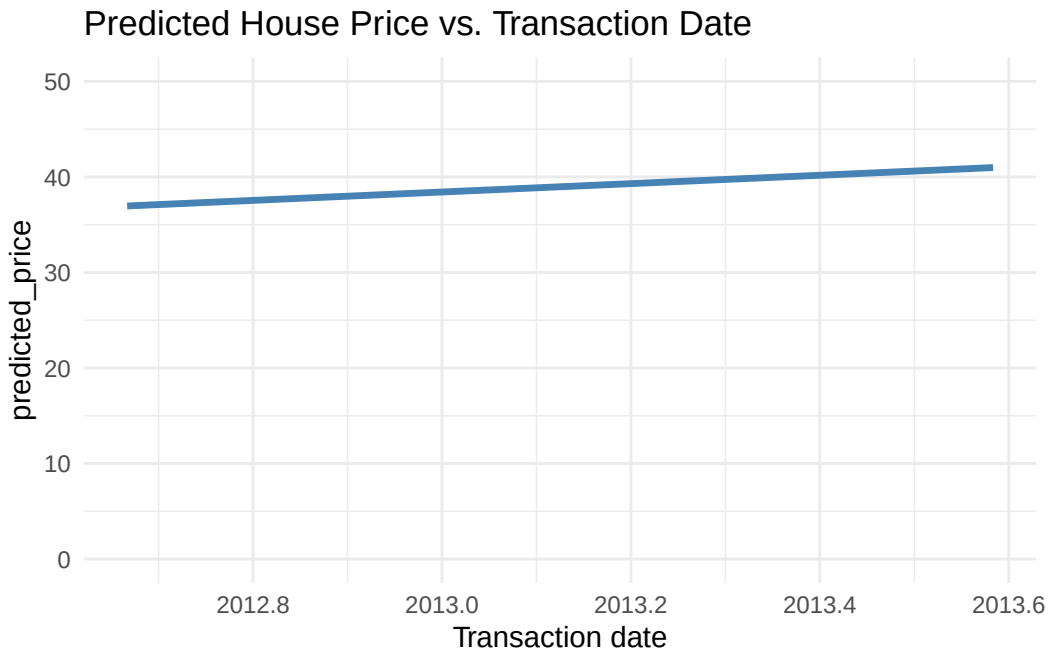
```r
# Create prediction data (same as above)
predict_data <- data.frame(
  transaction_date = mean(data$transaction_date),
  house_age = seq(min(data$house_age), max(data$house_age), length.out = 100),
  mrt_distance = mean(data$mrt_distance),
  convenience_stores = mean(data$convenience_stores),
  latitude = mean(data$latitude)
)

# Add predictions
predict_data$predicted_price <- predict(model2, newdata = predict_data)

# Plot
ggplot(predict_data, aes(x = house_age, y = predicted_price)) +
  geom_line(color = "steelblue", linewidth = 1.2) +
  labs(title = "Predicted House Price vs House Age",
       x = "House Age", y = "Predicted House Price") +
  theme_minimal()+
  ylim(0,45)
```

```
Warning: Removed 10 rows containing missing values or values outside the scale range
(`geom_line()`).
```



The predicted house price decreases as house age increases.

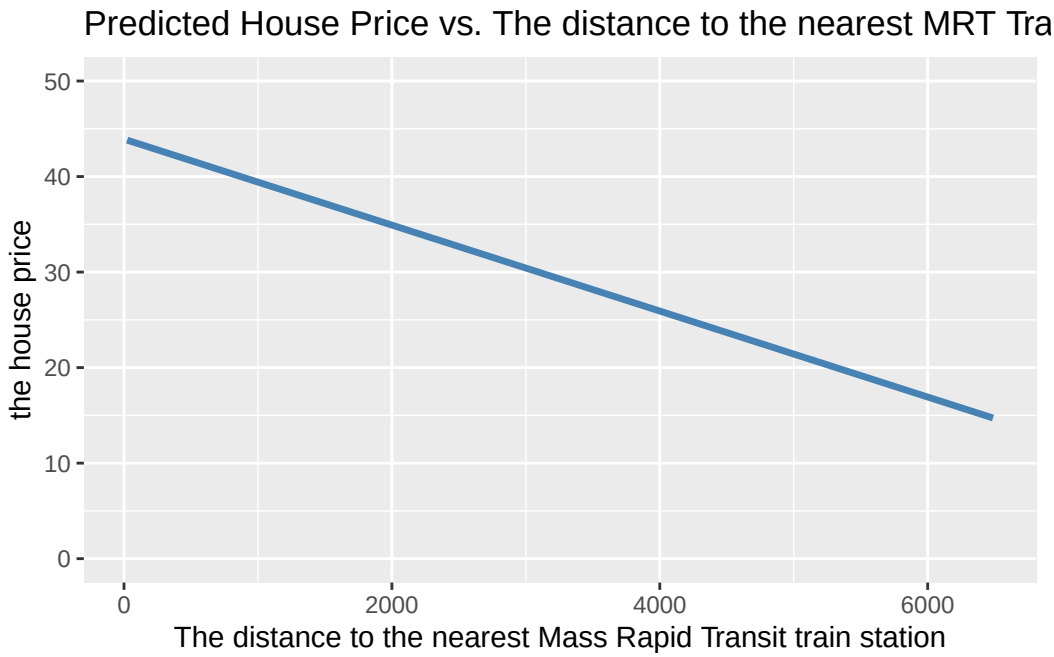
```
predict_data<-data.frame(
  transaction_date<-seq(min(data$transaction_date), max(data$transaction_date), length.out
  house_age=mean(data$house_age),
  mrt_distance=mean(data$mrt_distance),
  convenience_stores=mean(data$convenience_stores),
  latitude=mean(data$latitude)
)
predict_data$predicted_price<-predict(model2, newdata=predict_data)

ggplot(predict_data, aes(x=transaction_date,y=predicted_price))+
  geom_line(color="steelblue", linewidth=1.2)+
  ylim(0,50)+
  theme_minimal()+
  labs(x="Transaction date",
       title="Predicted House Price vs. Transaction Date")
```



The predicted price increases with a later transaction date.

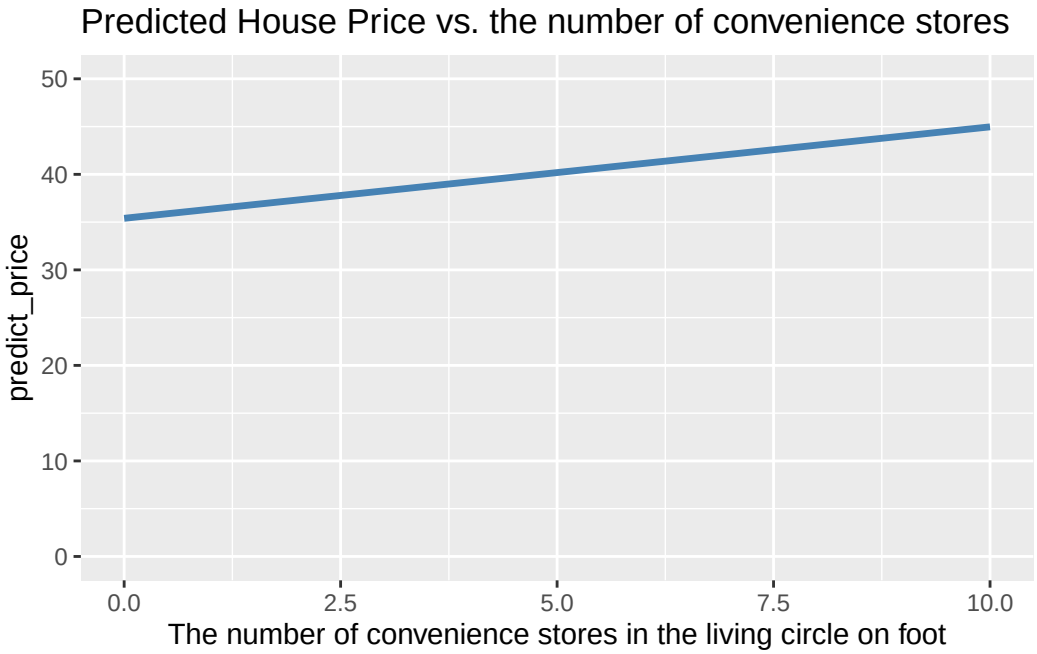
```
predict_data<-data.frame(
  transaction_date=mean(transaction_date),
  house_age=mean(data$house_age),
  mrt_distance=seq(min(data$mrt_distance), max(data$mrt_distance), length.out=100),
  convenience_stores=mean(data$convenience_stores),
  latitude=mean(data$latitude)
```

```
)
predict_data$predicted_price=predict(model2, predict_data)

ggplot(predict_data, aes(x=mrt_distance, y=predicted_price))+
  geom_line(color="steelblue", linewidth=1.2)+
  ylim(0, 50)+
  labs(x="The distance to the nearest Mass Rapid Transit train station",
       y="the house price",
       title="Predicted House Price vs. The distance to the nearest MRT Train Station")
```



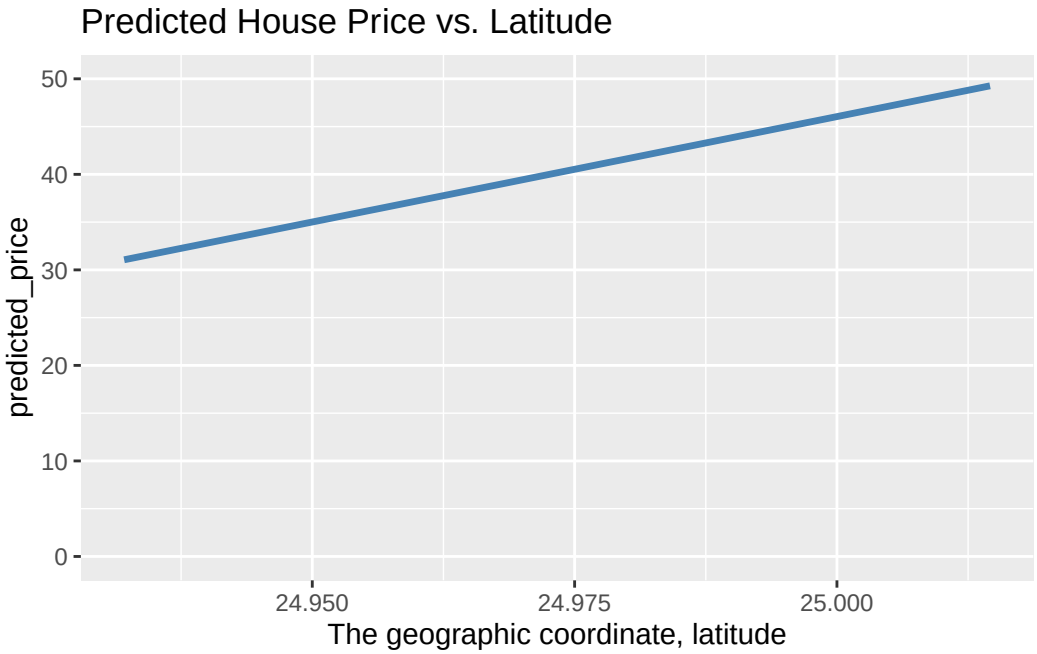
The closer a house is to the mass rapid transit (MRT) station, the higher its price.

```
predict_data<-data.frame(
  transaction_date=mean(data$transaction_date),
  house_age=mean(data$house_age),
  mrt_distance=mean(data$mrt_distance),
  convenience_stores=seq(min(data$convenience_stores), max(data$convenience_stores), lengt
  latitude=mean(data$latitude)
)
predict_data$predict_price=predict(model2, predict_data)

ggplot(predict_data, aes(x=convenience_stores, y=predict_price))+
  geom_line(color="steelblue", linewidth=1.2)+
  ylim(0,50)+
  labs(x="The number of convenience stores in the living circle on foot",
       title="Predicted House Price vs. the number of convenience stores")
```



The house price increases as the number of convenience stores increases.

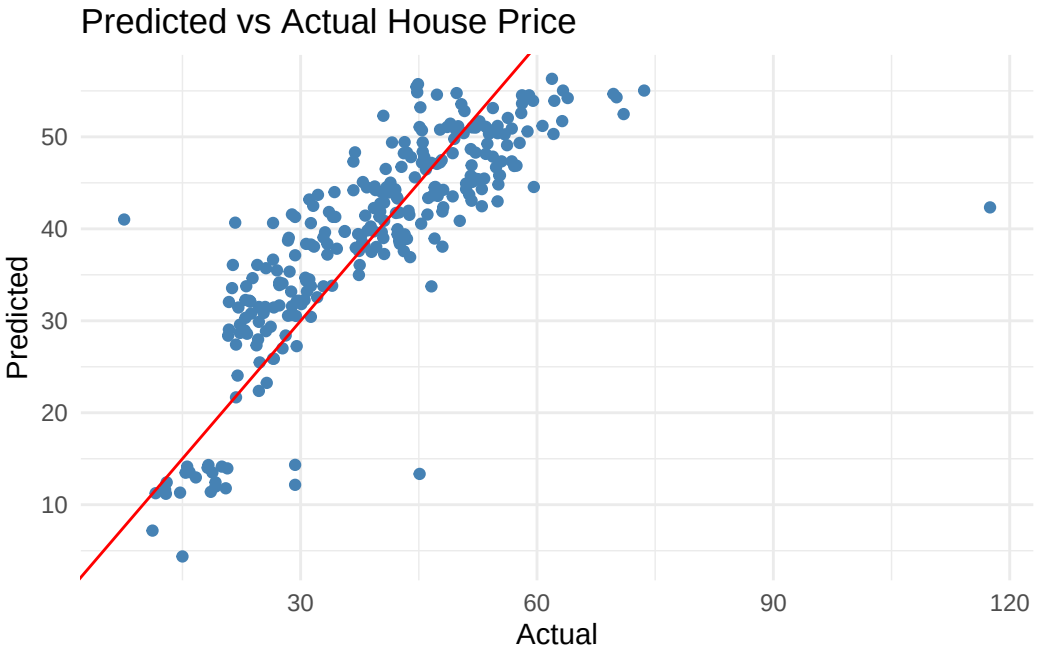
```
predict_data<-data.frame(
  transaction_date=mean(data$transaction_date),
  house_age=mean(data$house_age),
  mrt_distance=mean(data$mrt_distance),
  convenience_stores=mean(data$convenience_stores),
  latitude=seq(min(data$latitude), max(data$latitude), length.out=100)
)
predict_data$predicted_price=predict(model2, predict_data)

ggplot(predict_data, aes(x=latitude, y=predicted_price))+
  geom_line(color="steelblue", linewidth=1.2)+
  ylim(0,50)+
  labs(x="The geographic coordinate, latitude",
       title="Predicted House Price vs. Latitude")
```



As latitude increases, house prices increase.

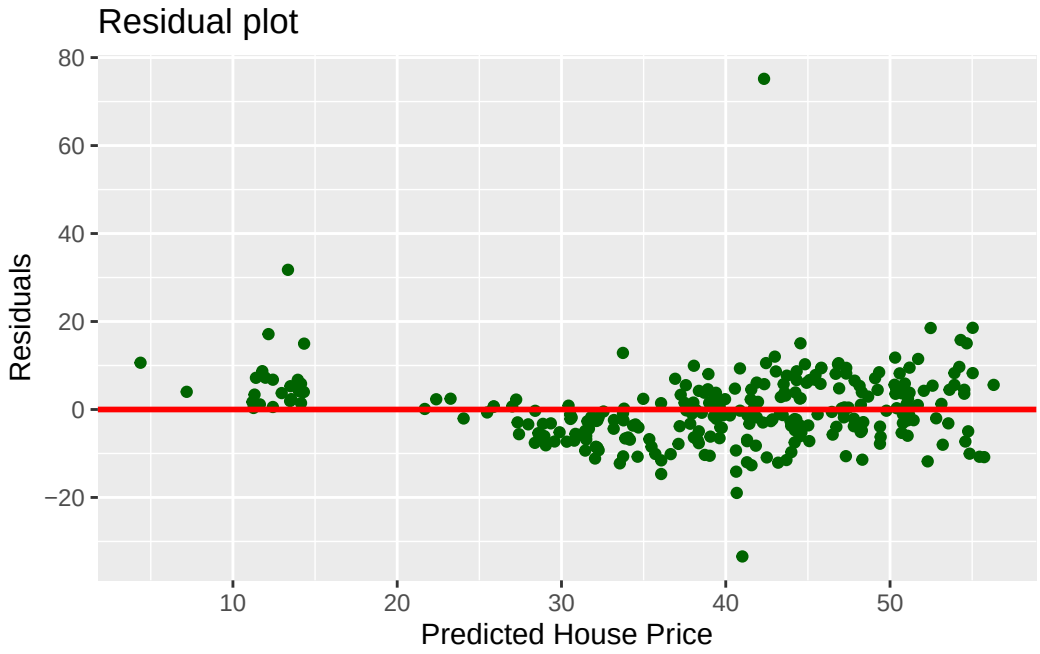
4.4 Predicted vs. Actual House Price

```
# Add predicted values
data$predicted <- predict(model2)

# Plot with ggplot2
ggplot(data, aes(x = house_price, y = predicted)) +
  geom_point(color = "steelblue") +
  geom_abline(slope = 1, intercept = 0, color = "red") +
  labs(title = "Predicted vs Actual House Price",
       x = "Actual",
       y = "Predicted") +
  theme_minimal()
```




```
data$predicted <-predict(model2)
data$residuals <- residuals(model2)

ggplot(data, aes(x=predicted, y=residuals))+
  geom_point(color="darkgreen")+
  geom_hline(yintercept=0, color="red", linetype="solid", size=1)+
  labs(x="Predicted House Price",
       y="Residuals",
       title="Residual plot")
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



4.5 Model validation

4.5.1 Root Mean Square Error (RMSE)

```
predicted <- predict(model2)
actual <- data$house_price
rmse <- sqrt(mean((predicted - actual)^2))
rmse
```

```
[1] 8.495034
```

4.5.2 Mean Absolute Error (MAE)

```
mae <- mean(abs(predicted - actual))
mae
```

```
[1] 5.855627
```

4.5.3 Mean Absolute Percentage Error (MAPE)

```r
mape <- mean(abs((actual - predicted) / actual)) * 100
mape
```

```
[1] 17.71169
```

On average, the model's predictions are within 17.71% of the actual house prices.