Regression modelling for prediction of House Prices

Sanjana Suresh

Sanjana Suresh

Dataset: Factors affecting housing prices in an area

About:

Size of dataset: 528 KB

Features:

The dataset consists of the attributes such as

1. Number of bedrooms, bathrooms, floors

2. Condition - Rated in a range from 1 to 5

2. Square feet of the house

3. Year built

4. Location

5. Price of the property, etc..

All of the attributes are real in this data

Step 1: Load & View the Data

```
#Loading house prices csv file and loading libraries
options(repos = "https://cran.rstudio.com/")
housing_csv<-read.csv("/Users/sanjana/Desktop/MS/UTA/PS/Dataset/housePricing.csv")
library(ggplot2)
library(lmtest)
```

```
## Loading required package: zoo
```

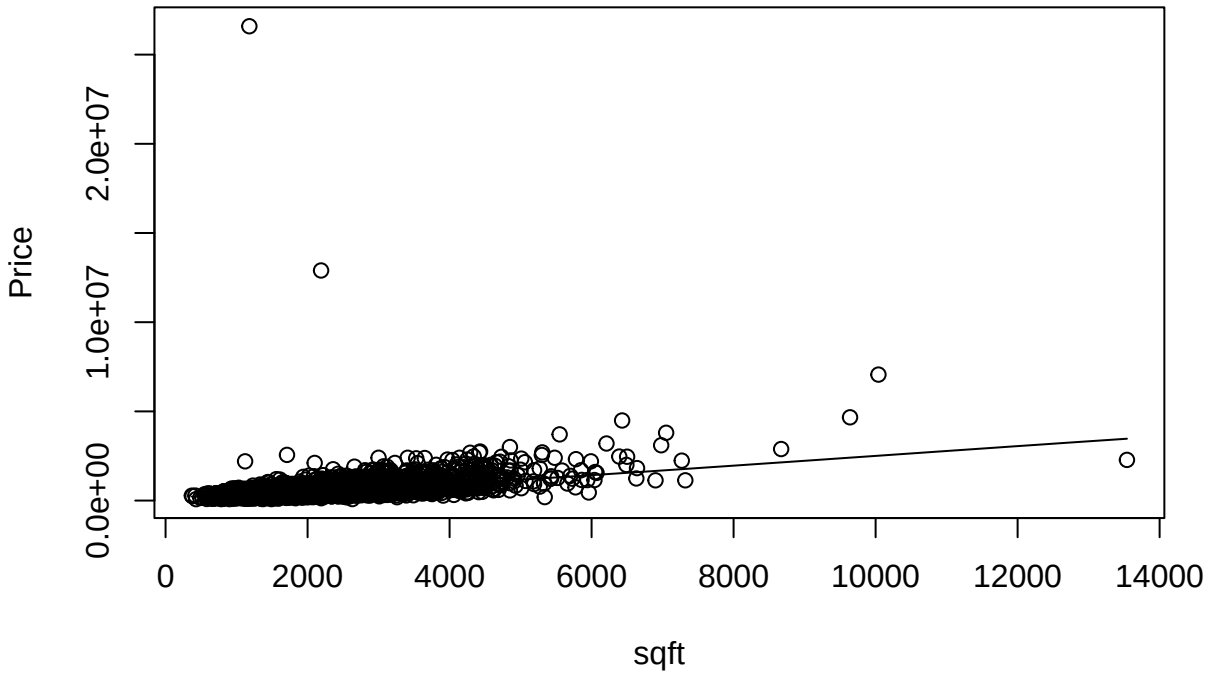
```
##
## Attaching package: 'zoo'
```

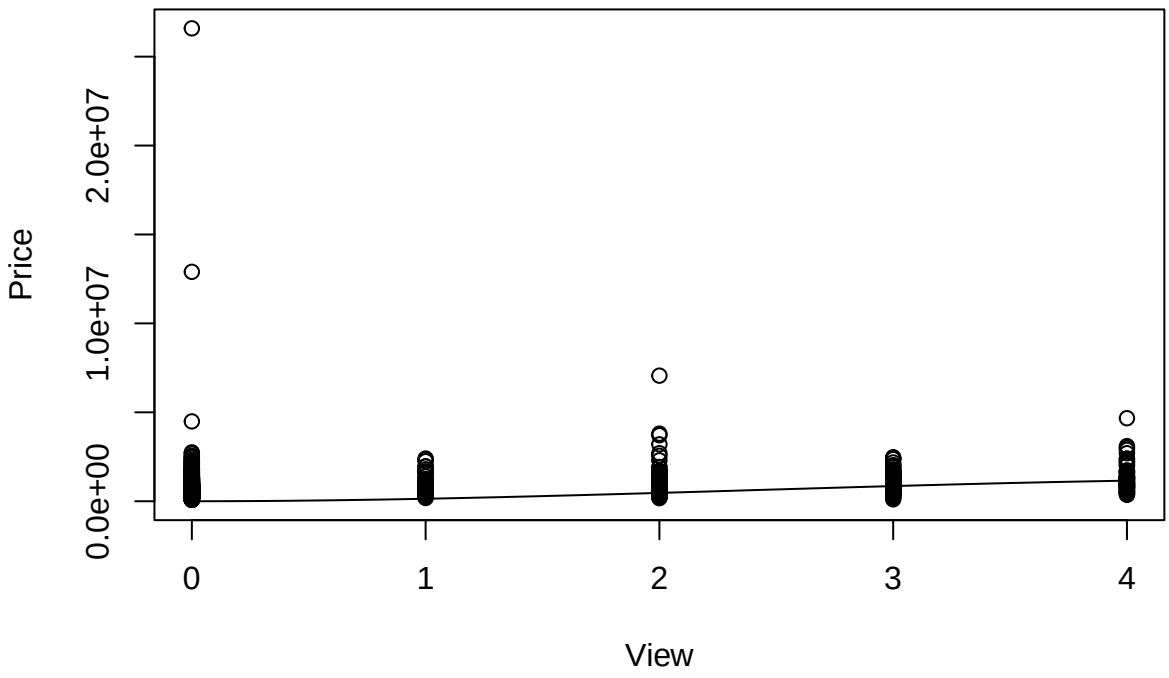
```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(MASS)
```

```
suppressWarnings({
scatter.smooth(housing_csv$sqft_living, housing_csv$price,xlab = "sqft", ylab = "Price")

scatter.smooth(housing_csv$view, housing_csv$price, xlab = "View", ylab = "Price")})
```





- During EDA of the House price prediction dataset, it was observed that

1. square feet has a strong correlation with the prices based on the scatter plot
2. Floor, bedrooms and view have a relationship which was observed from the correlation matrix.
3. **Question 1:** The above scatter plots also indicate a linear relationship between factors such as sqft_living and view with the price

Lets us perform multiple linear regression for the given dataset using the following factors(Chosen based on EDA as important features): 1. sqft_living 2. floor 3. bedrooms 4. view

```
# Perform multiple linear regression
model <- lm(price ~ sqft_living  + bedrooms+ floors+ view, data = housing_csv)

cat("Interpretation:\n")
```

```
## Interpretation:
```

```
# Print summary of the regression model
summary(model)
```

```
##
## Call:
## lm(formula = price ~ sqft_living + bedrooms + floors + view,
##     data = housing_csv)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1476776  -144644   -25519    90903 26290970
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114778.83   33446.18   3.432 0.000605 ***
## sqft_living    267.09      10.67  25.036  < 2e-16 ***
## bedrooms    -45937.56   10262.88  -4.476 7.79e-06 ***
## floors        6897.35   14728.25   0.468 0.639587
## view         81616.71   10269.22   7.948 2.38e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 500200 on 4545 degrees of freedom
## Multiple R-squared:  0.214,  Adjusted R-squared:  0.2133
## F-statistic: 309.4 on 4 and 4545 DF,  p-value: < 2.2e-16
```

```
# Extract coefficients
coefficients <- coef(model)
coefficients
```

```
## (Intercept) sqft_living    bedrooms      floors        view
## 114778.8300    267.0902 -45937.5586   6897.3537  81616.7088
```

Question 2: Obtain Equation of the line:

```
# Get the names of independent variables
independent_vars <- names(coefficients)[-1]  # Excluding intercept

# Create the equation string
equation <- paste("y =", coefficients[1], paste(paste0("+", coefficients[-1], "*", independent_vars), c
```

Question 3: Write the equation of the regression line

```
# Print equation of the line
cat("Equation of the line:\n")
```

```
## Equation of the line:
```

```
cat(equation, "\n")
```

```
## y = 114778.830006796 +267.090164120658*sqft_living +-45937.5586223139*bedrooms +6897.35370269307*floo
```

Question 4: How do you interpret the intercept?

```
summary(model)
```

```
##
## Call:
## lm(formula = price ~ sqft_living + bedrooms + floors + view,
##     data = housing_csv)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1476776  -144644   -25519    90903 26290970
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114778.83   33446.18   3.432 0.000605 ***
## sqft_living    267.09      10.67  25.036  < 2e-16 ***
## bedrooms    -45937.56   10262.88  -4.476 7.79e-06 ***
## floors        6897.35   14728.25   0.468 0.639587
## view         81616.71   10269.22   7.948 2.38e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 500200 on 4545 degrees of freedom
## Multiple R-squared:  0.214,  Adjusted R-squared:  0.2133
## F-statistic: 309.4 on 4 and 4545 DF,  p-value: < 2.2e-16
```

The intercept value is 114778.83. It is the average price of the house when the values of independent variables such as sqft_living,no. of bedrooms, no. of floors and view is equal to 0.

Question 5: How do you interpret the slope?

SLOPES:

1. sqft_Living = 267.09
2. no. of bedrooms = -45937.56
3. no. of floors = 6897.35
4. view = 81616.71

From the above slopes we understand that

1. For a unit increase in the sqft_living, the price of the house increases by 267.09 dollars
2. When the no. of bedrooms increases by 1, the price of the house decreases by 45937 dollars approximately for the given data
3. As the number of floors in the house increases, price of the house increases by 6897.35 dollars
4. For every improvement in the view, the house proce increases by 81616.71 dollars.

Question 6: Are the coefficients statistically significant? Significance based on values: Factors with p value<0.05 are statistically significant. Based on this we infer that: 1. sqft_living

2. bedrooms
3. view

are all statistically significant. Since floors has a p-value>0.05, it is NOT statistically significant.

Question 6 i):What is the null and alternative hypothesis that you are testing?

Based on p-value coefficient:

Null Hypothesis1:

sqft_living H0 -> Coefficient of sqft_living = 0

Alternative Hypothesis1:

sqft_living H1 -> Coefficient of sqft_living not equal to 0

Null Hypothesis2:

Bedrooms H0 -> Coefficient of Bedrooms = 0

Alternative Hypothesis2:

Bedrooms H1 -> Coefficient of Bedrooms not equal to 0

Null Hypothesis3:

Bedrooms H0 -> Coefficient of Floors = 0

Alternative Hypothesis3:

Bedrooms H1 -> Coefficient of Floors not equal to 0

Null Hypothesis4:

View H0 -> Coefficient of view = 0

Alternative Hypothesis4:

View H1 -> Coefficient of view not equal to 0

Based on intercept coefficient:

Intercept H0 -> Intercept Coefficient = 0

Intercept H1-> Intercept coefficient not equal 0

Question 6 ii):What are your conclusions and why?

If p-value of coefficient < 0.05, we reject the H0 and conclude H1 and say that the coefficient is significant otherwise we say that we don't have sufficient evidence to conclude that the coefficient is significant

- For intercept: p-value is 0.000605 and hence we reject the null hypothesis.
- For sqft_living: p-value is nearly 0 and hence we reject H0 and conclude the coefficient is significant
- For bedrooms: p-value is nearly 0 and hence we reject H0 and conclude the coefficient is significant
- For floors: p-values is more than 0.05 and hence we fail to reject H0 and hence we don't have sufficient evidence to conclude that the coefficient is significant
- For view: p-value is nearly 0 and hence we reject H0 and conclude coefficient is significant

Question 7: What is the variance of the model?

variance = (Residual standard error)^2

```
RSE = 500200
Variance = (RSE)^2
cat("The variance of the model is:\n")
```

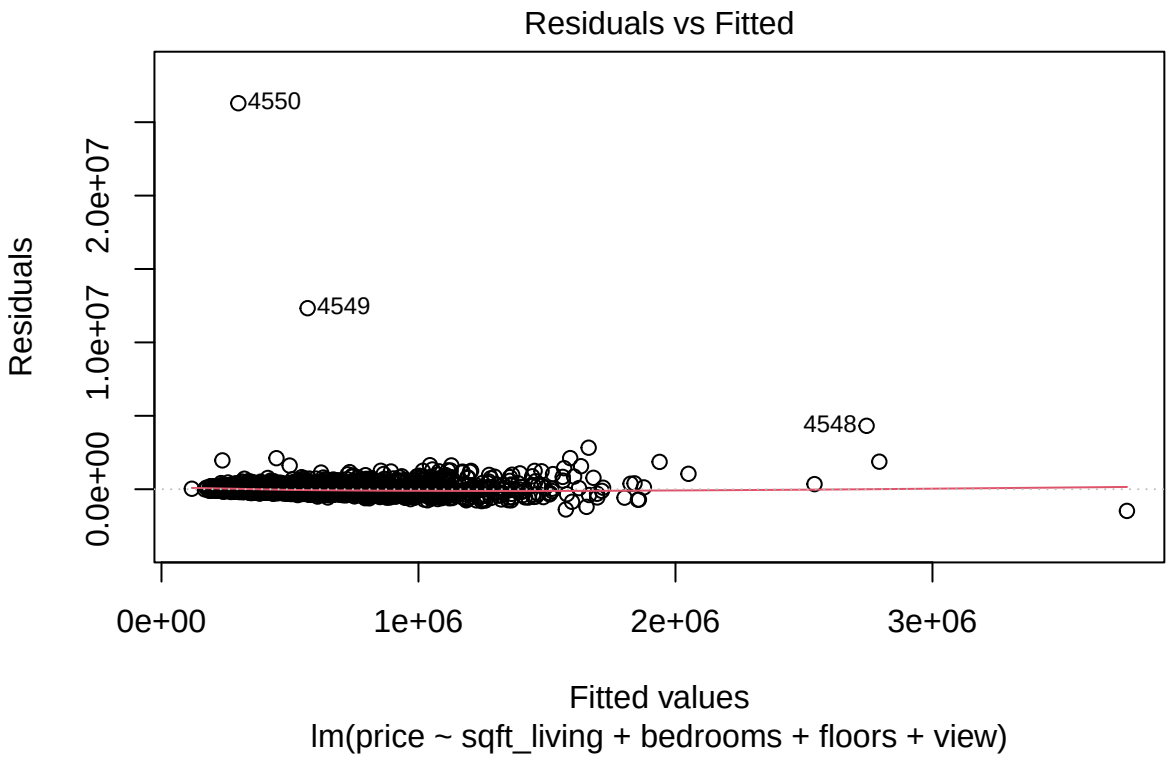
```
## The variance of the model is:
```

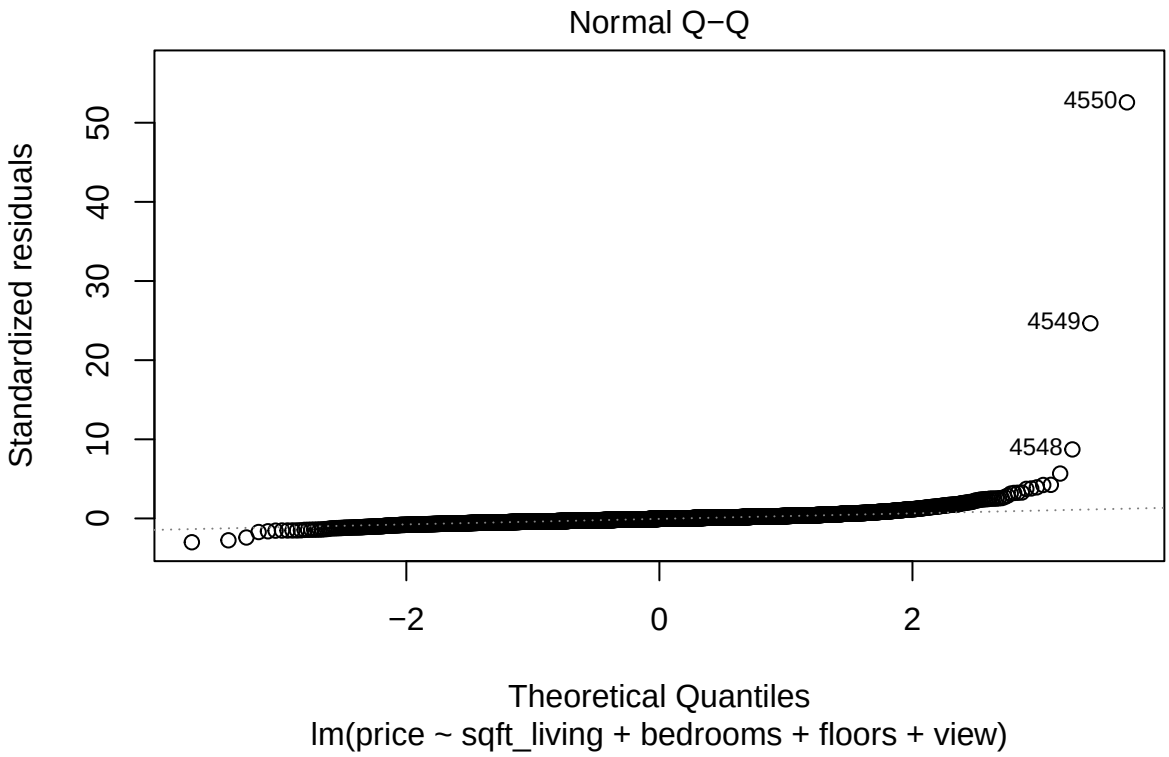
```
cat(Variance, "\n")
```

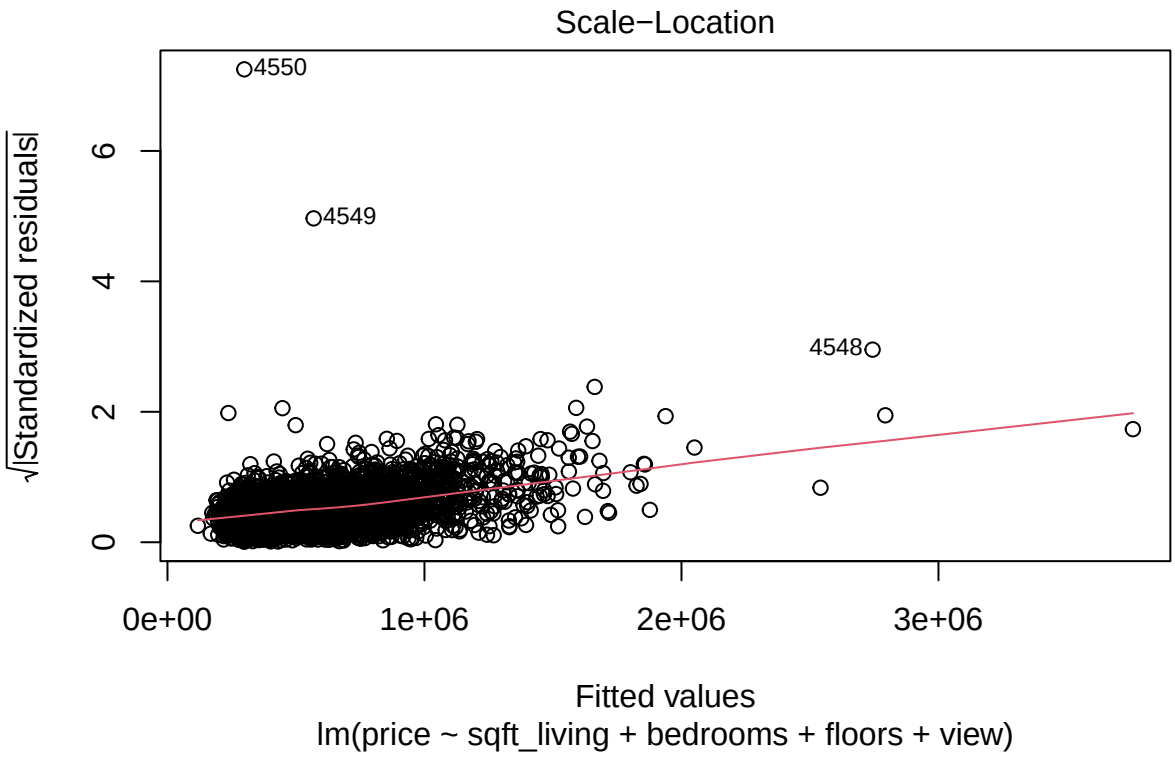
```
## 2.502e+11
```

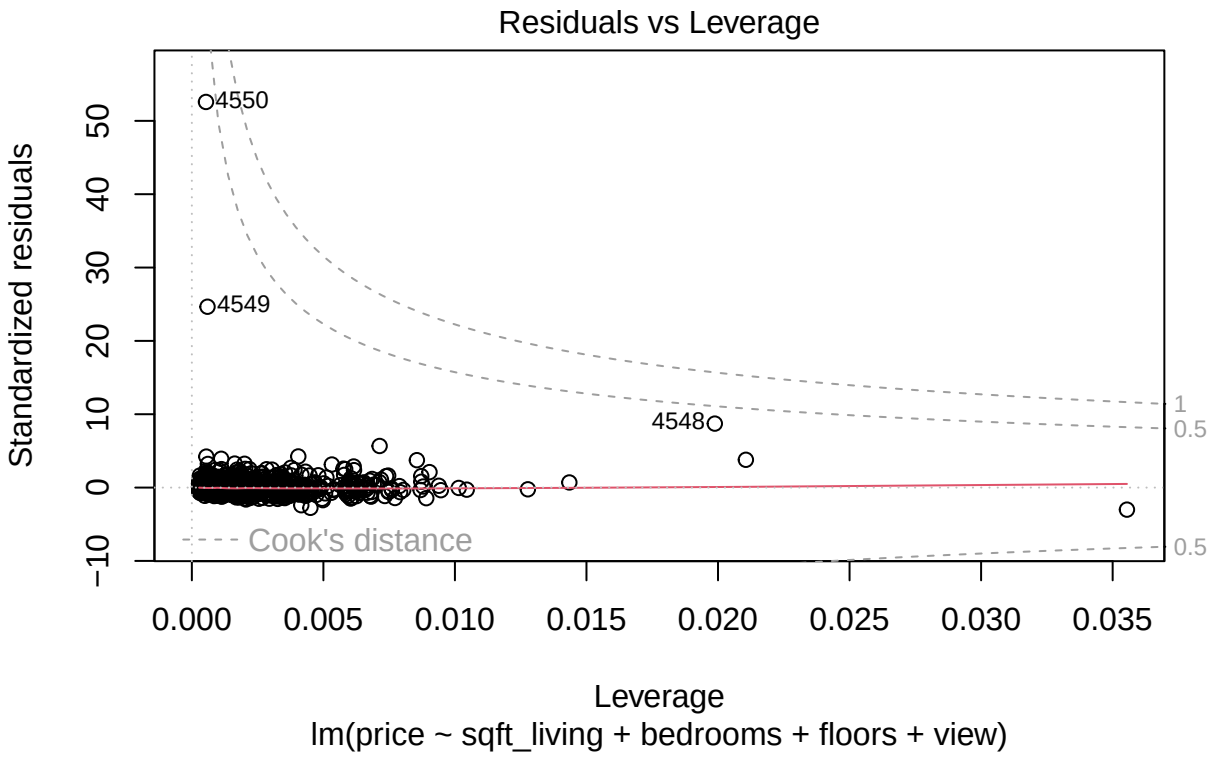
Question 8: Plot the regression fitted line on the scatterplot.

```
plot(model)
```




Normal Q−Q
Standardized residuals
0
10
20
30
40
50
4550
4549
4548
−2
0
2
Theoretical Quantiles
lm(price ~ sqft_living + bedrooms + floors + view)


Scale−Location
4550
4549
4548
√|Standardized residuals|
0
2
4
6
0e+00
1e+06
2e+06
3e+06
Fitted values
lm(price ~ sqft_living + bedrooms + floors + view)



Question 9,10: Model Assumptions

1. Linearity - The relationship between the independent variables and the dependent variable should be linear.
2. Independence - The residuals (the differences between observed and predicted values) should be independent of each other. T
3. Homoscedasticity- Residuals should have constant variance across all levels of the independent variables.
4. Normality of Residuals - The residuals should be normally distributed. Departures from normality might affect the statistical tests associated with the model coefficients.

Question 11: How do you test for them? Do they hold?

To check if they hold, lets analyse the graphs from the above

1. Test for Linearity-From the Residual vs fitted plot, we can see that the points are randomly scattered around the zero indicating linearity
2. Test for Independence - Durbin-Watson test

```
dwtest(model)
```

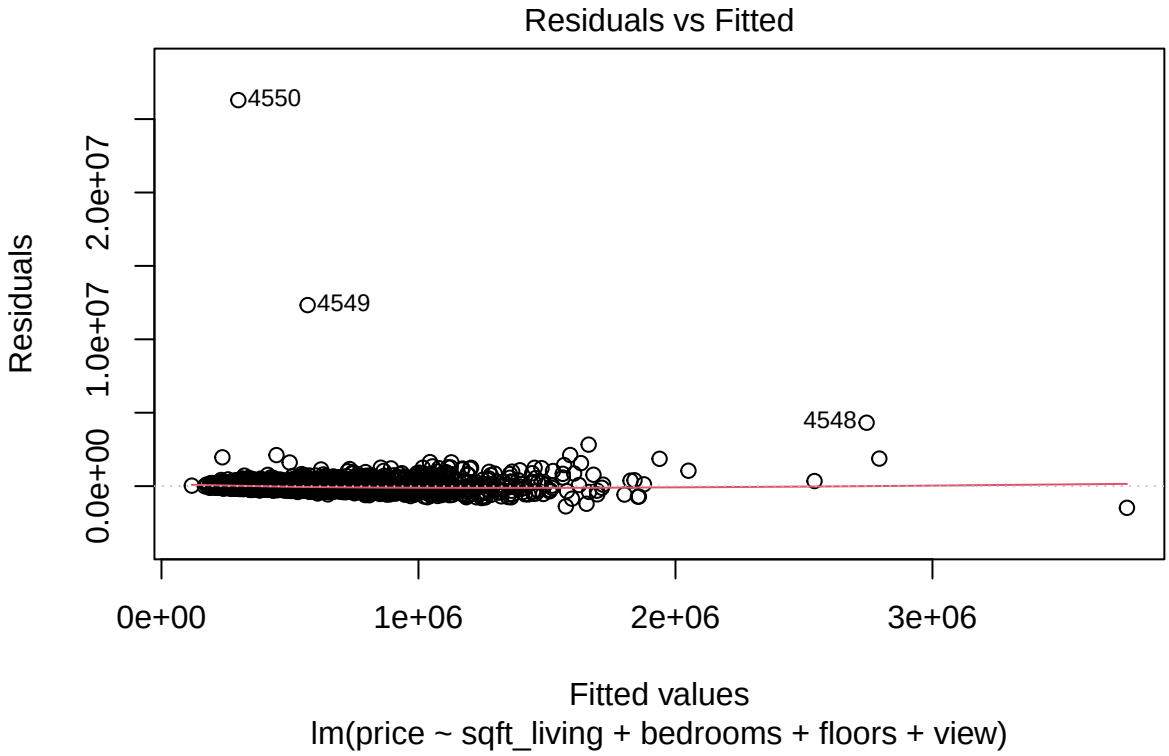
```
##
##  Durbin-Watson test
##
```

```
## data:  model
## DW = 0.48189, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

A value around 2 of the DW test indicates no autocorrelation and thereby they are independent. Here, DW=0.48189 which is not around 2 and hence independence does not hold for the given model

3. Test for Homoscedasticity - Breusch-Pagan test

```
plot(model, which=1)
```



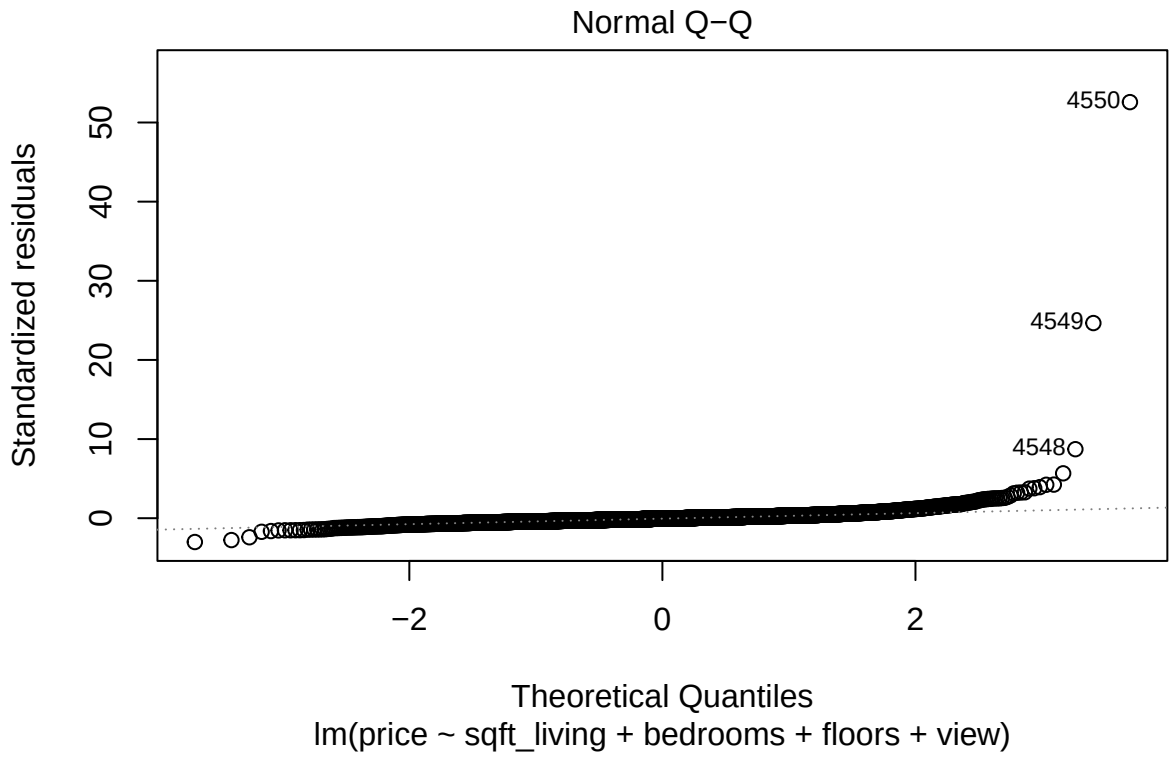

```
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 1.0172, df = 4, p-value = 0.9072
```

- Since the p-value as a result of the Breusch-Pagan test is 0.9072 which is high, it indicates that it holds Homoscedasticity.

4. Test for Normality of Residuals

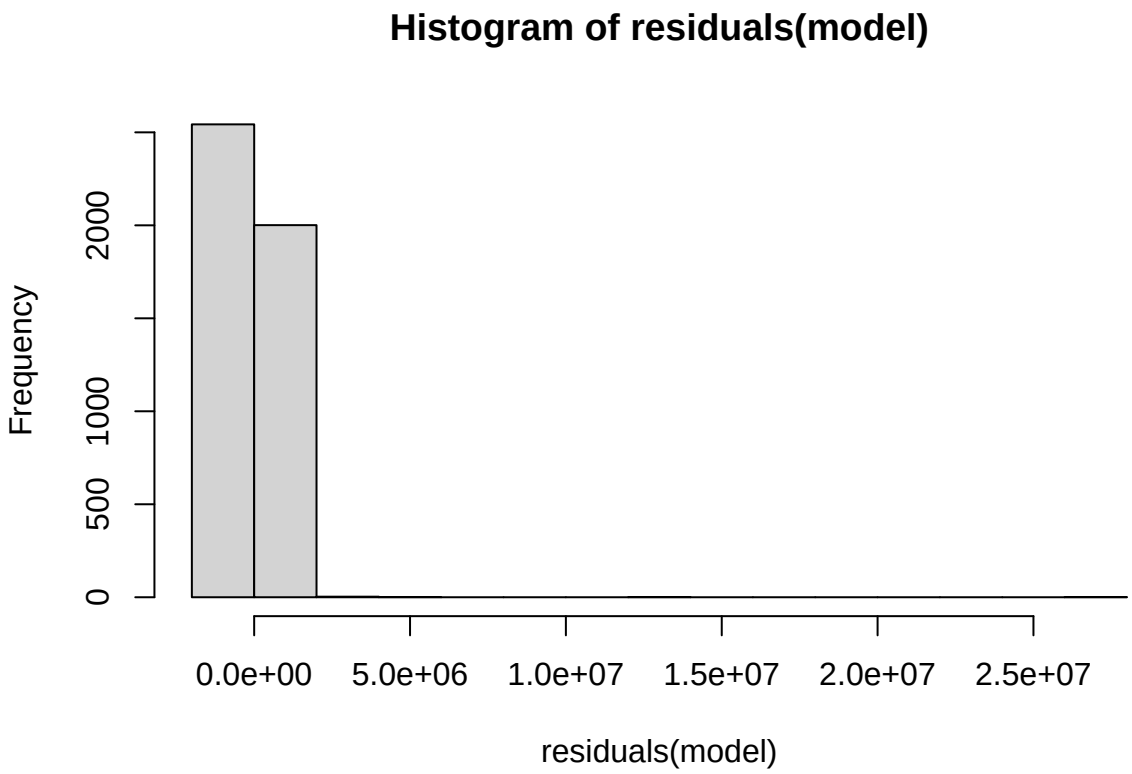
```
plot(model, which=2)
```




```
shapiro.test(residuals(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.27376, p-value < 2.2e-16
```

```
hist(residuals(model))
```



Based on 1. Q-Q plot- Though majority points are around 0, there are a few outliers that disables us from concluding that we can conclude normality

2. Shaporo Test- Low value of p indicates that there is no normality
3. Histogram of Residuals- The histogram also clearly indicates the model does not hold normality

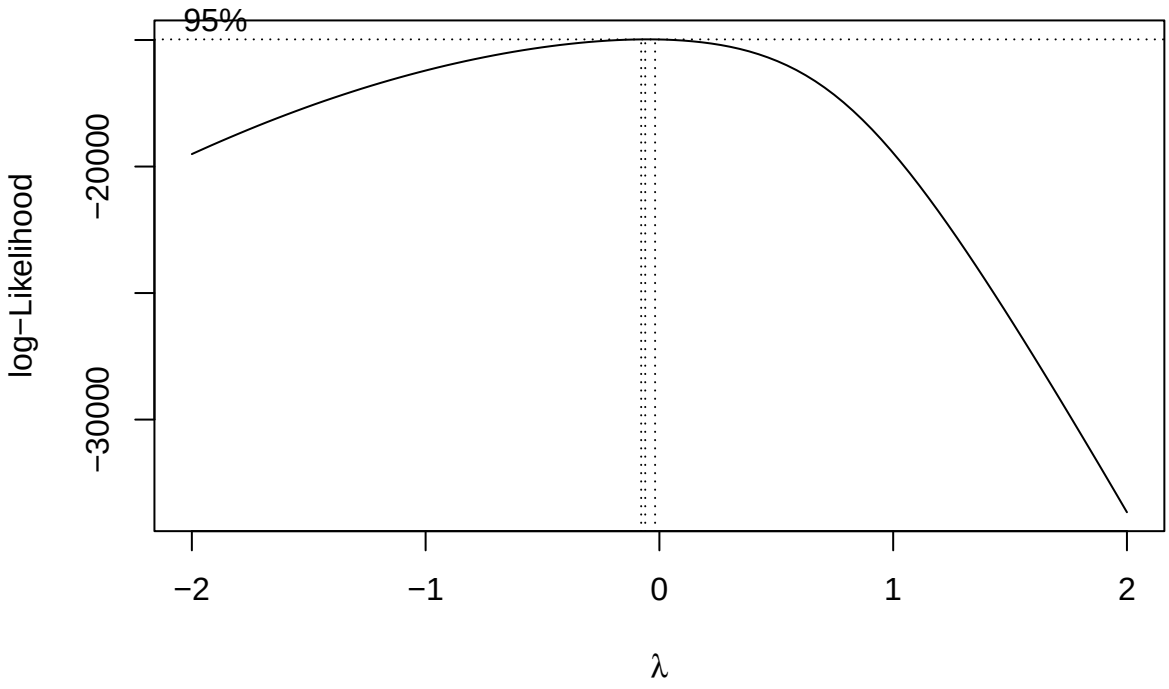
Question 12:Do you see outliers?

Based on the Q-Q plot, we can observe that there are multiple outliers

Question 13:What does the Box-Cox transformation suggest you do?

Plot Box-Cox transformation for different lambda values

```
bc <- boxcox(price~sqft_living, data=housing_csv)
```




```
lambda <- bc$x[which.max(bc$y)]
lambda
```

```
## [1] -0.06060606
```

Since lambda is negative i.e -0.0606, a mirrored-transformation is suggested as it would help to stabilise variance and normalise the data.

Question 14:What is the correlation between the independent and dependent variable?

Correlation is the square root of the R^2 which we have from the summary table as 0.214

Therefore correlation = 0.04578 ‘

Question 15:What is the value of the coefficient of determination and how do you interpret it?

Coefficient of Determination = Multiple R-squared From the summary table, we know that the Multiple R-squared=0.214

Question 15: Do you have any other suggestions to improve your model?

- During the box-cox transformation, we got a negative lambda value and hence a suggestion would be to apply a mirror transformation by taking the reciprocal of the lambda and taking the absolute value of the same so as to get a better model
- Since there are multiple outliers in the dataset, a suggestion would be to work on cleaning up the data and optimising so as to get a better prediction
- Transform the continous variables