Housing Analysis Project R

Sheng Mai

2025-03-01

Introduction

This report analyzes housing prices in King County using a dataset that includes 21,613 observations with various attributes such as price, square footage, number of bedrooms and bathrooms, house condition, and geographic location. Project goal: understanding the key drivers of house prices. The results reveal that square footage and grade have the strongest positive impact on price, while additional bedrooms unexpectedly correlate with lower prices—perhaps indicating that larger homes with more open space are more valuable than those with many small rooms.

Key Findings 1.House Prices Distribution The median house price is $450,000, with an average of $540,088. The most expensive property is valued at $7.7 million, while the least expensive is $75,000. The price distribution is likely right-skewed due to the high maximum value.

2.Size and Layout of Homes The average living area is 2,080 sqft, while the median is 1,910 sqft, indicating that a few large properties skew the mean. The number of bedrooms typically ranges from 3 to 4, with a maximum of 33, which suggests outliers. Most homes have 1.75 to 2.5 bathrooms, with a maximum of 8.

3.Property Conditions and Grades The condition of homes ranges from 1 (poor) to 5 (excellent), with most homes rated around 3 or 4. The grade, which measures overall construction and design quality, ranges from 1 to 13, with a median of 7.

4.Geographical Trends Properties are spread across ZIP codes ranging from 98001 to 98199. The latitude and longitude data indicate the dataset covers a wide range of locations within King County.

5.Renovation Trends Most homes were built between 1951 and 1997 (interquartile range). The majority of homes have not been renovated, as the median renovation year is 0 (not renovated).

6.Waterfront and View Influence Only 0.75% of homes have waterfront access, which may indicate a premium pricing factor. The view rating ranges from 0 to 4, with most homes having a rating of 0 (no significant view).

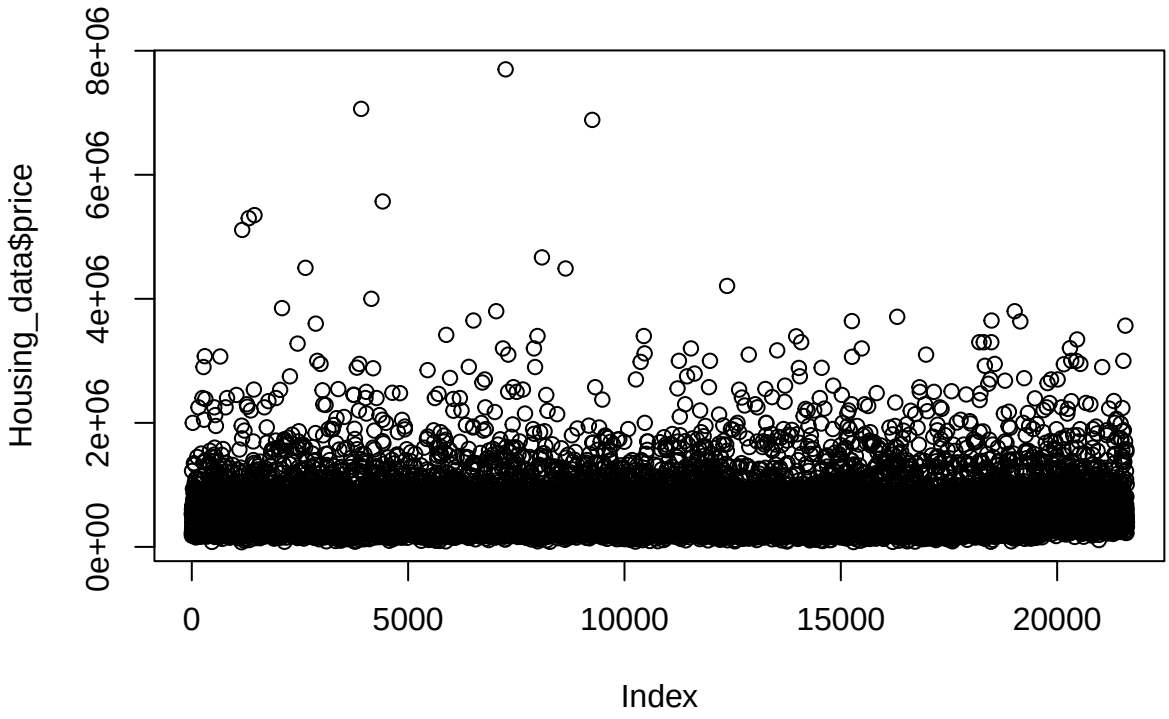
We will explore the data, find correlation between variables, and explore how different factors influence house prices.

```
Housing_data <- read.csv("C:\\Users\\15157\\Desktop\\Work\\project1\\kc_house_data.csv")
summary(Housing_data)
```

```
##        id                 date               price            bedrooms
##  Min.   :1.000e+06   Length:21613       Min.   :  75000   Min.   : 0.000
##  1st Qu.:2.123e+09   Class :character   1st Qu.: 321950   1st Qu.: 3.000
##  Median :3.905e+09   Mode  :character   Median : 450000   Median : 3.000
##  Mean   :4.580e+09                      Mean   : 540088   Mean   : 3.371
##  3rd Qu.:7.309e+09                      3rd Qu.: 645000   3rd Qu.: 4.000
##  Max.   :9.900e+09                      Max.   :7700000   Max.   :33.000
##    bathrooms      sqft_living       sqft_lot           floors
```

```
##  Min.   :0.000   Min.   :  290   Min.   :    520   Min.   :1.000
##  1st Qu.:1.750   1st Qu.: 1427   1st Qu.:   5040   1st Qu.:1.000
##  Median :2.250   Median : 1910   Median :   7618   Median :1.500
##  Mean   :2.115   Mean   : 2080   Mean   :  15107   Mean   :1.494
##  3rd Qu.:2.500   3rd Qu.: 2550   3rd Qu.:  10688   3rd Qu.:2.000
##  Max.   :8.000   Max.   :13540   Max.   :1651359   Max.   :3.500
##    waterfront            view          condition         grade
##  Min.   :0.000000   Min.   :0.0000   Min.   :1.000   Min.   : 1.000
##  1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.: 7.000
##  Median :0.000000   Median :0.0000   Median :3.000   Median : 7.000
##  Mean   :0.007542   Mean   :0.2343   Mean   :3.409   Mean   : 7.657
##  3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 8.000
##  Max.   :1.000000   Max.   :4.0000   Max.   :5.000   Max.   :13.000
##    sqft_above   sqft_basement       yr_built     yr_renovated
##  Min.   : 290   Min.   :   0.0   Min.   :1900   Min.   :   0.0
##  1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.0
##  Median :1560   Median :   0.0   Median :1975   Median :   0.0
##  Mean   :1788   Mean   : 291.5   Mean   :1971   Mean   :  84.4
##  3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:   0.0
##  Max.   :9410   Max.   :4820.0   Max.   :2015   Max.   :2015.0
##     zipcode           lat             long        sqft_living15
##  Min.   :98001   Min.   :47.16   Min.   :-122.5   Min.   : 399
##  1st Qu.:98033   1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1490
##  Median :98065   Median :47.57   Median :-122.2   Median :1840
##  Mean   :98078   Mean   :47.56   Mean   :-122.2   Mean   :1987
##  3rd Qu.:98118   3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2360
##  Max.   :98199   Max.   :47.78   Max.   :-121.3   Max.   :6210
##    sqft_lot15
##  Min.   :   651
##  1st Qu.:  5100
##  Median :  7620
##  Mean   : 12768
##  3rd Qu.: 10083
##  Max.   :871200
```

```
plot(Housing_data$price)
```



Outlier removal

we see from previous plot, that most houses lies below 400,000 price range. Any prices above 400,000 can be considered an outlier. We remove it and see how that influences the data. The removal of outliers significantly reduced the range of house prices, which likely made the model more representative of the majority of homes. However, a warning indicates that some groups with fewer than two data points were dropped, which might affect interpretability.

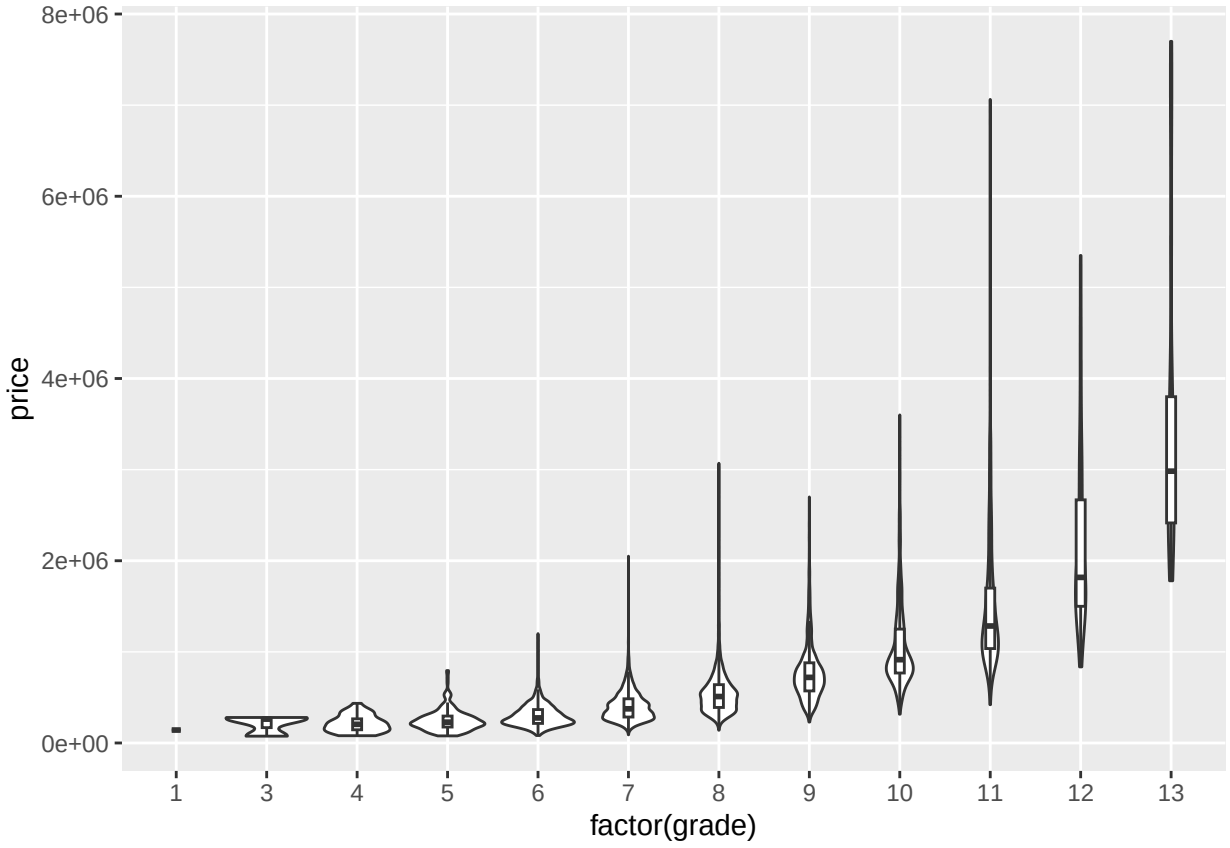
In the regression summary, Bedrooms have a negative coefficient (-51947.07), meaning that, after controlling for other variables, more bedrooms are associated with lower house prices. This might be due to smaller homes in premium areas having higher prices or other unobserved factors.

Bathrooms have a positive coefficient (9884.13), indicating that increasing the number of bathrooms is associated with higher house prices, though the effect is smaller than that of square footage.

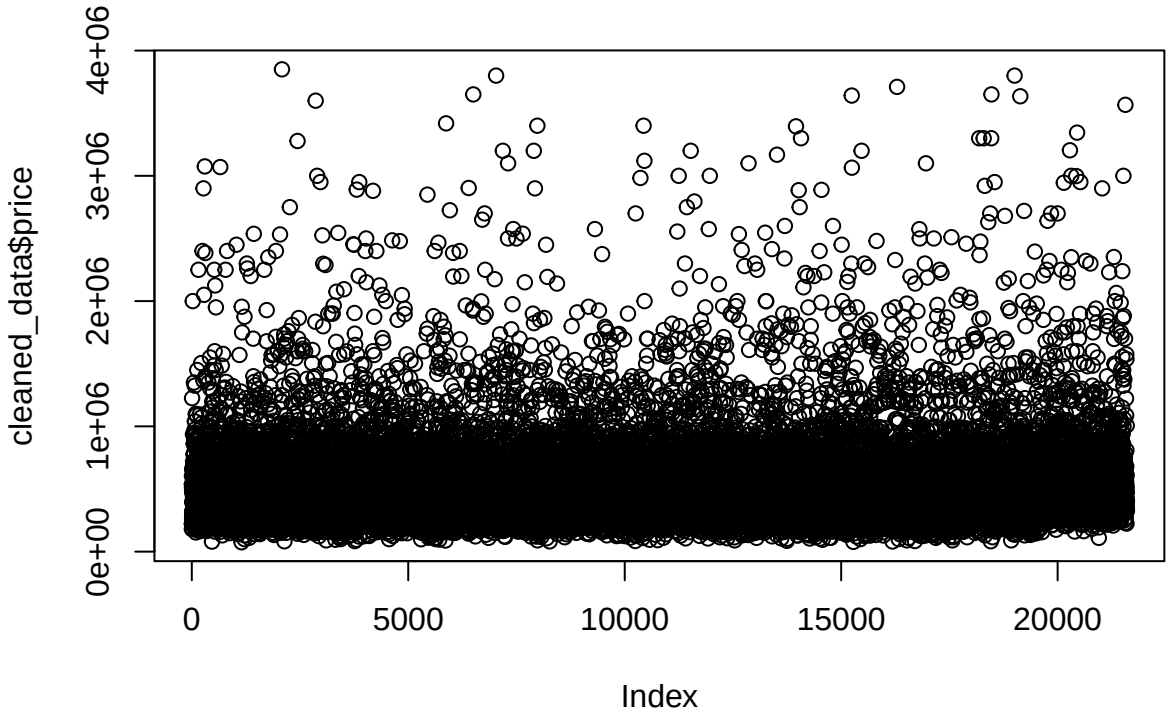
Sqft_living has the largest positive impact (291.27 per square foot), meaning that larger living spaces significantly increase house prices.

```
ggplot(Housing_data, aes(x = factor(grade), y = price)) + geom_violin() + geom_boxplot(width=0.1, outli
```

```
## Warning: Groups with fewer than two datapoints have been dropped.
## i Set 'drop = FALSE' to consider such groups for position adjustment purposes.
```




```
cleaned_data = Housing_data[Housing_data$price < 4000000, ]
plot(cleaned_data$price)
```

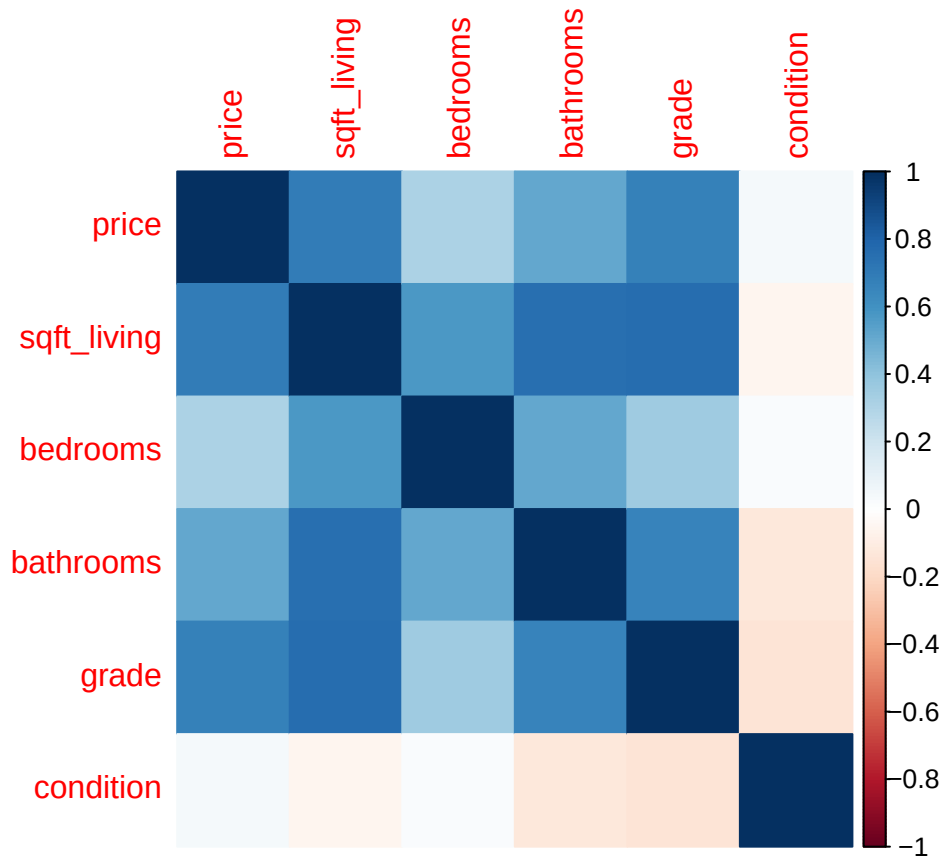



```
summary(lm(price ~ bedrooms + bathrooms + sqft_living, data = cleaned_data))
```

```
##
## Call:
## lm(formula = price ~ bedrooms + bathrooms + sqft_living, data = cleaned_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1466040  -142989   -24665    99192  2323706
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  86780.400   6638.444  13.072  < 2e-16 ***
## bedrooms    -51947.065   2243.930 -23.150  < 2e-16 ***
## bathrooms     9884.129   3370.631   2.932  0.00337 **
## sqft_living    291.272      2.992  97.338  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 247300 on 21597 degrees of freedom
## Multiple R-squared:  0.4946, Adjusted R-squared:  0.4946
## F-statistic:  7046 on 3 and 21597 DF,  p-value: < 2.2e-16
```

Correlation plots

```
corr_model = cor(cleaned_data %>% select(price, sqft_living, bedrooms, bathrooms, grade, condition), us
corrplot::corrplot(corr_model, method = "color")
```



Price per square foot -> further analysis

```
cleaned_data = cleaned_data %>% mutate(price_per_sqft = price / sqft_living)
log_data = cleaned_data %>% mutate(log_price = log(price))
```

Multivariable regression

1. Variables In this regression, we are predicting house price using five explanatory variables:

sqft_living: Total square footage of the living space. Larger homes generally have higher prices.

bedrooms: Number of bedrooms in the house.

bathrooms: Number of bathrooms in the house.

grade: A categorical variable that rates the quality of construction and design, with higher values indicating better quality.

condition: A categorical variable representing the condition of the house (e.g., poor to excellent).

2. Understanding Regression and Model Fit

Regression seeks a best-fit line that predicts the dependent variable (house price) using independent variables (sqft_living, bedrooms, etc.). The difference between the predicted and actual values is called the residual.

To evaluate model accuracy, we calculate:

Sum of Squared Residuals (SSR): Measures the total squared error in predictions.

Mean Squared Error (MSE): Average squared residuals, useful for comparing models.

R^2(coefficient of determination): Explains the proportion of variance in house prices that our model captures. The closer to 1, the better.

p-value: Indicates the statistical significance of each variable.

3. Interpretation of Regression Output

Key Findings: The multiple R-squared (0.5559) suggests that the model explains about 55.59% of the variance in house prices. The F-statistic (5407, $p < 2.2e-16$) confirms that at least one variable significantly contributes to predicting price.

Coefficients:

sqft_living (Estimate = 193.2, $p < 2e-16$): Each additional square foot adds $193.20 to the house price. This is the strongest predictor in the model.

bedrooms (Estimate = -35,510, $p < 2e-16$): Unexpectedly negative, meaning more bedrooms decrease price when controlling for other factors. This might be due to multicollinearity with sqft_living or differences in home design.

bathrooms (Estimate = -18,300, p = 2.3e-08): Also negative, which is unusual. It could be due to correlation with other features.

grade (Estimate = 110,500, $p < 2e-16$): Higher grade significantly increases price. A one-unit increase in grade raises the price by $110,500.

condition (Estimate = 65,030, $p < 2e-16$): A better condition increases price, but it has less impact than grade.

4. Plots

First we have the Residual vs Fitted Plot, where the clustering of plots forms a cone shape that lies within 0 - 1500000 on x and -1e+06 - 2e+06 on y, x = fitted values lm(price ~ sqft_living + bedrooms + bathrooms + grade + conditions) and y = residuals. This suggests that house price variance increases for expensive houses, meaning our model might be missing some key predictors for high-value homes.

For our Q-Q Plot, Most points follow the line, indicating that residuals roughly follow a normal distribution. However, the trailing off at high quantiles (above 2) suggests that the model struggles with extreme price values (potential outliers).

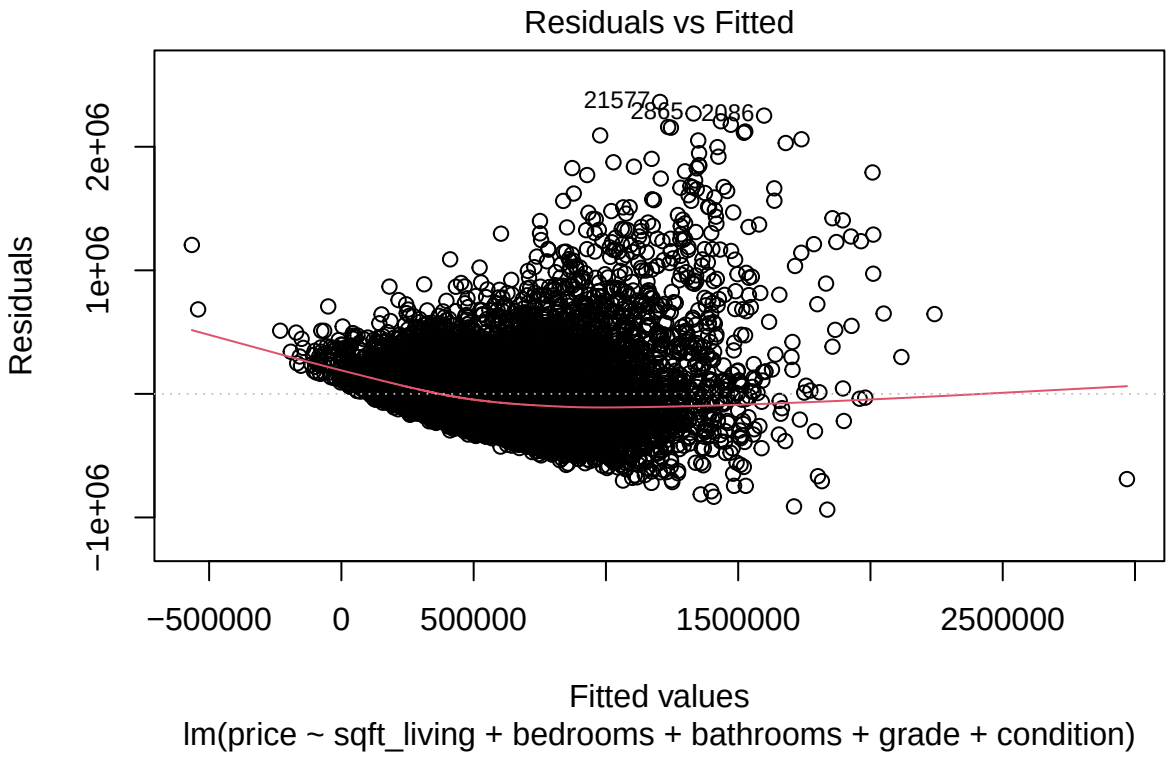
We then have a scale location plot that looks like a bit hair ball between x = 0 - 1500000 and y = 0 - 3.0 where y = sqrt(standard residual). Residuals are randomly scattered, meaning the model does not show severe patterns of heteroscedasticity. However, the slight fan shape at higher fitted values suggests increasing variance.

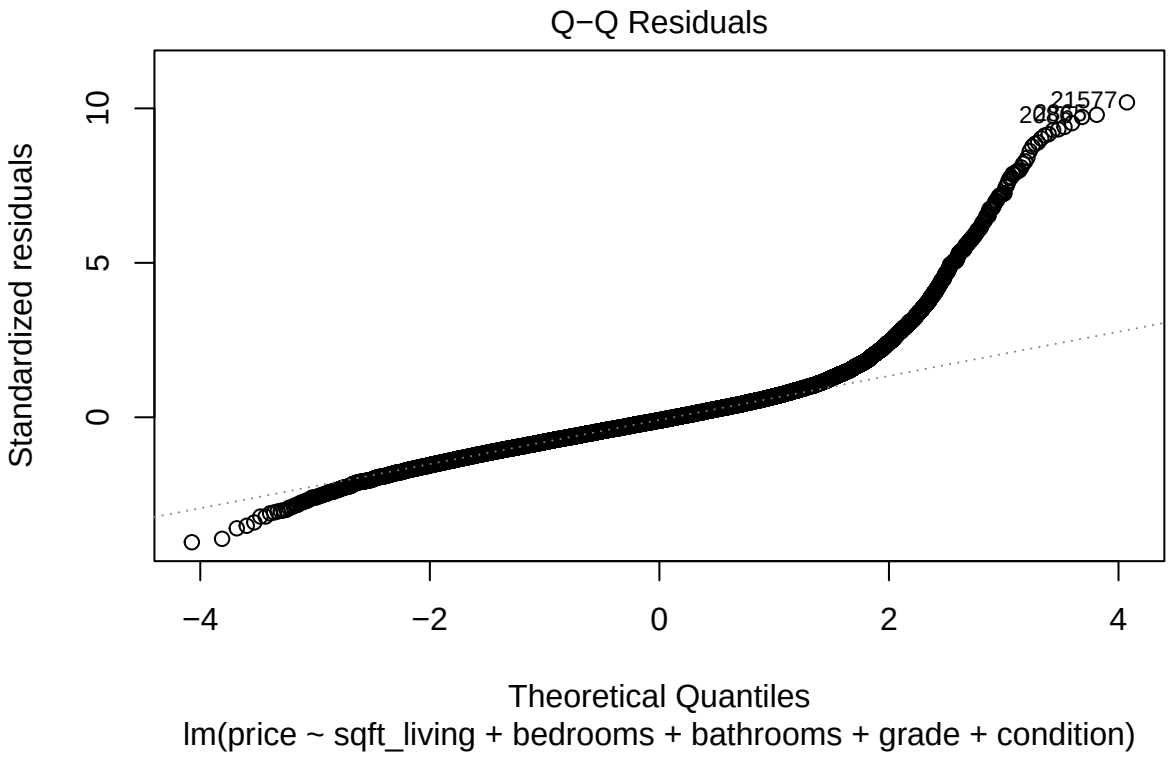
Last, we have our Residuals vs Leverage Plot. Most points cluster around low leverage, indicating no extreme influential points.One outlier at leverage = 15,870 suggests a highly influential observation—possibly a luxury home that doesn't fit the pattern.

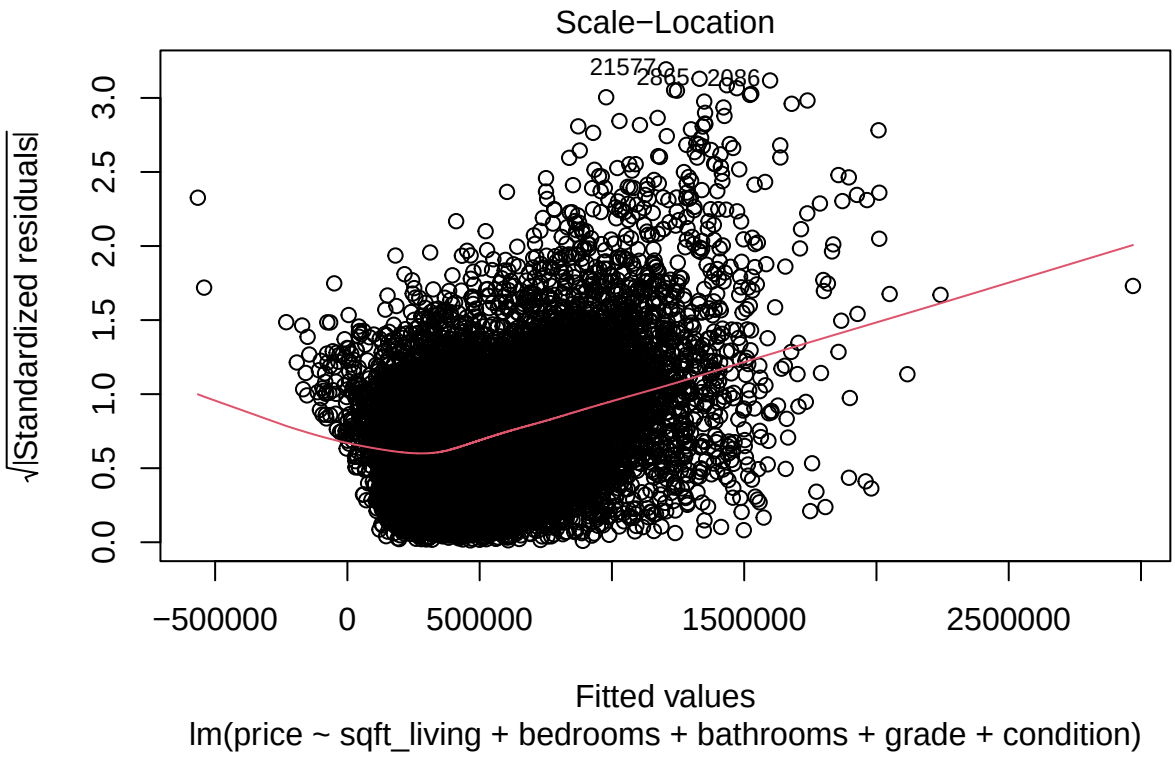
```r
reg_model = lm(price ~ sqft_living + bedrooms + bathrooms + grade + condition, data = cleaned_data)
summary(reg_model)
```

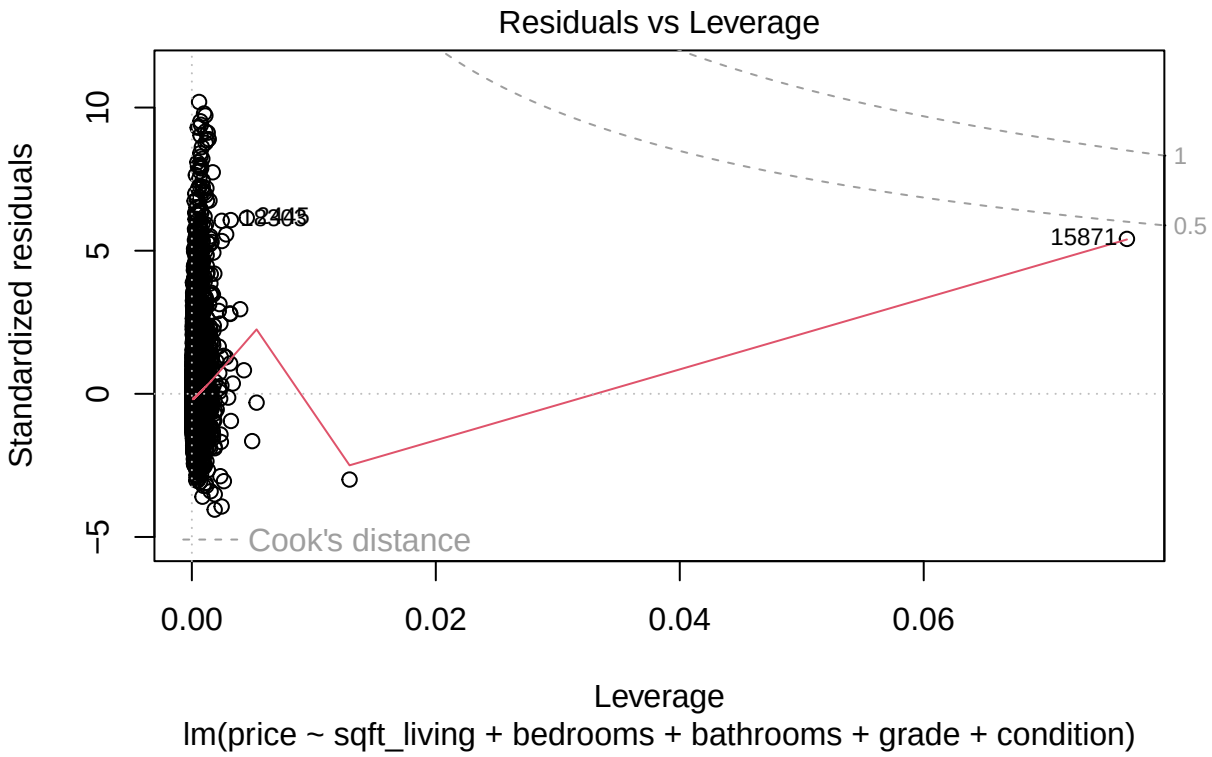
```
##
## Call:
## lm(formula = price ~ sqft_living + bedrooms + bathrooms + grade +
##     condition, data = cleaned_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -936710 -131791  -22890   91400 2362567
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.734e+05  1.717e+04  -45.05  < 2e-16 ***
## sqft_living  1.932e+02  3.398e+00   56.87  < 2e-16 ***
## bedrooms    -3.551e+04  2.153e+03  -16.50  < 2e-16 ***
## bathrooms   -1.830e+04  3.273e+03   -5.59  2.3e-08 ***
## grade        1.105e+05  2.185e+03   50.59  < 2e-16 ***
## condition    6.503e+04  2.475e+03   26.28  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 231800 on 21595 degrees of freedom
## Multiple R-squared:  0.5559, Adjusted R-squared:  0.5558
## F-statistic:  5407 on 5 and 21595 DF,  p-value: < 2.2e-16
```

```r
plot(reg_model)
```


Residuals vs Fitted
21577
2865
2086
Residuals
2e+06
1e+06
−1e+06
−500000
0
500000
1500000
2500000
Fitted values
lm(price ~ sqft_living + bedrooms + bathrooms + grade + condition)


Q−Q Residuals
Standardized residuals
10
5
0
21577
−4
−2
0
2
4
Theoretical Quantiles
lm(price ~ sqft_living + bedrooms + bathrooms + grade + condition)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
21577
2885
2086
−500000
0
500000
1500000
2500000
Fitted values
lm(price ~ sqft_living + bedrooms + bathrooms + grade + condition)

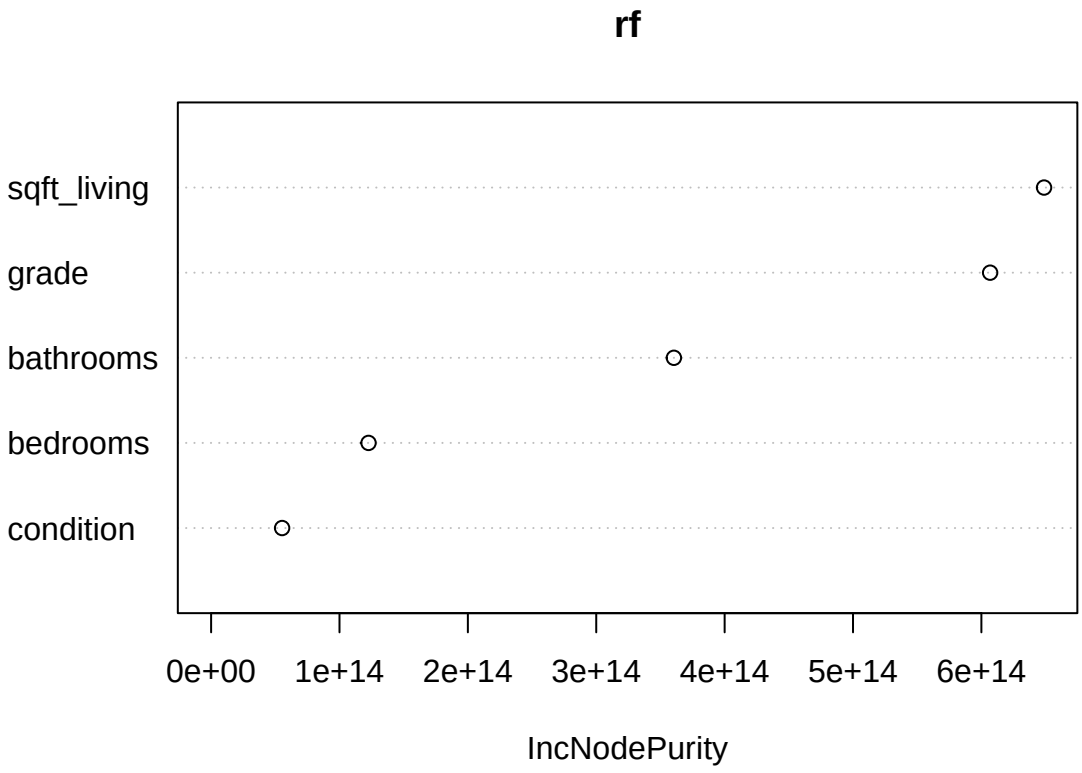


Compare linear regression vs lasso vs ridge

```
x = model.matrix(price ~ sqft_living + bedrooms + bathrooms + grade + condition, data = Housing_data)[,
y = Housing_data$price
ridge = cv.glmnet(x, y, alph = 0)
lasso = cv.glmnet(x, y, alpha = 1)
```

Predictive modeling

```
rf = randomForest(price ~ sqft_living + bedrooms + bathrooms + grade + condition, data=Housing_data, nt
varImpPlot(rf)
```



Model evaluation

```
predicted = predict(rf, Housing_data)
RMSE = sqrt(mean((Housing_data$price - predicted)^2))
print(RMSE)
```

```
## [1] 209338.9
```

This project analyzes the key factors influencing house prices using multiple statistical and machine learning techniques in R. Through linear regression, we identified that square footage, grade, and condition are the most significant predictors, while additional bedrooms have a counterintuitive negative effect, likely due to space trade-offs. Diagnostic plots revealed heteroscedasticity and potential outliers, suggesting that a simple linear model may not fully capture the complexity of housing prices. To improve model performance, we applied Ridge and Lasso regression, which helped reduce multicollinearity and refine feature selection. Finally, Random Forest provided a more flexible, non-linear approach, highlighting square footage and grade as the most important factors while achieving a reasonable RMSE of 208,667, though some error remained.