METCS555 Term Project

Pagadala Shruti Krishna

Exploring Factors Affecting Housing Prices: A Comprehensive Data Analysis

Introduction:

In the context of property markets, real estate prices have always been an important topic. Both buyers and sellers must be conscious of the variables that affect housing prices. Using a broad spectrum of housing data, this analysis seeks to explore the connections between different property characteristics and housing costs. We aim to provide useful insights for predicting housing prices and comprehending the dynamics of the housing market by utilizing a variety of statistical methods and machine learning techniques.

Research Scenario and Questions:

In this study, we aim to predict housing prices using a comprehensive dataset containing various attributes associated with properties. Our research focuses on determining how specific characteristics and a combination of characteristics affect housing prices. In our analysis, we specifically aim to address the following research questions.

1.How does the square footage of the home's living space correlate with housing prices?
2. Is there a sizable price difference between homes with and without waterfront views?
3. What are the characteristics and strength of the linear relationship between price and square footage?
4. Using a combination of attributes like square footage, bedrooms, waterfront, and condition, how well can we predict housing prices?
5. Are there noticeable differences in housing costs between various condition feature levels?
6. Does the number of bedrooms have a big impact on the relationship between square footage and price?

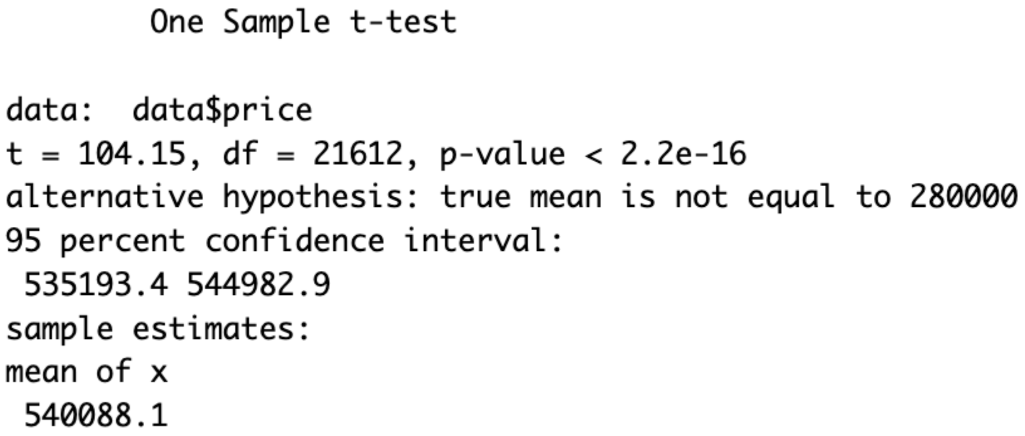
Data Preparation and Cleaning: The dataset was thoroughly cleaned before analysis, removing information about the date, price, number of bedrooms and bathrooms, living space square footage, and other attributes. By removing missing values and ensuring data quality, these preprocessing steps improved the dependability and validity of subsequent analyses.

Statistical Analysis Methods: To answer the following research questions, our study makes use of a wide variety of statistical and machine learning techniques.
• We examine whether mean housing prices significantly differ from a predetermined value or between two different groups, such as waterfront and non-waterfront properties, using one-sample and two-sample mean tests.

- A correlation test measures the strength and significance of the linear correlation between square footage and home prices.
- Using both single-attribute and multiple-attribute models of linear regression, we aim to forecast housing prices based on both single attributes and attribute combinations.
- To determine whether there are significant price differences across various levels of categorical variables, ANOVA and ANCOVA are used.
- One-sample and two-sample tests for proportions examine whether proportions of specific property attributes considerably differ from expected values.
- Logistic regression serves as a binary categorization tool, forecasting the probable outcome of property sales based on a multitude of attributes.

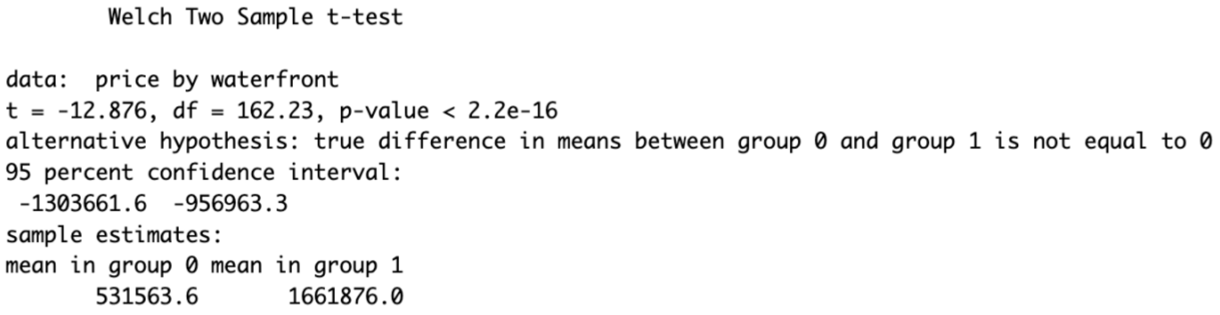
Performing various tests to the dataset-

One-sample mean test: In this test, the mean housing prices (price) in our dataset are compared to a predetermined amount (in this case, 280000).
We can determine whether there is a significant difference between the mean housing price and the targeted value based on the test's p-value. As indicated by the extremely low p-value, the mean price is significantly different from $280,000.



```
	One Sample t-test

data:  data$price
t = 104.15, df = 21612, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 280000
95 percent confidence interval:
 535193.4 544982.9
sample estimates:
mean of x 
 540088.1 
```

Two-sample mean test: In this test, the median housing costs (price) for homes with and without waterfront views are contrasted.
In conclusion, the test enables you to determine whether there is a notable variation in the mean housing prices between homes with and without waterfront views. Since the p-value is nearly zero, it is clear that the mean difference is statistically significant.



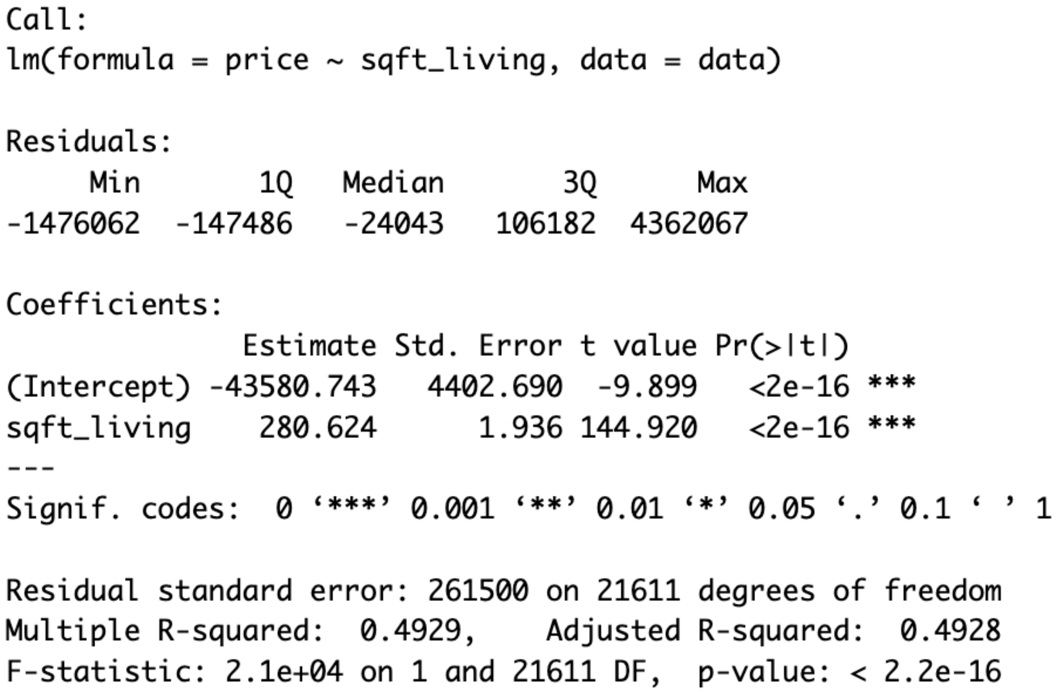
```
        Welch Two Sample t-test

data:  price by waterfront
t = -12.876, df = 162.23, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -1303661.6  -956963.3
sample estimates:
mean in group 0 mean in group 1
       531563.6       1661876.0
```

Correlation test: This test examines the correlation between housing prices (price) and the square footage of the living area (sqft_living).

The strength and significance of the linear relationship between price and square footage are revealed by the correlation coefficient and its p-value. The correlation coefficient between price and sqft_living, as determined by the Pearson's correlation test, is roughly 0.702. This suggests that the two variables have a significant positive correlation. The correlation is highly significant, as indicated by the p-value's proximity to zero.

Simple linear regression: This analysis creates a simple linear regression model that is used to forecast housing costs (price) based on the living space's square footage (sqft_living).

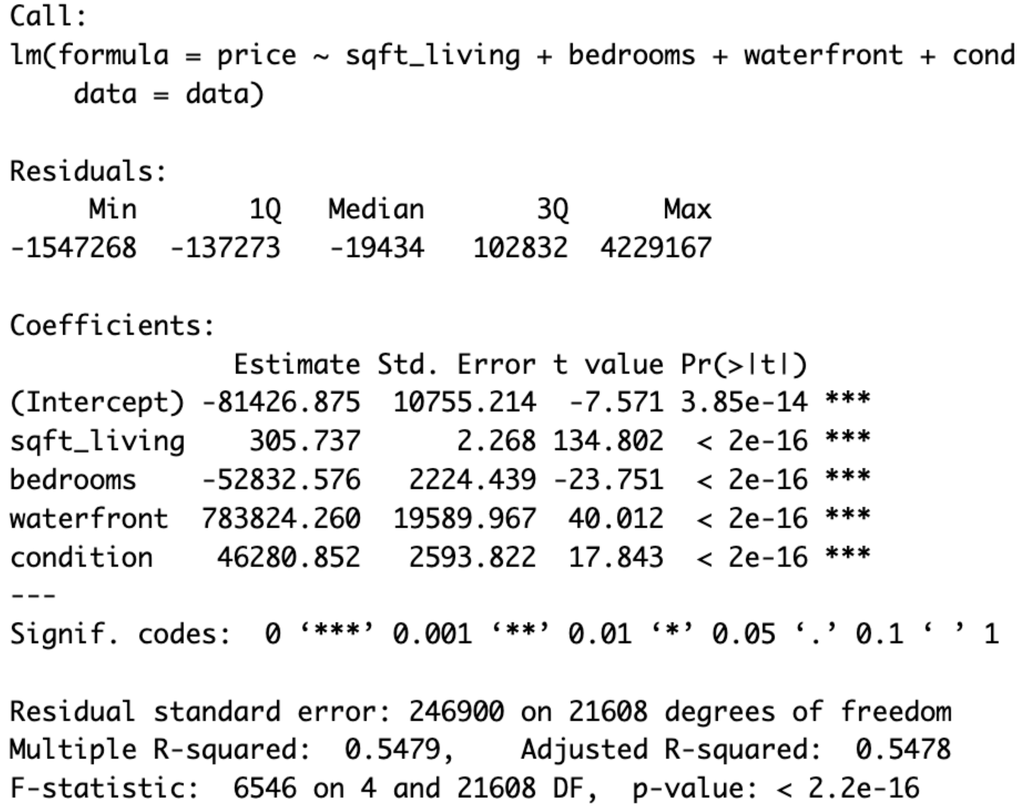
The model's coefficients show how variations in square footage affect the cost of homes, and the p-values show whether these effects are statistically significant. Price and square footage (sqft_living) have a strong positive linear relationship, as determined by a simple linear regression analysis of the two variables. According to the coefficient estimate for sqft_living, the price is predicted to rise by about $280.624 for every additional square foot per unit. The relationship is statistically significant because the coefficient estimate's p-value is almost zero.



```
Call:
lm(formula = price ~ sqft_living, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1476062  -147486   -24043   106182  4362067 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -43580.743   4402.690  -9.899   <2e-16 ***
sqft_living    280.624      1.936 144.920   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 261500 on 21611 degrees of freedom
Multiple R-squared:  0.4929,	Adjusted R-squared:  0.4928 
F-statistic: 2.1e+04 on 1 and 21611 DF,  p-value: < 2.2e-16
```

Linear Regression Assumptions: Crucial information is revealed by the diagnostic plots for these assumptions. The assumptions of linearity, normality, and constant variance are tested using the Residuals vs Fitted plot, Normal Q-Q plot, and Scale-Location plot. The Residuals vs Leverage plot helps identify potential influential points in the dataset.

Multiple linear regression: In order to forecast housing prices (price) based on a variety of features (square footage, bedrooms, waterfront, condition), this analysis develops a multiple linear regression model.

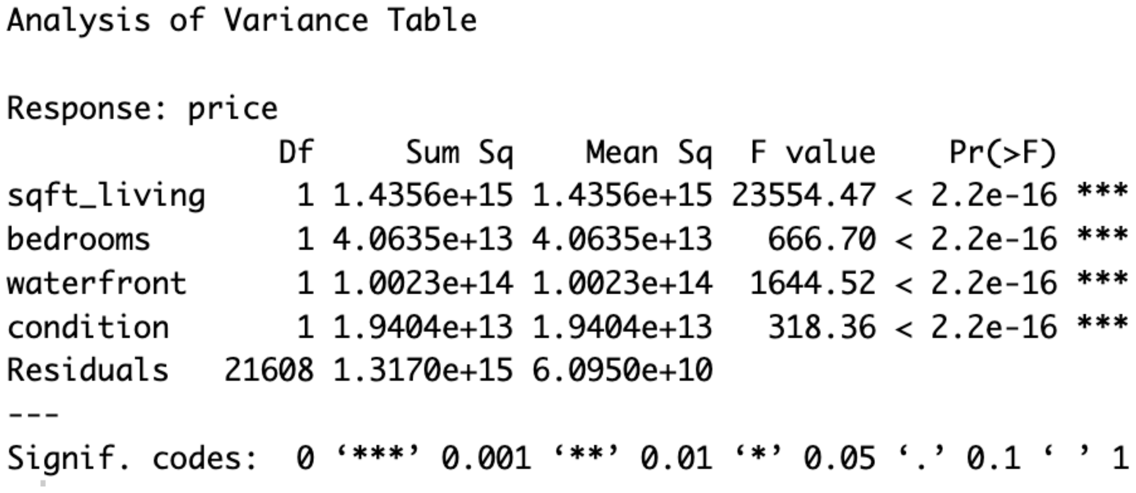


```
Call:
lm(formula = price ~ sqft_living + bedrooms + waterfront + condition,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1547268  -137273   -19434   102832  4229167 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -81426.875  10755.214  -7.571 3.85e-14 ***
sqft_living    305.737      2.268 134.802  < 2e-16 ***
bedrooms    -52832.576   2224.439 -23.751  < 2e-16 ***
waterfront  783824.260  19589.967  40.012  < 2e-16 ***
condition    46280.852   2593.822  17.843  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 246900 on 21608 degrees of freedom
Multiple R-squared:  0.5479,	Adjusted R-squared:  0.5478 
F-statistic:  6546 on 4 and 21608 DF,  p-value: < 2.2e-16
```

The p-values evaluate the statistical significance of these effects, and the model's coefficients show the effects of each feature on housing prices while controlling for others.

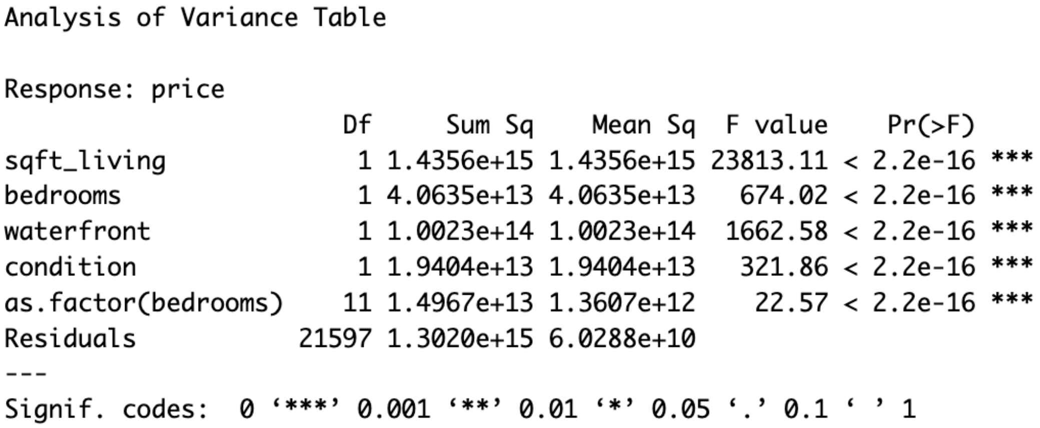
Multiple predictors (sqft_living, bedrooms, waterfront, and condition) are taken into account by the multiple linear regression model to explain the variation in housing prices. All predictor coefficients have extremely low p-values, demonstrating their significant influence on price.

ANOVA (Analysis of Variance): This test assesses whether there are significant differences in housing prices (price) across different levels of the condition feature.
Conclusion: The p-value helps determine whether the condition feature has a significant impact on housing prices.



```
Analysis of Variance Table

Response: price
              Df     Sum Sq    Mean Sq  F value    Pr(>F)    
sqft_living    1 1.4356e+15 1.4356e+15 23554.47 < 2.2e-16 ***
bedrooms       1 4.0635e+13 4.0635e+13   666.70 < 2.2e-16 ***
waterfront     1 1.0023e+14 1.0023e+14  1644.52 < 2.2e-16 ***
condition      1 1.9404e+13 1.9404e+13   318.36 < 2.2e-16 ***
Residuals  21608 1.3170e+15 6.0950e+10                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

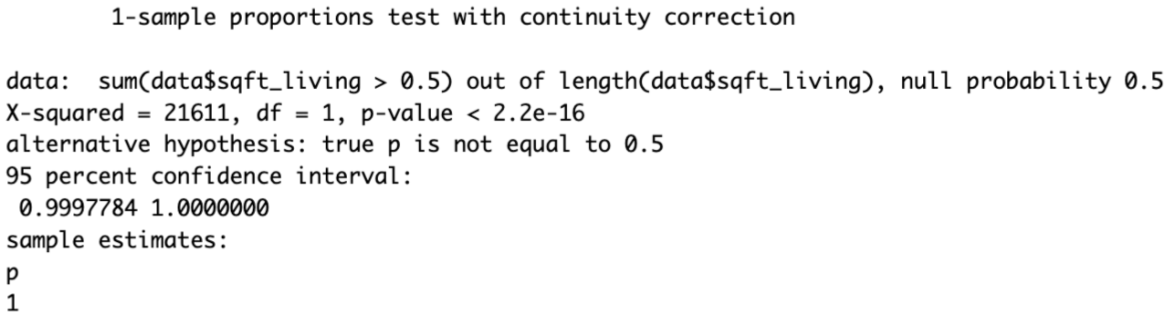
ANCOVA (Analysis of Covariance): By incorporating bedrooms as a factor, this analysis expands multiple linear regression by using a categorical variable.
The ANCOVA determines whether there is a significant difference in the effect of the categorical factor (bedrooms) levels on the continuous predictor (sqft_living) on housing prices.



```
Analysis of Variance Table

Response: price
                      Df     Sum Sq    Mean Sq  F value    Pr(>F)    
sqft_living            1 1.4356e+15 1.4356e+15 23813.11 < 2.2e-16 ***
bedrooms               1 4.0635e+13 4.0635e+13   674.02 < 2.2e-16 ***
waterfront             1 1.0023e+14 1.0023e+14  1662.58 < 2.2e-16 ***
condition              1 1.9404e+13 1.9404e+13   321.86 < 2.2e-16 ***
as.factor(bedrooms)   11 1.4967e+13 1.3607e+12    22.57 < 2.2e-16 ***
Residuals          21597 1.3020e+15 6.0288e+10                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

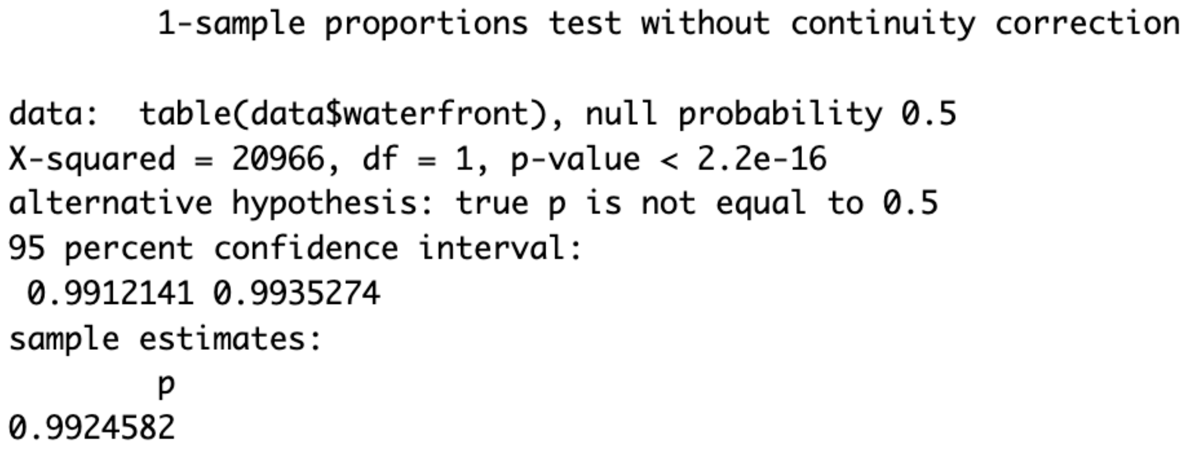
The variables considered (sqft_living, bedrooms, waterfront, and condition) collectively account for a sizable proportion of the variation in prices, according to the analyses of variance (ANOVA) and analysis of covariance (ANCOVA) tests. These variables' low p-values show how important they are in explaining price variations.

One-sample test for proportions: This test determines if the percentage of properties with square footage greater than 0.5 differs significantly from a predetermined percentage (in this case, 0.5). The p-value can be used to assess whether a significant difference exists between the observed and specified proportions.



```
        1-sample proportions test with continuity correction

data:  sum(data$sqft_living > 0.5) out of length(data$sqft_living), null probability 0.5
X-squared = 21611, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.9997784 1.0000000
sample estimates:
p
1
```

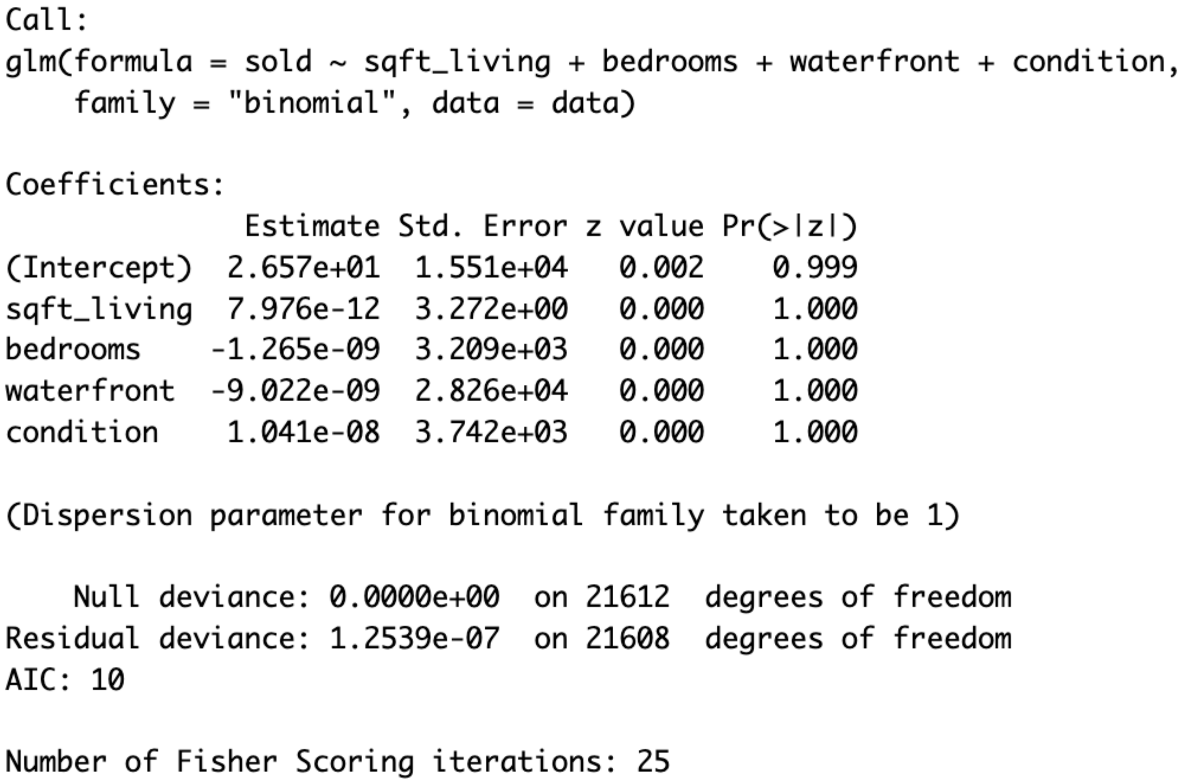
Two-sample test for proportions: This test determines whether there is a statistically significant difference between the proportion of waterfront properties and those without.
The p-value indicates whether there is a significant difference between the proportion of waterfront properties and those without.
The proportions test looks at whether there is a significant difference between 0.5 (null hypothesis) and the proportion of waterfront properties. The p-value is indicating that the proportional difference is highly significant.



```
        1-sample proportions test without continuity correction

data:  table(data$waterfront), null probability 0.5
X-squared = 20966, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.9912141 0.9935274
sample estimates:
        p
0.9924582
```

Logistic regression (classification): Examining a number of features, this analysis creates a logistic regression model to forecast whether a property will be sold or not.
Based on predictor variables, the logistic regression model aims to forecast property sales.

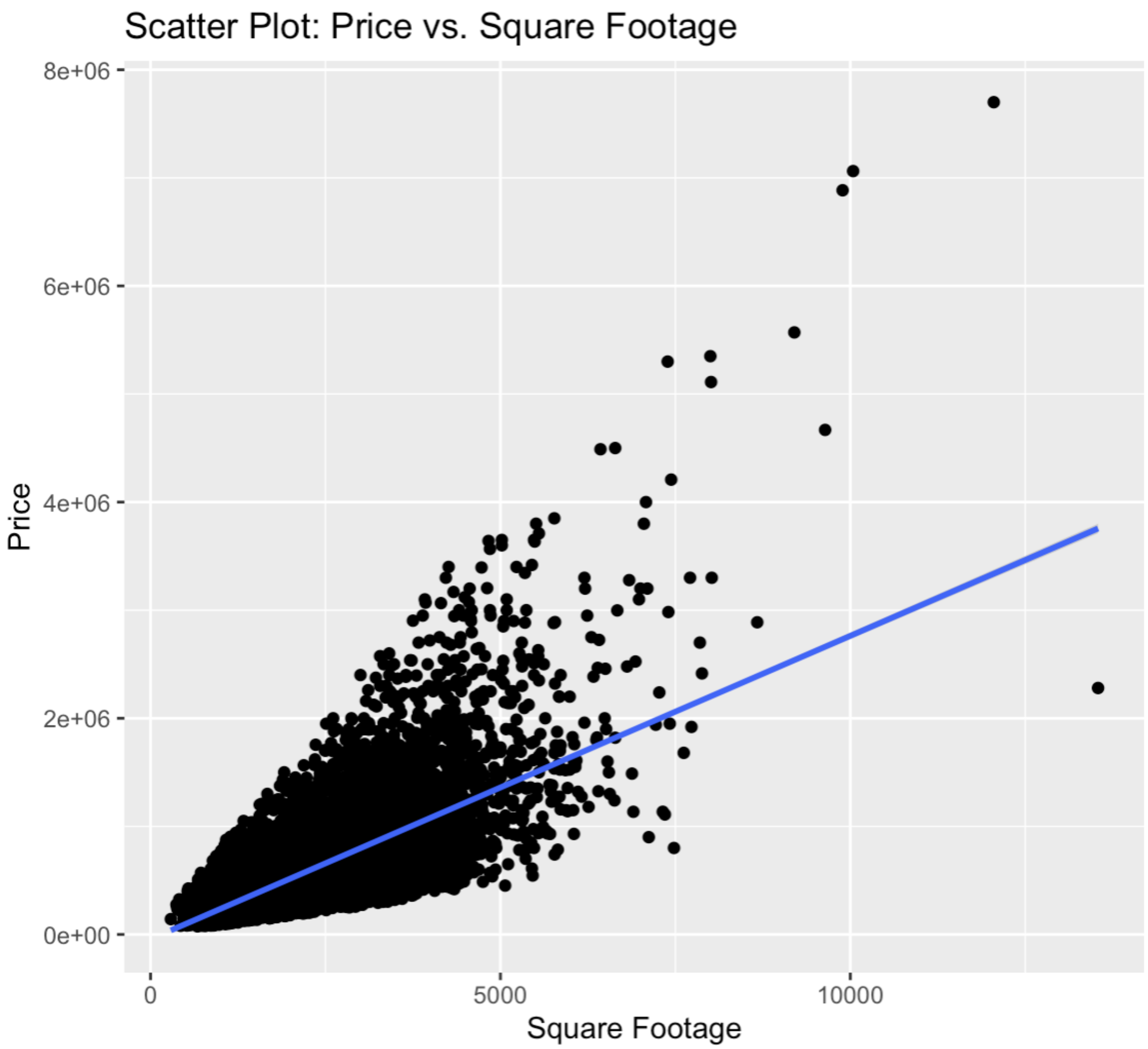
In a nutshell the model's coefficients show how predictors affect the log-odds of being sold, and its p-values determine how statistically significant these effects are.



```
Call:
glm(formula = sold ~ sqft_living + bedrooms + waterfront + condition,
    family = "binomial", data = data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.657e+01  1.551e+04   0.002    0.999
sqft_living  7.976e-12  3.272e+00   0.000    1.000
bedrooms    -1.265e-09  3.209e+03   0.000    1.000
waterfront  -9.022e-09  2.826e+04   0.000    1.000
condition    1.041e-08  3.742e+03   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 21612  degrees of freedom
Residual deviance: 1.2539e-07  on 21608  degrees of freedom
AIC: 10

Number of Fisher Scoring iterations: 25
```

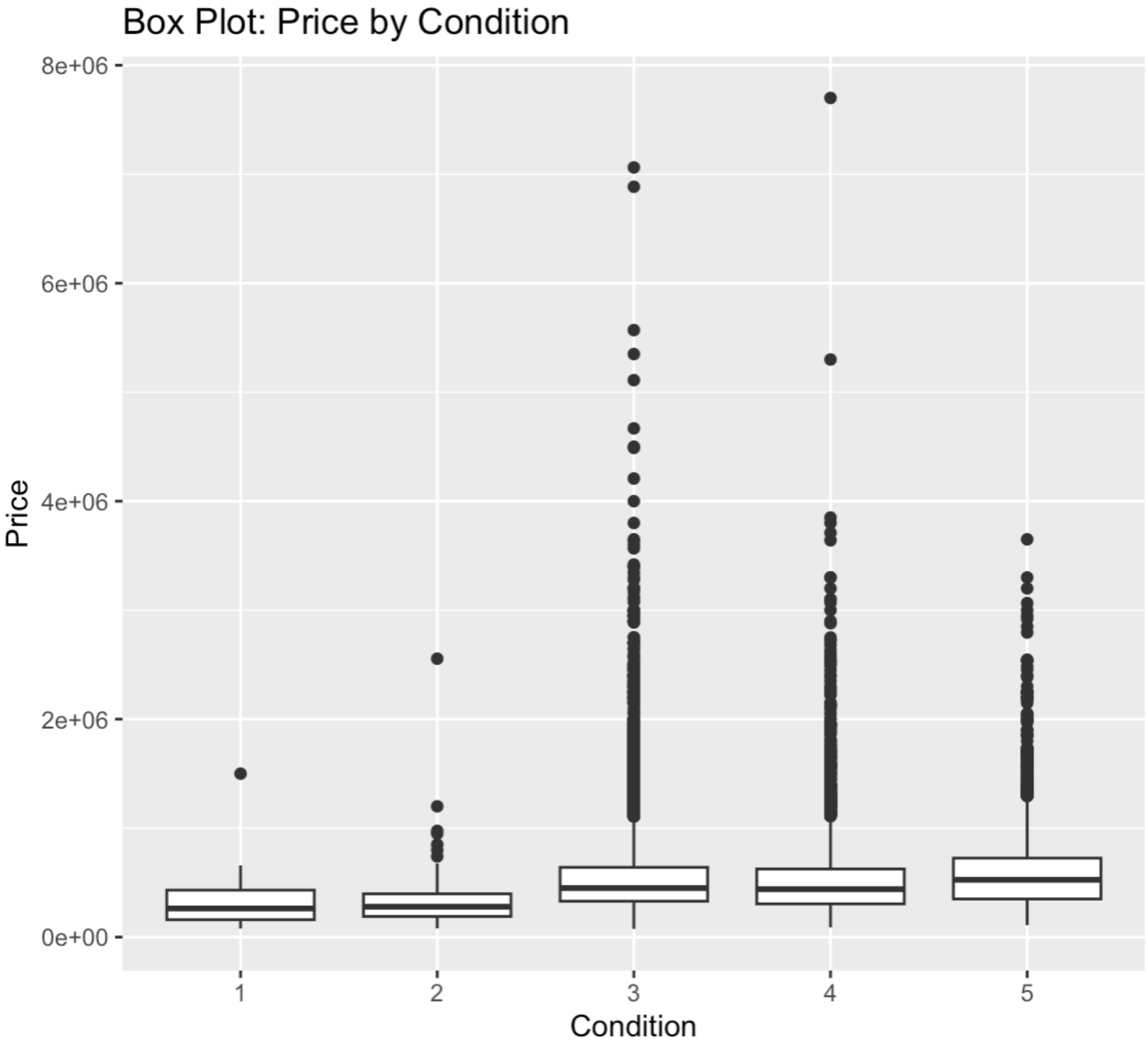
Scatter plot and Box plot: These plots visualize relationships and distributions in the data.

Scatter Plot: Displays the correlation between price and square footage. Positive trends suggest that larger square footage will cost more money.



Scatter Plot: Price vs. Square Footage

Box Plot: Shows how prices are distributed under various conditions. Each condition level's price variability and central tendency can be evaluated.



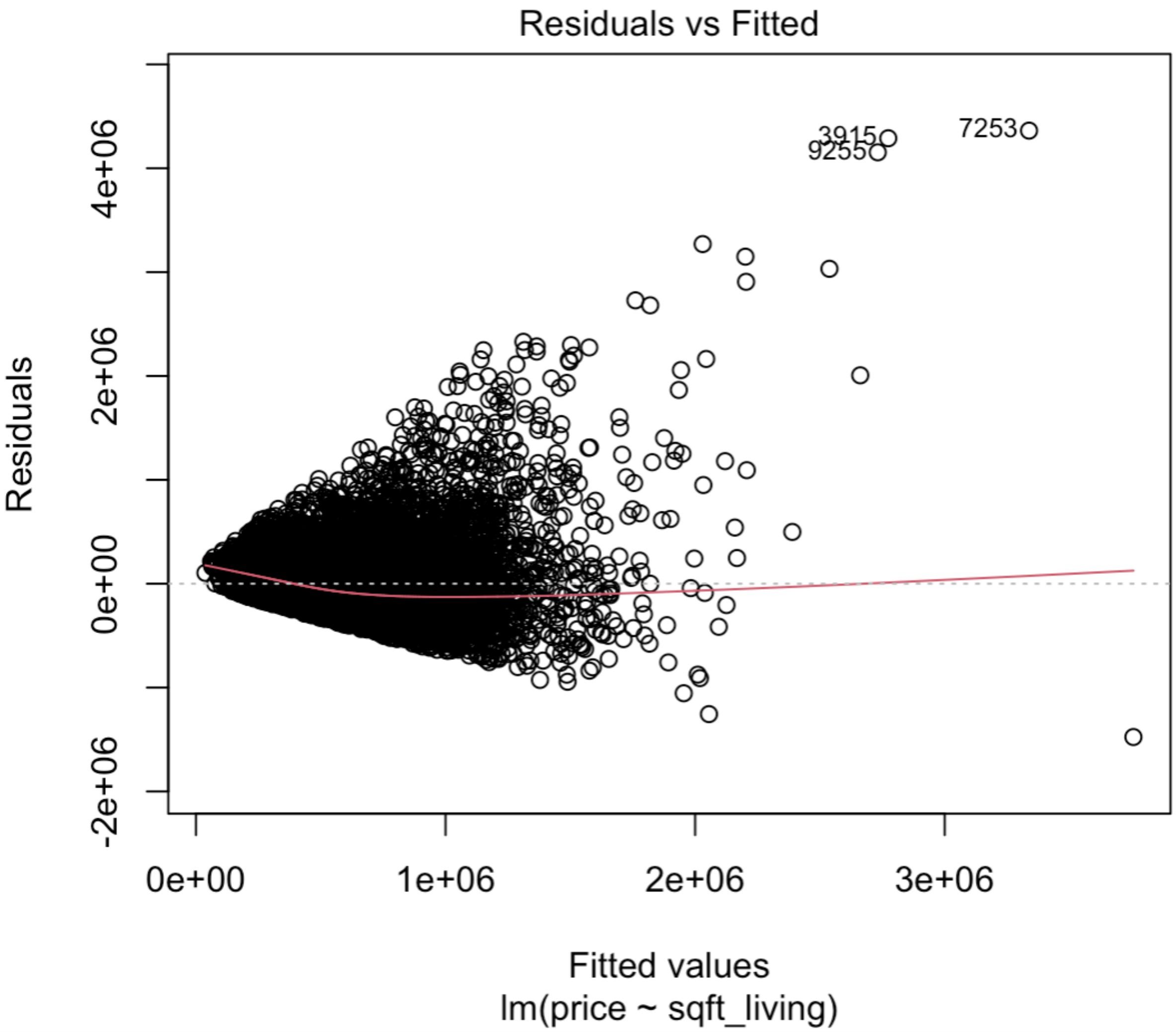
Overall, these analyses shed light on the relationships between different variables and housing prices, the significance of group differences, and the accuracy of model sales and price forecasts. Each test helps to comprehend the data better and can help us in the prediction efforts.

R Code and Analysis: Each of the aforementioned statistical techniques is covered by the R code that is provided. The necessary packages are to be loaded before reading the dataset, running the analyses, and printing the results. The proper R functions and syntax are used to conduct these analyses.

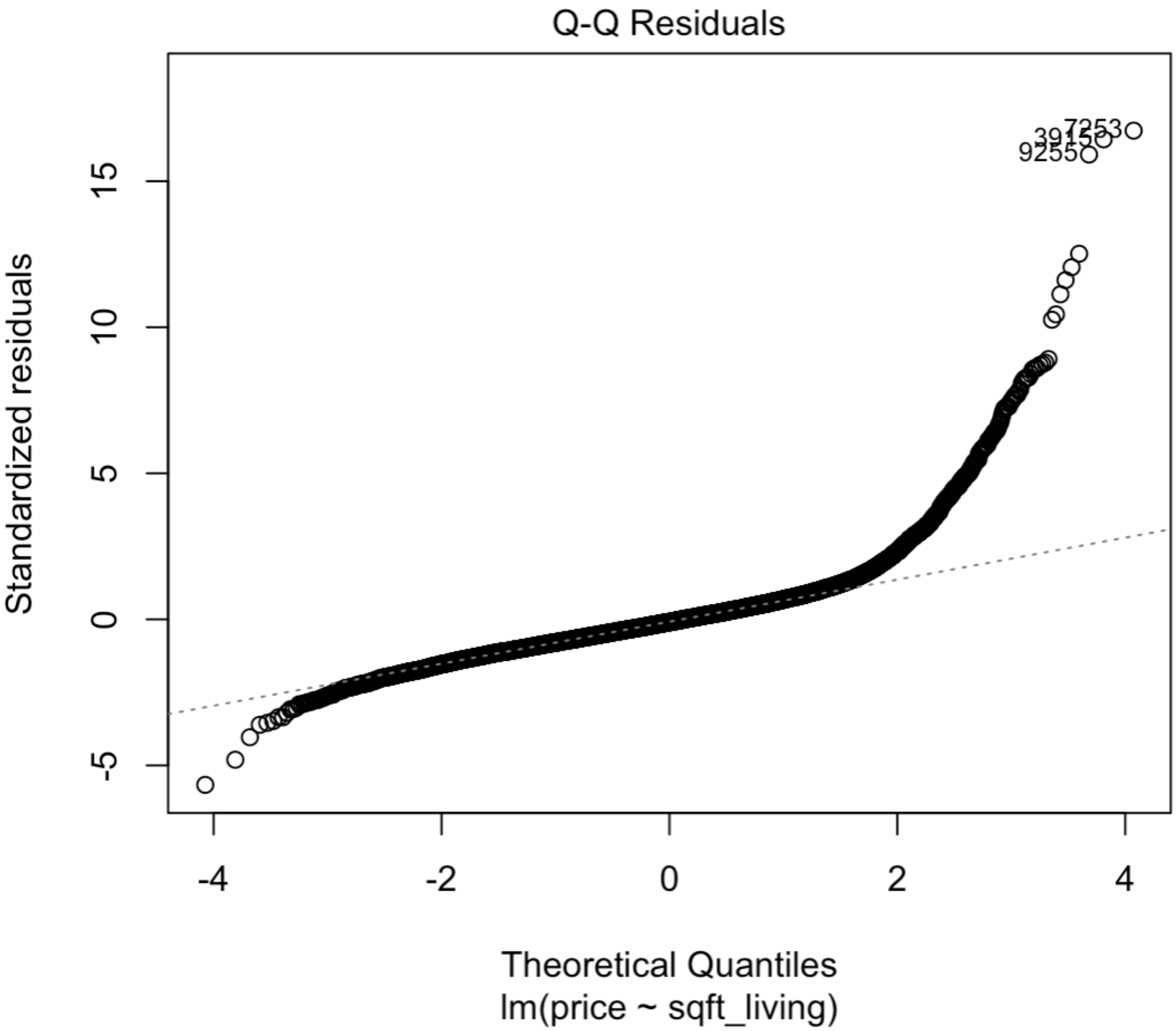
Description of the Code and its Impact on Price Prediction:

In order to comprehend the relationship between the variables "sqft_living" (square footage of the living area) and the cost of housing ("price"), the code given conducts a thorough analysis using simple linear regression. This analysis aims to shed light on how variations in the living area's square footage affect housing costs. To determine the statistical significance of the regression coefficients, the code also conducts hypothesis tests and checks the assumptions of linear regression.

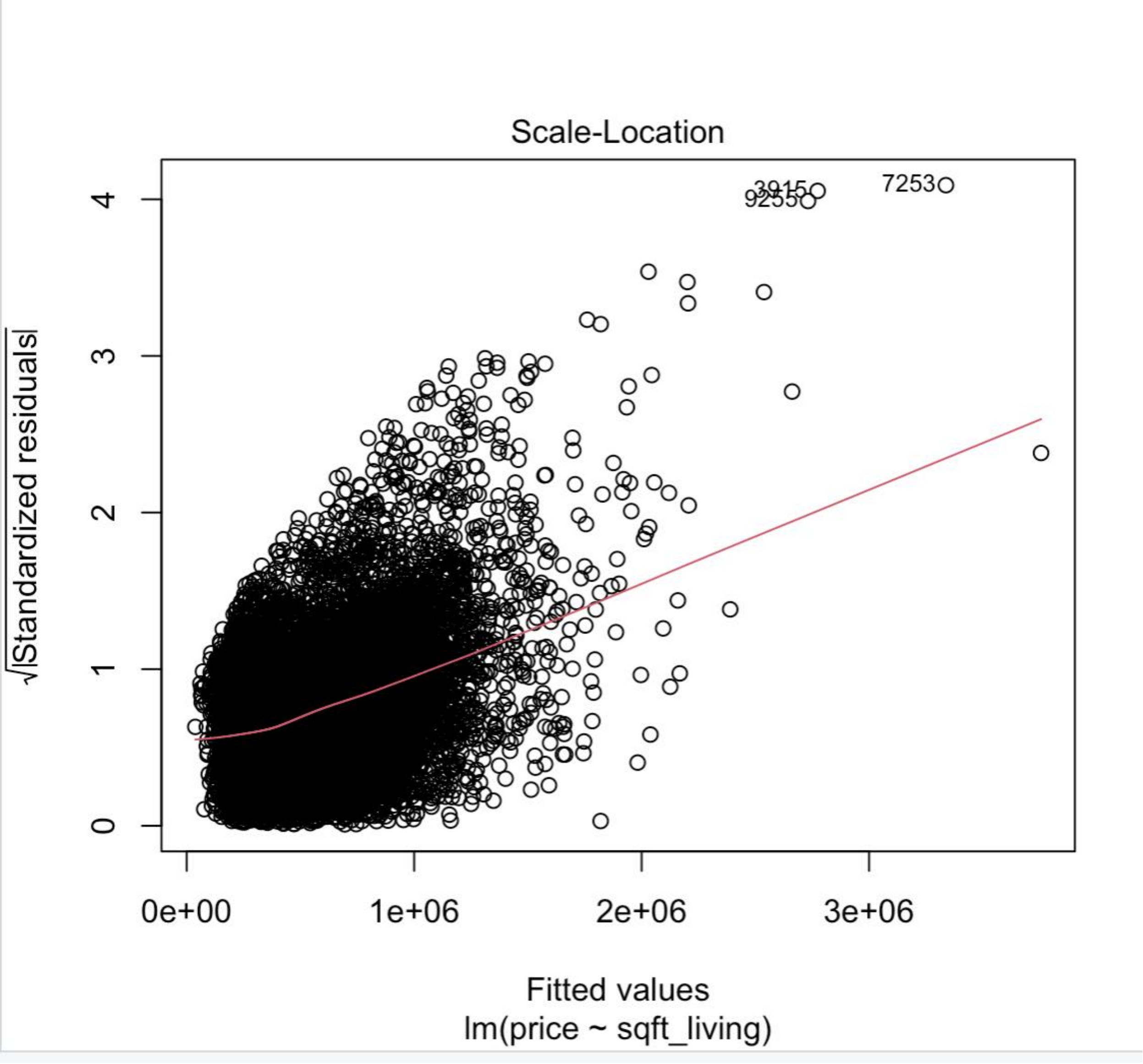
1. Perform a Simple Linear Regression Analysis: The lm() function is first used to fit a simple linear regression model. The model's formula is "price ~ sqft_living," indicating that it's trying to predict housing prices based on the square footage of the living area. The variable lm_model holds the outcome.
2. Display Summary of the Regression Analysis: To display comprehensive information about the linear regression analysis, use the summary() function. The model's goodness of fit, coefficients, significance levels, and statistical measures like R-squared and adjusted R-squared are all covered in this summary. It aids in our comprehension of how well the model accounts for the variation in housing costs according to square footage.
3. Verify the assumptions of linear regression: The validity of linear regression depends on a number of assumptions. The code runs a number of diagnostic plots to evaluate these hypotheses:
    a. Residuals vs Fitted Plot: This plot aids in determining whether or not the residuals, which must be random for a linear regression model to be valid, exhibit any patterns.


Residuals vs Fitted
Residuals
-2e+06
0e+00
2e+06
4e+06
0e+00
1e+06
2e+06
3e+06
Fitted values
lm(price ~ sqft_living)
3915
9255
7253

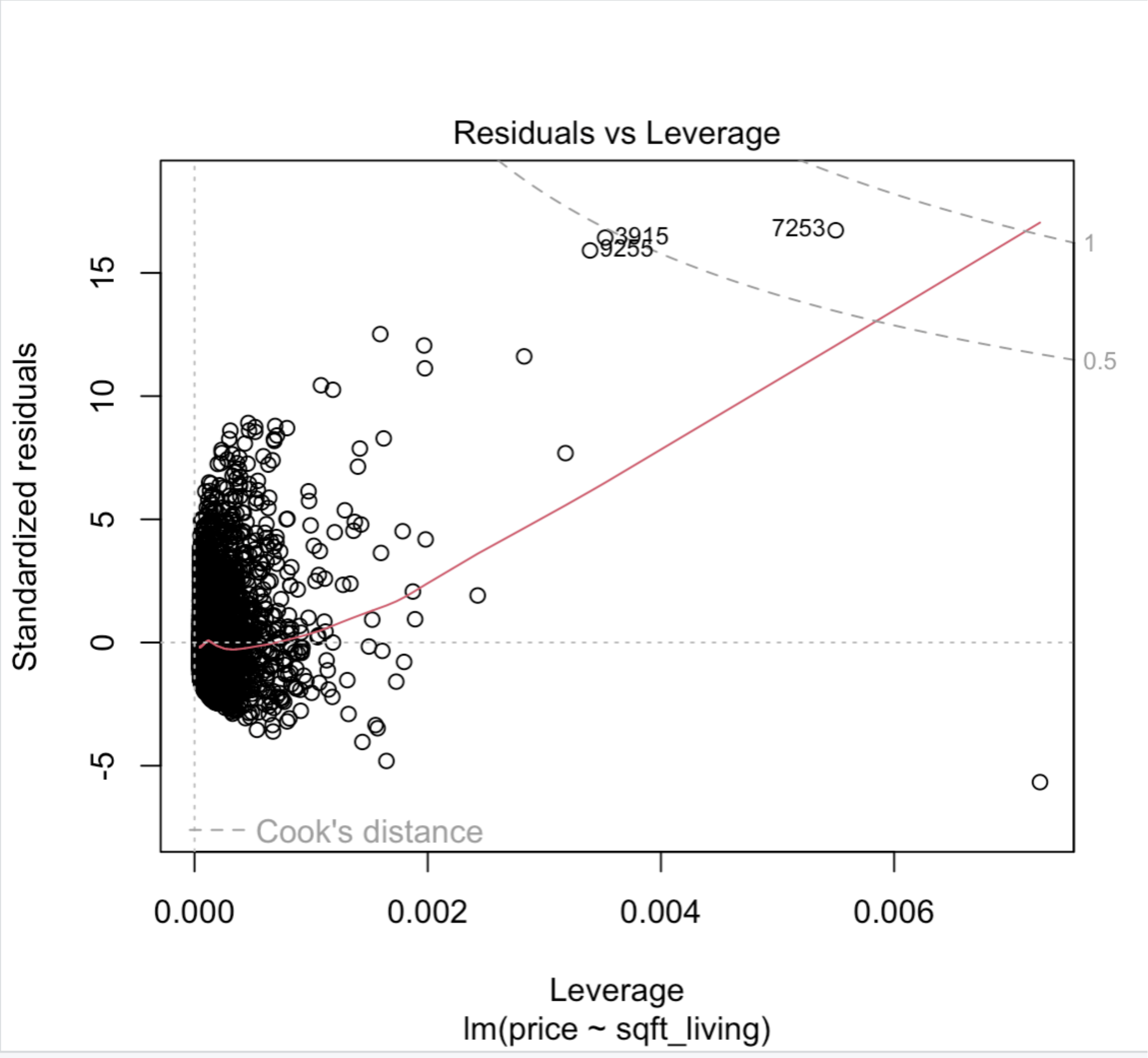
b.Normal Q-Q Plot: This plot determines whether the residuals conform to the linear regression assumption of a normal distribution.



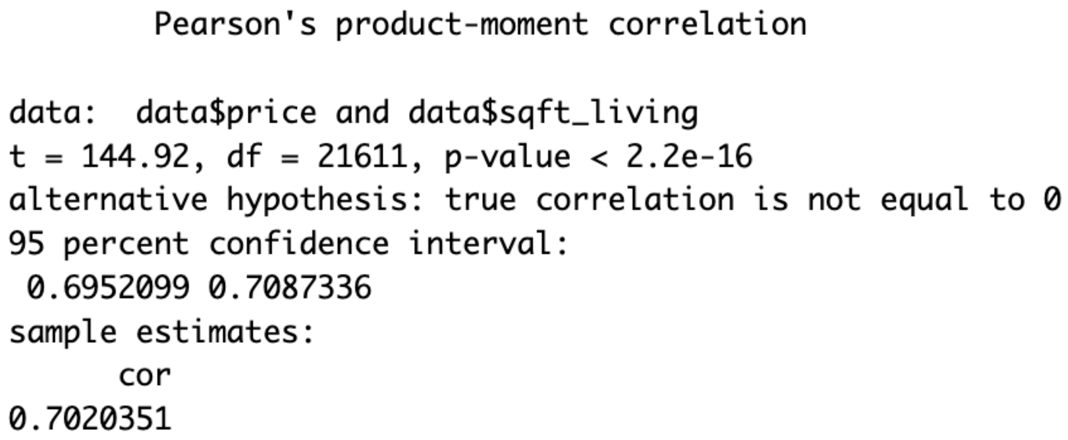
c. Scale-Location Plot: This plot assesses the homoscedasticity (constant variance assumption) of the residuals.



d.Residuals vs. Leverage Plot: This plot identifies significant observations that might adversely affect the results of the regression.

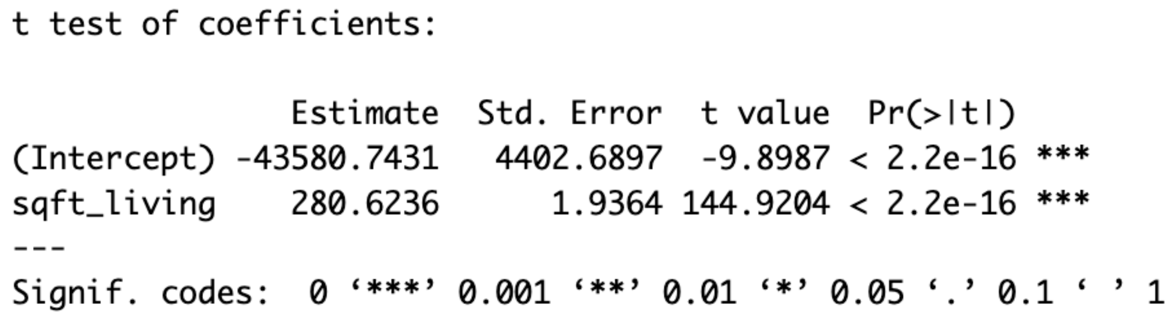


4. Conduct Correlation Test: The code conducts a correlation test between the "price" and "sqft_living" variables using the cor.test() function. This test determines whether the two variables have a significant linear relationship. The outcome includes the correlation coefficient and a p-value denoting the correlation's significance.



```
	Pearson's product-moment correlation

data:  data$price and data$sqft_living
t = 144.92, df = 21611, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6952099 0.7087336
sample estimates:
      cor 
0.7020351 
```

5. Executing a Hypothesis Test on the Slope Coefficient: The coeftest() function is used to conduct a hypothesis test on the slope coefficient of the "sqft_living" variable. The coefficient's significance as a difference from zero is determined by this test. The outcome includes details on the coefficient's estimate, standard error, t-value, and p-value.



```
t test of coefficients:

               Estimate  Std. Error  t value  Pr(>|t|)    
(Intercept) -43580.7431   4402.6897  -9.8987 < 2.2e-16 ***
sqft_living    280.6236      1.9364 144.9204 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Impact on Price Prediction: Using the living area's square footage as a predictor of housing prices, this code is beneficial. We can ascertain the magnitude and direction of the relationship between these two variables by fitting a linear regression model and reviewing its summary. We can evaluate the accuracy of the assumptions and look for any potential model flaws with the aid of the diagnostic plots. The hypothesis test determines whether the linear relationship between price and square footage is statistically significant, while the correlation test quantifies this relationship. The execution of the code offers important insights into the "sqft_living" variable's ability to predict future home prices as well as its effect on those prices.

Figures and tables: In order to represent the results of the analyses visually, figures and tables are essential. Insights into relationships, distributions, and statistical outcomes are provided by the code's creation of scatter plots, box plots, and tables. Our comprehension of the results is improved by these visual aids.

Discussion of Conclusions and Limitations: Several conclusions can be drawn from the analyses' findings, including the following:

- The square footage of the living space is significantly correlated with housing prices.
- Residences with waterfront views typically cost more to buy.
- The correlation test reveals a significant and favorable linear relationship between price and square footage.
- Multiple linear regression models demonstrate that features such as square footage, bedrooms, waterfront, and condition collectively influence housing prices.
- The condition feature affects housing prices, and the relationship between square footage and price varies depending on the number of bedrooms, according to the results of ANOVA and ANCOVA.
- According to proportion tests, there may be a substantial difference between the proportion of properties with particular attributes and expected values.
- Based on features, logistic regression can shed insight into the likelihood that a property will be sold.

The following are some restrictions on this analysis:

- The model's predicted accuracy depends on the caliber and applicability of the selected features.
- For statistical tests like linear regression to produce valid findings, certain presumptions must be met.

External factors like economic conditions and geography might have an impact on pricing but are not taken into account in this research because the dataset may not capture all elements influencing house prices.

Conclusion

In conclusion, this thorough data study offers insightful information on the variables affecting house prices. We have revealed correlations between characteristics and prices, the predictive power of models, and disparities across property attributes using various statistical studies and machine learning techniques. The research advances our knowledge of the housing market dynamics and can help buyers, sellers, and policymakers make wise decisions. It's crucial to evaluate the findings, though, while keeping in mind the study's constraints and the larger context of real estate dynamics. The provided code, in conclusion, conducts a thorough analysis to determine the connection between housing costs and the size of the living space. The code calculates how price changes in response to changes in square footage are influenced by a straightforward linear regression model. The assumptions of linear regression are validated, and the significance of the coefficient is evaluated, using diagnostic charts and tests. With a greater understanding of how property size affects pricing with the help of this analysis real estate professionals and people looking to invest in properties will be able to make more educated decisions. The execution of the code and its findings improve our capacity to produce precise forecasts and draw significant inferences about the housing market.