Assignment 8

Surenther

2024-10-21

Load XLSX file

```
#Load XLSX
library(readxl)
mydata <- read_excel("week-6-housing.xlsx", 1)
```

Data Cleanup

```
library(dplyr,warn.conflicts = FALSE)
# Remove unwanted Columns
housing_data <- select(mydata,-c(sale_reason,sale_instrument,sale_warning,sitetype,
                                 ctyname,present_use))
```

Explanation
The columns such as Sale Reason, Sale Instrument, Sale Warning, and Present Use contain references without any accompanying explanation, making their inclusion unclear. Additionally, columns like Sitetype and City Name are duplicates of Postalcty and Prop Type, respectively. Therefore, these columns have been removed for clarity and to avoid redundancy.

```
# Rename Columns
housing_data <- rename(housing_data,c(sale_date='Sale Date',sale_price='Sale Price',
                      address=addr_full,zip=zip5,city=postalctyn,longtitude=lon,
                      lattitude=lat))
```

Explanation
Rename the columns by converting all characters to lowercase and assigning more meaningful names for clarity and consistency.

```
# Add Year Column
housing_data <- housing_data %>%
  mutate('sale_year' = format(housing_data$'sale_date','%Y'))
```

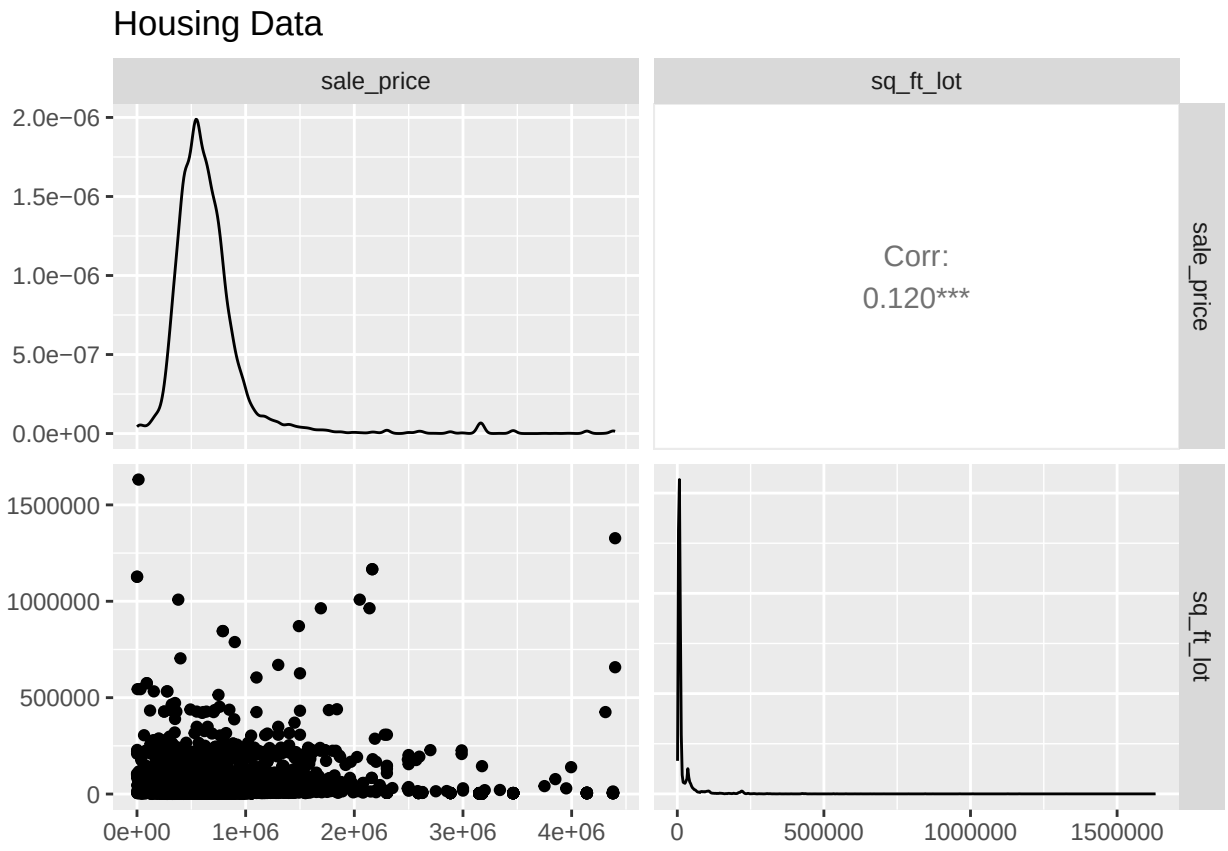
Explanation
A new "year" column was generated from the sale date to facilitate year-based calculations.

Correalation with Square feet

```r
#Identify Relationship between Sale Price and Sq_ft
library("ggplot2")
library("GGally")
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
ggpairs(housing_data[,c(2,17)], title="Housing Data")
```



Findings

The size of the lot (sq_ft_lot) is positively correlated with the sale price, but the relationship is relatively weak. This suggests that other factors besides lot size might be more significant determinants of the sale price.

The right-skewed distribution of sale_price indicates that there are a few high-priced houses that are pulling the average sale price up. This could be due to other factors.

Linear Regression Model with Sq_ft_lot (Model1)

```r
#Linear Regression Model
lm_sale_sqft <- lm(sale_price ~ sq_ft_lot, data = housing_data)
```

```r
#Summary of Lm
summary(lm_sale_sqft)
```

```
##
## Call:
## lm(formula = sale_price ~ sq_ft_lot, data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2016064  -194842   -63293    91565  3735109
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.418e+05  3.800e+03  168.90   <2e-16 ***
## sq_ft_lot   8.510e-01  6.217e-02   13.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 401500 on 12863 degrees of freedom
## Multiple R-squared:  0.01435,    Adjusted R-squared:  0.01428
## F-statistic: 187.3 on 1 and 12863 DF,  p-value: < 2.2e-16
```

Findings

Coefficients

(Intercept): This is the predicted sale_price when sq_ft_lot is 0. In this case, it's $641,800.

sq_ft_lot: This represents the change in sale_price for every one-unit increase in sq_ft_lot. So, for each additional square foot of lot size, the sale_price increases by $0.851. Both coefficients are statistically significant (indicated by the Pr(>|t|) values being much smaller than 0.05).

Model Fit

Residual standard error: This is the average distance between the actual sale_price and the predicted sale_price by the model. In this case, it's $401,500.

Multiple R-squared: This measures the proportion of variance in sale_price explained by the model. Here, it's 0.01435, meaning that only about 1.435% of the variation in sale_price can be explained by sq_ft_lot.

Adjusted R-squared: This is a similar metric to R-squared but penalizes for the number of predictors in the model. In this case, it's 0.01428, indicating that adding more predictors might not significantly improve the model's fit.

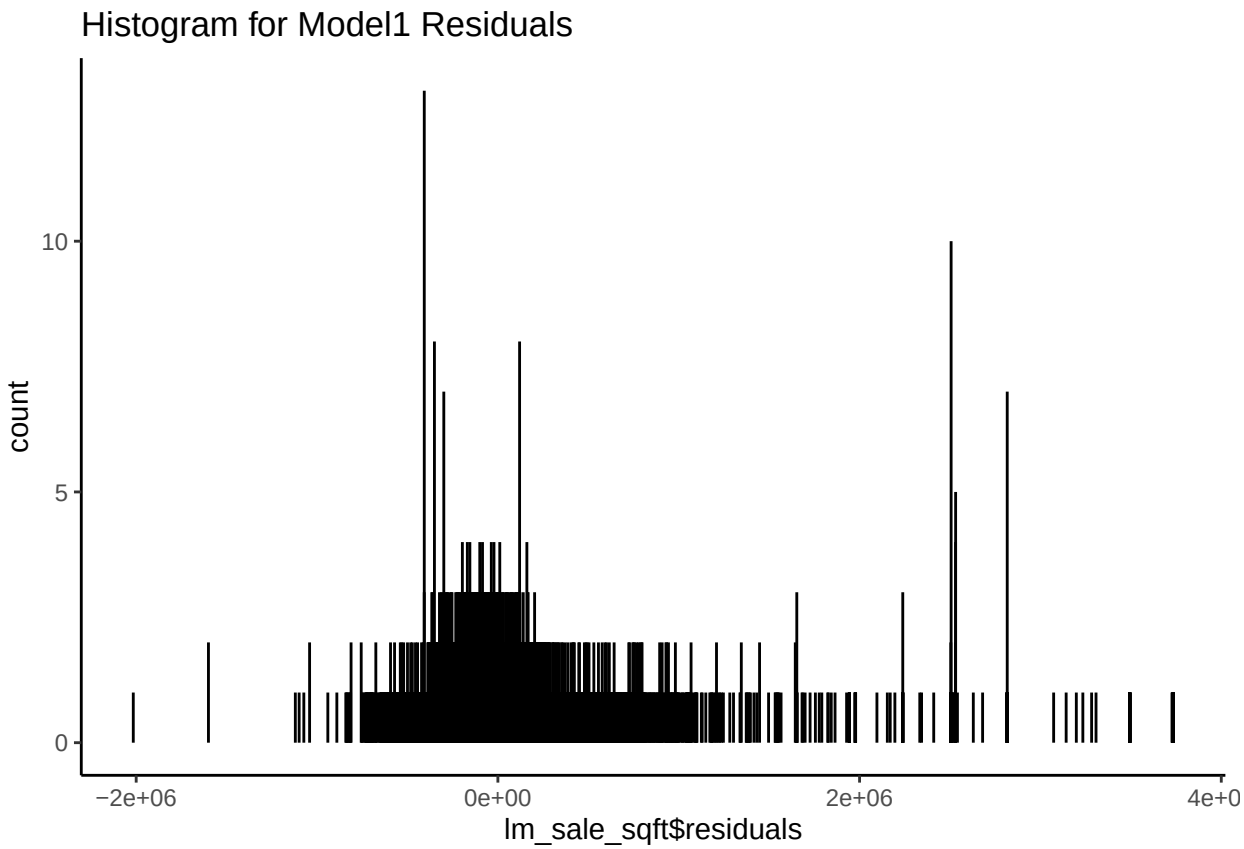
F-statistic and p-value: These test the overall significance of the model. A p-value of less than 0.05 suggests that the model is statistically significant, meaning that at least one predictor (in this case, sq_ft_lot) is significantly related to the sale_price.

In Summary

The model indicates that there is a significant positive relationship between sale_price and sq_ft_lot. However, only a very small portion of the variation in sale_price can be explained by this variable. This suggests that other factors may also influence the sale_price much more significantly.

Residuals for Model 1

```
ggplot(data=housing_data, aes(lm_sale_sqft$residuals)) +
geom_histogram(binwidth = 10, color = "black", fill = "purple4") +
theme(panel.background = element_rect(fill = "white"),
axis.line.x=element_line(),
axis.line.y=element_line()) +
ggtitle("Histogram for Model1 Residuals")
```

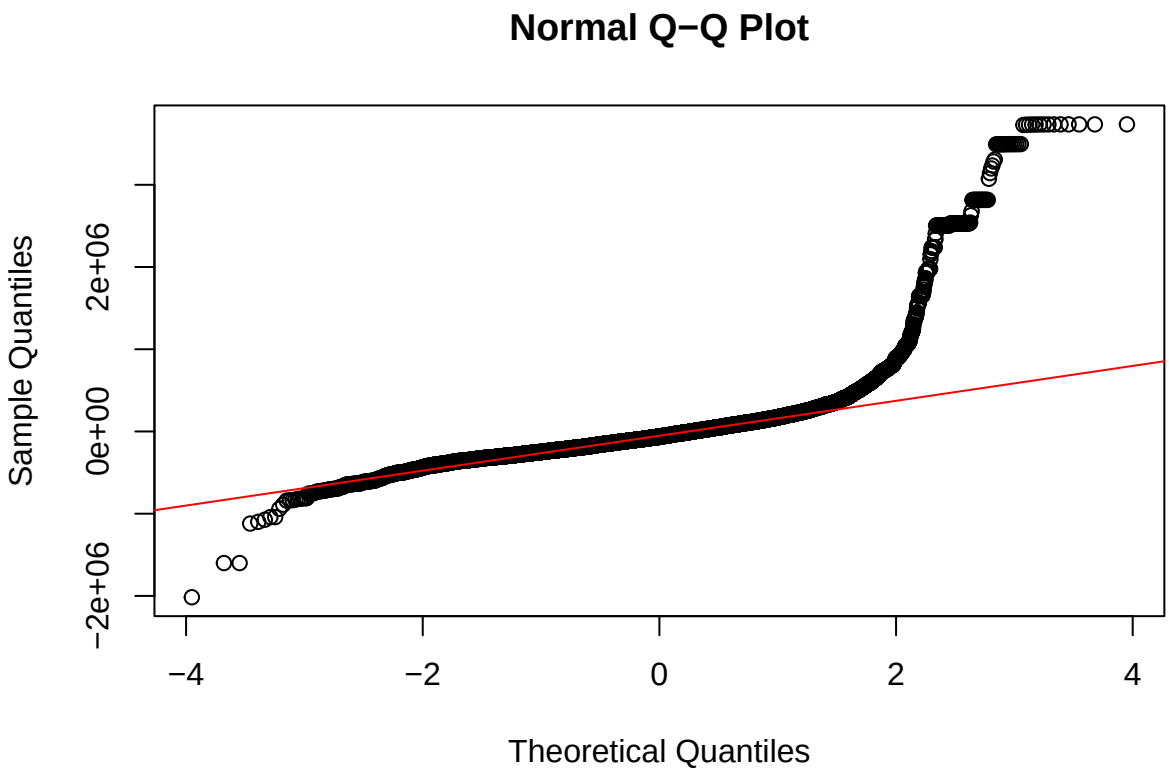


Findings

Central Tendency: The majority of residuals are clustered around 0, indicating that the model's predictions are generally accurate for most data points.
Skewness: There's a slight right-skewness, indicating that there are a few larger positive residuals.
Outliers: There appear to be some outliers, particularly on the right side of the histogram.

Residuals QQ Plot for Model 1

```
qqnorm(lm_sale_sqft$residuals)
qqline(lm_sale_sqft$residuals, col = "red")
```

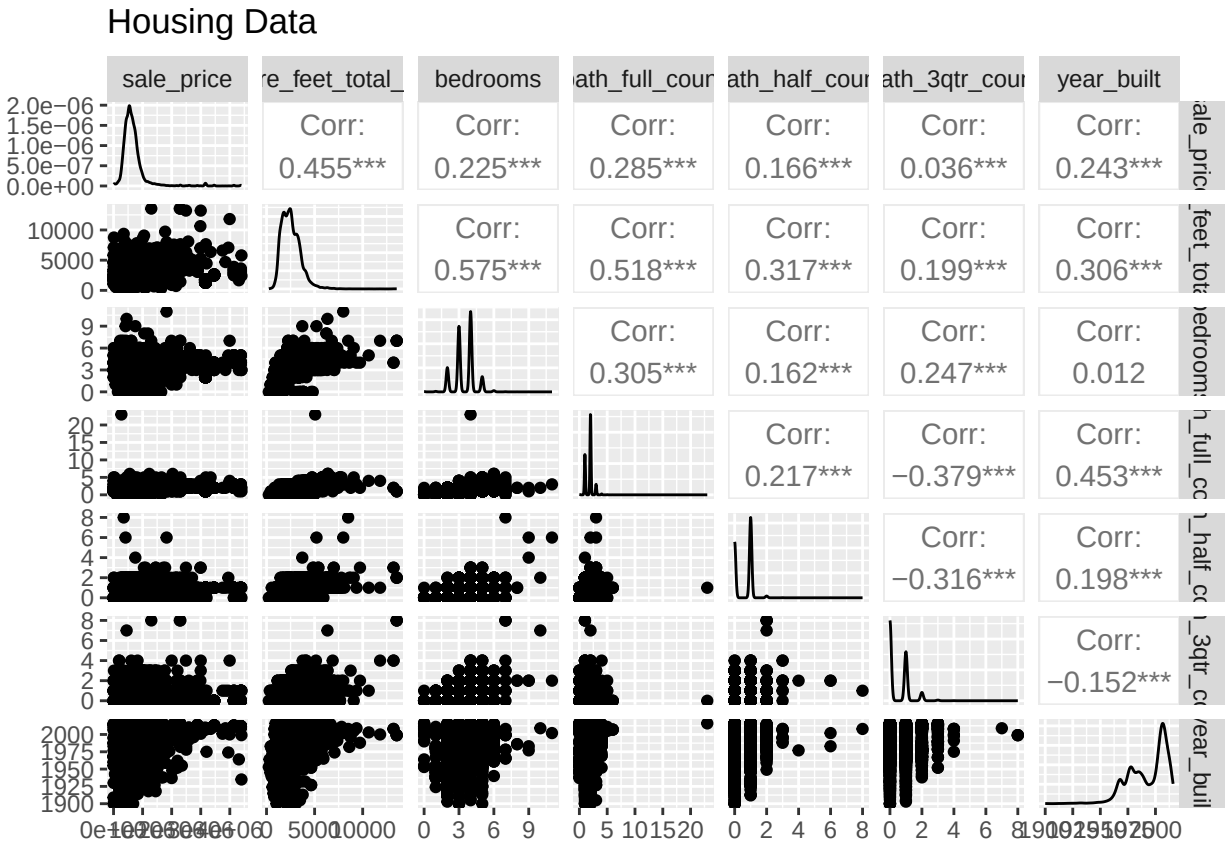


Findings

Deviation from Normality: The points in the QQ plot deviate from the red line (representing the normal distribution) in the tails, especially in the upper tail. This suggests that the residuals are not normally distributed.

Finding Correlation with other

```
#Identify Relationship between Sale Price and other
ggpairs(housing_data[,c(2,9,10,11,12,13,14)], title="Housing Data")
```



Findings

Based on the analysis, we have identified significant relationships between sale_price and the following variables: square_feet_total_living, bedrooms, bath_full_count, and year_built. However, bath_half_count and bath_3qtr_count appear to have negligible or no impact on sale_price. For our subsequent modeling efforts, we will focus on the variables that have demonstrated a meaningful association with the target variable.

Linear regression model with multiple varaibles (Model 2)

Explanation

Although our preliminary analysis revealed a positive correlation between lot size and sale price, the correlation coefficient indicates a moderate association. This suggests that other factors significantly contribute to housing prices. To explore these additional influences, we will construct a regression model incorporating variables such as the number of bedrooms, full bathrooms, overall living area, and property age. This expanded model aims to provide a more comprehensive understanding of the factors driving sale prices.

```
#Linear Regression Model
lm_sale_multiple <- lm(sale_price ~ square_feet_total_living + bedrooms + bath_full_count
                       + year_built, data = housing_data)
#Summary of Lm
summary(lm_sale_multiple)
```

```
##
## Call:
## lm(formula = sale_price ~ square_feet_total_living + bedrooms +
##     bath_full_count + year_built, data = housing_data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1719151  -120511   -42398    45744  3904824
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -4.430e+06  4.195e+05 -10.559  < 2e-16 ***
## square_feet_total_living  1.744e+02  4.423e+00  39.424  < 2e-16 ***
## bedrooms                 -1.375e+04  4.517e+03  -3.045  0.00234 **
## bath_full_count           1.730e+04  6.095e+03   2.838  0.00454 **
## year_built                2.340e+03  2.117e+02  11.053  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 357300 on 12860 degrees of freedom
## Multiple R-squared:  0.2194, Adjusted R-squared:  0.2192
## F-statistic: 903.7 on 4 and 12860 DF,  p-value: < 2.2e-16
```

Findings

Coefficients

(Intercept): This is the predicted sale_price when all predictor variables are 0. In this case, it's –$4,430,000. However, interpreting this intercept directly might not be meaningful, as it's unlikely to have a practical interpretation.

square_feet_total_living, bedrooms, bath_full_count, year_built: These coefficients represent the change in sale_price for every one-unit increase in the corresponding predictor variable, holding all other variables constant.For example, for each additional square foot of living area (square_feet_total_living), the sale_price increases by $174.40

Model Fit

Residual standard error: This is the average distance between the actual sale_price and the predicted sale_price by the model. In this case, it's $357,300.

Multiple R-squared: This measures the proportion of variance in sale_price explained by the model. Here, it's 0.2194, meaning that about 21.94% of the variation in sale_price can be explained by the combined effect of the predictor variables.

Adjusted R-squared: This is a similar metric to R-squared but penalizes for the number of predictors in the model. In this case, it's 0.2194, indicating that adding more predictors might not significantly improve the model's fit.

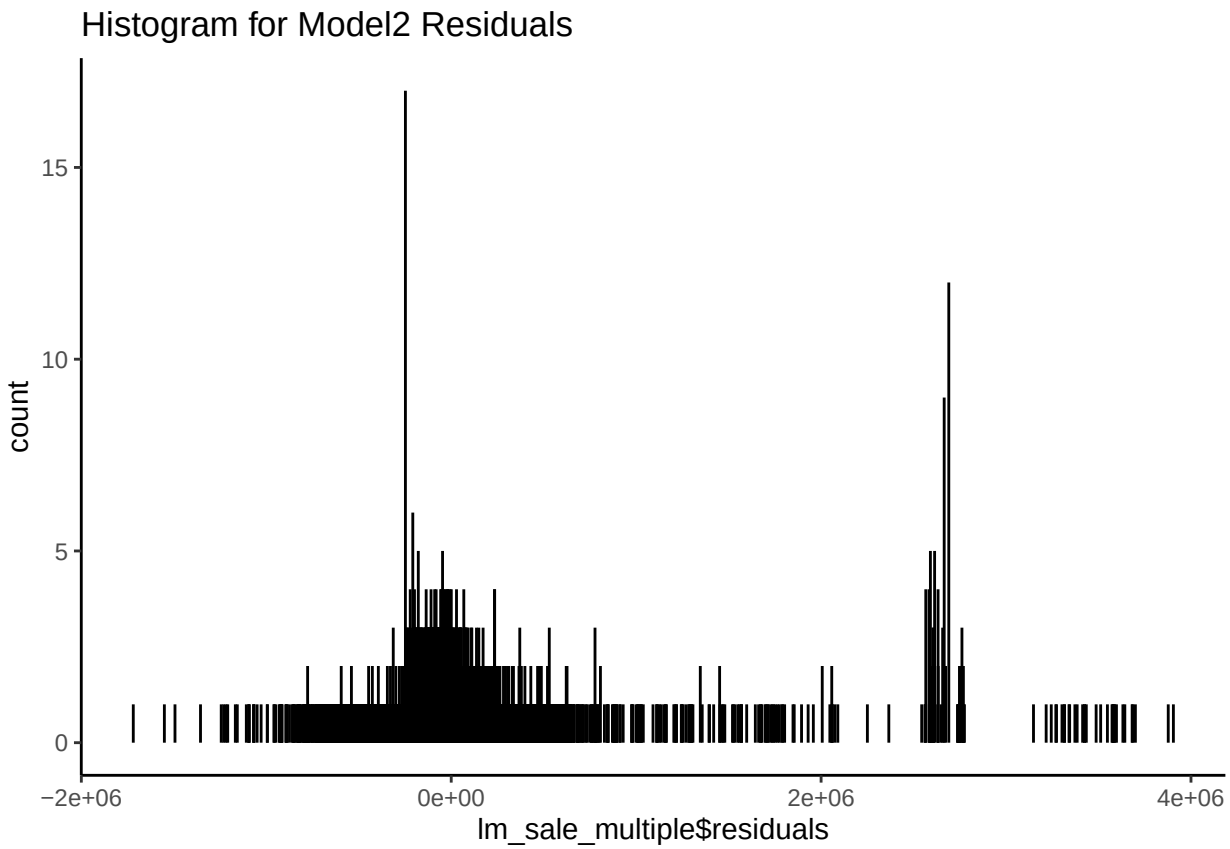
F-statistic and p-value: These test the overall significance of the model. A p-value of less than 0.05 suggests that the model is statistically significant, meaning that at least one predictor is significantly related to the sale_price.

In Summary

The model indicates that all predictor variables are significantly related to sale_price, with square_feet_total_living and year_built having the strongest effects.

Residuals for Model 2

```
ggplot(data=housing_data, aes(lm_sale_multiple$residuals)) +
geom_histogram(binwidth = 10, color = "black", fill = "purple4") +
theme(panel.background = element_rect(fill = "white"),
axis.line.x=element_line(),
axis.line.y=element_line()) +
ggtitle("Histogram for Model2 Residuals")
```

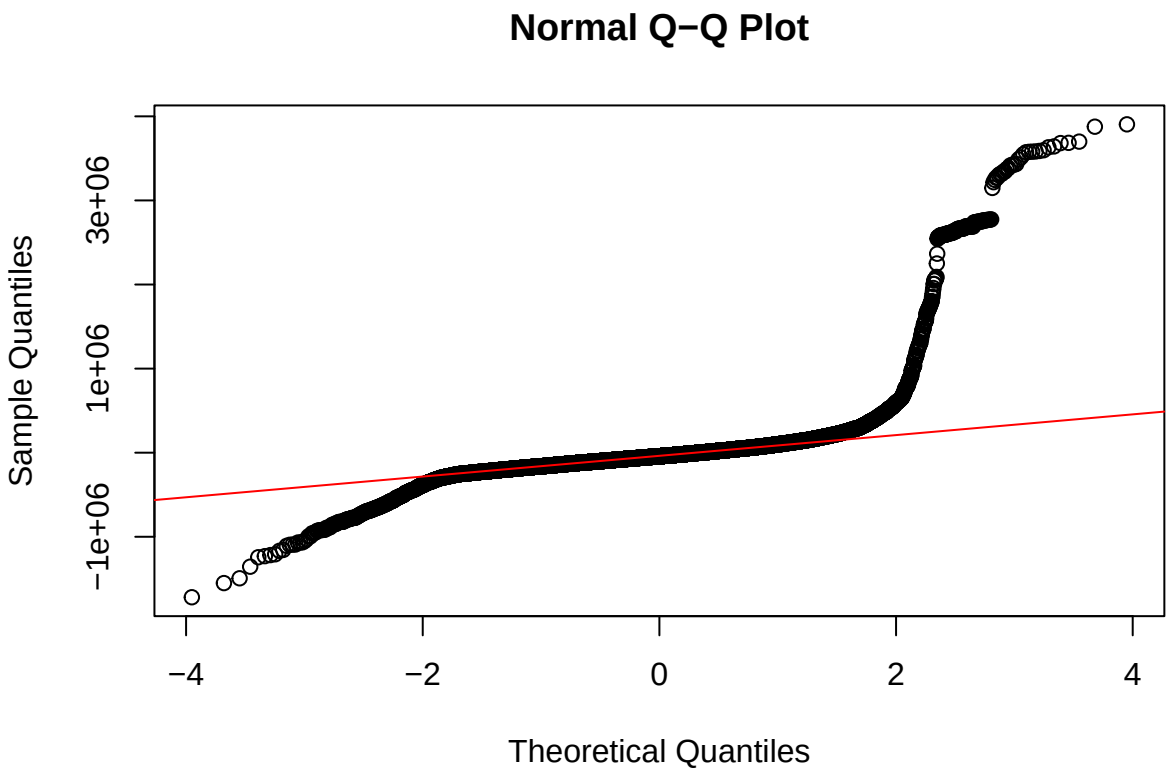


Findings

Central Tendency: A large number of residuals are clustered around 0.0000E+00, which is a positive sign as it indicates that the model's predictions are accurate for many data points.
Skewness: The histogram is slightly skewed to the right, suggesting that there are a few larger positive residuals.
Outliers: There appear to be some outliers, particularly on the right side of the histogram.

Residuals QQ Plot for Model 2

```
qqnorm(lm_sale_multiple$residuals)
qqline(lm_sale_multiple$residuals, col = "red")
```



Findings *Deviation from Normality: The points in the QQ plot deviate from the red line (representing the normal distribution) in the tails, especially in the upper tail. This suggests that the residuals are not normally distributed.*

Comparision between Model 1 and Model 2

Model 1:
Formula: sale_price ~ sq_ft_lot
Residual Standard Error (RSE): 401,500
Multiple R-squared: 0.01435
Adjusted R-squared: 0.01428
F-statistic: 187.3 (p-value < 2.2e-16)

Model 2:
Formula: sale_price ~ square_feet_total_living + bedrooms + bath_full_count + year_built
Residual Standard Error (RSE): 357,300
Multiple R-squared: 0.2194
Adjusted R-squared: 0.2192
F-statistic: 903.7 (p-value < 2.2e-16)

Findings:
R-Squared and Adjusted R-Squared: Model 1 has a very low R-squared value of 0.01435, meaning that only 1.4% of the variance in sale_price is explained by sq_ft_lot. Model 2 shows a significant improvement, with an R-squared of 0.2194, meaning 21.9% of the variance in sale_price is explained by the four predictors: square_feet_total_living, bedrooms, bath_full_count, and year_built. The increase in adjusted R-squared

from 0.01428 (Model 1) to 0.2192 (Model 2) suggests that Model 2 better fits the data, even after accounting for the number of predictors.

Residual Standard Error (RSE): The residual error decreased from 401,500 in Model 1 to 357,300 in Model 2, indicating that Model 2 produces more accurate predictions (lower spread of residuals).

F-statistic: The F-statistic in Model 1 is 187.3, whereas in Model 2, it jumps to 903.7, with a p-value < 2.2e-16 in both cases. This suggests that Model 2 is significantly better than Model 1 in explaining the variance in sale_price.

Anova Comparision

```
anova(lm_sale_sqft,lm_sale_multiple)
```

```
## Analysis of Variance Table
##
## Model 1: sale_price ~ sq_ft_lot
## Model 2: sale_price ~ square_feet_total_living + bedrooms + bath_full_count +
##     year_built
##   Res.Df        RSS Df  Sum of Sq    F    Pr(>F)
## 1  12863 2.0734e+15
## 2  12860 1.6420e+15  3 4.3134e+14 1126 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Findings

F-statistic (1126): This value is significantly large, indicating that the additional predictors in Model 2 (compared to Model 1) significantly improve the model's fit.
p-value (< 2.2e-16): The p-value is very small, confirming that the improvement in Model 2 is statistically significant at a very high level of confidence.

The ANOVA results strongly suggest that the second model, which includes additional predictors, is a significantly better fit to the data compared to the first model. This implies that the inclusion of square_feet_total_living, bedrooms, bath_full_count, and year_built provides a more comprehensive explanation of the variation in sale_price.

Residual Normality Comparision

Both models show signs of non-normal residuals and potential bias due to this deviation from normality.This suggests that the assumptions of normality for the residuals are violated in both models.This non-normality might imply that the models are missing some key features or interactions, or that transformations (e.g., log-transformation) may be necessary to address skewness and improve the fit of the model.

RMSE for Model 1

```
library(Metrics)
pre_model1 <- predict(object=lm_sale_sqft,newdata=housing_data)
rmse(housing_data$sale_price, pre_model1)
```

```
## [1] 401452.5
```

RMSE for Model 2

```
library(Metrics)
pre_model2 <- predict(object=lm_sale_multiple,newdata=housing_data)
rmse(housing_data$sale_price, pre_model2)
```

```
## [1] 357262
```

Findings

RMSE is a measure of the average error between the predicted and actual values. A lower RMSE generally indicates a better-fitting model.The RMSE of Model 2 (357262) is lower than the RMSE of Model 1 (401452.5). This indicates that Model 2, which includes additional predictor variables, has a better fit to the data. In other words, Model 2 makes more accurate predictions of the sale price compared to Model 1