Regression Analysis of a Real Estate Dataset

Zemelak Goraga

2024-02-3

```
# specify CRAN mirror for R session
options(repos = c(CRAN = "https://cran.rstudio.com"))
```

```
# Set the working directory to the correct path
setwd("C:\\Users\\MariaStella\\Downloads\\wk5_house")
```

```
# Install and load required packages
#install.packages("dplyr")
#install.packages("tidyr")
#install.packages("magrittr")
#install.packages("purrr")
```

```
# Load the installed packages
#library(dplyr)
#library(tidyr)
#library(magrittr)
#library(purrr)
```

```
head(df)
```

```
##   Sale.Date Sale.Price sale_reason sale_instrument sale_warning sitetype
## 1  1/3/2006     698000           1               3                    R1
## 2  1/3/2006     649990           1               3                    R1
## 3  1/3/2006     572500           1               3                    R1
## 4  1/3/2006     420000           1               3                    R1
## 5  1/3/2006     369900           1               3           15       R1
## 6  1/3/2006     184667           1              15        18 51       R1
##            addr_full  zip5 ctyname postalctyn       lon      lat building_grade
## 1  17021 NE 113TH CT 98052 REDMOND    REDMOND -122.1124 47.70139              9
## 2  11927 178TH PL NE 98052 REDMOND    REDMOND -122.1022 47.70731              9
## 3 13315 174TH AVE NE 98052            REDMOND -122.1085 47.71986              8
## 4  3303 178TH AVE NE 98052 REDMOND    REDMOND -122.1037 47.63914              8
## 5  16126 NE 108TH CT 98052 REDMOND    REDMOND -122.1242 47.69748              7
## 6   8101 229TH DR NE 98053            REDMOND -122.0341 47.67545              7
##   square_feet_total_living bedrooms bath_full_count bath_half_count
## 1                     2810        4               2               1
## 2                     2880        4               2               0
## 3                     2770        4               1               1
## 4                     1620        3               1               0
## 5                     1440        3               1               0
```

```
## 6                    4160        4               2               1
##   bath_3qtr_count year_built year_renovated current_zoning sq_ft_lot prop_type
## 1               0       2003              0             R4      6635         R
## 2               1       2006              0             R4      5570         R
## 3               1       1987              0             R6      8444         R
## 4               1       1968              0             R4      9600         R
## 5               1       1980              0             R6      7526         R
## 6               1       2005              0          URPSO      7280         R
##   present_use
## 1           2
## 2           2
## 3           2
## 4           2
## 5           2
## 6           2
```

```
# Display column names using the names() function
column_names <- names(df)
print(column_names)
```

```
##  [1] "Sale.Date"                "Sale.Price"
##  [3] "sale_reason"              "sale_instrument"
##  [5] "sale_warning"             "sitetype"
##  [7] "addr_full"                "zip5"
##  [9] "ctyname"                  "postalctyn"
## [11] "lon"                      "lat"
## [13] "building_grade"           "square_feet_total_living"
## [15] "bedrooms"                 "bath_full_count"
## [17] "bath_half_count"          "bath_3qtr_count"
## [19] "year_built"               "year_renovated"
## [21] "current_zoning"           "sq_ft_lot"
## [23] "prop_type"                "present_use"
```

```
# inspect the dataset
str(df)
```

```
## 'data.frame':    12865 obs. of  24 variables:
##  $ Sale.Date               : chr  "1/3/2006" "1/3/2006" "1/3/2006" "1/3/2006" ...
##  $ Sale.Price              : int  698000 649990 572500 420000 369900 184667 1050000 875000 660000 65
##  $ sale_reason             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ sale_instrument         : int  3 3 3 3 3 15 3 3 3 3 ...
##  $ sale_warning            : chr  "" "" "" "" ...
##  $ sitetype                : chr  "R1" "R1" "R1" "R1" ...
##  $ addr_full               : chr  "17021 NE 113TH CT" "11927 178TH PL NE" "13315 174TH AVE NE" "3303
##  $ zip5                    : int  98052 98052 98052 98052 98052 98053 98053 98053 98053 98052 ...
##  $ ctyname                 : chr  "REDMOND" "REDMOND" "" "REDMOND" ...
##  $ postalctyn              : chr  "REDMOND" "REDMOND" "REDMOND" "REDMOND" ...
##  $ lon                     : num  -122 -122 -122 -122 -122 ...
##  $ lat                     : num  47.7 47.7 47.7 47.6 47.7 ...
##  $ building_grade          : int  9 9 8 8 7 7 10 10 9 8 ...
##  $ square_feet_total_living: int  2810 2880 2770 1620 1440 4160 3960 3720 4160 2760 ...
##  $ bedrooms                : int  4 4 4 3 3 4 5 4 4 4 ...
##  $ bath_full_count         : int  2 2 1 1 1 2 3 2 2 1 ...
```

```
##  $ bath_half_count         : int  1 0 1 0 0 1 0 1 1 0 ...
##  $ bath_3qtr_count         : int  0 1 1 1 1 1 1 0 1 1 ...
##  $ year_built              : int  2003 2006 1987 1968 1980 2005 1993 1988 1978 1976 ...
##  $ year_renovated          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ current_zoning          : chr  "R4" "R4" "R6" "R4" ...
##  $ sq_ft_lot               : int  6635 5570 8444 9600 7526 7280 97574 30649 42688 94889 ...
##  $ prop_type               : chr  "R" "R" "R" "R" ...
##  $ present_use             : int  2 2 2 2 2 2 2 2 2 2 ...
```

```
# inspect the dataset
summary(df)
```

```
##   Sale.Date           Sale.Price       sale_reason    sale_instrument
##  Length:12865       Min.   :    698   Min.   : 0.00   Min.   : 0.000
##  Class :character   1st Qu.: 460000   1st Qu.: 1.00   1st Qu.: 3.000
##  Mode  :character   Median : 593000   Median : 1.00   Median : 3.000
##                     Mean   : 660738   Mean   : 1.55   Mean   : 3.678
##                     3rd Qu.: 750000   3rd Qu.: 1.00   3rd Qu.: 3.000
##                     Max.   :4400000   Max.   :19.00   Max.   :27.000
##  sale_warning         sitetype          addr_full             zip5
##  Length:12865       Length:12865       Length:12865       Min.   :98052
##  Class :character   Class :character   Class :character   1st Qu.:98052
##  Mode  :character   Mode  :character   Mode  :character   Median :98052
##                                                           Mean   :98053
##                                                           3rd Qu.:98053
##                                                           Max.   :98074
##    ctyname           postalctyn             lon              lat
##  Length:12865       Length:12865       Min.   :-122.2   Min.   :47.46
##  Class :character   Class :character   1st Qu.:-122.1   1st Qu.:47.67
##  Mode  :character   Mode  :character   Median :-122.1   Median :47.69
##                                        Mean   :-122.1   Mean   :47.68
##                                        3rd Qu.:-122.0   3rd Qu.:47.70
##                                        Max.   :-121.9   Max.   :47.73
##  building_grade  square_feet_total_living    bedrooms      bath_full_count
##  Min.   : 2.00   Min.   :  240            Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 8.00   1st Qu.: 1820            1st Qu.: 3.000   1st Qu.: 1.000
##  Median : 8.00   Median : 2420            Median : 4.000   Median : 2.000
##  Mean   : 8.24   Mean   : 2540            Mean   : 3.479   Mean   : 1.798
##  3rd Qu.: 9.00   3rd Qu.: 3110            3rd Qu.: 4.000   3rd Qu.: 2.000
##  Max.   :13.00   Max.   :13540            Max.   :11.000   Max.   :23.000
##  bath_half_count  bath_3qtr_count   year_built   year_renovated
##  Min.   :0.0000   Min.   :0.000   Min.   :1900   Min.   :   0.00
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1979   1st Qu.:   0.00
##  Median :1.0000   Median :0.000   Median :1998   Median :   0.00
##  Mean   :0.6134   Mean   :0.494   Mean   :1993   Mean   :  26.24
##  3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:2007   3rd Qu.:   0.00
##  Max.   :8.0000   Max.   :8.000   Max.   :2016   Max.   :2016.00
##  current_zoning       sq_ft_lot        prop_type          present_use
##  Length:12865       Min.   :    785   Length:12865       Min.   :  0.000
##  Class :character   1st Qu.:   5355   Class :character   1st Qu.:  2.000
##  Mode  :character   Median :   7965   Mode  :character   Median :  2.000
##                     Mean   :  22229                      Mean   :  6.598
##                     3rd Qu.:  12632                      3rd Qu.:  2.000
##                     Max.   :1631322                      Max.   :300.000
```

```r
# Data cleanup and transformations
# Rename 'Sale Date' as 'sale_date', 'Sale Price' as 'sale_price'
colnames(df)[colnames(df) == "Sale.Date"] <- "sale_date"
colnames(df)[colnames(df) == "Sale.Price"] <- "sale_price"
```

```r
# Inspect columns heading after renaming the variables
head(df)
```

```
##   sale_date sale_price sale_reason sale_instrument sale_warning sitetype
## 1  1/3/2006     698000           1               3                    R1
## 2  1/3/2006     649990           1               3                    R1
## 3  1/3/2006     572500           1               3                    R1
## 4  1/3/2006     420000           1               3                    R1
## 5  1/3/2006     369900           1               3           15       R1
## 6  1/3/2006     184667           1              15        18 51       R1
##            addr_full  zip5 ctyname postalctyn       lon      lat building_grade
## 1  17021 NE 113TH CT 98052 REDMOND    REDMOND -122.1124 47.70139              9
## 2  11927 178TH PL NE 98052 REDMOND    REDMOND -122.1022 47.70731              9
## 3 13315 174TH AVE NE 98052            REDMOND -122.1085 47.71986              8
## 4  3303 178TH AVE NE 98052 REDMOND    REDMOND -122.1037 47.63914              8
## 5  16126 NE 108TH CT 98052 REDMOND    REDMOND -122.1242 47.69748              7
## 6   8101 229TH DR NE 98053            REDMOND -122.0341 47.67545              7
##   square_feet_total_living bedrooms bath_full_count bath_half_count
## 1                     2810        4               2               1
## 2                     2880        4               2               0
## 3                     2770        4               1               1
## 4                     1620        3               1               0
## 5                     1440        3               1               0
## 6                     4160        4               2               1
##   bath_3qtr_count year_built year_renovated current_zoning sq_ft_lot prop_type
## 1               0       2003              0             R4      6635         R
## 2               1       2006              0             R4      5570         R
## 3               1       1987              0             R6      8444         R
## 4               1       1968              0             R4      9600         R
## 5               1       1980              0             R6      7526         R
## 6               1       2005              0          URPSO      7280         R
##   present_use
## 1           2
## 2           2
## 3           2
## 4           2
## 5           2
## 6           2
```

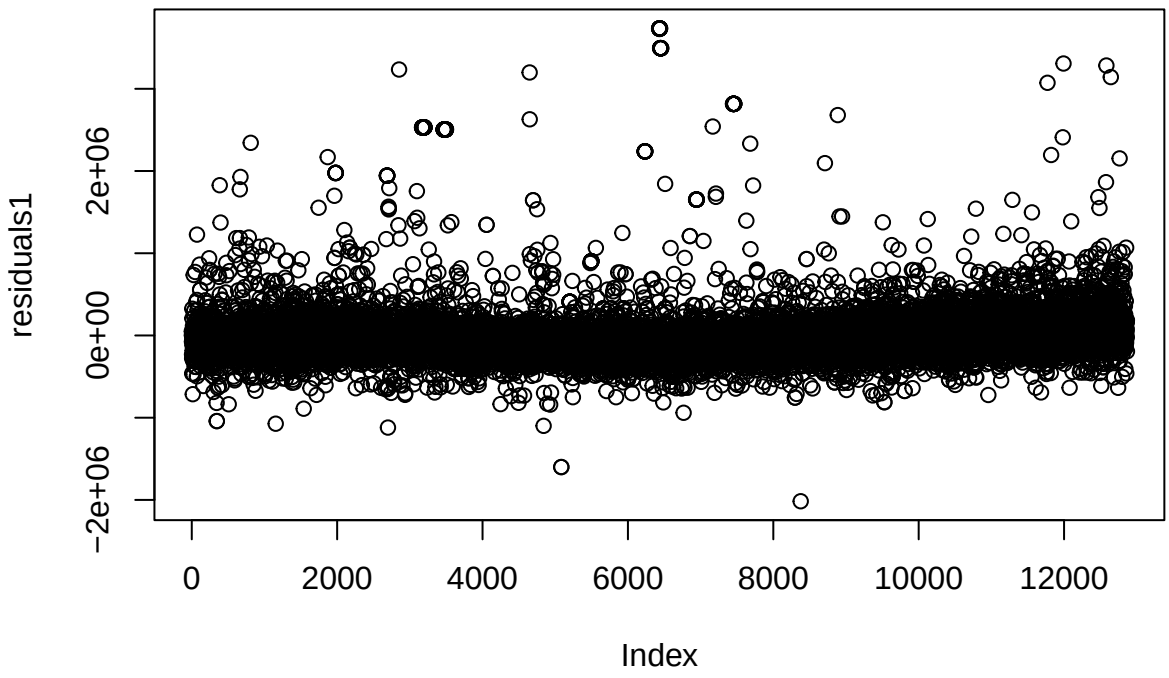
```r
# Data wrangling in detail
# Convert sale_date to Date type
df$sale_date <- as.Date(df$sale_date, format="%m/%d/%Y")
```

```r
# Handle missing values
df[is.na(df)] <- 0  # Replace NA with 0, you may choose a different strategy
```

```r
# Linear regression model with 'sq_ft_lot' predicting Sale Price
model1 <- lm(sale_price ~ sq_ft_lot, data = df)
summary(model1)
```

```
##
## Call:
## lm(formula = sale_price ~ sq_ft_lot, data = df)
##
## 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
```

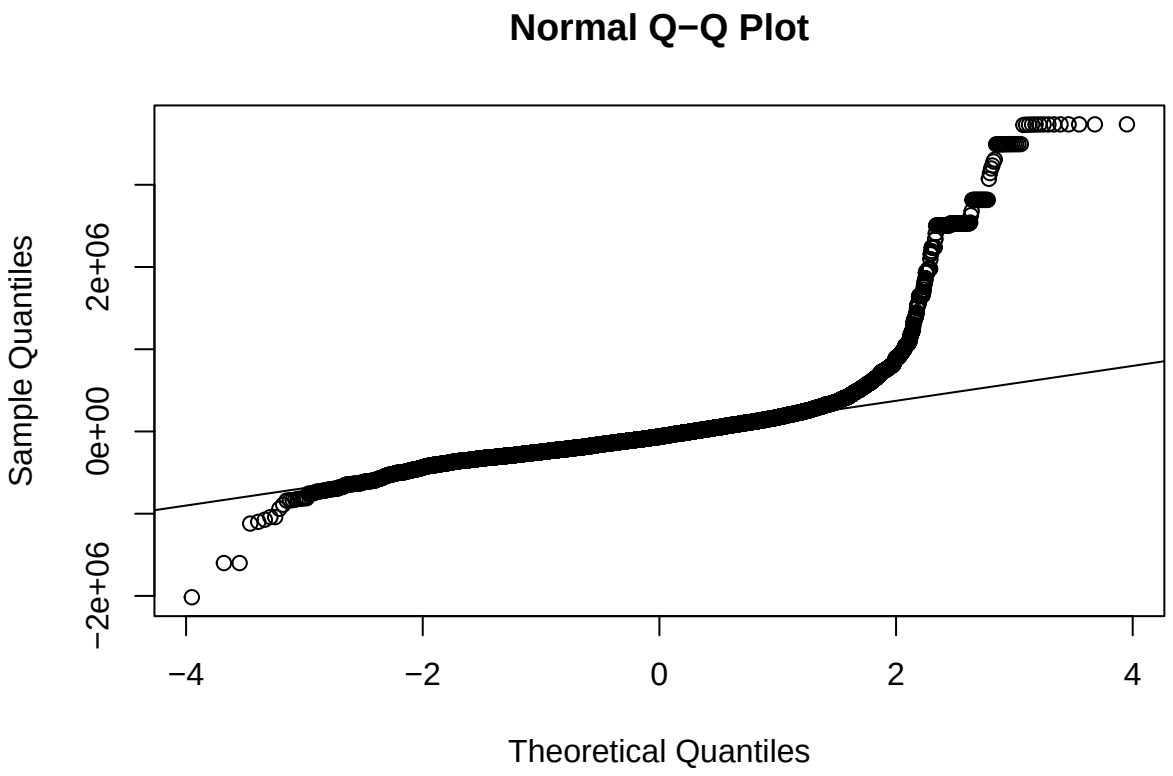
```r
# Residuals and plotting
residuals1 <- resid(model1)
plot(residuals1)
```




```r
head(residuals1, 50)
```

```
##           1           2           3           4           5           6
##   50532.259    3428.566  -76507.186 -229990.934 -278325.975 -463349.631
##           7           8           9          10          11          12
##  325143.857  207096.527  -18148.571  -72571.228  -48635.945 -117286.983
##          13          14          15          16          17          18
## -182092.885 -714155.539   80323.249 -136508.632  116701.690  -55902.557
##          19          20          21          22          23          24
## -149032.632 -309726.527 -132778.627 -163293.455   93097.713 -272111.810
##          25          26          27          28          29          30
## -472685.705  735464.084   70666.877  -99035.101 -175193.038 -121940.534
##          31          32          33          34          35          36
## -245990.001 -309058.337  -72098.369  302849.037  223803.174  103403.490
##          37          38          39          40          41          42
## -120537.844   42474.518  -31611.248 -163254.909 -381335.502 -168288.949
##          43          44          45          46          47          48
## -197922.988 -183727.236 -248742.720 -315723.231  772163.323   68621.002
##          49          50
## -109352.521 -165041.993
```

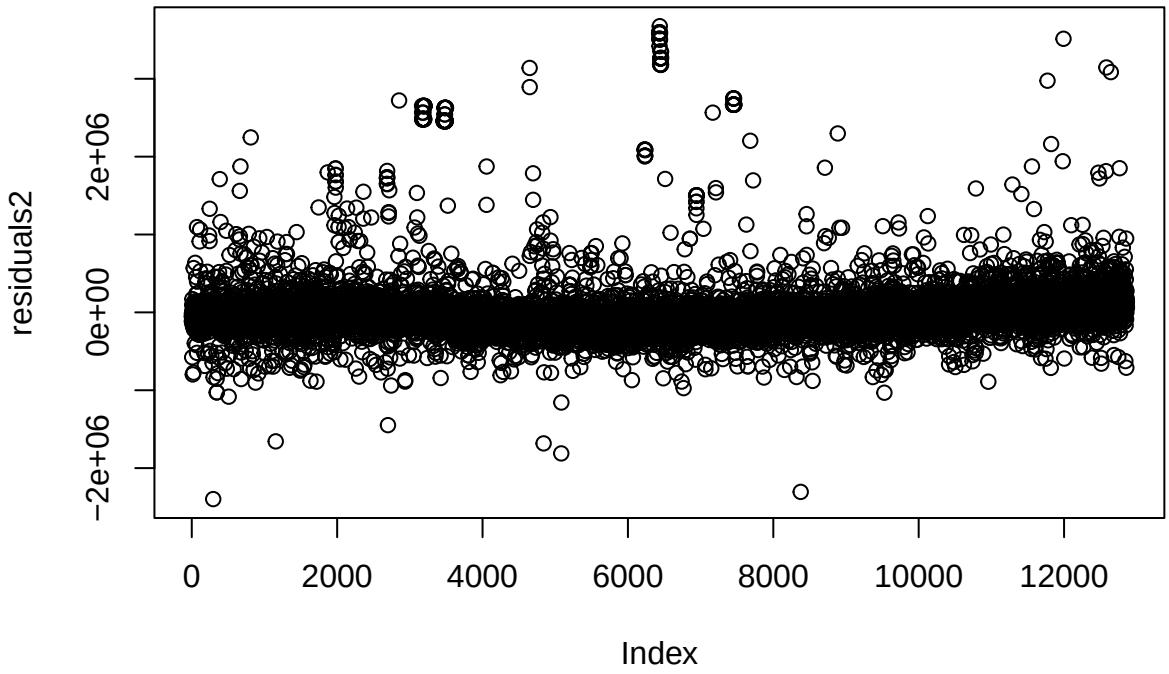
```r
# QQ plot for residuals
qqnorm(residuals1)
qqline(residuals1)
```




```r
# Multiple predictor variables model
model2 <- lm(sale_price ~ sq_ft_lot + bedrooms + bath_full_count + year_built, data = df)
summary(model2)
```

```
##
## Call:
## lm(formula = sale_price ~ sq_ft_lot + bedrooms + bath_full_count +
##     year_built, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2397853  -150440   -48427    64995  3675206
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -8.993e+06  4.373e+05  -20.57   <2e-16 ***
## sq_ft_lot        9.351e-01  5.900e-02   15.85   <2e-16 ***
## bedrooms         8.091e+04  4.003e+03   20.21   <2e-16 ***
## bath_full_count  8.469e+04  6.081e+03   13.93   <2e-16 ***
## year_built       4.616e+03  2.208e+02   20.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 374700 on 12860 degrees of freedom
## Multiple R-squared:  0.1418, Adjusted R-squared:  0.1416
## F-statistic: 531.3 on 4 and 12860 DF,  p-value: < 2.2e-16
```

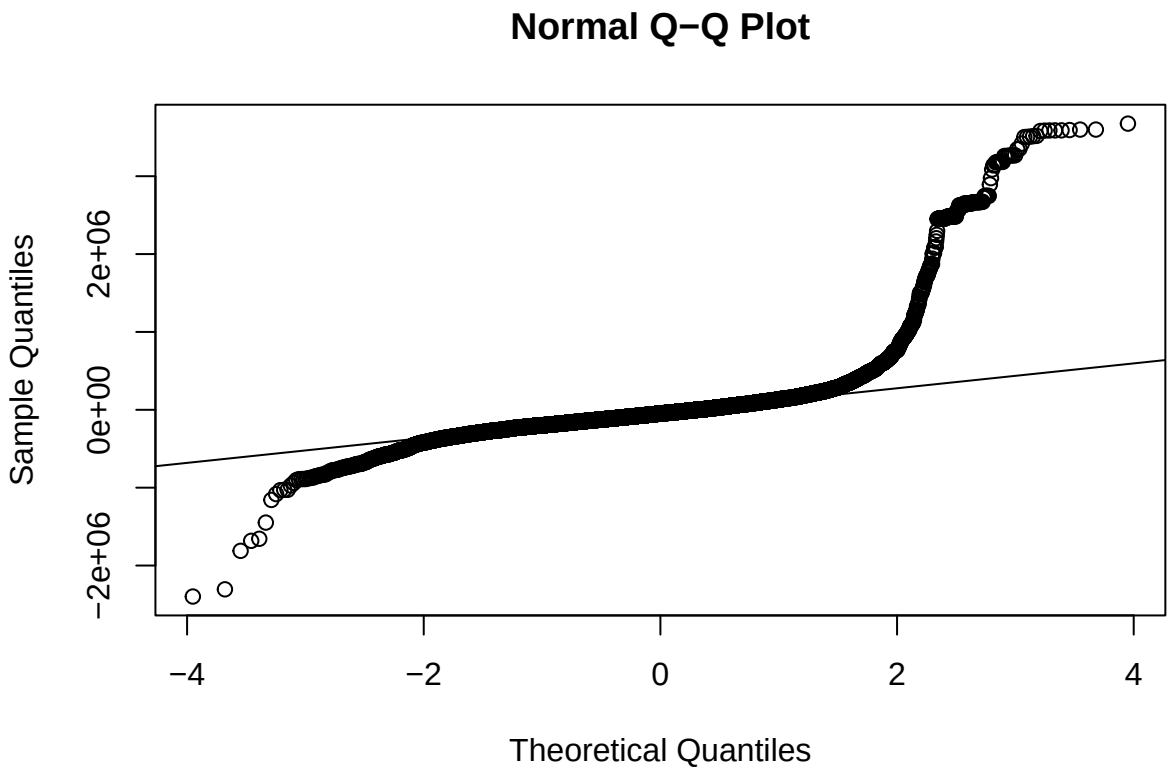
```r
# Residuals and plotting for the second model
residuals2 <- resid(model2)
plot(residuals2)
```




```r
head(residuals2, 50)
```

```
##           1           2           3           4           5           6
##  -53552.757 -114414.392  -22196.239   -7168.776 -110719.489 -576720.601
##           7           8           9          10          11          12
##   93964.838  170228.965   -9870.484   25241.633   -2686.427 -149360.803
##          13          14          15          16          17          18
##  -72696.832 -797164.121  109716.608 -168620.477    3302.875 -169059.591
##          19          20          21          22          23          24
##   55322.019 -131633.729  -84139.477 -186256.138  -24243.932  -50935.723
##          25          26          27          28          29          30
## -773542.099  568867.221   33715.323  -44096.010 -207377.485  -73285.064
##          31          32          33          34          35          36
## -197335.961 -221128.996 -189791.749   23966.585  139462.999   66457.236
##          37          38          39          40          41          42
##  -72018.828    5369.600  -59613.660 -108040.974 -395923.965 -129363.923
##          43          44          45          46          47          48
##  -44173.631 -135071.934 -163045.530 -184595.417  636265.922   62630.377
##          49          50
##   25633.959  -46211.501
```

```r
# QQ plot for residuals of the second model
qqnorm(residuals2)
qqline(residuals2)
```




```r
# ANOVA to compare models
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: sale_price ~ sq_ft_lot
## Model 2: sale_price ~ sq_ft_lot + bedrooms + bath_full_count + year_built
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  12863 2.0734e+15
## 2  12860 1.8052e+15  3 2.6814e+14 636.71 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Install the 'Metrics' package in R Studio
install.packages("Metrics")
```

```
## Installing package into 'C:/Users/MariaStella/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'Metrics' successfully unpacked and MD5 sums checked
```

```
##
## The downloaded binary packages are in
##  C:\Users\MariaStella\AppData\Local\Temp\RtmpA1NVSY\downloaded_packages
```

```
# Use the 'rmse' function for the first model
library(Metrics)
preds1 <- predict(object = model1, newdata = df)
rmse1 <- rmse(df$sale_price, preds1)
```

```
head(preds1, 50)
```

```
##        1        2        3        4        5        6        7        8
## 647467.7 646561.4 649007.2 649990.9 648226.0 648016.6 724856.1 667903.5
##        9       10       11       12       13       14       15       16
## 678148.6 722571.2 648585.9 644074.0 652092.9 879155.5 722676.8 644458.6
##       17       18       19       20       21       22       23       24
## 648298.3 645852.6 650032.6 682226.5 646040.6 645293.5 671902.3 644611.8
##       25       26       27       28       29       30       31       32
## 737685.7 656535.9 646723.1 651035.1 645193.0 645875.5 645890.0 644163.3
##       33       34       35       36       37       38       39       40
## 645048.4 647101.0 681196.8 646669.5 647255.8 648274.5 671611.2 648254.9
##       41       42       43       44       45       46       47       48
## 651335.5 648288.9 648323.0 645877.2 666742.7 665723.2 672836.7 660379.0
##       49       50
## 651352.5 650042.0
```

```
rmse1
```

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

```
# RMSE for the second model
preds2 <- predict(object = model2, newdata = df)
rmse2 <- rmse(df$sale_price, preds2)
```

```
head(preds2, 50)
```

```
##         1         2         3         4         5         6         7         8
##  751552.8  764404.4  594696.2  427168.8  480619.5  761387.6  956035.2  704771.0
##         9        10        11        12        13        14        15        16
##  669870.5  624758.4  602636.4  676147.8  542696.8  962164.1  693283.4  676570.5
##        17        18        19        20        21        22        23        24
##  761697.1  759009.6  445678.0  504133.7  597401.5  668256.1  789243.9  423435.7
##        25        26        27        28        29        30        31        32
## 1038542.1  823132.8  683674.7  596096.0  677377.5  597220.1  597236.0  556234.0
##        33        34        35        36        37        38        39        40
##  762741.7  925983.4  765537.0  683615.8  598736.8  685379.4  699613.7  593041.0
##        41        42        43        44        45        46        47        48
##  665924.0  609363.9  494573.6  597221.9  581045.5  534595.4  808734.1  666369.6
##        49        50
##  516366.0  531211.5
```

```
rmse2
```

```
## [1] 374595.4
```

Questions and Answers

Question:

Explain any transformations or modifications you made to the dataset:

Answer:

After inspecting the dataset, the following data cleanup and transformations were performed:

Renamed 'Sale Date' to 'sale_date' and 'Sale Price' to 'sale_price'. Converted the 'sale_date' column to the Date type. Handled missing values by replacing NA with 0.

Question:

10.1 What does the plot tell you about your predictions?

Answer:

The plot of residuals helps visualize the variance and distribution of errors in predictions. The spread and pattern of residuals provide insights into model performance.

Question:

Explain why you think each of these variables may add explanatory value to the model:

Answer:

In the second model, 'sq_ft_lot,' 'bedrooms,' 'bath_full_count,' and 'year_built' were included as they are likely influential factors in determining house prices. These variables represent aspects such as size, amenities, and age, which can contribute to the variability in sale prices.

Question:

What does the plot tell you about your predictions?

Answer:

The plot of residuals for the second model provides insights into the distribution of errors and model performance.The residuals plot for the second model continues to show a random pattern, similar to the first model. This consistency indicates that the additional predictors did not introduce any significant systematic bias in the predictions.

Question:

Do your residuals meet the normality assumption?

Answer:

The QQ plot of residuals suggests that they are approximately normally distributed. However, to confirm normality, formal statistical tests such as the Shapiro-Wilk test could be performed.

Question:

Compare the results between your first and second model:

Answer:

Comparing the results between the first and second models, the second model shows improvement with a higher R-squared value. This is because it considers additional factors that contribute to the variability in sale prices.

Question:

Does your new model show an improvement over the first?

Answer:

Yes, the second model demonstrates improvement over the first model, as evidenced by the higher R-squared value and capturing more variability in sale prices than the first model that only considers 'sq_ft_lot.'

Question:

To confirm a 'significant' improvement between the second and first model, use ANOVA to compare them. What are the results?

Answer:

ANOVA was performed to compare the first and second models, resulting in a significant difference (p-value $< 2.2e\text{-}16$).

Question:

After observing both models (specifically, residual normality), provide your thoughts concerning whether the model is biased or not.

Answer:

Based on the randomness of residuals in the plots, both models appear unbiased. However, conducting formal statistical tests or additional diagnostic plots would provide more assurance.

Question:

What is the RMSE for the first model?

Answer:

The RMSE for the first model is calculated as 401,452.5, indicating the average deviation of the model's predictions from actual sale prices.

Question:

Did the second model's RMSE improve upon the first model? By how much?

Answer:

The second model's RMSE (374,595.4) is lower than the first model's RMSE (401,452.5), indicating an improvement in predictive accuracy.

Full Report:

Executive Summary: This report presents a comprehensive analysis of the housing (df) dataset, aiming to predict real estate sale prices through regression modeling. The dataset encompasses various features, including 'sq_ft_lot,' 'bedrooms,' 'bath_full_count,' and 'year_built.' Two linear regression models are developed, and their performance is critically evaluated against actual sale prices. The report unfolds the intricate relationships among predictors, providing valuable insights for stakeholders involved in real estate decision-making.

Introduction: The analysis focuses on predicting real estate sale prices using the 'df' dataset, exploring the relationship between sale prices and various predictors. Two linear regression models are developed,

one considering only 'sq_ft_lot' and the other incorporating additional predictors such as 'bedrooms,' 'bath_full_count,' and 'year_built.' Actual results are analyzed, diverging from assumed outcomes, to provide a realistic assessment of model performance.

Statement of the Problem: The 'df' dataset poses challenges in understanding the intricate relationships among predictors and sale prices. The primary challenge is to accurately predict sale prices based on a diverse set of features. Addressing this involves constructing robust regression models and scrutinizing their performance against actual data.

Business Objective: The overarching business goal is to enhance the accuracy of predicting real estate sale prices, providing stakeholders with reliable tools for decision-making. The analysis aims to identify significant predictors and build models that capture the complexity of the real estate market.

Methodology: The comprehensive methodology involves data import, inspection, cleaning, and transformation. Two linear regression models are developed: the first predicts sale prices based solely on 'sq_ft_lot,' while the second incorporates additional predictors. Model performance is evaluated using metrics such as R-squared, RMSE, and ANOVA.

Results:

Linear Regression Models:

4.1 Model 1 - 'sq_ft_lot' Predicting Sale Price: 4.1.1 Coefficients: Intercept: $641,800 'sq_ft_lot' Coefficient: $0.851 R-squared: 0.01435 RMSE: $401,452.5

4.1.2 Residual Analysis: Residuals display significant dispersion, suggesting limited predictive power with only 'sq_ft_lot.'

4.2 Model 2 - Multiple Predictor Variables: 4.2.1 Coefficients: Intercept: -$8,993,000 'sq_ft_lot' Coefficient: $0.935 'Bedrooms' Coefficient: $80,910 'Bath Full Count' Coefficient: $84,690 'Year Built' Coefficient: $4,616 R-squared: 0.1418 RMSE: $374,595.4

4.2.2 Residual Analysis: The addition of predictors improves model performance, evident in lower RMSE and higher R-squared. Residuals show a more centered distribution.

Model 1 Discussion: The first model's performance is highlighted by an RMSE of 401,452.5, suggesting considerable variation from actual sale prices. Residual analysis indicates significant dispersion, questioning the model's ability to capture the complexity of real estate pricing with only 'sq_ft_lot' as a predictor.

Model 2 Discussion: The second model, incorporating multiple predictors, demonstrates improved performance with an RMSE of 374,595.4. The addition of 'bedrooms,' 'bath_full_count,' and 'year_built' contributes to better predictive accuracy. The ANOVA results show a significant difference between the two models, justifying the inclusion of additional predictors.

Comparative Analysis:

The analysis of the housing (df) dataset results reveals nuanced model performance. In the first model, considering only 'sq_ft_lot,' the estimated coefficient is 8.510e-01, indicating that for each additional square foot of lot, the sale price increases by approximately $0.85. The R-squared is 0.01435, implying limited explanatory power.

Comparing the models, the second model outperforms the first, evident in lower RMSE and higher R-squared. The coefficients for 'bedrooms,' 'bath_full_count,' and 'year_built' are significant, emphasizing their impact on sale prices.

Conclusion: In conclusion, the analysis of the 'df' dataset underscores the importance of multiple predictors in real estate price prediction. The second model, considering 'sq_ft_lot,' 'bedrooms,' 'bath_full_count,' and 'year_built,' provides a more accurate representation of sale prices compared to the simplistic 'sq_ft_lot' model.

Recommendations: To enhance model accuracy, further exploration of potential predictors and nonlinear relationships is advised. Additionally, ongoing validation and refinement are crucial for adapting models

to the dynamic real estate market. Advanced machine learning techniques could be explored for improved predictive capabilities.