Week 8 - 8.2 Exercise

Maxim Bilenkin

2025-01-30

1. Housing Data

a. Cleaning the data file.

```
## [1] "Original column names:"
```

```
##  [1] "Sale Date"       "Sale Price"      "sale_reason"     "sale_instrument"
##  [5] "sale_warning"    "sitetype"        "addr_full"       "zip5"
##  [9] "ctyname"         "postalctyn"
```

```
## [1] "First 5 rows of selected columns of the original dataset:"
```

```
## # A tibble: 5 x 4
##   `Sale Price` `Sale Date`         square_feet_total_living bedrooms
##          <dbl> <dttm>                                 <dbl>    <dbl>
## 1      4400000 2010-03-02 00:00:00                     5790        3
## 2      4400000 2010-03-02 00:00:00                     2410        3
## 3      4380542 2011-11-17 00:00:00                     3290        4
## 4      4380542 2011-11-17 00:00:00                     2450        4
## 5      4380542 2011-11-17 00:00:00                     2750        4
```

```
## [1] "'Sale Price' column is present in the original dataset."
```

```
## [1] "'Sale Price' column is present in the simplified dataset."
```

```
## [1] "Simplified dataset:"
```

```
## # A tibble: 5 x 5
##   `Sale Price` `Sale Date`         square_feet_total_living bedrooms
##          <dbl> <dttm>                                 <dbl>    <dbl>
## 1      4400000 2010-03-02 00:00:00                     5790        3
## 2      4400000 2010-03-02 00:00:00                     2410        3
## 3      4380542 2011-11-17 00:00:00                     3290        4
## 4      4380542 2011-11-17 00:00:00                     2450        4
## 5      4380542 2011-11-17 00:00:00                     2750        4
## # i 1 more variable: bath_full_count <dbl>
```

```
## [1] "Simplified dataset after transformation:"
```

```
## # A tibble: 5 x 5
##   `Sale Price` `Sale Date`         square_feet_total_living bedrooms
##          <dbl> <dttm>                                 <dbl>    <dbl>
## 1      4400000 2010-03-02 00:00:00                     5790        3
## 2      4400000 2010-03-02 00:00:00                     2410        3
## 3      4380542 2011-11-17 00:00:00                     3290        4
## 4      4380542 2011-11-17 00:00:00                     2450        4
## 5      4380542 2011-11-17 00:00:00                     2750        4
## # i 1 more variable: bath_full_count <dbl>
```

```
##
## Call:
## lm(formula = `Sale Price` ~ `Sale Date` + square_feet_total_living +
##     bedrooms + bath_full_count + bath_half_count + bath_3qtr_count +
##     year_built + year_renovated + sq_ft_lot, data = simplified_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2272372  -110876   -24192    63919  3729455
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -6.643e+06  4.753e+05 -13.976  < 2e-16 ***
## `Sale Date`               1.642e-04  3.199e-05   5.134 2.89e-07 ***
## square_feet_total_living  1.523e+02  5.634e+00  27.027  < 2e-16 ***
## bedrooms                 -8.563e+02  4.944e+03  -0.173   0.8625
## bath_full_count           4.310e+03  8.026e+03   0.537   0.5913
## bath_half_count           3.691e+03  7.521e+03   0.491   0.6236
## bath_3qtr_count          -1.722e+04  7.322e+03  -2.352   0.0187 *
## year_built                3.349e+03  2.388e+02  14.020  < 2e-16 ***
## year_renovated            6.890e+01  1.514e+01   4.550 5.43e-06 ***
## sq_ft_lot                 3.481e-01  6.220e-02   5.597 2.23e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 374700 on 12855 degrees of freedom
## Multiple R-squared:  0.1936, Adjusted R-squared:  0.193
## F-statistic: 342.8 on 9 and 12855 DF,  p-value: < 2.2e-16
```

b. Complete the following:

1. Explain any transformations or modifications you made to the dataset.

Answer:

I loaded the transaction file with house sale transactions from 1964 to 2016. I performed an initial data exploration to understand the data structure, column names and content. For all the missing cities for various zip codes, I manually filed all the cities in ‘ctyname’ column. Using unitedstateszipcodes.org portal, I was able to identify each city by zip code. In most cases it was Redmond.

In addition, I ensured that the ‘Sale Price’ column was formatted correctly as numeric and removing any none numeric values. I removed rows with incomplete data and converted none-numeric data to numeric. It was necessary in order to prepare the data for further statistical calculation and regression modeling.

Finally, I added extra zeros manually to all properties that had reported prices under 100K. Its impossible that big houses and in some cases fancy sold for 9K or 90K back in 2006. I researched several addresses price histories using Zillow.com and found out that they were sold at least at six digit prices. So, the best I could do is to add one extra zero at the end of each price in order to convert all the five digit prices into six. This way I removed extreme outliers making subsequent statistics outcome more precise.

2. Create a linear regression model where “sq_ft_lot” predicts Sale Price.

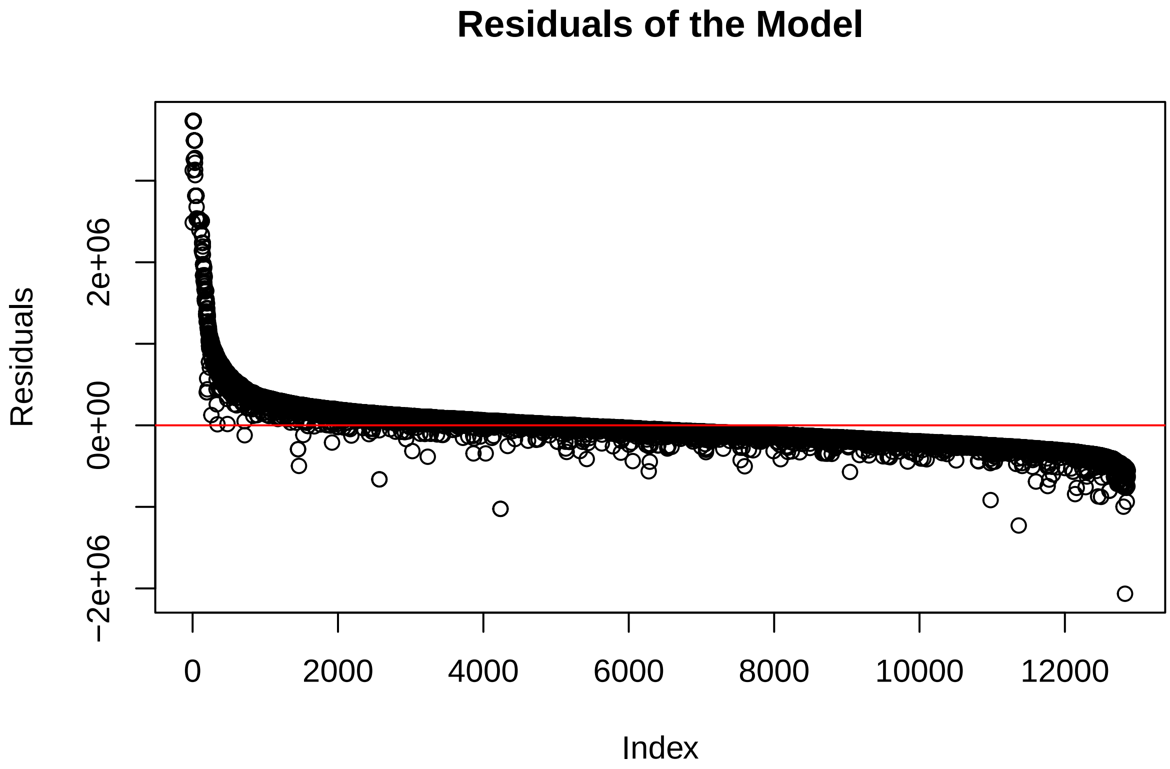
```
##
## Call:
## lm(formula = `Sale Price` ~ sq_ft_lot, data = housing_data_file)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2065450  -195362   -64778    89961  3734113
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.424e+05  3.771e+03  170.35   <2e-16 ***
## sq_ft_lot   9.582e-01  6.170e-02   15.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398400 on 12863 degrees of freedom
## Multiple R-squared:  0.01841,    Adjusted R-squared:  0.01833
## F-statistic: 241.2 on 1 and 12863 DF,  p-value: < 2.2e-16
```

3. Get a summary of your first model and explain your results (i.e., R2, adj. R2, etc.)

Answer:

Looking at the summary we can see the intercept is 6.424 or $642,400. It means when the sq_ft_lot is zero, the estimated Sale Price is 642,400. The sq_ft_lot coefficient is 9.582e-01 or 0.9582 cents. It means that for each additional square foot, the price will increase by 0.9582 cents. Although, we can see that the calculated statistics has significance since p-value < 2.2e-16. Nevertheless, in reality its has limited effect because the increase in price of square foot is very small of 0.9582 cents. By looking at Residuals, we can see that the median residual is -64778 which indicates that the model tends to underestimate the Sale Price. We can see the maximum residual is 3734113 and the minimum is -2065450. It looks like this model has some extreme values because the spread is big between the two numbers. So, we can conclude that this model tends to underestimate the house sale price because the median has negative sign. This model has high variability because the residual errors run from -2065450 to 3734113.That's a big variation in price which suggesting there is a big deviation between observed and predicted prices. The Multiple R-square of 0.01841 indicates that only 1.841% of the variance in the Sale Price is explained by the movement in sq_ft_lot(square foot per lot). This is not a big predictor. We can conclude that there are probably other better variables that can have more prediction power. The adjusted R-square of 0.01833 or 1.833% is slightly lower which only reinforcing our claim that the model's predictive power is limited. The Residual standard error of 398400 tells us that the predictive house price deviates by 398,400 from the actual price. This is a big difference. Looks like there are other variables that were not present in this file and that have predictive power. The F-statistic of 241.2 and p-value: < 2.2e-16, indicating that the relationship between Sale Price and sq_ft_lot(square foot per lot) is statistically significant. But because we have low R-squared value of 0.01841(1.841%) we can conclude that in reality the significance doesn't have much predictive power.

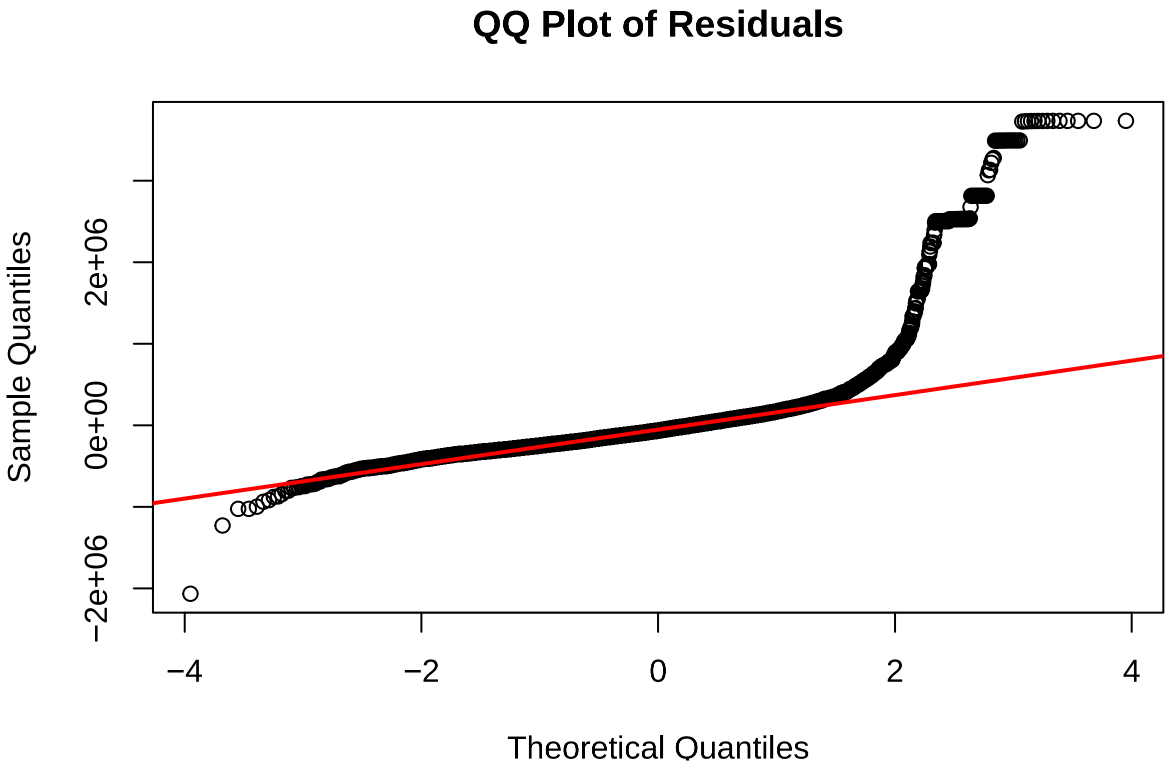
4. Get the residuals of your model (you can use 'resid' or 'residuals' functions) and plot them. What does the plot tell you about your predictions?



Answer:

By looking at the plot I can see there are many big residuals that deviate a lot from the zero line. This model doesn't seem to be a good predictor. Looks like there are many house Sale Prices that are outliers that make the model not so well predictable. We can see how the Residuals hugely deviate at index 0. Also, starting at approximately index 10000 through 12000 and beyond we can see how the residuals deviate considerably from the red line. Additionally, there are quite a few large outliers throughout the plot on the negative side. The plot indicates that my predictions are not accurate. There are likely other factors with greater predictive power on Sale Price that are not included in this dataset.

5. Use a qq plot to observe your residuals. Do your residuals meet the normality assumption?



Answer:

By observing the QQ Plot, we can conclude that my model residuals do not follow the normality assumption. The plot shows deviations from the red reference line on both sides. On the negative tail, the residuals deviate somewhat significantly, while on the positive tail(right tail), they deviate extremely from the red line. Thus, we conclude that the residuals in my model are not normally distributed.

6. Now, create a linear regression model that uses multiple predictor variables to predict Sale Price (feel free to derive new predictors from existing ones). Explain why you think each of these variables may add explanatory value to the model.

```
##
## Call:
## lm(formula = `Sale Price` ~ `Sale Date` + square_feet_total_living +
##     bedrooms + bath_full_count + bath_half_count + bath_3qtr_count +
##     year_built + year_renovated + sq_ft_lot, data = housing_data_file)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1474751  -363540  -143779    53391  3842465
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.150e+07  1.146e+07   1.003 0.315772
## `Sale Date`              -1.666e+04  5.534e+03  -3.010 0.002639 **
## square_feet_total_living  1.068e+02  2.178e+01   4.906 9.94e-07 ***
## bedrooms                  8.805e+04  2.271e+04   3.878 0.000108 ***
## bath_full_count          -5.415e+04  3.660e+04  -1.479 0.139167
## bath_half_count          -5.556e+04  3.214e+04  -1.729 0.084009 .
## bath_3qtr_count          -1.270e+05  3.310e+04  -3.835 0.000129 ***
## year_built                1.120e+04  9.752e+02  11.489  < 2e-16 ***
## year_renovated           1.340e+02  5.383e+01   2.488 0.012906 *
## sq_ft_lot                3.910e-01  2.063e-01   1.895 0.058166 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 735500 on 2287 degrees of freedom
## Multiple R-squared:  0.1282, Adjusted R-squared:  0.1247
## F-statistic: 37.36 on 9 and 2287 DF,  p-value: < 2.2e-16
```

Answer:

By selecting the 'Sale Date' variable I could see the changes of the home prices over time. The 'square_feet_total_living' can influence a home price. As we know, larger living areas usually derive higher prices. Similarly goes for the 'bedrooms' variable. The more bedrooms, the higher value of a property. The more full baths a property has the more significantly it will increase the house price. Thus, the 'bath_full_count' variable probably a good house price predictor. The 'year_built' variable is good idea to include because the newer houses in most cases sell for higher prices because they have newer features and newly materials. Similarly for the 'year_renovated' variable. We are all know that newly renovated property commands a higher value. The 'sq_ft_lot' variable also good idea to include. Since, the size of the lot can impact the overall value of the property.

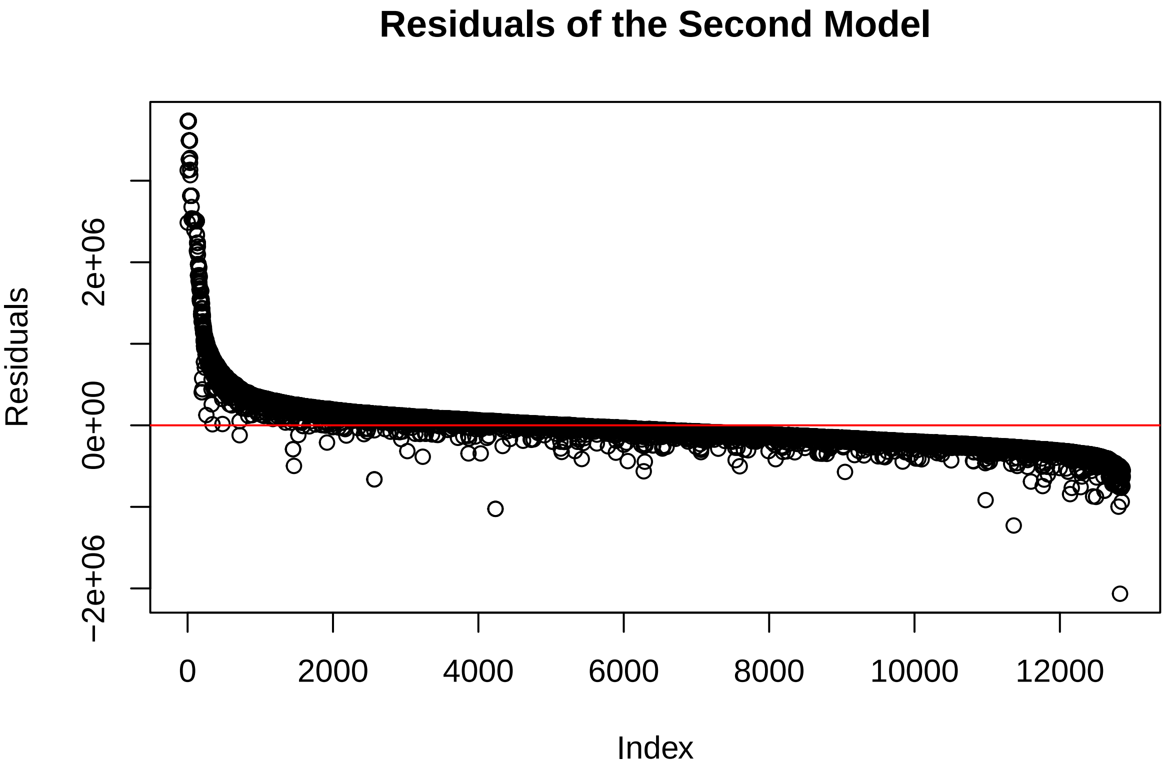
7. Get a summary of your next model and explain your results.

Answer:

From the summary of the next model we can see Intercept is 1.150e+07(or 11,500,000). This is the base Sale Price when all other predictable variables are set to zero. Sale Date -1.666e+04(or -16,660). This means that for each additional year the sale price decreases by -16,660. The calculated value of 1.068e+02(or 106.8) for 'square_feet_total_living' is large and positive which indicates that the larger living properties command higher values or to put it another way, for each additional square foot the sale price increases by $106.8. The calculated 8.805e+04 (or 88,050) for 'bedrooms' is positive and large which indicates that more bedrooms tend to increase the property sale prices by 88,050 dollars. Since 'bath_full_count' of -5.415e+04 (-54,150) is negative and not significant at p-value = 0.139167 then, we can conclude that it doesn't have much effect on property sale price. Similar goes for 'bath_half_count'. The calculated value of -5.556e+04 (-55,560) is marginally higher and negative which means it doesn't have much impact on sale price. The value of -1.270e+05 (-127,000) for 'bath_3qtr_count' is negative and significant at p-value of 0.000129. This variable seems to have meaningful predictability on sale prices. It does make sense. The bath 3 quarter count is the one that have shower or tub, sink, toilet, but no bathtub. In today's world probably most people would prefer to have bathtub. We can have considerable assumption that more bath_3qtr_count in a property would probably decrease value of a property.The 'year_built' is significant and positive at1.120e+04 (11,200) and p < 2e-16. This means that the year build has a significant impact on home prices. The same for 'year_renovated' with 1.340e+02 (134) and p < 2e-16 have positive and significant relationship on sale price. It does make sense. We know that newly renovated properties tend to sell for higher prices. Finally, 'sq_ft_lot' also positive and significant with coefficient of 3.910e-01 (0.391) indicating that for each additional foot the sale price will increase approximately by 0.391.

Residuals indicate large range which means there is a substantial variability between predicted and actual property sale prices. Both, Multiple and Adjusted R-squares are almost equal of 0.1282 and 0.1247 or 12%, indicating that approximately 12% of the variation in the Sale Price is explained by predictors. The residual standard error is 735500. This is the mostly the standard deviation between the actual and predicted sale price. With the F-statistic of 37.36 and p-value: < 2.2e-16, we can conclude that the overall model is statistically significant.

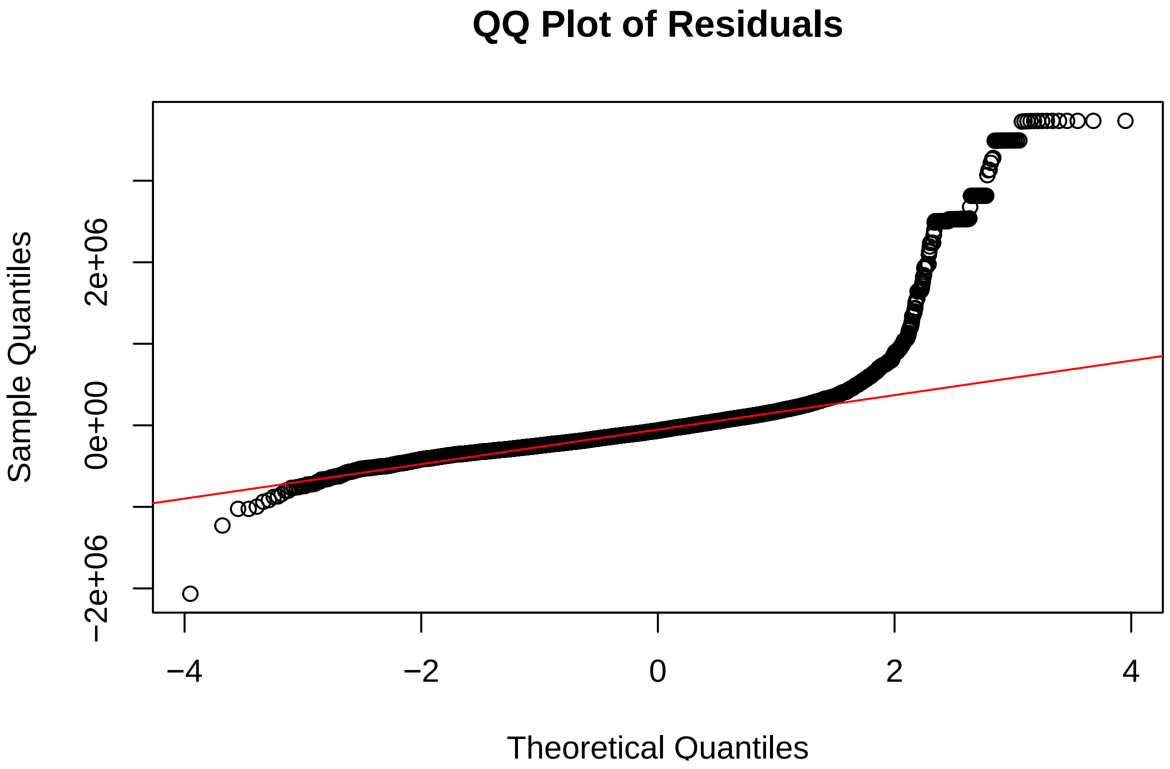
8. Get the residuals of your second model (you can use 'resid' or 'residuals' functions) and plot them. What does the plot tell you about your predictions?



Answer:

The “Residuals of the Second Model” does much better job than original model. Although, there are still extreme outliers at the begging index 0. However, the outliers closer now to the red line. Throughout the whole red line residuals are closer to the red zero line and on the right side (the positive tail) the majority residuals are attached to the red line as oppose to the original model where residuals detached from the red line on the negative side. We can conclude that the second model better predicts the sale price than the original model because the residuals are closer to the red line.

9. Use a qq plot to observe your residuals. Do your residuals meet the normality assumption?



Answer:

No, the residuals don't meet the normality distribution because on the left tail(negative side) and right tail(positive side) the residuals substantially deviate from the red line. Especially on the right side the residuals deviate extremely.

10. Compare the results (i.e., R2, adj R2, etc) between your first and second model. Does your new model show an improvement over the first? To confirm a 'significant' improvement between the second and first model, use ANOVA to compare them. What are the results?

```
## Analysis of Variance Table
##
## Model 1: `Sale Price` ~ sq_ft_lot
## Model 2: `Sale Price` ~ `Sale Date` + square_feet_total_living + bedrooms +
##     bath_full_count + bath_half_count + bath_3qtr_count + year_built +
##     year_renovated + sq_ft_lot
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  12863 2.0416e+15
## 2  12855 1.6067e+15  8 4.3494e+14 434.98 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

By comparing the two residual models we can see that the second model is better than the first one. For the first model, The Multiple R-squared(0.01841), Adjusted R-squared(0.01833) and Residual Standard Error(398400).

For the second model, Multiple R-squared(0.1282), Adjusted R-squared(0.1247) and Residual Standard Error(735500). The second model has both Multiple and Adjusted R-Squared higher than for the first model. This means the second model indicates more variability in the sale price. The reason second model has higher Residual Standard Error might be because it has more predictors.

Finally, we can confirm the significant improvement by using the ANOVA table. Looking at the above "Analysis of Variance Table", we can see that the P-value is extremely small, which means that the improvement from model 1 to model 2 is statistically significant. This means that additional inclusion of predictors improved model 2 ability to predict better sale price. Model 2 has 12855 degrees of freedom vs 12863 for model 1. This is due to the additional predictors in model 2. The F-Statistics of 434.98 is high and indicates that the reduction in RSS from model 1 to model 2 is significant. We can conclude that model 2 has better predictive power than model 1.

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

Answer:

Overall the model 2 had improved significantly in predicting sale price. By including more predictors we reduced bias between the actual and predictive sale price. However, I believe both models are still biased because both have many outliers and many are extreme once. They make both models biased. Even after improving residual normality for model 2, they are still many extreme outliers of residuals that make very hard accurately predict sale prices. Definitely, further investigation on the outliers needed. At the same time, it is hard to identify what were the actual true sale prices without having the actual records that could confirm the data. A lots of houses in the dataset had sale price below 100K which is hard to believe. I searched about ten addresses and saw nice and considerably big houses that don't look like they would trade under 100K. Without clear guidance that would give some clear price range on the lower and higher end so anything below and above could be deleted it is hard to clean the data.

12. Another important aspect of regression tasks is determining the accuracy of your predictions. For this section, we will look at root mean square error (RMSE), a common accuracy metric for regression models.

13. Install the 'Metrics' package in R Studio

```
## Installing package into 'C:/Users/maxim/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)
```

```
## package 'Metrics' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\maxim\AppData\Local\Temp\Rtmpk9ZUu5\downloaded_packages
```

2. Using the first model, we will make predictions on the dataset using the predict function. An example would look like this (will vary for you based on variable names):

   1. `‘preds <- predict(object = modelName, newdata = dataset)’`
   2. `Use the ‘rmse’ function to get RMSE for the model (‘rmse(actual, predicted)’)`

3. What is the RMSE for the first model?

```
## [1] "RMSE for the first model: 398368.267743496"
```

Answer:

The RMSE(Root Mean Square Error) for the first model is 398,368.27. This is the average error or deviation between predicted sale price and the actual price in dollar terms for the first model. The lower the RMSE, the better.

4. Perform the same task for the second model. Provide the RMSE for the second model.

```
## [1] "RMSE for the second model: 353397.289317687"
```

Answer:

The RMSE(Root Mean Square Error) for the second model is 353,397.29. This is the average error or deviation of predicted sale price from the actual price.

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

As stated before, lower RMSE is better. Since, the second model has lower RMSE of 353,397.29 vs 398,368.27 for the first model. We can conclude that second model has better predictive power. Including more predictors in the second model significantly improved predictive power by $44,970.98 (398,368.27 - 353,397.29).