Cordero_week8.2

Joaquin Cordero

2024-07-28

Housing Date

```
housing_data <- read_xlsx("week-6-housing.xlsx") %>%
  rename(sale_price = 'Sale Price')

my_data <- housing_data %>%
  select(sale_price, addr_full, sq_ft_lot) %>%
  distinct(addr_full, .keep_all = TRUE)
```

1.

I selected only 3 columns from the original data that is currently needed. Created a new data set using Sale Price, addr_full, and sq_ft_lot from the original. To remove duplicates, I used addr_full to remove any duplicated rows because a unique entry should not have the same address. I used duplicated() function prior to and after using distinct() function to check for any duplicate entries. Lastly, renamed column 'Sale Price' to sale_price.

2.

```
sq_ft_price_lm <- lm(sale_price ~ sq_ft_lot, data = my_data)
sq_ft_price_lm
```

```
##
## Call:
## lm(formula = sale_price ~ sq_ft_lot, data = my_data)
##
## Coefficients:
## (Intercept)    sq_ft_lot
##   6.464e+05    8.261e-01
```

3.

```
summary(sq_ft_price_lm)
```

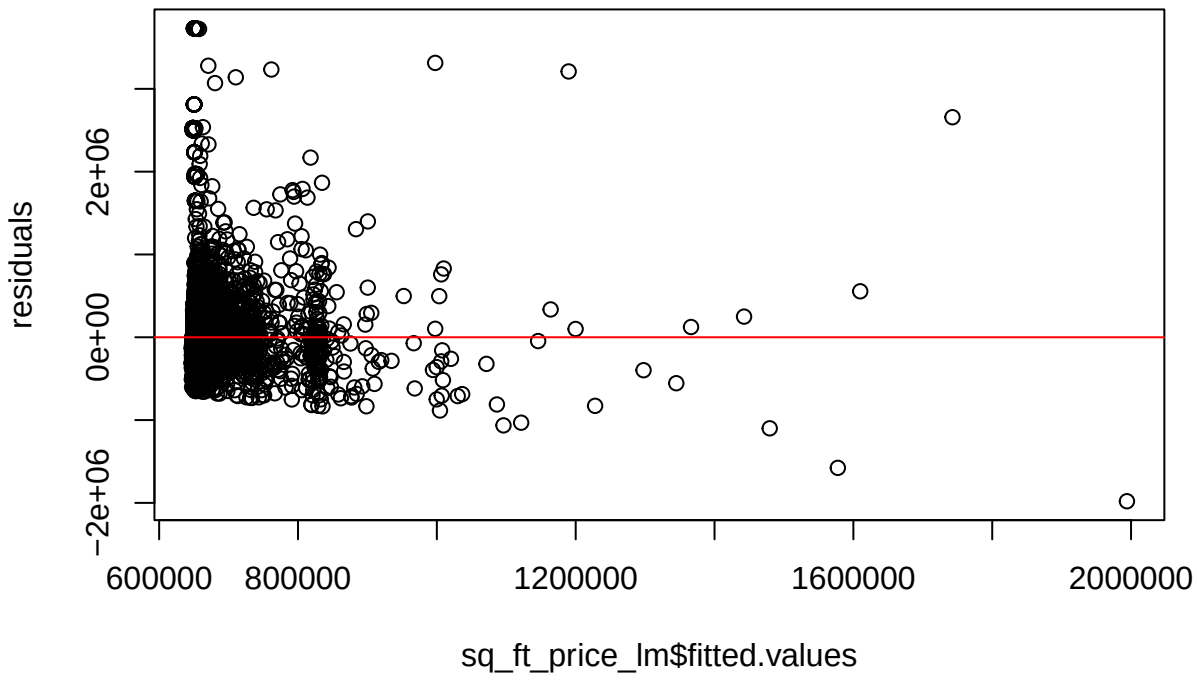
```
##
## Call:
## lm(formula = sale_price ~ sq_ft_lot, data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1980092  -192506   -69416    86073  3730602
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.464e+05  4.453e+03  145.16   <2e-16 ***
## sq_ft_lot   8.261e-01  7.500e-02   11.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 408900 on 9735 degrees of freedom
## Multiple R-squared:  0.01231,    Adjusted R-squared:  0.01221
## F-statistic: 121.3 on 1 and 9735 DF,  p-value: < 2.2e-16
```

For every additional square foot, price is expected to increase by 0.826.

4.

```
residuals <- resid(sq_ft_price_lm)

plot(sq_ft_price_lm$fitted.values, residuals) + abline(h=0, col = "red")
```

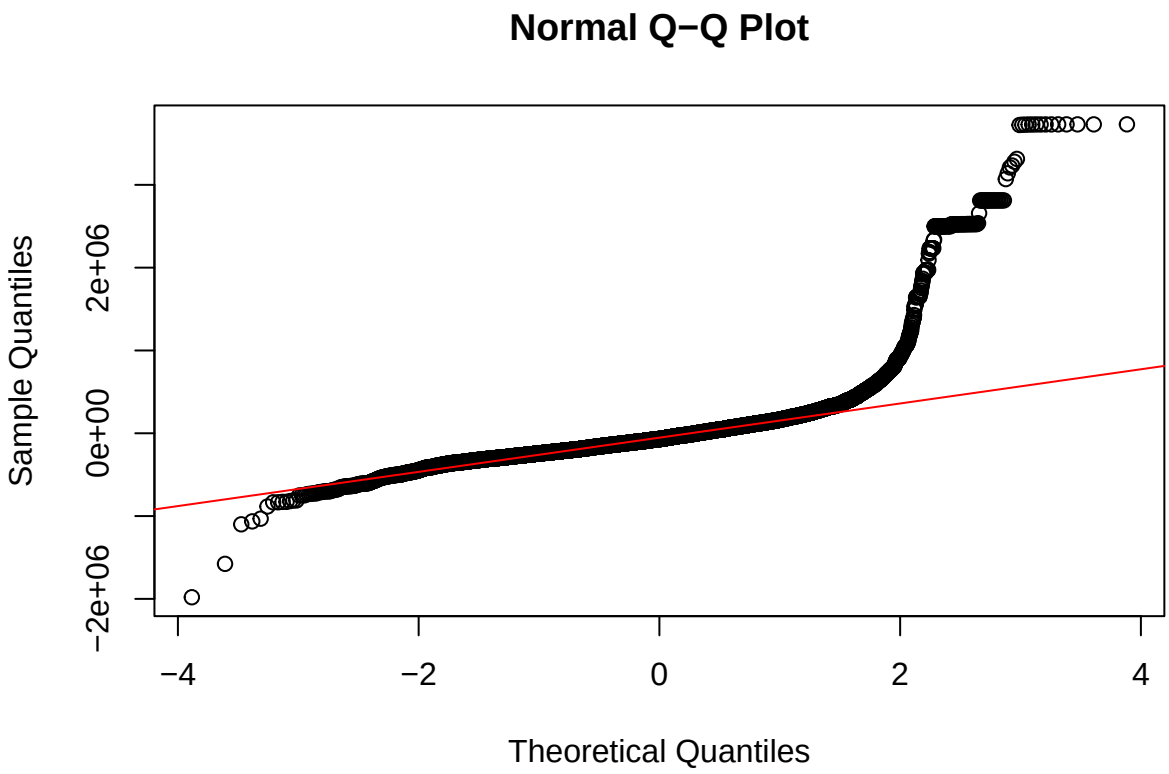



```
## integer(0)
```

The model indicates a cluster and not evenly distributed on the horizontal line at 0. May not be representing the relationship properly.

5.

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



The residuals do not meet the normality assumption since it forms an S-shaped curve.

6.

```
second_model <- housing_data %>%
  select(sale_price, addr_full, square_feet_total_living, bedrooms,
         bath_full_count, year_built) %>%
  distinct(addr_full, .keep_all = TRUE)

second_model_lm <- lm(sale_price ~ square_feet_total_living + bedrooms +
                        bath_full_count + year_built, data = second_model)
second_model_lm
```

```
##
## Call:
## lm(formula = sale_price ~ square_feet_total_living + bedrooms +
##     bath_full_count + year_built, data = second_model)
##
## Coefficients:
##              (Intercept)  square_feet_total_living                  bedrooms
##                -6081490.8                     167.6                  -16151.2
##          bath_full_count                year_built
##                   3881.8                    3196.2
```

Picking these variables may add explanatory value to the model because I believe these are variables buyers would be interested in. Having these variables might increase sale price.

7.

```
summary(second_model_lm)
```

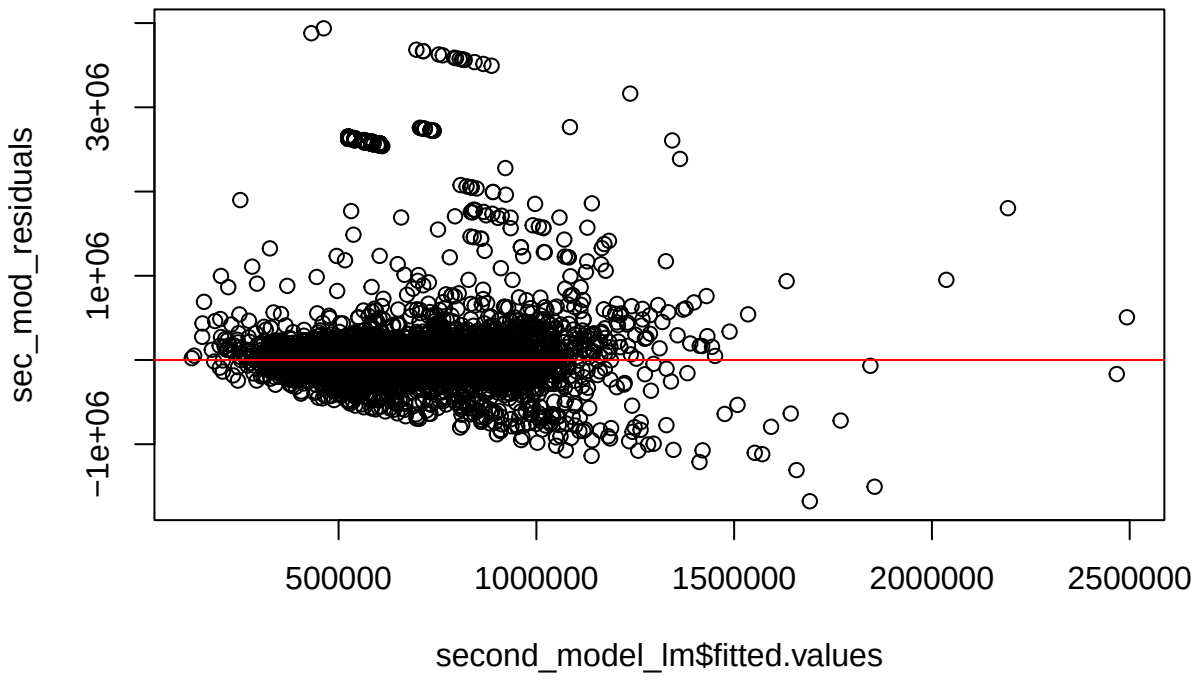
```
##
## Call:
## lm(formula = sale_price ~ square_feet_total_living + bedrooms +
##     bath_full_count + year_built, data = second_model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1677298  -122742   -47751    39000  3937530
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -6.081e+06  5.011e+05 -12.136  < 2e-16 ***
## square_feet_total_living  1.676e+02  5.380e+00  31.144  < 2e-16 ***
## bedrooms                 -1.615e+04  5.367e+03  -3.009  0.00263 **
## bath_full_count           3.882e+03  7.194e+03   0.540  0.58948
## year_built                3.196e+03  2.532e+02  12.625  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 369300 on 9732 degrees of freedom
## Multiple R-squared:  0.1945, Adjusted R-squared:  0.1942
## F-statistic: 587.6 on 4 and 9732 DF,  p-value: < 2.2e-16
```

Variables square_feet_total_living and year_built are highly significant predictors. While variable bath_full_count is a low significant predictor. For each additional square foot of living sale price is expected to increase by $167.6. For each additional bedroom sale price is actually expected to decrease by -$16,150. For each additional year price is expected to increase by $3,196 meaning newer homes are priced higher.

8.

```
sec_mod_residuals <- resid(second_model_lm)

plot(second_model_lm$fitted.values, sec_mod_residuals) + abline(h=0, col = "red")
```

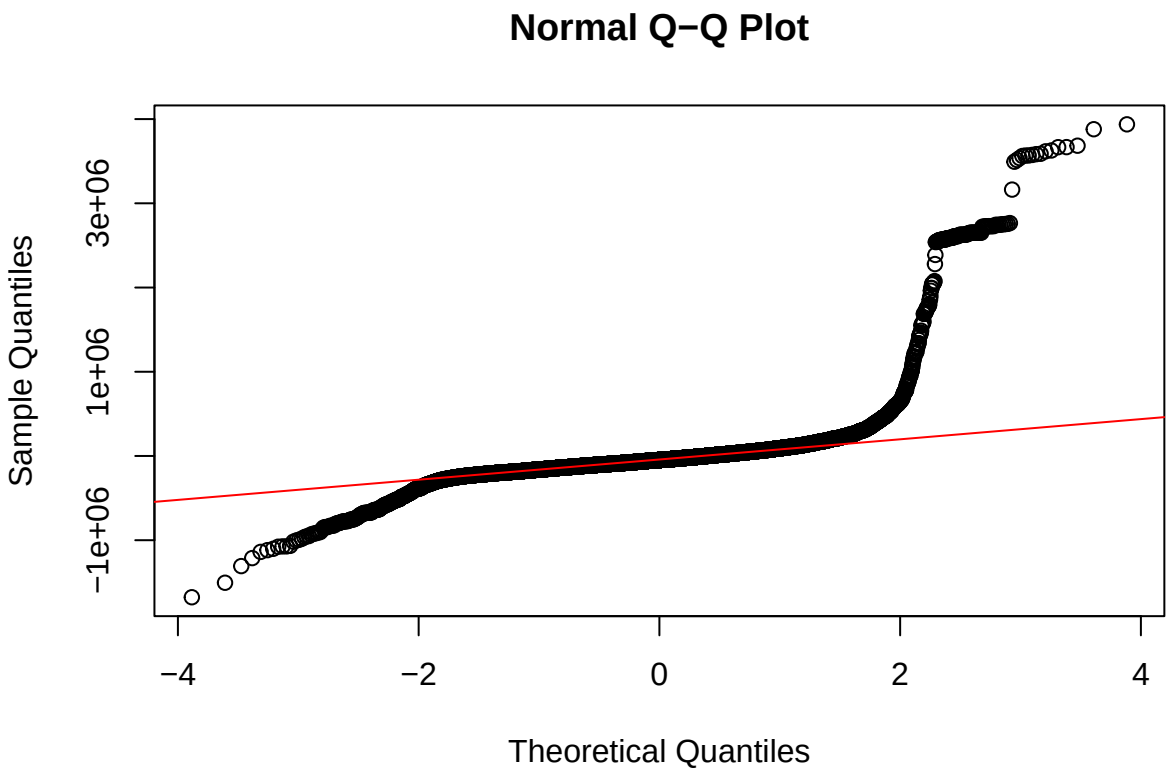



```
## integer(0)
```

The second model is clustered and not evenly distributed on the horizontal line at 0. Again, this may not be representing the relationship properly.

9.

```
qqnorm(sec_mod_residuals)
qqline(sec_mod_residuals, col = 'red')
```



My residuals do not meet the normality assumption since it forms an S-shaped curve as well.

10.

```
anova(sq_ft_price_lm, second_model_lm)
```

```
## 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   9735 1.6277e+15
## 2   9732 1.3274e+15  3 3.003e+14 733.91 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, there is significant improvements between the two models. The p-value shows that the additional variables are important to the model.

11.

The model is biased since it shows a clear pattern that does not meet normal distribution.

12.2.1

```
preds_model1 <- predict(object = sq_ft_price_lm, newdata = housing_data)
```

12.2.2

```
rmse_model1 <- rmse(housing_data$sale_price, preds_model1)
```

12.3

```
rmse_model1
```

```
## [1] 401475.6
```

12.4

```
preds_model2 <- predict(object = second_model_lm, newdata = housing_data)

rmse_model2 <- rmse(housing_data$sale_price, preds_model2)

rmse_model2
```

```
## [1] 357641.1
```

12.5

```
rmse_model1 - rmse_model2
```

```
## [1] 43834.47
```

Yes the second model's RMSE improved from the first model's RMSE. It improved by 43834.47.