

Kansas City Housing Data Report

Travis & Ryan

January 8, 2020


Abstract

Our research objective was to find the optimal multiple linear regression model for our dataset, which discovered 2015 prices for houses in Kansas City, Missouri. Our methods included a first-order linear regression model, then utilizing a centered predictor model with interactions for all the predictor variables. To optimize this updated centered model, we put it through a stepwise regression to find the optimal model to help predict the housing price based on the predictors in the data set. Our final model concluded by using each predictor from the first order model plus a few interactions between them which helped optimize the model (per the AIC stepwise regression).

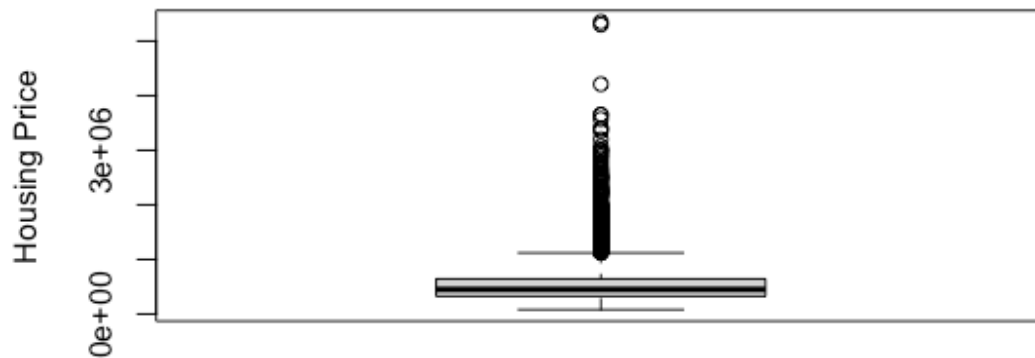
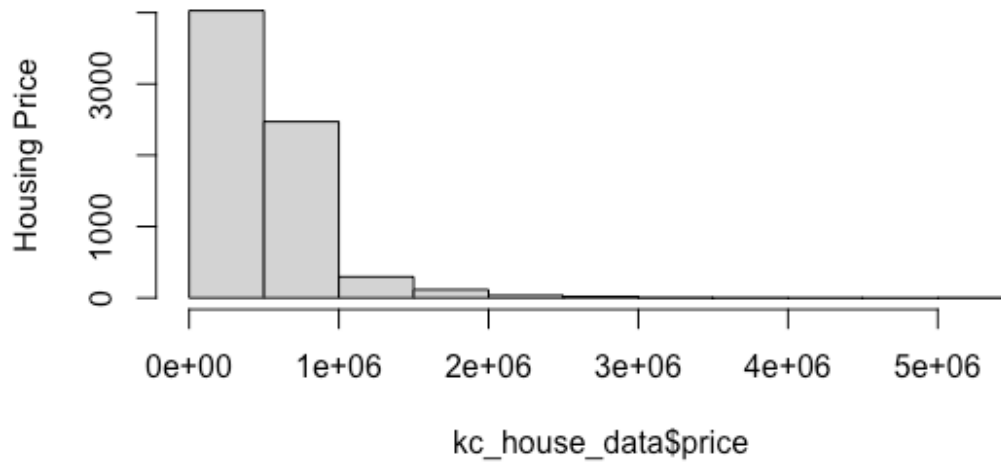
Introduction

This report examines data on housing prices and their relative features within the Kansas City metropolitan area. All data was taken from the year of 2015. The overall goal is to determine which housing features are most effective in predicting the price of the house. The predictor variables include number of floors, square feet of the living room, square feet of the land, year built, if it is waterfront property, number of bedrooms, and number of bathrooms. There are 6,975 observations in the data set.

Exploratory Analysis

```
##  
## — Column specification —————  
## cols(  
##   .default = col_double(),  
##   date = col_character()  
## )  
##  Use `spec()` for the full column specifications.
```

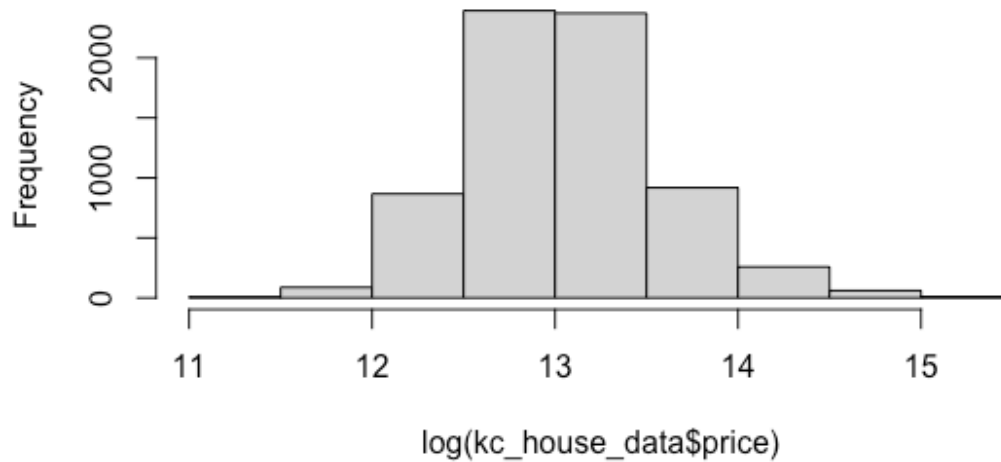
Histogram of kc_house_data\$price



Since price is our main focus, and will be the response variable, this histogram gives us a quick look at the variable's distribution. Through both the histogram and the box plot, we can see that this response variable seems to be right skewed, and may require a transformation. Especially when analyzing the box plot, there are lots of outliers since price seems to have a large range of values. To account for this in the report, we may look to put these values in logarithmic form.

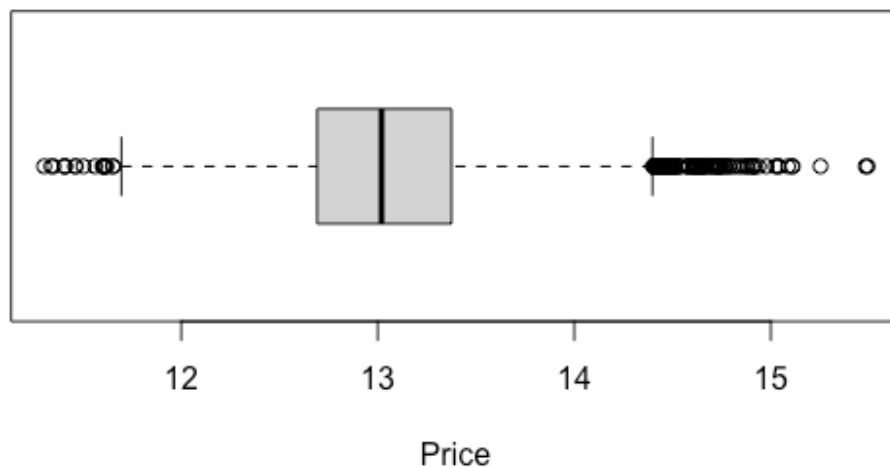
```
hist(log(kc_house_data$price))
```

Histogram of log(kc_house_data\$price)



```
boxplot (log (kc_house_data$price), horizontal = T, xlab="Price", main="Boxplot of Log(price)")
```

Boxplot of Log(price)

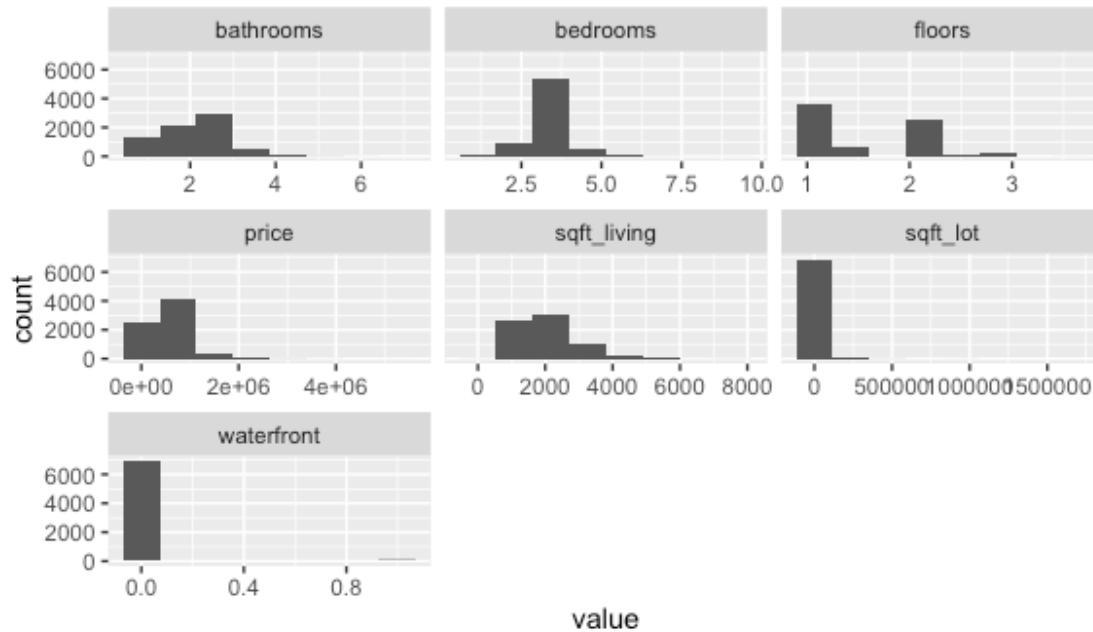


The distribution of log(price) is way more symmetric as seen through both the updated box plot and histogram. Logarithmic form is likely to be form of the response variable in this report.

Distributions of the other predictor variables:

```
library (ggplot2)  
library (tidyr)
```

```
ggplot(gather(kc_house_data[, 3:9]), aes(value)) +
  geom_histogram(bins = 8) +
  facet_wrap(~key, scales = 'free_x')
```

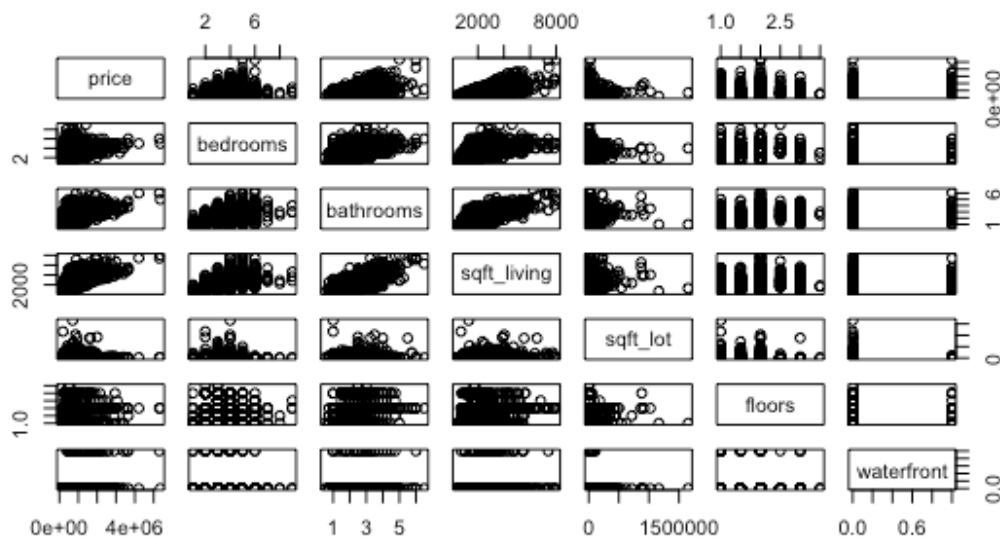


None of the predictor variables are extremely right or left skewed. Excluding the price (already discussed response variable above), the variables of slight concern seem to be square feet of the land, waterfront, and floors. One thing to note, is waterfront is not much of a shock. This variable is a categorical variable, and reads as 0 or 1 only. Therefore, the lack of normal distribution is not of a concern to me. To help scale the data for the rest of the variables of interest, we may look to consider a square root or log transformation for either square feet of the land or floors.

Next, we will analyze simple linear associations between each pair of variables

check for simple linear associations between each pair of variables

```
pairs (kc_house_data[,3:9])
```

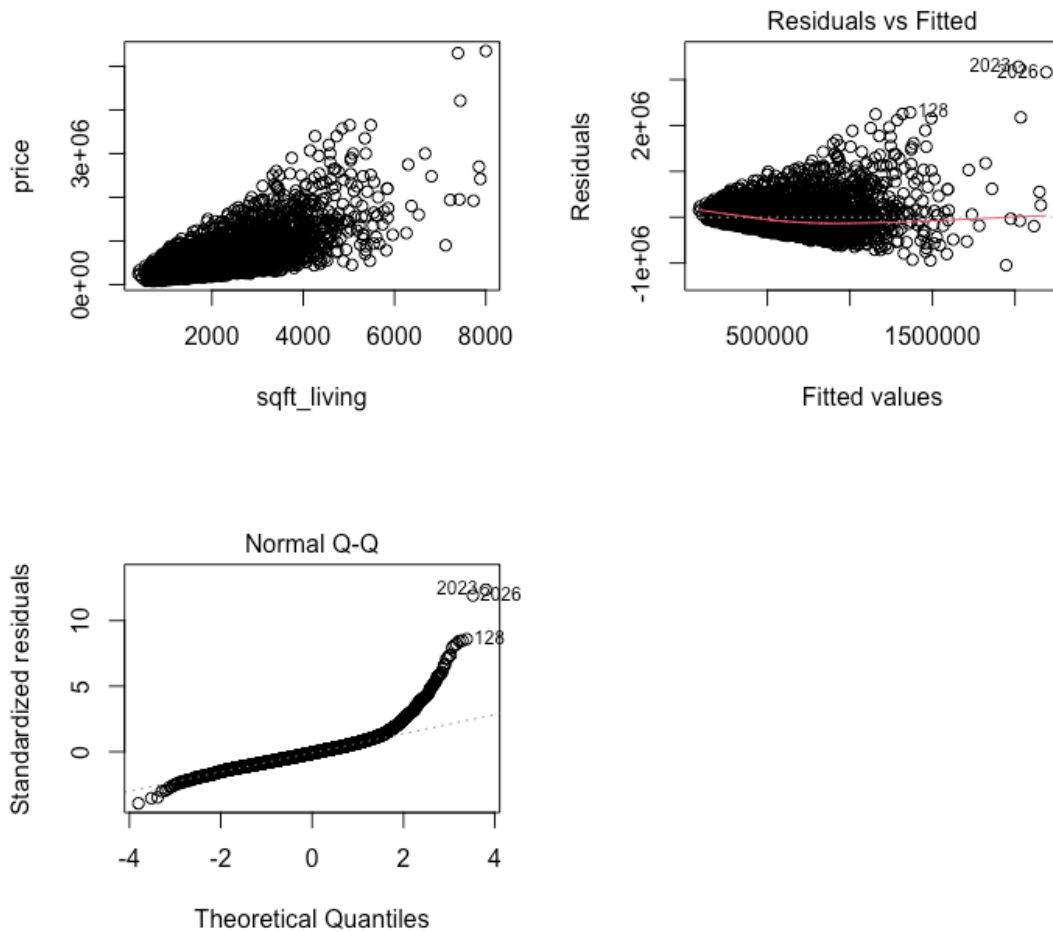


```
cormat = cor(kc_house_data[,3:9], use = "complete.obs")
round(cormat, 2)
```

```
##           price bedrooms bathrooms sqft_living sqft_lot floors waterfront
t
## price      1.00    0.31    0.51    0.68    0.10    0.25    0.25
5
## bedrooms   0.31    1.00    0.52    0.60    0.03    0.18   -0.01
1
## bathrooms  0.51    0.52    1.00    0.76    0.10    0.50    0.00
6
## sqft_living 0.68    0.60    0.76    1.00    0.17    0.35    0.00
9
## sqft_lot   0.10    0.03    0.10    0.17    1.00   -0.01    0.00
1
## floors     0.25    0.18    0.50    0.35   -0.01    1.00    0.00
4
## waterfront 0.25   -0.01    0.06    0.09    0.01    0.04    1.00
0
```

The pairs plot appears to show a linear relationship between price and bathrooms. It also appears that there is a linear relationship between price and bedrooms, as well as a linear relationship between price and sqft_living. There appears to be a very strong linear relationship between bathrooms and sqft_living, this should be further looked into for problems with collinearity.

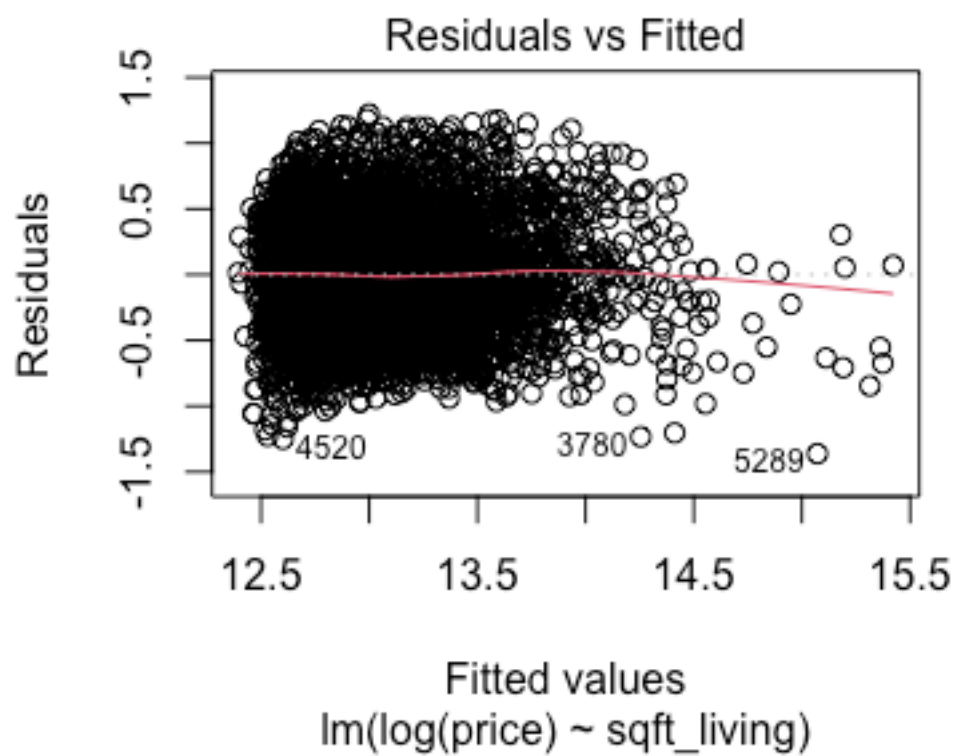
```
par(mfrow=c(1,2))
plot(price ~ sqft_living, data=kc_house_data)
fit = lm(price ~ sqft_living, data=kc_house_data)
plot(fit,which=1:2)
```

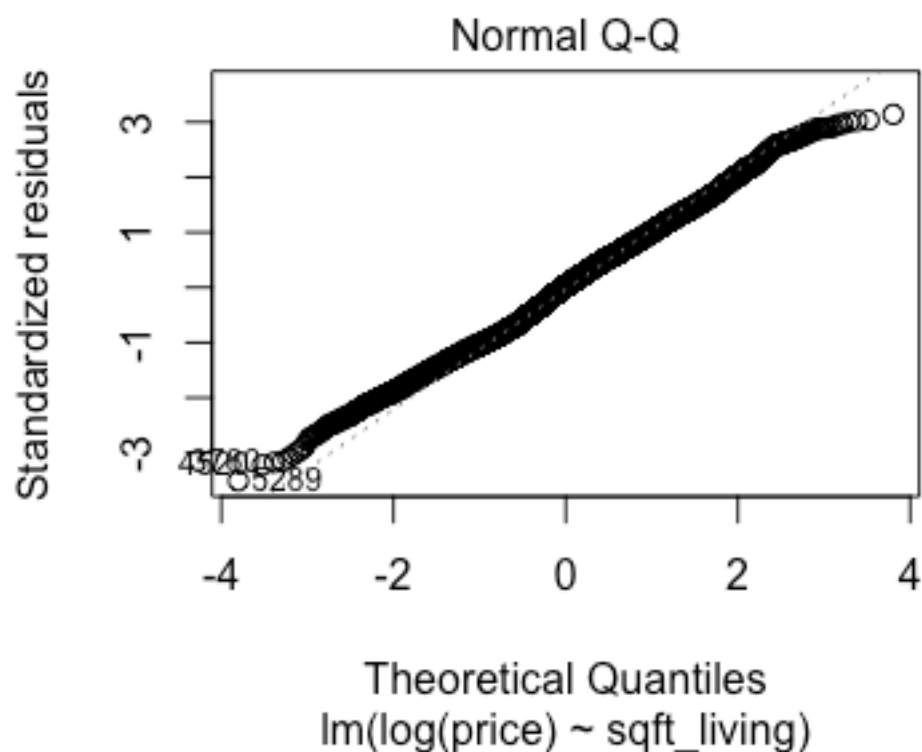


The plots for the simple regression between these express non-linearity and obvious curvature. As expressed above, there may be a need for a logarithmic form on price and `sqft_living` to fix this.

This will be shown below.

```
fit = lm(log(price) ~ sqft_living, data=kc_house_data)
plot(fit, which=1:2)
```





These plots above show linearity and constant variance within this simple linear model. Through this conclusion, the rest of the report will continue to analyze data when the price is in logarithmic form.

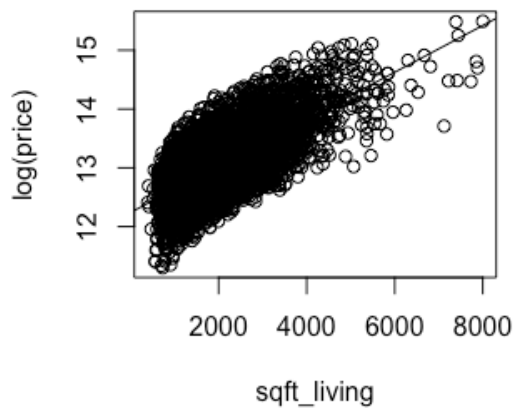
We will start with a simple linear regression log(price) vs sqft_living

```
par (mfrow = c(1,2))
plot (log(price) ~ sqft_living, data=kc_house_data)
fit0 = lm (log(price) ~ sqft_living, data=kc_house_data)
abline(fit0)
summary(fit0)

##
## Call:
## lm(formula = log(price) ~ sqft_living, data = kc_house_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.36150 -0.29994  0.01042  0.27017  1.22137
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.224e+01  1.162e-02 1053.04  <2e-16 ***
```

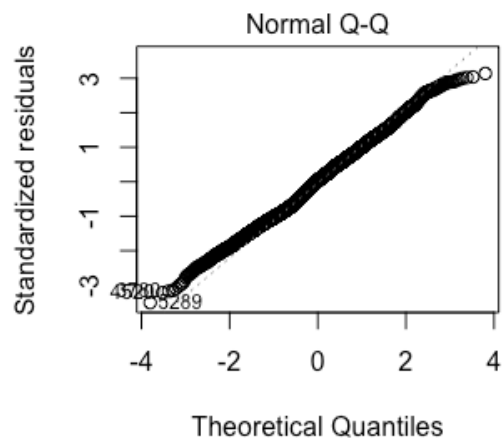
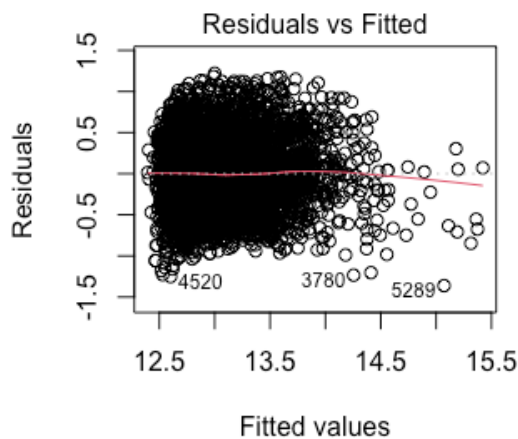


```
## sqft_living 3.977e-04  5.216e-06  76.25  <2e-16  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3891 on 6973 degrees of freedom
## Multiple R-squared:  0.4547, Adjusted R-squared:  0.4546
## F-statistic: 5815 on 1 and 6973 DF, p-value: < 2.2e-16
```



The log(price) increases 0.0003977 log(dollars) per square feet of living space. This is statistically significant with a p-value $> .000001$. The residual standard error is 0.3891 log(price). The square feet of living space explains 45.47% of the variation of log(price).

```
par (mfrow = c(1,2))
plot (fit0, which=1:2)
```



The residuals for the Log(Price) vs sqft_living model are consistent with linearity, constant variance, and a normal distribution.

First Order Model

fit a first-order linear model with all seven variables

```
first_fit = lm(log(price) ~ sqft_living + bathrooms + bedrooms + log(sqft_lot) + floors + yr_built + waterfront, data=kc_house_data)
summary(first_fit)

##
## Call:
## lm(formula = log(price) ~ sqft_living + bathrooms + bedrooms + log(sqft_lot) + floors + yr_built + waterfront, data = kc_house_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54354 -0.25763  0.01613  0.25277  1.28461
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.204e+01  3.632e-01  60.686  <2e-16 ***
## sqft_living  4.106e-04  8.744e-06  46.956  <2e-16 ***
## bathrooms    1.101e-01  1.026e-02  10.730  <2e-16 ***
## bedrooms    -6.785e-02  6.147e-03 -11.039  <2e-16 ***
## log(sqft_lot) -4.785e-02  5.756e-03  -8.313  <2e-16 ***
## floors       1.014e-01  1.070e-02   9.477  <2e-16 ***
## yr_built    -4.846e-03  1.890e-04 -25.645  <2e-16 ***
## waterfront   5.740e-01  5.253e-02  10.928  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3621 on 6967 degrees of freedom
## Multiple R-squared:  0.5283, Adjusted R-squared:  0.5278
## F-statistic: 1115 on 7 and 6967 DF, p-value: < 2.2e-16
```

Through analyzing the summary statistics of the first-order model, it seems that all slope parameters have high t-values, with statistically significant p-values at the 5% level. (Note: we will be using 5% significance level for our p-values within this report.) There also seems to be an adequate R-Squared (52,83%) and adjusted R-Squared (52.78%). This tells us that over half of the variability in $\log(\text{price})$ is being explained by this model. The residual standard error is 0.3621 $\log(\text{price})$. Our first-order model looks to be off to a good start by having all slope parameters statistically significant, and our model expressing a decent goodness-of-fit.

Next, we will look at the ANOVA table.

```
anova(first_fit)

## Analysis of Variance Table
##
## Response: log(price)
```

```
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## sqft_living    1  880.39   880.39 6715.649 < 2.2e-16 ***
## bathrooms     1    2.35    2.35  17.908 2.348e-05 ***
## bedrooms      1   12.89   12.89   98.305 < 2.2e-16 ***
## log(sqft_lot)  1   17.73   17.73  135.275 < 2.2e-16 ***
## floors        1    0.23    0.23    1.782  0.1819
## yr_built      1   93.57   93.57  713.742 < 2.2e-16 ***
## waterfront    1   15.66   15.66  119.425 < 2.2e-16 ***
## Residuals    6967 913.34    0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The analysis of variance table suggests that all but floors are statistically significant predictors at the 5% level. This explains that we may need to look to remove floors as a predictor in a separate model.

```
library (car)

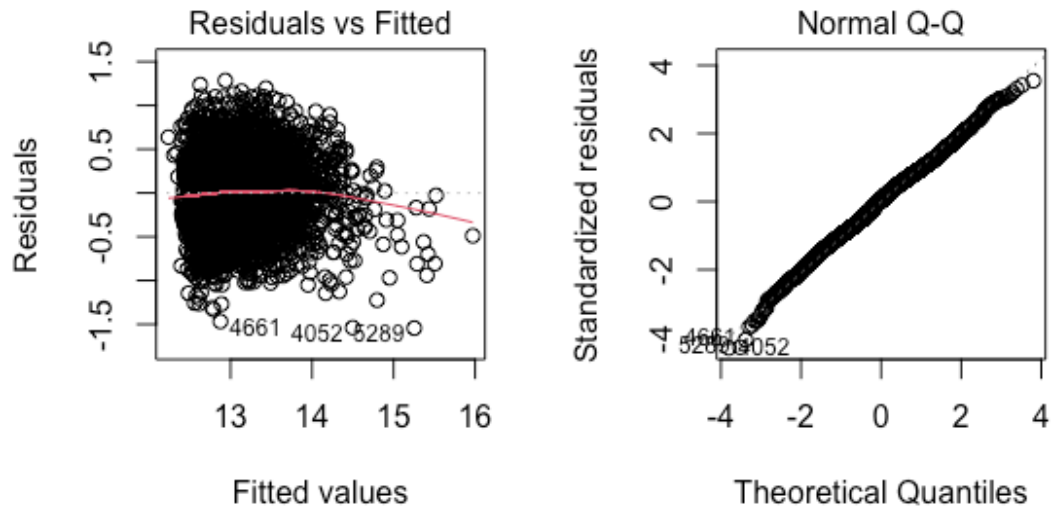
## Loading required package: carData

vif (first_fit)

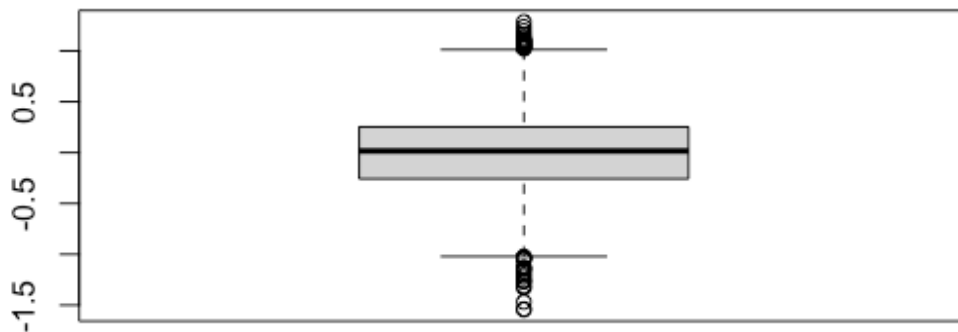
##   sqft_living    bathrooms    bedrooms log(sqft_lot)    floors
##   3.246064      3.278020      1.625683      1.420264      1.763757
##   yr_built      waterfront
##   1.611774      1.024123
```

None of the adjusted VIF's are higher than the standard cut-off of 5. All of the VIF's are less than 5 and are considered adequate.

```
par(mfrow = c(1, 2))
plot (first_fit, which = c(1, 2))
```



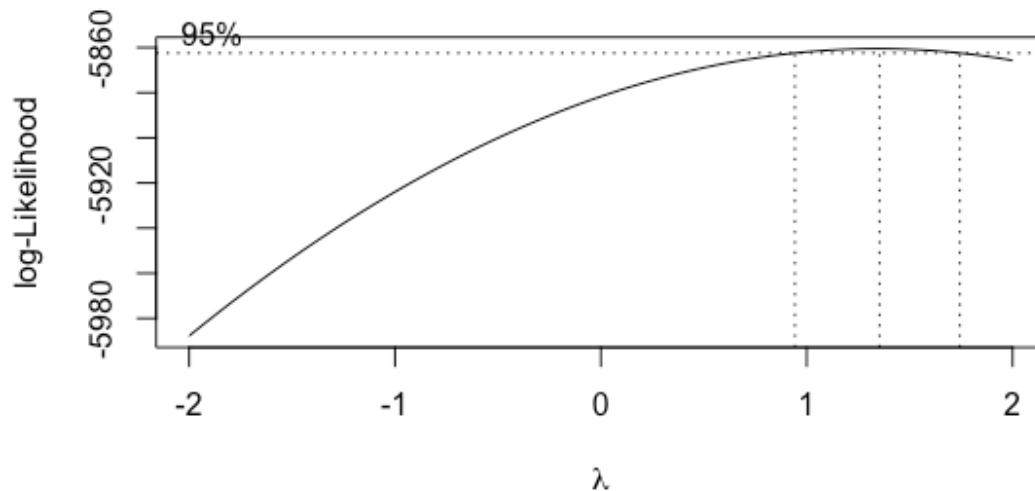
```
boxplot (first_fit$residuals)
```



The residual plots show slight right skewness in the residual distribution. There is also a decrease in variance from left to right. The quantile plot shows little deviation. Though there is slight curvature seen from the plot, there doesn't seem to be any obvious concerns with linearity, constant spread, or normal distribution.

Box-Cox analysis

```
library (MASS)
boxcox (first_fit)
```



The Box-Cox plot shows the confidence interval includes 1, which suggests that there is no need for further transformations of the response variable. Therefore, we can confirm $\log(\text{price})$ is the correct transformation.

We will remove some predictor variables one at a time and observe how results change.

```
second_fit = lm(log(price) ~ sqft_living + bathrooms + bedrooms + log(sqft_lot) + yr_built + waterfront, data=kc_house_data)
summary(second_fit)
```

```
##
## Call:
## lm(formula = log(price) ~ sqft_living + bathrooms + bedrooms +
##     log(sqft_lot) + yr_built + waterfront, data = kc_house_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60452 -0.25822  0.01525  0.25728  1.27710
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.117e+01  3.537e-01  59.86  <2e-16 ***
## sqft_living  4.235e-04  8.692e-06  48.73  <2e-16 ***
## bathrooms    1.273e-01  1.016e-02  12.53  <2e-16 ***
```

```
## bedrooms      -7.134e-02  6.175e-03  -11.55   <2e-16 ***
## log(sqft_lot) -6.874e-02  5.351e-03  -12.85   <2e-16 ***
## yr_built      -4.260e-03  1.797e-04  -23.71   <2e-16 ***
## waterfront     5.950e-01  5.282e-02   11.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3644 on 6968 degrees of freedom
## Multiple R-squared:  0.5222, Adjusted R-squared:  0.5218
## F-statistic: 1269 on 6 and 6968 DF,  p-value: < 2.2e-16
```

```
anova(second_fit)
```

```
## Analysis of Variance Table
##
## Response: log(price)
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## sqft_living  1  880.39   880.39 6631.134 < 2.2e-16 ***
## bathrooms    1    2.35    2.35  17.683 2.642e-05 ***
## bedrooms     1   12.89   12.89  97.068 < 2.2e-16 ***
## log(sqft_lot) 1   17.73   17.73 133.572 < 2.2e-16 ***
## yr_built     1   80.84   80.84 608.866 < 2.2e-16 ***
## waterfront   1   16.85   16.85 126.896 < 2.2e-16 ***
## Residuals   6968 925.11    0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The removal of the floors predictor variable decreases the adjusted R-Squared from 52.78% to 52.18%. The residual standard error is also slightly increased from the first-order model. It is also good to note that the t-values and corresponding p-values for each slope coefficient are all statistically significant at the 5% level. Along with this, the F-value and corresponding p-values are also all statistically significant within the ANOVA table for this model at the 5% level.

let's apply a model selection method to our original first order model.

```
first_fit = lm(log(price) ~ sqft_living + bathrooms + bedrooms + log(sqft_lot)
) + floors + yr_built + waterfront, data=kc_house_data)
summary(first_fit)
```

```
##
## Call:
## lm(formula = log(price) ~ sqft_living + bathrooms + bedrooms +
##     log(sqft_lot) + floors + yr_built + waterfront, data = kc_house_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54354 -0.25763  0.01613  0.25277  1.28461
##
## Coefficients:
```

```

##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.204e+01  3.632e-01  60.686  <2e-16 ***
## sqft_living  4.106e-04  8.744e-06  46.956  <2e-16 ***
## bathrooms    1.101e-01  1.026e-02  10.730  <2e-16 ***
## bedrooms    -6.785e-02  6.147e-03 -11.039  <2e-16 ***
## log(sqft_lot) -4.785e-02  5.756e-03  -8.313  <2e-16 ***
## floors       1.014e-01  1.070e-02   9.477  <2e-16 ***
## yr_built     -4.846e-03  1.890e-04 -25.645  <2e-16 ***
## waterfront   5.740e-01  5.253e-02  10.928  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3621 on 6967 degrees of freedom
## Multiple R-squared:  0.5283, Adjusted R-squared:  0.5278
## F-statistic: 1115 on 7 and 6967 DF,  p-value: < 2.2e-16

```

With the original first order model, lets find the best model selection through utilizing stepwise regression using AIC criterion.

```

fit1aic = step (first_fit, direction='both')

## Start: AIC=-14164.06
## log(price) ~ sqft_living + bathrooms + bedrooms + log(sqft_lot) +
##   floors + yr_built + waterfront
##
##           Df Sum of Sq      RSS   AIC
## <none>                913.34 -14164
## - log(sqft_lot)      1     9.059  922.40 -14097
## - floors              1    11.773  925.11 -14077
## - bathrooms           1    15.092  928.43 -14052
## - waterfront          1    15.656  928.99 -14048
## - bedrooms            1    15.975  929.31 -14045
## - yr_built             1    86.217  999.55 -13537
## - sqft_living         1   289.047 1202.38 -12248

summary(fit1aic)

##
## Call:
## lm(formula = log(price) ~ sqft_living + bathrooms + bedrooms +
##   log(sqft_lot) + floors + yr_built + waterfront, data = kc_house_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54354 -0.25763  0.01613  0.25277  1.28461
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.204e+01  3.632e-01  60.686  <2e-16 ***
## sqft_living  4.106e-04  8.744e-06  46.956  <2e-16 ***
## bathrooms    1.101e-01  1.026e-02  10.730  <2e-16 ***

```

```

## bedrooms      -6.785e-02  6.147e-03 -11.039  <2e-16 ***
## log(sqft_lot) -4.785e-02  5.756e-03  -8.313  <2e-16 ***
## floors        1.014e-01  1.070e-02   9.477  <2e-16 ***
## yr_built      -4.846e-03  1.890e-04 -25.645  <2e-16 ***
## waterfront    5.740e-01  5.253e-02  10.928  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3621 on 6967 degrees of freedom
## Multiple R-squared:  0.5283, Adjusted R-squared:  0.5278
## F-statistic: 1115 on 7 and 6967 DF, p-value: < 2.2e-16

```

According to the AIC stepwise regression model, it is most beneficial to not change anything in our model. Though this seems odd, we did already transformed the response variable and even a predictor variable. Just to further dive into the best model, let's see the best model according to the SBC criterion.

```

# Get the sample size from the dim function
n = dim(kc_house_data)[1]
fit2bic = step(first_fit, direction='both', k=log(n))

## Start: AIC=-14109.26
## log(price) ~ sqft_living + bathrooms + bedrooms + log(sqft_lot) +
##   floors + yr_built + waterfront
##
##           Df Sum of Sq    RSS    AIC
## <none>                913.34 -14109
## - log(sqft_lot)    1     9.059  922.40 -14049
## - floors           1    11.773  925.11 -14029
## - bathrooms        1    15.092  928.43 -14004
## - waterfront        1    15.656  928.99 -14000
## - bedrooms          1    15.975  929.31 -13997
## - yr_built           1    86.217  999.55 -13489
## - sqft_living        1   289.047 1202.38 -12200

summary(fit2bic)

##
## Call:
## lm(formula = log(price) ~ sqft_living + bathrooms + bedrooms +
##   log(sqft_lot) + floors + yr_built + waterfront, data = kc_house_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.54354 -0.25763  0.01613  0.25277  1.28461
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.204e+01  3.632e-01  60.686  <2e-16 ***
## sqft_living  4.106e-04  8.744e-06  46.956  <2e-16 ***
## bathrooms    1.101e-01  1.026e-02  10.730  <2e-16 ***

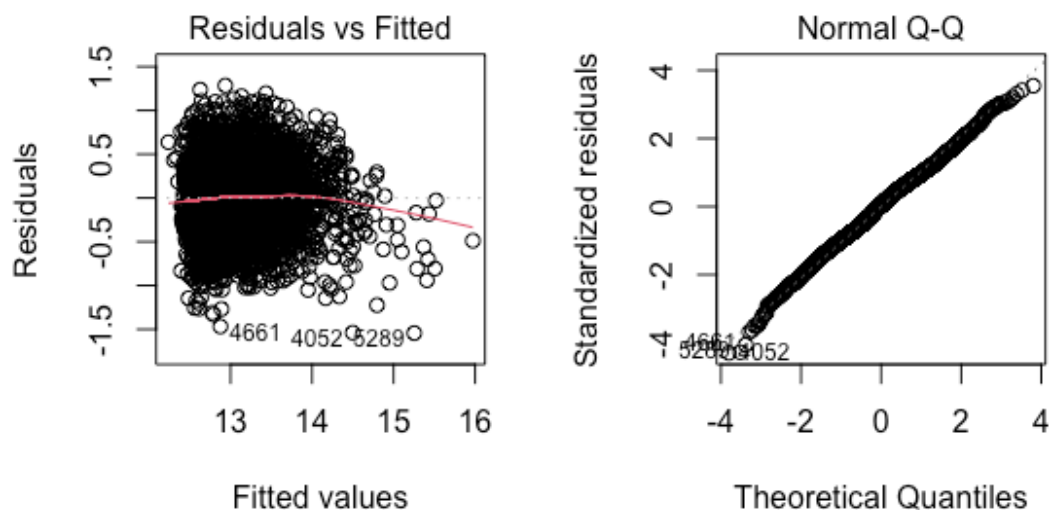
```



```
## bedrooms      -6.785e-02  6.147e-03 -11.039  <2e-16 ***
## log(sqft_lot) -4.785e-02  5.756e-03  -8.313  <2e-16 ***
## floors        1.014e-01  1.070e-02   9.477  <2e-16 ***
## yr_built      -4.846e-03  1.890e-04 -25.645  <2e-16 ***
## waterfront    5.740e-01  5.253e-02  10.928  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3621 on 6967 degrees of freedom
## Multiple R-squared:  0.5283, Adjusted R-squared:  0.5278
## F-statistic: 1115 on 7 and 6967 DF, p-value: < 2.2e-16
```

According to the BIC stepwise-regression model, the first-order model is once again the optimal model to utilize. Now let's check the residual analysis of this first-order model which succeeded to make it through both types of stepwise regression. Note: This plot was shown and interpreted earlier.

```
par(mfrow = c(1, 2))
plot (first_fit, which = c(1, 2))
```



As explained earlier, there doesn't seem to be any obvious concerns with linearity, constant spread, or normal distribution.

Now let's create centered interaction effects into our model

```
kc_house_data$price.c= kc_house_data$price - mean(kc_house_data$price)
kc_house_data$sqft_living.c= kc_house_data$sqft_living - mean(kc_house_data$sqft_living)
kc_house_data$bathrooms.c= kc_house_data$bathrooms - mean(kc_house_data$bathrooms)
kc_house_data$bedrooms.c= kc_house_data$bedrooms - mean(kc_house_data$bedrooms)
```

```

s)
kc_house_data$logsqft_lot= log(kc_house_data$sqft_lot)
kc_house_data$logsqft_lot.c= kc_house_data$logsqft_lot - mean(kc_house_data$logsqft_lot)
kc_house_data$floors.c= kc_house_data$floors - mean(kc_house_data$floors)
kc_house_data$yr_built.c= kc_house_data$yr_built - mean(kc_house_data$yr_built)
kc_house_data$waterfront.c= kc_house_data$waterfront - mean(kc_house_data$waterfront)

```

With new centered variables for each desired attribute, let's put them into new regression model with interaction effects

```

centeredfit = lm(log(price) ~ (sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c + floors.c + yr_built.c + waterfront.c)^2, data=kc_house_data)
summary(centeredfit)

```

```

##
## Call:
## lm(formula = log(price) ~ (sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c + floors.c + yr_built.c + waterfront.c)^2, data = kc_house_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41089 -0.25215  0.01064  0.24596  1.24525
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.305e+01  6.167e-03 2115.416 < 2e-16 ***
## sqft_living.c  4.500e-04  1.010e-05  44.556 < 2e-16 ***
## bathrooms.c    1.075e-01  1.060e-02  10.144 < 2e-16 ***
## bedrooms.c    -7.600e-02  6.645e-03 -11.438 < 2e-16 ***
## logsqft_lot.c -3.988e-02  6.280e-03  -6.351 2.27e-10 ***
## floors.c       6.283e-02  1.185e-02   5.301 1.19e-07 ***
## yr_built.c    -4.844e-03  2.082e-04 -23.272 < 2e-16 ***
## waterfront.c   8.293e-01  9.782e-02   8.478 < 2e-16 ***
## sqft_living.c:bathrooms.c -1.466e-05  8.934e-06  -1.641 0.10091
## sqft_living.c:bedrooms.c -4.387e-05  9.847e-06  -4.455 8.52e-06 ***
## sqft_living.c:logsqft_lot.c -3.652e-06  7.590e-06  -0.481 0.63039
## sqft_living.c:floors.c    1.976e-05  1.922e-05   1.028 0.30375
## sqft_living.c:yr_built.c -4.623e-07  3.649e-07  -1.267 0.20520
## sqft_living.c:waterfront.c -8.689e-05  7.945e-05  -1.094 0.27417
## bathrooms.c:bedrooms.c    2.594e-02  1.108e-02   2.340 0.01929 *
## bathrooms.c:logsqft_lot.c -2.569e-02  1.198e-02  -2.144 0.03208 *
## bathrooms.c:floors.c     -5.595e-02  2.213e-02  -2.528 0.01150 *
## bathrooms.c:yr_built.c    1.148e-03  3.518e-04   3.263 0.00111 **
## bathrooms.c:waterfront.c   3.818e-02  9.926e-02   0.385 0.70052
## bedrooms.c:logsqft_lot.c  4.676e-02  7.942e-03   5.887 4.11e-09 ***
## bedrooms.c:floors.c      4.621e-02  1.443e-02   3.203 0.00137 **

```

```

## bedrooms.c:yr_built.c      -1.199e-03  2.357e-04  -5.087  3.73e-07  ***
## bedrooms.c:waterfront.c    5.353e-02  7.040e-02   0.760  0.44708
## logsqft_lot.c:floors.c     -8.490e-02  1.138e-02  -7.458  9.85e-14  ***
## logsqft_lot.c:yr_built.c   1.789e-03  2.723e-04   6.569  5.44e-11  ***
## logsqft_lot.c:waterfront.c -8.213e-02  8.183e-02  -1.004  0.31557
## floors.c:yr_built.c        1.578e-03  4.798e-04   3.289  0.00101  **
## floors.c:waterfront.c      -9.587e-02  1.119e-01  -0.857  0.39158
## yr_built.c:waterfront.c    4.564e-03  2.295e-03   1.989  0.04676  *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3572 on 6946 degrees of freedom
## Multiple R-squared:  0.5424, Adjusted R-squared:  0.5405
## F-statistic: 294 on 28 and 6946 DF, p-value: < 2.2e-16

```

The model above with centered predictors and interaction effects has a slightly higher adjusted R-Squared (0.5405 vs .5278) and slightly lower residual standard error (.3572 vs .3621)

Apply stepwise regression to the model with centered interaction effects

```

step.aic = step(centeredfit, direction = "both")

## Start: AIC=-14333.68
## log(price) ~ (sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##   floors.c + yr_built.c + waterfront.c)^2
##
##
##           Df Sum of Sq    RSS    AIC
## - bathrooms.c:waterfront.c    1    0.0189 886.06 -14336
## - sqft_living.c:logsqft_lot.c  1    0.0295 886.07 -14336
## - bedrooms.c:waterfront.c     1    0.0737 886.12 -14335
## - floors.c:waterfront.c       1    0.0936 886.14 -14335
## - logsqft_lot.c:waterfront.c  1    0.1285 886.17 -14335
## - sqft_living.c:floors.c      1    0.1349 886.18 -14335
## - sqft_living.c:waterfront.c  1    0.1526 886.19 -14334
## - sqft_living.c:yr_built.c    1    0.2048 886.25 -14334
## <none>                        886.04 -14334
## - sqft_living.c:bathrooms.c   1    0.3434 886.38 -14333
## - yr_built.c:waterfront.c     1    0.5046 886.55 -14332
## - bathrooms.c:logsqft_lot.c  1    0.5862 886.63 -14331
## - bathrooms.c:bedrooms.c     1    0.6987 886.74 -14330
## - bathrooms.c:floors.c       1    0.8151 886.86 -14329
## - bedrooms.c:floors.c        1    1.3084 887.35 -14325
## - bathrooms.c:yr_built.c     1    1.3584 887.40 -14325
## - floors.c:yr_built.c        1    1.3803 887.42 -14325
## - sqft_living.c:bedrooms.c   1    2.5318 888.57 -14316
## - bedrooms.c:yr_built.c      1    3.3012 889.34 -14310
## - bedrooms.c:logsqft_lot.c   1    4.4214 890.46 -14301

```

```

## - logsqft_lot.c:yr_built.c      1      5.5040 891.55 -14292
## - logsqft_lot.c:floors.c        1      7.0952 893.14 -14280
##
## Step: AIC=-14335.53
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##   floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
##   sqft_living.c:bedrooms.c + sqft_living.c:logsqft_lot.c +
##   sqft_living.c:floors.c + sqft_living.c:yr_built.c + sqft_living.c:wate
rfront.c +
##   bathrooms.c:bedrooms.c + bathrooms.c:logsqft_lot.c + bathrooms.c:floor
s.c +
##   bathrooms.c:yr_built.c + bedrooms.c:logsqft_lot.c + bedrooms.c:floors.
c +
##   bedrooms.c:yr_built.c + bedrooms.c:waterfront.c + logsqft_lot.c:floors
.c +
##   logsqft_lot.c:yr_built.c + logsqft_lot.c:waterfront.c + floors.c:yr_bu
ilt.c +
##   floors.c:waterfront.c + yr_built.c:waterfront.c
##
##
##           Df Sum of Sq    RSS    AIC
## - sqft_living.c:logsqft_lot.c  1      0.0336 886.09 -14337
## - floors.c:waterfront.c        1      0.0750 886.14 -14337
## - bedrooms.c:waterfront.c      1      0.1097 886.17 -14337
## - sqft_living.c:floors.c       1      0.1268 886.19 -14336
## - logsqft_lot.c:waterfront.c   1      0.1313 886.19 -14336
## - sqft_living.c:waterfront.c   1      0.1379 886.20 -14336
## - sqft_living.c:yr_built.c     1      0.2028 886.26 -14336
## <none>                          886.06 -14336
## - sqft_living.c:bathrooms.c    1      0.3339 886.39 -14335
## + bathrooms.c:waterfront.c     1      0.0189 886.04 -14334
## - yr_built.c:waterfront.c      1      0.5142 886.57 -14334
## - bathrooms.c:logsqft_lot.c   1      0.5765 886.64 -14333
## - bathrooms.c:bedrooms.c      1      0.6850 886.75 -14332
## - bathrooms.c:floors.c        1      0.8014 886.86 -14331
## - bedrooms.c:floors.c         1      1.3080 887.37 -14327
## - bathrooms.c:yr_built.c      1      1.3461 887.41 -14327
## - floors.c:yr_built.c         1      1.3782 887.44 -14327
## - sqft_living.c:bedrooms.c    1      2.5132 888.57 -14318
## - bedrooms.c:yr_built.c       1      3.2914 889.35 -14312
## - bedrooms.c:logsqft_lot.c    1      4.4133 890.47 -14303
## - logsqft_lot.c:yr_built.c    1      5.4950 891.56 -14294
## - logsqft_lot.c:floors.c      1      7.0800 893.14 -14282
##
## Step: AIC=-14337.27
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##   floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
##   sqft_living.c:bedrooms.c + sqft_living.c:floors.c + sqft_living.c:yr_b
uilt.c +
##   sqft_living.c:waterfront.c + bathrooms.c:bedrooms.c + bathrooms.c:logs
qft_lot.c +

```

```

##      bathrooms.c:floors.c + bathrooms.c:yr_built.c + bedrooms.c:logsqft_lot
.c +
##      bedrooms.c:floors.c + bedrooms.c:yr_built.c + bedrooms.c:waterfront.c
+
##      logsqft_lot.c:floors.c + logsqft_lot.c:yr_built.c + logsqft_lot.c:wate
rfront.c +
##      floors.c:yr_built.c + floors.c:waterfront.c + yr_built.c:waterfront.c
##
##
##              Df Sum of Sq      RSS      AIC
## - floors.c:waterfront.c      1      0.0755 886.17 -14339
## - bedrooms.c:waterfront.c      1      0.1167 886.21 -14338
## - logsqft_lot.c:waterfront.c      1      0.1353 886.23 -14338
## - sqft_living.c:waterfront.c      1      0.1453 886.24 -14338
## - sqft_living.c:floors.c          1      0.1632 886.26 -14338
## - sqft_living.c:yr_built.c        1      0.2145 886.31 -14338
## <none>                          886.09 -14337
## - sqft_living.c:bathrooms.c      1      0.3575 886.45 -14336
## + sqft_living.c:logsqft_lot.c    1      0.0336 886.06 -14336
## + bathrooms.c:waterfront.c      1      0.0229 886.07 -14336
## - yr_built.c:waterfront.c        1      0.5199 886.61 -14335
## - bathrooms.c:bedrooms.c         1      0.7519 886.85 -14333
## - bathrooms.c:floors.c           1      0.8793 886.97 -14332
## - bathrooms.c:logsqft_lot.c      1      1.0141 887.11 -14331
## - bedrooms.c:floors.c            1      1.3035 887.40 -14329
## - floors.c:yr_built.c            1      1.3519 887.45 -14329
## - bathrooms.c:yr_built.c         1      1.3673 887.46 -14328
## - sqft_living.c:bedrooms.c       1      2.6275 888.72 -14319
## - bedrooms.c:yr_built.c          1      3.2579 889.35 -14314
## - bedrooms.c:logsqft_lot.c       1      4.4690 890.56 -14304
## - logsqft_lot.c:yr_built.c       1      5.4884 891.58 -14296
## - logsqft_lot.c:floors.c         1      7.1158 893.21 -14284
##
## Step:  AIC=-14338.68
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##      floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
##      sqft_living.c:bedrooms.c + sqft_living.c:floors.c + sqft_living.c:yr_b
uilt.c +
##      sqft_living.c:waterfront.c + bathrooms.c:bedrooms.c + bathrooms.c:logs
qft_lot.c +
##      bathrooms.c:floors.c + bathrooms.c:yr_built.c + bedrooms.c:logsqft_lot
.c +
##      bedrooms.c:floors.c + bedrooms.c:yr_built.c + bedrooms.c:waterfront.c
+
##      logsqft_lot.c:floors.c + logsqft_lot.c:yr_built.c + logsqft_lot.c:wate
rfront.c +
##      floors.c:yr_built.c + yr_built.c:waterfront.c
##
##
##              Df Sum of Sq      RSS      AIC
## - logsqft_lot.c:waterfront.c      1      0.0975 886.27 -14340
## - bedrooms.c:waterfront.c         1      0.1109 886.28 -14340

```

```

## - sqft_living.c:floors.c      1    0.1477 886.32 -14340
## - sqft_living.c:waterfront.c  1    0.1503 886.32 -14340
## - sqft_living.c:yr_built.c    1    0.2171 886.39 -14339
## <none>                          886.17 -14339
## - sqft_living.c:bathrooms.c   1    0.3443 886.51 -14338
## + floors.c:waterfront.c       1    0.0755 886.09 -14337
## - yr_built.c:waterfront.c     1    0.4466 886.62 -14337
## + sqft_living.c:logsqft_lot.c 1    0.0341 886.14 -14337
## + bathrooms.c:waterfront.c    1    0.0008 886.17 -14337
## - bathrooms.c:bedrooms.c      1    0.7550 886.92 -14335
## - bathrooms.c:floors.c        1    0.8977 887.07 -14334
## - bathrooms.c:logsqft_lot.c   1    1.0449 887.21 -14332
## - bathrooms.c:yr_built.c      1    1.3764 887.55 -14330
## - bedrooms.c:floors.c         1    1.3773 887.55 -14330
## - floors.c:yr_built.c         1    1.3974 887.57 -14330
## - sqft_living.c:bedrooms.c    1    2.6441 888.81 -14320
## - bedrooms.c:yr_built.c       1    3.2890 889.46 -14315
## - bedrooms.c:logsqft_lot.c    1    4.5044 890.67 -14305
## - logsqft_lot.c:yr_built.c    1    5.6278 891.80 -14296
## - logsqft_lot.c:floors.c      1    7.1643 893.33 -14284
##
## Step: AIC=-14339.91
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##   floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
##   sqft_living.c:bedrooms.c + sqft_living.c:floors.c + sqft_living.c:yr_b
##   uilt.c +
##   sqft_living.c:waterfront.c + bathrooms.c:bedrooms.c + bathrooms.c:logs
##   qft_lot.c +
##   bathrooms.c:floors.c + bathrooms.c:yr_built.c + bedrooms.c:logsqft_lot
##   .c +
##   bedrooms.c:floors.c + bedrooms.c:yr_built.c + bedrooms.c:waterfront.c
##   +
##   logsqft_lot.c:floors.c + logsqft_lot.c:yr_built.c + floors.c:yr_built.
##   c +
##   yr_built.c:waterfront.c
##
##
##              Df Sum of Sq    RSS    AIC
## - sqft_living.c:floors.c      1    0.1489 886.42 -14341
## - bedrooms.c:waterfront.c     1    0.1637 886.43 -14341
## - sqft_living.c:yr_built.c    1    0.2227 886.49 -14340
## <none>                          886.27 -14340
## - sqft_living.c:bathrooms.c   1    0.3313 886.60 -14339
## + logsqft_lot.c:waterfront.c   1    0.0975 886.17 -14339
## - sqft_living.c:waterfront.c   1    0.4115 886.68 -14339
## + floors.c:waterfront.c       1    0.0377 886.23 -14338
## + sqft_living.c:logsqft_lot.c 1    0.0375 886.23 -14338
## + bathrooms.c:waterfront.c    1    0.0044 886.26 -14338
## - bathrooms.c:bedrooms.c      1    0.7632 887.03 -14336
## - bathrooms.c:floors.c        1    0.8995 887.17 -14335
## - yr_built.c:waterfront.c     1    0.9351 887.20 -14335

```

```

## - bathrooms.c:logsqft_lot.c      1      1.0926 887.36 -14333
## - bedrooms.c:floors.c            1      1.3583 887.63 -14331
## - bathrooms.c:yr_built.c         1      1.3759 887.64 -14331
## - floors.c:yr_built.c            1      1.4023 887.67 -14331
## - sqft_living.c:bedrooms.c       1      2.6432 888.91 -14321
## - bedrooms.c:yr_built.c          1      3.2869 889.55 -14316
## - bedrooms.c:logsqft_lot.c       1      4.5416 890.81 -14306
## - logsqft_lot.c:yr_built.c       1      5.8046 892.07 -14296
## - logsqft_lot.c:floors.c         1      7.2025 893.47 -14286
##
## Step: AIC=-14340.74
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##   floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
##   sqft_living.c:bedrooms.c + sqft_living.c:yr_built.c + sqft_living.c:wa
terfront.c +
##   bathrooms.c:bedrooms.c + bathrooms.c:logsqft_lot.c + bathrooms.c:floor
s.c +
##   bathrooms.c:yr_built.c + bedrooms.c:logsqft_lot.c + bedrooms.c:floors.
c +
##   bedrooms.c:yr_built.c + bedrooms.c:waterfront.c + logsqft_lot.c:floors
.c +
##   logsqft_lot.c:yr_built.c + floors.c:yr_built.c + yr_built.c:waterfront
.c
##
##
##           Df Sum of Sq    RSS    AIC
## - sqft_living.c:yr_built.c      1     0.1348 886.55 -14342
## - bedrooms.c:waterfront.c       1     0.1496 886.57 -14342
## <none>                           886.42 -14341
## - sqft_living.c:bathrooms.c     1     0.2616 886.68 -14341
## + sqft_living.c:floors.c         1     0.1489 886.27 -14340
## - sqft_living.c:waterfront.c    1     0.3916 886.81 -14340
## + logsqft_lot.c:waterfront.c     1     0.0987 886.32 -14340
## + sqft_living.c:logsqft_lot.c   1     0.0734 886.34 -14339
## + floors.c:waterfront.c         1     0.0272 886.39 -14339
## + bathrooms.c:waterfront.c      1     0.0025 886.41 -14339
## - bathrooms.c:bedrooms.c        1     0.6978 887.11 -14337
## - bathrooms.c:floors.c          1     0.7610 887.18 -14337
## - yr_built.c:waterfront.c       1     0.9457 887.36 -14335
## - bathrooms.c:logsqft_lot.c     1     1.1342 887.55 -14334
## - bathrooms.c:yr_built.c        1     1.2312 887.65 -14333
## - floors.c:yr_built.c           1     1.3256 887.74 -14332
## - bedrooms.c:floors.c           1     1.7465 888.16 -14329
## - sqft_living.c:bedrooms.c      1     2.5777 888.99 -14322
## - bedrooms.c:yr_built.c         1     3.3710 889.79 -14316
## - bedrooms.c:logsqft_lot.c      1     4.5022 890.92 -14307
## - logsqft_lot.c:yr_built.c      1     5.7120 892.13 -14298
## - logsqft_lot.c:floors.c        1     7.5157 893.93 -14284
##
## Step: AIC=-14341.68
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +

```

```

## floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
## sqft_living.c:bedrooms.c + sqft_living.c:waterfront.c + bathrooms.c:be
drooms.c +
## bathrooms.c:logsqft_lot.c + bathrooms.c:floors.c + bathrooms.c:yr_buil
t.c +
## bedrooms.c:logsqft_lot.c + bedrooms.c:floors.c + bedrooms.c:yr_built.c
+
## bedrooms.c:waterfront.c + logsqft_lot.c:floors.c + logsqft_lot.c:yr_bu
ilt.c +
## floors.c:yr_built.c + yr_built.c:waterfront.c
##
##
## Df Sum of Sq RSS AIC
## - bedrooms.c:waterfront.c 1 0.1549 886.71 -14342
## <none> 886.55 -14342
## - sqft_living.c:waterfront.c 1 0.3675 886.92 -14341
## + sqft_living.c:yr_built.c 1 0.1348 886.42 -14341
## + logsqft_lot.c:waterfront.c 1 0.1029 886.45 -14340
## + sqft_living.c:logsqft_lot.c 1 0.0726 886.48 -14340
## + sqft_living.c:floors.c 1 0.0610 886.49 -14340
## + floors.c:waterfront.c 1 0.0311 886.52 -14340
## - sqft_living.c:bathrooms.c 1 0.4912 887.04 -14340
## + bathrooms.c:waterfront.c 1 0.0026 886.55 -14340
## - bathrooms.c:floors.c 1 0.6658 887.22 -14338
## - bathrooms.c:bedrooms.c 1 0.7890 887.34 -14338
## - yr_built.c:waterfront.c 1 0.9184 887.47 -14336
## - bathrooms.c:logsqft_lot.c 1 1.0344 887.59 -14336
## - bathrooms.c:yr_built.c 1 1.1720 887.72 -14334
## - floors.c:yr_built.c 1 1.1915 887.74 -14334
## - bedrooms.c:floors.c 1 1.6777 888.23 -14330
## - sqft_living.c:bedrooms.c 1 2.5470 889.10 -14324
## - bedrooms.c:yr_built.c 1 4.3081 890.86 -14310
## - bedrooms.c:logsqft_lot.c 1 4.4532 891.00 -14309
## - logsqft_lot.c:yr_built.c 1 5.7296 892.28 -14299
## - logsqft_lot.c:floors.c 1 7.9364 894.49 -14282
##
## Step: AIC=-14342.46
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
## floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
## sqft_living.c:bedrooms.c + sqft_living.c:waterfront.c + bathrooms.c:be
drooms.c +
## bathrooms.c:logsqft_lot.c + bathrooms.c:floors.c + bathrooms.c:yr_buil
t.c +
## bedrooms.c:logsqft_lot.c + bedrooms.c:floors.c + bedrooms.c:yr_built.c
+
## logsqft_lot.c:floors.c + logsqft_lot.c:yr_built.c + floors.c:yr_built.
c +
## yr_built.c:waterfront.c
##
##
## Df Sum of Sq RSS AIC
## - sqft_living.c:waterfront.c 1 0.2162 886.92 -14343

```



```

## <none>                                886.71 -14342
## + logsqft_lot.c:waterfront.c          1    0.1552 886.55 -14342
## + bedrooms.c:waterfront.c             1    0.1549 886.55 -14342
## + sqft_living.c:yr_built.c            1    0.1402 886.57 -14342
## + sqft_living.c:logsqft_lot.c         1    0.0832 886.62 -14341
## + sqft_living.c:floors.c              1    0.0514 886.65 -14341
## + bathrooms.c:waterfront.c            1    0.0317 886.67 -14341
## - sqft_living.c:bathrooms.c           1    0.4862 887.19 -14341
## + floors.c:waterfront.c               1    0.0213 886.68 -14341
## - bathrooms.c:floors.c                1    0.6684 887.37 -14339
## - bathrooms.c:bedrooms.c              1    0.7677 887.47 -14338
## - yr_built.c:waterfront.c             1    0.8537 887.56 -14338
## - bathrooms.c:logsqft_lot.c           1    1.0427 887.75 -14336
## - bathrooms.c:yr_built.c              1    1.1656 887.87 -14335
## - floors.c:yr_built.c                 1    1.2006 887.91 -14335
## - bedrooms.c:floors.c                 1    1.7021 888.41 -14331
## - sqft_living.c:bedrooms.c            1    2.4899 889.20 -14325
## - bedrooms.c:yr_built.c               1    4.3557 891.06 -14310
## - bedrooms.c:logsqft_lot.c            1    4.4806 891.19 -14309
## - logsqft_lot.c:yr_built.c            1    5.7552 892.46 -14299
## - logsqft_lot.c:floors.c              1    7.9447 894.65 -14282
##
## Step: AIC=-14342.76
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##   floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
##   sqft_living.c:bedrooms.c + bathrooms.c:bedrooms.c + bathrooms.c:logsqf
t_lot.c +
##   bathrooms.c:floors.c + bathrooms.c:yr_built.c + bedrooms.c:logsqft_lot
.c +
##   bedrooms.c:floors.c + bedrooms.c:yr_built.c + logsqft_lot.c:floors.c +
##   logsqft_lot.c:yr_built.c + floors.c:yr_built.c + yr_built.c:waterfront
.c
##
##                                     Df Sum of Sq    RSS    AIC
## + logsqft_lot.c:waterfront.c        1    0.3393 886.58 -14343
## <none>                                886.92 -14343
## + sqft_living.c:waterfront.c        1    0.2162 886.71 -14342
## + sqft_living.c:yr_built.c          1    0.1067 886.82 -14342
## + sqft_living.c:logsqft_lot.c       1    0.0959 886.83 -14342
## + sqft_living.c:floors.c            1    0.0559 886.87 -14341
## + bathrooms.c:waterfront.c          1    0.0473 886.87 -14341
## + floors.c:waterfront.c             1    0.0113 886.91 -14341
## + bedrooms.c:waterfront.c           1    0.0036 886.92 -14341
## - sqft_living.c:bathrooms.c         1    0.5797 887.50 -14340
## - bathrooms.c:floors.c              1    0.6269 887.55 -14340
## - yr_built.c:waterfront.c           1    0.6786 887.60 -14339
## - bathrooms.c:bedrooms.c            1    0.7429 887.66 -14339
## - bathrooms.c:logsqft_lot.c         1    1.0287 887.95 -14337
## - floors.c:yr_built.c               1    1.1687 888.09 -14336
## - bathrooms.c:yr_built.c            1    1.2275 888.15 -14335

```

```

## - bedrooms.c:floors.c          1    1.6562 888.58 -14332
## - sqft_living.c:bedrooms.c     1    2.4230 889.34 -14326
## - bedrooms.c:yr_built.c       1    4.2873 891.21 -14311
## - bedrooms.c:logsqft_lot.c    1    4.4209 891.34 -14310
## - logsqft_lot.c:yr_built.c    1    5.8549 892.78 -14299
## - logsqft_lot.c:floors.c      1    7.9925 894.91 -14282
##
## Step:  AIC=-14343.43
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##   floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
##   sqft_living.c:bedrooms.c + bathrooms.c:bedrooms.c + bathrooms.c:logsqf
t_lot.c +
##   bathrooms.c:floors.c + bathrooms.c:yr_built.c + bedrooms.c:logsqft_lot
.c +
##   bedrooms.c:floors.c + bedrooms.c:yr_built.c + logsqft_lot.c:floors.c +
##   logsqft_lot.c:yr_built.c + floors.c:yr_built.c + yr_built.c:waterfront
.c +
##   logsqft_lot.c:waterfront.c
##
##              Df Sum of Sq    RSS    AIC
## <none>                                886.58 -14343
## - logsqft_lot.c:waterfront.c          1     0.3393 886.92 -14343
## - yr_built.c:waterfront.c             1     0.3535 886.94 -14343
## + sqft_living.c:yr_built.c            1     0.1207 886.46 -14342
## + sqft_living.c:logsqft_lot.c         1     0.0765 886.51 -14342
## + floors.c:waterfront.c               1     0.0674 886.52 -14342
## + sqft_living.c:floors.c              1     0.0554 886.53 -14342
## + sqft_living.c:waterfront.c          1     0.0321 886.55 -14342
## + bedrooms.c:waterfront.c             1     0.0100 886.57 -14342
## + bathrooms.c:waterfront.c            1     0.0026 886.58 -14341
## - sqft_living.c:bathrooms.c           1     0.5462 887.13 -14341
## - bathrooms.c:floors.c                1     0.6573 887.24 -14340
## - bathrooms.c:bedrooms.c              1     0.7516 887.33 -14340
## - bathrooms.c:logsqft_lot.c           1     0.9716 887.55 -14338
## - floors.c:yr_built.c                  1     1.1877 887.77 -14336
## - bathrooms.c:yr_built.c              1     1.2042 887.79 -14336
## - bedrooms.c:floors.c                  1     1.7161 888.30 -14332
## - sqft_living.c:bedrooms.c            1     2.4875 889.07 -14326
## - bedrooms.c:yr_built.c               1     4.3166 890.90 -14312
## - bedrooms.c:logsqft_lot.c            1     4.4070 890.99 -14311
## - logsqft_lot.c:yr_built.c            1     5.5368 892.12 -14302
## - logsqft_lot.c:floors.c              1     7.8901 894.47 -14284

```

```
summary(step.aic)
```

```

##
## Call:
## lm(formula = log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c +
##   logsqft_lot.c + floors.c + yr_built.c + waterfront.c + sqft_living.c:b
athrooms.c +

```

```

## sqft_living.c:bedrooms.c + bathrooms.c:bedrooms.c + bathrooms.c:logsqf
t_lot.c +
## bathrooms.c:floors.c + bathrooms.c:yr_built.c + bedrooms.c:logsqft_lot
.c +
## bedrooms.c:floors.c + bedrooms.c:yr_built.c + logsqft_lot.c:floors.c +
## logsqft_lot.c:yr_built.c + floors.c:yr_built.c + yr_built.c:waterfront
.c +
## logsqft_lot.c:waterfront.c, data = kc_house_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40999 -0.25231  0.01182  0.24631  1.24384
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.305e+01  5.988e-03 2178.674 < 2e-16 ***
## sqft_living.c  4.492e-04  9.945e-06  45.168 < 2e-16 ***
## bathrooms.c    1.079e-01  1.049e-02  10.288 < 2e-16 ***
## bedrooms.c    -7.596e-02  6.524e-03 -11.642 < 2e-16 ***
## logsqft_lot.c -4.065e-02  6.195e-03  -6.561 5.73e-11 ***
## floors.c       6.339e-02  1.176e-02   5.392 7.19e-08 ***
## yr_built.c    -4.862e-03  2.063e-04 -23.565 < 2e-16 ***
## waterfront.c   7.373e-01  6.397e-02  11.527 < 2e-16 ***
## sqft_living.c:bathrooms.c -1.656e-05  7.999e-06  -2.070 0.038527 *
## sqft_living.c:bedrooms.c -4.283e-05  9.698e-06  -4.417 1.02e-05 ***
## bathrooms.c:bedrooms.c    2.608e-02  1.074e-02   2.428 0.015217 *
## bathrooms.c:logsqft_lot.c -2.755e-02  9.980e-03  -2.760 0.005787 **
## bathrooms.c:floors.c     -4.189e-02  1.845e-02  -2.270 0.023211 *
## bathrooms.c:yr_built.c    8.714e-04  2.836e-04   3.073 0.002127 **
## bedrooms.c:logsqft_lot.c  4.537e-02  7.717e-03   5.879 4.32e-09 ***
## bedrooms.c:floors.c       5.046e-02  1.375e-02   3.669 0.000246 ***
## bedrooms.c:yr_built.c    -1.283e-03  2.204e-04  -5.818 6.21e-09 ***
## logsqft_lot.c:floors.c    -8.202e-02  1.043e-02  -7.866 4.21e-15 ***
## logsqft_lot.c:yr_built.c  1.697e-03  2.575e-04   6.590 4.73e-11 ***
## floors.c:yr_built.c       1.383e-03  4.532e-04   3.052 0.002282 **
## yr_built.c:waterfront.c   3.109e-03  1.867e-03   1.665 0.095930 .
## logsqft_lot.c:waterfront.c -1.069e-01  6.553e-02  -1.631 0.102860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3571 on 6953 degrees of freedom
## Multiple R-squared:  0.5421, Adjusted R-squared:  0.5407
## F-statistic: 392 on 21 and 6953 DF, p-value: < 2.2e-16

```

None of the predictor variables were removed. Fourteen interaction effects were kept. Not all of them are significant ($p > .05$). Next we will try the BIC criterion.

```
step.bic = step(centeredfit, direction = "both", k=log(n))
```

```

## Start: AIC=-14135.03
## log(price) ~ (sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##   floors.c + yr_built.c + waterfront.c)^2
##
##           Df Sum of Sq   RSS   AIC
## - bathrooms.c:waterfront.c  1    0.0189 886.06 -14144
## - sqft_living.c:logsqft_lot.c  1    0.0295 886.07 -14144
## - bedrooms.c:waterfront.c  1    0.0737 886.12 -14143
## - floors.c:waterfront.c  1    0.0936 886.14 -14143
## - logsqft_lot.c:waterfront.c  1    0.1285 886.17 -14143
## - sqft_living.c:floors.c  1    0.1349 886.18 -14143
## - sqft_living.c:waterfront.c  1    0.1526 886.19 -14143
## - sqft_living.c:yr_built.c  1    0.2048 886.25 -14142
## - sqft_living.c:bathrooms.c  1    0.3434 886.38 -14141
## - yr_built.c:waterfront.c  1    0.5046 886.55 -14140
## - bathrooms.c:logsqft_lot.c  1    0.5862 886.63 -14139
## - bathrooms.c:bedrooms.c  1    0.6987 886.74 -14138
## - bathrooms.c:floors.c  1    0.8151 886.86 -14138
## <none>                                886.04 -14135
## - bedrooms.c:floors.c  1    1.3084 887.35 -14134
## - bathrooms.c:yr_built.c  1    1.3584 887.40 -14133
## - floors.c:yr_built.c  1    1.3803 887.42 -14133
## - sqft_living.c:bedrooms.c  1    2.5318 888.57 -14124
## - bedrooms.c:yr_built.c  1    3.3012 889.34 -14118
## - bedrooms.c:logsqft_lot.c  1    4.4214 890.46 -14109
## - logsqft_lot.c:yr_built.c  1    5.5040 891.55 -14101
## - logsqft_lot.c:floors.c  1    7.0952 893.14 -14088
##
## Step: AIC=-14143.73
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##   floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
##   sqft_living.c:bedrooms.c + sqft_living.c:logsqft_lot.c +
##   sqft_living.c:floors.c + sqft_living.c:yr_built.c + sqft_living.c:water
rfront.c +
##   bathrooms.c:bedrooms.c + bathrooms.c:logsqft_lot.c + bathrooms.c:floor
s.c +
##   bathrooms.c:yr_built.c + bedrooms.c:logsqft_lot.c + bedrooms.c:floors.
c +
##   bedrooms.c:yr_built.c + bedrooms.c:waterfront.c + logsqft_lot.c:floors
.c +
##   logsqft_lot.c:yr_built.c + logsqft_lot.c:waterfront.c + floors.c:yr_bu
ilt.c +
##   floors.c:waterfront.c + yr_built.c:waterfront.c
##
##           Df Sum of Sq   RSS   AIC
## - sqft_living.c:logsqft_lot.c  1    0.0336 886.09 -14152
## - floors.c:waterfront.c  1    0.0750 886.14 -14152
## - bedrooms.c:waterfront.c  1    0.1097 886.17 -14152
## - sqft_living.c:floors.c  1    0.1268 886.19 -14152
## - logsqft_lot.c:waterfront.c  1    0.1313 886.19 -14152

```

```

## - sqft_living.c:waterfront.c 1 0.1379 886.20 -14152
## - sqft_living.c:yr_built.c 1 0.2028 886.26 -14151
## - sqft_living.c:bathrooms.c 1 0.3339 886.39 -14150
## - yr_built.c:waterfront.c 1 0.5142 886.57 -14148
## - bathrooms.c:logsqft_lot.c 1 0.5765 886.64 -14148
## - bathrooms.c:bedrooms.c 1 0.6850 886.75 -14147
## - bathrooms.c:floors.c 1 0.8014 886.86 -14146
## <none> 886.06 -14144
## - bedrooms.c:floors.c 1 1.3080 887.37 -14142
## - bathrooms.c:yr_built.c 1 1.3461 887.41 -14142
## - floors.c:yr_built.c 1 1.3782 887.44 -14142
## + bathrooms.c:waterfront.c 1 0.0189 886.04 -14135
## - sqft_living.c:bedrooms.c 1 2.5132 888.57 -14133
## - bedrooms.c:yr_built.c 1 3.2914 889.35 -14127
## - bedrooms.c:logsqft_lot.c 1 4.4133 890.47 -14118
## - logsqft_lot.c:yr_built.c 1 5.4950 891.56 -14110
## - logsqft_lot.c:floors.c 1 7.0800 893.14 -14097
##
## Step: AIC=-14152.32
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
## floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
## sqft_living.c:bedrooms.c + sqft_living.c:floors.c + sqft_living.c:yr_b
uilt.c +
## sqft_living.c:waterfront.c + bathrooms.c:bedrooms.c + bathrooms.c:logs
qft_lot.c +
## bathrooms.c:floors.c + bathrooms.c:yr_built.c + bedrooms.c:logsqft_lot
.c +
## bedrooms.c:floors.c + bedrooms.c:yr_built.c + bedrooms.c:waterfront.c
+
## logsqft_lot.c:floors.c + logsqft_lot.c:yr_built.c + logsqft_lot.c:wate
rfront.c +
## floors.c:yr_built.c + floors.c:waterfront.c + yr_built.c:waterfront.c
##
## Df Sum of Sq RSS AIC
## - floors.c:waterfront.c 1 0.0755 886.17 -14161
## - bedrooms.c:waterfront.c 1 0.1167 886.21 -14160
## - logsqft_lot.c:waterfront.c 1 0.1353 886.23 -14160
## - sqft_living.c:waterfront.c 1 0.1453 886.24 -14160
## - sqft_living.c:floors.c 1 0.1632 886.26 -14160
## - sqft_living.c:yr_built.c 1 0.2145 886.31 -14160
## - sqft_living.c:bathrooms.c 1 0.3575 886.45 -14158
## - yr_built.c:waterfront.c 1 0.5199 886.61 -14157
## - bathrooms.c:bedrooms.c 1 0.7519 886.85 -14155
## - bathrooms.c:floors.c 1 0.8793 886.97 -14154
## - bathrooms.c:logsqft_lot.c 1 1.0141 887.11 -14153
## <none> 886.09 -14152
## - bedrooms.c:floors.c 1 1.3035 887.40 -14151
## - floors.c:yr_built.c 1 1.3519 887.45 -14150
## - bathrooms.c:yr_built.c 1 1.3673 887.46 -14150
## + sqft_living.c:logsqft_lot.c 1 0.0336 886.06 -14144

```

```

## + bathrooms.c:waterfront.c      1      0.0229 886.07 -14144
## - sqft_living.c:bedrooms.c      1      2.6275 888.72 -14140
## - bedrooms.c:yr_built.c         1      3.2579 889.35 -14136
## - bedrooms.c:logsqft_lot.c      1      4.4690 890.56 -14126
## - logsqft_lot.c:yr_built.c      1      5.4884 891.58 -14118
## - logsqft_lot.c:floors.c        1      7.1158 893.21 -14105
##
## Step:  AIC=-14160.57
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##   floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
##   sqft_living.c:bedrooms.c + sqft_living.c:floors.c + sqft_living.c:yr_b
uilt.c +
##   sqft_living.c:waterfront.c + bathrooms.c:bedrooms.c + bathrooms.c:logs
qft_lot.c +
##   bathrooms.c:floors.c + bathrooms.c:yr_built.c + bedrooms.c:logsqft_lot
.c +
##   bedrooms.c:floors.c + bedrooms.c:yr_built.c + bedrooms.c:waterfront.c
+
##   logsqft_lot.c:floors.c + logsqft_lot.c:yr_built.c + logsqft_lot.c:wate
rfront.c +
##   floors.c:yr_built.c + yr_built.c:waterfront.c
##
##
##              Df Sum of Sq    RSS    AIC
## - logsqft_lot.c:waterfront.c  1      0.0975 886.27 -14169
## - bedrooms.c:waterfront.c    1      0.1109 886.28 -14169
## - sqft_living.c:floors.c     1      0.1477 886.32 -14168
## - sqft_living.c:waterfront.c 1      0.1503 886.32 -14168
## - sqft_living.c:yr_built.c   1      0.2171 886.39 -14168
## - sqft_living.c:bathrooms.c  1      0.3443 886.51 -14167
## - yr_built.c:waterfront.c    1      0.4466 886.62 -14166
## - bathrooms.c:bedrooms.c     1      0.7550 886.92 -14164
## - bathrooms.c:floors.c       1      0.8977 887.07 -14162
## - bathrooms.c:logsqft_lot.c  1      1.0449 887.21 -14161
## <none>                        886.17 -14161
## - bathrooms.c:yr_built.c     1      1.3764 887.55 -14159
## - bedrooms.c:floors.c        1      1.3773 887.55 -14159
## - floors.c:yr_built.c        1      1.3974 887.57 -14158
## + floors.c:waterfront.c      1      0.0755 886.09 -14152
## + sqft_living.c:logsqft_lot.c 1      0.0341 886.14 -14152
## + bathrooms.c:waterfront.c   1      0.0008 886.17 -14152
## - sqft_living.c:bedrooms.c   1      2.6441 888.81 -14149
## - bedrooms.c:yr_built.c      1      3.2890 889.46 -14144
## - bedrooms.c:logsqft_lot.c   1      4.5044 890.67 -14134
## - logsqft_lot.c:yr_built.c   1      5.6278 891.80 -14125
## - logsqft_lot.c:floors.c     1      7.1643 893.33 -14113
##
## Step:  AIC=-14168.66
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##   floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
##   sqft_living.c:bedrooms.c + sqft_living.c:floors.c + sqft_living.c:yr_b

```

```

uilt.c +
## sqft_living.c:waterfront.c + bathrooms.c:bedrooms.c + bathrooms.c:logs
qft_lot.c +
## bathrooms.c:floors.c + bathrooms.c:yr_built.c + bedrooms.c:logsqft_lot
.c +
## bedrooms.c:floors.c + bedrooms.c:yr_built.c + bedrooms.c:waterfront.c
+
## logsqft_lot.c:floors.c + logsqft_lot.c:yr_built.c + floors.c:yr_built.
c +
## yr_built.c:waterfront.c
##
##
## Df Sum of Sq RSS AIC
## - sqft_living.c:floors.c 1 0.1489 886.42 -14176
## - bedrooms.c:waterfront.c 1 0.1637 886.43 -14176
## - sqft_living.c:yr_built.c 1 0.2227 886.49 -14176
## - sqft_living.c:bathrooms.c 1 0.3313 886.60 -14175
## - sqft_living.c:waterfront.c 1 0.4115 886.68 -14174
## - bathrooms.c:bedrooms.c 1 0.7632 887.03 -14172
## - bathrooms.c:floors.c 1 0.8995 887.17 -14170
## - yr_built.c:waterfront.c 1 0.9351 887.20 -14170
## - bathrooms.c:logsqft_lot.c 1 1.0926 887.36 -14169
## <none> 886.27 -14169
## - bedrooms.c:floors.c 1 1.3583 887.63 -14167
## - bathrooms.c:yr_built.c 1 1.3759 887.64 -14167
## - floors.c:yr_built.c 1 1.4023 887.67 -14166
## + logsqft_lot.c:waterfront.c 1 0.0975 886.17 -14161
## + floors.c:waterfront.c 1 0.0377 886.23 -14160
## + sqft_living.c:logsqft_lot.c 1 0.0375 886.23 -14160
## + bathrooms.c:waterfront.c 1 0.0044 886.26 -14160
## - sqft_living.c:bedrooms.c 1 2.6432 888.91 -14157
## - bedrooms.c:yr_built.c 1 3.2869 889.55 -14152
## - bedrooms.c:logsqft_lot.c 1 4.5416 890.81 -14142
## - logsqft_lot.c:yr_built.c 1 5.8046 892.07 -14132
## - logsqft_lot.c:floors.c 1 7.2025 893.47 -14121
##
## Step: AIC=-14176.33
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
## floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
## sqft_living.c:bedrooms.c + sqft_living.c:yr_built.c + sqft_living.c:wa
terfront.c +
## bathrooms.c:bedrooms.c + bathrooms.c:logsqft_lot.c + bathrooms.c:floor
s.c +
## bathrooms.c:yr_built.c + bedrooms.c:logsqft_lot.c + bedrooms.c:floors.
c +
## bedrooms.c:yr_built.c + bedrooms.c:waterfront.c + logsqft_lot.c:floors
.c +
## logsqft_lot.c:yr_built.c + floors.c:yr_built.c + yr_built.c:waterfront
.c
##
## Df Sum of Sq RSS AIC

```

```

## - sqft_living.c:yr_built.c      1    0.1348 886.55 -14184
## - bedrooms.c:waterfront.c      1    0.1496 886.57 -14184
## - sqft_living.c:bathrooms.c    1    0.2616 886.68 -14183
## - sqft_living.c:waterfront.c   1    0.3916 886.81 -14182
## - bathrooms.c:bedrooms.c       1    0.6978 887.11 -14180
## - bathrooms.c:floors.c         1    0.7610 887.18 -14179
## - yr_built.c:waterfront.c      1    0.9457 887.36 -14178
## <none>                          886.42 -14176
## - bathrooms.c:logsqft_lot.c    1    1.1342 887.55 -14176
## - bathrooms.c:yr_built.c       1    1.2312 887.65 -14176
## - floors.c:yr_built.c          1    1.3256 887.74 -14175
## - bedrooms.c:floors.c          1    1.7465 888.16 -14172
## + sqft_living.c:floors.c        1    0.1489 886.27 -14169
## + logsqft_lot.c:waterfront.c    1    0.0987 886.32 -14168
## + sqft_living.c:logsqft_lot.c  1    0.0734 886.34 -14168
## + floors.c:waterfront.c        1    0.0272 886.39 -14168
## + bathrooms.c:waterfront.c     1    0.0025 886.41 -14168
## - sqft_living.c:bedrooms.c     1    2.5777 888.99 -14165
## - bedrooms.c:yr_built.c        1    3.3710 889.79 -14159
## - bedrooms.c:logsqft_lot.c     1    4.5022 890.92 -14150
## - logsqft_lot.c:yr_built.c     1    5.7120 892.13 -14140
## - logsqft_lot.c:floors.c       1    7.5157 893.93 -14126
##
## Step: AIC=-14184.12
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##   floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
##   sqft_living.c:bedrooms.c + sqft_living.c:waterfront.c + bathrooms.c:be
drooms.c +
##   bathrooms.c:logsqft_lot.c + bathrooms.c:floors.c + bathrooms.c:yr_buil
t.c +
##   bedrooms.c:logsqft_lot.c + bedrooms.c:floors.c + bedrooms.c:yr_built.c
+
##   bedrooms.c:waterfront.c + logsqft_lot.c:floors.c + logsqft_lot.c:yr_bu
ilt.c +
##   floors.c:yr_built.c + yr_built.c:waterfront.c
##
##
##              Df Sum of Sq    RSS    AIC
## - bedrooms.c:waterfront.c      1    0.1549 886.71 -14192
## - sqft_living.c:waterfront.c    1    0.3675 886.92 -14190
## - sqft_living.c:bathrooms.c    1    0.4912 887.04 -14189
## - bathrooms.c:floors.c         1    0.6658 887.22 -14188
## - bathrooms.c:bedrooms.c       1    0.7890 887.34 -14187
## - yr_built.c:waterfront.c      1    0.9184 887.47 -14186
## - bathrooms.c:logsqft_lot.c    1    1.0344 887.59 -14185
## <none>                          886.55 -14184
## - bathrooms.c:yr_built.c       1    1.1720 887.72 -14184
## - floors.c:yr_built.c          1    1.1915 887.74 -14184
## - bedrooms.c:floors.c          1    1.6777 888.23 -14180
## + sqft_living.c:yr_built.c     1    0.1348 886.42 -14176
## + logsqft_lot.c:waterfront.c   1    0.1029 886.45 -14176

```



```

## + sqft_living.c:logsqft_lot.c 1 0.0726 886.48 -14176
## + sqft_living.c:floors.c 1 0.0610 886.49 -14176
## + floors.c:waterfront.c 1 0.0311 886.52 -14176
## + bathrooms.c:waterfront.c 1 0.0026 886.55 -14175
## - sqft_living.c:bedrooms.c 1 2.5470 889.10 -14173
## - bedrooms.c:yr_built.c 1 4.3081 890.86 -14159
## - bedrooms.c:logsqft_lot.c 1 4.4532 891.00 -14158
## - logsqft_lot.c:yr_built.c 1 5.7296 892.28 -14148
## - logsqft_lot.c:floors.c 1 7.9364 894.49 -14131
##
## Step: AIC=-14191.76
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
## floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +
## sqft_living.c:bedrooms.c + sqft_living.c:waterfront.c + bathrooms.c:be
drooms.c +
## bathrooms.c:logsqft_lot.c + bathrooms.c:floors.c + bathrooms.c:yr_buil
t.c +
## bedrooms.c:logsqft_lot.c + bedrooms.c:floors.c + bedrooms.c:yr_built.c
+
## logsqft_lot.c:floors.c + logsqft_lot.c:yr_built.c + floors.c:yr_built.
c +
## yr_built.c:waterfront.c
##
##
## Df Sum of Sq RSS AIC
## - sqft_living.c:waterfront.c 1 0.2162 886.92 -14199
## - sqft_living.c:bathrooms.c 1 0.4862 887.19 -14197
## - bathrooms.c:floors.c 1 0.6684 887.37 -14195
## - bathrooms.c:bedrooms.c 1 0.7677 887.47 -14195
## - yr_built.c:waterfront.c 1 0.8537 887.56 -14194
## - bathrooms.c:logsqft_lot.c 1 1.0427 887.75 -14192
## <none> 886.71 -14192
## - bathrooms.c:yr_built.c 1 1.1656 887.87 -14191
## - floors.c:yr_built.c 1 1.2006 887.91 -14191
## - bedrooms.c:floors.c 1 1.7021 888.41 -14187
## + logsqft_lot.c:waterfront.c 1 0.1552 886.55 -14184
## + bedrooms.c:waterfront.c 1 0.1549 886.55 -14184
## + sqft_living.c:yr_built.c 1 0.1402 886.57 -14184
## + sqft_living.c:logsqft_lot.c 1 0.0832 886.62 -14184
## + sqft_living.c:floors.c 1 0.0514 886.65 -14183
## + bathrooms.c:waterfront.c 1 0.0317 886.67 -14183
## + floors.c:waterfront.c 1 0.0213 886.68 -14183
## - sqft_living.c:bedrooms.c 1 2.4899 889.20 -14181
## - bedrooms.c:yr_built.c 1 4.3557 891.06 -14166
## - bedrooms.c:logsqft_lot.c 1 4.4806 891.19 -14165
## - logsqft_lot.c:yr_built.c 1 5.7552 892.46 -14156
## - logsqft_lot.c:floors.c 1 7.9447 894.65 -14138
##
## Step: AIC=-14198.9
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
## floors.c + yr_built.c + waterfront.c + sqft_living.c:bathrooms.c +

```

```

## sqft_living.c:bedrooms.c + bathrooms.c:bedrooms.c + bathrooms.c:logsqf
t_lot.c +
## bathrooms.c:floors.c + bathrooms.c:yr_built.c + bedrooms.c:logsqft_lot
.c +
## bedrooms.c:floors.c + bedrooms.c:yr_built.c + logsqft_lot.c:floors.c +
## logsqft_lot.c:yr_built.c + floors.c:yr_built.c + yr_built.c:waterfront
.c
##
##
## Df Sum of Sq RSS AIC
## - sqft_living.c:bathrooms.c 1 0.5797 887.50 -14203
## - bathrooms.c:floors.c 1 0.6269 887.55 -14203
## - yr_built.c:waterfront.c 1 0.6786 887.60 -14202
## - bathrooms.c:bedrooms.c 1 0.7429 887.66 -14202
## - bathrooms.c:logsqft_lot.c 1 1.0287 887.95 -14200
## <none> 886.92 -14199
## - floors.c:yr_built.c 1 1.1687 888.09 -14199
## - bathrooms.c:yr_built.c 1 1.2275 888.15 -14198
## - bedrooms.c:floors.c 1 1.6562 888.58 -14195
## + logsqft_lot.c:waterfront.c 1 0.3393 886.58 -14193
## + sqft_living.c:waterfront.c 1 0.2162 886.71 -14192
## + sqft_living.c:yr_built.c 1 0.1067 886.82 -14191
## + sqft_living.c:logsqft_lot.c 1 0.0959 886.83 -14191
## + sqft_living.c:floors.c 1 0.0559 886.87 -14190
## + bathrooms.c:waterfront.c 1 0.0473 886.87 -14190
## + floors.c:waterfront.c 1 0.0113 886.91 -14190
## + bedrooms.c:waterfront.c 1 0.0036 886.92 -14190
## - sqft_living.c:bedrooms.c 1 2.4230 889.34 -14189
## - bedrooms.c:yr_built.c 1 4.2873 891.21 -14174
## - bedrooms.c:logsqft_lot.c 1 4.4209 891.34 -14173
## - logsqft_lot.c:yr_built.c 1 5.8549 892.78 -14162
## - logsqft_lot.c:floors.c 1 7.9925 894.91 -14145
##
## Step: AIC=-14203.2
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
## floors.c + yr_built.c + waterfront.c + sqft_living.c:bedrooms.c +
## bathrooms.c:bedrooms.c + bathrooms.c:logsqft_lot.c + bathrooms.c:floor
s.c +
## bathrooms.c:yr_built.c + bedrooms.c:logsqft_lot.c + bedrooms.c:floors.
c +
## bedrooms.c:yr_built.c + logsqft_lot.c:floors.c + logsqft_lot.c:yr_buil
t.c +
## floors.c:yr_built.c + yr_built.c:waterfront.c
##
##
## Df Sum of Sq RSS AIC
## - bathrooms.c:bedrooms.c 1 0.5522 888.05 -14208
## - yr_built.c:waterfront.c 1 0.6160 888.12 -14207
## - bathrooms.c:yr_built.c 1 0.8714 888.37 -14205
## <none> 887.50 -14203
## - floors.c:yr_built.c 1 1.5542 889.06 -14200
## + sqft_living.c:bathrooms.c 1 0.5797 886.92 -14199

```

```

## - bathrooms.c:floors.c          1      1.6876 889.19 -14199
## + logsqft_lot.c:waterfront.c     1      0.3729 887.13 -14197
## + sqft_living.c:yr_built.c       1      0.3400 887.16 -14197
## + sqft_living.c:waterfront.c     1      0.3097 887.19 -14197
## + sqft_living.c:logsqft_lot.c    1      0.1261 887.38 -14195
## + bathrooms.c:waterfront.c       1      0.0939 887.41 -14195
## - bedrooms.c:floors.c            1      2.1697 889.67 -14195
## + bedrooms.c:waterfront.c        1      0.0177 887.48 -14194
## + floors.c:waterfront.c           1      0.0087 887.49 -14194
## + sqft_living.c:floors.c         1      0.0002 887.50 -14194
## - bathrooms.c:logsqft_lot.c      1      2.6370 890.14 -14191
## - sqft_living.c:bedrooms.c       1      3.7045 891.21 -14183
## - bedrooms.c:yr_built.c          1      4.1108 891.61 -14180
## - bedrooms.c:logsqft_lot.c       1      5.5173 893.02 -14169
## - logsqft_lot.c:yr_built.c       1      6.7588 894.26 -14159
## - logsqft_lot.c:floors.c         1      7.7875 895.29 -14151
##
## Step: AIC=-14207.71
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##   floors.c + yr_built.c + waterfront.c + sqft_living.c:bedrooms.c +
##   bathrooms.c:logsqft_lot.c + bathrooms.c:floors.c + bathrooms.c:yr_built.c +
##   bedrooms.c:logsqft_lot.c + bedrooms.c:floors.c + bedrooms.c:yr_built.c +
##   logsqft_lot.c:floors.c + logsqft_lot.c:yr_built.c + floors.c:yr_built.c +
##   yr_built.c:waterfront.c
##
##
##           Df Sum of Sq    RSS    AIC
## - yr_built.c:waterfront.c     1     0.6456 888.70 -14212
## - bathrooms.c:yr_built.c      1     0.7943 888.85 -14210
## <none>                          888.05 -14208
## - bathrooms.c:floors.c        1     1.4786 889.53 -14205
## - floors.c:yr_built.c         1     1.4851 889.54 -14205
## + bathrooms.c:bedrooms.c      1     0.5522 887.50 -14203
## + sqft_living.c:yr_built.c    1     0.3904 887.66 -14202
## + sqft_living.c:bathrooms.c   1     0.3891 887.66 -14202
## + logsqft_lot.c:waterfront.c  1     0.3591 887.69 -14202
## + sqft_living.c:waterfront.c  1     0.2663 887.79 -14201
## + sqft_living.c:logsqft_lot.c 1     0.1953 887.86 -14200
## + bathrooms.c:waterfront.c    1     0.0946 887.96 -14200
## + bedrooms.c:waterfront.c     1     0.0145 888.04 -14199
## + floors.c:waterfront.c       1     0.0123 888.04 -14199
## + sqft_living.c:floors.c      1     0.0061 888.05 -14199
## - bathrooms.c:logsqft_lot.c   1     2.3245 890.38 -14198
## - bedrooms.c:floors.c         1     2.3316 890.39 -14198
## - bedrooms.c:yr_built.c       1     3.5808 891.63 -14188
## - sqft_living.c:bedrooms.c    1     4.3806 892.43 -14182
## - bedrooms.c:logsqft_lot.c    1     5.0386 893.09 -14177
## - logsqft_lot.c:yr_built.c    1     6.5259 894.58 -14166

```

```

## - logsqft_lot.c:floors.c      1      7.9653 896.02 -14154
##
## Step: AIC=-14211.49
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##   floors.c + yr_built.c + waterfront.c + sqft_living.c:bedrooms.c +
##   bathrooms.c:logsqft_lot.c + bathrooms.c:floors.c + bathrooms.c:yr_built.c +
##   bedrooms.c:logsqft_lot.c + bedrooms.c:floors.c + bedrooms.c:yr_built.c +
##   logsqft_lot.c:floors.c + logsqft_lot.c:yr_built.c + floors.c:yr_built.c
##
##
##           Df Sum of Sq    RSS    AIC
## - bathrooms.c:yr_built.c      1      0.8723 889.57 -14214
## <none>                          888.70 -14212
## - bathrooms.c:floors.c        1      1.4161 890.12 -14209
## - floors.c:yr_built.c         1      1.5710 890.27 -14208
## + logsqft_lot.c:waterfront.c   1      0.6780 888.02 -14208
## + yr_built.c:waterfront.c     1      0.6456 888.05 -14208
## + bathrooms.c:bedrooms.c      1      0.5819 888.12 -14207
## + sqft_living.c:yr_built.c    1      0.3595 888.34 -14206
## + sqft_living.c:bathrooms.c   1      0.3339 888.37 -14205
## + sqft_living.c:logsqft_lot.c 1      0.2137 888.49 -14204
## + floors.c:waterfront.c       1      0.0720 888.63 -14203
## + sqft_living.c:waterfront.c  1      0.0639 888.64 -14203
## + sqft_living.c:floors.c      1      0.0022 888.70 -14203
## + bathrooms.c:waterfront.c    1      0.0001 888.70 -14203
## + bedrooms.c:waterfront.c     1      0.0000 888.70 -14203
## - bedrooms.c:floors.c         1      2.3077 891.01 -14202
## - bathrooms.c:logsqft_lot.c   1      2.5025 891.20 -14201
## - bedrooms.c:yr_built.c       1      3.8123 892.51 -14190
## - sqft_living.c:bedrooms.c    1      4.3913 893.09 -14186
## - bedrooms.c:logsqft_lot.c   1      5.1917 893.89 -14180
## - logsqft_lot.c:yr_built.c    1      7.3342 896.03 -14163
## - logsqft_lot.c:floors.c      1      8.1742 896.87 -14156
## - waterfront.c                1     19.2291 907.93 -14071
##
## Step: AIC=-14213.5
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
##   floors.c + yr_built.c + waterfront.c + sqft_living.c:bedrooms.c +
##   bathrooms.c:logsqft_lot.c + bathrooms.c:floors.c + bedrooms.c:logsqft_
##   lot.c +
##   bedrooms.c:floors.c + bedrooms.c:yr_built.c + logsqft_lot.c:floors.c +
##   logsqft_lot.c:yr_built.c + floors.c:yr_built.c
##
##
##           Df Sum of Sq    RSS    AIC
## - bathrooms.c:floors.c        1      0.9383 890.51 -14215
## <none>                          889.57 -14214
## + bathrooms.c:yr_built.c      1      0.8723 888.70 -14212
## + yr_built.c:waterfront.c     1      0.7236 888.85 -14210

```

```

## + logsqft_lot.c:waterfront.c 1 0.7160 888.86 -14210
## + bathrooms.c:bedrooms.c 1 0.5007 889.07 -14209
## + sqft_living.c:logsqft_lot.c 1 0.1819 889.39 -14206
## - bedrooms.c:floors.c 1 2.1068 891.68 -14206
## - floors.c:yr_built.c 1 2.1495 891.72 -14206
## + sqft_living.c:bathrooms.c 1 0.0943 889.48 -14205
## - bathrooms.c:logsqft_lot.c 1 2.1852 891.76 -14205
## + floors.c:waterfront.c 1 0.0740 889.50 -14205
## + sqft_living.c:waterfront.c 1 0.0684 889.50 -14205
## + sqft_living.c:yr_built.c 1 0.0049 889.57 -14205
## + sqft_living.c:floors.c 1 0.0039 889.57 -14205
## + bedrooms.c:waterfront.c 1 0.0001 889.57 -14205
## + bathrooms.c:waterfront.c 1 0.0001 889.57 -14205
## - bedrooms.c:yr_built.c 1 2.9563 892.53 -14199
## - sqft_living.c:bedrooms.c 1 4.2604 893.83 -14189
## - bedrooms.c:logsqft_lot.c 1 4.8593 894.43 -14184
## - logsqft_lot.c:yr_built.c 1 7.7854 897.36 -14162
## - logsqft_lot.c:floors.c 1 8.1367 897.71 -14159
## - waterfront.c 1 18.9085 908.48 -14076
##
## Step: AIC=-14214.99
## log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c +
## floors.c + yr_built.c + waterfront.c + sqft_living.c:bedrooms.c +
## bathrooms.c:logsqft_lot.c + bedrooms.c:logsqft_lot.c + bedrooms.c:floors.c +
## bedrooms.c:yr_built.c + logsqft_lot.c:floors.c + logsqft_lot.c:yr_built.c +
## floors.c:yr_built.c
##
##
## Df Sum of Sq RSS AIC
## <none> 890.51 -14215
## + bathrooms.c:floors.c 1 0.9383 889.57 -14214
## - bedrooms.c:floors.c 1 1.3803 891.89 -14213
## - floors.c:yr_built.c 1 1.4402 891.95 -14213
## + logsqft_lot.c:waterfront.c 1 0.6607 889.85 -14211
## + yr_built.c:waterfront.c 1 0.6430 889.87 -14211
## + sqft_living.c:bathrooms.c 1 0.6006 889.91 -14211
## - bathrooms.c:logsqft_lot.c 1 1.7909 892.30 -14210
## + bathrooms.c:yr_built.c 1 0.3944 890.12 -14209
## + sqft_living.c:floors.c 1 0.3756 890.13 -14209
## + bathrooms.c:bedrooms.c 1 0.3512 890.16 -14209
## + sqft_living.c:logsqft_lot.c 1 0.2743 890.24 -14208
## + sqft_living.c:waterfront.c 1 0.0761 890.43 -14207
## + floors.c:waterfront.c 1 0.0341 890.48 -14206
## + sqft_living.c:yr_built.c 1 0.0065 890.50 -14206
## + bathrooms.c:waterfront.c 1 0.0043 890.51 -14206
## + bedrooms.c:waterfront.c 1 0.0002 890.51 -14206
## - bedrooms.c:yr_built.c 1 2.7584 893.27 -14202
## - bedrooms.c:logsqft_lot.c 1 4.8021 895.31 -14186
## - sqft_living.c:bedrooms.c 1 5.0065 895.52 -14185

```

```

## - logsqft_lot.c:yr_built.c      1    7.5487 898.06 -14165
## - logsqft_lot.c:floors.c        1    8.9491 899.46 -14154
## - waterfront.c                  1   18.6368 909.15 -14079

summary(step.bic)

##
## Call:
## lm(formula = log(price) ~ sqft_living.c + bathrooms.c + bedrooms.c +
##     logsqft_lot.c + floors.c + yr_built.c + waterfront.c + sqft_living.c:b
##     edrooms.c +
##     bathrooms.c:logsqft_lot.c + bedrooms.c:logsqft_lot.c + bedrooms.c:floo
##     rs.c +
##     bedrooms.c:yr_built.c + logsqft_lot.c:floors.c + logsqft_lot.c:yr_buil
##     t.c +
##     floors.c:yr_built.c, data = kc_house_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.39515 -0.25311  0.01195  0.24806  1.24770
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.304e+01  5.664e-03 2302.867 < 2e-16 ***
## sqft_living.c  4.388e-04  9.255e-06  47.418 < 2e-16 ***
## bathrooms.c    1.063e-01  1.019e-02  10.438 < 2e-16 ***
## bedrooms.c    -6.961e-02  6.361e-03 -10.944 < 2e-16 ***
## logsqft_lot.c -4.276e-02  6.158e-03  -6.944 4.14e-12 ***
## floors.c       6.274e-02  1.152e-02   5.445 5.34e-08 ***
## yr_built.c    -4.819e-03  2.041e-04 -23.607 < 2e-16 ***
## waterfront.c   6.333e-01  5.248e-02  12.068 < 2e-16 ***
## sqft_living.c:bedrooms.c -3.670e-05  5.867e-06  -6.255 4.21e-10 ***
## bathrooms.c:logsqft_lot.c -3.106e-02  8.303e-03  -3.741 0.000185 ***
## bedrooms.c:logsqft_lot.c  4.484e-02  7.319e-03   6.126 9.51e-10 ***
## bedrooms.c:floors.c    4.085e-02  1.244e-02   3.284 0.001027 **
## bedrooms.c:yr_built.c  -8.835e-04  1.903e-04  -4.643 3.50e-06 ***
## logsqft_lot.c:floors.c  -8.640e-02  1.033e-02  -8.363 < 2e-16 ***
## logsqft_lot.c:yr_built.c  1.905e-03  2.480e-04   7.681 1.80e-14 ***
## floors.c:yr_built.c     1.368e-03  4.078e-04   3.355 0.000799 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3577 on 6959 degrees of freedom
## Multiple R-squared:  0.5401, Adjusted R-squared:  0.5391
## F-statistic: 544.8 on 15 and 6959 DF,  p-value: < 2.2e-16

```

According to the BIC stepwise regression model summary above, we can see that once again all the original centered fields have been included in the updated model. Along with these fields, there are also eight other interactions between the original centered fields that are also included in the updated optimal model. Between these two models, the numbers

seem to indicate that the AIC stepwise regression model may be bit better than the BIC model. Between these two models, the AIC model has a higher adjusted R-Squared (54.07% vs. 53.91%). However, the AIC model also has a slightly larger residual standard error (0.3621 vs. 0.3577). Overall, these two models are very similar statistically to each other. Because the AIC model seemed to have a better fit for the model, we will use the AIC generated model as the final model.

Final Model

```
finalmodel = lm(log(price) ~ (sqft_living.c + bathrooms.c + bedrooms.c + logsqft_lot.c + floors.c + yr_built.c + waterfront.c+sqft_living.c*bathrooms.c + sqft_living.c*bedrooms.c+bathrooms.c*bedrooms.c+bathrooms.c*logsqft_lot.c+bathrooms.c*floors.c+bathrooms.c*yr_built.c+bedrooms.c*logsqft_lot.c+bedrooms.c*floors.c+bedrooms.c*yr_built.c+logsqft_lot.c*floors.c+logsqft_lot.c*yr_built.c+floors.c*yr_built.c+yr_built.c*waterfront.c+logsqft_lot.c*waterfront.c), data=kc_house_data)
summary(finalmodel)
```

```
##
## Call:
## lm(formula = log(price) ~ (sqft_living.c + bathrooms.c + bedrooms.c +
##   logsqft_lot.c + floors.c + yr_built.c + waterfront.c + sqft_living.c *
##   bathrooms.c + sqft_living.c * bedrooms.c + bathrooms.c *
##   bedrooms.c + bathrooms.c * logsqft_lot.c + bathrooms.c *
##   floors.c + bathrooms.c * yr_built.c + bedrooms.c * logsqft_lot.c +
##   bedrooms.c * floors.c + bedrooms.c * yr_built.c + logsqft_lot.c *
##   floors.c + logsqft_lot.c * yr_built.c + floors.c * yr_built.c +
##   yr_built.c * waterfront.c + logsqft_lot.c * waterfront.c),
##   data = kc_house_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40999 -0.25231  0.01182  0.24631  1.24384
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.305e+01  5.988e-03 2178.674 < 2e-16 ***
## sqft_living.c   4.492e-04  9.945e-06  45.168 < 2e-16 ***
## bathrooms.c     1.079e-01  1.049e-02  10.288 < 2e-16 ***
## bedrooms.c     -7.596e-02  6.524e-03 -11.642 < 2e-16 ***
## logsqft_lot.c  -4.065e-02  6.195e-03  -6.561 5.73e-11 ***
## floors.c        6.339e-02  1.176e-02   5.392 7.19e-08 ***
## yr_built.c     -4.862e-03  2.063e-04 -23.565 < 2e-16 ***
## waterfront.c    7.373e-01  6.397e-02  11.527 < 2e-16 ***
## sqft_living.c:bathrooms.c -1.656e-05  7.999e-06  -2.070 0.038527 *
## sqft_living.c:bedrooms.c  -4.283e-05  9.698e-06  -4.417 1.02e-05 ***
## bathrooms.c:bedrooms.c    2.608e-02  1.074e-02   2.428 0.015217 *
## bathrooms.c:logsqft_lot.c -2.755e-02  9.980e-03  -2.760 0.005787 **
## bathrooms.c:floors.c     -4.189e-02  1.845e-02  -2.270 0.023211 *
```

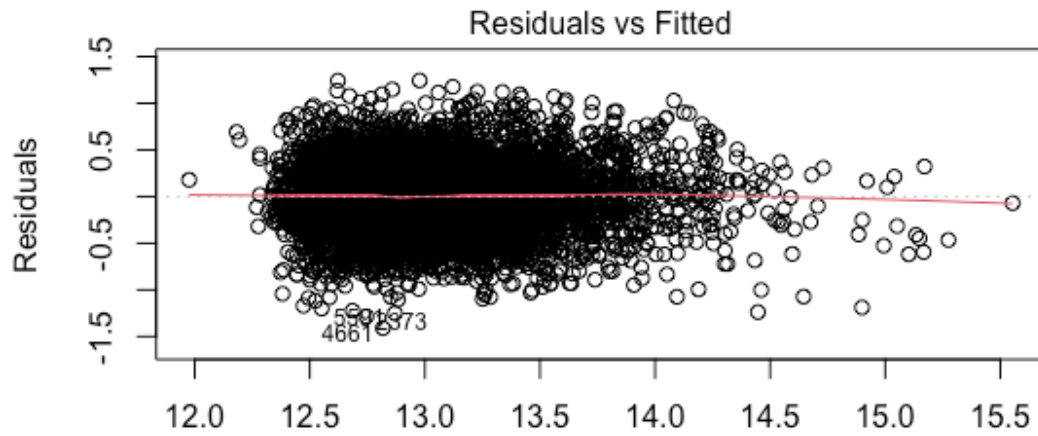
```

## bathrooms.c:yr_built.c      8.714e-04  2.836e-04   3.073 0.002127 **
## bedrooms.c:logsqft_lot.c   4.537e-02  7.717e-03   5.879 4.32e-09 ***
## bedrooms.c:floors.c       5.046e-02  1.375e-02   3.669 0.000246 ***
## bedrooms.c:yr_built.c     -1.283e-03  2.204e-04  -5.818 6.21e-09 ***
## logsqft_lot.c:floors.c    -8.202e-02  1.043e-02  -7.866 4.21e-15 ***
## logsqft_lot.c:yr_built.c   1.697e-03  2.575e-04   6.590 4.73e-11 ***
## floors.c:yr_built.c       1.383e-03  4.532e-04   3.052 0.002282 **
## yr_built.c:waterfront.c    3.109e-03  1.867e-03   1.665 0.095930 .
## logsqft_lot.c:waterfront.c -1.069e-01  6.553e-02  -1.631 0.102860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3571 on 6953 degrees of freedom
## Multiple R-squared:  0.5421, Adjusted R-squared:  0.5407
## F-statistic:   392 on 21 and 6953 DF,  p-value: < 2.2e-16

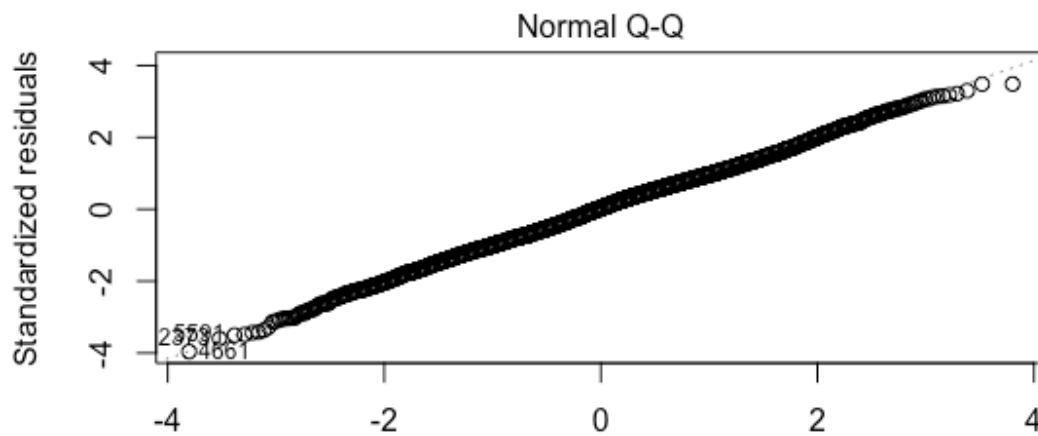
```

Now let's analyze the residual plots for the final model.

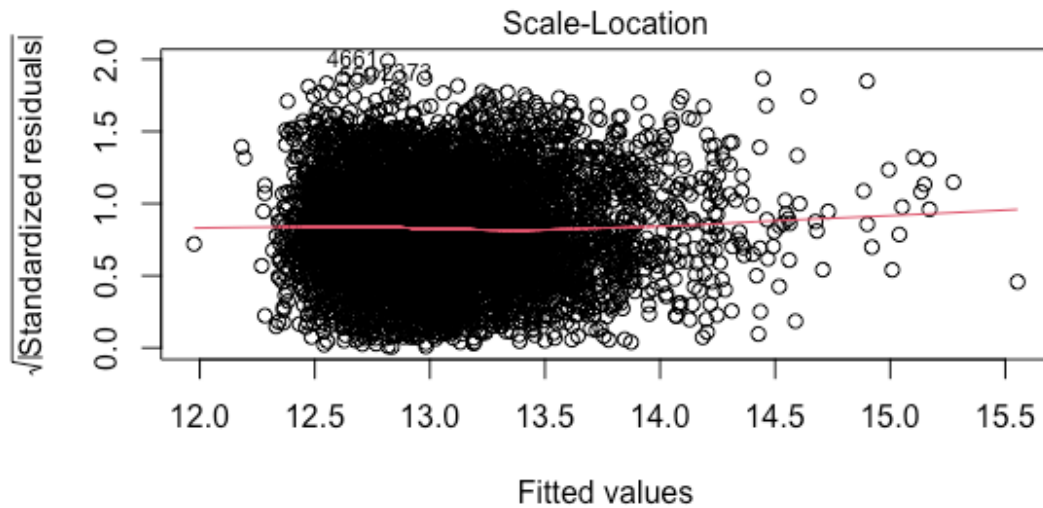
```
plot(finalmodel)
```

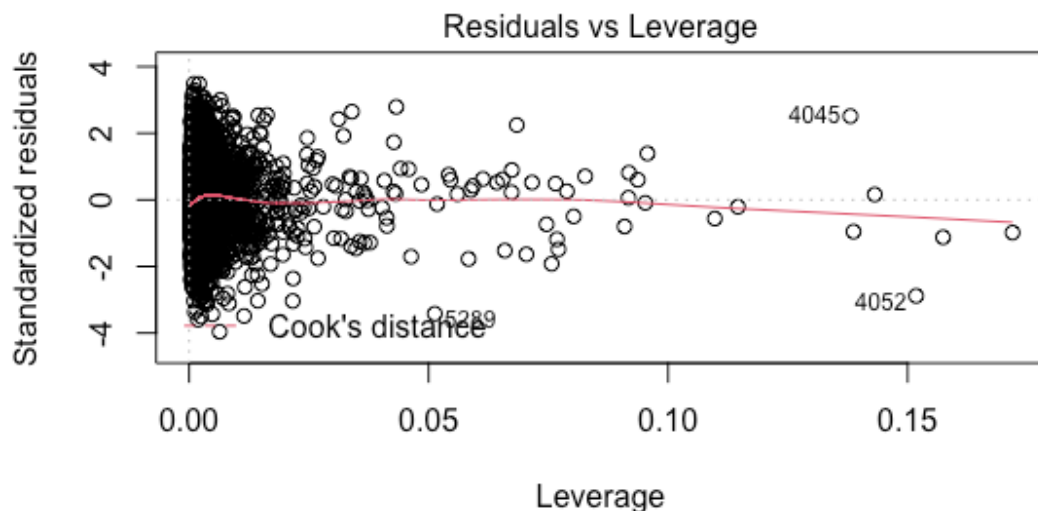
$\text{lm}(\log(\text{price}) \sim (\text{sqft_living.c} + \text{bathrooms.c} + \text{bedrooms.c} + \text{logsqft_lot.c} + \dots$



$\text{lm}(\log(\text{price}) \sim (\text{sqft_living.c} + \text{bathrooms.c} + \text{bedrooms.c} + \text{logsqft_lot.c} + \dots$



$\text{lm}(\log(\text{price}) \sim (\text{sqft_living.c} + \text{bathrooms.c} + \text{bedrooms.c} + \text{logsqft_lot.c} + \dots$

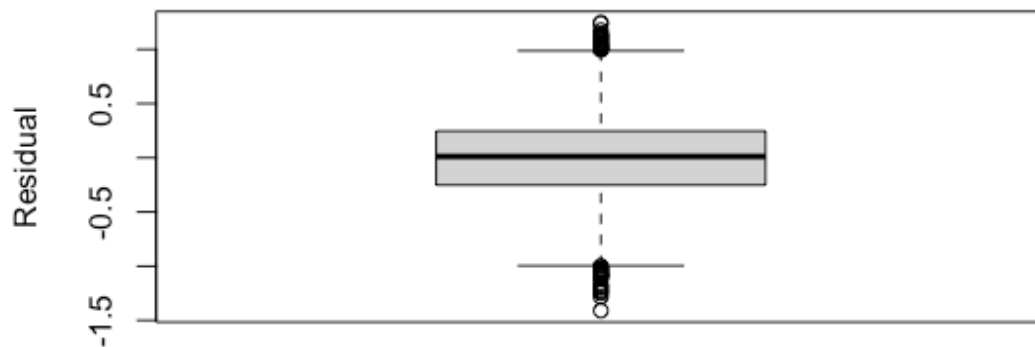


$\text{lm}(\log(\text{price}) \sim (\text{sqft_living.c} + \text{bathrooms.c} + \text{bedrooms.c} + \text{logsqft_lot.c} + \dots$

The residuals for the final model seem to be consistent with linearity, constant spread, and a normal distribution. By analyzing the residuals vs fitted plot, we see a plot with a best-fit line that seems to be both centered and a straight line with a slope of 0. There also is no obvious signs on non-constant variance in this plot. The normal quantile plot could not be too much straighter on the zero-line. This expresses that our final model seems to be normally distributed. There also doesn't seem to be any obvious irregularities within the Scale-Location or the Cook's distance plots. The Scale-Location plot has a best-fit line that is very straight down the middle of the scatter plot. And the Cook's distance plot doesn't show any flagged outliers. There are a few points that have more leverage than the others, but there are not any points that are flagged as influential to the model. Going deeper into the

Cook's Distance (4th residual plot), we can see that the majority of the spread of points is centered between 0 and 0.05. Though the majority of the points are around this leverage range, there are a few points that reach further to the right on the x-axis. These points that reach out further to the right, are more likely to have an influence on the slope of the best fit line. The plot highlights three points with the highest Cook's Distance: 4052, 4045, and 5289. However, I do not see any obvious need to remove these from the model. Though they have more leverage than the others, with the number of observations in the data set, I do not see them being too influential.

```
boxplot (finalmodel$residuals, ylab="Residual")
```



The box plot of the final model's residuals show a fairly centered mean around zero. There are a few points both above and below the box plot's "whiskers", however, there doesn't seem to be any that are noticeably high above the others. With as many observations as this data set had, we should expect a few right above and below the "whiskers" like we see in the plot above.

Now let's check the variance inflation factors from this final model.

```
library (car)
vif (finalmodel)

##          sqft_living.c          bathrooms.c
##          4.317512          3.521472
##          bedrooms.c          logsqft_lot.c
##          1.883130          1.691591
##          floors.c          yr_built.c
##          2.189022          1.975602
##          waterfront.c sqft_living.c:bathrooms.c
##          1.561280          4.571603
## sqft_living.c:bedrooms.c bathrooms.c:bedrooms.c
```

```
##          4.859479          4.292747
## bathrooms.c:logsqft_lot.c    bathrooms.c:floors.c
##          3.071051          2.439027
##    bathrooms.c:yr_built.c    bedrooms.c:logsqft_lot.c
##          2.044479          2.027091
##      bedrooms.c:floors.c    bedrooms.c:yr_built.c
##          2.306289          2.189312
##    logsqft_lot.c:floors.c    logsqft_lot.c:yr_built.c
##          2.142797          2.640999
##      floors.c:yr_built.c    yr_built.c:waterfront.c
##          2.190902          1.314158
## logsqft_lot.c:waterfront.c
##          1.734792
```

From the VIF analysis table above, we can see that each utilized field and interaction from this model seems reasonable. The standard cut-off for the VIF table is being above 5. As seen above, all of the utilized fields are below the cut-off, and we can claim that there is no multi-collinearity, or at least not enough to believe our model is invalid due to multi-collinearity.

Now let's go back and analyze the slope parameters.

```
summary(finalmodel)
```

```
##
## Call:
## lm(formula = log(price) ~ (sqft_living.c + bathrooms.c + bedrooms.c +
##    logsqft_lot.c + floors.c + yr_built.c + waterfront.c + sqft_living.c *
##    bathrooms.c + sqft_living.c * bedrooms.c + bathrooms.c *
##    bedrooms.c + bathrooms.c * logsqft_lot.c + bathrooms.c *
##    floors.c + bathrooms.c * yr_built.c + bedrooms.c * logsqft_lot.c +
##    bedrooms.c * floors.c + bedrooms.c * yr_built.c + logsqft_lot.c *
##    floors.c + logsqft_lot.c * yr_built.c + floors.c * yr_built.c +
##    yr_built.c * waterfront.c + logsqft_lot.c * waterfront.c),
##    data = kc_house_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40999 -0.25231  0.01182  0.24631  1.24384
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  1.305e+01  5.988e-03 2178.674 < 2e-16 ***
## sqft_living.c  4.492e-04  9.945e-06  45.168 < 2e-16 ***
## bathrooms.c    1.079e-01  1.049e-02  10.288 < 2e-16 ***
## bedrooms.c    -7.596e-02  6.524e-03 -11.642 < 2e-16 ***
## logsqft_lot.c -4.065e-02  6.195e-03  -6.561 5.73e-11 ***
## floors.c       6.339e-02  1.176e-02   5.392 7.19e-08 ***
## yr_built.c    -4.862e-03  2.063e-04 -23.565 < 2e-16 ***
## waterfront.c   7.373e-01  6.397e-02  11.527 < 2e-16 ***
```

```

## sqft_living.c:bathrooms.c -1.656e-05 7.999e-06 -2.070 0.038527 *
## sqft_living.c:bedrooms.c -4.283e-05 9.698e-06 -4.417 1.02e-05 ***
## bathrooms.c:bedrooms.c 2.608e-02 1.074e-02 2.428 0.015217 *
## bathrooms.c:logsqft_lot.c -2.755e-02 9.980e-03 -2.760 0.005787 **
## bathrooms.c:floors.c -4.189e-02 1.845e-02 -2.270 0.023211 *
## bathrooms.c:yr_built.c 8.714e-04 2.836e-04 3.073 0.002127 **
## bedrooms.c:logsqft_lot.c 4.537e-02 7.717e-03 5.879 4.32e-09 ***
## bedrooms.c:floors.c 5.046e-02 1.375e-02 3.669 0.000246 ***
## bedrooms.c:yr_built.c -1.283e-03 2.204e-04 -5.818 6.21e-09 ***
## logsqft_lot.c:floors.c -8.202e-02 1.043e-02 -7.866 4.21e-15 ***
## logsqft_lot.c:yr_built.c 1.697e-03 2.575e-04 6.590 4.73e-11 ***
## floors.c:yr_built.c 1.383e-03 4.532e-04 3.052 0.002282 **
## yr_built.c:waterfront.c 3.109e-03 1.867e-03 1.665 0.095930 .
## logsqft_lot.c:waterfront.c -1.069e-01 6.553e-02 -1.631 0.102860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3571 on 6953 degrees of freedom
## Multiple R-squared:  0.5421, Adjusted R-squared:  0.5407
## F-statistic: 392 on 21 and 6953 DF, p-value: < 2.2e-16

```

By analyzing the coefficients for each field, we can see that all of the original fields have p-values that are all significant at the 5% level, and there are only two interactions that are not significant at the 5% level. However, these still should stay in because the BIC stepwise regression model stated that these helped give us the optimal model. By looking at the t-values, we can see that square feet of the house (sqft_living.c) has the highest statistical effect on the price of the house. This seems reasonable because the more land a house house should have a big effect on how much a house costs. The next largest statistical effect on price is the year the house was built. This also seems reasonable as the newer houses may tend to be a bit more expensive than older houses. Now let's dig into the coefficient's stated effect on housing price per our model.

As the total house square footage increases by 1, the price of the house should increase by 4.492e-04 logged dollars.

As the bathrooms increases by 1, the price of the house should increase by 1.079e-01 logged dollars.

As the bedrooms increases by 1, the price of the house should decrease by -7.596e-02 logged dollars.

As the lot size of the house increases by 1, the price of the house should decrease by -4.065e-02 logged dollars.

As the floors in the house increases by 1, the price of the house should increase by 6.339e-02 logged dollars.

As the year the house was built increases by 1, the price of the house should decrease by -4.862e-03 logged dollars.

If the house is a waterfront property (by a body of water), the price of the house should increase by $7.373e-01$ logged dollars.

We will provide some example response predictions with confidence intervals

selected houses for prediction consist of two highly priced houses(>\$1,300,000), two middle priced houses(\$500,000

1 3750 2.25 4 8.52 2 1924 0 1.31e6

2 3320 3 5 8.59 2 2004 0 1.70e6

3 2100 2.5 3 8.71 2 2013 0 5.55e5

4 2300 2.5 3 8.37 2 1998 0 5.00e5

5 970 1 3 9.07 1 1962 0 1.70e5

6 800 1 2 9.15 1 1958 0 1.80e5

... with 3 more variables: pred.price , pred.lower , pred.upper

``` Among the six houses whose predictions are shown above, five of the homes have prediction intervals that contain the observed price.

```
kc_house_data$in.interval = ifelse (kc_house_data$pred.lower <= kc_house_data
$pred.price & kc_house_data$price <= kc_house_data$pred.upper,1,0)
```

```
mean(kc_house_data$in.interval)
```

```
[1] 0.9753405
```

Among all of the houses in the data set, 97.53% have prediction intervals that contain the observed price of the house.