# Redmond Housing Models

David Culhane

2024-05-02

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##
## Attaching package: 'pastecs'

## The following object is masked from 'package:magrittr':
##
##     extract

## The following objects are masked from 'package:dplyr':
##
##     first, last
```

## Making the Models

```
housing <- read_xlsx(path = "week-6-housing.xlsx")
```

The housing data has 24 variables and a total of 12,865 observations from home sales in the town of Redmond, WA. The data goes from 2006-2016, so major property events will follow in that span including the bubble burst of the Great Recession. It will be a good idea to select data within the same era for analysis. For that reason, I will use data from 2014-2016.

```
housing1416 <- housing %>% filter(`Sale Date` >= as.POSIXct("2014-01-01"))
```

This brings the total number of entries down from 12,865 to 3,646, but that should still be enough data for analysis.

### Model 1

It would not be out of the question to believe that the size of a property's lot has an effect on the price of the property. To test that, we can run a linear regression on the housing1416 data.

```
reg1 <- lm(`Sale Price` ~ sq_ft_lot, data = housing1416)
# Creates the linear model to model Sale Price using lot square footage
summary(reg1)
```

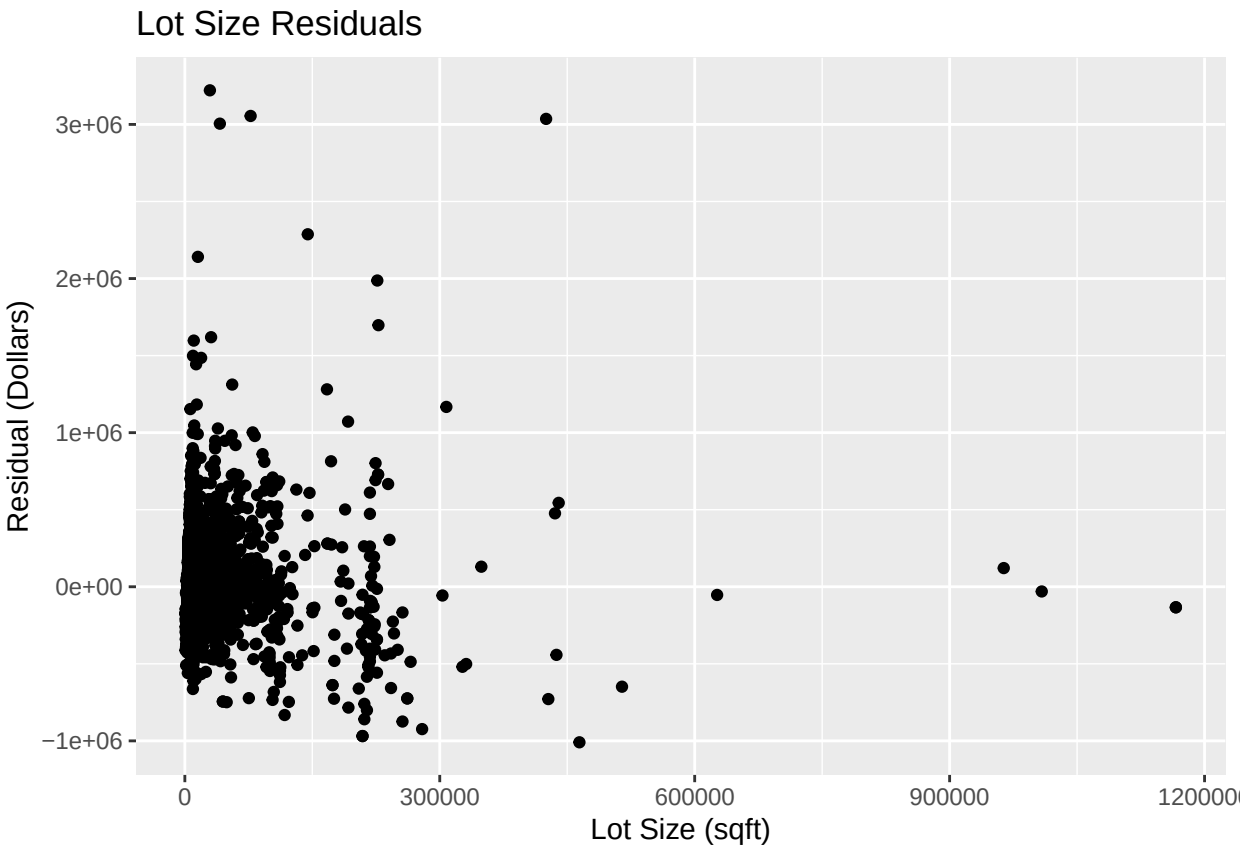
```
##
## Call:
## lm(formula = `Sale Price` ~ sq_ft_lot, data = housing1416)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1009211  -176944   -24386   130647  3221234
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.880e+05  5.177e+03  132.89   <2e-16 ***
## sq_ft_lot   1.381e+00  8.491e-02   16.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 288300 on 3644 degrees of freedom
## Multiple R-squared:  0.06766,    Adjusted R-squared:  0.0674
## F-statistic: 264.4 on 1 and 3644 DF,  p-value: < 2.2e-16
```

The summary of reg1's regression shows that this regression could be viewed as problematic. While the p-value t-test values for the coefficient and intercept appear to be significant, the given R-squared values are atrocious. The value of .06766 and .0674 would indicate that the total variance explained by lot square footage comes to explain only just under 7% of how sale price varies with lot size. The coefficient for the lot size comes to 1.381. This seems unrealistic given that what's being talked about is land. The plot for the line of best fit with this data doesn't look fantastic either.

```
housing1416$reg1Fits <- reg1$fitted.values
ggplot(housing1416, aes(x=sq_ft_lot,
                        y=`Sale Price`)
       ) + geom_point() + geom_line(aes(x=sq_ft_lot,
                                        y=reg1Fits)
                                    ) + labs(x="Lot Size (sqft)",
                                             y="Sale Price (Dollars)",
                                             title="Lot Size vs Sale Price Regression")
```


Now that we have the linear regression model, we can grab the residuals and store them and plot them.

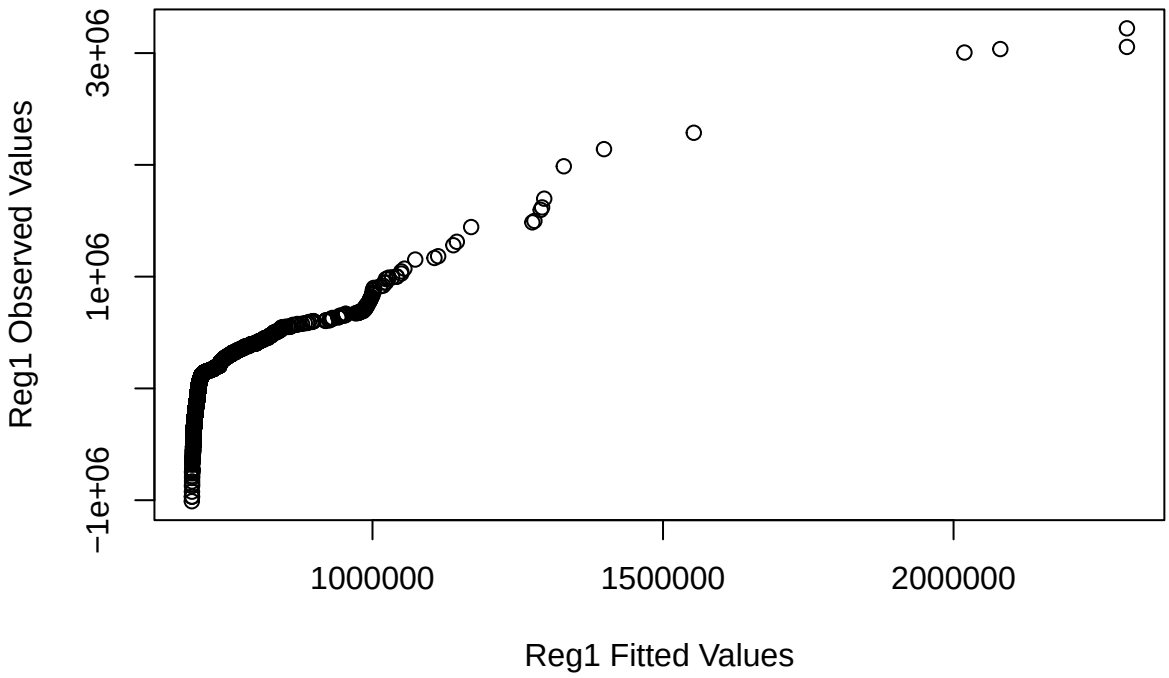
```
housing1416$reg1Residuals <- resid(reg1)
residual1Plot <- ggplot(housing1416, aes(x=sq_ft_lot, y=reg1Residuals))
residual1Plot + geom_point() + labs(x="Lot Size (sqft)",
                                    y="Residual (Dollars)",
                                    title="Lot Size Residuals")
```



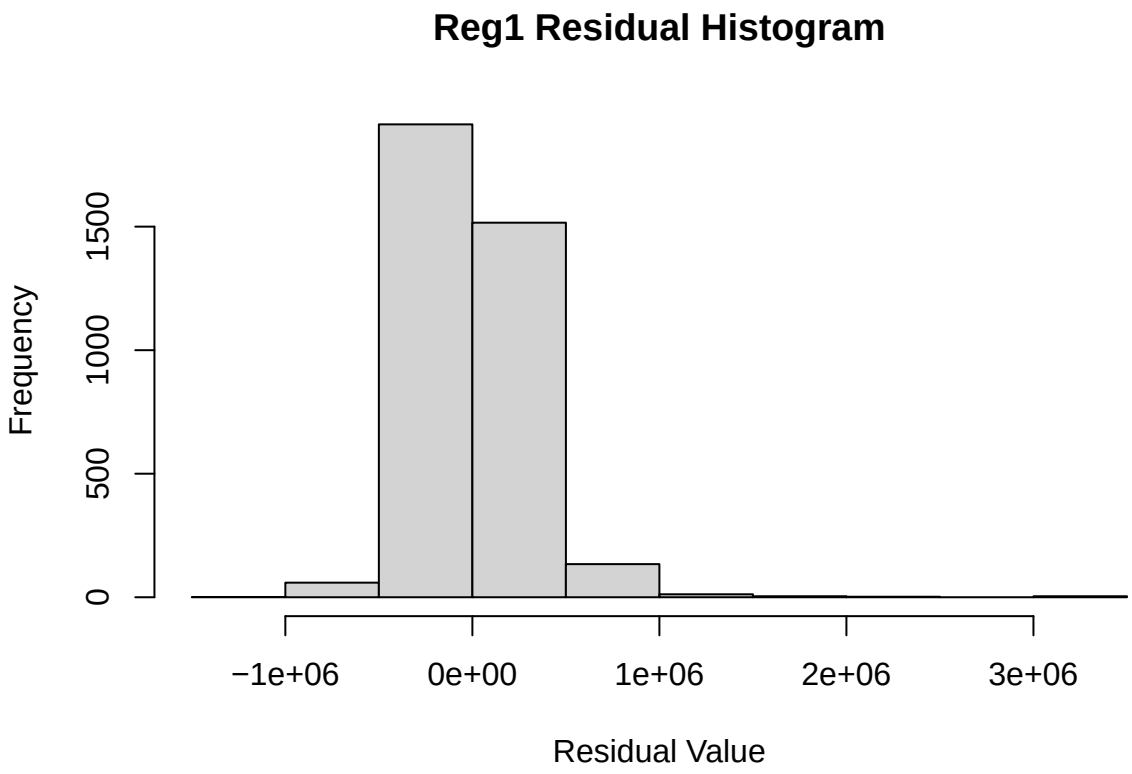

In this plot, the residual value is the difference between the model generated by the reg1 regression and the actual sale price of the property. The majority of the residuals appear to be between -$500,000 and $1,000,000 from the model and extends from lots ranging from 0 square feet to around 150,000 square feet. The points in the upper left corner of the graph seem to be much further away from the model, indicating that they are potentially outliers.

A Q-Q Plot can help visualize if the distribution of residuals is normal. If the distribution is normal, it will follow a linear pattern. Normally distributed residuals can also be seen as such in a histogram.

```
residual1_qqplot <- qqplot(housing1416$reg1Fits,
                           housing1416$reg1Residuals,
                           xlab="Reg1 Fitted Values",
                           ylab="Reg1 Observed Values")
```




```
hist(housing1416$reg1Residuals,
     main = "Reg1 Residual Histogram",
     xlab = "Residual Value")
```




The Q-Q Plot's distribution resembles a logarithmic curve instead of a linear curve here, so I would assume that this set is not linearly distributed. To be sure, I also plotted a histogram of the residuals. While the residuals appear to have a single mode, it could also be argued that the data might be slightly skewed to the left A brief look at the norm statistics from pastecs' stat.desc function can illuminate that.

```
stat.desc(housing1416$reg1Residuals, norm=TRUE, basic=FALSE)
```

```
##        median          mean       SE.mean  CI.mean.0.95           var
## -2.438594e+04 -2.620192e-11  4.774379e+03  9.360718e+03  8.310944e+10
##       std.dev      coef.var      skewness      skew.2SE      kurtosis
##  2.882871e+05 -1.100252e+16  2.333372e+00  2.877169e+01  1.829943e+01
##      kurt.2SE    normtest.W    normtest.p
##  1.128516e+02  8.720643e-01  1.140288e-47
```

The residuals have a skew value of around 2.3, which could be acceptable since that's not far from zero. However, the kurtosis is around 18.3, which is definitely not normal. The distribution is positively pointy.

### Model 2

To find a better model, we can create a multiple linear regression model. The original model only incorporated the square footage of the lot. Other factors that should be included should be the number of bedrooms, bathrooms, square footage in the home itself. The bathroom data is separated out by the number of types of bathrooms. For ease of use, this can be consolidated into a new single variable.

```
housing1416$bath_total = housing1416$bath_full_count + (0.5 * housing1416$bath_half_count) + (0.75 * hou
```

Now that the bathroom totals have been added as a single variable, we can use it for the multiple linear regression. The most valuable variables, in my opinion, would follow in the order of:

- Lot size
- Living space size
- Number of bedrooms
- Number of bathrooms

The numbers of bedrooms and bathrooms go hand-in-hand, so these should be treated as if they were interacting together.

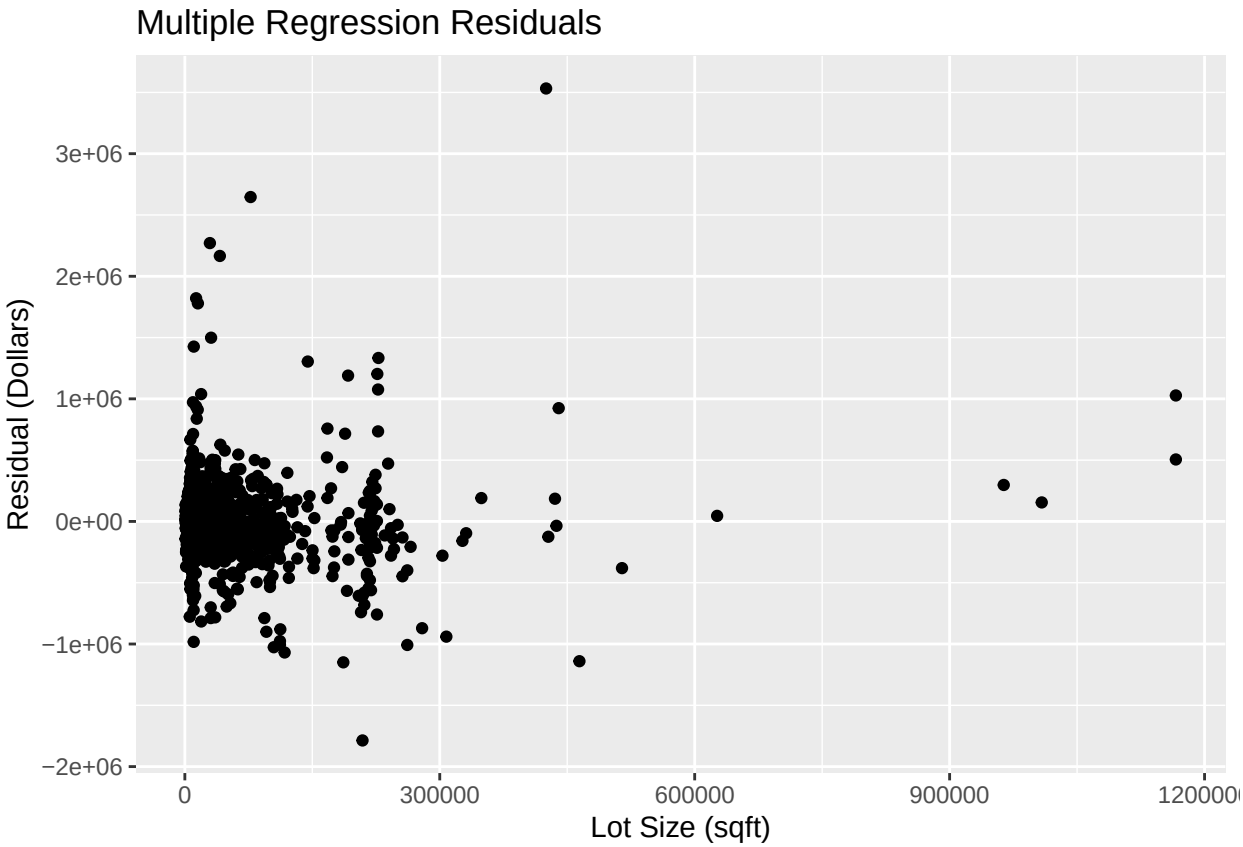
```
reg2 <- lm(`Sale Price` ~ sq_ft_lot + square_feet_total_living + bedrooms:bath_total,
           data = housing1416)
summary(reg2)
```

```
##
## Call:
## lm(formula = `Sale Price` ~ sq_ft_lot + square_feet_total_living +
##     bedrooms:bath_total, data = housing1416)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1786793   -77450    -5352    76035  3531982
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.672e+05  9.436e+03  17.716   <2e-16 ***
## sq_ft_lot                  6.098e-01  6.232e-02   9.785   <2e-16 ***
## square_feet_total_living 2.103e+02  5.593e+00  37.592   <2e-16 ***
## bedrooms:bath_total      3.265e+02  1.299e+03   0.251    0.802
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 204200 on 3642 degrees of freedom
## Multiple R-squared:  0.5325, Adjusted R-squared:  0.5322
## F-statistic:  1383 on 3 and 3642 DF,  p-value: < 2.2e-16
```

This model seems much stronger than the previous one. For starters, the multiple R-squared value is 0.5325 and the adjusted value is 0.5322. These are fairly close together and much better than the previous model's 0.067s. The coefficients are all positive, indicating increases in value when one is increased and the others stay the same, which also makes good sense. The t-value and its probability are significant for the size of the lot and livable square footage, but aren't that great for the combination/interaction between bedrooms and total number of bathrooms. That may have something to do with the total for their interaction in each observation being MUCH smaller than the square footage numbers.

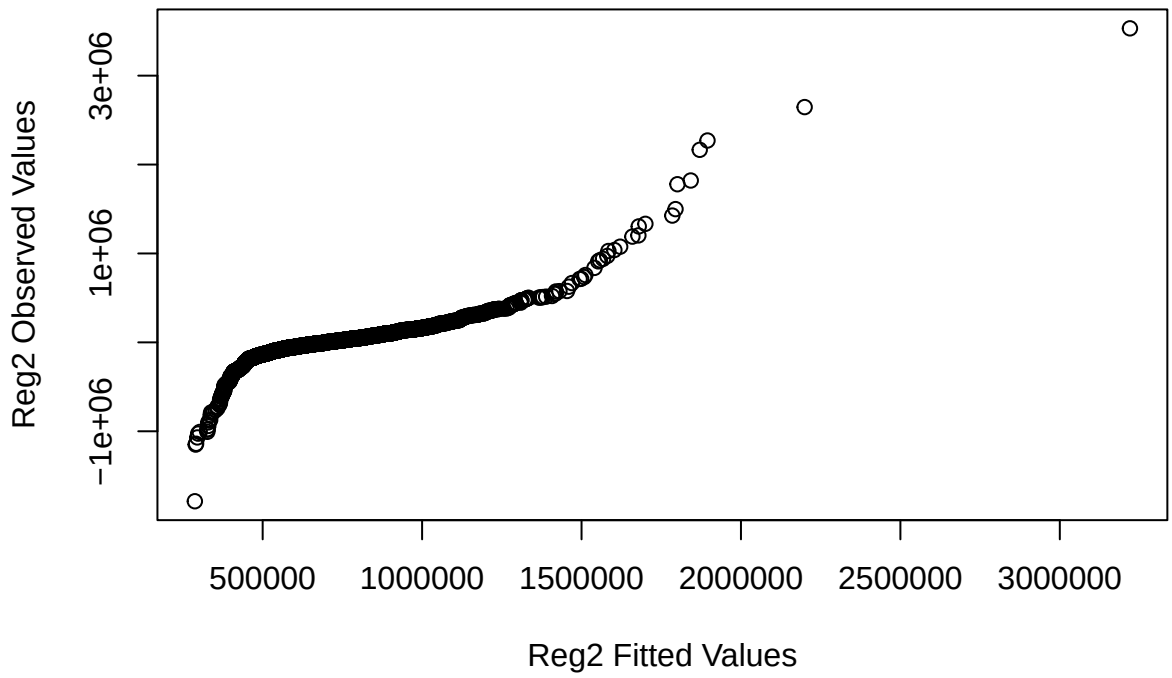
Now that the model has been created, its values and residuals can be plotted and stored. That will make plotting the residuals easier. Four variables were used as predictors, so creating a plot using all four to predict Sale Price would be difficult to make. So to plot the residuals, I will use the size of the lot as the horizontal axis to make comparisons easier.

```
housing1416$reg2Fits <- reg2$fitted.values
housing1416$reg2Residuals <- resid(reg2)
residual2Plot <-ggplot(housing1416, aes(x=sq_ft_lot, y=reg2Residuals))
residual2Plot + geom_point() + labs(x="Lot Size (sqft)", y="Residual (Dollars)", title="Multiple Regress
```

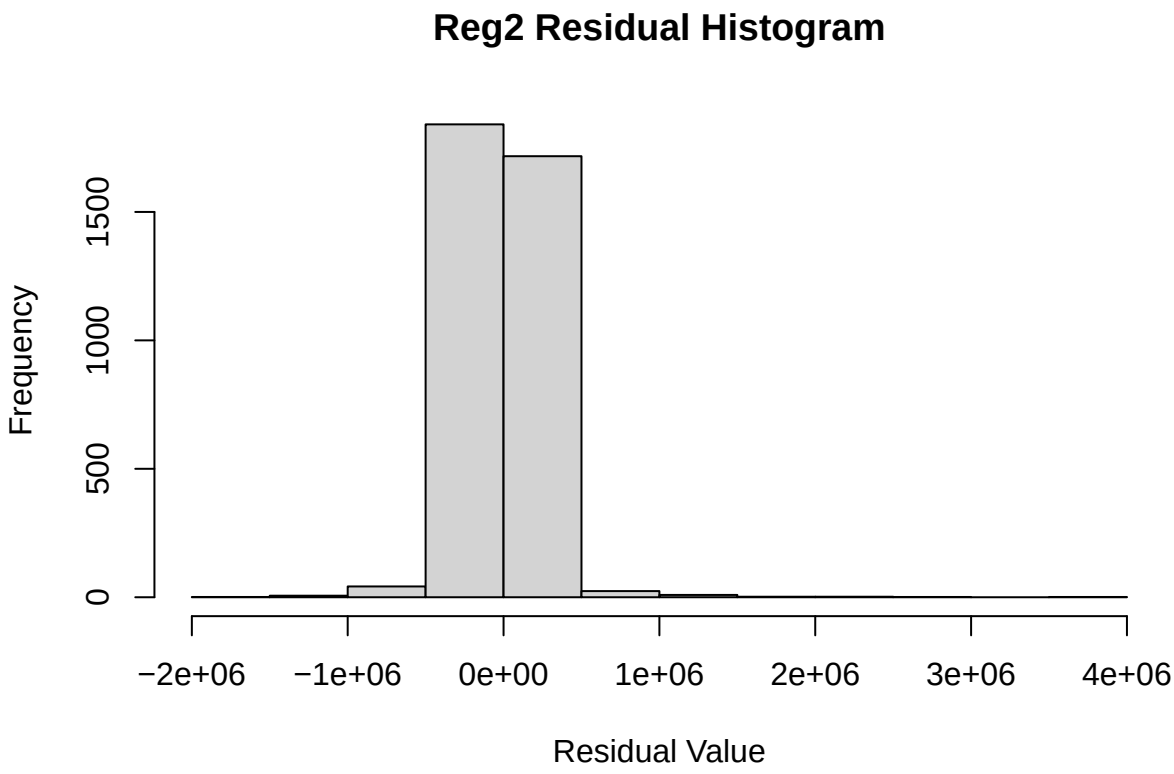



When comparing the residual plots between the two models, we can see with reg2's residuals that the cloud condenses slightly more towards 0 and the furthest points away, vertically, come down quite a bit as well. The residuals off to the right do change somewhat, and another appears.

```
residual1_qqplot <- qqplot(housing1416$reg2Fits,
                           housing1416$reg2Residuals,
                           xlab="Reg2 Fitted Values",
                           ylab="Reg2 Observed Values")
```




```
hist(housing1416$reg2Residuals,
     main = "Reg2 Residual Histogram",
     xlab = "Residual Value")
```




```
stat.desc(housing1416$reg2Residuals, basic=FALSE, norm=TRUE)
```

```
##        median          mean       SE.mean  CI.mean.0.95           var
## -5.352300e+03 -1.019932e-10  3.380674e+03  6.628199e+03  4.166996e+10
##       std.dev      coef.var      skewness      skew.2SE      kurtosis
##  2.041322e+05 -2.001430e+15  3.068194e+00  3.783242e+01  5.021058e+01
##      kurt.2SE    normtest.W    normtest.p
##  3.096459e+02  7.371210e-01  3.164879e-60
```

The data from reg2 would also not appear to be normal. This could probably be taken care of by using standardized versions of each variable. If that were done, then each unit up or down for each variable would indicate an increase in that variable by a standard deviation instead of a dollar, square foot, or bedroom-bathroom. The former two of those units are orders of magnitude larger than the latter of those. That likely significantly influences the kurtosis of the histogram being used to assess normality of the residuals.

The two models, reg1 and reg2, can be compared using ANOVA as well.

```
anova(reg1, reg2)
```

```
## Analysis of Variance Table
##
## Model 1: `Sale Price` ~ sq_ft_lot
## Model 2: `Sale Price` ~ sq_ft_lot + square_feet_total_living + bedrooms:bath_total
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   3644 3.0293e+14
```

```
## 2   3642 1.5189e+14  2 1.5105e+14 1810.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The size of the F-statistic, 1810.9, combined with its small probability shows a significant improvement when going from reg1 to reg2 when using it to predict a property's sale prince.

## Thoughts on the Analyses

For any model to be biased, it means that there are factors that cause systemic error baked into the model. For these analyses, I would imagine that the nature of most of the variables being used would cause the model to be skewed. The units being used, whole dollars or square feet instead of thousands of dollars or hundreds of square feet, could end up causing the significance of the coefficients to skew towards the lot size and living space size.

Reg1 follows that, in my opinion. That model only took the square footage of the lot into account. Its coefficients came out to 688,041.40 and around 1.381. These results paint an outlandish picture of Redmond, Washington. The intercept of this regression is the value of a property on a lot with 0 square feet of space. That would be the equivalent of either renting or buying a box under a bridge. $688,041.40 seems like an outlandishly steep cost for just that. The 1.381 slope term multiplied by the value of the lot's square footage says that for each square foot of lot the cost of the property should increase by $1.38. The R-squared value of this model came to around .06, further illustrating the low quality of the model. Reg1's skew and kurtosis of its residuals were previously discussed. The kurtosis of this model's residuals is what makes this non-normal distribution apparent.

Reg2's model has more sensible coefficient values. The intercept value was 167,173.50, the lot square footage coefficient was 0.6098, the home square footage coefficient was 210.25, and the coefficient for the bedroom/bathroom total interaction (only their product since : was used instead of *) was 326.54. These values seem much more sensible than those of Reg1's model. First, the intercept is much lower at $167,173.50 to buy a property on a 0 square foot lot with no living space, bedrooms, or bathrooms. Not a terrible starting cost for Redmond, Washington. The living space's coefficent of $210.25 seems like it might be a bit high, (I'm basing that opinion off the cost per square foot of my own home, a condo in NC, of $126.81 that was purchased 6 years later) but Redmond is also a suburb of Seattle and home to companies like Microsoft and Amazon. Each square foot of lot space costs a hair under 61 cents. The bedroom-bathroom coefficient was $210.25. The concept of a bedroom-bathroom seems odd, but the product of the two means that each additional bedroom and/or bathroom increases the value of the home almost exponentially. For example, a 2 bed, 2 bath home would have a product of 4 times the $210.25 coefficient while a 3 bed, 3 bath home would be 9 times the coefficient. That makes a significant difference when comparing properties of different types.

Reg2's adjusted R-squared value of 0.5322 is much higher than reg1's .06. I think reg1 suffers from bias due to the large units, evident in its kurtosis. Reg2 suffers from the same effect as well, but additional terms have mitigated that effect somewhat. Re-doing the analyses using smaller units, like thousands of dollars or hundreds of square feet, or standardized units, in terms of standard deviations, might hopefully address that problem and tame the kurtosis issue.

## RMSEs of the Models and Metrics

We can check the accuracies of each model by generating data using predict and evaluating RMSE with the Metrics libary's rmse() function.

```
reg1_predictions <- predict(object=reg1, newdata=housing1416)
reg2_predictions <- predict(object=reg2, newdata=housing1416)
```

```r
reg1_RMSE <- rmse(housing1416$`Sale Price`, reg1_predictions)
reg2_RMSE <- rmse(housing1416$`Sale Price`, reg2_predictions)
paste("Reg1's RMSE comes to", round(reg1_RMSE, digits = 3))
```

```
## [1] "Reg1's RMSE comes to 288247.548"
```

```r
paste("Reg2's RMSE comes to", round(reg2_RMSE, digits = 3))
```

```
## [1] "Reg2's RMSE comes to 204104.226"
```

We can see that Reg2's model lowered RMSE by around $80,000. That's a significant chunk of change in any market and further solidifies reg2 as a better model than reg1.