Bountiful Home Prices

Parker Holzer

3/24/2021

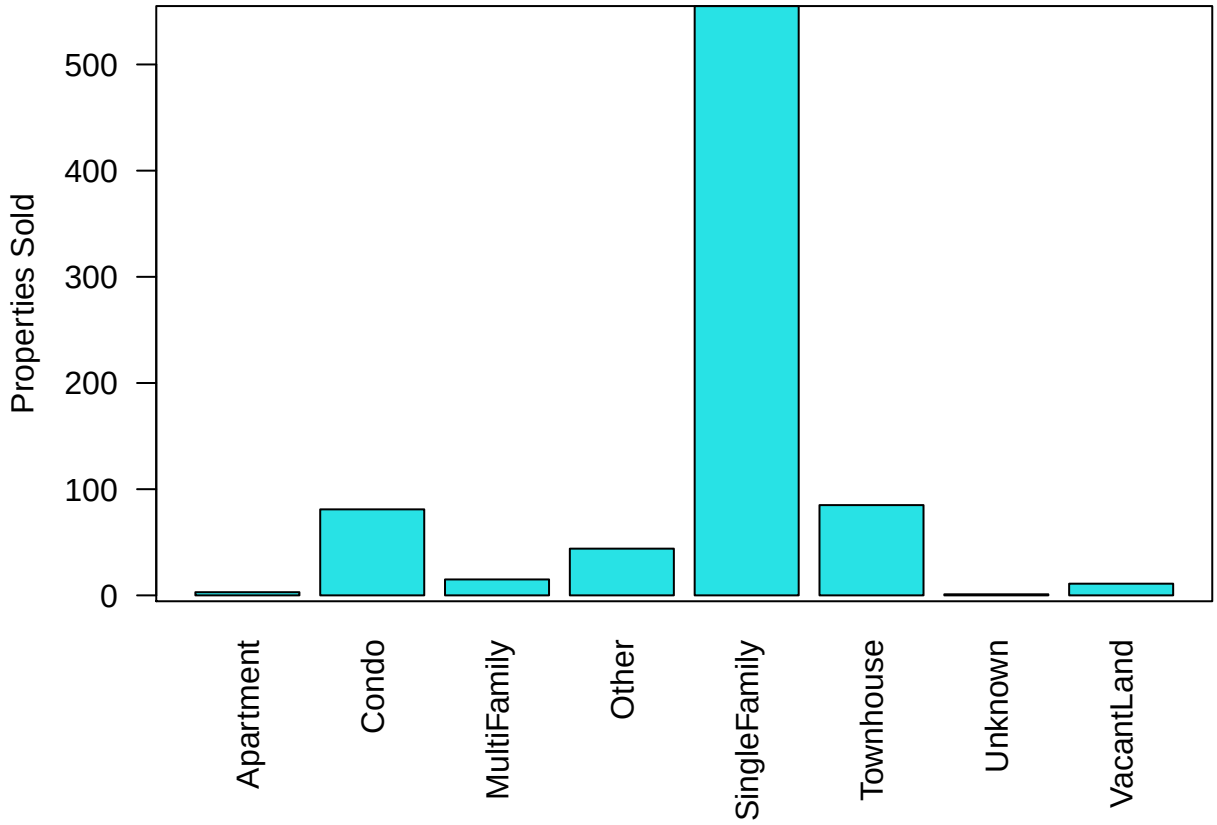
Introduction

What factors influence the selling price of houses and in what way? What houses are reasonably priced and which ones are likely overpriced? Questions like these are valuable for those of us (yes including me) who are looking into buying a house one day and want to get the inside scoop from data science.

This data set comes from scraping Zillow.com on March 25, 2021. In particular, this is a set of properties sold in Bountiful, Utah in the months leading up to March 2021. So, let's start by introducing the data!

Property Type

```
df = read.csv("Bountiful_UT_3-25-2021.csv")
df$Type[is.na(df$Type)] = "Other"
tbl = table(df$Type)
par0 = par()
par(mar=c(6,4,1,1))
barplot(tbl, las=2, col = 5, ylab = "Properties Sold")
box()
```



So the majority of the 795 properties sold were SingleFamily, which is the type we're interested in here. From this point on we don't analyze the other types.

Next we need to clean up the data a bit and take a look at each of the variables to see what we have to work with.

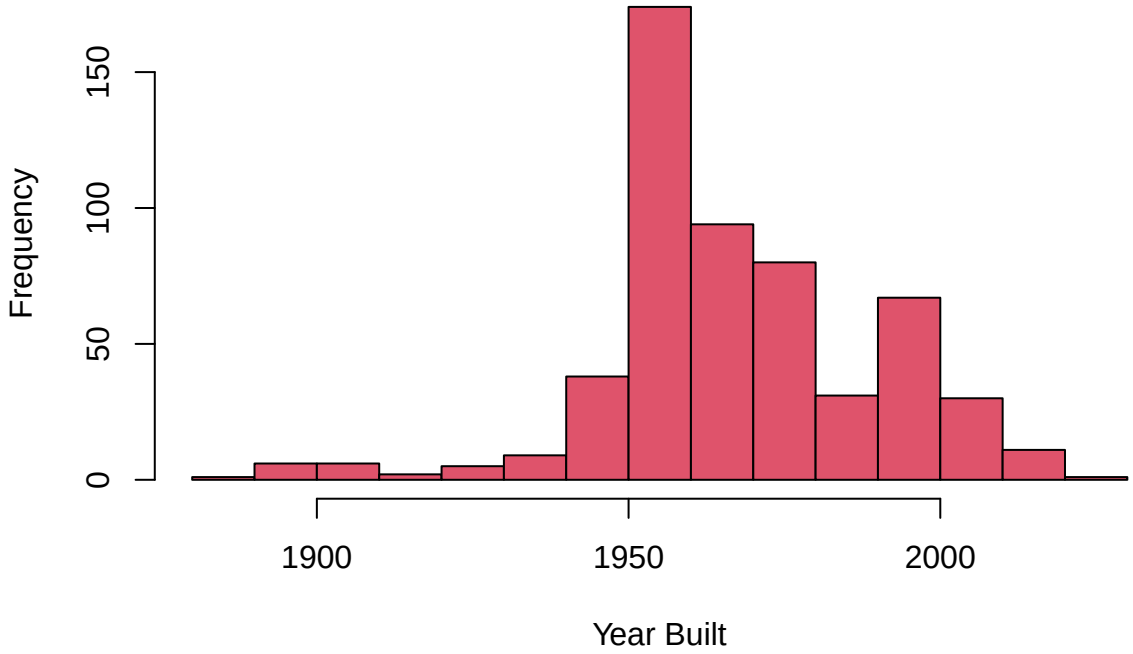
```
myfunc1 = function(l){
  if(grepl("Acres", l)){
    return(as.numeric(strsplit(l, " ")[[1]][1]))
  }
  if(grepl("sqft", l)){
    sqft = as.numeric(gsub(",", "",strsplit(l, " ")[[1]][1]))
    return(sqft/43560)
  }
  if(is.na(l)){return(NA)}
}
df$Lot = sapply(df$Lot, myfunc1)
df$Area = as.numeric(gsub(",", "", df$Area))
df$Cost = as.numeric(gsub("\\$", "", gsub(",", "", df$Cost)))

houses = df[df$Type == "SingleFamily",2:7]
houses$Lot = as.numeric(houses$Lot)
names(houses)
```

```
## [1] "Built" "Lot"   "Bed"   "Bath"  "Area"  "Cost"
```

Year Built

The variable *Built* is the year the house was built. Although Bountiful, UT traces its roots back to the mid 1800s, most of the existing properties have been built within the last 50 years. Let's look at a histogram to get a more precise idea of this variable in the data.

```
hist(houses$Built[!is.na(houses$Built)], col=2,
     xlab = "Year Built", main = "")
```

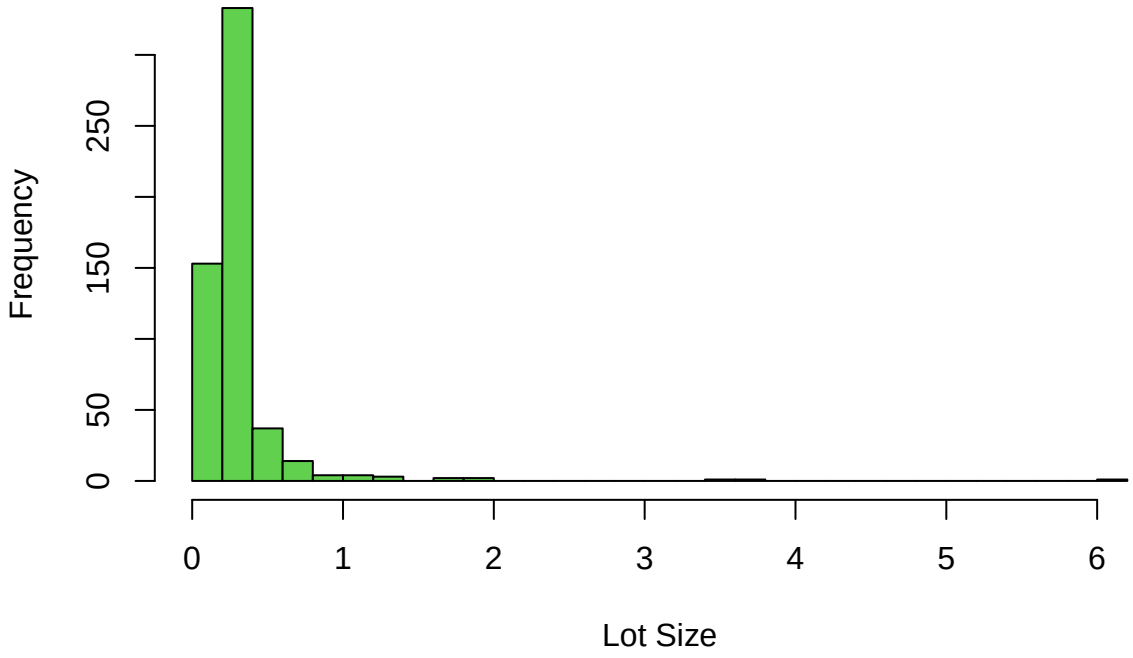


Interesting! There are 12 of these sold properties that were built before 1910 with the oldest being built in 1889. Most were built in the later half of the 1900's. And there is a good chunk of them that are very new. From the looks of the histogram, there are no obvious outliers though. Definitely a skew to the left that we might want to keep in mind. Also, as a sidenote, there were no houses that did not have a recorded year built in this data set.

Lot Size

The variable *Lot* is the lot size of the property in units of acres. Not all properties had the lot size in units of acres, so part of the data cleaning that was done earlier was converting the occasional squarefeet lot size to acres. What does the distribution of lot sizes look like?

```
hist(houses$Lot, breaks = 30, col=3,
     xlab = "Lot Size", main = "")
```



So most of the lot sizes are less than 0.5 acres, with a couple very strong outliers above 3 acres. Let's take a quick look at those outliers.

```
houses[houses$Lot > 3,]
```

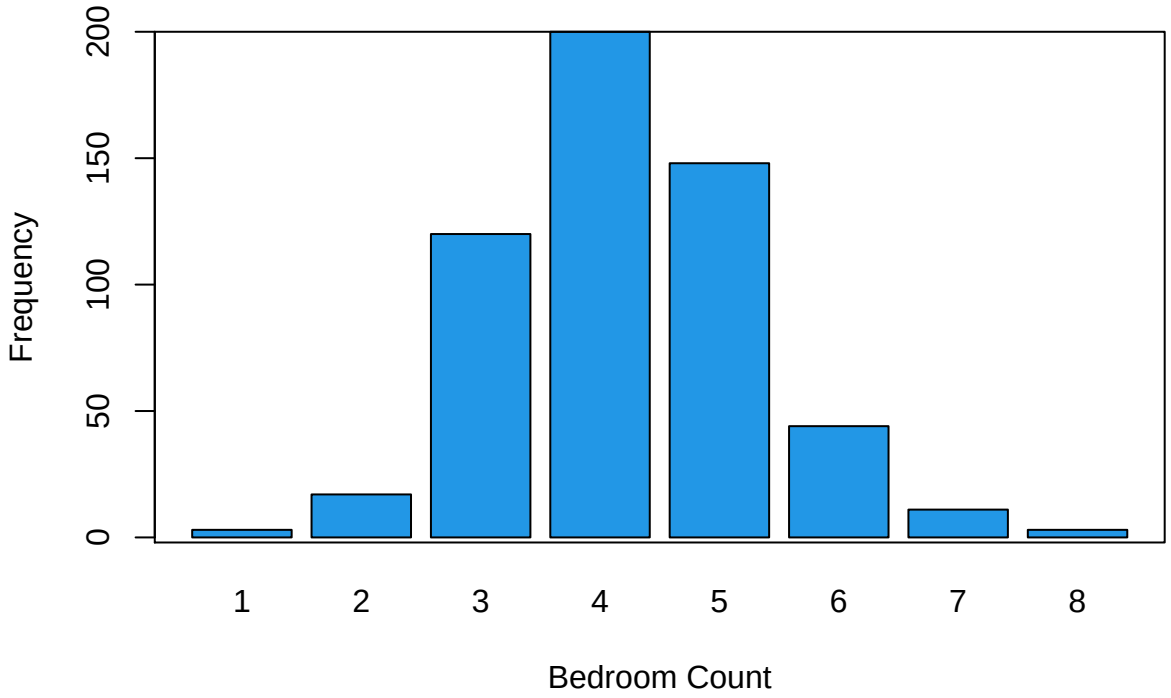
```
##     Built  Lot Bed Bath Area    Cost
## 102  1980 3.79   4    6 6887 1878196
## 266  1996 6.13   4    5 6741 1079659
## 578  2007 3.55   7    5 6705 1566829
```

So apparently three mansions were sold! Interesting, but probably not the typical house we are thinking of building our model around. Let's set them aside for now.

```
houses = houses[houses$Lot <= 3,]
```

Bedroom count

```
barplot(table(houses$Bed), col=4,
        xlab = "Bedroom Count", ylab = "Frequency")
box()
```



Nothing particularly unusual stands out about the bedroom counts. And it looks suprisingly bell-shaped! Also, there were six houses with missing values for the bedroom count:

```
houses[is.na(houses$Bed),]
```

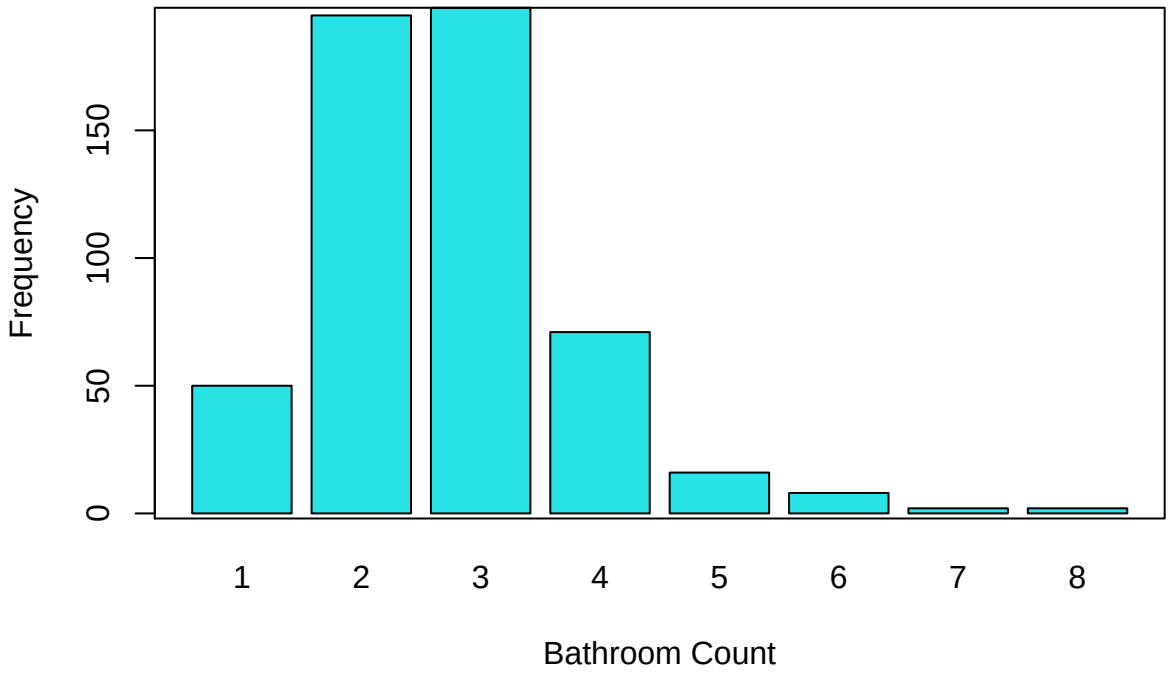
```
##     Built       Lot Bed Bath Area   Cost
## 316  1970 0.4100000  NA   NA 5760     NA
## 317  1969 0.1799816  NA   NA 1212     NA
## 318  1961 0.2600000  NA   NA 2557     NA
## 609  1997 1.7000000  NA   NA 2705 364532
## 649  2021 0.0500000  NA   NA 5614     NA
## 771  1959 0.2500000  NA   NA 1170     NA
```

Since these six also have missing values for Bathroom count, and all but one have a missing cost, we will drop these at this point.

```
houses = houses[!is.na(houses$Bed),]
```

Bathroom count

```
barplot(table(houses$Bath), col=5,
        xlab = "Bathroom Count", ylab = "Frequency")
box()
```



Same story for the bathroom counts. Except that there is a little skew right. I've personally never lived in a place that had 8 bathrooms, but I guess they exist! Four additional houses had missing values for the bathroom count:

```
houses[is.na(houses$Bath),]
```

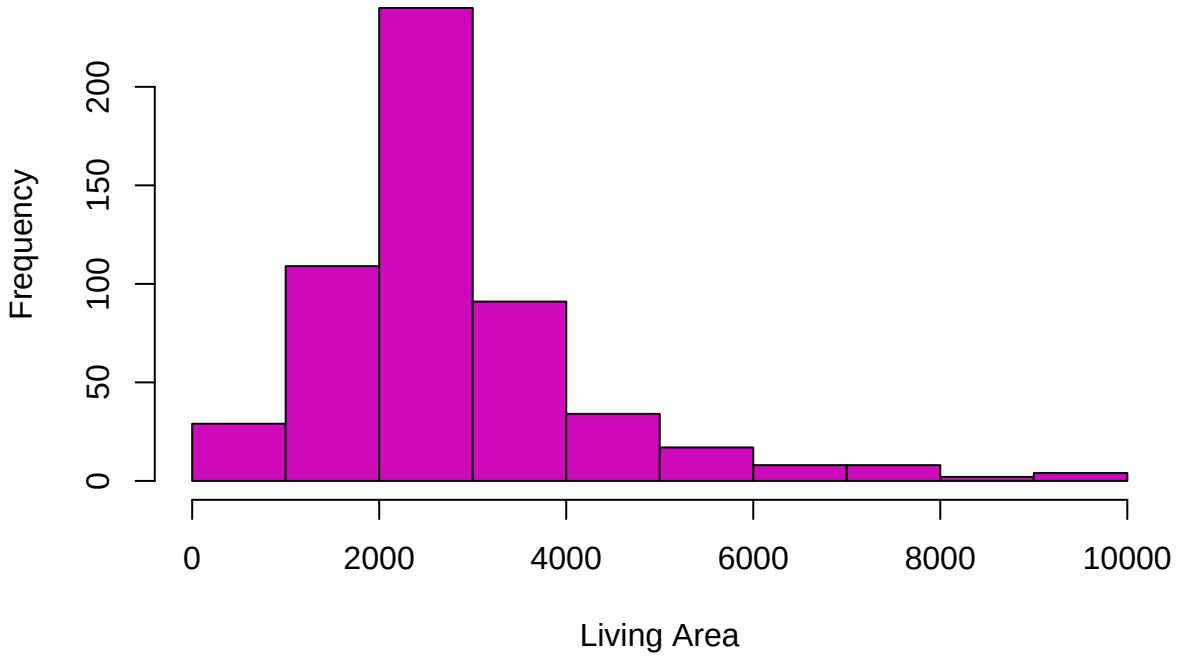
```
##     Built        Lot Bed Bath Area   Cost
## 231  1980 0.25000000   4   NA    2 532120
## 428  1997 0.20000000   5   NA 2300 502647
## 590  1963 0.25000000   5   NA 2200 382594
## 598  1938 0.01999541   3   NA  938 270578
```

These are all rather small properties, so perhaps they have no bathroom. But just for sanity's sake, we won't impute the missing values with 0 here. Since we would like to include the bathroom count in our model, we will also drop these four cases. The first one looks awfully suspicious with an area of 2 square feet anyway!

```
houses = houses[!is.na(houses$Bath),]
```

Area

```
hist(houses$Area, col=6,
     xlab = "Living Area", main="")
```



Obviously a strong skew off the the right, but no extreme outliers of high area. Are there any properties left that were like the suspicious case above with only 2 square feet of area?

```
min(houses$Area)
```

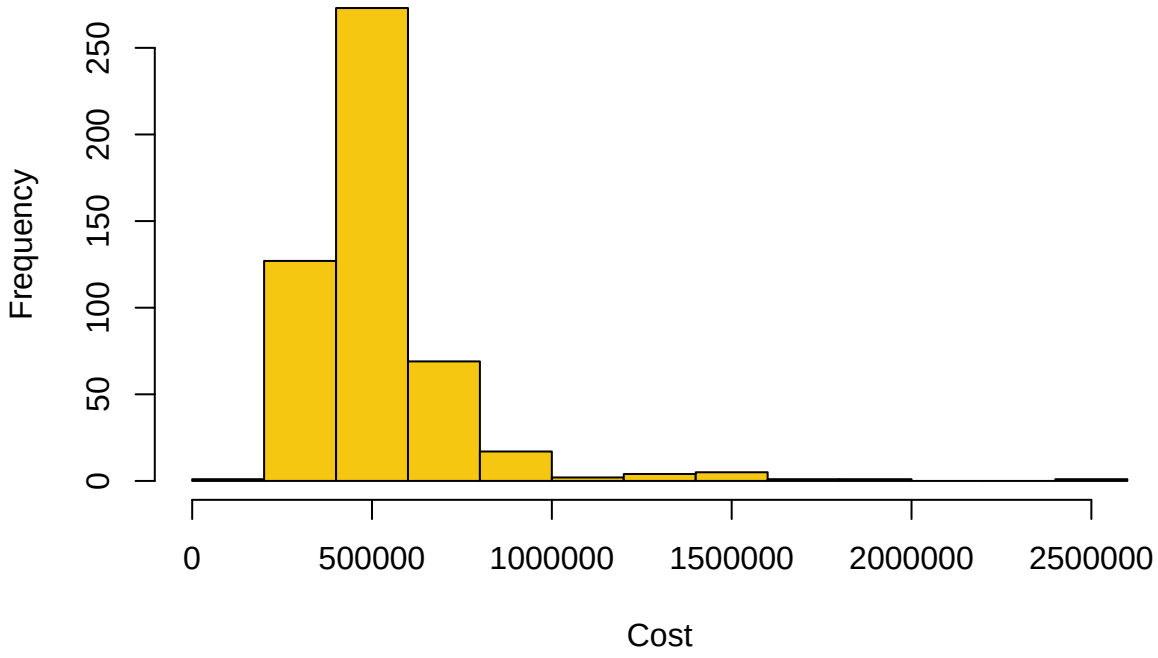
```
## [1] 130
```

It doesn't look like there are any suspiciously low cases left. And no missing values for living area are left either.

Cost

There are 41 cases that have missing values for the Cost. Since we are using that as the response variable, we will drop these cases for building our model.

```
houses = houses[!is.na(houses$Cost),]
row.names(houses) = 1:dim(houses)[1]
hist(houses$Cost, col=7,
     xlab = "Cost", main="")
```



There are some very high cost properties in this dataset, but most seem to center around 500,000 dollars.

Linear Model

To start let's see what we get with a linear model of the cost as a linear combination of all other explanatory variables. This will help to answer the question of how each of these variables influences the final cost.

```
mdl = lm(Cost ~ Built + Lot + Bed + Bath + Area, data=houses)
summary(mdl)
```

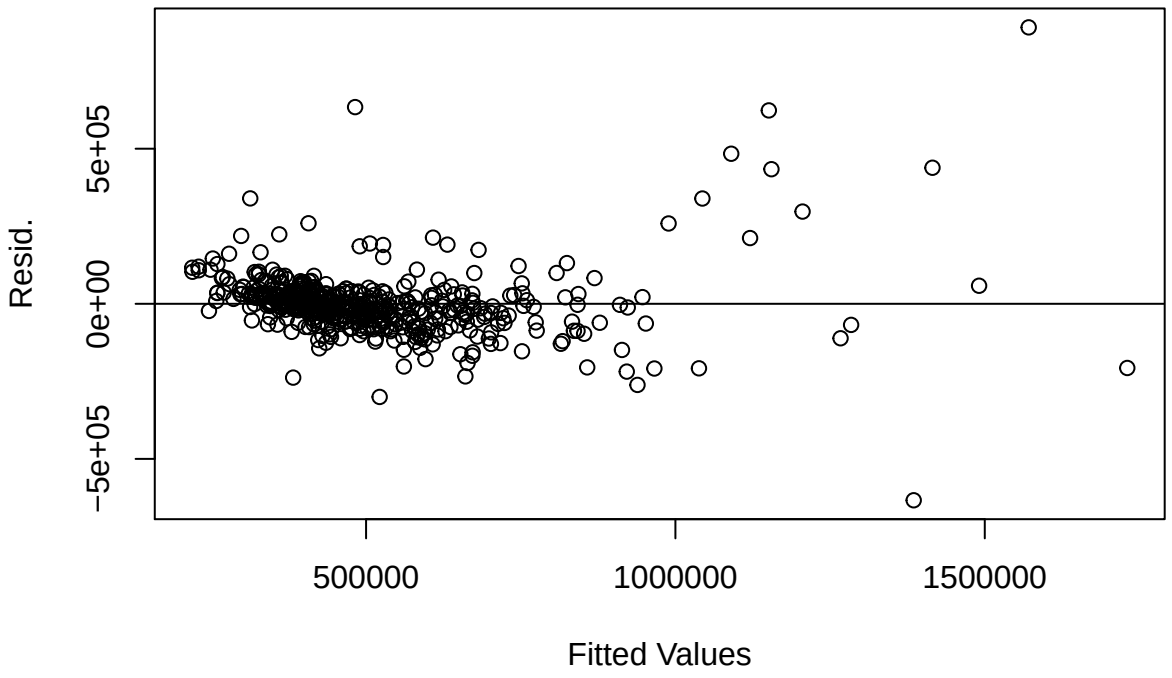
```
##
## Call:
## lm(formula = Cost ~ Built + Lot + Bed + Bath + Area, data = houses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -633376  -45936   -4636   31686  891079
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.903e+06  5.193e+05  -7.516 2.67e-13 ***
## Built        2.082e+03  2.673e+02   7.787 4.06e-14 ***
## Lot          4.003e+05  2.962e+04  13.514  < 2e-16 ***
## Bed         -1.577e+04  5.539e+03  -2.847   0.0046 **
## Bath         3.569e+04  7.647e+03   4.668 3.93e-06 ***
## Area         6.498e+01  5.564e+00  11.677  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 106100 on 495 degrees of freedom
## Multiple R-squared:  0.7755, Adjusted R-squared:  0.7732
## F-statistic:   342 on 5 and 495 DF,  p-value: < 2.2e-16
```

All of the coefficients appear to be statistically significant. The correct interpretation of these coefficients is

important to note. For example, the coefficient of the *Bed* count is as follows: after accounting for effects of year built, lot size, bathroom count, and area on both the cost and bedroom count, the cost decreases by approximately $15,000 for each additional bedroom. Wow! That's actually rather surprising. I would have thought that the cost should go up, but that's not what the data is suggesting. All other variables have positive coefficients, which is what we would expect though.

Now if our end goal is to predict the cost of a house given its characteristics, we should analyze this model closer before proceeding. Let's look at the plot of residuals vs. fitted values.

```
plot(mdl$residuals ~ mdl$fitted.values,
     xlab = "Fitted Values", ylab = "Resid.")
abline(h=0)
```

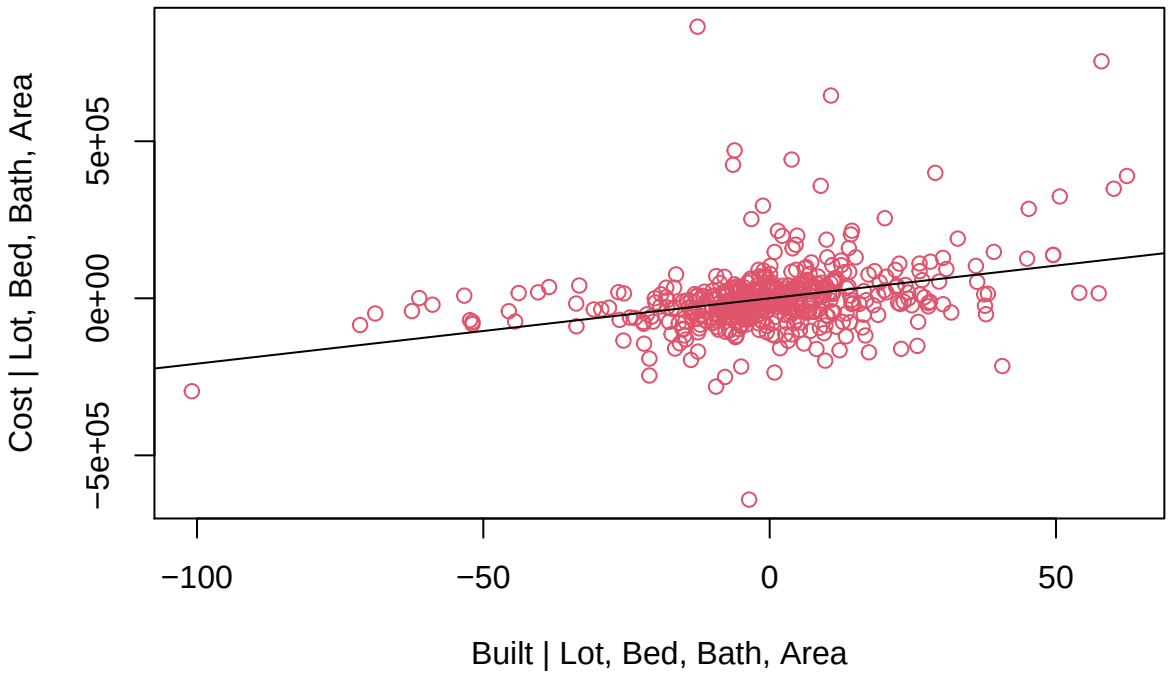


Hmmm... there does appear to be a slight pattern in the residuals on the left. This means we should probably adjust the model to improve our prediction. A type of plot that is very useful for this is called the Added Variable Plot.

Added Variable Plots

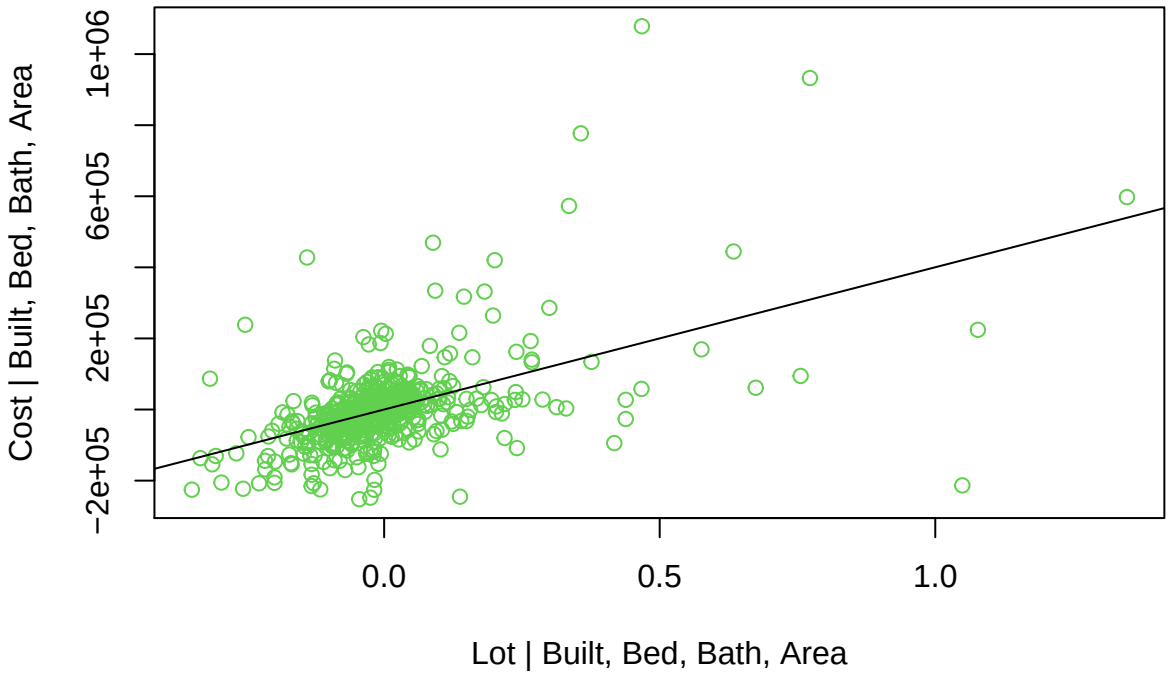
These are plots that take the response variable (which is the cost in our case), model out the effects of all but one explanatory variable, and plot those against the one explanatory variable after removing effects of all other explanatory variable from it as well. Mathematically, the slope of the best fit line in this new plot is the same as the coefficient in the original linear model.

```
c1 = lm(Cost ~ Lot + Bed + Bath + Area, data=houses)$residuals
x1 = lm(Built ~ Lot + Bed + Bath + Area, data=houses)$residuals
m1 = lm(c1 ~ 0 + x1)
plot(c1 ~ x1, xlab="Built | Lot, Bed, Bath, Area",
     ylab = "Cost | Lot, Bed, Bath, Area", col=2)
abline(m1)
```



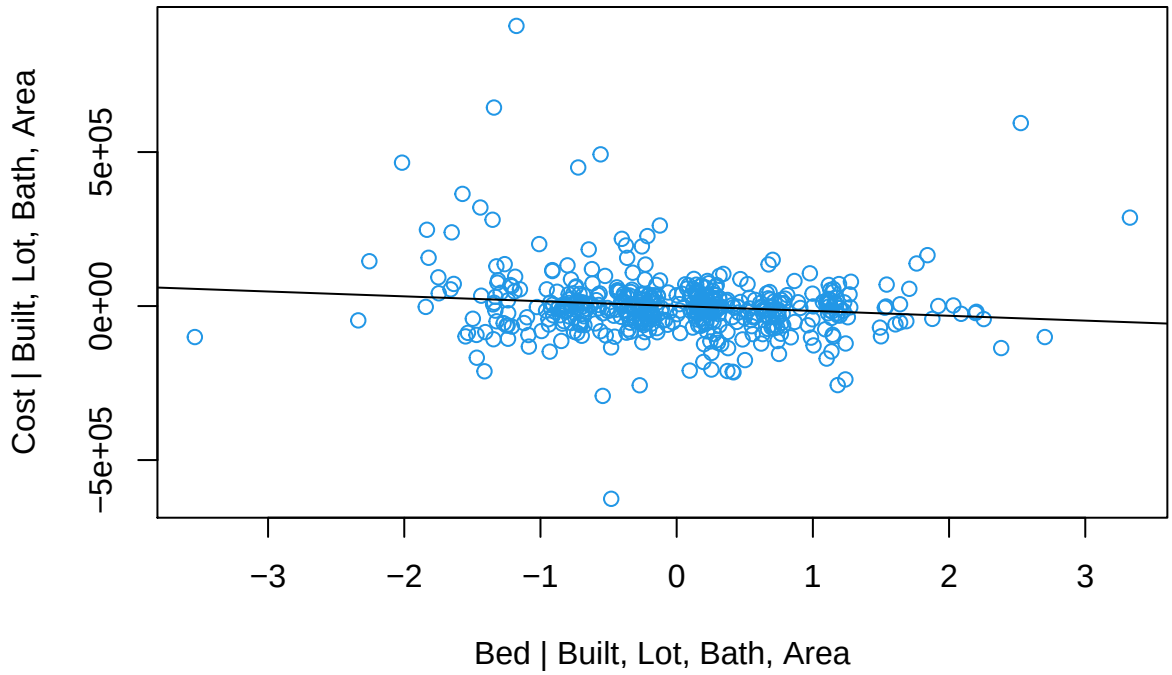
This added variable plot of the *Built* variable looks pretty good. Certainly nothing obvious that resembles the pattern we saw in the residual plot. Let's do the same thing for the *Lot* variable.

```
c1 = lm(Cost ~ Built + Bed + Bath + Area, data=houses)$residuals
x1 = lm(Lot ~ Built + Bed + Bath + Area, data=houses)$residuals
m1 = lm(c1 ~ 0 + x1)
plot(c1 ~ x1, xlab="Lot | Built, Bed, Bath, Area",
     ylab = "Cost | Built, Bed, Bath, Area", col=3)
abline(m1)
```



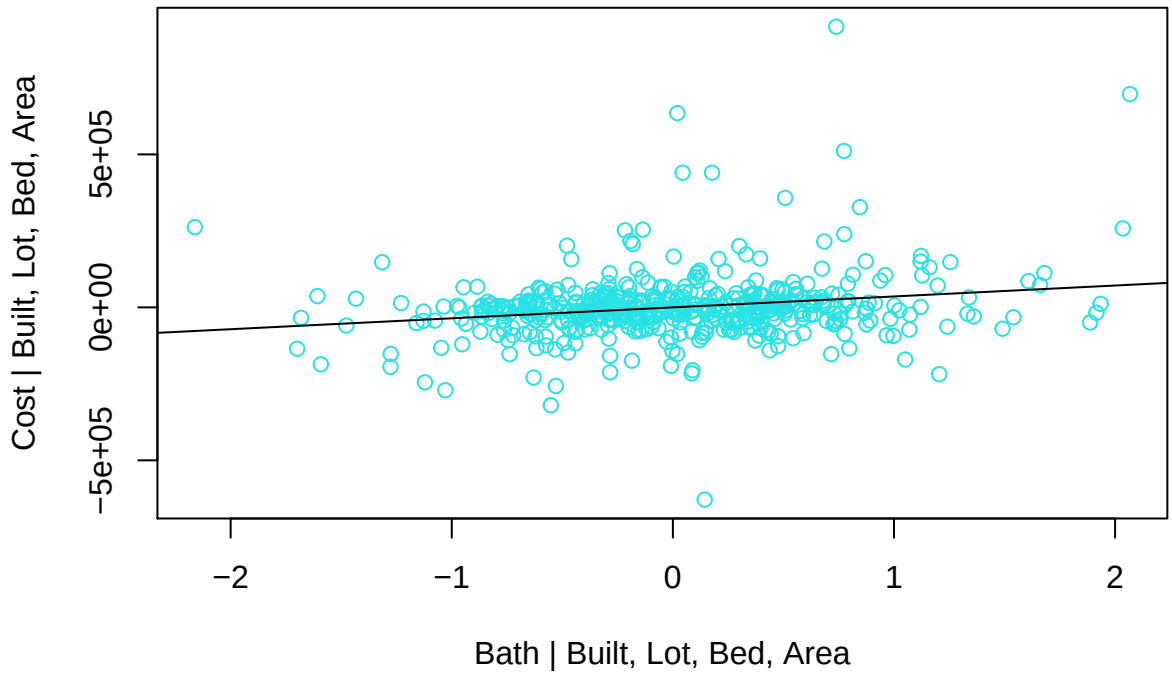
Nothing unusual here either. Let's look at the Added Variable Plot of *Bed* next.

```
c1 = lm(Cost ~ Built + Lot + Bath + Area, data=houses)$residuals
x1 = lm(Bed ~ Built + Lot + Bath + Area, data=houses)$residuals
m1 = lm(c1 ~ 0 + x1)
plot(c1 ~ x1, xlab="Bed | Built, Lot, Bath, Area",
     ylab = "Cost | Built, Lot, Bath, Area", col=4)
abline(m1)
```



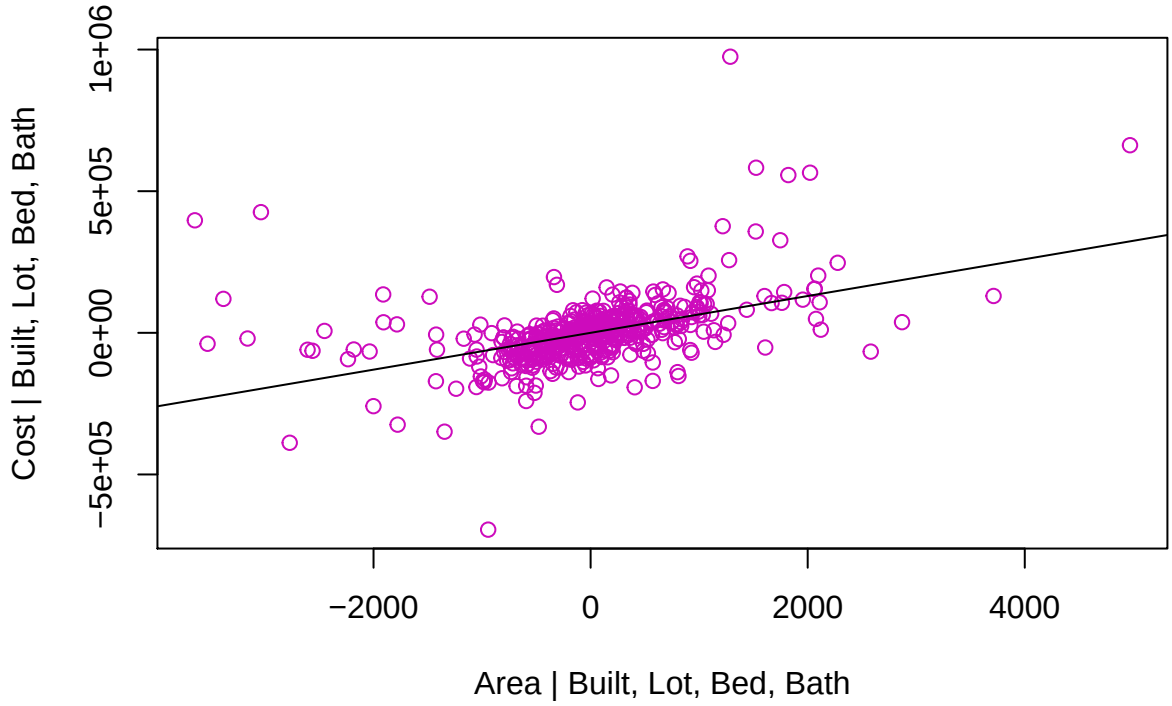
There's the negative slope that we saw before. Still quite an interesting thing to see, but no patterns here that suggest a change to the model.

```
c1 = lm(Cost ~ Built + Lot + Bed + Area, data=houses)$residuals
x1 = lm(Bath ~ Built + Lot + Bed + Area, data=houses)$residuals
m1 = lm(c1 ~ 0 + x1)
plot(c1 ~ x1, xlab="Bath | Built, Lot, Bed, Area",
     ylab = "Cost | Built, Lot, Bed, Area",col=5)
abline(m1)
```



Same thing with the *Bath* variable.

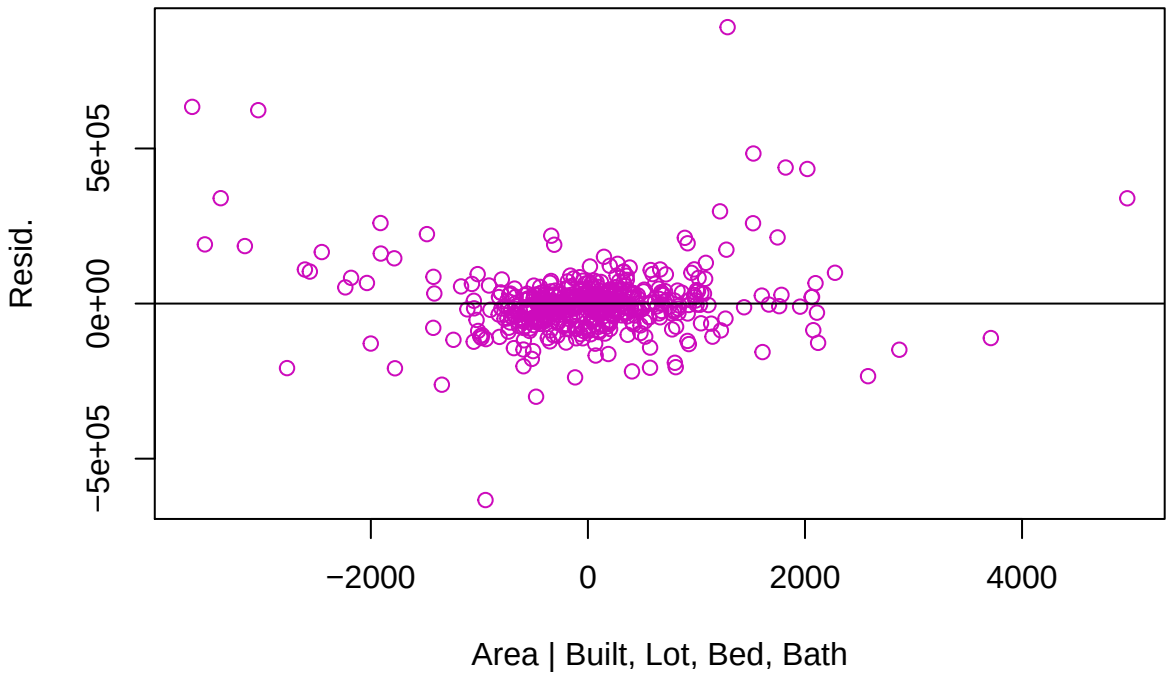
```
c1 = lm(Cost ~ Built + Lot + Bed + Bath, data=houses)$residuals
x1 = lm(Area ~ Built + Lot + Bed + Bath, data=houses)$residuals
m1 = lm(c1 ~ 0 + x1)
plot(c1 ~ x1, xlab="Area | Built, Lot, Bed, Bath",
     ylab = "Cost | Built, Lot, Bed, Bath",col=6)
abline(m1)
```



Aha! There's a bit of a pattern around the fitted line for the *Area* variable! The points in this plot start off all being above the fitted line, then start centering around the line and go back to being a little bit more

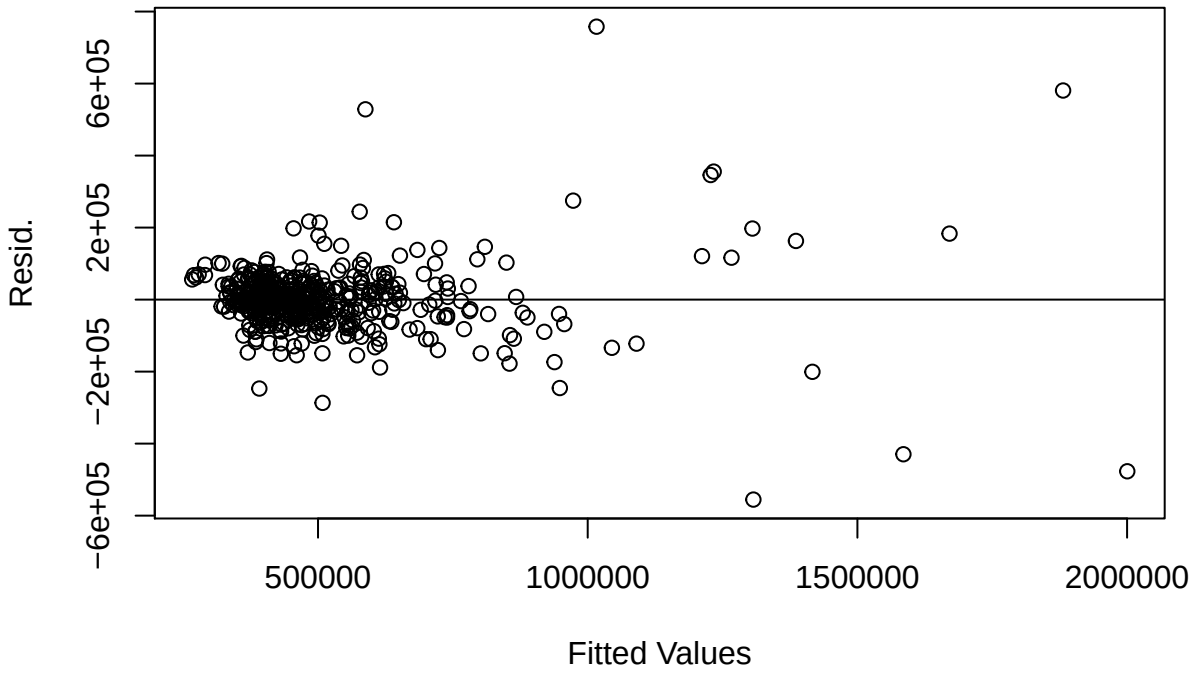
above than below the line. Let's take a look at the residual plot for this just to be sure though.

```
plot(m1$residuals ~ x1, ylab = "Resid.",
     xlab="Area | Built, Lot, Bed, Bath",col=6)
abline(h=0)
```



Yep! There's the trend that was probably driving the pattern we saw in the residual plot for the overall model. This plot now suggests that we need to add in a quadratic term of the *Area* variable. Let's try that and see if the overall residual plot looks better.

```
mdl2 = lm(Cost ~ Built + Lot + Bed + Bath + Area + I(Area^2), data=houses)
plot(mdl2$residuals ~ mdl2$fitted.values, ylab = "Resid.",
     xlab="Fitted Values")
abline(h=0)
```



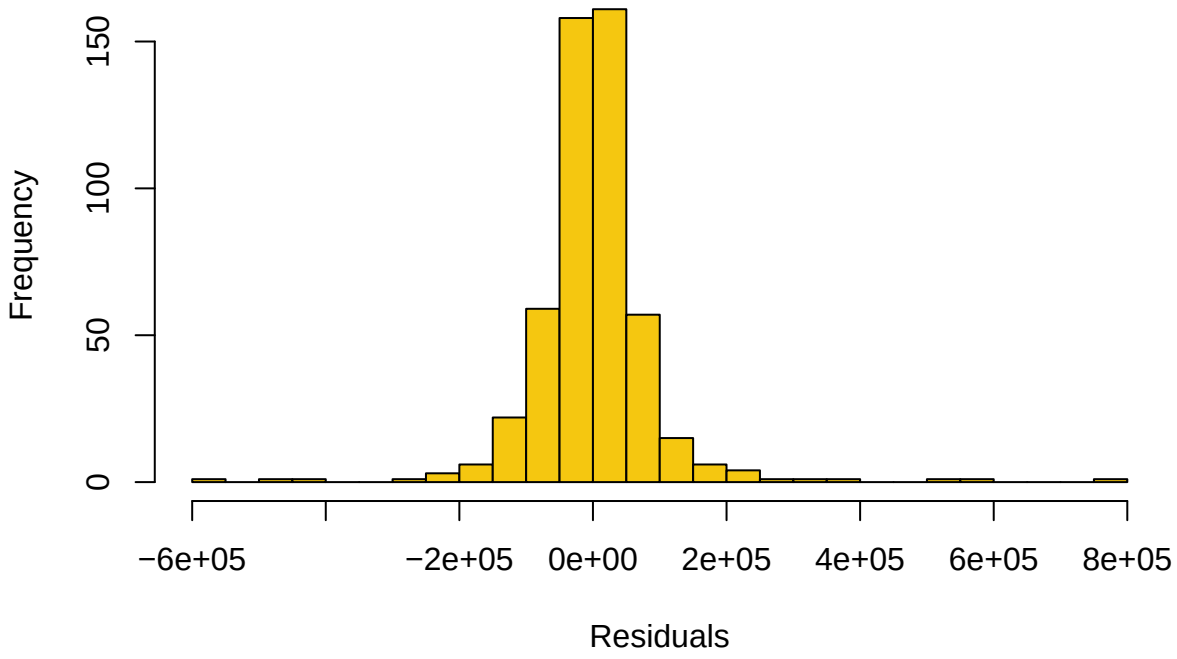
That looks much better! The pattern has disappeared and the residuals look pretty much randomly scattered about 0. Let's take a look at the summary of this new model.

```
summary(md12)
```

```
##
## Call:
## lm(formula = Cost ~ Built + Lot + Bed + Bath + Area + I(Area^2),
##     data = houses)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -555405  -34934    -442   33577  758007
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.368e+06  4.635e+05  -7.267 1.45e-12 ***
## Built        1.871e+03  2.381e+02   7.858 2.45e-14 ***
## Lot          2.998e+05  2.771e+04  10.818  < 2e-16 ***
## Bed         -6.876e+03  4.980e+03  -1.381  0.16798
## Bath         3.760e+04  6.794e+03   5.534 5.08e-08 ***
## Area        -2.941e+01  9.548e+00  -3.080  0.00218 **
## I(Area^2)    1.282e-02  1.110e-03  11.554  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 94260 on 494 degrees of freedom
## Multiple R-squared:  0.8233, Adjusted R-squared:  0.8211
## F-statistic: 383.5 on 6 and 494 DF,  p-value: < 2.2e-16
```

Yep, the quadratic term that we added in is actually a very statistically significant part of the model. Let's look at a histogram of the residuals to see if their roughly bell-shaped (i.e. normally distributed). This will mostly be a check to see how much we can trust the p-values for each coefficient.

```r
hist(mdl2$residuals, breaks = 30,
     xlab = "Residuals", main = "", col=7)
```



That looks pretty bell-curve shaped to me, with a few strong outliers on both ends. The only drawback of including the quadratic term in the model is that the other coefficients are (slightly) less interpretable.

Further Work

Many more steps could be taken to improve this model for the purposes of prediction. First, taking a closer look at the houses that have large residuals would be good. Perhaps they are influencing the model more than they should. Second, there is certainly a possibility that the cost is not linearly related to some of the other variables. This could be addressed by including other polynomial terms in the model, or by using a more sophisticated approach like random forest regression. From the looks of the added variable plots, however, it does not appear in advance that nonlinear models will improve the prediction by much.

Another use of this model would be to look at some of the houses on the market in Bountiful, UT right now and see if they are reasonably priced according to this model. Or perhaps someone is looking to buy, or sell, a particular kind of house in Bountiful and would like to know a reasonable range for the price.

Conclusions

Overall, we conclude that, after accounting for effects of year built, lot size, bathroom count, and living area, the value of a house in Bountiful decreases for each additional bedroom by about $15,000 on average. We also conclude that the cost depends quadratically, not linearly, on the living area.