Predicting the cost of housing in the bay area using a linear model

Abhi Mehta

12/3/2017

Introduction & Background

In post two, I wanted to expand upon post 1 with new information and new models to understand the broad topic of housing. I foucsed on housing in the San Francisco Bay Area and I wanted to see if we can use a statistical method in predicting how exactly a house might cost in this expensive market.

The data set can be found here: http://eeyore.ucdavis.edu/sts198/Data/SFHousing.rda

The dataset I used came from SF Chronicoles records on homes and apartments in the bay area. In particular, this report focused on building square feet, bedrooms, & location, price per sqft to be able to predict how expensive housing might be.

The housing market in San Francisco is one of the most expensive in the world due to its image as the center of innovation and startups. This unique area has created some of largest fluctuations in house pricing as more and more people come to the San Francisco Bay area. This has made it near impossible to not only live in the bay area and find a home for a family without a stellar income, but it has also become difficult to predict what a house might sell for.

In this post, we will use a linear model explore these discrepenies in the housing market by looking at factors including building square feet (bsqft), and the number of bedrooms (br) in various locations in the bay area as a factor that goes in as an input in the model in city of San Francisco and the city of Ricmond. Last time, I looked at just the city of Fremont in the South Bay but this time I wanted to go a step further and analyze a few different areas in the bay area to get a better feel for the factors that go into the linear model may vary.

Housing costs are a way great benchmark to understand an important aspect of the total cost of living as they serve as great indicator of how comfortabledy one may be able to live or how many people are in a house. Intuitively, one would think that as the number of bedrooms and square footage go up that the price also goes up. We all know that the bay area is expensive - but the question is that is the same for all of the bay area? How much do prices go up as the number of bedrooms change in a home or square footage depending on where you are? Let's try to use a linear model and find out!

Motivation

The housing market very much is important for all of us to figure out as we currently live in the bay area and must find a place to live once we graduate. Understanding these differences can motivate all of us to make better decisions based on real-life statistics!

Concepts

Linear models are functions in R that lets you create categorical interactions between categorical variables, continuous variables, and between numeric variables.

The argument to lm is a model formula, which has the response on the left of the tilde ~ and a Wilkinson-Rogers model specification formula on the right

In r, we can see the following specifications:

- **to combine elementary terms, as in A+B**
  for interactions, as in A:B;

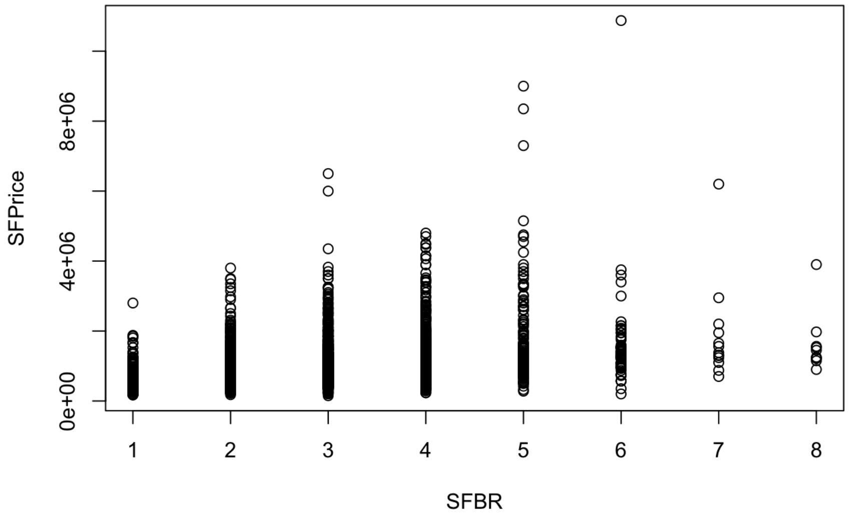
- for both main effects and interactions, so A*B = A+B+A:B

In our case, as shown later, our linear model is as follows - lm(formula = price ~ bsqft * city * brf, data = cities3)

A closer look at two cities - (San Francisco & Richmond)

Last time, we looked at Fremont. This time we wanted to pick more atypical cities in the bay area. In Richmond and in San Francisco, we looked at the cost of a house. The total cost of the home was then divided by the number of bedrooms to get the price per bedroom in each of these regions. The average price per bedroom varied per region and was used as a measure to estimate how much each of these homes in the bay area should cost given the number of bedrooms. The average price per bedroom was then multiplied by the number of bedrooms in each hoome to predict how much a home in that particular region should cost.

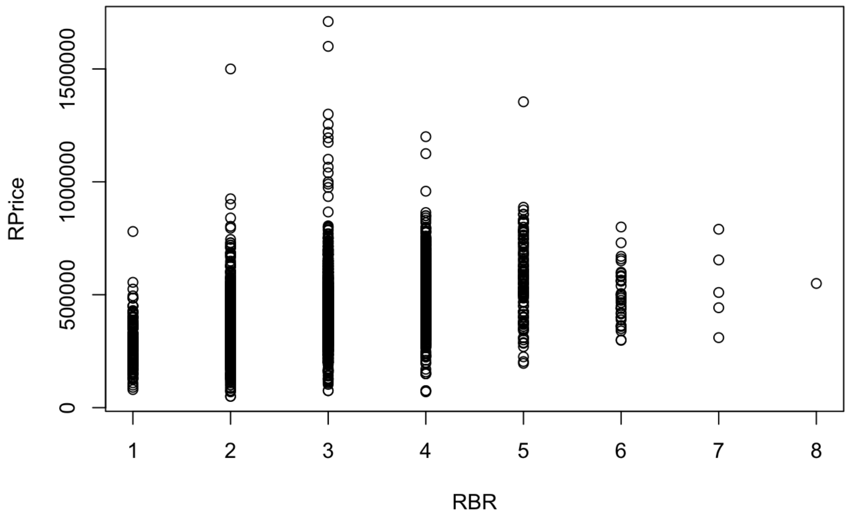
```r
#San Francisco
#I first created a subset of the data of housing prices
SFPrice = housing$price[housing$city == 'San Francisco']
# I took the average price
SFAvg = mean(SFPrice)
# I created a subset of the data of bedrooms in SF
SFBR = housing$br[housing$city == 'San Francisco']
SFAvgBR = mean(SFBR)
SFPriceBR = SFPrice / SFBR
# I then created a correlation and plot
plot(SFBR, SFPrice)
```



```
cor(SFPrice, (SFBR * mean(SFPriceBR)))
```

```
## [1] 0.4341846
```

Figure 1 - The scatterplot and correlation above reflect the data for San Francisco’s housing market with the price on the Y axis and the number of bedrooms on the X Axis. There is far more data centered around the lower number of bedrooms and that there is a clear trend of the price increasing as the number of bedrooms goes up. However, the correlation here isn’t as strong as one might expect.



```
## [1] 0.4122046
```

Figure 2 - The scatterplot and correlation above reflect the data for Richmond’s housing market with the price on the Y axis and the number of bedrooms on the X Axis. There is far more data centered around the lower number of bedrooms and that there is a clear trend of the price increasing as the number of bedrooms goes up. The correlation for Richmond seems to be weaker than Fremont.

Linear Model for Housing

After taking a closer look at these two cities, We wanted to see if we could build linear model to see if we can predict the cost of a home given three variables, the city, the building square feet & the number of bedrooms. This linear model would help us understand discrepnecies in the housing market based on where one might be in the bay area. This would also help us predict what a house might cost if you were in the cities of San Francisco, Fremont, or Richmond.

```
##
## Call:
## lm(formula = price ~ bsqft * city * brf, data = cities3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4767183  -104449   -19687    73826  3300080
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.194e+05  6.882e+04   1.736   0.0826 .
## bsqft                         2.095e+02  8.605e+01   2.435   0.0149 *
## cityRichmond                  5.058e+04  1.057e+05   0.478   0.6323
## citySan Francisco             4.255e+05  6.915e+04   6.154 7.70e-10 ***
## brf2                         -1.068e+04  7.351e+04  -0.145   0.8845
## brf3                          5.360e+04  7.038e+04   0.762   0.4463
## brf4                         -6.043e+04  7.120e+04  -0.849   0.3961
## brf5+                        -1.954e+05  7.885e+04  -2.479   0.0132 *
## bsqft:cityRichmond           -6.783e+01  1.317e+02  -0.515   0.6066
## bsqft:citySan Francisco      -2.020e+02  8.608e+01  -2.347   0.0189 *
## bsqft:brf2                    6.414e+01  8.911e+01   0.720   0.4717
## bsqft:brf3                    6.226e+01  8.662e+01   0.719   0.4723
## bsqft:brf4                    1.261e+02  8.651e+01   1.457   0.1450
## bsqft:brf5+                   1.878e+02  8.705e+01   2.158   0.0310 *
## cityRichmond:brf2             7.721e+04  1.106e+05   0.698   0.4851
## citySan Francisco:brf2       -2.307e+03  7.446e+04  -0.031   0.9753
## cityRichmond:brf3            -5.693e+04  1.087e+05  -0.524   0.6006
## citySan Francisco:brf3       -9.743e+04  7.192e+04  -1.355   0.1755
## cityRichmond:brf4             9.917e+04  1.128e+05   0.879   0.3792
## citySan Francisco:brf4       -6.829e+05  7.784e+04  -8.773  < 2e-16 ***
## cityRichmond:brf5+            2.308e+05  1.346e+05   1.714   0.0865 .
## citySan Francisco:brf5+      -1.671e+05  9.353e+04  -1.787   0.0740 .
## bsqft:cityRichmond:brf2      -8.440e+01  1.351e+02  -0.625   0.5320
## bsqft:citySan Francisco:brf2  9.670e+01  8.945e+01   1.081   0.2796
## bsqft:cityRichmond:brf3      -7.974e+00  1.330e+02  -0.060   0.9522
## bsqft:citySan Francisco:brf3  1.828e+02  8.695e+01   2.102   0.0355 *
## bsqft:cityRichmond:brf4      -1.036e+02  1.334e+02  -0.777   0.4373
## bsqft:citySan Francisco:brf4  4.824e+02  8.758e+01   5.509 3.66e-08 ***
## bsqft:cityRichmond:brf5+     -1.698e+02  1.364e+02  -1.245   0.2132
## bsqft:citySan Francisco:brf5+  2.227e+02  8.839e+01   2.520   0.0117 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 241400 on 22773 degrees of freedom
##   (38 observations deleted due to missingness)
## Multiple R-squared:  0.546,  Adjusted R-squared:  0.5454
## F-statistic: 944.5 on 29 and 22773 DF,  p-value: < 2.2e-16
```

Figure 3 - Above we can see a linear model as the variables "city", "bsqft", and "data" from the three cities in our data are considered as inputs to predict the price when different factors are considered to see how much a house might actually cost in a given area.

Conclusion & Take-Home Message

The first blog post showed us one city, Fremont, and its relation to bedroom price. However, this is just a small part of the bay area and it would be much more helpful to take it a step further and understand these discrepencies in the bay area depending on where in the bay someone might be.

We saw that the number of bedrooms was a better predictor in Richmond, a more residential area In San Francisco, the correlation between home price & the number of bedrooms was not nearly as high. However, just by looking at bedrooms, we weren't getting very strong correlations.

We included building square feet (bsqft) into the analysis. Using a linear model that took in the inputs of building square feet, the number of bedrooms, and location (San Francisco, Fremont, or Richmond) let us build a model to help us predict how much each factor may actually affect the price of a house in the Bay Area. We found that location was a huge determinate. In a different location, the price per bedroom, and the price per squarefoot both had an impact on the actual price of a house. These results came as expected but to be able to visaulize which factors affect house prices help us better understand what the bay area housing market really looks like.

It was difficult to generalize that data into one clear data set. It's important to understand just how different the bay area is and this post put us closer to understanding the SF Bay area's volatile housing market beyond just bedrooms while better understanding linera models.

References

https://sf.curbed.com/2017/9/29/16385146/san-francisco-price-per-square-foot

http://www.stanford.edu/~vcs/StatData/SFHousing.rda

http://data.princeton.edu/R/linearModels.html

http://eeyore.ucdavis.edu/sts198/Data/SFHousing.rda

https://www.rdocumentation.org/packages/stats/versions/3.4.1/topics/lm

https://www.bizjournals.com/sanjose/news/2017/05/08/bay-area-housing-market-fremont-oakland-san-jose.html

https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R

https://smartech.gatech.edu/bitstream/handle/1853/31763/Corsini_Kenneth_R_200912_mast.pdf