The Cost of Living In The SF Bay Area - Understanding Data Through Correlations and Plots

Abhi Mehta

10/31/2017

Introduction

The world around is full of data. One area where we have tons of data is housing - from house sales to the square foot of homes. Every year, we collect more and more data about how we live and how these conditions are changing. Housing is one area that affects everyone. One way we can look at these changes is by looking at the cost of bedrooms in the San Francisco Bay Area. As someone who is new to R, it is important to understand how we can use data to visualize things that are applicable in our own lives. R makes it easy to see such through basic analysis tools such as tables, correlations, boxplots, x-y plots, and a variety of other built in tools. In this blog post, I will focus on correlations & plots as a method for understanding the San Francisco Bay Area housing market.

Motivation

How exactly can someone go about this process and look at so much data? My motiation focuses on being able to learn from these varianves I find in this huge data set accross the bay area so that I can learn more about the housing market and one day plan where I might live in the bay area after college. Learning more about how much each area costs in the bay area through basic data analysis using correlations and tables would allow us to get a much better picture of this problem that everyone must consider when choosing the right place to live in the SF bay area.

Background - Concepts

First, I'd like to talk about correlations. Correlations are used to test the relationship between two varaibles. It is a single number that may prove some kind of relationship but can typically not prove a causal relationship. Correlations in r can be done through the cor() function. In the case of applications, correlations can help us understand if two variables are related such as prices and the number of bedrooms in a house.

Second, let's talk about plots - Scatterplots & boxplots are two basic plots in R. They come built in and can be used via the function plot() or boxplot(). Both have an X and Y axis that can help an audience understand a data set beyond just numbers. Plots show a series of various data points such as all the number of bedrooms in a house and the price. Boxplots give more information in that they show the spread in terms of the different quartiles.

Background - Data Set (SF Housing)

The dataset I used came from SF Chronicoles records on homes and apartments in the bay area including their selling price, # of bedrooms, home square feet, lot squsare feet & location. This data set also included data on when the house was last sold with most of the prices avaliable in the past decade.

Our data set included homes in all of the bay area and in particular, 9414 homes in Fremont. Note that by homes, we mean houses because to limit external factors that may exist due to housing conditions such as shared bedrooms etc.

Using prices from different areas such as Fremont allowed us to get a better overall feel for the bay area. We looked at homes & bedrooms in these locations to see if the number of bedrooms affected the house price.

Making Sense of Data - Examples of Correlations & Plots

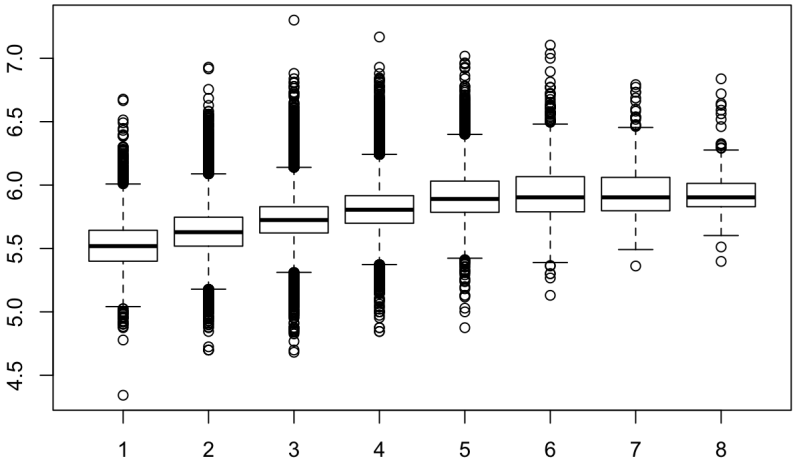
1. Correlations & Plots (All of Bay Area)

Let's take a look at the relationship between the number of bedrooms for the overall bay area. First, we took a look at the correlation between the price of homes and the number of bedrooms. For the visualization, I chose a boxplot because it offered a complete view of the bay while allowing us to see the spread.

```
#Bay Area
cor(housing$price, housing$br)
```

```
## [1] 0.3577729
```

```
#Bay Area Housing Price & Bedrooms
boxplot(log10(housing$price)~ as.factor(housing$br), title = 'Bay Area Housing Prices & Bedrooms')
```



The boxplot above shows the price in hundreds of thousands ranging from 450,000 to 700,000 in price and the data ranges from 1 bedroom to 8 bedrooms for a home in the San Francisco Bay Area. This graph gives us an overall picture of how much it costs in the bay area for a home of different sizes. We can see that the price of homes level off after reaching 6 bedrooms.

2. Correlations & Plots (Fremont - City in Bay Area)

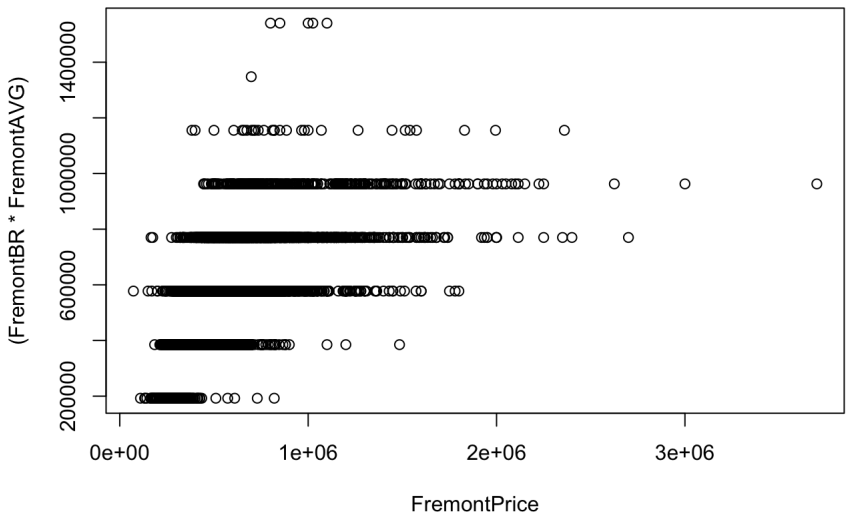
Although the trend seems to match expectations, we wanted to better understand individual areas of the bay area given how different cities like San Francisco and Fremont might be.

In Fremont, we looked at the cost of a home whether it was an apartment or 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. We used the function cor() to understand this relationship through one number and then through a scatterplot to see all of the data points through subsetting.

```
#Fremont
FremontPrice = housing$price[housing$city == 'Fremont']
FremontAvg = mean(FremontPrice)
FremontBR = housing$br[housing$city == 'Fremont']
AvgFremontBR = mean(FremontBR)
FremontPriceBR = FremontPrice / FremontBR
FremontAVG = mean(FremontPriceBR)
cor(FremontPrice, FremontBR)
```

```
## [1] 0.5914326
```

```
plot(FremontPrice, (FremontBR * FremontAVG))
```



Conclusion & Take-Home Message

Although very simply, correlations and plots are a great way of understanding large amounts of data quickly. They give a great picture of

relationships between two things that aren't causally related to understand how one might potentailly influence the other variable. In this blog post, I took these two basic concepts to a very real situation of being able to understand the San Francisco Bay Area housing market through bedrooms.

Intuitively, one would think that as the number of bedrooms go up, that the price would also go up. However, although this was the trend, we learned far more from this data. After looking at the bay area as a whole, I saw that the price of homies increased up until 6 bedrooms and then flattened. As expected, most of the data centered around homes with less bedrooms given the high costs of living in the bay area. Additionally, the correlation between home prices and bedrooms was far lower than predicted at only .3577 for the entire data set. Since the bay area was big, I wasn't sure if this was conclusive personally so I took a deeper dive.

We wanted to dive deeper to understand the data as the pricing may not be the same for all of the cities in the bay area. Subsetting and utilizing basic functions in R such as correlation of a smaller subset of the data allowed us to see this. We chose to look at Fremont, a residential city in the San Francisco Bay Area. Here through the plot we saw that the price and and number of bedrooms was much closer. In fact, by using correlation function in this subset, we saw that the correlation was .5914, significantly higher than all of the bay area. This indicates that we really cannot just look at the bay area as one place. As we all think about housing, we must consider area as a primary factor.

The take away here is that there is a large fluctuation based on where one might actually be in the bay area as we can see by using basic correlations and plots in R giving us a great first glimpse at how important relationships are in understanding data.

References

SF Chronicle Data Set - http://www.stanford.edu/~vcs/StatData/SFHousing.rda

https://www.socialresearchmethods.net/kb/statcorr.php

https://smartasset.com/mortgage/what-is-the-cost-of-living-in-san-francisco

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

http://www.businessinsider.com/san-francisco-pending-home-sales-plunge-2017-8

http://www.theanalysisfactor.com/r-tutorial-part-13/

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