Challenge 3

Martin Gascon

Jul 20, 2015

Introduction

Motivation and Goals

Let's suppose I have a construction company that want to build houses in San Francisco. The question would be where to build in the city in order to get the maximum profit out of it. If the property value is rising in some points of the city, those points will likely to push up the price of houses in adjacent neighborhoods. We also have to consider that there is an inertia, so the prices won't go up immediately. This fact makes some adjacent neighborhoods to be the candidates to be growing at a much higher rate to catch up the relative distance between them and the already hot neighborhoods.

I’d like to take data analysis tools to decide which neighborhood have the chance to grow faster than the others.

Some of the various questions that I would like to answer include:

- How are neighborhoods related in terms of price change?
- Are they geographically close?
- Are there factors (bridges, rivers, criminal activity, etc.) that prevent the price to go up even when the closest neighborhood is getting too expensive?
- What is the effect of Gentrification and Ethnicity ratio affects the actual correlation?

Considering that San Francisco is one of the most expensive cities in the US, I believe that this project has the potential to be extremely interesting. This project could contribute to the future development of the city and its urbanization projects.

About the Data

For this project, I’ve downloaded data from Zillow (www.zillow.com) which provides the monthly mean value per square feet of all homes in U.S. This data includes values per month organized by zip codes for every major city in the country. I’ve selected those corresponding to San Francisco but the analysis could be extended to any major city in the U.S.

Analysis

Exploratory analysis

The exploratory analysis using R language reveals that we have around 9000 zipcodes and 235 variables from which 229 correspond to monthly prices. I selected those corresponding to San Francisco (23) and eliminated one of them since it had mostly NA values (Zip code 94117).

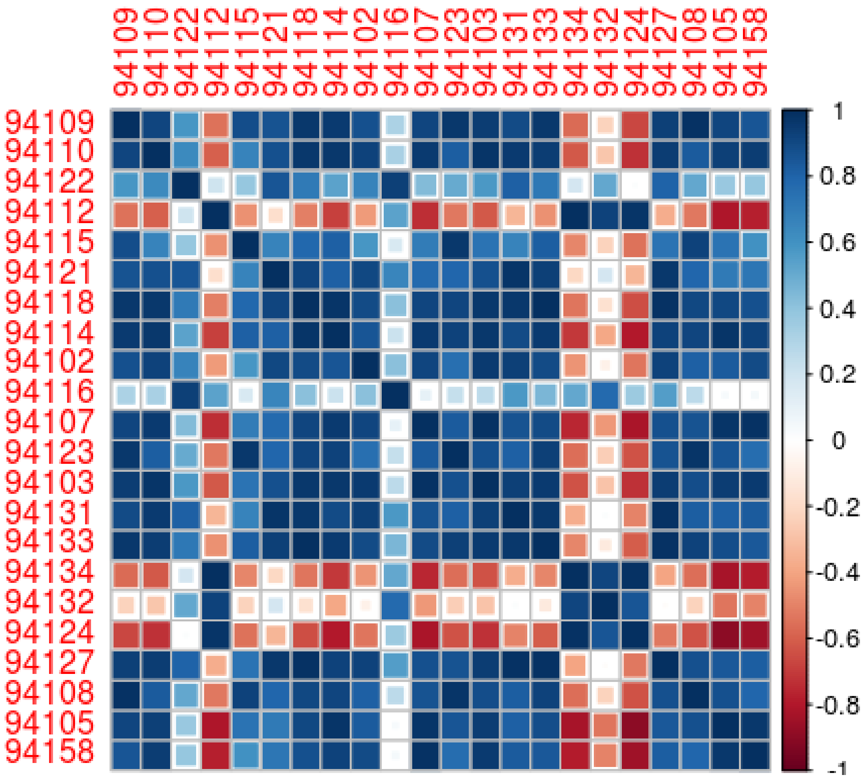
Data Processing

After removing some irrelevant columns in our data, I loaded a matrix with the price information of 22 zip codes. Then, I calculated the correlation matrix for this data and print the first 5x5 sector of the matrix.

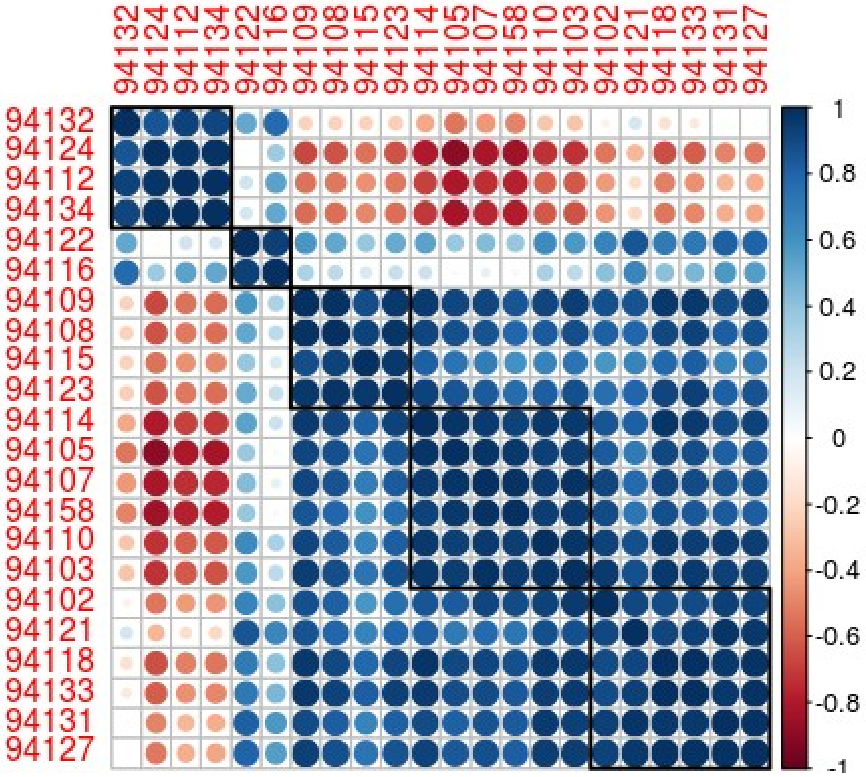
```
  # calculate the correlation matrix
  set.seed(1)
  cormat <- apply(data1, MARGIN=1, FUN=function(z) apply(data1,
MARGIN=1, FUN=function(y) cor(z, y)))
  print(cormat[1:5,1:5])
  M <- cor(cormat)

       94109  94110  94122  94112  94115
94109 1.0000 0.9879 0.9843 0.9560 0.9878
94110 0.9879 1.0000 0.9864 0.9444 0.9669
94122 0.9843 0.9864 1.0000 0.9738 0.9642
94112 0.9560 0.9444 0.9738 1.0000 0.9446
94115 0.9878 0.9669 0.9642 0.9446 1.0000
```

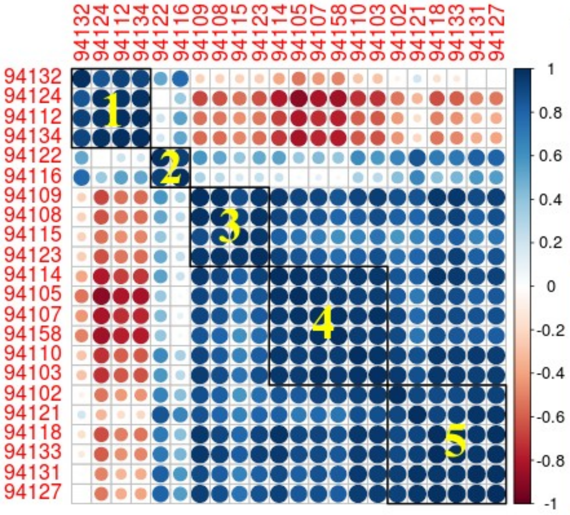
Now, using the corrplot package which is a graphical display of correlation matrices, I obtained the following plot:

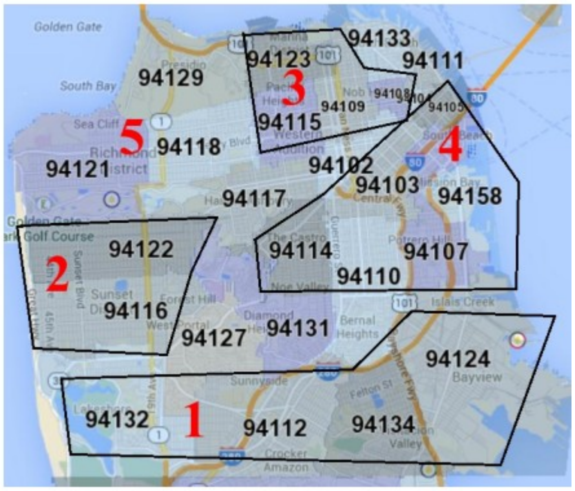


Now, if I perform a matrix reordering, the plot shows the correlation of zipcodes grouped in five main blocks.



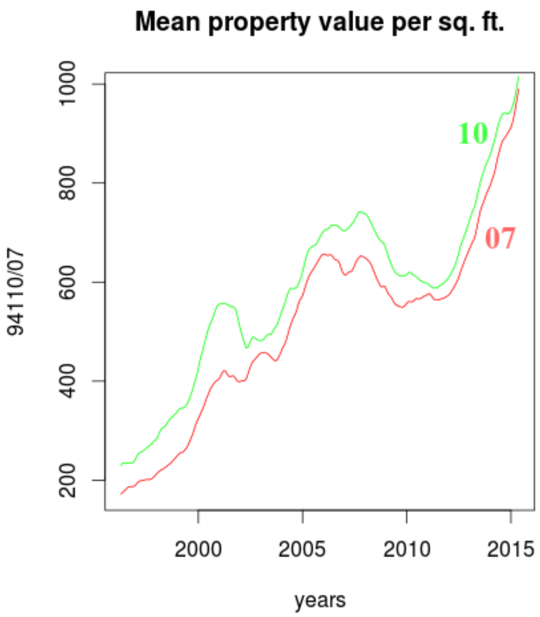
Let's assign a number to each block of correlated zip codes. We have to remember that zip-code 17 was excluded. Now if we plot the zip code map of San Francisco we observe a clear correlation to those 5 regions. In this plot we can see five or six different regions of interest. Region 1 corresponds to the four south neighborhoods (zip codes 32-12-34-24), while Region 2 is the one on the west part of the city (16-22). Region 3 groups the neighborhoods on the north part of the city (23-15-08-09) and Region 4 corresponds to the ones in the east (14-10-07-03-58). Finally Region 5 could be spitted in two correlated zones. One on the top-left corner of San Francisco (02-21-18) and another one in the middle (27-31).

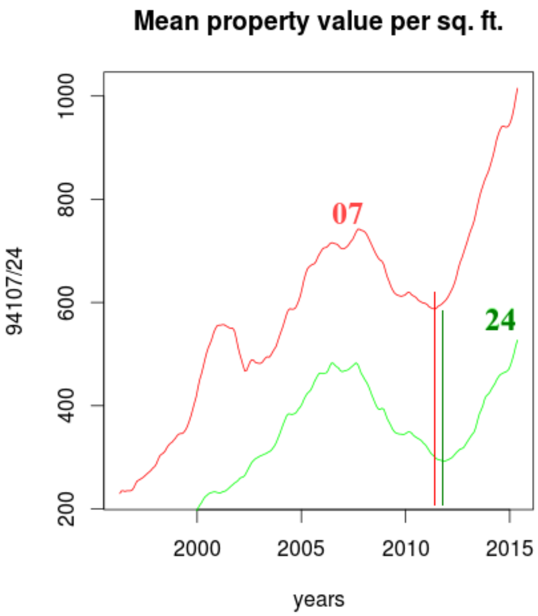




Considering the clear correlation that we have between neighbohrs we can extract some conclusions:

- If the mean property price of Region 5 goes up, the neighborhoods corresponding to Region 1, do not observe the same change at the same time. Let's put an example: In this case, since 94107 is in the same region as 94110, there is a perfect correlation and it is hard to expect a different behavior one from the other. However, if we look at 94124 and compare it with 94107, which is its neighborhood, we can see that 24 was going to go up around 2011 a few months later than 94107 because their correlation was quite poor.





Next Steps

I’d like to perform more exploratory analysis to determine if this trend is similar for cities with a larger set of data.