Data Analysis and Visualization in R (IN2339)

Case Study

Abdo, Arda Andirin, Denis, Hui Zheng

2022-01-20

Motivation

All around the world, a trend towards increasing housing prices has been observable. Barcelona has been leading the list of most expensive cities in terms of price per square meter for many consecutive years now.[1] This effect can usually be traced back and allocated to different causes: Rural exodus, Immigration, Inflation, and many more. It can be demand-driven, but also supply-driven if there is simply a lack of housing offers. The goal of this analysis is to check if disposable household income, since it can be a good indicator for the wealth of an area, and immigration (an increase in demand), are associated with rent prices in Barcelona.

Data Preparation

We mainly use four data tables, including *immigrants_by_nationality.csv* and *population.csv* from the given Barcelona dataset, and disposable household income per person[2] and average monthly rent per surface[3] from Barcelona's City Hall Open Data Service[4]. The data we use are all from the year 2017.

First, we rename the columns, as many of them were in Catalan. Then we check for NAs in all the data tables we use. After removing NAs, we look for neighborhoods that are common in all datasets, which left us 70 neighborhoods. We take the average of the quarterly rent prices to make sure the results are in (€/m2 per month) on a yearly basis. To eliminate the influence of population size, we further divide the total number of immigrants by the total population of the neighborhood. Finally we merge all the information we need for further analysis into one data table, named *rent_immi*. Here is the first row from *rent_immi*:

```
head(rent_immi, 1)
```

```
##    Neighborhood Code ave_rent Immigrants District Name total_population
## 1:                 1   13.895      15713  Ciutat Vella            47608
##    immigrants_ratio Income
## 1:        0.3300496  11407
```

Data Analysis

Barplot

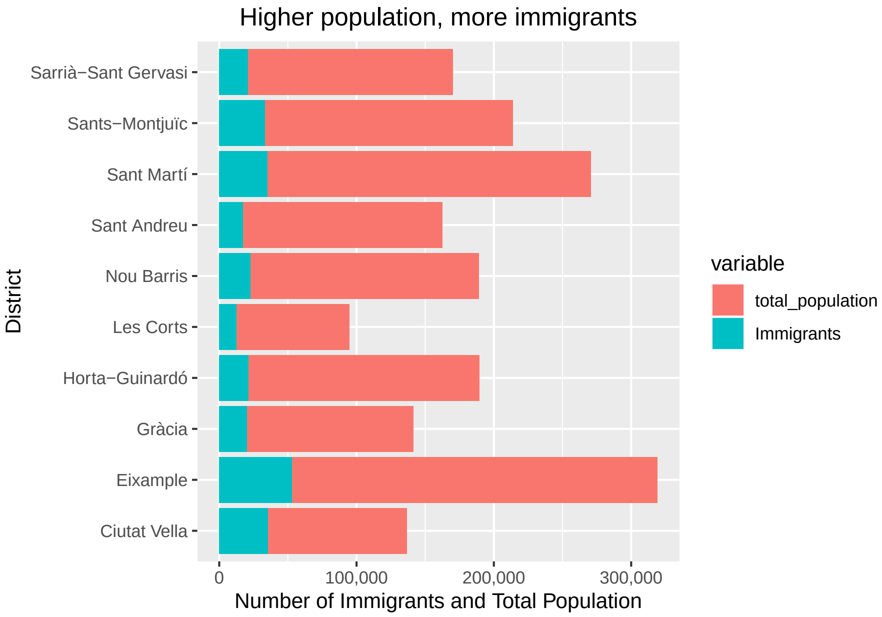
To gain first insights into the data, we create one barplot for the number of immigrants and total population by district.

[1] https://housinganywhere.com/rent-index-by-city

[2] https://opendata-ajuntament.barcelona.cat/data/en/dataset/renda-disponible-llars-bcn,

[3] https://opendata-ajuntament.barcelona.cat/data/en/dataset/est-mercat-immobiliari-lloguer-mitja-mensual

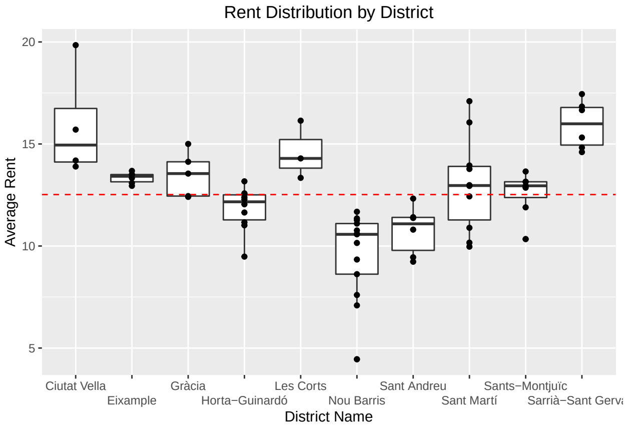
[4] https://opendata-ajuntament.barcelona.cat/

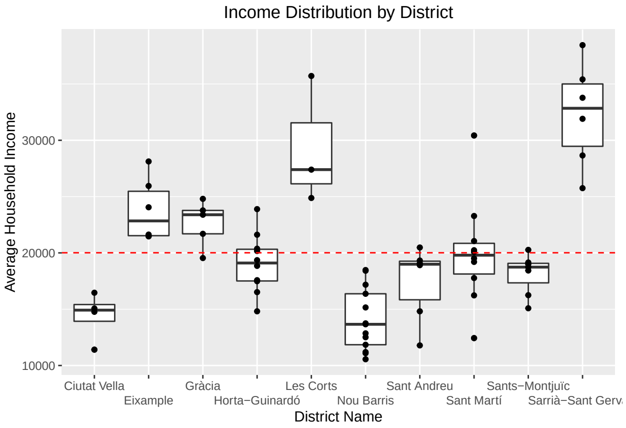


From this barplot, we can observe the number of immigrants and the total population in each district of Barcelona and the relationship between them. It's clear that the higher the population of the district, the higher the number of immigrants.

Boxplot

To obtain good graphical insights into the distribution of the data, we create two boxplots on average monthly rent per surface and average household income by district.



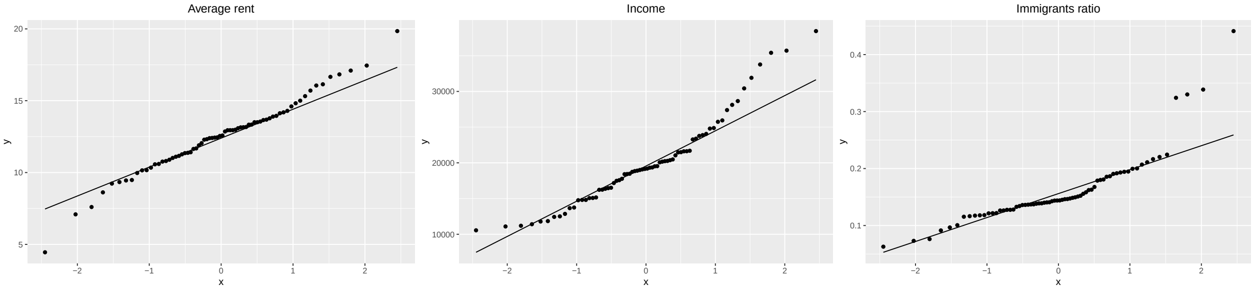


The left figure represents that the distribution average rent (€/m2 per month) of the city of Barcelona and their relationship to the global average in the dashed red line. The average rent seems to be evenly distributed around the global average. We also observe that the three most populated districts, namely Eixample, Sant Martí, and Sants-Montjuïc are closer to the global average, which is self-evident, as they contribute more to the mean.

From the right figure, we can see how the average household income in the city of Barcelona is distributed among the 10 different districts. We see that they are not as evenly distributed around the global average as the average rent.

Q-Q plot

Then we further use Q-Q plots to explore the distribution of the three most important variables, namely average rent, income and immigrants ratio by neighborhood, for our further statistical testing. We assume here the Normal distribution (Gaussian with mean 0 and variance 1) as reference theoretical distribution.

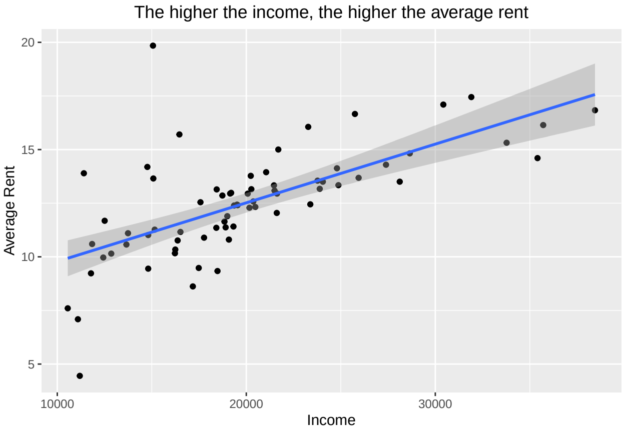


As we can see from the Q-Q plots, the distribution of average rent, income and immigrants ratio by neighborhood is very close to the Normal distribution, which allows us to use pearson correlation test later. An observation is that especially income and immigrants ratio have more extreme values.

Hypotheses and testing

Income and rent

Generally speaking, wealthier areas have higher rents. So our null hypothesis is that disposable household income and average rent are not correlated. Then we check for the relationship between them.



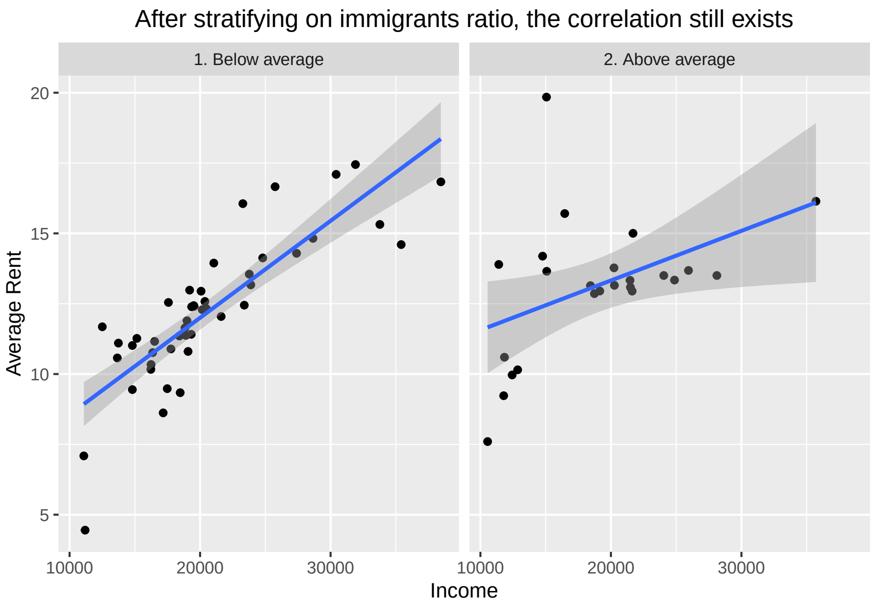
Here is an associative plot of disposable household income and average rent. We can see a positive correlation between two variables. The higher the income of a neighborhood, the higher the rent price. To see if there actually exists a correlation statistically, we apply spearman and pearson correlation tests. Here is the result of spearman correlation test.

```r
cor.test(rent_immi$Income, rent_immi$ave_rent, method="spearman")$p.value
```

```
## [1] 5.364338e-12
```

As the p-values are quite low, we can reject the null hypothesis and say that there exists a strong correlation between disposable household income and average rent.

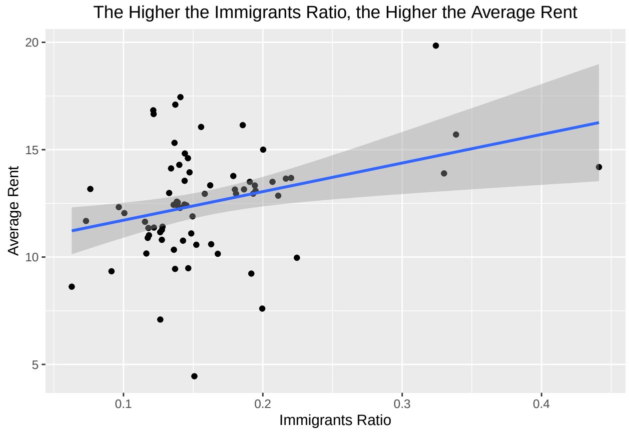
Now we consider immigration as a confounding factor. Because we only have data for 70 neighborhoods in Barcelona, we stratify it into “Below average” and “Above average” instead of more strata to avoid contingency.



After stratifying the data on immigrants ratio we can still see a correlation, with pearson test p-values of 0.03 for above-average income and 6.476e-13 for below-average income. So after controlling for the confounding variable immigration, we can reject our null hypothesis and claim that there is a positive correlation between household income and average rent.

Immigration and rent

Since immigration can lead to an increase in demand of housing market, we consider immigration as a significant variable in the housing market of Barcelona. So our null hypothesis is that immigration is not correlated to the average rent. Then we check for the relationship between immigration and average rent.



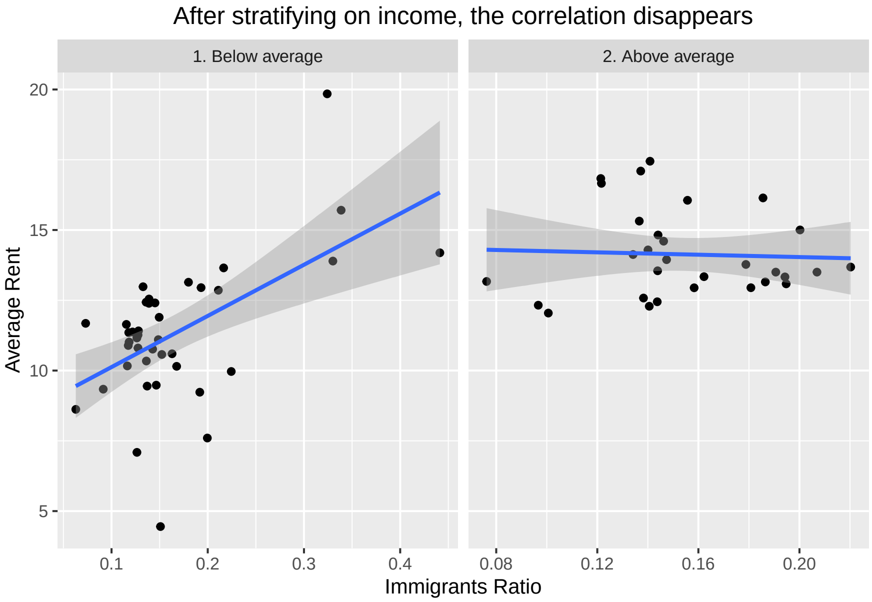
Here is an associative plot of immigrants ratio and average rent. We can see a positive correlation between two variables. The higher the immigrants ratio, the higher the rent price. To see if there actually exists a correlation statistically, we use spearman and pearson correlation tests. Here is the result of pearson correlation test.

```
cor.test(rent_immi$immigrants_ratio, rent_immi$ave_rent, method="pearson")$p.value
```

```
## [1] 0.006694315
```

As the p-values are quite low, we can reject the null hypothesis and say that there exists a correlation between immigration and average rent.

Now we consider income as a confounding factor.



After stratifying the data on income, we find that for neighborhood with income above average, the p-values are 0.85 for spearman test and 0.81 for pearson test, which means that the correlation disappears.

Hence, after controlling for confounding variable income, we fail to reject our null hypothesis.

Conclusion

With the performed analysis, we find that there is a strongly positive correlation between disposable household income and average rent by neighborhood. After stratifying the data on immigrants ratio, we can still see a correlation. Then we can reject our null hypothesis and claim that there is a positive correlation between household income and average rent.

Also, we find that there is a positive correlation between immigrants ratio and average rent by neighborhood. Then we check for disposable household income as a confounding factor. However, after stratifying on income, we find that for neighborhoods with income above average, the correlation disappears. Hence, we fail to prove our claim that immigration is a significant variable in the housing market of Barcelona.

In conclusion, looking back at our initial motivation, we can establish that disposable income is correlated with the average rent prices while immigration ratio is not.

Version control (will be deleted)

v1, 2022-01-19, add 1 Q-Q plot, footnotes, proofread, Hui

v2, 2022-01-20, line 116 neighborhoods->district, line 172(district names corrected) , small grammatical details, line 193 last sentence, line335 ,qqplot titles, proofread Arda,

v3, . . . , add name,