Barcelona

Lyla Dejoua, Chiara Menini, Lidia Torres, Daniel Moreno

22/1/2022

Motivation

Barcelona is the second most populated municipality in Spain. It was therefore of high interest analyzing its population regarding the year 2015.

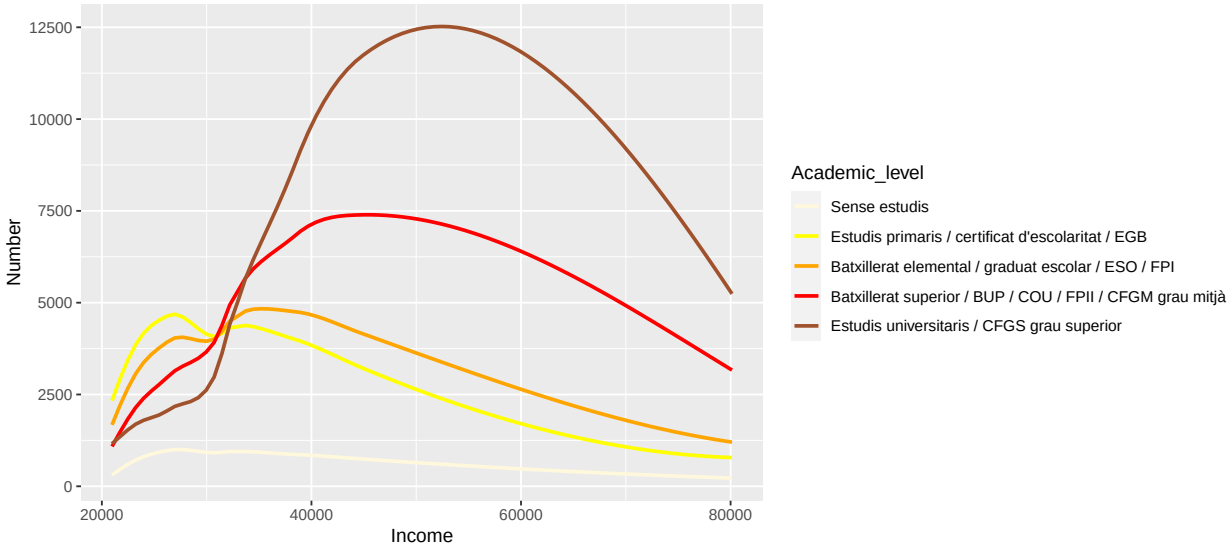
The goal of this analysis is to have a better understanding of the level of education of the population in the city. More precisely, the aim is to study whether and how the different educational levels reached by the population influence the registered income.

Observe the first and last rows and str of the datasets

Neighborhood	Academic_level	Number	Income	Inhabitants
Baró de Viver	Batxillerat elemental / graduat escolar / ESO / FPI	697	20971.5	2030
Baró de Viver	Batxillerat superior / BUP / COU / FPII / CFGM grau mitjà	279	20971.5	2030
Baró de Viver	Estudis primaris / certificat d'escolaritat / EGB	785	20971.5	2030
Baró de Viver	Estudis universitaris / CFGS grau superior	115	20971.5	2030
Baró de Viver	Sense estudis	154	20971.5	2030
Can Baró	Batxillerat elemental / graduat escolar / ESO / FPI	1817	31002.0	7812

Data Analysis

In this section we explore our datasets. The first descriptive plot shows the relationship between a certain value of income and the number of people earning that income divided by their educational level.



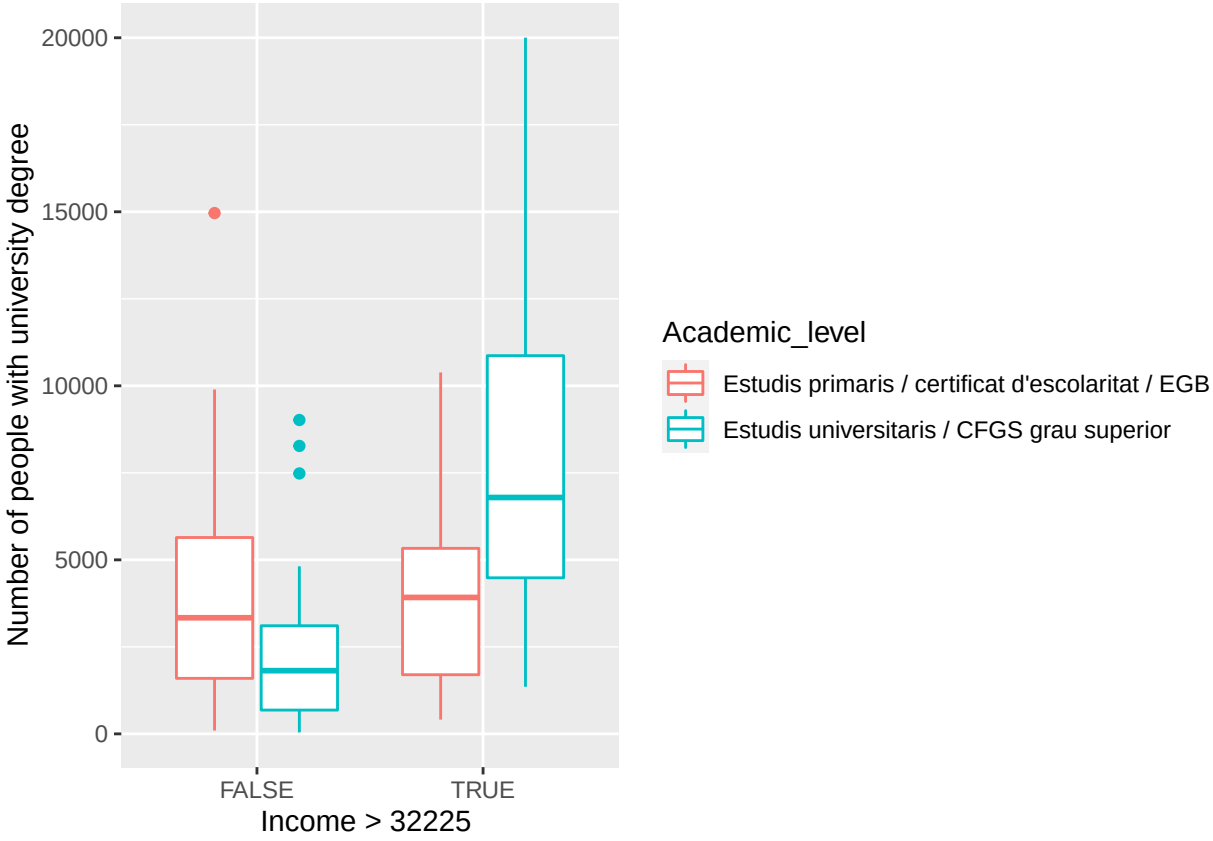
Few observations can be made:

1. As expected, the higher income is mainly earned by a greater number of people with the highest educational level.
2. We can see a crucial point around an income value a bit higher than 32225. That is a pivotal point, where we can see a switch in the preponderant class of people earning the considered income.
3. Lastly we can see how the peaks of the different curves shift towards higher incomes the more we consider higher levels of education. This indicates that, by increasing the educational level, also the income that the majority of the people earn increases.

To further analyze...

To further analyze our findings, by focusing only on the higher and lower educational level, we plot a boxplot to see the big difference around the previously mentioned pivotal point (around income=32225). We chose the value 32225 as it is the median income in Barcelona (all neighborhoods taken into account).

```
## [1] 32225
```

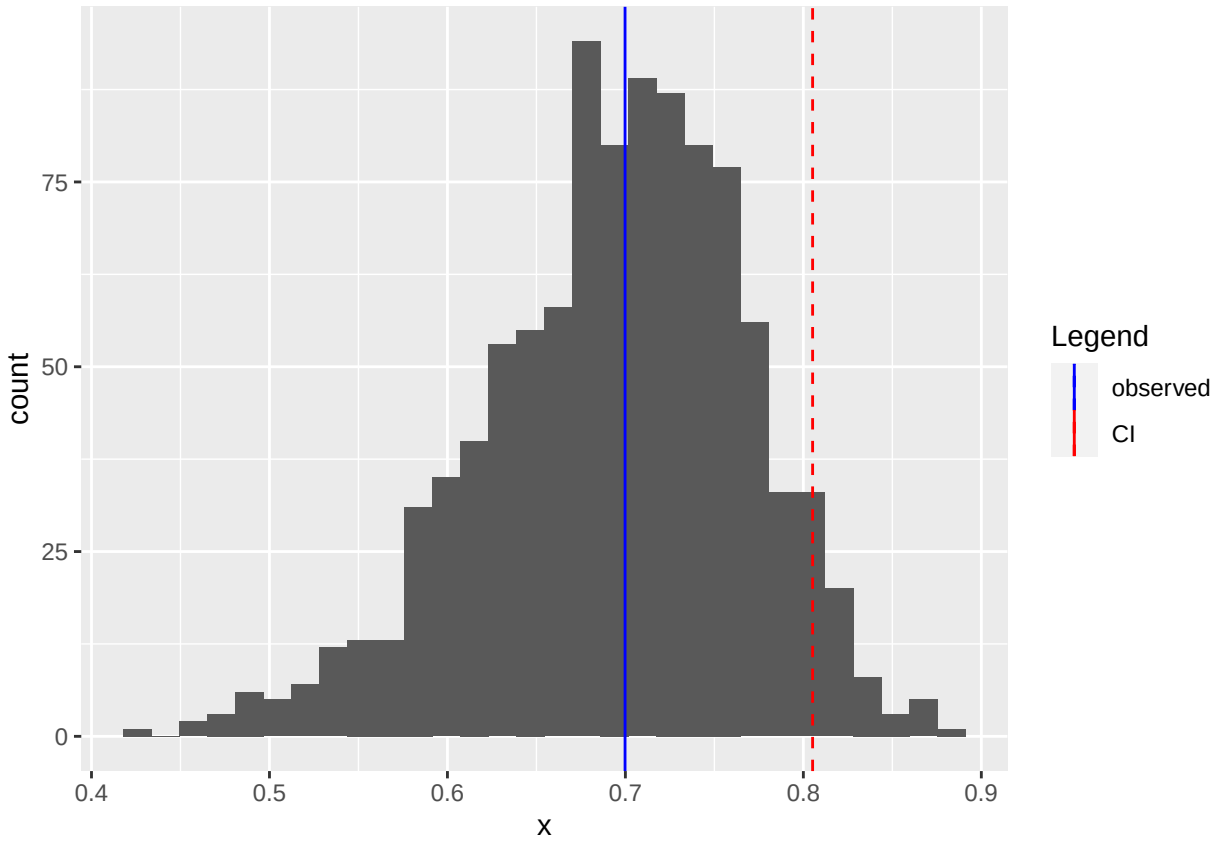


This boxplot shows us that for households with an income lower than 32225 euros a year, they have a lower median of people with low education level than university educated people. For those with an income higher than 32225, it's the opposite, they have more university educated people.

Statistically supported claim and visualize it

We have already seen some evidence about the positive correlation between income and the level of studies. However, how sure can we be that the greater the median income in a district the more people with university degree. A good approach for trying to quantify this concept of how sure is a test hypothesis. We are using a well-known test called Spearman's correlation test. Our null hypothesis is that the Spearman's correlation coefficient is greater than 0, which is true if and only if the correlation between both variables is greater than 0.

Our null hypothesis would be rejected if and only if the confidence interval corresponding to our estimation of the Spearman's correlation coefficient, which is of the form (-infinity,b), b real number, does not intersect with (0,+infinity). As we see in the graph, the intersersect is not empty, therefore we can not reject the null hypothesis.



Introduction of a third variable

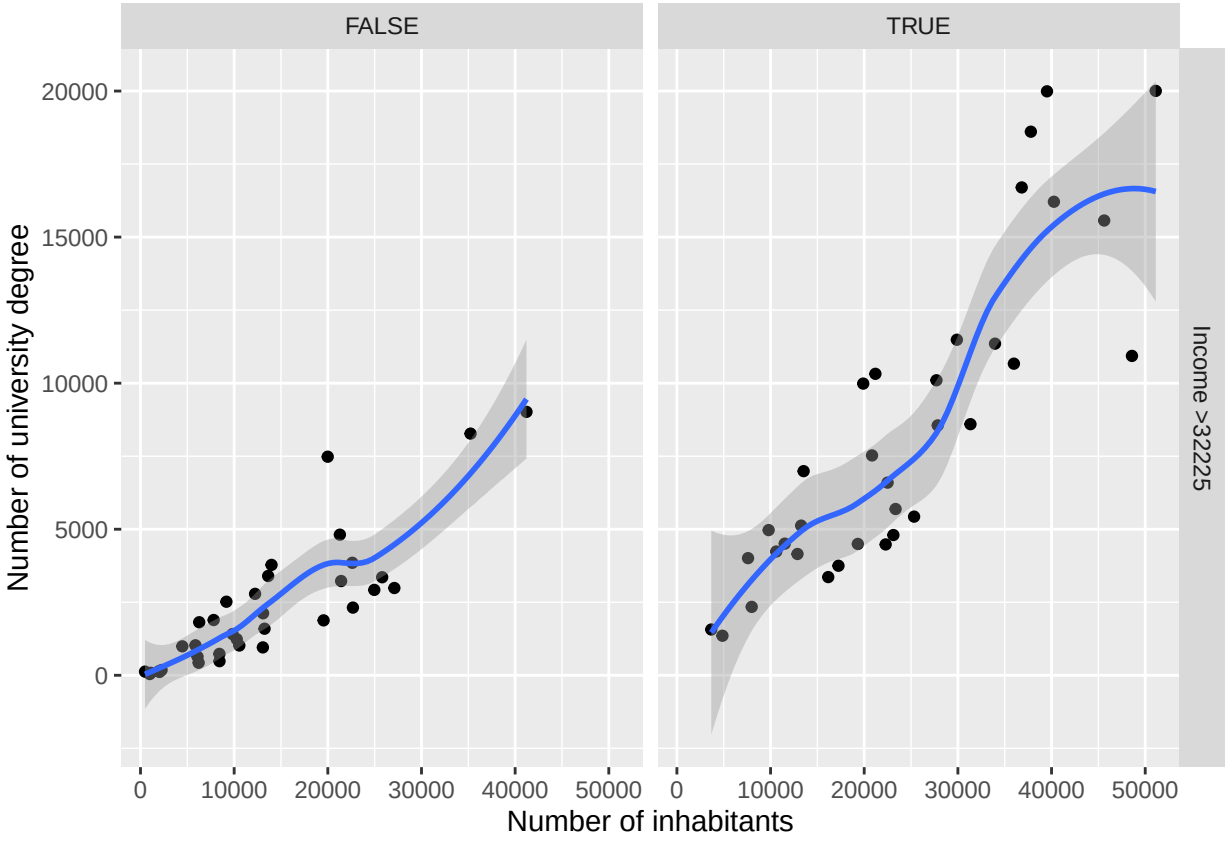
According to those results, the average income highly depends on the number of people with a university diploma. Logically, the 10 neighborhoods with the highest income would correspond to the 10 neighborhoods with the highest number of people with a university degree. To check it, we compute these two top 10, and we verify that we find most of the same neighborhoods in both of them.

Neighborhoods that have the higher income and also have the highest proportion of university-educated people :

```
##                      Neighborhood Number  Income Inhabitants proportion
## 1:          Sant Gervasi - Galvany  19991 58807.0       39513  0.5059348
```

```
## 2: l'Antiga Esquerra de l'Eixample  16701 41875.5       36815  0.4536466
## 3:          la Dreta de l'Eixample  18609 47099.0       37791  0.4924188
## 4:                       les Corts  16210 44693.0       40252  0.4027129
```

We observe that only 4 out of 10 variables are in both top 10. Consequently, we can make the hypothesis that there is another parameter intervening. As our Income data are average within neighborhoods while the number is just a plain number, not taking into account the main difference between neighborhoods, the population. It seems obvious that a neighborhood with 1000 people won't have the same amount of people with a university degree that a neighborhood with 10 people.



We indeed can see that the relation between the number inhabitants and the number of people with a university degree is almost linear. We need to take the population into account.

To address the issue, we remove this "population" parameter and replace it with a less arbitrary parameter: the proportion of people with a university degree and compute once again the Spearman correlation.

```
cor.test(high_degree$Income,high_degree$proportion,method="spearman",alternative = "greater")
```

```
##
##  Spearman's rank correlation rho
##
## data:  high_degree$Income and high_degree$proportion
## S = 5660, p-value < 2.2e-16
## alternative hypothesis: true rho is greater than 0
## sample estimates:
##       rho
## 0.8966021
```

The correlation coefficient is now very high, getting close to 1 with a still very low p-value. We can then check again the top 10 to see if we get better results with the proportion instead of the Inhabitants.

Neighborhoods that have the higher income and also have the highest proportion of university-educated people:

```
##                               Neighborhood Number  Income Inhabitants proportion
##  1:                               Pedralbes   4970 77013.0        9796  0.5073499
##  2:                  Sant Gervasi - Galvany  19991 58807.0       39513  0.5059348
##  3:              Sant Gervasi - la Bonanova  10320 64225.5       21192  0.4869762
##  4:                                  Sarrià   9985 59159.5       19896  0.5018597
##  5: Vallvidrera, el Tibidabo i les Planes   1567 53382.0        3673  0.4266267
##  6:       l'Antiga Esquerra de l'Eixample  16701 41875.5       36815  0.4536466
##  7:                  la Dreta de l'Eixample  18609 47099.0       37791  0.4924188
##  8:          la Vila Olímpica del Poblenou   4010 61253.0        7595  0.5279789
##  9:                               les Corts  16210 44693.0       40252  0.4027129
## 10:                         les Tres Torres   6988 80135.0       13529  0.5165201
```

We now notice that all (10 out of 10) of the neighborhoods with the highest proportion of people with a university degree are also the ones with the highest income.

Conclusion

With the performed analysis we conclude that the higher the level of education, the higher the average income. We have been able to verify it in the example, where the richest neighborhoods are those that have people with the highest academic degrees.

In the first part of the work, "your data exploration including at least one descriptive plot with meaningful descriptions", we have studied the correlation that exists between a certain value of income and the number of people earning that income divided by their educational level. It is here where we can verify that the higher the educational level, the more people earn this income.

Then, in "your data analysis including at least one hypotheses/claim supported by a demonstrative plot and at least one statistically supported claim and its visualization and an example where controlling for confounding factors was necessary to support a claim or invalidate the hypothesis or implement one prediction task and show its performance", we have created the null hypothesis to corroborate what we said before, implying that it is also true.

Finally, we have made an example with real data, taken from the 10 richest neighborhoods and coincides with those with the highest educational level. Therefore, it is fully proven.