Data Analysis and Visualization in R (IN2339)

Case Study

Aditi Arora, Smarth Bakshi, Ishan Bhattacharya, Francisca Andrea Lagos López

2022-01-22

Motivation

Open to the Mediterranean Sea and famous for Gaudí and his modernist architecture, Barcelona is revealed as one of the most important European capitals. The city is a focus of new trends in the world of culture, fashion and gastronomy. It combines the creativity of its artists and designers with respect and care for the traditional places of all time. In it, the charm and pause of its historic center, the avant-garde of its most modern neighborhoods and the urban rhythm of one of the most visited cities in the world coexist.

Notwithstanding the foregoing, is Barcelona a good place to migrate? In the following report, an analysis of the allocation of its inhabitants will be carried out, in particular of the immigrants who have decided to settle there. As Malgesini (2006) states, most of the immigrants settle down in centric urban areas due to the economic, social and cultural opportunities that main cities can offer. Figure 1.2 supports the above showing that l'Eixample -home district of the famous Sagrada Familia church and the Milà and La Pedrera Gaudì houses- is the most popular destination for immigrants to allocate.

Figure 1.1: Total population by districts

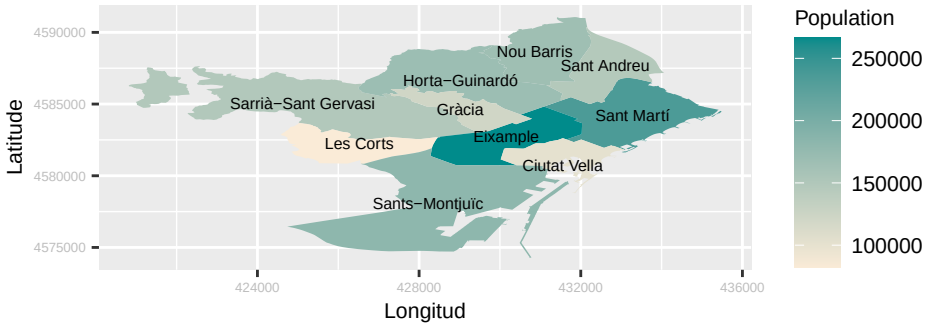
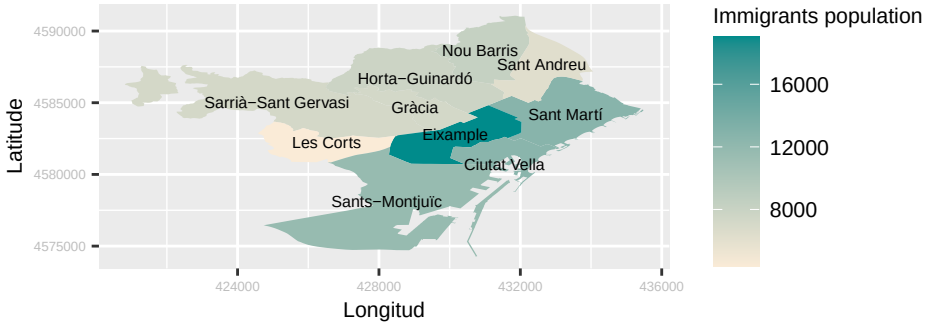


Figure 1.2: Total immigrant population by districts



Literature widely reveals allocation patterns for immigrants taking into account socio demographic variables, like education, salary, age of arrival to the country, sex, local language profency, ethnicity and others. In this case study, the objective is to glimpse if there is any immigrant allocation pattern with respect to the inherent characteristics of each of the 73 neighborhoods of the city of Barcelona, such as connectivity, rate of accidents, unemployment, life expectancy, population, etc. For the above four correlation analyses will be conducted, considering the number of immigrants per neighborhood in 2017 and the different characterizing explanatory variables. Afterwards, the existence of a linear relationship between the dependent variable number of immigrants and the explanatory ones will be tested.

Data Preparation

In this section, the needed data preparation steps required for the analysis is performed. Unnecessary chunk codes are omitted in the compiled pdf-file. First of all, the actual allocation of the immigrants in the city is revealed, for which the "immigrants_by_nationality.csv" file was used. Due to the lack of data and the impossibility to differentiating them, all types of migrations were considered, including the internal one, even when this group represents 36.32% of all migrations data in 2017 (35,354/97,327).

The data preparation steps for the different variables are as follow:

For understanding the connectivity of each neighborhood, the "bus_stop.csv" file was used. After tidying up missing or mistaken values, and deleting duplicate ones, the final count of metro stations and bus stops per neighborhood was calculated. For counting the number of bus stop per area, just the "Day bus stop" and "Night bus stop" variables were considered and sum up in the "dayandnigth_count" column, leaving appart the "Airport bus stop" and "Bus station" ones because they are just a few observations and do not really impact in the daily connectivity of the hood.

The "accidents_2017.csv" file contains the number of accidents, exact location and severity details for 10,339 observations during the year 2017. Since areas with higher population would have higher accidents, the number of accidentes was normalized using the total population of that neighborhood and accidents per 1,000 persons.

The "life_expectancy.csv" file contains life expectancy data for different neighborhoods of Barcelona. For this purpose, the mean life expectancy per neighborhood was taken. To do so, the average of life expectancy over the years 2006-2014 was calculated. Finally, the average of male and female population for each neighborhood was estimated. To achieve the above the melting and casting operations on the original data table were used.

Last but not least, the "unemployment.csv" and "population.csv' ' contains the number of unemployed individuals and population at districts and neighborhood level for the years 2013 - 2017. The immigrant_emigrants_age consists of immigrants data for the years 2015-2017. The registered unemployed data at neighborhood level is extracted and merged with neighborhood's population data for the 2015-2017 time period. The resultant table is then merged with immigrants data at neighborhood level, also the unemployment is then normalized by population and consequently used for analysis and visualization.

Data Analysis

In this section, the analysis on number of bus stops, accidents, number of unemployed people and life expectancy per neighborhood in relation to the immigrant population are conducted. For each variable the p-value using Spearman correlation test is given. The reason why Spearman correlation was used, is because the immigrant count per neighborhood does not follow a Gaussian distribution (Figure 2). For all the tests, the null hypothesis states that there is no correlation between the variable and the immigrant count.

Figure 2: Immigrant population distribution.

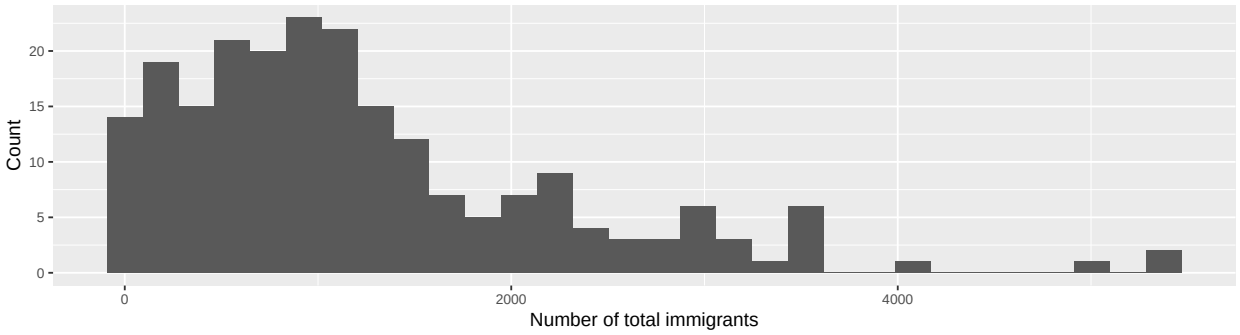


Figure 3.1: Connectivity vs Immigrants

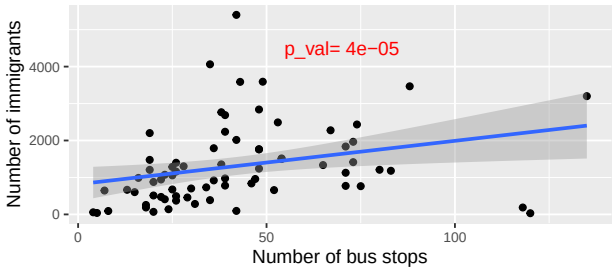


Figure 3.2: Accidents vs Immigrants

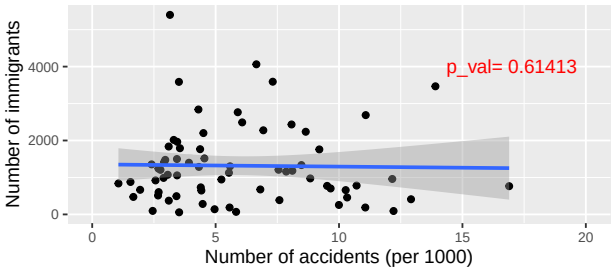


Figure 3.3: Unemployment vs Immigrants

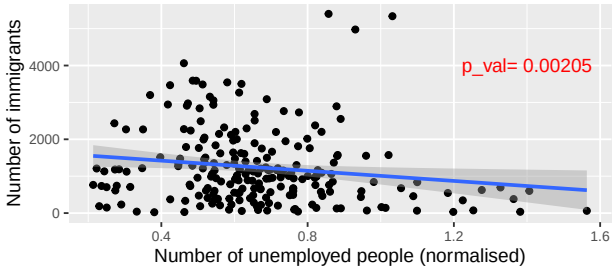
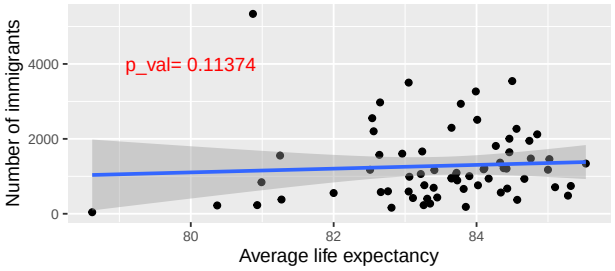


Figure 3.4: Life Expectancy vs Immigrants



From the above four plots a number of observations can be made. Firstly, the number of accidents (Figure 3.2) and average life expectancy (Figure 3.4) of a neighborhood do not have any correlation with the immigrant population. This is confirmed by the correlation tests which return high p-values and hence it is failing to reject the null hypothesis. However, for the number of bus stops (Figure 3.1) and unemployment (Figure 3.3) the relationship is the other way around. This can be confirmed by the low p-values obtained, which means that the null hypothesis can be rejected at a 5% level of confidence, therefore some correlation between the variables exists. In consequence, for these variables further testing to figure out a linear relationship between the number of immigrants in a given neighborhood will be carried out, but first of all the three assumptions of linear regression should be fulfilled.

Figure 4.1: Testing Heteroscedascity for Unemployment Model

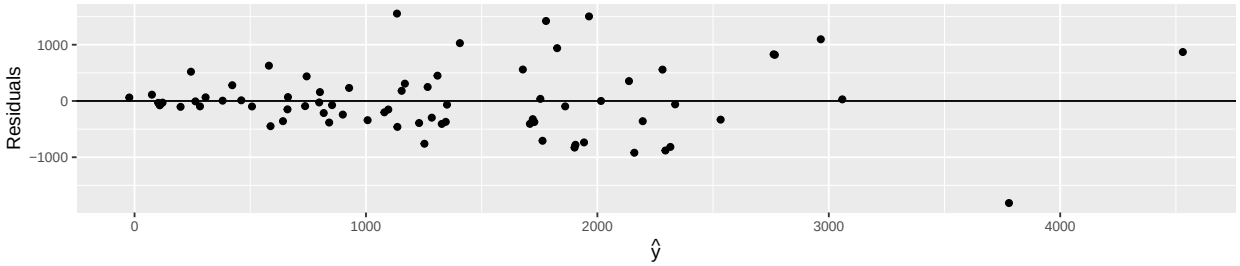
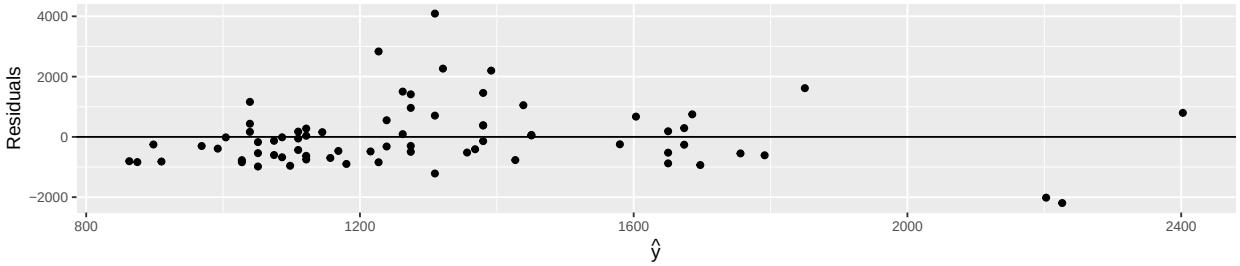


Figure 4.2: Testing Heteroscedascity for Bus Stops Model



For both of the models, the variance of residuals is not constant, violating the assumption that the errors are identically and independently distributed and hence the resultant models will not be very effective. Nevertheless, on going ahead with the linear models the following results were obtained:

```
## Analysis of Variance Table
##
## Model 1: immigrant_count.x ~ dayandnight_count
## Model 2: immigrant_count.x ~ unemployment + dayandnight_count
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     70 75164197
## 2     69 23068968  1  52095229 155.82 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: immigrant_count.x ~ unemployment
## Model 2: immigrant_count.x ~ unemployment + dayandnight_count
##   Res.Df      RSS Df Sum of Sq      F  Pr(>F)
## 1     70 24877107
## 2     69 23068968  1   1808139 5.4082 0.02299 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the current data it appears that the complex Model 2 (immigrant count ~ unemployment + bus stops) works better than the simpler models Model 1 (immigrant count ~ bus stops) for the upper table, and Model 1 (immigrant count ~ unemployment) for the downer one. This is ascertained by the p-values of the ANOVA test which are 1.62e-15 and 0.02 respectively. Hence the null hypothesis can be rejected, which states that the complex model is not better than the models on individuals factors.

Conclusion

From the performed analysis is proceed to conclude that unemployment and bus stops (connectivity measurement) do have some correlation with the number of immigrants allocated in a certain neighborhood. In the case of unemployment the correlation (-0.207) tells that immigrants do tend to live in neighborhoods with better employability. For the bus stops the positive correlation (0.463) is consistent with reality because of the necessity of being easily connected. This would be particularly important for immigrants who rely heavily on public transport as compared to private vehicles.

On verifying the assumptions for linear regression, is noticed that both unemployment and bus stops, do not conform to them strictly. However, since the variation in residuals is not huge, the linear models for these data points were compared and the ANOVA test showed that a model using both of these factors performed better than a simpler model with either one of them.

For accidents is seen that there is not much correlation, negative or positive, with immigrant population. This could imply that immigrants don't necessarily care about these numbers and are more focused on factors like cost of living and connectivity. The assumption for taking into account this variable, was that better life expectancy could relate to better health services or higher safety in a neighborhood. It seems that this is not a contributing factor in the distribution of immigrants in the city of Barcelona.

Some of the shortcomings of the analysis are the lack of data on factors such as cost of living that might have a significant correlation with the immigrant population. In addition, as the literature states, the presence of other immigrants with the same cultural background in the neighborhood would also have been a useful explanatory variable. Regarding the model, an additional linear analysis could be performed using Poisson regression, considering that the dependent variable "immigrant count per neighborhood", possesses such a distribution (Figure 2) and, therefore, this more complex model, in conjunction with the other explanatory variables mentioned above, could provide a better understanding of immigrant allocation patterns.

References

For creating the district maps, shape files for the city of Barcelona were obtained from https://opendata-ajuntament.barcelona.cat/data/en/dataset.

Bartel, Ann P., (1989). "Where Do the New U.S. Immigrants Live?" Journal of Labor Economics, Vol. 7, No. 4, pp. 371-391.

Chiswick, Barry R., Yew Liang Lee and P.W. Miller (2001). "Geographical Concentration Among Immigrants in Australia," Australasian Journal of Regional Studies, Vol. 7, No. 2, 2001, pp.125-150.

Chiswick, B., Miller, P. (2004), "Where Immigrants Settle in the United States", Journal of Labor Economics, Discussion Paper No. 1231.

Malgesini, G. (2006), "Immigrants from Urban to Rural Areas in Spain: The Impact of Transnationalism", online resource, Centro de Investigación para la Paz.