MOVING TO MADRID OR TO BARCELONA?

A DATA-DRIVEN APPROACH

SERGIO HIDALGO

MARCH 27, 2021

TABLE OF CONTENTS

ABSTRACT

Whether you would want to migrate to one of the two main cities of Spain, Madrid or Barcelona, from the same country or from abroad, you may need to take the most documented decision possible.

If we consider a specific migrant ‘persona’ integrated by young people, with no children, who are animal lovers, and look for a natural way of life even in a downtown area, Madrid seems to be a better choice.

In fact, there seems to be always an option for migrants of such kind, regardless of their economic level, to find a neighborhood which may fit their needs. Of course, the higher the salary level the more options available, as well as those having a more ‘natural’ richness and accommodations of bigger size, but options are always available in any case.

INTRODUCTION

BACKGROUND

In Spain there is a regular and historical trend for local people willing to progress moving to one of the two main cities of the country [1], [2]: Madrid or Barcelona. Exactly the same happens to foreign people looking to move to Spain.

In either case, this is mainly based on their huge facilities portfolio, their overall business structure and a rich high-quality jobs offer.

But, where to move? There are countless online and offline resources from which you can pull information about where your dreamed job may be. But little to no options are available to check rationalized and objective information about what you can do in your personal life, when the working time is over, and how to compare those two cities in regards to such aspects of your life.

It is true that some headhunting and talent acquisition companies provide certain degree of advisory about this, but there is no way you can manage to get such information yourself remotely in an easy way.

Taking a blind decision to move to one or another place cannot be just driven by the images of the flashing city lights and the night-life of Madrid, or by the sun and the sea of Barcelona.

PROBLEM

Everybody has their own needs, so we will focus on a given ‘persona’ profile very likely to be a national and/or international migrant. This ‘persona’ profile consists of:

- young people without children [3],
- likely with pets [4],
- and lovers of the natural life [5], [6], [7].

These people usually have a higher-than-average level of spending, which is a driver for local entities to attract such kind of talent too.

We will use data science to select which of the two cities is the most appropriate one for migrants looking for a downtown 'natural way of life'. Later, we will provide options on the most promising neighborhoods based on this criterion, and the economic capabilities of the possible diverse migrants.

INTEREST

The idea of this study is to provide people with a useful tool to remotely take decisions and choose the right city and neighborhood to move to.

But this is not all. This study may be of value for talent acquisition companies, helping them provide a more holistic service to their possible customers.

This study could be also used by the local public entities of the recommended city to leverage their skills to attract high-quality migration.

METHODOLOGY

DATA SOURCES

The following data sources were used to extract the required information:

- Wikipedia sites for districts and neighborhoods’ structure of both Madrid [8] and Barcelona [9] were parsed with the **BeautifulSoup** library,
- Neighborhoods’ coordinates from a mix of Wikipedia and **Geohack** [10] sites,
- **Foursquare** API to extract the most relevant venues per neighborhood,

- .csv format datasets with diverse complementary information like the average building measures by neighborhood coming from sub-municipal information of the Spanish National Statistics Institute (**INE**) [11], and economic information coming from the **INEBase** datasets [12],
- average renting price by m^2 and neighborhood from **Idealista** [13] real state portal,
- and average salaries for a set of alternative jobs in the destination city from **Glassdoor** [14] job-seeking portal.

DATA CLEANING

In order to define the geographic location of both Madrid and Barcelona neighborhoods several tweaks must be done.

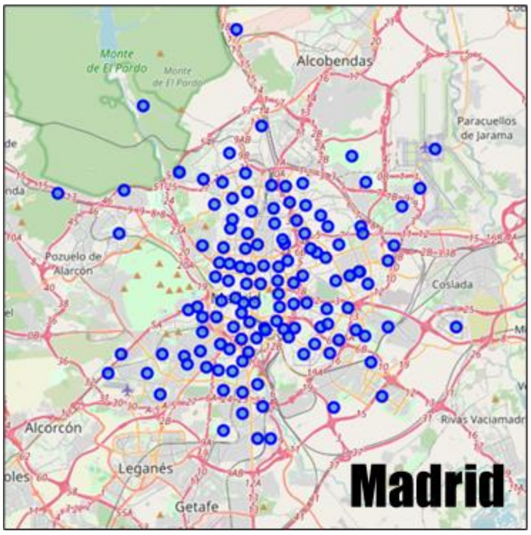
In Spain there is no direct correlation between postal codes and neighborhoods. This is the reason why, if we extract the list of districts and neighborhoods for both cities from any online source, there will be no straight way to correlate those with postal codes and to use any geolocation package as **geopy** or **pgeocode** to return the expected coordinates.

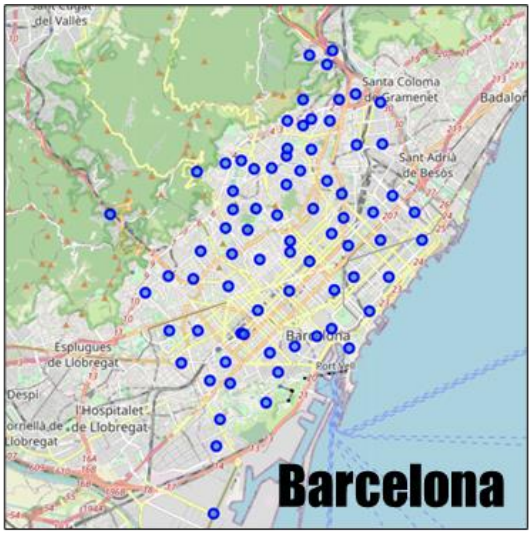
Thus, after parsing the structure of districts and neighborhoods for Madrid and Barcelona (see table below), and cleaning those tables by i.e., exploding into different rows multi-element lists, we needed to run some additional searches, based on the **Geohack** site, and manage to build additional tables correlating neighborhoods and coordinate values. Once these files were generated, they were imported from .csv files as pandas dataframes and merged with the original dataframes generated from the parsed info.

City	**Districts**	**Neighborhoods**
Madrid	21	131
Barcelona	10	75

It is relevant to highlight we needed to do some string substitutions in the names of the districts and neighborhoods. This was made to avoid the Spanish letter 'ñ' and the typical accent marks in both Spanish and Catalan being a barrier for later matching and merging processes.

Once we created the dataframes with their respective locations for each city, we could evaluate the position of their neighborhoods via the **Folium** library.





In order to list the diverse venues per neighborhood in both Madrid and Barcelona we used the **Foursquare** API. We defined 1 km as the radius of search around the neighborhood's center coordinates. This could be somehow too much for some neighborhoods (based on their geometric shape and size), but we intended to be the most exhaustive possible with the venues' extraction process.

To avoid geographic overlapping we used a de-duping process over the full content generated dataframes and just kept one single occurrence of each venue in the datasets. A dataframe with 3356 venues and 290 unique venue categories was built for Madrid and a dataframe with 2582 venues and 285 unique venue categories for Barcelona.

Once we had ready a dataframe for each city with information about the district, neighborhood, venue category and venue, we did a grouping by the first three variables and got cumulated counts for the last.

After that, we generated a list of venues which may be related to 'natural life' and filtered the dataframes based on those, to later aggregate by categories into a fewer amount like:

- healthy restaurants (including i.e., 'Juice Bar', 'Salad Place', etc.)
- healthy shops (including i.e., 'Farmers Market', 'Grocery Store', etc.)
- pet places (including i.e., 'Dog Run', 'Pet Café', 'Pet Store', etc.)
- sports (including i.e., 'Athletics & Sports', 'Basketball Court', etc.)
- and walkways (including i.e., 'Beach', 'Garden', 'Park', etc.).

In this stage of the project, we can get a couple of dataframes per city, only including natural category venues:

- one dataframe including absolute venue counts per natural category,

- and one dataframe grouped by district including natural venue counts for each.

Merging the two dataframes of the same type for each city into one will allow to:

- display graphically the differences between cities
- run statistical analyses to compare inter-group variability vs intra-group variability.

When the decision was made between which city is the most suitable for the 'natural way of life' migrant persona we are targeting at, we needed to generate new dataframes using one-hot encoding processes to match the a-priori requisites of a cluster analysis.

Top 10 venue categories per neighborhood information was consolidated into a dataframe with a record for every neighborhood in the city. This dataframe was used for the clustering analysis and its structure was preserved when the filter by cluster was executed later on.

Finally, from the shortlist of 10 possible neighborhoods with a more suitable profile for our analysis, we generated a dataframe which was merged with info from other social-economic sources. Thus, by including information on the average flat size in m^2 and the average renting price by m^2 in each neighborhood we were able to create a calculated column with the average renting price (in EUR) per neighborhood in the destination city.

ANALYSES DESCRIPTION

Once our datasets were correctly built, we proceeded with the required analyses to lead to a documented conclusion.

As a first approach we generated some visual displays comparing the number of 'natural' venues per category and city, and the variability of 'natural' venues by district and city. For this purpose, we used both **matplotlib** and **seaborn** libraries.

To understand which was the most suitable of the two cities based on possible statistical differences, we run a t-test using the **scipy** library over the 'natural' venues by district dataframes. In order to understand which type of t-test was required in this case, we needed to run some exploratory statistics with the **statistics** library and obtain the mean and variance for each group.

When the decision was made about which was the best candidate city, it was the time to funnel down the process and look to provide a shortlist of neighborhoods to the possible migrants with the best 'natural' venues score possible and a renting price at hand of their economic capacity.

In order to aggregate neighborhoods by common venues profiles we used a **k-means** clustering process run with **SciKit-learn** library. As a pre-requisite of this analysis, it is required to define the number of clusters (k). To get that number we iterated over a clustering process the same dataset with a range of k values going from 1 till 6, and we found that the most balanced results on the number of clusters and the distribution of neighborhoods by cluster appeared when k = 4.

With a visual inspection of resulting clusters, we defined the one with an overall presence of ‘natural’ venues within the top 10 venues per neighborhood which seemed to be the biggest.

Since the chosen cluster had still many neighborhoods included and could not provide a straightforward recommendation pattern, we proceeded to generate a shortlist of the top 10 ones within such cluster which were ranking higher for more 'natural' venues within their top venues.

Finally, a set of 3 salary ranges was established based on possible jobs of the migrants when in destination, and diverse scenarios on possible neighborhood selection were built based on their economic capacity (taking into account it is never recommended to spend more than a 40% of the monthly gross salary in accommodation).

RESULTS

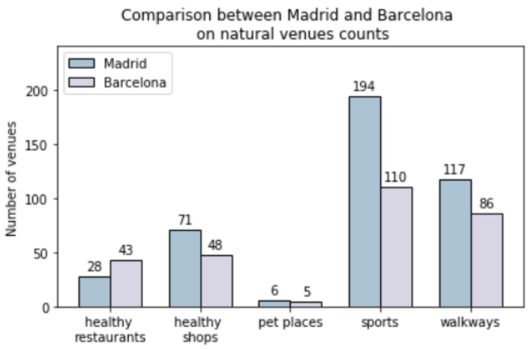
CHOOSING THE RIGHT CITY

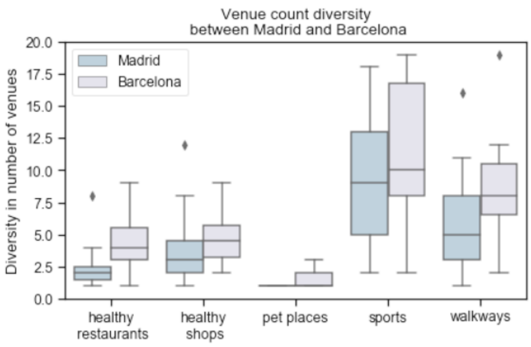
Based on the definition of our problem, the main factors to decide which of the two cities was the most suitable destination were:

- which one had the biggest number of venues linked to ‘natural way of life’ (healthy restaurants and shops, sport areas, natural walkways, etc.)
- which one had those ‘natural’ venues more equally distributed across neighborhoods, providing a more consistent offer to the migrant.

Madrid showed to have a bigger number of ‘natural’ venues regardless of the category (except for healthy restaurants, where Barcelona led), with a total of 416 venues vs 292 in Barcelona.

By looking at a visual representation like a box plot, we were also able to see that Barcelona had a bigger variability in the distribution of ‘natural’ venues amongst districts than Madrid. This means that there were districts of Barcelona with a high concentration of ‘natural’ venues whilst others did not have so many. In Madrid the distribution of such venues seemed to be more homogeneous.





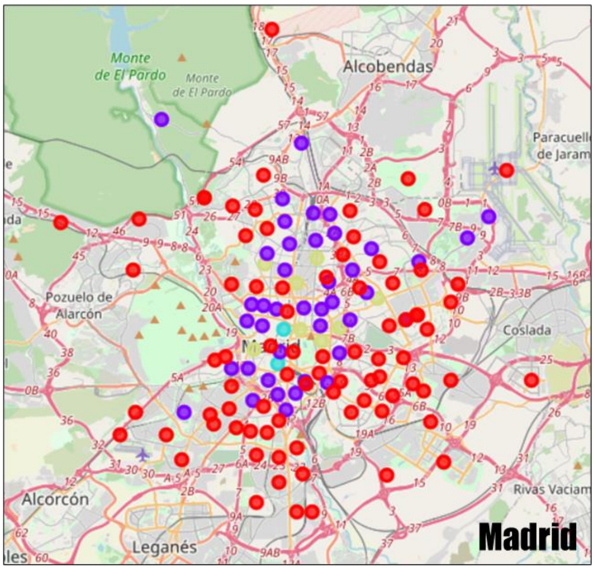
When inspecting some descriptive statistics for the two cities, we saw complete unequal groups, with different sample sizes, means and variances:

City	Districts	Mean	Variance
Madrid	21	19.81	119.36
Barcelona	10	29.20	131.51

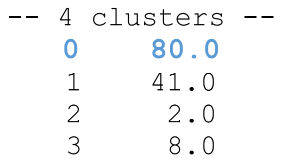
When we run a t-test for unequal samples we found that there was a statistical difference between Madrid and Barcelona in the amount and distribution of ‘natural’ venues amongst districts (t = -2.16, p = 0.04).

CHOOSING THE RIGHT NEIGHBORHOOD

Once Madrid was chosen as the preferred destination city, neighborhoods of similar characteristics were clustered in four groups during the k-means process:

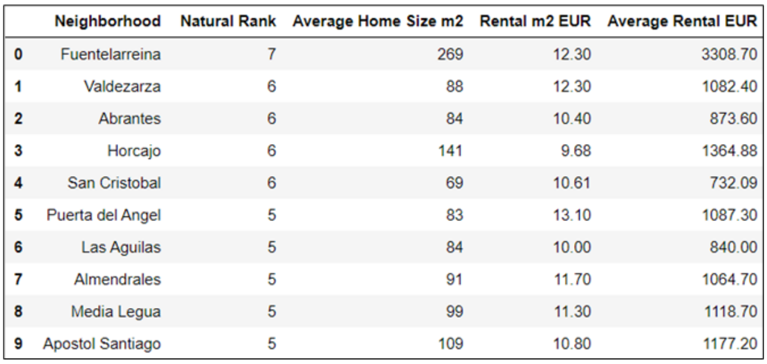


Despite the k-means process run with k = 4 was the one offering the most balanced outcome, since we found that the most representative cluster of the kind of 'natural way of life' neighborhoods we were looking for was the first one, which was still accounting for 80 out of the 131 neighborhoods in Madrid, we needed to think on a way to shortlist the final number of options to 10.



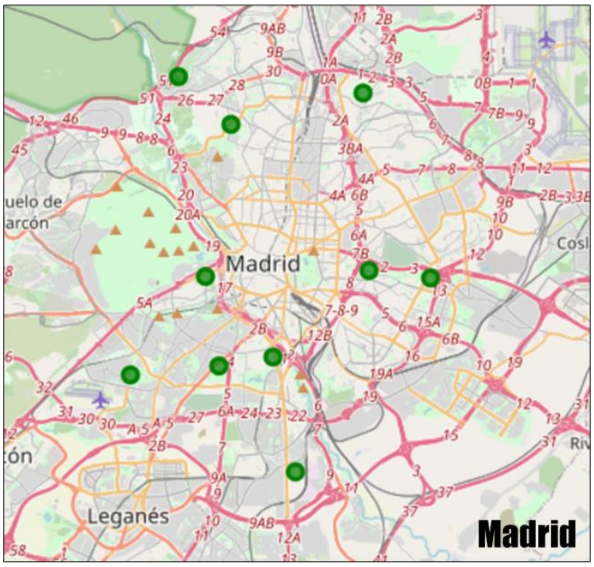
```
-- 4 clusters --
   0    80.0
   1    41.0
   2     2.0
   3     8.0
```

When we assigned a Natural Rank to every neighborhood of the cluster and we sorted those by the value of such rank, we were able to identify and split a dataset for the top 10 most 'natural' neighborhoods in Madrid.



	Neighborhood	Natural Rank	Average Home Size m2	Rental m2 EUR	Average Rental EUR
0	Fuentelarreina	7	269	12.30	3308.70
1	Valdezarza	6	88	12.30	1082.40
2	Abrantes	6	84	10.40	873.60
3	Horcajo	6	141	9.68	1364.88
4	San Cristobal	6	69	10.61	732.09
5	Puerta del Angel	5	83	13.10	1087.30
6	Las Aguilas	5	84	10.00	840.00
7	Almendrales	5	91	11.70	1064.70
8	Media Legua	5	99	11.30	1118.70
9	Apostol Santiago	5	109	10.80	1177.20

Surprisingly, when we plotted the top 10 options over a map, no specific geographic concentration was found.

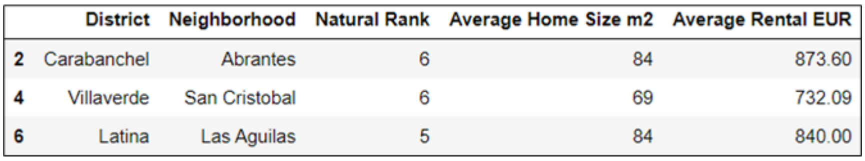


Finally, with the top 10 neighborhoods set defined, we generated 3 alternative and exclusive scenarios for migrant earning a yearly gross salary of:

- EUR20k – EUR30k,
- EUR30k – EUR40k,
- and EUR40k – EUR50k.

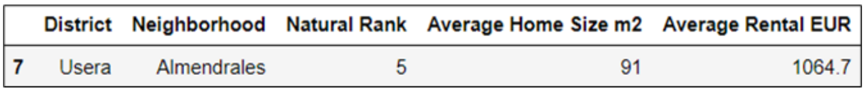
In neither case, the neighborhood Fuentelarreina, in the North, with the highest Natural Score (7) but a monthly average renting price of EUR3.3k, was included on the final propositions.

For migrants with a yearly gross salary of EUR20k – EUR30k three possible options could be offered, all of them in the South and South-West areas of Madrid.



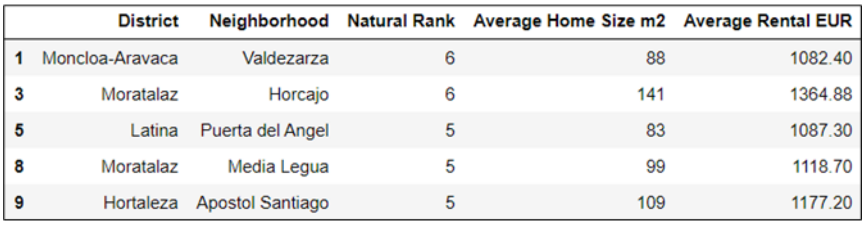
	District	Neighborhood	Natural Rank	Average Home Size m2	Average Rental EUR
2	Carabanchel	Abrantes	6	84	873.60
4	Villaverde	San Cristobal	6	69	732.09
6	Latina	Las Aguilas	5	84	840.00

Only one option was proposed as new one for migrants earning a yearly gross salary of EUR30k – EUR40k. This was also a neighborhood in the South of Madrid.



	District	Neighborhood	Natural Rank	Average Home Size m2	Average Rental EUR
7	Usera	Almendrales	5	91	1064.7

For migrants earning a yearly gross salary of EUR40k – EUR50k a richer proposition set was offered, not only on the number of options but also in the placement of such options, located in some of the best areas of Madrid downtown.



	District	Neighborhood	Natural Rank	Average Home Size m2	Average Rental EUR
1	Moncloa-Aravaca	Valdezarza	6	88	1082.40
3	Moratalaz	Horcajo	6	141	1364.88
5	Latina	Puerta del Angel	5	83	1087.30
8	Moratalaz	Media Legua	5	99	1118.70
9	Hortaleza	Apostol Santiago	5	109	1177.20

DISCUSSION

Our analysis showed that Madrid is a better city of choice in Spain for migrants looking for an environment which is full of venues and places where to live following a more 'natural' way of living, even if doing so in the downtown area.

This was proved statistically, being also noticeable that Madrid has more 'natural' venues under all the expected categories (except restaurants), and also that these 'natural' venues have a more homogeneous distribution across the diverse areas of the city there than in Barcelona.

But choosing one city or the other was only one of the targets of this analysis. After getting Madrid as the suitable candidate city, we followed a funneling process to first determine via a k-means clustering analysis which was the best candidate pool of neighborhoods to migrate to within the city, to later go more granular into where to really land based on the migrants' economic level.

We determined 4 as the best and more balanced number of clusters to define a priori during the process. From those, cluster 1, with a big number of neighborhoods included (80 out of the total 131 neighborhoods in Madrid), seemed the one with a most homogeneous and natural-driven profile of areas included.

Anyway, as said, having 80 elements in this group would not help much with doing suitable recommendation. For this reason, we needed to shortlist those to a more manageable amount of top 10 more 'natural' neighborhoods. Surprisingly, there was no specific geographic concentration of these neighborhoods in any given area of the city, with the best candidates being spread all across Madrid downtown.

From those top 10 candidate neighborhoods we found that one, Fuentelarreina, in the North, with huge average house sizes and renting prices, was completely out of scope of any migrant with annual gross salaries below EUR50k. The other 9 candidate neighborhoods appeared in any of the 3 alternative (and exclusive) proposition groups based on salary ranges.

Of course, being that said, for example, someone in the salary range 3 (between EUR40k and EUR50k yearly), like a Doctor or an Architect, would likely choose as a preference to live in the neighborhood of Horcajo. That person might be willing to spend money on accommodation in the upper bound of the range, but would be getting in return the highest possible value of 'natural' likeliness (6) and the average biggest houses (141 m^2) to live in.

CONCLUSSIONS

The purpose of this project was to help possible migrant candidates (under a specific 'natural way of life' persona group) to Spain to choose the best city for them, whether being Madrid or Barcelona.

Once the best candidate city was chosen (herein Madrid), we tried to provide the most accurate recommendations possible to narrow down the destination neighborhoods based on their 'natural life' profile likeliness. We also framed alternative scenarios based on the migrants' economic capacity, so that they could choose to spend a fair amount on accommodation while still having the kind of life they would be looking for.

Of course, this is just a help for migrants to make a documented choice beyond just using a possible job opportunity to decide where to move to. Many other factors like proximity to job place, access to major roads, levels of noise, contamination levels, etc., would also need to be taken into account, and these may need to be handled by the own migrants when trying to make the final decision.

What’s next? This is the kind of question which may come out after reading about this project.

There is a chance to promote this kind of analysis between local public entities looking to attract high-value immigration based on specific local assets they could use in marketing campaigns.

Also headhunting and talent acquisition companies may be interested on including this kind of information on their portfolios of services and differentiate themselves from the rest in the marketplace.

Finally, a follow-up broader project, incorporating other ‘persona’ profiles, could provide a complete backend solution for a web portal, whose activity could be monetized via marketing from the own incumbent venues promoted from the dataset.

REFERENCES

[1] **El Confidencial:** https://www.elconfidencial.com/economia/2019-09-27/exodo-urbano-espana-llegadas-madrid-ciudades_2240155/

[2] **ABC:** https://www.abc.es/sociedad/abci-espanoles-espana-lugar-origen-mas-frecuente-inmigracion-interna-cada-provincia-201901310239_noticia.html?ref=https:%2F%2Fwww.google.com%2F

[3] **World Population Review:** https://worldpopulationreview.com/country-rankings/birth-rate-by-country

[4] **PetSecure:** https://www.petsecure.com.au/pet-care/a-guide-to-worldwide-pet-ownership/

[5] **Eurostat - European Commission**: https://ec.europa.eu/eurostat/statistics-explained/index.php/Statistics_on_sport_participation

[6] **Wikipedia:** https://en.wikipedia.org/wiki/Vegetarianism_by_country

[7] **Boomberg:** https://www.bloomberg.com/news/articles/2019-02-24/spain-tops-italy-as-world-s-healthiest-nation-while-u-s-slips

[8] **Wikipedia:** https://en.wikipedia.org/wiki/List_of_neighborhoods_of_Madrid

[9] **Wikipedia:** https://en.wikipedia.org/wiki/Districts_of_Barcelona

[10] **Geohack:** https://geohack.toolforge.org/geohack.php?pagename=Corralejos&params=40_27_52_N_3_35_24_W_type:city(150000)_region:ES

[11] **INE - sub-municipal data:** https://www.ine.es/jaxiT3/Tabla.htm?t=30139

[12] **INEBase - economy:** https://www.ine.es/dyngs/INEbase/es/categoria.htm?c=Estadistica_P&cid=1254735570541

[13] **Idealista:** https://www.glassdoor.es/index.htm

[14] **Glassdoor:** https://www.glassdoor.es/index.htm

ACKNOWLEDGEMENTS

I would like to thank to my wife, Laura, who has provided some very interesting ideas along the way, and to whole Stack Overflow community, as any time I got stacked with the code I was able to find a way out amongst the countless tips shared there.