Analysing store locations

Finding optimal location for a low-cost market in Madrid

Emilio Macias

December 23, 2020

Introduction

Background

Opening a physical store is not without risks. One the most obvious risks to evaluate before opening a new store is receiving enough customers to make the business profitable. Therefore it is essential to pick the right location to make sure it is convenient for the customers and there is enough demand for a new store in the area.

Identifying the target customers is also key taking into account the behaviour of the different groups of consumers and the stores they usually buy at.

Business problem

In this project, we tackle the problem of a low-cost supermarket chain trying to decide in which area of Madrid (Spain) they should open their new store in order to maximise the revenue. It is important to note that the city of Madrid consists of 21 districts and 131 neighbourhoods with great differences between them.

The goal is to identify the optimal neighbourhood for opening a store taking different factors into consideration such as the types of neighbourhood (a residential area would be ideal), the amount of people living in those areas (the higher the population the higher the food demand), their average income (working class people are preferred) and the stores that are already available (avoiding areas with a big density of supermarkets).

Using data science, geospatial analysis and machine learning techniques, this project aims to provide a solution for this problem and recommending the best neighbourhood for opening the low-cost supermarket.

Data

The following sections describe the data that is needed for answering this business question.

Wikipedia

The first data that we need is the list of neighbourhoods in Madrid. Even though this information could have been directly obtained from a CSV file from Madrid city council, it has been decided to use web scraping on Wikipedia for learning purposes.

The Wikipedia page “List of neighbourhoods of Madrid” **[1]** shows a table with the name of each neighbourhood for each of the 21 districts. In our project, we will work directly with the neighbourhoods and ignore the districts since this way we can perform a more granular analysis of the areas.

Geospatial data

Since the plan is to target residential areas, we need to analyse the type of food venues present in each neighbourhood. With the Foursquare API **[2]** we can explore the different food venues, considering that a big density of bars and restaurants over very few supermarkets will most likely refer to a business or recreational area where people don’t usually buy at supermarkets. In the other hand, a large proportion of supermarkets over the rest of food venues might indicate it is a residential area where people normally make their food shopping.

Before we can make use of the Foursquare API we need to convert the neighbourhood names into a pair of latitude and longitude coordinates. We can query the Foursquare API using the HTTP GET method on the explore endpoint indicating the geographical coordinates, venue categories and radius.

The following figure shows the location of all the neighbourhoods within the city of Madrid, it is important to remark that any areas and towns outside the city have not been considered for this project.

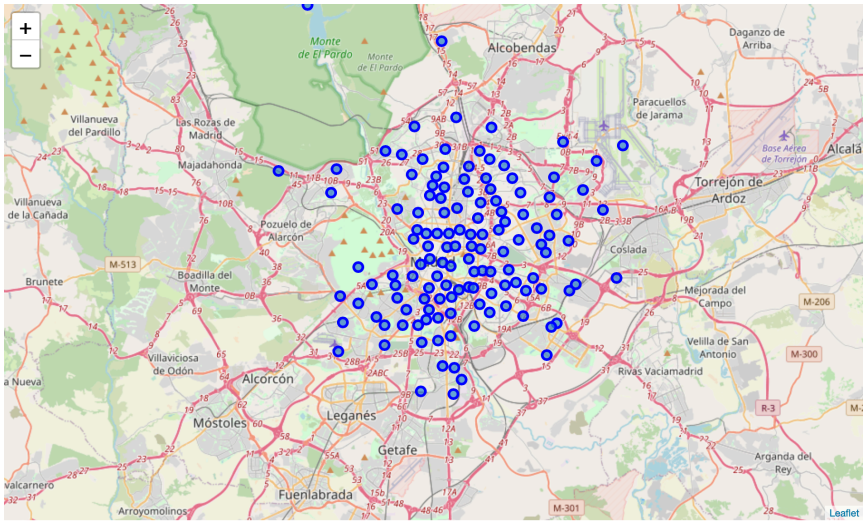

Fig 1: Madrid neighbourhoods

Census data

Finally we will need data from the census of Madrid. We can obtained this data from Excel files that are accessible from the Madrid city council website. Particularly we are interested in the population **[3]** and the average income **[4]** of each neighbourhood.

Below we can see the type of information that is required for this analysis.

	Neighbourhood	Latitude	Longitude	Income	Population
0	PALACIO	40.41517	-3.71273	34675	23691.0
1	EMBAJADORES	40.40803	-3.70067	25999	47460.0
2	UNIVERSIDAD	40.42565	-3.70726	30701	33527.0
3	ACACIAS	40.40137	-3.70669	44669	36690.0
4	CHOPERA	40.39536	-3.69833	31933	20241.0

Table 1: Geographical and census data

Methodology

Neighbourhood segmentation

Since initially we don't know how many different types of neighbourhood we can find in Madrid, we are using the elbow method to obtain the optimal number (k) of clusters.

Although the figure below shows 3 as the optimal number of clusters, we are using 5 (the second best k) since this way we can break down more the number of neighbourhoods that we are going to analyse against the census data.

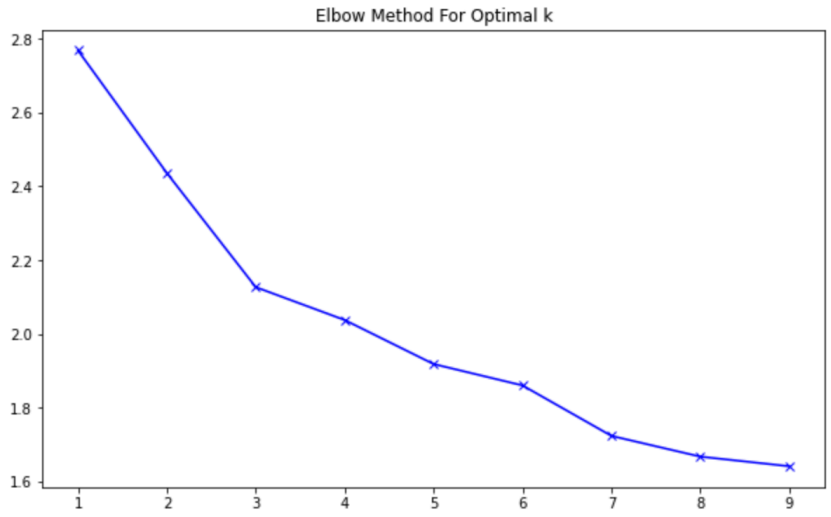

Fig 2: Elbow method for the optimal number of clusters

With K-means algorithm we can group the 131 neighbourhoods into 5 clusters depending on the most popular venues in those areas. Below we can visualise what are the most common venues and the cluster (from 0 to 4) that has been assigned to each area.

	Neighbourhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	PALACIO	40.41517	-3.71273	2	Church	Restaurant	Spanish Restaurant	Mexican Restaurant	Plaza	Monument / Landmark	Historic Site	Office	Korean Restaurant	Furniture / Home Store
1	EMBAJADORES	40.40803	-3.70067	2	Spanish Restaurant	Tapas Restaurant	Bar	Indian Restaurant	Restaurant	Salon / Barbershop	Coffee Shop	Food & Drink Shop	Theater	Pizza Place
2	CORTES	40.41589	-3.69636	0	Spanish Restaurant	Office	Restaurant	Government Building	Tapas Restaurant	Capitol Building	Hotel	Japanese Restaurant	Hotel Bar	Building
3	JUSTICIA	40.42479	-3.69308	1	Art Gallery	Office	Courthouse	Boutique	Food	Restaurant	Embassy / Consulate	Building	Women's Store	Post Office
4	UNIVERSIDAD	40.42565	-3.70726	2	Café	Bookstore	Bar	Government Building	Office	Hotel	Bed & Breakfast	Coworking Space	Mexican Restaurant	Metro Station

Table 2: Neighbourhood clusters and most popular venues

In order to name the different clusters it is necessary to explore the most representative venues in the neighbourhoods of each cluster. Finally we come with the following representation:

Cluster	Type of area	Most popular venues
0	Recreation	Restaurants, bars, shops
1	Business	Offices, co-working spaces
2	Residential	Bars, barber shops, banks
3	Airport	Airport gates, rental cars
4	Student	Universities, colleges, libraries

Table 3: Types of neighbourhood

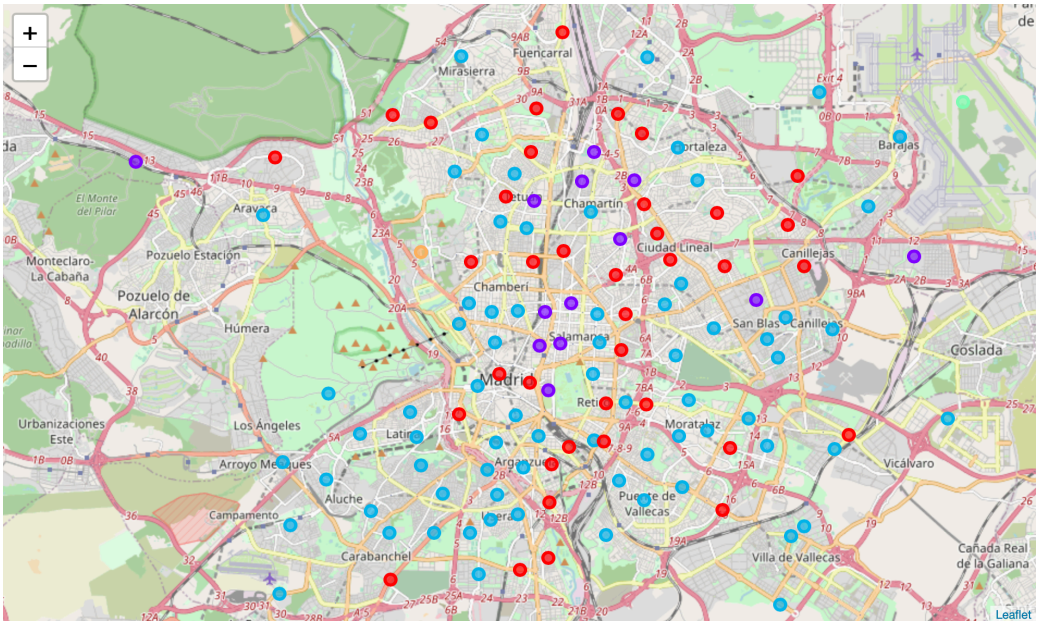

Fig 3: Map of Madrid with different clusters of neighbourhoods

Therefore we will analyse the cluster 2 (light blue in the map above) in detail since this is the one that refers to residential areas. This cluster contains 76 neighbourhoods.

Census analysis

Merging the geographical data of the residential neighbourhoods together with the census data we can first count the number of markets that exist in each neighbourhood and also the ratio of people per market. This will help us identify the neighbourhoods where the offer of supermarkets is not large enough to satisfy the demand of the population.

	Neighbourhood	Latitude	Longitude	Number of Markets	Income	Population	People per Market
0	PALACIO	40.41517	-3.71273	33	34675	23691.0	717.0
1	EMBAJADORES	40.40803	-3.70067	34	25999	47460.0	1395.0
2	UNIVERSIDAD	40.42565	-3.70726	49	30701	33527.0	684.0
3	ACACIAS	40.40137	-3.70669	21	44669	36690.0	1747.0
4	CHOPERA	40.39536	-3.69833	20	31933	20241.0	1012.0

Table 4: Geographical, census and market data for residential neighbourhoods

Results and discussion

The following scatterplot represents the residential neighbourhoods considering the ratio people — markets and the household income. The highlighted quadrant covers neighbourhoods with a household income below average (since we are targeting working class areas) and a ratio of people per market above average, making these neighbourhoods good candidates for the location of the new store.

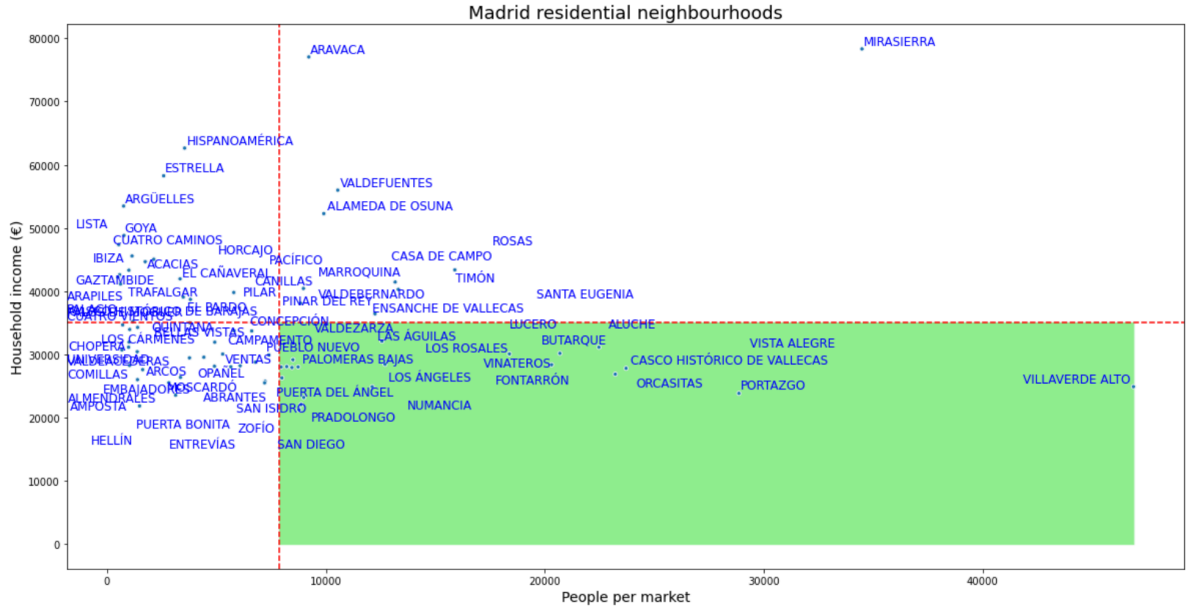

Fig 4: Residential areas: household income vs ratio people per market

Above we can clearly see that Villaverde Alto looks the optimal neighbourhood since the household income is clearly under average (around 26K €) and there are very few supermarkets for the amount of population in the area.

However it would also be recommended to consider other similar areas such as Portazgo that is also located on the bottom-right corner of the plot.

It is also interesting to mention that even though the low amount of existing supermarkets in Mirasierra does clearly not cover the demand of the neighbourhood, it would be a convenient place for a high class market since the household income there is at the very top.

Conclusion

The neighbourhoods of Madrid were analysed with the purpose of finding the ideal location for a new low-cost supermarket. We applied machine learning techniques such as k-means clustering to find different clusters so that we could focus in only one type of neighbourhood (residential). Further data such as population and market venues have been used to reduce the number of potential areas.

This project could be improved by only taking certain venue categories into consideration when performing the clustering segmentation. We could for example identify the key types of venue that define a residential area such as schools, pharmacies, small markets and corner shops, and the types of venue that discard a residential area such as night clubs, theatres and so on. Another improvement could be achieved by only handling certain groups of ages and social classes that would normally shop in a supermarket.

Although this project focuses particularly in a low-cost supermarket, it could easily be amended for any type of business and city, as long as the corresponding data are available to be included in the analysis.

References

[1] Wikipedia article: https://en.wikipedia.org/wiki/List_of_neighborhoods_of_Madrid

[2] Foursquare API: https://developer.foursquare.com

[3] Madrid city council (population data): http://www-2.munimadrid.es/TSE6/control/seleccionDatosBarrio

[4] Madrid city council (economic data): https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=d029ed1e80d38610VgnVCM2000001f4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default