# BUSINESS EMPLACEMENT LOCATOR

## IBM Data Science Professional Certificate

Abstract

The right location for a physical business can make the difference between its failure and success. For an entrepreneur interested in finding the right location in an unknown city, this can become a big challenge.

The following document presents a machine-learning-based tool designed to explore the different venues across a target city and find those areas with the highest business potential.

Antonio Sanchez Calderon

Antonio.scr@outlook.com

## Table of contents

## Table of figures

# 1. Introduction

## 1.1 Scope

The right location for physical business can make the difference between its success and its failure. It is well known for companies and entrepreneurs when expanding their business in a new city the challenge that represents finding the right emplacement for their business.

The scope of this project is to provide recommendations about which areas are more valuable for a specific business to be located. From there, it's up to the user to continue with the exploration on other aspects, such as real state pricing or specific existing regulations, to find the best one.

## 1.2 Use case

In this project, the focus will be on finding the best possible locations for a social fast food restaurant, specialized in small dishes. In other words: a tapas bar. The location chosen as a target will be Madrid city, but due to the flexible logic applied in it, it could also be extended to other type of businesses and locations.

## 1.3 Target group

This use case is especially relevant to those companies and entrepreneurs looking to expand their business or interested on making their entry into a city to which they are not familiarized with.

# 2. Resources

## 2.1 Data sources & availability

For the location and use case specified for this project, the data sources will be the following:

- Wikipedia: Via the use of the *BeautifulSoup* repository, information about the different districts of Madrid city will be fetched for later exploration, such as name, area or population density (see Fig. 2).
- Foursquare: This location data platform provides an API to perform queries and fetch useful information about different venues, such as coordinates or categories (see Fig. 4).

# 3. Methodology

## 3.1 Process methodology

The principle to be followed in order to identify the best areas for the new business emplacement will be quite simple, based on simply weighting pros and cons:

- Two lists of venue categories will be prepared by the user according to the nature of the business:

  - **Source venues**: This list includes those location categories that attract people to them, and therefore, increase the number of potential customers. In this case the categories chosen for this list are "Arts & Entertainment", "Music Venue", "Shopping Mall" and "Stadium", but of course, additional ones could be added.

  - **Sink venues**: It will contain those categories that attract similar customers to ours, causing a drop in the number of potential customers. In other words, direct competitors or extensive green areas that reduce the population and customer density. The list will be made of venues that fit one of the following categories from Foursquare: "Bar", "Tapas Restaurant", "Fast Food Restaurant" and "Irish Pub".

- Cluster of sources will be identified through different machine learning methods (see section 3.2) as areas of interest.

- Each area of interest will be evaluated by searching the total source and sink venues. Its value will be based on the subtraction of the total counts of each search.

- The cluster showing a positive value will then be presented in green, while those with negative values will be displayed in red.

### 3.2 Machine learning methods

According to the methodology of the analysis, two algorithms stand out, both belonging to the unsupervised learning category, with different characteristics and applications:

- **K-Means clustering:** Especially useful if the user already knows how many emplacements, they would like to set up to maximize their coverage, or simply play around with the number of recommendations.

- **DBSCAN**: This methodology will identify high density areas with high value potential. The number of areas recommended may vary depending on the hyperparameters chosen and how restrictive they are. Clusters obtained with this method are more precise and less sensitive to outliers, but the irregular shape of them makes them more challenging to work with.

## 4. Analysis and results

### 4.1 Exploratory analysis

The first step is to get familiarised with the data that we will be using for the analysis. As mentioned in the section 2.1, the sources will be essentially two: Wikipedia to extract the district information and Foursquare for the venue data.

The reason why the exploration will be done via the districts, is due to the maximum number of venues that can be fetched from Foursquare at a time, which is limited to 50. A request for each location category will be conducted for each district. Since each district has an arbitrary shape and the searches can only be performed in a circular format, a big radius (1.5km) has been defined. Afterwards, the duplicated will be removed, in case some searches had elements overlapping.

This type of exploration may lead to blinds spots depending on the sizes of the districts (see Fig. 1), which will have to be taken into consideration once the first results are obtained.

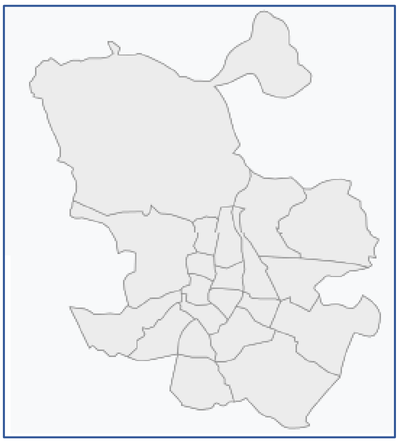
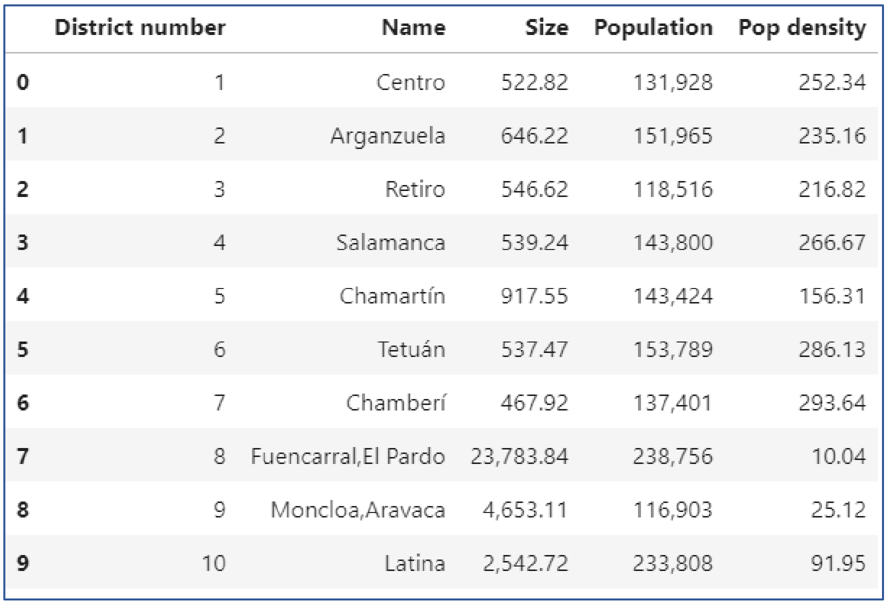


***Fig. 1.*** *Districts of Madrid*

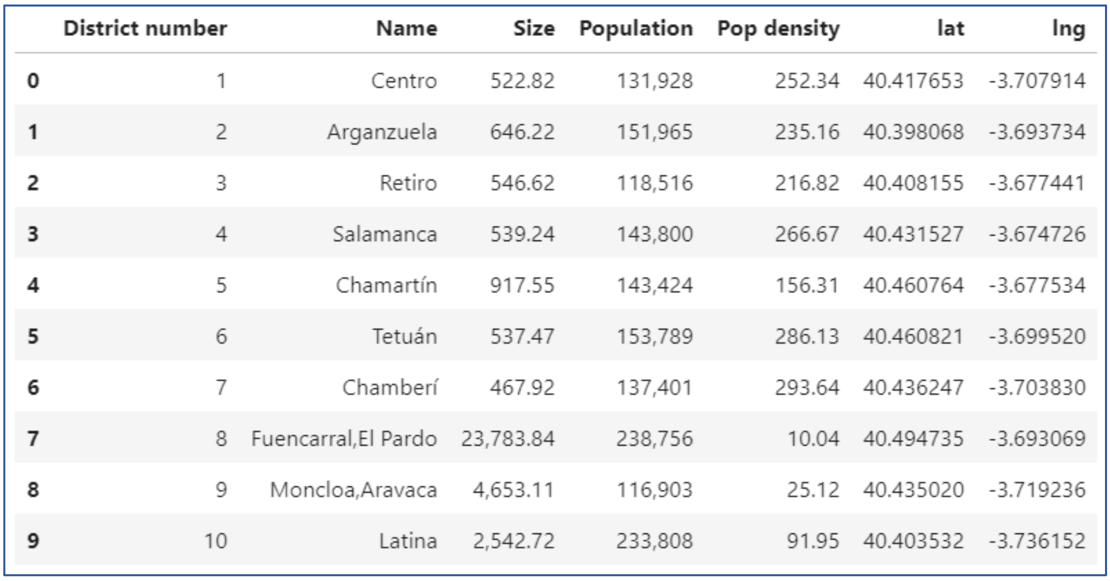
Thanks to the *BeautifulSoup* library, we are able to access the data directly from the website itself. For the target city of Madrid, the data frame resulting from this process looks as shown in the following table (Fig. 2):



| | District number | Name | Size | Population | Pop density |
|---|---|---|---|---|---|
| **0** | 1 | Centro | 522.82 | 131,928 | 252.34 |
| **1** | 2 | Arganzuela | 646.22 | 151,965 | 235.16 |
| **2** | 3 | Retiro | 546.62 | 118,516 | 216.82 |
| **3** | 4 | Salamanca | 539.24 | 143,800 | 266.67 |
| **4** | 5 | Chamartín | 917.55 | 143,424 | 156.31 |
| **5** | 6 | Tetuán | 537.47 | 153,789 | 286.13 |
| **6** | 7 | Chamberí | 467.92 | 137,401 | 293.64 |
| **7** | 8 | Fuencarral,El Pardo | 23,783.84 | 238,756 | 10.04 |
| **8** | 9 | Moncloa,Aravaca | 4,653.11 | 116,903 | 25.12 |
| **9** | 10 | Latina | 2,542.72 | 233,808 | 91.95 |

***Fig. 2.*** *First ten districts of Madrid data frame*

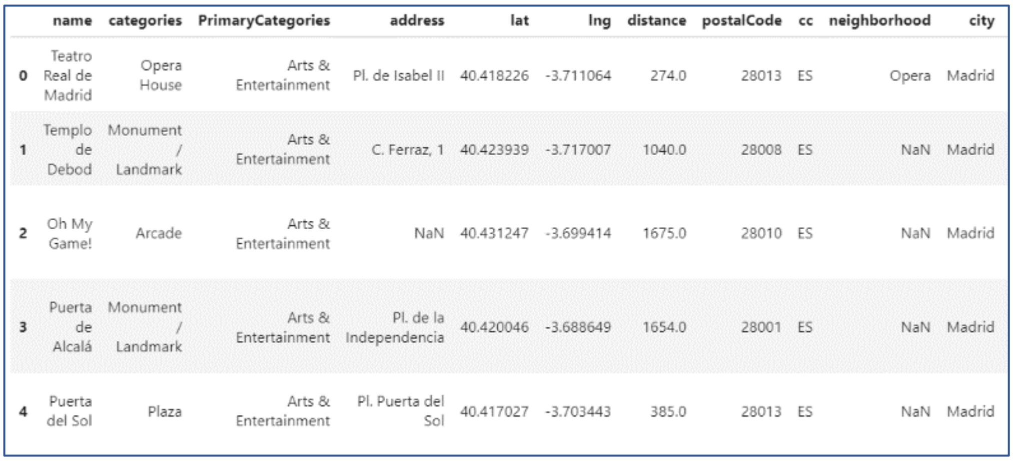
Based on the data frame shown in Fig. 2, each district will be scanned for *source venues* with a radius of 1.5 km. To do so, the coordinates of each districts are needed. These coordinates will be obtained thanks to the Geopy API, that will return the coordinates for a given address. In this case, the district name, followed by the city and country coordinates. The result of the final data frame can be seen in Fig. 3, as follows:



| | District number | Name | Size | Population | Pop density | lat | lng |
|---|---|---|---|---|---|---|---|
| **0** | 1 | Centro | 522.82 | 131,928 | 252.34 | 40.417653 | -3.707914 |
| **1** | 2 | Arganzuela | 646.22 | 151,965 | 235.16 | 40.398068 | -3.693734 |
| **2** | 3 | Retiro | 546.62 | 118,516 | 216.82 | 40.408155 | -3.677441 |
| **3** | 4 | Salamanca | 539.24 | 143,800 | 266.67 | 40.431527 | -3.674726 |
| **4** | 5 | Chamartín | 917.55 | 143,424 | 156.31 | 40.460764 | -3.677534 |
| **5** | 6 | Tetuán | 537.47 | 153,789 | 286.13 | 40.460821 | -3.699520 |
| **6** | 7 | Chamberí | 467.92 | 137,401 | 293.64 | 40.436247 | -3.703830 |
| **7** | 8 | Fuencarral,El Pardo | 23,783.84 | 238,756 | 10.04 | 40.494735 | -3.693069 |
| **8** | 9 | Moncloa,Aravaca | 4,653.11 | 116,903 | 25.12 | 40.435020 | -3.719236 |
| **9** | 10 | Latina | 2,542.72 | 233,808 | 91.95 | 40.403532 | -3.736152 |

***Fig. 3.*** *First 10 districts of Madrid data frame with their coordinates*

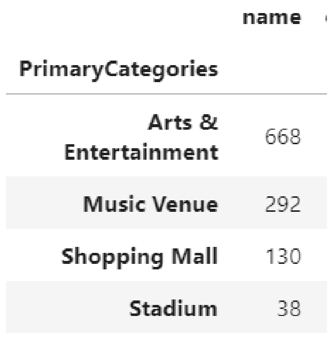
Due to the search method (circle based, per district) it is expected an overlap in the central ones that will have to be treated, while those situated in the city outskirts or of great size, will probably not be entirely covered. The result of such exploration returns a data frame as illustrated in Fig. 4.



| | name | categories | PrimaryCategories | address | lat | lng | distance | postalCode | cc | neighborhood | city |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Teatro Real de Madrid | Opera House | Arts & Entertainment | Pl. de Isabel II | 40.418226 | -3.711064 | 274.0 | 28013 | ES | Opera | Madrid |
| **1** | Templo de Debod | Monument / Landmark | Arts & Entertainment | C. Ferraz, 1 | 40.423939 | -3.717007 | 1040.0 | 28008 | ES | NaN | Madrid |
| **2** | Oh My Game! | Arcade | Arts & Entertainment | NaN | 40.431247 | -3.699414 | 1675.0 | 28010 | ES | NaN | Madrid |
| **3** | Puerta de Alcalá | Monument / Landmark | Arts & Entertainment | Pl. de la Independencia | 40.420046 | -3.688649 | 1654.0 | 28001 | ES | NaN | Madrid |
| **4** | Puerta del Sol | Plaza | Arts & Entertainment | Pl. Puerta del Sol | 40.417027 | -3.703443 | 385.0 | 28013 | ES | NaN | Madrid |

***Fig. 4.*** *First 5 entries districts of the district venue exploration*

To facilitate the interpretation of these results, a summary (Fig. 5) and a visualization was prepared thanks to the Folium maps library, (see Fig. 6).



| PrimaryCategories | name |
|---|---|
| Arts & Entertainment | 668 |
| Music Venue | 292 |
| Shopping Mall | 130 |
| Stadium | 38 |

*__Fig. 5.__ Venue exploration summary table*

From the table in Fig. 5, we can observe the venue distribution per category. A total of 1128 venues have been fetched from the Foursquare data bank, out of which 668 venues belong to the dominant category which is "Arts & Entertainment".

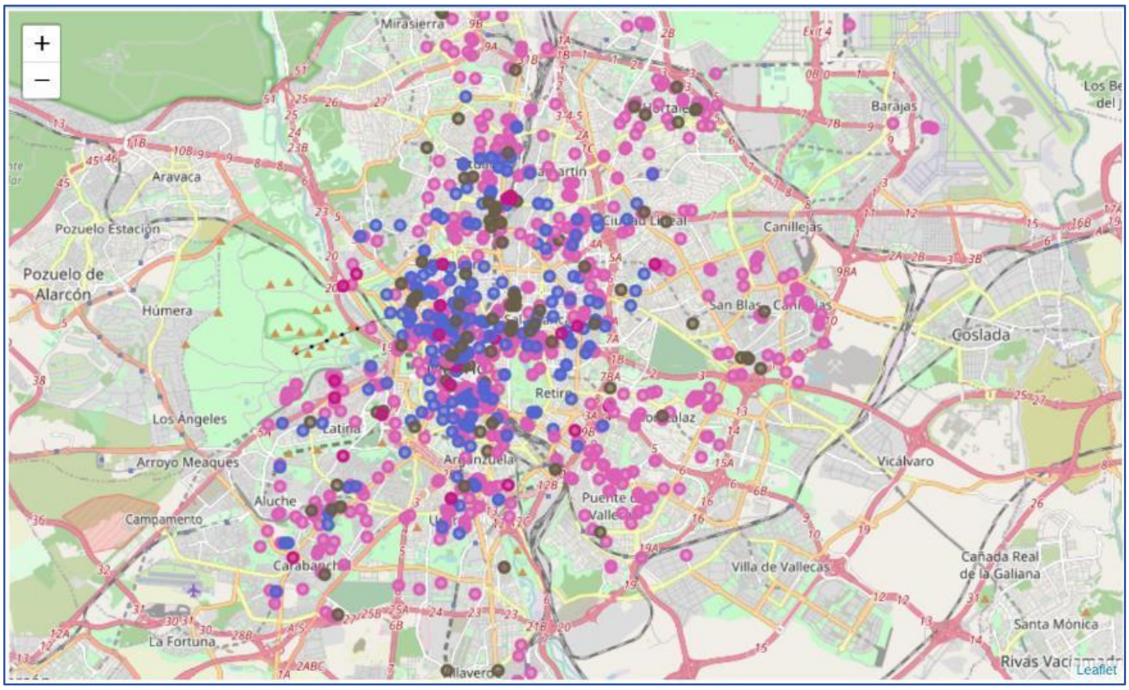


*__Fig. 6.__ Venue exploration map*

As it can be seen in Fig. 6, there are no obvious gaps except perhaps on the districts situated in the city outskirts. Therefore, it is acceptable to assume that the whole city is covered by the venue search method used and the data is reliable enough to conduct the analysis.

## 4.2 Results (K-Means and DBSCAN)

The results will be divided into two parts, one for each machine learning method used. The idea is to compare the similarities and differences between each, to elaborate more solid conclusions.

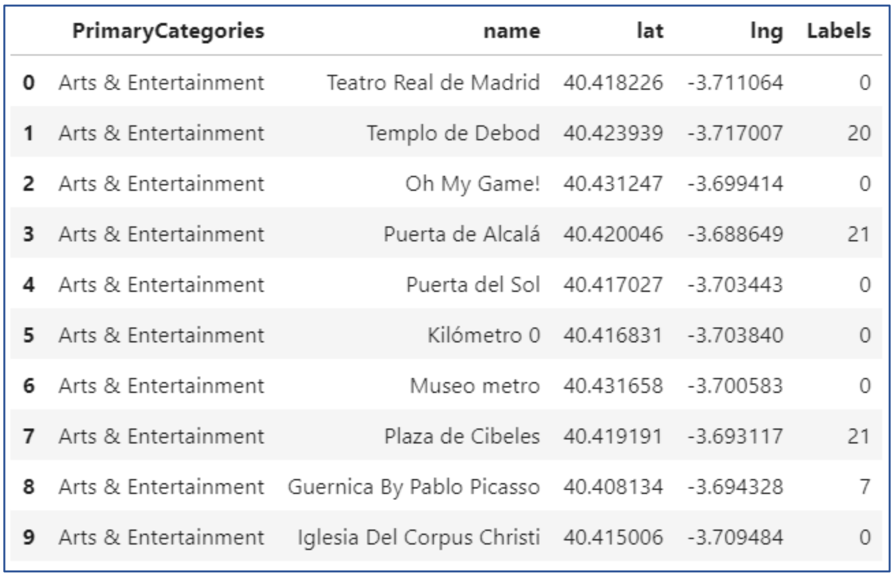
### 4.2.1 K-Means analysis

As explained in the methodology section, this method has the positive side of returning results in a circular area, which is easy to analyse. It also allows the user to modify the number of clusters and get different proposals, which may produce different results, going from specific to broad areas.

On the downside, since all points are assigned to a cluster, the presence of outliers may increase dramatically the size of the clusters, providing the user a huge area as a recommendation which could be misleading.

In this case, the analysis was performed first on 12, which turned out to be overly broad due to the size of the city. In the present document we will illustrate the result done with 24 clusters, which will segment further the city, providing more specific recommendations. Initial centroids are set by the "*k-means++*" method.

The K-Means algorithm is applied to the venue data frame and the resulting cluster labels are added to each element in it, resulting into a data frame as shown in Fig. 7.



| | PrimaryCategories | name | lat | lng | Labels |
|---|---|---|---|---|---|
| 0 | Arts & Entertainment | Teatro Real de Madrid | 40.418226 | -3.711064 | 0 |
| 1 | Arts & Entertainment | Templo de Debod | 40.423939 | -3.717007 | 20 |
| 2 | Arts & Entertainment | Oh My Game! | 40.431247 | -3.699414 | 0 |
| 3 | Arts & Entertainment | Puerta de Alcalá | 40.420046 | -3.688649 | 21 |
| 4 | Arts & Entertainment | Puerta del Sol | 40.417027 | -3.703443 | 0 |
| 5 | Arts & Entertainment | Kilómetro 0 | 40.416831 | -3.703840 | 0 |
| 6 | Arts & Entertainment | Museo metro | 40.431658 | -3.700583 | 0 |
| 7 | Arts & Entertainment | Plaza de Cibeles | 40.419191 | -3.693117 | 21 |
| 8 | Arts & Entertainment | Guernica By Pablo Picasso | 40.408134 | -3.694328 | 7 |
| 9 | Arts & Entertainment | Iglesia Del Corpus Christi | 40.415006 | -3.709484 | 0 |

***Fig. 7**. Venues data frame with K-Means labels*

To understand how the clusters are distributed across the city, Fig. 8 visualization has been created, highlighting the centroid of each cluster in red.

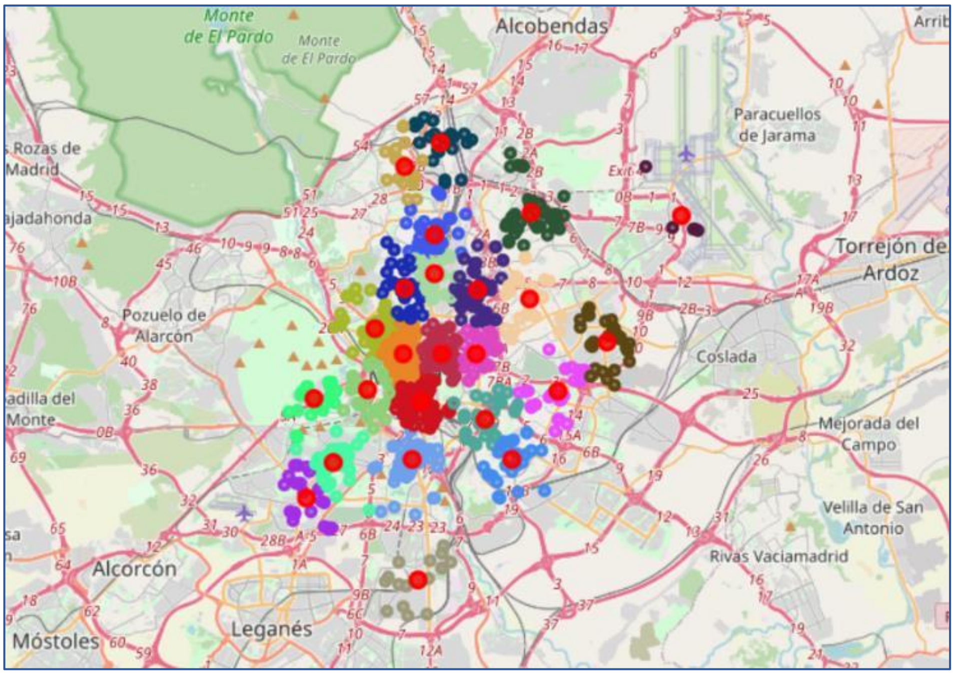


***Fig. 8.*** *K-Means cluster distribution*

At this point, it is a good moment to cross-check how similar the cluster centroids are to the district centres (see Fig 9). As explained in the methodology section, the search process may lead to biasing in the data clusters, as not the full district is scanned, but only 1.5km around the centre.

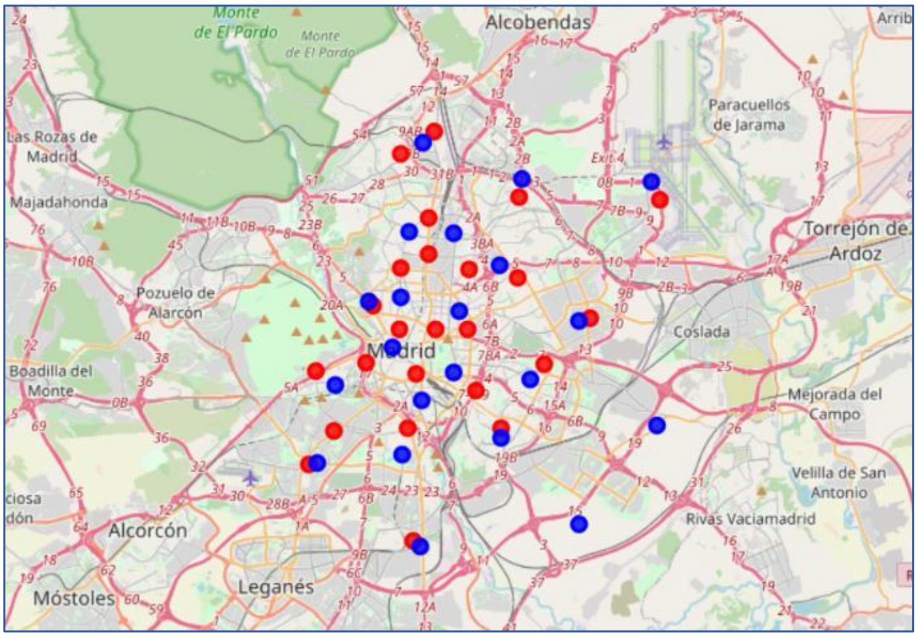


***Fig. 9.*** *K-Means centroids (red) vs District centres (blue)*

As it can be seen in Fig. 9, in the outer districts the cluster centroids are relatively close the centres. In the city core, the centroids are more distributed and there is no clear pattern in most of the cases.

Before the cluster values can be calculated, the search radius of each must be defined. The selected method will be the average radius, calculated by averaging the distance of the centroid to all the points. This way the impact of potential outliers will be mitigated.

This task proved to be very challenging, as calculating the distance between coordinates couldn't be performed by the simple Euclidean distance, but through conversion via the radius of the earth and trigonometry, into meters.

The search areas that will be used for the different clusters are shown in Fig. 10.

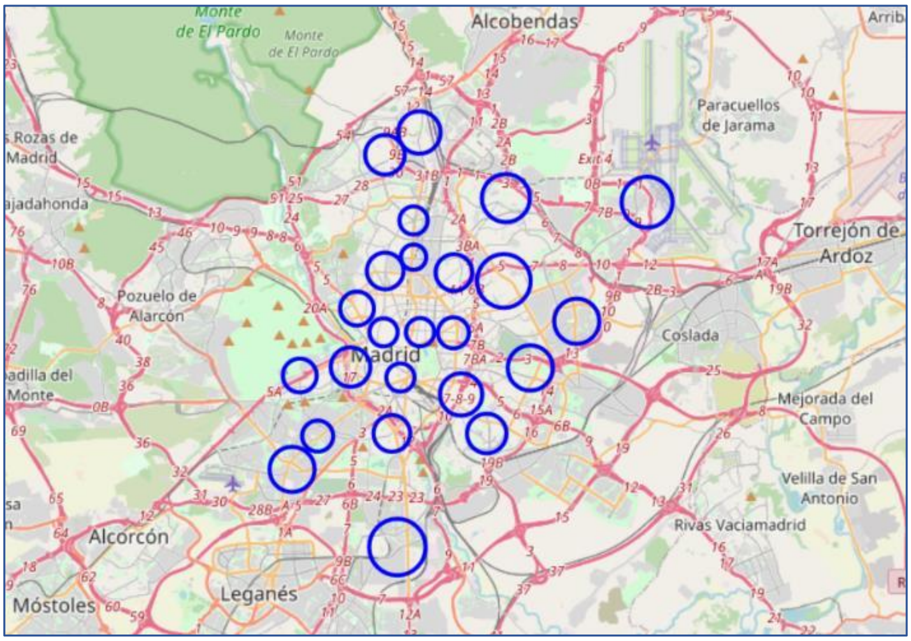
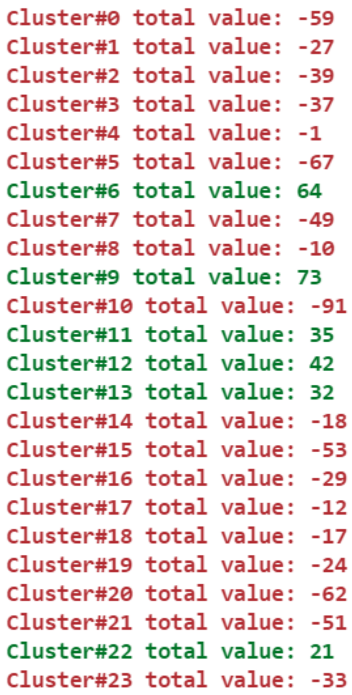


***Fig. 10.*** *K-Means cluster search areas*

The next step is to calculate each cluster value, based on the total number of source and sink venues contained in each. The result was as follows:



```
Cluster#0 total value: -59
Cluster#1 total value: -27
Cluster#2 total value: -39
Cluster#3 total value: -37
Cluster#4 total value: -1
Cluster#5 total value: -67
Cluster#6 total value: 64
Cluster#7 total value: -49
Cluster#8 total value: -10
Cluster#9 total value: 73
Cluster#10 total value: -91
Cluster#11 total value: 35
Cluster#12 total value: 42
Cluster#13 total value: 32
Cluster#14 total value: -18
Cluster#15 total value: -53
Cluster#16 total value: -29
Cluster#17 total value: -12
Cluster#18 total value: -17
Cluster#19 total value: -24
Cluster#20 total value: -62
Cluster#21 total value: -51
Cluster#22 total value: 21
Cluster#23 total value: -33
```

In order to get a better impression on were these are located, the results over the map can be observed in Fig. 11.

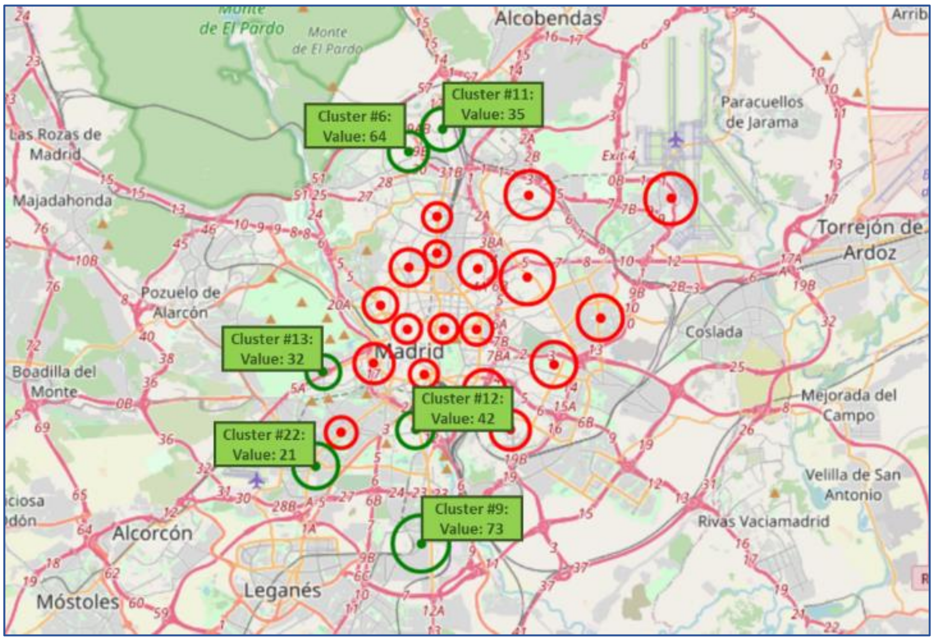


*Fig. 11. K-Means cluster value results*

Since the number of areas is not remarkably high, it would be advisable for the user to explore in detail the different clusters in green at a later stage. The goal should be to understand their feasibility, instead of opting directly for the one with the highest value.

For instance, Cluster #13 is located on top of an amusement park. Since it is a private location, there is a lack of competitors and the high number of venues related to entertainment explains why this location was recommended by the algorithm.

### 4.2.2 DBSCAN analysis

This method has the advantage of finding clusters of irregular shape, making it more precise. The user has the flexibility in this case to define how restrictive the model is by defining epsilon and the minimum number of samples of each cluster.

On the downside, given the fact that information from Foursquare can only be fetched in circular shapes, analysing the cluster values becomes a more challenging.

The DBSCAN algorithm was configured to be restrictive, by setting the epsilon to 0.15 and the minimum number of samples in each cluster to 10. The result is shown in Fig 12.

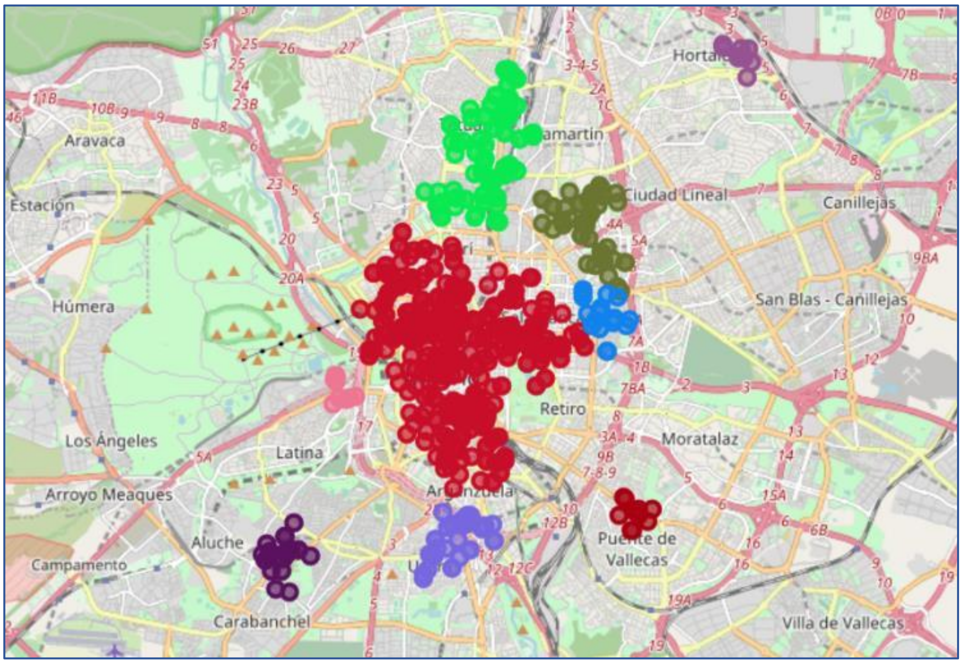
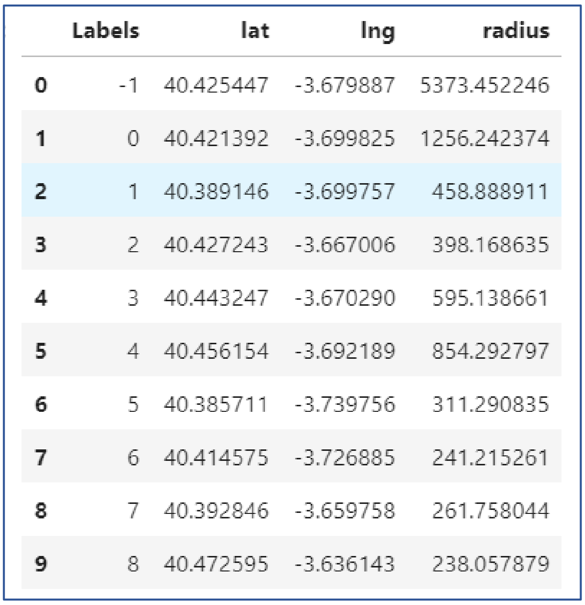


***Fig. 12.*** *DBSCAN clusters visualization*

The assumption now will be to approximate the search area of each cluster to a circle, located in the centre of mass of each cluster with a radius equal to the average distance of the centroid to each element of the cluster.

After the calculations were done, a data frame was created (see Fig. 13) to store the information.



| | Labels | lat | lng | radius |
|---|---|---|---|---|
| 0 | -1 | 40.425447 | -3.679887 | 5373.452246 |
| 1 | 0 | 40.421392 | -3.699825 | 1256.242374 |
| 2 | 1 | 40.389146 | -3.699757 | 458.888911 |
| 3 | 2 | 40.427243 | -3.667006 | 398.168635 |
| 4 | 3 | 40.443247 | -3.670290 | 595.138661 |
| 5 | 4 | 40.456154 | -3.692189 | 854.292797 |
| 6 | 5 | 40.385711 | -3.739756 | 311.290835 |
| 7 | 6 | 40.414575 | -3.726885 | 241.215261 |
| 8 | 7 | 40.392846 | -3.659758 | 261.758044 |
| 9 | 8 | 40.472595 | -3.636143 | 238.057879 |

***Fig. 13.*** *DBSCAN clusters circular approximation*

The resulting cluster simplification can be visualised in Fig. 14. These areas are assumed to be representative for each cluster.

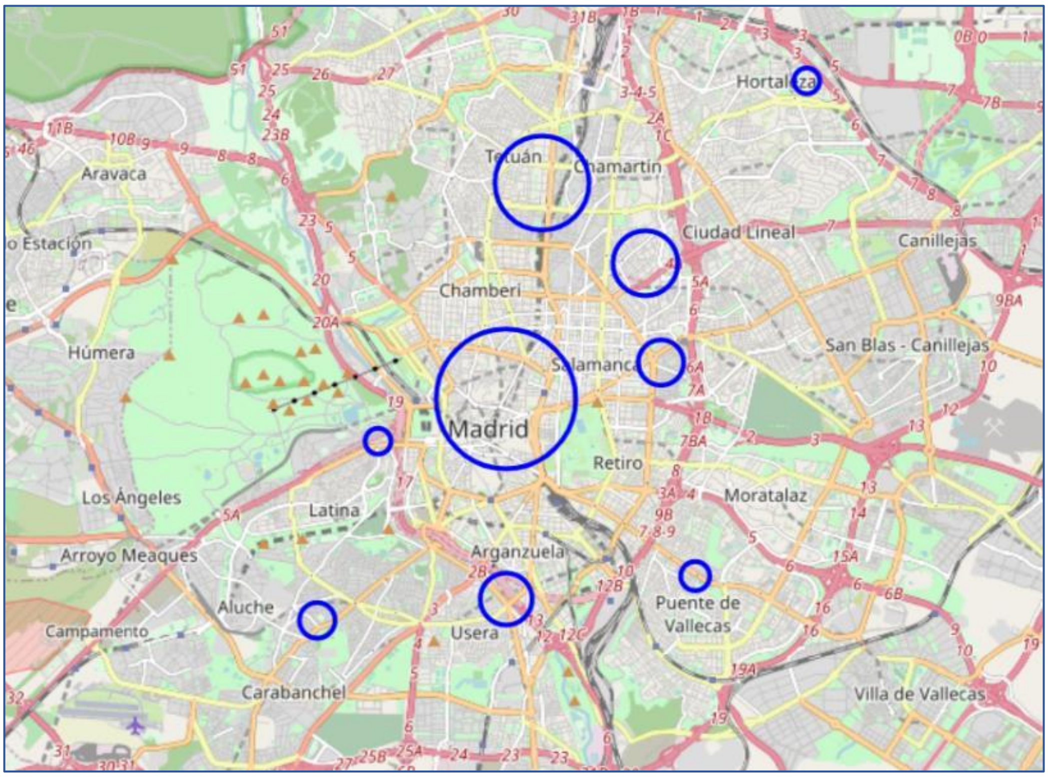
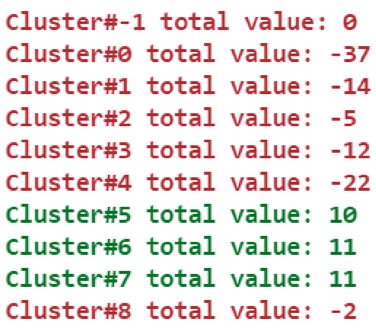


***Fig. 14.*** *DBSCAN clusters circular approximation visualization*

The calculation of the cluster values will be the same as in the K-Means analysis: Subtracting the sink venues from the source ones.



```
Cluster#-1 total value: 0
Cluster#0 total value: -37
Cluster#1 total value: -14
Cluster#2 total value: -5
Cluster#3 total value: -12
Cluster#4 total value: -22
Cluster#5 total value: 10
Cluster#6 total value: 11
Cluster#7 total value: 11
Cluster#8 total value: -2
```

Cluster#-1 corresponds to the outliers, or those venues that were not part of any cluster. This cluster is excluded from the visualizations.

The outcome of the cluster value calculations can be visualized in Fig. 15.

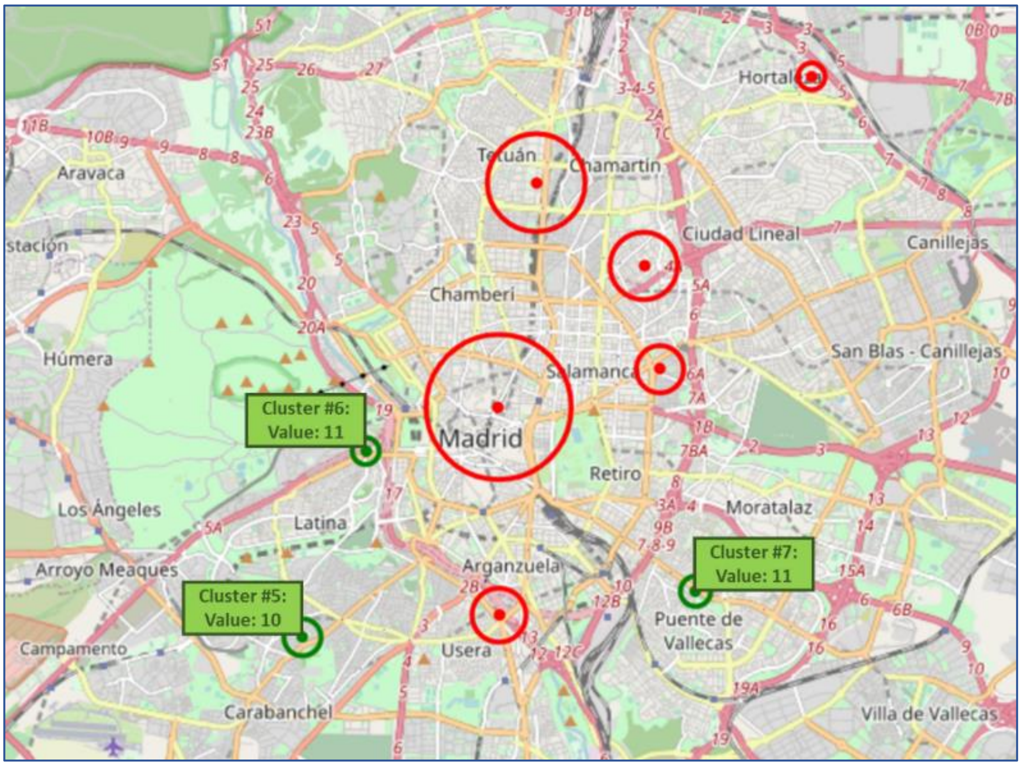


*Fig. 15.* *DBSCAN cluster values visualization*

Now that the recommendations have been prepared, the user can now carry on the manual research and exploration of these areas or modify the DBSCAN parameters to get another recommendation if this one doesn't meet the original expectations.

### 4.3 Observations & recommendations

It is clear from the analyses with DBSCAN and K-means, that the south east part of the city may contain areas that *a priori* present a positive scenario for the potential "tapas bar" defined in the use case. Therefore, the recommendation for the business/entrepreneur would be to take these parts of the city into consideration for the location of their business.

It is clear that this recommendation is not sufficient, as it is important now to understand in detail the kind of businesses that operate there, since the categories can be very broad or the cluster biased towards a specific category, and if these still fit to the taste of the target customers.

If these areas were not sufficient, by modifying the parameters, like increasing the number of clusters to be used in K-Means, it would be possible to make the search more specific and find new areas of smaller size also worth looking into.

### 4.3.1 Additional developments

The methodology presented in this analysis provides only a rough estimation of the area based on a simple calculation. A great additional add-on to this methodology, would be to further develop the code to fine tune some aspects, that were discarded in first place for sake of simplicity, as for example:

- **City densities**: The population density of each neighbourhood may have a dramatic impact on the performance of our business. For instance, it may happen that a highly populated area with few source venues, is still quite successful due to the high number of residents.
- **Venues value**: In this analysis, all venues were given the same value, one point each. But on reality, the capacity of each venue to attract potential customers to their surroundings varies from one category to another. For instance, a touristic spot or train station may attract more people than a theatre quantitatively. Similar logic also applies to the sink venues.
- **Venue performance**: The capacity of a venue to attract customers, may also vary with time. For example, a museum could be more regular, while a stadium has a strong seasonality depending on the day of the week or the month of the year.
- **Venue size**: The idea behind it is that the size of each venue can make a difference as well. As within the same category, the bigger it its, the higher the number of potential customers.

## 5. Conclusions

A machine learning project is never black or white, but the combination of assumptions, business rules and intuition that drive the technical development and lead to the results, which can be sometime the result of several project iterations. Thus, these may vary depending on the person running the analysis and how the problem is perceived.

This analysis for instance, tries to find the gaps in demand of a service by weighting the venues that attract potential customers to their surroundings with those that satisfy that similar need and act as competitors. This logic is not entirely wrong *per se* but is not taking into consideration something that might be obvious for somebody familiar with that industry: *the city hot spots.*

Let us assume that our use case tries to find the best area to emplace a new Chinese restaurant. Based on the current logic, an area such as *"Chinatown"* will be discarded with almost certainty due to the existing number of competitors there. But when thinking from a customer perspective, for somebody craving for Chinese food, this will probably be the first area that comes to mind.

Therefore, it is of tremendous importance for a machine learning project not only to have at disposal great technical skills, but also to ensure that the best business knowledge is part of the project development from the very beginning and that both work side by side.

## 6. References

- Wikipedia – Districts of Madrid
- Fig. 1 (Modified)- Districts of Madrid
- Foursquare.com – Venue categories