# Coursera Capstone: Analyzing neighborhoods of Madrid, Spain

## Introduction

The city of Madrid is the capital of Spain and one of the largest cities in this country. Being a big city it offers more opportunities to start a new business like a restaurant. But it is a big issue to decide where to locate that new business. It makes sense to locate the restaurant in a neighborhood where there are more eating places because many people is used to going to that location to eat. But also if there is a big amount of restaurants the business will have a higher competence to rise.

With this idea in mind this project explored the neighborhoods in the city of Madrid to know if opening a restaurant is a good choice and where to do it. So anyone interested in starting a new business of this type could make a better choice based on data.

## Methodology

- Data:

The data was extracted from Wikipedia. This web page contains a list of all the neighborhoods in Madrid separated by districts. This unstructured data was processed to obtain structured data (for this step BeautifulSoup was used). Only the columns “District name” and “Name” where saved in a data frame. From this data the latitude and longitude of each neighborhood were calculated using Geopy and later on information about nearby venues in each neighborhood was extracted using the Foursquare API. This information about venues was used to cluster the neighborhoods in Madrid and look for the best cluster to open a new restaurant.

The data from Wikipedia was downloaded making a request and retrieving the html file. The file was parsed using BeautifulSoup. The blank rows were skipped and there were two types of rows with data: rows with a district and a neighborhood and rows with only a neighborhood. To differentiate the district from the neighborhood in the first case a parenthesis is searched in the last position (the district always ends in a closing parenthesis). To extract the district the cell containing the name and the district number were split and the district name was selected. To differentiate the neighborhood a string not empty, different from a number and different from “[[ ]]” (the images in the table) was searched. To extract the neighborhood the spaces before and after the name were removed.

The next step was calculating the latitude and longitude of each neighborhood. As said before Geopy was used for this step. The locations were searched as “*District name*, *neighborhood name,* Madrid, España”. After this calculation a null value was generated and that row was deleted. The resulting data frame looks like the following one:

| | District | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|
| **0** | Centro | Palacio | 40.417821 | -3.715111 |
| **1** | Centro | Embajadores | 40.421058 | -3.707185 |
| **2** | Centro | Cortes | 40.416389 | -3.696463 |
| **3** | Centro | Justicia | 40.424246 | -3.693508 |
| **4** | Centro | Universidad | 40.422753 | -3.709849 |

The names, latitudes, longitudes and categories of nearby venues of each neighborhood were extracted making a request to Foursquare. The resulting data frame looks like the following one:

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| **0** | Palacio | 40.417821 | -3.715111 | Palacio Real de Madrid | 40.417940 | -3.714259 | Palace |
| **1** | Palacio | 40.417821 | -3.715111 | Plaza de la Almudena | 40.416320 | -3.713777 | Plaza |
| **2** | Palacio | 40.417821 | -3.715111 | Santa Iglesia Catedral de Santa María la Real ... | 40.415767 | -3.714516 | Church |
| **3** | Palacio | 40.417821 | -3.715111 | Plaza de Oriente | 40.418326 | -3.712196 | Plaza |
| **4** | Palacio | 40.417821 | -3.715111 | Jardines de Sabatini | 40.419954 | -3.713126 | Garden |