

Correlation between Malaga and Toronto Neighborhoods

Juan Antonio Garcia Fernandez

Coursera | IBM Data Science Capstone Project | January 2020

Contents

[**Introduction** 2](#_Toc36410341)

[**Problem description** 2](#_Toc36410342)

[**Target audience** 2](#_Toc36410343)

[**Success criteria** 2](#_Toc36410344)

[**Data Description** 3](#_Toc36410345)

[**Data sources** 3](#_Toc36410346)

[**Methodology** 4](#_Toc36410347)

[**Web scraping** 4](#_Toc36410348)

[**Data wrangling / Data cleaning** 4](#_Toc36410349)

[**Add latitude and longitude** 4](#_Toc36410350)

[**Merging data** 4](#_Toc36410351)

[**Explore neighborhoods** 5](#_Toc36410352)

[**Analyze each neighborhood** 5](#_Toc36410353)

[**Cluster neighborhoods** 6](#_Toc36410354)

[**Examine clusters** 8](#_Toc36410355)

[**Results** 9](#_Toc36410356)

[**Discussion** 9](#_Toc36410357)

[**Conclusion** 10](#_Toc36410358)

# **Introduction**

After centuries of technological progress and advances in international cooperation, the world is more connected than ever. **Globalization** is the word used to describe the growing interdependence of the world’s economies, cultures, and populations, brought about by cross-border trade in goods and services, technology, and flows of investment, people, and information.

**“Flows of people”** means that people can now travel more often in less time, but also **live in many different countries**. This is not always easy because the new country can be a great unknown to that level at the beginning. When you move from one city to another in a distant country, you would like to find a place, a neighborhood, relatively similar to the one you were living in. **People usually like to keep their living standards**.

Therefore, the idea of this project comes from the effort required for finding a neighborhood to live after moving to another city in a different country. Specifically, this report is focused on **people moving from Toronto (Canada) to Malaga (Spain)** in order to illustrate the process for developing a **Machine Learning model to identify the best possible neighborhood to move in to.**

## **Problem description**

The goal of this project is to explore the neighborhoods of Malaga and Toronto in order to identify similarities and correlations between them. In short, the question to be answered is: **Given a neighborhood X in the city of Toronto, which neighborhoods in Malaga have the same kind of venues?**

## **Target audience**

The target audience of this project are:

* **Real estate agencies in Málaga or Toronto** that are paid by people who are moving from Malaga to Canada (or vice versa) to search for a proper accommodation, either sale or rental, in a convenient neighborhood similar to the one they are coming from.
* **International companies** hiring people from another country (Spain/Canada) need to offer to their new employees a place to stay. Preferably, a place where they feel comfortable and at home.
* **Families** that want to live an adventure abroad, but also maintain an environment close to their lifestyle.

## **Success criteria**

The success criteria of this project will be a good recommendation of similar neighborhoods, i.e., same kind of venues, leisure opportunities, green areas and so on, to people moving from Malaga to Canada (or vice versa).

# **Data Description**

Malaga and Toronto neighborhoods are the main scope of the project. We need to explore, segment and cluster them based on the types and amounts of venues offered by each of them. Accordingly, we will need the following data:

* **List of neighborhoods in Malaga and Toronto.** Note that Malaga is a medium city in the South of Spain, while Toronto is a big city located on Lake Ontario in Canada.
* **Coordinates of the neighborhoods.** This information is needed to determine the available venues in each neighborhood, and also to be able to plot the neighborhoods together with the clustering outcome provided by the algorithm on a map.
* **Venue data.** This is key to feed the clustering algorithm that aims to process the data and find clusters of neighborhoods, if they exist in the data.

## **Data sources**

* For the **Toronto neighborhood data**, a **Wikipedia** page exists that has all the information we need to explore and cluster the neighborhoods in Toronto (<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>). On the other hand, for the **Malaga neighborhood data**, there is an **OpenStreetMap** page with the same kind of information that we need to explore and cluster the neighborhoods in Malaga (<https://wiki.openstreetmap.org/wiki/ES:Lugares_en_M%C3%A1laga>). We will scrape these webpages, wrangle the data, clean it, and then read it into a pandas dataframe so that it is in a structured format. Once the data is in a structured format, we will analyze the dataset.
* **Geopy** is an excellent Python library for (among others) geocoding and reverse geocoding that supports many APIs. In this project, we will use the **Nominatim** API, which is based on OpenStreetMap (OSM) data. Nominatim allows us **to convert an address into latitude and longitude values**. In this manner, we will get the coordinates of the neighborhoods.
* **FourSquare API** is an API to interact with the Foursquare platform. Foursquare is a social location service that allows users to explore the world around them. One of the available API methods **returns the surrounding venues given a specific location** (<https://developer.foursquare.com/docs/api/venues/search>).
* **Folium** is a Python library for producing maps with Leaflet.js. With this, we will be able to create maps where to visualize the clusters provided by the Machine Learning model.

# **Methodology**

## **Web scraping**

Firstly, we need to scrape a couple of webpages to obtain the Toronto and Malaga neighborhood data. The web scraping will be done via one of the most used Python libraries, ***pandas***. *Pandas* allow us to read a **HTML table into a Dataframe**.

The list of neighborhoods in Toronto is available in a Wikipedia webpage (<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>). There are 3 tables in this webpage; however, just the table of postal codes is of our interest since it presents the different boroughs and neighborhoods in Toronto.

On the other hand, we can find the directory of districts in Malaga, along with their corresponding neighborhoods, in a Wiki page where there are again 3 tables (<https://wiki.openstreetmap.org/wiki/ES:Lugares_en_M%C3%A1laga>). In this case, we are interested in the second one where the list of districts is introduced. The same procedure followed for Toronto is done here to scrape the required information.

## **Data wrangling / Data cleaning**

The data acquired in the previous stage is not perfect in terms of content and format. This is why it’s needed to apply some basic processing to prepare the data by ensuring that it is correct and consistent. Below are listed the different actions performed in that sense:

* **Rename columns** (Malaga and Toronto data)
* **Drop rows with *Not assigned* Boroughs** (Toronto Data)
* **Fill *Not assigned* Neighborhoods** (Toronto Data)
* **Group data** (Toronto Data)
* **Remove useless columns** (Malaga Data)

## **Add latitude and longitude**

We need to know the neighborhood coordinates (latitude and longitude) so that we can input later this information into the data venues provider (Foursquare). To do this, we go through all Postal Code/District names and call the **ArcGIS API** for Python. Once this is done, we will add this information into our already cleaned Dataframes.

## **Merging data**

Firstly, we can go through an optional step to reduce drastically the number of boroughs in Toronto data. In this project, it is recommended to do so because the number of boroughs in Toronto is much larger than in Malaga. Therefore, we are going to keep only those Toronto boroughs that contain the word *Toronto*, i.e., boroughs from Toronto City.

Finally, we **concatenate our two datasets** (Malaga and Toronto) into one single Dataframe, as shown in Figure 1.

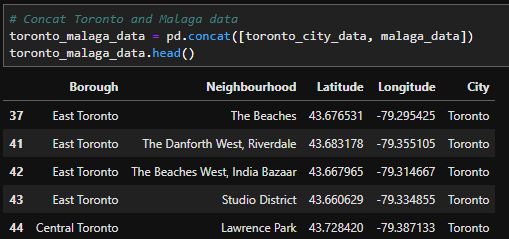


Figure 1 – Dataframe containing Toronto and Malaga data

## **Explore neighborhoods**

In this section, we will use **Foursquare API** to get the top 100 venues that are in all neighborhoods of Toronto City and Malaga within a radius of X meters. By default, a radius of 1000 meters is defined. The use of this API requires registration as a *Developer* in order to achieve the Foursquare ID and Foursquare secret key.

Next, we will call the Foursquare API (GET request) by passing in the geographical coordinates of each of the neighborhoods under study. Venue data will be returned in JSON format from where we will extract the venue name, the venue location and the venue category as presented in Figure 2. This information allows us to check how many venues were returned for each neighborhood as well.



Figure 2 – Nearby Venues

## **Analyze each neighborhood**

Firstly, a **One-Hot Encoding** process is used to convert venues (categorical variables) into numerical data.

Secondly, Dataframe rows are grouped by Neighborhood by taking the mean of the frequency of occurrence of each venue category. The output can be checked by printing each neighborhood along with the top 5 most common venues. See an example of one of the Malaga’s districts in Figure 3.

Finally, a new Dataframe is created with the top 10 venues for each neighborhood. See Figure 4.

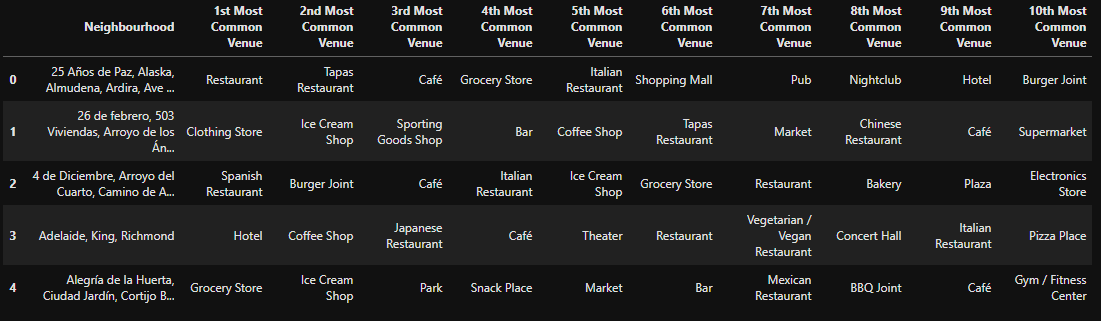


Figure 3 – Dataframe with Top 10 Venues

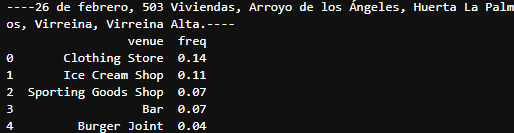


Figure 4 – Five Most Common Venues

## **Cluster neighborhoods**

At this point we have gathered, cleaned and prepared all data we need. The following step is to perform some suitable **Machine Learning** technique to get insights from this data. The goal is to employ a **Clustering Algorithm** that provides us some potential relations and similarities between Malaga and Toronto City neighborhoods. In this case, we have opted for a straightforward and well-known method: **K-Means.**

The most important parameter to run K-Means is the “k”, i.e., the **number of cluster**s we want to obtain. A priori, we don’t know how many clusters we need to ensure that we have achieved a good result, but what we do know is that we want to maximize the number of clusters where there is at least one Malaga neighborhood and one Toronto neighborhood. In this way, we will be able to establish relations between both cities. This is why the first step is to search for the optimal “k” value from the “Malaga and Toronto sharing clusters” point of view. Once we have the optimal “k” value, we will run K-Means once more.

Next, we will create a Dataframe containing all neighborhoods along with the cluster label provided by the algorithm and the top 10 venues, as shown in Figure 5.

Finally, we can also visualize these results on a map through the **Folium** Python library. A map for each city and its corresponding clusters is presented in Figure 6 and Figure 7.



Figure 5 – Clustered Neighborhoods

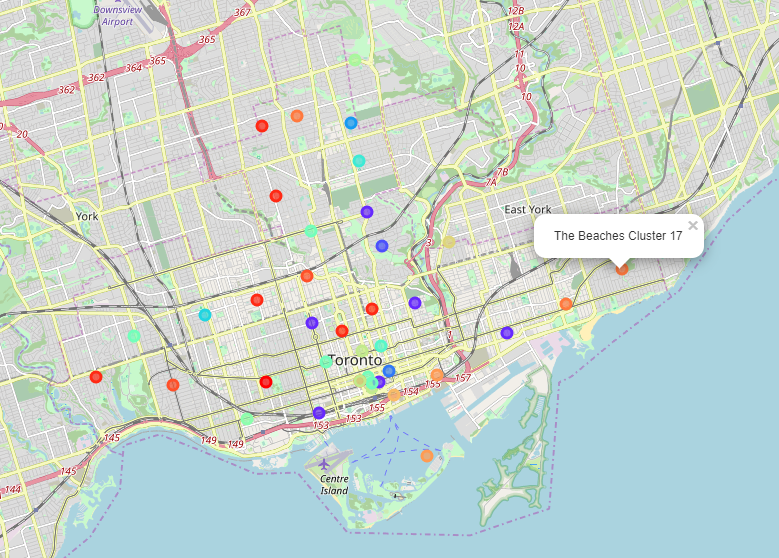


Figure 6 - Map with Toronto City Clusters

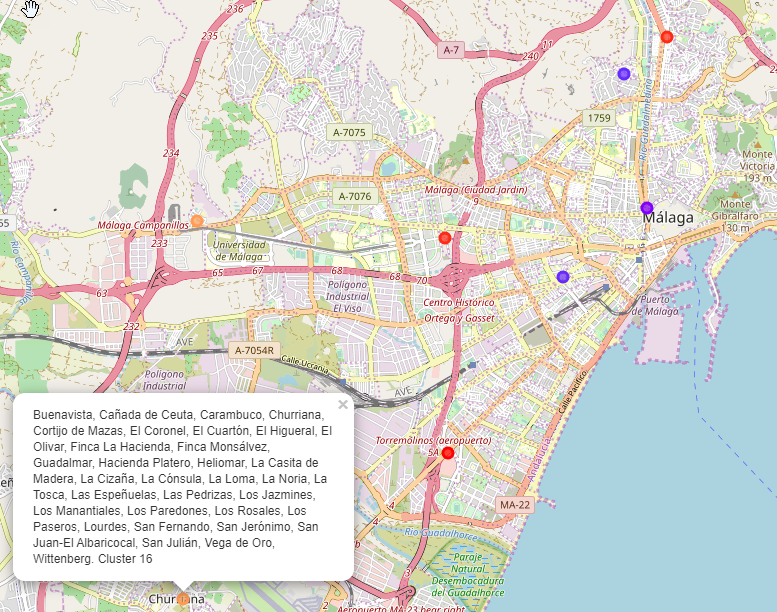


Figure 7 - Map with Malaga Clusters

## **Examine clusters**

To conclude, we can examine each cluster and determine which one has a correlation between Malaga and Toronto City neighborhoods. Then, we can understand which are the common venues and similarities in these areas. In that manner, we can advise people who want to move from one city to another about the areas where the lifestyle is in somehow very similar.

# **Results**

The first result that needs to be described is the optimal “k”. The optimal number of clusters to be used in our case was **20** as it provided up to **5 clusters** containing neighborhoods from both cities under study.

Regarding the clusters identified by K-Means, we can see the outcome below:

* **Clusters 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 17, 18** 🡪 THESE CLUSTERS HAVE NO CORRELATION BETWEEN MALAGA AND TORONTO
* **Cluster 0:** There is 1 Toronto Neighborhood and 1 Malaga Neighborhood
* **Cluster 2:** There are 6 Toronto Neighborhoods and 2 Malaga Neighborhoods
* **Cluster 13:** There are 3 Toronto Neighborhoods and 1 Malaga Neighborhood.
* **Cluster 16:** There are 2 Toronto Neighborhoods and 3 Malaga Neighborhoods.
* **Cluster 19:** There are 6 Toronto Neighborhoods and 2 Malaga Neighborhoods.

Clusters 0, 2, 13 ,16 ,19 are grouped together based on similar nearby venues in each of the neighborhoods. This information is highly relevant so that we can find “mirror” neighborhoods between Malaga and Toronto.

# **Discussion**

Based on the results above, the first thing that we can observe is that not all Malaga neighborhoods have been related to some of the Toronto neighborhoods. This means that there are **2 Malaga neighborhoods where there is no correlation**, in terms of nearby venues, with Toronto neighborhoods. Specifically, they are the neighborhoods from **District “Este”** and **District “Centro”**.

Now, we can analyze the other 5 clusters where we have a correlation between cities. If we take a carefully look at the top 10 most common venues in those clusters, we notice the following:

* **Cluster 0:** main spots are cheap Restaurants and Pubs. Due to this, we can name this cluster as **Teenagers and Nightlife.**
* **Cluster 2:** main spots are Restaurants (Italian, Fast Food, Japanese ...) and Cafés. Due to this, we can name this cluster as **Restaurants and Cafés.**
* **Cluster 13:** here there is a huge mix of everything (shops, cafés, department stores, hotels, pubs ...). Due to this, we can name this cluster as **Regular Neighborhood.**
* **Cluster 16:** main spots are places close or related to the sea (Harbors, Ferries, Seafood Restaurants ....) and Farmer Markets. Due to this, we can name this cluster as **Sea and Farms.**
* **Cluster 19:** main spots are Asian restaurants (sushi, Japanese, Korean ...), Cafés (or Coffee Shops) and Parks. Due to this, we can name this cluster as **Asian Restaurants, Cafés and Parks.**

These cluster names are just a possible proposal to describe somehow what is mainly expected from these neighborhoods, or in other words, which are the principal common interests in these areas. With this information, companies can advise their customer on the better places to move in depending on where they are coming from.

# **Conclusion**

In this project we have been able to build a complete system that searches for similar neighborhoods between two cities based on nearby venues by taking advantage of the power of machine learning techniques. Specifically, we have focused on the cities of Malaga and Toronto to develop the idea during the project, managing to propose similar neighborhoods between these 2 cities.

We have only considered the frequency of occurrence of nearby venues to feed the clustering algorithm, but there are other factors that could be also introduced to improve or influence the clustering, such as the number of inhabitants or the average rent for an apartment. Future research could go into that direction to include new features in the system.

Besides, other clustering techniques could be explored as well as tuning the hyperparameters available for K-Means algorithm. These ideas open interesting lines of research for the future.

In conclusion, this project has been an amazing (and highly recommended) journey to produce an valuable end-to-end solution based on machine learning from scratch.