COURSERA

IBM CAPSTONE PROJECT

# STUDY OF LAUSANNE NEIGHBOURHOODS

Lavinia Saccoccio

Data Scientist

Sapienza University of Rome , EPFL

## INTRODUCTION - BUSINESS CASE

For this project, I will introduce a hypothetical business case.

Imagine that you are a business man/student/retired couple who wants to move to the centre of Lausanne for personal or job reasons. But before moving in you want to find the perfect place to live, so that all your needs are fulfilled. Let's say you are a young worker, so you want to find a place near pubs and restaurants, close to the heart of the nightlife of the city.

The idea is to study the city of Lausanne in order to find a suitable place to live. The assumption behind the analysis is that we can use unsupervised machine learning (in particular the K-means algorithm) to create clusters of venues that will provide us with a list of areas for consideration.

## DATA

The data used to solve this problem is geolocation data collected from the FourSquare API. There is no data for the neighbourhoods in Lausanne, so I decided only to use the ones I got from the API.

The data will be structured in one dataframe, which will contain all the data we will need for the analysis. The dataframe is a simple one, and will contain only the necessary data, such as the name of the venue, the category and the geospatial coordinates (so that we can plot them on a map with the folium library)

| | name | categories | lat | lng |
|---|---|---|---|---|
| 0 | Sleepy Bear Coffee | Coffee Shop | 46.515338 | 6.631369 |
| 1 | Les Trois Rois | French Restaurant | 46.515515 | 6.631223 |
| 2 | Pasta e Sfizi | Italian Restaurant | 46.516374 | 6.633477 |
| 3 | Café du Simplon | Mediterranean Restaurant | 46.515961 | 6.629948 |
| 4 | Bar Tabac | Bar | 46.518066 | 6.634775 |

**Figure 1:** Example of dataframe present in the study.

Regarding the analysis we will use K-Means, un unsupervised machine learning algorithm to cluster the different kind on venues retrieved by Foursquare. Before that, we will try to find the best number of cluster with the elbow method, and then we will proceed with the study.

## METHODOLOGY

In this Section we will explain the methodology used to obtain the wanted results. In detail, we do the following.

After cleaning up and exploring the data, we will apply the K-means unsupervised algorithm for creating clusters of venues. Then we will use the elbow method for choosing the optimal number of clusters.
First, we create a table, that contains the list of venues of Lausanne with the respective geospatial coordinates. When done, it will look like Figure 1. For this part of the project, the geocode Python library was used. Then, just to explore visually the venues in the city, we plotted a heat map of the city:

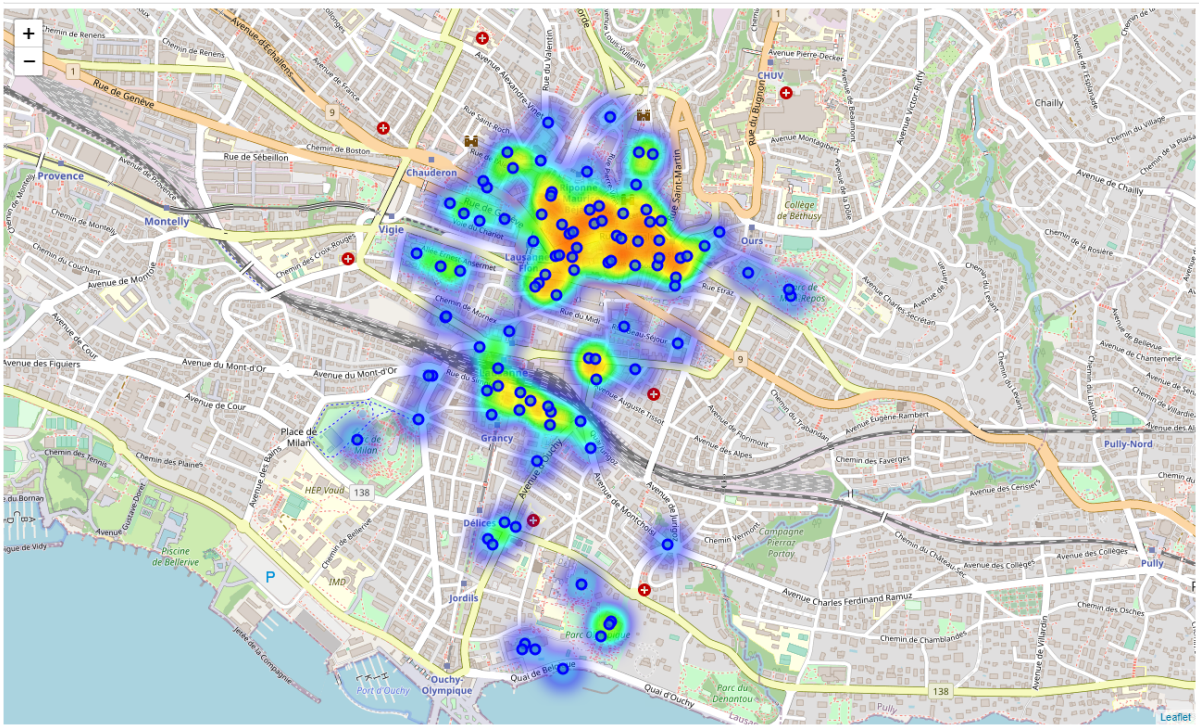


**Figure 2:** Heath map of the venues in Lausanne

This can help people decide where to move, because you can have a fast idea where the

city is more crowded or more calm. After visualising the entire dataframe on the map, we moved on and started preparing the data to be analysed.
In order to run K-Means, we have to transform the dataframe into something that the machine could understand. This is because K-Means cannot work with categorial data, but only with numerical data. A way to transform the data this way is to do a **One Hot Encoding**: One hot encoding is a process by which categorical variables are converted into binary variables (0,1). The dataframe after the one hot encoding looks like this:

| | name | Art Museum | Bar | Bistro | Bookstore | Breakfast Spot | Burger Joint | Café | Candy Store | Chinese Restaurant | Church | Coffee Shop | Creperie |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Sleepy Bear Coffee | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **1** | Les Trois Rois | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | Pasta e Sfizi | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | Café du Simplon | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | Les Gosses du Québec | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Figure 3:** Dataframe after the one hot encoding

For the clustering process, as we said earlier, the K-means algorithm was used, which is an unsupervised machine learning algorithm. This process also requires to set the parameter for the number of clusters. To be able to identify the optimal number for k (k is the number of clusters), the **elbow method** was used. We should choose the k which makes the slope sharply shift, but in this case is not that clear. In any case, we will choose the value 7 as the best number to be used for clusters. The (not so helping) elbow method curve is the following:

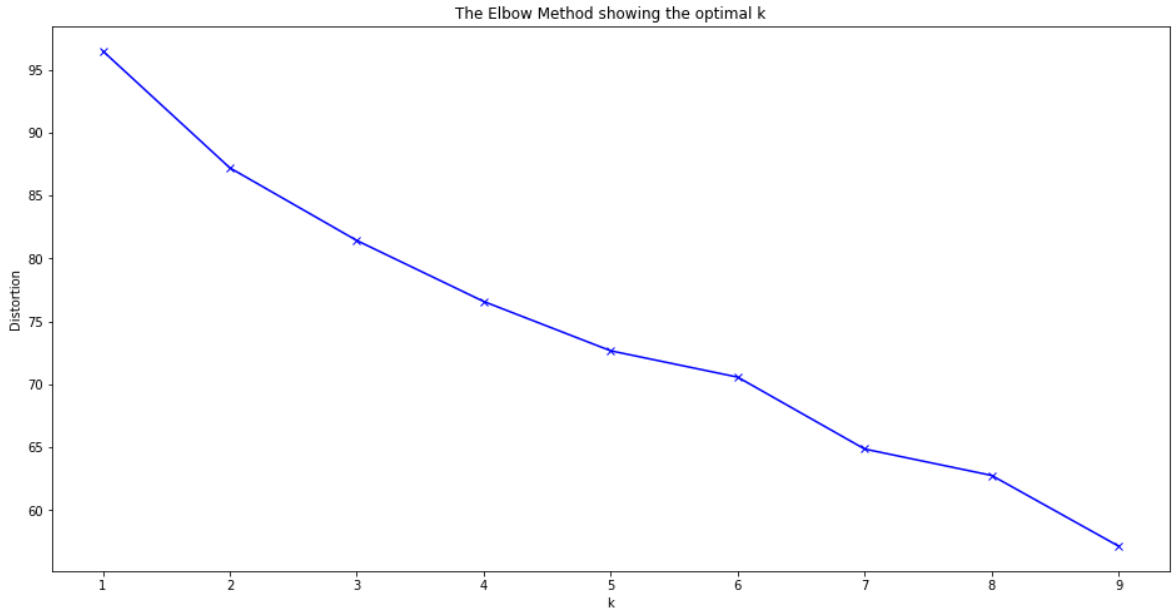


**Figure 4:** Elbow method plot

After choosing the value for K, we are ready to run K-Means. The resulting dataframe, after running the algorithm wit K = 7 is the following:

| | Cluster Labels | name | categories | lat | lng |
|---|---|---|---|---|---|
| **0** | 0 | Sleepy Bear Coffee | Coffee Shop | 46.515338 | 6.631369 |
| **1** | 1 | Les Trois Rois | French Restaurant | 46.515515 | 6.631223 |
| **2** | 5 | Pasta e Sfizi | Italian Restaurant | 46.516374 | 6.633477 |
| **3** | 0 | Café du Simplon | Mediterranean Restaurant | 46.515961 | 6.629948 |
| **4** | 2 | Les Gosses du Québec | Bar | 46.517072 | 6.633122 |

**Figure 5:** Dataframe with cluster labels

We can finally visualise the clustered venues on the map, thanks to the folium library.

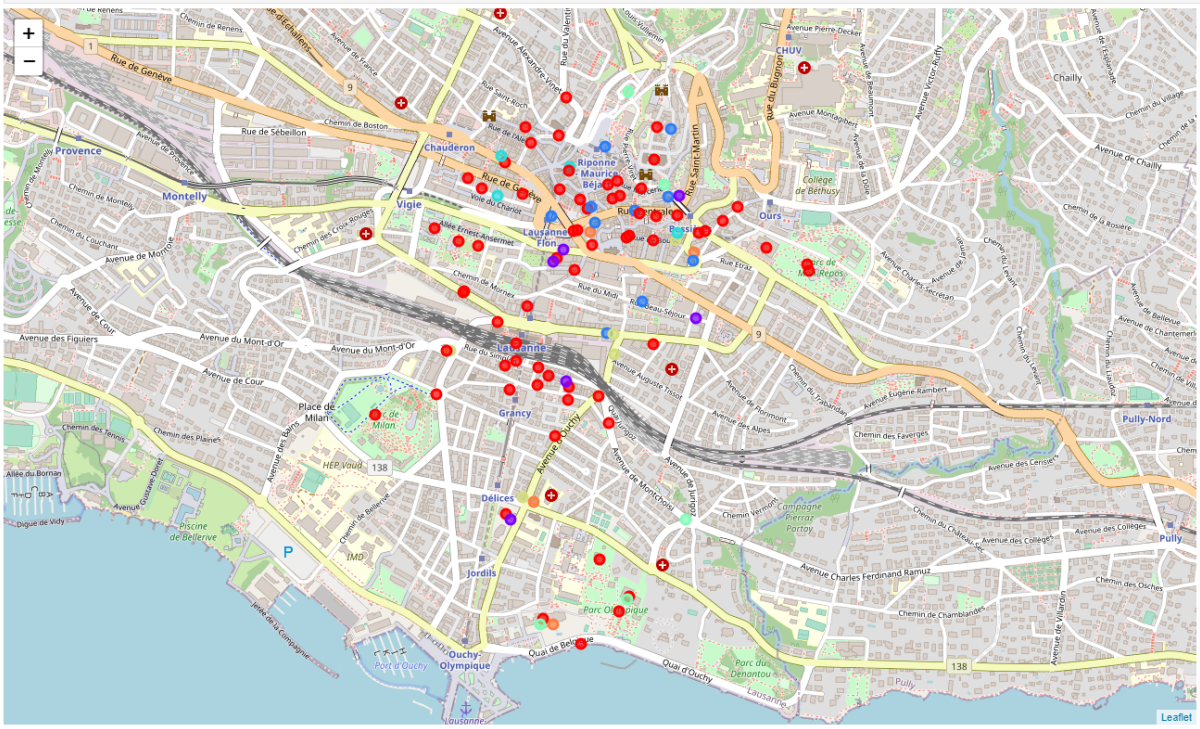


**Figure 6:** Final results

## RESULTS

The first cluster (Cluster label 0) is the biggest cluster and is a cluster with miscellaneous venues. Cluster 2,6 and 7 (Cluster label n-1) are small clusters and contain French, Italian and other restaurants.
Cluster 3 contains bars, Cluster 4 contains mainly fast food places and Cluster 5 contains Cafes .

## CONCLUSIONS

Based on what we learned about the clusters, we now can suggest a place to live, if people ask us. As an example, let's take again the business case we introduced at the beginning of the project: a young worker, who wants to find a place near pubs and cheap restaurants, close to the heart of the nightlife of the city. In this case, analysing the cluster, we would suggest him/her to find a place in Cluster 3 or 4. In addition to that we could also analyse the heat map, to see where places are more dense, meaning they are closer to one another (and that's the recipe for nightlife!).