IBM DATA SCIENCE

COURSERA CAPSTONE FINAL PROJECT

SPORT NUTRITION STORE @ GRANADA, SPAIN

Enrique Benítez López

February 2020

Introduction

Nowadays, there can be no doubt that diet and sport practice play an important role in our lives. They have been scientifically proven to be the major factors in the promotion, restoration and maintenance of good health throughout the entire life course. There is currently a huge amount of information accessible to everybody as well as an increase on sport facilities. People do more exercise and are more aware of the importance of a good diet in his purpose to live inside a healthier body.

The problem

The objective of this capstone project is to analyze and select the best location in the city of Granada, Spain to open a new sports nutrition store. Using data science methodology and machine learning techniques like clustering, this project aims to fin what is the best location in Granada to start a sports nutrition business.

The audience

As a healthy lifestyle is becoming more and more popular in our current society, lots of investors are deciding to put their money in projects related to wellbeing, such as gyms, sport infrastructure or specific nutrition stores.

About the city…

Granada is the capital city of the province of Granada, in the autonomous community of Andalusia, Spain. Granada is located at the foot of the Sierra Nevada mountains, at the confluence of four rivers, the Darro, the Genil, the Monachil and the Beiro. It sits at an average elevation of 738 m (2,421 ft) above sea level, yet is only one hour by car from the Mediterranean coast, the Costa Tropical. Nearby is the Sierra Nevada Ski Station.

In the 2005 national census, the population of the city of Granada proper was 236,982, and the population of the entire urban area was estimated to be 472,638, ranking as the 13th-largest urban area of Spain. Its nearest airport is Federico García Lorca Airport.

The Alhambra, an Arab citadel and palace, is located in Granada. It is the most renowned building of the Islamic historical legacy with its many cultural attractions that make Granada a popular destination among the tourist cities of Spain. The Almohad influence on architecture is also preserved in the Granada neighborhood called the Albayzín with its fine examples of Moorish and Morisco construction. Granada is also well-known within Spain for the University of Granada which has an estimated 82,000 students spread over five different campuses in the city.

Data

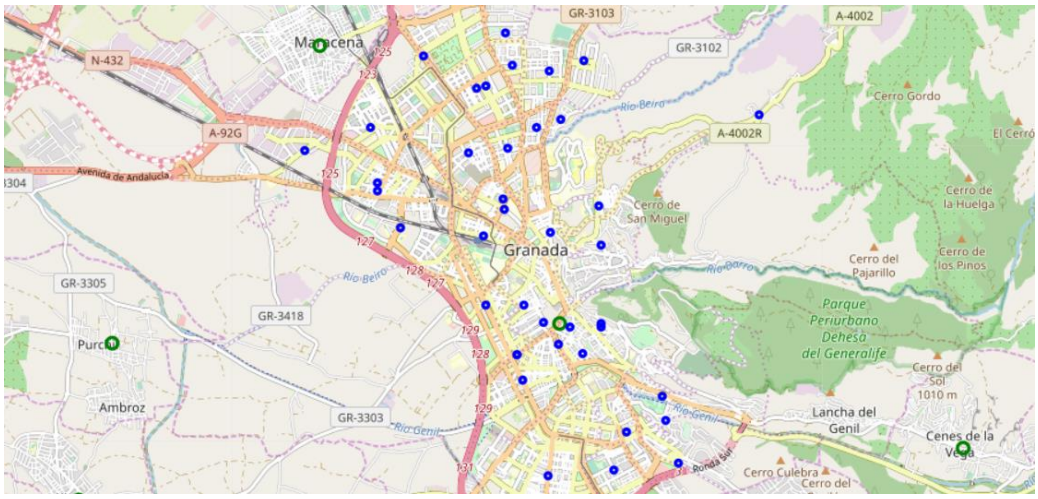
In order to solve the problem, three different types of data will be needed:

- **List of neighborhoods and municipalities in Granada**

  Neighborhoods of Granada are available in this link of Wikipedia whereas the XLS file with the list of municipalities has to be downloaded from the SIMA platform of the Junta de Andalucía, which is the organization that manages all official data of Granada.

- **Latitude and longitude coordinates of those locations**

  We will get the geographical coordinates of the neighborhoods using python Geocoder, paying attention to it in order to get the right geographical coordinates of our neighborhoods of interest.



- **Venue data, particularly data related to Gyms**

  We will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API Will provide many categories of the venue data, we are particularly interested in the **Gym** category in order to help us to solve the business problem put forward.

This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of neighborhoods in the city of Granada, that we can found in the following link: https://es.wikipedia.org/wiki/Distritos_de_Granada. We also will download the coordinates of all the municipalities of Andalusia through, although after considering include them or not, we finally will discard them.

We will scrap the data from the Wikipedia page by using python requests, what will give a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Granada.

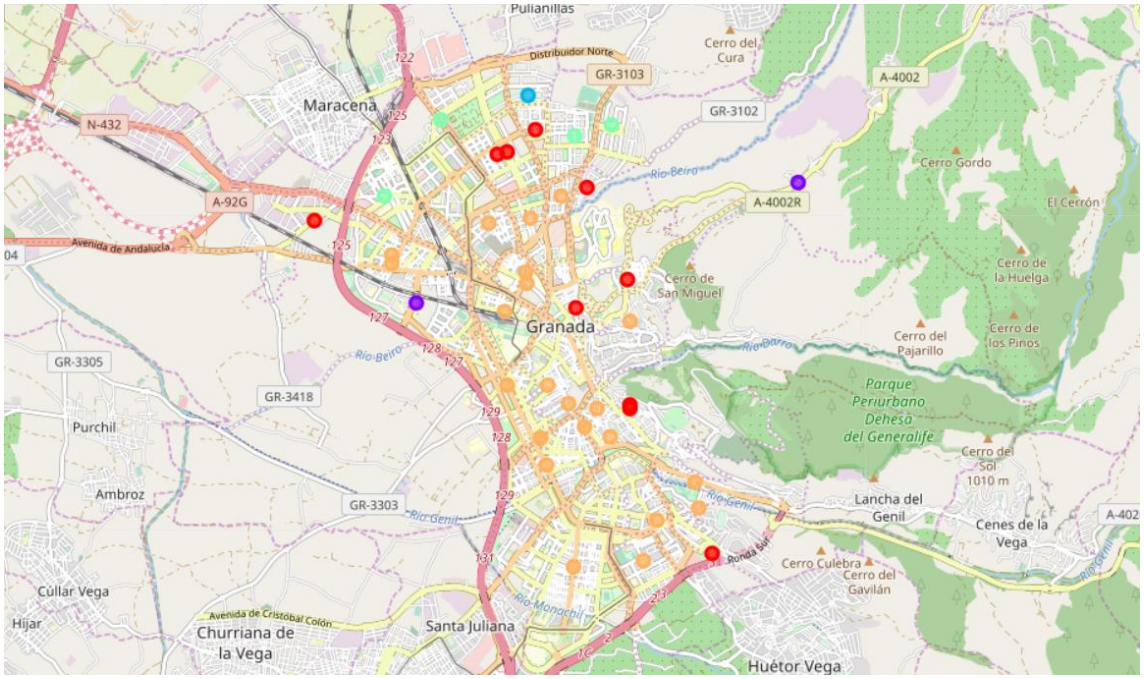
Next, we will use Foursquare API to get the top 100 venues that are within a radius of 500 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the Gym data, we will filter the Gym as venue category for the neighborhoods.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 5 clusters based on their frequency of occurrence for Gym. The results will allow us to identify which neighborhoods have higher and fewer concentration of gyms.

Based on the occurrence of gyms in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open a sports nutrition store.

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 5 different cluster:



The information of the clusters is as follows:

- Cluster 0: Red, 11 neighborhoods, mainly shops and restaurants.
- Cluster 1: Purple, 2 neighborhoods, mainly restaurants.
- Cluster 2: Blue, 1 neighborhood, a shop.
- Cluster 3: Green, 4 neighborhoods, mainly gyms!
- Cluster 4: Orange, 23 neighborhoods, mainly restaurants.

Let's see the information of cluster 3:

Barrio	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Almanjáyar	Bar	Breakfast Spot	Winery	Campground	Gastropub	Garden	Food	Dry Cleaner	Diner	Deli / Bodega
Cartuja	Gym	Bar	Lake	College Residence Hall	Park	Castle	Gastropub	Garden	Food	Dry Cleaner
Cerrillo de Maracena	Gym	Stadium	Breakfast Spot	Park	Diner	Bar	Food	Dry Cleaner	Garden	Deli / Bodega
Parque Nueva Granada	Gym	Lake	College Residence Hall	Park	General College & University	Gastropub	Garden	Food	Dry Cleaner	Diner

Discussion

As observations noted from the map in the Results section, most of the gyms are concentrated in cluster 3. This represents a great opportunity and a high potential area to open a sport nutrition store as there is a big number of potential clients, due to the density of gyms in the area. From another perspective, the results also show that there could be also opportunities to open a gym in lots of neighborhoods in Granada, as there aren't other areas with great numbers of gyms.

Therefore, this project recommends property developers to capitalize on these findings to open new gyms in neighborhoods in cluster 3 with little to no competition. We also recommend the option to invest in gyms in other parts of the city, as it does not seem to be a great number of options there.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 5 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new sport nutrition store. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 3 are the most preferred locations to open a new sport nutrition store as the have the higher number of gyms. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new sport nutrition store.