Finding the optimal location for a business

MADRID NEIGHBOURS DATA

IBM APPLIED DATA SCIENCE CAPSTONE - RJS

Problem description

In this project, the problem attempted to solve will be to find the best possible location or the most optimal, for a Mexican restaurant in the city of Madrid, Spain. To achieve this task, an analytical approach will be used, based on advanced machine learning techniques and data analysis, concretely clustering and perhaps some data visualization techniques.

During the process of analysis, several data transformations will be performed, in order the find the best possible data format for the machine learning model to ingest. Once the data is set up and prepared, a modeling process will be carried out, and this statistical analysis will provide the best possible places to locate the Mexican restaurant.

Data presentation

1. The Foursquare API: This data will be accessed via Python and used to obtain the most common venues per neighborhood in the city of Madrid. This way, it is possible to have a taste of how the city's venues are distributed, what are the most common places for leisure, and in general, it will provide an idea of what people's likes are.

2. The Madrid City Hall's Web Portal: This site provides several data sources of great utility to solve this problem. The files are provided in Excel format, and they are built over a statistical exploitation and use basis. The data contains updated information about the immigrant population per country and per nationality. This data will be analyzed in such a way that one could determine the best location of r anew venue/restaurant/other based on people's nationalities. For the sake of simplicity, it will be assumed for this exercise that people's likes vary according to their nationality, and that people from one specific country will be more attracted to place that matches the environment and culture of their own countries, rather than the ones from foreign countries.

Methodology

The methodology used to approach this problem includes some statistical exploration of the data and some visualizations. The main machine learning technique involved in the development of this project is clustering, in concrete the K-Means algorithm was used, implemented with Python.

https://github.com/RjimenezS/IBM-Capstone-Project/blob/master/RJS-FinalCapstone_notebook.ipynb

Results

The results obtained were five clusters of very different population and venues distribution. The following is a description of the clusters:

- Cluster One:

  Mostly inhabited by south Americans, Europeans, and north Americans. The most common venues are tapas restaurants, Argentinian restaurants, pizza places, supermarkets and Spanish restaurants, among many others.

- Cluster Two:

  This cluster is composed only by 2 different population kinds: Ukrainian people and Dominican Republic people. The most common venues are gyms, Asian restaurants, eastern European restaurants, grocery stores and bakeries among others.

- Cluster Three:

  This cluster is only composed by Bangladeshi people. The most common places are falafel restaurants, fish markets, fast food restaurants and electronic stores.

- Cluster Four:

  Again, only people from two countries seems to live in this clusters. Ecuador and Bolivia. The most common venues are nightclubs, soccer fields, falafel restaurants and fast food restaurants.

- Cluster Five:

  This is a very variate cluster. Some of the main countries here are Rumania, France, Honduras, Philippines, Paraguay and Morocco among others. The most common venues do also variate. Some of them are Mexican restaurants, Chinese restaurants, breweries, sandwich places, seafood restaurants, coffee shops, Mediterranean restaurants, etc.…

Conclusions (I)

As far as we can see with this data, there are no Mexican populations registered in Madrid. However, in Cluster 1, it is possible to notice that there's a Mexican restaurant located in the "Centro" neighborhood, which is the town center.

If a deeper exam is performed into this cluster, it is noticeable that its living population are mostly Latinos, mixed with some other Europeans, but mainly, the people living in this cluster come from south American countries. Apart of this fact, other kinds of Latin restaurants can be found, like Argentinian restaurants, tapas restaurants, and Italian restaurants. So, it is possible to tell that the inhabitants of this area like these kinds of food.

By following this logic, if we would like to open a new Mexican restaurant in the city or any kind of restaurant in fact, it would only be necessary to find a where are the restaurants similar the one we want to open, study the population in that area, and find similar clusters of population in the city that don't have yet or have very few restaurants like the one we would like to open.

Conclusions (II)

In this example, clusters 4 and 5 could make a good match for our target population. Looking at the venues in these clusters, it is possible to find one Mexican restaurant, and a good bunch of fast food, Argentinian, and south American restaurants. So, in these clusters, it is possible to state that the existing restaurants matches the population's nationalities and tastes.

In conclusion and taking into consideration the explanations given above as well as the data, it is highly possible that clusters 4 and five could be a good place to open our Mexican restaurants. As explained above, the same logic could apply to open other kind of restaurant or business in any other area of the city. It is only necessary to examine the existing businesses in our target area, and study the population, then compare these two factors with the same ones in areas where there are existing businesses like the one we want to open, and then verify if the matching is correct.