Portugal an opportunity

Data Science by IBM/Coursera

Final Capstone

by

Cláudio Ferreira

Introduction

State of art

Portugal is located in southwestern Europe, comprising a continental part and two autonomous regions: the Azores and Madeira archipelagos. Due to a prime location by the Atlantic Ocean. The Portuguese climate is temperate maritime or Mediterranean, temperatures are mild all year round and with many sunny days allied to the vast sea coast becomes a great invitation for foreign tourists to visit. In addition Portugal offers a rich gastronomy based on the Mediterranean diet, excellent wines, differentiated hotels, beautiful landscapes, historical monuments, popular culture, nature sports etc. Proof of this are the numerous awards won in the tourism like this year Portugal has won 23 World Travel Awards.

Portugal is trendy, according to the latest statistics in 2018, Portugal received over 12 million foreign tourists, which translated into an 8.1% increase over the previous year (Observador.pt). Portugal is also one of the countries with the largest increase in visitors from the European Union. More important than the amount of tourists received annually is the quality of tourism offered by Portugal, which attracts not only foreigners but also Portuguese residents.

Problem

Imagine that an investment group wants to invest in a successful business in Portugal to take advantage of the growth of tourism. The aim of the project will be to answer the following questions:

1. What kinds of businesses the cities already offer?
2. In which cities they could invest?
3. What kind of investment should the investment group make??

This project will be based on the segmentation and description of each city with the most visitors. The cities will be characterized according to the different types of commercial establishments (venues). Finally there will be a reflection and discussion on what kind of investment could be made and in which cities this investment should be made in order to maximize its potential. To achieve the three main goals I will use some techniques of data cleaning, data processing, data visualization and machine learning.

Data acquisition and cleaning

Data acquisition

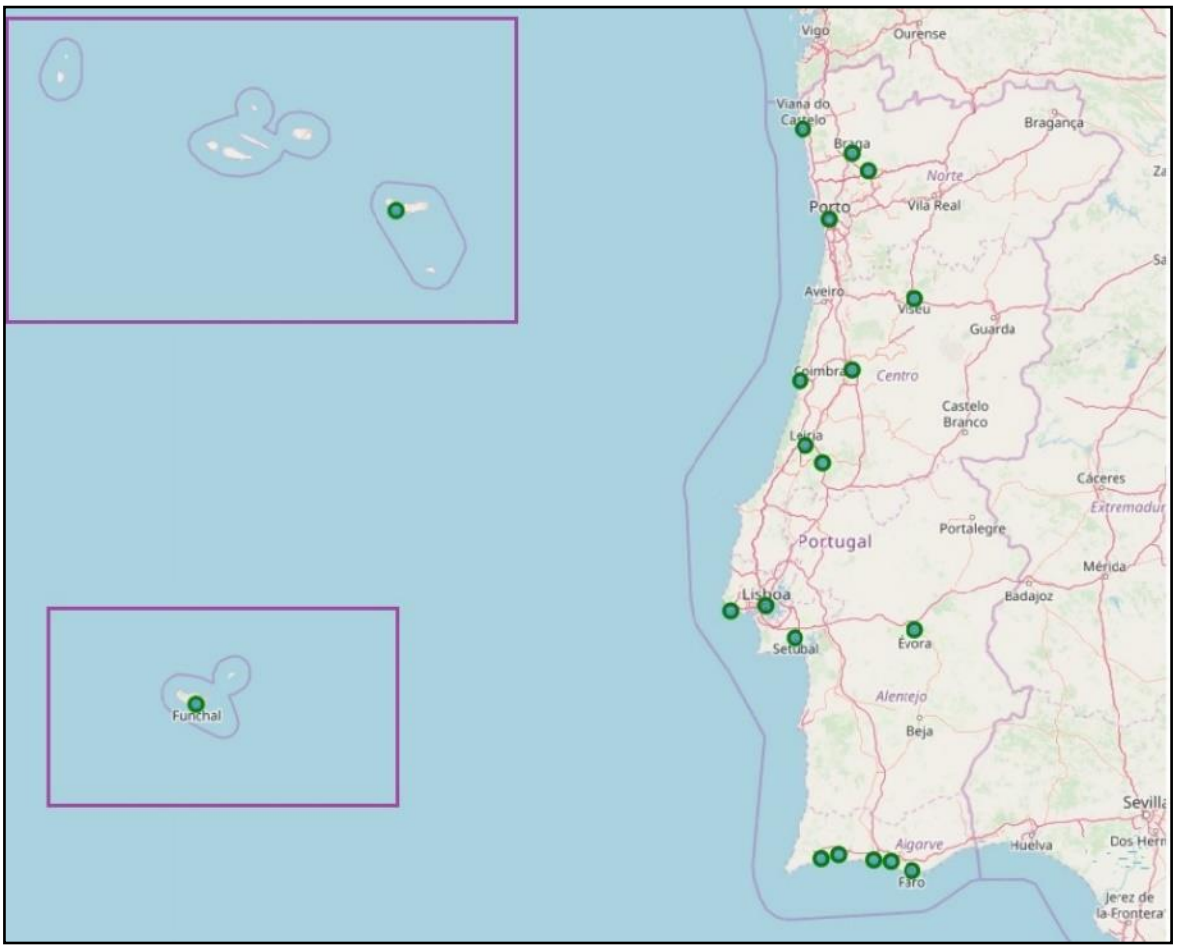
To answer to the problem posed, I used a dataset that we can find in Pordata. The dataset had data with the number of hotels and accommodations in Portugal, districts and cities, I choose this information as an index to know the most visited cities. I obtain the cities coordinates from geopy and google maps to fix some coordinates of geopy. In order to get a description of the venues of each cities I used the foursquare API database.

Cities	Hotels	latitude	longitude
Lisboa	189	38.725147	-9.144997
Porto	89	41.149451	-8.610788
Ourém	54	39.633666	-8.670023
Funchal	43	32.649650	-16.908678
Albufeira	42	37.089506	-8.246843
Cascais	27	38.696812	-9.424695
Ponta Delgada	24	37.739434	-25.668362
Braga	24	41.551058	-8.428005
Loulé	22	37.079227	-8.110704
Portimão	20	37.126283	-8.538482
Leiria	19	39.743790	-8.807112
Évora	18	38.570774	-7.909281
Lagos	18	37.099555	-8.676873
Coimbra	17	40.210980	-8.429206
Guimarães	15	41.441768	-8.295571
Viana do Castelo	15	41.694867	-8.831088
Figueira da Foz	15	40.148282	-8.855414
Viseu	13	40.657471	-7.913866
Faro	13	37.019869	-7.929246
Setúbal	13	38.523410	-8.894496

Table 1 – Dataset with cities and coordinates.

Data cleaning

The dataset from Pordata is very complete, have the information of several years and about all Portuguese districts and cities. For this project I drooped the data about the districts and I focus only at the 20 cities with the most number of hotels. Once selected the 20 cities I used the geopy to obtain the coordinates of each city. I had to change some coordinates that are incorrect, and I created a new dataset. The city of Matosinhos was discarded because it is an adjacent city of Porto, strongly influenced by the tourist flow of Porto. Not being in itself a city of tourist interest, it works as an expansion of Porto.



Map 1 – Folium map of the 20 Portuguese cities.

Methodology

Data processing

To achieve the goal of making a successful investment, enjoy the growth of tourism in Portugal. I need to know firsthand which cities have more visitors. After intense research, I discovered without much surprise that several newspaper articles indicate that Lisbon, Porto, Fatima (Ourem), Évora and other cities in the Algarve region are the cities with the most visitors daily. Unfortunately I was unable to obtain the original data in order to know the full list and the authenticity of the data. I decided to opt for a representative index of tourist affluence, the number of hotels per city registered by INE. I use panda's library to achieve the list of 20 cities present in Table 1.

The next step was to create a table that grouped the different venues categories by city, which will possibility the use of a clustering technique. To obtain the coordinates of which city I used the geocode and the methodology described here. The coordinates will use to do the request venues in Foursquare API database. The request was restricted to a radius of 3000m and a maximum of 200 venues, around the center city.

Data visualization

Folium library was used to two times to confirm the correct location and identification of each city on the Portugal map, and to observe possible similarities in the distribution of cities belonging to the same cluster. Because some coordinates was wrong and to improve the search for venues on Foursquare and to be more reliable and standardized, it was necessary change some coordinates to the cities central areas. The Map 1 show the final localization of all twenty cities.

Data Clustering

In order to identify similarities between the selected cities, I resorted to k-means clustering technique. The Table 2 was the data fit to cluster the 20 cities. I use Pandas and SKLearn (scikits-learn) to implement the K-Means. The number of cluster was two decided based on elbow method (fig 1).

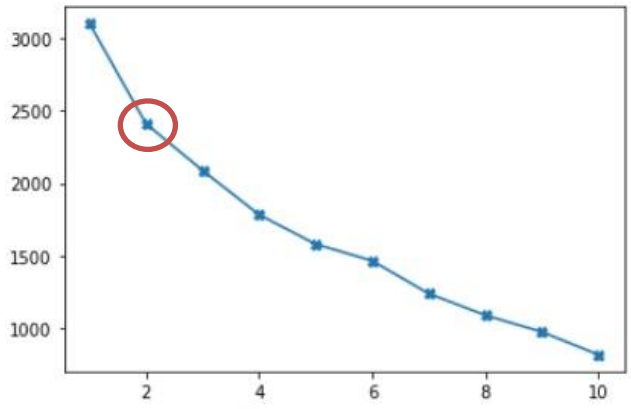


Figure 1 – Number of K, elbow method.

Results

Determination of cities with the greatest tourism potential

Portugal has it in 2018, three hundred and eight municipalities (municipalities and cities) with accommodation available to receive resident or international tourists, according to INE data. Since the goal is to invest in a business that is open to the visiting public of cities, the more popular and attractive the city is for tourists, the greater will be the success and income that this business will generate. On this premise the study focused on the cities that receive the most visitors and who spend the most time in each city. For this reason I analyzed the number of hotels per city to determine which cities have the most tourists.

The twenty cities are distributed throughout the Portuguese territory including in the archipelago of Madeira and Azores. The cities represented are the largest urban centers of Portugal, such as Lisbon, Porto and Setubal, but also cities in the interior of the country such as Évora, Viseu, Coimbra and Guimarães, which are important historical centers. As is shown in Figure 2 Lisbon is the city with more visitors they spend on average 2.4 nights (by INE). On the other hand Faro and Setubal are the territory with the least hotels among the selected cities. In 2018 Setubal owns thirteen hotels and its visitors spend an average of 1.9 nights. I highlight the presence of the city of Ourem, which is not in itself a focus of tourism, but the high number of hotels in the municipality of Ourem is due to the strong religious tourism to the Fátima Sanctuary, so the coordinates used to obtain the request are from the center of the Fátima Sanctuary, instead of the center of Ourem.

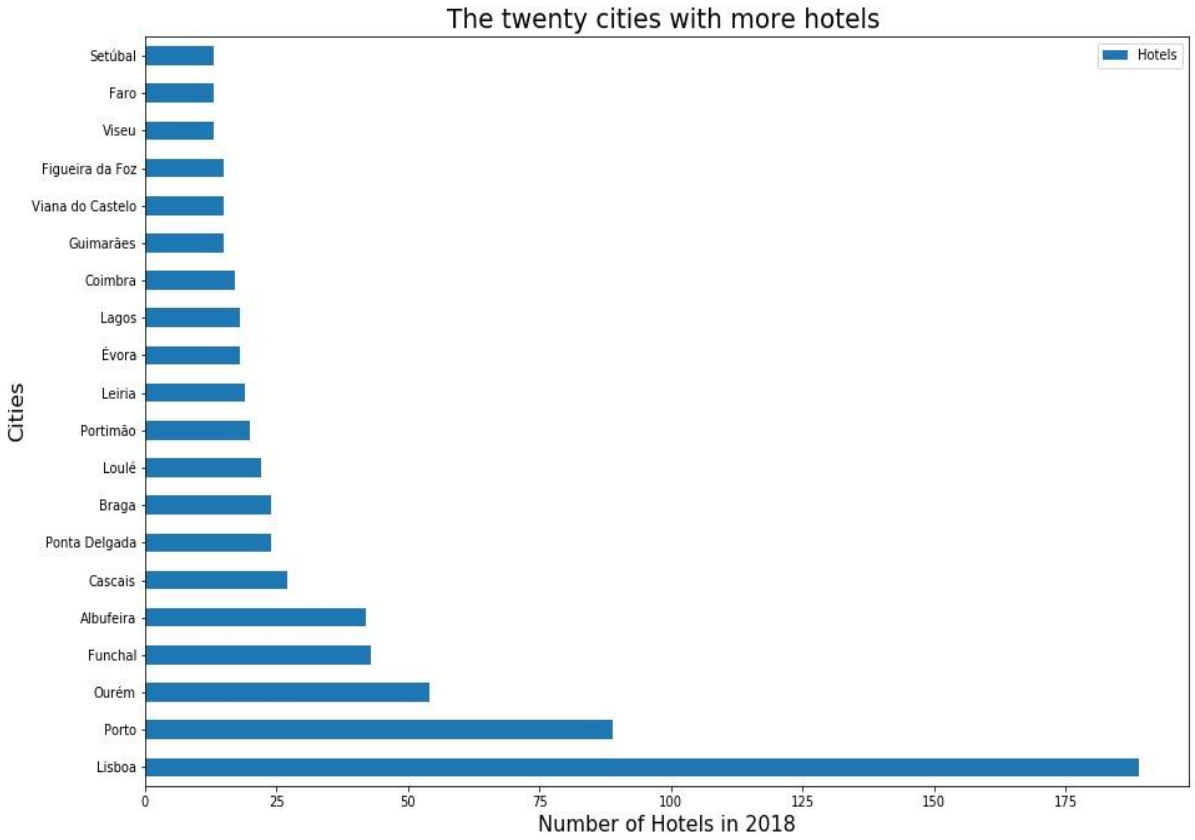


Figure 2 – Bar plot with cities vs hotel number in 2018.

Exploring Cities Venues

To understand and characterize each city, was coded a Foursqare loop request based on table 1. The result was a total of 1779 venues, distribuided for 178 venues categories. Most cities have a maximum number of 100 venues, other cities as Ourem, Loule and Figueira da Foz have a lower number of venues 55, 52 and 40 respectively. To put it simply, all venues was grouped by category in a new dataset PT_KM. Some cataegories was aggregataded in just one, as like the differents kinds of fast food restaurants there were grouped into a single Fast Food Restaurant category and equally Asian restaurants were all grouped under Asian Restaurant category, which reduced from 178 to 165 categories.

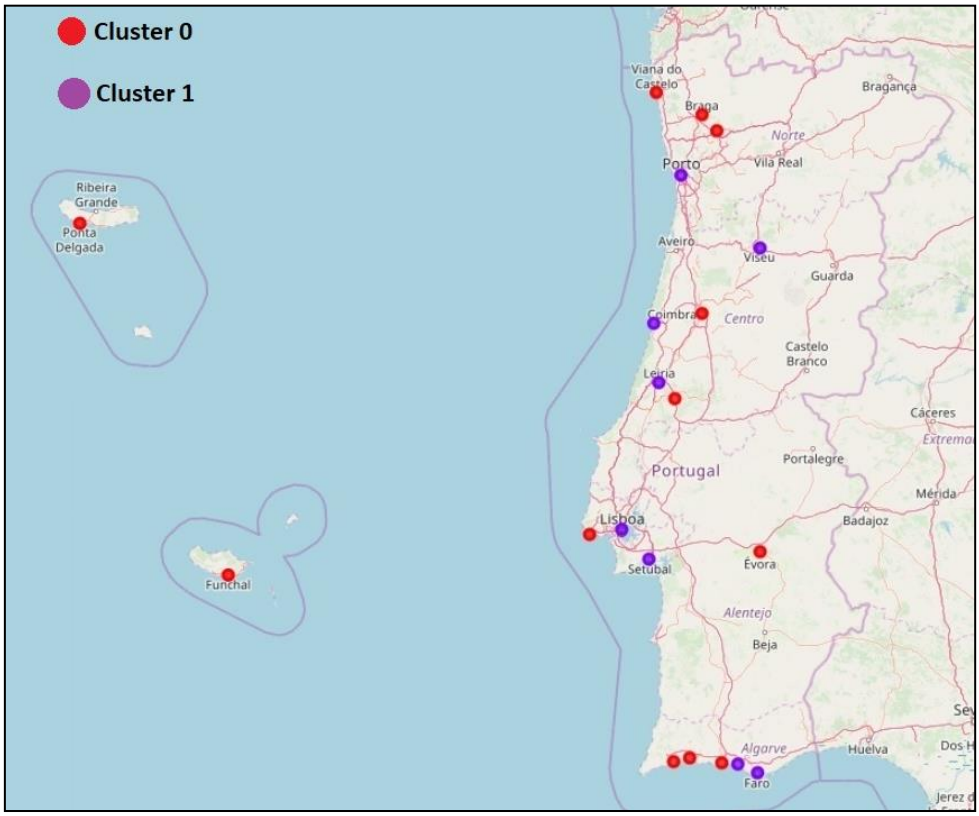
Analyzing each city individually we can see that the most reprehended Venue category in all cities except Lisbon and Figueira da Foz are the Portuguese restaurants. The most represented categories are cafes (coffes) and hotels. I highlight that in Porto, Faro and Leiria the hotels are not one of most common categories (Table 2).

Cities	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Albufeira	Portuguese Restaurant	Hotel	Bar	Beach	Italian Restaurant
Braga	Portuguese Restaurant	Bar	Café	Asian Restaurant	Hotel
Cascais	Portuguese Restaurant	Asian Restaurant	Hotel	Seafood Restaurant	Beach
Coimbra	Portuguese Restaurant	Plaza	Café	Hotel	Bar
Faro	Portuguese Restaurant	Fast Food Restaurant	Bakery	Bar	Tapas Restaurant
Figueira da Foz	Café	Portuguese Restaurant	Beach	Fast Food Restaurant	Hotel
Funchal	Portuguese Restaurant	Hotel	Bakery	Café	Bar
Guimarães	Portuguese Restaurant	Café	Bar	Hotel	Plaza
Lagos	Portuguese Restaurant	Fast Food Restaurant	Beach	Resort	Hotel
Leiria	Portuguese Restaurant	Café	Fast Food Restaurant	Plaza	Asian Restaurant
Lisboa	Hotel	Portuguese Restaurant	Asian Restaurant	Scenic Lookout	Café
Loulé	Portuguese Restaurant	Hotel	Café	Beach	Lounge
Ourém	Portuguese Restaurant	Hotel	Café	Historic Site	Trail
Ponta Delgada	Portuguese Restaurant	Hotel	Café	Seafood Restaurant	Fast Food Restaurant
Portimão	Portuguese Restaurant	Hotel	Beach	Fast Food Restaurant	Seafood Restaurant
Porto	Portuguese Restaurant	Bar	Hostel	Café	Ice Cream Shop
Setúbal	Fast Food Restaurant	Café	Portuguese Restaurant	Seafood Restaurant	Fish & Chips Shop
Viana do Castelo	Portuguese Restaurant	Bakery	Café	Fast Food Restaurant	Bar
Viseu	Portuguese Restaurant	Café	Bar	Fast Food Restaurant	Hotel
Évora	Portuguese Restaurant	Hotel	Historic Site	Café	Bed & Breakfast

Table 2 – Top 5 most common venue.

Exploring the Clusters

The next step was cluster the 20 cities based on PT_KM dataset the result was 2 clusters, “0” and “1” (Map 2). The Cluster “0” is characterized by a greater avarage per city of Portuguese Restaurants, Hotels, Bar, Bakery, Asian Restaurant. These cities have a smaller area, fewer residents and are cities more focused on tourism as major source of income. On other hand the Group “1” is characterized by a higher avarage of Coffes and Fast Food restaurant, Parks and Hostels (Figure 3), in this group are present the biggest cities, like Lisboa, Porto and Setubal, so they are cities with a great resident population and also are important business centers.



Map 2 – Representation of Cluster “0” and “1”

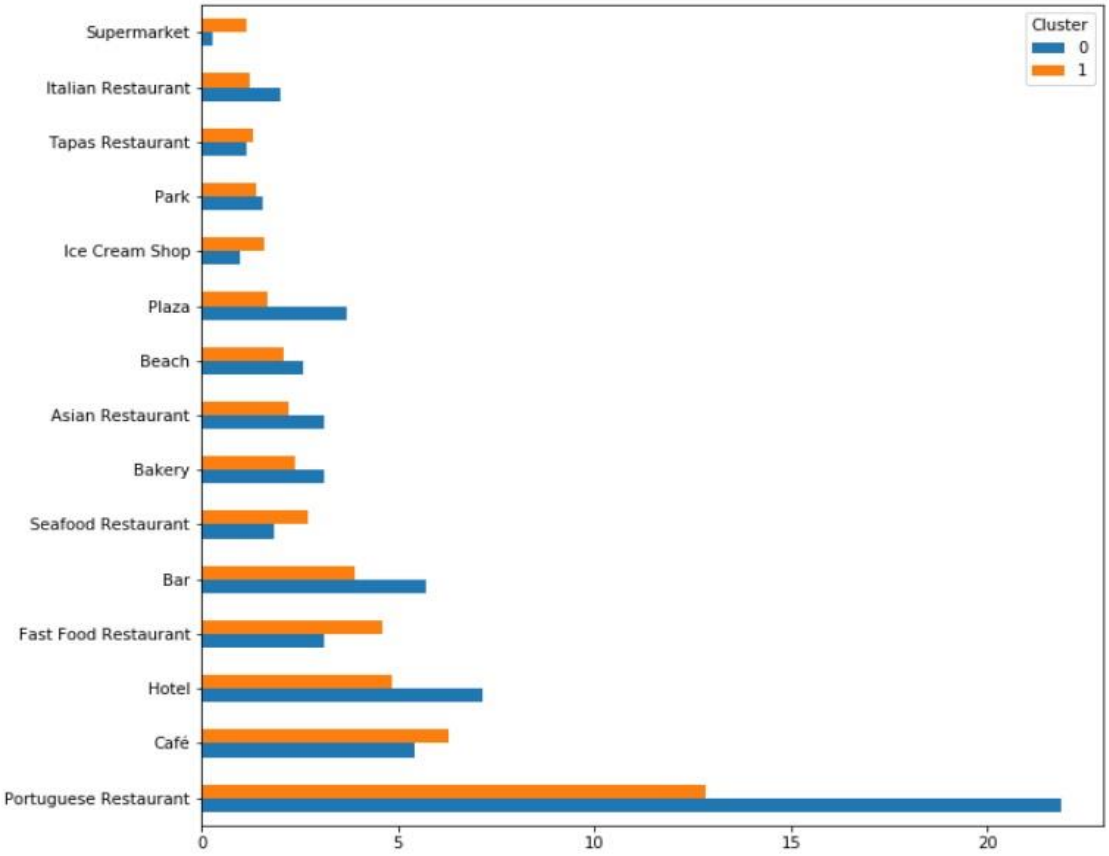


Figure 3 – Representation of the 2 clusters.

Discussion

Recalling that the main objective of this project is to find one or more cities with the potential to invest in a business that exploits the success of tourism in Portugal. In fact Portugal has several cities with great tourist potential, and the success of an investment depends on several factors, such as product quality, customer service, creativity, demand rate, supply value among other factors. Considering the results obtained in this case study, cities in cluster 0 are more prepared to receive visitors and offer a wider range of experiences, such as local restaurants, hotels, bars or tradidional bakeries. In order to proceed with an investment in one of these cities, further study is necessary to create a differentiated business from the existing ones.

Analyzing the cities belonging to cluster "1" they are the cities with the largest resident population and the largest number of annual visitors (INE, IPDT), the amount of venues offered in the central areas of the cities is less than cluster cities "0". I highlight that Faro and Porto did not have a relevant number of hotels when compared to other categories of venues. This becomes more relevant when we find that Porto is the second Portuguese city with more hotels, and that Porto and Faro are two cities that have an airport, that is, a gateway for tourists in the country. It would be expected that these two cities had a higher visitor retention capacity. In addition, in Lisbon there are nine tourists for each resident, in Porto there are eight visitors for each resident, ie the difference between the two cities is small but the supply of accommodation in Lisbon is more than twice as large as Porto. In my opnion Porto and Faro will be the best bet to invest.

Another point to note is that coastal cities have greater access to fresh fish and seafood, so it is natural that there is a greater supply of seafood restaurants in these regions. However we can see in table 2 and figure 3 that there is little seafood restaurant supply in some coastal cities. Although there may be seafood restaurants classified as Portuguese food restaurants, generating false positives. Investing in a seafood restaurant can be a good investment in cities like Porto, Faro, Lisboa or Figueira da Foz but further study of the existing offer is required.

Future directions

I believe that the work done on this project may be the beginning of a more in-depth future study. With access to more data and a more reliable database than Fousquare, it is possible to use machine learning techniques to help the decision making of an investor group.

It should be noted that the Foursquare platform does not have a large representation in Portugal, ie has few users and is certainly not the best database to carry out this type of projects. It is necessary to find a platform with a database similar to Foursquare but with a larger number of Portuguese users and more updated.

Conclusions

The analysis of the twenty cities shows that Portuguese restaurants are the most represented business in practically all cities. For this reason, I believe that investing in catering will have a high risk of failure.

Given the results of K-means clustering and the prior knowledge of each city, I can conclude that cities in cluster "1" have more potential to invest. These cities receive thousands of tourists annually, but also many business visitors. Venues found on average a lower number compared to cities in cluster "0", ie there is potential for investment growth.

In short, given the results I recommend that the investment made in the construction of a hotel in the cities of Faro and especially Porto.