**A successful investment**

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. In which cities they could invest?
2. What kinds of businesses the cities already offer?
3. What are the existing needs in this cities?

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 (veneus). 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](https://www.pordata.pt/Municipios/Alojamentos+tur%c3%adsticos+total+e+por+tipo+de+alojamento-746). 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.

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.

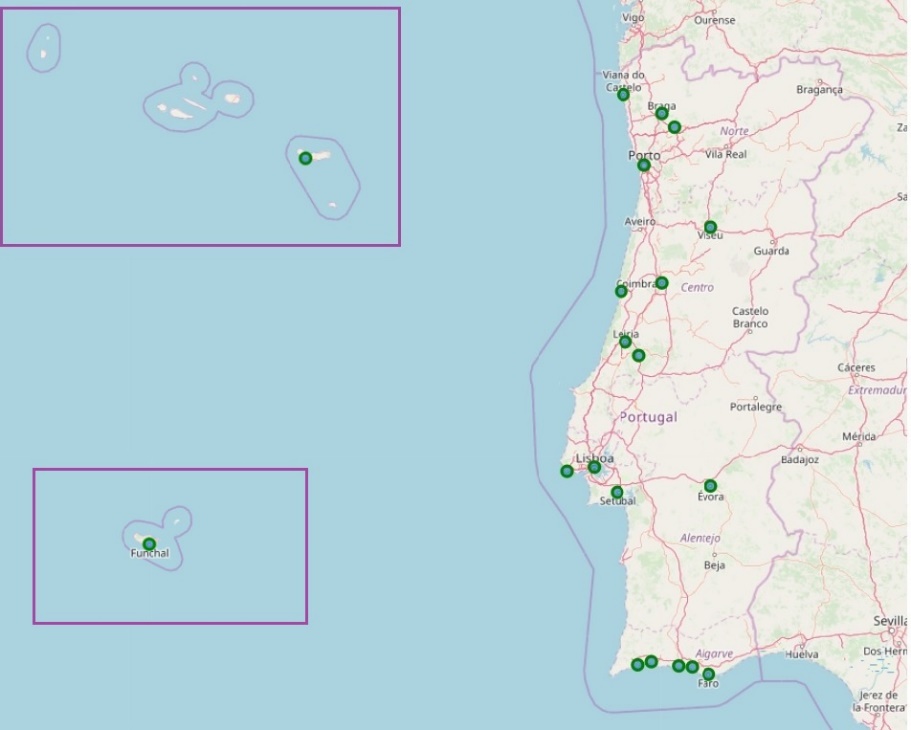


Fig 1 – Folium map of the 20 Portuguese cities.