IBM Data Science Professional Certificate

Capstone Project - The Battle of Neighborhoods

Introduction

I am from Portugal as such I know that Algarve, located at the south part of the country, is a bustling tourist destination, especially for sun-starved northerners, this makes it Portugal's biggest tourist attraction.
Hence it is ripe with business opportunities, be it in form of new hotels/inns or expansion of the already existing nightlife.

Objective

Here I will study the distribution and concentration of the above mentioned businesses . With this in mind it will be possible to determine the best geographic location for a new business.
As such this is an excellent opportunity for a small business owner who wishes to expand into the thriving tourist area of Algarve.
This work will provide an overview of Algarve and locations where the ratio of hotels to restaurants, or vice -versa, are low enough to warrant a construction of a new one, be it a hotel or restaurant.
The type of expansion will be dependent on the concentration of existing businesses in a certain radius.
The same techniques could be used for any other region and any venues, not just hotels and restaurants.

Data

The data acquired from Portuguese Post Office (www.ctt.pt) and restructure to csv file for easier manipulation and reading.

Another aspect of this project is the Foursquare data. I believe that the data provided is sufficiently accurate, meaning although we are using Foursquare data for segmentation and clustering, the amount and accuracy of data captured can't 100% determine correct classification in real world.

Methodology

The work was divided in two parts, the first (The Battle of Neighborhoods - Data Preparation) obtained and reduced the raw data into the desired data. The second (The Battle of Neighborhoods - Final Capstone) analysed the data to obtain the desired conclusions.

First Part

Using the previously mentioned data from the Portuguese Post Office (www.ctt.pt), I obtained the locations of all Districts and their constituents (Concelhos), afterwards I focused on the District of Algarve and obtained the geographic coordinates of all available locations. These were used as centers for hotel searches with Foursquare with a radius of 10km. A filtering algorithm was also used to remove all duplications, since some location could be located within 10km of one another. Around each of the unique hotels a search for 1)foodstuffs: restaurants, cafe, diner, bistro, etc and 2) nightlife: bars,nightclubs, pub, etc was done, here in the radius of 1km. Due to a limitation of Foursquare only 50 venues could be determined at a time, meaning that around each hotel a maximum of 50 type (nightlife/foodstuff) venues were found.
In order to have a better grasp on the data, all data was outputted as a ratio to the maximum of venues found. All data was saved and further analysed in the second part.

Second part

First a geographic center of Algarve was calculated, in order to have a centered reference point for plotting.
Next the the geographic coordinates (latitude,longitude) of all the hotels were extracted from the “location.labeledLatLngs”, since the location data returned by Foursquare is not directly accessible.
Finally Algarve was over-plotted with a density heat map of restaurants around hotels (Fig.1), a similar plot was done with nightlife heat map (Fig.2). The hotels are also visible as black dots.

Results and Discussion

From the graphs shown in Fig.1 and Fig.2 we can see that most of the hotels and neighboring venues are located near the cost. It is also observable that the density of food venues is higher than that of nightlife venues. As such the location and type of venue for a potential expansion would be somewhere along the coast, and in the category of a nightlife venue. In my opinion somewhere near the southwest coast of Algarve.

Limitations and Conclusion

As previously mentioned a small limitation of Foursquare is that it does not allow for more than 50 venues to be requested, as such in the case of high density areas, there is an underestimation of venues, most of them in the food venues.
Another limitation is that in the case that a hotel also includes a built-in restaurant, the restaurant part will not be marked as such.
In conclusion we can assume that despite the limitations the results obtained are still accurate and as such can still be used for prospective investors.

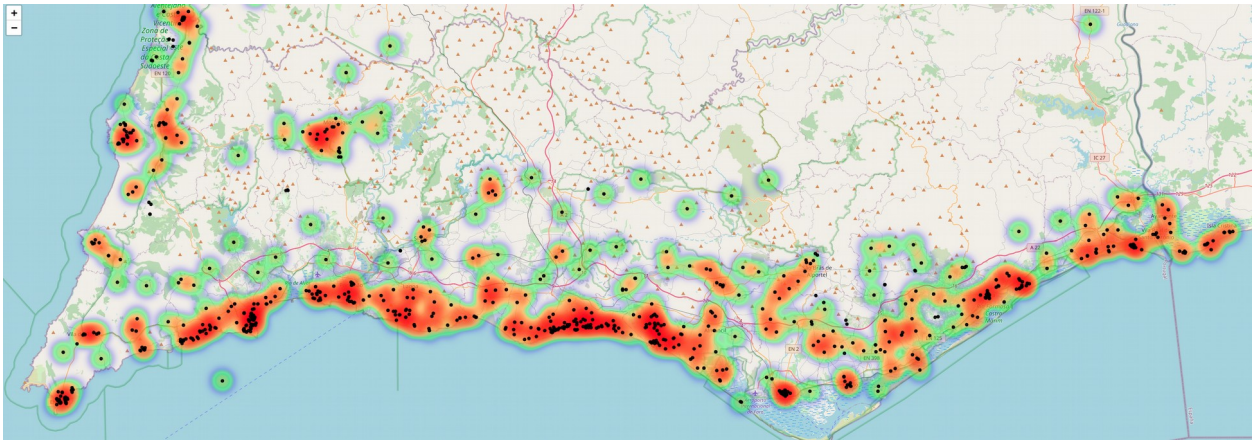

Fig.1 Heat-map of food venues.

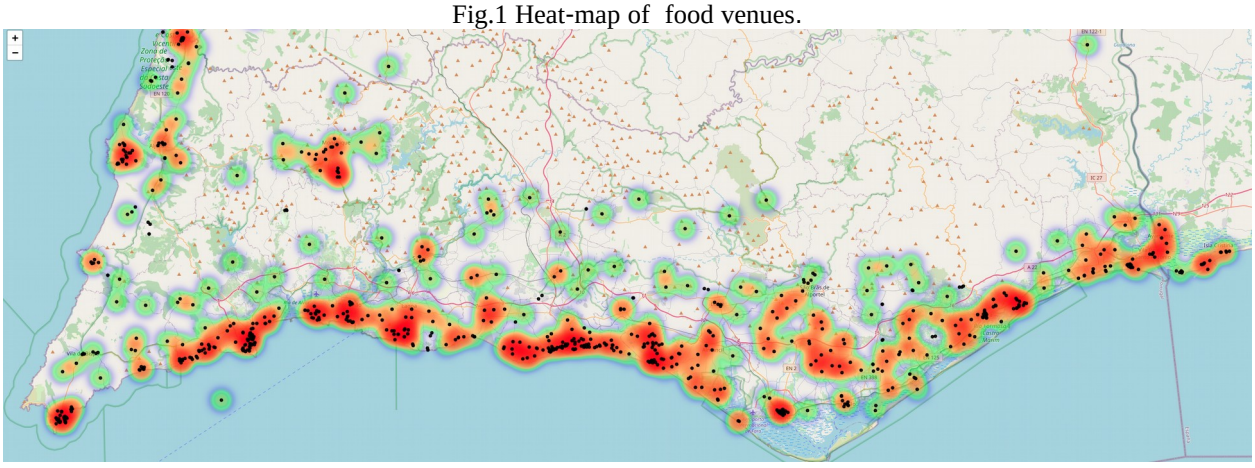

Fig.2 Heat-map of nightlife venues.