IBM Capstone

IBM Applied Data Science Capstone

Sheep-oriented Steakhouse in Florence, Italy

Solano Campos

July 2019

Introduction

For many tourists, enjoy the real culture of the country that you are visiting is a goal. We all know that Italy is very famous for the pizza and the pasta, but, essentially in Toscana Region, a region where the tradition of the agriculture and livestock have maintained along of the years, the creation of high quality sheep's have been develop to a free system of native alimentation, which corresponds to the highest quality of the meat. Italians love meat, being it from cows, sheep's, pigs, birds or deer's, but most part of the meats are used in sophisticated plates or as a side dish, this way, just the better cuts from the meat are utilized, and the others are sold for banana's price. In Uruguay and South of Brazil, the climate and the creation of the sheep are very similar to the Toscana model, in the other hand, the meat is the main dish of the lunch or dinner, ate in a barbecue without any side dish. This has happened in Italy when the hungry have hit Europe, but, as the conditions were getting better, the European stopped to innovate and maintain the tradition and started to eat just noble parts of meat. Regain this culture, in the hearth of Toscana, would be not just lucrative, but a very important step towards perpetuating the roots of a people with a lot of history.

Business Problem

The objective of this Capstone Project is to analyze and select the best location in Florence to open a new Steakhouse. Building all the Project from zero, passing thought the entire methodology, data architecture and machine learning techniques to improve the algorithms along time. Providing the best solution that answers the question: At Florence, if a business man is looking to open a new restaurant, what kind and where you will recommend that he open it?

Target Audience of this project

This project is useful for any entrepreneur with will to create history, any investor that wants to enter in one of the most competitive markets from Italy, that is the Food Industry, with a different idea that will connect people through past to future, perpetuating a culture that would be forbidden. Sheep Breeders are also interested in this project regarding their production and the will that they have to serve their own country instead of keep exporting most part of the meat produced in their fields, as the historical quote says: From the earth, to the earth.

Data

- List of neighborhoods in Florence. This defines the scope of this project, which is restricted to Florence city in Toscana District at Italy.
- Latitude and Longitude coordinates of those neighborhoods, it's required in order to plot maps, discover venues, explore the city and develop the clusters.
- Venue Data, the most important data of the project, will be used as the main variable to perform machine learning clustering and all exploration of the region.

Sources of Data and Methods to extract them

The Wikipedia page (https://commons.wikimedia.org/wiki/Category:Neighborhoods_in_Florence) contains a list of neighborhoods in Florence, with a total of 33. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and *beautifulsoup* packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give the latitude and longitude coordinates of the neighborhoods.

After that, will be used 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 various categories of the venue data, we are particularly interested in the food industries categories 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, will be presented the Methodology section where will be discussed the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Starting from the scratch, we got the list of neighborhoods that exists in the city of Florence. The list is available at https://commons.wikimedia.org/wiki/Category:Neighborhoods_in_Florence. We have used web scraping with Python requests and BeautifulSoup package to extract the list of neighborhood data. With the list ready, and all the

modifications to make the names readable for the future calls that we made (excluding special characters and useless words), it's time to get the geographical coordinates in order to be able to use the FourSquare API. Geocoder package was used, it enabled us to get latitude and longitude for each neighborhood. After that, a Pandas Dataframe was build with all this data, this was our first Dataframe with all the complete data of Florence city. With that, we were able to create a Folium map, pointing each square with marker, allowing us to perform a sanity check to make sure the geographical coordinates returned by geocoder were correct.

Using the FourSquare API, we got the top 100 venues that were in a radius of 1000 meters. To access it, we needed to register in a developer account in order to obtain the ID and Secret Key to run the calls in the API. This way, we were able to make the calls in the API using the credentials and the latitudes and longitudes that we got for each neighborhood in a For-In loop. Foursquare returned the venue data in a JSON format where we'd extracted the venue name, venue category, venue latitude and longitude. With the data, we can check now how many venues were returned for each neighborhood, examine the categories and the relation between venues in each square. It is time to prepare the data to create the clusters, analyzing each square by grouping the rows and taking the mean of the frequency of occurrence of each venue category, this way we will be creating a second important Dataframe that will be used later.

The K-means cluster algorithm was choose to make the analysis and segregate the clusters, the 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 alg. And is particularly suited to solve problems like the one in this project. We will cluster the neighborhoods into 5 clusters based on the similarity of the categories and the number of venues in each neighborhood. These groups will help us to develop the business solution that we have planned.

Results

The results from the k-means show that we must categorize the neighborhoods into 5 clusters:
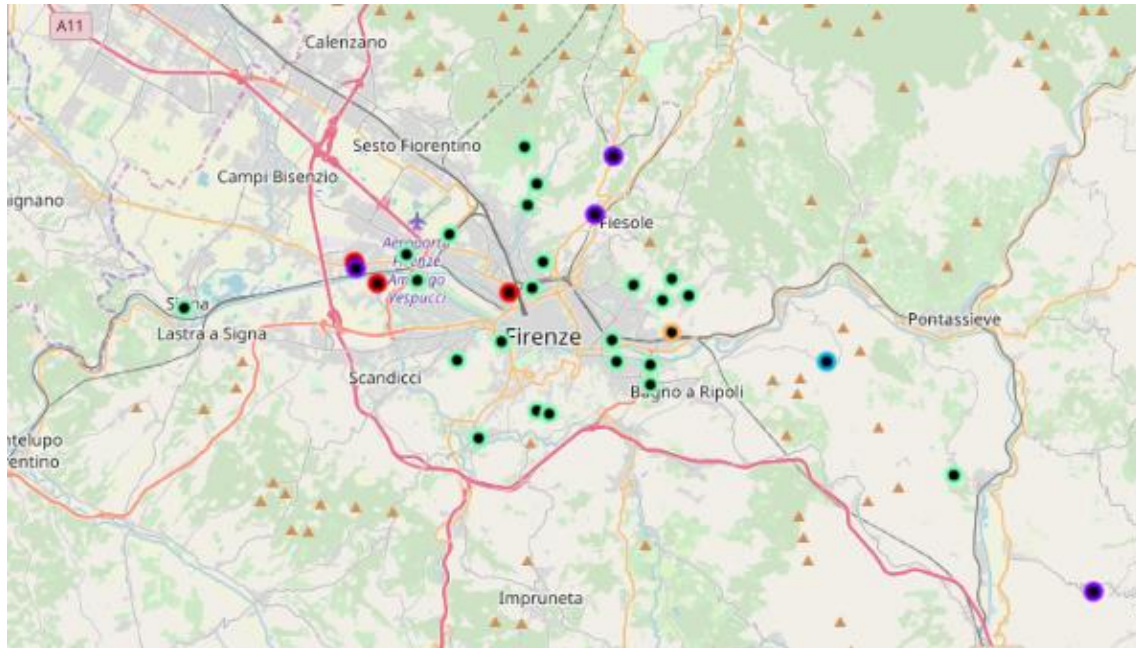
- Cluster 1 (Red) : At cluster 1 we can find the most part of nightclubs, mixed with malls and some food, more residential areas.
- Cluster 2 (Purple): It is a tourist area, with a lot of hotels, restaurants and shopping malls. Includes the airport.
- Cluster 3 (Blue): Lonely cluster, this is an isolated neighborhood.
- Cluster 4(Green): Urban area, most part of the city center is here, most part of the restaurants and hotels are in this area, also the stadium and green areas.
- Cluster 5 (Orange): It's the suburb of the city, closest to the trails and with some green area.

Discussion

Toscana, the major region where Florence is in, is very well known by the quality of the sheep meat with a large production. Italians are very traditional, they love eat pasta, pizza and even barbecue, but for sheep's, they use to serve it in the plate, as a side dish, with some pasta for main meal. In Uruguay and South of Brazil, where the climate is very similar, the sheep's have the same meat quality, but in their culture, they eat the meat as a barbecue, as the main dish, performing a very different and extraordinary experience in steakhouses. The main idea of this project was to identify where is the best place to open a specific Steakhouse that could develop again the sheep culture inside Italy, and the best results was in neighborhoods of cluster 2, mainly in *Bellosguardo, LePiagge* and *Monterinaldi*. Due to the concentration of hotels and the already existence of some restaurants that run away from the traditional pizza and pasta, where people can go and have another good option, that will be the Sheep-house.

Limitations and Suggestions for Future Research

In this project, would be very helpful if we have also the GDP per neighborhood per family, this would drive us to better results in the market, knowing where the money is and measuring how far we must be from each rich or poor neighborhood, other important point will be the pricing of the dishes, as the sheep alone is not a common plate nowadays. Another important data would be the security level of each neighborhood, could be measured by robberies per month, this would help to identify if we would need any special security for the store and the clients.

Conclusion

In this project, we have 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, sheep breeders and entrepreneurs regarding the best location and the best market to open a new business.

To answer the final question that was raised in the introduction section, the solution of the problem is: The best's neighborhoods to develop a new steakhouse (or any new non-Italian restaurant) are *Bellosguardo*, *LePiagge* and *Monterinaldi,* due to their presence of hotels and tourist attraction, their location near from the center, and their pre-disposed food courts, which includes already non-Italian food, but any of them containing a history of the development of the country and their people and culture, this way, the sheep-oriented steakhouse would be the lead of a hole new operation in the food industry at Florence, having just 4 steakhouses, being 3 of them from the same brand, that serves pizza, pasta and cow meat. The players in this market are not ready to a competitivity like that, and the Florence citizens deserves the best meat quality of Europe, that is their production, but is exported in most part of the time. The ecosystem to create this steakhouse is very wealth, creating strong points since the breeder until the final consumer, giving everyone the possibility to honor their own land.