Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Italian Restaurant in Rome, Italy

Introduction

Italy is one of the best tourist destinations, Tourist around the globe love to travel Italy. Italy is one of the best, lovable and enjoyable destination. As per the report Italy experience highest amount of tourist every year, if we compare with other European countries. Invest on food industry in Italy would be beneficial as people love Italian cousin.

Opening Italian restaurant allows investor to earn consistent rental income. Of course, as with any business decision, opening a new restaurant requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the restaurant is one of the most important decisions that will determine whether the restaurant will be a success or a failure.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city Rome ,Italy to open a new authentic Italian restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Rome, Italy, if a investor or individual is looking to open a new restaurant, where would you recommend that they open it?

Data

To solve the problem, we will need the following data:

- List of neighbourhoods in Rome. This defines the scope of this project which is confined to the city of Rome, the capital city of the country of Italy.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to restaurant. We will use this data to perform clustering on the neighbourhoods.

- **Sources of data and methods to extract them**

  Rome Neighborhoods Guide contains a list of neighbourhoods in Rome, with a total of 31 neighbourhoods. We will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

  After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the restaurant category 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 working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Rome. Fortunately, the list is available Rome Neighborhoods Guide. We collect the list, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Rome.

Next, we will use Foursquare API to get the top 500 venues that are within a radius of 5000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Italian restaurant” data, we will filter the “Italian restaurant” as venue category for the neighbourhoods.

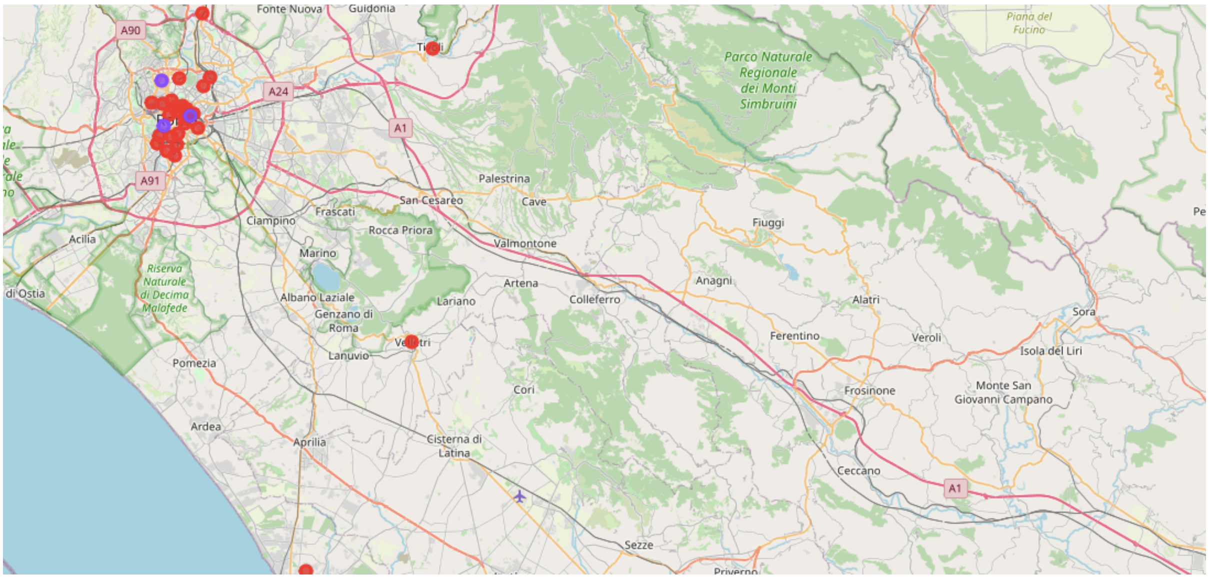
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 2 clusters based on their frequency of occurrence for “Italian restaurant”. The results will allow us to identify which neighbourhoods have higher concentration of restaurant while which neighbourhoods have fewer number of Italian restaurant. Based on the occurrence of Italian restaurant in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Italian restaurant.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 2 clusters based on the frequency of occurrence for "Italian restaurant":

- Cluster 0: Neighbourhoods with moderate number of Italian restaurant
- Cluster 1: Neighbourhoods with high no. of Italian restaurant

  The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour.



Discussion

As observations noted from the map in the Results section, most of the restaurant are concentrated in the central area of Rome city, with the highest number in cluster 1 and moderate number in cluster 0. This represents a great opportunity and high potential areas to open new restaurant as there is very little to no competition from existing restaurant. Meanwhile, restaurant in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of restaurant.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of restaurant, there are other factors such as population and income of residents that could influence the location decision of a restaurant . However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new restaurant. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone 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 2 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 0 are the most preferred locations to open a new restaurant. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new restaurant.