REPORT

Introduction

This is my capstone project for IBM Data Science Professional Certificate. In this project we tried to find the best location to open an Italian restaurant in Toronto.

The most suitable location not only depends on how many other Italian restaurants are in the area. We should also consider expenses like rents and potential customers who live in the area.

DATA AND EXTRACTING THE DATA

To solve this problem, we will use the data, we already looked at in the previous weeks, namely:

- List of neighborhoods in Toronto, Canada scrapped from Wikipedia.
- Latitude and Longitude of these neighborhoods from the Geocoder package for python
- Venue data related to restaurants. This data we will receive with the Foursquare API. This will help us find the neighborhoods that are more suitable to open an Italian Restaurant.

METHODOLOGY

Everything has been carried out in Python. The first thing we do is to get the list of neighborhoods of Toronto. As seen in the previous weeks we can scrape them from Wikipedia:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

This we could get by using requests and BeautifulSoup to scrape the webpage and at the end we stored the data in a pandas dataframe.

To get the respective coordinates for each neighborhood we used the webpage:
http://cocl.us/Geospatial_data

With the same methodology as before we got from there the postal codes, latitudes, longitudes and stored them in a pandas dataframe.

To find the neighborhoods and nearby venues we used the functional API from Foursquare ID. For this we created a Foursquare developer account in order to obtain an account ID and an API key to pull the data (for the notebook to work you will have to put in our credentials).

Once we had the date, we analyzed each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. In this matter we prepared the data and normalized it so we could use it for our machine learning algorithms. Moreover we only included Italian restaurants. As we wanted to find out were most of them are located.

Then we used a clustering method in particular the k-means clustering, which we learned during the previous weeks.

We choose to have 3 clusters for the unsupervised machine learning algorithm. When we get the clusters we can identify where a suitable location for a Italian restaurant could be.

RESULTS

	Venue Category
Cluster Labels	
0	16
1	1
2	12
3	11

Cluster 1 has the fewest restaurants with only one restaurant. This could be a possible candidate. Whereas Cluster 0 has the most. But the neighborhoods are centered around the city center, where of course the city is most densely populated and the demand is highest.
Cluster 2 & 3 are similar in size but they are located on the outskirts of the city especially cluster 2.

DISCUSSION

For further investigation one may need more data, hence, results could differ significantly.

RESULTS

In conclusion this project recommends cluster 3 because one should also consider that opening restaurants in the city center may lead to high expenses like rents. Whereas in the suburbs there may not live enough potential customers.