Course IBM Data Science Capstone Project: Opening a new Peruvian Restaurant In Toronto

Report Prepared by: Yonatan Tacza
(www.github.com/itacza2020)

Introduction:

For this Capstone project, I am creating a hypothetical scenario for a concept Peruvian Restaurateur who wants to explore opening an authentic Peruvian restaurant in Toronto area. The idea behind this project is that there may not be enough Peruvian restaurants in Toronto and it might present a great opportunity for this entrepreneur who is based in Canada. As Peruvian food is very similar to other Latin American cuisines, this entrepreneur is thinking of opening this restaurant in locations where Latin American food is popular (many Asian restaurants in the neighborhood). With the purpose in mind, finding the location to open such a restaurant is one of the most important decisions for this entrepreneur and I am designing this project to help him find the most suitable location.

Business Problem :

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Peruvian restaurant in Toronto, Canada. By using data science methods and machine learning methods such as clustering, this project aims to provide solutions to answer the business question: In Toronto, if an entrepreneur wants to open a Peruvian restaurant, where should they consider opening it?

Target Audience:

The entrepreneur who wants to find the location to open authentic Peruvian restaurant

Data

To solve this problem, I will need below data:

- List of neighborhoods in Toronto, Canada.
- Latitude and Longitude of these neighborhoods.
- Venue data related to Latin America restaurants. This will help us find the neighborhoods that are most suitable to open a Peruvian restaurant.

Extracting the Data

- Scrapping of Toronto neighborhoods via Wikipedia
- Getting Latitude and Longitude data of these neighborhoods via Geocoder package
- Using Foursquare API to get venue data related to these neighborhoods

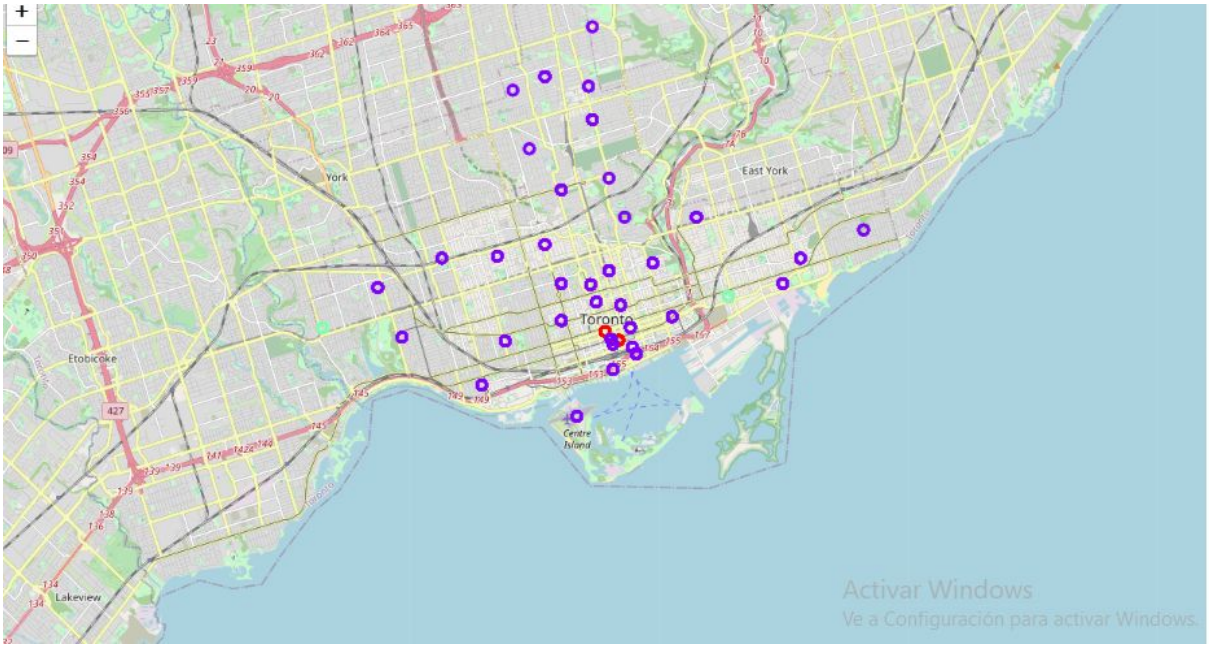
Methodology :

First, I need to get the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from wikipedia page("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M ") I did the web scraping by utilizing pandas html table scraping method as it is easier and more convenient to pull tabular data directly from a web page into dataframe. However, it is only a list of neighborhood names and postal codes. I will need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I tried using Geocoder package but it was not working so I used the csv file provided by IBM team to match the coordinates of Toronto neighborhoods. After gathering all these coordinates, I visualized the map of Toronto using Folium package to verify whether these are correct coordinates. Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later. Here, I made a justification to specifically look for "Peruvian Restaurant". Previously, when I ran the model, I was looking for "Latin American restaurants" but there are very few results (maybe due to Foursquare categorization) so I looked for the restaurants closest to Peruvian cuisine taste

(side note: Peruvian food food and Latin American food are very similar in taste, so my justification is that if there are people who enjoyed Latin American food, they likely are going to enjoy Peruvian food too!)
Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centeriods, 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 it is highly suited for this project as well. I have clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for “Latin food”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

Results :

Clusters



The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Latin american restaurants are in each neighborhood:

- Cluster 0: Neighborhoods with little or no Latin American restaurants
- Cluster 1: Neighborhoods with high number of Thai restaurants

● Cluster 2: Neighborhoods with no Thai restaurants

The results are visualized in the above map with Cluster 0 in red color, Cluster 1 in purple color and Cluster 2 in light green color.

Recommendations :

Most of Latin restaurants are in Cluster 1 which is around Adelaide, King, Richmond areas and lowest (close to zero) in Cluster 0 areas which are North Toronto West and Parkdale areas. Also, there are good opportunities to open near Studio Distric , Runnymeda and Swansea as the competition seems to be low. Looking at nearby venues, it seems Cluster 2 might be a good location as there are not a lot of Latin Amercican restaurants in these areas. Therefore, this project recommends the entrepreneur to open an authentic Peruvian restaurant in these locations with little to no competition. Nonetheless, if the food is authentic, affordable and good taste, I am confident that it will have great following everywhere :)

Limitations and Suggestions for Future Research :

In this project, I only take into consideration of one factor: the occurrence / existence of Latin American restaurants in each neighborhood. There are many factors that can be taken into consideration such as population density, income of residents, rent that could influence the decision to open a new restaurant. However, to put all these data into this project is not possible to do within a short time frame for this capstone project. Future research can take into consideration of these factors. In addition, I am relying on the existence of Thai restaurants only for this project but future research can take into consideration of other variables such as existence of Latin American restaurants, Latin population level in each neighborhood etc.

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 the machine learning by utilizing k-means clustering and providing recommendation to the stakeholder.

References :

List of neighborhoods in Toronto:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
Foursquare Developer Documentation:
https://developer.foursquare.com/docs