

IBM Data Science Professional Certificate

Applied Data Science Capstone

A CARIBBEAN RESTAURANT IN MONTREAL

By Samson JULES; July 4th, 2020

1. INTRODUCTION

1.1. Business Problem

Montreal is one of the major cities in Canada and Quebec in particular. Its population is the second largest French-speaking population in the world after France, composed largely of immigrants. One of the most immigrant-friendly cities in North America with more than half of the entire Haitian Canadian population residing in Montreal it is one of the best places to start a Caribbean restaurant.

In this project we will go through step by step process to make a decision whether it is a good idea to open a Caribbean restaurant. We analyze the neighborhoods in Montreal to identify the most profitable area since the success of the restaurant depends on the people, environment, ambience and other. Since we already know that Montreal shelter a greater number of Haitian than any other city in Canada, it is a good idea to start the restaurant here, but we just need to make sure whether it is a profitable idea or not. If so, where we can place it, so it yields more profit to the owner.

1.2. Interest or Target audience

Who will be more interested in this project? What type of clients or a group of people would be benefitted?

1. Business personnel who wants to invest or open a Caribbean restaurant in Montreal. This analysis will be a comprehensive guide to start or expand restaurants targeting the Caribbean crowd.
2. Freelancer who loves to have their own restaurant as a side business. This analysis will give an idea, how beneficial it is to open a restaurant and what are the pros and cons of this business.
3. Haitian or Caribbean crowd who wants to find neighborhoods with lots of option for Caribbean restaurants.
4. Business Analyst or Data Scientists, who wish to analyze the neighborhoods of Montreal using Exploratory Data Analysis and other statistical & machine learning techniques to obtain all the necessary data, perform some operations on it and, finally be able to tell a story out of it.

2. DATA ACQUISITION AND CLEANING

2.1. Data acquisition

A. We use the "Open Data" on <http://donnees.ville.montreal.qc.ca/dataset/quartiers> for data on the reference boroughs, neighbourhoods and cities of Montreal. For the geospatial coordinates of the neighbourhoods, we use the geolocator(geocoder). This data includes geospatial information (latitude and longitude). Neighbourhoods are reference neighbourhoods in housing.

B. The data on Montreal's immigrant population comes from the 2016 Census through the document entitled in French: "ANNUAIRE STATISTIQUE DE L'AGGLOMERATION DE MONTREAL / MONTREAL AGGLOMERATION STATISTICAL YEARBOOK". The document is available here: http://ville.montreal.qc.ca/portal/page?_pageid=6897,68087755&_dad=portal&_schema=PORTAL

C. To get location and other information about various venues in Montreal I'm using Foursquare's API. Using the Foursquare's explore API (which gives venues recommendations), I'm fetching details about the venues up present in Montreal and collected their names, categories and locations (latitude and longitude).

From Foursquare API (<https://developer.foursquare.com/docs>), I retrieved the following for each venue:

- Name: The name of the venue.
- Category: The category type as defined by the API.
- Latitude: The latitude value of the venue.
- Longitude: The longitude value of the venue.

2.2. Data Cleaning

After importing some libraries such as numpy, pandas, geocoder matplotlib and others, we proceeded to read the first csv file containing information on the boroughs of reference of Montreal. We must specify that the databases of the different sites of Montreal are in French and the names of these neighborhoods or boroughs are also in French. We have only changed the titles of the columns in English. The number of reference districts is 90.

The latitude and longitude of the districts are obtained from Geocoder (geolocator). A CSV file has been created for this purpose. Once finished with these preparations, we concat these dataframes.

As for the demography database in Montreal, as for the others, we have modified, adapted or translated the columns to finally obtain the distribution of immigrants of Haitian origin. It should also be noted that the data available in open data contains only the data for the districts and not for the neighbourhoods as we wished. We should make do with that.

Finished this first round, it's time to find the categories of the different venues from Montreal. To do so, we first proceeded to the geospatial location of Montreal, then from our Foursquare identifier locate in an area with a radius of 1000 meter, we get the JSON file of the categories and Nearby Venues from around 1174. This work resulted in 211 unique categories in which Caribbean Restaurant is one of them. Then follows the One hot encode for getting dummies of venue category. So that we calculate mean of all venue groupby their neighborhoods.

3. EXPLORATORY DATA ANALYSIS

3.1. Folium Library and Leaflet Map

As we said Folium is a python library. We used it to draw an interactive leaflet map using coordinate data.

3.2. Relationship between neighborhood and Haitian Restaurant

In this section, we want to establish the relationship between Montreal's neighbourhoods and the Caribbean restaurants that exist there. The study of this relationship consists first of taking the columns of the Caribbean neighborhoods and restaurants established above. By merging the new dataframe with the one containing the geospatial coordinates and by using tools such as matplotlib, seaborn and others, we observe that :

- Only Côte-des-Neiges-Notre-Dame-de-Grâce and Le Sud-Ouest have one or two Caribbean restaurants.

3.3. Relationship between neighborhood and Haitian population

The purpose of this approach is to establish the relationship between the Haitian population in Montreal and the neighbourhoods. Montreal-North, Rivière-des-Prairies-Pointe-aux-Trembles and Villeray-Saint-Michel-Parc-Extension are the top 3 boroughs with the most Haitians in their neighbourhoods.

3.4. Relationship between Haitian population and Caribbean restaurant

With only 2 restaurants in 2 boroughs, we cannot establish a meaningful relationship between the Haitian population and the number of Caribbean restaurants in Montreal. However, we can look in which borough where the population is much denser to place a Caribbean restaurant. This will be a certain guarantee for the good start and performance of the restaurant.

After performing the data cleaning & data analysis we can identify that there is no big relationship established in terms of the Caribbean population & the popular Caribbean restaurants.

Thus, this marks the end of the data cleaning & analyses step in this project. Next, we will look into the predictive modeling. In the predictive modelling we are going to use Clustering techniques since this is an analysis of unlabeled data. K-Means clustering is used to perform the analysis of the data at hand.

4. PREDICTIVE MODELING

4.1. Clustering Neighborhoods of Montreal

We proceed in this section to the grouping of the following Montreal neighborhoods that have more or less Caribbean restaurants. First step in K-means clustering is to identify the best K value meaning the number of clusters in a given dataset. We use in this case the dataframe "*montreal_res*" containing the Montreal dataset with Caribbean restaurant percentage. The Elbow method will be used to do so.

After analysing using elbow method using distortion score & Squared error for each K value, looks like K = 4 is the best value. The next step is the resumption on a map this clustering with 3 colors: yellow, blue and purple.

We have 4 clusters such as 0,1,2,3:

- Cluster 0 has no rows meaning no data points or neighborhood was near to this centroid.
- Cluster 1 contains the neighborhoods which are more populated with Caribbean restaurants. It is shown in purple in the map.
- Cluster 2 contains the neighborhoods which are sparsely populated with Caribbean restaurants. It is shown in blue in the map.
- Cluster 3 contains all the neighborhoods which have the least number of Caribbean restaurants. It is shown in yellow color in the map.

5. RESULTS AND DISCUSSION

5.1. Results

This section presents the results of our work. Our approach consists in analyzing the best possibility of implementing a Caribbean restaurant in Montreal. To do so, we have documented geospatial data of the reference (residential) neighbourhoods of Montreal and their population size, to finally arrive at the neighbourhood that would be the most appropriate to open a Caribbean restaurant. By using Statistics Montreal data for the geospatial coordinates of the neighbourhoods and their respective populations, as well as the Foursquare API tool for the analysis of our data, we found the following results:

- Among the the 33 boroughs or cities linked to Montreal, only Côte-des-Neiges–Notre-Dame-de-Grâce and Le Sud-Ouest have one or two Caribbean restaurants.
- Boroughs such as Montréal-Nord, Rivière-des-Prairies-Pointe-aux-Trembles, Villeray-Saint-Michel-Parc-Extension, Ahuntsic-Cartierville and Saint-Léonard are the most populous in terms of the number of Haitian immigrants.
- With the help of clusters examing & violin plots looks like Côte-des-Neiges–Notre-Dame-de-Grâce and Le Sud-Ouest are already densely populated with Caribbean restaurants. So it is better idea to leave those boroughs out and consider only Rivière-des-Prairies–Pointe-aux-Trembles, Montréal-Nord, Villeray–Saint-Michel–Parc-Extension, Ahuntsic-Cartierville and Saint-Léonard for the new restaurant's location.
- After careful consideration it is a good idea to open a new Caribbean restaurant in **Montreal-Nord** borough since it has high number of Caribbean population which gives a higher number of customers possibility and lower competition since very less Caribbean restaurants in the neighborhoods.

5.2. Discussion

The results obtained allow us to consider the Montreal-Nord borough as being the least competitive with the largest community of Haitian immigrants and among the boroughs with the fewest Caribbean restaurants. In addition, the distribution of the population gives the impression of a high density of Caribbean population, particularly Haitian, which favours the new restaurant by offering great opportunities for customers to visit. It is therefore certain that this region could be a perfect place to open a quality Caribbean or Haitian restaurant.

However, our work meets certain limits. First of all, with the clustering which is based solely on data obtained by Foursquare API. In addition, the distribution of the Haitian immigrant population in each neighbourhood or district is also based on the 2016 census, which is not up to date. Thus, the population distribution would certainly have changed by 2020, given the four-year lag in the data. Given that the distribution of the Haitian population in each neighbourhood and the number of Haitian restaurants are the main elements of this analysis and that the data is not entirely up to date, this analysis is not far from conclusive and can be improved in many areas. However, it certainly provides us with good insights, preliminary information on possibilities and a head start on this business problem by setting the right milestones. In addition, it can also potentially vary depending on the type of aggregation techniques we use to examine the data.

6. CONCLUSION

This project gave us the opportunity to work on a project with real data like a real data scientist. Such work required the use of different tools such as python libraries for data retrieval, manipulation, visualization and analysis. Matplotlib and seaborn were also very useful for visualization as well. For the prediction of the result of the data, we used a machine learning technique, reinforced by Folium for visualization on a map. Although there are some limitations inherent in our work that lead us to believe that our analysis can be further improved, we still believe that they have not undermined the quality of our work and that it will prove to be very useful. Thus, this project can be used for different purposes or scenarios such as the opening of a football club for example. Furthermore, this project is only the introduction of a young Data scientist into the immense universe of Data science, we hope that it can serve as a guide for dealing with other realities involving Data science.