Capstone Project – IBM Professional Certificate in Data Science – Coursera

Montréal's Neighborhoods Analysis – Where to open a wellness facility?

By

Antoine Morin

January 29th, 2021

ABSTRACT

In this report, we analyze the venues of each neighborhood in the city of Montréal using the FourSquare API in order to determine which neighborhood is better suited to invest in the development of a new wellness facility. After modeling the data using multiple linear regression on the most significant venues categories, it is found that this neighborhood is Ahuntsic Sud-Est, where only a single fitness center currently exists. These findings could be the starting point for a lucrative business opportunity.

INTRODUCTION

Montréal is the second-most populous city in Canada with a population of almost 2 million inhabitants. It goes without saying that the city offers a lot of opportunities for whoever wants to invest in some sort of business, either small or big. Montréal is home to all kind of venues, from restaurants to cafés, museums, artisanal shops, etc. There is now a clear tendency for people to take care of their health, and Montréal has seen in the past years a steady growth in the number of wellness facilities, for example gyms and yoga studios. As this trend is likely to continue to grow, we are interested in taking a share in this business by opening a wellness facility in Montréal. To have better chances of success, we want to know where the best place is to open this facility. More precisely, in which neighborhood should we invest in our business? To answer this question, we will use the power of data along with machine learning to get crucial insights about the business ecosystem of the city of Montréal.

DATA REQUIEREMENTS

To solve our problem, we will first need to get data about the neighborhoods of Montréal. To identify the neighborhoods, we will use the postal codes of the city, available at https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_H. We will also need the latitude and longitude of each neighborhood to later get the venues with the FourSquare API. Because Nominatim cannot provide us with these latitudes and longitudes using only postal codes and because the Google Maps API needs a paid subscription, we will get manually these latitudes and longitudes using Google Maps. Finally, we need to get the venues in each neighborhood, based on the latitudes and longitudes and a specified radius. To accomplish this step, we will use the FourSquare API. The best radius can be determined with the latitudes and longitudes of each neighborhood by calculating the mean distance between each closest neighborhood. See the Methodology section for more information.

Regarding the venues that we will get through the FourSquare API, we are mainly interested in the number of venues in each category. We will need to define broad categories in order to extract meaningful relationships between the venues. In particular, we want to see correlations between the “Wellness” venues category and other categories. The broad categories that we will use are: restaurants, arts, wellness, fast food, artisanal food, bars, cafés, public area, and essentials. For example, gyms will be in wellness, while Chinese restaurants will be in restaurants, etc…For more specificity, see the Methodology section.

METHODOLOGY

To get the postal codes and the names of the neighborhoods in Montréal, we scraped the associated Wikipedia page given in the Data Requirements section and we used the library BeautifulSoup to parse the HTML code. Some of the postal codes were in fact situated in Laval, an adjacent city to Montréal. The postal codes of Laval all contain the number 7 in them, and this fact was used to get rid of them. Some remaining postal codes, such as H0H 0H0 reserved to Santa Claus, were manually removed.

As was mentioned before, to get the latitude and longitude of each neighborhood, Nominatim doesn't work with only postal codes, and the Google Maps API is a paid service. We had to manually get the latitudes and longitudes using the website Google Maps.

To find venues using the FourSquare API, we need to specify a certain radius to look for venues around a specify geographic point described by a latitude and a longitude. To avoid too much overlap between the neighborhoods but to cover at the same time most of the city, we need to choose the radius carefully. One way to do so is to calculate the distance between each neighborhood using the latitudes and longitudes, and then find, for each neighborhood, the closest one. Looking at the distribution of the distances, we will be able to find an almost optimal radius.

We will use this radius to get at most 100 venues for each neighborhood using the FourSquare API. Looking at the categories of all the venues, we will classify them in 9 broad categories. The following associations are used:

Restaurants = ["Afghan Restaurant","American Restaurant","Arepa Restaurant","Argentinian Restaurant","Asian Restaurant","Australian Restaurant","Cajun / Creole Restaurant","Cambodian Restaurant","Caribbean Restaurant","Chinese Restaurant","Comfort Food Restaurant","Dumpling Restaurant","Eastern European Restaurant","Empanada Restaurant","English Restaurant","Falafel Restaurant","Filipino Restaurant","French Restaurant","German Restaurant","Greek Restaurant","Hawaiian Restaurant","Indian Restaurant","Indonesian Restaurant","Italian Restaurant","Japanese Restaurant","Jewish Restaurant","Korean Restaurant","Latin American Restaurant","Lebanese Restaurant","Mediterranean Restaurant","Mexican Restaurant","Middle Eastern Restaurant","Modern European Restaurant","Moroccan Restaurant","New American Restaurant","North Indian Restaurant","Persian Restaurant","Peruvian Restaurant","Polish Restaurant","Portuguese Restaurant","Ramen Restaurant","Restaurant","Russian Restaurant","Salvadoran Restaurant","Seafood Restaurant","South American Restaurant","Spanish Restaurant","Sri Lanka Restaurant","Steakhouse","Sushi Restaurant","Swiss Restaurant","Szechuan Restaurant","Tapas Restaurant","Tex-Mex Restaurant","Thai Restaurant","Tibetan Restaurant","Turkish Restaurant","Vegetarian / Vegan Restaurant","Vietnamese Restaurant"]

Arts = ["Art Gallery","Art Museum","Arts & Crafts Store","Arts & Entertainment","Bookstore","Botanical Garden","College Bookstore","Comedy Club","Concert Hall","Garden","Garden Center","General Entertainment","Historic Site","History Museum","Indie Movie Theater","Jazz Club","Movie Theater","Museum","Music Store","Music Venue","Opera House","Performing Arts Venue","Planetarium","Public Art","Recording Studio","Theater","Video Store"]

Wellness = ["Athletics & Sports","Boxing Gym","Climbing Gym","College Gym","Dance Studio","Gym","Gym / Fitness Center","Gym Pool","Martial Arts School","Massage Studio","Spa","Sports Club","Yoga Studio"]

Fast Food = ["BBQ Joint","Burger Joint","Burrito Place","Fast Food Restaurant","Fish & Chips Shop","Food Truck","Fried Chicken Joint","Hot Dog Joint","Mac & Cheese Joint","Pizza Place","Poutine Place","Sandwich Place","Taco Place"]

Artisanal Food = ["Bagel Shop","Bakery","Breakfast Spot","Chocolate Shop","Creperie","Cupcake Shop","Dessert Shop","Donut Shop","Gourmet Shop","Ice Cream Shop","Juice Bar","Pastry Shop","Pie Shop","Sausage Shop"]

Bars = ["Bar","Beer Bar","Bistro","Brewery","Cocktail Bar","Dive Bar","Gastropub","Gay Bar","Hookah Bar","Irish Pub","Karaoke Bar","Lounge","Nightclub","Pub","Smoke Shop","Speakeasy","Sports Bar","Whisky Bar","Wine Bar"]

Cafe = ["Café","Coffee Shop","Gaming Cafe"]

Public Area = ["Baseball Field","Beach","Canal","Cemetery","Church","Dog Run","Event Space","Football Stadium","Hockey Arena","Lake","Monument / Landmark","Mountain","Nature Preserve","Park","Playground","Plaza","Pool","Scenic Lookout","Soccer Field","Soccer Stadium","Stadium","Tennis Court","Tennis Stadium","Trail","Zoo"]

Essentials = ["Auto Garage","Auto Shop","Bank","Clothing Store","Convenience Store","Deli / Bodega","Drugstore","Electronics Store","Farmers Market","Fish Market","Flea Market","Food & Drink Shop","Fruit & Vegetable Store","Furniture / Home Store","Gas Station","Grocery Store","Hardware Store","Liquor Store","Market","Paper / Office Supplies Store","Pharmacy","Shopping Mall","Shopping Plaza","Supermarket"]

We will use these 9 categories and the given associations to count the number of venues that fits in these broad categories for each neighborhood. The goal here is to be able to determine if there are any correlations in the number of venues between the wellness category and the other categories. Because some neighborhoods have a low count of total venues, we will use a certain criterion of the total number of venues to discard the neighborhoods that are too small, because in their case, it won't be possible to get significant statistical relationships. To compare the remaining neighborhoods, we will normalize each category of a given neighborhood by its total numbers of venues.

This will allow us to calculate the correlation matrix between each category. It will indicate if there exist any linear relationship between the categories that we can exploit. For the categories showing the biggest Pearson correlation coefficients with regards to the wellness category, we will calculate the p-value to ensure statistical significance.

Finally, all the previous steps will allow us to focus on the categories that have the highest correlation with the wellness category. We will use these categories in a machine learning model, specifically a multiple linear regression in order to predict the number of wellness facilities that we should find in a certain neighborhood given the number of venues from the other significant categories. The goal is to be able to find a neighborhood which actually have a significantly lower number of wellness facilities than should be expected. This is a strong indication that there exists a business opportunity there.

RESULTS

From the scraping of the Wikipedia page, we got 124 postal codes. From these, 19 were postal codes in Laval, leaving 105 postal codes in Montréal. Four postal codes were manually removed, leaving a total of 101 postal codes and neighborhoods in Montréal. They are shown in Fig. 1.

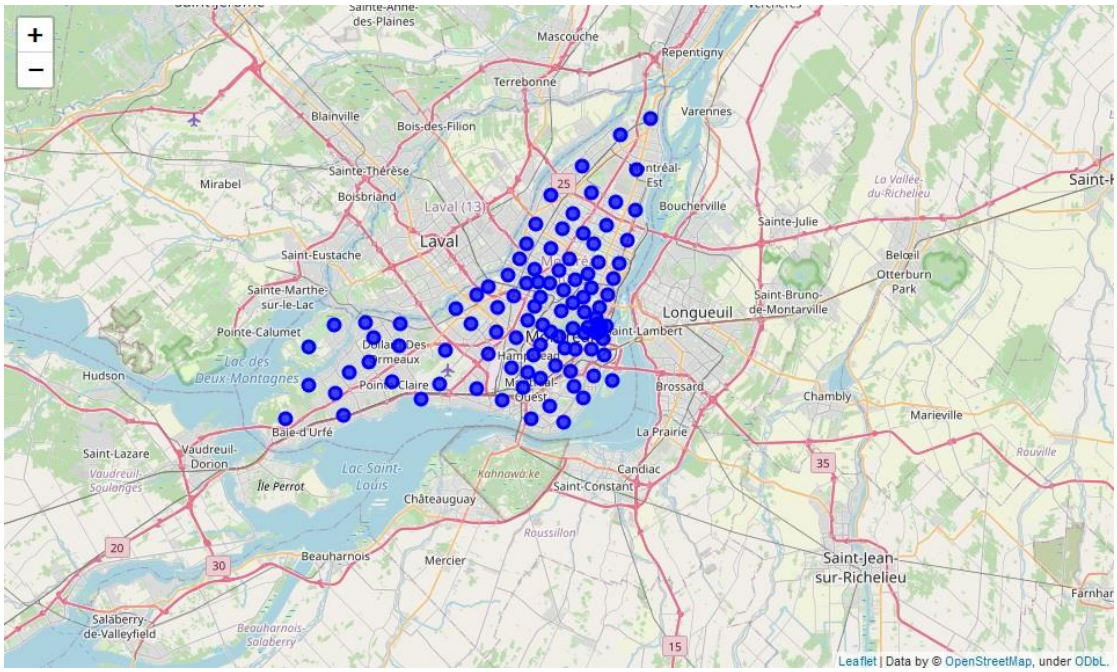

Figure 1: All the neighborhoods of Montréal.

With the latitude and longitude of each neighborhood, we determined that the minimum distance between two neighborhoods is 349 meters, the maximum distance is 4427 meters, and the mean distance between nearest neighborhoods is 1940 meters. Fig. 2 shows the distribution of these distances.

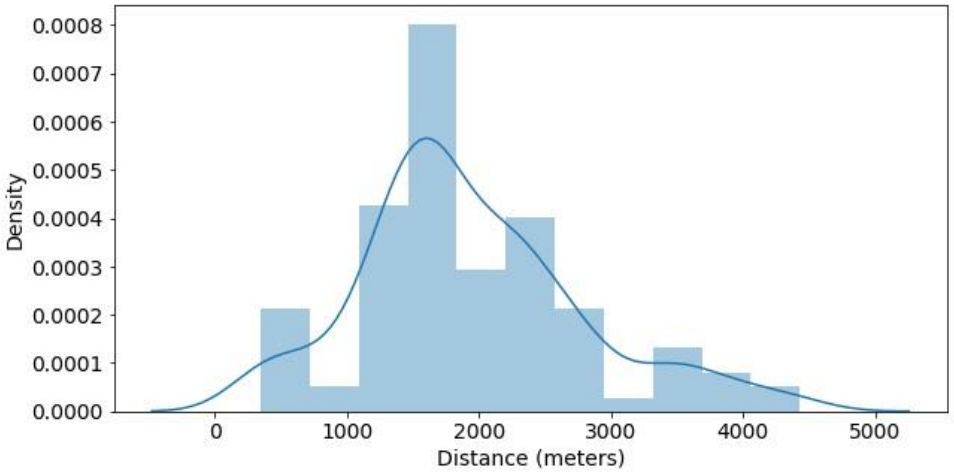

Figure 2: Distribution of distances between nearest neighborhoods.

From this, we determined that a radius of 1000 meters is a good parameter for the FourSquare API to cover most of the city while limiting the overlap between the neighborhoods.

Using the FourSquare API to get venues in each neighborhood with a limit of 100, we got a mean number of venues returned of 41. This means that some neighborhoods have a very small number of venues and need to be discarded. We chose a total number of venues of 50 as the discriminant, which decreased the number of neighborhoods from 101 to 31.

Looking at all the venues categories returned by the FourSquare API, we were able to group the venues in 9 broad categories, described in the Methodology section. By grouping all the venues in these 9 broad categories, we found that 27.1% of them were restaurants, 7.1% were arts related, 4.7% were wellness facilities, 9.7% were fast foods, 10.0% were artisanal food shops, 9.0% were bars, 12.7% were cafés, 6.4% were public area, and 13.4% were essentials shops. The neighborhood with the smallest ratio of wellness facilities is Outremont with 1.1% and the neighborhood with the biggest ratio of wellness facilities is Maisonneuve with 11.32%.

The next step is to look at the correlations between the categories. Fig.3 shows a heatmap of the Pearson correlation coefficients.

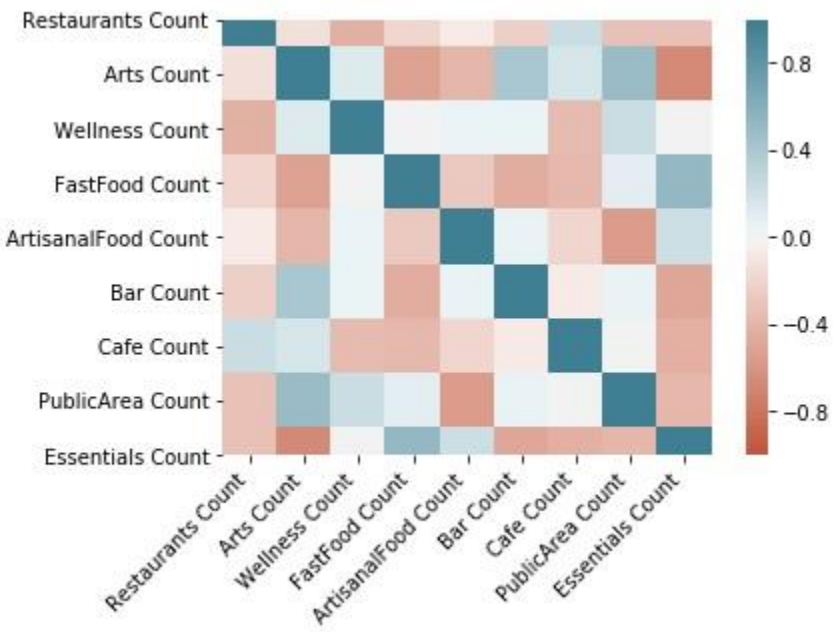

Figure 3: Pearson Correlation Coefficients between the broad categories.

From this figure, we can see that the wellness category doesn't seem to be correlated with the other categories that much. However, we can still see that the highest correlations concern the following categories: restaurants, arts, cafés, and public areas.

More concretely, the Pearson Correlation Coefficient between Restaurants and Wellness is -0.43 with a P-value of 0.017. The Pearson Correlation Coefficient between Arts and Wellness is 0.13 with a P-value of 0.502. The Pearson Correlation Coefficient between Cafés and Wellness is -0.36 with a P-value of 0.044. The Pearson Correlation Coefficient between Public Areas and Wellness is 0.23 with a P-value of 0.210.

From the above p-values, it seems that the only significant correlations involve Restaurants and Cafés. The p-values are respectively 0.017 and 0.044 (< 0.05), indicating a moderate evidence that the correlation is significant. The Pearson Correlation Coefficients are respectively -0.43 and -0.36, indicating a negative

linear relationship between the number of restaurants and cafés and the number of wellness centers. In other words, we would expect more wellness centers in neighborhoods having less restaurants and cafés.

We proceeded to do a multiple linear regression with the categories restaurants and cafés as independent variables and the category wellness as the dependent variable. Denoting *y* as the actual number of wellness facilities and *yhat* as the predicted number of wellness facilities by the model, we then looked directly at the error *y-yhat* for each neighborhood. This is important because this contrast we the usual error measures involving the absolute error or the mean squared error. Here, we we want to know where the error is negative, indicating that there should be more wellness centers (*yhat*) than observed (*y*). The largest negative error was found for the neighborhood called Ahuntsic Sud-Est, with an error value of -0.045. Fig. 4 shows where this neighborhood is situated.

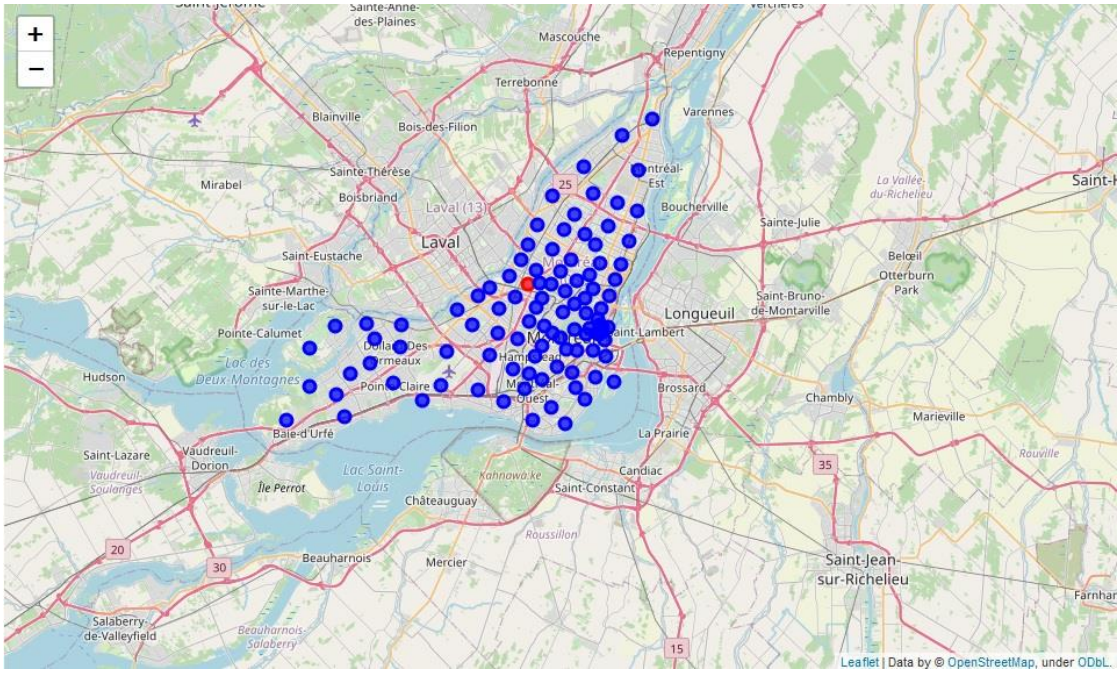

Figure 4: The neighborhood Ahuntsic Sud-Est (red point) would be the best option to open a wellness facility.

DISCUSSION

The previous analysis has identified the neighborhood called Ahuntsic Sud-Est as the best option to open a wellness facility. The first thing to do now is to look more closely at this particular neighborhood. Table 1 shows the count of each specific wellness facility in Ahuntsic Sud-Est.

Table 1: Count of each specific wellness facility in Ahuntsic Sud-Est.

	Wellness Facility	Count
0	Athletics & Sports	0
1	Boxing Gym	0
2	Climbing Gym	0
3	College Gym	0
4	Dance Studio	0
5	Gym	0
6	Gym / Fitness Center	1
7	Gym Pool	0
8	Martial Arts School	0
9	Massage Studio	0
10	Spa	0
11	Sports Club	0
12	Yoga Studio	0

We can see that there only exists at the moment a gym or fitness center in Ahuntsic Sud-Est. Our work indicates that there should be room for more wellness facilities, but it doesn't specify which one. Is it better to open another gym? Or should we try to complement the services of the already existing gym a, for example, a yoga studio? Or course, these questions would be the starting point of an entirely new project and they are outside the scope of the present project. For such a project, it could be interesting to go directly in the neighborhood and look at different variables. For example, where is the actual gym? Where is the majority of the venues? What services are offered at the gym? What could be possible locations for the new wellness facility? Also, it could be very useful to survey the population living in Ahuntsic Sud-Est to get their opinions about the neighborhoods, its services, and the possibility of a new wellness facility.

The analysis conducted in this small project could obviously be improved on many points. The most obvious one is the venues returned by the FourSquare API. How much overlapping exists in the data? It might be better to change the radius parameter for each call to the API, each call corresponding to a different neighborhood. Also, how are the venues returned by the API selected? Is there a bias in the selection process when a limit is imposed on the number of venues returned? Are wellness facilities underrepresented compared to restaurants? These are questions that would need to be addressed in the future.

Also, it would be helpful to conduct an analysis on the impact of our choice of broad categories. Should we make more smaller broad categories, or less bigger categories? And what about our choice of

discriminant for the number of total venues, which helped us discard some neighborhoods? Should we set it higher or lower? In which way would the results differ with these changes?

Another interesting path to follow is to look not at a single neighborhood with a large negative error, but at many and comparing them. For example, according to our model, the next neighborhood with a large negative error is Verdun Nord. How does this neighborhood compare to Ahuntsic Sud-Est? Which wellness facilities already exist there? Maybe this neighborhood would be a better choice for various reasons.

With so many questions remaining to be answered, I think it shows the potential of the present analysis and its importance. It could really lead to some interesting discoveries. As we already mentioned, this is a good starting point.

CONCLUSION

This project aimed to answer a simple direct question: Where is the best place in Montréal to open a new wellness facility, for example a gym or a yoga center? To achieve the goal of this project, we first had to describe the data that would be required to answer the question. This data comprises the different neighborhoods in Montréal, to address the “where” in the question, and comprises the different venues that we can find in these neighborhoods, to address the “wellness facility” part of the question. The data related to the neighborhoods was scraped from a Wikipedia page, while the data related to the venues was retrieved using the FourSquare API. After preparing the data to classify all the venues in broad categories, one of them being “Wellness”, we were able to calculate statistical relationships amongst the data, specifically the Pearson Correlation Coefficients and the P-Values. This indicated that the wellness category is moderately correlated to the restaurants category and the cafés category. We used a multiple linear regression model to predict the number of wellness facilities that should be found in each neighborhoods of interest. By quantifying the errors of the model, we were able to determine that the neighborhood Ahuntsic Sud-Est is the one where the actual number of wellness facilities is farthest from the predicted number by the model, indicating a strong business potential for this neighborhood. Many questions are yet to be answered, but this work is an excellent starting point for anyone looking to take part in the huge business ecosystem of the beautiful city of Montréal.