**Determining the best**

**place for my**

**Own Venture!**

**Author**: Joaquín Santini

**Date**: 06-12-2019

**Conceptual Framework**: Applied Data Science Capstone

**Objective**: analyze enough information to choose the best place to open my business

1. **Introduction**
2. **Background**

Today, we live in an era where globalization allows us to inform ourselves in very different ways than we knew just 15 or 20 years ago. Consequently, today it’s easier to learn and understand how to start a business autonomously and not fail in the attempt.

Given that one of the first questions we ask ourselves is "Where do I open my business?" The intention of this document is to help answer that question by analyzing DATA that, for our benefit, is publicly accessible to all!

In my case, I want to open a Coffee Shop.

1. **Problem**

Taking advantage of the DATA that we have available on the web, the idea is to help us make the best decision to the answer described above.

In this sense, the best decision is to avoid competing with other businesses of the same category to be able to capture more "New Customers". The location analyzed in this case will be Toronto, Canada.

1. **Interest Audience**

Just like me, there are many people who think about the idea of their own entrepreneurship and I understand that the problem I pose arises for everyone.

Therefore, I invite any interested person to use the proposed analysis and succeed also in their own idea of entrepreneurship! The business item does not matter as it can be applied to anyone!

1. **Data Section: acquisition and preparation**
2. **Data Sources**

The first DATA SOURCE is [this Wikipedia article](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) where I can find the information about all the neighborhoods of Toronto, Canada.

Our second DATA SOURCE is Foursquare. For those who don't know what it is, I briefly comment that Foursquare is "a location technology platform dedicated to improving how people move through the real world" according to [their website](https://foursquare.com/about). They make available an API through which you can obtain relevant information about different places around the world.

The idea is to obtain information about all the other coffee shops in every neighborhood and start the analysis process.

1. **Data Preparation**

First, we obtain the information about the neighborhoods of Toronto. This information was obtained through web scraping and stored in a table, where for each neighborhood, we have different information such as zip code and borough.

The first data cleaning that was done was to leave aside the postal codes that were not defined in any neighborhood: "Not Assigned" in our table.

Then all the neighborhoods are grouped, since for us the postal code is not relevant and we know that there are neighborhoods that have more than 1 postal code. The idea is to have our table without repeated data.

At this point we are able to obtain, for each neighborhood, its geographical coordinates. For this, the free Nominatim API that we have in Python was used with the help of the geopy library. This process consists of going through our table and for each neighborhood, making up to 3 geolocation attempts, since in certain occasions, the API answers with empty data. We are also keep in mind the restriction described on [this page](https://operations.osmfoundation.org/policies/nominatim/) where we cannot make more than 1 request per second to the Nominatim API.

Despite having tried up to 3 times for each neighborhood, there were some cases where the API answered empty results. This is because they could not geolocate them under the name of the neighborhood we use. These cases are detected and analyzed that do not represent an impediment to our analysis since they represent less than 10% of the total Toronto neighborhoods. It is a very small sample to contradict any results we obtain at the end of our project.

We continue now analyzing whether we have duplicate neighborhoods in our table. At this point the coordinates obtained by the API were analyzed and in cases where we find 2 exactly the same coordinates, we assume that they correspond to the same neighborhood. What happened at this point is that there are neighborhoods in Toronto such as Harbourfront that are divided into more than 1, in this case in North and South. For the API, the 3 neighborhoods are the same and return the same geographic coordinate for all cases.

As a last point regarding the coordinates of the neighborhoods, it was analyzed in a map visualization, using the folium library, if we had many uncovered areas. The visualization did not determine many Toronto gaps without analyzing, so the analysis continues without additional comments.

To start using the Foursquare API and obtain our second set of data, it is important that for each neighborhood we define the analysis radios. This would be to understand how many meters around each neighborhood, in the form of a circle, we will use since it is a parameter that the Foursquare API needs to give us information.

Therefore, the logic to define the radios is as follows:

1) The distances between each neighborhood are calculated

2) The closest joint neighborhood is taken for each neighborhood

3) The distance to the nearest neighborhood is divided into 2

We do this to avoid overlap between each neighborhood and not duplicate data from venues.

Now we are going to get our second data set using the Foursquare API. What we get for each neighborhood, are the list of venues within the defined radius. The fields of our second table are:

- Venue

- Latitude

- Length

- Category

In addition to our first table, we now have all the data we need to start with their exploratory analysis.

1. **Methodology Section: exploratory data analysis**
2. **Venues Categories of interest**

The Foursquare API returns us an extensive list of venues for each neighborhood and each listing in turn contains many categories to classify each venue. What we do then, is to pass all those categories to a single list and understand which of those categories are the ones that interest us.

As mentioned in previous sections, the objective of our project is to determine the best place to open a coffee business, so what we analyze from the entire list of previously defined categories is if we have to take more than 1 venue category to determine which correspond to our future competition.

For this analysis, it was observed that a category that we have to take into account is "Coffee Shop". But, in order not to be left alone with that one, we look for all the categories that begin with the letter "c". It follows the second category that we will consider that it will be "Café".

|  |  |
| --- | --- |
| **Category** | **Count** |
| Coffee Shop | 236 |
| Café | 96 |
| Clothing Store | 60 |
| Chinese Restaurant | 52 |
| Convenience Store | 26 |
| Cosmetics Shop | 20 |
| Caribbean Restaurant | 17 |
| Cocktail Bar | 10 |
| Cantonese Restaurant | 7 |
| Cheese Shop | 4 |
| Comfort Food Restaurant | 4 |
| Concert Hall | 3 |
| Chiropractor | 2 |
| Costume Shop | 2 |
| Chocolate Shop | 2 |
| Curling Ice | 2 |
| Candy Store | 2 |
| Comic Shop | 2 |
| Cuban Restaurant | 2 |
| Cajun / Creole Restaurant | 2 |
| College Rec Center | 2 |
| Camera Store | 1 |
| Cupcake Shop | 1 |
| Creperie | 1 |
| Campground | 1 |
| Climbing Gym | 1 |
| Comedy Club | 1 |
| Cycle Studio | 1 |
| Convention Center | 1 |
| Construction & Landscaping | 1 |
| Country Dance Club | 1 |

We do not intend that our business corresponds only to 1 single category since it would cease the analysis of our competition. That is why we determine that, since a venue of the "Coffee Shop" category is not exactly the same as a venue of the "Café" category, we take them all as future competition.

1. **Distribution of interested venues categories**

With our 2 previously defined interest categories, now what we want to understand is how the venues that belong to these categories are distributed.

For this, what we calculate is the amount of sales for each neighborhood and we store them in a third table. This can already give us a better idea of what are the neighborhoods where we are going to find more competition if we open our business.

Then, we can determine a top 20 neighborhoods where we do not want to start our venture in order to avoid the aforementioned competition.

At this point we are not going to do a correlation analysis to understand if the amount of competition increases or decreases in the hands of some other category since then in the methodology part we will apply KNN to determine the neighborhoods similar to each other.

1. **Define new variable: density of venues**

Now that we have determined the number of venues per neighborhood that we are interested in analyzing, we can think of a new type of variable that will also determine which neighborhoods resemble each other.

We call this new variable "density of venues" where the calculation we use is the amount previously calculated and the radius of analysis of each neighborhood.

Then, if the amount of venues we divide it by the radius meters, we obtain the density defined previously. This variable helps us to relate the analysis radius to the number of venues since we can have a very large radius with a small amount of competition or a very small radius also with a small amount of competition.

The density can determine that the neighborhood with a small radius is not a candidate even though it has a small amount of venues. In contrast, the neighborhood where we analyze a larger radius, having a small number of premises will give us a much lower density and is a better indicator for our main objective.

This is the top 20 non-candidates defined before:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Neighborhood** | **Quantity\_Venues** | **Latitude** | **Longitude** | **Radius** | **Venues\_Density** |
| First Canadian Place | 3 | 43.6487681 | -79.3816917928303 | 78.804068 | 0.038069 |
| Harbourfront | 9 | 43.6400801 | -79.3801495 | 256.488743 | 0.035089 |
| Yorkville | 14 | 43.6713861 | -79.3901677 | 414.454103 | 0.033779 |
| Kensington Market | 7 | 43.6552136 | -79.4022604 | 212.473580 | 0.032945 |
| Central Bay Street | 5 | 43.660912 | -79.3858973 | 188.795342 | 0.026484 |
| Queen's Park | 4 | 43.6606092 | -79.3905725 | 188.795342 | 0.021187 |
| Deer Park | 8 | 43.68809 | -79.3940935 | 384.619903 | 0.020800 |
| Commerce Court | 1 | 43.64809515 | -79.3790207204735 | 48.202675 | 0.020746 |
| Union Station | 3 | 43.6446934 | -79.3801320058597 | 156.945027 | 0.019115 |
| Toronto Dominion Centre | 1 | 43.64736955 | -79.3813733580709 | 54.697236 | 0.018282 |
| Runnymede | 7 | 43.6517026 | -79.4759978 | 386.608289 | 0.018106 |
| Cabbagetown | 6 | 43.6644734 | -79.3669861 | 357.777862 | 0.016770 |
| Roselawn | 11 | 43.7098517 | -79.4042948 | 704.952281 | 0.015604 |
| South of Bloor | 1 | 43.6644447 | -79.3987444 | 64.912373 | 0.015405 |
| Willowdale | 15 | 43.7615095 | -79.4109234 | 986.305994 | 0.015208 |
| Ryerson | 2 | 43.65846945 | -79.3789932724589 | 133.075058 | 0.015029 |
| Garden District | 2 | 43.6564995 | -79.3771141 | 133.075058 | 0.015029 |
| CN Tower | 3 | 43.6425637 | -79.3870871832047 | 203.213085 | 0.014763 |
| Little Portugal | 6 | 43.64741325 | -79.4311163254605 | 407.796462 | 0.014713 |
| Christie | 3 | 43.6641106 | -79.4184051 | 216.637201 | 0.013848 |

1. **Methodology Section: clustering modeling**

As mentioned in the previous section, what we will use as a methodology is the KNN algorithm. The objective will be to determine 5 clusters of neighborhoods similar to each other in order to find 1 cluster of interest and then find the candidate neighborhoods for our business.

The cluster we are looking for is the one where coffee businesses are in greater quantity. The question that anyone would ask at this time is "if we are looking for neighborhoods with LOWER competition, why are we interested in the cluster with GREATER amount of competition?"

The answer is simple: the more you come from our competition coexist in the same neighborhood, it means that they stay with customers on a frequent basis.

This can be due to several factors, including for example if they are neighborhoods with more offices or universities or tourist places where there is more flow of people who can approach to a coffee business.

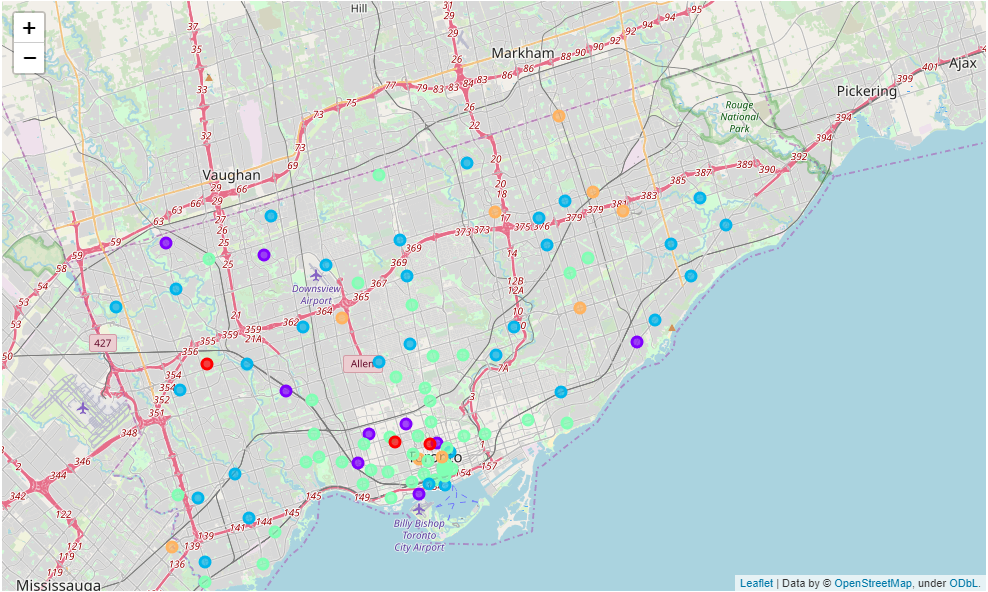
Anyway, as these variables we are not analyzing them, if we are analyzing almost a consequence of all the above: we have the rest of the venues of other categories that also coexist with coffee businesses in the same neighborhood.

Therefore, neighborhoods similar to each other by "categories of venues" are also a direct consequence of these variables that we are not analyzing.

We begin with determining the categories with the technique of "one-hot encoding" and thus have a new data set where we have a column by venue category for each neighborhood.

For each neighborhood, we determine the average of each category. At this point we already have our set of input variables to apply KNN but before that we will add the new variable that we defined in the previous section: the density of venues.

Once we add it to our table, we apply the model with 5 clusters and we have already defined to which each neighborhood belongs. If we see them on a map:



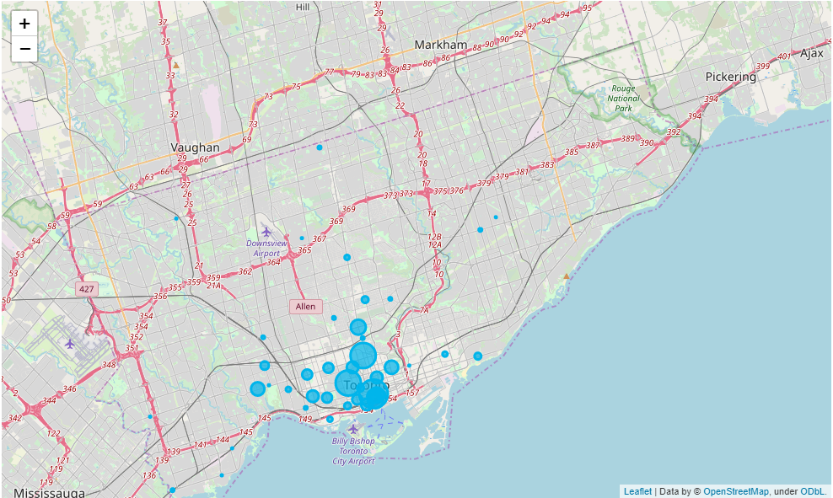
We can see that there are 2 clusters in particular with quite a lot of neighborhoods (light blue and green). Taking up our "non-candidates" table, what happens if we add the "cluster number" column:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Neighborhood** | **Quantity\_Venues** | **Latitude** | **Longitude** | **Radius** | **Venues\_Density** | **Cluster\_Label** |
| First Canadian Place | 3 | 43.6487681 | -79.3816917928303 | 78.804068 | 0.038069 | 3.0 |
| Harbourfront | 9 | 43.6400801 | -79.3801495 | 256.488743 | 0.035089 | 2.0 |
| Yorkville | 14 | 43.6713861 | -79.3901677 | 414.454103 | 0.033779 | 3.0 |
| Kensington Market | 7 | 43.6552136 | -79.4022604 | 212.473580 | 0.032945 | 3.0 |
| Central Bay Street | 5 | 43.660912 | -79.3858973 | 188.795342 | 0.026484 | 1.0 |
| Queen's Park | 4 | 43.6606092 | -79.3905725 | 188.795342 | 0.021187 | 0.0 |
| Deer Park | 8 | 43.68809 | -79.3940935 | 384.619903 | 0.020800 | 3.0 |
| Commerce Court | 1 | 43.64809515 | -79.3790207204735 | 48.202675 | 0.020746 | 3.0 |
| Union Station | 3 | 43.6446934 | -79.3801320058597 | 156.945027 | 0.019115 | 3.0 |
| Toronto Dominion Centre | 1 | 43.64736955 | -79.3813733580709 | 54.697236 | 0.018282 | 3.0 |
| Runnymede | 7 | 43.6517026 | -79.4759978 | 386.608289 | 0.018106 | 3.0 |
| Cabbagetown | 6 | 43.6644734 | -79.3669861 | 357.777862 | 0.016770 | 3.0 |
| Roselawn | 11 | 43.7098517 | -79.4042948 | 704.952281 | 0.015604 | 2.0 |
| South of Bloor | 1 | 43.6644447 | -79.3987444 | 64.912373 | 0.015405 | 3.0 |
| Willowdale | 15 | 43.7615095 | -79.4109234 | 986.305994 | 0.015208 | 2.0 |
| Ryerson | 2 | 43.65846945 | -79.3789932724589 | 133.075058 | 0.015029 | 3.0 |
| Garden District | 2 | 43.6564995 | -79.3771141 | 133.075058 | 0.015029 | 2.0 |
| CN Tower | 3 | 43.6425637 | -79.3870871832047 | 203.213085 | 0.014763 | 3.0 |
| Little Portugal | 6 | 43.64741325 | -79.4311163254605 | 407.796462 | 0.014713 | 3.0 |
| Christie | 3 | 43.6641106 | -79.4184051 | 216.637201 | 0.013848 | 3.0 |

We find a very curious fact: of these 20 neighborhoods, there are 14 that belong to cluster number 3. This is telling us that the neighborhoods with the highest density of competition are concentrated in a single cluster, number 3. Then, with the reasoning explained above, this is the cluster we were looking for!

With our cluster of defined interest, what we do then is to keep the list of all the neighborhoods belonging to it (a total of 44). The final analysis that we will do will be to visualize them on the map but adding the "density" factor to visually see how they are distributed.

To do this, for each neighborhood we make a circle where the radius of that circle will be directly related to density, the higher the density of competition we have, the larger the circle will be. This is what we see:



1. **Results and Discussion Section**

As you can see in the methodology section, we already find a reduced number of candidate neighborhoods to open our business.

It was concluded that we find 44 neighborhoods similar to each other where coffee businesses today have a fluid amount of customers, that is, since they started, they continue today and that tells us they did well.

As our main objective is not to compete directly with a lot of other businesses, what we do with that list of 44 neighborhoods is to reduce it even more. For this, we use the density of venues to determine the neighborhoods where there is less density, that is, from the last visualization we look for the smallest circumference radii. We find the following list:

- Forest Hill North (Group 1)

- Parkdale (Group 1)

- Leaside (Group 1)

- Riverdale (Group 1)

- High Park (Group 2)

- Silverthorn (Group 2)

- Wexford (Group 2)

- Wilson Heighs (Group 2)

We define 2 groups of neighborhoods because if you look at the map, we see that there are points near where there is greater competition density and others that are further away.

We understand that any of these 2 groups are candidates but refer to a different objective.

Group 1 is closer to the "density of competition" but we would be located in neighborhoods with low density of coffee shops, therefore, we would be the new place that will have close but not direct competition and we are in a "area of ​​success" where other businesses did well.

On the contrary, in group 2 we find neighborhoods in the areas furthest from the largest concentration of coffee businesses, so if we open our business in that area, we will almost have no competition, but it will be in an area where we will not have our permanence is so assured, since if other premises were not located nearby, it must be due to some factor or variable that we have not analyzed. However, we can take this as a challenge with greater risk, but at the same time more likely to have a higher profit.

1. **Conclusion Section**

Those interested in the project (or starting their own business) now have a better idea of how to use public data to prepare a first analysis of one of the biggest doubts they have at the beginning, which would be "Where to locate my business?".

It is not an easy question to answer nor can 100% efficiency be guaranteed, but we can get close enough.

The final decision will be taken not only with our analysis in mind, there are also other factors such as culture of each neighborhood, attractive building, tourist density, or others that can further limit the list of candidate neighborhoods.

We hope it is a contribution to help make the difficult decision to enter the world of entrepreneurs and to be able to say the famous phrase "I am my own boss".