Where should a chef open a restaurant in Toronto?

Paul Vidal
August 2019

1 Introduction

French Chef Jacques Durand wishes to open a new French restaurant in Toronto called *La Bonne Cuisine*.
In order to ensure the restaurant's success, he hired a data scientist to analyze the most successful restaurants in Toronto, their location.
One particularly interesting aspect will be to see if French restaurants are more successful in one neighborhood or another.

2 Data preparation

To perform this analysis, the project leveraged three sources:

- The Foursquare Venue API (link)
- The Wikipedia list of Postal Codes in Canada (link)
- A list of latitude and longitude per Canada postal code (link)

Data preparation for this project consisted of the following steps:

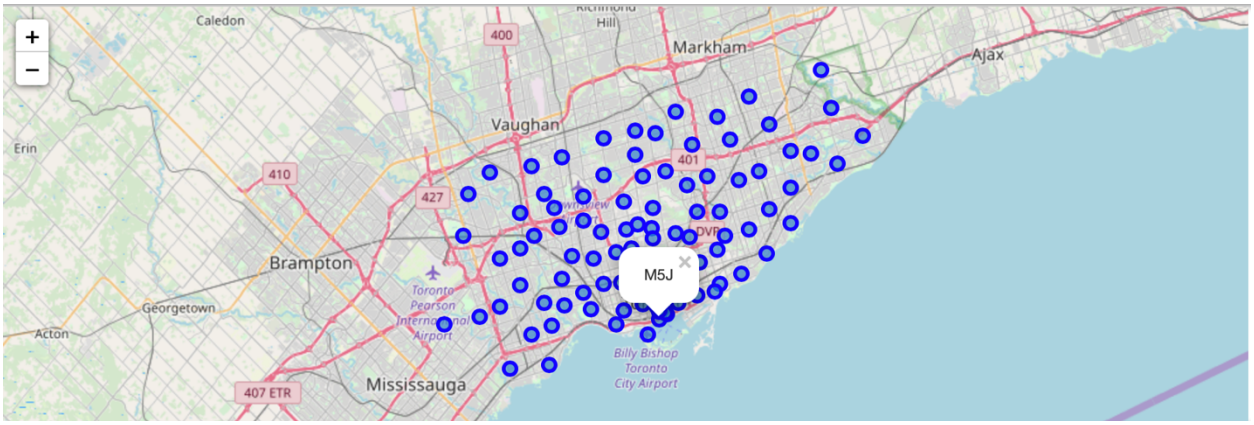
1. Scrape postal code from the Wikipedia page using python library BeautifulSoup
2. Combine the extracted postal codes with their latitude and longitude
3. Use the explore Foursquare API call to get all venues associated with a Postal Code.

While scraping is a straightforward process, eliminating the “Not assigned” PostalCodes was important for data preparation.

The table below shows an overview of the end result of the combination of the scraped Wikipedia data and the latitude and longitude per Canada postal code:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	[North York]	[Parkwoods]	43.753259	-79.329656
1	M4A	[North York]	[Victoria Village]	43.725882	-79.315572
2	M5A	[Downtown Toronto, Downtown Toronto]	[Harbourfront, Regent Park]	43.654260	-79.360636
3	M6A	[North York, North York]	[Lawrence Heights, Lawrence Manor]	43.718518	-79.464763
4	M7A	[Queen's Park]	[Not assigned]	43.662301	-79.389494

The map below shows a visualization of the location of these different Postal Codes:



Iterating through this list of PostalCodes and Latitude and Longitude, and leveraging the Foursquare API, it is easy to extract the different location nearby a Postal Code, as depicted below:

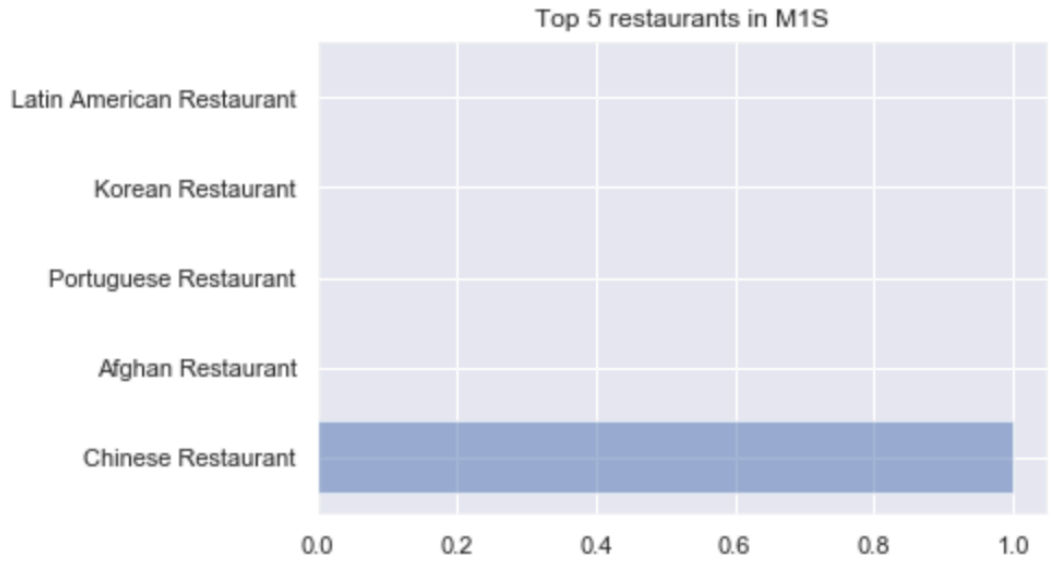
	PostalCode	PostalCode Latitude	PostalCode Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue ID
0	M3A	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park	4e8d9dcdd5fbbbb6b3003c7b
1	M3A	43.753259	-79.329656	KFC	43.754387	-79.333021	Fast Food Restaurant	4e6696b6d16433b9ffff47c3
2	M3A	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop	4cb11e2075ebb60cd1c4caad
3	M4A	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena	4c633acb86b6be9a61268e34
4	M4A	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop	4bbe904a85fbb713420d7167

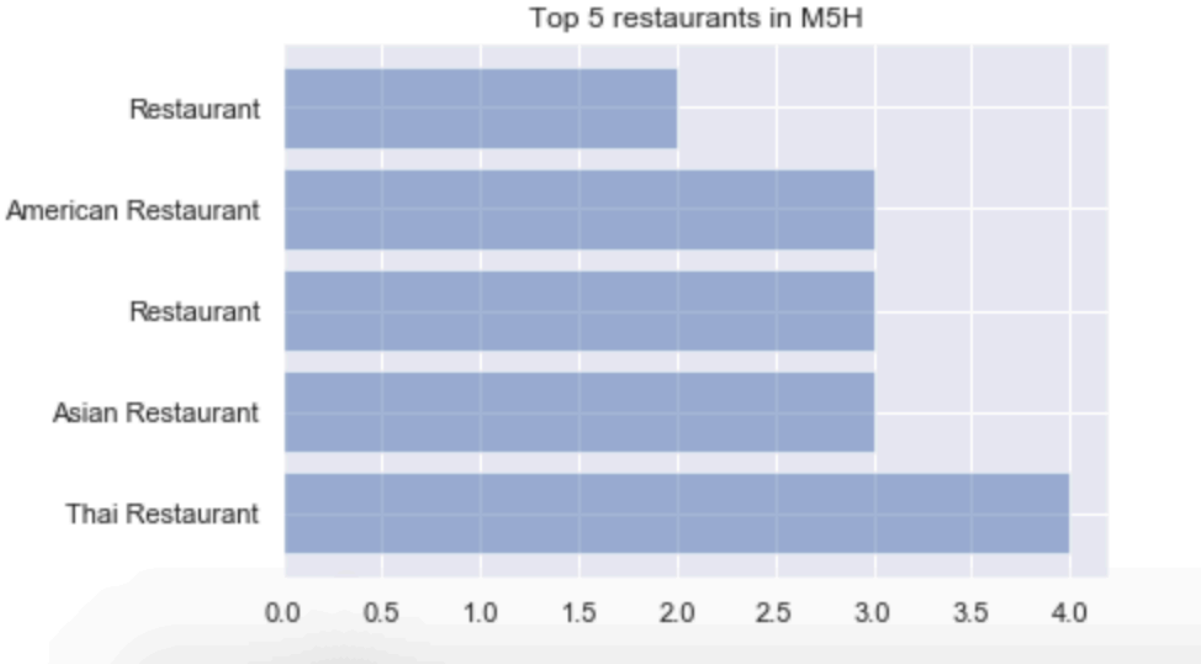
The final stage of this data preparation was to only extract the Restaurant venues. This was done by searching for every venue category containing the word Restaurant.

3 Exploratory Data Analysis

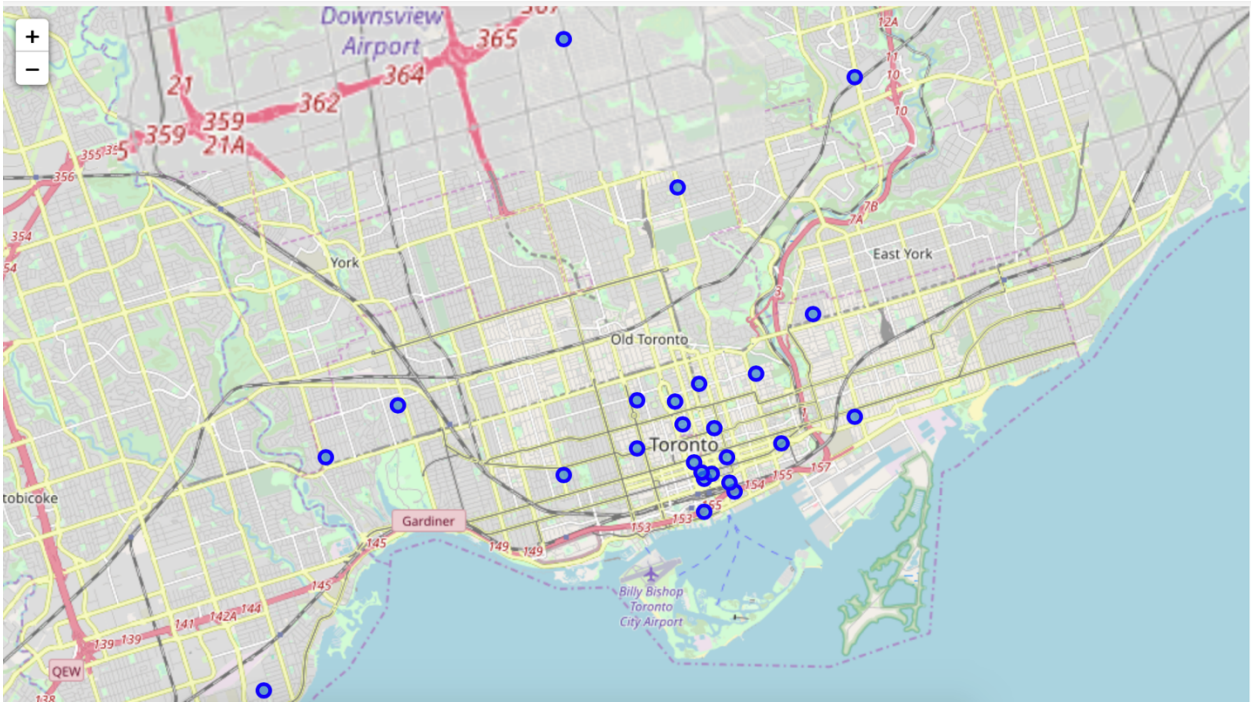
This section of the project explores the neighborhoods with the most restaurant, per category of restaurant.

Leveraging one hot encoding, the result of the search found that many areas had very little restaurant presence, when others had many:





Mapping the postal codes with the most restaurants, the areas that seem the most popular for a restaurant are close to the center of Toronto and not on its outskirts:



4 Predictive Modeling

4.1 Model building

Based on this initial exploration, it seems that being close to the bay is where most restaurants open. The next goal of the project is to build a predictive model that will determine the potential success of a restaurant based on its latitude and longitude.

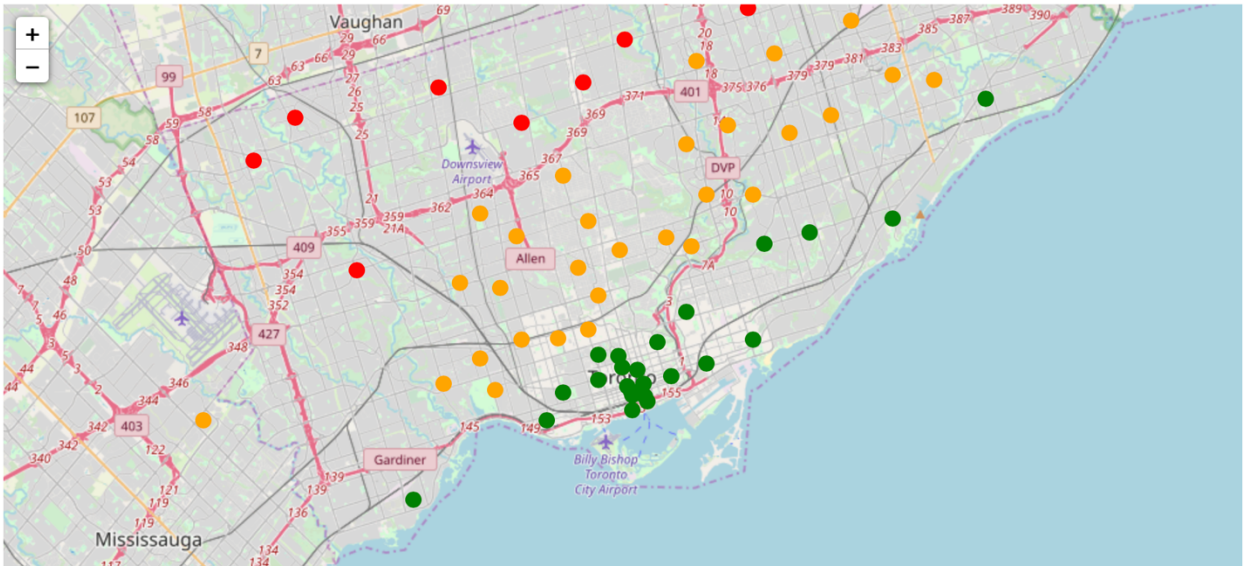
In this project, the rating of a venue was used to qualify the success of a restaurant.

The model built where therefore a linear regression model using:

- X: Latitude and Longitude of a Venue
- Y: Venue rating

By splitting the data set obtained from the Foursquare API into 75% training and 25% testing we obtained a linear model with a Mean squared error of 2.81 and a variance error of 0.38.

By using postal codes latitude and longitude, we can map out the predicted ratings of a restaurant based per postal code. The maps below shows these predictions (red = rating <5, orange = rating between 5 and 7, green = rating > 7):



This map shows that the model of rating seems to be in accordance with the number of restaurants in Toronto: the most popular predicted ratings are in the same area where most restaurants are!

4.2 Known Limitations

This model has a lot of limitations:

1. Because of Foursquare limitation of premium calls, not all venue data was retrieved
2. Using rating as the only measure of success of a restaurant is very limiting (some venues have not been rated, etc.). Using additional metrics such as check-in counts could help but the data found in the Foursquare API didn't have enough check-in information.
3. As a result the model created does not have good predictability (see MSE and R2 above)

5 Conclusion

Based on this preliminary analysis, the neighborhood to suggest to Chef Jacques Durand are all located right in the center of Toronto.

With more data and cross referencing over time, better predictions could be made.