Predicting the success of a venue in Toronto.

Mario Andre Molina Caballero
mario.andre.molina@gmail.com

Applied data science capstone project, first deliverable.
IBM data science professional certificate.

August 14, 2020

1. Description of the problem and background

Over the course of the last six months working in this certification, we have reviewed different topics in data science fields, including methodologies, tools and data analysis, modelling, evaluation and visualization; lastly, we explored how to use the Foursquare API in order to query its comprehensive venues database. The purpose of this project is to apply all this acquired knowledge in order to 'attempt' to solve a real-world problem, I have decided to try to predict the success of a venue in Toronto, using the venue rating as the parameter for success.

Data science is an interdisciplinary field that combines knowledge from well-known and developed disciplines such as statistics, probability and mathematics with the capabilities that modern tools and software provide for us, including data mining, data bases and higher computing power and storage. All of this allows us to manipulate huge amounts of often instructed data in order to identify trends, obtain previously hidden insights and even predicting future outcomes.
It is the job of the data scientist to obtain value from the data and communicate the findings to stakeholders in the project in order to allow them to make better decisions or decide a new strategy. In order to correctly convey the most important information obtained, a data scientist must build a narrative and employ data visualization techniques such as charts, maps and plots.

In the last course of the certification, we learned about Foursquare and how to query its database using the publicly available API. Foursquare is the leading independent location technology platform, its core product is the 'Places database' which has information for more than 60 million points of interest including restaurants, shops and services across more than 190 countries. The information provided for each of these points of interest includes the venue name, category, location, rating and even user generated data such as photos and tips. All of this data is available through the 'Places API', after generating the credentials you can send API requests to the different Foursquare endpoints, each of them providing a different response.

Finally, I wanted to apply all of this knowledge to try to predict the rating a venue in Toronto would obtain given the location and its category. Toronto is the most populous city in Canada, with 2.7 million habitants as of 2016, a truly multicultural and cosmopolitan city. Given its importance as a cultural and economic hotspot in Canada and the world, it seemed interesting to try and predict the success of a new business in this city, for it will surely be a challenging enterprise for any entrepreneur.

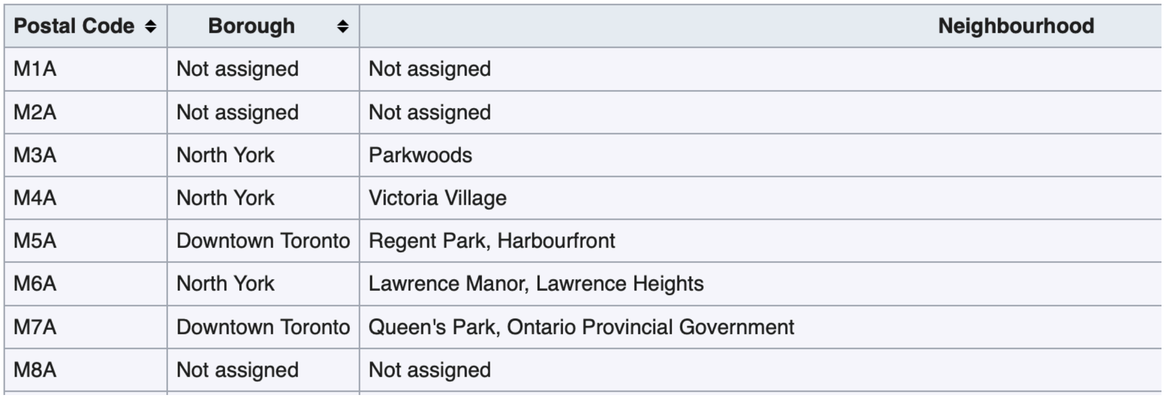
2. Description of the data and how it will be used to solve the problem

We will be using a Jupyter notebook running python hosted in the 'Skill Networks Lab' by IBM.

Some of the libraries that we will be using in our python script are:

- BeautifulSoup: for web scraping.
- Requests: for http request handling.
- Numpy: numerical computing tools.
- Pandas: data analysis and manipulation tools.
- IPython.display: for displaying images and web content.
- Geopy: python client that enables coordinate location.
- Matplotlib: for visualizations.
- Sklearn: machine learning tools.
- Folium: for map visualization.
- Seaborn: for visualization.

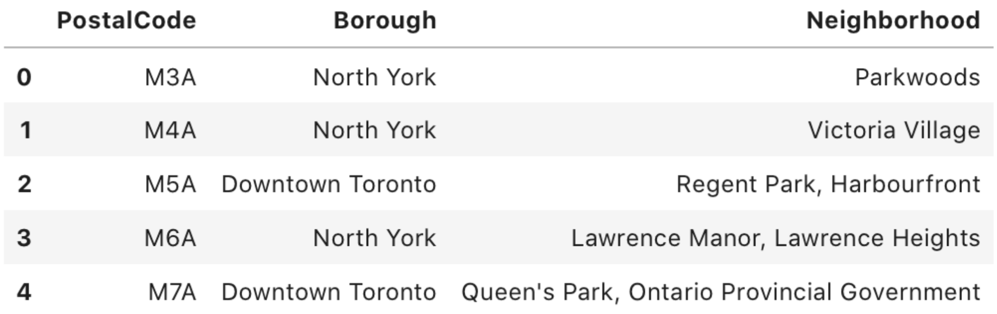
The first thing was to obtain the list of neighborhoods in Toronto, with the 'BeautifulSoup' library, we could send a request directly Wikipedia to read the following table into the Jupyter notebook:



Postal Code ⬍	Borough ⬍	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
M8A	Not assigned	Not assigned

Figure 2.1: Wikipedia list for Canadian postal codes.

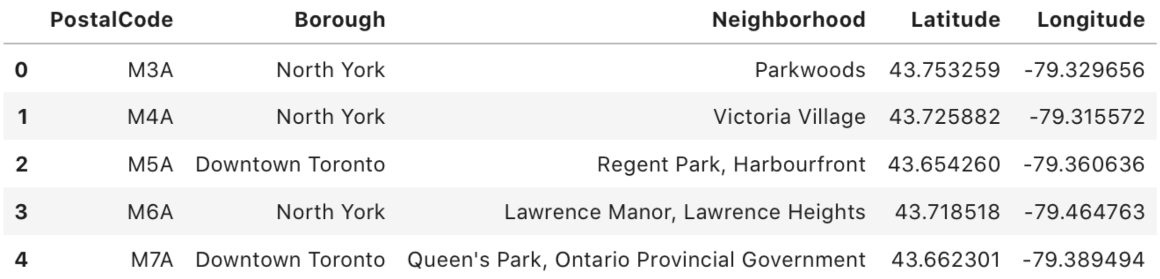
After some manipulation, we obtained the following pandas dataframe:



	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Figure 2.2: First dataframe obtained

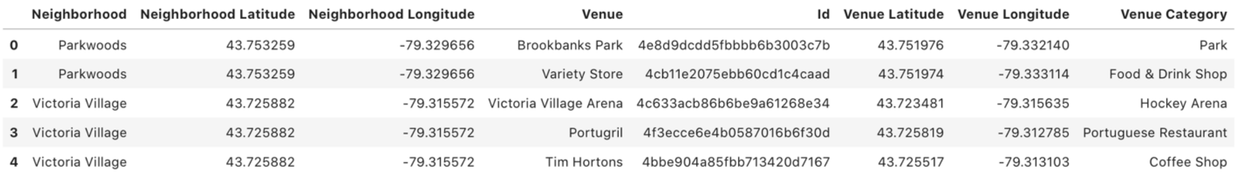
The next step was to obtain the coordinates for the postal codes obtained, downloaded from the csv file: https://cocl.us/Geospatial_data, after merging the data, this was the dataframe obtained:



	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	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Figure 2.3: Dataframe with neighborhood and coordinates data.

The next step was to get venue information for Toronto, so using the Foursquare API, we queried their Places Database in order to obtain at most 100 venues for each neighborhood, after cleaning the response and merging with our previous dataframe, we obtained this new one, including venue name, ID, coordinates and category:



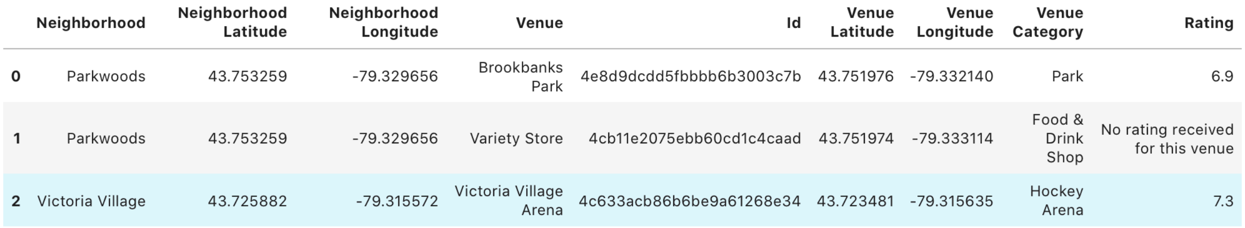
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Id	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	4e8d9dcdd5fbbbb6b3003c7b	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	4cb11e2075ebb60cd1c4caad	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	4c633acb86b6be9a61268e34	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Portugril	4f3ecce6e4b0587016b6f30d	43.725819	-79.312785	Portuguese Restaurant
4	Victoria Village	43.725882	-79.315572	Tim Hortons	4bbe904a85fbb713420d7167	43.725517	-79.313103	Coffee Shop

Figure 2.4: Dataframe with venue data.

It’s important to mention that this request made to the places database was pointing to the ‘explore’ endpoint, a ‘regular call’, which means we get to do 99,500 of such calls each day.

In order to get the ratings for the venues retrieved, we had to query the 'details' endpoint, this is a 'premium call' and as such we only get 500 requests a day. From the previous steps we had obtained venue data for about 2,100 venues in Toronto, so in order to get the ratings for each of them, we had to split the full dataframe in 5 smaller ones, in order to query for the ratings 500 venues a day, and at the end, concatenate them back together.

Finally, we ended up with this final data set of venues for Toronto:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Id	Venue Latitude	Venue Longitude	Venue Category	Rating
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	4e8d9dcdd5fbbbb6b3003c7b	43.751976	-79.332140	Park	6.9
1	Parkwoods	43.753259	-79.329656	Variety Store	4cb11e2075ebb60cd1c4caad	43.751974	-79.333114	Food & Drink Shop	No rating received for this venue
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	4c633acb86b6be9a61268e34	43.723481	-79.315635	Hockey Arena	7.3

Figure 2.5: Final dataframe with full venue data, including rating.

We have 2,384 rows and 11 columns in our dataframe:

- 2384 venues.
- 96 neighborhoods.
- 264 categories.

References

- van der Aalst W. (2016) Data Science in Action. In: Process Mining. Springer, Berlin, Heidelberg.
  https://doi.org/10.1007/978-3-662-49851-4_1

- (2020). Places Database. Foursquare.
  https://developer.foursquare.com/docs/places-database/

- (2020). Places API. Foursquare.
  https://developer.foursquare.com/docs/places-api/