Where is my new store in York Region

Brian Su

Feb 14th, 2019

Introduction

Background

My business is expanding so I want to open a new grocery store in York Region. But what neighborhood should I choose?

If I open it at a rural area where I don't have enough potential customers, I can't make the business to run properly.

If I open it at a great area but there already have enough similar stores, I may not be able to get enough revenue.

I want to find a neighborhood where (1) have a fair number of households, (2) doesn't have enough competitors.

I believe I can have the best business at there.

Problem

How to get the data I need?

How to analyze the data I got?

How to make the correct decision based on the analysis?

What are the tools I can use?

Data Acquisition and Cleaning

Data Sources

- ▶ York Region neighborhood data can be obtained from Wikipedia.
- ▶ The latitudes and longitudes data can be acquired from the Geocoder Python data or Google.
- ▶ The location & venues data can be retrieved from Foursquare Places API.

Data Cleaning

For the neighborhood data I scraped from Wikipedia, I need to remove useless data, fill the NULL cells, and group some records. I also need to de-duplicate when needed.

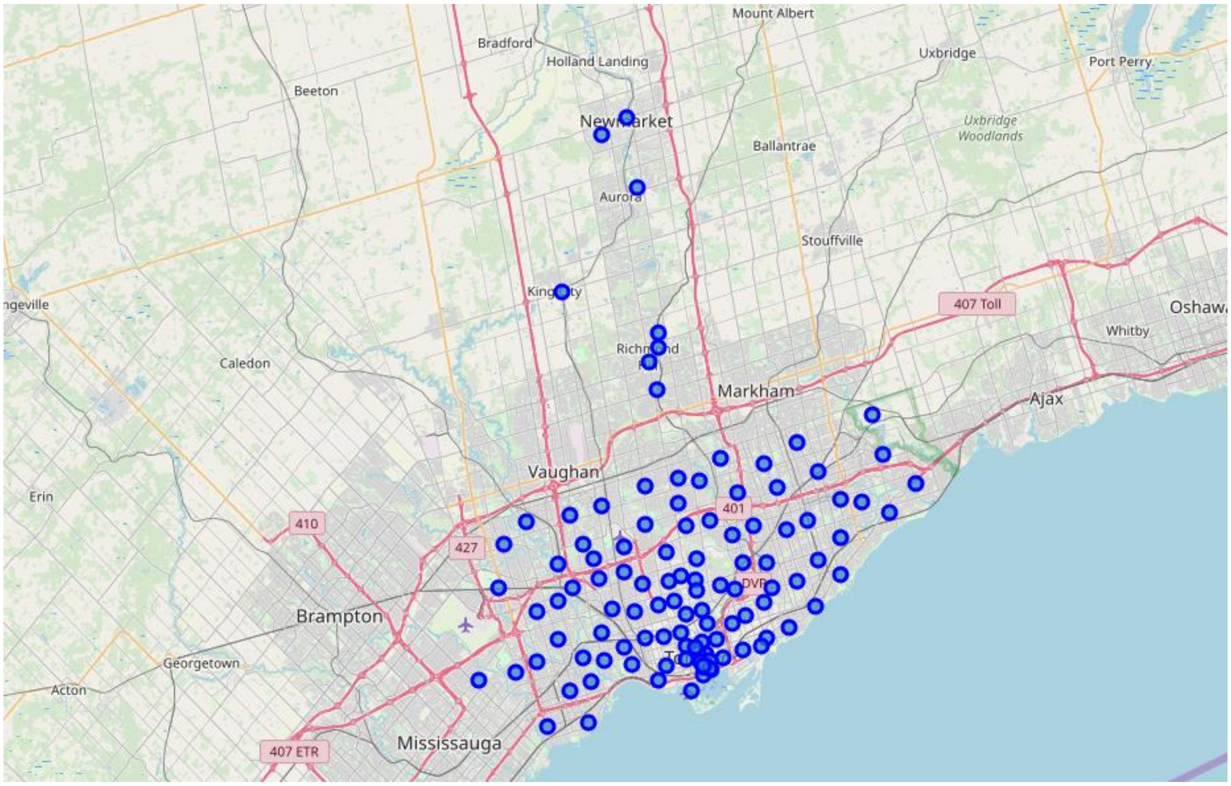
For the location data I got from Foursquare, the venue categories are too fine-grained to what I really need. So I have to downgrade the granularity.

Data Analysis

Data Visualization

Because this is very location-wise, I need to visualize the data in a map.

This is too intuitive and helpful.

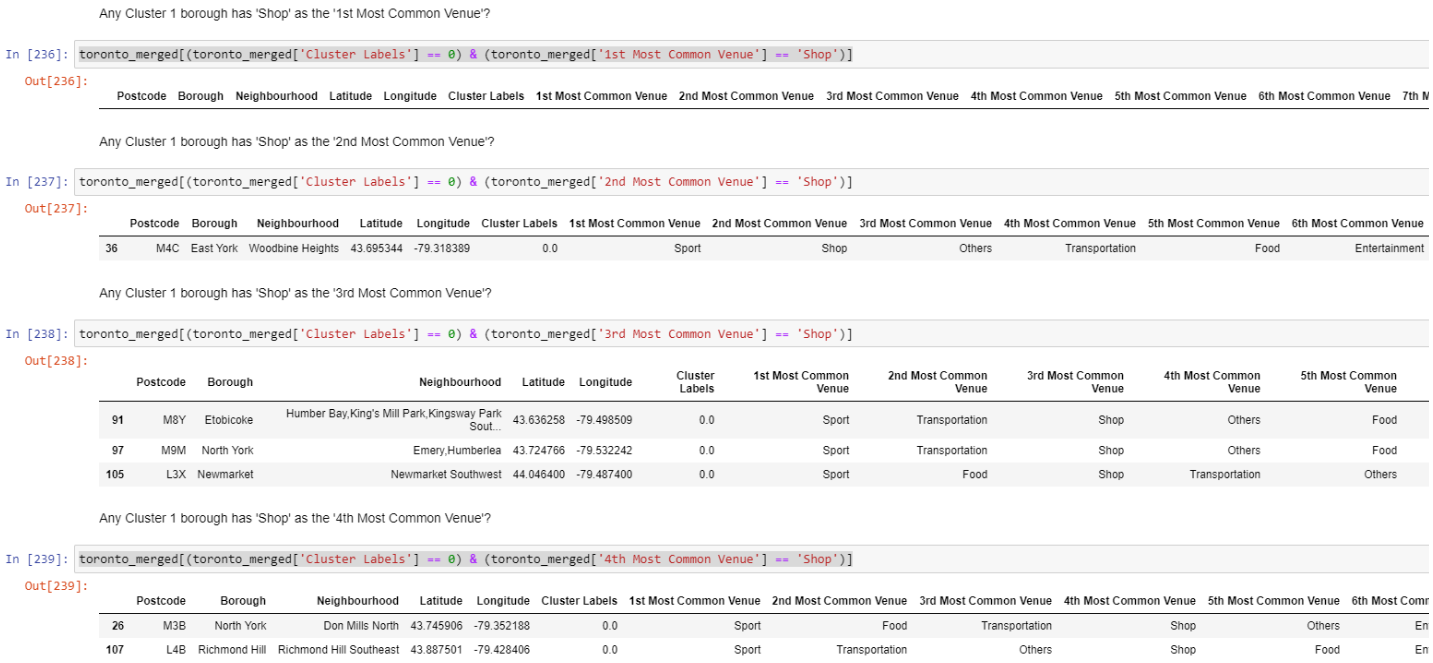


Data Mining

But it is still not enough. I need the hot spots only.

So I need to get the location venue data, then use k-means clustering to categorize all the neighborhoods.

I will create 5 clusters, and Cluster 1 is the hottest.



Any Cluster 1 borough has 'Shop' as the '1st Most Common Venue'?

```
In [236]: toronto_merged[(toronto_merged['Cluster Labels'] == 0) & (toronto_merged['1st Most Common Venue'] == 'Shop')]
```

Out[236]:

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th M

Any Cluster 1 borough has 'Shop' as the '2nd Most Common Venue'?

```
In [237]: toronto_merged[(toronto_merged['Cluster Labels'] == 0) & (toronto_merged['2nd Most Common Venue'] == 'Shop')]
```

Out[237]:

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
36	M4C	East York	Woodbine Heights	43.695344	-79.318389	0.0	Sport	Shop	Others	Transportation	Food	Entertainment

Any Cluster 1 borough has 'Shop' as the '3rd Most Common Venue'?

```
In [238]: toronto_merged[(toronto_merged['Cluster Labels'] == 0) & (toronto_merged['3rd Most Common Venue'] == 'Shop')]
```

Out[238]:

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
91	M8Y	Etobicoke	Humber Bay,King's Mill Park,Kingsway Park Sout...	43.636258	-79.498509	0.0	Sport	Transportation	Shop	Others	Food
97	M9M	North York	Emery,Humberlea	43.724766	-79.532242	0.0	Sport	Transportation	Shop	Others	Food
105	L3X	Newmarket	Newmarket Southwest	44.046400	-79.487400	0.0	Sport	Food	Shop	Transportation	Others

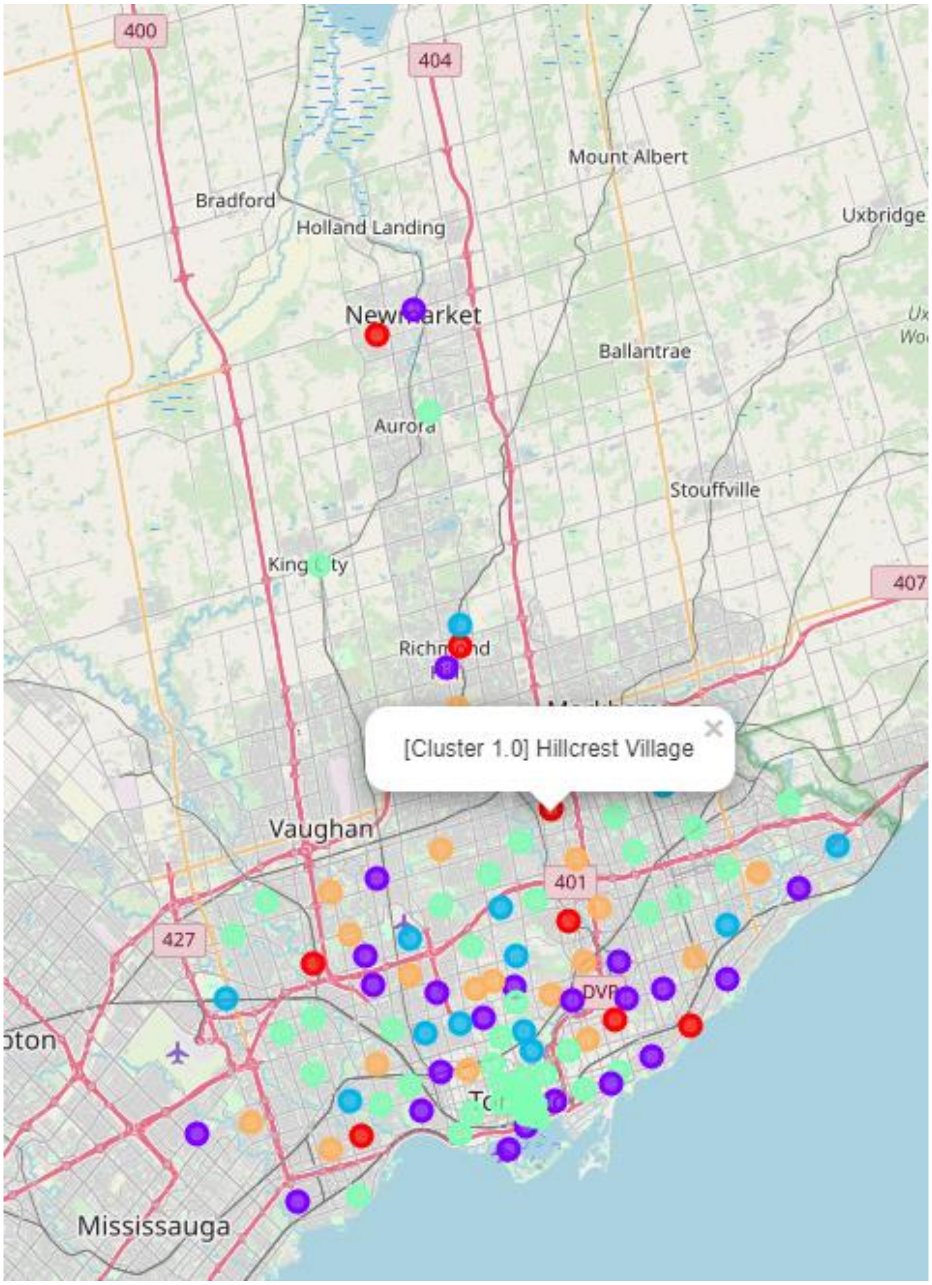
Any Cluster 1 borough has 'Shop' as the '4th Most Common Venue'?

```
In [239]: toronto_merged[(toronto_merged['Cluster Labels'] == 0) & (toronto_merged['4th Most Common Venue'] == 'Shop')]
```

Out[239]:

	Postcode	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Comm
26	M3B	North York	Don Mills North	43.745906	-79.352188	0.0	Sport	Food	Transportation	Shop	Others	En
107	L4B	Richmond Hill	Richmond Hill Southeast	43.887501	-79.428406	0.0	Sport	Transportation	Others	Shop	Food	En

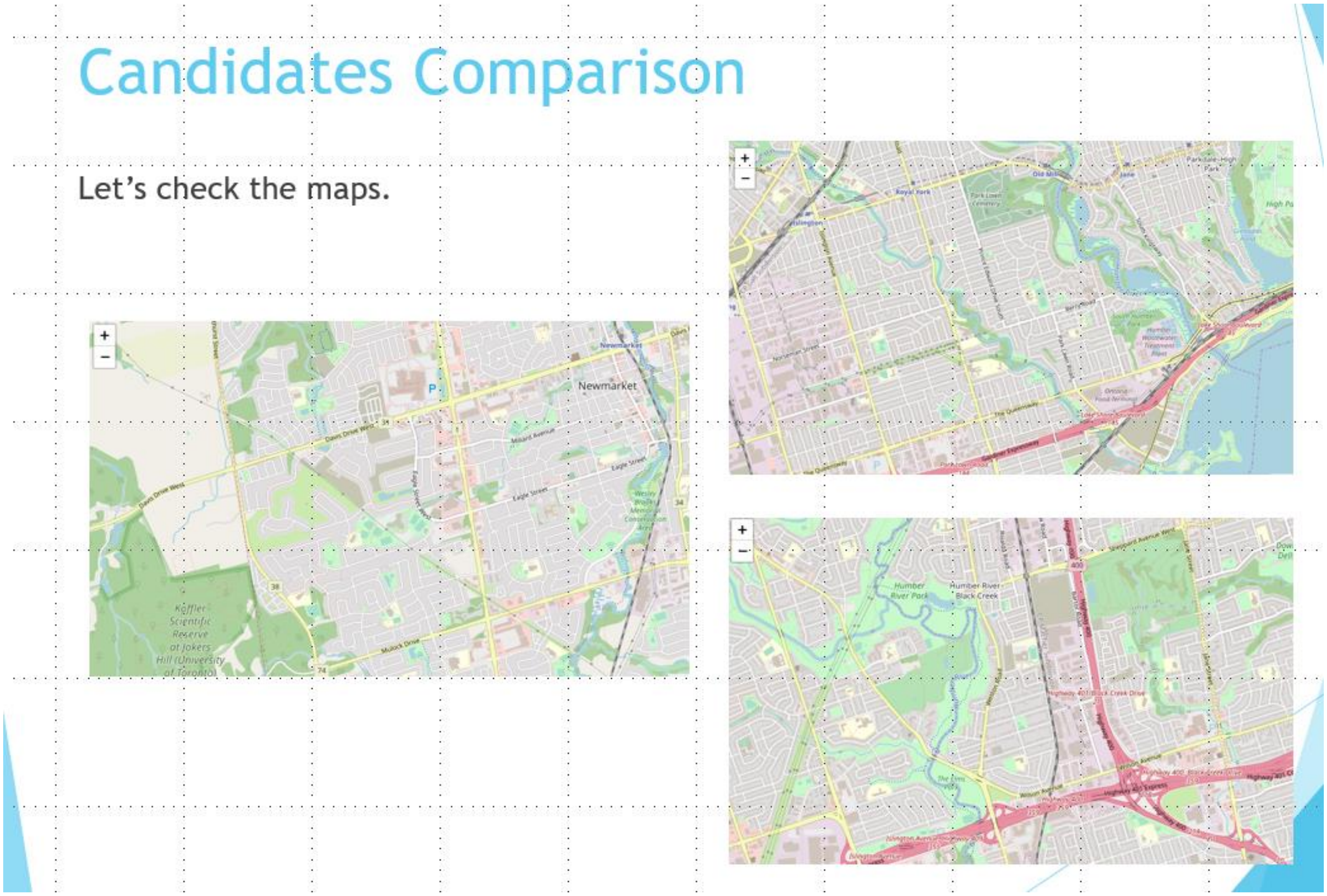
With this, I am able to get even better visualization.



Conclusion

Based on my analysis, I chose a borough in Cluster 1 who has 'Shop' as the '3rd Most Common Venue'.

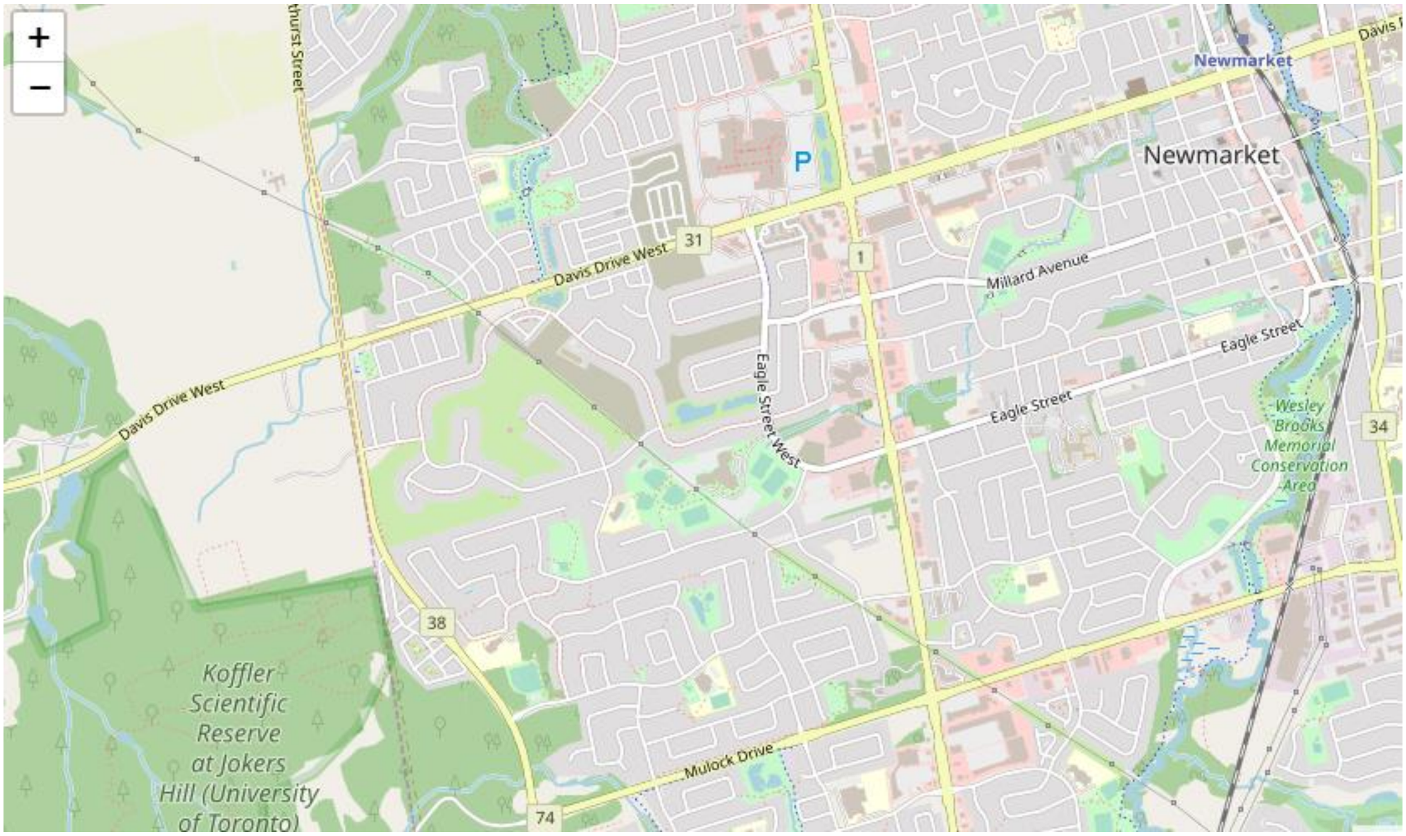
There were 3 candidates who meet these 2 criteria.



Candidates Comparison

Let's check the maps.

After further comparison, I chose 'L3X - Newmarket Southwest'.



Problem solved!

Thanks for Data Science.

References

[1] List of postal codes of Canada: L - Wikipedia

[2] List of postal codes of Canada: M - Wikipedia

[3] Foursquare Developer

[4] Google Map