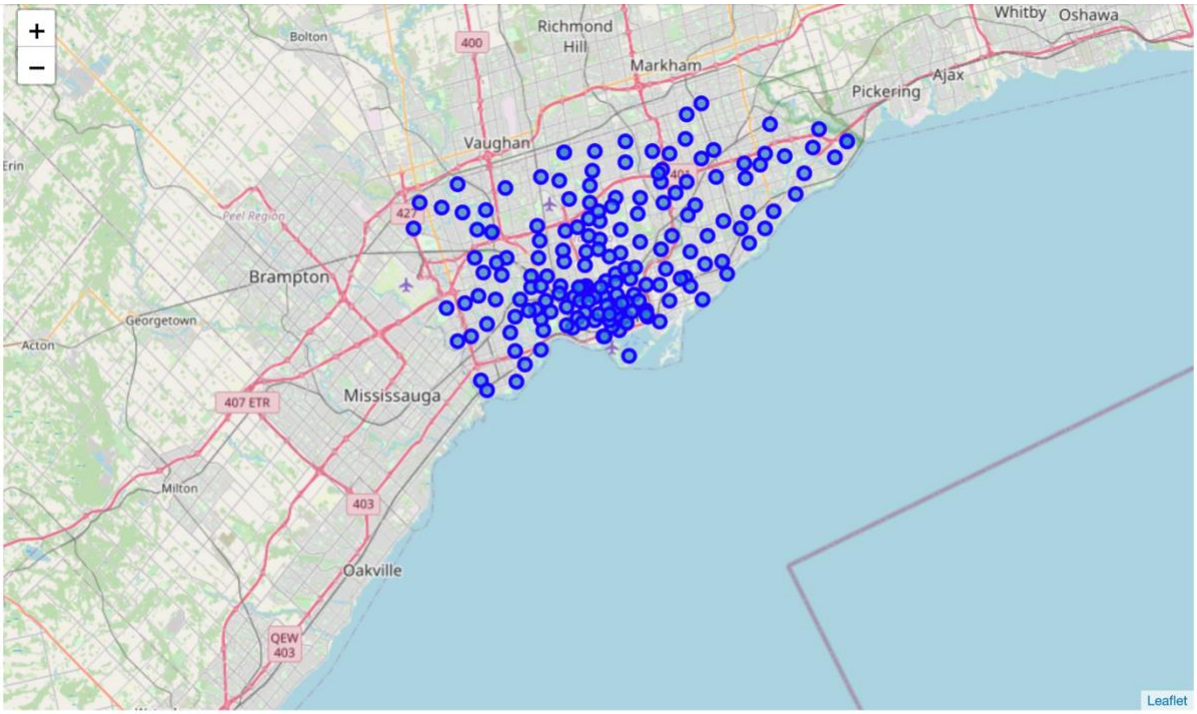
Coursera Capstone - Final Assignment

Anton Möller

Introduction & Business Problem



Background

As the capital city of Canada, Toronto has a flourishing business landscape and consistently attracts new players because of the attractive market and potential opportunities. This also means that the market is very competitive which lowers margins and increases the effort required to be successful. Rash investment decision making will not get you anywhere here, but strategic and well-planned actions stands to make your business and lower the investment risk.

Aims

In this report, we will consider the fictive retail chain **Good Purchase** who are well-established in other regions of Canada but has yet to establish themselves in Toronto. With a well-functioning business model and concept, they have proven themselves as a viable business in other markets but as with all new market establishment, they need their first store opening in Toronto to be a successful one. This would enable them to continue expansion throughout Toronto, replicating the opening strategies from their first store.

To ensure success, Good Purchase has enlisted the help of a data scientist team with the objective to **determine and recommend the most suitable neighbourhood of Toronto to open their first store in.** Along with the engagement letter, Good Purchase has hinted that from their previous operations in other geographies they know that the **two most important factors for retail success are:**

1. Low local competition (=low # of stores),

2. High demand (=high population).

Data

Data description and collection

For this project, we need to collect, investigate and cluster the neighbourhoods of Toronto along with their corresponding demographic data and geodata of existing retail stores.

Throughout this project, we will be collecting the data by scraping (mainly Wikipedia) web pages containing the demographic and neighbourhood segmentations. We will append this dataset using open-source coordinate API:s to collect the geographical coordinates of each neighbourhood. This, in turn, will allow us to utilise the Foursquare API to collect information about existing competing businesses.

One important aspect when collecting data in this manner is to properly clean, check and verify the integrity of the dataset. Hence, in every collection step we will pause and ensure that the newly collected data meets the standards we want, and if not, retrace our steps and recollect missing data.

Ideally, we will end up with a dataset containing:

- Neighbourhood ID
- Neighbourhood Name
- Neighbourhood geographical coordinates
- Neighbourhood population
- Competing Venue Name
- Competing Venue Category
- Competing Venue geographical coordinates

To be able to answer the posed question, we will build our analysis on the **two fundamental pillars** set out by Good Purchase, namely that a good neighbourhood will have low competition and high demand.

Instead of finding the neighbourhood with the lowest competition and highest demand, we will use a clustering technique (e.g. k-means or DBSCAN) to identify a cluster of neighbourhoods that stands out with respect to the potential of opening a store. This will in addition to answer the client's immediate question also give a road map to potential future openings.

Methodology

To acquire the data, we used web scraping from two wikipedia pages, one containing a list of the neighbourhoods of Toronto and the other containing demographic information.

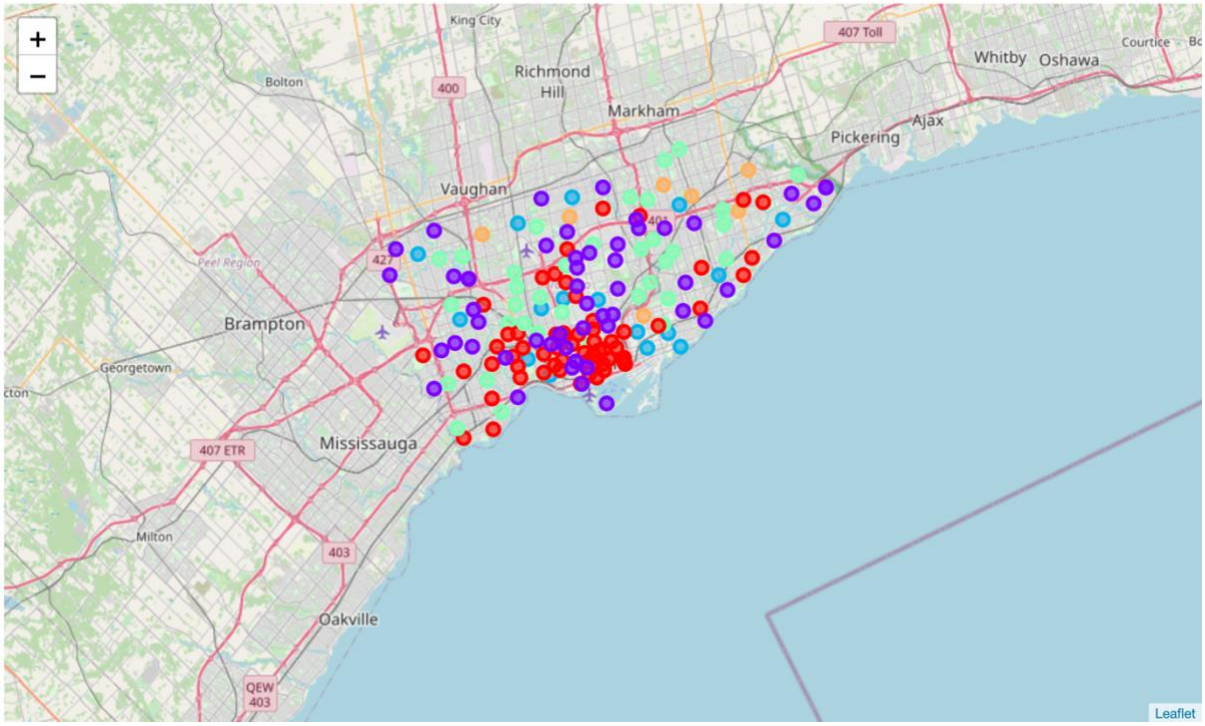
In order to then map the level of competition within each neighbourhood, we then used a Geocoder API to extract the geographical coordinates for each neighbourhood and extended our dataset with them.

Finally, we collected detailed information on the various businesses in each neighbourhood using the Foursquare API. After cleaning and proccessing this dataset we extracted a competition score which represents the level of competing businesses within the market segment that our client Good Purchase operates in.

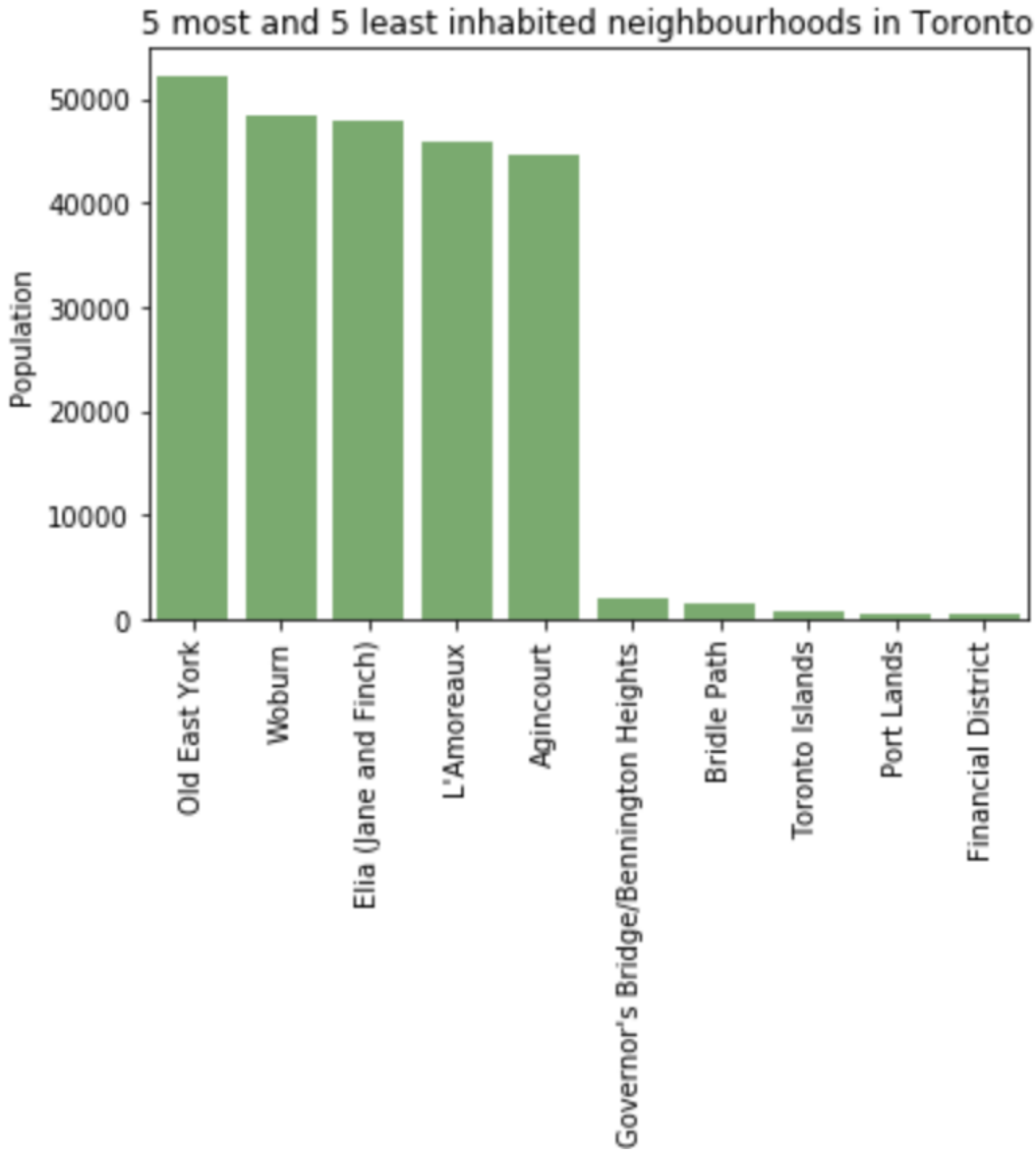
The final step in our analysis was to use k-means clustering to cluster the neighbourhoods into segments with similar competition and population. Our client wants to expand into a high demand - low supply environment, so this analysis allowed us to identify not only one suitable candidate, but an entire set of candidates.

Results

After performing the data cleaning, we applied a k-means clustering algorithm with k=5 clusters, the results of which are shown below. The clustering was performed using a 50/50 weighted population score (higher population = higher score) and a competition score (lower competition = higher score), which allowed our algorithm to determine the best neighbourhood cluster to expand into.



We also investigated the distribution of the population in Toronto neighbourhoods, and as a sample we plotted the neighbourhoods with the largest and smallest populations. The key takeaways from this is that while there are many areas with near 50000 population, there are also very small districts with around 1000 inhabitants.



In addition, the Foursquare API yielded much information about the venues within each neighbourhood. This enabled us to append the existing dataset with information of the level of competition within each neighbourhood. An excerpt from the collected and cleaned data is displayed below, and just from glancing over it - one can immediately identify more and less suitable areas for establishing a new store.

	Neighbourhood	1 Most Common Venue	2 Most Common Venue	3 Most Common Venue	4 Most Common Venue	5 Most Common Venue	6 Most Common Venue	7 Most Common Venue	8 Most Common Venue	9 Most Common Venue	10 Most Common Venue
0	Agincourt	Coffee Shop	Yoga Studio	Electronics Store	Ethiopian Restaurant	Event Service	Event Space	Exhibit	Falafel Restaurant	Farmers Market	Fast Food Restaurant
1	Alderwood	Pizza Place	Pool	Dance Studio	Pub	Pharmacy	Skating Rink	Sandwich Place	Coffee Shop	Gym	Event Space
2	Alexandra Park	Bar	Café	Furniture / Home Store	Yoga Studio	Vegetarian / Vegan Restaurant	French Restaurant	Caribbean Restaurant	Coffee Shop	Gym	Park
3	Allenby	Sushi Restaurant	Coffee Shop	Italian Restaurant	Deli / Bodega	Gym	Café	Liquor Store	Gastropub	Tea Room	Spa
4	Amesbury	Café	Construction & Landscaping	Park	Fast Food Restaurant	Bakery	Flea Market	Flower Shop	Fish Market	Fish & Chips Shop	Filipino Restaurant

Discussion

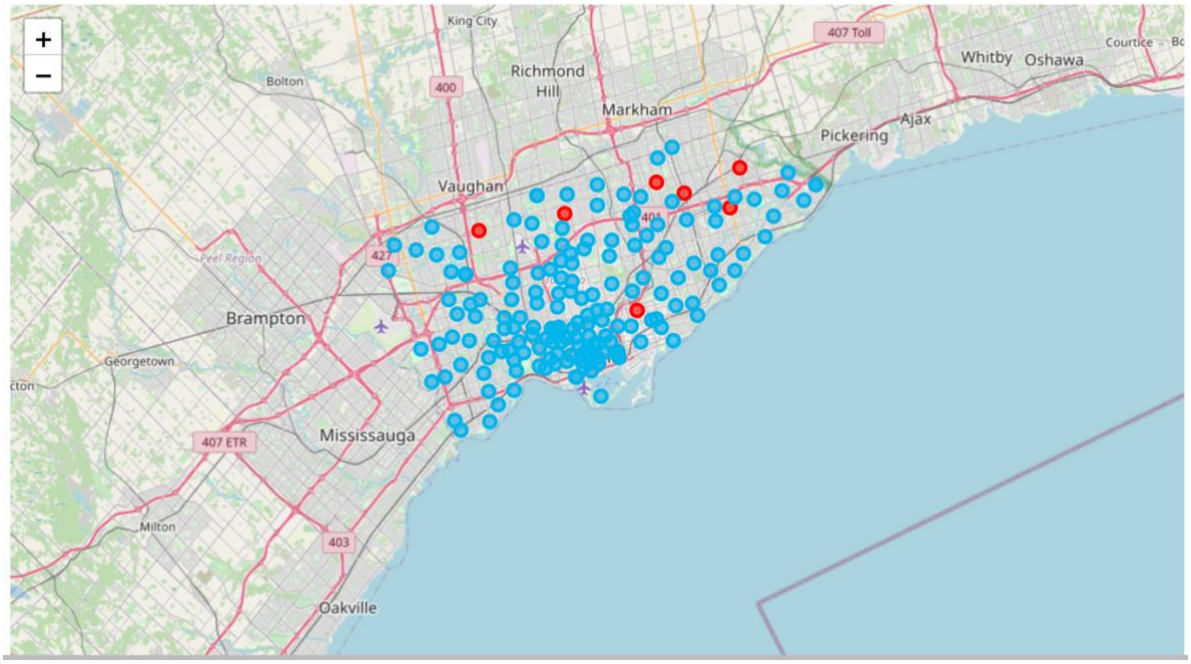
Based on a cluster by cluster investigation, it was determined that cluster number 4 share the desired similarities of being highly populated while having a lower rate of competition than the other clusters. Below is an excerpt from the dataset used by the clustering algorithm, with specific emphasis on the last three columns.

	Neighborhood	Population	Latitude	Longitude	Population (norm)	Competition score	Score
1	Woburn	48507	43.7765	-79.2317	2.007928	1.0	1.503964
0	Old East York	52220	43.692	-79.3378	2.161626	0.7	1.430813
4	Agincourt	44577	43.788	-79.2839	1.845247	1.0	1.422624
2	Elia (Jane and Finch)	48003	43.7573	-79.5177	1.987065	0.8	1.393533
10	Fairbank	34121	43.6979	-79.4511	1.412425	1.0	1.206213

As shown in the image below, cluster 4 (highlighted in red) are scattered throughout Toronto which makes expansion into them suitable as good geographical coverage can be achieved. Specifically, the neighbourhoods in cluster 4 include those of

- Woburn,
- Old East York,
- Agincourt,
- Elia,
- L'Amoureaux,

and others.



Conclusion

In conclusion, we have investigated the suitability of a new retail opening within the neighbourhoods of Toronto. We did this by first collecting data, specifically demographic

and competitor data, after which we utilised k-means clustering to find similar neighbourhoods.

Specifically, we searched for neighbourhoods with high population (=high demand) and low competition (=low supply). Our analysis showed that cluster number 4 had these desired properties, and our **final recommendation to the client therefore is to consider opening within one of Woburn, Old East York, Agincourt, Elia or L'Amoureaux.**

Further expansion after this can preferably take place within the same cluster (i.e. the neighbourhoods listed above) to provide adept geographical coverage, but the main recommendation for this case is to re-evaluate the neighbourhoods when this is up for discussion to account for any recent changes in the market landscape.