Introduction

Background

Although I spent my entire career as an engineer, I pursued my master's degree in Management of Technology while having full time job over in my current workplace. One of the most memorable things that I recall from the lectures that I attended was about the difference between the physical store and internet based retail stores. In a physical stores, the sales are mainly determined by the location; for example, a large pedestrian traffic translates to greater customer visits, and hence the greater sales and revenue.

Problem

Since the main focus of the lecture that I took was about business strategy in e-commerce, there was not in depth analysis done on the correctness of the correlation between the customer traffic versus the store sales. Though I would imagine it does make sense that a retail business of a physical stores would depend upon the customer access. In fact, many apparel retail stores in Japan, opens their stores in a building that is located close to the major stations that can expect to have greater customer traffic.

In order to make this a meaningful research, I would perform an investigation to perform an analysis of the number of stores open vs the customer traffic near subway stations in Toronto, and set the goal of research to determine the possible locations in which a apparel retails store (Clothing Store in Foursquare API category) should open.

The main audience of this research would therefore be someone who is interested in opening an apparel store in the Toronto metropolitan area, alongside the subway and RT to be specific.

Data

Data sources

There are three data that I need to complete my project. The first data that I need would be the traffic data, or to be more specific, ridership data of the subway over in Toronto. Luckily, the city government of Toronto provides the Station Usage data[1] in XLS (MS Excel) format. This Excel spreadsheet provides the listing of all Subway stations that are present in Toronto, and their typical business day platform usage in "To Trains", "From Trains", and "Totals" format.

---

[1] https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/#75d6b4a2-7f29-b0df-f1eb-cc5bc7f53b68

The second source of the data that I used is the GPS locations of all the stations in Toronto. Fortunately, I was able to find the listing of all the stations and their geographical locations[2] in the all four lines (Bloor-Danforth, Yonge-University-Spadina, Sheppard, and Scarborough RT) in the "We Saw a Chicken" blog.

The third and the last data is the number of Apparel stores collected through the Foursquare API. The query used is the venues search query and was ran on foursquare API v2 / version '20180604'. The category ID for the Clothing Store is 4bf58dd8d48988d103951735[3], and searched for an empty string so that every venues that has the category is returned.

Data cleaning

It required a couple of steps of data cleaning to make the data available for analysis. For example, the excel spreadsheet available from the city government of Toronto had unneeded rows and had to specify head=3 in order to ensure that the appropriate headers were set. Together with that, some stations were stops for the multiple subway lines. For example, "St. George" station existed for both line1 and line2. I have aggregated the multi-access stations into one station so that the totals is the sum of the ridership of the two appearances in the spreadsheet.

The subway station GPS locations were split into four files and they needed to be merged. Once files were merged and loaded into dataframe, the data frame was used to add the GPS location to the subway traffic (ridership) dataframe that was loaded above.

Finally, the store information needed to be added to the ridership table. The first problem that I encountered was that the foursquare API only allowed the maximum of 50 search results per API calls. The default radius of 500m returned 50 search results on multiple geo locations, meaning that the radius being too big. I therefore reduced the radius to the point that the maximum number of the stores was at most 50. The optimal radius for getting this result was 130m.

Methodology

Calculations of ridership per store ratio

When finding the correct locations for starting a new Clothing Store, the best metric would be the ratio of ridership per store. This, however would become a problem when there is zero existing stores, as number may not be divided by a zero. Therefore, I would separate a list of stations that has zero stores nearby from the rest of the stations.

---

[2] http://scruss.com/blog/2005/12/14/toronto-subway-station-gps-locations/
[3] https://developer.foursquare.com/docs/resources/categories

I have created two tables, one with the People per Store (or People per Store in the table) column, and the other without the column.

	Latitude	Longitude	Station	Totals	Stores	People per Store
0	43.750054	-79.462343	Downsview	37670	1	37670
1	43.775565	-79.346936	Don Mills	33756	1	33756
2	43.657142	-79.452678	Dundas West	29617	1	29617
3	43.698123	-79.397331	Davisville	25328	1	25328
4	43.725422	-79.401878	Lawrence	24555	1	24555

People per Store vs Station

	Latitude	Longitude	Station	Totals	Stores
0	43.781490	-79.415673	Finch	100819	0
1	43.638020	-79.536388	Kipling	52925	0
2	43.645950	-79.523948	Islington	37412	0
3	43.662663	-79.426157	Ossington	31614	0
4	43.660665	-79.435956	Dufferin	29937	0

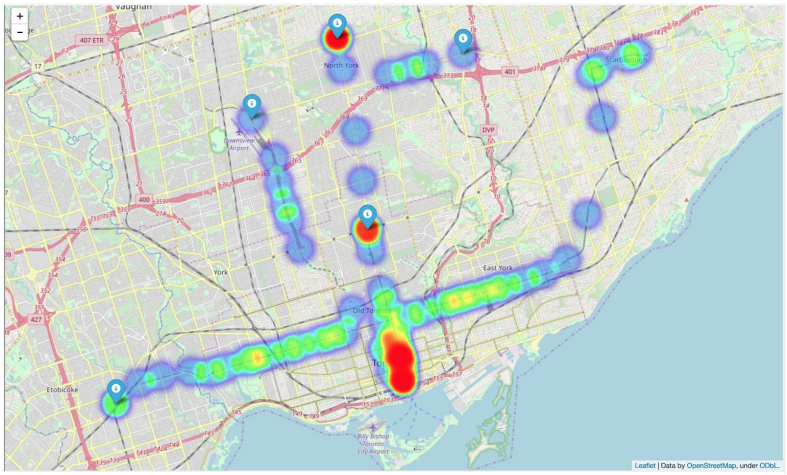
Number of Ridership vs Station

Looking at the table, the Downsview has the largest number of People per Store, 37,670, and the Finch stations has 100,819 ridership, and zero existing stores. Even the second store in the zero store station, Kipling, has 52,925 ridership. By opening a store over in stations with existing stores, the store will be splitting the ridership with the existing store, so the effective people per store would decrease so the “People per Store” figure is actually half of what it currently represent if the store is to open, so the Downsview station would have 18,835 people per store, which is lower than the 5th station in the zero store stations.

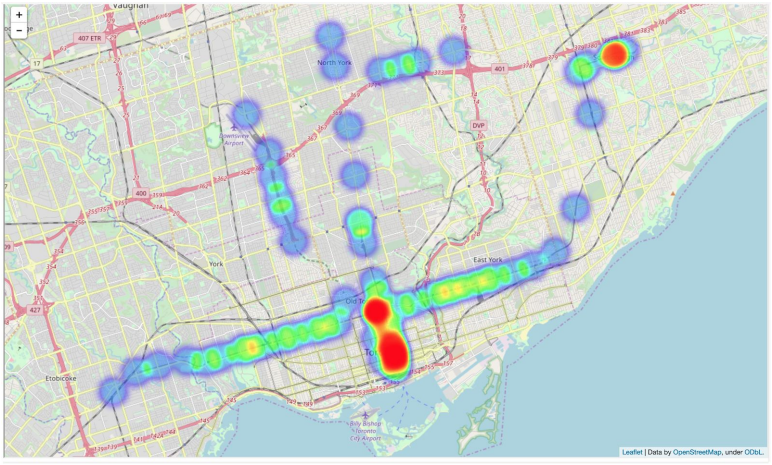
Heatmap comparison

Two heatmaps was generated based on the ridership and the store concentration. What you notice at the first look is that there is severe concentration of the Ridership and the Store concentration at the lower center of the map. This is the downtown area of Toronto, and is obvious that the two overlaps each other.

In contrary to the popular belief that stores are concentrated on the stations with larger ridership, the heatmap uncovered the fact that there were some exceptions to this rule.



Heatmap of the ridership / station location



Heatmap of store concentration / station location

To support the findings from the “Calculations of ridership per store ratio” section, Finch, (Red area at the top of “Heatmap of the ridership / station location”) and Kipling (Green area at the far west of “Heatmap of the ridership / station location”)

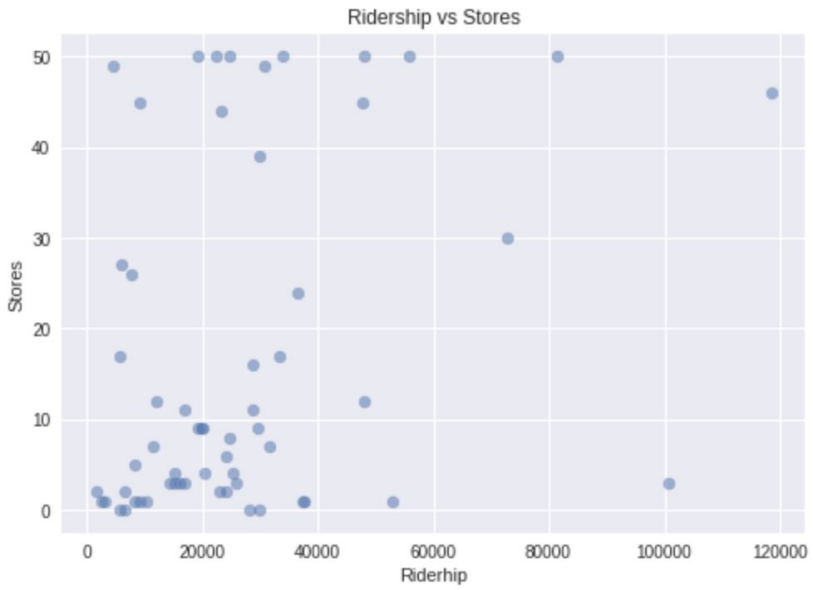
What stands out in the heatmap is Eglinton station(Red area at the center of “Heatmap of the ridership / station location”). Eglinton has the ridership of 72,746, which makes it a strong candidate for the location for store opening, but it also has 30 stores near by.

	Latitude	Longitude	Station	Totals	Stores
45	43.706646	-79.399158	Eglinton	72746	30

Eglinton : Ridership and number of stores

Relationship between Ridership and the number of Stores

The previous sections uncovered the fact that there are some exceptions to the belief that there exists correlations between store concentration and the ridership. It would make sense to generate a scatter plot to verify that ridership and the store concentration has no clear correlation.



Ridership vs Stores

Results

Not surprisingly, the downtown area had large number of ridership and the store concentration. While opening an apparel shop in this area may attract the potential customers, the store would likely to face a fierce competition among nearby retail stores.

Looking further into the heatmap generated, it became apparent that some ridership intensive stations had less competing stores. Namely, Finch and Kipling. Both Finch and Eglinton has no apparel stores in the 130m radius meaning there will be no competition when the store newly opens, and yet the business day ridership is 100,819 for Finch and 52,925 for Eglinton.

What stands out when comparing the ridership heatmap and the store count heatmap is that Eglinton had large number of ridership and yet the store count is seemingly moderate. Looking at the store count for the 500m radius, there are 30 stores.

Conclusion

The conclusion is fairly simple, I would strongly advise that the new opening store to start its business over in the Finch station. There are 100,819 ridership in the Finch station for typical business days, and yet there are zero stores on 130m radius. Another possible candidate would be Kipling station. Kipling has 52,925 ridership on a business day, and yet, like Finch, has no competing stores nearby.

Future Directions

While this report provides insights into the plausible location in which the store should open based on the customer access and the number of competing stores, there had not been any measurements on the cost. It is most likely that the cost profit analysis would be a necessity when opening a store, and it is likely that the rent accounts for a large proportion of the cost. Therefore, I suggest that the future research to review how much the rent is, say in dollar per square meter, and then compare it with the number of ridership. For example, suppose that it would cost $5,000 per month to rent out 100 square meters of a shopping building, and the daily traffic is 5000, then the cost ($ per person) would be $1 per month. Compare this number among the potential store locations and the cheapest location would be the best option in terms of cost performance.