Battle of the Businesses in a Neighborhood

Applied Data Science Capstone Project

By Yashodhara Thakur

Why is it important to analyze different factors before starting a business

- New start-ups pop up every day all around the world, each of them hoping to get acquired by a larger company or make it big in their own right
- There are thousands which fall into obscurity
- The following picture shows the reason for which the startups fail and the top reason is “No market need”.
- To avoid this analysis is necessary

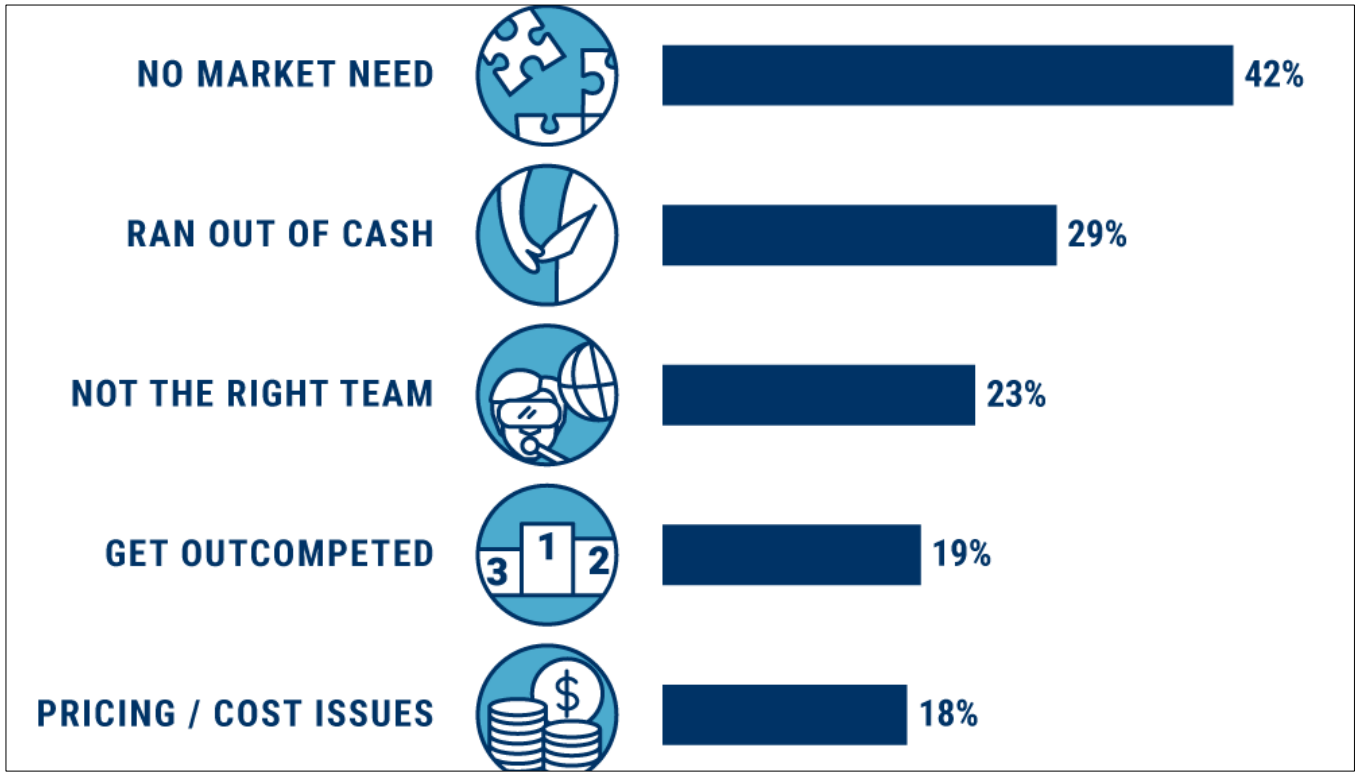


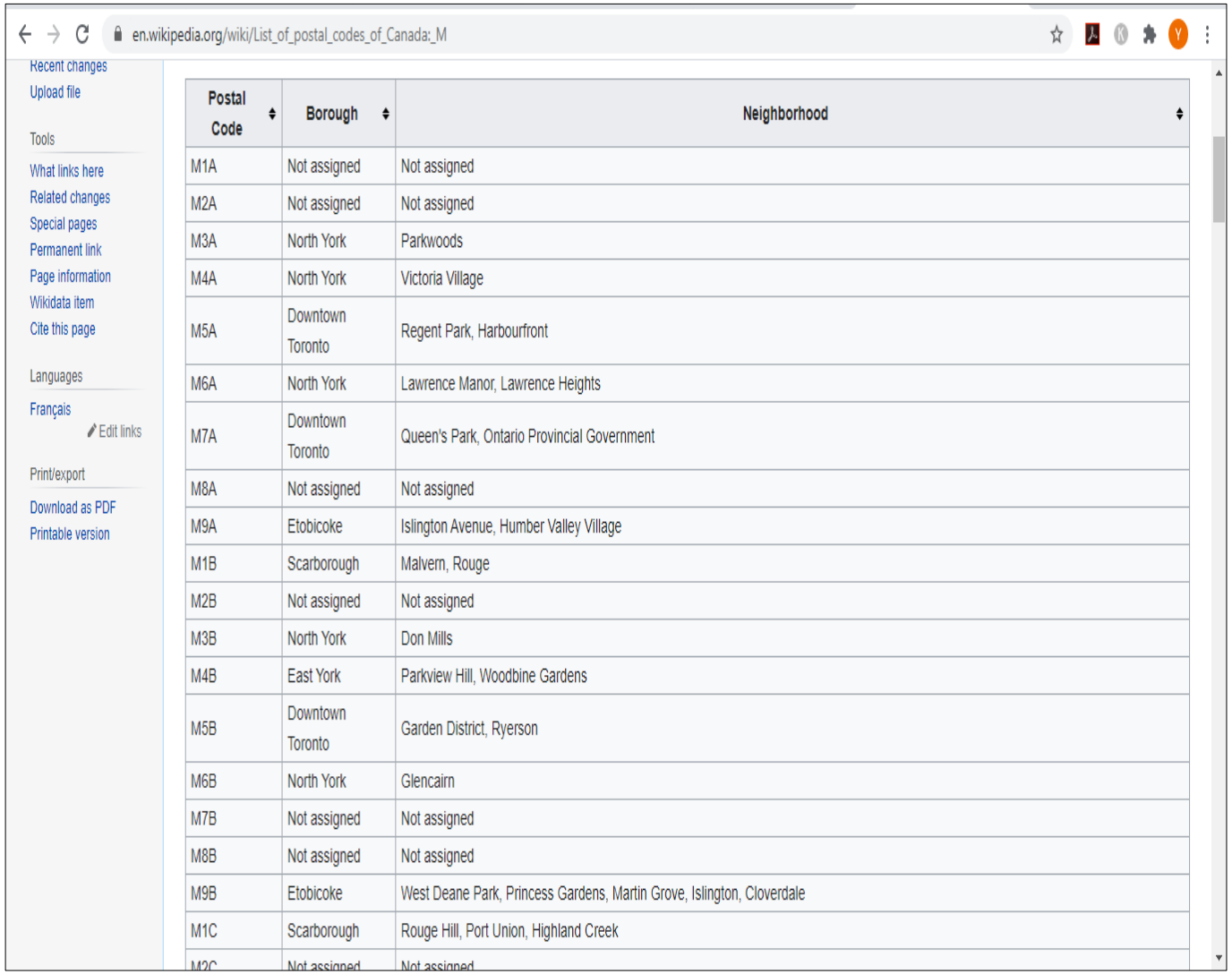
Image source : https://www.cbinsights.com/research/startup-failure-reasons-top/

Introduction to the project

- Building a model to determine best local business to start in a popular Neighborhood
- In addition to this, model also provides with most competitive and least competitive business
- In this project I am going to explore top -3 local businesses in the city of Toronto (as an example).
- It will help the business owners to analyse which location is best for which type of business
- The business owners can select whether they want to start a business which has huge competition because it is popular business or they want to start a business which has less competition and gain competitive advantage

Data Acquisition

- The data for “Toronto” was found on Wikipedia and the table was scraped
- Also the “Geospatial_Coordinates” data was used which was obtained during the course period. These 2 datasets were used in the model.



Postal Code	Borough	Neighborhood
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
M9A	Etobicoke	Islington Avenue, Humber Valley Village
M1B	Scarborough	Malvern, Rouge
M2B	Not assigned	Not assigned
M3B	North York	Don Mills
M4B	East York	Parkview Hill, Woodbine Gardens
M5B	Downtown Toronto	Garden District, Ryerson
M6B	North York	Glencairn
M7B	Not assigned	Not assigned
M8B	Not assigned	Not assigned
M9B	Etobicoke	West Deane Park, Princess Gardens, Martin Grove, Islington, Cloverdale
M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
M2C	Not assigned	Not assigned

Data source : https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Data Cleaning

- Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values in the Toronto dataset. I decided to drop the values where Borough were ‘Not assigned’ and mask the Neighbourhood where values were ‘Not assigned’.

- Second, multiple entries existed for similar Postal Code with different neighbourhoods. This cause their data to represent multiple samples with incomplete data. I wrote script to extract the unique Postal Code, I grouped the data according to the ‘Postal Code’ and ‘Borough’.

- Now I wanted to merge the Toronto dataset with the Geospatial_Coordinates. To do this I merged them on the basis of ‘Postal Code’.

After cleaning and merging the data

```
In [7]: data.rename(columns = {'PostalCode':'Postal Code'}, inplace = True)
        df = pd.merge(df,data, how ='inner', on ='Postal Code')
        df.head()
```

Out[7]:

	Postal Code	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

Et Voilà, we have our dataset ready!!!

Data Analysis

- During this analysis I found out that there were 10 boroughs and 103 neighbourhoods
- Further analysing the data I found that “North York” was the top borough of all because it had the maximum neighbourhoods than all.
- Then I created a dataset which consist of only “North York” data as the below

The unique Boroughs with their neighborhood counts are shown above
Let us analyse the top borough that is "North York"

In [13]:
```
ny_data = df[df['Borough'] == 'North York'].reset_index(drop=True)
ny_data
```

Out[13]:

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
3	M3B	North York	Don Mills	43.745906	-79.352188
4	M6B	North York	Glencairn	43.709577	-79.445073
5	M3C	North York	Don Mills	43.725900	-79.340923
6	M2H	North York	Hillcrest Village	43.803762	-79.363452
7	M3H	North York	Bathurst Manor, Wilson Heights, Downsview North	43.754328	-79.442259
8	M2J	North York	Fairview, Henry Farm, Oriole	43.778517	-79.346556
9	M3J	North York	Northwood Park, York University	43.767980	-79.487262
10	M2K	North York	Bayview Village	43.786947	-79.385975
11	M3K	North York	Downsview	43.737473	-79.464763

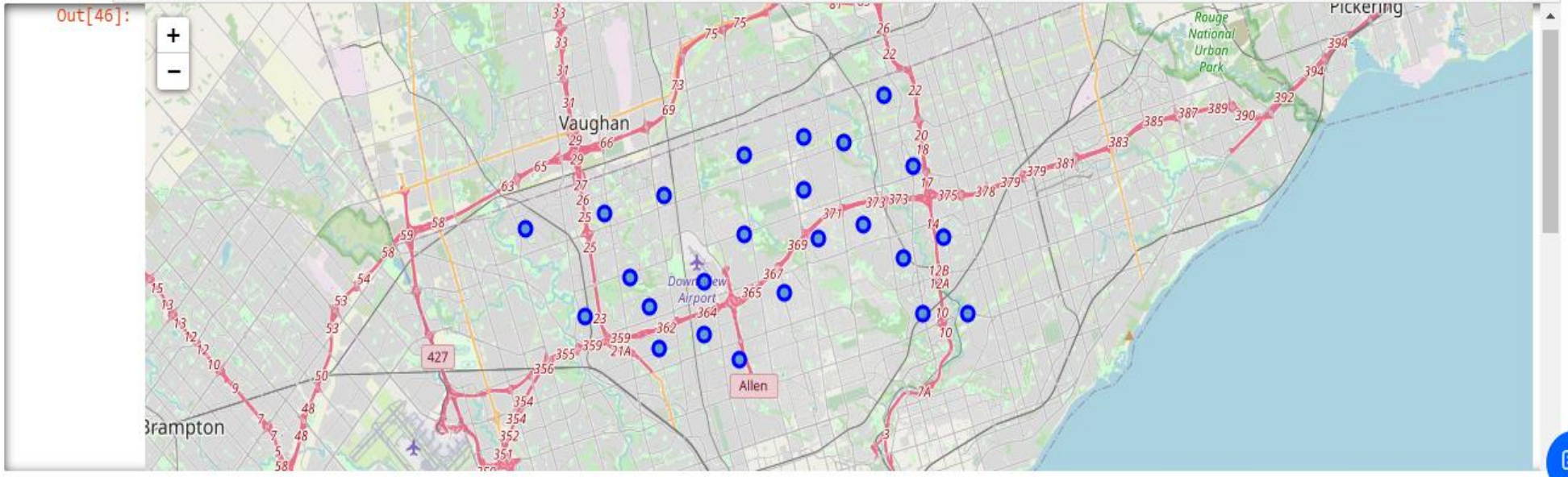
North York map

- I also created a clusterd map of “North York” for visual representation.

```
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_ny)

map_ny
```

Out[46]:



Further analysis

- Now we wanted to find the best neighbourhoods in North York.
- To do that I used foursquare. I used foursquare to get the venues for all the neighbourhoods. Once this was done I had to get the top 2 neighbourhoods with maximum venues
- After computing the top 2 neighbourhoods were: [Fairview, Henry Farm, Oriole] and [Willowdale and Willowdale East].
- Now we had to find the popular businesses in these neighbourhoods.

To find popular businesses in the neighborhood

- To do that we first made 2 dataframes which consist the information of these 2 neighbourhoods and then grouped them according to the “Venue Category”, which looked something like this

```
In [31]: top1ny.groupby('Venue Category').count()
```

Out[31]:

Venue Category	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
American Restaurant	1	1	1	1	1	1
Asian Restaurant	1	1	1	1	1	1
Bakery	2	2	2	2	2	2
Bank	2	2	2	2	2	2
Bar	1	1	1	1	1	1
Baseball Field	1	1	1	1	1	1
Boutique	1	1	1	1	1	1
Burger Joint	1	1	1	1	1	1
Burrito Place	1	1	1	1	1	1
Bus Station	1	1	1	1	1	1
Chinese Restaurant	1	1	1	1	1	1
Chocolate Shop	1	1	1	1	1	1
Clothing Store	8	8	8	8	8	8

Results

- Now we analysed the data to find the Top 3 most common business categories .The example of the output is shown below

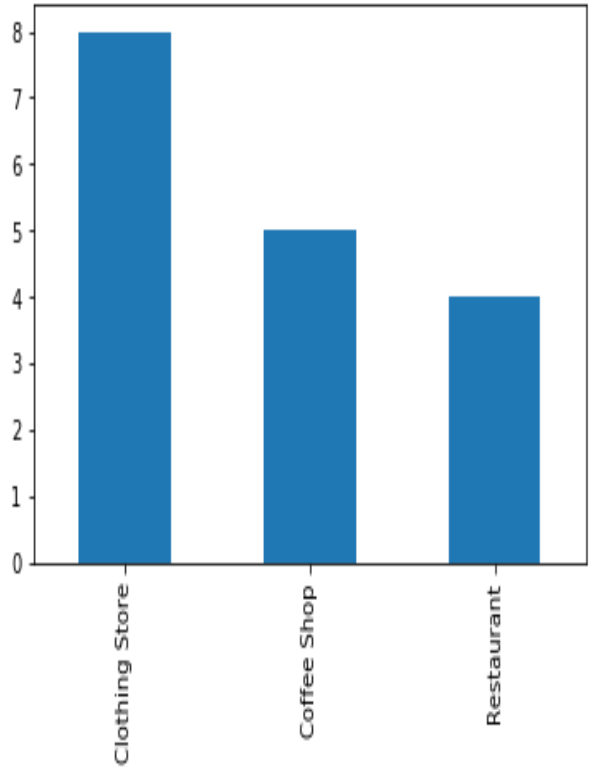
Top 3 most common business categories

```
In [33]: b1 = top1ny['Venue Category'].value_counts()
         b1=b1.head(3)
         b1

Out[33]: Clothing Store    8
         Coffee Shop       5
         Restaurant        4
         Name: Venue Category, dtype: int64
```

```
In [35]: b1.plot(kind = 'bar')
```

```
Out[35]: <matplotlib.axes._subplots.AxesSubplot at 0x7f6ae42f5748>
```



These are the common businesses in the area

Let's see least competitive businesses in the area

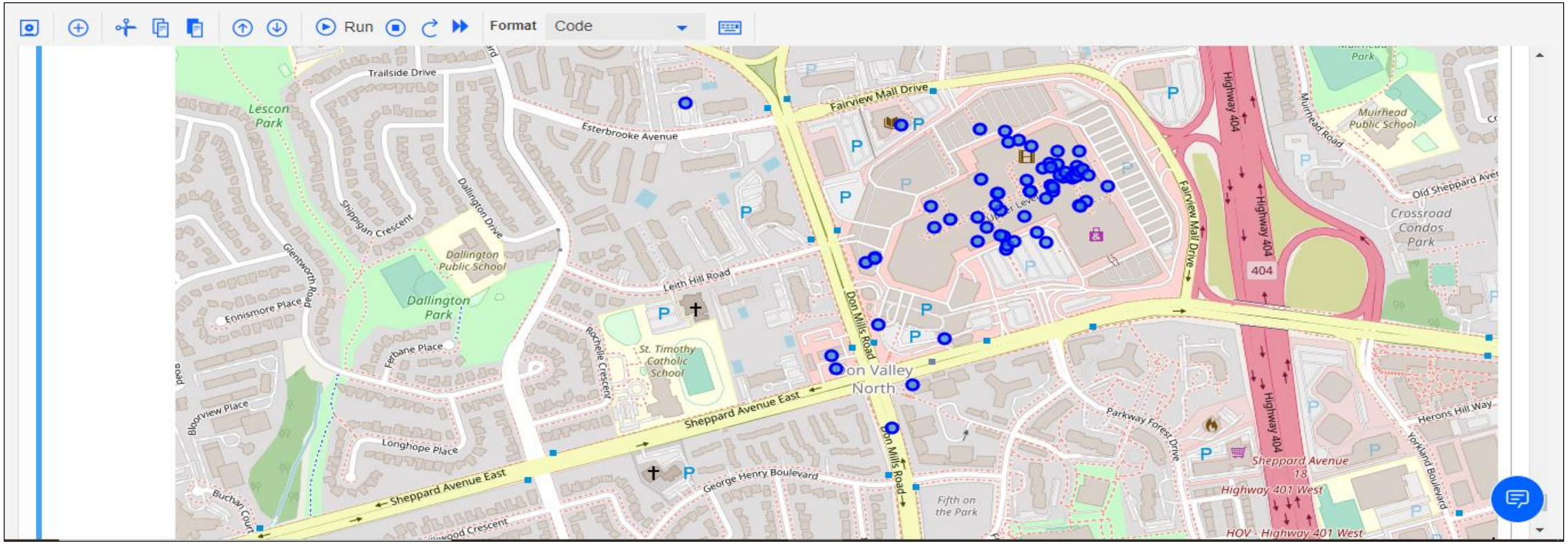
In [36]:
```
blueocean1 = top1ny['Venue Category'].value_counts()
blueocean1=blueocean1.tail(3)
blueocean1
```

Out[36]:
```
Luggage Store    1
Bar              1
Theater          1
Name: Venue Category, dtype: int64
```

The least competitve businesses are :

1. Luggage store
2. Bar
3. Theater

- Now at last we append both the datasets as one to find the best neighbourhood cluster for business and plot it on the map as shown below



Conclusion and Further Direction

- In this study, I analysed the different neighbourhoods in the city of Toronto.
- Also analysed the popular boroughs and in those boroughs we further analysed the neighbourhoods
- We further analysed these neighbourhoods to find the best neighbourhood to start a business.
- . Not only that we also found out the most competitive and the least competitive local business categories in these neighbourhoods.
- This model will be very helpful for people who want to start their own business and want to do competitive analysis
- I have prepared this model on the basis of neighbourhoods / boroughs for a particular city. In future this model can be used to find the best city to start a business as well as a best country.
- Also further competitive analysis is possible using this model by adding and analysing the current businesses information