# Business Plan for Opening a New Club in Toronto

I am a data analyst working on research, data collection, collation and analysis, problem solution scenario development, etc. I have a project on business settlement in Toronto, collecting and analyzing data and coming up with recommendations on the problem defined.

## Problem statement

The owner of popular clubs all over Europe is aiming at opening new branches in other continents, particularly in Toronto, Canada. I am asked to make an analysis and tell him where would it be best to open the clubs. I am also asked to be as specific as possible and name not only boroughs but also neighborhoods which might work the best for club opening. The club owner also mentioned, that the strategy of the company is to settle in areas where the competition is lower, but it's still accessible from more overloaded parts of the city.

## Research summary

I am starting with the analysis of the problem. Clubs are mainly popular with the youth. Thus I am making a small research on neighborhoods/boroughs of Toronto which have the highest concentration of young people. A number of articles state, that most of the top neighborhoods (such as Yonge and Eglinton, Distillery District, Liberty Village, etc.) for the youth are located in East York. An example of such articles is the following: https://www.torontorentals.com/blog/toronto-neighbourhoods-for-young-professionals.

But as my client mentioned strategy-wise they are aiming at less loaded parts of the city, thus the nest step was to find the borough with lower competition on clubs but at the same time accessible from East-York and surroundings. It was obvious from the research that a potential borough would be Scarborough, which is just 20-minute drive away from East York.

## Data

Having researched the problem and brought up specifications needed for the analysis, we come to the point of the data. Initially having the location of Toronto, we narrowed it up to Scarborough based on the research. Thus at this point we need to come up with the specific locations (namely neighborhoods) in Scarborough, Toronto where opening a club would be more efficient. And as clubs are mostly visited by young people we need to discover neighborhoods with active youth.

For that we will be concentrated on exploring the most common venues in different neighborhoods of Toronto. Thus we need the following steps of data collection, collation and analysis:

1. All neighborhoods in Toronto with their longitudes and latitudes. We have data on the coordinates of Postal Codes, but as we will need to make the research on the neighborhoods and boroughs, we are to create the necessary database. Thus we will:

a. First the list of the boroughs, neighborhoods and their Postal Codes will be extracted from the following page:
   https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
b. The data will be collated with the geospatial coordinates of the Postal Codes.
2. Based on the database, we will collect data on most common venues with the help of the Foursquare location data.
3. Finally we will use clustering to understand the best neighborhoods of Scarborough for opening a club.

## Data collection and collation

As mentioned in the section above, for the analysis we need to have a database on each of the Toronto neighborhoods with their longitudes and latitudes. We have a database on the coordinates of different neighborhoods, but the thing is that they are mentioned with the postal codes only.

| | PostalCode | Latitude | Longitude |
|---|---|---|---|
| 0 | M1B | 43.806686 | -79.194353 |
| 1 | M1C | 43.784535 | -79.160497 |
| 2 | M1E | 43.763573 | -79.188711 |
| 3 | M1G | 43.770992 | -79.216917 |
| 4 | M1H | 43.773136 | -79.239476 |
| 5 | M1J | 43.744734 | -79.239476 |
| 6 | M1K | 43.727929 | -79.262029 |
| 7 | M1L | 43.711112 | -79.284577 |
| 8 | M1M | 43.716316 | -79.239476 |
| 9 | M1N | 43.692657 | -79.264848 |
| 10 | M1P | 43.757410 | -79.273304 |
| 11 | M1R | 43.750072 | -79.295849 |
| 12 | M1S | 43.794200 | -79.262029 |
| 13 | M1T | 43.781638 | -79.304302 |
| 14 | M1V | 43.815252 | -79.284577 |

Thus, we are going to need to add to this data corresponding information on neighborhoods and boroughs. On the Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) we get a list of all the boroughs and neighborhoods of Toronto, which also have postal codes mentioned. We are going to use the postal codes as references for data collation of two different datasets.

Both databases need to be cleaned and prepared. We start from the wikipedia list.

```python
#reading data from the newly created file
dtable=pd.read_csv('list.csv')
#renaming the colomns
dtable=dtable.rename(columns={"Postcode": "PostalCode","Borough": "Borough","Neighbourhood\n": "Neighborhood"})
#removing '\n's from the Neighborhood column values
dtable['Neighborhood'] = dtable['Neighborhood'].str[:-1]
#Filtering 'Not assigned' Boroughs out
dtable=dtable[dtable.Borough != 'Not assigned']
#grouping the values of Neighborhood by PostalCode and setting ',' as a separator
dtable = dtable.groupby(['PostalCode','Borough']).agg(lambda x: ", ".join(x))
#reseting the indexes
dtable.reset_index(level=['PostalCode','Borough'], inplace=True)
#Filling the 'Not assigned' Neighborhood cells with the corresponding values of Borough
dtable.loc[dtable['Neighborhood'] =='Not assigned', 'Neighborhood'] = dtable['Borough']
#cheching the shape
dtable.shape
```

(103, 3)

The first step will be removing the rows with boroughs valued as "Not assigned". It is important for us to work with neighbors at Scarborough borough only, thus missing information about the borough makes the data on the specific postal code useless, or even harmful for the analysis.

Secondly, we see that we have repeating postal codes due to the fact that different boroughs refer to the same postal code. Thus we will make sure all the relevant data is grouped by postal codes and the neighborhoods of the later are attached to each other.

And lastly, we also notice a "Not assigned" value in the neighborhood feature. Here we will make sure that all such cases get the value of the corresponding borough.

So now we have a proper database with all postal codes, neighborhoods and boroughs of Toronto.

| | PostalCode | Borough | Neighborhood |
|---|---|---|---|
| 0 | M1B | Scarborough | Rouge, Malvern |
| 1 | M1C | Scarborough | Highland Creek, Rouge Hill, Port Union |
| 2 | M1E | Scarborough | Guildwood, Morningside, West Hill |
| 3 | M1G | Scarborough | Woburn |
| 4 | M1H | Scarborough | Cedarbrae |
| 5 | M1J | Scarborough | Scarborough Village |
| 6 | M1K | Scarborough | East Birchmount Park, lonview, Kennedy Park |
| 7 | M1L | Scarborough | Clairlea, Golden Mile, Oakridge |
| 8 | M1M | Scarborough | Cliffcrest, Cliffside, Scarborough Village West |
| 9 | M1N | Scarborough | Birch Cliff, Cliffside West |

The second dataset also needs to be worked on.

Although it seems from the first sight that it looks quite good, the number of observations is also matching the first dataset, we notice, that the names of our referral feature, namely postal codes, differs from the was it is mentioned in the first dataset. Thus, we rename is to exactly the way it is named in the later.

```
[48]:  #reading the file into a df
       df_geo=pd.read_csv('http://cocl.us/Geospatial_data')
       #changing the column name on the one in the inital table
       df_geo=df_geo.rename(columns={"Postal Code": "PostalCode"})
       #merging two tables with respect to their Postal Codes
       dtable_geo= pd.merge(dtable, df_geo, on='PostalCode')
       dtable_geo.shape

[48]:  (103, 5)
```

Afterwards we make the last step and merge the two datasets and receive the final one, which includes 103 observations with 5 features each.

```
[15]: dtable_geo.head(10)
```

| | PostalCode | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | M1B | Scarborough | Rouge, Malvern | 43.806686 | -79.194353 |
| 1 | M1C | Scarborough | Highland Creek, Rouge Hill, Port Union | 43.784535 | -79.160497 |
| 2 | M1E | Scarborough | Guildwood, Morningside, West Hill | 43.763573 | -79.188711 |
| 3 | M1G | Scarborough | Woburn | 43.770992 | -79.216917 |
| 4 | M1H | Scarborough | Cedarbrae | 43.773136 | -79.239476 |
| 5 | M1J | Scarborough | Scarborough Village | 43.744734 | -79.239476 |
| 6 | M1K | Scarborough | East Birchmount Park, Ionview, Kennedy Park | 43.727929 | -79.262029 |
| 7 | M1L | Scarborough | Clairlea, Golden Mile, Oakridge | 43.711112 | -79.284577 |
| 8 | M1M | Scarborough | Cliffcrest, Cliffside, Scarborough Village West | 43.716316 | -79.239476 |
| 9 | M1N | Scarborough | Birch Cliff, Cliffside West | 43.692657 | -79.264848 |

Finally, we are importing from the Foursquare database all relevant data on venues in the surroundings of each of the neighborhoods. We are taking top 100 venues only in the radius of 500 meters.

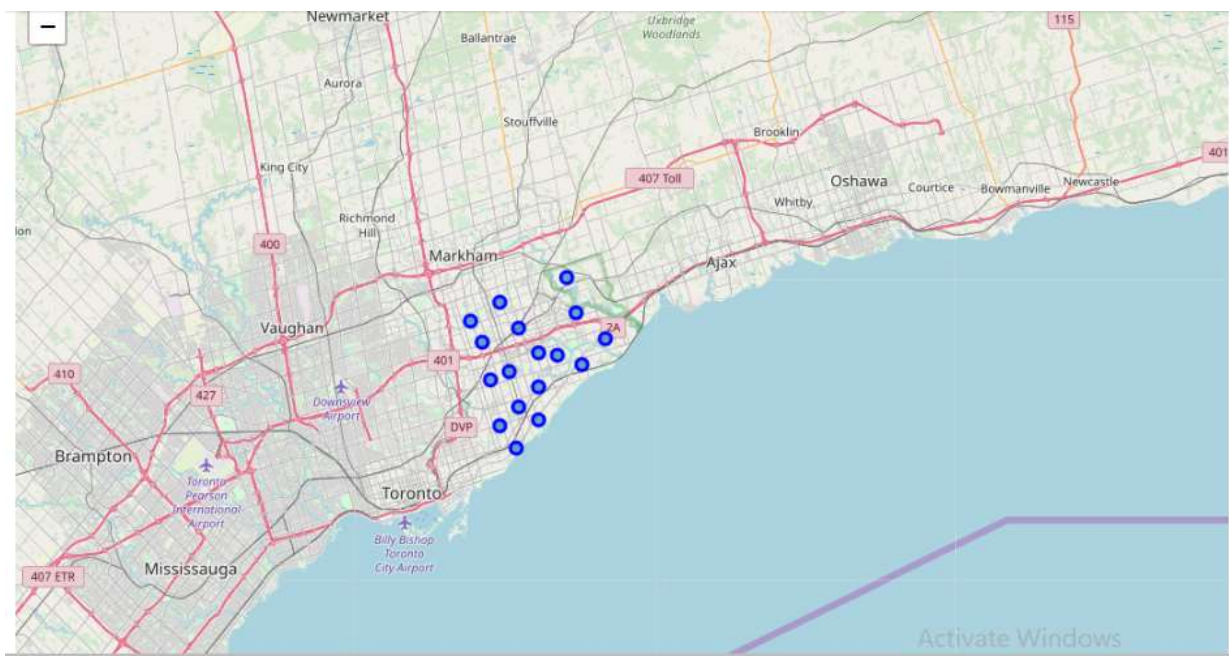| Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|
| Agincourt | 4 | 4 | 4 | 4 | 4 | 4 |
| Agincourt North, L'Amoreaux East, Milliken, Steeles East | 3 | 3 | 3 | 3 | 3 | 3 |
| Birch Cliff, Cliffside West | 4 | 4 | 4 | 4 | 4 | 4 |
| Cedarbrae | 8 | 8 | 8 | 8 | 8 | 8 |
| Clairlea, Golden Mile, Oakridge | 10 | 10 | 10 | 10 | 10 | 10 |
| Clarks Corners, Sullivan, Tam O'Shanter | 10 | 10 | 10 | 10 | 10 | 10 |
| Cliffcrest, Cliffside, Scarborough Village West | 3 | 3 | 3 | 3 | 3 | 3 |
| Dorset Park, Scarborough Town Centre, Wexford Heights | 6 | 6 | 6 | 6 | 6 | 6 |
| East Birchmount Park, Ionview, Kennedy Park | 4 | 4 | 4 | 4 | 4 | 4 |
| Guildwood, Morningside, West Hill | 9 | 9 | 9 | 9 | 9 | 9 |
| Highland Creek, Rouge Hill, Port Union | 2 | 2 | 2 | 2 | 2 | 2 |
| L'Amoreaux West | 12 | 12 | 12 | 12 | 12 | 12 |
| Maryvale, Wexford | 7 | 7 | 7 | 7 | 7 | 7 |
| Rouge, Malvern | 2 | 2 | 2 | 2 | 2 | 2 |
| Scarborough Village | 1 | 1 | 1 | 1 | 1 | 1 |
| Woburn | 4 | 4 | 4 | 4 | 4 | 4 |

Activate Windows

## Methodology

We are starting with exploratory data analysis. To have the overall image of what we are working with we first are mapping it all down.
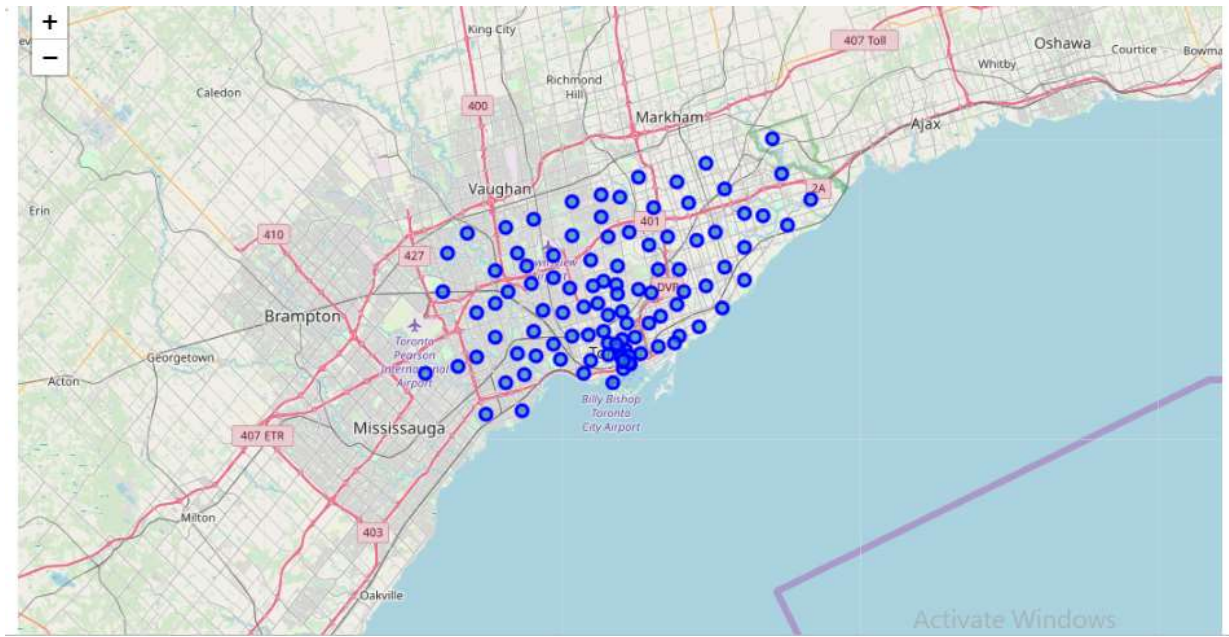


The map shows all the neighborhoods in Toronto, Canada.

As we want to concentrate of Scarborough only, we are leaving the neighborhoods in Scarborough only.



## Methodology

If we observe the top 100 venues located within 500meters of the neighborhoods mentioned in the map we will receive 89 venues referring to 53 different categories, such as restaurants, banks, bakeries, bars, etc.

| Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|
| Agincourt | 4 | 4 | 4 | 4 | 4 | 4 |
| Agincourt North, L'Amoreaux East, Milliken, Steeles East | 3 | 3 | 3 | 3 | 3 | 3 |
| Birch Cliff, Cliffside West | 4 | 4 | 4 | 4 | 4 | 4 |
| Cedarbrae | 8 | 8 | 8 | 8 | 8 | 8 |
| Clairlea, Golden Mile, Oakridge | 10 | 10 | 10 | 10 | 10 | 10 |
| Clarks Corners, Sullivan, Tam O'Shanter | 10 | 10 | 10 | 10 | 10 | 10 |
| Cliffcrest, Cliffside, Scarborough Village West | 3 | 3 | 3 | 3 | 3 | 3 |
| Dorset Park, Scarborough Town Centre, Wexford Heights | 6 | 6 | 6 | 6 | 6 | 6 |
| East Birchmount Park, Ionview, Kennedy Park | 4 | 4 | 4 | 4 | 4 | 4 |
| Guildwood, Morningside, West Hill | 9 | 9 | 9 | 9 | 9 | 9 |
| Highland Creek, Rouge Hill, Port Union | 2 | 2 | 2 | 2 | 2 | 2 |
| L'Amoreaux West | 12 | 12 | 12 | 12 | 12 | 12 |
| Maryvale, Wexford | 7 | 7 | 7 | 7 | 7 | 7 |
| Rouge, Malvern | 2 | 2 | 2 | 2 | 2 | 2 |
| Scarborough Village | 1 | 1 | 1 | 1 | 1 | 1 |
| Woburn | 4 | 4 | 4 | 4 | 4 | 4 |

```
#now let's see how many unique ones are here
print('There are {} uniques categories.'.format(len(scarborough_venues['Venue Category'].unique())))
```

There are 53 uniques categories.

What we want to do is to understand which are the most popular venues in each of the neighborhoods. For example, for Agincourt we see that Lounge, Breakfast Spot, Sandwich Place and Chinese Restaurant are the most popular venue types/categories, which are all equally popular (have same occurrence frequencies).

```
----Agincourt----
                venue  freq
0              Lounge  0.25
1      Breakfast Spot  0.25
2       Sandwich Place  0.25
3   Chinese Restaurant  0.25
4  American Restaurant  0.00
```

Playing with the viewing format a little we can observe the most common vanues for each of the neighborhoods in a common table.

| | Neighborhood | 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 Most Common Venue | 8th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Agincourt | Lounge | Sandwich Place | Breakfast Spot | Chinese Restaurant | Vietnamese Restaurant | Coffee Shop | General Entertainment | Fried Chicken Joint |
| 1 | Agincourt North, L'Amoreaux East, Milliken, St... | Park | Asian Restaurant | Playground | Vietnamese Restaurant | General Entertainment | Fried Chicken Joint | Fast Food Restaurant | Electronics Store |
| 2 | Birch Cliff, Cliffside West | General Entertainment | Skating Rink | Café | College Stadium | Vietnamese Restaurant | Coffee Shop | Grocery Store | Fried Chicken Joint |
| 3 | Cedarbrae | Caribbean Restaurant | Thai Restaurant | Athletics & Sports | Fried Chicken Joint | Bakery | Bank | Lounge | Hakka Restaurant |
| 4 | Clairlea, Golden Mile, Oakridge | Bakery | Bus Line | Intersection | Fast Food Restaurant | Soccer Field | Bus Station | Metro Station | Park |

Finally, we are ready to perform the clustering. We are using partitioning-based clustering K-means algorithm, which suggests the set number of mutually exclusive clusters. We are going to opt for 5 clusters, which are presented in the map below. Each color identifies a specific cluster, thus the neighborhoods of the same color are in the same cluster, namely are similar to each other and dissimilar to others.

Now let's have a closer look at each of the clusters.

*Cluster 1.* Most common venues in the first cluster mainly include places for some family time, thus I'd call it 'Family'.

```
[48]: scarborough_merged.loc[scarborough_merged['Cluster Labels'] == 0, scarborough_merged.columns[[2] + list(range(5, scarborough_merged.shape[1]
```

[48]:

| | Neighborhood | 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 Most Common Venue | 8th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Scarborough Village | 0 | Playground | Vietnamese Restaurant | Chinese Restaurant | Grocery Store | General Entertainment | Fried Chicken Joint | Fast Food Restaurant | Electronics Store |
| 14 | Agincourt North, L'Amoreaux East, Milliken, St... | 0 | Park | Asian Restaurant | Playground | Vietnamese Restaurant | General Entertainment | Fried Chicken Joint | Fast Food Restaurant | Electronics Store |

*Cluster 2.* Here we see, that most of he venues, such as Coffee Shop, Discount Store, Restaurants, etc., are popular with the youth, hence I'll name it 'youth'.

```
]: scarborough_merged.loc[scarborough_merged['Cluster Labels'] == 1, scarborough_merged.columns[[2] + list(range(5, scarborough_merged.shape[1]
```

]:

| | Neighborhood | 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 Most Common Venue | 8th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | Guildwood, Morningside, West Hill | 1 | Intersection | Breakfast Spot | Mexican Restaurant | Tech Startup | Spa | Electronics Store | Medical Center | Pizza Place |
| 3 | Woburn | 1 | Coffee Shop | Korean Restaurant | Convenience Store | Vietnamese Restaurant | Grocery Store | General Entertainment | Fried Chicken Joint | Fast Food Restaurant |
| 4 | Cedarbrae | 1 | Caribbean Restaurant | Thai Restaurant | Athletics & Sports | Fried Chicken Joint | Bakery | Bank | Lounge | Hakka Restaurant |
| 6 | East Birchmount Park, Ionview, Kennedy Park | 1 | Discount Store | Department Store | Playground | Coffee Shop | Vietnamese Restaurant | Chinese Restaurant | Grocery Store | General Entertainment |
| 7 | Clairlea, Golden Mile, Oakridge | 1 | Bakery | Bus Line | Intersection | Fast Food Restaurant | Soccer Field | Bus Station | Metro Station | Park |
| 9 | Birch Cliff, Cliffside West | 1 | General Entertainment | Skating Rink | Café | College Stadium | Vietnamese Restaurant | Coffee Shop | Grocery Store | Fried Chicken Joint |
| 10 | Dorset Park, Scarborough Town Centre, Wexford ... | 1 | Indian Restaurant | Chinese Restaurant | Latin American Restaurant | Vietnamese Restaurant | Pet Store | Bakery | Grocery Store | General Entertainment |
| 11 | Maryvale, Wexford | 1 | Middle Eastern Restaurant | Auto Garage | Bakery | Shopping Mall | Sandwich Place | Breakfast Spot | Vietnamese Restaurant | College Stadium |
| 12 | Agincourt | 1 | Lounge | Sandwich Place | Breakfast Spot | Chinese Restaurant | Vietnamese Restaurant | Coffee Shop | General Entertainment | Fried Chicken Joint |
| 13 | Clarks Corners, Sullivan, Tam O'Shanter | 1 | Pizza Place | Noodle House | Thai Restaurant | Fried Chicken Joint | Fast Food Restaurant | Italian Restaurant | Bank | Chinese Restaurant |
| 15 | L'Amoreaux West | 1 | Chinese Restaurant | Fast Food Restaurant | Coffee Shop | Grocery Store | Pharmacy | Pizza Place | Breakfast Spot | American Restaurant |

Activate W

**Cluster 3.** Fast Food Restaurant, Print Shop, Asian Restaurants, Grocery Stores are liked budget-concious visiors, so it's gonna be named as 'Budget-concious'.

```
[50]: scarborough_merged.loc[scarborough_merged['Cluster Labels'] == 2, scarborough_merged.columns[[2] + list(range(5, scarborough_merged.shape[1]
```

[50]:

| | Neighborhood | 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 Most Common Venue | 8th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Rouge, Malvern | 2 | Fast Food Restaurant | Print Shop | Vietnamese Restaurant | Chinese Restaurant | Grocery Store | General Entertainment | Fried Chicken Joint | Electronics Store |

**Cluster 4.** Bar,Construction & Landscaping, Restaurants, Coffee Shops,Grocery Stores are of main inerest to man, thus we have here the 'Man' cluster.

```
[51]: scarborough_merged.loc[scarborough_merged['Cluster Labels'] == 3, scarborough_merged.columns[[2] + list(range(5, scarborough_merged.shape[1]
```

[51]:

| | Neighborhood | 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 Most Common Venue | 8th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Highland Creek, Rouge Hill, Port Union | 3 | Bar | Construction & Landscaping | Vietnamese Restaurant | Coffee Shop | Grocery Store | General Entertainment | Fried Chicken Joint | Fast Food Restaurant |

**Cluster 5.** And the last cluster with 1st Most Common Venue as motel is named 'traveler'.

```
[52]: scarborough_merged.loc[scarborough_merged['Cluster Labels'] == 4 , scarborough_merged.columns[[2] + list(range(5, scarborough_merged.shape[1
```

[52]:

| | Neighborhood | 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 Most Common Venue | 8th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|
| 8 | Cliffcrest, Cliffside, Scarborough Village West | 4 | Motel | American Restaurant | Coffee Shop | Grocery Store | General Entertainment | Fried Chicken Joint | Fast Food Restaurant | Electronics Store |

Activate W

As we can see clusters differ depending on main customers/visitor of the most common venues of the cluster. Thus, the names of the clusters are based on the group of the main visitors.

## Results

Now we have five main clusters as a final result. The most common venues of the neighborhoods in these cluster are Playgrounds, Parks, Restaurants And Grocery Stores. We can conclude that these cluster is mostly popular with families.

The most popular venues In the second cluster turned out to be General Entertainment, Café's, Coffee Shops, Fast Food Restaurants, Discount Stores, Skating Rink, etc. It is obvious, that here the target group are young professionals, students, etc. Thus, we have here the "Youth" cluster.

Fast Food Restaurant, Print Shop, Asian Restaurants, Grocery Stores were the most common venues for the third cluster, which are liked by budget-conscious visitors, so it is going to be named as 'Budget-conscious'.

The fourth cluster might be mainly popular with man, as we can conclude from Bars, Construction & Landscaping, Restaurants, Coffee Shops, Grocery Stores, which are of main interest to man, thus we have here the 'Man' cluster.

And the last fifth cluster is going to be called 'Traveler' due to Motel, General Entertainment, Coffee Shops being the most common venues.

## Conclusion

Having gone through a long journey of data collection, collation, analysis and visualization, using K-means clustering algorithm and analyzing the results, we received five main groups of neighborhoods:

1. Family
2. Youth
3. Budget-conscious
4. Man
5. Traveler

Our task was to determine the neighbors where it would be most efficient to open a club. Our results make it clear, that the neighborhoods of the cluster "Youth" would be the perfect locations for clubs.

The analysis turned out really successful and hit all mentioned goals. On one hand these neighborhoods were indicating as common ones venues which are popular with the Youth (e.g. General Entertainment, Coffee Shops, Café's, Discount Stores or Fast Food Restaurants), on the other hand at the same time no clubs were mentioned. Here we meet the strategic decision of the business owner to settle in places with max potential demand and minimal competition.

The map introduces in purple all the neighborhoods which are going to work the best for this case.



## Future directions

In this work we have segmented our analysis to one borough only. It might be useful to also analyze the data for the whole Toronto area and see if the clusters look similar and the logic behind them can be related to the one noted here.

Another interesting point would be to make the same analysis within the chosen cluster only. This will give us more information about dissimilarities within the clusters and help to make more specific recommendations.