Find a suitable location to open a restaurant in Toronto

Abhishek Deb

April 2020

1. Introduction

1.1. Background :
As the population is constantly increasing, the requirement of Different services are also increasing. Hence, Opening a Restaurant or a chain of Restaurant is a very common business idea For any entrepreneur nowadays. The Success and profit of the restaurant are dependent on several factors like locality, type, services etc. To Maximize the profit of the Restaurant, It is important to find the answer of a few fundamental questions. One of the very first problems is to find a suitable locality for the new Restaurant as location plays a big role in the success of Restaurant. In this project, I am trying to solve this problem using Data Science and Machine learning Algorithm.

1.2. Business Problem:
The objective of my project is to find a suitable location to open a new restaurant in Toronto, the financial capital of Canada. I have used data science methods and machine learning algorithms using the programming language python in this project. The goal of this project is to provide the answer to the business question "In Toronto if an entrepreneur wants to open a new restaurant, which neighbourhood/neighbourhoods will be suitable for this?

1.3. Target Audience:
The entrepreneur who wants to find a good location to open a new restaurant in Toronto.

2. Data:

2.1. Data Acquisition:
We dont use to see ongoing and successful restaurant everywhere in a city. The concept behind this project is, As locality of a restaurant is a key factor if we can identify the similar locality based on different Characteristic of those localities, we can choose our suitable location as well.

As an example, If neighbourhood N1, N2, N3 is similar and if Restaurants in N1 and N2 are successful then there is a high possibility that N3 is also a suitable location for Restaurant.

In this project, we have used the below data to find the similar Neighbourhood.

i. Different neighbourhood information of Toronto City along with their popularity, income and other details: We have extracted the data from The official website for the City of Toronto (www.toronto.ca). please Note, up to the year 2016 data is available here hence we have used the latest data (i.e. the data of 2016) only.

ii. Latitude and Longitude of all neighbourhoods: we have extracted the Latitude and Longitude data of all neighbourhoods using Geocoder.

iii. Different venue details based on all neighbourhoods: we have extracted the data based on Foursquare API

2.2. Data Cleaning and Modification :

The initial data we have Sourced from the official website of Toronto was filled with thousands of detail.

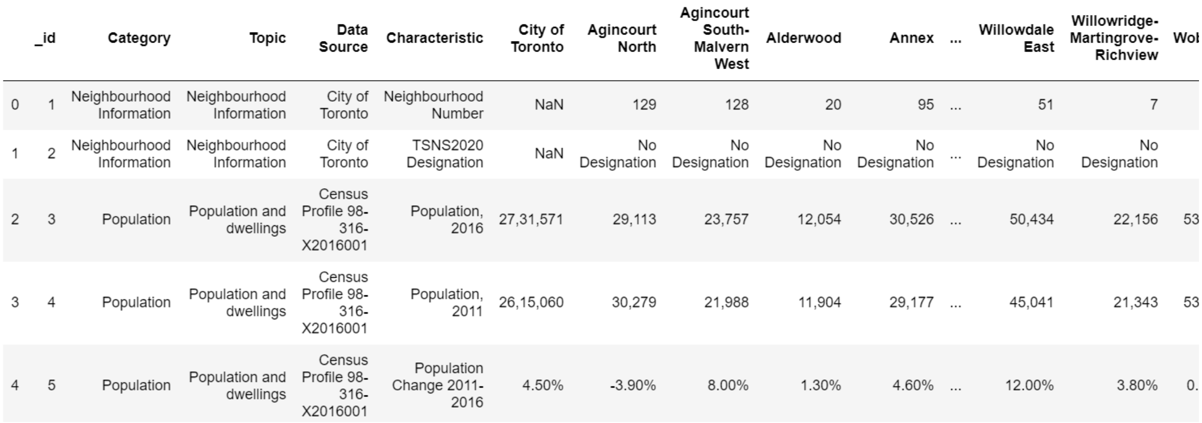
First Column in the dataset '-id', a number to define each observation.

From Second to Fifth Columns are describing what data is present in that row.

Sixth Column is containing information about Toronto city as a whole.

For each column in Seventh to last column is representing one Neighbourhood's information for different categories.

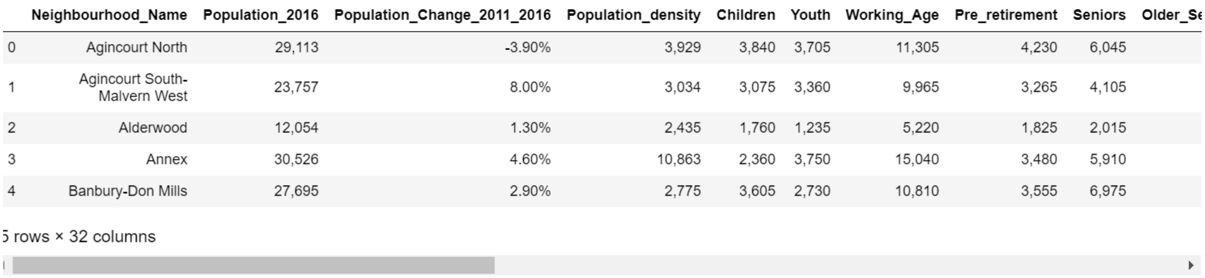
Pic 1. Initial data from Toronto's website



	_id	Category	Topic	Data Source	Characteristic	City of Toronto	Agincourt North	Agincourt South-Malvern West	Alderwood	Annex	...	Willowdale East	Willowridge-Martingrove-Richview	Wol
0	1	Neighbourhood Information	Neighbourhood Information	City of Toronto	Neighbourhood Number	NaN	129	128	20	95	...	51	7	
1	2	Neighbourhood Information	Neighbourhood Information	City of Toronto	TSNS2020 Designation	NaN	No Designation	No Designation	No Designation	No Designation	...	No Designation	No Designation	
2	3	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2016	27,31,571	29,113	23,757	12,054	30,526	...	50,434	22,156	53
3	4	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2011	26,15,060	30,279	21,988	11,904	29,177	...	45,041	21,343	53
4	5	Population	Population and dwellings	Census Profile 98-316-X2016001	Population Change 2011-2016	4.50%	-3.90%	8.00%	1.30%	4.60%	...	12.00%	3.80%	0.

I have cleaned up those columns which will not be required in this project. As an example, we can understand the observation with 5th column only, so I have deleted 2nd -4th columns. Also, I am going to work based on the Neighbourhood . so it will be easier if each row represents the data of each neighbourhood. Hence I have rotated our data. Along with that, I have given a more understandable column name.

Pic 2. It is showing data after few modification

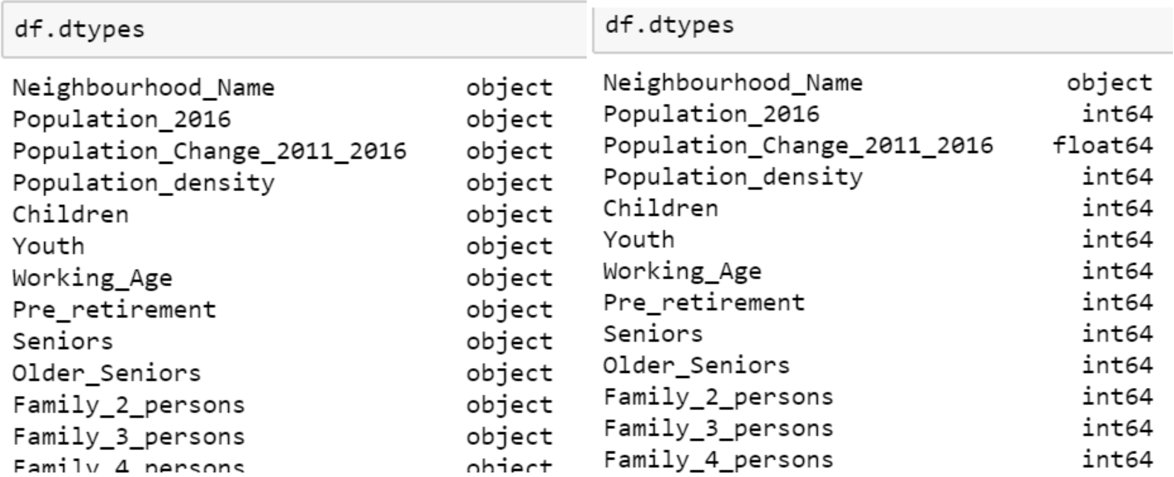


	Neighbourhood_Name	Population_2016	Population_Change_2011_2016	Population_density	Children	Youth	Working_Age	Pre_retirement	Seniors	Older_Se
0	Agincourt North	29,113	-3.90%	3,929	3,840	3,705	11,305	4,230	6,045	
1	Agincourt South-Malvern West	23,757	8.00%	3,034	3,075	3,360	9,965	3,265	4,105	
2	Alderwood	12,054	1.30%	2,435	1,760	1,235	5,220	1,825	2,015	
3	Annex	30,526	4.60%	10,863	2,360	3,750	15,040	3,480	5,910	
4	Banbury-Don Mills	27,695	2.90%	2,775	3,605	2,730	10,810	3,555	6,975	

5 rows × 32 columns

It Has been observed that almost all columns are containing numeric values only though their type in the Data is not numeric. I have made the changes to numeric so that it will be easier to use different process/ algorithm in the data.

Pic 3. Please find before and after datatypes



```
df.dtypes
```

```
Neighbourhood_Name               object
Population_2016                  object
Population_Change_2011_2016      object
Population_density               object
Children                         object
Youth                            object
Working_Age                      object
Pre_retirement                   object
Seniors                          object
Older_Seniors                    object
Family_2_persons                 object
Family_3_persons                 object
Family_4_persons                 object
```

```
df.dtypes
```

```
Neighbourhood_Name              object
Population_2016                  int64
Population_Change_2011_2016    float64
Population_density               int64
Children                         int64
Youth                            int64
Working_Age                      int64
Pre_retirement                   int64
Seniors                          int64
Older_Seniors                    int64
Family_2_persons                 int64
Family_3_persons                 int64
Family_4_persons                 int64
```

The dataset is containing the number of families based on members of each family. except 'Couples' and 'living alone', let me divide these into two categories
i. If the family contains less than 5 members, it is a small family
ii. If the Family has 5 or more members, it's a big family

In a similar way, I have reduced the columns related to income by dividing income ranges into 3 columns. The average income of Canda in 2016 was around $ 58,000 for a year, so I am going to create below three category

i. If Income is lower than $ 40,000, its Low Income
ii. If Income is more than $ 40,000 but less than $ 70,000 its AVG income
iii.More than $ 70,000 its a high income

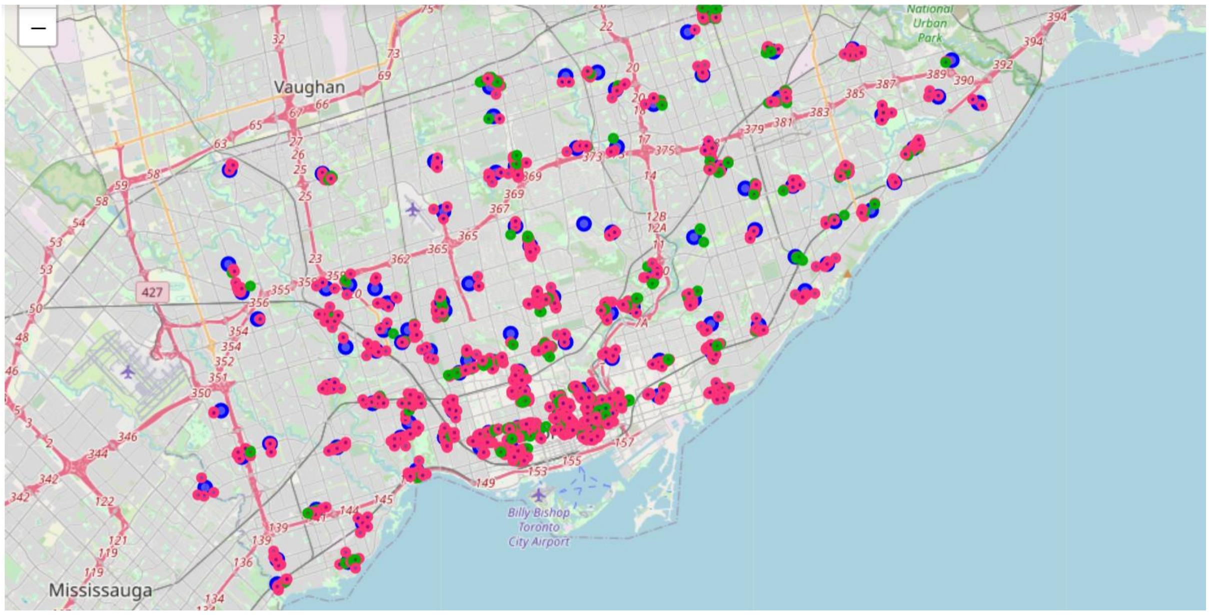
In the second step of data acquisition while extracting latitude and longitude we have not received geocode for a few neighbourhoods and received one wrong geocode. As part of the final cleaning steps, we have removed those neighbourhoods before proceeding to machine learning .

3. Methodology:

3.1. Exploratory Data Analysis: as part of data analysis we have plotted the data in map using folium library to show the distribution of restaurant and other venues

In the below map, I have marked three points on a map .
i. Blue circle: It is representing the centre point of the Neighbourhood
ii. Green Point: The location of a restaurant
iv. Purple point: The location of other venues

Pic 4. Toronto's map with highlighting neighbourhoods, restaurant and other venues

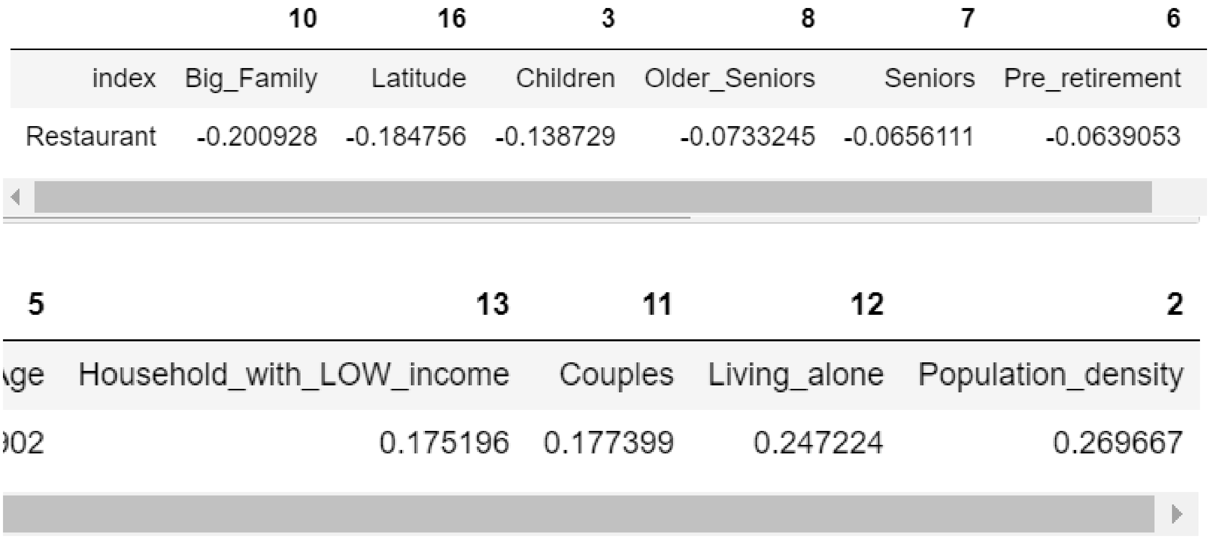


3.2. Feature Selection and scaling :

In our Final consolidated dataset, we have data of neighbourhood from two categories mainly.
i.Population related data such as Population_2016, Population_density, Age and salary-related information etc
ii.Venue related data such as number of markets, Cafe, grocery store etc present in that neighbourhood

It has observed in machine learning that if we use important features instead of all features, the machine learning algorithm's output is better. So I will filter out those entities which have more effect in Restaurant. For This, I have used the correlation matrix.
I will select features from two categories which I have mentioned. For the population-related category, I have selected two features which have more effect.
i. Living_alone (high positive correlation)
ii. Population_density (high positive correlation)

pic 5. Corelation score between Restaurant and population-related data

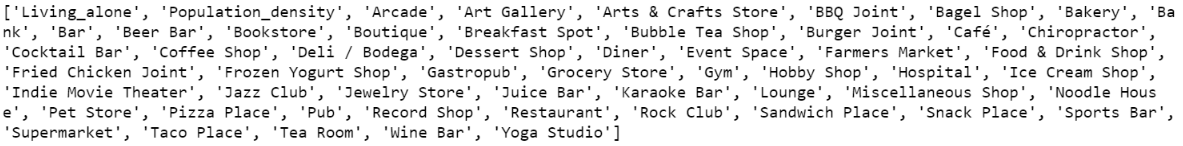


	10	16	3	8	7	6
index	Big_Family	Latitude	Children	Older_Seniors	Seniors	Pre_retirement
Restaurant	-0.200928	-0.184756	-0.138729	-0.0733245	-0.0656111	-0.0639053

5	13	11	12	2
ge	Household_with_LOW_income	Couples	Living_alone	Population_density
902	0.175196	0.177399	0.247224	0.269667

For venue related category, I have used the same correlation matrix, but this time. As we have several columns in venue details, I have selected only those features which have correlation score more than 0.24 or less than -0.24(i.e. high positive or negative correlation). Below are the final selected features which will be used in the later phase.

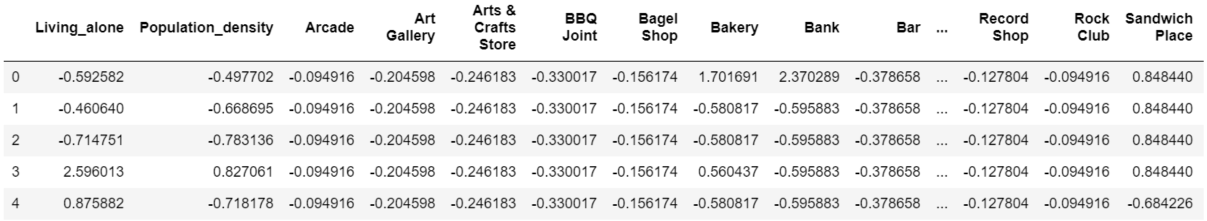
Pic 6. Final Selected features



```
['Living_alone', 'Population_density', 'Arcade', 'Art Gallery', 'Arts & Crafts Store', 'BBQ Joint', 'Bagel Shop', 'Bakery', 'Ba
nk', 'Bar', 'Beer Bar', 'Bookstore', 'Boutique', 'Breakfast Spot', 'Bubble Tea Shop', 'Burger Joint', 'Café', 'Chiropractor',
'Cocktail Bar', 'Coffee Shop', 'Deli / Bodega', 'Dessert Shop', 'Diner', 'Event Space', 'Farmers Market', 'Food & Drink Shop',
'Fried Chicken Joint', 'Frozen Yogurt Shop', 'Gastropub', 'Grocery Store', 'Gym', 'Hobby Shop', 'Hospital', 'Ice Cream Shop',
'Indie Movie Theater', 'Jazz Club', 'Jewelry Store', 'Juice Bar', 'Karaoke Bar', 'Lounge', 'Miscellaneous Shop', 'Noodle Hous
e', 'Pet Store', 'Pizza Place', 'Pub', 'Record Shop', 'Restaurant', 'Rock Club', 'Sandwich Place', 'Snack Place', 'Sports Bar',
'Supermarket', 'Taco Place', 'Tea Room', 'Wine Bar', 'Yoga Studio']
```

Feature scaling in ML algorithms is one of the most important steps, which use to affect the output of the algorithm. It helps to normalise the data within a particular range. we have used StandardScaler from SKlearn library to perform feature scaling on our data before feeding it to ML algorithm.

Pic 7. Dataset after scaling



	Living_alone	Population_density	Arcade	Art Gallery	Arts & Crafts Store	BBQ Joint	Bagel Shop	Bakery	Bank	Bar	...	Record Shop	Rock Club	Sandwich Place
0	-0.592582	-0.497702	-0.094916	-0.204598	-0.246183	-0.330017	-0.156174	1.701691	2.370289	-0.378658	...	-0.127804	-0.094916	0.848440
1	-0.460640	-0.668695	-0.094916	-0.204598	-0.246183	-0.330017	-0.156174	-0.580817	-0.595883	-0.378658	...	-0.127804	-0.094916	0.848440
2	-0.714751	-0.783136	-0.094916	-0.204598	-0.246183	-0.330017	-0.156174	-0.580817	-0.595883	-0.378658	...	-0.127804	-0.094916	0.848440
3	2.596013	0.827061	-0.094916	-0.204598	-0.246183	-0.330017	-0.156174	0.560437	-0.595883	-0.378658	...	-0.127804	-0.094916	0.848440
4	0.875882	-0.718178	-0.094916	-0.204598	-0.246183	-0.330017	-0.156174	-0.580817	-0.595883	-0.378658	...	-0.127804	-0.094916	-0.684226

3.3. Applying ML Algorithm : At first, let me explain what logic I have used to achieve the answer to the business problem.

Let's say, we have 5 neighbourhoods such as N1, N2, N3, N4 and N5 and number of restaurant in those neighbourhoods are 10, 2, 2, 13, 4 respectively.

And After using a clustering algorithm based on its features (excluding the restaurant number), let's say we have 2 clusters such as (N1, N3, N4) and (N2, N5)

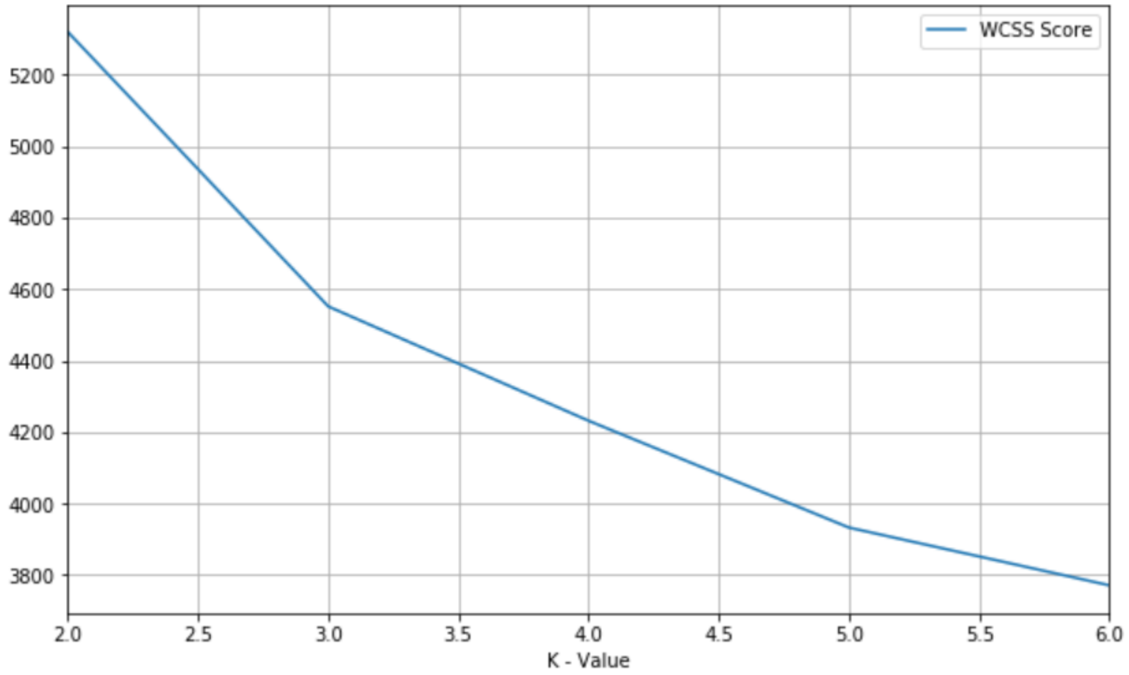
As N1, N3 and N4 are similar and we have a high number of successful restaurants in N1 and N4, we can expect N3 is also a good location for a restaurant with less competition.

To find the cluster in our neighbourhood data (after feature selection and scaling), I have used unsupervised machine learning algorithm K-means. Kmeans algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It is one of the simplest, popular and effective algorithms and highly suited for this project as well.

As user needs to feed the number of cluster K into the algorithm, it is important to identify the right K. here I have used Elbow- Method to find the right K for this problem. I have applied this algorithm for different K values and plotted K vs WCSS(within-cluster sum of square) to find the right K. In our case, the right value for K was 5. hence we have divided the neighbourhoods into 5 clusters.

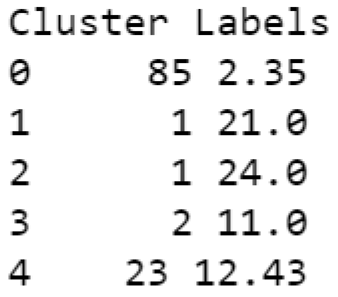
Pic 8. K vs WCSS graph

```
<matplotlib.axes._subplots.AxesSubplot at 0x2cebc277f08>
```



4. **Results and Discussion:** After applying K means algorithm for the value of K=5, I have received 5 clusters. To understand and find the best cluster I have collected below information about the cluster.
i. How many neighbourhoods belongs to each cluster
ii. The average number of Restaurant for each cluster.

Pic 9. Some details about cluster



```
Cluster Labels
0      85 2.35
1       1 21.0
2       1 24.0
3       2 11.0
4      23 12.43
```

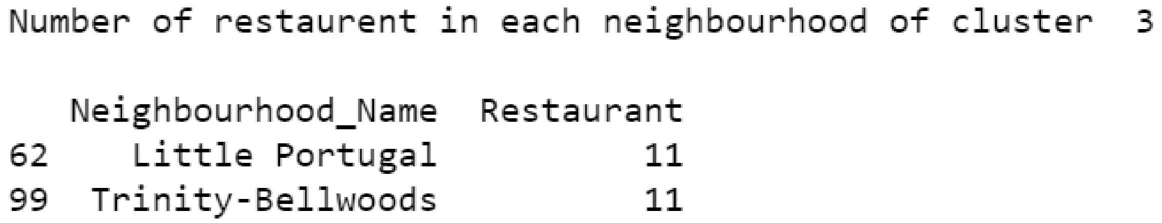
In the above table, the first column is representing the cluster number and the second column is showing two values, the number of neighbourhoods in that cluster and the average number of restaurant in that cluster.

We can see, Cluster no 2 and 3 have only one Neighbourhood only with high Restaurant numbers. Must be these neighbourhoods are very different than others. so we can exclude it as creating a restaurant here will be full of tough competitions.

Also, cluster 0 has very low average restaurant numbers , so we need to exclude this .
Among others we can see the average number of Restaurant is high is cluster 3 and 4.

To explore cluster 3 and 4 a little more, I have checked restaurant for each neighbourhood in that cluster.

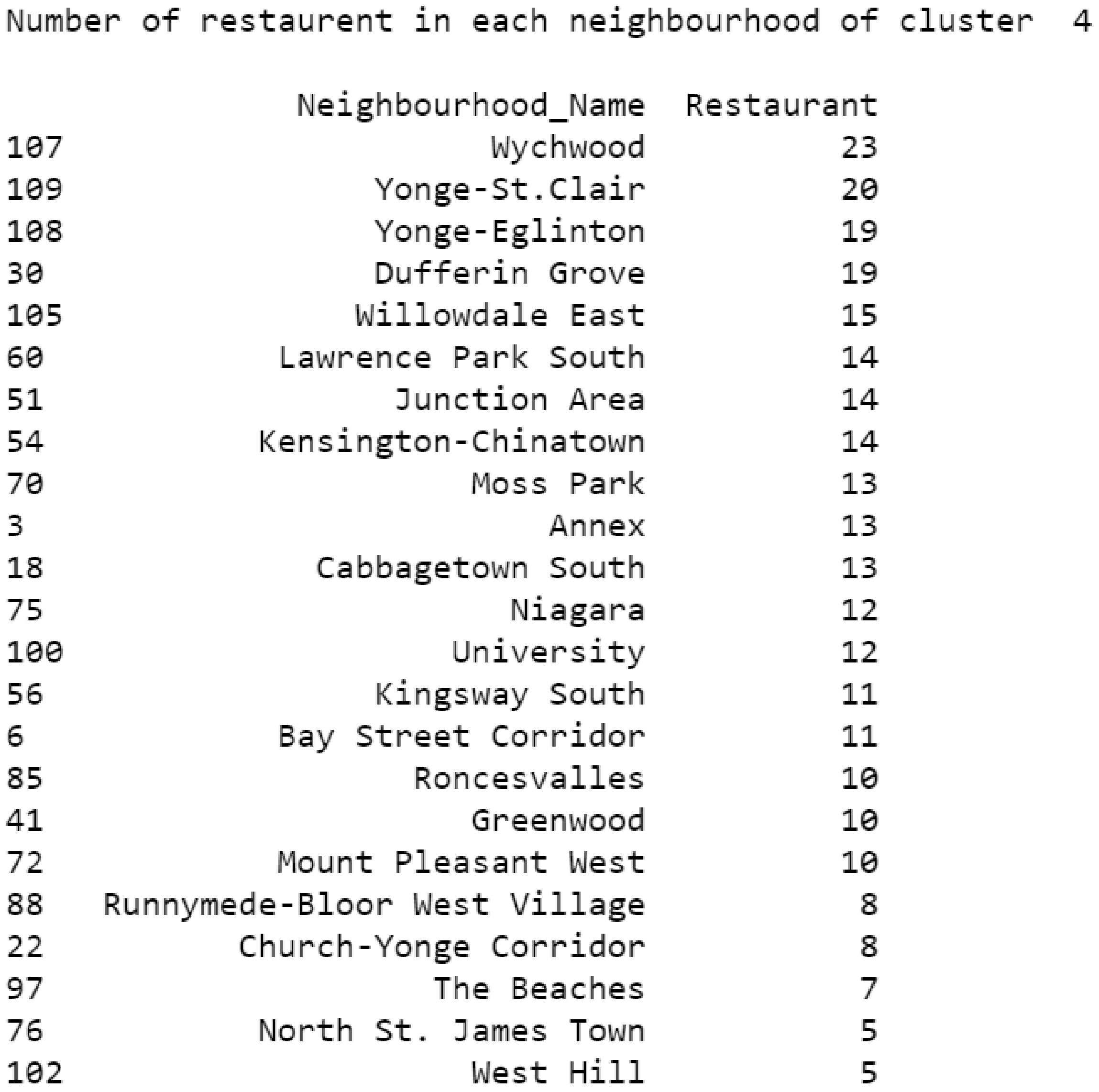
Pic 10. Cluster 3 related information



```
Number of restaurent in each neighbourhood of cluster  3

   Neighbourhood_Name  Restaurant
62    Little Portugal          11
99  Trinity-Bellwoods          11
```

Pic 11. Cluster 3 related information



```
Number of restaurent in each neighbourhood of cluster  4

                Neighbourhood_Name  Restaurant
107                       Wychwood          23
109                 Yonge-St.Clair          20
108                 Yonge-Eglinton          19
30                  Dufferin Grove          19
105                Willowdale East          15
60             Lawrence Park South          14
51                   Junction Area          14
54            Kensington-Chinatown          14
70                       Moss Park          13
3                            Annex          13
18               Cabbagetown South          13
75                         Niagara          12
100                     University          12
56                  Kingsway South          11
6               Bay Street Corridor         11
85                     Roncesvalles         10
41                        Greenwood         10
72             Mount Pleasant West          10
88   Runnymede-Bloor West Village            8
22            Church-Yonge Corridor          8
97                      The Beaches          7
76            North St. James Town           5
102                       West Hill          5
```

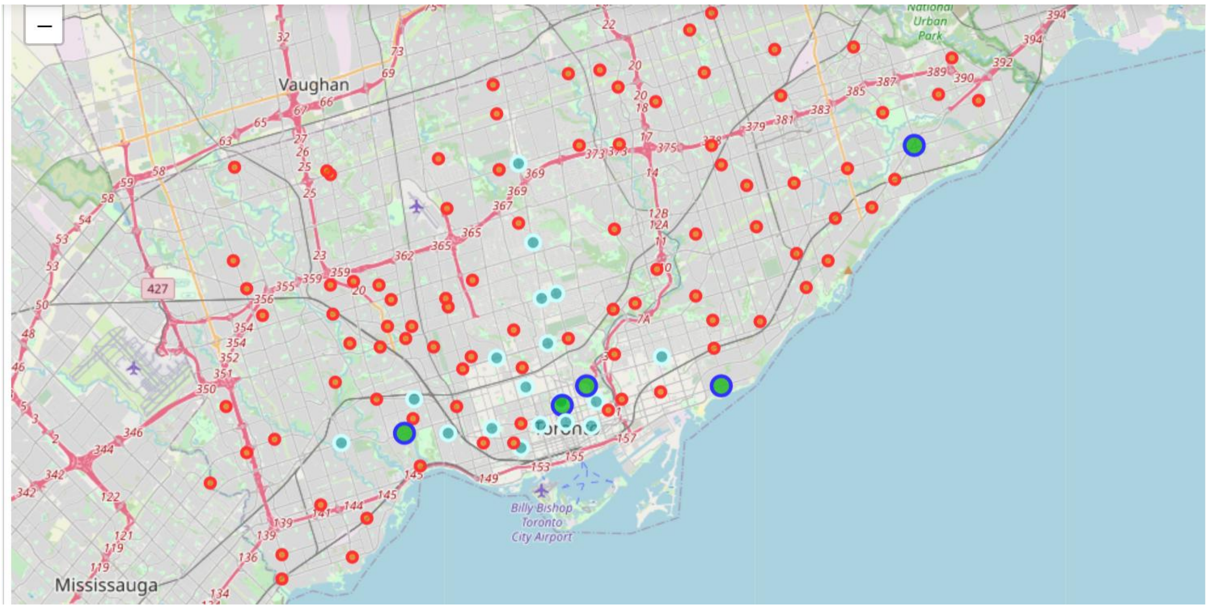
5. **Recommendations:** After exploring the above clusters, we can see each neighbourhood in cluster 3 has 10+ restaurants. Again there must be high competition for a new restaurant.
But, in cluster number 4 we can see few neighbourhoods with less than 10 restaurants.
Those are : 'Runnymede-Bloor West Village', 'Church-Yonge Corridor ', 'The Beaches', 'North St. James Town' and 'West Hill'.

These 5 Neighbourhoods have less number of restaurants although the neighbourhood is similar to other neighbourhood which has a higher number of restaurants.

So, An entrepreneur can think of opening a new restaurant in there because those neighbourhoods have that element which can be helpful towards the restaurant business.

At Final step, I have plotted and highlighted these neighbourhoods in Toronto map using Folium library.

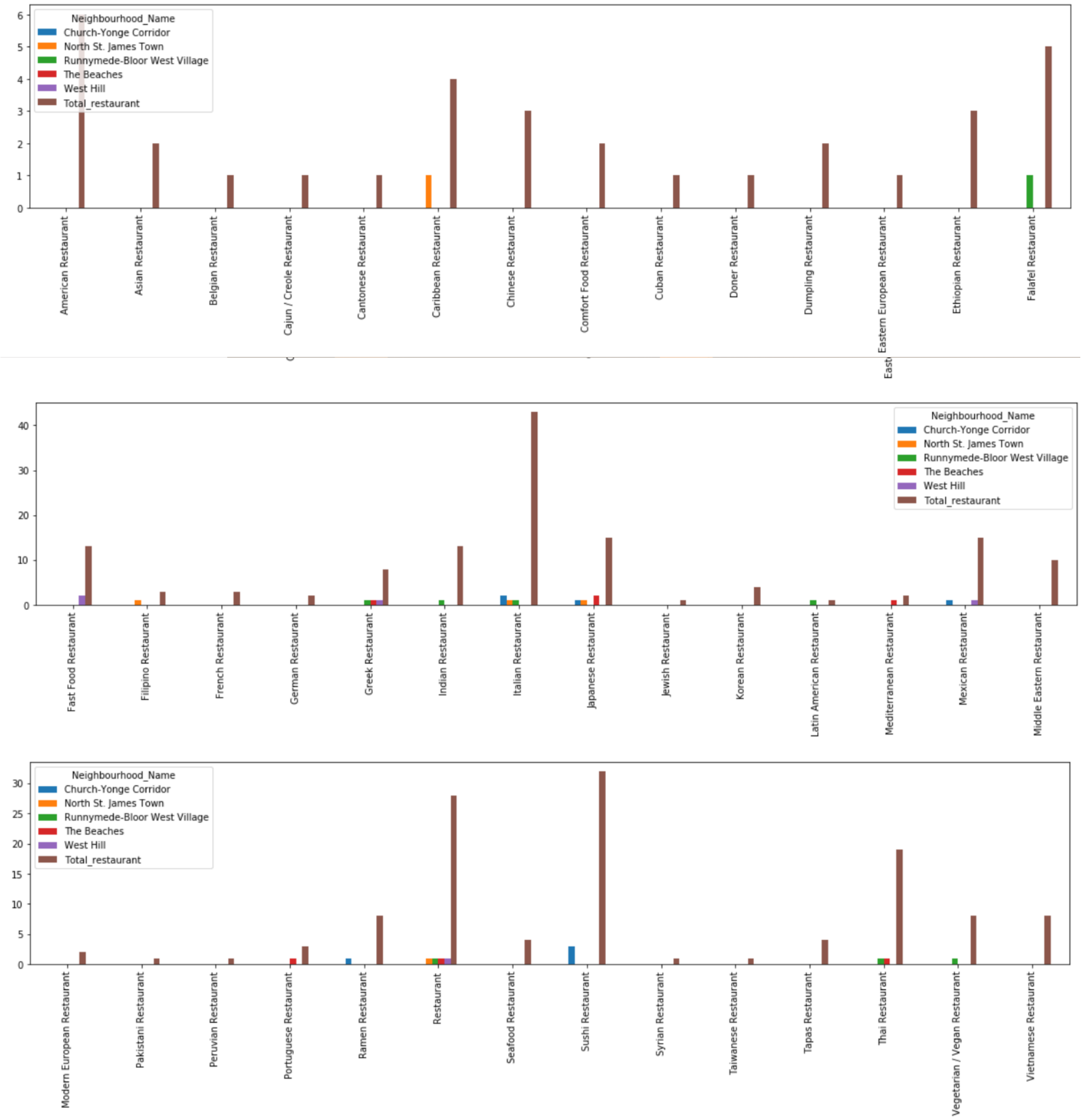
Pic 12. Final Toronto's map with highlighting preferred Neighbourhoods



In the above map, The largest cycles are representing the best 5 location according to this project.
the small blue circle is representing, other Neighbourhoods in cluster 4.
And all other Neighbourhoods are showing as a red circle.

Additional: I have also plotted 'comparison between 5 selected neighbourhood's restaurant vs cluster no 4 total restaurants' using barplot which may help to select the perfect type of restaurant.

Pic 13. Comparison between 5 selected neighbourhood's restaurant vs cluster no 4 total restaurants



6. **<u>Future directions:</u>** In this project, I have considered a few aspects such as few population-related data and few venue related data. There are many factors that can be taken into consideration such as connectivity of the neighbourhood, rating of the existing venues, how much the neighbourhood is developing etc. Additionally, if we can categorize existing venues in a better way, it will affect the result as well. Future research can take into consideration these factors.

7. **<u>Conclusion:</u>** In this study to find a solution to the business problem, I have gone through the process of identifying the business

problem. I have analyzed, collected and prepared to find a solution performing the machine learning (i.e. k-means clustering). In the end, I have provided recommendations to the entrepreneur based on the result.