Location prediction of the eastern European cuisine restaurant in the Toronto area

Capstone Project

Author
Vladimir Dikić

Keywords
Location prediction
Toronto income and demographics analysis
Clustering
Classification

Report issued : 02-28-2020

1. Introduction

Starting and running a successful business is a multifactor dependant process, whi ch requires both good initial screening, marketing and continuous improvement.

The market analysis is an important first step in predicting the viability of an endeavour. Accordingly, opening a restaurant with a niche cuisine is highly dependent on the target group of customers to help the business grow, before expanding and capturing the customers that are outside of the target culture. Likewise, it is important to capture the market where the income level of customers matches the offer business provides. Last but not least, it is important to evaluate the competition in the targeted area.

1.1 Problem statement

Stakeholder description/ interest

The *family of immigrants from an eastern European country that are the residents of Toronto for past 25 years have decided to pursue their own business and open a restaurant in Toronto area. The idea is to open an affordable restaurant for the middle-income customers that would especially cater to the customers of the similar background. For this reason, the family has hired a data analyst to help them determine the best location for opening the venue.

Problem tackling through data

The best restaurant location is dependent on several factors. The data required to tackle the issue needs to contain the geographical parameters necessary to utilize the Foursquare platform to obtain the competing venues. It also needs to contain the demographic data for each neighbourhood: population, average income, ethnic makeup (or similar). The acquisition of aforementioned data is described in the next chapter.

**Note from the author: This project is somewhat based on the real world problem. Yours truly has a family in Toronto, and at one family meeting some time ago, they mentioned that they were contemplating whether to open an eastern European restaurant. So this whole project is a nice opportunity to test what would be the best place to do that, and if the idea is viable.*

2. Data acquisition and processing

2.1 Data acquisition

The necessary data and the sources are described in the list below:

- Neighbourhoods, Borough and postal code data:
    - The information is obtained from canadapost.ca website, the grouped data is taken from [1]. This data is primarily needed to access the geographical data and Foursquare API. Tables were scraped using the BeautifulSoup toolbox

- Neighbourhood geo data:
    - The coordinates data for each neighbourhood was obtained using geocoder-arcgis method on postal code information from the Toronto neighbourhood data

- Population, average income and demographic data:
    - The demographics data was obtained from the [2]. This will be used to determine the best-suited neighbourhood in terms of number and income of target customer group (i.e. immigrants from Eastern Europe). Tables were scraped using the BeautifulSoup toolbox

- Venue data:
    - Foursquare API is used to obtain the venues in Toronto area. This will be used to determine the amount of competition in each neighbourhood

2.2 Data processing

Initial step of data processing was scrapping the tables from [1] and [2] using the BeautifulSoup toolbox. Subsequently raw data was pre-cleaned and placed in the corresponding dataframes. Data was additionally cleaned removing duplicates, NaN values and transforming numbers of string type into real numbers

Dataframe obtained from data [1] will be referred to as “postal code dataframe”, and dataframe [2] will be referred to as “demographics dataframe”.

The main difficulties in merging the postal code and demographics data came from the significant difference in neighbourhood naming. This consisted of differently grouped neighbourhoods in both data sets (e.g. neighbourhood X and Y in data set 1 are grouped together while in the data set 2, X and Y are separate, and the other way around). In addition, the level of detail was different in the data sets, where table 1 can contain one neighbourhood but split into west and east side, while the table 2 only accounts for the entire area (e.g. Steeles vs South and East Steeles).

First, it was necessary to separate all the grouped neighbourhoods. This was made slightly more difficult by inconsistent grouping symbols (e.g. “\”, “-“, “,”, “/”). The loop was utilized to account for all the aforementioned symbols, where in order to retain the other data (e.g. postal code while separating neighbourhoods), grouped data was separated by using split, expand and stack function.

Second, in some cases the data in the demographic data set only accounted for whole neighbourhoods, while postal code data set also accounted for geographical areas of the neighbourhood, as was previously mentioned. This was solved by appending the demographic dataframe with the disaggregated neighbourhoods. Each row of the postal code data frame was tested if it contains the neighbourhood that includes the name of non-joinable neighbourhood from the demographics dataframe (e.g. East Steeles contains Steeles). Then, the demographics data frame was appended using the following rules:

- Neighbourhood is split into segments that correspond to number of segments in postal code dataframe. Names are equalized
- Population data is separated in even amounts depending on the number of segments
- Demographics and the average income are assumed to stay the same for each separated segment

Demographics of the area were determined by the majority spoken second language in the neighbourhood (outside of English). Subsequently, these language groups were assigned to geographical / cultural groups. These assignments do not yield in a precise segmentation (e.g. grouping Persian and Somali or Filipino and Mandarin), but will allow for easier clustering in order to find the best location for an eastern European restaurant.

Table 2-1 Segmentation based on geographical area and cultural similarity

Geo/Cultural group	Language group
South Asian / Indian	[Tamil, Gujarati, Bengali, Urdu, Punjabi]
East Asian	[Filipino, Cantonese, Mandarin, Korean]
West Asian and African	[Persian, Somali]
Unspecified	[Unspecified]
Eastern European/Slavic	[Russian, Bulgarian, Ukrainian, Polish]
South European / Romance	[Greek, French, Spanish, Portuguese, Italian]

3. Data analysis

The problem and the possible solutions need to be evaluated and understood through data, in order to utilize the clustering/ classification algorithms properly. For this reason, several aspects of the data set were evaluated.

Population data

Population data was obtained from [2], it can give a good insight in the size of the market a certain business is entering in. This can show the amount of available customers in near proximity of the venue.

The following tables illustrate the most and least populated areas of Toronto.

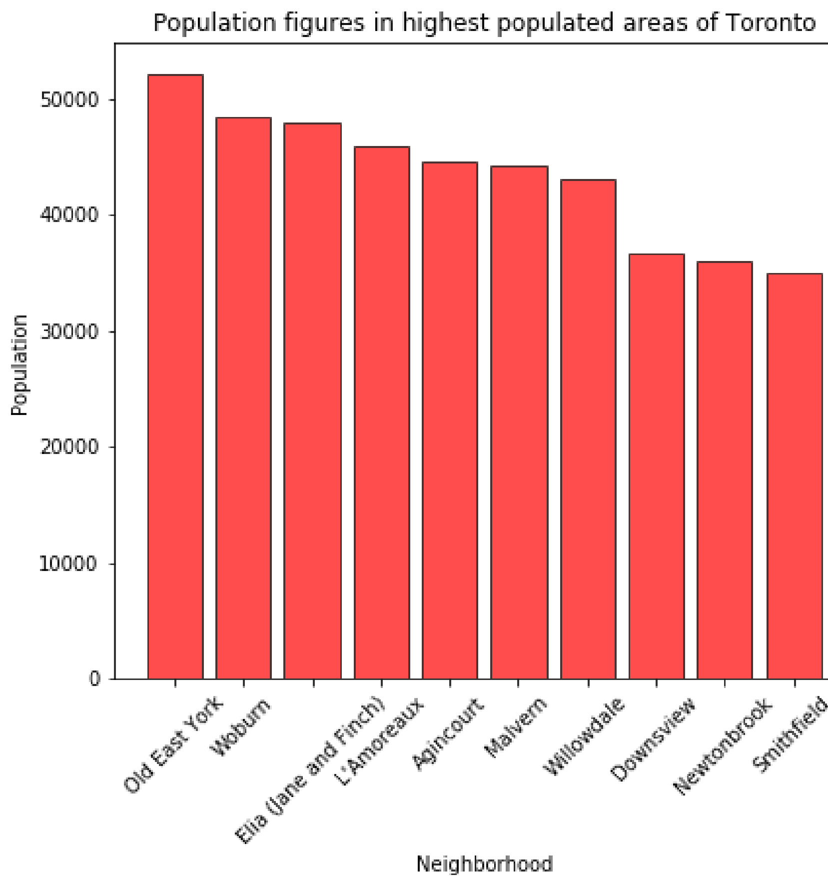


Figure 3-1 Highest populated neighbourhoods of Toronto

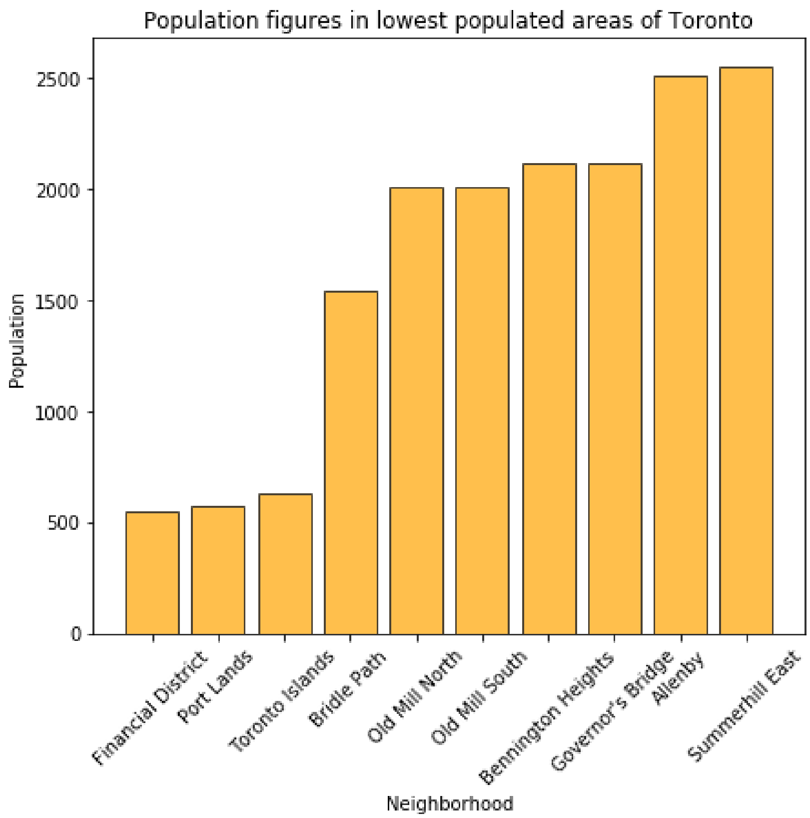


Figure 3-2 Lowest populated neighbourhoods of Toronto

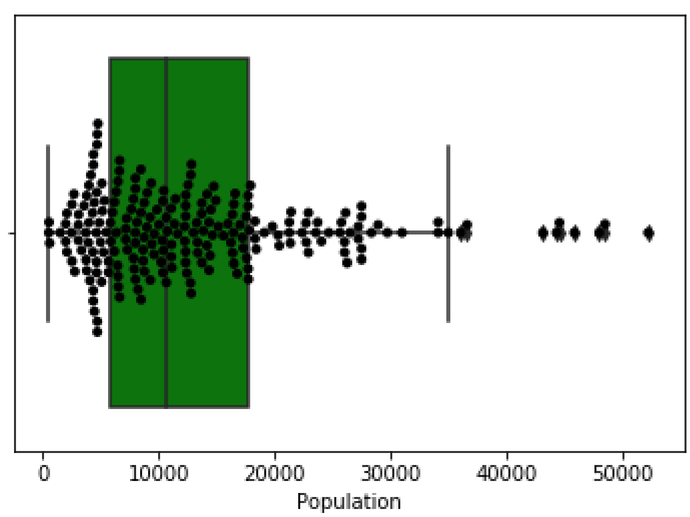


Figure 3-3 Box and swarm plot of neighbourhoods population

It can be seen from the graphs that the most of neighbourhoods have approximately 10 000 residents, with the outliers being in the 40 000+ range. The targeted neighbourhoods will be the ones with moderate-high population while still satisfying other conditions.

Average income data

The restaurant offer is targeted towards middle income customers, for this reason it is necessary to evaluate the income state of different neighbourhoods.

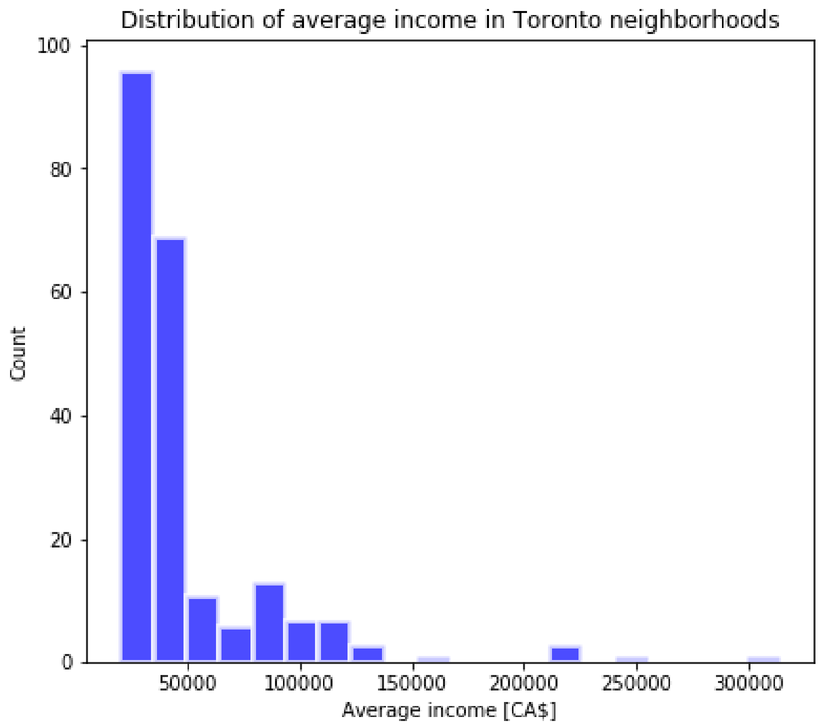


Figure 3-4 Distribution of average income in Toronto neighbourhoods

Table 3-1 Descriptive statistics of Population and Average income data

	Population	Average income
count	113.000000	113.000000
mean	14135.557522	48619.380531
std	9794.219106	33966.522451
min	627.000000	21155.000000
25%	7672.000000	28403.000000
50%	12348.000000	36361.000000
75%	17602.000000	48965.000000
max	48507.000000	214110.000000

The mean average income is in the range of 50 000 Canadian dollars with several outliers above 150 000 CA$. The target areas will cover the middle income neighbourhoods.

Geo/Cultural demographics

The demographics of each neighbourhood were determined based on the majority second spoken language besides English indicating the immigrant population in the area.

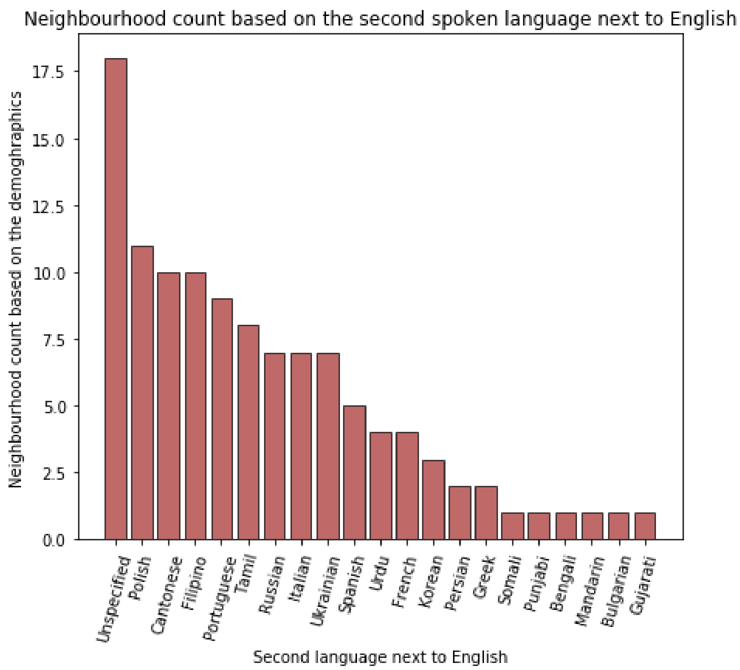
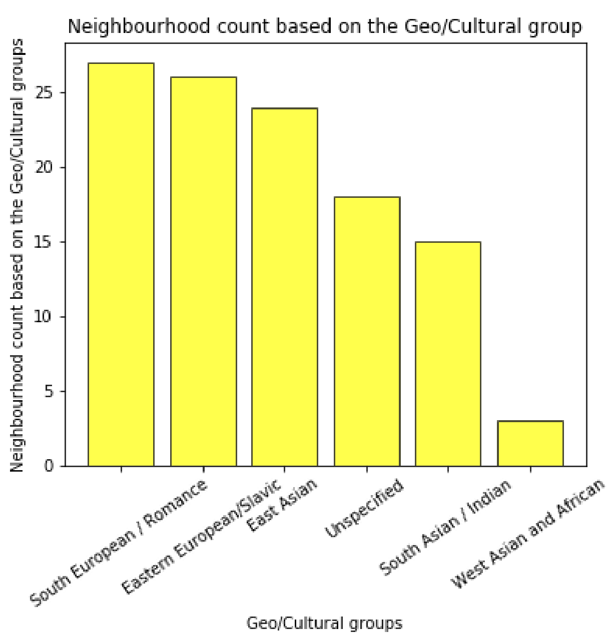


Figure 3-5 Neighbourhood count of certain language group

The method of aggregating groups was discussed in the section 2.2. See the table 2-1 for the grouping information.



Data visualisation using folium maps

Folium maps were used to visualise the demographics data. In the following maps it can be seen that there are circle markers inside circle markers, this is due to several neighbourhoods being in the same postal code, thus having the same latitude and longitude.

Each colour of the circle represents a geo/cultural group.

O – Represents "South European / Romance" geo/cultural group
O – Represents "Eastern European / Slavic" geo/cultural group
O – Represents "South Asian / Indian" geo/cultural group
O – Represents "East Asian" geo/cultural group
O – Represents "Unspecified" geo/cultural group
O – Represents "West Asian and African" geo/cultural group

The size of the circle is dependent on the evaluated parameter (i.e. Population or Average income).

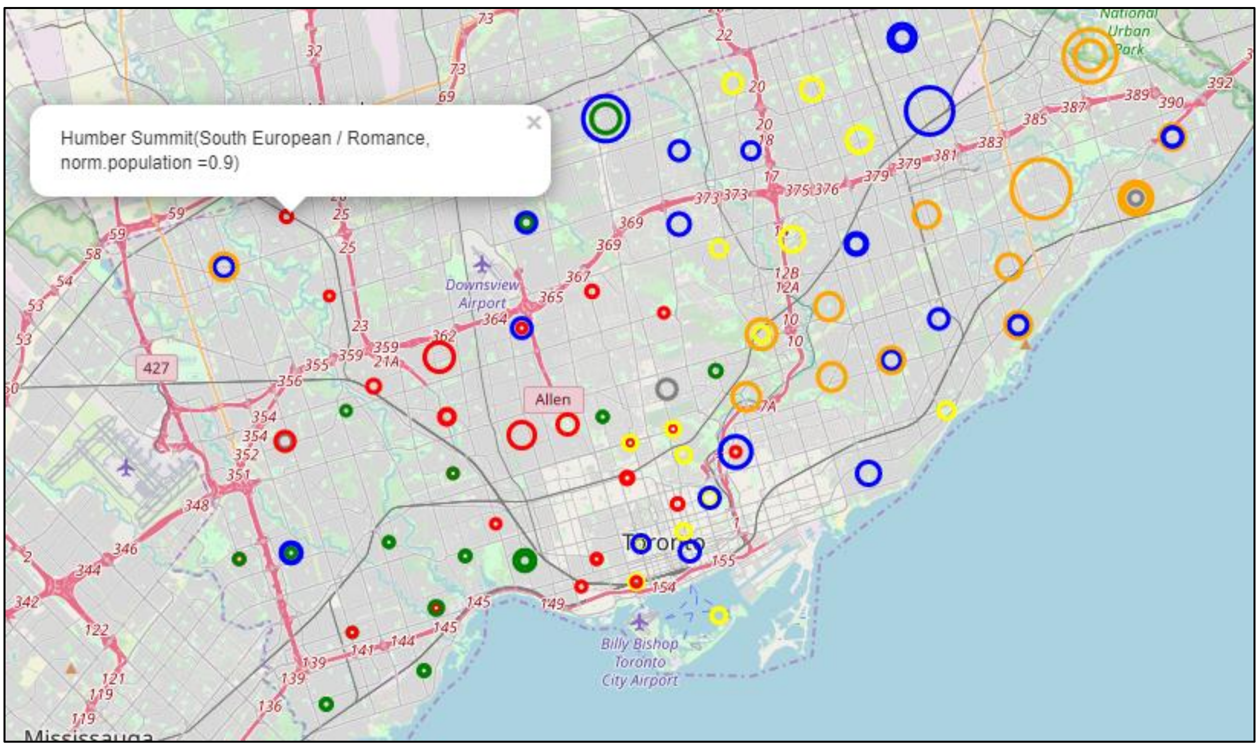


Figure 3-6 Map of Toronto representing different neighbourhoods where marker colour = Demographic group, and marker size = Population

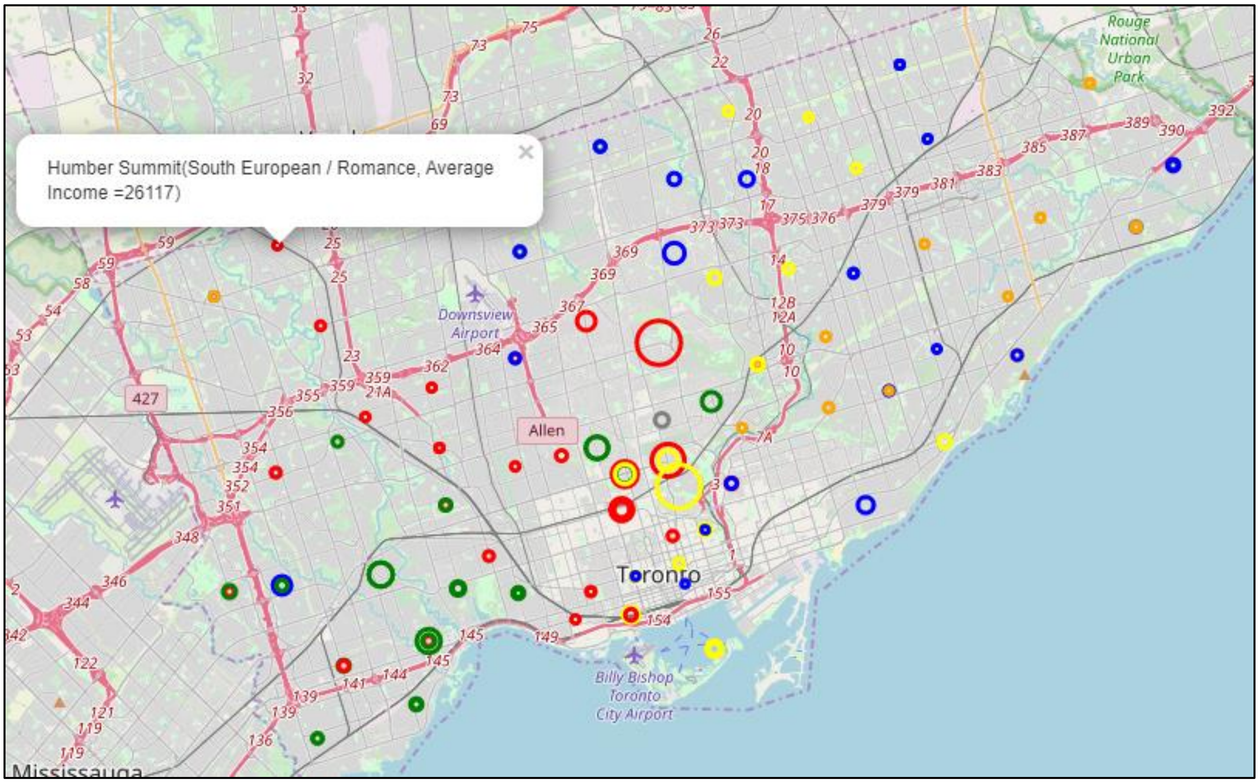


Figure 3-7 Map of Toronto representing different neighbourhoods where marker colour = Demographic group, and marker size = Average Income

4. Methodology and results

The procedure of obtaining the best location for the eastern European restaurant consisted of two steps. Where initially the choice was reduced to several restaurant locations based on the use of clustering and classification algorithms. Subsequently, the competition was evaluated in the pruned list of locations. Finally, by assigning a specific cluster to each location and utilizing logistic regression, the location with the highest probability of satisfying the need of the venue was chosen.

Step 1 consisted of:

- Assigning labels to each geo/cultural group
- Constructing the “ideal” location based on venue requirements
- Choosing the best clustering algorithm
- Clustering the neighbourhoods based on the population, average income and geo/cultural group label data
- Choosing the best classifying algorithm
- Classifying the “ideal” location based on the previously obtained clusters. By doing so we can choose the cluster where “ideal” location would be
- Filtering the locations and choosing the locations from the “ideal” cluster

Step 2 consisted of:

- Determining the competition in the filtered neighbourhoods using FourSquare API
- Filtering out locations with the competitive market
- Assigning unique labels to each remaining neighbourhood
- Classifying the "ideal" neighbourhood based on the labels of each neighbourhood using logistic regression
- Evaluating probabilities to which neighbourhood (label/cluster), "ideal" one corresponds to. Three locations with the highest probability are taken and visualized

4.1 Step 1 - Methodology

Geographical / cultural group labelling for clustering

The labels were assigned to each geo/cultural group in order to cluster the data. Zeros were assigned to everything besides Eastern European group in order to focus the preference for clustering.

Table 4-1 Group labels

	Geo/Cultural group	Language group	Label
0	South Asian / Indian	[Tamil, Gujarati, Bengali, Urdu, Punjabi]	0
1	East Asian	[Filipino, Cantonese, Mandarin, Korean]	0
2	West Asian and African	[Persian, Somali]	0
3	Unspecified	[Unspecified]	0
4	Eastern European/Slavic	[Russian, Bulgarian, Ukrainian, Polish]	1
5	South European / Romance	[Greek, French, Spanish, Portuguese, Italian]	0

The "ideal" neighbourhood construction (prediction parameter)

The ideal neighbourhood is the neighbourhood with middle average income, high population and a majority eastern European population.

The ideal population was assumed to be moderately high population.

$$P_{ideal} = P_{mean} + \frac{(P_{max} - P_{mean})}{2}$$

The ideal average income was assumed to be equal to mean average income.

$$I_{ideal} = I_{mean}$$

The ideal neighbourhood would be the one with majority eastern European population (label = 1).

$N_{ideal} = 1$

Clustering algorithm selection

Two clustering algorithms were tested.

Table 4-2 Clustering algorithms

Algorithms
KMeans
DBSCAN

Both algorithms gave identical results. Kmeans was chosen for further work

$$Init = \mathrm{k-means++}$$

$$n_{clusters} = 10$$

Classifying algorithm selection

Three classifying algorithms were tested.

Table 4-3 Classifying algorithms

Algorithms
KN-neighbours
Logistic regression
Support-vector machine

K nearest neighbour and logistic regression algorithms predicted the similar outcome with the focus on the geo/cultural demographic side of data. On the other hand, SVM with the rbf kernel predicted the outcome with the focus on the population data, by using the sigmoid kernel the predicted results were more in line with KNN and LR.

It should also be noted that KNN was highly dependent on the number of clusters that were used, with the n = 8+ results were more accurate. In the end logistic regression algorithm with the liblinear solver was used for the convenience sake.

4.2 Step 1 - Results

Classification algorithm predicted the cluster where the “ideal” neighbourhood would belong to. This cluster consisted primarily of eastern European neighbourhoods with moderate population and middle income residents. The accuracy of the classifying algorithm is assumed to be sufficient, as the similarity between the “ideal” and obtained was strongly correlated. The neighbourhoods were filtered based on the predicted results.

Figure 4-1 Location of filtered neighbourhoods

Table 4-4 Filtered neighbourhood information

PostalCode	Borough	Neighborhood	Population	Income	Second Language	Geo/Cultural group
M8Y	Etobicoke	Sunnylea	17602	51398	Polish	Eastern European/Slavic
M6R	West Toronto	Roncesvalles	15996	46820	Polish	Eastern European/Slavic
M3H	North York	Bathurst Manor	14945	34169	Russian	Eastern European/Slavic
M8W	Etobicoke	Alderwood	11656	35239	Polish	Eastern European/Slavic
M6S	West Toronto	Swansea	11133	58681	Polish	Eastern European/Slavic
M8V	Etobicoke	Humber Bay Shores	10775	39186	Russian	Eastern European/Slavic
M8V	Etobicoke	New Toronto	10455	33415	Polish	Eastern European/Slavic
M9C	Etobicoke	Markland Wood	10240	51695	Polish	Eastern European/Slavic
M8W	Etobicoke	Long Branch	9625	37288	Polish	Eastern European/Slavic
M9P	Etobicoke	Westmount	5857	35183	Ukrainian	Eastern European/Slavic
M8V	Etobicoke	Mimico South	4732	47011	Polish	Eastern European/Slavic
M8Y	Etobicoke	Mimico NE	4732	47011	Polish	Eastern European/Slavic
M8Z	Etobicoke	Mimico NW	4732	47011	Polish	Eastern European/Slavic
M9B	Etobicoke	West Deane Park	4395	41582	Ukrainian	Eastern European/Slavic
M6N	York	Runnymede	4382	42635	Ukrainian	Eastern European/Slavic
M6S	West Toronto	Runnymede	4382	42635	Ukrainian	Eastern European/Slavic

4.3 Step 2 – Methodology

Foursquare API and competing venues

After obtaining the filtered list of neighbourhoods, next step was to find out the amount of competitiveness in each. Foursquare API was used to obtain the list of venues in each neighbourhood, and filter out the eastern European venues.

Four eastern European venues were found, where 3/4 were contained in the single neighbourhood (Roncesvalles). These 2 neighbourhood will be removed from the further evaluation citing their competitive market.

Table 4-5 Neighbourhoods containing eastern European restaurants

	Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category
56	Roncesvalles	Inter Steer	43.649796	-79.450310	Eastern European Restaurant
59	Roncesvalles	Café Polonez	43.645113	-79.448517	Eastern European Restaurant
88	Roncesvalles	Chopin Restaurant	43.644165	-79.448162	Eastern European Restaurant
196	Mimico NW	Zam	43.620798	-79.528265	Eastern European Restaurant

Manual clustering of neighbourhoods for the purpose of finding the best location

Each neighbourhood was considered to be it's own cluster for the final classification step. In this way, it would be possible to determine the best location out of 14 final neighbourhoods. Labels from 0 to 13 were assigned to each area.

	PostalCode	Borough	Neighborhood	Population	Income	Second Language	Geo/Cultural group	latitude	longitude	Labels
0	M3H	North York	Bathurst Manor	14945	34169	Russian	Eastern European/Slavic	43.757875	-79.448688	0
1	M6N	York	Runnymede	4382	42635	Ukrainian	Eastern European/Slavic	43.676125	-79.481932	1
2	M6S	West Toronto	Runnymede	4382	42635	Ukrainian	Eastern European/Slavic	43.649620	-79.476141	2
3	M6S	West Toronto	Swansea	11133	58681	Polish	Eastern European/Slavic	43.649620	-79.476141	3
4	M8V	Etobicoke	Humber Bay Shores	10775	39186	Russian	Eastern European/Slavic	43.612200	-79.495146	4
5	M8V	Etobicoke	Mimico South	4732	47011	Polish	Eastern European/Slavic	43.612200	-79.495146	5
6	M8V	Etobicoke	New Toronto	10455	33415	Polish	Eastern European/Slavic	43.612200	-79.495146	6
7	M8W	Etobicoke	Alderwood	11656	35239	Polish	Eastern European/Slavic	43.601131	-79.538785	7
8	M8W	Etobicoke	Long Branch	9625	37288	Polish	Eastern European/Slavic	43.601131	-79.538785	8
9	M8Y	Etobicoke	Mimico NE	4732	47011	Polish	Eastern European/Slavic	43.632835	-79.489550	9
10	M8Y	Etobicoke	Sunnylea	17602	51398	Polish	Eastern European/Slavic	43.632835	-79.489550	10
11	M9B	Etobicoke	West Deane Park	4395	41582	Ukrainian	Eastern European/Slavic	43.650347	-79.555040	11
12	M9C	Etobicoke	Markland Wood	10240	51695	Polish	Eastern European/Slavic	43.648573	-79.578250	12
13	M9P	Etobicoke	Westmount	5857	35183	Ukrainian	Eastern European/Slavic	43.696505	-79.530252	13

Classification using logistic regression probabilities

Logistic regression was chosen as a classification tool due to the ability to predict the probabilities. Doing so, the classification of a single item can give information on good class alternatives (high probability clusters that were not chosen)

The final step consisted of using the logistic regression to predict the location with the highest similarity with the “ideal” neighbourhood. Afterwards, the probabilities were evaluated and the neighbourhood with the highest probability was chosen as the best location.

4.4 Step 2 - Results

Due to inherent similarity between the neighbourhoods, the values of probabilities were very similar. Even so, the results predicted by logistic regression were logical, in the sense that best locations were the ones closest to the objective (ideal) item (i.e. large population, and the mean income).

Table 4-6 Probability analysis of best neighbourhoods

	PostalCode	Borough	Neighborhood	Population	Income	Geo/Cultural group	Labels	Probabilities %
10	M8Y	Etobicoke	Sunnylea	17602	51398	Eastern European/Slavic	10	7.311821
3	M6S	West Toronto	Swansea	11133	58681	Eastern European/Slavic	3	7.236099
0	M3H	North York	Bathurst Manor	14945	34169	Eastern European/Slavic	0	7.207163
12	M9C	Etobicoke	Markland Wood	10240	51695	Eastern European/Slavic	12	7.196573
4	M8V	Etobicoke	Humber Bay Shores	10775	39186	Eastern European/Slavic	4	7.159521
7	M8W	Etobicoke	Alderwood	11656	35239	Eastern European/Slavic	7	7.159083
8	M8W	Etobicoke	Long Branch	9625	37288	Eastern European/Slavic	8	7.134445
6	M8V	Etobicoke	New Toronto	10455	33415	Eastern European/Slavic	6	7.133471
5	M8V	Etobicoke	Mimico South	4732	47011	Eastern European/Slavic	5	7.092504
9	M8Y	Etobicoke	Mimico NE	4732	47011	Eastern European/Slavic	9	7.092504
1	M6N	York	Runnymede	4382	42635	Eastern European/Slavic	1	7.071060
2	M6S	West Toronto	Runnymede	4382	42635	Eastern European/Slavic	2	7.071060
11	M9B	Etobicoke	West Deane Park	4395	41582	Eastern European/Slavic	11	7.067436
13	M9P	Etobicoke	Westmount	5857	35183	Eastern European/Slavic	13	7.067259

The three best locations are:

1) Sunnylea in the Etobicoke area
2) Swansea in the West Toronto area
3) Bathurst Manor in the North York area

Table 4-7 Three best locations for the eastern European restaurant

	PostalCode	Borough	Neighborhood	Population	Income	Second Language	Geo/Cultural group	latitude	longitude	Labels	Probabilities %
0	M8Y	Etobicoke	Sunnylea	17602	51398	Polish	Eastern European/Slavic	43.632835	-79.489550	10	7.311821
1	M6S	West Toronto	Swansea	11133	58681	Polish	Eastern European/Slavic	43.649620	-79.476141	3	7.236099
2	M3H	North York	Bathurst Manor	14945	34169	Russian	Eastern European/Slavic	43.757875	-79.448688	0	7.207163

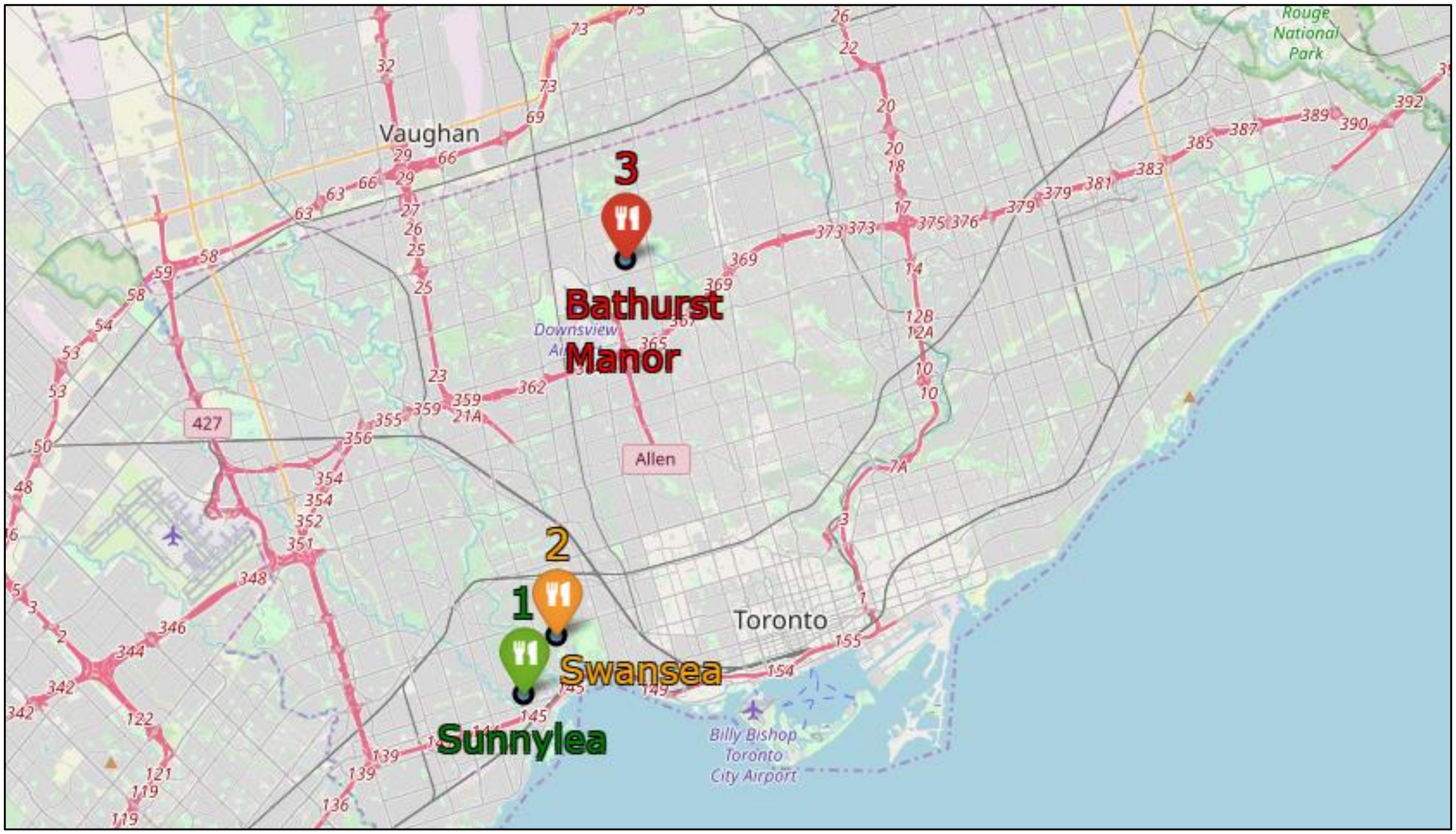


Figure 4-2 Three best locations for the eastern European restaurant visualised on the folium map

5. Conclusions

The subject of this study was to determine the best location for the eastern European restaurant in the Toronto area. The study included the demographic analysis of the Toronto neighbourhoods, including the population data, income information and cultural makeup.

Methodology of obtaining the solution for the problem included utilization of both clustering and classification algorithms in the two step procedure. Where by, in the initial step, the list of possible neighbourhoods was significantly pruned, and in the second step 3 best locations were obtained: 1) Sunnylea in the Etobicoke area, 2) Swansea in the West Toronto area and 3) Bathurst Manor in the North York area. With the Sunnylea being the best location according to this study

6. Recommendations

Some directions for the future work should include:

- Accounting for the proximity of the neighbourhoods, as this study looked in to neighbourhoods as separate entities

- More detailed grouping should be performed in order to determine the similarity of cuisines between different nationalities (as Bulgarian cuisine can be more similar to Greek than eastern European, and some other examples).

- Large portion of Neighbourhoods was not included in the study, due to faulty and unavailable data. This can be overcome with the higher workload and examination of different databases.

Literature

[1] https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

[2] https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods