Recommendation System

of

Locations for Opening a New Italian Restaurant

Coursera Capstone Project

Name: **Nouman Aftab**

Email: **nom.aftab@gmail.com**

Date: **September 3, 2019**

1. Introduction:

In the world of business, it is observed that running a restaurant is much beneficial and lasts for decades. But about 90% of the restaurant setups fail to gather required number of customers to remain in business. There are many factors which are needed to be stressed before opening a new restaurant. One of them is the “right location” for new installation which still have too many things to focus on in itself. Selection of a new location for such a business in an unfamiliar city is totally a risk, which needs too much efforts, time, expenses, collaborations, information, and decision making.

For instance, a foreign investor/businessman to the city, who want to open a new Italian restaurant, can avoid these problems by approaching a data analysis organization which work on all these factors and have a system to recommend the most likely places according to requirements.

1.1. Interest:

Local businessman would be very interested in such a recommendation system who can’t keep an eye on multiple factors constantly changing around for expanding his business. Furthermore, foreign investor as well as entrepreneur, who are altogether unaware of the society or multiple domains for location selection would be interested. Researcher on restaurants may also be interested in already built system having multiple features and visual relationships.

2. Data Description and Acquisition:

Location for opening a restaurant is the most critical feature, which in turn needs many factors to work on.

2.1. Data Description:

Different types of factors affect in selecting a location for restaurant. In this problem, best possible three locations are required to be prescribed consisted on the following data:

2.1.1. Population & Density

For a restaurant to run successfully, population factor plays and important role. More the population of an area, more will be the likelihood of customers coming to restaurant. It is in fact directly proportional to the success of a restaurant.

Type of population also plays a vital role in running a themed business, in this case, Italian food is the theme of restaurant. Though, this theme is not very specific instead liked worldwide giving it a generalized level but in the premises of Italian population there is more likelihood of success.

2.1.2. Average Income of Neighbourhood

More the income more will be the probability of people having food from outside for taste and ease. There could be more events in that specific area needing a good restaurant for food. So average income of neigbhourhoods may be a good factor to put stress on.

2.1.3. Crime Rate

Safety is another component for any business to run. Therefore, the crime rate of boroughs/neigbhourhoods is essential to be considered.

2.1.4. Recreation Places in Neighbourhood

Neighbourhood with shopping mall, university/college, or other entertainment facilities increases the chances of people to use those facilities through walk, cycling or by car/train. This increases the activity in the city especially, the foot or car traffic specifically in that neighbourhood.

2.1.5. Employment in Neighbourhood

Employment in a neighbourhood gives an insight of the likelihood of number of people that can get benefited from a restaurant. There could be a very wealthy neighbourhood but wealth could be in hands of certain people instead of general public, decreasing the probability of a successful restaurant business.

2.1.6. Commuting on any Transport

Commuting percentage on any transport is a good description of activity in the neighbourhood which in turn let the restaurant to run smoothly with good profits. Though this doesn’t include pedestrian but is a very important feature for crowd in restaurant.

2.1.7. Percentage of Relevant Population (Italian in this case)

Relevant population play an important role for any restaurant but the Italian food is world-wide acceptable so this feature can’t be considered the critical one.

2.1.8. Postal codes of Neighbourhood with Boroughs

Postal codes, neighbourhood names with boroughs are required for accessing the latitude and longitude data and detailed information. A vital feature for joining different datasets.

2.1.9. Latitude and Longitude of Neighbourhood

Latitude and longitude of neighbourhood for accessing the corresponding available data within certain radius.

2.2. Data Acquisition:

Data acquisition is done using multiple websites. Web scraping was used to extract the data from the Wikipedia and Toronto police websites. Geo spatial data, venues data to extract recreational data, neighbourhood crime data and wellbeing Toronto data for employment is downloaded into csv files. In total, six datasets are used in the project, whose links are given below:

Sr. No.	Feature	Dataset Link
1.	Population & Density	https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods
2.	Average Income of Neighbourhood	https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods
3.	Crime Rate	http://data.torontopolice.on.ca/datasets/neighbourhood-crime-rates-
4.	Recreation Places in Neighbourhood	Foursquare.com
5.	Employment in Neighbourhood	http://map.toronto.ca/wellbeing
6.	Commuting on any Transport	https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods
7.	Percentage of Relevant Population	https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods
8.	Postcode, Latitude & Longitude	http://cocl.us/Geospatial_data
9.	Postcode, Boroughs & Neighbourhood	https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
	Total Distinct Dataset Links	**6**

3. Data Cleaning:

After the data acquisition & scraping, there comes data cleaning step. In this project, different sources are used to gather data so each data is cleaned separately.

3.1. Employment Data:

The wellbeing data from Toronto downloaded had multiple columns but out of all only local employment data is required which is extracted through dataframe but the neighbourhood names have some non-alphabetic characters as well as some extra information attached like old or alternate neighbourhood name. This is required to get separated for further joining the data with other data to be extracted for all possible data availability. Therefore, the data get split and only current neighbourhood name is kept. Still the data was not in the form for joining with other data because of the placement of word “The” and “East” / ”West” etc. therefore they were removed for ease. Furthermore, the growth rate of employment in Toronto is given till year 2016 but from the year 2016 till present (2019) is 9.0 approx. as per certain statistics so it is calculated and inserted into the column.

a. For year 2016 employment growth rate ---> 1.2 %

b. For year 2017 employment growth rate ---> 3.9 %

c. For year 2018 employment growth rate ---> 1.8 %

Therefore, for consistency, employment in each neighbourhood is approximated with available employment growth rate. Employment data is saved to CS_EmploymentData.csv file.

3.2. Average Crime Data:

The file named Neighbourhood_Crime_Rates_Boundary_File_.csv which is taken from wellbeing data of Toronto, has average crime rate data which is extracted into dataframe. The neighbourhood data is extracted and processed as in step 3.1, but 'Assault_AVG', 'AutoTheft_AVG', 'BreakandEnter_AVG', 'Robbery_AVG', 'TheftOver_AVG' and 'Homicide_AVG' columns are used to find the total neighbourhood crime. The extracted and calculated data is saved into CS_CrimeData.csv file.

3.3. Population and Multiple Data:

The data for population, density, average income, transit commuting percent, second most common language percent for each neighbourhood is taken from demographics of Toronto on Wikipedia website. Extra rows are removed which don’t represent the neighbourhood and/or having no data. Second language column splits and percentage for Italian people is kept. Same process is done again for neighbourhood names and data is saved to CS_Population_Multiple_Data.csv

3.4. Borough and Neighbourhood Names Data:

Wikipedia is used to extract the data for borough and neighbourhood for information about each neighbourhood as well as is used for joining all the data sets. The rows with not assigned values in neighbourhood as well as in borough are removed where not assigned neighbourhood is assigned its borough data. The data is saved in CS_combinedNeighbourhoodData.csv file for later use.

3.5. Latitude and Longitude Data:

Above csv file is used to extract the data of neighbourhood and cognitive class labs provided geospatial data is downloaded into Geospatial_DATA.csv. Data from both is joined for respective neighbourhood latitude and longitude. The lat/long data is saved in CS_combinedNeighbourhoodLatLongData.csv file for later use.

3.6. Recreational Places Count Data:

Foursquare API is used to extract 100 venues in a radius of 500 meters from each neighbourhood. Dataframe having returned JSON data is flattened using normalization of JSON and different columns are selected. Categories are extracted through web scraping. Another column is made as recreational places count data using string matching by extracting only data which is essentially a recreational place like Spa, Bank, Art Gallery, Theater and Park etc. and their total count is set against each neighbourhood. The data is saved in CS_RecreationData.csv.

4. Data Merging:

Sometimes, one of the most difficult steps in data processing is data merging. This is because of data inconsistencies or not having a common column. Data extracted and save in CS_XYZ.csv files are read and corresponding data frames are made to get merged into a single data frame. The inner join property of sets is used to get unique neighbourhood and corresponding data entries. Because of certain mismatching in neighbourhood naming and also because of selected neighbourhood of Toronto in

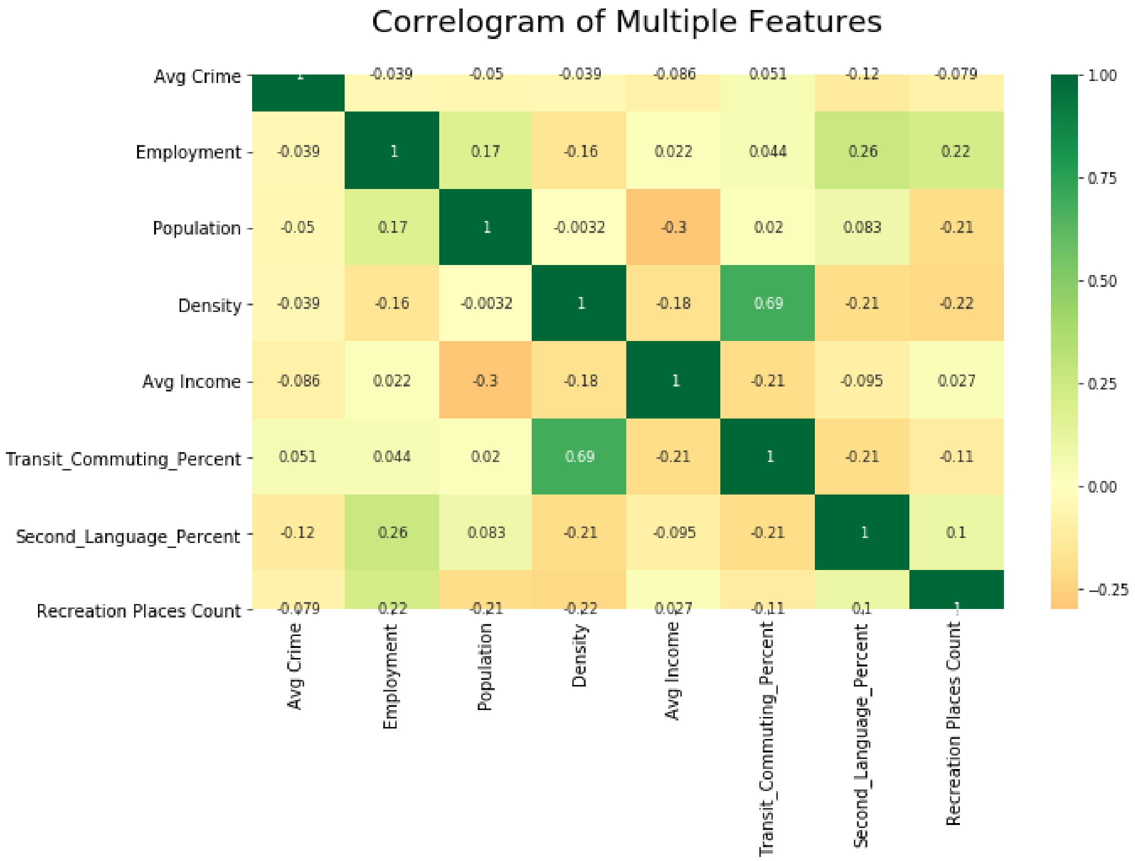
Wikipedia website source, only 60 neighbourhood data is successfully joined out of total 140 Toronto city neighbourhoods. Total neighbourhood which are unique and having corresponding assigned borough in en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M are itself 103 at minimum in any other csv file.

5. Exploratory Data Analysis

After reading the merged data, exploratory data analysis is the step to follow. The data is read and converted to data frame using pandas package.

5.1. Heatmap for Correlation and Pearson coefficient:

Correlation of features is done using heatmap plot. It gives a general insight of the relations among features. Highest positive correlation is between density and commuting transit percent where, the negative most correlation is between average income and population.

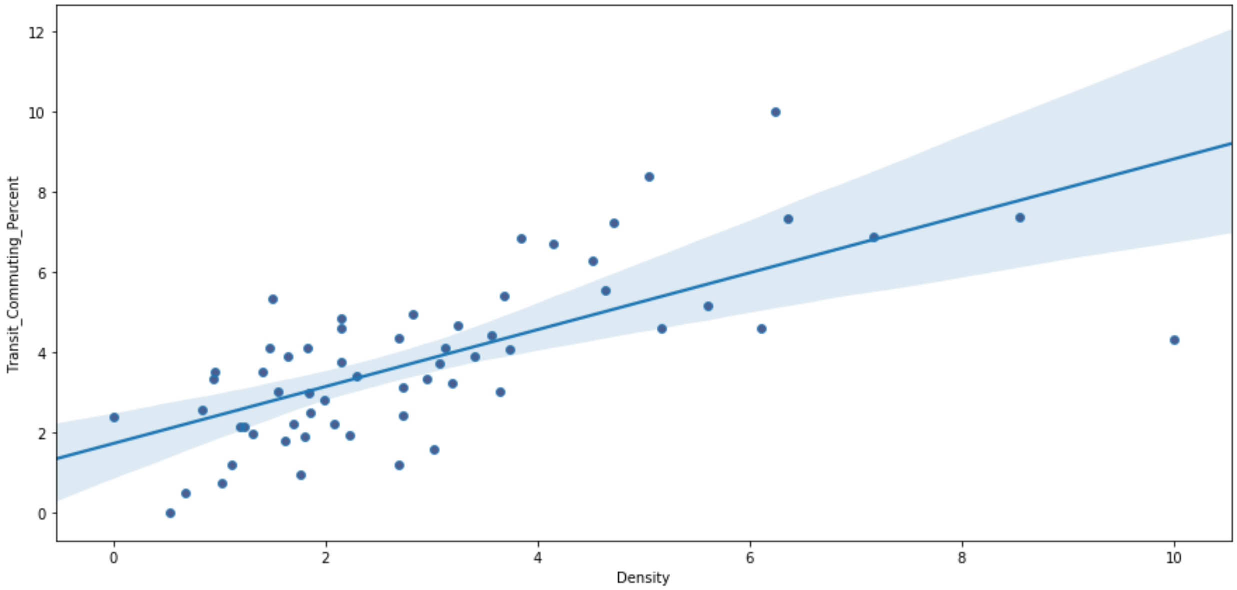


Correlation coefficient is close to 1 as it is 0.69 and p-value is 7.939413292581887e-10# which is far < 0.001 so it has a strong correlation.

On the other hand, F-value is not considerably large as only 5.514682857019 showing mean of both distributions is not far apart and the p-value is 0.02052128232031738 which is less than 0.05. Pearson correlation coefficient and p-value tells that there is strong positive correlation. Where, in ANOVA Test, F-Value and p-value tells that features are inter-dependent and strongly indistinct able. Therefore, one of them can be removed from the feature set. The one which I kept is density feature.

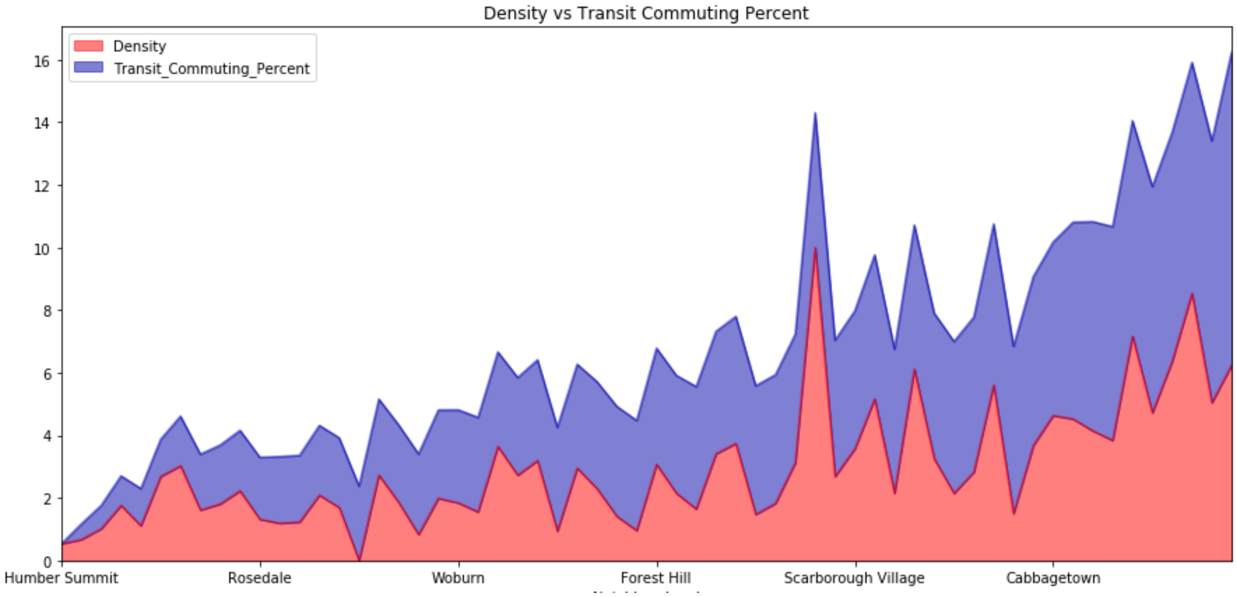
5.1.1. Pearson correlation coefficient Effects

Following figure shows the regression line in a scatter plot of the same density variable vs transit commuting percent. The effect of Pearson correlation coefficient can be considered a scatter plot where there exists a strong positive slope of the data. But this doesn't give any information about any measurable variable for distinction.



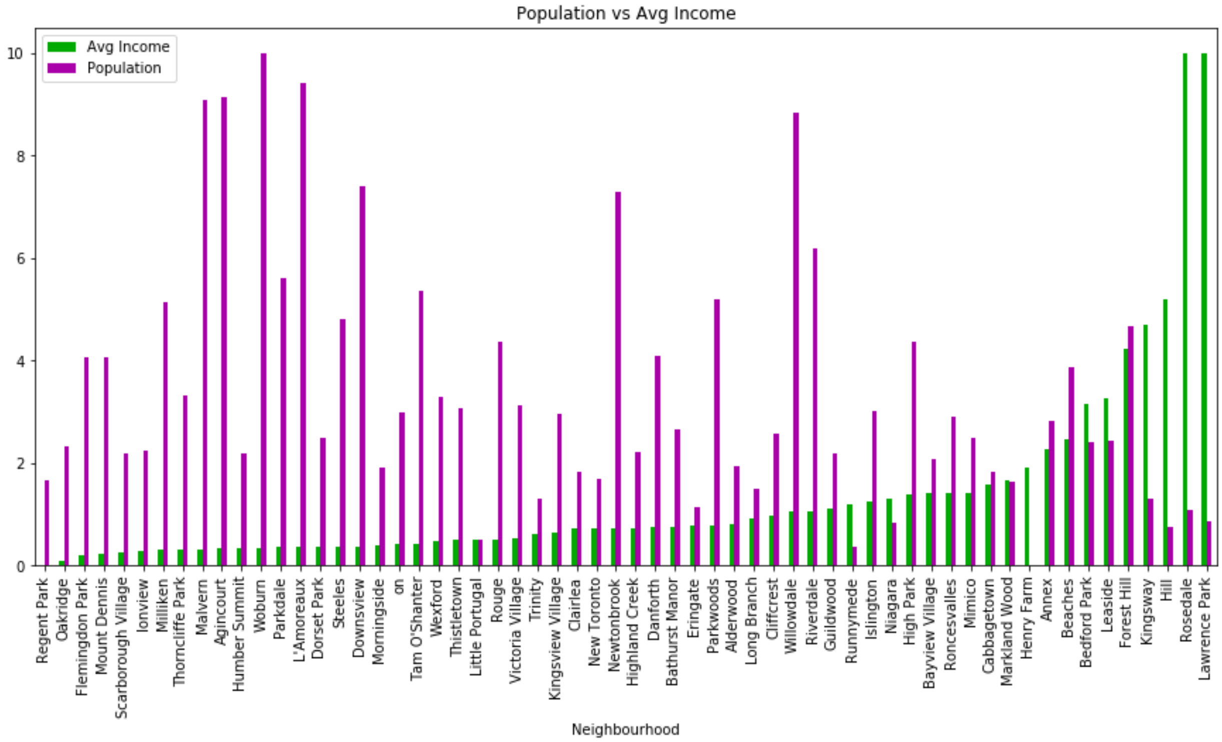
5.1.2. ANOVA Test Effects

ANOVA test can be seen as area plot shows a proportional relationship of density with transit commuting percentage of people having almost similar graph which is not clearly distinguished.



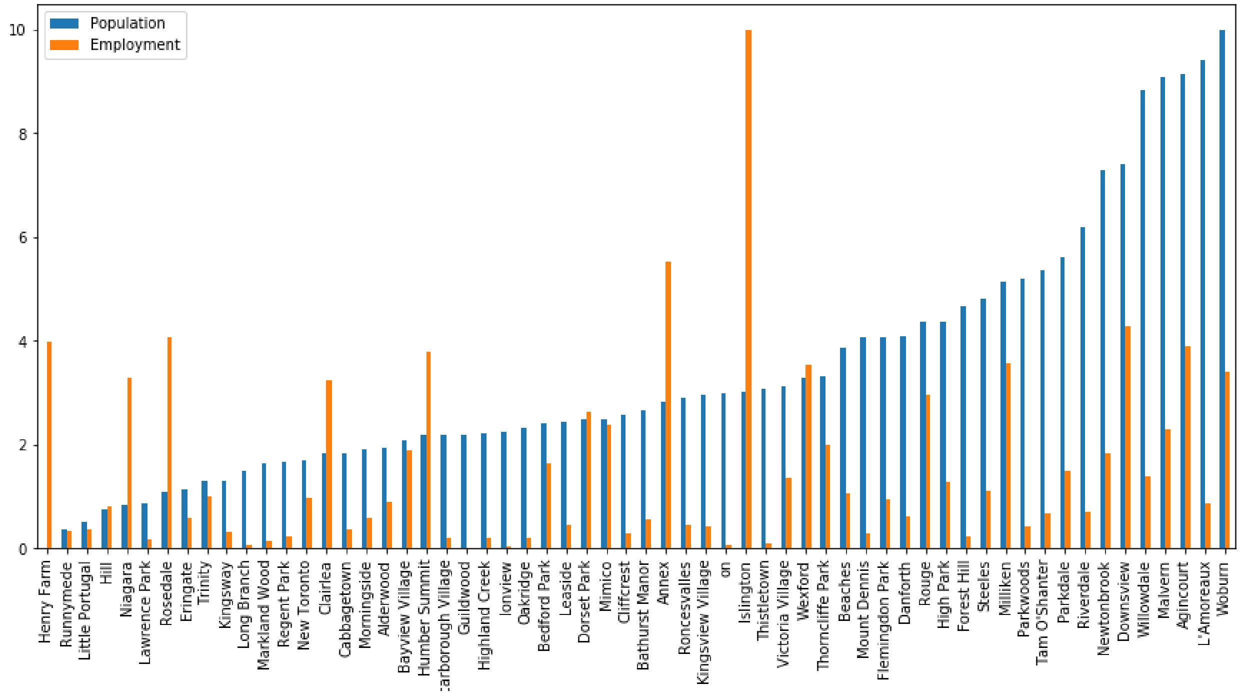
5.2. Population vs Average Income:

Average Income is more in the areas where there is less population, showing wealthy people living which could be a good location for opening a theme restaurant or with unique idea.



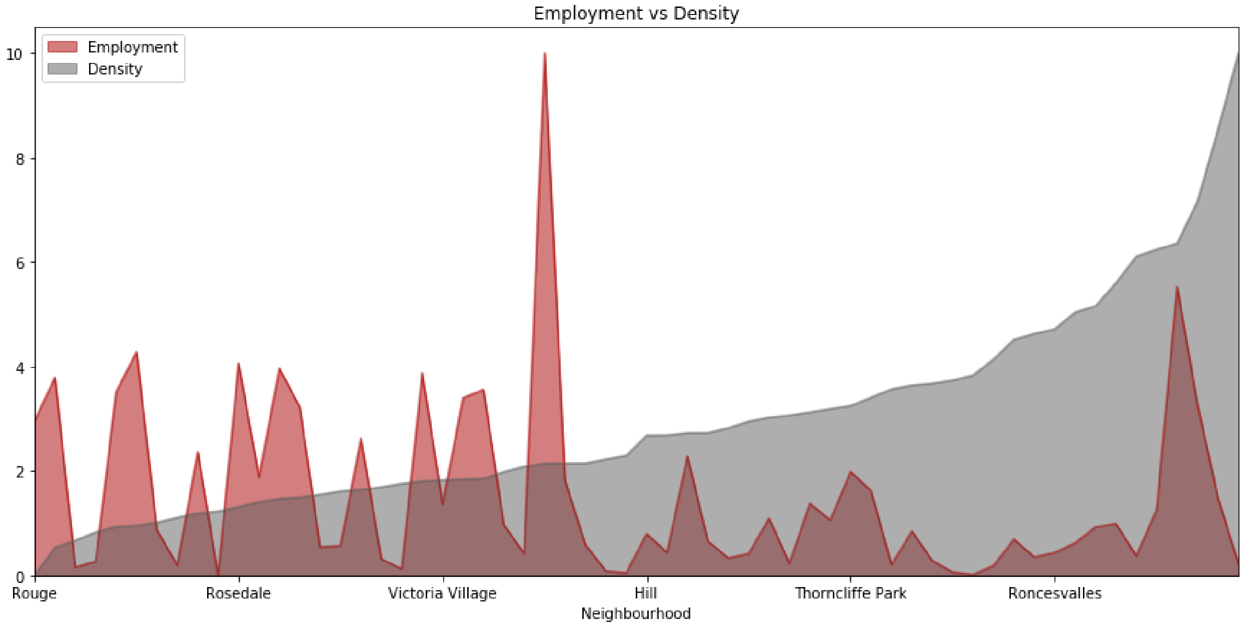
5.3. Population vs Employment:

In some Neighbourhoods, the total population is far less to the total employment therefore, those are the neighbourhoods with greater job opportunities. This gives an insight of the neighbourhood being more commercial and/or industrial areas. Those areas are very good for opening a specific time restaurant.



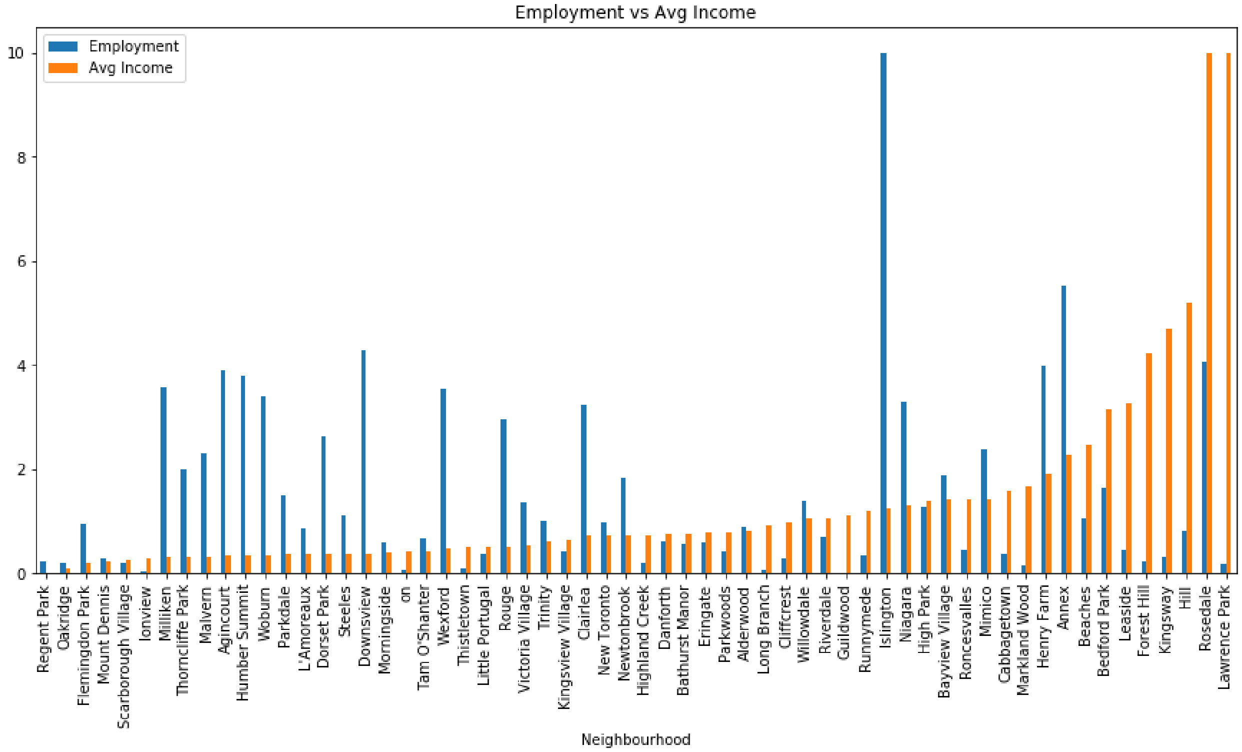
5.4. Employment vs Density

More density means there is less employment opportunities as can be observed from the graph. The jobs are fully packed but can be the ideal place to open a restaurant having more likelihood of a new restaurant to run easily at start.



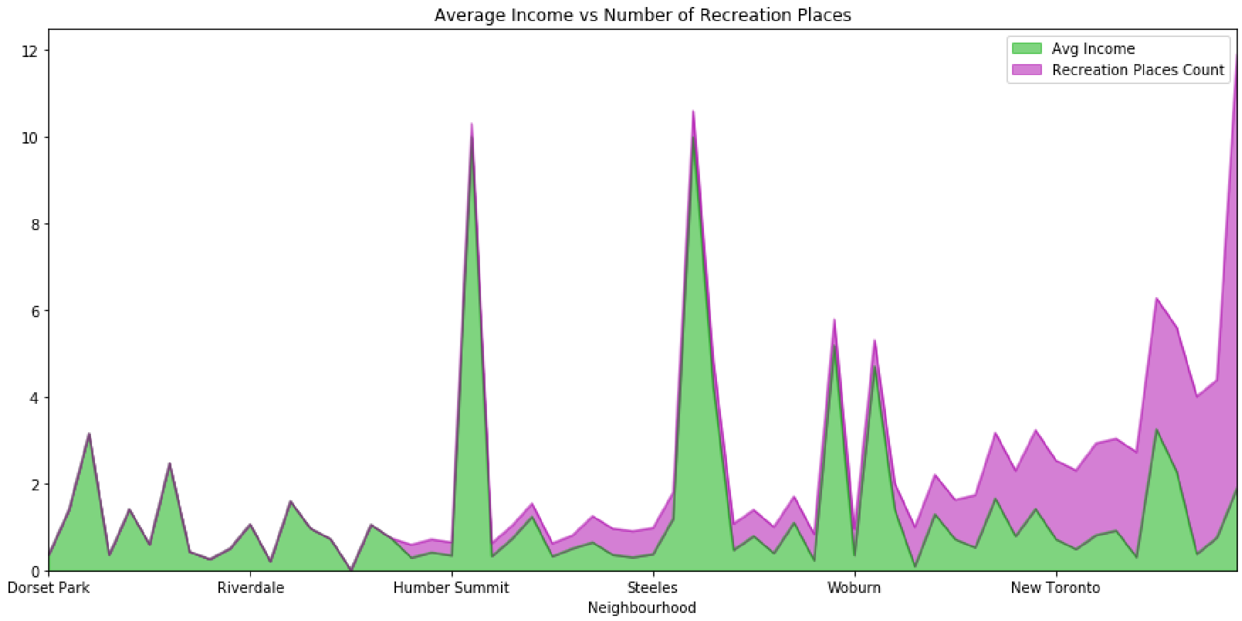
5.5. Employment vs Average Income:

With high average income areas, we can see that there is not considerable increase in the employment giving an insight of a residential area of wealthy people. For the restaurant, it may not be a very good place until with a unique attraction. Because of the reason that it also has very less population as can be seen in population vs average income.



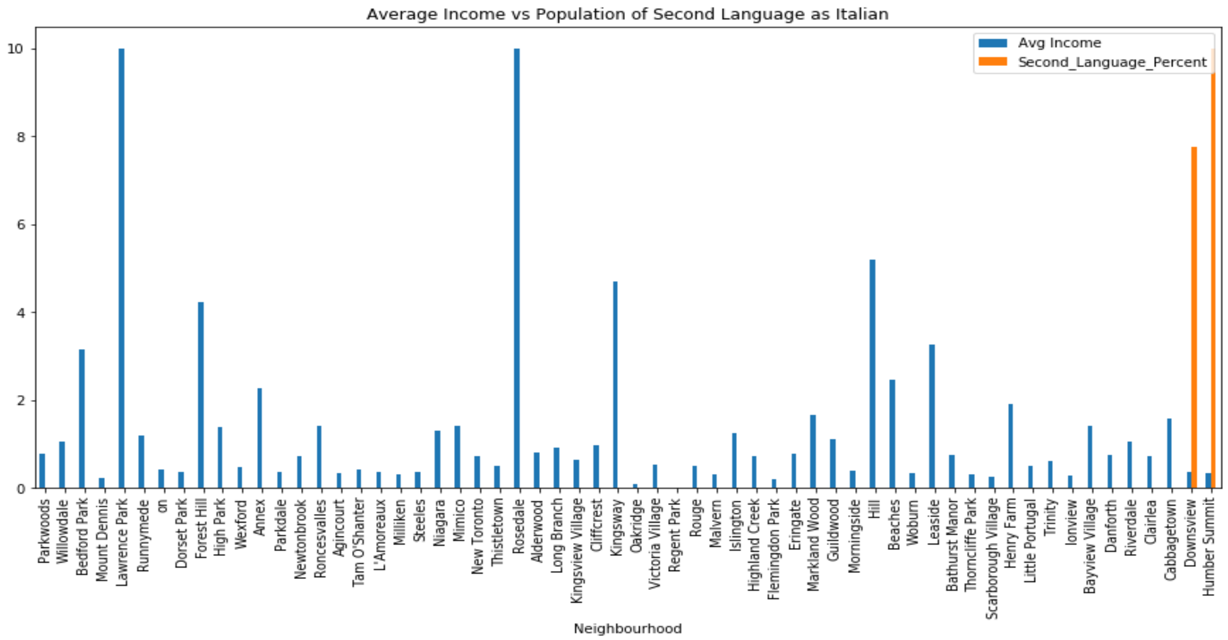
5.6. Average Income vs Total Recreation Places:

The graph shows a big difference in the average income to the total of recreation places nearby but to a certain limit. This describes the shortage of time of that neighbourhood population which in turn would be the least favorable place to open a business.



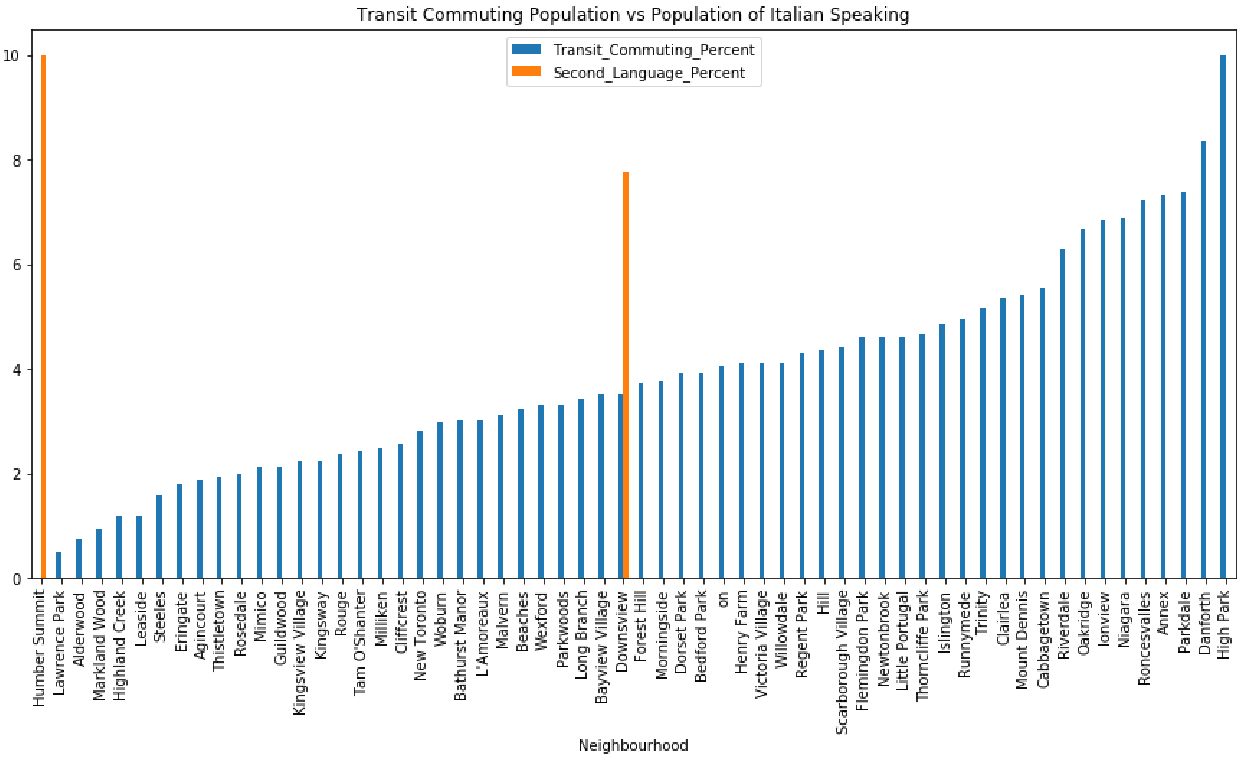
5.7. Average Income vs Second Language Percent:

The second language percent feature is describing the Italian population percentage in the neighbourhood already filtered out showing very less corresponding average income as per bar chart.



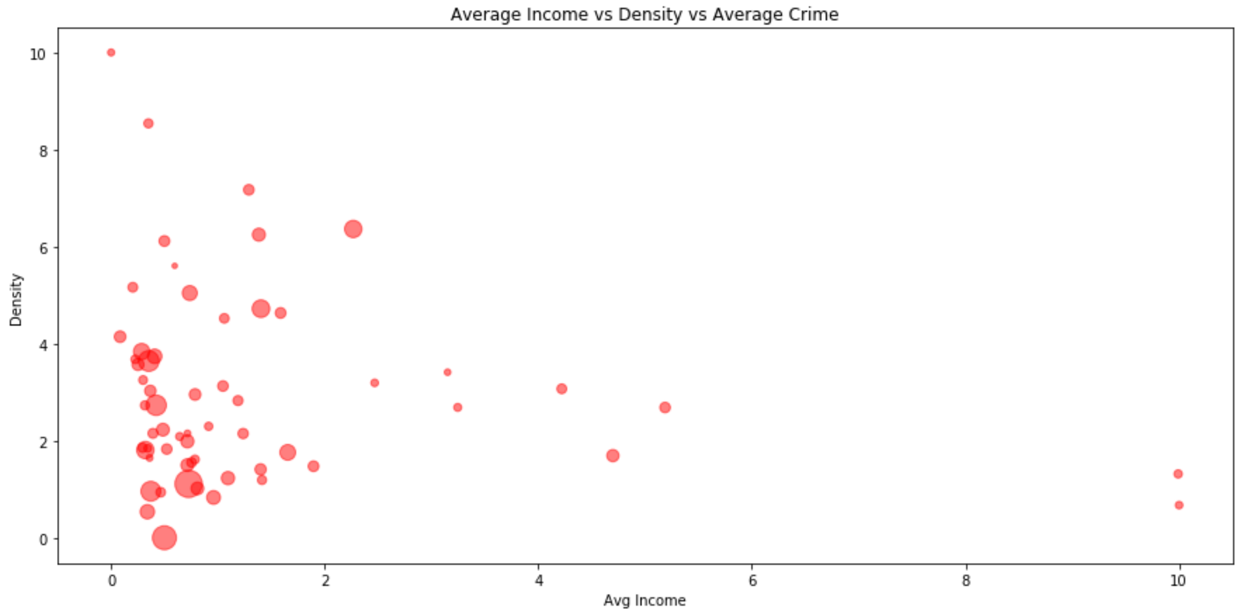
5.8. Transit Commuting Percent vs Second Language Percent:

From the graph, it is clearly visible that most of the Italians don't commute therefore the neighbourhood (in this case Downsview) should be selected where more Italians are commuting for restaurant.



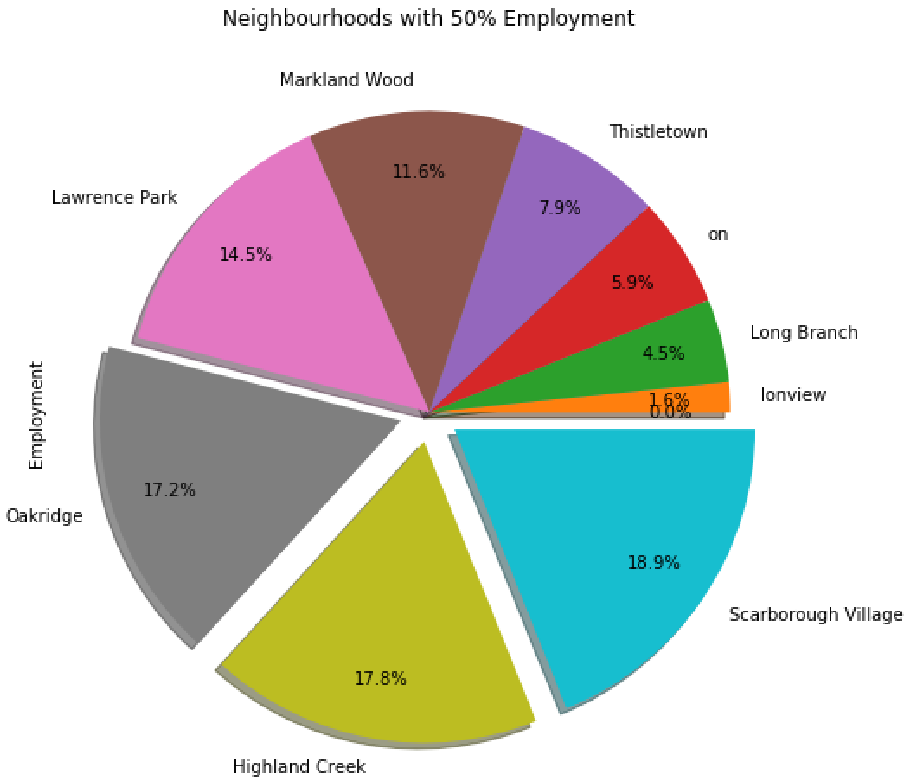
5.9. Average Income vs Density vs Average Crime:

The Bubble plot shows that the neighbourhood with low average income and low to medium density there is more crime. It can be seen as with low average income and lesser people commuting for job through transport has more criminal activities in the area having low income at the same time. Keeping in mind that density is directly proportional to Transit Commuting Percentage.



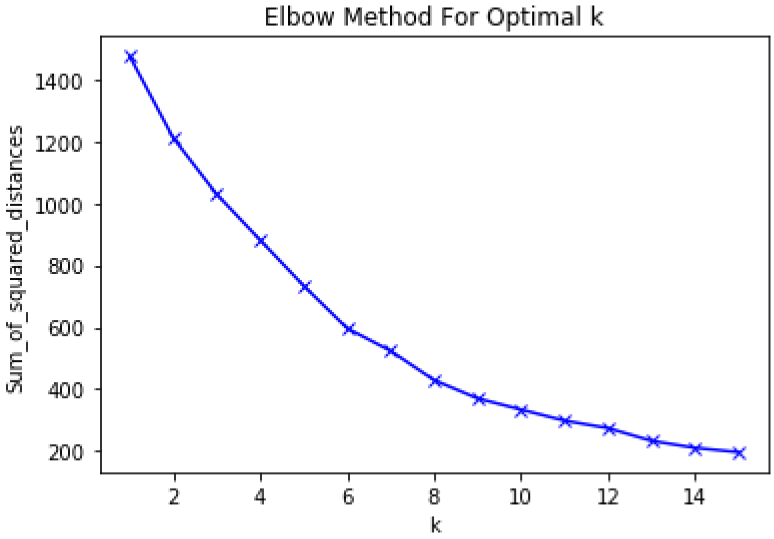
5.10. Neighbourhood Offering Major Portion of Employment:

Among all the other neighbourhoods, ten neighbourhoods cover 50% of the employment where Oakridge, Highland Creek and Scarborough Village cover even further 50% of the employment among the ten.



6. Clustering:

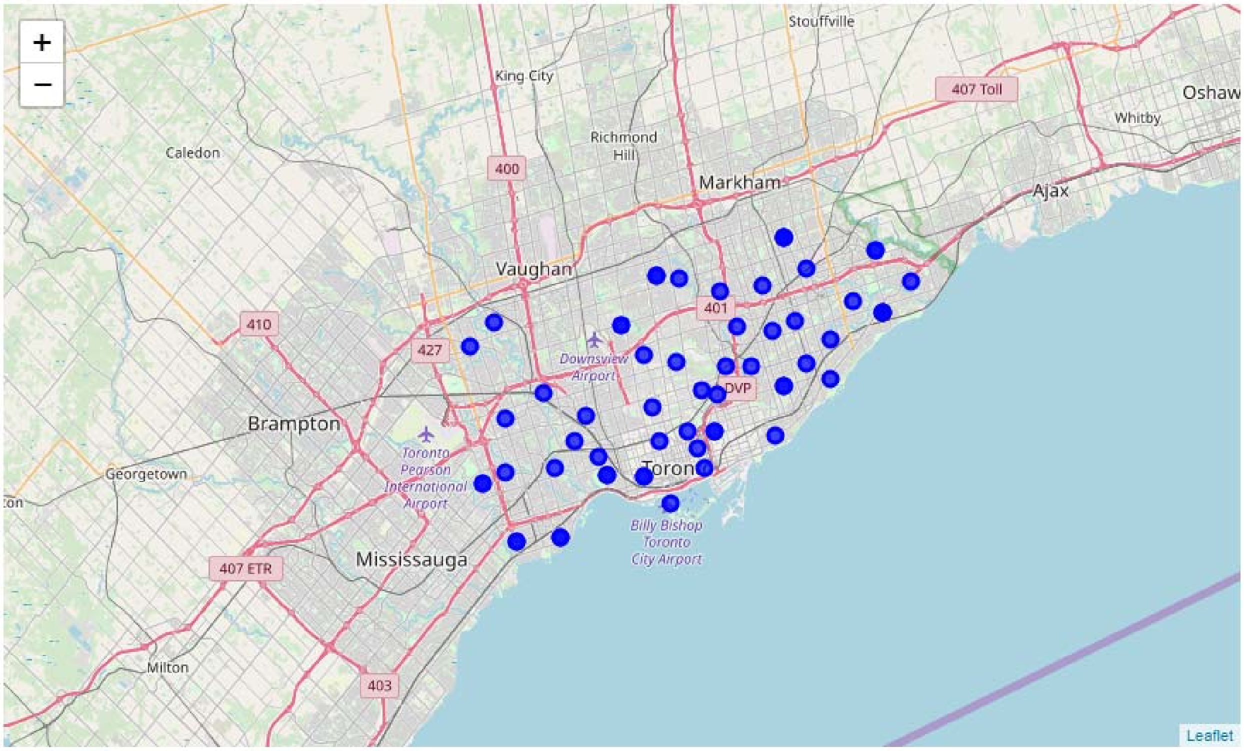
Clustering is an unsupervised learning technique in machine learning. the clusters made are based on similarity of one or more than one features. Total number of clusters “k” can be set but is manual and is set by user. The value of “k” is dependent on the type of dataset. To find out the optimal value of “k” there exist an Elbow Method, in which, multiple values of k = 1,2,3… are taken and where the elbow in the graph response is visible is taken as the optimal “k” value. In this project, the optimal “k” is taken as 6, though it is not very clearly making an elbow because of the data intrinsic nature.



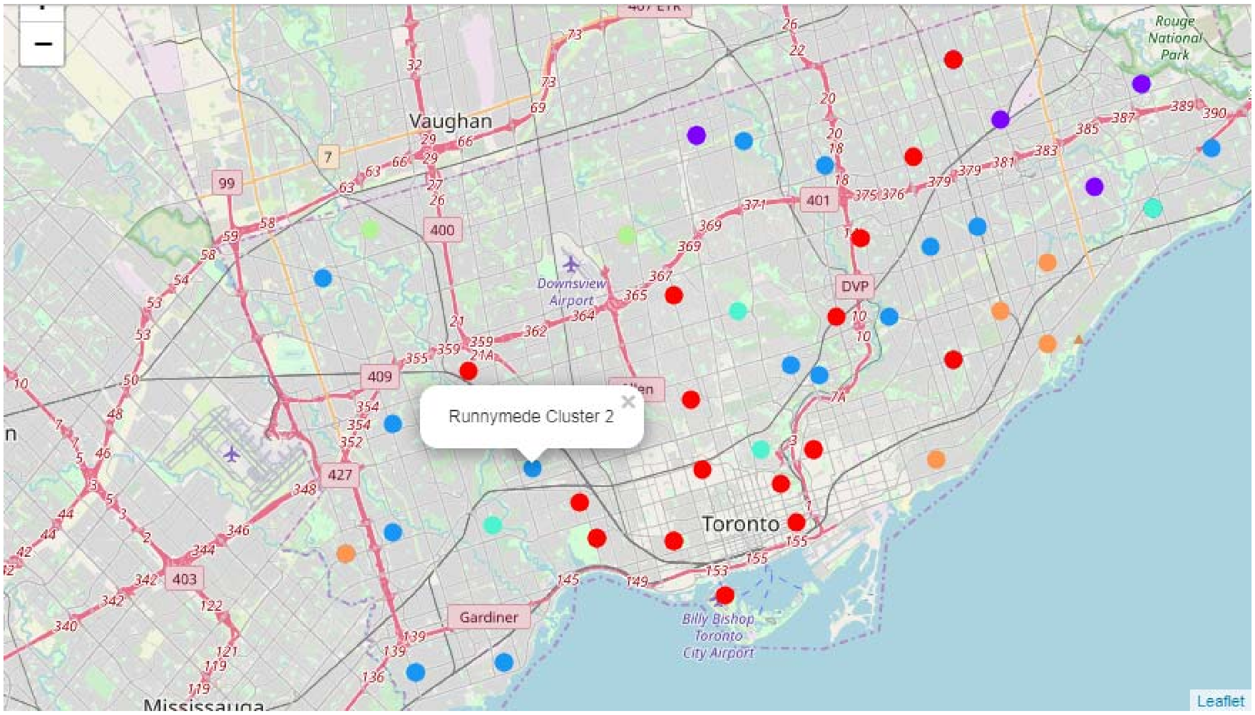
Finally, the selected “k” is used for clustering again on normalized data.

7. Toronto Map Plot:

Python package Folium is used to plot the world map besides geolocator to point out the Toronto city and its latitude-longitude for focusing on it with zoom level set at 10. Each neighbourhood having corresponding latitude and longitude is provided to folium object to plot on the map. Every neighbourhood location is represented by blue circle as can be seen:



After clustering, each cluster of possible restaurant locations is presented in distinct colors as:



8. Weighed Clustering:

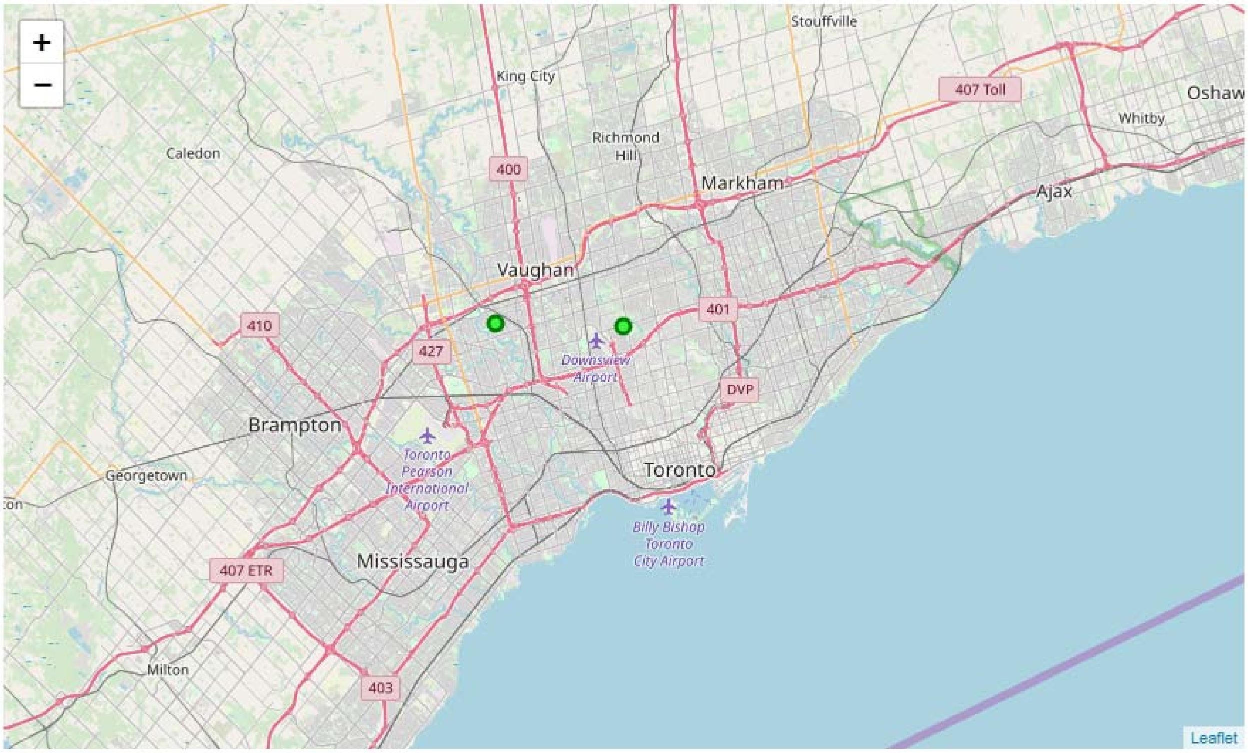
Each column is given a weight for retrieval of corresponding data. In this project, following feature requirements:

i. Required Avg Crime in Neighbourhood = low, better the safety more people for outing
ii. Required Employment in Neighbourhood = high, so people could more likely to go to restaurant
iii. Required Population in Neighbourhood = high, more population more crowd in restaurant
iv. Required Density in Neighbourhood = high, more density more crowd in restaurant
v. Required Avg Income in Neighbourhood = high, to spend more
vi. Required Second Language Percent in Neighbourhood = high, more Italian people good chance of selling
vii. Required Recreation Places in Neighbourhood = high, more pedestrian more selling

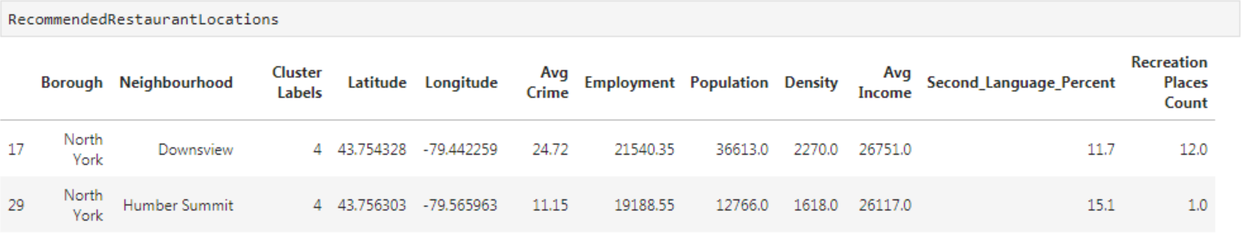
Note: Required Transit Commuting Percent in Neighbourhood = high, more activity more probability of dining (Though it is not used but worth keeping in mind)

10. Recommended Locations:

Scipy package is used to find the distance for cluster to be recommended. Another data frame is made for distance measure which is sorted and the cluster with minimum distance is selected for recommendation. In this project cluster 4 is selected for recommendation giving 2 most relevant locations both in North York and plotted as shown in green circles:



The data against the Cluster Number 4 can be seen hare:



```
RecommendedRestaurantLocations
```

	Borough	Neighbourhood	Cluster Labels	Latitude	Longitude	Avg Crime	Employment	Population	Density	Avg Income	Second_Language_Percent	Recreation Places Count
17	North York	Downsview	4	43.754328	-79.442259	24.72	21540.35	36613.0	2270.0	26751.0	11.7	12.0
29	North York	Humber Summit	4	43.756303	-79.565963	11.15	19188.55	12766.0	1618.0	26117.0	15.1	1.0

11. Conclusion:

In this project, I gathered multiple but related data to recommend the best locations for opening an Italian restaurant in Toronto. Web scraping is done, and besides latitude-longitude, 7 other features are used for clustering the data. Average Crime, employment, population, density, average income, second language percentage, and recreational places count are the features used. Total 7 csv files are created and 3 others are accessed for data storage or as a source. Data cleaning is done as a preprocessing step.

Heat map is used to get a better insight of data similarity/dissimilarity and strength. Pearson correlation coefficient and ANOVA Test are used to remove a feature. Multiple visualizations are used to plot data for analysis K-Means clustering is used to cluster the data with getting optimal number of clusters using Elbow Method. Weighted clustering is used to get requirements met of the stakeholder. Euclidean distance gave the closeness of cluster to recommend.

12. Future Work:

I got 60 neighbourhood in Toronto after utilizing the join method. They can be increased by adding more neighbourhood from other data source or manually. Further, the neighbourhood names can be made consistent manually.

Though parking place is handled in the recreation places count feature but more indirect features like real estate prices can be utilized for having an insight of total cost for own attached parking place. On the other hand, it can also be used for stakeholder’s capacity for investment. Paying special stress on the entry and exit points of crowd and transit commuting highway junctions can provide more strength to the recommendation system. Based on research, some thresholds can be made for features like in case of minimum population base a value can be set.

These are some possible measures for improving the recommendation system.