AN EVALUATION OF TORONTO NEIGHBOURHOODS TO OPTIMIZE THE LOCATION OF A CARIBBEAN THEMED RESTAURANT

Table of Contents

1.0 INTRODUCTION

The purpose of the project is to evaluate various neighbourhoods in Toronto in order to select the best neighbourhood for the client to open a new branch of a Caribbean Restaurant. The restaurant known as "George's Tastee Foods" is currently located in the Milliken neighbourhood of Toronto and the owners have requested that the location of the second restaurant is similar to the current neighbourhood with relatively low competition, a growing population base and an average income level similar to where they currently operate.

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of ~ 2.7 million in 2016 while the Toronto greater metropolitan area has a population of ~ 6 million people. Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. The diverse population of Toronto reflects its current and historical role as an important destination for immigrants to Canada. More than 50 percent of residents belong to a visible minority population group, and over 200 distinct ethnic origins are represented among its inhabitants. While the majority of Torontonians speak English as their primary language, over 160 languages are spoken in the city.

Immigrants to Canada from the Caribbean region are one of the largest non-European groups in the country, numbering over 500,000 in 2001 or 2% of the total population. This is also a fast-growing group as numbers has increased by 11% between 1996 – 2001 as compared to overall Canadian population growth of 4%. Of these immigrants who have Caribbean origins 42% were from Jamaica, 16% from Haiti, 10% from Guyana, 10% from Trinidad and Tobago and 5% from Barbados. The majority of Canadians of Caribbean origin live in either Toronto or Montreal. In 2001 60% or ~ 280,000 reported that they lived in Toronto and 20% or ~ 100,000 in Montreal comprising of 6% and 3% of the population of the respective cities.

Caribana (Toronto Caribbean Carnival)

2.0 METHODOLOGY

2.1 Sources of Data

Data for this project was obtained from public sources or free sources and included:

i) **List of Canadian Postal Codes** which contained information about neighbourhoods and their respective boroughs.

ii) **Geospatial Data** which contained a list of latitude and longitude coordinates for Toronto neighbourhoods

iii) **Demographic Data** which contained a list of neighbourhoods which information about population, population growth, income etc

iv) **Venue Data** from Foursquare which contained venues and venue categories for the list of selected Toronto neighbourhoods

2.2 Data Preprocessing

A number of steps were taken to massage the data into a format which the machine learning algorithm could process. In fact, these steps made up the majority of the code.

2.2.1 List of Canadian Postal Codes

- The data was imported into a pandas dataframe, column names were recorded on the 1st row so this was changes to column labels and the 1st row deleted.
- The spelling of neighbourhood was found to be the American version “Neighborhood”, hence this was changes to prevent compatibility issues
- Some neighbourhoods did not have a borough assigned and these were filtered from the dataframe
- The data was then grouped by Postal Code and Borough and the index was reset.

```
In [178]: df2.head()
Out[178]:
  Postal Code      Borough                              Neighbourhood
0         M1B  Scarborough                             Malvern, Rouge
1         M1C  Scarborough  Rouge Hill, Port Union, Highland Creek
2         M1E  Scarborough       Guildwood, Morningside, West Hill
3         M1G  Scarborough                                    Woburn
4         M1H  Scarborough                                 Cedarbrae
```

2.2.2 **Geospatial Data**

- The data was imported into a pandas dataframe
- This dataframe was then merged with the Postal Codes data to add the latitudinal and longitudinal coordinates for each neighbourhood
- The duplicate column “Postal Code” was dropped
- The dataframe was then stacked so that each neighbourhood was on one row, because most postal codes contained multiple neighbourhoods
- Temporary columns created during the stacking procedure were dropped

```
In [196]: df4.head()
Out[196]:
    Neighbourhood      Borough   Latitude  Longitude
0         Malvern  Scarborough  43.806686 -79.194353
1           Rouge  Scarborough  43.806686 -79.194353
2      Rouge Hill  Scarborough  43.784535 -79.160497
3      Port Union  Scarborough  43.784535 -79.160497
4  Highland Creek  Scarborough  43.784535 -79.160497
```

2.2.3 **Demographic Data**

- The data was imported into a pandas dataframe, column names were recorded on the 1st row so this was changes to column labels and the 1st row deleted.
- The spelling of neighbourhood was found to be the American version “Neighborhood”, hence this was changes to prevent compatibility issues
- Demographic Data was then combined with the merged Postal Codes and Geospatial Data
- Rows with duplicate neighbourhood names were dropped from the dataframe

```
In [198]: df5.head()
Out[198]:
     Neighbourhood    FM                                      Census Tracts  \
0        Agincourt     S  0377.01, 0377.02, 0377.03, 0377.04, 0378.02, 0...
1        Alderwood     E                                0211.00, 0212.00
2   Bathurst Manor    NY                          0297.01, 0310.01, 0310.02
3  Bayview Village    NY                                0305.01, 305.02
4     Bedford Park  OCoT                          0141.01, 0141.02, 0142.00

  Population Land area (km2) Density (people/km2)  \
0      44577           12.45                3580
1      11656            4.94                2360
2      14945            4.69                3187
3      12280            4.14                2966
4      13749            2.27                6057

  % Change in Population since 2001 Average Income Transit Commuting %  \
0                               4.6          25750                11.1
1                              -4.0          35239                 8.8
2                              12.3          34169                13.4
3                              41.6          46752                14.4
4                              -1.4          80827                15.2

  % Renters Second most common language (after English) by name  \
0       5.9                                Cantonese (19.3%)
1       8.5                                    Polish (6.2%)
2      18.6                                   Russian (9.5%)
3      15.6                                 Cantonese (8.4%)
4      10.1                                     Greek (0.7%)

  Second most common language (after English) by percentage  Map      Borough  \
0                                  19.3% Cantonese         NaN  Scarborough
1                                     06.2% Polish         NaN    Etobicoke
2                                    09.5% Russian         NaN   North York
3                                   08.4% Cantonese         NaN   North York
4                                       00.7% Greek         NaN   North York

    Latitude  Longitude
0  43.794200 -79.262029
1  43.602414 -79.543484
2  43.754328 -79.442259
3  43.786947 -79.385975
4  43.733283 -79.419750
```

2.2.4 **Venue Data**

- A custom function which utilized the Haversine formula was used to determine the distance between two points via their latitudinal and longitudinal coordinates. For each neighbourhood in the combined dataframe the distance to every other neighbourhood was calculated to determine which was the closest. This nearest distance was then divided by 2 to determine the radius for the venues search for each neighbourhood to ensure there was no overlapping search areas. This was also done to ensure neighbourhoods in suburban areas has a larger search radius. This information was added to the combined dataframe.
- Once the custom radius information was calculated for each neighbourhood another custom function was used to perform an API call to Foursquare to retrieve venue name and category information based on the coordinated of each neighbourhood. This information was written to a dataframe.

- The venues dataframe was then grouped by neighbourhood so that the number of venues per neighbourhood could be visually inspected
- Onehot encoding was used on the venue categories dataframe to convert it to a format which could be read by the KMeans algorithm. This information was then grouped by neighbourhood to develop a mean score for each venue category per neighbourhood.
- The index of the grouped onehot array was then reset

```
In [200]: toronto_venues.head()
Out[200]:
  Neighbourhood  Neighbourhood Latitude  Neighbourhood Longitude  \
0     Agincourt                43.7942               -79.262029
1     Agincourt                43.7942               -79.262029
2     Agincourt                43.7942               -79.262029
3     Agincourt                43.7942               -79.262029
4     Agincourt                43.7942               -79.262029

                          Venue  Venue Latitude  Venue Longitude  \
0                   The Roti Hut       43.787277       -79.258724
1                    Mona's Roti       43.791613       -79.251015
2        Babu Catering & Take Out       43.791721       -79.251132
3  Fahmee Bakery & Jamaican Foods       43.810170       -79.280113
4                   Strength-N-U       43.784888       -79.251685

         Venue Category
0   Caribbean Restaurant
1   Caribbean Restaurant
2  Sri Lankan Restaurant
3   Caribbean Restaurant
4   Gym / Fitness Center
```

```
In [201]: toronto_grouped.head()
Out[201]:
     Neighbourhood  Afghan Restaurant  Airport  American Restaurant  \
0        Agincourt                0.0     0.00                 0.01
1        Alderwood                0.0     0.00                 0.00
2   Bathurst Manor                0.0     0.01                 0.01
3  Bayview Village                0.0     0.00                 0.00
4     Bedford Park                0.0     0.01                 0.01

   Amphitheater  Art Gallery  Arts & Crafts Store  Asian Restaurant  \
0           0.0          0.0                 0.01              0.01
1           0.0          0.0                 0.01              0.00
2           0.0          0.0                 0.00              0.01
3           0.0          0.0                 0.00              0.00
4           0.0          0.0                 0.01              0.00

   Athletics & Sports  Auto Dealership  Automotive Shop  BBQ Joint  \
0                0.00              0.0              0.0       0.00
1                0.00              0.0              0.0       0.01
2                0.02              0.0              0.0       0.00
3                0.00              0.0              0.0       0.00
4                0.00              0.0              0.0       0.00
```

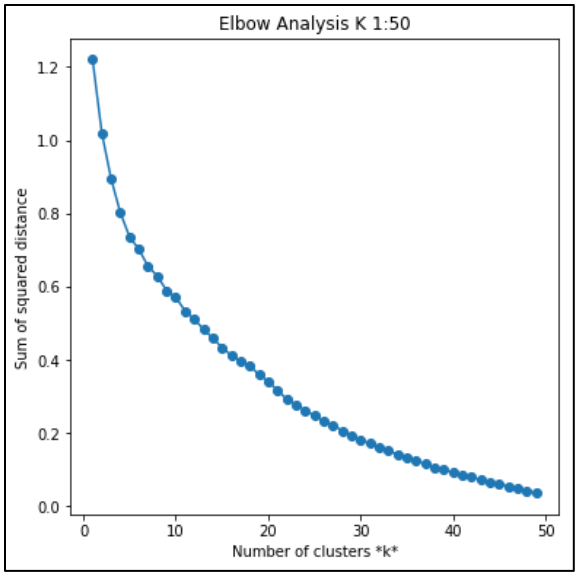
2.3 Machine Learning Algorithm

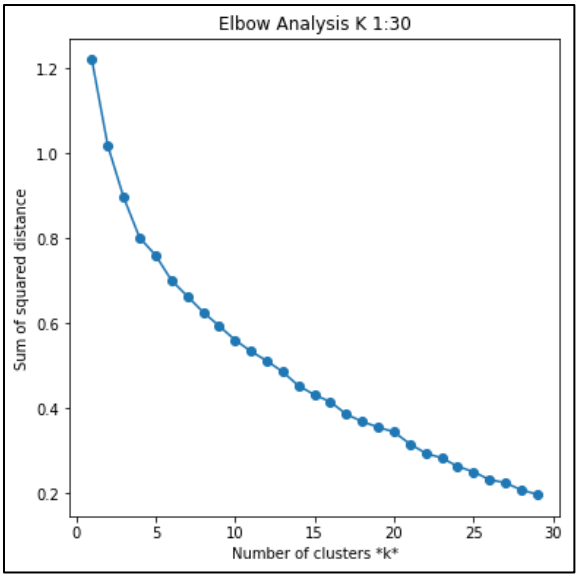
The Kmeans algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way Kmeans algorithm works is as follows:

- Specify number of clusters K.
- Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

In this instance the KMeans algorithm was used to partition the neighbourhoods into K groups based on the score for each category in the grouped dataframe. Here the goal was to find neighbourhoods similar to the one where the client is currently operating. The elbow method was used to determine the optimal number for the K hyperparameter, however results were inconclusive hence a value of 10 was used in the analysis.





2.4 Evaluation Metrics

Once the KMeans algorithm had classified the neighbourhoods the main dataframe was filtered to only include neighbourhoods in the same cluster as Milliken, the neighbourhood in which the client currently operates a restaurant. The venue category score for the category “Caribbean Restaurant” was then added and sorted by score to determine the neighbourhoods where competition was expected to be low. (Where the score was low or zero due to a small number of venues in that category existing in the neighbourhood). The demographics data was then used to discriminate between the remaining neighbourhoods. Three metrics were used; Total Population, % change in population since the last census and average income. The values were ranked/normalized between 0 – 1 across a linear scale from the maximum to minimum for each metric. The score was then added for each row/neighbourhood and the one with the best score was selected.

$$Metric\ Score = \frac{value - \min value}{\max value - \min value}$$