The Battle of Neighbourhoods — Coursera Capstone Project
OPENING A SALAD BAR IN TORONTO

ADEKALU ADEDAYO
April 17, 2021

1. Introduction

1.1 Background

What is the best location for salad bar in Toronto?
In what Neighbourhood should I open a salad bar to have the best chance of becoming profitable as soon as possible?
Bright and bustling, Toronto is a cosmopolitan city whose residents have roots across the globe. Art, food, beaches, nightlife – in Toronto, you've got it all. It is a destination for young talent from all over Canada and the globe.
The main objective of this project is to pick a profitable location for a new low calorie salad bar that focuses on young fitness enthusiasts that already show an existing commitment to fitness

1.2 Contributing Data

For this project we need the following data:

a. Toronto city data that contains Borough, Neighbourhoods along with their latitudes and longitudes

Data Source: Toronto location data frame from the previous week that was developed using the Toronto postal code wiki page and the geospatial data csv Description: This data set contains the required information. And we will use this data set to explore various neighbourhoods of Toronto.

b. Gyms in Toronto. Since our target market is fitness lovers, setting up the business in areas saturated with fitness centres will increase the likely hood of attracting the right clientele

Data Source: Foursquare API Description: By using this API we will get all available venues in the city and filter it to include on venues in "venue category"= Gym then narrow it down to the neighbourhood with the highest number of gyms and lowest number.

a. Interest

Data driven entrepreneurs looking to establish a healthy food restaurant and has identified his preferred clientele as fitness enthusiasts will are the target for this project.

c. Data acquisition and cleaning

a. General Approach

Collect the Toronto city data from the previous week's notebook.
Using Foursquare API, we will get all venues for each neighbourhood then narrow down to neighbourhoods in the borough with the highest gyms

The feature to be extracted
The "star" feature to be extracted is the number of Gyms in the each in each neighbourhood, the frequency data will enable to discover the neighbourhood with the highest. This will be done by one-

hot encoding the venue data gotten directly from foursquare

The option of franchising
clustering based on the singular feature that is gyms, the goal is find similar neighbourhood that can serve as upcoming location should the business attempt to grow

Data Visualization and some statistical analysis.
Analysing using Clustering (Specially K-Means):

Find the best value of K
Visualize the neighbourhoods with a number of Gyms Restaurants. Compare the Neighbourhoods to Find the Best Place for Starting up a Restaurant. Inference From these Results and related Conclusions

b. Data acquisition and cleaning

Data downloaded from the Cousera site and scraped from table from the wiki page 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M' were combined into one table. I was able to avoid missing values by avoiding the scraping of postal codes with non-assigned boroughs/ neighbourhoods.

```
df.head()
```

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Queen's Park	Ontario Provincial Government

Figure 1 The wiki data converted to a Pandas data frame

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.81	-79.19
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.78	-79.16
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.76	-79.19
3	M1G	Scarborough	Woburn	43.77	-79.22
4	M1H	Scarborough	Cedarbrae	43.77	-79.24

Figure 2 Merged with the coursera provided csv file

Nearby venues were derived in JSON format then converted into a Pandas data frame and the important columns filtered out

nue.location.address	venue.location.crossStreet	venue.location.lat	venue.location.lng	venue.location.labeledLatLngs	v
1530 Albion Rd	Albion Mall	43.74	-79.58	[{'label': 'display', 'lat': 43.741685, 'lng':...	
80-1530 Albion Rd	at Kipling Ave. (Albion Centre)	43.74	-79.58	[{'label': 'display', 'lat': 43.74120870478487...	
6210 Finch Ave West, Store 103	at Albion Rd.	43.74	-79.59	[{'label': 'display', 'lat': 43.74264512142215...	

Figure 3 data as derived from Foursquare API

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.81	-79.19	Wendy's	43.81	-79.20	Fast Food Restaurant
1	Rouge Hill, Port Union, Highland Creek	43.78	-79.16	Chris Effects Painting	43.78	-79.16	Construction & Landscaping
2	Rouge Hill, Port Union, Highland Creek	43.78	-79.16	Royal Canadian Legion	43.78	-79.16	Bar
3	Guildwood, Morningside, West Hill	43.76	-79.19	RBC Royal Bank	43.77	-79.19	Bank
4	Guildwood, Morningside, West Hill	43.76	-79.19	G & G Electronics	43.77	-79.19	Electronics Store

Figure 4 cleaned foursquare data

.

c. Feature selection

By One hot encoding the venue categories, I determined that the data set of interest had 274 features.

	Accessories Store	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum
0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0

Figure 5 encoded data

However only one feature was clearly stated in the objective as the target audience i.e., the density of Gyms, thus that feature was filtered out of the dataset and sorted in descending order of magnitude to provide the best location, this was determined to be the Willowdale, Newton brook in the North York borough

	Neighborhood	Gym
95	Willowdale, Newtonbrook	0.50
21	Don Mills North	0.25
1	Alderwood, Long Branch	0.14
22	Don Mills South	0.10
58	New Toronto, Mimico South, Humber Bay Shores	0.08

Figure 6 Gym data by frequency'

d. Clustering

In order to determine similar neighbourhoods in the event of company franchising desired by the executives of the target audience, the neighbourhoods were clustered with gym density as the only criteria

a. K-means Clustering

i. Applying standard algorithms and their problems

I applied the Kmeans Clustering algorithm using Create a function that calculates Inertia for n times by sweeping through 1 to n to find the optimal cluster number. The optimal cluster number was

determined to be 4 and that was utilize to create clusters, there were 83 neighbourhoods that were highly suited for expansion.

```
done sorting
end inner loop
Iteration 1, inertia 0.00016526477255519856
center shift 0.000000e+00 within tolerance 3.506871e-07
```

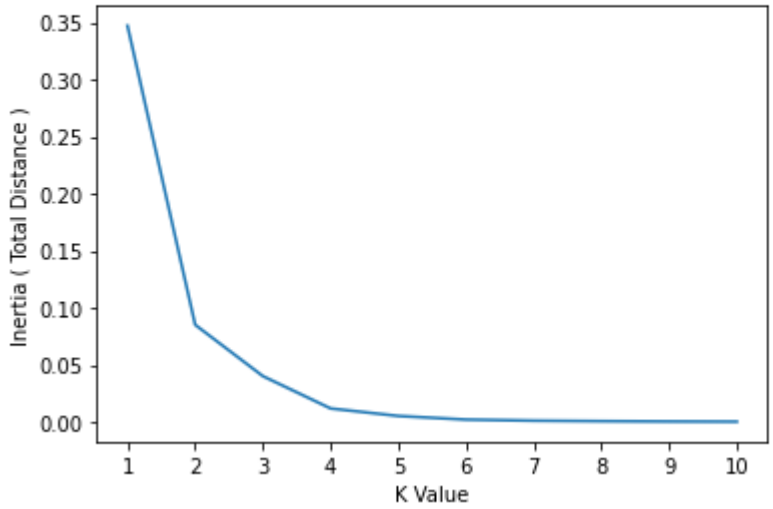

Figure 7 Selecting the K value

ii. Visualisation of Solutions

I created a folium map with the cluster points as markers to visualise the clusters on the Toronto map,

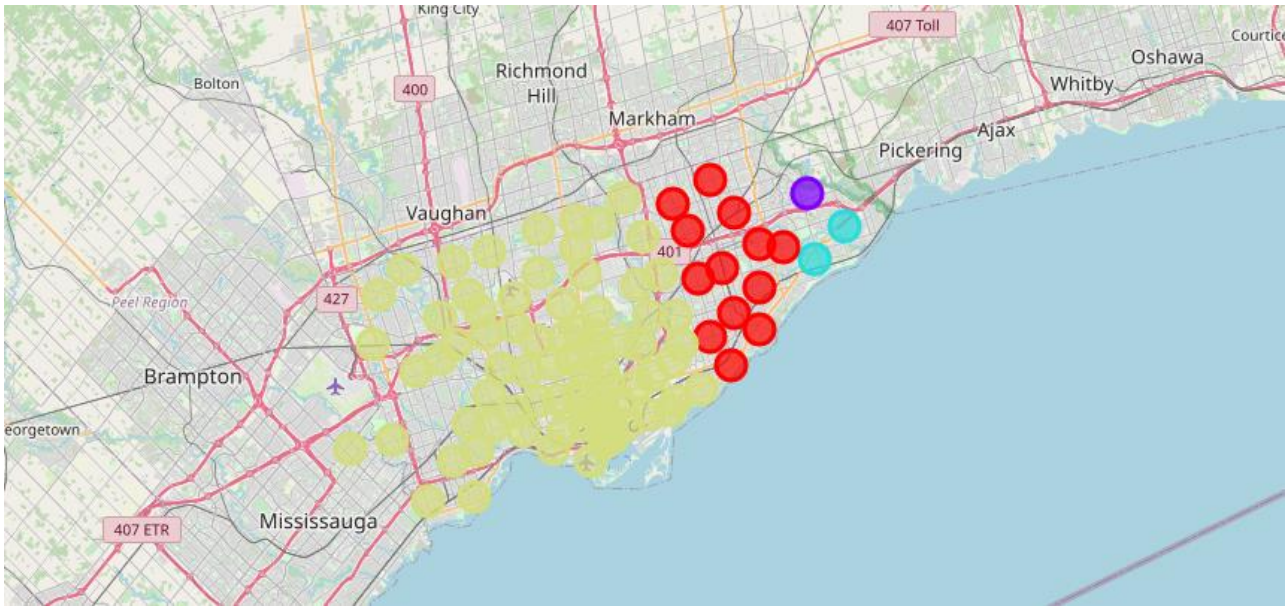

Figure 8 Map of Toronto showing clusters

e. Conclusions

The client is best advised to open their first venue in Willowdale, Newton brook, most of Toronto's neighbourhood appear to have a high density of gyms and the client is advised to open venues in order of gym frequencies in cluster 4

f. Opportunities for Improvement

a. Increasing the number of features utilized would have that would have improved the precision of our clustering, this could be made possible by
    i. The inclusion of gym adjacent venues such as yoga studios and sports centres
    ii. The study of similar businesses such as other low calories salad bars and juice places hoping to selected cluster based on the inverse relationship between all forms of similar and competing business
    iii. This would result in the ideal location being one with a high density of fitness related facilities (gyms, yoga studios, sports centres) and a low density of healthy food options.

b. Further exploratory analysis to discover relationships between venue categories, preferably to use this to determine features to be selected in combination with the Gyms

c. The use of multiple clustering algorithms Such as mean shifting, hierarchical clustering method among others. This would have allowed for accuracy comparisons between mode