SETTING UP A NEW AFRICAN RESTAURANT IN A TORONTO NEIGHBORHOOD

Samuel Onekutu

06 May 2021

Table of contents

Introduction: Business Problem

A successful Restaurant business board from Nigeria_Africa is looking to break into the market space around The Canadian capital Toronto , given that only very few African restaurant exist in the area, being foreign investor; they want to know what area is will have less competition and importantly a viable market
They have approached a data scientist for data back recommendations before taking the leap into the unknown;

Data

Data sourced from **Wikipedia web page** using webscrapping and venue information from **Four Square API** and **lastly Geodata file**, all this was combined/cleansed into a location data for the Toronto Area. to get : Neighborhood around Toronto, Venues in each neighborhood, total Restaurant per hood and total venues per hood half akilometer from neighborhood centers
This location Data will be used to create different models, after determining number of restaurants per neighborhood as the dependent variable, this is excluded and the rest of the data it clustered and compared against the dependent variable for correlation and tested again for best model that can accurately find viable neighborhoods with lest competition Some models to use include Linear Regression and Kmeans Clustering

Data was collected from wikipedia.org using beautiful soup library in python result below shows to 4 roles

After inserting the latitudes and longitudes then filtering only boroughs

	Postal Code	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

around Toronto we get,

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M5A	Downtown Toronto	Regent Park , Harbourfront	43.654260	-79.360636
1	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
2	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
3	M4E	East Toronto	The Beaches	43.676357	-79.293031
4	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306

I thereafter used the Four Square API call to get venues 500meters around the neighborhood centers, I sourced the complete data I needed

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park , Harbourfront	43.65426	-79.36063	Roselle Desserts	43.6534	-79.3620	Bakery
1	Regent Park , Harbourfront	43.65426	-79.36063	Tandem Coffee	43.6535	-79.3618	Coffee Shop
2	Regent Park , Harbourfront	43.65426	-79.36063	Cooper Koo Family YMCA	43.6532	-79.3580	Distribution Center
3	Regent Park , Harbourfront	43.65426	-79.36063	Body Blitz Spa East	43.6547	-79.3598	Spa
4	Regent Park , Harbourfront	43.65426	-79.36063	Impact Kitchen	43.6563	-79.3569	Restaurant

METHODOLOGY

two major factors are put to play that is:
commercially viable neighborhoods*¶
area with least competition in restaurant*¶
in this project with the data available and gathered by me,(four squared venue data) I intend use three models to determine the best location for a new African restaurant based on the two factor above

- first to run a *Regression Model* to determine a relationship or correlation between thriving commercial hood and number of restaurants
- second cluster neighborhood together base on their venue setup similarities then

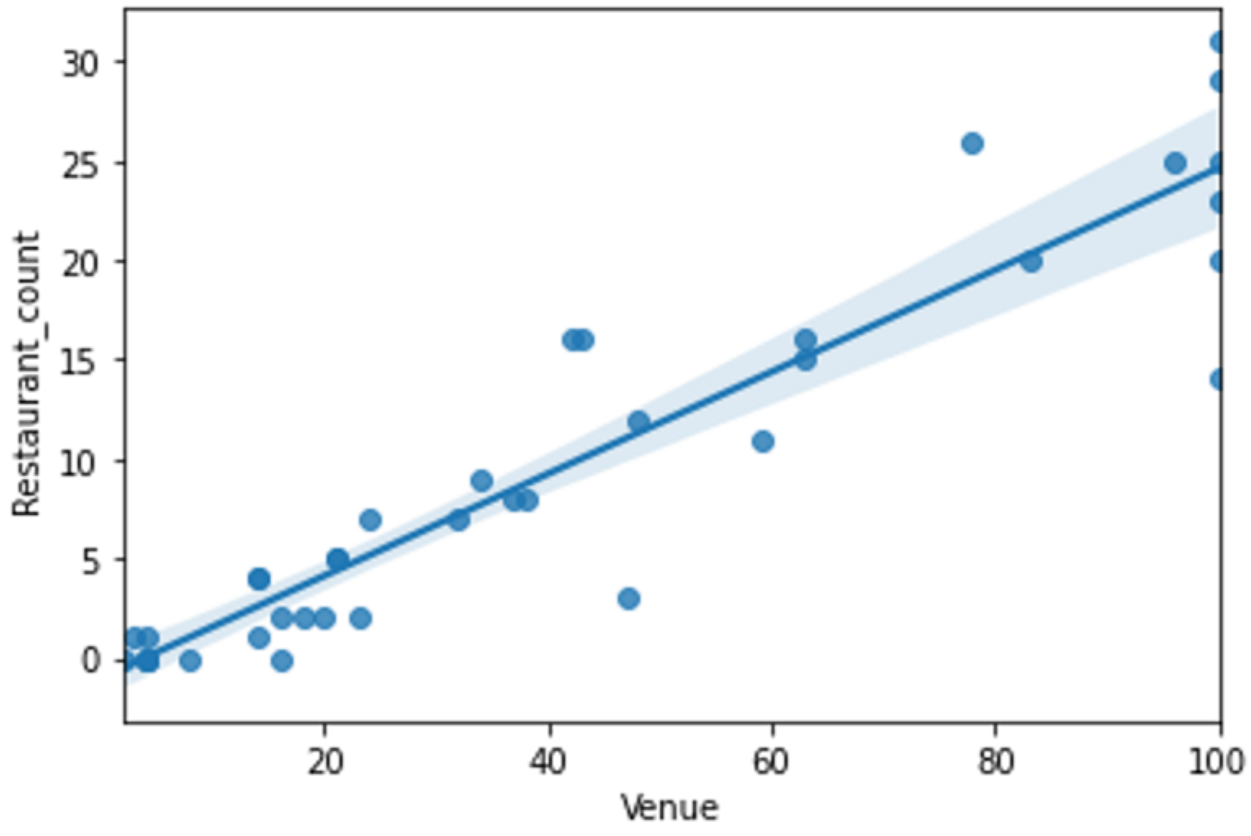
- Thirdly to combine the effects of the first two results there by creation a cluster of neighborhoods with commercial activities and similarity of venue

all this done excluding the **Dependent y** *variable we know to be number of restaurants per Hood* so that we can evaluate our models accuracy testing how well it can predict our dependent variable: the number of restaurants the best model is found and used to get results

DATA ANALYSIS:

Below are the 3 models

1. *Linear Regression: is there a correlation between no. venues and no. of restaurants. we will merge and plot to see on restaurants count and venue count per neighborhood*

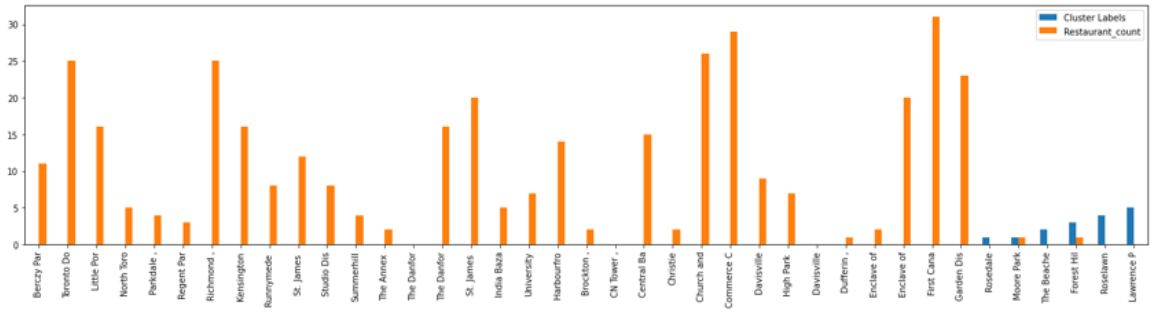


	Neighborhood	Venue	Restaurant_count
0	Berczy Park	59	11.0
1	Brockton , Parkdale Village , Exhibition Place	23	2.0
2	CN Tower , King and Spadina , Railway Lands ...	16	NaN
3	Central Bay Street	63	15.0
4	Christie	16	2.0

with Correlation at **0.91** we can say except for the few out layers that we on to something lets saves this result on go on to use

2. CLUSTERING to further analysis trends between Neighborhood Similarities and Number of Restaurants
because restaurant y is our independent variable lets remove all 45 of it

We first employ the use of get_dummies to transpose each venue category into columns and compute a ratio of that venue category amongst the rest of the venues .this was done so the clustering algorithm can work



	Neighborhood	Museum	Adult Boutique	Airport	Airport Fod Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	Antique Shop	...	Tanning Salon	Tea Room	Tennis Court	Theater	Toy / Game Store	Trail	Train Station	Video Game Store	Wine Bar	Yoga Studio
0	Berczy Park	0.020	0.0	0.000	0.000	0.000	0.000	0.000	0.000	0.0	...	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.000
1	Brockton , Parkdale Village , Exhibition Place	0.000	0.0	0.000	0.000	0.000	0.000	0.000	0.000	0.0	...	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.047
2	CN Tower , King and Spadina , Railway Lands ...	0.000	0.0	0.062	0.062	0.062	0.125	0.187	0.125	0.0	...	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.000
3	Central Bay Street	0.000	0.0	0.000	0.000	0.000	0.000	0.000	0.000	0.0	...	0.0	0.020	0.0	0.0	0.0	0.0	0.0	0.0	0.020	0.020
4	Christie	0.000	0.0	0.000	0.000	0.000	0.000	0.000	0.000	0.0	...	0.0	0.000	0.0	0.0	0.0	0.0	0.0	0.0	0.000	0.000

At 6 clusters I got the optimal results

Out of the six clusters 0 shows the most promise but the results are to open and unlike accurate so.

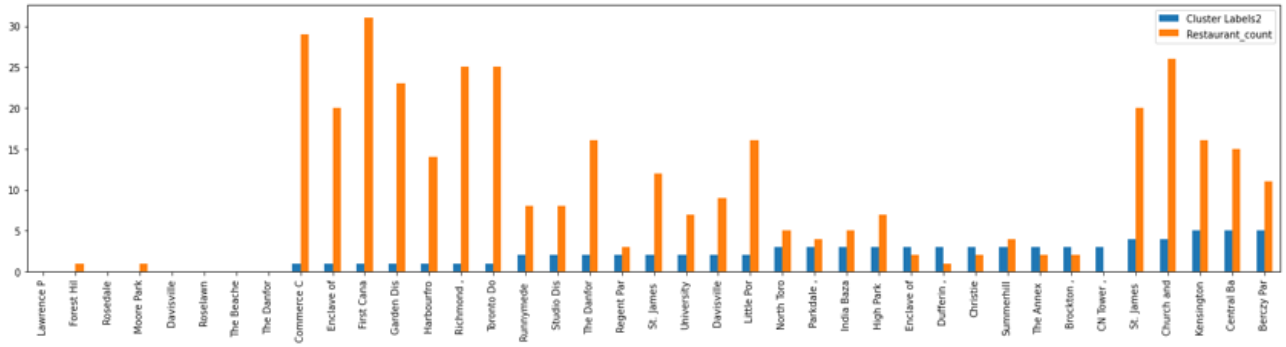
3 **COMBINED_CLUSTER**

some trend is appearing, now what if we factor total number on venues like we found earlier to our clustering algorithm combine both neighborhood venue similarity and number total number of venues...

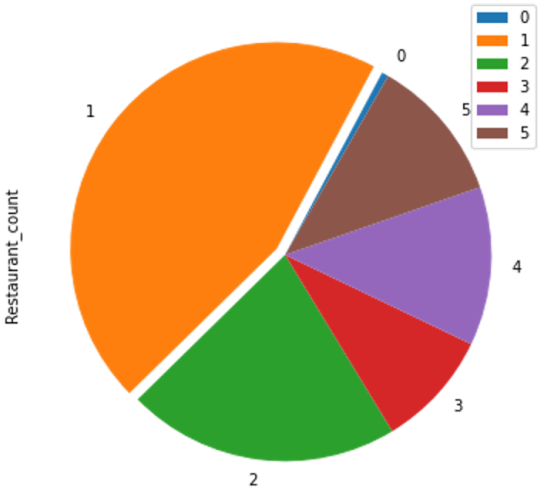
ultimately we can find that neighborhood out layer that need to have more restaurants to attain equilibrium too

Now out neighborhood are splitting into six new clusters this time much more accurate results

`Neighborhood by No. of restaurants and 2nd cluster labeled cluster_1 is seen`



By comparing our results cluster with restaurants again we see a dominant cluster



shows that cluster 1 stands out followed by cluster 0 as cluster with thriving Restaurant Business

RESULT AND DISCUSSION

interpreting data we can see that Harbourfront East and Enclave of M5E neighborhoods¶ are the most promising because they are not yet saturated with a many restaurants yet and a have the attributes of all the places with teaming Restaurants¶

	index	Neighborhood	Cluster Labels2	Restaurant_count	Venue
0	12	Harbourfront East , Union Station , Toronto ...	1	14.0	100
1	9	Enclave of M5E	1	20.0	100

2	11	Garden District, Ryerson	1	23.0	100

harbourFront having only 14 restaurants and over a 100 our venues means that is ranks top on our list of neighborhoods follow by the Enclaves and Garden District and after review of this areas we find that there are no African restaurants. we pick Enclaves as the second runner up because it has 2 restaurants less compared with Garden Districts

Conclusion

in this project we set out to use location data to determine the most viable location for an African Restaurant in a neighborhood commercial enough for patronage and with little as much competition arising from presence of other restaurant using four square venue data along with some machine learning clustering algorithm we arrived at two promising neighborhoods Harbourfront East and Enclave of M5E.