Determining the best neighborhood to set up a indian restaurant.

Anish S Ghiya

1. Introduction

1.1 : Background

A restaurant , or an eatery, is a business that prepares and serves food and drinks to customers. Meals are generally served and eaten on the premises, but many restaurants also offer take-out and food delivery services. Restaurants vary greatly in appearance and offerings, including a wide variety of cuisines and service models ranging from inexpensive fast food restaurants and cafeterias, to mid-priced family restaurants, to high-priced luxury establishments.

Indian cuisine consists of a wide variety of regional and traditional cuisines native to the Indian subcontinent. Given the range of diversity in soil type, climate, culture, ethnic groups, and occupations, these cuisines vary substantially from each other and use locally available spices, herbs, vegetables, and fruits. Indian food is also heavily influenced by religion, in particular Hinduism, cultural choices and traditions. The cuisine is also influenced by centuries of Islamic rule, particularly the Mughal rule. Samosas and pilafs can be regarded as examples.

Toronto has the largest Indo-Canadian population in Canada. Almost 51% of the entire Indo-Canadian community resides in the Greater Toronto Area. Most Indo-Canadians in the Toronto area live in Brampton, Markham, Scarborough, Etobicoke, and Mississauga. Indo-Canadians, particularly, Punjabi Sikhs, have a particularly strong presence in Brampton, where they represent about a third of the population . The area is middle and upper middle class, home ownership is very high. The Indo-Canadians in this region are mostly of Punjabi, Telugu, Tamil, Gujarati, Marathi, Malayalee and Goan origin.

1.2 : Problem

Looking at he background of toronto we observe that the there are a lot of indians in the city of toronto.So the more the indian folk the more would be the buisiness for the indian restaurants. But there are few areas where there are indian restaurants which are very popular. So the new setup must be in the areas where there is no competition or at places where the population is mostly indian with high population density and the average income is also decently high so there is high chances of profitable buisiness , the other acse would be to set up in a place where the competition exists but is not that strong but that would not be advised as there might be some other reasons also why the business might not be running

properly in that area.

1.3 : Interest

The restaurant franchise owners might be interested who would be wanting to set up an indian restaurant. Any startup, in the feild of hotel buisiness who would want to set up a new buisiness in the restaurant field and would be especially helpfull to people who want to set up an indian restaurant. With the same code we can also find out for various types of the cuisines.

2. Data acquisition and cleaning

2.1 Datasources

The main data was obtained from wikipedia ie the postal codes of Canada.But this was not sufficient to get the required attributes that were required. So for the remaining data ie the latitudes and longitudes were taken from the geospatial data that was provided in the previous modules.The remainong columns were from wikipedia ie Demographics of toronto by neighborhood.

The fousquare api was alo used for this to attain the datasets for the venues in the particular neighborhood.

The links to each of theses are :

i) https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
ii) https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

2.2 Data cleaning :

Data downloaded and the scraped ones from wikipedia had some data elemnts which were not defined.So for that set of data i had to remove the rows where the borough of the neighborhood was not assigned and then use the remaining data.

The next obstacle was with the demographic data which had the neighborhoods in a different pattern so i had to convert the other negihborhood data into csv and then manually change the data and get it to that particular format that was compatible.But with that also i only managed to get the data for 47 neighborhoods in the city of toronto.

The demographic data had the secong language as the combintion of the percentage of the people who comrised of that population. So to change that i had to convert it to a list and the convert then into 2 parts. The first one was to to use the split funtion based on space and then use the list created to extract the language. Similarly there was a problem with the population % , to handle it i used the same approach but the change was just that i used the deliminator for the spit function as the “% “symbol.

Now the next problem was that the percentages got saves as a object instead of a float so i had to change the variable type to numeric

There were no outliers as such in the data as this was census data and so there was not much of a trouble there.

2.3 Feature Selection :

The features set had 18 features which were not enough to understand the data so after dropping the columns of the second highest spoke language and the percentage the data was removed of the redndant data of these.

Now the other features like the map was not available for all the values and so that column had to be dropped.now this left me with just 43 rows ie 43 neighborhoods. Don't worry this is just the basic data. The main data comes in when the foursquare calls are made.

Once the foursqaure calls are made the total number of columns increase by 4 and all of them are important as they include the venue name and the venue latitiude and longitiude and the category of that shop

I also designed another feature of population percentage ie the actual population which spoke the second highest language as there might be areas with less population density to which might not be so significant.

Name	FM	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	% Renters	Postal code	Borough	Latitude	Longitude	Language	Percentag
Agincourt	S	44577	12.45	3580	4.6	25750	11.1	5.9	M1S	Scarborough	43.794200	-79.262029	Cantonese	19.
Bayview Village	NY	12280	4.14	2966	41.6	46752	14.4	15.6	M2K	North York	43.786947	-79.385975	Cantonese	08.
bagetown	OCoT	11120	1.40	7943	5.3	50398	18.5	29.6	NaN	NaN	43.667967	-79.367675	Unspecified	01.
hurch and Wellesley	OCoT	13397	0.55	24358	8.8	37653	25.1	57.0	M4Y	Downtown Toronto	43.665860	-79.383160	Spanish	01.
Davisville	OCoT	23727	3.14	7556	4.5	55735	26.0	31.7	M4S	Central Toronto	43.704324	-79.388790	Persian	01.

Fig1 : Table depicting the toronto data

Out[81]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Agincourt	43.7942	-79.262029	Panagio's Breakfast & Lunch	43.792370	-79.260203	Breakfast Spot
1	Agincourt	43.7942	-79.262029	El Pulgarcito	43.792648	-79.259208	Latin American Restaurant
2	Agincourt	43.7942	-79.262029	Twilight	43.791999	-79.258584	Lounge
3	Agincourt	43.7942	-79.262029	Mark's	43.791179	-79.259714	Clothing Store
4	Agincourt	43.7942	-79.262029	Commander Arena	43.794867	-79.267989	Skating Rink

Fig 2 : Table showing the venues in each neighborhood

3. Exploratory Data Analysis

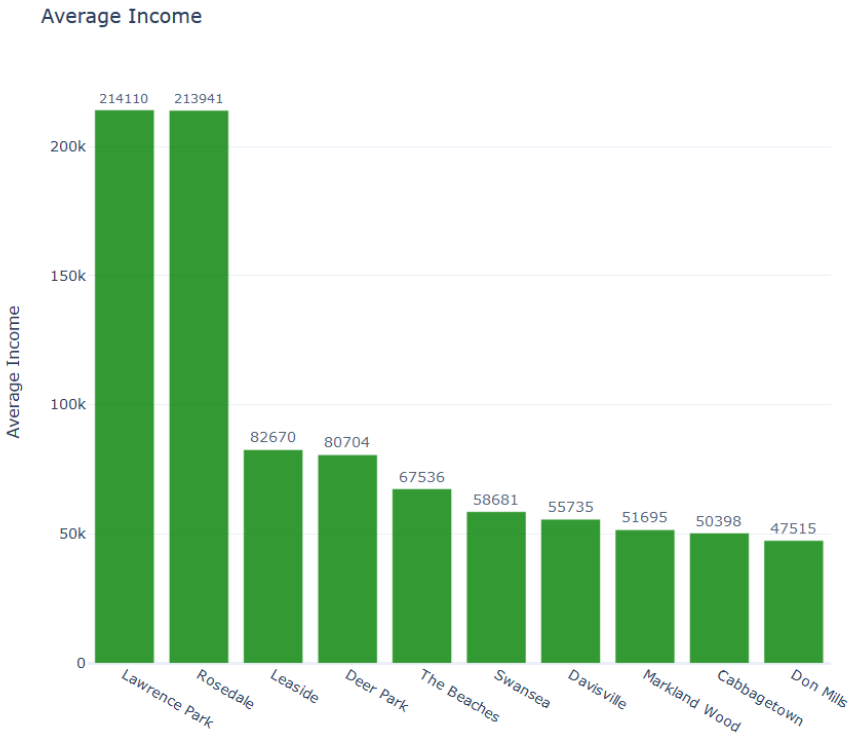
3.1 Understanding the essential information from the data

3.1.1 Average income

Average income of the particular neighborhood is an essential feature when it comes to understand the data, because if the plan is to open a restaurant then we must be able to understand that the people who have higher income group will be more interested in the more fancy restaurants ie 5 star restaurants etc.

The middle class family usually goes to the normal restaurant for regular weekend meals.Since i am from a middle class family i know that only on special occations like birthdays and anniversaries is the time when we go o ut to the fancy restaurants.

So average income is a major contributor the kinf of restaurant especially if u have the the specific neighborhood already set in your mind.



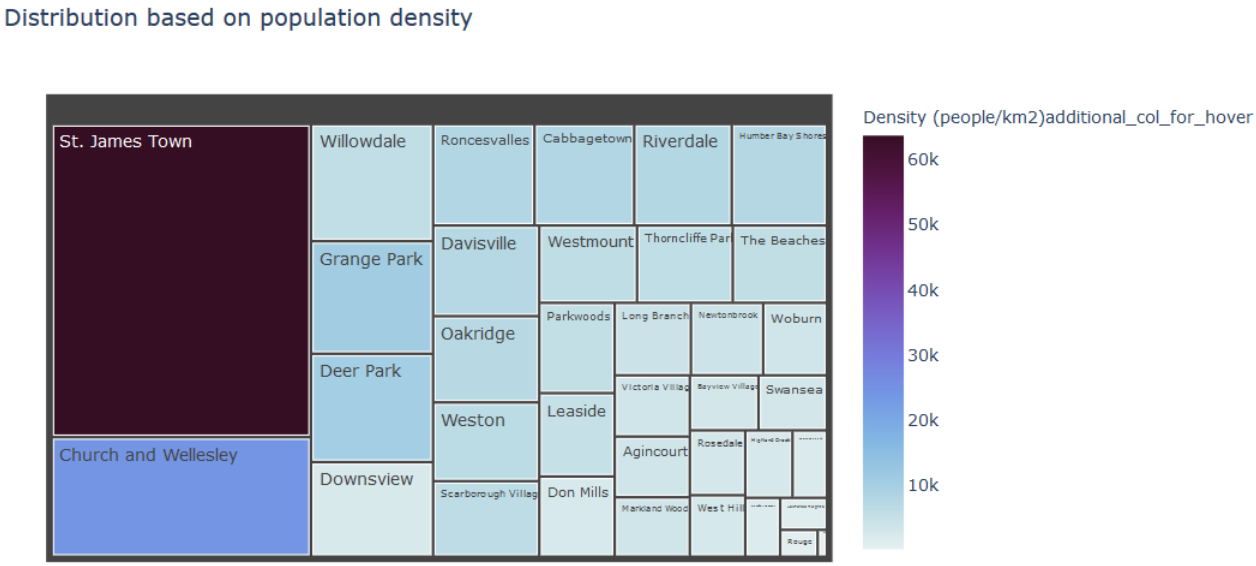
From the above graph we understand that the 2 neighborhood Lawrence Park and Rosedale have much higher average income when compared to the other neighborhoods

Following neighborhood come under the middle class income group and so this are the areas we should focus when it comes for the restaurant business

3.1.2 Population density

Population density is another one of the important data feature so as to find the regions where the restaurant might function better. With the greater

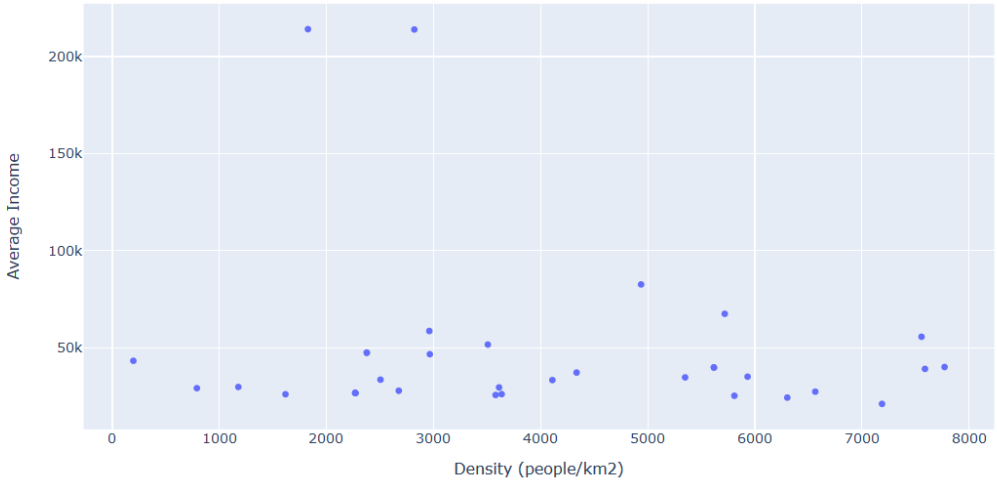
population density there are more people in one particular neighborhood and hence gives us a better neighborhood to set up the restaurant.



From the tree map we obsreve that the St. James Town has the highest population density but this might not be sufficient to undertand the population type in the neighborhood and it is not alone that helpful

3.1.3 Density versus average income

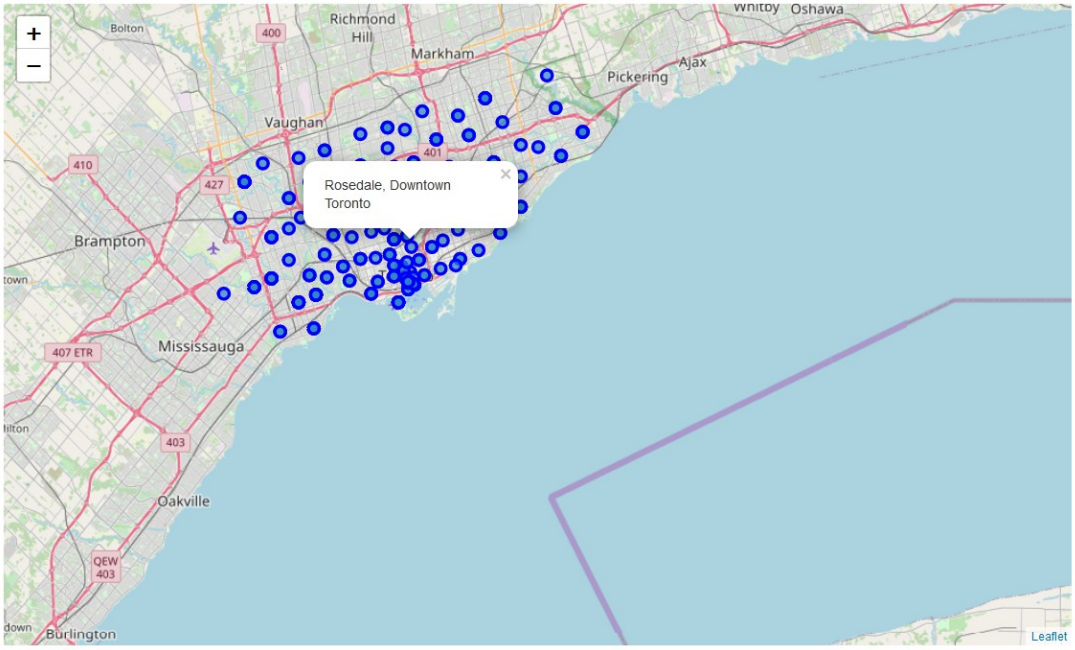
The average income as we have seen plays a major role in the location picking as well as the population density but now let us see the relation between the population density and the average income which helps in understand the population of that neighborhood.



3.1.4 Ploting the neighborhoods

Lets now plot a map of toronto for the neighborhoods of the city and see the density of the plot where the points are the maximum for this i ahve the toronto data with the neighborhoods only on clicking the points u can notice the neighborhood and the borouhgh will pop out.

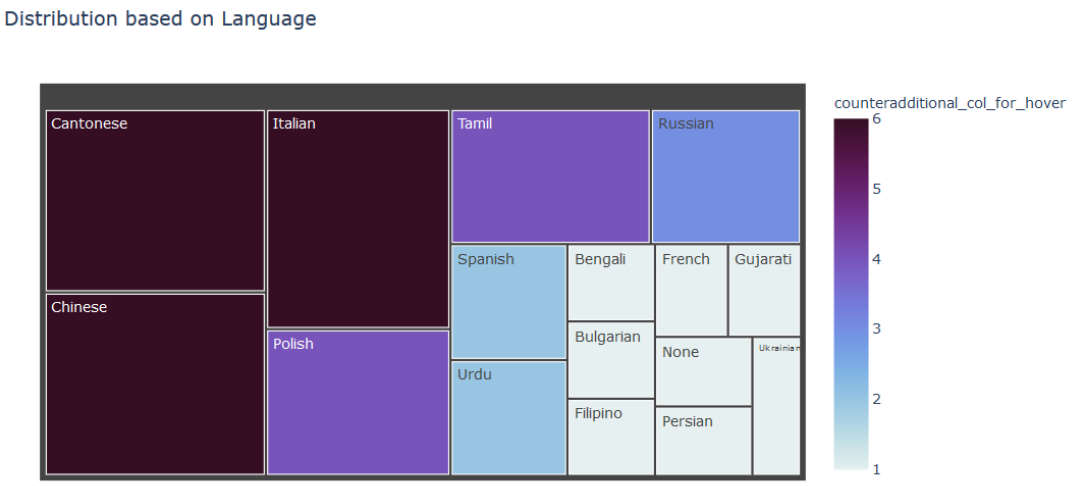
There are chances that u will not be able to see the plot in the ipny notebook that is why i will aslo upload the html file for this as map1 in the github page



From the above graph we observe that the main data is pertaining to toronto main suburb and there are few in the outskirts of the dataframe in terms of the city

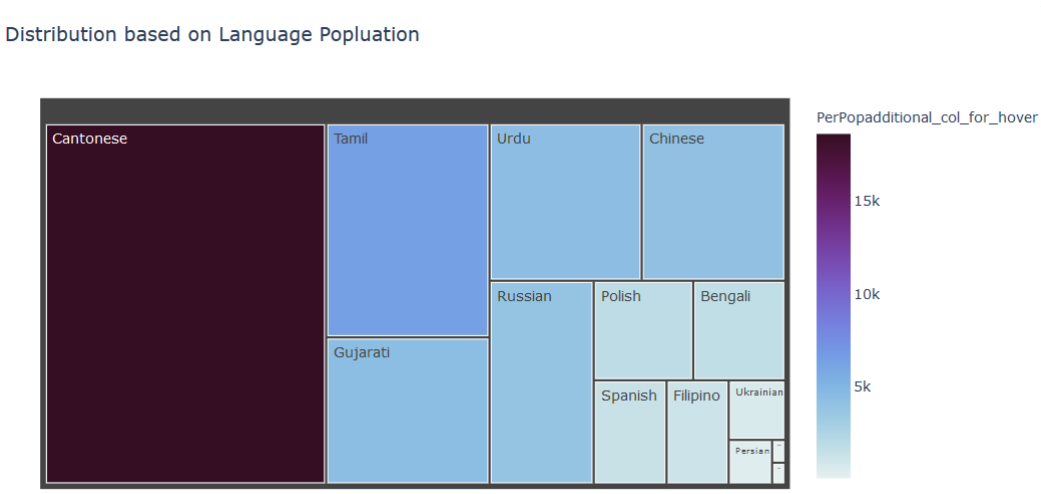
3.1.5 Language distribution

The language will play a major role in detemining the population type of the particular neighborhood and so there will be a lot of emphasis on this as the more the population of that language in the region they will like that type of cuisine more.

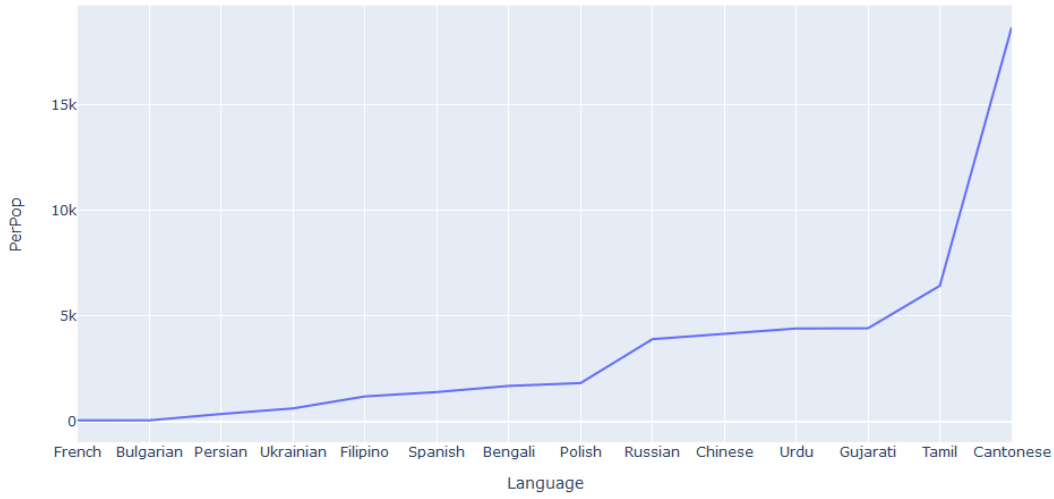


From the tree map we can see that cantonese,chinese,italian are the major second languages in the cty of toronto
Indians are not behind but as we speak different languages in different states

so we are diversified by language but if we sum all of these we can get ahead of all the other languages.

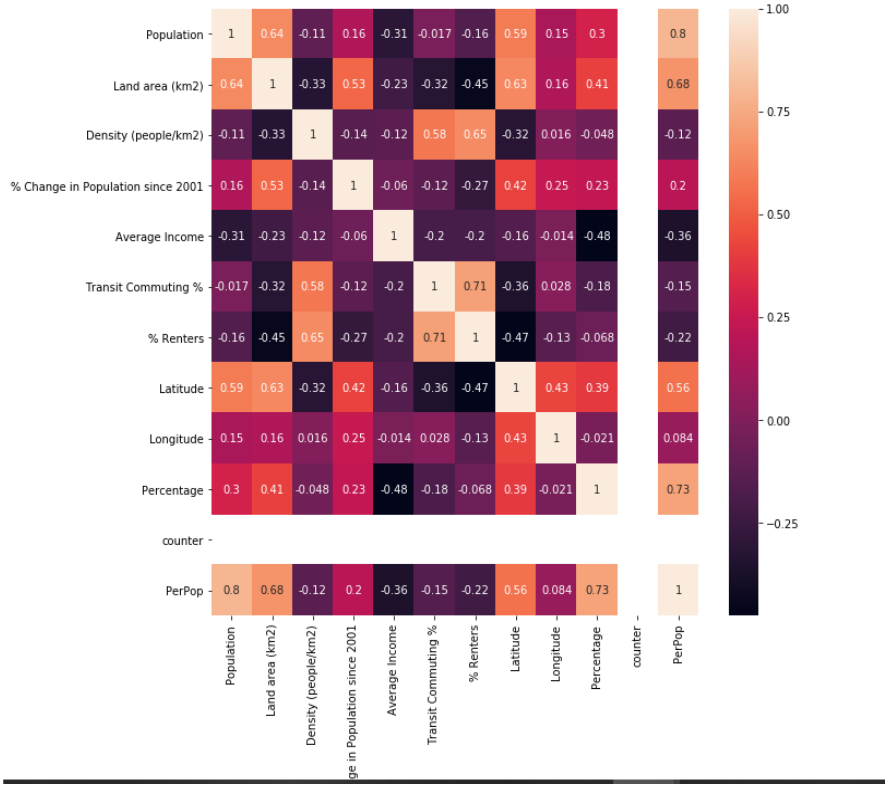


The above tree map is a descrption of how there a change when we see the population which speak the language amd here we can see that cantonese reamins at the top with the max population speaking the language and then is followed by tamil and Gujarati.



3.1.6 Heatmap of the correlation between the features

It is essential to understand the correlation between various data features so as to get a better understanding of the dataframe and with this correlation we can understand whatdetrmines the paarticular feature in the best possible way.



3.1.7 Understanding the venues in neighborhood

Now let us see the number of each type of venue present in each neighborhood and then i have sorted in the desceneding order to see the maximum in fe of the neighborhoods

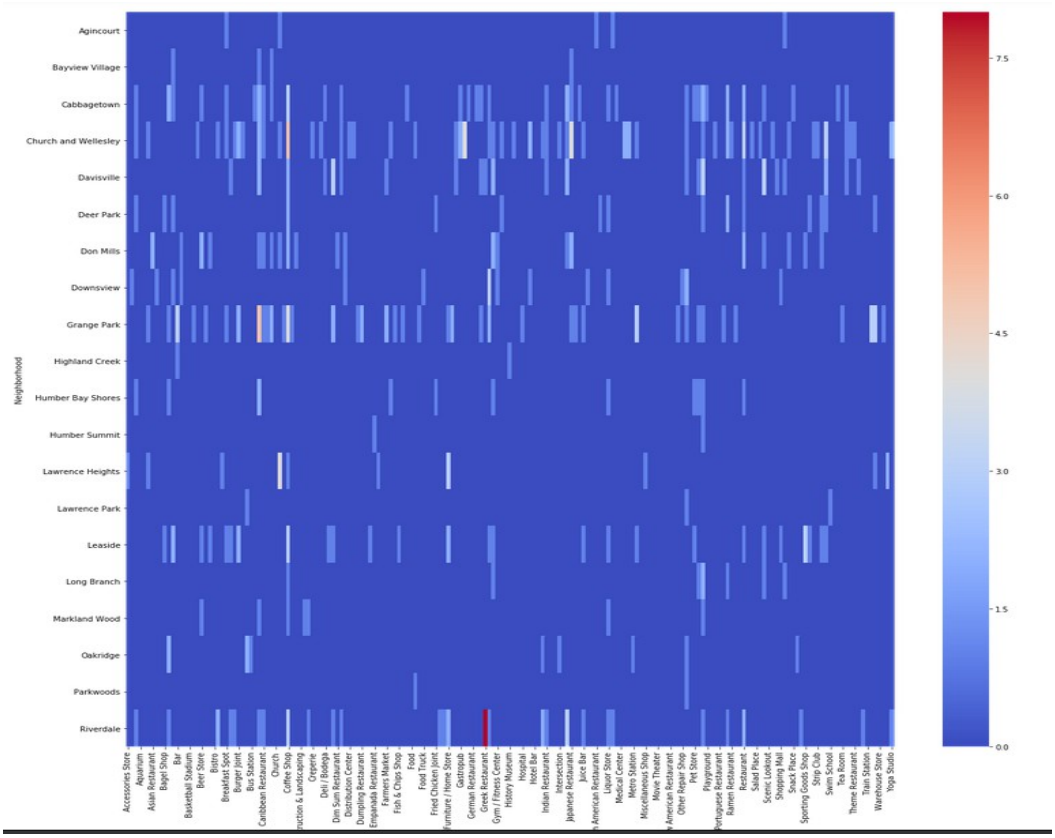


Fig 8 Heatmap showing venue v=category of each neighborhood

The followin plot shows the different venue categories in each of the neighborhoods

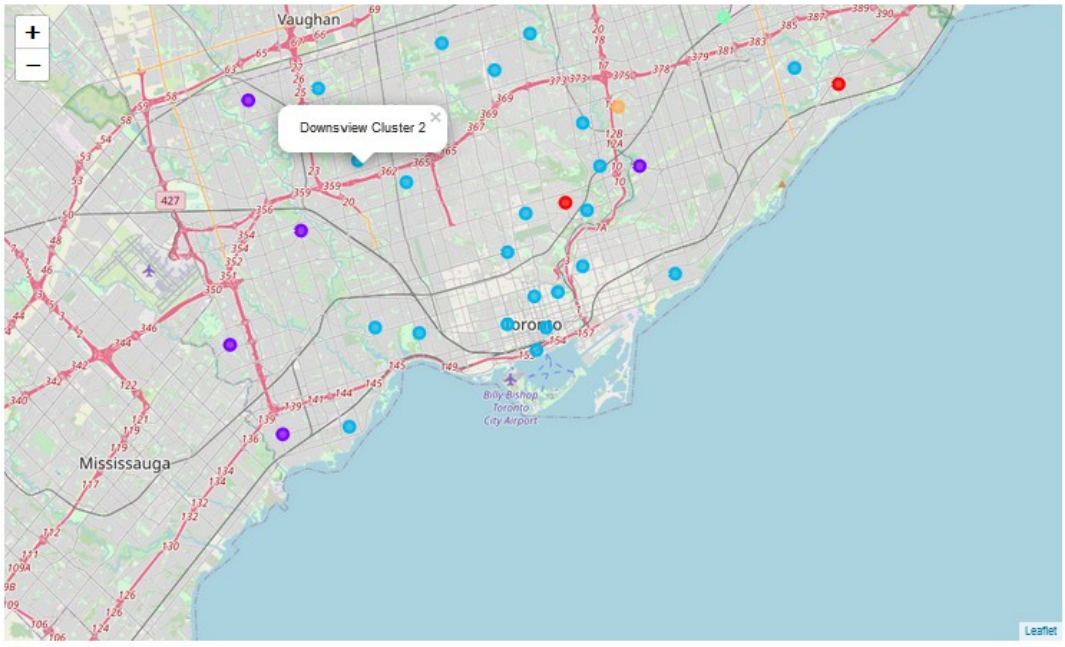
So for example there are a lot of greek restaurants in Riverdale which in the heatmap is represented as a red block in the diagram and so similarly we can read the data from this heat map

4. Predictive Modelling :

4.1 Model used :

Kmeans algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-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.

4.2 Modelling Output :



The cluster 2 contains most of the restaurants do the further analysis was done on this cluster for which i created a measure :

The average income must be high in so that the buisiness remains profitable.

- The transit population must be less
- The transit population must be less so as to understand that the population outflux is not high

5.Result :

The Rouge neighborhood in the Scarborough would be best to set up a south indian restaurant in this cluster

```
In [ ]: 1 toronto_clustered_india.iloc[0]
```

```
Out[ ]: FM_x                                                S
        % Change in Population since 2001_x                175
        Average Income_x                                 29230
        Transit Commuting %_x                             12.1
        % Renters_x                                        2.7
        Postal code_x                                      NaN
        Borough_x                                   Scarborough
        Latitude                                        43.8067
        Longitude                                      -79.1944
        Language_x                                        Tamil
        Percentage_x                                       15.6
        counter_x                                             0
        PerPop_x                                        3544.94
        Transit Comm pop_x                               2749.6
        Rental pop_x                                     613.548
        Cluster Labels                                         2
        1st Most Common Venue                Fast Food Restaurant
        2nd Most Common Venue                Vietnamese Restaurant
        3rd Most Common Venue          Eastern European Restaurant
        4th Most Common Venue                  German Restaurant
        5th Most Common Venue                Fried Chicken Joint
        Name                                               Rouge
        FM_y                                                   S
        Population                                         22724
        Land area (km2)                                    28.72
        Density (people/km2)                                 791
        % Change in Population since 2001_y                  175
        Average Income_y                                   29230
        Transit Commuting %_y                               12.1
        % Renters_y                                          2.7
        Postal code_y                                        NaN
        Borough_y                                     Scarborough
        Language_y                                          Tamil
        Percentage_y                                         15.6
        counter_y                                               0
        PerPop_y                                          3544.94
        Transit Comm pop_y                                 2749.6
        Rental pop_y                                       613.548
        Measure                                            1.14142
        Name: 16, dtype: object
```