Segmenting, Clustering and Recommending to avoid bad choice of Neighborhoods for New Restaurants in Toronto

Austin Howard

1. Introduction

1.1 Background

Running a restaurant is more than just offering food in exchange for money. You will be offering an experience to customers based on items such as the decor, food, and service. As the owner, it is up to you to determine the type of experience you want to give customers. Even after you've decided on the type of restaurant you want to open, you need to make sure there is a market for it's **cuisine**, and you can find **the right location**.

Recommendation systems are now very popular in many aspects such as e-commerce services, movie rating (Netflix), social networking and a lot of services helping users finding preferred items, in this work i will focus on (LBSN) location-based social networks.

Location-based social networks, e.g., **Foursquare**, Gowalla and Whrrl4, have seen soaring popularity, attracting millions of users. People are increasingly using these location-based social networking services to connect with friends, explore places (e.g., restaurants, stores, cinema theaters, etc.) and share their locations.

1.2 Business Problem

choosing a Location for Your Restaurant Location is vital to the success of any restaurant. There are several factors to consider when searching for that perfect restaurant location, including population base, local employment figures, and accessibility and a critically important factor is to **select an area where the competition is not intense for the food you are offering**.

1.3 Interest

Out goal is to find a solution made for who is looking to open a restaurant in Toronto, by asking him what kind of food (cuisine) he will offer, i will recommend him which neighborhoods will not be a good choice for him.
In this work, I will design a restaurant recommender system based on k-means an unsupervised clustering model, which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean of the frequency of occurrence of each category (Restaurant Cuisines).

2. Data acquisition and cleaning

2.1 Data sources

i will use a combination of data types:

- List of postal codes of Toronto from WIKIPEDA.
- Geospatial Coordinates CSV file for Toronto postal codes from http://cocl.us/Geospatial_data.
- Foursquare API

2.2 Data cleaning

I will read first the HTML text using BeautifulSoup library of the List of postal codes of Toronto to group neighborhoods for each zip code.

Toronto - FSAs [edit]

Note: There are no rural FSAs in Toronto, hence no postal codes start with M0.

Postcode	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Harbourfront
M5A	Downtown Toronto	Regent Park
M6A	North York	Lawrence Heights
M6A	North York	Lawrence Manor
M7A	Queen's Park	Not assigned

Group each zip code with all neighborhoods in:

Out[13]:

	PostalCode	Borough	Neighborhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

2.3 Insert Geo-spatial Coordinates

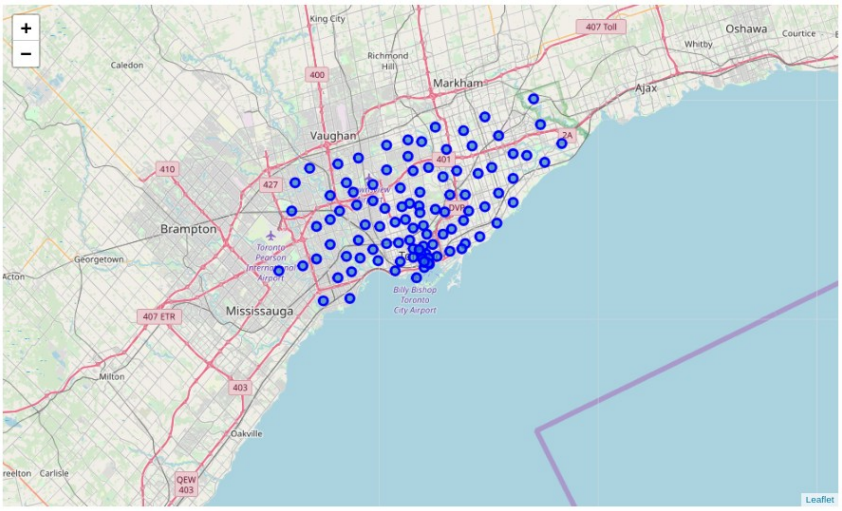
i will insert Geospatial Coordinates csv file to the code from my IBM Cloud Object Storage, to add Latitude and Longitude for each zip code area and store it into pandas dataframe and I will perform a join with the previous dataframe to get a new dataframe with each zip code and neighborhood and the latitude and longitude coordinates.

Out[20]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Then we can plot a map of the neighborhoods using Folium Library:

Out[18]:



2.4 Foursquare API utilizing

finally i will utilize Foursquare API to explore neighborhoods and segment them.

We can get the top 100 venues that are within a radius of 500 meters for a zip code area, the result is a json object.

For example the result of Rouge, Malvern area will be.

Out[23]:


```
{'meta': {'code': 200, 'requestId': '5ca786cadd57977cb5c7a670'},
 'response': {'groups': [{'items': [{'reasons': {'count': 0,
       'items': [{'reasonName': 'globalInteractionReason',
         'summary': 'This spot is popular',
         'type': 'general'}]},
     'referralId': 'e-0-4bb6b9446edc76b0d771311c-0',
     'venue': {'categories': [{'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/fastfood_',
         'suffix': '.png'},
        'id': '4bf58dd8d48988d16e941735',
        'name': 'Fast Food Restaurant',
        'pluralName': 'Fast Food Restaurants',
        'primary': True,
        'shortName': 'Fast Food'}],
      'id': '4bb6b9446edc76b0d771311c',
      'location': {'cc': 'CA',
       'city': 'Toronto',
       'country': 'Canada',
       'crossStreet': 'Morningside & Sheppard',
       'distance': 387,
       'formattedAddress': ['Toronto ON', 'Canada'],
       'labeledLatLngs': [{'label': 'display',
         'lat': 43.80744841934756,
         'lng': -79.19905558052072}],
       'lat': 43.80744841934756,
       'lng': -79.19905558052072,
       'state': 'ON'},
      'name': "Wendy's",
      'photos': {'count': 0, 'groups': []}}},
    {'reasons': {'count': 0,
      'items': [{'reasonName': 'globalInteractionReason',
        'summary': 'This spot is popular',
        'type': 'general'}]},
     'referralId': 'e-0-5539e7d2498edaf4b02673ca-1',
```


We need to clean the json to extract venues categories and structure it into pandas dataframe.

Out[25]:

	id	name	categories	lat	lng
0	4bb6b9446edc76b0d771311c	Wendy's	Fast Food Restaurant	43.807448	-79.199056
1	5539e7d2498edaf4b02673ca	Interprovincial Group	Print Shop	43.805630	-79.200378

Then I will perform top venue exploration for all neighborhoods and merge it to one data frame, that will include neighborhoods and their top rated venues, and because our main goal is to recommend a neighborhood for a new restaurant, we need to filter our data, so we will have only venues categories that contains the word 'Restaurant'

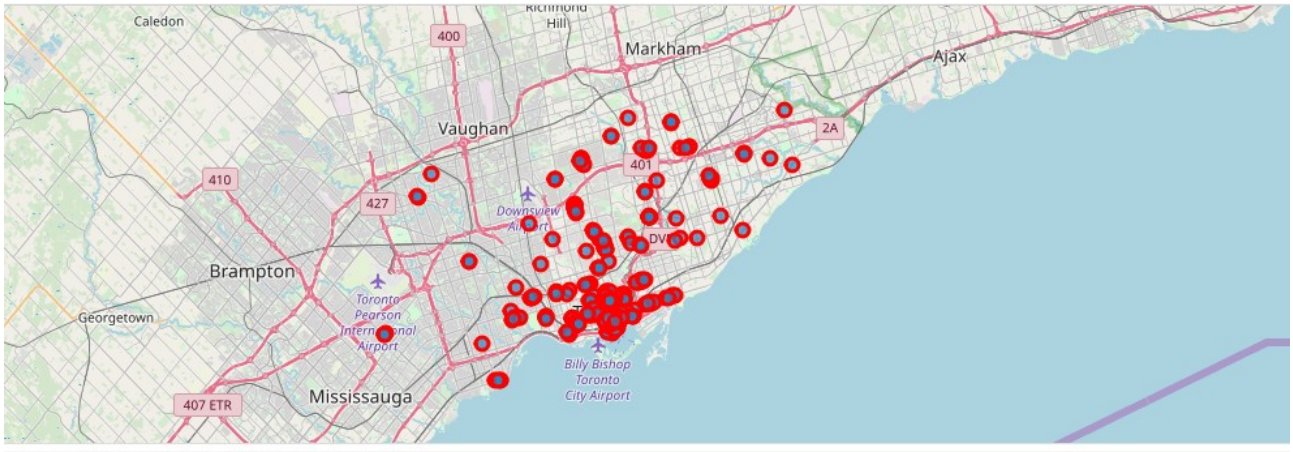
Out[32]:

	index	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Id	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	0	Rouge, Malvern	43.806686	-79.194353	4bb6b9446edc76b0d771311c	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	6	Guildwood, Morningside, West Hill	43.763573	-79.188711	5411f741498e9ebd5e35d8bd	Big Bite Burrito	43.766299	-79.190720	Mexican Restaurant
2	13	Woburn	43.770992	-79.216917	4de0403ed4c040523ea079f4	Korean Grill House	43.770812	-79.214502	Korean Restaurant
3	14	Cedarbrae	43.773136	-79.239476	4b1711a6f964a520cbc123e3	Federick Restaurant	43.774697	-79.241142	Hakka Restaurant
4	15	Cedarbrae	43.773136	-79.239476	4e261f261f6eb1ae13930699	Drupati's Roti & Doubles	43.775222	-79.241678	Caribbean Restaurant

3. Methodology section (EDA - Exploratory Data Analysis)

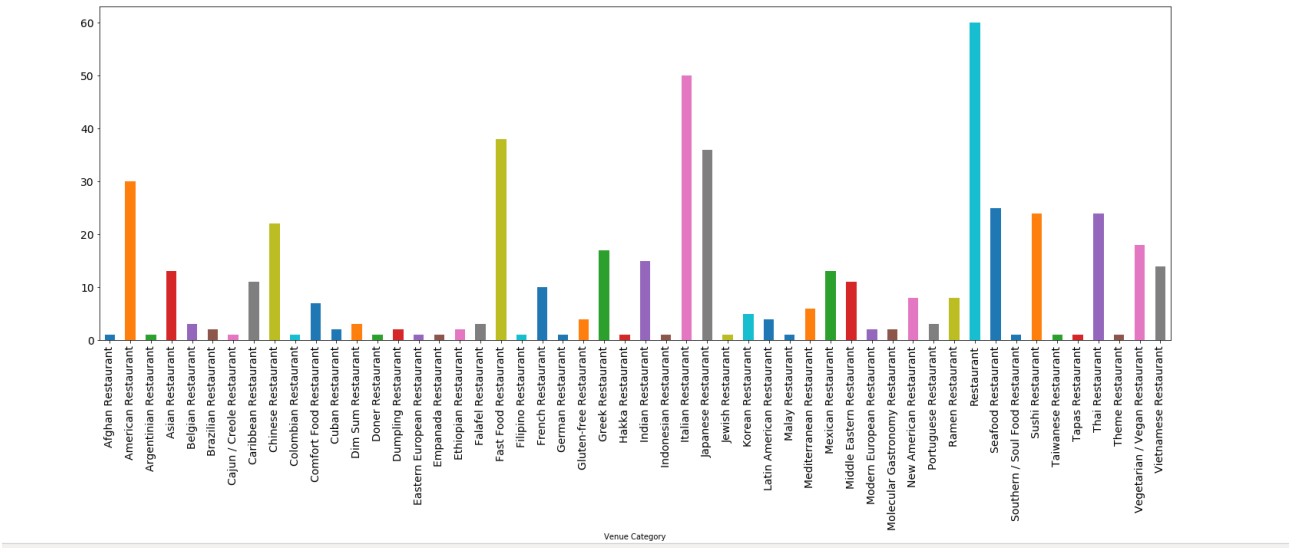
From the dataset, we filtered out the neighborhood which have venue contains the word 'Restaurant' . After filtering out the the data, our dataset shape is (514, 9) that means that the data contains 514 rows (Restaurant venues) and 9 relevant columns.

Here is a map of all the top rated restaurant in Toronto:



Here is the distribution of the restaurant cuisines:

Out[56]: <matplotlib.axes._subplots.AxesSubplot at 0x7f41e86bc208>



4. Results section

we can see the is the most frequent restaurant type is the a ‘General Restaurant’ show in light blue.

Then for each one of the neighborhood we can extract the top frequent restaurant categories:

```
----Adelaide, King, Richmond----
                              cuisine  freq
0      Venue Category_Thai Restaurant  0.15
1  Venue Category_American Restaurant  0.11
2     Venue Category_Sushi Restaurant  0.11
3           Venue Category_Restaurant  0.11
4  Venue Category_Japanese Restaurant  0.07


----Albion Gardens, Beaumond Heights, Humbergate, Jamestown, Mount Olive, Silverstone, South Steeles, Thistletown----
                                cuisine  freq
0    Venue Category_Japanese Restaurant   0.5
1   Venue Category_Fast Food Restaurant   0.5
2      Venue Category_Afghan Restaurant   0.0
3  Venue Category_Indonesian Restaurant   0.0
4      Venue Category_Jewish Restaurant   0.0
```

we can see for ‘Adelaide, King’ Neighborhoods Thai Restaurant is the most frequent category with 0.15 frequency mean of all the top rated restaurants, and then in the second place is the American Restaurant with 0.11 frequency mean.

Out[40]:

	Neighborhood	1st Most Common cuisine	2nd Most Common cuisine	3rd Most Common cuisine	4th Most Common cuisine	5th Most Common cuisine
0	Adelaide, King, Richmond	Venue Category_Thai Restaurant	Venue Category_American Restaurant	Venue Category_Sushi Restaurant	Venue Category_Restaurant	Venue Category_Japanese Restaurant
1	Albion Gardens, Beaumond Heights, Humbergate, ...	Venue Category_Japanese Restaurant	Venue Category_Fast Food Restaurant	Venue Category_Vietnamese Restaurant	Venue Category_Greek Restaurant	Venue Category_German Restaurant
2	Bathurst Manor, Downsview North, Wilson Heights	Venue Category_Fast Food Restaurant	Venue Category_Sushi Restaurant	Venue Category_Restaurant	Venue Category_Vietnamese Restaurant	Venue Category_Dim Sum Restaurant
3	Bayview Village	Venue Category_Japanese Restaurant	Venue Category_Chinese Restaurant	Venue Category_Vietnamese Restaurant	Venue Category_Doner Restaurant	Venue Category_German Restaurant
4	Bedford Park, Lawrence Manor East	Venue Category_Fast Food Restaurant	Venue Category_Italian Restaurant	Venue Category_Comfort Food Restaurant	Venue Category_American Restaurant	Venue Category_Thai Restaurant

5. Clustering Modeling

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

- The centroids of the K clusters, which can be used to label new data.

- Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined.

I will use popularity recommendation filtering approach in order to make recommendations.

Popularity Recommendation : Given the number of play counts (occurrences) for items songs, movies, whatever, just sort by descending ordering and recommend everyone what’s popular.This actually makes a good shelf for any content platform, such as Netflix, Spotify or Amazon.

I performed a qualitative comparison of the existing techniques used in the Data I got from Foursquare, followed by a clustering on the type of the services and the location features the Foursquare utilizes to perform the recommendations.

First of all I performed one hot encoding on the data to get a binary represented data set for each neighborhood venue(Restaurant)

Out[35]:

	Neighborhood	Venue Category_Afghan Restaurant	Venue Category_American Restaurant	Venue Category_Argentinian Restaurant	Venue Category_Asian Restaurant	Venue Category_Belgian Restaurant	Venue Category_Brazilian Restaurant	Venue Category_Cajun / Creole Restaurant	Venue Category_Caribbean Restaurant	Venue Category_Chinese Restaurant	...	Venue Category_Restaurant	Venue Category_Seafood Restaurant
0	Rouge, Malvern	0	0	0	0	0	0	0	0	0	...	0	0
1	Guildwood, Morningside, West Hill	0	0	0	0	0	0	0	0	0	...	0	0
2	Woburn	0	0	0	0	0	0	0	0	0	...	0	0
3	Cedarbrae	0	0	0	0	0	0	0	0	0	...	0	0
4	Cedarbrae	0	0	0	0	0	0	0	1	0	...	0	0

5 rows × 53 columns

and then grouped rows by neighborhood and by taking the mean of the frequency of occurrence of each category I got :

Out[36]:

	Neighborhood	Venue Category_Afghan Restaurant	Venue Category_American Restaurant	Venue Category_Argentinian Restaurant	Venue Category_Asian Restaurant	Venue Category_Belgian Restaurant	Venue Category_Brazilian Restaurant	Venue Category_Cajun / Creole Restaurant	Venue Category_Caribbean Restaurant	Venue Category_Chinese Restaurant	...	Venue Category_Restaurant	Venue Category_Seafood Restaurant	C
0	Adelaide, King, Richmond	0.0	0.111111	0.0	0.074074	0.0	0.037037	0.0	0.0	0.0	...	0.111111	0.037037	
1	Albion Gardens, Beaumond Heights, Humbergate, ...	0.0	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	...	0.000000	0.000000	
2	Bathurst Manor, Downsview North, Wilson Heights	0.0	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	...	0.333333	0.000000	
3	Bayview Village	0.0	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0	0.5	...	0.000000	0.000000	
4	Bedford Park, Lawrence Manor East	0.0	0.090909	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	...	0.090909	0.000000	

5 rows × 53 columns

a dataframe representing each neighborhood and the mean frequency of occurrence for each restaurant cuisine in it.

In my solution I choosed K = 10, so I want to split neighborhoods to 10 clusters:

for the first 10 neighborhoods, the model result is :

```
Out[41]: array([2, 3, 8, 3, 2, 2, 2, 8, 2, 1], dtype=int32)
```

first neighborhood is in cluster number 2, second neighborhood in cluster number 3, third neighborhood in cluster number 8 and so on...

6. Discussion section

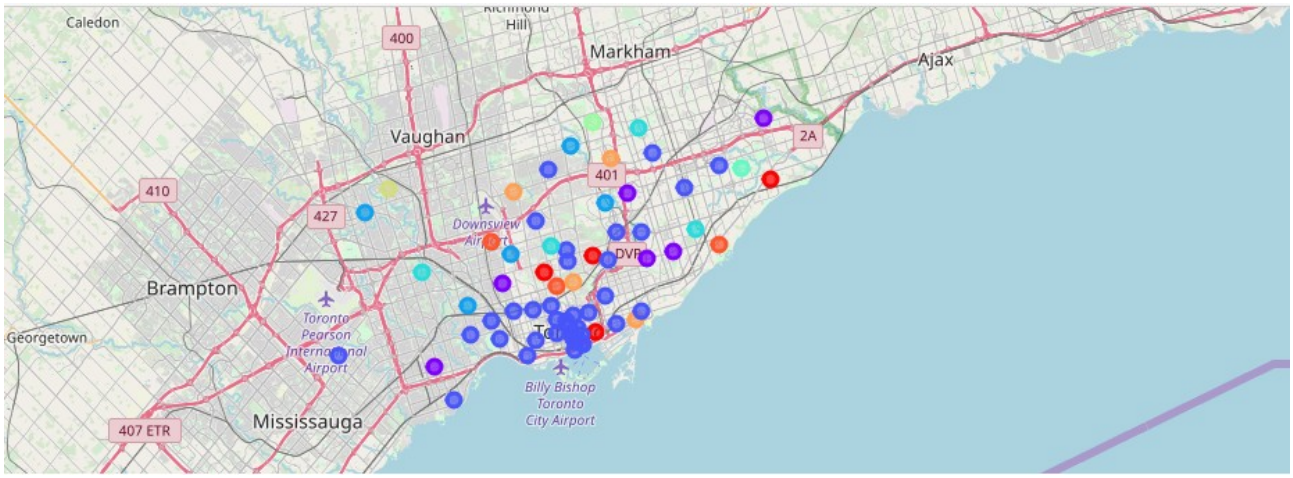
for simplicity we can see clusters size:

clusters sizes

```
In [44]: toronto_merged.groupby('Cluster Labels').count()['Neighborhood']

Out[44]: Cluster Labels
         0     4
         1     6
         2    38
         3     5
         4     4
         5     1
         6     1
         7     1
         8     4
         9     3
         Name: Neighborhood, dtype: int64
```

Easily we can see that cluster 2 is the biggest one, that means that all neighborhoods in cluster 2 are similar with top rated restaurants cuisines, Now, let’s plot a map of neighborhood and colored clusters:

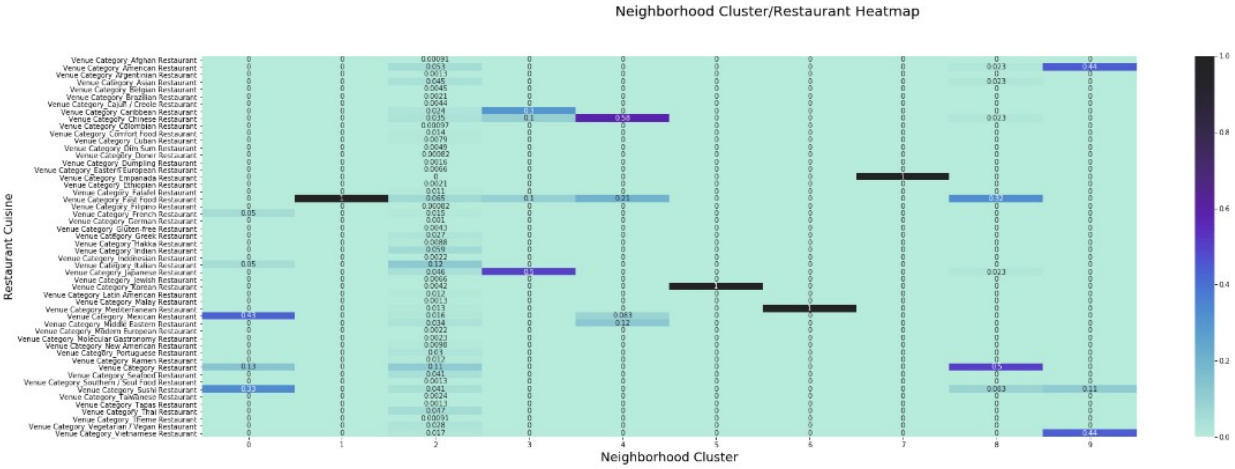


in this stage, I merged cluster labels to the dataset, so the new dataset will be :

Out[51]:

Cluster Labels	Venue Category_Afghan Restaurant	Venue Category_American Restaurant	Venue Category_Argentinian Restaurant	Venue Category_Asian Restaurant	Venue Category_Belgian Restaurant	Venue Category_Brazilian Restaurant	Cate
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
2	0.000907	0.053188	0.001316	0.044569	0.004468	0.002119	
3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
5	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
6	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
7	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
8	0.000000	0.022727	0.000000	0.022727	0.000000	0.000000	
9	0.000000	0.444444	0.000000	0.000000	0.000000	0.000000	

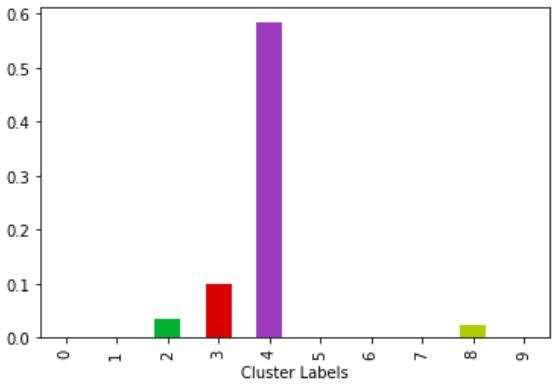
then I grouped the dataset by cluster to get the mean of the frequency occurrence of occurrence of each cluster, we then I examined each cluster by a heatmap:



7. Conclusion section

conclusion can be taken from the previous heatmap, Dark color means that the restaurant cuisine is very common in the corresponding neighborhood cluster, for example if someone is willing to open a new Chinese Restaurant, we can recommend him where not to open his new restaurant.

```
In [54]: toronto_clusters_trans.loc['Venue Category_Chinese Restaurant'].plot(kind='bar')
Out[54]: <matplotlib.axes._subplots.AxesSubplot at 0x7f93a004db00>
```



cluster 4 has a high frequent of Chinese restaurants, so try not to open a new Chinese restaurant in cluster 4's neighborhoods.