Growing from Local to Global: A Recommender System for Restaurant Business Expansion

IBM Data Science Professional Capstone Project

Junkang Gu – 2020
jgu8@u.rochester.edu
University of Rochester, NY

I. Introduction

With the rise of social media, insta-famous restaurants are fast growing among younger customers. In contrast to mature franchises like McDonald's and Burger King, which already have established networks of existing stores, emerging franchises such as Shake Shack are growing and expanding their business from cities where they are originally located, to a wider range of places by opening new stores. However, choosing the location for opening a new store may turn out complicated, as cities and neighborhoods vary from each other, sometimes significantly. The project aims to develop a machine learning based recommender system advising restaurant owners on potential locations for business expansion.

Business Problem

This project will focus on Shake Shack in Rochester as an example, while the recommender system can be generalized and applied to any similar business scenarios. Shake Shack is an insta-famous burger restaurant originated as a hotdog cart in Madison Square Park, New York City. Shake Shack quickly grew its business in NYC, mostly in the busy regions in Manhattan. It would not be challenging to open another Shake Shack in NYC, as Shake Shack is originated and rooted in the culture of NYC. However, if its owner looked into a wider scope of cities they are less familiar with, such as Beijing in China with a completely different city layout pattern from NYC, or Rochester, NY which is a lot less populated, it would be difficult to find the perfect location that best fit their franchise and brand.

Here we suppose that Shake Shack is considering the possibility of business expansion to the city of Rochester for example. Rochester is a city a lot less populated than the NYC and differs significantly in many aspects. Unlike the NYC which is an economic center as well as a metropolis for tourist attractions, Rochester is a less populated city suffering a decline since the optics tycoon Kodak collapsed. However, it might still be worth consideration, due to multiple universities and colleges which lead to a huge number of young college students who are potential customers for Instagrammable restaurants. "Rochesterfoodies", an Instagram account, has 11.6K followers, signaling great potential for these restaurants. Therefore, with data analysis, we would like to figure out the best neighborhood in Rochester, if any, for opening a new Shake Shack store.

Target Audience

The target audience would be entrepreneurs and growing business owners who are looking for business expansion, especially in cities they are less familiar with, due to different cultural, economic, and geographic backgrounds. This recommender system will analyze data of a given target city and help identify potential locations that are best fit for their business brand.

II. Data Description

The following data are required for our analysis.

Determine the Research Subjects

In this example, we want to study the Shake Shack in New York City and look for potential business expansion in Rochester. Therefore, the research subjects are prepared as the following:

```
In [5]: #Existing store samples proven to be successful:
existing_stores_query = "Shake Shack in NYC"
sample_city = "New York City"
#Target City for expansion:
target_city_query = "Rochester"
```

Sample Stores Proven to be Successful

In order to predict the potential locations that are best fit for business expansion, we need to acquire data of the existing stores which are proven to be located in profitable neighborhoods. Here we assume that any Shake Shack stores still in business are making profits, with their location carefully and professionally selected by their owners and advisors. We acquire a list of Shake Shack stores in NYC by calling geolocator with query “shake shack in NYC”:

```
In [6]: #Location list for samples
sample_loc_list = geolocator2.geocode(query=existing_stores_query, exactly_one=False)
sample_loc_list
```

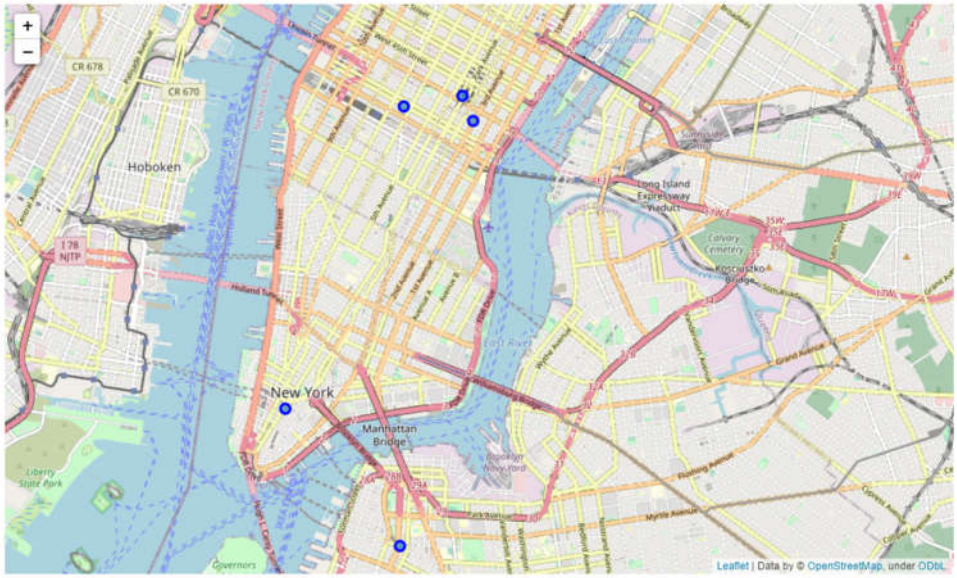
```
Out[6]: [Location(Shake Shack, Broadway, Financial District, Manhattan, Manhattan Community Boa
rd 1, New York County, New York, 10038, United States of America, (40.7105601, -74.0090
139, 0.0)),
 Location(Shake Shack, Broadway, Morningside Heights, Manhattan, Manhattan Community Bo
ard 9, New York County, New York, 10025, United States of America, (40.8079479, -73.964
3022, 0.0)),
 Location(Shake Shack, Ring Road, Staten Island Mall, Staten Island, New York, Richmond
County, New York, 10314-3903, United States of America, (40.5817907, -74.1676659, 0.
0)),
 Location(Shake Shack, Terminal 4 Departures, Bayswater, New York, Queens County, New Y
ork, 11430, United States of America, (40.6380016, -73.781199, 0.0)),
```

Out[7]:

	name	latitude	longitude
0	Shake Shack in NYC	40.710560	-74.009014
1	Shake Shack in NYC	40.807948	-73.964302
2	Shake Shack in NYC	40.581791	-74.167666
3	Shake Shack in NYC	40.638002	-73.781199
4	Shake Shack in NYC	40.643742	-74.075875

```
map_sample
```

Out[11]:



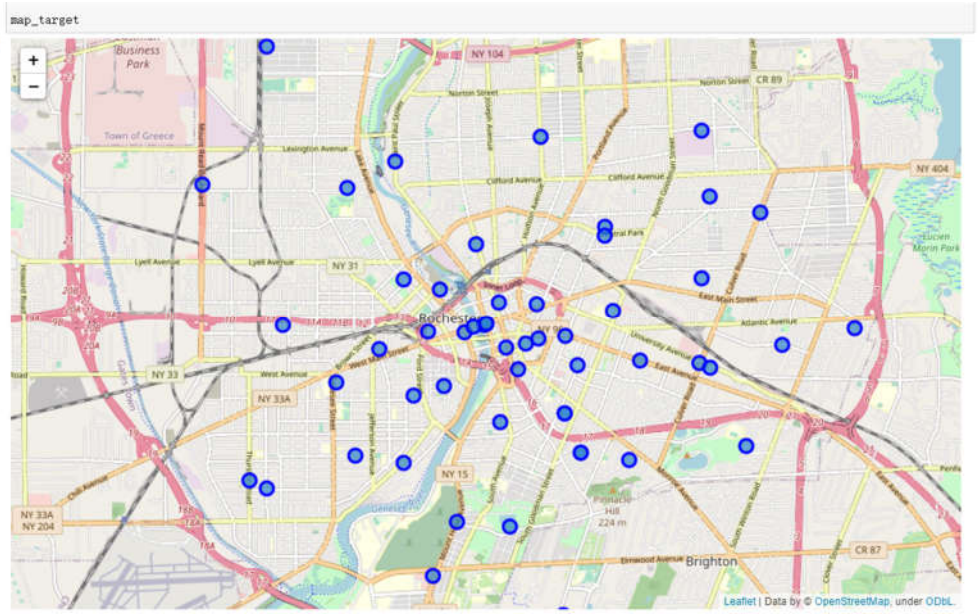
Target City for Business Expansion

Here we call the city for business expansion as the “target city”, which is Rochester in this example. As we are looking for the best neighborhood for a new Shake Shack, we need to acquire the information of all neighborhoods in Rochester. The list of neighborhood can be found here: https://rocwiki.org/Rochester_Neighborhoods Using the BeautifulSoup crawler we are able to load them into a data frame.

Out[17]:

	Neighborhood	Latitude	Longitude
0	Cascade District	43.155540	-77.619205
1	Convention District	43.156389	-77.609167
2	East End	43.154971	-77.594786
3	Four Corners	43.155428	-77.612713
4	Grove Place	43.159181	-77.599871

Out[19]:



Venue Data

The idea of this recommender system is based on venue data, which are the venues (airports, shopping malls, schools, etc.) surrounding the places we are interested in. Our hypothesis here is that the types of venues surrounding a neighborhood determines the characteristics of that neighborhood. If Shake Shack is mainly operating in busy commercial districts, these districts may have share similarities in the venue data, such as the high density of shopping malls and public transportation. Therefore, intuitively we will find similar neighborhoods by looking for areas with high density of shopping malls and public transportation. What machine learning makes a difference is that it can process hundreds of categories of venues and do clustering analysis, which is too complicated for a human analyst to perform.

Venue data for both sample stores and target neighborhoods are acquired via Foursquare API:

Out[8]:

	Name	Center Latitude	Center Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Shake Shack in NYC	40.71056	-74.009014	Shake Shack	40.710703	-74.009024	Burger Joint
1	Shake Shack in NYC	40.71056	-74.009014	Anthropologie	40.710618	-74.009661	Women's Store
2	Shake Shack in NYC	40.71056	-74.009014	Chick-fil-A	40.710419	-74.008550	Fast Food Restaurant
3	Shake Shack in NYC	40.71056	-74.009014	The Assemblage John Street	40.710104	-74.008574	Coworking Space
4	Shake Shack in NYC	40.71056	-74.009014	Nobu Downtown	40.710532	-74.009593	Japanese Restaurant

Out[20]:

	Name	Center Latitude	Center Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Convention District	43.156389	-77.609167	Morton's The Steakhouse	43.156566	-77.608455	Steakhouse
1	Convention District	43.156389	-77.609167	Rochester Riverside Hotel	43.156826	-77.609954	Hotel
2	Convention District	43.156389	-77.609167	Starbucks	43.156616	-77.608549	Coffee Shop
3	Convention District	43.156389	-77.609167	Hyatt Regency Rochester	43.156469	-77.608561	Hotel
4	Convention District	43.156389	-77.609167	Hyatt Focus Lounge	43.156549	-77.608646	Bar

A data preprocessing and cleaning is also required. We transform the venue data into one-hot encoding, taking averages and get the following:

	Name	Airport Lounge	BBQ Joint	Bakery	Bar	Basketball Stadium	Bubble Tea Shop	Burger Joint	Café	Clothi St
0	Shake Shack in NYC	0.013889	0.013889	0.013889	0.027778	0.013889	0.013889	0.138889	0.013889	0.013

	Name	Airport Lounge	BBQ Joint	Bakery	Bar	Basketball Stadium	Bubble Tea Shop	Burger Joint	Café	Clothing Store
0	ABC Streets Neighborhood	0	0	0.000000	0.2	0	0	0.000000	0.000000	0
1	Bull's Head	0	0	0.000000	0.0	0	0	0.000000	0.000000	0
2	Changing of the Scenes	0	0	0.000000	0.0	0	0	0.000000	0.000000	0
3	College Town	0	0	0.076923	0.0	0	0	0.076923	0.076923	0
4	Convention District	0	0	0.000000	0.2	0	0	0.000000	0.000000	0

III. Methodology

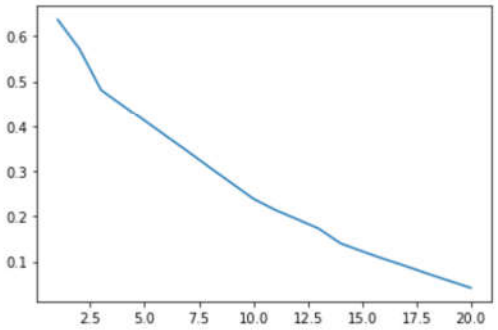
Identify the Optimal K Value

The K-means clustering clusters data into K groups based on their features. Neighborhoods with similar venues are more likely to be grouped into the same group. Before training the model, we need to determine the best K value for neighborhoods in Rochester.

Firstly, we try the common “Elbow Method”. We test the K value from 1-20.

In [31]:
```
kmax = 20
errors = []
for k in range(1, kmax+1):
    error = 0
    kmeans = KMeans(n_clusters=k, random_state=0).fit(k_clustering)
    centroids = kmeans.cluster_centers_
    label = kmeans.labels_
    for i in range(0, len(k_clustering)):
        groupid = label[i]
        centroid = centroids[groupid]
        dist = linalg.norm(k_clustering.iloc[i].to_numpy() - centroid)
        error += dist
    errors.append(error/len(k_clustering))
print(errors)
```

In [32]:
```
x_axis = range(1,kmax+1)
plt.plot(x_axis, errors)
plt.show()
```



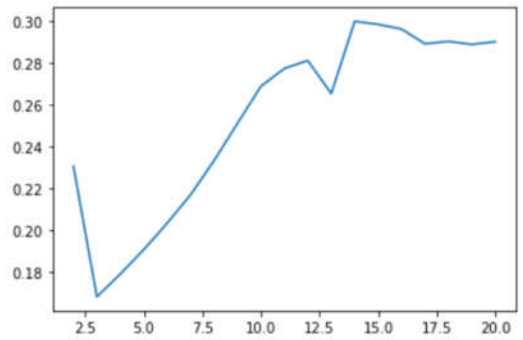
Unfortunately, there is no clear “elbow point” visible in this graph. Therefore, another approach, the Silhouette Method, is attempted.

In [33]:
```
from sklearn.metrics import silhouette_score

sil = []
kmax = 20

# dissimilarity would not be defined for a single cluster, thus, minimum number of cluster
for k in range(2, kmax+1):
  kmeans = KMeans(n_clusters=k, random_state=0).fit(k_clustering)
  labels = kmeans.labels_
  sil.append(silhouette_score(k_clustering, labels, metric = 'euclidean'))
```

In [34]:
```
x_axis = range(2,kmax+1)
plt.plot(x_axis, sil)
plt.show()
```



The global maximum point is the optimal k value, which is 14 groups. That is, neighborhoods in Rochester can be clustered into 14 categories.

K-means Clustering

With the selected K value 14, we use sklearn kit for k-means training. Neighborhoods in Rochester are clustered into 14 different groups, with label 0 to label 14.

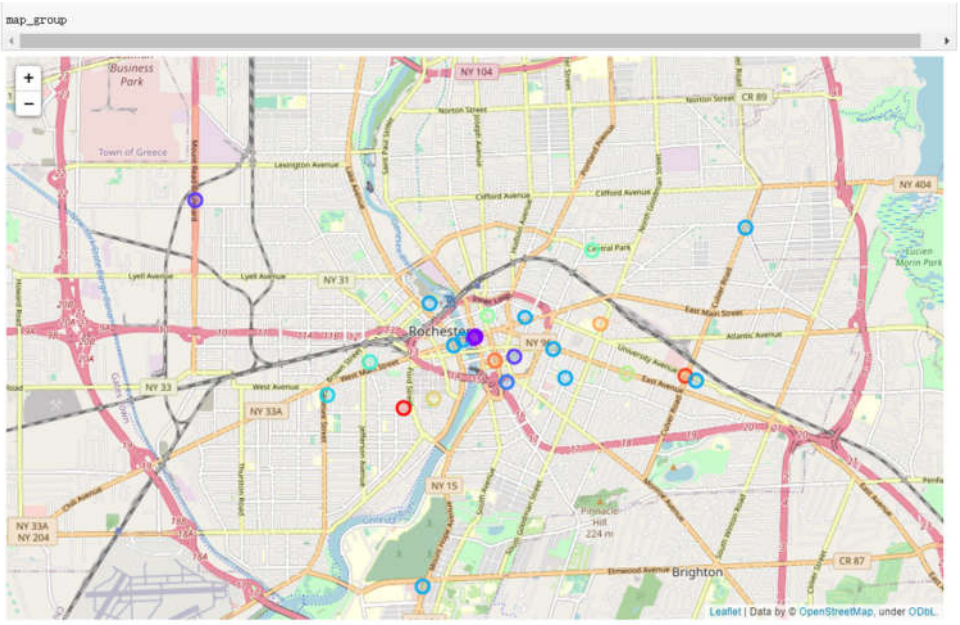
```
In [35]: kclusters = 14
         kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(k_clustering)
         kmeans.labels_
```

```
Out[35]: array([ 1,  5,  0,  4,  1, 10, 13,  4,  9,  4,  4,  4,  4,  1,  4,  1,  2,
                 2, 11,  4,  7,  4,  1,  8,  6,  4,  3, 12,  1])
```

Out[38]:

	Cluster_Labels	Name	Latitude	Longitude
0	1	ABC Streets Neighborhood	43.156578	-77.608846
1	5	Bull's Head	43.148933	-77.635535
2	0	Changing of the Scenes	43.147237	-77.621773
3	4	College Town	43.123889	-77.618333
4	1	Convention District	43.156389	-77.609167

Out[39]:



Recommendation

With the K-means clustering result, the program is able to generate automated

recommendation on the best location for business expansion. It compares the centroids of each group with the Shake Shack data and rank the 14 groups from the most similar to the least similar. After that, it marks the most similar group as “Top Recommended”, first 25% groups as “Strongly Recommended”, first 50% as “Weakly Recommended”, and the rest as “Not Recommended”.

In [41]:
```
min_dist = sys.maxsize
index = 0
for i in range(len(target_centroids)):
    d = linalg.norm(target_centroids[i] - sample_centroid)
    if min_dist > d:
        index = i
        min_dist = d
print(f'Group {index} is the closest cluster to the sample centroid.')
```

```
Group 4 is the closest cluster to the sample centroid.
```

In [42]:
```
group_rank = {}
for i in range(len(target_centroids)):
    d = linalg.norm(target_centroids[i] - sample_centroid)
    group_rank[i] = d

sort_orders = sorted(group_rank.items(), key=lambda x: x[1], reverse=False)
rank_list = []
for i in sort_orders:
    print(i[0], i[1])
    rank_list.append(i[0])
```

```
4 0.21682063561003495
1 0.510597568093709
13 0.6079463507844247
6 0.7323014625813371
8 0.7323014625813371
12 0.7323014625813371
0 1.0179712334338165
2 1.0179712334338165
3 1.0179712334338165
5 1.0179712334338165
7 1.0179712334338165
9 1.0179712334338165
10 1.0179712334338165
11 1.0179712334338165
```

Out[43]:
```
{4: 'Top Recommended',
 1: 'Strongly Recommended',
 13: 'Strongly Recommended',
 6: 'Weakly Recommended',
 8: 'Weakly Recommended',
 12: 'Weakly Recommended',
 0: 'Weakly Recommended',
 2: 'Not Recommended',
 3: 'Not Recommended',
 5: 'Not Recommended',
 7: 'Not Recommended',
 9: 'Not Recommended',
 10: 'Not Recommended',
 11: 'Not Recommended'}
```

IV. Results

The results obtained are shown below. We give each neighborhood a recommendation level for opening a Shake Shack store. There’s also a visualization of our recommendation results.

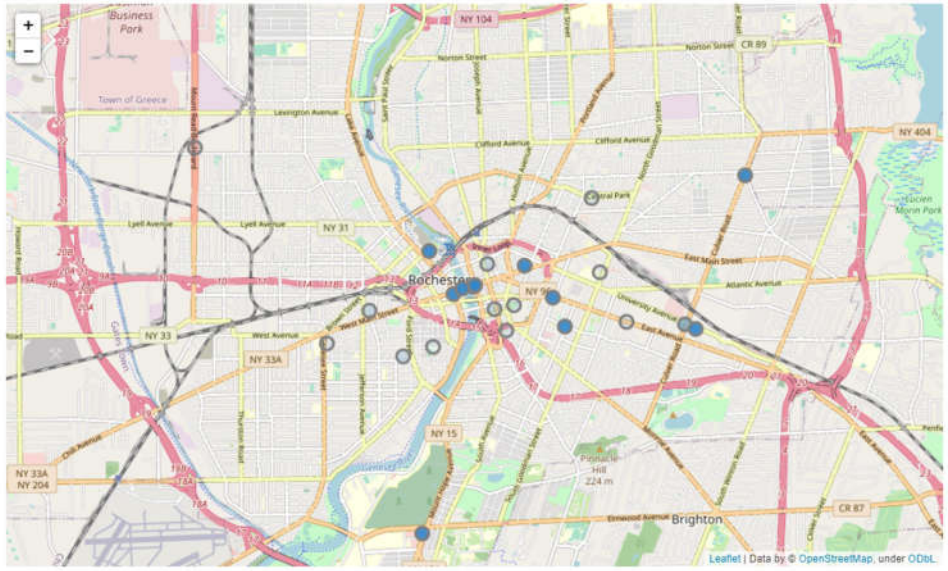
Circles with darker colors are more recommended and circles with lighter colors are less recommended.

Out[44]:

	Cluster_Labels	Name	Latitude	Longitude	Recommendation
0	1	ABC Streets Neighborhood	43.156578	-77.608846	Strongly Recommended
1	5	Bull's Head	43.148933	-77.635535	Not Recommended
2	0	Changing of the Scenes	43.147237	-77.621773	Weakly Recommended
3	4	College Town	43.123889	-77.618333	Top Recommended
4	1	Convention District	43.156389	-77.609167	Strongly Recommended

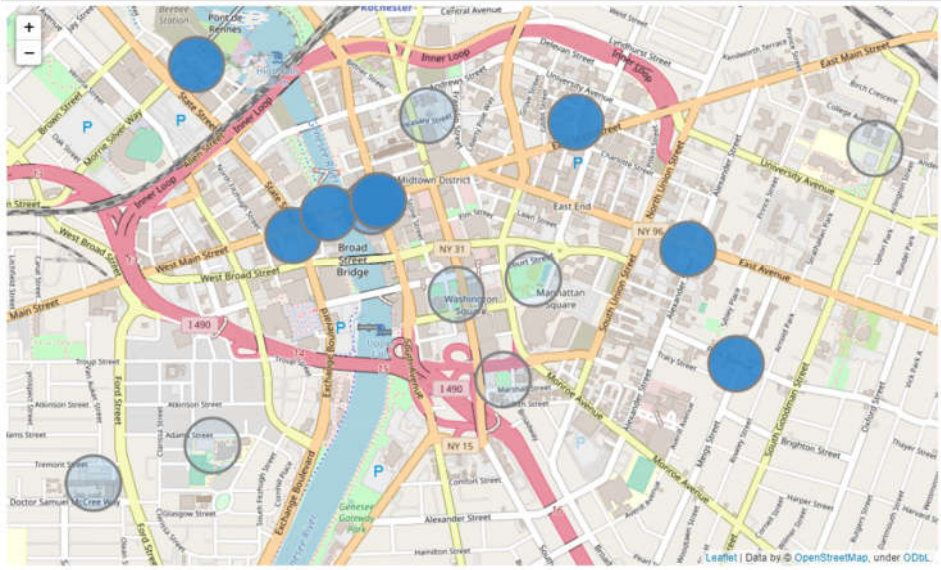
```
map_rec
```

Out[45]:




```
map_rec
```

Out[45]:



V. Discussion

From the "Shake Shack in Rochester" example, we can observe that the recommended places are mostly clustered in the downtown of Rochester, as well as a few other places, such as the College Town and Marketplace, outside the downtown. This is because according to the venue data we obtained, these places are the most similar to the neighborhoods of existing Shake Shack stores in NYC, based on the analysis of the surrounding environment. If Shake Shack decided to expand its business to Rochester, the top-recommended placed above would likely to be the potential locations for new Shake Shack stores. This matches our hypothesis, as Shake Shack commonly sells in busy, middle-to-upper class, urban areas, and the downtown and college town are the places in Rochester most likely to fit in that category.

VI. Conclusion

This project explores a location-based recommender system on potential business expansions. It tests with the “Shake Shack in Rochester” example and effectively identifies potential locations that best fit Shake Shack’s franchise. The pipeline and code in this project are highly generic and can be easily and widely applied to any similar business problems. It can help business owners, market analysts, and decision makers quickly locate the areas best fit for expansions, helping their business growing from the local to the global.