Finding an optimal location for opening an Indian Restaurant in New York

Shahbaaz Khan

February 16, 2019

I. Introduction

In this project we will try to find an optimal location for a restaurant. Specifically, this report will be targeted to stakeholders interested in opening an **Indian restaurant** in **New York**, USA.

This report will try to answer questions such as: Should I place my restaurant in a place where there are a lot of other Indian Restaurants or in a place where there is none or should I try to find some kind of balance between these two extremes?

We will use our data science powers to generate a few most promising neighborhoods based on these criteria.

II. Data

Based on definition of our problem, factors that will influence our decision are:

- find the high rating Indian restaurants in NY and identify them visually on a map
- create several "clusters" of these Indian Restaurants to identify potential districts to choose from

We should choose one from these business districts and locate our restaurant in the center of it. The closer to the center, the high will be the rent. This is another choice that will be made by us referring to our budget.

To simplify the problem, we will only use the geographic location to cluster the cafes. For this we will use the "explore" endpoint. The request url is "https://api.foursquare.com/v2/venues/explore". According to the document, in the request, we should pass following parameters: "section=Indian restaurant" and "near=New York, NY". In the response, we are interested in groups.items.categories, groups.items.venue.name, group.items.venue.location. To simplify the problem, we will only use the geographic location to cluster the restaurants.

Following data sources will be needed to extract/generate the required information:

- number of restaurants and their type and location in every neighborhood will be obtained using **Foursquare API**

III. Methodology

The basic idea behind the solution is that for locating a new Indian restaurant, you should find where the successful restaurants are now in, since that location is tested by experience. When we get a bunch of restaurants, the data is clustered to find several concentrations that will be targeted by the new restaurants. The **intended location** is the **geographic average of the location** of the currently successful restaurants.

a. Getting the New York location through Geolocator

```
address = '102 North End Ave, New York, NY'

geolocator = Nominatim()
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print(latitude, longitude)
```

b. Getting the Indian restaurants location details from Foursquare API

```
search_query = 'Indian Restaurant'
radius = 5000
print(search_query + ' .... OK!')
```

```
Indian Restaurant .... OK!
```

```
url = 'https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&ll={},{}&v={}&query={}&radius={}&limit={}'.
```

c. Storing the Foursquare results (.json) in a file

```
results = requests.get(url).json()
# results
```

d. Transforming the file into a pandas dataframe

```
df=pd.DataFrame()
df['venue_ID']=dataframe['venue.id']
df['name']=dataframe['venue.name']
df['lat']=dataframe['venue.location.lat']
df['lng']=dataframe['venue.location.lng']
df.head()
```

]:

	venue_ID	name	lat	lng
0	4bbb9dbded7776b0e1ad3e51	Tamarind TriBeCa	40.719211	-74.008727
1	5a1e961c1987ec47beed877d	Baar Baar	40.724534	-73.991624
2	5b770657c0cacb002c89bc63	The Kati Roll Company	40.709114	-74.009091
3	575dea4c498e2739e43a27e2	Aahar Indian Cuisine	40.713307	-74.007994
4	4593ed04f964a52050401fe3	The Kati Roll Company	40.729570	-74.000861

e. Using K-means clustering to cluster the locations

```
k_means = KMeans(init = "k-means++", n_clusters = 4, n_init = 12)
df_array=np.array(df[['lat','lng']])
k_means.fit(df_array)
k_labels=pd.DataFrame(k_means.labels_)
k_labels.info()
df['cluster']=k_labels
df.groupby('cluster').count()
```

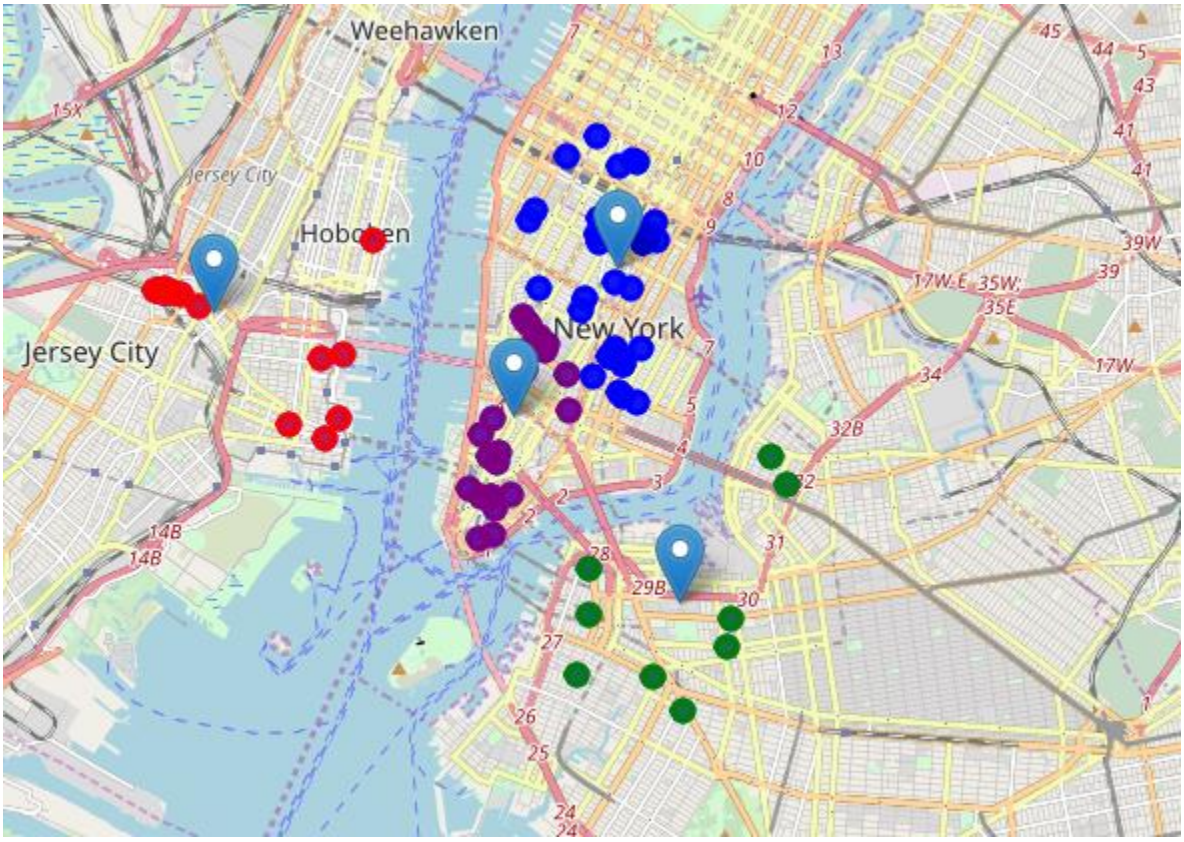
cluster	venue_ID	name	lat	lng
0	44	44	44	44
1	21	21	21	21
2	9	9	9	9
3	25	25	25	25

f. Finding the recommended locations by using geographic average of the current location

```
center_location=df.groupby('cluster').mean()
center_location
```

IV. Results

I use the dataset to create four clusters, as shown in the below figure. The geographic average of this location is marked and recommended the location for opening the business.



V. Discussion

This solution is based purely on the geographic location, which is quite good in most cases. Since in the business of restaurants, the right location is one of the most important decisions to make. Nevertheless, if we could find more information on each cluster, it will improve the solution further.

VI. Conclusion

The clustering is very useful and intuitive to solve the problem related to geographic data. We should also pay attention to the outcomes of the clustering, especially when it yields a result of different ideas of a 'business circle' in a city. It has two potential reasons. One is that the clustering is simply wrong. But more likely, the result may present a new business opportunity. In the latter case, we should examine the features we included in the model to find the cause of

the result. If we can identify the reason that causes the clustering, it might be a valuable business insight.

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.