Coursera Applied Data Science Capstone Project

Opening a new Business in city of Toronto

Author: Arjun P

1. Introduction:

Toronto is the provincial capital of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America. The diverse population of Toronto reflects its current and historical role as an important destination for immigrants to Canada. The city continues to grow and attract immigrants. A study by Ryerson University showed that Toronto was the fastest-growing city in North America. The city added 77,435 people between July 2017 and July 2018. The Toronto metropolitan area was the second-fastest-growing metropolitan area in North America.

Toronto is Canada's business and financial capital, a growing financial hub in North America, and a top ten global financial centre. Toronto's Gross Domestic Product (GDP) growth is significantly outpacing the national average.

So, it's always a good opportunity for a prospective entrepreneur to start a new business in the city of Toronto. Because of the ever-increasing population and the rapid growth rate of city, coupled with its diverse population, there are ample business opportunities in the city of Toronto. But the entrepreneur needs to know which area is suitable for starting his business, so the selection of location for the business venture is one of the most important decisions that will determine the success of the business.

2. Business Problem:

The objective of this project is to propose the most suitable location for starting a new business in the city of Toronto. Here I make use of Data science methods and tools to list out the business ventures/venues in the city of Toronto and based on the business venture selected by the user, suitable locations in the city of Toronto can be suggested to the entrepreneur using Foursquare location data and Machine Learning techniques like K-Means clustering and also visualization of maps using Folium. Thus, this project aims to find solution to the business question: In Toronto, if an entrepreneur wants to start a business which location can be suggested for opening the business for its success?

3. Target Audience:

The project is targeted at entrepreneurs and investors who consider opening a new business or investing in a new business in the city of Toronto, since, Toronto is competitive in almost every major business sector from technology and life sciences to green energy; from fashion and design to food and beverage; from film and television production to music and digital media.

4. Data section:

The objective of this project is to propose the most suitable location for starting a new business in the city of Toronto. Since Toronto is a large metropolis with diverse business venues and ventures, we require acute and exhaustive data corresponding to Toronto neighborhoods and venues.

To solve this problem, the following Data are needed:

1. List of Boroughs and Neighborhoods in Toronto, Canada
    - Data source: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
    - Description: This is a list of postal codes in Canada where the first letter is M. Postal codes beginning with M are located within the city of Toronto in the province of Ontario.
    - Data extraction: A python code is written to scrap this Wikipedia page and extract the required data using pandas read_html method.

2. Latitude and Longitudinal co-ordinates of the Neighborhoods in Toronto,Canada
    - Data source: http://cocl.us/Geospatial_data , which is a .csv file
    - Description: This is a .csv file that contains the list of neighborhoods in Toronto with their Latitude and Longitudinal co-ordinates which is used to make Foursquare API calls.
    - Data extraction: A python code is written to read the .csv file into a pandas dataframe for further processing

3. The data corresponding to various venues/businesses and their categories in Toronto neighborhoods. This data will be used for clustering the neighborhoods.
    - Data source: https://api.foursquare.com. We use Foursquare API to get the venue and venue category data related to each neighborhood. The exact url for data extraction is: https://api.foursquare.com/ v2/venues/explore?&client_id=CLIENT_ID& client_secret=CLIENT_SECRET&v=VERSION&ll=lat,lng&radius=1000&limit=100
    - Description: This is a .json file that contains the details about a particular neighborhood location and 100 different venues around each neighborhood within a radius of 1km.

- Data extraction: A python code is written to read the required fields in the .json file into a pandas dataframe for further processing.

5. Methodology:

5.1 First requirement is to get the list of neighborhoods in the city of Toronto. This was doneby extracting the data from the Wikipedia page:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

which contains the neighborhood data for the city of Toronto. I have used the pandas read_html method to do web scraping and the resulting table was read into a pandas dataframe. But the initial list contained only the Postal code, Borough and Neighborhood names. In order to do analysis using Foursquare API, I need the location co-ordinates of the neighborhoods. This was obtained from http://cocl.us/Geospatial_data , which is a .csv file provided by IBM team. This data was also read into dataframe. The next step was to add the latitude and longitudinal values of the neighborhoods to the original data frame containing Borough and Neighborhood names using pandas left merging technique. A snippet of the final dataframe would look like as shown below:

[9]:
```
tor_neigh=tor_cord.copy()
tor_neigh = tor_neigh[tor_neigh.Borough.str.contains("Toronto")]
tor_neigh = tor_neigh.reset_index(drop=True)
tor_neigh
```

[9]:

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
4	M4E	East Toronto	The Beaches	43.676357	-79.293031
5	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306
6	M5G	Downtown Toronto	Central Bay Street	43.657952	-79.387383
7	M6G	Downtown Toronto	Christie	43.669542	-79.422564
8	M5H	Downtown Toronto	Richmond, Adelaide, King	43.650571	-79.384568
9	M6H	West Toronto	Dufferin, Dovercourt Village	43.669005	-79.442259
10	M5J	Downtown Toronto	Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752
11	M6J	West Toronto	Little Portugal, Trinity	43.647927	-79.419750
12	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188
13	M5K	Downtown Toronto	Toronto Dominion Centre, Design Exchange	43.647177	-79.381576
14	M6K	West Toronto	Brockton, Parkdale Village, Exhibition Place	43.636847	-79.428191
15	M4L	East Toronto	India Bazaar, The Beaches West	43.668999	-79.315572

5.2 After that I selected only those Boroughs that contain the word Toronto, so as to focus the search on Toronto neighborhoods. Then I used geopy library to get the latitude and longitude values of Toronto and visualized the map of Toronto with neighborhoods superimposed on top using Folium package.

5.3 Next I used the Foursquare API to pull the list of top 100 venues within 1km radius of each neighborhood. For this first I created a Foursquare developer account to obtain a client ID and client secret which is required for pulling data using Foursquare. From Foursquare I got the name, latitude, longitude and category of venues of all the neighborhoods as a json file which was read into a pandas dataframe for further processing. Then I analyzed each neighborhood by grouping the dataframe rows by neighborhoods and then taking the mean of the frequency of occurrence of each venue category using onehot encoding technique.

5.4 Here the idea is to let the user(entrepreneur) select a business of his choice and the algorithm will suggest him which cluster of neighborhoods is suitable for him to start the new business. For this first I found out the total number of unique venues/businesses in Toronto and then put that as a Table containing all of the Unique Venues/Businesses Categories in Toronto with an index number assigned to each venue/business category using python and pandas methods. A snippet of the table is shown below:

```
list_of_business.sort_values([ 'Business/Venue Category' ], inplace=True)
list_of_business = list_of_business.reset_index(drop=True)
list_of_business.index.name = 'Business Index No:'
list_of_business
```

[24]:

Business Index No:	Business/Venue Category
0	Afghan Restaurant
1	Airport
2	Airport Food Court
3	Airport Lounge
4	Airport Service
5	Airport Terminal
6	American Restaurant
7	Antique Shop
8	Aquarium
9	Art Gallery
10	Art Museum
11	Arts & Crafts Store
12	Asian Restaurant
13	Athletics & Sports

5.5 Next the algorithm will prompt the user to enter the Venue/Business category Index No: from the above table, that matches to the Business he wishes to open. Once the user enters the Business index number, a function defined and called in the algorithm will list the selected business and return a Dataframe containing the Neighborhoods and the selected

business frequency in each neighborhood, based on the Business Index No. selected by the user as shown below:

```
]: business_cat = int(input("Enter the Business Index No: corresponding to the Business you wish to open by referring the list_of_business table:"))
   tor_new_business = business_name(business_cat)
   tor_new_business.head()
Enter the Business Index No: corresponding to the Business you wish to open by referring the list_of_business table:
```

Suppose an entrepreneur wishes to open an Italian Restaurant. From the Business Index Table, the index no of Italian Restaurant is 132 and so all the user needs to do is to input 132 when prompted and the function will list out the selected business along with a dataframe that contains the details of the frequency of occurrence of the Business category entered by user, here for eg: Italian Restaurant.

```
[57]: business_cat = int(input("Enter the Business Index No: corresponding to the Business you wish to open by referring the list_of_b
      tor_new_business = business_name(business_cat)
      tor_new_business.head()
Enter the Business Index No: corresponding to the Business you wish to open by referring the list_of_business table: 132
The selected Business is: Italian Restaurant
```

[57]:

	Neighborhoods	Italian Restaurant
0	Berczy Park	0.000000
1	Brockton, Parkdale Village, Exhibition Place	0.045455
2	Business reply mail Processing Centre, South C...	0.000000
3	CN Tower, King and Spadina, Railway Lands, Har...	0.000000
4	Central Bay Street	0.060606

5.6 Now we perform clustering of the data in the dataframe returned by the function using K-Means clustering algorithm, which is one of the most popular unsupervised machine

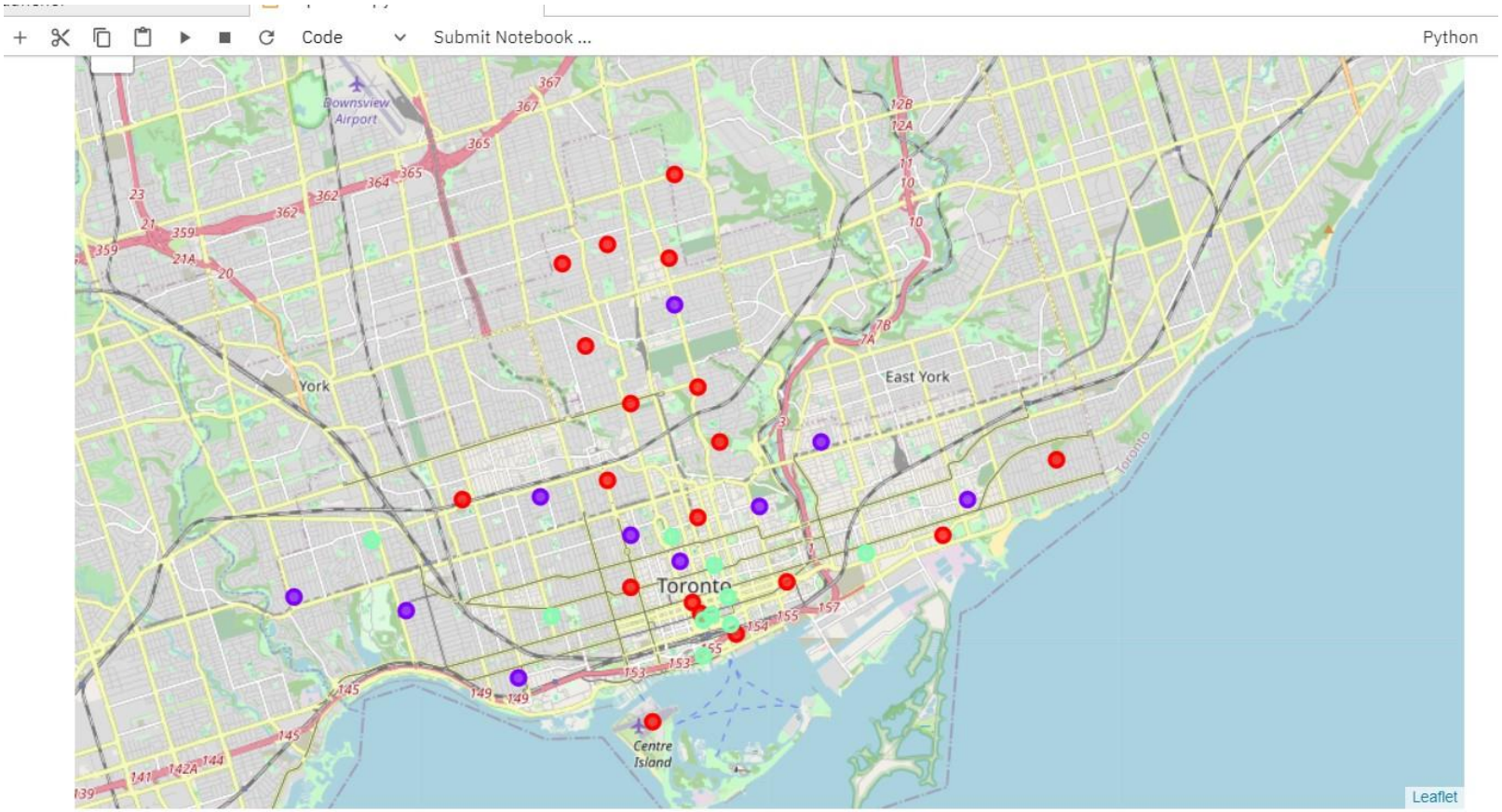
learning algorithms. It is particularly suited for solving the clustering problems like the one discussed in this project. Here we cluster the neighborhoods into 3 Clusters,viz, Cluster 0, Cluster 1 and Cluster 2, based on the frequency of occurrence of the selected business category. The cluster labels and latitude and longitudinal values are merged to the dataframe.

5.7 Next I have visualized the clusters using Folium package and then displayed each cluster as a dataframe to examine the clusters.

5.8 Then I have defined a function to automatically recommend the Best cluster and the second best cluster to start the Business. This function automatically suggests the Best Cluster and second best cluster using the Mean value of the frequency of occurrence of the selected Business in each Cluster. So it will automatically suggest the user which cluster of neighborhoods is best suited for starting the new business.

6. Results:

6.1 The clustered neighbourhoods based on the frequency of occurrence of the business category selected (here for eg. Italian Restaurant) are visualized in the Folium map as shown below:



Here Red markers indicate Cluster 0, Purple markers indicate Cluster 1 and Green markers indicate Cluster 2.

6.2 Each of the individual clusters are listed as a dataframe for analysis and conclusion. First 5 neighborhoods in each cluster are as follows:

25. Examine Clusters.

25.1 Cluster 0

[72]:
```
clus_0 = tor_new_business_merged.loc[tor_new_business_merged['Cluster Labels'] == 0]
clus_0.head()
```

[72]:

	Neighborhood	Italian Restaurant	Cluster Labels	Latitude	Longitude
0	Berczy Park	0.0	0	43.644771	-79.373306
24	Regent Park, Harbourfront	0.0	0	43.654260	-79.360636
18	Lawrence Park	0.0	0	43.728020	-79.388790
17	Kensington Market, Chinatown, Grange Park	0.0	0	43.653206	-79.400049
25	Richmond, Adelaide, King	0.0	0	43.650571	-79.384568

25.2 Cluster 1

[73]:
```
clus_1 = tor_new_business_merged.loc[tor_new_business_merged['Cluster Labels'] == 1]
clus_1.head()
```

[73]:

	Neighborhood	Italian Restaurant	Cluster Labels	Latitude	Longitude
30	St. James Town, Cabbagetown	0.044444	1	43.667967	-79.367675
28	Runnymede, Swansea	0.050000	1	43.651571	-79.484450
36	The Danforth West, Riverdale	0.069767	1	43.679557	-79.352188
38	University of Toronto, Harbord	0.055556	1	43.662696	-79.400049
16	India Bazaar, The Beaches West	0.047619	1	43.668999	-79.315572

25.3 Cluster 2

[74]:
```
clus_2 = tor_new_business_merged.loc[tor_new_business_merged['Cluster Labels'] == 2]
clus_2.head()
```

[74]:

	Neighborhood	Italian Restaurant	Cluster Labels	Latitude	Longitude
23	Queen's Park, Ontario Provincial Government	0.030303	2	43.662301	-79.389494
37	Toronto Dominion Centre, Design Exchange	0.030000	2	43.647177	-79.381576
15	High Park, The Junction South	0.040000	2	43.661608	-79.464763
14	Harbourfront East, Union Station, Toronto Islands	0.030000	2	43.640816	-79.381752
13	Garden District, Ryerson	0.020000	2	43.657162	-79.378937

6.3 Thus we observe that:

- Cluster 0 has very few number of the selected business category, viz, the Italian restaurant.
- Cluster 1 has got a large concentration of the selected business category, viz, the Italian restaurant.
- Cluster 2 has got a moderate concentration of the selected business category, viz, the Italian restaurant.

6.4 Then the algorithm will automatically suggest the Best Cluster and the second best cluster using the Mean value of the frequency of occurrence of the selected Business in each Cluster. For this I defined a function 'clus_choice' to recommend the Best cluster and the second best cluster to start the Business. The mean value of frequency of occurrence of the selected business in each cluster is found out and is assigned to a list. This list is the given

as the argument of the 'clus_choice' function which returns the Best cluster and the second best cluster to start the Business. The function selects the best cluster based on the logic that the cluster with the lowest mean value of frequency will have least count of the selected business and hence they offer an ideal opportunity for the entrepreneur to start the new business because of less competition and early mover advantage.

26. To Automatically suggest the Best Cluster using the Mean value of the frequency of occurrence of the selected Business in each Cluster.

26.1 To find the mean value of frequency of occurence of the selected business in each cluster and assign them to k0,k1 and k2 for Clusters 0,1 and 2 respectively.

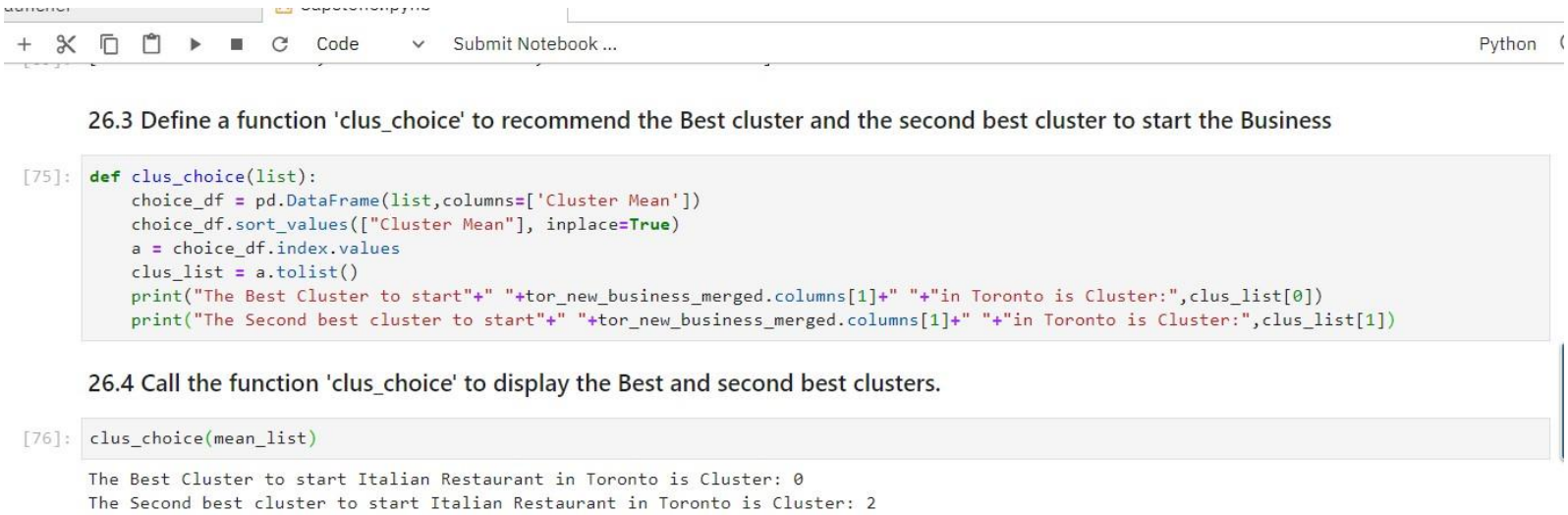
[68]:
```
k0 = clus_0.iloc[:,1].mean()
k1 = clus_1.iloc[:,1].mean()
k2 = clus_2.iloc[:,1].mean()
print("Mean value of frequency of occurence of"+" "+tor_new_business_merged.columns[1]+" "+"in Cluster 0 is k0 =",k0)
print("Mean value of frequency of occurence of"+" "+tor_new_business_merged.columns[1]+" "+"in Cluster 1 is k1 =",k1)
print("Mean value of frequency of occurence of"+" "+tor_new_business_merged.columns[1]+" "+"in Cluster 2 is k2 =",k2)
```

```
Mean value of frequency of occurence of Italian Restaurant in Cluster 0 is k0 = 0.0005263157894736842
Mean value of frequency of occurence of Italian Restaurant in Cluster 1 is k1 = 0.05562753783684017
Mean value of frequency of occurence of Italian Restaurant in Cluster 2 is k2 = 0.028459916372287508
```

26.2 Save the mean values in a list called clus_list.

[69]:
```
mean_list=[k0,k1,k2]
mean_list
```

[69]:
```
[0.0005263157894736842, 0.05562753783684017, 0.028459916372287508]
```



7. Discussion and Recommendation:

7.1. As discussed in the Results section we observe that, for the case of Italian Restaurant selected by the user/entrepreneur, most of them are concentrated in neighborhoods in Cluster 1 and the least concentration is observed in neighborhoods in Cluster 0. Neighborhoods in Cluster 2 seems to have a medium concentration of Italian restaurants.

7.2. So looking at nearby venues, it seems any neighborhood in Cluster 0 might be a good location for the entrepreneur to start the selected business, viz, an Italian restaurant as there are very few Italian restaurants in these areas and hence they will benefit from the early mover advantage. Also the competition will be next to null in Cluster 0, so that's an added advantage especially if the entrepreneur is a novice in the business. So the

neighborhoods in Cluster 0 are recommended as the Best locations to start the selected business, viz, an Italian restaurant.

7.3. Cluster 2 has a medium concentration of the selected business, viz, an Italian restaurant and hence this location is suggested only some to experienced and moderately experienced entrepreneurs with some unique selling propositions who has the ability and guts to stand out from the competition. Since Cluster 2 already has some Italian restaurants, the customer base is almost guaranteed and hence the entrepreneur needs to focus on capturing that customer base and attracting new ones. So neighborhoods in Cluster 2 are recommended as the second best locations to start the selected business, viz, an Italian restaurant

7.4. Cluster 1 which already has a large concentration of the selected business, viz, an Italian restaurant has got fierce competition. So neighborhoods in this cluster are only suggested to a few highly experienced and established brands who, with their experience and brand value, can devise methods to capture customers.

7.5. Thus we can also verify that this analysis exactly matches with the predictions of the "clus_choice" function which automatically suggests the Best and second best Clusters to start the business.

8. Suggestions for future research:

8.1. The analysis can be expanded by expanding the data set to include the population data and income data of the residents and also by including the migrant population data in that area. That is expected to a give more focused and accurate analysis and when we rope in more data sets, the number of clusters can also be increased to get a more detailed analysis of the neighborhoods, which in turn can improve the accuracy of our predictions.

9. Conclusion:

9.1. In this project we have gone through the extensive process of Identifying the business problem, extracting the required data sets, data preparation, performing machine learning techniques on the data set, visualize the results and finally recommending the best options to the relevant stakeholders, in this case an entrepreneur who wishes to start a new business in the city of Toronto.

9.2. Thus this project aims to help a prospective entrepreneur who wishes to start a new business in the city of Toronto to select a particular neighborhood, based on the distribution and concentration of the selected business in Toronto neighborhoods.

10. References:

1. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
2. http://cocl.us/Geospatial_data
3. Foursquare developer documentation : https://developer.foursquare.com/docs/
4. Pandas documentation : https://pandas.pydata.org/docs/
5. Folium documentation : https://pypi.org/project/folium/
6. Scikit-learn KMeans documentation: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
7. Geopy documentation : https://pypi.org/project/geopy/