Coursera Applied Data Science Capstone Project

Opening A New Business In The City Of Toronto

By: ARJUN P

Introduction:

- Toronto is the provincial capital of Ontario.
- Toronto is Canada’s business and financial capital, a growing financial hub in North America, and a top ten global financial centre.
- 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

Business Problem:

- Project Objective: To propose the most suitable location for starting a new business in the city of Toronto.
- Use of Data science methods and tools to list out the business ventures/venues in the city of Toronto.
- Prompts user/entrepreneur to enter his choice of Business category
- Use Foursquare location data and Machine Learning techniques like K-Means clustering to suggest suitable locations in the city of Toronto.
- Visualization of maps using Folium.
- Target Audience: Entrepreneurs and investors who consider opening a new business or investing in a new business in the city of Toronto

Data Acquisition & Cleaning:

- List of Boroughs and Neighborhoods in Toronto, Canada
  - Data source: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
  - Description: 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 read_html method is used to scrap this Wikipedia page and extract the required data.

- Latitude and Longitudinal co-ordinates of the Neighborhoods in Toronto, Canada
  - Data source: http://cocl.us/Geospatial_data , which is a .csv file

Data Acquisition & Cleaning (contd.):

- 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

➢ The data corresponding to various venues/businesses and their categories in Toronto neighborhoods. This data will be used for clustering the neighborhoods.

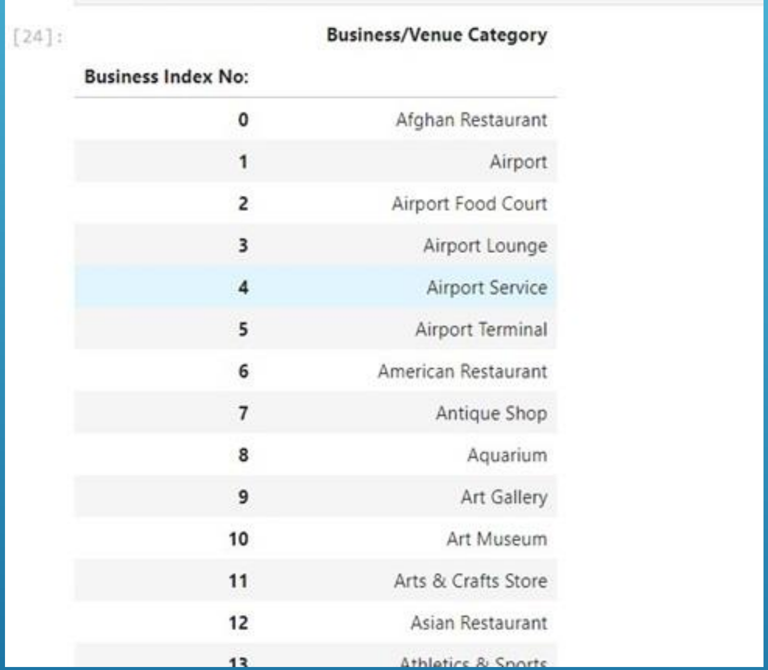
- Data source: https://api.foursquare.com. We use Foursquare API to get the venue and venue category data related to each neighborhood.
- Description: This is a .json file that contains the details about a particular neighborhood location and 100 different venues around each neighborhood within a radius 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.

Methodology:

- Extracting the data from the Wikipedia page using web scraping method of pandas
- Get location co-ordinates (latitudes & longitudes) of the neighborhoods. This was obtained from http://cocl.us/Geospatial_data
- Use 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.
- Use Foursquare API to pull the list of top 100 venues within 1km radius of each neighborhood
- Group the dataframe rows by neighborhoods and then take the mean of the frequency of occurrence of each venue category using onehot encoding technique.

Methodology (contd.):

- Display all of the Unique Venues/Businesses Categories in Toronto as a Table with an index number assigned to each venue/business category.



[24]:

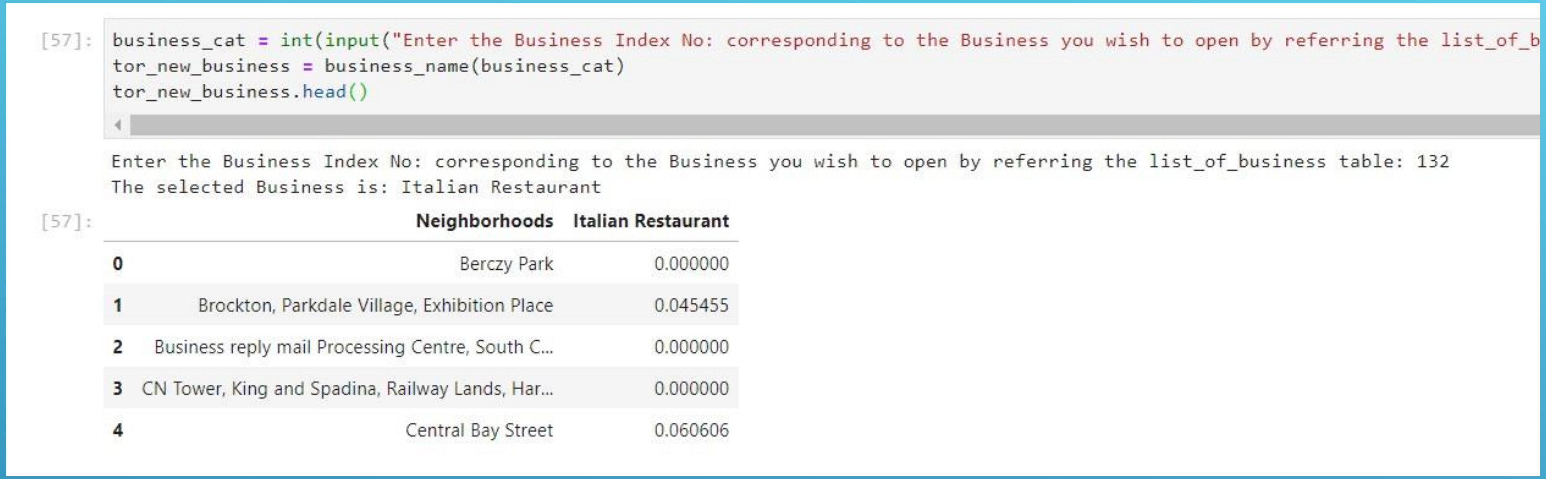
	Business/Venue Category
Business Index No:	
0	Afghan Restaurant
1	Airport
2	Airport Food Court
3	Airport Lounge
4	Airport Service
5	Airport Terminal
6	American Restaurant
7	Antique Shop
8	Aquarium
9	Art Gallery
10	Art Museum
11	Arts & Crafts Store
12	Asian Restaurant
13	Athletics & Sports

- Prompt the user to enter the Venue/Business category Index No: from the above table, that matches to the Business he wishes to open.

Methodology (contd.):

- 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
- Suppose an entrepreneur wishes to open an Italian Restaurant. So user needs input 132 which is the index no. of Italian Restaurant, 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.

Methodology (contd.):



```
[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
```

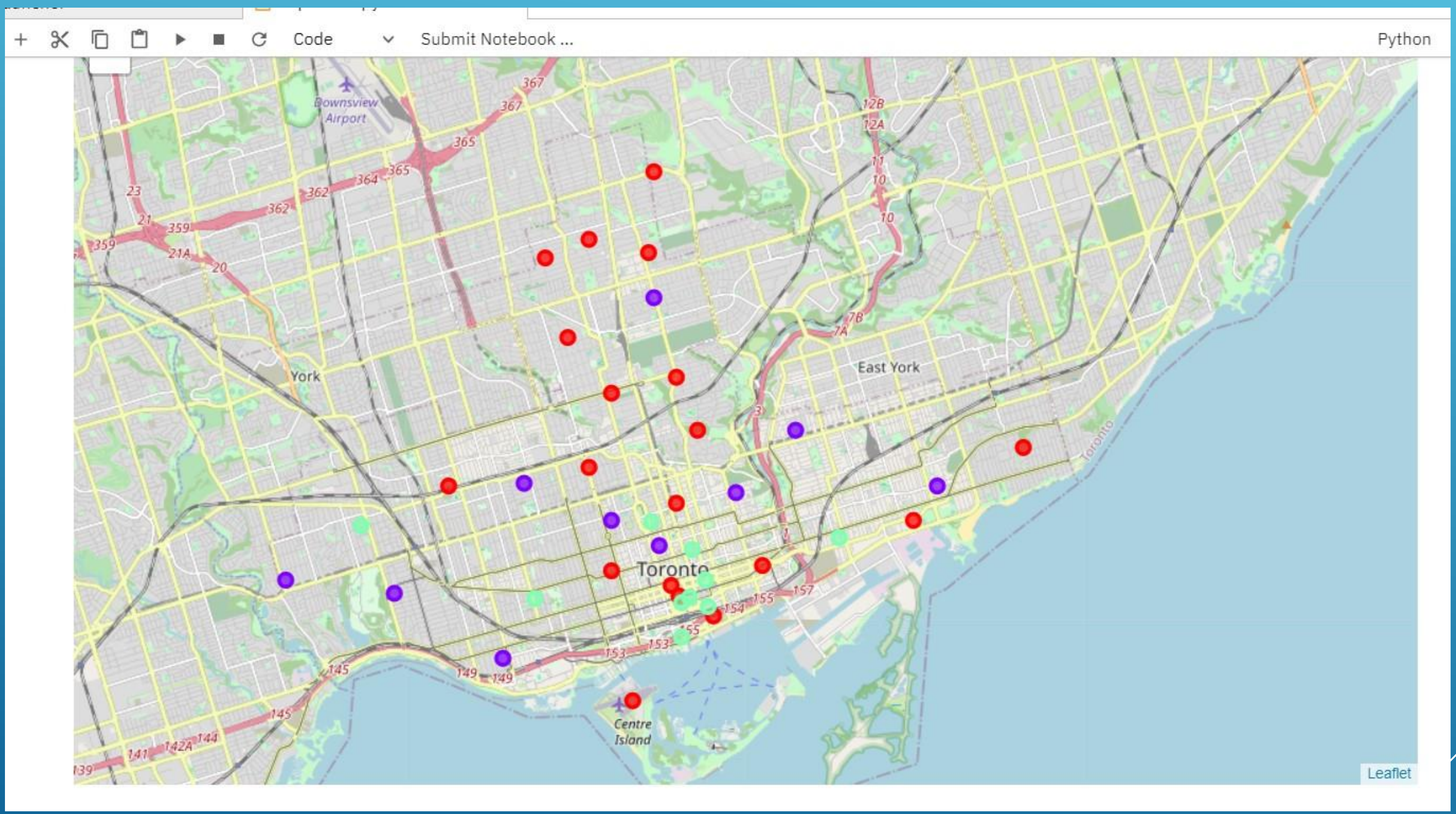
[57]:

	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

- ➢ Perform clustering of the data using K-Means clustering algorithm to cluster the neighborhoods into 3 Clusters.
- ➢ Visualize the clusters using Folium package and then display each cluster as a data frame to examine the clusters.
- ➢ Define and call a function to automatically recommend the Best cluster and the second best cluster using the Mean value of the frequency of occurrence of the selected Business in each Cluster.

Results:

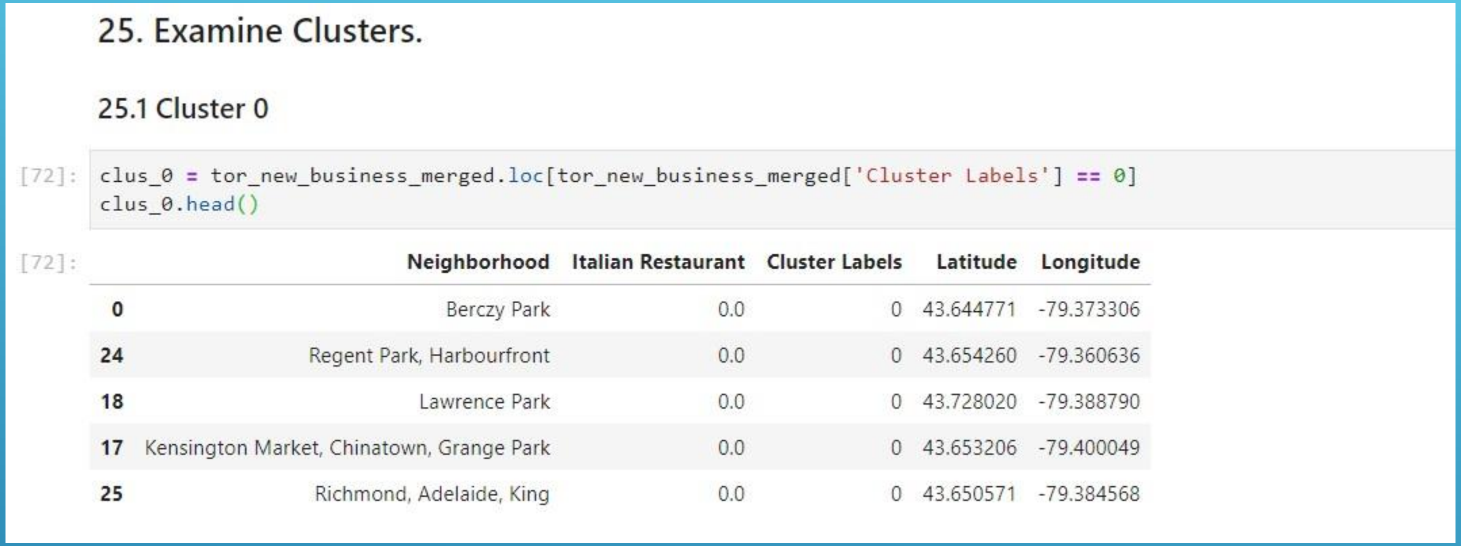
- Categorized the neighborhoods into 3 clusters based on the frequency of occurrence of the business category selected (here for eg. Italian Restaurant)



Results (contd.):

- Cluster details:
  - Cluster 0: Indicated with Red marker has very few number of the selected business category, viz, the Italian restaurant.
  - Cluster 1: Indicated with Purple marker has got a large concentration of the selected business category, viz, the Italian restaurant.
  - Cluster 2: Indicated with Green marker has got a moderate concentration of the selected business category, viz, the Italian restaurant.
- First 5 neighborhoods in each cluster are listed:

Results (contd.):



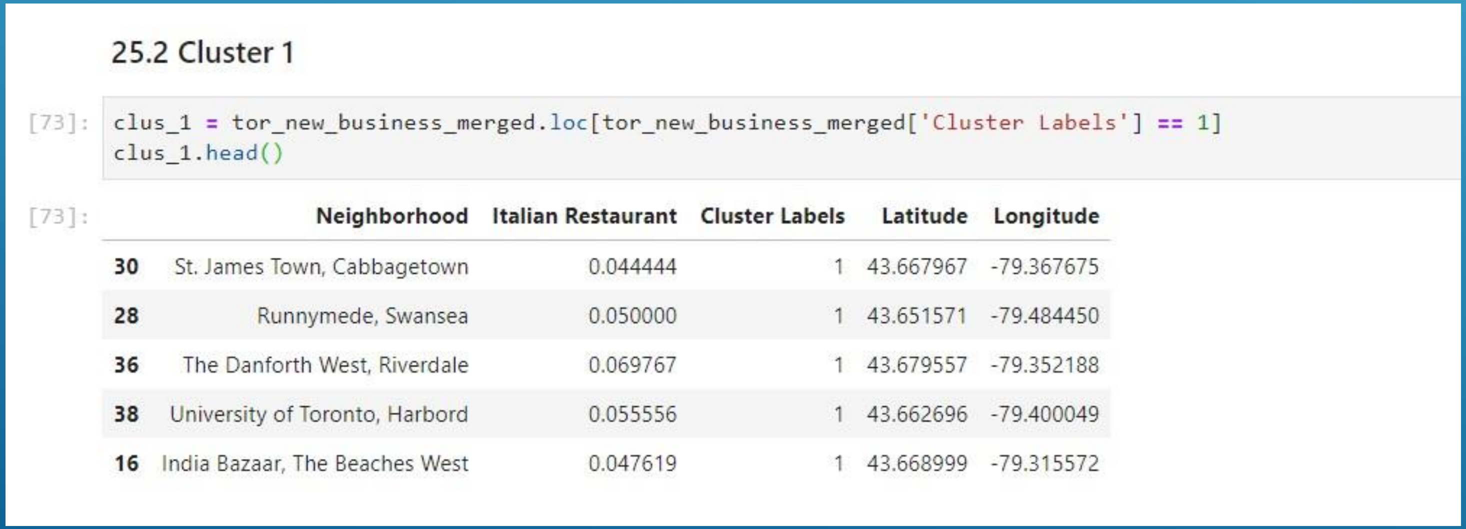
25. Examine Clusters.

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



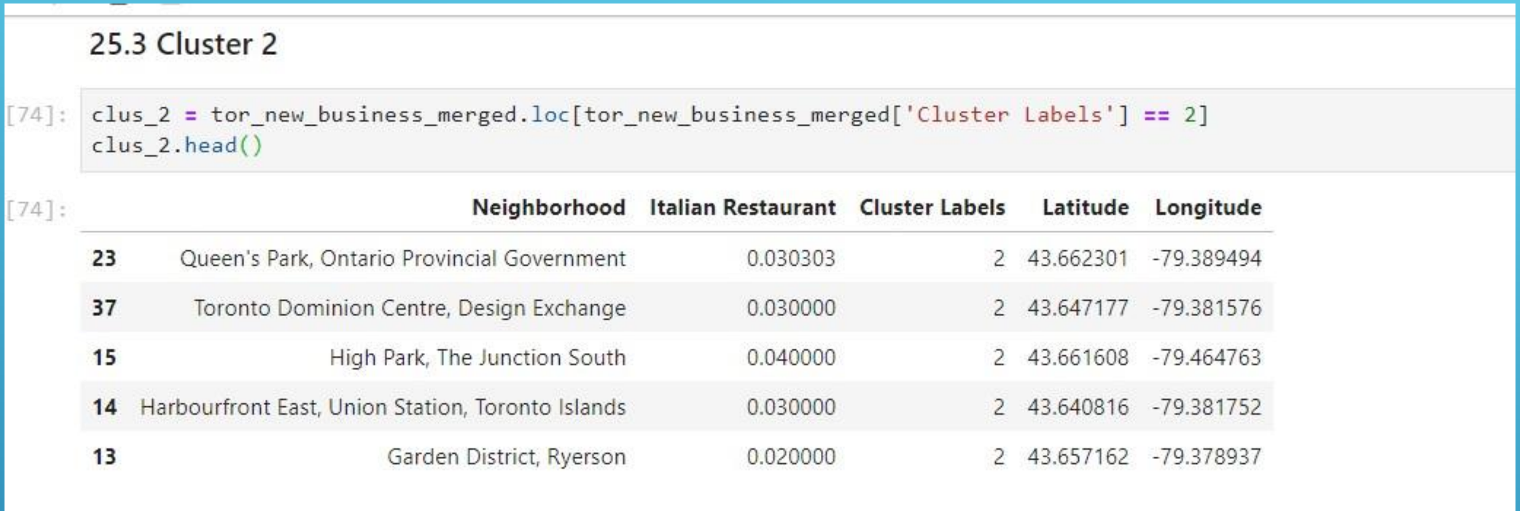
25.2 Cluster 1

```
[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

Results (contd.):



25.3 Cluster 2

[74]:

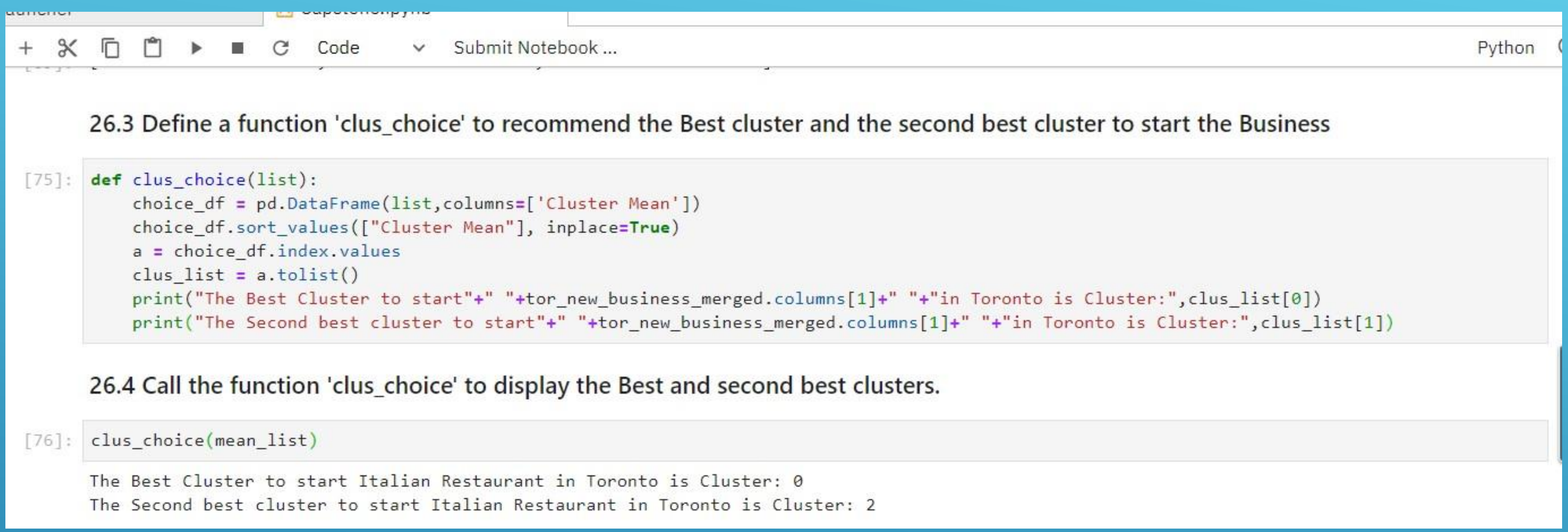
```
clus_2 = tor_new_business_merged.loc[tor_new_business_merged['Cluster Labels'] == 2]
clus_2.head()
```

[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

- Define and call a function 'clus_choice' which will automatically suggest the Best Cluster and the second best cluster to start the new business using the Mean value of the frequency of occurrence of the selected Business in each cluster.

Results (contd.):



26.3 Define a function 'clus_choice' to recommend the Best cluster and the second best cluster to start the Business

```
[75]: def clus_choice(list):
          choice_df = pd.DataFrame(list,columns=['Cluster Mean'])
          choice_df.sort_values(["Cluster Mean"], inplace=True)
          a = choice_df.index.values
          clus_list = a.tolist()
          print("The Best Cluster to start"+" "+tor_new_business_merged.columns[1]+" "+"in Toronto is Cluster:",clus_list[0])
          print("The Second best cluster to start"+" "+tor_new_business_merged.columns[1]+" "+"in Toronto is Cluster:",clus_list[1])
```

26.4 Call the function 'clus_choice' to display the Best and second best clusters.

```
[76]: clus_choice(mean_list)
```

```
The Best Cluster to start Italian Restaurant in Toronto is Cluster: 0
The Second best cluster to start Italian Restaurant in Toronto is Cluster: 2
```

Discussion and Recommendation:

- Neighborhoods in Cluster 0 are recommended as the Best locations to start the selected business, viz, an Italian restaurant, as there are very few Italian restaurants in these areas
  - Early mover advantage
  - Less competition
  - Recommended for an entrepreneur who is a novice in the business
- Neighborhoods in Cluster 2 are recommended as the second best locations to start the selected business, viz, an Italian restaurant, as this cluster has a medium concentration of Italian restaurants.
  - Recommended to experienced and moderately experienced entrepreneurs with some unique selling propositions.
  - Customer base is almost guaranteed.

Discussion and Recommendation (contd.):

- The analysis exactly matches with the predictions of the "clus_choice" function which automatically suggests the Best and second best Clusters to start the business.

Conclusion & Future Improvements:

- The 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.
- **Future Improvements:** Expand 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