DATA SCIENCE PROJECT
WEEK 5 REPORT

INTRODUCTION

This is a capstone project .In this project, I am creating a hypothetical scenario for a concept that there may not be enough Indian Restaurants in Toronto Area. Therefore it might be a great opportunity for an entrepreneur who is based in Canada. As the Indian food is popular among Asian community,so this entrepreneur might think of opening its business in areas where asian community resides.With the purpose in mind, finding the location to open such a restaurant is one of the mostimportant decisions for this entrepreneur and I am designing this project to help him find the most suitable location.

BUSINESS PROBLEM

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Indian Restaurant in Toronto, Canada. By using data science methods and tools along with machine learning algorithms such as clustering, this project aims to provide solutions to answer the business question: In Toronto, if an entrepreneur wants to open an Indian Restaurant, where should they consider opening it?

TARGET AUDIENCE

The entrepreneur who wants to find the location to open an authentic Indian restaurant.

DATA

To solve this problem, we will need below data:

- List of neighborhoods in Toronto, Canada
- Latitude and Longitude of these neighborhoods
- Venue data related to Indian restaurants. This will help us find neighborhoods that are more suitable to open an Indian Restaurant.

EXTRACTING THE DATA

- The scrapping of Toronto neighborhoods via Wikipedia
- Getting Latitude and Longitude data of these neighborhoods via Geocoder package
- Using Foursquare API to get venue data related to these neighborhoods

METHODOLOGY

I got the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from Wikipedia: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

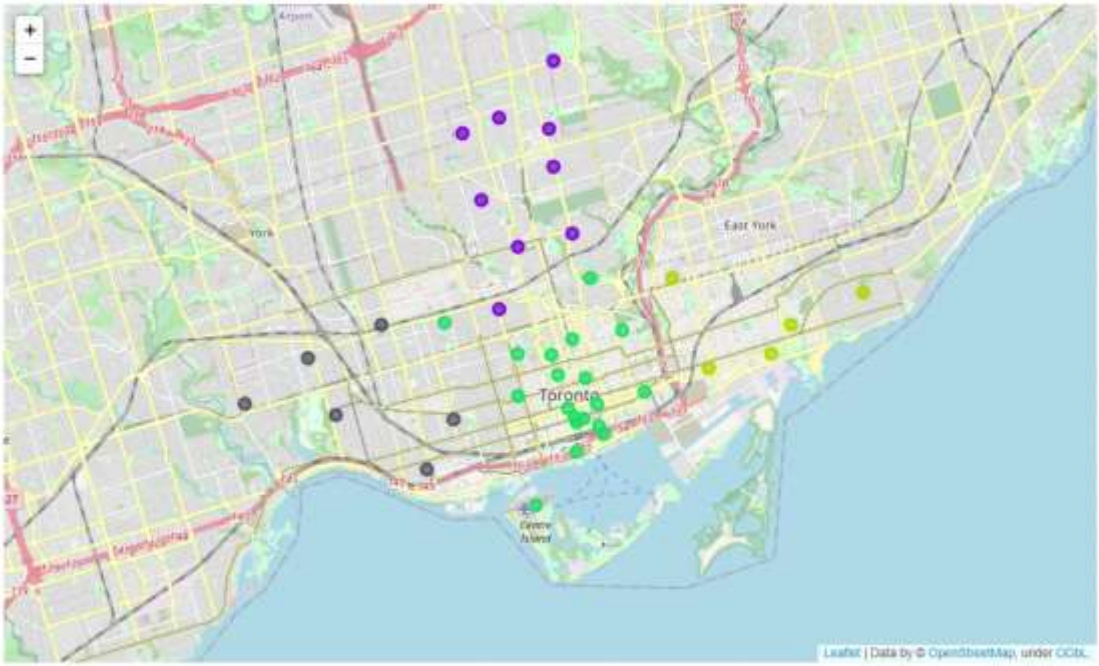
I did the web scraping by utilizing pandas HTML table scraping method as it is easier and more convenient to pull tabular data directly from a web page into the data frame.

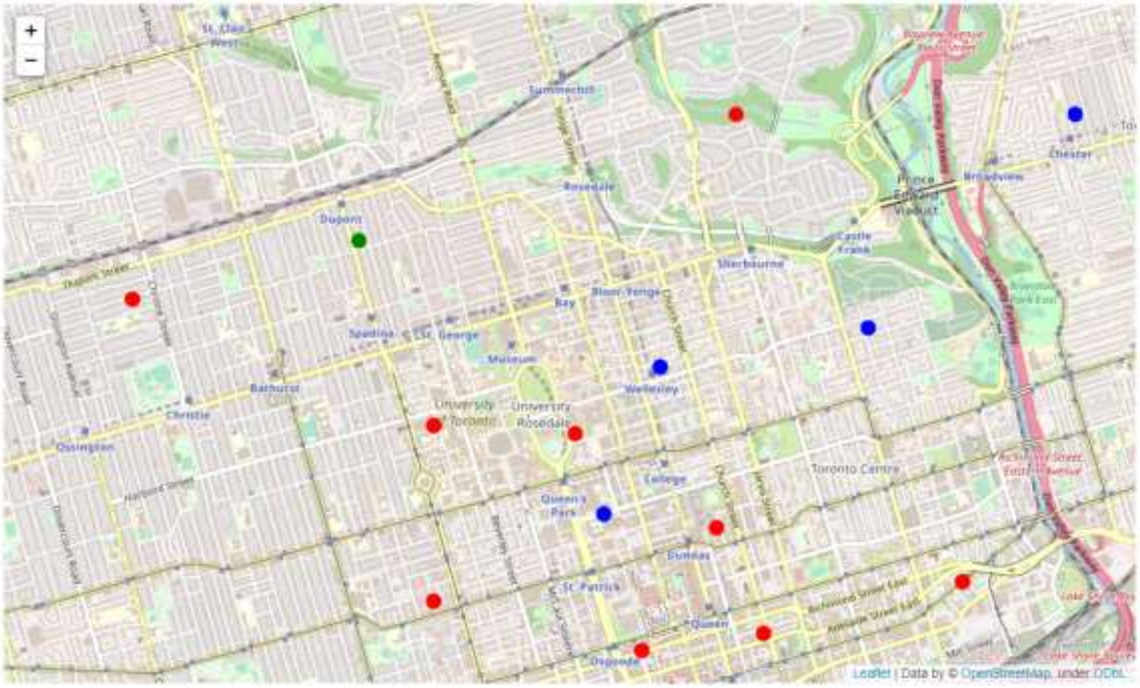
However, it is only a list of neighborhood names and postal codes. I need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I tried using Geocoder Package but it was not working so I used the CSV file provided by IBM team to match the coordinates of Toronto neighborhoods. After gathering these coordinates, I visualize the map of Toronto using Folium package to verify whether these are correct coordinates. Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude, and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

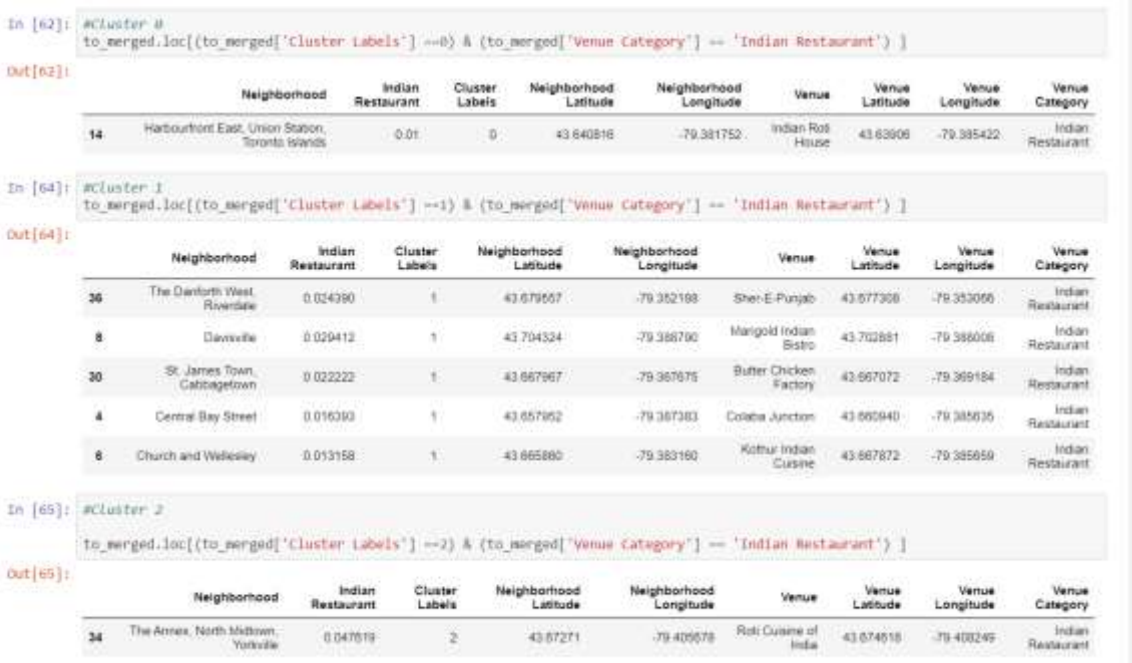
Here, I made a justification to specifically look for “Indian restaurants”. Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. I have clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for “Indian food”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

RESULT

CLUSTERS


+
−
York
East York
Toronto
Leaflet | Data by © OpenStreetMap, under ODbL.


+
−
St. Clair West
Summerhill
Rosedale
Dupont
Prince Edward Viaduct
Broadview
Chester
Castle Frank
Sherbourne
Bloor-Yonge
Bay
Spadina
St. George
Museum
Wellesley
Bathurst
Christie
Ossington
University of Toronto
Queen's Park
College
Toronto Centre
Dundas
St. Patrick
Queen
Osgoode
Leaflet | Data by © OpenStreetMap, under ODbL.



```
In [62]: #Cluster 0
         to_merged.loc[(to_merged['Cluster Labels'] ==0) & (to_merged['Venue Category'] == 'Indian Restaurant') ]
```

Out[62]:

	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
14	Harbourfront East, Union Station, Toronto Islands	0.01	0	43.640816	-79.381752	Indian Roti House	43.63906	-79.385422	Indian Restaurant

```
In [64]: #Cluster 1
         to_merged.loc[(to_merged['Cluster Labels'] ==1) & (to_merged['Venue Category'] == 'Indian Restaurant') ]
```

Out[64]:

	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
36	The Danforth West, Riverdale	0.024390	1	43.679557	-79.352188	Sher-E-Punjab	43.677308	-79.353056	Indian Restaurant
8	Davisville	0.029412	1	43.704324	-79.388790	Marigold Indian Bistro	43.702881	-79.388008	Indian Restaurant
30	St. James Town, Cabbagetown	0.022222	1	43.667967	-79.367675	Butter Chicken Factory	43.667072	-79.369184	Indian Restaurant
4	Central Bay Street	0.016393	1	43.657952	-79.387383	Colaba Junction	43.660940	-79.385635	Indian Restaurant
6	Church and Wellesley	0.013158	1	43.665860	-79.383160	Kothur Indian Cuisine	43.667872	-79.385659	Indian Restaurant

```
In [65]: #Cluster 2

         to_merged.loc[(to_merged['Cluster Labels'] ==2) & (to_merged['Venue Category'] == 'Indian Restaurant') ]
```

Out[65]:

	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
34	The Annex, North Midtown, Yorkville	0.047619	2	43.67271	-79.405678	Roti Cuisine of India	43.674618	-79.408249	Indian Restaurant