WEEK3-IBM CAPSTONE PROJECT-PEER REVIEW ASSIGNMENT

Finding the Best Location to Establish a Location for a Clinic

1. Introduction

For this Project, I am going to explore finding out the best neighbourhood to open a new clinic as a professional Doctor in Toronto area. We will look into a scenario where a doctor after getting his medical licence wants to open his clinic and which neighbourhood will be suitable depending upon the concentration of clinics , demography and other parameters. As the decision to invest in a space and open a clinic is crucial for the doctor we would like to help him get the best venue.

2. Business Problem

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

3. Target Audience

The Doctor who wants to find the neighbourhood/location to open his clinic

4. Data

To solve this problem, I will need below data:

- List of neighborhoods in Toronto, Canada.
- Latitude and Longitude of these neighborhoods.
- Venue data related to clinics. This will help us find the neighborhoods that are most suitable to a clinic

a. **Extracting the data**

i. Scrapping of Toronto neighborhoods via Wikipedia
ii. Getting Latitude and Longitude data of these neighborhoods via Geocoder package
iii. Using Foursquare API to get venue data related to these Neighborhoods

5. **Methodology**

First, I need to get the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from wikipedia page ("https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M")

Data Preparation

```
[1]: #defing URL to scrape
     url_wiki = 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'
     url_wiki

Out[1]: 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'
```

```
[2]: #Scraping the data and tabulating
     import pandas as pd

     pd_page = pd.read_html(url_wiki)

     df_TOR = pd_page[0]
     df_TOR.head()
```

Out[2]:

	Postcode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront

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 dataframe.

However, it is only a list of neighborhood names and postal codes. 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.

To plot the co-ordinates of the neighbourhoods we will have to get the geospatial data of the neighborhoods from http://cocl.us/Geospatial_data.

Out[12]: (103, 3)

Getting Geospatial Data for Neighborhood

In [13]:
```
url_postcode_TOR = 'http://cocl.us/Geospatial_data'
url_postcode_TOR
```

Out[13]: 'http://cocl.us/Geospatial_data'

In [14]:
```
pd_postcode = pd.read_csv(url_postcode_TOR)
df_PC_TOR = pd.DataFrame(pd_postcode)
df_PC_TOR.head()
```

Out[14]:

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

After gathering all these coordinates, The neighbourhood data and location data are merged to form a single dataframe to get the neighbourhood data.

Merging the 2 tables

In [15]:
```
df_merged = df_TOR
df_merged = df_merged.join(df_PC_TOR.set_index('Postal Code'), on='Postcode')
df_merged.head(10)
```

Out[15]:

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M1A	Not assigned	Not assigned	NaN	NaN
1	M2A	Not assigned	Not assigned	NaN	NaN
2	M3A	North York	Parkwoods	43.753259	-79.329656
3	M4A	North York	Victoria Village	43.725882	-79.315572
4	M5A	Downtown Toronto	Harbourfront, Regent Park	43.654260	-79.360636
5	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763
6	M7A	Queen's Park	Queen's Park	43.662301	-79.389494
7	M8A	Not assigned	Not assigned	NaN	NaN
8	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
9	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353

Here, I made a justification to specifically look for “Clinics”. Previously, when I ran the model, I was looking for “Medical Centres” but there are very few results (maybe due to Foursquare categorization) so I looked for the clinics in the Toronto area.

Next, I used 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 was 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 analysed each neighbourhood by grouping the rows by neighbourhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

Initializing Foursquare request type for Clincs around Toronto

In [21]:
```
!conda install -c conda-forge geopy --yes
from geopy.geocoders import Nominatim # module to convert an address into latitude and longitude values
```

```
Solving environment: done

# All requested packages already installed.
```

In [22]:
```
address = 'Toronto'

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

```
43.653963 -79.387207
```

In [23]:
```
search_query = 'Clinic'
radius = 500
print(search_query + ' .... OK!')
```

```
Clinic .... OK!
```

In [24]:
```
CLIENT_ID = 'YQFF03YT4BOY32U4XIVELI3PUXTABGYD2W1AFDBHLWQFU1TV' # your Foursquare ID
CLIENT_SECRET = 'LLX3X4P1CXMWFTA4DLJJXZSY5U1MQIARDGLL3IECBNR4ZOTL' # your Foursquare Secret
VERSION = '20190804'
LIMIT = 30
print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

```
Your credentails:
CLIENT_ID: YQFF03YT4BOY32U4XIVELI3PUXTABGYD2W1AFDBHLWQFU1TV
CLIENT_SECRET:LLX3X4P1CXMWFTA4DLJJXZSY5U1MQIARDGLL3IECBNR4ZOTL
```

In [25]:
```
url = 'https://api.foursquare.com/v2/venues/search?client_id={}&client_secret={}&ll={},{}&v={}&query={}&radius={}&limit=
url
```

Out[25]:
```
'https://api.foursquare.com/v2/venues/search?client_id=YQFF03YT4BOY32U4XIVELI3PUXTABGYD2W1AFDBHLWQFU1TV&client_secret
```

The result data set was normalized into a dataframe and readied for analysing.

Transform the JSON into Pandas Dataframe

In [30]:
```
# assign relevant part of JSON to venues
#venues = results1['response']['venues']

# tranform venues into a dataframe
#dataframe = json_normalize(venues)
#dataframe.head()

# assign relevant part of JSON to venues
venues = results['response']['venues']

# tranform venues into a dataframe
dataframe = json_normalize(venues)
dataframe.head()
```

Out[30]:

	categories	hasPerk	id	location.address	location.cc	location.city	location.country	location.crossStreet
0	[]	False	4f6334e0e4b04d4b145f5b83	NaN	CA	NaN	Canada	NaN
1	[{'id': '4bf58dd8d48988d177941735', 'name': 'D...	False	4de68e89e4cde71744c308da	123 Edward Street	CA	Toronto	Canada	NaN
2	[{'id': '4bf58dd8d48988d196941735', 'name': 'H...	False	507ff5a8e4b03cfb613822e9	The Hospital for Sick Children (SickKids)	CA	Toronto	Canada	NaN
3	[{'id': '4bf58dd8d48988d104941735', 'name': 'M...	False	4b8e90ecf964a520122933e3	595 Bay St.	CA	Toronto	Canada	in Atrium on Bay
4	[{'id': '4bf58dd8d48988d177941735', 'name': 'D...	False	51c0df108bbd5e5f422621d1	101-133 Hazelton Avenue	CA	Toronto	Canada	NaN

	location.distance	location.formattedAddress	location.labeledLatLngs	location.lat	location.lng	location.postalCode	location.state	name	ref
0	421	[Canada]	[{'label': 'display', 'lat': 43.65746938169752...	43.657469	-79.389178	NaN	NaN	Transplant Clinic	15650
1	219	[123 Edward Street, Toronto ON M5G 1E2, Canada]	[{'label': 'display', 'lat': 43.65589356806434...	43.655894	-79.386638	M5G 1E2	ON	Rudd-PES Endoscopy Clinic	15650
2	378	[The Hospital for Sick Children (SickKids), To...	[{'label': 'display', 'lat': 43.65734334277486...	43.657343	-79.387732	NaN	ON	Clinic 9	15650
3	387	[595 Bay St. (in Atrium on Bay), Toronto ON M5...	[{'label': 'display', 'lat': 43.65613745484842...	43.656137	-79.383454	M5G 2C2	ON	MCI Medical Clinic	15650
4	484	[101-133 Hazelton Avenue, Toronto ON M5R 0A6, ...	[{'label': 'display', 'lat': 43.65072566250946...	43.650726	-79.391225	M5R 0A6	ON	Visage Clinic	15650

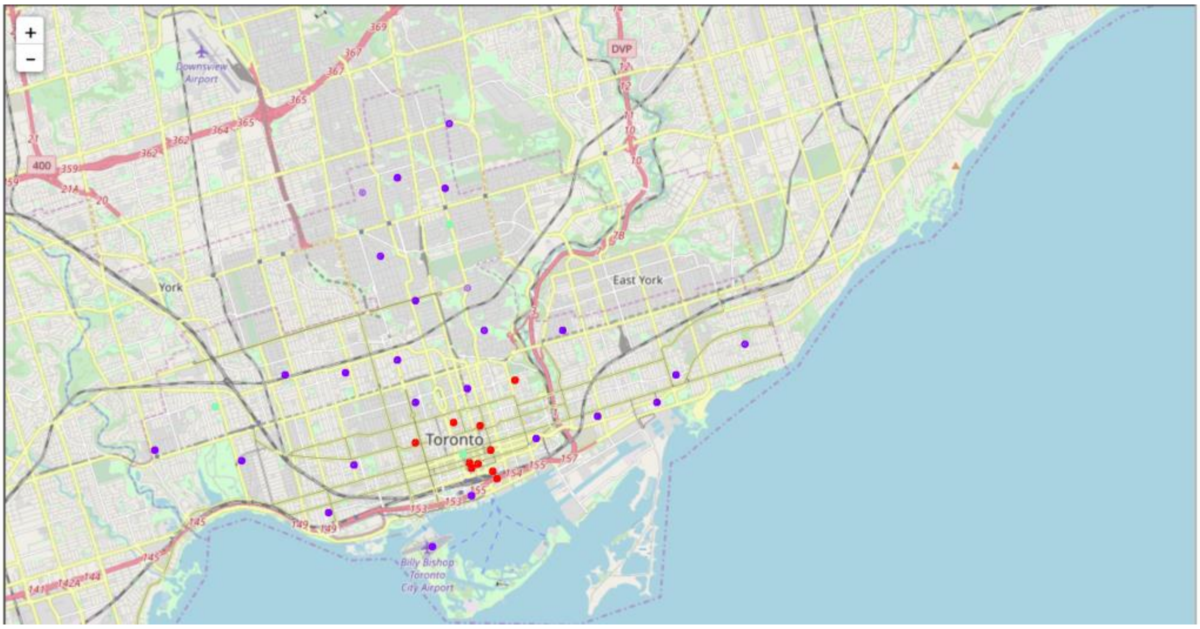
The raw dataframe is then filtered down and merged with neighbourhood data fro clinic clustering.

	name	lat	lng	postalCode
1	Rudd-PES Endoscopy Clinic	43.655894	-79.386638	M5G 1E2
3	MCI Medical Clinic	43.656137	-79.383454	M5G 2C2
4	Visage Clinic	43.650726	-79.391225	M5R 0A6
6	Dundas University Health Clinic	43.654196	-79.388166	M4P 2K8
7	Dundas West Chiropractic Clinic	43.654866	-79.387836	M6R 3A9
8	The Voice Clinic	43.655368	-79.386429	M7A 0A1
11	Cystoscopy Clinic	43.658806	-79.389568	M5G 2N2
15	Gastrointestinal Clinic	43.658706	-79.388775	M5G 0A3
17	The Mindfulness Clinic	43.652069	-79.382722	M5G 1Z6
18	Toronto Foot Clinic	43.653187	-79.382181	M5G 2A3
24	Grow Legally Marijuana Clinic and Consulting	43.656043	-79.381403	M5G 1Z3
26	Tuina Health Clinic	43.655100	-79.380500	M5C 2L7

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 neighbourhoods in Toronto into 3 clusters based on their frequency of occurrence for “Clinic”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the clinic.

6.Results

Clusters



```
In [61]: # import k-means from clustering stage
         from sklearn.cluster import KMeans

         # run k-means clustering
         kmeans = KMeans(n_clusters = 5, random_state = 0).fit(clinic_onehot)
```

```
In [62]: means_df = pd.DataFrame(kmeans.cluster_centers_)
         means_df.columns = clinic_onehot.columns
         means_df.index = ['G1','G2','G3','G4','G5']
         means_df['Total Sum'] = means_df.sum(axis = 1)
         means_df.sort_values(axis = 0, by = ['Total Sum'], ascending=False)
```

Out[62]:

	Rudd-PES Endoscopy Clinic	MCI Medical Clinic	Visage Clinic	Dundas University Health Clinic	Dundas West Chiropractic Clinic	The Voice Clinic	Cystoscop Clini
G3	1.0	1.0	0.0	0.0	0.0	0.0	1.
G1	0.0	0.0	1.0	0.0	0.0	0.0	0.
G2	0.0	0.0	0.0	1.0	0.0	0.0	0.
G4	0.0	0.0	0.0	0.0	1.0	0.0	0.
G5	0.0	0.0	0.0	0.0	0.0	0.0	0.

```
In [63]: neigh_summary = pd.DataFrame([means_df.index, 1 + kmeans.labels_]).T
         neigh_summary.columns = ['Neighbourhood', 'Group']
         neigh_summary
```

Out[63]:

	Neighbourhood	Group
0	G1	3
1	G2	2
2	G3	4
3	G4	5
4	G5	1

Best Neighborhood to open a clinic

The results from k-means clustering show that we can categorize Toronto neighbourhoods into 5 clusters based on how many clinics are in each neighbourhood:

- Cluster G1: Has Medium Number of Clinics
- Cluster G2: Has Medium Number of Clinics
- Cluster G3: Neighbourhoods have high number of clinics
- Cluster G4: Neighbourhoods have low number of clinics
- Cluster G5: Neighbourhoods have medium number of clinics

7.Recommendations

Most of Clinics are in G3 which is around Central Bay District areas and lowest (close to zero) in G4 areas which is around St James Town and Parkdale areas.

8.Limitations and Suggestions for Future Research

In this project, I only take into consideration of one factor: the occurrence / existence of Clinics in each neighbourhood. There are many factors that can be taken into consideration such as population density, income of residents, rent that could influence the decision to open a new clinic. However, to put all these data into this project is not possible to do within a short time frame for this capstone project. Future research can take into consideration of these factors. In addition, I am relying on the existence of clinics only for this project but future research can take into consideration of other variables such as existence of Hospitals as per population level in each neighbourhood etc.

9.Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing the machine learning by utilizing k-means clustering and providing recommendation to the stakeholder.

10.References

List of neighbourhoods in Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Foursquare Developer Documentation: https://developer.foursquare.com/docs