Coursera Capstone Project

Battle of the Neighbourhoods

1. Introduction/Business Problem

If you are looking to open a Chain of Cafes in Canada and sell coffee as an entrepreneur, you will be facing tough competition in a potentially saturated market. Therefore a key problem to solve is:

Where is the best place to open up your multiple Cafes?

Solving this problem gives us information and prevents opening a Cafe in a non-saturated area may provide a new business the best opportunity to thrive without competition. This is a very difficult question to solve and another issue arises:

How do we measure/determine which is the best place to start a Cafe?

For a given n number of chains you wish to open, how do you distribute them such that you can get the greatest coverage and or exposure in a city with an established market. Traditionally one would think the best way to approach this is to maybe find places of high traffic of people. This data however is extremely difficult to come by. However, an important piece of information available is the location of current Venues. This tells us many things:

- Which places or locations already have a steady population of customers that drink coffee. This is inferred based on density of venues in a given location.
- Customer base and or demographics of customers. One would think that places with lots of Cafes would be a business district or a place of high traffic for sales.

Data science and exploration of location data will be key to finding the solution.

2. Data

To solve the problem, we will need the following data:

- List of neighbourhoods in Canada. This defines the scope of the project to the city of Sydney. This data can be sourced from web scraping Wikipedia pages and or some location data service.
- Latitude and longitude coordinates of these neighbourhoods for visualization, clustering and other purposes. This data can be sourced from location data services.
- Venue data for finding saturated neighbourhoods of Cafes. This data can be sourced from FourSquare's API.

The culmination of this data will provide insights into saturation of venues in specific neighbourhoods and potential vacancies in locations.

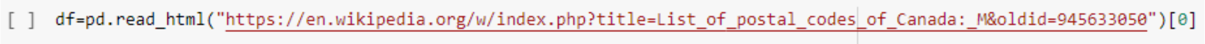
What we expect to find are:

- Metrics of where cluster centroids are given n number of shops to open in Toronto such that;
- We can determine the competition each of these stores will be facing to then allocate resources.

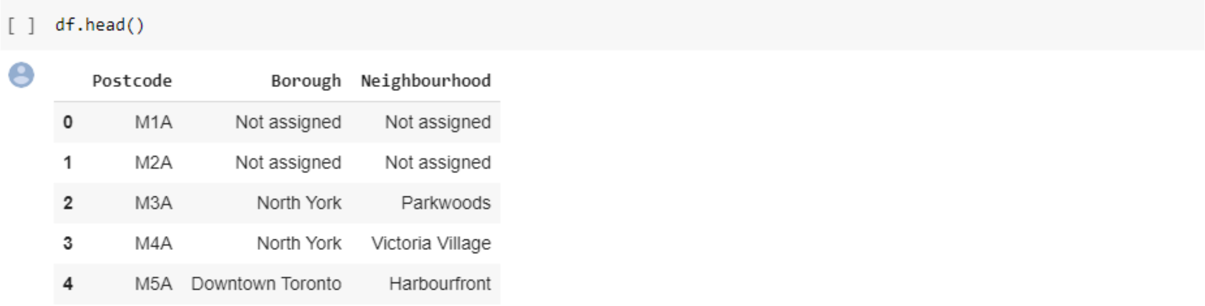
3. Methodology

The begin, we require data of a list of neighbourhoods in Canada and their respective latitude and longitudes such that we can query the FourSquare API to receive venue data.

We source our neighbourhood data by scraping Wikipedia using pandas.read_html function:



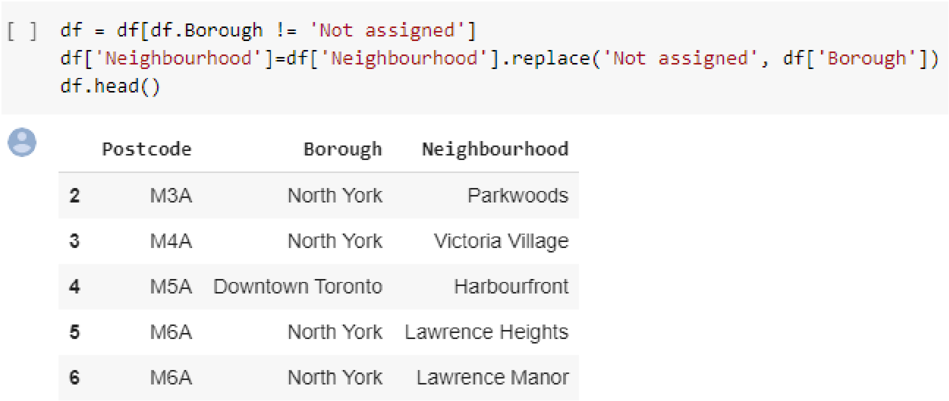
```
[ ]  df=pd.read_html("https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=945633050")[0]
```



```
[ ]  df.head()
```

	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

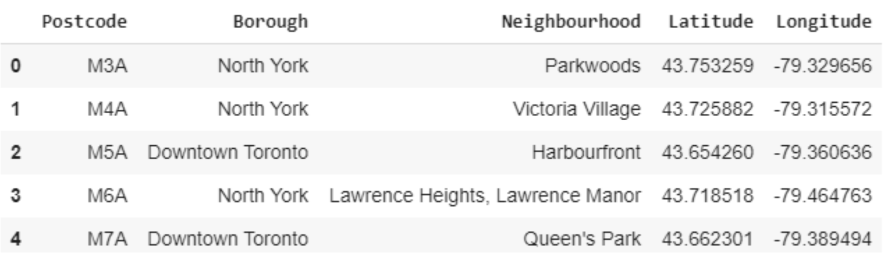
We then clean the data and remove any entries with Boroughs unassigned. If neighbourhoods are not assigned, we assume that they are the Boroughs:



```
[ ]  df = df[df.Borough != 'Not assigned']
     df['Neighbourhood']=df['Neighbourhood'].replace('Not assigned', df['Borough'])
     df.head()
```

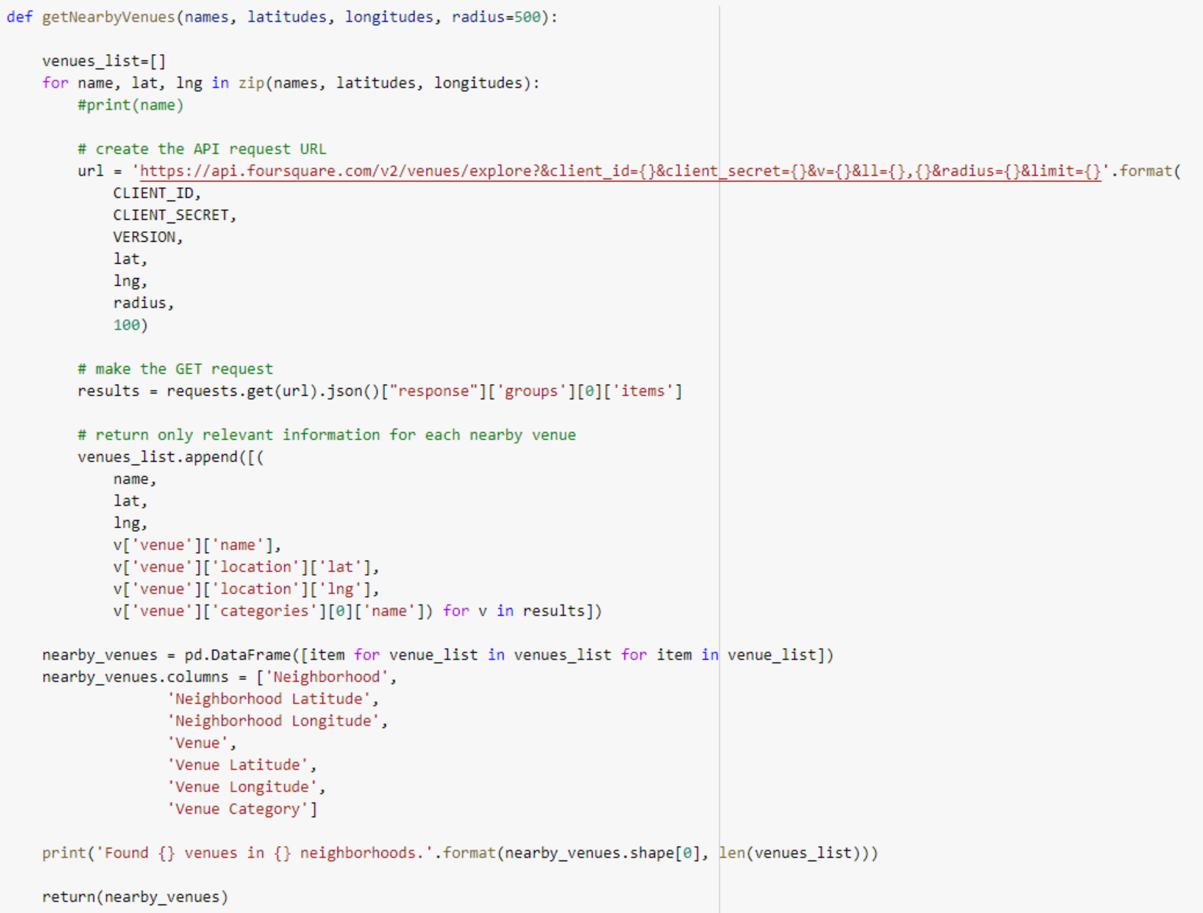
	Postcode	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront
5	M6A	North York	Lawrence Heights
6	M6A	North York	Lawrence Manor

For geospatial, we are provided that through the course and we download it, available from http://cocl.us/Geospatial_data. After some cleaning, we can merge the neighbourhood and geospatial data to make our dataset for querying venues in this structure:



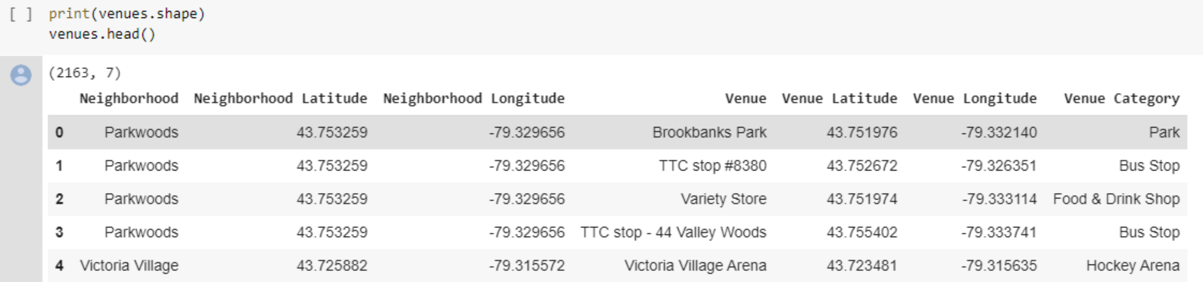
	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494

For our venue data, we create a function that takes our FourSquare credentials and the neighbourhood latitude and longitude data created:



```
def getNearbyVenues(names, latitudes, longitudes, radius=500):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        #print(name)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            100)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    print('Found {} venues in {} neighborhoods.'.format(nearby_venues.shape[0], len(venues_list)))

    return(nearby_venues)
```

Once called, we find that there are over 2163 venues in 103 neighbourhoods. The structure of this dataset is as below:



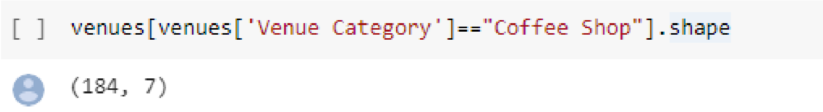
```
[ ] print(venues.shape)
    venues.head()
```

```
(2163, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	TTC stop #8380	43.752672	-79.326351	Bus Stop
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Parkwoods	43.753259	-79.329656	TTC stop - 44 Valley Woods	43.755402	-79.333741	Bus Stop
4	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena

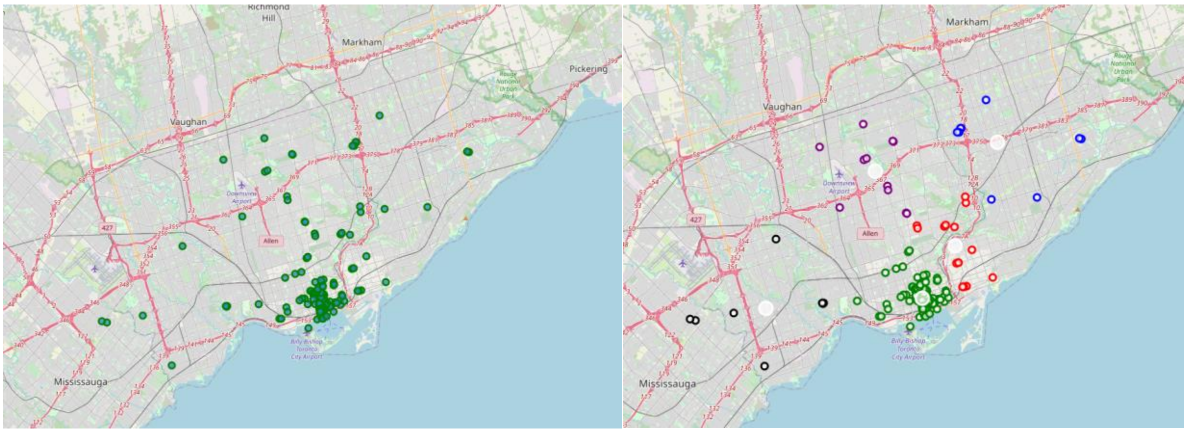
Importantly, we wish to use the fields “Venue”, “Venue Latitude”, “Venue Longitude” and “Venue Category”. We will be clustering based on the latitude and longitudes and filtering by category to determine coffee shops.

Once filtered we find that there are 184 coffee shops within Toronto:



```
[ ]  venues[venues['Venue Category']=="Coffee Shop"].shape

     (184, 7)
```

We use folium to visualise these on the map:

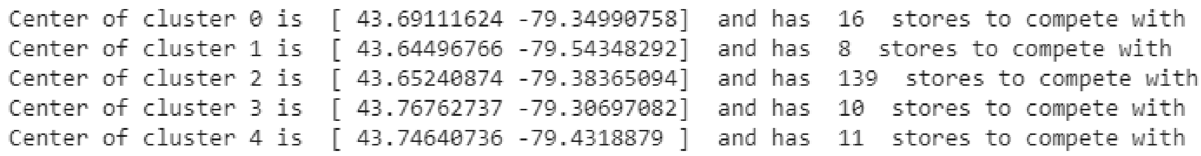


The distribution is heavily within the city of Toronto. For our study, we specify that we are interested in opening 5 coffee shops. Therefore our K is equal to 5 for our KMeans clustering method. Once we apply this clustering method, we are given the cluster centres as well as how many shops will each of these stores be competing with within the locale.

The white larger circles, denote the cluster centre and their respective colours surrounding them are the shops they will be competing with.

4. Results

The following metrics are available for each cluster:



```
Center of cluster 0 is  [ 43.69111624 -79.34990758]  and has  16  stores to compete with
Center of cluster 1 is  [ 43.64496766 -79.54348292]  and has  8  stores to compete with
Center of cluster 2 is  [ 43.65240874 -79.38365094]  and has  139  stores to compete with
Center of cluster 3 is  [ 43.76762737 -79.30697082]  and has  10  stores to compete with
Center of cluster 4 is  [ 43.74640736 -79.4318879 ]  and has  11  stores to compete with
```

It is particularly important to note that cluster 2 which is located in Toronto has a staggering 139 other stores to compete with, while cluster 1 contains only 8 other stores to compete with.

5. Discussion

It is important to note that these are purely metrics that aid in the decision making of where in the city to open a Café and what competition exists there. The centre of the clusters denote just that and not where to actually open up the store as the sparsity in less dense clusters heavily effect the centroid of KMeans clustering.

There is also business strategy that may be required of whether or not it actually is a good idea to open in a less or more dense area. Some business follow the strategy of getting exposure off competition whilst others find new oceans to open up businesses where competition is low to gain a monopoly.

Unless your product is amazing, I recommend not to open a store in Toronto (green clusters) and instead open a Café chain in the red clustered location. Key landmarks may also play a role in this as previously mentioned as increased foot traffic may lead to success.

6. Conclusion

To conclude, Toronto is a highly competitive location to open up a coffee store and would be strongly advised to open on the outskirts of Toronto. Sparsity of shops is another key consideration within these clusters as it can be seen that within clusters, there are minor clusters that may be due to landmarks within the city.