Data Science Capstone Project

Introduction

Problem Description:

Chennai, one of the metropolitan cities in India is a multi-cultural city. People from across the world come and live here. Chennai is also famous for its choice of food. One of my friends, Robinson, is relocating to India and is interested in setting up an Italian Restaurant in Chennai. He has been a Pizza lover since his childhood. He doesn't want to setup a pizza alone store. He is looking at locations where presence of Italian flavor is minimal. He has asked me to leverage my data science skills and come up with ideal locations for his new venture i.e. Italian Restaurant.

While many prefer south Indian as default choice, Robinson wants to invest on his passion. To start with, I gathered the list of neighborhoods in Chennai.

Data Description:

- I gathered the necessary neighborhood details of Chennai from Wikipedia.
- I used BeautifulSoup library in Python to scrape the Wikipedia page and obtain the list of neighborhoods.
- Then, I used the google maps API and using geolocator, fetched the latitude and longitude of all the neighborhoods.
- Used Four Square API to fetch venue details

Methodology:

Using Google API, fetched the coordinates for neighborhoods scraped from Wiki.

For each neighborhood in the list, get the latitude and longitude and add to a dataframe. If the latitude or longitude values are NaN, delete them.

In [7]:
```
address_geocode_list=[]
for address in neighborhood:
    latitude, longitude = get_lat_long(address)
    address_geocode_list.append([address,latitude,longitude])
#address_geocode_list
df_geo_list = pd.DataFrame(address_geocode_list,columns=['Neighborhood','Latitude','Longitude'])
df_geo_list = df_geo_list.dropna()
df_geo_list.head()
```

Out[7]:

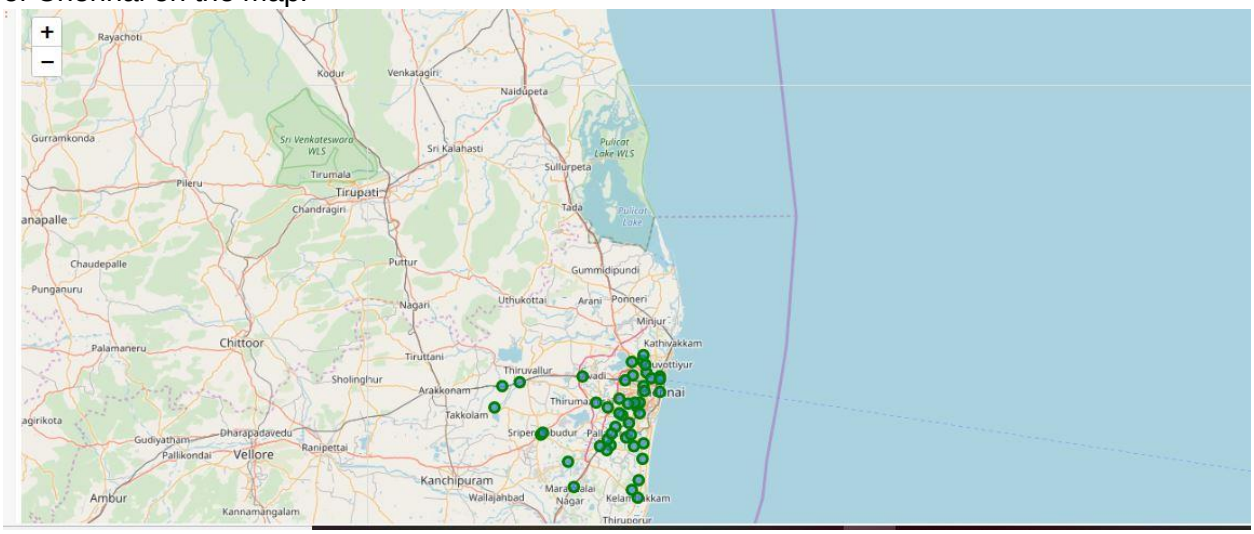
	Neighborhood	Latitude	Longitude
0	Alandur	12.994373	80.194284
1	Anna Nagar	11.170349	77.351114
2	Ashok Nagar, Chennai	13.040073	80.215925
3	Assisi Nagar	13.164610	80.233000
4	Ayanavaram	13.094616	80.235410

Get latitude and longitude of Chennai

In [8]:
```
latitude, longitude = get_lat_long('Chennai')
print ('Latitude and Longitude of Chennai is {},{}'.format(latitude,longitude))
```

```
Latitude and Longitude of Chennai is 13.0801721,80.2838331
```

Using folium library, created a map of Chennai and superimposed all the neighborhoods of Chennai on the map.



I also leverage Four Square API to get all the venues near the given coordinates. Fetched the venue details of all the neighborhoods in Chennai. The criteria I used was to get all the nearby venues within a radius of 1000m and number of returned values restricted to 100.

I also segregated data as **Is Restaurant** and **Is Italian**. After segregating, the total Restaurant count returned by four square API is 161 and among them 28 are Italian which included Pizza Stores also.

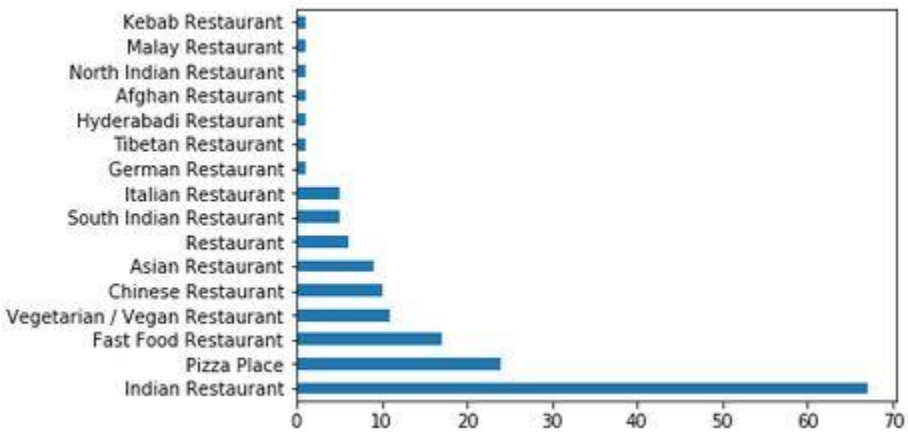
	Neighborhood	Latitude	Longitude	Venue_Name	Venue_Latitude	Venue_Longitude	Venue_Category	Is_Resturant	Is_Italian
0	Alandur	12.994373	80.194284	Pizza Republic	12.990987	80.198613	Pizza Place	True	True
1	Alandur	12.994373	80.194284	Hotel Saravana Bhavan	12.996520	80.190224	South Indian Restaurant	True	False
2	Alandur	12.994373	80.194284	Sukkkubai Beef Biryani Shop	12.998769	80.201381	Indian Restaurant	True	False
3	Alandur	12.994373	80.194284	The Great Kebab Factory	12.994200	80.187495	Indian Restaurant	True	False
4	Alandur	12.994373	80.194284	Aasife & Brothers Biryani Centre	13.000457	80.200635	Indian Restaurant	True	False

```
print('Total Restaurants in Chennai under given conditions:',df_venue.Is_Resturant.count())
print ('Total Italian Restaurants in Chennai under given conditions:', df_venue[df_venue.Is_Italian==True].Is_Italian.count())
```

```
Total Restaurants in Chennai under given conditions: 161
Total Italian Restaurants in Chennai under given conditions: 28
```

```
df_venue['Venue_Category'].value_counts().plot(kind='barh')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd9cd1563c8>
```



From the above graph, it can be seen that type Indian Restaurant have maximum presence across Chennai followed by Pizza stores.

I then applied one hot encoding to see the details of each venue category. Filtered the data to top three venues and Italian looking at the above graph. Then, took the mean of each category and came up with a table that shows top venue category for neighborhood.

	Neighborhood	Latitude	Longitude	Venue_Name	Venue_Latitude	Venue_Longitude	Venue_Category	Is_Resturant	Is_Italian	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Alandur	12.994373	80.194284	Pizza Republic	12.990987	80.198613	Pizza Place	True	True	2	Indian Restaurant	South Indian Restaurant	Chinese Restaurant	Italian Restaurant
1	Alandur	12.994373	80.194284	Hotel Saravana Bhavan	12.996520	80.190224	South Indian Restaurant	True	False	2	Indian Restaurant	South Indian Restaurant	Chinese Restaurant	Italian Restaurant
2	Alandur	12.994373	80.194284	Sukkkubai Beef Biryani Shop	12.998769	80.201381	Indian Restaurant	True	False	2	Indian Restaurant	South Indian Restaurant	Chinese Restaurant	Italian Restaurant
3	Alandur	12.994373	80.194284	The Great Kebab Factory	12.994200	80.187495	Indian Restaurant	True	False	2	Indian Restaurant	South Indian Restaurant	Chinese Restaurant	Italian Restaurant
4	Alandur	12.994373	80.194284	Aasife & Brothers Biryani Centre	13.000457	80.200635	Indian Restaurant	True	False	2	Indian Restaurant	South Indian Restaurant	Chinese Restaurant	Italian Restaurant

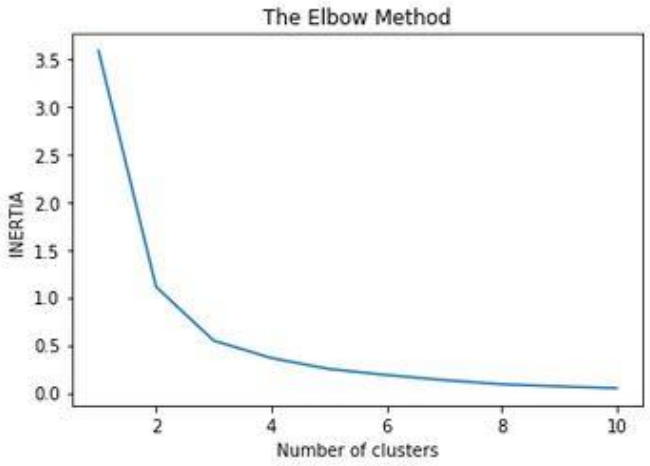
```
# create map
```

I have saved all my work in github repository

Results:

To segregate the venues based on their characteristics, I have used k means clustering algorithm. K- Means is one of the best unsupervised algorithm in classifying the data. As a first step, found the K value using the elbow joint method. Value obtained is 3.

```
In [151]: chennai_cluster = neighborhood_group.drop('Neighborhood', 1)
          inertia=[]
          for i in range(1, 11):
              km = KMeans(n_clusters=i)
              km = km.fit(chennai_cluster)
              inertia.append(km.inertia_)
          plt.plot(range(1, 11), inertia)
          plt.title('The Elbow Method')
          plt.xlabel('Number of clusters')
          plt.ylabel('INERTIA')
          plt.show()
```

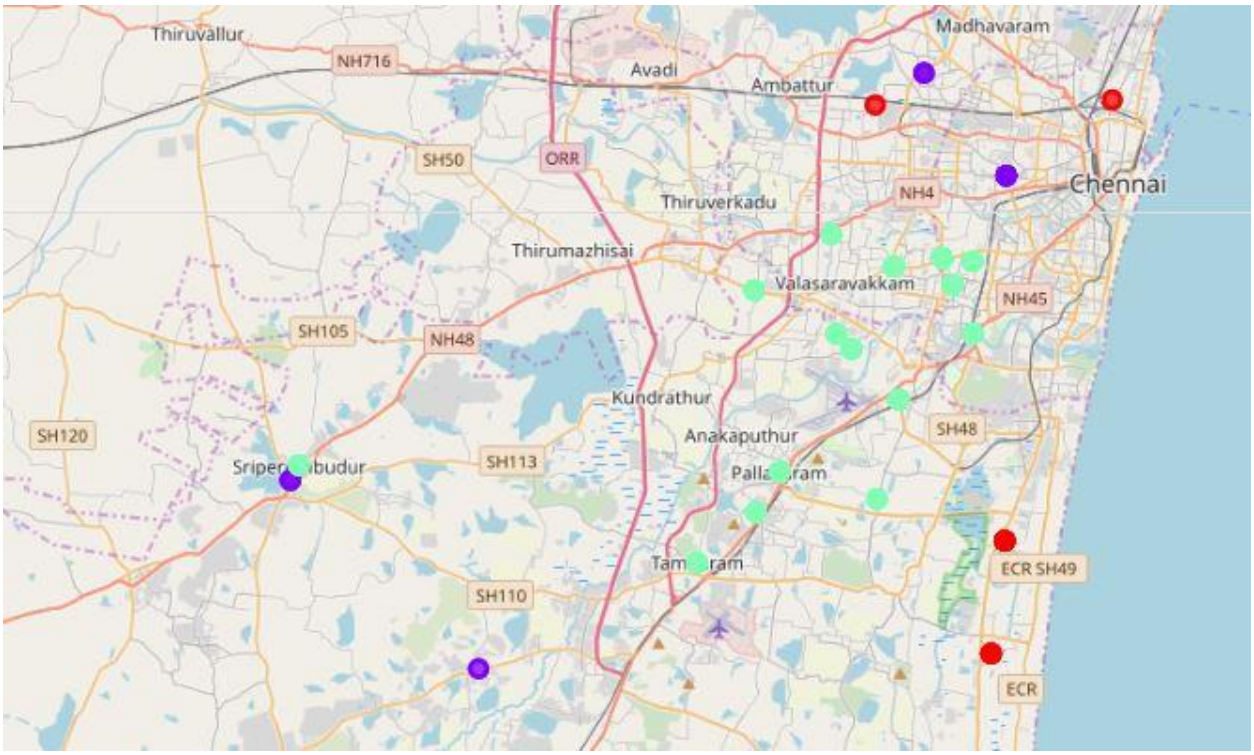


Using the K value=3, ran the k means algorithm on the data and got the classified labels.

```
In [152]: kclusters=3
          # run k-means clustering
          kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(chennai_cluster)
          kmeans.labels_

Out[152]: array([2, 2, 1, 2, 2, 2, 1, 1, 2, 1, 0, 2, 0, 2, 2, 2, 1, 2, 0, 1, 2, 0,
                 1, 2, 2, 2, 0, 2, 2, 0], dtype=int32)
```

Using folium, mapped the clusters on Chennai map to analyze how close or far the venues are clustered.



Conclusion:

From the analysis of cluster data, the following can be interpreted:

1. Across all clusters, Indian Restaurant is the top venue.
2. Among the 3 clusters, Cluster 0 has presence of Italian Restaurants which includes Pizza stores also.
3. Cluster 1 and 2 has minimal presence of Italian Restaurant

In a multi-cultural city like Chennai, opening an exclusive Italian Restaurant is a viable option. The city is over flooded with Indian and Chinese Restaurants. My friend indeed made a good choice in opening an Italian Restaurant.

References:

- Wikipedia - https://en.wikipedia.org/wiki/Category:Suburbs_of_Chennai
- Four Square API - https://developer.foursquare.com/docs/resources/categories