IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

APPLIED DATA SCIENCE CAPSTONE – WEEK 5 REPORT

About the City:

Coimbatore is known as “The Manchester Of South India” because of its textile production volume. It has got the best weather to live in, various nature spots where people go on a road trip, and many religious sites in addition. In recent years, it has got the best name for the number of colleges and hospitals present. Many students from across India stay at Coimbatore for their studies thereby increasing the population of the youth in the city. Hence, there has been a great increase in the demand for restaurants, cafes, etc.

Target Audience:

As mentioned earlier, there has been a greater number of restaurant openings in the city to meet the demands. The target audience for this project are those who are looking for the perfect location to open their restaurants, cafes, etc. It is a common point of view that the localities with a greater number of restaurants present currently, have the greatest demand thereby earning high profits.

Question to be answered:

Which is the best location to open a restaurant (food place) at Coimbatore?

Dataset:

1) Coimbatore Data:

The dataset that will be used for this project is from,

https://news.abplive.com/pincode/tamil-nadu/coimbatore.html

This webpage consists of a table that includes the name of the Post Offices (Similar to neighborhoods), Taluks (Similar to Boroughs), District, State, and Postal Code.

2) Geospace Data:

The table present in the link mentioned above does not contain the latitudes and longitudes of the Post Offices. To fetch the coordinates, geospace data will be employed.

3) Foursquare Data:

As mentioned earlier, localities with a large number of restaurants indicate the high demand. Hence, to find the restaurants, cafes, and any kind of food places near the locality, four square data will be used.

Web Scrapping and Pre-processing:

The dataset is scrapped from the link mentioned in the above section using the requests library. Following that, the text scrapped from the link is then parsed using Beautiful Soup thereby creating a beautiful soup object. By employing the read_html method, the table from the soup object is converted into a pandas dataframe. The first five rows in the datafram are shown in figure 1.

	Office	Taluk	District	State	Pincode
0	15 Velampalayam	Tiruppur	Coimbatore	TAMIL NADU	641652
1	63 Velampalayam	Palladam	Coimbatore	TAMIL NADU	641663
2	A Nagore	Udamalpet	Coimbatore	TAMIL NADU	642205
3	Achipatti	Pollachi	Coimbatore	TAMIL NADU	642002
4	Agrahara Kannadiputhur	Udumalaipettai	Coimbatore	TAMIL NADU	642111

Fig -1: First five rows in the dataframe

There are 587 unique post offices and 20 unique taluks. In the dataset, the office column is similar to neighborhood and Taluks are similar to boroughs. The state and district columns are not necessary and hence, they are removed from the dataset. The Office column is renamed as 'PostOffice'. The objective of the project is to find a suitable location to start any food places. Therefore, the main part of the city is concentrated which includes Coimbatore, Coimbatore North and Coimbatore South Taluks. The resulting dataframe with 183 instances is shown in figure 2.

	PostOffice	Taluk	Pincode
6	Agraharasamakulam	Coimbatore North	641110
13	Alandurai	Coimbatore North	641101
22	Anaikatti	Coimbatore North	641108
38	Athipalayam	Coimbatore North	641110
45	Bharathiyar University	Coimbatore North	641046

Fig -2: Processed Dataframe

Fetching the Coordinates of PostOffices:

To plot the localities and to fetch the nearby venues, the coordinates of the localities are essential. This is achieved using geospace data. Initially, the coordinates of the first 30 places are fetched to check if the code is executed correctly. Following the process, other coordinates are fetched using the similar

approach. The indices of localities whose coordinates are not fetched are stored in a separate list. Here, all the coordinates are perfectly obtained without any glitches except for index 113. This index is dropped from the dataframe and the resulting dataframe with shape 182 x 5 is obtained.

	PostOffice	Taluk	Pincode	Latitude	Longitude
0	Agraharasamakulam	Coimbatore North	641110	11.0776	76.9253
1	Alandurai	Coimbatore North	641101	10.9472	76.8305
2	Anaikatti	Coimbatore North	641108	11.0821	76.8566
3	Athipalayam	Coimbatore North	641110	11.0776	76.9253
4	Bharathiyar University	Coimbatore North	641046	11.039	76.8764
...	...	...	...	...	...
177	Kuttagam	Coimbatore	638462	11.0081	76.9795
178	Malumichampatti	Coimbatore	641050	10.9196	76.9985
179	Merkupathi	Coimbatore	638103	11.0081	76.9795
180	Vadavalli	Coimbatore	641041	11.0273	76.9116
181	Vallipuram	Coimbatore	638103	11.0081	76.9795

182 rows × 5 columns

Fig -3: Dataframe incorporating Latitudes and Longitudes

The locations of these Post Offices are marked geographically by employing folium as seen in figure 4.

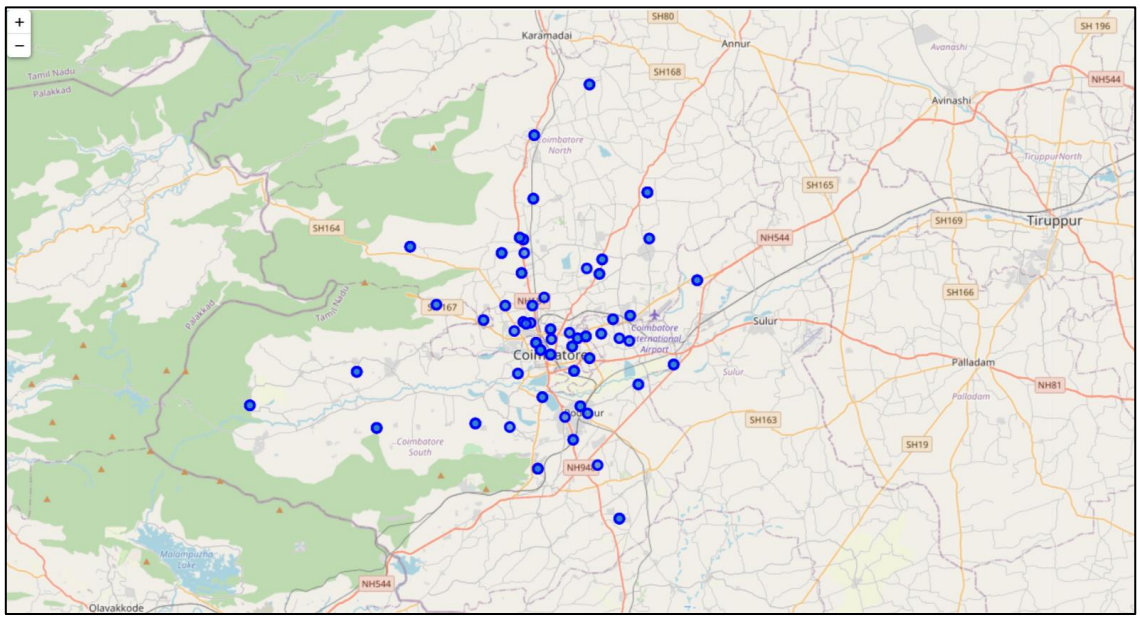

Fig -4: Locations of Post Offices

Fetching Venues using Foursquare:

Using the foursquare API, the venues that are around 500 mts from the localities are fetched and the results are displayed as follows,

	PostOffice	Taluk	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Cherannagar	Coimbatore North	11.062781	76.940771	Kada Peru Theriyala	11.059936	76.940182	Food Truck
1	Cherannagar	Coimbatore North	11.062781	76.940771	Nandhini Bakery	11.060019	76.939815	Bakery
2	Cherannagar	Coimbatore North	11.062781	76.940771	Shree Kulfi	11.059670	76.940990	Ice Cream Shop
3	Cherannagar	Coimbatore North	11.062781	76.940771	Linda	11.063611	76.944130	Fast Food Restaurant
4	Edayarpalayam	Coimbatore North	11.038393	76.928186	Edayarpalayam	11.038498	76.925066	Bus Station

Fig -5: Venues Dataframe

As seen in the first row, the name of the venue is 'Kada Peru Theriyala' which means 'Name Unknown'. All the rows with no names of the venues are removed from the dataframe. The resulting dataframe consisted of 508 venues with 61 unique categories. Out of all the venue categories, all the categories related to food are alone considered. The dataframe is then encoded and the total number of eating places are calculated for all the localities by summing up the columns. The latitudes and longitudes are then merged with the resultant dataframe which can be seen in figure 6.

```
merged
```

	PostOffice	Total Eating Places	Latitude	Longitude
0	Amritanagar	3	11.001812	76.962842
3	CBE Mpl Central Busstand	3	11.015528	76.989695
6	Cherannagar	3	11.062781	76.940771
9	Chettipalayam	3	11.001812	76.962842
12	Coimbatore Aerodrome	2	11.030835	77.023088
...	...	...	...	...
307	Vellakinar	4	11.062781	76.940771
311	Vellalapalayam Podanur	1	10.979933	77.029073
312	Vellalore	1	10.979933	77.029073
313	Venkitapuram	6	11.056904	77.073897
319	Vilankurichi	4	11.072893	77.001949

83 rows × 4 columns

Fig -6: Final Dataframe

Clustering:

K-Means clustering is performed on the resulting dataframe with the aim of clustering the whole dataset into 3 clusters thereby we can visualize the areas with high/moderate/low demands. After performing clustering operation, the post offices are visualized using folium, with different colors for each cluster. The final map is seen in figure 7.

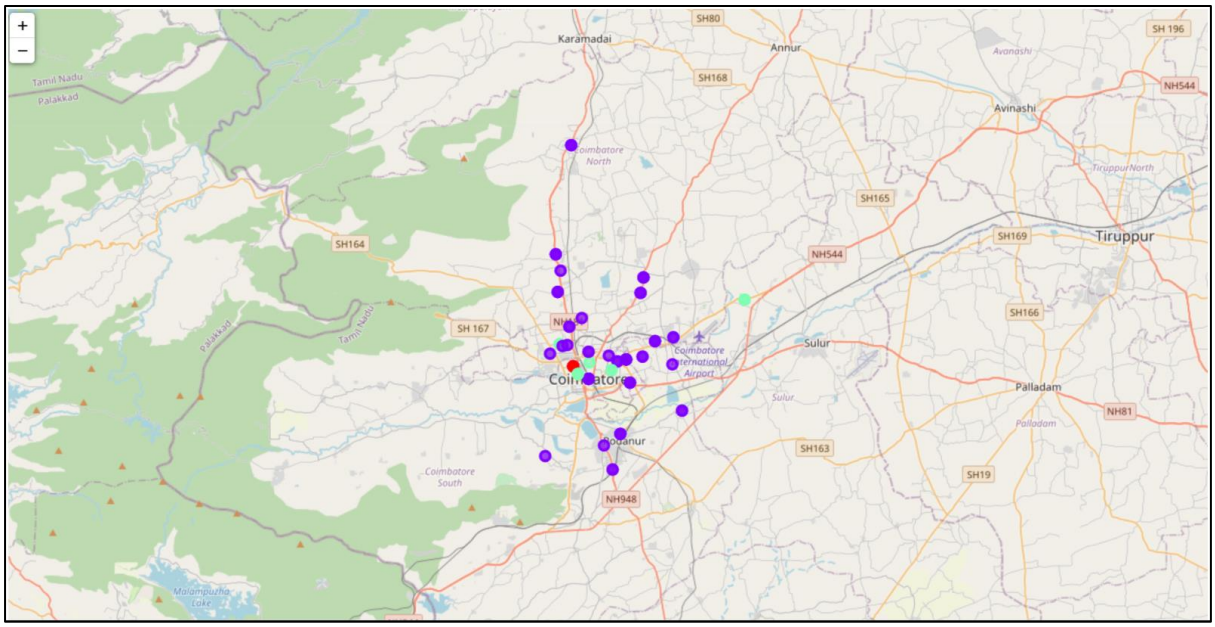

Fig -7: Map after clustering

Discussion:

When the localities of each clusters are analyzed, post offices with cluster label 0 have the highest demand (red), those with cluster label 1 have the least demand (violet), and those with cluster label 2 have a moderate demand. Thus, the areas with red markers have the highest demand for food places and hence, they are the regions to be considered while starting a restaurant.

Conclusion:

Thus, the main part of Coimbatore is analysed using various parameters and the resulting localities are grouped into 3 clusters according to the total number of venues nearby.