CAPSTONE PROJECT

ON

APPLIED DATA SCIENCE

By,

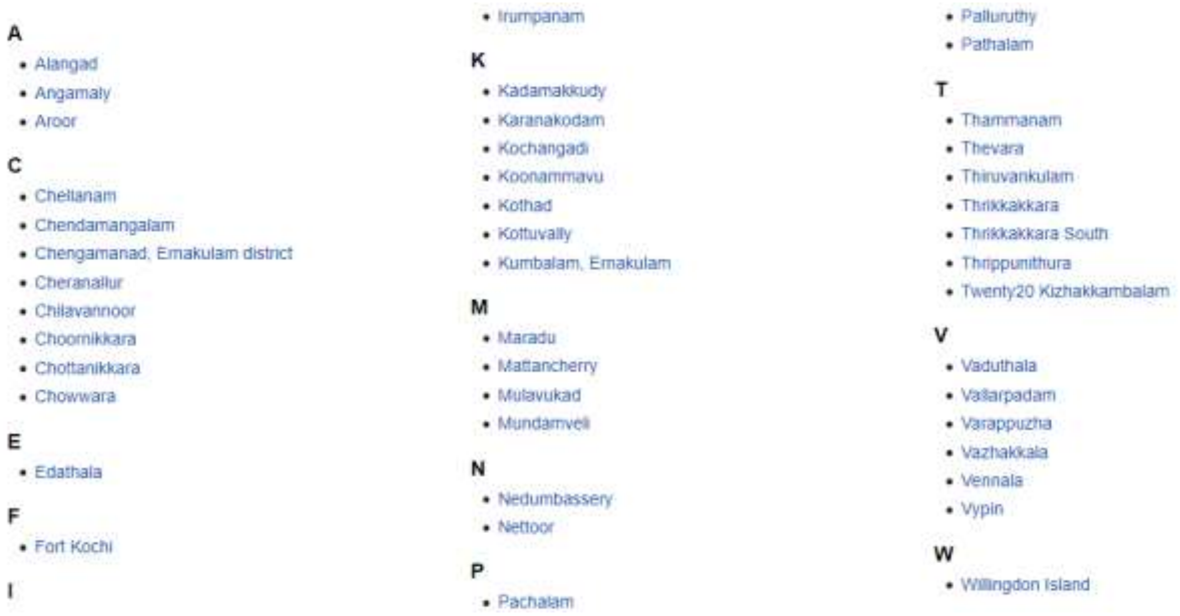
VINEETH R

Introduction

A client is interested in opening a **bakery in the city of Kochi in India**. Opening a bakery presents many unique challenges that are different from other types of businesses as there is high degree of competition. To minimise the competition, and to explore areas that do not have many bakeries, Data Science and Machine Learning tools are used to identify the best cluster of neighborhoods for opening a bakery in Kochi, India.

Data

List of neighborhoods in Kochi, India is available in Wikipedia at https://en.wikipedia.org/wiki/Category:Suburbs_of_Kochi.



A

- Alangad
- Angamaly
- Aroor

C

- Chellanam
- Chendamangalam
- Chengamanad, Ernakulam district
- Cheranallur
- Chilavannoor
- Choornikkara
- Chottanikkara
- Chowwara

E

- Edathala

F

- Fort Kochi

I

- Irumpanam

K

- Kadamakkudy
- Karanakodam
- Kochangadi
- Koonammavu
- Kothad
- Kottuvally
- Kumbalam, Ernakulam

M

- Maradu
- Mattancherry
- Mulavukad
- Mundamveli

N

- Nedumbassery
- Nettoor

P

- Pachalam
- Palluruthy
- Pathalam

T

- Thammanam
- Thevara
- Thiruvankulam
- Thrikkakkara
- Thrikkakkara South
- Thrippunithura
- Twenty20 Kizhakkambalam

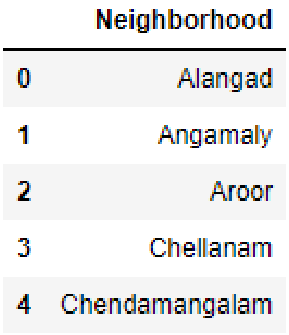
V

- Vaduthala
- Vallarpadam
- Varappuzha
- Vazhakkala
- Vennala
- Vypin

W

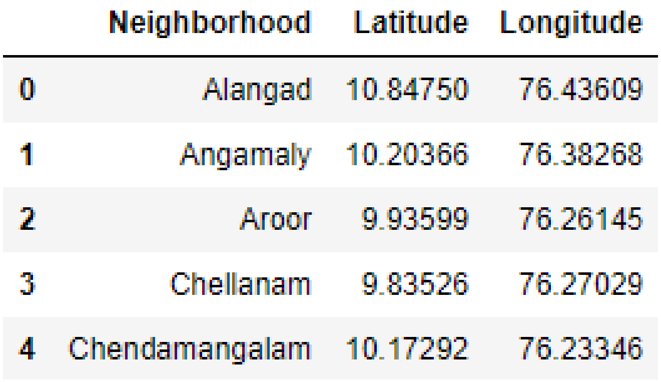
- Willingdon Island

Data frame of neighborhoods in Kochi, India can be made by scraping the data from Wikipedia page using **BeautifulSoup** library.



	Neighborhood
0	Alangad
1	Angamaly
2	Aroor
3	Chellanam
4	Chendamangalam

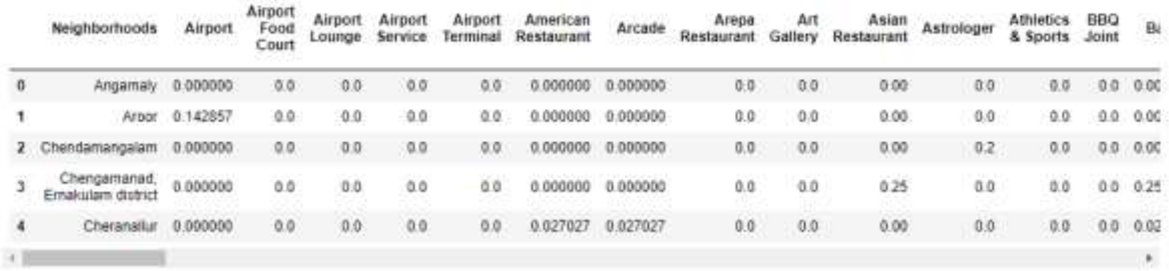
Geocoder library is used to extract the coordinates of the list of neighborhoods in Kochi. Once the Data Frame of neighborhoods in Kochi, India is made by scraping the data from Wikipedia page using **BeautifulSoup** library, the neighborhood addresses are converted into their equivalent latitude and longitude values using geocoder library



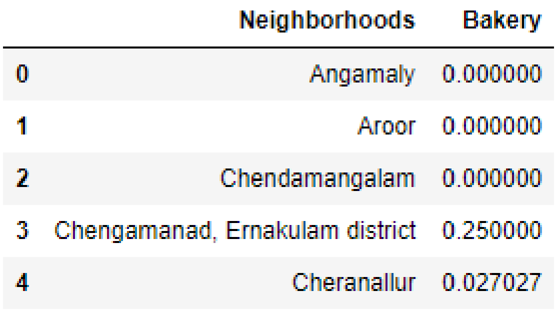
	Neighborhood	Latitude	Longitude
0	Alangad	10.84750	76.43609
1	Angamaly	10.20366	76.38268
2	Aroor	9.93599	76.26145
3	Chellanam	9.83526	76.27029
4	Chendamangalam	10.17292	76.23346

Methodology

Using the latitude & longitude coordinates, **Foursquare API** is invoked to explore neighborhoods in Kochi, India. Explore function is used to get the common venue categories in each neighbourhood.

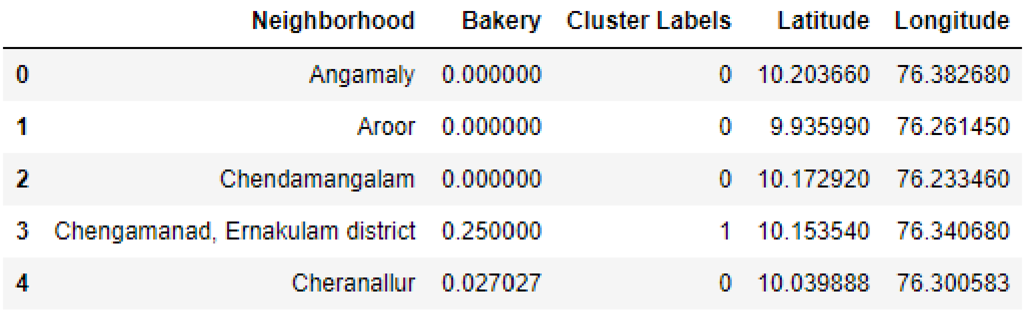


	Neighborhoods	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Arcade	Arepa Restaurant	Art Gallery	Asian Restaurant	Astrologer	Athletics & Sports	BBQ Joint	Ba
0	Angamaly	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.00	0.0	0.0	0.0	0.00
1	Aroor	0.142857	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.00	0.0	0.0	0.0	0.00
2	Chendamangalam	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.00	0.2	0.0	0.0	0.00
3	Chengamanad, Ernakulam district	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.25	0.0	0.0	0.0	0.25
4	Cheranallur	0.000000	0.0	0.0	0.0	0.0	0.027027	0.027027	0.0	0.0	0.00	0.0	0.0	0.0	0.02



	Neighborhoods	Bakery
0	Angamaly	0.000000
1	Aroor	0.000000
2	Chendamangalam	0.000000
3	Chengamanad, Ernakulam district	0.250000
4	Cheranallur	0.027027

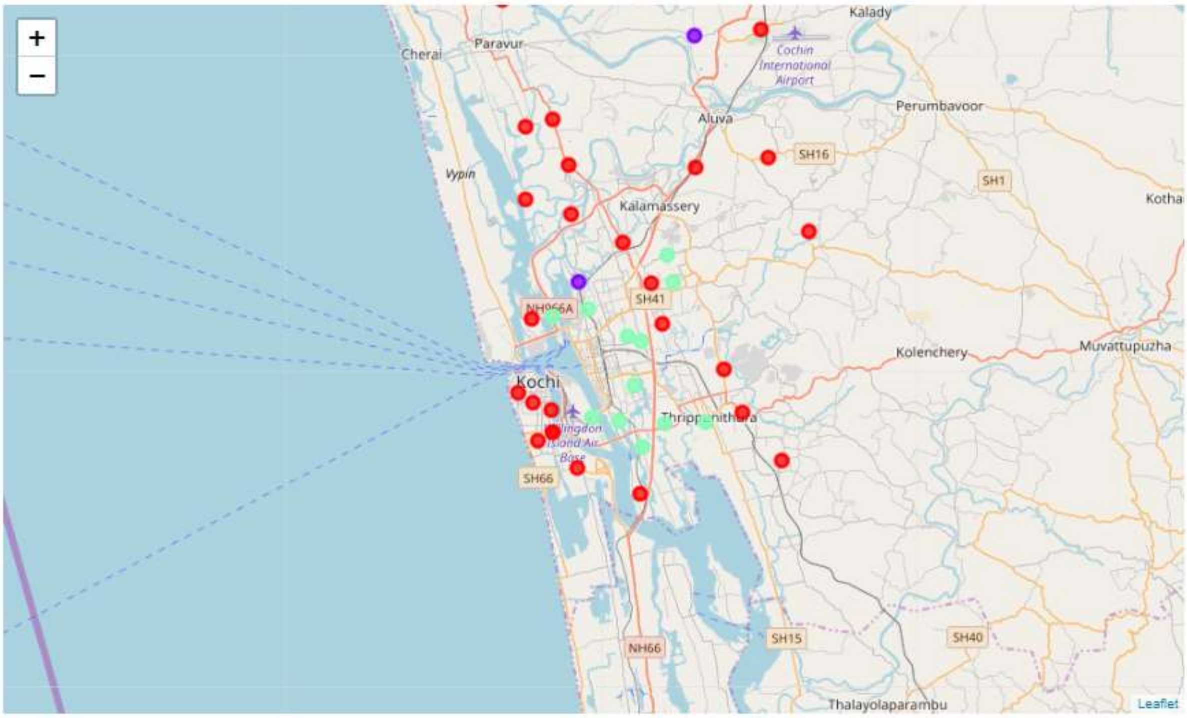
The above results are used to group the neighborhoods into clusters. **k-means** clustering algorithm is used to cluster the neighborhoods into three based on number of Bakeries: High (2), Medium (1), Low (0).



	Neighborhood	Bakery	Cluster Labels	Latitude	Longitude
0	Angamaly	0.000000	0	10.203660	76.382680
1	Aroor	0.000000	0	9.935990	76.261450
2	Chendamangalam	0.000000	0	10.172920	76.233460
3	Chengamanad, Ernakulam district	0.250000	1	10.153540	76.340680
4	Cheranallur	0.027027	0	10.039888	76.300583

Finally, **Folium** library is used to visualize the clusters of neighborhoods in Kochi India based on bakeries.

Cluster Label	**No. of Bakeries**	**Colour**
0	Low	Red
1	Medium	Violet
2	High	Green



Results

As evident from the map and chart above, it can be observed that the Neighborhoods in cluster with label 0 are the best locations to open a Bakery as the number of Bakeries are less. Neighborhoods in cluster with label 1 have medium number of Bakeries while Neighborhoods in cluster with label 2 has the highest number of Bakeries.

Discussion

It is to be noted that the information provided by the Foursquare app depends on the popularity of the application in that geographical region. Here in Kochi, India the usage of foursquare app is observed to be less which is evident from the low number of results provided by the explore query. Hence the accuracy of the analysis can be done by incorporating data from sources that are popular in that specific geographical region.

Conclusion

Data Science and Machine Learning tools were used to meet the requirements of the client in identifying the best cluster of neighborhoods for opening a bakery in Kochi, India.