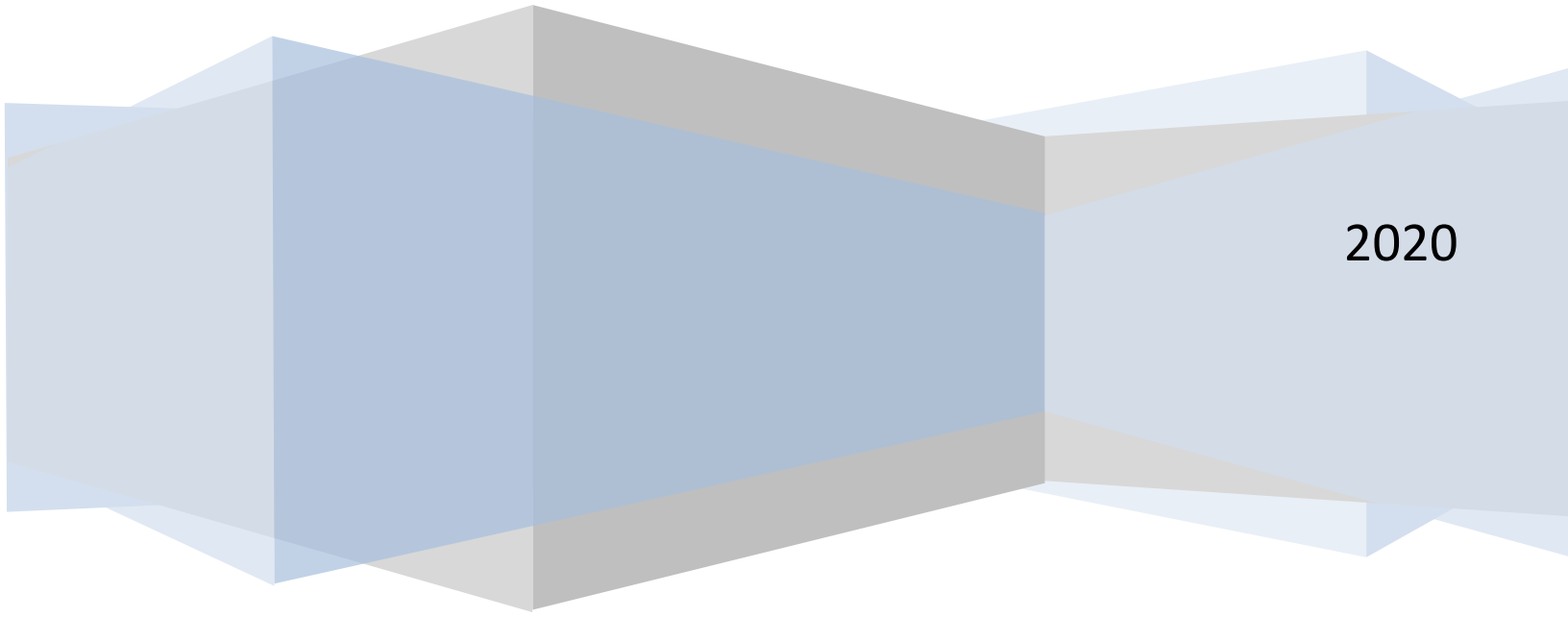
Applied Data Science Capstone

IBM Data Science Professional Certificate

Vaibhav Bodhe



2020

Introduction :

Pune has been an Educational center for India as well as Maharashtra for over decades and also the introduction of IT sector in the city has populated a huge amount of Engineers and Students in the city , these specific population has a lot technical buff on their plate , also that involves long hours of sleeplessness and focusing .

Coffee Shops/Café serving coffee is the only thing that keeps the student working on assignments for long hours and IT professionals committing to near deadlines .

Although the city is blessed with such dense population but the Coffee Shops aren't well distributed. Some neighborhoods are loaded with Cafés but others don't even have a single Coffee shop around.

Therefore through this project we try to segment the neighborhoods of Pune according to the availability of coffee shops and help the Coffee Shops/ Café franchise willing to open an outlet in Pune by availing the areas with less Competition and more customer turn over.

Business Problem :

The main objective of this project is to help the investors and Coffee Shop franchises by segmenting the area of Pune and providing them the list of areas where there is lack of Coffee Shop/Café so as to defend them from high Competition and also helping them achieve great turnout of customers. **If a Coffee shop/Café franchise is interested to open an outlet in Pune what are the areas they should be targeting..??**

Target Audience :

- Coffee Shop / Café franchise interested for an outlet in Pune.
- Coffee lovers who like to spend time working at Café**.(Secondary Audience)**

Data :

- List of Neighborhoods in Pune .This basically sets a scope of our project .
- Area Markers i.e Longitudes and Latitude of the Scope and also the neighborhoods targeted for analysis.
- Venue data , for calculation of customer occurrences and availability of specific venue in an area.

Sources for Data Availability:

- First and foremost Source for list of neighborhood in Pune can be made available through Wikipedia . A csv file is created containing the list which is further merged with area markers.
- The Coordinates of Neighborhood can be extracted from Python Geocoder Package.
- Venue data can be extracted from Foursquare.com site , which is prestigious site four venue details , ratings etc. A Developer account needs to be created so as to avail API calls.

Methodology :

Project Starts from gathering the data , initial list of neighborhoods is obtained by creating a delimited file manually the list of neighborhoods was obtained from Wikipedia (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Pune) . All the Areas (incl.Pimpri-chinchwad)were included in this file , the next step was to gather the coordinates of these neighbourhoods along with coordinates of Pune , this was swiftly done by using python geocoder package which easily helped us to convert addresses to respective latitudes and longitudes , which inturn where fed to Foursquare API , which primarily depends upon coordinates to return the venues and details specifically residing in those coordinate ,for this a developer account was created beforehand with limited API calls .

Next, we obtain top 100 venues allocated around (2km) radius , API calls were used by feeding in the coordinates using a python loop which returned the venue details in JSON format and key elements like venue name , venue category ,venue latitude ,venue longitude were extracted . After extraction starts the Analyzing part ,where each neighborhood was analyzed by calculating the mean of frequency of occurrence of each venue category. Also after analyzing we filter the data according to Coffee Shop venues as the project is limited to this scope.

After all the data preparation K-means clustering is applied on prepared data frame. *k*-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes. Value of k is set to three (k=3) , i.e 3 clusters of varied frequency coffee shop occurrences in different neighborhood . Result will allow us to identify which neighborhood have higher or lower concentrations of

coffee shops ,it will help us answer the question as to which neighborhoods are suitable to open a new shopping mall.

Results :

The results after clustering (k=3) clearly exhibits that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Coffee shop / café ”:

- Cluster 0(RED): Neighborhoods with moderate number of Coffee Shops .
- Cluster 1(PURPLE): Neighborhoods with low number to no existence of Café.
- Cluster 2(MINT GREEN): Neighborhoods with high concentration of Coffee shops .

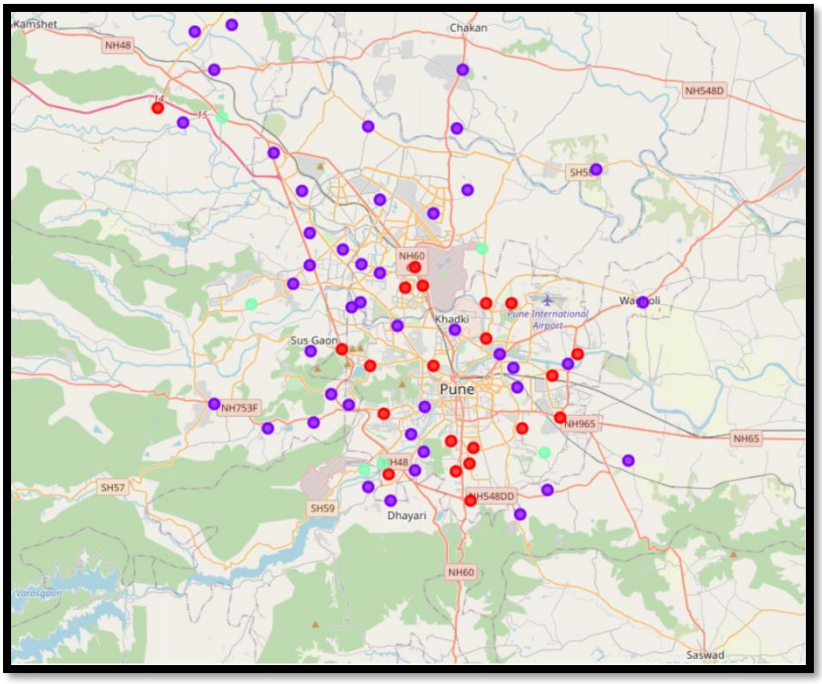


Fig.All the clusters are visualized with markers.

Clusters :

	Neighborhood	Coffee Shop	Cluster Labels	Latitude	Longitude
36	Manjri	0.055556	0	18.48194	73.86562
44	Parvati	0.062500	0	18.48696	73.85006
34	Kothrud	0.059701	0	18.50517	73.80245
45	Pashan	0.086957	0	18.53674	73.79290
46	Phugewadi	0.066667	0	18.59017	73.83032
31	Kharadi	0.086957	0	18.54462	73.93922
47	Pimple Gurav	0.062500	0	18.58900	73.81815
29	Katraj	0.076923	0	18.44732	73.86405
28	Kasarwadi	0.055556	0	18.60263	73.82435
26	Kalas	0.083333	0	18.57845	73.87489
23	Hadapsar	0.111111	0	18.50253	73.92706
53	Shivajinagar	0.070000	0	18.53723	73.83808
17	Dhanori	0.050000	0	18.57856	73.89264
16	Dhankawadi	0.047619	0	18.46628	73.85326
40	Mundhwa	0.049383	0	18.53017	73.92125
70	Wanowrie	0.081633	0	18.49538	73.90009
11	Bibvewadi	0.068966	0	18.47187	73.86336
62	Urse	0.111111	0	18.70864	73.64322
5	Baner	0.054054	0	18.54820	73.77316
66	Vishrantwadi	0.086957	0	18.55533	73.87492
64	Vadgaon Khurd	0.083333	0	18.46458	73.80617

	Neighborhood	Coffee Shop	Cluster Labels	Latitude	Longitude
38	Mohammed Wadi	0.130435	2	18.478670	73.915940
54	Shivane	0.250000	2	18.467810	73.788970
35	Maan	0.125000	2	18.577920	73.709530
71	Warje	0.153846	2	18.472110	73.802130
55	Somatne	0.142857	2	18.701912	73.688788
19	Dighi	0.166667	2	18.615220	73.872410

	Neighborhood	Coffee Shop	Cluster Labels	Latitude	Longitude
58	Talegaon	0.000000	1	18.734130	73.683340
42	Panmala	0.000000	1	18.876470	73.897080
43	Parandwadi	0.000000	1	18.699230	73.661130
69	Wakad	0.000000	1	18.604050	73.750290
68	Wagholi	0.000000	1	18.579530	73.985290
67	Vadgaon Sheri	0.035714	1	18.537890	73.932670
65	Vadgaon Maval	0.000000	1	18.771930	73.404300
48	Pimple Nilakh	0.020000	1	18.579080	73.786530
63	Vadgaon Budruk	0.000000	1	18.467280	73.824730
49	Pimple Saudagar	0.000000	1	18.598540	73.800240
50	Pirangut	0.000000	1	18.511230	73.683170
51	Rahatani	0.000000	1	18.604660	73.787250
52	Ravet	0.000000	1	18.653650	73.744730
41	Nanded	0.000000	1	18.456420	73.792000
57	Talawade	0.000000	1	18.696425	73.792049
61	Undri	0.000000	1	18.454270	73.917880
60	Thergaon	0.000000	1	18.614490	73.773860
59	Tathawade	0.000000	1	18.625480	73.750690
56	Sus	0.000000	1	18.546700	73.751130
0	Akurdi	0.000000	1	18.764080	73.695730
72	Yerwada	0.020408	1	18.544829	73.884679
8	Bhosari	0.000000	1	18.638730	73.837480
9	Bhugaon	0.000000	1	18.499220	73.753160
10	Bhukum	0.000000	1	18.495100	73.721240
12	Chakan	0.000000	1	18.734150	73.858560
13	Chimbali	0.000000	1	18.695120	73.854060
14	Chinchwad	0.000000	1	18.647430	73.800020
15	Dehu Road	0.000000	1	18.678880	73.724820
18	Dhayari	0.000000	1	18.447020	73.807570
39	Moshi	0.000000	1	18.653760	73.861950
4	Balewadi	0.018519	1	18.576020	73.779830
3	Aundh	0.017857	1	18.563450	73.812270
20	Erandwane	0.020000	1	18.509650	73.831240
21	Fursungi	0.000000	1	18.473650	73.974730
22	Ghorpadi	0.012195	1	18.522320	73.897120
24	Hingne Khurd	0.025000	1	18.479790	73.830750
25	Hinjawadi	0.000000	1	18.591420	73.738950
27	Karve Nagar	0.025000	1	18.491500	73.821720
30	Khadki	0.000000	1	18.561140	73.853000
32	Kondhwa	0.000000	1	18.438250	73.898950
33	Koregaon Park	0.020000	1	18.535330	73.893820
2	Ambi	0.000000	1	18.759300	73.669720
1	Ambegaon	0.000000	1	19.004960	73.945830
37	Markal	0.000000	1	18.667570	73.952570
7	Bavdhan Khurd	0.027027	1	18.511100	73.777730
6	Bavdhan Budruk	0.000000	1	18.518270	73.765570

Discussion :

As inferred from the visualization of clusters it is clear that the franchise targeting the areas in Cluster2 will be facing a lot of competition from its peers as the number of Coffee shop in this areas is highly concentrated whereas Cluster1 serves to be the new market for cafes as these areas gravely lack in those amenities , also Cluster0 can be a fair choice but moderation in number wouldn't hype up the sales but can surely attract moderate amount of customers with fair amount of competition.

Best Areas to be targeted : Areas in Cluster 1 (PURPLE)

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 machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e Coffee franchise eager to invest in Pune region by opening a new outlet. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations to open a new Café . The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Coffee shop / Café .

References:

- Category: Neighborhoods in Pune , India .Retrieved from (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Pune)

- Foursquare Developers Documentation. Foursquare. Retrieved from https://developer.foursquare.com/docs.