COURSERA - IBM

PROFFESIONAL CERTIFICATE COURES

APPLIED DATA SCIENCE CAPSTONE

TO LOCATE A SUITABLE PLACE TO OPEN A NEW SHOPPING MALL

ANANDAKRISHNAN

APRIL 2020

INTRODUCTION

For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. They can do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more activities. Shopping malls are like a one-stop destination for all types of shoppers. For retailers, the central location and the large crowd at the shopping malls provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more shopping malls to cater to the demand. As a result, there are many shopping malls in the city of Bangalore and many more are being built. Opening shopping malls allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping mall is one of the most important decisions that will determine whether the mall will be a success or a failure

BUSINESS PROBLEM

The objective of this capstone project is to analyse and select the best locations in the city of BANGALORE, KARANATAKA to open a new shopping mall. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question.

BUSINESS QUESTION

In the city of BANGALORE, KARANATAKA if a property developer is looking to open a new shopping mall, where would you recommend that they open it?

TARGET AUDIENCE OF THIS PROJECT

This project is particularly useful to property developers and investors looking to open or invest in new shopping malls in the IT HUB of INDIA, i.e, BANGALORE KARNATAKA. This project is timely as the city is currently suffering from oversupply of shopping malls.

DATA

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

1. List of neighborhoods in Bangalore. This defines the scope of this project which is confined to the city of Bangalore.

2. Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.

3. Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the neighborhoods.

SOURCES OF DATA AND METHODS TO EXTRACT THEM

This Wikipedia page (https://commons.wikimedia.org/wiki/Category:Suburbs_of_Bangalore) contains a list of neighbourhoods in Bangalore, with a total of 59 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use **Foursquare API** to get the venue data for those neighbourhoods.

- Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers.
- Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order to help us to solve the business problem put forward.
- This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with **API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium)**.
- In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

METHODOLOGY

Firstly, we need to get the list of neighborhoods in the city of Bangalore. Fortunately, the list is available in the Wikipedia page (https://commons.wikimedia.org/wiki/Category:Suburbs_of_Bangalore). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude.

After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinate's data returned by Geocoder are correctly plotted in the city of Kuala Lumpur.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 3000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key.

We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude

and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues.

Then, we will analyse each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Shopping Mall” data, we will filter the “Shopping Mall” as venue category for the neighborhoods.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project.

We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Shopping Mall”. The results will allow us to identify which neighborhoods have higher concentration of shopping malls while which neighborhoods have fewer number of shopping malls.

Based on the occurrence of shopping malls in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new shopping malls.

DATA FRAME CREATED FROM DATA IN WIKIPEDIA

	Neighborhood
0	⊢ Agara, Bangalore (2 C, 6 F)
1	⊢ Arekere (5 F)
2	⊢ Banashankari (1 C, 4 F)
3	⊢ Banaswadi (2 F)
4	⊢ Basavanagudi (5 C, 11 F)

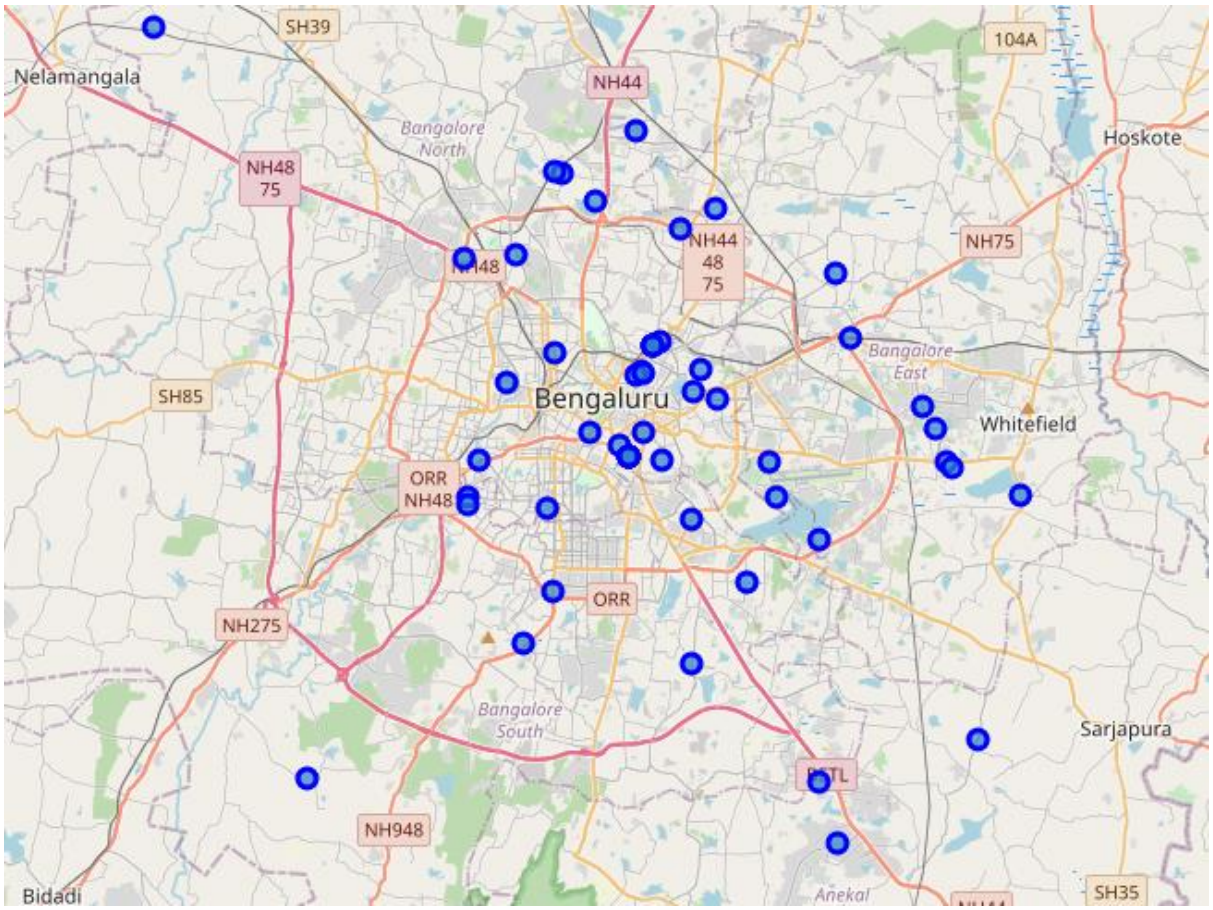
FINDING THE GEOGRAPHICAL COORDINATES

```
]: [[12.841269941511115, 77.4815905253185],
    [12.997973409660101, 77.61047545115127],
    [12.90876015015293, 77.57288203719202],
    [12.997820409660104, 77.61038642561601],
    [12.938980000000072, 77.57137000000006],
    [12.882450000000063, 77.62475000000006],
    [12.927350000000047, 77.67185000000006],
    [12.981176497085812, 77.62506386955391],
    [12.966180000000065, 77.58690000000007],
    [12.958034857060447, 77.6009343090484],
    [12.817530000000033, 77.67879000000005],
    [12.966235877087087, 77.6067910877087],
    [12.855496201771626, 77.73111547336825],
    [13.250110000000063, 77.70788000000005],
    [12.997856409660098, 77.61040795115127],
    [12.943290000000047, 77.65602000000007],
    [12.839884329698537, 77.67221065989045],
    [12.998940000000061, 77.61276000000004],
    [12.942790000000059, 77.54122000000007],
    [13.047455598389988, 77.63327823980035],
    [13.049810000000036, 77.58903000000004],
    [12.957454844018699, 77.60090848251485],
    [12.912160000000029, 77.64490000000006],
    [12.978220000000022, 77.633970000000003]
```

ADDING LATITUDE & LONGITUDE TO DATAFRAME

	Neighborhood	Latitude	Longitude
0	⊢ Agara, Bangalore (2 C, 6 F)	12.841270	77.481591
1	⊢ Arekere (5 F)	12.997973	77.610475
2	⊢ Banashankari (1 C, 4 F)	12.908760	77.572882
3	⊢ Banaswadi (2 F)	12.997820	77.610386
4	⊢ Basavanagudi (5 C, 11 F)	12.938980	77.571370

CREATE A MAP OF BANGALORE WITH NEIGHBORHOODS SUPERIMPOSED ON TOP



AFTER USING FOURSQUARE API WE HAVE MODIFIED THE DATAFRAME

:	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	⊢ Agara, Bangalore (2 C, 6 F)	12.841270	77.481591	Art of Living International Center	12.844607	77.507343	Spiritual Center
1	⊢ Agara, Bangalore (2 C, 6 F)	12.841270	77.481591	rachenamadu	12.850793	77.505317	Nature Preserve
2	⊢ Agara, Bangalore (2 C, 6 F)	12.841270	77.481591	ѕнιvα ѕαι ιηdυѕтяy αηd тяαdeяѕ	12.816927	77.491872	Outdoor Supply Store
3	⊢ Arekere (5 F)	12.997973	77.610475	Mangalore Pearl - Seafood Restaurant	12.994472	77.615551	Seafood Restaurant
4	⊢ Arekere (5 F)	12.997973	77.610475	Naturals Icecream	12.996912	77.611268	Ice Cream Shop

CLUSTERING PROCESS BEGINS

6]:

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
29	← Kettohalli (1 C)	0.010000	0	12.957455	77.600908
27	← Jayanagar, Bangalore (1 C, 7 F)	0.010000	0	12.961606	77.597723
57	← Yeshwantpur (1 C, 6 F)	0.018519	0	13.029540	77.540220
31	← Konanakunte (1 F)	0.018868	0	12.890437	77.561763
35	← Madiwala (1 C, 6 F)	0.020000	0	12.956603	77.613550
36	← Magadi (2 C, 10 F)	0.020000	0	12.988227	77.605822
37	← Mahadevapura (2 C)	0.010000	0	12.958035	77.600934
38	← Majestic (Bangalore) (1 C)	0.010000	0	12.957455	77.600908
39	← Malleswaram (4 C, 2 F)	0.020000	0	12.995000	77.573460
40	← Marathahalli (8 C, 1 P, 30 F)	0.020619	0	12.955740	77.719419
25	← J. P. Nagar (2 C)	0.010000	0	12.958035	77.600934

EXAMINING CLUSTER 0

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
29	← Kettohalli (1 C)	0.010000	0	12.957455	77.600908
27	← Jayanagar, Bangalore (1 C, 7 F)	0.010000	0	12.961606	77.597723
57	← Yeshwantpur (1 C, 6 F)	0.018519	0	13.029540	77.540220
31	← Konanakunte (1 F)	0.018868	0	12.890437	77.561763
35	← Madiwala (1 C, 6 F)	0.020000	0	12.956603	77.613550
36	← Magadi (2 C, 10 F)	0.020000	0	12.988227	77.605822
37	← Mahadevapura (2 C)	0.010000	0	12.958035	77.600934
38	← Majestic (Bangalore) (1 C)	0.010000	0	12.957455	77.600908
39	← Malleswaram (4 C, 2 F)	0.020000	0	12.995000	77.573460
40	← Marathahalli (8 C, 1 P, 30 F)	0.020619	0	12.955740	77.719419
25	← J. P. Nagar (2 C)	0.010000	0	12.958035	77.600934
41	← Mathikere (1 C)	0.010000	0	13.030327	77.559672

EXAMINING CLUSTER 1

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
48	← Seetharampalya (1 C, 14 F)	0.0	1	13.113200	77.424630
46	← Ramamurthy Nagar (1 C, 20 F)	0.0	1	13.023820	77.677850
45	← Rajarajeshwari Nagar, Bangalore (1 C, 4 F)	0.0	1	12.940380	77.541539
10	← Bommasandra (33 F)	0.0	1	12.817530	77.678790
12	← Chandapura (4 F)	0.0	1	12.855496	77.731115
13	← Devanahalli (5 C, 2 F)	0.0	1	13.250110	77.707880
3	← Banashankari (1 C, 4 F)	0.0	1	12.908760	77.572882
0	← Agara, Bangalore (2 C, 6 F)	0.0	1	12.841270	77.481591
16	← Electronics City (2 C, 34 F)	0.0	1	12.839884	77.672211
32	← Koramangala (1 C, 12 F)	0.0	1	12.935130	77.624450
18	← Girinagar (1 C, 11 F)	0.0	1	12.942790	77.541220
30	← Kodihalli, Bangalore (1 C, 4 F)	0.0	1	13.059765	77.576729

EXAMINING CLUSTER 2

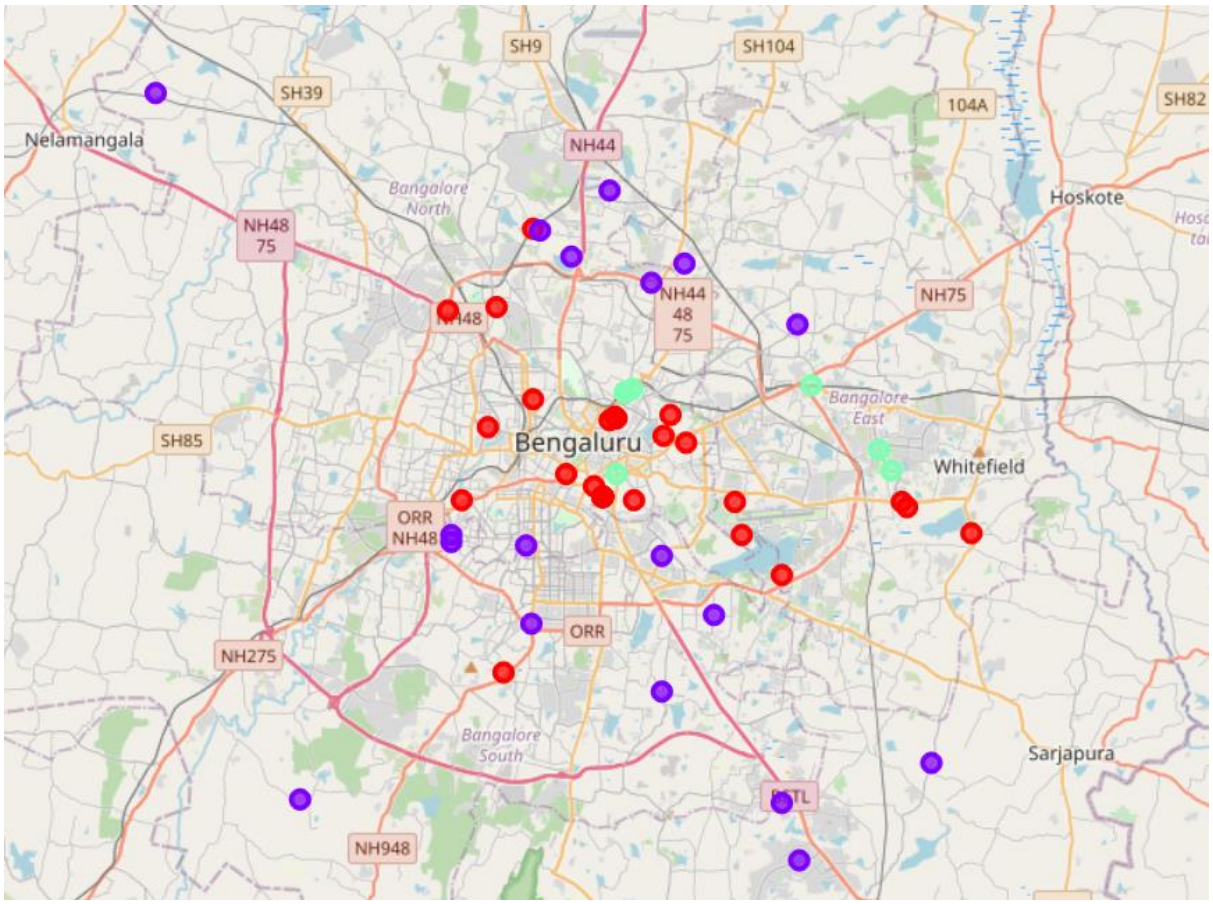
	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
1	← Arekere (5 F)	0.030000	2	12.997973	77.610475
55	← Whitefield, Bangalore (4 C, 1 P, 22 F)	0.050000	2	12.975970	77.710317
24	← Ittamadu (3 F)	0.030000	2	12.997856	77.610408
14	← Dhobi Ghat (Bangalore) (3 F)	0.030000	2	12.997856	77.610408
17	← Fraser Town, Bangalore (1 C, 10 F)	0.030000	2	12.998940	77.612760
34	← Kundalahalli (96 F)	0.030000	2	12.967520	77.715000
33	← Krishnarajapura (3 C, 3 F)	0.030769	2	13.000390	77.683680
28	← Jeevanbheemanagar (3 F)	0.030000	2	12.997856	77.610408
4	← Banaswadi (2 F)	0.030000	2	12.997820	77.610386
11	← Brigade Road, Bangalore (3 C, 8 F)	0.030000	2	12.966236	77.606791

RESULTS

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for "Shopping Mall":

- Cluster 0: Neighborhoods with moderate number of shopping malls
- Cluster 1: Neighborhoods with low number to no existence of malls
- Cluster 2: Neighborhoods with high concentration of shopping malls

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.



DISCUSSION

As observations noted from the map in the Results section, most of the shopping malls are concentrated in the central area of Bangalore city, with the highest number in cluster 2 and moderate number in cluster 0.

On the other hand, cluster 1 has very low number to no shopping mall in the neighborhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls.

Meanwhile, shopping malls in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. From another perspective, the results also show that the oversupply of shopping malls mostly happened in the central area of the city, with the suburb area still have very few shopping malls.

Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in neighborhoods in cluster 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighborhoods in cluster 0 with moderate competition.

Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of shopping malls and suffering from intense competition.

LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

In this project, we only consider one factor i.e. frequency of occurrence of shopping malls, there are other factors such as population and income of residents that could influence the location decision of a new shopping mall.

However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new shopping mall.

In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

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. property developers and investors regarding the best locations to open a new shopping mall.

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 shopping mall.

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 shopping mall.