IBM APPLIED DATA SCIENCE CAPSTONE COURSE BY COURSERA

Submitted By: Nikhila Joshy

Start your shopping mall in Kochi, India

Project Report - 2 Jan 2021

1. Introduction

A storm of global trends are coming together at the same time to cause malls to change the role they play in people's lives. The main reason shopping malls are so popular is that they are very convenient. Most of them include clothing stores, a food court, movie theatres, arcade rooms, and much more. Not every shopping mall is the same, but they all basically have the same things. Shopping for basic needs is an old trend, but it is changing rapidly and shopping malls, these days, are primarily recreational places for many. This new trend has increased in the past years. For instance, many children prefer to meet up in a shopping mall, watch a movie and later have snacks in the food court. Besides, with the decline of open places and parks, many families gather at a large shopping mall to enjoy their holidays.

2. Business Problem

Starting a mini shopping mall can be quite rewarding and also demanding because it needs proper planning and complete dedication. Today in cities, the mall structure is quite popular and most of the people love to purchase things from mall. This is not just because of the ambience but also the facility. All the things that you need in your daily life, starting from home product to luxury items, everything is available under one roof. This project aims to tackle the problem of finding a suitable location to build a new shopping mall in Kochi. It can be extended to any place other than Kochi and can also be used to find suitable locations for buildings other than shopping malls like gyms, hotels etc.

Footfall is very important for success in the mall business. We should get in touch with a franchise having good customer base. Large number of people in rural area has access to internet and they have learned to buy things that they need from shopping sites available online. Thus before venturing into the business, we have to keep in mind that we not only have to deal with the competition in the retail market but also plan a unique strategy so that customers are attracted towards our mall and an increase in footfall is seen.

Things to consider while starting a shopping mall:

Anybody who has proper knowledge about running a store can open a mini shopping mall. Here are a few things that you would need to pay attention to before venturing into the business.

1. Decide on the niche you would like to target. Usually if your target is teenagers and women below the age of 40 then make sure that the mall you open looks trendy.

2. Normally, when a mall is opened people in the locality just come in with a motive to spend time and see what the stores have on shelf but if a particular franchise has strong

customer base then the foot traffic will naturally increase and you will find people dropping in from different area as well.

3. Location plays a vital role in the business. Look for an area where there are no competitors around. If there are more malls in the locality you are planning to set your business, then competition will be high and profit will be less.

This project will take care of the making an informed decision about the location.

3. Data

To complete this project, we need the following:

1. Since the scope of this project is confined to the district of Kochi in Kerala, India, we need the list of all the neighbourhoods of Kochi.

    *   The neighbourhood information will be scraped from the Wikipedia webpage https://en.wikipedia.org/wiki/Category:Suburbs_of_Kochi.

2. Latitude and longitude coordinates of those neighbourhoods to plot the map and to get venue details.

    *   Geocoder package in python is used to get the coordinates of the neighbourhoods.

3. Venue data to perform clustering

   *   Foursquare API will be exploited to get the venue details for all the neighbourhoods.

These steps will explore many data science skills including web scraping, K-means clustering, map visualization using Folium, working with Foursquare API, data cleaning and data wrangling.

4. Methodology

- Firstly, the list of neighbourhoods in the city of Kuala Lumpur is extracted from the Wikipedia page https://en.wikipedia.org/wiki/Category:Suburbs_of_Kochi using web scraping with beautifulsoup and other python packages. However, this is just a list of names. Geocoder package will allow us to convert address into geographical coordinates in the form of latitude and longitude to be able to use the Foursquare API. This concludes the gathering of data.

	Neighborhood	Latitude	Longitude
0	Alangad	10.847500	76.436090
1	Angamaly	10.203660	76.382680
2	Aroor	9.936010	76.261420
3	Chellanam	9.835260	76.270290
4	Chendamangalam	10.172920	76.233460
5	Chengamanad, Ernakulam district	10.153540	76.340680
6	Cheranallur	10.060850	76.289010
7	Chilavannoor	9.961180	76.306590
8	Choornikkara	10.081320	76.341550
9	Chottanikkara	9.939106	76.375036
10	Chowwara	10.128150	76.372170
11	Edathala	10.086410	76.381810
12	Fort Kochi	9.957580	76.242390
13	Irumpanam	9.966870	76.357200
14	Kadamakkudy	10.063520	76.246600
15	Karanakodam	9.988447	76.303427
16	Kochangadi	9.947600	76.260790
17	Koonammavu	10.107690	76.261710
18	Kothad	10.055610	76.271640
19	Kottuvally	10.103150	76.246150
20	Kumbalam, Ernakulam	9.902200	76.310640

- Then, the data is populated into a pandas DataFrame and we can visualize the neighbourhoods in a map using Folium package. This ensures that the geographical coordinates returned by Geocoder are correctly plotted in the city of Kochi.
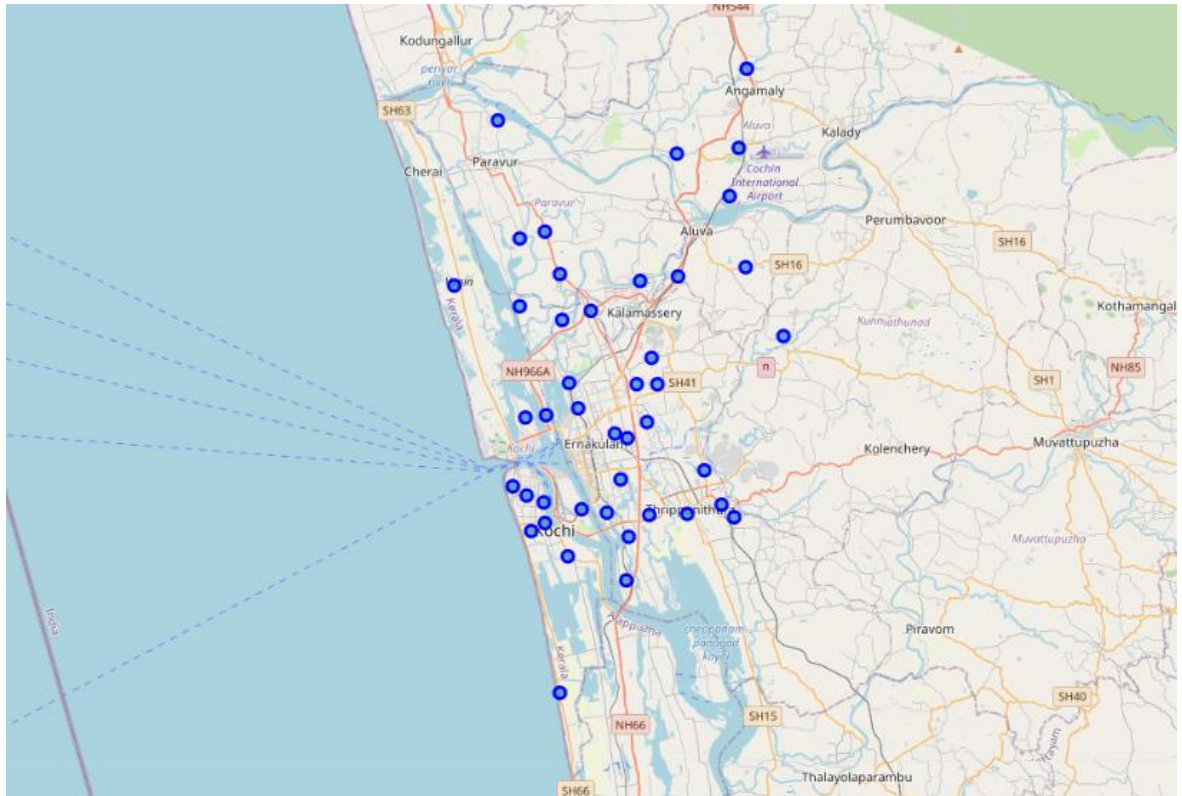
- Next, Foursquare API is used to get the top 100 venues that are within a radius of 2000 meters. API calls are made to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues.

```
# Let's find out how many unique categories can be curated from all the returned venues
print('There are {} uniques categories.'.format(len(venues_df['VenueCategory'].unique())))
```

```
    There are 125 uniques categories.
```

```
# print out the list of categories
venues_df['VenueCategory'].unique()[:125]
```

```
]: array(['Multiplex', 'Bus Station', 'Indian Restaurant', 'Hotel',
        'Fried Chicken Joint', 'Indie Movie Theater', 'Hotel Bar',
        'Restaurant', 'Gym', 'Halal Restaurant', 'Fish Market',
        'Department Store', 'Smoke Shop', 'Fast Food Restaurant',
        'Tea Room', 'Light Rail Station', 'Auditorium', 'Asian Restaurant',
        'Historic Site', 'River', 'Boat or Ferry', 'Astrologer',
        'Comfort Food Restaurant', 'Bakery', 'Food Truck',
        'South Indian Restaurant', 'Café', 'Thai Restaurant',
        'French Restaurant', 'Nightclub', 'Stadium', 'Athletics & Sports',
        'Sandwich Place', 'Ice Cream Shop', 'Chinese Restaurant',
        'Motorcycle Shop', 'Burger Joint', 'Donut Shop', 'Park',
        'Gym / Fitness Center', 'Juice Bar', 'Pizza Place',
        'Middle Eastern Restaurant', 'Clothing Store', 'Coffee Shop',
        'Vegetarian / Vegan Restaurant', 'Dhaba', 'Bar', 'Lounge',
```

- Then, each neighbourhood is analysed by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. The data is now prepared for clustering. Since we are analysing the "Shopping Mall" data, we will filter the "Shopping Mall" as venue category for the neighbourhoods. Lastly, we will perform clustering on the data by using k-means clustering.

- 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.
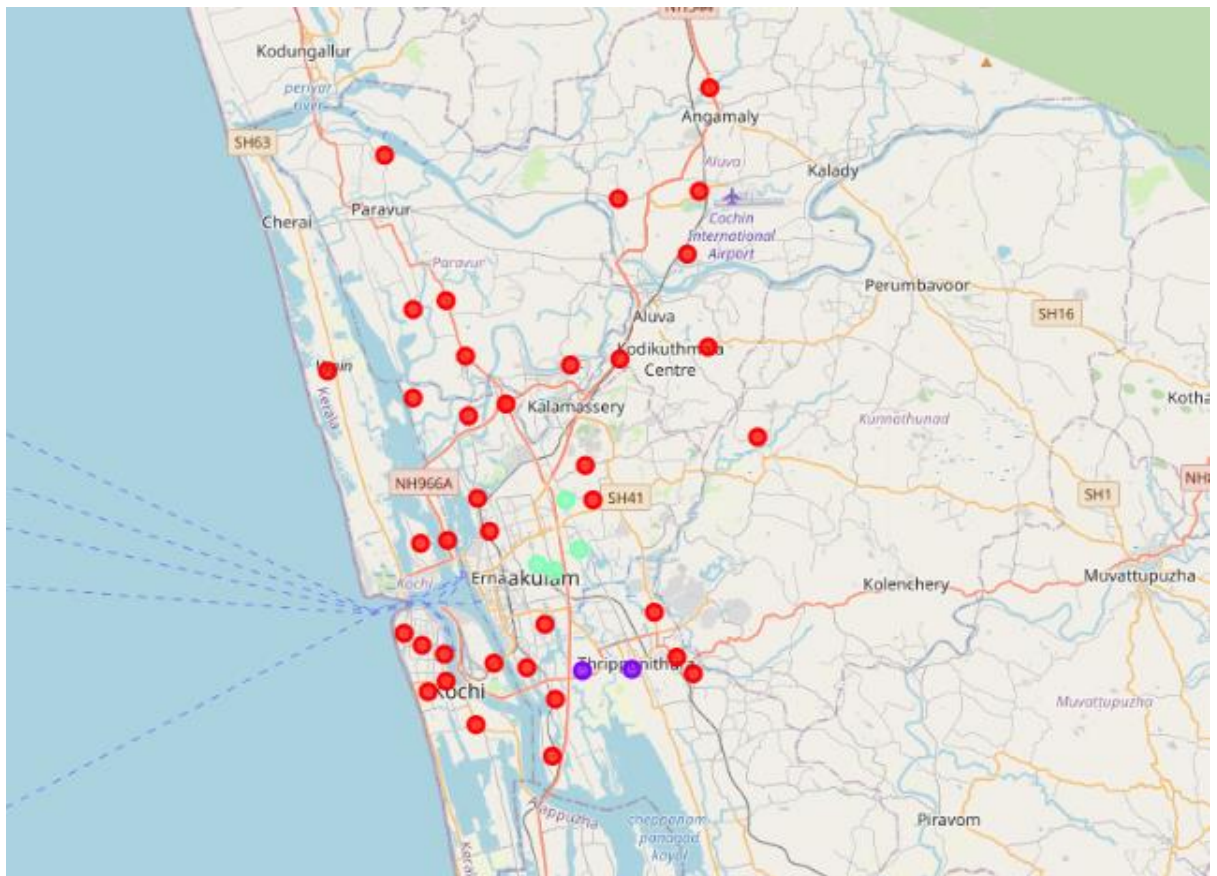
- K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. The neighbourhoods are grouped into 3 clusters based on their frequency of occurrence for "Shopping Mall". The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new shopping malls.

5. Results

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

- Cluster 0: Neighbourhoods with no shopping malls.

```
kl_merged.loc[kl_merged['Cluster Labels'] == 0]
```

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
0	Angamaly	0.0	0	10.203660	76.382680
21	Mulavukad	0.0	0	9.999140	76.262410
22	Mundamveli	0.0	0	9.930700	76.253200
23	Nedumbassery	0.0	0	10.156690	76.377800
24	Nettoor	0.0	0	9.927260	76.311810
25	Pachalam	0.0	0	10.003470	76.281220
26	Palluruthy	0.0	0	9.916420	76.275670
27	Pathalam	0.0	0	10.078170	76.318570
29	Thevara	0.0	0	9.942090	76.298390
30	Thiruvankulam	0.0	0	9.946350	76.367460
32	Thrikkakkara South	0.0	0	10.033240	76.325190
34	Twenty20 Kizhakkambalam	0.0	0	10.046260	76.404110
35	Vaduthala	0.0	0	10.018250	76.275860
36	Vallarpadam	0.0	0	9.997890	76.249810
37	Varappuzha	0.0	0	10.082610	76.270410
38	Vazhakkala	0.0	0	10.017830	76.329080
40	Vypin	0.0	0	10.075380	76.207020
18	Kumbalam, Ernakulam	0.0	0	9.902200	76.310640
20	Mattancherry	0.0	0	9.952060	76.250800
7	Chottanikkara	0.0	0	9.939106	76.375036
1	Aroor	0.0	0	9.936010	76.261420
2	Chendamangalam	0.0	0	10.172920	76.233460
3	Chengamanad, Ernakulam district	0.0	0	10.153540	76.340680
4	Cheranallur	0.0	0	10.060850	76.289010
5	Chilavannoor	0.0	0	9.961180	76.306590
6	Choornikkara	0.0	0	10.081320	76.341550
17	Kottuvally	0.0	0	10.103150	76.246150
8	Chowwara	0.0	0	10.128150	76.372170

- Cluster 1: Neighbourhoods with highest concentration of shopping malls.

```
kl_merged.loc[kl_merged['Cluster Labels'] == 1]
```

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
33	Thrippunithura	0.058824	1	9.94111	76.34698
19	Maradu	0.043478	1	9.94051	76.32395

- Cluster 2: Neighbourhoods with moderate to high number of shopping malls.

```
kl_merged.loc[kl_merged['Cluster Labels'] == 2]
```

7]:

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
31	Thrikkakkara	0.022472	2	10.017360	76.316370
13	Karanakodam	0.014286	2	9.988447	76.303427
39	Vennala	0.029412	2	9.995380	76.322430
28	Thammanam	0.021739	2	9.985570	76.311300

6. Discussion

- Kochi is a place with its fair number of shopping destinations for people. It can be seen that shopping malls are concentrated in cluster 1 with almost as many shopping malls as cluster 1 in cluster 2 and none in cluster 0.

- This shows that there is a lot of potential for shopping malls in areas that belong to cluster 0 because of the lack of competition and a need for a common destination point with restaurants, banks, theatres, professional offices, service stations, and other establishments. The area required to build a shopping mall can also be acquired at a low cost in cluster 0 and hence it would be easier to break even and start getting profits.

- If a shopping mall is brought about in cluster 1 or cluster 2, it would suffer from the intense competition between the already existing shopping outlets and the upcoming ones. Therefore, this project can be utilized by project architects to capitalize on these observations in building shopping malls in different neighbourhoods of Kochi.

7. 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 urbanization, 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 neighbourhood 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.

- This project can also be extended to find suitable locations to open a new restaurant, gym or universities. It can also be reused to check for locations outside India by very little modification to the code.

8. Conclusion

In this project, we specified the business problem definition and identified the methodology to be used, extracted and prepared the data and then performed K - means clustering algorithm to cluster the data into three clusters based on the similarities in the frequencies of shopping malls and finally drawing inference from the results that were obtained. This can then be used by the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new shopping mall. The findings of this project will help to capitalize on the opportunities on high potential locations while avoiding overcrowded areas and the competition that follows it. It was found out from the above data that the neighbourhoods in cluster 0 are the most preferred locations to open a new shopping mall in Kochi, Kerala.