Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Café in Kochi, Kerala, India

By: Job Thomas T

June 2020

Introduction

Kochi is becoming one of the busiest places in India. Kochi, known (a little tritely) as The Queen of the Arabian Sea, has much to offer the curious traveler, whatever their interests might be. While a lot of this has to do with its affinity for the arts, and embracing of the new and modern, its rich and colorful history cannot be denied. Kochi was one of the most important port cities in the international spice trade from the late 1400s, drawing traders from Europe, West Asia and China. Besides these influences, the city also had a small but thriving population of Jews and Anglo-Indians. The cuisine, architecture, and general culture features all of these influences including the traditions of the Portuguese, Dutch and English colonizers.

The city has always been quietly obsessed with food. However, in recent years—mostly thanks to the Kochi Muziris Biennale—the rest of the world has been let in on this open secret. And they just can't get enough.

As for an entrepreneur this is the right time to start a food joint such as a café in this vibrant city. Of course, as with any business decision, opening a new café requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the café 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 analyze and select the best locations in the city of Kochi, Kerala, India to open a new café. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Kochi, Kerala, India, if a property developer is looking to open a new café, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful for young entrepreneurs and investors looking to open or invest in café in Kochi. This project is timely as the city is becoming more popular with tourists and is one of the best place in Kerala with a night life.

Data

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

- List of neighborhoods in Ernakulam district. This defines the scope of this project.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to cafe. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

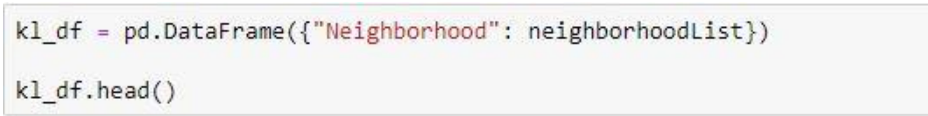
This Wikipedia page ("https://en.wikipedia.org/wiki/Category:Suburbs_of_Kochi") contains a list of neighborhoods in Kochi, with a total of 44 neighborhoods. I myself have added one more which seemed relevant. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers.

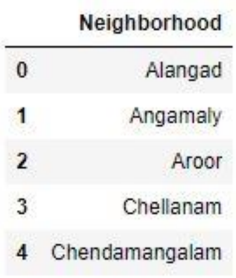
Foursquare API will provide many categories of the venue data, we are particularly interested in the Café category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), 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 Kochi. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Suburbs_of_Kochi). 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 Kochi.

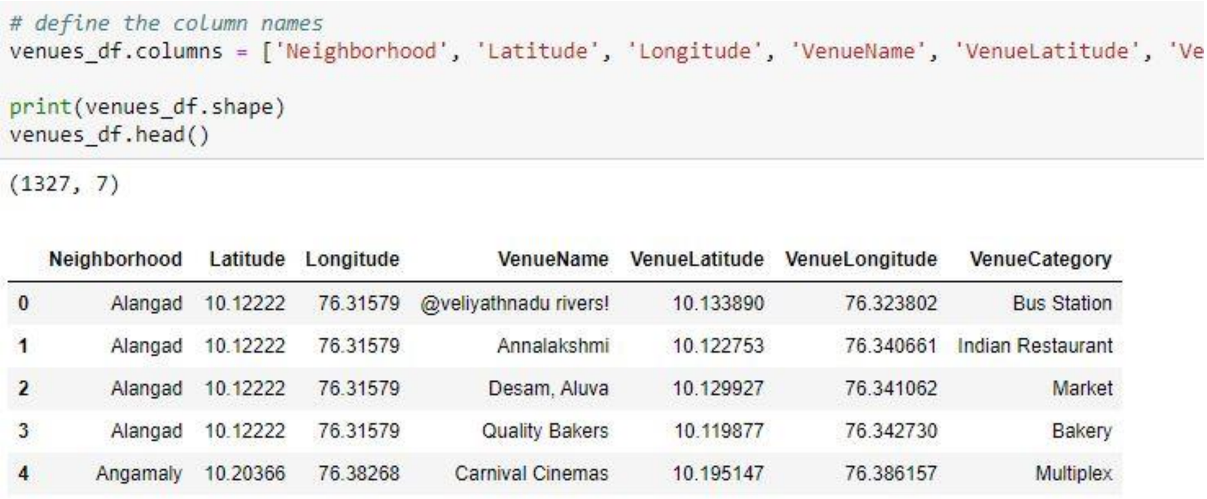


```
kl_df = pd.DataFrame({"Neighborhood": neighborhoodList})

kl_df.head()
```



	Neighborhood
0	Alangad
1	Angamaly
2	Aroor
3	Chellanam
4	Chendamangalam

Next, we will use Foursquare API to get the top 200 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 analyze 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 analyzing the "Café" data, we will filter the "Café" as venue category for the neighborhoods.



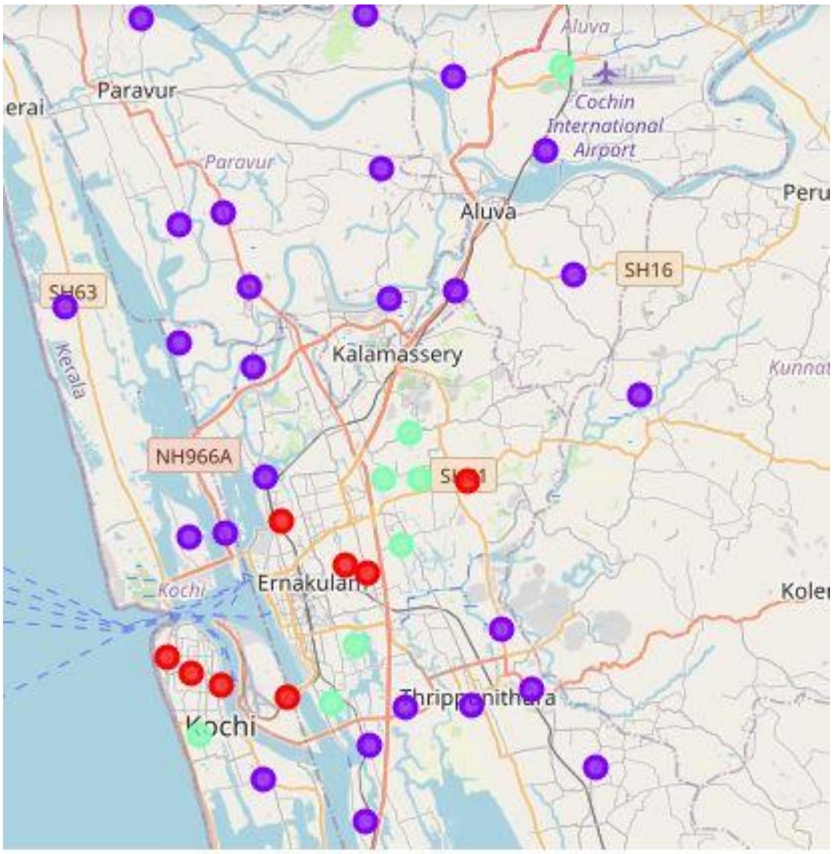
```
# define the column names
venues_df.columns = ['Neighborhood', 'Latitude', 'Longitude', 'VenueName', 'VenueLatitude', 'Ve

print(venues_df.shape)
venues_df.head()
```

```
(1327, 7)
```

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Alangad	10.12222	76.31579	@veliyathnadu rivers!	10.133890	76.323802	Bus Station
1	Alangad	10.12222	76.31579	Annalakshmi	10.122753	76.340661	Indian Restaurant
2	Alangad	10.12222	76.31579	Desam, Aluva	10.129927	76.341062	Market
3	Alangad	10.12222	76.31579	Quality Bakers	10.119877	76.342730	Bakery
4	Angamaly	10.20366	76.38268	Carnival Cinemas	10.195147	76.386157	Multiplex

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 "Café". The results will allow us to identify which neighborhoods have higher concentration of Cafés while which neighborhoods have fewer number of Cafés. Based on the occurrence of Cafés in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Cafés.

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 "Café":

- Cluster 0: Neighborhoods with high number of Cafés (Red)
- Cluster 1: Neighborhoods with low number to no existence of Cafés (Purple)
- Cluster 2: Neighborhoods with moderate concentration of Cafés (Light Blue)

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in light Blue color.

Discussion

As observations noted from the map in the Results section, most of the Cafés are concentrated in cluster 0 and 2 which is the central area of Kochi in the Ernakulam district. Cluster 1 has low number of cafe. But these areas are the outskirts of the city and the possibility of people going to café is very low. Also such areas have very less night life which is disadvantageous to open up a café. Entrepreneurs and challenging people who has the capability to stand out from the competition can open new Cafés in neighborhoods in **cluster 2** with moderate competition. Since this cluster area is within the city, they have a good prospects of becoming successful. Lastly, Entrepreneurs are advised to avoid neighborhoods in cluster 0 which already have high concentration of Cafés 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 Cafés, which itself is not the exact data. 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 Café. 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. entrepreneurs and business man regarding the best locations to open a new Café. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 2 are the most preferred locations to open a new Café. The findings of this project will help the relevant entrepreneurs to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Café.

References

Category:Suburbs in Kuala Lumpur. *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Category:Suburbs_of_Kochi

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

Appendix

Cluster 0

- Bangsar South
- Bukit Bintang
- Bukit Nanas
- Bukit Tunku
- Chow Kit
- Damansara Heights
- Damansara Town Centre
- Damansara, Kuala Lumpur
- Dang Wangi
- Jalan Cochrane, Kuala Lumpur
- Jalan Duta
- Kampung Baru, Kuala Lumpur
- Medan Tuanku
- Mont Kiara
- Segambut
- Setiawangsa
- Shamelin
- Taman Desa
- Taman Tun Dr Ismail

Cluster 1

- Alam Damai
- Ampang, Kuala Lumpur
- Bandar Menjalara
- Bandar Sri Permaisuri
- Bandar Tasik Selatan
- Bandar Tun Razak
- Batu 11 Cheras
- Batu, Kuala Lumpur
- Bukit Jalil
- Bukit Kiara
- Bukit Petaling
- Desa Petaling
- Federal Hill, Kuala Lumpur
- Happy Garden
- Jinjang
- Kampung Datuk Keramat
- Kepong
- Kuchai Lama
- Maluri
- Miharja
- Pantai Dalam
- Putrajaya
- Salak South
- Semarak
- Sentul Raya
- Setapak
- Sri Hartamas
- Sri Petaling
- Sungai Besi
- Taman Bukit Maluri
- Taman Cheras Hartamas
- Taman Connaught
- Taman Ibukota
- Taman Len Seng
- Taman Melati
- Taman Midah
- Taman OUG
- Taman P. Ramlee
- Taman Sri Sinar
- Taman Taynton View
- Taman Wahyu
- Titiwangsa
- Wangsa Maju

- Cheras, Kuala Lumpur

Cluster 2

- Bangsar
- Bangsar Park
- Brickfields
- KL Eco City
- Lembah Pantai
- Pudu, Kuala Lumpur
- Taman U-Thant