IBM Applied Data Science Capstone Project

Analysis of Restaurants near Hospitals in Bangalore, India

Maria Franklin Judia

September 2020

Introduction

Hospitals are one of the most essential services that is required for the proper functioning of any community. There are people travelling even from different countries in order to access specialized medical services provided by the respective Hospitals. In such a scenario, extended periods of stay are often envisioned and good restaurants are sought out near the Hospitals for the food requirements of family members and visitors. While nutritious, appealing food in hospitals may not have yet evolved to the point that all stakeholders would like, advances are being made. Concerns persist with respect to many issues including insufficient budgets and human resources; local and sustainable food procurement challenges; ensuring food safety and sustainability; balancing nutrition and taste; plate waste; and barriers to patient eating.

Business Problem

The objective of this capstone project is to analyse Hospitals in the city of Bangalore, India to get a statistics of the Restaurants around it. 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 Bangalore, if a patient's family or bystander is looking for food venues, do they have enough options and if not, for investing in a new restaurant, where can we recommend it be opened?

Target audience for this project

The target audience of this project is twofold: First, the family and friends of patients visiting, who would get a clear idea of the restaurants around them and second, to property developers and investors looking to open or invest in new restaurants in Bangalore. We can analyse the top venues around the Hospitals and check the significance of each place relative to the Hospital. Comfortable and spacious dining areas with a choice of healthy foods would be of good demand near Multi-Speciality Hospitals. These places will give ample options for people visiting, if they prefer not to eat Hospital food for the various reasons discussed above.

Data

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

• List of hospitals in Bangalore. This defines the scope of this project which is confined to the city of Bangalore, India.

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

• Venue data, particularly data related to Restaurants. We will use this data to perform clustering on the Hospitals.

- From Foursquare API, the following were retrieved for each venue:
  - **Name:** The name of the venue.
  - **Category:** The category type as defined by the API.
  - **Latitude:** The latitude value of the venue.
  - **Longitude:** The longitude value of the venue.

Sources of data and methods to extract them

https://en.wikipedia.org/wiki/Category:Hospitals_in_Bangalore is the Wikipedia page that contains a list of Hospitals in Bangalore, with a total of 27 Hospitals. 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 Hospitals using Python Geocoder package which will give us the latitude and longitude coordinates of the Hospitals. After that, we will use Foursquare API to get the venue data for those Hospitals. 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 Restaurant 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

We begin our analysis by getting the list of Hospitals in the city Bangalore. The list is available in the Wikipedia page:

https://en.wikipedia.org/wiki/Category:Hospitals_in_Bangalore

i. Web Scraping and getting geographical co-ordinates

We proceed onto web scraping using Python requests and BeautifulSoup packages to extract the list of Hospitals data from the webpage. Since the result is a list of names, we require the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. For this, we will use the 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 Hospitals in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical co-ordinates data of Hospitals, returned by Geocoder, are correctly plotted in the map of the city of Bangalore.

ii. Explore venues using Foursquare API

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 5000 meters. For this we use a Foursquare Developer Account to obtain the Foursquare Client ID and secret key. We then make API calls to Foursquare passing in the geographical coordinates of the Hospitals 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 in the neighbourhood of each Hospital and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each Hospital by grouping the rows 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 Restaurant data, we will filter the ‘Restaurant’ as venue category for the Hospitals.

iii. Clustering

In the concluding steps, we 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 most popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the Hospitals into 3 clusters based on their frequency of occurrence for ’Restaurants’ near them.

The results will allow us to identify which Hospitals have higher concentration of Restaurants and which of them don't. Based on the occurrence of Restaurants around different Hospitals, it will help us to answer the question as to which areas near Hospitals are most suitable to open new Restaurants and which of them already are brimming with the required food services

Exploratory Data Analysis

a) Plotting the location of Hospitals in Bangalore

We get the geographical co-ordinates of Bangalore using the Geocoder package in Python. We then create the map of Bangalore using Folium and add markers on it that show the location of the Hospitals that we use for our analysis as show in Fig 1.1

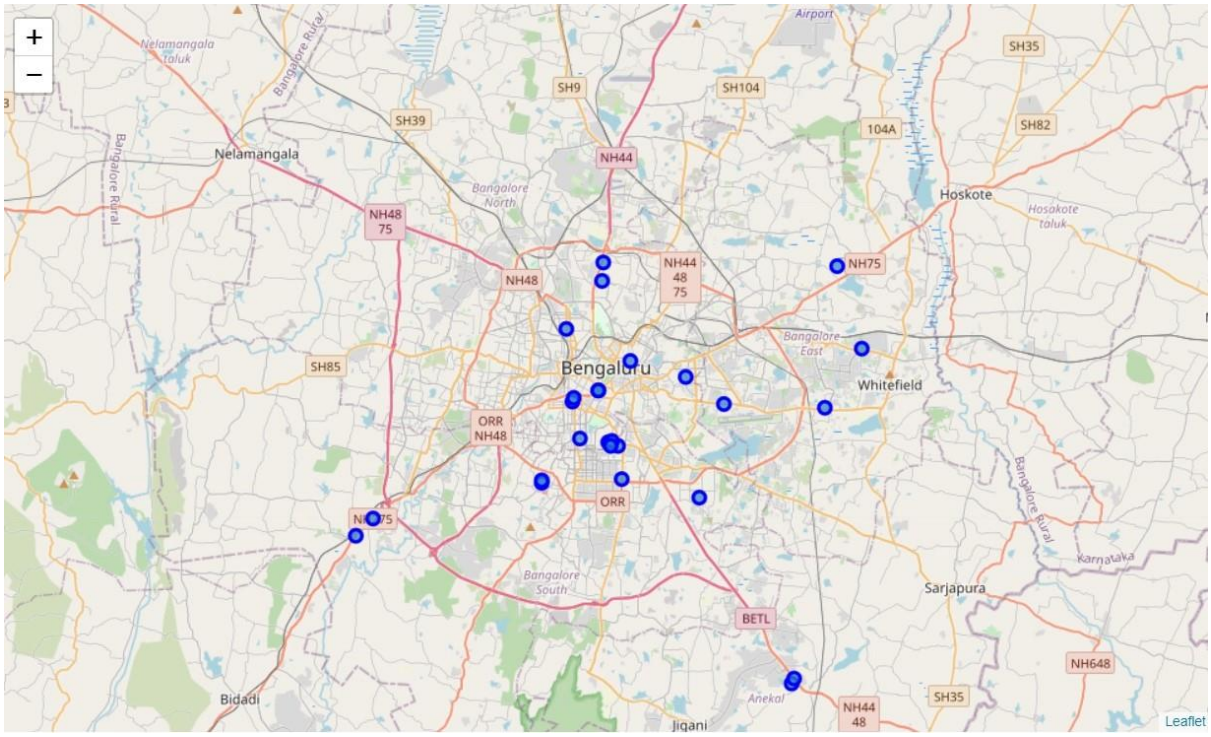

Figure 1.1

b) Number of venues in each category near the Hospitals

After using Foursquare API to get the venues and their categories around the required Hospital co-ordinates, we use the groupby() and unique() function to plot the various types of venues versus their count on a bar plot as shown in Fig 1.2

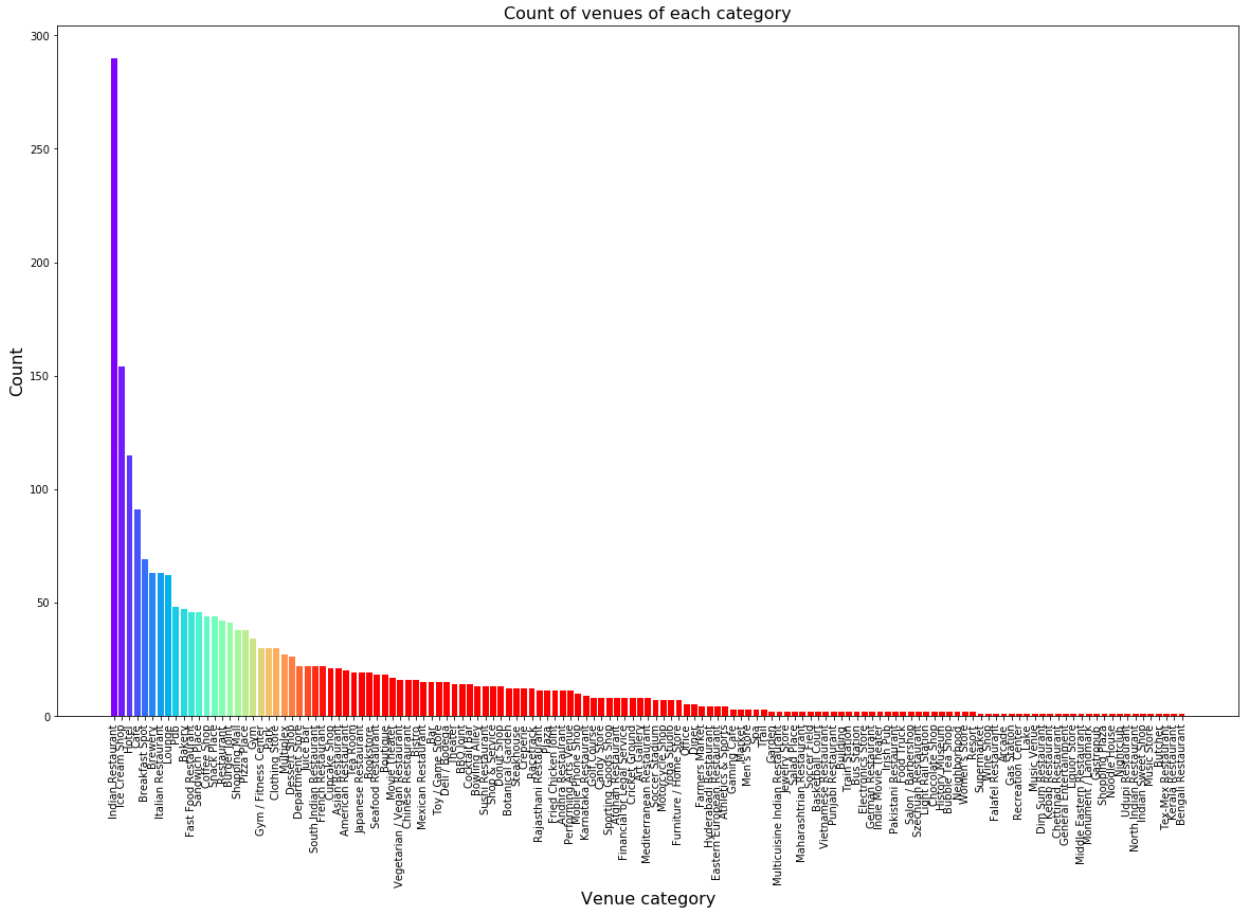

Figure 1.2

c) Mean of Restaurants around the various Hospitals

After one-hot encoding, we group the hospitals as per the mean of the number of restaurants around them. Then we plot the Hospital vs the mean of Restaurants around them as shown in Fig 1.3

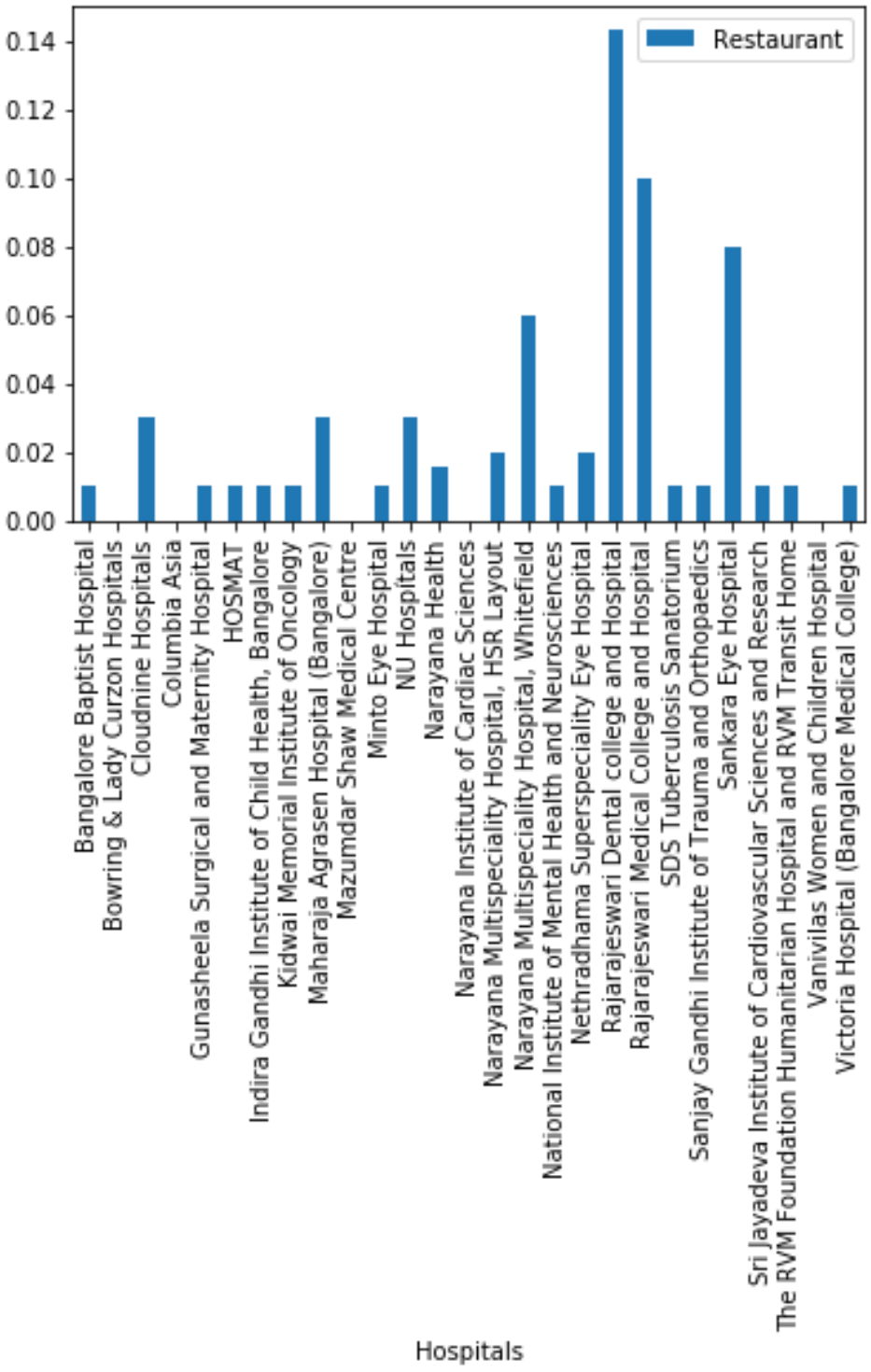

Figure 1.3

Results

The results from the k-means clustering show that we can categorize the Hospitals into 4 clusters based on the frequency of occurrence for “Restaurant”:

- Cluster 0: Hospitals with a very low number to no existence of Restaurants
- Cluster 1: Hospitals with moderate number of Restaurants
- Cluster 2: Hospitals with low number of Restaurants
- Cluster 3: Hospitals with high concentration of Restaurants

The results of the clustering are visualized in the map shown in Fig 2.1 with cluster 0 in red colour, cluster 1 in purple colour, cluster 2 in blue colour and cluster 3 in mint green colour.

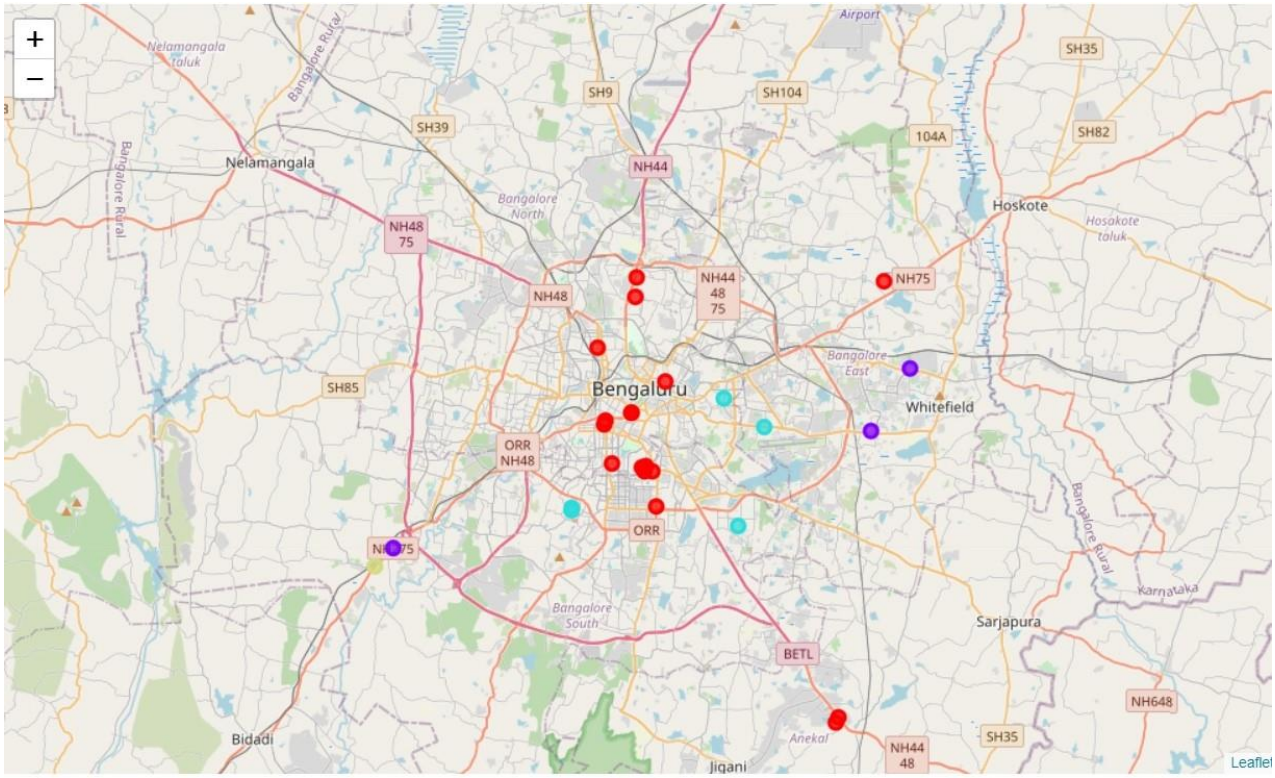

Figure 2.1

Discussion

As observations noted from the map in the Results section, most of the Hospitals are concentrated in the central area of Bangalore city, with the highest number of Restaurants in cluster 3 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to no Restaurants near the Hospitals. This represents a great opportunity and high potential areas to open new Restaurants as there is very little to no competition from existing places. Meanwhile, Restaurants in cluster 3 are likely suffering from intense competition due to oversupply and high concentration of the same.

From another perspective, the results also show that the concentration of Hospitals mostly happened in the central area of the city, with the suburb area having very few Hospitals. Therefore, this project recommends property developers to capitalize on these findings to open new Restaurants near Hospitals in cluster 0 and 2 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new Restaurants near Hospitals in cluster 1 with moderate competition. Lastly, property developers are advised to avoid Restaurants in cluster 3 which already have high concentration of Restaurants 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 Restaurants, there are other factors such as Stations and Hotels nearby, population and income of residents that could influence the location decision of a new Restaurant. However, to the best knowledge of this researcher such data are not available to the Hospital 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 Restaurant. 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 4 clusters based on their similarities, and lastly providing recommendations to the bystanders and visitors of patients as well as the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new Restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The Hospitals in clusters 0 and 2 are the most preferred locations to open a new Restaurant. 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 Restaurant near a Hospital.

References

1. Category: Suburbs in Kuala Lumpur. *Wikipedia*. Retrieved from

   https://en.wikipedia.org/wiki/Category:Hospitals_in_Bangalore

2. Foursquare Developers Documentation. *Foursquare*. Retrieved from

   https://developer.foursquare.com/docs

3. Hospital News. Bringing local food to health care food service. Retrieved from

   http://hospitalnews.com/bringing-local-food-to-health-care-food-service/

4. Pell, Amanda. 2017. Millennial tastes are driving marketers crazy, but it's doing the food industry good. Upworthy February 7, 2017. Retrieved from

   http://www.upworthy.com/millennial-tastes-are-driving-marketers-crazy-but-its-doingthe-food-industry-good

5. The national voice of healthcare organizations and hospitals across Canada website:

   https://www.healthcarecan.ca/news-events/press-releases/