IBM DATA SCIENCE CAPSTONE PROJECT

Opening Department Store in Optimised Places of Bangalore, India

By: Swagata Chakraborty

May 2021

1. Introduction:

1.1 Motivation:

Groceries are the basic needs of day-to-day life. From monthly ration to everyday needs like milk, curd, bread etc everything is bought from a departmental store. People often prefer to have such stores near their residence. Thus, every colony, or society requires a standard departmental store to satisfy their needs. This results in the need for a large number of such stores everywhere. Industrialists, manufacturer, and many tradesmen often invests in the evergreen grocery business. Such a business has very little risk factors, even entrepreneurs can invest in it. However, location of the store can be a factor that affects this business. In this project, we will try to find an optimised place for opening a departmental store.

1.2 Business Problem:

The purpose of this capstone project is to analyse, visualize, cluster areas and predict where should we place our store. We will be using data science methodologies and machine learning predictive techniques in order to make our prediction as accurate as possible. The project aims to answer the business question: Where should a person open a new department store in the city of Bangalore, India.

1.3 Target Audience:

Industrialists, manufactures, tradesmen, entrepreneurs and small shopkeepers can also find this project useful. This

project targets all audiences who are interested in opening departmental stores in Bangalore, India.

2. Data:

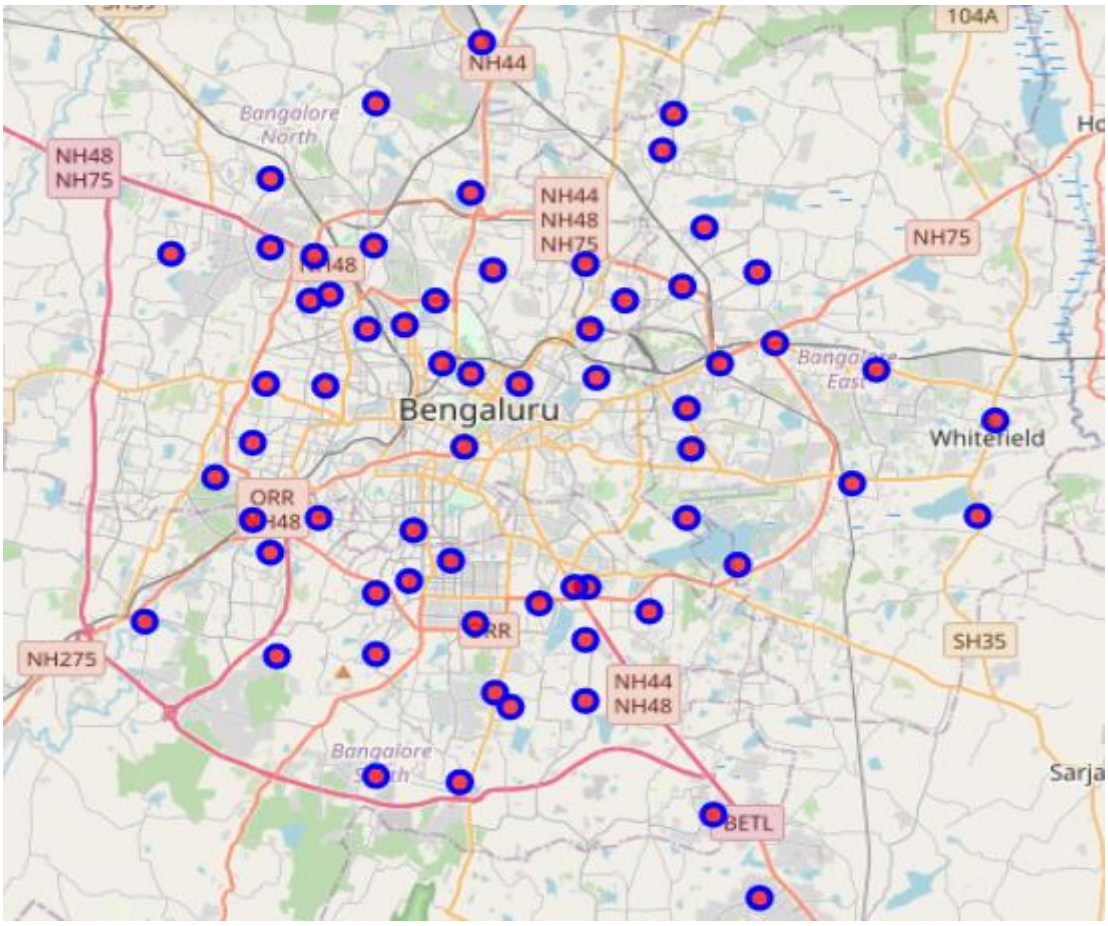
2.1 Data and its source:

- First, we need a list of neighbourhoods in Bangalore, India. I found this list on Wikipedia here.
- Next, we had to generate the latitude and longitude of these neighbourhoods.
- The latitude and longitude of Bangalore was also required during map generation.
- Lastly, we had to explore the surrounding 1000m area of each neighbourhood in order to get venues. For this we will use FourSquare API thus need a developer account in FourSquare.

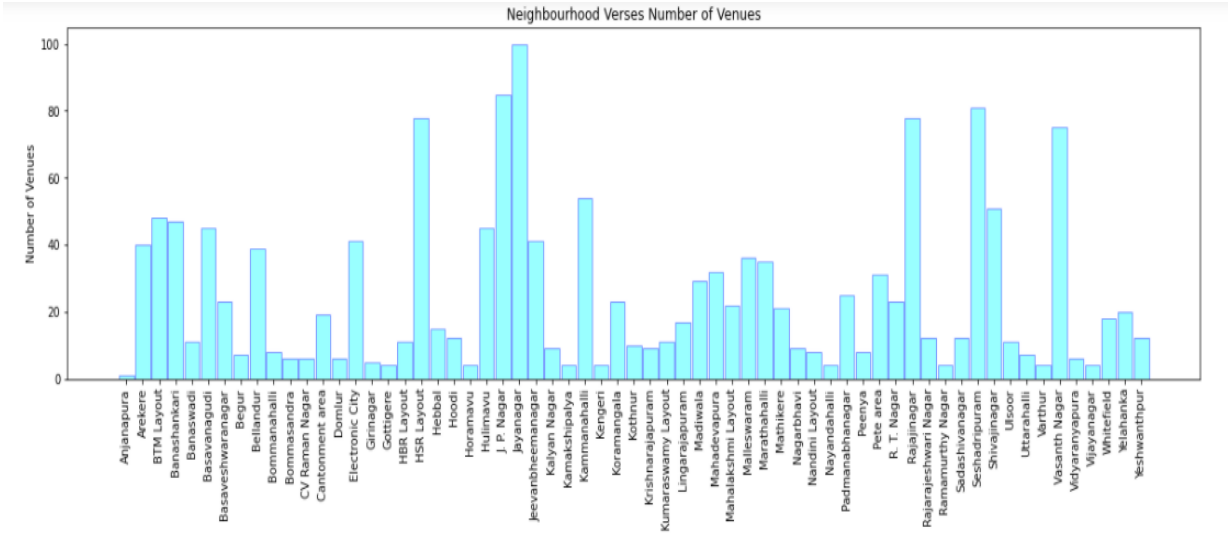
2.2 Data Extraction and Preparation Methodologies:

The Wikipedia page provided us with tables containing 65 different neighbourhoods. Using the BeautifulSoup library, we scrap data from those tables and convert them into a dataframe using the Pandas library. We keep an array "neighborhood" where we append the table data present in 0^{th} position if and only if its size is more than 0. We drop the last few rows which contain wrong data and null values. Next, we use the Geocoder library to generate the geographical coordinates of the different neighbourhoods present in the dataframe "df_neighborhood". The

coordinates are stored in a list called “cords” and later that list is appended with the dataframe.
Before, we visualize the different neighbourhoods of Bangalore, we need the exact geographical coordinates of the city. We use the Nominatim package of the geopy.geocoders library to get the latitude and longitude of Bangalore. Then we use the Folium library to generate the map of Bangalore and CircleMarker function was used to mark all the neighbourhoods on the map.



Now, the last step is to explore the neighbourhoods of Bangalore and extract venue data. Here, we use the FourSquare API which has one of the largest database of 105+ million places and is used by many developers. It provides with a large category of venues. However, here in India getting venues were a bit difficult.



As you can see in some places, we got very few categories of venues. With the developer account, we can make 500 calls per day and with each call a JSON file was returned. Lastly, the list of venues was appended with the neighbourhood dataframe. This new dataframe namely "df_venues" is our final dataset on which we will be doing our further analysis.

3. Data Analysis

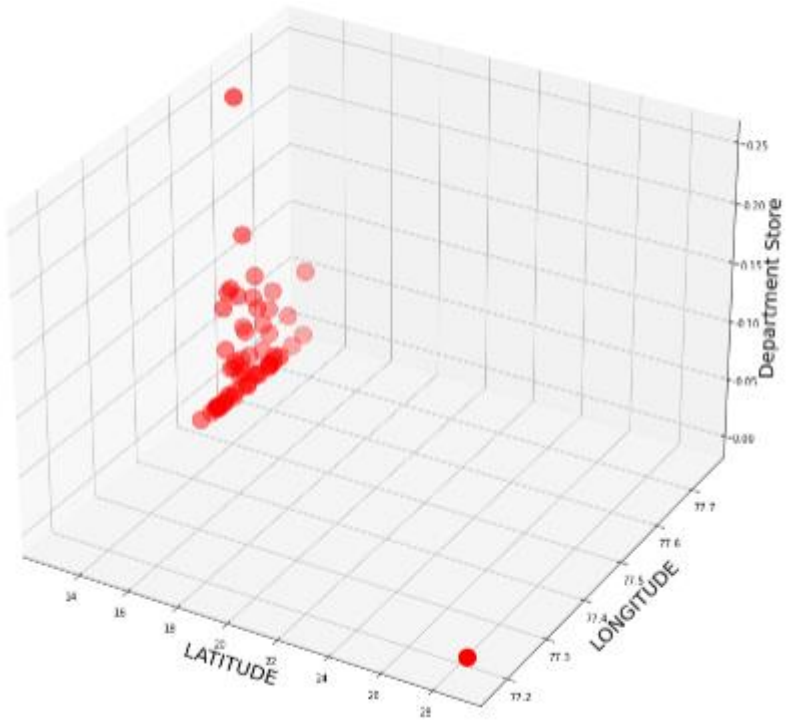
The data analysis part was challenging because we were dealing with nominal categorical variables. Categorical variables are of two types – Nominal and Ordinal. Ordinal categorical variables have numeric relationship among themselves. Hence, they can be encoded with numbers and ranked easily. For example, graduate, masters, PhD can be encoded as 1,2,3 where 1<2<3 because graduation degree is the most basic degree, master degree is higher than graduation and PhD is the highest degree. Thus, ranking ordinal categorical variables is easier.
But nominal categorical variables have no numeric relationship among themselves. Hence, they cannot be encoded with numbers and ranked. For example, names of different venue categories. Thus, we will use OneHot encoding method here which will

identify each venue category with a distinct binary number, also called encoding with dummy variables. Using this method, we will find the top 10 venues of every neighbourhood and store them in a dataframe.

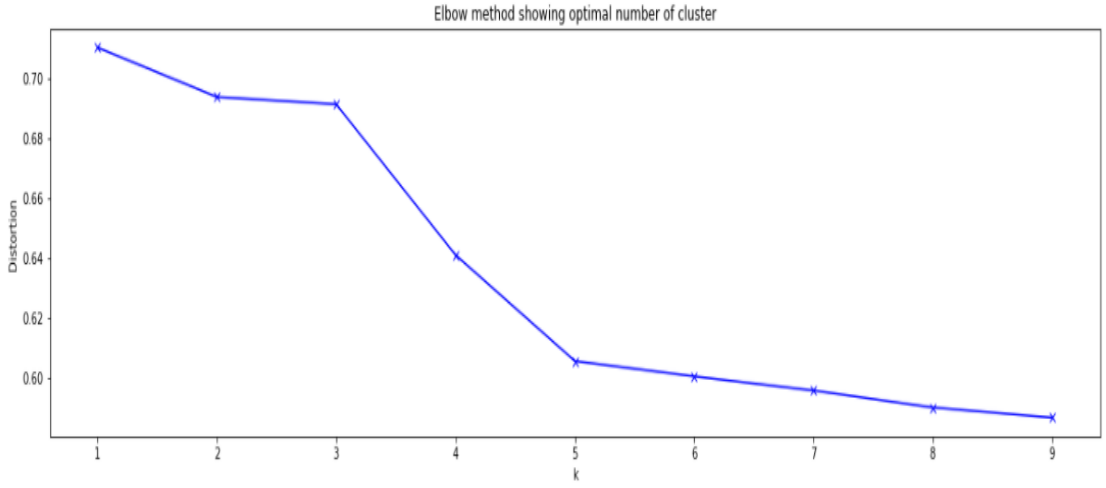
Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Shivajinagar	Indian Restaurant	Café	Clothing Store	Hotel	Tea Room	Women's Store	Candy Store	Donut Shop	Fast Food Restaurant	Salad Place
Ulsoor	Café	Indian Restaurant	BBQ Joint	Dessert Shop	Pub	Seafood Restaurant	Burger Joint	Bar	Coffee Shop	Asian Restaurant
Uttarahalli	Badminton Court	Hotel Bar	Convenience Store	Dhaba	Auto Garage	Rock Climbing Spot	Supermarket	Eastern European Restaurant	Fish Market	Fish & Chips Shop
Varthur	ATM	Bus Station	Supermarket	Lake	Dessert Shop	Dhaba	Flea Market	Fish Market	Fish & Chips Shop	Fast Food Restaurant
Vasanth Nagar	Indian Restaurant	Coffee Shop	Hotel	Chinese Restaurant	Café	Pub	Italian Restaurant	Lounge	Nightclub	Pizza Place
Vidyaranyapura	Indian Restaurant	Basketball Court	Pizza Place	Coffee Shop	Bus Station	Ice Cream Shop	Arcade	Event Space	Food & Drink Shop	Food
Vijayanagar	Pub	Brewery	Stadium	Rock Climbing Spot	Women's Store	Eastern European Restaurant	Fish Market	Fish & Chips Shop	Fast Food Restaurant	Farmers Market

4. Predictive Modelling:

For modelling purpose, we will used unsupervised machine learning techniques.

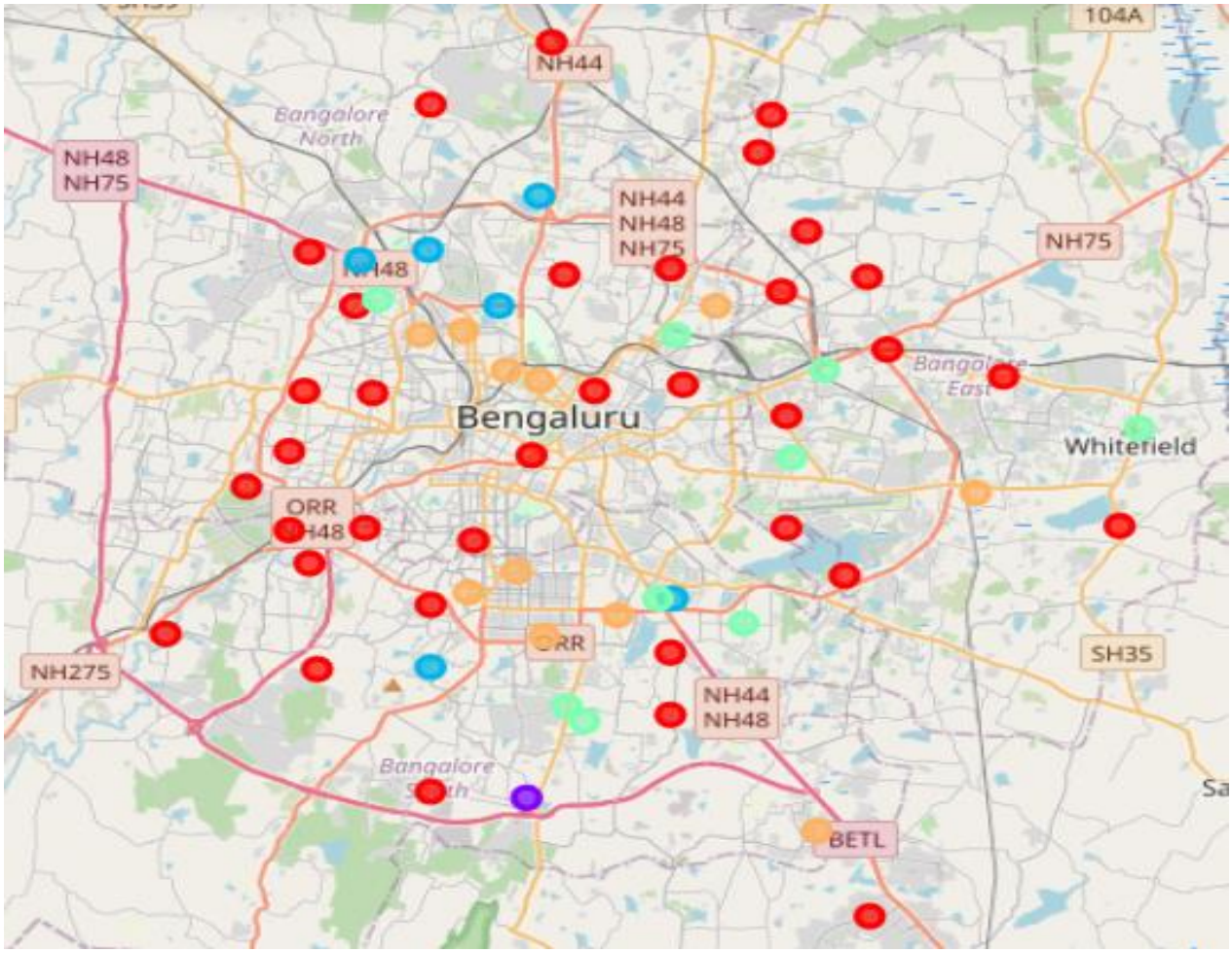


This graph shows how departmental stores are scattered over different latitudes and longitudes. They do not follow any particular pattern. Hence, it is definitely an unsupervised machine learning problem. Moreover, the data that is used here i.e. venue categories and neighbourhoods are unlabelled data and no classes for these data are predefined. Thus, we need to cluster the different neighbourhoods depending on their similar or dissimilar venues. A cluster is a group of objects (here neighbourhoods) that are similar to objects present in the same cluster and dissimilar to objects present in different cluster. We often cluster people based on their choices in customer segmentation. Here we will cluster neighbourhoods based on their venues.
We will use the partition based clustering algorithm, KMeans present in the sklearn.cluster package. KMeans divided the neighbourhoods in k subsets(clusters) depending on their similarities. It places k different centroids on for each cluster. More the number of clusters more accurate is our model. In most of the cases, we receive k non-overlapping clusters. But in our case, clusters are overlapping, this is because two neighbourhoods are clustered together based on their similarity in venues and not on distance between them. We have used the elbow method to determine the best value of k.



From this elbow method we can conclude that 5 clusters would give us better accuracy. Next, we have appended the cluster labels

with the dataframe. And finally, we have visualized the clusters using folium. We have assigned different colours to each cluster so that it can be identified better.



5. Result:

In the end, we have 5 different clusters.

- Cluster 0 (in red colour): Neighbourhoods with **no existence** of department stores. The top most common venues in these neighbourhoods are café, restaurants, hotel, parks, bus and stations, clothing stores etc. Neighbourhoods under this cluster are:

Anjanapura	Kengeri	Yelahanka	Krishnarajapuram

Nagarbhavi	Nandini Layout	Nayandahalli	Padmanabhanagar
Peenya	Kamakshipalya	Pete Area	Rajarajeshwari Nagar
Shivajinagar	Ulsoor	Uttarahalli	Ramamurthy Nagar
Vijayanagar	Varthur	RT Nagar	Vidyaranyapura
Kalyan Nagar	Kothnur	Horamavu	Cantonment Area
Basavanagudi	Girinagar	Banaswadi	Basaveshwaranagar
Bommasandra	Hoodi	HBR Layout	CV Raman Nagar
Bellandur	Domlur	Begur	Bommanahalli

- Cluster 1 (in purple colour): Neighbourhood with **low existence** of department store. Neighbourhood under this cluster is:

Gottigere

- Cluster 2 (in sea green colour): Neighbourhoods with **high existence** of department store. Neighbourhood under this cluster is:

Mathikere	Yeshwanthpur	Koramangala
Hebbal	Sadashivanagar	Kumaraswamy Layout

- Cluster 3 (in light green colour): Neighbourhoods with **moderate existence** of department store. Neighbourhoods under this cluster are:

Arekere	Whitefield	Hulimavu
Mahalakshmi Layout	Mahadevapura	Madiwala
Lingarajapuram	HSR Layout	Jeevanbheemanagar

- Cluster 4 (in orange colour): Neighbourhoods with **low existence** of department store. Neighbourhoods under this cluster are:

Malleswaram	JP Nagar	Jayanagar
BTM Layout	Vasanth Nagar	Kammanahalli
Marathahalli	Seshadripuram	Rajajinagar
Bnashankari	Electronic City	

6. Precautions:

In this project, the density of civilians in a neighbourhood has not been taken into consideration. I believe that this is an important factor which can affect the sales of the store. Some areas have no existence of department stores which may (or may not) be due to no existence of civilians in that area. Hence, recommending to open store in such an area is risky. The financial status of the civilians can also affect the rate of sales. This project has to be worked upon more in order to remove such risks.

7. Discussions:

On reading the result section, it is clearly visible that a very high number of departmental stores are present in neighbourhoods of cluster 2. One will have to face high marketing competition in this cluster. Hence **neighbourhoods under cluster 2 are not recommended**. However, the high density of departmental stores confirms that this area has high density of civilians as well. Neighbourhoods under cluster 3 has moderate presence of departmental stores. This suggests lesser marketing competition as compared to cluster 3 and also confirms the presence of civilians. Hence **neighbourhoods under cluster 3 can be recommended**. Neighbourhoods under cluster 4 and cluster 1 has low presence of departmental stores. This suggests very less marketing competition

as compared to cluster 2 and 3 and also suggests that fair number of civilians must be present. If we look into the top venues of these clusters, we understand that it is also a hotspot of eating places. Other than the residents many other people must be visiting this place. Hence **neighbourhoods under cluster 4 and 1 are highly recommended**.
Neighbourhoods under cluster 0 has no existence of departmental stores. This confirms no competition as compared to all other clusters but the presence of civilians is very uncertain. However, on looking into the top most common places of cluster 0, we get an idea that it more of a hotspot for eating and amusement. Hence **it is difficult to say anything about cluster 0**.

8. Conclusion:

The answer to the business question raised by this project is: For opening new department store recommended neighbourhoods are clustered under 1,3,4. Target audiences will find this project helpful. However, the precautions must be kept in mind.