Chapter -1

Introduction/Business Problem

1. Introduction

Bangalore, is a very rapidly growing city. It is a IT hub and rightly referred to as a Silicon Valley of India. With its rapid growth attracting many companies to open its center here and also many start-ups opening up, it is attracting a lot of people all of the world to come here looking for jobs. This has led to many opportunities for business like restaurants, café, transports, homestays/hostels, medical, hypermarkets and even malls to quickly capitalize this moment to capture and establish the hold in the market by strategically positioning them in accordance the requirements of the locality.

2. Business Problem

Now, although we understand there is an opportunity for business to gain huge profits by coming in Bangalore but we still need to find the most appropriate positions/locations for their establishment to get maximum profit. The problems can be classified in following points:-

- Understanding if a particular kind of business is in demand from the city
- Finding best location for the business

This can only be done by totally analyzing the whole location of Bangalore, from its every locality to its neighborhood and for this proper data are also required which is also one of the big problems.

Here, we will try to search for the locality data from the web and scrape it. Then, after proper cleaning and filtration of useful data we would use FourSquare API to look over neighborhood to find the businesses and type of businesses present there. After that we could analyze the data and cluster the similar localities together and then present/suggest the location best suited for the business.

Chapter -2

Data Acquisition and Cleaning

1. Data Acquisition

First, I had to find the data for all the localities in Bangalore. For that I had to search through google and I reached this site – click here to visit the site. The source is from a well reputed news channel in India so I thought to continue with it.

I scrapped the data from their site into a csv file for its further use. At first look itself I could understand that it had some inconsistencies like null data, similar data also sub location of the locality listed as a separate locality. So, these were the first things in my mind to sort.

Before starting data cleaning process I even added Latitude and Longitude columns in the dataframe with the values added into them using geopy package of Python.

	Office	Taluk	District	State	Pincode	Latitude	Longitude
0	A F Station Yelahanka	Bangalore North	Bangalore	KARNATAKA	560063	NaN	NaN
1	Agram	Bangalore South	Bangalore	KARNATAKA	560007	NaN	NaN
2	Air Force Hospital	Bangalore North	Bangalore	KARNATAKA	560007	12.964027	77.627500
3	Amruthahalli	Bangalore North	Bangalore	KARNATAKA	560092	13.066513	77.596624
4	Anandnagar Bangalore	Bangalore North	Bangalore	KARNATAKA	560024	13.033377	77.589523

Table 2.1: Look of the data after 1st stage of Data Acquisition

2. Data Cleaning

I started with searching for duplicate data and null values. First, I removed all the null valued areas as they were mostly the inner neighborhood of the localities already present in the dataset.

Secondly, I combined the localities which had their same latitude and longitude into one row.

Third, I removed the unnecessary columns i.e 'Taluk' and 'State' and renamed the column 'Office' as 'Locality' which is a better suited name for the column.

After this I visualized the data using folium package on the map and found many outliers i.e the locations which are far away from the Bangalore.

To get rid of those locations as well as outskirt areas of Bangalore, I removed all those locations which were more than 35kms apart from the center of Bangalore using 'Haversine distance' method so that our analysis gets concentrated in urban areas of Bangalore which offers much better scope of business.

After this a total of 142 localities were left which are now ready to be used for further work.

At a later stage after using FourSquare API , I even found that some localities had no venue within specified radius as a result I had to remove them too as they couldn't be clustered in any group.

3. FourSquare API

Now, after the data cleaning phase we were left with 142 localities. I used the FourSquare API and ran through all those localities to find the venues present inside the 500 meters radius from their location and to store and maintain a different dataframe for it. It had total of 1032 venues in it.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1027	Kannur	13.103799	77.601617	Udupi Grand, Kogilu Cross	13.103455	77.600627	Fast Food Restaurant
1028	Kannur	13.103799	77.601617	Yelahanka lake	13.103729	77.602205	Lake
1029	Kannur	13.103799	77.601617	Gobi Adda	13.103732	77.602401	Food Truck
1030	Madanayakanahalli	13.059247	77.461511	Khimaj Caterers	13.060353	77.460115	Restaurant
1031	Neriga	12.911766	77.776525	Shristhi Village	12.913261	77.773636	Resort

Table 2.2: Results returned from using FourSquare API on the localities

Chapter -3

Methodology

1. Data Analysis

After gathering all the data including the venues near every neighborhoods, I tried to find the total count of venues that each locality had. I found the following result (table 3.1):

:

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Adugodi	5	5	5	5	5	5
Air Force Hospital	2	2	2	2	2	2
Amruthahalli	4	4	4	4	4	4
Anandnagar Bangalore	6	6	6	6	6	6
Anekal	2	2	2	2	2	2
Anjanapura	1	1	1	1	1	1
Arabic College	1	1	1	1	1	1
Attur	1	1	1	1	1	1
Austin Town	4	4	4	4	4	4
Bagalgunte	4	4	4	4	4	4
Bagalur Bangalore	1	1	1	1	1	1
Banashankari	5	5	5	5	5	5
Banashankari III Stage	6	6	6	6	6	6
Banaswadi	5	5	5	5	5	5
Bangalore Corporation Building	12	12	12	12	12	12
Bangalore International Airport	10	10	10	10	10	10

Table 3.1: A sample of locality with their total count of venues

From this I confirmed that most of the profile areas and the locations with the IT park and most venues nearer to them like Jayanagar, HSR , Indiranagar, Electronic City etc.

Now, that we know which locations have majority of venues now I tried to find out what kind of venue is most common in Bangalore. For this I grouped the dataframe on venue categories and sorted them in Descending order. The following result was observed (Table 3.2).

From this I concluded that Indian Restaurants, Café and Fast Food Restaurants are most common among Bangalore people as it had most shops for it with Indian Restaurants being by far the number one venue in Bangalore with a total of 174 outlets.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category						
Indian Restaurant	174	174	174	174	174	174
Café	70	70	70	70	70	70
Fast Food Restaurant	38	38	38	38	38	38
Coffee Shop	34	34	34	34	34	34
Bakery	33	33	33	33	33	33
Department Store	28	28	28	28	28	28
Hotel	26	26	26	26	26	26
Pizza Place	25	25	25	25	25	25
Chinese Restaurant	22	22	22	22	22	22
Ice Cream Shop	21	21	21	21	21	21
Juice Bar	17	17	17	17	17	17
Dessert Shop	17	17	17	17	17	17
Restaurant	17	17	17	17	17	17
Snack Place	15	15	15	15	15	15

Table 3.2: A sample of results from the total no. of different venues present in Bangalore

After this I performed the one hot encoding for different venues present on basis of their category, then grouped each neighborhood and mentioned each category with its frequency of occurrence in the corresponding neighborhood (Table 3.3).

	Neighborhood	ATM	Accessories Store	Airport Lounge	American Restaurant	Andhra Restaurant	Antique Shop	Arts & Crafts Store	Arts & Entertainment	A R
0	Adugodi	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.
1	Air Force Hospital	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.
2	Amruthahalli	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.
3	Anandnagar Bangalore	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.
4	Anekal	0.500000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.
5	Anjanapura	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.
6	Arabic College	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.
7	Attur	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.
8	Austin Town	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.
9	Bagalgunte	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.
10	Bagalur Bangalore	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.

Table 3.3: A sample of neighborhood with the frequency of various kinds of venues in it

Now that I had the frequency of the venues I could easily find out and arrange the venues in descending order of their occurrence in the given locality (Table 3.4).

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Adugodi	Ice Cream Shop	Athletics & Sports	Bus Station	Café	Playground
1	Air Force Hospital	Multiplex	Casino	Yoga Studio	Dry Cleaner	Flower Shop
2	Amruthahalli	Convenience Store	Shoe Store	Indian Restaurant	Department Store	Flower Shop
3	Anandnagar Bangalore	Hotel	Pharmacy	Flea Market	Garden Center	Motorcycle Shop
4	Anekal	ATM	Camera Store	Cosmetics Shop	Dry Cleaner	Flower Shop

Table 3.4: Sample of neighborhoods with their most common venues

2. Clustering model

To analyze the localities of Bangalore and to understand the trends of it, the best way would be to use a Clustering model to cluster the similar localities on the basis of the frequency of the venues.

Here, the clustering model which I used was K-Means clustering. For this I ran the K-means algorithm on the same dataset for many times using different values of K and came to a decision that the best suited K value for this problem is 5 as further increasing the number didn't yield in any insightful results.

After applying the algorithm I added a column named 'Cluster Labels' which would contain the cluster group number of the given locality/neighborhood , to the previous dataframe. The resultant dataframe looked like this. (Table 3.5)

	Locality	District	Pincode	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th I Com Venu
118	Hunasamaranahalli	Bangalore	562157	13.144010	77.619124	0.0	Indian Restaurant	Coffee Shop	Lake	Donut Shop	Flow Shop
119	Kadabagere	Bangalore	562130	12.988910	77.449245	0.0	Indian Restaurant	Yoga Studio	Dry Cleaner	Food	Flow Shop
120	Kannur	Bangalore	562149	13.103799	77.601617	2.0	Food Truck	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Lake	Dry Clea
121	Neriga	Bangalore	562125	12.911766	77.776525	1.0	Resort	Yoga Studio	Donut Shop	Flower Shop	Flea Mark
122	Sarjapura	Bangalore	562125	12.860087	77.786135	1.0	Snack Place	Yoga Studio	Dry Cleaner	Flower Shop	Flea Mark

Table 3.5: Sample of Locality with their cluster labels

Chapter -4

Results and Discussion

1. Results

The result of the K-Means Clustering algorithm was stored in the dataframe as another column named as ‘Cluster Labels’ and then was visualized using the Folium package to see the actual distribution of the clusters over the map of Bangalore (Fig. 4.1).

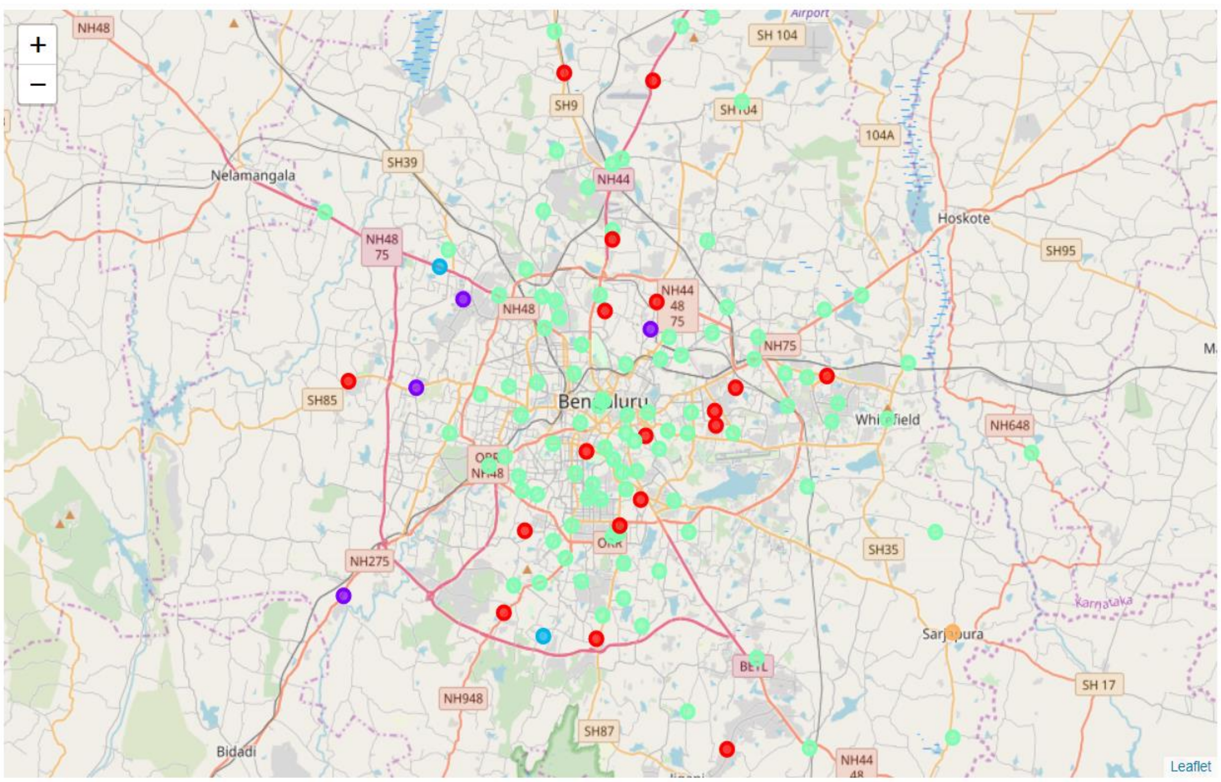

Figure 4.1: Showing distribution of localities in different Clusters.

After this each cluster was analyzed individually. First cluster, the cluster value equal to 0 yielded this result (Table 4.1).

	Locality	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
11	Byatarayanapura	0.0	Indian Restaurant	Wine Shop	Cosmetics Shop	Andhra Restaurant	Department Store
18	Hoodi	0.0	Indian Restaurant	Yoga Studio	Furniture / Home Store	Dry Cleaner	Flower Shop
32	NAL	0.0	Indian Restaurant	Café	Chinese Restaurant	Dessert Shop	Kerala Restaurant
40	Singanayakanahalli	0.0	Indian Restaurant	Café	Yoga Studio	Food	Flower Shop
68	Gottigere	0.0	Indian Restaurant	Department Store	Dessert Shop	Yoga Studio	Dumpling Restaurant
81	Mavalli	0.0	Indian Restaurant	Hotel	Food Truck	Snack Place	Restaurant
88	Thalaghattapura	0.0	Indian Restaurant	Yoga Studio	Dry Cleaner	Food	Flower Shop
103	Msrit	0.0	Indian Restaurant	Fast Food Restaurant	Bus Station	Yoga Studio	Dumpling Restaurant
118	Hunasamaranahalli	0.0	Indian Restaurant	Coffee Shop	Lake	Donut Shop	Flower Shop
119	Kadabagere	0.0	Indian Restaurant	Yoga Studio	Dry Cleaner	Food	Flower Shop

Table 4.1: Localities clustered in first cluster

Second cluster, the cluster value equal to 1 yielded this result (Table 4.2).

	Locality	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
3	Arabic College	1.0	Coffee Shop	Yoga Studio	Dry Cleaner	Food	Flower Shop
4	Attur	1.0	Juice Bar	Yoga Studio	Dry Cleaner	Food	Flower Shop
8	Bellandur,Ashoknagar Bangalore,Shanthinagar	1.0	Hotel	Vineyard	Capitol Building	Park	Donut Shop
9	Benson Town	1.0	Candy Store	Café	Dessert Shop	Pakistani Restaurant	Bar
10	Bhattarahalli	1.0	Ice Cream Shop	Hotel	Japanese Restaurant	Café	Dry Cleaner
13	CMP Centre & School	1.0	Hotel	Soccer Field	Café	Sporting Goods Shop	Health Food Store
20	Indiranagar Bangalore	1.0	Lounge	Pub	Café	Ice Cream Shop	Restaurant
22	Kadugodi	1.0	Playground	Platform	Indian Restaurant	Park	Train Station

Table 4.2: Sample of Localities clustered in second cluster

Third cluster, the cluster value equal to 2 yielded this result (Table 4.3).

	Locality	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Air Force Hospital	2.0	Multiplex	Casino	Yoga Studio	Dry Cleaner	Flower Shop
25	Kundalahalli	2.0	Fast Food Restaurant	Hotel	Sports Bar	Dessert Shop	Bakery
29	Maruthi Sevanagar	2.0	Fast Food Restaurant	Clothing Store	Multiplex	Department Store	Shopping Mall
58	Bolare	2.0	Fast Food Restaurant	Yoga Studio	Dry Cleaner	Food	Flower Shop
59	Chamrajpet Bangalore	2.0	Fast Food Restaurant	Indian Restaurant	Vegetarian / Vegan Restaurant	Shop & Service	Yoga Studio
74	Kathriguppe	2.0	Breakfast Spot	Fast Food Restaurant	Smoke Shop	Bus Station	Yoga Studio
84	Nayandahalli	2.0	IT Services	Accessories Store	Metro Station	Fast Food Restaurant	Train Station
115	Chikkajala	2.0	Hotel	Bike Shop	Fast Food Restaurant	Bar	Café
120	Kannur	2.0	Food Truck	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Lake	Dry Cleaner

Table 4.3: Localities clustered in third cluster

Fourth cluster, the cluster value equal to 3 yielded this result (Table 4.4).

	Locality	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Amruthahalli	3.0	Convenience Store	Shoe Store	Indian Restaurant	Department Store	Flower Shop
2	Anandnagar Bangalore	3.0	Hotel	Pharmacy	Flea Market	Garden Center	Motorcycle Shop
5	Austin Town	3.0	Tibetan Restaurant	Indian Restaurant	Italian Restaurant	Bakery	Dry Cleaner
6	Banaswadi	3.0	Vegetarian / Vegan Restaurant	Indian Restaurant	Bakery	BBQ Joint	Yoga Studio
7	Bangalore International Airport	3.0	Coffee Shop	Airport Lounge	Indian Restaurant	Vegetarian / Vegan Restaurant	Italian Restaurant
12	C.V.Raman Nagar	3.0	Indian Restaurant	Pizza Place	Shop & Service	Park	Department Store
14	Doddanekkundi	3.0	Indian Restaurant	Food	Fried Chicken Joint	Clothing Store	Restaurant
15	Domlur	3.0	Indian Restaurant	Café	Food & Drink Shop	Pub	Sports Bar
16	Dr. Ambedkar Veedhi	3.0	Coffee Shop	Indian Restaurant	Hotel	Capitol Building	Dry Cleaner

Table 4.4: Sample of Localities clustered in Fourth cluster

Fifth cluster, the cluster value equal to 4 yielded this result (Table 4.5).

	Locality	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
21	Jakkur	4.0	ATM	Dry Cleaner	Food	Flower Shop	Flea Market
41	Venkateshapura	4.0	ATM	Indian Restaurant	Motorcycle Shop	Cosmetics Shop	Dumpling Restaurant
63	Deepanjalinagar	4.0	ATM	Metro Station	Travel & Transport	Dry Cleaner	Flower Shop
71	Hulimavu,Hulimavu	4.0	ATM	Department Store	Lake	Badminton Court	Dumpling Restaurant
78	Kumbalagodu	4.0	ATM	Café	Dhaba	Flower Shop	Flea Market
96	Jalahalli West	4.0	ATM	Dry Cleaner	Food	Flower Shop	Flea Market
100	Magadi Road	4.0	ATM	Photography Studio	Restaurant	Donut Shop	Flea Market
112	Anekal	4.0	ATM	Camera Store	Cosmetics Shop	Dry Cleaner	Flower Shop

Table 4.5: Localities clustered in fifth cluster

2. Discussion

A total of 5 cluster was obtained each unique in its own way. The first cluster (Table 4.1) is a hub of Indian Restaurants, most of them having some yoga centers and some flower shops.

The second cluster (Table 4.2) has variety of things and places with non-traditional Indian restaurants and many fitness places.

The third cluster (Table 4.3) are places having chain of Fast food Restaurants and multiple entertainment places like multiplex, casino, sports bar etc.

The fourth cluster (Table 4.4) are food hubs of the city with enough restaurants and cafes present.

The fifth cluster (Table 4.5) are the locations with majority of ATMs of city and many flea shops.

The following recommendations could be provided according to the results:

- First Cluster- Non Indian restaurants or fast food restaurants can be opened there.
- Second Cluster- It could be mixed with some traditional Indian Restaurants.
- Fifth Cluster- It has a lot of ATMs so it could be used as a reference before installing new ATMs.

Chapter -5

Conclusion

This project of 'Bangalore Analysis' brought some insights on the distribution of various places with its pattern. Here, the venues were not just limited to food but also parks, ATMs, shops etc. Even with this general/ high-level analysis I was able to recommend some business.

For a better analysis i.e specific to a particular locality or specific to particular type of venues we could build a similar model by confining our datasets to that locality or venue.

The Foursquare API used in the analysis of the localities of Bangalore to find out venues was good for the high level analysis but from my experience I would not prefer it for any specific analysis as I feel it lacks many venues and doesn't give the correct insight venues present in the locality (It's Specifically for the Bangalore city not for any other place). Rather I would be using and would propose the use of Google map APIs for such analysis in future.

Lastly, I would like to conclude this report by highlighting its major points:

- It included all the localities of Bangalore.
- To protect my analysis from all the outliers I kept only the urban areas by keeping the localities which were within 35kms from the center.
- All the venues were extracted near every locality using FourSquare API.
- At last, neighborhoods were clustered together using K-Means Clustering.