PREDICTING THE BEST LOCALITY TO LIVE & START A RESTAURANT BUSINESS IN BANGALORE

Abdullah Shahab

28 July'20

1. Introduction

1.1 Business Problem

Bangalore is one of the fastest growing metropolitan in India and is also know as the Silicon Valley of India. It has the second highest number of start-ups in India after Delhi-NCR region. Because of this growth many people are planning to move to Bangalore for new business opportunities and accommodation.

If someone is looking to start a restaurant business in Bangalore, that person might look to start the restaurant in a region where mostly people in the city go for food, basically which has most of the restaurants.

Also, for accommodation, people generally look for a place which is near to market, ATMs, theatre, metro station, park etc.

Machine learning clustering algorithm can help such people in identifying the region as per their need by dividing the regions of Bangalore into various clusters.

1.2 Who would be interested?

As, already mentioned, I am trying to help the people who are looking to start a restaurant business in Bangalore or who are looking to move to Bangalore just for accommodation. ML clustering algorithm can help these people in selecting the region to start the restaurant or region to settle down.

2. Data to solve the problem

To identify the various regions of Bangalore and explore the venues in various regions of Bangalore, I have used the GeoJson file of the constituencies of Bangalore. The GeoJson file contains the name of the various constituencies of Bangalore and the wards that are present under each constituency. It also contains the latitude and longitude information of each ward.

3. Source of data

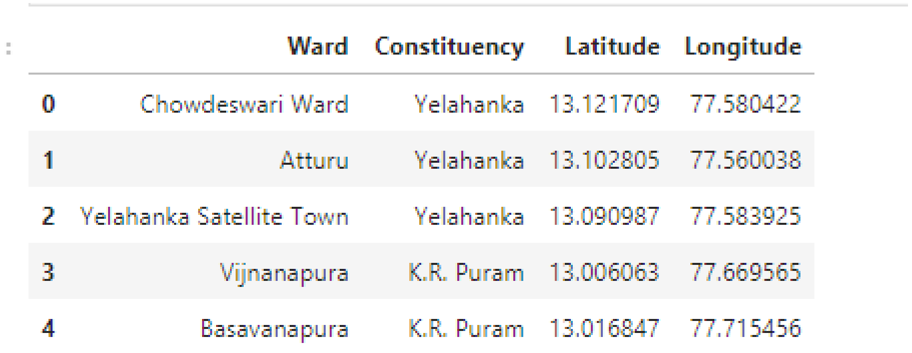
I have downloaded the data from
http://projects.datameet.org/Municipal_Spatial_Data/bangalore/

Actual GeoJson
https://raw.githubusercontent.com/datameet/Municipal_Spatial_Data/master/Bangalore/BBMP.GeoJSON

4. Methodology

To capture the data, I researched various websites and finally found the required data on DataMeet website. I downloaded the data and uploaded it into the notebook only, to make the retrieval easy.
After this I fetched the content of json file and stored it in a variable using json.load() function. After analysing the first node of json file for its various attributes, I converted the json file into dataframe. In this process I kept only those attributes which were required and dropped the rest of the attributes of json. The attributes kept are Ward, Constituency, Latitude, and longitude.

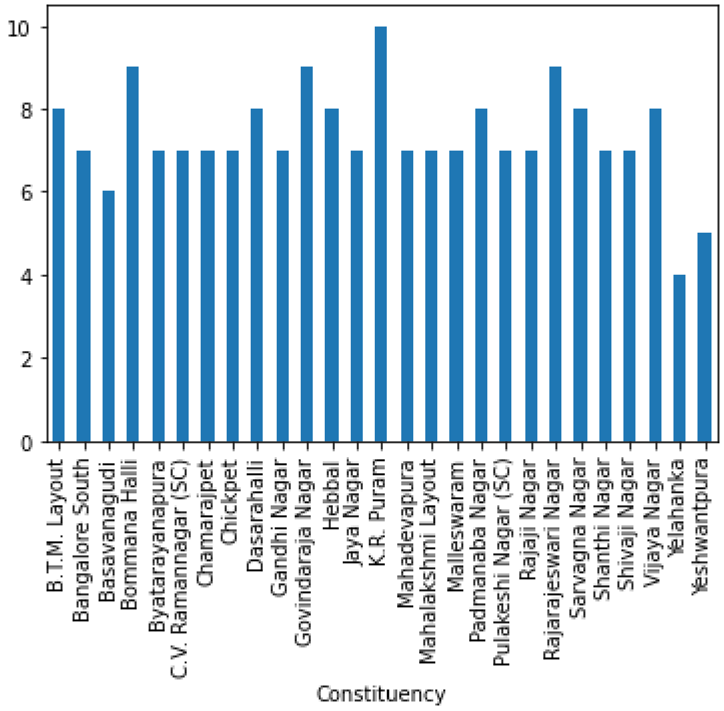
Master data set after converting json to dataframe and dropping unnecessary attributes:



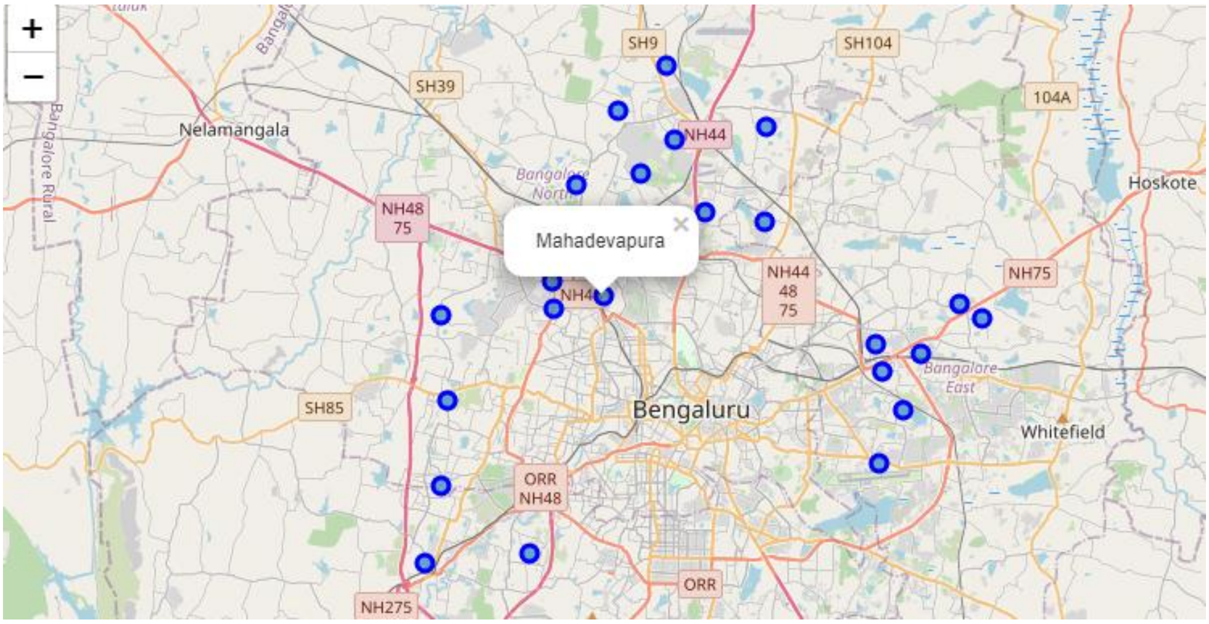
	Ward	Constituency	Latitude	Longitude
0	Chowdeswari Ward	Yelahanka	13.121709	77.580422
1	Atturu	Yelahanka	13.102805	77.560038
2	Yelahanka Satellite Town	Yelahanka	13.090987	77.583925
3	Vijnanapura	K.R. Puram	13.006063	77.669565
4	Basavanapura	K.R. Puram	13.016847	77.715456

This is only a small portion. There are total 27 constituencies and 198 wards.

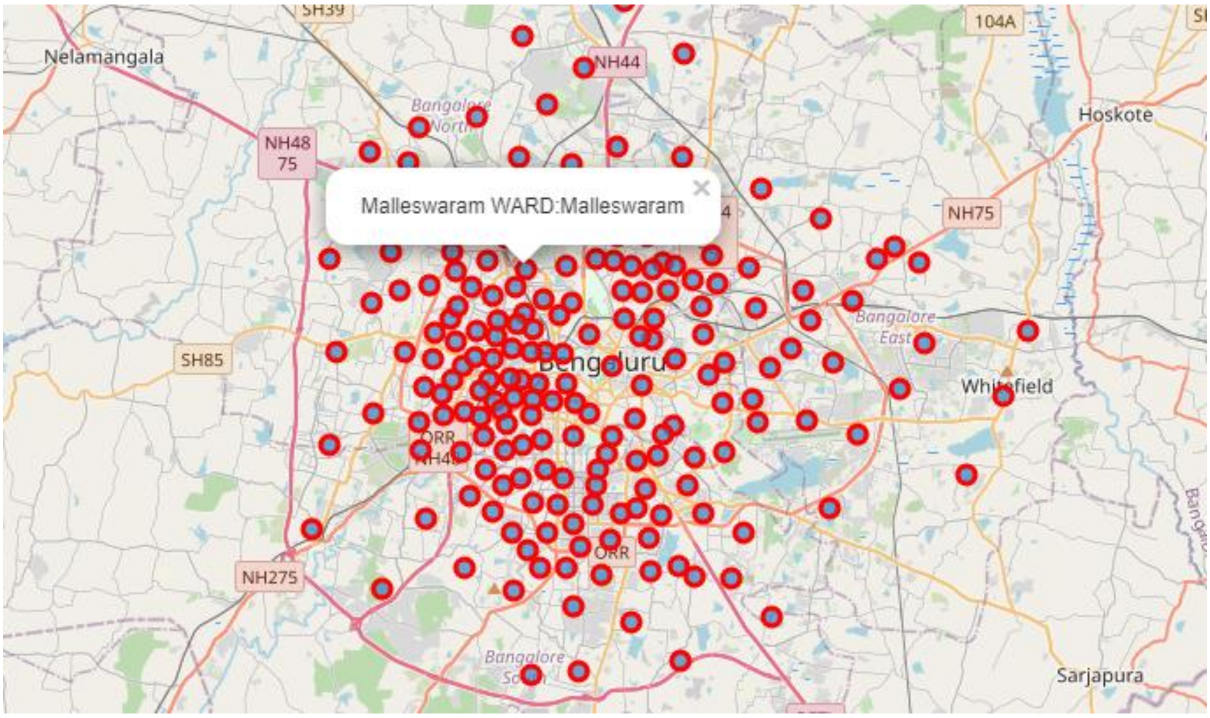
Bar chart to show number of wards under each constituency.



After that, for better visualization, I have tried to plot the constituencies on map of Bangalore. This circles are clickable, and popup will appear with the name of constituency on clicking. To create this visualization, I have used **folium library**. Map is plotted with the help of map () function and for circle markers, I have used the CircleMarker () function.



Ward under each constituency



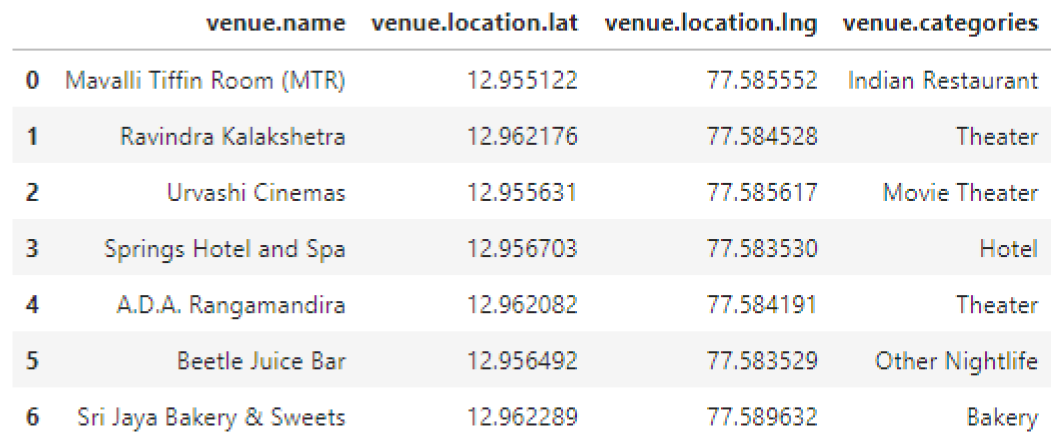
Now comes the next major part. To find and explore the venues near a ward, I have used the **Foursquare API.** Firstly, I tried to analyse single ward and took the *Sudham Nagar ward.*

To make call to the foursquare API, a URL is created with fixed format as given below: https://api.foursquare.com/v2/venues/explore?&client_id= xx&client_secret= xx&v= xx&ll= xx&radius=xx&limit=xx

A request call is made using this URL and then json data is fetched from the foursquare API. This json data needs to be converted to dataframe for better handling. This is done using **json_normalize()** function of pandas library.

The data frame receive contains lot of unnecessary attributes and we need to drop all those attributes. We only need Venue name, latitude, longitude, and Category. Venue name, latitude and longitude could be fetched directly from dataframe, but venue category is in nested form and I then used get_category() function to fetch the category.

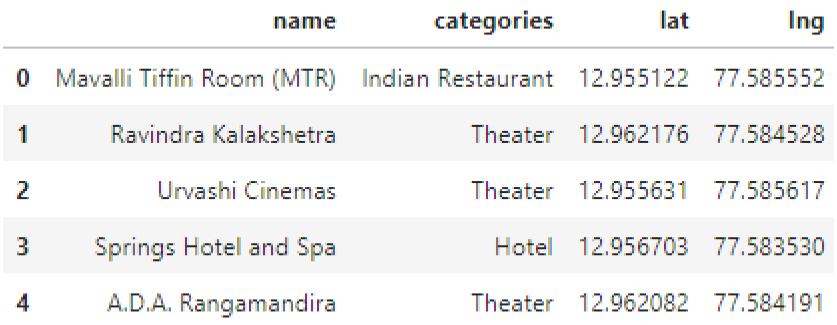
Venue table of Sudham Nagar Ward. There are total 7 venues in this ward.



	venue.name	venue.location.lat	venue.location.lng	venue.categories
0	Mavalli Tiffin Room (MTR)	12.955122	77.585552	Indian Restaurant
1	Ravindra Kalakshetra	12.962176	77.584528	Theater
2	Urvashi Cinemas	12.955631	77.585617	Movie Theater
3	Springs Hotel and Spa	12.956703	77.583530	Hotel
4	A.D.A. Rangamandira	12.962082	77.584191	Theater
5	Beetle Juice Bar	12.956492	77.583529	Other Nightlife
6	Sri Jaya Bakery & Sweets	12.962289	77.589632	Bakery

After analysing the categories, I noticed “Movie Theatre” and “Theatre” are same category and information is **redundant**. So, I renamed “Movie Theatre” as “Theatre”.

After this I **renamed** the columns to make more sense. Final table of *Sudham Nagar*.



	name	categories	lat	lng
0	Mavalli Tiffin Room (MTR)	Indian Restaurant	12.955122	77.585552
1	Ravindra Kalakshetra	Theater	12.962176	77.584528
2	Urvashi Cinemas	Theater	12.955631	77.585617
3	Springs Hotel and Spa	Hotel	12.956703	77.583530
4	A.D.A. Rangamandira	Theater	12.962082	77.584191

Analysing all the venues of all the wards.

Now I will try to delve deeper and try to explore the venues near all the wards of all constituencies.

For this, firstly I have created a function which returns the URL corresponding to each ward. These URLs will be used to make call to foursquare API

Secondly, I have defined a separate function to get all the nearby venues of each ward which is passed into the function.

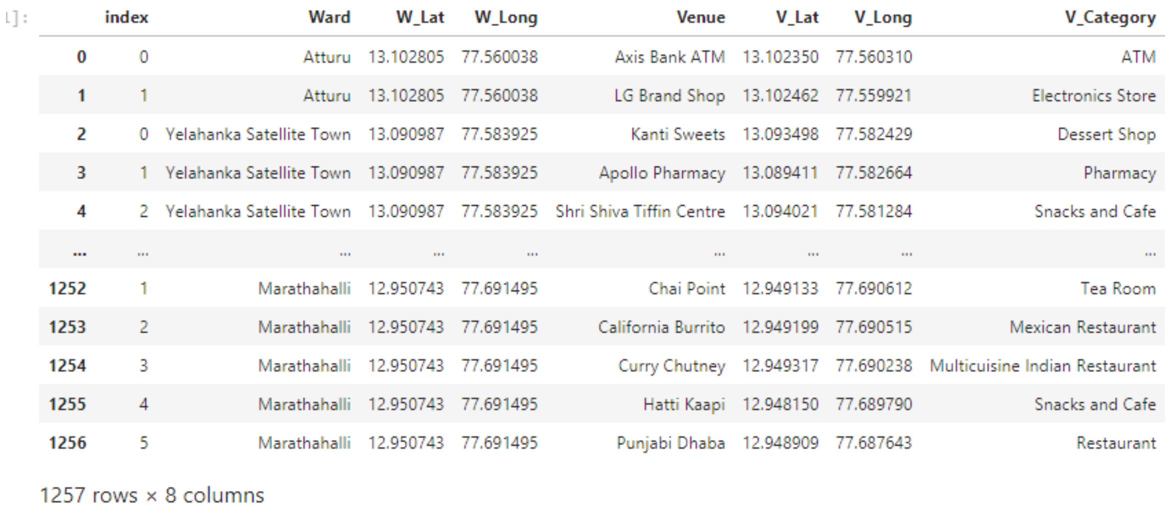
Finally, I have created the data frame of the venues of all wards by making calls to above two functions.

Repeating the process mentioned above, I have cleaned the data frame and created final master data frame of all the venues in each ward.

Total categories of venues before removing redundant categories are 162.

After this, to remove the **redundancy**, I have merged different types of restaurants into one "*restaurant*" category and various quick bite corners into "*snack and coffee*" category.

Final venue table of all wards:



	index	Ward	W_Lat	W_Long	Venue	V_Lat	V_Long	V_Category
0	0	Atturu	13.102805	77.560038	Axis Bank ATM	13.102350	77.560310	ATM
1	1	Atturu	13.102805	77.560038	LG Brand Shop	13.102462	77.559921	Electronics Store
2	0	Yelahanka Satellite Town	13.090987	77.583925	Kanti Sweets	13.093498	77.582429	Dessert Shop
3	1	Yelahanka Satellite Town	13.090987	77.583925	Apollo Pharmacy	13.089411	77.582664	Pharmacy
4	2	Yelahanka Satellite Town	13.090987	77.583925	Shri Shiva Tiffin Centre	13.094021	77.581284	Snacks and Cafe
...	...	...	...	...	...	...	...	...
1252	1	Marathahalli	12.950743	77.691495	Chai Point	12.949133	77.690612	Tea Room
1253	2	Marathahalli	12.950743	77.691495	California Burrito	12.949199	77.690515	Mexican Restaurant
1254	3	Marathahalli	12.950743	77.691495	Curry Chutney	12.949317	77.690238	Multicuisine Indian Restaurant
1255	4	Marathahalli	12.950743	77.691495	Hatti Kaapi	12.948150	77.689790	Snacks and Cafe
1256	5	Marathahalli	12.950743	77.691495	Punjabi Dhaba	12.948909	77.687643	Restaurant

1257 rows × 8 columns

Finding top 5 venues category in each Ward

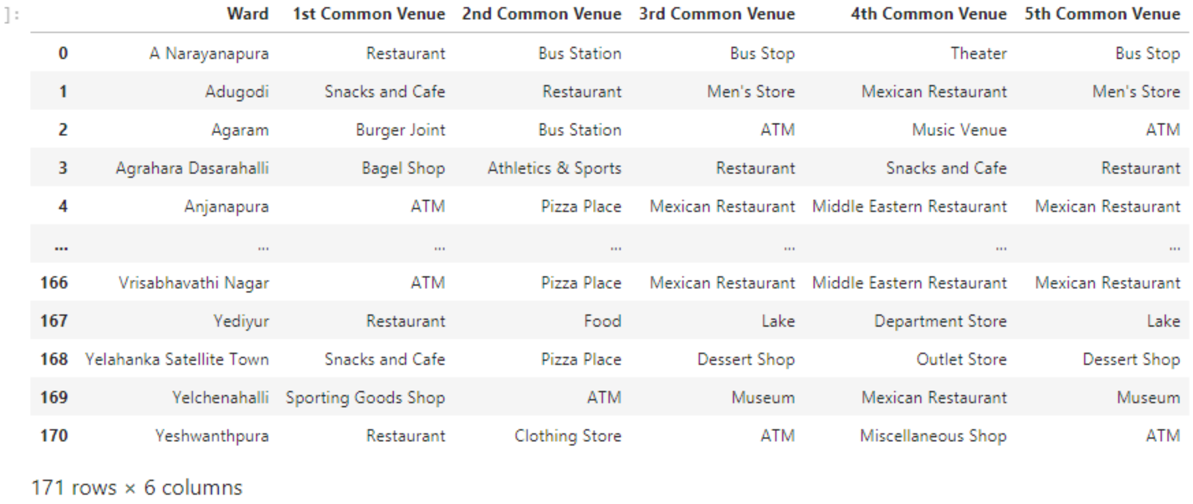
To analyse the categorical data, I have converted the categories of each venue in binary form i.e. now the categories are attributes of each venue. If a venue' category is ATM, then only ATM attribute will have value equal to 1 and other attributes will be 0. I have used get_dummies() function for this work.

Since I am going to use only ward/location for type of category, I have dropped the venue details in below step.

After this, I grouped the dummy table by "*Ward*" and took the mean of each venue for each ward.

Now, to find the top 5 venues in each category, I have created a function that will arrange the dummy table in descending order for each ward one by one. In this way, we will be able to fetch top 5 venues of each ward in each iteration and add this data in a new data frame.

Final table with top 5 venues category in each ward



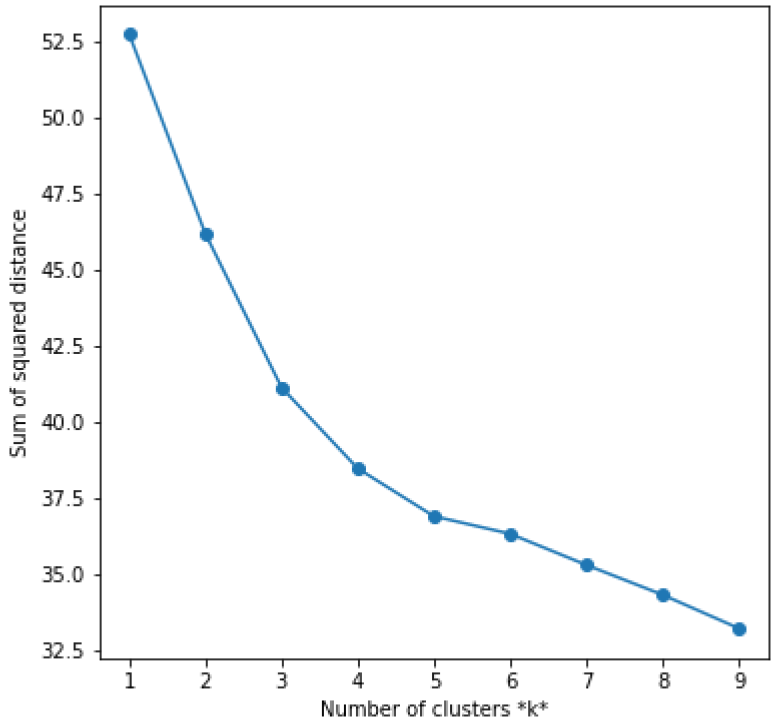
]:

	Ward	1st Common Venue	2nd Common Venue	3rd Common Venue	4th Common Venue	5th Common Venue
0	A Narayanapura	Restaurant	Bus Station	Bus Stop	Theater	Bus Stop
1	Adugodi	Snacks and Cafe	Restaurant	Men's Store	Mexican Restaurant	Men's Store
2	Agaram	Burger Joint	Bus Station	ATM	Music Venue	ATM
3	Agrahara Dasarahalli	Bagel Shop	Athletics & Sports	Restaurant	Snacks and Cafe	Restaurant
4	Anjanapura	ATM	Pizza Place	Mexican Restaurant	Middle Eastern Restaurant	Mexican Restaurant
...	...	...	...	...	...	...
166	Vrisabhavathi Nagar	ATM	Pizza Place	Mexican Restaurant	Middle Eastern Restaurant	Mexican Restaurant
167	Yediyur	Restaurant	Food	Lake	Department Store	Lake
168	Yelahanka Satellite Town	Snacks and Cafe	Pizza Place	Dessert Shop	Outlet Store	Dessert Shop
169	Yelchenahalli	Sporting Goods Shop	ATM	Museum	Mexican Restaurant	Museum
170	Yeshwanthpura	Restaurant	Clothing Store	ATM	Miscellaneous Shop	ATM

171 rows × 6 columns

K-Means Clustering

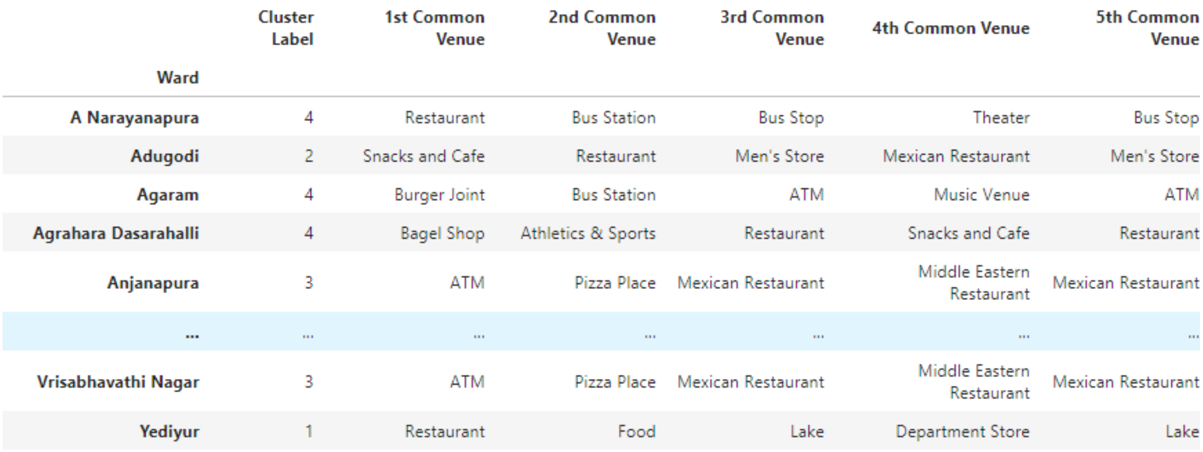
To start the clustering, I have first used the **elbow method** to find the optimum value of k to cluster this data.



From the elbow plot, it is clear that **optimum value of K for clustering this data is 5** as there is a bit sharp turn in graph for k=5.

Now I have used the KMeans function of **sklearn** library to initiate the instance of KMeans and then used fit function to model the data using KMeans clustering algorithm.

After this I added the label of each cluster to the above table created for top 5 venues and set the index to “Ward”.

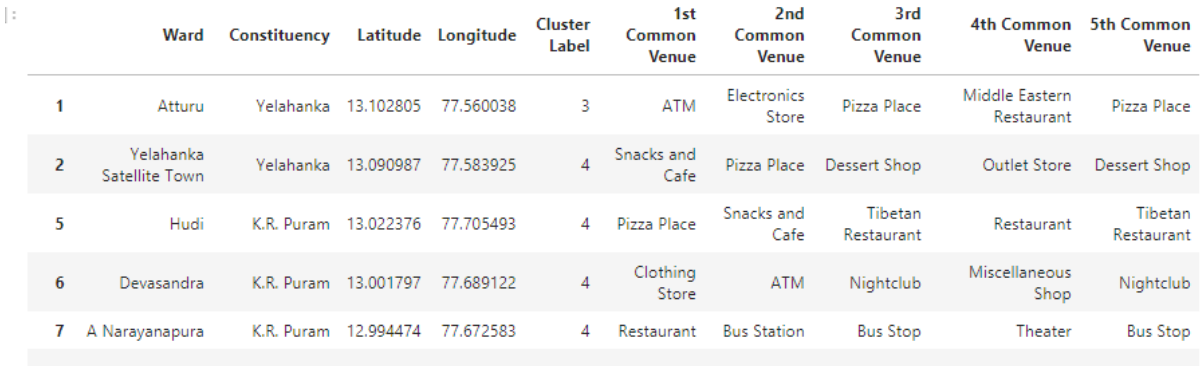


	Cluster Label	1st Common Venue	2nd Common Venue	3rd Common Venue	4th Common Venue	5th Common Venue
Ward						
A Narayanapura	4	Restaurant	Bus Station	Bus Stop	Theater	Bus Stop
Adugodi	2	Snacks and Cafe	Restaurant	Men's Store	Mexican Restaurant	Men's Store
Agaram	4	Burger Joint	Bus Station	ATM	Music Venue	ATM
Agrahara Dasarahalli	4	Bagel Shop	Athletics & Sports	Restaurant	Snacks and Cafe	Restaurant
Anjanapura	3	ATM	Pizza Place	Mexican Restaurant	Middle Eastern Restaurant	Mexican Restaurant
...	...	...	...	...	...	...
Vrisabhavathi Nagar	3	ATM	Pizza Place	Mexican Restaurant	Middle Eastern Restaurant	Mexican Restaurant
Yediyur	1	Restaurant	Food	Lake	Department Store	Lake

5. Result

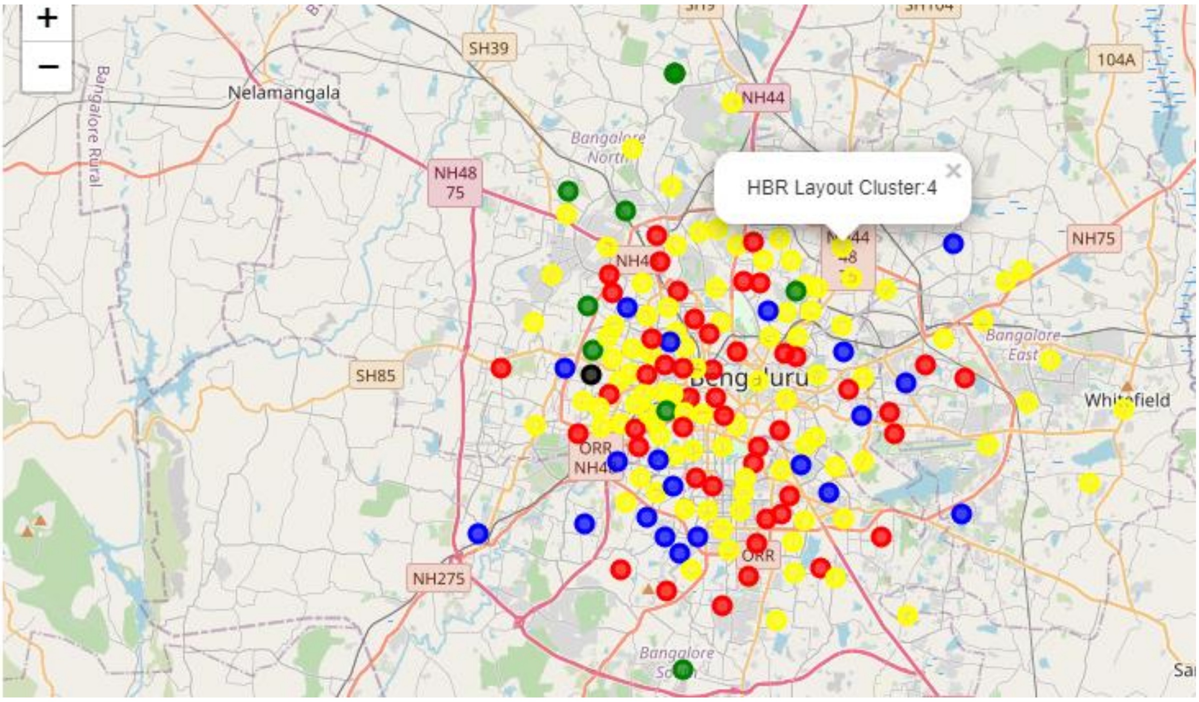
5.1 Plotting the data on Folium Map

Next, I will try to prepare the table for plotting. For this I would need the name of Constituency and the longitude , latitude of each ward in the cluster labelled table. So, merging the master data set with above table. The final table that I got after **merging** the two tables:



	Ward	Constituency	Latitude	Longitude	Cluster Label	1st Common Venue	2nd Common Venue	3rd Common Venue	4th Common Venue	5th Common Venue
1	Atturu	Yelahanka	13.102805	77.560038	3	ATM	Electronics Store	Pizza Place	Middle Eastern Restaurant	Pizza Place
2	Yelahanka Satellite Town	Yelahanka	13.090987	77.583925	4	Snacks and Cafe	Pizza Place	Dessert Shop	Outlet Store	Dessert Shop
5	Hudi	K.R. Puram	13.022376	77.705493	4	Pizza Place	Snacks and Cafe	Tibetan Restaurant	Restaurant	Tibetan Restaurant
6	Devasandra	K.R. Puram	13.001797	77.689122	4	Clothing Store	ATM	Nightclub	Miscellaneous Shop	Nightclub
7	A Narayanapura	K.R. Puram	12.994474	77.672583	4	Restaurant	Bus Station	Bus Stop	Theater	Bus Stop

Again, using the folium library we will plot the wards on map. This time map will contain the cluster information as well. I have given the **different colour of each cluster**. I have also used the CircleMarker and popup function to make the map more interactive.

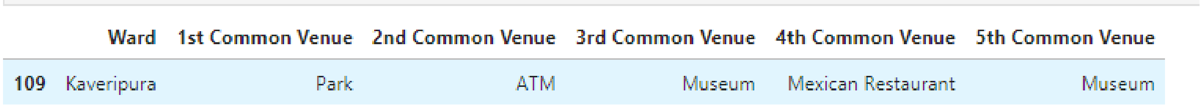


So, one of the major aim of this report was to create various clusters and plot them on map using KMeans clustering algorithm and that has been achieved.

Now I would like to analyse each cluster in a bit detail to understand which cluster of regions is suitable for Restaurants business and which is suitable for living.

5.2 Analysing Cluster One

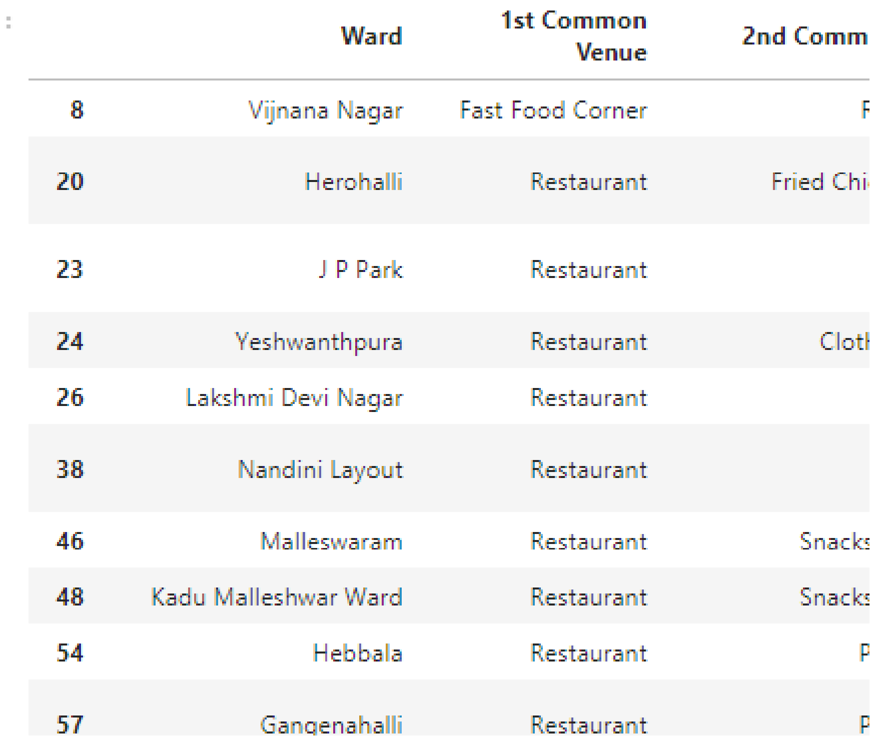
There is only one ward in this cluster. So, I cannot make a general assumption about this cluster. Let us look at other clusters.



	Ward	1st Common Venue	2nd Common Venue	3rd Common Venue	4th Common Venue	5th Common Venue
109	Kaveripura	Park	ATM	Museum	Mexican Restaurant	Museum

5.3 Analysing Second cluster

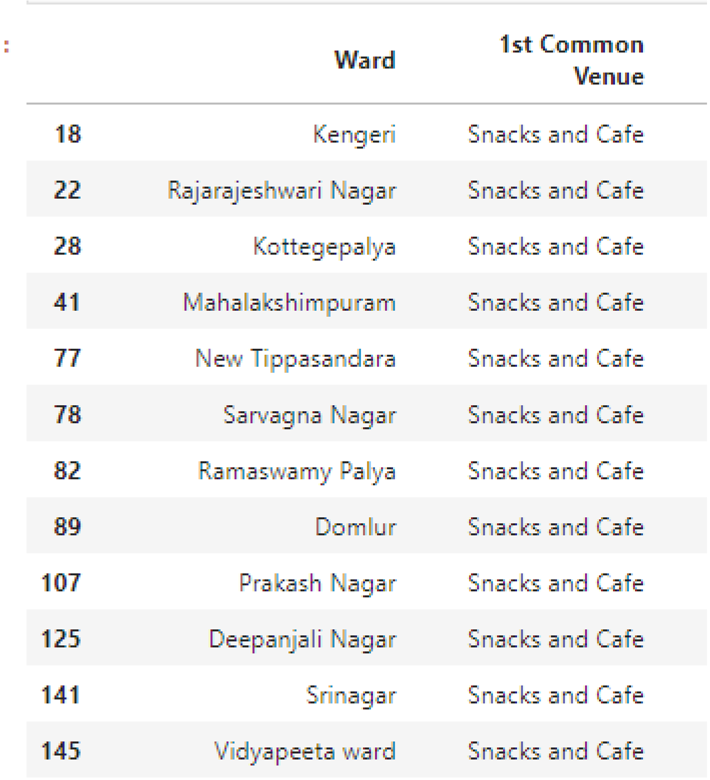
It is evident from the table below of second cluster that topmost common venue category in these region is Restaurant and we can easily say that this is a hub for food. So, these regions/wards are ideal for opening Restaurants. We can name this Cluster as **Restaurant Cluster.**



	Ward	1st Common Venue	2nd Comm
8	Vijnana Nagar	Fast Food Corner	F
20	Herohalli	Restaurant	Fried Chi
23	J P Park	Restaurant	
24	Yeshwanthpura	Restaurant	Clotl
26	Lakshmi Devi Nagar	Restaurant	
38	Nandini Layout	Restaurant	
46	Malleswaram	Restaurant	Snacks
48	Kadu Malleshwar Ward	Restaurant	Snacks
54	Hebbala	Restaurant	P
57	Gangenahalli	Restaurant	P

5.4 Analysing Cluster three

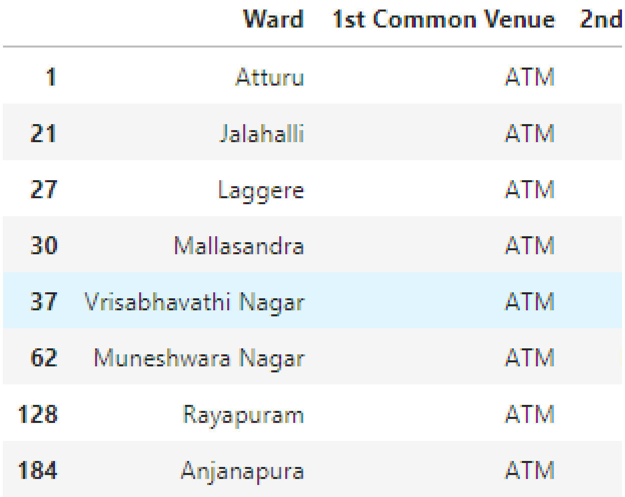
The topmost common venue in this cluster is Snacks and Cafe. Snacks and Cafe is a merged category that contains Breakfast places, quick bites, and Coffee Place. We can name this cluster as **Quick Bite.**



	Ward	1st Common Venue
18	Kengeri	Snacks and Cafe
22	Rajarajeshwari Nagar	Snacks and Cafe
28	Kottegepalya	Snacks and Cafe
41	Mahalakshimpuram	Snacks and Cafe
77	New Tippasandara	Snacks and Cafe
78	Sarvagna Nagar	Snacks and Cafe
82	Ramaswamy Palya	Snacks and Cafe
89	Domlur	Snacks and Cafe
107	Prakash Nagar	Snacks and Cafe
125	Deepanjali Nagar	Snacks and Cafe
141	Srinagar	Snacks and Cafe
145	Vidyapeeta ward	Snacks and Cafe

5.5 Analysing Cluster Four

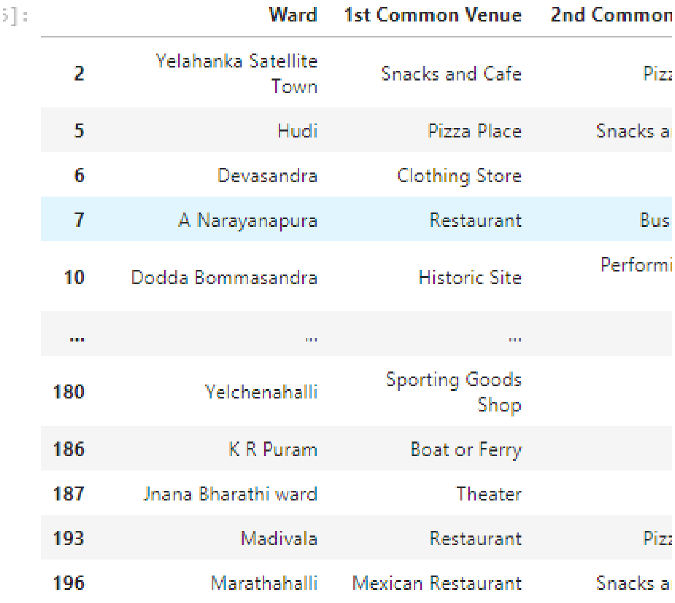
In the cluster, the topmost common venue is ATM. So, it can work as guide for people who are looking for ATMs. Also, the clusters are great for Middle Eastern, Mexican Restaurants and pizza places. We can name this Cluster as **ATM Hub.**



	Ward	1st Common Venue	2nd
1	Atturu	ATM	
21	Jalahalli	ATM	
27	Laggere	ATM	
30	Mallasandra	ATM	
37	Vrisabhavathi Nagar	ATM	
62	Muneshwara Nagar	ATM	
128	Rayapuram	ATM	
184	Anjanapura	ATM	

5.6 Analysing Cluster Five

Looking at the second cluster, it is clear that there is no topmost common venue category in this region. It is a mix of Gym, Bakery, Theatre, clothing store, Dance Studio, Sporting Goods shop, and ATM. There are 45 unique categories of venues in topmost common venue. This cluster seems perfect for living as all the basic amenities are available nearby. This cluster could be named as **Housing Cluster.**



5]:

	Ward	1st Common Venue	2nd Common
2	Yelahanka Satellite Town	Snacks and Cafe	Pizz
5	Hudi	Pizza Place	Snacks a
6	Devasandra	Clothing Store	
7	A Narayanapura	Restaurant	Bus
10	Dodda Bommasandra	Historic Site	Performi
...	...	...	
180	Yelchenahalli	Sporting Goods Shop	
186	K R Puram	Boat or Ferry	
187	Jnana Bharathi ward	Theater	
193	Madivala	Restaurant	Pizz
196	Marathahalli	Mexican Restaurant	Snacks a

6. Discussion

Since, Bangalore is one of the fastest growing place in India and is IT hub of the country as well, the city is expanding fast, and the structure of city has become overly complex. There are areas dedicated to IT firms and then there are residential localities. Different areas are famous for different things such as for marketing, for dining, for nightlife etc. As per my data, there are 27 constituencies and 198 wards, which makes the analysis process overly complex. Various clustering approaches could be used to analyse such data and there could be some difference in the result of each method.

For my project, I have used the KMeans clustering algorithm. To get the optimum value of K, I have evaluated the model using elbow method. The value that come out of elbow method was 5. The data used might be a bit outdated and some more constituencies might have been added to this date. However, same code could be used for the analysis of new data as well by just changing the initial json file (if format of new json is same)

I used some visualization through bar chart and folium map to better analyse the data. I have also used foursquare API for exploring the neighbourhoods. Further, information about schools in various localities could also be fetched from net and then analysis could be done on that as well.

8. Conclusion

Just like Bangalore, there are many more cities in India such as New Delhi, Mumbai, Hyderabad etc that are getting lots of migrants every year. These migrants come to these cities in search of business opportunities or better standard of living.

This type of project could help them in understanding that which region is suited for which type of business and which region is fit for accommodation. They can get this information even before visiting the city with the help of data analysis and machine learning.

Note: Please go through code to better understand this project.

References:

https://www.linkedin.com/pulse/housing-sales-prices-venues-data-analysis-ofistanbul-sercan-y%C4%B1ld%C4%B1z/

https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/DP0701EN/sample_submission/Predicting_the_Improvement_of_NBA_players_Report.pdf

https://raw.githubusercontent.com/datameet/Municipal_Spatial_Data/master/Bangalore/BBMP.GeoJSON