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

Abdullah Shahab

28 July’20

1. **Introduction**
   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 can help in dividing the regions of Bangalore into various cluster and this can help in identifying the region with fewer restaurants.

* 1. **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.

1. **Data to solve the problem**

To identify the various regions of Bangalore and explore the venues in various regions of Bangalore, I gave 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.

1. **Source of data**

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

Actual GeoJson

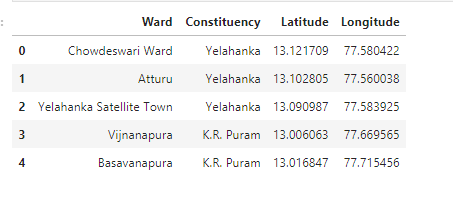
<https://raw.githubusercontent.com/datameet/Municipal_Spatial_Data/master/Bangalore/BBMP.GeoJSON>

1. **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.

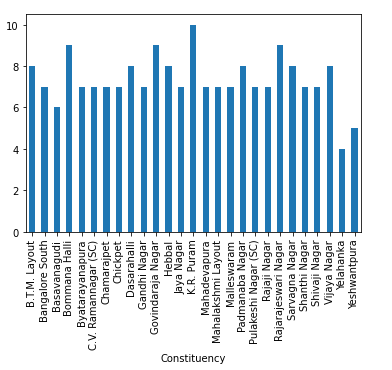
After this I fetched the content of json file and stored it in a variable using json.load() function. After analysing the first node of json file for its various attributes, I converted the json file into dataframe. In this process I kept only those attributes which were required and dropped the rest 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:

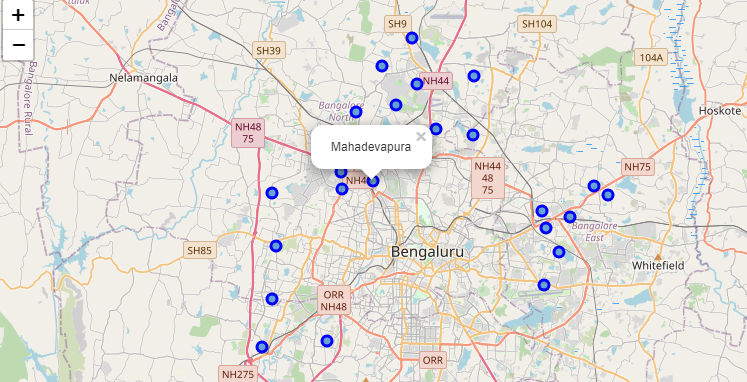


This is only a small portion. There are total 27 constituencies and 198 wards.

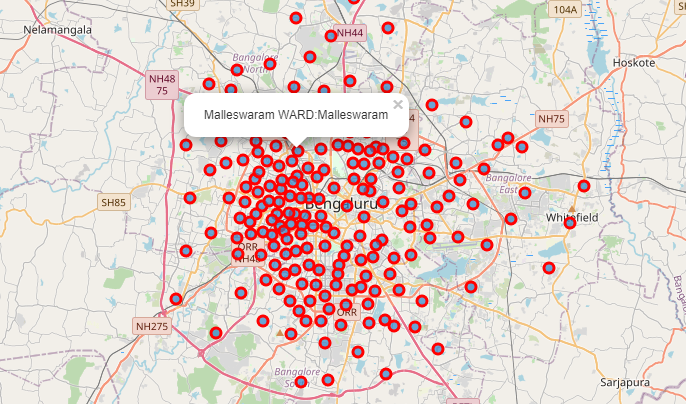
Bar chart to show number of wards under each constituency.



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, we will be using the **Foursquare API. Firstly, I tried to analyse single ward and took the *Sudham Nagar ward.***

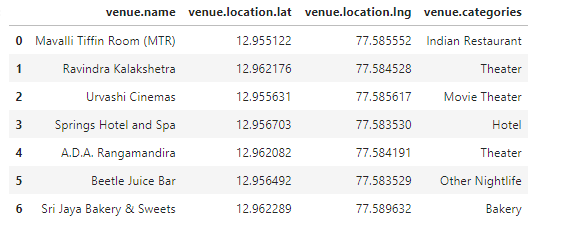
**To make call to the foursquare API, a URL is created with fixed format as given below:**

[https://api.foursquare.com/v2/venues/explore?&client\_id= xx&client\_secret= xx&v= xx&ll= xx&radius=xx&limit=xx](https://api.foursquare.com/v2/venues/explore?&client_id=%20xx&client_secret=%20xx&v=%20xx&ll=%20xx&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.

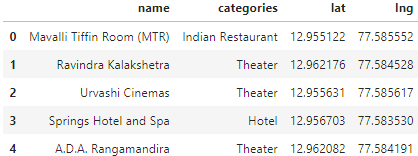
The data frame receive contains lot of unnecessary attributes and we need to drop all those attributes. We only need Venue name, latitude, longitude, and Category. Venue name, latitude and longitude could be fetched directly from dataframe, but venue category is in nested form and I then used get\_category()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

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



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*.



**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:



**Finding top 5 venues 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.

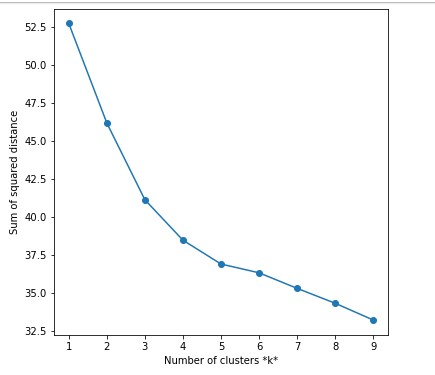
Now, to find the top 5 venues in each category, I have created a function that will arrange the dummy table in descending order for each ward one by one. In this way, we will be able to fetch top 5 venues of each ward in each iteration and add this data in a new data frame.

Final table with top 5 venues in each category



**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.

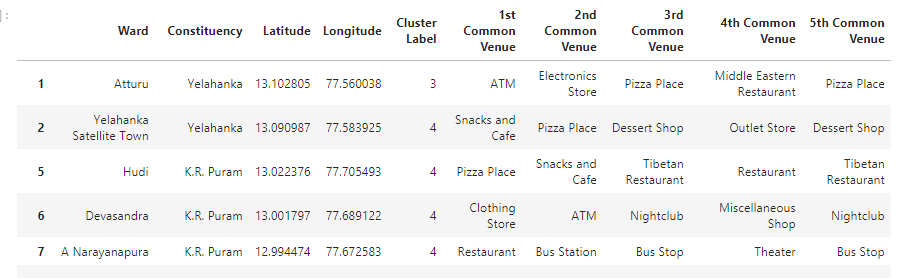
After this I added the label of each cluster to the above table created for top 5 venues and set the index to “Ward”.



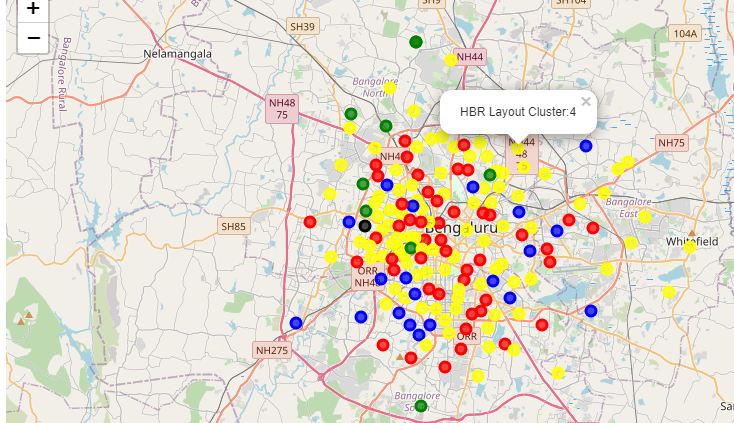
1. **Result**

**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:



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.

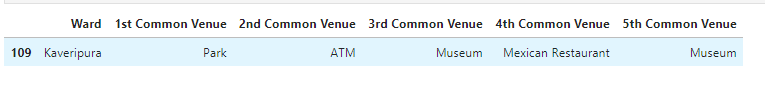


So, one of the major aim of this report was to create various clusters and plot them on map using KMeans clustering algorithm and that has been achieved.

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.

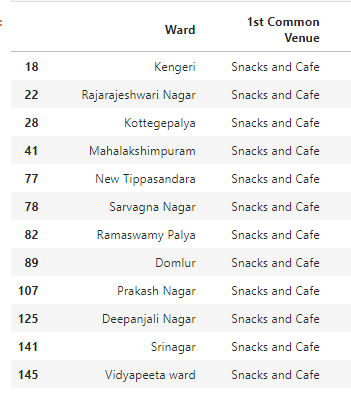
**5.3 Analysing Second cluster**

It is evident from the table below of second cluster that topmost common venue in this 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.**



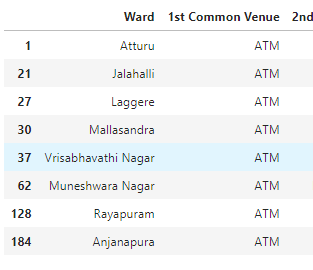
**5.4 Analysing Cluster three**

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.**



**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.**



**5.6 Analysing Cluster Five**

Looking at the second cluster, it is clear that there is no topmost common venue 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 are there 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.**



1. **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.

1. **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 understand this project better.

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>