

THE BATTLE OF NEIGHBOURHOOD

CAPSTONE PROJECT REPORT FOR APPLIED DATA
SCIENCE CERTIFICATION

GAWALI.SAGAR@GMAIL.COM

Contents

1. Introduction/Business Problem	3
2. Data	4
3. Methodology.....	5
4. Results & Observations	8
5. Conclusion.....	9

1. INTRODUCTION/BUSINESS PROBLEM

Bengaluru is the capital of the Indian state of Karnataka. It has a population of more than 11 million, making it the third-most populous city and fifth-most populous urban agglomeration in India. Located in southern India on the Deccan Plateau, at a height of over 900 m (3,000 ft) above sea level, Bangalore is known for its pleasant climate throughout the year. The city is also widely regarded as the "Silicon Valley of India" (or "IT capital of India") because of its role as the nation's leading information technology (IT) exporter. Indian technological organisations such as ISRO, Infosys, Wipro, and HAL are headquartered in the city. A demographically diverse city, Bangalore is the second-fastest-growing major metropolis in India. Recent estimates of the metro economy of its urban area have ranked Bangalore either the fourth or fifth-most productive metro area of India. It is home to many educational and research institutions in India. Numerous state-owned aerospace and defence organisations, such as Bharat Electronics, Hindustan Aeronautics, and National Aerospace Laboratories are located in the city. The city also houses the Kannada film industry.

Owing to the factors mentioned above, there is a high demand for residential apartments/bungalows/villas, etc. (both rented/owned). And most of the time a newcomer in the city or someone who is trying to relocate within the city doesn't know which place is good for staying and which is not based on personal criteria. As part of this project, we would try to answer this question - Which area is suitable for different categories of people. We will also try to answer whether the area is good for staying for a bachelor's or families.

We can classify people into different categories. E.g. - someone who loves native food (i.e Indian Food, or western food), someone who prefers other sorts of entertainment, gym and fitness center, etc.

2. DATA

To answer the question raised in section, we will need the following set of data:

- Classification of areas of Bengaluru city into smaller area i.e. ward (a ward is an administrative unit of the city region;)
- Venue data (Malls, market, Supermarket, Parks, Schools, recreational activities, restaurants)

We will use the following sources for the above-mentioned data categories

1. Ward data: <https://opencity.in/data/bbmb-wards>

Sample:

Ward No	Ward Name	BBMP Zone Name	BBMP Division	BBMP Sub Division	Assembly constituency	MP Constituency
1	Kempegowda Ward	Yelahanka	Yelahanka	Yelahanka	Yelahanka	Chikballapur
2	Chowdeswari Ward	Yelahanka	Yelahanka	Yelahanka	Yelahanka	Chikballapur
3	Atturu	Yelahanka	Yelahanka	Yelahanka	Yelahanka	Chikballapur
4	Yelahanka Satellite Town	Yelahanka	Yelahanka	Yelahanka	Yelahanka	Chikballapur
5	Jakkuru	Yelahanka	Byatarayanapura	Byatarayanapura	Byatarayanapura	Bangalore North
6	Thanisandra	Yelahanka	Byatarayanapura	Byatarayanapura	Byatarayanapura	Bangalore North
7	Byatarayanapura	Yelahanka	Byatarayanapura	Byatarayanapura	Byatarayanapura	Bangalore North
8	Kodigehalli	Yelahanka	Byatarayanapura	Byatarayanapura	Byatarayanapura	Bangalore North
9	Vidyaranyapura	Yelahanka	Byatarayanapura	Vidyaranyapura	Byatarayanapura	Bangalore North
10	Dodda Bommasandra	Yelahanka	Byatarayanapura	Vidyaranyapura	Byatarayanapura	Bangalore North

2. Venue Data: Foursquare APIs

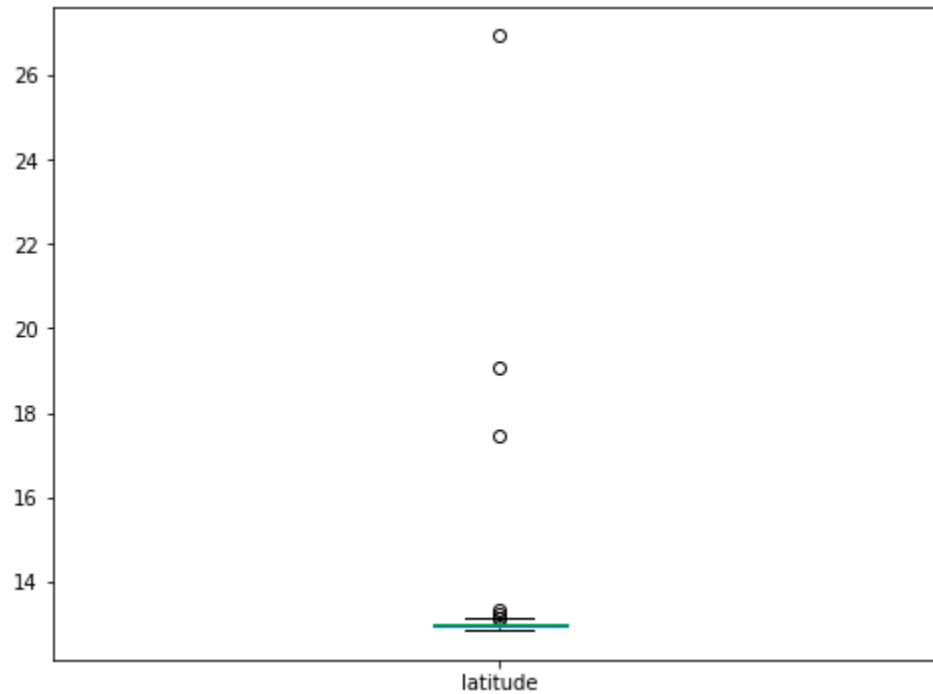
	Ward Name	Ward Latitude	Ward Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Kempegowda Ward	12.9564316	77.5741194	Brahmins Coffee Bar	12.95398336	77.5688621	Breakfast Spot
1	Kempegowda Ward	12.9564316	77.5741194	VB Bakery	12.95168678	77.57712392	Bakery
2	Kempegowda Ward	12.9564316	77.5741194	Food Street	12.9509559	77.57779773	Snack Place
3	Kempegowda Ward	12.9564316	77.5741194	Hari Super Sandwich	12.94906313	77.57493508	Sandwich Place
4	Kempegowda Ward	12.9564316	77.5741194	Bangalore City Institute	12.9569247	77.57291529	General Entertainment
5	Kempegowda Ward	12.9564316	77.5741194	Karnataka Bhel House	12.95941243	77.56726435	Fast Food Restaurant

We will be using Venue Category to drive analysis.

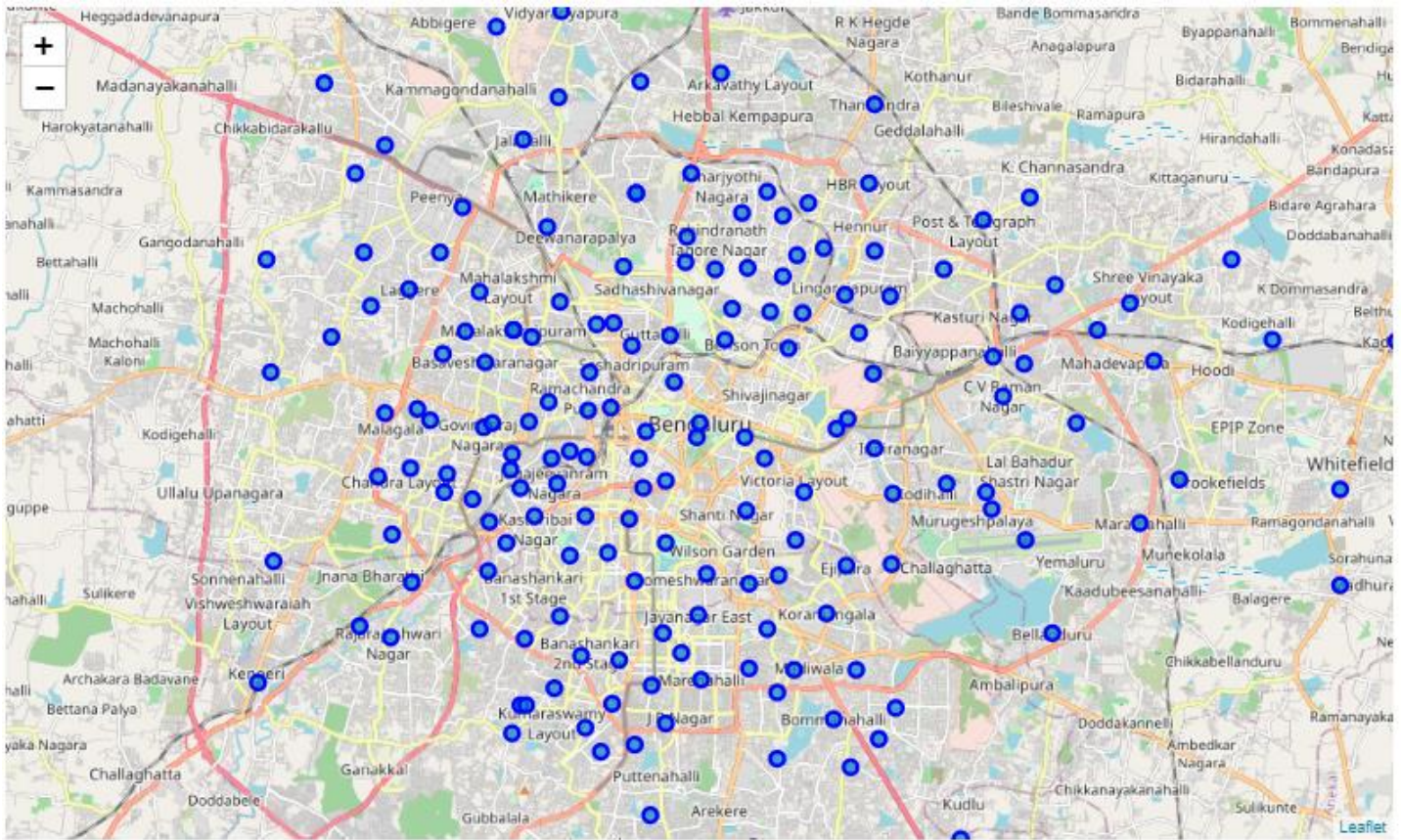
3. METHODOLOGY

Following methodology was followed to perform the analysis:

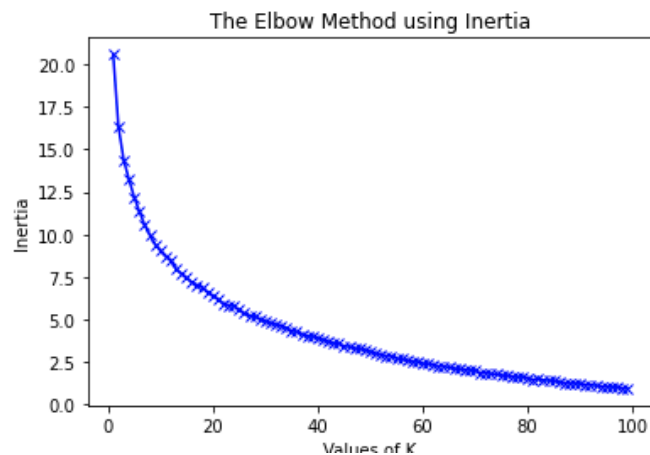
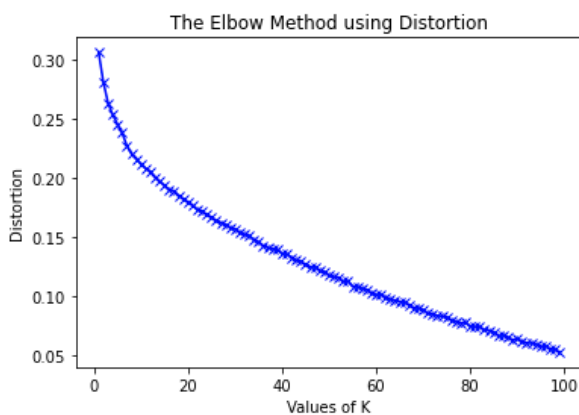
1. Data collection from the opencity portal for list of wards.
 - a. Manual Data cleansing of the list of wards
 - b. Identifying latitude and longitude of the wards for analysis
 - i. Remove missing values of latitude and longitude.
 - ii. Identified and Remove outliers in the data
 1. Box plot showing outliers in the data



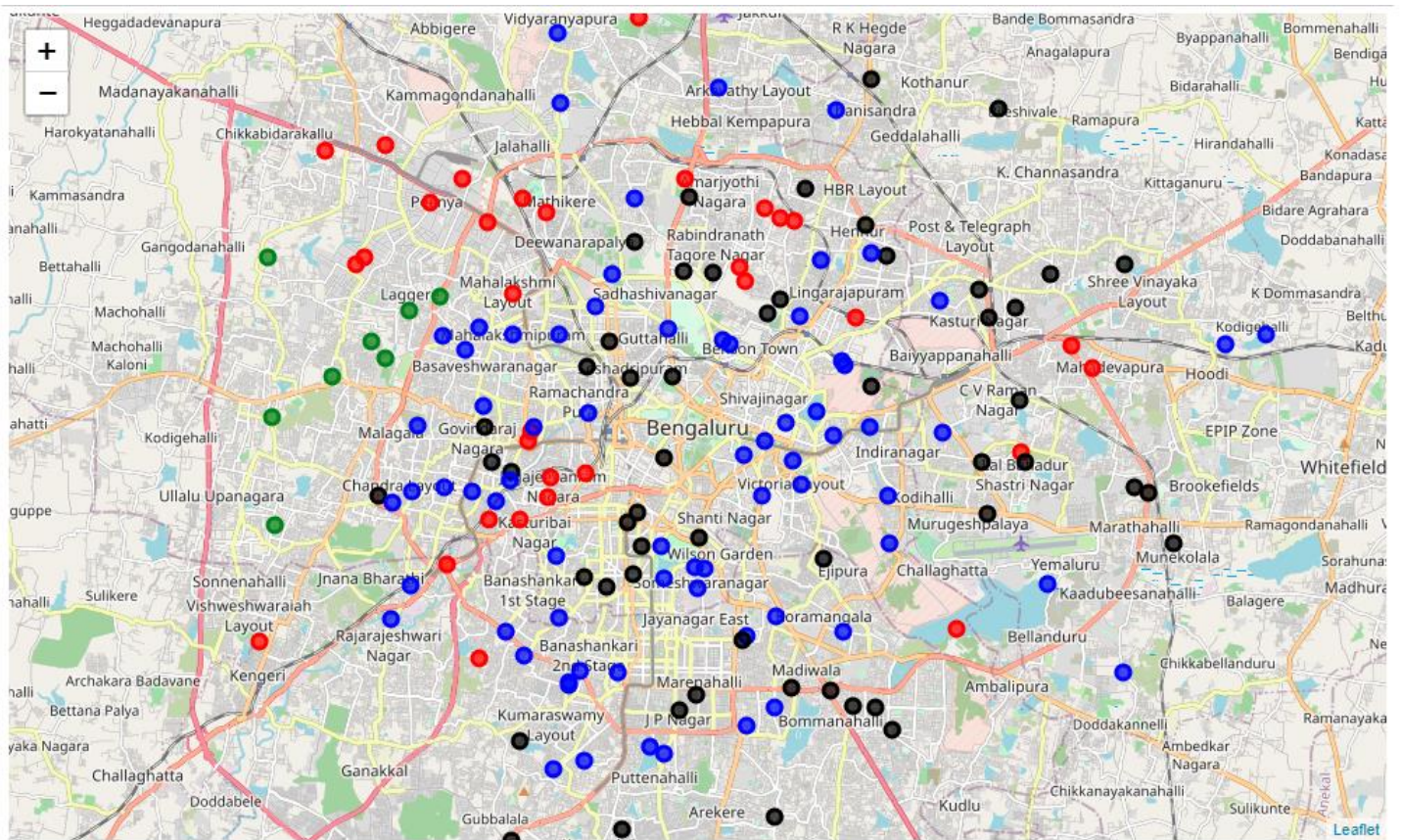
- c. Plot the ward on folium maps



2. Based on list of wards, latitude and, longitude, capturing venue data from Foursquare APIs
3. Perform ward & Venue level analysis
 - a. It was identified that there were too many unique venue categories present in the data. Consolidate the venue categories based on business understanding and reduced the categories for further analysis.
4. Performed one-hot coding i.e. created dummy variables
5. Applied K-means algorithm for the clustering
 - a. Tried to identify optimal 'K' value for clustering



- b. Performed clustering by keeping value of $K = 4$.
 - i. Plotted clusters on folium map



6. Examined and analysed the cluster to provide observation, results and conclusion.

4. RESULTS & OBSERVATIONS

Based on the data, every ward in the city was classified into one of the four clusters. This was achieved through unsupervised learning algorithm of clustering using K- means. These four cluster were created from the data fetched from the foursquare API's related to venues.

Cluster 1: In this cluster, first 3 'most common venues' are related to the food.

Cluster 2: In this cluster, first 3 'most common venues' are related to Sports, Gym & Fitness Centre, Entertainment, Travel, Supermarket and restaurant, park etc

Cluster 3: In this cluster, first 3 'most common venues' are related to food and more specifically to Indian Restaurants, snack & refreshments and Groceries

Cluster 4: In this cluster, first 4 'most common venues' are related to ATM, Sports, Gym & Fitness Centre, Automobile.

5. CONCLUSION

1. Based on clusters obtained and observations made, following conclusion can be drawn:

Cluster	Suitable for
1	For the people loving different types food or bachelors who love to eat outside instead of cooking at home
2	For the people who are extroverts, who love to go out and enjoy or families who need all rounded facilities like Gym, sports, entertainment, parks, easy travel arrangement for schools and offices
3	For the people who love traditional food or native food, mostly to the higher age people who are accustomed to the native/traditional food items
4	For the people who enjoys food prepared at home but like to go out for sports, gym and fitness center, have personal vehicles

2. Though these conclusions are drawn based on cluster, these don't feel distinctive clusters. Some of these clusters feel overlapping and clear demarcation is absent.
3. Also, when we tried to identify optimal 'K' value for clustering using K-means, there was no clear output using elbow method.
4. Foursquare venue data had different categories present, however some of the key categories were missing e.g. Hospitals, Schools, Banks, Parks, Day Care Centers and Amusement Parks in the locality. There is high presence of food related categories. Thus, clustering is heavily biased towards food category. More elaborate venue data is required for better analysis of the neighborhood.