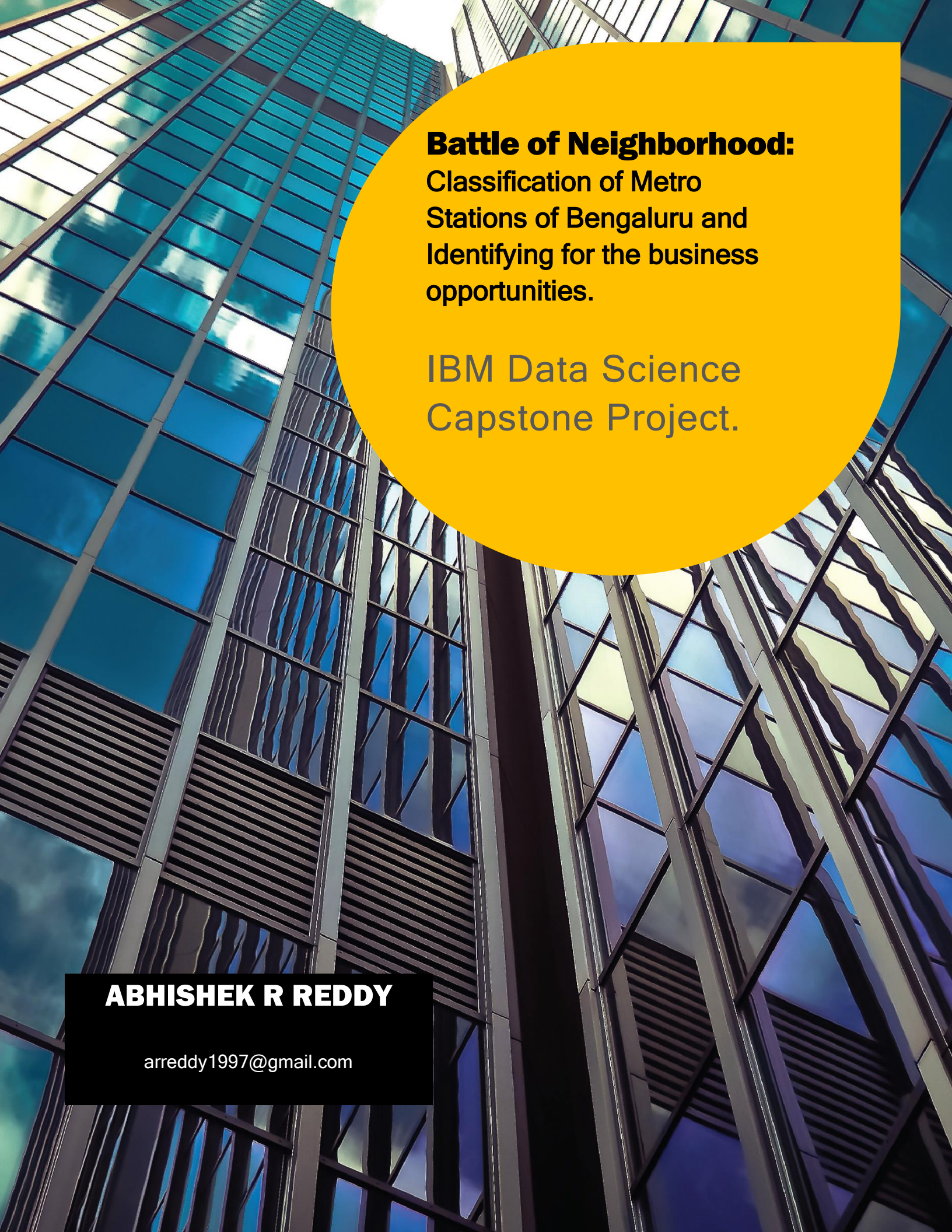


Battle of Neighborhood:

Classification of Metro Stations of Bengaluru and Identifying for the business opportunities.

IBM Data Science Capstone Project.

ABHISHEK R REDDY

arreddy1997@gmail.com

Table of Contents

1. Introduction 3
1.1 Problem Statement 3
1.2 Target Audience 4
2. Data 4
3. Methodology 5
3.1 Fetching data from Foursquare 5
3.2 Exploratory Data analysis 5
3.2 K-Means Clustering 7
3.2.1 Elbow method 7
3.2.1 Classification of station using the K Means clustering 8
4. Results 10
5. Discussion and Conclusion 12

Table of Figures

Figure 1 Data which contains Station Names with Co-Ordinates 4
Figure 2 The Venue Categories 5
Figure 3 The Venues categories count 5
Figure 4 Box plot of the categorical variable 6
Figure 5 Boxplot after dropping few columns and Normalization 6
Figure 6 WCCS graph 8
Figure 7 Box Plot of the all the four cluster 8

1. Introduction

Bangalore officially known as Bengaluru (12.97° N 77.56° E) is the capital of the Indian state of Karnataka which is located in the southern part of India. Bangalore is located on the heart of Mysore Plateau which is a part of Deccan Plateau. Bangalore metro known as " Namma Metro " which translates into English as ' Our Metro' is a rapid transit system serving the city of Bangalore, India. It is the fourth longest operational metro network in India after the Delhi Metro, Hyderabad Metro and Chennai Metro. The metro network consists of two color-coded lines with a total length of 42.3 kilometers serving 40 stations. By 2023, the system is expected to complete its phase 2 network and provide connectivity to the city's important tech hubs of Electronic City and Whitefield.

1.1 Problem Statement

Business opportunities of the around the metro stations is very attractive. It attracts large sets of people as it is commute. So for the business is around the metro station will be profitable. But having the same type of business around the metro station will leads to competitive market. As competition grows the profits will eventually less.

This project mainly focuses on what kind of the businesses which are currently existing around the existing as well as upcoming metro stations and clustering to identify the gaps in the market. By considering the metro stations which are still in construction phase we can plan the for business which are currently nor existing around those metro stations.

1.2 Target Audience

Current project will be useful for the entrepreneurs who are looking for business opportunities in Bengaluru city. It will be also useful to the construction builders of the localities of the metro stations, so that they can estimate their future real estate values.

2. Data

We need the data of all the existing metro stations as well as the upcoming metro stations. The data should consists of the Station name, Co-ordinates. But unfortunately there no websites which directly give this information. So the data scrapping is the only solution. In this link, Future Expansion Plans of Bangalore Metro Phase 2 we can find all the list of existing and upcoming metro stations of Bangalore. All the metro station name/location address will be scrapped from this web page using BeatifulSoup library. The location coordinates to the corresponding names will be obtained using the GeoPy API.

Station_names	Latitude	Longitude
Bangalore International Exhibition Centre (BIE...	13.060757	77.474365
Whitefield, Bangalore	12.969637	77.749745
Ragigudda Temple, Bangalore	12.914189	77.593330
Nagawara, Bangalore	13.043141	77.620909
Jindal, Bangalore	13.053283	77.483814
Jayadeva Hospital Interchange, Bangalore	12.916721	77.599941
Arabic College, Bangalore	13.030009	77.620866
Kadugodi, Bangalore	12.998577	77.760972
BTM Layout, Bangalore	12.915177	77.610282
Nagasandra, Bangalore	13.047950	77.500135
ITPL, Bangalore	12.984569	77.737665

Figure 1 Data which contains Station Names with Co-Ordinates

3. Methodology

3.1 Fetching data from Foursquare

- The Venue categories is fetched from Foursquare API.

```
Arts & Entertainment (4d4b7104d754a06370d81259)
College & University (4d4b7105d754a06372d81259)
Event (4d4b7105d754a06373d81259)
Food (4d4b7105d754a06374d81259)
Nightlife Spot (4d4b7105d754a06376d81259)
Outdoors & Recreation (4d4b7105d754a06377d81259)
Professional & Other Places (4d4b7105d754a06375d81259)
Residence (4e67e38e036454776db1fb3a)
Shop & Service (4d4b7105d754a06378d81259)
Travel & Transport (4d4b7105d754a06379d81259)
```

Figure 2 The Venue Categories

- The corresponding venue categories count is fetched from the Foursquare API.

Station_names	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
Bangalore International Exhibition Centre (BIE...	0	1	1	5	0	0	4	0	1	1
Whitefield, Bangalore	0	5	0	15	5	5	15	0	6	2
Ragigudda Temple, Bangalore	4	7	1	84	13	10	28	1	27	12
Nagawara, Bangalore	5	2	2	25	5	3	29	0	5	5
Jindal, Bangalore	0	0	0	4	0	0	3	0	1	1

Figure 3 The Venues categories count

3.2 Exploratory Data analysis

- In order to view the the categorical variables relationship among them, the boxplot will be plotted as shown in below.

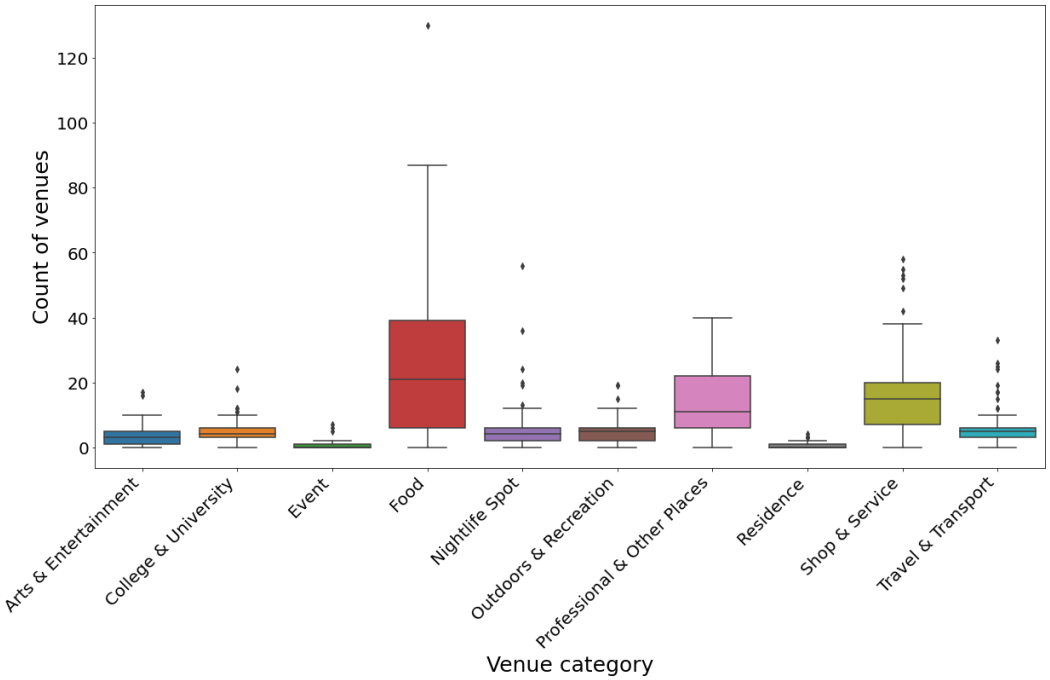


Figure 4 Box plot of the categorical variable

- It has been observed that we have very less data about the 'Event' and 'Residence' categories so both has been dropped from the data frame.
- As the data is distributed with different limits it is difficult to compare them, so we need to normalize the data. After normalization again the boxplot has been plotted.

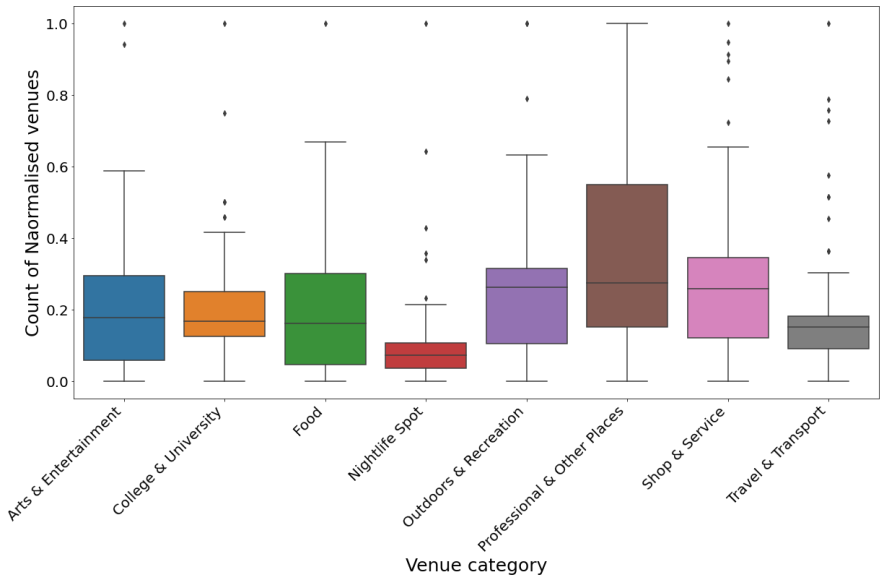


Figure 5 Boxplot after dropping few columns and Normalization

3.2 K Means Clustering

K-means clustering is a clustering method that subdivides a single cluster or a collection of data points into K different clusters or groups.

The algorithm analyzes the data to find organically similar data points and assigns each point to a cluster that consists of points with similar characteristics. Each cluster can then be used to label the data into different classes based on the characteristics of the data. K-Means clustering works by constantly trying to find a centroid with closely held data points. This means that each cluster will have a centroid and the data points in each cluster will be closer to its centroid compared to the other centroids.

Choosing the Right Number of Clusters is crucial in K-means clustering. An ideal way to figure out the right number of clusters would be to calculate the Within-Cluster-Sum-of-Squares (WCSS). WCSS is the sum of squares of the distances of each data point in all clusters to their respective centroids.

3.2.1 Elbow method

We can find the optimum value for K using an Elbow point graph. We randomly initialize the K-Means algorithm for a range of K values and will plot it against the WCSS for each K value.

The WCSS plot for the data set has been plotted and shown below

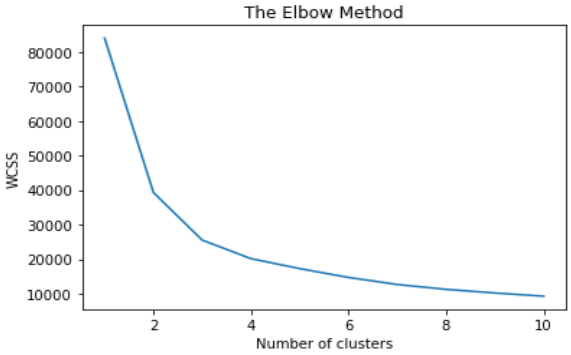


Figure 6 WCCS graph

From the above we can conclude that the ideal number of cluster will be 4.

3.2.1 Classification of station using the K Means clustering

The K means has been trained with available data. After the training the box plot of the all the four cluster has been plotted. Based on that we can draw the conclusion about the metro stations.

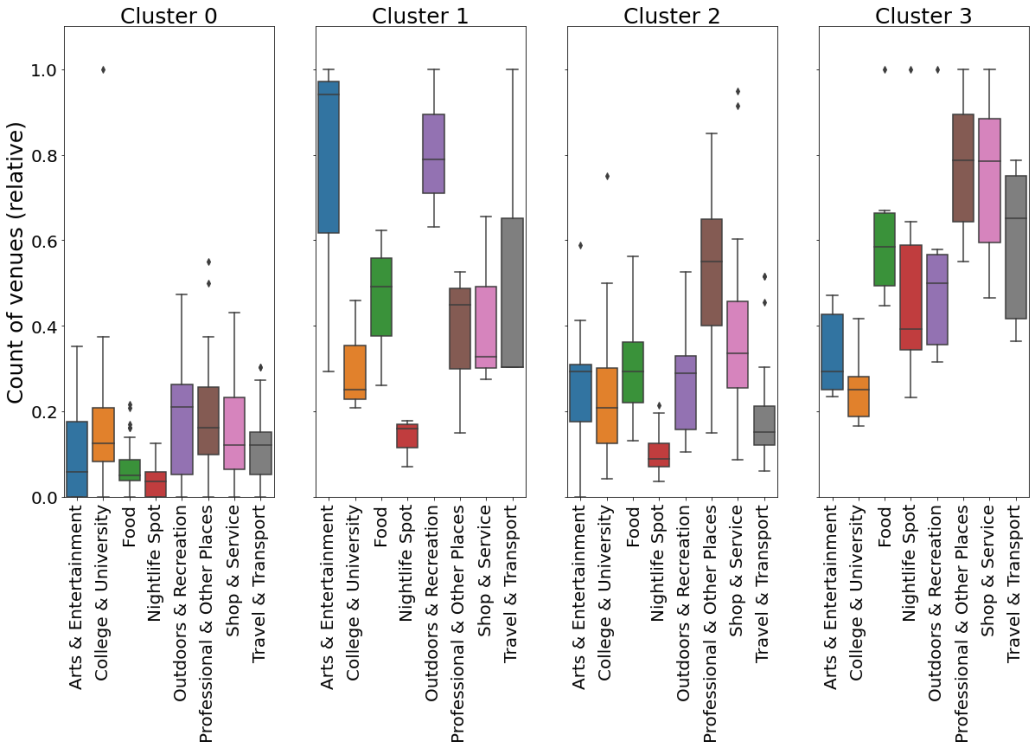
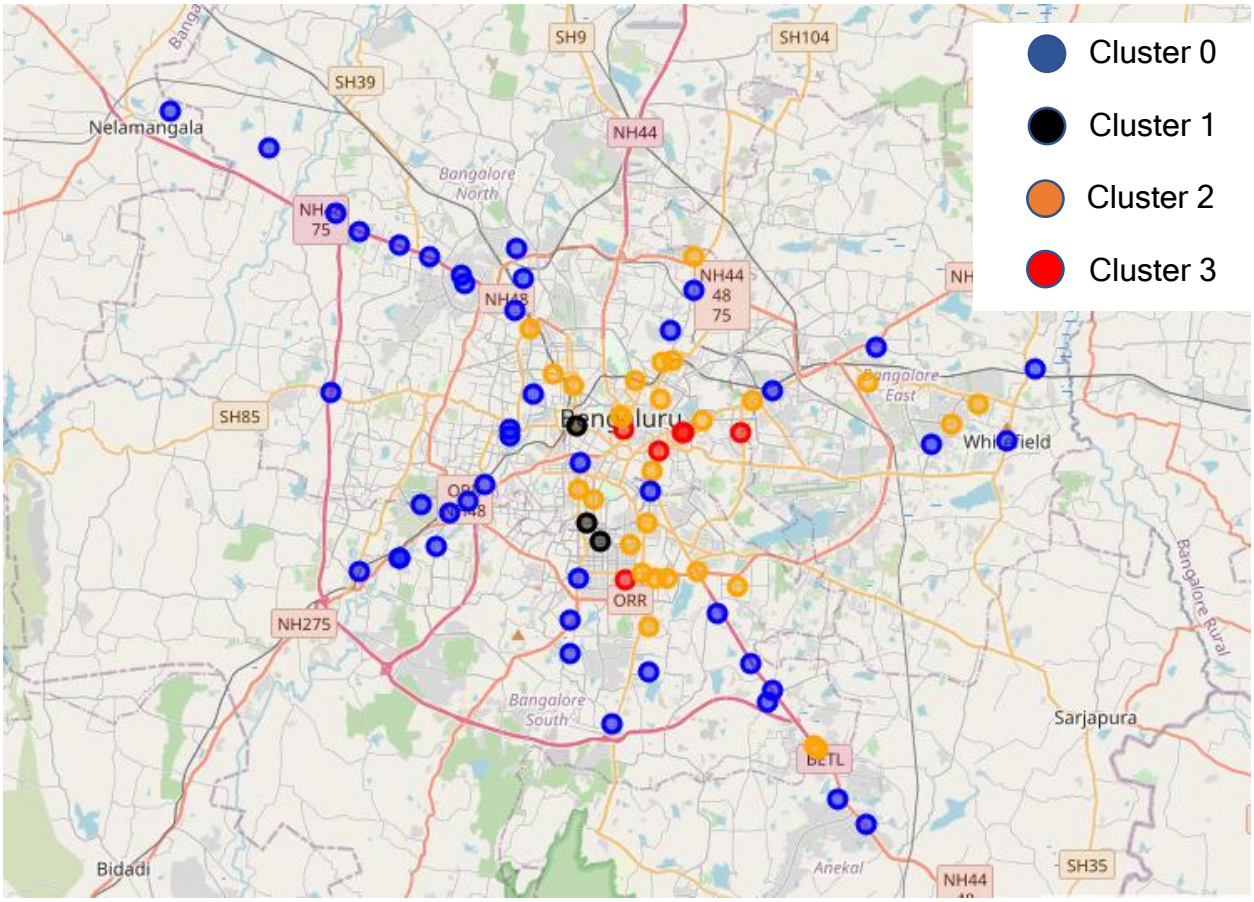


Figure 7 Box Plot of the all the four cluster

We can compare these cluster across the categories within them.

The Bangalore map with metro station clustering has been plotted. It is shown below.



4. Results

The following conclusions were drawn from clustering

Cluster 0: Under Developed zones

Cluster 1: Entertainment and Recreation Dominated zones

Cluster 2: Professional dominated zones

Cluster 3: Developed zones with proper distribution of venues.

The following tables represents the station names with their respective cluster.

Sl.No	Cluster 0	Cluster 1	Cluster 2	Cluster3
1	BIEC, Bangalore	Southend Circle, Bangalore	Nagawara, Bangalore	Ragigudda Temple, Bangalore
2	Whitefield, Bangalore	Jayanagar, Bangalore	Jayadeva Hospital , Bangalore	Vellara Junction, Bangalore
3	Jindal, Bangalore	Majestic, Bangalore	BTM Layout, Bangalore	Indiranagar, Bangalore
4	Arabic College, Bangalore	-	ITPL, Bangalore	Trinity, Bangalore
5	Kadugodi, Bangalore	-	Silk Board Interchange,	Mahatma Gandhi Road, Bangalore
6	Nagasandra, Bangalore	-	Tannery Town, Bangalore	Cubbon Park, Bangalore
7	Dasarahalli, Bangalore	-	HSR Layout, Bangalore	-
8	Jalahalli, Bangalore	-	Pottery Town, Bangalore	-
9	Oxford College, Bangalore	-	Vydehi Hospital, Bangalore	-
10	Peenya Industry, Bangalore	-	Cantonment Railway StN	-
11	Kundalahalli, Bangalore	-	Shivajinagar, Bangalore	-

12	Muneshwara Nagar, Bangalore	-	Langford Town, Bangalore	-
13	Peenya, Bangalore	-	Sandal Soap Factory, Bangalore	-
14	Chikkabegur, Bangalore	-	Mahadevapura, Bangalore	-
15	Basapura Road, Bangalore	-	Electronics City 1, Bangalore	-
16	Yeshwanthpur, Bangalore	-	Mahalaxmi, Bangalore	-
17	Hosa Road, Bangalore	-	Electronics City 2, Bangalore	-
18	MICO Bosch, Bangalore	-	Dairy Circle, Bangalore	-
19	KR Puram, Bangalore	-	Swagath Road Cross, Bangalore	-
20	Rajajinagar, Bangalore	-	Srirampura, Bangalore	-
21	Huskur Road, Bangalore	-	Sampige Road, Bangalore	-
22	Baiyappanahalli, Bangalore	-	Swami Vivekananda Road,ba	-
23	Hebbagodi, Bangalore	-	IIMB, Bangalore	-
24	Bommasandra, Bangalore	-	Halasuru, Bangalore	-
25	Hulimavu, Bangalore	-	National College, Bangalore	-
26	KR Market, Bangalore	-	Lalbagh, Bangalore	-
27	Gottigere, Bangalore	-	Vidhana Soudha, Bangalore	-
28	RV road, Bangalore	-	City Railway Station, Bangalore	-
29	Banashankari, Bangalore	-	-	-
30	Jayaprakash Nagar, Bangalore	-	-	-
31	Magadi Road, Bangalore	-	-	-

32	Yelachenahalli, Bangalore	-	-	-
33	Hosahalli, Bangalore	-	-	-
34	Anjanapura Road Cross, Bangalore	-	-	-
35	Vijayanagar, Bangalore	-	-	-
36	Vajarahalli, Bangalore	-	-	-
37	Attiguppe, Bangalore	-	-	-
38	Deepanjali Nagar, Bangalore	-	-	-
39	Mysore Road, Bangalore	-	-	-
40	Nayandahalli, Bangalore	-	-	-
41	Rajarajeshwari Nagar, Bangalore	-	-	-
42	Bangalore University, Bangalore	-	-	-
43	R.V.C.E, Bangalore	-	-	-
44	Kengeri, Bangalore	-	-	-

5. Discussion and Conclusion

Foursquare data regarding Bangalore city is very limited. It definitely affect the results in cluster. Considering the accuracy of foursquare data, many cluster usually dominated by one to two categories. Hence there will be ample amount of scope to start new business where the competition will be very less.

In underdeveloped region the venues are very less in number, So starting the professional organization in that area will indirectly affects the other venues also. Since we considered upcoming stations also, An entrepreneurs can analyze the results he can choose the appropriate location for his business plan. The property builders can identify the future lacking of the buildings which will be suited to certain business and they can come up with planning a construction in that area.

6. References

1. https://towardsdatascience.com/classification-of-moscow-metro-stations-using-foursquare-data-fb8aad3e0e4