Coursera Capstone

# IBM Data Science Professional Certificate

## **Clustering Areas of Bangalore on the basis of number of Night Life Venues in the Area**

### By: Manash Pratim Borah



#### **1. Introduction:**

Bangalore, the capital of the state of Karnataka, is not only known as the ‘Silicon Valley of India’ but it is also the ‘Pub Capital of India. Ever since the IT boom in India, Bangalore has been flooded with people from all of India and abroad and with the advent of this young energetic group of individuals, the night life of Bangalore grew, with the establishment of numerous clubs, pubs and restaurants.

With the hope of riding the wave of young people’s advent into the city, most of the venues were constructed near the areas where population of young and office going people were high. But with time there has been a saturation in the market as many business’ have made their mark in the night life industry with multiple chains around the city like SOCIAL, House of Commons, Truffles, Sherlock’s etc.

For someone who wants to start a restaurant or other night life venues, choice of location is very critical as in can make or break them. This project can help with this problem as it will cluster different neighborhoods of the city based on the number of venues already available.

#### **1.1 Business Problem:**

The objective here is to use the KMeans algorithm to cluster the different areas based on the number of night life venues present in them. The algorithm will help to differentiate areas with High, Medium and Low density of such venues.

It can help a businessman or anyone who wants to build such venues take a small step forward by atleast showing them the areas where they can get the lion’s share of profit instead of stepping into a place where it is already crowded with old established brands

#### **1.2 Target Audience:**

This project is particularly useful to property developers and investors looking to open or invest in new night life projects in the pub capital of India. It may also be helpful to anyone who considers night life venues as a considerable factor to where they stay or where they hang out.

#### **2. Data:**

For this project, we needed:

* List of all of the areas of Bangalore
  + Sources:
    - <https://pincode.net.in/>
    - <https://www.indiatvnews.com/pincode/karnataka/bangalore/bangalore-city>
* The area’s respective Latitude and Longitude
  + Using the Geopy Library
* List of all the venues within a certain radius of the coordinates
  + Using the FourSquare API

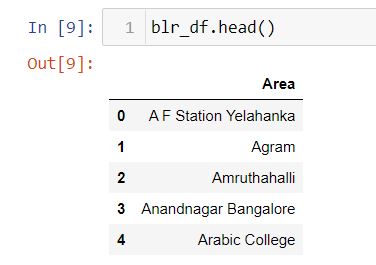
***2.1 Data Collection:***

It wasn’t a direct process to get the areas along with their coordinates as it isn’t available on the internet, so we first get a list of all the areas in Bangalore and then by the geocoder API we try to get their respective coordinates. There might be certain minor flaws as the exact coordinates might be difficult to get via the API and in certain examples, I had to manually change the coordinates to their correct ones.

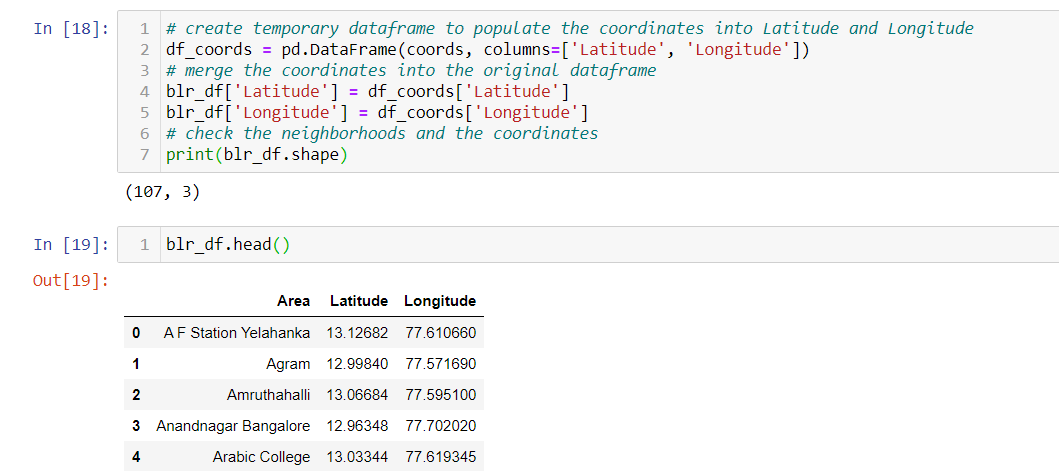
1. First, we make a csv file with all the areas of Bangalore we could find from the websites mentioned above, I did it simply by copying the



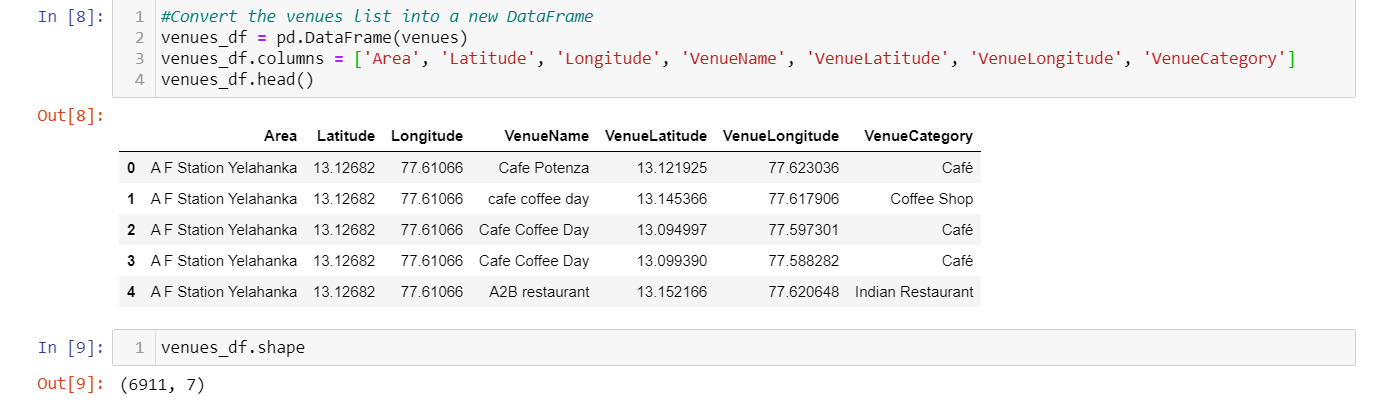
1. We then read this file in Jupyter using Pandas



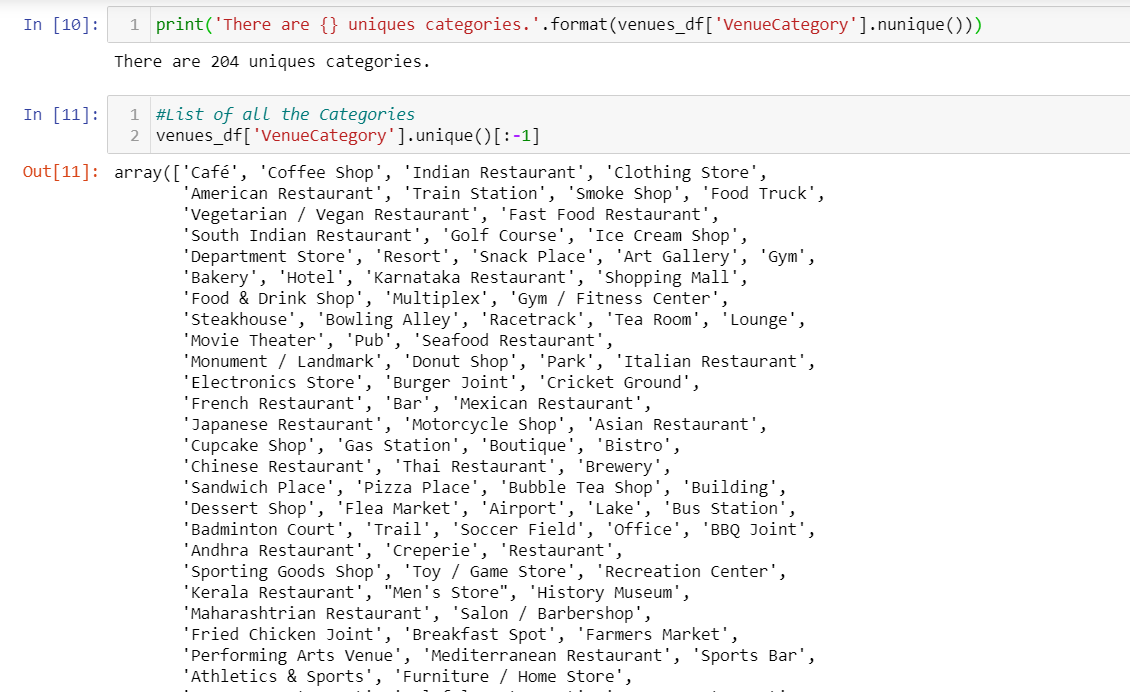
1. We then use the geocoder API to get the coordinates of all the respective areas.
2. We then convert the coordinates into a dataframe and copy the columns of this dataframe into the main dataframe.



1. We then store this to a csv file and read that csv as the main csv for the project.
2. After getting this lot of data another data that had to be collected were the venues around the areas of Bangalore, that was done using the Foursquare API.
3. For the foursquare API we have to enter our CLIENT\_ID and CLIENT\_SECRET.
4. We then have to send a GET request to their server and they’ll return a json file and the entries have been appended to a list called ‘venues’.



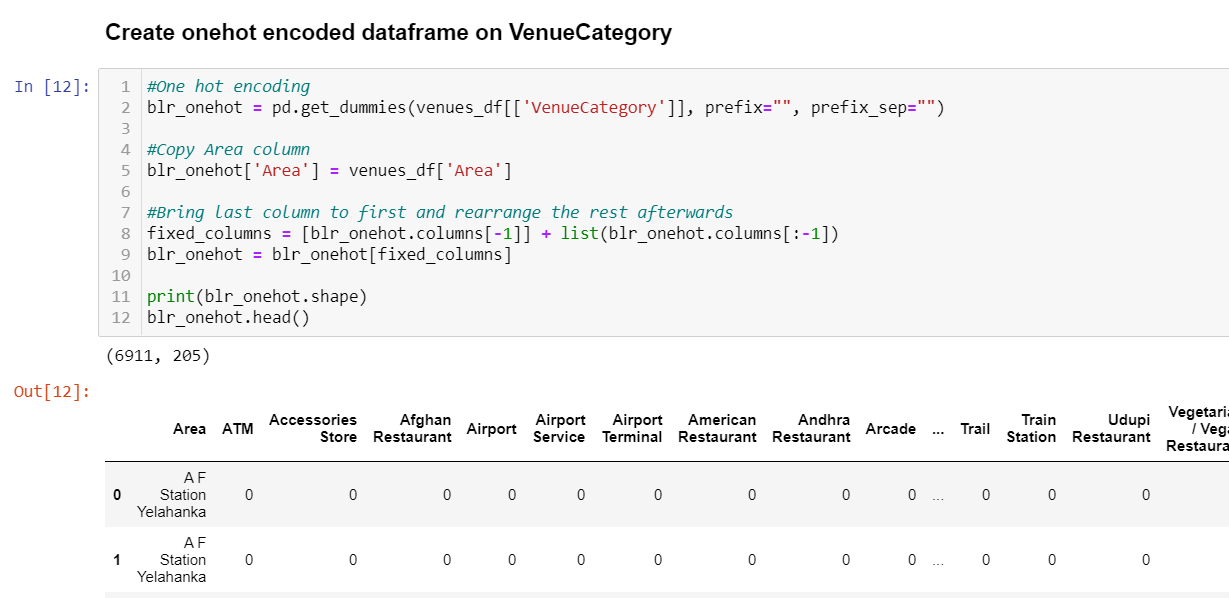
1. There are 204 unique venue categories



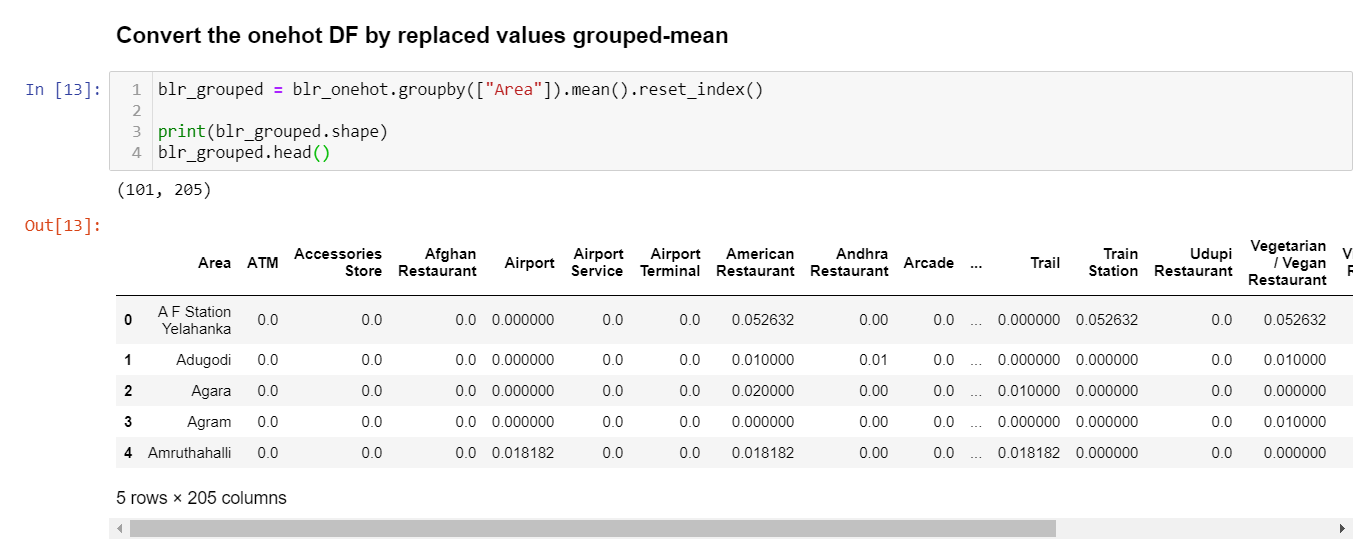
1. Now we have all the data that is necessary, after this we’ll be cleaning and feature engineering as per the problem

#### **3. Methodology:**

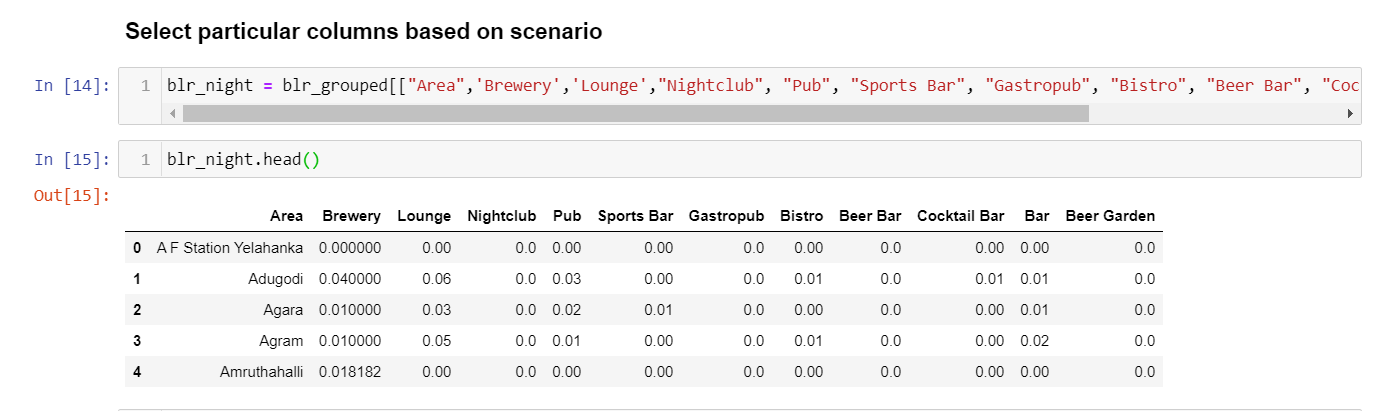
1. For the first step we have to one-hot encode our dataframe to get categorical data into numerical.



1. Then we group the dataframe on the ‘Area’ column and then take the mean of all the entries.



1. From the list of categories, we now select all the necessary ones, as our problem statement includes night life venues, we select all the night life related categories we could find and create a new dataframe.



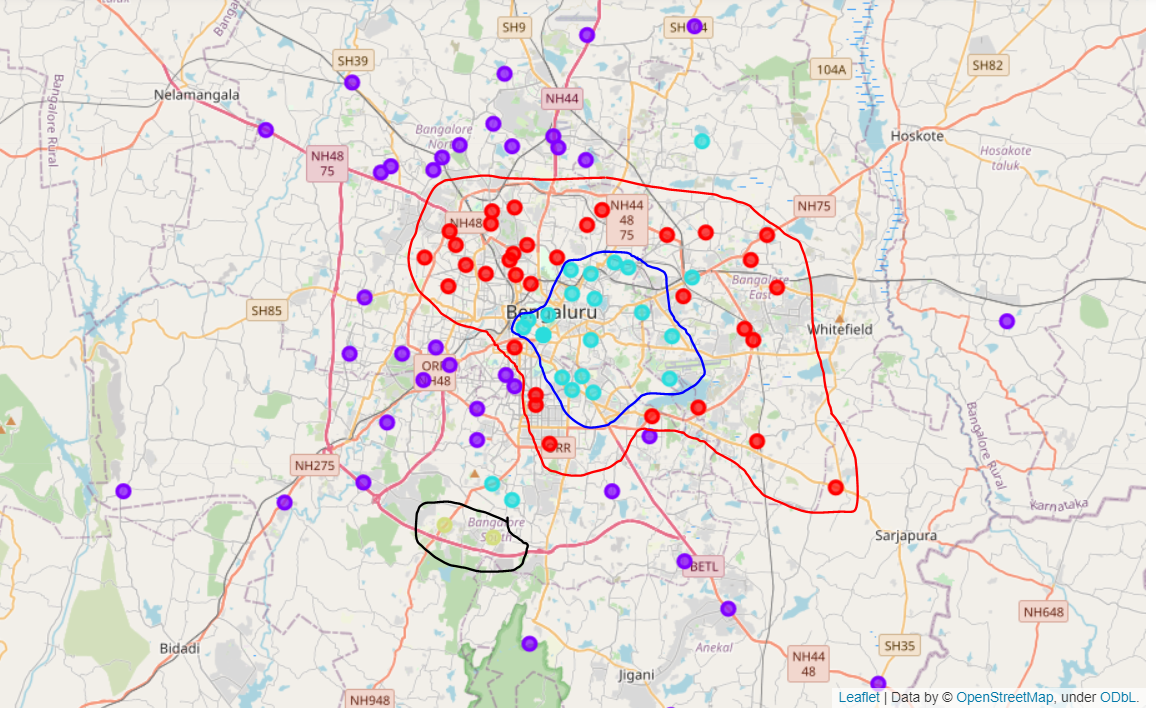
1. We now make use of K-Means algorithm to cluster the areas into 4 clusters. We selected 4 as we could find better results that were much better interpretable. We also add the cluster labels from the labels\_ returned from kmeans.



1. After this we merge it the foremost dataframe to get the coordinates of all the areas.

#### **4. Result:**

For the results we will plot the clusters on the Bangalore map using folium library.

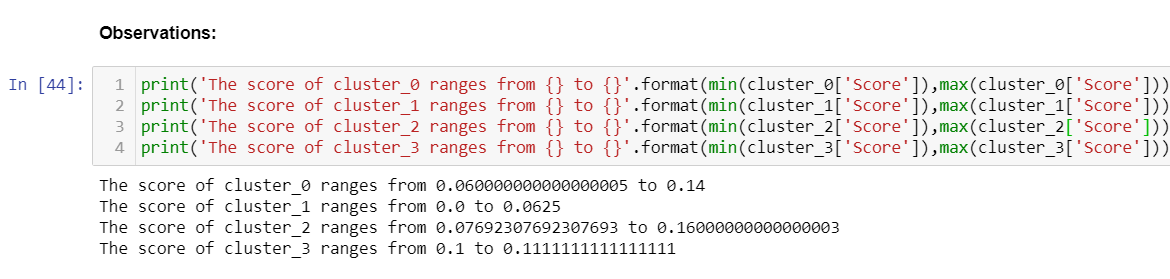


As from this map we can clearly say that there is an interior cluster, a mid-outer cluster, an outer cluster and a small cluster towards the lower left of the map.

1. The inner cluster of cyan dots represent – Cluster 2
2. The mid-outer cluster of red dots represent – Cluster 0
3. The outer cluster of violet dots represent – Cluster 1
4. The small cluster towards lower left of map represent – Cluster 3

**4.1 Ranking by score:**

To make the analysis easier we make a ‘score’ by summing up half the value of the one-hot encoded columns for each cluster.



This shows that if we consider the maximum score of all clusters.

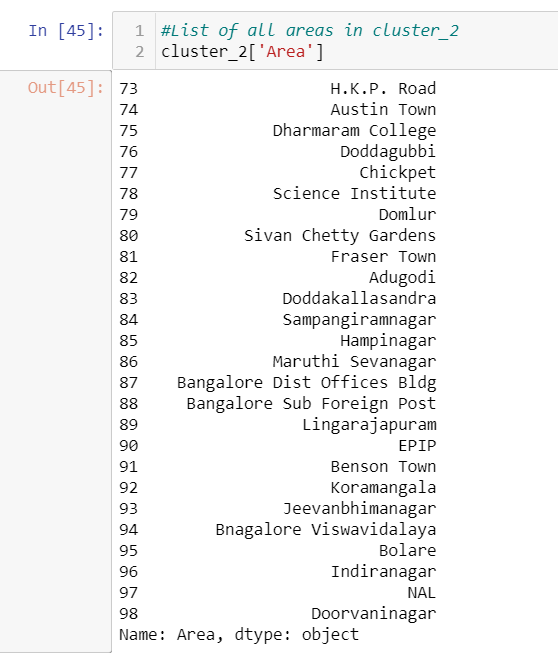
The ranking will be sort of like this

Cluster\_2 > Cluster\_0 > Cluster\_1 > Cluster\_3

1. As we know the inner cluster is represented by cluster \_2 which also has the highest max score, which inferences that the inner cluster has maximum number of night life venues.
2. The mid-outer cluster represented by cluster\_0 has the second highest max score as it is not that far from the hotspots.
3. Cluster\_1 is third as it is farther away from the city but more venues than cluster\_3
4. Cluster\_3 has the least with lowest number of venues

**4.2 Cluster Wise Analyzing**

*Cluster\_2:*

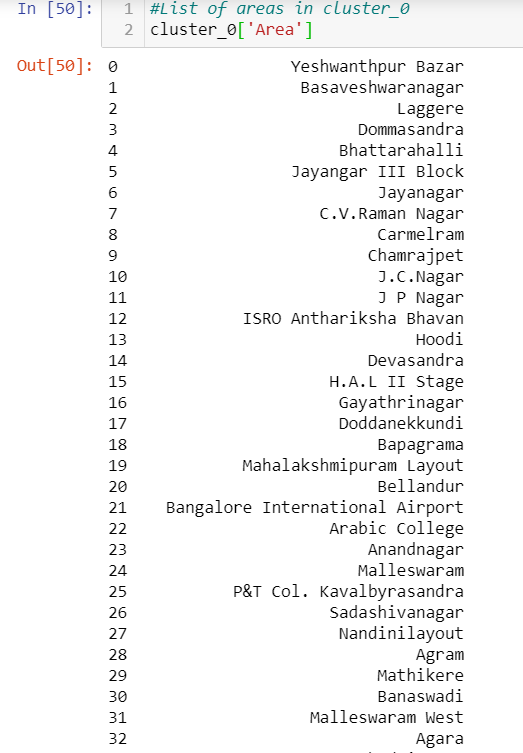


Cluster\_2 has areas like Koramangala, Indiranagar which are hotspots of night life activity.

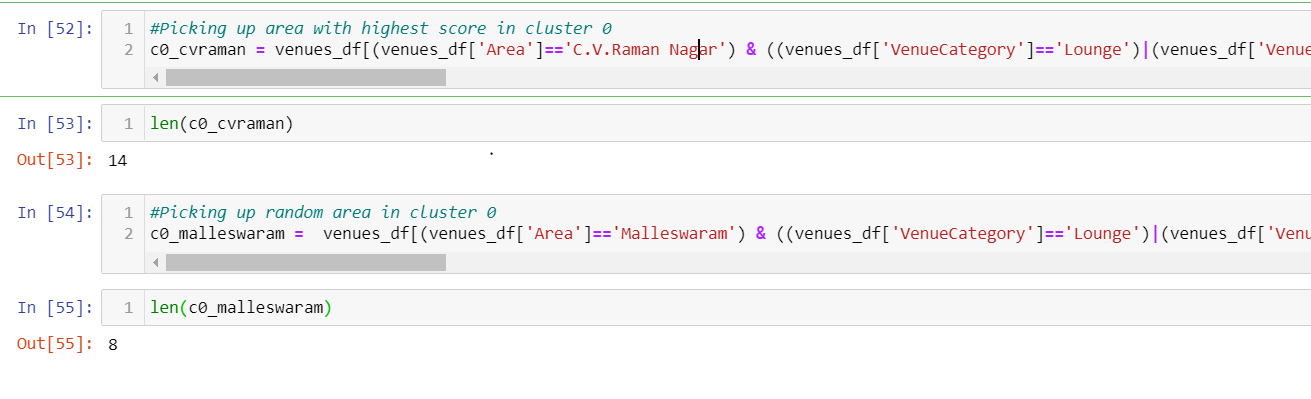


We can also check this out by the number of venues that comes under these 2 areas with Koramangala having 14 venues and Indiranagar having 16 venues which ranks among the most number of venues.

*Cluster\_0:*

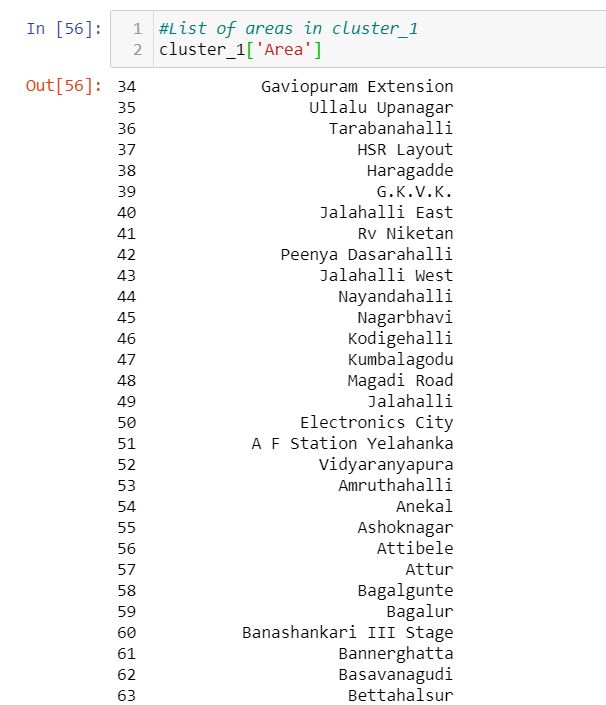


These are some of the areas in cluster\_0

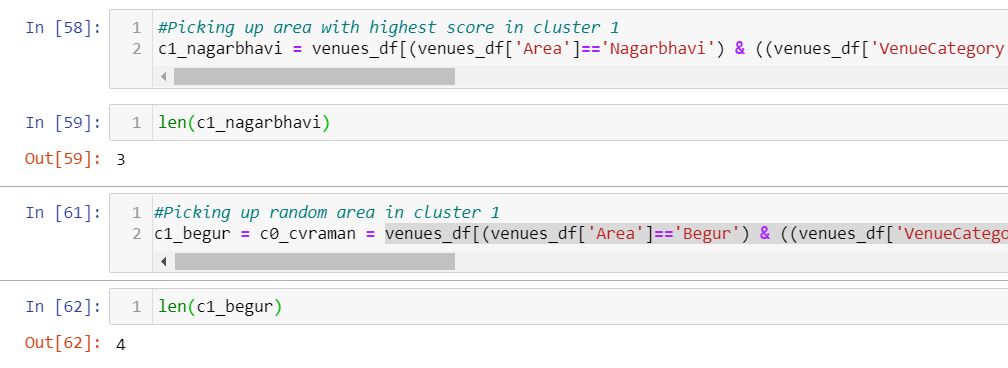


We can also verify the score of cluster\_0 by taking the area with the highest score among all in cluster\_0 and a random area from the same cluster and seeing that the number of venues is quiet high but not as high as cluster\_2.

*Cluster\_1:*

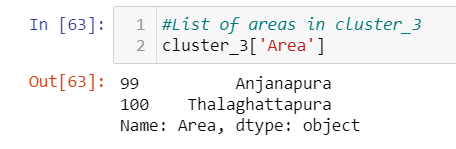


These are some of the areas in cluster\_1

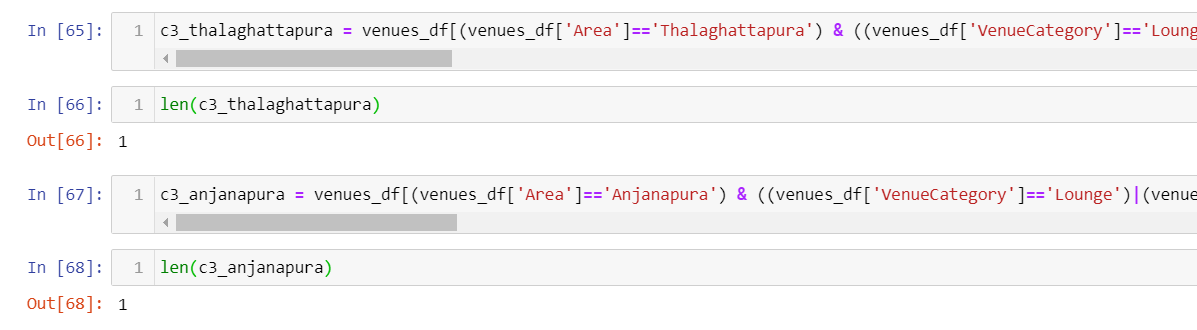


As we see these areas have lesser number of venues than cluster\_0 and cluster\_2.

*Cluster\_3:*



Cluster\_3 has only 2 areas within it and also significantly lesser number of venues.



***4.3 Final Result:***

As verified from the number of venues per cluster taking the highest scoring area and a random area from the cluster, we can say that the clustering is correct as we have predicted that according to number of venues the ranking is:

Cluster\_2 > Cluster\_0 > Cluster\_1 > Cluster\_3

|  |  |  |
| --- | --- | --- |
| **Cluster** | **Area** | **Number of Venues** |
| 0 | CV Raman Nagar | 14 |
| Malleswaram | 8 |
| 1 | Nagarbhavi | 3 |
| Begur | 4 |
| 2 | Koramangala | 14 |
| Indiranagar | 16 |
| 3 | Thalaghattapura | 1 |
| Anjanapura | 1 |

#### **5. Discussion**

In this section, I would be discussing the observations I have noted and the recommendation that I can make based on the results.

This analysis is performed on limited data. There may be some discrepancies based on coordinate data. But if good amount of data is available there is scope to come up with better results.

* There is high competition in Cluster\_2 so it is very risky to open business in these areas and is probably a saturated market with brands with strong hold on customer loyalty.
* Cluster\_0 has potential as it is not as saturated as cluster\_2 but not as far from the city heart as cluster\_1.
* Cluster\_1 is a tricky call to make as with the ever growing population of a city like Bangalore with major traffic issues, many people who are about the age of 35+ and well settled financially are moving away from the city to more peaceful areas, so it wouldn’t be a bad choice to set up a branch sub chain or a major hub where people can come out of the city as a weekend trip.
* Cluster\_3 is not a very good call as far as I can tell but if there is a influx of people there than it would be a great place to set up a new business as it is almost untouched.

#### **6. Conclusion**

What we wanted to achieve from this project was met and we could cluster Bangalore into potential clusters for the target audience and we clearly showed which cluster is a saturated market and which is not and also showed clusters that had future potential.

For future branching of this project into bigger projects we can first of all try to get more accurate coordinate data and also use other data like demographic data that includes population, income, age-group etc. to find out even better clusters and could be a full-fledged application with a bit of hardwork and patience.