Clustering of Hotels based on their proximity to Multiple Cuisines

By

B. Dinesh Bharadwaj

(as part of the IBM capstone final assignment)

The proposal

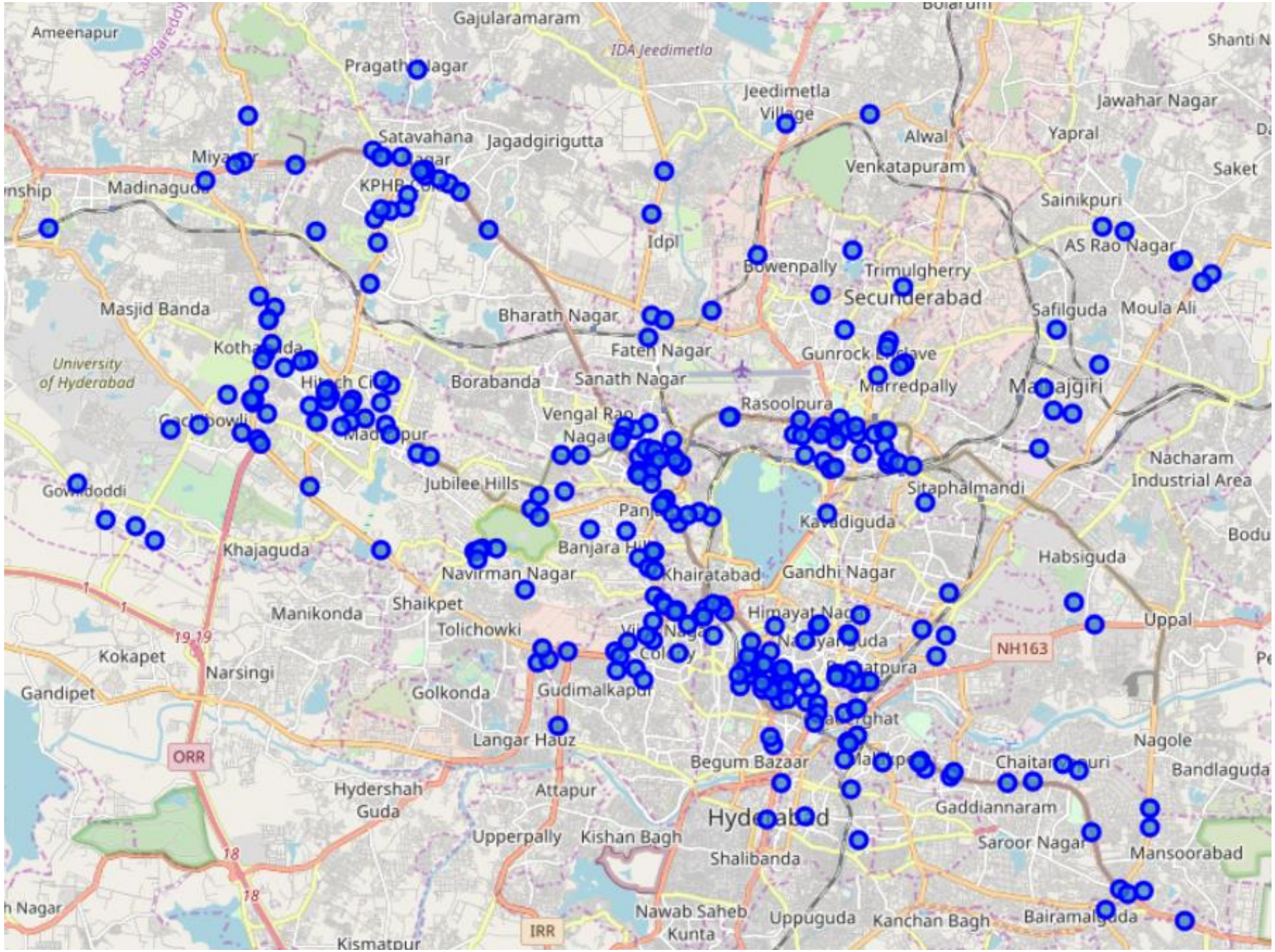
- The idea is to cluster the hotels in Hyderabad based on their proximity to the following types of Cuisines
  - (South Indian, North Indian, Cafes, Fast food, Biryani, Pizza, Burger, Chinese, Bakery and Continental)

- This should help visitors to the city be able a make a choice on their hotel and its location based on their food preferences.

Data Collection

- A list of Hotels (n=312) has been obtained using Foursquare API by searching for hotels around 10 locations (5 kms around each location) in Hyderabad spread around the city.
- Further, the list of available number of food places within 1.5 kms of each hotel is acquired using same Foursquare API.
- This data collection had to be done over the course of days because of limitations in number of queries to Foursquare that can be done per day.
- The loaded data is saved on IBM database for further use.

The hotels on the map



Data Collection

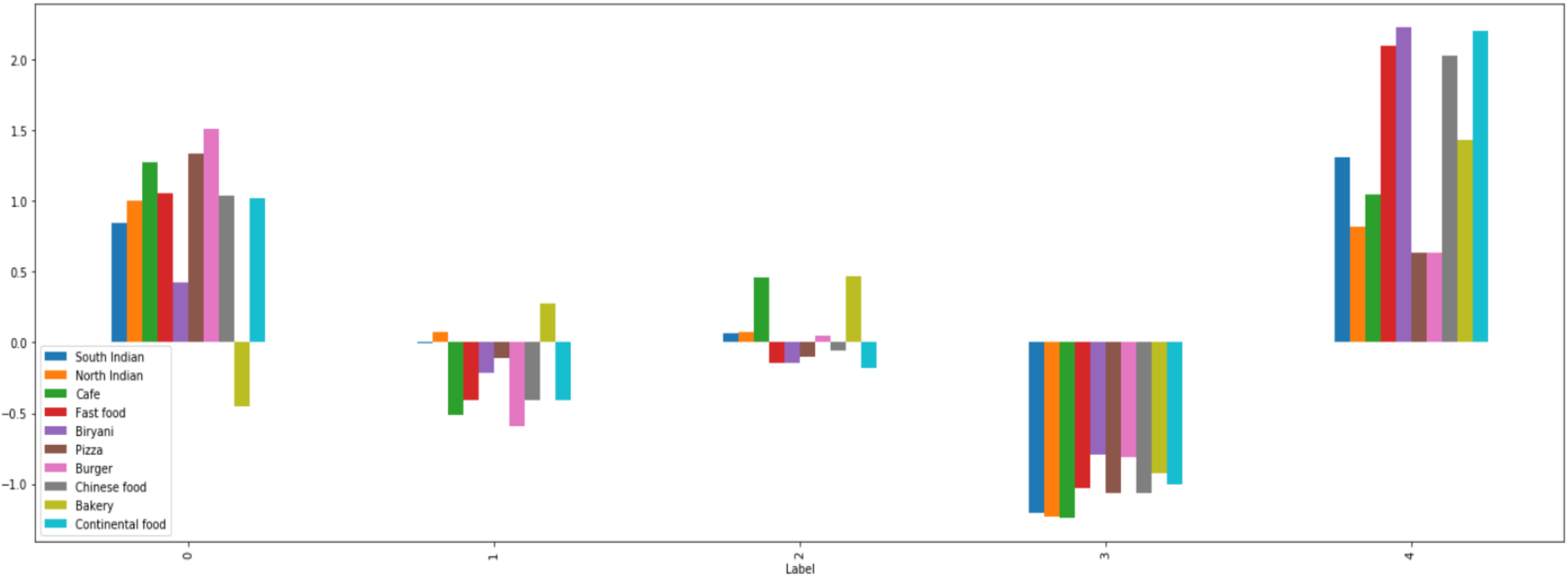
- By the end, a feature set of 312 rows and 10 columns is generated.
- Every row corresponds to a Hotel and each column corresponds to the frequency of a particular cuisine.
- The data is scaled and normalized to avoid biases

Machine Learning

- K-means clustering has been used to cluster the hotels into 5 clusters.
- sk-learn toolkit in python has been used for this purpose.
- The algorithm has been run with 12 different initializations and the best one is selected.
- The distribution of Hotels among the clusters is as follows

```
Size of cluster  0  is  58
Size of cluster  1  is  82
Size of cluster  2  is  69
Size of cluster  3  is  75
Size of cluster  4  is  28
```

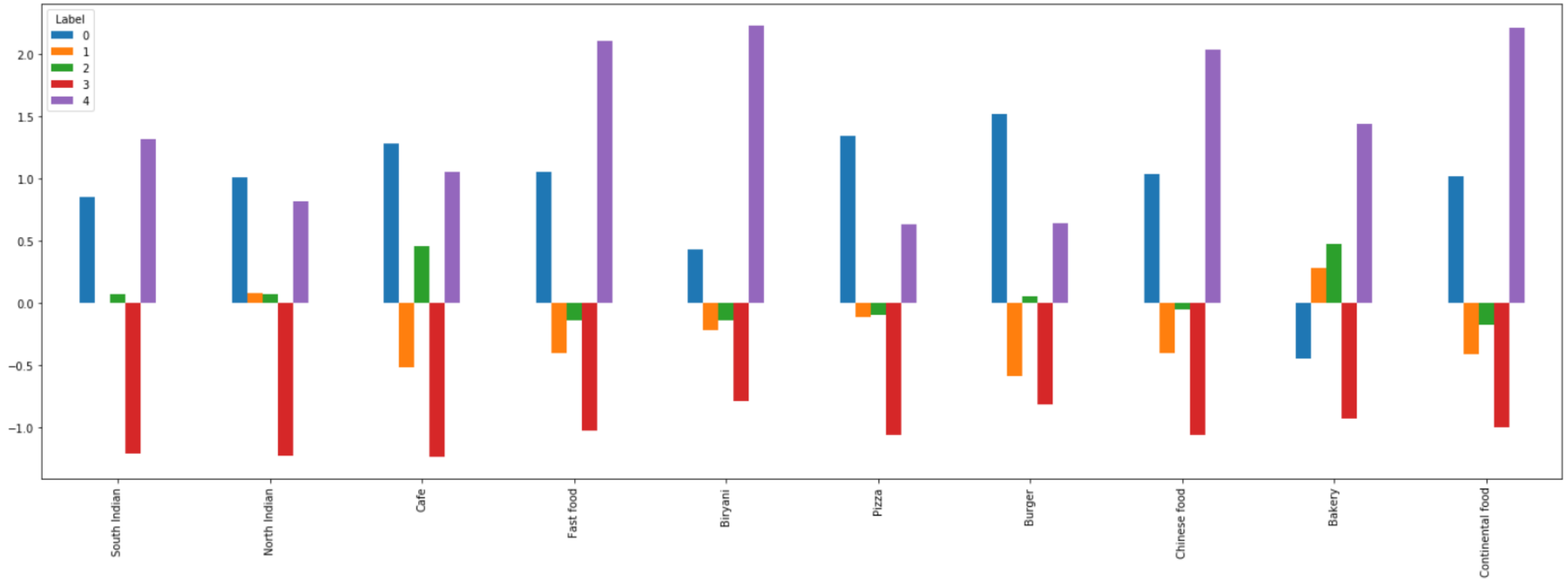
Discussion – Bar plot 1



Discussion – Bar plot 1

- Cluster 3 seems to have the lowest frequency of all types of cuisines in its proximity. Not recommended.
- Cluster 0 and 4 seem to be the best spot for most cuisines.
- Cluster 1 and 2 are pretty much in the middle of the bar. i.e they have the almost mean frequency on all cuisines making them just okay

Discussion – Bar plot 2

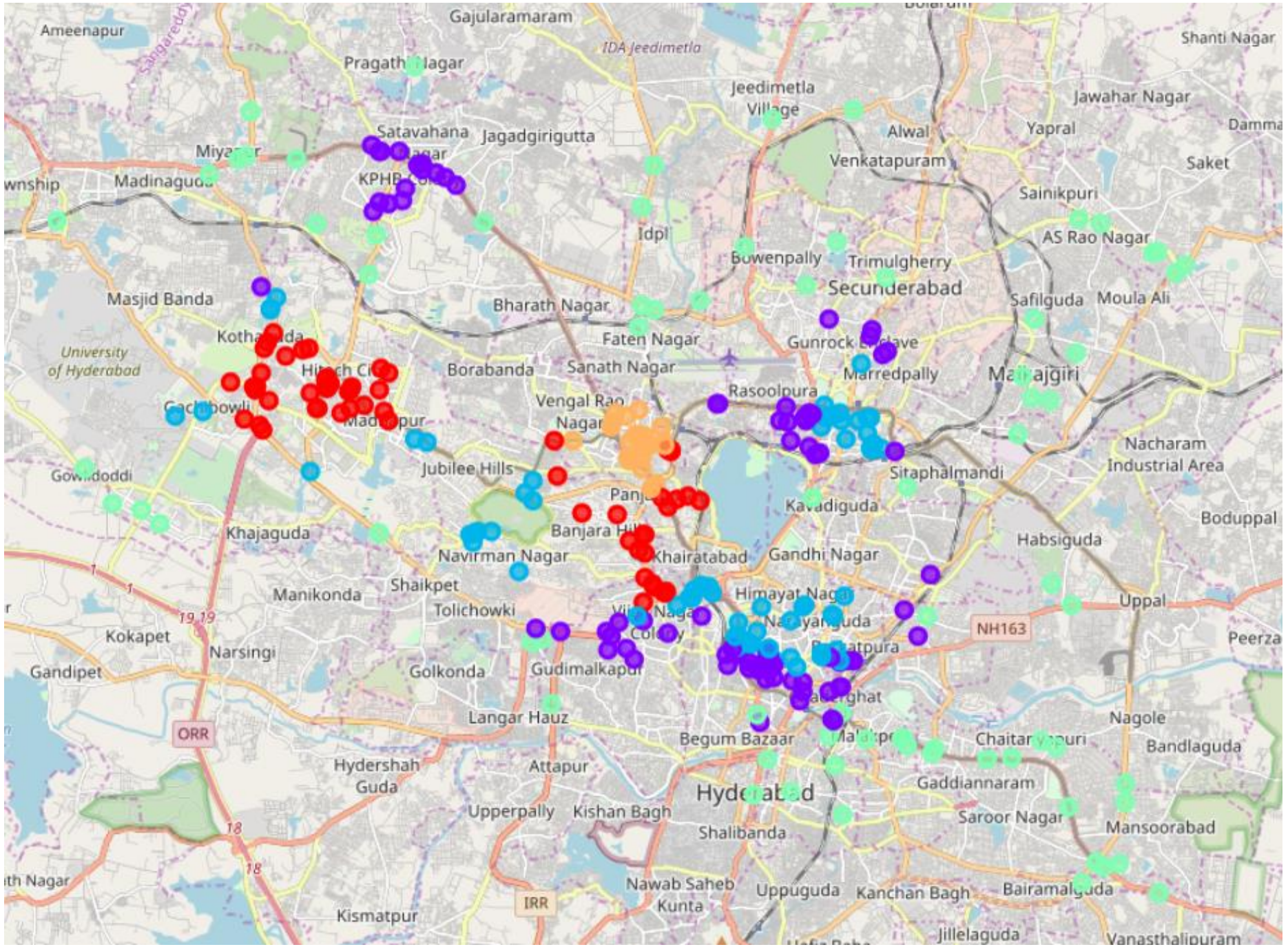


Discussion – Bar plot 2

- Cluster 3 is clearly, the worst spot for any cuisine on the list. Not recommended.
- Cluster 4 seems to be the best spot for South Indian food followed by cluster 0.
- Cluster 0 takes over cluster 4 as the best spot to have North Indian, Cafes, Pizzas and Burgers.
- When it comes to Fast food, Biryani, Chinese and Continental food, Cluster 4 dominates again by a higher margin followed by Cluster 0.
- Cluster 2 shows a high frequency for Bakery almost after Cluster 4.

The clustered hotels on the map

- the cluster 0 is represented in red
- the cluster 1 is represented in violet,
- the cluster 2 is represented in blue
- the cluster 3 is represented in radium blue
- the cluster 4 is represented in orange



Discussion

- The cluster 3 which is least recommended, seems to be located around the border of the city. (Radium blue spots on the map)
- The most recommended hotels (Cluster 4) seems to be confined to the Ameerpet region of the city.
- The next best hotels (Cluster 0) seems to be around the Hi tech city area (Jubilee hills, Madhapur, Gachibowli areas) of the city and a little around Banjara hills.
- The rest of the "Okay" hotels, with mean frequencies of all cuisines (Clusters 1 and 2) seems to be spread across the city with small spreads around Kukatpally and Abids.

Conclusion and future steps

- We've built a model that clustered hotels based on their proximity to multiple cuisines to inform the user in making a choice.
- The Next step in this project is to include the ratings of the food places and hotels into the machine learning so that the user is more informed about the hotels.