CLUSTER ANALYSIS ON THE NEIGHBOURHOODS OF WEST HYDERABAD

Chetan Sunkara

May 20, 2020

1. Introduction

1.1 Background

There are more than 2000 restaurants in *Greater Hyderabad region*. Hence, to open any kind of restaurant in the city will be a challenging task for the investor. Here we concentrate on the west region of the Hyderabad City for convenience. Choosing a restaurant type and a good spot, an entrepreneur usually carelessly relies on common sense, gut feeling and domain knowledge. Needless to say, that too often an inconsiderate decision leads to a poor income and inevitable bankruptcy. According to several surveys, up to 50% of such start-ups fail in the very first year. Let's suppose, an investor has enough time and money, as well as a passion to open the best eating spot in Hyderabad. For the past 5 years we have seen a significant increase in the business related to canteens and restaurants. So, we will have a better insight if we can answer the following.

- *What type of restaurant would it be?*
- *What would be the best place for it?*

What if there is a way to cluster city neighbourhoods, based on their near-by restaurant similarity? What if we can visualize these clusters on a map? What if we might find what type of restaurant is the most and least popular in each location? Equipped with that knowledge, we might be able to make a smart choice from a huge number of restaurant types and available places. Let us allow machine learning to get the job done. Using reliable venue data, it can investigate the city neighbourhoods, and show us unseen dependencies. Dependencies that we are not aware of.

1.2. Problem statement

Choosing a type of restaurant based on the area where to open it is a challenging task for the investor. Performing analysis to get insight on the geographical properties of each neighbourhood may help determine the place and type of the restaurant.

1.3. Target audience

The investors, entrepreneurs, and chefs interested in opening a restaurant in Hyderabad, who may need a piece of objective advice of what type of restaurant would be more successful and where exactly it should be opened.

2. Data Acquisition and Cleaning

2.1. Data Source

Initially we have to go through the selection of the neighbourhoods in west Hyderabad. Collection of the information for the selected neighbourhoods can be done by using *Geopy* and *Foursquare API*. To visualize the information, we use *Folium* library.

Using *Foursquare API*, collect the top 100 restaurants maximum and their categories for each neighbourhood within a radius 1500 meters. Group collected restaurants by location and by taking the mean of the frequency of occurrence of each type, preparing them for clustering. Cluster restaurants by k-means algorithm and analyse the top 10 most common restaurants in each cluster. Visualize clusters on the map, thus showing the best locations for opening the chosen restaurant.

2.2. Feature Selection

After collecting the coordinates for the particular neighbourhood in west Hyderabad, we have a data frame of 14 neighbourhoods and 3 features.

	Neighborhood	latitude	longitude
0	Gachibowli	17.443622	78.351964
1	Lingampally	17.488050	78.316068
2	Hitech City	17.469814	78.385378
3	Madhapur	17.440858	78.391629
4	KPHB Colony	17.492207	78.397364
5	BHEL	25.351179	69.391200
6	Chandanagar	17.487298	78.332214
7	Miyapur	17.498161	78.356763
8	Madinaguda	17.491795	78.342306
9	Masjid Banda	17.467037	78.341517
10	Kothaguda	17.458705	78.363880
11	Nizampet	17.497656	78.392507
12	Hafeezpet	17.482400	78.363014
13	Kondapur	17.458791	78.373056

Fig 2.1 Data frame of Neighbourhoods

Later we will collect the features of the top 100 venues around 1500 meters of range for each neighbourhood with the help of Foursquare API. Then a data frame is constructed with the features as *Neighbourhood*, *Neighbourhood latitude*, *Neighbourhood longitude*, *Venue*, *Venue latitude*, *Venue longitude* and *Venue category*. Sample of the data frame with venues is as follows.

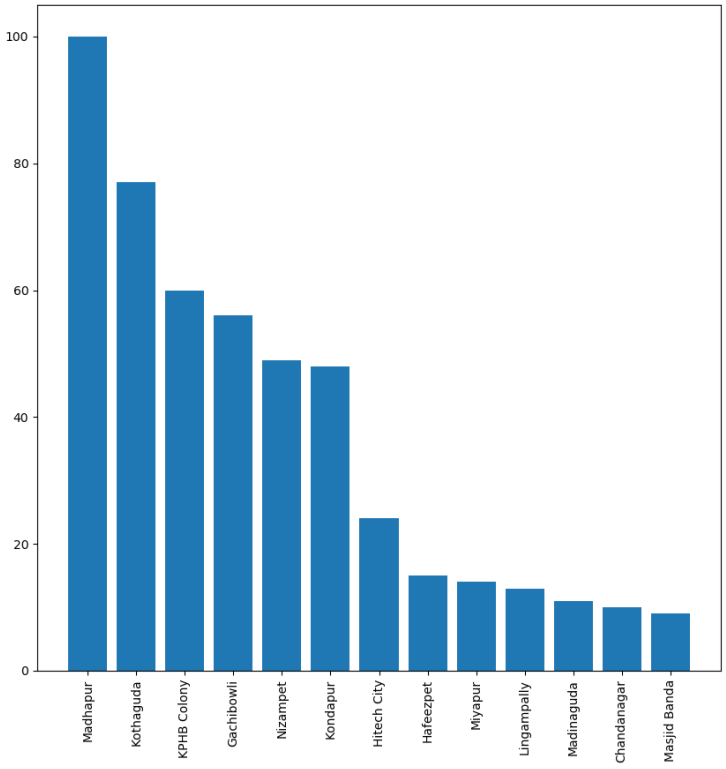
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Gachibowli	17.443622	78.351964	Absolute Barbecues	17.442922	78.357302	BBQ Joint
1	Gachibowli	17.443622	78.351964	Karachi Bakery	17.442930	78.355336	Bakery
2	Gachibowli	17.443622	78.351964	Chettinaduvilas	17.442858	78.356053	Food Truck
3	Gachibowli	17.443622	78.351964	Aviyal	17.441939	78.357185	Vegetarian / Vegan Restaurant
4	Gachibowli	17.443622	78.351964	creamstone	17.442998	78.355475	Ice Cream Shop

Fig 2.2 Sample rows of venues

3. Exploratory Data Analysis

3.1. Frequency of venues in each neighbourhood

Let us explore the venues in each neighbourhood. We can visualize the frequency of venues and make note of the top 5 neighbourhood for further analysis.




```
Madhapur        100
Kothaguda        77
KPHB Colony      60
Gachibowli       56
Nizampet         49
Kondapur         48
Hitech City      24
Hafeezpet        15
Miyapur          14
Lingampally      13
Madinaguda       11
Chandanagar      10
Masjid Banda      9
```

Fig 3.1 Frequency of venues in each neighbourhood

Venues in *Madhapur*, *Kothaguda*, *KPHB colony*, *Gachibowli*, *Nizampet* and *Kondapur* have significantly more venues than other neighbourhoods. This analysis will help us to name the clusters and explore the clusters further with the proportionate weightages. Totally we got 486 venues in 14 neighbourhoods of the west Hyderabad region.

3.2. Visualisation of the Venues

Apart from the frequency bar chart the visualization of the venues in the west Hyderabad region in the map will appeal to watch the density of the observations.

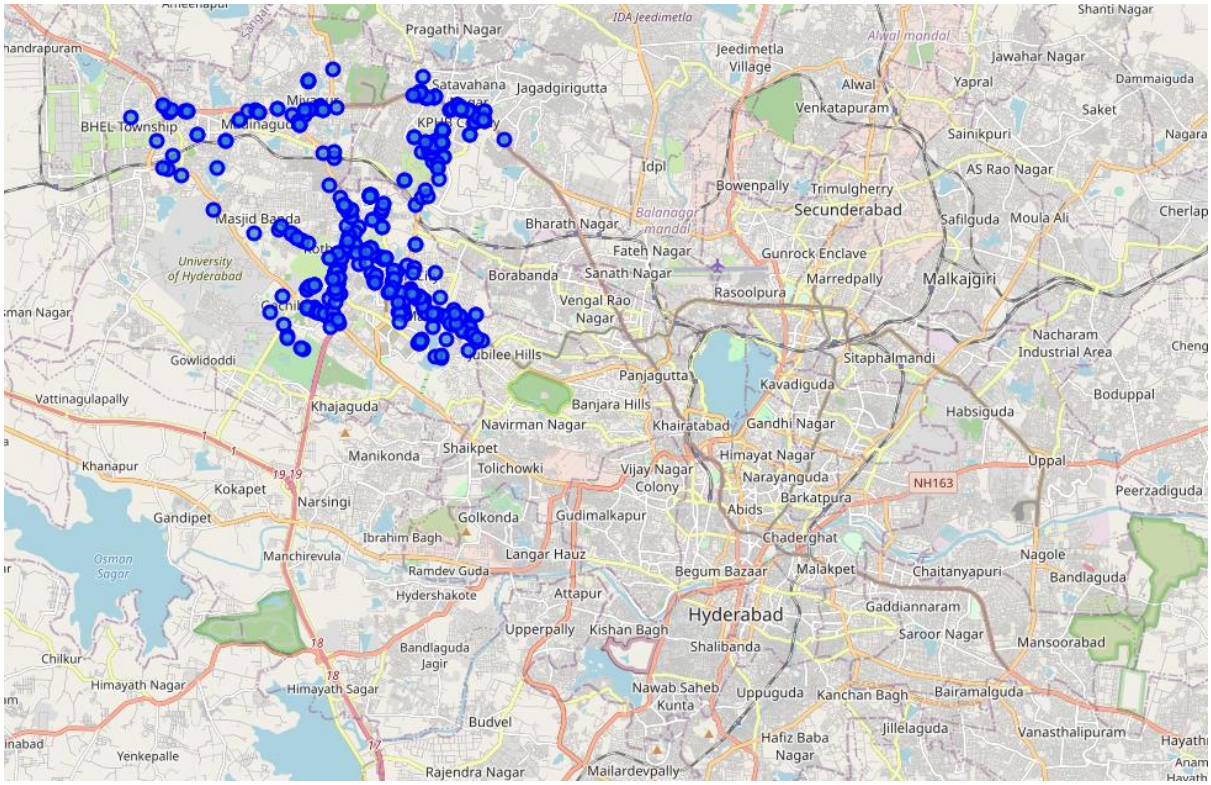


Fig 3.2 Visualization of density in the neighbourhoods

In the above geographical visualization, the blue markers indicate the venue locations. They are concentrated towards the west region of Hyderabad in accordance with our analysis. Later while modelling, we are going to use clustering, an unsupervised learning technique to find the hidden insights in the above data. The neighbourhoods will be visualized on the map similarly, according to the cluster they belong. We will indicate different cluster with different colour to differentiate them.

4. Modelling

4.1. Cluster analysis

Clustering comes under unsupervised learning type of machine learning. Unsupervised learning is a machine learning technique, where you do not need to supervise the model. Instead, you need to allow the model to work on its own to discover information. It mainly deals with the unlabelled data and algorithms allows you to perform more complex processing tasks compared to supervised learning. They find all kinds of unknown patterns in data and help you to find features which can be useful for categorization. There are three types, the clustering techniques are classified into. They are

1. Partition based
2. Hierarchical
3. Density based

In our analysis we are going to use *k-means* clustering technique to dig insights from the data collected.

4.2. K-means Clustering

K-means is one of the simplest unsupervised learning algorithms that solves the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.

The main idea is to define k centres, one for each cluster. These centroids should be placed in a smart way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycentre of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words, centroids do not move any more.

Finally, this algorithm aims at minimizing an *objective function*, in this case a squared error function. The objective function is a chosen distance measure between a data point x_i and

the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centres.

4.3. Application of the clustering model

We are going to apply k- means clustering and see the performance. Before we get into clustering analysis, we have to do one-hot encoding. In this, we will calculate the frequency of venue category with each neighbourhood by grouping with neighbourhood feature and also identify the top 10 common venues in each neighbourhood. Now, we can fit the k-means clustering model to the one-hot encoded data frame. The sample data of the one-hot encoded is shows below in fig 4.1.

	Neighborhood	Afghan Restaurant	American Restaurant	Arcade	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Badminton Court	Bakery
0	Gachibowli	0	0	0	0	0	0	1	0	0
1	Gachibowli	0	0	0	0	0	0	0	0	1
2	Gachibowli	0	0	0	0	0	0	0	0	0
3	Gachibowli	0	0	0	0	0	0	0	0	0
4	Gachibowli	0	0	0	0	0	0	0	0	0

Fig 4.1 Sample one-hot encoded data

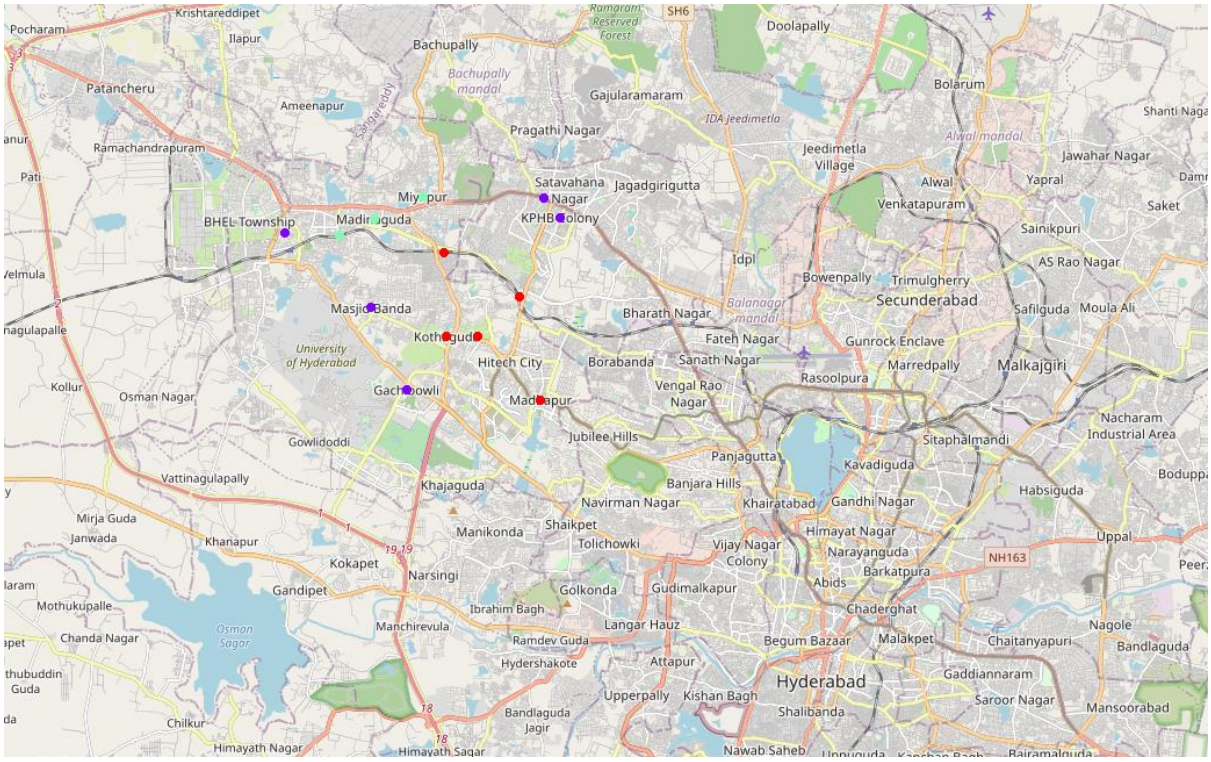


Fig 4.2 Cluster visualization

The cluster categories are allocated to each neighbourhood. We can visualize the clustering of neighbourhoods in the map as shown above in fig 4.2. The markers are colour coded according to the cluster.

Cluster Labels	
0	264
1	187
2	35

Fig 4.3 Cluster labels and venue frequency

The three clusters gave the above result and the performance is satisfactory. Highest number of venues fall into cluster 0 and second highest into cluster 1. The cluster details are seen as follows.

Madhapur	100	KPHB Colony	60		
Kothaguda	77	Gachibowli	56		
Kondapur	48	Nizampet	49	Miyapur	14
Hitech City	24	Lingampally	13	Madinaguda	11
Hafeezpet	15	Masjid Banda	9	Chandanagar	10

Cluster 0 cluster 1 cluster 2

Now, we have to name the clusters based on their properties. Cluster 0 is named as "S*hopping"*, because the most common venues at each neighbourhood include Night club, Gym, Lounge, Shopping mall and coffee shops apart from restaurants. Similarly, cluster 1 is named as "*Recreational*", because apart restaurants these venues include stadium, multiplex, movie theatre, ice-cream shops, pizza place and hotel bars as the most common venues. Similarly, cluster 2 is names as "*Residential*", because apart from restaurants these venues include Departmental stores, Train stations, Yoga studio, Bus stations, and Bakery as most common venues.

5. Conclusion

Analysing the most popular restaurants in each cluster, the investor should prefer the *least* popular types as a safe choice. There is no sense in opening same type of business in the same street as competitor. But bear in mind that descending on the most common venue list we might face an absence of demand for this type of food, and open a restaurant that is not needed in this particular location. Presence of interested customers is a must for a successful business.

That is why in our recommendations we offer to stop on 10th and 9th positions. In this report we worked out a methodology to determine the best type of restaurant to open in a promising location. The cluster wise recommendations can help the investors to watch the analysis and take a decision. Recommendations, based on description of each cluster are as follows.

1. **Cluster 0 or shopping:** Madhapur and Kondapur
2. **Cluster 1 or Recreational:** Gachibowli and Nizampet
3. **Cluster 2 or Residential:** Miyapur and Chandanagar