Analyze neighbourhood data in Hyderabad

Arjun Agarwal - 18 March 2020

Introduction / Business Problem

Location is most important aspect for success of a restaurant. A not well placed location can lead to footfall and less revenue, and eventual closing down of restaurants which will result in investor loss. Hence it is highly valuable for investors to have a data driven approach while choosing a location.

The problem statement here is to identify location characteristics across Hyderabad that can be used to judge suitability of location for opening a Restaurant.

Data

As part of solving the problem, Hyderabad city will be divided in squares with the city center as the central point. From the city center point, Hyderabad sprawls to 15 KM each side. hence we will divide it in 30x30 square with each block size of 1x1 square-km. I will use geopy library to find **coordinate of center of Hyderabad** (center Hyderabad address will be Lumbini Park). Based on the central coordinate, I will divide the city in blocks using algorithm. and will determine the latitude and longitude of each block. I will use **geopy library to find neighborhood name for each block**. Once neighbourhood information is available, I will use **foursquare venue API** to get the popular venues in vicinity.

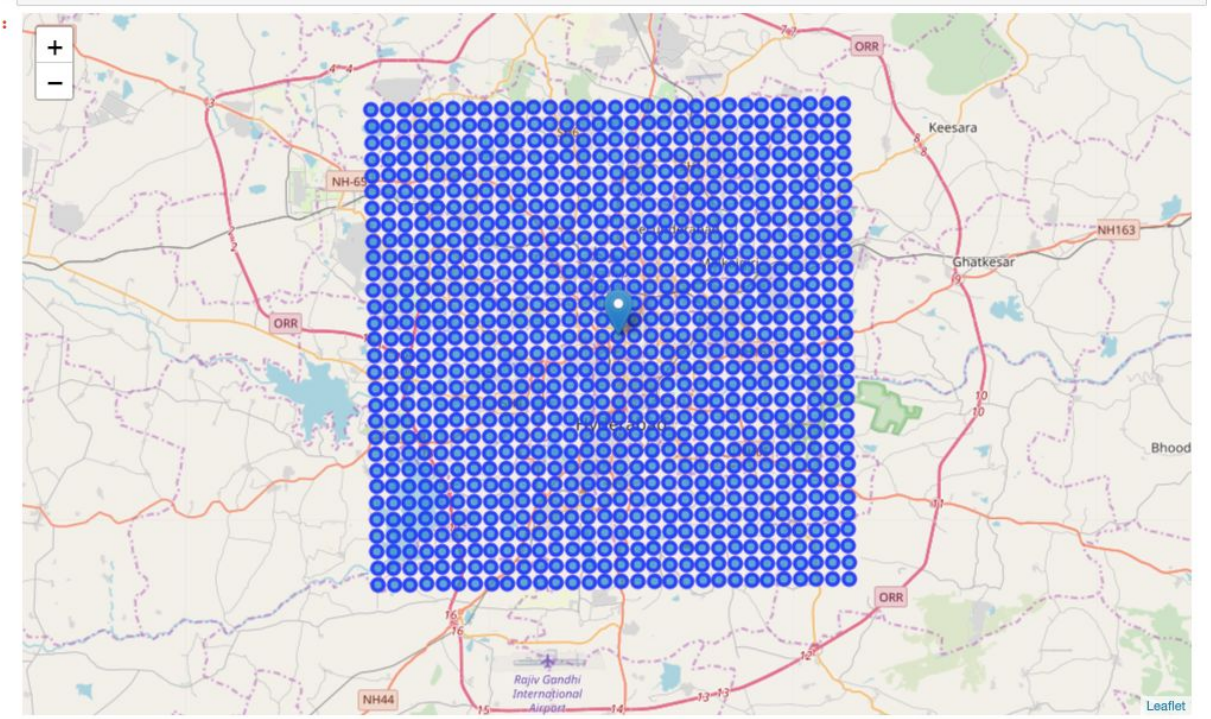
Once I have all data in place (i.e. Neighbourhood name, Latitude, Longitude, Venue Name, Venue Category etc), I will explore the data, apply ML model, evaluate the Model and will finally present results and conclusions.

Methodology

We will identify the optimal location to open a restaurant in Hyderabad. Let's first identify coordinates for Lumbini Park, The city center of Hyderabad.

Next, Let's divide the square block of 15 KM in each direction of our city center to 1x1 grid. Note that, using square is not optimal, as we will miss out venues at corner while using Foursquare venue API. We will address this problem in future versions. For now let's go with square only as the algorithm to produce square blocks is comparatively easy. We will create a block center overlapped on Hyderabad city center using Folium.

As our city center is in geographic coordinates, we will need to convert it to cartesian. Once we have a cartesian coordinate of the city center, we can use that information to compute a cartesian coordinate of 1x1 blocks. The cartesian coordinate of 1x1 block will converted back to geographic coordinate.

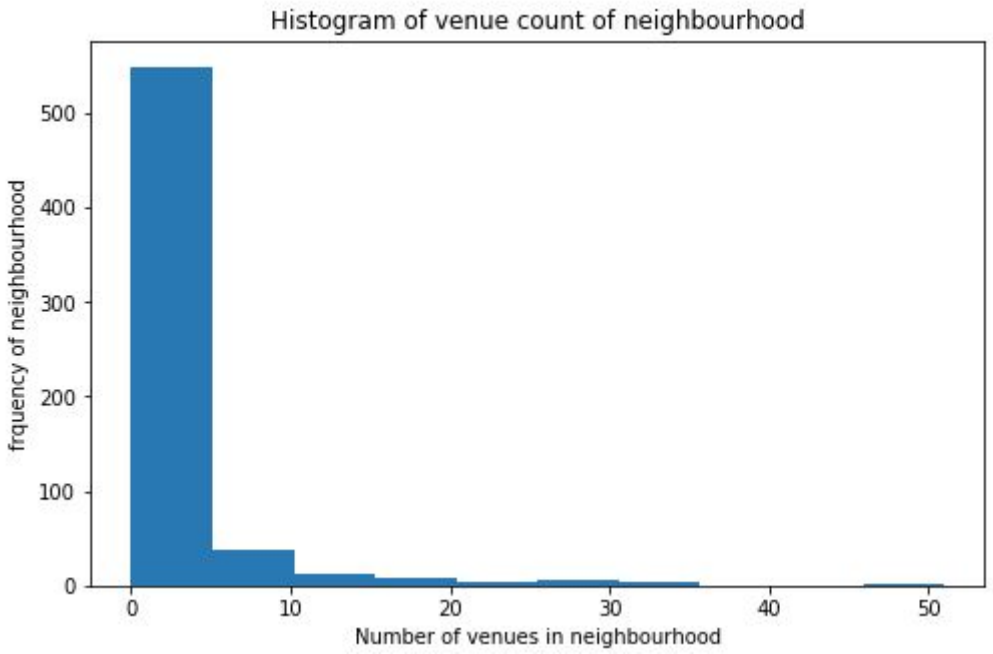


Now that we have our blocks with center coordinates, We will find our neighbourhood names of block center using geopy library. We will use longitude and lattitude information for blocks.

	Address	Latitude	Longitude
0	Marlaguda, Kavvaguda, Shamshabad mandal, Ranga...	17.272612	78.333794
1	Nagireddyguda, Moinabad mandal, Rangareddy, Te...	17.281642	78.333664
2	Nagireddyguda, Moinabad mandal, Rangareddy, Te...	17.290671	78.333534
3	Bangaliguda, Nagireddyguda, Moinabad mandal, R...	17.299701	78.333404
4	Bangaliguda, Nagireddyguda, Moinabad mandal, R...	17.308730	78.333273

Now that we have block geographic coordinates, Let's find the top 100 venues in a radius of 500 meter for each block. I will use Foursquare API for this purpose. Foursquare API sometimes does not have category information of the venue, we will drop such records afterwards.

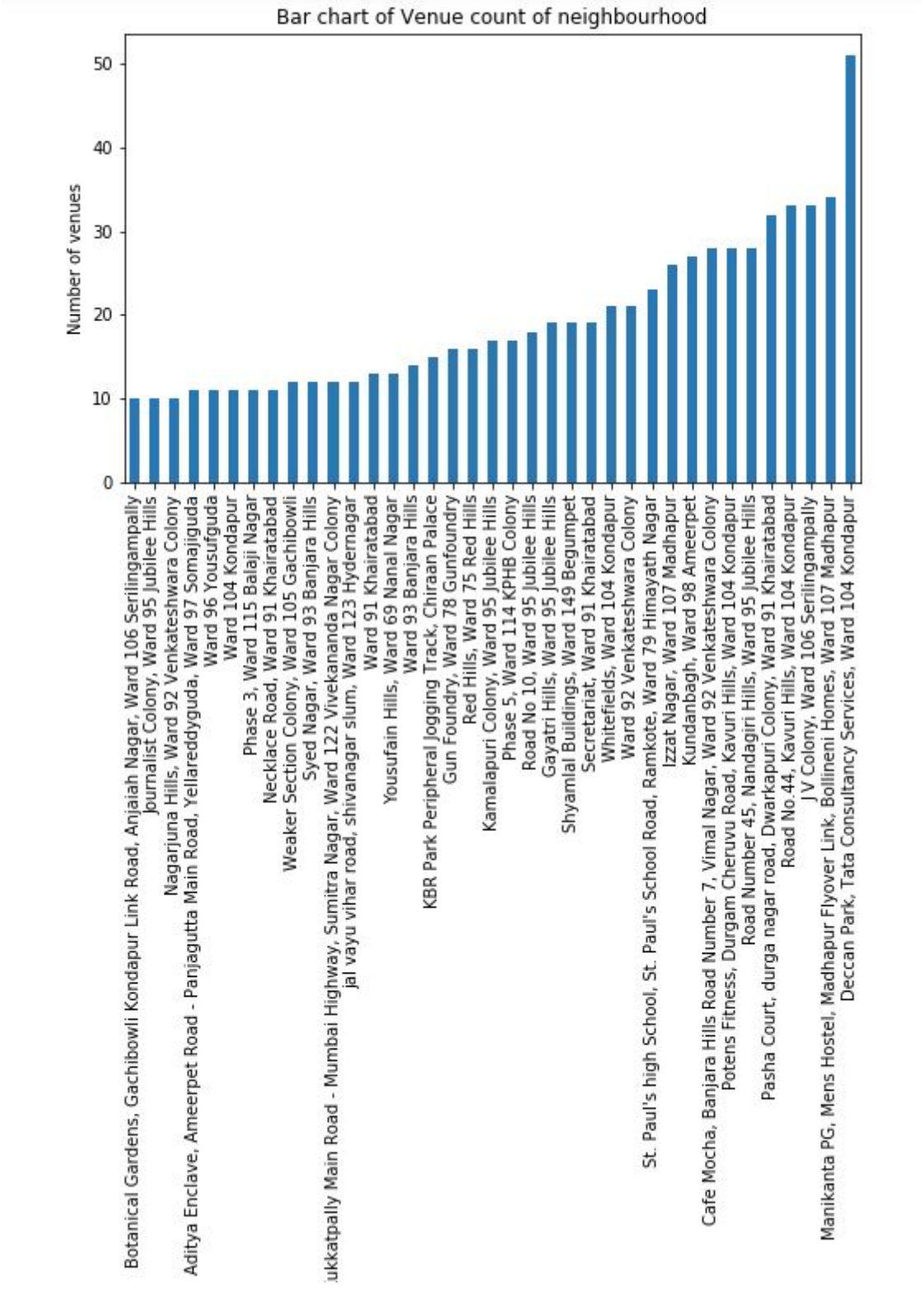
We have got 1517 venues with categories. Now, Let's see a histogram distribution to see distribution of venues neighbourhood.



From the histogram above, it's evident that most of the neighbourhood has 10 or less venues. This indicates a lack of data for such neighbourhoods and hence let's discard them now.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
71	Weaker Section Colony, Ward 105 Gachibowli, Gr...	17.417334	78.350516	Domino's Pizza	17.418600	78.348195	Pizza Place
72	Weaker Section Colony, Ward 105 Gachibowli, Gr...	17.417334	78.350516	Subway	17.419278	78.347612	Sandwich Place
73	Weaker Section Colony, Ward 105 Gachibowli, Gr...	17.417334	78.350516	Cafe Coffee Day	17.418786	78.347679	Café
74	Weaker Section Colony, Ward 105 Gachibowli, Gr...	17.417334	78.350516	Waverock Cafeteria	17.418998	78.348175	Food Court
75	Weaker Section Colony, Ward 105 Gachibowli, Gr...	17.417334	78.350516	Axis Bank ATM	17.420349	78.350220	ATM

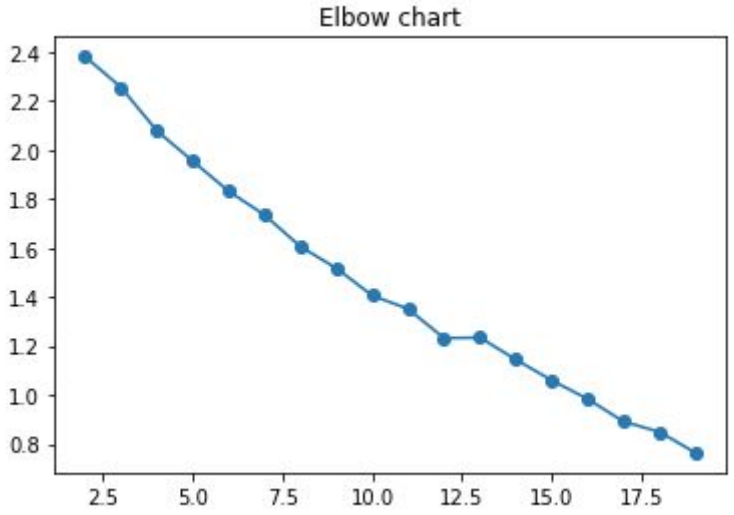
We are now left with potentially 37 neighbourhoods to choose from. Let's see the venue distribution across these neighbourhoods.



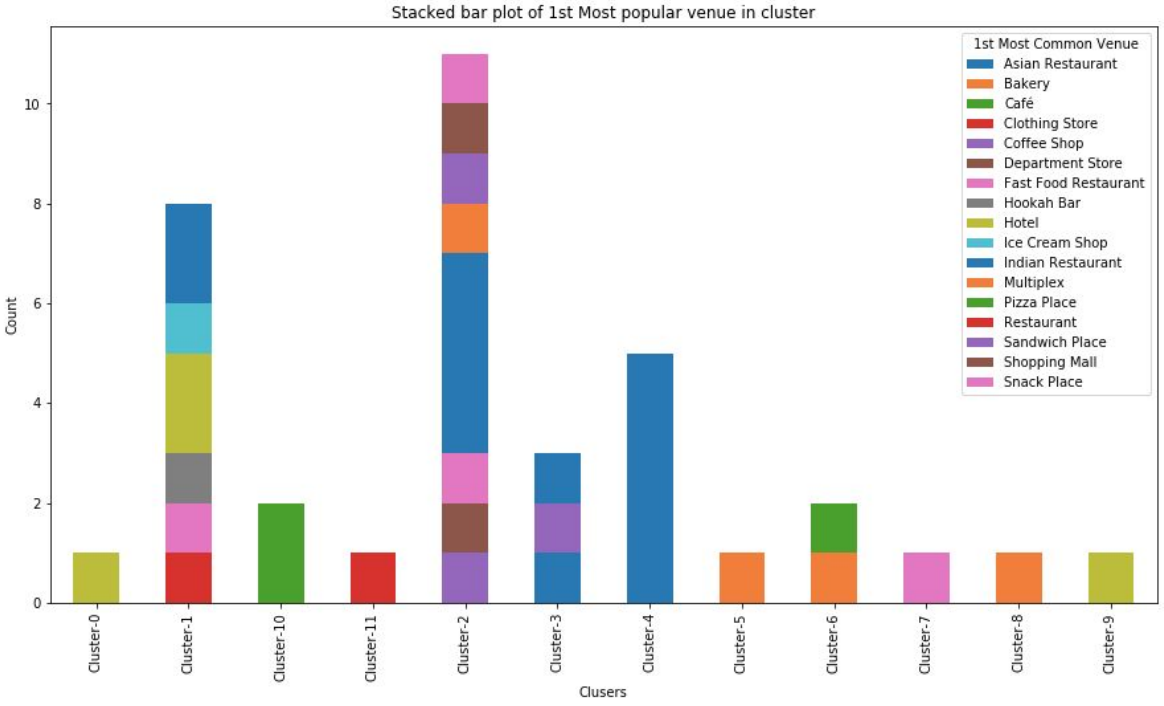
From the above chart, we can see that Kondapur and Madhapur have more venues, and Journalist colony and Serilingampally have relatively less venues.

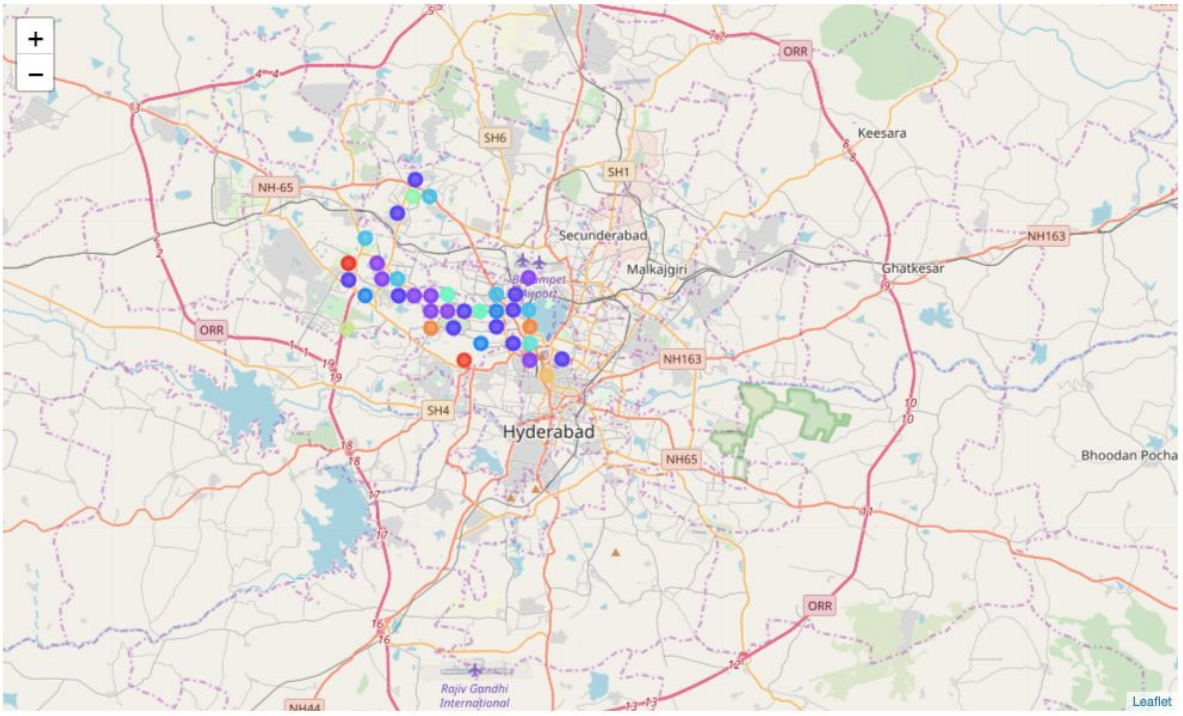
Now that we have data in good shape, we need decide on our model. Most important feature in our data are Category of venue and their neighbourhood. We need to label our neighbourhood

data. Classification and Clustering are two ways to label such data. Classification is a supervised algorithm and requires training data which we don't have. Hence we will go with the Clustering approach. We will try k means clustering algorithm and choose optimal k using elbow method.



The elbow point is visible at k=12. We will cluster our neighbourhood with k=12. And, then will analyze each cluster by creating bar plot of 1st most common venue across cluster.





Results

Investors would like to open restaurants in underserved places or venues that attract large crowds. From our analysis Hotel and Multiplex are venues that attract large crowds. Let's look at the data from these two types of clusters.

Hotel

Neighbourhood	Longitude	Lattitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Cluster
Botanical Gardens, Gachibowli Kondapur Link Ro...	78.359402	17.453577	Hotel	Bar	Nightclub	Donut Shop	Indian Restaurant	Restaurant	Italian Restaurant	Garden	Department Store	Electronics Store	0
Gun Foundry, Ward 78 Gunfoundry, Greater Hyder...	78.473168	17.391831	Hotel	Bakery	Fast Food Restaurant	Bookstore	Mobile Phone Shop	Shopping Mall	Neighborhood	Diner	Shoe Store	Fried Chicken Joint	9

Multiplex

	Neighbourhood	Longitude	Lattitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Cluster
24	Secretariat, Ward 91 Khairatabad, Greater Hyde...	78.463514	17.409773	Multiplex	Indian Restaurant	Coffee Shop	Fast Food Restaurant	Movie Theater	Platform	Hotel	Vegetarian / Vegan Restaurant	Boat or Ferry	Convenience Store	5

From both above
1. The Hotel cluster has mostly mixed types of venues at the 2nd to 10th place of popular venues.

2. The Multiplex cluster has one more movie theatre at 5th spot and a hotel at 7th spot.

Hence it would make more sense to open a restaurant at Cluster 5 neighbourhood
that is **Secretariat , Khariatabad, Hyderabad**.

Conclusion

We did an analysis of neighbourhoods in Hyderabad based on Top 10 popular venues, and gave weightage to most popular venues to identify suitable neighbourhoods.
Though we came up with an answer, there is a lot of scope of further improvement. In future we will focus on following improvement to come up with more accurate results.

1. Change blocks to hexagon are the one that is closer to a circle. This way we will avoid missing venues in square corners.
2. Introduce restaurant score while deciding suitability of opening a restaurant in a neighbourhood. The restaurant score should factor into the popularity of the venue as well.
3. Introduce DBSCAN clustering algorithm that can create clusters of arbitrary shape to give better results.

Thanks,