**Geo-spatial analysis of Delhi using Clustering (Unsupervised learning) in Python**

Aryan Agarwal

May 03, 2020

**Business Problem:**

A retail company wants to set up supermarket stores in Delhi city but is not exactly sure about which Neighborhood(s) to open the store(s) in. The chosen locations should ideally have a considerable population so that there is more store footfall & near to work centers/residential districts for easier access to a large number of citizens.

There are 2 business questions that need to be answered.

1. Which part (area) of the city should the company open the supermarket first.

2. Which Neighborhood(s) would be ideal in that part (as in point 1) setting up such a supermarket in the city.

The company would ideally prefer to open the store/s in Neighborhoods where there is a comparatively lower real estate prices (not absolutely low). But the same time, they want to choose the Neighborhoods with a high population and more number of venues, since it should result in more footfall for the store. When we consider the business problem, we can create a map and information chart where the real estate prices are placed on Delhi and each area is clustered according to the venue density.

# Background:

I have selected Delhi for my project since I am familiar with the same, being a resident of the city. Delhi district is a metropolitan with a population of roughly 5 million and 150 Neighborhoods (GHMC) . The city has a high population and population density. Being a crowded city leads the owners of shops and social sharing places in the city where the population is dense. This clustering will ensure that Neighborhoods with moderate real estate price and more number of venues will be in single clustered together and hence would be used to answer the business problem.

# Data Description:

In order to solve the business problem, I have decided to use the following data as listed below, which includes the Foursquare Location data API.

Geographical co-ordinates data of Neighborhoods in Delhi city by zip code from GitHub repository.

Source : <https://github.com/sanand0/pincode/blob/master/data/IN.csv>

Venue data for each Neighborhood in the city using Foursquare API. I included venues within a 500 meter radius from each neighborhood.

The data helps us to identify similar Neighborhoods using venues and also helps in clustering algorithm.

Geo-Json data for GHMC (Delhi Municipality) for Choropleth Maps (to show real estate prices).

**Use:**

Mapping Neighborhoods on Folium Map. Generating centers for each Neighborhood using geo co-ordinates.

The data helps us to show real estate prices on Choropleth/Folium Maps.

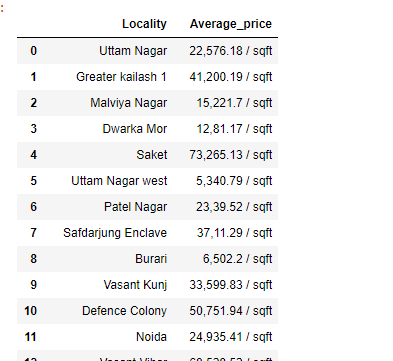
Average House prices (per square feet) for each Neighborhood in Delhi city.

Source: [https://www.makaan.com/price-trends/property-rates-for-buy-in-delhi](https://www.makaan.com/price-trends/property-rates-for-buy-in-hyderabad)

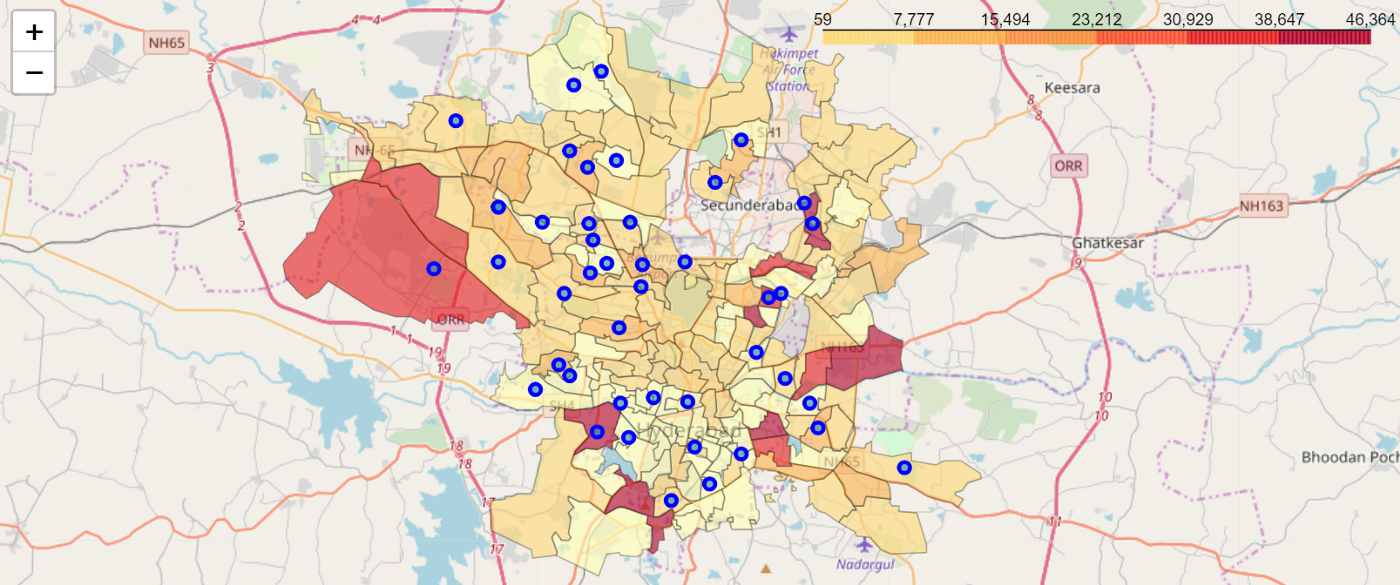
Use:

The data helps us to show real estate prices on Choropleth Maps and to identify potential Neighborhoods where stores can be opened.

# Methodology

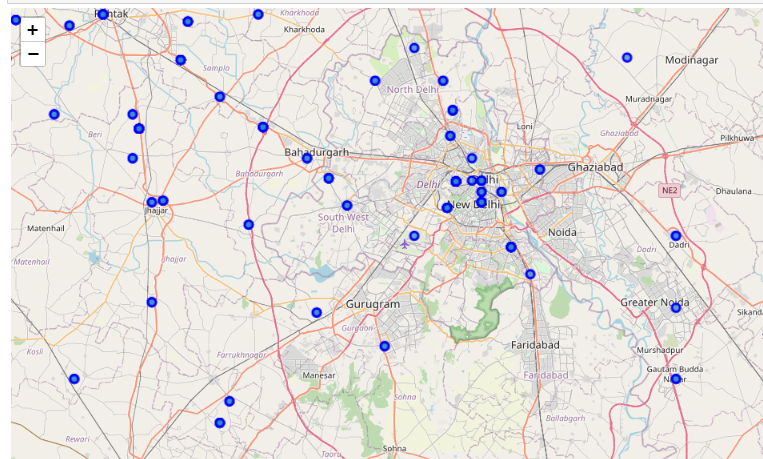
For the House prices, I have used web scraping to extract data from a house finder website in my project. A part of the table shown below.I have used python folium library to visualize geographic details of Delhi by creating a map of Delhi with Neighborhoods superimposed on top. I used latitude and longitude values to get the visual as below:

I have used python folium library to visualize geographic details of Delhi by creating a map of Delhi with Neighborhoods superimposed on top. I used latitude and longitude values to get the visual as below:



Using Delhi Geojson data (with boundary co-ordinates for Neighborhoods), I calculated the center co-ordinates for each Neighborhoods using python code & list comprehension. Then, I used Folium Library to visualize the centers on map.

Below is the image of map showing all the Neighborhoods (in blue) and their centers (as red dots).



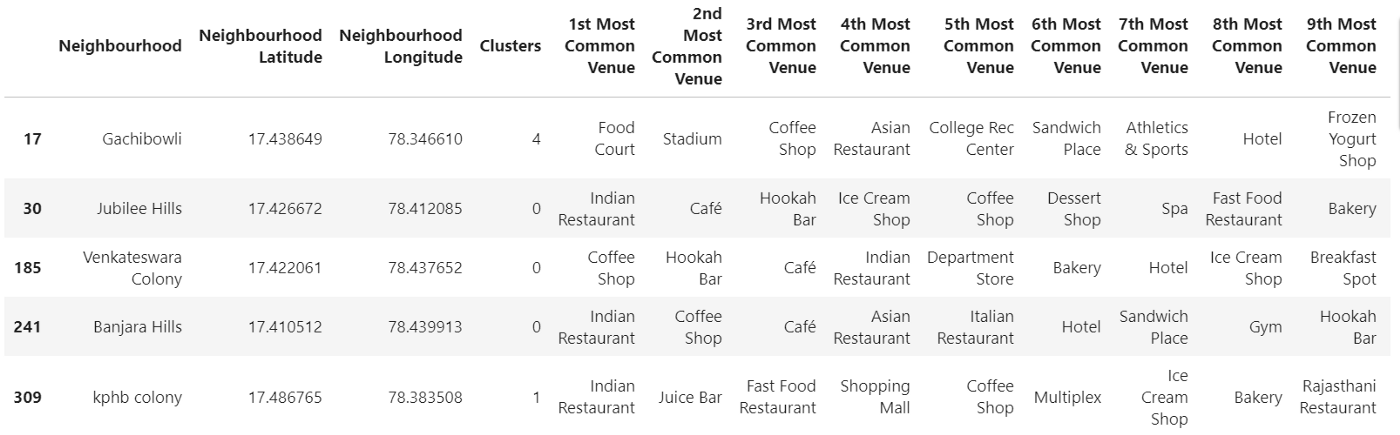
I utilized the Foursquare API to explore Neighborhoods and segment them. I kept the limit as 100 venues and the radius 1000 meters for each Neighborhood centers (calculated above) from their given latitude and longitude data. Here is a head of the list Venues name, category, latitude and longitude information from Foursquare API.



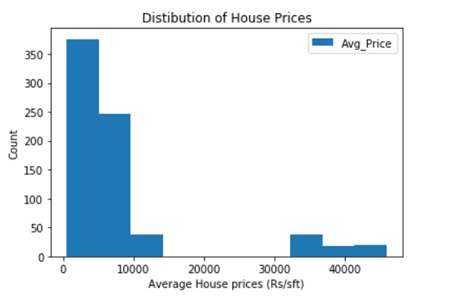
In summary of this data ~ 1400 venues were returned by Foursquare for Neighborhoods in Delhi.

Since we have some common venue categories in Neighborhoods, I have used unsupervised learning K-means algorithm to cluster the Neighborhoods. K-Means algorithm is one of the most common clustering method for unsupervised learning. I used K-Means with elbow method to find optimal k value as 5 for clustering.

Below is the merged table with cluster labels for each Neighborhoods.



We can also examine that the frequency of average housing sales prices in different Neighborhoods using histogram plot.



# Results

The analysis shows that although there is a great number of venues in Delhi (~1400 venues for 145 Neighborhoods), which are concentrated in western & west-central Neighborhoods.

I considered only western part of Delhi (13 Neighborhoods discussed above) for clustering since they have are highly populated and avg house prices are also high, indicating comparatively high income of inhabitants. Also west part of Delhi has large number of work centers and offices (Gachibowli/Kondapur Neighborhoods).

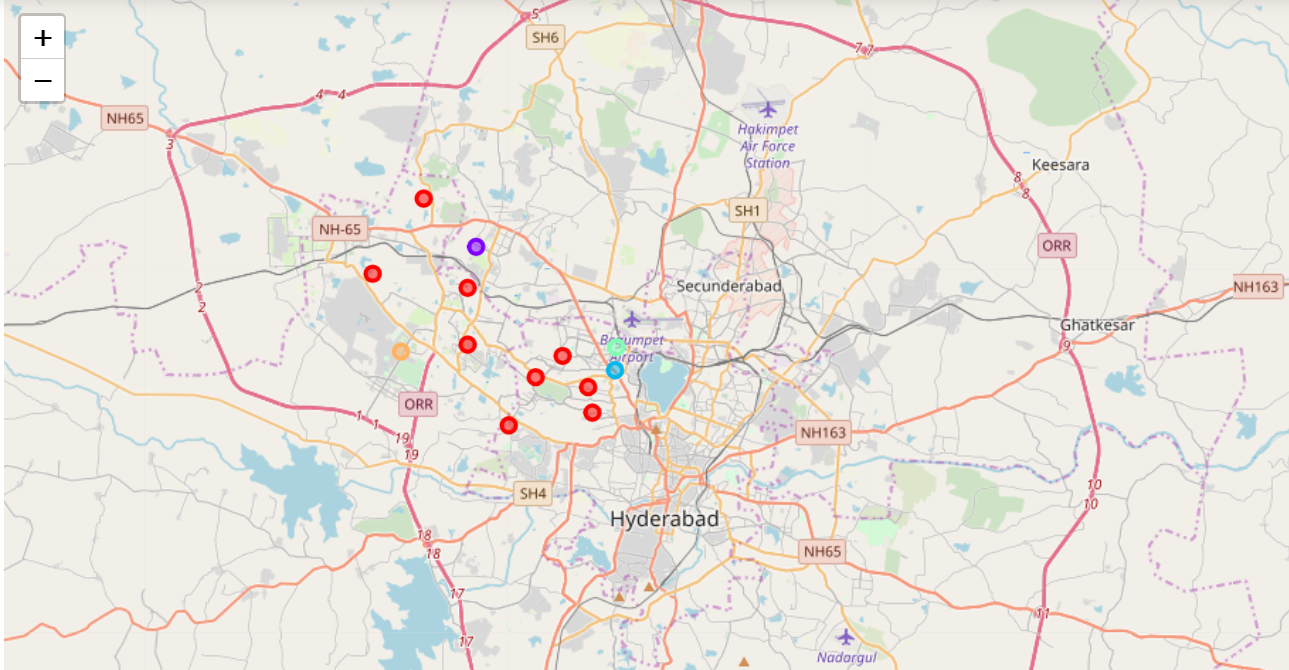
Some points:

* Somajiguda, Banjara Hills have number of supermarkets, convenience stores and department stores.
* Kondapur, Jubilee Hills & Venkateswara Colony (which is an extension of Banjara hills) also have high number of venues and high population.

As per the business problem discussed in Introduction section, we can recommend answers as follows:

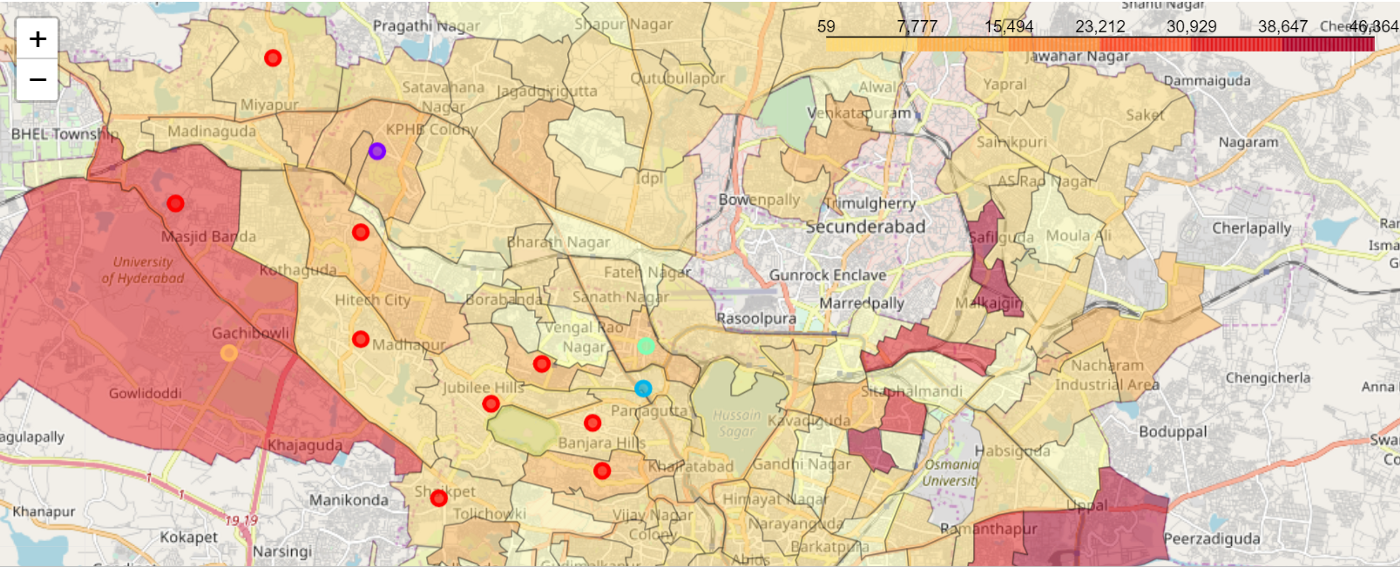
1. Western part of Delhi would be suitable for opening of supermarket/Hypermarket due to reasons discussed above.
2. In the western part, Neighborhoods such as Jubilee hills/Banjara Hills, Somajiguda & Kondapur could be recommended to open
3. supermarkets. You can also see a clustered map of west neighborhoods in Delhi below.

Map of Neighborhoods in west Delhi with Clusters shown below.



In summary section, one of my aim was also visualize the Average Prices for per square feet with choropleth style map. In final section, I created choropleth map which also has the below information for each borough:

1. Cluster name
2. Housing Sales Price (Avg\_Price) as Choropleth,



# Conclusion

Purpose of this project was to identify Delhi Neighborhoods close to center with low number of restaurants to aid stakeholders in narrowing down the search for optimal location for a setting supermarket. Clustering of the Neighborhoods in western Delhi was performed in order to create major zones of interest to be used as starting points for final exploration by stakeholders.

Recommended areas should be considered only as starting point for more detailed analysis which could eventually result in location which has other factors taken into account and all other relevant conditions met.

Final decision on supermarket restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to population/work centers & markets), proximity to major roads, real estate prices, social and economic dynamics of every neighborhood etc.

Thanks for reading!