Analyzing Venues and Real Estate Prices of Delhi

1. Introduction

1.1. Background

Delhi, officially the National Capital Territory of Delhi (NCT), is a city and a union territory of India containing New Delhi, the capital of India. The NCT covers an area of **1,484 square kilometers** (573 sq mi). According to the 2011 census, Delhi's city proper population was over **16 million**, the second-highest in India. The city is divided into **272 wards**. As a resident of this city, I decided to use this for my project.

As evident from the figures, Delhi is a city with a huge population and a high population density. Being such a crowded and huge city, it must be advantageous for shop keepers, local businesses and/or investors to accurately be able to determine the locality suitable for their type of business based on several factors such as:

1. level of the population: higher the better
2. price of real estate: lower prices will be preferred

At the same time, they may want to choose the district according to the density of the social places. However, it is difficult to obtain information that will guide investors in this direction, nowadays.

When we consider all these problems, we can create a map and information chart where the real estate index is placed on Delhi and each ward is clustered according to the venue density. Another idea is to check for correlations between the real estate prices and the population and the venue density of the wards.

1.2. Data Description

1. A geo-JSON file containing all the Delhi wards is used and it can be found here. The *Ward Number*, *Ward Name* and *Coordinates* features were selected.

2. Real estate prices were manually added from here. There were a lot of localities for which there was no real estate price data available. It was either due to that locality being an industrial zone or it not having and residential land or there being no residential land open to the public such as government-owned properties or army cantonments. I removed those localities and continued with only those localities for which we have the real estate price data.

3. The population data was scraped from a table found here.

4. Foursquare API was used to get the most common venues of given localities.

2. Methodology

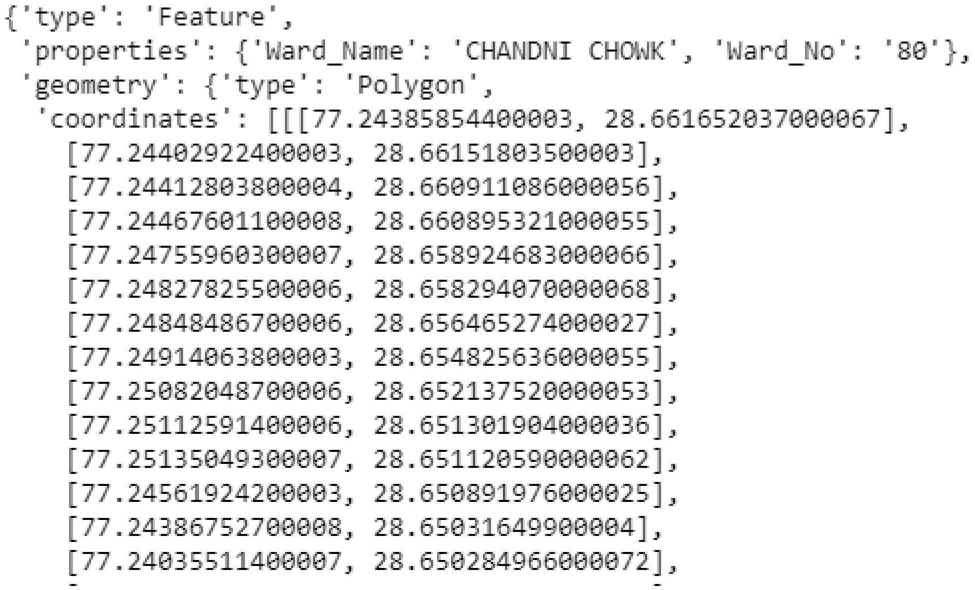
2.1. Collecting and Processing the Data

As discussed above, there will be three parts to the collection of data. It will be as follows:

- Part 1: Collecting the Delhi Wards Data
- Part 2: Adding Real Estate Prices
- Part 3: Adding Population Data
- Part 4: Collecting Venue data using FourSquare A

Part 1: Collecting the Delhi Wards Data

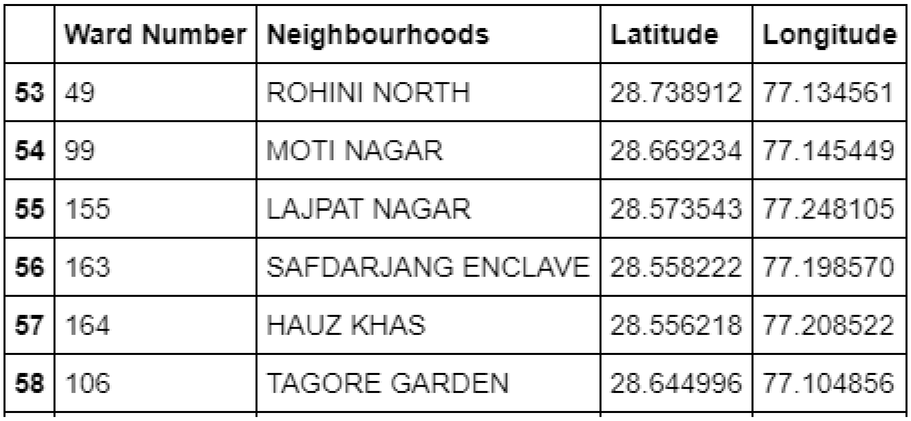
The geo-JSON file containing the ward data for Delhi had the following structure:



```
{'type': 'Feature',
 'properties': {'Ward_Name': 'CHANDNI CHOWK', 'Ward_No': '80'},
 'geometry': {'type': 'Polygon',
  'coordinates': [[[77.24385854400003, 28.661652037000067],
    [77.24402922400003, 28.66151803500003],
    [77.24412803800004, 28.660911086000056],
    [77.24467601100008, 28.660895321000055],
    [77.24755960300007, 28.658924683000066],
    [77.24827825500006, 28.658294070000068],
    [77.24848486700006, 28.656465274000027],
    [77.24914063800003, 28.654825636000055],
    [77.25082048700006, 28.652137520000053],
    [77.25112591400006, 28.651301904000036],
    [77.25135049300007, 28.651120590000062],
    [77.24561924200003, 28.650891976000025],
    [77.24386752700008, 28.65031649900004],
    [77.24035511400007, 28.650284966000072],
```

Delhi GEOJSON file structure

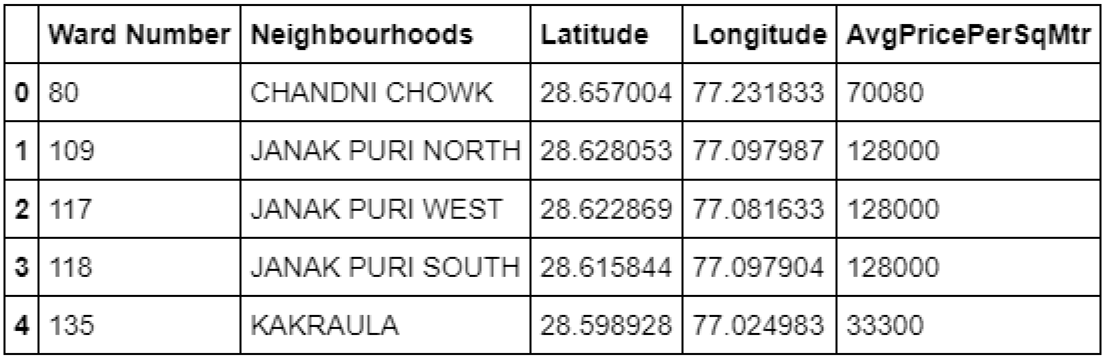
I had previously used the Bing API to geocode the Neighborhoods of Delhi but it was severely inaccurate. Rather than doing that, I took the average of the boundary coordinates hoping for the average to be somewhat lying in the center. I was hoping that somebody could comment on the accuracy of this method. Finally, we had the following data frame:



	Ward Number	Neighbourhoods	Latitude	Longitude
53	49	ROHINI NORTH	28.738912	77.134561
54	99	MOTI NAGAR	28.669234	77.145449
55	155	LAJPAT NAGAR	28.573543	77.248105
56	163	SAFDARJANG ENCLAVE	28.558222	77.198570
57	164	HAUZ KHAS	28.556218	77.208522
58	106	TAGORE GARDEN	28.644996	77.104856

Part 2: Adding Real Estate Prices for each locality

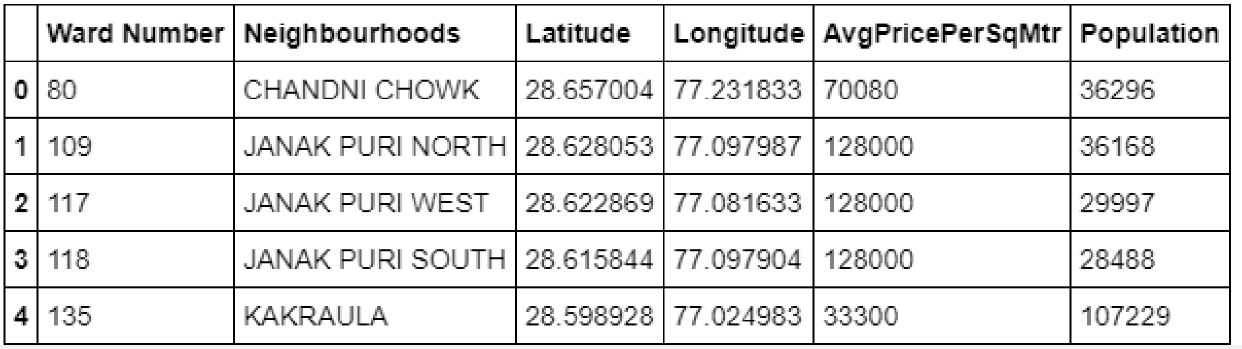
I used the bankbazaar website to get the average real estate price per square meter. I had to manually add it against each locality due to the differences in spellings of the names. The output was as follows:



	Ward Number	Neighbourhoods	Latitude	Longitude	AvgPricePerSqMtr
0	80	CHANDNI CHOWK	28.657004	77.231833	70080
1	109	JANAK PURI NORTH	28.628053	77.097987	128000
2	117	JANAK PURI WEST	28.622869	77.081633	128000
3	118	JANAK PURI SOUTH	28.615844	77.097904	128000
4	135	KAKRAULA	28.598928	77.024983	33300

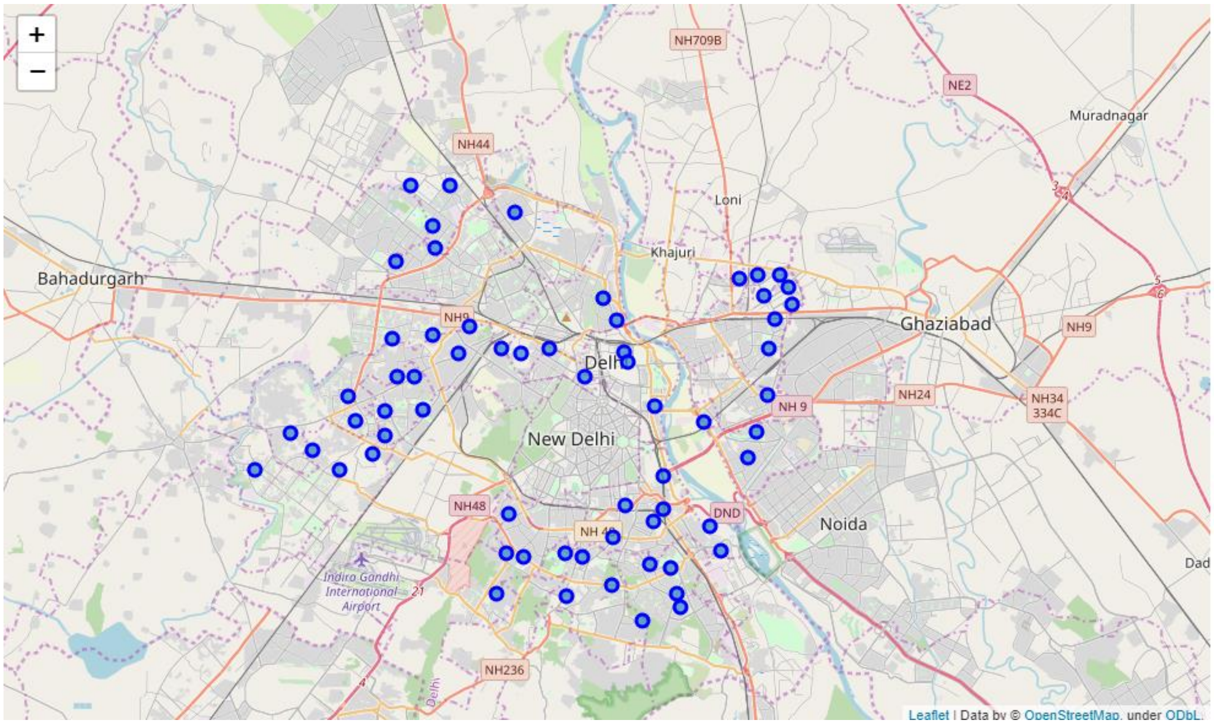
Part 3: Adding the Population Data

To add the population data I scraped the indikosh website which contained a table with the population of each ward against the ward number. For this, I used the BeautifulSoup package. I finally merged the two tables on the ward number to get the final output of the data on which further analysis was performed. It contained 63 entries.



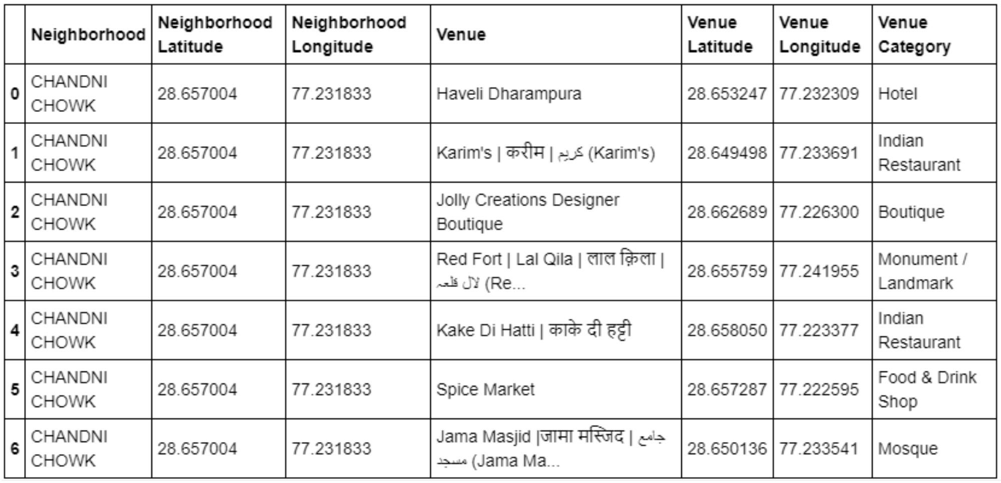
	Ward Number	Neighbourhoods	Latitude	Longitude	AvgPricePerSqMtr	Population
0	80	CHANDNI CHOWK	28.657004	77.231833	70080	36296
1	109	JANAK PURI NORTH	28.628053	77.097987	128000	36168
2	117	JANAK PURI WEST	28.622869	77.081633	128000	29997
3	118	JANAK PURI SOUTH	28.615844	77.097904	128000	28488
4	135	KAKRAULA	28.598928	77.024983	33300	107229

To get an idea of what neighborhoods and locations we are going to be performing our analysis on I visualized Delhi and the neighborhoods. For this, I plotted a map of Delhi and superimposed the neighborhoods on top of it. I used Nominatim to get the latitude and longitude of Delhi.



Part 4: Collecting Venues for each locality using FourSqaure API

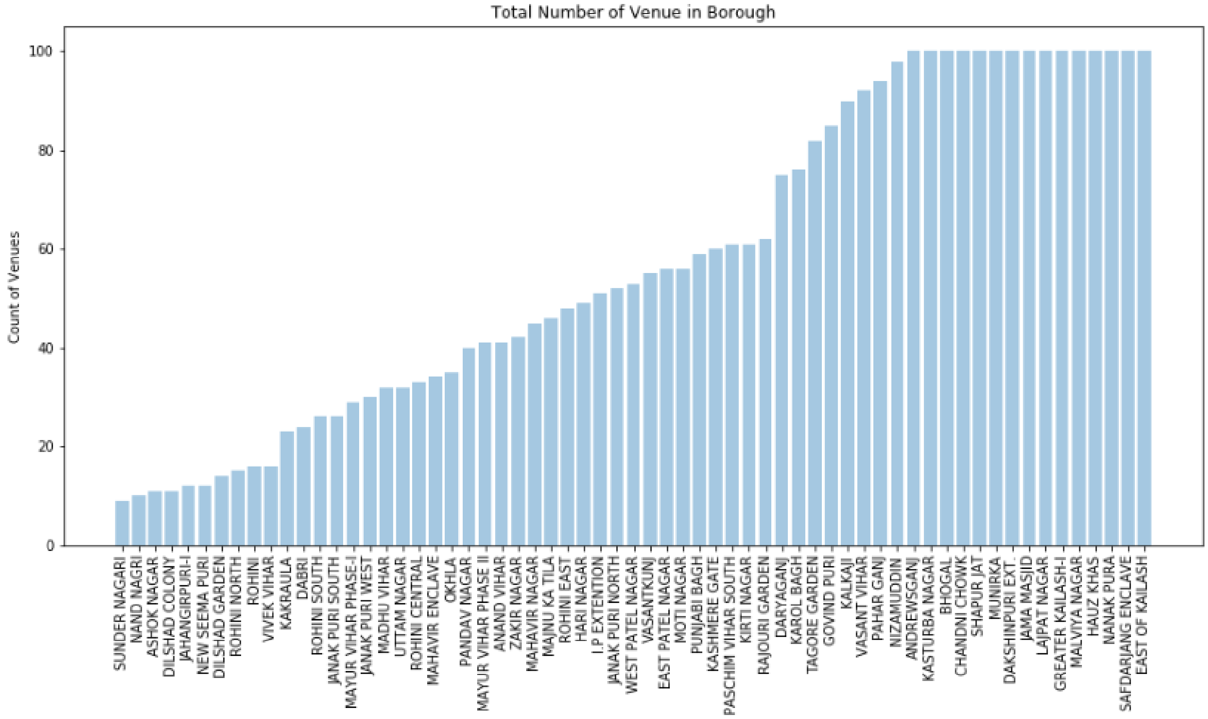
Now, that I had all the localities on which I want to do the analysis, all that was left was the venue data. To get this, I leveraged the FourSquare API to get the most popular venues in a 2km radius since the neighborhoods are quite spread out. The API used was as following:
https://api.foursquare.com/v2/venues/explore?&client_id=**CLIENT_ID**&client_secret=**CLIENT_SECRET**&v=**VERSION**&ll=**LATITUDE**,**LONGITUDE**&radius=**RADIUS**&limit=**LIMIT**
The resulting data frame had 3,620 entries.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	CHANDNI CHOWK	28.657004	77.231833	Haveli Dharampura	28.653247	77.232309	Hotel
1	CHANDNI CHOWK	28.657004	77.231833	Karim's | करीम | کریم (Karim's)	28.649498	77.233691	Indian Restaurant
2	CHANDNI CHOWK	28.657004	77.231833	Jolly Creations Designer Boutique	28.662689	77.226300	Boutique
3	CHANDNI CHOWK	28.657004	77.231833	Red Fort | Lal Qila | लाल क़िला | لال قلعہ (Re...	28.655759	77.241955	Monument / Landmark
4	CHANDNI CHOWK	28.657004	77.231833	Kake Di Hatti | काके दी हट्टी	28.658050	77.223377	Indian Restaurant
5	CHANDNI CHOWK	28.657004	77.231833	Spice Market	28.657287	77.222595	Food & Drink Shop
6	CHANDNI CHOWK	28.657004	77.231833	Jama Masjid |जामा मस्जिद | جامع مسجد (Jama Ma...	28.650136	77.233541	Mosque

2.2. Exploratory Data Analysis

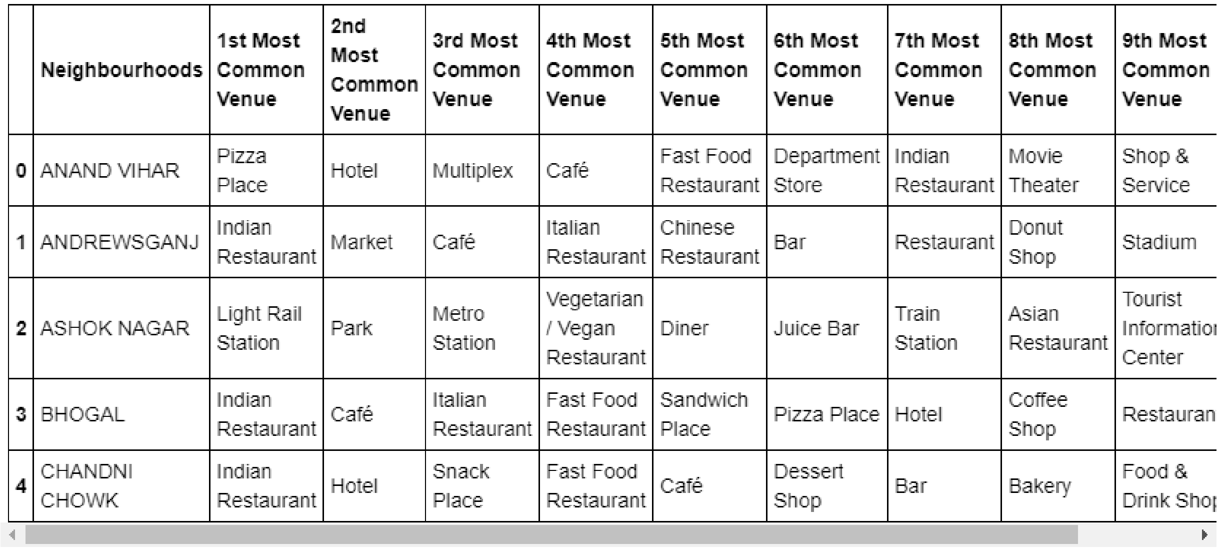
2.2.1 Analyzing the Venues returned for each Ward



We can see that from Sunder Nagari to Vivek Vihar, which includes 10 different wards, the number of venues returned by the foursquare API is less than 20. On the other hand, there are almost 15 locations for which the number of venues returned reached the limitation of a hundred results. We can increase or decrease the number of venues returned by playing around with the radius size but we are satisfied with the result for now. There were **170** unique categories out of the **3,620** venues.

2.2.2 Analyzing each Ward

I analyzed each ward against the venue information. To do this, I first did a one-hot encoding of the venues against their categories. Next, I grouped them all on the *Neighbourhoods* took the mean of the frequency of occurrence. I ordered the result in descending order and took the top 10 venue categories for each neighborhood.

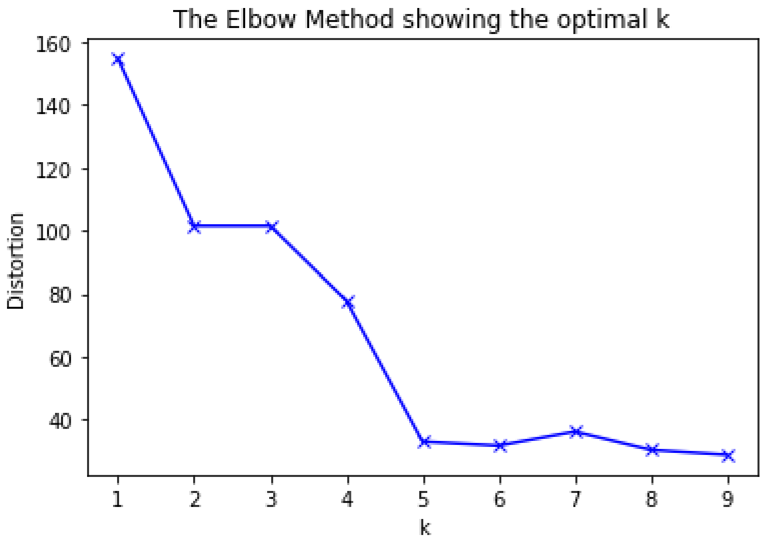


	Neighbourhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	ANAND VIHAR	Pizza Place	Hotel	Multiplex	Café	Fast Food Restaurant	Department Store	Indian Restaurant	Movie Theater	Shop & Service
1	ANDREWSGANJ	Indian Restaurant	Market	Café	Italian Restaurant	Chinese Restaurant	Bar	Restaurant	Donut Shop	Stadium
2	ASHOK NAGAR	Light Rail Station	Park	Metro Station	Vegetarian / Vegan Restaurant	Diner	Juice Bar	Train Station	Asian Restaurant	Tourist Informatio Center
3	BHOGAL	Indian Restaurant	Café	Italian Restaurant	Fast Food Restaurant	Sandwich Place	Pizza Place	Hotel	Coffee Shop	Restauran
4	CHANDNI CHOWK	Indian Restaurant	Hotel	Snack Place	Fast Food Restaurant	Café	Dessert Shop	Bar	Bakery	Food & Drink Sho

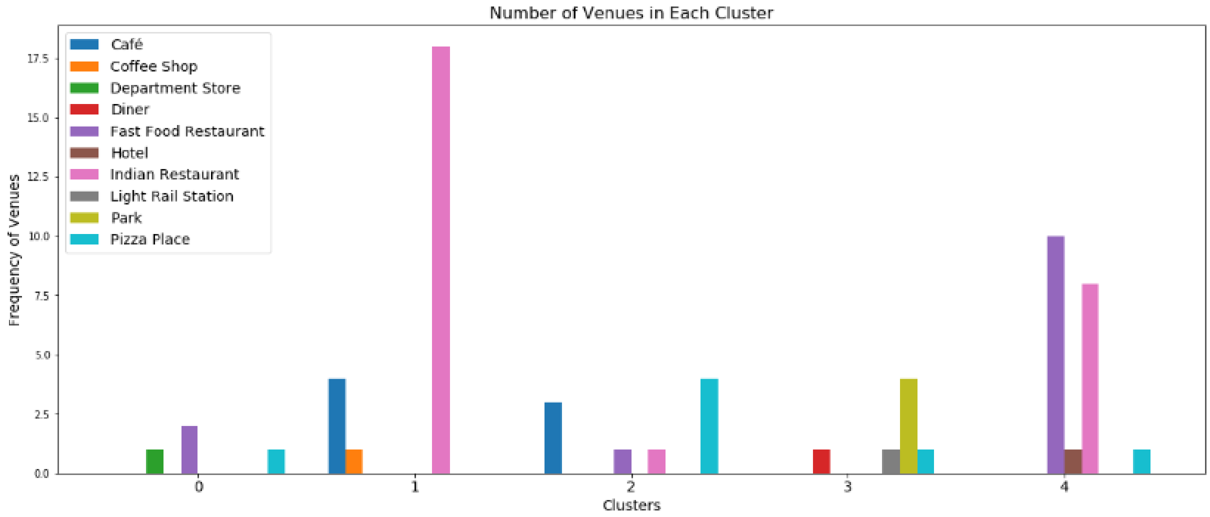
2.2.3 Clustering the Neighborhoods

There are some common venue categories in each neighborhood. For this reason, the next step was to do an unsupervised clustering of the neighborhoods. I used the **K-Means algorithm** to do this. K-Means algorithm is one of the most common clustering methods of unsupervised learning.

To ensure that the K-Means is used most efficiently, we use the elbow method to correctly identify the K value for the clustering. From the graph below, I select the value for **K equal to 5** as beyond the rate at which the distortion drops slows down the most significantly.



To help me find proper label names for each cluster, I decided to plot a bar chart containing the number of 1st Most Common Venue in each cluster.

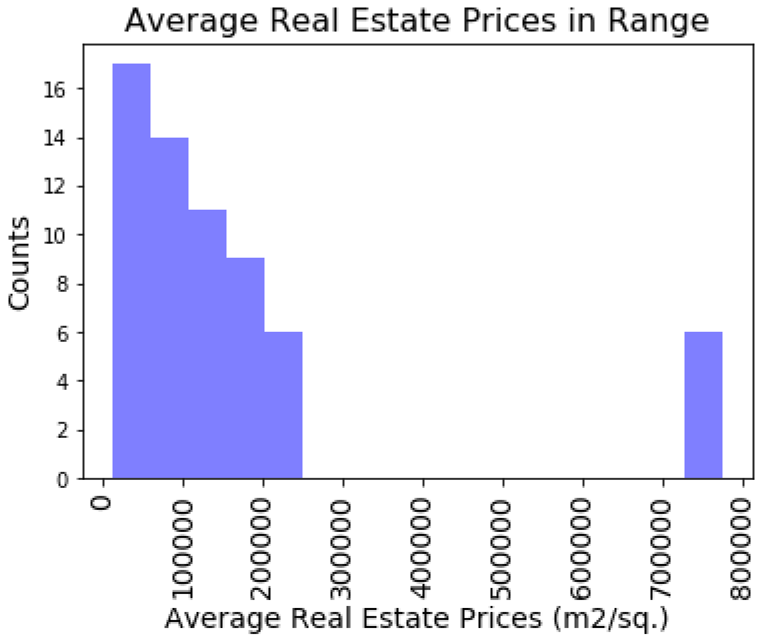


After examining the above graph I decided to label each cluster as follows:

- Cluster 0: ‘Departmental Stores
- Cluster 1: ‘Cafe and Indian Restaurants’
- Cluster 2: ‘Restaurants’
- Cluster 3: ‘Fast-food’
- Cluster 4: ‘Parks and Recreation’

2.2.4 Analyzing real estate prices

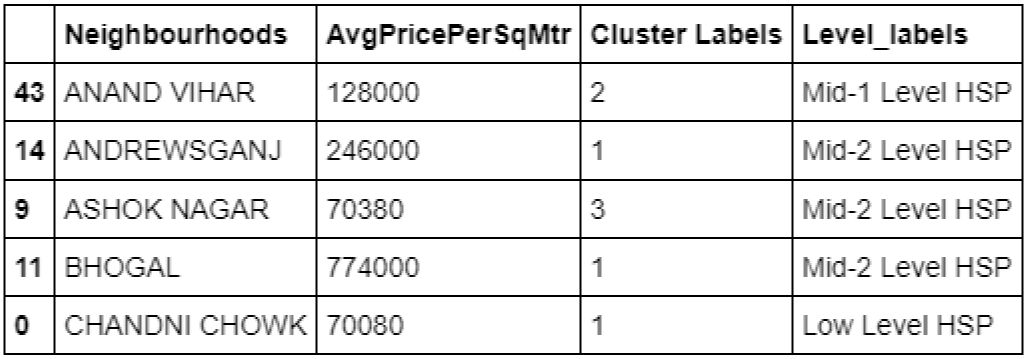
I created a histogram to visualize the frequency of real estate sales prices in different ranges.



From the above histogram, I decided to define the range as follows:

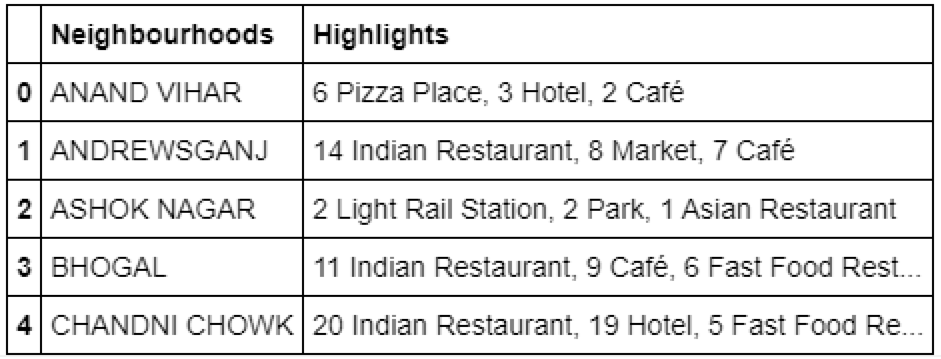
- <40,000 : “Low Level”
- 40,000–80,000 : “Mid-1 Level”
- 120,000–160,000 : “Mid-2 Level”
- 240,000–280,000 : “High-1 Level”
- >320,000 : “High-2 Level”

After adding the cluster labels and real estate price labels the resulting data frame looked as follows:



	Neighbourhoods	AvgPricePerSqMtr	Cluster Labels	Level_labels
43	ANAND VIHAR	128000	2	Mid-1 Level HSP
14	ANDREWSGANJ	246000	1	Mid-2 Level HSP
9	ASHOK NAGAR	70380	3	Mid-2 Level HSP
11	BHOGAL	774000	1	Mid-2 Level HSP
0	CHANDNI CHOWK	70080	1	Low Level HSP

One of the aims was also to show the number of top 3 venue information for each borough on the map. Thus, I grouped each borough by the number of top 3 venues and I combined those pieces of information in the **Highlights** column.

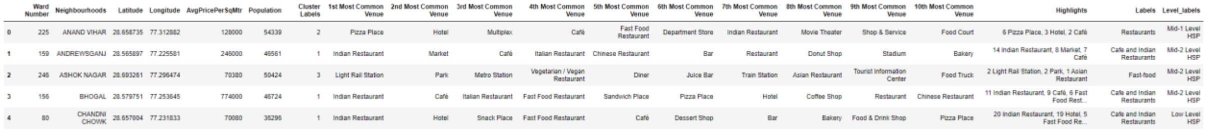


	Neighbourhoods	Highlights
0	ANAND VIHAR	6 Pizza Place, 3 Hotel, 2 Café
1	ANDREWSGANJ	14 Indian Restaurant, 8 Market, 7 Café
2	ASHOK NAGAR	2 Light Rail Station, 2 Park, 1 Asian Restaurant
3	BHOGAL	11 Indian Restaurant, 9 Café, 6 Fast Food Rest...
4	CHANDNI CHOWK	20 Indian Restaurant, 19 Hotel, 5 Fast Food Re...

3. Results

3.1. Main Table with Results

I merged the new variables with related cluster information in the main data frame.

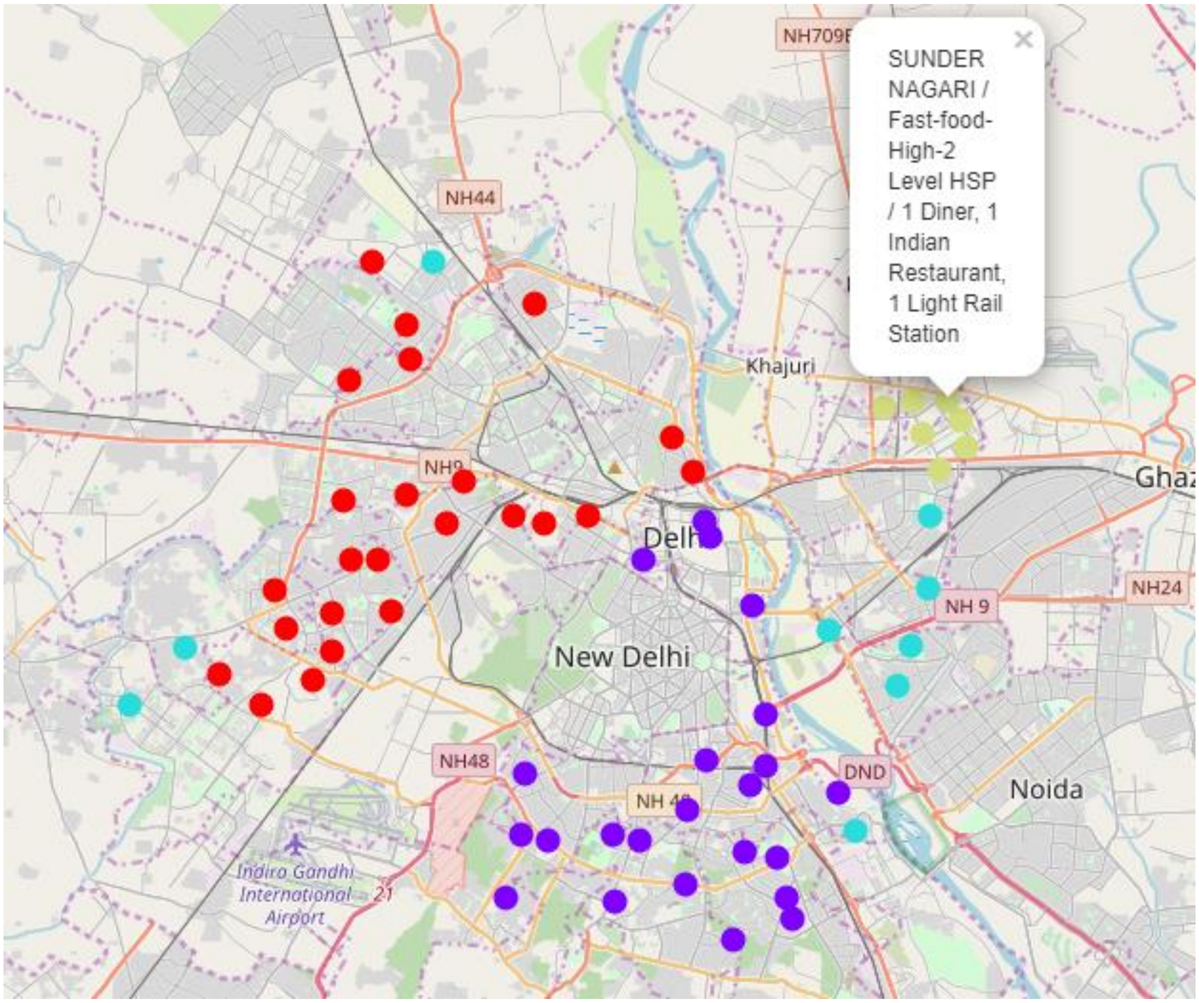


	Ward Number	Neighbourhoods	Latitude	Longitude	AvgPricePerSqMtr	Population	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Highlights	Labels	Level_labels
0	225	ANAND VIHAR	28.658735	77.312882	128000	54339	2	Pizza Place	Hotel	Multiplex	Café	Fast Food Restaurant	Department Store	Indian Restaurant	Movie Theater	Shop & Service	Food Court	6 Pizza Place, 3 Hotel, 2 Café	Restaurants	Mid-1 Level HSP
1	159	ANDREWSGANJ	28.565897	77.225581	246000	46561	1	Indian Restaurant	Market	Café	Italian Restaurant	Chinese Restaurant	Bar	Restaurant	Donut Shop	Stadium	Bakery	14 Indian Restaurant, 8 Market, 7 Café	Cafe and Indian Restaurants	Mid-2 Level HSP
2	246	ASHOK NAGAR	28.693261	77.296474	70380	50424	3	Light Rail Station	Park	Metro Station	Vegetarian / Vegan Restaurant	Diner	Juice Bar	Train Station	Asian Restaurant	Tourist Information Center	Food Truck	2 Light Rail Station, 2 Park, 1 Asian Restaurant	Fast-food	Mid-2 Level HSP
3	156	BHOGAL	28.579751	77.253645	774000	46724	1	Indian Restaurant	Café	Italian Restaurant	Fast Food Restaurant	Sandwich Place	Pizza Place	Hotel	Coffee Shop	Restaurant	Chinese Restaurant	11 Indian Restaurant, 9 Café, 6 Fast Food Rest...	Cafe and Indian Restaurants	Mid-2 Level HSP
4	80	CHANDNI CHOWK	28.657004	77.231833	70080	36296	1	Indian Restaurant	Hotel	Snack Place	Fast Food Restaurant	Café	Dessert Shop	Bar	Bakery	Food & Drink Shop	Pizza Place	20 Indian Restaurant, 19 Hotel, 5 Fast Food Re...	Cafe and Indian Restaurants	Low Level HSP

The final data frame

3.2. Map representing Clusters

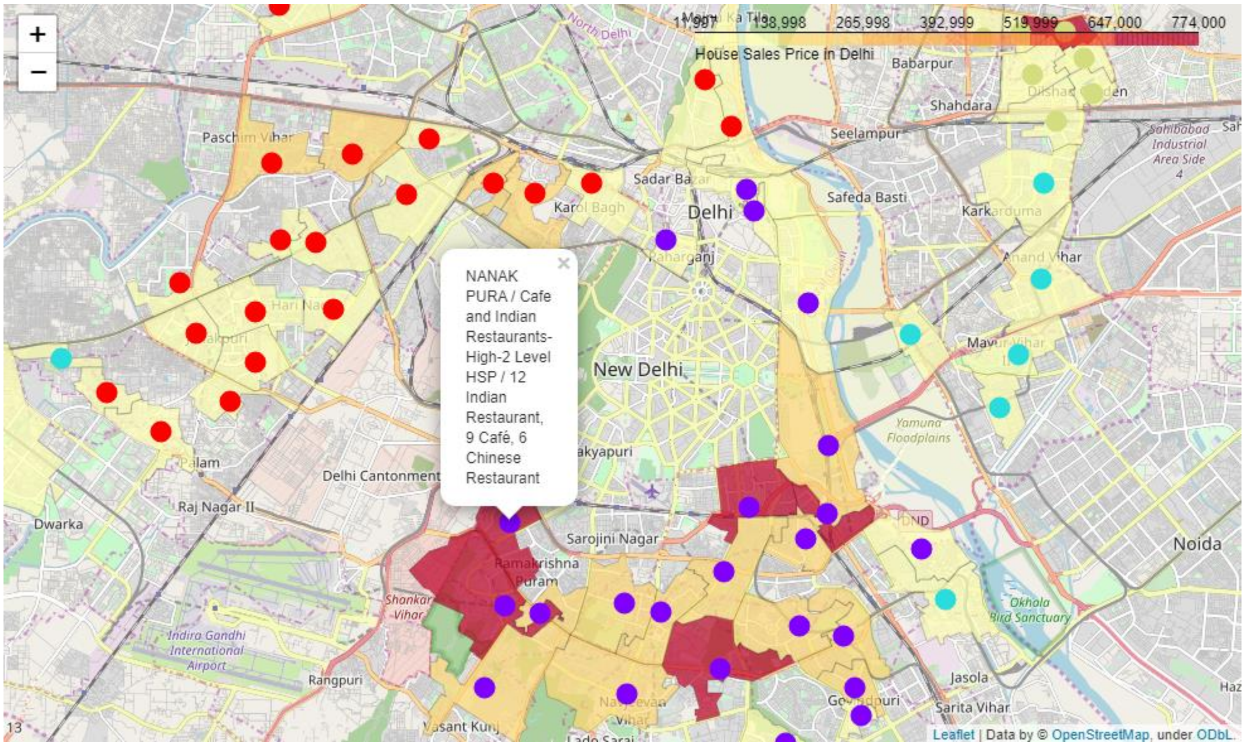
I visualized the resulting clusters and the way they are spread out. It is interesting to notice that the clusters are quite spatially separated.


SUNDER NAGARI / Fast-food-High-2 Level HSP / 1 Diner, 1 Indian Restaurant, 1 Light Rail Station
NH709E
NH44
Khajuri
NH9
Delhi
Ghaz
New Delhi
NH 9
NH24
NH48
NH 48
DND
Noida
Indira Gandhi International Airport
21

3.3. Map of Real Estate Prices with Cluster information

One of the aims was also to visualize the Average Real Estate Prices per square meter for the localities under speculation using a choropleth map. For this, I used the Delhi Wards JSON file which was also the file used to obtain the wards of Delhi in the very beginning. The link for the same can be found under the data section in the introduction.

I had to clean up the JSON file to only include the neighborhoods we were analyzing.

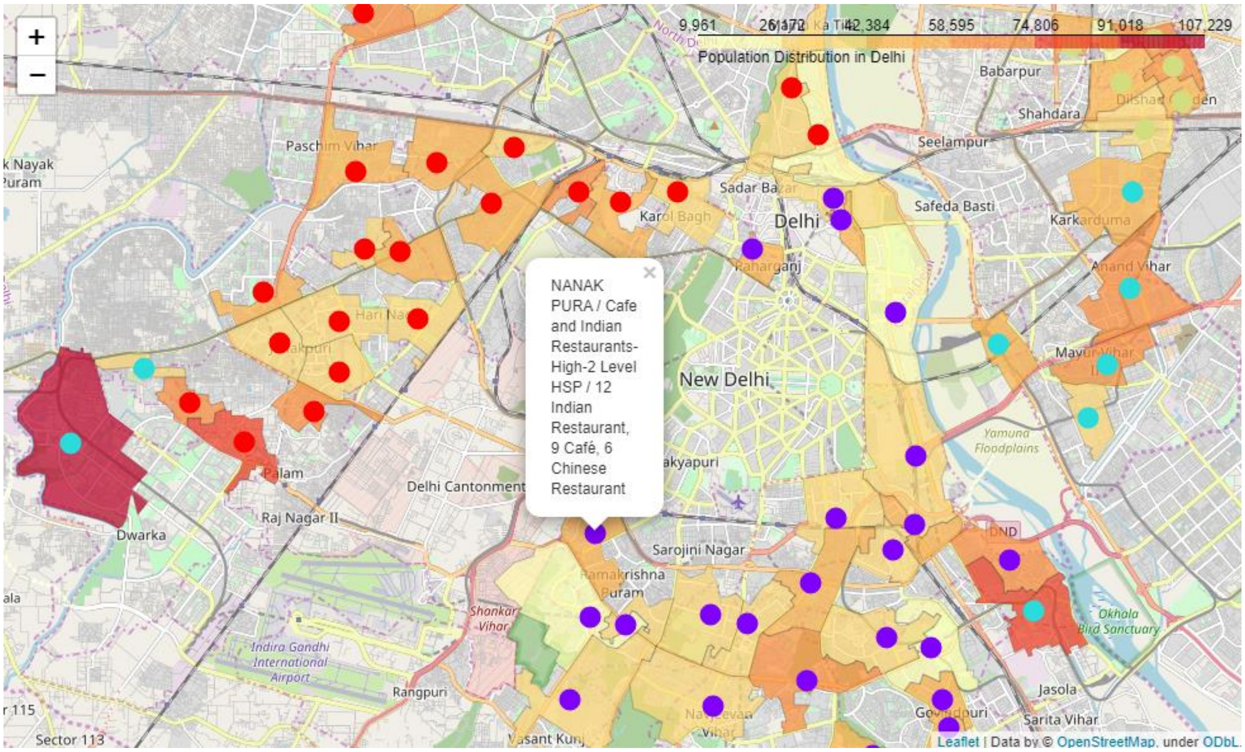


3.4. Map of Population with Cluster information

Another aim was to visualize the Population for the localities under speculation using a choropleth map.

In the final section, I created a choropleth map which also has the below pieces of information for each borough:

- Neighborhood name,
- Cluster name,
- Real Estate Price Levels,
- Top 3 number of venue



4. Discussion

As I mentioned before, Delhi is a big city with a high population density. The total number of measurements and population densities of the different wards in total can vary. As there is such a complexity, very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high-quality results for this metropolis.

I used the Kmeans algorithm as part of this clustering study. When I tested the Elbow method, I set the optimum k value to 5. However, only 63 ward coordinates were used. For more detailed and accurate guidance, the data set can be expanded and the details of the localities or street can also be drilled.

I also performed data analysis through this information by adding the coordinates of the wards, real estate price averages and population as static data on GitHub. In future studies, these data can also be accessed dynamically from specific platforms or packages.

I ended the study by visualizing the data and clustering information on the Delhi map. In future studies, web or telephone applications can be carried out to direct investors.

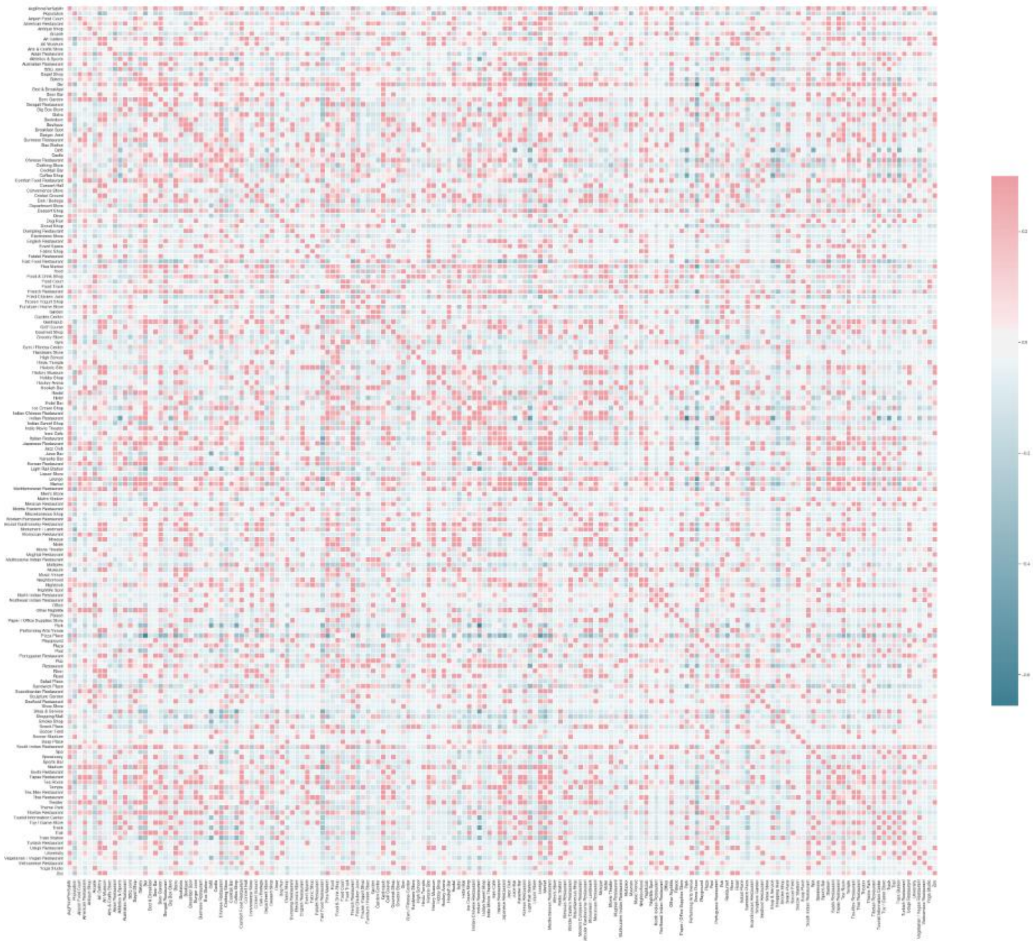
5. Conclusion

As a result, people are turning to big cities to start a business or work. For this reason, people can achieve better outcomes through their access to platforms where such information is provided.

Not only for investors but also city managers can manage the city more regularly by using similar data analysis types or platforms.

6. Future Interest and Exploration

After looking at the above choropleth map depicting the population levels and the clusters, I notice a trend for certain types of clusters and the population levels. It could give an insight into what kind of businesses flourish in areas with different kinds of population density. For this, I think we should see if there is any sort of correlations between the different venues and the population. I chose to visualize a correlation heat map to get me started with what variables to play around with.



7. References

- [1] Delhi Wikipedia
- [2] Github Repository containing Wards Data
- [3] Foursquare API
- [4] Population Data
- [5] Real Estate Prices