# Coursera Capstone - Applied Data Science Final Report

## Opening a new Shopping Mall in Delhi NCR

19.03.2019

Yashwanth P

## Overview

Delhi being a very vast location and with a very huge population, it makes it difficult to place the location of a new Shopping Mall, given that it might already have a few.

The project aims to leverage Clustering and Machine Learning to find different locations where a new mall or malls would actually be a great fit in North Capital Region (Delhi and its surrounding cities included).

Target Audience - Any property investor who would want to build any type of Public Entertainment. The analysis would also give us an understanding of the locations for other venues such as restaurants, pubs and nightclubs etc.

## Background

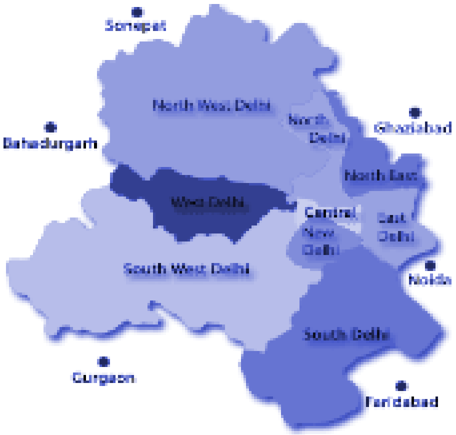
- ❖ Shopping Mall - A Shopping Mall is a modern term for a multi level Shopping Precinct or Shopping Center in which one or more buildings form a complex of shops representing merchandisers.Typical Components include
  - ➢ Food Courts
  - ➢ Department Stores
  - ➢ Movie Theatres

- ❖ Delhi, officially the National Capital Territory of Delhi is the capital of India. The National Capital Region (NCR) is Delhi’s urban area which also include the satellite cities of Faridabad, Gurgaon, Ghaziabad & Noida.

  NCR has an area of around 1,484 square KM (573 sq mi). The population of NCR is estimated to be over 26 Million people, making it the 2nd largest Urban area according to the United Nations.

  It’s also the 2nd most productive metro area of India - home to 18 Billionaires and 23k Millionaires

***DELHI NCR Map:***



## Goals

1. Delhi being a very vast location and with a very huge population, it makes it difficult to place the location of a new Shopping Mall, given that it might already have quite a few.

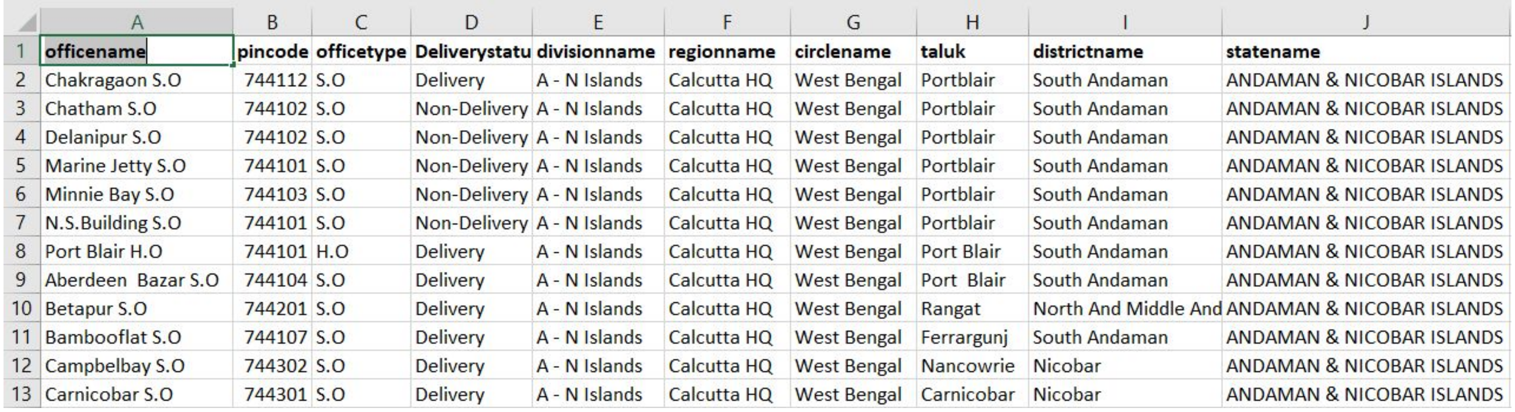
2. The project aims to leverage Clustering and Machine Learning to find different locations where a new mall or malls would actually be a great fit in North Capital Region (Delhi and its surrounding cities included).

## Target Audience

Any property investor who would want to build any type of Public Entertainment. The analysis would also give us an understanding of the locations for other venues such as restaurants, pubs and nightclubs etc.

## Data Sources

1. Postal Code - The Indian Postal Data is present at (https://www.indiapost.gov.in/vas/pages/findpincode.aspx) as a csv file which would enable us to identify all the postal codes in NCR. Postal Codes in India are a 6 - Digit Code (Ex. 122002). Below is the sample of the same data:

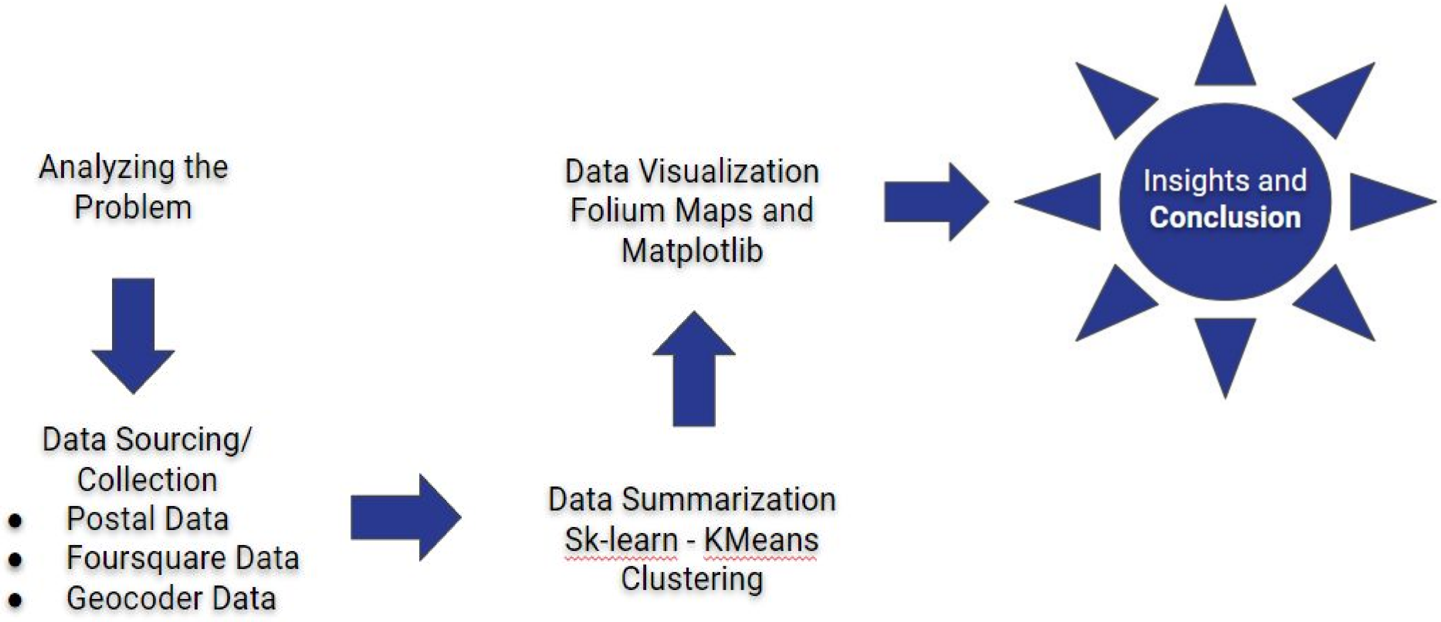


| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **officename** | **pincode** | **officetype** | **Deliverystatu** | **divisionname** | **regionname** | **circlename** | **taluk** | **districtname** | **statename** |
| 2 | Chakragaon S.O | 744112 | S.O | Delivery | A - N Islands | Calcutta HQ | West Bengal | Portblair | South Andaman | ANDAMAN & NICOBAR ISLANDS |
| 3 | Chatham S.O | 744102 | S.O | Non-Delivery | A - N Islands | Calcutta HQ | West Bengal | Portblair | South Andaman | ANDAMAN & NICOBAR ISLANDS |
| 4 | Delanipur S.O | 744102 | S.O | Non-Delivery | A - N Islands | Calcutta HQ | West Bengal | Portblair | South Andaman | ANDAMAN & NICOBAR ISLANDS |
| 5 | Marine Jetty S.O | 744101 | S.O | Non-Delivery | A - N Islands | Calcutta HQ | West Bengal | Portblair | South Andaman | ANDAMAN & NICOBAR ISLANDS |
| 6 | Minnie Bay S.O | 744103 | S.O | Non-Delivery | A - N Islands | Calcutta HQ | West Bengal | Portblair | South Andaman | ANDAMAN & NICOBAR ISLANDS |
| 7 | N.S.Building S.O | 744101 | S.O | Non-Delivery | A - N Islands | Calcutta HQ | West Bengal | Portblair | South Andaman | ANDAMAN & NICOBAR ISLANDS |
| 8 | Port Blair H.O | 744101 | H.O | Delivery | A - N Islands | Calcutta HQ | West Bengal | Port Blair | South Andaman | ANDAMAN & NICOBAR ISLANDS |
| 9 | Aberdeen Bazar S.O | 744104 | S.O | Delivery | A - N Islands | Calcutta HQ | West Bengal | Port Blair | South Andaman | ANDAMAN & NICOBAR ISLANDS |
| 10 | Betapur S.O | 744201 | S.O | Delivery | A - N Islands | Calcutta HQ | West Bengal | Rangat | North And Middle And | ANDAMAN & NICOBAR ISLANDS |
| 11 | Bambooflat S.O | 744107 | S.O | Delivery | A - N Islands | Calcutta HQ | West Bengal | Ferrargunj | South Andaman | ANDAMAN & NICOBAR ISLANDS |
| 12 | Campbelbay S.O | 744302 | S.O | Delivery | A - N Islands | Calcutta HQ | West Bengal | Nancowrie | Nicobar | ANDAMAN & NICOBAR ISLANDS |
| 13 | Carnicobar S.O | 744301 | S.O | Delivery | A - N Islands | Calcutta HQ | West Bengal | Carnicobar | Nicobar | ANDAMAN & NICOBAR ISLANDS |

As you can see the data is not Processed as required. It has a lot of duplicates and a lot of unnecessary columns. Had to create a final name column from the combination of 'divisionname' and 'districtname' columns present in the data

2. Foursquare API - The foursquare data would enable us to understand the postal code area venues which would become the base for our clustering. The data would be something like bars or pharmacies near a particular Lat/Long location
3. Geocoder data - To get the Latitude and Longitude of the Postal Codes taken for consideration
4. Folium, Pandas, Matplotlib, Json, Sklearn libraries which help in data wrangling, data visualization, data manipulation for the given clustering analysis in Python

## Methodology

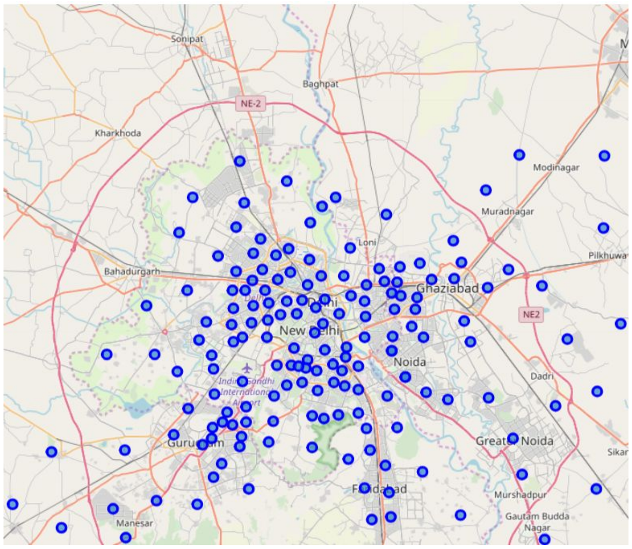

Analyzing the Problem
Data Sourcing/ Collection
● Postal Data
● Foursquare Data
● Geocoder Data
Data Summarization Sk-learn - KMeans Clustering
Data Visualization Folium Maps and Matplotlib
Insights and Conclusion

The process flow shows the methodology followed for arriving at the conclusions for the analysis

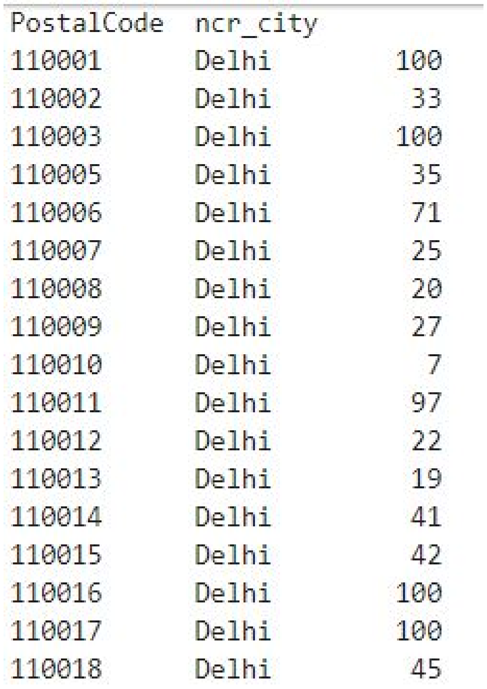
1. From the data sources we get all the Postal data for complete India as a csv file. We use the read_csv function to read csv file to a Pandas Dataframe.
2. We then proceed to filter out the cities Delhi, Gurgaon, Faridabad, Ghaziabad and Gautam Buddha Nagar (Noida) from the complete list.
3. We do create a new derived column for the city name instead of multiple columns and remove all duplicates if any exist. We did find specific nuances where the same postal code was mapped to different cities and updated the data so that it would actually point to the relevant city

4. A total of 212 Postal codes were a part of the analysis region. Used the geocoder to get the Lat-Long data for the selected data and then used Folium to plot the mapping of the data as shown below.

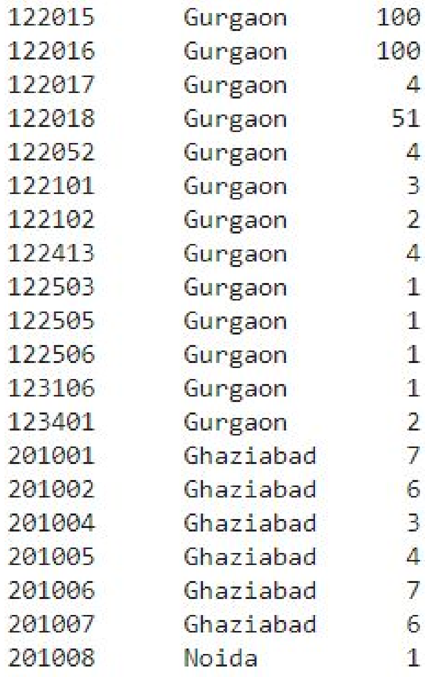
Delhi NCR Postal Code Mapping



5. Now that we have all the geo data we proceed with getting the Venue data from the FourSquare API
6. I did try with multiple radii as we had very less fill rate for most of the locations with some locations having no venues nearby. Finally I did proceed with the 2 Km as my radius and got 4,286 venues for all the 212 locations.

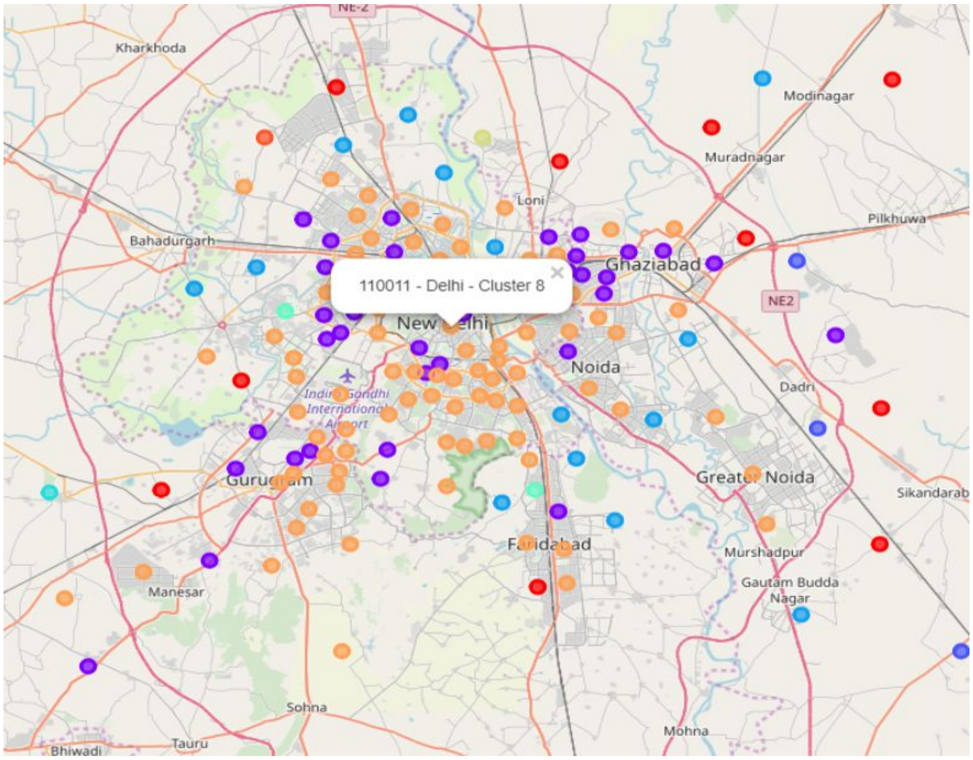


| PostalCode | ncr_city | |
|---|---|---|
| 110001 | Delhi | 100 |
| 110002 | Delhi | 33 |
| 110003 | Delhi | 100 |
| 110005 | Delhi | 35 |
| 110006 | Delhi | 71 |
| 110007 | Delhi | 25 |
| 110008 | Delhi | 20 |
| 110009 | Delhi | 27 |
| 110010 | Delhi | 7 |
| 110011 | Delhi | 97 |
| 110012 | Delhi | 22 |
| 110013 | Delhi | 19 |
| 110014 | Delhi | 41 |
| 110015 | Delhi | 42 |
| 110016 | Delhi | 100 |
| 110017 | Delhi | 100 |
| 110018 | Delhi | 45 |



| | | |
|---|---|---|
| 122015 | Gurgaon | 100 |
| 122016 | Gurgaon | 100 |
| 122017 | Gurgaon | 4 |
| 122018 | Gurgaon | 51 |
| 122052 | Gurgaon | 4 |
| 122101 | Gurgaon | 3 |
| 122102 | Gurgaon | 2 |
| 122413 | Gurgaon | 4 |
| 122503 | Gurgaon | 1 |
| 122505 | Gurgaon | 1 |
| 122506 | Gurgaon | 1 |
| 123106 | Gurgaon | 1 |
| 123401 | Gurgaon | 2 |
| 201001 | Ghaziabad | 7 |
| 201002 | Ghaziabad | 6 |
| 201004 | Ghaziabad | 3 |
| 201005 | Ghaziabad | 4 |
| 201006 | Ghaziabad | 7 |
| 201007 | Ghaziabad | 6 |
| 201008 | Noida | 1 |

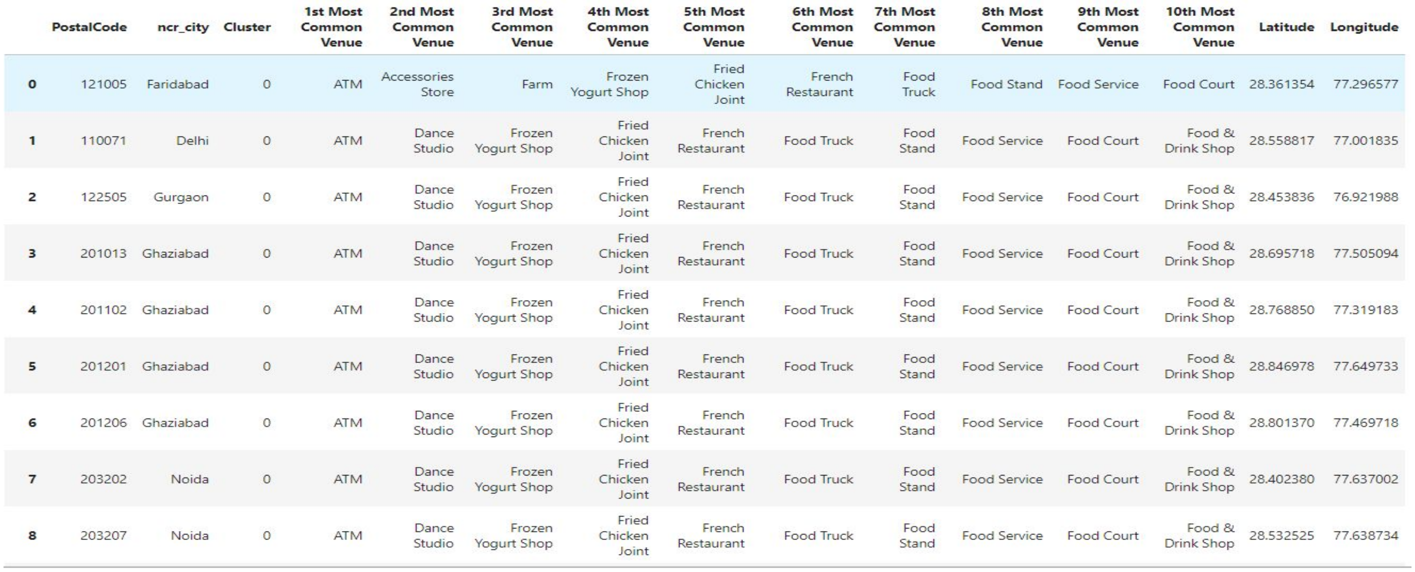
7. Analyzed all neighbourhoods by grouping rows on the Postal code and mean of the frequencies of occurrence of each venue category. Sorted them on the basis of the top 10 occuring common venues
8. Clustered them using the K-Means clustering Algorithm with the number of clusters as 10, as the number of locations were high in number.



**Cluster Mapping on All Venues**

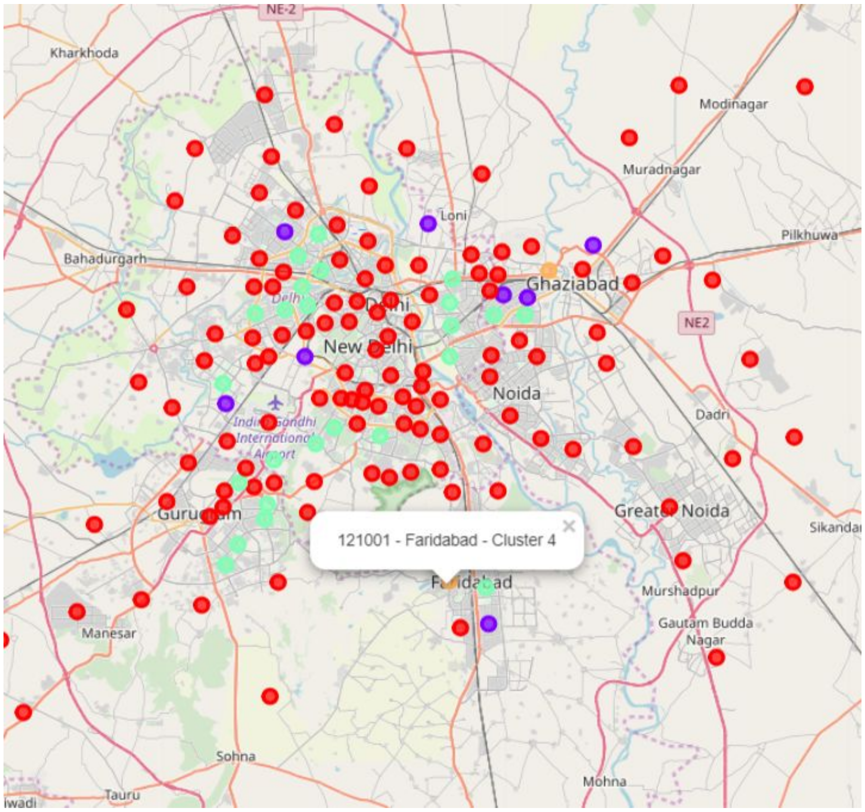
**10 Clusters formed**

9. The above created clusters can be used for solving any other such as finding the competition in any location or good location for any restaurant etc. The data can be viewed as below:



| | PostalCode | ncr_city | Cluster | 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 | Latitude | Longitude |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 121005 | Faridabad | 0 | ATM | Accessories Store | Farm | Frozen Yogurt Shop | Fried Chicken Joint | French Restaurant | Food Truck | Food Stand | Food Service | Food Court | 28.361354 | 77.296577 |
| **1** | 110071 | Delhi | 0 | ATM | Dance Studio | Frozen Yogurt Shop | Fried Chicken Joint | French Restaurant | Food Truck | Food Stand | Food Service | Food Court | Food & Drink Shop | 28.558817 | 77.001835 |
| **2** | 122505 | Gurgaon | 0 | ATM | Dance Studio | Frozen Yogurt Shop | Fried Chicken Joint | French Restaurant | Food Truck | Food Stand | Food Service | Food Court | Food & Drink Shop | 28.453836 | 76.921988 |
| **3** | 201013 | Ghaziabad | 0 | ATM | Dance Studio | Frozen Yogurt Shop | Fried Chicken Joint | French Restaurant | Food Truck | Food Stand | Food Service | Food Court | Food & Drink Shop | 28.695718 | 77.505094 |
| **4** | 201102 | Ghaziabad | 0 | ATM | Dance Studio | Frozen Yogurt Shop | Fried Chicken Joint | French Restaurant | Food Truck | Food Stand | Food Service | Food Court | Food & Drink Shop | 28.768850 | 77.319183 |
| **5** | 201201 | Ghaziabad | 0 | ATM | Dance Studio | Frozen Yogurt Shop | Fried Chicken Joint | French Restaurant | Food Truck | Food Stand | Food Service | Food Court | Food & Drink Shop | 28.846978 | 77.649733 |
| **6** | 201206 | Ghaziabad | 0 | ATM | Dance Studio | Frozen Yogurt Shop | Fried Chicken Joint | French Restaurant | Food Truck | Food Stand | Food Service | Food Court | Food & Drink Shop | 28.801370 | 77.469718 |
| **7** | 203202 | Noida | 0 | ATM | Dance Studio | Frozen Yogurt Shop | Fried Chicken Joint | French Restaurant | Food Truck | Food Stand | Food Service | Food Court | Food & Drink Shop | 28.402380 | 77.637002 |
| **8** | 203207 | Noida | 0 | ATM | Dance Studio | Frozen Yogurt Shop | Fried Chicken Joint | French Restaurant | Food Truck | Food Stand | Food Service | Food Court | Food & Drink Shop | 28.532525 | 77.638734 |

10. Proceeding with the analysis for the new shopping mall in the Delhi NCR data, I filtered only the shopping mall data from the venue data which we had retrieved
11. Now applied the group by Postal Code and mean of frequency of the Shopping Malls in the Location
12. Applied the K-Means Clustering on the Malls Data with the number of clusters as 5 to view the variance amongst the clusters



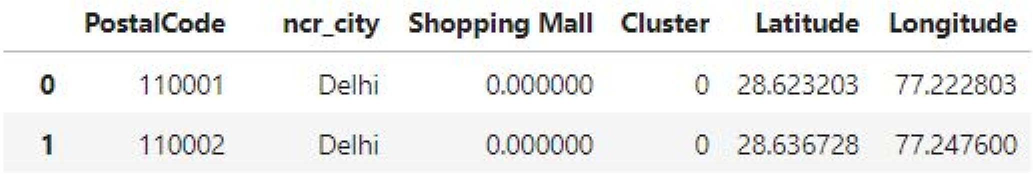
**Cluster Mapping on Shopping Mall Data**

**5 Clusters formed**

## Results

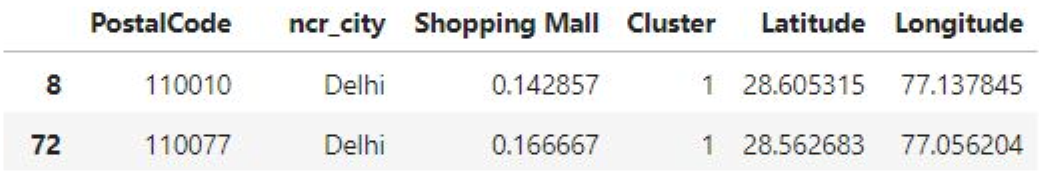
Now proceeding to Examining all the clusters one after the other: (Example Data)

a. Cluster - 0



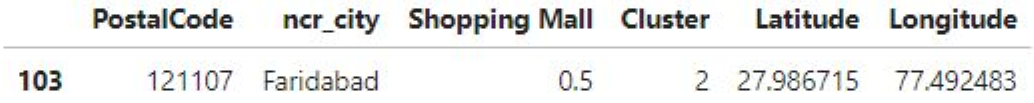
| | PostalCode | ncr_city | Shopping Mall | Cluster | Latitude | Longitude |
|---|---|---|---|---|---|---|
| **0** | 110001 | Delhi | 0.000000 | 0 | 28.623203 | 77.222803 |
| **1** | 110002 | Delhi | 0.000000 | 0 | 28.636728 | 77.247600 |

b. Cluster - 1



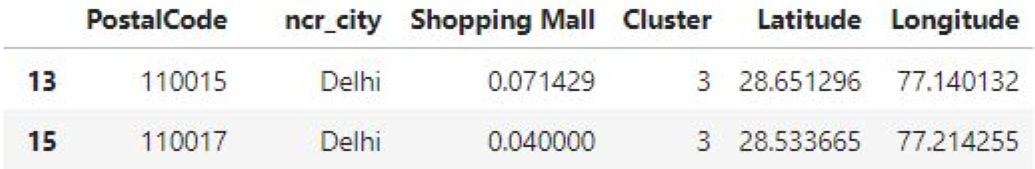
| | PostalCode | ncr_city | Shopping Mall | Cluster | Latitude | Longitude |
|---|---|---|---|---|---|---|
| **8** | 110010 | Delhi | 0.142857 | 1 | 28.605315 | 77.137845 |
| **72** | 110077 | Delhi | 0.166667 | 1 | 28.562683 | 77.056204 |

c. Cluster-2



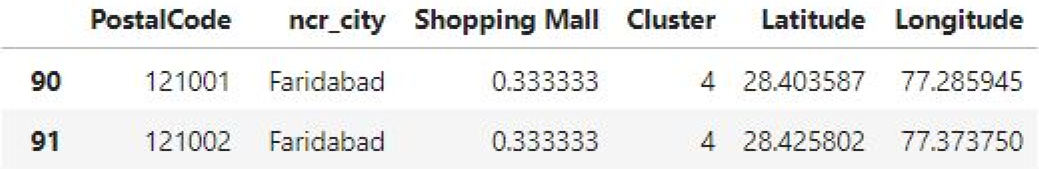
| | PostalCode | ncr_city | Shopping Mall | Cluster | Latitude | Longitude |
|---|---|---|---|---|---|---|
| **103** | 121107 | Faridabad | 0.5 | 2 | 27.986715 | 77.492483 |

d. Cluster-3



| | PostalCode | ncr_city | Shopping Mall | Cluster | Latitude | Longitude |
|---|---|---|---|---|---|---|
| **13** | 110015 | Delhi | 0.071429 | 3 | 28.651296 | 77.140132 |
| **15** | 110017 | Delhi | 0.040000 | 3 | 28.533665 | 77.214255 |

e. Cluster-4



| | PostalCode | ncr_city | Shopping Mall | Cluster | Latitude | Longitude |
|---|---|---|---|---|---|---|
| **90** | 121001 | Faridabad | 0.333333 | 4 | 28.403587 | 77.285945 |
| **91** | 121002 | Faridabad | 0.333333 | 4 | 28.425802 | 77.373750 |

From all the Cluster examples we see above, we can infer that

- Cluster 0 - Locations which have no shopping malls in the vicinity
- Cluster 3 - Locations which have less shopping malls in the vicinity
- Cluster 1 - Locations which have shopping malls in the vicinity
- Cluster 4 - Locations which have a good number shopping malls in the vicinity
- Cluster 2 - Locations which have abundant shopping malls in the vicinity

## Conclusion

- Cluster 2 and Cluster 4 already have many shopping malls in their vicinities
- Cluster 0 has no shopping Malls
- I would suggest the builder or property investor to go build near the ***Cluster 1 and Cluster 3*** as malls which are near Cluster 0 would give rise to Dead Malls as people would not visit a location specifically for a single mall.