Opening a Hotel in Delhi (INDIA) Neighborhood

Business Problem

The objective of this capstone project is to analyze and select the best locations in the city of New Delhi, India to open a new Hotel. Using data science and machine learning techniques this project alms to provide solutions to answer the business problem. The solution recommend the best neighborhood for opening the Hotel.

Target Audience

This project is particularly useful to property developers and investors looking to open or invest in new Hotel in New Delhi, India. This project is timely as the city is currently suffering from oversupply of Hotel. This project will lead to proper availability of Hotels throughout the city. It will result in the benefit of the Investors and the customers.

Data

The data set that I have used for solving the problem is:

- A complete list of neighborhoods in New Delhi, India. Source of the data is Wikipedia.org
- Geographical coordinates (latitude and longitude) of those neighborhoods. Source of the data will be FourSquare.
- FourSquare provided Venue data which is related to Hotels. Machine Learning Technique called Clustering will be used for solving the problem.

Data Sources

This wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi) contains a list of neighborhoods in New Delhi, with a total of 137 neighborhoods. we will use web scraping techniques to extract the data from the wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbors. After that, we will use the Foursquare API to get the venue data for those neighborhoods.

Methodology

Methodology for finding a suitable location for opening a new Hotel in Delhi, India is based on Clustering of venues and places in Neighbourhoods of Delhi. I have grouped the similar venues together on the basis of availability of venues of the different categories.

I have used Machine Learning technique called Clustering for the analysis of venues and places of different categories in the Neighbourhoods of Delhi.

First step is pulling and preprocessing of the data. In have used web scraping python library for pulling the data from wikipedia pages and then I have performed the preprocessing of the data using python data munging library called pandas. The data received from the wikipedia was in JSON format so I have to use JSON parsing for extracting the different parameters of the data.

Second step is all about Exploratory Data Analysis, where I have used few statistical measures and data visualization techniques for understanding the internal structure of the data and the relationship between different parameters of the data.

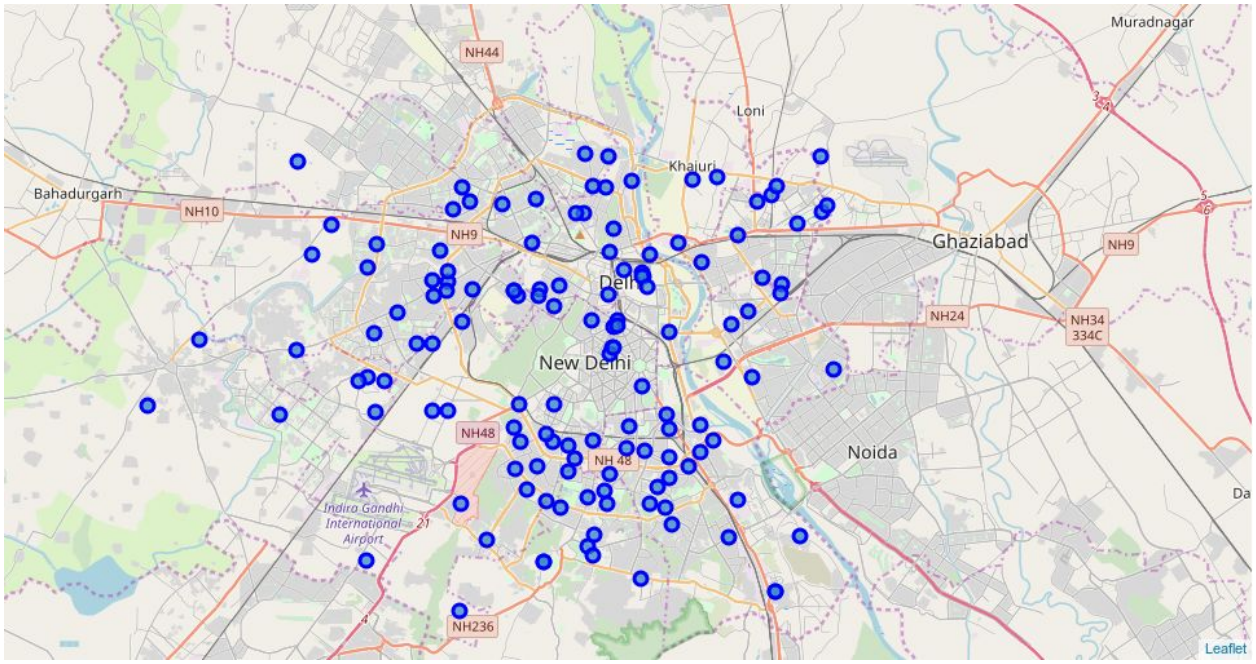
Exploratory Data Analysis has helped me to decide which machine learning technique will be suitable for solving the business problem. I have also used map visualization library (folium) and python library geocoder for fetching the geographical coordinates of different venues and places. I have used Machine Learning technique called Clustering. Clustering unsupervised machine learning where have unlabelled dataset.

Here is a data table consisting of initial 5 neighborhood and their geographical coordinates:

	Neighborhood	Latitude	Longitude
0	Ashok Nagar (Delhi)	28.692230	77.301270
1	Ashok Vihar	28.690420	77.176060
2	Ashram Chowk	28.710568	77.326949
3	Babarpur	28.507370	77.303470
4	Badarpur, Delhi	28.507370	77.303470

The dataset consists of 136 neighbourhoods of Delhi. In the above table only intial five neighbourhoods are shown. (For complete table please refer to IPython notebook).

After geographical coordinates of all the venues and places in the neighborhood of Delhi, I have plot the places data on the map of Delhi. I have used *geocoder* and folium python libraries for visualizing the places data on the map:



For creating the complete data table that I have used for analysis I pulled the data from FourSquare Places data using FourSquare REST APIs. After fetching the data I have parsed the data from the JSON format and created a table:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ashok Nagar (Delhi)	28.69223	77.301270	Neeraj Kumar Garg	28.692731	77.298772	Spa
1	Ashok Vihar	28.69042	77.176060	Domino's Pizza	28.693000	77.177000	Pizza Place
2	Ashok Vihar	28.69042	77.176060	Sagar Ratna | सागर रत्ना	28.693381	77.177977	South Indian Restaurant
3	Ashok Vihar	28.69042	77.176060	Kay's Bar-Be-Que	28.693278	77.173177	BBQ Joint
4	Ashok Vihar	28.69042	77.176060	Kays, Ashok Vihar	28.693572	77.173003	Indian Restaurant
5	Ashok Vihar	28.69042	77.176060	J Block Murga Market	28.687144	77.173035	Market
6	Bali Nagar	28.65218	77.129775	Gianis Ice Cream Parlor	28.651737	77.129924	Dessert Shop
7	Bali Nagar	28.65218	77.129775	Vidhan Sabha metro station	28.654045	77.129745	Light Rail Station
8	Bali Nagar	28.65218	77.129775	Raja garden	28.650680	77.126284	Garden
9	Bali Nagar	28.65218	77.129775	Respawn Gaming Cafe	28.649474	77.133211	Arcade

For applying Machine Learning I have converted categories into new features and I have also converted text data into numerical form so that the algorithm can perform mathematical operations which are an integral part of any machine learning algorithm.

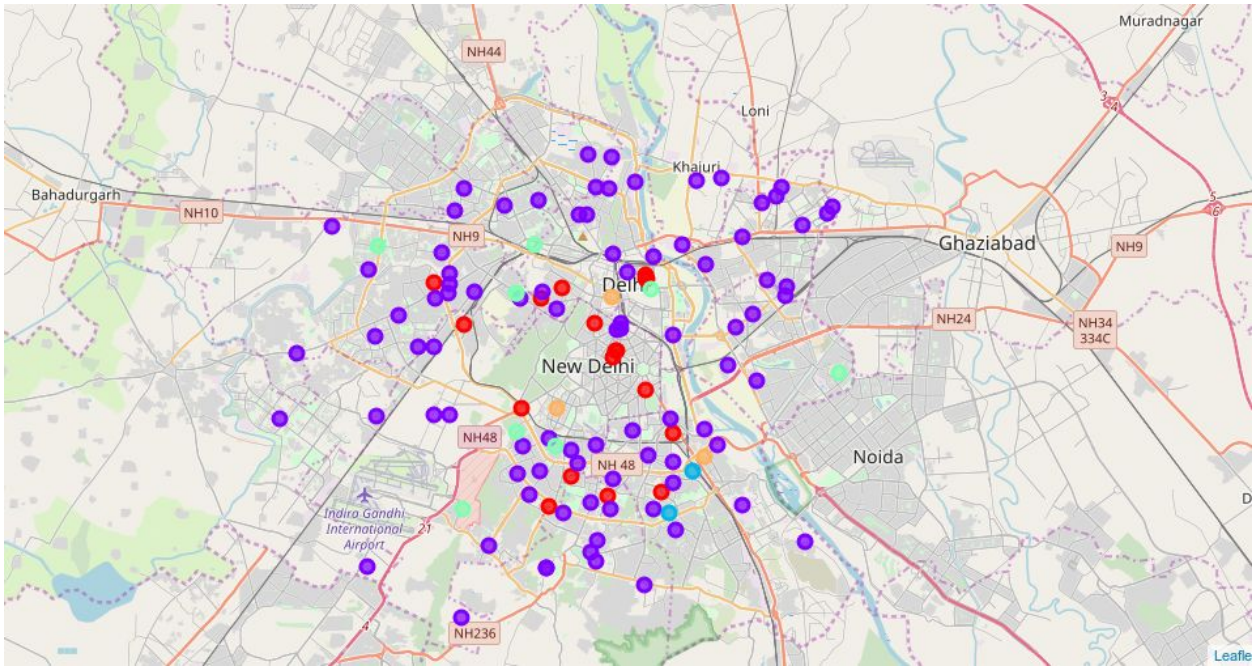
I have used KMeans Clustering algorithm for understanding internal complexity of the data and cluster the similar places and venues together. Total number of clusters are 5.
Here is the data table after applying KMeans Clustering:

	Neighborhood	Hotel	Cluster Label	Latitude	Longitude
0	Ashok Nagar (Delhi)	0.000000	1	28.69223	77.301270
1	Ashok Vihar	0.000000	1	28.69042	77.176060
2	Bali Nagar	0.000000	1	28.65218	77.129775
3	Ber Sarai	0.142857	0	28.54954	77.181700
4	Bhajanpura	0.000000	1	28.69980	77.259170

In the above table Cluster Label column represents the cluster number assigned to the palces.

Result

I have categorized all the venues and places in the Neighborhood of Delhi into 5 different clusters on the basis of availability of different categories of places.
Here is the visualization of all 5 clusters on the map of Delhi, India :



Observation

After carefully looking at the different clusters it can be easily deduced that which locations are suitable for opening a new Hotel. In the below tables the Hotel column shows the information of availability of Hotels nearby respective venues and places

First Cluster:

	Neighborhood	Hotel	Cluster Label	Latitude	Longitude
48	Lutyens' Delhi	0.185185	0	28.621190	77.216710
37	Karol Bagh	0.117647	0	28.650450	77.188730
32	Jangpura	0.200000	0	28.583370	77.247140
119	Yamuna Pushta	0.185185	0	28.621050	77.217100
86	Raisina Hill	0.150000	0	28.618397	77.215478
28	Gulmohar Park	0.083333	0	28.554390	77.212520
24	Golf Links, New Delhi	0.100000	0	28.603040	77.232690
23	Gole Market	0.100000	0	28.634080	77.205760
43	Krishna Nagar, Delhi	0.200000	0	28.563640	77.193670
76	Old Delhi	0.125000	0	28.654320	77.232590
34	Kailash Colony	0.190476	0	28.556090	77.240610
14	Dhaula Kuan	0.142857	0	28.594895	77.167263
88	Rajendra Place	0.090909	0	28.645460	77.177760
6	Chandni Chowk	0.153846	0	28.656240	77.232330
108	Shivaji Place	0.071429	0	28.652650	77.121390
3	Ber Sarai	0.142857	0	28.549540	77.181700
9	Dariba Kalan	0.133333	0	28.654568	77.233419
67	Naraina Area	0.111111	0	28.633650	77.136740

First cluster contains locations which has lower to moderate availability of Hotels.

Second Cluster: This cluster has largest no of places (Only few of them are shown here)

	Neighborhood	Hotel	Cluster Label	Latitude	Longitude
63	Munirka	0.000000	1	28.555030	77.171270
64	Nanakpura	0.000000	1	28.622830	77.113360
87	Rajendra Nagar, Delhi	0.000000	1	28.640658	77.185701
65	Nand Nagri	0.000000	1	28.696700	77.303860
66	Nangloi Jat	0.000000	1	28.678560	77.067640
85	Punjabi Bagh	0.000000	1	28.666330	77.125250
84	Pitam Pura	0.000000	1	28.695890	77.137260
72	New Friends Colony	0.000000	1	28.578100	77.269990
71	New Delhi	0.014925	1	28.630950	77.217220
68	Narela	0.000000	1	28.839770	77.076930
79	Palika Bazaar	0.015152	1	28.631580	77.219590
78	Palam	0.000000	1	28.591080	77.091190
75	Nizamuddin West	0.000000	1	28.589730	77.245220
69	Naveen Shahdara	0.000000	1	28.673690	77.283260
74	Nigambodh Ghat	0.000000	1	28.664750	77.236360
73	New Moti Bagh	0.000000	1	28.580997	77.181823

As seen in the Hotel column of the data this cluster has the Very Low/None availability of Hotels.

Third Cluster

	Neighborhood	Hotel	Cluster Label	Latitude	Longitude
111	Sriniwaspuri	0.666667	2	28.565680	77.257330
80	Pamposh Enclave	0.571429	2	28.546776	77.244759

This cluster has only two places but the availability score of the hotels is very high.

Fourth Cluster

	Neighborhood	Hotel	Cluster Label	Latitude	Longitude
118	West Patel Nagar	0.250000	3	28.64780	77.164470
114	Urdu Bazaar	0.285714	3	28.64989	77.235145
57	Mayur Vihar Phase - 3	0.333333	3	28.61125	77.334060
60	Moti Bagh	0.250000	3	28.58363	77.164720
51	Mahipalpur	0.250000	3	28.54843	77.136360
104	Shakti Nagar, Delhi	0.333333	3	28.67037	77.174140
82	Paschim Vihar	0.333333	3	28.66933	77.091730
83	Patel Nagar	0.250000	3	28.64780	77.164470
70	Netaji Nagar, Delhi	0.250000	3	28.57747	77.185160

This cluster also contains less number of places but the availability of hotels is good if not the best.

Fifth Cluster

	Neighborhood	Hotel	Cluster Label	Latitude	Longitude
5	Chanakyapuri	0.500000	4	28.59506	77.18573
77	Paharganj	0.473684	4	28.64596	77.21493
50	Maharani Bagh	0.428571	4	28.57223	77.26357

Conclusion

After careful analysis of all five clusters It's clear that the Places which are part of Second cluster (Cluster Label - 1) are most suitable for opening a new Hotel. Second cluster has the least no of existing hotels but It has very good connectivity to other popular public places.

This analysis suggests to **open a new Hotel in the Second cluster (Cluster Label - 1)**