APPLIED DATA SCIENCE CAPSTONE PROJECT BY IBM AND COURSERA

FINDING LOCATIONS TO OPEN A HOTEL IN KAROL BAGH (DELHI), INDIA

Table of contents:

* Introduction / Business Problem
* Data
* Methodology
* Results
* Discussion
* Conclusion

Introduction / Business Problem:

The aim of this project is to find suitable locations to open a hotel in karol bagh greater metropolitan area [Karol bagh and its suburbns],Delhi (India).

The first requirement is that the new hotel should be easily accessible by its prospective customers and more specifically it should be located near a metro station. The number of hotels already existing in an area should also be considered so that fierce competition be avoided if possible.

Apart from the obvious intended stakeholders, entrepreneurs looking to start a hotel business, similar methodology could be used for other specific types of businesses. It can serve as an initial starting point of locations to consider to start their business.

For the project objectives to be achieved, python geolocation libraries will be used, along with the Foursquare API. Also, in order to create clusters of similar candidate locations, the K-Means machine learning clustering algorithm will be used.

Data:

The necessary data for this project, based on the above stated requirements, are:

The metro stations in the karol bagh greater metropolitan area
Number of existing hotels near each station
In addition, the distance to the nearest hotels for every metro station will be used
In order to obtain the data, a combination of the geopy Python library and the Foursquare API will be used:

1. 'Karol Bagh Market' will be considered as the center of Karol Bagh,Delhi. It is indeed one of the most central location in the city. I will obtain its geospatial coordinates using the geopy library

2. Having the coordinates of the 'center' of Karol Bagh, the Foursquare API will be used to retrieve data for all the metro stations in Athens greater area in a radius of 15 km

3. To find the existing hotels near the metro stations, the Foursquare API will again be used for every station. I will obtain data for all the hotels located in a radius of 1000 meters of every metro station

Using the collected data, I will calculate the number of existing hotels near each station. I will also be able to determine the minimum distance to a hotel for every metro station from the 3rd step of the above process. This minimum distance to every metro station from a hotel, along with the number of already existing hotels near the station will be used as input to K-Means clustering algorithm to obtain the clusters of areas (metro stations).

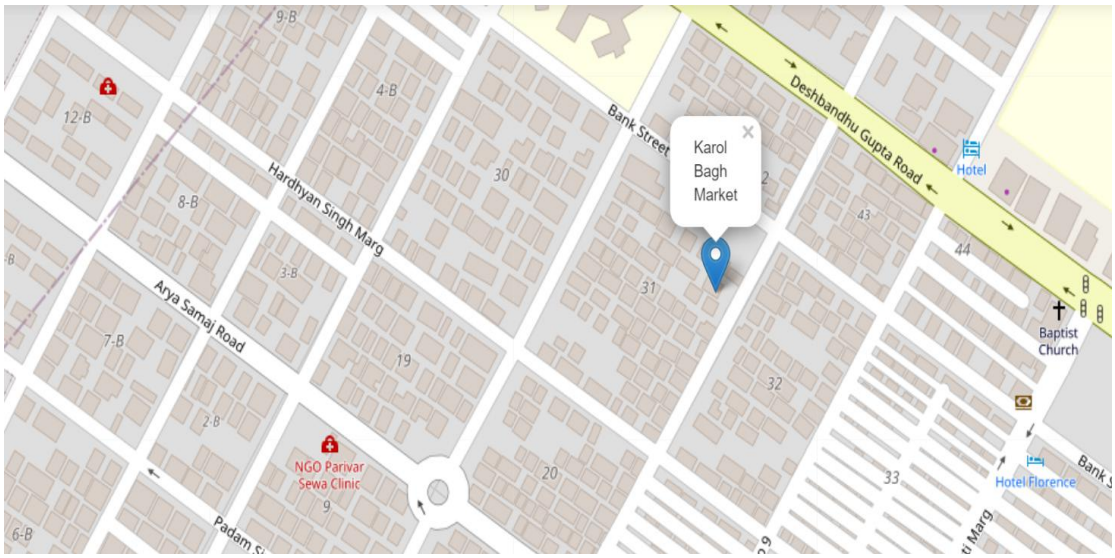
Methodology:

The objective of this project is to obtain information about metro stations in the greater metropolitan karol bagh area with potential for opening a hotel, and having as criteria:

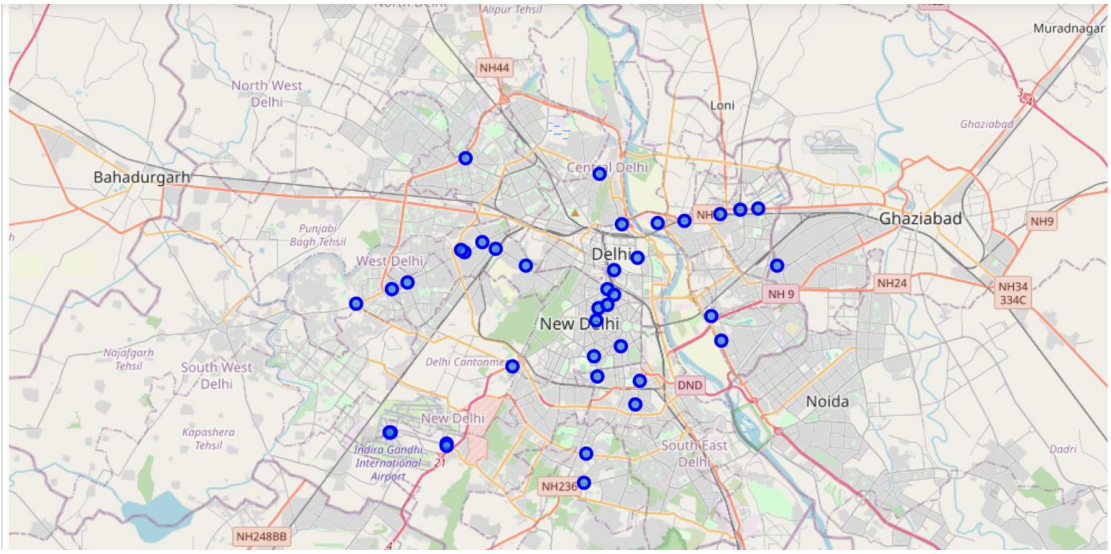
* Low number of already existing hotels
* Minimum distance of existing hotels to its nearby station

The steps I will follow to identify potential areas (metro stations) are:

1. Consider '**Karol Bagh Market**' as the center of **karol bagh**.



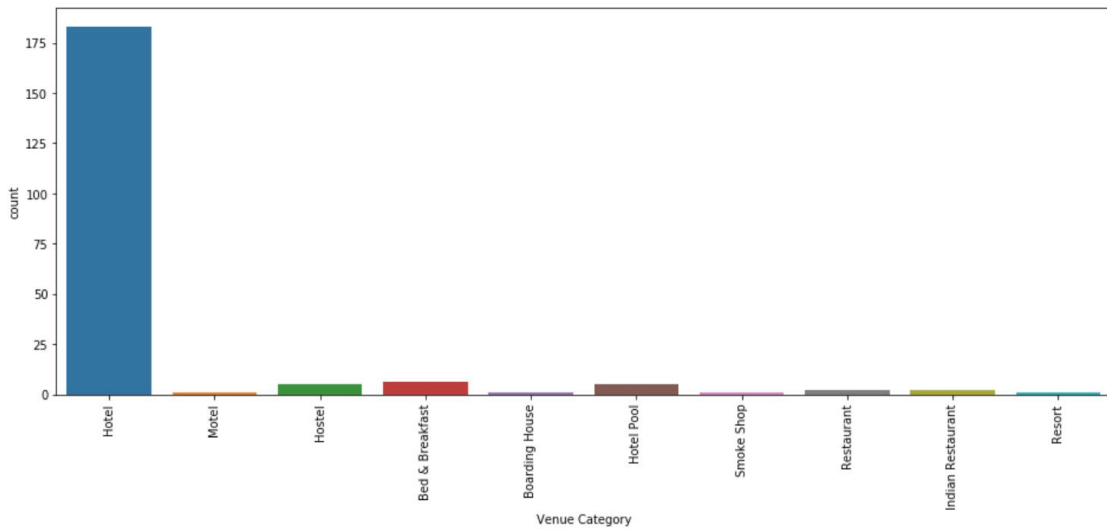
2. Based on the coordinates of **Karol Bagh Market,** I will obtain information about **metro stations in a radius of 15 km** using the **Foursquare API**. At this stage I will remove from the above dataset points that although they are identified as 'Metro stations' by the Foursquare API, they are only used as depots or maintenance gathering for the metro carriages.



```
RangeIndex: 38 entries, 0 to 37
Data columns (total 6 columns):
station                 38 non-null object
Venue Type              38 non-null object
lat                     38 non-null float64
lng                     38 non-null float64
Distance from Center    38 non-null int64
Postal Code             38 non-null object
dtypes: float64(2), int64(1), object(3)
memory usage: 1.9+ KB
None
```

	station	Venue Type	lat	lng	Distance from Center	Postal Code
0	Rajiv Chowk | राजीव चौक Metro Station	Metro Station	28.632981	77.219288	3702	110001
1	New Delhi | नई दिल्ली Metro Station	Metro Station	28.642803	77.223351	3540	110001
2	Central Secretariat Metro Station	Metro Station	28.615908	77.212558	4725	110001
3	Kirti Nagar Metro Station	Metro Station	28.654676	77.151075	3711	110015
4	Kashmere Gate Metro Station	Metro Station	28.667746	77.228304	4173	110006

3. After the collection of metro stations information, I will again use the Foursquare API to locate all the existing hotels in a radius of 1000 meters from each station



Venue Category	station	lat	lng	Venue	Venue Latitude	Venue Longitude	Distance from Station
Bed & Breakfast	6	6	6	6	6	6	6
Boarding House	1	1	1	1	1	1	1
Hostel	5	5	5	5	5	5	5
Hotel	183	183	183	183	183	183	183
Hotel Pool	5	5	5	5	5	5	5
Indian Restaurant	2	2	2	2	2	2	2
Motel	1	1	1	1	1	1	1
Resort	1	1	1	1	1	1	1
Restaurant	2	2	2	2	2	2	2
Smoke Shop	1	1	1	1	1	1	1

4. At this stage I will only keep as my data set the results that correspond only to 'Hotel Pool','Bed & Breakfast' and 'Hotel' subcategories. I will remove the rest of the subcategories such as 'Boarding House','Indian Restaurant','Motel', etc.

Venue Category	station	lat	lng	Venue	Venue Latitude	Venue Longitude	Distance from Station
Bed & Breakfast	6	6	6	6	6	6	6
Hotel	183	183	183	183	183	183	183
Hotel Pool	5	5	5	5	5	5	5

5. I will ignore for the purpose of clusttering any metro stations that based on the results of the Foursquare API don't have any existing hotels in their vicinity. For those, there can either be no existing data in the Foursquare database, or indeed there no existing hotels near the corresponding station

	station	lat_x	lng_x	lat_y	lng_y	Min Distance from Station	Hotels Count
3	Kirti Nagar Metro Station	28.654676	77.151075	NaN	NaN	NaN	NaN
6	Vishwavidyalaya Metro Station	28.695030	77.214718	NaN	NaN	NaN	NaN
10	Moti Nagar Metro Station	28.657830	77.142445	NaN	NaN	NaN	NaN
12	Ramesh Nagar Metro Station	28.652722	77.131657	NaN	NaN	NaN	NaN
13	Shastri Park Metro Station	28.668430	77.250559	NaN	NaN	NaN	NaN
16	Mayur Vihar Phase I Metro Station	28.604851	77.289577	NaN	NaN	NaN	NaN
17	Dhaula Kuan Airport Metro Express Station	28.591429	77.161017	NaN	NaN	NaN	NaN
20	Jor Bagh Metro Station	28.585828	77.213219	NaN	NaN	NaN	NaN
21	Pitampura Metro Station	28.703198	77.132263	NaN	NaN	NaN	NaN
23	Tilak Nagar Metro Station	28.636513	77.096429	NaN	NaN	NaN	NaN
24	Seelampur Metro Station	28.669805	77.266846	NaN	NaN	NaN	NaN
26	Akshardham Metro Station	28.618362	77.283052	NaN	NaN	NaN	NaN
30	Shahdara Metro Station	28.673344	77.289011	NaN	NaN	NaN	NaN
31	Jhilmil Metro Station	28.675876	77.312309	NaN	NaN	NaN	NaN

So the seneteen metro stations for which no data were returned from the Foursquare API are:

- Kirti Nagar Metro Stationn
- Kashmere Gate Metro Station
- Vishwavidyalaya Metro Station
- Moti Nagar Metro Station
- Ramesh Nagar Metro Station
- Shastri Park Metro Station
- Mayur Vihar Phase I Metro Station
- Dhaula Kuan Airport Metro Express Station
- Azadpur | आजादपुर Metro Station
- Jor Bagh Metro Station
- Pitampura Metro Station
- Tilak Nagar Metro Station
- Seelampur Metro Station
- Akshardham Metro Station
- Janak Puri East Metro Station
- Shahdara Metro Station
- Jhilmil Metro Station

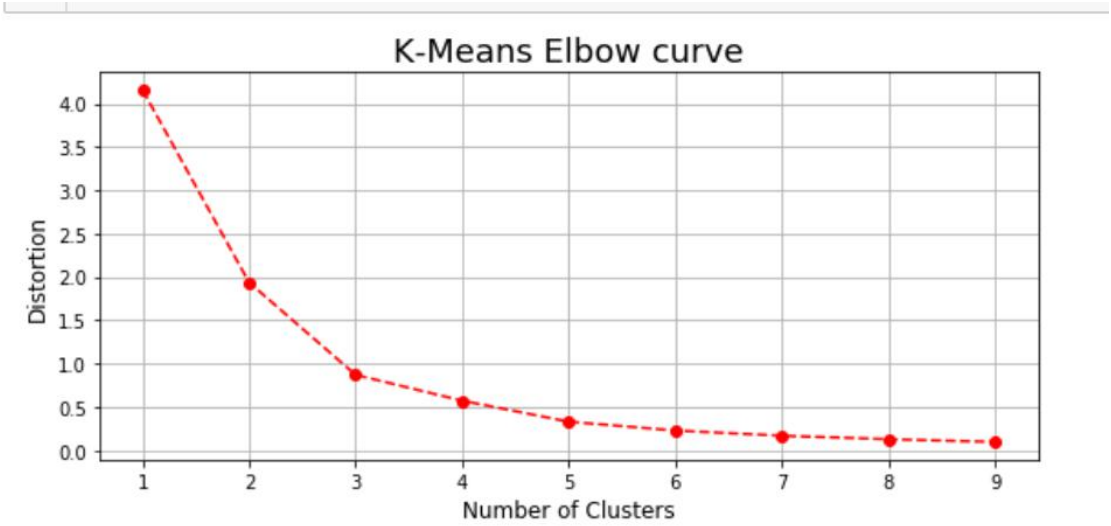
6. Having the information about hotels around metro stations, I will calculate the number of existing hotels near each station as well as the minimum distance from each station to a hotel using available python statistical functions.

	station	lat	lng	Min Distance from Station	Hotels Count
0	Rajiv Chowk | राजीव चौक Metro Station	28.632981	77.219288	167	19
19	New Delhi | नई दिल्ली Metro Station	28.642803	77.223351	205	29
48	Central Secretariat Metro Station	28.615908	77.212558	621	9
57	Kashmere Gate Metro Station	28.667746	77.228304	759	1
58	Barakhamba Road Metro Station	28.629993	77.223830	267	13
71	Patel Nagar Metro Station	28.645115	77.169256	345	6
77	Patel Chowk Metro Station	28.622511	77.213679	347	12
89	Janpath Metro Station	28.624426	77.219187	135	14
103	Jama Masjid Metro Station	28.649631	77.237812	536	5
108	Race Course Metro Station	28.596674	77.211206	692	2

7. The data will be normalized so that both factors (minimum distance, number of existing hotels) will have equal weight when they will be used by a machine learning method

	station	lat	lng	Min Distance from Station	Hotels Count	Norm Min Distance from Station	Norm hotel Count
0	Rajiv Chowk | राजीव चौक Metro Station	28.632981	77.219288	167	19	-0.251636	0.389881
1	New Delhi | नई दिल्ली Metro Station	28.642803	77.223351	205	29	-0.212379	0.747024
2	Central Secretariat Metro Station	28.615908	77.212558	621	9	0.217373	0.032738
3	Kashmere Gate Metro Station	28.667746	77.228304	759	1	0.359935	-0.252976
4	Barakhamba Road Metro Station	28.629993	77.223830	267	13	-0.148330	0.175595
5	Patel Nagar Metro Station	28.645115	77.169256	345	6	-0.067751	-0.074405
6	Patel Chowk Metro Station	28.622511	77.213679	347	12	-0.065685	0.139881
7	Janpath Metro Station	28.624426	77.219187	135	14	-0.284694	0.211310
8	Jama Masjid Metro Station	28.649631	77.237812	536	5	0.129563	-0.110119
9	Race Course Metro Station	28.596674	77.211206	692	2	0.290720	-0.217262
10	Khan Market Metro Station	28.602285	77.227860	130	3	-0.289859	-0.181548
11	Azadpur | आजादपुर Metro Station	28.654047	77.129642	993	1	0.601670	-0.252976
12	Lajpat Nagar Metro Station	28.570607	77.236510	403	5	-0.007834	-0.110119
13	Jangpura Metro Station	28.583052	77.239316	176	3	-0.242338	-0.181548
14	Hauz Khas | हौज खास Metro Station	28.543979	77.206263	675	4	0.273158	-0.145833

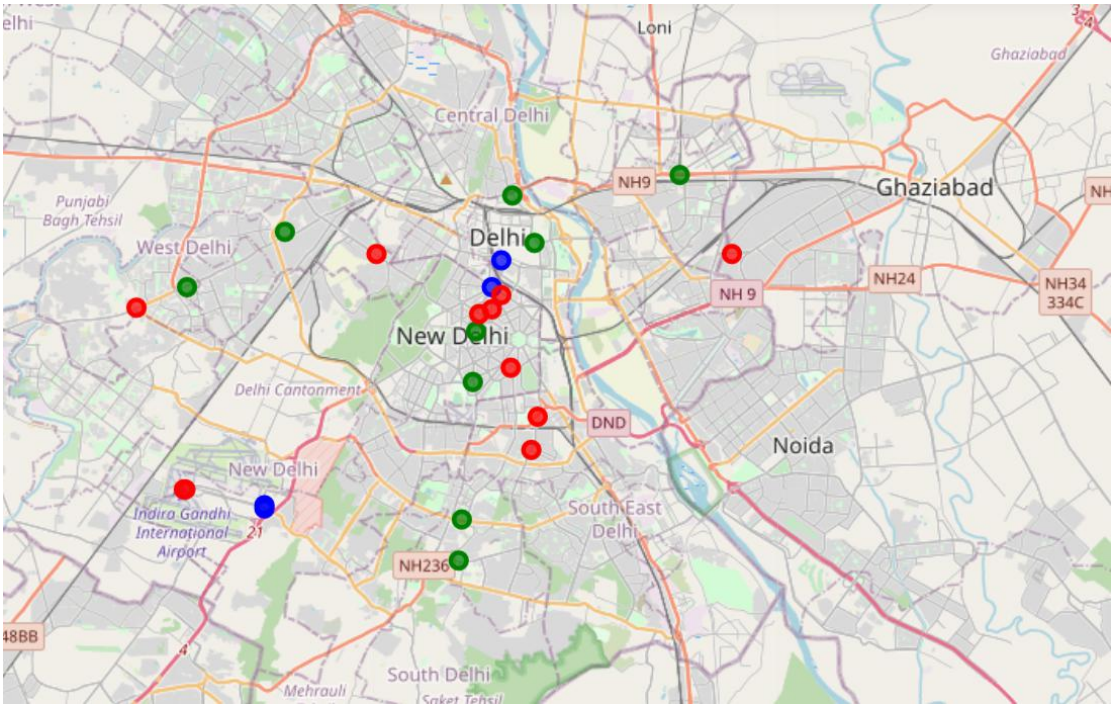
8. The **K-Means Machine Learning clustering algorithm** will be used to divide the stations and hotels data set into clusters of similar locations.
9. The Elbow method will be used to find the most suitable number of clusters.



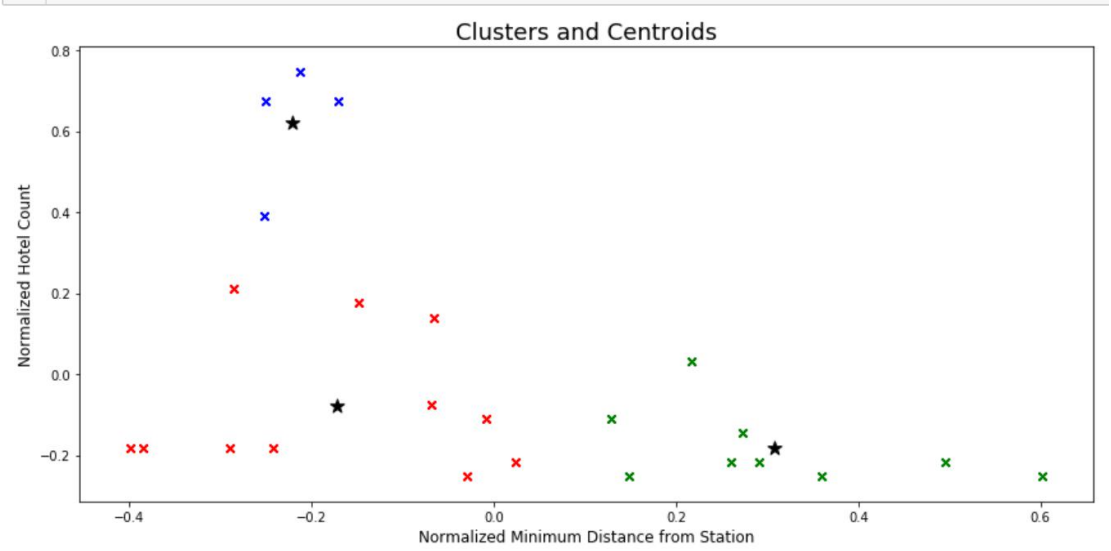
From this cluster we can decide the most suitable location in that area

Result:

After executing K-Mean clustering algorithm three cluster found near metro station ,identified by their respecting colour on following map:



Graph of distribution of clusters and their centroid in black based on normalizes values



The following are main three type of cluster

Cluster 1 (AVERAGE potential)

* Although not a prohibitive metro station to open a hotel in its vicinity, there is already a fair number of hotels in the area and the nearest one is not far from the metro station

	station	Cluster Label	lat	lng	Min Distance from Station	Hotels Count	Norm Min Distance from Station	Norm hotel Count
4	Barakhamba Road Metro Station	0	28.629993	77.223830	267	13	-0.148330	0.175595
5	Patel Nagar Metro Station	0	28.645115	77.169256	345	6	-0.067751	-0.074405
6	Patel Chowk Metro Station	0	28.622511	77.213679	347	12	-0.065685	0.139881
7	Janpath Metro Station	0	28.624426	77.219187	135	14	-0.284694	0.211310
10	Khan Market Metro Station	0	28.602285	77.227860	130	3	-0.289859	-0.181548
12	Lajpat Nagar Metro Station	0	28.570607	77.236510	403	5	-0.007834	-0.110119
13	Jangpura Metro Station	0	28.583052	77.239316	176	3	-0.242338	-0.181548
15	Kaushambi Metro Station	0	28.645475	77.324144	434	2	0.024191	-0.217262
17	Indira Gandhi International Airport Metro Station	0	28.555696	77.086223	25	3	-0.398330	-0.181548
21	Uttam Nagar East Metro Station	0	28.624648	77.065229	383	1	-0.028495	-0.252976
22	IGI Airport	0	28.555872	77.085881	39	3	-0.383867	-0.181548

Cluster 2 (LOW potential)

* There are already many existing hotels in the area and the nearest hotels is in most cases in a relatively short distance from the metro station

	station	Cluster Label	lat	lng	Min Distance from Station	Hotels Count	Norm Min Distance from Station	Norm hotel Count
0	Rajiv Chowk | राजीव चौक Metro Station	2	28.632981	77.219288	167	19	-0.251636	0.389881
1	New Delhi | नई दिल्ली Metro Station	2	28.642803	77.223351	205	29	-0.212379	0.747024
19	Airport Express Line Metro Station	2	28.549347	77.120869	169	27	-0.249570	0.675595
23	Delhi Aerocity Airport Metro Express Station	2	28.548632	77.120558	246	27	-0.170024	0.675595

Cluster 3 (HIGH potential)

* There are only few already existing hotels in the area and the nearest hotels is in most cases relatively not in a short distance to the metro station

	station	Cluster Label	lat	lng	Min Distance from Station	Hotels Count	Norm Min Distance from Station	Norm hotel Count
2	Central Secretariat Metro Station	1	28.615908	77.212558	621	9	0.217373	0.032738
3	Kashmere Gate Metro Station	1	28.667746	77.228304	759	1	0.359935	-0.252976
8	Jama Masjid Metro Station	1	28.649631	77.237812	536	5	0.129563	-0.110119
9	Race Course Metro Station	1	28.596674	77.211206	692	2	0.290720	-0.217262
11	Azadpur | आजादपुर Metro Station	1	28.654047	77.129642	993	1	0.601670	-0.252976
14	Hauz Khas | हौज खास Metro Station	1	28.543979	77.206263	675	4	0.273158	-0.145833
16	Janak Puri East Metro Station	1	28.632935	77.086751	890	2	0.495265	-0.217262
18	Mansarover Park Metro Station	1	28.675370	77.300932	555	1	0.149191	-0.252976
20	Malviya Nagar Metro Station	1	28.528696	77.204997	662	2	0.259728	-0.217262

Discussion:

Possible areas that were not in the Foursquare database should also be examined so that it can be determined if it is just lack of data about these stations or indeed there are no hotels in the vicinity of the stations. A lot more factors can be considered when choosing an appropriate location. Some examples of extra factors can be:

* Area near Airport side so travelling connecting people can wait at airport or take a room in hotel
* Population density in the area
* Per day rent in this area is this area expensive or not
* Near by surrouding Airport
* Number of people daily visitng this area
* Property prices in the area
* Number of businesses operating in the area
* Final decission on optimal hotel 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 park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.

Conclusion:

* The above results can be a good starting point for a prospective businessman that is interested in opening a hotel. Similar methodology can be used for other types of businesses probably with customized criteria.

* Delhi is Main city in India ,and mostly all connecting international flights go from Delhi Airport .so many people can stay in this hotel this area is near to airport area and only few hotels are there .

* Due to few hotels their is more chance to ppeople can book their.

* With the availability of a number of different tools and Machine Learning algorithms, it is possible to find solutions (or possible solutions) to an ever increasing number of problems and queries.

* And it is getting better and better!

References:

* Karol Bagh - Wikipedia
* Google (many postal code are NaN so for cleaning the data)
* Forsquare API
* Google Map
* K-Mean Cluster