2021

DATA SCIENTIST

MANISH KUMAR

[BATTLE OF NEIGHBORHOODS]

[This report consists of a applied Data Science project. This project consists of a problem and a solution to it , the problem of finding the best location for my new hotel/restaurant in Chandigarh.]

Contents

1. Introduction/Business Problem

(a) Discussion of the business problem and the audience who would be interested in this project.

(i). Something about the tourist place - CHANDIGARH, INDIA

I am from Chandigarh, a beautiful tourist spot in nothern India. Chandigarh, the capital of the nothern Indian states of Punjab and Haryana, was designed by the Swiss-French mordernist architect, Le Carbusier. His buildings include the capitol complex with its High Court, Secretariat and Legislative Assembly, as well as giant Open Hand Monument. The nearby Rock Garden is a park featuring sculptures made of stones, recycled ceramics and industrial relics. The most soothning location is the Sukhna Lake, one the most attracting location for tourists.

(ii). Opening of Hotel/Restaurant Shop

Coming down to business problem , I would like to open a hotel/restaurant in the centre most part of it. As it is a famous tourist spot, there is already lots of attention towards it. I know there will be many competitors in terms of hotel and restaurant. But keeping them in mind, i need to locate my hotel in place where more people are attracted and comfortable for a stay and a good meal. I want to bring foreign and local peoples attention towards my new hotel. I would like to flavor my retaurant recipe with Italian, American, typical south & north indian foods to grab their taste.

The challenge is to find a suitable location for opening a new hotel / restaurant attracted to all local and foreign people in the centre of all famous venues

(iii). Expected / Interested Audience

85% indians and 15% foreign peoples visit Chandigarh once in a year. Some people stay for couple of days or more. Also they find some place for hangout or a good meal. Their main focus might be belonging to stay somewhere near to reach venues or to the main hub where everything is easily available. Apart from these set of people, students and working professionals are common audience here. So we may need to fascinate them all.

2. DATA DESCRIPTION

(a) Data used

We will be completely working on Foursquare data to explore and try to locate our new hotel, as its prime data gathering source as it has a database of millions of places, especially their places API which provides the ability to perform location search, location sharing and details about a business.

We will need data about different venues in different neighborhoods of that specific borough. In order to gain that information we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

- Area
- Area Latitude
- Area Longitude
- Venue
- Name of the venue e.g. the name of a store or restaurant
- Venue Latitude
- Venue Longitude
- Venue Category

How it will be used to solve the problem?

We will looking for midpoint area of venues to locate our new hotel. Before that our major focus will be on all venues present in and around the core place of Chandigarh.

Just a heads up on how many hotels are distributed now around Chandigarh. We will perform some EDA on hotels & restaurants present in the tourist spot. On furthur notebook we will use Foursquare data to determine other venues as well. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of venues, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 1000 meter.

The data of the hotels in the radius of 1000 meter is extracted using Foursquare API and listed as in above table.

	name	categories	distance	lat	lng	id
0	Taj Hotel	Hotel	636	30.745340	76.785161	4bc8db62762beee1d69a3d38
1	Hotel Shivalik View	Hotel	583	30.739890	76.776496	4e3ec5391495bf24a5ec0c51
2	Hotel Corporate Inn	Hotel	746	30.746130	76.785923	4bd79397304fce7226b833ab
3	Hotel Piccadily	Hotel	937	30.733332	76.776710	4bd10e6620cd9960c4ae2e9e
4	Hotel Komfort Inn	Hotel Pool	103	30.739159	76.782327	5167015ee4b0e36021fdc146

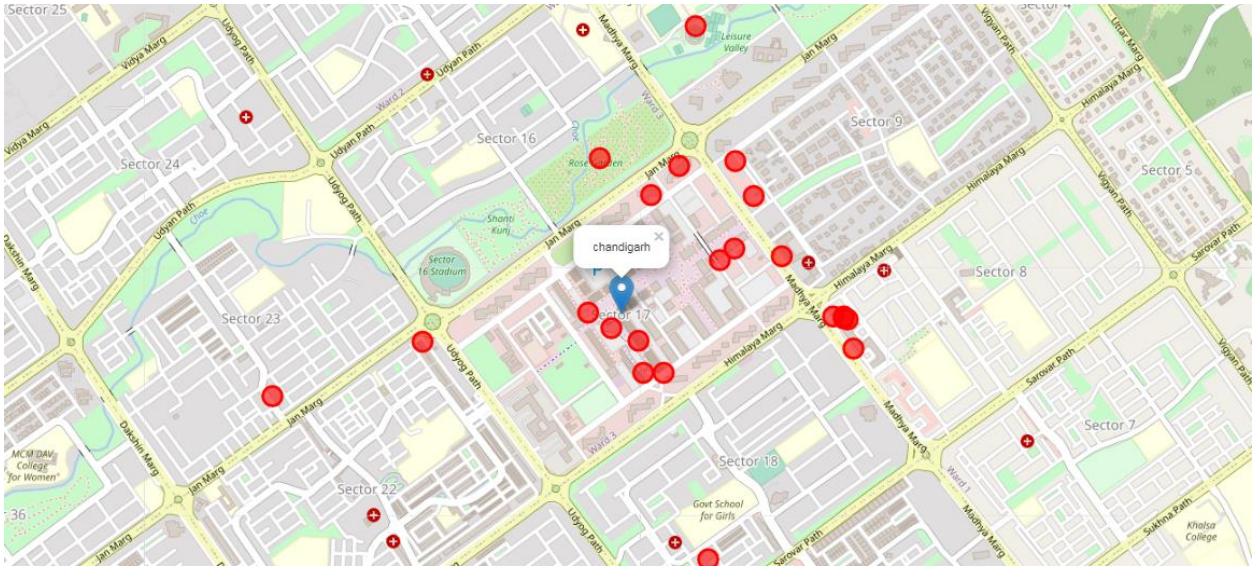
Fig. Data of nearby Hotels extracted by Foursquare

Further the famous venues data is also extracted as per our need and the ratings and tips are also extracted from the Foursquare API. Some places weren't listed in the provided data so we performed a search query .

3. Methodology section

In this sections we will perform some data analysis and EDA to find insight from data. We will try to understand the current stats of all given data. Probably, clustering or centroid of all venues will help us to locate new hotel.

So now using folium we plotted all the venues successfully on the map as shown above.



Unsupervised Learning

Unsupervised learning was carried out in order to find out the similarities between found similarities between neighborhoods. K-Means, a clustering algorithm, was implemented. In this case K-Means is used due to its simplicity and its similarity approach to find patterns.

K-Means: K-Means is a clustering algorithm. This algorithm search clusters within the data and the main objective function is to minimize the data dispersion for each cluster. Thus, each group found represents a set of data with a pattern inside the multi-dimensional features. It is necessary for this algorithm to have a prior idea about the number of clusters since it is considered an input of this algorithm.

```
In [108]: # Clustering

          # set number of clusters
          kclusters = 3

          neighbor_grouped_clustering = neighbor_grouped.drop('name', 1)

          # run k-means clustering
          kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(neighbor_grouped_clustering)

          # check cluster labels generated for each row in the dataframe
          kmeans.labels_[0:10]


          # add clustering labels
          neighborhoods_venues_sorted.insert(0, 'Clustersss', kmeans.labels_)

          neighbor_merged = final_venues

          # merge toronto_grouped with toronto_data to add latitude/longitude for each neighborhood
          neighbor_merged = neighbor_merged.join(neighborhoods_venues_sorted.set_index('name'), on='name')

          kmeans

Out[108]: KMeans(n_clusters=3, random_state=0)
```

Fig. Clustering using K-Means

Plotting

Various plotting techniques we used as well in order to visualize the data. Visualizing data often gives a clear understanding of the data as it is easier to spot patterns in a visualized data as compares to quantitative data.

Folium: Folium library was used to plot maps of Manchester city as well as neighborhoods. Folium was also used to visualize the cluster data.

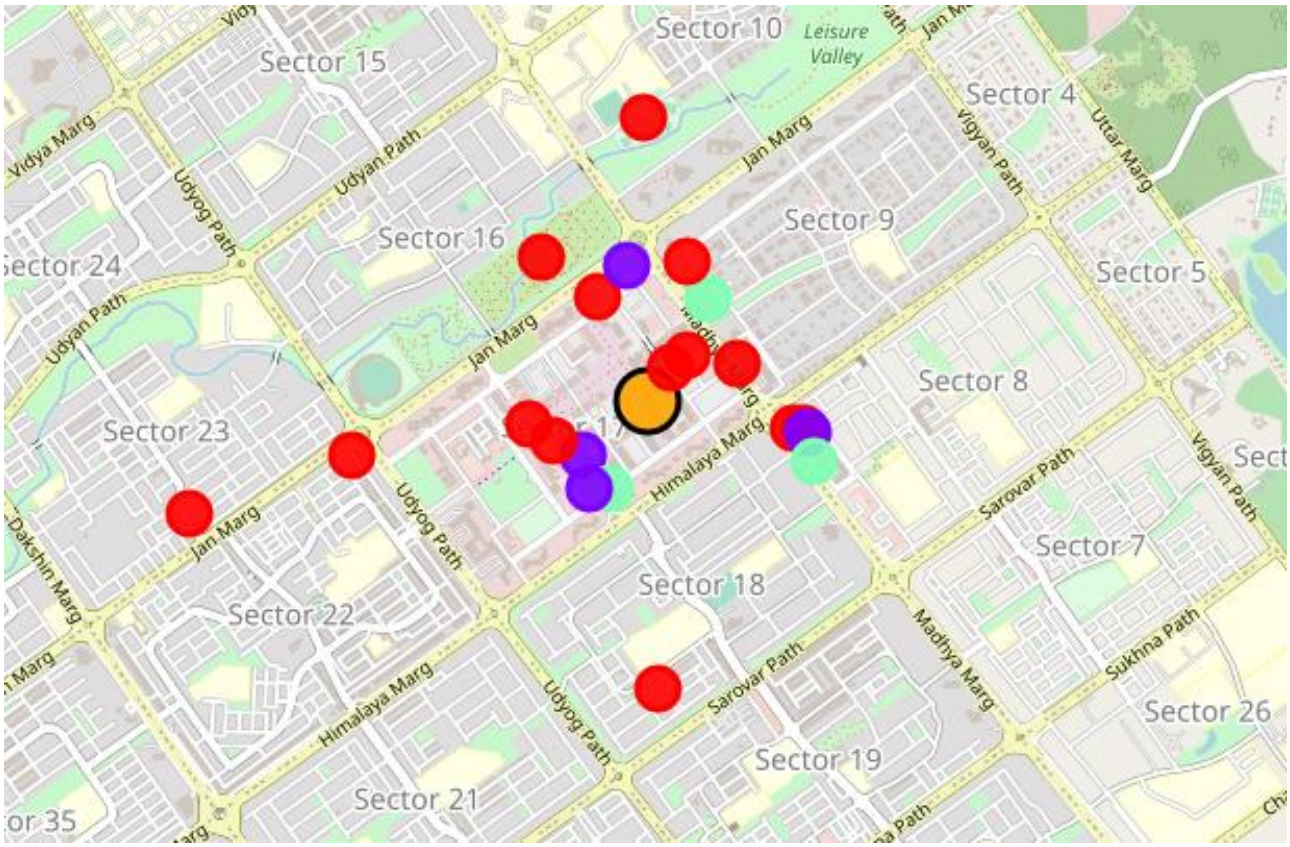


Fig. Clustered map

In this above map you can see the orange marker is our hotel location that we got from clustering and the other markers of red, purple, and light green shows the different clusters formed using k-means clustering.

4. Results section

(a) My hotel location

- Final location is pointed at 30.7409581,76.7860204
- This location is at the centre of the main market of sector 17.
- Located at exact junction , which can give more attention to people who pass by.

(b) Top Rated Venues

- Ghazal
- Girl in cafe
- Sector-17
- Rose Garden

All these venues are rated well than other and located within 600 metres to core location of Chandigarh . So tourists may like to visit these places.

(c) Spot my hotel against others

- popup marker - My hotel location
- Green - Chandigarh core location.
- orange - Venues.
- Blue - Other hotels.
- My predicted location and core location are very close to each other which is expected. As this has central attraction, the predicted one almost matched with the core.

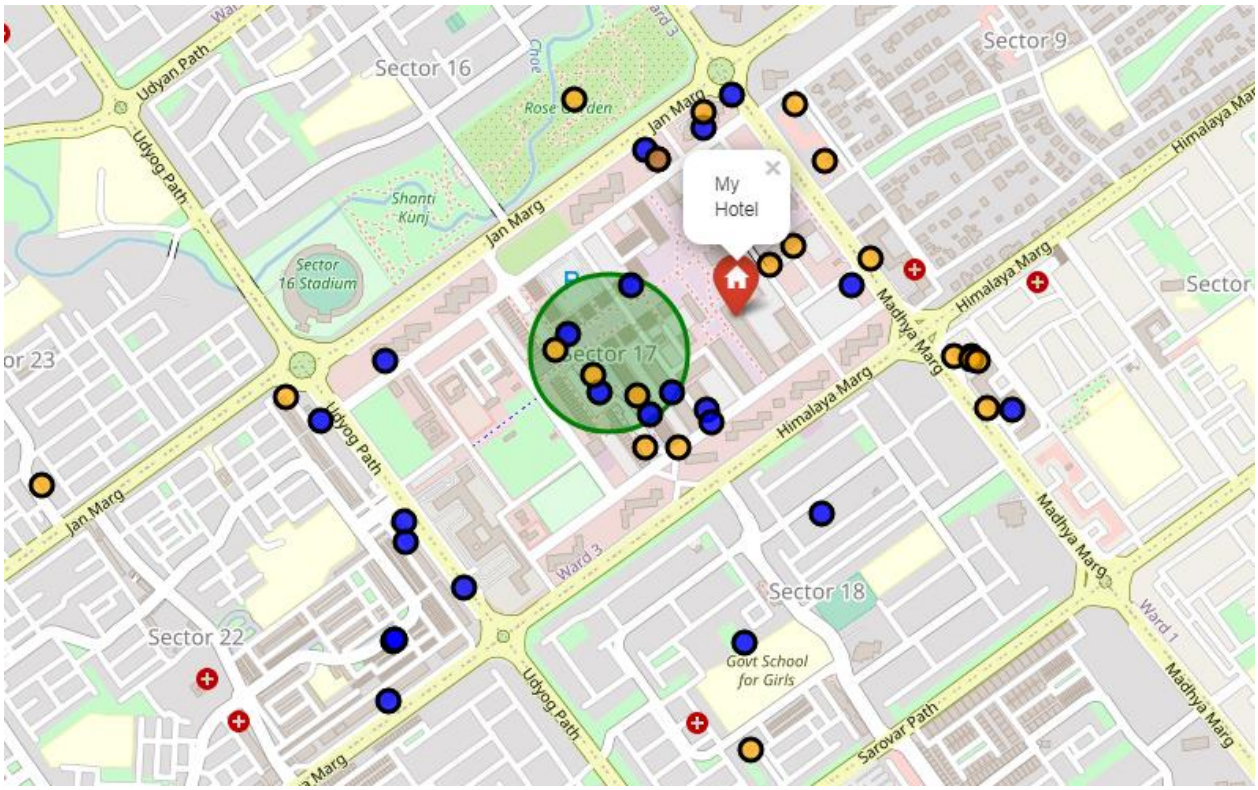


Fig. Map showing venues, hotels, and our Hotel

• The big green marker indicates the core location .

• The blue markers indicate the hotels/restaurants in nearby places.

• The yellow markers indicates the famous venues around the area.

• The pop-up shows the location we got for our new hotel.

5. Discussion section

From above reports, we could get an idea why the predicted one is pointed/clustered on the given spot. First most thing could be the center of attraction for the place.

K-Means have figured out the most common place for all the venues. This output was very adjacent to the core location. This proves the accurate spotting of our predicted algorithm.

Despite of the findings, there were some lack in data. Tips and ratings were missing for most of the venues. Also when I compared foursquare data with google map , i could see there were many hotels and venues found missing in foursquare.

6. Conclusion section

As a business person, one would be able to set up a hotel/restaurant on given spot.This will bring revenue automatically as we have located in very near to core one.We proved this with Kmeans.

Future Expectation:

As mentioned earlier ,most of data needs to be extracted from google maps. Even though we got somewhat accurate prediction. To be very confident on concluding our output, we may need more data to analyse.

Research based on hotel reviews and restaurant menus could be used for future purpose.

My Experience:

It was wonderful journey for me in IBM capstone and other courses. It can aid to layman people as well who don't know a pinch of Data science. Thanks to Coursera for keeping Skilful instructors with their awesome materials.