Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Hotel for Tourism in the most populous cities of India

By: Nakul Lakhotia

July 2020

1. Introduction

Historically, the concept of hospitality is about receiving guests in a spirit of goodwill—especially strangers from other lands. Hospitality implies warmth, respect and even protection; it builds understanding and appreciation among cultures. The Latin root hospes is formed from hostis, which means “stranger” or “enemy.” Related words are host, hospital, hostel and hotel.

Today, hospitality also refers to a segment of the service industry that includes hotels, restaurants, entertainment, sporting events, cruises and other tourism-related services. As such, the hospitality industry is important not only to societies—but to economies, customers and employees.

2. Business Problem

About 10.89 million foreign tourists visited India in 2019, an increase of 3.1% from the year before. Forex earnings from inbound tourists rose 8.2% to Rs 2.2 lakh crore in the past year. The growth was 9.6% in 2018 and 15% in 2017, according to figures from the ministry and Reserve Bank of India.

Since the domestic and international tourist inflow has been increasing every year, there is a bright future for the hospitality industry in India. Hotels are definitely one of the fastest-growing sectors in the tourism sector and it is truly justified as accommodation is the key part in the development of any country or region’s tourism. Tourism and Hotel Industry always go hand in hand and the presence of enough hotels also adds value and quite a lot of factors and punches it within the region’s economy. The Existence of a Hotel isn’t enough to single-handedly boost a region’s tourism but they also give out a symptom of health tourism.

3. Data

To solve the problem, we will need the following data:

- List of India neighbourhoods/cities which are highly populated according to Census 2011. This defines the scope of this project which is confined to the country of India.

- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.

- Venue data, particularly data related to Hotels. We will use this data to perform clustering on the neighbourhoods

4. Sources of Data and Extraction Methods

This page (https://en.wikipedia.org/wiki/List_of_cities_in_India_by_population) contains a list of most populous cities in India, with a total of 100 neighbourhoods. 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 neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

UA[a]	State/Territory	Population (2011)[4]
Mumbai	Maharashtra	18,394,912
Delhi	Delhi	16,349,831
Kolkata	West Bengal	14,112,536
Chennai	Tamil Nadu	8,696,010
Bangalore	Karnataka	8,520,435
Hyderabad	Telangana	7,749,334
Ahmedabad	Gujarat	6,361,084
Pune	Maharashtra	5,057,709
Surat	Gujarat	4,591,246
Jaipur	Rajasthan	3,073,350
Kanpur	Uttar Pradesh	2,920,496

Figure 1- Wikipedia table

5. Methodology

5.1. Extraction of data

We scrape the data from the Wikipedia page of the most populous Indian cities according to census 2011. Python libraries like BeautifulSoup and requests library are used to scrape the table from the Wikipedia page. The extracted data is converted to a dataframe for further processing.

2. Scrap data from Wikipedia page into a DataFrame

```
# get the response in the form of html
wikiurl="https://en.wikipedia.org/wiki/List_of_cities_in_India_by_population"
table_class="wikitable sortable jquery-tablesorter"
response=requests.get(wikiurl).text
response
```

...

```
# parse data from the html into a beautifulsoup object
soup = BeautifulSoup(response, 'html.parser')
indiatable=soup.find('table',{'class':"wikitable"})
```

```
df=pd.read_html(str(indiatable))
```

```
# convert list to dataframe
df=pd.DataFrame(df[0])
```

```
df.head()
```

Figure 2-Code for Data Scraping

5.2. Geographical coordinates of the cities & Mapping with Folium

We will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted.

Folium builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of the leaflet.js library. All cluster visualization are done with help of Folium which in turn generates a Leaflet map made using OpenStreetMap technology.

	Neighborhood	Population	Latitude	Longitude
0	Mumbai	18394912	18.94017	72.83483
1	Delhi	16349831	28.63410	77.21689
2	Kolkata	14112536	22.57053	88.37124
3	Chennai	8696010	13.08362	80.28252
4	Bangalore	8520435	12.96618	77.58690

Figure 3- Dataframe with coordinates of cities

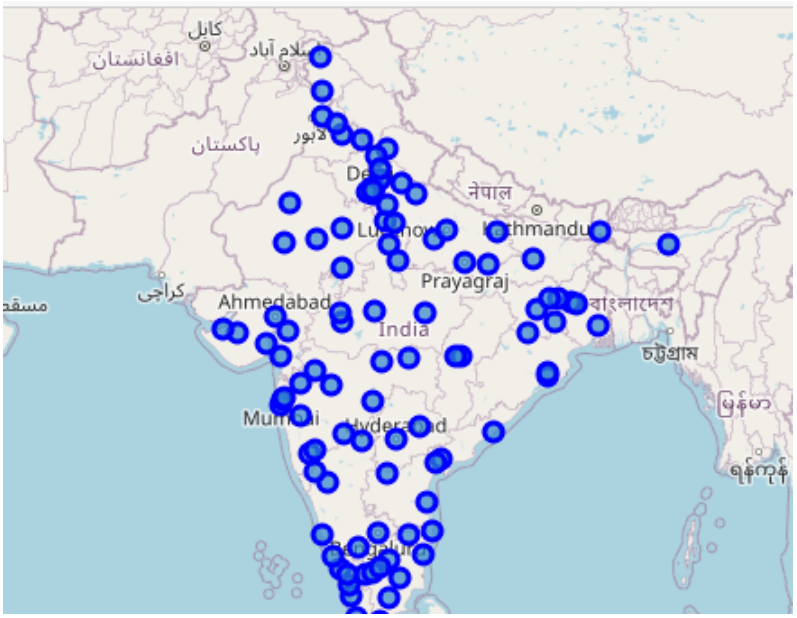

Figure 4-Folium mapping of India Cities

5.3. Foursquare API

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 5000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the "Hotel" data, we will filter the "Hotel" as venue category for the neighbourhoods.

5.4. Clustering

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project.

We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for "Hotel". The results will allow us to identify which neighbourhoods have higher concentration of hotels within a radius of 5km while which neighbourhoods have fewer number of hotels in the same radius. Based on the presence of hotels in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new hotels according to their tourism popularity.

6. Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for "Hotel":

- Cluster 0: Neighbourhoods with moderate number of hotels
- Cluster 1: Neighbourhoods with low number to no existence of hotels
- Cluster 2: Neighbourhoods with high concentration of hotels

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.

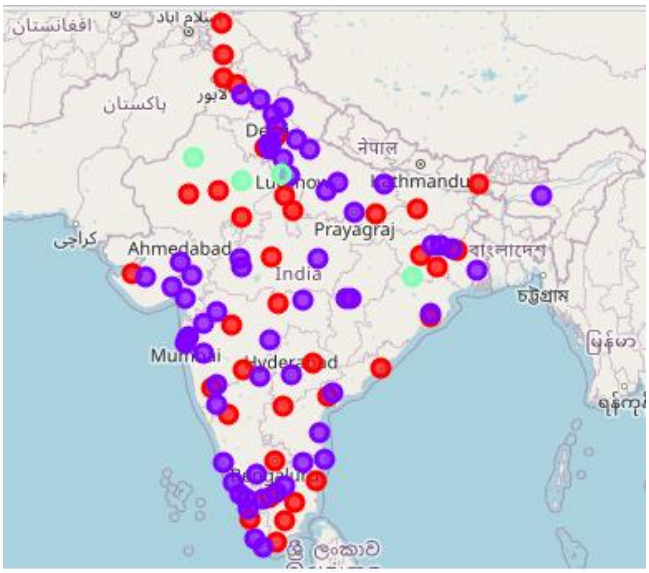

Figure 5- Cluster of Hotels in Indian cities

7. Discussion

As observations noted from the map in the Results section, very few hotels are concentrated in the north western area of the country, with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to no hotels in the neighbourhoods.

This represents a great opportunity and high potential areas to open new hotels as there is very little to no competition from existing hotels. Meanwhile, hotels in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of hotels. From another perspective, the results also show that the oversupply of hotels mostly happened in the north western part of the country, with the extreme north and extreme southern regions still have very few hotels. Therefore, this project recommends hotel builder and chains to capitalize on these findings to open new hotels in neighbourhoods in cluster 1 with little to no competition. Hotel chains with unique selling propositions to stand out from the competition can also open new hotels in neighbourhoods in cluster 0 with moderate competition. Lastly, existing hotel chains are advised to avoid neighbourhoods in cluster 2 which already have high concentration of hotels and suffering from intense competition.

8. Limitations and Future Scope

In this project, we only consider one factor i.e. frequency of occurrence of hotels, there are other factors such as population, tourism places, market places, and hospitals that could influence the location decision of a new hotel. However, in this research such data are not available in the foursquare api to the neighbourhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new hotel. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain better results.

9. Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. hotel chains and investors regarding the best locations to open a new hotel. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 are the most preferred locations to open a new hotel. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new hotel.