COL865 Project Report
Hospitality and Tourism

Pratik Nimbalkar (2020CS10607)
Garvit Dhawan (2020CS50425)
Nikhil Unavekar (2020CS10363)

Problem Statement

The hospitality industry, encompassing hotels and restaurants, is a crucial component of the tourism sector, as it directly influences travellers' experiences and satisfaction. It might be possible to gain valuable insights into the relationship between hospitality and tourism in India by analysing data pertaining to various hotels and restaurants across different states.

Collect relevant data from sites such as goibibo, Airbnb, MakeMyTrip, TripAdvisor, etc. Some desirable features can be (Hotel/Restaurant) Name, Location, Text Reviews, Ratings, Location, etc.

Analyse the text data and other complementary features to evaluate the quality of the hospitality industry across states and perform a comparative analysis. The Ministry of Tourism releases Indian tourism statistics every year, including state-wise statistics such as revenue, number of visits, etc. Interesting correlations between the hospitality industry in India and MoT's statistics based on the analysis or even observe some long-term trends in this relationship can be noted.

Objective

- India has a lot of opportunities in the **Tourism** and **Hospitality** sector. There is a need for a Comprehensive analysis of the industry which will provide insights for research and development **planning** purposes.
- We aim to gain valuable insights into the relationship between hospitality and tourism in India by analyzing data pertaining to various **hotels** and **restaurants** across different **states** and across different **years**.
- We have analysed the **review** data and other complementary features to evaluate the **quality** of the hospitality industry across states and perform a **comparative** analysis.

Targets Achieved

- Performed web **scraping** on various Hotel and restaurant websites like MakeMyTrip, Goibibo, Zomato and extracted data on ratings, customer reviews, location, amenities, etc
- Analyzed **government** and **makemytrip** Hotel Data to highlight the **hotspots** and quality of Hospitality industry in different cities of India at different times (pre and post **COVID**) of different ratings hotels.
- Analyzed Restaurant Data using Statistical Methods- Forming **heatmaps**, analysing city wise distribution, analysing top restaurant chains and cuisine types of India, etc.
- Performed **NLP** based Analysis of Hotel and Restaurant Reviews - Formed **WordCloud** based on hotel reviews, analyzed **sentiment** distribution., Segregated Reviews into different groups using **Topic Modelling**, Implemented **Aspect Based** Sentiment Analysis.
- Analyzed Tourism data (like **monumental visits**) from Government websites for different states and performed **correlation analysis** with that of **hotel** data.
- Build a Graphical User Interface (**GUI**) for easy, user-friendly data visualization for further needs adding functionalities like **BARD** API.

Web Scraping

```
            s4=t.text
        except:
            s4=""
    hotel_address=soup2.find("span",attrs={"itemprop":"streetAddress"})
    hotel_rating=soup2.find("div",attrs={"itemprop":"aggregateRating"})
    hotel_review=soup2.find("span",attrs={"class":"UserReviewstyles__UserReviewTextStyle-sc-1y0517z-4 UNLBr"})
    i=0
    s1=""
    s2=""
    for link in soup2.findAll("a",attrs={"itemprop":"item","class":"Breadcrumbstyles__BreadCrumbsTagItemLink-ily95m-5 kTxYic"}):
        x=link.findAll('span')
        if i==0:
            try:
                s1=(x[2].contents[0])
            except:
                s1=""
        else:
            try:
                s2=(x[2].contents[0])
            except:
                s2=""
        i+=1
    amneties=[]
    amneties=soup2.findAll("span",attrs={"class":"Amenitiesstyles__AmenityItemText-sc-10opy4a-8 izeHgc"})
    s3=""
    for a in amneties:
        s3=s3+a.text
        s3=s3+","
    print(s1,s2,s4)
    datas['HOTEL_name'].append(hotel_name.text if hotel_name else '')
    datas['HOTEL_address'].append(hotel_address.text if hotel_address else '')
    datas['PRICE_range'].append(s4)
    datas['HOTEL_rating'].append(hotel_rating.text[0:5] if hotel_rating else '')
    datas['HOTEL_review'].append(hotel_review.text if hotel_review else '')
    datas['COUNTRY_name'].append(s1)
    datas['CITY_name'].append(s2)
    datas['amneties'].append(s3)
    k=k+1
    if k>1000:
        break
if k>1000:
    break

table = pd.DataFrame(datas, columns=['HOTEL_name','CITY_name','COUNTRY_name','PRICE_range','HOTEL_rating','amneties','HOTEL_address','HOTEL_review'])
table.index = table.index + 1
table.to_csv(f'my_albums.csv', sep=',', encoding='utf-8', index=False)
print(table)
```

Implemented **Web Scraping** using BeautifulSoup library to obtain customer reviews, Hotel/restaurants price, amenities (if available), Hotel/restaurant ratings, etc.

Scrapper uses requests from **.xml** file and **BeautifulSoup** to generate excel.

Our Analysis can be broadly divided into 2 categories - Hotel Data Analysis and Restaurant Data Analysis

Hotel Data Analysis

Statistical Analysis of Government Data

We used **government-approved** (Classified) hotel data available on the National Database for Accommodation Units released By the Ministry of Tourism of India to perform analysis.

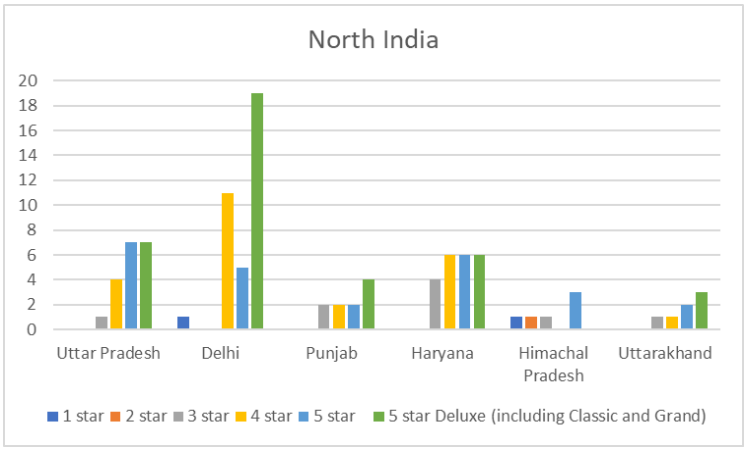
Note - Only cities which have the highest hotel number count from each region are included.

We have divided Indian cities into 5 parts - North India, South India, East and Central India, Maharashtra & Goa, Rajasthan & Gujarat.
We have included -

- Punjab, Haryana, Uttarakhand, Himachal Pradesh, Uttar Pradesh, Delhi in North India,
- Telangana, Tamil Nadu, Karnataka, Kerala, Andhra Pradesh in South India,
- Madhya Pradesh, Chhattisgarh, Odisha in Central India.
- Bihar, Assam, Arunachal Pradesh, Jharkhand, West Bengal, Jharkhand, Tripura in East India.
- Goa, Gujarat, Maharashtra, Rajasthan in West India.

Distribution of Hotels across Indian states

1. North India

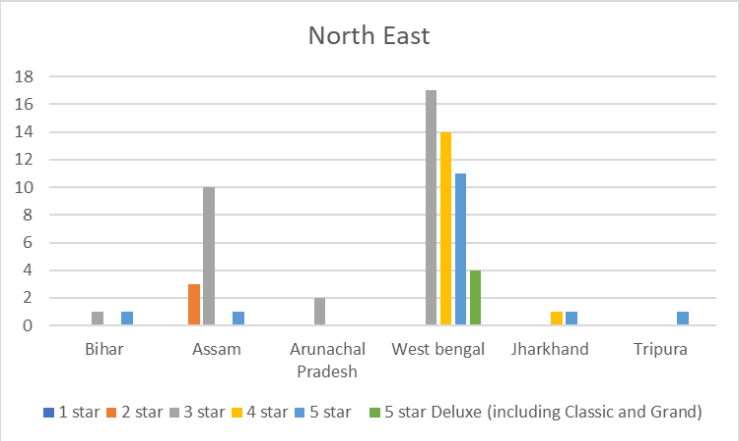


We observed from the plot that we have many **more 5-star** and **deluxe hotels** than other rated hotels in all the North Indian Cities. Especially in Delhi, Deluxe hotel numbers are high.

Uttarakhand and Himachal Pradesh, despite being tourist spots, have a **low** number of classical (approved) hotels due to scattered settlements and fewer work client guests.

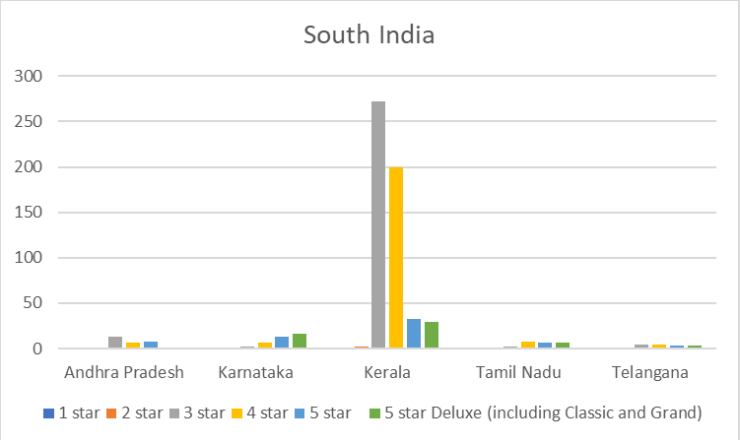
In Himachal Pradesh, hotels were majorly concentrated in **Shimla, Mandi**, and **Manali**, which are the most visited tourist cities. Similarly, Uttarakhand's hotels are also majorly located in cities like Rishikesh, Dehradun etc.

2. East



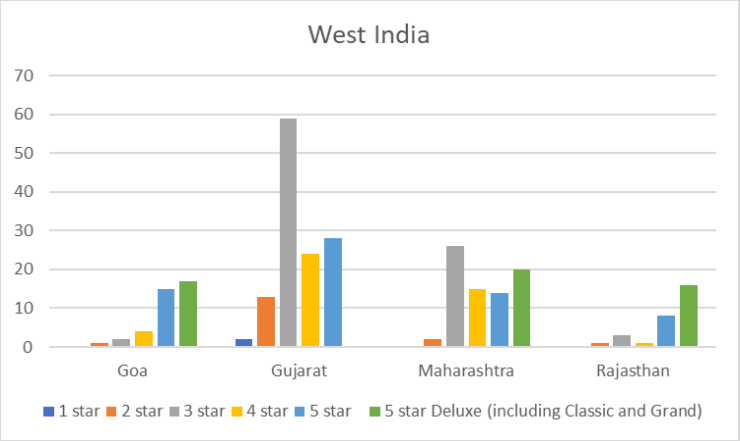
We can see that the east and northeast regions have very **few** hotels. Some states like Manipur and Nagaland had even less than 5 approved hotels. This shows the **extreme lack of tourism** in the northeastern states.
Comparatively, there are more hotels in West Bengal due to the presence of **Kolkata**, which is the commercial hub of the northeast and is famous for its culture.

3. South India



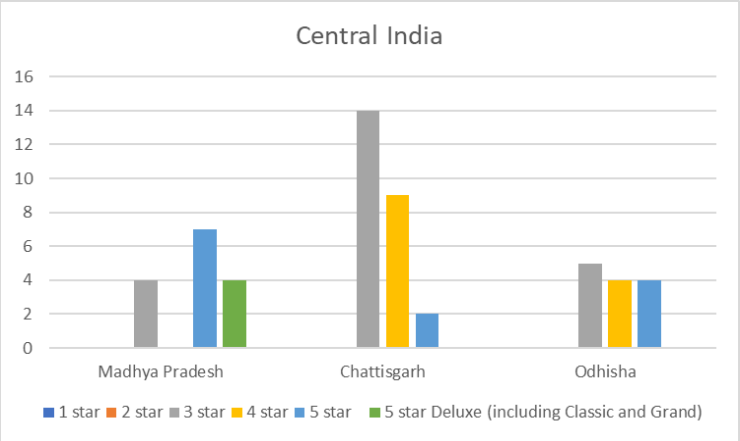
Kerala has the **highest** count of approved hotels in India by a huge margin. The distribution suggests more 3 and 4-star hotels but fewer 5 and deluxe hotels. The reason for this is the **government's policy** to link bar licenses with the classification. This has resulted in a parallel hotel lobby that thrives on **liquor** sales. These so-called ‘bar hotels’ in Kerala, which bag 4-star and 5-star licenses through corrupt means, set up bars of different levels in the property to cater to all customers' budgets and make money.
However, in the rest of South India, we do not see as many approved hotels.

4. West India



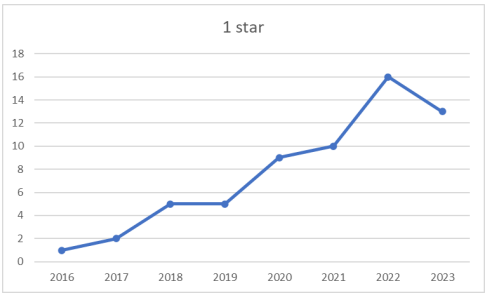
West India has **many** hotels, especially **3-star** ones, indicating a fair amount of **middle-class** customers.
Gujarat does not have deluxe hotels at all. There is a lack of fresh investments. Industry sources say **prohibition law, insufficient tourism** opportunities and a recently unfavourable **tax regime** have played spoilsport for high-end hotels in the state.

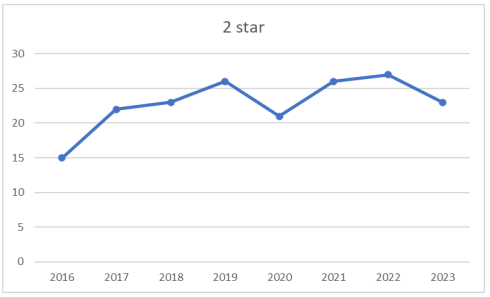
5. Central India

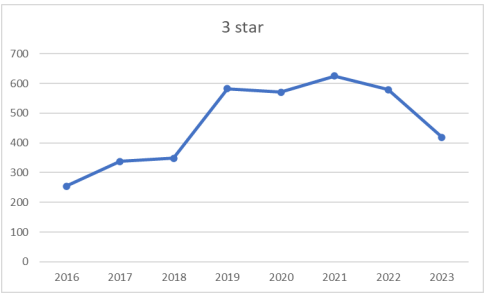


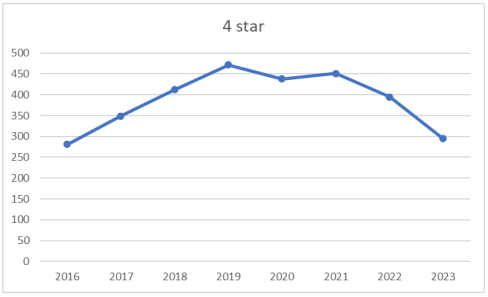
Central India has **few hotels** due to a lack of Tourism and Business hubs in the states.
Deluxe hotels are almost **negligible** in count.

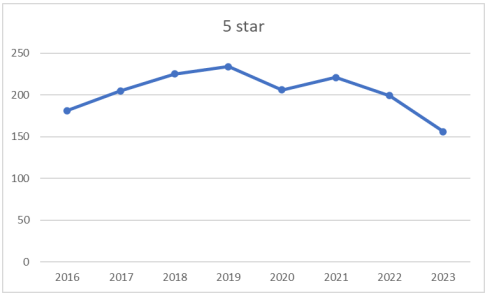
Time Series of Hotels Across India

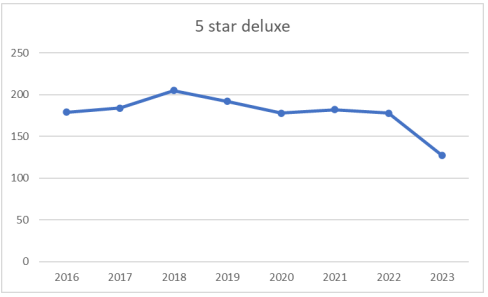










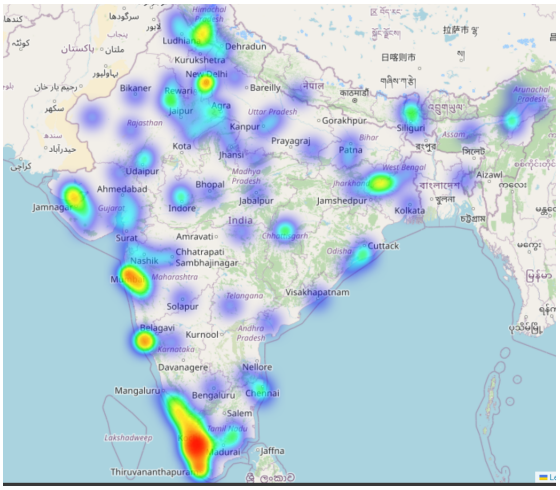


1-star hotels are **few**; still, their count is increasing continuously despite COVID. The number of 2-star hotels is also low, the counts are more or less consistent throughout the years. 3-start hotels have the **highest** count among all. There is an **increase** up to 2019, after which there is a **significant decrease**, attributed to **COVID**. They have good facilities with **affordable prices**, hence not much demand for 1 and 2-star hotels, therefore the low-end hotels are fewer.

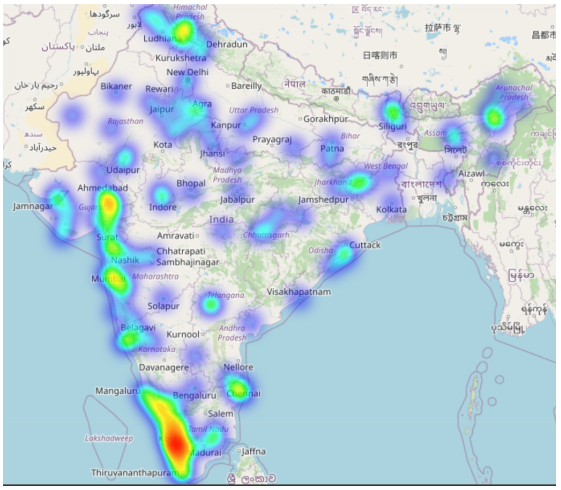
4-star hotels are also high in number. **Regular growth** is observed, with COVID-19 breaking the trend, after which there is a decrease. This is probably due to many hotels being **shut down** during the pandemic. 5-star hotels show a very **similar trend**, just the overall count being a little less. Deluxe hotels haven't had much growth since 2016, COVID **didn't affect** their count much since they are **well established**.

Spatial-Temporal Distribution

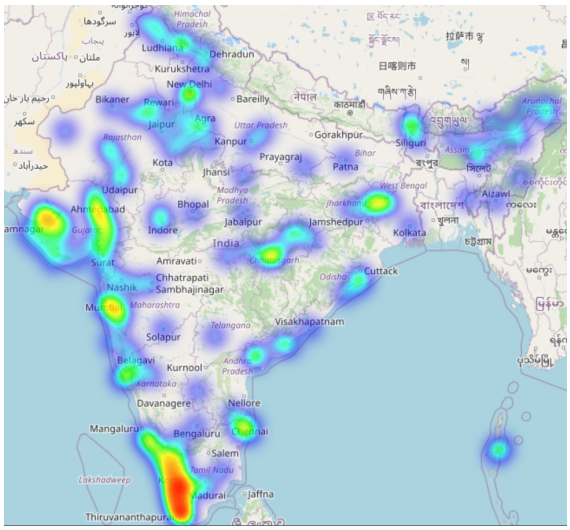
We generated heat maps of hotel concentrations throughout India from 2016 to 2023 using the **folium** library in Python. These were the results we obtained:



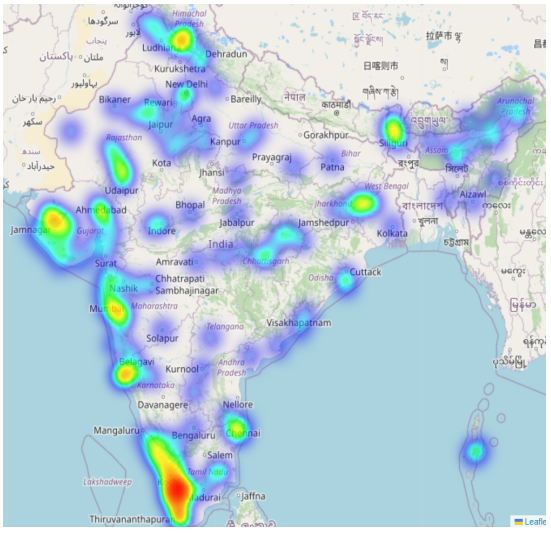
2016



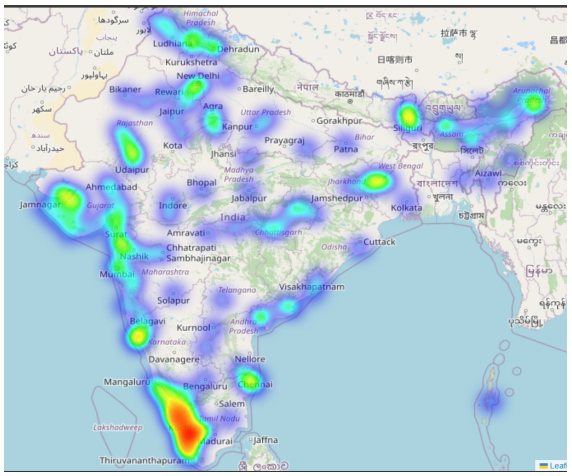
2017



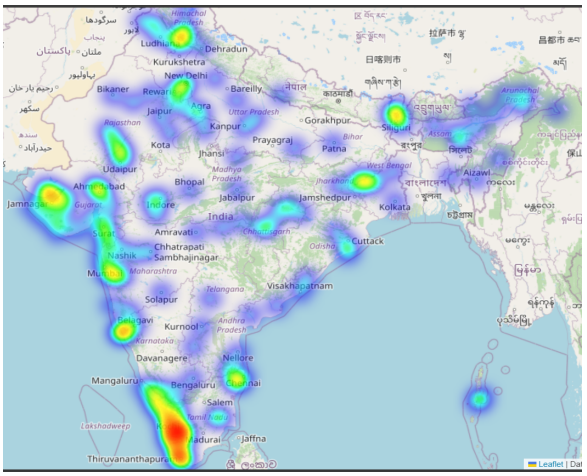
2018



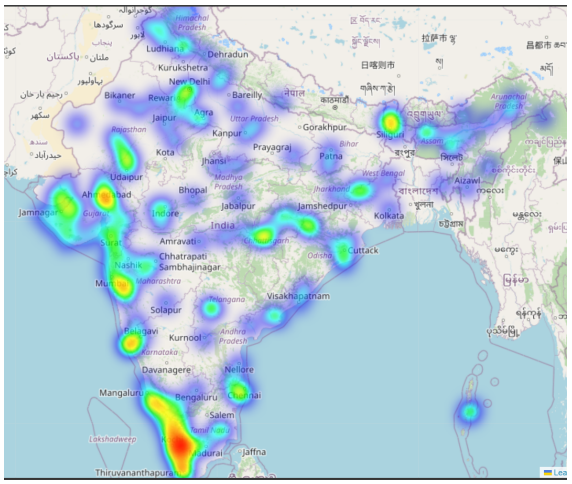
2019



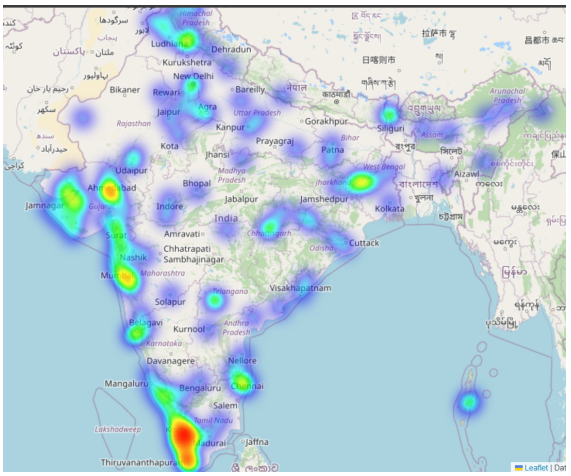
2020



2021



2022



2023

Hotspots and distributions of hotels all over India are highlighted through these hotspots.

Some Key Observations are:

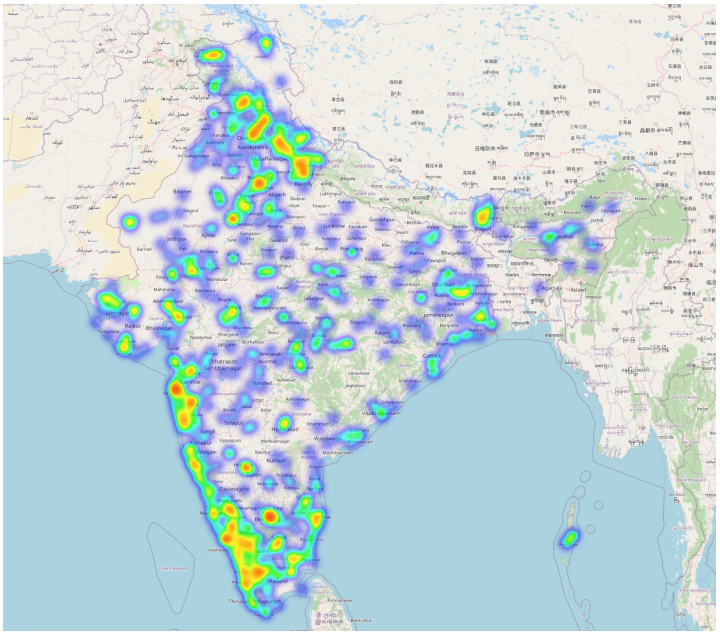
- In Himachal Pradesh and Punjab, the density has overall **decreased**
- In Chhattisgarh, the density is **constantly low** from 2018 onwards, after some minor fluctuations in 2016, 2017
- Kerala shows the **highest density**, with a decrease after 2019 due to COVID
- A small hotspot temporarily appeared in Jamnagar in 2018

The following table shows the variation in the density of some regions in the pre-COVID, during COVID, and post-COVID times:

Region	Pre-Covid	During Covid	Post-Covid
Ahmedabad	Increase	Decrease	Increase
Sikkim	Constant	Increase	Significant decrease
North East	Low count	Increase	Decrease
Jharkhand	Increase	Constant	Slight Decrease
Chennai	Constant (moderate)	Constant	Decrease
Mumbai	Increase	Decrease	Increase

Statistical Analysis of MakeMyTrip Data

Heat Map Observations



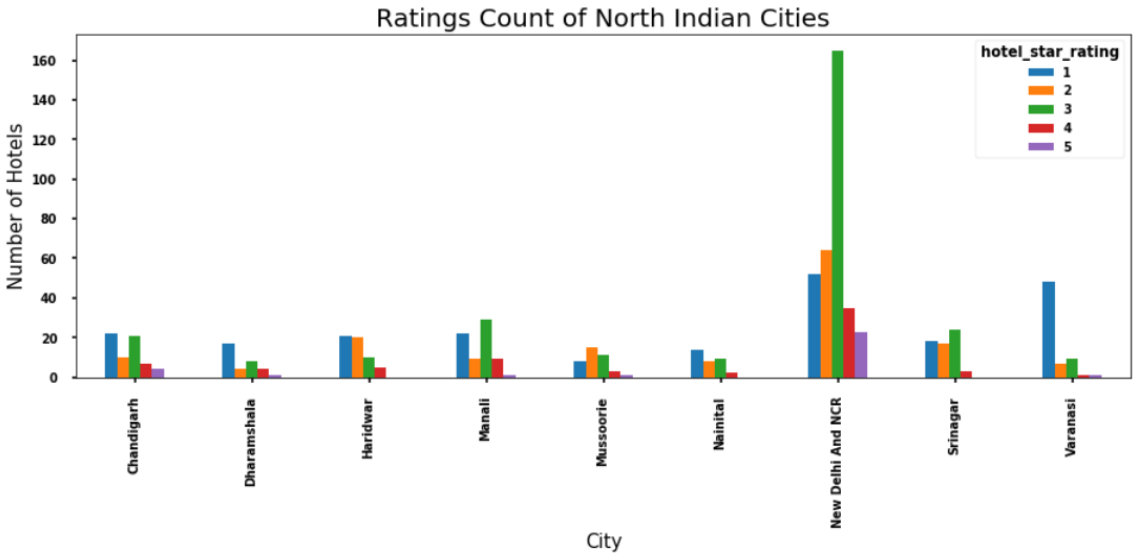
Govt data had very few (nearly **1500**) hotels as compared to the real scenario (eg MakeMyTrip had around **20,000** hotels) since government data only has **classical** (approved) hotels. Therefore,

- Many more cities are now covered, and **new hotspots** are visible.
- Though **Kerala** still has a lot of hotels(hotpot distribution nearly the same), we also have comparable hotspots in **northern, western** and **southern** India.

Distribution of Hotels across Indian Cities

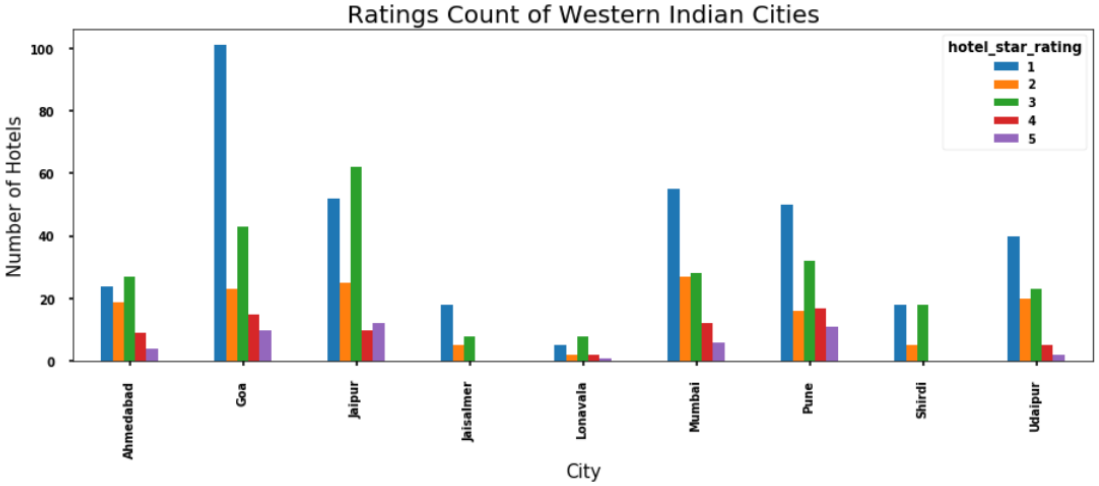
In each plot, we have only displayed the cities with the most hotels in their regions. We have divided Indian cities into 5 parts - North India, South India, East and Central India, Maharashtra & Goa, Rajasthan & Gujarat. The distribution of states is similar to that used in government data analysis.

1. North India



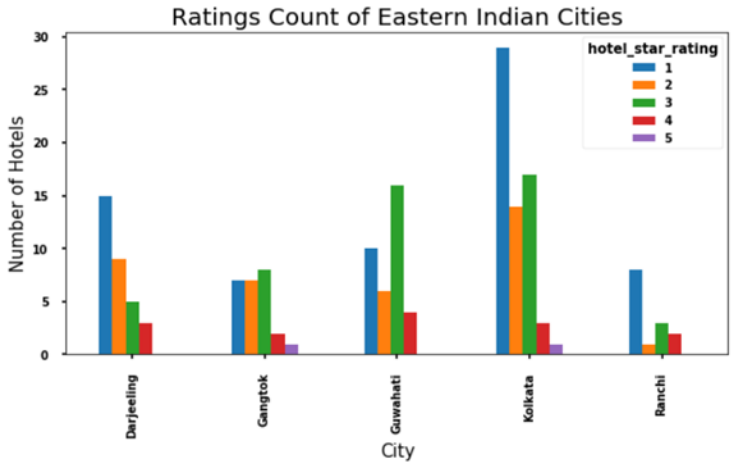
In North India, **Delhi** is the **dominant** city as also observed in govt data, but here, the most abundant hotels are the **3-star** hotels (compared to 5-star hotels being dominating in the govt data). The cities of Uttarakhand and Himachal Pradesh also have a good number of hotels due to a sufficient presence of tourism for hill stations in them.

2. West India



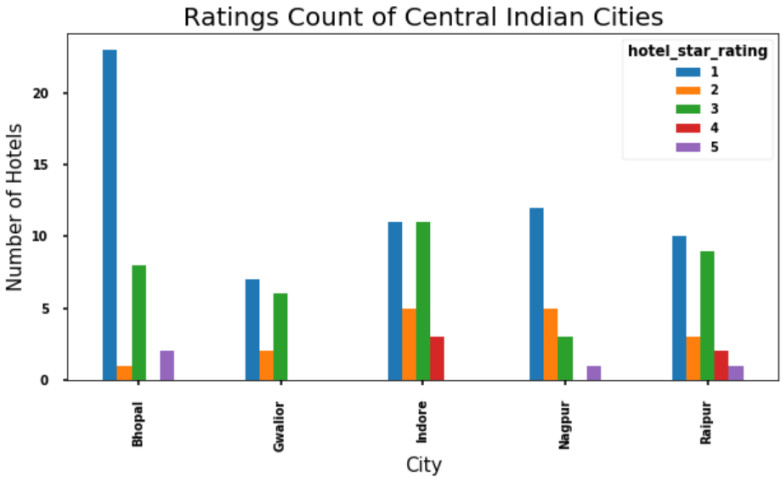
In West India, **Goa** has the **highest** number of hotels because it is a famous holiday spot. Western Maharashtra and Rajasthan cities also have a good count of hotels due to the presence of **business hubs** and various **tourist attractions**. Overall, the average number of hotels per city is good in West India.

3. East India



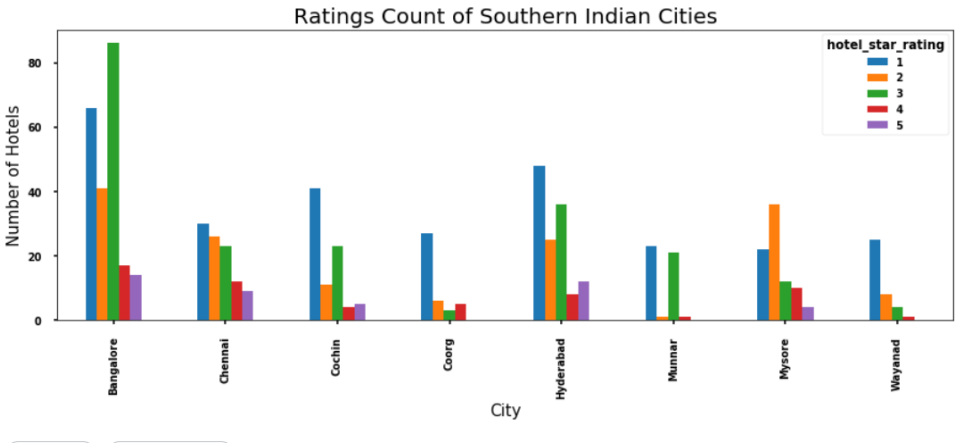
In East India, **Kolkata** has the **highest** number of hotels (the same as observed in the government data), while the Northeastern cities have a **low** overall count.

4. Central India



Central Indian cities have a **low** average count, with even fewer high-end hotels (similar to what was observed in the govt. data).

5. South India



Here Kerala does **not** have the highest count as in govt data. **Bangalore** and **Hyderabad** have the **highest** counts instead which tells that the distribution is more **even** in South India. This may be because they are the most dense **business hubs** of the country.

Conclusions of city wise MakeMyTrip Data Analysis

- Hospitality industry is concentrated in **North, West** and **South** India
- **3-star** hotels have the **highest** count among all the ratings due to affordable prices and decent facilities
- There is growth in the industry, **hindered** by **COVID**, with the total count decreasing from 2019
- The distribution of hotspots over time remains quite consistent with only **minor variations**
- Govt. data is **biased** towards **Kerala** in South India
- The number of **approved** hotels in India is significantly **lower** than the actual number of hotels.
- There is a significant disparity in the count of 1-star hotels on MMT compared to the govt. records.
- There are **limited 5-star** hotels in **Central** India, primarily due to a lack of significant tourism, limited hospitality infrastructure, and the economic profile of the region
- **North East** has a significantly **lower** count except Kolkata due to geographic challenges, while **Kolkata** is a trade, educational and cultural hub.

NLP-based Analysis of Chennai Hotel Reviews

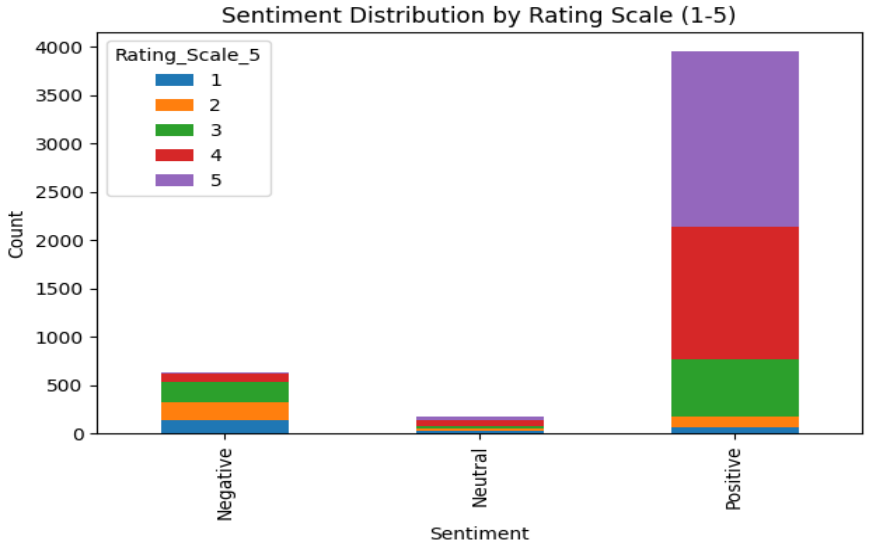
We have analysed the reviews of **Chennai** hotels. We chose Chennai hotels since there are a lot of hotels in Chennai and this was the only good data available.
We have made several **EDA** (Exploratory Data Analysis) Analysis like **WordCloud, Sentiment Based Distribution, Aspect Based Analysis and Topic Modelling**.

The Bangalore restaurant reviews data which we used had the following headers- Hotel_name, Review_Title, Review_Text and Rating_Percentage.
The **Review_text** column has reviews of the restaurants with each entry having several reviews for a restaurant, combined together. There are more than **5,000** hotel reviews in the data.

1. WordCloud

- Used **Wordcloud** and **Textblob** libraries in python to create the wordcloud.
- WordCloud suggests that **room, service, location, breakfast, facilities** are the aspects which are frequently mentioned in the reviews.
- Aspects like **swimming pool, hot water, TV, ambience,** and **spacious** are not frequently used by customers in the reviews.
- This suggests that customers are more interested in the room quality, service and location of the restaurant.

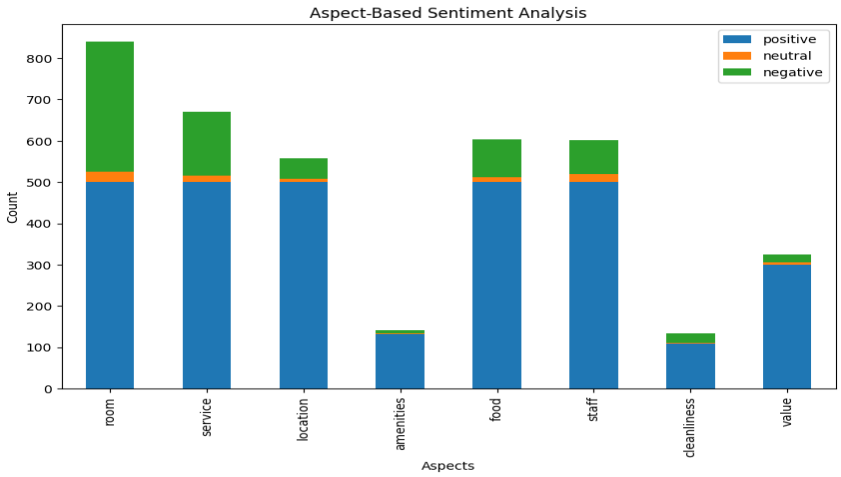
2. Sentiment Distribution



- Using the **Textblob** and **matplotlib** python libraries, we plot the reviews dividing them into 3 - positive reviews, negative reviews and neutral reviews.
- Textblob library segregates the reviews depending on the words (which get converted into **vectors**) of the review, whether the review is neutral or positive or negative.
- From the graph, we see that **most** of the reviews have a **positive** sentiment. The y axis shows how many restaurant reviews have a particular sentiment.
- Also, we see that Most of the 5, 4 star reviews have **positive** sentiment while 2, 3 star hotels reviews have **all kinds** of sentiments.
- As expected **Neutral** count is the least since people either write a positive review (eg - “The food was amazing”) of the restaurant or criticise the same (eg - “The food quality was extremely poor”), but hardly anyone writes a neutral review like “The food was fine”.

3. Aspect Based Sentiment Analysis

- The graph shows the Sentiment distribution of different aspects.
- Aspects are attributes or features of the reviews that we are interested in, such as **"room," "service," "location,"** etc.
- We use an Aspect Sentiments **Dictionary** which stores **sentiment** counts for each aspect (positive, negative, neutral) in a dictionary. Our **analyze_sentiment** function assesses sentiment in text and categorises it using the **Textblob** library of python. The code loops through reviews, checks if an aspect is mentioned, and counts sentiments

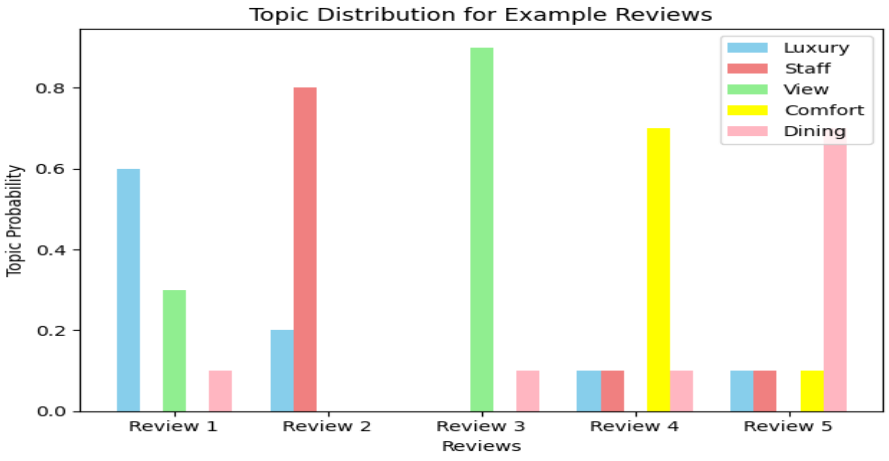


.

- Attached graph shows that Indian people mostly focus on **room, service, and food aspects**, while cleanliness, amenities don't matter much.
- All the aspects have **more positive** reviews than negative or neutral reviews since overall, the positive reviews count is highest.
- Moreover, aspects like **Staff, Ambience, Kitchen** have a very high percentage of **positive** sentiment. Similar is also true for restaurant review Aspects as we will see in the next part.

4. Topic Modelling

For Topic Modelling, we use the **sklearn** python library which is a ML library.
We are directly using the **LDA** (Linear Discriminant Analysis) function of sklearn to extract key words and hence topics.
In LDA, we convert each word into a **vector** and put correlated words together which are then grouped to form one **topic** like "quantity".
Whenever a new review is analysed, we correlate it with already existing topics and find the probability of it belonging to different topics.



- Using Topic Modelling, we found that in our reviews, **‘Luxury, Staff, View, Comfort, Dining’** were the most prominent topics.
  The topics mentioned here are the most important ones.

The 5 reviews used by us to plot this graph are -

Review 1: "The hotel was really fancy! The view from the room was amazing, especially during the sunset. The food at the hotel's dining was superb."
Review 2: "Our stay at this hotel was fantastic! The staff was very cooperative and had a nice attitude."
Review 3: "I had a pleasant experience at this hotel. The view from our room was stunning. The dining options were great."
Review 4: "The hotel was comfortable and luxurious and the staff was very polite. The view from our room was spectacular, especially in the evening. The room was comfortable, and the dining options offered a range of delicious dishes."
Review 5: "We had a great time at the hotel. The staff was helpful and the view from our room was outstanding. The dining options were delightful with a variety of tasty dishes."

For example, Review 3 ("I had a pleasant experience at this hotel. The view from our room was stunning. The dining options were great.") is majorly focused on the view of the hotel and partly on the Dining of the hotel. Thus we have probability of “View” topic = 0.9 and that of “Dining” topic = 0.1.

5. Point-Biserial Correlation between Discrete Rating and Sentiment

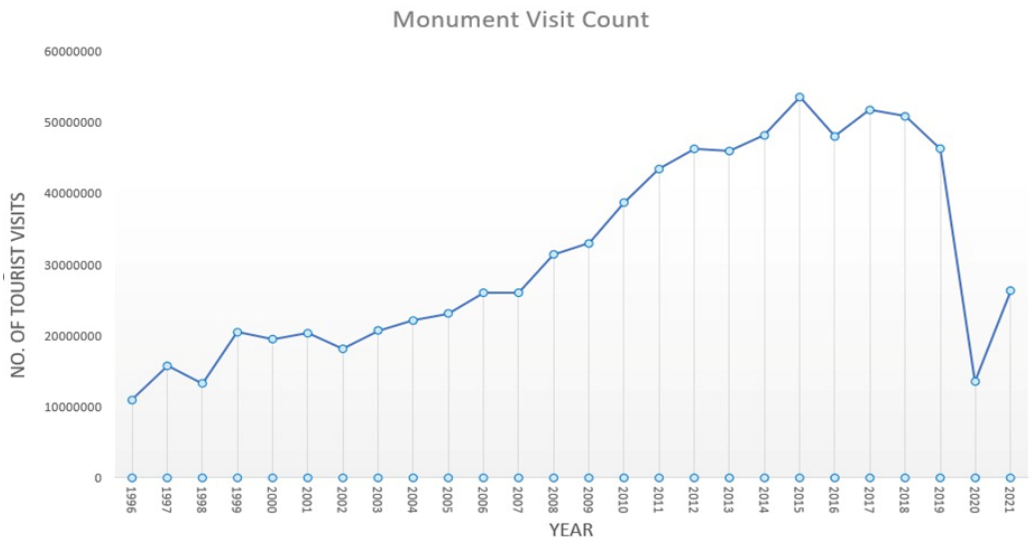
- The Point-Biserial Correlation between Discrete Rating and Sentiment was found be to 0.5149004828069564.
- Point-Biserial Correlation is used for comparing **dichotomous** data. We assumed positive reviews as 1 and both negative and neutral reviews as 0. We used the scipy python library to find this correlation. Using this library, we can find a direct formula to link dichotomous and continuous data.
- Correlation = **0.51** suggests that the relation between Ratings as per given data and the sentiment distribution which we calculated is **positive**, i.e. both are **directly proportional** to each other to a certain extent.
- If the value was negative, we would have said that both are indirectly proportional and if the value was 1 (which is highest) we would have said that both are strongly proportional to each other

Conclusions of Hotel Data Analysis

- Hospitality industry concentrated in **North, West** and **South** India
- **3 star** hotels have the **highest** count among all the ratings due to affordable prices and decent facilities
- There is growth in industry, **hindered** by **COVID**, total count has decreased from 2019
- The distribution of hotspots over time remains quite consistent with only **minor variations**
- Govt. data is **biased** towards **Kerala** in South India
- The number of **approved** hotels in India is significantly **lower** compared to the actual number of hotels.
- Many **more data points**, more evenly distributed than Govt data
- There is a significant disparity in the count of 1-star hotels on MMT compared to the government. records.
- There are **limited 5-star** hotels in **Central** India, primarily due to a lack of significant tourism, limited hospitality infrastructure, and the economic profile of the region
- **North East** have significantly **low** count except Kolkata due to geographic challenges, while **Kolkata** is a trade and educational and cultural hub
- Ratings and review sentiments have correlation of **0.51** indicating a positive relation but not strong
- **EDA** and aspect analysis show similar results in terms of topics of concern
- **Topic modelling** predicts that most reviews focus towards a single dominant topic

Tourism Data Analysis

Monumental Visits Data Analysis



To make an analysis of the Tourism sector data of India, we found the Monumental Visits Data to be the most suitable. We have taken this Data from the **Indian Tourism Statistics Report** released by the **Ministry of tourism**.
Above graph shows that Monumental visits count **increased consistently** till Covid years. A sudden drop in the count is observed after **2019** as a result of **Covid**.

We can see that just like the hotel industry, Covid has also affected the tourism industry as can be seen by the monumental visits number over the years.

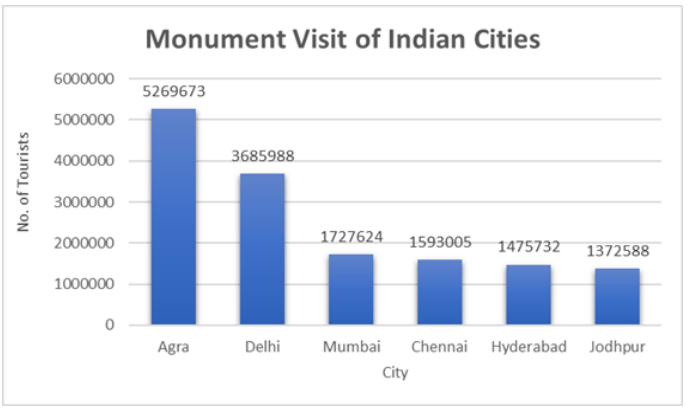
Also, the number of domestic and international **tourists**, both were **lowered** after 2019.

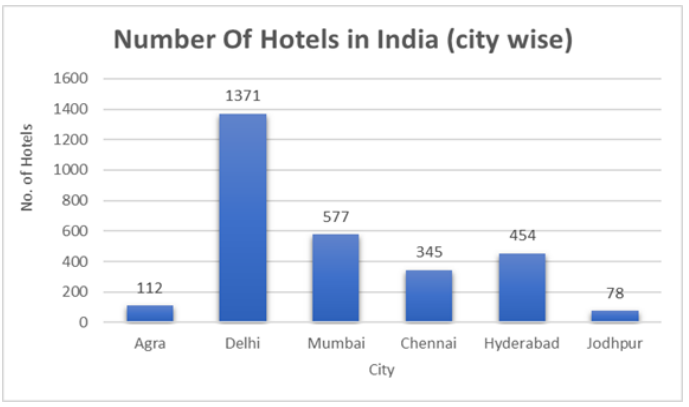
Correlation of City wise Monumental Visits & Hotel Data

Now, we try to see a **correlation** between the **hotel** industry and the **tourism** industry. For this, we have used scrapped Hotel Data from **MakeMyTrip** which is available publicly on Kaggle. For the Monumental Visits data, we used the same data as mentioned before - the **Indian Tourism Statistics Report** released by the Ministry of tourism.

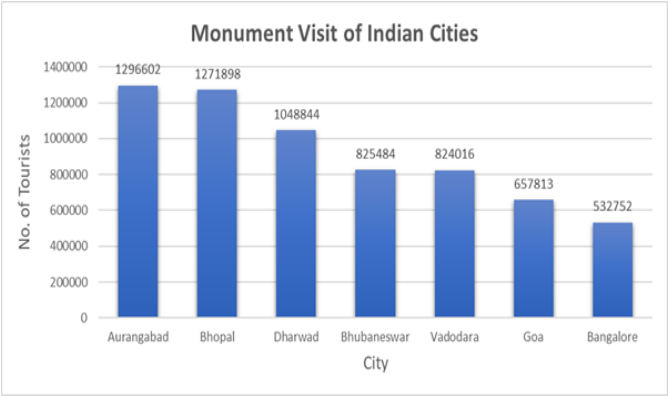
Note - We have only included Cities which have highest monumental visits in India as per MoT.

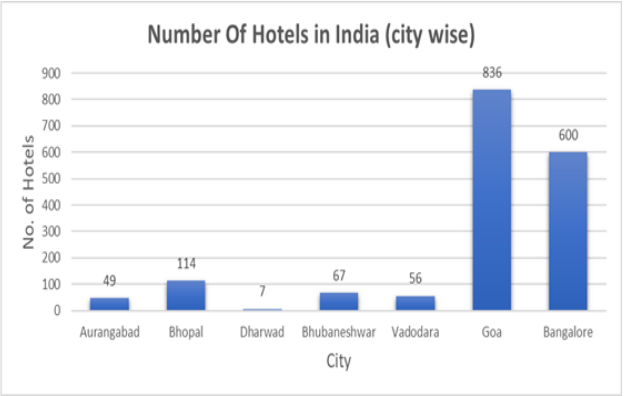
Starting from the cities having the highest monumental visits, the comparison analysis is as follows-





- **Agra** has the **highest** tourist attraction (>**50** Lakh annual monumental visits) but the no. of hotels are very less (only **112**).
- Similarly **Jodhpur** also has a potential market for the Hotel Industry. The tourist number is >**13** lakhs but the hotel number is just **78**.



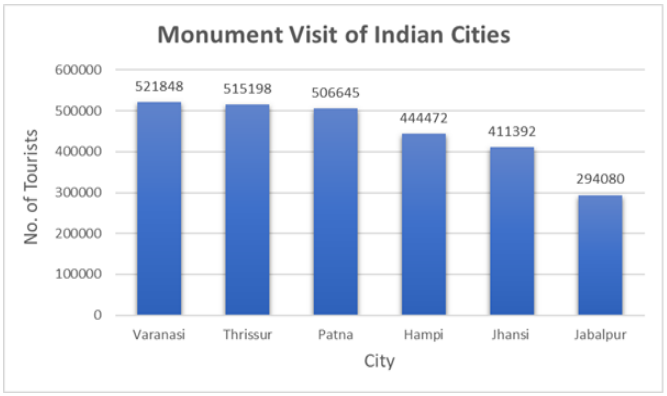


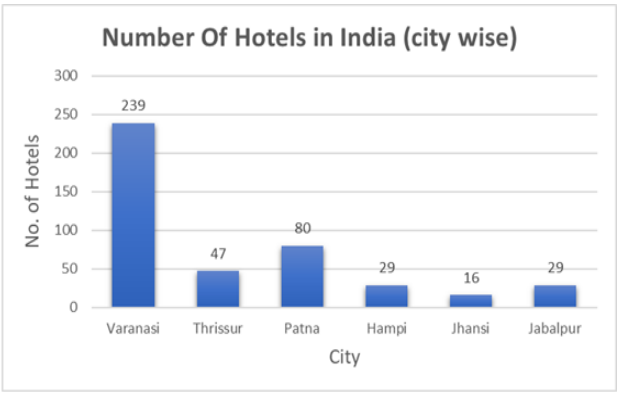
These are the next cities which have the highest monumental tourists attraction.

- **Bangalore** and **Goa** have **low** monumental visits but have a huge number of hotels. The reason being that people generally come to Bangalore for **business** and **work** related purposes and to Goa due to its **beaches**, which results in a demand for Hotels.
- **Dharwad** has a monumental visit number around **10 lakhs** but the hotel count is **extremely low** giving us a great **potential** for expansion of Hotel industry in the city.

- Also, **Aurangabad, Bhubaneshwar, Vadodara** have a good number of tourists due to monumental visits but hotel count is low.

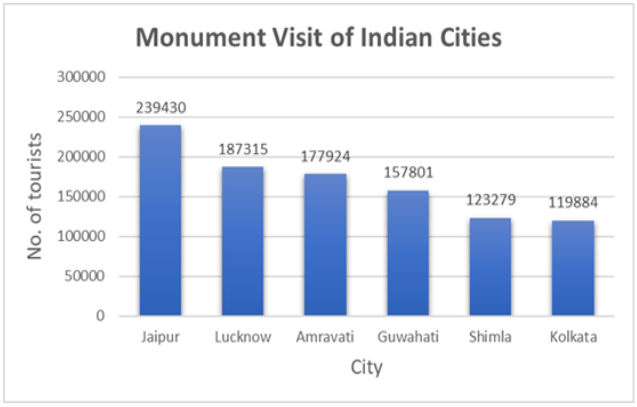
Then we have the following cities-

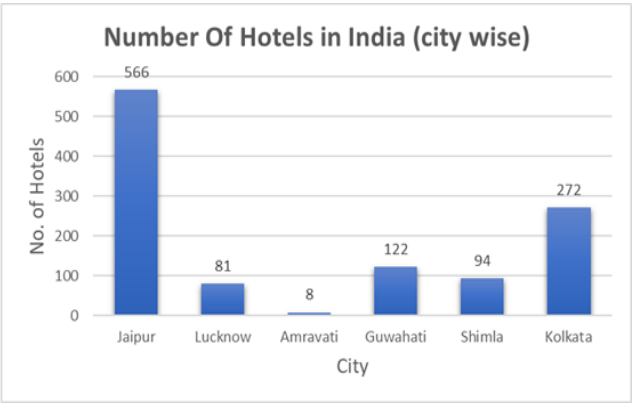




- All cities except Varanasi are in need of more hotels especially **Thrissur** and **Jhansi**.
- Places like **Hampi**, Jhansi and Thrissur have a rich monumental and **historic heritage**, thus providing an opportunity to expand the hotel industry.

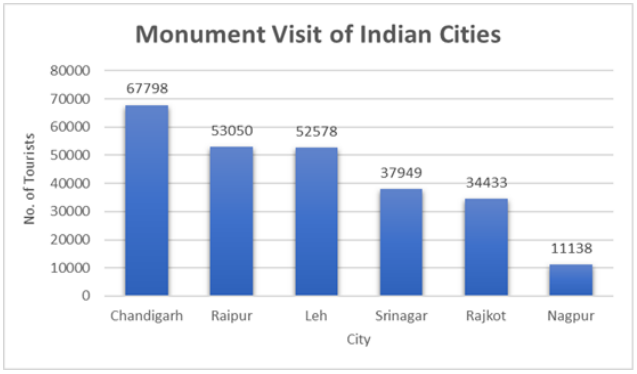
Next,

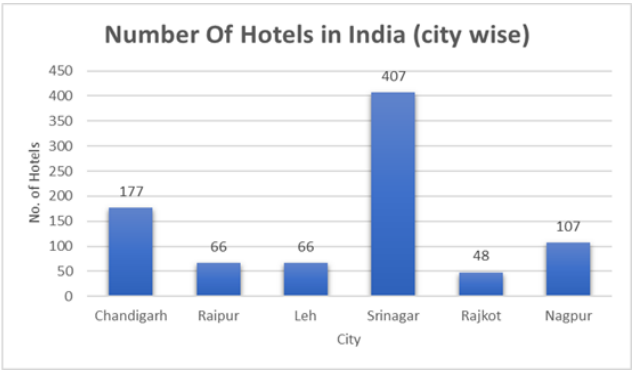




- **Amravati's** hotel count (**8**) is very less as compared to its tourist visit number (**1.8 lakhs**).
- **Lucknow** also has a potential of having more hotels due to its high monumental visit count (**1.9 lakh**) as well as its **business** and **work** related visitors.

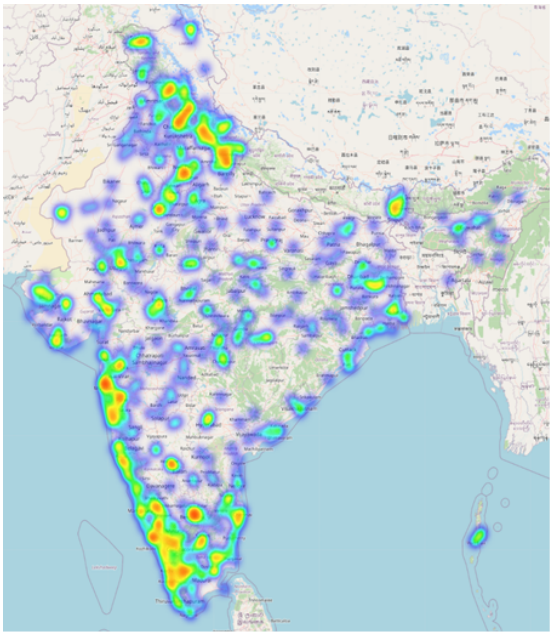
At the end, we have -



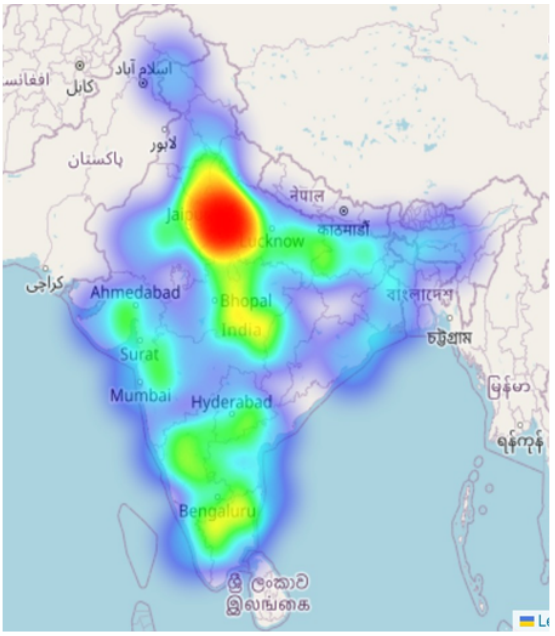


- We see that surprisingly there are more hotels in **Srinagar** (**407**) than **Chandigarh** (**177**) despite the fact that the monumental visit count is higher in Chandigarh (**68k**) than in Srinagar (**38k**).
- Also, we observe a need for more hotels in **Raipur, Leh and Rajkot**.

Hotel and Monumental Visits Distribution in India using Heatmap



Heatmap for MakeMyTrip Hotels Data



Heatmap for Monumental Visits of Tourists Data

We used the **folium** python library to generate heatmaps for monumental visit city wise data (released by MoT) and MakeMyTrip city wise data of India.

Observations-

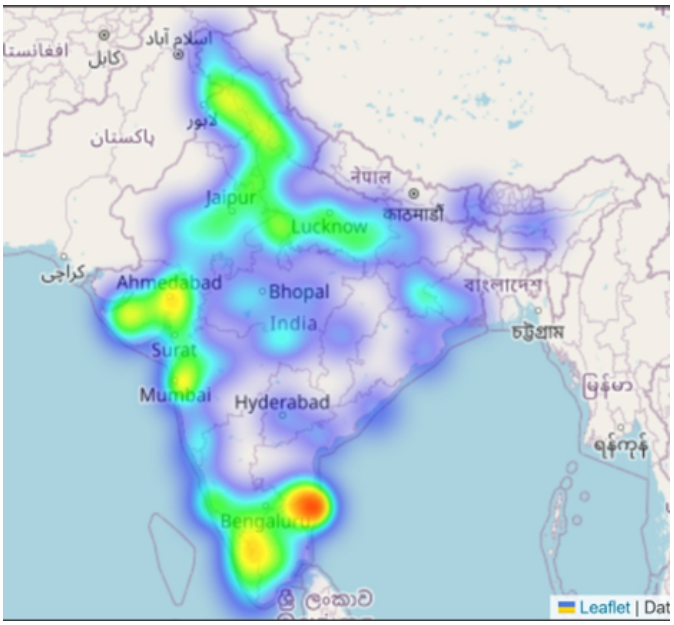
- Monumental Visits Heatmap is highly concentrated in **Delhi** and **Agra** regions due to monuments like **Taj Mahal, Agra Fort, Qutub Minar, Red Fort**, etc. which attract an extremely high number of both domestic and international tourists.
- Apart from North India regions, the Hotel industry is also dense in **South** and **West** Indian regions of **Kerala** and **Maharashtra**.
- Thus , we see a need for the Hospitality Industry to invest in **Monumental rich** cities like Agra.
- We also see that as the tourist industry fuels the hotel industry, vice-versa is also true. Many times people visit some monuments or cities because there is a **good availability** of hotels there. We saw some examples in the data which had a lower count of total monuments but more monumental visits due to hotel facilities in that city.

Restaurant Data Analysis

We used **Zomato India** Restaurants Data which had around **1 lakh** hotels to analyse the Restaurant industry of India.
This data has the following headers - res_id, name_of_establishment, url_address, city, city_id, locality, latitude, longitude, zipcode, country_id, locality_verbose, cuisines, timings, average_cost_for_two, price_range, currency, highlights, aggregate_rating, rating_text, votes, photo_count, opentable_support, delivery, takeaway, Ratings

Statistical Analysis

City wise Distribution using Heatmap

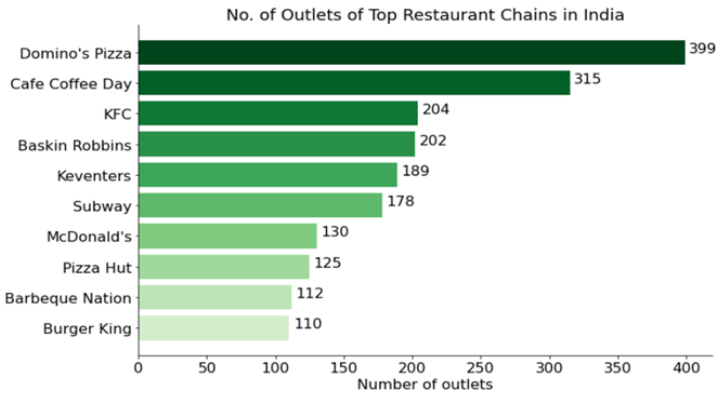


As expected, **metro** cities have more restaurants than others with the domination of **South India**.
Heatmap is densely populated in the Southern part including **Chennai, Bangalore**. Gujarat, Maharashtra, Punjab and Delhi regions are also concentrated.

Top Restaurant Chains

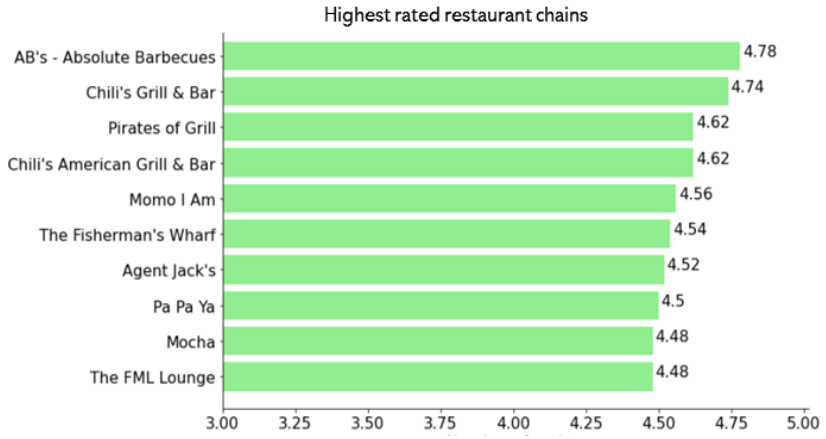
Total Restaurants after removing duplicates = 55568
Total Restaurants that are part of some chain = 19358
Percentage of Restaurants that are part of a chain = **35.0 %**

1. Based on Number of Outlets in the Chains-



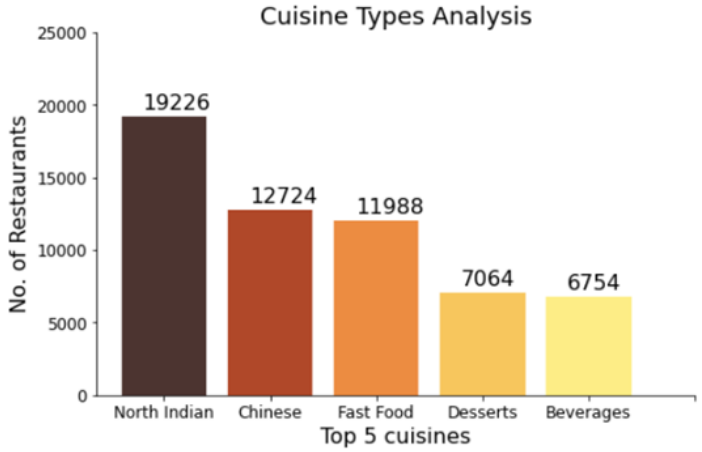
We see that the restaurant industry of India Majorly dominated by **big fast food chains** like **Dominos**, CCD, KFC which have the highest number of outlets in India.

2. Based on Ratings of the Chains



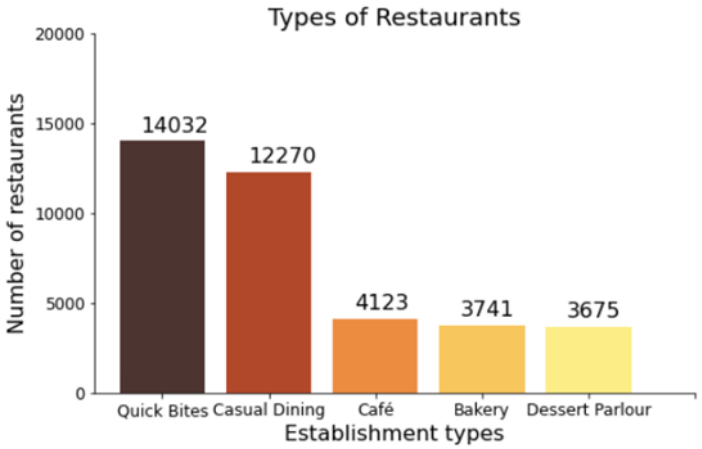
- Interestingly, **no fast food chain** is present in this graph. Most **bars** are present.
- To maintain a high rating, restaurants need to provide **excellent service** which becomes impossible with booming fast food restaurants in every street due to the **less-hygienic** food of fast food chains.

Cuisines Types Analysis



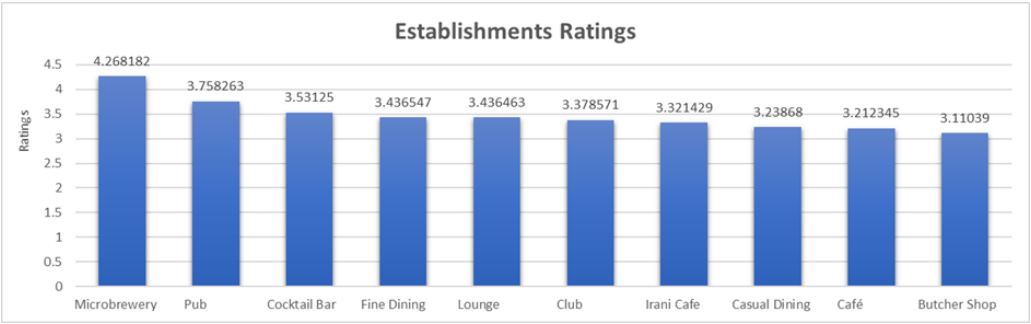
- We see that **North Indian** food has the highest outlet number.
- Surprisingly, South Indian cuisine is not in the top 5.
- Moreover, **Chinese** food comes second in the list, even ahead of fast food, desserts and South Indian food.

Establishments Types Analysis



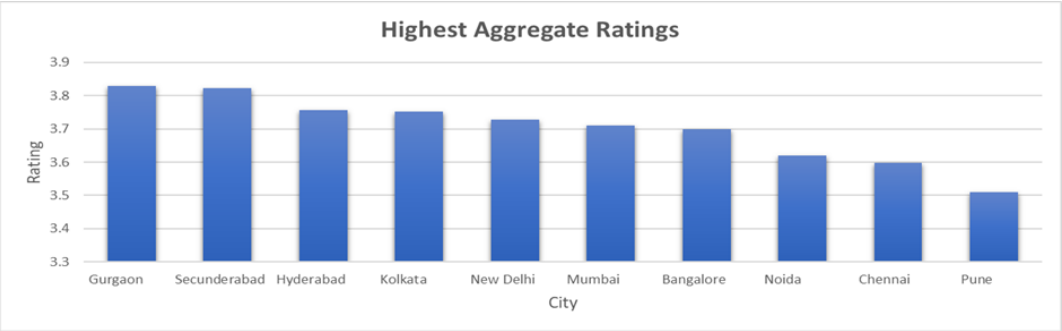
- Top 3 (Quick Bites, Casual Dining, Cafe) represents more casual and quick service restaurants, then at 4 and 5 we have dessert based shops.
- We have a **higher** number of **casual and quick** restaurant types in India as compared to others.

Establishments Ratings



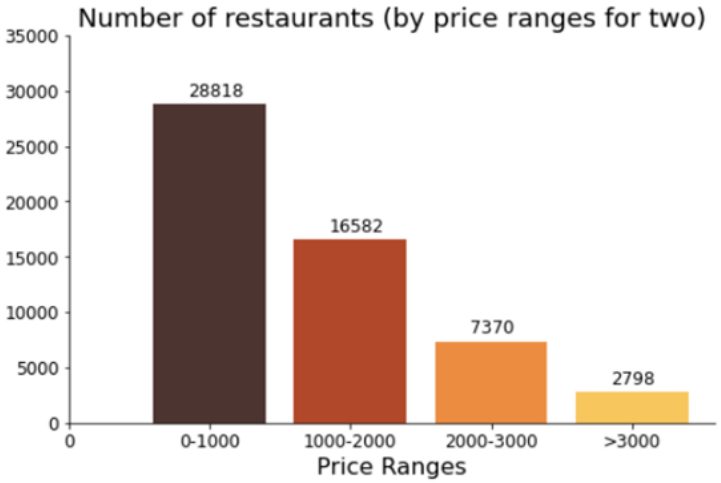
- **Breweries, Pubs, Bars** have the best aggregate ratings in India (all **>3.5** star). It can be concluded that establishments with **alcohol** availability have the highest average ratings.
- Fast food establishments average ratings are generally not very high since to maintain a high rating, restaurants need to provide excellent service which becomes impossible with booming fast food restaurants in every street due to the less-hygienic food of fast food chains

City wise Aggregate Ratings



- **Gurgaon** has the highest rated restaurants (**3.82** star rating) followed by Secunderabad and **Hyderabad** which have one of the best **Biryani** and Haleem Restaurants of India.
- Here also, we can see that all the cities present are **metro cities**.
- Also, **Hyderabad** has the highest **critics**.

Price Range count



- The average cost for two was between **Rs.350 to Rs. 900**. Hence, we can say that the Majority of restaurants are budget friendly.
- Number of restaurants **decreases** with increase in price range which is obvious since we have a larger number of people who prefer budget friendly restaurants with price <1000 Rs for two as compared to people who prefer highly priced restaurants.

Relationship Between Price and Ratings

- Relation between Average price for two and Rating = **0.25**. This suggests that the relationship between the two is **positive** which means that as the average price for two increases, the ratings of the restaurant increases.
- The higher the price a restaurant charges, more services they provide and hence more chances of getting good ratings from their customers.

Distribution of Restaurants across Indian Cities

We have used the same data for this - All **India Zomato** Restaurants Data.
We rounded off the Ratings to the nearest integer.
Also, 0 star restaurants are the ones which have not received any rating till now.

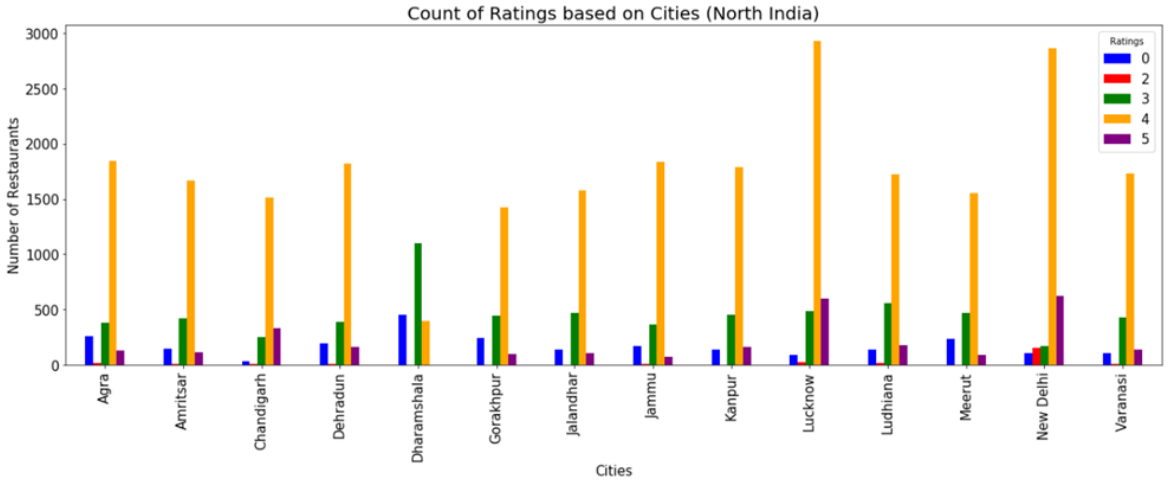
Note - Only cities which have the highest restaurant number count from each region (which have **>2000** restaurants in a city) are included.

We observe that for almost all the cities of India, the number of **4 star** restaurants were the **highest** and **2 stars** were **lowest**.

We have divided Indian cities into 5 parts - North India, South India, East and Central India, Maharashtra & Goa, Rajasthan & Gujarat.
We have included -

- Punjab, Haryana, Jammu & Kashmir, Uttarakhand, Himachal Pradesh, Uttar Pradesh, Delhi in North India,
- Telangana, Tamil Nadu, Karnataka, Kerala, Andhra Pradesh in South India,
- Madhya Pradesh, Chhattisgarh, Odisha, North East states, Bihar and West Bengal in East and Central India.

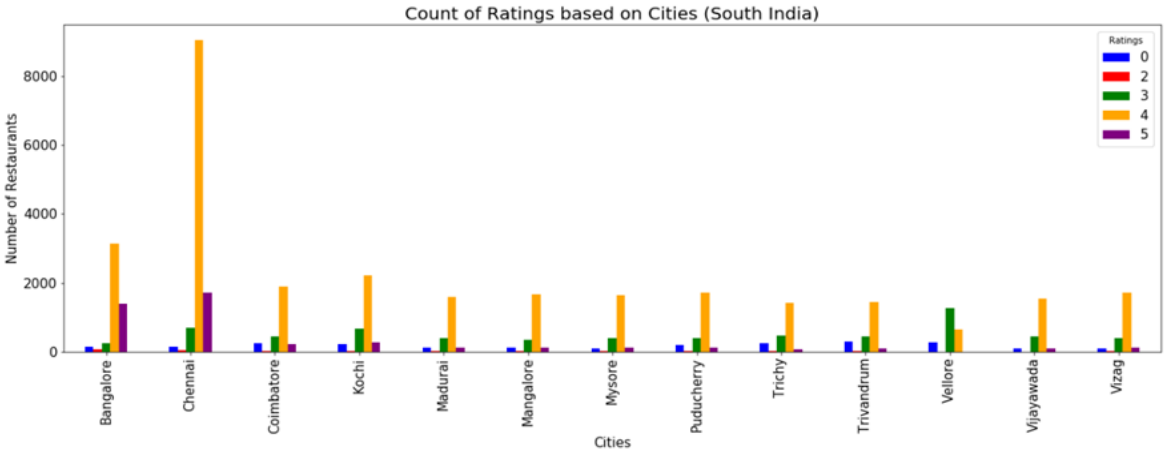
1. North India



Observations-

- Overall 4 star restaurants count is the highest (**Dharmashala** being an exception) and 2 star count is very less.
- Dharamshala features a significant number of non-rated restaurants.
- In **UP**, **Lucknow's** count is highest followed by Agra, Kanpur and Varanasi.
- In **Punjab** Chandigarh is not the one with the highest count; **Ludhiana**, Amritsar are the leading cities.

2. South India

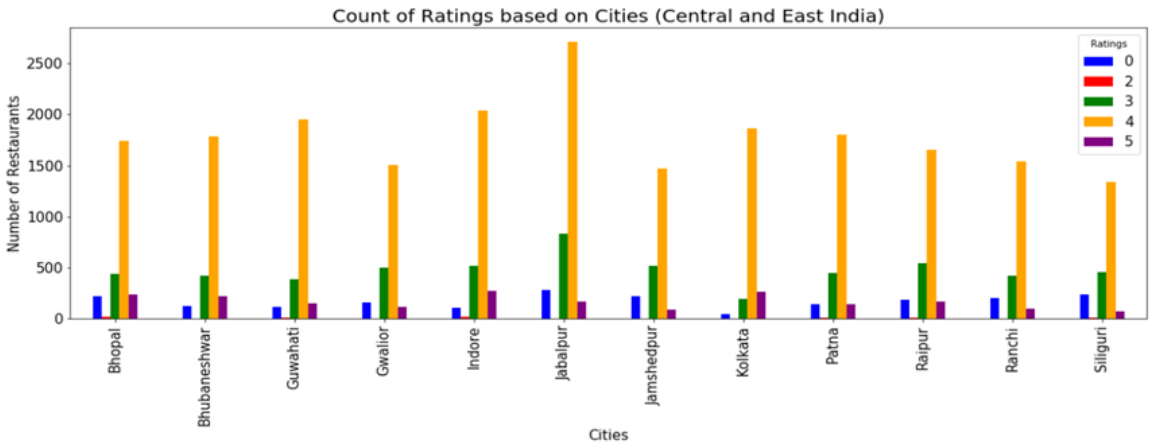


Observations -

- **Chennai** has the highest restaurant count in India, surpassing Delhi by **three** times.
- **Bangalore**, Kochi, Coimbatore's count is also high. Overall, the scenario of restaurants in South India is pretty good.

- Surprisingly, **Hyderabad** is not in the list, despite being the Biryani hub of India
- South India sees very few non-rated (0-star) restaurants.
- **Chennai** and **Bangalore** have a good number of **5 star** restaurants as well whereas other cities don't have much 5 star restaurants.

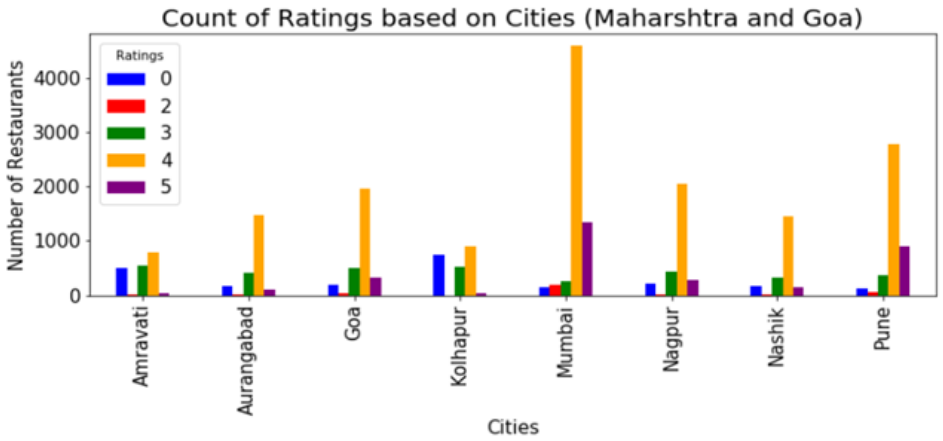
3. Central and East India



Observations -

- There is **no** City from the **North East** part of India (except **Guwahati**) in the top 12 cities with respect to restaurant numbers in Central and East India combined.
- MP's **Jabalpur**, Indore, Bhopal have the highest count in the given graph followed by Bhubaneshwar, Patna, Kolkata.

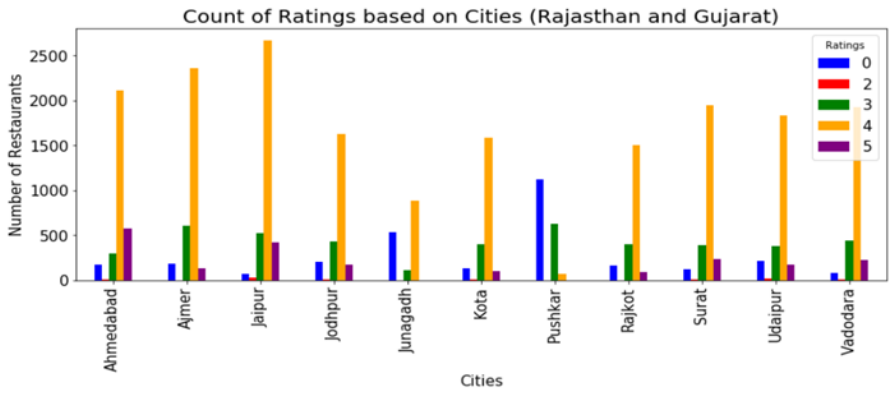
4. Maharashtra and Goa



Observations -

- Maharashtra has the **highest** number of cities which have **>2000** no. of restaurants.
- **Mumbai, Pune** have good no. of **5 star** restaurants. While in other cities, we don't find a high number of 5 star restaurants.
- **Kolhapur** has a substantial number of **non rated** restaurants.

5. Rajasthan and Gujarat



Observations -

- We can see that **Gandhinagar**, despite being the capital of Gujarat is not present in the top restaurant count cities of Gujarat and Rajasthan
- Infact, the count of restaurants in Gandhinagar is only **152**.
- **Pushkar** has maximum non rated restaurants.
- There aren't any 5 star hotels in Pushkar and Junagadh.

NLP based Analysis of Bangalore Restaurants Reviews

We have analysed the reviews of **Bangalore** restaurants. We chose Bangalore restaurants since there are a lot of restaurants in Bangalore and this was the only good data available. Moreover, we thought of analysing more cities' restaurant review data as well and comparing the same with each other, but we were not able to find the suitable data.

We have made several **EDA** (Exploratory Data Analysis) Analysis like **WordCloud, Sentiment Based Distribution, Aspect Based Analysis and Topic Modelling**.

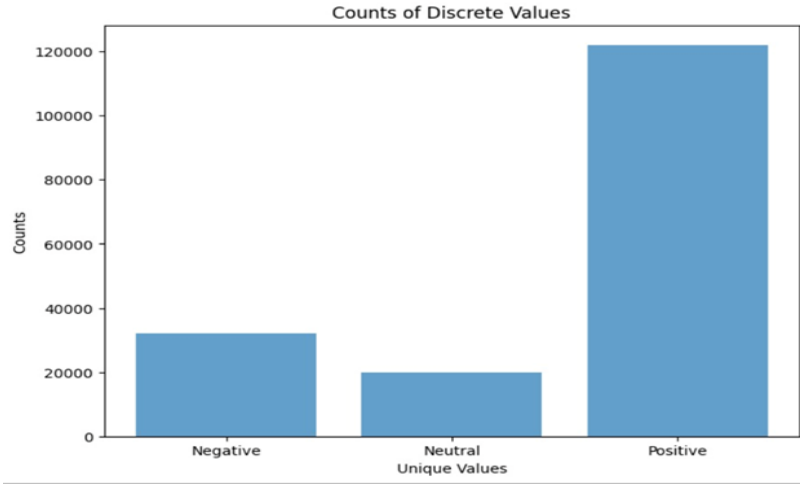
The Bangalore restaurant reviews data which we used had the following headers- url, address, name, online_order, book_table, rate, votes, phone, location, rest_type, dish_liked, cuisines, approx_cost(for two people), reviews_list, menu_item, listed_in(type), listed_in(city)

The **reviews_list** column has reviews of the restaurants with each entry having several reviews for a restaurant, combined together. There are more than **50,000** restaurants in the data.

1. WordCloud

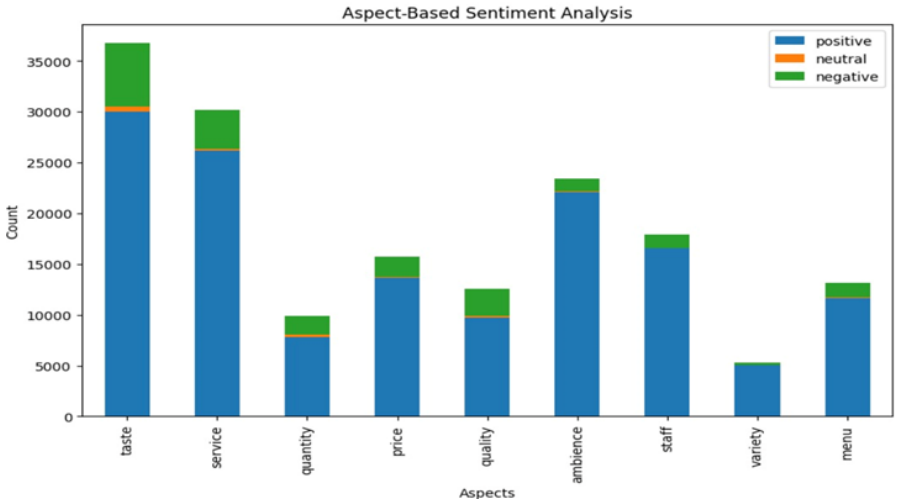
- Used **Wordcloud** and **Textblob** libraries in python to create the wordcloud.
- WordCloud suggests that **food, place, quality, service, quantity, price**, and **taste** are the aspects which are frequently mentioned in the reviews.
- Aspects like **flavour, pocket friendly, buffet, spicy, and location** are not much frequently used by customers in the reviews.
- This suggests that customers are more interested in the quality, quantity of food and the price and place of the restaurant.

2. Sentiment Distribution



- Using the **Textblob** and **matplotlib** python libraries, we plot the reviews dividing them into 3 - positive reviews, negative reviews and neutral reviews.
- Textblob library segregates the reviews depending on the words (which get converted into **vectors**) of the review, whether the review is neutral or positive or negative.
- From the graph, we see that **most** of the reviews have a **positive** sentiment. The y axis shows how many restaurant reviews have a particular sentiment.
- As expected **Neural** count is the least since people either write a positive review (eg - “The food was amazing”) of the restaurant or criticise the same (eg - “The food quality was extremely poor”), but hardly anyone writes a neutral review like “The food was fine”.

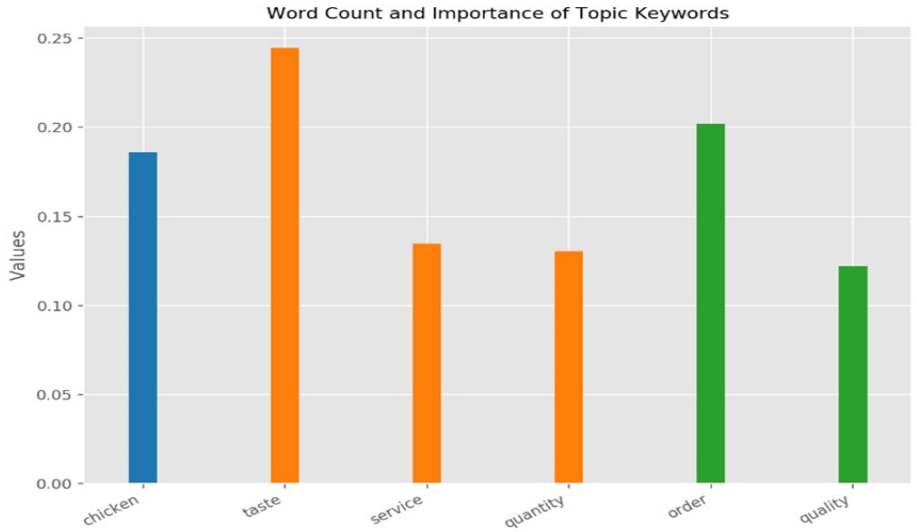
3. Aspect Based Sentiment Analysis



The graph shows the Sentiment distribution of different aspects.
Aspects are attributes or features of the reviews that we are interested in, such as **"taste", "service", "ambience,"** etc.
We use an Aspect Sentiments **Dictionary** which stores **sentiment** counts for each aspect (positive, negative, neutral) in a dictionary. Our **analyze_sentiment** function assesses sentiment in text and categorises it using the **Textblob** library of python.
The code loops through reviews, checks if an aspect is mentioned, and counts sentiments.
Attached graph shows that Indian people mostly focus on taste, service, ambience, while **variety, quantity** don't matter much.
All the aspects have **more positive** reviews than negative or neutral reviews since overall, the positive reviews count is highest.
Moreover, aspects like **Staff, Ambience, Variety** have a very high percentage of **positive** sentiment.

4. Topic Modelling

For Topic Modelling, we use the **sklearn** python library which is a ML library.
We are directly using the **LDA** (Linear Discriminant Analysis) function of sklearn to extract key words and hence topics.
In LDA, we convert each word into a **vector** and put correlated words together which are then grouped to form one **topic** like "quantity".
Whenever a new review is analysed, we correlate it with already existing topics and put the same into a particular topic.



- Using Topic Modelling, we found that in our reviews, **'taste, Chicken, order, service, quantity, quality'** were the most prominent topics.
- The topics mentioned here are the most important ones and summation of all topics values is 1.

Conclusions of Tourism and Restaurant Data Analysis

- India's restaurant landscape is shaped significantly by key players like **Domino's** Pizza, **CCD**, and KFC, with **35%** of establishments being a part of recognized chains.
- **Bangalore** stands out as a culinary hub with the **highest** restaurant density.
- Most restaurants are rated between 3 and 4 star with **alcohol** availability restaurants having **highest** average ratings and review votes.
- In India, we observe a preference towards **North Indian** and **Chinese** food, with **Barbeques** being the highest rated food chain.
- **Gurgaon** has the highest rated restaurants (average 3.83) whereas **Hyderabad** has more number of **critics** (votes).
- There are comparatively **less** number of restaurants at **higher price** ranges and their rating is generally high.
- In the culinary landscape, **Chennai** emerges as a standout with a **threefold** number of restaurants compared to Delhi, the **highest** in the nation.
- Only big **metro cities** like Mumbai, Bangalore, Chennai, Delhi, Pune, Lucknow have substantial amounts of **5 star** restaurants.
- **South Indian** Cities have almost negligible non rated restaurants.
- **Covid** has affected the **Tourism** industry which in return affected the **Hotel** Industry.
- Cities like **Agra**, Jodhpur, Hampi, Dharwad, Amravati have high monumental visit tourists but **poor Hotels** availability, thus creating a **potential** for Hotel Industry.
- **EDA** and **Aspect analysis** show similar results in terms of topics of concern.

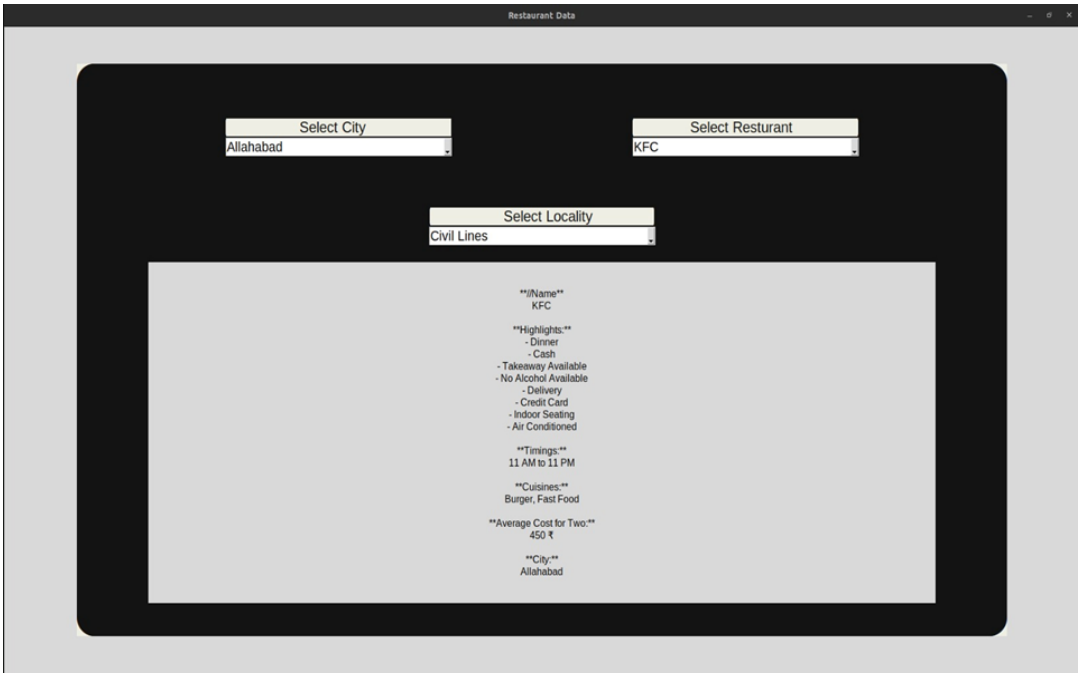
Graphical User Interface (GUI)

The Python script is a Tkinter-based GUI application that allows users to query restaurant or hotel information based on city, state, and locality. It utilizes the **Bard API** to generate a detailed and well-formatted response for the selected restaurant criteria.

Key Features:
1. **Dropdown Menus:** Users can select a city, restaurant/hotel, and locality using dropdown menus.
2. **Dynamic Suggestions**: As users type, the dropdown menus dynamically suggest matching options to assist in selection.
3. **Response Generation:** Upon selecting the criteria, the application generates a formatted response using the Bard API.
4. **User-Friendly Interface**: The GUI is designed with labeled sections for selecting city, restaurant/hotel, and locality, providing a clear and organized layout.
5. **Output Display**: The detailed information about the selected restaurant/hotel is displayed in a labeled output section.

Note: The code assumes the existence of a '**data.pickle**' file containing restaurant information which is scrapped from **Zomato API** and 'customtkinter' and 'bardapi' modules for enhanced GUI and Bard API functionality, respectively.

This application facilitates easy exploration of restaurant/hotel details, providing a user-friendly experience for accessing information based on specific criteria.



References

- https://tourism.gov.in/sites/default/files/202209/India%20Tourism%20Statistics%202021%20(1).pdf – Government Tourism Report
- https://www.kaggle.com/code/shahules/zomato-complete-eda-and-lstm-model/notebook - Kaggle's Bangalore Restaurants Data
- https://arslanr369.medium.com/exploring-indian-restaurants-dataset-a-comprehensive-eda-on-kaggle-d2fcf34b6f3f - EDA of Indian Restaurants
- https://www.kaggle.com/code/devarsheesandilya/zomato-delhi-eda/notebook - Delhi Restaurants Analysis
- https://nidhi.nic.in/MOT/Catrpt.aspx – Government Hotel Data
- https://www.kaggle.com/code/kiranreddyrebel/hotel-city-visualization/notebook – Kaggle's MakeMyTrip Data
- https://www.kaggle.com/code/rjtmehta99/e-d-a-of-chennai-hotel-reviews – Kaggle's Chennai Reviews Data
- https://tourism.gov.in/sites/default/files/202209/India%20Tourism%20Statistics%202021%20(1).pdf – Govt Tourism Report