

IBM Applied Data Science Capstone

*Recommending a Business at a particular Tourist
Venue*

Introduction

Tourism has always been a booming sector across the globe. No matter which country you live in, you can always come across a group of people, big or small, who always like to visit places. I being an adventurer myself can acknowledge this fact as to how tourism plays a salient role for a traveler/explorer. Tourism is not only an important aspect of a country's economy but also for its global standing.

Why Tourism is important to any country?

The tourism industry is important for the benefits it brings and due to its role as a commercial activity that creates demand and growth for many more industries. Tourism not only contributes to more economic activities but also generates more employment, revenues, and play a significant role in development.

- i. Tourism activity creates demand.
- ii. Tourism industry value chain meets & spreads demand across industries & boosts more economic activities.
- iii. Tourism induces more consumption.

Business Problem

All the benefits of tourism tend to reflect on the employment opportunity which it gives to the people of that country. The objective of this project is to analyze the tourist places of a given state in India and try to recommend the best location where they can open a restaurant or lodging to make the best use of the opportunity.

The target audience for this project includes people who are interested in opening a restaurant, lodging, transport services, or any other similar businesses which fall within the tourism industry. This also recommends travelers' tourist venues to be visited in a given state of a country.

Data Anatomization

To tackle the above mentioned problem, we need to have the dataset that contains -

- i. All the districts of a particular country.
- ii. Latitude and longitudes of all the districts.

The Wikipedia page https://en.wikipedia.org/wiki/List_of_districts_in_India is the major source of data that is being used to obtain all the districts of India. We then use beautifulsoup4 package, a Python module that helps to scrape information from the web pages to extract all the tables from this Wikipedia page and convert it into a pandas dataframe. Then we use Python's geopy package to obtain the latitude and longitude of all the districts present in the dataframe.

Description of the data -

The output shows the final dataset. The dataset consists of a single Dataframe with 9 columns containing state, districts, latitude and longitudes of the particular district etc. Other columns like code, headquarters, population, area and density have also been scraped from the website which can be used for further analysis.

| | State | Code | District | Headquarters | Population(2011) | Area(km2) | Density(/km2) | Latitude | Longitude |
|---|---------------------|------|--------------------------|--------------|------------------|-----------|---------------|-----------|-----------|
| 0 | Andaman and Nicobar | NI | Nicobar | Car Nicobar | 36842 | 1841.0 | 20 | 7.000000 | 93.000000 |
| 1 | Andaman and Nicobar | NaN | North and Middle Andaman | Mayabunder | 105597 | 3736.0 | 28 | 12.611239 | 92.831654 |
| 2 | Andaman and Nicobar | SA | South Andaman | Port Blair | 238142 | 2872.0 | 89 | 10.705890 | 92.487468 |
| 3 | Andhra Pradesh | AN | Anantapur | Anantapur | 4083315 | 19130.0 | 213 | 14.654623 | 77.556260 |
| 4 | Andhra Pradesh | CH | Chittoor | Chittoor | 4170468 | 15152.0 | 275 | 13.160105 | 79.155551 |

Literature Review

There are specific factors within the characteristics of the population which makes the tourism industry lead to an improvement of the socio-economic conditions of the population [1]. This will eventually result in low rates of unemployment and a higher percentage of the working population. The former improves the socioeconomic conditions of the population whereas the latter helps finance, through different tax burdens, public policies aimed at achieving a higher level of economic development. It also demonstrates that countries with regressive population pyramids have greater difficulties for tourism growth to improve their socio-economic conditions.

The survey from Annual Report Tourism of India provides us with the following facts –

- i. Tourism has contributed around 5.06% share in GDP during 2016-17
- ii. There were 1854.93 million domestic tourist visits all over the country during the year 2018.
- iii. Foreign Tourist Arrivals during 2019 were 10.89 million (Provisional) with a growth of 3.2% over the same period of the previous year
- iv. Foreign Exchange Earnings during the period during Jan 2019 – Dec 2019 were Rs.2,10,981 crores (Provisional estimates) with a growth of 8.3% over the same period of the previous year.
- v. According to Tourist Satellite Account, the tourism industry has provided around 87.50 million people employment opportunities in the year 2018-19
- vi. The above-obtained statistics highlight the importance of the Tourism Industry in the overall development of the country.

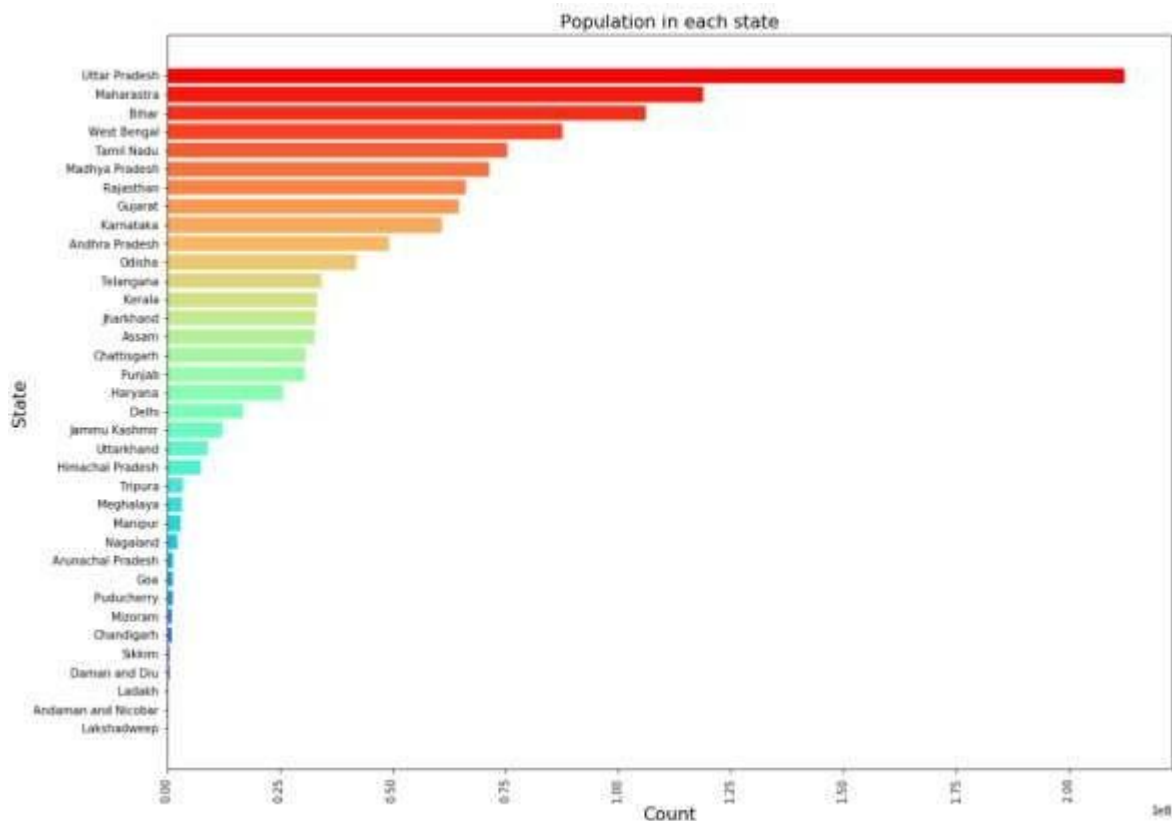
Methodology

The first step is to collect the data. This is done by scraping the Wikipedia page https://en.wikipedia.org/wiki/List_of_districts_in_India. Then we use geopy API to get the latitude and longitude of all the districts of the country. There existed some missing values in the dataset which were removed. The final dataset has nine columns as shown –

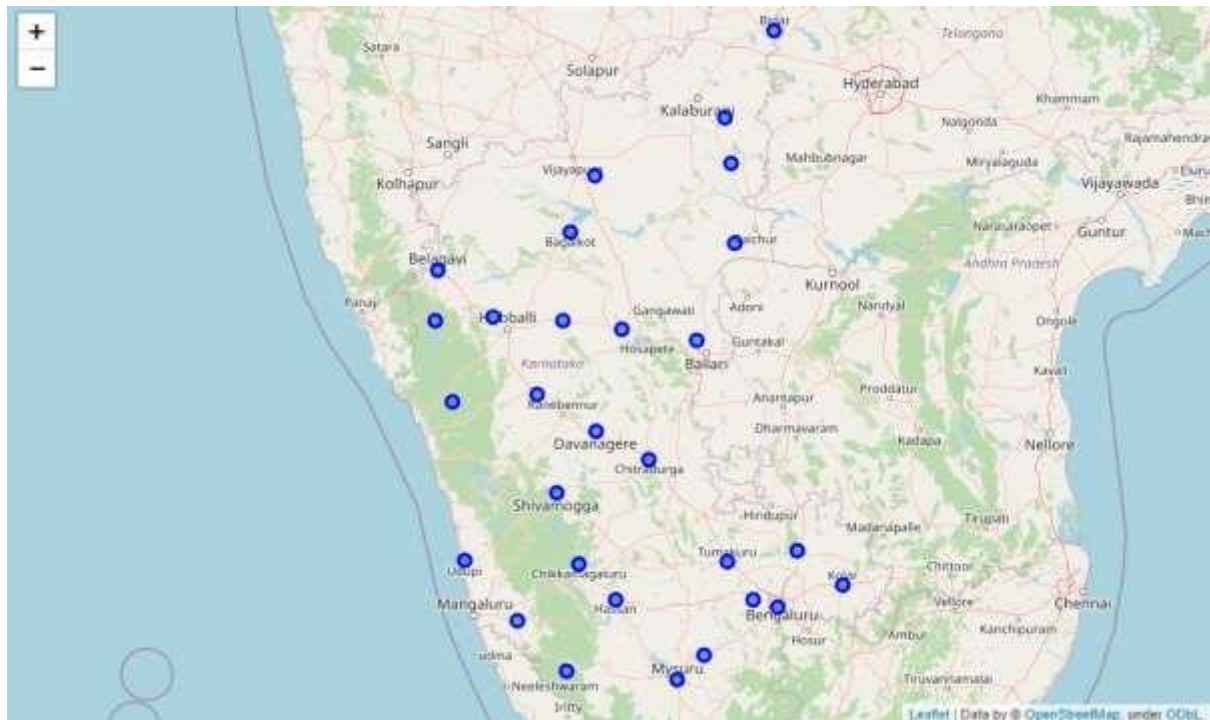
| | State | Code | District | Headquarters | Population(2011) | Area(km2) | Density(/km2) | Latitude | Longitude |
|---|---------------------|------|--------------------------|--------------|------------------|-----------|---------------|-----------|-----------|
| 0 | Andaman and Nicobar | NI | Nicobar | Car Nicobar | 36842 | 1841.0 | 20 | 7.000000 | 93.000000 |
| 1 | Andaman and Nicobar | NaN | North and Middle Andaman | Mayabunder | 105597 | 3736.0 | 28 | 12.611239 | 92.831654 |
| 2 | Andaman and Nicobar | SA | South Andaman | Port Blair | 238142 | 2872.0 | 89 | 10.705890 | 92.487468 |
| 3 | Andhra Pradesh | AN | Anantapur | Anantapur | 4083315 | 19130.0 | 213 | 14.654623 | 77.556260 |
| 4 | Andhra Pradesh | CH | Chittoor | Chittoor | 4170468 | 15152.0 | 275 | 13.160105 | 79.155551 |

There are 36 states (including Union Territories) which have been retrieved from the webpage and stored in the dataset.

As mentioned in the literature review, there can be some impacts of the population of a state on tourism. The below graph shows the population in each state.

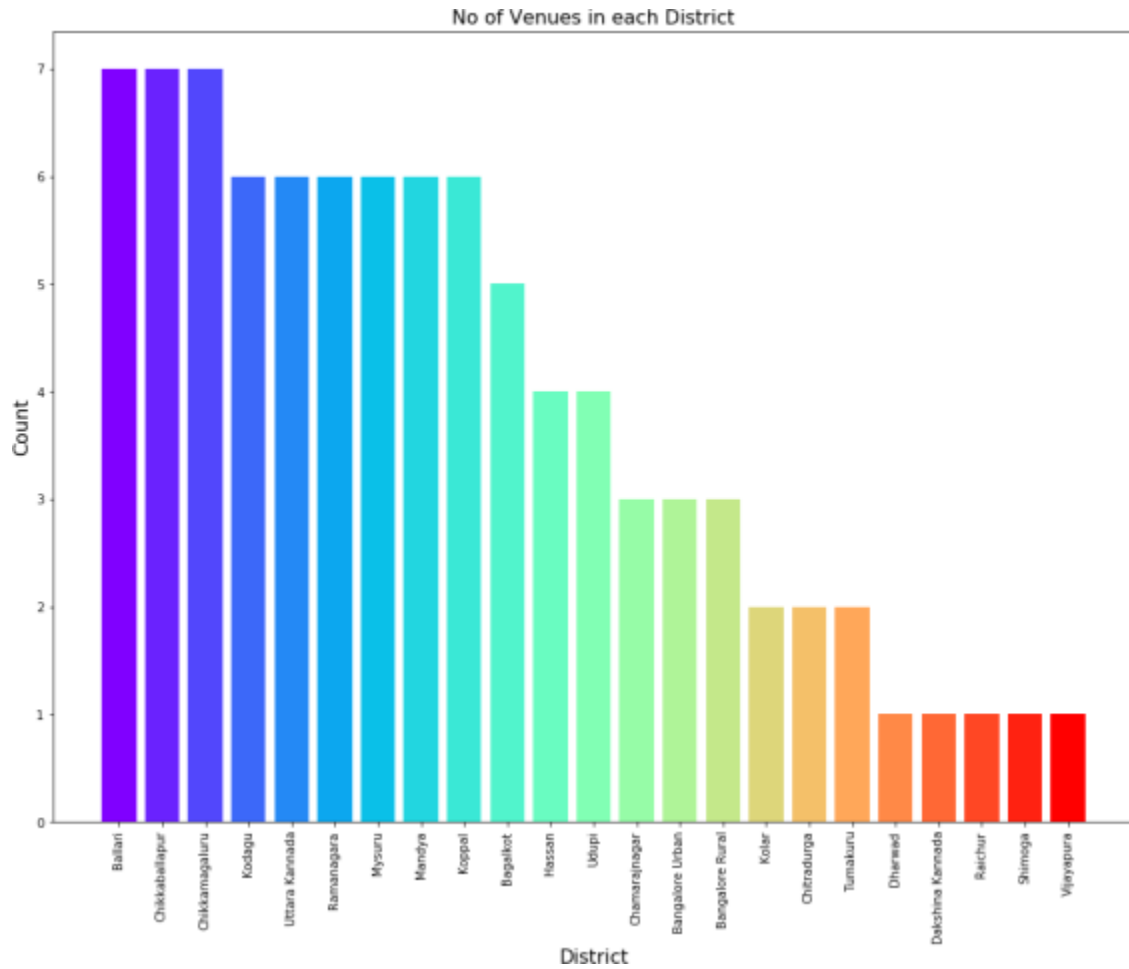


The user can enter the state of his choice among the given states. Here Karnataka is taken as a choice. A visualization with all the districts of the given state will be displayed as shown below –



Using the Foursquare API, we acquire only the categories which are related to tourism for tourists category and which are related to tourist services for employment opportunities to people separately. The former includes Arts & Entertainment, Nightlife Spot, Outdoors & Recreation, whereas the latter includes Food, Shop & Service, Travel & Transport services.

The next step is to obtain the nearby tourist venues within a radius of 50km. This gives us multiple tourist spots if there are in a particular district. We visualize a bar graph by plotting District v/s count to obtain the number of venues in each district. The visualization can be shown below –



We then organize the unique venue categories obtained and create a one-hot encoding to analyze each district. This results in a Dataframe that displays the most common venue category in a particular district. The result is as shown –

| | District | 1st Most Common Venue Category | 2nd Most Common Venue Category | 3rd Most Common Venue Category | 4th Most Common Venue Category | 5th Most Common Venue Category | 6th Most Common Venue Category | 7th Most Common Venue Category |
|---|-----------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|
| 0 | Bagalkot | Sculpture Garden | Historic Site | Scenic Lookout | Zoo | Mountain | Bowling Alley | Brewery |
| 1 | Ballari | Historic Site | Mountain | Scenic Lookout | Zoo | Bowling Alley | Brewery | Campground |
| 2 | Bangalore Rural | Theme Park | Bowling Alley | Zoo | Mountain | Brewery | Campground | Castle |
| 3 | Bangalore Urban | Plaza | Performing Arts Venue | Park | Zoo | Island | Bowling Alley | Brewery |
| 4 | Chamarajnagar | Mountain | Garden | National Park | Zoo | Bowling Alley | Brewery | Campground |

We then aggregate all the venues which belong to the particular category in a particular district.

| | District | Venue Category | Venue |
|---|----------|------------------|---|
| 0 | Bagalkot | Historic Site | Aihole, Pattadakal - World Heritage Site |
| 1 | Bagalkot | Scenic Lookout | Badami |
| 2 | Bagalkot | Sculpture Garden | Aihole Rock Cut Shiva Temple, Pattadakal Temple |
| 3 | Ballari | Historic Site | Hampi, Hanuman Temple, Vitthala temple, Queens Bath |
| 4 | Ballari | Mountain | Martanga Hill, Hemakuta Hill |

After obtaining the most common venue categories in all the districts, we replace the categories with the venues if they are present in the district.

| | District | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue |
|---|-----------------|---|--|------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| 0 | Bagalkot | Aihole Rock Cut Shiva Temple, Pattadakal Temple | Aihole, Pattadakal - World Heritage Site | Badami | | | | |
| 1 | Ballari | Hampi, Hanuman Temple, Vitthala temple, Queens Bath | Martanga Hill, Hemakuta Hill | Lotus Mahal | | | | |
| 2 | Bangalore Rural | Wonderla Amusement Park, Wonder La | PVR bioO | | | | | |
| 3 | Bangalore Urban | M G Road Boulevard | Rangashankara | Cubbon Park | | | | |
| 4 | Chamaraj Nagar | Mudhumalai Forest | Rose Garden | Bandipur National Park | | | | |
| 5 | Chikkaballapur | Bhartiya City | The Druid Garden | Gangamma Circle | Richard's Park | Our Native Village | Nandi Hills | nandi hills |

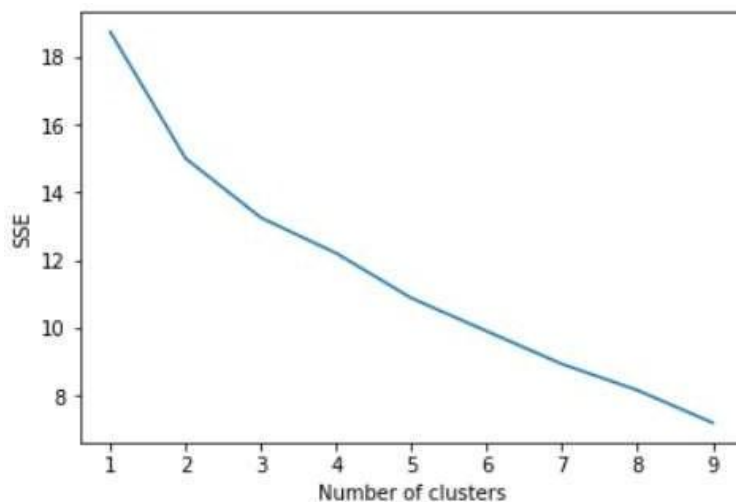
This gives an idea to a person as to where he could start his business in a particular district. But still, he can be not sure or have any idea as to what type of business he could open up at a given tourist venue. So to make sure that his business attracts many tourists as possible, we then attempt to find the most sought business at the tourist spot. So, then we acquire the top businesses which are being established at the tourist venue within the range of 500 meters.

| | Venue | Business | BLatitude | BLongitude | Business Category |
|----|--------|-------------------------|-----------|------------|-------------------------------|
| 13 | Badami | Sangam Restaurant | 15.924083 | 75.679891 | Vegetarian / Vegan Restaurant |
| 15 | Badami | Hotel New Satkar Deluxe | 15.924083 | 75.679891 | Hotel |
| 17 | Badami | Hotel Mookambika Deluxe | 15.922389 | 75.683092 | Hotel |
| 19 | Hampi | Gopi Roof Restaurant | 15.336163 | 76.480259 | Indian Restaurant |
| 20 | Hampi | Funky monkey | 15.336225 | 76.481525 | Indian Restaurant |
| 21 | Hampi | Laughing Buddha | 15.338600 | 76.456436 | Cafe |
| 22 | Hampi | Mango Tree Restaurant | 15.335544 | 76.480337 | Indian Restaurant |

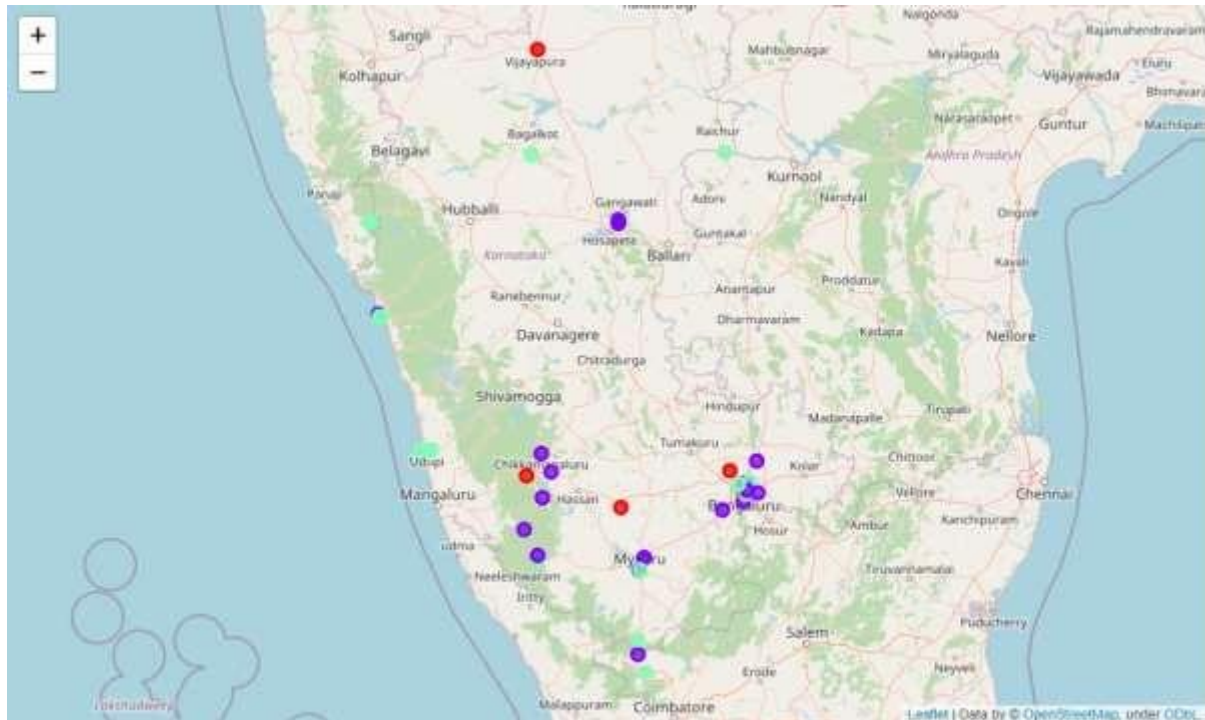
We then perform similar one hot encoding and analyse each venue to get the top businesses at a venue.

| | Venue | 1st Most Common Business | 2nd Most Common Business | 3rd Most Common Business | 4th Most Common Business | 5th Most Common Business | 6th Most Common Business | 7th Most Common Business | 8th Most Common Business | 9th Most Common Business | 10th Most Common Business |
|---|------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|---------------------------|
| 0 | Abbey Falls | Hotel | Indian Restaurant | Vegetarian / Vegan Restaurant | Comfort Food Restaurant | Fast Food Restaurant | Electronics Store | Dry Cleaner | Donut Shop | Diner | Dessert Shop |
| 1 | Badami | Hotel | Vegetarian / Vegan Restaurant | Food Court | Fast Food Restaurant | Electronics Store | Dry Cleaner | Donut Shop | Diner | Dessert Shop | Department Store |
| 2 | Bandipur National Park | Rest Area | Vegetarian / Vegan Restaurant | Coffee Shop | Fast Food Restaurant | Electronics Store | Dry Cleaner | Donut Shop | Diner | Dessert Shop | Department Store |
| 3 | Bhartiya City | Garden Center | Pizza Place | Cafe | Vegetarian / Vegan Restaurant | Comfort Food Restaurant | Fast Food Restaurant | Electronics Store | Dry Cleaner | Donut Shop | Diner |
| 4 | Channakeshwara Temple | Vegetarian / Vegan Restaurant | Indian Restaurant | Food Court | Fast Food Restaurant | Electronics Store | Dry Cleaner | Donut Shop | Diner | Dessert Shop | Department Store |

We then use the K-means clustering algorithm to group the businesses into clusters that aim to partition 'n' observations into k clusters in which each observation belongs to the cluster. Here elbow method is used to determine the optimum value of k to perform K-means clustering. The graph obtained is –



Results and Discussion



The colors purple, green, and red represents cluster 0, 1, and 2 respectively.

The results show that the most common business in cluster one at the respective venues are Indian Restaurants. So Indian Restaurants are popular in these tourist venues and opening up a similar one can attract many tourists. This is because India is a land with many cultures. Tourists always like to experience the flavor of local dishes available at a particular location and so this could be a nice opportunity to open up a business at that locality.

Cluster 0 –

| | Venue | 1st Most Common Business | 2nd Most Common Business | 3rd Most Common Business | 4th Most Common Business | 5th Most Common Business | 6th Most Common Business | 7th Most Common Business | 8th Most Common Business | 9th Most Common Business | 10th Most Common Business |
|----|----------------|--------------------------|-------------------------------|--------------------------|-------------------------------|--------------------------|--------------------------|--------------------------|-------------------------------|--------------------------|---------------------------|
| 19 | Hampi | Indian Restaurant | Restaurant | Cafe | Vegetarian / Vegan Restaurant | Comfort Food Restaurant | Fast Food Restaurant | Electronics Store | Dry Cleaner | Donut Shop | Diner |
| 31 | Hanuman Temple | Indian Restaurant | Vegetarian / Vegan Restaurant | Food Court | Fast Food Restaurant | Electronics Store | Dry Cleaner | Donut Shop | Diner | Dessert Shop | Department Store |
| 38 | Martanga Hill | Indian Restaurant | Vegetarian / Vegan Restaurant | Food Court | Fast Food Restaurant | Electronics Store | Dry Cleaner | Donut Shop | Diner | Dessert Shop | Department Store |
| 45 | Hemakuta Hill | Indian Restaurant | Restaurant | Cafe | Vegetarian / Vegan Restaurant | Comfort Food Restaurant | Fast Food Restaurant | Electronics Store | Dry Cleaner | Donut Shop | Diner |
| 54 | Wonder La | Indian Restaurant | Restaurant | Pizza Place | Vegetarian / Vegan Restaurant | Coffee Shop | Electronics Store | Dry Cleaner | Donut Shop | Diner | Dessert Shop |
| 84 | Rangashankara | Indian Restaurant | Fast Food Restaurant | Seafood Restaurant | Electronics Store | Diner | Coffee Shop | Sandwich Place | Vegetarian / Vegan Restaurant | Cafe | Snack Place |

Whereas in cluster two the most sought business is the Hotel, Seafood Restaurants, and Cafeterias. This is clearly visible in the map above. The green clusters at the seaside clearly indicate that opening a seafood restaurant would help a person make the best use of the opportunity. Also, there are some green clusters in the middle of the map, which indicates Hotels and Cafeterias would be the best business at that tourist spot.

Cluster 1 –

| | | | | | | | | | | | |
|-----|------------------|--------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------|-------------------------------|-------------------|-------------|--------------|------------------|
| 435 | St Mary's Island | Seafood Restaurant | Fast Food Restaurant | Vegetarian / Vegan Restaurant | Coffee Shop | Electronics Store | Dry Cleaner | Donut Shop | Diner | Dessert Shop | Department Store |
| 442 | End Point | Hotel | Sandwich Place | Convenience Store | Vegetarian / Vegan Restaurant | Coffee Shop | Fast Food Restaurant | Electronics Store | Dry Cleaner | Donut Shop | Diner |
| 447 | Kudle Beach | Café | Mediterranean Restaurant | Seafood Restaurant | Hotel | Pizza Place | Vegetarian / Vegan Restaurant | Electronics Store | Dry Cleaner | Donut Shop | Diner |
| 468 | Ori Beach | Hotel | Cafeteria | Café | Vegetarian / Vegan Restaurant | Comfort Food Restaurant | Fast Food Restaurant | Electronics Store | Dry Cleaner | Donut Shop | Diner |
| 472 | Half Moon Beach | Seafood Restaurant | Vegetarian / Vegan Restaurant | Coffee Shop | Fast Food Restaurant | Electronics Store | Dry Cleaner | Donut Shop | Diner | Dessert Shop | Department Store |

Finally, in cluster three Fast Food/Vegetarian Restaurants have been given a top priority.

Cluster 2 –

| | Venue | 1st Most Common Business | 2nd Most Common Business | 3rd Most Common Business | 4th Most Common Business | 5th Most Common Business | 6th Most Common Business | 7th Most Common Business | 8th Most Common Business | 9th Most Common Business | 10th Most Common Business |
|-----|----------------------|--------------------------|-------------------------------|-------------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|---------------------------|
| 113 | Cubbon Park | Fast Food Restaurant | Sandwich Place | Vegetarian / Vegan Restaurant | Coffee Shop | Electronics Store | Dry Cleaner | Donut Shop | Diner | Dessert Shop | Department Store |
| 191 | Our Native Village | Dry Cleaner | Vegetarian / Vegan Restaurant | Toy / Game Store | Fish & Chips Shop | Fast Food Restaurant | Electronics Store | Donut Shop | Diner | Dessert Shop | Department Store |
| 217 | Mudigere | Fast Food Restaurant | Vegetarian / Vegan Restaurant | Toy / Game Store | Fish & Chips Shop | Electronics Store | Dry Cleaner | Donut Shop | Diner | Dessert Shop | Department Store |
| 282 | Lord Bahubali Temple | Asian Restaurant | Vegetarian / Vegan Restaurant | Comfort Food Restaurant | Fish & Chips Shop | Fast Food Restaurant | Electronics Store | Dry Cleaner | Donut Shop | Diner | Dessert Shop |
| 479 | Gol Gumbaz | Comfort Food Restaurant | Vegetarian / Vegan Restaurant | Toy / Game Store | Fish & Chips Shop | Fast Food Restaurant | Electronics Store | Dry Cleaner | Donut Shop | Diner | Dessert Shop |

Conclusion

In this project, an attempt has been made to make use of the Foursquare API to get the famous tourist locations situated in a particular district of a State. K-means clustering algorithm has been used to cluster these tourist spots based on exploring the frequency of the businesses that are present which could help us indicate a business opportunity that could be established in the locality so that the business could attract as many tourists as possible.

Future possible research could make use of other significant factors which includes the foot traffic where the tourists are likely to bypass the area i.e a high traffic area, competition i.e the number of similar businesses that could impact the new business being established, accessibility, and average business rates that could be incurred for a particular business. These above-mentioned factors could help the system make the analysis more accurate.

References

[1] Rivero, Marcelino & Cárdenas-García, Pablo. (2014). Population characteristics and the impact of tourism on economic development. *Tourism Geographies*. 16. 10.1080/14616688.2014.889207.