Applied Data Science Capstone

IBM

Shubham Chhetri

Business Finder to open a business in India

Introduction:

India is a large country full of opportunities. The objective of this project is to analyze the various regions in our country based on latitude and longitude values and give an overview of the place based on the various businesses currently running in the particular region with up to 12th most common business in the region and clustering the similar regions based on their most common business and visually depicting it on the map with the help of Follium library. By looking at the map, we can easily visualize the region and observe the different regions and their most common business hotspot to overcome the problem of opening a business with lack of idea.

Business Problem:

Opening a business by investing a huge sum of our savings or loan amount, we are taking a chance. It can make us more prosperous or we might be in a huge debt or incur huge losses. So there should be some knowledge about a place where we are opening a business like what businesses are popular in that area and things like that. With the information from this project, we can significantly get a head start in our business and the chance of risk is significantly reduced.

Data Source:

The dataset for this project was borrowed from the Kaggle datasets, contributed by the user Kiran Reddy. You can view the dataset by clicking on the link: https://www.kaggle.com/kiranreddy24/indian-states-latitudes-and-longitudes. The dataset was then stored on IBM Db2 database and was accessed through the notebook on IBM Watson Studio where this whole project was created.

The data contains three columns with column name: State, Latitude and Longitude which helped in properly communicating with the Foursquare API for gathering the information of the region.

The major portion of the data that we will be working on will come from the Foursquare API and it will be stored in a pandas data frame.

Data Description:

First five rows of the Kaggle Dataset is shown below, the data contains three columns with column name: State, Latitude and Longitude. The data is pretty concise and needed no modification.

	State	Latitude	Longitude
0	Andaman and Nicobar Islands	11.667026	92.735983
1	Andhra Pradesh	15.910429	79.747003
2	Arunachal Pradesh	27.100399	93.616601
3	Assam	26.749981	94.216667
4	Bihar	25.785414	87.479973

Methodology:

The first step is to retrieve the data. The data was burrowed from the Kaggle website. The data that we need required the coordinates of a state in India to communicate with the Foursquare API for gathering the information of the region since the major portion of the data that we will be working on will come from the Foursquare API and it will be stored in a pandas data frame.

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

The user can enter the state of his choice among the given states. Using the Foursquare API, we acquire only the categories data which are related to business and exclude other categories.

People cannot be sure nor have any idea as to what type of business he could open up at a given venue or place. So to make sure that his business attracts many customers as possible, we attempt to find the most sought business at the particular region. So, then we acquire the top businesses which are being established at the tourist venue within the range of 500 meters.

Tabular visualization of businesses nearby:

	Business Category	No of Businesses
0	Airport	1
1	American Restaurant	5
2	Asian Restaurant	15
3	Auto Dealership	1
4	BBQ Joint	2
5	Bakery	11
6	Bistro	3
7	Bookstore	2
8	Breakfast Spot	1
9	Buffet	1
10	Burger Joint	1
11	Café	45
12	Caribbean Restaurant	2
13	Clothing Store	6
14	Coffee Shop	22

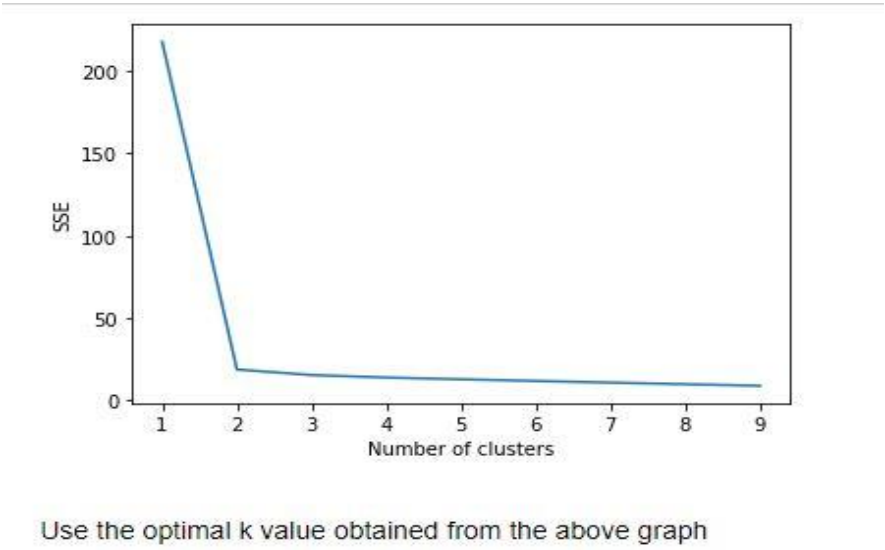
We then perform similar one hot encoding and analyze each venue to get the top businesses at a venue.

	Venue	Airport	American Restaurant	Asian Restaurant	Auto Dealership	BBQ Joint	Bakery	Bistro	Bookstore	Breakfast Spot	Buffet	Burger Joint	Café	Caribbean Restaurant	Clothing Store	Coffee Shop	Convenience Store	Cosmetics Shop	De Bode
1	Andaman & Nicobar Islands	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Andaman & Nicobar Islands	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	Sector 36	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
9	Sector 36	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
11	Sector 36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

After that created columns according to number of top business (up to 12th most common business) and then created a new data frame.

	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	11th Most Common Business	12th Most Common Business
0	Ana Sagar Lake	Hotel	Airport	Music Store	Furniture / Home Store	Gastropub	Indian Restaurant	Italian Restaurant	Juice Bar	Market	Mediterranean Restaurant	Middle Eastern Restaurant	Miscellaneous Shop
1	Andaman & Nicobar Islands	Airport	Auto Dealership	Pharmacy	Gastropub	Hotel	Indian Restaurant	Italian Restaurant	Juice Bar	Market	Mediterranean Restaurant	Middle Eastern Restaurant	Miscellaneous Shop
2	Barapani Lake	Indian Restaurant	Airport	Music Store	Furniture / Home Store	Gastropub	Hotel	Italian Restaurant	Juice Bar	Market	Mediterranean Restaurant	Middle Eastern Restaurant	Miscellaneous Shop
3	Bharathi Park	Coffee Shop	Hotel	Café	Indian Restaurant	Bakery	Pizza Place	Italian Restaurant	Department Store	Vegetarian / Vegan Restaurant	French Restaurant	Sandwich Place	Tourist Information Center
4	Biju Pattnaik Park (Forest Park)	Pharmacy	Music Store	Furniture / Home Store	Gastropub	Hotel	Indian Restaurant	Italian Restaurant	Juice Bar	Market	Mediterranean Restaurant	Middle Eastern Restaurant	Miscellaneous Shop

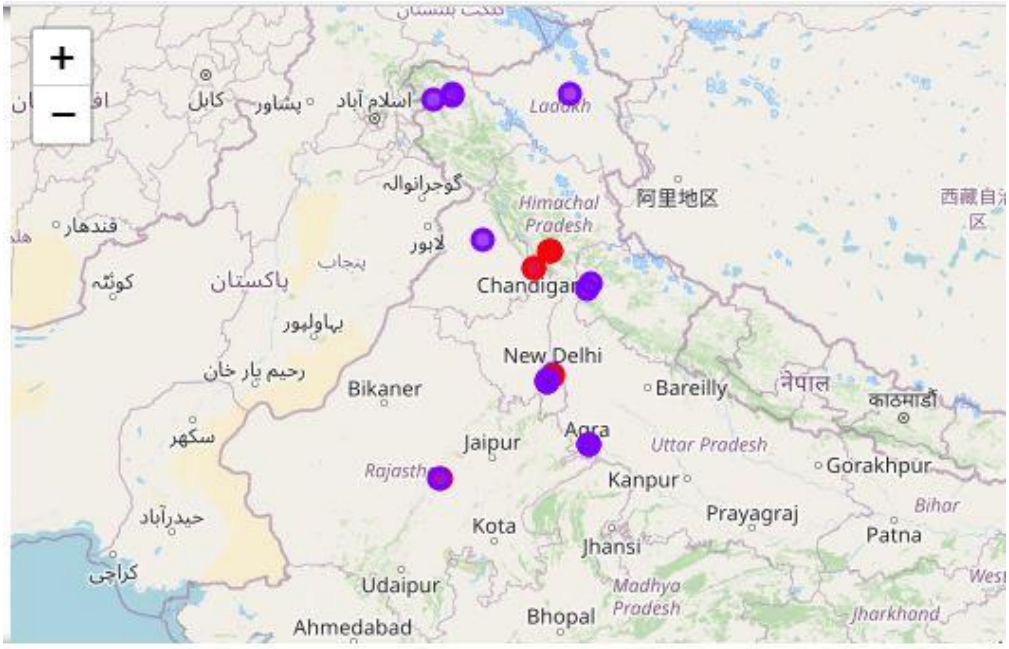
We then use the K-means clustering algorithm to group the businesses into clusters that aim to partition ‘n’ observations into k clusters in which each observation belongs to the cluster. Here elbow method is used to determine the optimum value of k to perform K-means clustering. The graph obtained is:



Use the optimal k value obtained from the above graph

Hence, k=2

Results and Discussion:



The colors purple and red represents cluster 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 venues and opening up a similar one can attract many customers.

Cluster 1:

Cluster 1

```
nearby_business_merged.loc[nearby_business_merged['Cluster Labels'] == 0, nearby_business_merged.columns[[0] + list(range(4, nearby_business_merged.shape
```

33	Resort & Water Park	Indian Restaurant	Airport	Music Store	Furniture / Home Store	Gastropub	Hotel	Italian Restaurant	Juice Bar	Market	Mediterranean Restaurant	Eastern Restaurant	Miscellaneous Shop
36	Connaught Place | कनॉट प्लेस (Connaught Place)	Indian Restaurant	Café	Food & Drink Shop	Molecular Gastronomy Restaurant	Portuguese Restaurant	Coffee Shop	Clothing Store	Food Truck	Bistro	Bakery	Asian Restaurant	Spa
63	Sinquerium Beach	Restaurant	Italian Restaurant	Hotel	Indian Restaurant	Spa	Seafood Restaurant	Portuguese Restaurant	Caribbean Restaurant	Airport	Molecular Gastronomy Restaurant	Gastropub	Juice Bar
86	Candolim Beach	Indian Restaurant	Asian Restaurant	Restaurant	American Restaurant	BBQ Joint	Italian Restaurant	Clothing Store	Seafood Restaurant	Airport	Miscellaneous Shop	Middle Eastern Restaurant	Juice Bar
113	Hilltop	Burger Joint	Restaurant	Airport	Music Store	Gastropub	Hotel	Indian Restaurant	Italian Restaurant	Juice Bar	Market	Mediterranean Restaurant	Middle Eastern Restaurant
118	DLF CyberHub	Indian Restaurant	Café	Asian Restaurant	Portuguese Restaurant	American Restaurant	Italian Restaurant	Bistro	Pizza Place	Gastropub	Coffee Shop	Falafel Restaurant	Middle Eastern

Finally in cluster 2, hotels and music stores should be given the importance.

Cluster 2

```
earby_business_merged.loc[nearby_business_merged['Cluster Labels'] == 1, nearby_business_
```

	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	7 C B
13	Rose Garden	Hotel	Airport	Music Store	Furniture / Home Store	Gastropub	Indian Restaurant	Re
61	Lodhi Gardens (लोधी बाग़) (Lodhi Gardens)	Food Court	Music Store	Furniture / Home Store	Gastropub	Hotel	Indian Restaurant	Re
257	Indian Institute Of Advanced Studies	Cosmetics Shop	Airport	Music Store	Gastropub	Hotel	Indian Restaurant	Re
259	Viceregal Lodge	Hotel	Airport	Music Store	Furniture / Home Store	Gastropub	Indian Restaurant	Re
292	Mysore Palace	Hotel	Airport	Music Store	Furniture / Home Store	Gastropub	Indian Restaurant	Re

Conclusion:

In this project, an attempt has been made to make use of the Foursquare API to get the information of locations situated in a particular State. K-means clustering algorithm has been used to cluster these 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 customer as possible.

Future Scope:

Future possible research could make use of other significant factors which includes the foot traffic where the customers are likely to bypass the area (e.g.: a high traffic area), competition (e. g.: 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:

- https://developer.foursquare.com
- https://www.kaggle.com/kiranreddy24/indian-states-latitudes-and-longitudes
- www.wikipedia.org