IBM Data Science Professional certificate Capstone

-Harshad Pednekar

Introduction/Business Problem

Goa is a state in Southern-India and is my birthplace. So the project focuses on the state of Goa and through this project i hope to analyse the trends of Goan people based on their frequently visited places.
Let me provide a brief overview about the state of Goa. Goa is the smallest state in India having an area of only 3702 sqkm. It is a multicultural state that has a Portuguese culture which is a vestige of the Portuguese rule 70 years ago. Over time Goa has changed drastically in terms of its residents and the types of occupations it offers. Goa is home to Christians, Hindus, Muslims and other religious belief system. Therefore it makes it an ideal location for Business that caters to needs of a variety of people. It is also good location for tourists who like to experience different cultures in a single place. Goa is one of the popular tourist destinations of India due to its vast heritage and location near the coast whose beaches attract tourists from India as well as from foreign countries.

Past few years Goa has seen an increase in the tourists visiting it and thus provides a great opportunity to businessmen to cater the needs of these tourists. However, not every location in Goa is ideal for business like resorts, hotels and restaurants. Businessmen from outside Goa try to get in their share of market place by starting their business here. The business location plays a vital role in determining the success of the business. Businesses poorly located can lead to significant loss for the business owner since it does not provide significant returns. A Chinese restaurant in a residential area will be less likely to gain customers than if it were located in a area which has entertainment and leisure services like movie theaters, shopping malls,theme parks etc. Indian culture favors homemade food and thus Goans are less likely to visit a restaurant unless they are visiting such entertainment and leisure areas.

The target audience of this analysis are business owners looking to start their business in Goa but do not know which is the ideal location for their business which would help them maximize their profits. One example of this is, where would a person start a restaurant such that he is sure enough people will visit it to sustain the business.

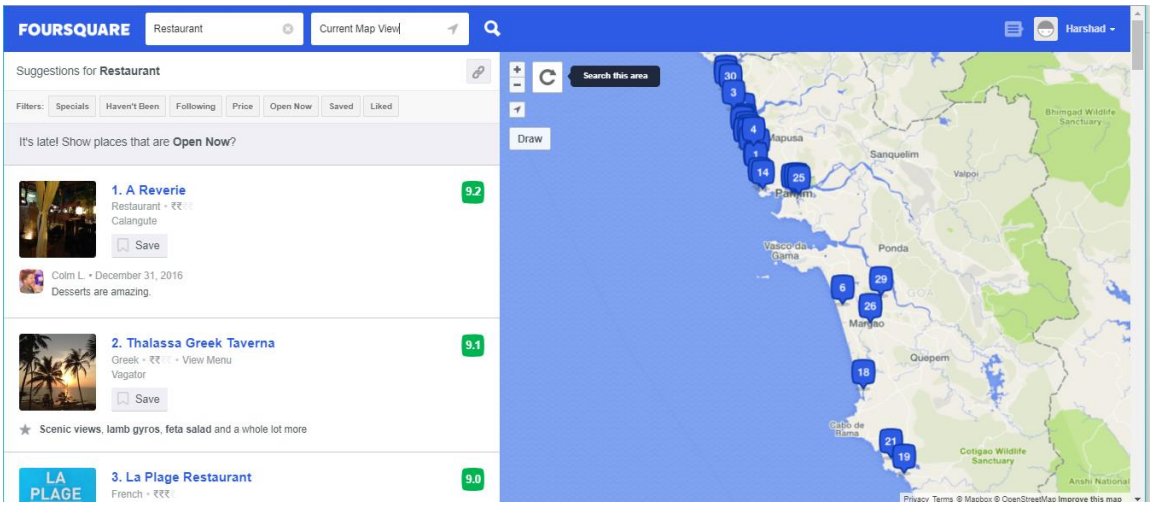
Data

To solve the above problem we will require location data of Goa. We will get this data from https://www.mapsofindia.com/pincode/india/goa/south-goa/ . Goa has 2 districts hence we will also need to get the data from https://www.mapsofindia.com/pincode/india/goa/north-goa/ to get the pincodes(postal codes) and the locations from north Goa. To retrieve this data we will use the request and BeautifulSoup library to fetch the html pages and then parse the data into respective dataframes. The final dataframe after retrieving data (using the BeautifulSoup library) from both webpages and combining will look as shown below

```
new_df.head(10)
```

	pincode	location	district
0	403801	A.P.dabolim	South Goa
1	403801	A.P.terminal	South Goa
2	403503	Advalpal	North Goa
3	403530	Advoi	North Goa
4	403401	Agapur Adpoi	South Goa
5	403512	Agarwada	North Goa
6	403702	Agonda	South Goa
7	403508	Aldona	North Goa
8	403521	Alto Betim	North Goa
9	403521	Alto-porvorim	North Goa

Next we will use the Nominatim function from the geopy library to get the coordinates of the the locations based on the location name. After getting the coordinates the foursquare API is used to retrieve the most visited locations for each of the towns/cities in the dataframe. We can then apply clustering algorithmns to segment the towns based on the frequently visited locations.



Methodology

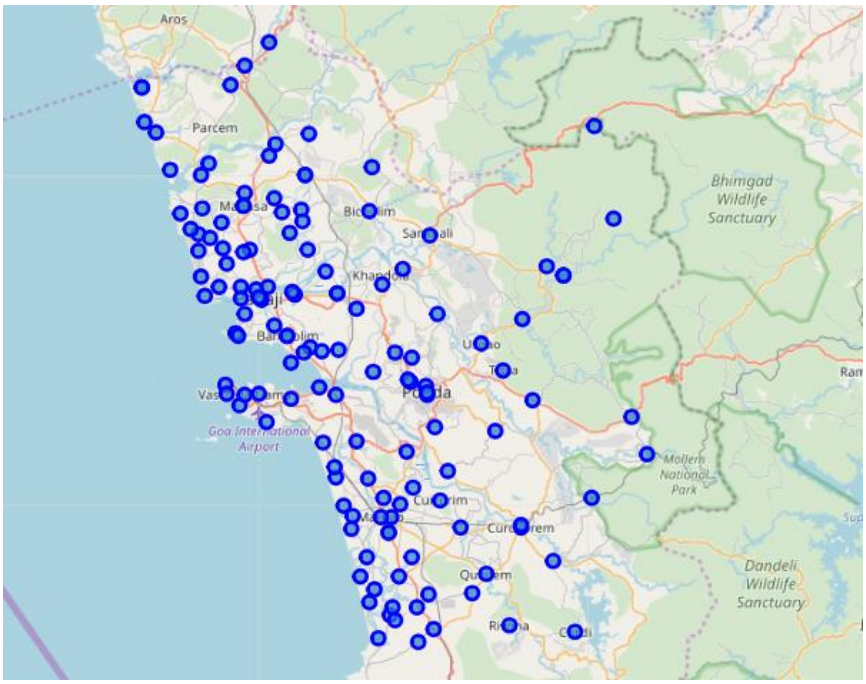
1. **Getting the coordinates of the towns of Goa.**

   We get the location coordinates of Goa using the Nominatim function from Geopy

   Some of the towns coordinates are not available. These are useless entries and we will remove them. Our final dataframe with the coordinates looks like:

pincode	location	district	latitude	longitude
403702	Agonda	South Goa	15.040267	73.990318
403508	Aldona	North Goa	15.588825	73.874076
403705	Ambaulim	South Goa	15.194188	74.055423
403723	Ambelim	South Goa	15.171813	73.968711
403107	Amona	South Goa	15.528204	73.981450
403509	Anjuna	North Goa	15.584865	73.743944
403601	Aquem	South Goa	15.272229	73.970029
403524	Arambol	North Goa	15.678065	73.705411
403516	Arpora	North Goa	15.563097	73.764037
403507	Assagao	North Goa	15.590412	73.768097

   Visualising the Goan locations



2. **Using Foursquare api to retrieve the venues**

   Using our api credentials – client id and key we define a function to retrieve the venues of the location coordinates in the dataframe.

location	pincode	location Latitude	location Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Agonda	14	14	14	14	14	14	14
Aldona	2	2	2	2	2	2	2
Anjuna	18	18	18	18	18	18	18
Aquem	6	6	6	6	6	6	6
Arambol	29	29	29	29	29	29	29
Arpora	17	17	17	17	17	17	17
Assagao	4	4	4	4	4	4	4
Assolna	1	1	1	1	1	1	1
Baga	9	9	9	9	9	9	9
Baina	2	2	2	2	2	2	2
Balli	1	1	1	1	1	1	1
Bandora	4	4	4	4	4	4	4
Batim	1	1	1	1	1	1	1
Benaulim	13	13	13	13	13	13	13
Betim	6	6	6	6	6	6	6
Betul	2	2	2	2	2	2	2
Bicholim	4	4	4	4	4	4	4
Bogmalo	7	7	7	7	7	7	7
Borim	1	1	1	1	1	1	1
Calangute	35	35	35	35	35	35	35
Camurlim	2	2	2	2	2	2	2
Candepar	26	26	26	26	26	26	26
Cansarvornem	3	3	3	3	3	3	3
Cansaulim	1	1	1	1	1	1	1
Carambolim	4	4	4	4	4	4	4

Next in order to work with these different types of locations we encode the categorical venues using the onehotencoder.

	location	pincode	ATM	Airport Terminal	American Restaurant	Arcade	Art Gallery	Art Museum	Asian Restaurant	BBQ Joint	Bakery	Bar	Beach	Beach Bar	Bed & Breakfast	Bistro	Boat or Ferry	Bookstore	Boutique	Breakfast Spot	Buff
0	Agonda	403702	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.071429	0.0	0.000000	0.071429	0.0	0.0	0.0	0.000000	(
1	Aldona	403508	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.000000	(
2	Anjuna	403509	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.055556	0.0	0.000000	0.000000	0.0	0.0	0.0	0.000000	(
3	Aquem	403601	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.000000	0.0	0.0	0.0	0.000000	(
4	Arambol	403524	0.0	0.0	0.0	0.0	0.0	0.0	0.068966	0.0	0.0	0.034483	0.000000	0.0	0.034483	0.000000	0.0	0.0	0.0	0.034483	(

We then group the data based on location , which will give us the frequency of occurances of the venue types per location. We will then use this grouped dataframe to retrieve the commonly visited locations. Since we need the most frequently visted locations we will sort these locations based on the frequency rates and get the top 5 most visited location of each town.

3. Clustering

Now that we have our most visited locations we perform clustering since we want to group locations with identical venue visit patterns . We use the K-means Clustering algorithm to cluster the locations into categories of

locations which will be profitable for a new restraurant and those that will likely result in a loss.

Visualizing the clusters

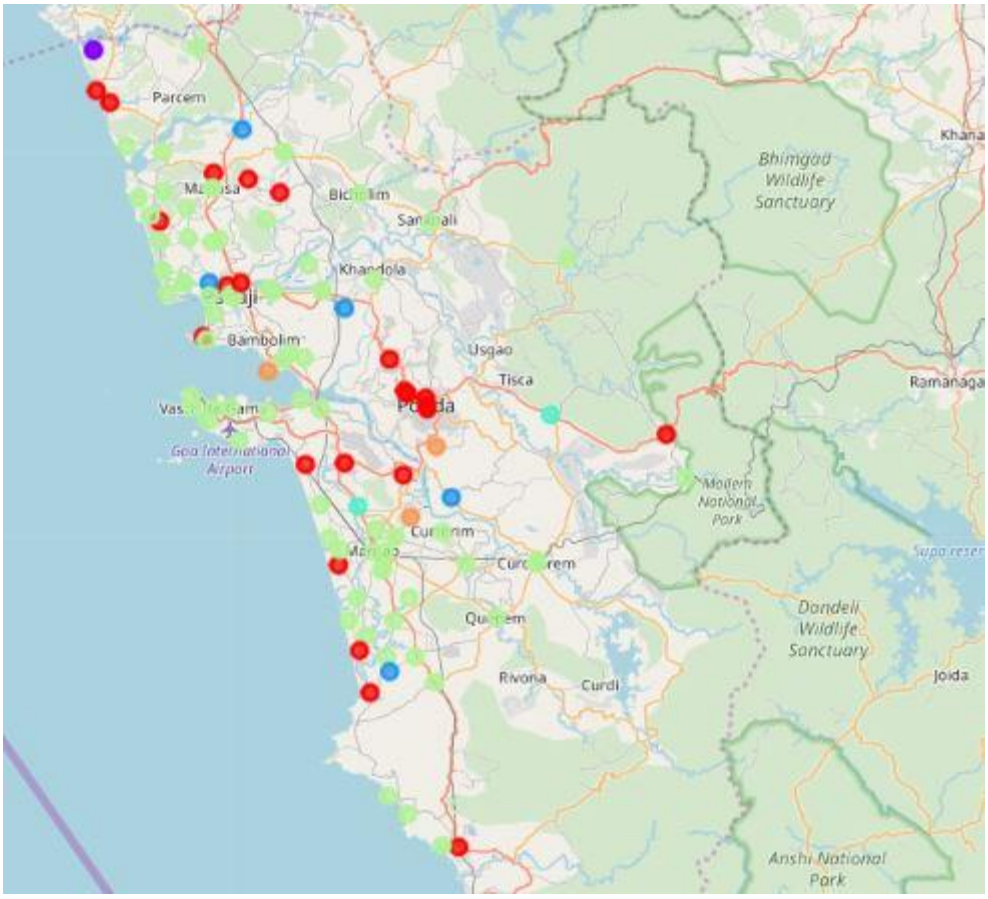
We now create a new dataframe by merging the venues dataframe with the Goa locations dataframe to get the coordinates of these cluster points.

```
goa_merged.head()
```

```
(113, 11)
```

	location	Cluster	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	pincode	district	latitude	longitude
0	Agonda	4	Indian Restaurant	Resort	Beach	Snack Place	Bistro	403702	South Goa	15.040267	73.990318
1	Aldona	0	Indian Restaurant	Market	Food Court	Food & Drink Shop	Food	403508	North Goa	15.588825	73.874076
2	Anjuna	4	Indian Restaurant	Café	Burger Joint	Italian Restaurant	Juice Bar	403509	North Goa	15.584865	73.743944
3	Aquem	4	Deli / Bodega	Train Station	Coffee Shop	Clothing Store	Multiplex	403601	South Goa	15.272229	73.970029
4	Arambol	0	Indian Restaurant	Resort	Restaurant	Asian Restaurant	Café	403524	North Goa	15.678065	73.705411

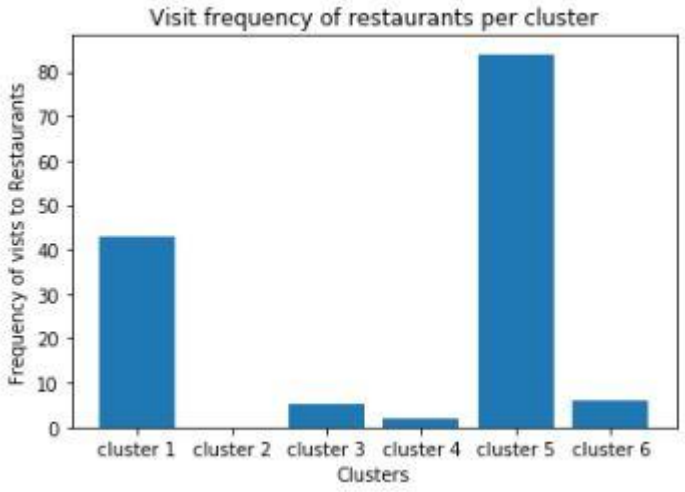
Now using this dataframe we visualize the clusters using the folium visualization library.



Results

We see that locations from cluster 1, cluster 6 and cluster 5 show a very good rate of visits to restaurants. Locations from these clusters are ideal to startup an restraurant since people are more likely to visit a new restaurant in these areas rather then the ones in cluster 2,3 and 4. These locations provide a

very high chance of profit and investors are less likely to suffer loss.



Locations from cluster 5 shows the highest frequency of visits to restaurants and these locations are likely to provide the most profit for an restaurant owner as people from these location show a high tendency to visit different restaurants and food places especially Indian Restaurants.

Additionally we can also infer that if someone wants to start an residency the locations from cluster 5 are ideal since these locations have the different day-to-day life services available and are suitable for tourists as well as people with a high class lifestyle who would like to have quick access to leisure services on day to day basis.

Discussion

Since location of a business is very crucial it is important to choose it wisely. We see that some regions favour restaurants whereas others do not.
This study could have been taken further by considering the tourist visit rates to these locations and the different types of families that live in Goa. For example, the Older section of the society prefers Goan restaurants whereas the younger generation likes the western food.
Since some locations were omitted due to unavailability of their coordinates, the results could have been better.

Conclusion

The study shows us that a few of the locations from Goa are not suitable for new restaurants.
If a person wishes to start a restaurant then it will much more profitable if he chooses an location from clusters 1 and 5. These locations would be the ones that I would recommend.