INTRODUCTION

A. BUSINESS UNDERSTANDING

Image: Victoria Memorial, a colonial structure

Kolkata (earlier known as Calcutta) was a colonial city under British rule. It later became the seat of Indian renaissance lead by people such as Swami Vivekananda and Raja Ram Mohan Roy during the 1800s. The city witnessed immense participation during the Indian freedom struggle. When India got independence, it became the state capital of West Bengal. The city, thus, has a combination of colonial and modern architecture. It is one of India's largest cities and one of its major ports. The city is centered on the east bank of the Hugli (Hooghly) River. There the port city developed as a point of trans-shipment from water to land and from river to sea. A city of commerce, transport, and manufacture, Kolkata is the dominant urban centre of eastern India.

Image: Howra Bridge

Foreign	2012	2013	2014
Domestic	22830205	25547300	49039890
Foreign	1218310	1245202	1375795
Total	**24048515**	**26792502**	**50415685**

Table: Tourism data in West Bengal, Tourism Survey Final Report, Gov. of India

Kolkata covers an area of 100km^2 and has a wide variety of cuisines to explore. Due to its proximity to China and South East Asia, different dishes from those regions are available. The city also contains western delicacies due to its colonial history.

Image: Kolkata Biriyani

Image: Chinese pasta

B. TARGET GROUP

This project will be helpful for visitors and foodies in the vibrant city of Kolkata whose numbers are increasing day by day. Using the clusters, people can effectively utilize their time and money. The clusters will tell you the best food venues and their cuisines, top 10 must-see places in Kolkata and other venues that you can visit.

DATA

Data sources used in this project include:

1. **Foursquare** website
2. **Trip Advisor** website
3. **Geo coder** data

The data from **Trip Advisor** website will be used to obtain the top places to visit in Kolkata. Method of data collection include web scraping using **Beautiful Soup** library in Python.

Foursquare API will be used to obtain the top 100 venues around the city of Kolkata. They will again be filtered and split into food venues and other venues. The food venues are the input to the API, to return the number of likes registered in each venue in foursquare website. Four categorical variables each will be created to cluster the data. One will be related to the quality of the venues and other to the cuisines available.

Geo coder in **GeoPy** library will be utilized to convert location address to geographical co-ordinates throughout the project. Since this project utilizes open sources, some geocoding didn't return any results. Thus those locations have been removed from the list.

METHODOLOGY

The top locations obtained from the **Trip Advisor** website and the other venues had some overlapping. Thus they were removed accordingly.

	name	Lat	Long
0	Victoria Memorial Hall	22.545080	88.342643
2	Mother House	22.553101	88.363662
3	Park Street	22.555159	88.350117
5	Howrah Bridge	22.585118	88.346744
6	Eden Gardens	22.564588	88.342290
7	Science City	22.539925	88.395810
8	Quest Mall	22.539027	88.365656
10	Prinsep Ghat	22.556573	88.331418
11	Birla Planetarium	22.545507	88.347318
12	New Market	22.560119	88.356735

Table: Top Locations

The venue data acquired from Foursquare had to be categorized into four each as mentioned above. The venues were split into four, using number of likes obtained as a measure of quality. The boundaries for the classification were obtained by **statistical analysis** of the 'likes' data.

```
count     66.000000
mean      23.318182
std       31.065271
min        1.000000
25%        8.000000
50%       13.000000
75%       21.750000
max      175.000000
Name: Likes, dtype: float64
```

Fig: Statistical description of the data

According to the analysis above, the 1st, 2nd and 3rd quartiles were used to define the Quality category. Thus the categories were defined as:

1. **Poor**
2. **Below average**
3. **Above average**
4. **Excellent**

Another classification was needed for the cuisines available in the city since our target group was assumed to be as diverse as possible. Thus after manually checking the type of restaurants available, they were grouped into 4:

1. **Indian**
2. **Other Asian**
3. **Western**
4. **Beverages**

After the data manipulation, one hot encoding was done to convert the categorical variables into numerical variables. The resulting data frame was used as input to the **K-means clustering** algorithm. The algorithm was so selected because of its proven reliability and popularity

	Name	Beverages	Indian	Other Asian	Western food	above avg	below avg	excellent	poor
0	The Blue Poppy	0	0	1	0	0	0	1	0
1	Jai Hind Dhaba	0	1	0	0	0	0	1	0
3	Balwant Singh's Eating House	0	1	0	0	0	0	1	0
4	Peter Cat	0	1	0	0	0	0	1	0
5	Oh! Calcutta	0	1	0	0	1	0	0	0

Table: One hold encoded variables

The number of clusters was limited to 4 for simplicity. This was because our categorization of venues was not large and did not warrant a large clustering of data points.

RESULT

The clusters were obtained and checked manually for properties.

Cluster 0:
Venues Categories: Indian and Western foods dominate
Quality:All of them are **below average**

Cluster 1:
Venues Categories: Indian foods dominate
Quality:All of them are **poor**

Cluster 2:
Venues Categories: Indian and Western foods dominate
Quality:All of them are **above average**

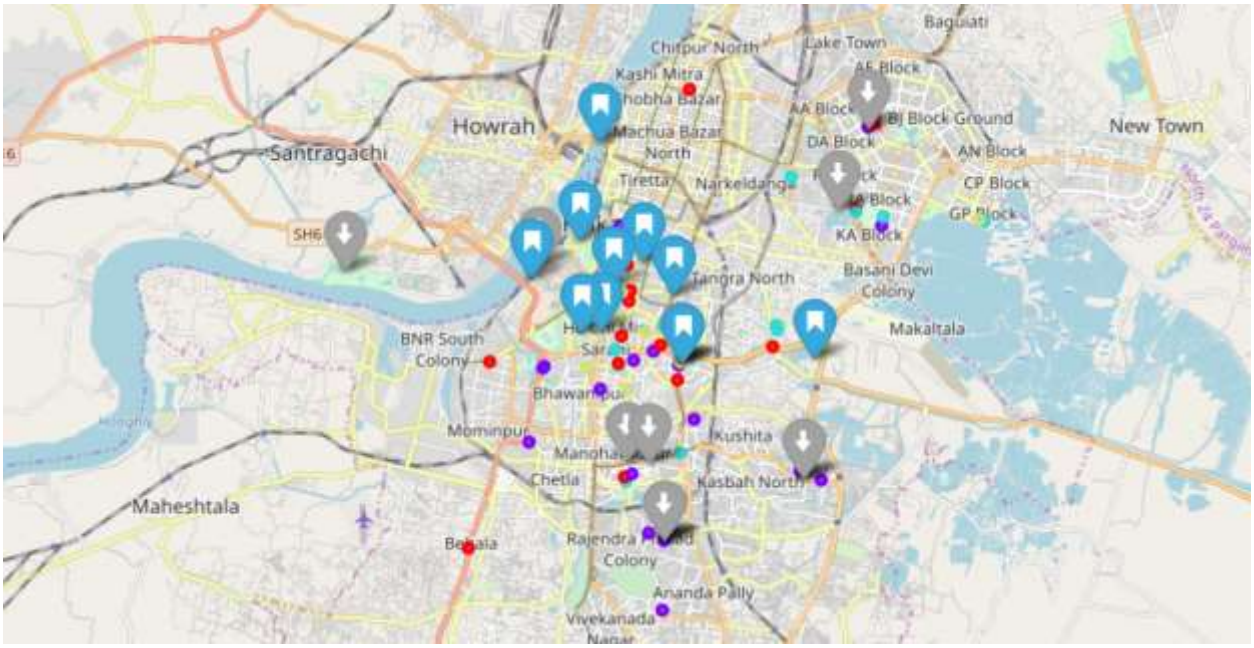
Cluster 3:
Venues Categories: Indian, Other Asian and Western foods dominate
Quality:All of them are **Excellent**

Label	Number of Venues
0	15
1	20
2	14
3	17

Table: Labels and their corresponding venues included

The clusters were divided on the basis of quality automatically. One thing to note is that Indian food dominated in all the categories and Beverages were present in all except the poor category.

Data visualization using Folium maps returned the following image:



Cluster	Color	Quality	Food Category
0	Red	Below Average	Indian and Western
1	Violet	Poor	Indian
2	Cyan	Above Average	Indian and Western
3	Yellow	Excellent	Indian, Other Asian and Western

Table: Color coding for the clusters

The **color coded circles** represents the four clusters

The **Blue Bookmark icon** shows the top must-see places in Kolkata

The **Gray Arrow icon** represents the venues other than the above categories. These include parks, shopping malls etc.

DISCUSSION

The map obtained through folium gave valuable insights into the location of the venues. The best places to explore will be the near to the **center of the city** and it included the best food venues as well as the must-see sights. Thus, I would recommend investing one's time around these areas.

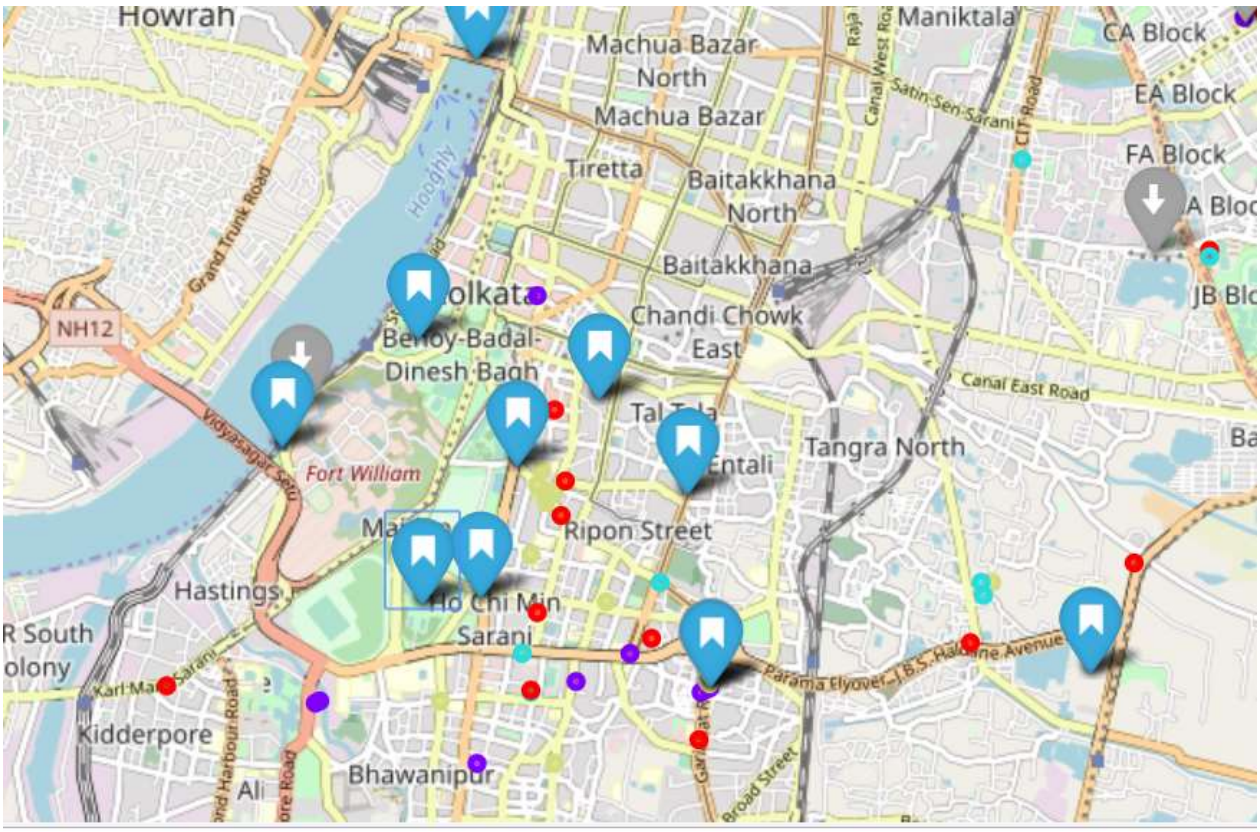


Fig: Yellow marked circles represent venues with excellent quality

name	categories	Likes	Quality
The Blue Poppy	Asian Restaurant	23	excellent
Jai Hind Dhaba	Dhaba	48	excellent
Balwant Singh's Eating House	Dhaba	74	excellent
Peter Cat	Indian Restaurant	175	excellent
Nocturne	Nightclub	36	excellent
Flurys	Bakery	151	excellent
6 Ballygunge Place	Bengali Restaurant	39	excellent
Bar-B-Q	BBQ Joint	100	excellent
Mocambo	Restaurant	80	excellent
Aqua	Lounge	24	excellent
Big Boss	Chinese Restaurant	22	excellent
The Irish House	Irish Pub	22	excellent
Chili's Grill & Bar	Tex-Mex Restaurant	40	excellent
Arsalan	Mughlai Restaurant	35	excellent
Olypub	Pub	54	excellent
Mainland China	Chinese Restaurant	41	excellent
Barbecue Nation	Indian Restaurant	40	excellent

Table: Best Food Venues in Kolkata

CONCLUSION

The project thus encourages the target group to go to the **cluster 3** where the best of what the city has to offer is within reachable distance. The cluster suggestions will have some short comings because of various factors such as the quality of data, availability of data, the assumptions involved etc. and there is room for improvement.