Kolkata - A Scope for Food Lovers and Businesses

Sayak Chakrabarti

April 16, 2020

1.Introduction

1.1 Background

Kolkata is a fabulous place to visit. Irrespective of seasons, the city attracts people from various parts of the country and the globe. One of the most famous attributes of its culture is the food. The diversity and vibrancy of food that one can get in Kolkata is unparallel anywhere in India.

From the visitor's point of view, it makes a lot of sense to look into reviews/ratings and price points and other attributes. Also, most visitors would prefer to stay in a centrally located spot in the city and best if the best restaurants are right across the corner. They select places of stay or dining as per their budget.

Therefore, restaurants/food-based service providers/hoteliers, who contribute to one of the larger sections of the economy in the city consider it important to utilise data for identifying locations or price points which can attract more footfall and therefore generate more profit for their businesses.

Also, it makes a lot of sense for aggregator-based-online-service providers to do either the above or to suggest new visitors with updated information on the same.

1.2 Problem

Past data involving location-ratings-price adds a lot of credence. This project analyses all of these and associated factors and hence aims to identify which places are a person more likely to visit and on what grounds ideally, which can be used by upcoming visitors as well as restaurateurs in their respective fields.

1.3 Interest

Evidently, visitors would be interested in selecting places with good ratings or proximity or pocket-friendly budgets. Restaurateurs would also like to avail such data to align their business models. Online aggregators might utilize such data for its clientele to make better decisions.

2. Data Acquisition and Cleaning

2.1 Data Sources

Data has been obtained from two APIs, Foursquare and Zomato Developers. The centre point of Kolkata was selected vide its latitude-longitude and then a radius of 4 Kilometres was searched for finding out venues using Foursquare API. Thereafter, the latitude-longitude values were used to fetch more details on the venues using Zomato API

2.2 Data Cleaning

Data was downloaded using two API sources as discussed above and finally combined into a single table. The initial data called via Foursquare API comprised of several attributes and was filtered to consider just the venue name-category-latitude-longitude and thereafter, the second set of data was called using Zomato API which is an aggregator for restaurants and food joints, to match the earlier latitude-longitude values and present somewhat-similar data.

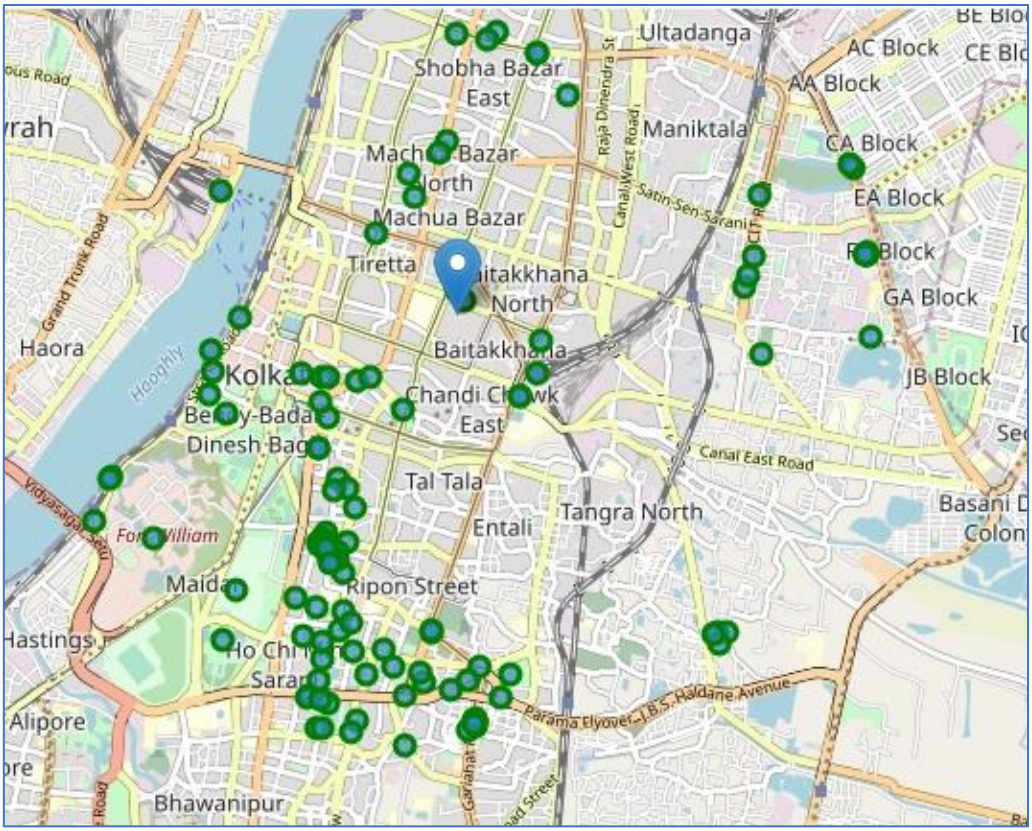

Fig1. Map showing list of venues as per Foursquare API

The second dataframe comprised of the venue name, latitude, longitude, price for two, price range, rating and address. Thereafter, both the dataframes were combined to one by matching latitude-longitude values.

2.3 Feature Selection

After data cleaning, a couple of erroneous events were observed and the same rectified sequentially

a. While most of the venues overlapped, some of them did not. Therefore, it was decided to drop all corresponding venues from the two datasets that had their latitude and longitude values different by more than 0.0004. This resulted in 59 rows of data.
b. Post this, it was found in the dataframe that
    i. There are venues like multiplexes which have restaurants inside them
    ii. Two locations are so close that they share the same latitude and longitude values
    iii. There were one or two duplicate data entries

    While it was ok to keep i, it did not make any sense to keep ii and iii and therefore, the corresponding rows were dropped from the dataframe.

c. A new column had to be introduced, called average price, which was an important attribute
d. Redundant columns like latitude, longitude, latitude and longitude differences, price for two were dropped. Name was also dropped as there was already a column called venue which sufficed. This resulted in 41 rows of data.
e. Ratings are an important feature of the dataset and therefore, rows with nil ratings were removed.

The final dataset therefore had 39 rows of data and 8 columns.

3. Exploratory Data Analysis

Now that the final dataset has been obtained, the data is analysed based on ratings and price of each venue. There are different ways of viewing the same

a. The top category types can be analysed.
b. Places where multiple venues are located can be analysed so that whenever the visitor goes to such a place, he/she can choose among multiple options
c. Venues can be analysed basis ratings and price points

Thereafter, the venues would be clustered depending on the availability of information of each venue. This would enable us to identify the target set of venues with certain attributes that can be recommended.

3.1 Count of Venues of Each Category

A bar chart was plotted between the venues of different categories vs their counts

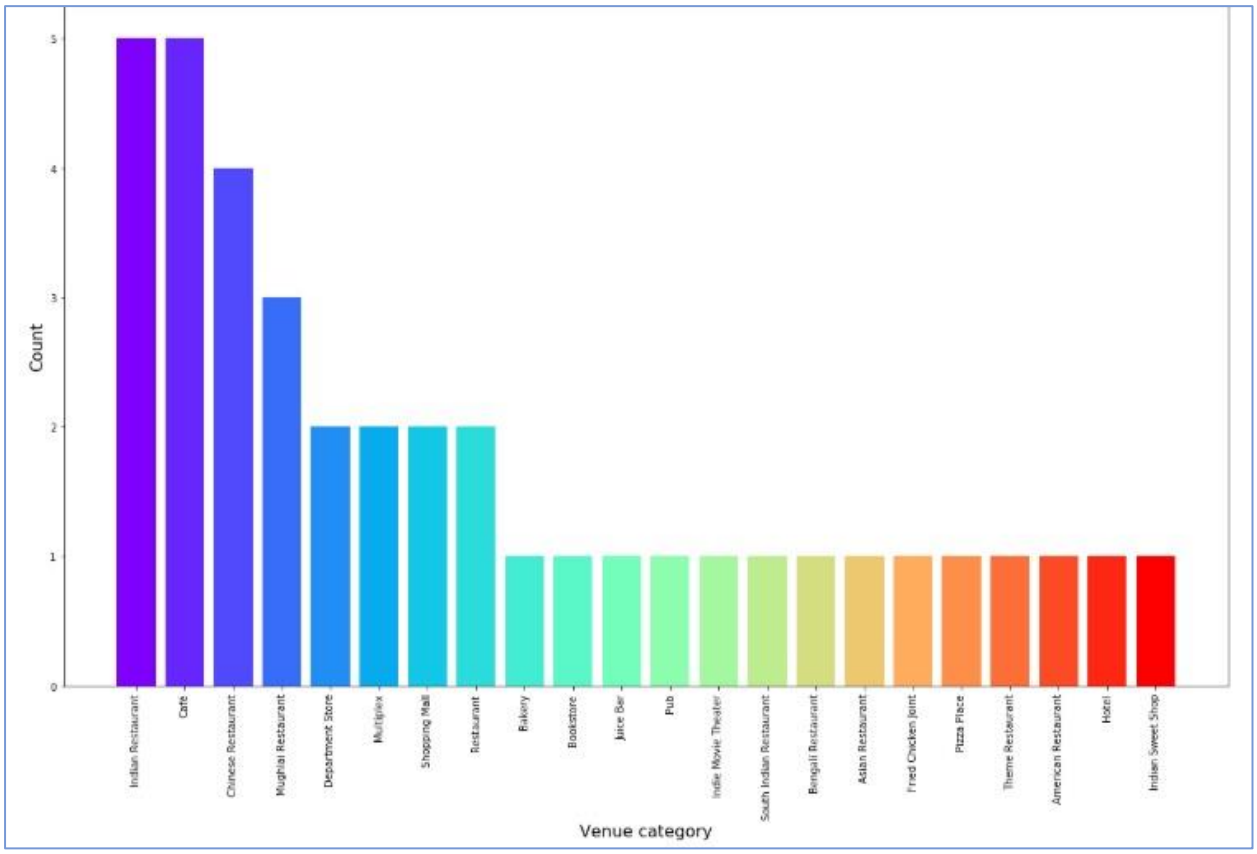

Fig2. Bar Plot for Venue Category vs Count

It was observed that the most frequent visits / popularities were amongst Indian restaurants, cafes and Chinese restaurants in the decreasing order of magnitude. Therefore, a tourist/visitor looking for these cuisines would be really lucky.

3.2 Count of Ratings

Whether a venue is to be tried out or not certainly depends on past reviews / ratings. Therefore, a bar plot was done between the ratings of different venues and their counts so as to visualize what could be the average ratings across venues.

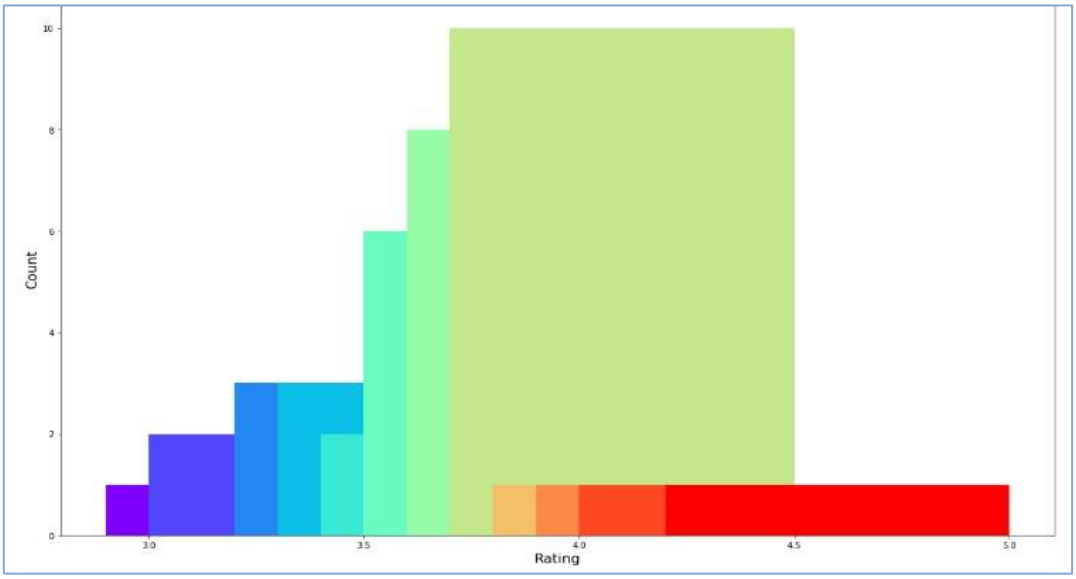

Fig3. Bar Plot for Ratings Vs Count

The ratings range from 2 to 5 on a scale of 1-5 and there is a substantial chunk that has been rated 5. Visitors might be interested to know the places with the highest ratings.

Fig4. Map showing Distribution of Ratings Basis Count

From the map, it was observed that most of the highest rated places in dark green are closer to AJC Bose Road – Camac Street – Park Street. Therefore, a visitor should definitely try out these areas.

3.3 Price Versus Venue

The price of a venue is perhaps one of the most important determinants of venue selection. Therefore, a scatter plot was constructed between average price points and the respective venue counts.

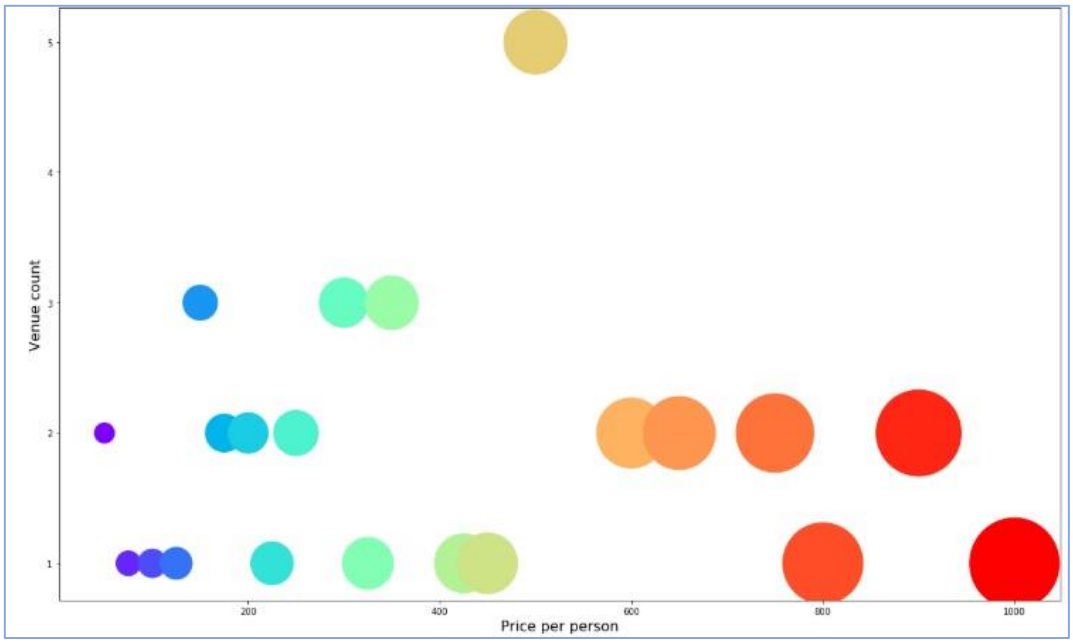

Fig5. Scatter Plot for Average Price Vs Venue Count

From the above plot, it was observed that the preferred price range for restaurants was in the range of Rs 200 to Rs 600. This means that a visitor with a pocket friendly budget can tend to find a lot of options in Kolkata.

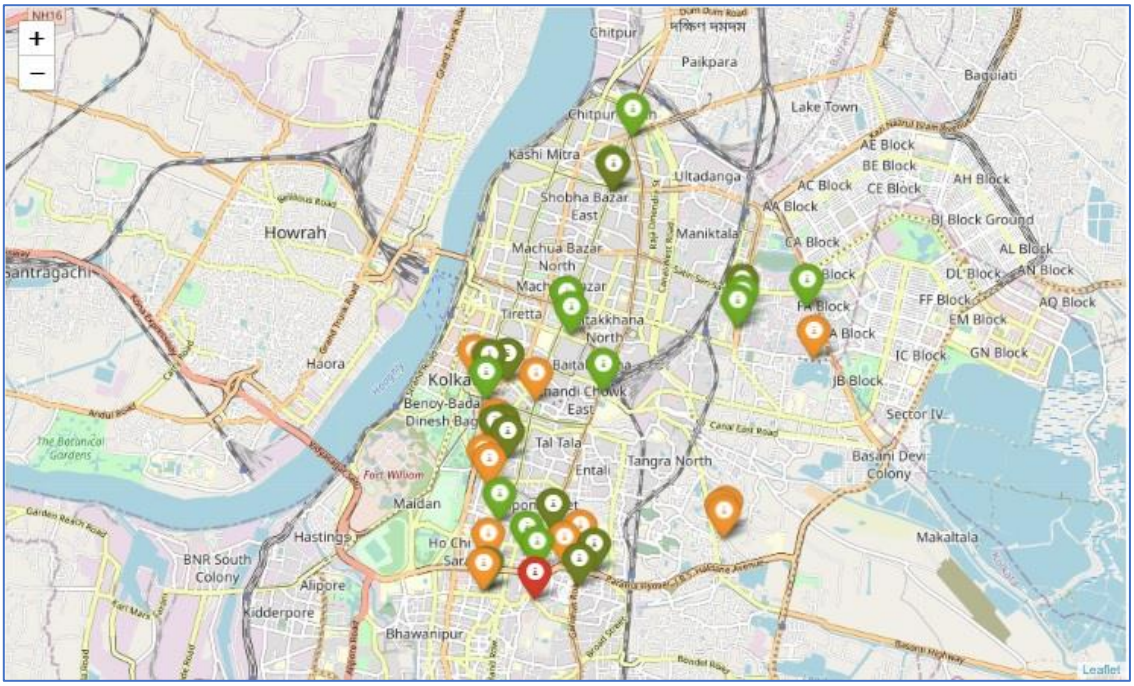

Fig6. Map showing Distribution of Venues Basis Price

It was observed from the above data that the earlier zone of AJC Bose Road – Camac Street – Park Street that was identified with the highest ratings also seem to have the most expensive restaurants on a relative basis. Therefore, a visitor might weigh the ratings versus price before making a judgement.

Also, restaurateurs might also analyse more as to why the price points are high in these areas. Maybe they are prime locations in the city and therefore attract exorbitant crowds but also are higher on the rental side and so a cost benefit analysis may be required before proceeding with business.

4. Predictive Modelling – Classification – Clustering

Predictive modelling is a technique in machine learning where-in the machine learns from a given set of attributes and past data and therefore either predicts the next possible scenario or classifies the data into different zones and tries analysing each separately.

Clustering is an unsupervised learning algorithm where we break the data into different zones or clusters on the basis that elements in a cluster are highly similar to each other and those in different clusters are highly different from each other.

4.1 Applying Clustering

Having done all the exploratory analyses, the venues were then clustered into 3 separate clusters keeping in mind the important attributes of price-location-ratings in order to identify similar venues and the relationship amongst them

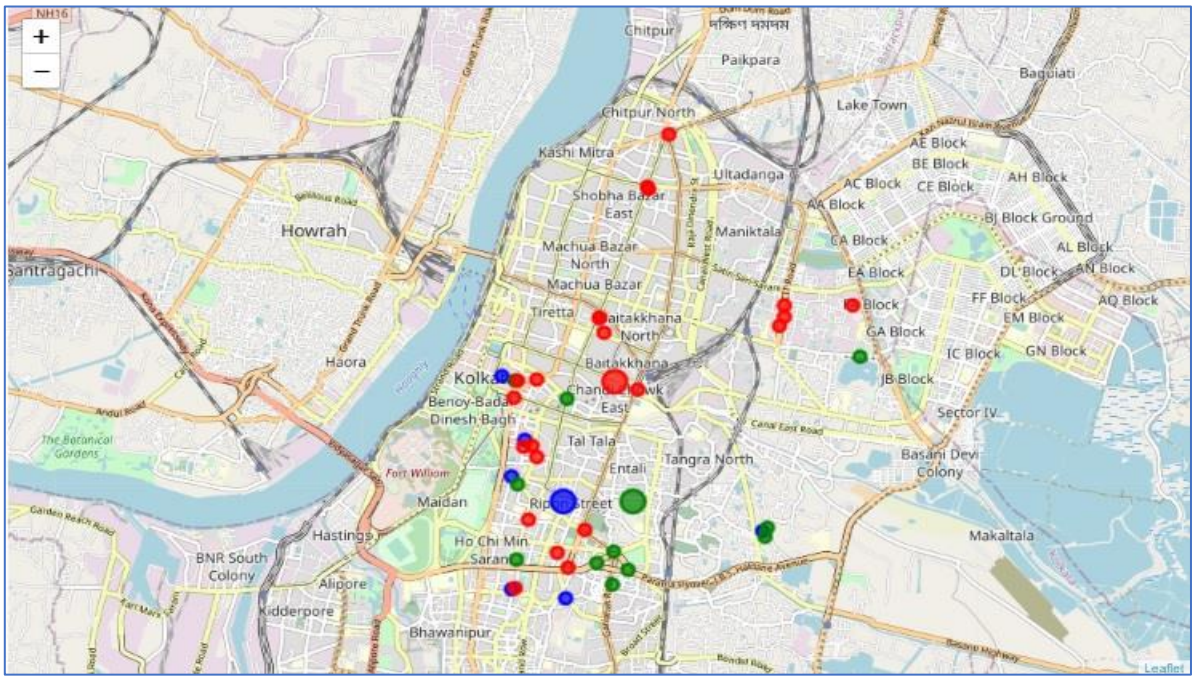

Fig7. Map showing 3 Different Clusters of Venues

Three Clusters were observed

a. **Cluster 0 (Green)** is sparsely spread to the boundaries of the city centre and includes some venues. These venues have **mean price range of 2.82 and rating spread around 3.97**.

b. **Cluster 1 (Red)** is evenly spread across the city. These venues have **mean price range of 1.41 and rating spread around 3.88**.

c. **Cluster 2 (Blue)** is sparsely spread to the boundaries of the city centre and includes some venues. These venues have **mean price range of 3.17 and rating spread around 4.08**.

5. Results and Future Directions

Basis the cluster analysis of venues within a certain reachable radius from the city centre, we found that the data could fit well amidst 3 clusters.

The data for Kolkata was collected using a combination of Foursquare API and Zomato API and was matched using latitude and longitude values being common attributes.

a. Majority of the venues were Indian restaurants and Cafes and Chinese restaurants. A visitor coming to Kolkata with a craving for good food would surely be lucky.
b. A substantial amount of venues had been rated 5 on 5 which strengthens the fact that there are ample restaurants offering good quality of food and service. Most of these restaurants were in the posh zone of AJC Bose Road – Camac Street – Park Street which is frequented by business travellers, tourists and party lovers.
c. It was observed from the above data that the earlier zone of AJC Bose Road – Camac Street – Park Street that was identified with the highest ratings also seem to have the most expensive restaurants on a relative basis. Therefore, a visitor might weigh the ratings versus price before making a judgement. Also, restaurateurs might also analyse more as to why the price points are high in these areas. Maybe they are prime locations in the city and therefore attract exorbitant crowds but also are higher on the rental side and so a cost benefit analysis may be required before proceeding with business.
d. **Cluster 0 (Green)** is dominated by Indian restaurants. These are moderately priced and moderately rated. So, someone who wishes to taste good Indian food but is not too finnicky about service parameters can try out Park Circus and adjoining areas.

e. **Cluster 1 (Red)** is spread across small fast food joints and bakeries and multi-cuisine restaurants which are pretty much pocket friendly but have a decent rating. So, someone who

wishes to stay pretty low on pockets but enjoy a good meal can flock to these places in and around Esplanade.

f. **Cluster 2 (Blue)** is dominated by Chinese restaurants which are few in number in the city but are highly rated and priced as well. So, someone who wishes to enjoy authentic Chinese can flock to these places, though there is not a single zone which has concentration of these restaurants in the city.

Apart from visitors, a restauranteur can utilise all the above information to suit his/her target audience and then plan as to what could be his mode of operation and where it wishes to operate.

An online aggregator company can use the above information to build up a good customer database by simply allowing the customer to browse venues in the city basis various search criteria like name or ratings or price.

6. Conclusion

The purpose of this project was to explore the places that a person visiting Kolkata could visit. It also intended to identify venues that could help a restauranteur or an online aggregator to capitalise on their businesses. The venues were identified using Foursquare and Zomato API and plotted on the map. The maps and the clusters revealed that a visitor could either visit Park Circus for good Indian food or Esplanade and adjacent for good street food or fast food and bakery and in selected areas in the city for authentic Chinese food. However, the preferred restaurants include the ones with low price and good taste of food.