Top 10 Places to Visit in Kolkata– Capstone project

By ABHISHEK SINHA

IT Engineer, UK

Table of Contents

- Introduction
- Business Problem Statement
- Data
    - Neighborhoods
    - Geocoding
    - Venue Data
- Methodology
    - Accuracy of Geocoding API
    - Folium
    - One hot Coding
    - Top 10 most common venues
    - Optimum Number of Clusters
    - K-Means Clustering
- Results
- Discussion
- Conclusion

Introduction

Kolkata, also known as Calcutta, is the capital of the Indian state of West Bengal. it is the seventh most populous city in India; the city had a population of 4.5 million, while the suburb population brought the total to 14.1 million, making it the third-most populous metropolitan area in India.

Kolkata see a heavy influx of tourist throughout the year due to its diversity in Food, culture, language, art and heritage buildings and museums.

The objective of this project is to provide best venues or location to visit, along with important places like banks., ATMs, Hospitals etc

Business Problem Statement

1

Kolkata is a big city spanning over several kilometres in radius. Being one of the oldest city Kolkata has a complex system of roads and intersections.

2

As this is not a planned city often tourists find themselves lost now and then if one wrong turn is made while travelling within in the city.

3

It’s very difficult to roam around the city without a guide or without taking help from local people asking about where what lies.

Data

Neighbourhoods:

The data of the neighbourhoods in Kolkata can be extracted out by web scraping using BeautifulSoup library for Python. The neighbourhood data is scraped from a Wikipedia webpage.

Geocoding

The file contents from ***kolkata.csv*** is retrieved into a Pandas DataFrame. The latitude and longitude of the neighbourhoods are retrieved using Google Maps Geocoding API. The geometric location values are then stored into the initial dataframe.

Venue Data

From the location data obtained after Web Scraping and Geocoding, the venue data is found out by passing in the required parameters to the **FourSquare** API and creating another DataFrame to contain all the venue details along with the respective neighbourhoods.

Methodology

Accuracy of the Geocoding API

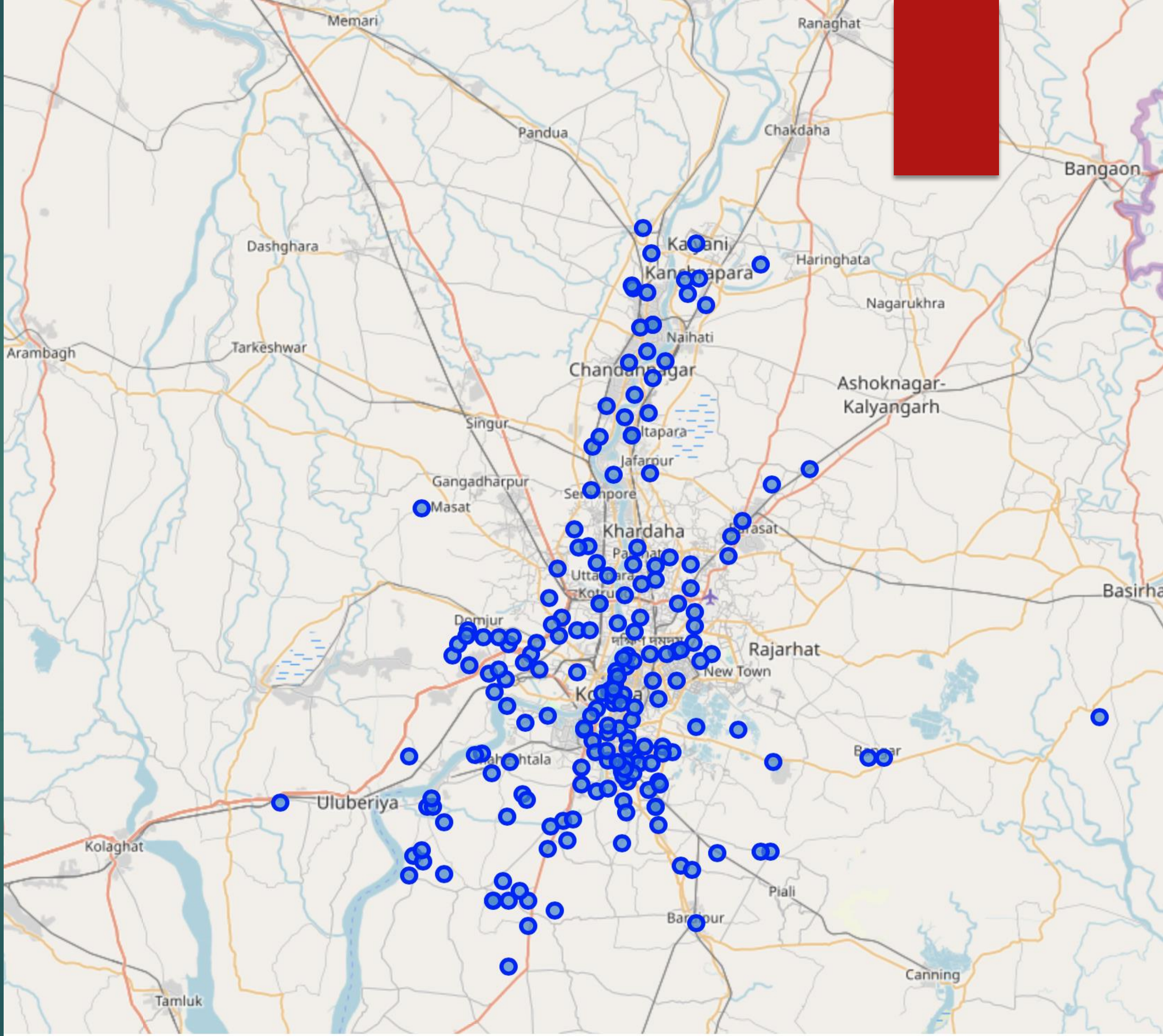
In the initial development phase with OpenCage Geocoder API, the number of erroneous results were of an appreciable amount, which led to the development of an algorithm to analyse the accuracy of the Geocoding API used.

In the algorithm developed, Geocoding API from various providers were tested, and in the end, Google Maps Geocoder API turned out to have the least number of collisions (errors) in our analysis.

Folium

Folium builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of the **leaflet.js** library. All cluster visualization are done with help of Folium which in turn generates a Leaflet map made using OpenStreetMap technology.

Methodology (map)



Methodology contd.

One hot encoding

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. For the K-means Clustering Algorithm, all unique items under Venue Category are one-hot encoded.

Top 10 most common venues

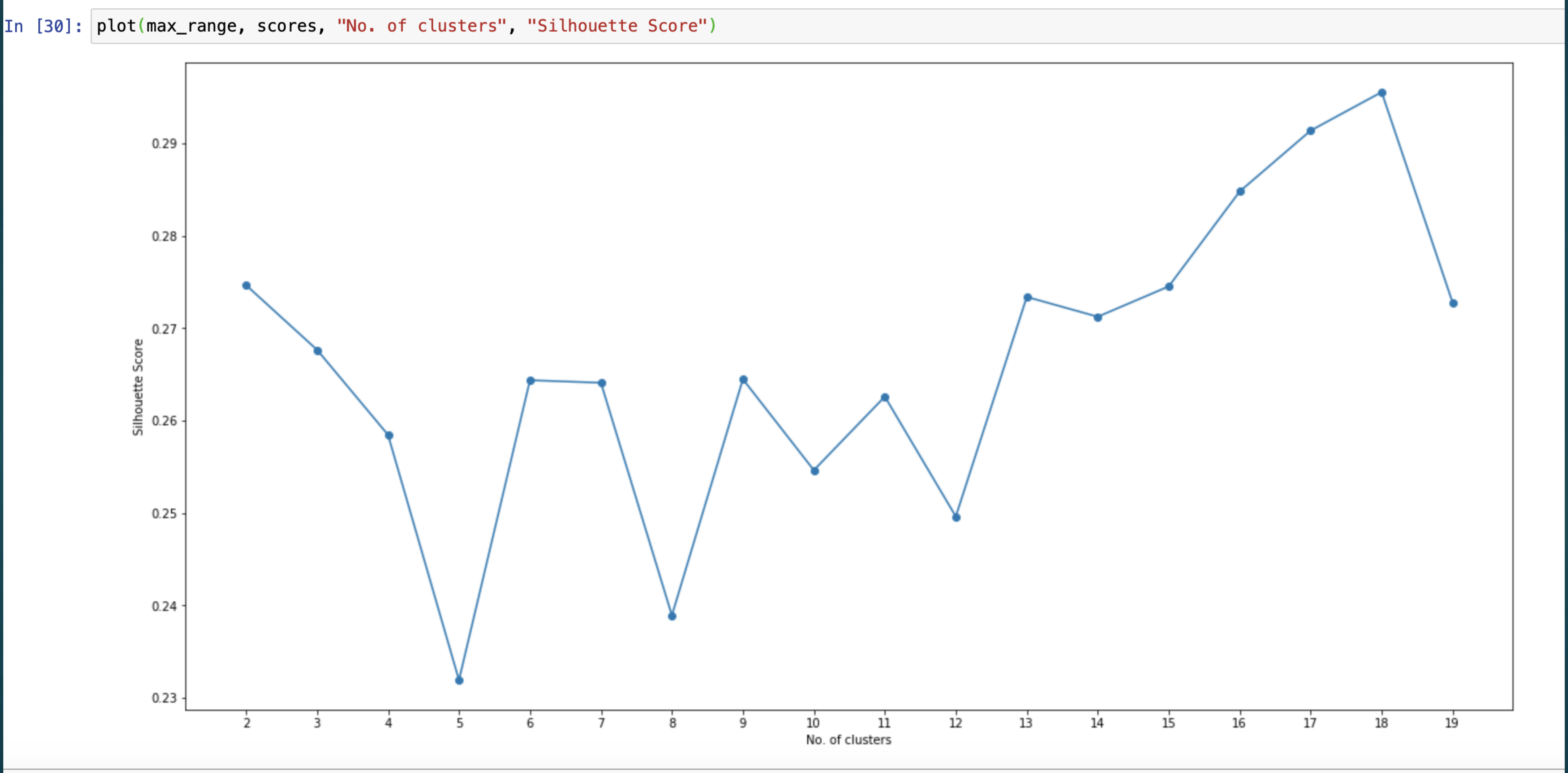
Due to high variety in the venues, only the top 10 common venues are selected and a new DataFrame is made, which is used to train the K-means Clustering Algorithm.

Optimal number of clusters

Silhouette Score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

Based on the Silhouette Score of various clusters below 20, the optimal cluster size is determined.

Methodology (optimum cluster size)



```
In [30]: plot(max_range, scores, "No. of clusters", "Silhouette Score")
```

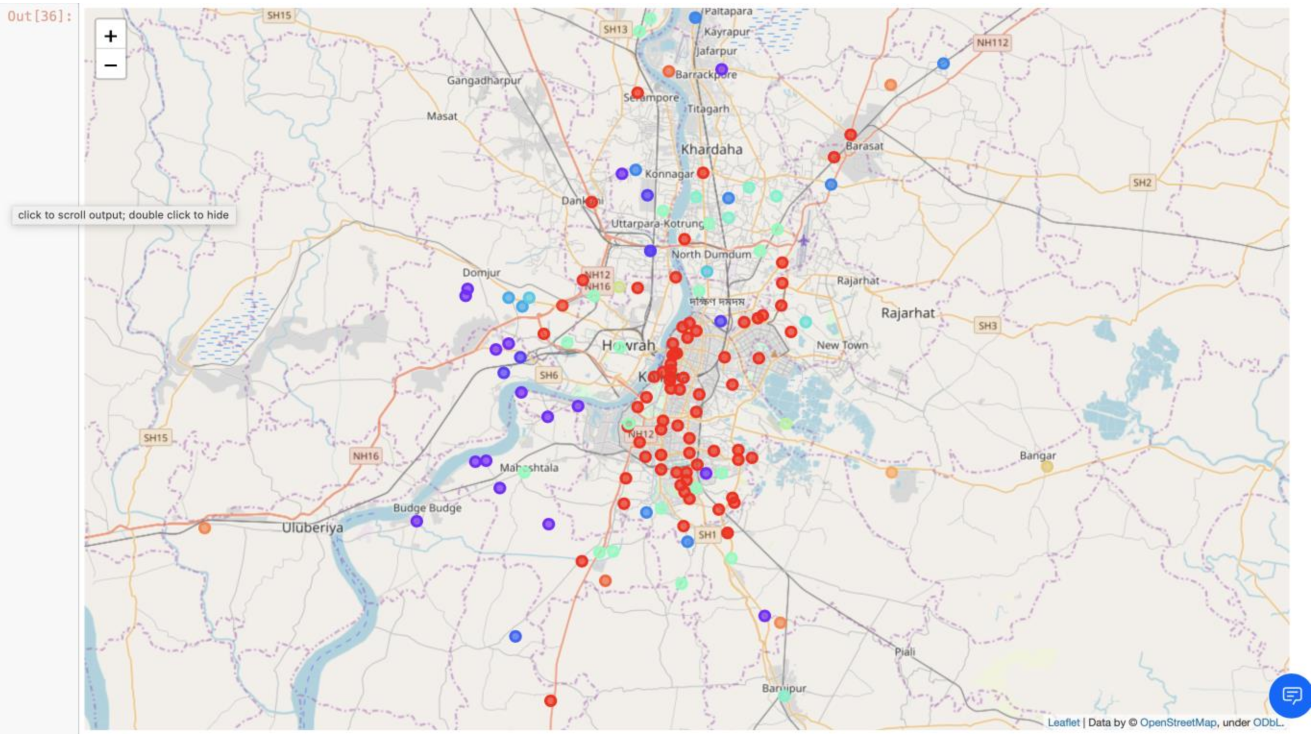
Methodology contd.

K-means clustering

The venue data is then trained using K-means Clustering Algorithm to get the desired clusters to base the analysis on. K-means was chosen as the variables (Venue Categories) are huge, and in such situations K-means will be computationally faster than other clustering algorithms.

Results

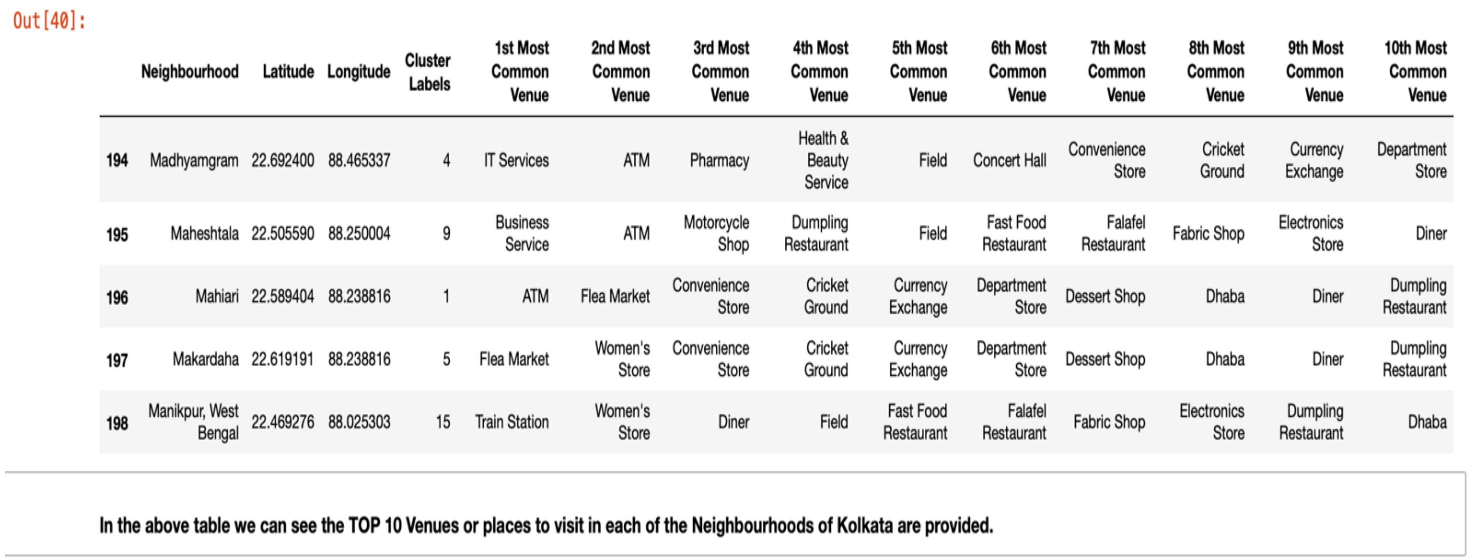
The neighbourhoods are divided into n clusters where n=18, is the number of clusters found using the optimal approach . The clustered neighbourhoods are visualized using different colours so as to make them distinguishable.



Discussion

After analysing the various clusters produced by the Machine learning algorithm, **cluster no.18,** is a prime fit to solving the problem of finding a cluster with common venue mentioned.

Kolkata is a big city and this derivation will help tourist in visiting the famous places in their neighbourhood along with providing information on the what's around the corner.



Out[40]:

	Neighbourhood	Latitude	Longitude	Cluster Labels	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
194	Madhyamgram	22.692400	88.465337	4	IT Services	ATM	Pharmacy	Health & Beauty Service	Field	Concert Hall	Convenience Store	Cricket Ground	Currency Exchange	Department Store
195	Maheshtala	22.505590	88.250004	9	Business Service	ATM	Motorcycle Shop	Dumpling Restaurant	Field	Fast Food Restaurant	Falafel Restaurant	Fabric Shop	Electronics Store	Diner
196	Mahiari	22.589404	88.238816	1	ATM	Flea Market	Convenience Store	Cricket Ground	Currency Exchange	Department Store	Dessert Shop	Dhaba	Diner	Dumpling Restaurant
197	Makardaha	22.619191	88.238816	5	Flea Market	Women's Store	Convenience Store	Cricket Ground	Currency Exchange	Department Store	Dessert Shop	Dhaba	Diner	Dumpling Restaurant
198	Manikpur, West Bengal	22.469276	88.025303	15	Train Station	Women's Store	Diner	Field	Fast Food Restaurant	Falafel Restaurant	Fabric Shop	Electronics Store	Dumpling Restaurant	Dhaba

In the above table we can see the TOP 10 Venues or places to visit in each of the Neighbourhoods of Kolkata are provided.

Conclusion

This is a self-guide for people about each Neighbourhood to find the most popular places to visit, to eat, for shopping, banks, ATMs etc.

This will also help the boosting the local economy as tourists will go to the famous places around their area of visit. This will also increase the competition among the vendors dealing in similar items as the better their services and ratings the better chance they have in appearing in **TOP 10** list.