

The Battle of Neighbourhood | Best possible localities for buying or renting a property in Kolkata

ANIRBAN PAUL

25-07-2021

Table of Contents

A. Introduction/Business Problem Statement	1
A.1 Target Audience:.....	2
B. Data Section	2
B.1 The Location	2
B.2 Location Data.....	2
B.3 Intertwining Foursquare API	2
C. Methodology	3
C.1 Workflow.....	3
C.2 Clustering Approach.....	5
C.3 Libraries used.....	5
D. Result	5
E. Discussion	6
F. Conclusion.....	7
F.1 Future Works:.....	7

A.Introduction/Business Problem Statement

Kolkata, the City of Joy and the capital of West Bengal is one of the largest metropolises in the world where over 15 million people live, and it has a population density of 2.813 people per square kilometre

The purpose of this project is to assist people in exploring better facilities around their neighborhood. it'll help people making smart and efficient decisions on selecting great neighborhoods out of variety of other neighborhoods in Kolkata, West Bengal.

Lots of people come to Kolkata for employment or educational purposes and needed much research for good housing prices and reputed schools for his or her children. This project is for those people that are trying to find better neighborhoods. For simple accessing to Cafe, School, Supermarket, medical shops, grocery shops, mall, theatre, hospital, likeminded people, etc.

It will help people to urge the notice of the world and neighborhood before moving to a new and unknown city, state, country, or place for his or her work or to start out a new fresh life.

A.1 Target Audience:

The major purpose of this project is to suggest a far better neighborhood/place for a new city for the one who is shifting there. The report tries to focus on the following things:

- Social presence in society in terms of likeminded people.
- Connectivity to the airport, bus stand, city centre, markets, and other daily needs things nearby.
- Sorted list of homes in terms of housing prices in an ascending or descending order Sorted list of faculties in terms of location, fees, rating, and reviews

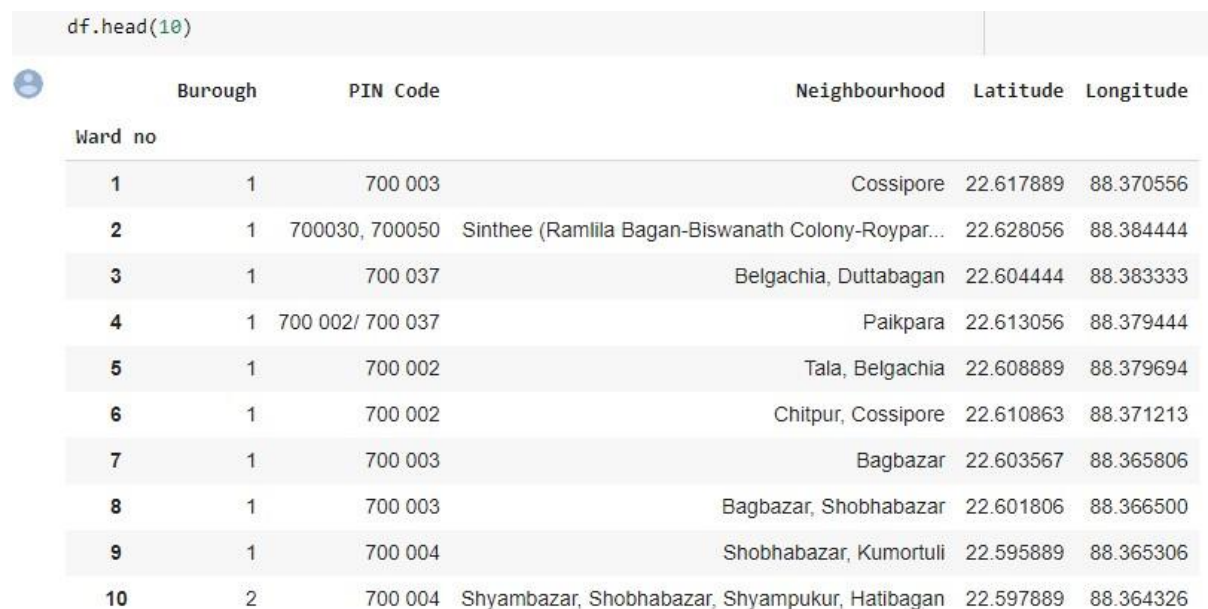
B. Data Section

B.1 The Location

Kolkata may be a popular destination for brand spanking new immigrants in West Bengal to reside. Being home to varied religious groups and places of worship. Although immigration has become a hot topic over the past few years with more governments seeking more restrictions on immigrants and refugees, the overall trend of immigration into Kolkata has been on the increase.

B.2 Location Data

The location data of Kolkata postal code and ward wise is not easily available and we have to use web scraping to get the data in required format by iterating 144 webpages ([Kolkata wards](#)). This dataset contains 6 columns (Ward no, Borough, Pin Code, Neighbourhood, Latitude and Longitude)



```
df.head(10)
```

	Borough	PIN Code	Neighbourhood	Latitude	Longitude
Ward no					
1	1	700 003	Cossipore	22.617889	88.370556
2	1	700030, 700050	Sinthee (Ramlila Bagan-Biswanath Colony-Roypar...	22.628056	88.384444
3	1	700 037	Belgachia, Duttabagan	22.604444	88.383333
4	1	700 002/ 700 037	Paikpara	22.613056	88.379444
5	1	700 002	Tala, Belgachia	22.608889	88.379694
6	1	700 002	Chitpur, Cossipore	22.610863	88.371213
7	1	700 003	Bagbazar	22.603567	88.365806
8	1	700 003	Bagbazar, Shobhabazar	22.601806	88.366500
9	1	700 004	Shobhabazar, Kumortuli	22.595889	88.365306
10	2	700 004	Shyambazar, Shobhabazar, Shyampukur, Hatibagan	22.597889	88.364326

Fig 1. Data collected via web scrapping

B.3 Intertwining Foursquare API

This project would use Foursquare API because it's prime data gathering source as it features a database of many places, especially their places API which provides the power to perform location search, location sharing, and details a few businesses.

We will need data about different venues in different neighbourhoods of that specific borough. In order to gain that information, we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of neighbourhoods, we then connect to the Foursquare API to gather information about venues inside each neighbourhood. For each neighbourhood, we have chosen the radius to be 2000 meter.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Neighbourhood
2. Neighbourhood Latitude
3. Neighbourhood Longitude
4. Venue
5. Name of the venue e.g., the name of a store or restaurant
6. Venue Latitude
7. Venue Longitude
8. Venue Category

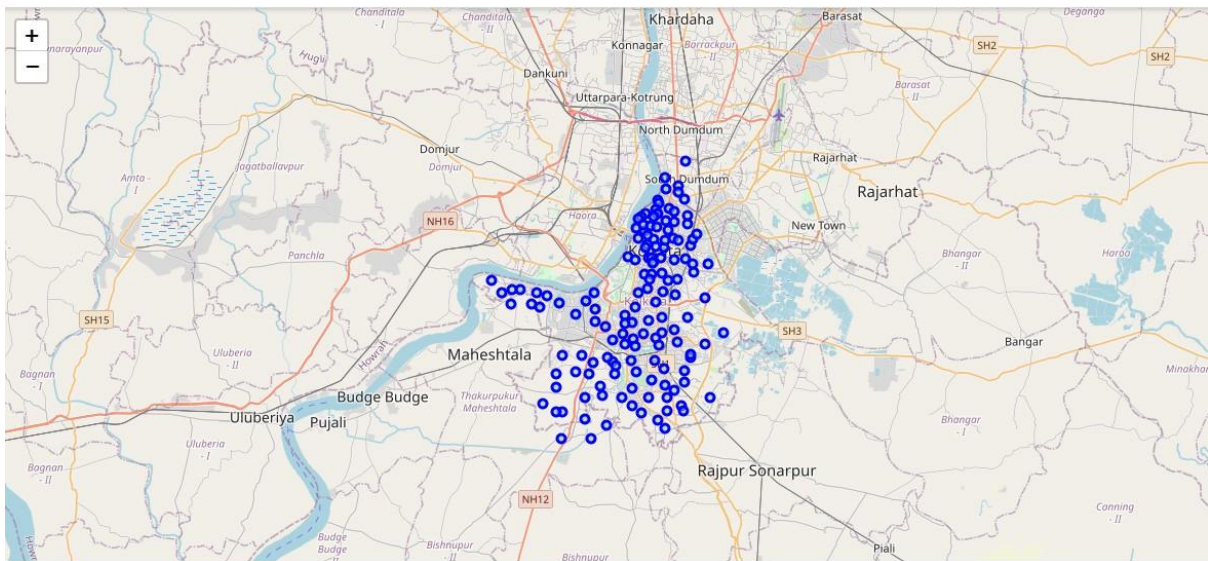


Fig 2. Folium Map of wards in Kolkata

C. Methodology

C.1 Workflow

C.1.a Data Collection:

Data is collected via two methods: A) Web scraping B) Using Foursquare API

A) The location data is collected by web scraping 144 Wikipedia webpages in tabular format containing 6 columns (Ward no, Borough, Pin Code, Neighbourhood, Latitude and Longitude)

B) Using credentials of Foursquare API features of nearby places of the neighbourhoods would be mined. Due to HTTP request limitations, the number of places per neighbourhood parameter would reasonably be set to 100 and therefore the radius parameter would be set to 2000 which results 132 unique venues for this study.

C.1.b Data Processing:

I used python folium library to visualize geographic details of Kolkata and its boroughs, and I created a map of Kolkata with boroughs superimposed on top.

We have some common venue categories in boroughs. In this reason I used unsupervised learning K-means algorithm to cluster the boroughs. K-Means algorithm is one of the most popular clustering method of unsupervised learning.

The optimal K value is determined using the elbow point method (i.e., K=5) and using that the clusters are formed accordingly.

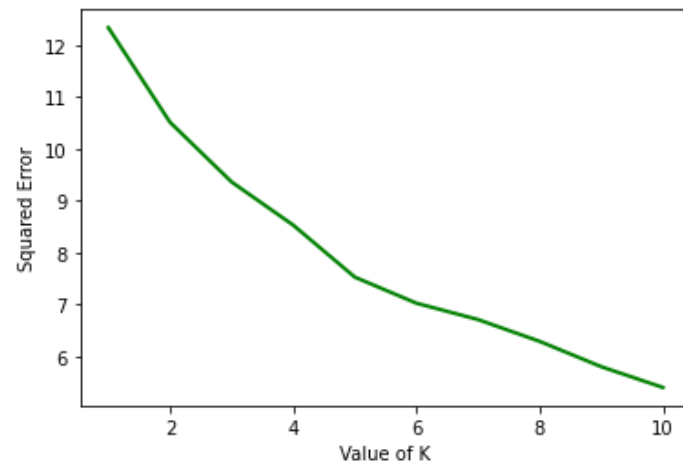


Fig 3. Using Elbow point method to determine the value of K

After that by creating dummy data using one hot encoding and merging clustered label data with the data-frame we got the 10 most common venues ranked in ascending order for each ward.

	Neighborhood	Zoo	ATM	American Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Awadhi Restaurant	BBQ Joint	Bakery	Bank	Bar	Beer Bar	Beer Garden	Bengali Restaurant	Bistro
0	Alipore	0.035714	0.000000	0.0	0.0	0.0	0.0	0.035714	0.035714	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0
1	Ashok Nagar, Kudghat, Tollygunge Club, Regent	0.000000	0.142857	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0
2	Badartala, Rajabagan	0.000000	0.750000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.250000	0.0	0.0	0.0	0.0	0.000000	0.0
3	Bagbazar	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.071429	0.0
4	Bagbazar, Shobhabazar	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.071429	0.0
...
37	Tollygunge Circular Road (Sirity-Senhati Colony)	0.000000	0.166667	0.0	0.0	0.0	0.0	0.083333	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0
38	Ultadanga (Daspara-Muchi Bazar-Telenga)	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.040000	0.0	0.0	0.0	0.0	0.040000	0.0

Fig 4. Using K-Means Clustering and One-Hot Encoding for getting the mean of frequency

C.2 Clustering Approach

To compare the similarities and dissimilarities of two postal code neighborhoods, we decided to segment them and group them into clusters to find similar neighborhoods in a big city like Kolkata. To be able to perform that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm.

C.3 Libraries used

- Pandas: For creating and manipulating data frames.
- Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.
- Scikit Learn: For importing k-means clustering.
- JSON: Library to handle JSON files.
- XML: To separate data from presentation and XML stores data in plain text format.
- Geocoder: To retrieve Location Data.
- Beautiful Soup and Requests: To scrap and library to handle http requests.
- Matplotlib: Python Plotting Module.

D.Result

The resulting master table looks like this which has 10 most common venues ranked in ascending order for each ward.

	Borough	PIN Code	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
Ward no														
1	1	700 003	Cossipore	22.617889	88.370556	0	Plaza	Playground	River	Bank	Indian Restaurant	Vegetarian / Vegan Restaurant	Train Station	Metro Station
2	1	700030, 700050	Sinthee (Ramilla Bagan-Biswanath Colony-Roypar...	22.628056	88.384444	0	Plaza	Gift Shop	Playground	Fast Food Restaurant	Burger Joint	Liquor Store	Park	Music Store
3	1	700 037	Belgachia, Duttabagan	22.604444	88.383333	0	Indian Restaurant	Train Station	Indian Sweet Shop	Indie Movie Theater	Market	Asian Restaurant	Pizza Place	Bengali Restaurant
4	1	700 002/ 700 037	Paikpara	22.613056	88.379444	0	Plaza	Train Station	Metro Station	Platform	Bakery	Playground	Indian Restaurant	Diner
5	1	700 002	Tala, Belgachia	22.608889	88.379694	0	Train Station	Indian Restaurant	Asian Restaurant	Bengali Restaurant	Plaza	Playground	Platform	Metro Station

Fig 5. Most Common Venues

And finally, the clustered data are superimposed on the Kolkata Map with the clusters labelled as colour coding

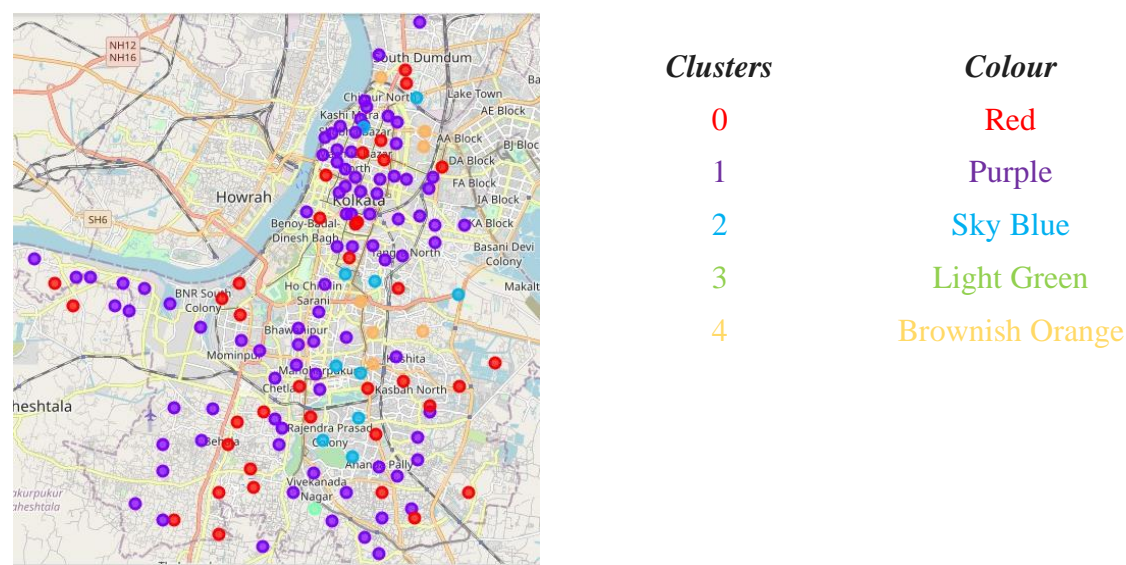


Fig 6. Map of Clusters in Kolkata

E. Discussion

Analysing each cluster and calculating the statistical mode of from the list of first most common venue and the corresponding count of ward we can put that data in tabular format and from that we can make inference for each cluster.

Analysing Cluster 0

▶

⬆

⬇

🔍

⚙

📄

🗑

⋮

kolkata_merged.loc[kolkata_merged['Cluster Labels'] == 0]

⬆

	Borough	PIN Code	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
Ward no														
1	1	700 003	Cossipore	22.617889	88.370556	0	Plaza	Playground	River	Bank	Indian Restaurant	Vegetarian / Vegan Restaurant	Train Station	
2	1	700030, 700050	Sinthee (Ramliia Bagan-Biswanath Colony-Roypar...	22.628056	88.384444	0	Plaza	Gift Shop	Playground	Fast Food Restaurant	Burger Joint	Liquor Store	Park	
3	1	700 037	Belgachia, Duttabagan	22.604444	88.383333	0	Indian Restaurant	Train Station	Indian Sweet Shop	Indie Movie Theater	Market	Asian Restaurant	Pizza Place	Rest
		700												

Fig 7. Cluster Analysis

	MostCommonVenue	Count
Cluster no		
Cluster 0	Metro Station	5
Cluster 1	Indian Restaurant	14
Cluster 2	Café	42
Cluster 3	Mughlai Restaurant	4
Cluster 4	ATM	6

Fig 8. Inference from the cluster analysis

F. Conclusion

In this project, using k-means cluster algorithm I separated the neighbourhood into 5 different clusters and for 144 different latitude and longitude from dataset, which have very-similar neighbourhoods around them. Using the charts above results presented to a particular neighbourhood based on average house prices and school rating have been made.

I feel rewarded with the efforts and believe this course with all the topics covered is well worthy of appreciation. This project has shown me a practical application to resolve a real situation that has impacting personal and financial impact using Data Science tools. The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision better with confidence.

F.1 Future Works:

This project can be continued for making it more precise in terms to find best house in Kolkata. Best means based on all required things (daily needs or things we need to live a better life around and in terms of cost effective. But there is not enough data about schools and hospitals in foursquare API for the location Kolkata. So, thinking to use google maps API for getting the best results.