Applied Data Science—Clustering Localities in Kolkata, India

Arghyadeep Ganguly - 29/01/2020

Introduction/Business Problem

In our city of Kolkata(India), famously known as City Of Joy, Kolkata is the commercial and financial hub of East and North-East India and home to the Calcutta Stock Exchange.It is a major commercial and military port, and is the only city in eastern India, apart from Bhubaneswar to have an international airport.

Around 6 million people live in this city with a population density of 22 thousand people per square kilometre. This city has been divided into 140 wards. When we think of it by the investor, we expect from them to prefer the districts where the type of business they want to install is less intense. If we think of the city residents they may want to choose the district according to the density of the social place. However, it is difficult to obtain information that will guide investors in this direction, nowadays.

Methodology

Data Collection

There is no available dataset in csv/json/xls format which has enlisted all Kolkata neighbourhoods with latitude and longitude data.This data is only available in Wikipedia. Every Kolkata Corporation Wards have separate Wiki pages and that contains geospatial data. I am planning to run a process that will hit every Wikipedia page (for each Kolkata ward) and will collect relevant information.

An Example Kolkata Corporation Ward Page — "https://en.wikipedia.org/wiki/Ward_No._1,_Kolkata_Municipal_Corporation"

Example Data Set, That can be extracted from Wiki Pages

Coordinates: 22.617889°N 88.370556°E Coordinates: 22.617889°N 88.370556°E Country: India State: West Bengal City: Kolkata Neighbourhood covered: Cossipore Police station: Cossipore Parliamentary constituency: Kolkata Uttar Assembly constituency: Kashipur-Belgachhia

We will consider pulling Coordinates, Ward No and Assembly constituency(as Neighbourhood).

Our next dataset would be Foursquare venue data. After collecting nearby venues for available Kolkata geographical points, those will be filtered to consider only categories which can be marked as eateries.

After analyzing Foursquare Explore api data set which can be accessed with below API calls

https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}

Extracting Coordinates from Web scrapped Data

Now, We will start collecting/ preparing our data from Kolkata Corporation Wiki Pages. We have used the Pandas read_html(link) method to read data from Wiki pages. We are interested to collect data of “Ward no”, “Ward Name”, “Borough”, “Postal Code”, “locality”,” latitude”,” longitude”.In this data preparation phase, We will iterate through all available Wiki pages for all 144 Wards.

Issues We have faced during Extraction:-

1. Extracted text data for coordinates include junk characters like — “\ufeff”
2. Coordinates include “°E” and “°N”. This suffix information has to be removed to work with Folium maps.

Extracted Data frame

The below output will show, how the initial data frame with extracted data will look like

	ward_no	ward	borough	postalcode	locality	latitude	longitude
1	1	Cossipore	1	700 003	Kashipur-Belgachhia	22.617889	88.370556
2	2	Sinthee (Ramlila Bagan-Biswanath Colony-Roypar...	1	700 050	Kashipur-Belgachhia	22.628056	88.384444
3	3	Belgachia, Duttabagan	1	700 037	Kashipur-Belgachhia	22.604444	88.383333
4	4	Paikpara	1	700 002/ 700 037	Kashipur-Belgachhia	22.613056	88.379444
5	5	Tala	1	700 002	Kashipur-Belgachhia	22.608889	88.379694
6	6	Chitpur, Cossipore	1	700 002	Kashipur-Belgachhia	22.610863	88.371213
7	7	Bagbazar	1	700 003	Shyampukur	22.603567	88.365806
8	8	Bagbazar, Shyampukur	1	700 003	Shyampukur	22.601806	88.366500
9	9	Shobhabazar, Kumortuli	1	700 004	Shyampukur	22.595889	88.365306
20	20	Shobhabazar, Ahiritola	2	700 049	Shyampukur	22.593556	88.354667
19	19	Ahiritola and Beniatola	2	700 004	Shyampukur	22.595667	88.357667

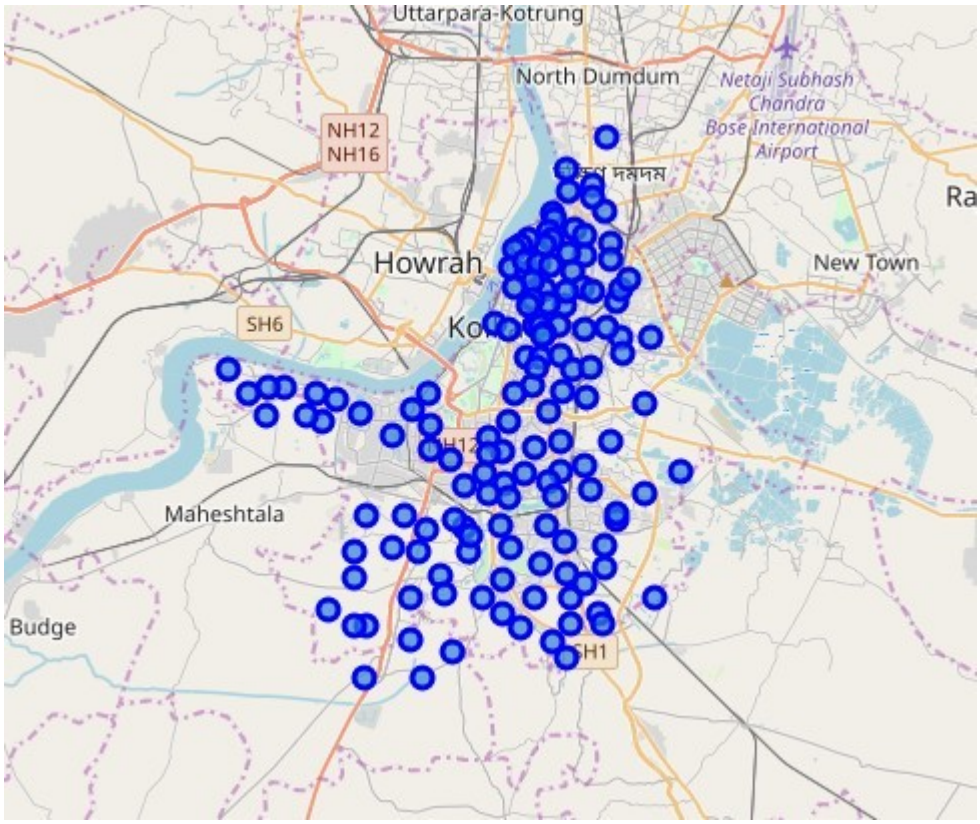
Populating Missing Data with dummy values

After observing the data frame data, We have identified 3 Wards (Jika, Garden Reach (Akshay Kanan),Paikpara) which do not contain any postal code information. We have manually assigned values for those.

	ward_no	ward	borough	postalcode	locality	latitude	longitude
1	1	Cossipore	1	700 003	Kashipur-Belgachhia	22.617889	88.370556
2	2	Sinthee (Ramlila Bagan-Biswanath Colony-Roypar...	1	700 050	Kashipur-Belgachhia	22.628056	88.384444
3	3	Belgachia, Duttabagan	1	700 037	Kashipur-Belgachhia	22.604444	88.383333
4	4	Paikpara	1	700 002, 700 037	Kashipur-Belgachhia	22.613056	88.379444
5	5	Tala	1	700 002	Kashipur-Belgachhia	22.608889	88.379694

Projection of Data Points on Folium Map

Here we see a map of Kolkata with blue circles pointing each ward. Folium CircleMaker method has been used to create these circles by passing latitude and longitude data from above data frame. Folium Popup method has been used to create pop-ups with Ward name and neighbourhood/locality name data. This popup will be visible whenever a user will click on these blue circles



Example Data Frame

This is how an example data frame may look. There are numerous fields returned by foursquare but we will only consider venue-name, venue-categories,venue-location-lat,venue-location-lang

	name	categories	lat	lng
0	Balaram Mullick & Radharaman Mullick	Indian Sweet Shop	22.533097	88.347082
1	Balwant Singh's Eating House	Dhaba	22.537714	88.344220
2	Balwant Singh's Snack Corner	Fast Food Restaurant	22.537709	88.344228
3	Tyre Patty	Café	22.538048	88.349169
4	Privy Ultra Lounge	Nightclub	22.538145	88.351121
5	The Bikers Cafe	American Restaurant	22.537719	88.348993
6	McDonald's	Fast Food Restaurant	22.536206	88.346353
7	Chai Break	Tea Room	22.539382	88.347070
8	Cafe Coffee Day	Café	22.525476	88.345297
9	Harish Park	Park	22.530499	88.343569

Creating broader categories

This is a very important step in our reporting. We do not want to get confused by hundreds of different venue categories available in Foursquare. We have planned to create few bigger/major categories and want to include available categories to Any of these major categories. We have planned to mark these data to below major categories.

'Hospital & Medical','Fitness & Sports' 'Eateries' 'Transport', 'Plaza' 'Stores' 'Financial Institution' 'River & Waterbody', 'Neighborhood' 'Business' 'Hotel', 'Health & Beauty', 'Hostel', 'Campground', 'Nightclub',

Here we have used "Venue Category" in get_dummies method and created columns for each venue category type.

	Hostel	Bed & Breakfast	Business	Campground	Eatries	Financial Institution	Fitness & Sports	Health & Beauty	Hospital & Medical	Hotel	Neighborhood	Nightclub	Plaza	River & Waterbody	Stores	Transport
0	0	0	0	0	0	0	1	0	0	0	Sinthee (Ramlila Bagan-Biswanath Colony-Roypar...	0	0	0	0	0
1	0	0	0	0	1	0	0	0	0	0	Sinthee (Ramlila Bagan-Biswanath Colony-Roypar...	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	Belgachia, Duttabagan	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0	0	Belgachia, Duttabagan	0	0	0	0	1
4	0	0	0	0	0	0	1	0	0	0	Paikpara	0	0	0	0	0

The data frame has been grouped again with neighborhood and Mean() calculated

Then I created a table which shows list of top 10 venue category for each neighbourhood in below table.

	Neighborhood	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
0	Ahiritola and Beniatola	Eatries	Transport	Stores	River & Waterbody	Plaza	Nightclub	Hotel	Hospital & Medical	Health & Beauty	Fitness & Sports
1	Alipore	Eatries	Hotel	Transport	Stores	River & Waterbody	Plaza	Nightclub	Hospital & Medical	Health & Beauty	Fitness & Sports
2	Ashok Nagar, Kudghat, Tollygunge Club, Regent ...	Eatries	Stores	Transport	River & Waterbody	Plaza	Nightclub	Hotel	Hospital & Medical	Health & Beauty	Fitness & Sports
3	Badartala, Rajabagan	Financial Institution	Transport	Stores	River & Waterbody	Plaza	Nightclub	Hotel	Hospital & Medical	Health & Beauty	Fitness & Sports
4	Bagbazar	River & Waterbody	Fitness & Sports	Financial Institution	Eatries	Transport	Stores	Plaza	Nightclub	Hotel	Hospital & Medical

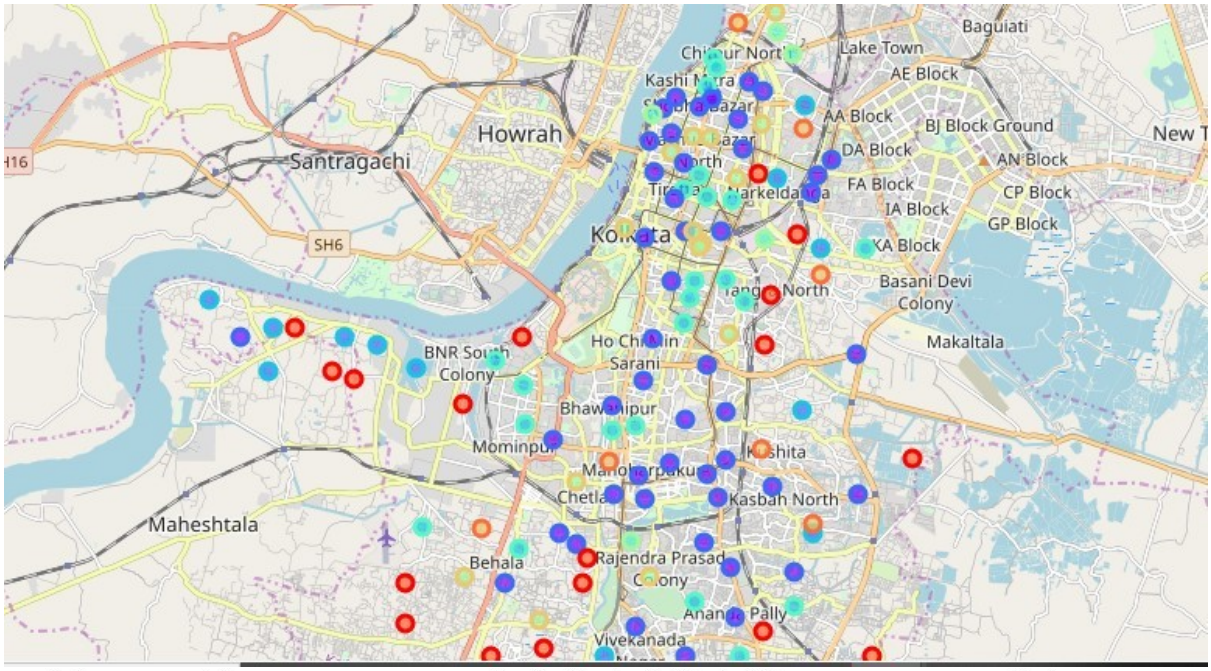
K Mean clustering

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 common cluster method of unsupervised learning.

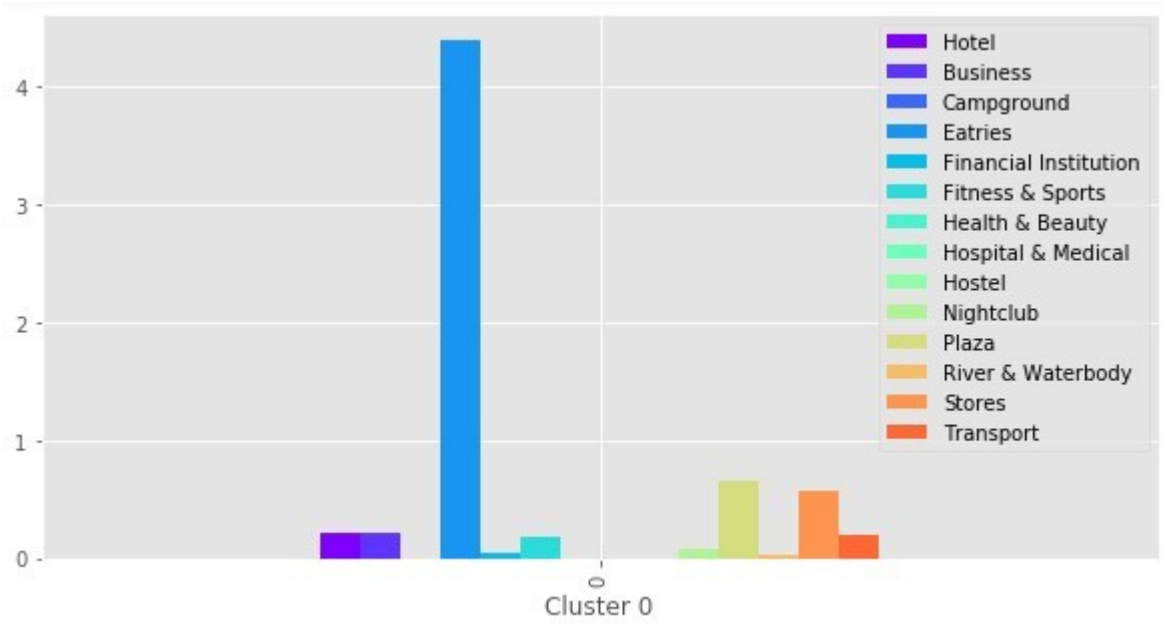
First, I will run K-Means to cluster the wards into 6 clusters

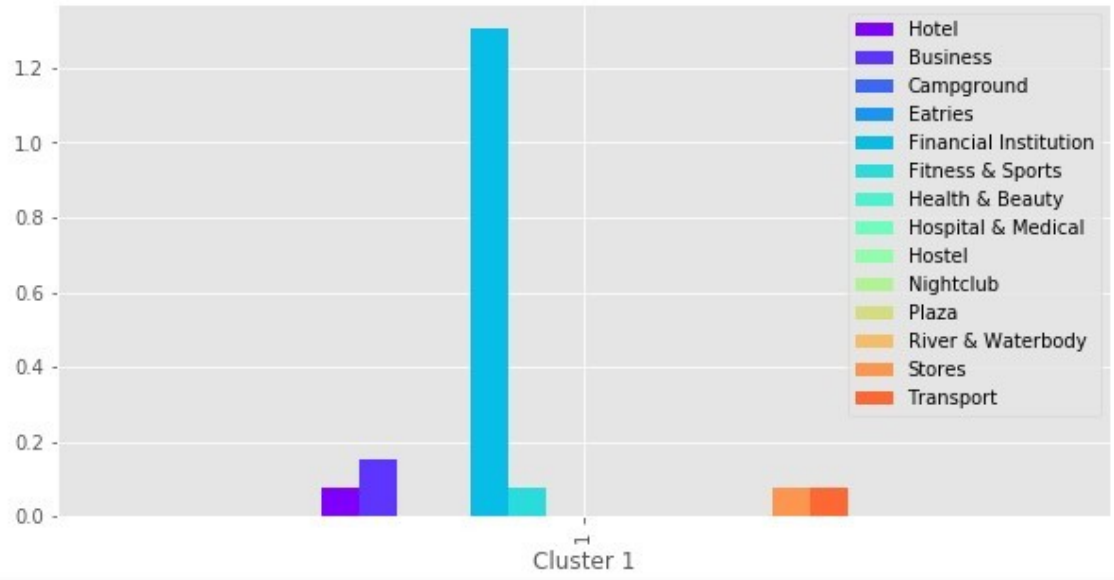
	ward_no	ward	borough	postalcode	locality	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
1	1	Cossipore	1	700 003	Kashipur-Belgachhia	22.617889	88.370556	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	2	Sinthee (Ramlila Bagan-Biswanath Colony-Roypar...	1	700 050	Kashipur-Belgachhia	22.628056	88.384444	0.0	Fitness & Sports	Eatries	Transport	Stores	River & Waterbody	Plaza	Nightclub	Hotel	Hospital & Medical
3	3	Belgachia, Duttabagan	1	700 037	Kashipur-Belgachhia	22.604444	88.383333	3.0	Transport	Stores	River & Waterbody	Plaza	Nightclub	Hotel	Hospital & Medical	Health & Beauty	Fitness & Sports
4	4	Paikpara	1	700 002, 700 037	Kashipur-Belgachhia	22.613056	88.379444	4.0	Plaza	Fitness & Sports	Transport	Stores	River & Waterbody	Nightclub	Hotel	Hospital & Medical	Health & Beauty
5	5	Tala	1	700 002	Kashipur-Belgachhia	22.608889	88.379694	3.0	Transport	Plaza	Stores	River & Waterbody	Nightclub	Hotel	Hospital & Medical	Health & Beauty	Fitness & Sports

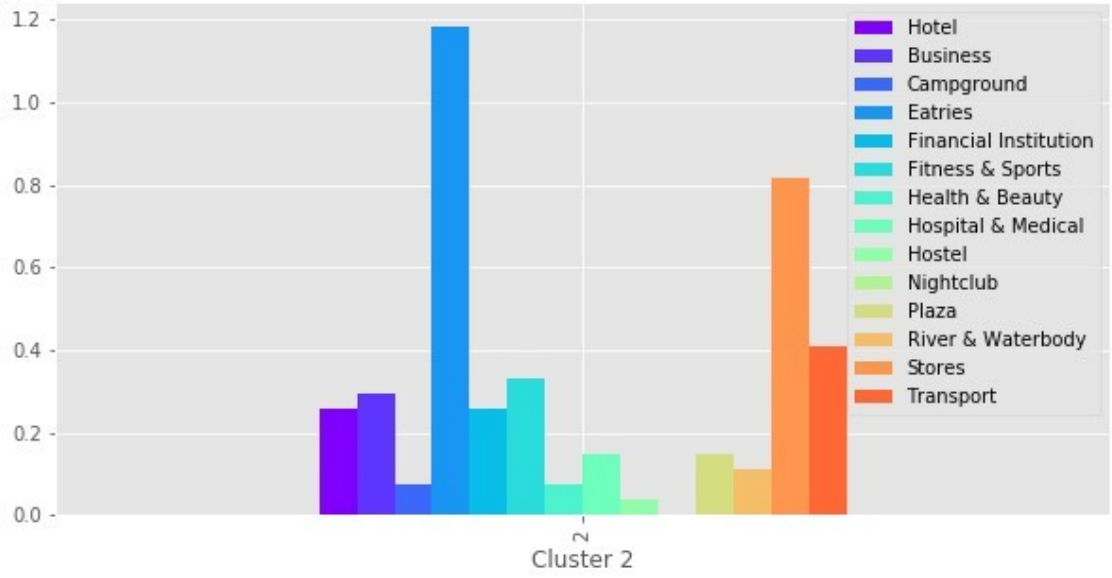
Exploring Clusters with Maps and charts

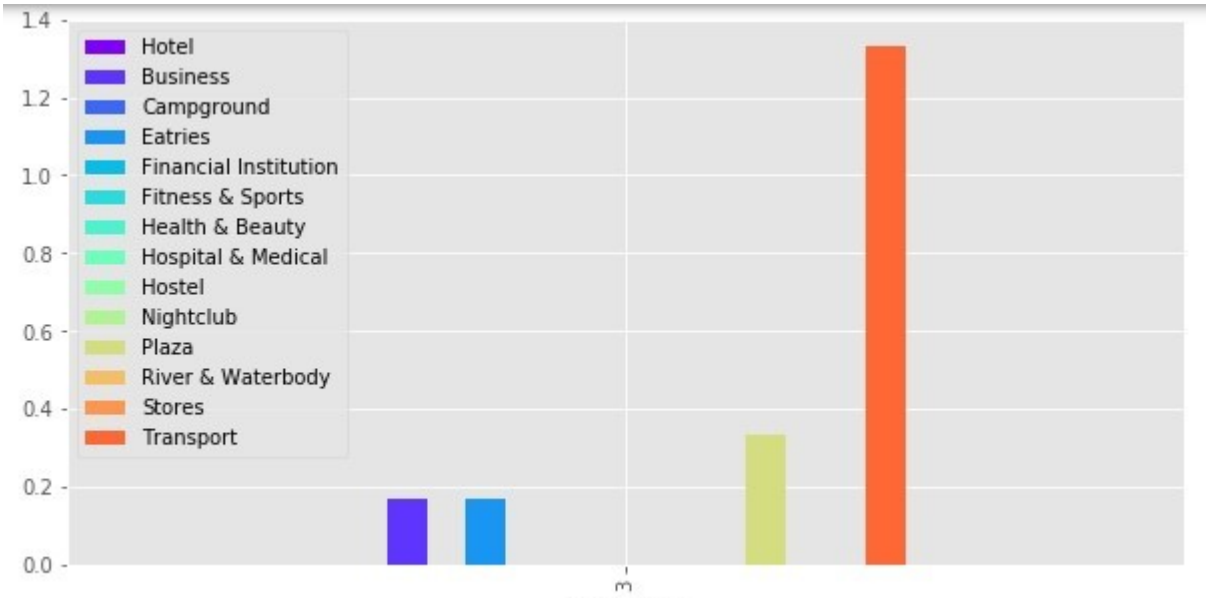


	Cluster Labels	Hostel	Bed & Breakfast	Business	Campground	Eatries	Financial Institution	Fitness & Sports	Health & Beauty	Hospital & Medical	Hotel	Nightclub	Plaza	River & Waterbody	Stores	Transport
0	0	0.000000	0.019231	0.211538	0.000000	4.403846	0.038462	0.173077	0.000000	0.000000	0.211538	0.076923	0.653846	0.019231	0.576923	0.192308
1	1	0.000000	0.000000	0.153846	0.000000	0.000000	1.307692	0.076923	0.000000	0.000000	0.076923	0.000000	0.000000	0.000000	0.076923	0.076923
2	2	0.037037	0.000000	0.296296	0.074074	1.185185	0.259259	0.333333	0.074074	0.148148	0.259259	0.000000	0.148148	0.111111	0.814815	0.407407
3	3	0.000000	0.000000	0.166667	0.000000	0.166667	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.333333	0.000000	0.000000	1.333333
4	4	0.000000	0.000000	0.000000	0.000000	0.600000	0.000000	0.933333	0.000000	0.000000	0.133333	0.000000	1.066667	0.000000	0.133333	0.200000

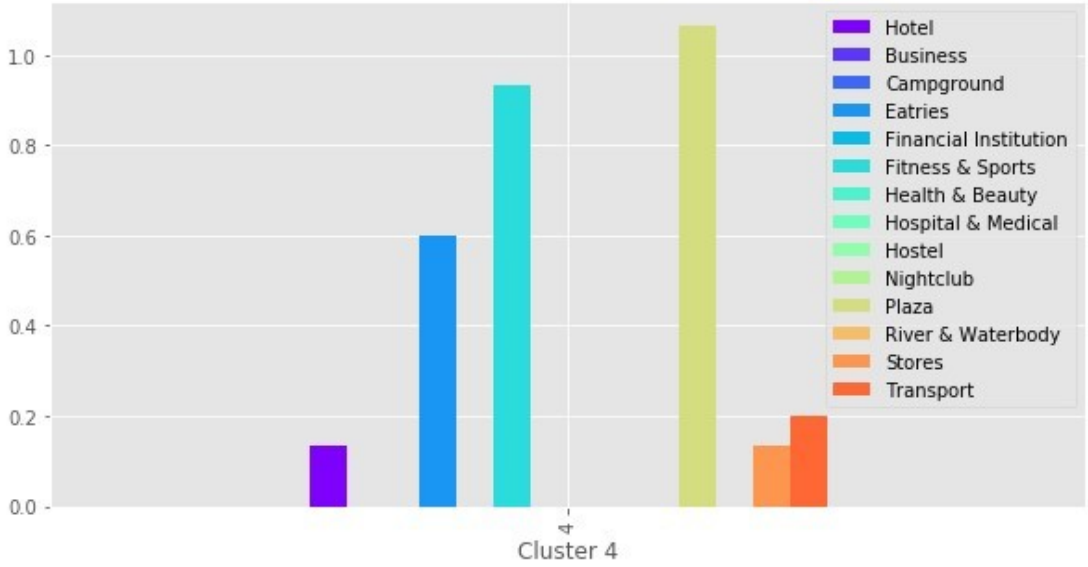

Hotel
Business
Campground
Eatries
Financial Institution
Fitness & Sports
Health & Beauty
Hospital & Medical
Hostel
Nightclub
Plaza
River & Waterbody
Stores
Transport
0
Cluster 0


Hotel
Business
Campground
Eatries
Financial Institution
Fitness & Sports
Health & Beauty
Hospital & Medical
Hostel
Nightclub
Plaza
River & Waterbody
Stores
Transport
1
Cluster 1

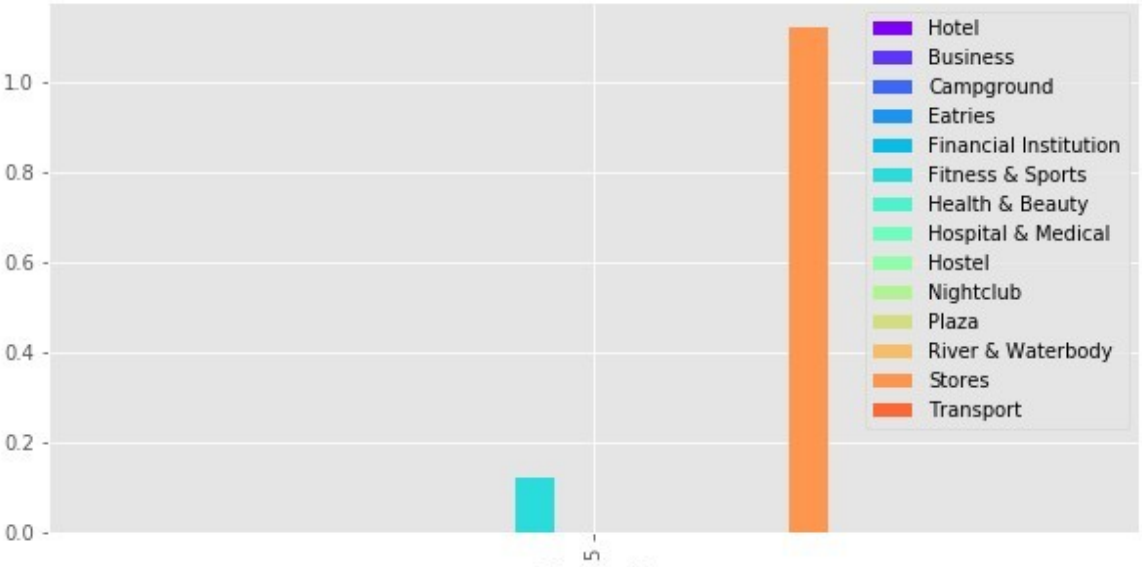

Hotel
Business
Campground
Eatries
Financial Institution
Fitness & Sports
Health & Beauty
Hospital & Medical
Hostel
Nightclub
Plaza
River & Waterbody
Stores
Transport
2
Cluster 2



Cluster 3



Cluster 4



Cluster 5

Results

In the result section, I would like to highlight three major outcomes of this project.

1. We could segregate kolkata wards depending upon different available amenities
2. We could identify most common venues for all kolkata Wards.
3. Different characteristics of every cluster was shown using bar charts. One cluster might have more number of eateries and other might have high on financial institutions.

- Cluster 0 - More Eateries and more than other major type of venues available
- Cluster 1 - More financial institution but less other venues
- Cluster 2 - Best Clusters, nearly all major categories of venues are available.
- Cluster 3 - Good transportation but very few other venues
- Cluster 4 - Many Plazas that suggest lots of shops an business with higher location desirability
- Cluster 5 - Only Stores...

New residents or new investors can identify there cluster of choice by identifying abundance or scarcity of type of venues in each cluster.

Conclusion

This project has tried to provide venue data with different view points. New residents or new investors can identify there cluster of choice by identifying abundance or scarcity of type of venues in each cluster.

References

Wikipedia -https://en.wikipedia.org/wiki/Kolkata

Wikipedia (Ward level) - "https://en.wikipedia.org/wiki/Ward_No._1,_Kolkata_Municipal_Corporation"

Foursquare API