Coursera Capstone Project: Applied Data Science

Suman Kumar Mal
sumank343@gmail.com

April 2020

Opening a new Shopping Mall in Kolkata, West Bengal, India

1 Introduction

The importance of shopping malls as retailing formats has become increasingly remarkable, and today malls plays a significant role in consumers' lifestyle. But nowadays shopping malls have become not just a place to shop it has become a place where social factors get deeper. Shopping malls are like one-stop destination for not only various types of shoppers but also for dine at restaurants, watch movies, celebrate etc. This gives retailers a central location and large crowd at the shopping mall which in turn provides a great distribution channel to market their products and services. Real Estate, builders are also taking interest to build shopping malls to cater the demand. This also becomes a consistent rental income for the owners. As a result, there are many shopping malls in the Kolkata and more and more malls are being built. But to open a shopping and such that the shopping mall succeeds lot of factors come into play. Among them one of the major factors is the location of the shopping mall.

2 Business Problem

Since, shopping malls are main interest to various groups therefore they are built and real estate investors invest in these projects. But for shopping malls to attract large crowd there are few major factors, one of those is location. The objective of this capstone project is to analyse and select the best locations in Kolkata, West Bengal, India to open a new shopping mall with high chances of success using data science methodology and machine learning techniques. We will make this decision in this project.

This will specially benefit the real estate builders since Indian retail sector has metamorphosed significantly over last few decades. Rapid urbanization and digitization, rising disposable incomes and lifestyle changes of particularly the middle-class has led to a major revolution in the retail sector, projected to grow from US $672 billion in 2017 to US $1.3 trillion in 2020. Evolving rapidly from usual 'kirana shops' to large multi-format stores offering global experience to the e-commerce model that is highly technology-driven, the Indian retail sector has evolved.

3 Data

To solve the problem, we will need the following data:

1. List of neighbourhoods in Kolkata. This defines the scope of this project which is confined to Kolkata, the capital city of the state West Bengal, India.
2. Latitude and Longitude coordinates of the neighbourhoods. This required to plot the map and get the venue data
3. Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the neighbourhoods.

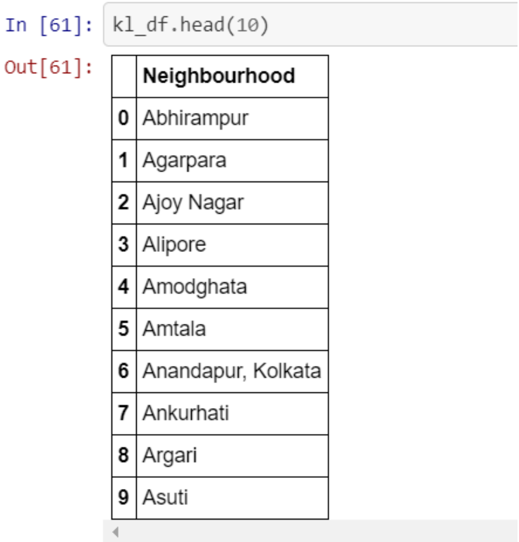
Data Sources:

- The data of the neighbourhoods in Kolkata can be extracted from Wikipedia page. (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Kolkata)
- Then the latitude and longitude data can be retrieved from Python geocoder package.
- Then using latitude and longitude data venues can be fetched from Foursquare API.

4 Methodology

Method to extract data:

- We will do web scraping using *BeautifulSoup* a library of python to get the neighbourhoods from the Wikipedia page. We will send get request to get the html page and then extract the list and store it in a csv file.
  Shape of our data frame is (323 × 1)



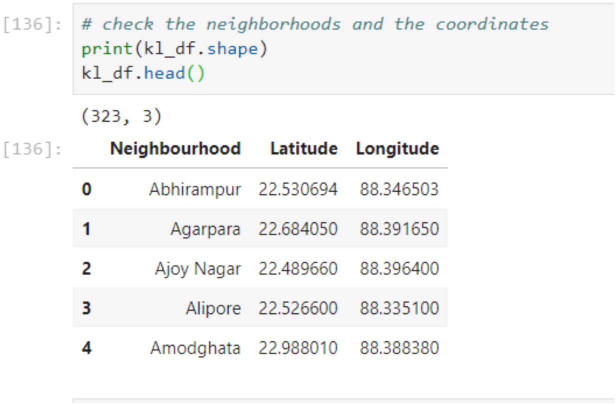
```
In [61]: kl_df.head(10)
```

Out[61]:

	Neighbourhood
0	Abhirampur
1	Agarpara
2	Ajoy Nagar
3	Alipore
4	Amodghata
5	Amtala
6	Anandapur, Kolkata
7	Ankurhati
8	Argari
9	Asuti

Method get Longitude and Latitude and Visualize :

- The list will be just names of the places but we will need their geographical coordinates so we will pass the data from the csv file to *Geocoder* to get the latitude and longitude. Then we will append the latitude and longitude of individual location to the dataframe.



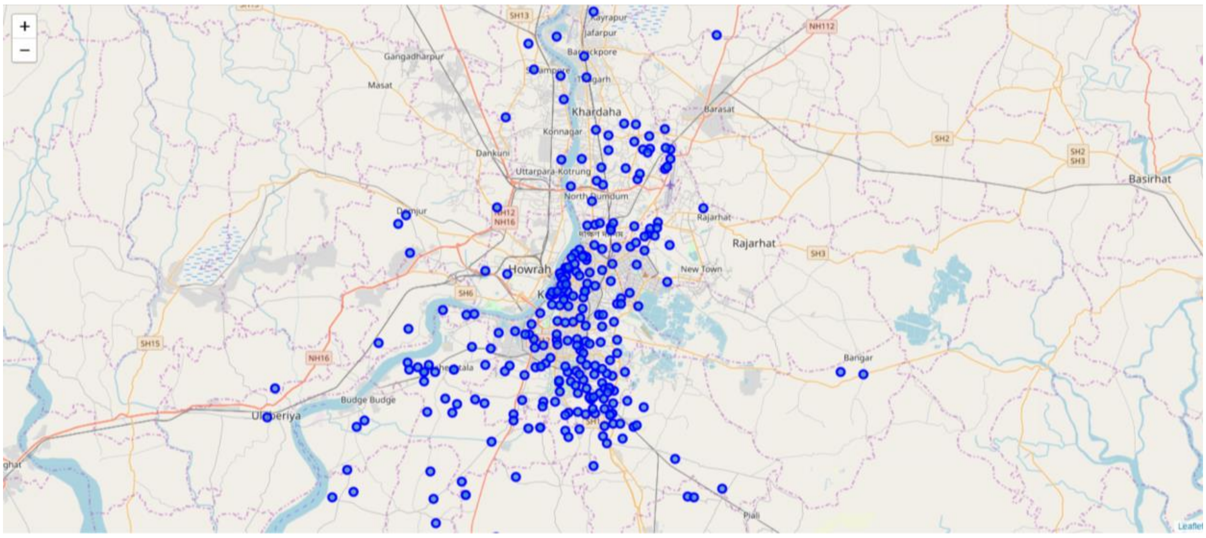
```
[136]: # check the neighborhoods and the coordinates
       print(kl_df.shape)
       kl_df.head()

       (323, 3)
```

[136]:

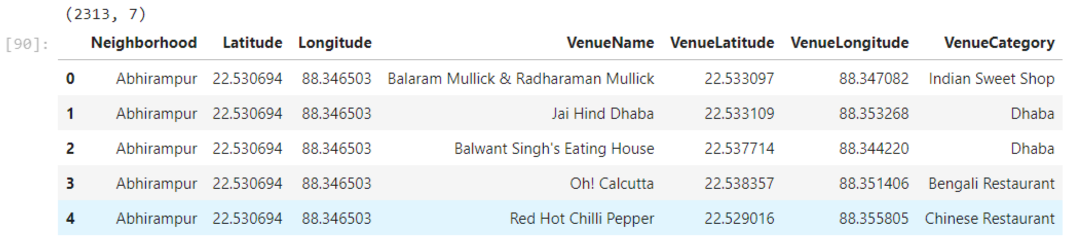
	Neighbourhood	Latitude	Longitude
0	Abhirampur	22.530694	88.346503
1	Agarpara	22.684050	88.391650
2	Ajoy Nagar	22.489660	88.396400
3	Alipore	22.526600	88.335100
4	Amodghata	22.988010	88.388380

- Now we can visualize these locations on map using *Folium*. This allows us to perform a sanity check to make sure that the geographical coordinates returned by *Geocoder* are correctly plotted.



Method to get venue data from FourSquare API:

- Next, we will use *Foursquare API* to get the top 30 venues that are within a radius of 1000 meters. For this a *Foursquare* developer account is needed.
  We make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. We use the venues/explore API endpoint to request the data. Foursquare returns the venue data in JSON format which is then decoded to extract venue names, venue category, venue longitude and venue latitude.

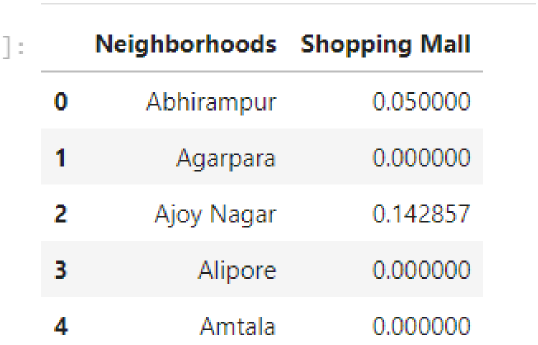


```
(2313, 7)
[90]:
```

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Abhirampur	22.530694	88.346503	Balaram Mullick & Radharaman Mullick	22.533097	88.347082	Indian Sweet Shop
1	Abhirampur	22.530694	88.346503	Jai Hind Dhaba	22.533109	88.353268	Dhaba
2	Abhirampur	22.530694	88.346503	Balwant Singh's Eating House	22.537714	88.344220	Dhaba
3	Abhirampur	22.530694	88.346503	Oh! Calcutta	22.538357	88.351406	Bengali Restaurant
4	Abhirampur	22.530694	88.346503	Red Hot Chilli Pepper	22.529016	88.355805	Chinese Restaurant

Grouping data and Performing Clustering:

- Then we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the "Shopping Mall" data we will filter the "Shopping Mall" as venue category for the neighbourhoods.



```
]:
```

	Neighborhoods	Shopping Mall
0	Abhirampur	0.050000
1	Agarpara	0.000000
2	Ajoy Nagar	0.142857
3	Alipore	0.000000
4	Amtala	0.000000

- Lastly, we will perform clustering on the data by using K-means Clustering. K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping centroid as small as possible. In this project we have clustered the neighbourhoods into 3 clusters based on their frequency of occurrence for "Shopping Mall". Now, lets go towards the results and discuss them.

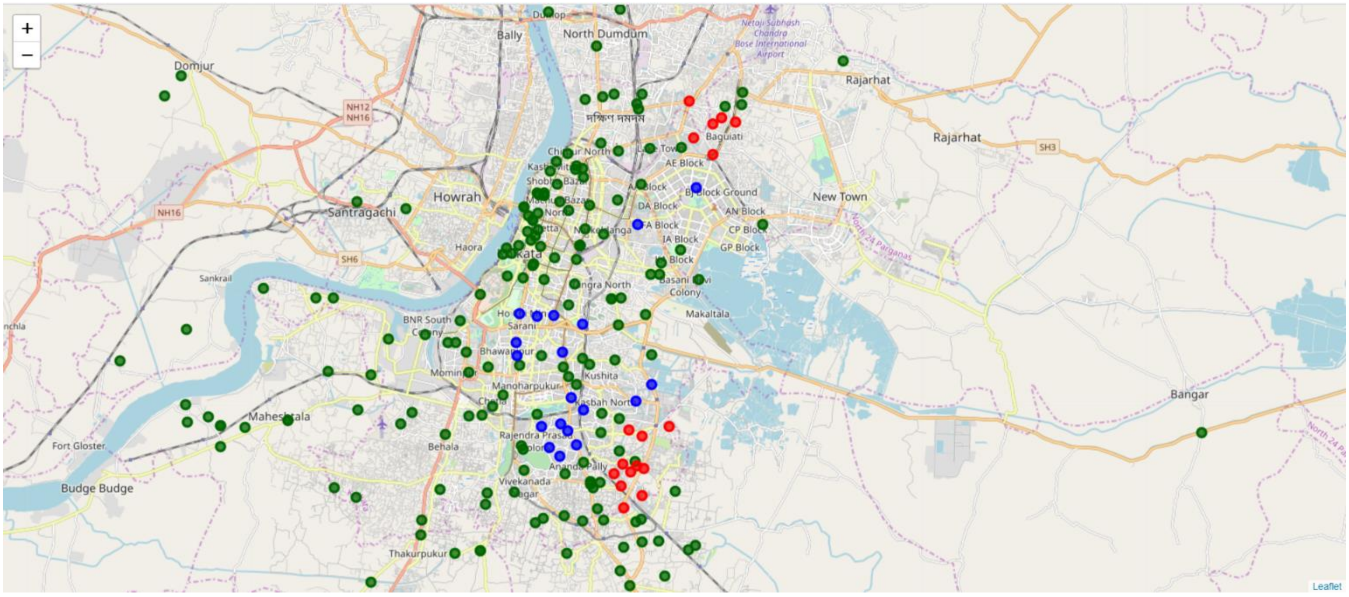
5 Result

Results from K-means Clustering show that:

- Cluster 0: Neighbourhoods with low to no existence of shopping malls.
- Cluster 1: Neighbourhoods with high concentration of shopping malls.
- Cluster 2: Neighbourhoods with moderate number of shopping malls.

Colour Code

Cluster 0 – Dark Green
Cluster 1 – Red
Cluster 2 – Blue



6 Discussion

Most of the shopping malls are concentrated in the central east, southern and few on the northern part of Kolkata, with the highest number in cluster 1 and moderate number in cluster 2. On the other hand, cluster 0 has very low number to totally no shopping mall in the neighbourhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls. Meanwhile, shopping malls in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. From another perspective, this also shows that the oversupply of shopping malls mostly happened in the central eastern and southern part of the city, with the suburb area still have very few shopping malls. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in neighbourhoods in cluster 0 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighbourhoods in cluster 2 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 1 which already have high concentration of shopping malls and suffering from intense competition.

7 Conclusion

In this project, we have gone through all the data science methodology. We first identified a business problem, then collected the required data, preparing the data and then performing machine learning by clustering the data into 3 clusters based on their similarities. Finally, we have also provided recommendations to the relevant stakeholders regarding the best locations to open a new shopping mall.

To answer the problem which was raised at the Introduction: The neighbourhoods in cluster 0 are most preferred location to open a new shopping mall.

Appendix

Cluster 0

- Kamalgazi
- Mankundu
- ModelTown,Kolkata
- Mukundapur,Kolkata
- NabagramColony
- Nadabhanga
- Naihati
- Naktala
- Naldanga,Hooghly
- Nalpur
- NandiBagan
- Narendrapur
- Narkeldanga
- Nawabganj,North24Parganas
- NetajiNagar,Kolkata
- NewAlipore
- NewBarrackpore
- NewGaria
- NewTown,Kolkata
- Nibra
- Nimta
- Noapara,India
- Noapara,WestBengal
- NorthBarrackpur
- NorthDumdum
- Nungi
- Palta,North24Parganas
- Panchpara
- Panchpota
- Panihati
- Maniktala
- Manikpur,WestBengal
- Mahiari
- Maheshtala
- Kadamtala
- Kaikhali
- Kalara
- Kalighat
- Kalua,Maheshtala
- Kalyani,WestBengal
- Uttarpara
- Kamarhati
- Kanaipur
- Kanchrapara
- Kanganbaria
- Kantlia
- Kasba,Kolkata
- Keota,Hooghly
- ParnasreePally
- Kesabpur,India
- Khantora
- Khardaha
- Kidderpore
- Kodalia
- KolkataWestInternationalCity
- Konnagar
- Kriparampur
- Kudghat
- Kulihanda
- Kumortuli
- LakeTown,Kolkata
- Lalbazar
- Liluah
- Madhyamgram
- Khalia
- Jorasanko
- Pathuriaghata
- PicnicGarden
- Shyamnagar,WestBengal
- Shyampukur
- Simla,WestBengal
- Sinthee
- Sodepur
- Sonagachi
- SouthDumdum
- SouthEasternRailwayColony
- SouthWestKolkata
- Subhashgram
- Sukchar
- Tala,Kolkata
- Talbandha
- Taltala
- Tangra,Kolkata
- Taratala
- TechnoCity,Kolkata
- Teghoria
- Tentulberia
- Thakurpukur
- Tiljala
- TirettaBazaar
- Titagarh
- Tollygunge
- Topsia
- Tribeni,Hooghly
- Ultadanga
- Uluberia
- UttarRaypur
- Shyambazar
- Shobhabazar
- Shibpur
- Sheoraphuli
- Poali

- PoddarNagar
- Podrah
- Posta,Burrabazar
- RabindraSarobar
- Raghudebbati
- Raghunathpur(PS-Dankuni)
- Rahara,Kolkata
- Raigachhi
- Rajabagan
- Rajabazar,Kolkata
- Ramchandrapur,Maheshtala
- Ramchandrapur,Sonarpur
- Ramchandrapur,WestBengal
- Phoolbagan
- Ramkrishnapur,Bishnupur
- RegentPark,Kolkata
- Rishra
- Rishra,SreerampurUttarpara
- Salkia
- Samali
- Sankrail
- Santoshpur,Kolkata
- Santoshpur,Uluberia
- Santragachhi
- Saptagram
- Sarenga
- Sarsuna
- Sealdah
- Serampore
- Ramrajatala
- Jorabagan
- Watgunge
- ChakBaria
- Bijoygarh
- Bijpur,North24Parganas
- Baghajatin
- Birati
- Bagbazar
- Bishnupur,South24Parganas
- BowBarracks
- Bowali
- Bowbazar
- Brahmapur,WestBengal
- BudgeBudge
- Buita
- Bidyadharpur
- Burrabazar
- Badartala
- CalcuttaRiverside
- Joka,Kolkata
- Babughat
- Chakapara
- 69Champdani
- 70Chandannagar
- Chandpur,Ghola
- CharuMarket
- B.B.D.Bagh
- Chinsurah
- Chitpur
- Burtolla
- Argari
- Baidyabati
- Bhasa,Bishnupur
- Bally,Bally-Jagachha
- BallygungeCircularRoad
- Bandel
- BangurAvenue
- Bankra
- Bansdroni
- Bantala
- Banupur
- Baranagar
- Barasat
- Barisha,India
- Barrackpore
- Bhatpara
- BarrackpurCantonment
- Baruipur
- Batanagar
- Behala
- Belgachia
- Belgharia
- Beliaghata
- Balarampur,BudgeBudge
- Beniapukur
- Bhadrakali,Hooghly
- Bhadreswar,Hooghly
- BhangarRaghunathpur
- BalaramPota
- Bartala
- Cossipore
- Chetla
- DakshinRajyadharpur
- Ghola,North24Parganas
- DakshinJhapardaha
- GirishPark
- Gobindapur,Bhangar
- Alipore
- Gondalpara
- Halisahar
- Hanspukuria
- Haridevpur
- Harinavi
- Hastings,Kolkata
- Hatgachha
- Hatibagan
- Garulia
- HindMotor

- Hridaypur
- Agarpara
- Hugli-Chuchura
- Ichapore
- IchhapurDefenceEstate
- Jadavpur
- Jagacha
- Jagadishpur
- Jagatdal
- Janbazar
- JaynagarMajilpur
- Jetia
- Jhorhat
- Howrah
- Garshyamnagar
- Amtala
- Anandapur,Kolkata
- Dakshineswar
- Dankuni
- DhalaiBridge
- Dhapa,India
- Dharmapur,WestBengal
- Dharmatala
- Dhuilya
- DumDum
- Durganagar,Kolkata
- EastKolkata
- Ekbalpur
- Eksara
- Duttapukur
- Esplanade,Kolkata
- Garia
- Garfa
- GardenReach
- FortWilliam,India
- Entally
- Bally,Howrah

Cluster 1

- Sahaganj
- Pujali
- Nagerbazar
- Ankurhati
- SurveyPark
- Baguiati
- Ranikuthi
- AjoyNagar
- Makardaha
- Kestopur
- Rajarhat
- Patulia
- Barkalikapur
- DumDumPark
- ChakGaria
- Panchasayar
- Haltu

Cluster 2

- Jodhpur Park
- Belur, West Bengal
- Kankurgachi
- Katju Nagar
- Golf Green
- Lake Gardens
- Gariahat Road
- Dhakuria
- Chowringhee
- North Kolkata
- Park Circus
- Calcutta International School
- Birlapur
- Bikramgarh
- Bidhannagar
- Bhowanipore
- Ruiya
- Ballygunge
- Abhirampur