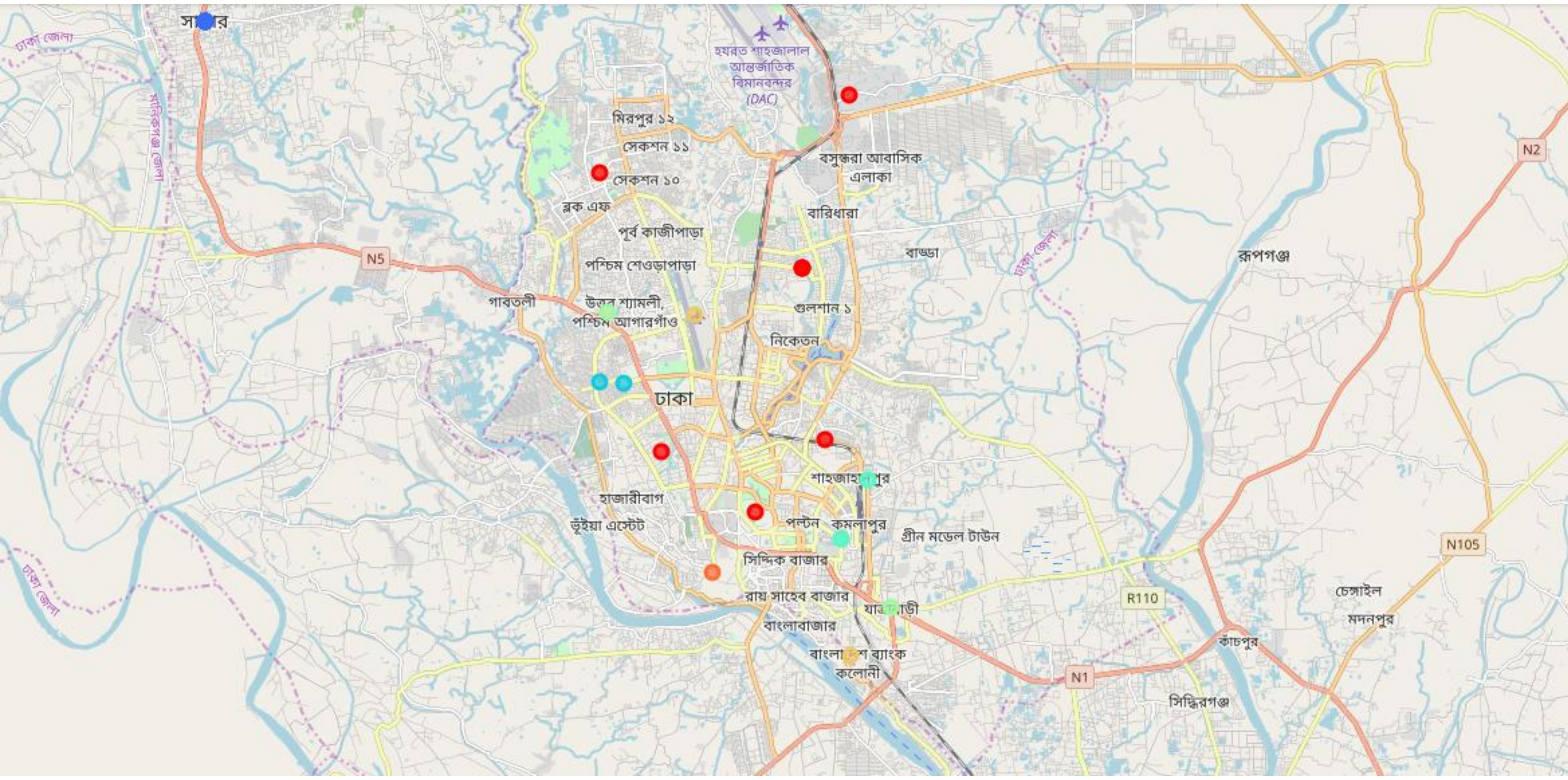
Finding a Suitable Place for Buying an Apartment in Dhaka, Bangladesh



Coursera Capstone

IBM Applied Data Science Capstone

Ragib Shahariar Ayon

Rajshahi University of Engineering and Technology

Table of Contents

Abstract

This is a project report for Coursera's Applied Data Science Capstone project: The Battle of Neighborhoods. This report will explain how a potential real-state buyer can choose a desired neighborhood efficiently within a short period of time. We have collected data for neighborhoods in Dhaka city and utilized K mean algorithm to find out similarities within the dataset and cluster them on the basis of similarities. This project utilized numerous data science techniques and methodologies such as web scraping, data acquisition, data wrangling, machine learning and telling a story based on the result. This is a pilot project which will open doors to numerous other project ideas based on location data and how to utilize it to use the data more efficiently.

Chapter 1: Introduction

Every day potential real state buyers spend many hours to find out which area will be perfect for their living. They want everything at their fingertips and they also look for similar neighborhoods as they were been to. So, finding the similarities between neighborhoods and clustering them can be a solution to choose where to buy an apartment or real state easily and efficiently within a short time.

Where should I buy it?

The objective of this capstone project is, to analyze and select locations in Dhaka, Bangladesh to cluster them based on their top ten most common venues. This project will find out hidden patterns between these locations and based on this pattern one will be able to choose a cluster of areas with the same features and narrow down his search for a specific purpose. This project aims to provide the answer for the following business question: **In Dhaka, Bangladesh if a buyer wants to buy an apartment where would you recommend them to buy based on their current area's characteristics?**

Conclusion

This is a project that will make use of many data science skills, such as, web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Chapter 2: Methodology

In this chapter, we will discuss how we acquired the data and manipulated it to reach a solution

Dataset Acquisition

To solve the problem, we need the following data:

- List of neighborhoods in Dhaka. This defines the scope of this project which is confined to the Dhaka city, the capital city of Bangladesh.
- The latitude and longitude coordinates of those neighborhoods. This is required in order to find out more about the surrounding venues of these areas and to plot a map as well.
- Venue data, particularly data that are within 1 KM radius of the neighborhood. We will use this data to perform clustering on the neighborhoods.

Finding Neighborhood Location

This Wikipedia page[1] contains a list of postal codes in Bangladesh. This page contains tables with District, Thana, Sub-Office and postal codes according to divisions. From this page, we want to extract data pertaining to the Dhaka division only. Here we have chosen Sub-Office of these postal areas as neighborhoods.

We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python Get requests and bs4 packages. The request will give us an HTML format of the wiki page and we have to extract the required information by cleaning it using the BeautifulSoup module. Then we will get the geographical coordinates of these neighborhoods using another python package- geopy. From geopy, we will use Nominatim which will give us the required latitude and longitude coordinates of the neighborhoods.

Finding nearby venues

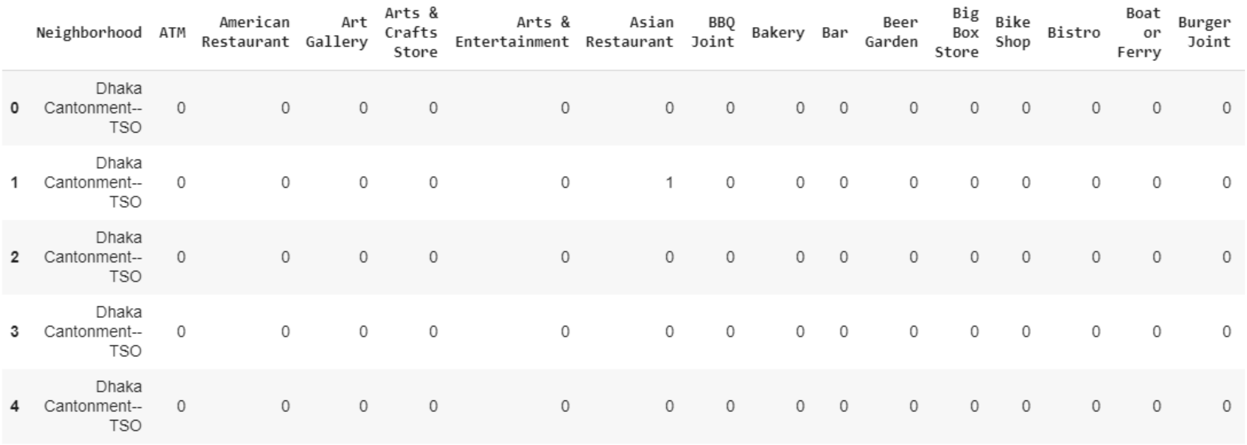
After getting the latitude and longitude data of the neighborhoods, we will use **Foursquare**'s API to get the venue data for those neighborhoods. Foursquare has one of the largest databases of over 105 million places and is used by over 125,000 developers. Foursquare API will provide the nearest venue data including their name, location, and venue type.

[1] https://en.wikipedia.org/wiki/List_of_postal_codes_in_Bangladesh

Data Processing

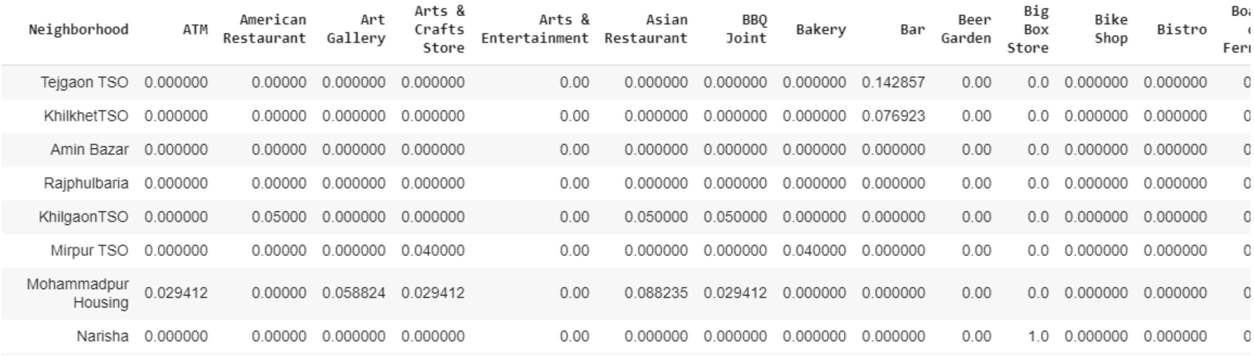
The dataset contains 1 District, 48 Thana, 48 Sub office (neighbor), 809 unique venues and 35 unique venue categories. We extracted the venue and venue categories by using the Foursquare API.

The Data frame consists of categorical values which we have to convert into numerical values for analyzing it. We used a one-hot encoding to convert the data set from the categorical variable. The Shape of the Dataset after one hot encoding was (849, 95). After that, we grouped the data set based on neighborhoods and took the mean of every column. The shape after grouping by neighbors was. (39, 95) after processing the data we can cluster the data set using K mean Clustering.



	Neighborhood	ATM	American Restaurant	Art Gallery	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	BBQ Joint	Bakery	Bar	Beer Garden	Big Box Store	Bike Shop	Bistro	Boat or Ferry	Burger Joint
0	Dhaka Cantonment--TSO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Dhaka Cantonment--TSO	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
2	Dhaka Cantonment--TSO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Dhaka Cantonment--TSO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Dhaka Cantonment--TSO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 1 Dataset: one hot encoding



Neighborhood	ATM	American Restaurant	Art Gallery	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	BBQ Joint	Bakery	Bar	Beer Garden	Big Box Store	Bike Shop	Bistro	Boa Ferr
Tejgaon TSO	0.000000	0.00000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.142857	0.00	0.0	0.000000	0.000000	0
KhilkhetTSO	0.000000	0.00000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.076923	0.00	0.0	0.000000	0.000000	0
Amin Bazar	0.000000	0.00000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.0	0.000000	0.000000	0
Rajphulbaria	0.000000	0.00000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	0.0	0.000000	0.000000	0
KhilgaonTSO	0.000000	0.05000	0.000000	0.000000	0.00	0.050000	0.050000	0.000000	0.000000	0.00	0.0	0.000000	0.000000	0
Mirpur TSO	0.000000	0.00000	0.000000	0.040000	0.00	0.000000	0.000000	0.040000	0.000000	0.00	0.0	0.000000	0.000000	0
Mohammadpur Housing	0.029412	0.00000	0.058824	0.029412	0.00	0.088235	0.029412	0.000000	0.000000	0.00	0.0	0.000000	0.000000	0
Narisha	0.000000	0.00000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.00	1.0	0.000000	0.000000	0

Figure 2 Dataset after grouping by neighborhood and taking by mean

We normalized the data so that the each attribute has equal impact and the result is not skewed.

K mean Clustering

Optimizing the number of k

We used K mean clustering for clustering the dataset. At first, we evaluated SSE(Sum Of Squares Error) and found that we should take 7 clusters so that the SSE remains low but the data set is not overly clustered.

We used 'k-means++' as initializer. It selects initial cluster centers for k-mean clustering in a smart way to speed up convergence.

n_init was set to 1000. n_init is the number of times the k-means algorithm will be run with different centroid seeds. The final results will be the best output of n_init consecutive runs in terms of inertia.

The max_iter: was set to 1000. max_iter is the maximum number of iterations of the k-means algorithm for a single run.

So we can say that the output of the algorithm was good enough to take.

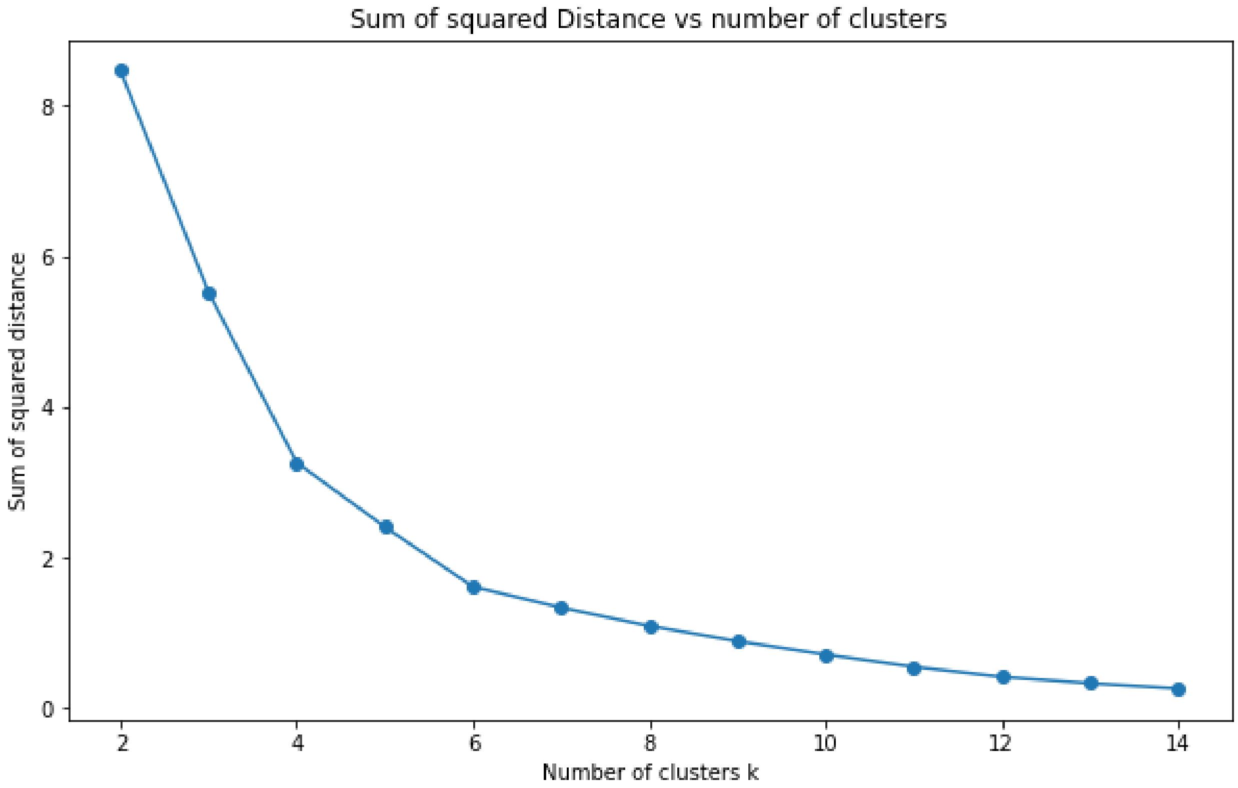


Figure 3 SSE of the K mean algorithm

Mapping the Data

After clustering, we can plot our neighborhood in the Folium map to visualize which areas are similar.

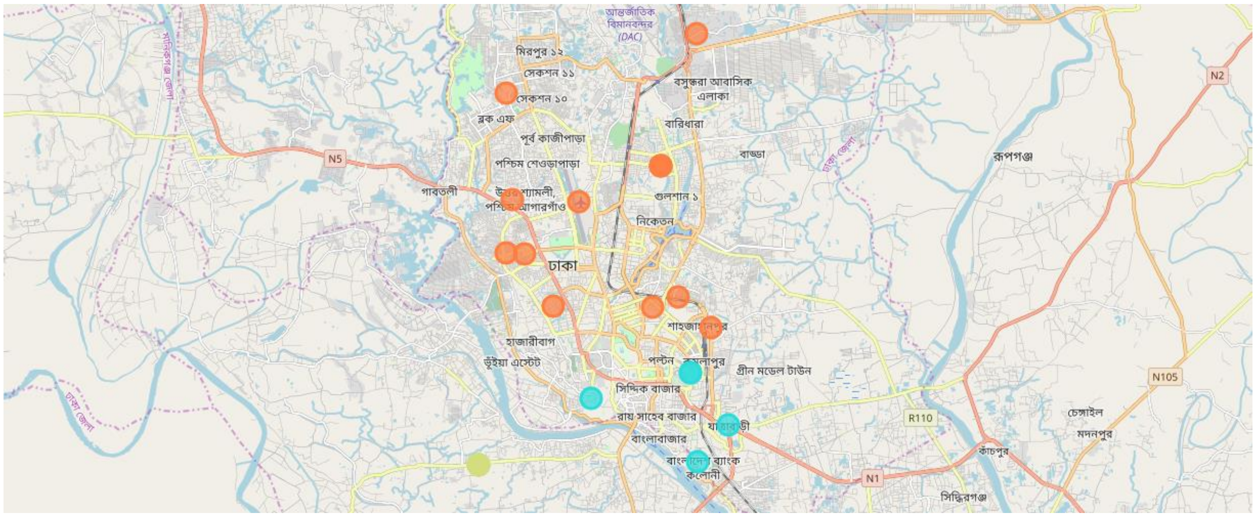
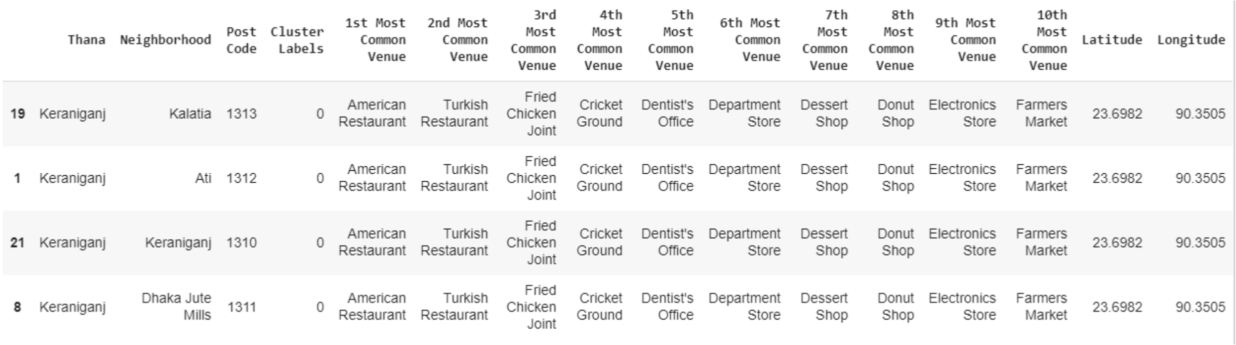


Figure 4 Map of clustered areas.

Chapter 3: Result analysis

Cluster 0 is in Thana Keraniganj with Ati, Keraniganj, Dhaka Jute mills, and kalatia locations which have common attributes within but they have distinct differences with other clusters. the most common restaurant categories for this cluster are American and Turkish restaurants. They also have a cricket ground Department store farmers market within 1.5 KM range.

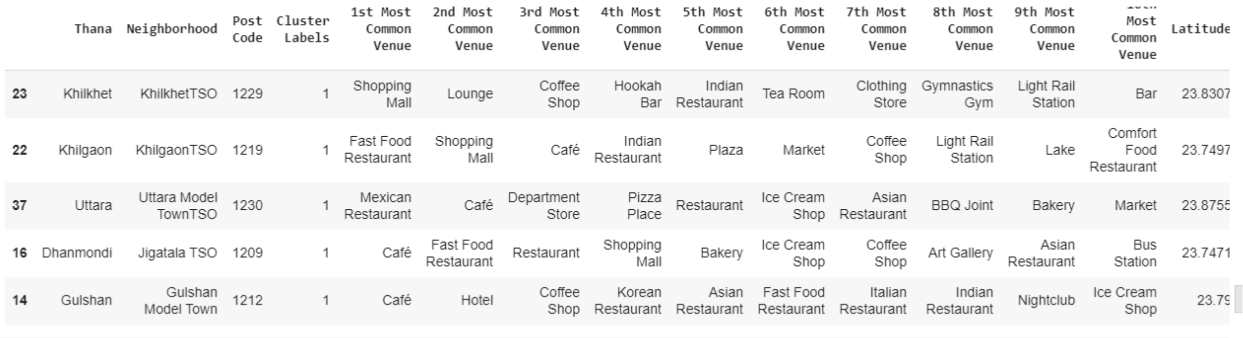


	Thana	Neighborhood	Post Code	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Latitude	Longitude
19	Keraniganj	Kalatia	1313	0	American Restaurant	Turkish Restaurant	Fried Chicken Joint	Cricket Ground	Dentist's Office	Department Store	Dessert Shop	Donut Shop	Electronics Store	Farmers Market	23.6982	90.3505
1	Keraniganj	Ati	1312	0	American Restaurant	Turkish Restaurant	Fried Chicken Joint	Cricket Ground	Dentist's Office	Department Store	Dessert Shop	Donut Shop	Electronics Store	Farmers Market	23.6982	90.3505
21	Keraniganj	Keraniganj	1310	0	American Restaurant	Turkish Restaurant	Fried Chicken Joint	Cricket Ground	Dentist's Office	Department Store	Dessert Shop	Donut Shop	Electronics Store	Farmers Market	23.6982	90.3505
8	Keraniganj	Dhaka Jute Mills	1311	0	American Restaurant	Turkish Restaurant	Fried Chicken Joint	Cricket Ground	Dentist's Office	Department Store	Dessert Shop	Donut Shop	Electronics Store	Farmers Market	23.6982	90.3505

Figure 1CLuster 0

Cluster 1 consists of 11 Thanas. They are Gulshan, Mohammadpur, Khilkhet, Uttara, Khilgaon, Ramna, Tejgaon, Dhaka Sadar Dhanmondi, shabujbag, and Mirpur. we have analyzed 14 neighborhoods in this area. they are Khilkhet-TSO, Khilgaon-TSO, Uttara Model Town-TSO, Jigatala TSO, Gulshan Model Town, Sangsad Bhaban TSO, Mirpur TSO, Mohammadpur Housing, Shantinagr TSO, Dhaka Cantonment TSO, Basabo TSO, Tejgaon TSO, Banani TSO, and Badda.

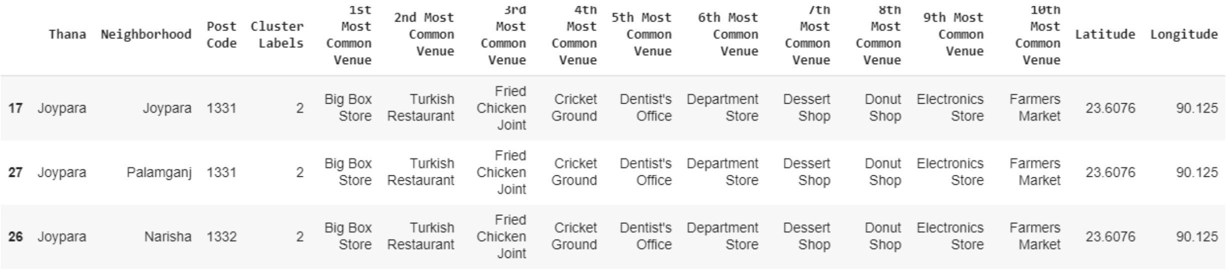
the most common item in these areas is Café, Fast Food Restaurant, Shopping Mall, and Hotel.



	Thana	Neighborhood	Post Code	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	Most Common Venue	Latitude
23	Khilkhet	KhilkhetTSO	1229	1	Shopping Mall	Lounge	Coffee Shop	Hookah Bar	Indian Restaurant	Tea Room	Clothing Store	Gymnastics Gym	Light Rail Station	Bar	23.8307
22	Khilgaon	KhilgaonTSO	1219	1	Fast Food Restaurant	Shopping Mall	Café	Indian Restaurant	Plaza	Market	Coffee Shop	Light Rail Station	Lake	Comfort Food Restaurant	23.7497
37	Uttara	Uttara Model TownTSO	1230	1	Mexican Restaurant	Café	Department Store	Pizza Place	Restaurant	Ice Cream Shop	Asian Restaurant	BBQ Joint	Bakery	Market	23.8755
16	Dhanmondi	Jigatala TSO	1209	1	Café	Fast Food Restaurant	Restaurant	Shopping Mall	Bakery	Ice Cream Shop	Coffee Shop	Art Gallery	Asian Restaurant	Bus Station	23.7471
14	Gulshan	Gulshan Model Town	1212	1	Café	Hotel	Coffee Shop	Korean Restaurant	Asian Restaurant	Fast Food Restaurant	Italian Restaurant	Indian Restaurant	Nightclub	Ice Cream Shop	23.79

Figure 2 Cluster 1

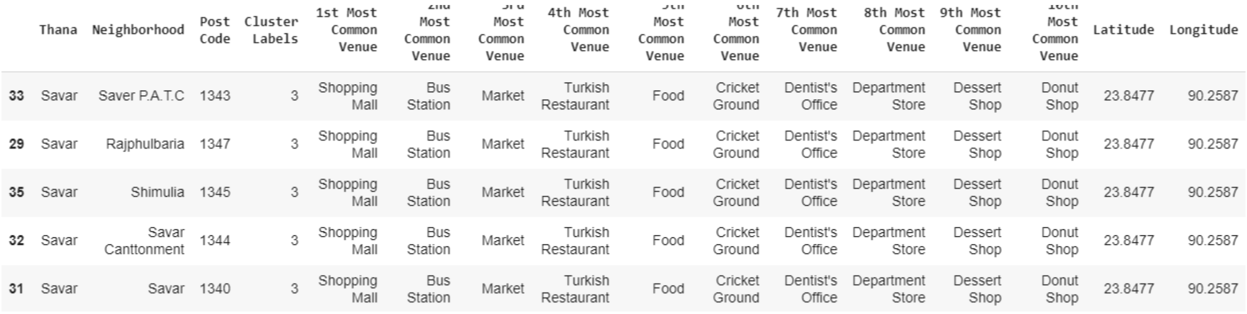
In this study, Cluster 2 consists of one thana, Jaypara, with three neighborhood Joypara, Palamganj, and Narisha. The most common items in this location are Big Box store, Turkish Restaurant, and Fried Chicken Joint.



	Thana	Neighborhood	Post Code	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	Latitude	Longitude
17	Joypara	Joypara	1331	2	Big Box Store	Turkish Restaurant	Fried Chicken Joint	Cricket Ground	Dentist's Office	Department Store	Dessert Shop	Donut Shop	Electronics Store	Farmers Market	23.6076	90.125
27	Joypara	Palamganj	1331	2	Big Box Store	Turkish Restaurant	Fried Chicken Joint	Cricket Ground	Dentist's Office	Department Store	Dessert Shop	Donut Shop	Electronics Store	Farmers Market	23.6076	90.125
26	Joypara	Narisha	1332	2	Big Box Store	Turkish Restaurant	Fried Chicken Joint	Cricket Ground	Dentist's Office	Department Store	Dessert Shop	Donut Shop	Electronics Store	Farmers Market	23.6076	90.125

Figure 3Cluster 2

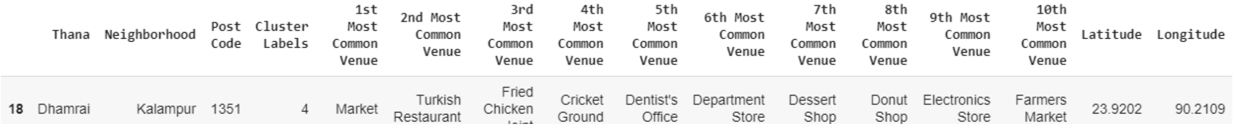
In this study, Cluster_3 consists of one thana, Savar, with ten neighborhoods Amin Bazar, Dairy Farm, Saver P.A.T.C, Rajphulbaria, EPZ, Kashem Cotton Mills, Jahangirnagar University, Savar Cantonment, Savar, and Shimulia. The most common items in this location are Shopping mall, Bus station, and market



	Thana	Neighborhood	Post Code	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	Latitude	Longitude
33	Savar	Saver P.A.T.C	1343	3	Shopping Mall	Bus Station	Market	Turkish Restaurant	Food	Cricket Ground	Dentist's Office	Department Store	Dessert Shop	Donut Shop	23.8477	90.2587
29	Savar	Rajphulbaria	1347	3	Shopping Mall	Bus Station	Market	Turkish Restaurant	Food	Cricket Ground	Dentist's Office	Department Store	Dessert Shop	Donut Shop	23.8477	90.2587
35	Savar	Shimulia	1345	3	Shopping Mall	Bus Station	Market	Turkish Restaurant	Food	Cricket Ground	Dentist's Office	Department Store	Dessert Shop	Donut Shop	23.8477	90.2587
32	Savar	Savar Canttonment	1344	3	Shopping Mall	Bus Station	Market	Turkish Restaurant	Food	Cricket Ground	Dentist's Office	Department Store	Dessert Shop	Donut Shop	23.8477	90.2587
31	Savar	Savar	1340	3	Shopping Mall	Bus Station	Market	Turkish Restaurant	Food	Cricket Ground	Dentist's Office	Department Store	Dessert Shop	Donut Shop	23.8477	90.2587

Figure 4 Cluster 3

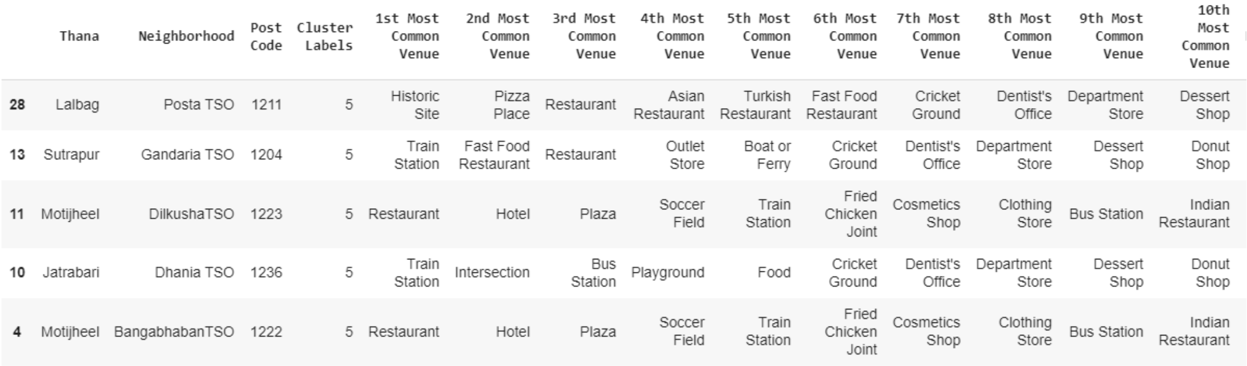
The Dhamrai thana's Kalampur is a cluster of its own, we can say this is an anomaly. the most common item of kalampur is market, Turkish Restaurant, and fried chicken joint.



	Thana	Neighborhood	Post Code	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	Latitude	Longitude
18	Dhamrai	Kalampur	1351	4	Market	Turkish Restaurant	Fried Chicken Joint	Cricket Ground	Dentist's Office	Department Store	Dessert Shop	Donut Shop	Electronics Store	Farmers Market	23.9202	90.2109

Figure 5 Cluster 4

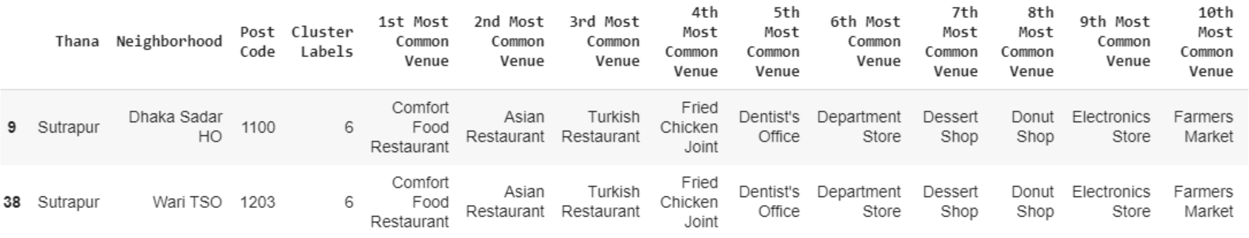
Cluster 5 is a little different it has some of the historical sites. We call this cluster old Dhaka. It consists of Lalbag, Sutrapur, Motijheel and Jatrabari thana. The neighborhoods analyzed are Posta, Gandaria, Dilkusha, Dhania, and Bangabhaban. The most common items in this cluster are train stations, restaurants, and historic sites.



	Thana	Neighborhood	Post Code	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
28	Lalbag	Posta TSO	1211	5	Historic Site	Pizza Place	Restaurant	Asian Restaurant	Turkish Restaurant	Fast Food Restaurant	Cricket Ground	Dentist's Office	Department Store	Dessert Shop
13	Sutrapur	Gandaria TSO	1204	5	Train Station	Fast Food Restaurant	Restaurant	Outlet Store	Boat or Ferry	Cricket Ground	Dentist's Office	Department Store	Dessert Shop	Donut Shop
11	Motijheel	DilkushaTSO	1223	5	Restaurant	Hotel	Plaza	Soccer Field	Train Station	Fried Chicken Joint	Cosmetics Shop	Clothing Store	Bus Station	Indian Restaurant
10	Jatrabari	Dhania TSO	1236	5	Train Station	Intersection	Bus Station	Playground	Food	Cricket Ground	Dentist's Office	Department Store	Dessert Shop	Donut Shop
4	Motijheel	BangabhabanTSO	1222	5	Restaurant	Hotel	Plaza	Soccer Field	Train Station	Fried Chicken Joint	Cosmetics Shop	Clothing Store	Bus Station	Indian Restaurant

Figure 6Cluster 5

Cluster 6 has two neighborhoods and they are both from sutrapur thana, they are Dhaka Sadar and wari. the most common item in this cluster in a food restaurant.



	Thana	Neighborhood	Post Code	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
9	Sutrapur	Dhaka Sadar HO	1100	6	Comfort Food Restaurant	Asian Restaurant	Turkish Restaurant	Fried Chicken Joint	Dentist's Office	Department Store	Dessert Shop	Donut Shop	Electronics Store	Farmers Market
38	Sutrapur	Wari TSO	1203	6	Comfort Food Restaurant	Asian Restaurant	Turkish Restaurant	Fried Chicken Joint	Dentist's Office	Department Store	Dessert Shop	Donut Shop	Electronics Store	Farmers Market

Figure 7 Cluster 6

Chapter 4 Discussion and Conclusion

Our analysis shows clusters of similar locations in Dhaka Bangladesh. based on their distinctive venue categories. We have used 1000 maxim iterations for the k mean algorithm and the SSE is bellow 1. So we can say that the clusters are of acceptable accuracy.

some clusters have values everything in common. this may be possible due to having the same latitude and longitude or the range is high. so the foursquare API listed the same items and they were sorted into the same cluster. This project doesn't show any visualization maybe in the future I will learn more and create more promising projects to showcase.

Conclusion

the purpose of this project was to do everything a data scientist does every day in his life. The project gives enough opportunity to learn and exercise exciting new things. though this project is not impactful enough, it is a start. I want to thank IBM and Coursera to bring such a good course for learning data science.