Introduction

Bangladesh is a country of over 150 million people with a GDP of 250 US Billion dollars. After its independence from Pakistan in 1971, Bangladesh has moved itself from being an under-developed to a developing country within a span of 30 years. As the economy has grown and still growing at a fast pace (600% increase in GDP within 60 years), the buying power of general people is also increasing rapidly. One of the indicators of this event is the burgeoning growth of restaurants around the country, especially in Dhaka, the Capital city of Bangladesh. While it was hard to find a decent coffee place around the capital around 1980-90s, now-a-days big chains like Starbucks, Gloria-Jeans are found at every corner of the city.

In a statistical report by Bangladesh Bureau of Statistics, it is cited that the restaurant market in Bangladesh will reach a whopping 56 Million USD, contributing around 2.1% of the total GDP of Bangladesh by the year 2021. Several social changes can be selected as the reason of this upward trend. As the economy is growing, most of the people in the city are working and they are looking for affordable options for their daily eating routine. At the same time, people under 25, which is 50% of the total population, prefers the ever-growing market of fast foods. Another interesting reason can be attributed to the dating habit of the younger generation. In Dhaka, it is hard to find a calm and quiet place where you can spend time with your loved ones as it is the most densely populated city in the world. The restaurants provide an alternative where people can go spend some quality time with their better halves.

Dhaka, being the most densely populated city in the world, is a place where the availability of jobs is scarce compared to the number of people. That is why, a great number of people is looking for entrepreneurship opportunities to make a living. Restaurant business, being one of the most burgeoning market in Dhaka, is where the people are trying to invest most. In this report, I plan to identify the preference of people around different neighborhoods in Dhaka city over different types of restaurants. I believe, this report will help a lot of young and upcoming entrepreneurs who are looking to invest in restaurant business around the neighborhood they are living in, to help them decide on the type of restaurant they should invest.

Data

The Datasets I have used for this exercise are the following:

1. List of postal codes in Bangladesh: List of postal codes in Bangladesh are available in the following Wikipedia page. I will only use the postal codes for Dhaka district as I will plan to identify the preferences of people over different restaurants around Dhaka district. https://en.wikipedia.org/wiki/List_of_postal_codes_in_Bangladesh
2. Geocoder: I have used the geocoder from geopy library to collect the latitude and longitude information.
3. Foursquare data: I will use foursquare to data to find out relevant information about the restaurants around Dhaka district, i.e., location, restaurant type, etc.

The steps I have followed to perform data collection and cleaning are as follows:

1. Import the relevant libraries
2. Read the table from the Wikipedia page
   Using BeatifullSoup, I read the table from the Wikipedia page containing the postal codes of Dhaka district.
3. Transform the table into pandas data frame

   In this section, I transform the wiki table into data frame. I also create a feature "ADDRESS" using the columns of the data frame. This feature will be used to extract the latitude and longitude values of the locations.

4. Collect Latitude and Longitude information for the locations
   In this section, I collect the lattitude and longitude of the locations saved as the feature "ADDRESS" in the data frame "dhaka". I have used the library "Geopy" to conduct this exercise. After collecting the information, I have appended the information as new features in my original data frame "dhaka".
   To learn about the procsess further, plick in the following link: https://towardsdatascience.com/geocode-with-python-161ec1e62b89.

Methodology

Data visualization

Create initial map of Dhaka city

Now that I have all the relevant data for this exercise, I have created initial visualizations on the spatial data in the data frame. At first, I have created a folium map where the addresses are shown in blue markers. Please see figure 1. This map gives us a general idea about the sparsity of the locations.

It can be easily seen that most of the locations are concentrated around Dhaka metropolitan area. As we move away from the metro area, the locations get sparse. This observation is intuitive and logical.

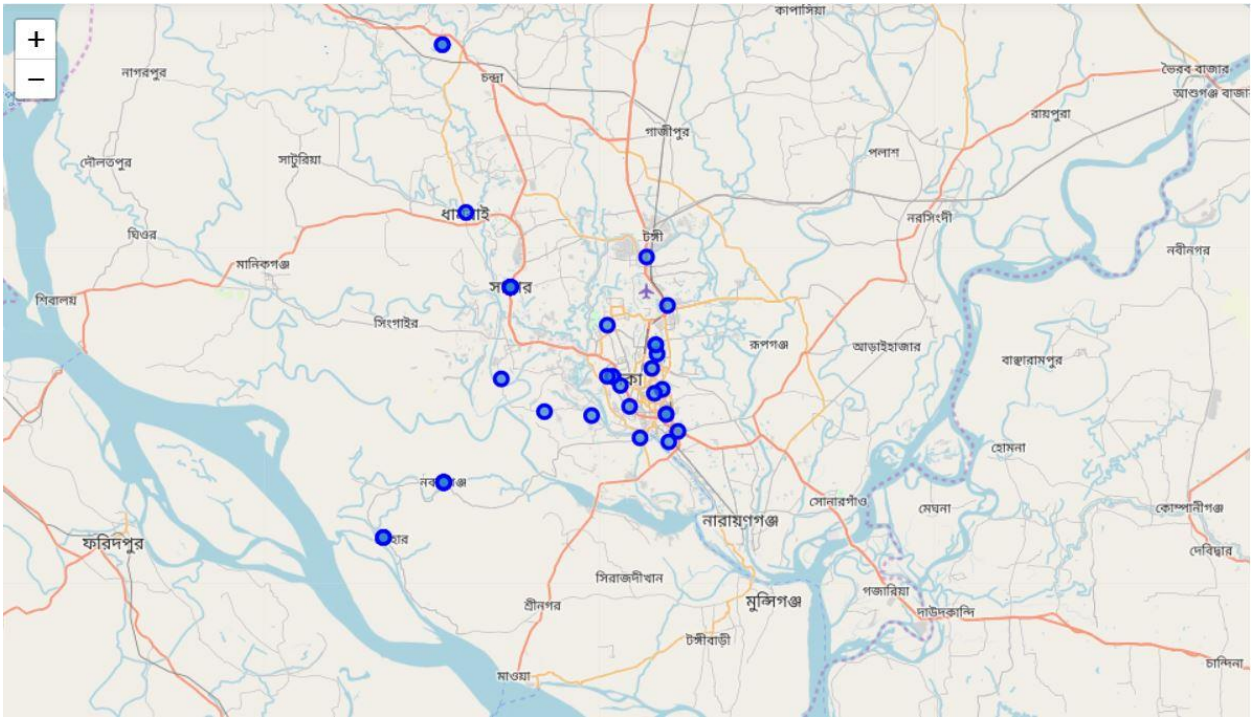

Figure 1: Restaurants in Dhaka district

Collect venues data using Foursquare

After creating some initial visualizations, I have collected the venue data around the location in the data frame 'dhaka' using "Foursquare" API. Figure 2 has a snapshot of venues for some addresses enlisted on the data frame “dhaka”.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	01350, Dhamrai, Dhaka, Bangladesh	23.920162	90.210870	Dhamrai Bazar	23.919938	90.211445	Market
1	1209, Dhanmondi, Dhaka, Bangladesh	23.753550	90.373124	Sausly's	23.755445	90.375762	Sandwich Place
2	1209, Dhanmondi, Dhaka, Bangladesh	23.753550	90.373124	Drik Gallery	23.752380	90.369972	Art Gallery
3	1209, Dhanmondi, Dhaka, Bangladesh	23.753550	90.373124	Nando's	23.753045	90.369766	Portuguese Restaurant
4	1209, Dhanmondi, Dhaka, Bangladesh	23.753550	90.373124	BFC	23.755495	90.375534	Fried Chicken Joint
5	1209, Dhanmondi, Dhaka, Bangladesh	23.753550	90.373124	Dhanmondi Rd 27	23.753419	90.372098	Scenic Lookout

Figure 2: Venues around Dhaka City

From the collected venues information, it can be observed that only 248 venues are found for 46 locations in the data frame 'dhaka'. This is also logical. "Foursquare" is not popular in Bangladesh or around Dhaka city. That is why, the data collected from FourSquare is quite small. This is one of the setbacks of this exercise as 246 is too small a sample size for representing the total number of establishments situated in Dhaka.

One-hot encoding

As the venues around Dhaka can be divided into 56 unique categories, I have applied one-hot encoding to the data, so that proper clustering method can be applied. The final matrix obtained after one-hot encoding has a shape of 248*57.

Clustering

Before applying clustering algorithms, I have developed a data frame that contains the top-10 venues around each address enlisted in the data frame “dhaka”. Then I have merged the new data

frame with the data frame “dhaka” so that I can apply K-means clustering. Figure 3 shows a snapshot of the merged data frame.

	District	Thana	SubOffice	Post Code	ADDRESS	location	point	latitude	longitude	altitude	...	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Dhaka	Dhamrai	Kalampur	01350	01350, Dhamrai, Dhaka, Bangladesh	(ধামরাই, ধামরাই উপজেলা, ঢাকা জেলা, ঢাকা বিভাগ,...	(23.920162, 90.2108702, 0.0)	23.920162	90.210870	0.0	...	Market	Turkish Restaurant	Hobby Shop
1	Dhaka	Dhanmondi	Jigatala TSO	1209	1209, Dhanmondi, Dhaka, Bangladesh	(ধানমন্ডি আ/এ, ঢাকা, ঢাকা জেলা, ঢাকা বিভাগ, 12...	(23.7535496, 90.37312384681817, 0.0)	23.753550	90.373124	0.0	...	Art Gallery	Asian Restaurant	Shopping Mall
2	Dhaka	Gulshan	Banani TSO	1213	1213, Gulshan, Dhaka, Bangladesh	(গুলশান, ঢাকা, ঢাকা জেলা, ঢাকা বিভাগ, 1213, Ba...	(23.78346, 90.4122658, 0.0)	23.783460	90.412266	0.0	...	Café	Italian Restaurant	Hotel
3	Dhaka	Gulshan	Badda	1212	1212, Gulshan,	(গুলশান, ঢাকা, ঢাকা জেলা,	(23.7930078,	23.793008	90.410661	0.0	...	Indian	Café	Hotel

Figure 3: Merged data frame

The merged data frame contained the information from the data frame “dhaka” and the most common categories of restaurants around each address. I have applied K-means clustering in the merged dataset with a maximum of 5 possible clusters.

Result

The resultant clusters are shown in figure 4. Applying K-means clustering, I have been able to create 4 different clusters for the addresses listed in the data frame “dhaka”

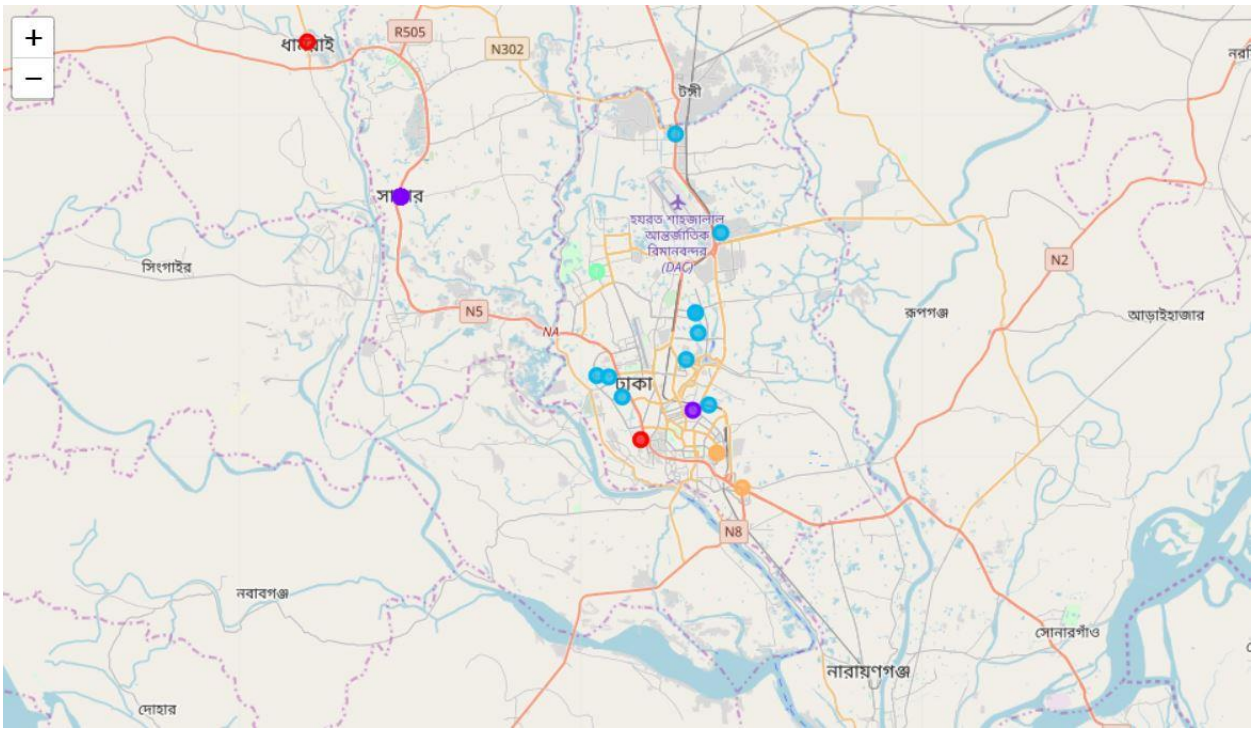

Figure 4: Clustered addresses

Observation

1. First, the datapoints are sparse as it can be seen from figure 4. One of the reasons for this problem can be attributed to the unpopularity of FourSquare in Bangladesh. One way to resolve this issue is to use data source that is used by most people in Bangladesh.
2. Cluster 2 is the most spread-out cluster.

Conclusion

This exercise was aimed to create a cluster of different restaurant venues situated across the district of Dhaka, the capital of Bangladesh. At first the postal codes of Dhaka district were collected from Wikipedia using "BeautifulSoup". Then the latitude and longitude data were collected using the python library "GeoPy". After the dataset had complete information about the latitudes and longitudes of each location, the restaurant venues were collected using "FourSquare" API. Then K-Means clustering was applied to create at most 5 clusters of the dataset. From the map above, we can see that, cluster-2 is the most spread-out cluster among others.

The clustering of the locations of different venues will help the young entrepreneurs around Dhaka city to decide on the type of restaurant to invest on around their neighborhood. One possible direction for future work is to use a data-source for venues that has more data than the Foursqure platform. Dhaka district most certainly has more than 246 restaurants. The more data we can collect about the restaurants around Dhaka district, the clustering process will provide more tangible separation.