Clustering Venues of Dhaka

Abidur Rahman

June 8, 2020

# Introduction

## Background:

Dhaka is the capital and the largest city of Bangladesh. It is the ninth-largest, and the sixth-most densely populated city in the world, with a population over 21 million in the Greater Dhaka Area. It is the largest city in Eastern South Asia. As a resident of this city, I decided to do my project on Dhaka.

## Problem:

Despite being a large city, a metropolitan, Dhaka doesn’t have enough representation in the databases. Before Android smartphones became a thing, was nowhere to be found in any kind of database. Even so many years later, there is not enough data or data analysis about Dhaka, especially about the different venues in Dhaka. However, to start a new business, it is very important to have adequate data, refined data about the neighborhood. So an analysis of the type of venues in an area, and the competition level, turn out to be very important. Considering all of these, I decided to create a map of different venues in Dhaka, and cluster them using unsupervised machine learning.

## Interest:

This clustering will help us figure out the pattern in different areas, which can later help to determine the types of businesses to be opened there. This analysis will be important to any entrepreneur who wants to open a new business and needs to know which location would be the best.

# Data

For this project, I will need the following data:

* Google Maps
  + for finding nearby venues and some details about them
* Wikipedia
  + details about different areas of Dhaka
* Foursquare API
  + common venues of Dhaka with location and other details

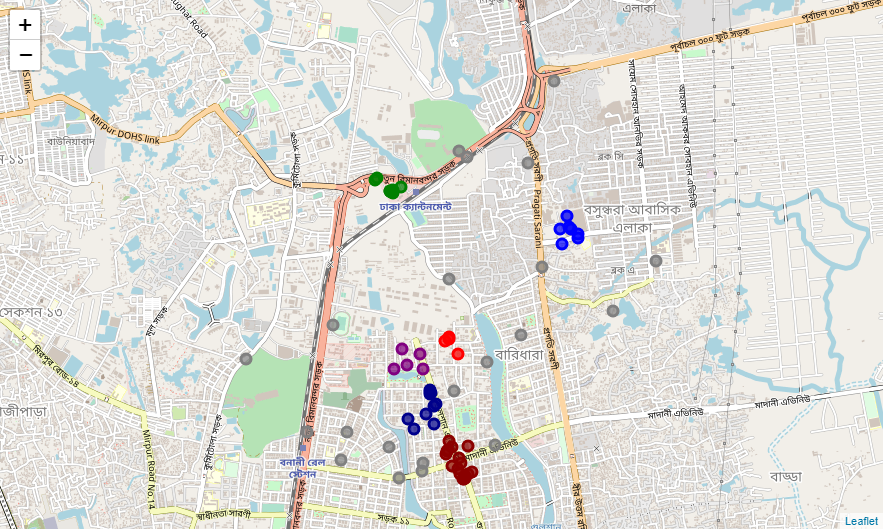
# Methodology

At first I collected the data from Foursquare API, which contained the venues with their locations in longitude and latitude, and their categories.

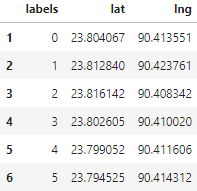
Then I used Machine Learning to divide the data into a few clusters and outliers. DBSCAN was perfect for this, as I was looking for dense clusters and that’s what DBSCAN does, look for dense clusters.

DBSCAN requires two parameters, epsilon and min samples. Optimizing epsilon was important. KNN distance can be used to determine the epsilon, however this was not applicable to this dataset (details in code). So epsilon was chosen using trial and error method.

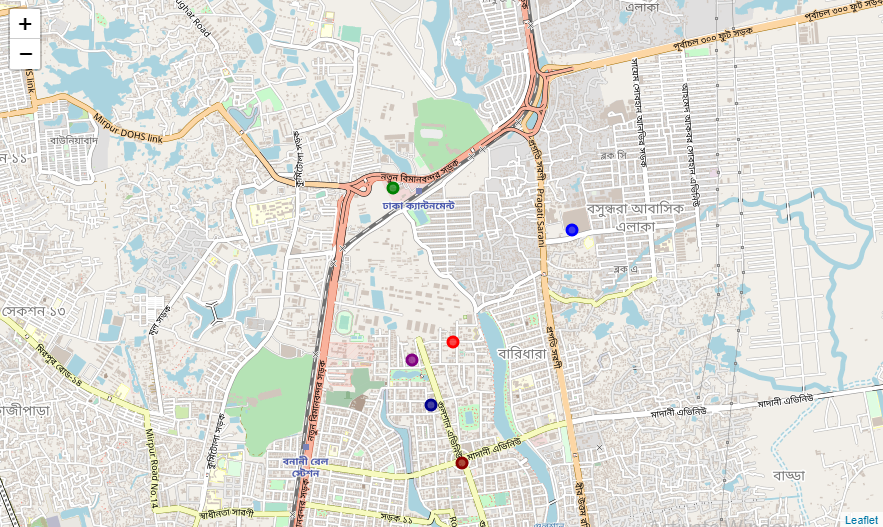
After clustering the data, the dataset was displayed on a map using different color for different clusters.



After that, the center of each cluster was determined.



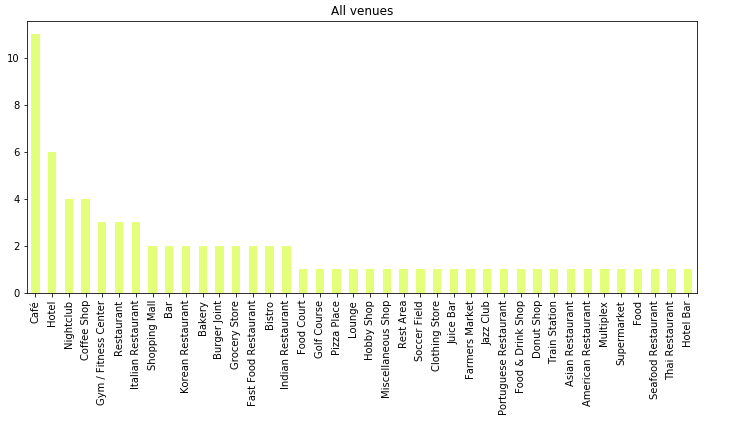
These locations were also mapped separately.

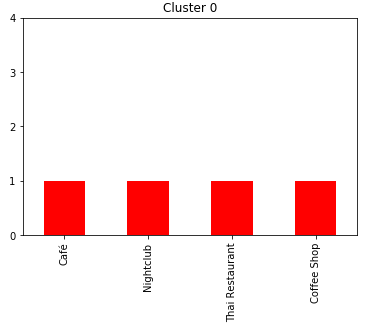


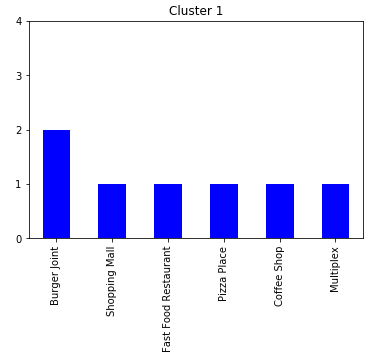
This map shows the location of the center of each of the clusters and the colors used here are used for the graphs also.

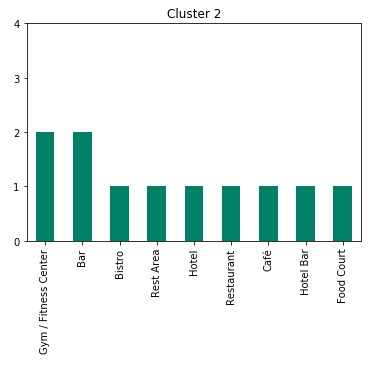
# Results

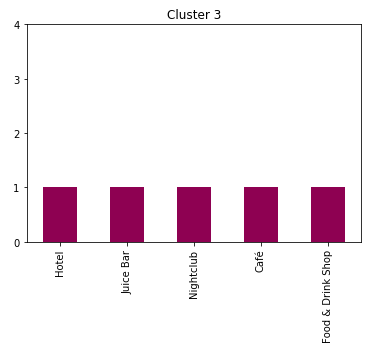
As the answer to the main question, I determined the type of venues among all the venues, then all the clusters, and then among the outliers, and plotted them in bar charts.

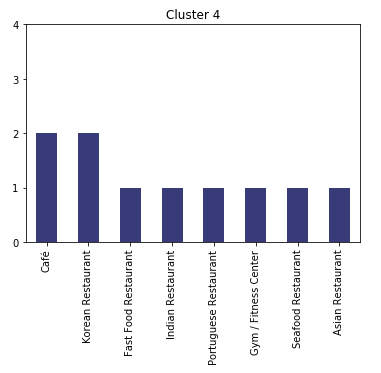


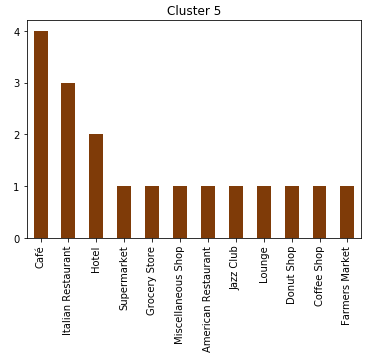


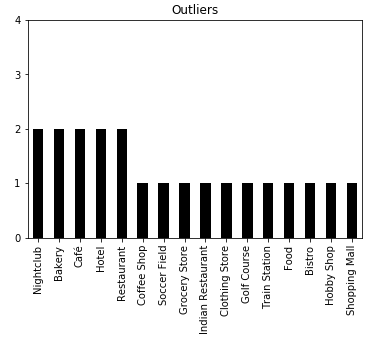












This data shows that cluster 5 is largest/densest, and cluster 0 is smallest. Checking from Google Maps and Wikipedia, we can see that this is the truth.

# Discussion

As I mentioned before, Dhaka is a big city but the data available is sparse. As a result, the number of venues received from Foursquare API was really low. Still, I tried to assess the trend from the available data assuming that it was representative, which is most likely not true. So more data needs to be collected which can be easily clustered using the same method/code I used here.

I used DBSCAN algorithm for clustering in this study. However the optimum epsilon had to be determined by trial and error and I found 6 clusters. The 6 clusters are shown in the map.

Among the 6 cluster, cluster 5 was the largest and more varied, while cluster 0 was the opposite. This indicates the fierce competition near the area of cluster 5 and the opposite around cluster 0.

I ended the study by showing frequency of different categories in different clusters. This will help the entrepreneurs determine which business to start.

# Conclusion

As the world goes forward, there’s more need for organized data and their proper use. That’s especially true for the entrepreneurs in Dhaka as the economy is growing as new businesses open and flourish, and the need for data increases exponentially.

Hopefully more data on Dhaka will be collected so that the clustering done here becomes more appropriate and more people are benefited from the analysis.

# References

* Wikipedia
* Google Maps
* Foursquare API
* Github