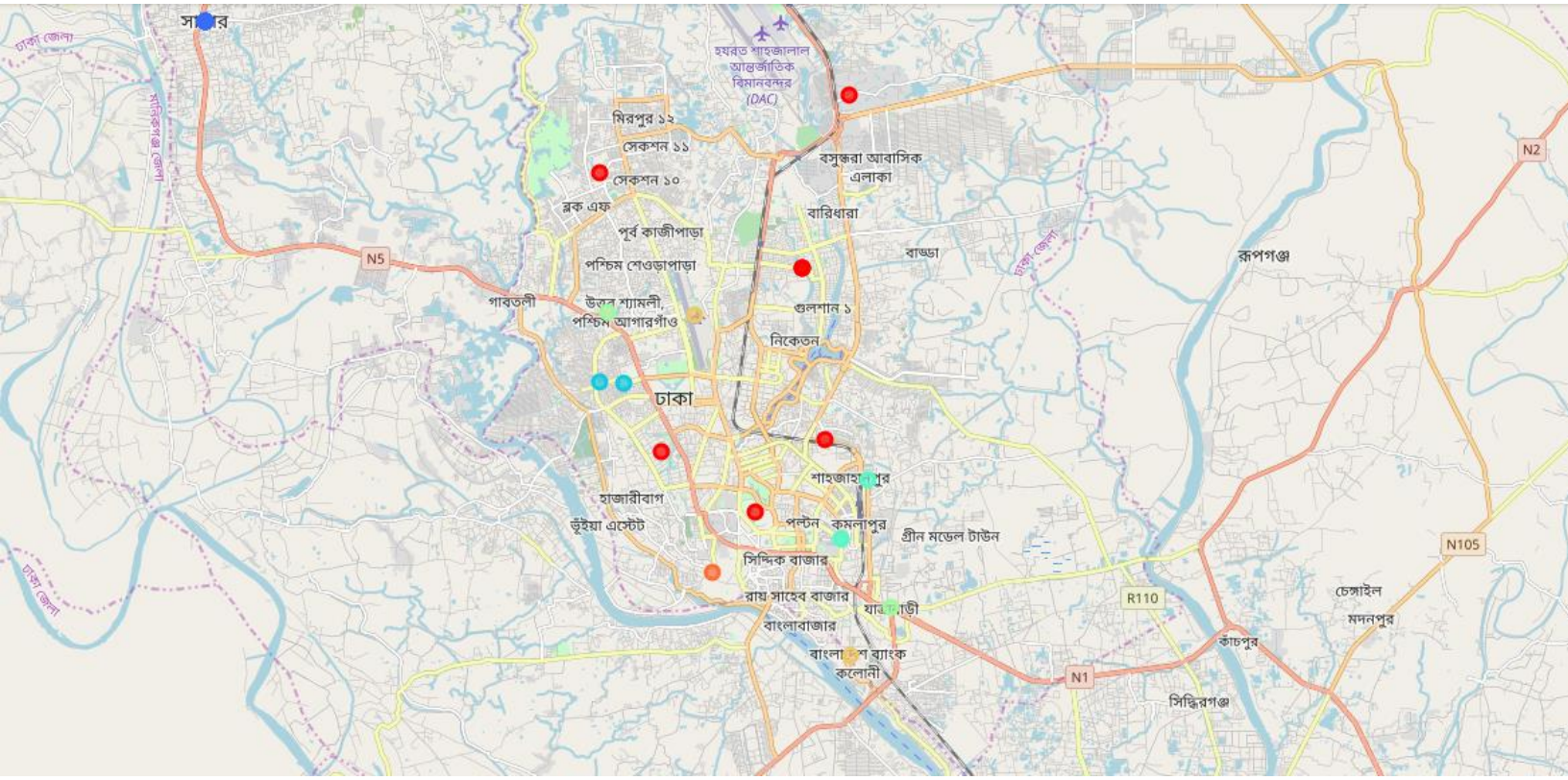
Finding a Suitable Place for Buying an Apartment in Dhaka, Bangladesh



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?

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 same features and narrow down his search for specific purpose. This project aims to provide 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 about 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 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 required information by cleaning it using `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 nearest venue data including their name, location and venue type.

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