Coursera Capstone

Applied Data Science Capstone

Project Report
On

Opening Juice Bars in Dhaka, Bangladesh

By

Amdadul Haque

April 2020

1. Introduction

Fresh fruits and vegetables can change everything. They boost our immune system, they protect us against cancer, and they help with weight management. Once consumers become aware and experience the amazing feeling that accompanies a healthy diet, their buying habits change forever. As education is making consumers extremely health conscious and hungry for a revamped diet, juice bars can be handy to meet the needs of those in their communities. However, in crowded cities like Dhaka, peoples are not habituated with taking fruits in general.

Education is changing everything. More than ever, people are aware that it is on the cusp of alarming unhealthy eating habits. In busy city life, people are aware of healthy lifestyle and food habits and rapidly changing preferences of consumers. Eventually, Juices and smoothies are becoming a vital part of consumer health and convenience.

In densely populated area like Dhaka, settings up multiple juice bar is promising. But location is a good indicator for this type of business to grow faster. It requires serious consideration of analysis of neighborhoods and a better location as well.

2. Business Plan

This capstone project will analyze the neighborhood of Dhaka city and help investors to come up with best locations to setup multiple Juice Bar in the city. Using data science methodology and machine learning techniques specially clustering, this projects aims to find best locations for opening multiple Juice Bar in Dhaka City by analyzing neighborhoods.

3. Data

Performance of machine learning algorithms highly depend of proper datasets. This project will require following datasets.

- List of neighborhood of Dhaka city
- Location information of the neighborhoods for better visualization
- Venue data specially related to Juice Bar Industries in Dhaka.

Data Acquisition

Data will be collected from multiple sources available in web using popular Python packages.

- Neighborhood data is not so much publicly available for Dhaka city. However, The Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Dhaka) of Dhaka, Bangladesh contains 42 neighborhoods. I will use BeautifulSoup and Python requests packages for scrapping.
- The geographical coordinates for the neighborhoods will be extracted using Python Geocoder Packages which will give latitude and longitude information.
- And finally venues information will be collected using Foursqure API. Foursqure has 150+ million places and used by large amount of developers around the Globe.

4. Methodology

In this capstone project course, we already explores the neighborhood data analysis of city of New York, US and Toronto, Canada. It has come up with bunch of powerful Python packages for data analysis and machine learning algorithms. To analyze statistical significance, learn datasets, and finally finds the suitable venues for open up Juice Bar in the city using below steps.

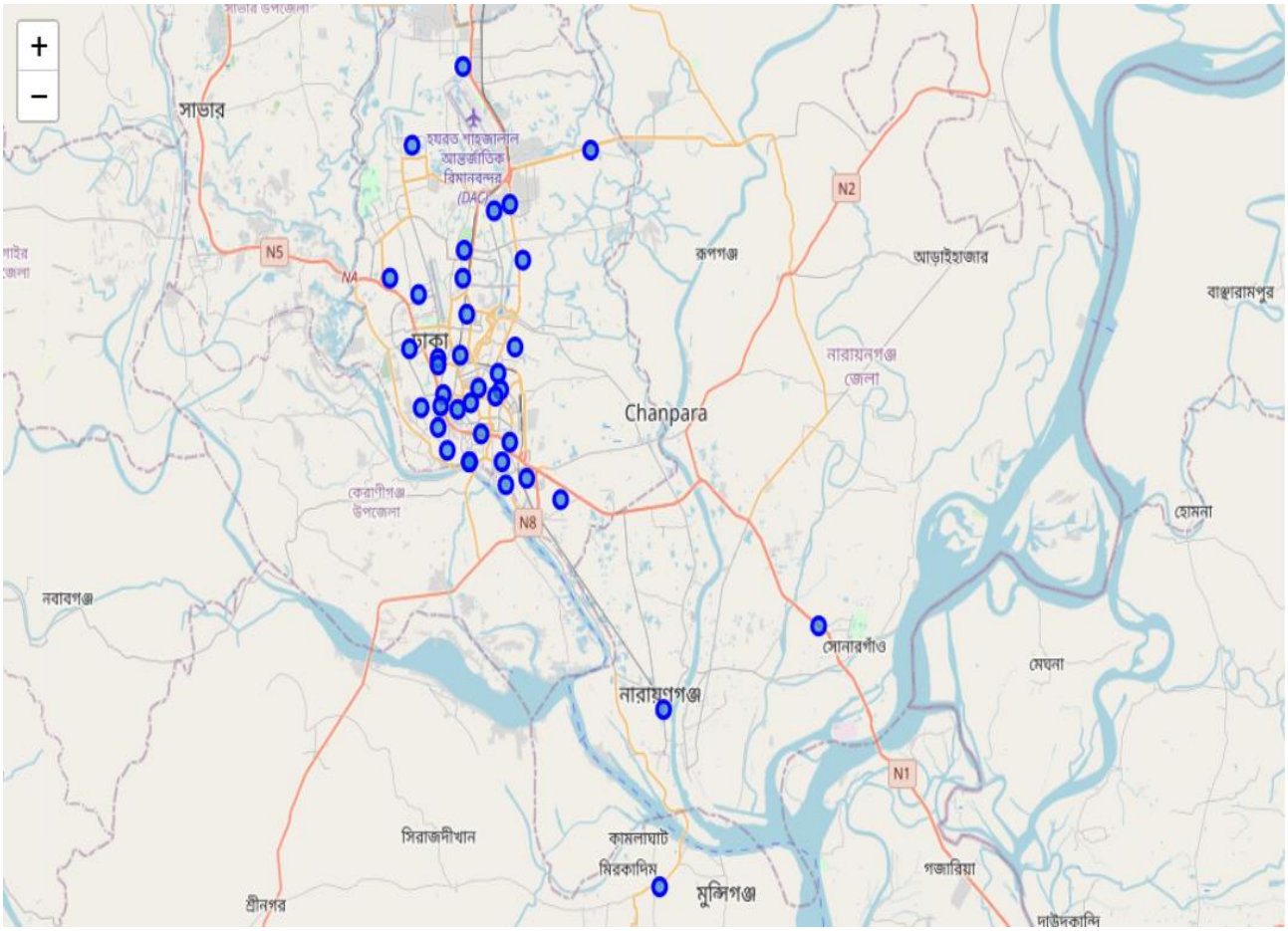
4.1 Dataset Preparation

- Download and install required Python packages i.e. Numpy, Pandas, geopy, geocoder, matplotlib, Scikit-learn, folium and BeautifulSoup.
- Scrap Wikipedia page of Dhaka City, parse using BeautifulSoup package.
- Create pandas dataframe from scrapped web data. It come up with a dataset of 42 neighborhood of Dhaka city.
- Getting the geographical data i.e. latitude, longitude of neighborhoods using geocoder package and update dataframe.

	Neighborhood	Latitude	Longitude
0	Agargaon	23.77731	90.37273
1	Armanitola	23.73895	90.38594
2	Azimpur, Dhaka	23.72612	90.38296
3	Bailey Road, Dhaka	23.74134	90.40411
4	Banani DOHS	23.79388	90.39656

4.2 Create map of Dhaka City superimposed on top

The python package folium makes it easy to visualize data that's been manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map. Dhaka city superimposed map has been created using above geographical data of the neighborhoods.



4.3 Explore the neighborhoods using Foursquare API

Foursquare is a location technology platform dedicated to improving how people move through the real world. We used Foursquare API to get the top 100 venues that are within a radius of 2000 meters and found 1598 venues information. Finally updated the dataframe from applying machine learning algorithms.

4.4 Explore the neighborhoods Data

The dataset come up with 115 unique venue categories.

```
print('There are {} uniques categories.'.format(len(df_venues['VenueCategory'].unique())))
# print out the list of categories
df_venues['VenueCategory'].unique()[:100]
```

```
There are 115 uniques categories.

array(['Electronics Store', 'Asian Restaurant', 'Theme Park',
       'Fast Food Restaurant', 'Shopping Mall', 'Steakhouse',
       'Indian Restaurant', 'Market', 'Department Store', 'Bus Line',
       'Bookstore', 'Shawarma Place', 'Bakery', 'Restaurant', 'Nightclub',
       'Art Gallery', 'Street Food Gathering', 'Bike Shop', 'Plaza',
       'Multiplex', "Dentist's Office", 'Convenience Store', 'BBQ Joint',
       'Café', 'History Museum', 'Fried Chicken Joint', 'Clothing Store',
       'Park', 'Ice Cream Shop', 'Italian Restaurant', 'Pharmacy',
       'Pizza Place', 'Food', 'Burger Joint', 'Thai Restaurant',
       'Scenic Lookout', 'Chinese Restaurant', 'North Indian Restaurant',
       'Bistro', 'Hotel', 'Diner', 'Boutique', 'Recreation Center',
       'Food Court', 'Soccer Field', 'Arts & Crafts Store', 'Shoe Store',
       'Historic Site', 'Coffee Shop', 'Bus Station',
       'Other Great Outdoors', 'Performing Arts Venue',
       'American Restaurant', 'Flea Market', 'Fish Market', 'Juice Bar',
       'Golf Course', 'Seafood Restaurant', 'Gym / Fitness Center',
       'Japanese Restaurant', 'Supermarket', 'Hobby Shop', 'Lounge',
       'Turkish Restaurant', 'Mexican Restaurant', 'Sushi Restaurant',
       'Donut Shop', 'Grocery Store', 'Portuguese Restaurant',
```

4.4 Explore the neighborhoods Data

- Perform data preprocessing of categorical data like venues using one hot encoder algorithm.
- Analyze each neighborhood data and group venue categories by neighborhood and by taking the mean of the frequency of occurrence of each category.
- Create a new DataFrame for Juice Bar data only

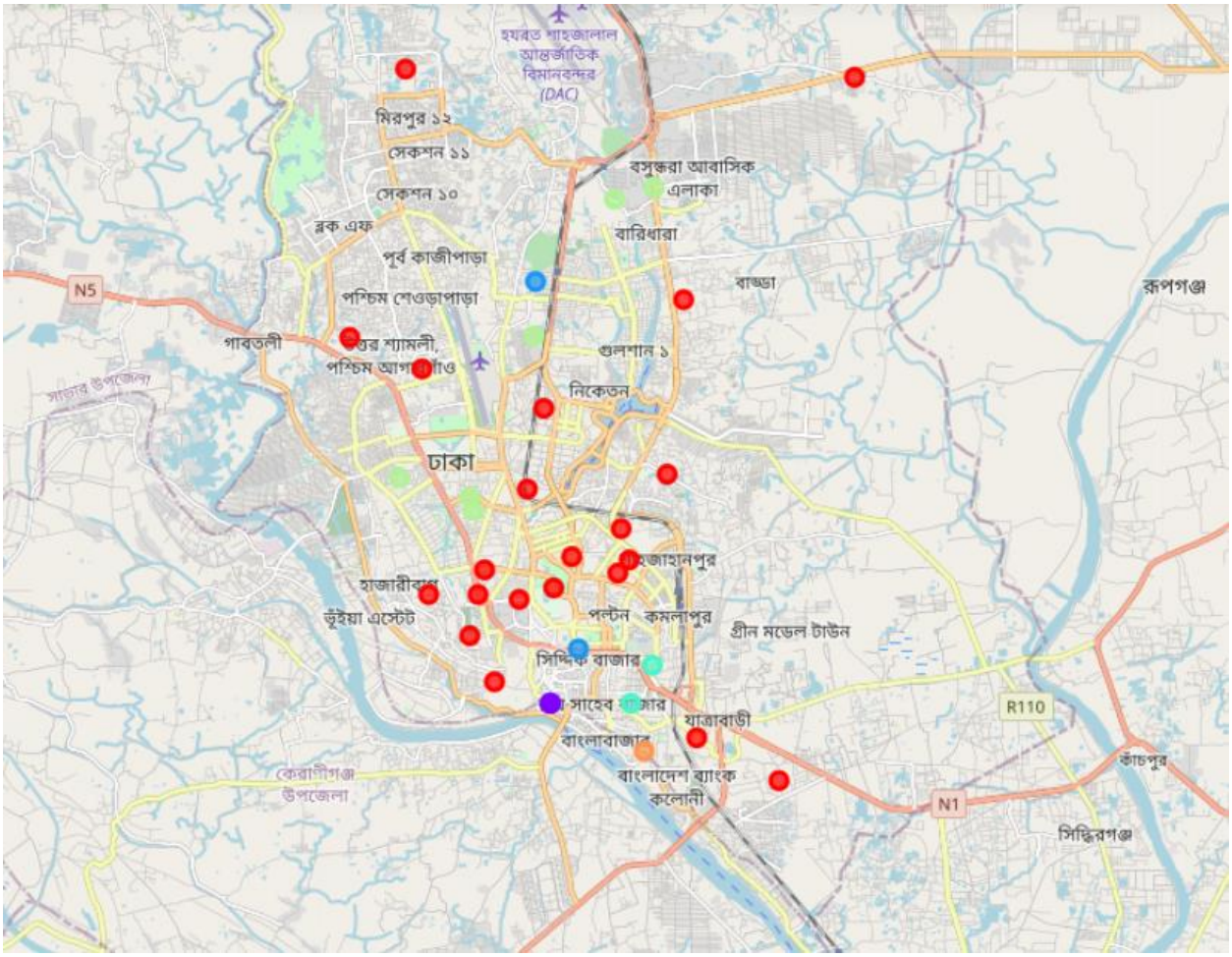
4.5 Clustering using KMeans Algorithm

- We run k-means to cluster the neighborhoods in Dhaka into 6 clusters on newly created grouped dataset of Juice Bar.
- Fit the algorithm on dataset
- Find the clusters with clusters levels.

5. Results and Findings

Results of 6 clusters of k-means shows that we have found the venues of the city into the similar categories in same clusters based on the frequencies of Juice Bar in the neighborhoods. Clusters are represented in different colors superimposed in the map.

- Cluster 0: It is the biggest cluster but has **ZERO** concentration of Juice Bars in the neighborhoods.
- Cluster 1: This cluster has high concentration number of Juice Bars in the neighborhoods
- Cluster 2 and Cluster 3: Moderate Concentration
- Cluster 4 and Cluster 5: These clusters are small and has low concentration of Juice Bars in the neighborhoods.



6. Recommendation and Future Aspects

As results and findings are represented in section 5, there are no Juice Bar in the city of cluster 0. The biggest cluster has no juice bar which is a good indication that business prospects are very promising in these areas. Any investor can open up multiple locations as marked using red color in the map. On the other hand, It depicts that cluster 1 to cluster 5 has very few number of Juice bar with good concentration. That is, Juice bar is quite popular some areas neighborhoods in Dhaka city. So opening up multiple Juice bar in the areas of cluster 0 with have chance of flourishing business.

We found only 42 neighborhood information in the Dhaka city but the city is quite bigger and mostly populated. It would be fine if data can be collected from multiple sources and compiled. And this project uses Foursquare Sandbox account having limited number of API calls and result returned. Enterprise account could be very handy of results like analyzing users in the neighborhood and come up with more specified areas.

7. Conclusion

In this work, a prospect of opening Juice bar in the Dhaka city with most popular clustering algorithm and tools. From identifying business area, extract data from available sources, analyze dataset, explore location API and finally found cluster that has zero concentration of Juice bar in cluster 0 and It also depicts that Juice bar is quite popular in the city. Thus, opening multiple Juice bar has a good prospect of business expansion. Investors can come up with this plan of opening Juice bar in cluster 0 areas.

8. References

1. Category:Neighbourhoods in Dhaka, WiKipedia, https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Dhaka

2. Foursquare Developers Documentations.

https://developer.foursquare.com/docs/

3. Applied Data Science Capstone projects of week 1-4