COURSERA CAPSTONE PROJECT: APPLIED DATA SCIENCE

A perspective view on local businesses in Dhaka, Bangladesh

MD ISLAM
SEATTLE, WASHINGTON, USA
tomislam86@gmail.com
https://www.linkedin.com/in/bornohin/

Overview

- Introduction
- Business Problem
- Data
- Methodology
- Results
- Discussion
- Conclusion

Introduction

- Dhaka is the capital of Bangladesh and is also the most densely populated (estimated 8.5 million) city in Bangladesh.
- Dhaka is also the most diverse place in Bangladesh. And with lot of diversity and population comes lot of business opportunities.
- Purpose of this project is to find the target audience and type of business one should consider using exploratory data analysis method.

Business Problem

- Pick an area of interest,
- Explorer businesses in this location,
- What different business are offered in this region,
- Is it possible to do segmentation and clustering based on number of similar businesses?
- What business ideas can come up with the most common venues in these locations?

Data

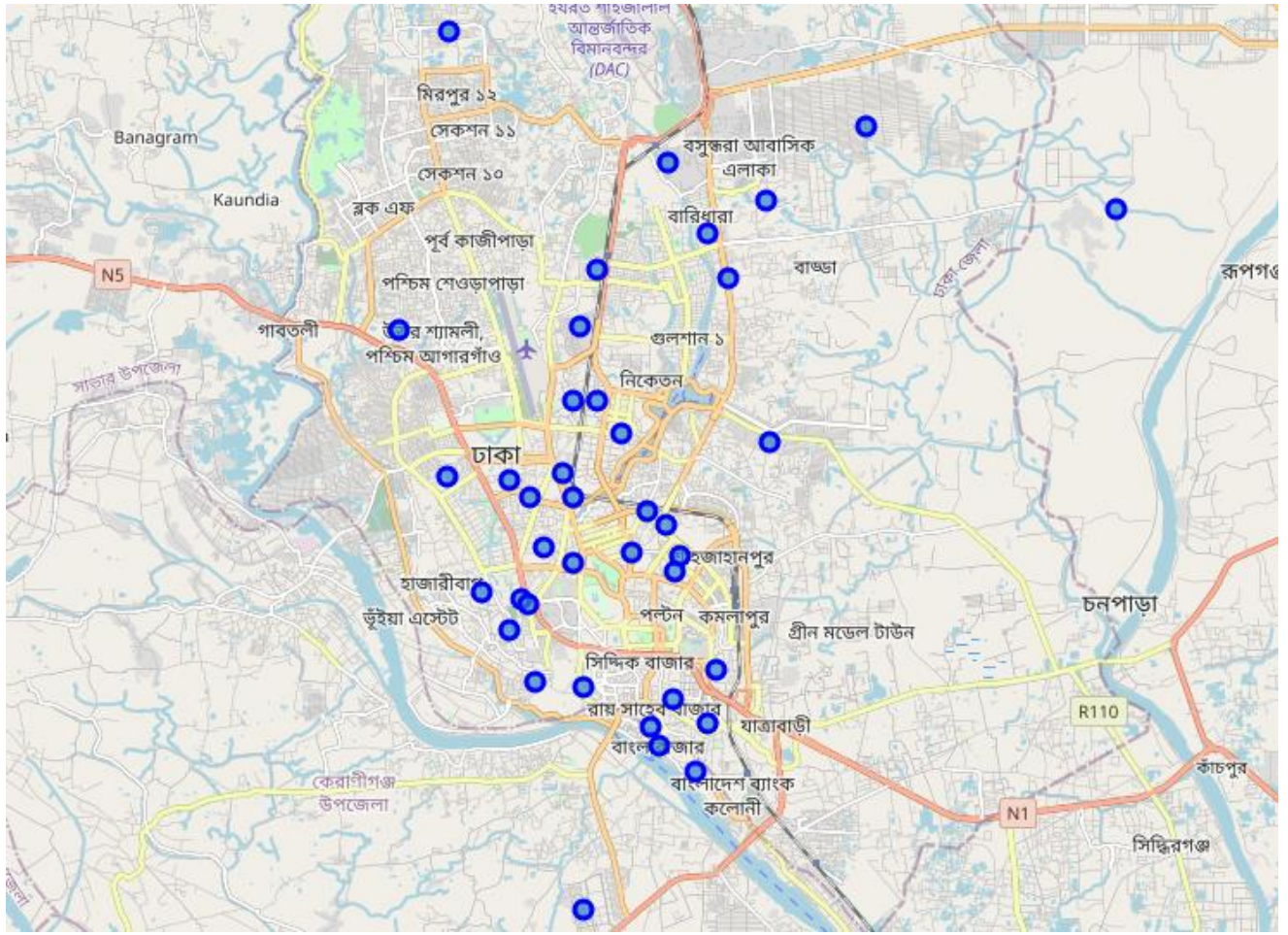
- Neighbourhoods
  - Data on neighborhoods of Dhaka was scraped from Wikipedia page using BeautifulSoup and saved into dhaka.csv
- Geocoding
  - We used Google Maps Geocoding API for these neighbourhoods to find associated latitude and longitude and saved them into a Pandas DataFrame.
- Venue Data
  - We used FourSquare API for retrieving venues using the DataFrame created above passing appropriate parameters and saved them into another DataFrame.

Methodology

- Accuracy of the Geocoding API
  - After much research and different factors into consideration, Google Geocoding API was ahead of any available APIs for this project. This was also extremely easy to use and less prone to error. A collision error was also performed.
- Folium
  - All cluster visualization are done with help of Folium which in turn generates a Leaflet map made using OpenStreetMap technology.

Methodology

- Figure: Neighbourhoods of Dhaka.



Methodology

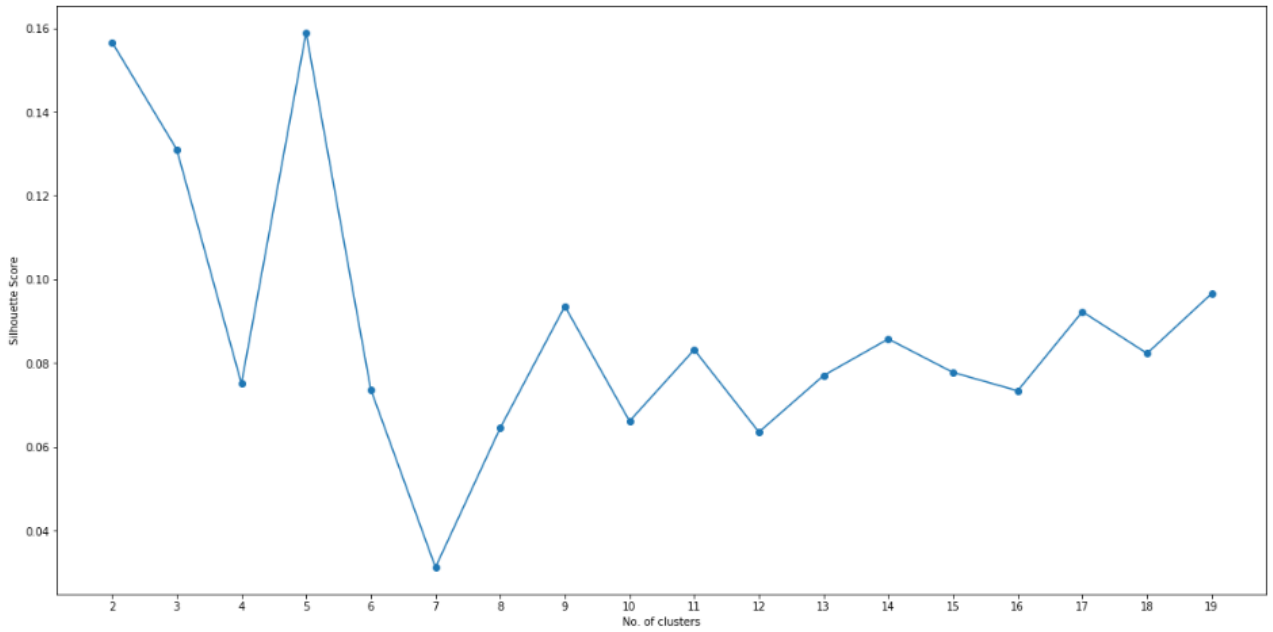
- One hot encoding
  - One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. For the K-means Clustering Algorithm, all unique items under Venue Category are one-hot encoded.
- Top 10 most common venues
  - We will try to determine the top 10 most common venues using our DataFrame. This region has a variety of venues hence only top 10 were selected to use it in K-means clustering algorithm.

Methodology

- Optimal number of clusters
  - Silhouette Score is the measurement of how similarity of an object to its own cluster (cohesion) compared to other clusters (separation). It ranges from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. Based on the Silhouette Score of various clusters below 20, the optimal cluster size is determined.
- K-means clustering
  - Venue data is trained using K-means Clustering Algorithm. This algorithm was chosen because K-means will be computationally faster than other clustering algorithms in our scenario.

Methodology

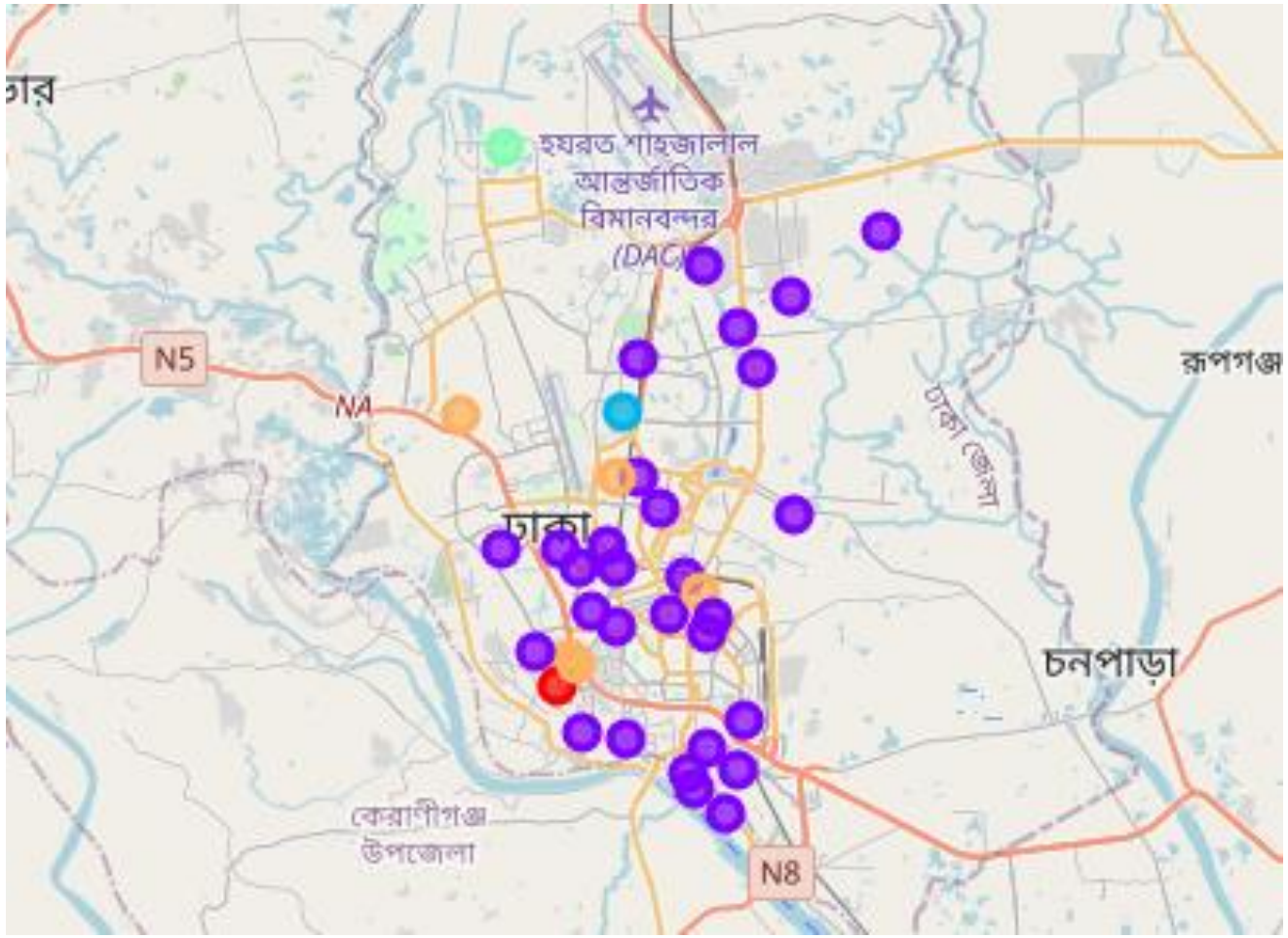
- Figure: Silhouette score vs Number of clusters.



Results

The neighbourhood are color coded in n clusters to separate from one another.

- Figure: Neighbourhoods of Dhaka (Clustered).



Discussion

- We analyzed the K-Means Clustering and it appears Shopping Mall/Markets are the most common venue. This gives us idea that a range of business services can be setup within a Mall which also can be close to a Bus Station.
- An interesting finding is the demography of these 5 locations. It ranges in the middle-class income families who can be the target audience for any product or service. This can be invaluable information for someone with a business plan who is looking for a location of the business.

Discussion

	Neighbourhood	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
15	Kallyanpur	Bus Station	Theater	Diner	Comfort Food Restaurant	Convenience Store	Cosmetics Shop	Cupcake Shop	Department Store	Dessert Shop	Electronics Store
20	Malibagh	Shopping Mall	Market	BBQ Joint	Theater	Diner	Comfort Food Restaurant	Convenience Store	Cosmetics Shop	Cupcake Shop	Department Store
23	Nakhalpara	Shopping Mall	Bookstore	Theater	Diner	Comfort Food Restaurant	Convenience Store	Cosmetics Shop	Cupcake Shop	Department Store	Dessert Shop
25	New Market, Dhaka	Market	Bus Station	Bookstore	Theater	Coffee Shop	Convenience Store	Cosmetics Shop	Cupcake Shop	Department Store	Dessert Shop
27	Nilkhet	Bus Station	Market	Bookstore	Theater	Coffee Shop	Convenience Store	Cosmetics Shop	Cupcake Shop	Department Store	Dessert Shop

Figure: Cluster having most common venue.

Discussion

- Demography of the selected 5 areas.

Area	Demography
Kallyanpur	Middle Class
Malibagh	Upper Middle Class
Nakhalpara	Lower Middle Class
New Market	Middle Class
Nilkhet	Middle Class

Conclusion

- As the current pandemic looms on us, a big part of the families who frequently commute to locations like Shopping malls or markets, will stay back at home. This is another opportunity to provide them with what they need with online services. A delivery service in these areas will prove to be extremely profitable among other possibilities. This is just an overview. This project has a lot of scopes for future growth and based on that we can further analyze and explore other possibilities.

Thank You