Coursera Capstone Project : Applied Data Science

Pritthijit Nath

Jadavpur University, Kolkata, India

pritthijit.nath@ieee.org

Overview

Introduction

- The Kolkata Suburban Railway is a suburban rail system serving the suburbs surrounding the city of Kolkata. Railways such as these are important and heavily used infrastructure in India. It is the largest suburban railway network in India by track length and number of stations.

Introduction

- Train stations are ideal locations for small businesses to set up shops, because they are hubs of human interaction where hundreds or even thousands of people day and night come and go.
- Each person in this flow of foot traffic is a potential customer who might need a specific item or purchase on impulse while waiting for a train.
- To succeed with retail at a train station, one must provide an accessible and affordable shopping experience offering merchandise or services that travellers might not quickly find elsewhere en route while travelling.

Business Problem

- Train passengers as well as station and train employees need to eat breakfast, lunch, dinner and snacks. Foods that attract busy people on the go include egg sandwiches, fries, pizza, burgers, microwaveable or cold prepared meals. Beverages such as coffee, tea, wraps, bottled water, soda and juice also sell well.
- The main objective is to find ideal spots in the city where fast food retail chains can be put up, aiming at the above demographic and maximize profits out of them.

Data

Neighbourhoods

The data of the neighbourhoods in Kolkata can be extracted out by web scraping using `BeautifulSoup` library for Python. The neighbourhood data is scraped from a `Wikipedia` webpage.

Geocoding

The file contents from `kolkata.csv` is retrieved into a `Pandas DataFrame`. The latitude and longitude of the neighbourhoods are retrieved using `Google Maps Geocoding API`. The geometric location values are then stored into the intial dataframe.

Data

Venue Data

From the location data obtained after Web Scraping and Geocoding, the venue data is found out by passing in the required parameters to the `FourSquare API`, and creating another `DataFrame` to contain all the venue details along with the respective neighbourhoods.

Methodology

Accuracy of the Geocoding API

In the initial development phase with `OpenCage Geocoder API`, the number of erroneous results were of an appreciable amount, which led to the development of an algorithm to analyze the accuracy of the Geocoding API used.

In the algorithm developed, Geocoding API from various providers were tested, and in the end, `Google Maps Geocoder API` turned out to have the least number of collisions (errors) in our analysis.

Folium

`Folium` builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of the `leaflet.js` library. All cluster visualization are done with help of Folium which in turn generates a `Leaflet` map made using `OpenStreetMap` technology.

Methodology

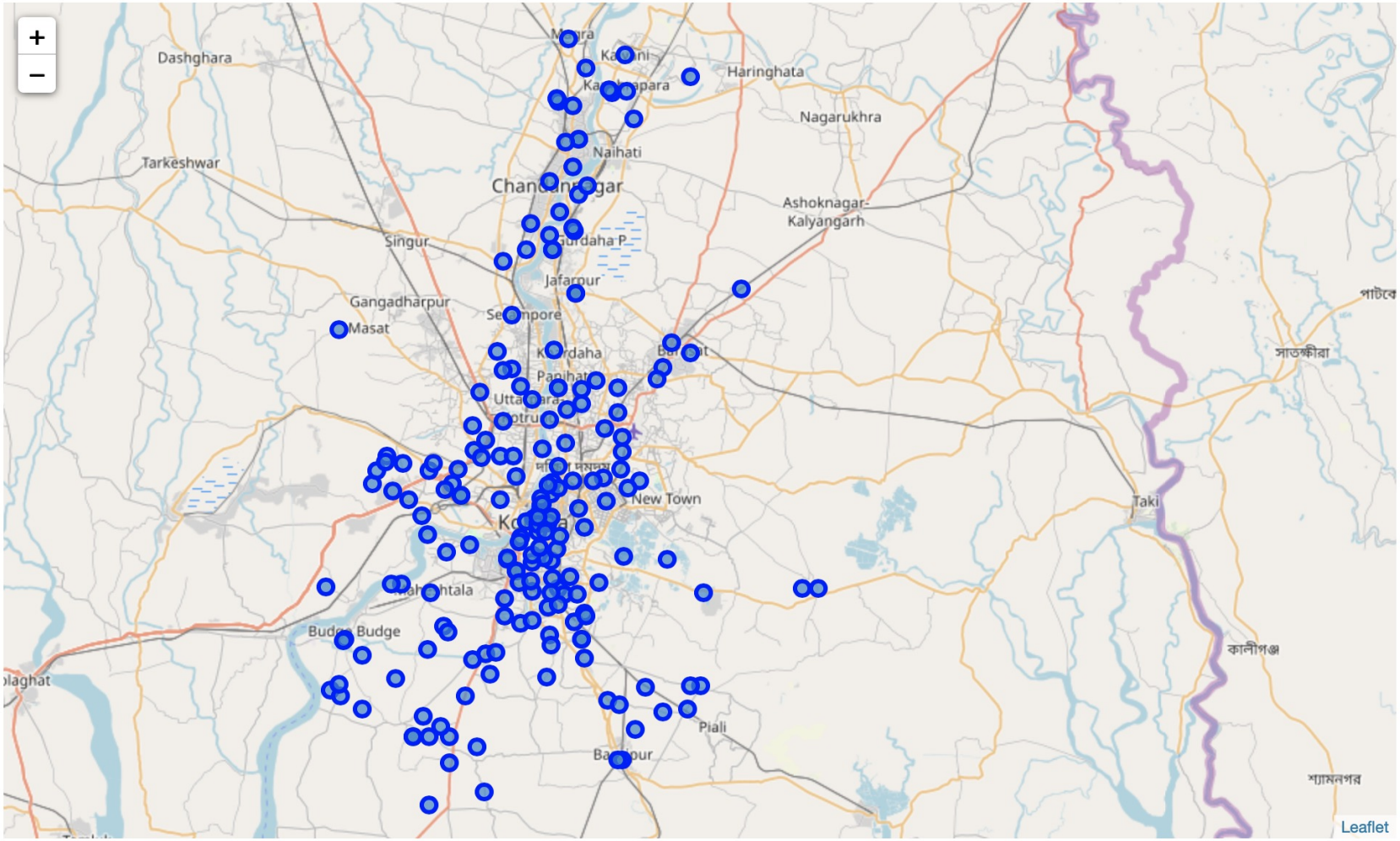

Figure: Neighbourhoods of Kolkata.

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

Due to high variety in the venues, only the top 10 common venues are selected and a new `DataFrame` is made, which is used to train the `K-means` Clustering Algorithm.

Methodology

Optimal number of clusters

`Silhouette Score` is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette 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.

Methodology

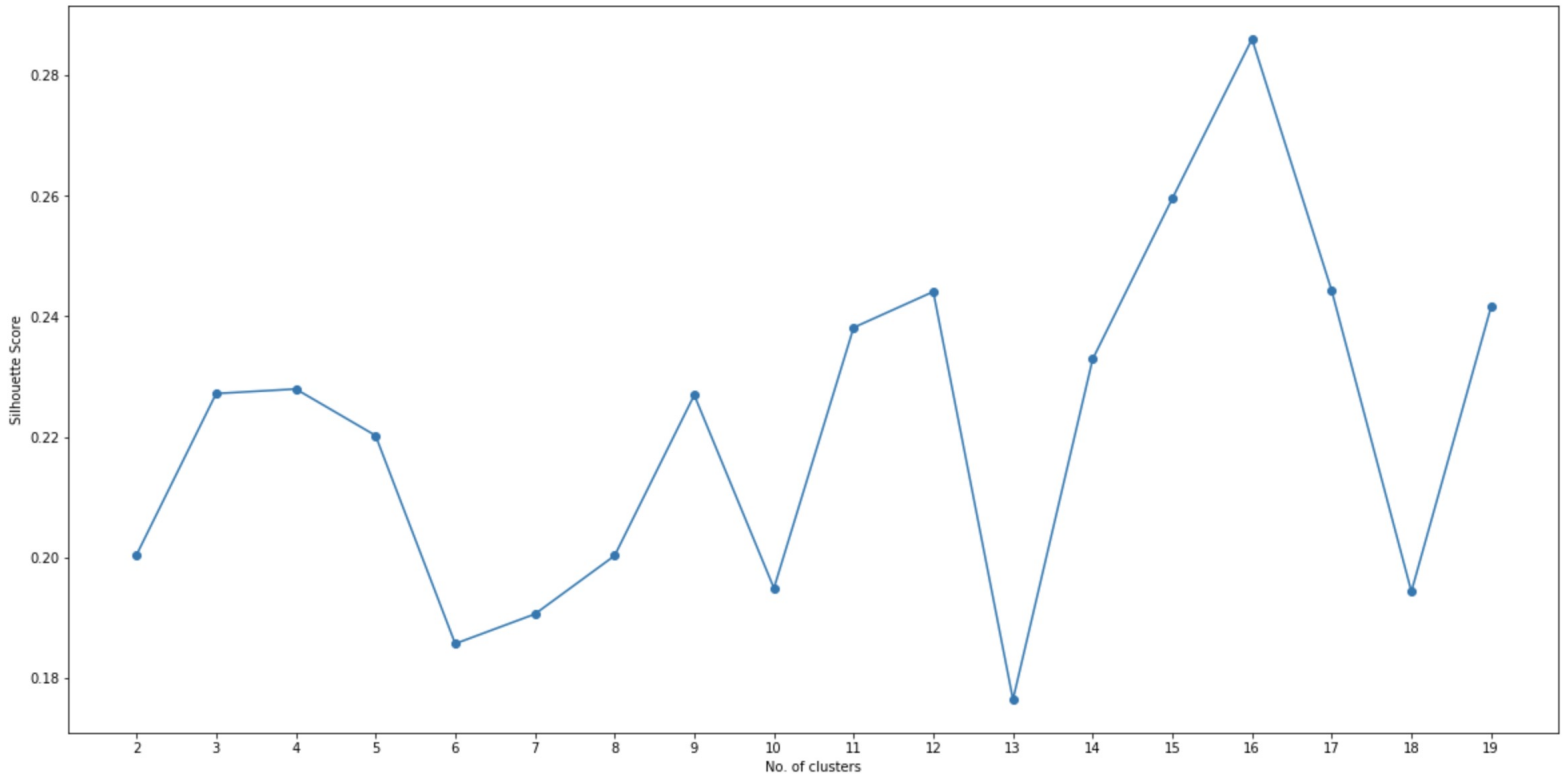

Figure: Silhouette score vs No.of clusters.

Methodology

K-means clustering

The venue data is then trained using `K-means` Clustering Algorithm to get the desired clusters to base the analysis on. `K-means` was chosen as the variables (Venue Categories) are huge, and in such situations `K-means` will be computationally faster than other clustering algorithms.

Results

The neighbourhoods are divided into `n` clusters where `n` is the number of clusters found using the optimal approach. The clustered neighbourhoods are visualized using different colours so as to make them distinguishable.

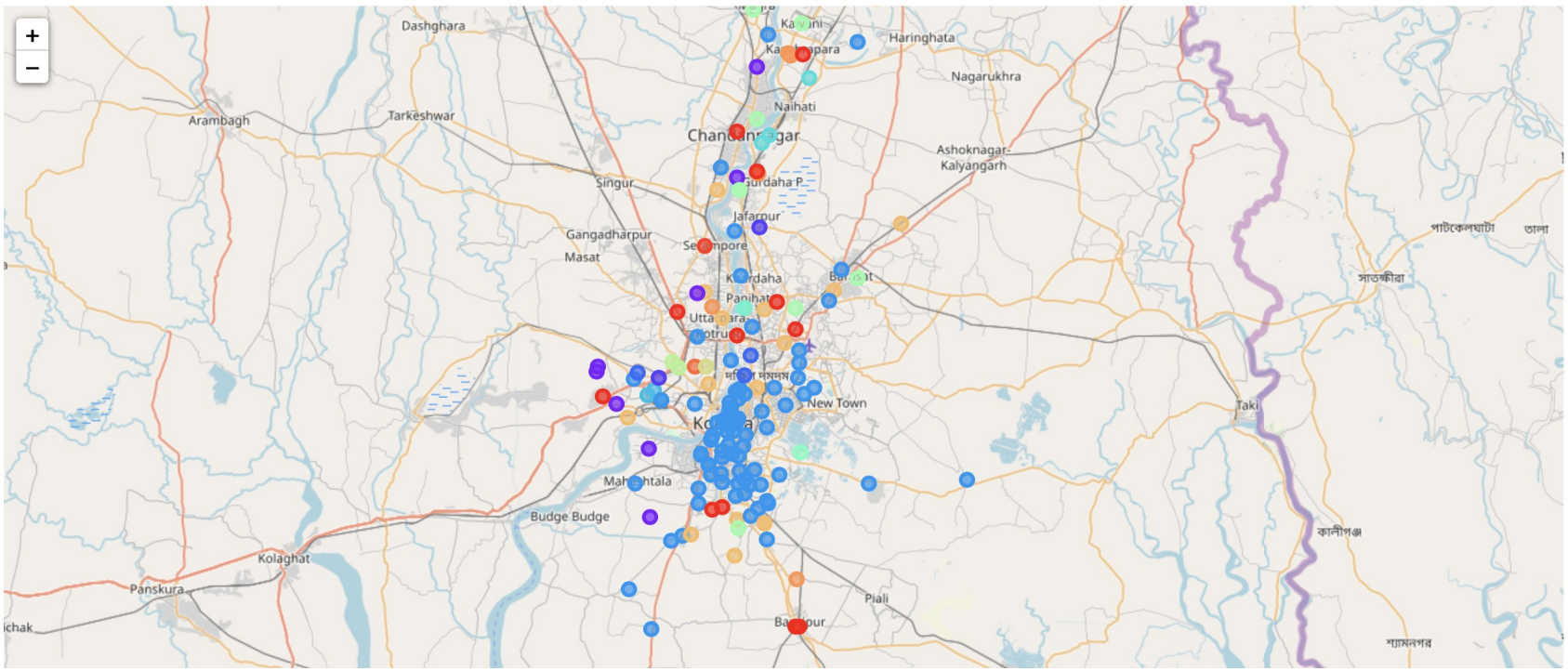

Figure: Neighbourhoods of Kolkata (Clustered).

Discussion

The five places namely `Bijpur, Garshyamnagar, Halisahar, Hind Motor and Kodalia` fall in the outskirts of the city of Kolkata, hence the demographic of the population in these areas fall under the lower middle class of the society.

	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
60	Bijpur, North 24 Parganas	Train Station	Women's Store	Electronics Store	Film Studio	Field	Fast Food Restaurant	Falafel Restaurant	Fabric Shop	Event Service	Eastern European Restaurant
123	Garshyamnagar	Train Station	Platform	Women's Store	Eastern European Restaurant	Field	Fast Food Restaurant	Falafel Restaurant	Fabric Shop	Event Service	Electronics Store
131	Halisahar	Train Station	Women's Store	Electronics Store	Film Studio	Field	Fast Food Restaurant	Falafel Restaurant	Fabric Shop	Event Service	Eastern European Restaurant
141	Hind Motor	Light Rail Station	Train Station	Women's Store	Electronics Store	Field	Fast Food Restaurant	Falafel Restaurant	Fabric Shop	Event Service	Eastern European Restaurant
186	Kodalia	Train Station	Women's Store	Electronics Store	Film Studio	Field	Fast Food Restaurant	Falafel Restaurant	Fabric Shop	Event Service	Eastern European Restaurant

Figure: Cluster having Train Station as most common venue

Discussion

- India is expected to see a dramatic growth in the middle class, from 5 to 10 percent of the population in 2005 to 90 percent in 2039, by which time a billion people will be added to this group.
- In 2005, the mean per capita household expenditure was just US $3.20 per day, and very few households exceeded incomes of US $5 per day. Yet, by 2015, half the population had crossed this threshold. By 2025, half the Indian population is expected to surpass US $10 per day.

Source : Diana Farrell and Eric Beinhocker. "Next Big Spenders: India's Middle Class," McKinsey Inc., May 19, 2007.

Discussion

Table 1
Occupations of India's Lower Middle Class
(Percent of Population)

Occupation	%
Vendors	30
Food Industry	13
Leather Work	8
Painters/Carpenters	7
Construction	6
Miscellaneous	7
Cloth/Shop Washing	5
Security Services	5
Unspecified	4
Welding & Repairing	4
Bangle Workers	2
Cable/Electrical Work	2
Data Entry	2
Driver/Transport Services	2
Imitation Jewelery Makers	2
Bangle workers	2

Source: *ATLAS: The Local Impact of Globalization in South and Southeast Asia.*

Figure: Occupation of Indian's Lower Middle Class

Conclusion

- As the middle class will grow at a rapid rate in the next upcoming years, opening food outlets catered to the needs of that section of the society will see a massive increase in footfall, which would lead to a further increase in business.
- If the food outlets have an average rate of US $0.5 equivalent to 15 percent of the per capita household expenditure, for their items, then profits can be expected to be high as the food rate is neither too low or too high for a person of the concerned demographic to spend.
- Assuming a footfall of 30 people getting off at these stations for each train, 100 trains passing through these stations daily, and a conversion rate of 20 percent, ordering only one meal, a daily turnover of around US $300 can be expected from these outlets per station.