Coursera Capstone Project : Applied Data Science

Zoya Fahad Khan

zoyashaikh.0894@gmail.com

Overview

Introduction

- The Mumbai Suburban Railway (colloquially called local trains or simply locals) consists of exclusive inner suburban railway lines augmented by commuter rail on main lines serving outlying suburbs to serve the Mumbai Metropolitan Region. Spread over 390 kilometers (240 mi), the suburban railway operates 2,342 train services and carries more than 7.5 million commuters daily. By annual ridership (2.64 billion), the Mumbai Suburban Railway is one of the busiest commuter rail systems in the world and it has the most severe overcrowding in the world. Trains run from 04:00 until 01:00, and some trains also run up to 02:30. It is the second largest suburban rail network in terms of route length after the Kolkata Suburban Railway.

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 a new restaurant chain can be put up and also finding out the category of the same, aiming at the above demographic and maximize profits out of them.

Data

Neighbourhoods

The data of the neighbourhoods in Mumbai 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 mumbai.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.
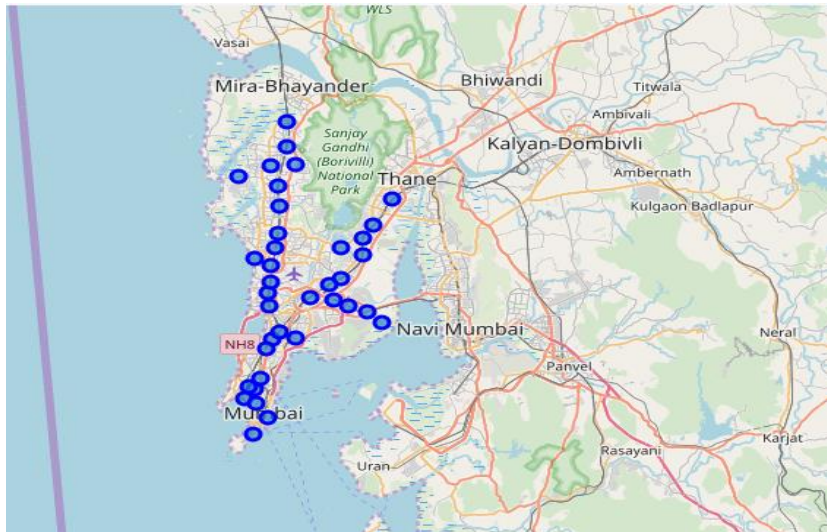
Methodology

Figure: Neighbourhoods of Mumbai.

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.
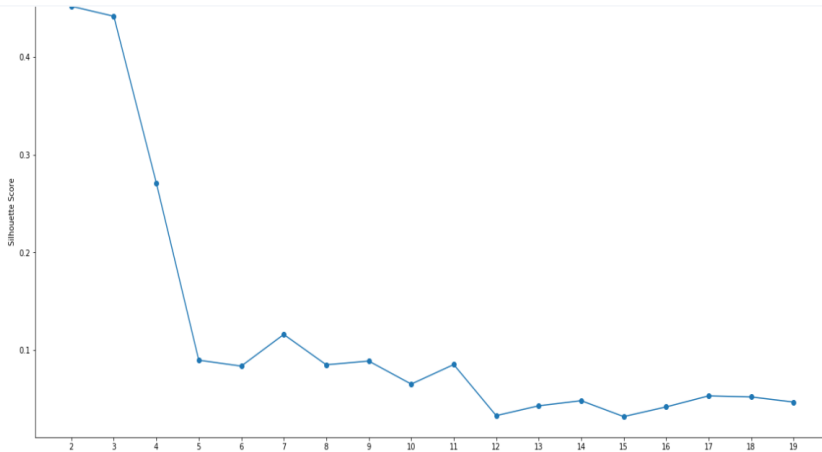
Methodology

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.
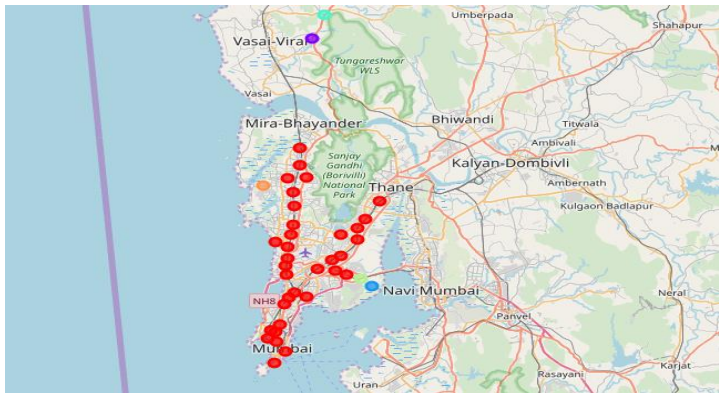

Figure: Neighbourhoods of Mumbai (Clustered).

Discussion

After analysing the various clusters produced by the Machine learning algorithm, cluster no.1, is a prime fit to solving the problem of finding a cluster with common venue as an Indian Restaurant.

	Neighborhood	Latitude	Longitude	Cluster Labels	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
0	Andheri	19.1197	72.8464	0	Indian Restaurant	Fast Food Restaurant	Café	Ice Cream Shop	Sandwich Place	Camera Store	Falafel Restaurant	Coffee Shop	Pizza Place	Burger Joint
1	Bhayandar	19.1972	72.8114	5	Hot Dog Joint	Diner	Resort	Dance Studio	Dessert Shop	Design Studio	Donut Shop	Electronics Store	Department Store	Food & Drink Shop
2	Bandra	19.055	72.8402	0	Indian Restaurant	Snack Place	Gourmet Shop	Café	Seafood Restaurant	Park	Indie Movie Theater	Food Truck	Korean Restaurant	Lounge
3	Borivali	19.2291	72.8574	0	Ice Cream Shop	Chinese Restaurant	Indian Restaurant	Clothing Store	Park	Pizza Place	Scenic Lookout	Sandwich Place	Restaurant	Café
4	Dahisar	19.2572	72.8575	0	Indian Restaurant	Fast Food Restaurant	Department Store	Café	Coffee Shop	BBQ Joint	Soccer Field	Gym	Restaurant	Sandwich Place
5	Goregaon	19.1648	72.85	0	Indian Restaurant	Fast Food Restaurant	Snack Place	Design Studio	Indie Movie Theater	Restaurant	Sandwich Place	Seafood Restaurant	Lounge	Bookstore
6	Jogeshwari	19.1349	72.8488	0	Ice Cream Shop	Indian Restaurant	Coffee Shop	Mughlai Restaurant	Fast Food Restaurant	Smoke Shop	Sandwich Place	Asian Restaurant	Pharmacy	Pizza Place
7	Juhu	19.107	72.8275	0	Hotel	Café	Chinese Restaurant	Lounge	Seafood Restaurant	Cocktail Bar	Bar	Indian Restaurant	Spanish Restaurant	Restaurant
8	Kandivali West	19.2084	72.8422	0	Electronics Store	Fast Food Restaurant	Ice Cream Shop	Pizza Place	Coffee Shop	Gym / Fitness Center	Food	Snack Place	Restaurant	Chinese Restaurant
9	Kandivali East	19.2102	72.8649	0	Pizza Place	Fast Food Restaurant	Indian Restaurant	Lounge	Department Store	Restaurant	Café	Sports Bar	Multiplex	Chinese Restaurant
10	Khar, Mumbai	19.0697	72.8399	0	Bar	Gym / Fitness Center	Ice Cream Shop	Pub	Café	Seafood Restaurant	Gym	Frozen Yogurt Shop	Fish & Chips Shop	Coffee Shop

Figure: Cluster having Indian Restaurant 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.

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.