Coursera Capstone Project : Applied Data Science

Sarthak Srivastava

Overview

Introduction

- Mumbai is India's commercial capital, its entertainment capital, and is famously known as the city that never sleeps. Mumbai is India's busiest and most-populous city, with the 2018 Census of India estimating that 12 million call it home.

A city that never sleeps, people are always working tirelessly for their work! And to minimize the time, many even try to grab their meals from restaurants near their work place to save time.

Business Problem

- Each person in the corporate sector is in a rush right from the morning. As a result, there are many restaurant outlets opening up in the proximity of such areas and locations.

- And this is exactly what my project is about! To open up a new Food Outlet in the best proximal region around a corporate sector. Let us call this restaurant “Mr. Brown”.

- Local train commuters rely heavily on their breakfast, sometimes even lunch, on light meals. These can be on the go sandwiches, fries, momos, pizza, burgers, some microwaveable or cold prepared meal along with beverages.

- Our goal is to find the optimal location where this restaurant can be set up and flourish with it’s light ready-to-go snacks.

- A location, in Mumbai City, where a food outlet can easily survive with a healthy competition.

- A location, where food outlets are present in scarcity and are hugely needed. A location where it attracts the mass gentry of people.

- We can also try to find more than one location, and who knows we might be able to set up a ‘Mr. Brown’ chain of food outlets!

Target Audience

- The entrepreneur who wants to set up a new food outlet or perhaps a new food chain in the city of Mumbai

- A food outlet for the mass gentry people, a food outlet for the corporates. A food outlet affordable for the youth. A food outlet for everyone!

Data

Area, Location

The data of Area and Location in Mumbai can be extracted out by scraping the web using the BeautifulSoup library of python. The data used in the project is scraped from a Wikipedia page.

Geocoding

To get the latitude and longitude of the extracted area and location, I tried to use Foursquare API, but the results were mostly returned as ‘None’. I also tried to use the Google Maps Geocoding API, but was unsuccessful in the same way. Hence, the latitude and longitudes of the area are scraped from Wikipedia as well.

Venue Data

From the Area, Location and Region data, the venue data is found out by passing in the required parameters to the FourSquare API, and thereby creating a pandas DataFrame to contain all the venue details along with the location details.

How data will be used to solve the problem?

From the venues extracted using the Foursquare API, venues which are most demanded in the choice of people will be selected.

From the data analysis, we will group all these regions in different clusters according to category of venues.

The weakest cluster, i.e. the cluster which has met least demands of the people in the food category will be selected as our optimal region.

Methodology

Accuracy of Geocoding API and Foursquare API

In the initial development phase, the API returned a bunch of erroneous results and continued to do so in the later stages of the project as well.

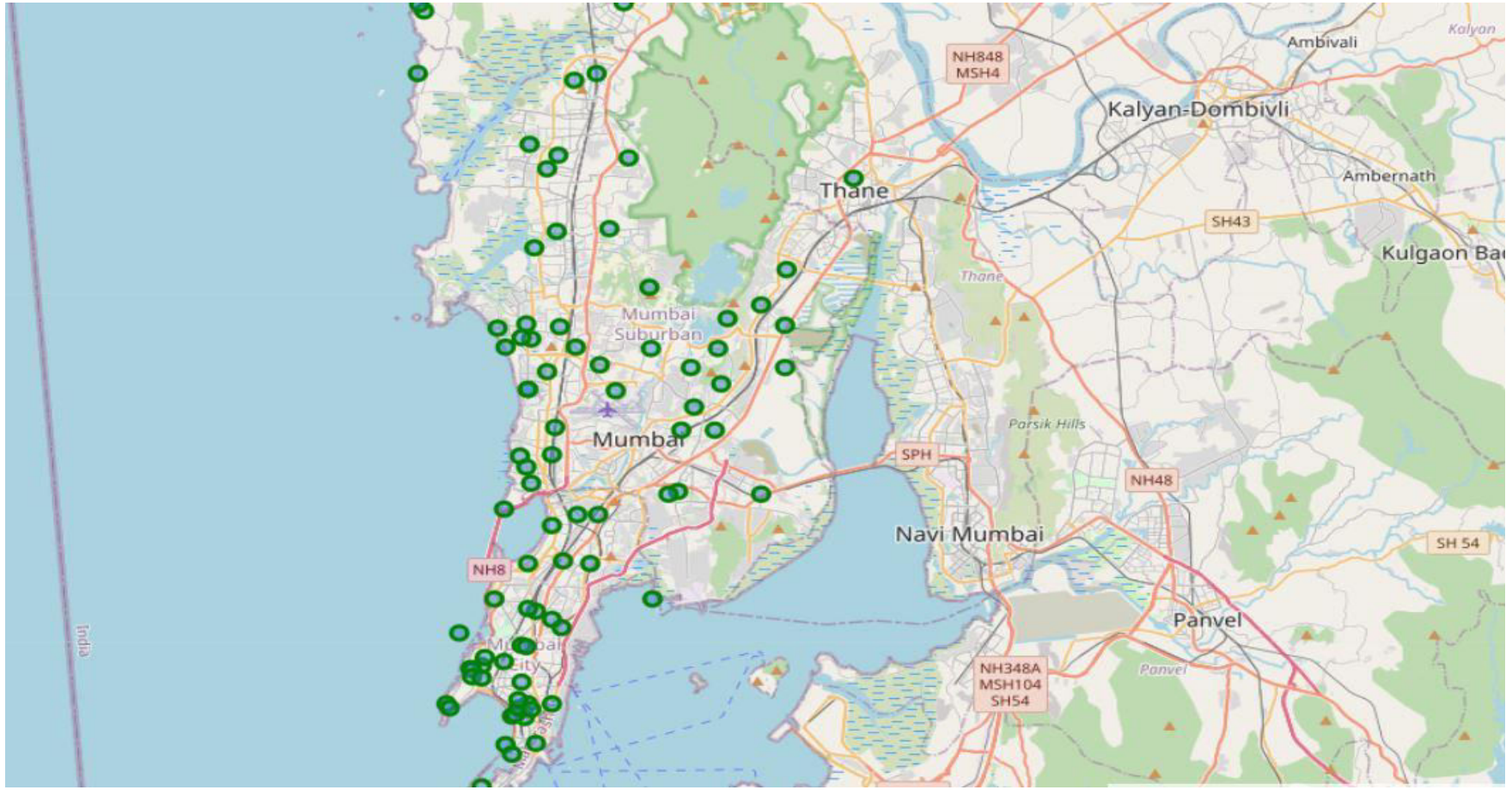
The Foursquare API was not much help as well when it encountered some areas which were unpopular. As a result, I had to scrape the coordinates from the Wikipedia page as well, using the BeautifulSoup library of python.

Methodology

Folium

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

Methodology



Areas in Mumbai, Maharashtra

Methodology

One-Hot Encoding of all the Venues

One hot encoding is a process by which categorical variables are converted into a form that could be provided to Machine Learning 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 15 Most Common Venues!

There is a high variety in the venues collected from the Foursquare API, and to avoid a sparse distribution only the top 15 common venues and a new pandas DataFrame is made. This dataframe is then used to train the K-Means Clustering Algorithm to find the optimal number of clusters!

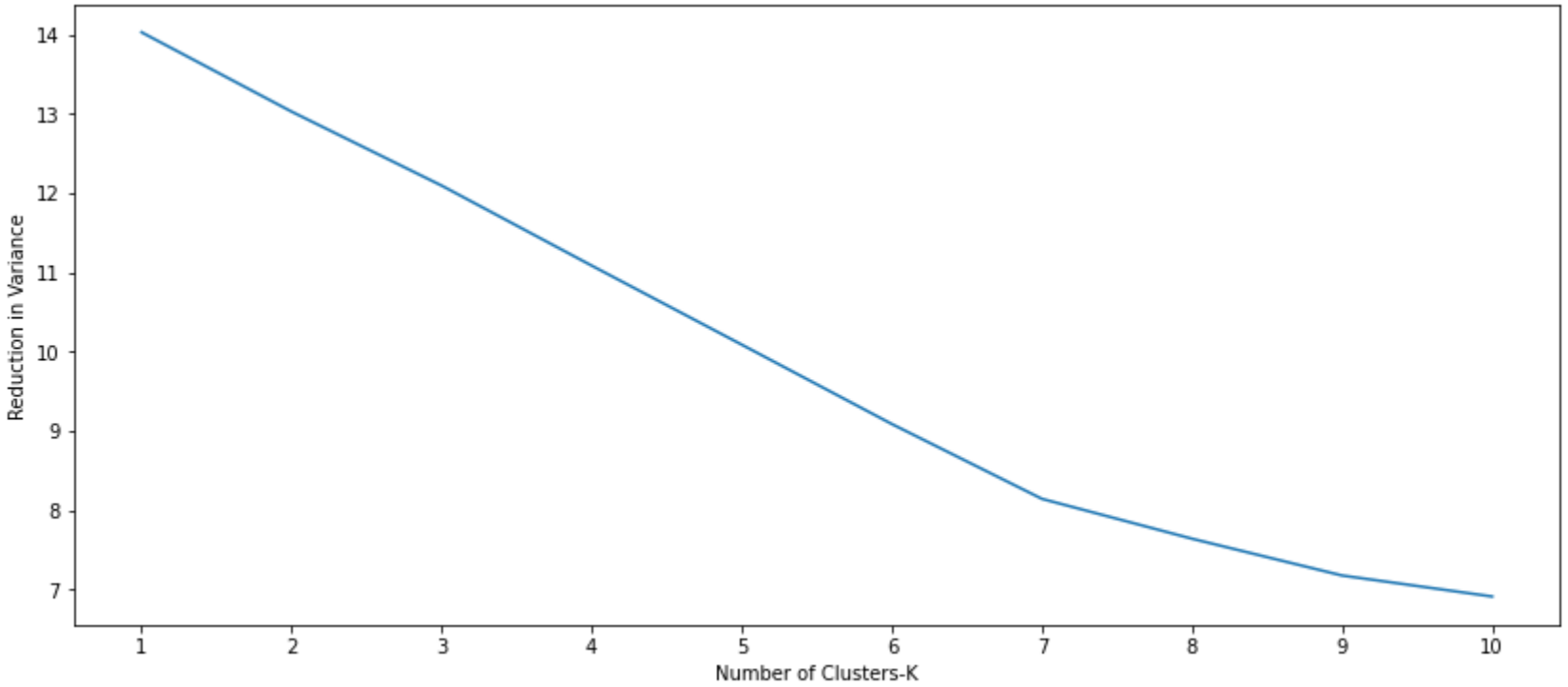
Methodology

Optimal number of Clusters- This is done in 2 ways!

- By plotting the a graph between reduction in variance and increasing number of clusters. From the graph, we can point out the "**elbow point**" visually, and this determines the optimal number of clusters.

- **Silhouette Score-** 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.

Methodology



Locating the Elbow Point

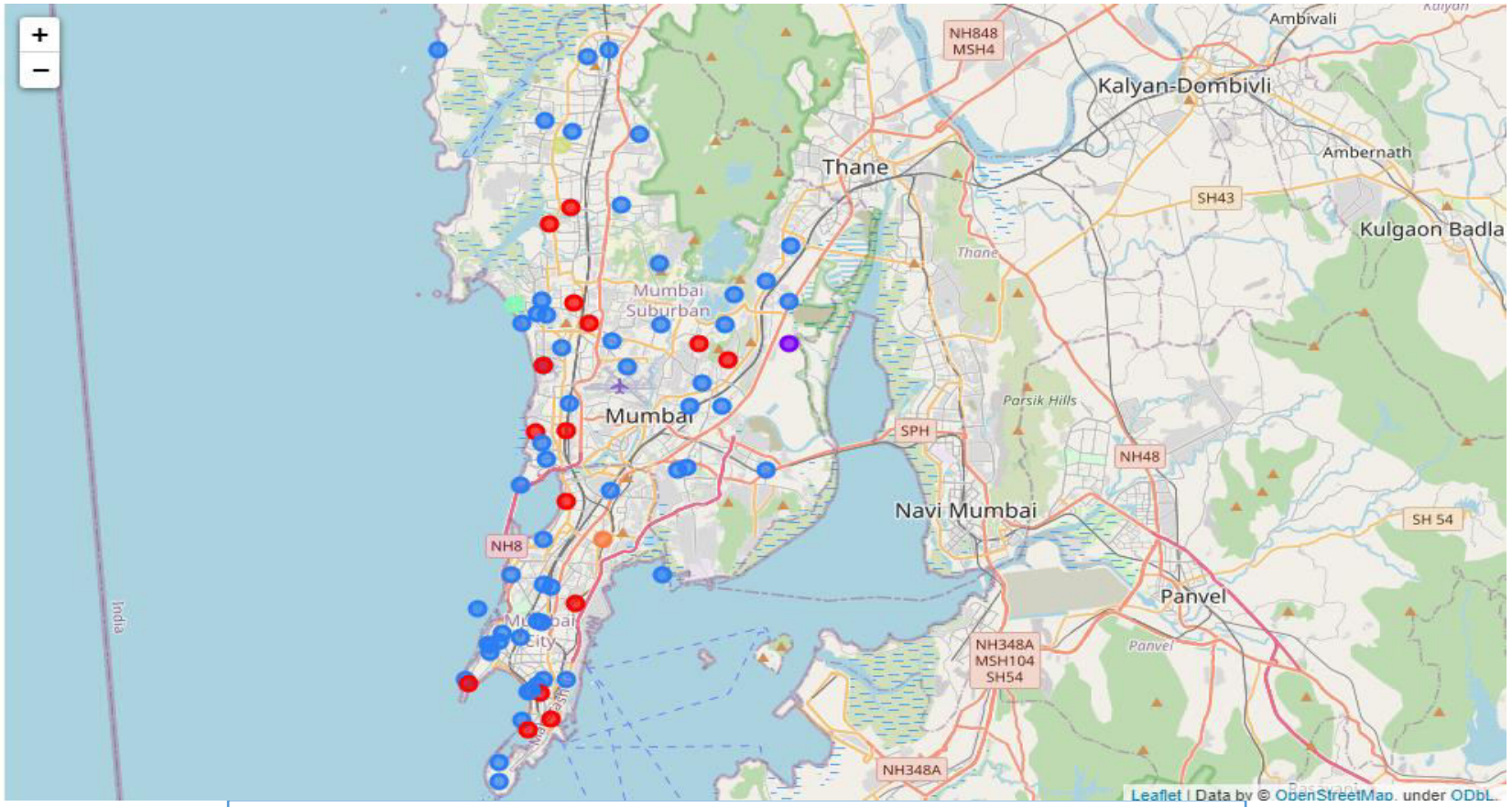
Methodology

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 number of Venue Categories were huge in number, and in such situations K-means will be computationally faster than other clustering algorithms.

Results

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



Clustered Areas in Mumbai

Discussion

- Considering from a Business point of view, we need to take care of the fact that Mr. Brown is a new setup of food outlet. It does not have any chains or links or connection in the market whatsoever.

- So we need a location which is not highly competitive but competitive enough for our restaurant to flourish from a **differential point of view**.

- Now, looking closely at the map we can see that the blue cluster is highly concentrated. On close analysis of this cluster we see that it consists of high end venues such as Seafood and Thai Restaurants along with other posh venues. This does not attract our target people.

Discussion

- Looking at the sparse clusters such as the orange one and the sky blue one, they are located in remote areas of Mumbai. Hence, it would make them a poor target given they would not attract the corporate and mass audience

- The highly optimised cluster to place our food outlet is the RED cluster. This region is dense enough to encourage a healthy competition, not very high end in the Top 15 venues of people and even attracts a good mass of audience. A highly detailed analysis of this is provided in the final project presentation.

Discussion

Cluster 1, our target audience!

```
In [64]: val = 1
mumbai_merged.loc[mumbai_merged['Cluster Labels'] == (val - 1), mumbai_merged.columns[[0] + np.arange(4, mumbai_merged.shape[1]).
```

Out[64]:

	Area	Region	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	11th Cor
1	Amboli	Western Suburbs	0	Coffee Shop	Bakery	Gym	Indian Restaurant	Falafel Restaurant	Event Space	Electronics Store	Dumpling Restaurant	Donut Shop	Diner	Din Resta
11	Bhayandar	Western Suburbs	0	Indian Restaurant	Market	Fast Food Restaurant	American Restaurant	Cheese Shop	Restaurant	Ice Cream Shop	Food	Dim Sum Restaurant	Event Space	Elect
15	Pali Hill	Western Suburbs	0	Indian Restaurant	Café	Bakery	Cheese Shop	Fast Food Restaurant	Food Truck	Bar	Market	Coffee Shop	Chinese Restaurant	S
20	Bangur Nagar	Western Suburbs	0	Indian Restaurant	Chinese Restaurant	Train Station	Café	Jewelry Store	Yoga Studio	Dhaba	Falafel Restaurant	Event Space	Electronics Store	Dur Resta
21	Jogeshwari West	Western Suburbs	0	Indian Restaurant	Dessert Shop	Restaurant	Juice Bar	Market	BBQ Joint	Ice Cream Shop	Breakfast Spot	Antique Shop	Department Store	Dur Resta
28	Khar Danda	Western Suburbs	0	Indian Restaurant	Art Gallery	Cupcake Shop	Diner	Yoga Studio	Dhaba	Farmers Market	Falafel Restaurant	Event Space	Electronics Store	Dur Resta
30	Sunder Nagar	Western Suburbs	0	Indian Restaurant	Cheese Shop	Restaurant	Bridal Shop	Middle Eastern Restaurant	Café	Bar	Bakery	Electronics Store	Ice Cream Shop	Fast Resta
33	Nalasopara	Western Suburbs	0	Indian Restaurant	Bakery	Café	Bar	Coffee Shop	Cheese Shop	Boutique	Market	Mexican Restaurant	Soccer Field	Resta
36	Vile Parle	Western Suburbs	0	Food	Ice Cream Shop	Indian Restaurant	Fast Food Restaurant	Yoga Studio	Dhaba	Falafel Restaurant	Event Space	Electronics Store	Dumpling Restaurant	
38	Amrut Nagar	Eastern Suburbs	0	Resort	Seafood Restaurant	Indian Restaurant	Dessert Shop	Falafel Restaurant	Event Space	Electronics Store	Dumpling Restaurant	Donut Shop	Diner	Din Resta
42	Nehru Nagar	Eastern Suburbs	0	Indian Restaurant	Coffee Shop	Concert Hall	Event Space	Café	Dhaba	Falafel Restaurant	Electronics Store	Dumpling Restaurant	Donut Shop	

Cluster having Restaurant, Food Outlets as most common venue

Recommendations based on Observation

The best location to set up a restaurant or a food outlet in Mumbai would be: Powai, Goregaon (East and West), Thane, Dharavi, Vile Parle and IIT Bombay Campus.

These are the locations that attract the majority people and are a frequent visiting place for the commuters.

Conclusion

- The areas of Powai and Gogegaon East and Goregaon West are chosen as the solution as the best location to set up a restaurant or a food outlet.

- As the middle class and corporate sector will grow at a rapid rate in the next upcoming years, opening food outlets cateres to the needs of this section of the society: especially to these two regions in the Western Suburbs of Mumbai city.

- Our food outlet also attracts a large mass as students: all these factors are described in details in the final project report. The factors that lead to the prosperity of Mr. Brown Bakery.

Conclusion

'According to the National Restaurant Association of India's (NRAI) Food Services Report 2019, the hospitality sector had a compounded annual growth rate (CAGR) of 11 per cent between 2015-16 and 2018-19. The organised segment, which is 35 per cent of this sector, had a CAGR of 13 per cent during the same period, its market share growing from Rs 1,01,475 crore to Rs 1,48,353 crore. This segment ***was projected to grow at a CAGR of 15 per cent to reach a market size of Rs 2,57,907 crore by 2022-23****. Clearly, the outlook was positive.'*

Source : The Indian Express

With these statistics, we can say that our food outlet would surely be encouraged by the people it has been targeting in the right locations. What can we say more? Happy Eating :D