Coursera Capstone Project : Applied Data Science

Sarthak Srivastava

1 Introduction

Mumbai is an amalgamation of seven islands namely Parel, Mazagaon, Colaba, Mahim, Old Woman’s Island (Little Colaba) and Bombay Island.

The trains in Mumbai are not just a mere mode of transport for commuters, it’s the lifeline which breathes life into the mundane life of workers and the corporate community. For hawkers and trinket sellers, trains are a means of livelihood. Trains provide a varied assortment of contrasting personalities and people from different walks of life. Mumbai is renowned as the city that never sleeps.

A city that never sleeps requires all the energy to put in all the energy towards it’s work. The energy is food, they are the take-away snacks, they are the sips of coffee between two corporate meetings. They are the small bite of food and sometimes even a meal for the busy people, who are always in rush.

With a population of 1.84 crores, the city is pouring with people who work in the corporate sector. Just like it’s rains, the mass of youth is flooding the city! And it is never not going to decrease. A financial capital, it is a home to many hard workers.

2 Business Problem

The commuters, the local train travellers, the students who commute daily to their college, and the hawkers heavily rely on food outlets for their meals. Sometimes for snacks, and sometimes even for their meals. These food outlets could be located near railway stations, near colleges, near apartment buildings. Any accessible place for the mass people.

This is our target audience! The mass gentry of people. These are the people that the owner of Mr. Brown Bakery is looking for. To set up chains of such food outlets that provides snacks, meals, and even affordable lunch and dinner for the commuters. Ready-to-go food, not time consuming and homemade.

Thus, the main objective of the project is to find ideal locations in the city of Mumbai where fast food retail chains can be put up, aiming at the middle-class demographic, and helping the owners to extract maximum profit out of them.

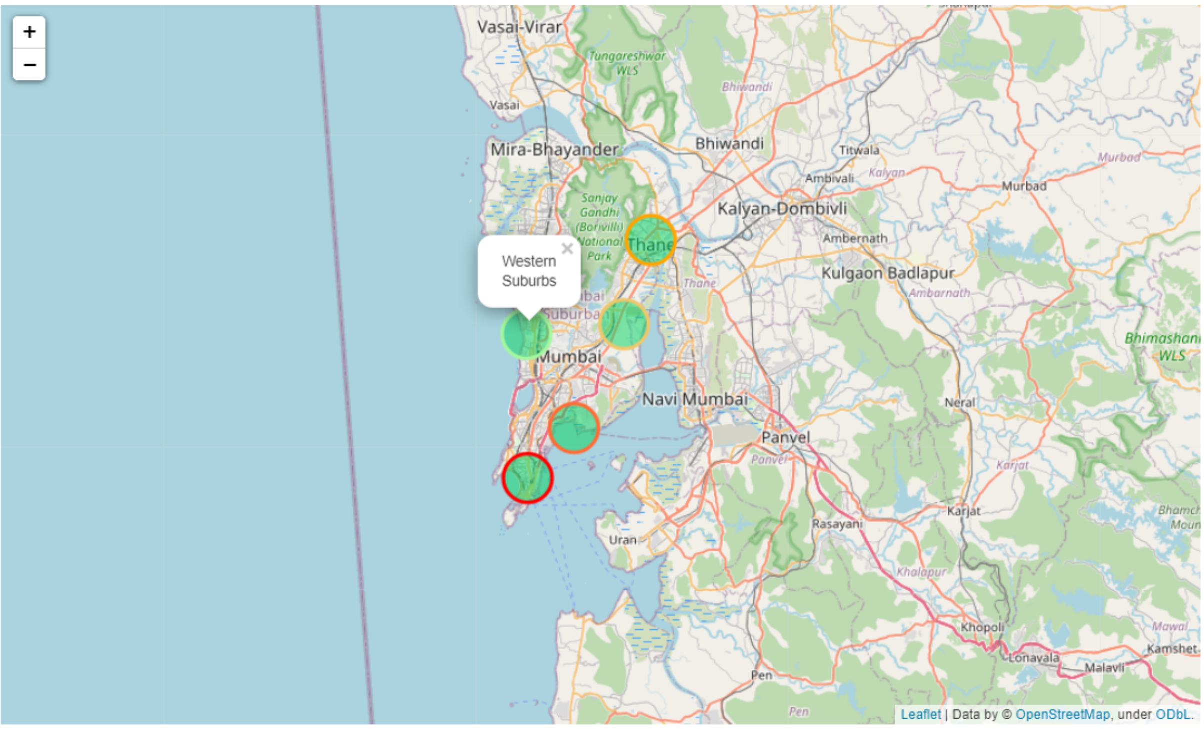
3 Data

The data for this project has been extracted out by scraping the Wikipedia page and using multiple but accurate sources, taking care of the accuracy of the methods used.

To extract the data, the BeautifulSoup library of python is used, and the Foursquare API is used to extract the venues near the areas by passing in the latitude and longitudes.

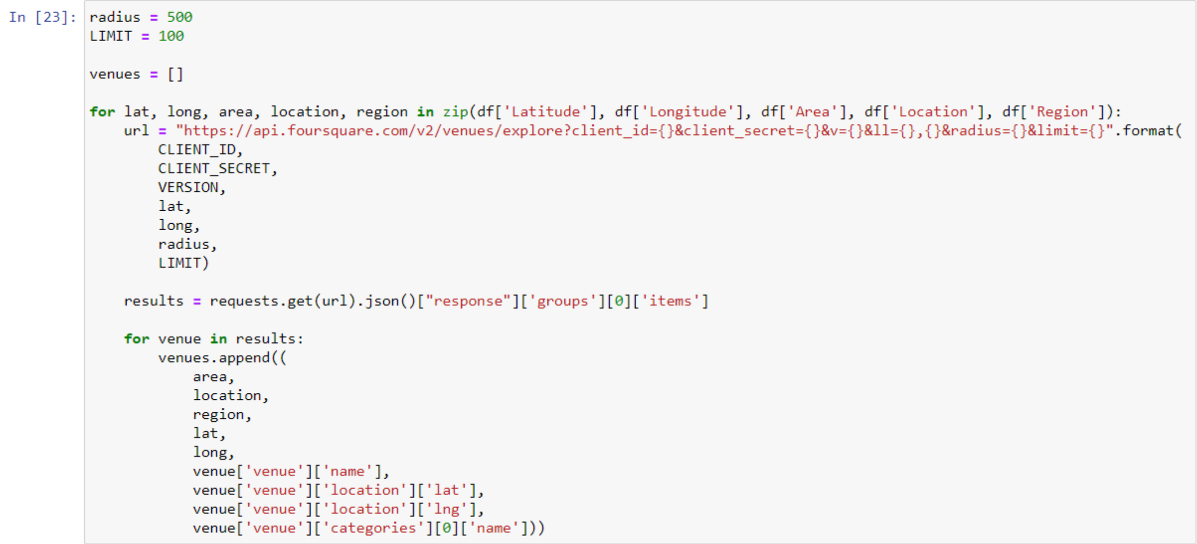
3.1 Area, Location and Region

This data of Area and Location is extracted using the BeautifulSoup Library of python from a Wikipedia page. From this data, I have performed augmentation of data to create a column ‘Region’ which clusters these areas, into five big clusters as five different regions in Mumbai: Western Suburbs, Eastern Suburbs, Trombay, Mumbai and South Mumbai.



3.2 Venue Data

From the locations data obtained after web scraping, 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 their respective locations and coordinates.



```
In [23]: radius = 500
         LIMIT = 100

         venues = []

         for lat, long, area, location, region in zip(df['Latitude'], df['Longitude'], df['Area'], df['Location'], df['Region']):
             url = "https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}".format(
                 CLIENT_ID,
                 CLIENT_SECRET,
                 VERSION,
                 lat,
                 long,
                 radius,
                 LIMIT)

             results = requests.get(url).json()["response"]['groups'][0]['items']

             for venue in results:
                 venues.append((
                     area,
                     location,
                     region,
                     lat,
                     long,
                     venue['venue']['name'],
                     venue['venue']['location']['lat'],
                     venue['venue']['location']['lng'],
                     venue['venue']['categories'][0]['name']))
```

Figure 1: Collecting Venues data using Foursquare API

4 Methodology

A thorough analysis of the principles of methods, rules, and postulates employed have been made in the project in order to insure that the inferences that are made are as accurate as possible.

4.1 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.

4.2 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.

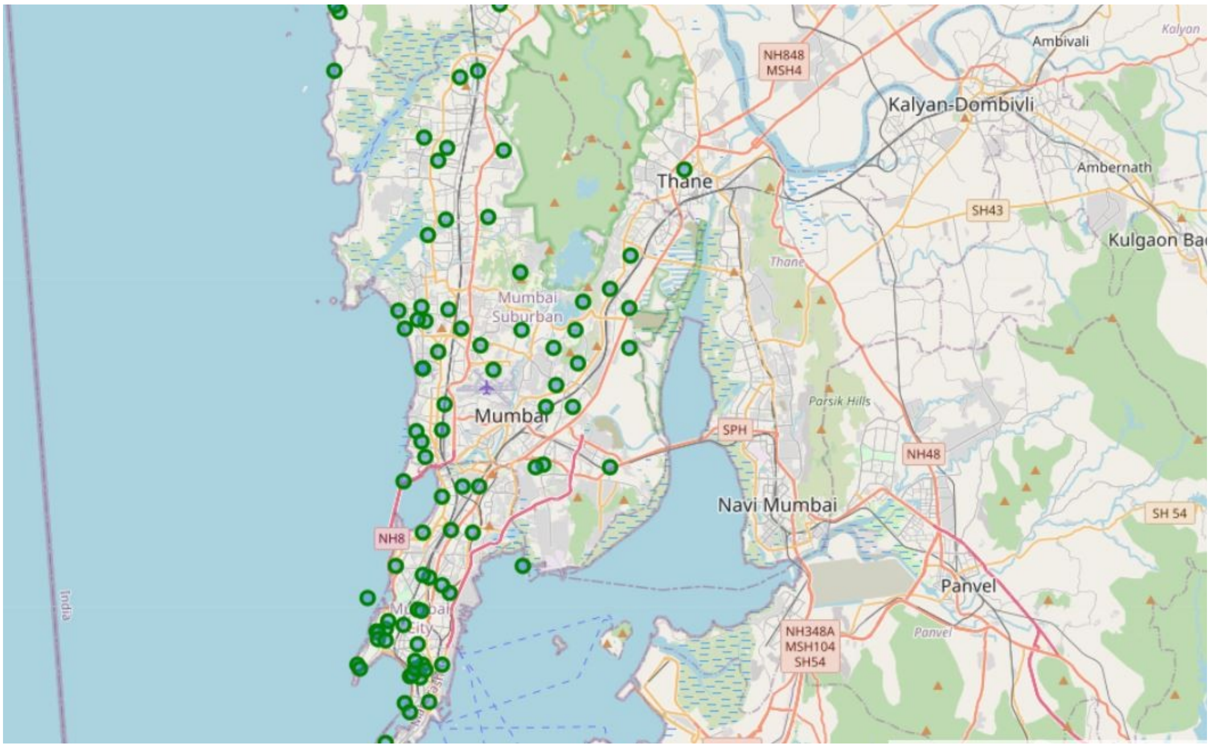


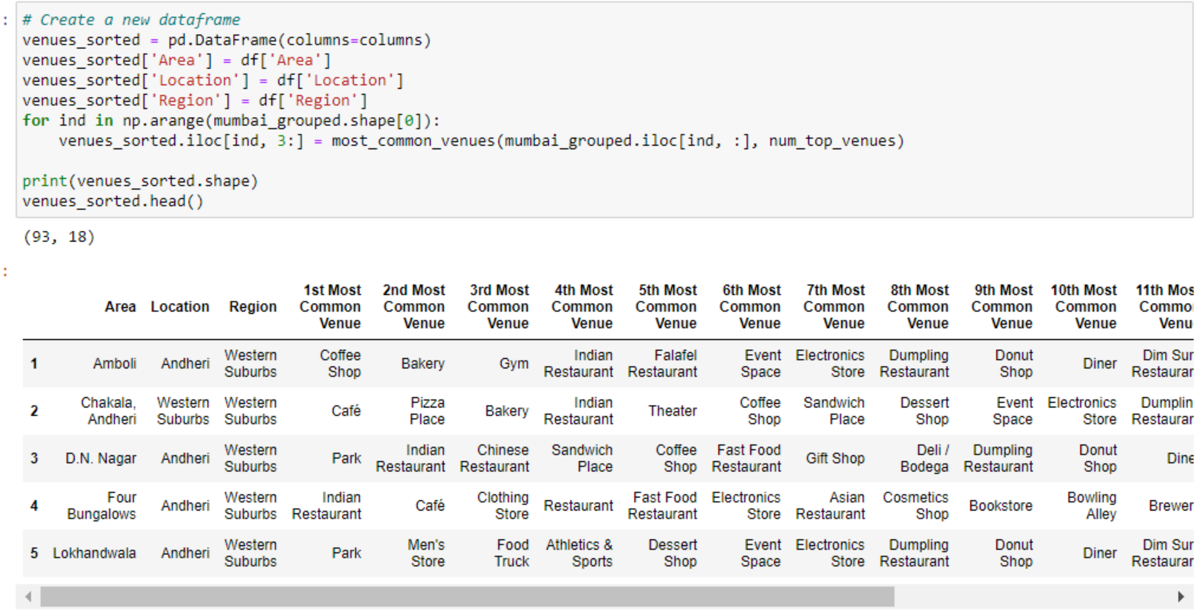
Figure 2: Areas in Mumbai City

4.3 One-Hot Encoding

For categorical variables where no such ordinal relationship exists, the integer encoding is not enough. One hot encoding is the most widespread approach, and it works very well unless your categorical variable. For the K-Means Clustering Algorithm used later in the project, all unique items under Venue Category are one-hot encoded.

4.4 Top 15 Most Common Venues

A total of 173 venues were returned when we used the Foursquare API was called within a radius of 500 metres of each area. Due to high variety in the venues, only the top 15 venues are selected and a new DataFrame is created, which is used for K-Means unsupervised clustering algorithm.



```
# Create a new dataframe
venues_sorted = pd.DataFrame(columns=columns)
venues_sorted['Area'] = df['Area']
venues_sorted['Location'] = df['Location']
venues_sorted['Region'] = df['Region']
for ind in np.arange(mumbai_grouped.shape[0]):
    venues_sorted.iloc[ind, 3:] = most_common_venues(mumbai_grouped.iloc[ind, :], num_top_venues)

print(venues_sorted.shape)
venues_sorted.head()
```

```
(93, 18)
```

	Area	Location	Region	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 Mos Commo Venu
1	Amboli	Andheri	Western Suburbs	Coffee Shop	Bakery	Gym	Indian Restaurant	Falafel Restaurant	Event Space	Electronics Store	Dumpling Restaurant	Donut Shop	Diner	Dim Sur Restaurar
2	Chakala, Andheri	Western Suburbs	Western Suburbs	Café	Pizza Place	Bakery	Indian Restaurant	Theater	Coffee Shop	Sandwich Place	Dessert Shop	Event Space	Electronics Store	Dumplin Restaurar
3	D.N. Nagar	Andheri	Western Suburbs	Park	Indian Restaurant	Chinese Restaurant	Sandwich Place	Coffee Shop	Fast Food Restaurant	Gift Shop	Deli / Bodega	Dumpling Restaurant	Donut Shop	Dine
4	Four Bungalows	Andheri	Western Suburbs	Indian Restaurant	Café	Clothing Store	Restaurant	Fast Food Restaurant	Electronics Store	Asian Restaurant	Cosmetics Shop	Bookstore	Bowling Alley	Brewer
5	Lokhandwala	Andheri	Western Suburbs	Park	Men's Store	Food Truck	Athletics & Sports	Dessert Shop	Event Space	Electronics Store	Dumpling Restaurant	Donut Shop	Diner	Dim Sur Restaurar

Figure 3: Top 15 Venues in every area

4.5 Cluster Optimization

- In the project, to find the optimum number of clusters two ways are used. One, 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. The logic behind this is that after the elbow point, the reduction in variance keeps on decreasing but steadily now each datapoint is assigned to it’s own unique cluster.

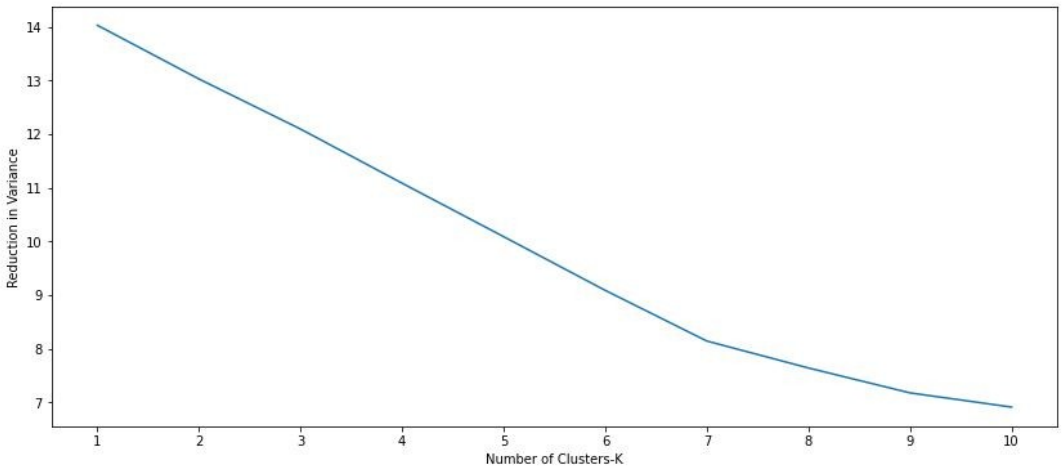


Figure 4: Locating the Elbow Point Visually: 7

- **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 and poorly matched to the neighbouring clusters. Based on the Silhouette Score of various clusters below 11, the optimal cluster size is determined.

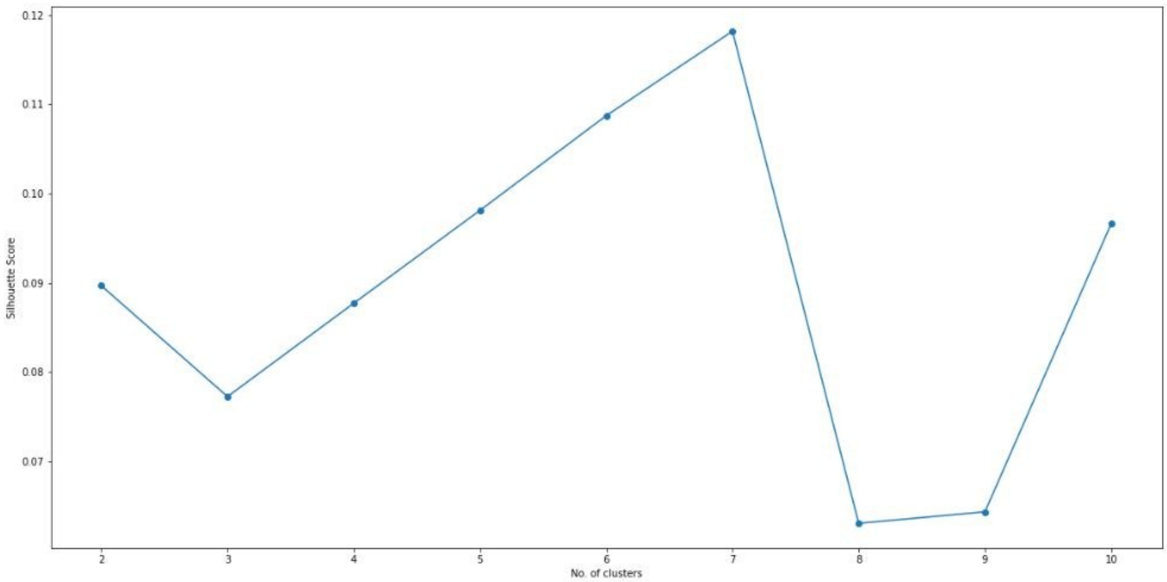


Figure 5: Clusters Optimization using Silhouette Score

4.6 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.

5 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.

The division of clusters is done on the basis of the set of venues which are visited most by a group of people. The graph below shows the **seven** clusters that we found form the K-means algorithm.

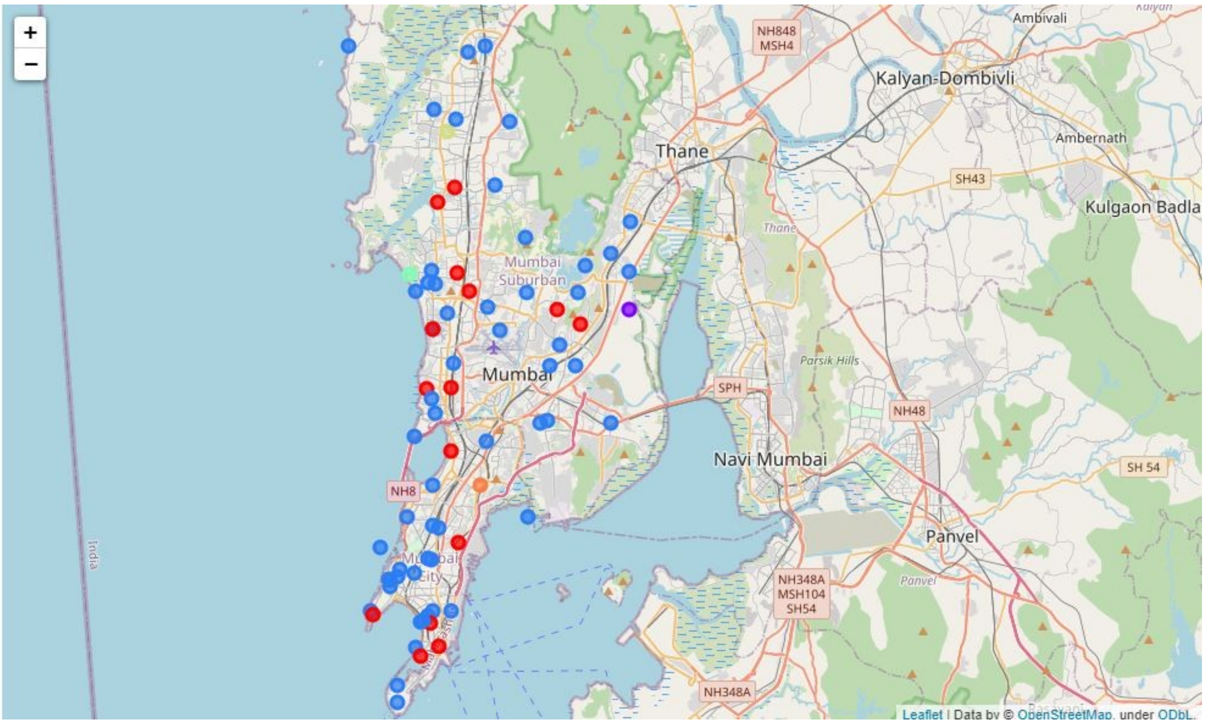


Figure 6: Clustering of areas

6 Discussion

Inferential Statistics:

- 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.

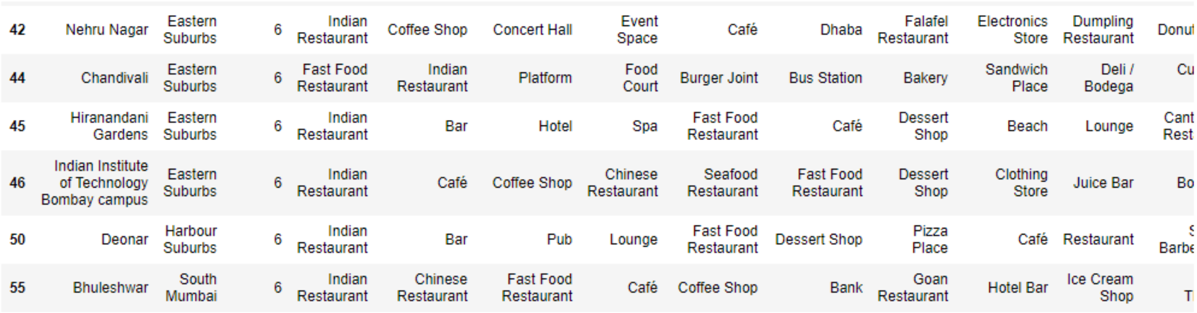
- So we need a location which is not highly competitive but competitive enough for our restaurant to flourish from a **differential point of view**.
- Now, looking closely at the map we can see that the purple cluster is highly concentrated. On close analysis of this cluster we see that it consists of high end venues such as Seafood and Thai Restaurants along with other posh venues. This does not attract our target people.
- Looking at the sparse clusters such as the red 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

After analysing the various clusters produced by the Machine Learning Algorithm, we find that cluster number seven (i.e. the light spots in the map) is a prime fir to solving the business problem set up. This is the Western Suburban region of Mumbai consisting of areas such as Powai, Goregaon, DN Nagar, Vile Parle, Bandra East and areas of South Mumbai.

Reasoning:

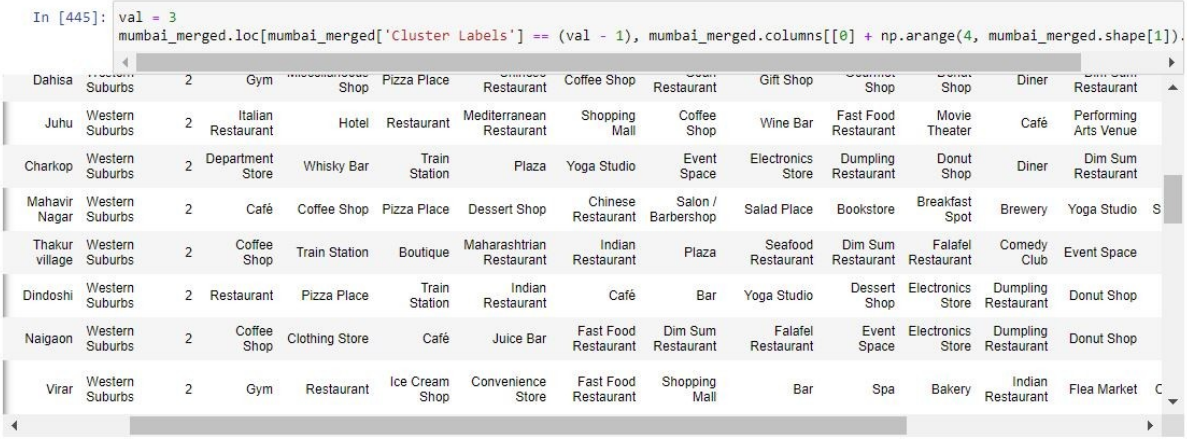
Our target was to attract a mass of audience to our food outlet. Now, let us pick up an area to analyse it. Let us take Powai for example. Powai is the educational centre of Mumbai city with a cluster of colleges in the area. The famous Indian Institute of Technology, Bombay , NITIE and ICFAI Business School of Management is located in Powai as well.

This clustering of colleges in Powai leads to an influx of students and youth to Powai every single day. The gentry of youth, by statistics, is attracted to local food outlets for meals and for replenishment. We can see this by comparing the most visited venues in Powai (red cluster) and in South Bombay (a highly posh area in Mumbai) i.e. the blue cluster.



42	Nehru Nagar	Eastern Suburbs	6	Indian Restaurant	Coffee Shop	Concert Hall	Event Space	Café	Dhaba	Falafel Restaurant	Electronics Store	Dumpling Restaurant	Donut
44	Chandivali	Eastern Suburbs	6	Fast Food Restaurant	Indian Restaurant	Platform	Food Court	Burger Joint	Bus Station	Bakery	Sandwich Place	Deli / Bodega	Cu
45	Hiranandani Gardens	Eastern Suburbs	6	Indian Restaurant	Bar	Hotel	Spa	Fast Food Restaurant	Café	Dessert Shop	Beach	Lounge	Cant Rest
46	Indian Institute of Technology Bombay campus	Eastern Suburbs	6	Indian Restaurant	Café	Coffee Shop	Chinese Restaurant	Seafood Restaurant	Fast Food Restaurant	Dessert Shop	Clothing Store	Juice Bar	Bo
50	Deonar	Harbour Suburbs	6	Indian Restaurant	Bar	Pub	Lounge	Fast Food Restaurant	Dessert Shop	Pizza Place	Café	Restaurant	S Barbe
55	Bhuleshwar	South Mumbai	6	Indian Restaurant	Chinese Restaurant	Fast Food Restaurant	Café	Coffee Shop	Bank	Goan Restaurant	Hotel Bar	Ice Cream Shop	T

Figure 7: The Red Cluster has affordable places as its top venues!



```
In [445]: val = 3
          mumbai_merged.loc[mumbai_merged['Cluster Labels'] == (val - 1), mumbai_merged.columns[[0] + np.arange(4, mumbai_merged.shape[1]).
```

Dahisa	Suburbs	2	Gym	Shop	Pizza Place	Restaurant	Coffee Shop	Restaurant	Gift Shop	Shop	Shop	Diner	Restaurant
Juhu	Western Suburbs	2	Italian Restaurant	Hotel	Restaurant	Mediterranean Restaurant	Shopping Mall	Coffee Shop	Wine Bar	Fast Food Restaurant	Movie Theater	Café	Performing Arts Venue
Charkop	Western Suburbs	2	Department Store	Whisky Bar	Train Station	Plaza	Yoga Studio	Event Space	Electronics Store	Dumpling Restaurant	Donut Shop	Diner	Dim Sum Restaurant
Mahavir Nagar	Western Suburbs	2	Café	Coffee Shop	Pizza Place	Dessert Shop	Chinese Restaurant	Salon / Barbershop	Salad Place	Bookstore	Breakfast Spot	Brewery	Yoga Studio	S
Thakur village	Western Suburbs	2	Coffee Shop	Train Station	Boutique	Maharashtrian Restaurant	Indian Restaurant	Plaza	Seafood Restaurant	Dim Sum Restaurant	Falafel Restaurant	Comedy Club	Event Space
Dindoshi	Western Suburbs	2	Restaurant	Pizza Place	Train Station	Indian Restaurant	Café	Bar	Yoga Studio	Dessert Shop	Electronics Store	Dumpling Restaurant	Donut Shop
Naigaon	Western Suburbs	2	Coffee Shop	Clothing Store	Café	Juice Bar	Fast Food Restaurant	Dim Sum Restaurant	Falafel Restaurant	Event Space	Electronics Store	Dumpling Restaurant	Donut Shop
Virar	Western Suburbs	2	Gym	Restaurant	Ice Cream Shop	Convenience Store	Fast Food Restaurant	Shopping Mall	Bar	Spa	Bakery	Indian Restaurant	Flea Market	C

Figure 8: The Blue cluster has high end Restaurants and Dining Bars as its top venues

The latter is not the gentry mass of people that we wanted to target, as discussed in the Business Problem. The area of Powai has affordable outlets of food such as Desert Shops, Café, Coffee Shops which the youth mostly likes to visit. Hence, the cluster is a good choice.

Let us take another example, the area of Vile Parle. Now, this area is located closely to the airport. Loads travellers and tourists land in Mumbai every day. Setting up a restaurant would highly benefit the influx of people and also open it up to a new audience.

Hence, the cluster chosen is an optimal and appropriate choice for setting up Mr. Brown Bakery.

7 A Hidden Insight!

While analysing and cleaning the DataFrame, we grouped areas according to the top 15 venues in the area. Now, there were nine rows returned which consisted of NaN values. On data analysis, it was found out that these areas had some venues, but just because of the fact that they did not match with the top ones in the list- they were segregated.

Now, anyone who lives in Mumbai would know that Thane is one of the top five busiest stations in Mumbai- for it connects Mumbai to Navi Mumbai and faces a heavy heavy influx of commuters each day! Now, imagine how profitable it would be to set up a food outlet or even a restaurant while expanding the food chains of Mr. Brown Bakery!

Hence, I would like to add Thane city as well to the areas listed in the Discussion section, even though it does not belong to the red cluster that we have chosen! But it will highly boost and benefit the outlet if we decide to open it here.

8 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 organized 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 ***<u>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</u>.*** *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