Expansion Strategy in Mumbai City for Beriyan Biryani

Applied Data Science Capstone by IBM/Coursera

Table of contents

Introduction: Business Problem

Biryani is one of the most consumed foods in India. It is estimated that the biryani market in India is estimated to be approximately Rs 1,500 crore (USD200M) in the organised sector, and Rs 15,000 crore (USD2B) in the unorganised sector.

Beriyan Biryani, a Biryani restaurant chain started out of Hyderabad, India has quickly garnered popularity and revenue growth. As they look to scale up, it sees Mumbai as the logical next choice of the city to expand to. Beriyan Biryani stakeholders are looking to choose a few localities in Mumbai to do a street level analysis, in order to come up with the prospecting address in Mumbai.

This project attempts to find optimal locations for restaurants in a city. Specifically, this report will be targeted to Beriyan Biryani stakeholders interested in opening an **Biryani restaurant** in **Mumbai**, India.

We will use data visualization and machine learning methods to generate a few most promissing neighborhoods based on this criteria. Each selection will be comprehensively outlined for the aforementioned stakeholders to make the choice

Data

Based on definition of our problem, some key factors that will influence the decision are:

- Density of Restaurants in a neighborhood
- Restaurant Ratings in a neighborhood
- Commercial Space Rates in a neighborhood

A map encapsulating all the major neighborhoods would be a good starting point.

Also, pertinent information about the neighborhoods would be necessary

Following data sources will be needed to extract/generate the required information:

- Mumbai neigborhoods will be availed by web-scraping the **Mumbai Wikipedia** page

- Neighborhood co-ordinates (Latitude and Longitude) will be retrieved using **Nominatim** module from **GeoPy** library
- Neighborhood data related to restaurants, venues etc. will be retrieved using **Foursquare API**
- Commercial Space Rates will be availed by web-scraping **MagicBricks** and **99Acres** (major online real estate platforms in India) database

Mumbai Neighborhoods Dataframe

Mumbai Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai)) lists all of the major neighborhoods in Mumbai, sorted by the suburbs, namely- Western, Eastern, Harbor and South Mumbai.

We have a basic csv file prepared made by copying this data. We start by importing it, along with the necessary libraries for data analysis and visualization.

Out[2]:

	Suburb	Neighborhood	Commercial Rates
0	Western	Andheri	18639
1	Western	Mira Bhayandar	8369
2	Western	Bandra	33466
3	Western	Borivali	17238
4	Western	Dahisar	13064

Out[3]:

	Suburb	Neighborhood	Commercial Rates	Latitude	Longitude
0	Western	Andheri	18639	19.119698	72.846420
1	Western	Mira Bhayandar	8369	NaN	NaN
2	Western	Bandra	33466	19.054979	72.840220
3	Western	Borivali	17238	19.229068	72.857363
4	Western	Dahisar	13064	19.249450	72.859621

Out[4]:

	index	Suburb	Neighborhood	Commercial Rates	Latitude	Longitude
0	0	Western	Andheri	18639	19.119698	72.846420
1	2	Western	Bandra	33466	19.054979	72.840220
2	3	Western	Borivali	17238	19.229068	72.857363
3	4	Western	Dahisar	13064	19.249450	72.859621
4	5	Western	Goregaon	11121	19.164803	72.850045

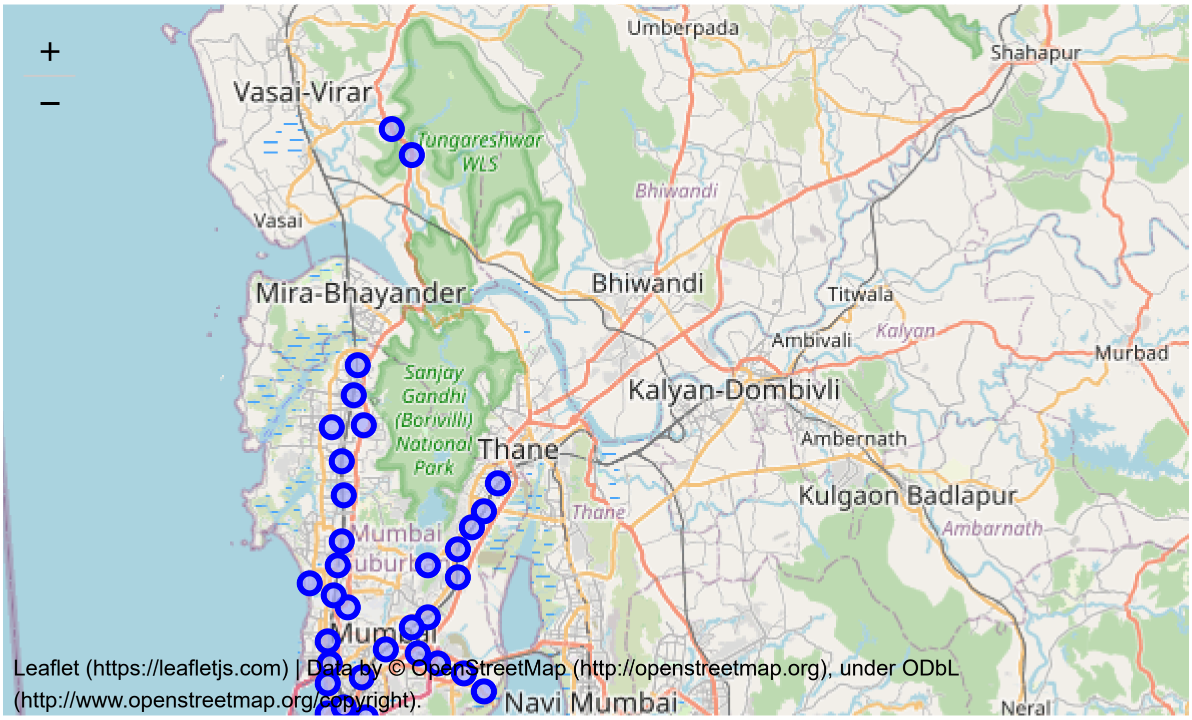
Out[5]:

	Suburb	Neighborhood	Commercial Rates	Latitude	Longitude
0	Western	Andheri	18639	19.119698	72.846420
1	Western	Bandra	33466	19.054979	72.840220
2	Western	Borivali	17238	19.229068	72.857363
3	Western	Dahisar	13064	19.249450	72.859621
4	Western	Goregaon	11121	19.164803	72.850045

Now that the data is cleaned up and ready, let us create a map visualization of the Mumbai neighborhoods, so as to get the geographical overview of the city.

```
Mumbai co-ordinates are: 19.1648029 72.8500454
```

Out[7]:



Leaflet (https://leafletjs.com) | Data by © OpenStreetMap (http://openstreetmap.org), under ODbL (http://www.openstreetmap.org/copyright).

Methodology

Data Mining

Foursquare is a social location service that allows users to explore the world around them. The Foursquare API allows application developers to interact with the Foursquare platform.

We have Mumbai neighobrhoods ready, now we use **Foursquare API** to get info on venues in each Mumbai neighborhood. The venues are categorized in several types such as Indian Restaurants, Internet Cafes, Car Showrooms etc.

We will be gathering **location (co-ordinates), venue name and venue category** for each Mumbai neighborhood.

For the perview of this project, we will be considering venues relavant to food only, unrelated venue categories such as Convenience store, Gym, Library etc. will be removed. This will remove the noise and improve the quality of the clustering and of the insights gathered from it.

Exploratory Analysis

We will visualize the **venue density and commercial space rates** in the neighborhoods. We will also take a look at what kind of venues are the most popular for a given neighborhood. This will help us understand the distribution of venue categories and even identify a trend in the data.

Clustering

In this final step we will find **clusters of locations** based on the data and try to gain an understanding of those clusters from the stakeholders' perspective. We will present map of all such locations but also create clusters (using **k-means clustering**) of those locations to identify neighborhoods which should be a starting point for final 'street level' exploration and search for optimal venue location by stakeholders.

We start by setting up Client credentials for Foursquare API

```
Your credentails:
CLIENT_ID: BEOXN0050X0Q2TAMAV1G4WG5AMT0ESFSM1O21CTY5OWTB00D
CLIENT_SECRET:5HWTCTKTJ0KR1W0T2GBC0EFWMGZTE3P1IQQ21XDLNMTKFT3R
```

Out[9]:

```
'https://api.foursquare.com/v2/venues/explore?&client_id=BEOXN0050X0Q2TAMAV1
G4WG5AMT0ESFSM1O21CTY5OWTB00D&client_secret=5HWTCTKTJ0KR1W0T2GBC0EFWMGZTE3P1
IQQ21XDLNMTKFT3R&v=20180604&ll=19.1648029,72.8500454&radius=500&limit=100'
```

Now we define a function that will return venues within specific radius of a neighborhood, from Foursquare API. This will help us populate the data on restaurant name, categories, location etc.

As we can see, a total of 150 venue categories are retrieved. Now, as discussed in the Methodology section, for the perview of this project, we will be considering venues relavant to food only.

Out[13]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Andheri	19.119698	72.84642	Merwans Cake shop	19.119300	72.845418	Bakery
1	Andheri	19.119698	72.84642	Narayan Sandwich	19.121398	72.850270	Sandwich Place
2	Andheri	19.119698	72.84642	McDonald's	19.119691	72.846102	Fast Food Restaurant
3	Andheri	19.119698	72.84642	Cafe Alfa	19.119667	72.843560	Indian Restaurant
4	Andheri	19.119698	72.84642	McDonald's	19.118411	72.848002	Fast Food Restaurant

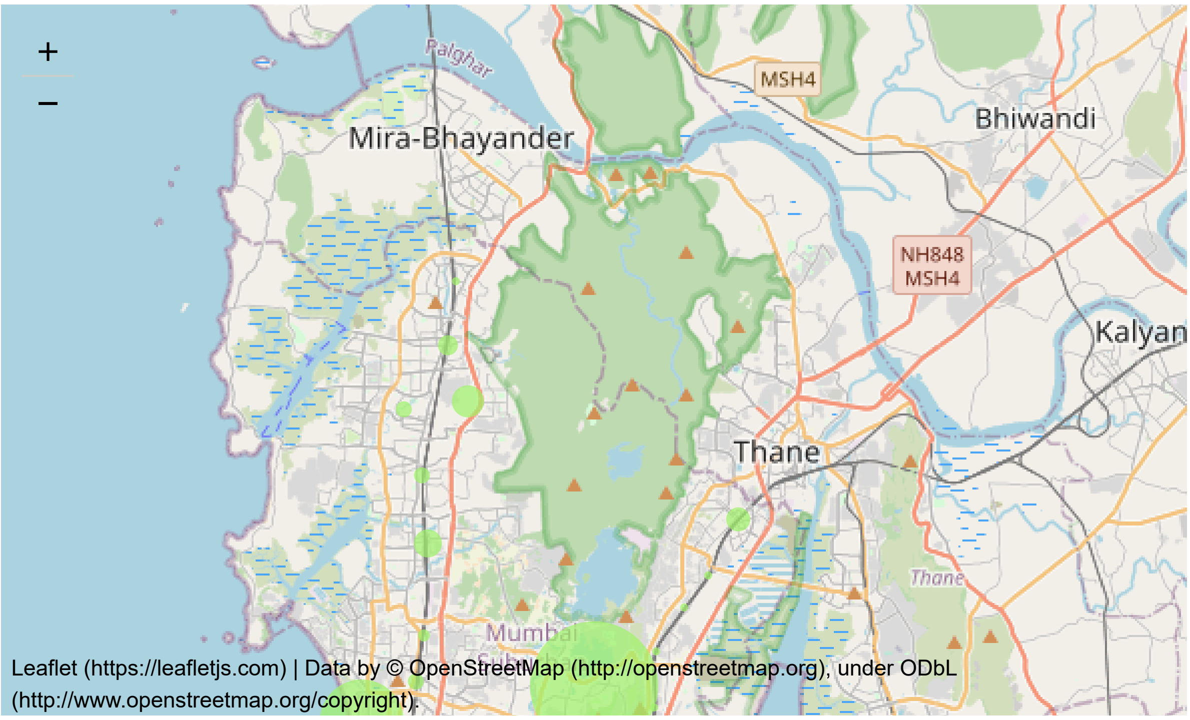
Out[15]:

	Suburb	Neighborhood	Commercial Rates	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude
0	Western	Andheri	18639	19.119698	72.84642	Merwans Cake shop	19.119300	72.845418
1	Western	Andheri	18639	19.119698	72.84642	Narayan Sandwich	19.121398	72.850270
2	Western	Andheri	18639	19.119698	72.84642	McDonald's	19.119691	72.846102
3	Western	Andheri	18639	19.119698	72.84642	Cafe Alfa	19.119667	72.843560
4	Western	Andheri	18639	19.119698	72.84642	McDonald's	19.118411	72.848002

Now we have the venue data that we can use to cluster the neighborhoods and analyze. We have gathered the venues from Foursquare API for Mumbai neighborhoods and filtered by the categories relevant to our analysis, i.e. the Food category.

Let's take a view at which neighborhoods have greater density of food places, we will visualize it using a bubble map.

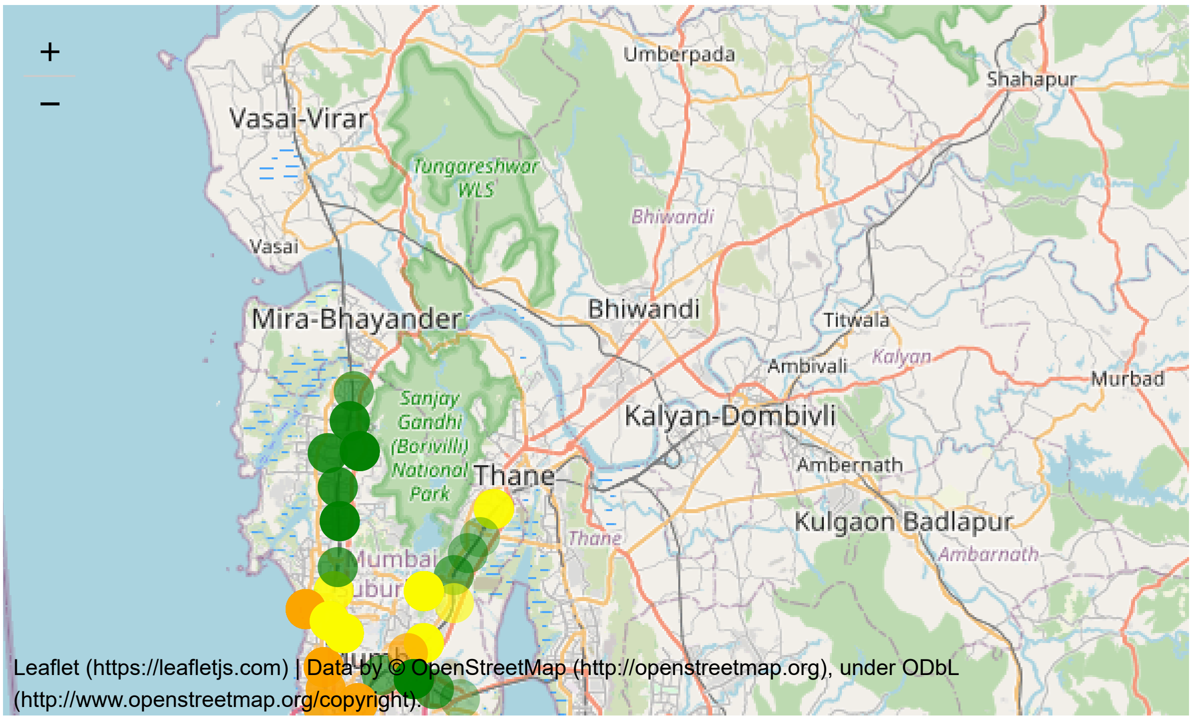
Out[16]:



As we can see, prominent residential neighborhoods such as Powai, Colaba, Juhu, Khar, Santacruz etc. have higher density of food places. One key insight would be to avoid selecting such a locality that has higher density of food places.

It would also be a good idea to visualize the commercial property prices for neighborhoods, hence we map them by color scale.

Out[17]:



It's no surprise that South Mumbai neighborhoods fetch greatest price for the commercial property, they have greater business activity and higher level of quality of life. Also, South-Westerm neighborhoods are highly priced, owing to the well planned neighborhoods and more recreational venues.

Analysis

Now we will proceed with the analysis part. As discussed in the Methodology section, we will be doing **K-means clustering** on our dataframe, taking into consideration a variety of parameters such as Venue Categories, Commercial Rates, number of Venues etc. The k-means clustering algorithm attempts to split a given anonymous data set (a set containing no information as to class identity) into a fixed number (k) of clusters. These clusters often exhbit a unique characteristic trend that we can identify with analysis of the clusters.

We start by setting up our data appropriately for the K-means clustering.

Out[18]:

	Neighborhood	Commercial Rates	American Restaurant	Asian Restaurant	BBQ Joint	Bakery	Bar	Bed & Breakfast	Beer Bar	Ga
0	Andheri	18639	0	0	0	1	0	0	0	
1	Andheri	18639	0	0	0	0	0	0	0	
2	Andheri	18639	0	0	0	0	0	0	0	
3	Andheri	18639	0	0	0	0	0	0	0	
4	Andheri	18639	0	0	0	0	0	0	0	

5 rows × 69 columns

Out[19]:

```
(553, 69)
```

Out[20]:

	Neighborhood	Commercial Rates	American Restaurant	Asian Restaurant	BBQ Joint	Bakery	Bar	Bed & Breakfast	Beer Bar	(
0	Andheri	18639	0.0	0.0	0.0	0.111111	0.0	0.0	0.0	
1	Antop Hill	28571	0.0	0.0	0.0	0.000000	0.5	0.0	0.0	
2	Bandra	33466	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	
3	Bhandup	14890	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	
4	Borivali	17238	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	

5 rows × 69 columns

In order to further remove the noise, we will be considering only the top ten venue categories from a neighborhood. This will de-clutter the data further and clustering will be much more neat, which will help us identify the trends even further.

Out[22]:

	Neighborhood	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 Mos Commo Venu
0	Andheri	Commercial Rates	Fast Food Restaurant	Indian Restaurant	Sandwich Place	Food Court	Bakery	Pizz Plac
1	Antop Hill	Commercial Rates	Diner	Bar	Fast Food Restaurant	Cupcake Shop	Dessert Shop	Dhab
2	Bandra	Commercial Rates	Café	Indian Restaurant	Brewery	Pub	Restaurant	Italia Restaurar
3	Bhandup	Commercial Rates	Falafel Restaurant	Hotel	Indian Restaurant	Cupcake Shop	Dessert Shop	Dhab
4	Borivali	Commercial Rates	Ice Cream Shop	Chinese Restaurant	Snack Place	Burger Joint	Pizza Place	Restaurar

Out[23]:

	Neighborhood	1st Most Common Venue	Venue Count	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	
0	Andheri	Commercial Rates	9	Fast Food Restaurant	Indian Restaurant	Sandwich Place	Food Court	Bakery	
1	Antop Hill	Commercial Rates	2	Diner	Bar	Fast Food Restaurant	Cupcake Shop	Dessert Shop	
2	Bandra	Commercial Rates	10	Café	Indian Restaurant	Brewery	Pub	Restaurant	F
3	Bhandup	Commercial Rates	3	Falafel Restaurant	Hotel	Indian Restaurant	Cupcake Shop	Dessert Shop	
4	Borivali	Commercial Rates	10	Ice Cream Shop	Chinese Restaurant	Snack Place	Burger Joint	Pizza Place	F

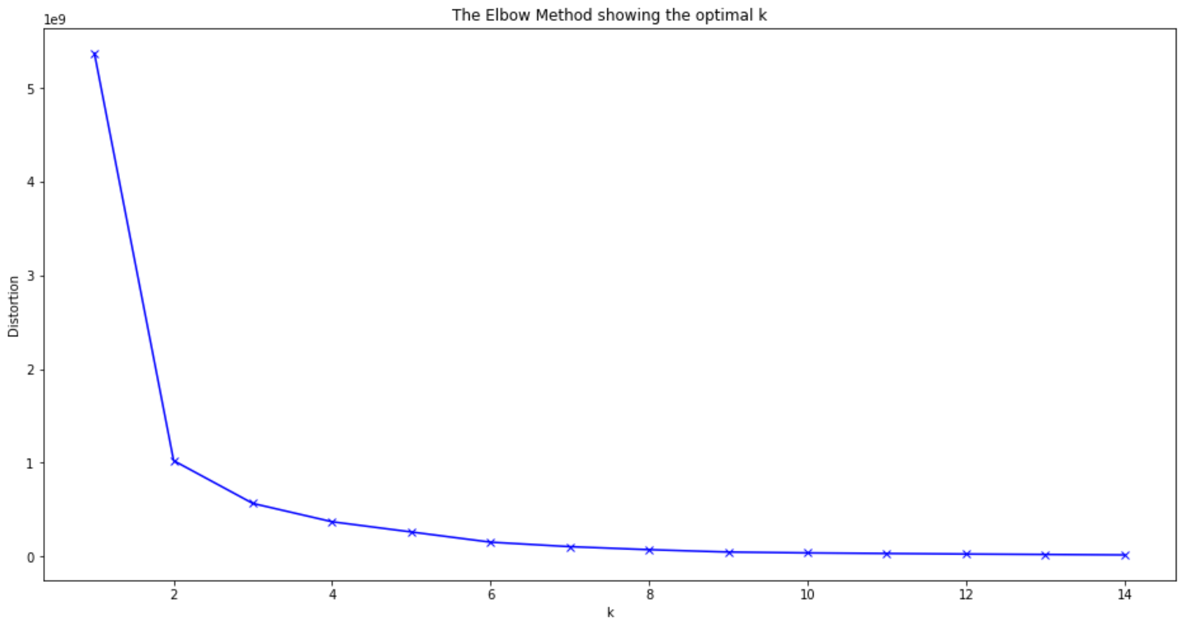
Now, one important thing to consider is that the K-means algorithm (discussed further) splits data into as many clusters as we specify. After all, we could split the data into as many datapoints there are and that would be a perfect split! Alas, we need to find as few significant clusters as we can, in order to make sense of them.

Determining the optimal number of clusters in a data set is a fundamental issue in partitioning clustering, such as k-means clustering, which requires the user to specify the number of clusters k to be generated.

Unfortunately, there is no definitive answer to this question. The optimal number of clusters is somehow subjective and depends on the method used for measuring similarities and the parameters used for partitioning.

Some approches to determine number of clusters consist of optimizing a criterion, such as the within cluster sums of squares or the average silhouette. Most commonly used methods are elbow and silhouette methods.

We will be choosing optimum number of clusters by using **Elbow method**.



The Elbow Method suggests 2 as the optimum number of clusters, as the distortion in the data drops suddenly at 2 clusters and very slowly decreases from there. But considering insights from the exploratory data analysis, we will proceed with grouping the data into **3 clusters**, as it will be easier to gather insights that way.

Out[26]:

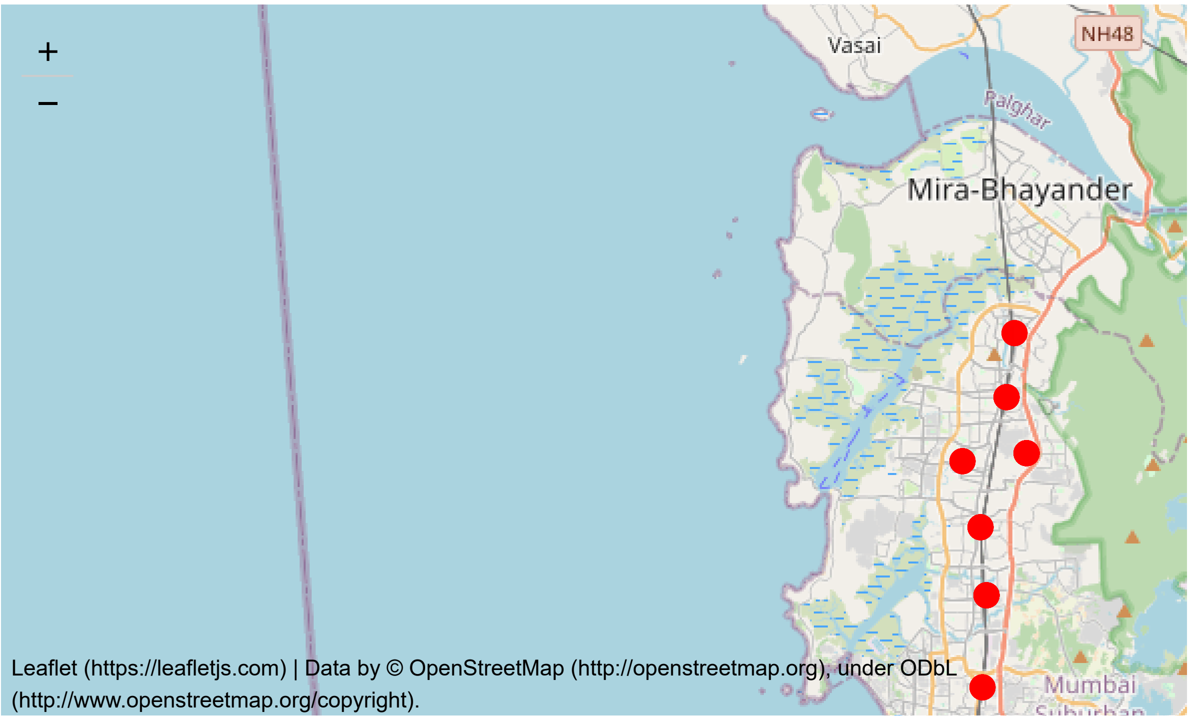
	Suburb	Neighborhood	Commercial Rates	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude
0	Western	Andheri	18639	19.119698	72.84642	Merwans Cake shop	19.119300	72.845418
1	Western	Andheri	18639	19.119698	72.84642	Narayan Sandwich	19.121398	72.850270
2	Western	Andheri	18639	19.119698	72.84642	McDonald's	19.119691	72.846102
3	Western	Andheri	18639	19.119698	72.84642	Cafe Alfa	19.119667	72.843560
4	Western	Andheri	18639	19.119698	72.84642	McDonald's	19.118411	72.848002

5 rows × 21 columns

Now we have the entire datapoints grouped into three clusters, we can view the cluster label in the column 'Cluter Labels'. They take values in [0, 1, 2].

Now we map the neighborhoods by color coding them according to the cluster they belong in.

Out[27]:

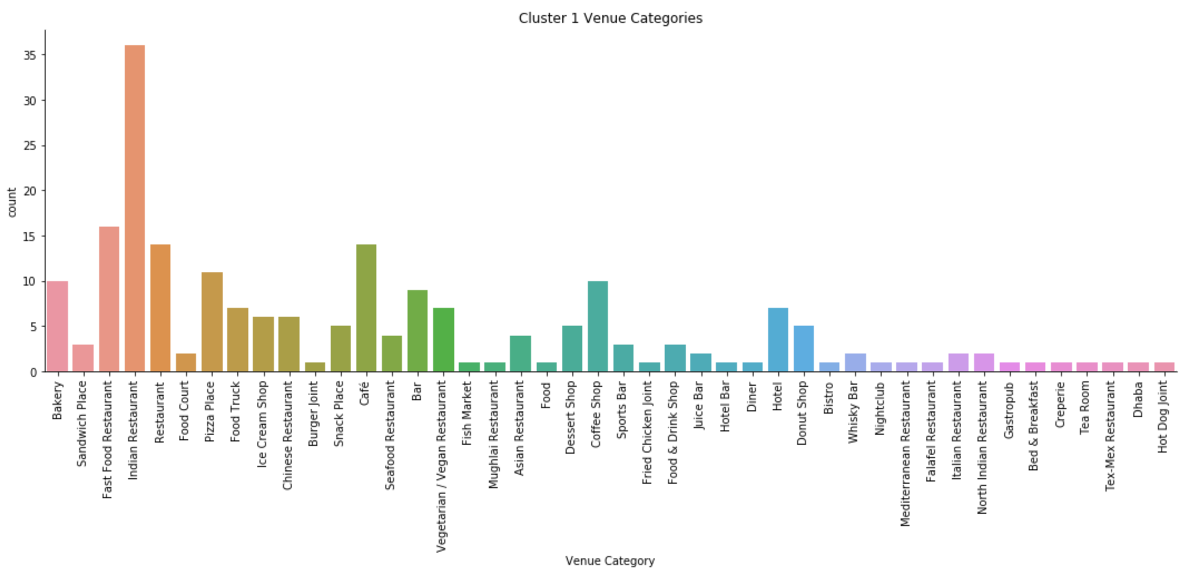


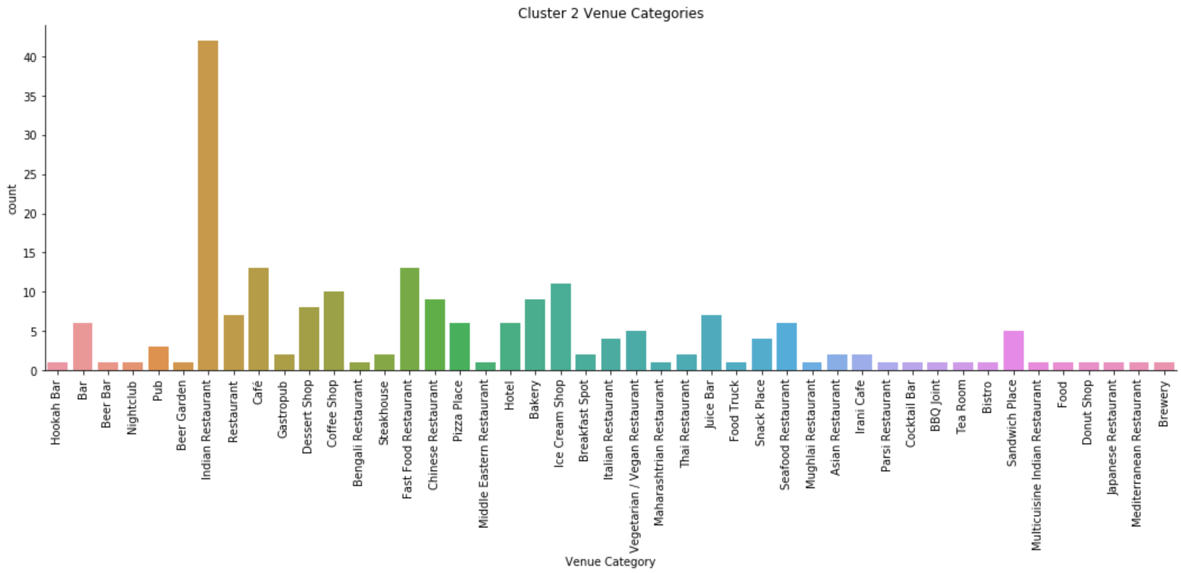
From the first glance, it appears that high-commercial rate neighborhoods have been allocated to the cluster with purple markers and so on. But let's dive into the results of this clustering analysis and derive some insights from it.

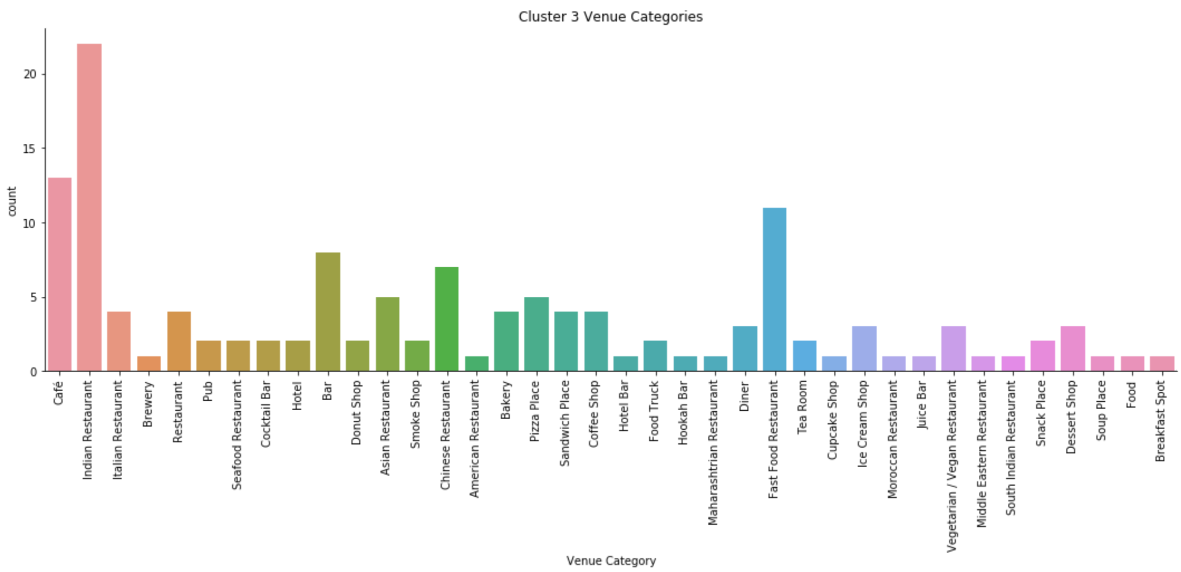
Results and Discussion

We will analyze each cluster with regards to the two important metrics we have at hand: the type of restaurant in the neighborhood and the commercial rates in the neighborhood.

Now we create a distribution plot for venue categories in each cluster. This will give us insight into the kind of food places a cluster contains.

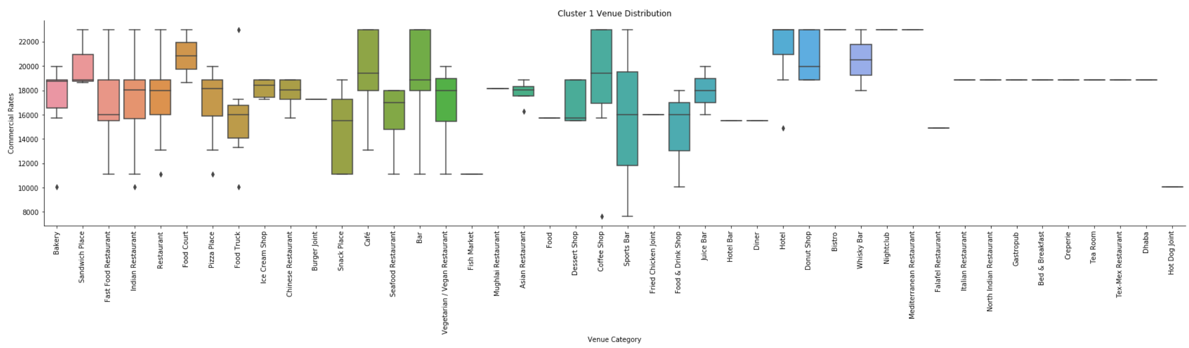


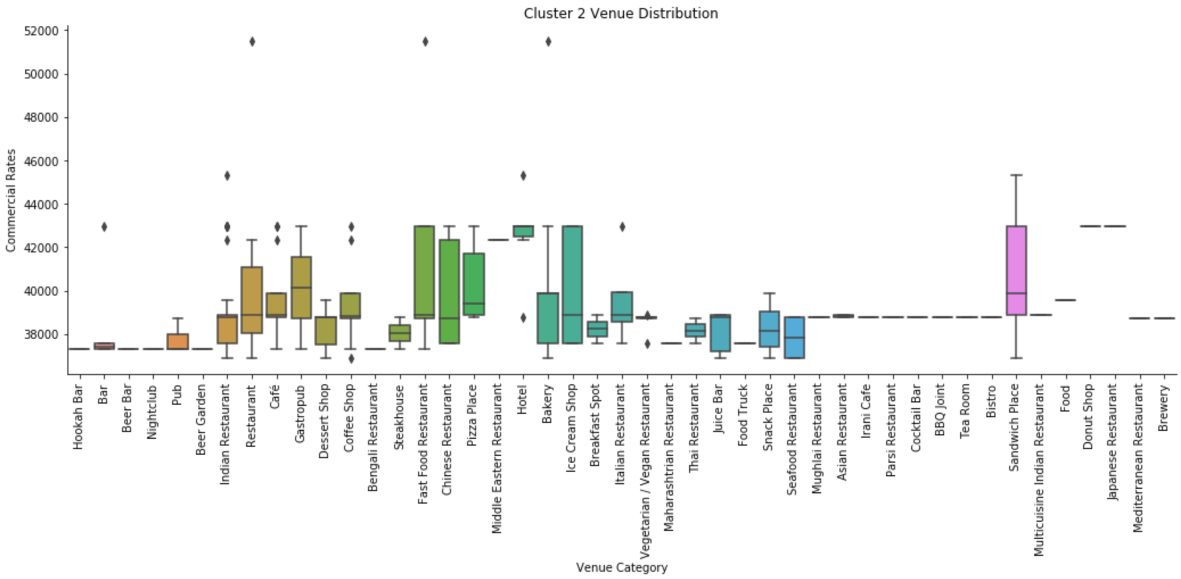


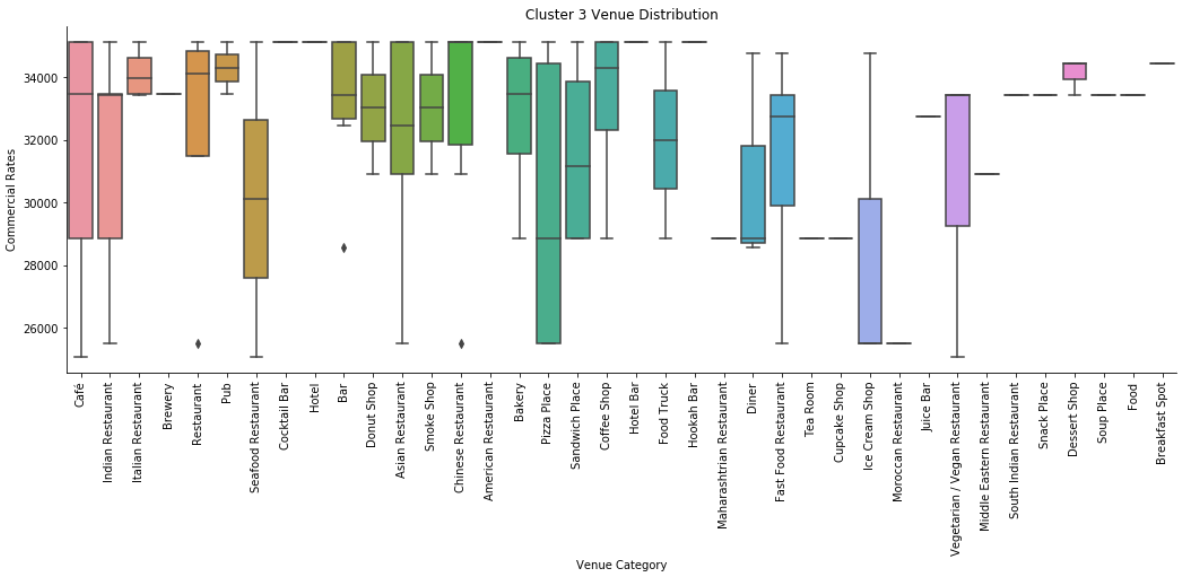


Looking at these charts, it is amply evident that Indian Restaurants, Fast Food Restaurants and Cafes are the three most popular food venues in all of the clusters. One notable observation is that although cluster 3 has fewer number of venues, it has relatively much more balanced distribution overall. This means that neighborhoods in cluster 3 contain a much broader variety of food places as compared to clusters 1 and 2.

Now we take a look at what kind of distribution we can find when we consider commercial rates of different clusters.







We can clearly see that cluster 1 has low commercial rates and a wide variety of food venues. Cluster 2 is the priciest one with numerous outliers, indicating presence of very few luxury food places. This goes hand in hand with the observation that South Mumbai (the area mainly covered by cluster 3) is a pricey business district. Cluster 3, as discussed has a balanced representation of all kinds of food venues. It is demographically richer than cluster 1 and the commercial property rates are moderate.

Recommendation

Out[37]:

```
array(['Bandra', 'Juhu', 'Vile Parle', 'Mulund', 'Vidyavihar', 'Vikhroli',
       'Antop Hill', 'Byculla', 'Kamathipura', 'Matunga', 'Tardeo',
       'Mahim'], dtype=object)
```

A good location recommendation for Beriyan Biryani expansion is South Western Mumbai. Neighborhoods **Bandra, Juhu, Vile Parle, Vikhroli, Byculla, Matunga and Mahim** stand out. Stakeholders are advised to proceed with street level analysis of these neighborhoods for Beriyan Biryani expansion.

Some recommendations to make this analysis even more effective, by enriching the data:

1. Get User specific data from **Google Maps API about keyword - "Biryani"**. For example, how many users in a neighborhood search for Biryani, how many food places in a neighborhood have word "Biryani" in their reviews.
2. Get the online order volume data for "Biryani" from food-delivery platforms such as **Zomato and Swiggy**.

Conclusion

This project analyzed Mumbai neighborhoods for the patterns related to food places. The purpose was to narrow down the Stakeholders' search for optimal location for company expansion to select neighborhoods. Based on the property price trends and food venue distribution across the neighborhoods, we were able to come up with important insights that helped us narrow down the options and give confident recommendations.

Stakeholders were further advised to gather more information by partnerships with appropriate entities so that this analysis can be made more effective. Also, feedback from the street level analysis of neighborhood will be very effective in refining this study to come up with better insights.