# Best location for a Thai restaurant

## In MUMBAI

Based on k-means clustering of geospatial data

Avanti Gogate - 06 July 2020

# 1. Introduction

**Fig 1: A view of the Sea Link in Mumbai**

This project is defined on the basis of available data on the city of Mumbai. The metropolitan city of Mumbai is the most populous city in India and is also the financial capital of the country. It is a major economic hub, and with a touch of both, the rustic and modern elements, it attracts people from all over the world.

In a hypothetical scenario, a successful hotelier and owner of a chain of Asian restaurants called "Thai Jasmin", wishes to open another one, in India. This is his first restaurant in the country, and after several business visits, Mumbai seemed the most obvious choice due to its vibrancy and charm. It is not a cheap investment though, and hence he needs to be careful in his selection of a location in Mumbai: a non-trivial process.

### 1.1 Defining the problem

The idea behind this exercise is to find the best neighbourhood in Mumbai for the restaurant to run successfully and attract both foreigners as well as local residents. There are several factors to consider here, and some more that are out-of scope for this project, but should be considered nonetheless.

- - Finding the most popular neighbourhoods based on all venues: Monuments, shopping places, restaurants etc.

- In these popular neighbourhoods, narrowing down to the number of Asian restaurants
- Choosing a location that our restaurant will be exclusive to (no other Asian restaurants nearby) and is also easily accessible from other popular venues

Other things to consider (out of scope due to lack of data)

- Commercial Rental prices: Whether within budget or not
- Deciding the price tier of the restaurant: Affordable or Expensive?
- Density of population around the location: Segregate residential population from non-residential
- Analyse locations also based on the income of residents: Can they afford the price scale of the restaurant?
- etc

## 2. Methodology

I started by importing all the necessary libraries, such as pandas, folium, sklearn etc.

### 2.1 Loading the data

Next, the geospatial data of Mumbai was loaded. I scraped the table given on this website: https://www.mapsofindia.com/pincode/india/maharashtra/mumbai/ but I retrieved the coordinates of these locations through private communication. The two tables were merged beforehand in preparation for this exercise. An example of the loaded

| | Location | Pincode | State | District | Latitude | Longitude |
|---|---|---|---|---|---|---|
| 0 | A I staff colony | 400029 | Maharashtra | Mumbai | 19.078700 | 72.860050 |
| 1 | Aareymilk Colony | 400065 | Maharashtra | Mumbai | 19.016681 | 72.881740 |
| 2 | Agripada | 400011 | Maharashtra | Mumbai | 18.984810 | 72.816860 |
| 3 | Airport | 400099 | Maharashtra | Mumbai | 19.090052 | 72.868672 |
| 4 | Ambewadi | 400004 | Maharashtra | Mumbai | 18.963010 | 72.823040 |
| 5 | Andheri | 400053 | Maharashtra | Mumbai | 19.115525 | 72.835262 |
| 6 | Andheri East | 400069 | Maharashtra | Mumbai | 19.113657 | 72.869722 |
| 7 | Andheri Railway station | 400058 | Maharashtra | Mumbai | 19.113668 | 72.869711 |
| 8 | Antop Hill | 400037 | Maharashtra | Mumbai | 19.022230 | 72.866554 |
| 9 | Anushaktinagar | 400094 | Maharashtra | Mumbai | 19.035017 | 72.925371 |
| 10 | Asvini | 400005 | Maharashtra | Mumbai | 18.895200 | 72.808790 |

Fig 2: Spatial information for all major localities in Mumbai.

dataframe, df_mum, is shown in Fig 2. There were 173 available neighbourhoods spread across Mumbai.

## 2.2 Use geolocator to calculate the coordinates of Mumbai

```
In [3]: address = 'Mumbai, India'

        geolocator = Nominatim(user_agent="my-application")
        location = geolocator.geocode(address)
        latitude = location.latitude
        longitude = location.longitude
        print('The geograpical coordinate of Mumbai are {}, {}.'.format(latitude, longitude))

        #latitude = 18.9387711
        #longitude=72.8353355

        The geograpical coordinate of Mumbai are 19.0759899, 72.8773928.
```

Fig 3: Latitude and longitude of Mumbai using geocoder

## 2.3 Visualise our Mumbai dataset using Folium

Using the coordinates returned by geocoder, the 173 neighbourhoods were projected on the map of Mumbai ( Fig. 4).

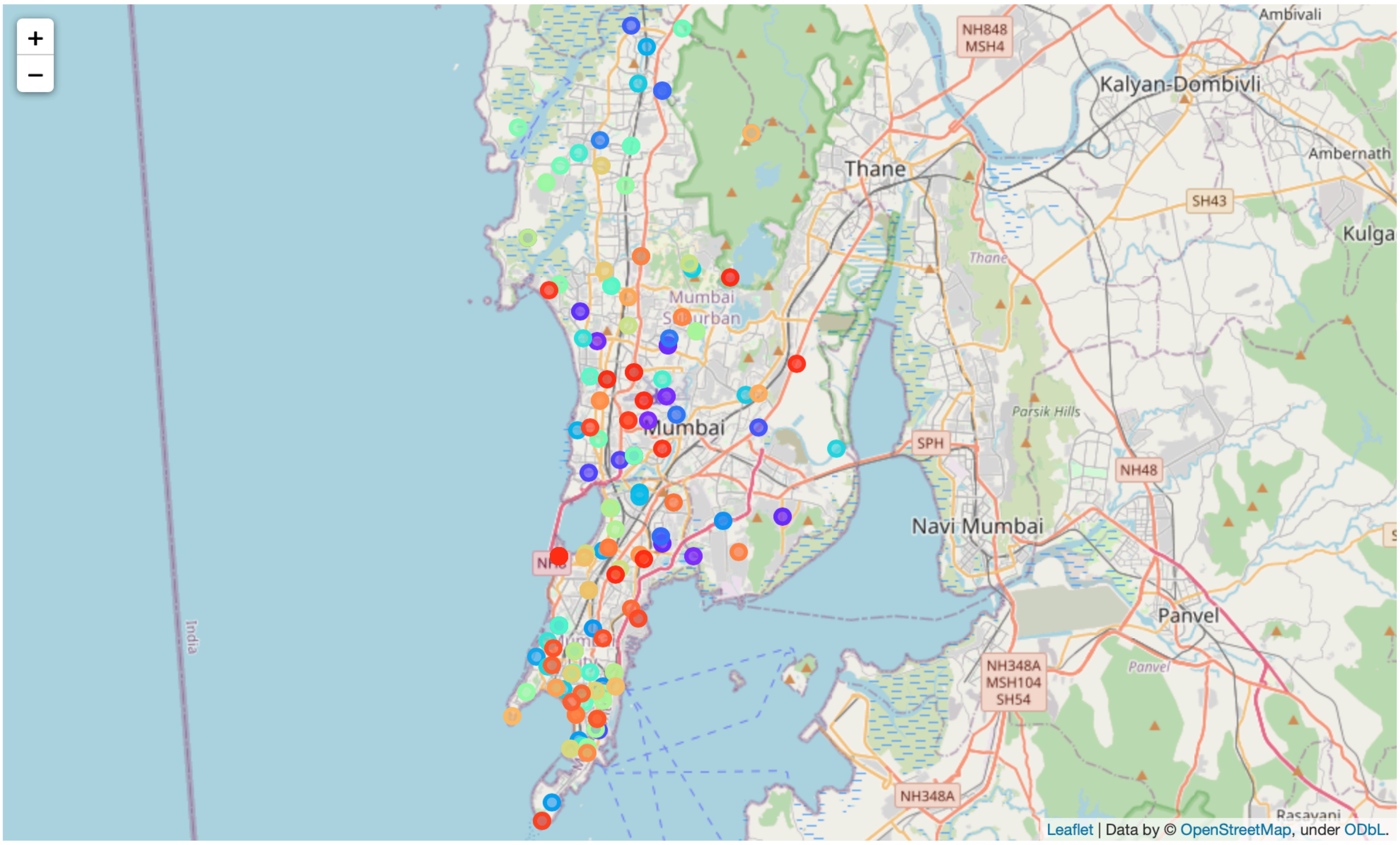


Fig 4: A visualisation of the various neighbourhoods in Mumbai using Folium. Colors are chosen at random and do not have any significance.

## 2.4 The Foursquare API

Using the Foursquare API, the venue data for all the existing neighbourhood were retrieved. I limited the search to a radius of 1 km and the number of venues returned by Foursquare to 100. A pre-defined function was then used to convert the raw JSON files returned by Foursquare into a workable dataframe. Given below are a some entries from the foursquare data frame (referred to as *df_mum_all)*.

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category | Venue Check-in |
|---|---|---|---|---|---|---|---|---|
| 0 | A I staff colony | 19.07870 | 72.86005 | Natural's Ice Cream | 19.077560 | 72.863035 | Ice Cream Shop | 0 |
| 1 | A I staff colony | 19.07870 | 72.86005 | Nilesh Dry Fruits | 19.077578 | 72.864080 | Department Store | 0 |
| 2 | A I staff colony | 19.07870 | 72.86005 | City bakery,vakola | 19.078248 | 72.859005 | Bakery | 0 |
| 3 | A I staff colony | 19.07870 | 72.86005 | Geeta Vihar Hotel | 19.078022 | 72.862714 | Indian Restaurant | 0 |
| 4 | Agripada | 18.98481 | 72.81686 | Gallops | 18.981231 | 72.817178 | Restaurant | 0 |
| 5 | Agripada | 18.98481 | 72.81686 | Little Italy | 18.983885 | 72.815241 | Italian Restaurant | 0 |
| 6 | Agripada | 18.98481 | 72.81686 | Nehru Centre Art Gallery | 18.988617 | 72.814981 | Art Gallery | 0 |
| 7 | Agripada | 18.98481 | 72.81686 | Nehru Auditorium | 18.988726 | 72.814814 | Theatre/Multiplex | 0 |
| 8 | Agripada | 18.98481 | 72.81686 | Viceroy @ NSCI | 18.984465 | 72.815192 | Restaurant | 0 |
| 9 | Agripada | 18.98481 | 72.81686 | Sardar Vallabhai Patel Stadium | 18.986212 | 72.815040 | Stadium | 0 |

Fig 5: An example of the dataset after extracting venue data from Foursquare.

## 2.4 Preliminary analysis of the dataset and data wrangling

Figs 6 and 7 help us visualise the *df_mum_all* dataset. In Fig. 6, we look at all the venues in terms of count, to see which venue categories are the most popular in Mumbai.

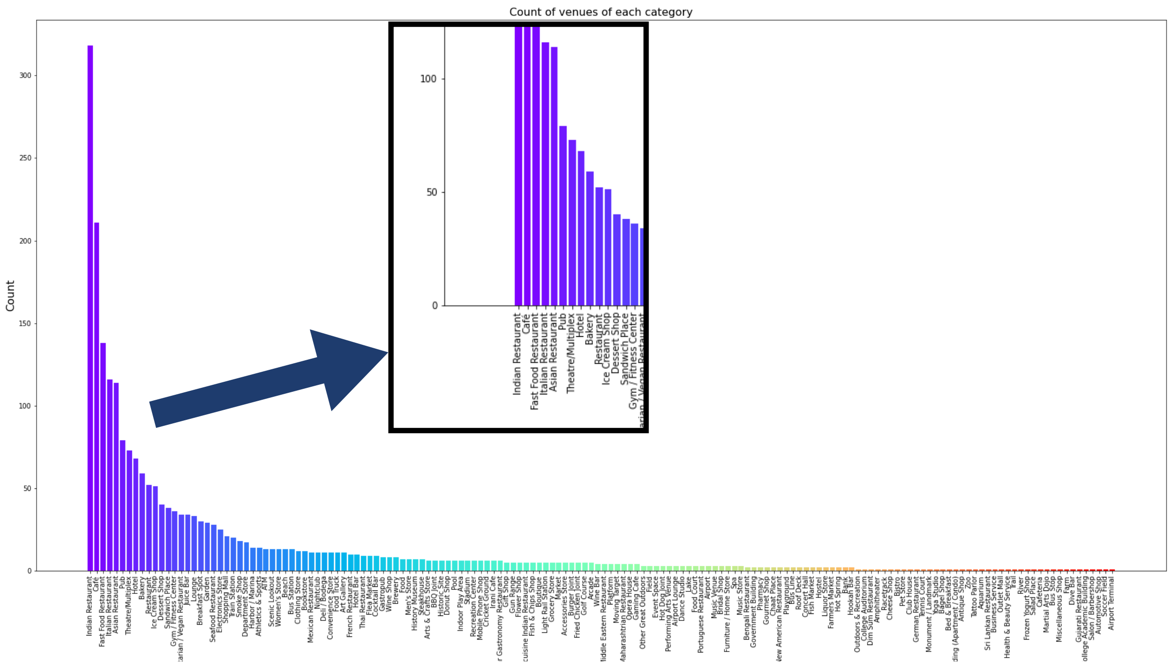


Fig 6: Venue data categories vs count: Summarising the most popular venues in Mumbai

We can see from this plot that Indian restaurants are the most popular, followed by cafes, Italian restaurants and Asian restaurants. It is also observed that there are more than 100 Asian restaurants in Mumbai.
Similar to Fig 6, Fig. 7 shows the number of venues hosted by each neighbourhood.

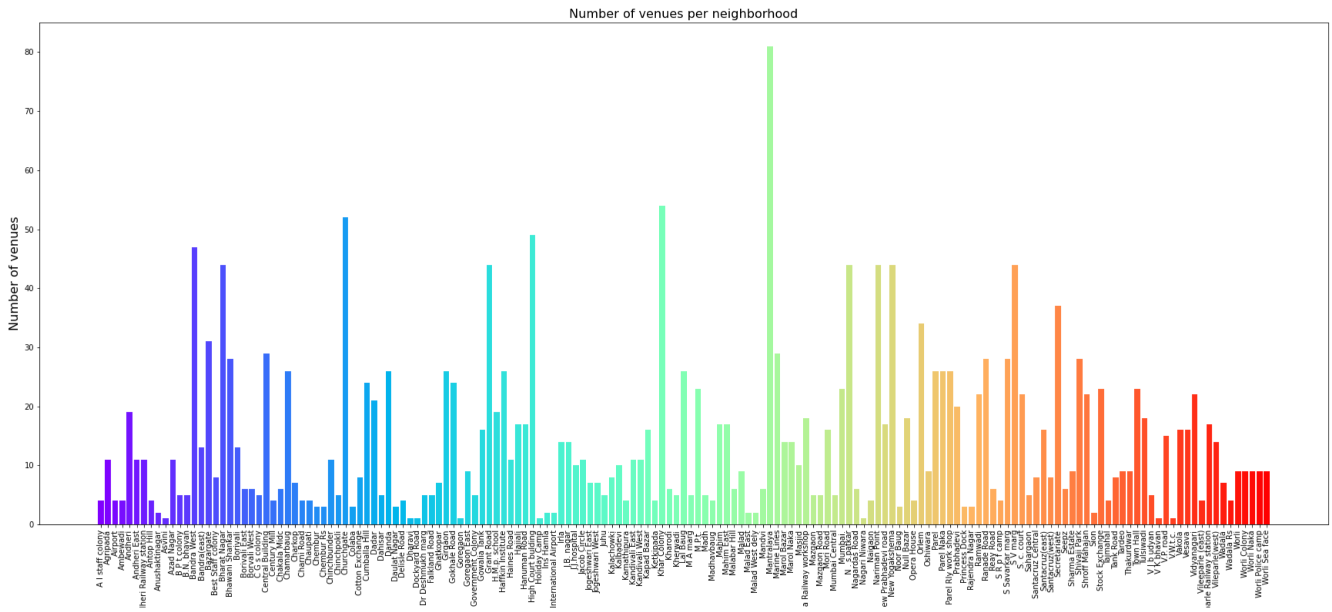


Fig 7: Venue data categories vs count: Summarising the most popular neighbourhoods in Mumbai

From the figure, it is clear that "Mantralaya" seems to be the most popular location with more than 80 Venues. There are several neighbourhoods hosting more than 40 venues, and any of them could be our choice of interest.

Data Wrangling: I used one-hot-encoding to create a dataset that is suitable for clustering analysis. The dataset consists of all venue categories, such as Indian restaurants, Asian restaurants, cafes, pubs, theatres etc.

### 2.5 k-means Clustering

As mentioned earlier, the clustering algorithm takes into account all venue categories, and also includes neighbourhood coordinates. However, finding the best K is a non-trivial process. The "Elbow" method was used to find the best k, which takes into account the sum-of-square distances calculated for a range of k values and the best k is chosen at the elbow point of the graph, where the curve declines exponentially.

Fig. 8 Shows our Elbow method plot. It was not easy to determine which k corresponded to the elbow, but looking closely it could be k = 8 or k =3. I chose k=8 for this analysis.

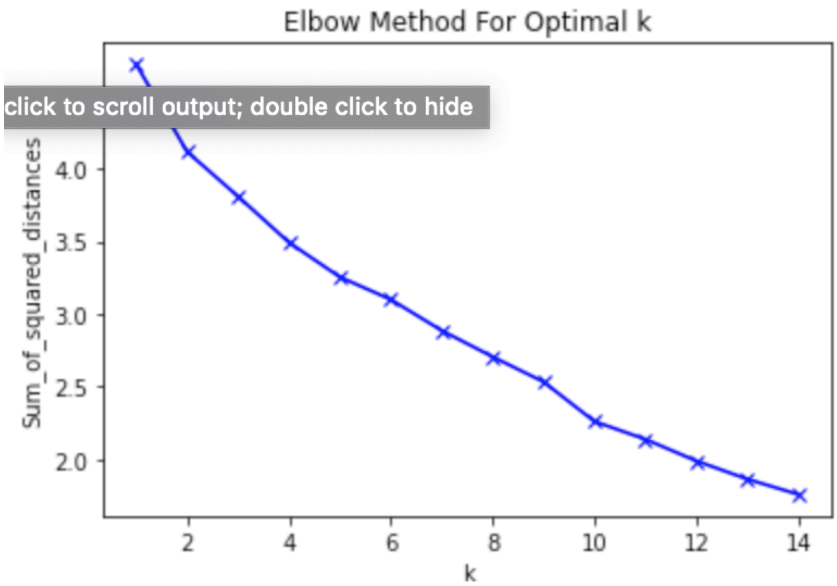


Fig 8: Elbow method for best k.

## Clustering results:


Fig 9: Clustering results. The colours are indicated in the label.

The clustering algorithm returned 8 clusters with varying number of neighbourhoods (see Fig. 9). Cluster 1 was the largest cluster with 27 neighbourhoods. In the table (left) below are given all the neighbourhoods in cluster 1 which already have Asian restaurants. Column 2 indicates the number of Asian restaurants per neighbourhood.

| Neighbourhoods with Asian Restaurants | Freq |
|---|---|
| MANTRALAYA | 6 |
| HIGH COURT BULDING | 5 |
| BANDRA WEST | 4 |
| SECRETARIATE | 3 |
| CHURCHGATE | 3 |
| TULSIWADI | 3 |
| NARIMAN POINT | 2 |
| TOWN HALL | 2 |
| NEW YOGAKSHEMA | 2 |
| MUMBAI. | 2 |
| STOCK EXCHANGE | 2 |
| DANDA | 2 |
| M.P.T. | 2 |

| Neighbourhoods with other venues | Freq. |
|---|---|
| MANTRALAYA | 72 |
| CHURCHGATE | 49 |
| N . S.PATKAR | 46 |
| S V MARG | 46 |
| CHOWPATTY | 46 |
| GRANT ROAD | 46 |
| BANDRA WEST | 43 |
| HIGH COURT BULDING | 42 |
| NARIMAN POINT | 40 |
| NEW YOGAKSHEMA | 40 |
| SECRETARIATE | 33 |
| GIRGAON | 26 |
| CUMBALLA HILL | 24 |
| DANDA | 22 |
| SHROFF MAHAJAN | 20 |
| RAMWADI | 20 |
| S. C. COURT | 20 |
| TOWN HALL | 18 |
| MUMBAI. | 18 |
| M.P.T. | 18 |
| STOCK EXCHANGE | 18 |
| VILEPARLE(WEST) | 17 |
| IRLA | 17 |
| HAJIALI | 17 |
| GOWALIA TANK | 17 |
| TULSIWADI | 14 |
| J.J.HOSPITAL | 11 |

Fig 10: Neighbourhoods with Asian restaurants and their counts (left) and all neighbourhoods in cluster (right)

In table on the right in Fig. 10, I indicate all the neighbourhoods in Cluster 1, with the ones without an asian restaurant marked in green. Comparing the two neighbourhood columns, it is found that NS Patkar, SV Marg, Chowpatty, Grant Road, Girgaon, Cumballa Hill, Shroff Mahajan, Ramwadi, S. c. court, Vile Parle West, Irla, Hajiali, Gowalia Tank and JJ hospital are the remainder neighbourhoods, in decreasing order of number of venues.

# 4. Conclusions

- SV Marg, Chowpatty, Grant road, NS. Patkar and Girgaon are the top 5 locations to open an Asian (Thai restaurant) in a exclusive, yet popular place which has no other Asian restaurants within a search radius of 1 km.

- This conclusion is based on the largest cluster returned by the algorithm. Other clusters need to be examined further for any other conclusions to be drawn.

- This analysis is fairly limited since it does not take into account the commercial aspects such as the rent, menu price, income etc. More data is required for further analysis. Venue data can also be improved upon by using the premium version of Foursquare which is unavailable for this project.

- The choice of k can be tricky and needs to be looked at closely in order for the clustering to give the right results.