CAPSTONE PROJECT REPORT

IBM Applied Data Science Capstone

Opening a new RESTAURANT in MUMBAI, INDIA

By

Likhith Harish

INTRODUCTION

For many citizens, dining at restaurants is a major part of their lives. So, this has been a huge market for investors and business people to invest into many restaurants. So, there is a high competition into this market. Hence opening a new restaurant is a difficult task. The first problem comes with the location. A new market player can not open his restaurant in a location where many restaurants are already triumphing over competition. So, they should find a place which has low concentration of restaurants and with more customers. If a new comer opens a restaurant in a location with high competition, the business wont be feasible and they finally have to shut down .

BUSINESS PROBLEM:

The objective of this capstone project is to find an appropriate location in city of Mumbai, INDIA. So, we use data science methodologies and machine learning techniques like clustering to solve a very important question for this business. That is, what will be the most feasible location to open a new restaurant in city of Mumbai, India?

TARGET AUDIENCE:

This data and output of the project is most useful to new business entries that are thinking of opening a new restaurant in the Mumbai city. This project is most important to people because of oversupply of restaurants in the city.

DATA

To solve the problem, we will need the following data:

- List of neighbourhoods in Mumbai. This defines the scope of this project which is confined to the city of Mumbai, the most populated metropolitan city of INDIA.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to various restaurants. We will use this data to perform clustering on the neighbourhoods. Venue data will include the latitude and longitude also.

SOURCES OF DATA & METHODS TO EXTRACT THEM:

The Wikipedia page https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Mumbai contains the list of neighbourhoods in Mumbai. There are about 135 neighbourhoods in Mumbai. With the help of python requests and Beautiful soup package, we will scrape data from the Wikipedia page and store it in a data frame for further evaluation. Next, we will use Geocoder package in python to get the geographical coordinates of all the neighbourhoods and add them as latitude and longitude columns to the data frame.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Restaurant category in order to help us to solve the business problem put forward. So, using Foursquare API, we will obtain various venue names, their latitude and longitude and along with the venue categories and add them to our dataset. We will then utilise the dataset and choose those neighbourhoods with a venue category of Restaurant. But we can anticipate various categories with Restaurant being included in them as Afghan Restaurant, Indian Restaurant etc. As the location is specified to INDIA and reduce complexity venue category has been limited to Restaurant & Indian Restaurant.

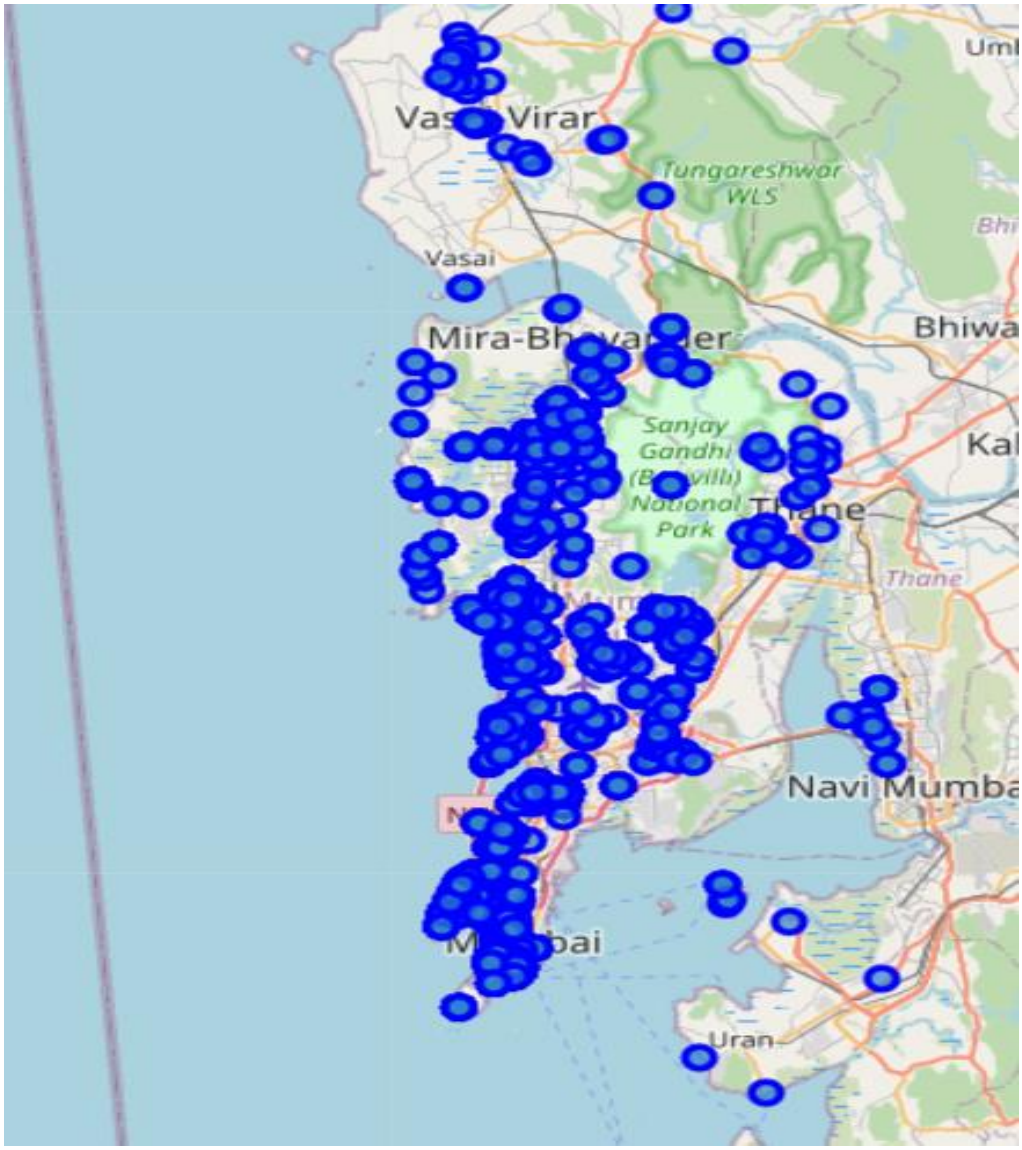
METHODOLOGY

We need certain packages to proceed, so we first have to import geocoder, bs4(Beautiful Soup), folium and certain other packages. Next, we will use beautiful soup package, to create a soup object and parse data from the Wikipedia page and store the data into a list. Then we have to convert this list into a dataframe. This dataframe currently has two columns indexing and 135 Neighbourhoods. Using pandas, add two more columns latitude and longitude with values as '0'.

We will use the geocode method of geolocator to get the latitude and longitude values, venue names, venue category of all the neighbourhoods and store them in the dataframe. The resulting dataframe will be:

	Neighborhood	latitude	longitude
1	Aarey Milk Colony	19.156129	72.870722
2	Agripada	18.975302	72.824898
3	Altamount Road	18.966355	72.809163
4	Amboli, Mumbai	19.131992	72.849960
5	Amrut Nagar	19.100845	72.911820

Again, we will use same method to obtain the coordinates of MUMBAI city. To visualise the neighbourhoods over the city, we will use folium package to insert all the neighbourhoods over the city as markers and save it as map_mi.



Next, we will use Foursquare API to get the top 500 venues that are within a radius of 10000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues.

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Aarey Milk Colony	19.156129	72.870722	The Westin Mumbai Garden City	19.172654	72.860518	Hotel
1	Aarey Milk Colony	19.156129	72.870722	PVR Cinemas	19.174016	72.860485	Multiplex
2	Aarey Milk Colony	19.156129	72.870722	Starbucks	19.174177	72.860350	Coffee Shop
3	Aarey Milk Colony	19.156129	72.870722	British Brewing Company (BBC)	19.174186	72.860504	Brewery
4	Aarey Milk Colony	19.156129	72.870722	MadOverDonuts	19.173902	72.860185	Donut Shop

Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Restaurant” data, we will filter the “Restaurant” as venue category for the neighbourhoods. Later we will filter “Indian Restaurant” as venue category also.

We shall use the following data to perform k-Means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 5 clusters based on their frequency of occurrence for “Restaurant”. We will also perform the same with “Indian Restaurant” Category.

RESULTS

The K-Means Clustering has divided our neighbourhoods into 5 clusters namely.

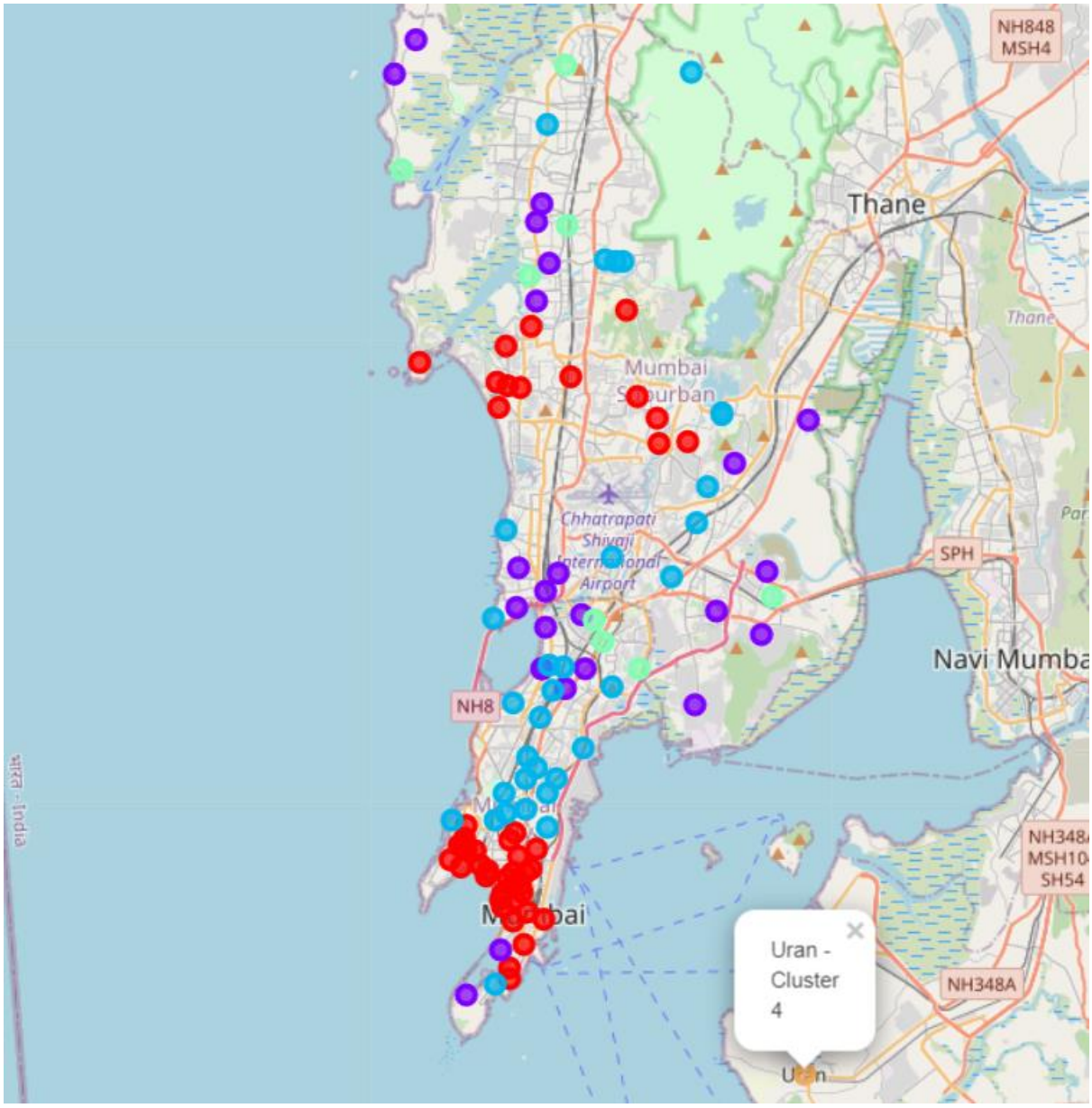
Cluster 0(Red): Neighbourhoods with low number of restaurants.

Cluster 1(Purple): Neighbourhoods with moderate number of restaurants.

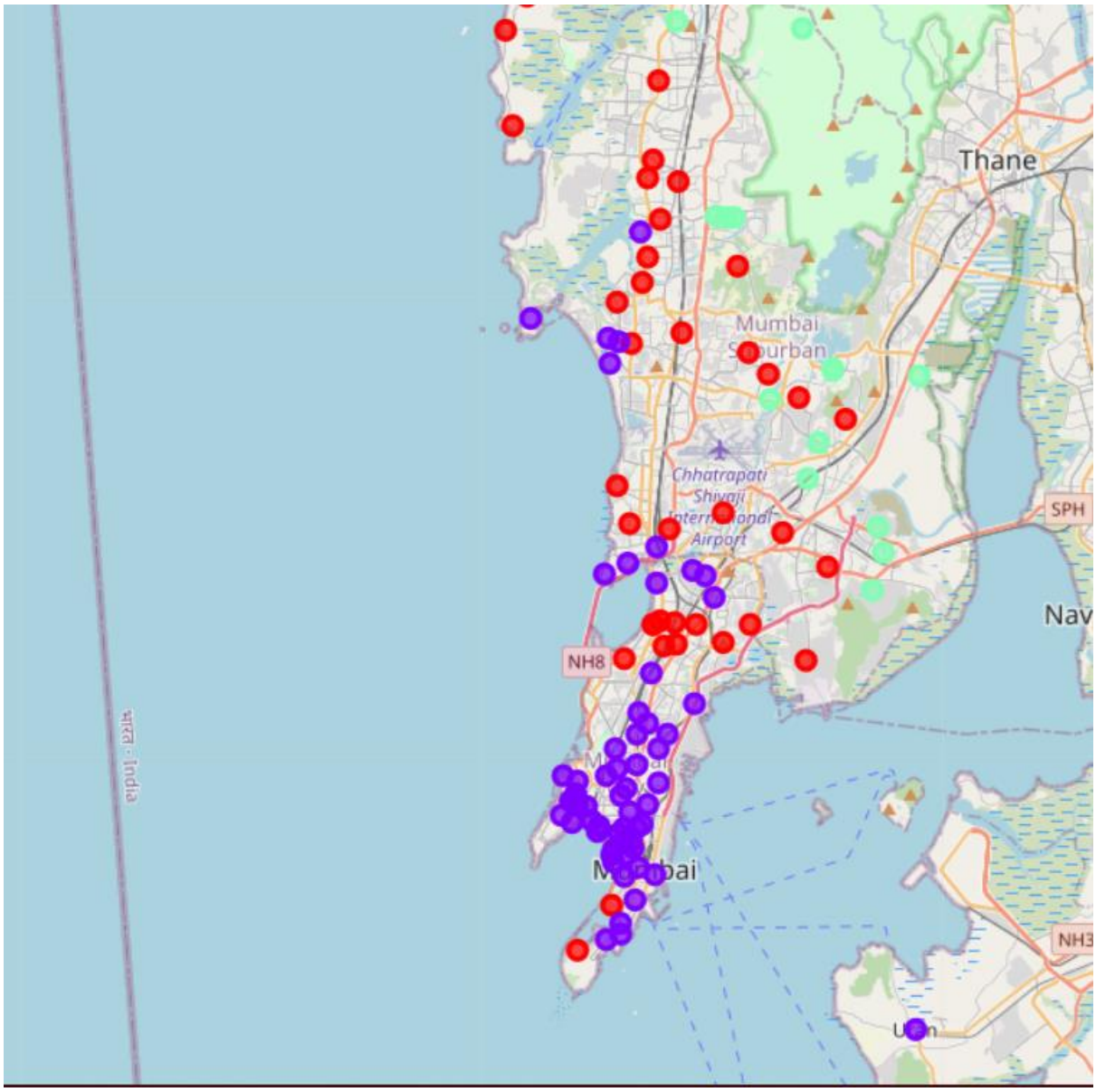
Cluster 2(Blue): Neighbourhoods with semi-moderate number of restaurants.

Cluster 3(Green): Neighbourhoods with high number of restaurants.

Cluster 4(Orange): Neighbourhoods with no / significantly low number of restaurants.



With “Indian Restaurant” venue category, we set the number of clusters to three and we obtain:



The k-Means clustering doesn’t have much use in this category. Because the cluster frequency difference between the three clusters is very less significant.

DISCUSSIONS

As we look into the category of “Restaurant”, most of the neighbourhoods are concentrated in the part of the city surrounded by the sea. But the cluster 0 falls into that area with moderate number of restaurants and cluster 4 with no restaurant category at all.

If we look into the results from Indian Restaurant category, most of the neighbourhoods contain Indian restaurants and since Mumbai is a huge city with Indian diversities too. So, we can assume there are variety of Indian Restaurants in the Mumbai. So as a positive market player, our business can go with being an Indian Restaurant and try to remain in the cluster 0,2 & 4. But the location can be somewhere in the city, where the cluster density is low.These clusters include a huge number of neighbourhoods too, compared to Cluster 3&5. So, we can minimise the task of picking a neighbourhood in this cluster by choosing neighbourhood with low frequency of restaurants in the neighbourhood of cluster.

FUTURE SCOPE OF PROJECT

We only used one factor, frequency of restaurants. But there are certain other factors as Population density, income of the people etc. With more factors coming into play, the efficiency may be increased and we can recommend a more exact location along with latitude and longitude info to the business. However, obtaining this data is out of scope, but can be very helpful.

CONCLUSION

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 5 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. restaurant operators regarding the best locations to open a new restaurant.

To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 0 are the most preferred locations to open a new restaurant preferably of type Indian. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new restaurant.

REFERENCES

- Category – Neighbourhoods in Mumbai
  https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Mumbai
- FourSquare API developer documents
  https://developer.foursquare.com/docs