Capstone Project- The Battle of Neighbourhoods

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE

Suggestions for a one-day stay.

Introduction

In this project I want to tackle the problem which I've faced myself many times. Often you would go to a city for a business trip or maybe you have a flight layover, or any other way through which you get a few spare hours in an unknown city. Now, instead of browsing on your cellphone, you might want to go out and look at the city. I would like to make a framework that gives you the best suggestions depending on your company(group), your budget, the amount of time you have and most importantly, on the basis of your interests.

This problem is quite common and the foremost audience is going to be the people who have a busy schedule and don't have a lot of time to do this research; or it could be the people who got into the position of this outing spontaneously. In such situation it can a be a very handy tool.

I'm going to start this project with **Mumbai, India**– the financial capital of India. Two reasons for choosing this are, firstly that I live in the city so I'll have first hand knowledge to cross verify the results of this subjective problem and secondly that it's frequented place and has a good amount of data available.

If the framework works well it can be used for any other location in India as well.

Currently it'll be limited to India because in the framework, along with Foursquare data, I've used Zomato data, that has a comprehensive data of reviews of restaurants and many other similar venues.

Data Acquisition and Cleaning

Data Sources

To get location and other information about various venues in Mumbai, I have used two APIs and decided to proceed with a combined data from them together.

Using the Foursquare's explore API (which gives venues recommendations), I fetched venues up to a range of 15 kilometers from Mumbai International Airport, given the fact that it's huge city and most of North Mumbai is essentially residential, and collected their names, categories and locations (latitude and longitude).

Mumbai is port city having a shape more longer than circular. We know that the northern part of the city is mostly residential complexes, so I've set the centre of Mumbai to a little more towards the south than the actual coordinates of the airport.

Using the name, latitude and longitude values, I will use the Zomato search API to fetch venues from its database. This API allows to find venues based on search criteria (usually the name), latitude and longitude values and more.

From Foursquare API (https:// https://foursquare.com/developers/apps), I retrieved the following for each venue:
• Name: The name of the venue.
• Category: The category type as defined by the API.
• Latitude: The latitude value of the venue.
• Longitude: The longitude value of the venue.

From Zomato API (https://developers.zomato.com/api), I retrieved the following for each venue:
• Name: The name of the venue.
• Address: The complete address of the venue.
• Rating: The ratings as provided by many users.
• Price range: The price range the venue belongs to as defined by Zomato.
• Price for two: The average cost for two people dining at the place.
• Latitude: The latitude value of the venue.
• Longitude: The longitude value of the venue

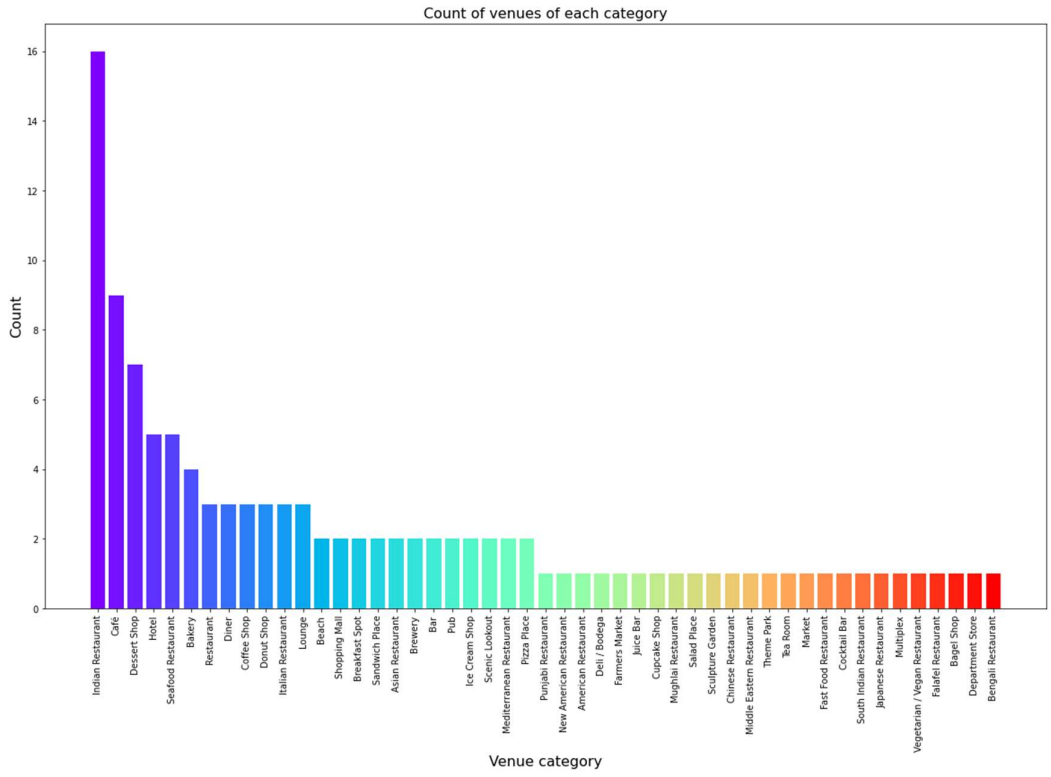
Data Cleaning

This data will be cleaned.
The data from multiple resources might not always align. Thus, it is important to combine the data retrieved from multiple resources properly.

We'll first plot the two data points on the map. We'll then try to combine data points that have their latitude and longitude values very close to one another. From the remaining selected venues, we will inspect the venues to ensure that any remaining mismatched venues are also removed from the final dataset of venues before we begin any analysis.

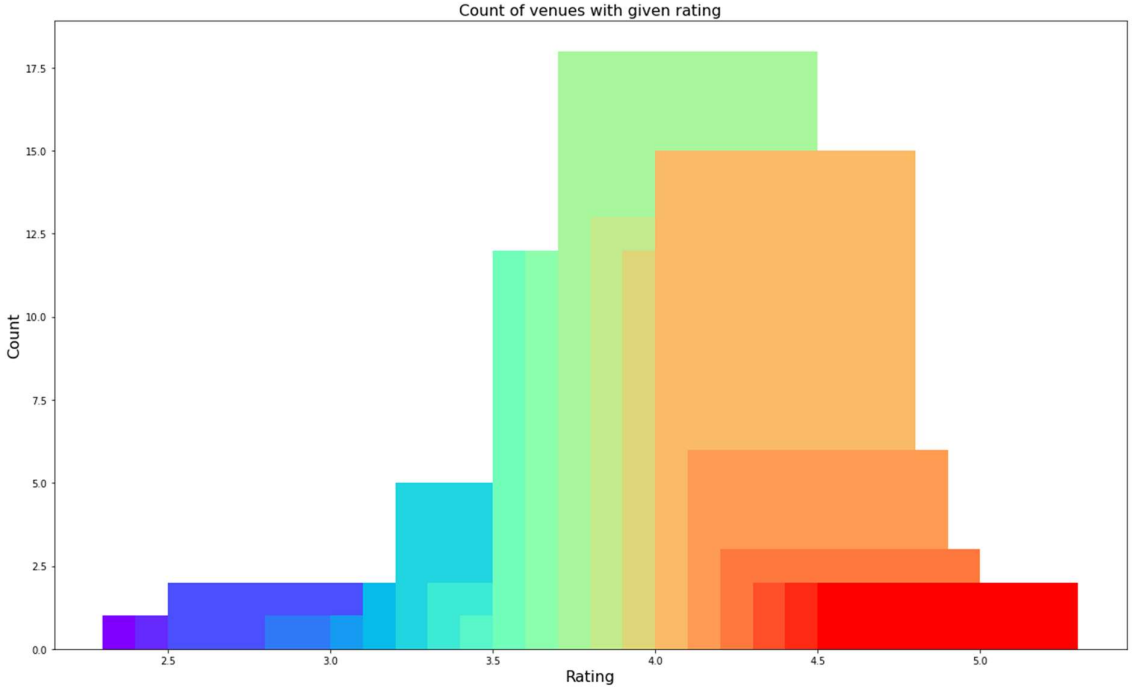
Exploratory Data Analysis

We tried to find the frequency of the venue types. This will give us an idea of the type of locations and the variation in options for the users,



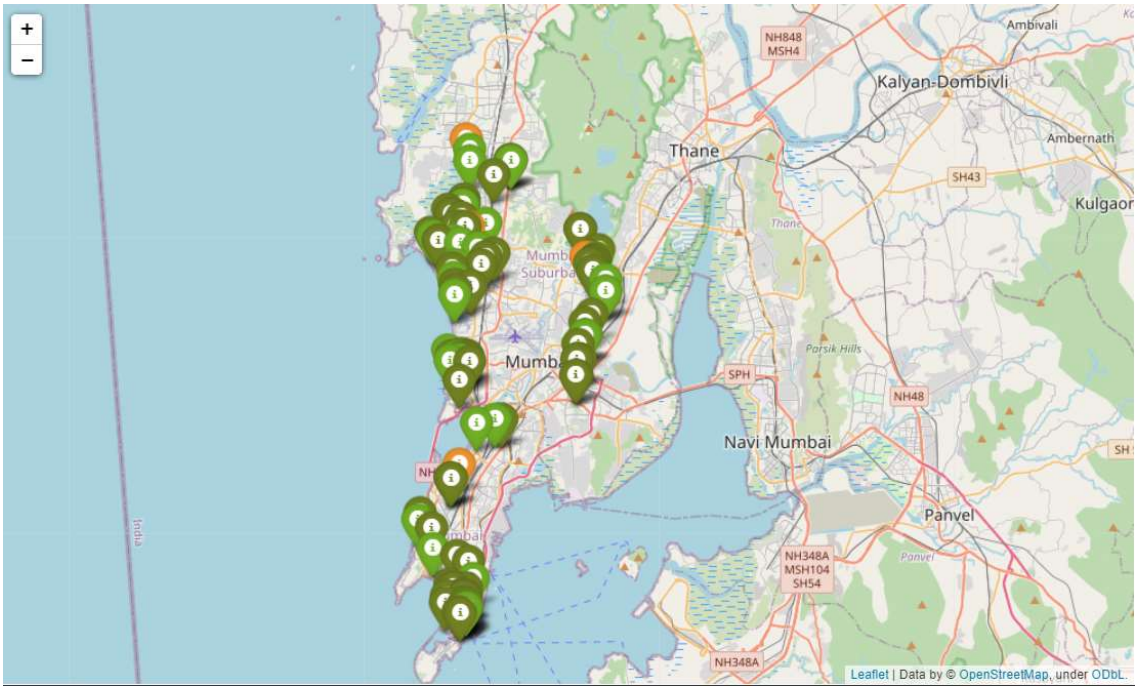
As we can see the maximum number of counts is for Indian Restaurants and Cafes.

Further exploratory analysis was conducted between the rating and number of venues.



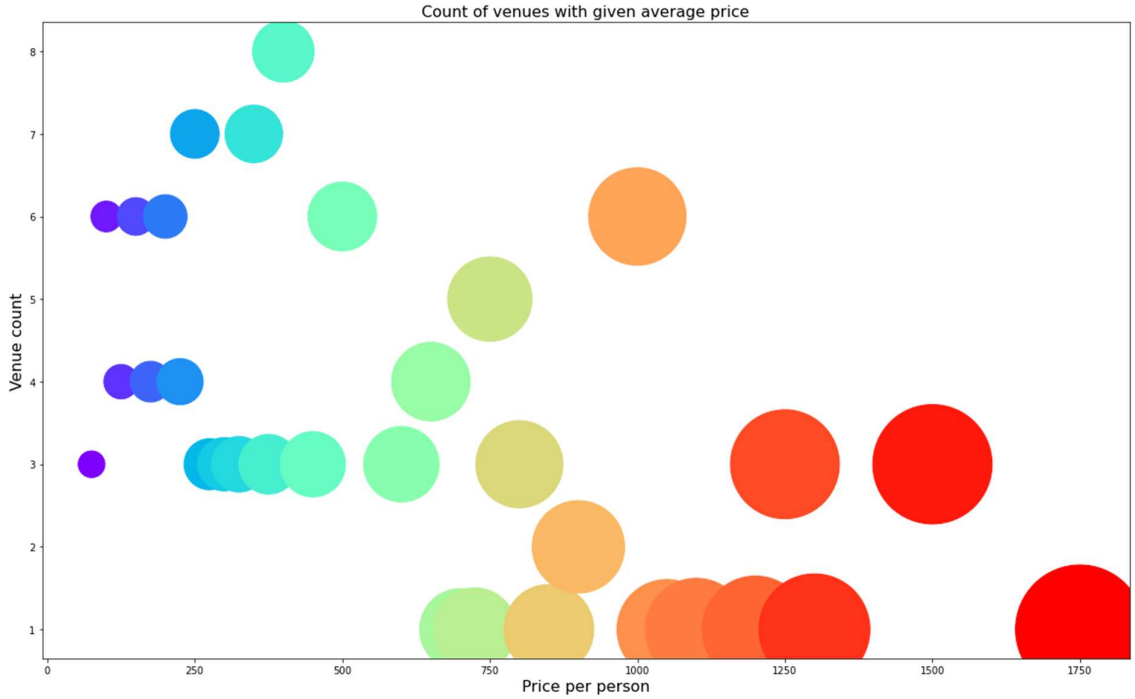
As we can see from the representation, most venues have a rating of above 3.7 to 4 which is a pretty decent rating.

When we colour-coded the locations on the basis of the rating range then we got this.



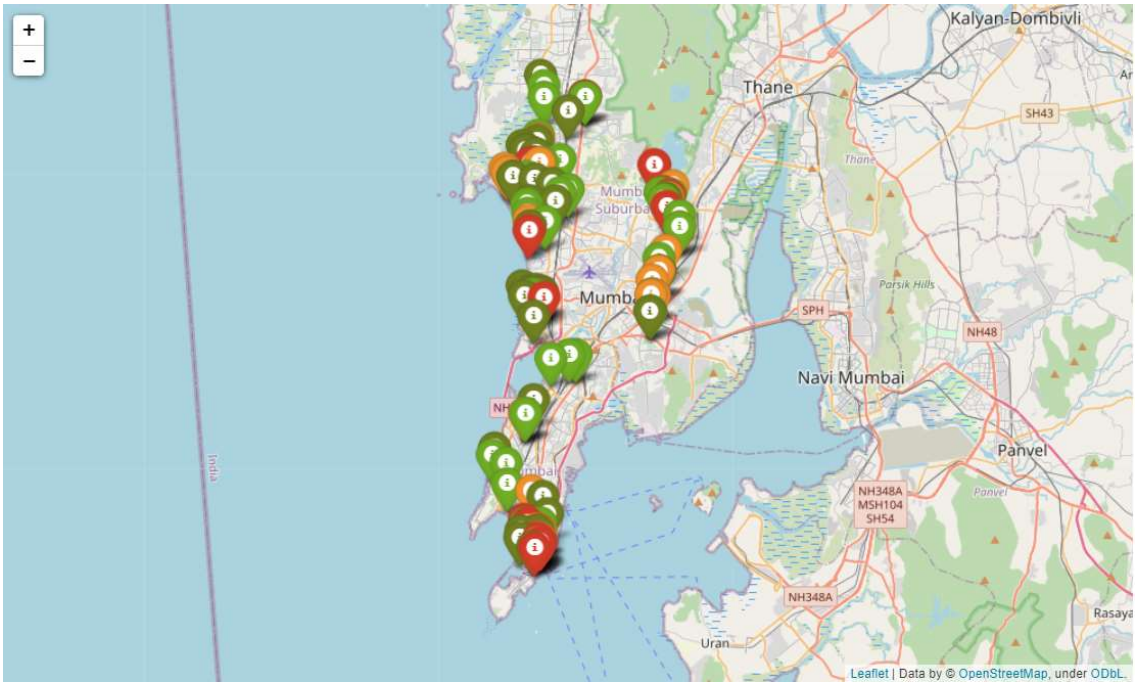
Here, greener reflects more rating for the venue. As we can see, most venues have a good rating.

We now looked at the pricing of the venues in the location.



As we can see that there is a huge range in price. The lowest price is somewhere around Rs. 50 while it goes as high as Rs. 1750 per person. More number of venues are in the average zone, which is still a high number, compared to the other cities in the country.

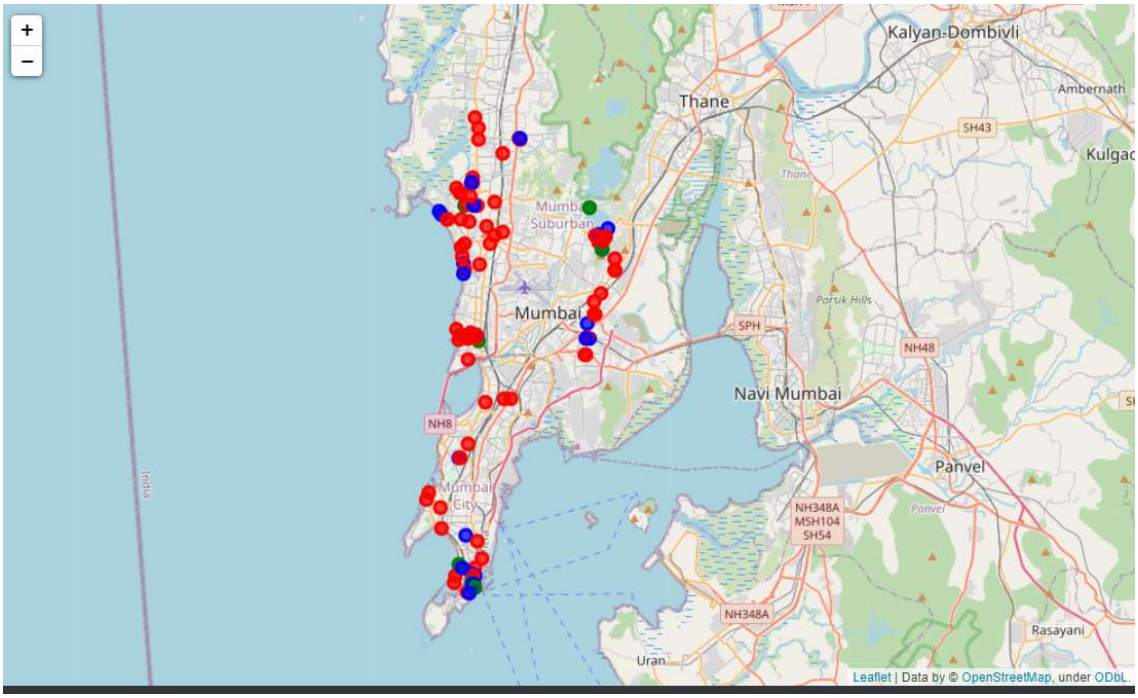
Representing this data on the map, we can see that the expensive locations (marked red) are more spread out and are sparse. The moderate orange ones are in shopping malls and the cheap green venues are the street food shops closer to the beach line.



Further Analysis

We now perform the further clustering analysis on the venues. The clustering is done twice on the basis of two different sets of features. Clustering A is done to make similar clusters on the basis of rating, price and location. These clusters are spread throughout the city and their centres are close to the starting point. This clustering gives us three kinds of clusters.

1. Expensive venues with good rating (4.21)
2. Cheap venues with decent rating (4.04)
3. Moderately priced with excellent rating (4.27)



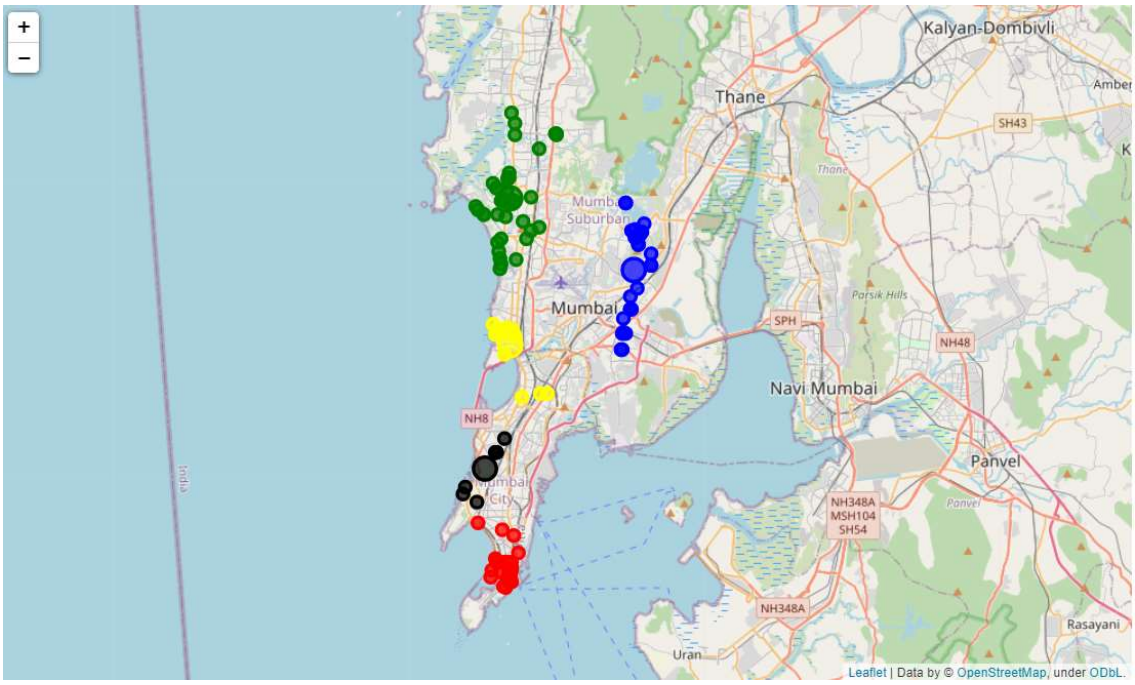
The green dots mark the venues of the first cluster.
The red dots mark the venues of the second cluster.
The blue dots mark the venues of the third cluster.

The second clustering is done on the basis of location. As the city is pretty huge, we need to divide the venues in to zones and then depending on the zone a person can decide on the options available.

On performing the clustering, 5 zones are obtained.

The clusters are:
Juhu-Versova area. This area is near the beaches.
Powai-Ghatkopar area. This area is near the Powai lake and is pivotal for entry from New Mumbai.
Bandra area. This is the area where most Bollywood stars have their residences.
Parel area. This is mostly business offices cum residential areas with lots of cafes and bars.
Colaba area. This is South Bombay, one of the most expensive areas in the city.

These areas are colour coded in the map below.
1. Green- Juhu-Versova
2. Blue- Powai-Ghatkopar
3. Yellow- Bandra
4. Black- Parel
5. Red- Colaba



Results and Conclusion

Based on our analysis above, we can draw a number of conclusions that will be useful to aid any visitor visiting the city of Mumbai, India.

After collecting data from the Foursquare and Zomato APIs, we got a list of 227 different venues. However, not all venues from the two APIs were identical. Hence, we had to inspect their latitude and longitude values as well as names to combine them and remove all the outliers. This resulted in a total venue count of 113.

We identified that from the total set of venues, majority of them were Cafes and Indian Restaurants.

While the complete range of ratings range from 1 to 5, the **majority venues have ratings close to 4**. This means that most restaurants provide **good quality food and service** which is liked by the people of the city and the tourists, thus indicating the high rating.

When we take a look at the price values of each venue, we explore that many venues have prices which are in the range of Rs 250 to Rs 1000 for one person. This also indicates the popular fact that the financial capital of India is indeed an expensive place. However, there are plenty of choices for cheaper venues as well. This is good thing as we have an option for every demand.

Through the first clustering we identified that there are **few venues** which are **high rated (4.21) and have a high price as well**. These belong to the Zeroth cluster. Then there are **many venues** which are **low priced and have average rating of 4.04**. These belong to the First cluster. The Second cluster has **limited venues** which are **moderately priced and are very high rated(4.27)**.

Future Directions

A company can use this information to build up an online website/mobile application, to provide users with up to date information about various venues in the city based on the search criteria (name, rating and price).