Capstone Project Report

Applied Data Sciences by IBM

Opening a Shopping Mall in suburbs of Mumbai, India

By:

Shulin Saraswat

Introduction

Visiting shopping malls is a good way to enjoy, have fun and relax during holidays or weekends. Grocery shopping, dining, watching movies, etc. some of the activities that a shopping mall provides to its customers. Building new malls allows the developer to earn consistent rental income as well as retailers that buy shops in the mall earn from providing from channelling their products and services to the market. Location of the mall place a critical role to determine the success or failure of the mall.

Business Problem

The aim of this capstone project is to analysis and suggest best locations in Mumbai for the developers to build a new mall. Using different data science techniques like clustering, this project aims to answer the business problem.

Target Audience

This project is helpful for the property developers looking to open a shopping mall in the suburbs and fast growing city of Mumbai, India.

Data

- List of suburbs neighbourhood in Mumbai
- Latitude and Longitude coordinates of the neighbourhood.
- Venue data related to shopping malls.

Sources of data and methods to extract

The Wikipedia page https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai contains list of suburbs neighbourhood in Mumbai. With a total of 40 neighbourhoods. To extract the data we will be using Python requests and beautifulsoup packages. Python geocoder package is used to get the latitude and longitude coordinates of the neighbourhood.

Foursquare API will provide us with many venue data like shopping mall data.

Methodology

We get the list of suburbs neighbourhood from Wikipedia page https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai. We used python request and beautifulsoup packages for web scrapping. To get the geographical coordinated of the neighbourhood, we will use GEOCODER package that converts address into coordinates.

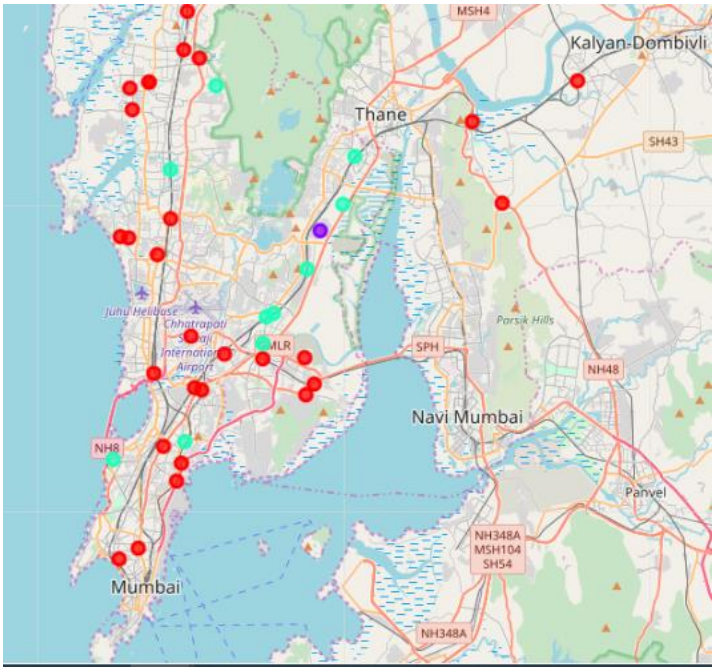
We will use Foursquare API to get top 100 venues within 2Km. We make API calls to foursquare passing the coordinates of the neighbourhood. We will extract he venue name, venue category, latitude and longitude from the venue data returned by foursquare. Since, we are analysing the Shopping mall data, we will filter the Shopping mall as venue category for the neighbourhood.

We perform clustering on the data by using K-means clustering technique. We clustered the neighbourhood into 3 cluster based on the number of occurrence of Shopping mall. This will allow us to identify which cluster has more or less number of shopping mall in the neighbourhood.

Results

The results from k-mean clustering show that we can group the neighbourhood into 3 clusters on the basis of the number of occurrence for Shopping mall

- Cluster 0: It has the cluster 0 has a low number of shopping malls.
- Cluster 1: It has moderate number of shopping malls.
- Cluster 2: It has the highest number of shopping malls.



Discussion

From the observation from the developed results we can say that most shopping malls in Mumbai are in cluster 2. Cluster 0 has very low number of malls in neighborhood. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls. Meanwhile, shopping malls in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. Property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of shopping malls and suffering from intense competition.

Conclusion

In this project, we were able to identify a business problem and provide answers to the problem by using data science methodology and techniques such as clustering. The answer to the problem raised would be a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls in cluster 0.