COURSERA CAPSTONE

IBM Data Science Module

Research to open a Mall in the city of India

Author: Swapnil, Feb 2021

Introduction

Malls are like a one-stop destination for all types of shoppers. For retailers, the central location and the large crowd at the shopping malls provides a great distribution channel to market their products and services. Developers are also taking advantage of this trend to build more malls to cater to the demand. As a result, there are many malls in India and many more are being built. Opening malls allows property developers to earn consistent rental income. Of course, the location of the mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business Problem

The objective of this project is to analyze and select the best locations in the city of India to open a mall. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In India, if a developer is looking to open a mall, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful for developers and investors looking to open or invest in new malls in the Indian city.

Data

- List of neighbourhoods in Indian city. This defines the scope of this project which is confined to the 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 malls. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

This Wikipedia page (https://commons.wikimedia.org/wiki/Category:Suburbs_of_India) contains a list of neighbourhoods in India, with a total of 58 neighbourhoods. I will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautiful-soup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database 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 Mall category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

I will do web scraping using Python requests and beautiful-soup packages to extract the list of neighbourhoods data. However, this is just a list of names. I need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of India. Based on the occurrence of shopping malls in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new malls.

Results
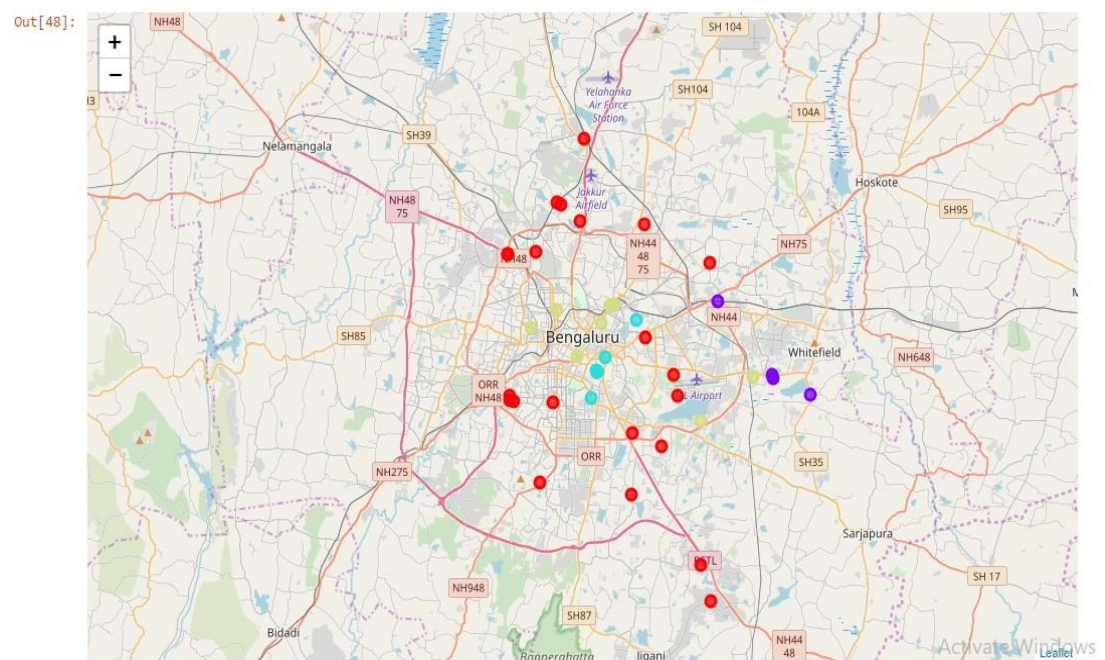
- Cluster 0: Neighborhoods with low number to no existence of malls
- Cluster 1: Neighborhoods with high concentration of malls
- Cluster 2: Neighborhoods with equal concentration malls
- Cluster 3: Neighborhoods with moderate malls

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, cluster 2 in mint green color, and cluster 3 whitish yellow.

Discussion

As observations noted from the map in the Results section, most of the malls are concentrated in the central area of Indian electronic city, with the highest number in cluster 2 and moderate number in cluster 2. On the other hand, cluster 0 has very low number to no mall in the neighborhoods. 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, malls in cluster 3 are likely suffering from intense competition due to oversupply and high concentration of malls. From another perspective, the results also show that the oversupply of shopping malls mostly happened in the central area of the city, with the suburb area still have very few malls. Therefore, this project recommends property developers to capitalize on these findings to open new malls in neighborhoods in cluster 0 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new malls in neighborhoods in cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of malls and suffering from intense competition.

Limitations / Future Scope

In this project, we only consider one factor i.e. frequency of occurrence of shopping malls, there are other factors such as population and income of residents that could influence the location decision of a new shopping mall.

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 "4" clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. developers and investors regarding the best locations to open a new mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 2 are the most preferred locations to open a new mall. 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 mal

References

Category:Suburbs in India. *Wikipedia*. Retrieved from
https://commons.wikimedia.org/wiki/Category:Suburbs_of_India