Coursera Capstone Project

The Battle of Neighborhoods

Malls in Mumbai

Description

Mumbai is the capital city of the Indian state of Maharashtra. Mumbai is the second most populous city in India after Delhi and the seventh most populous city in the world with a population of 19.98 million. With this much population and a lot of tourist that visit Mumbai, Shopping Malls are good places to enjoy, relax and do shopping. In the shopping malls, visitors can watch movies, eat food, play games at the game zone, do shopping of almost anything. Therefore, shopping malls are an all in one destination for tourists and local residents. Mumbai already has a large number of shopping malls but with the growing living standard of residents and tourist visiting Mumbai there are demands of more shopping malls. Shopping malls are a very good investment if opened in a premium location. It will generate good revenue for the investor for a long time and will help people to get everything in just one place.

Problem Statement

To find the best location for opening a Shopping Mall in Mumbai, India. Opening a shopping mall is a big decision and requires a large investment and long time for its construction. Therefore, location of the shopping mall must be chosen carefully.

Target Audience

The main target audience for this project are the real estate builders who want to invest in shopping malls for long term investment or other businessmen who want to shift to investing in real estate.

Data Required

The following data will be used to solve this problem

- Neighborhoods of Mumbai.

- Coordinates of the neighborhoods. Latitude and longitude.

- List of already opened malls in Mumbai.

- Coordinates of these malls.

Data utilization and Techniques used

The data of neighborhoods of Mumbai is collected from Wikipedia. The link https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai is used. The data from the Wikipedia page is extracted using web scrapping technique utilizing the BeautifulSoup package. Geocoder package is used to find out the coordinates of neighborhoods. Folium is used for plotting the map and FourSquare for finding the venues, in our case shopping malls in the neighborhoods.

Methodology

We collected the neighborhoods in the city of Mumbai. The list is available on Wikipedia (https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai). We used web scraping using Python requests and BeautifulSoup packages to extract the list of neighborhoods. We need the geographical coordinates in the form of latitude and longitude in order to use Foursquare API. Therefore, we used the Geocoder package that allowed us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we populated it into a pandas DataFrame and visualized the neighborhoods in a map using Folium. This allowed us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Mumbai. Next, we used Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We made API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare would return the venue data in JSON format and we extracted the venue name, venue category, venue latitude and longitude.

With the data, we checked how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we analyzed each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we also prepared the data for use in clustering. Since we are analyzing the "Shopping Mall" data, we filtered the "Shopping Mall" as venue category for the neighborhoods. Lastly, we performed clustering on the data by using k-means clustering. It is one of the simplest and popular unsupervised machine learning algorithms. We clustered the neighborhoods into 3 clusters based on their frequency of occurrence for "Shopping Mall". The results allowed us to identify which neighborhoods have higher concentration of shopping malls while which neighborhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new shopping malls.

Results

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for "Shopping Mall":

- Cluster 0: Neighborhoods with the moderate number of shopping malls

- Cluster 1: Neighborhoods with the low number to no existence of shopping malls

- Cluster 2: Neighborhoods with the high concentration of shopping malls

Discussion

As observations noted from the map, most of the shopping malls are concentrated in the southern area of Mumbai city, with the highest number in cluster 2 and moderate number in cluster 0. On the other hand, cluster 1 has very low number to no shopping 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, shopping malls in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of shopping 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 shopping malls. Therefore, this project recommends real estate developers to capitalize on these findings to open new shopping malls in neighborhoods in cluster 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping 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 shopping malls and suffering from intense competition.

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 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. real estate developers and investors regarding the best locations to open a new shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations to open a new shopping 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 shopping mall.

Thank You