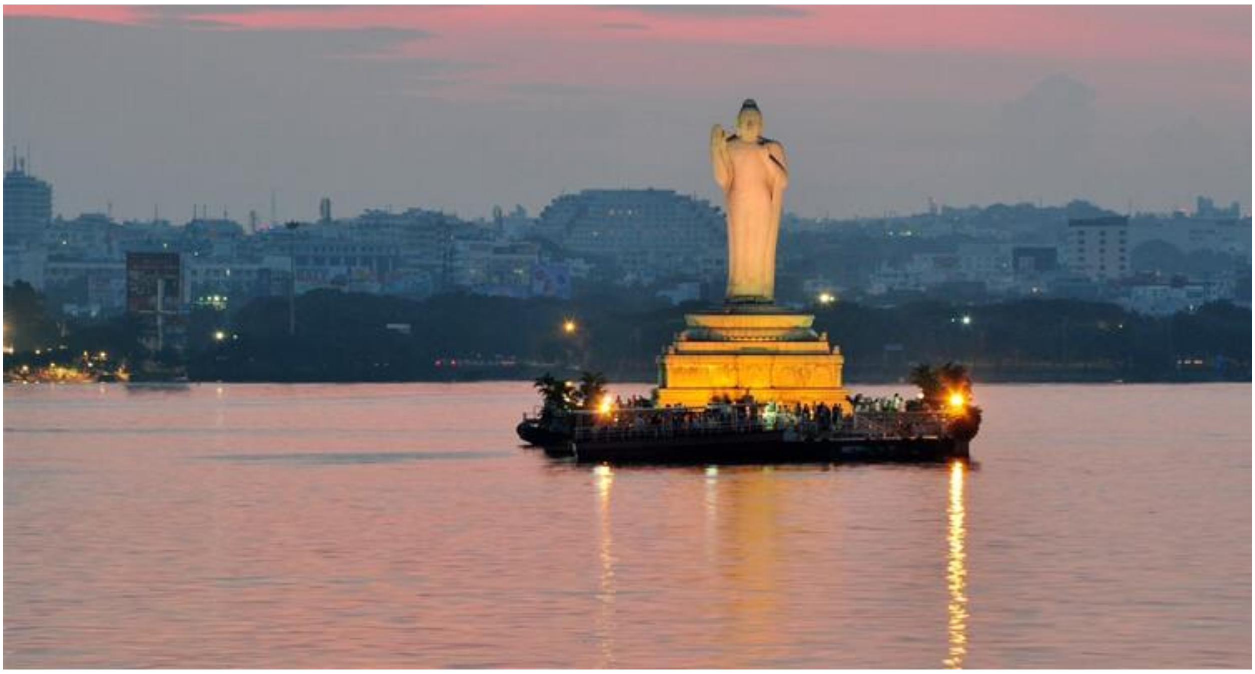
Coursera

IBM Applied Data Science Capstone

Opening a New Shopping Mall in Hyderabad, India



Author: Sourabh Koul

June 2020

Introduction

Shopping 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 provide a great distribution channel to market their products and services. Property developers take advantage of this trend to build more shopping malls to cater to the demand. As a result, there are many shopping malls in the city of Hyderabad and many more are being built. Opening shopping malls allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business Problem

Now arises the business question: In the city of Hyderabad, India, if a property developer is looking to open a new shopping mall, where would you recommend that they open it?

The objective of this capstone project is to analyze and select the best locations in the city of Hyderabad, India to open a new shopping mall.

Target Audience

This project is particularly useful for property developers and investors looking to open or invest in new shopping malls in the city of Hyderabad, India.

Data

To solve the above business problem, we need the following data:

1. List of neighborhoods in Hyderabad, India:

   This data can be obtained from the below Wikipedia page by web scraping data:

   https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India

2. Coordinates (Longitude, Latitude) of the neighborhoods in Hyderabad, India:

   Python Geocoder Package can be used to extract this data

3. Venue data (related to shopping malls):

   Foursquare API can be used to extract the venue details i.e.

   the places data nearby the neighborhoods.

Methodology

First, we need to get the list of neighborhoods in the city of Hyderabad from the page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India). We will do web scraping using Python requests and Beautiful Soup packages to extract the list of neighborhoods data. However, this is just a list of names. We 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 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 Data Frame and then visualize the neighborhoods 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 Hyderabad. Next, we will use Foursquare API to get the top 100 venues that are within a radius of 1500 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 neighborhoods 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 neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze 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 are also preparing the data for use in clustering the places later.

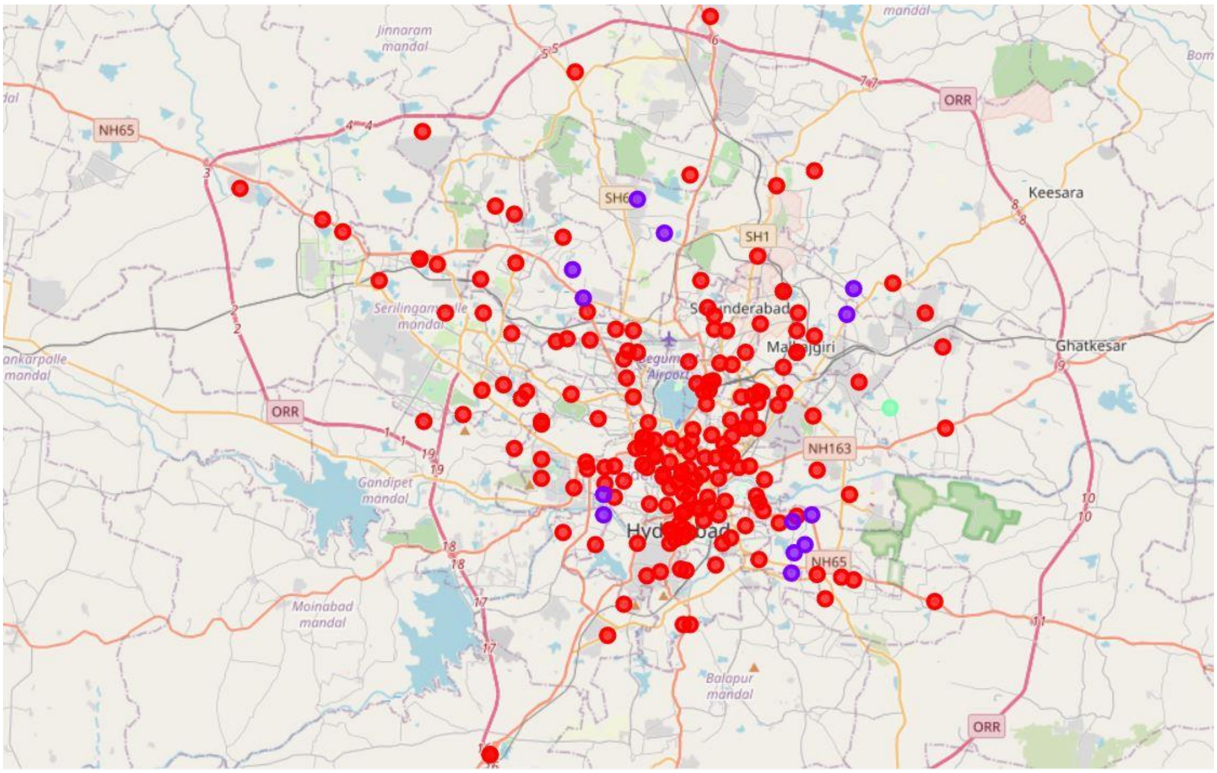
Lastly, we will perform clustering on the data by using K-Means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for "Shopping Mall". The results will allow us to identify which neighborhoods have higher concentration of shopping malls while which neighborhoods have fewer number of shopping malls.

Results

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

- Cluster 0: Neighborhoods with low concentration of shopping malls
- Cluster 1: Neighborhoods with moderate to high concentration of shopping malls
- Cluster 2: Neighborhoods with high concentration of shopping malls

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



Discussion

As we can see from the maps and the table data, most of the shopping malls are present in Cluster 1. Also, there is a high frequency of shopping malls in Cluster 2. On the other hand, Cluster 0 has very few shopping malls. This cluster present a huge opportunity for opening new malls as there will be very little to no competition in trying to open a new shopping mall and generating revenue.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the required data, extracting and preparing the data, clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new shopping mall.

Thus, this project recommends property developers and investors to consider the areas in Cluster 0 for opening a new shopping mall in Hyderabad.