Coursera Capstone

IBM Applied Data Science Capstone

Starting new Multiplex in Hyderabad, India

Harish Reddy Kunduru

March 2019

Hyderabad

Introduction

Many people are interested in visiting multiplex as it is a place where everyone can enjoy. Major concentration of multiplex is having multiple theaters so that people with different tastes can have entertainment at same place. Since it is one stop entertainment place to many people lot of other business like restaurants, children parks and other relay on crowd coming to multiplex. Hyderabad is now having 10 million population and people here have great purchasing capabilities. So, adequate number of Multiplex are required. Hyderabad have good record in ease of doing business. Starting a new Multiplex attracts people, it also enhances other business. Majorly location of Multiplex is key factor in success or failure.

Business Problem

The objective of this capstone project is to analyze and select best locations in the city of Hyderabad for starting new Multiplex. Using Data science methodologies and Machine Learning techniques like Clustering this project aims to provide solutions to answer the business question: In the city of Hyderabad, India, if a property developer is looking to open a new Multiplex, where would you recommend that they open it?

Data

To solve the problem, we will need the following data:

- List of neighborhoods in Hyderabad. This defines the scope of this project which is confined to the city of Hyderabad, the capital city of the country of India in Southern Asia.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to Multiplex. We will use this data to perform clustering on the neighborhoods.

Link (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Hyderabad,_India) contains a list of neighborhoods in Hyderabad, with a total of 199 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.
After that, we will use Foursquare API to get the venue data for those neighborhoods. 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 Multiplex 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.

Method

Firstly, we need to get the list of neighborhoods in the city of Hyderabad. Fortunately, the list is available in the Wikipedia page. We will do web scraping using Python requests and beautifulsoup 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. 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 co-ordinates data returned by Geocoder are correctly plotted in the city of Hyderabad.

Next, we will use Foursquare API to get the top 80 venues that are within a radius of 2500 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. Since we are analyzing the “Multiplex” data, we will filter the “Multiplex” as venue category for the neighborhoods.

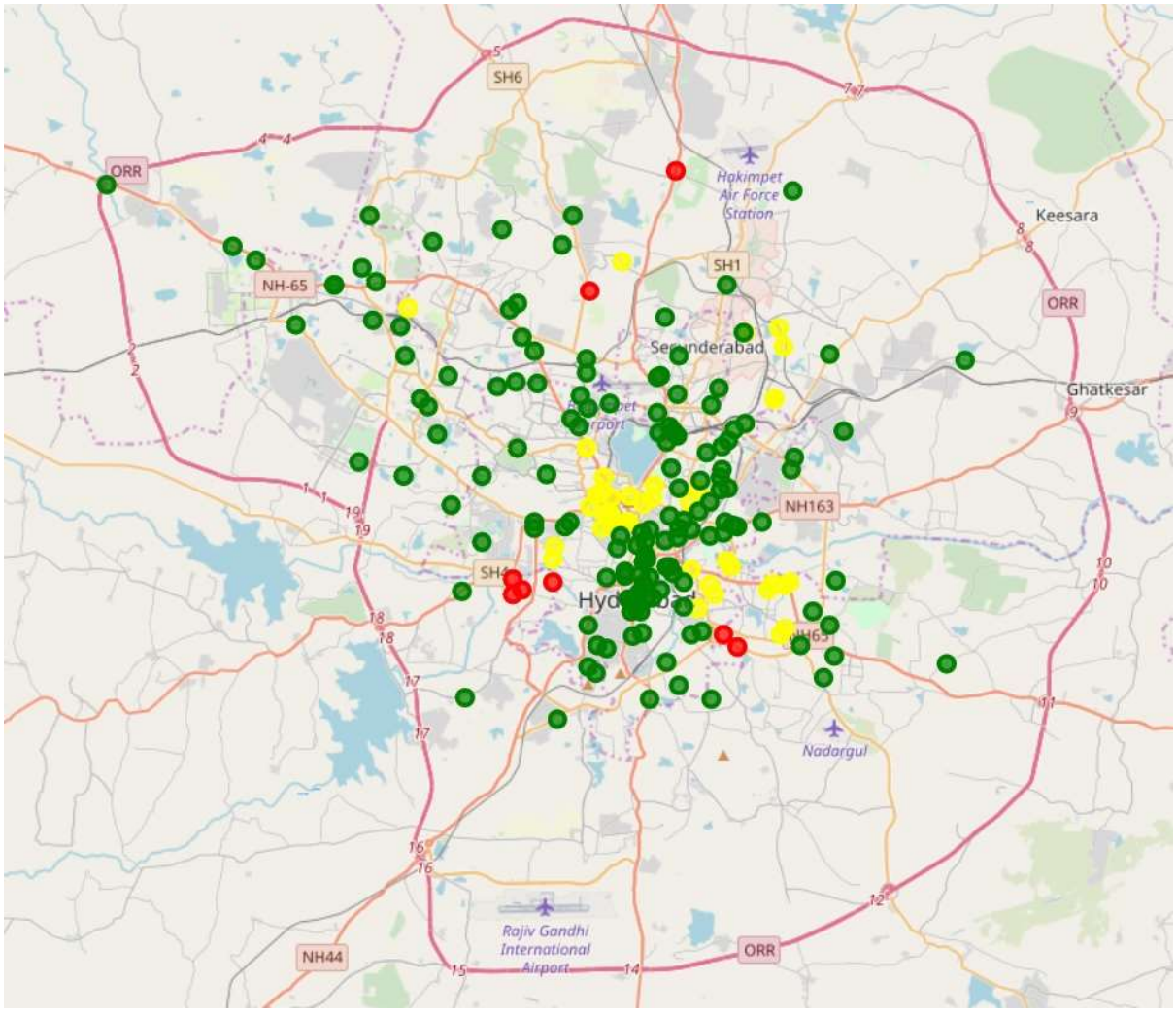
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine 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 “Multiplex”. The results will allow us to identify which neighborhoods have higher concentration of Multiplexes while which neighborhoods have fewer number of Multiplexes. Based on the occurrence of Multiplexes in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Multiplexes.

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 “Multiplex”:

- Cluster 0: neighborhoods with moderate number of Multiplexes
- Cluster 1: neighborhoods with low number to no existence of Multiplexes
- Cluster 2: neighborhoods with high concentration of Multiplexes

The results of the clustering are visualized in the map below with cluster 0 in Yellow color, cluster 1 in Green color, and cluster 2 in mint Red color.



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. property developers and investors regarding the best locations to open a new Multiplex. 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 start a new Multiplex. 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 Multiplex.