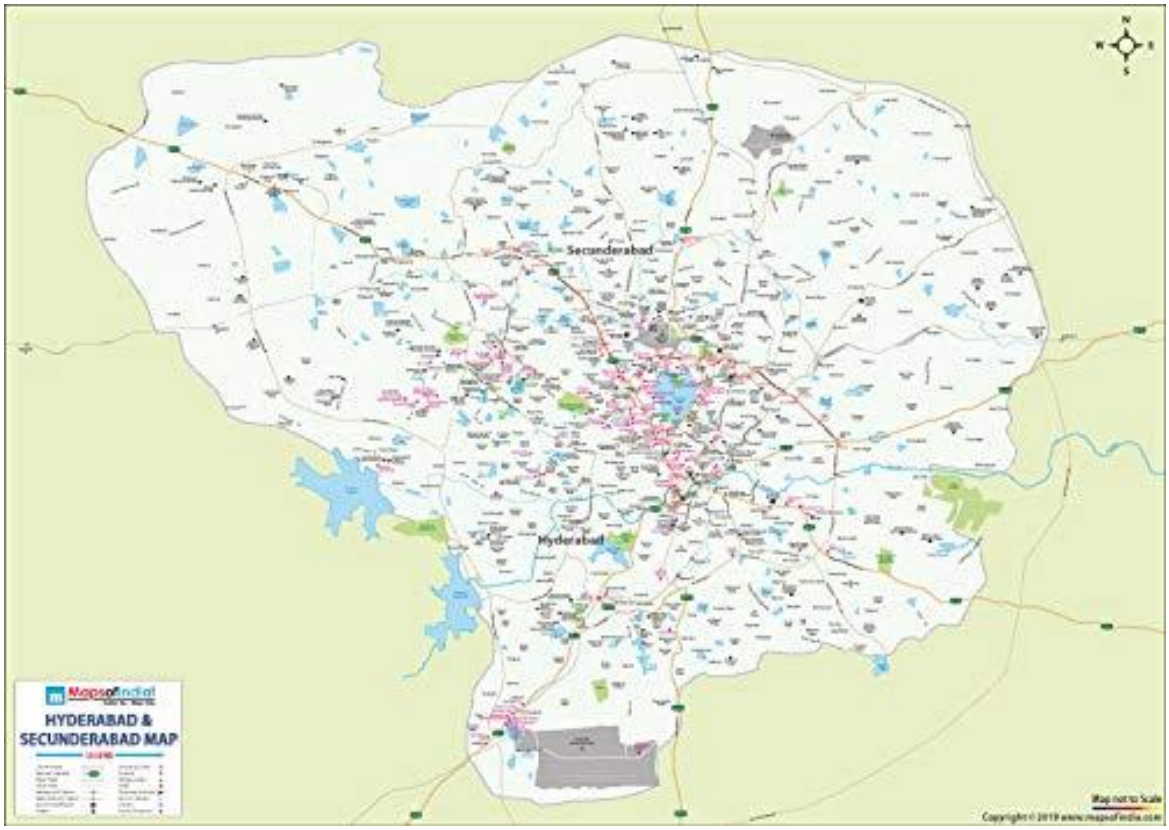
Coursera Capstone IBM Applied Data Science Capstone

Opening a New Multiplex in Hyderabad, India

By: Sharath Kumar V

Nov 2019



Introduction

Multiplex segment is growing while single screen segment is declining. As of March 2005, there were approximately 13000 cinemas in India of which 73 were multiplexes with a total of 276 screens. Multiplexes constitute only 0.6% of about 12000 cinemas halls, but account for28% to 34% of box office collection for top 50films in 2004(Yes bank, 2004). Growing film industry is the key driver for generating more footfalls for film exhibition industry growing faster (IBEF,2013). The modern shopping malls offer variety of entertainment services, life style products, gaming hubs, food courts and cinemas (Ibrahim and Ng, 2002; Friedberg, 1993). Shopping trips can have many purposes (O'Kelly,1981). If we consider the view point of Davies (1995) people enter in a theatre or cinema for leisure. People are turning towards multiplexes due to various reasons, some of them are safety, better ambience, eateries, security etc. (ooi and sim,2007). One will be surprised to see the number of women walking out of multiplexes after nightfall but saying that with the boom of multiplexes it is not a happy time for single screen theatres. When it comes to enjoying a movie with a woman companion most of the people prefer multiplexes (Eliashberg etal.,2005).

The purpose of this is to highlight the various aspects which lead to the preference of multiplexes over single screen theatres. The survey would be conducted across a minimum of ten malls across Hyderabad.

For many movie lovers, visiting multiplexes is a great way to relax and enjoy themselves during weekends and holidays. Multiplexes are like a one-stop destination for all types of movie watchers. To enhance the attraction of the shopper malls investors will start multiplexes also promote the many products as adds as part of the movie show. The location of the multiplex is one of the most important decisions that will determine whether the multiplex will be a success or a failure.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Hyderabad, India to open a new multiplex mall. Using data science methodology 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?

Target Audience of this project

This project is particularly useful to property developers and investors looking to open or invest in new multiplex in the one of the fastest growing capital of Telangana, India i.e. Hyderabad. This project is timely as the city is currently suffering from oversupply of Multiplexes due to most of the film cities around the city. Some research gone to understand where to start the multiplex and how to enhance the attraction in the city, but many are failed to create the attraction, this problem statement may help to the financial investors and developers to select the right place to start the multiplexes.

Data

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

• List of neighbourhoods in Hyderabad. This defines the scope of this project which is confined to the city of Hyderabad, the capital city of the Telangana and one of the fastest growing cities in India

• Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and to get the venue data.

• Venue data, particularly data related to Multiplexes. We will use this data to perform clustering on the neighbourhoods. Sources of data and methods to extract them This Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India) contains a list of neighbourhoods in Hyderabad, with a total of 200 neighbourhoods. 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 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 databases 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

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Hyderabad. Fortunately, (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhood's 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 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 Hyderabad. Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 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 neighbourhoods 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 neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood 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 analysing the "Multiplex" data, we will filter the "Multiplex" as venue category for the neighbourhoods. 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 neighbourhoods into 3 clusters based on their frequency of occurrence for "Multiplex". The results will allow us to identify which neighbourhoods have higher concentration of Multiplexes while which neighbourhoods have fewer number of

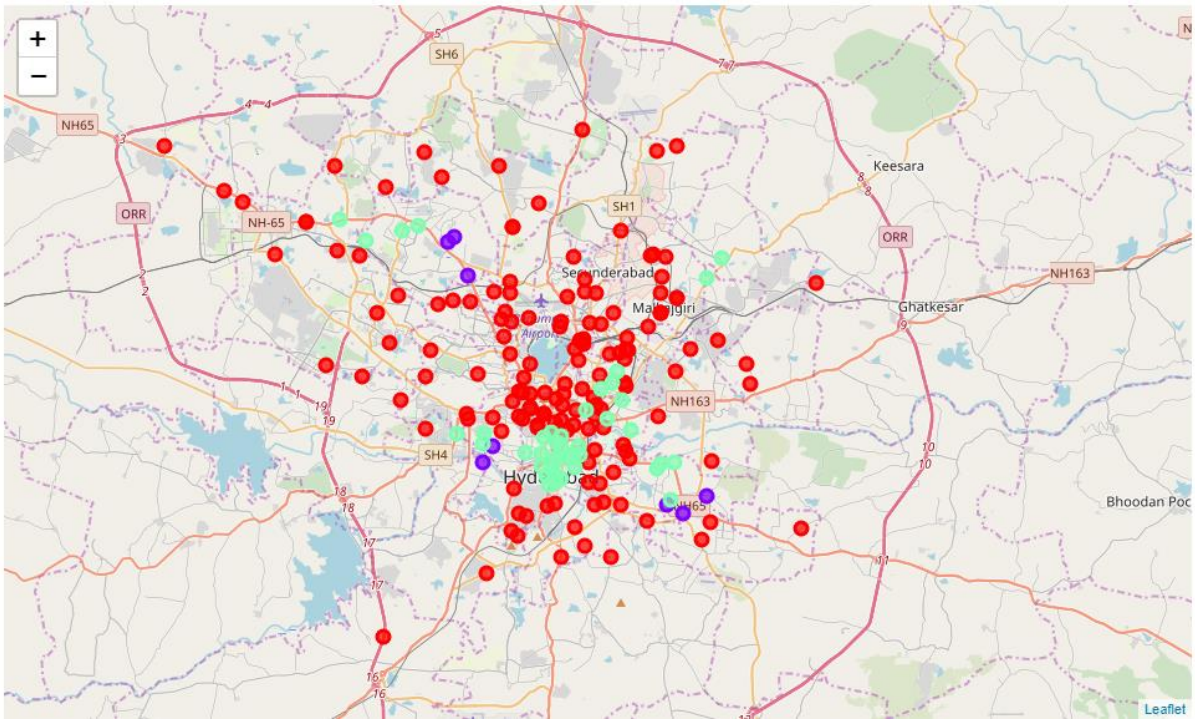
Multiplexes. Based on the occurrence of Multiplexes in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Multiplexes.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Multiplex”:

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

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



Discussion

As observations noted from the map in the Results section, most of the Multiplexes are concentrated in the central area of Hyderabad 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 Multiplex in the neighbourhoods. This represents a great opportunity and high potential areas to open new Multiplexes as there is very little to no competition from existing malls. Meanwhile, Multiplexes in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of Multiplexes. From another perspective, the results also show that the oversupply of Multiplexes mostly happened in the central area of the city, with the suburb area still have very few Multiplexes. Therefore, this project recommends property developers to capitalize on these findings to open new Multiplexes in neighbourhoods in cluster 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new Multiplexes in neighbourhoods in cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 2 which already have high concentration of Multiplexes and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of multiplexes, there are other factors such as population and income of residents that could influence the location decision of a new multiplex. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new multiplex. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results

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 neighbourhoods in cluster 1 are the most preferred locations to open 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.

References

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

Foursquare Developers Documentation. https://developer.foursquare.com/docs

Study of the factors influencing cinegores preference for multiplex compared to single screen cinemas in Pune, by Amit Mohan Sharma and Komal Chopra