Coursera Capstone

IBM Applied Data Science Capstone

Opening a Shopping Mall in Hyderabad, India

By

Syed Asif Ali

2020

Table of Contents

INTRODUCTION

For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. They can do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more activities. 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 provides a great distribution channel to market their products and services. Property developers are also taking 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

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. 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 shopping mall, 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 shopping malls in the city Hyderabad. This project is timely as the city is currently suffering from oversupply of shopping malls. The city has attracted private equity (PE) inflows of $200 million (over Rs 1,400 crore) into the retail segment in 2019 alone. Not only high-street stores are being set up by international retail giants here, these new malls coming up are the examples of city’s ongoing commercial growth and recent uptake of office space by multi-nationals and national firms. Hyderabad city is also expected to see good growth in residential and retail segments.

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, India in 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 shopping malls. We will use this data to perform clustering on the neighborhoods.

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 neighborhoods in Hyderabad. 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 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 Shopping 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

Firstly, we need to get the list of neighborhoods in the city of Hyderabad, India. Fortunately, the list is available in the Wikipedia

(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 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 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 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, India.

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 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 “Shopping Mall” data, we will filter the “Shopping Mall” as venue category for the neighborhoods.

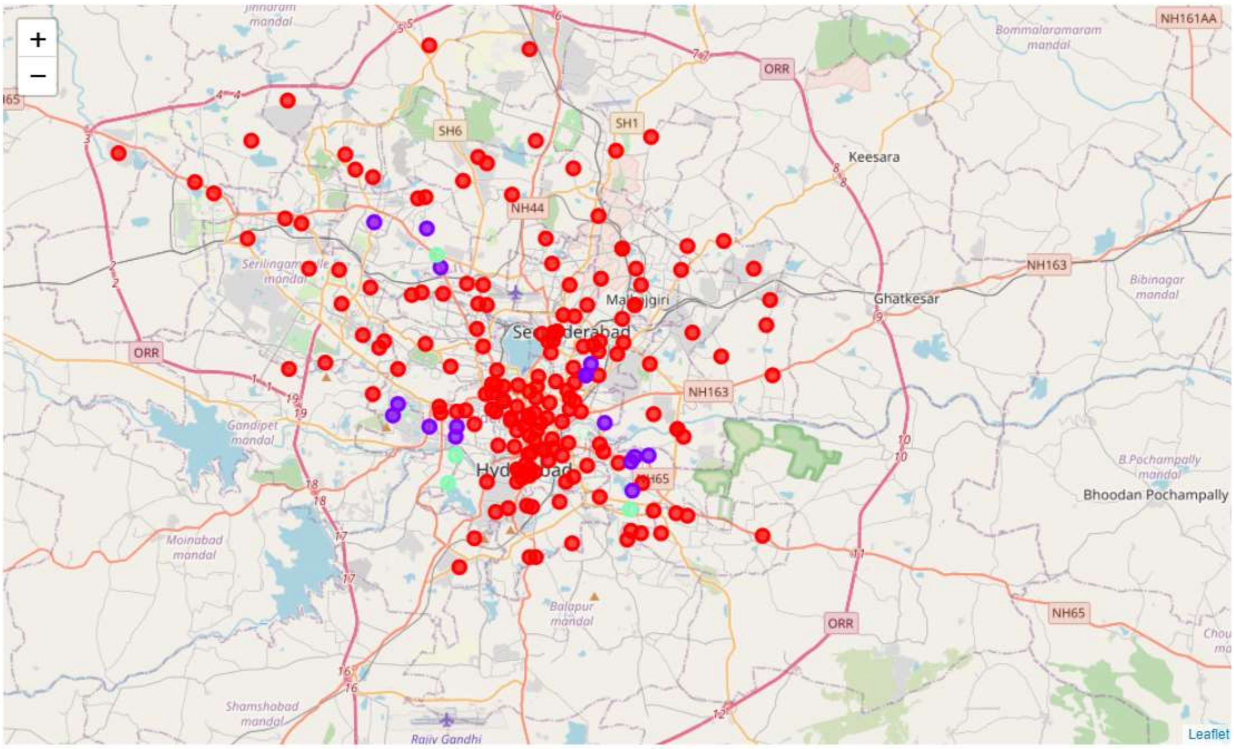
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine 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 “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. 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 very less number of shopping malls
- Cluster 1: Neighborhoods with a moderate 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 observations noted from the map in the Results section, most of the shopping malls are concentrated in the central area of Hyderabad, India city, with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 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 property 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 1 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.

Limitations and Suggestions for Future Research

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. However, to the best knowledge of this researcher such data are not available to the Neighborhood 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 shopping mall. 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.

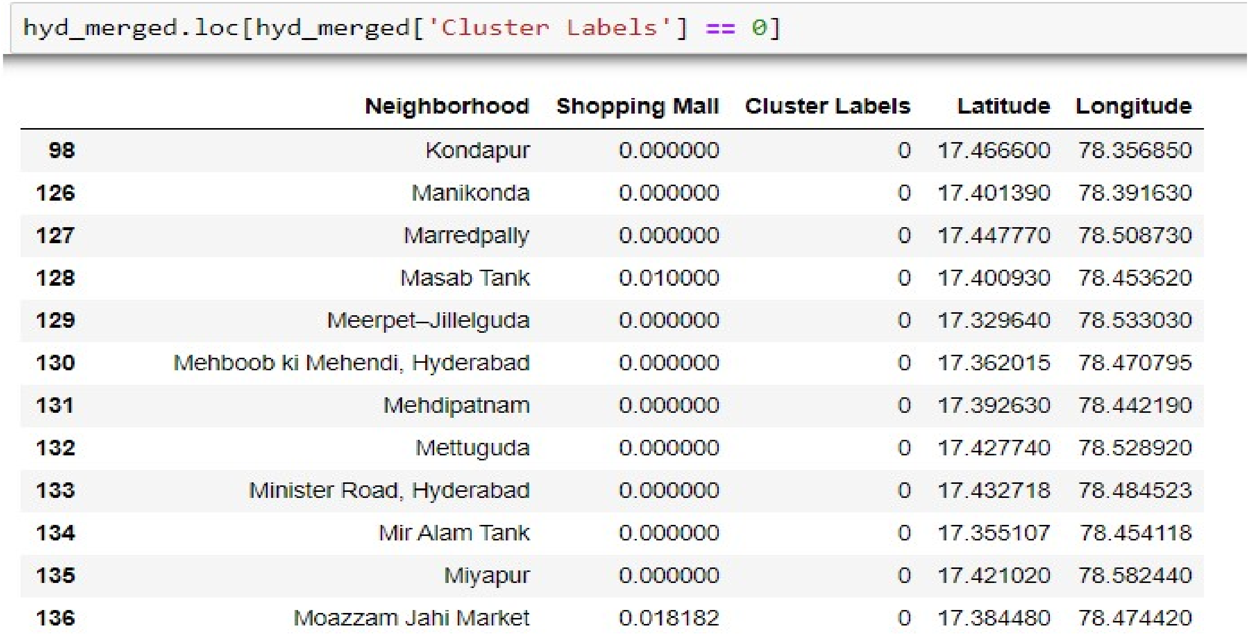
CONCLUSION

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 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 0 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.

REFERENCES

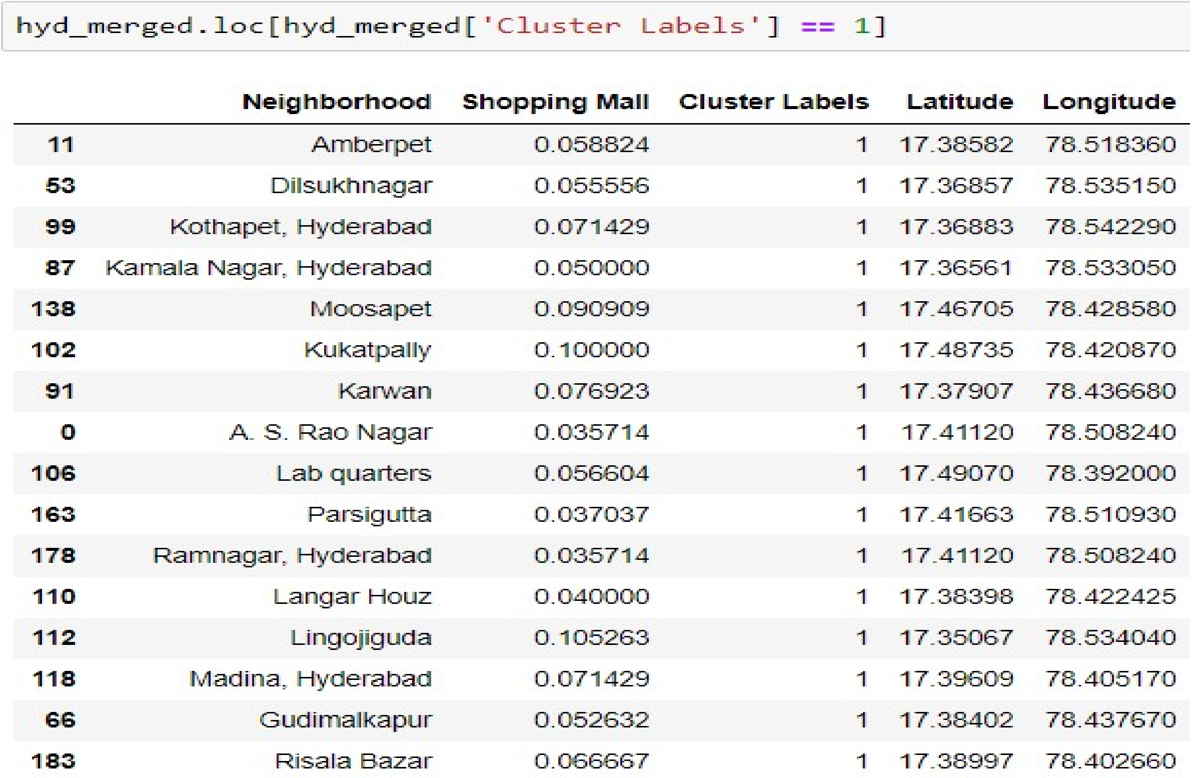
- Category: Suburbs in Hyderabad, India. *Wikipedia*. https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India
- Foursquare Developers Documentation. *Foursquare*. https://developer.foursquare.com/docs
- Geocoding service https://developers.google.com/maps/documentation/javascript/geocoding

APPENDIX



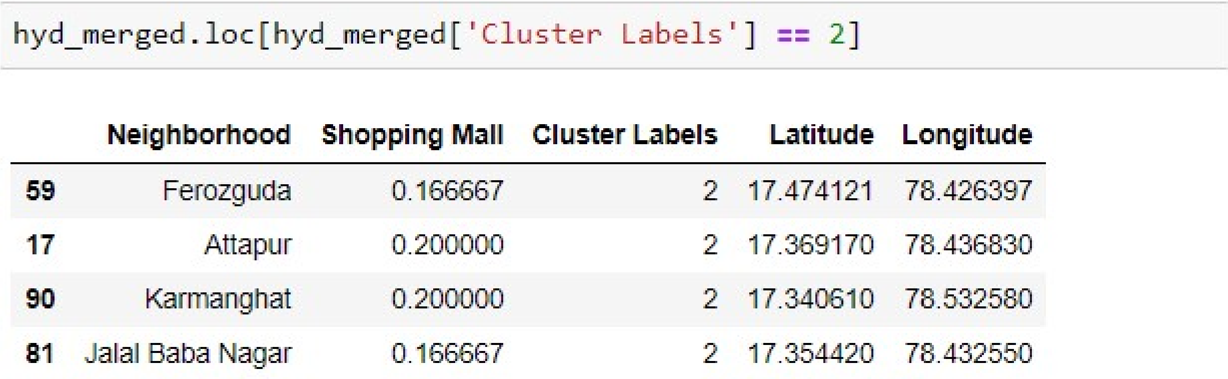
```
hyd_merged.loc[hyd_merged['Cluster Labels'] == 0]
```

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
98	Kondapur	0.000000	0	17.466600	78.356850
126	Manikonda	0.000000	0	17.401390	78.391630
127	Marredpally	0.000000	0	17.447770	78.508730
128	Masab Tank	0.010000	0	17.400930	78.453620
129	Meerpet–Jillelguda	0.000000	0	17.329640	78.533030
130	Mehboob ki Mehendi, Hyderabad	0.000000	0	17.362015	78.470795
131	Mehdipatnam	0.000000	0	17.392630	78.442190
132	Mettuguda	0.000000	0	17.427740	78.528920
133	Minister Road, Hyderabad	0.000000	0	17.432718	78.484523
134	Mir Alam Tank	0.000000	0	17.355107	78.454118
135	Miyapur	0.000000	0	17.421020	78.582440
136	Moazzam Jahi Market	0.018182	0	17.384480	78.474420



```
hyd_merged.loc[hyd_merged['Cluster Labels'] == 1]
```

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
11	Amberpet	0.058824	1	17.38582	78.518360
53	Dilsukhnagar	0.055556	1	17.36857	78.535150
99	Kothapet, Hyderabad	0.071429	1	17.36883	78.542290
87	Kamala Nagar, Hyderabad	0.050000	1	17.36561	78.533050
138	Moosapet	0.090909	1	17.46705	78.428580
102	Kukatpally	0.100000	1	17.48735	78.420870
91	Karwan	0.076923	1	17.37907	78.436680
0	A. S. Rao Nagar	0.035714	1	17.41120	78.508240
106	Lab quarters	0.056604	1	17.49070	78.392000
163	Parsigutta	0.037037	1	17.41663	78.510930
178	Ramnagar, Hyderabad	0.035714	1	17.41120	78.508240
110	Langar Houz	0.040000	1	17.38398	78.422425
112	Lingojiguda	0.105263	1	17.35067	78.534040
118	Madina, Hyderabad	0.071429	1	17.39609	78.405170
66	Gudimalkapur	0.052632	1	17.38402	78.437670
183	Risala Bazar	0.066667	1	17.38997	78.402660



```
hyd_merged.loc[hyd_merged['Cluster Labels'] == 2]
```

	Neighborhood	Shopping Mall	Cluster Labels	Latitude	Longitude
59	Ferozguda	0.166667	2	17.474121	78.426397
17	Attapur	0.200000	2	17.369170	78.436830
90	Karmanghat	0.200000	2	17.340610	78.532580
81	Jalal Baba Nagar	0.166667	2	17.354420	78.432550