Coursera Capstone Project: Applied Data Science

K Prudhvi

kprudhvi.52@gmail.com

Coursera Capstone Project: Applied Data Science

1 Introduction

Hyderabad, the city of Pearls is one of the largest city and capital of Telangana State, India. The dream city of Hyderabad is currently home to 1,22,17,956 people in Telangana. According to recent estimates, Hyderabad Metropolitan area or Hyderabad Urban Agglomeration is all set to cross 14.5 million (1.45 Crore) populations by the end of 2019. This figure was recorded at 77, 49,334 in 2011 census. Most of this population is youth only who came to Hyderabad in search of employment. Shopping and be in line with present style trend is what youth mainly focuses on these days. For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. 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. For this, 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.

2. 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: “What are the best and recommended locations in the city of Hyderabad to open a new shopping mall?”

3 Data

The data for this project has been retrieved and processed through multiple sources, giving careful considerations to the accuracy of the methods used.

The following are the major data required and the corresponding sources of them:

- **Neighbourhood Data**:

The data of the neighbourhoods in Hyderabad can be extracted out by web scraping using BeautifulSoup library for Python. The neighbourhood data is scraped from a Wikipedia webpage

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

- **Coordinates of those Neighbourhoods:**

The latitude and longitude of the neighbourhoods are retrieved using Geocoder Module. The geometric location values are then stored into the initial dataframe.

- **Venue Data for those neighbourhoods:**

From the location data obtained after Web Scraping and Geocoding, the venue data is found out by passing in the required parameters to the FourSquare API, and creating another Data Frame to contain all the venue details along with the respective neighbourhoods.

4 Methodology

Firstly, we need to get the list of neighbourhoods in the city of Hyderabad through Web Scrapping by the web URL mentioned in above section. We will do web scraping using Python requests and BeautifulSoup packages to extract the list of neighbourhoods data.

Coursera Capstone Project: Applied Data Science

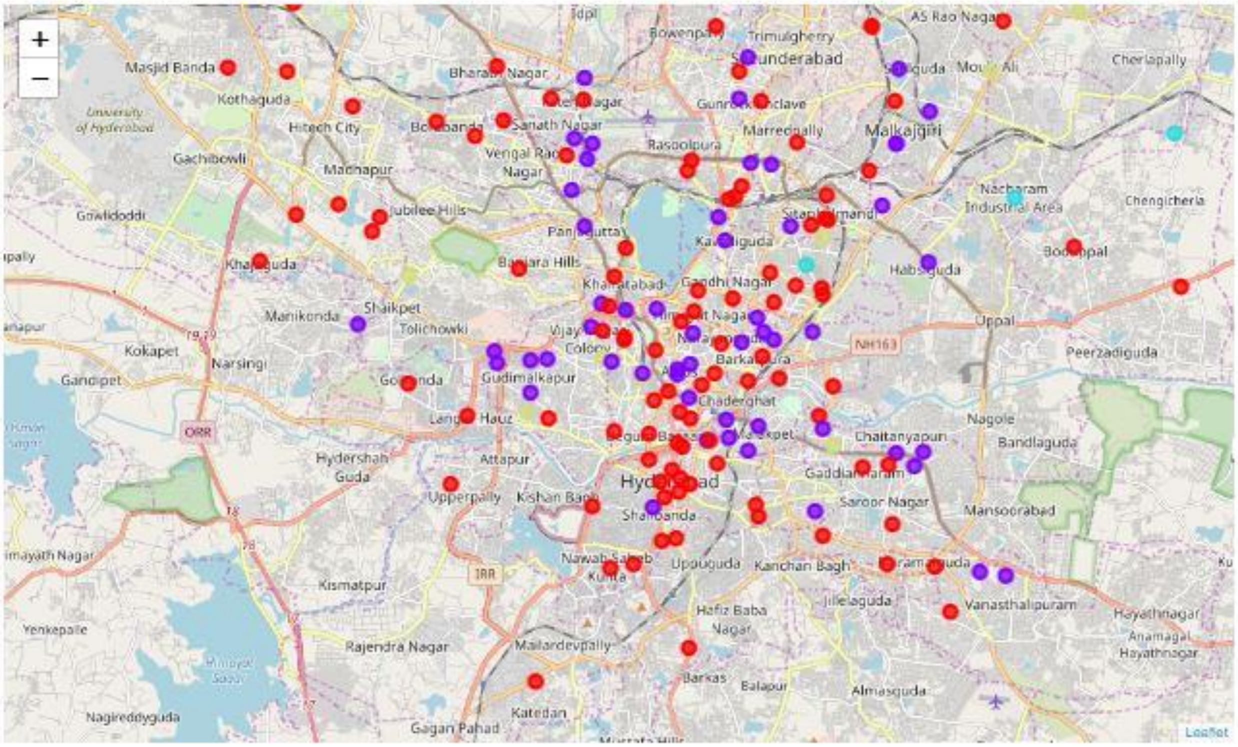
We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 500 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 analyze each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category.

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 identify the optimum number of clusters by checking the Elbow Point for distortions. The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new shopping malls.

5 Results

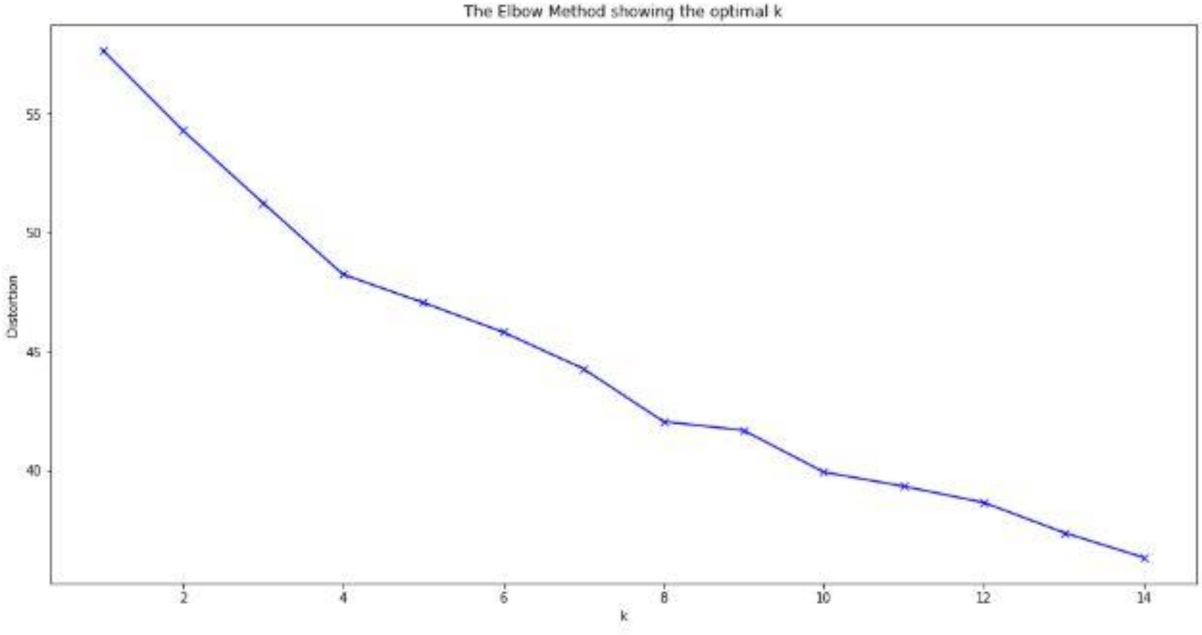
The neighbourhoods are divided into k clusters where k is the number of clusters found using the optimal approach. The clustered neighbourhoods are visualized using different colours so as to make them distinguishable. The clusters in each neighbourhood has similar characteristics when it comes to setting up of a new shopping mall.



- Cluster 0 and 1 (Red and Navy Blue) – These clusters are already having shopping malls in higher number
- Cluster 2(Sky Blue) – does not have shopping malls, but these are outskirts of city and are sparsely populated.
- Cluster 3(Light Yellow) – Does not have shopping malls, but small shopping initiatives like Women’s Stores are present.

Looking at all the clusters, it is very clear that new shopping mall should be set up in Cluster 3 neighbourhoods.

Also, we have chosen k=4 as no. of clusters based on the Elbow Point in the plot of Distortions vs ‘k’.



6 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 4 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. Please note that Population and Income of residents are two important factors which can be considered for future research purpose on this topic.