Applied Data Science Capstone

Final Project Assignment Report

Introduction

Gurgaon, officially named Gurugram, is a city located in the northern Indian state of Haryana. It is situated near the Delhi-Haryana border, about 30 kilometres southwest of the national capital New Delhi It is one of the major satellite cities of Delhi and is part of the National Capital Region of India. Gurgaon has a population of about 10 lacs.

Gurgaon has become a leading financial and industrial hub with the third-highest per capita income in India. Gurgaon has local offices of so many Fortune 500 companies.

For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during

weekends and holidays. Shopping malls are like a one-stop place for all. Real estate 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 already in the city. Opening shopping malls allows property developers to earn consistent rental income. The location of the shopping mall is one of the most key question in-front of real estate developers, which will determine if the mall will be a success or a failure.

So in this project we will try to build a solution which will help answering this question.

Data

To explore we will need the following data which is publicly available Data:

1. Gurgaon city Neighbourhoods: https://nominatim.openstreetmap.org/details.php?place_id=263703796
2. Foursquare Developers Access to venue data: https://foursquare.com/

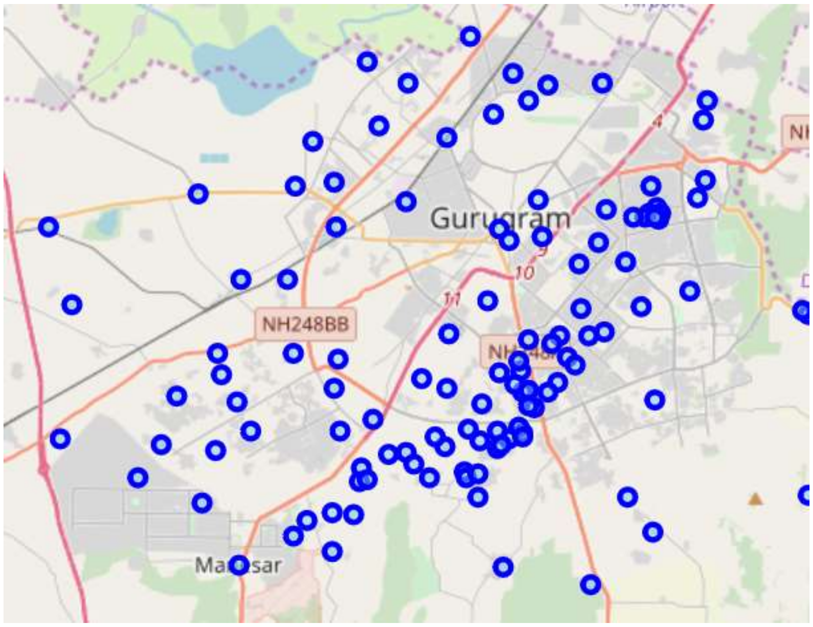
Using this data, we will do exploration first and try to find answer our question. The neighbourhood data will enable us to properly group shopping malls by neighbourhood.

Methodology

At the start, we need to get the list of neighbourhoods of Gurgaon city, though there are many public sites which has these data , we used the list from available in the **Nominatim.org** site (https://nominatim.openstreetmap.org/details.php?place_id=263703796).

We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. However, there are many more columns apart from list of names on this page, as well so many different info(in rows) which are not relevant, hence we have to performed data cleaning activities .
Now after getting the list of neighbourhoods, we need to get the geographical coordinates (latitude and longitude) for each neighbourhood in order to be able to use Foursquare API to gather venues data. 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.

This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted for Gurgaon city.



Next, we will use Foursquare API to get the top 30 venues that are within a radius of 500 meters for each neighbourhoods.

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 “Shopping Mall” data, we will filter the “Shopping Mall” 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 5 clusters based on their frequency of occurrence for “Shopping Mall”. The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while 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.

Results and Discussion

The results from the k-means clustering show that we can categorize the neighbourhoods into 5 clusters based on the frequency of occurrence for “Shopping Mall”:

• Cluster 0: Neighbourhoods low number of shopping malls

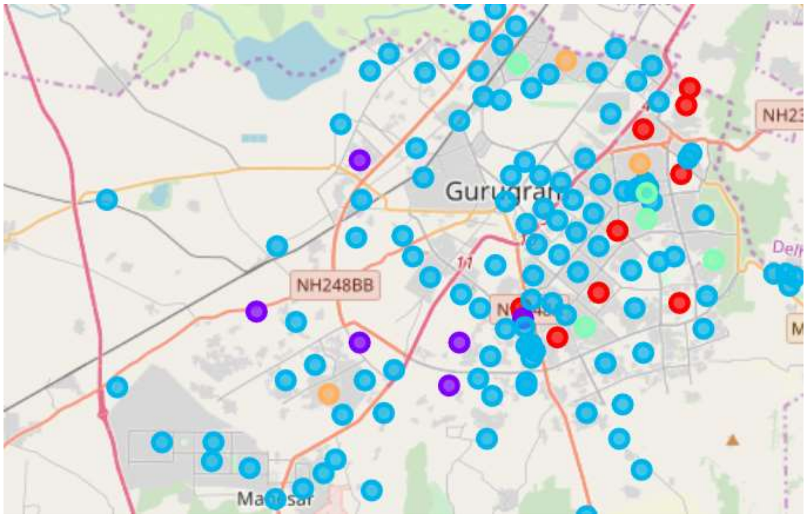
• Cluster 1: Neighbourhoods with high numbers of shopping malls

• Cluster 2: Neighbourhoods very low numbers or non-existence of shopping malls

• Cluster 3: Neighbourhoods with moderate concentration of shopping malls

- Cluster 4: Neighbourhoods with moderate concentration of shopping malls

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, cluster 2 in light blue colour, cluster 3 in light green colour and cluster 4 in light orange colour.



Most of the shopping malls are scatter around the central part of the city, with the highest number in cluster 1 and moderate number in cluster 3/4. On the other hand, cluster 2 has very rare instances of Malls in these areas. 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.

Therefore, this project recommends real estate developers to capitalize on these findings to open new shopping malls in neighbourhoods in cluster 2 with little to no competition.

Conclusion

Purpose of this project was to identify less/negligible numbers of Shopping Malls in Gurgaon city in order to guide real estate developers and city planner, with locations where a new Shopping Mall can be planned. So Cluster 2 areas plotted in map which are having most potential for laying a new shopping mall.