Coursera Capstone Project: Applied Data Science

Jaskirat Singh Nandhra

jaskiratnandhra@gmail.com

Predicting the best Place to setup a restaurant in Delhi

Introduction

Delhi is the capital city of India and is regarded as the heart of the nation. The city is popular for its enriched culture and heritage. The city hosts some famous historical monuments and is developing with the passing of time.

The influence of religious diversity can be seen in the city along with the cultural impact of the Mughal, the ancient Indian and the British. There are many beautiful gardens in the city, away from pollution and busy city life that provide opportunities to walk leisurely in the midst of greenery.

Delhi's urban area is now considered to extend beyond the NCT boundaries, and include the neighbouring satellite cities of Ghaziabad, Faridabad, Gurgaon and Noida in an area now called National Capital Region (NCR) and had an estimated 2016 population of over 26 million people, making it the world's second-largest urban area according to the United Nations. As of 2016, recent estimates of the metro economy of its urban area have ranked Delhi either the most or second-most productive metro area of India. Delhi is the second-wealthiest city in India after Mumbai and is home to 18 billionaires and 23,000 millionaires. Delhi ranks fifth among the Indian states and union territories in human development index.

Delhi has the second-highest GDP per capita in India. Furthermore, it is considered one of the world's most polluted cities by particulate matter concentration.

Problem Description

The objective of this Capstone Project is to analyse and select the best locations in the city of Delhi , India to open a new restaurant . Using data Science Methodology and machine learning techniques like clustering. This project aims to provide solutions to the business Question "What is the best and recommended locations in the city of Delhi to open a new restaurant?"

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 Delhi can be extracted out by web scraping using Beautiful Soup library for Python. The neighbourhood data is scraped from a Wikipedia webpage https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi and https://www.mapsofindia.com/pincode/india/delhi/
- 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.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Delhi 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.
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 restaurants in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new restaurants.

Target Audience

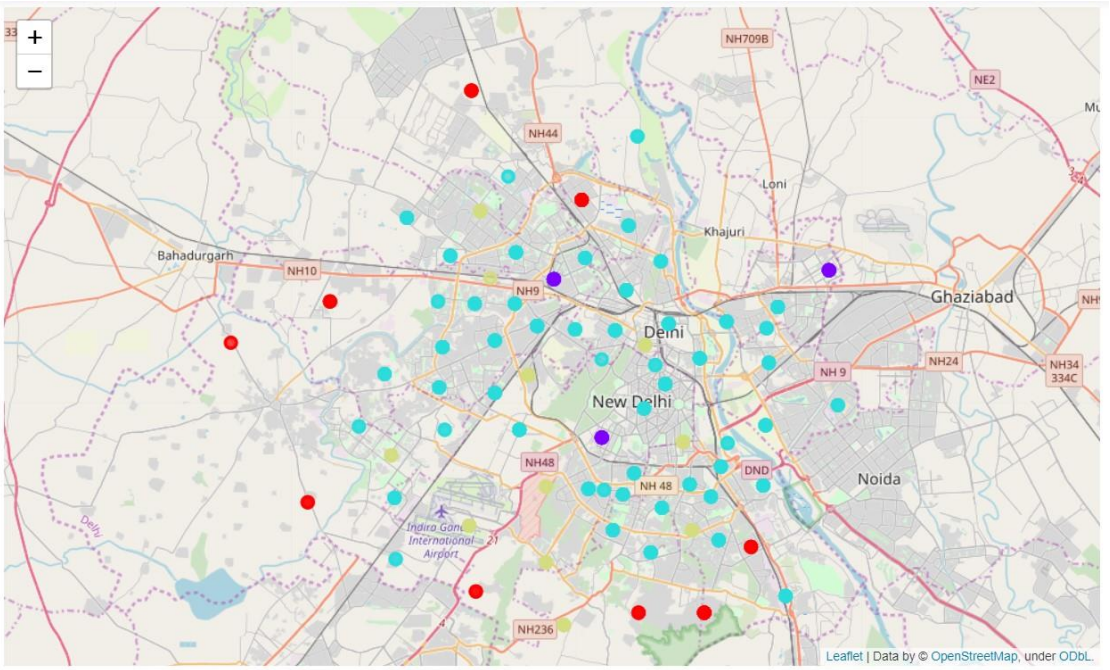
The objective of this Project is to locate and recommend to the management which neighbourhood of Delhi will be best choice to start a new restaurant in Delhi. The Management also expects to understand the rationale of the recommendations made.

This would interest anyone who wants to build a new restaurant in Delhi.

Results

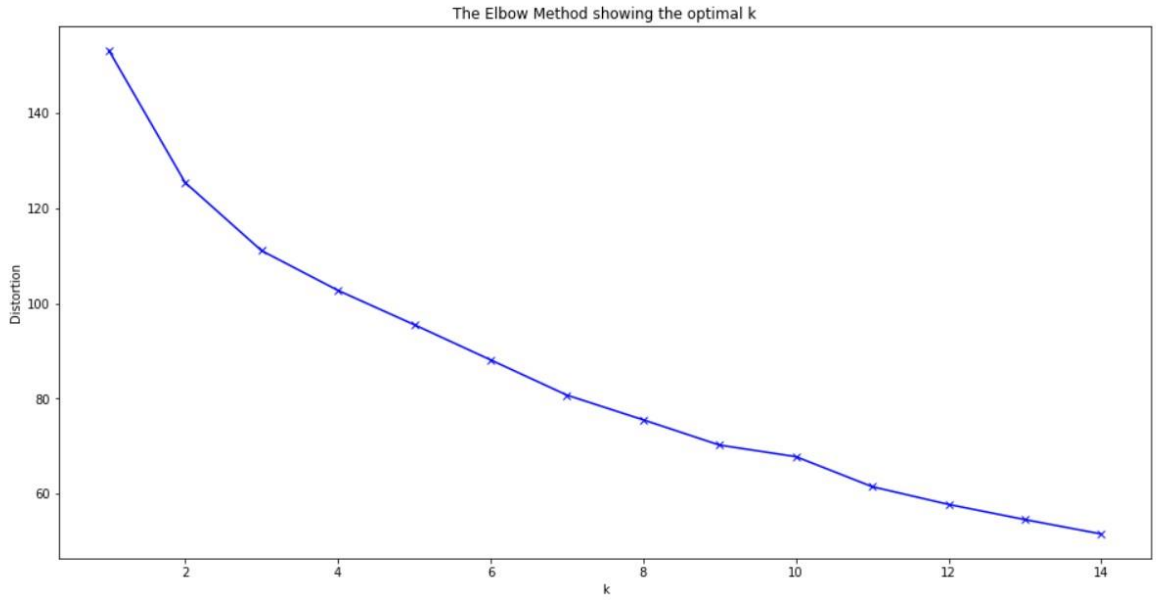
The neighbourhoods are divided into k clusters where k in sthe number if clusters found using optimal approach. The clustered neighbourhoods are visualized using different colour so as to make them distinguishable. The clusters in each neighbourhood has similar characteristics when it comes to setting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

Out[109]:



- Cluster 0 (Red) – These clusters are present on the outskirt of Delhi and have fewer number of restaurants since they are sparsely populated
- Cluster 1(Navy Blue) – These clusters are present besides the heart of the capital city of Delhi and have large number shops and fast food joints. But still the number of restaurants present in this cluster are less than all the other clusters.
- Cluster 2(Sky Blue) – These clusters are present with in the heart of the capital city of Delhi and have large number of restaurants since they are largely populated
- Cluster 3(Yellow)- These clusters are present through the city and the number of restaurants is good but are less than the Cluster number 2.

Also we have chosen k=4 as the number of cluster based on the Elbow Metho d in the plot of distortion vs ‘k’.



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 recommendation to the relevant stake holder i.e. property developers and investors regarding the best location to open a new restaurants . Please note that Population and Income of residents are two important factors which can be considered for future research purpose on this topic .