Coursera Capstone Project

IBM Applied Data Science Capstone

New Cineplex in London, United Kingdom

Author: Niranjan Ganesan

August 2019

1. Introduction

Whenever our favourite movie releases we always wish to watch it in its first day of release. Cinema is not just an entertainment any more, it has become an integral part of our lives where we spend time with our loved ones. From 5-year-old kid to 90-year-old person, cinema has been the No.1 entertainment factor and not to forget that it has also become an inspiration to many. The demand for Cineplex in more crowded areas has always been high. Property developers can take this as an advantage to build new Cineplex theatres to attract new customers but choosing the location of building the new Cineplex has always been challenging. The new Cineplex must be built at more crowded areas in the city where other Cineplex's doesn't exist. Determining this location is very challenging and contributes to success of the new Cineplex being built.

1.1 Business Problem

The objective of this project is to analyse and select the best locations in the city of London, United Kingdom to open a new Cineplex. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In a multi-cultural city like London, if a property developer is looking to open a new Cineplex, where would you recommend, they open it?

1.2 Target Audience of this project

This project is particularly useful for property developers and investors looking to open or invest in new Cineplex in the Capital city of England. General statistics shows that number of Cinemas sites in the UK has been increased steadily over a period of time. In 2017 UK has a total of 774 cineplex's in the country. The country's continued obsession with Cinemas will always welcome new Cineplex's in the city.

2. Data acquisition and cleaning

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

- List of Neighbourhoods in London. This defines the scope of this project which is confined to the city of London.
- Latitude and Longitude coordinates of those neighbourhoods. This is required to plot the map and get the venue data.
- Venue data, particularly data related to Cineplex. We will use this data to perform clustering on neighbourhoods.

2.2 Data Sources

This Wikipedia page(https://en.wikipedia.org/wiki/List_of_areas_of_London) contains a list of neighbourhoods in London. 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 using python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

Then, we will use Foursquare API to get the venue data for those neighbourhoods. The venue data includes the number of Movie theatres in each location. 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 Movie Theatre category in order to help us solve the business problem. This project will make use of many data science skills from web scraping, working with API, data cleaning, data wrangling, to machine learning and map visualization. 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.

2.3 Data Cleaning

The London neighbourhood data downloaded from Wikipedia are scraped into a single pandas dataframe. The area names were present as a link to its respective Wikipedia page. The text of these links alone were scraped using python BeautifulSoup package and was stored as a table into a dataframe.

After scraping there were lot of duplicate values which also included values such as “None”. In order to eliminate the duplicates and keep only one instance of each values, the drop_duplicates function while keeping only the first instance. The first instance of all the duplicate values and other unique values consists of the final dataset which has all the names of neighbourhoods in London.

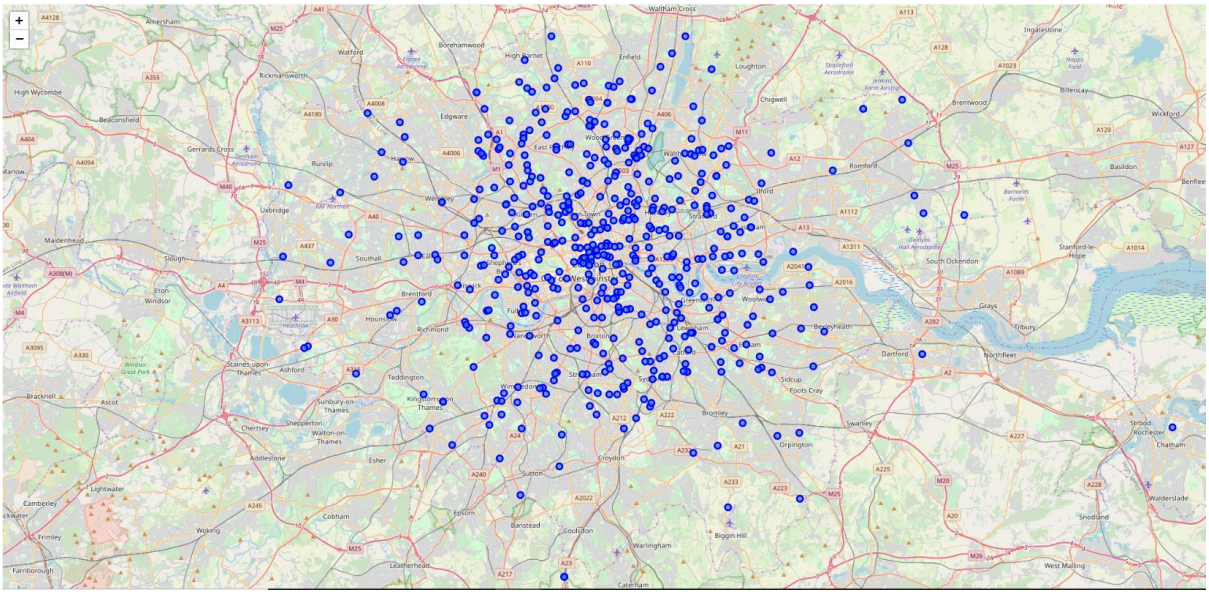
Now that the duplicates have been removed the index values must be reset in order to process the dataset sequentially. So, the index values were reset using reset_index python function. Finally, after web scraping and cleaning up of dataset we have the list of neighbourhoods in London stored in a dataframe for further processing.

Methodology

Firstly, we need to get the list of neighbourhoods in the City of London. This information is readily available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_areas_of_London). To extract the list of neighbourhood data we use web scraping using Python requests and Beautifulsoup packages to extract the list of neighbourhoods’ data. Next the geographical coordinates are required in order to use the Foursquare API method to analyse the neighborhoods. To get the geographical coordinates we will use the Geocoder package which will allow us to convert

address into geographical coordinates in the form of latitude and longitude. After getting the required data we will populate the data into pandas dataframe and then visualize the neighbourhoods in the map using Folium package. This allows us to ensure that the geographical coordinates data returned by the Geocoder are correctly plotted in the city of London.

After plotting the data on the map using folium it looks like this,



Now we will use Foursquare API method to get the top 100 venues that are within the radius of 2000 meters. A foursquare developer account is required in order to explore the city of London using its Geographical coordinates. We explore by making API calls to foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare returns the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With this 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 values. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking 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 “Movie Theatre” data, we will filter the “Movie Theatre” as venue category for the neighbourhoods.

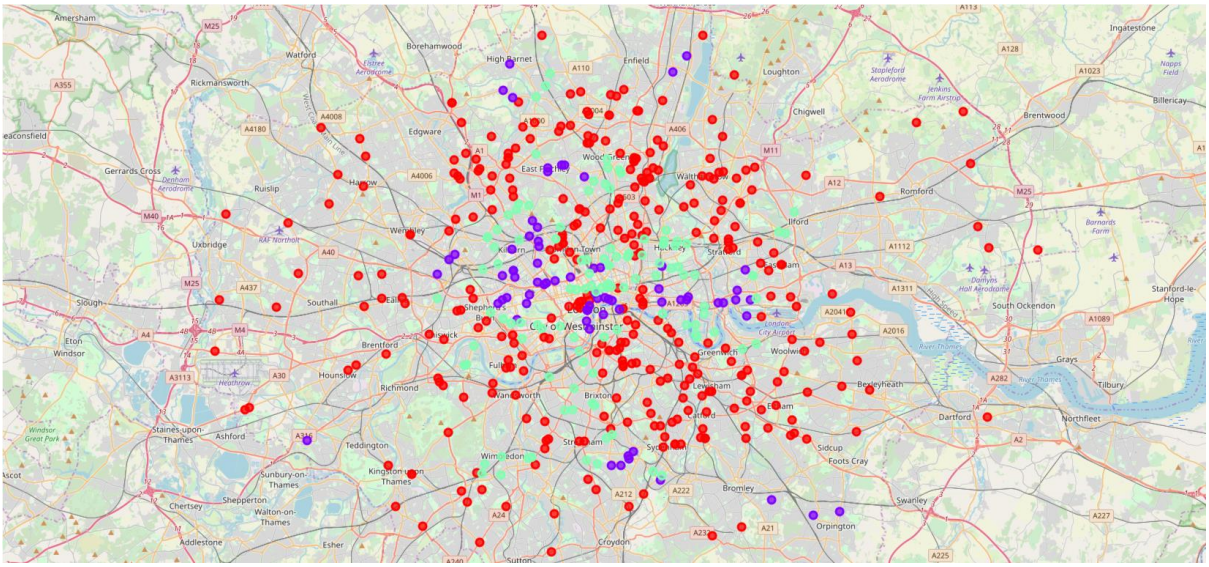
Finally, we will perform clustering on the data using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine 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 “Movie Theatre”. The results will allow us to identify which neighbourhoods have higher concentration of Movie Theatres while which neighbourhoods have fewer number of Movie Theatres. Based on the occurrence of Movie Theatres in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open a new Cineplex.

Results

The results from K-means clustering shows that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Movie Theatre”:

1. Cluster 0: Neighbourhoods with absolutely no number of Movie Theatres
2. Cluster 1: Neighbourhoods with high number of Movie Theatres
3. Cluster 2: Neighbourhoods with moderate number of Movie Theatres

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



Observations:

Our Analysis shows that although there are great number of Movie Theatres in London (~200 in our initial area of interest around London). Looking at the clusters the maximum number of Cineplex's are present in Cluster 1, very small number of Cineplex's are present in Cluster 2 and absolutely no Cineplex's are present in Cluster 0. Building a Movie Theatre in any one of the locations in Cluster 0 will be a right decision. Examining the clusters more Movie Theatres are present in Central London near Thames and northern part of Central London. So opening a Movie Theatre here will not be the right choice.

The south eastern part of London such as Lewisham, Catford, Eltham., etc does not have any Movie Theatres nearby. Opening a Cineplex in any one of these will definitely attract more customers.

Also closely analysing the areas near Walthamstow, West Hackney., etc there are few numbers of Movie Theatres. These areas could also be the right place for opening a new Cineplex as there are few Movie Theatres in the nearby areas.

So to conclude on a shorter note, areas near Central London such as Marylebone, Kensington, Westminster, Waterloo are not the right place to open a Cineplex as there are more number of movie theatres around. Highly recommended places to open a Cineplex are places in South

East London such as Lewisham, Catford, Eltham., etc. Also, areas near North London such as Walthamstow, West Hackney., etc are quite suitable places to open a new Movie Theatre.

Conclusion

Purpose of this project was to identify London areas with low number of Movie Theatres in order to aid stakeholders or Property developers in narrowing down the search for optimal location for a new Cineplex. By Calculating Movie Theatre density distribution from Foursquare data, we have first identified the areas that have movie theatres. Then these identified locations were clustered based on the density of the Movie theatres present. Clusters gave the visualization of the density of Cineplex which gave the insight to places which has very less or absolutely no movie theatres. These addresses are chosen for optimal locations to build a new Cineplex which will then be used as starting point for final exploration by Stakeholders or Property developers.

Final decision on optimal Cineplex location will be made by Property developers based on specific characteristics of neighbourhoods and locations in every recommended zone, taking into consideration additional factors like total population of each location (more number of people preferred), proximity to major roads, real estate availability, prices, social and economic dynamics of every neighbourhood etc.

References

Category: Areas in London, United Kingdom. Retrieved from

https://en.wikipedia.org/wiki/List_of_areas_of_London

Foursquare Developers Documentation. Foursquare retrieved from

https://developer.foursquare.com/docs