Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Cinema, London, United Kingdom

Bharath

Jun, 2020

Introduction

London is one of the main entertainment avenues in the United Kingdom and in western Europe. Millions of people from around the world visit London each year and they look for various entertainment outlets in London. Local population too is very vibrant and one of the main entertainment for this huge influx of people is visiting Cinemas. London too has a prolific movie industry and many Hollywood blockbusters use London for their movie premieres. Movie going is an experience in itself which many people experience week by week. Opening of a cinema hall is a huge investment from finding the ideal location and installing state of the art equipment. To make a profitable investment, finding an ideal location is crucial.

Business Requirement:

The objective of this capstone project is to analyse the councils and neighbourhoods within the Greater London area to open a new cinema hall in a council which has greater prospect of return on investment. Using data science methodology and machine learning techniques, the project will aim to answer the business requirement. Though there are various features that can be considered to solve the defined requirement such as looking into population in the council, population density in the council etc… the approach taken here is to identify existing councils with existing entertainment avenues and identify a neighborhood in the cluster.

Target Audience:

Media conglomerates, Cinema chain and for anyone interested to get into entertainment avenue business.

Data:

Following data will be used in the analysis:

- List of councils/boroughs in Greater London
- Latitude and Longitude coordinates of the neighbourhoods within the councils
- Venue data of existing cinema halls

Wikipedia page (https://en.wikipedia.org/wiki/List_of_areas_of_London) will be used as the main data source.

Following actions will be performed on the data source:

- Python Requests library will be used to extract the data from the Wiki page and BeautifulSoup module will be used to scrap the data from the response.
- The geographical coordinates are retrieved using Python Geocoder package
- Foursquare APIs will be used to retrieve the venues in the neighbourhoods.
- With the retrieved data, pre-processing will be done and using K-Means Clustering algorithm clusters are defined

Methodology:

First task is to get a list of boroughs and the neighborhoods in the borough in the greater London area. Borough and neighbourhood data is available in wikipedia and the foremost task is to scrape the data from the wiki page. Once the data is scraped from wiki, data is filtered to have only Borough and Neighbourhood to generate a Dataframe.

The wiki data doesn't have the neighbourhood coordinates which is needed to visualize the data. Using an appropriate Python library, the latitude and longitude details of the neighborhood are extracted and added to the Dataframe. Now the Dataframe can be passed to appropriate visualization techniques to generate maps.

Next step is to use Foursquare APIs to retrieve top 100 venue details from each neighbourhood. To achieve this, we should have a valid Foursquare account and appropriate keys to access their APIs. Once the right credentials are set up in the Foursquare account, we can iterate over the Dataframe to retrieve top 100 venues of each neighborhood. The retrieved venue details are to the Dataframe. Now we have the Boroughs, neighborhoods in the boroughs and the popular venues in the neighbourhoods.

Then we will analyze each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing for Cinema hall data, we will filter out the Dataframe for Cinemas.

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 6 clusters based on their frequency of Cinemas. The results will allow us to identify which neighbourhoods have higher concentration of Cinemas. Based on the occurrence in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open Cinemas.

Recommendation:

The results from the k-means clustering show that we can categorize the neighbourhoods into 6 major clusters. Going back to the business requirement, we are interested in neighbourhoods which already have densely packed cinema venues. Going by that requirement, neighbourhoods in cluster 1 and cluster 3 have more cinema venues and these neighborhoods are ideal for opening up an

cinema venue. In this project, we only consider one factor i.e. frequency of occurrence of Cinemas, there are other factors such as population and income of residents that could influence the location decision. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations. 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.