Coursera Capstone

IBM Applied Data Science Capstone

Opening Cafe in London

Introduction

People in urban area especially those neighborhoods with primary central business district need some place to hangout with their friends. They can also come to lunch or do serious thing such as work, study and read a book. Because of those reason to visit the Cafe, there are opportunities for person or maybe company to open Cafe business. As a result, there are Cafes in the city of London which is metropolitan city for financial, business, education, and government headquarters. Opening a Cafe requires deep knowledge of the location, business plan, and of course the coffee. In terms of location, opening a Cafe is a tricky process that will determine whether the Cafe able to attract customer around the area to visit or not able to compete with other restaurants.

Business Problem

The objective of this project is to analyze and choose the best location around London to open a new Cafe. Using data science methodology and machine learning techniques like clustering, this project perhaps able to provide solutions to answer the business question, "Where would you recommend the location around London to open a Cafe for someone or some company?"

Target Audience

This project is particularly useful to someone or some company looking to open a Cafe in London. With many Cafes have already opened in some neighborhood, this analysis hopefully will help them to decide the location based on the data

Data

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

1. List of neighborhoods in London. (https://en.wikipedia.org/wiki/List_of_areas_of_London)
2. Latitude and Longitude coordinates of those neighborhoods. This is required to plot those neighborhoods in map
3. Venue Data. Especially data related to Cafe that later we need this data to perform clustering on the neighborhood

Methodology

First, the list of neighborhoods in London are available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_areas_of_London). Web scraping technique using python language such as Request and BeautifulSoup library is used to extract the list of neighborhoods data. However, the list only contains text names without any information to analyze. In order to get information from Foursqure API, coordinates from those neighborhoods are needed that can be get from Arcgis Geocoder. After gathering the information, converting to table with Pandas dataframe is performed then visualizing the neighborhood in map is done with Folium libarary. This plotting is also useful to make sure the coordinates of data correctly plotted.

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 to obtain the Foursquare credential information. Then we do API calls to Foursquare passing in coordinates of the neighborhoods. Foursquare will return the surrounding venue data within the radius in JSON format and need to be extracted later to gather the only information that we need such as venue name, venue category, and venue latitude and longitude. With the data we can check how many venues are returned for each neighborhood and examine how many unique venue categories can be curated. Then we will analyze each neighborhood by grouping rows by neighborhood and taking the mean of the frequency of occurrences of each venue categories. By doing so, we are also preparing the data to use in clustering technique. Since we are analyzing the Cafe category, we should filter Cafe.

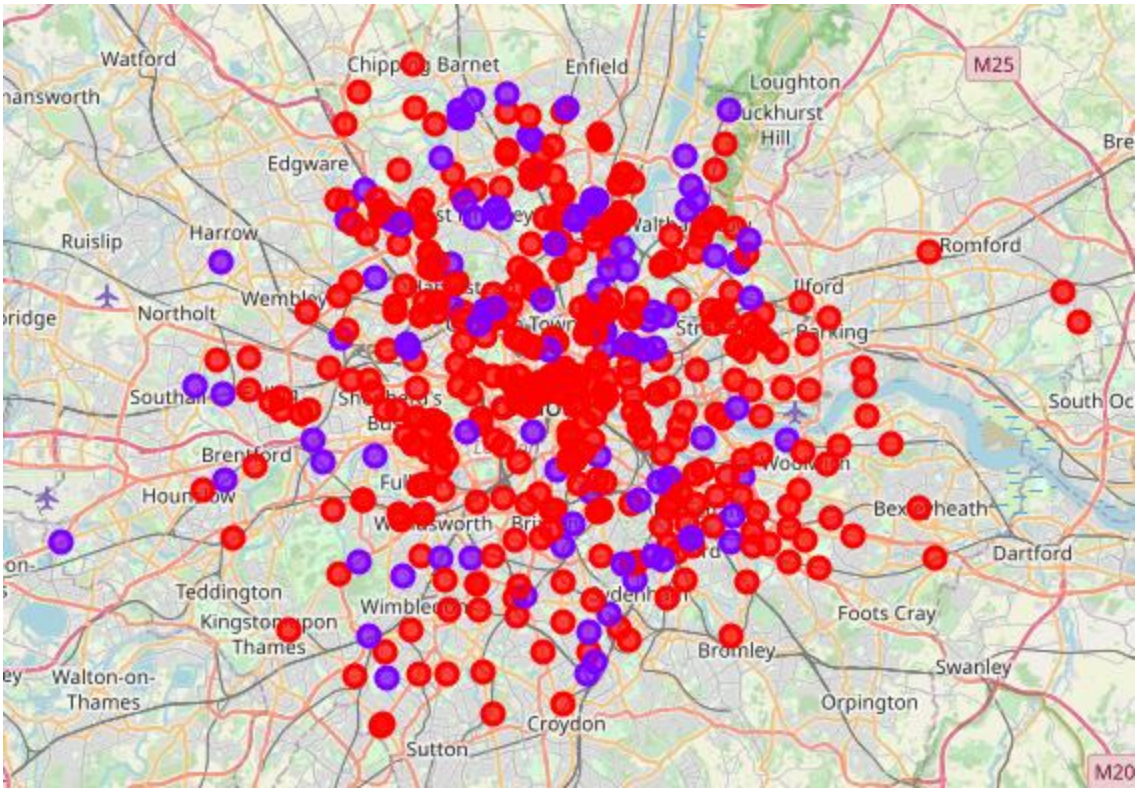
Finally, we can perform clustering on the data using K-Means clustering. K-Means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve our problem. We cluster the neighborhoods into 3 cluster based on the occurrence for Cafe. The results will allow us to identify which neighborhoods have higher concentration of Cafe while which neighborhood have fewer number of Cafe. Based on the occurrence of Cafe in different neighborhoods, it helps us to answer the question to which neighborhood are most preferred to open new Cafe.

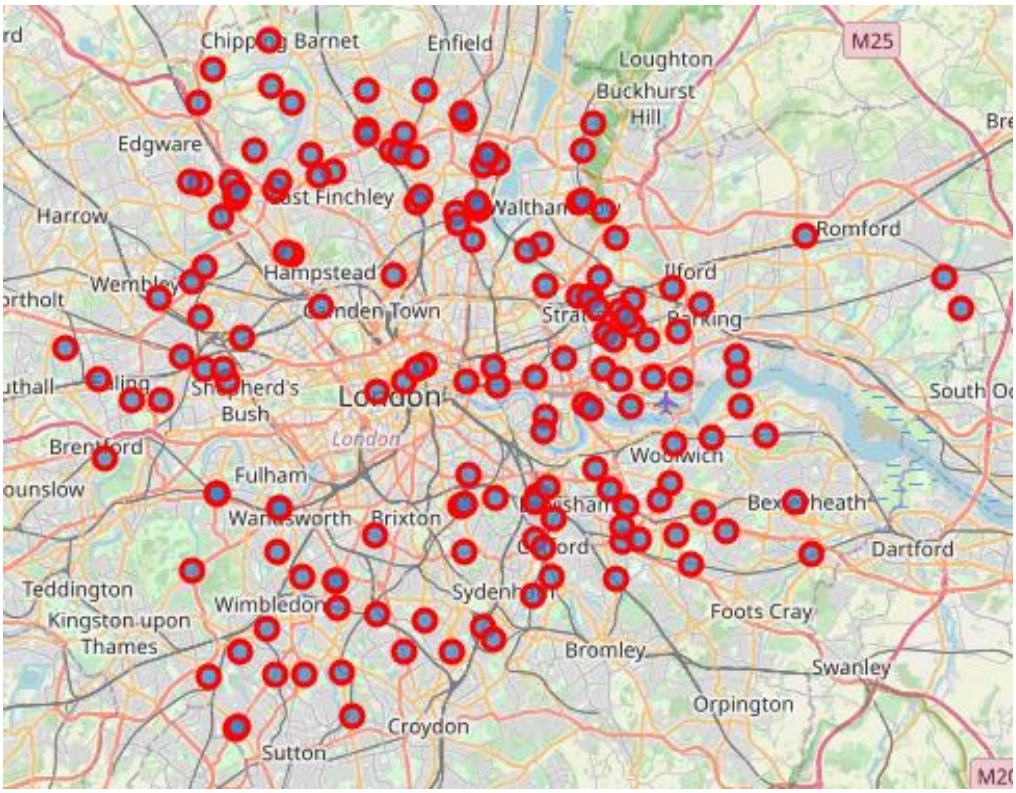
Result

The results from the KMeans clustering show that there are 2 cluster of neighborhoods based on the frequency of occurrence for Cafe:

1. Cluster 0 : Neighborhoods with moderate number of Cafes
2. Cluster 1 : Neighborhoods with high number of Cafes

The results of the clustering are visualized in the map below with Cluster 0 in red color and Cluster 1 in blue color. The next picture is neighborhood that is great to open a Cafe because there is still no Cafe in those area.





Limitation and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of Cafe, there are other factors such as type of district in the neighborhood, income of people in the surrounding area and other foods and beverages business that could influence the location decision of a new Cafe. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new Cafe. 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. Future research could make use of paid account to bypass these limitations and obtain more results.

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 algorithm by clustering the data into 2 cluster based on the similarities, and finally providing recommendation to the relevant stakeholders. To answer the business question that has been questioned in introduction, the answer proposed by this project is, the neighborhoods in cluster 0 are the most preferred locations to open a new Café. The findings of this project will help someone or company that want to capitalized the opportunities in food and beverages business to open a new Cafe on high potential locations without worrying about high competition due to many Cafés have been opened.