Coursera Capstone

IBM Applied Data Science Capstone

Opening a New African Restaurant in London

By: Robbin DIlles

June 2020

Introduction

Considering how diverse the city of London is, it can be difficult to find nice African dining experiences. This project will try to find the best spots within London for opening a new restaurant for African Cuisine. This will mainly target African (black) people and as such will focus on the demographics of London for finding a spot.

The restaurant itself could be anything ranging from Senegalese to Cameroonian, Nigerian, South African, Ghanaian etc, or a combination of them of course.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of London to open a new African Restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of London, if a Chef or branch manager is looking to open a new African Restaurant, where would you recommend that they open it?

Target Audience of this project

This project is particularly useful to chefs, cooks, branch managers, entrepreneurs and investors looking to open or invest in new restaurants in the capital city of London. This project is timely as the city is currently suffering from an undersupply of African Restaurants.

Data

The following data will be used to solve the problem:

- A list of neighbourhoods in London.
- Data of demographics in London.
- Latitude and Longitude coordinates of those neighbourhoods. Required to plot the map and get venue data.
- Venue data, particularly related to restaurants. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

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 of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods. After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database 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 Shopping Mall category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). 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.

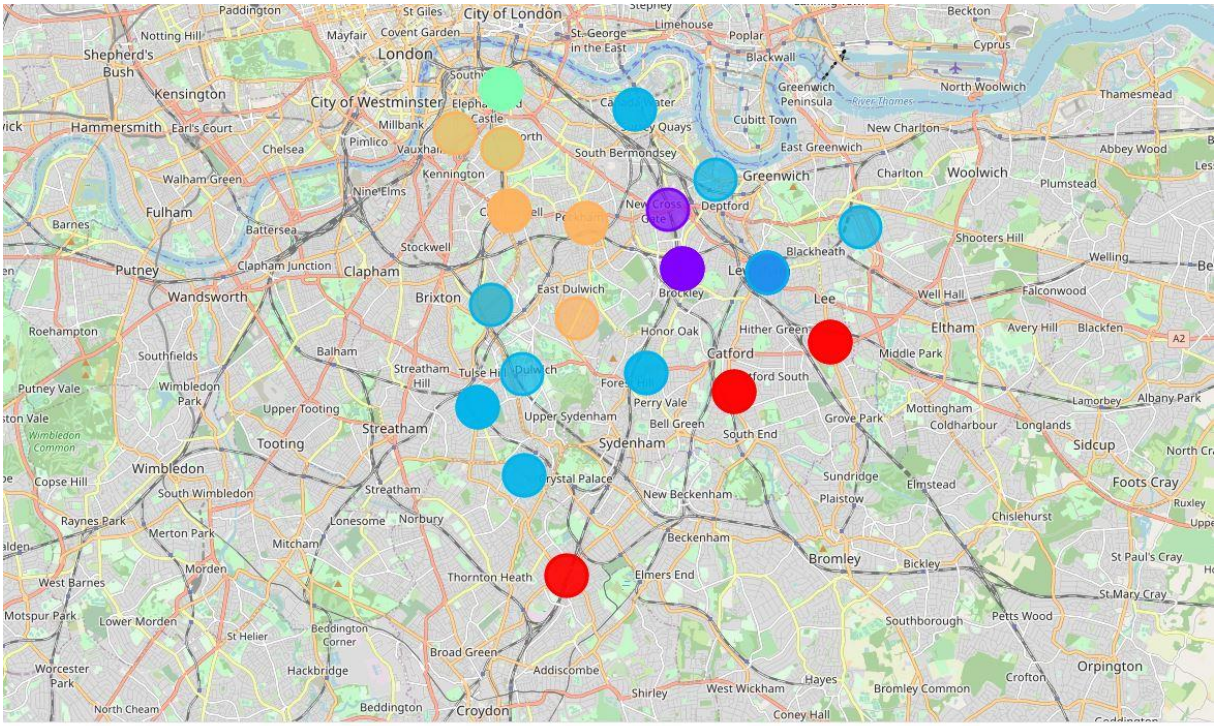
Methodology

Firstly, we need to get the list of neighbourhoods in the city of London. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_areas_of_London). We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. However, this is just a list of names. 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 wonderful 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 in the city of London. Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 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 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 “Restaurant” data, we will filter the “Restaurantl” 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.

Results

Some important highlights of the 5 clusters:

1. Drinking establishments like Pubs, Cafe's and Coffee Shops are popular in the South East London Area.
2. For restaurants it looks like the Italian Restaurants are the most popular. Especially in Southwark and Lambeth.
3. Considering the Lewisham area is the most condensed area of Africans in the South East Area, it is surprising to see how you can barely see restaurants in the top 5 venues.
4. In all clusters, it is easy to see a predominance of pubs.



Discussion and Conclusion

We find 2 clusters that look like the most viable clusters to establish an African Restaurant. Their proximity to other amenities and to the station are of high importance. These 2 clusters do not have top restaurants that could rival a new restaurant if it were established. The proximity to much needed resources is also important as Lewisham and Lambeth are not far out from Peckham.

Ultimately, this project would have had better results if it were possible to analyse all neighbourhoods, and get more data in terms of crime within the area, traffic access, store and warehouse proximity and ability to explore more venues with the Foursquare api.

Of course, getting ratings and feedbacks of the current establishments within the clusters would provide more insight as well.