Battle of the Neighborhoods

Finding the best neighborhood for opening an Arts & Crafts store in London using Data Science

This project aims to use Data Science concepts learned in the IBM Data Science Professional Course to solve a business problem.
In particular, we will use Foursquare location data and clustering of venue information to determine what might be the 'best' borough in London to open an Arts & Crafts store.
In this project we will proceed in a step by step manner, from problem definition to a conclusion that can be leveraged by the business stakeholders to make their decisions.

Table of Contents

1. Introduction

The city of London is famous for many things – its magnificent ancient buildings, its nightlife, its multiculturalism, its role in international trade,

its growing foodie scene, its being the setting of many films and novels, and so on.

As one of the world's creative capitals it is not surprising that London:

- is home to more than 1.000 art galleries, many of which are rated amongst the best in the world;
- has also been the home of and inspiration for countless artists and creative movements;
- plays hosts to multiple world-renowned art schools, with students that travel across the globe to join their varied and inspiring courses.

The objective of this project is to determine which might be the 'best' borough in London for an entrepreneur to open an Arts & Crafts store for selling artist materials and where to organize creative workshops.

2. Data Overview

The data we will use to conduct our analysis comes from multiple sources:

- a list of boroughs in London (via Wikipedia. See link https://en.wikipedia.org/wiki/List_of_London_boroughs). The Wikipedia page provides different information about each borough, including its name, its geographical location, its area, its population (2013 estimation) and so on. Since the data is not in a format that is suitable for direct analysis, scraping of the data was done from this site;
- a population projection for year 2020 for each London borough (via London Datastore, a free and open data-sharing portal where

anyone can access data relating to the capital. See link https://data.london.gov.uk/dataset/london-borough-profiles). The file was in Excel format, so we could load it directly;

- GIS boundaries for each borough in London (via London Datastore. See link https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london). The file was in ESRI format, so we could load it directly;
- venue data pertaining to Arts & Crafts store, but also Art Schools and Art Museums (via Foursquare). The venue data will help us find which borough is best suitable to open our store.

3. Methodology

After all the data was collected and put into data frames, cleansing and merging of the data was required to start the process of analysis.
When getting the data from Wikipedia, the population of each borough was an estimation relative to year 2013, so we used the data gathered from London Datastore to have a population projection for the year 2020.

Next, we used the Foursquare API to get a list of all the venues in London belonging to specific categories. In addition to Arts & Crafts Store we included: Art Gallery, Art Museum, College Arts Building and Art Studio. In fact, in deciding where to open a new Arts & Crafts store, we think it would convenient to consider the presence, in that particular borough, of other activities related to Art that can contribute to the success of our store.

We used the boundaries of each borough to assign each venue returned by Foursquare API to the borough it's located into. In fact, if you query

Foursquare to search for venues nearby the coordinates of each borough you can't tell if all the venues returned are located into that specific borough; it can also happen that some venues are outside all London borough. If a venue is not assigned to any borough then it is ignored.

Then, to analyze the venue's data returned by Foursquare we performed a technique in which categorical data is transformed into numerical data for Machine Learning algorithms: this technique is called One hot encoding.
After this, we grouped rows of our data frame by borough and, by taking the mean of the frequency of occurrence of each venue's category, we created a new data frame listing the top 10 venues for each borough, making the data much simpler to analyze.
We can see that only 28 boroughs out of 33 have at least one venue falling in one of the categories we are interested into.

To make the analysis more interesting, we decided to cluster the boroughs based on the similarities of the frequency of occurrence of each venue's category in that borough. To do this we used *K-Means* clustering. To get our optimum K value we ran a test with different values of K and measured the accuracy of our model. In our case, we had the optimum at $K = 6$. That means we will have a total of 6 clusters.
After, we merged this data with the borough data frame to have a global vision on the distribution of venues of our interest among the various London boroughs. This new data frame represents the basis for analyzing new opportunities for opening a new Arts & Crafts store in London.

Then we created a map using the Folium package in Python and each borough was colored based on the cluster label. This map shows the different clusters that had a similar mean frequency of the venue's categories we were interested into.

4. Results

From cluster analysis we get that only the following results:

Cluster 1

Cluster 1 has three boroughs: Croydon, Brent and Enfield.
It we look at the most common venues one could suppose that this cluster has the highest average of Art Studios and Universities, but if you look at raw data you can find that there are only to Art Studios among all the venues we are interested to.
In the map, we can see that boroughs of this cluster are dispersed all throughout London making it one of the most sparsely populated cluster.

Cluster 2

Cluster 2 has five boroughs: Lambeth, Greenwich, Ealing, Havering and Waltham Forest.
The most common venues are Art Galleries and Arts & Crafts stores: we have 11 Art Galleries and 8 Arts & Crafts stores.

Cluster 3

Cluster 3 has thirteen boroughs.
The most common venues are Art Galleries and College Arts Buildings.
In this cluster we have a total of 11 Arts & Crafts stores.
Boroughs of cluster 3 are mainly located in Inner London.

Cluster 4

Cluster 4 has only one borough: Redbridge.
In this cluster, among all the venues we are interested into, there is only a Discount store.

Cluster 5

Cluster 5 has two boroughs: Lewisham and Haringey.
In this cluster there are two College Arts Buildings, one for each borough. No other venue of our interest.

Cluster 6

Cluster 6 has four boroughs: Hounslow, Bromley, Islington and Hammersmith and Fulham.
In this cluster the most common venue are Arts & Crafts stores, with a total of 9 such stores, of which 5 in the Islington borough.

5. Discussion

All of the Arts & Crafts stores are located in cluster 2, 3 and 6, with a total of 28 stores, divided in the following way:

- 8 stores in cluster 2;
- 11 stores in cluster 3;
- 9 stores in cluster 6.

We have a total of 40 College Arts Buildings, of which 30 are located in cluster 3 and 6 in cluster 2.
We can note that in cluster 3, and in particular in the Westminster borough, we have the most part of Art Galleries and Art Museums.

In the Westminster borough there are also 7 College Arts Buildings, but here we can find only 2 Arts & Crafts stores. This gives us a good opportunity for opening in this borough a new Arts & Crafts store.
One of the main drawbacks of this analysis is that it is completely based on venue's data obtained from the Foursquare API, and I can tell for sure that many venues belonging to the category we were interested into are missing or misclassified. For example, I found that one of the biggest Arts & Crafts store in London, which is also renowned for its online store, is classified as a miscellaneous store, so it has not be included in our analysis.

6. Conclusion

In conclusion, we had an opportunity to face a business problem and to tackle it using data science methods.
We used different Python libraries to fetch the information, control the content and visualize those datasets.
We learned how to gather data needed for our analysis from different sources like:

- Foursquare API, to investigate the venues in boroughs of London;
- Wikipedia, to get a list of all the London boroughs;
- London Datastore, to get some relevant information about London boroughs (GIS boundaries, population, etc.).

This project can be greatly improved with a better data source, rather than Foursquare, from which to retrieve venue's information filtered by a better classification into categories.

We can indeed use other information to guide our choice to where to open a new Arts & Crafts store, for example the proximity to public transport.

Anyway, this project is a starting point to investigate similar situations, for example opening a new restaurant, a new bookshop and so forth. This project acts as an initial direction to tackle more complex real-life problems using data science.