2020

Opening a new hotel in London

Coursera Capstone

IBM Applied Data Science Capstone

By: Evangelos Dragoumanos

June 2020

Introduction

Tourism is one of the biggest industries throughout the world and as long as the flights from country to country are increased, more and more people visit other countries either for leisure or business.

The capital of United Kingdom is both an economical and a touristic center and as a result London attracts around 30 million visitors from around the world every year. Consequently, the demand of accommodation is high and a large number of venues offer this convenience, from hotels to AirBnb apartments and hostels to student rooms.

Business problem

A client company which would like to expand its business to the accommodation industry has asked us to evaluate the possibility of opening a new hotel in London. Using data science methodology and analysis as well as machine learning techniques, this project aims to provide suggestions for the optimal location which a new hotel should be start welcoming visitors.

Data

Due to the fact that the greater London covers an area of approximately 1,570km, for this problem, it is necessary at first step to locate the districts/neighborhoods. The following link to the corresponding Wikipedia page has been used to get the data according to the Postal Codes of districts: https://en.wikipedia.org/wiki/List_of_areas_of_London

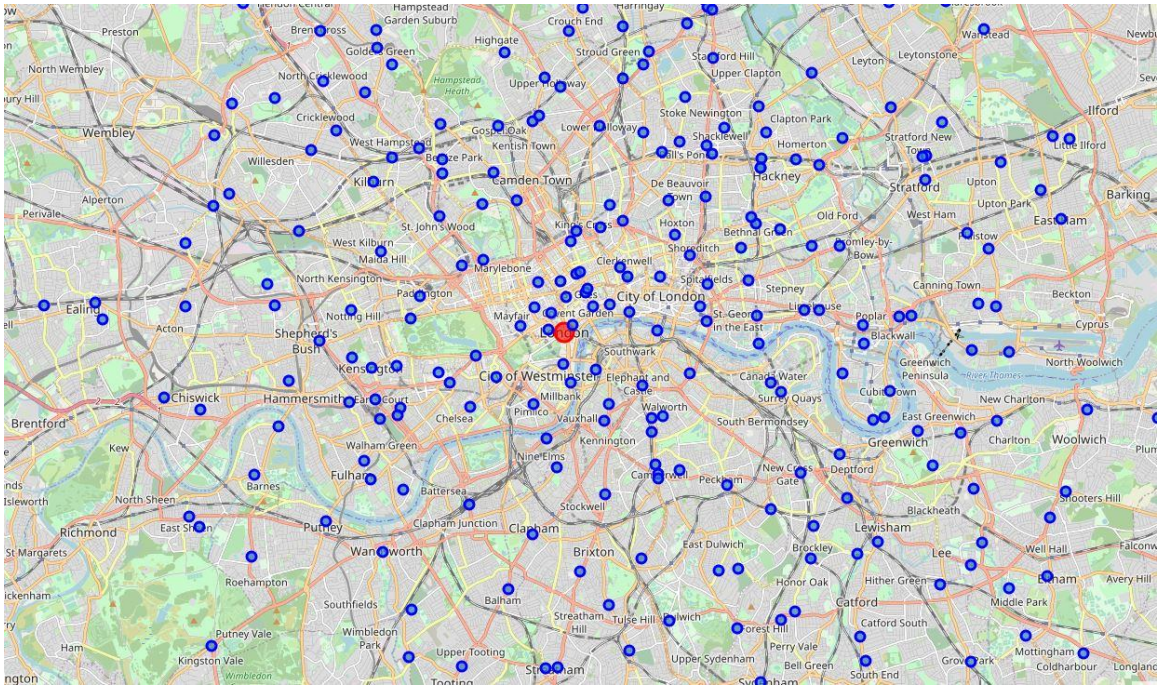
The above data will be scrapped in proper form and then it will be converted to a data frame in order to be suitable for processing in Jupyter notebook. Afterwards, we will reduce the data frame to London only, due to the fact that the areas of great distance from the city of London are not suggested to opening a new hotel.

Afterwards, based on the name, we will retrieve the geographical coordinates of each neighborhood using geocoder and a new data frame will be created with this information.

In order to explore the neighborhoods, Foursquare API will be used to get the latitude, longitude and location of all of the hotels in London area.

Methodology

A map of the London neighborhoods using the data frame after scrapping the data and selecting London based only locations was created as indicated below. The centers of each neighborhood is indicated in small blue circles while the bold red dot corresponds to the coordinates of the city center.



1 Figure - Map of Neighborhoods

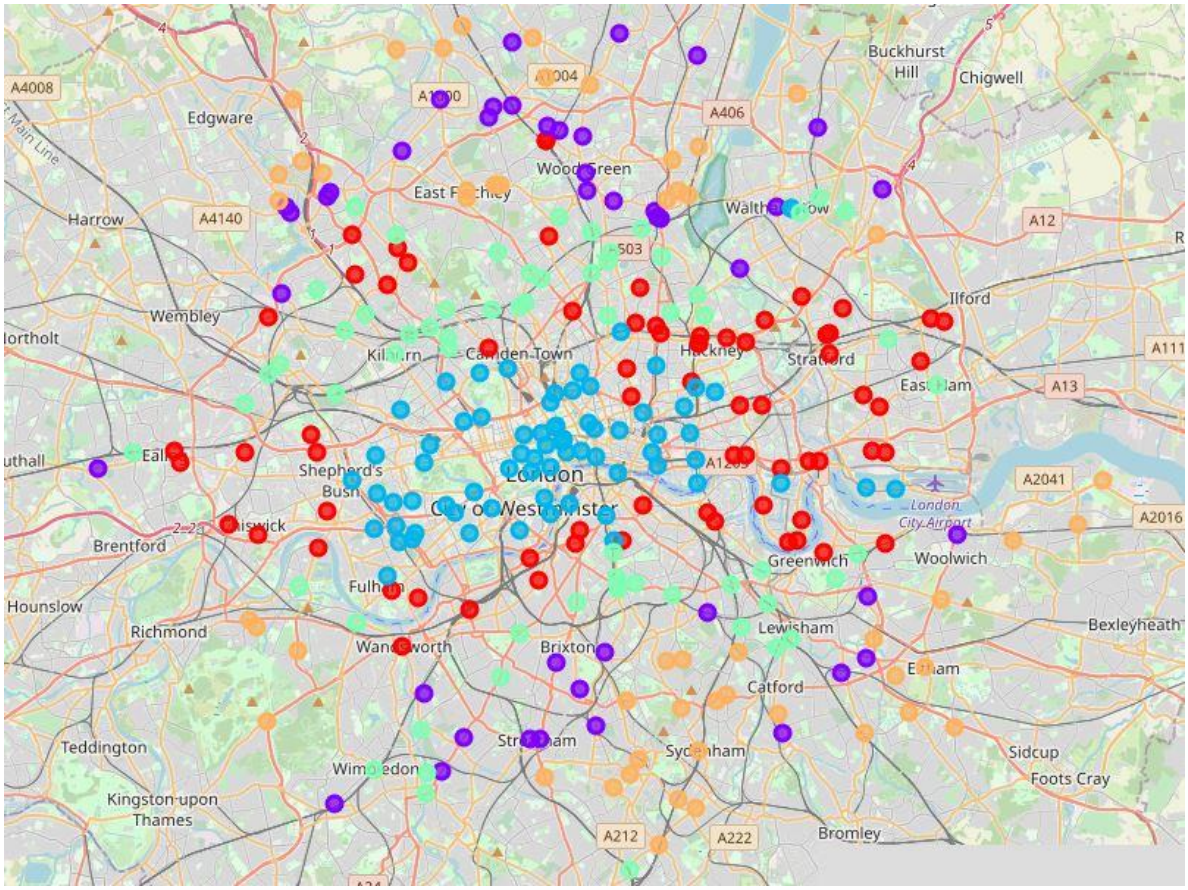
After searching in Foursquare API from the hotels in each district, we noticed that the category 'Hotel' contains many subcategories such as 'Bed & Breakfast', 'Hostel', 'Motel', 'Vacation Rental' et cetera.

As a consequence, the filtering of the data to 'Hotel' only is of great importance because having other category venues than the one we are interested for will affect the analysis of data.

At next step, we will use k-means clustering algorithm for clustering our data. 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 neighborhoods into 5 clusters, after examining the results of clusters for various numbers of k. The results give us the opportunity to identify which neighborhoods have higher density of hotels and which neighborhoods have the lowest.

Results

After clustering our data according to the total number of hotels per neighborhood, we will visualize the resulting clusters on a map as it is indicated at the following image.

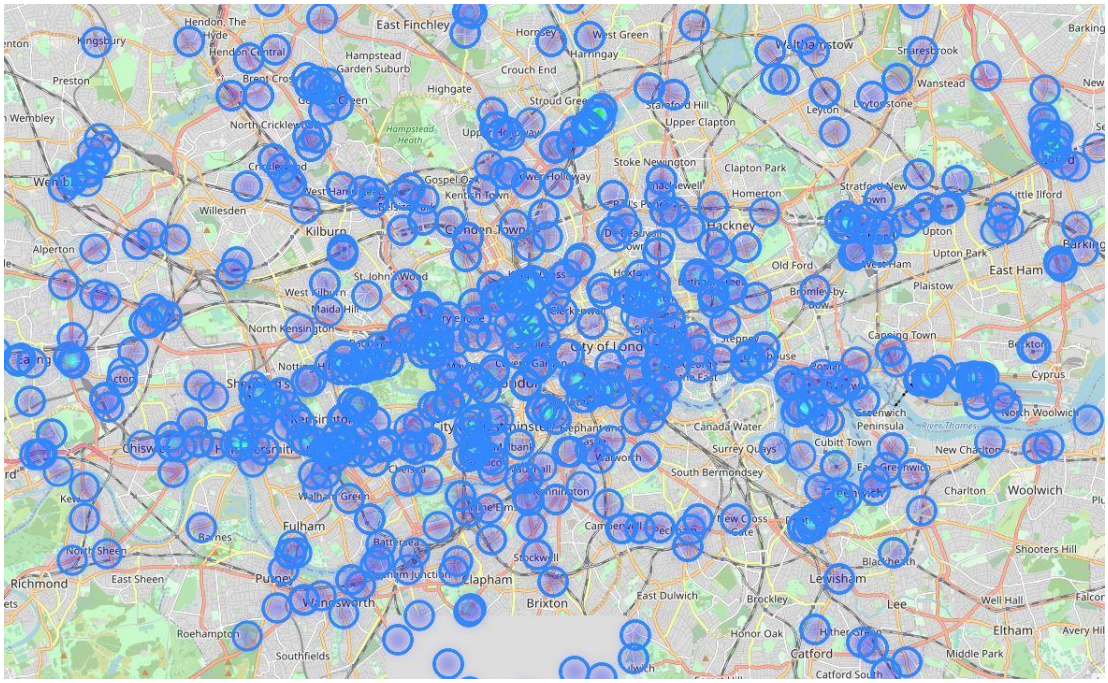


2 Figure - Map of clusters

Cluster 1: **red** color	~ 20 hotels per neighborhood
Cluster 2: **purple** color	~ 8 hotels per neighborhood
Cluster 3: **light blue** color	~ 25 hotels per neighborhood
Cluster 4: **light green** color	~ 14 hotels per neighborhood
Cluster 5: **orange** color	~ 3 hotels per neighborhood

By having a look at the cluster, the cluster 3 has the largest mean per neighborhood and the cluster 5 the smallest one. This means that the light blue colored dots on the above image correspond to the areas with high density of hotels and the orange are referring to the lowest density of hotels.

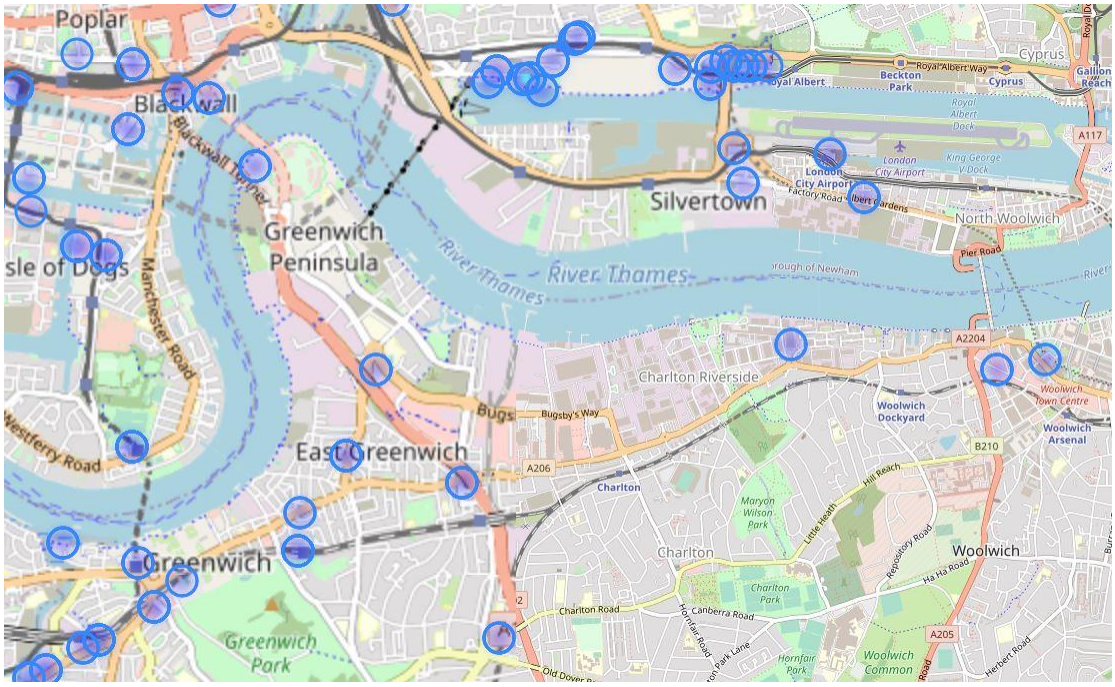
A heat map was also generated using the exact locations of all of the hotels and it is illustrated below.



3 Figure - Heat Map of Hotels' coordinates

Fortunately, the map is interactive and we can choose around each neighborhood to check in detail the areas of our interest.

An example to this, is the Figure 4 which depicts the hotels around Greenwich and City Airport.



4 Figure - Greenwich Area Zoom

It is easy to conclude that areas around airports and points of interest (such as Greenwich Observatory) have more choices for accommodation in hotels.

Conclusion

The purpose of this project was to identify neighborhoods in London and then explore the hotels and their exact locations. Clustering of those locations was then performed in order to create major zones of interest.

Keeping this in mind, we noticed after many trials that the k-means algorithm has better performance for creating 5 clusters to the neighborhoods of London.

The neighborhoods of cluster 5 has the lowest number of hotels (mean= ~ 3 hotels) while the clusters 3 and 1 are the most crowded neighborhoods according to Hotel venues added in Foursquare API from which the exact coordinates have been exported. Regarding the cluster 3 of which the mean of hotels is ~25 hotels, it is very risky to open a new hotel because the number of already existing hotels is high. The less lower risky is the neighborhoods in cluster 1 which contains approximately 20 hotels.

In addition, the heatmap would be better to be examined thoroughly to locate the most suitable and optimal location.

As a suggestion to our client it would be to choose areas between clusters 2 (purple) and 4 (light green) of which the mean number of hotels per neighborhood is ~8 and ~14 and the distance from the city center is not that high with more preference to cluster number 2.

Further research

The final decision for the optimal location for a new hotel will be made by the client based on specific characteristics of neighborhoods and distances from special locations, such as landmarks, airports, train and/or tube stations, as well as museums and restaurants. In general, the distances from the points of interest are actually factors of attractiveness of each location.

Finally, there is no doubt that further analysis has to be taken place considering the number of stars of the hotel that the client is thinking to open, by collecting the relevant data for the hotel categories by stars. In that case, probably data from *booking.com* could be retrieved for further analysis and evaluation.