Finding Suitable area to build a new Hotel in Bristol, UK

Sriram Rajagopala

31/08/2019

1. Introduction

1.1 Business Problem

The objective of this project is to use readily available data about Bristol, like post codes, geo spatial codes, top venues etc to explore the city, segment and cluster its various post code areas and help identify a suitable place to open a Hotel Business.

1.2 Background

Bristol is one of the UK's most popular tourist destinations and was selected in 2009 as one of the world's top ten cities by international travel publishers Dorling Kindersley in their Eyewitness series of travel guides. The city is steeped in history and the historical and heritage sites, like Bristol Cathedral and the Lord Mayor's Chapel, are very popular among tourists. World famous Festivals like Bristol balloon festival, Bristol Harbour Festival draw lot of tourists all year. Bristol is represented by professional teams in all the major national sports so there are regular sporting events held here as well. It is well connected by Air, Rail and Road making it ideal city to open a hotel business.

According to the British Hospitality Association (BHA), the hospitality industry now supports more than 2.7 million people. The hotel industry represents a hugely important part of it, with more than 45,000 establishments across the UK responsible for jobs in hotels and related services. Total combined turnover for the hotel industry is estimated to exceed £40bn – a significant portion of the £127bn tourist economy.
By 2025 this figure is forecast by accountancy firm Deloitte to rise to £257bn, which is around 10% of the UK's GDP. It will support 3.8 million jobs at that point, which is around 11% of the UK workforce.

Actual tourism spending in 2013 reached £113bn, with £24bn via international visitors and £89bn from domestic residents. And for every £1,000 generated directly from tourists into the industry, a further £1,800 goes into the economy via the supply chain and consumer spending, Deloitte's report Tourism: jobs and growth states.
This suggests that despite threats of terrorism, flooding, and bad summer weather in recent years, which have undoubtedly hampered the UK tourism industry, there has been strong growth, particularly following some key international events.

1.3 Target audience

Based on the background, opening a hotel in this city looks like a good business proposition.
If a hotel chain wants to open its new hotel or a startup is trying to open their first hotel, this project will aim to provide recommendation on best place to open one.

2. Data

To identify various post codes and cluster them, we will be using the following data:

- List of Post codes and their coverage - This will define the area/city in scope for this project which is Bristol, UK. Wikipedia will be used to source this data. This page has a table with Post code districts, Post town, their coverage and the Local authority area. Python's Beautifulsoup package will be used to scrape data from this page.
- Latitude and Longitude co-ordinates of these post codes which will be used to plot the map and also to get the popular venues list. This can either be found using Python's Geocoder package or pulled from a compiled list readily available on various sites. 2-Outcode file from the link was used to get the co-ordinate details for Bristol.
- Venue data will be used to perform clustering of the post code areas - Foursquare is a location technology platform which has one of the largest data of over 105 million+ locations and is used by over 150k developers. Explore function of this API will be used to get the most common venue categories in each post code, and then use this feature to group the various post codes of Bristol into clusters. K-means will be used to cluster the various post codes based on similarity to gain insights.

3. Methodology

3.1 Data Collection

Firstly Wikipedia page was used to scrape data related to Post code areas and their Coverage areas. Wikipedia page has a table containing the post code, coverage areas, post town and Local authority areas of Bristol. BeautifulSoup package was be used to scrape this data into a dataframe. Any blanks/nulls were removed to retain just the data related to post codes. We now had a dataframe consisting of 37 post codes/coverage areas.

The next step was to add geo spatial details to above dataframe. Using the spreadsheet downloaded from this site, we got Latitude and Longitude details of each of the post code areas. This data was then merged with the dataframe containing Wikipedia data. Folium map was used to view the map of the city with coverage areas superimposed on it.

Foursquare API was used to get venues data. A developer account was created on their website to request this information. Foursquare offers regular and premium services. Getting venue data is a regular request. We did not require additional details like an image of the venue, tip for the venue etc., about each venue which are categorised as premium requests. We set a radius of 1000m around each post code and a limit of 100 venues while requesting the information from Foursquare. We got a total of 824 different venues which translated to 149 different venue categories. Initial look at the data showed that there were some areas like the city centre which

were very busy returning all 100 venues requested, but there were also some areas where we were struggling to get much data. The next step was to check how many of these venues were hotels. We noticed that there were only 19 entries for Hotel in the entire set.

3.2 Segmenting and Clustering

This is a case of unlabelled data and we are trying to segment these 824 rows (or 37 rows, If we group them by post codes). K-means is an unsupervised clustering algorithm that can be used in such situations to divide the data into k non-overlapping clusters. If we can segment the venues data into K clusters, we will be able to see if there is any cluster that is more suitable to open a hotel.
The next steps were to, prepare the data for this algorithm, identify suitable k that can be used and then apply the algorithm.

a. To prepare the data for clustering algorithm we used one-hot encoding. This converts 'categorical' venue data containing 824 rows to 'numeric' data. This was then grouped by post code/coverage area and mean of the values were used to create a dataframe that was then eventually used in clustering algorithm. We then identified top 10 venues for each of the areas and populated a new dataframe.
b. There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. We used this method to find K=4 provides best result. We also made sure to set a value for random_state while calculating the best k value because when we actually got to apply the algorithm, k=4 was still best. With K-Means, the results may vary even if you run the function with the same inputs' values, hence, in order to make the results reproducible, we need to specify a value for the random_state parameter.
c. K-means algorithm was applied on the data to create 4 clusters. The cluster labels were superimposed on a folium map of Bristol to show where each of the cluster points are.

4. Results

We managed to segment and cluster Bristol city into 4 clusters using K-means. Examining the clusters we noticed that:

- Cluster 1 is busy city centre area/similar areas as it is evident from the high venue_counts in these rows. These are also the areas that have lots of cafe's, restaurants, tourist places etc.
- Clusters 2 3 and 4 are not so busy and are outside the main city region with a mix of Golf courses, pubs, restaurants etc.

5. Discussion

If we look at the count_hotel column in these clusters we can see that clusters 2, 3 and 4 have no entry. These coverage areas could be good candidates for building a new hotel. However, looking at the map these are not central to city so, stakeholders will have to think if there will be enough demand in these areas. Cluster 1 on the other hand already has a few hotels, but based on the high demand, it may be a good option to open one here. Being well connected by train, air and road helps to support this notion.

6. Conclusion

The main purpose of the project was to help identify suitable areas in Bristol where a client can open a Hotel. We have used foursquare api to identify venues in each of the areas and used that to get count of hotels in each of the areas. We identified the best k that can be used to cluster the venue data and then applied the algorithm to form 4 clusters. Analysis of these 4 clusters provided some insights to so that anyone willing to approach this problem can use this project as a starting point.

This was an attempt to collect readily available information and come up with best available options. This project can be enhanced by analysing additional points like:

- Number of hotels in each area
- Distance between each of them
- Average price of rooms in a given area
- What facilities are being planned to be included in the hotel etc