IBM Data Science Capstone Project

Title: Analysing London Boroughs

Author: Jake Smith

1. Introduction

1.1. Background

London is a large, diverse city in the United Kingdom. Its scope in terms of geography can be defined in different ways, though typically it is divided into 32 London boroughs along with the City of London district. This area is often referred to as Greater London and covers an area of 1572 square kilometres.

London has a large resident population of approximately 9 million inhabitants. In addition to this, a vast number of tourists visit the city each year; over 20 million international visitors were received in 2018. Tourism is a key industry for the economy of the city and this is emphasised by the fact that tourists who visited London in 2011 were estimated to spend £9.4 billion [1].

1.2. Description of business problem

The resident population of London, in combination with tourists who visit each year, offer lucrative opportunities for the hospitality industry. The business problem I will address in this project is related to this; can specific boroughs within the city be identified as suitable locations to set up a new craft beer bar?

For someone wishing to set up a new hospitality business venture such as this, a key concern would be the frequency that various types of business can be found across different neighbourhoods. When setting up a new bar, an individual would most likely want to know that the location that they identify is in an area in which hospitality businesses are popular and therefore can thrive. Another important factor is knowing where the most popular tourist areas are. This can be assessed in terms of the total tourist traffic they experience per year; a reasonable assumption to make is that areas more popular with tourists are likely to generate more income for hospitality businesses. Investigating both factors can help to identify areas in London that have the right target market.

To address this business problem, I will create a data frame that combines borough names, latitude and longitude coordinates for each borough scraped from Wikipedia, the 10 most commonly found venues in each borough taken from the Foursquare API, tourist trip data sourced from data websites, and finally a geo json file that marks out the separate London boroughs. The boroughs are to be clustered using the unsupervised machine learning technique of K-Means clustering according to the most commonly found venues. This data frame will be used to create an interactive map with Folium that visualises both venue data and tourist trip data for each borough. It will take the form of a choropleth map, which is an ideal way to

visualise the different levels of tourist trips across the boroughs, with coloured labels indicating how the boroughs are clustered based on the frequency with which different venues are found. The labels will also provide descriptive information including cluster descriptions and the popularity of each borough based on categories.

1.3. Interest in the problem

This business problem is likely to be of interest to anyone who intends to enter the hospitality industry in London, specifically bars, pubs and other types of drinking establishment. Similarly, those who already have a hospitality business but are considering expanding into more locations would find this information useful.

This information would also be interesting for tourists who want to gain a good idea of the different areas in London before they visit, including popularity of each borough alongside the typical venues they are likely to encounter in different places.

2. Data

2.1. Data description and sources

- The names, latitude and longitude co-ordinates of each London Borough and the City of London district (which is not officially classed as a borough even though it is in central London) will be sourced by web scraping from Wikipedia using BeautifulSoup [2].

- The Foursquare API will be used to obtain the most common types of venue found within the different London boroughs [3].

- Tourism trips to each London borough (thousands per year) for the year 2007 is to be sourced from the UK government data website [4].

- Geo json data that marks out the boundaries for each London borough will be used to create a choropleth map; this is to be sourced from London Data Store website [5].

References

[1] Wikipedia – London

[2] Wikipedia – London boroughs

[3] Foursquare API

[4] UK Government data site - tourism trips to London boroughs

[5] London Data Store - geo json data for London boroughs