Applied Data Science Capstone Project

Finding the most attractive neighborhood in London based

1. Introduction

1.1. Background

London is the capital of and largest city in England and the United Kingdom. From an administrative point of view, the city is structured in 32 boroughs that have been created in 1965. The city stands on the River Thames in the south-east of England, at the head of its 50-mile (80 km) estuary leading to the North Sea.

London is one of the world's most important global cities. It exerts a considerable impact upon the arts, commerce, education, entertainment, fashion, finance, healthcare, media, professional services, research and development, tourism and transportation.

London contains four World Heritage Sites: The Tower of London; Kew Gardens; the site comprising the Palace of Westminster, Westminster Abbey, and St Margaret's Church; and the historic settlement in Greenwich where the Royal Observatory, Greenwich defines the Prime Meridian (0° longitude) and Greenwich Mean Time. Other landmarks include Buckingham Palace, the London Eye, Piccadilly Circus, St Paul's Cathedral, Tower Bridge, Trafalgar Square and The Shard. London has numerous museums, galleries, libraries and sporting events. These include the British Museum, National Gallery, Natural History Museum, Tate Modern, British Library and West End theatres. The London Underground is the oldest underground railway network in the world.

1.2. Problem

Data regarding venues and neighborhood attractions may determine where a tourist picks a hotel. For the scope of this analysis, other potential influencing factors will be excluded such as accommodation rating, rate per night etc. The aim of this project is to find the most attractive neighborhood based on how many and what type of venues are located within a specific area.

1.3. Interest

The main interest is from a touristic point of view e.g. tourists or travel agencies.
This is not limited to external visitors and the analysis could be helpful to current residents of the area that perhaps moved recently and did not discover the surrounding area, or long-term residents that are not aware of all the attractions developed in the recent years.

2. Methodology

2.1. Data sources and pre-processing

As data source, for neighborhood data will be scraped from the Wikipedia page https://en.wikipedia.org/wiki/List_of_areas_of_London using BeautifulSoup. Raw table provides information about location, London borough, post town, postcode district, dial code and OS grid ref.

Additional information required about the geographical coordinates data will be extracted using ArcGis. ArcGis has no limitation in regards of potential API calls, which fits our needs very well.

We start the data preprocessing by replacing the spaces with _ and remove the [] from borough column.

In feature selection we select only the columns needed for the analysis e.g., London borough, post town, postcode district and rename these columns with a simpler name: borough, town, postcode.

Further for feature engineering, because our dataset includes a lot of information about the inner and outer London area, we'll select for further analysis only the info about London town which will be the City of London.

Analysis continues with ArcGis gathering the geographical coordinates to plot our initial map of London without any markers for visualization purposes and adding the initial markers for neighborhoods.

Using KMeans clustering method, we cluster based on top venues categories and then plot the map of London again. We have selected the top 10 most frequent venue categories and clustered based on this in 10 clusters.

3. Results

By analyzing each cluster, we can see the most common venue category for each cluster. Cluster 2 includes in the most frequent Coffee shops, pubs and restaurant which makes it the most attractive area compared with cluster 6 where the most common venue is supermarket which from residents' point of view would be of interest, but not as much for a tourist.

4. Discussion

The neighborhoods of London are multicultural. There are a lot of different cuisines including Indian, Italian, Turkish and Chinese found in the large variety of restaurants, bars, juice bars, coffee shops, fish and chips shops and breakfast spots.

London has as well a lot of shopping options: flower shops, fish markets, fishing stores, clothing stores and supermarkets.

The main modes of transport seem to be buses and trains.

For leisure, the neighborhoods are set up to have lots of parks, golf courses, zoo, gyms and historic sites.

5. Conclusion

Having such a variety available in all areas, it’s hard to pick one neighborhood based only on this aspect so for further analysis in order to choose the optimal neighborhood for visiting, other factors need to be taken in consideration which leaves our decision open for discussion. This is an opportunity for further processing including rating of accommodation, perhaps criminal rates in neighborhoods and of course proximity to public transport for accommodation.

For the moment, based on this project, I would personally pick a center location just out of conveniency to avoid spending too much money on transportation and enjoy walking and exploring all these areas by foot while keeping fit.