Clustering London, a data analysis approach

A. Introduction/Business problem

London, the capital of England and the United Kingdom, is a bustling, lively city considered to be one of the world's most important global cities with a diverse range of people and cultures and more than 300 languages spoken in the region. It exerts a considerable impact upon arts, commerce, education, entertainment, fashion and finance.

Administratively, London is divided in 32 boroughs. Together with City of London, these boroughs form 33 local authority districts, which make up Greater London. The region covers 1,572 km2 (607 sq mi) and had a population of 8,173,920 according to 2019 data.

Due to its significance in the global scene and the variety of opportunities it offers, London still attracts people from all over the world seeking for new endeavors. An established Greek company, which runs its own microbrewery as well as several delicatessen stores, is exploring the idea of setting up a new business in London and needs help in terms of identifying areas with possible opportunities. Data analysis tools could help us providing an answer to a couple of strategic questions regarding this business idea:

- Can we identify boroughs with potential of setting up a microbrewery and/or a delicatessen store?
- How could household income and population affect our decision?

B. Data Description

The following data sources will be used to get an insight into our business problem:

1. First dataset consists of London Boroughs with respective median household incomes. This dataset will be used primarily to depict the variability of the median household income across London

boroughs, with the use of a choropleth map. We will also use it in conjunction with the second dataset to build the final dataset to be used for our analysis. It is compiled from two different sources: CACI & https://www.trustforlondon.org.uk/ and can be found on my GitHub page.

2. Second dataset consists of London boroughs with geographical co-ordinates and population data. As described above, it will be used in conjunction with the first dataset to build the final dataset needed for our analysis. Source : https://github.com
3. GeoPy enables us to get geographical coordinates of London Boroughs in a json file. This will serve as the primary data source for building a choropleth map. Source: https://githb.com
4. Foursquare API will be used to explore London boroughs and retrieve their respective most common venues. Results from these queries will be utilized as key feature for clustering London boroughs.