CAPSTONE PROJECT - FINAL REPORT

ANALYZING THE DIFFERENT NEIGHBOURHOODS OF LONDON TO OPEN A SUPERMARKET

INTRODUCTION

London is considered to be one of the world's most important global cities and has been called the world's most powerful, most desirable, most influential, most visited, most expensive, sustainable, most investment-friendly, and most-popular-for-work city. It exerts a considerable impact upon the arts, commerce, education, entertainment, fashion, finance, healthcare, media, professional services, research and development, tourism and transportation. London ranks 26th out of 300 major cities for economic performance. It is one of the largest financial centres and has either the fifth- or the sixth-largest metropolitan area GDP. It is the most-visited city as measured by international arrivals and has the busiest city airport system as measured by passenger traffic. It is the leading investment destination, hosting more international retailers and ultra high-net-worth individuals than any other city.

London has a diverse range of people and cultures, and more than 300 languages are spoken in the region. Its estimated mid-2018 municipal population (corresponding to Greater London) was 8,908,081, the third most populous of any city in Europe and accounts for 13.4% of the UK population. London's urban area is the third most populous in Europe, after Moscow and Paris, with 9,787,426 inhabitants at the 2011 census. The London commuter belt is the second-most populous in Europe, after the Moscow Metropolitan Area, with 14,040,163 inhabitants in 2016.

BUSINESS PROBLEM

Analyse the different neighbourhoods of London and to find the areas which have the least number of supermarkets to establish a new one. Opening a new supermarket in areas where the present number of supermarkets is low reduces the competition between the supermarkets and increases the sales.

DATA SECTION

The data needed for performing an analysis on the neighbourhood of London to find a suitable area to establish a supermarket are -

1. **Neighbourhood data**
    a. **Data source** - https://en.wikipedia.org/wiki/Category:Areas_of_London
    b. **Data Description** - This contains the list of all the areas in London.

2. **Venues in each Neighbourhood**
    a. **Data source** - Foursquare API
    b. **Description** - Using the Foursquare API we get the different venues in each neighbourhood. From which we can determine which neighbourhood has the least number of supermarkets.

3. **Geographical Coordinates**

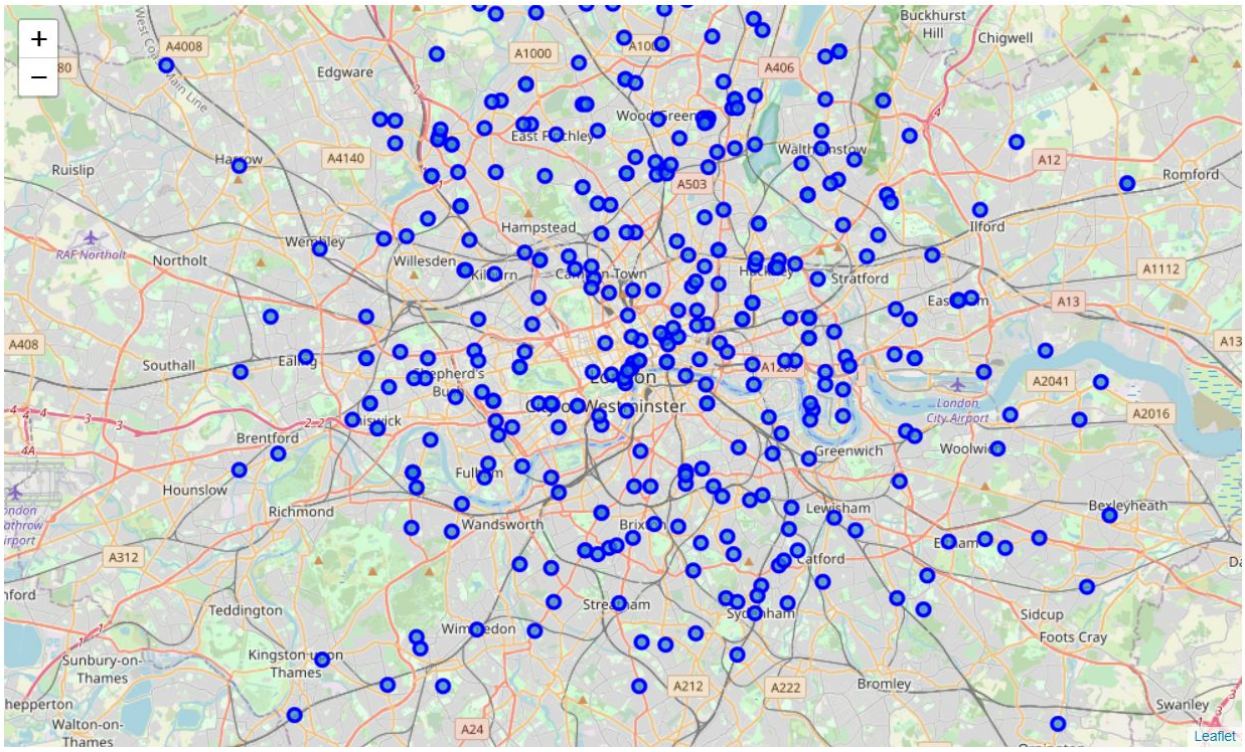
    The geographical coordinates ie. Latitude and Longitude are extracted using the GeoPy library in Python.

METHODOLOGY

Step 1 - Folium Maps

Folium makes it easy to visualize data that's been manipulated in Python on an interactive Leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing Vincent/Vega visualizations as markers on the map. The library has a number of built-in tilesets from OpenStreetMap, MapQuest Open, MapQuest Open Aerial, Mapbox, and Stamen, and supports custom tilesets with Mapbox or Cloudmade API keys. Folium supports both GeoJSON and TopoJSON overlays, as well as the binding of data to those overlays to create choropleth maps with color-brewer color schemes.

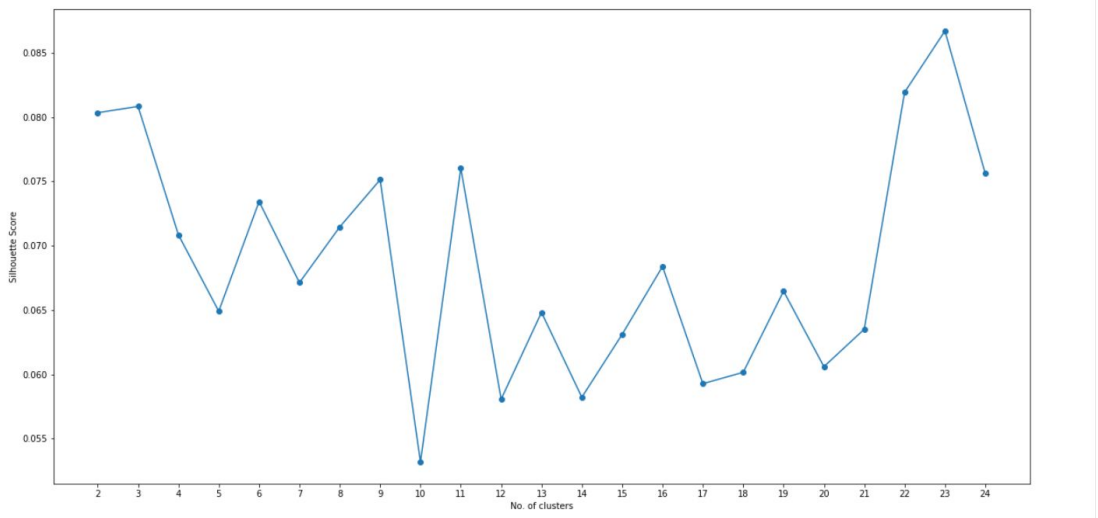
With the help of folium maps and the latitude and longitude details obtained from the geospace data, the different neighbourhoods in London are plotted on a map.



Step 2 K-means clustering

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. It is popular for cluster analysis in data mining. *k*-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.

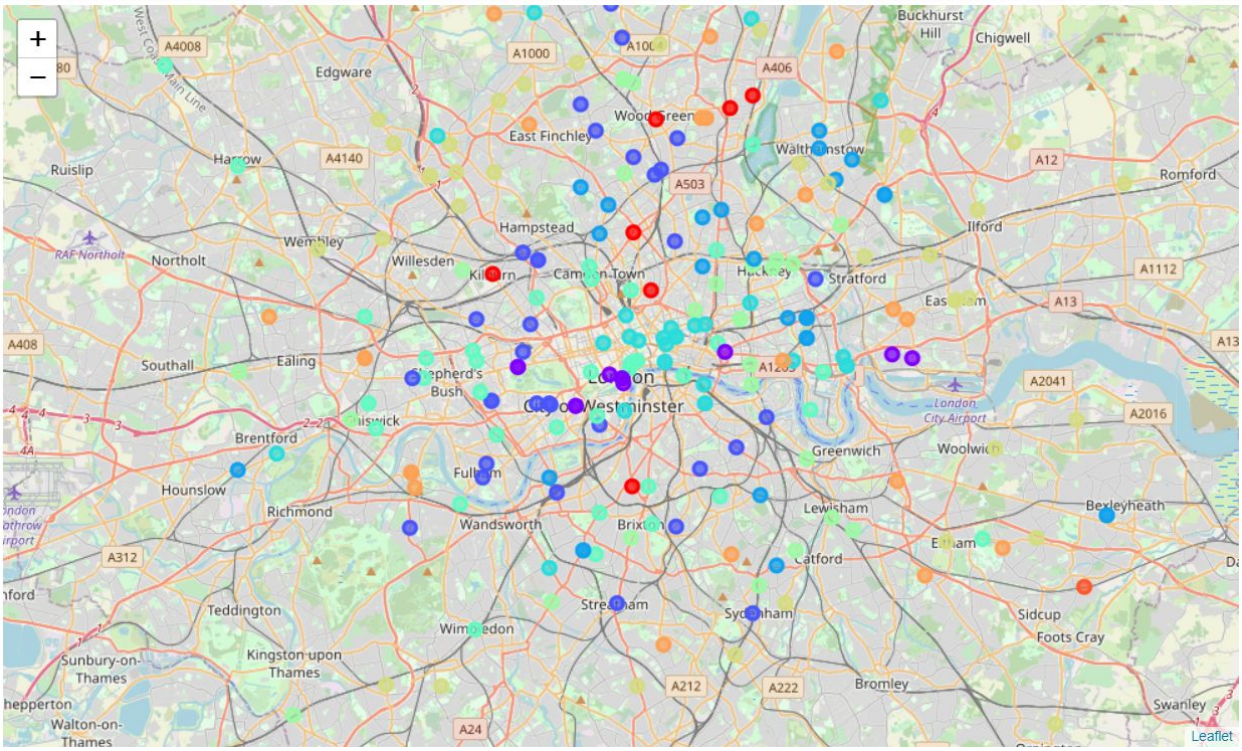
After applying the K-means clustering Algorithm to our analysis we obtained the following graph from which the optimal number of clusters is derived.



From the graph it is observed that the optimal number of clusters needed for the city of London is 10.

Step 3 - Applying Clusters with Folium maps

The different clusters are highlighted using folium maps to show the different neighbourhoods in a specific cluster. The map of London after clustering is therefore obtained.



RESULT

After finding the top 10 categories in each neighbourhood, analysis is done by checking the number of supermarkets in all the clusters.

The number of supermarkets in the 10 neighbourhoods was found as -

Neighbourhood 1 - 4

Neighbourhood 2 - 2

Neighbourhood 3 - 11

Neighbourhood 4 - 3

Neighbourhood 5 - 2

Neighbourhood 6 - 8

Neighbourhood 7 - 7

Neighbourhood 8 - 26

Neighbourhood 9 - 6

Neighbourhood 10 - Nil

DISCUSSION

It is observed that Cluster number 8 has the maximum number of supermarkets and Cluster numbers 2,4,5 and 10 have the least to nil supermarkets in the neighbourhood. Therefore, it is recommended to start a supermarket in the clusters of 2,4,5 or 10 since the competition will be low and the sales will be higher.

CONCLUSION

Therefore using Python and machine learning analysis was performed to find areas with the least number of supermarkets in London. This same type of analysis can be done to find various other categories of business in any other place in the world.