Data

To identify the problem and quantify the results the following data have been used:

1. Population Projection https://data.london.gov.uk/dataset/trend-based-population-projections
2. Foursquare Developers Access to venue data: https://foursquare.com/

Methodology

The methodology includes:

- Data retrieval, exploration and wrangling
- Performing K-means clustering algorithm to segment neighborhoods
- Visualizing population projections and neighborhood segments
- Understand growth pattern and urban shift

Data retrieval, exploration and wrangling

1. The population dataset, an excel file, is provided by GLA and aggregates population projections for different geographical divisions from 2011 to 2050. Spatial geographies used in this dataset are:

    - London Government Statistical Service Codes (gss_code) of the 33 Local District Authorities
    - London Government Statistical Service Names (gss_name) of the 33 Local District Authorities

    Data of Population Projection from 2011- 2050 is retrieved in a Pandas DataFrame and

    grouped by gss-name (renamed as Area).

    From this data a new column is created containing % of population growth from 2011 to 2050.

    Latitude and Longitude of each district is retrieved using Geocoder from Geopy Libraby in Python.

	Area	Latitude	Longitude
0	Barking and Dagenham, London, United Kingdom	51.554117	0.150504
1	Barnet, London, United Kingdom	51.653090	-0.200226
2	Bexley, London, United Kingdom	51.441679	0.150488
3	Brent, London, United Kingdom	51.563826	-0.275760
4	Bromley, London, United Kingdom	51.402805	0.014814

2. **Foursquare API is used to explore types of venues and their frequencies in each district. This data is issued to classify districts based on their urban development.**
   Foursquare identifies 10 top-level categories, with some sub-categories that have not been used in this project.

Arts & Entertainment	College & University	Events	Food	Nightlife Spot
Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport

The results were analyzed through a Box Plot.

Some outliers for Art & Entertainment and Outdoors & Recreation. Not enough data for Event, so this category will be dropped.

Performing K-means clustering algorithm to segment local districts

The following conclusions have been made for the different number of clusters:

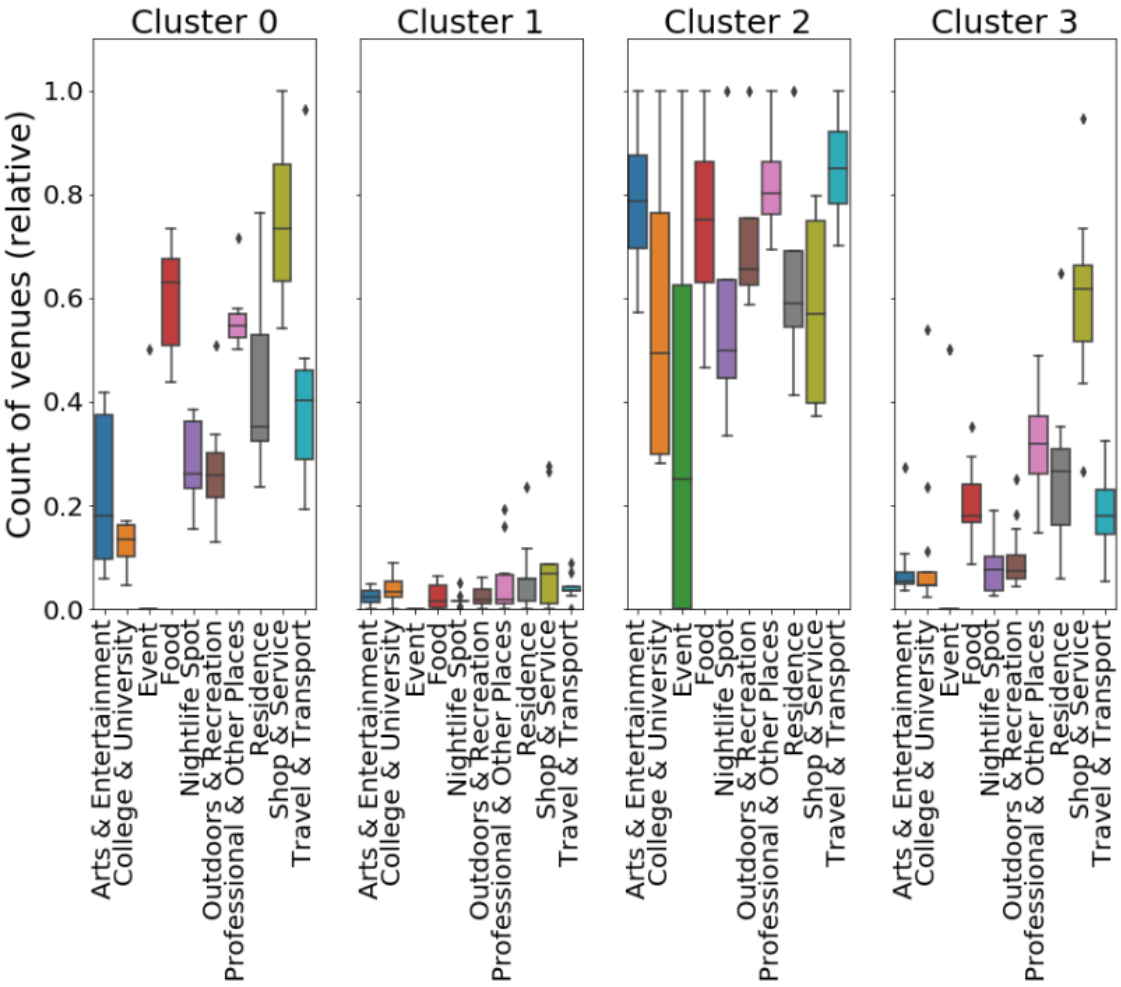
• 2 clusters only show the uptown/downtown divide

• 3 clusters add clustering within the downtown

• 4 clusters add clustering within uptown and downtown

• 5 and more clusters are difficult to interpret

For the final analysis, 4 clusters have been selected because they are the best fit for this project.



Some interpretations of the clusters, based on the Box Plot above:

• Cluster 0 (blue) has moderate scores with shops and services being the most popular. These are developed residential suburbs

• Cluster 1 (green) has low frequencies for all venue categories. These appear to be underdeveloped neighborhoods

• Cluster 2 (red) has consistently high frequencies for all venue categories. This is the most diversely developed part of city

• Cluster 3 (magenta) has high frequencies but with less residential places and more professional places. These are the developed professional or industrial suburbs

Plotting latitude and longitude confirmed that the most developed area of the city is concentrated in most central zone of the city and the outer area are relatively less developed.