The Battle of Neighbourhoods in London

Applied Data Science Captstone – part of the IBM Data Science Professional Certificate

Introduction

London is one of the most diverse and populated city in the world with approximately 9 million people living in Greater London. London itself is divided into 32 boroughs plus the City of London, each of which is different in terms of, for example, infrastructure, housing prices and average income as well as ethnic distribution.

The purpose of this study is to analyse whether different clusters of boroughs, as determined by a k-means algorithm on Foursquare venue category data, also have similar social, demographic and economic attributes. For example, do boroughs of the same cluster have similar household income, well-being levels and crime rates? Further, how would London's boroughs be clustered based on socio-economic data?

Data Description

Wikipedia will be used initially to obtain location data for each borough. This data is readily available here: https://en.wikipedia.org/wiki/List_of_London_boroughs.

Foursquare data will be used to collect venue information for each borough, group the data by venue category and cluster London's boroughs using a k-means algorithm.

The London Datastore will be used to obtain additional information for each borough which will be used to compare boroughs within each cluster. The following data sets will be used:

- Average housing price data (https://data.london.gov.uk/dataset/average-house-prices)
- Income (https://data.london.gov.uk/dataset/average-income-tax-payers-borough)
- Employment (https://data.london.gov.uk/dataset/economic-activity-rate-employment-rate-and-unemployment-rate-ethnic-group-national)
- Education (https://data.london.gov.uk/dataset/qualifications-working-age-population-nvq-borough)
- Social integration figures (https://data.london.gov.uk/dataset/social-integration-headline-measures)
- Personal well-being statistics (https://data.london.gov.uk/dataset/recorded_crime_rates?resource=c051c7ec-c3ad-4534-bbfe-6bdfee2ef6bb)
- Recorded crime figures (https://data.london.gov.uk/dataset/subjective-personal-well-being-borough)

The aforementioned data is subject to the open government licence (http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/).

Methodology

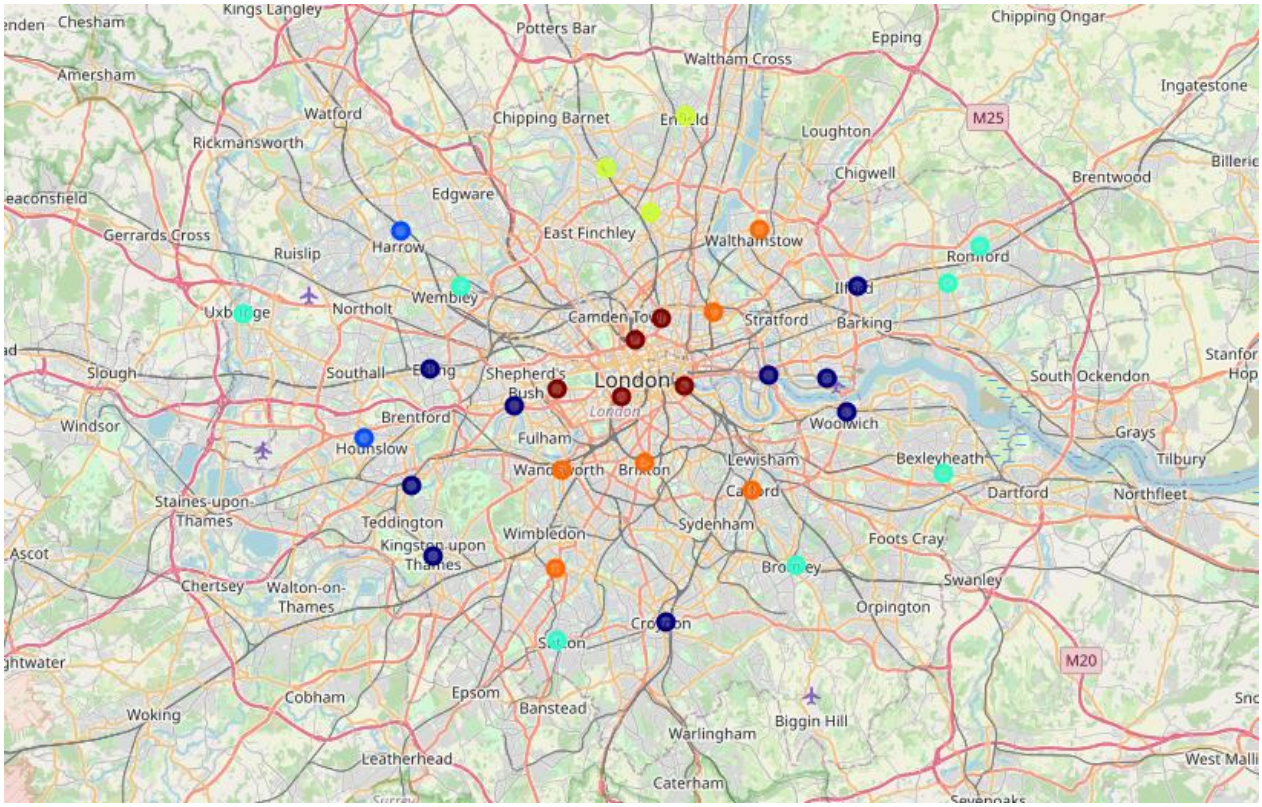
After obtaining location data for each borough using BeautifulSoup, Foursquare venue data will be obtained using a 5 kilometre radius and a 1,000 venue limit. This data will then be transformed by aggregating the number of venues in each venue category at borough level. This table of the most common venues will then be used for the clustering exercise. Clusters are determined by using a k-means algorithm with k = 6.

A number of maps will then be created based on the results of the aforementioned k-means algorithms (see below). Further, additional overlays will be used to investigate how socio-economic data matches to the determined clusters.

In the second part of the study, socio-economic data from the London Datastore will be used to determine a second set of clusters based on these data. Again, a k-means algorithm with k = 6 will be used.

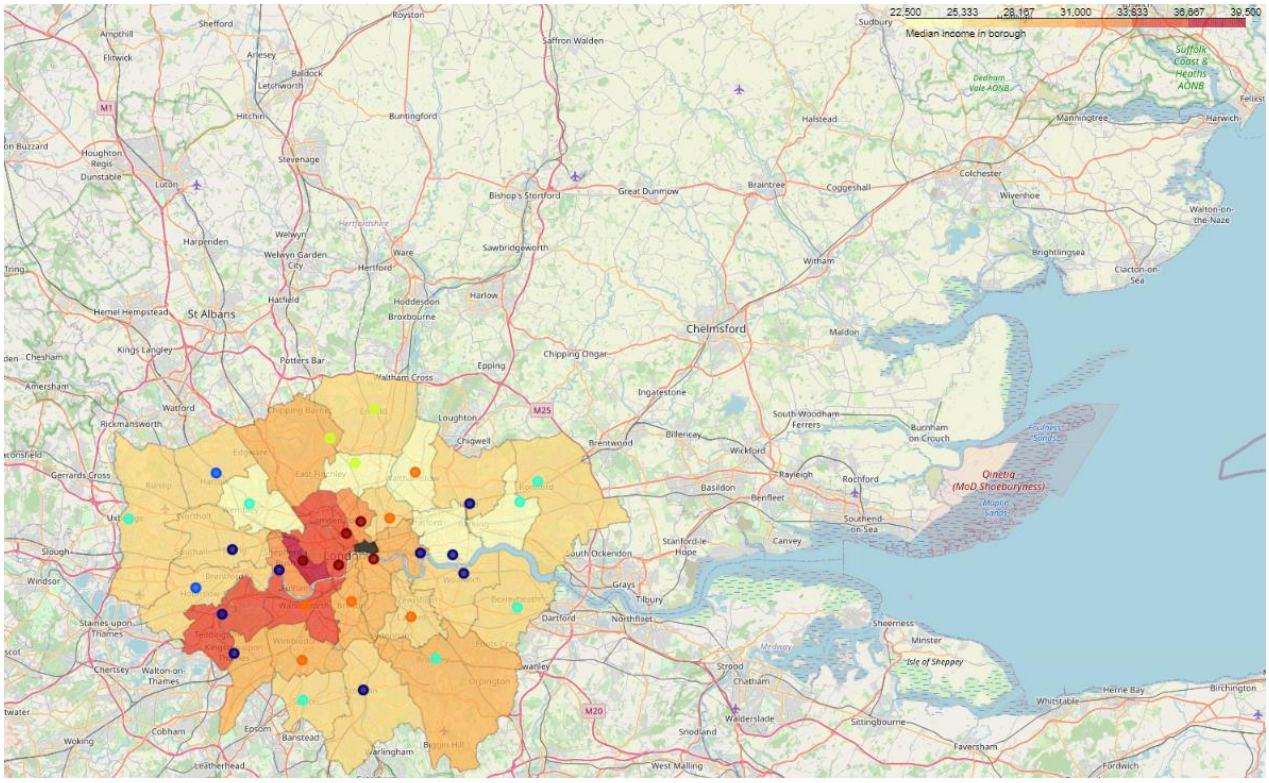
Results and Discussion

Clusters based on Foursquare Venue Data

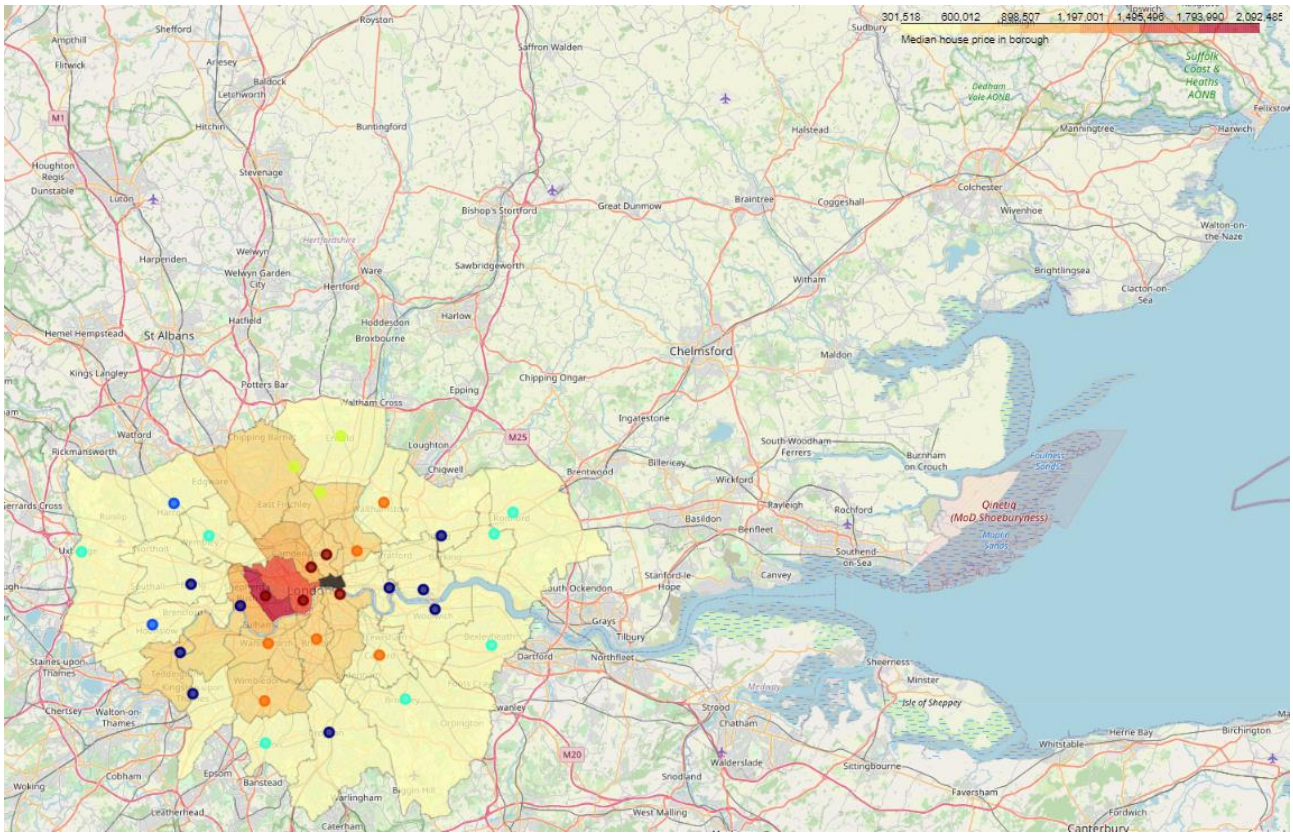


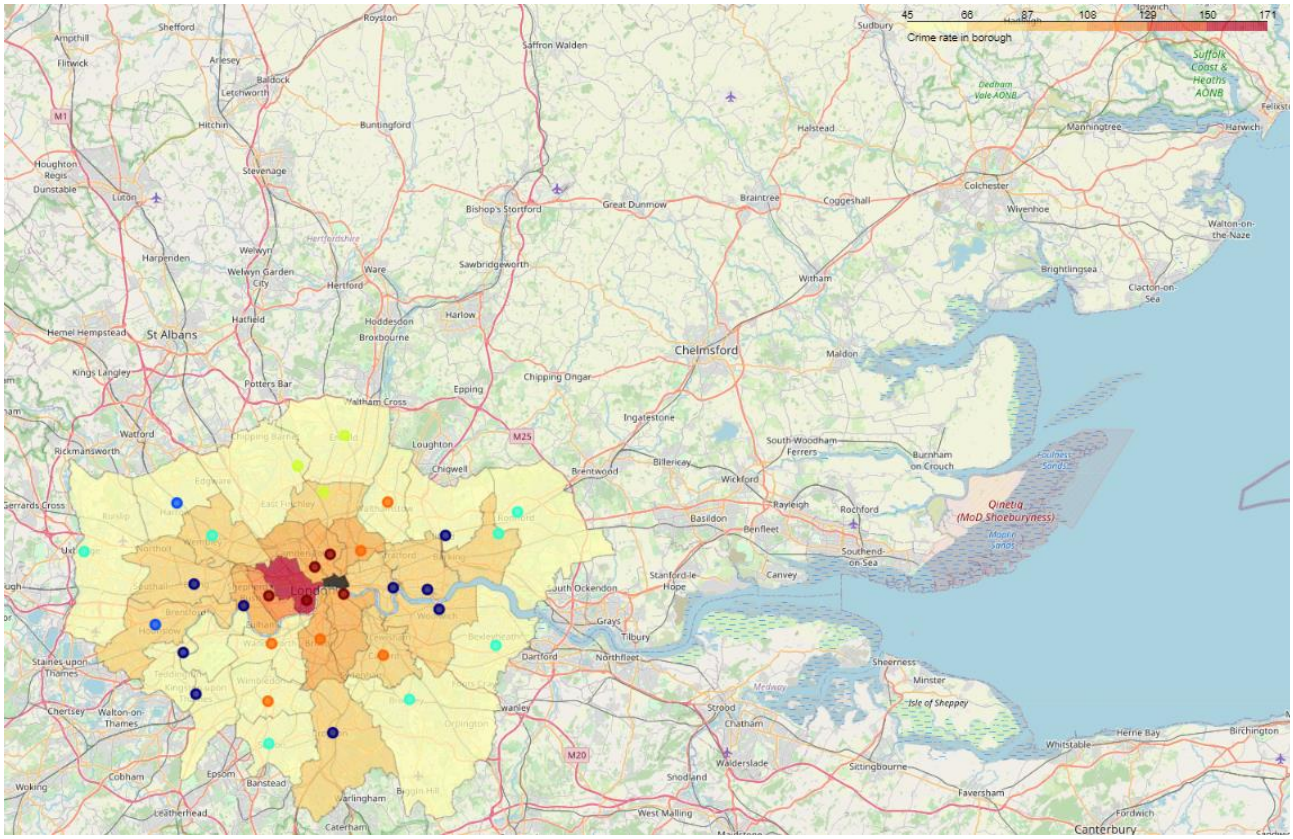
We can see the following clusters:

- **Cluster 0 (red):** concentrated at the centre of London and includes, for example, Westminster and Camden.
- **Cluster 1 (purple):** divided into West London and East London boroughs with cluster 0 in-between.
- **Cluster 2 (blue):** concentrated in West London.
- **Cluster 3 (pink):** various (outer) boroughs around London.
- **Cluster 4 (yellow):** concentrated in North London.
- **Cluster 5 (orange):** concentrated in South-East London

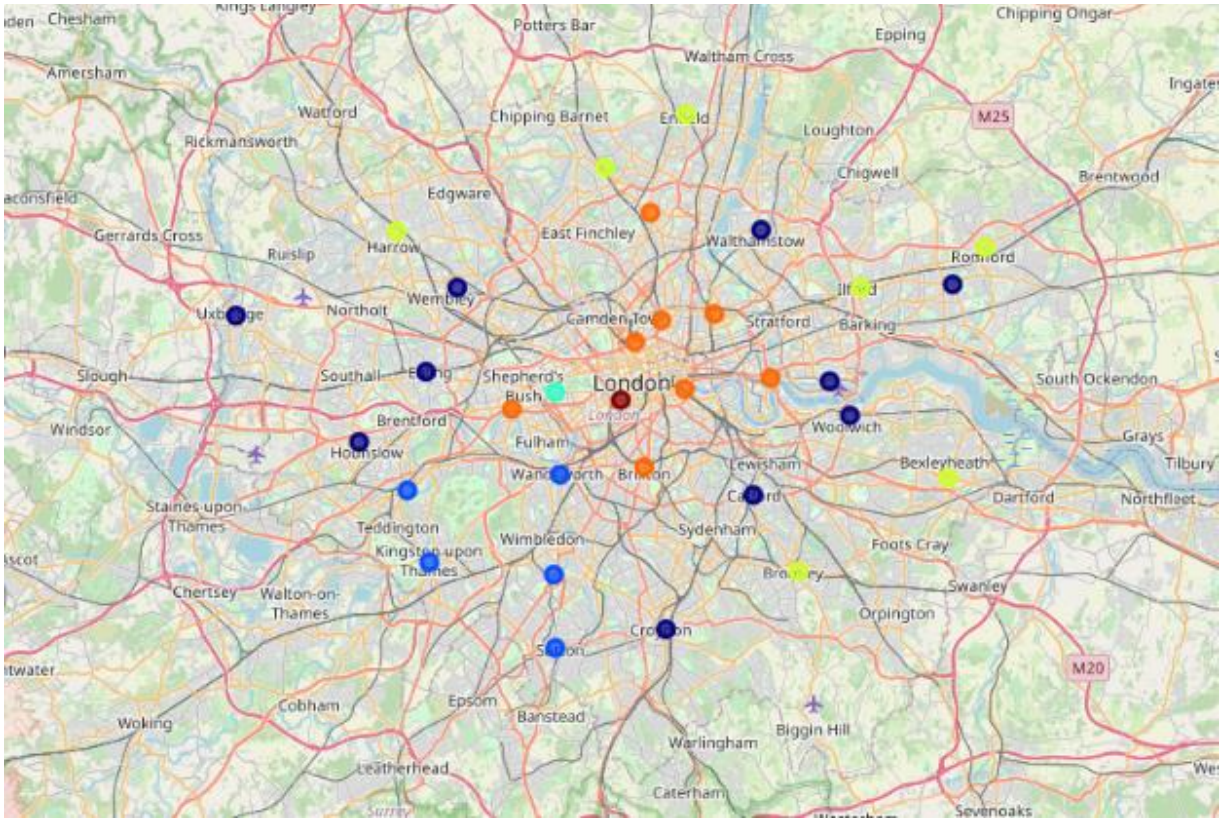


There does not appear to be a clear relationship between median income, median house price nor crime rates and clusters based on Foursquare venue data.





Clusters based on socio-economic Data



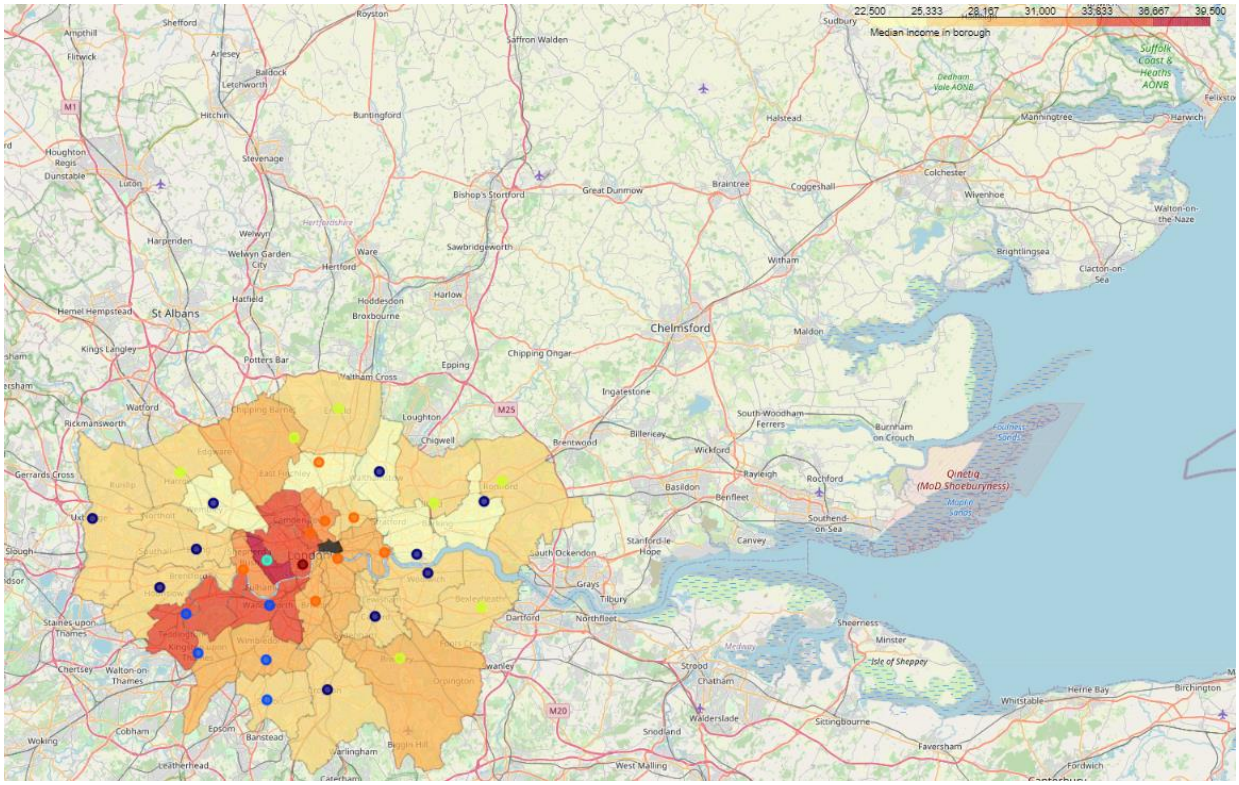
We can see the following clusters:

- **Cluster 0 (red):** single borough, only consists of Westminster.
- **Cluster 1 (purple):** some of the outer boroughs.
- **Cluster 2 (blue):** concentrated in (South-)West London.

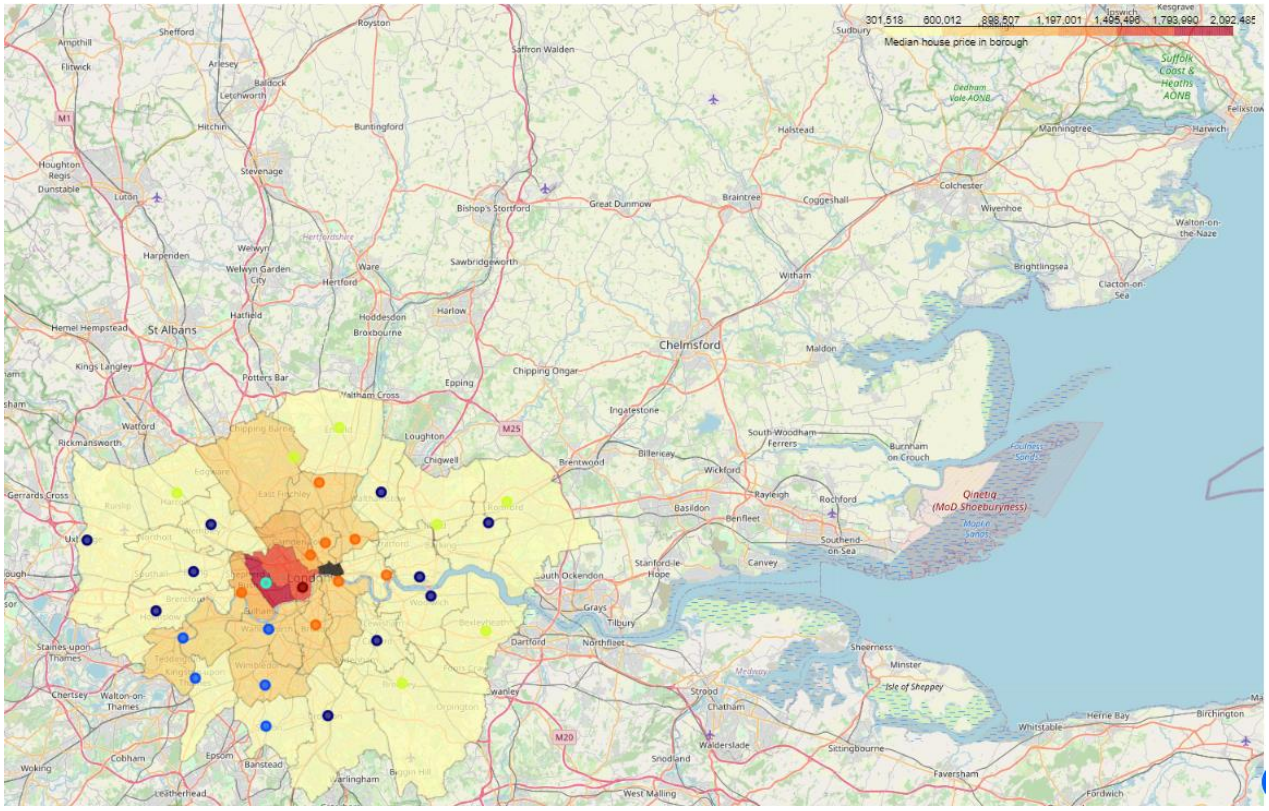
- **Cluster 3 (pink):** single borough, only consists of Kensington and Chelsea.
- **Cluster 4 (yellow):** outermost boroughs in (North-/South-)East London.
- **Cluster 5 (orange):** mainly consists of central London boroughs

In comparison to the clusters based on Foursquare venue data, boroughs in central London still appear to be together in the same cluster. One noticeable difference is that Westminster (red) and Kensington and Chelsea (pink) are in different clusters to the other central London boroughs.

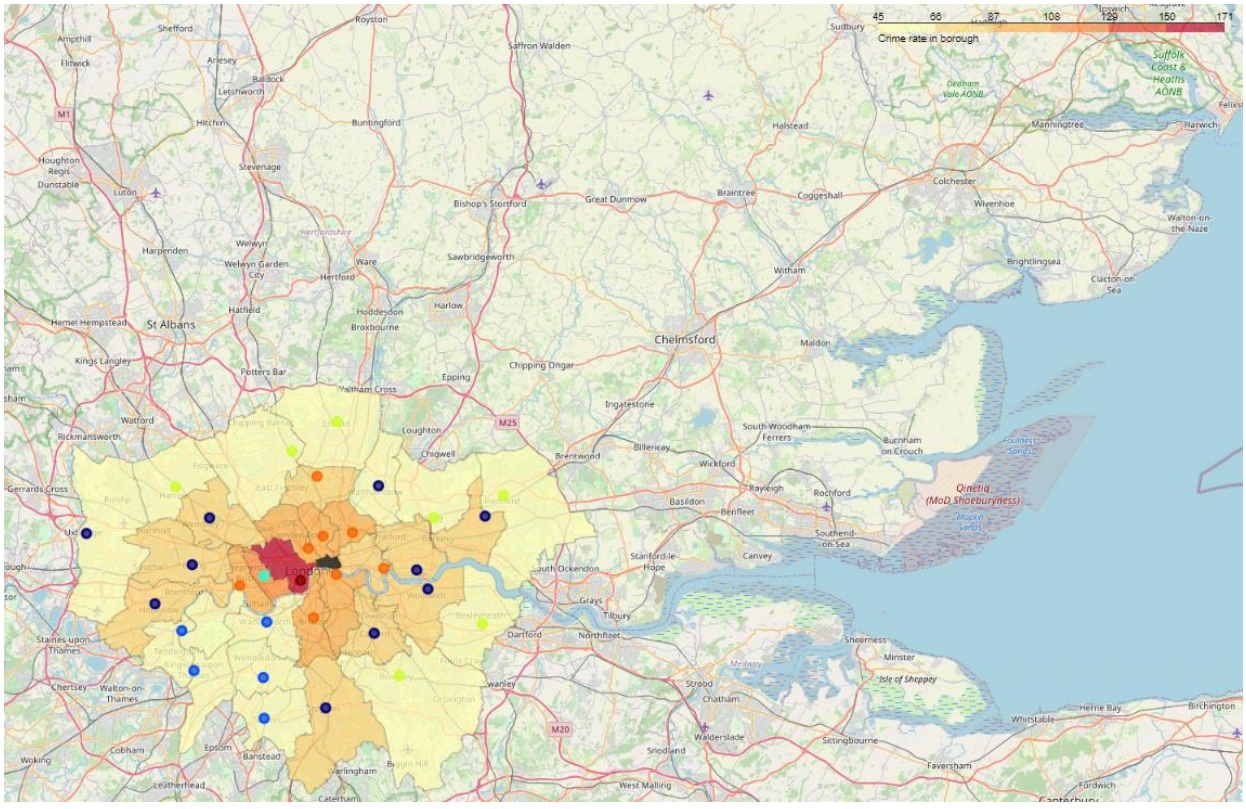
Further, there appears to be a concentration of clusters in certain parts of London, for example, blue boroughs are exclusively in South-West London while yellow boroughs consist of outer boroughs in North-East London.



The does not appear to be a clear trend for different clusters when looking at median income. However, in general, central London boroughs and a handful of West London boroughs have the highest median income in London. Median Income is lower in South and East London and lowest in a handful of boroughs in North London.

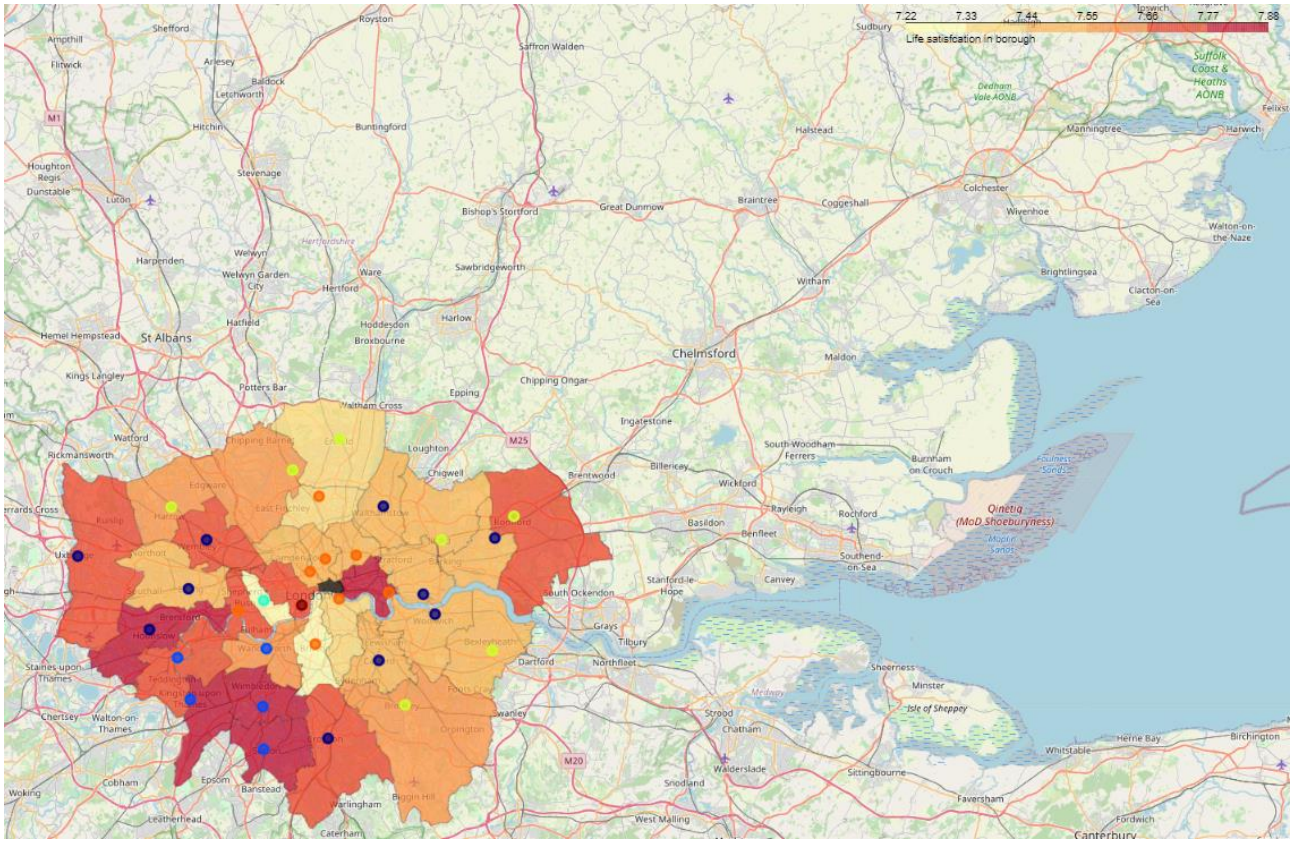


Westminster (red) and Kensington and Chelsea (pink) stand out with the highest median house prices. In contrast to median income, there is a more profound trend for median house prices. Blue and orange boroughs have (mostly) similar levels of median house prices. They are the most expensive boroughs after Westminster and Kensington and Chelsea. On the other hand, yellow and purple boroughs which mainly consist of outer London boroughs have the lowest median house prices. We would expect this to be the case.



Similar to median house prices, crime rates appear to be similar within each cluster. Again, Westminster (red) and Kensington and Chelsea (pink) stand out with significantly higher crime rates which might be due to an

increased number of robberies and break-ins given their affluent status. In contrast, blue boroughs appear to have low crime rates despite having high median income and house prices. Orange boroughs have slightly higher crime rates than purple boroughs. Both of which have higher crime rates than blue and yellow boroughs.



In terms of life satisfaction there appears to be no definitive trend within each cluster. In general, (South-)West London boroughs appear to have the highest levels of life satisfaction. Interestingly, despite having the high levels of median income and house prices, life satisfaction is lowest in Kensington and Chelsea (pink).

Conclusion

Clustering London's boroughs based on Foursquare venue data does not necessarily reflect the underlying socio-economic attributes. Given that venue data is dynamic, clusters change regularly thereby making it difficult to reach definitive conclusions.

Future studies should look at additional socio-economic attributes as well as combining these attributes with, for example, Foursquare venue data. Further, it should also be investigated whether the number