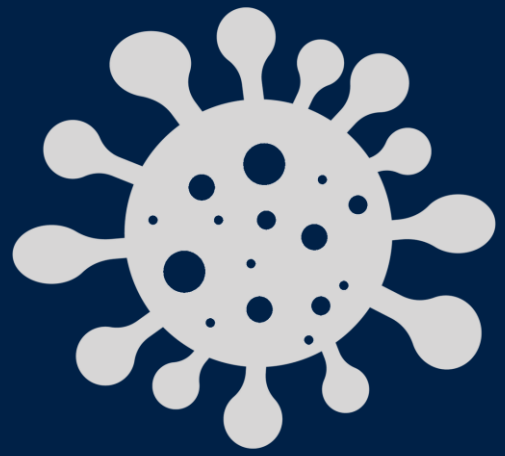


# Using Demographics and Location-Based Social Network Data to Understand the Spread of COVID-19 in London

Author:
Thomas Hall

Date:
Thursday, 07 May 2020

# The Problem

- The virus, SARS-CoV-2, and the disease it causes, COVID-19, are poorly understood due to their novelty.

- Uncertainty over the disease has led to the shutdown of entire economies.

- By understanding risk factors and transmission characteristics, can boroughs at high risk of a serious outbreak be pre-empted and planned for accordingly?

# Example Questions

- Are deprived areas at more risk of developing COVID-19?
- Are areas with an older population or those with higher BAME populations at greater risk of having more COVID-19 cases?
- Is prevalence increased by the number of certain venue types, e.g. pubs, restaurants, or parks?

# Data Sources

- **Demographics:** comprehensive profiles for each London borough, sourced from the Greater London Authority.

- **Venues**: top two-hundred venues for each borough, sourced from the Foursquare Places API.

- **Coronavirus:** cumulative lab-confirmed cases for each borough, sourced from Public Health England.

# Methodology

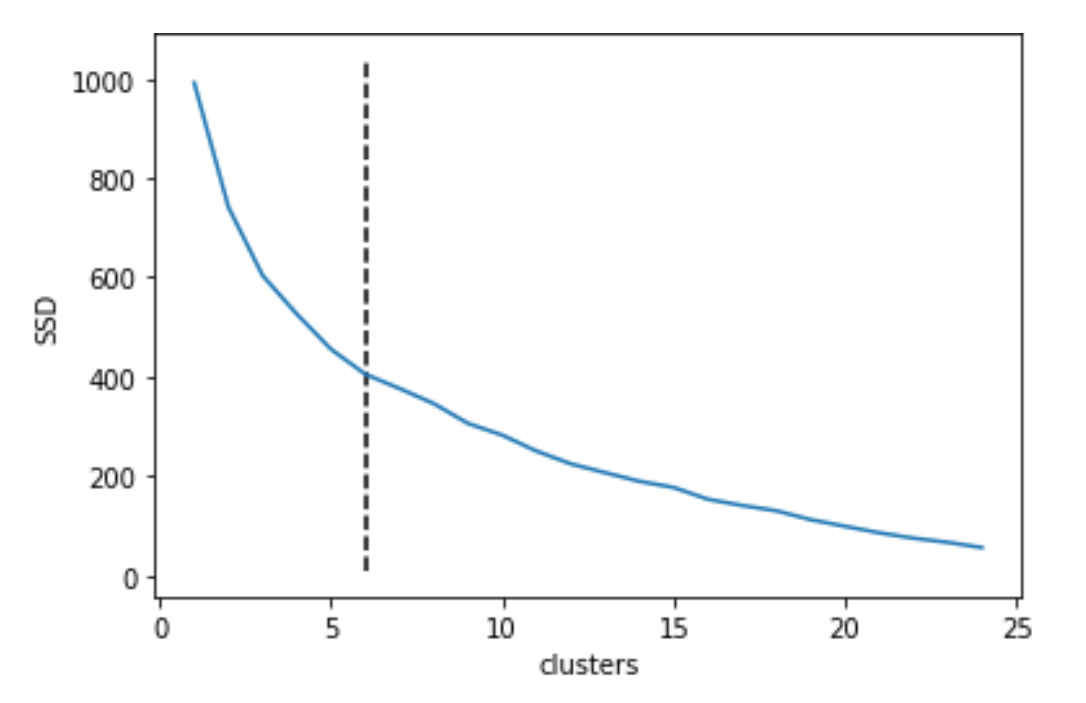
- Two clustering studies are performed;  by **demographics** and by venues.

- K-means clustering is performed with the elbow method used to define the number of clusters and twelve initialisations per number of clusters to avoid local minima.

- Inter-cluster trends in COVID-19 prevalence rates are assessed for statistical significance.

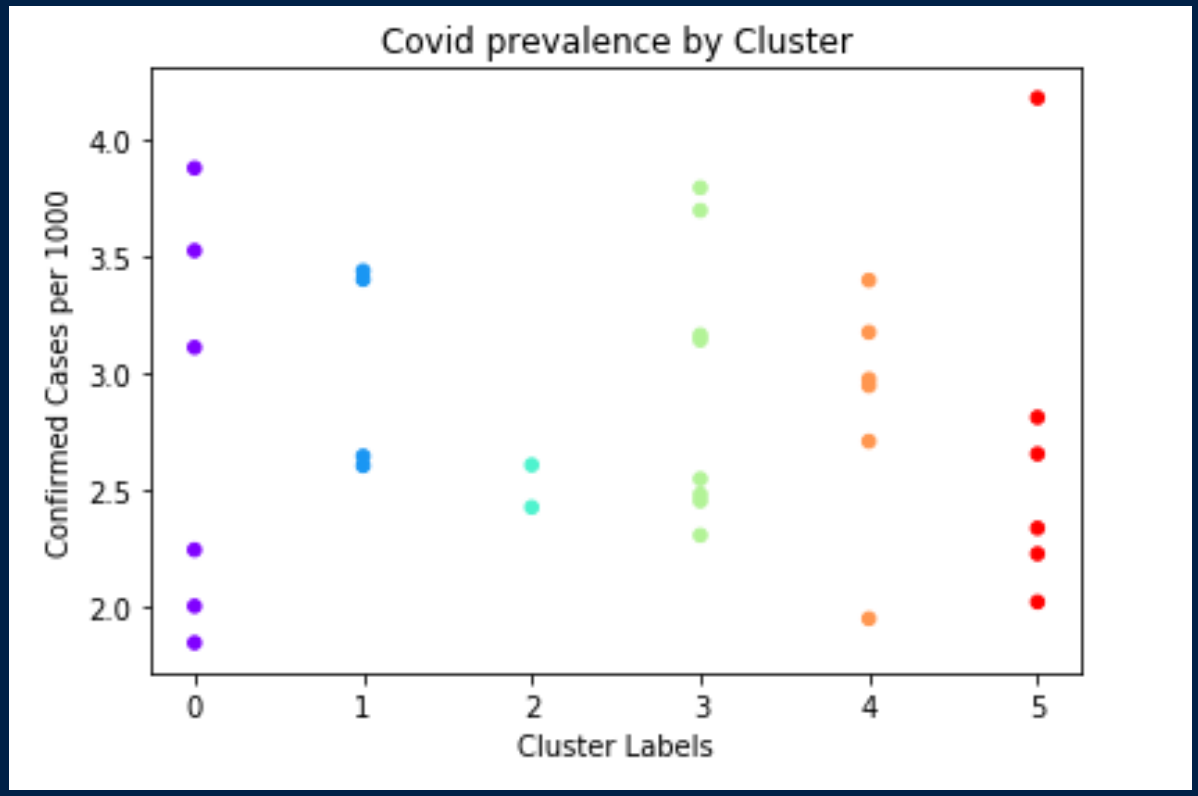
# Results: Demographics I

Elbow point method used to determine number of clusters (n=6).



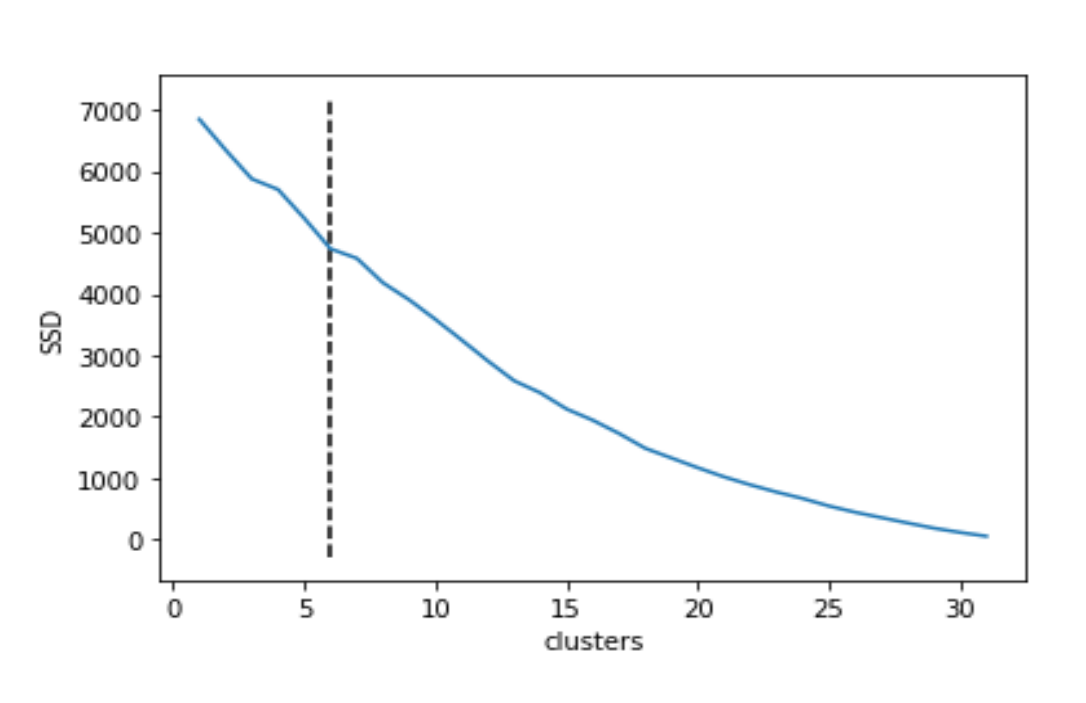

# Results: Demographics II

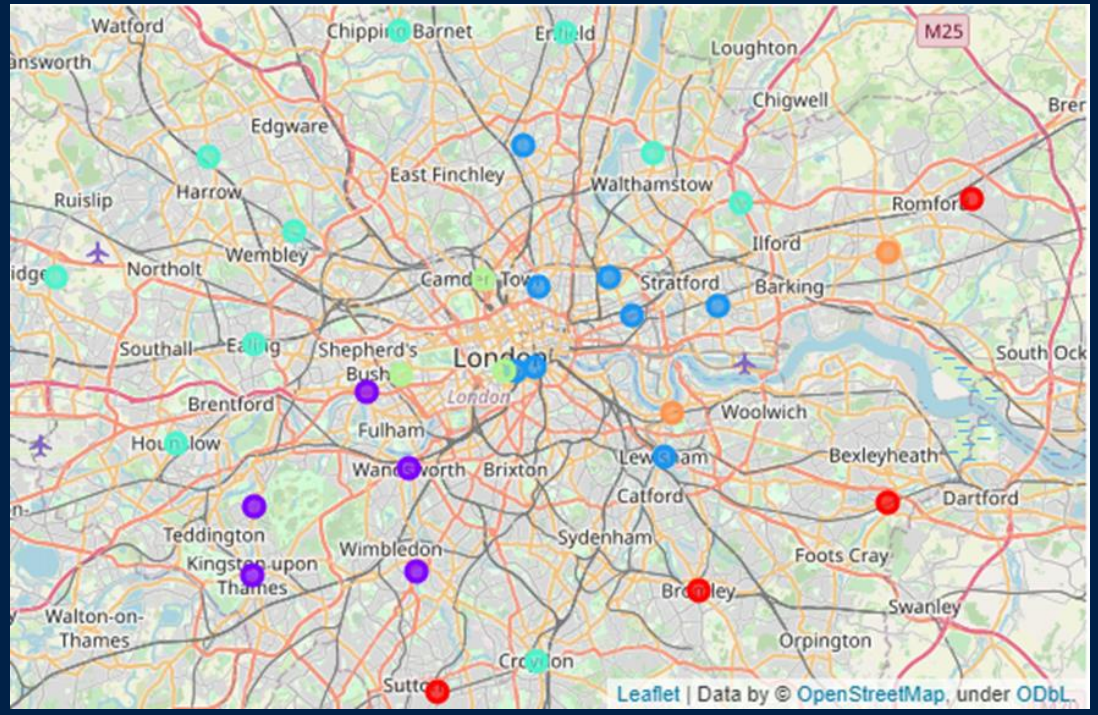
- High intra- and low inter-cluster variation.
- Sample size (n=32) is low.
- No statistically significant effect between demographic clusters.



# Results: Venues I

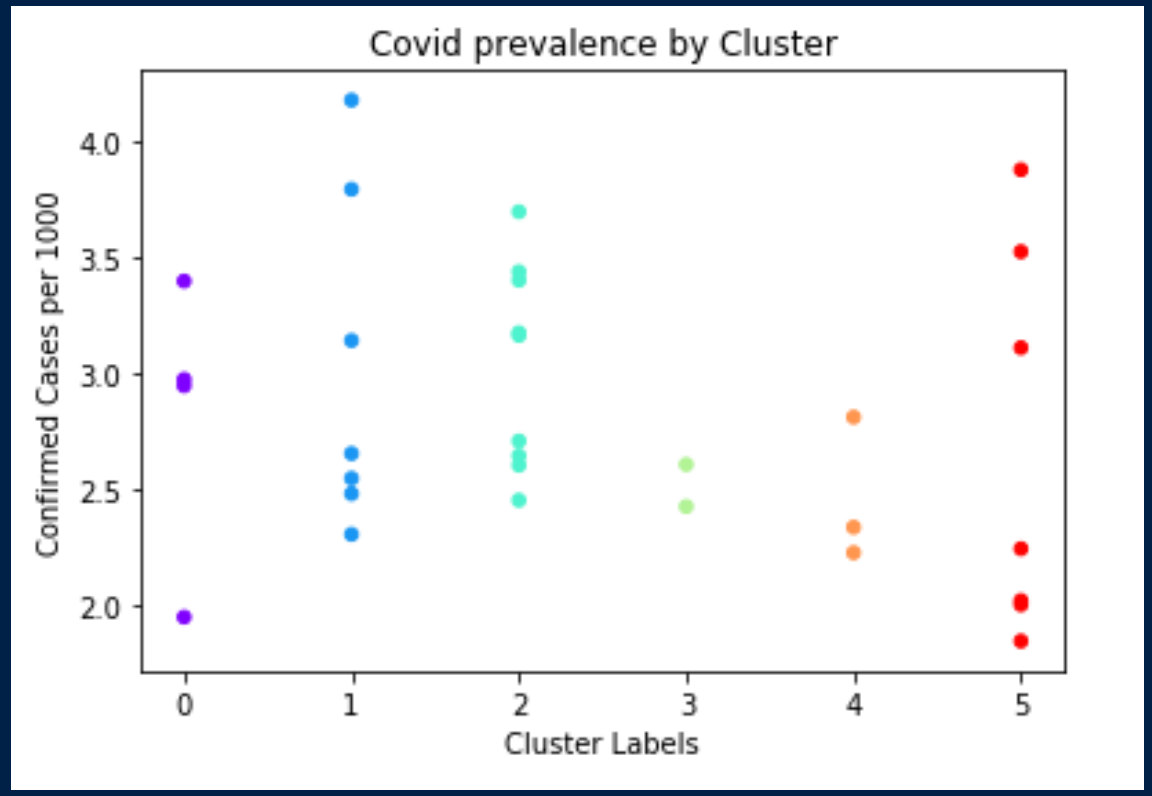
Highly homogeneous dataset, hard to identify elbow point (n=6).



# Results: Venues II

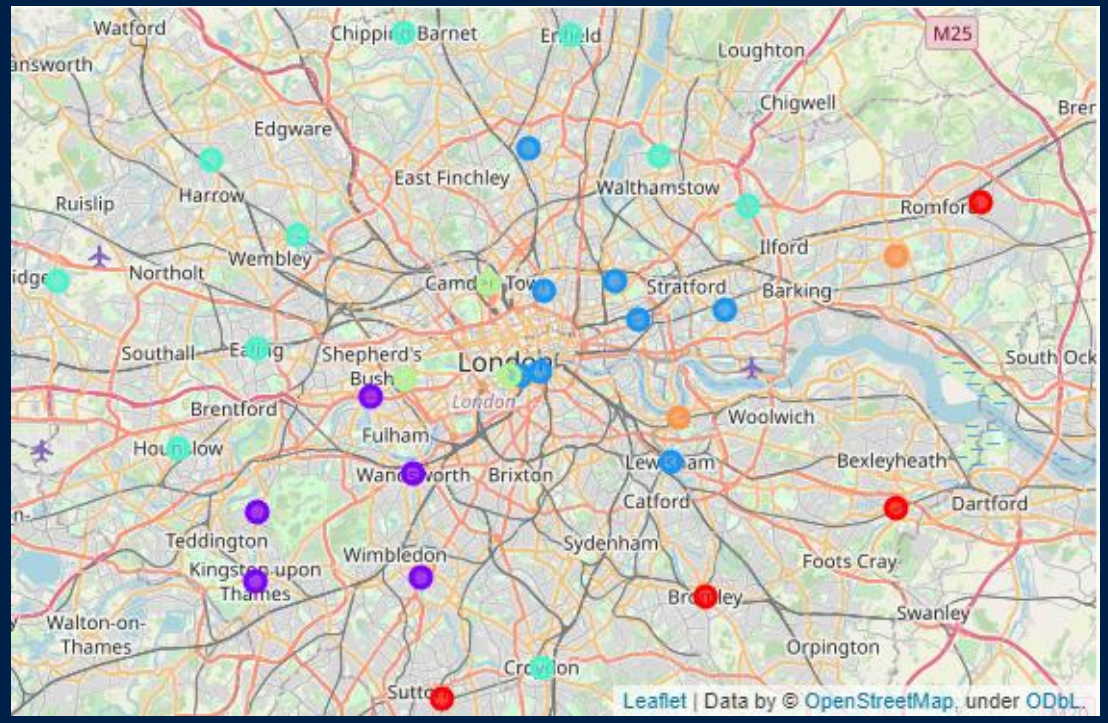
- High intra- and low inter-cluster variation.
- Sample size (n=32) is low.
- No statistically significant effect between demographic clusters.



# Results: Comparison

## Clustered by Demographics


## Clustered by Venues




Observation: demographic clusters resemble venue clusters.

# Discussion

- Neither demographics nor venue profiles appear to have a strong effect on COVID-19 prevalence in a borough.
- Low variation (homogeneity) and low size (n=32) of the datasets leads to low statistical significance.
- Clustering of geospatial neighbours results in interdependence of prevalence rates due to cross-transmission.
- Classification approaches such as decision trees, support vector machines, and logistic regression are also likely to struggle with the same limits.

# Conclusion

- Using demographics and venue profiles is of limited use to pre-empting COVID-19 outbreaks.
- Reactive rather than pre-emptive tactics should be employed when compiling a lockdown exit strategy.
- Fast response to outbreaks will be the key to avoiding a second surge.