Multivariate Analysis of London during the COVID-19 Pandemic

Capstone Project

Background

- The COVID-19 Pandemic has caused unprecedented challenges to retail businesses, who are now faced with uncertain government regulation and customer sales.
- London, with a population of nearly 10 million people, is largely heterogeneous, and hence a prime location for us to conduct data analysis on.
- We will attempt to do so by both cluster and factor analysis, to give a comprehensive description of the situation.

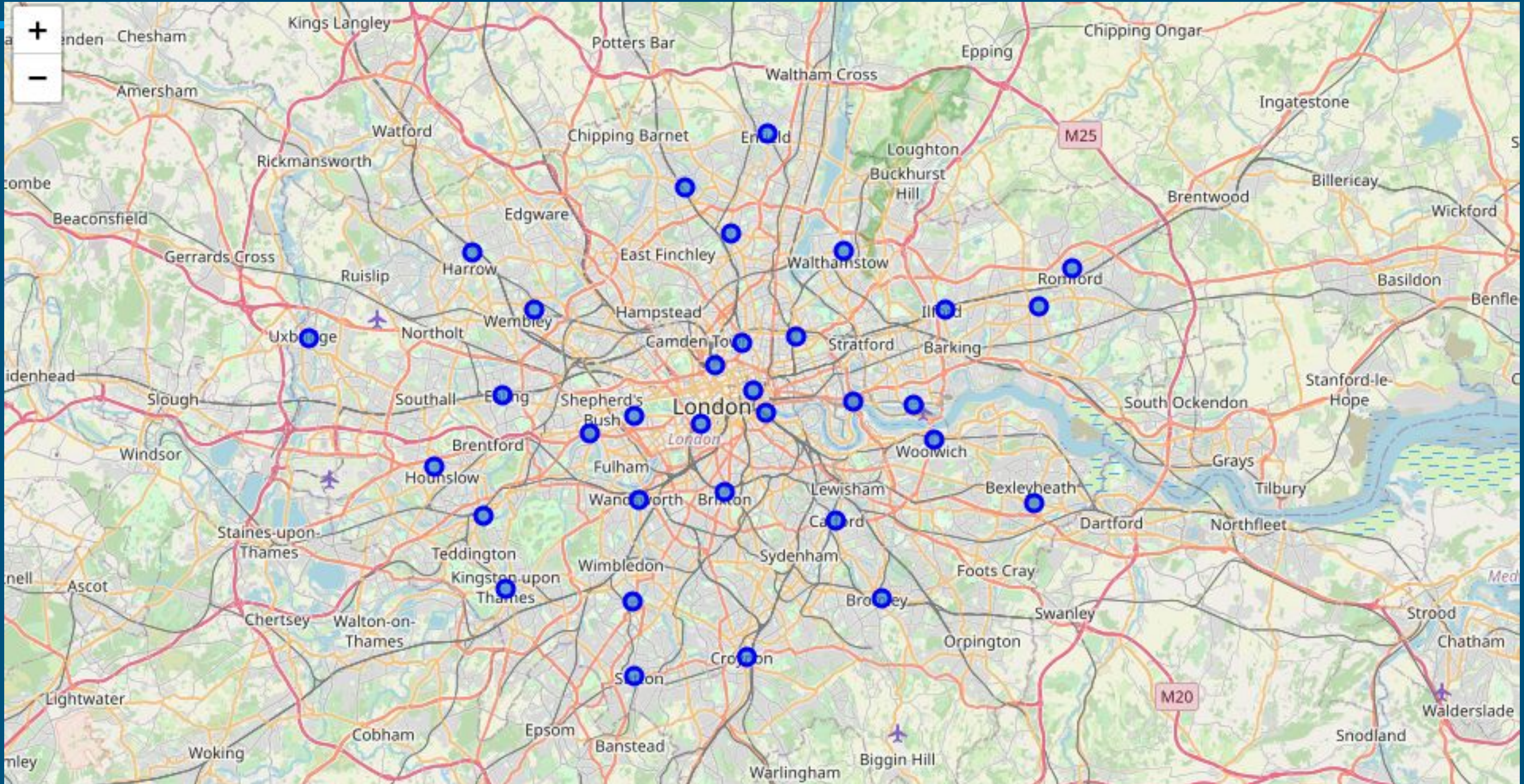
Data

Our final dataset will consists of the following variables :

- Identification data (name, code)
- Economic data (unemployment, income, etc)
- Socio-Demographic data (Proportion of youths, etc)
- COVID-19 Data (Incidence, death rate, etc)

Giving a final dataset size of 33 rows x 15 columns

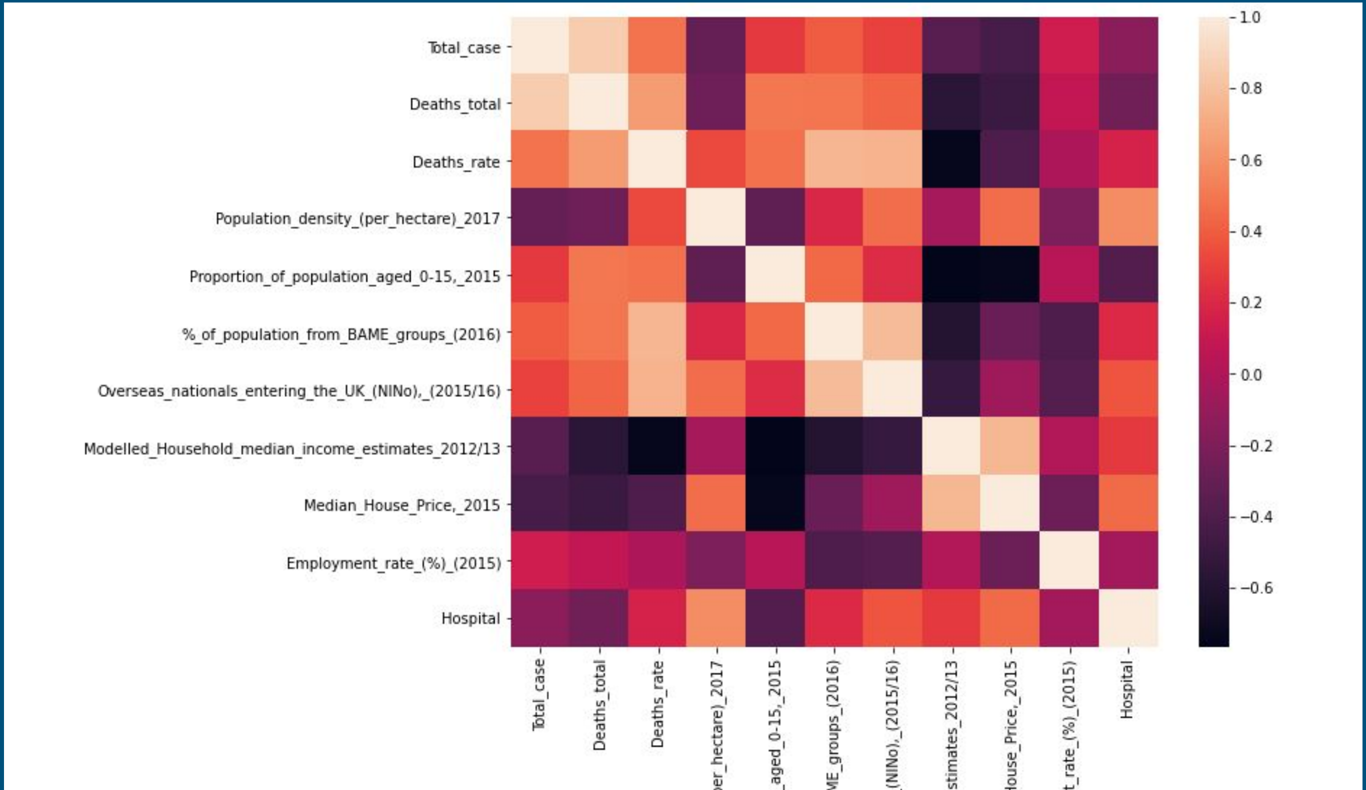
Map of London



Exploratory Data Analysis

- Overall, we identify the following sets of correlations. As expected, economic indicators, such as median house price and median household income, are strongly negatively correlated with the various COVID-19 indicators.
- Similarly, social indicators, such as proportion of youths are highly correlated with COVID-19 indicators. Interestingly, the population density seems to be negatively correlated with COVID-19 indicators.

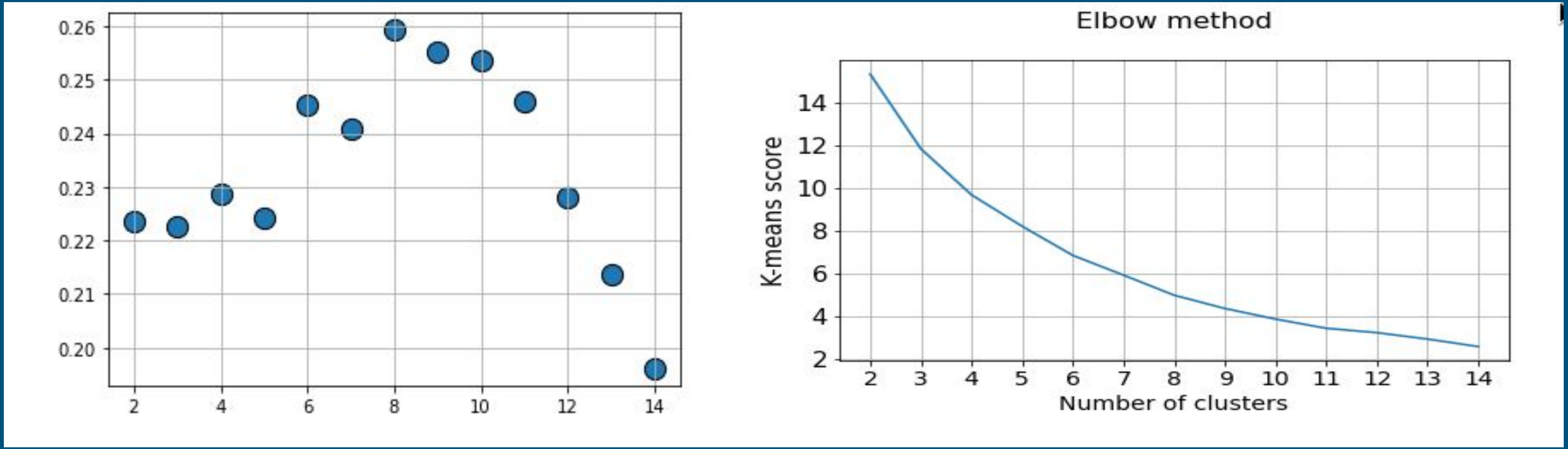
Correlation Matrix



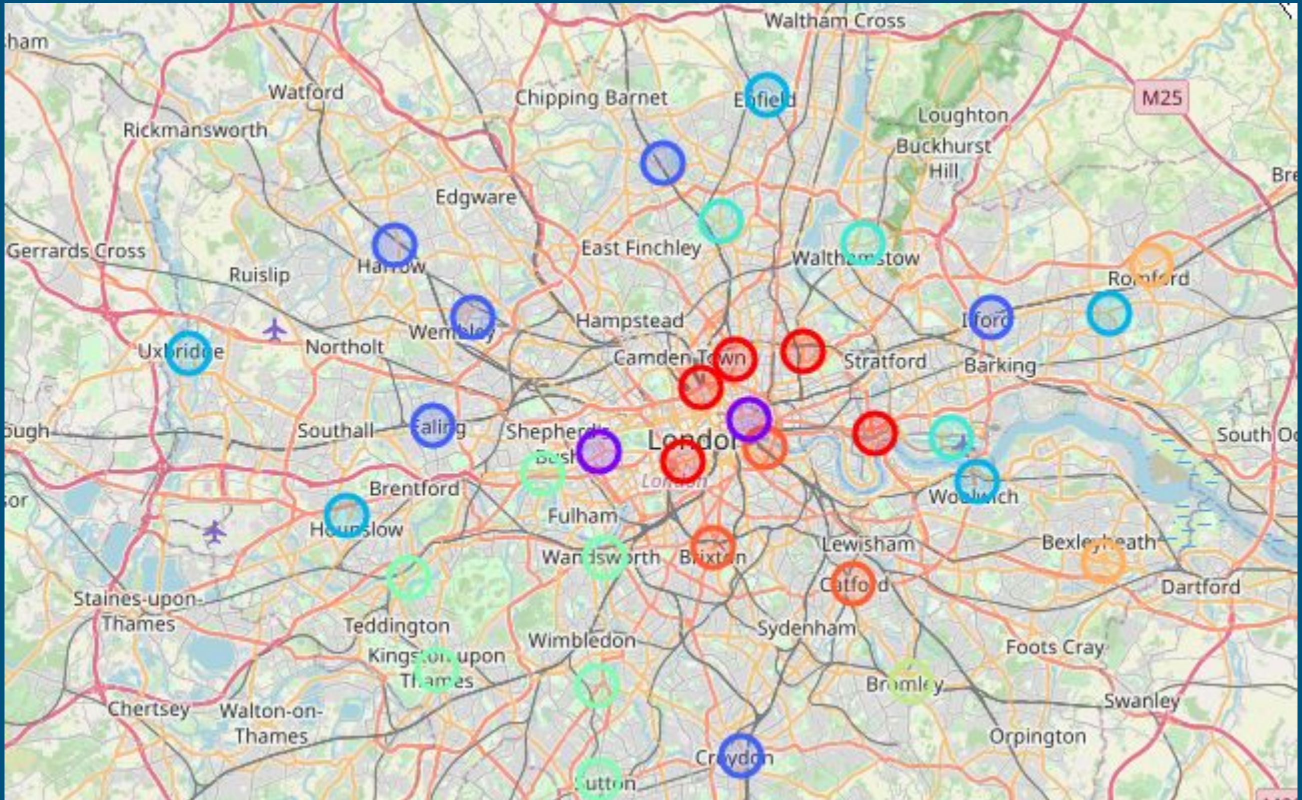
Cluster Analysis

- We carry out the k-means clustering algorithm. However, before that, we need to carry out the feature normalisation and the identification of the number of clusters to use.
- To identify the number of clusters, we use both the traditional elbow method and the Silhouette score.
- Quite clearly, the elbow method and the Silhouette coefficient both indicate that k=9 is the optimal number of cluster. While the elbow method is less obvious, as it seems like k=4 could be a candidate as well, the silhouette coefficient optimality is clear as day.

Scores for Cluster Analysis



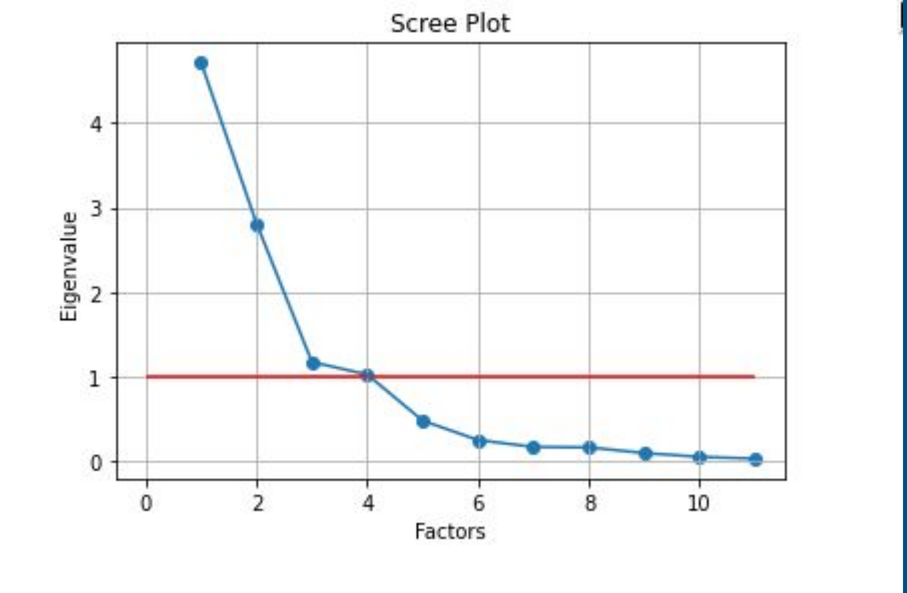
Classification of London Areas



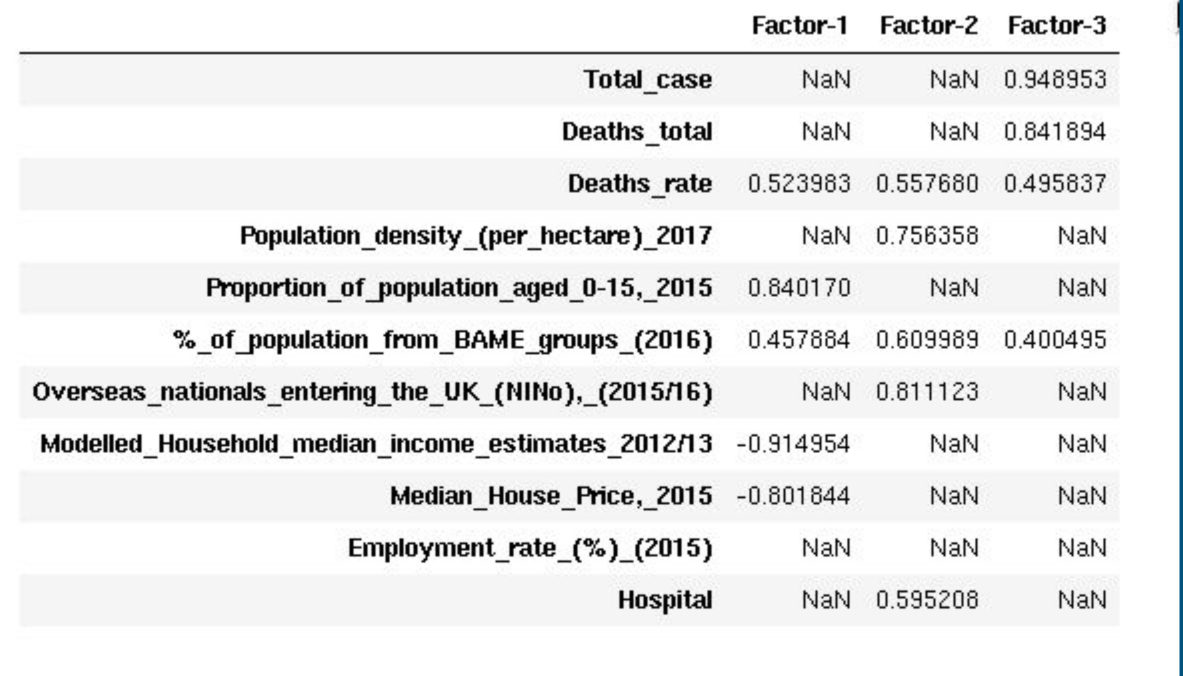
Factor Analysis

- Given the weak structure of our data in clustering, we move on to factor analysis to identify a possible set of underlying variables for the observed data. However, before that we carry out two test to ensure the suitability of the data.
- The KMO value at ~0.65 is fairly decent, while the p-value for Bartlett's test is highly significant. We move on to obtaining a scree plot to identify the number of factors to use.
- We use k=3 as shown in the Scree Plot, which confirms with our own theoretical intuition that the variables are broadly classified into such categories.

Factor Analysis Scree Plot



Factor Analysis Result



	Factor-1	Factor-2	Factor-3
Total_case	NaN	NaN	0.948953
Deaths_total	NaN	NaN	0.841894
Deaths_rate	0.523983	0.557680	0.495837
Population_density_(per_hectare)_2017	NaN	0.756358	NaN
Proportion_of_population_aged_0-15,_2015	0.840170	NaN	NaN
%_of_population_from_BAME_groups_(2016)	0.457884	0.609989	0.400495
Overseas_nationals_entering_the_UK_(NINo),_(2015/16)	NaN	0.811123	NaN
Modelled_Household_median_income_estimates_2012/13	-0.914954	NaN	NaN
Median_House_Price,_2015	-0.801844	NaN	NaN
Employment_rate_(%)_(2015)	NaN	NaN	NaN
Hospital	NaN	0.595208	NaN

Results

- Our cluster analysis results indicated to us that it was possible to classify the boroughs into 9 different clusters, which were, unsurprisingly, strongly spatially dependent as can be seen from our map. We expect that areas which are closer together would exhibit similar characteristics, and hence, be classified into similar clusters.
- Our factor analysis result indicated to us that variables could broadly be classified into 3 different factors, which further confirms our theoretical classification of the variables, and these are broadly, socio-economic, socio-demographic and COVID-19 related data.

Discussion and Conclusion

- For further analysis, we recommend the usage of more comprehensive socio-economic data, and the integration of temporal data, which would allow us to observe the evolution of these heterogeneity over time. Due to our limited resources, we were not able to carry out such analysis in this project.
- Our analysis provided both cluster and factor analysis on a broad dataset in London. Given our usage of current COVID-19 data, we hope this will support and inform businesses and policy-makers in their decisions.
- Thank you!