Multivariate Analysis of London during COVID-19 pandemic

Sherman Khoo

December 22, 2020

1 Introduction

In this project, we perform a multivariate analysis of a dataset consisting of socio-economic, geographical and COVID-19 data in the 33 areas of London (32 Boroughs adn the City of London). Thus, we attempt to provide a comprehensive description of the currently on-going COVID-19 epidemic taking into account the various differences between areas. Our analysis will consists primarily of cluster analysis and factor analysis.

1.1 Background

London, as a region of nearly 10 million people, contains a varied and heterogeneous population. In our dataset containing a broad mix of different socio-economic and demographics data, across the 33 areas of London, we hoped to capture, to a certain extent, the heterogeneity of the different areas, and thus be able to provide a description during this period of pandemic of firstly, the heterogeneity between the areas, and secondly, to extract the underlying factors that described these heterogeneities. These two issues were tackled with our cluster analysis and factor analysis respectively.

1.2 Business Problem

Because of the varying restrictions and rules of the lockdown regulations, multi-branch retail businesses are faced with the difficult issue of deciding on which of their branches to keep open or close. In their business decision, retailers are faced with multiple trade-offs, for example, COVID-19 incidence and socio-economic factors. We thus conduct cluster analysis to aid businesses in making such decisions. Furthermore, factor analysis is conducted to find underlying factors within the different variables in our dataset. Often, businesses are faced with multiple variables/metrics, but have very limited means of identifying which variable are truly relevant or to what extent multiple variables are

similar. This is especially salient now with COVID-19 forcing businesses to make significant changes with large amount of dynamic data in a short period of time. We hope that with factor analysis, we can identify underlying factors that will help businesses make better decisions.

2 Data

Our final dataset is a dataframe, with a size of 33 rows x 15 columns.

- Borough name and the statistical geography code as given by the ONS
- **COVID-19 Data** from a single period: total cases, total deaths and the death rate
- **Economic Data**: Modelled household median income estimates (2012/2013), median house price (2015), employment rate (2015)
- **Geographical, Socio-Demographic Data**: Coordinate of the borough centre (latitutde and longitude), number of hospitals within a 5000m radius from the city centre, Population density (per HA, 2017), proportion of population aged 0-15 (2017), percentage of population from BAME groups (2016), Overseas nationals entering the UK (2015/16)

2.1 Data source

As this is a self-compiled dataset, there are multiple sources for this data. Chiefly, this data is obtained from the Office of National Statistics (*ONS London Data*, N.d.), Foursquare API (*Foursquare API*, N.d.), and the Greater London Authority (*GLA COVID-19 Data*, N.d.).

2.2 Data Cleaning and Preparation

Minor data preparation and cleaning was carried out to ensure that the dataframe was in the right format and setting for our analysis. Furthermore, standardisation was carried out with a Min-Max Scalar to ensure proportional equivalence between each of the variables.

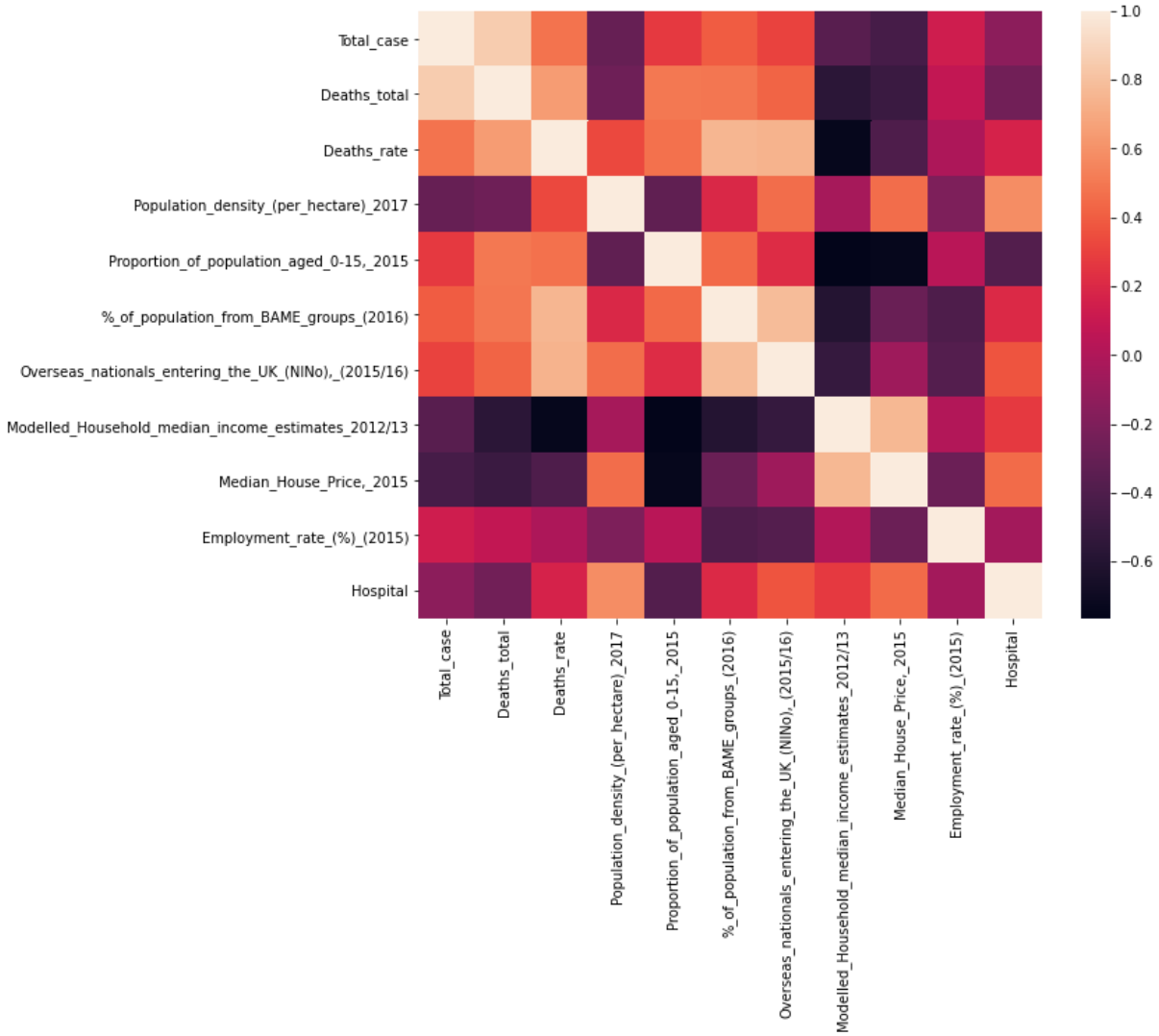
3 Methodology and Analysis

We carry out our analysis in three parts.
Firstly, we carry out exploratory data analysis, using a heatmap to look a possible correlations within the different variables in our dataset.

Secondly, a non-hierarchical k-means clustering algorithm is used to group the given group data points into predefined clusters. We have to decide on the number of clusters used in the division of data. We will use an elbow data chart and the silhouette score to aid us in a decision.
Finally, factor analysis will be carried out. Before this is done, we need to verify the sampling adequacy. We use both the Kaiser-Meyer-Olkin statistic and the Bartlett's test of sphericity.

3.1 Exploratory Data Analysis

To begin, due to the large number of variables in our dataset, we begin with a correlation matrix to identify dependence between the variables in our data.

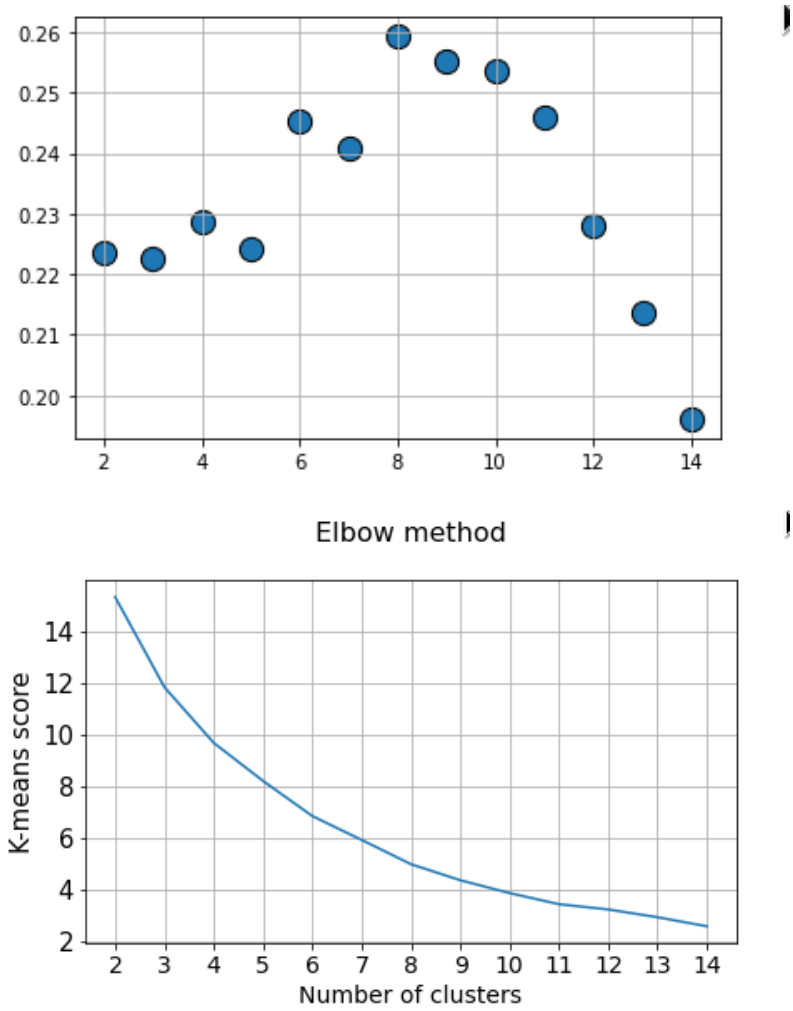


Overall, we identify the following sets of correlations. As expected, economic indicators, such as median house price and median household income, are strongly negatively correlated with the

various COVID-19 indicators. Hence, this correlation confirms our usual intuition that poorer regions are disproportionately affected by the pandemic.
Similarly, social indicators, such as proportion of youths, percentage of BAME and number of overseas nationals are highly correlated with COVID-19 indicators. Interestingly, the population density seems to be negatively correlated with COVID-19 indicators.

3.2 Cluster Analysis

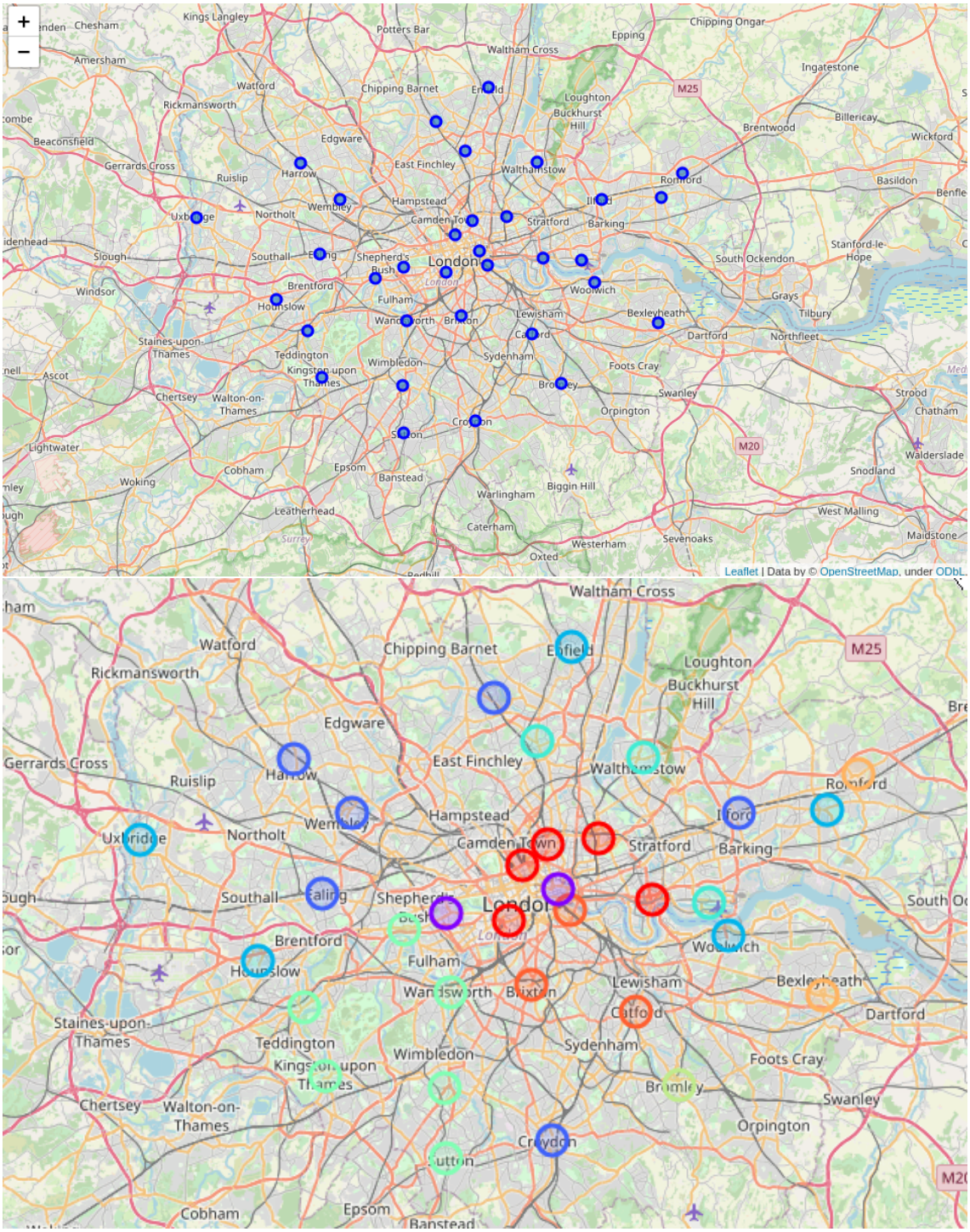
To begin our cluster analysis, we first looked at the silhouette score and the k-means score to help us to identify the optimal number of clusters.



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, with a significant peak at k=8 and k=9.

Given the low silhouette coefficient, this indicates a possible weak structure in the data, thus motivating our usage of factor analysis in the next section.

We continue on to carry out the algorithm with k=9. The results are shown below, firstly with a map of the boroughs of London, and then, a color coded map, with each distinct color representing a cluster. To begin, we note the strong spatial dependence of the clustering, which is in line with our intuition that geographically similar regions exhibit similar characteristics.

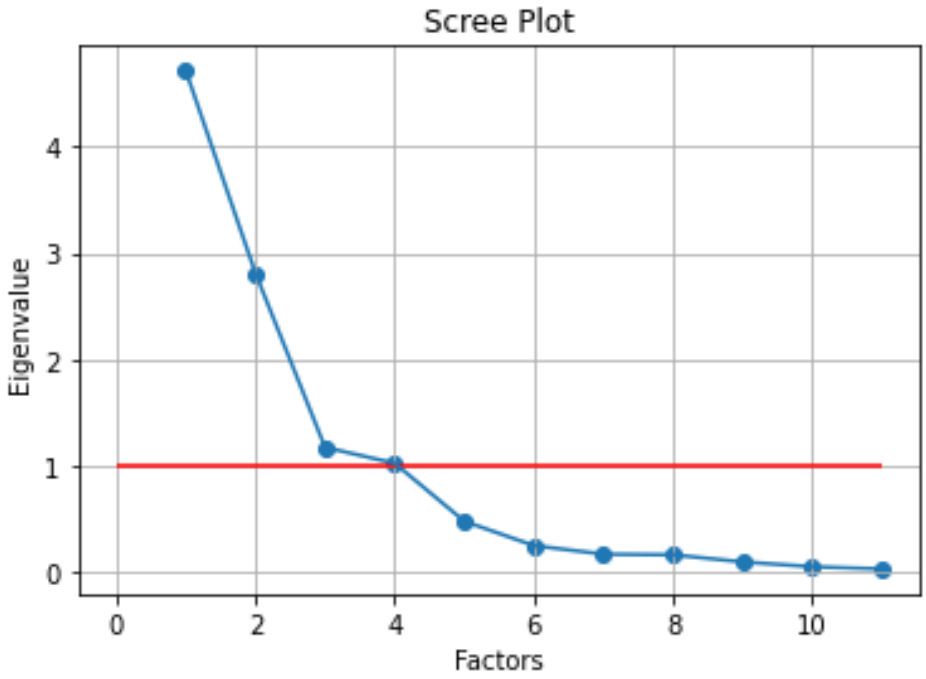


3.3 Factor Analysis

Given the weak structure of our data in clustering, we move on to factor analysis to identify a possible set of underlying factors for the observed data. However, before that we carry out two test to ensure the suitability of the data.

Bartlett's test of sphericity is carried out on our normalised dataset, and we obtain a highly significant result. Similarly, the KMO score is calculated, and at 0.68, we conclude our data is fairly robust and is suitable for further factor analysis.

We move on with a scree plot to identify the number of factors to use in our analysis.



The usual cut-off used is to identify factors with an eigenvalues of greater than 1, as this indicates that the factor accounts for at least as much variance as a single variable. The factors at k=3 also confirms with our own theoretical intuition that the variables are broadly classified into such categories.

Next, we carry out the analysis, using the default method of MINRES (similar to the standard OLS method for factor extraction) in the factor_analyzer package in Python and we specify to use the common varimax rotation.

Thus, we classify the 3 factors based on the variables identified within each factor. Broadly, they are: Economic, Demographic, and COVID-19 factors. Finally, we check the variance explained from the 3 factors. Altogether, they explain 73% of the variance, which is a good portion.

	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

	0	1	2
Variance	3.095470	2.572976	2.370473
Proportional Var	0.281406	0.233907	0.215498
Cumulative Var	0.281406	0.515313	0.730811

4 Results and Discussion

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, economic, demographic and COVID-19 related data.

For businesses, our result provides a general guideline on a way to distinguish the London areas, in a data-driven and robust manner. This would, for example, allow retail businesses to make better informed decisions on where to open or close their businesses. Our factor analysis also provides a concrete method to compare and identify the variety of metrics available to businesses to allow

them to better understand the demographics of the area, which we hope will better aid them in integrating these metrics into their decision making process.
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.

5 Conclusion

Our analysis provided both cluster and factor analysis on a broad dataset in London. While our result for cluster analysis indicated weak structure, our result for factor analysis was fairly robust, and we wish that our result could be taken as an indicative guide to better support and inform businesses in their decision making process during the greatly uncertain times of the COVID-19 pandemic.

References

Foursquare API. N.d. https://developer.foursquare.com/developer/. Accessed: 2020-12-22.

GLA COVID-19 Data. N.d. https://data.london.gov.uk/dataset/coronavirus–covid-19–cases. Accessed: 2020-12-22.

ONS London Data. N.d. https://www.ons.gov.uk/, note = Accessed: 2020-12-22.