Overall Limitations on Methodology

1. The different nature of data in our datasets

The decision to normalise our data comes from the different nature of the figures in all of our datasets, as our study includes very different data, from gas emissions to population and GCSE results as well as data between 2000 and 2019 depending on its public availability (we understand the importance of analysing data during the Covid-19 pandemic, however most data after or during 2020 is not yet public or available). Therefore, in order to scale all our scores we applied the feature scaling normalisation formula (Google Developers, 2021). This technique doesn’t affect the relationship and proportions within the boroughs that we studied, which is the main focus of our project, but provides accurate scores within a controlled range. However, the normalisation of our data is not exempt of limitations:

* Outliers in our data affect every single datapoint, since, by compressing all the values within the range marked by the normalisation formula, all values far from the outlier will be grouped at one end of the range. Moreover, visualisations with this data could be considered misleading, as, for example, in the variable gas emissions, some boroughs are awarded the score 0, but they still have a very high score of gas emissions, not significantly lower from the borough awarded with the 1.
* We encountered another type of limitation, specifically relating to the borough of City of London, since, due to the City of London’s small population size, values for Hackney and the City of London are sometimes combined (Public Health England, 2017). While we tried to find the most adequate and complete datasets for each variable, for the variables of noise pollution and healthy life expectancy, we were unable to find a value for City of London. For each borough, we tried to estimate this value using other available data:
  + For the variable of noise pollution, we created our own scores for Hackney and City by splitting the complaints proportionally. This means that neither the City nor the Hackney score is valuable compared within each other, but they are significant when compared to the other boroughs.
  + For the variable of healthy life expectancy, we awarded City of London the lowest score, considering it being the healthiest borough and the borough with the highest life expectancy. This presents a clear limitation, as City of London is one of the key boroughs of this study, being the borough with the BSe and the lowest BSh.

1. The process of calculating each borough score

Other limitations included calculating each borough score, what factors to include and how the equation should look like, since we initially intended to use more socio-economic and environmental factors but realised that in the limited time we have to complete the project that might not be attainable.

* For example, in calculating socio-economic factors, we included a margin of error in our calculations (ε) because of the possibility of errors included in some of our data sets, for example the case for Cost of Living (CL).
* For the Cost of Living (CL) the data from Numbeo (2022) was crowd-sourced and unavailable for half of the London boroughs. Moreover, for the missing data we used a method similar to the Hot-Deck imputation to calculate the missing values by using the average mean for each variable of the neighbouring boroughs of the borough whose data was missing. Hot Deck imputation is a statistical method for missing data processing where compromised or non-existent data is replaced with observed responses from “similar” units (Andridge and Little, 2010). The core of the method is to replace the missing values with plausible data for further analysis not to generate fully accurate missing data replacement (Little, Rubin and Bayes, 2002).
* All initial Income (IN), Education (ED), and Race & Migration (RM) formulas had different equations which had to be replaced since most included a division that after data normalisation gave always one variable that had to be divided by zero and hence would not have had a quantifiable number as a result.
* For example, here are the initial equations for the Income and Education scores and then the equation we ended up using.

*Initial Income equation*:

IN = (E - CL) ÷ Unemployment Rate + ε

Where IN represents Income, E is Earnings per Head of boroughs, CL is Cost of Living, and Unemployment Rate does not have an shortened variable name since we ended up using Employment Rate (ER) instead because it worked better with our new equation.

*Final Income Equation*:

IN = E + MP + ER - CL + ε

Where IN represents Income, E is Earnings per Head of boroughs, CL is Cost of Living, and ER is Employment Rate.

*Initial Education equation:*

ED = (GCSE score + A-level score of public schools) ÷ no. of public schools

+ (GCSE score + A-level score of independent schools) ÷ no. of independent schools + ε

, where ED is Education score.

However, since we couldn't find any data regarding the exams attainment in terms of public or independent schools and we needed to avoid the division by zero we arrived to the final Education equation:

*Final Education Equation*:

ED = GCSE + ALVL + TS + ε

Where ED is Education, GCSE is score of GCSE attainment per borough, ALVL is score of A-level attainment per borough, TS is number of total schools (public and independent) in each London borough, and ε is a margin of error which will not be counted numerically towards the final calculations.

1. Limitations on MLR:

To improve the rigour of adopting multiple linear regression, it is suggested that we test the correlation between each two independent variables before conducting MLR. If there are two independent variables strongly correlated with one another, it will influence the accuracy of the final statistical results.