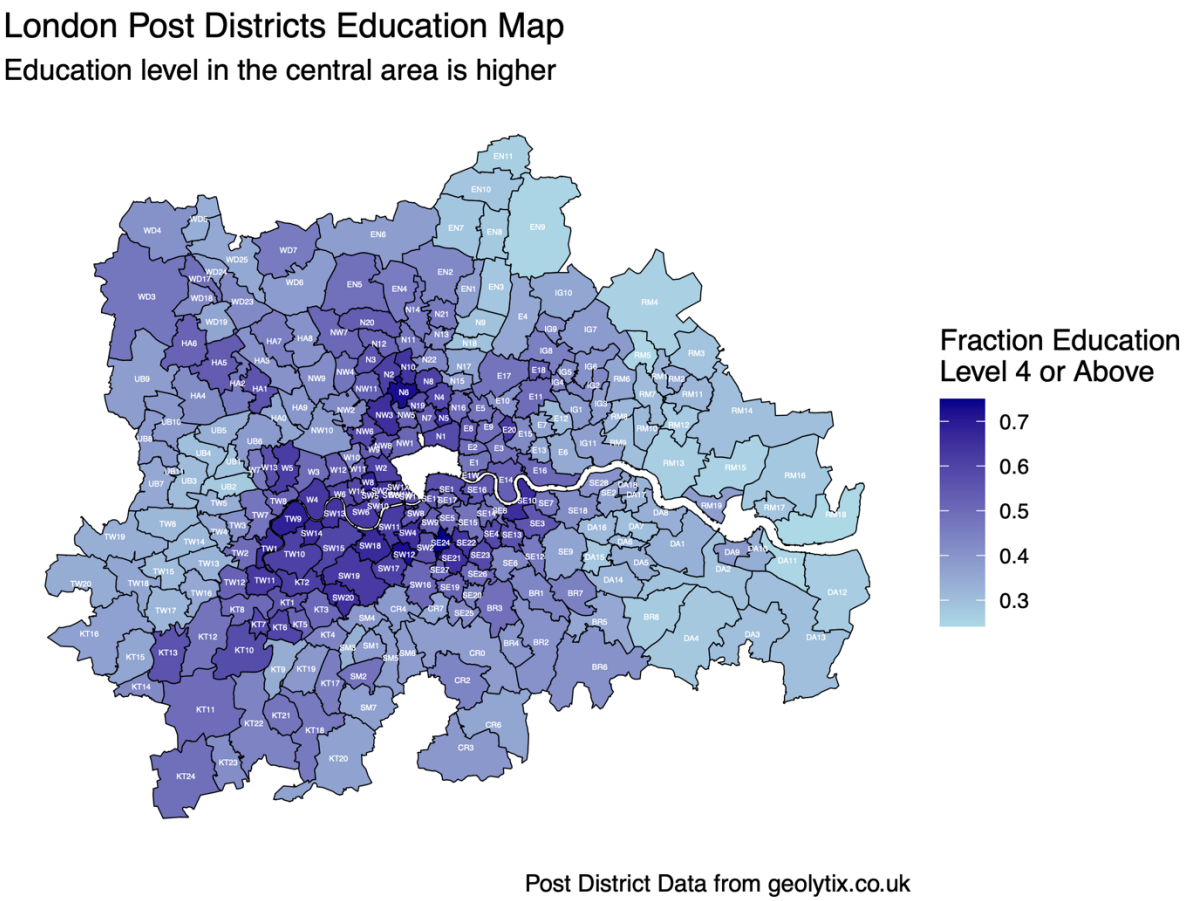
ECON0128 Coursework (Siyuan-Harry)

1 Introduction

Urban crime rates are a pressing concern for policymakers and researchers worldwide, as they significantly impact social stability and public safety. Traditional crime analysis methods often rely on aggregated data that may not capture the spatial distribution and underlying factors affecting crime rates at finer geographic levels. However, Fatehkia, O'Brien and Weber (2019) proposed a new methodology, utilizing residents' interest data to predict crime rates in different zip codes areas (neighborhoods), which showcases a feasible way to predict crime rate by integrating geographic data.

London, as one of the most densely populated cities in the world, faces unique challenges in managing crime and addressing the needs of its residents. Some scholars have revealed the relationship between dropout, employment exclusion and participation in crime through interviews (Spencer and Scott, 2013). Inspired by these studies, I wonder if we can also find the correlation between crime and education and employment status in a wider geographical distribution and verify the consistency with their conclusions. I firstly tried to visualize a map of the distribution of education in London and wondered about the distribution of crime rates and whether there was an intrinsic relationship between them.



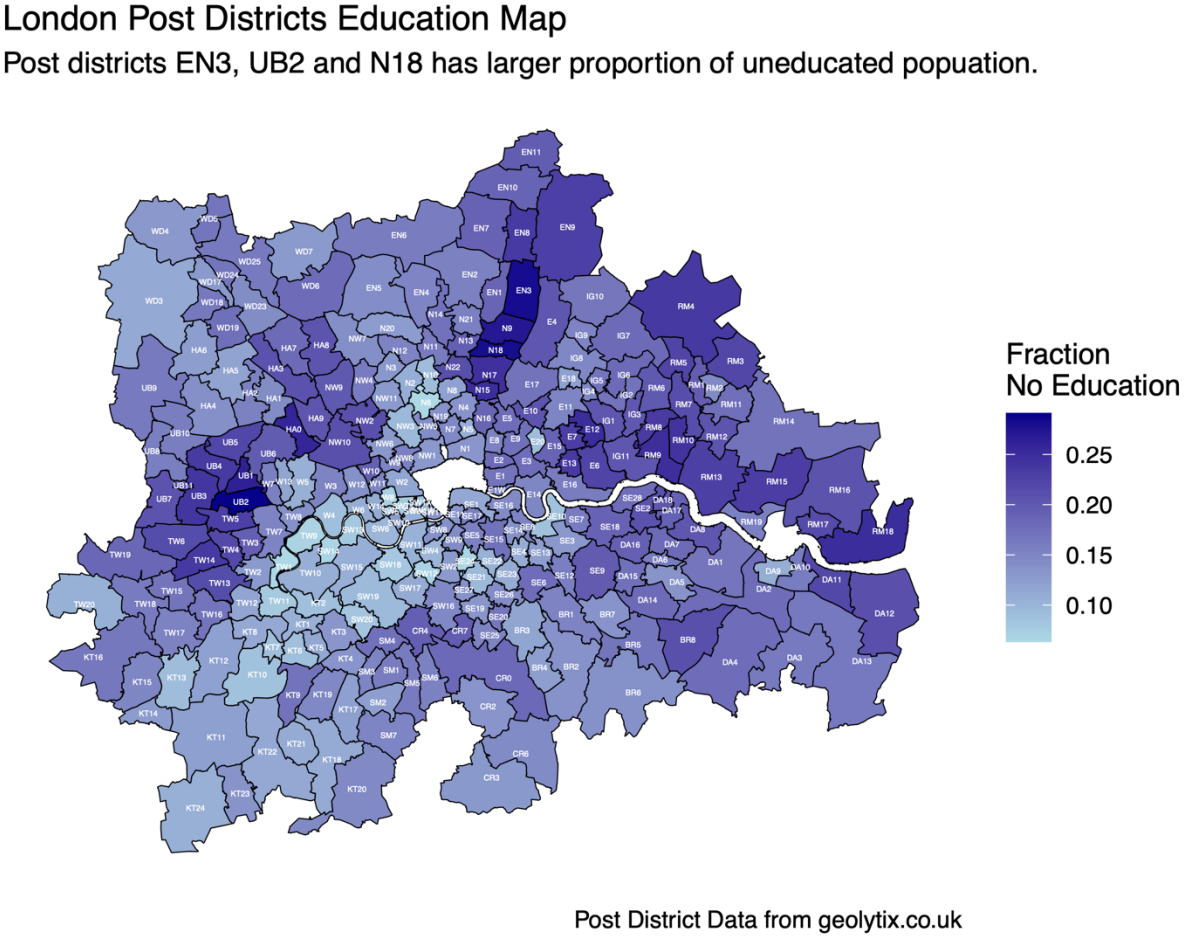


Figure 1: London Post District Map of Educational Status

1.1 Hypothesis

According to conclusion of Spencer and Scott (2013), I assume that low educational level area might have higher crime rate. To be more specific: (1) *the larger the fraction of population owns high-tiered education qualifications, the lower the crime rate in the district* and (2) *the fewer poor-educated residents should indicate a lower crime rate*.

2 Data

2.1 Data collection

The data used in this project is mainly selected from four different sources. The data from the first source is of ShapeFile type (.shp), which contains London's geographical information with respect to each of its post district (e.g. WC1 is the post district of UCL). This ShapeFile data comes from Grolytix's London Postal 2024 product.

The second data source is data.police.uk, from which all crime data (.csv format) from September 2023 to August 2024 is downloaded. This data contributes to the statistics of crime rate of each London's LSOAs (Lower layer Super Output Areas, they comprise between 400 and 1,200 households and have a usually resident population between 1,000 and 3,000 persons).

The third data source is the UK 2021 Census bulk data, which provides this research descriptive demographic information (e.g. education level, employment level, unemployment rate) for each LSOA geographical area, to evaluate the relationship of them to the crime rate.

The fourth data is downloaded from geoportal.statistics.gov.uk, which provides the connection between LSOAs and Post Districts. This connection is crucial because one Post District could include dozens of LSOAs, making the Post District-level data more suitable for visualisation and LSOAs-level data more suitable for constructing model. The connection this dataset provides allows this study to be conducted on both scales simultaneously.

2.2 Data Preprocessing

When engineering the crime rate data, an algorithm is designed to traverse the monthly crime report data and sum them at two levels and matched by LSOA region into the appropriate position, resulting in a data frame called *yearly_crime_london_by_lsoa*. This data table is then summed by each row to get the total number of crimes committed in each London LSOAs for the year from September 2023 to August 2024. This number is then calculated together with the total residents in each LSOAs from the 2021 UK Census to yield the crime rate per 1000 capita. The method of calculating crime rate follows UK crime stats official method:

$$\text{Crime Rate}_{\text{per 1000 capita}} = \frac{\text{Total Crime of The Year}}{\text{Total Residents}/1000}$$

The other demographic variables (predictors) are mainly calculated in the following way:

$$\text{Fraction of residents with status X} = \frac{\text{Estimated residents with status X}}{\text{Estimated users in the geographical area}}$$

Finally, the calculated crime rate data, high level education population data, low level education population data, unemployment data, etc. are combined at the LSOAs level to yield the *Education_Crime_Employment_LSOA* data frame, which has 32 fields and 5,897 records. This data frame is obtained for data modeling, regression and visualization.

In the visualization process, the data set is grouped by the Post District variable, and some indicators (such as proportion and crime rate) are recalculated. The *sf* package is used to represent the geographical features of the data, to obtain visual data that can be plotted with *ggplot2*.

2.3 Variable Selection

I picked several related variables: *crime_rate* stands for the crime rate, *frac_no_edu* stands for the fraction of people in the district that don't have any educational qualifications, *frac_edu_lv4* stands for the fraction of people in the district that have at least level 4 educational qualification (e.g. certificate of higher education), *frac_ne_nw_12m* stands for the fraction of people that have not been employed or worked for last 12 months, and lastly, *frac_une_ex_student* stands for the fraction of people aged over 16 and currently unemployed, excluding students. These variables are related to the hypothesis of this research.

Summary Statistics of Important Variables

Statistic	N	Mean	St. Dev.	Min	Max
crime_rate	5,897	131.950	245.054	0.000	8,370.075
frac_no_edu	5,897	0.161	0.063	0.015	0.400
frac_edu_lv4	5,897	0.452	0.139	0.174	0.872
frac_ne_nw_12m	5,897	0.199	0.056	0.040	0.458
frac_une_ex_student	5,897	0.039	0.014	0.004	0.117

Table 1: Summary Statistics of Important Variables

During the process of selecting variables, it is found that the *crime_rate* variable is extremely skewed. Thus, a log transformation has been made to it, and the distribution of some relevant variables are shown in Figure 3.

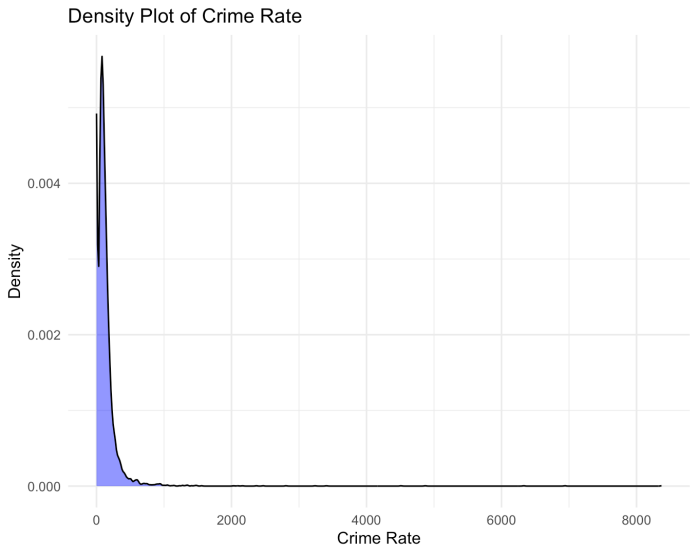


Figure 2: The *crime_rate* is variable is skewed to the right

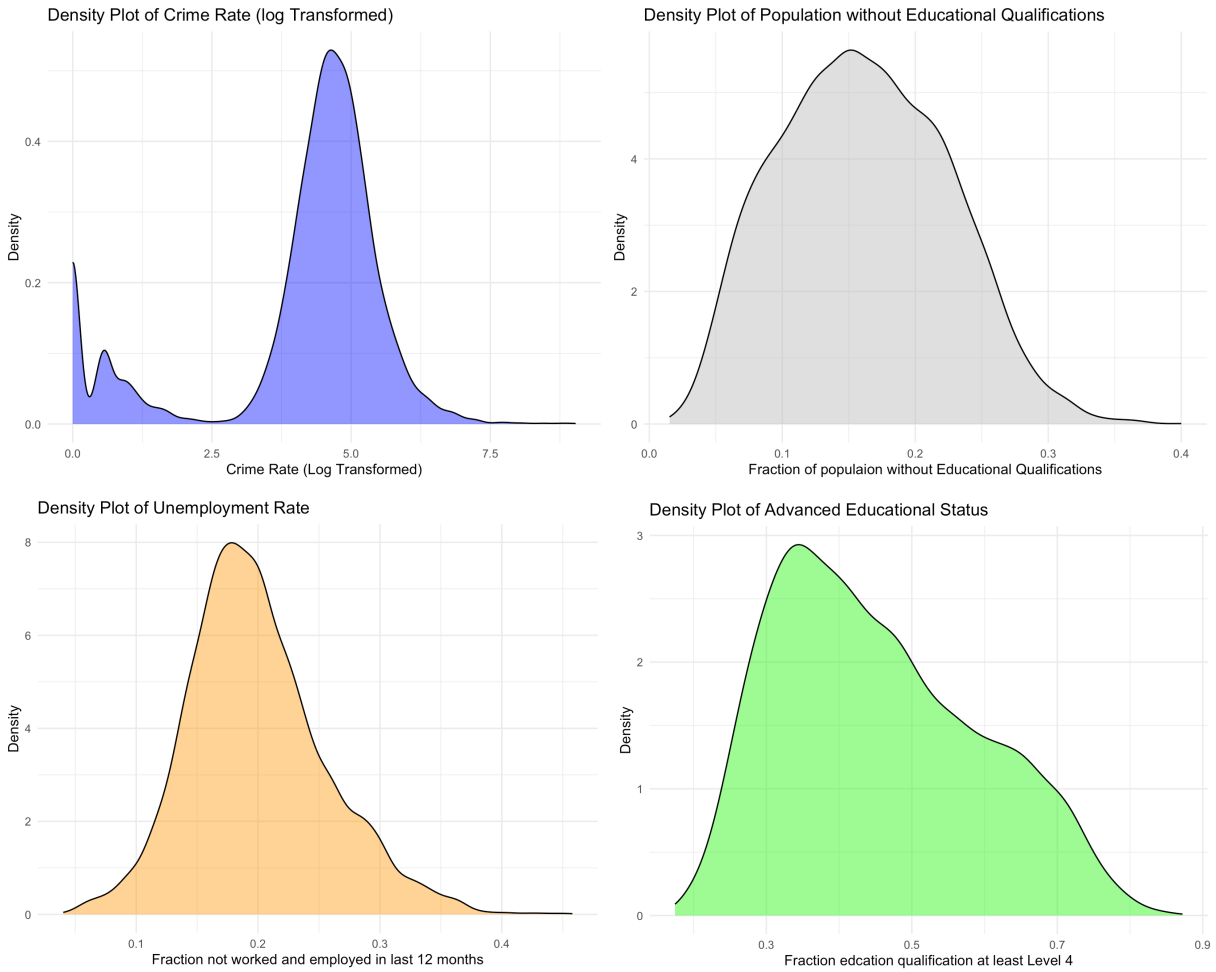


Figure 3: Density Plot of important variables

4 Result

4.1 Model

The first model (1) is relatively simple. It aims at modelling the linear relationship between educational status and crime rate. The F-Statistics is very significant, meaning this model is overall statistically significant. coefficient of *frac_no_edu* variable is 23.1425, which means that the proportion of uneducated population has a strong positive effect on the crime rate and is a major driver of crime.

$$\log(\text{crime\_rate}) = \beta_0 + \beta_1 \cdot \text{frac\_no\_edu} + \beta_2 \cdot \text{frac\_edu\_lv4} + \epsilon$$

Regression Results: lmmodel

	Dependent variable:
	crime_rate_log
frac_no_edu	23.143***
	(0.553)
frac_edu_lv4	11.895***
	(0.251)
Constant	−5.027***
	(0.196)
Observations	5,897
R^2	0.277
Adjusted R^2	0.277
Residual Std. Error	1.435 (df = 5894)
F Statistic	1,129.290*** (df = 2; 5894)
Note:	*p<0.05; **p<0.01; ***p<0.001

Table 2: Model 1's Regression Results

By introducing two new variables (*frac_une_ex_student* and *frac_ne_nw_12m*) and their interactions, combined with the previous education variables (*frac_no_edu* and *frac_edu_lv4*), I constructed model (2). The residual standard error (RSE) of this new model is 1.330, lower than previous model, meaning the error is reduced.

$$\begin{aligned}\log(\text{crime\_rate}) = \beta_0 &+ \beta_1 \cdot \text{frac\_no\_edu} + \beta_2 \cdot \text{frac\_edu\_lv4} \\ &+ \beta_3 \cdot (\text{frac\_no\_edu} \times \text{frac\_edu\_lv4}) + \beta_4 \cdot \text{frac\_une\_ex\_student} \\ &+ \beta_5 \cdot \text{frac\_ne\_nw\_12m} + \beta_6 \cdot (\text{frac\_une\_ex\_student} \times \text{frac\_ne\_nw\_12m}) \\ &+ \epsilon\end{aligned}$$

Moreover, the model now explains about 37.9% of the crime logarithm variance, a significant increase from the previous model (27.7%). This indicates that unemployment related variables significantly increase the explanatory power of the model. It is also noticeable that the interaction effect of *frac_une_ex_student* and *frac_ne_nw_12m* (232.36) is very strong, suggesting that the driving effect of lack of economic activity (whether unemployment or prolonged absence from work) on crime may be amplified under certain conditions.

Regression results: comparison between two models

	Dependent variable:	
	crime_rate_log	
	(1)	(2)
frac_une_ex_student		−15.991**
		(5.033)
frac_ne_nw_12m		−11.118***
		(0.855)
frac_no_edu	23.143***	5.995***
	(0.553)	(0.947)
frac_edu_lv4	11.895***	6.219***
	(0.251)	(0.344)
frac_une_ex_student:frac_ne_nw_12m		232.364***
		(24.619)
frac_no_edu:frac_edu_lv4		22.605***
		(2.175)
Constant	−5.027***	−0.049
	(0.196)	(0.318)
Observations	5,897	5,897
R^2	0.277	0.379
Adjusted R^2	0.277	0.378
Residual Std. Error	1.435 (df = 5894)	1.330 (df = 5890)
F Statistic	1,129.290*** (df = 2; 5894)	599.122*** (df = 6; 5890)
Note:		*p<0.05; **p<0.01; ***p<0.001

Table 3: Comparison between two models

4.2 Predictions and visualization

Utilizing the model (2), I made a prediction of crime rate for every LSOAs, sum them up and visualized them at the Post District level. The model correctly captures the changing crime trends within geographic areas, but because of the relatively simple structure of the model, the specific values predicted by it are not quite precise. In summary, the current predictive effect of the model shows that, in a general sense, we can model the crime rate of a region by its basic education level and unemployment rate. Further engineering the data features could lead to a better prediction outcome.

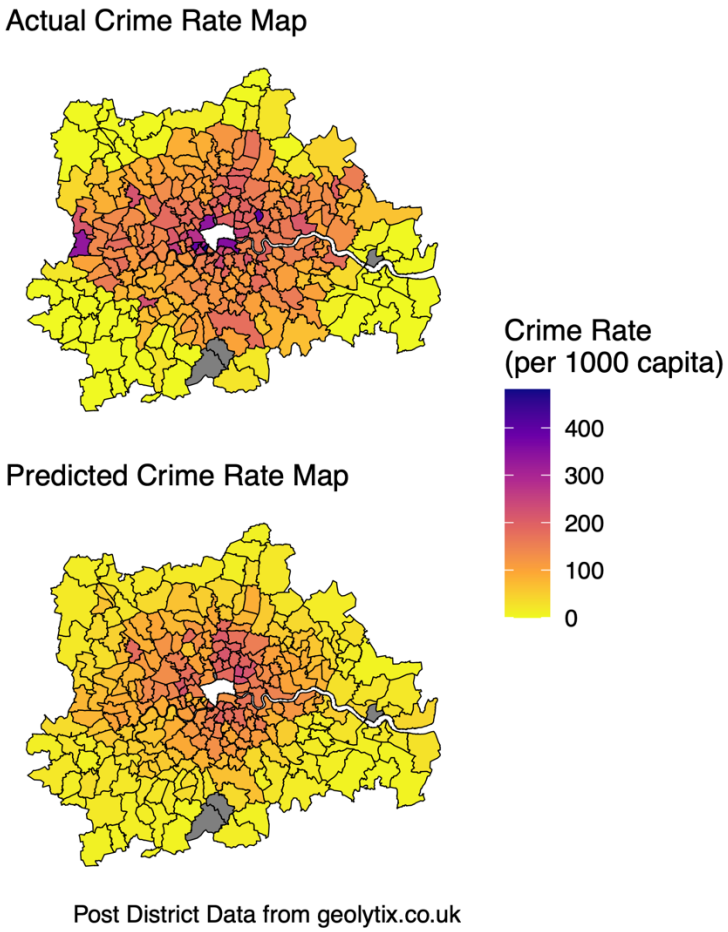


Figure 4: Comparison between predicted and actual crime rate map

5 Limitations

The first limitation is about this research is the time and space mismatch of the data. I assume the education and employment state 2021-2024 is relatively stable, and the crime data of 2024 was merged with 2021 census data. Furthermore, in visualisations, only the relatively outer parts of London were visualized, excluding the central and surrounding business districts of London. This is because the distribution of post districts in central London is too dense to visualize directly in the same scale with other parts.

Another limitation of this research is that, it had not explore deeper into several interesting phenomena occurred in the process. For instance, at the result of both model (1) and (2), we can clearly observe that the coefficient of *frac_edu_lv4* is positive. This is a bit counter-intuitive and challenged my hypothesis. It may be worth exploring why do areas with higher rates of higher education also have higher crime rates? Is it related to urbanization and the high concentration of economic activities? Also, in the model (2), the coefficient of interactive term *frac_no_edu:frac_edu_lv4* is positive – does it indicates that areas with highly differentiated levels of education have greater social inequality, leading to higher crime rates? These problems may need to be addressed further in future studies.

6 References

Fatehkia, M. et al. (2019) Correlated impulses: Using Facebook interests to improve predictions of crime rates in urban areas. *PloS one*. [Online] 14 (2), e0211350–e0211350.

Liz, Spencer., James, Wesley, Scott. (2013). 1. School meets street: Exploring the links between low achievement, school exclusion and youth crime among African-Caribbean boys in London. Research Papers in Economics,