To what extent is violent crime in London driven by deprivation?

2020209

Introduction

Between 2012 and 2018, three quarters of violent crime reported in London occurred within regions that are in the top decile for deprivation (London City Hall, 2019). Violent crime consists of a range of offences from harassment to murder (The Crown Prosecution Service, 2023), which not only indicate risk to personal safety but also gauge wider cohesion and wellbeing within communities. Deprivation is a consequence of a lack of income and resources which can be seen in the form of people living in poverty (Townsend, 1979). Further investigation into the link between deprivation and crime is important as deprivation rates within London are high, with half of its boroughs lying within the top third of the most deprived areas nationally (Bosetti, 2019). This investigation aims to quantify the extent to which deprivation causes violent crime within London through the *English indices of deprivation (*2019) dataset and the 'Recorded Crime: Geographic Breakdown - London Datastore' (2023), which record deprivation and crime measures by LSOA regions. Previous studies have found correlation and explained the causation linking deprivation and crime, however the connection between deprivation and violent crime in London is yet to be analysed in detail. This investigation maps violent crime rate and deprivation measures at LSOA levels across London before univariate, multivariate and optimised multivariate linear regression models are produced. The results of these models are presented and discussed, finding that deprivation accounts 51% of the variation in violent crime rate across London. The limitations of the final model such as multicollinearity between independent variables, the removal of outliers and some remaining spatial autocorrelation and heteroskedasticity of residuals are all discussed as accounted for as much as possible. Understanding the driving factors of violent crime will help governments, communities and individuals make the informed steps required to reduce violent crime.

To what extent is violent crime in London driven by deprivation?

What role do the following play:

a) Employment
b) Barriers to housing and services
c) Living environment

Literature review

It is well established that spatial variations in crime are linked to social structure and deprivation (Hannon and Defronzo, 1998; Congdon, 2013). These links may also be stronger than reported crime rate data suggest as a lack of social cohesion (often a result of deprivation) can also reduce the likelihood of populations to report crimes (Goudriaan, Wittebrood and Nieuwbeerta, 2006).

The employment domain makes up 22.5% of the Index of multiple deprivation (IMD) and measures deprivation through assessing the proportion of the working age population which have unwillingly been excluded from work (*English indices of deprivation 2019*, 2019). The link between employment opportunities and crime rate is reported across a range of literature, with Crutchfield and Pitchford (1997) found that people who are unemployed or have unstable jobs are more likely to commit crimes. Kohfeld and Sprague (1988) found that this individual effect is aggregated to areas as people in similar socio-economic conditions have a propensity to live close to each other.

Barriers makes up just 9.3% of the IMD and measures the accessibility both physically and financially to housing and services for an area population (*English indices of deprivation 2019*, 2019). Where barriers are higher less people own homes, Disney *et al.* (2021) found that the right to buy scheme in the UK increased homeownership rates from 60% to 70% between the 1990s and early 2000s, which was responsible for a 1.5% fall in crime rates. They argue that the gentrification and behavioural change of incumbent tenants was the main driver of crime reduction across the period. Home ownership encourages protecting the value of the property and thus increasing security and reducing antisocial behaviour, improving the local community, and driving a reduction in crime.

The environment domain of deprivation measures the quality of indoor and outdoor environments within an area and makes up 9.3% of the IMD (*English indices of deprivation 2019*, 2019). The quality of urban environment can influence crime, Glaeser and Sacerdote (1999) found that urban areas have higher crime rates as cities facilitate crime interactions and lower the chances of arrest or recognition by perpetrators. Additionally, the makeup of urban areas can influence crime rates; mixing commercial and residential zoning was found to reduce crime in one study (Anderson *et al.*, 2013), indicating that where there is poor urban planning, and a low environmental quality ranking, crime rates are higher.

Research into the role of both deprivation on violent crime and the link between deprivation and crime within London is sparce. There is a paper exploring the connection between deprivation and violent crime nationally (Congdon, 2013) however, there is very little London based literature.

Methods

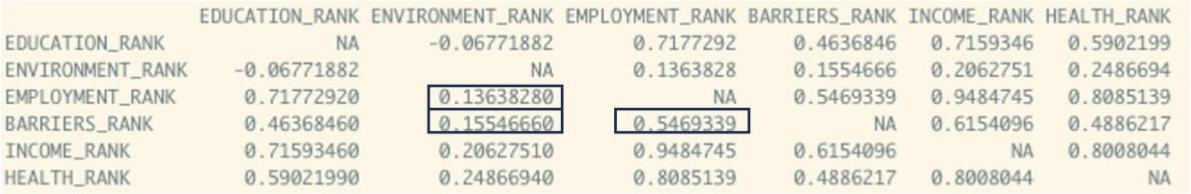
To investigate the role of deprivation on violent crime rate this report will use secondary datasets which are manipulated and modelled with RStudio (RStudio Team, 2020). The data is explored with maps and initial univariate regression models are made between variables. These are then combined into a multivariate regression model before being optimised by removing influential outliers identified by cook's distance. This improved multivariate model is used to make inferences on the role of deprivation on London's violent crime rate.

Datasets

Violent crime data has been subset from the 'Recorded Crime: Geographic Breakdown - London Datastore' (2023) which records the reported crime rate per 1000 people, from the metropolitan police. This is merged with the 2019 Index of Multiple deprivation dataset (*English indices of deprivation 2019*, 2019) which is used to compare the relative deprivation of small areas across England. Seven weighted domains of deprivation are used to score, rank and decile, each LSOA in the country from most to least deprived. Rank has been selected to compare deprivations as it places each LSOA relative to the other 32,843 regions, which is more precise than score which is done to 1 decimal place and to decile which places each LSOA in one of ten deciles.

Selecting appropriate independent variables

When creating multivariate regression models, the co-correlation of variables must be minimised to reduce multicollinearity where the same effect may be contributing to a model more than once (Kiely and Sergievsky, 1991). To minimise the impact of this effect only 3 of 7 deprivation domains were included in the model. In this case, the three variables selected are strong enough to predict violent crime rate. Even through employment and income are both strong predictors of crime rate, only one can be included in the model as they have a correlation coeffect of 0.94 as can be seen in figure 1. Of the remaining indices environment and barriers were chosen as they have the lowest correlation to each other and employment while providing a strong influence on variation in crime rate. The correlation of 0.54 between employment and barriers is reasonably high however some level of correlation must be accepted in this case, as all three independent variables are indicators of deprivation therefore some correlation between them is expected.



```
                 EDUCATION_RANK ENVIRONMENT_RANK EMPLOYMENT_RANK BARRIERS_RANK INCOME_RANK HEALTH_RANK
EDUCATION_RANK               NA      -0.06771882       0.7177292     0.4636846   0.7159346   0.5902199
ENVIRONMENT_RANK    -0.06771882               NA       0.1363828     0.1554666   0.2062751   0.2486694
EMPLOYMENT_RANK      0.71772920       0.13638280              NA     0.5469339   0.9484745   0.8085139
BARRIERS_RANK        0.46368460       0.15546660       0.5469339            NA   0.6154096   0.4886217
INCOME_RANK          0.71593460       0.20627510       0.9484745     0.6154096          NA   0.8008044
HEALTH_RANK          0.59021990       0.24866940       0.8085139     0.4886217   0.8008044          NA
```

Figure 1: Correlation matrix of IMD domains. The correlations between the three selected independent variables are highlighted. Crime is removed as a potential explanatory variable for violent crime

Initial data transformations:

Linear regression models require normally distributed, continuous data (*Assumptions of Linear Regression*, no date), a Shapiro-Wilk test is used to test for normality. For large datasets the p-value of Shapiro wilk tests are less indicative of normal distributions as a very small variation from normality can affect the value, therefore it is more useful to look at the w value (Choueiry, no date). A Shapiro-Wilk test suggests that the violent crime rate per 1000 is not normal (W = 0.5601, p-value < 2.2e-16) therefore it is logged, after which a more normal shape is observed (W = 0.98781, p-value < 2.2e-16), and the dependant variable is ready to be modelled. None of the independent variables require transformations before being used in the regression models as they all have normally shaped data (Employment w=0.9612, p-value < 2.2e-16, Barriers w = 0.91557, p-value < 2.2e-16 and environment w = 0.97392, p-value < 2.2e-16).

Univariate models

Univariate linear regression models are used to identify relationships between independent and dependant variables; a linear relationship is another requirement for multivariate regression models (*Assumptions of Linear Regression*, no date). The deprivation ranks are also mapped to show the distribution of each independent variable across London. All univariate models of independent variables show linear relationships to the dependant variables and thus they can be used in a multivariate linear regression model.

Multivariate model

Using the three independent variables an initial multivariate model is produced. The effectiveness of the model is explored through several measures, adjusted R^2, AIC score, Breusch-Pagen test and Moran's I test as well as investigating the normality of fitted, residual and Cooks distance values. This initial model is then optimised by removing outliers identified with Cooks distance. The Cook's value at which the cut-off was selected for was chosen to maximise improvement to the adjusted R^2 value, AIC score, Breusch-Pagen score and Moran's I score while minimising the number of samples removed through sensitivity studies.

Results

Violent crime rate map

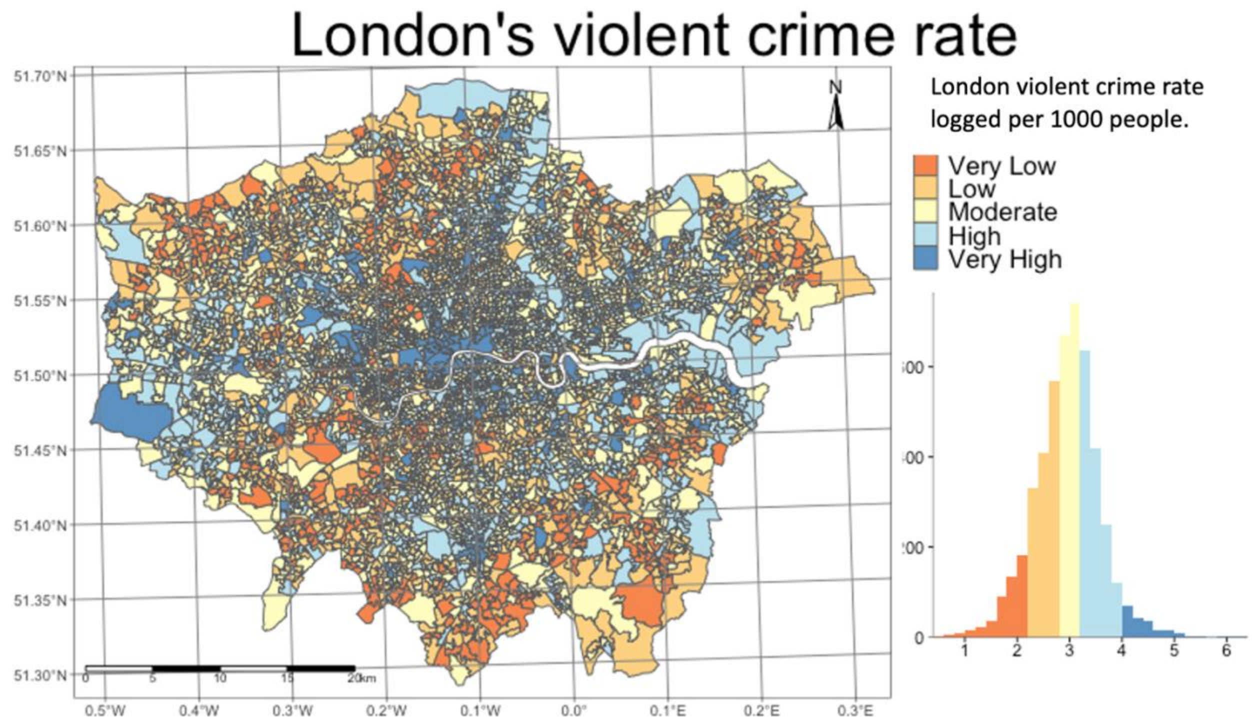
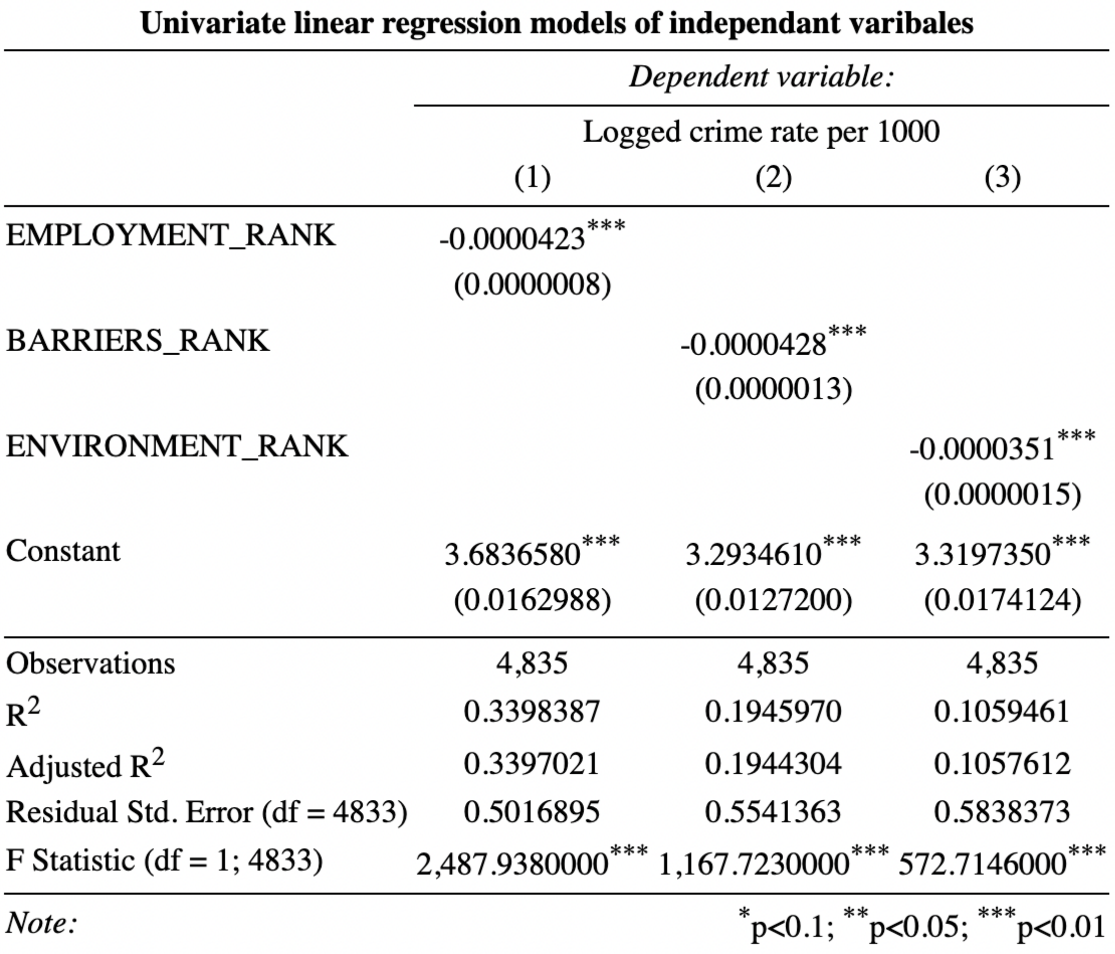


Figure 2: London's logged violent crime rate mapped by LSOA.

The map of London’s violent crime rate shows that there are areas with more reported events than others. The highest crime rates are in central London. The lowest crime rates are in the southeast and northwest boundaries of the city. There appears to be clustering with high crime rates being surrounded by high crime rates and low crime rates being surrounded by low crime rates.

Univariate regression results



Univariate linear regression models of independant varibales

	Dependent variable:		
	Logged crime rate per 1000		
	(1)	(2)	(3)
EMPLOYMENT_RANK	-0.0000423***		
	(0.0000008)		
BARRIERS_RANK		-0.0000428***	
		(0.0000013)	
ENVIRONMENT_RANK			-0.0000351***
			(0.0000015)
Constant	3.6836580***	3.2934610***	3.3197350***
	(0.0162988)	(0.0127200)	(0.0174124)
Observations	4,835	4,835	4,835
R^2	0.3398387	0.1945970	0.1059461
Adjusted R^2	0.3397021	0.1944304	0.1057612
Residual Std. Error (df = 4833)	0.5016895	0.5541363	0.5838373
F Statistic (df = 1; 4833)	2,487.9380000***	1,167.7230000***	572.7146000***
Note:			*p<0.1; **p<0.05; ***p<0.01

Figure 3: Summary table of univariate regression model output from independent variables. Produced with Stargazer. (Hlavac, 2022)

R^2 values show how much variation in the dependant variable is explained by an independent variable. Changes to employment explain the most variation in violent crime with 34%, then Barriers which account for 19% of variation and then environment with an R^2 of 11%. All p-values are below 0.05, suggesting these results are statistically significant.

Employment

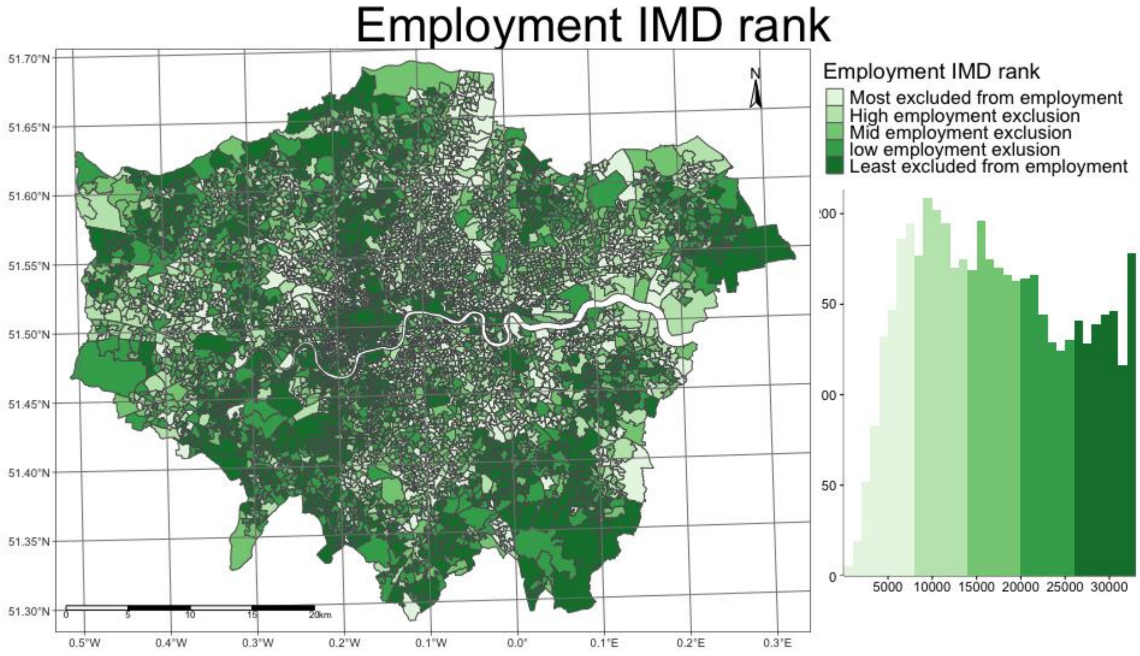


Figure 4: London's employment IMD rank mapped by LSOA region.

Of the three domains of deprivation, employment deprivation is lowest relative to national deprivation. Figure 4 shows the areas most excluded from employment are in the east of the city. The univariate regression model (figures 3 and 5) show that generally the LSOAs more deprived of employment opportunities in London are also those with higher violent crime rates.

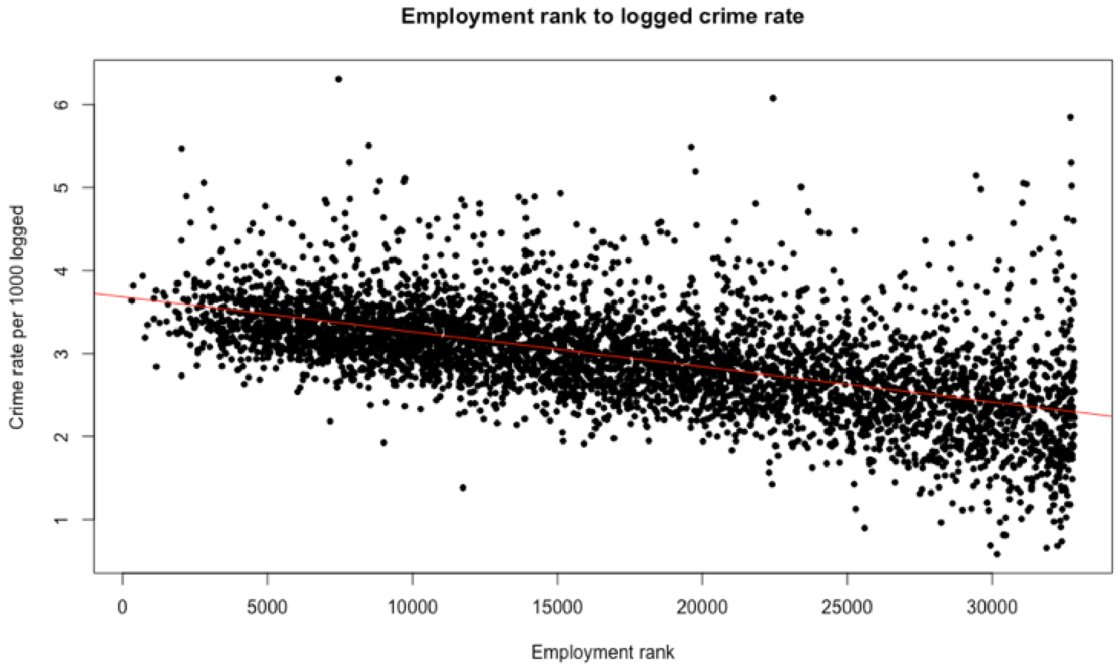


Figure 5: Scatter plot of employment IMD rank against the logged crime rate of London. Univariate regression model shown by red line.

Barriers to housing and services

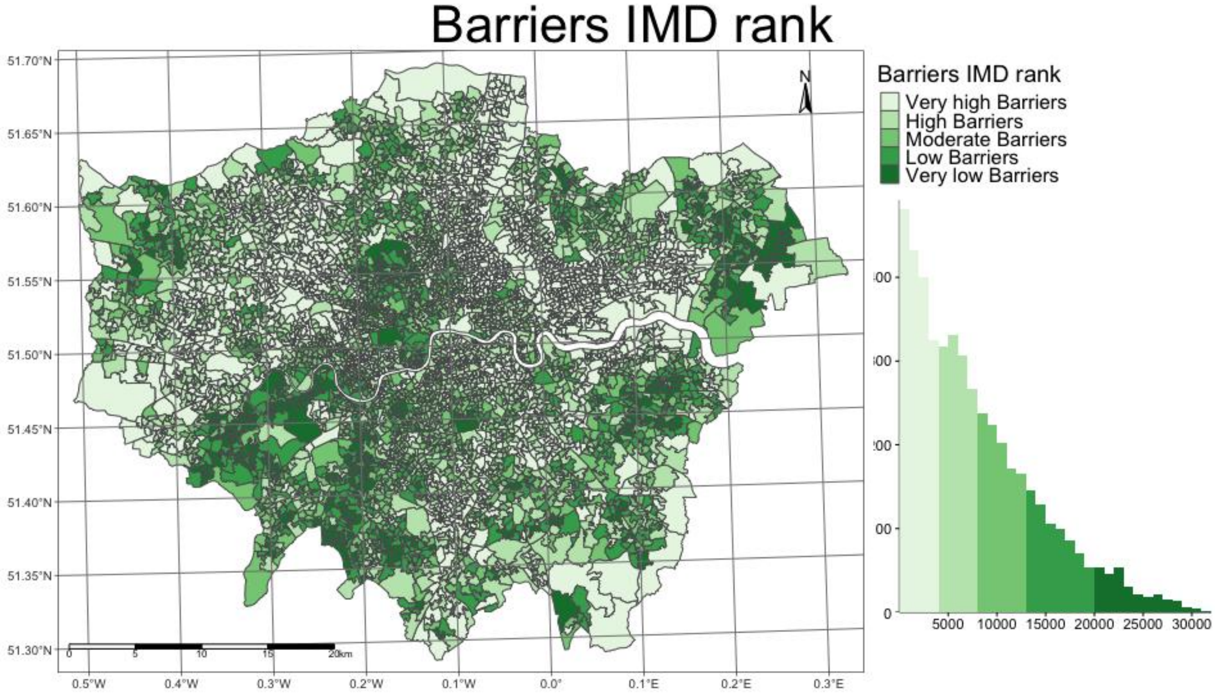


Figure 6: London's barriers IMD rank mapped by LSOA region.

Relative to the rest of the country, there is lots of barrier-based deprivation in London as shown by the histogram of deprivation rank on figure 6 . The map in figure 6 shows that there is a band of very high barriers to the northwest region of London and in the east. As shown by Figures 3 and 7, the model shows again that within London, the areas with higher barriers to services and housing have higher violent crime rates.

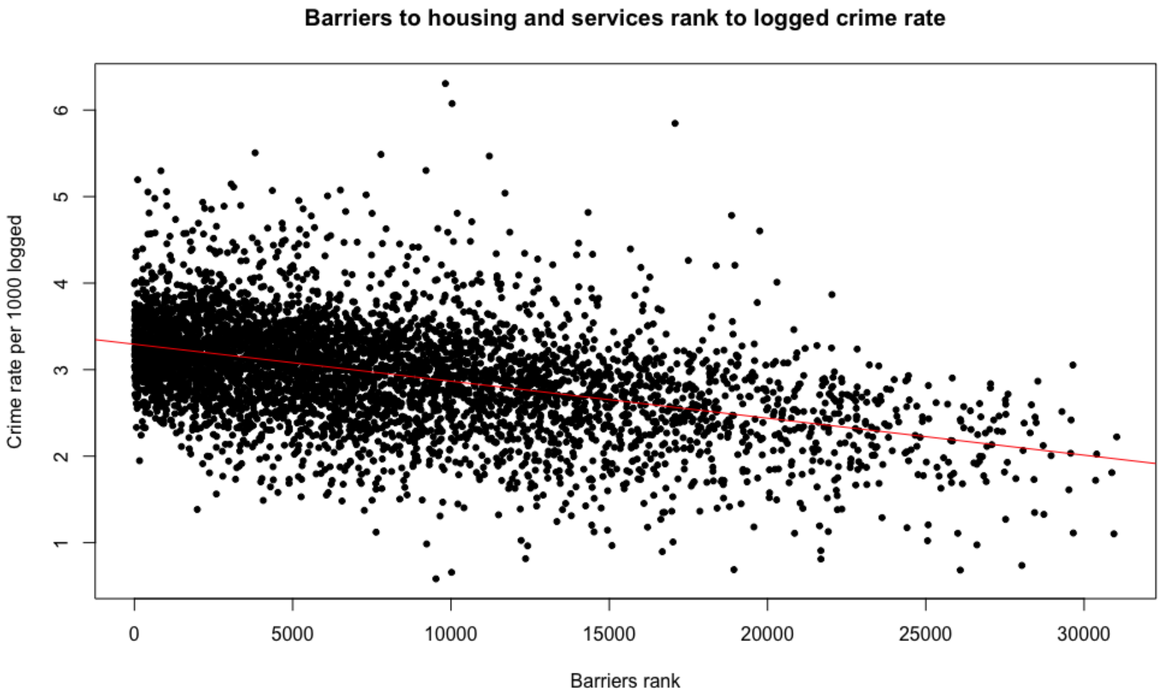


Figure 7: Scatter plot of barriers IMD rank against the logged crime rate of London. Univariate regression model shown by red line.

Living Environment

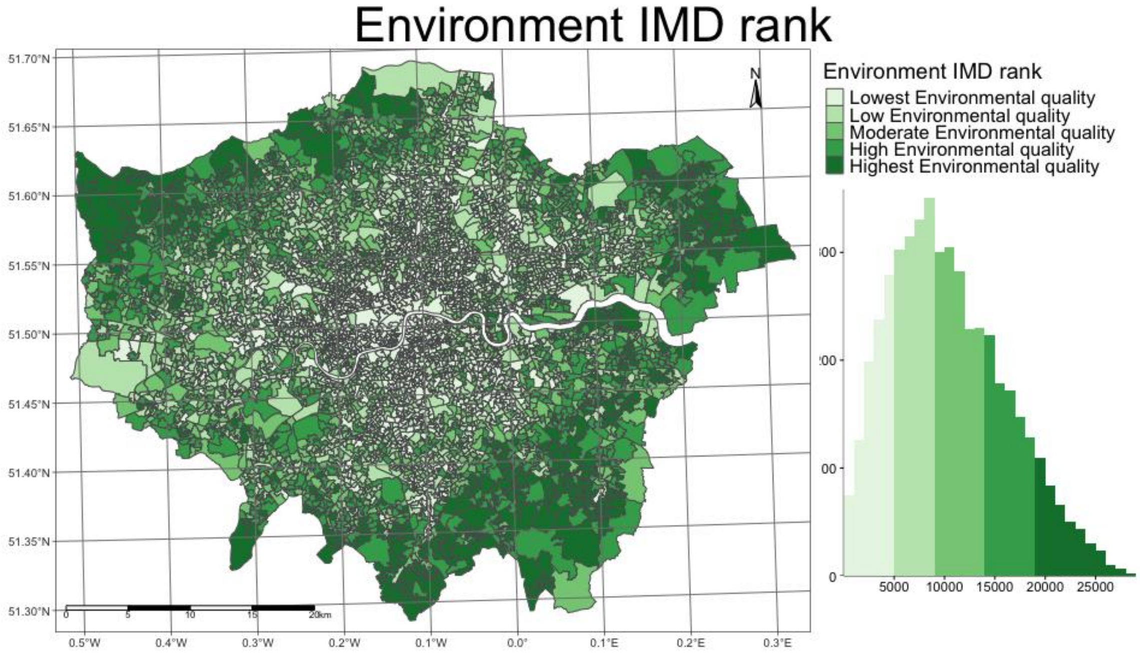


Figure 8: London's environment IMD rank mapped by LSOA region.

The spatial distribution of environment deprivation is shown in figure 8, there is clear spatial pattern to the ranking with the lowest environmental quality nearest the centre of London and higher quality environments (relative to the rest of the country). As can be seen by figure 3 and figure 9, areas with worse living environment have higher violent crime rates.

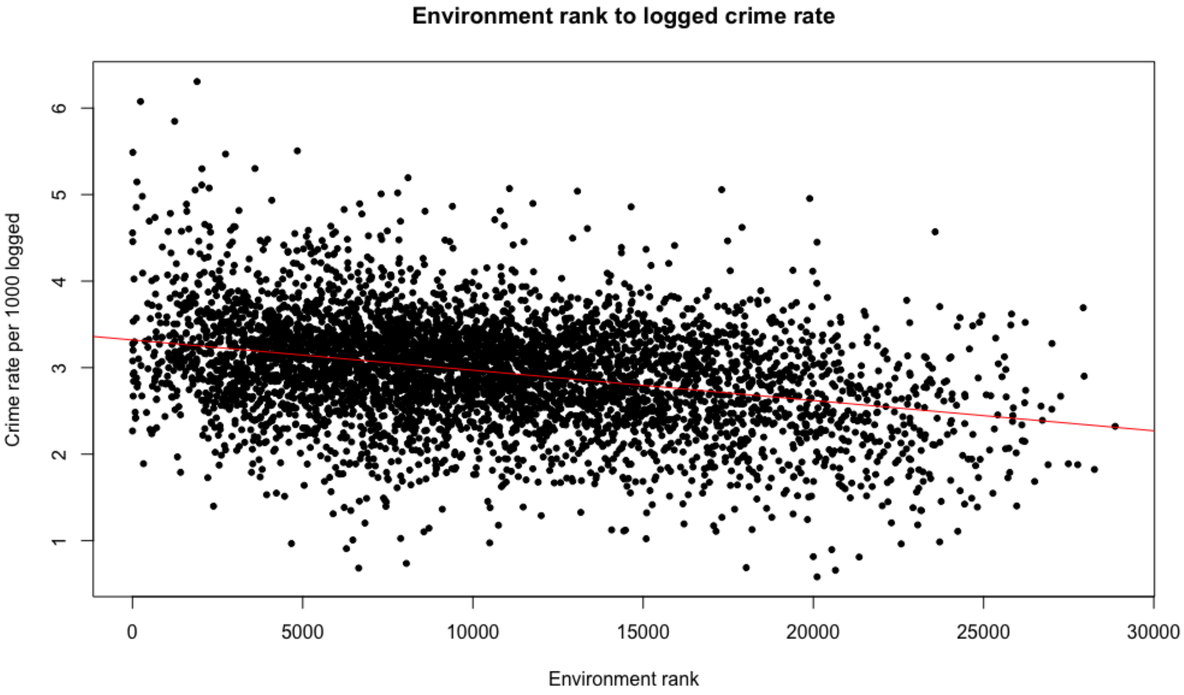
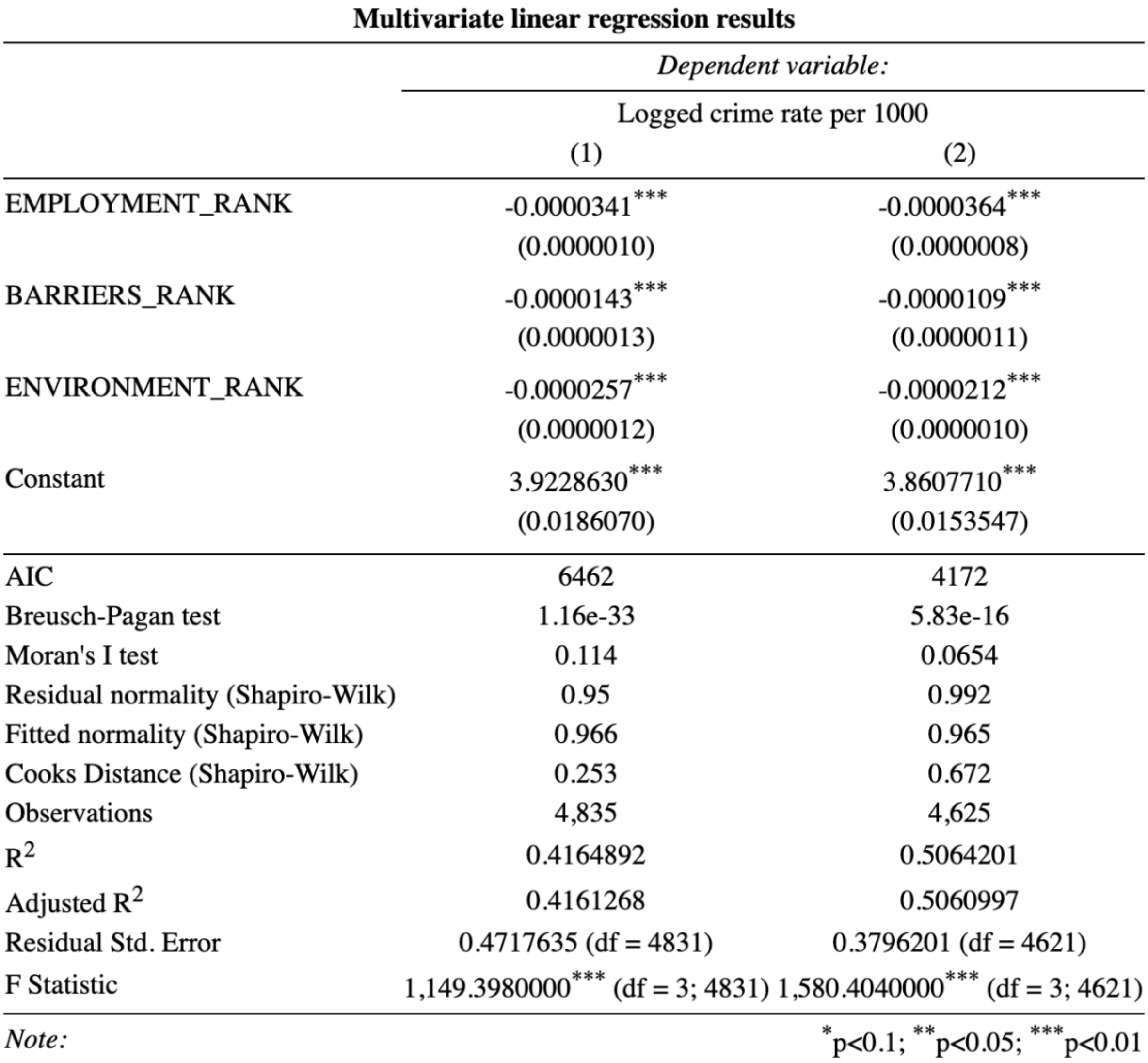


Figure 9: Scatter plot of barriers IMD rank against the logged crime rate of London. Univariate regression model shown by red line.



Multivariate linear regression results

	Dependent variable:	
	Logged crime rate per 1000	
	(1)	(2)
EMPLOYMENT_RANK	-0.0000341***	-0.0000364***
	(0.0000010)	(0.0000008)
BARRIERS_RANK	-0.0000143***	-0.0000109***
	(0.0000013)	(0.0000011)
ENVIRONMENT_RANK	-0.0000257***	-0.0000212***
	(0.0000012)	(0.0000010)
Constant	3.9228630***	3.8607710***
	(0.0186070)	(0.0153547)
AIC	6462	4172
Breusch-Pagan test	1.16e-33	5.83e-16
Moran's I test	0.114	0.0654
Residual normality (Shapiro-Wilk)	0.95	0.992
Fitted normality (Shapiro-Wilk)	0.966	0.965
Cooks Distance (Shapiro-Wilk)	0.253	0.672
Observations	4,835	4,625
R^2	0.4164892	0.5064201
Adjusted R^2	0.4161268	0.5060997
Residual Std. Error	0.4717635 (df = 4831)	0.3796201 (df = 4621)
F Statistic	1,149.3980000*** (df = 3; 4831)	1,580.4040000*** (df = 3; 4621)
Note:		*p<0.1; **p<0.05; ***p<0.01

Figure 10: Summary table of both multivariate linear regression models. (1) is the Initial model and (2) is the final model. Outputs demonstrate an improvement to the model. Produced with Stargazer (Hlavac, 2022)

Initial multivariate linear regression model

The initial multivariate regression results can be seen in figure 10. The adjusted R^2 value of 0.42 suggests that the model is effectively predicting 42% of the changes in crime rate with the independent variables. The AIC sore of this first model is 6462, the value is arbitrary however when compared to the improved model it shows that the improved model is better at predicting changes to crime rate as the score falls to 4625.

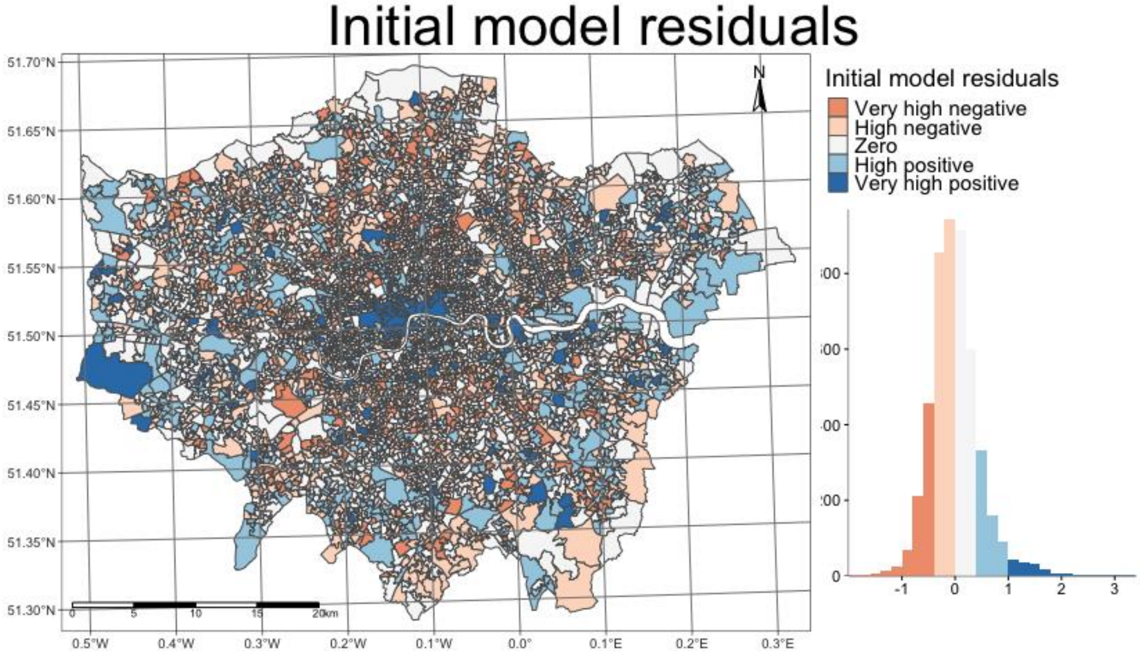


Figure 11: Map of residuals from the initial multivariate linear regression model.

Figure 11 shows that the model residuals are normal in shape, meaning that the model captures the main variations in crime rate at that the variation within the error is random. However there appears to be some spatial autocorrelation, where there are clusters of high positive residuals in the centre of London, where crime rate is underpredicted. This spatial autocorrelation of residuals is confirmed by the Moran's I statistic of 0.114; there is spatial autocorrelation of residuals across the whole dataset. Additionally, the Breusch-Pagen test statistic (156.3 and p < 2.2e -16) indicates that there is heteroskedasticity within the model. This can be visually confirmed by figure 13 which shows residuals departing from normality at the highest values of crime rate.

Improved multivariate regression model.

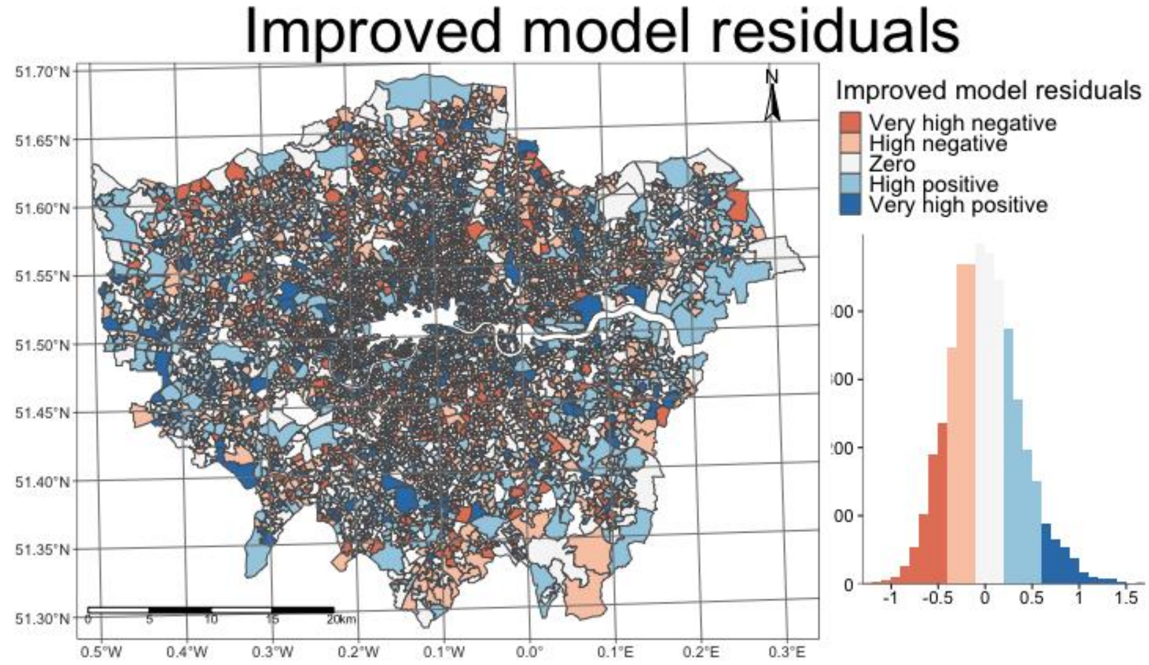


Figure 12: Residual map of improved multivariate regression model

The initial model is improved by removing the most influential outliers, gaps in the map of figure 12 show which LSOAs have been removed. They are mostly in central London; for these removed LSOAs deprivation is inaccurate in predicting crime rates. There is a more normal shape to residuals and there appears to be less spatial autocorrelation within the residuals when compared to the initial model. This is confirmed by the smaller Moran's I statistic of 0.069 however there is still spatial autocorrelation of residuals. The heteroskedasticity of the model is reduced as seen by the more normal black points on figure 13 and the lower Breusch-Pagen score of 74.1 (p-value = 5.828e-16).

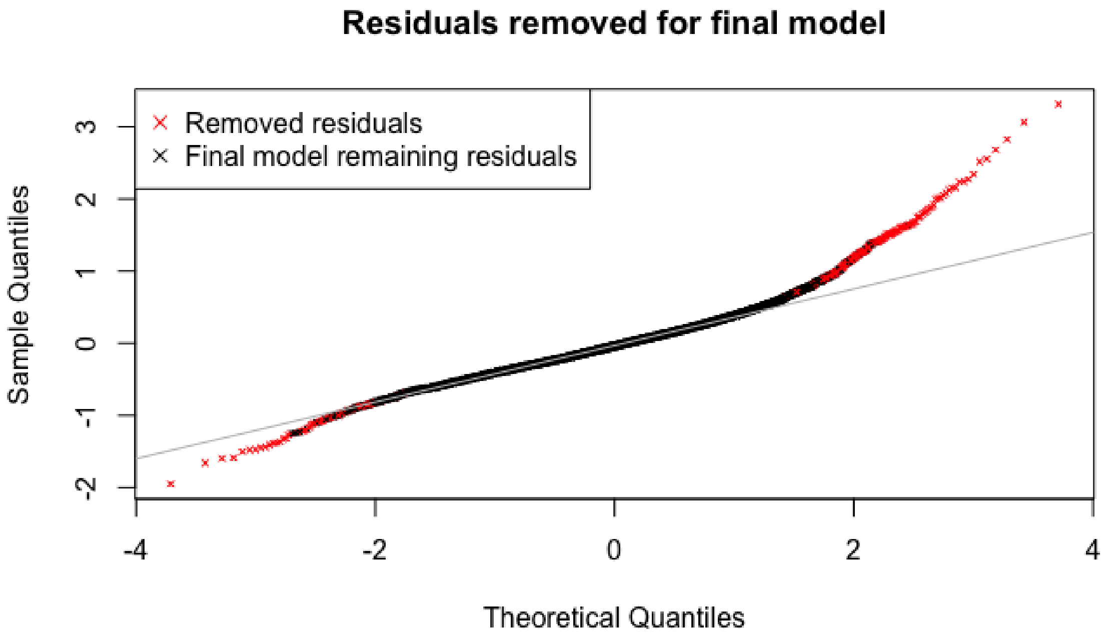


Figure 13: QQ plot showing the initial model residuals. Once the model is optimised only the black residuals remain; the red residuals are the 210 which are removed from the model, which are causing heteroskedasticity and spatial autocorrelation issues.

Discussion

The multivariate linear regression model produced demonstrates that approximately 51% of the variation in logged violent crime rate reported in London is associated with variations in the deprivation ranks of variables measured. This influence of deprivation is slightly higher than the national average found by Congdon (2013) who modelled deprivation changes to be driving 41% of the variation in violent crime.

Variations in employment are the largest influence on violent crime rate accounting for 34% of variation. One explanation of this is that individuals in unstable jobs are more likely to commit crime (Crutchfield and Pitchford, 1997).

Barriers to housing have the next largest influence on violent crime rates at 19%, which may be a result of home owners having more motivation to keep properties secure than those who are renting (Disney *et al.*, 2021).

Finally, while still providing a useful addition to the model, environmental deprivation rank contributes the least to the model. Where there is a poor quality environment, crime is more likely to occur (Anderson *et al.*, 2013) which was found to explain 11% of variation in crime rate.

The strength of the final model lies in its ability to produce an adjusted R^2 value which predicts more than half of the variation in violent crime rate with only three independent variables. This model also has less heteroskedasticity and less spatial autocorrelation of residuals than the initial model, which implies that it is more reliable to make inferences about the influences on crime rates from. Even though removing influential outliers doesn't eliminate heteroskedasticity (Breusch-Pagen score of 156, p-value < 2.2e-16 to a 74.1, p-value = 5.828e-16) or spatial autocorrelation of residuals (Moran's I 0.114 to 0.069) it does improve It.

There are three main limitations of the model; a lack of ability to predict all LSOA areas, the potential for multicollinearity between independent variables and some remaining heteroskedasticity and spatial autocorrelation of residuals.

By removing the influential outliers using cooks' distance, there are 210 LSOAs in London where the model doesn't fit as it was mis-predicting crime rate to an extent which skewed the overall accuracy of the model. Sensitivity studies were performed to ensure the minimum amount of data was removed while improving the model, which only lead to 4% of LSOAs being removed and an model improvement of 10% (by adjusted R^2). The removal improves the model overall through reducing heteroskedasticity and spatial autocorrelation and increasing adjusted R^2 but for those areas removed It is not reliable for predicting crime rate. Figure 11 shows that of the outliers which were removed a large group of them were in central London. One potential reason for this could be a limitation of the data set; crimes reported may be perpetrated by people who aren't contributing to deprivation measures in that area. For example, people may commute into the city to commit crime where they are less likely to be caught (Glaeser and Sacerdote, 1999) when they are included in the deprivation measure of a different LSOA.

Even though deprivation variables were selected to minimise correlation, any form of multi collinearity between independent variables can make adjusted R^2 values less accurate (Kiely and Sergievsky, 1991), in this case some level of multicollinearity must be accepted as the independent variables all contribute to the measure of the same factor (deprivation) and thus correlation between them is somewhat unavoidable.

The remaining heteroskedasticity and spatial autocorrelation of residuals indicates that there are more predictor variables influencing violent crime rate in London. This investigation is not arguing that deprivation is the only factor influencing violent crime rate; therefore, it is not expected to create a linear regression model which accounts for all variation, and removing too many outliers causing heteroskedasticity and spatial autocorrelation may increase the R^2 value but would also reduce the proportion of London it is successfully predicting.

Future research that models independent variables other than deprivation may provide stronger adjusted R^2 values however the focus of this investigation is to quantify the extent to which deprivation influences crime rather than finding all variables which influence violent crime. Additionally, the other factors suggested to be causing crime from literature such as inequality (Hooghe *et al.*, 2011), social capital (Congdon, 2013) or family dynamics of individuals (Ellis, Farrington and Hoskin, 2019) are more difficult to quantify and not easily accessible in the current datasets available.

Overall, the model produced shows strong, explainable links between variations in employment, barriers to housing and living environment and the variations in the logged violent crime rate across London's LSOAs. Additionally, consideration has been made for the limitations of the model, by carefully selecting the cooks' distance cut-off, choosing the least correlated variables and by considering other factors responsible for heteroskedasticity and spatial autocorrelation of residuals, allowing for an effective model in calculating the influence of deprivation on violent crime rate across London's LSOA's.

Conclusion

Using multivariate linear regression models this investigation found that 51% of the variation in violent crime across London can be attributed to changes in deprivation rank of employment, barriers to housing and lived environment.

These findings were strengthened by existing literature which provides links between the correlation of deprivation and crime to justify causation.

Considerations were taken to carefully mitigate issues of multicollinearity, minimise the level of data which required removal and explain heteroskedasticity and spatial autocorrelation of residuals.

A literature review found links between deprivation and crime were well established but hadn't been examined in the context of violent crime in London. R was used to create and refine a multivariate linear model which modelled the role of deprivation on London's violent crime rate. The results showed employment deprivation rank explained the largest variation in crime rate with 34%, then barriers which accounted for 19% and then living environment accounting for 11% of variation. When combined these deprivation variables accounted for 51% of variation in violent crime rate in London. The rest of the variation is likely driven by other non-deprivation variables such as inequality (Hooghe *et al.*, 2011), social capital (Congdon, 2013) or family dynamics of individuals (Ellis, Farrington and Hoskin, 2019).

Further research into modelling how these variables influence London's violent crime rate would improve the knowledge in the subject. Understanding the influences impacting violent crime rate is essential for planning policy to reduce it in the future. It is important to note that even though this investigation implies that reducing deprivation in London would reduce the rates of violent crime, which increases the safety and well-being of populations, social stability and encourages economic development, it is not arguing that all those that are deprived commit crime, only that where there is deprivation, violent crime rates are more likely to be higher.

3,498 words

Bibliography

Anderson, J.M. *et al.* (2013) 'Reducing Crime by Shaping the Built Environment with Zoning: An Empirical Study of Los Angeles', *University of Pennsylvania Law Review*, 161(3), pp. 699–756.

Assumptions of Linear Regression (no date) *Statistics Solutions*. Available at: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-linear-regression/ (Accessed: 25 May 2023).

Choueiry, G. (no date) 'How to Report the Shapiro-Wilk Test – QUANTIFYING HEALTH'. Available at: https://quantifyinghealth.com/report-shapiro-wilk-test/ (Accessed: 26 May 2023).

Congdon (2013) 'A Model for Spatially Varying Crime Rates in English Districts: The Effects of Social Capital, Fragmentation, Deprivation and Urbanicity', *International Journal of Criminology and Sociology* [Preprint]. Available at: https://doi.org/10.6000/1929-4409.2013.02.14.

Crutchfield, R.D. and Pitchford, S.R. (1997) 'Work and crime: The effects of labor stratification', *Social Forces*, 76, pp. 93–118. Available at: https://doi.org/10.2307/2580319.

Disney, R.F. *et al.* (2021) 'Does Homeownership Reduce Crime? A Radical Housing Reform from the UK'. Rochester, NY. Available at: https://doi.org/10.2139/ssrn.3759340.

Ellis, L., Farrington, D.P. and Hoskin, A.W. (2019) *Handbook of Crime Correlates*. Academic Press.

English indices of deprivation 2019 (2019) *GOV.UK*. Available at: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 (Accessed: 25 May 2023).

Glaeser, E.L. and Sacerdote, B. (1999) 'Why is There More Crime in Cities?', *Journal of Political Economy*, 107(S6), pp. S225–S258. Available at: https://doi.org/10.1086/250109.

Goudriaan, H., Wittebrood, K. and Nieuwbeerta, P. (2006) 'NEIGHBOURHOOD CHARACTERISTICS AND REPORTING CRIME: Effects of Social Cohesion, Confidence in Police Effectiveness and Socio-Economic Disadvantage1', *The British Journal of Criminology*, 46(4), pp. 719–742.

Hannon, L. and Defronzo, J. (1998) 'The Truly Disadvantaged, Public Assistance, and Crime*', *Social Problems*, 45(3), pp. 383–392. Available at: https://doi.org/10.2307/3097192.

Hooghe, M. *et al.* (2011) 'Unemployment, Inequality, Poverty and Crime: Spatial Distribution Patterns of Criminal Acts in Belgium, 2001–06', *The British Journal of Criminology*, 51(1), pp. 1–20. Available at: https://doi.org/10.1093/bjc/azq067.

Kiely, J.L. and Sergievsky, G.H. (1991) 'Some conceptual problems in multivariable analyses of perinatal mortality', *Paediatric and Perinatal Epidemiology*, 5(3), pp. 243–257. Available at: https://doi.org/10.1111/j.1365-3016.1992.tb00289.x.

2020209

Kohfeld, C.W. and Sprague, J. (1988) ‘Urban Unemployment Drives Urban Crime’, *Urban Affairs Quarterly*, 24(2), pp. 215–241. Available at: https://doi.org/10.1177/004208168802400203.

London City Hall (2019) *Revealed: full links between poverty and violent crime in London*. Available at: https://www.london.gov.uk/press-releases/mayoral/full-links-between-poverty-and-violent-crime (Accessed: 25 May 2023).

Nicolas Bosetti (2019) *Deprivation remains a major issue for London boroughs*. Available at: https://centreforlondon.org/blog/deprivation-london/ (Accessed: 25 May 2023).

Peter Townsend (1979) *Poverty in the United Kingdom: A Survey of Household Resources and Standards of Living*. Middlesex. Available at: https://www.journals.uchicago.edu/doi/10.1086/227691 (Accessed: 28 May 2023).

‘Recorded Crime: Geographic Breakdown - London Datastore’ (2023). Available at: https://data.london.gov.uk/dataset/recorded_crime_summary (Accessed: 25 May 2023).

RStudio Team (2020) ‘RStudio: Integrated Development for R’. PBC, Boston: RStudio. Available at: http://www.rstudio.com/.

The Crown Prosecution Service (2023) *Violent crime*. Available at: https://www.cps.gov.uk/crime-info/violent-crime (Accessed: 28 May 2023).

Appendix

```
## Setup --------------------------------------------------------------------
#Setting up working directory, packages and downloading data.
setwd("/Volumes/TOM's HD/sm2idwd")
library(sf)
library(sp)
library(tmap)
library(lmtest)
library(RColorBrewer)
library(stargazer)
library(spdep)
library(RANN)

crime = read.csv("LONDON_CRIME_RATE_2019.csv")
deprivation =read.csv("LONDON_DEPRIVATION_2019.csv")
cencus= read.csv("LSOA_DATA.csv")
lsoa= read_sf("LSOA.shp")

#Merging data into a single spatial data-set
lsoa_crime = merge(lsoa, crime, by.x = "LSOA", by.y = "LSOA", all.x = TRUE)
lsoa_deprivation= merge(lsoa, deprivation, by.x = "LSOA", by.y = "LSOA", all.x = TRUE)
lsoa_cencus= merge(lsoa, cencus, by.x = "LSOA", by.y = "LSOA", all.x = TRUE)
crime_dep_join = st_join(lsoa_crime, lsoa_deprivation, join = st_equals, left = TRUE)
fulldata = st_join(crime_dep_join, lsoa_cencus, join = st_equals, left = TRUE)

## Crime map ----------------------------------------------------------------
shapiro.test(fulldata$RATE_per_1000)
hist(fulldata$RATE_per_1000)
# 0.5601
# this isn't normal enough

fulldata$RATE_per_1000logged= log(fulldata$RATE_per_1000)
shapiro.test(fulldata$RATE_per_1000logged)
hist(fulldata$RATE_per_1000logged)
# w = 0.9878
# and histogram looks much more normal

tm_shape(fulldata) +

  tm_polygons("RATE_per_1000logged", title = "London violent Crime Rate
            \nlogged per 1000 people",
            palette = "RdYlBu",legend.hist = TRUE, labels =
              c("Very Low", "Low", "Moderate", "High", "Very High"),
            style = "jenks") +
  tm_shape(fulldata) +

  tm_borders(lwd = 0.2, col = "#4C4E52")+
  tm_compass(position = c(0.9, 0.9)) +
  tm_scale_bar(position = c(0.01, 0.016))+
  tm_legend(position = c("left", "top"), title.size = 0.6)+
  tm_layout(main.title = "London's violent crime rate",

            main.title.size = 2.5,
            main.title.position = c("center", "top"),
            legend.outside = TRUE,
            legend.title.size = 2.5,
            legend.text.size = 2,
            legend.hist.width = 10,
            legend.hist.height = 1)+

  tm_graticules()

# Correlation matrix -----------------------------------------------------------


cor_test_education_environment = cor.test(fulldata$EDUCATION_RANK,
                                          fulldata$ENVIRONMENT_RANK)
#-0.07
cor_test_environment_employment = cor.test(fulldata$ENVIRONMENT_RANK,
                                           fulldata$EMPLOYMENT_RANK)
#0.14
cor_test_environment_barriers = cor.test(fulldata$ENVIRONMENT_RANK,
                                         fulldata$BARRIERS_RANK)
#0.16
cor_test_environment_income = cor.test(fulldata$ENVIRONMENT_RANK,
                                       fulldata$INCOME_RANK)
#0.2
cor_test_health_environment = cor.test(fulldata$HEALTH_RANK,
                                       fulldata$ENVIRONMENT_RANK)
#0.24
cor_test_education_barriers = cor.test(fulldata$EDUCATION_RANK,
                                       fulldata$BARRIERS_RANK)
#0.46
cor_test_health_barriers = cor.test(fulldata$HEALTH_RANK,
                                    fulldata$BARRIERS_RANK)
#0.48
cor_test_employment_barriers = cor.test(fulldata$EMPLOYMENT_RANK,
                                        fulldata$BARRIERS_RANK)
#0.54
cor_test_education_health = cor.test(fulldata$EDUCATION_RANK,
                                     fulldata$HEALTH_RANK)
#0.59
cor_test_income_barriers = cor.test(fulldata$INCOME_RANK,
                                    fulldata$BARRIERS_RANK)
```

2020209

```
#0.65
cor_test_education_income = cor.test(fulldata$EDUCATION_RANK,
                             fulldata$INCOME_RANK)
#0.72
cor_test_education_employment = cor.test(fulldata$EDUCATION_RANK,
                                fulldata$EMPLOYMENT_RANK)
#0.72
cor_test_health_income = cor.test(fulldata$HEALTH_RANK,
                           fulldata$INCOME_RANK)
# 0.8
cor_test_health_employment = cor.test(fulldata$HEALTH_RANK,
                              fulldata$EMPLOYMENT_RANK)
#0.80
cor_test_employment_income = cor.test(fulldata$EMPLOYMENT_RANK,
                              fulldata$INCOME_RANK)
#0.96

# Creating a correlation matrix
cor_matrix <- matrix(NA, nrow = 6, ncol = 6)
row.names(cor_matrix) <- colnames(cor_matrix) <- c("EDUCATION_RANK",
                                      "ENVIRONMENT_RANK",
                                      "EMPLOYMENT_RANK",
                                      "BARRIERS_RANK",
                                      "INCOME_RANK",
                                      "HEALTH_RANK")

cor_matrix["EDUCATION_RANK", "ENVIRONMENT_RANK"] <- 0.06771882
cor_matrix["ENVIRONMENT_RANK", "EDUCATION_RANK"] <- 0.06771882
cor_matrix["ENVIRONMENT_RANK", "EMPLOYMENT_RANK"] <-  0.1363828
cor_matrix["EMPLOYMENT_RANK", "ENVIRONMENT_RANK"] <-  0.1363828
cor_matrix["ENVIRONMENT_RANK", "BARRIERS_RANK"] <- 0.1554666
cor_matrix["BARRIERS_RANK", "ENVIRONMENT_RANK"] <- 0.1554666
cor_matrix["ENVIRONMENT_RANK", "INCOME_RANK"] <- 0.2062751
cor_matrix["INCOME_RANK", "ENVIRONMENT_RANK"] <- 0.2062751
cor_matrix["HEALTH_RANK", "ENVIRONMENT_RANK"] <- 0.2486694
cor_matrix["ENVIRONMENT_RANK", "HEALTH_RANK"] <- 0.2486694
cor_matrix["EDUCATION_RANK", "BARRIERS_RANK"] <- 0.4636846
cor_matrix["BARRIERS_RANK", "EDUCATION_RANK"] <- 0.4636846
cor_matrix["HEALTH_RANK", "BARRIERS_RANK"] <- 0.4886217
cor_matrix["BARRIERS_RANK", "HEALTH_RANK"] <- 0.4886217
cor_matrix["EMPLOYMENT_RANK", "BARRIERS_RANK"] <- 0.5469339
cor_matrix["BARRIERS_RANK", "EMPLOYMENT_RANK"] <- 0.5469339
cor_matrix["EDUCATION_RANK", "HEALTH_RANK"] <- 0.5902199
cor_matrix["HEALTH_RANK", "EDUCATION_RANK"] <- 0.5902199
cor_matrix["INCOME_RANK", "BARRIERS_RANK"] <- 0.6154096
cor_matrix["BARRIERS_RANK", "INCOME_RANK"] <- 0.6154096
cor_matrix["EDUCATION_RANK", "INCOME_RANK"] <- 0.7159346
```

```
cor_matrix["INCOME_RANK", "EDUCATION_RANK"] <- 0.7159346
cor_matrix["EDUCATION_RANK", "EMPLOYMENT_RANK"] <- 0.7177292
cor_matrix["EMPLOYMENT_RANK", "EDUCATION_RANK"] <- 0.7177292
cor_matrix["HEALTH_RANK", "INCOME_RANK"] <- 0.8008044
cor_matrix["INCOME_RANK", "HEALTH_RANK"] <- 0.8008044
cor_matrix["HEALTH_RANK", "EMPLOYMENT_RANK"] <- 0.8085139
cor_matrix["EMPLOYMENT_RANK", "HEALTH_RANK"] <- 0.8085139
cor_matrix["EMPLOYMENT_RANK", "INCOME_RANK"] <- 0.9484745
cor_matrix["INCOME_RANK", "EMPLOYMENT_RANK"] <- 0.9484745
# Displaying the correlation matrix
#cor_matrix$Row_Sums <- rowSums(cor_matrix, na.rm = TRUE)
cor_matrix

# Checking normalicy --------------------------------------------------------
shapiro.test(fulldata$EMPLOYMENT_RANK)
#W = 0.9612, p-value < 2.2e-16
shapiro.test(fulldata$ENVIRONMENT_RANK)
#W = 0.97392, p-value < 2.2e-16
shapiro.test(fulldata$BARRIERS_RANK)
#W = 0.91557, p-value < 2.2e-16



#Employment -----------------------------------------------------------------
tm_shape(fulldata) +

  tm_polygons("EMPLOYMENT_RANK", title = "Employment IMD rank",
              palette = "Greens",legend.hist = TRUE, labels =
                c("Most excluded from employment  ", "High employment exclusion",
                  "Mid employment exclusion", "low employment exlusion",
                  "Least excluded from employment"),
              style = "fisher") +
  tm_shape(fulldata) +

  tm_borders(lwd = 0.2, col = "#4C4E52")+
  tm_compass(position = c(0.9, 0.9)) +
  tm_scale_bar(position = c(0.01, 0.016))+
  tm_legend(position = c("left", "top"), title.size = 0.6)+
  tm_layout(main.title = "Employment IMD rank",

            main.title.size = 2.5,
            main.title.position = c("center", "top"),
            legend.outside = TRUE,
            legend.title.size = 2,
            legend.text.size = 1.5,
            legend.hist.width = 2,
            legend.hist.height = 1)+
  tm_graticules()



employ_model= lm(RATE_per_1000logged ~EMPLOYMENT_RANK , data= fulldata)
summary(employ_model)

plot(fulldata$EMPLOYMENT_RANK, fulldata$RATE_per_1000logged, pch= 20,
     main= "Employment rank to logged crime rate",
     xlab= "Employment rank",
     ylab= "Crime rate per 1000 logged")
abline(employ_model, col= "red")



# Environment----------------------------------------
tm_shape(fulldata) +
  tm_polygons("ENVIRONMENT_RANK", title = "Environment IMD rank",
              palette = "Greens",legend.hist = TRUE, labels =
                c("Lowest Environmental quality ", "Low Environmental quality ",
                  "Moderate Environmental quality", "High Environmental quality",
                  "Highest Environmental quality"),
              style = "fisher") +
  tm_shape(fulldata) +

  tm_borders(lwd = 0.2, col = "#4C4E52")+
  tm_compass(position = c(0.9, 0.9)) +
  tm_scale_bar(position = c(0.01, 0.016))+
  tm_legend(position = c("left", "top"), title.size = 0.6)+
  tm_layout(main.title = "Environment IMD rank",
            main.title.size = 2.5,
            main.title.position = c("center", "top"),
            legend.outside = TRUE,
            legend.title.size = 2,
            legend.text.size = 1.5,
            legend.hist.width = 2,
            legend.hist.height = 1)+
  tm_graticules()

environ_model= lm(RATE_per_1000logged ~ENVIRONMENT_RANK , data= fulldata)
summary(environ_model)

plot(fulldata$ENVIRONMENT_RANK, fulldata$RATE_per_1000logged, pch= 20,
     main= "Environment rank to logged crime rate",
     xlab= "Environment rank",
     ylab= "Crime rate per 1000 logged")
abline(environ_model, col= "red")
```

```
# Barriers -------------------------------------------------------------------
tm_shape(fulldata) +
  tm_polygons("BARRIERS_RANK", title = "Barriers IMD rank",
            palette = "Greens",legend.hist = TRUE, labels =
              c("Very high Barriers ", "High Barriers ", "Moderate Barriers",
                "Low Barriers", "Very low Barriers"),
            style = "fisher") +
  tm_shape(fulldata) +

  tm_borders(lwd = 0.2, col = "#4C4E52")+
  tm_compass(position = c(0.9, 0.9)) +
  tm_scale_bar(position = c(0.01, 0.016))+
  tm_legend(position = c("left", "top"), title.size = 0.6)+
  tm_layout(main.title = "Barriers IMD rank",
          main.title.size = 2.5,
          main.title.position = c("center", "top"),
          legend.outside = TRUE,
          legend.title.size = 2,
          legend.text.size = 1.5,
          legend.hist.width = 2,
          legend.hist.height = 1)+
  tm_graticules()

Barriers_model= lm(RATE_per_1000logged ~BARRIERS_RANK, data= fulldata)
summary(Barriers_model)

plot(fulldata$BARRIERS_RANK, fulldata$RATE_per_1000logged, pch= 20,
     main= "Barriers to housing and services rank to logged crime rate",
     xlab= "Barriers rank",
     ylab= "Crime rate per 1000 logged")
abline(Barriers_model, col= "red")



#Univariate models of independent variables. ----------------------------------

stargazer( employ_model,Barriers_model,environ_model, type= "html",
          dep.var.labels = c("Logged crime rate per 1000"),
          title= "Univariate linear regression models of independant varibales ",
          digits=7,
          out= "Univariatemodels.html")




# Initial multiple linear regression model ------------------------------------
```

2020209

```
model_1 = lm(RATE_per_1000logged~
EMPLOYMENT_RANK+BARRIERS_RANK+ENVIRONMENT_RANK,
        data=fulldata)
summary(model_1)


# checking model performance -----------------------------------------------
fulldata$residuals_1 <- resid(model_1)
hist(fulldata$residuals_1)
fulldata$fitted1 = fitted.values(model_1)
hist(fulldata$fitted1)
fulldata$cooks1 = cooks.distance(model_1)
hist(fulldata$cooks1)


plot(model_1)
shapiro.test(fulldata$residuals_1)
shapiro.test(fulldata$fitted1)
shapiro.test(fulldata$cooks1)
bptest(model_1)




# Mapping residuals ----------------------------------------------------------
tm_shape(fulldata) +
  tm_polygons("residuals_1", title = "Initial model residuals",
          palette = "RdBu",legend.hist = TRUE, labels =
            c("Very high negative ", "High negative ", "Zero",
              "High positive", "Very high positive"),
          style = "fisher") +
  tm_shape(fulldata) +

  tm_borders(lwd = 0.2, col = "#4C4E52")+
  tm_compass(position = c(0.9, 0.9)) +
  tm_scale_bar(position = c(0.01, 0.016))+
  tm_legend(position = c("left", "top"), title.size = 0.6)+
  tm_layout(main.title = "Initial model residuals",
        main.title.size = 2.5,
        main.title.position = c("center", "top"),
        legend.outside = TRUE,
        legend.title.size = 2,
        legend.text.size = 1.5,
        legend.hist.width = 2,
        legend.hist.height = 1)+
  tm_graticules()
```

```
#accounting for autocorrelation ------------------------------------------------

#making geometirc data
fulldata$Centroids <- st_centroid(fulldata$geometry)

residuals = fulldata$residuals_1
# looking at the 100 closest values to each point
knear100 <- knearneigh(x = fulldata$Centroids, k = 100)


r <- sapply(1:100, function(i){
  cor(residuals, residuals[knear100$nn[,i]]) })

# this plots the correlation between values and their neighbours of k distance.

plot(x = 1:100, y = r, xlab = "kth nearest neighbour",
     ylab = "Correlation, r", type = "l" )

lines(smooth.spline(x = 1:100, y = r, lambda = 0.001), col = "red", lty = 2)

# the plot is showing that after around 27th nearest neighbors (line below)
# there is less spatial autocorrelation

abline(v= 27, lty= 3, col= "blue")

# we identify 27 as the number of neighbors to weight more heavily
knear27 = knearneigh(x= fulldata$Centroids, k=27)

# converting to spatial
coords27 = coordinates(as_Spatial(fulldata))

# nearest neighbors of spatialised data
knn27 = RANN::nn2(data = coords27, query = coords27, k = 27)

distances = knn27$nn.dists[,-1]

glist =lapply(1:nrow(distances), function(i) {
  max.d = distances[i,26]
  exp(-0.5*distances[i,]^2 / max.d^2)
})
```

2020209

```
knear = knearneigh(as_Spatial(fulldata$Centroids), k = 26)
knearnb = knn2nb(knear)
gau_w = nb2listw(knearnb, glist = glist)

# running a Morans I test with the guasian weighting of 27 in order to make the
#  suit the data in a way where values far from where they are correlated
# influence the results.
moran.test(x = fulldata$residuals_1, listw = gau_w)
# 0.1141


# Cooks distance---------------------------------------------------------------
#influential out-liers identified and removed.

 # finding the cooks distance of the model
Cooks = cooks.distance(model_1)
mean(Cooks)
Cooks_omit = 4.5 *mean(Cooks)


qqnorm(fulldata$residuals_1, main = "Residuals removed for final model",
      col = ifelse(cooks > cooks_omit, "red", "black"), cex=0.5, pch = 4)
qqline(fulldata$residuals_1, col = "grey")
legend("topleft", legend = c("Removed residuals", "Final model remaining residuals"),
      pch = c(4, 4), col = c("red", "black"))


Model_2= lm(RATE_per_1000logged
~EMPLOYMENT_RANK+BARRIERS_RANK+ENVIRONMENT_RANK,
      data= fulldata, subset = abs(Cooks)<Cooks_omit)
plot(Model_2)
summary(Model_2)

subset_data <- subset(fulldata, abs(Cooks) < Cooks_omit)
subset_data$residuals_2 <- resid(Model_2)
shapiro.test(subset_data$residuals_2)

fitted2 = fitted.values(Model_2)
hist(fitted2)
shapiro.test(fitted2)

cooks2 = cooks.distance(Model_2)
hist(cooks2)
shapiro.test(cooks2)

bptest(Model_2)
```

2020209

```
# this is the same code as above for the new subsetted dataset
# renaming variables from previous code.
subset_data$Centroids = st_centroid(subset_data$geometry)
residuals2 = subset_data$residuals_2

# we identify 27 as the number of neighbors to weight more heavily
knear272 = knearneigh(x= subset_data$Centroids, k=27)
# converting to spatial
coords272 = coordinates(as_Spatial(subset_data))
# nearest neighbors of spatialised data
knn272 = RANN::nn2(data = coords272, query = coords272, k = 27)
distances2 = knn272$nn.dists[,-1]
glist2 = lapply(1:nrow(distances2), function(i) {
  max.d =distances2[i,26]
  exp(-0.5*distances2[i,]^2 / max.d^2)
  })
knear2 = knearneigh(as_Spatial(subset_data$Centroids), k = 26)
knearnb2 = knn2nb(knear2)
gau_w2 = nb2listw(knearnb2, glist = glist2)

# running a morans I test with the guasian weighting of 27 in order to make the
# morans test suit the data better
moran.test(x = subset_data$residuals_2, listw = gau_w2)
#0.0654

tm_shape(subset_data) +
          tm_polygons("residuals_2", title = "Improved model residuals",
 palette = "RdBu",legend.hist = TRUE, labels =
  c("Very high negative ", "High negative ", "Zero", "High positive",
    "Very high positive"),
  style = "fisher") +
  tm_shape(subset_data) +


          tm_borders(lwd = 0.2, col = "#4C4E52")+
          tm_compass(position = c(0.9, 0.9)) +
          tm_scale_bar(position = c(0.01, 0.016))+
          tm_legend(position = c("left", "top"), title.size = 0.6)+
          tm_layout(main.title = "Improved model residuals",
                main.title.size = 2.5,
                main.title.position = c("center", "top"),
              legend.outside = TRUE,
                legend.title.size = 2,
                legend.text.size = 1.5,
                  legend.hist.width = 2,
                  legend.hist.height = 1)+
          tm_graticules()




# Initial vs Final model -------------------------------------------------------


# Compute AIC for each model
AIC_model_1 = AIC(model_1)
AIC_Model_2 = AIC(Model_2)

# Residual results
shap_resid_1= shapiro.test(fulldata$residuals_1)
shap_fitted_1= shapiro.test(fulldata$fitted1)
shap_cooks_1= shapiro.test(fulldata$cooks1)

shap_resid_2= shapiro.test(residuals2)
shap_fitted_2= shapiro.test(fitted2)
shap_cooks_2= shapiro.test(cooks2)


# Compute Breusch-Pagan test for each model
bptest_model_1 <- bptest(model_1)
bptest_Model_2 <- bptest(Model_2)

# Compute Moran's I test for each model (these values come from earlier code)

Moran_Mod_1= 0.1141
Moran_Mod_2= 0.0654


stargazer(model_1, Model_2,
        type = "html",
        dep.var.labels = c("Logged crime rate per 1000"),
        title = "Multivariate linear regression results",
        digits = 7,
        out = "Initial_vs_Final_model with extras.html",
        add.lines = list(c("AIC", format(AIC_model_1, digits = 3), format(AIC_Model_2, digits =
3)),
                    c("Breusch-Pagan test", format(bptest_model_1$p.value, digits = 3),
format(bptest_Model_2$p.value, digits = 3)),
                    c("Moran's I test", format(Moran_Mod_1, digits = 3), format(Moran_Mod_2,
digits = 3)),
                    c("Residual normality (Shapiro-Wilk)", format(shap_resid_1$statistic, digits = 3), format(shap_resid_2$statistic, digits = 3)),
                    c("Fitted normality (Shapiro-Wilk)", format(shap_fitted_1$statistic, digits = 3), format(shap_fitted_2$statistic, digits = 3)),
                    c("Cooks Distance (Shapiro-Wilk)", format(shap_cooks_1$statistic, digits = 3), format(shap_cooks_2$statistic, digits = 3))))
```