CHEER-UP's FSDS Group Project

Declaration of Authorship

We, CHEER UP Group, confirm that the work presented in this assessment is our own. Where information has been derived from other sources, we confirm that this has been indicated in the work. Where a Large Language Model such as ChatGPT has been used we confirm that we have made its contribution to the final submission clear.

Date:

Student Numbers:

Brief Group Reflection

What Went Well	What Was Challenging
A	B
C	D

Priorities for Feedback

Are there any areas on which you would appreciate more detailed feedback if we're able to offer it?

Response to Questions

1. Who collected the data?

Inside Airbnb was founded by Murray Cox who conceived the project, and keeps updating and analyzing Airbnb data (*About Inside Airbnb*, no date).

2. Why did they collect it?

Murray Cox and his partners aim to make Airbnb's data more transparent and accessible to the public. In addition to making visual dashboards of different cities, they also wrote many reports based on these data. For example, they criticized Airbnb's New York data as misleading (Cox and Slee, 2016).

For consumers, they can understand market trends more easily, which help them make more informed decisions and choose accommodation that fits their needs and budget; For landlords, transparency of information will promote fair competition, and maintain the stability of short-term rental market prices; For academics, they can use these data to explore how Airbnb relates to many factors in urban development; For the government, they can understand the current situation of the short-term rental market and try to control the phenomenon that is not conducive to the stable development of the city by establishing more supervision regulation.

3. How was the data collected?

Inside Airbnb uses web scraping scripts to extract publicly available information from Airbnb's website on a quarterly basis. These scripts navigate the Airbnb site, accessing various pages to retrieve data such as listing titles, descriptions, hosts, pricing, and more. The collected data is then verified, cleaned, and made available on the Inside Airbnb website for users to explore online or download for analysis.

In addition to providing a snapshot, Inside Airbnb conducts in-depth analyses using this raw data to derive more insightful information for users. For instance, the platform employs a proprietary "San Francisco Model" to estimate the frequency with which an Airbnb listing is rented out and to approximate a listing's income.

4. How does the method of collection impact the completeness and/or accuracy of its representation of the process it seeks to study, and what wider issues does this raise?

1. Pricing and availability among others are extremely dynamic attributes set by hosts that cannot be captured accurately with web-scraping, which only reflects a snapshot of the website at one specific moment in time.
2. Susceptible to changes in the structure of the website.
3. Web scraping only accesses publicly available listings.
4. Too frequent web scraping may affect site performance which in turns affect the data scraped
5. Has to ethical and legal implications
6. Data capture can only capture real-time AirBnb data, and long-term comparative analysis requires data collection in advance

5. What ethical considerations does the use of this data raise?

First, Airbnb is a profitable company, making their data public without interfering with normal operations makes it easier to regulate them. But as landlords and renters, they have the right to request that their personal information not be made public elsewhere. However, it is not possible to ask everyone's consent when the data is crawled. Therefore, using this data may be needed to handle sensitive personal information. For example, this could have implications for home privacy and home security.

Second, as the data is simply extracted and shown, users can only see the surface and cannot understand the context and logic of it. It may cause inaccuracies usage of data, which can be misused and misguide public opinion (D'Ignazio and Klein, 2020). For example, the rental activities of a particular group may be negatively analyzed, which will reinforce the stereotype of this group and then trigger discrimination and antagonism mood of society.

Third, data settings are always influenced by power (D'Ignazio and Klein, 2020). Airbnb often leaves out reviews that aren't friendly to properties and focuses on those that are profitable, such as setting its recommendation algorithm to favor partners. Therefore, users need to be fair when using such data.

6. With reference to the data (*i.e.* using numbers, figures, maps, and descriptive statistics), what does an analysis of Listing types suggest about the nature of Airbnb lets in London?

Background and research question

Airbnb provides indispensable accommodation for the development of tourism. As a short-term rental platform, the mobility of tenants is one of its major characteristics. Mobility is a factor in crime (Mburu and Helbich, 2016), so recently there has been increasing concern about whether Airbnb, which brings high mobility to the local community, could be linked to an increase or decrease in crime. For example, Xu, Pennington-Gray and Kim (2019) studied the relationship between Airbnb density and crime in Florida and found that certain types of listings do have a significant impact on crime. Therefore, we try to explore whether Airbnb in London could also have an impact on crime based on Lower Layer Super Output Areas (LSOAs) level. The reason why we choose LSOA level as this allows for a more detailed examination of localized patterns and variations within a borough. Different neighborhoods within a borough can exhibit distinct characteristics that might be overlooked in a broader borough-level analysis as crime rates and housing patterns can vary significantly from one neighborhood to another. Finally, our research question is: Whether the type of Airbnb listing have an impact on crime in the neighbourhood?

Data source

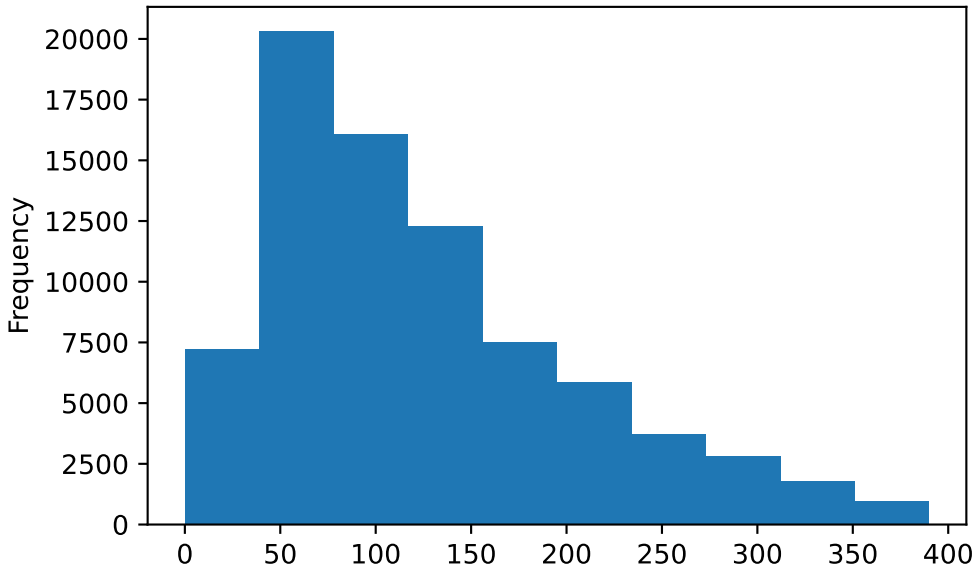
The Airbnb dataset is from the September 2023 updated version provided by Inside Airbnb. The crime dataset is from data.police.uk and has the same time dimension as Airbnb's. The spatial partitioning dimension in this study is LSOA and is from the Office of National Statistics website.

Defining the data

In order to examine the nature of airbnb and crime relationship, we want to first establish the parameters to focus on a subset of the datasets:

As Xu, Pennington-Gray, and Kim (2019) mentioned different types of listings have different impacts, we refer to their classification standards to divide Airbnb housing into private rooms, shared rooms and entire homes, which will be treated as independent variables. Each variable is defined as follows (*Data Dictionary*, nd):
a) Entire home: Tenants can enjoy an entire whole listing.
b) Private room: The tenant has a private bedroom but may need to share some space with others such as the kitchen.
c) Shared room: The tenant needs to share the bedroom and other spaces with others.
When preprocessing Airbnb listings data, we excluded hotel rooms, which are concentrated rather than spread across different residential buildings and have less of an impact on local neighbors. It is worth mentioning that we trim listing price outliers to control for potentially bogus listings.

```
<Axes: ylabel='Frequency'>
```

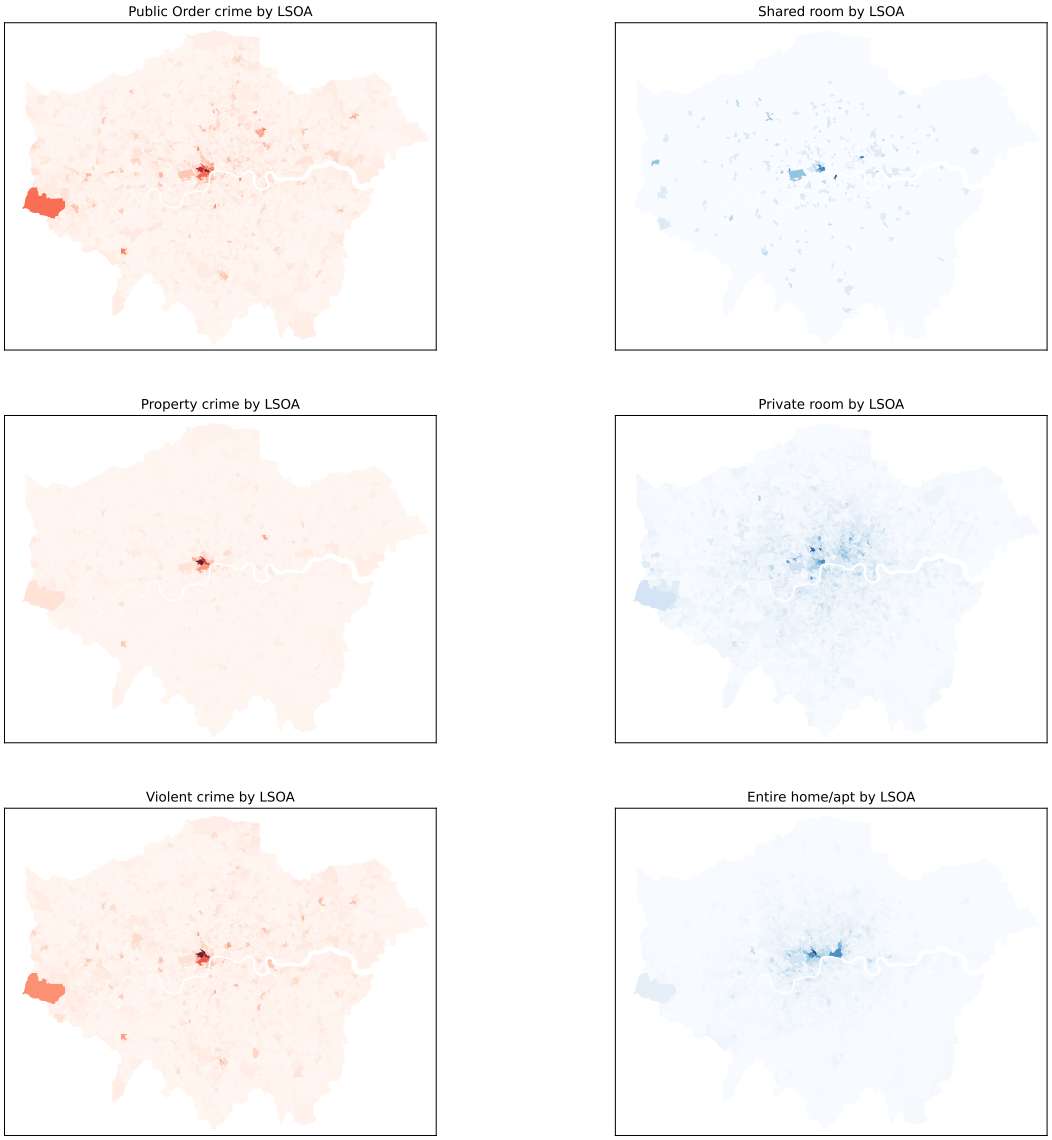


As for crime data, as we want to make recommendations more specific to different types and levels of crime, we classify crime into the following types (Xu, Pennington-Gray and Kim, 2019; Flatley, 2016): - Public order: public order, drugs, possession of weapons, anti-social behaviour. - Violent (involve force): robbery, violence and sexual offences. - Property (without force): bicycle theft, burglary, criminal damage and arson, shoplifting, theft from the person, vehicle cri These three types of crime will be treated as dependent variables.e. e.

The spatial visualization of all variables

1. Spatial distribution of each variable

In our spatial analysis of crime and housing patterns across LSOAs, distinct geographic concentrations emerge. Public order crimes are dispersed citywide, with notable pockets in Westminster and Hillingdon. Property crimes concentrate in central London, particularly in Westminster, while violent crimes exhibit a similar pattern, with Westminster and Hillingdon standing out. Concurrently, the distribution of shared rooms reveals widespread dispersion, with heightened concentrations in Hillingdon, Westminster, Tower Hamlets, and Southwark. Private rooms and entire home exhibit citywide dispersal, with Westminster and Camden featuring prominently in private room concentrations, and the City of London emerging as a hub for entire home. These spatial insights provide valuable data for informed decision-making in law enforcement, housing policies, and urban planning throughout the diverse neighbourhoods of London.



2. Relationship between Airbnb listings and crime:

In order to have a sense of the relationship between different airbnb listing type and crime, we plotted kernel density estimation (KDE) analysis heat map and correlation

matrix.

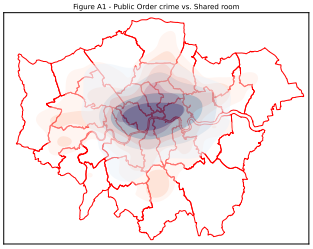
a. KDE analysis

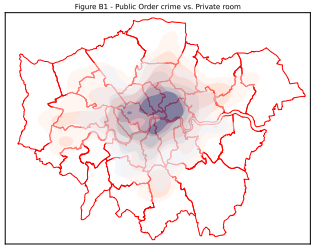
By categorising crime types and Airbnb room types, we do kernel density analysis of different types of variables spatially with simple spatial superposition, in order to explore the relationship between the spatial distribution of different room types and different crime types.

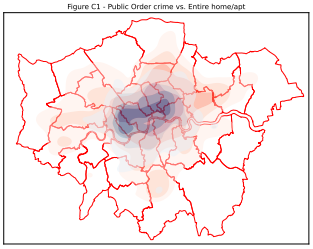
The Shared room shows spatial clustering centered on central of London and radiates mainly in the east-west direction. Figure A, Shared room and three types of crime spatial density overlay can be found, they all overlap in the central of London area, public crime and violent crime spatial density range is greater than shared room. while property crime and spatial density is less than Shared room.

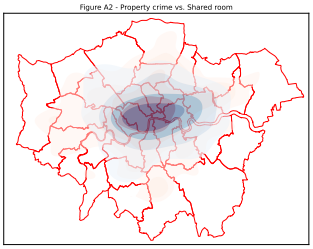
The Private room spatially shows clustering in the north-east direction of central of London, showing a semi-circular ring. Figure B, Private room with crime type spatial overlay in the middle direction, they both overlap in the north-west direction of central of London.

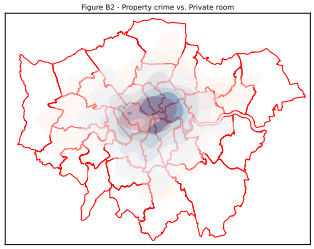
Entire home shows a spatial density that is striped across the central of London area. Figure C, overlaps with the central of London area north of the Thames for each offence type.

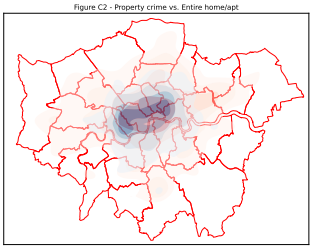


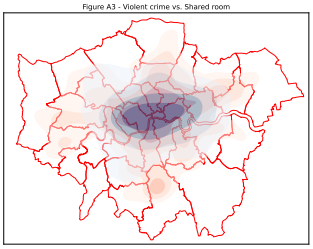


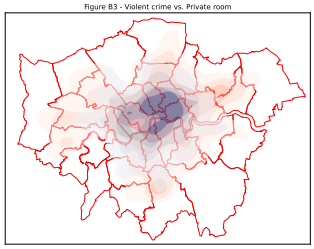


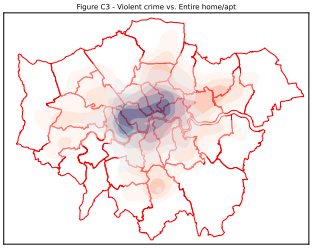




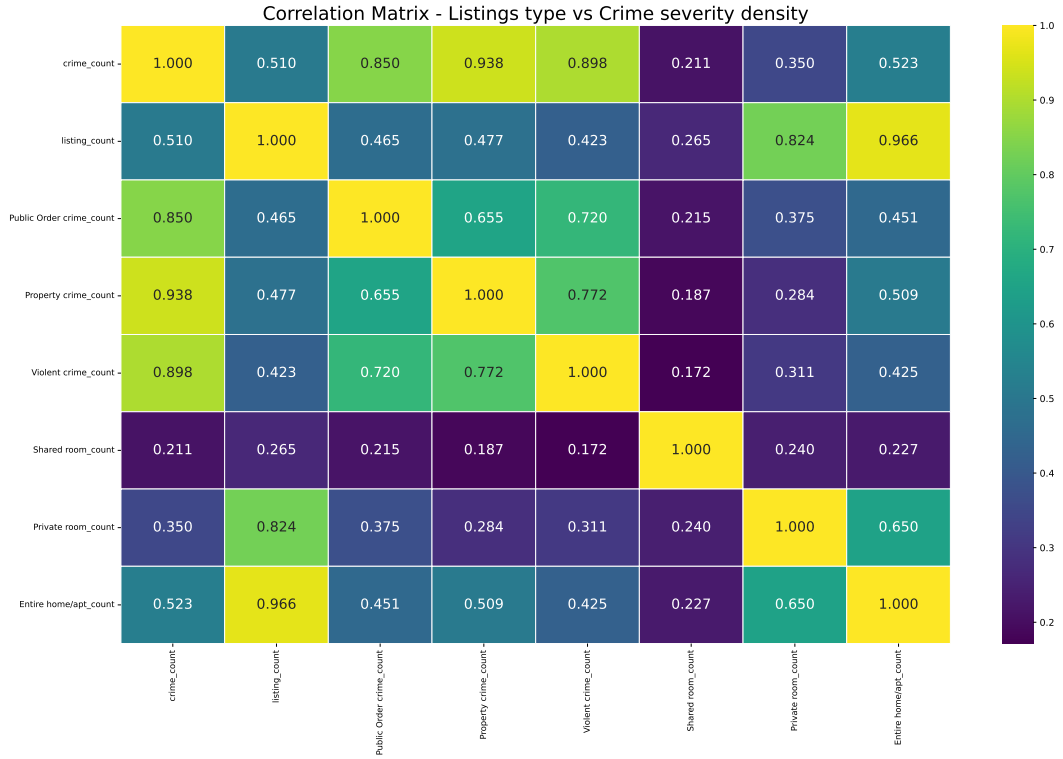








b. Correlation analysis



From the correlation matrix, the correlation coefficient between each Airbnb room type and each crime type is greater than 0, which means that they are positively correlated. The correlation between entire home and property crime is the strongest, with a correlation coefficient of 0.509. On the contrary, the correlation between shared room and violent crime is the weakest, with a correlation coefficient of 0.173.

3. Visualising correlation on a map (not visible on PDF file, please view the HTML file on github)

When looking at the spatial distribution of all Airbnb listings and crimes, it is not surprising to see that the areas that are brown in colour, which means there are a large number of both Airbnb listings and crimes, are mostly concentrated in the central of London such as Westminster. Unexpectedly, an area in Hillington shows a high concentration, which has 198 criminal records and 55 listings. Similarly, the colour of an area of Kingston upon Thames also is brown, with 224 criminal records and 41 listings. These areas received little attention before.

As the entire home and property crime show the strongest correlation, we pay separate attention to their correlation spatial distribution. From this map, we can see that the situation is similar to the previous one, in which the problem of an area in Hillington and Kingston upon Thames respectively still exists.

Since we now find that all of the independent variables we set are correlated with the dependent variables, and correlation doesn't mean causation, we would like to do a further regression analysis to determine whether different types of Airbnb listings have an impact on different types of crime.

7.

[Regression explanation of all the dependent and independent variables]

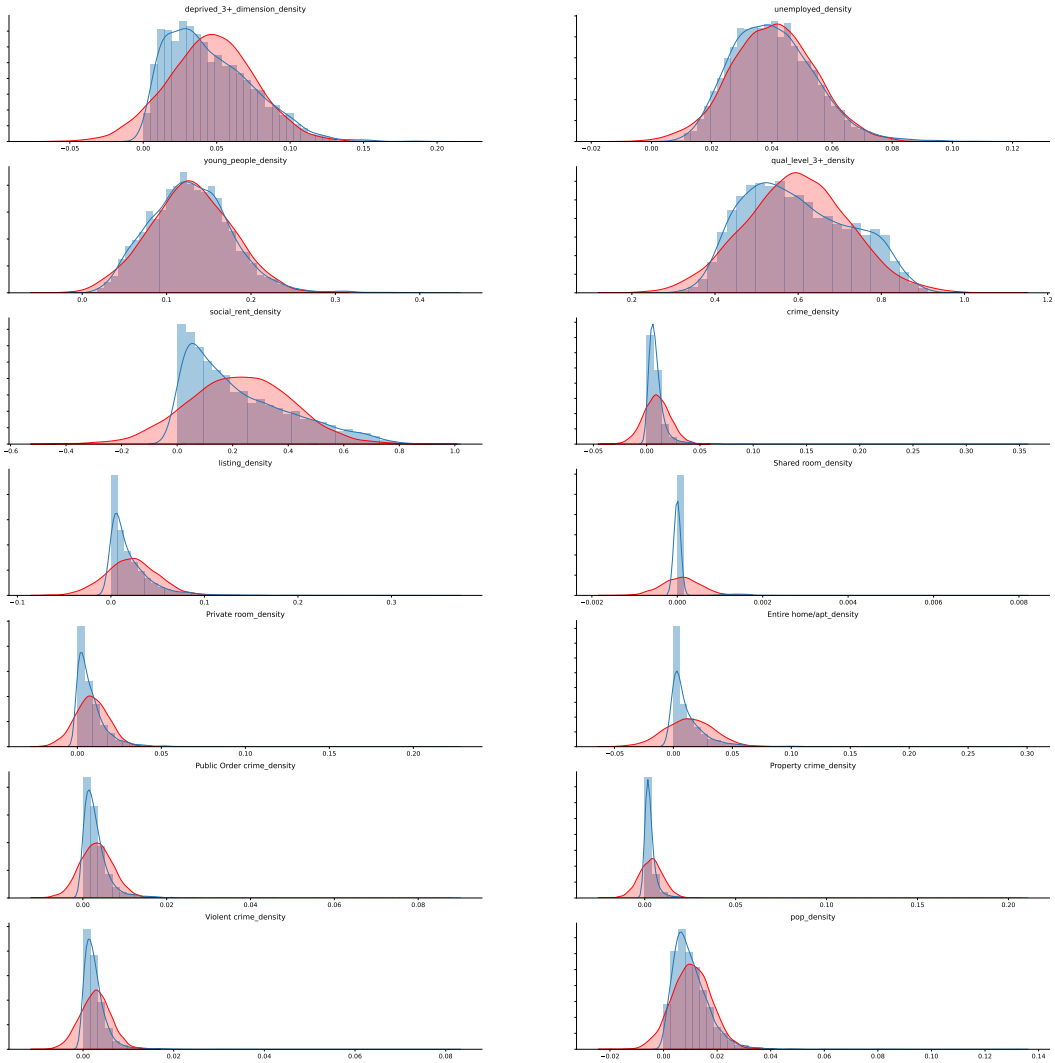
Chosen columns: X

- Population density: Usual resident (11) / Area (need to get)
- Bad+Very Bad Health (120+121) / Pop (11)
- deprived 2+ dimension (144+145+146) / Pop (11)
- Unemployed (72) / Pop (11)
- Young people (17+18+19) / Pop (11)
- 3x Airbnb listings (3 types) / Household number (3)

Y

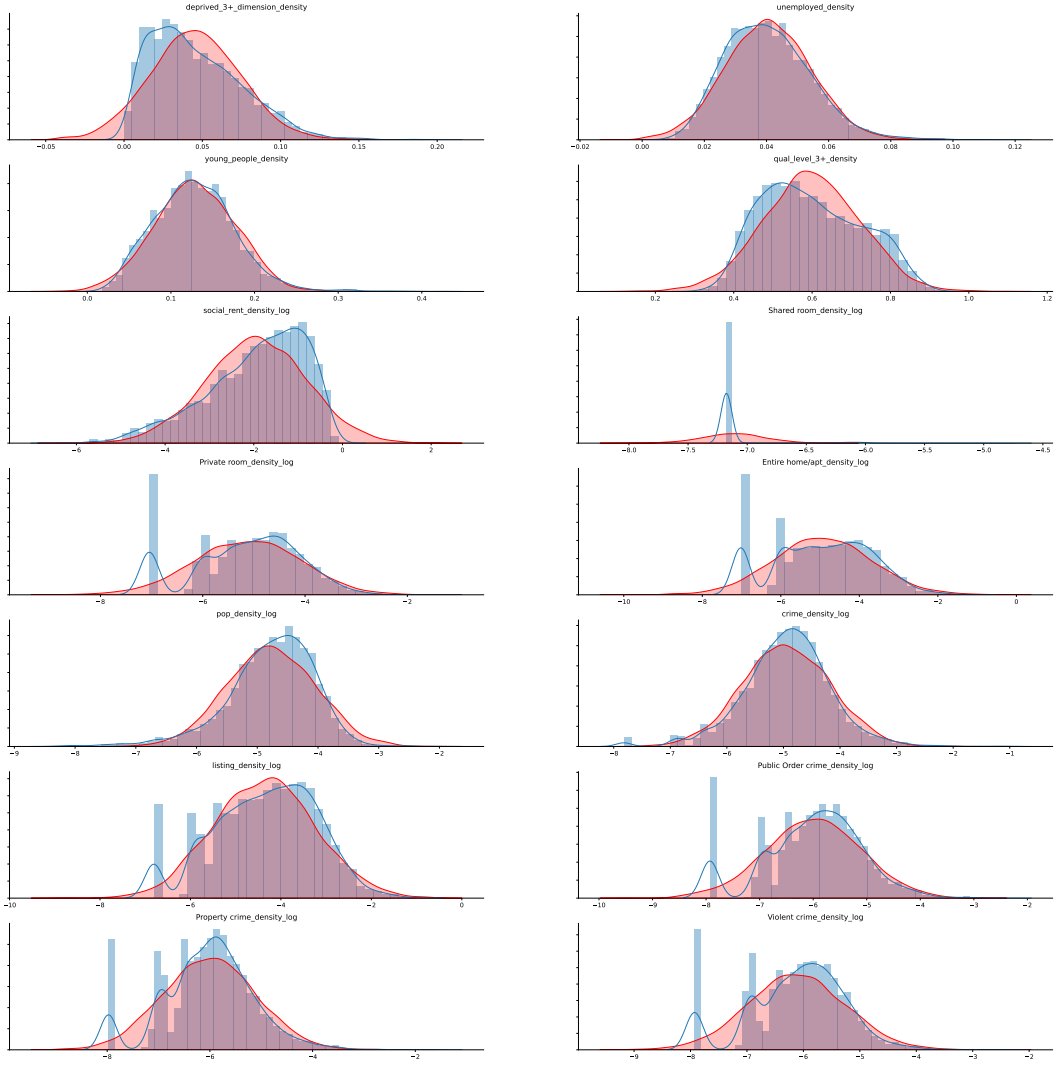
- Crime-relevant / Pop (11)

Preparing Regression Data



Need log transformation for - Crime density - Property crime density - Violent crime density - Public order crime density - Listing density - Social rent density - Shared room density - Private room density - Entire home/apt density - population density

```
Dropping: listing_density_log
```



Regression Model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      crime_density_log   R-squared:                       0.282
Model:                            OLS   Adj. R-squared:                  0.280
Method:                 Least Squares   F-statistic:                     217.0
Date:                Sun, 17 Dec 2023   Prob (F-statistic):               0.00
Time:                        06:57:34   Log-Likelihood:                -4969.9
No. Observations:                4994   AIC:                             9960.
Df Residuals:                    4984   BIC:                         1.002e+04
Df Model:                           9
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           -4.8298      0.324    -14.909      0.000      -5.465      -4.195
deprived_3+_dimension_density    1.6586      0.607      2.731      0.006       0.468       2.849
```

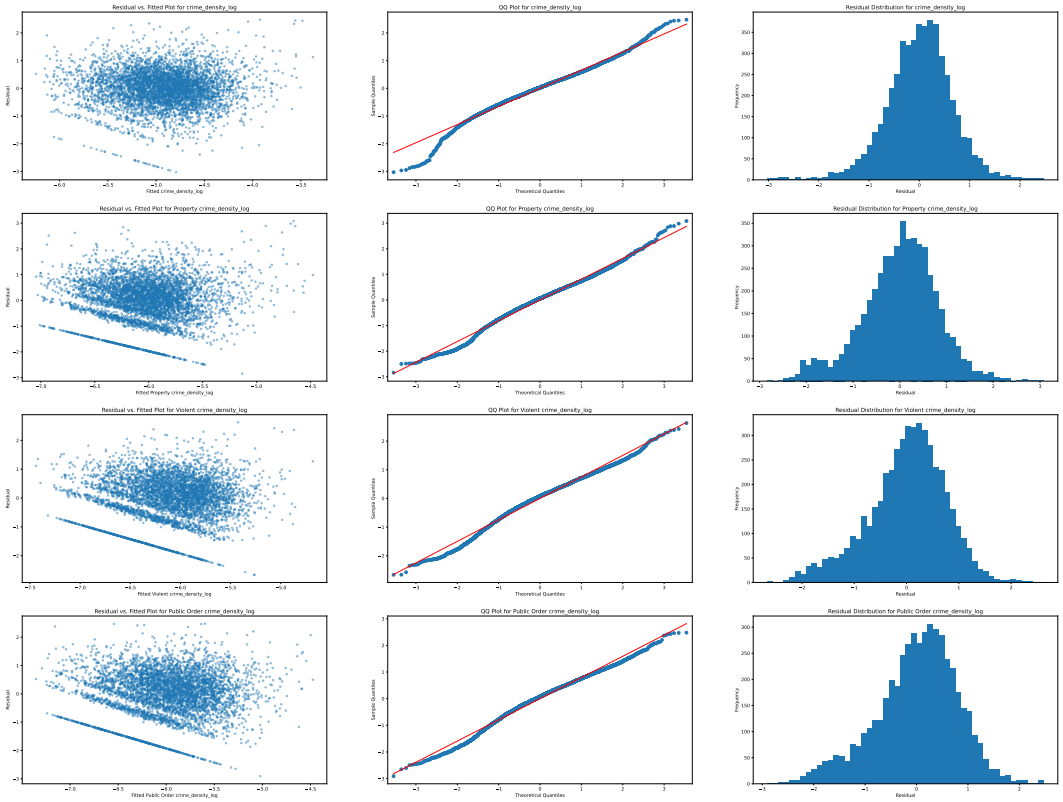
```
unemployed_density           3.9537    0.991    3.990    0.000     2.011     5.896
young_people_density         3.6913    0.230   16.058    0.000     3.241     4.142
qual_level_3+_density        -0.1423    0.128   -1.113    0.266    -0.393     0.108
social_rent_density_log       0.1555    0.014   11.270    0.000     0.128     0.183
Shared room_density_log       0.1040    0.037    2.823    0.005     0.032     0.176
Private room_density_log      0.0369    0.012    2.970    0.003     0.013     0.061
Entire home/apt_density_log    0.1635    0.013   13.038    0.000     0.139     0.188
pop_density_log            -0.2744    0.016  -17.332    0.000    -0.305    -0.243
==============================================================================
Omnibus:                      279.098   Durbin-Watson:                   1.810
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              624.459
Skew:                          -0.360   Prob(JB):                     2.51e-136
Kurtosis:                       4.576   Cond. No.                      1.25e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.25e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================================
Dep. Variable:    Property crime_density_log   R-squared:                       0.160
Model:                             OLS   Adj. R-squared:                  0.158
Method:                  Least Squares   F-statistic:                     105.4
Date:                 Sun, 17 Dec 2023   Prob (F-statistic):           2.72e-181
Time:                         06:57:34   Log-Likelihood:                -6045.0
No. Observations:                4994   AIC:                         1.211e+04
Df Residuals:                    4984   BIC:                         1.218e+04
Df Model:                           9
Covariance Type:                nonrobust
==============================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                    -6.8910      0.402    -17.152      0.000      -7.679      -6.103
deprived_3+_dimension_density    1.6432      0.753      2.182      0.029      0.167      3.120
unemployed_density           2.9141      1.229      2.371      0.018      0.505      5.323
young_people_density         3.3643      0.285     11.801      0.000      2.805      3.923
qual_level_3+_density        0.4425      0.159      2.791      0.005      0.132      0.753
social_rent_density_log       0.0796      0.017      4.653      0.000      0.046      0.113
Shared room_density_log       0.0578      0.046      1.264      0.206     -0.032      0.147
Private room_density_log      0.0048      0.015      0.309      0.757     -0.025      0.035
Entire home/apt_density_log    0.1824      0.016     11.729      0.000      0.152      0.213
pop_density_log            -0.3079      0.020    -15.679      0.000     -0.346     -0.269
==============================================================================
Omnibus:                      96.208   Durbin-Watson:                   1.814
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              114.365
Skew:                          -0.281   Prob(JB):                     1.47e-25
Kurtosis:                       3.484   Cond. No.                      1.25e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
[2] The condition number is large, 1.25e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:    Violent crime_density_log  R-squared:                       0.221
Model:                            OLS   Adj. R-squared:                  0.220
Method:                 Least Squares   F-statistic:                     157.5
Date:                Sun, 17 Dec 2023   Prob (F-statistic):          4.21e-263
Time:                        06:57:34   Log-Likelihood:                -5648.5
No. Observations:                4994   AIC:                         1.132e+04
Df Residuals:                    4984   BIC:                         1.138e+04
Df Model:                           9
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                            -5.8310      0.371    -15.713      0.000      -6.559      -5.104
deprived_3+_dimension_density     0.2802      0.696      0.403      0.687      -1.084       1.644
unemployed_density                4.9460      1.135      4.357      0.000       2.721       7.171
young_people_density              3.9444      0.263     14.979      0.000       3.428       4.461
qual_level_3+_density            -0.7365      0.146     -5.030      0.000      -1.024      -0.449
social_rent_density_log           0.1931      0.016     12.211      0.000       0.162       0.224
Shared room_density_log           0.0828      0.042      1.960      0.050      -2e-05       0.166
Private room_density_log          0.0477      0.014      3.354      0.001       0.020       0.076
Entire home/apt_density_log       0.1004      0.014      6.991      0.000       0.072       0.129
pop_density_log                  -0.2405      0.018    -13.260      0.000      -0.276      -0.205
==============================================================================
Omnibus:                      164.896   Durbin-Watson:                   1.927
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              184.526
Skew:                          -0.434   Prob(JB):                     8.53e-41
Kurtosis:                       3.367   Cond. No.                     1.25e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.25e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:    Public Order crime_density_log  R-squared:                       0.231
Model:                            OLS   Adj. R-squared:                  0.229
Method:                 Least Squares   F-statistic:                     166.1
Date:                Sun, 17 Dec 2023   Prob (F-statistic):          5.17e-276
Time:                        06:57:34   Log-Likelihood:                -5943.7
No. Observations:                4994   AIC:                         1.191e+04
Df Residuals:                    4984   BIC:                         1.197e+04
Df Model:                           9
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                            -5.0519      0.394    -12.832      0.000      -5.824      -4.280
```

```
deprived_3+_dimension_density    2.4469     0.738     3.316     0.001      1.000      3.894
unemployed_density               3.2879     1.204     2.730     0.006      0.927      5.649
young_people_density             3.6205     0.279    12.960     0.000      3.073      4.168
qual_level_3+_density            -0.3211    0.155    -2.067     0.039     -0.626     -0.017
social_rent_density_log           0.1501    0.017     8.947     0.000      0.117      0.183
Shared room_density_log           0.1515    0.045     3.383     0.001      0.064      0.239
Private room_density_log          0.0513    0.015     3.398     0.001      0.022      0.081
Entire home/apt_density_log       0.1671    0.015    10.966     0.000      0.137      0.197
pop_density_log                  -0.2096    0.019   -10.895     0.000     -0.247     -0.172
==============================================================================
Omnibus:                      153.954   Durbin-Watson:                   1.856
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              168.081
Skew:                          -0.435   Prob(JB):                     3.17e-37
Kurtosis:                       3.222   Cond. No.                     1.25e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.25e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Comment

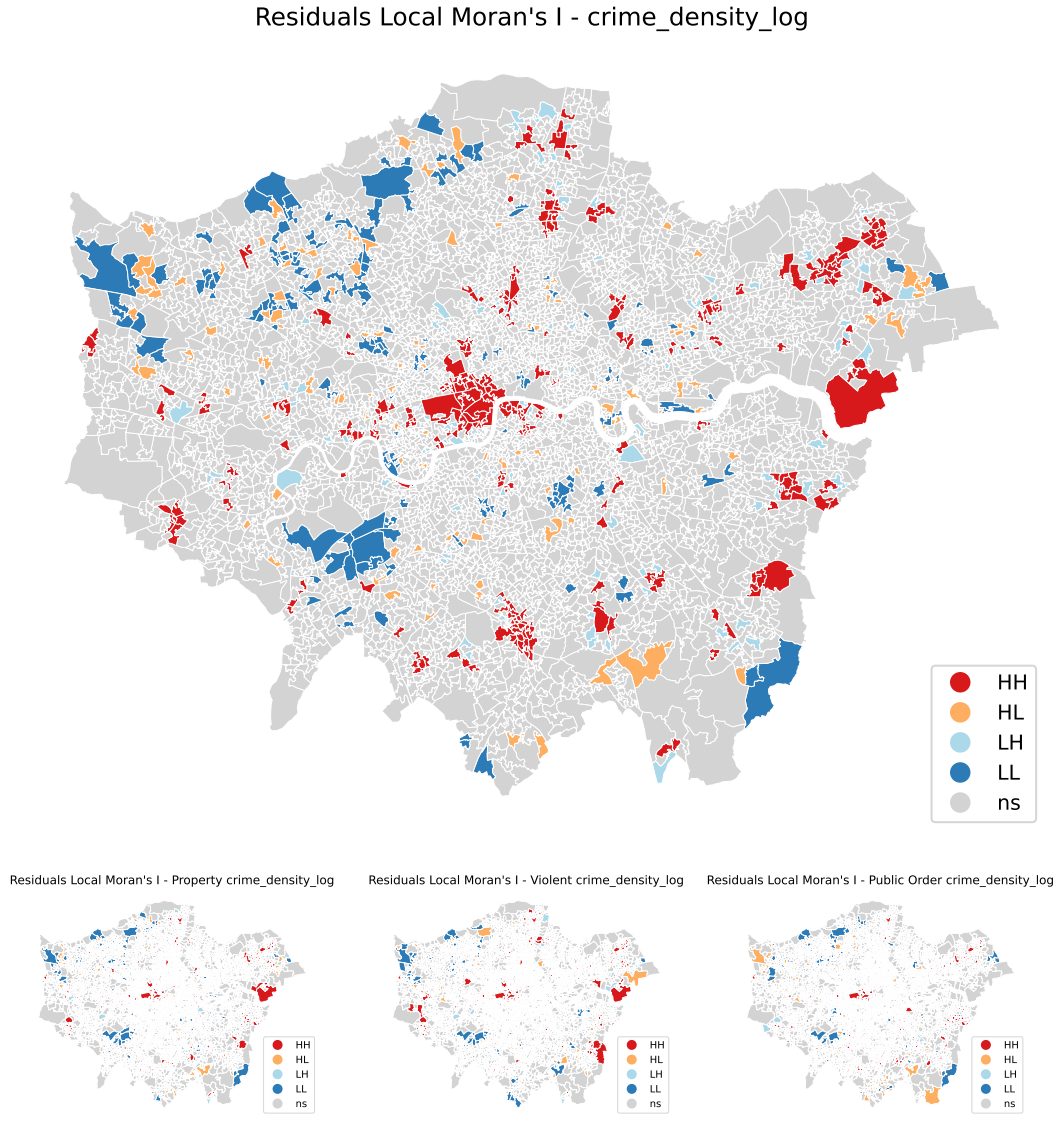
Residual Analysis to check for OLS assumptions



Comment

Residual Moran's I to check if there is Spatial Lag within the Residuals (spatial autocorrelation unaccounted for in the model)

```
Residual Moran's I - crime_density_log: 0.13749960376291004
Residual Moran's I - Property crime_density_log: 0.13749960376291004
Residual Moran's I - Violent crime_density_log: 0.13023456620902985
Residual Moran's I - Public Order crime_density_log: 0.083030538039212
```

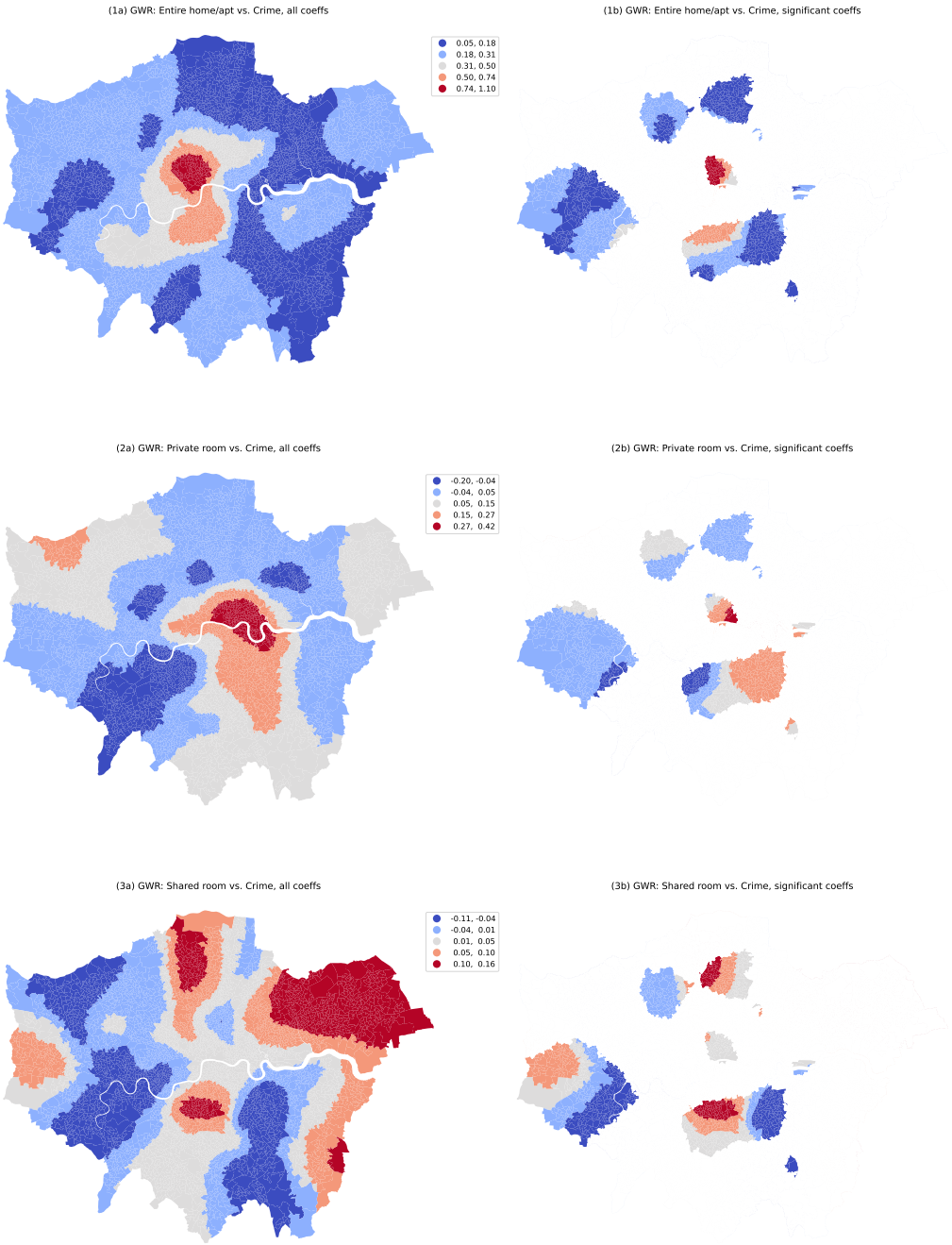



```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIALLY WEIGHTED LEAST SQUARES (HET)
-----------------------------------------------------------
Data set            :     unknown
Weights matrix      :     unknown
Dependent Variable  :crime_density_log          Number of Observations:      4994
Mean dependent var  :    -4.9364            Number of Variables   :        10
S.D. dependent var  :     0.7723            Degrees of Freedom    :      4984
Pseudo R-squared    :      0.2803
```

```
N. of iterations    :         1             Step1c computed      :        No

------------------------------------------------------------------------------
         Variable    Coefficient     Std.Error    z-Statistic    Probability
------------------------------------------------------------------------------
         CONSTANT     -5.4454587     0.3255627    -16.7262974     0.0000000
deprived_3+_dimension_density     1.8047169     0.6012056     3.0018297     0.0026836
  unemployed_density     4.6726235     1.0084322     4.6335523     0.0000036
young_people_density     3.8485811     0.2995302    12.8487255     0.0000000
qual_level_3+_density     0.0794200     0.1450277     0.5476194     0.5839533
social_rent_density_log     0.1475069     0.0146126    10.0945272     0.0000000
Shared room_density_log     0.0725050     0.0367991     1.9702945     0.0488046
Private room_density_log     0.0407930     0.0128976     3.1628293     0.0015624
Entire home/apt_density_log     0.1469145     0.0134250    10.9433410     0.0000000
    pop_density_log    -0.2991048     0.0188976    -15.8276603     0.0000000
            lambda      0.3363186      0.0198388     16.9525449      0.0000000
------------------------------------------------------------------------------
================================ END OF REPORT =====================================
```

Comment

Geographically Weighted Regression to visualize geo variation of the relationship between different airbnb types and crime_density



Comment

[Policy recommendation] TBD. Some ideas (NEED LITERATURE TO BACKUP)

If there are relationships

1. Reactive: Increase police presence in areas with high Airbnb density
2. Proactive: Limit Airbnb density in areas with high crime rates of a certain type
3. Collaborative: Airbnb to share data with local police departments to help them identify areas with high Airbnb density
4. Punitive: Airbnb to pay a fine for every crime incident that occurs in areas with high Airbnb density
5. Preventive: Airbnb to pay for additional security cameras in areas with high Airbnb density

If not: Other cities this but London doesn't see the same issue. Should look at improving crime by improving other aspects…

References

Add reference