Chocolate's Group Project

1. Who collected the InsideAirbnb data?

(2 points; Answer due Week 7)

The Inside Airbnb project, founded in 2014 by Murray Cox, who wanted to dispute Airbnb's claim that 87% of the hosts rent out the place in which they live (Alsudais, 2021). The project has benefited from the John Morris, who designed and refined the user experience, and Samantha Box, who supported coding and analysis while addressing key issues. Team members like Michael "Ziggy" Mintz and Anya Sophe Behn further enhanced automation and cloud migration ('Inside airbnb', n.d.). Operating independently, the project equips policymakers, researchers, and communities to address Airbnb's effects on housing and neighborhoods.

2. Why did they collect the InsideAirbnb data?

(4 points; Answer due Week 7)

The Inside Airbnb project is a mission-driven initiative that provides data and advocacy on Airbnb's impact on residential communities('Inside airbnb', n.d.). The collected data offers a more transparent and critical perspective, enabling stakeholders to assess Airbnb's role in local economies, housing markets, and community dynamics. By analyzing Airbnb listings, the project helps policymakers, researchers, and communities better understand the impact of Airbnb on issues such as housing affordability, availability, gentrification, and displacement.

3. How did they collect it?

(5 points; Answer due Week 8)

Inside Airbnb's data is extracted through web scraping technology from publicly available listing information on the Airbnb website. This process collects key details, including price, location, host information, and availability. To provide accurate and up-to-date insights, the data is regularly updated, with new information replacing previous records on a monthly basis (Prentice and Pawlicz, 2024). Additionally, Inside Airbnb estimates monthly occupancy rates and income based on historical data, providing market analysis, basic hospitality indicators, and information on the legal status of the short-term rental market. At the same time, Inside Airbnb integrates geospatial data through public GIS datasets to map community boundaries and analyze the local impact. The collected data is analyzed, cleaned, and aggregated to facilitate public discussion and support the analysis of short-term rental market trends, helping communities make informed decisions regarding short-term rental regulation.

4. How does the method of collection (Q3) impact the completeness and/or accuracy of the InsideAirbnb data? How well does it represent the process it seeks to study, and what wider issues does this raise?

(11 points; Answer due Week 9)

Since InsideAirbnb provides property information and updates it monthly, it can, to some extent, represent the overall situation of the short-term rental market and assist in analyzing dynamic changes related to the time dimension. However, InsideAirbnb's data still has temporal limitations, and the publicly available data only reflects the content uploaded by hosts, which may contain false information or omissions, such as inaccurate reviews or misleading property descriptions. These issues may cause the data to fail to fully reflect the real market situation, thus affecting research based on InsideAirbnb data and further influencing the formulation of short-term rental market regulatory policies.

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

(18 points; Answer due 2024-12-17)

6. With reference to the InsideAirbnb data (*i.e.* using numbers, figures, maps, and descriptive statistics), what does an analysis of Hosts and the types of properties that they list suggest about the nature of Airbnb lettings in London?

(15 points; Answer due 2024-12-17)

Host Analysis

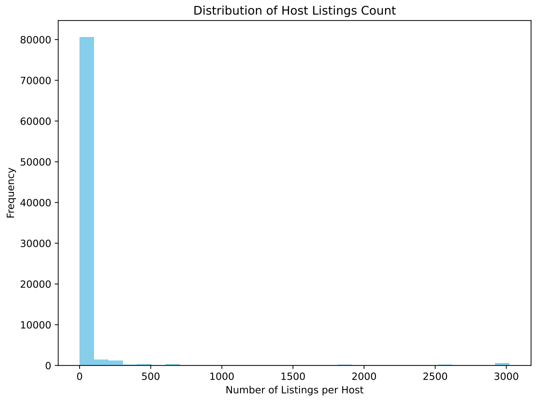
1.Host Listings Distribution:

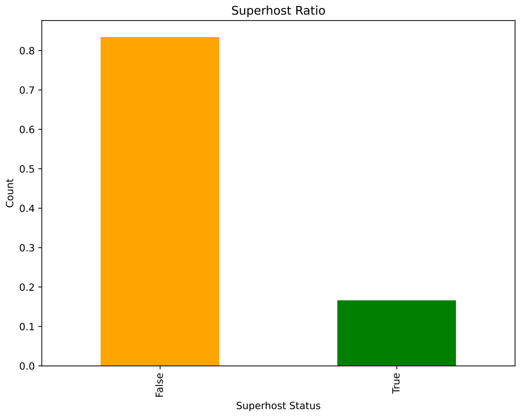
1.The majority of hosts own between 1 to 2 listings, with a few managing hundreds or even thousands of properties.

2.The distribution highlights the presence of professional operators in London's Airbnb market, with the largest host managing over 3000 listings.

2.Superhost Ratio:

Only 16.6% of hosts are Superhosts, suggesting that while a small percentage of hosts maintain high service standards, the majority are casual or infrequent hosts.





Property Type Distribution and Room Type Distribution

Property Type Distribution:

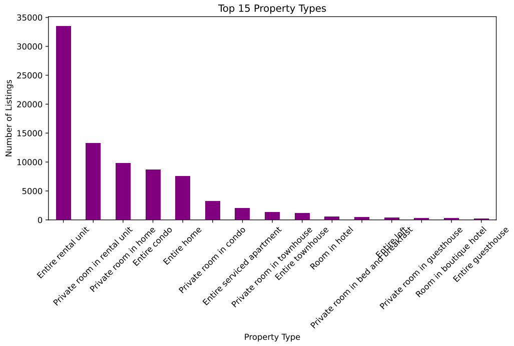
1.The most common property types are “Entire rental units” and “Private rooms in rental units”.

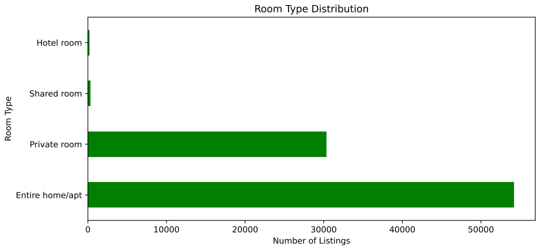
2.This indicates a mix of properties catering to both tourists seeking entire apartments and budget-conscious travelers looking for private rooms.

Room Type Distribution:

1.63.7% of listings are entire homes/apartments, while 35.7% are private rooms.

2.The dominance of entire homes suggests that Airbnb in London is often used for short-term vacation rentals rather than shared accommodation.

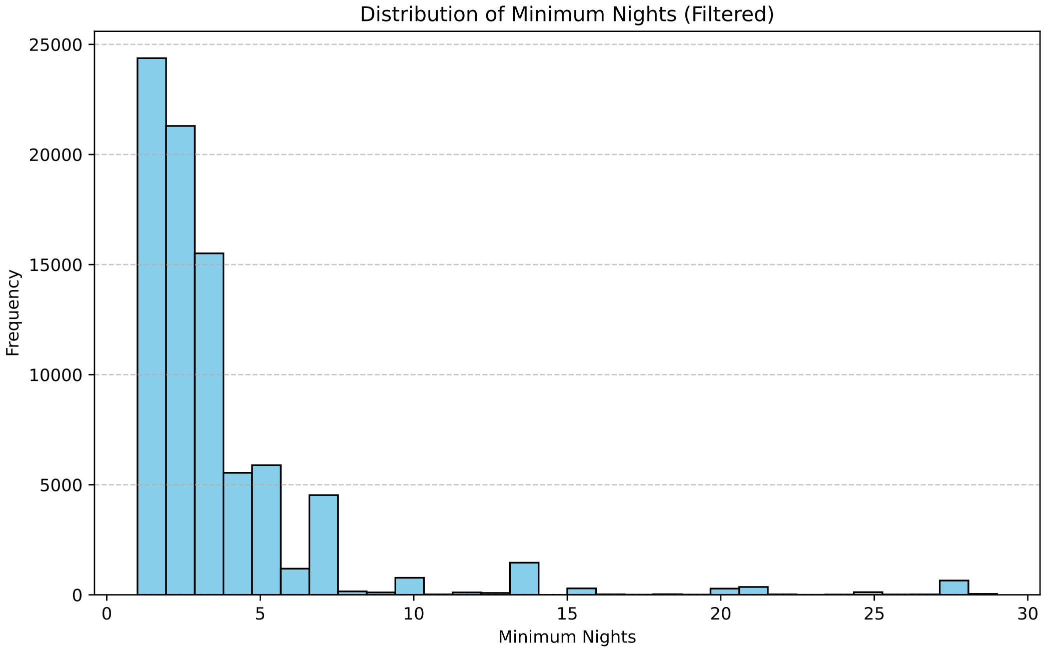




Price and Rental Duration

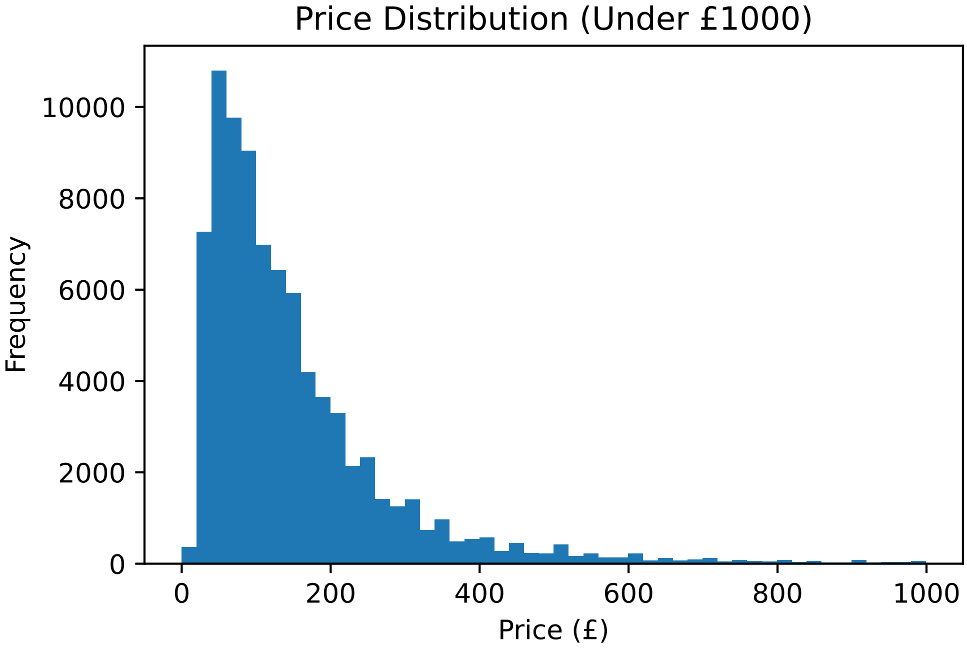
1.The average price of an Airbnb listing is £183, but most listings are priced between £50 and £200.

2.The presence of a few extremely high-priced listings skews the average, indicating luxury rental options are available.



Minimum Nights Distribution

Most properties have a minimum stay requirement of 1 to 5 nights, making Airbnb suitable for short-term rentals.



Geographical Distribution

Airbnb Listings Distribution:

1.Listings are concentrated in central boroughs such as Westminster, Camden, and Kensington.

2.This reflects high tourist demand in central London areas.

Listings by Borough:

1.Central boroughs have the highest density of listings, correlating with higher property prices and demand for short-term accommodation.

2.Outer boroughs have fewer listings, indicating lower demand or fewer hosts in these regions.

```
"\nfig, axes = plt.subplots(1, 2, figsize=(18, 8))\n\n# Airbnb Listing\nborough.plot(ax=a
```

Conclusion

The analysis of hosts and property types in the InsideAirbnb data reveals that Airbnb lettings in London are characterized by:

1.A significant proportion of professional hosts with multiple listings.

2.A dominance of entire home/apartment rentals, catering to tourists and short-term visitors.

3.Concentration in central boroughs with high demand for short-term rentals.

4.Price points that reflect both budget and luxury options.

7. Drawing on your previous answers, and supporting your response with evidence (*e.g.* figures, maps, EDA/ESDA, and simple statistical analysis/models drawing on experience from, e.g., CASA0007), how *could* the InsideAirbnb data set be used to inform the regulation of Short-Term Lets (STL) in London?

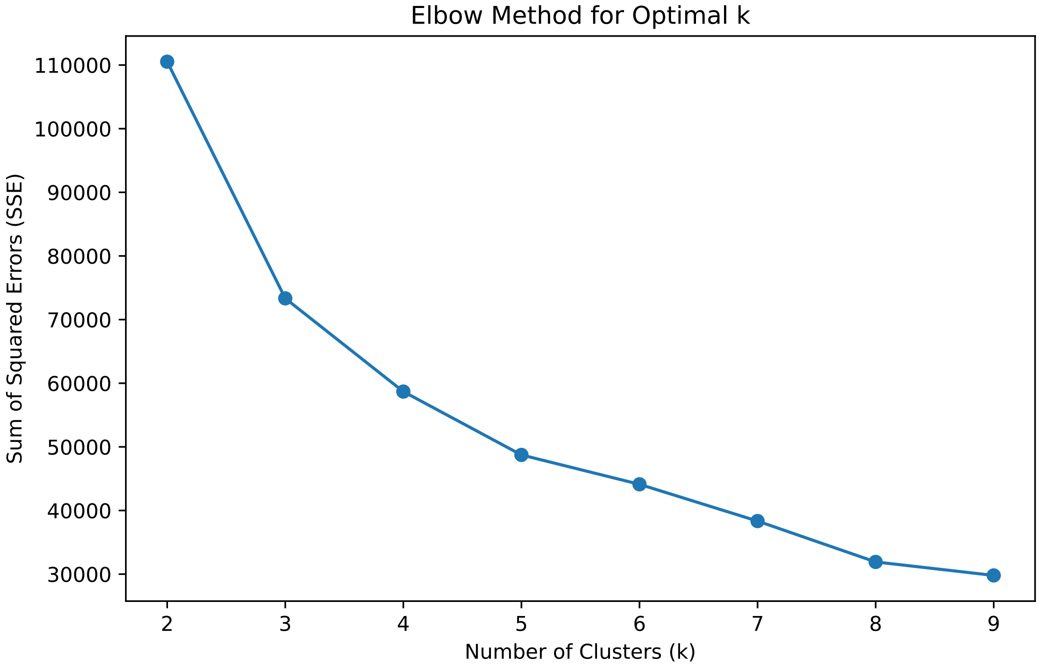
(45 points; Answer due 2024-12-17)

This section primarily explores the spatial differences of different property types and the social impacts brought by short-term rental density, using cluster analysis and geographically weighted regression. It also analyzes how such visualizations can inform the development of short-term rental (STL) regulation policies in London.

cluster analysis

This approach uses price, accommodates, and room type to classify the property situation across different boroughs.

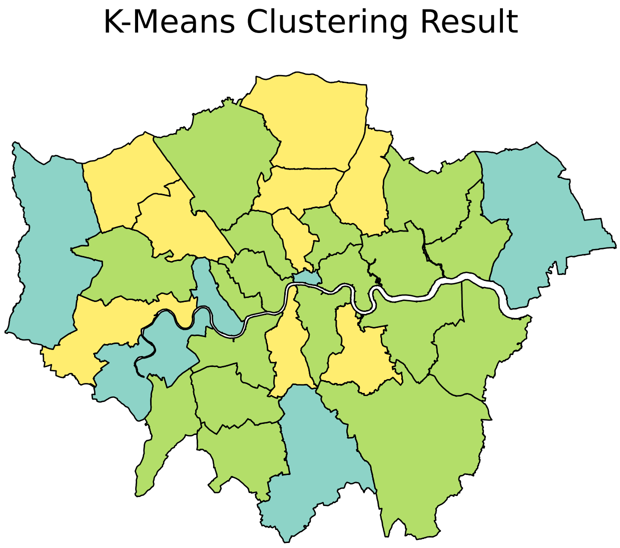
First, the elbow curve is plotted to determine the number of clusters for classification, with the final choice being 3 clusters.



The blue area in the final result represents the high-priced property zone. The average price is 464.46, which is much higher than the other two clusters. The standard deviation is large (1001.6), indicating significant price variation in this region, with a maximum price of 80,100, suggesting the presence of extreme high-priced properties. The average accommodation capacity is 5.93, indicating that these properties are often larger in size. Among them, 94.3% are Entire home/apt (entire properties), with very few Private room or Shared room listings. These areas are mainly concentrated in the wealthier, upscale regions on the map, likely in the city center or high-income residential areas. These characteristics align with the property market in central London, such as Westminster, Kensington and Chelsea, and the City of London. These economically developed areas are hotspots for high-end rentals. However, this category is also distributed in peripheral areas such as Harrow, Havering, and Croydon, likely due to the presence of large, low-density residential areas in London's outskirts, where entire homes are often rented out at high prices. Particularly in Havering, although the area is generally low-density, a small number of newly developed or high-spec entire homes (e.g., villas or resort-style properties) drive up prices, causing these areas to be included in this cluster.

The green area represents the low-priced property zone. The average price is 59.70, much lower than the other two clusters. The standard deviation is small (30.61), indicating a stable price distribution. The average accommodation capacity is 1.78, showing that these properties are small, typically catering to individuals or couples. Among them, 97.7% are Private room listings. The proportion of Shared room listings is slightly higher than in the other clusters. These properties are generally located in Outer London areas, such as Barking and Dagenham and Newham, where transportation is relatively convenient. However, the demand for these rentals primarily comes from budget-conscious tenants, such as students or solo travelers.

The yellow area represents the medium-priced property zone. The average price is 143.69, which falls between the other two clusters. The standard deviation is 74.67, indicating a relatively balanced price distribution. The average accommodation capacity is 2.98, suggesting that these properties are medium-sized and suitable for small families or groups of three. Among them, 97.0% are Entire home/apt listings, with very few Private room or Shared room listings. These properties are typically located in Inner London areas, such as Hackney, Islington, and Southwark, forming a transitional zone.



In summary, The blue area is primarily composed of high-end entire home rentals, characterized by extremely high prices and great diversity. Spatially, it may overlap with central London or upscale residential areas. The green area mainly consists of low-priced private rooms, catering to budget-conscious renters such as solo travelers or students. It is primarily located in outer or secondary market areas. The yellow area represents medium-priced rentals, mostly offering entire homes, suitable for small families or groups. Geographically, it has a broader distribution, with some overlap with tourist hotspot areas.

regression analysis

The distribution of short-term rental properties may lead to a series of social effects. This section takes social inequality as an example to explore the impact of short-term rental distribution on social outcomes. By visualizing the spatial differences in these effects, the analysis reveals the varying urgency for policy intervention across different areas.

Since factors such as economic conditions, population characteristics, and transportation accessibility may influence the distribution of short-term rentals—and these factors are not the focus of this study—their effects were removed from the independent variables.

The factors selected to describe economic conditions, population characteristics, and transportation accessibility include poverty rate, population density, density of subway stations, and density of rail lines. Principal Component Analysis (PCA) was applied to extract the most representative components from these variables.

The dependent variable chosen to represent social inequality is the 80:20 ratio of earnings, which indicates the level of income inequality within a given area. Specifically, this metric reflects the ratio of the total income of the wealthiest 20% of the population to the total income of the poorest 20%.

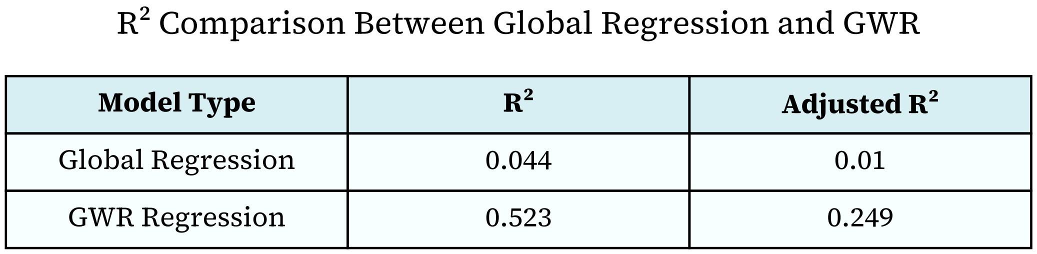
Administrative boroughs were used as the spatial units of analysis. The variables derived from PCA were first used to model the density of short-term rentals. The residuals—representing the portion of variation potentially linked to other effects associated with short-term rental distribution—were then used as predictors for the social inequality variable. The analysis proceeded as follows:

1.A global regression was performed using Ordinary Least Squares (OLS).

2.Geographically Weighted Regression (GWR) was applied in the second step.

```
/opt/conda/lib/python3.11/site-packages/geopandas/geodataframe.py:1819: SettingWithCopy


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/opt/conda/lib/python3.11/site-packages/geopandas/geodataframe.py:1819: SettingWithCopy


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/opt/conda/lib/python3.11/site-packages/geopandas/geodataframe.py:1819: SettingWithCopy


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/opt/conda/lib/python3.11/site-packages/geopandas/geodataframe.py:1819: SettingWithCopy


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

The following table describes the results of the second step, Geographically Weighted Regression (GWR). The results show that, after removing the aforementioned control factors, GWR performs better than the global regression, with the independent variables explaining more of the variation in the dependent variable. This also indicates, to some extent, that the impact of rental density on social inequality exhibits spatial variation.



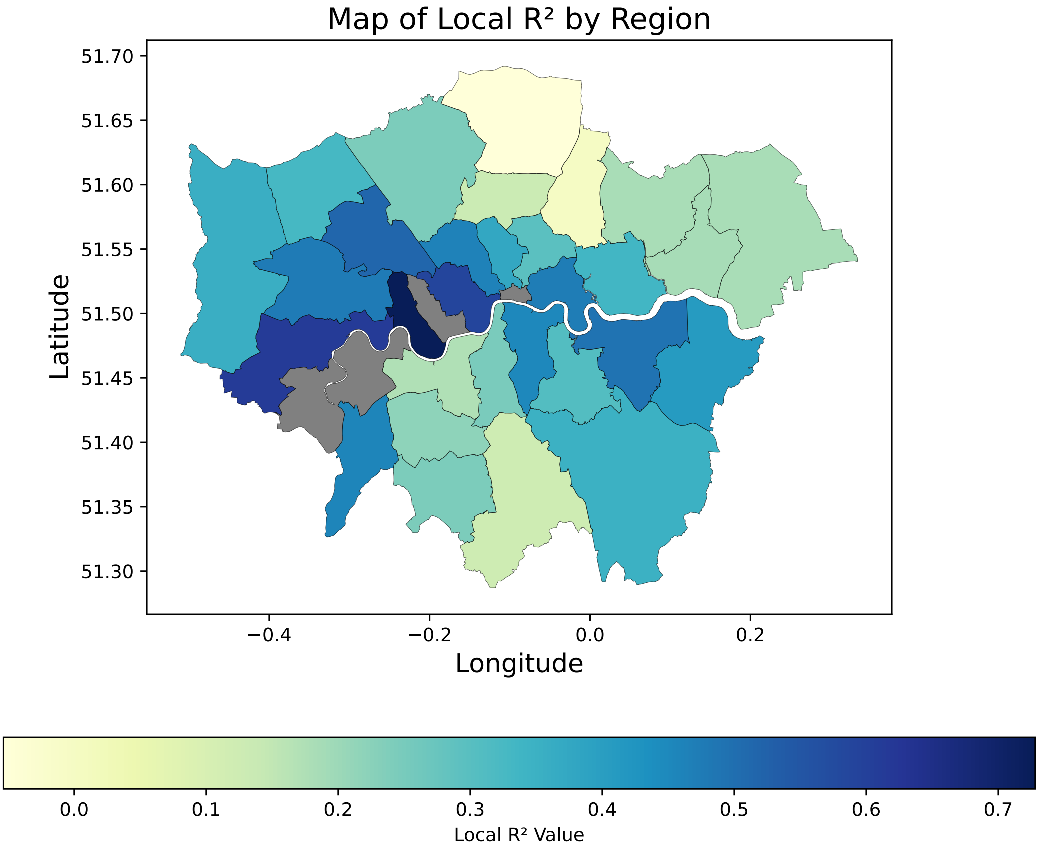
R^2 Comparison Between Global Regression and GWR

Model Type	R^2	Adjusted R^2
Global Regression	0.044	0.01
GWR Regression	0.523	0.249

By comparing the R^2 values of the GWR results across different boroughs, the geographical variation in the impact of short-term rental distribution on social inequality was assessed.

The results reveal that certain boroughs, particularly in the central and western areas, such as Hammersmith and Fulham, display high R^2 values, reaching up to 0.7. This indicates that the independent variables explain 70% of the variation in the dependent variable. In contrast, boroughs such as Enfield and Waltham Forest exhibit R^2 values close to 0, suggesting that the effects of short-term rental distribution on social inequality vary significantly across regions.

The results reveal the varying degrees to which social inequality in different regions is influenced by the distribution of short-term rentals, highlighting differences in the urgency of policy intervention across areas. For example, in Hammersmith and Fulham, the R-squared value is notably high, and the coefficient is positive, indicating a significant positive correlation between the increase in short-term rental density and the rise in social inequality. This suggests that the impact of short-term rental distribution on income inequality is particularly pronounced in this area, underscoring the greater necessity for policy intervention. Measures such as setting limits on short-term rentals, increasing taxes and fees, and encouraging properties to return to the long-term rental market can be implemented. However, the specific measures to be taken require further analysis of the underlying causes of this high correlation.



Nevertheless, there are several issues and limitations in this analysis:

1.The large spatial unit of the data may result in insufficient sample sizes, which can affect the stability and interpretability of the model's results.

2.Missing data in some regions led to the absence of R^2 values in certain areas.

3.Controlling for the influence of economic conditions, population, and transportation accessibility assumed linear effects, potentially overlooking nonlinear influences.

4.The residuals, after accounting for economic, population, and transportation factors, may still include unaccounted external effects, meaning they do not solely represent the impact of short-term rental distribution.

References

Alsudais, A. (2021) 'Incorrect data in the widely used inside airbnb dataset', *Decision Support Systems*, 141, p. 113453. Available at: https://www.sciencedirect.com/science/article/pii/S0167923620302086.

'Inside airbnb' (n.d.). Available at: http://insideairbnb.com.

Prentice, C. and Pawlicz, A. (2024) 'Addressing data quality in airbnb research', *International Journal of Contemporary Hospitality Management*, 36(3), pp. 812–832.