Spatial Determinants of Airbnb Prices in Milan

Sanzhar Sailaubek

GitHub Repository: https://github.com/sanzhkzz/Geospatial_analytics_DS

Abstract

This study investigates how location and neighbourhood characteristics influence Airbnb prices in Milan. Using georeferenced listings, we computed distance to Duomo, aggregated prices at the neighbourhood level, and assessed spatial patterns through regression analysis and Moran's I. Results reveal a clear center–periphery gradient: prices decline with distance from the city center and increase with accommodation capacity. Significant positive spatial autocorrelation confirms that similar price levels cluster geographically across neighbourhoods.

1. Introduction

The rapid growth of short-term rental platforms such as Airbnb has significantly reshaped urban housing markets across Europe. In major cities, short-term rentals influence tourism dynamics, neighborhood composition, and local housing affordability. As a result, understanding the determinants of Airbnb pricing has become an important topic in urban economics and spatial analysis.

Location is widely recognized as a key determinant of housing prices. In urban contexts, proximity to the city center, accessibility to services, and neighborhood characteristics often generate strong spatial price gradients. However, the extent to which Airbnb prices exhibit spatial structure within a city remains an empirical question that requires formal geospatial analysis.

1.2 Research question

This study investigates the following research question:

> **How do location and neighbourhood characteristics influence Airbnb prices in Milan?**

The further analysis examines how prices vary across neighbourhoods to determine whether certain areas consistently command higher listing prices and to identify the mechanisms driving these differences. The study evaluates whether central locations are more expensive primarily due to proximity to the city centre, or whether local amenities such as the Cattedrale Duomo di Milano contribute additional market value to short-term rental properties.

2. Description of the Data

Airbnb listing data were obtained from the website InsideAirbnb, an independent, open-data initiative that collects and publishes publicly accessible snapshots of Airbnb listings for major cities worldwide. The platform scrapes listing information directly from Airbnb's website and provides structured datasets for academic research and policy analysis. These data have been widely used in urban studies, housing market research, and spatial econometrics to examine the impact of short-term rentals on housing affordability, tourism dynamics, and neighborhood change.

The dataset used in this study includes georeferenced Airbnb listings in Milan at the time of data collection. Each observation represents an individual listing and contains both spatial and non-spatial attributes describing property characteristics, host activity, and review performance.

Key variables used in this analysis include:

- Price (in euros)

- Latitude and longitude

- Neighborhood identifier

- Room type (entire home/apartment, private room, shared room)

- Accommodation capacity and bedroom count

- Number of reviews and rating scores

The original coordinate reference system (CRS) of the dataset was EPSG:4326 (WGS84), which represents geographic coordinates in degrees of latitude and longitude. Since distance calculations in degrees are not meaningful for metric analysis, the dataset was transformed to EPSG:32632 (UTM Zone 32N), a projected coordinate system appropriate for Northern Italy. This transformation enables accurate computation of distances in meters, which is essential for spatial regression models and distance-based variables such as proximity to central landmarks.

Neighborhood boundary data were obtained in GeoJSON format and used to spatially aggregate listing-level prices. A spatial join was performed to assign each listing to its corresponding neighborhood polygon.

To ensure statistical robustness, neighborhoods with fewer than 20 listings were excluded from aggregated analysis. Median prices were used as the primary measure of central tendency to reduce sensitivity to extreme values.

3. Data Analysis Oriented by the Research Question

This section outlines the empirical framework adopted to evaluate spatial dependence and determinant factors of Airbnb prices in Milan. The analysis integrates exploratory statistical analysis, neighbourhood-level spatial aggregation, spatial autocorrelation assessment using Moran’s I, and log-linear regression modeling. Together, these methods allow for a comprehensive investigation of whether Airbnb prices are spatially clustered and how geographic centrality and property characteristics contribute to price formation.

3.1 Exploratory Analysis of Price Distribution

The exploratory data analysis (EDA) serves to examine the structural properties and underlying patterns of the Airbnb dataset prior to formal statistical modeling. This stage of analysis is essential for understanding the distributional characteristics of key variables, particularly price, and for identifying the presence of skewness, dispersion, and extreme values that may influence model specification.

Through descriptive statistics and visualizations, the exploratory phase provides empirical grounding for subsequent methodological decisions, including the selection of explanatory variables and the adoption of a logarithmic transformation. By systematically examining both non-spatial and spatial dimensions of the data, the EDA ensures that the regression and spatial statistical analyses are informed, justified, and analytically coherent.

The initial exploratory analysis began with boxplots of Airbnb prices across room types (Figure 1). These visualizations revealed substantial dispersion within each category and, more importantly, the presence of extreme price values far exceeding the general distribution range. In particular, several listings exhibited prices that were orders of magnitude higher than the typical market levels, suggesting potential data entry errors, atypical luxury listings, or irregular pricing structures.

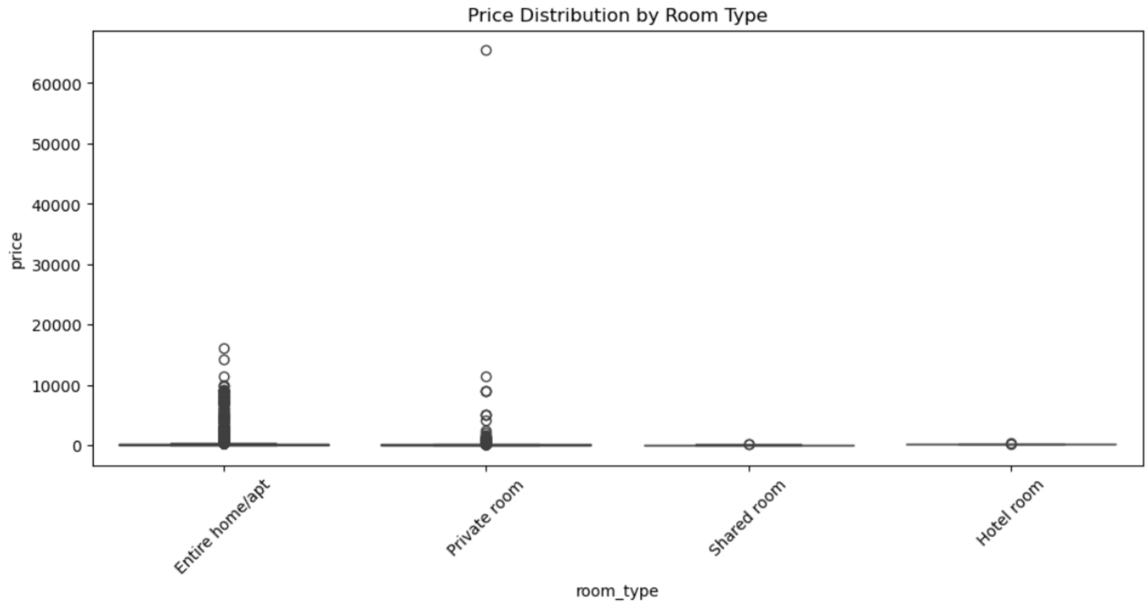


Figure 1. Boxplot of Prices

Such extreme observations exert disproportionate influence on summary statistics and regression estimates, potentially distorting the underlying structural relationships. To mitigate this effect while preserving the overall distributional structure, prices were trimmed at the 99th percentile. This approach removes only the most extreme 1% of observations, thereby reducing leverage from outliers without imposing arbitrary thresholds.

Following percentile trimming, the price distribution remained positively skewed but became substantially more stable and representative of the broader market. This adjustment provided a more robust foundation for subsequent modelling.

Figure 2 presents boxplots of Airbnb prices across room types after percentile trimming. As can be seen, entire home/apartment listings exhibit the highest median prices and the greatest variability. Private rooms show lower median prices but substantial dispersion, while shared rooms are consistently the least expensive category. Numerous high-price outliers are observed across all categories, particularly among entire home listings, indicating strong right-skewness in the overall price distribution.

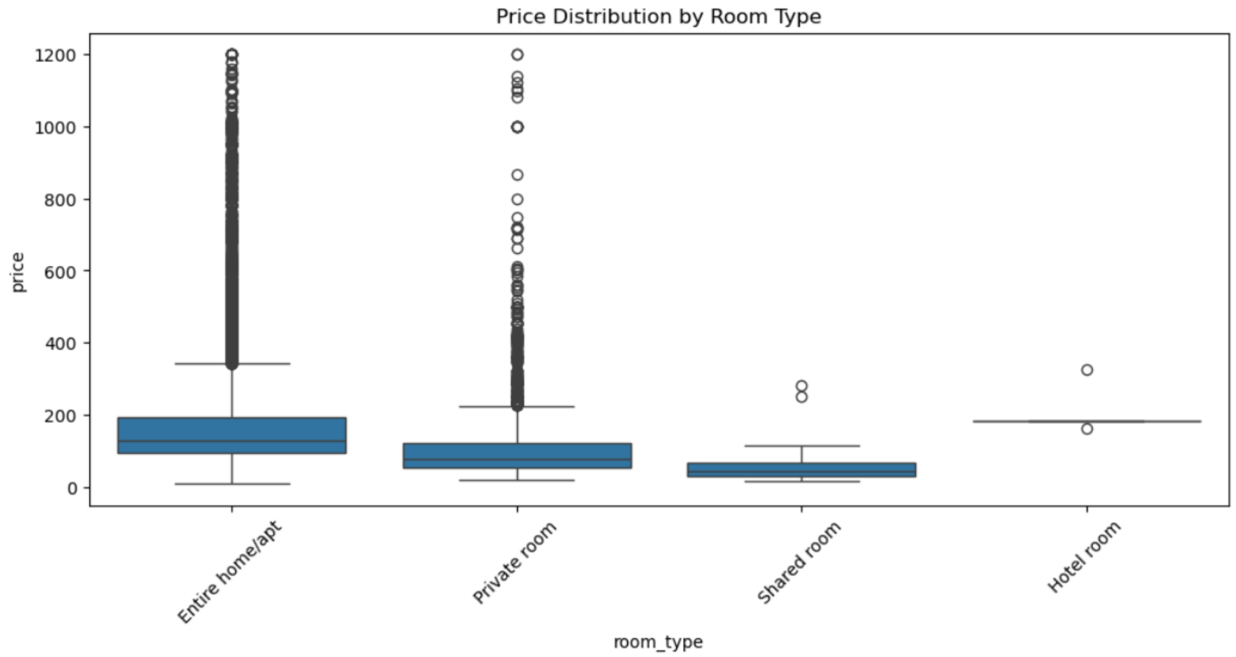


Figure 2. Price Distribution by Room Type

Figure 3 presents the overall shape of the distribution remains positively skewed, but extreme distortions are substantially reduced. This cleaned distribution better reflects the typical pricing structure of Airbnb listings .

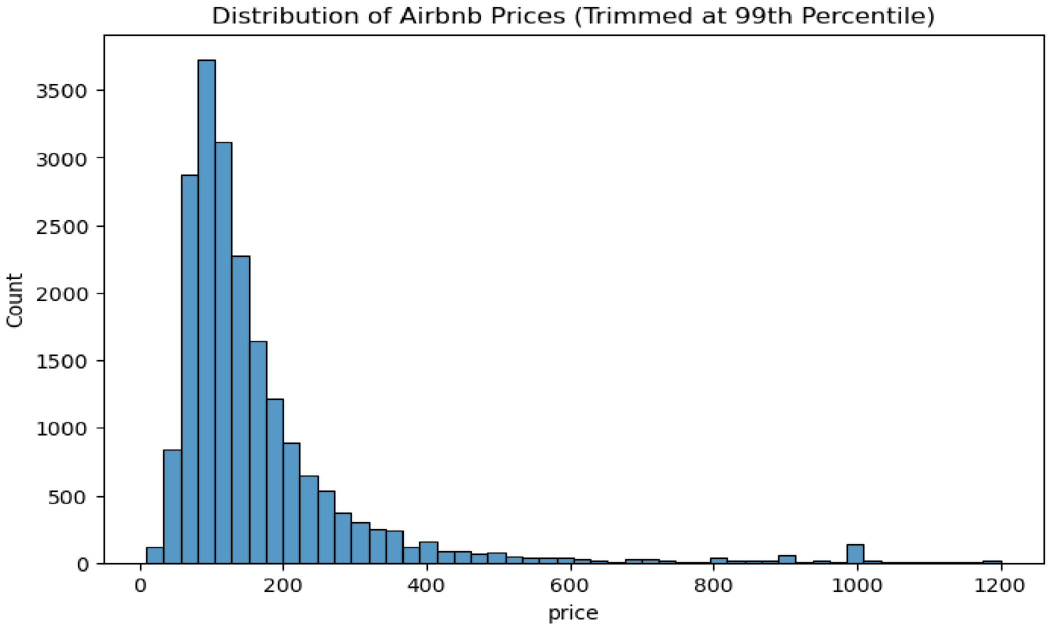


Figure 3. Price Distribution Trimmed at the 99th Percentile

Figure 4 shows the distribution of Airbnb prices truncated at €500 to improve visibility of the main mass of observations. The distribution is strongly right-skewed, with the majority of listings concentrated between €80 and €200 per night which is more representative population.

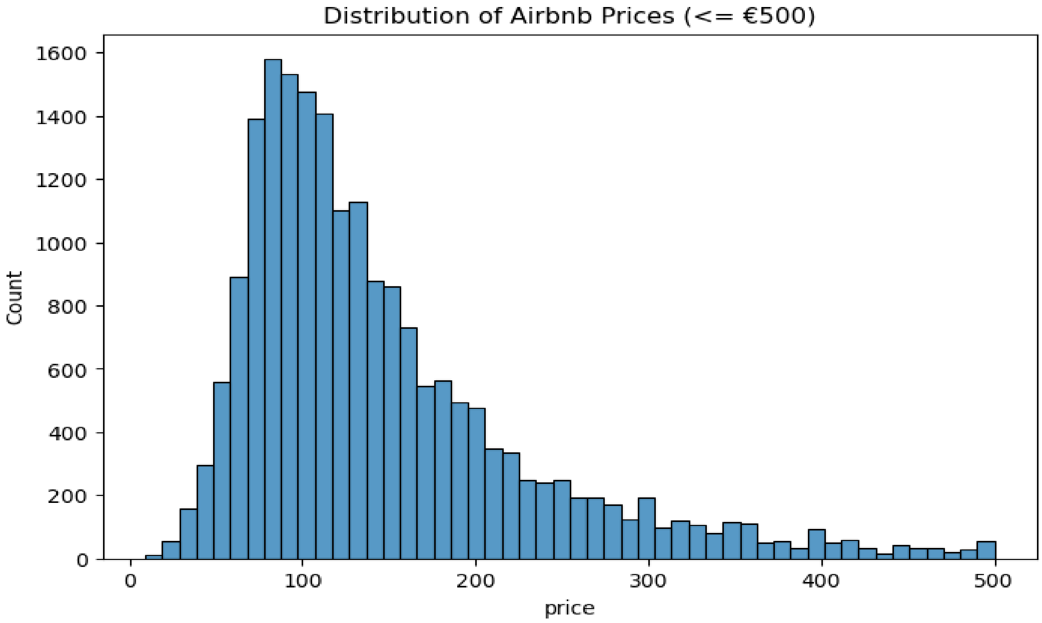


Figure 4. Histogram of Prices (≤ €500)

Figure 5 displays price distributions across distance bands from Duomo (0–1 km, 1–3 km, 3–5 km, and 5+ km). Central listings (0–1 km) exhibit both higher median values and greater dispersion compared to peripheral areas.

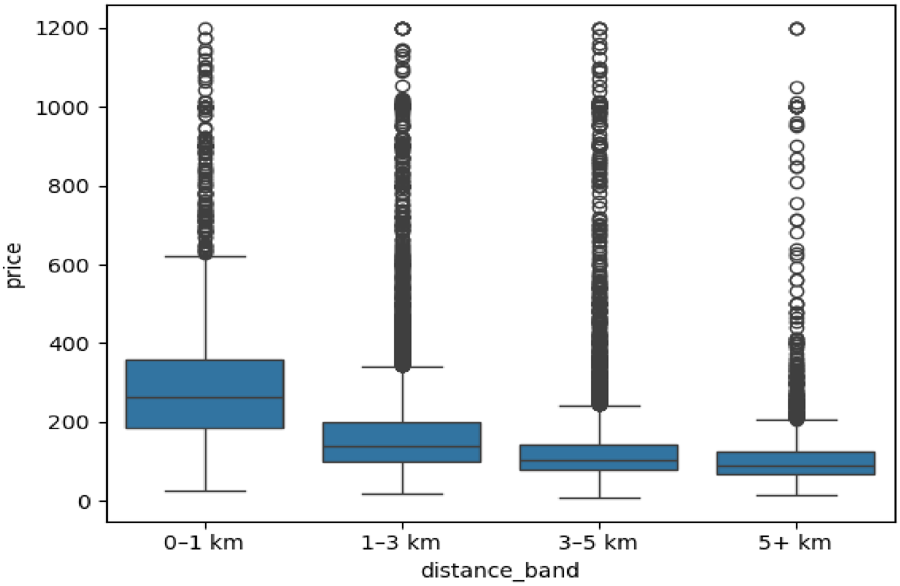


Figure 5. Price Distribution by Distance Band

Figure 6 illustrates the relationship between price and accommodation capacity. A positive association is visible, with larger properties generally commanding higher prices. However, substantial vertical dispersion suggests that capacity alone does not fully explain price variation, reinforcing the need for multivariate analysis.

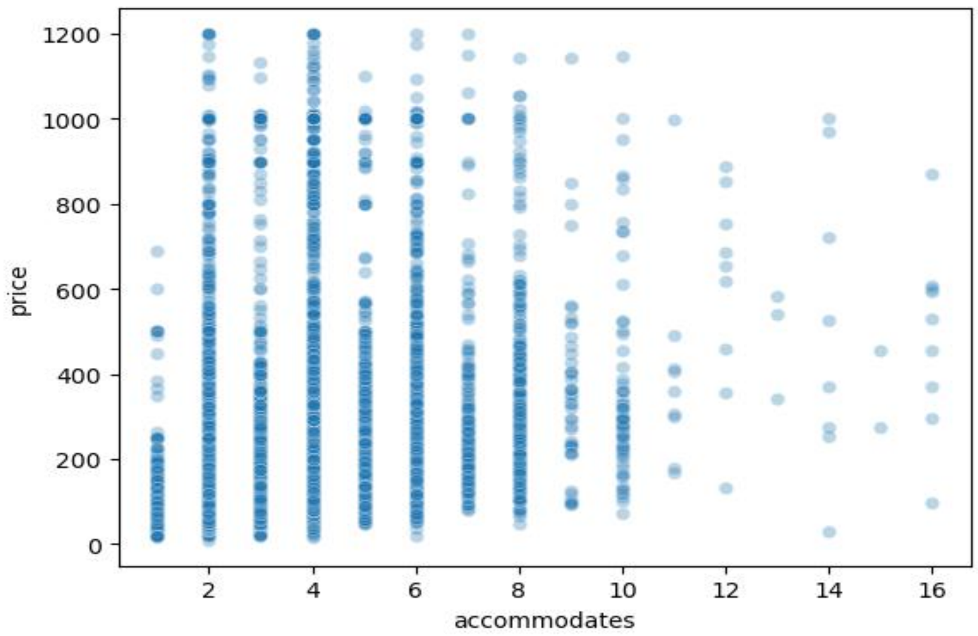


Figure 6. Price vs Accommodation Capacity

Figure 7 displays the relationship between listing price and distance to Duomo. A negative gradient is visible, with prices generally declining as distance increases. The concentration of higher-priced listings in proximity to the city center visually confirms the center-periphery pattern later quantified through regression analysis.

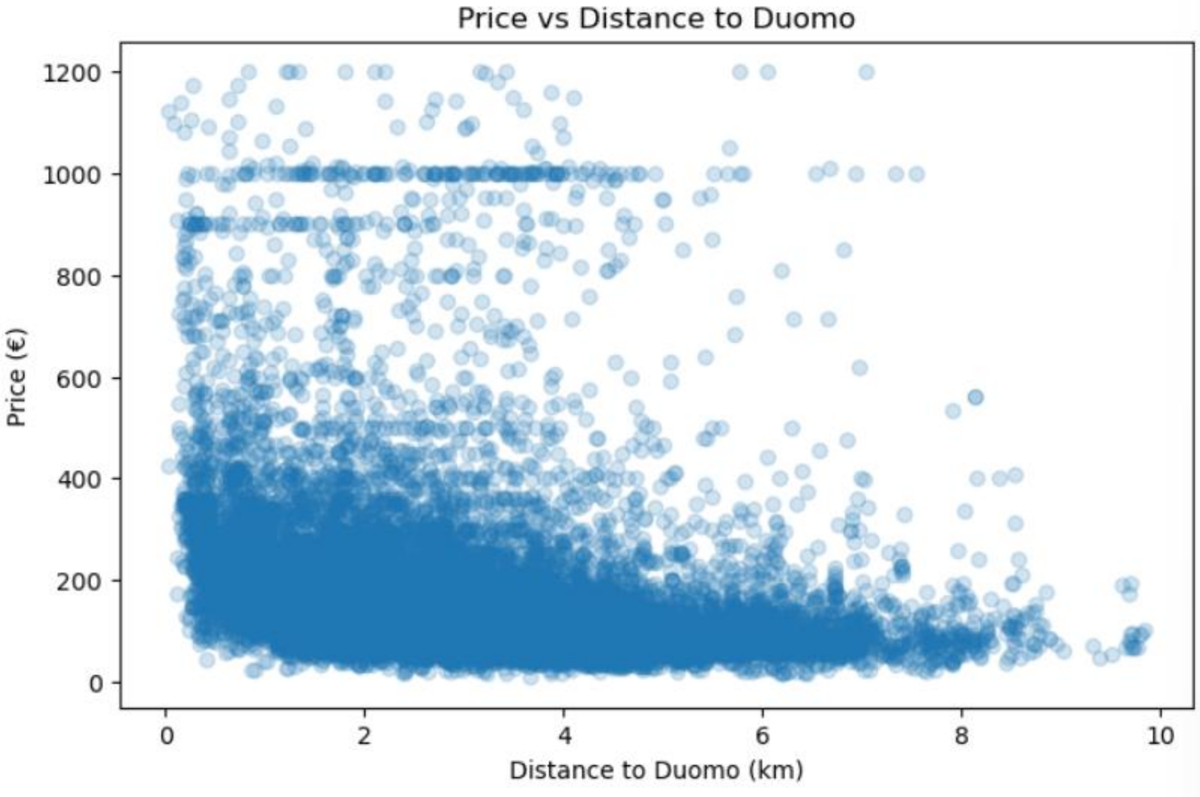


Figure 7. Price vs Distance to Duomo

To formally investigate spatial structure and assess clustering at a broader geographic scale, the analysis now shifts from individual listings to neighbourhood-level aggregation.

3.2 Spatial Aggregation and Neighbourhood-Level Analysis

Each Airbnb listing was spatially joined to Milan's neighbourhood polygons using a containment-based spatial join (within predicate). This operation assigns each point observation (listing) to a corresponding administrative unit (neighbourhood polygon), thereby enabling aggregation from micro-level observations to areal units. Such spatial aggregation is necessary because many spatial statistical techniques, including measures of spatial autocorrelation, are defined for polygon-level (areal) data rather than individual points.

To transition from point-level observations to an areal framework suitable for spatial analysis, the median price was computed for each neighbourhood. The median provides a more robust measure of central tendency. The resulting neighbourhood-level dataset was visualized through a choropleth map of median prices.

Choropleth mapping is a standard technique in spatial analysis for representing aggregated values across administrative units. By encoding price levels using a continuous color gradient, spatial variation in median prices becomes visually interpretable.

Figure 8 presents a choropleth map of median Airbnb prices aggregated at the neighbourhood level. Darker shades indicate higher median prices, while lighter tones represent lower-priced areas. A clear spatial gradient is visible, with central neighbourhoods surrounding Duomo exhibiting the highest prices, while peripheral ares display systematically lower values, reinforcing the presence of a geographical factor in pricing.

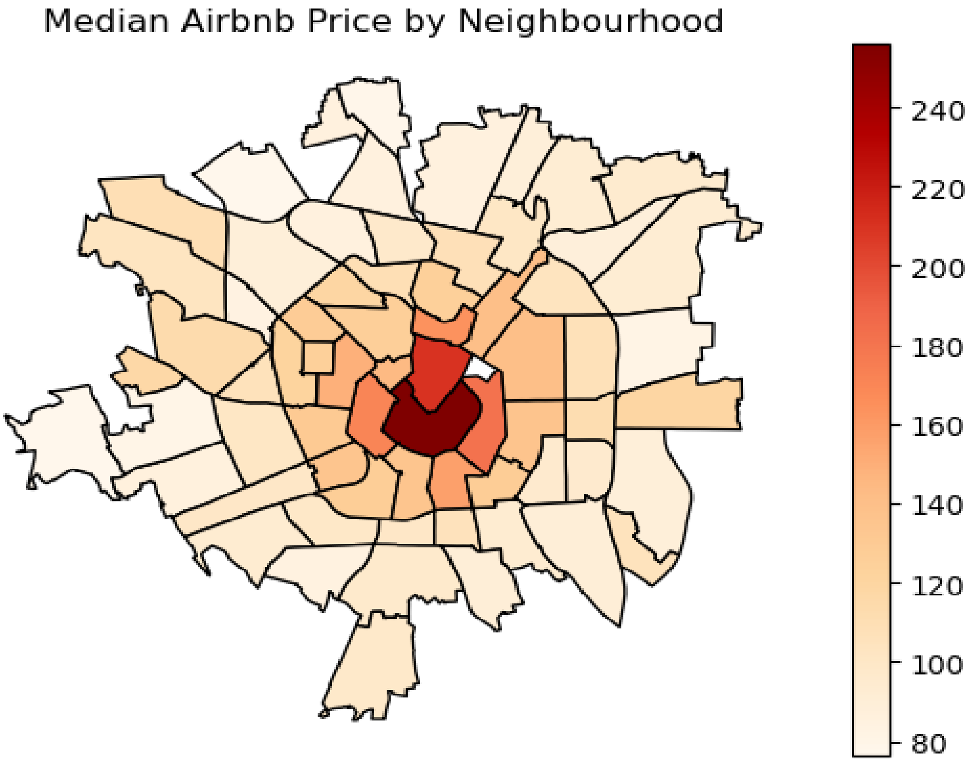


Figure 8. Median Airbnb Price by Neighbourhood

These patterns are consistent with classical urban economic theory, which predicts higher land and rental values in central areas due to accessibility, amenities, and concentration of economic activity.

However, visual inspection alone does not establish whether the observed spatial pattern is statistically significant or merely coincidental. Descriptive mapping can reveal apparent clusters, but it does not quantify spatial dependence. To formally assess whether similar price levels are spatially clustered, spatial autocorrelation was evaluated using Moran's I.

Moran's I provides a global measure of spatial dependence by comparing the similarity of values in neighbouring areas relative to overall variance. A statistically significant positive value indicates that neighbourhoods with high (or low) median prices tend to be located near areas with similar values, confirming the presence of structured spatial clustering rather than random dispersion.

3.3 Spatial Autocorrelation: Moran's I

In spatial data analysis, observations located close to each other are often not independent. This phenomenon is known as spatial autocorrelation. Traditional statistical methods assume independence between observations. However, when spatial dependence exists, this assumption may be violated. Therefore, it is necessary to formally test whether spatial clustering is present before interpreting spatial patterns.

Moran's I is one of the most widely used global measures of spatial autocorrelation. It evaluates whether similar values of a variable tend to occur near one another more frequently than would be expected under spatial randomness.

The statistic compares:

- The deviation of each spatial unit's value from the global mean
- The values observed in neighboring units
- The overall variance of the dataset

A spatial weights matrix defines which units are considered neighbors. In this study, a Queen contiguity matrix is used, meaning two neighbourhoods are considered neighbors if they share a boundary.

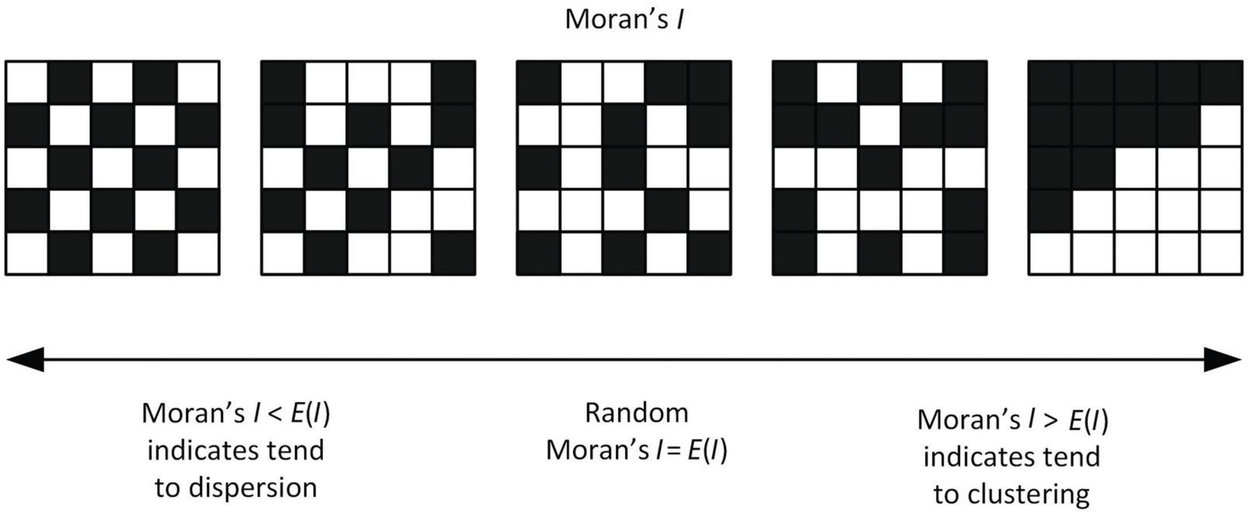

Moran's *I*
Moran's *I* < *E*(*I*) indicates tend to dispersion
Random Moran's *I* = *E*(*I*)
Moran's *I* > *E*(*I*) indicates tend to clustering

Figure 9. Moran's I example visualization

The value of Moran's I typically ranges between −1 and +1:

- $I > 0$ indicates positive spatial autocorrelation (similar values cluster together)
- $I \approx 0$ suggests spatial randomness
- $I < 0$ indicates negative spatial autocorrelation (high values are surrounded by low values)

Testing spatial autocorrelation is essential because it validates whether observed spatial patterns in the choropleth map reflect genuine geographic structure rather than visual coincidence.

Spatial dependence was evaluated using Moran's I with a Queen contiguity spatial weights matrix.

The Moran's I statistic is defined as:

$$I = \frac{n}{W} \cdot \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

where:

- $x_i$is the median price in neighbourhood i
- w_{ij}indicates whether two neighbourhoods share a boundary

The estimated Moran's I in our case is approximately **0.60**, with a permutation p-value of **0.001.**

This indicates:

- Strong positive spatial autocorrelation
- Statistically significant clustering of similar price levels

Therefore, Airbnb prices in Milan are not randomly distributed in space. High-price neighbourhoods tend to be located near other high-price neighbourhoods, and low-price areas cluster together which proves stated point before.

3.4 Centrality Analysis: Distance from Duomo

To evaluate whether centrality influences Airbnb pricing in Milan, the Euclidean distance from each listing to Duomo used as a proxy for the city center was calculated. Distances were computed using projected coordinates to ensure accurate measurement in kilometers.

The relationship between price and distance was initially assessed using Pearson correlation. The estimated correlation coefficient is approximately 0.27, indicating a moderate negative association. This suggests that listings located farther from Duomo tend to have lower prices on average.

Analyzing distance serves two purposes:

1. It provides an intuitive measure of spatial centrality.
2. It allows us to quantify how geographic accessibility contributes to price variation before formal regression modeling.

These findings suggest that proximity to Duomo plays a structural role in Airbnb pricing, motivating its inclusion as a key explanatory variable in the regression analysis.

3.5 Determinant Analysis: Log-Linear Regression Model

Background on the Regression Modeling Strategy

To investigate the determinants of Airbnb prices, a sequence of Ordinary Least Squares (OLS) regression models was estimated. OLS is a statistical method used to estimate the relationship between a dependent variable and one or more explanatory variables by minimizing the sum of squared residuals - the squared differences between observed and predicted values.

The modeling strategy followed a stepwise approach in order to progressively refine the specification and improve interpretability.

Model 1: Distance-Only Specification

The first model included only the distance from Duomo as an explanatory variable:

$$\text{price} = \beta_0 + \beta_1 \cdot \text{distance}_{km} + \varepsilon$$

This baseline model was designed to isolate the direct relationship between centrality and price. It provides an initial estimate of how price changes as listings move farther from the city center. However, this specification does not account for structural differences in listing characteristics.

Model 2: Distance and Accommodation Capacity

The second model extended the specification by including accommodation capacity:

$$\text{price} = \beta_0 + \beta_1 \cdot \text{distance}_{km} + \beta_2 \cdot \text{accommodates} + \varepsilon$$

This model allows us to evaluate the partial effect of distance while controlling for property size. Including capacity improves explanatory power and reduces omitted-variable bias, as larger properties naturally command higher prices.

Model 3: Log-Linear Specification

Because the price distribution exhibited strong right skewness and heteroskedasticity, a log transformation of the dependent variable was applied:

$$\log(\text{price}) = \beta_0 + \beta_1 \cdot \text{distance}_{km} + \beta_2 \cdot \text{accommodates} + \varepsilon$$

Log-linear models are commonly used in real estate and urban economics because housing prices often follow a log-normal distribution. The transformation stabilizes variance, reduces skewness, and allows coefficients to be interpreted as approximate percentage changes.

This progression from a simple linear model to a multivariate specification and finally to a log-transformed model ensures that the final results are both statistically robust and economically interpretable.

Table 1 presents three model specifications. Model 1 estimates the isolated effect of distance from Duomo on price, showing a significant negative relationship. Model 2 introduces accommodation capacity, which substantially improves explanatory power (R^2 increases from 0.072 to 0.159). Model 3 applies a log transformation to address skewness and allows percentage interpretation of coefficients. In the log-linear specification, each additional kilometer from the city center reduces price by approximately 14.6%, while each additional guest capacity increases price by approximately 15.2%. The log model explains nearly 30% of the variation in Airbnb prices.

Table 1: Model Performances

Variable	**Model 1: Price**	**Model 2: Price**	**Model 3: Log(Price)**
Distance (km)	−25.72***	−24.45***	−0.146***
	(0.65)	(0.62)	(0.002)
Accommodates	—	28.53***	0.152***
		(0.62)	(0.002)
Constant	244.41***	144.10***	4.818***
	(2.17)	(3.01)	(0.011)
R^2	0.072	0.159	0.295
Observations	20,342	20,342	20,342

Standard errors in parentheses.

**** $p < 0.01$*

Stars	***Meaning***	***p-value threshold***
***	*Highly significant*	$p < 0.01$
**	*Significant*	$p < 0.05$
*	*Marginally significant*	$p < 0.10$

3.6 Interactive Spatial Exploration

While regression analysis and spatial autocorrelation provide formal statistical evidence of price determinants and clustering, they do not fully capture the fine-grained spatial heterogeneity visible at the listing level. Statistical models summarize relationships numerically, but they abstract away from the geographic context in which these relationships unfold.

To complement the quantitative results and allow a more intuitive exploration of spatial patterns, an interactive web map was developed. Unlike static maps or aggregated statistics, the interactive visualization enables dynamic inspection of individual listings, price gradients, and room-type distributions across the city.

This tool serves two purposes. First, it visually confirms the gradient identified in the regression analysis. Second, it allows users to explore micro-level spatial variation that may not be fully captured by neighbourhood-level aggregation or global statistics such as Moran’s I.

The interactive component therefore bridges formal spatial analysis and exploratory geographic interpretation, enhancing both transparency and interpretability of the findings.

An interactive web map was developed to allow visual exploration of:

- Price gradient (continuous color scale)
- Room type categories (toggle layers)

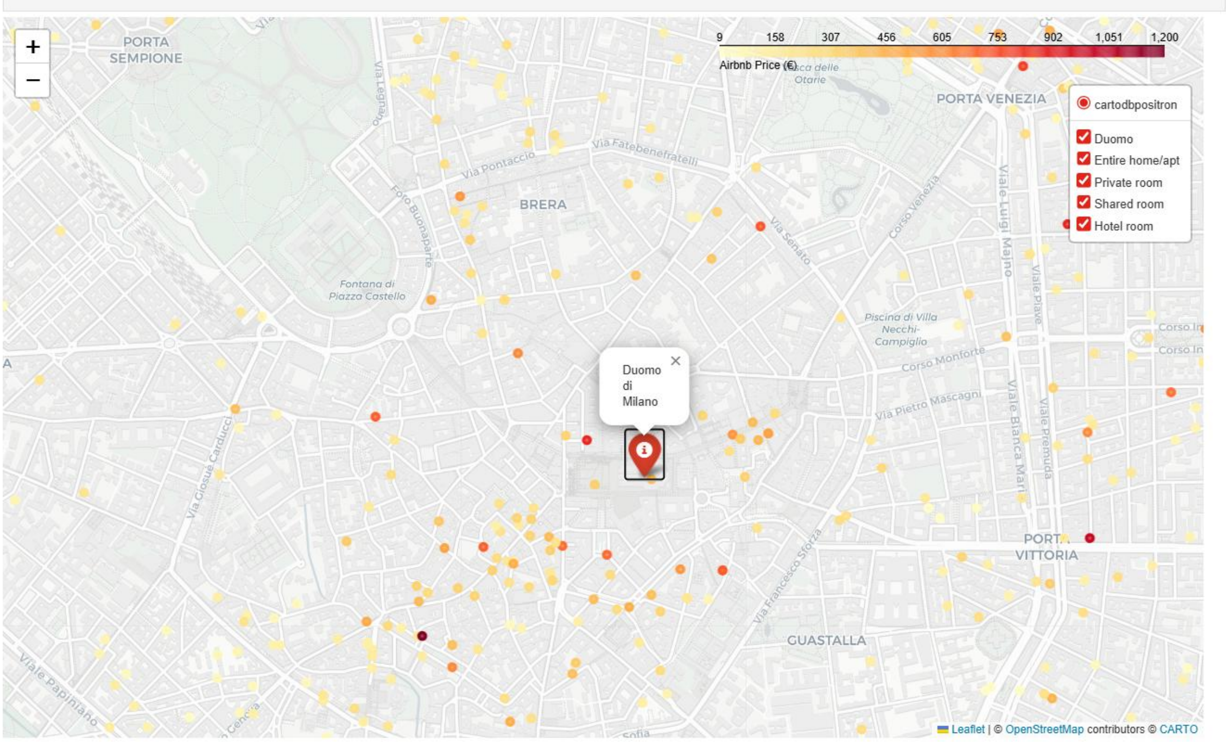


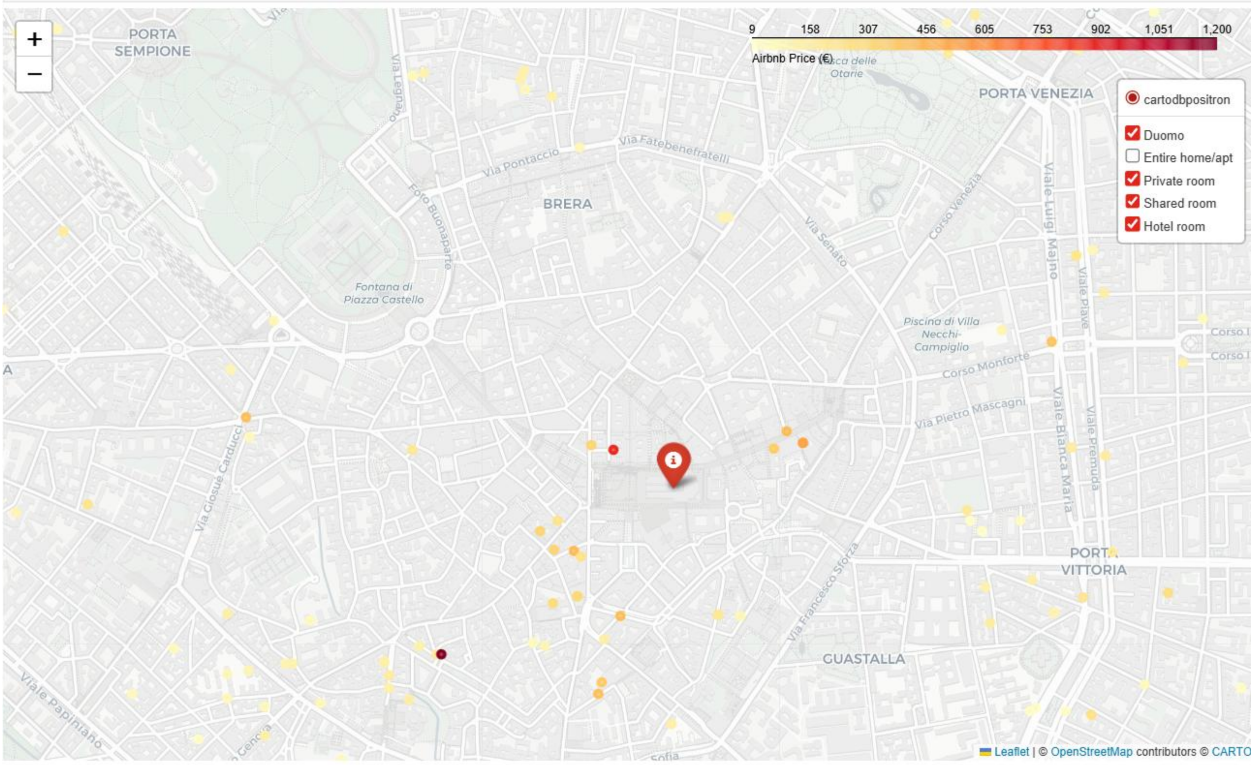
Figure 10. Overview on interactive map

At higher zoom levels, the interactive map enables inspection of individual Airbnb listings at street scale. Each marker reveals detailed property-level information through interactive tooltips, including:

- Price (€)
- Distance from Duomo (km)
- Accommodation capacity
- Room type

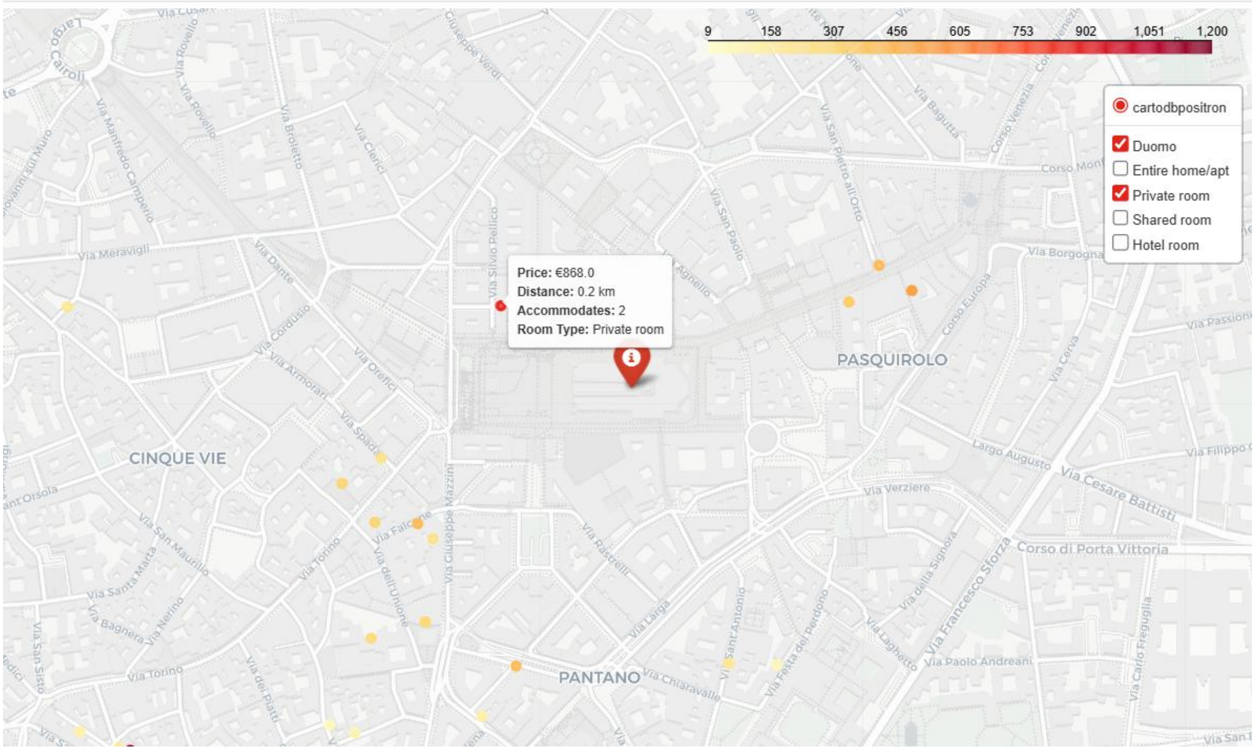
For example as shown on figure 11, a listing located approximately 0.2 km from Duomo is priced at €868 per night, accommodates two guests, and is categorized as a private room. This granular view demonstrates the premium pricing associated with central proximity, even for non-entire-unit listings.

The ability to zoom and inspect individual listings enhances exploratory analysis by linking macro-level spatial patterns to micro-level property characteristics. Users can directly observe how price levels vary within small urban areas and assess how centrality interacts with room type and capacity.



● cartodbpositron
☑ Duomo
□ Entire home/apt
☑ Private room
☑ Shared room
☑ Hotel room

Figure 11. Overview on interactive map with home/apt flag off



● cartodbpositron
☑ Duomo
□ Entire home/apt
☑ Private room
□ Shared room
□ Hotel room

Price: €868.0
Distance: 0.2 km
Accommodates: 2
Room Type: Private room

Figure 12. Zoom in to the closest private room to Duomo

4. Conclusions

This project examined the spatial structure and determinants of Airbnb prices in Milan by integrating geospatial analysis, regression modeling, and spatial autocorrelation techniques.

The combined evidence suggests that Airbnb pricing in Milan reflects both spatial structure and property-level characteristics. Centrality and accommodation capacity emerge as primary determinants, with prices decreasing systematically as distance from Duomo increases and rising with larger guest capacity. Furthermore, the significant Moran's I statistic confirms that prices are not randomly distributed but instead exhibit strong spatial clustering at the neighbourhood level.

Together, these findings indicate that Airbnb prices in Milan are shaped by both economic fundamentals such as location and size and broader neighbourhood dynamics. The results are consistent with urban economic theory that was stated earlier.

Overall, the project demonstrates how geospatial methods including spatial joins, choropleth mapping, regression modeling, and spatial autocorrelation analysis can be systematically integrated to uncover structured patterns in urban economic data and provide statistically grounded insights into spatial pricing mechanisms.

5. Limitations and Further Developments

Although the analysis provides strong evidence of spatial clustering and centrality effects, several limitations should be acknowledged.

First, the regression model includes a limited set of explanatory variables distance to Duomo and accommodation capacity. While these variables capture key structural determinants of price, other relevant factors such as amenities, review ratings, host characteristics, property quality, and neighbourhood socio-economic conditions were not incorporated. The omission of such variables may limit explanatory power and could introduce omitted variable bias.

Second, the analysis is cross-sectional and does not account for temporal dynamics. Airbnb prices may vary seasonally, during special events, or across years. A longitudinal dataset would allow examination of price dynamics over time.

Third, although significant spatial autocorrelation was detected at the neighbourhood level using Moran's I, the regression model does not explicitly incorporate spatial lag or spatial error components. Spatial econometric models (e.g., spatial lag or spatial error models) could more formally account for spatial dependence and potentially improve model performance.

Future research could integrate richer property attributes, socio-economic indicators, temporal variation, and spatial econometric techniques to provide a more comprehensive understanding of short-term rental pricing.

6. Reproducibility and Tools

This project was implemented using open-source geospatial and statistical tools in Python.

Key frameworks are described with their version and purpose in the Table 2.

Table 2. Version control and libraries

Library	**Version**	**Purpose**
pandas	2.2.3	Data manipulation and preprocessing
geopandas	1.1.2	Geospatial data handling and spatial joins
folium	0.20.0	Interactive web-based mapping
libpysal	4.14.1	Spatial weights construction
esda	2.8.1	Spatial autocorrelation (Moran’s I)
statsmodels	0.14.4	Regression modeling (OLS)

The full code, including data cleaning, spatial join, statistical modeling, and interactive map generation, is provided within the accompanying Jupyter Notebook to ensure reproducibility.