Urban Dynamics: Data Analysis and Machine Learning for Understanding Contemporary Cities. Learning from Airbnb data

Carmen Rubio Garcia
Architecture Bachelor Thesis
ETSAM Universidad Politécnica de Madrid
Tutor: Jose Ballesteros
carmenrubio1@hotmail.es

Abstract
This research proposes to explore the tools of artificial intelligence for urban analysis through the study of Airbnb data from three case studies: Madrid, Berlin, and Chicago.

Keywords: Urban Data, Machine Learning, Airbnb, Urbanism, Pricing prediction, Data analysis, K-nearest

1 Introduction

This thesis explores urban dynamics through the lens of data analysis and machine learning. The study investigates extensive datasets sourced from Airbnb to understand the contemporary urban environments. Subsequently, a machine learning model is developed not only to predict Airbnb pricing but also to construct a classification model, enhancing our ability to comprehend urban dynamics. Additionally, a comprehensive examination the relationships between the different variables of the Airbnb Data is undertaken to discern underlying patterns and correlations indicative of urban phenomena.

2 Methodology

The dataset is compiled by the group Inside Airbnb, which publishes information about Airbnb listings sourced from the Airbnb website. Specifically, data published on November 14, 2019, was utilized for this study. The dataset includes information such as listing prices, property types, neighbourhood/district, availability, minimum nights, and room types.

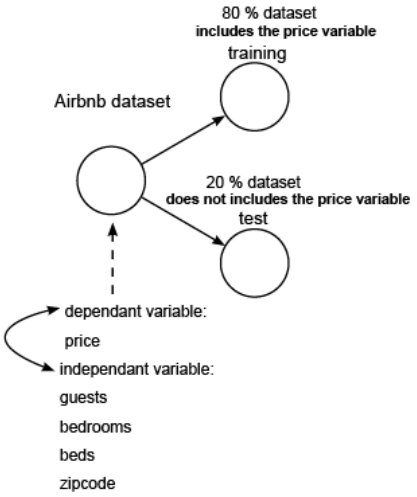


Figure 1: Data variables and set split diagram

Before conducting any analysis, the collected data was pre-processed to ensure its quality and usability. This involved several steps, including data cleaning, outlier detection, and feature engineering. Missing values were handled using appropriate imputation techniques, and outliers were identified and either corrected or removed from the dataset. The categorical variables were encoded, and numerical variables

were standardized to ensure consistency across the datasets. In addition, a data exploration was conducted.

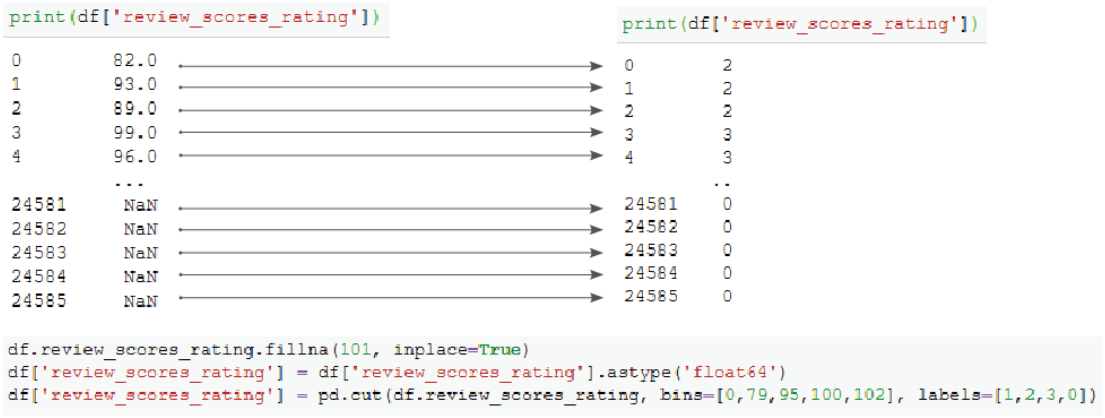


Figure 2. Redistribution of values in the variable review_scores_rating

The next step involves building predictive price models using K-nearest neighbours and linear regression algorithms. Additionally, a classification model will be created to accurately locate the optimal area for an Airbnb listing.

The tools for the data analysis and the ML model are Python, the Scikit Library, Matplot and Pandas and GIS for data mapping. These tools provided robust functionality for data manipulation, machine learning modelling, and visualization, facilitating the research process and analysis of results.

Once the data is prepared, and the ML results are obtained, an urban context analysis is conducted for the three cities to derive insights and facilitate comparison among the study cases. This analysis delves into various aspects of urban dynamics, including population distribution, housing market trends, neighbourhood characteristics, and tourism patterns. By examining these factors within the context of each city, we aim to extract meaningful conclusions and identify similarities and differences among Berlin, Madrid, and Chicago.

Following the development of predictive price models and classification algorithms, the analysis proceeds to leverage Python, the Scikit Library, Matplotlib, Pandas, and GIS tools for comprehensive data mapping and visualization. These tools play a pivotal role in facilitating the integration of machine learning outcomes with spatial and contextual data, enabling a nuanced exploration of urban dynamics across Berlin, Madrid, and Chicago.

Through this iterative process of data analysis and interpretation, the research endeavors to extract actionable insights to inform policy-making, urban planning, and strategic decision-making processes. By fostering a nuanced understanding of urban complexities, this interdisciplinary approach empowers stakeholders to devise targeted interventions that promote sustainable growth, equitable development, and enhanced quality of life across diverse urban contexts.

3 Data Preparation

Data Cleaning:

- **Handling Missing Values:** Any missing values in essential columns, such as price, availability, and neighbourhood, were addressed. Depending on the nature of the missing data, techniques such as imputation or removal were applied.
- **Removing Outliers:** Outliers, which could skew the predictive models, were identified and removed. This step was crucial to ensure the accuracy and reliability of the predictive models.
- **Feature Engineering:** New features were created to enhance the predictive power of the models. For example, features like the average price per neighbourhood, the number of amenities provided, and the proximity to landmarks or tourist attractions were derived from the existing dataset.

```
berlin_data['price'] = berlin_data['price'].str.replace(r'$','')
berlin_data['price'] = berlin_data['price'].str.replace(r',','')
berlin_data['price'] = berlin_data['price'].astype('float64')
berlin_data['zipcode'] = berlin_data['zipcode'].astype('str')
for i in berlin_data['zipcode']:
    berlin_data['zipcode'] = i[0:4]
berlin_data['zipcode'] = berlin_data['zipcode'].astype('int64')
for row in berlin_data:
    berlin_data[row] = berlin_data[row].astype('float64')
berlin_data = berlin_data.dropna()
```

Figure 3. Python code sample preparing the data

Data Encoding:
Categorical variables were encoded using techniques such as one-hot encoding or label encoding to convert them into numerical representations suitable for machine learning algorithms. This step was necessary as most machine learning algorithms require numerical inputs.

Data Splitting:
The dataset was split into training and testing sets to evaluate the performance of the predictive models. Typically, the data is divided into a training set (used to train the model) and a testing set (used to evaluate the model's performance).

Scaling:
Continuous variables were scaled to ensure that all features contributed equally to the model training process. Techniques like min-max scaling or standardization were applied to bring all feature values within a similar range.

Feature Selection:
Feature selection techniques, such as correlation analysis or recursive feature elimination, were employed to identify the most relevant features for predicting listing prices. This helped in reducing the dimensionality of the dataset and improving the efficiency of the predictive models.

	accommodates	bathrooms	bedrooms	beds	price	security_deposit	cleaning_fee	guests_included	extra_people
0	1.0	1.0	1.0	1.0	21.0	0.0	0.0	1.0	10.0
1	4.0	1.0	1.0	2.0	90.0	300.0	100.0	2.0	20.0
2	1.0	1.0	1.0	1.0	28.0	250.0	30.0	1.0	18.0
3	2.0	1.0	1.0	1.0	125.0	0.0	39.0	1.0	0.0
4	2.0	1.0	1.0	2.0	33.0	0.0	0.0	1.0	25.0
...	...	...	...	...	...	...	...	...	...
24581	3.0	1.0	0.0	1.0	77.0	0.0	0.0	1.0	25.0
24582	4.0	1.0	2.0	2.0	140.0	300.0	30.0	1.0	0.0
24583	1.0	1.5	1.0	1.0	20.0	0.0	15.0	1.0	14.0
24584	2.0	1.0	1.0	2.0	35.0	0.0	0.0	1.0	0.0
24585	3.0	1.0	1.0	2.0	35.0	0.0	50.0	2.0	5.0

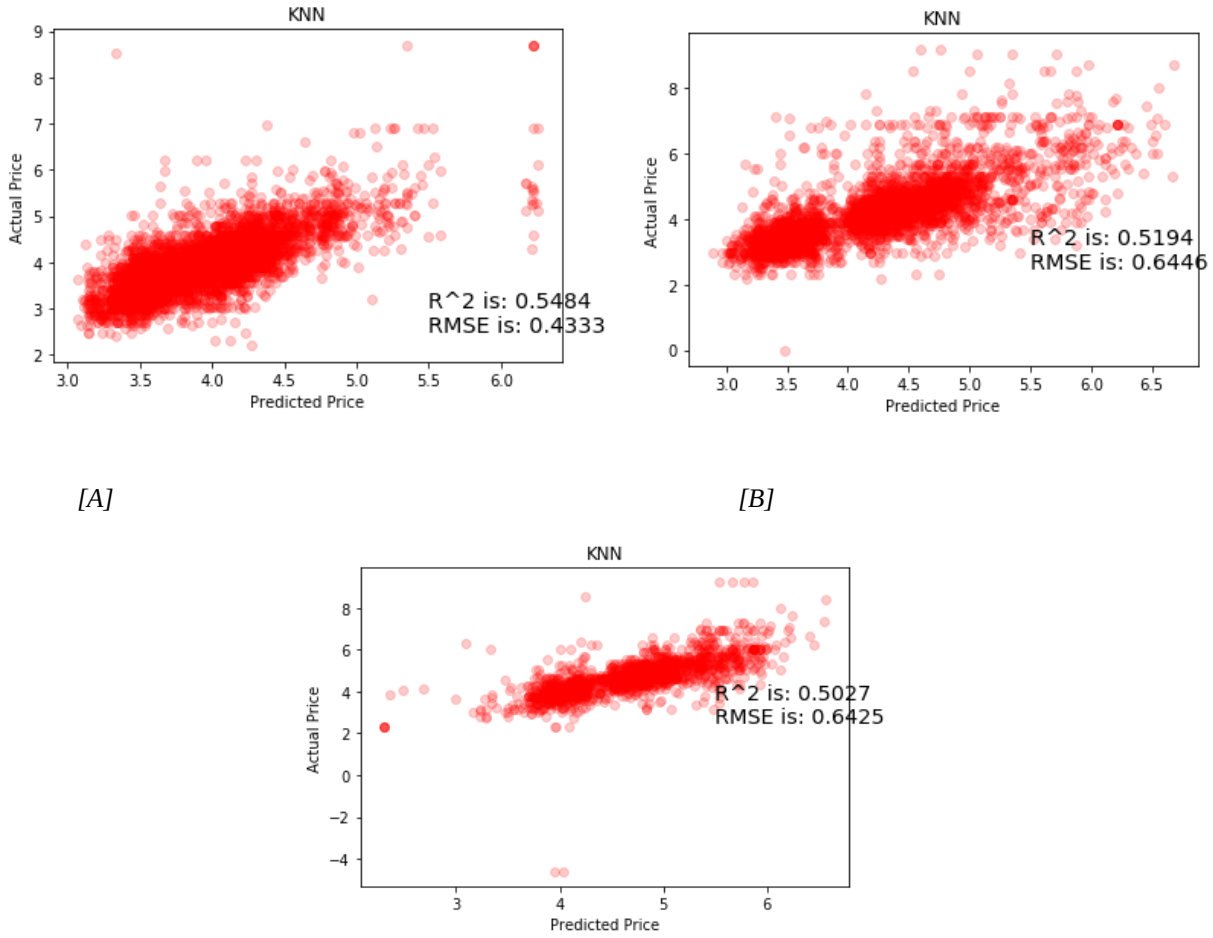
Figure 4. Processed Berlin dataset sample

4 ML Model Price Prediction development

Optimizing the Number of Neighbours: To determine the optimal number of neighbours, we employ the cross-validation method. This technique involves partitioning the dataset into multiple subsets and computing the mean accuracy of the prediction model across these partitions. By testing different ranges of neighbour values (0-20, 20-50, 50-80), we estimate the R2 score and identify the point where it reaches its maximum value.

Algorithm Training: Following data preparation and determination of the optimal number of neighbours, we proceed to train the k-nearest neighbour model using KNeighborsRegressor from Scikit-Learn. We allocate 80% of the dataset to the predictor variables (x_train) and the corresponding prices (y_train) for training purposes. Utilizing the 'auto' option, KNeighborsRegressor automatically selects the algorithm that best suits the dataset structure to compute the nearest neighbours. Subsequently, the trained model is tested on the remaining 20% of the dataset, where the predictor variables (x_test) are utilized as test data to generate predictions. In this instance, the 'kd-tree' algorithm is identified as the most suitable for the dataset structure.

Model Evaluation Metrics: We employ Root Mean Squared Error (RMSE) and R-squared (R2) to assess accuracy. Our model achieved an R2 value of 0.548, indicating that it explains approximately 55% of the variance in the data.



[A] *[B]*

[C]

Figure 6: Actual price and predicted price accuracy in the knn model. [A] Berlin, [B] Madrid, [C] Chicago

Testing Other Regression Algorithms: Linear Regression. To enhance the accuracy of price predictions, we explore the implementation of a linear regression model. Following the same data processing steps as the k-nearest neighbours' model, we allocate 80% of the data (x_train, y_train) for training, reserving the remaining 20% (x_test, y_test) for testing the prediction model. Once again, we assess the accuracy of the model using Root Mean Squared Error (RMSE) and R-squared (R2). Unfortunately, there is no significant improvement observed compared to the previous model. The recalculated RMSE value is slightly higher, and the R2 value is lower than that of the previous model, indicating that the linear regression model performs less accurately than the k-nearest neighbour's model.

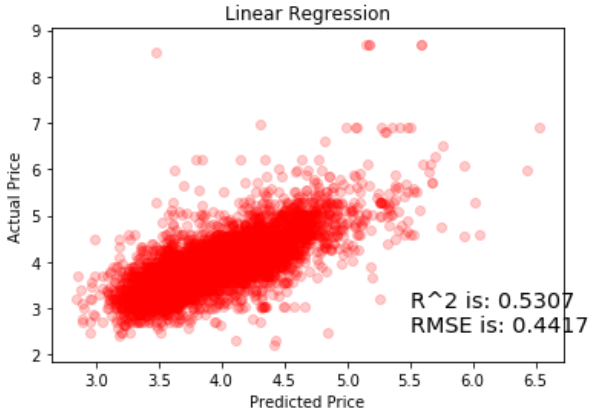


Figure 7: Actual price and predicted price accuracy in the linear regression model for Berlin

5 Analysis of Relationships

In this section, we delve into examining the relationships between different variables within the dataset to uncover underlying patterns and correlations indicative of urban phenomena. The analysis aims to discern how numerous factors interact and influence each other within the context of urban environments.

There is a clear clustering per colour in some charts, indicating that district location has a strong correlation with many variables such as the number of beds and bedrooms, which is likely related to the predominant size of apartments in different neighbourhoods.

Additionally, the number of reviews appears to be inversely proportional to the number of beds, indicating that accommodations with fewer beds tend to receive more reviews. This could imply that smaller accommodations are more frequently booked, leading to a higher turnover of guests and thus more opportunities for reviews. Conversely, larger accommodations may attract fewer guests and consequently receive fewer reviews. Further investigation into this relationship could provide insights into guest satisfaction levels and booking trends across different types of accommodations.

In addition, we observe that the score of the location and the prices have no discernible relationship, suggesting that the perceived quality of a location does not necessarily correlate with the pricing of accommodations. This finding raises questions about the factors that influence pricing decisions on the platform and underscores the complexity of pricing dynamics in urban environments. Further exploration into this discrepancy could shed light on the determinants of pricing strategies adopted by Airbnb hosts and the perceived value of location among guests.

Correlation Graphs can be found in the annex

6 Dataset Mapping

Berlin stands out with the highest number of Airbnb listings relative to its population, followed by Madrid and then Chicago. This suggests a bustling activity on the platform in Berlin despite its smaller population compared to the other two cities.

Chicago boasts the highest accommodation prices among the three cities, followed by Madrid and Berlin. This price disparity may stem from various factors such as supply and demand dynamics, local regulations, and the availability of luxury accommodations.

While the urban centre typically hosts the bulk of Airbnb listings in all three cities, interestingly, districts outside the downtown core in Berlin and Chicago also exhibit significant Airbnb activity.

Although a clear centre-periphery polarization exists in land and rental prices across the cities, this trend is less pronounced in Airbnb accommodation prices. However, prices tend to be higher in areas closer to the urban centre in all three cities.

Berlin leads in terms of Airbnb accommodation availability, followed by Madrid and Chicago. This suggests a relatively higher supply of accommodations in Berlin, possibly influenced by local regulations and tourism demand.

Full houses or apartments are the predominant type of accommodation in all three cities, with Berlin offering a higher proportion of private rooms compared to Madrid and Chicago.

Regarding minimum stay requirements, Berlin demonstrates greater homogeneity due to local regulations, while Chicago and Madrid exhibit wider ranges, with shorter minimum stays in peripheral areas and longer stays in the city centre.

Tourist activity is concentrated in specific areas within each city, particularly around the urban centre, which correlates with the distribution of Airbnb accommodations.

In summary, the comparison of Airbnb data across Berlin, Madrid, and Chicago highlights both similarities and differences in accommodation supply, pricing, availability, and location. Each city's short-term rental market dynamics are shaped by a combination of factors including local regulations, tourism demand, and socio-economic characteristics.

Conclusions from the analysis of the city through data

It is possible to interpret the layers of the city through data and draw conclusions about patterns or behaviours of the data in its urban context.

The exploration of the data has followed the criterion of urban organization of each city, namely the districts. Although the results show that this was an appropriate approach to understanding the city, it could have also been analysed using other criteria such as the distribution of data by area.

Due to the extensive nature of the data, the most relevant variables have been selected based on the Airbnb user’s perspective. However, there are other variables whose analysis in their urban context could be of great interest, such as the number of guests or the seasonality of prices.

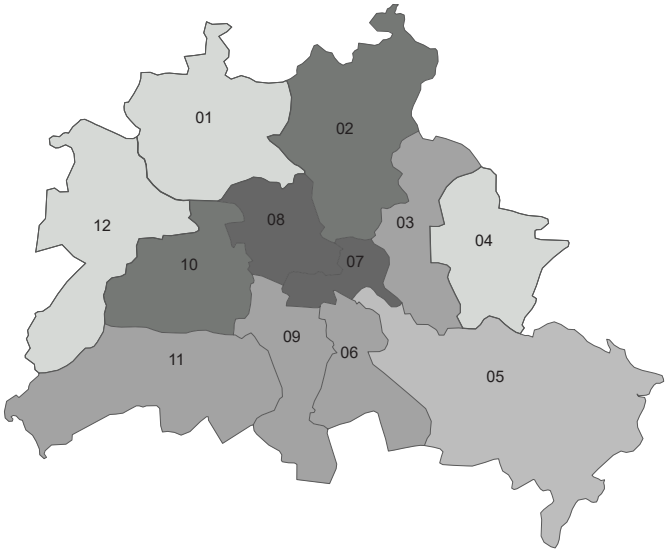
The behaviour of the distribution of Airbnb data and variables is different in each analysed city. However, the analysis results are predictable as the behaviour of Airbnb data follows the trends of the city, although there may be some anomalies. This reflects the reality of differences in urban fabric such as rental and land prices or the tourist polarization of the city (centre vs. periphery), which in turn can be related to the socioeconomic and demographic context in the urban setting.

Urban tourism is evolving in tandem with demographic, economic, and technological changes, which in turn impact cities. Through Airbnb data, trends can be observed, common behavioural areas can be detected, and patterns of tourism in the city can be identified. It is important to understand how this affects cities in order to promote positive changes and minimize negative impacts such as gentrification on urban fabric.

In addition to the insights gleaned from Airbnb data, it's crucial to consider broader urban dynamics and their interplay with tourism trends. Beyond just the distribution of Airbnb listings across different districts, analyzing factors like transportation infrastructure, cultural amenities, and economic development can provide a more comprehensive understanding of how tourism shapes and is shaped by the city.

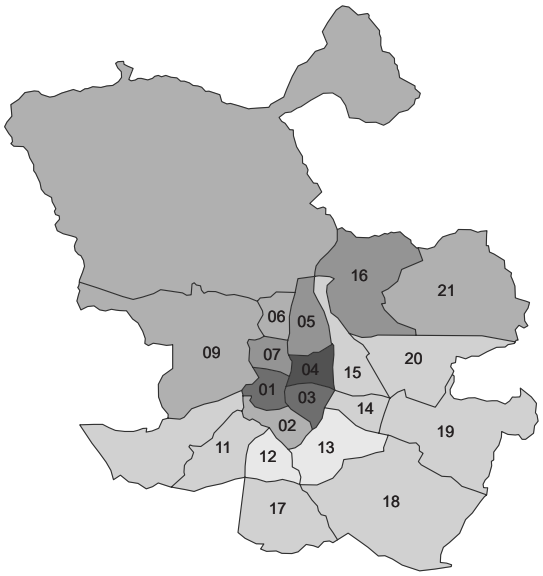
For instance, examining the relationship between the number of guests and transportation hubs or popular attractions can reveal key insights into visitor behavior and preferences. Moreover, delving into the seasonality of prices can shed light on the ebb and flow of tourism throughout the year, offering opportunities for targeted marketing campaigns or infrastructure improvements to accommodate peak periods.

Ultimately, leveraging data-driven insights from platforms like Airbnb can empower cities to foster sustainable tourism practices while preserving their unique character and supporting inclusive growth. By identifying common behavioral patterns, anticipating emerging trends, and proactively addressing challenges, cities can harness the transformative potential of tourism to enhance livability, economic vitality, and cultural vibrancy for residents and visitors alike.



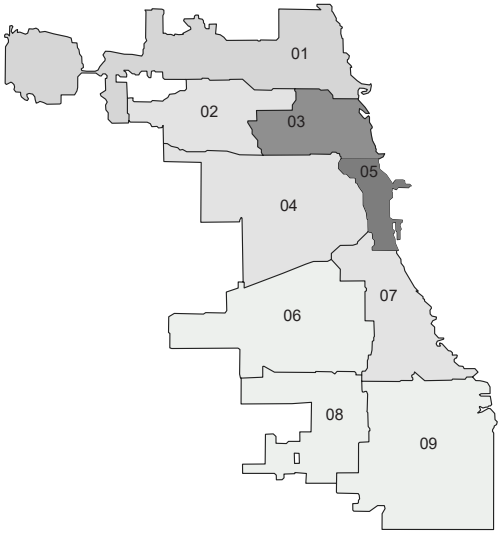
01 Reinickendorf
02 Pankow
03 Lichtenberg
04 Marzahn - Hellersdorf
05 Treptow - Köpenick
06 Neukölln
07 Friedrichshain-Kreuzberg
08 Mitte
09 Tempelhof - Schöneberg
10 Charlottenburg-Wilmersdorf
11 Steglitz - Zehlendorf
12 Spandau

2100-2600 €/m²
2600-3100 €/m²
3100-3600 €/m²
3600-4100 €/m²
4100-4600 €/m²



01 Centro
02 Arganzuela
03 Retiro
04 Salamanca
05 Chamartín
06 Tetuán
07 Chamberí
08 Fuencarral - El Pardo
09 Moncloa
10 Latina
11 Carabanchel
12 Usera
13 Pte. Vallecas
14 Moratalaz
15 Ciudad Lineal
16 Hortaleza
17 Villaverde
18 Villa de Vallecas
19 Vicálvaro
20 San Blas - Canillejas
21 Barajas

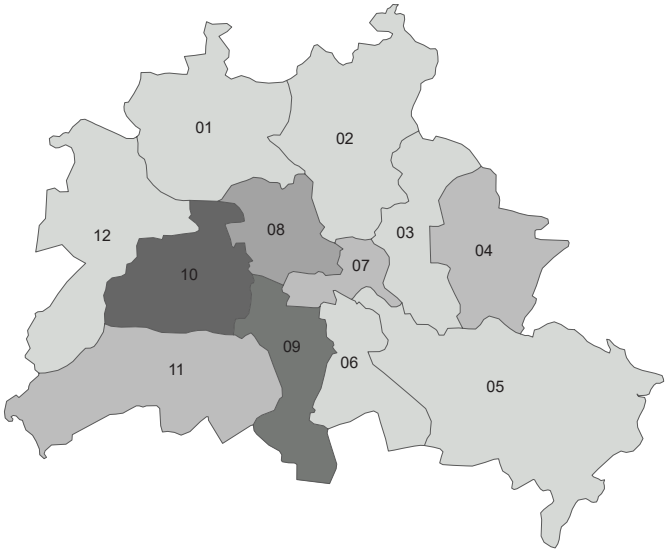
0-1500 €/m²
1500-2500 €/m²
2500-3500 €/m²
3500-4500 €/m²
4500-5500 €/m²
5500-6500 €/m²



01 Far North Side
02 Northwest Side
03 North Side
04 West Side
05 Central
06 Southwest Side
07 South Side
08 Far Southwest Side
09 Far Southeast Side

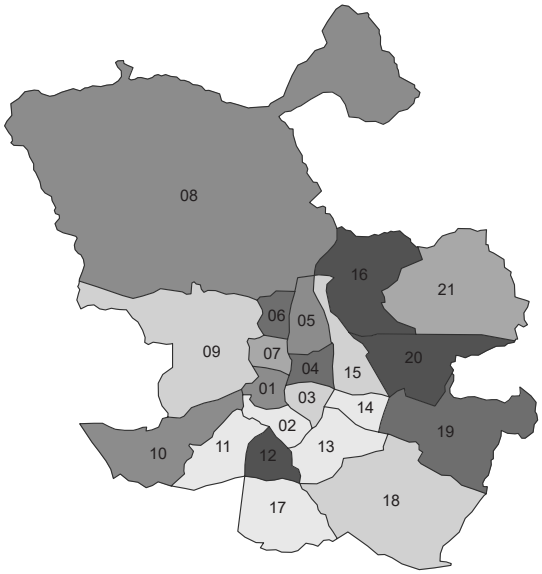
750-900 $/mes
900-1050 $/mes
1050-1200 $/mes
1200-1350 $/mes
1350-1500 $/mes
1500-1750 $/mes
1750-1900 $/mes

Fig. 8: Land and rent prices in Berlin, Madrid, and Chicago



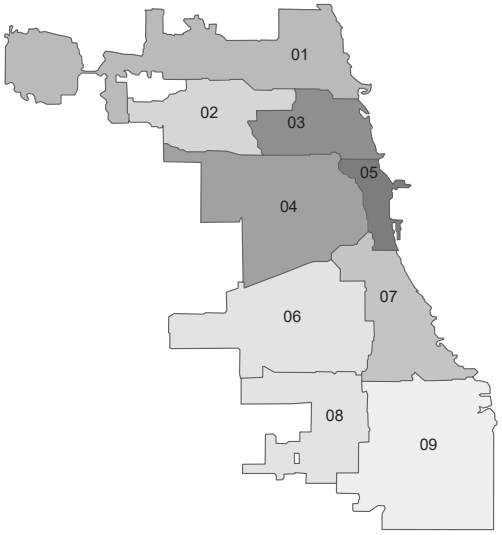
01 Reinickendorf
02 Pankow
03 Lichtenberg
04 Marzahn - Hellersdorf
05 Treptow - Köpenick
06 Neukölln
07 Friedrichshain-Kreuzberg
08 Mitte
09 Tempelhof - Schöneberg
10 Charlottenburg-Wilmersdorf
11 Steglitz - Zehlendorf
12 Spandau

50-60 € per night
60-70 € per night
70-80 € per night
90-100 € per night
100-120 € per night



01 Centro
02 Arganzuela
03 Retiro
04 Salamanca
05 Chamartín
06 Tetuán
07 Chamberí
08 Fuencarral - El Pardo
09 Moncloa
10 Latina
12 Usera
13 Pte. Vallecas
14 Moratalaz
15 Ciudad Lineal
16 Hortaleza
17 Villaverde
18 Villa de Vallecas
19 Vicálvaro
20 San Blas - Canillejas
21 Barajas

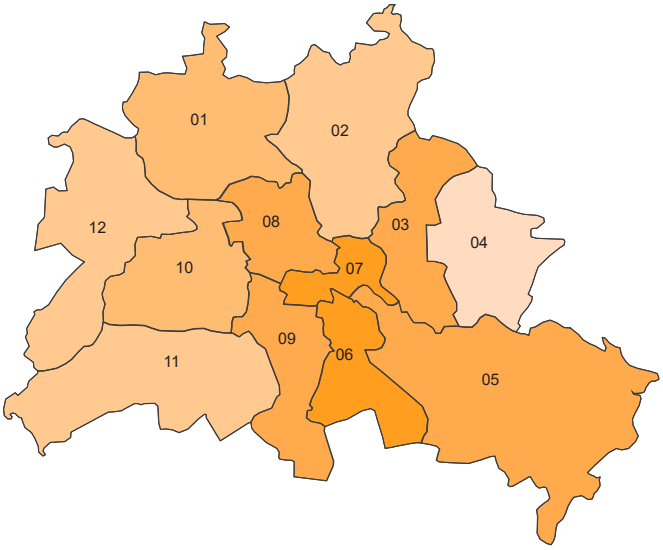
90-100 € per night
100-110 € per night
120-130 € per night
130-140 € per night
140-160 € per night
160-400 € per night



01 Far North Side
02 Northwest Side
03 North Side
04 West Side
05 Central
06 Southwest Side
07 South Side
08 Far Southwest Side
09 Far Southeast Side

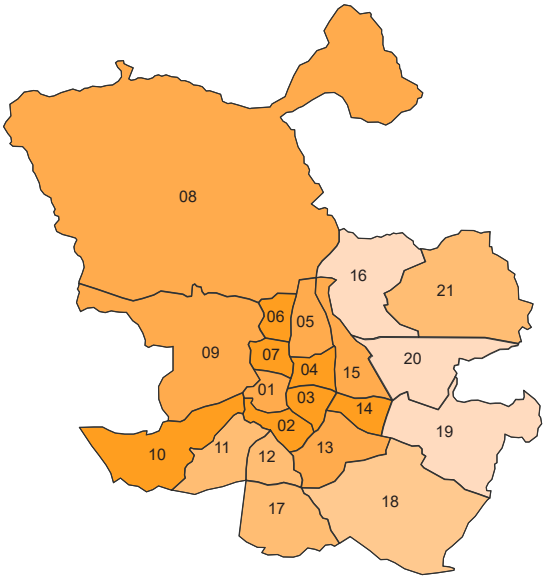
60-70 $ per night
70-80 $ per night
90-100 $ per night
100-110 $ per night
120-170 $ per night
200-300 $ per night
300-400 $ per night

Fig. 9: Accommodation prices per night in Berlin, Madrid, and Chicago.



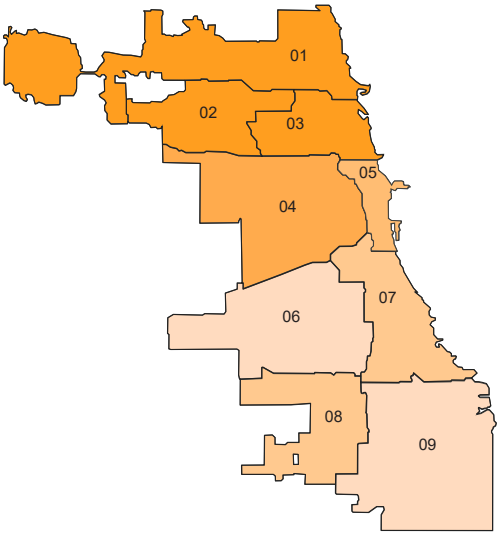
01 Reinickendorf
02 Pankow
03 Lichtenberg
04 Marzahn - Hellersdorf
05 Treptow - Köpenick
06 Neukölln
07 Friedrichshain-Kreuzberg
08 Mitte
09 Tempelhof - Schöneberg
10 Charlottenburg-Wilmersdorf
11 Steglitz - Zehlendorf
12 Spandau

39-45 available nights
32-39 available nights
24-32 available nights
17-24 available nights
0-17 available nights



01 Centro
02 Arganzuela
03 Retiro
04 Salamanca
05 Chamartín
06 Tetuán
07 Chamberí
08 Fuencarral
09 Moncloa
10 Latina
11 Carabanchel
12 Usera
13 Pte. Vallecas
14 Moratalaz
15 Ciudad Lineal
16 Hortaleza
17 Villaverde
18 Villa de Vallecas
19 Vicálvaro
20 San Blas - Canillejas
21 Barajas

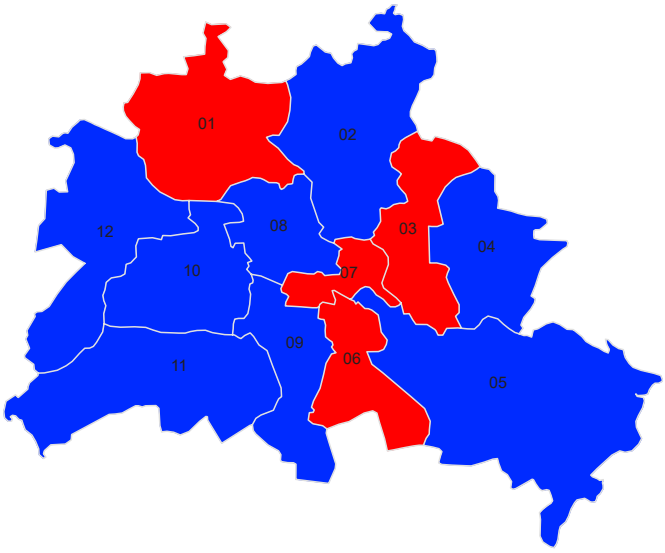
51-55 available nights
47-51 available nights
43-47 available nights
39-43 available nights
35-39 available nights



01 Far North Side
02 Northwest Side
03 North Side
04 West Side
05 Central
06 Southwest Side
07 South Side
08 Far Southwest Side
09 Far Southeast Side

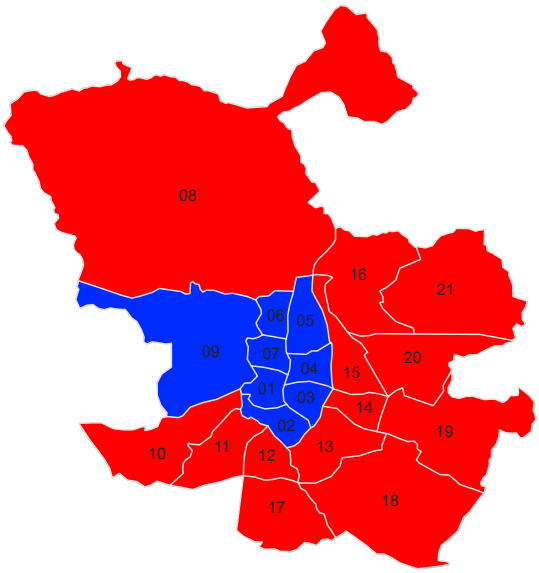
66-70 available nights
62-66 available nights
58-62 available nights
54-58 available nights
50-54 available nights

Fig. 10: Availability by district in Berlin, Madrid, and Chicago.



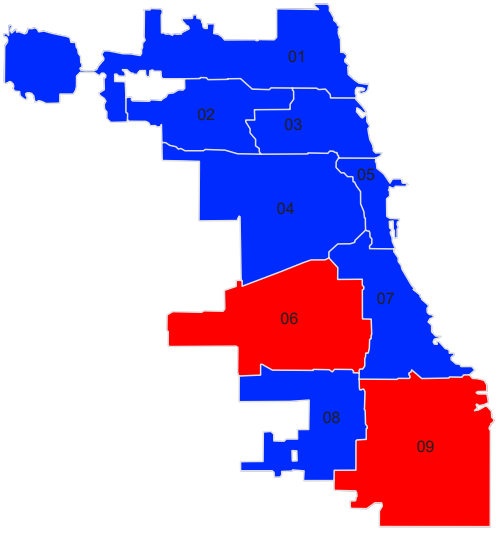
01 Reinickendorf
02 Pankow
03 Lichtenberg
04 Marzahn - Hellersdorf
05 Treptow - Köpenick
06 Neukölln
07 Friedrichshain-Kreuzberg
08 Mitte
09 Tempelhof - Schöneberg
10 Charlottenburg-Wilmersdorf
11 Steglitz - Zehlendorf
12 Spandau

Entire home/apt
Private Room



01 Centro
02 Arganzuela
03 Retiro
04 Salamanca
05 Chamartín
06 Tetuán
07 Chamberí
08 Fuencarral
09 Moncloa
10 Latina
11 Carabanchel
12 Usera
13 Pte. Vallecas
14 Moratalaz
15 Ciudad Lineal
16 Hortaleza
17 Villaverde
18 Villa de Vallecas
19 Vicálvaro
20 San Blas - Canillejas
21 Barajas

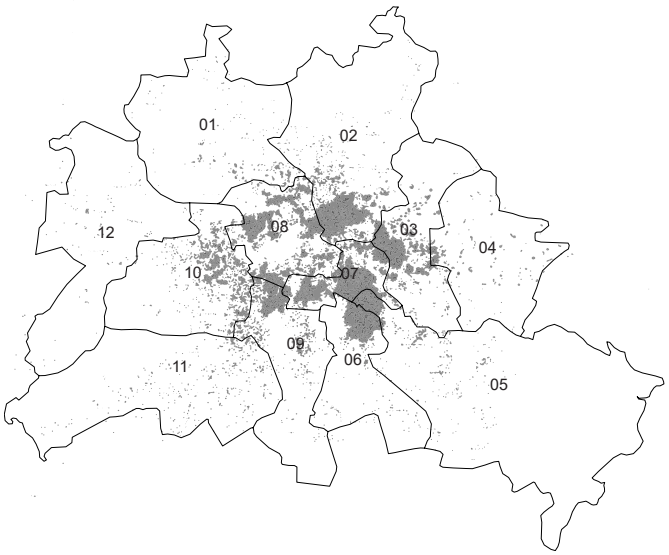
Entire home/apt
Private Room



01 Far North Side
02 Northwest Side
03 North Side
04 West Side
05 Central
06 Southwest Side
07 South Side
08 Far Southwest Side

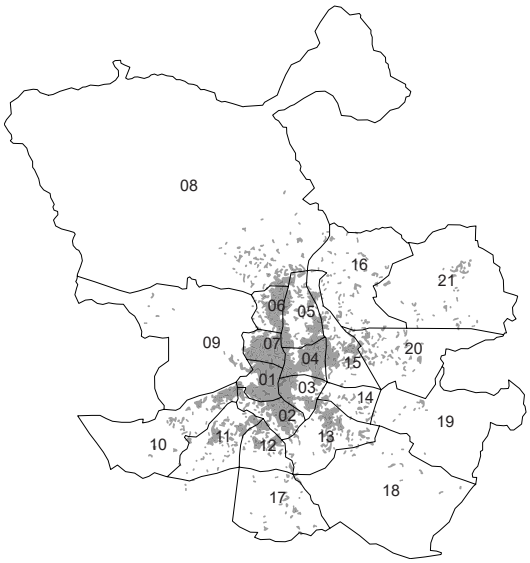
Entire home/apt
Private Room

Fig. 11: Type of accommodation by district in Berlin, Madrid, and Chicago.



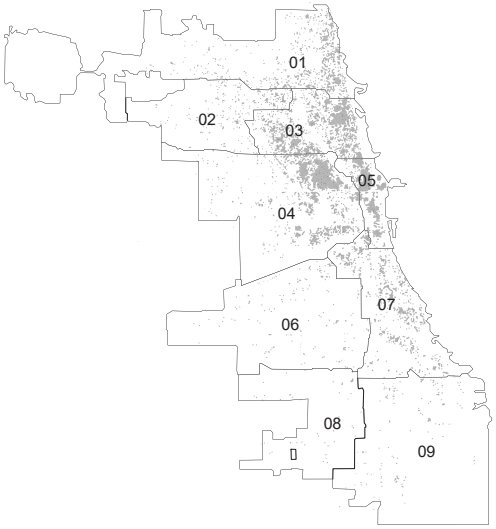
01 Reinickendorf
02 Pankow
03 Lichtenberg
04 Marzahn - Hellersdorf
05 Treptow - Köpenick
06 Neukölln
07 Friedrichshain-Kreuzberg
08 Mitte
09 Tempelhof - Schöneberg
10 Charlottenburg-Wilmersdorf
11 Steglitz - Zehlendorf
12 Spandau

● Accomodation



01 Centro
02 Arganzuela
03 Retiro
04 Salamanca
05 Chamartín
06 Tetuán
07 Chamberí
08 Fuencarral
09 Moncloa
10 Latina
11 Carabanchel
12 Usera
13 Pte. Vallecas
14 Moratalaz
15 Ciudad Lineal
16 Hortaleza
17 Villaverde
18 Villa de Vallecas
19 Vicálvaro
20 San Blas - Canillejas
21 Barajas

● Accomodation

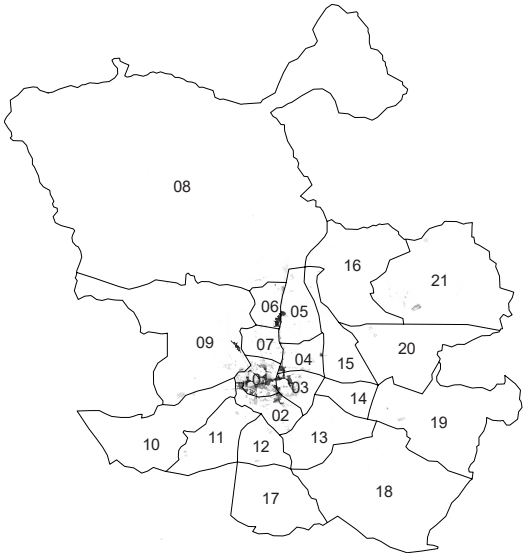


01 Far North Side
02 Northwest Side
03 North Side
04 West Side
05 Central
06 Southwest Side
07 South Side
08 Far Southwest Side
09 Far Southeast Side

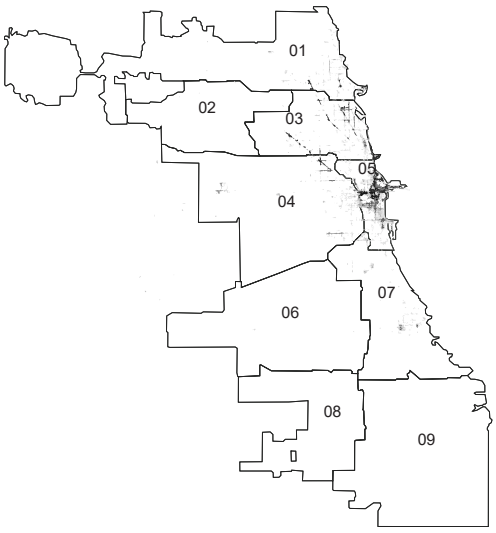
● Accomodation

Fig. 12: Distribution of Airbnb listings in Berlin, Madrid, and Chicago.

01 Reinickendorf
02 Pankow
03 Lichtenberg
04 Marzahn - Hellersdorf
05 Treptow - Köpenick
06 Neukölln
07 Friedrichshain-Kreuzberg
08 Mitte
09 Tempelhof - Schöneberg
10 Charlottenburg-Wilmersdorf
11 Steglitz - Zehlendorf
12 Spandau



01 Centro
02 Arganzuela
03 Retiro
04 Salamanca
05 Chamartín
06 Tetuán
07 Chamberí
08 Fuencarral
09 Moncloa
10 Latina
11 Carabanchel
12 Usera
13 Pte. Vallecas
14 Moratalaz
15 Ciudad Lineal
16 Hortaleza
17 Villaverde
18 Villa de Vallecas
19 Vicálvaro
20 San Blas - Canillejas
21 Barajas



01 Far North Side
02 Northwest Side
03 North Side
04 West Side
05 Central
06 Southwest Side
07 South Side
08 Far Southwest Side

Fig. 13: Tourism flows in Berlin, Madrid, and Chicago.

7 Prediction model mapping

Conclusions from Berlin Predictions

In the realm of price prediction, a noticeable narrowing of the range occurs. While the actual maximum average price per night falls within the range of 100-120 €, the predicted maximum is notably lower, ranging between 67-75 € per night. However, upon closer examination of the medians, we find that their ranges are more aligned. This discrepancy in the average price range is likely attributed to the presence of higher-priced outliers, which, though fewer in number, exert a significant influence on the regression model, ultimately leading to their loss in comparison to the more frequently occurring prices represented by the medians.

Regarding the distribution of prices across districts, the model's outcomes starkly contrast with the actual distribution. Nevertheless, a discernible zoning pattern in the price distribution is evident once again.

Conclusions from Madrid Predictions

In price prediction, we observe a notable narrowing of the range. However, there are discrepancies between the actual and predicted values. For instance, while the actual minimum average price per night falls within the range of 90-100 €, the predicted price is significantly lower, ranging between 30-50 € per night. Similarly, the actual maximum price spans from 160-400 € per night, whereas the predicted maximum ranges from 130-230 € per night. Although the medians exhibit closer ranges, they still deviate from the actual values. This disparity in average price range can be attributed to the presence of higher-priced outliers, which, although fewer in number, exert a substantial influence on the regression model, resulting in their loss compared to the more frequently occurring prices represented by the medians.

As for the distribution of prices by district, the model's outcomes diverge significantly from the actual distribution. However, there is still a discernible zoning pattern evident in the distribution of prices.

Conclusions from Chicago Predictions

In price prediction, while the range significantly narrows, the average prices per night tend to be higher. Specifically, although the maximum predicted average price (220-350 $ per night) falls below the actual maximum (300-400 $ per night), the predicted minimum (100-120 $ per night) surpasses the actual minimum (60-70 $ per night). Additionally, there is noticeable disparity between the actual and predicted medians.

Regarding the distribution of prices by district, the model's outcomes diverge considerably from the actual distribution. However, it remains evident that there is a clear zoning pattern in the distribution of prices, with the most expensive districts in the prediction being those where the actual price is lower (Far Southwest Side and Far Southeast Side).

The observed narrowing of price ranges in the predictive models suggests a degree of accuracy in forecasting average prices per night. However, the disparities between actual and predicted values highlight the inherent challenges in modeling complex urban phenomena. Factors such as changing market dynamics, local regulations, and unpredictable events can all contribute to deviations from predicted outcomes, necessitating ongoing refinement and validation of predictive algorithms.

Additionally, the identification of zoning patterns in the distribution of prices across districts underscores the utility of spatial analysis techniques in urban research. Understanding how prices vary geographically can inform targeted interventions and resource allocation strategies to address disparities and enhance the overall quality of life for residents. Moreover, by integrating spatial data with socio-economic indicators and demographic trends, policymakers can develop more holistic approaches to urban planning and development that prioritize equity and sustainability.

In conclusion, while predictive modeling offers valuable insights into urban dynamics, it is crucial to interpret results with caution and contextualize findings within broader socio-economic and spatial frameworks. By acknowledging the limitations of predictive algorithms and embracing a multidisciplinary approach to data analysis, researchers and policymakers can harness the transformative potential of data-driven insights to build more resilient, inclusive, and vibrant cities.

Maps with predicions can be found in the annex

8 Results and Conclusion

General conclusions

In conclusion, this study underscores the paramount importance of conducting thorough analysis and data preparation prior to constructing predictive models, particularly in the realm of urban dynamics and machine learning. The precision of results significantly improves after meticulous variable processing, highlighting the need for a deep understanding of the urban data context. Moreover, the size and quality of the dataset are pivotal determinants of model efficacy, with larger datasets enabling more robust training and predictions.

Furthermore, while some independent variables exhibit stronger correlations with the dependent variable, others may require more nuanced weighting adjustments to enhance predictive accuracy. As this study represents an initial foray into programming and machine learning, it underscores the learning curve inherent in these endeavours and the potential for optimization with greater expertise and algorithmic exploration.

However, it's essential to acknowledge the transient nature of urban dynamics and the imperative of continually updating models with fresh data to ensure their relevance over time. Despite these challenges, the developed model offers valuable insights into Airbnb pricing estimation and serves as a versatile tool for exploring urban behavioural patterns. By leveraging Airbnb data, we can gain deeper insights into evolving urban tourism trends and inform strategies for sustainable urban development.

9 References

- Diego Calvo ‘Aprendizaje supervisado’ (2019) http://www.diegocalvo.es/aprendizaje-supervisado/

- Onel Harrison ‘Machine Learning Basics with the K-Nearest Neighbors Algorithm’ (2018) https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

-Aishwarya Singh’ A Practical Introduction to K-Nearest Neighbors Algorithm for Regression (with Python code)’ (2018) https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

- Inside Airbnb. http://insideairbnb.com/

- ‘Codificación de One Hot Encoding de un conjunto de características categóricas con sklearn’ (2018) https://www.interactivechaos.com/python/scenario/codificacion-one-hot-encoding-de-un-conjunto-de-caracteristicas-categoricas-con

- Swetha Lakshmanan ‘How, When and Why Should You Normalize / Standardize / Rescale Your Data?’ (2019) https://medium.com/@swethalakshmanan14/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

- Scikit-Learn User’s Guide ‘Cross-validation: evaluating estimator performance’ https://scikit-learn.org/stable/modules/cross_validation.html

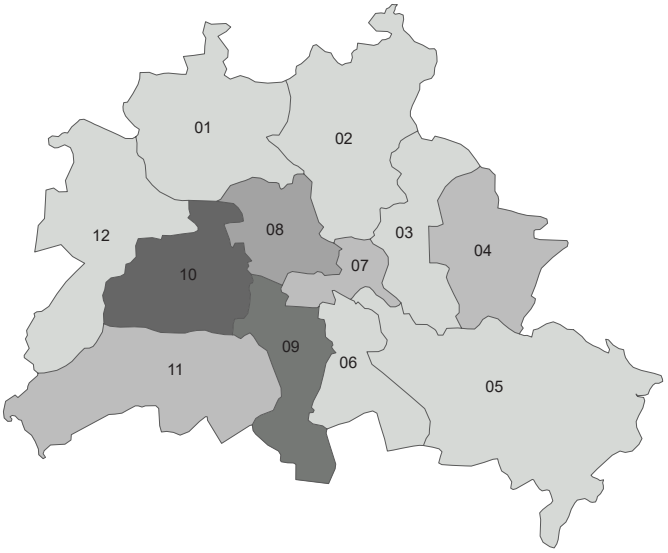
- Siemer, Matthews-Hunter. ‘The spatial pattern of gentrification in Berlin’ https://pcag.uwinnipeg.ca/Prairie-Perspectives/PP-Vol19/Siemer-MatthewsHunter.pdf

- Santiago Gómez Plata .‘El lenguaje arquitectónico de la inteligencia artificial’. Trabajo de Fin de Grado (2020) http://oa.upm.es/58140/1/TFG_20_Gomez_Plata_Santiago.pdf

- Artem M. Chirkin y Reinhard König .‘Concept of Interactive Machine Learning in Urban Design Problems’ (ETH Zurich)

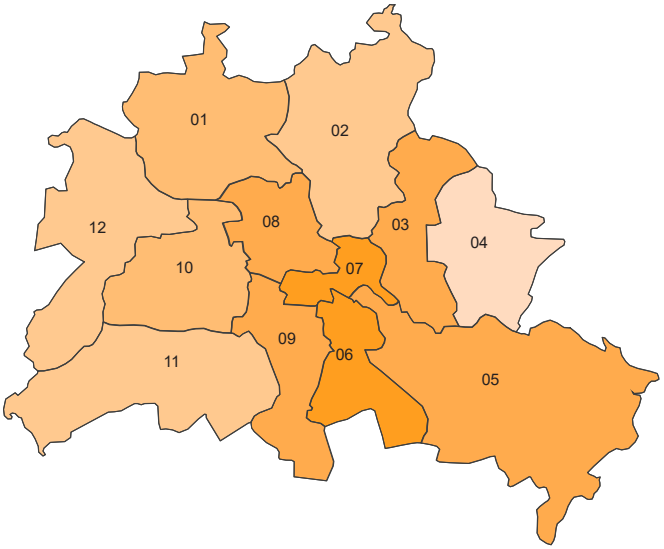
- Dr. Bharatendra Rai. ‘K-Nearest Neighbour (KNN) with R | Classification and Regression Examples’ (2018) https://www.youtube.com/watch?v=tSPg-JDAF4M&list=PLcTXSgeM4nvIFkPsrTd6FantYPZa8FqMW&index=11&t=484s

Annex: Charts and Maps



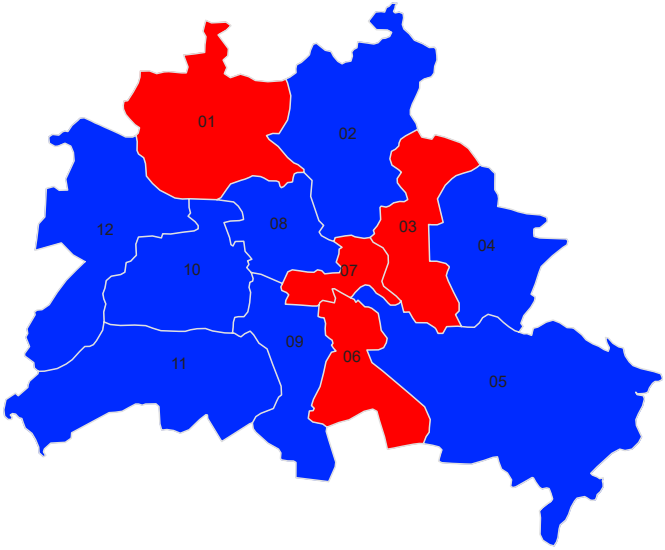
01 Reinickendorf
02 Pankow
03 Lichtenberg
04 Marzahn - Hellersdorf
05 Treptow - Köpenick
06 Neukölln
07 Friedrichshain-Kreuzberg
08 Mitte
09 Tempelhof - Schöneberg
10 Charlottenburg-Wilmersdorf
11 Steglitz - Zehlendorf
12 Spandau

50-60 € per night
60-70 € per night
70-80 € per night
90-100 € per night
100-120 € per night



01 Reinickendorf
02 Pankow
03 Lichtenberg
04 Marzahn - Hellersdorf
05 Treptow - Köpenick
06 Neukölln
07 Friedrichshain-Kreuzberg
08 Mitte
09 Tempelhof - Schöneberg
10 Charlottenburg-Wilmersdorf
11 Steglitz - Zehlendorf
12 Spandau

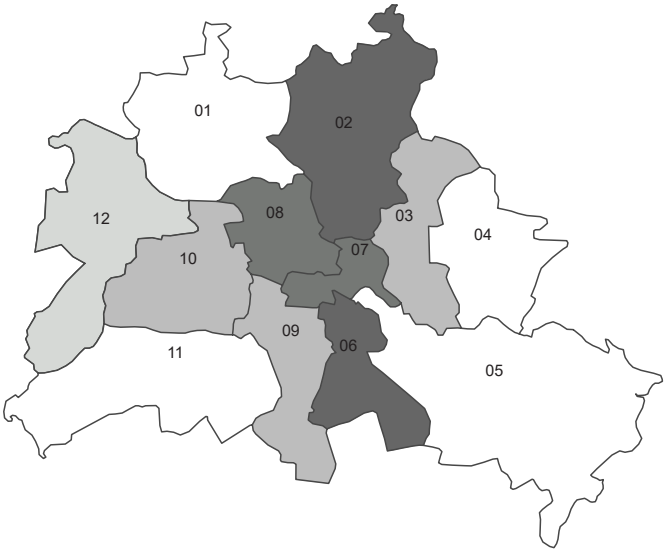
39-45 available nights
32-39 available nights
24-32 available nights
17-24 available nights
0-17 available nights



01 Reinickendorf
02 Pankow
03 Lichtenberg
04 Marzahn - Hellersdorf
05 Treptow - Köpenick
06 Neukölln
07 Friedrichshain-Kreuzberg
08 Mitte
09 Tempelhof - Schöneberg
10 Charlottenburg-Wilmersdorf
11 Steglitz - Zehlendorf
12 Spandau

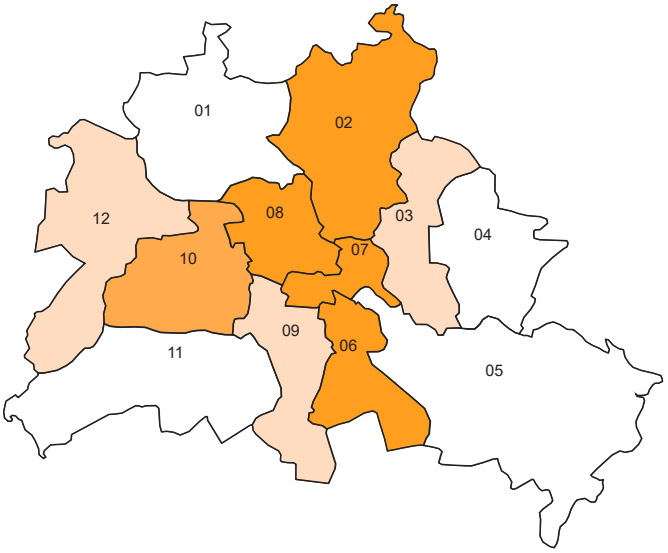
Entire home/apt
Private Room

Fig. 14: Top-down maps: distribution of average nightly price, demand, and accommodation type in Berlin from actual data.



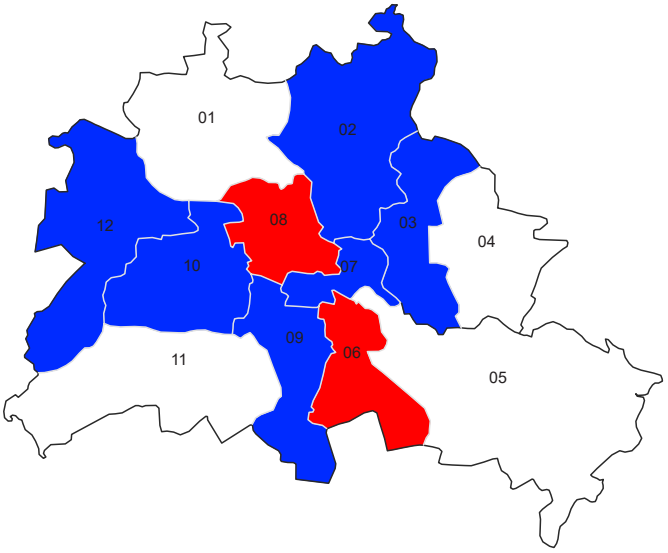
01 Reinickendorf
02 Pankow
03 Lichtenberg
04 Marzahn - Hellersdorf
05 Treptow - Köpenick
06 Neukölln
07 Friedrichshain-Kreuzberg
08 Mitte
09 Tempelhof - Schöneberg
10 Charlottenburg-Wilmersdorf
11 Steglitz - Zehlendorf
12 Spandau

40-46 € per night
46-52 € per night
52-58 € per night
58-64 € per night
64-70 € per night
no offer



01 Reinickendorf
02 Pankow
03 Lichtenberg
04 Marzahn - Hellersdorf
05 Treptow - Köpenick
06 Neukölln
07 Friedrichshain-Kreuzberg
08 Mitte
09 Tempelhof - Schöneberg
10 Charlottenburg-Wilmersdorf
11 Steglitz - Zehlendorf
12 Spandau

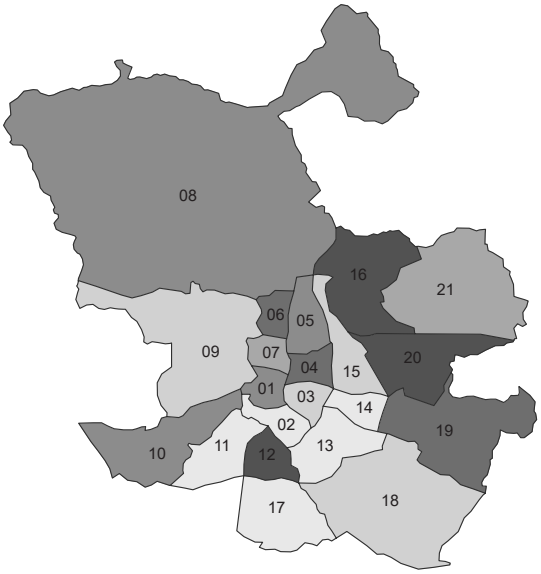
00-04 available nights
04-08 available nights
08-12 available nights
12-16 available nights
16-20 available nights
no offer



01 Reinickendorf
02 Pankow
03 Lichtenberg
04 Marzahn - Hellersdorf
05 Treptow - Köpenick
06 Neukölln
07 Friedrichshain-Kreuzberg
08 Mitte
09 Tempelhof - Schöneberg
10 Charlottenburg-Wilmersdorf
11 Steglitz - Zehlendorf
12 Spandau

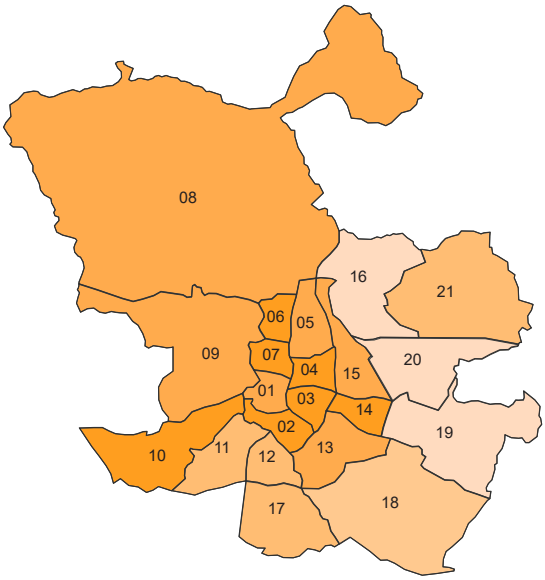
Entire home/apt
Private Room
no offer

Fig. 15: Top-down maps: distribution of average nightly price, demand, and accommodation type in Berlin from classification results.



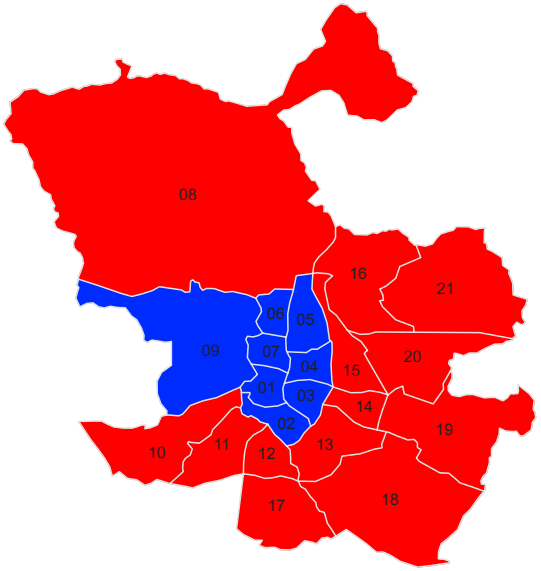
01 Centro
02 Arganzuela
03 Retiro
04 Salamanca
05 Chamartín
06 Tetuán
07 Chamberí
08 Fuencarral
09 Moncloa
10 Latina
11 Carabanchel
12 Usera
13 Pte. Vallecas
14 Moratalaz
15 Ciudad Lineal
16 Hortaleza
17 Villaverde
18 Villa de Vallecas
19 Vicálvaro
20 San Blas -
21 Barajas

90-100 € per night
100-110 € per night
120-130 € per night
130-140 € per night
140-160 € per night
160-400 € per night



01 Centro
02 Arganzuela
03 Retiro
04 Salamanca
05 Chamartín
06 Tetuán
07 Chamberí
08 Fuencarral
09 Moncloa
10 Latina
11 Carabanchel
12 Usera
13 Pte. Vallecas
14 Moratalaz
15 Ciudad Lineal
16 Hortaleza
17 Villaverde
18 Villa de Vallecas
19 Vicálvaro
20 San Blas -
21 Barajas

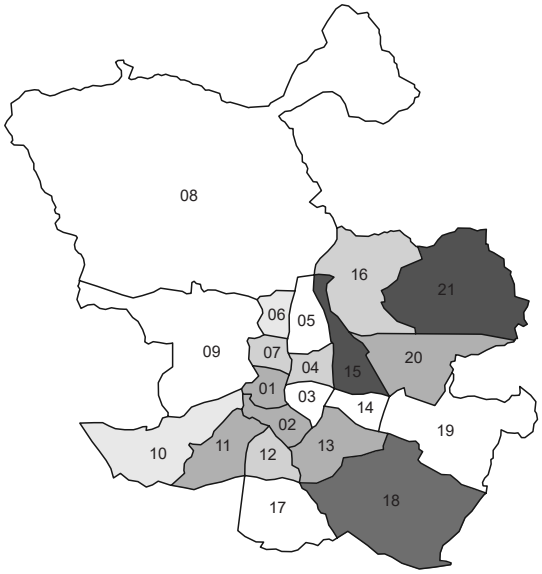
51-55 available nights
47-51 available nights
43-47 available nights
39-43 available nights
35-39 available nights



01 Centro
02 Arganzuela
03 Retiro
04 Salamanca
05 Chamartín
06 Tetuán
07 Chamberí
08 Fuencarral
09 Moncloa
10 Latina
11 Carabanchel
12 Usera
13 Pte. Vallecas
14 Moratalaz
15 Ciudad Lineal
16 Hortaleza
17 Villaverde
18 Villa de Vallecas
19 Vicálvaro
20 San Blas -
21 Barajas

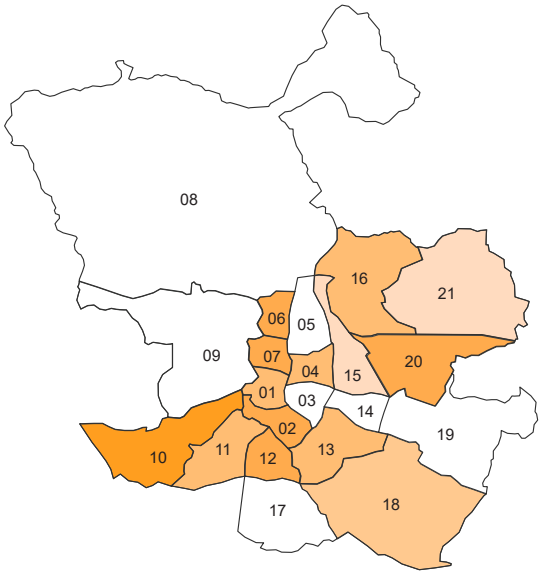
Entire home/apt
Private Room

Fig. 16: Top-down maps: distribution of average nightly price, demand, and accommodation type in Madrid from actual data.



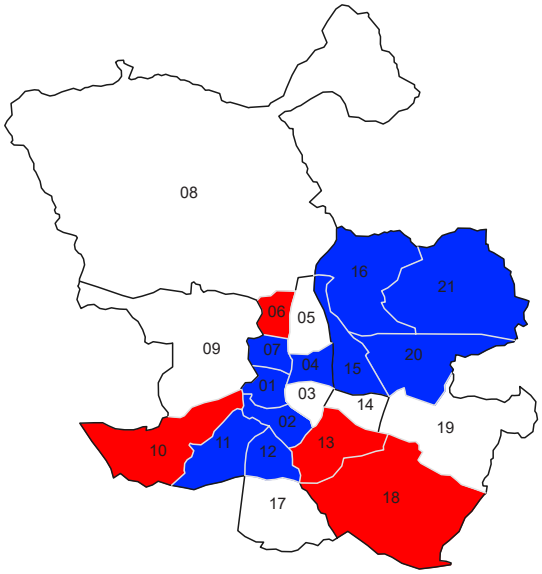
01 Centro
02 Arganzuela
03 Retiro
04 Salamanca
05 Chamartín
06 Tetuán
07 Chamberí
08 Fuencarral
09 Moncloa
10 Latina
11 Carabanchel
12 Usera
13 Pte. Vallecas
14 Moratalaz
15 Ciudad Lineal
16 Hortaleza
17 Villaverde
18 Villa de Vallecas
19 Vicálvaro
20 San Blas
21 Barajas

20-40 € per night
40-60 € per night
60-80 € per night
80-100 € per night
100-120 € per night
120-160 € per night
no offer



01 Centro
02 Arganzuela
03 Retiro
04 Salamanca
05 Chamartín
06 Tetuán
07 Chamberí
08 Fuencarral
09 Moncloa
10 Latina
11 Carabanchel
12 Usera
13 Pte. Vallecas
14 Moratalaz
15 Ciudad Lineal
16 Hortaleza
17 Villaverde
18 Villa de Vallecas
19 Vicálvaro
20 San Blas
21 Barajas

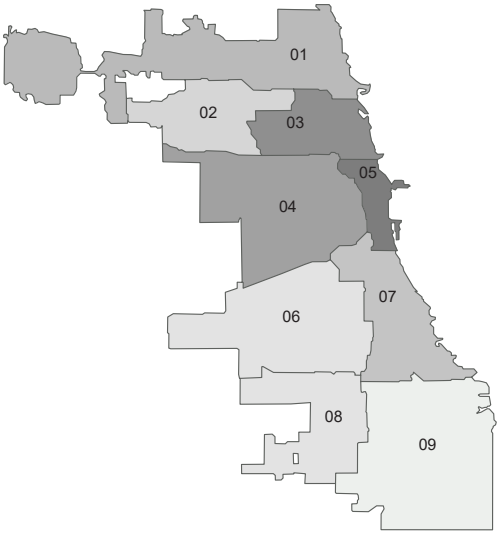
57-70 available nights
44-57 available nights
31-44 available nights
18-31 available nights
05-18 available nights
no offer



01 Centro
02 Arganzuela
03 Retiro
04 Salamanca
05 Chamartín
06 Tetuán
07 Chamberí
08 Fuencarral
09 Moncloa
10 Latina
11 Carabanchel
12 Usera
13 Pte. Vallecas
14 Moratalaz
15 Ciudad Lineal
16 Hortaleza
17 Villaverde
18 Villa de Vallecas
19 Vicálvaro
20 San Blas
21 Barajas

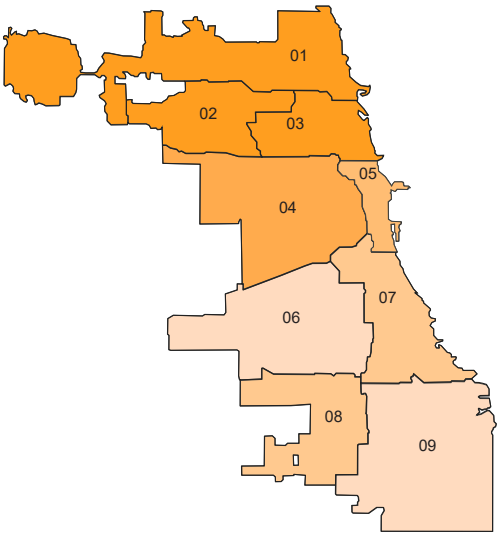
Entire home/apt
Private Room
no offer

Fig. 17: Top-down maps: distribution of average nightly price, demand, and accommodation type in Madrid from classification results.



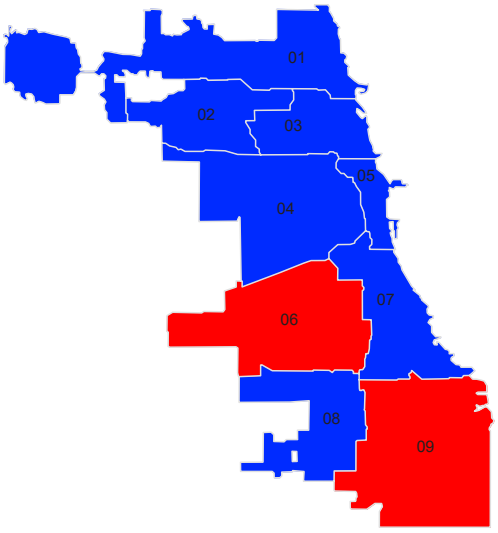
01 Far North Side
02 Northwest Side
03 North Side
04 West Side
05 Central
06 Southwest Side
07 South Side
08 Far Southwest Side
09 Far Southeast Side

60-70 $ per night
70-80 $ per night
90-100 $ per night
100-110 $ per night
120-170 $ per night
200-300 $ per night
300-400 $ per night



01 Far North Side
02 Northwest Side
03 North Side
04 West Side
05 Central
06 Southwest Side
07 South Side
08 Far Southwest Side
09 Far Southeast Side

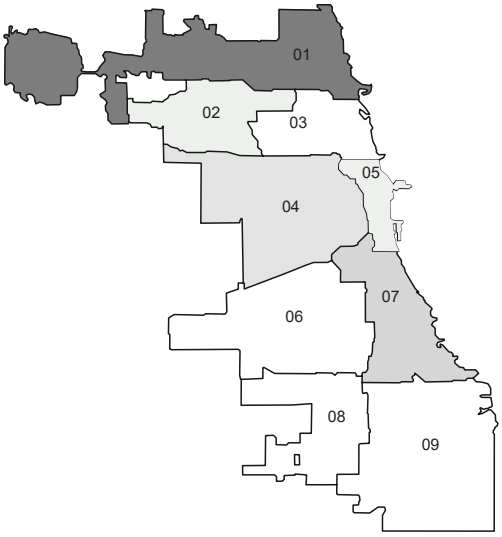
66-70 available nights
62-66 available nights
58-62 available nights
54-58 available nights
50-54 available nights



01 Far North Side
02 Northwest Side
03 North Side
04 West Side
05 Central
06 Southwest Side
07 South Side
08 Far Southwest Side
09 Far Southeast Side

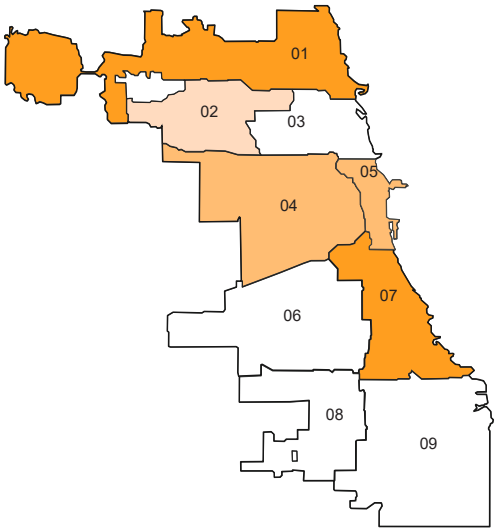
Entire home/apt
Private room

Fig. 18: Top-down maps: distribution of average nightly price, demand, and accommodation type in Chicago from actual data.



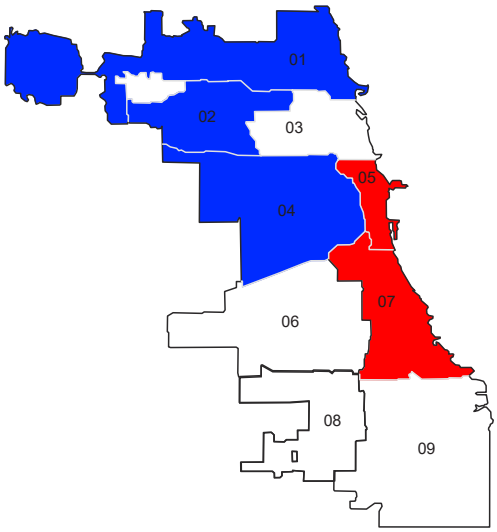
01 Far North Side
02 Northwest Side
03 North Side
04 West Side
05 Central
06 Southwest Side
07 South Side
08 Far Southwest Side
09 Far Southeast Side

100-108 $ per night
108-116 $ per night
116-124 $ per night
124-132 $ per night
132-140 $ per night
140-148 $ per night
148-170 $ per night
no offer



01 Far North Side
02 Northwest Side
03 North Side
04 West Side
05 Central
06 Southwest Side
07 South Side
08 Far Southwest Side
09 Far Southeast Side

62-66 available nights
58-62 available nights
54-58 available nights
50-54 available nights
46-50 available nights
no offer

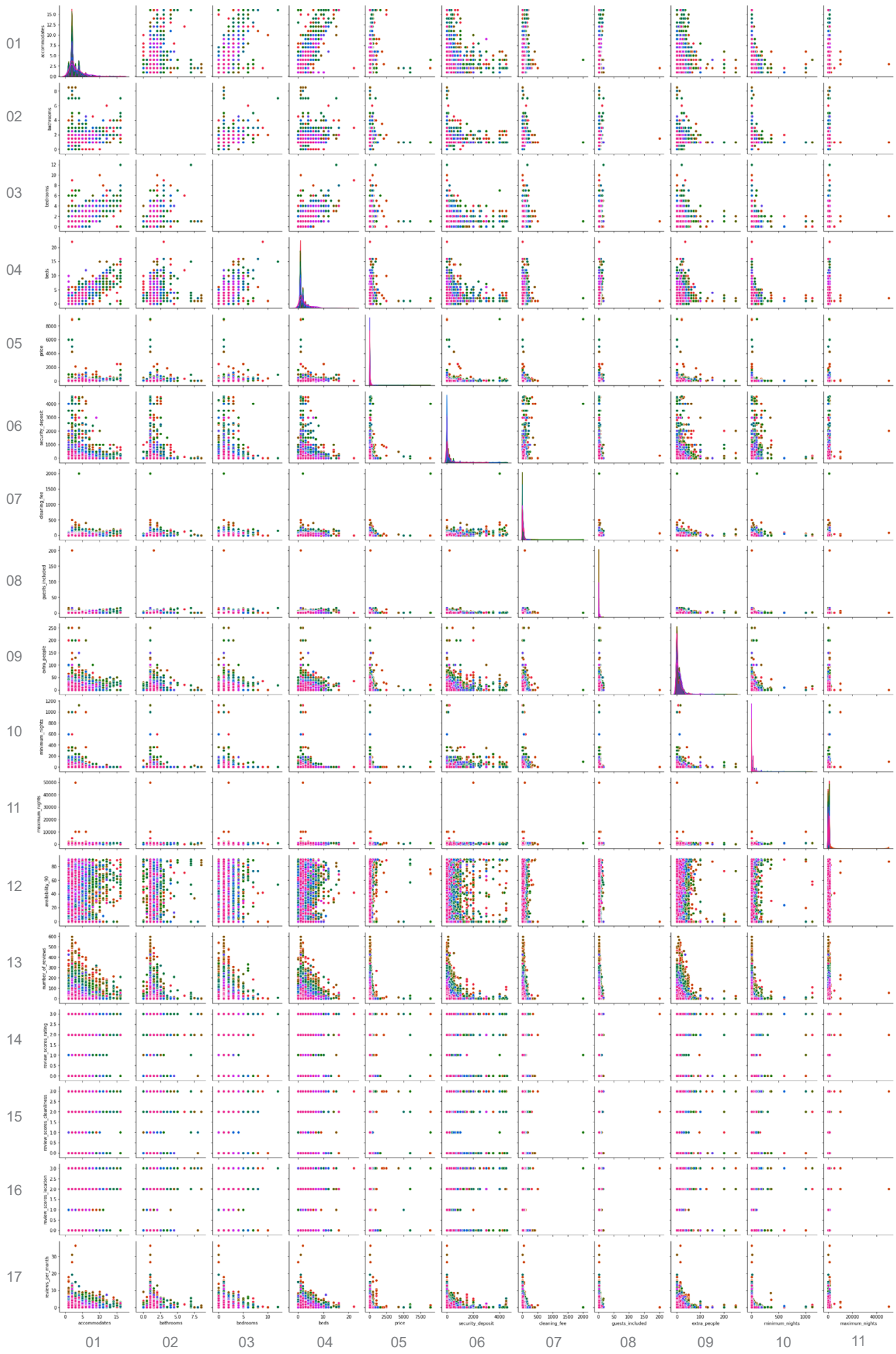


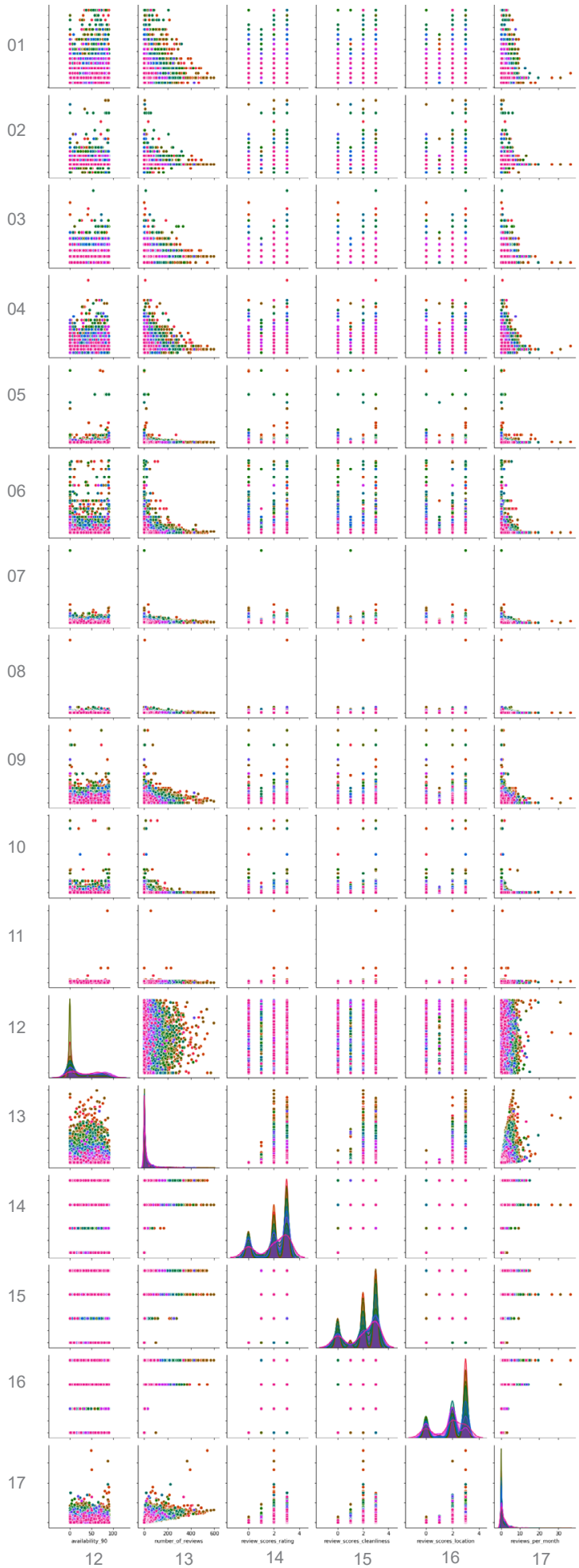
01 Far North Side
02 Northwest Side
03 North Side
04 West Side
05 Central
06 Southwest Side
07 South Side
08 Far Southwest Side
09 Far Southeast Side

Entire home/apt
Private room
no offer

Fig. 19: Top-down maps: distribution of average nightly price, demand, and accommodation type in Chicago from classification results.

Berlin


01
02
03
04
05
06
07
08
09
10
11
accommodates
bathrooms
bedrooms
beds
price
security_deposit
cleaning_fee
guests_included
extra_people
minimum_nights
maximum_nights
number_of_reviews
review_scores_rating
review_scores_cleanliness
review_scores_location
reviews_per_month

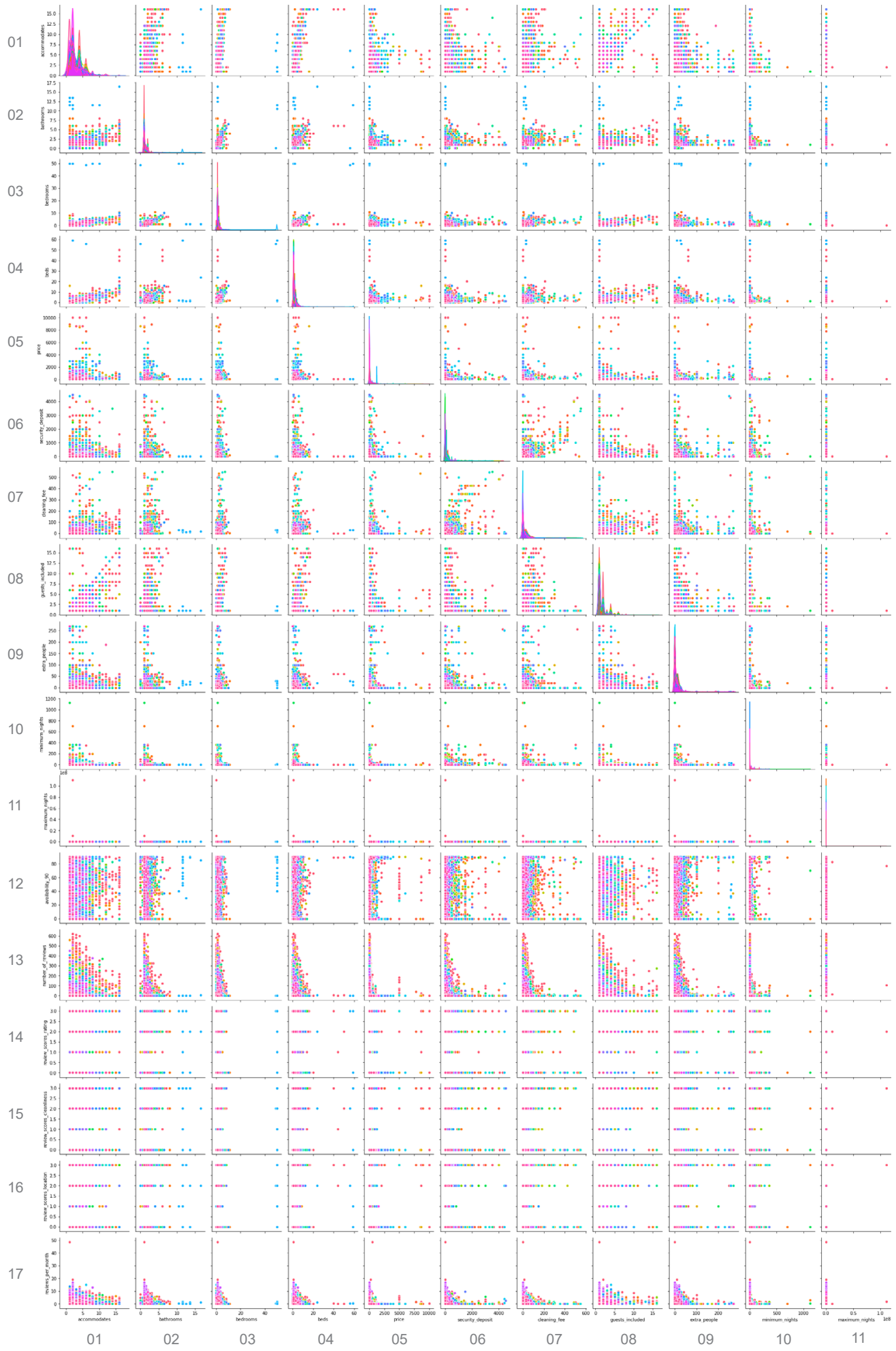


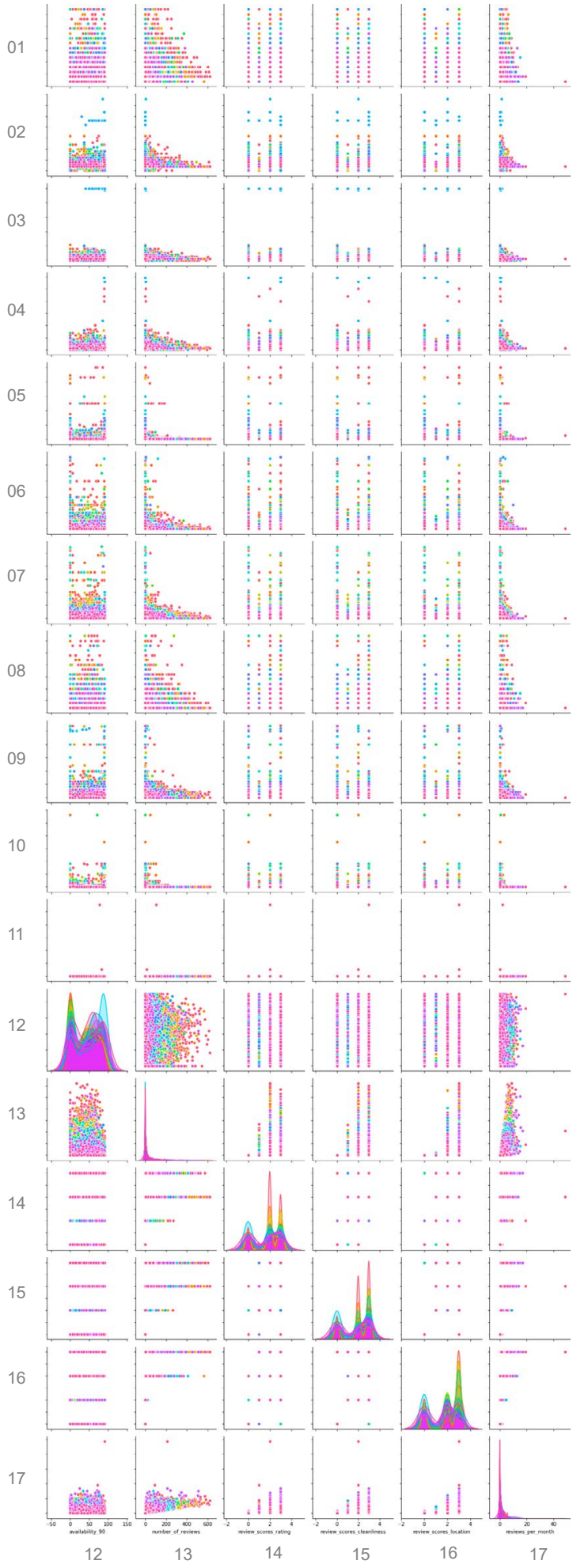
01 accomodates
02 bathrooms
03 bedrooms
04 beds
05 price
06 security_deposit
07 cleaning_fee
08 guests_included
09 extra_people
10 minimum_nights
11 maximum_nights
12 availability_90
13 number_of_reviews
14 review_scores_rating
15 review_scores_cleanliness
16 review_scores_location
17 reviews per month

Reinickendorf
Pankow
Lichtenberg
Marzahn - Hellersdorf
Treptow - Köpenick
Neukölln
Friedrichshain-Kreuzberg
08 Mitte
Tempelhof - Schöneberg
Charlottenburg-
Wilmersdorf
Steglitz - Zehlendorf

Fig. 20: Matrix of relationships between variables for Berlin.

Madrid


01
02
03
04
05
06
07
08
09
10
11
accommodates
bathrooms
bedrooms
beds
price
security_deposit
cleaning_fee
guests_included
extra_people
minimum_nights
maximum_nights
availability_90
number_of_reviews
review_scores_rating
review_scores_cleanliness
review_scores_location
reviews_per_month

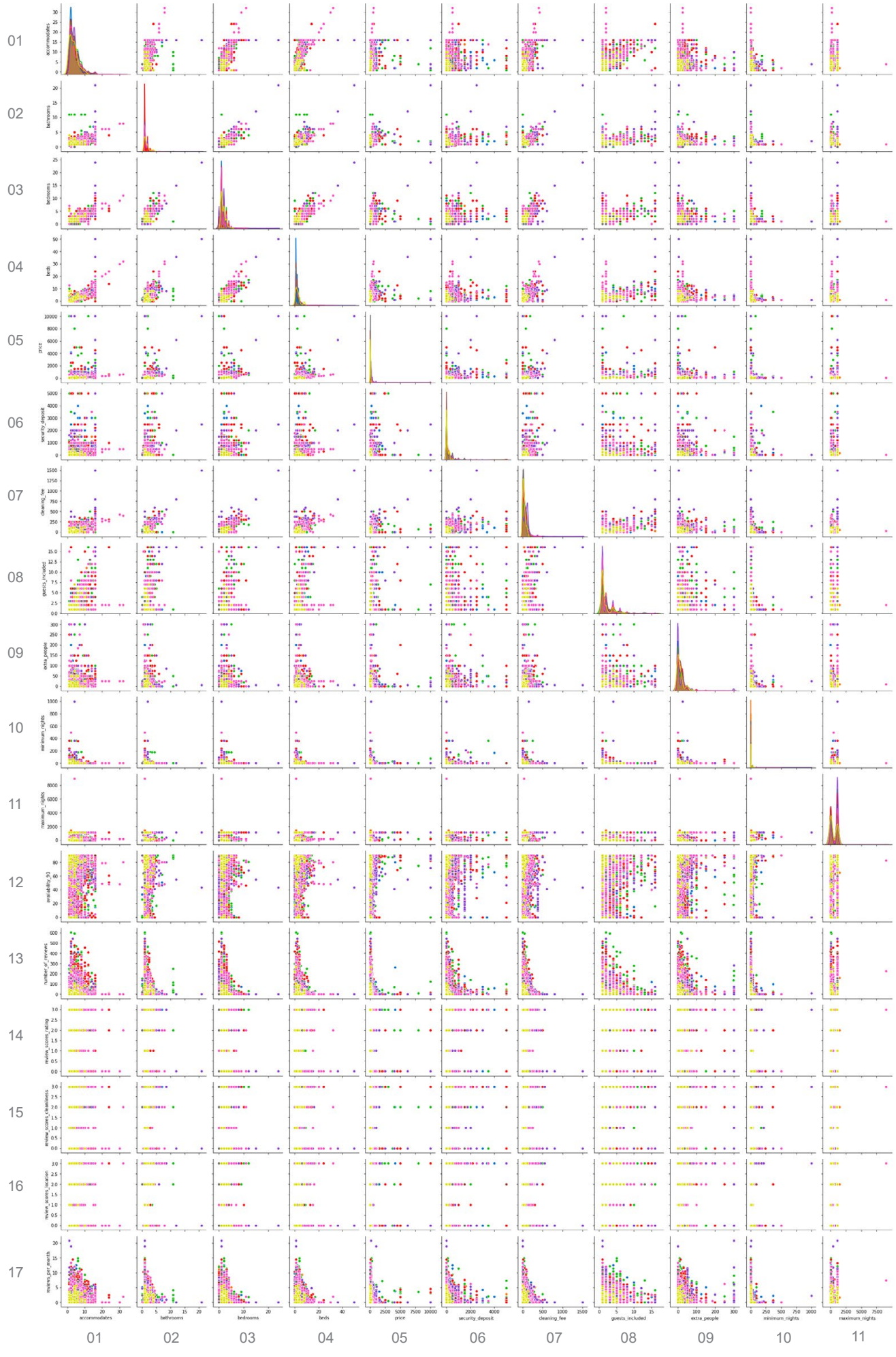


01 accomodates
02 bathrooms
03 bedrooms
04 beds
05 price
06 security_deposit
07 cleaning_fee
08 guests_included
09 extra_people
10 minimum_nights
11 maximum_nights
12 availability_90
13 number_of_reviews
14 review_scores_rating
15 review_scores_cleanliness
16 review_scores_location
17 reviews per month

Centro
Arganzuela
Retiro
Salamanca
Chamartín
Tetuán
Chamberí
Fuencarral - El Pardo
Moncloa - Aravaca
Latina
Carabanchel
Usera
Pte. Vallecas
Moratalaz
Ciudad Lineal
Hortaleza
Villaverde
Villa de Vallecas
Vicálvaro
San Blas - Canillehas
Barajas

Fig. 21: Matrix of relationships between variables for Madrid.

Chicago


01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
accommodates
bathrooms
bedrooms
beds
price
security_deposit
cleaning_fee
guests_included
extra_people
minimum_nights
maximum_nights
availability_30
number_of_reviews
review_scores_rating
review_scores_cleanliness
review_scores_location
reviews_per_month

01 accomodates
02 bathrooms
03 bedrooms
04 beds
05 price
06 security_deposit
07 cleaning_fee
08 guests_included
09 extra_people
10 minimum_nights
11 maximum_nights
12 availability_90
13 number_of_reviews
14 review_scores_rating
15 review_scores_cleanliness
16 review_scores_location
17 reviews per month

Far North Side
Northwest Side
North Side
West Side
Central
Southwest Side
South Side
Far Southwest Side
Far Southeast Side

Fig. 22: Matrix of relationships between variables for Chicago.