Analysis of Residential Flat Rent Prices in Major Polish Cities

Wojciech Krzos, 276264

June 10, 2024

Contents

Abstract

This paper presents an analysis of flat rental prices in six major cities in Poland: Warsaw, Kraków, Łódź, Wrocław, Poznań, and Gdańsk.

1 Introduction

In a dynamic market environment influenced by a conflict to the east of Poland, increased immigration, and government policies, flat rental prices have become a significant concern for the residents of major Polish cities. This paper examines rental prices in six major Polish cities most affected by the current state of affairs: Warsaw, Kraków, Łódź, Wrocław, Poznań, and Gdańsk.

1.1 Background of the Polish Flat Rental Market

Poland's rental market has undergone major changes after its post-communist transition - widely recognised to have started in 1989 **?**. Urbanisation and demographic shifts have greatly influenced the economy in the past years, starting in 2019, including the rental market **?**. The demand for rental properties has increased, especially in major cities, driven by factors i.e. job opportunities, education possibilities, and lifestyle preferences **?**. The six major cities are recognised to be the economic and cultural hubs, attracting a diverse population. That increased the population of professionals, students, and expatriates, who all contribute to the current state of the rental market. An understanding of the current market climate is crucial for real estate agencies and potential renters, to make informed decisions.

1.2 Goal of the Paper

The paper's goal is to provide a comprehensive analysis the current rental properties market in six major Polish cities. By comparing these cities, the paper aims to identify trends, disparities, and underlying factors influencing rental costs.

1.3 Research Questions

Considering the above, it has been decided to form the following research questions:

- **Research Question 1:** What are the current flat rental prices in Warsaw, Kraków, Łódź, Wrocław, Poznań, and Gdańsk?
- **Research Question 2:** How are rental properties spaced in a city's borders?
- **Research Question 3:** What factors contribute to the differences in rental prices across cities and to what extent can the rent prices be rpedicted using said factors?

By answering these questions, the paper seeks to uncover the key determinants of rental price variations and provide a detailed understanding of the rental landscape in these urban areas.

1.4 Structure of the Paper

The paper is structured as follows:

- **Section 2: Methodology** - Outlines the data collection methods and analytical techniques used to assess rental prices.
- **Section 3: Results** - The findings of the analysis, including comparative rental prices and trends across the six cities, are presented in this section.
- **Section 4: Discussion** - Interprets the results, discussing the factors influencing rental prices and the implications for different stakeholders.
- **Section 5: Conclusion** - Summary of key findings, their implications, suggestions for future research, and areas for improvement.
- **Section 6: Appendix** - Stores all figures relevant to the work that would otherwise disturb the above section's flow of reasoning.

2 Methodology

Data on rental prices were collected from OLX.PL, a child company of the Dutch OLX Group **?**. The collection process was implemented using a web-scraping Python application specifically designed to suit the needs of this study.

2.1 Tools

The following tools were utilized in this study:

- **Python**: Used for data collection, analysis, and visualization. Libraries such as pandas, geopandas, matplotlib, seaborn, contextily, and folium were employed.
- **Geopandas**: Used for geospatial analysis and visualization of the rental data.
- **Matplotlib Seaborn**: Used for plotting graphs and heatmaps to visualize the distribution of rental prices.
- **Folium**: Used for creating interactive heatmaps.
- **Scikit-learn**: Used for regression analysis to estimate rental prices based on available parameters.
- **LaTeX**: Used for documenting the methodology and findings, with Overleaf as the platform for compilation.
- **Zotero**: Used for documenting all the references used in the paper.
- **Chat GPT API by OpenAI**: Used for the retrieval of addresses from the descriptions of the offers. The GPT 3.5 turbo model was used due to its reliance and cost-efficiency.
- **Google Maps API**: Used for the retrieval of geolocation of the offers.

2.2 Methodology

The methodology involved several key steps:

1. **Data Collection**: Rental price data was collected from olx.pl.
2. **Data Preparation**: The collected data was cleaned and preprocessed. This involved handling missing values, encoding categorical variables, ensuring consistency in data formats, determining addresses, and correcting the addresses.
3. **Geospatial Analysis**: Geospatial analysis was conducted using geopandas and contextily to visualize the spatial distribution of rental prices in different cities. Heatmaps and distribution plots were generated for each city.
4. **Regression Analysis**: Regression models were developed using scikit-learn to estimate rental prices based on features such as floor, furniture, area, rooms, and building type. Model performance was evaluated using R-squared and Mean Absolute Error (MAE).
5. **Visualization and Documentation**: The results were visualized using matplotlib, seaborn, and folium. Figures were included in the LaTeX document to provide a clear representation of the findings.

2.3 Data Collection and Preparation

The data coveres six major Polish cities: Gdańsk, Kraków, Łódź, Poznań, Warszawa, and Wrocław. Those cities were chosen as a representative set of polish cities with population above 400,000 residents. Each dataset included attributes such as the price, floor level, furniture status, area, number of rooms, building type, and geographic coordinates (latitude and longitude).

- **Accessing olx.pl API**: The OLX.pl API was accessed to retrieve real-time data on rental listings.
- **Downloading .json files**: The data was downloaded in JSON format, which included detailed information about each rental listing.
- **Parsing JSON data**: The JSON files were parsed to extract relevant fields such as price, address, floor level, furniture status, area, number of rooms, building type, and geographic coordinates.
- **Extracting addresses**: The address of each property was extracted using Chat GPT 3.5. A prompt was sent for each description that returned the address.
- **Extracting locations**: The longitude and latitude were extracted using Google Maps API.
- **Storing data**: The extracted data was stored in CSV files for ease of analysis and processing. Additionally, a SQLite database has been introduced for long-term storage.
- **Data verification**: The data was verified for accuracy and completeness, ensuring that all necessary fields were present and correctly formatted.

- **Handling missing values**: Any missing or incomplete data entries were identified and appropriately handled, either through imputation or exclusion - the choice was case-specific.
- **Handling outliers**: Any outliers, whether due to their location being too far from the city or their prices being too low or too high, were manually removed using geospatial visualizations in section 7.

2.4 Constraints

Several constraints were encountered during the study:

- **Data Availability**: Not all offers provided complete data, leading to potential gaps in the dataset.
- **Computational limits**: Data has been collected using a personal computer and a virtual private server. Due to machine's hardware limitations, less data could have been collected in a given time-frame.
- **API access limit**: Due to limitations set by OLX.pl and potential rate limits, data collection was conducted in a phased manner to avoid overloading the API and ensure compliance with usage policies.
- **Geographic Coverage**: The data was limited to six major cities, which may not fully represent rental trends in smaller towns or rural areas.
- **Geographic Coverage within a City**: The data relating to a city has been cleaned off any offers from the suburbs that do not belong to the city itself.
- **Temporal Consistency**: The data was collected at the span of 2 weeks, which may not account for seasonal variations in rental prices.
- **Feature Limitations**: Certain potentially influential features (e.g., proximity to amenities, public transport) were not available in the dataset, which may impact the accuracy of the regression models.

3 Results

The results show a significant variation in rental prices across the six cities. Warsaw has the highest average rental prices, while Lodz has the lowest.

3.1 City-specific Analysis

It was decided to use 2 regression models for the prices: linear and random forest. Initially, only the linear model was used; however, it was noticed that the flat prices seem to reach a plateau in each city, after reaching a certain price. So, linear models estimate the prices closer to the median well. However, they are insufficient considering the whole data set. So, an additional, third, graph was added with data logarithmically normalised.

Contrary to the above, it could be said, that values on the plateau should not be included and should be considered outliers. However, this paper argues, that they must

be included. The issue lies in the insufficient data collected, not in those offers being outliers. Although less common, those offers tend to disappear from the market faster. This, at first, may seem unintuitive; however, it becomes clearer when one considers the scarce supply of such properties and the increasing number of middle-to-high-class residents in major Polish cities **?**.

3.1.1 Gdańsk

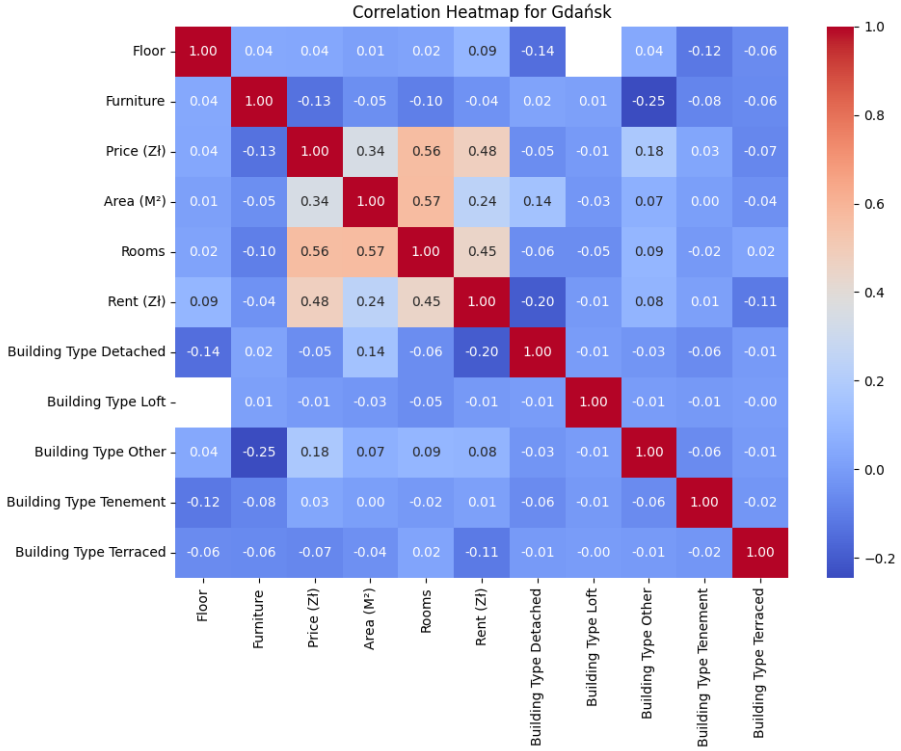

Figure 1: Correlation Heatmap for Gdańsk

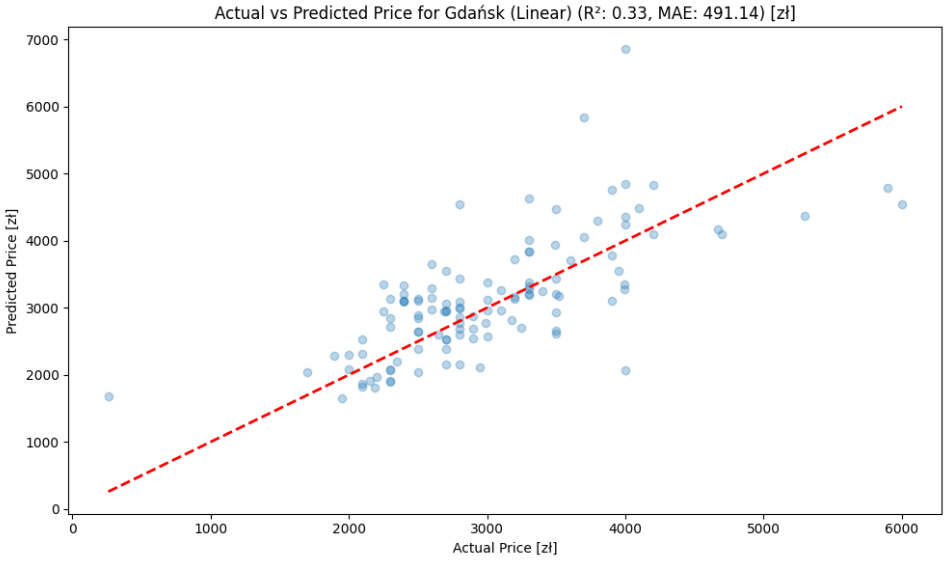

Figure 2: Price Prediction (Linear) for Gdańsk

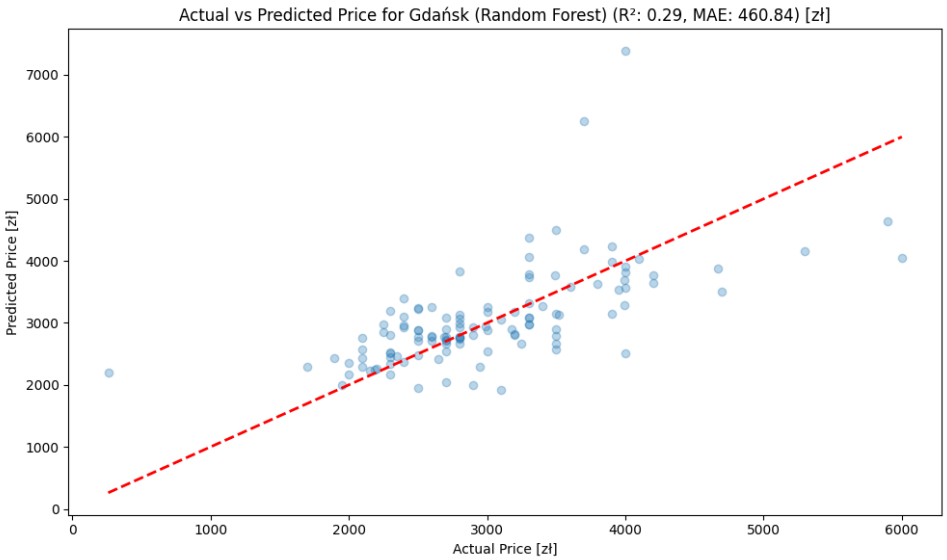

Figure 3: Price Prediction (Random Forest) for Gdańsk

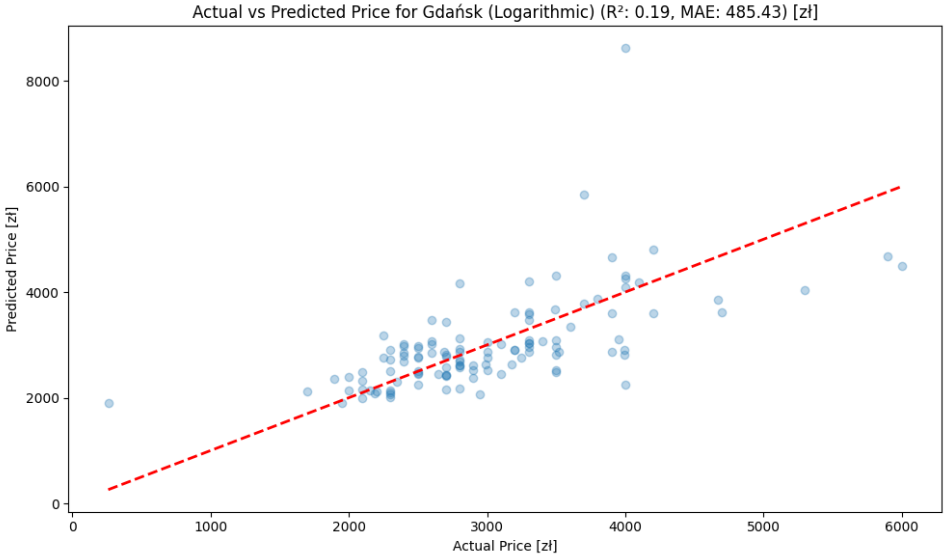

Figure 4: Price Prediction (Logarithmic) for Gdańsk

3.1.2 Kraków

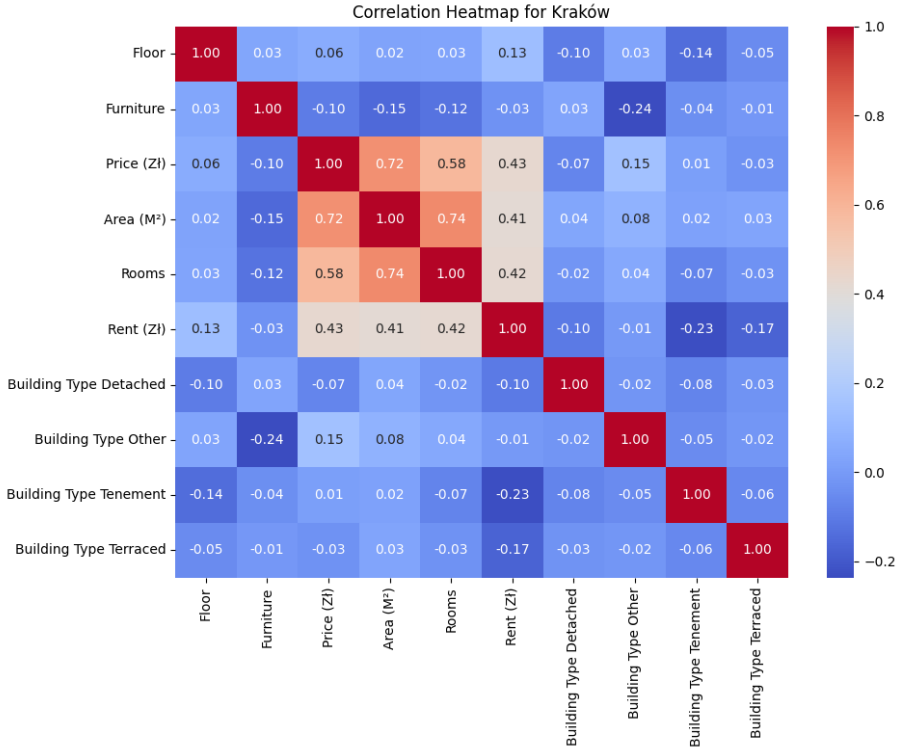

Figure 5: Correlation Heatmap for Kraków

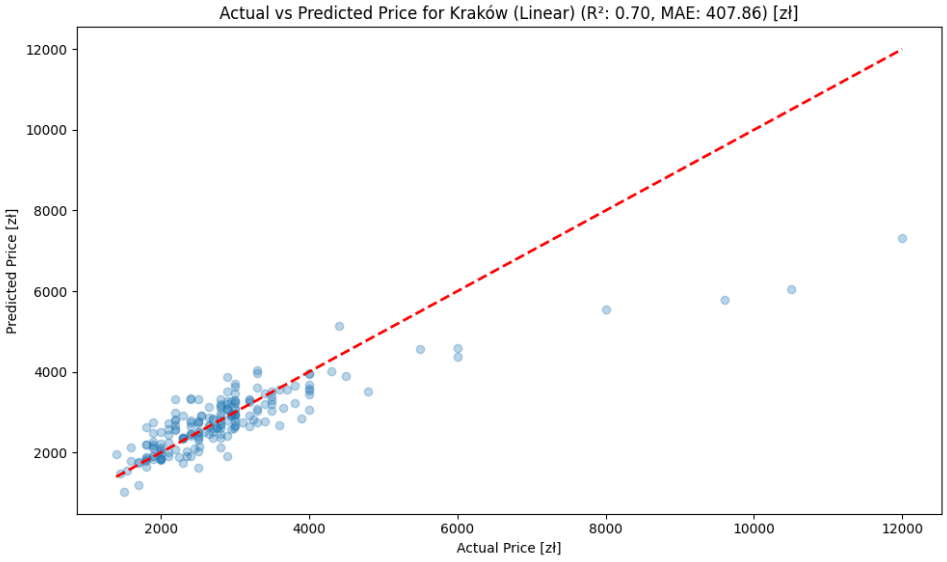

Figure 6: Price Prediction (Linear) for Kraków

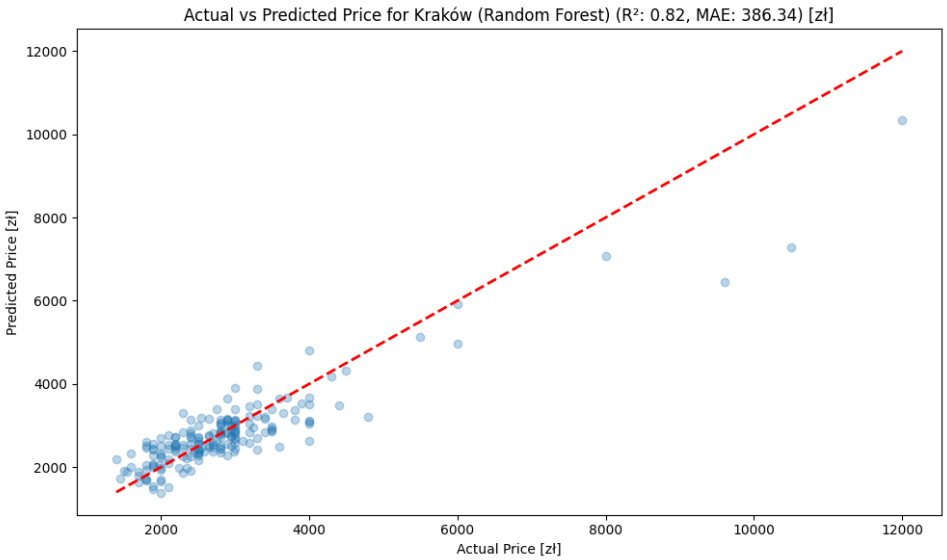

Figure 7: Price Prediction (Random Forest) for Kraków

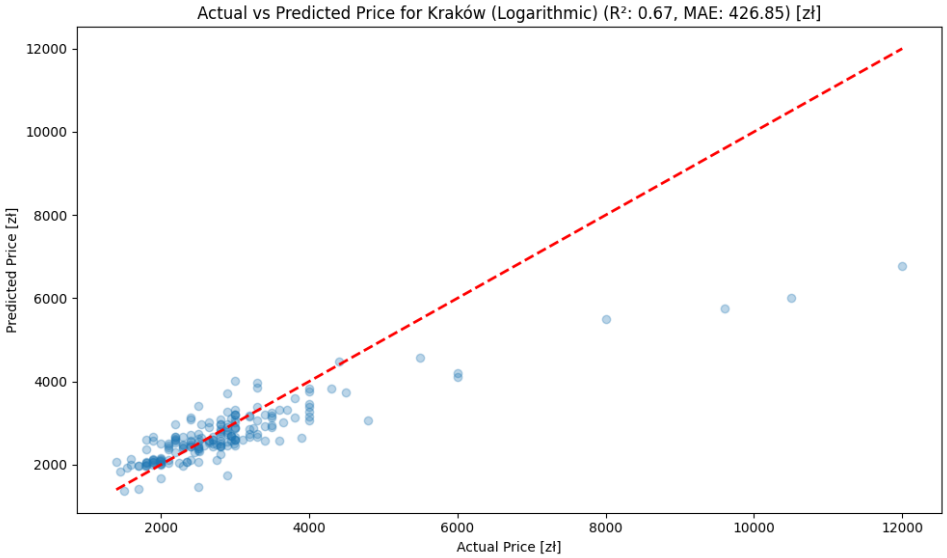

Figure 8: Price Prediction (Logarithmic) for Kraków

3.1.3 Łódź

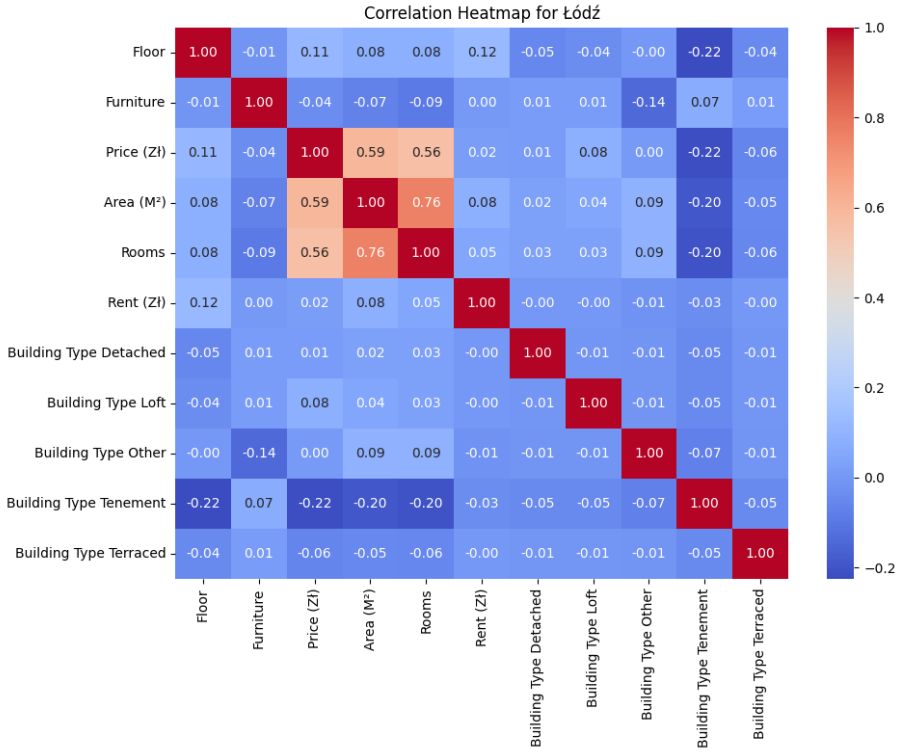

Figure 9: Correlation Heatmap for Łódź

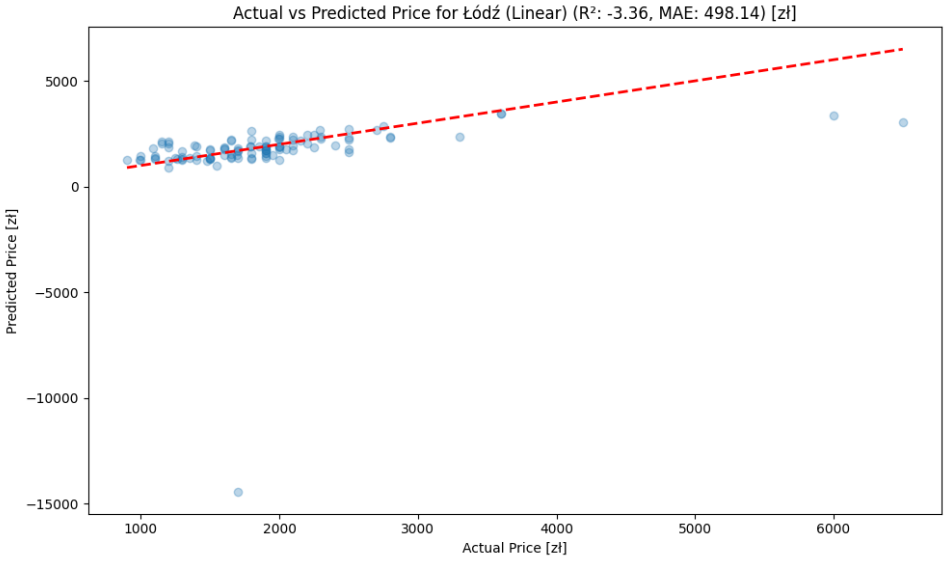

Figure 10: Price Prediction (Linear) for Łódź

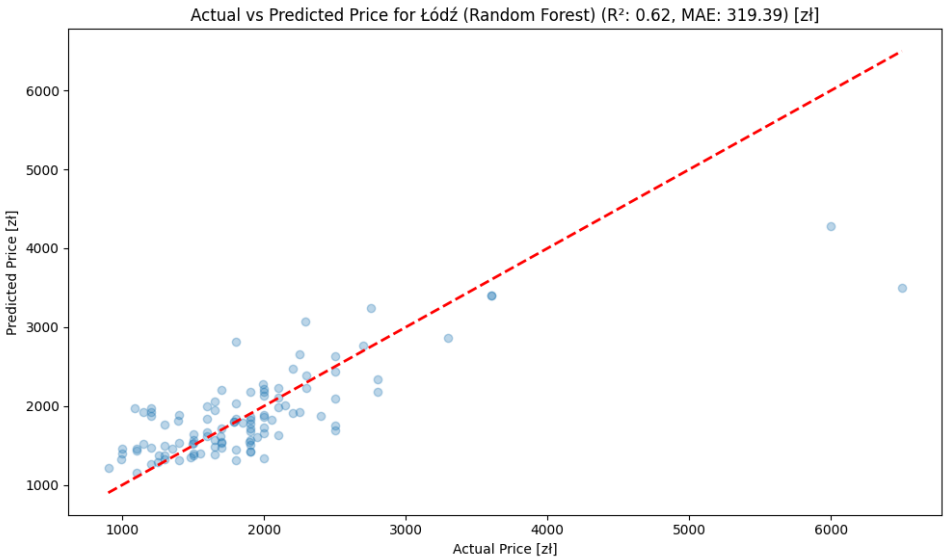

Figure 11: Price Prediction (Random Forest) for Łódź

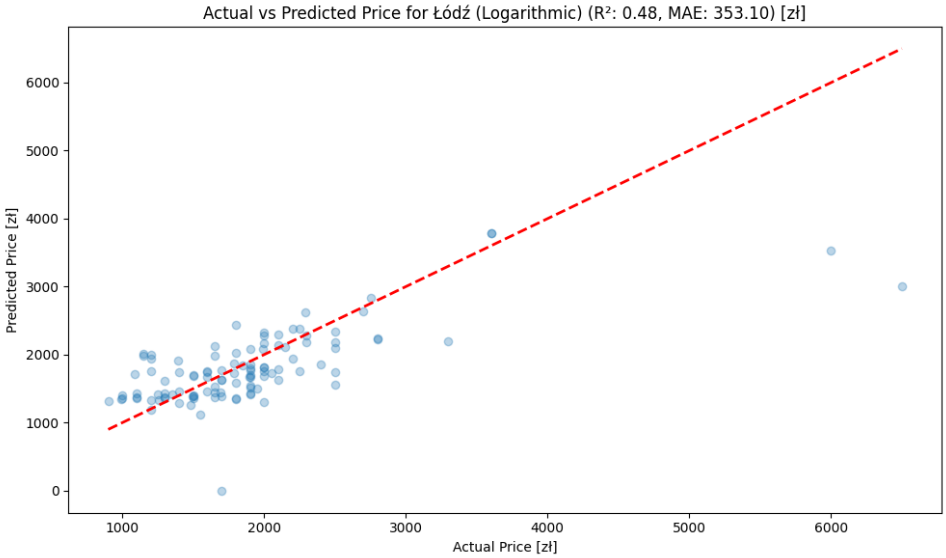

Figure 12: Price Prediction (Logarithmic) for Łódź

3.1.4 Poznań

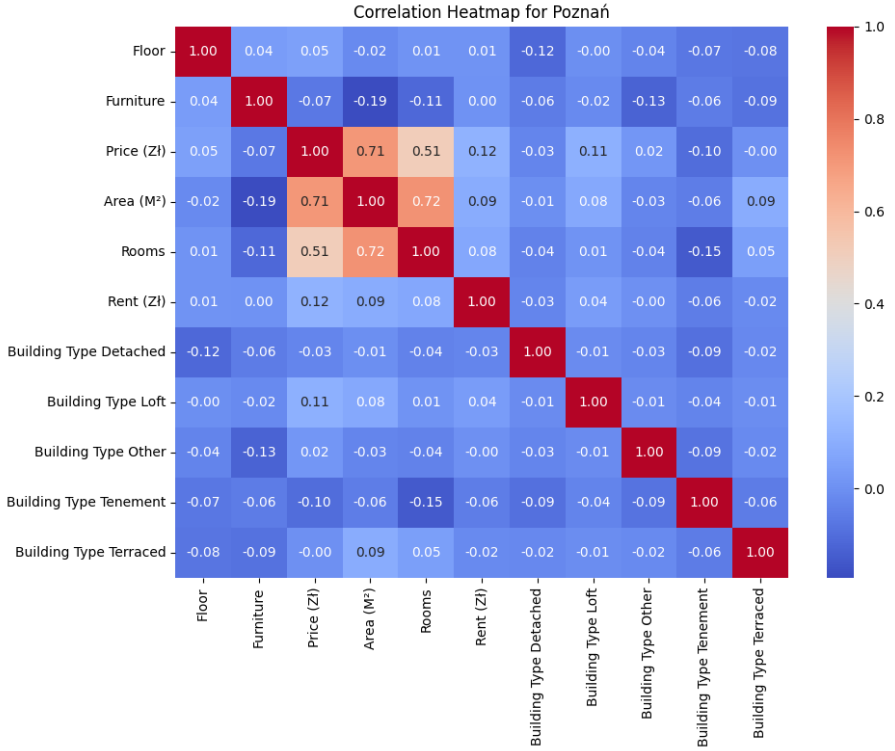

Figure 13: Correlation Heatmap for Poznań

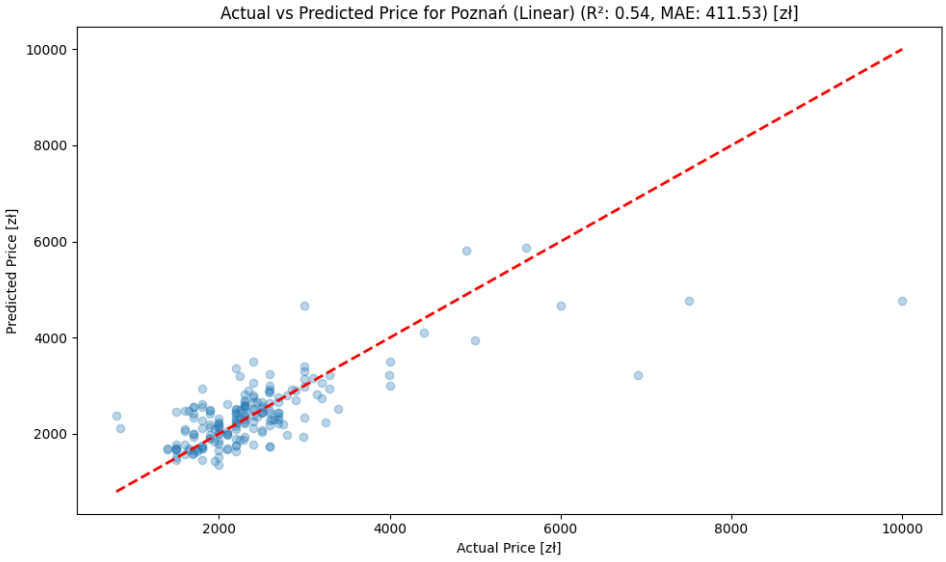

Figure 14: Price Prediction (Linear) for Poznań

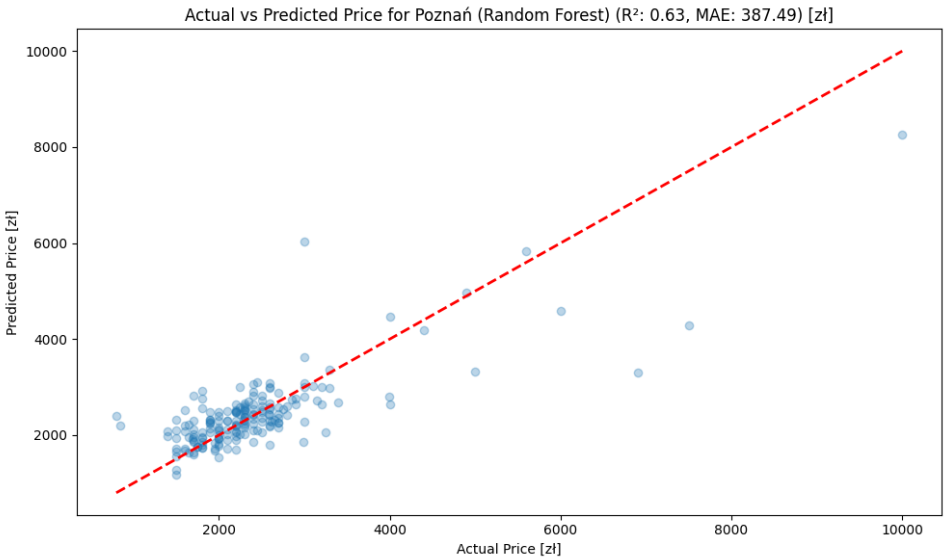

Figure 15: Price Prediction (Random Forest) for Poznań

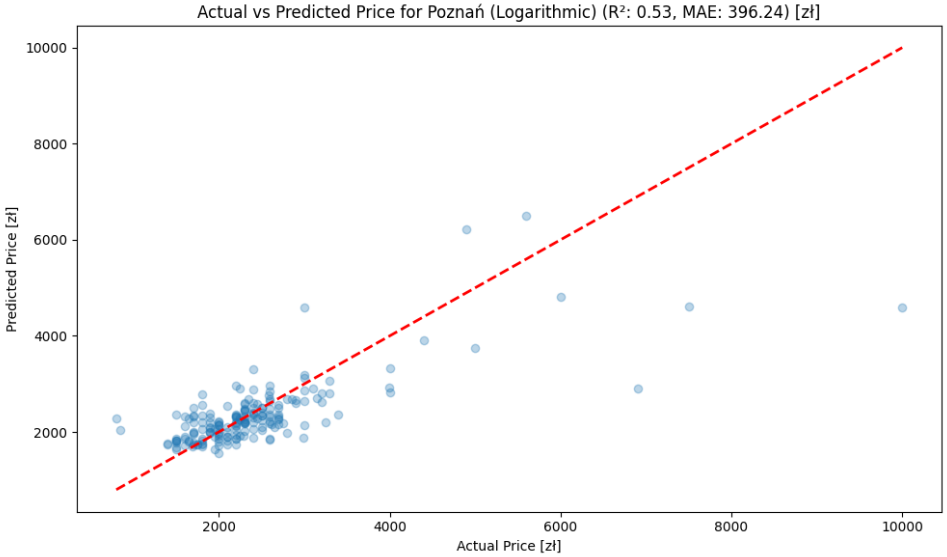

Figure 16: Price Prediction (Logarithmic) for Poznań

3.1.5 Warsaw

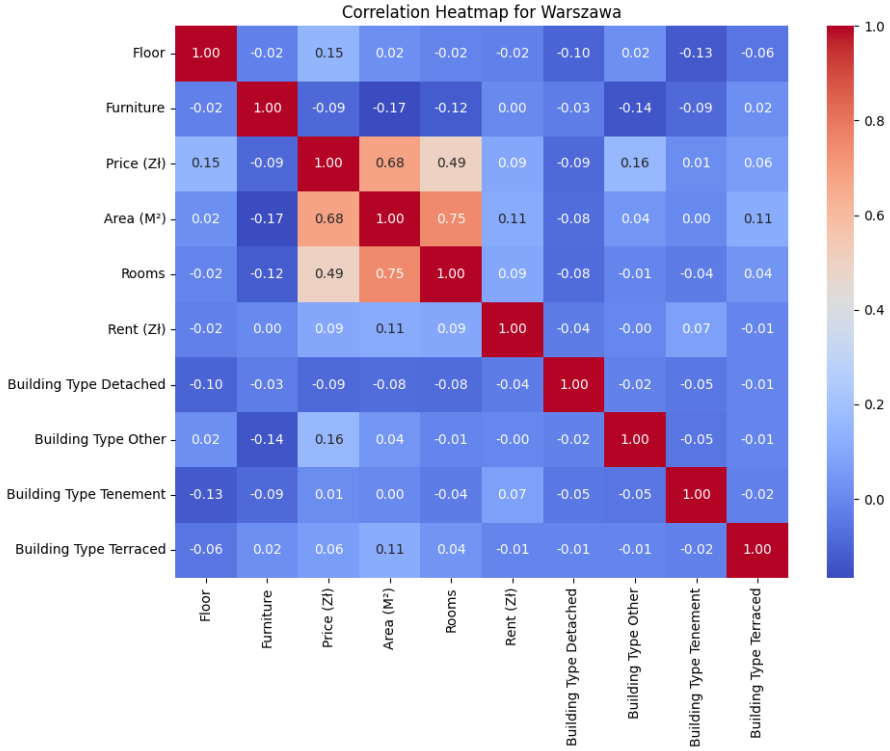

Figure 17: Correlation Heatmap for Warsaw

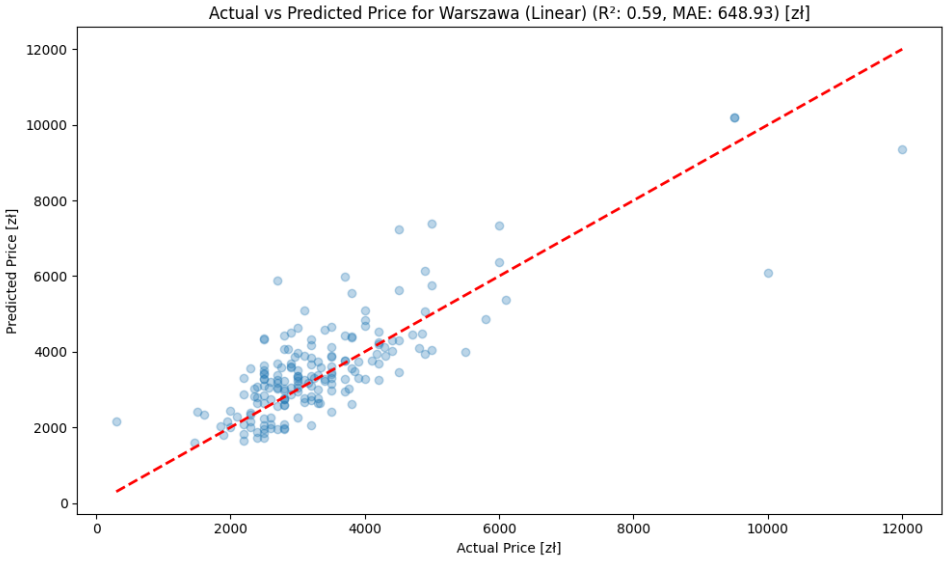

Figure 18: Price Prediction (Linear) for Warsaw

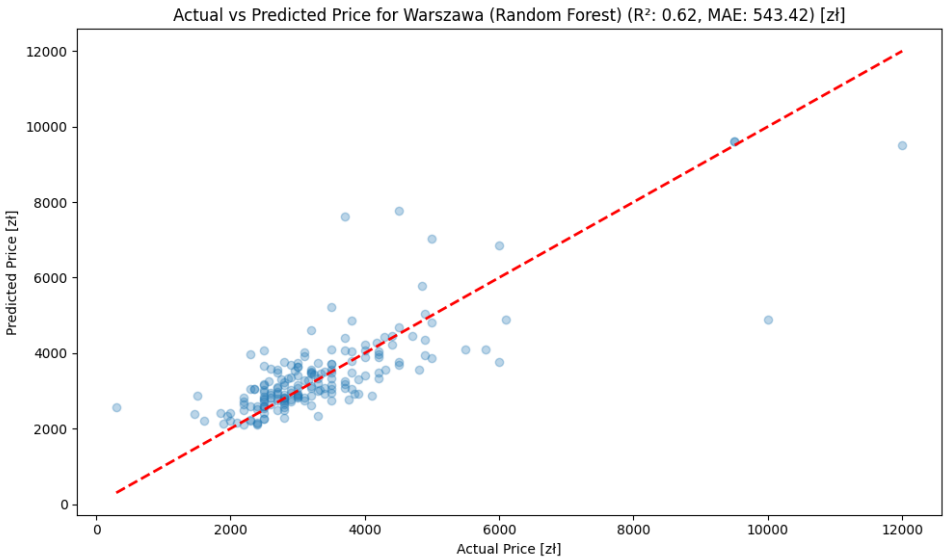

Figure 19: Price Prediction (Random Forest) for Warsaw

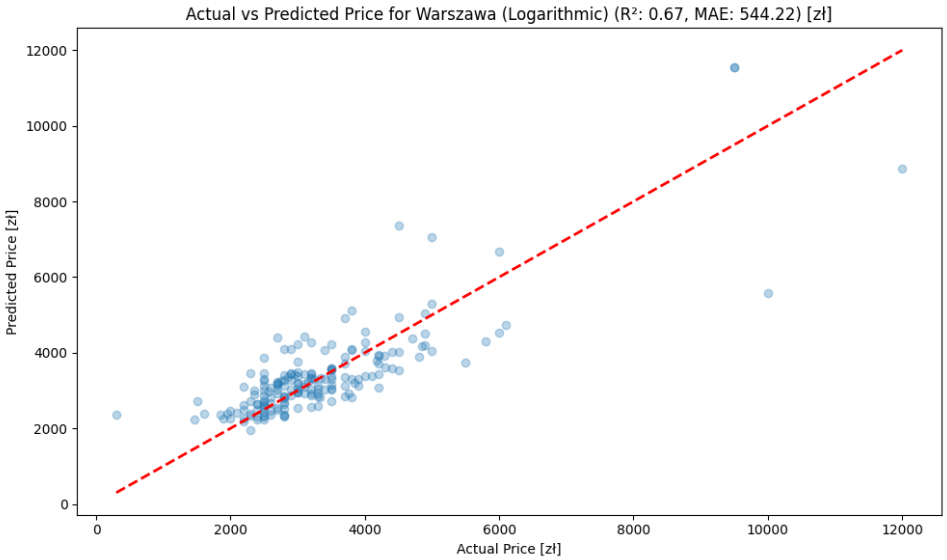

Figure 20: Price Prediction (Logarithmic) for Warsaw

3.1.6 Wrocław

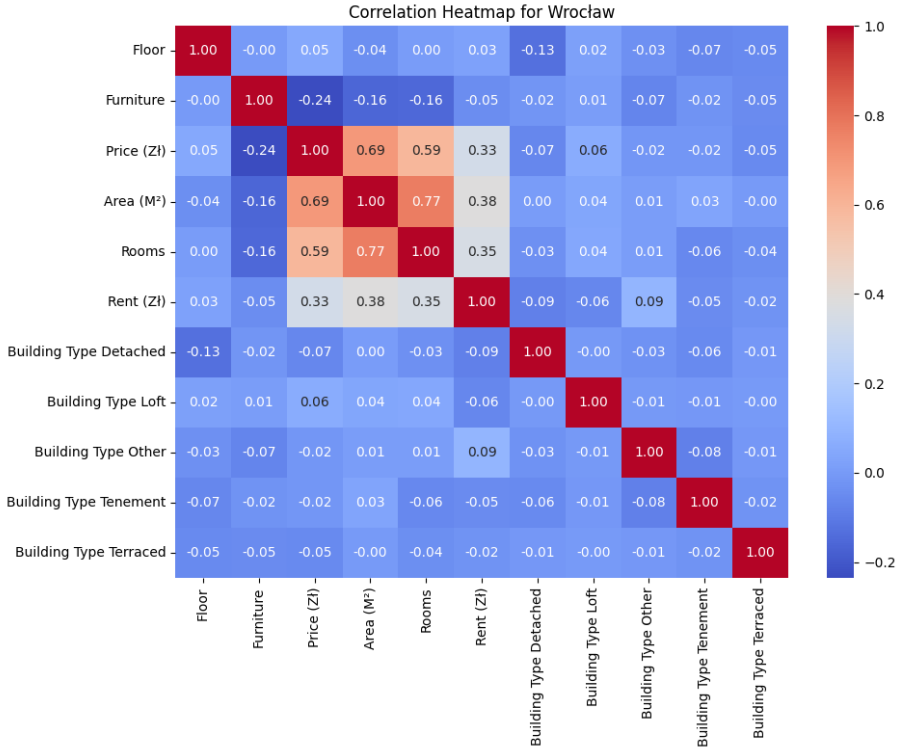

Figure 21: Correlation Heatmap for Wrocław

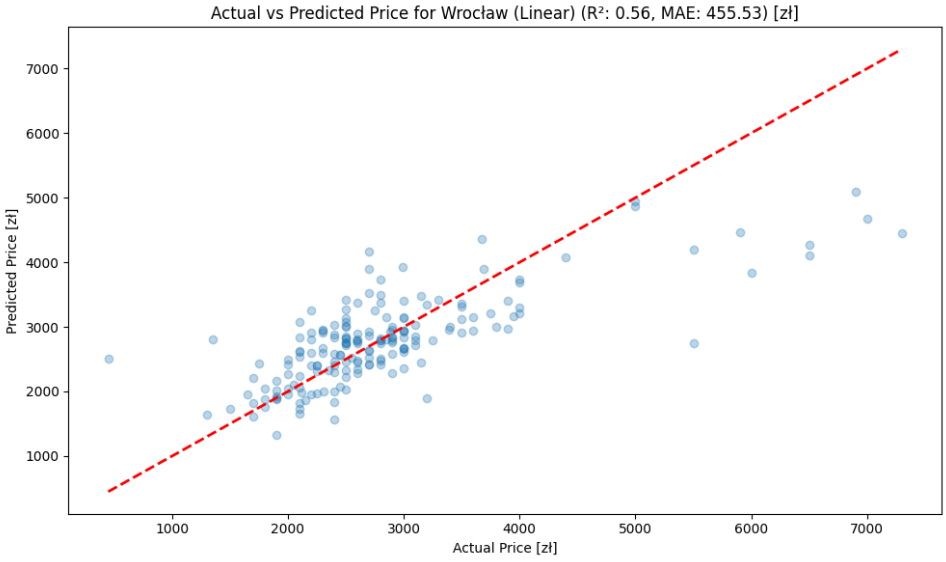

Figure 22: Price Prediction (Linear) for Wrocław

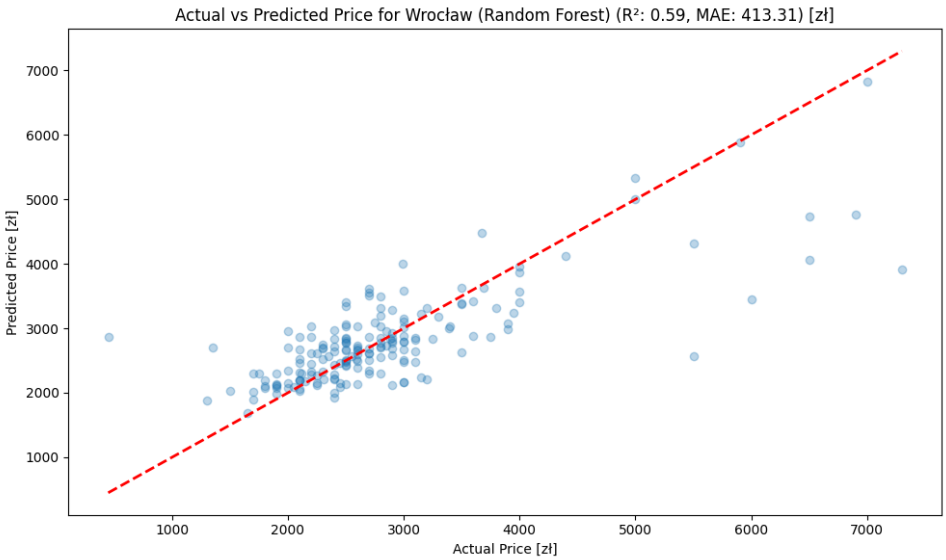

Figure 23: Price Prediction (Random Forest) for Wrocław

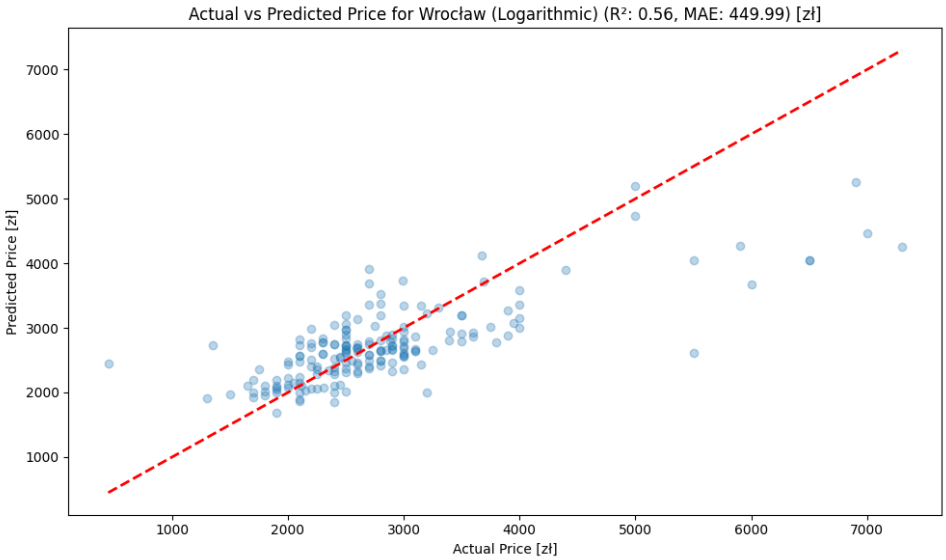

Figure 24: Price Prediction (Logarithmic) for Wrocław

3.2 City-joined Analysis

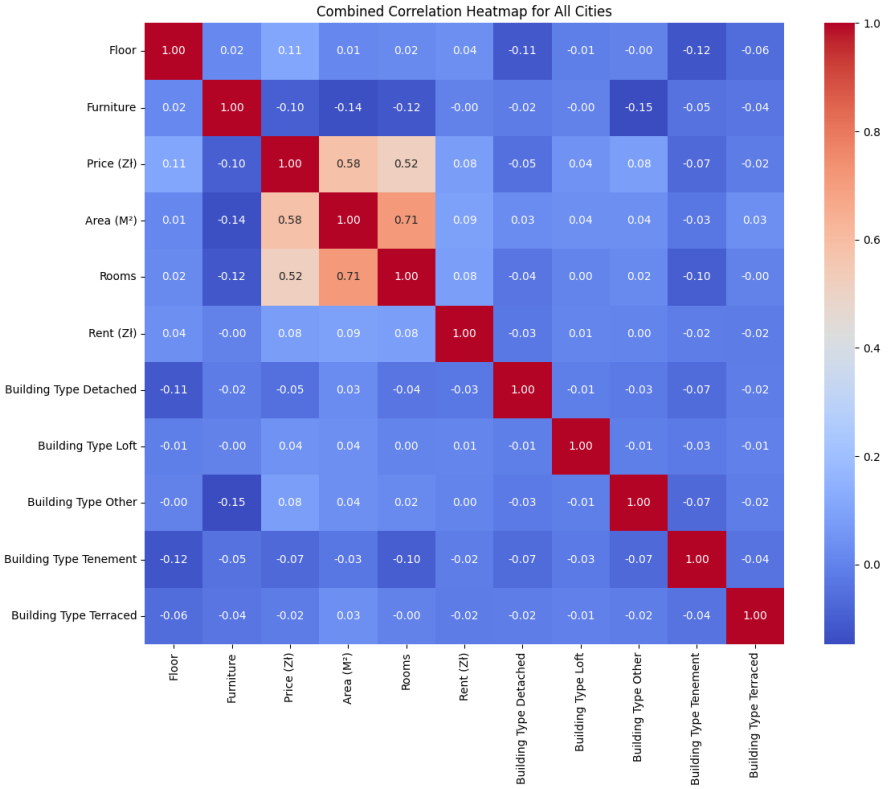

Figure 25: Combined Correlation Heatmap for All Cities

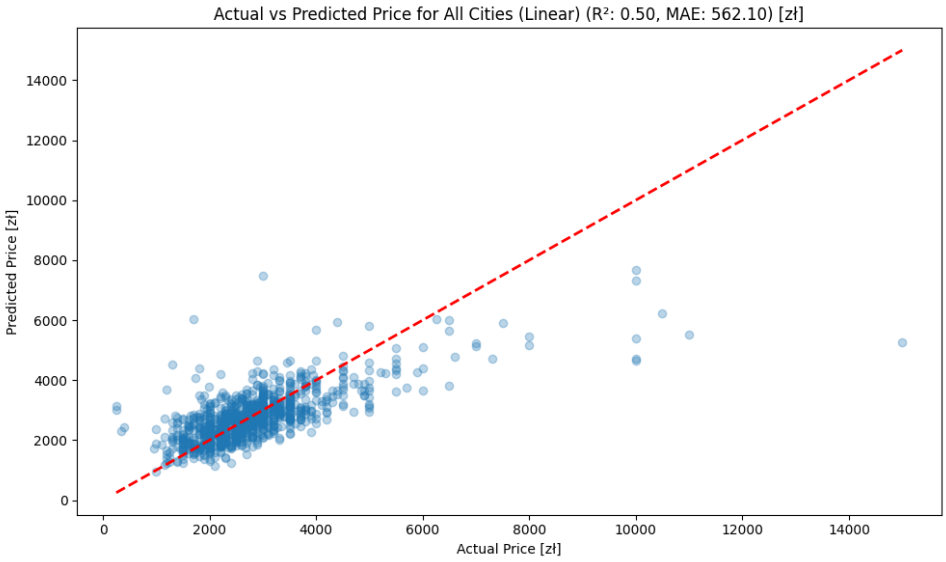

Figure 26: Price Prediction (Linear) for All Cities

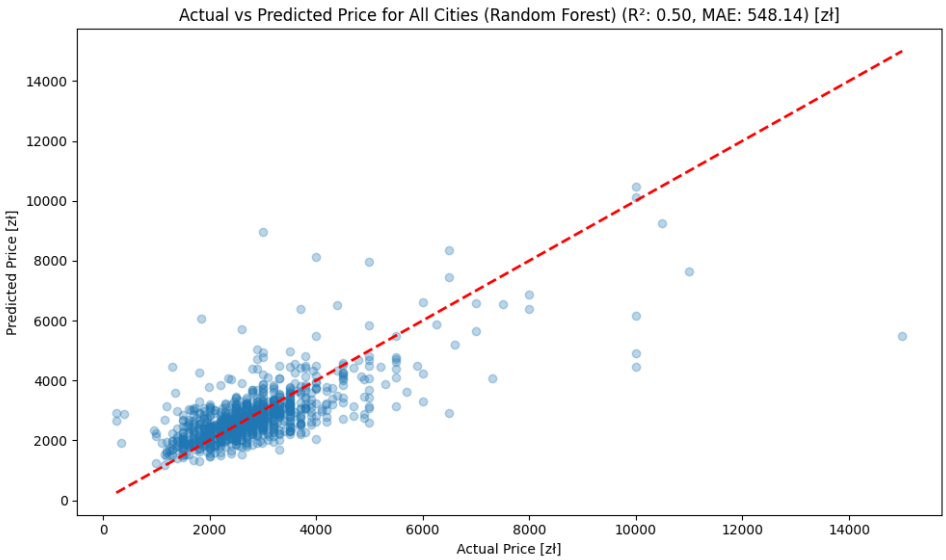

Figure 27: Price Prediction (Random Forest) for All Cities

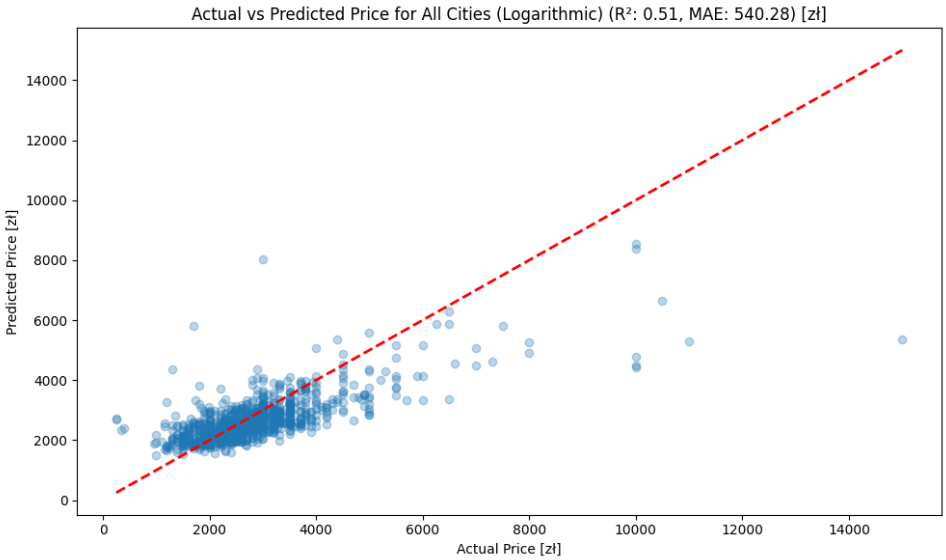

Figure 28: Price Prediction (Logarithmic) for All Cities

4 Discussion

4.1 Figures for the Joined Data

The combined data analysis for all cities provides a broader understanding of the factors affecting rental prices across Poland. The correlation heatmap and the regression analysis (linear, random forest, and logarithmic models) shed light on the relationships between various features and rental prices.

4.1.1 Correlation Heatmap Analysis

The combined correlation heatmap for all cities reveals several key relationships:

- **Area (m²)** shows a strong positive correlation with **Price (zł)** (0.58), indicating that larger apartments tend to have higher prices.
- **Rooms** also correlate positively with **Price (zł)** (0.52), suggesting that apartments with more rooms are more expensive. Also, apartments with bigger **Area (m²)** tend to have more **Rooms**.
- Interestingly, **Rent (zł)** has a very weak correlation with **Price (zł)** (0.08), implying that rent is not a strong predictor of the apartment's sale price in this dataset.
- Building types show varying degrees of correlation with **Price (zł)**, with **Building Type Loft** having a slight positive correlation (0.04) and **Building Type Tenement** showing a small negative correlation (-0.07).

These correlations suggest that while the size and number of rooms are significant factors in determining the price, other factors such as building type also play a role, albeit to a lesser extent.

4.1.2 Regression Analysis

The combined regression analysis for all cities highlights the influence of the factors on apartment prices. The logarithmic regression model, in particular, demonstrates the non-linear nature of price increases, effectively capturing the plateau effect at higher price points. These insights can inform future studies and policy decisions regarding housing and urban development in major Polish cities.

Linear Regression Model The linear regression model demonstrates a moderate fit with an R^2 value of 0.50 and a mean absolute error (MAE) of 562.10 zł. The predicted prices generally follow the trend of actual prices, but the model struggles with higher-priced properties, often underestimating them. This underestimation at higher price points suggests that the linear model may not fully capture the complexities of the data, particularly for more expensive apartments.

Random Forest Regression Model The random forest model shows a similar fit to the linear model with an R^2 value of 0.50 and an MAE of 548.14 zł. The random forest model captures the overall trend more robustly, but like the linear model, it also struggles with the highest-priced properties. The slight improvement in MAE suggests that the random forest model handles the non-linear aspects of the data slightly better than the linear model.

Logarithmic Regression Model The logarithmic regression model provides the best fit among the three, with an R^2 value of 0.51 and an MAE of 540.28 zł. This model better captures the plateau effect observed in the data, where prices increase rapidly up to a certain point and then level off. The logarithmic model's performance indicates that it is more suited to capturing the diminishing returns effect in apartment prices as they reach higher values.

Overall, the regression analysis suggests that while linear and random forest models offer reasonable approximations, the logarithmic model provides a more accurate representation of the data's underlying patterns, particularly for higher-priced apartments.

4.2 Geospatial Analysis

Note: All maps related to this analysis are located in section 7. The analysis indicates that Gdańsk has three strong peaks in location popularity that are close to each other. Warsaw exhibits multiple peaks, one in the city centre and others in tower block housing estates and the suburbs, showing widespread popularity.

In Wrocław, flat locations are more spread out from the centre, especially in Krzyki, Fabryczna, and Psie Pole. High popularity is observed near the city centres and main roads.

A similar pattern can be observed in Gdańsk, Poznań, Łódź and Kraków.

It can be observed, that all the cities share one feature of having a central point with the most flats for rent. It is observed to be near the city-center.

5 Conclusion

This study provides an overview of the rental market in major Polish cities, highlighting significant differences and underlying factors influencing rental prices.

5.1 Summary of the findings

The analysis reveals notable variations in rental prices across the six major Polish cities. Warsaw exhibits the highest average rental prices, likely due to its status as the capital and its economic opportunities. Krakow and Gdansk also demonstrate higher rental prices, reflecting their cultural and economic significance.

Key findings include:

- **Correlation Analysis**: The size of the apartment (area in $\mathbf{m^2}$) and the number of rooms are strong predictors of rental prices.
- **Regression Models**: The logarithmic regression model outperformed the linear and random forest models, better capturing the non-linear relationship between apartment features and prices, especially at higher price points.
- **Geospatial Analysis**: Rental popularity peaks vary by city, with central areas generally showing higher popularity. In Warsaw, multiple peaks are observed, including in the city centre and suburban tower blocks. In Wrocław, popular rental locations are more dispersed.

5.2 Evaluation of the paper

While the study successfully identifies key factors influencing rental prices and demonstrates the effectiveness of different regression models, several limitations constrain the findings. Those limitations were noted earlier in the Methodology section. For ease of reading, they were given once more with proposed solutions:

- **Data Availability**: Incomplete data from some offers may lead to gaps in the dataset, potentially affecting the robustness of the analysis. *To mitigate this, future studies could collaborate with rental platforms to ensure more comprehensive data collection. Implementing methods to handle missing data, such as data imputation techniques, could also enhance dataset completeness.*

- **Computational Limits**: Data collection was restricted by hardware limitations, which may have reduced the volume of data collected within the given timeframe. *Utilizing cloud-based computing resources or high-performance computing systems can overcome these limitations, enabling the processing of larger datasets more efficiently.*

- **API Access Limit**: The phased data collection approach was necessary due to API rate limits, which could have affected the continuity and completeness of the data. *Engaging with data providers to negotiate higher API rate limits or utilizing data scraping techniques with appropriate ethical considerations can help gather more continuous data streams.*

- **Geographic Coverage**: The study focused on six major cities, which may not fully represent rental trends in smaller towns or rural areas. Additionally, data from city suburbs was excluded to maintain focus on the urban areas. *Expanding the geographic scope to include smaller towns and rural areas, as well as suburban regions, would provide a more comprehensive understanding of rental trends across diverse locations.*

- **Temporal Consistency**: Data was collected over two weeks, not accounting for potential seasonal variations in rental prices. *Extending the data collection period to cover different seasons and ensuring temporal consistency would help capture seasonal variations in rental prices, providing a more accurate analysis.*

- **Feature Limitations**: Important features such as proximity to amenities and public transport were not included, which may impact the accuracy and comprehensiveness of the regression models. *Incorporating additional features, such as proximity to public amenities, transport links, and neighborhood quality, can enhance the predictive power and accuracy of the regression models.*

5.3 Future extensions

Future research could address the current study's limitations by:

- **Expanding Data Collection**: Including more cities and extending the geographic scope to cover suburban and rural areas would provide a more comprehensive understanding of rental trends across Poland.

- **Improving Data Completeness**: Ensuring more complete data from rental offers, possibly through collaboration with data providers, would enhance the dataset's robustness.
- **Accounting for Temporal Variations**: Conducting data collection over a longer period, including different seasons, would help capture seasonal variations in rental prices.
- **Incorporating Additional Features**: Adding data on proximity to amenities, public transport, and other relevant factors could improve the accuracy of the regression models.
- **Enhancing Computational Resources**: Utilizing more powerful computational resources could enable larger-scale data collection and more complex analyses.

These improvements would provide a more detailed and accurate picture of the rental market in Poland, aiding stakeholders in making informed decisions.

6 References

7 Appendix

7.1 Geospatial data

7.1.1 Gdańsk

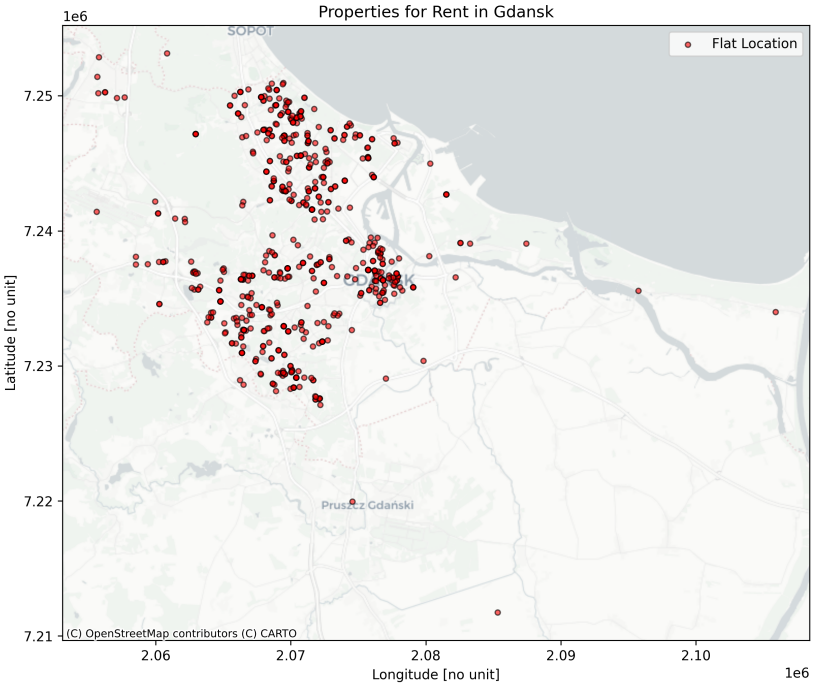

Figure 29: Spatial distribution of flats in Gdańsk with basemap.

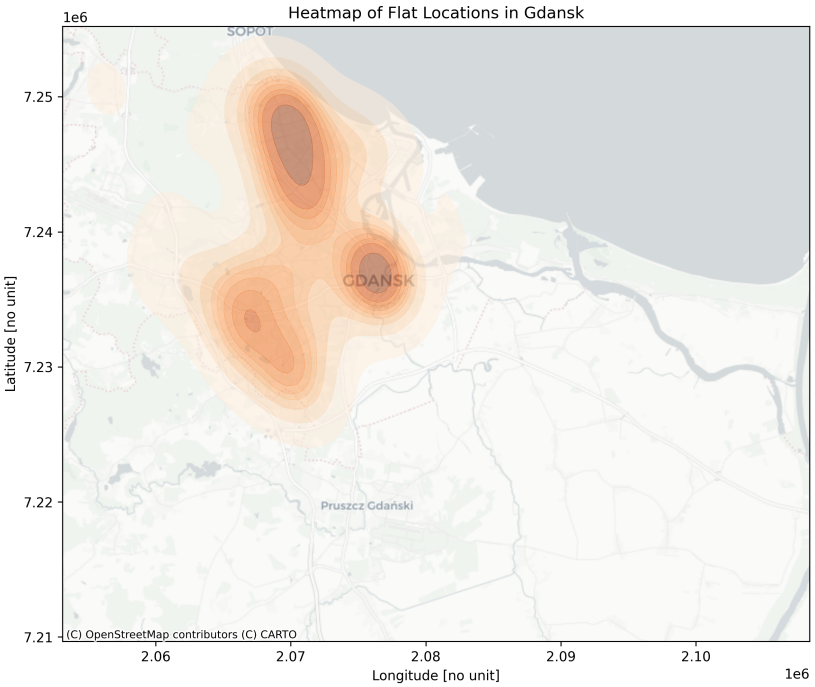

Figure 30: Heatmap of flat locations in Gdańsk.

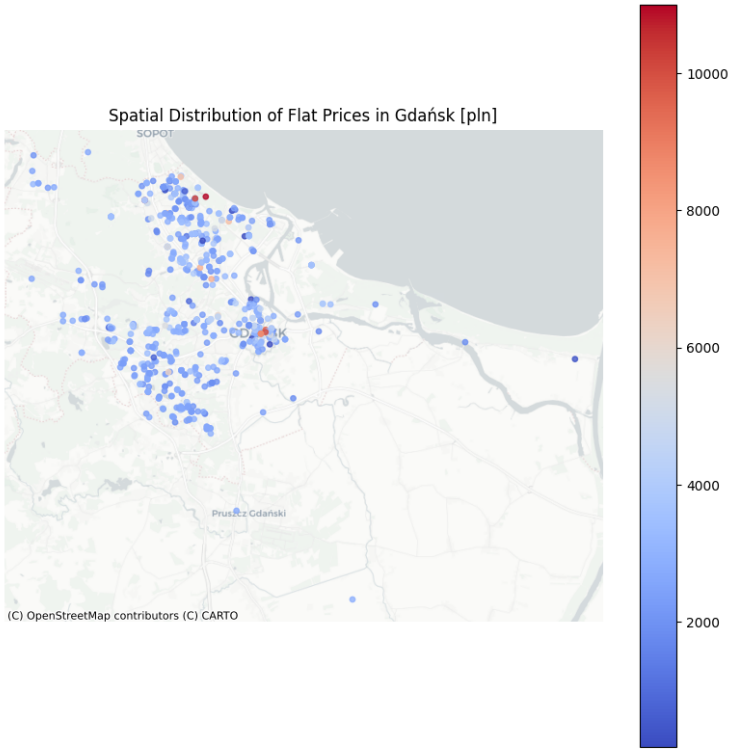

Figure 31: Price distribution of flats in Gdańsk.

7.1.2 Kraków

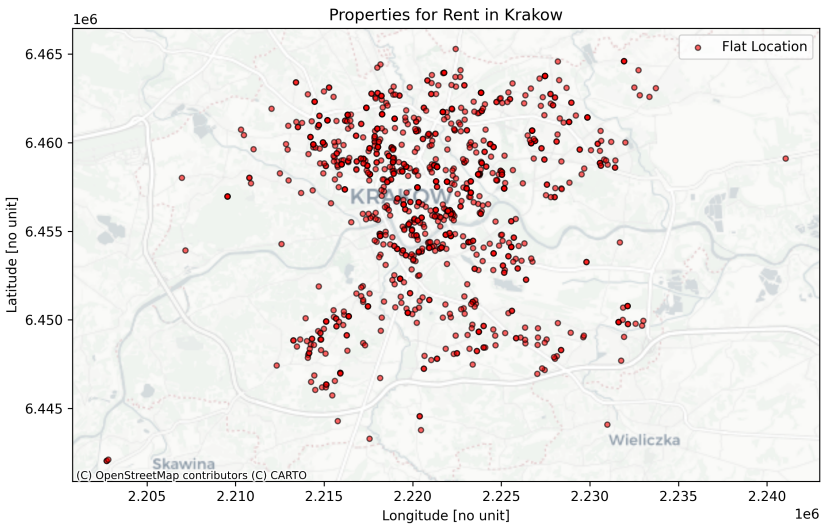

Figure 32: Spatial distribution of flats in Kraków with basemap.

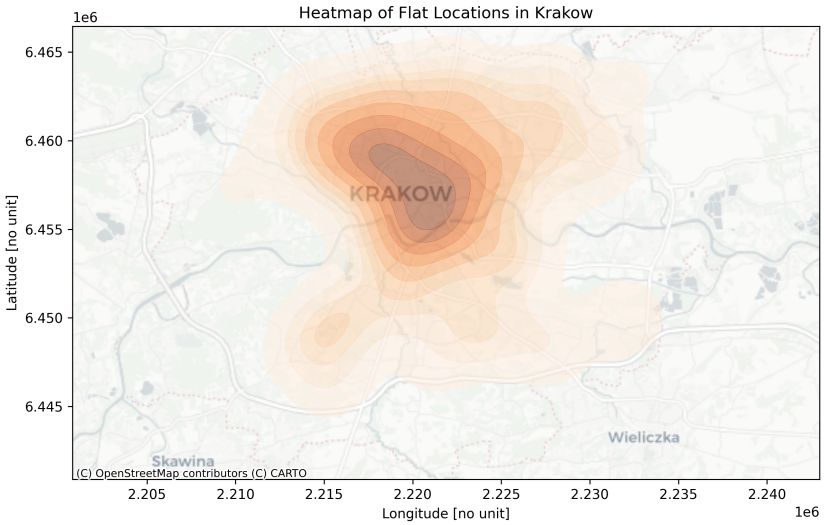

Figure 33: Heatmap of flat locations in Kraków.

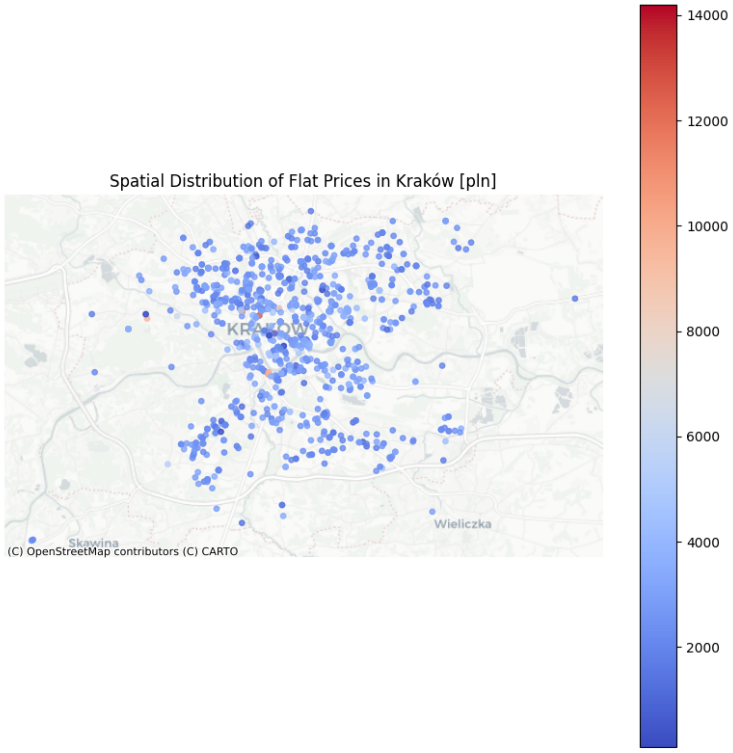

Figure 34: Price distribution of flats in Kraków.

7.1.3 Łódź

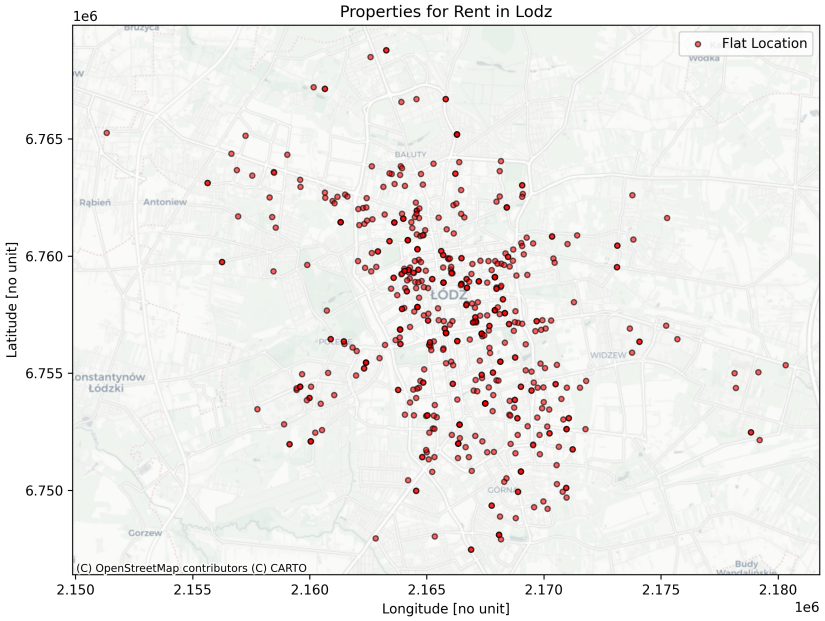

Figure 35: Spatial distribution of flats in Łódź with basemap.

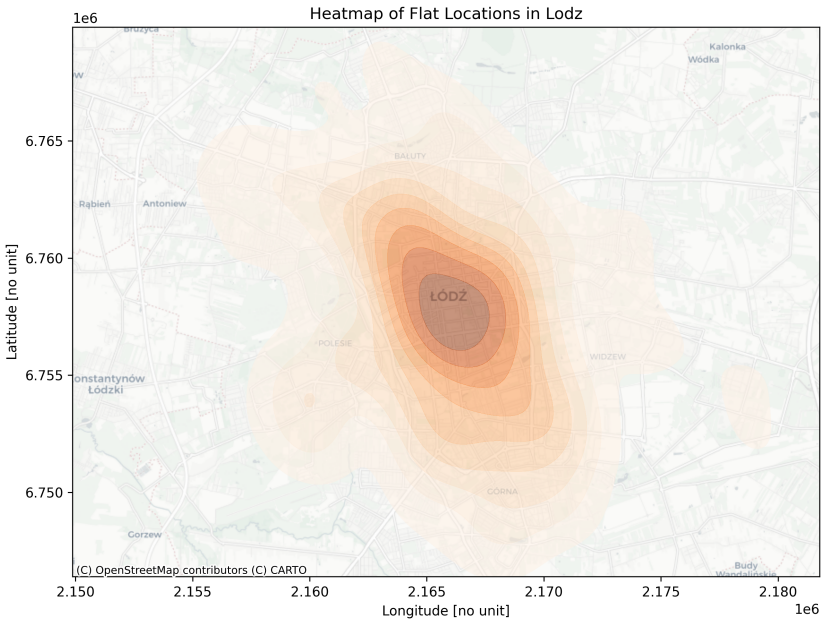

Figure 36: Heatmap of flat locations in Łódź.

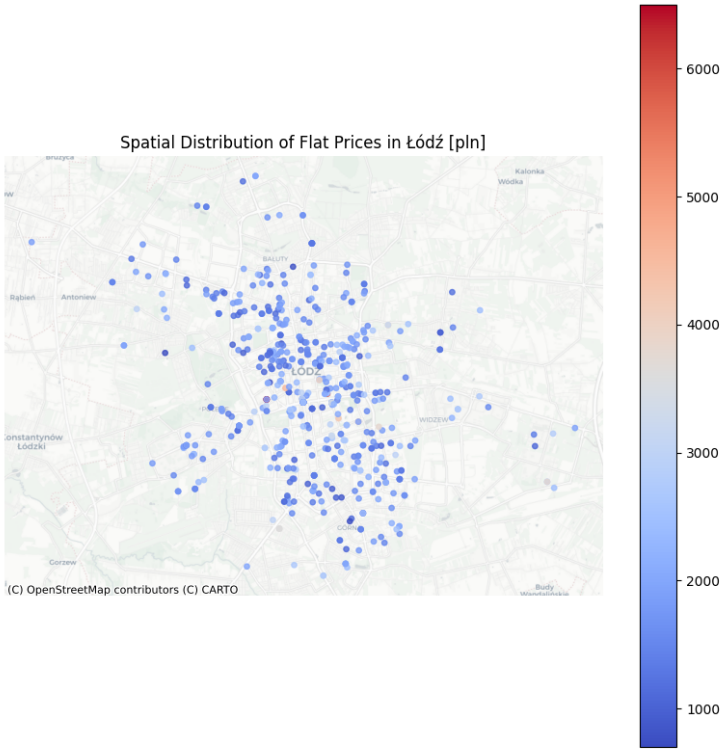

Figure 37: Price distribution of flats in Łódź.

7.1.4 Poznań

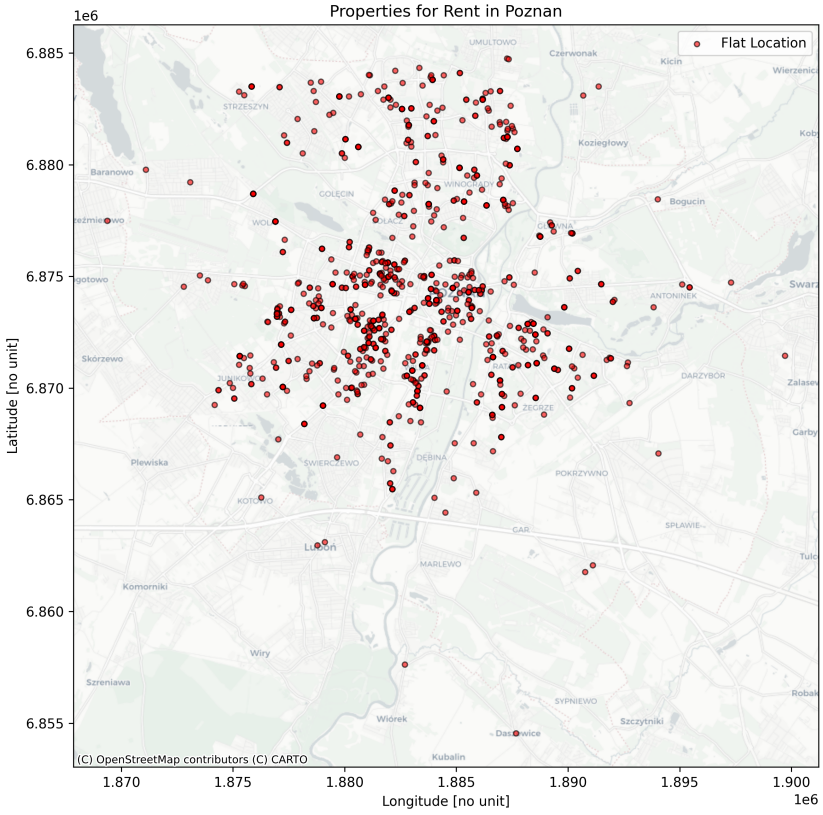

Figure 38: Spatial distribution of flats in Poznań with basemap.

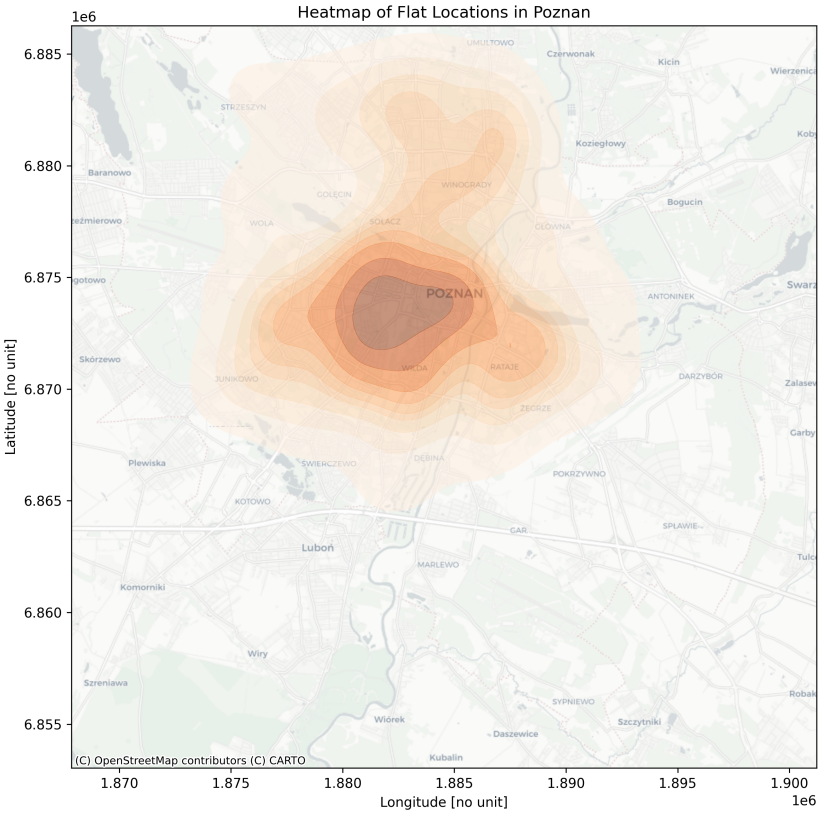

Figure 39: Heatmap of flat locations in Poznań.

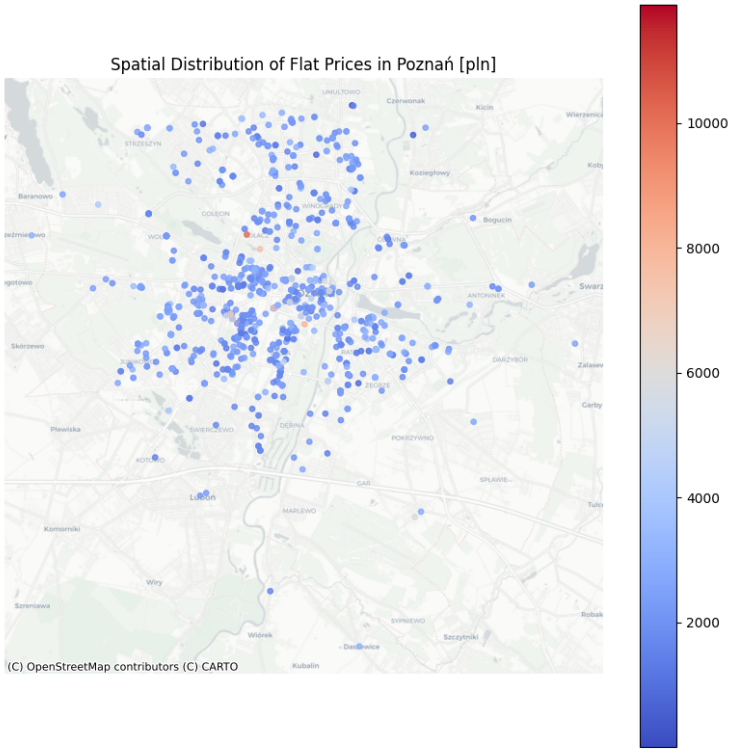

Figure 40: Price distribution of flats in Poznań.

7.1.5 Warszawa

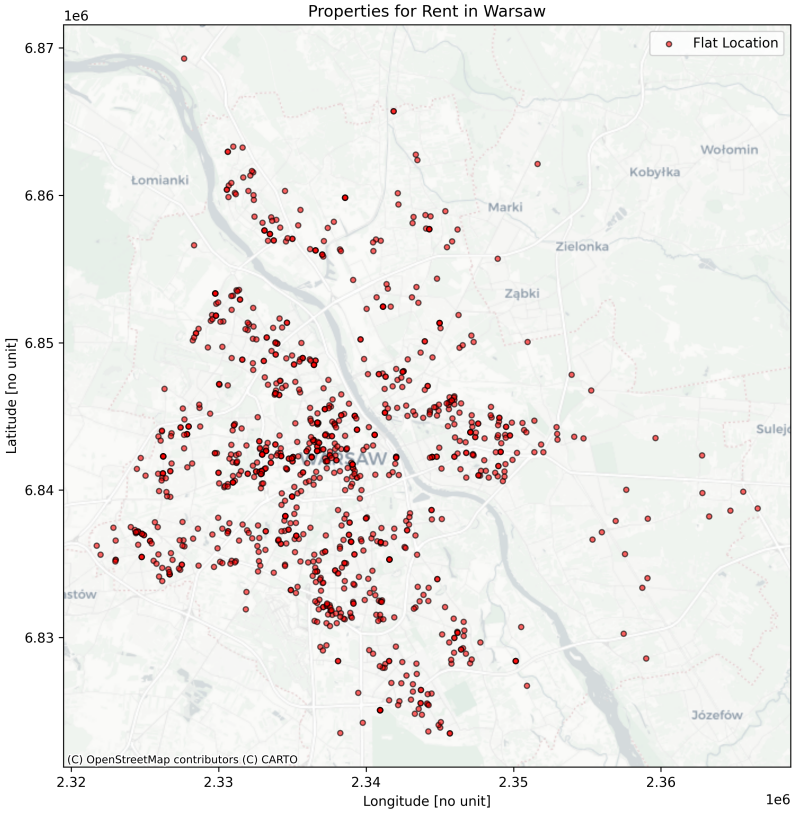

Figure 41: Spatial distribution of flats in Warszawa with basemap.

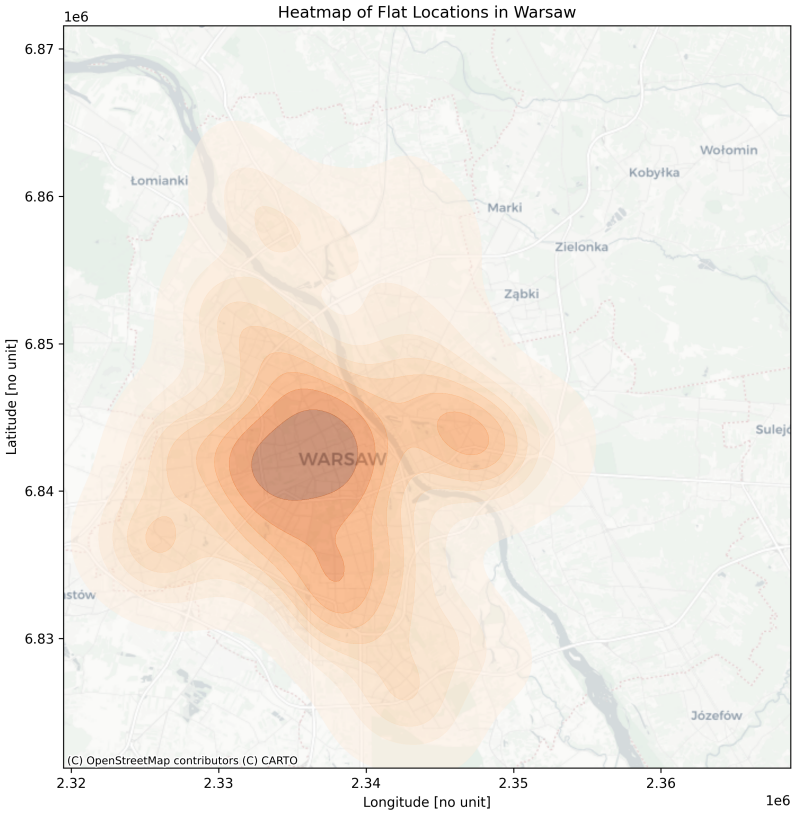

Figure 42: Heatmap of flat locations in Warszawa.

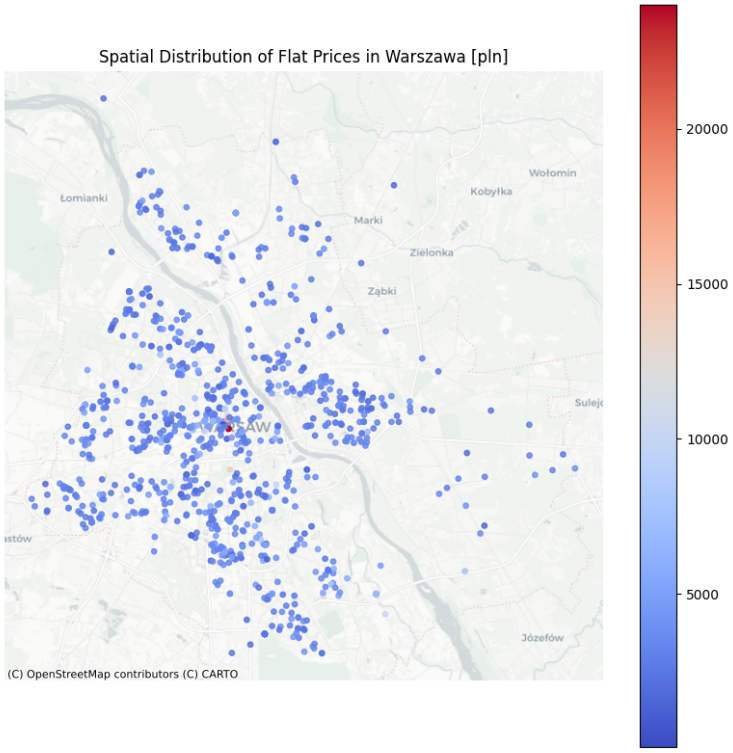

Figure 43: Price distribution of flats in Warszawa.

7.1.6 Wrocław

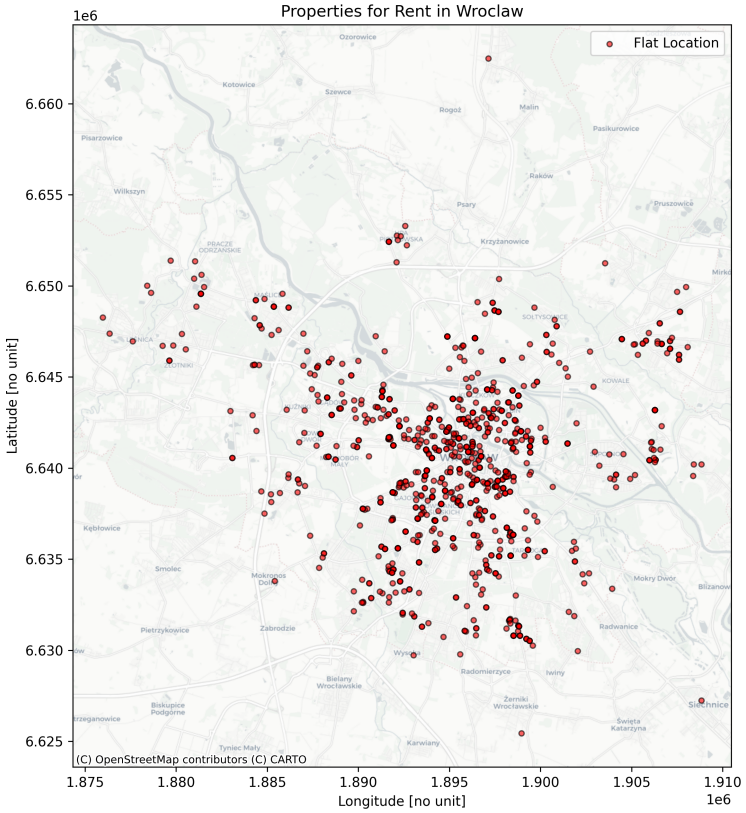

Figure 44: Spatial distribution of flats in Wrocław with basemap.

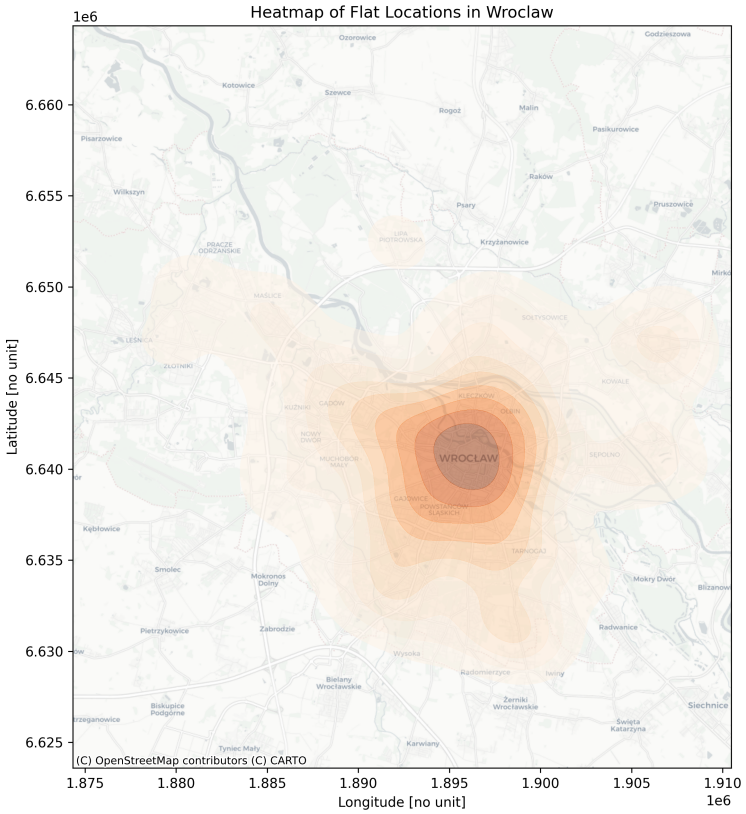

Figure 45: Heatmap of flat locations in Wrocław.

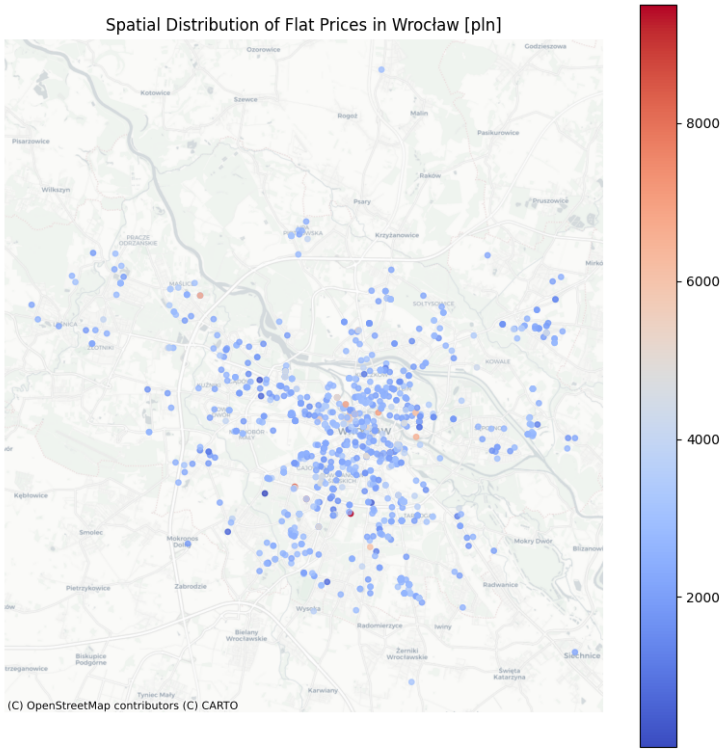

Figure 46: Price distribution of flats in Wrocław.