Computational Social Science

Analysis & Automated Valuation of German Real Estate

Department of Statistics
Ludwig-Maximilians-Universität München

Davit Martirosyan and Aaryan Mallayanmath

Munich, March, 2024

Contents

1 Abstract

Real estate plays a pivotal role in the German economy, experiencing dynamic growth and facilitated by the emergence of prominent online platforms for real estate advertising. This technological shift has mitigated information asymmetry, enhancing liquidity in both sales and rental markets. Leveraging detailed data from a substantial portion of the online real estate landscape, this study aims to monitor and analyze the German real estate market. In this study, an emphasis is placed on investigating the variations in apartment and house rental prices across different regions of Germany. The research seeks to identify the different factors influencing rental pricing, contributing valuable information to the understanding of the German real estate landscape. Moreover, we also isolated spatial features from the models to get better insight on location effects on the prices. An additional emphasis is placed on the application of machine learning models, particularly tree-based bagging and boosting ensembling methods, to accurately estimate the value of real estate rental prices based on internal and external features. Our research focuses on evaluating the efficacy of a specialized category of machine learning models—tree-based bagging and boosting ensembling methods—in predicting the rental prices of apartments and houses in Germany. To facilitate these objectives, scalable data collection pipelines were employed to establish a comprehensive German Real Estate database, serving as the foundation for robust models for price prediction. Leveraging OpenStreetMaps, we employed visualization techniques to depict rental prices across Germany. The analysis revealed that areas in and around prominent cities such as Munich, Hamburg, Frankfurt and Berlin exhibited the highest rental prices. This highlights the regional disparities within the real estate market in Germany. We also utilized a RandomForestRegressor model to identify the most important features influencing the rental prices. Our analysis determined that interior size and zip code were the two most important features for predicting apartment and house prices. The experimental findings highlight the superior performance of the **XGBoost** model achieving an R-squared value of **0.855** for apartments and **0.591** for houses model compared to **CatBoost** and **Random Forest Regressor**.

2 Introduction

Real estate is defined as a property, made up of land and any physical structures on it such as various types of buildings. The German real estate market size was valued at $372.77 billion in 2024, and is projected to reach $499.84 billion by 2029, growing at a compound annual growth rate of 3.06%.(1)

Growing transaction volumes and price fluctuations drive the development of automated valuation frameworks in the real estate sector. These frameworks aim to identify market opportunities and streamline valuation processes while minimizing errors and human intervention. Our research focuses on evaluating the performance of three tree-based machine learning ensembling methods: XGBoost, Catboost, and Random Forests.

Moreover, our project aims to enhance predictive modeling in real estate and lay the groundwork for future studies. By leveraging advanced machine learning techniques, efficient data collection methods, and spatial analysis, we aim to improve decision-making processes and support informed investment strategies in the German real estate market. Through innovative methodologies, our research seeks to refine property valuation accuracy and efficiency, empowering stakeholders with valuable insights for rental property management and investment decision-making

3 Related Work

Automated real estate valuation is a popular topic in applied machine learning, and recent work describes the effectiveness of different tree-based ensembling techniques for accurate price prediction. We decide to focus upon the recent advacements in the field and discuss the most prevalent and promising approaches.

In their recent paper, *Prediction and Analysis of Chengdu Housing Rent Based on XGBoost Algorithm* (5), the authors compare the performance of three techniques: LightGBM, XGBoost, and Random Forest Regressor. According to the paper, the most promising performance was obtained by the XGBoost model. Using parameter tuning, the latter attained a coefficient of determination (R-squared) of 0.85 (on a scale of 0 to 1, 1 indicating excellent performance) (5). Furthermore, they achieved the following results:

Model	MSE	R2
RandomForestRegressor	0.06	0.83
XGBoost	0.04	0.85
LightGBM	0.05	0.84

Another paper, titled *Product marketing prediction based on XGboost and LightGBM algorithm* (4), discusses the LightGBM and XGBoost models for product marketing prediction. The overall conclusion that the paper reached is that they both perform relatively similarly, however, the overall RME's of the XGBoost model is relatively smaller. (4)

4 Research Methodology

In this section we will describe the stages of this project along with the methods we implemented. The project can be divided into the following main stages: Data collection, data preprocessing, and modeling.

The first stage, as mentioned above was naturally data collection. We created end-to-end data collection pipelines for the following online real estate markets: "immonet.de", "kleinanzeigen.de", and "engelvoelkers.com/de" . As we will be describe the data in greater detail in the Data autorefsection:Data.

Following data collection we needed to complete the data preprocessing stage. We segmented the data into different datasets according to the websites from which they were scraped. One containing data that includes information about apartments and the other which contains information about houses. We implemented outlier detection with the interquartile range (IQR) method for the following features: Room count, interior size and price (for apartments) and Room count, floor count, interior size, exterior size and price (for houses). We filtered out any row that had values higher than the IQR multiplied by 1.5 plus the third quartile (Q3), or lower than the first quartile, Q1 subtracted by IQR multiplied by 1.5, for any of the above mentioned columns. Furthermore, a custom function was introduced to transform 'Zip Code' values based on their length, ensuring a standardized format and finally, missing values were addressed across different datasets by replacing them with the median of each respective column. Afterwards, we concatenated the different datasets into one normalised data set giving us 28740 observations for apartments and 2444 observations for houses.

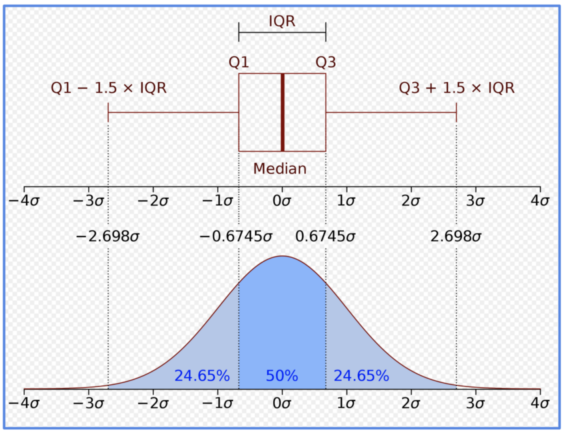

Figure 1: IQR method

Using the split datasets (apartments and houses) is a necessary step as the data for houses is generally more volatile compared to apartments which is why one model for both real estate types would result in a lower accuracy. Using hyperparameter tuning, we fed the data to three models: XGBoost, Random Forest Regressor, and Catboost. Also, we have used the *Random Forest Regressor* approach to obtain feature importance values for each case.

4.1 XGBoost

Extreme Gradient Boosting (XGBoost) (3) is an improved version of the Gradient Boosting Decision Tree (GBDT). This algorithm is composed of multiple decision trees, and the gradient descent method is used to "boost" each tree, meaning that we learn to adjust the regressor per the error found in a specific tree. Based on all single decision trees, the optimization is carried out by minimizing the loss function as the objective. Unlike the GBDT algorithm, the XGBoost algorithm can automatically use the CPU for multi-threaded parallel computation and carry out Taylor's second-order expansion on the loss function. Meanwhile, the tree model complexity is taken as a regular term in the target function to avoid over-fitting. The target function of the XGBoost algorithm is as follows (4):

$$L\left(f_t\right) = \sum_{i=1}^{\infty} l\left(y_i, \hat{y}_i^{t-1}\right) + \Omega\left(f_t\right) + C \tag{1}$$

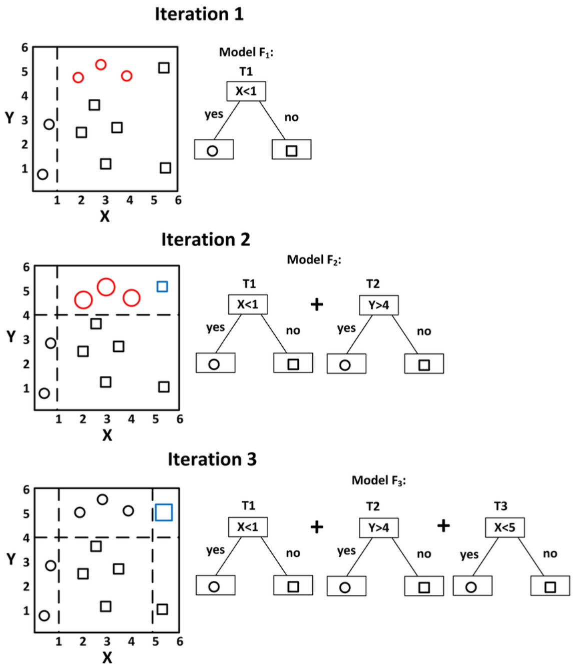

Figure 2: Gradient boosting

4.2 Random Forest

The Random Forest Algorithm (2) is composed of multiple decision trees, each with the same nodes, but using different data, leading to different leaves. Merging the decisions of multiple decision trees, the algorithm finds an answer, which represents the average of all these decision trees. When using the Random Forest Algorithm to solve regression problems, the mean squared error (MSE) is used to know how the data branches from each node:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (f_i - y_i)^2 \tag{2}$$

Where N is the number of data points, f_i is the value returned by the model and y_i is the actual value for data point i. (2)

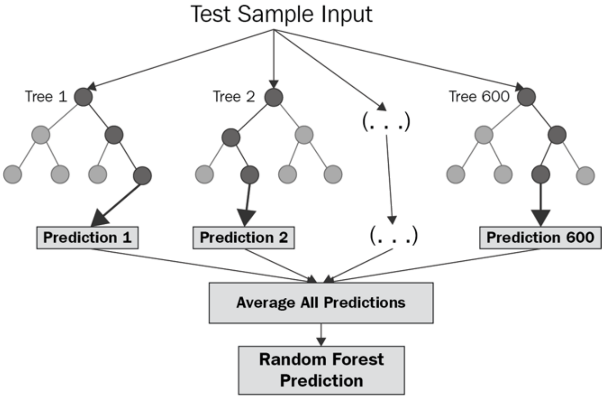

Figure 3: Random forest

4.3 Catboost

CatBoost builds upon the theory of decision trees and gradient boosting. The main idea of boosting is to sequentially combine many weak models and through greedy search create a strong competitive predictive model. One of CatBoost's core edges is its ability to integrate a variety of different data types, such as images, audio, or text features into one framework. CatBoost also offers an idiosyncratic way of handling categorical data, requiring a minimum of categorical feature transformation, opposed to the majority of other machine learning algorithms, that cannot handle non-numeric values. From a feature engineering perspective, the transformation from a non-numeric state to numeric values can be a very non-trivial and tedious task, and CatBoost makes this step obsolete. (6)

5 Data

In order to acquire the necessary data, we had to design data collection pipelines for the following online real estate websites: "immonet.de", "kleinanzeigen.de"and "engelvoelkers.com/de. Majority of the data came from "kleinanzeigen.de". Below is additional information about the data collected from each website:

Website	**Observations**	
	Apartments	**Houses**
kleinanzeigen.de	18184	4569
immonet.de	12710	963
engelvoelkers.com/de	597	123

The following features were extracted: Interior size, room count, zip code, latitude, longitude, floor level (which floor the apartment is on), balcony, storage, elevator and price (for apartments) and house type, year constructed (if available), interior size, exterior size (outside area of a house), room count, zip code, latitude, longitude, floor count (how many floors does the house have), parking, terrace, basic neccessities rating (if available), transportation availability rating(if available) and price (for houses).

In the concatenated data set for houses, basic neccessities rating and transportation availability rating have been excluded as the data wasn't available on every website.

6 Data Exploration

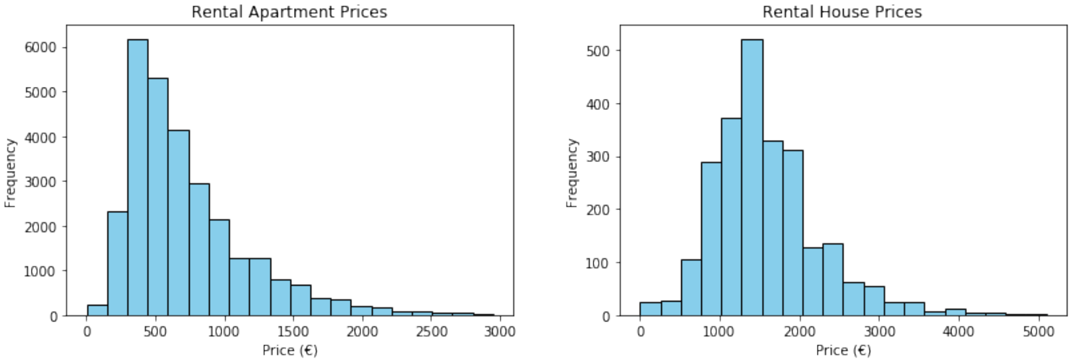

Figure 4: Left: Apartments Price Distribution, Right: Houses Price Distribution

Figure 4 illustrates the distribution of the rental prices of both apartments and houses based on data scraped from the above mentioned sources. This is the target variable that we aim to predict. The histograms peak at around the €350 to €600 and €1000 to €1700 price categories for apartments and houses respectively, and decline in reverse proportion to price, as expected.

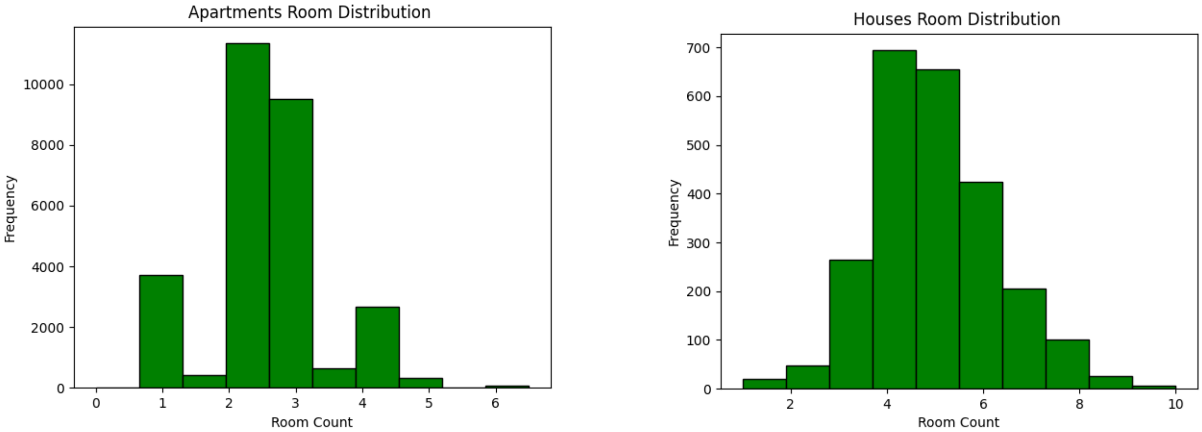

Figure 5: Left: Apartments Room Distribution, Right: Houses Room Distribution

In *Figure 5* it can be observed the distributions of rooms. Apartments mostly have 2 to 3 rooms, while houses 4 to 6. We can also see that in case of houses the total options for amount of rooms is more than that of apartments.

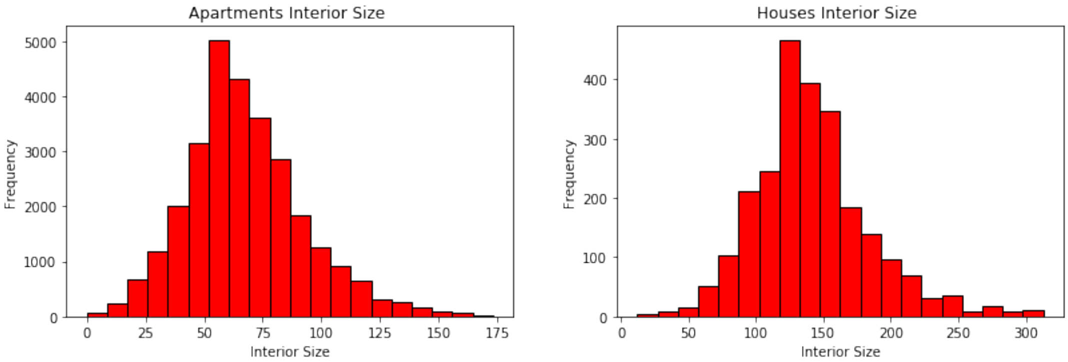

Figure 6: Left: Apartments Area Distribution, Right: Houses Area Distribution

Figure 6 illustrates the distribution of interior surface area. In the case of apartments, we can see a relatively normal distribution with most of the data gathered around the 50 to 70 square meters mark, while for houses the 125 to 175 square meter mark. This is expected, as houses usually tend to have a larger surface area in comparison to apartments.

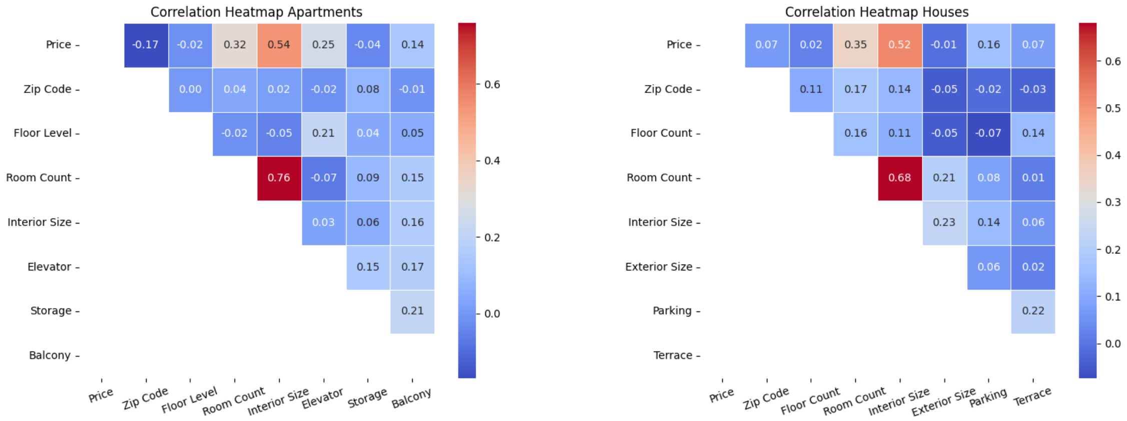

Figure 7: Left: Houses Correlation Heatmap, Right: Apartments Correlation Heatmap

In *Figure 7* the correlation heatmaps provide valuable insights into the relationships among key variables for houses and apartments. Notably, within the apartment category, a moderately strong positive correlation is observed between interior size and the target variable price, suggesting that the size of the living space significantly influences pricing. Furthermore, a strong positive correlation is observerd between the number of rooms and interior size, aligning with the expectation that larger apartments typically offer a greater number of rooms. This correlation pattern is consistently mirrored in houses, indicating a general trend across different property types.

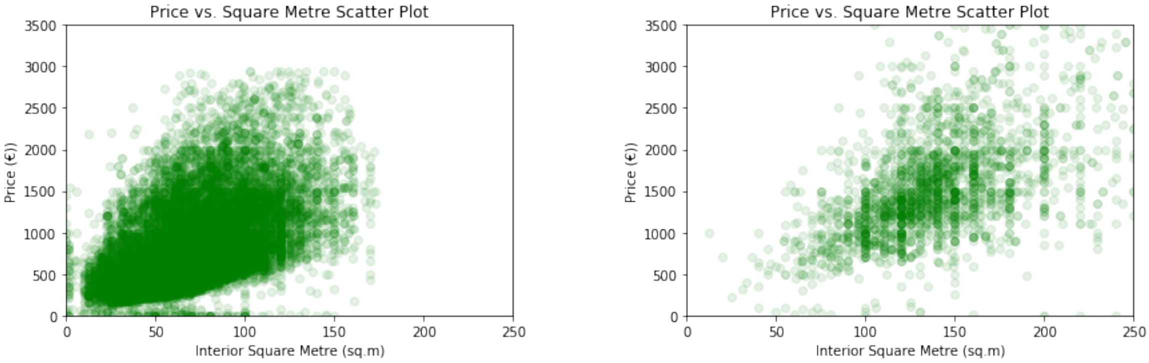

Figure 8: Left: Apartment Price vs Sq.m. scatter plot, Right: House Price vs Sq.m. scatter plot

In *Figure 8* the scatter plot belonging to apartments data exhibits a distinct funnel shape, with data points densely concentrated towards the center and gradually dispersing towards the outer edges.The densely packed nature of the data points suggests a high degree of consistency or correlation between the variables, potentially indicating a strong relationship or underlying trend .The second scatter plot belonging to the houses data also exhibits a funnel-like pattern, although the narrowing effect is less pronounced compared to the first plot. Here, the data points are notably dispersed, indicating greater variability across the range of values and potentially greater heterogeneity withing the data set. Despite the less condensed appearance, the general trend towards convergence at the center suggests a similar underlying pattern of association between the variables, albeit with greater dispersion.

7 Results & Analysis

7.1 Price Prediction

After subsetting the data into houses and apartments, we checked each column for missing values. We then imputed the median of each column (We used the median instead of the mean as this is better when the distribution of a given column is not completely normal). We then split the data into training and testing sets for both datasets. Moving forward, we conducted price prediction for each dataset, leveraging information from diverse real estate websites mentioned earlier. Notably, the data retrieved from 'immonet.de' enriched our analysis with additional features, specifically the 'Basic Necessities Rating' and 'Transport Availability Rating.' We were keen to examine the impact of these supplementary features alongside the existing variables on our target variable, price."

7.1.1 Feature Importance

As mentioned above we used *Random Forest Regressor* for feature importance. We implemented the latter on both datasets, subsetting houses and apartments. In the graphs below can be observed the results:

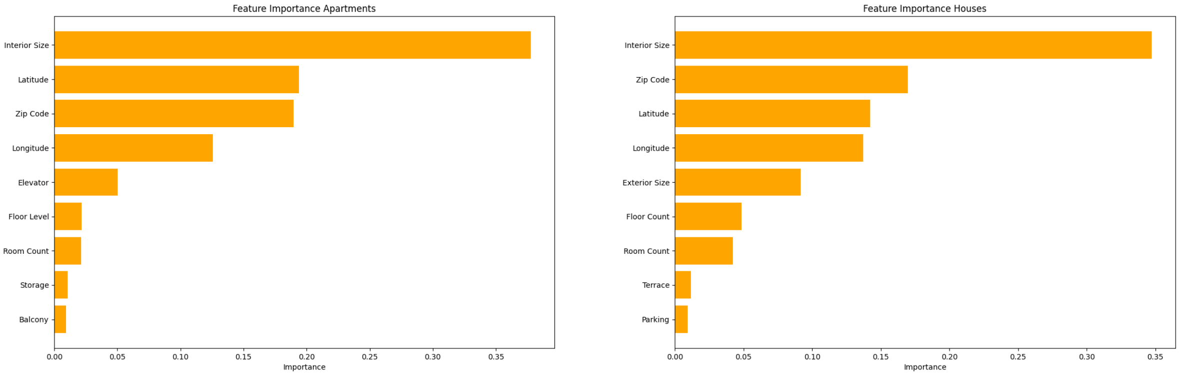

Figure 9: Left: Apartments Important Features, Right: Houses Important Features

In *Figure 9* we can see the important features for apartments and houses using the concatenated dataset. Notably, the 'interior size' feature emerges as consistently significant for both property types, indicating its strong influence on predicting prices. Equally noteworthy is the significance attributed to the 'Zip Code' feature, underlining its substantial impact on the pricing dynamics.

Interestingly, the role of the 'exterior size' feature in case of houses appears to be less significance, suggesting that, in this dataset, other factors may contribute more significantly to predicting house prices.

7.1.2 Regression Models

In this section, we will show the results attained using both datasets. As already noted, all models underwent hyperparameter tuning in order to get the optimal parameters for each. The evaluation metric mean absolute error (MAE) shows the absolute value of the average error (how far the model predicted from the actual value). Firstly, we will show the results from the apartments dataset with and without rating features available:

Model Apartments	MAE	R2
RandomForestRegressor	145.29	0.752
XGBoost	149.43	0.754
CatBoost	155.07	0.742

From the table above, it can be observed that the best results for the houses model was yielded by the XGBoost algorithm, achieving an R-squared of 0.754, and an MAE of 149.43

Next let us examine the regression models' performance on the houses dataset:

Model Houses	MAE	R2
RandomForestRegressor	341.53	0.553
XGBoost	322.53	0.591
CatBoost	353.18	0.53

The table above shows the results for the houses model. Similar to the apartments model (yet with a much lower accuracy), the XGBoost model again yields the highest accuracy with an R-squared of 0.591 and an MAE of 322.53

Finally, let us examine the regression models' performance on the apartment and houses dataset complemented with the rating features:

	Model	MAE	R2
Apartments	RandomForestRegressor	124.75	0.850
	XGBoost	125.05	0.855
	CatBoost	135.32	0.841
Houses	RandomForestRegressor	430.39	0.432
	XGBoost	415.73	0.531
	CatBoost	375.89	0.579

From the table above, it is evident that, once again, the XGBoost model stands out as the best performer for the apartments dataset, achieving an R-squared value of 0.855 and a MAE of 125.05. In contrast, for the houses dataset, all the models exhibit moderate performance, with CatBoost being the most effective among them, producing an R-squared of 0.531 and a MAE of 415.73.

7.1.3 Residuals Visualizations

Here we will look at some visualizations that will help better understand model performance:

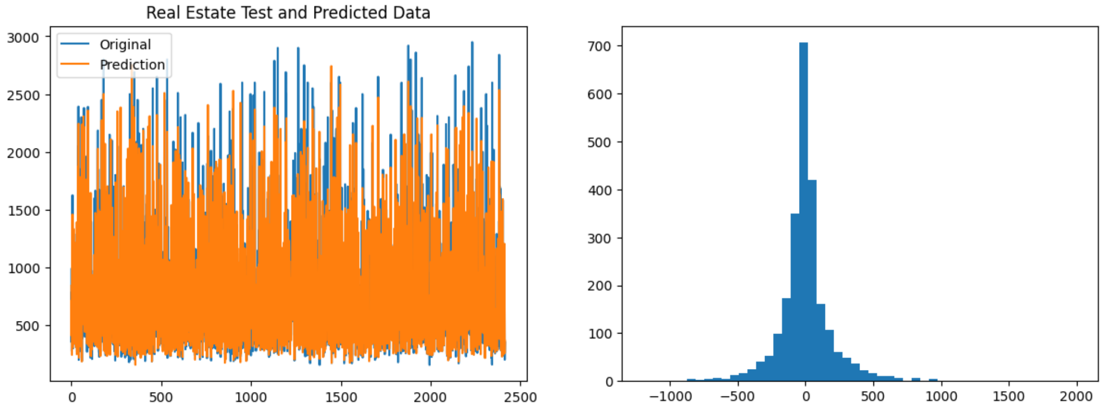

Figure 10: Left: Actual vs Prediction Apartments, Right: Residuals Distribution Apartments

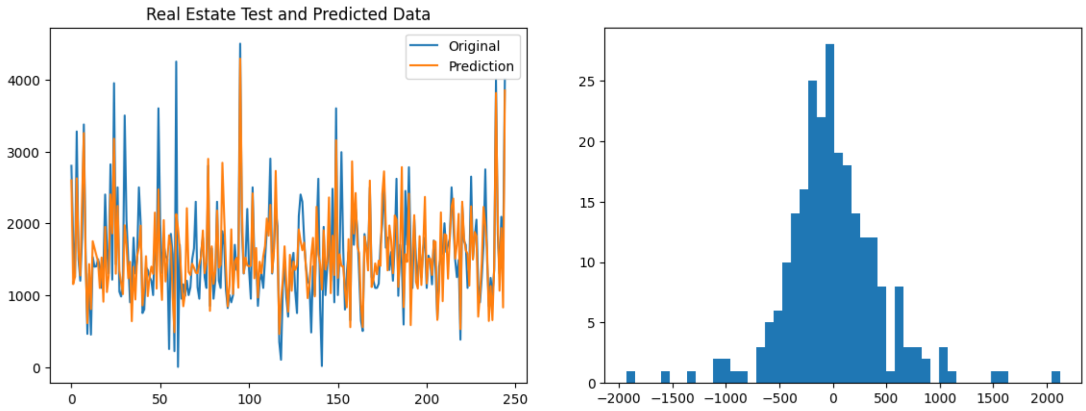

Figure 11: Left: Actual vs Prediction Houses, Right: Residuals Distribution Houses

The rightmost visualizations in *Figures 10* and *11* depict the distribution of residuals for the best apartments and houses models, respectively. In both cases, the plots exhibit a normal distribution, indicating a good fit.

7.1.4 Density Heatmaps for Spatial Analysis

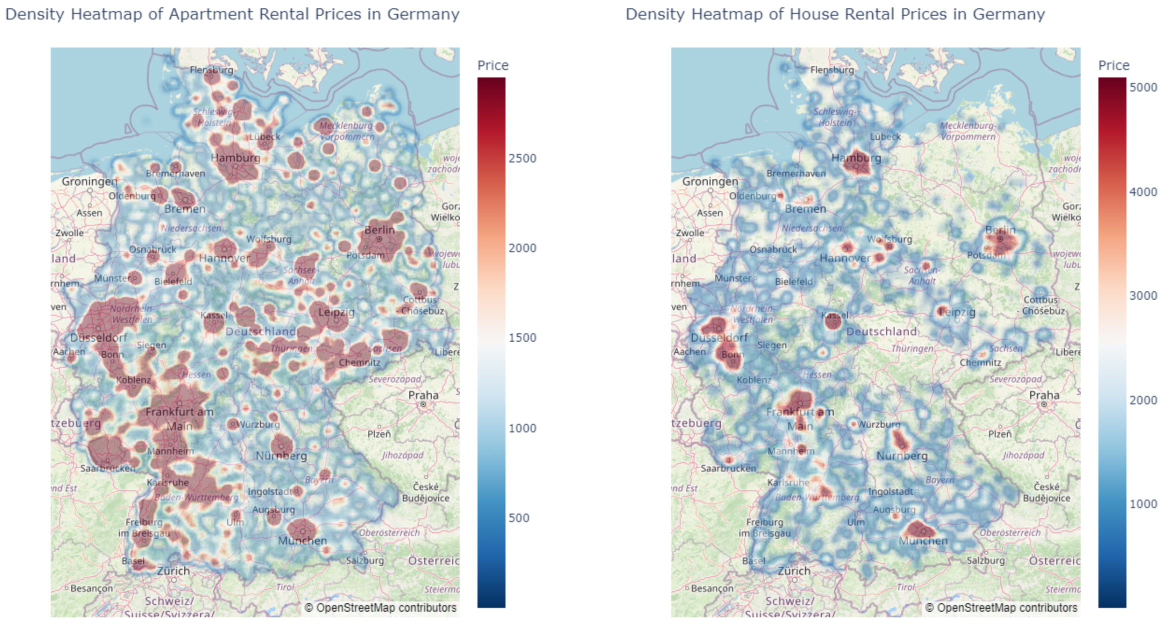

Figure 12: Left: Apartment Rental Prices Heatmap, Right: Houses Rental Prices Heatmap

Figure 12 represents the geographical distribution of rental prices of apartments and houses in Germany. The intensity of the color gradients on the map reflects the density of rental prices, with warmer color indicating higher prices and cooler colors indicating lower prices. Prominent cities such as Munich, Berlin, Frankfurt and Düsseldorf clearly have higher rental prices as compared to other areas in the country. Beyond these metropolitan cities, the heatmap also reveals other pockets of intensified colors, suggesting other regions where the rental prices are notably higher that its surroundings. This pattern in consistent for both rental apartments and rental houses. This correlation emphasizes the broader regional influence on the overall housing market and the impact of local economic factors on residential property values.

Further, we will compare price predictions generated by models with and without spatial data to understand the impact of spatial variables on predicting prices.

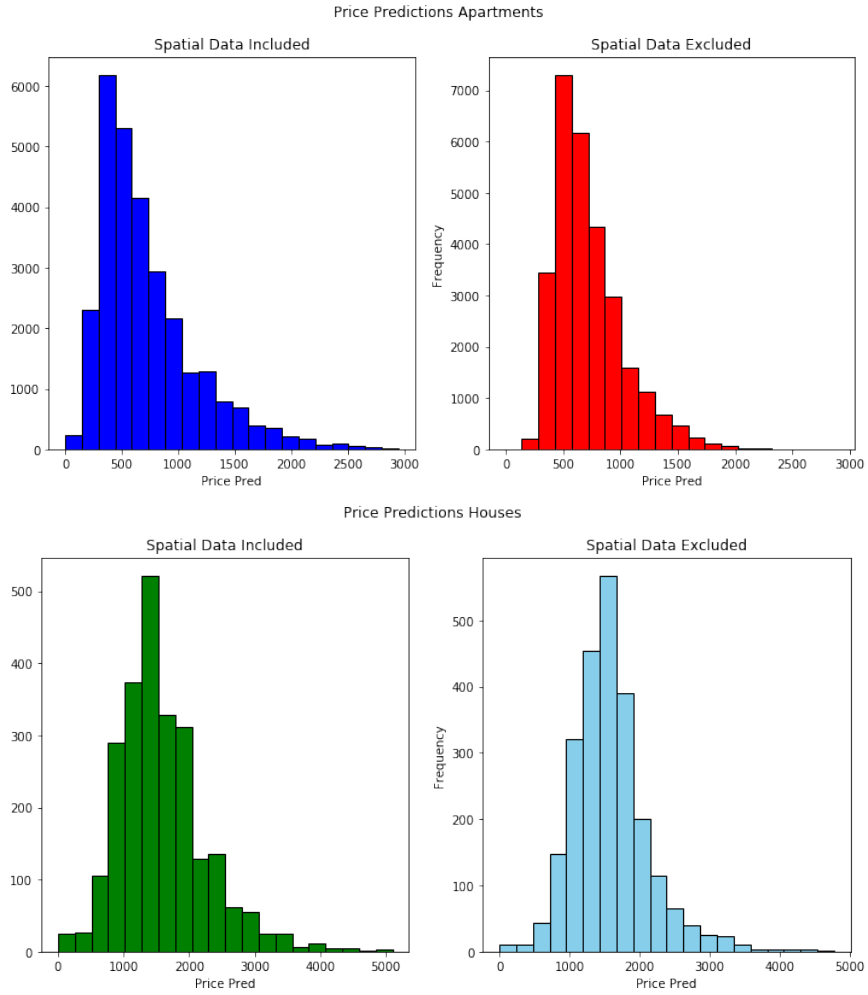

Figure 13: Comparison of Price Predictions

In *Figure 13*, the histograms illustrate rental price predictions for apartments and houses with and without spatial data. For apartments, although the shape remains consistent, notable differences in price distribution are observed. Without spatial data, prices peak around €500 to €700, with the highest frequency at 7000. Conversely, with spatial data, the peak shifts slightly to €350 to €600, with the highest frequency at 6000. Similarly, in houses, significant differences in price distribution are observed despite a consistent shape. Without spatial data, prices peak at €1200 to €1800, with the highest frequency at 520. With spatial data, prices peak slightly lower, around €1000 to €1700, with the highest frequency at 570. These findings suggest that spatial data influences price predictions, resulting in shifts in peak frequencies.

Lastly, we will look at density heatmaps for price predictions generated by models without spatial data.

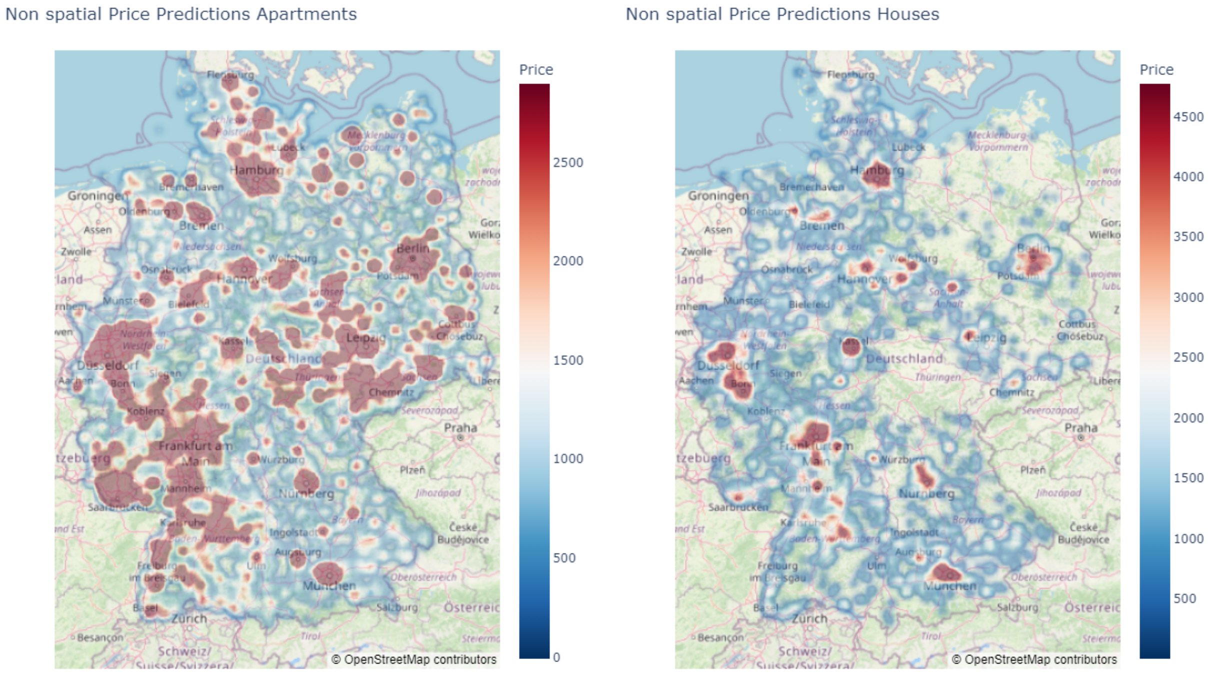

Figure 14: Left: Apartment Rental Prices Heatmap, Right: Houses Rental Prices Heatmap

Figure 14 offers a visual depiction of the geographical distribution of rental prices for both houses and apartments, derived from models developed without the incorporation of spatial data. Since the changes in price scales occurred in the model without spatial data, it indicates that non-spatial factors may be contributing to the observed variations in property prices. The variations in price scales may also indicate regional heterogeneity in housing markets, where different geographic areas exhibit unique pricing dynamics and market conditions. Regional differences in property types, market segments, or demographic characteristics of buyers and sellers could influence price dynamics and contribute to differences in predicted price scales. Without accounting for spatial effects, the model may fail to adequately adjust for these variations in sample composition, leading to differences in predicted price scales.

8 Discussion & Conclusions

In this project, we evaluated the performance of three families of tree-based ensemblers—XGBoost, Random Forest, and Catboost—on data extracted from the German real estate market. Through comprehensive analysis, we determined that XGBoost outperforms the other models in terms of predictive accuracy and computational efficiency. In addition to model evaluation, we developed a robust data collection pipeline capable of continuously expanding the dataset's size and iteratively enhancing model performance. This scalable approach ensures the sustainability and adaptability of our predictive modeling efforts, enabling us to accommodate evolving data landscapes and emerging trends in the real estate market. Moreover, our research extended beyond model architectures and hyperparameters to include the incorporation of spatial data—a critical component in real estate analysis. By integrating spatial information into our predictive models, we aimed to capture nuanced geographic patterns and spatial dependencies inherent in the German real estate market. While the impact of spatial data on model performance was inconclusive in this study, further exploration of spatial relationships and localized factors could yield valuable insights into pricing dynamics and market trends. Looking ahead, several avenues for future research could further enhance our understanding and utilization of spatial data in real estate analysis. Exploring more sophisticated spatial modeling techniques and conducting localized analyses at smaller geographic scales or within specific submarkets could uncover spatial dynamics and trends that may not be apparent at broader regional levels.

In summary, our project aims to improve predictive modeling in the real estate domain and provides a foundation for future research. Through the utilization of advanced machine learning techniques, efficient data collection methods, and exploration of spatial analysis, we seek to contribute to better decision-making processes and support informed investment strategies within the German real estate market.

A Appendix

A.0.1 Insights from Immonet Real Estate Data

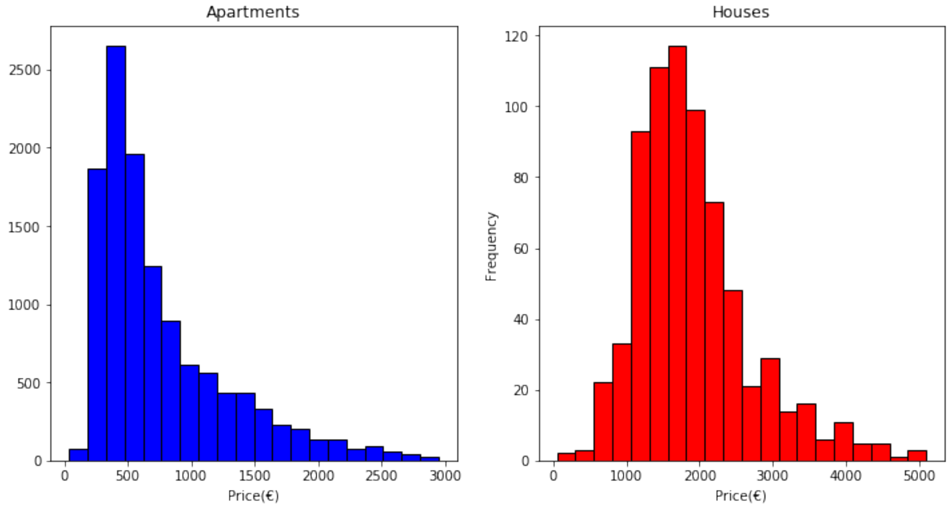

Figure 15: Left: Apartments Price Distribution, Right: Houses Price Distribution

Figure 15 illustrates the distribution of the rental prices of both apartments and houses based on data scraped from 'immonet.de'. The histograms peak at around €100 to €600 and €1500 to €2000 price categories for apartments and houses respectively, and decline in reverse proportion to price, as expected.

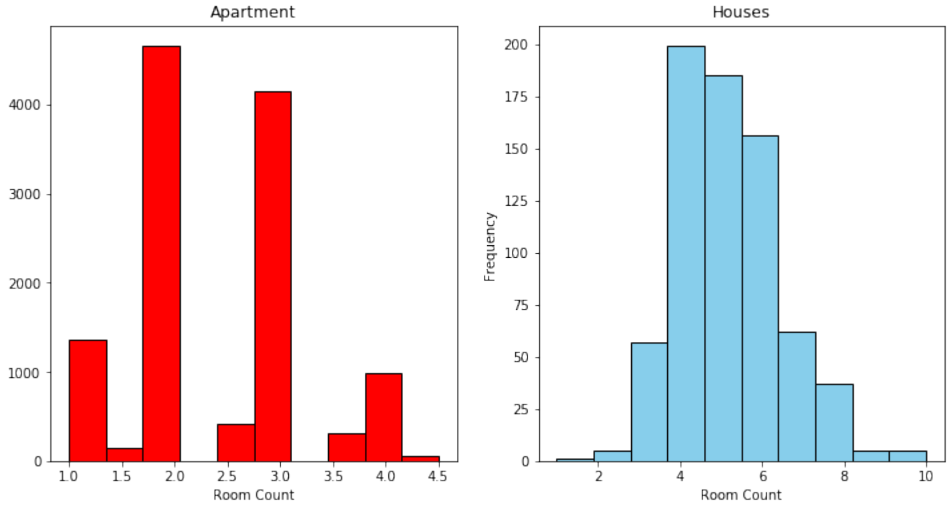

Figure 16: Left: Apartments Rooms Distribution, Right: Houses Rooms Distribution

In *Figure 16*, the distributions of rooms can be observed, with apartments predominantly having 2 to 3 rooms, while houses tend to have 4 to 6 rooms. Additionally, it is evident that houses generally have a higher total number of rooms compared to apartments.

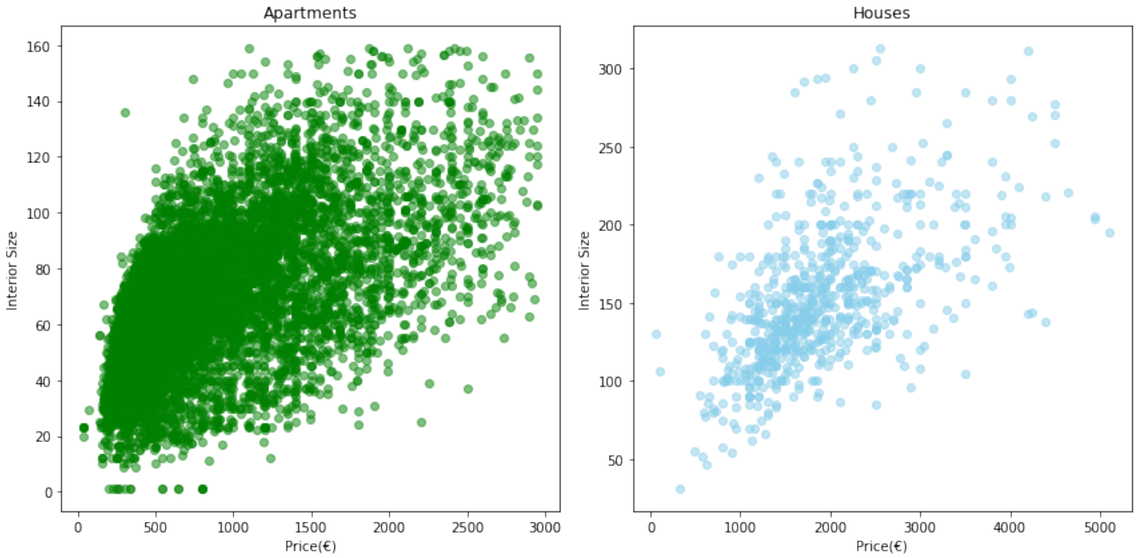

Figure 17: Left: Apartments Scatter Plot, Right: Houses Scatter Plot

Figure 16, describes the scatter plot of price versus interior size. For apartments, the plot exhibits a funnel shape, with data points concentrated more towards the central and upward portion, indicating a strong positive correlation between price and interior size. The scatter plot for houses also exhibits a funnel-like shape, though with more dispersed data compared to apartments. Despite the general trend of higher prices associated with larger interior sizes for houses, there is greater variability in prices among houses of similar sizes.

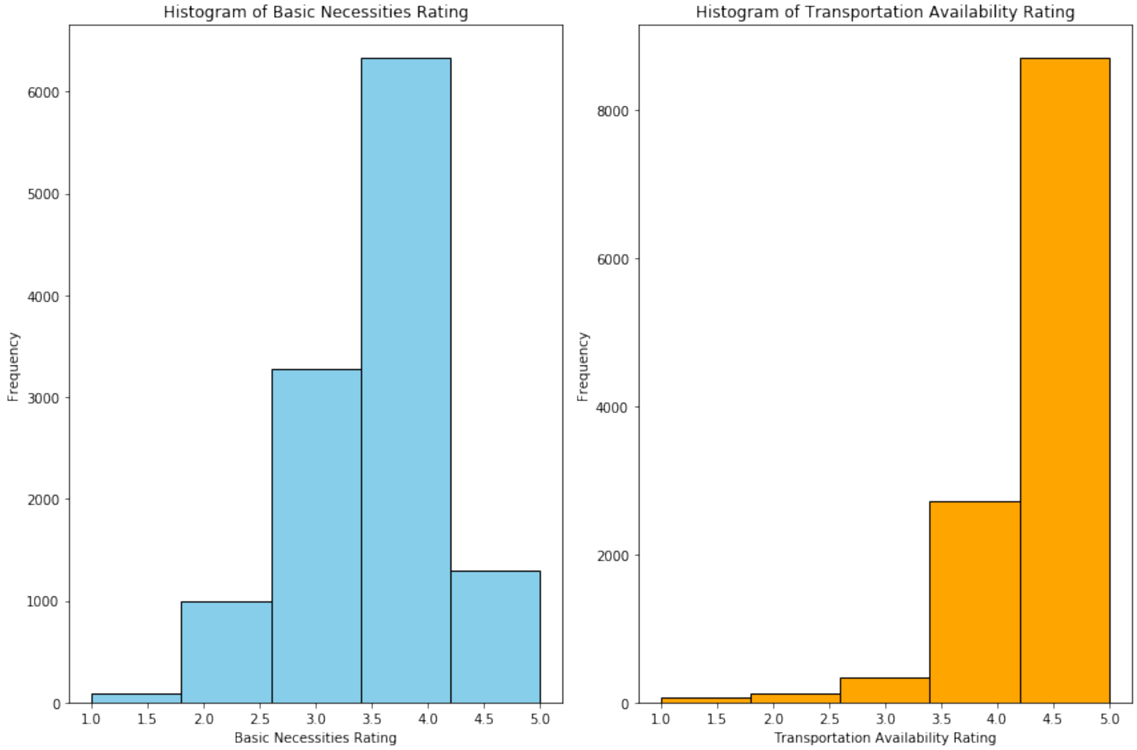

Figure 18: Left: Apartments Basic Necessities, Right: Apartments Transport Availibilty

Figure 18 illustrates the distribution of ratings for apartments, providing insights into the range and frequency of ratings assigned to different properties. For basic necessities, the peak around 3.5-4 indicates that most apartments likely have access to a satisfactory range of essential amenities such as grocery stores, schools, and healthcare facilities. Similarly, the peak in transportation ratings between 4.5-5 suggests that many apartments are situated in areas with excellent transportation options, such as proximity to public transit hubs, well-connected road networks, and short commute times. The presence of peaks at different ratings for basic necessities and transportation indicates that while some apartments may have satisfactory access to one aspect.

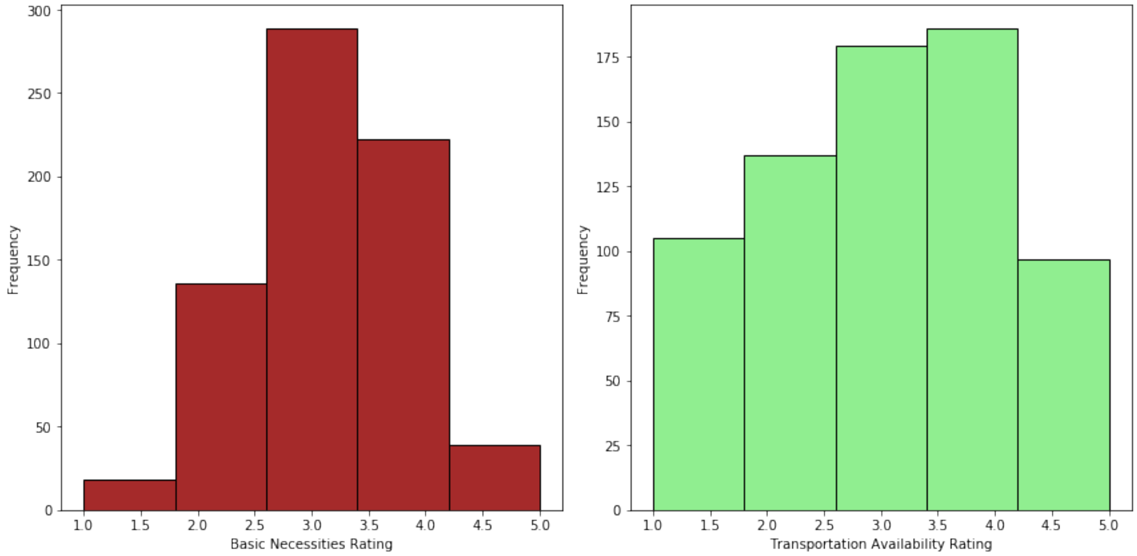

Figure 19: Left: Apartments Basic Necessities, Right: Apartments Transport Availibilty

Figure 19 illustrates the distribution of ratings for houses, providing insights into the range and frequency of ratings assigned to different properties. For basic necessities, the peak between 2.5-4 indicates that many houses have access to a moderate range of essential amenities, although the quality and availability may vary. Similarly, the peak in transportation ratings between 2-4.5 suggests that many houses are situated in areas with varying degrees of transportation options. Some houses may be located in neighborhoods with relatively efficient transportation networks, including public transit and road infrastructure, while others may have more limited access to transportation resources. Overall, the peaks in these histograms imply that the distribution of ratings for houses reflects the diverse range of neighborhoods and communities represented in the dataset. While some houses may offer relatively favorable access to both basic necessities and transportation, others may have more limited access to these amenities

A.0.2 Insights from Kleinanzeigen Real Estate Data

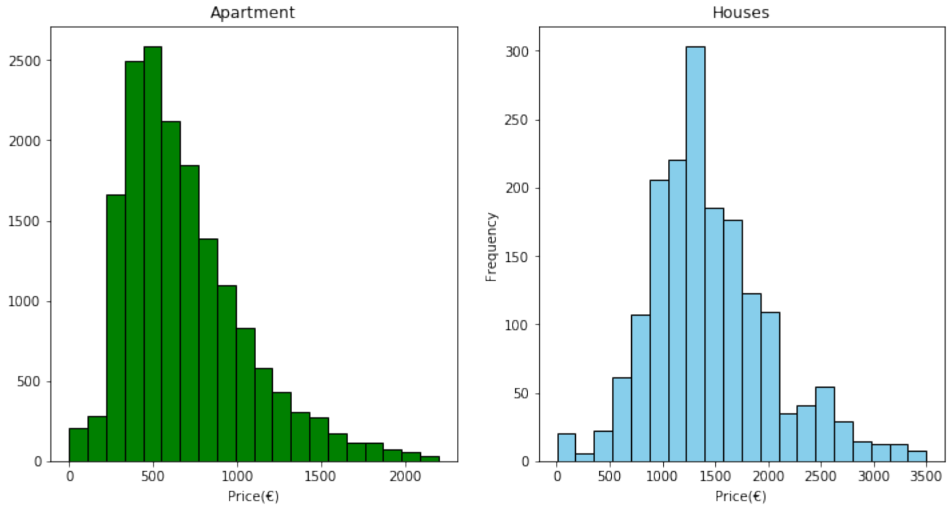

Figure 20: Left: Apartments Price Distribution, Right: Houses Price Distribution

Figure 20 showcases the distribution of rental prices for both apartments and houses sourced from 'immonet.de'. The histograms peak at approximately €400 to €600 and €1250 to €1500 price ranges for apartments and houses, respectively. Additionally, the histograms exhibit a decline in frequency as prices increase, aligning with expectations.

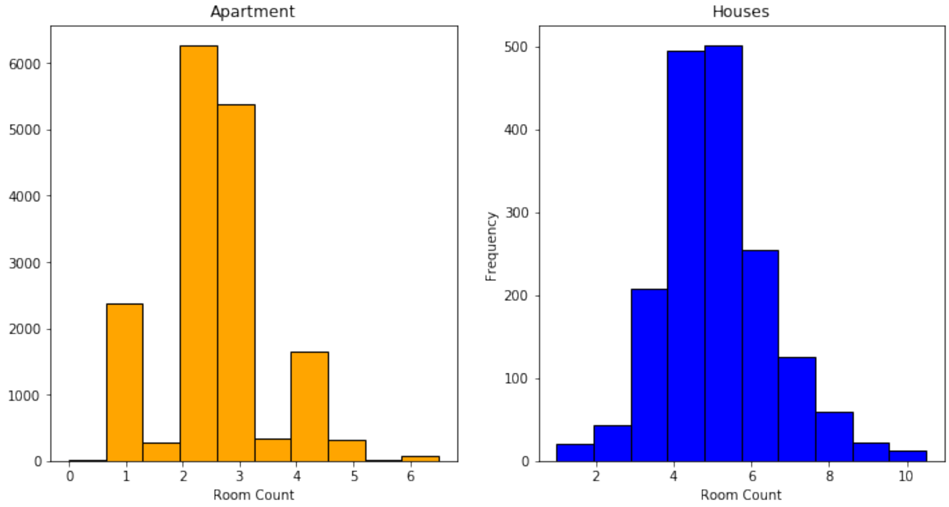

Figure 21: Left: Apartments Rooms Distribution, Right: Houses Rooms Distribution

In *Figure 21*, we can observe the distributions of rooms, where apartments typically feature 2 to

3 rooms, whereas houses tend to have 4 to 6 rooms. Moreover, houses generally have a greater total number of rooms compared to apartments as expected.

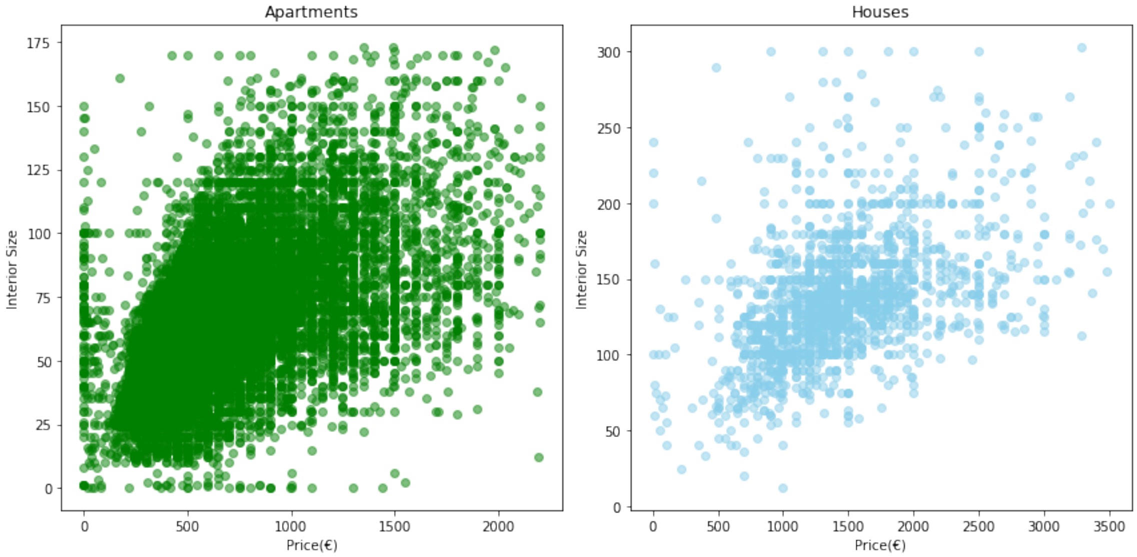

Figure 22: Left: Apartments Scatter Plot, Right: Houses Scatter Plot

Figure 22, depicts the scatter plot of price versus interior size. In the case of apartments, the plot exhibits a funnel shape, with the majority of data points concentrated in the central and upward regions. However, there are also instances of outliers, where properties have either a price of 0 with a large interior size or vice versa, indicating potential anomalies or errors in the dataset. For houses, the scatter plot similarly displays a funnel shape, with data concentrated towards the center. However, a significant portion of the data appears dispersed, suggesting greater variability in prices and interior sizes among houses.

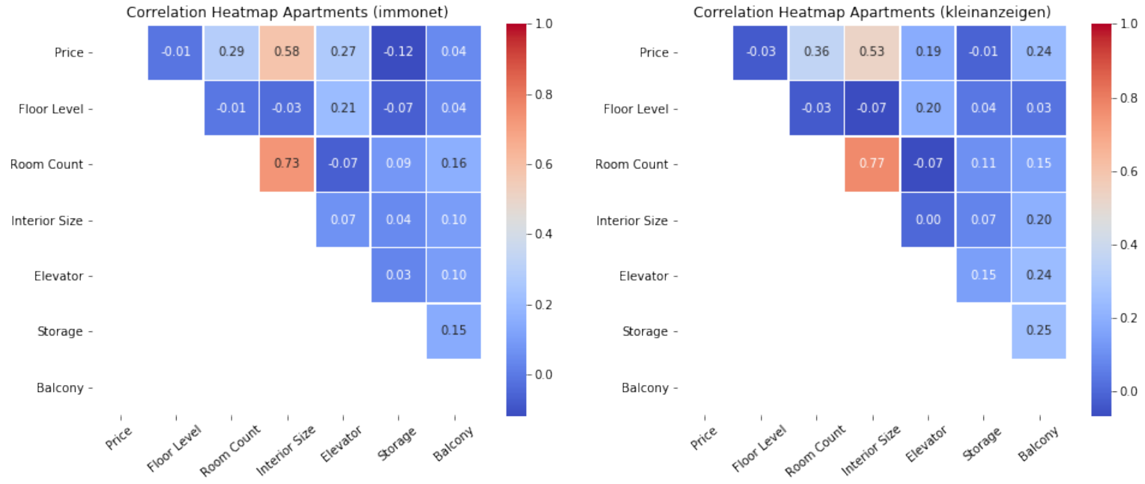

Figure 23: Left: Apartments Immonet Correlations, Right: Apartments Kleinanzeigen Correlations

In *Figure 23* the correlation heatmaps provide valuable insights into the relationships among key variables for apartments sourced from Kleinanzeigen and Immonet. A moderately strong positive correlation is observed between interior size and the target variable price, suggesting that the size of the living space significantly influences pricing. Furthermore, a strong positive correlation is observed between the number of rooms and interior size in both datasets, aligning with the expectation that larger apartments typically offer a greater number of rooms. This correlation pattern remains consistent across the data scraped from these two different websites.

A.0.3 Comparison with Concatenated Data

Both the individual datasets and the concatenated dataset exhibited similar trends in the distribution of prices and room counts, with peaks consistently observed in comparable ranges. While the general funnel shape remained consistent across all datasets, minor differences were noted upon analyzing the scatter plots. These variances included potential outliers and slight alterations in concentration patterns, particularly notable in the apartment dataset. Although the deviations were subtle, they underscore the importance of meticulously examining individual datasets for anomalies and understanding the potential impact of data aggregation on visual representations of underlying trends.

References

[1] Residential real estate market in germany. Online. `https://www.mordorintelligence.com/industry-reports/residential-real-estate-market-in-germany`.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[4] Y. Liang, J. Wu, W. Wang, Y. Cao, B. Zhong, Z. Chen, and Z. Li. Product marketing prediction based on xgboost and lightgbm algorithm. *Proceedings of the 2nd International Conference on Artificial Intelligence and Pattern Recognition - AIPR '19*, 2019.

[5] Y. Ming, J. Zhang, J. Qi, T. Liao, M. Wang, and L. Zhang. Prediction and analysis of chengdu housing rent based on xgboost algorithm. *Proceedings of the 2020 3rd International Conference on Big Data Technologies*, 2020.

[6] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.