





# Smart Renting: Harnessing Urban Data with Statistical and Machine Learning Methods for Predicting Property Rental Prices from a Tenant's Perspective

Francisco Louzada <sup>1,†</sup>, Kleython José Coriolano Cavalcanti de Lacerda <sup>2,†</sup>, Paulo Henrique Ferreira <sup>3,\*,†</sup>  
and Naomy Duarte Gomes <sup>1,†</sup>

<sup>1</sup> Institute of Mathematics and Computer Sciences, University of São Paulo, Av. Trab. São Carlsense, 400-Centro, São Carlos 13566-590, Brazil; louzada@icmc.usp.br (F.L.); naomy.gomes@usp.br (N.D.G.)

<sup>2</sup> School of Medicine, University of São Paulo, Av. Bandeirantes, 3900, Monte Alegre, Ribeirão Preto 14049-900, Brazil; kleython\_lacerda@usp.br

<sup>3</sup> Department of Statistics, Federal University of Bahia, Avenida Milton Santos s/n, Campus de Ondina, Salvador 40170-110, Brazil

\* Correspondence: paulohenri@ufba.br

† These authors contributed equally to this work.

**Abstract:** The real estate market plays a pivotal role in most nations' economy, showcasing continuous growth. Particularly noteworthy is the rapid expansion of the digital real estate sector, marked by innovations like 3D visualization and streamlined online contractual processes, a momentum further accelerated by the aftermath of the Coronavirus Disease 2019 (COVID-19) pandemic. Amidst this transformative landscape, artificial intelligence emerges as a vital force, addressing consumer needs by harnessing data analytics for predicting and monitoring rental prices. While studies have demonstrated the efficacy of machine learning (ML) algorithms such as decision trees and neural networks in predicting house prices, there is a lack of research specifically focused on rental property prices, a significant sector in Brazil due to the prohibitive costs associated with property acquisition. This study fills this crucial gap by delving into the intricacies of rental pricing, using data from the city of São Carlos-SP, Brazil. The research aims to analyze, model, and predict rental prices, employing an approach that incorporates diverse ML models. Through this analysis, our work showcases the potential of ML algorithms in accurately predicting rental house prices. Moreover, it envisions the practical application of this research with the development of a user-friendly website. This platform could revolutionize the renting experience, empowering both tenants and real estate agencies with the ability to estimate rental values based on specific property attributes and have access to its statistics.

**Keywords:** machine learning; statistics; artificial intelligence; real estate leasing; rental price



Academic Editor: Wei Zhu

Received: 29 August 2024

Revised: 20 January 2025

Accepted: 21 January 2025

Published: 27 January 2025

**Citation:** Louzada, F.; de Lacerda, K.J.C.C.; Ferreira, P.H.; Gomes, N.D. Smart Renting: Harnessing Urban Data with Statistical and Machine Learning Methods for Predicting Property Rental Prices from a Tenant's Perspective. *Stats* **2025**, *8*, 12. <https://doi.org/10.3390/stats8010012>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The real estate market is a cornerstone of economic growth and societal development, consistently adapting to evolving trends and challenges. The COVID-19 pandemic, while initially causing a slowdown in growth [1], catalyzed significant transformations in the sector, most notably the adoption of digital technologies. These innovations, such as 3D property visualization and seamless online contractual processes [2], have redefined how tenants and landlords interact, removing geographical barriers and enabling data-driven decision making in unprecedented ways.

Within this dynamic landscape, artificial intelligence (AI) and machine learning (ML) have emerged as indispensable tools. By enabling the precise collection and analysis of consumer and market data, AI facilitates the prediction of property values, empowering stakeholders across the real estate ecosystem. While extensive research has demonstrated the effectiveness of statistical and ML techniques in predicting property sale prices [3,4], rental markets, particularly in developing economies like Brazil, remain underexplored. This gap is critical given the growing emphasis on rental housing as a viable alternative to property ownership, driven by economic factors such as high interest rates and changing consumer preferences [5].

Focusing on the Brazilian city of São Carlos-SP—a hub of academic activity and a microcosm of diverse rental demands—this study addresses the pressing need for rental price estimation tools. São Carlos is uniquely positioned as a case study, given its demographic mix of students, professionals, and families, coupled with its robust real estate dynamics. By leveraging advanced statistical and ML models, this research aims to develop a comprehensive framework to analyze and predict rental prices. A user-friendly web platform is proposed to democratize access to these insights, empowering tenants and landlords alike with actionable, data-driven knowledge. By transcending geographical barriers and national borders and making renting properties a seamless experience for all, the website aims not only to offer accurate rental value estimates but also to provide valuable statistical insights.

This dual focus on academic contribution and practical application underscores the transformative potential of this study. It enriches the literature by addressing methodological gaps in rental market analysis, particularly in the application of ML to this sector. For practitioners, the study introduces innovative tools to optimize decision making, foster transparency, and promote fairness in rental transactions. By bridging the gap between research and practice, this work aspires to contribute to a more efficient, equitable, and accessible rental market.

This paper is structured as follows: the “Literature Review” section analyzes prior studies on ML applications in real estate, emphasizing the gaps that this research seeks to fill. The “Materials and Methods” section details the data collection and analytical approach employed. The “Results and Analysis” section presents findings, emphasizing their practical implications. Finally, the “Conclusions” section synthesizes key insights, discusses broader impacts, and outlines pathways for future research and development.

## 2. Literature Review

Given its economic and social significance, the real estate sector has been the subject of studies and analyses by experts [6–9]. In Brazil, this sector possesses unique characteristics and specific challenges arising from public policies, such as the Minha Casa Minha Vida program [10], for instance. However, parallels can be drawn with the real estate market in the United States [11]. A central feature of the real estate market is its volatility. Economic fluctuations, government policies, changes in financing conditions, and real estate cycles can impact property prices and demand. A prime example of this was the changes in property values during the COVID-19 pandemic [1]. Additionally, regulatory aspects, such as urban legislation and zoning regulations, can directly influence market development. Therefore, to make informed decisions and seize the opportunities offered by this sector, it is important to monitor and analyze both sale and rental prices.

Historically, property price prediction relied primarily on regression models, particularly the hedonic regression approach [12,13]. While these models laid a strong foundation for understanding property value determinants, the emergence of machine learning (ML) techniques has significantly enhanced predictive accuracy, enabling the capture of com-

plex, non-linear relationships inherent in real estate markets [14]. For instance, random forests have been successfully applied to predict apartment prices in Ljubljana, Slovenia, demonstrating their effectiveness in handling diverse property attributes [15]. Similarly, decision trees have been employed to explore the relationship between house prices and their characteristics, yielding reliable predictions [16]. Advanced ML models, including artificial neural networks, support vector regression (SVR), and XGBoost, have emerged as some of the most effective tools for property price prediction, further illustrating the transformative impact of these techniques on the field [17].

The real estate market in São Carlos, a city located in the state of São Paulo, Brazil, is a constantly expanding market, featuring characteristics that make it an excellent source of data, and has been the subject of several studies over the years. The author in [18] applied statistical techniques to a dataset of urban plots in the city of São Carlos in 2005, estimating an empirical function of land value. She explored the class of models, including linear models, generalized linear models, and generalized additive models for location, scale, and shape (GAMLSSs). Considering the existence of two types of real estate prices—the sold (observed) and the advertised (censored)—her thesis also proposed the use of survival analysis, considering left-censoring and GAMLSSs in the parameter estimation process. Additionally, she conducted a simulation study and a local influence study. Also applying statistical techniques, the work of [19] proposes a representative regression equation for determining the market value of urban lots in the municipality of São Carlos, also in the year 2005, using standard linear models (normal errors and constant variance). The study in [20] focuses on a social analysis, where the authors explore the impacts of the Minha Casa Minha Vida program in São Carlos, discussing the real estate market and social inequalities. Lastly, [21] seeks to understand students as social agents driving the growth of the real estate market. The study also analyzes the spatial influences of real estate agents and their implemented actions to meet this demand. Such studies demonstrate the importance of the real estate market in São Carlos, which offers a diverse range of properties, from residential apartments and houses to commercial and industrial properties. The city is home to renowned companies and educational institutions such as the University of São Paulo (USP) and the Federal University of São Carlos (UFSCar), attracting a varied demand for properties. The presence of these two major universities results in a significant influx of tenants every year, which makes the city an excellent case study.

### 3. Materials and Methods

The real estate data used in this study were obtained from the website of São Carlos' largest real estate agency, Imobiliária Cardinali [22], as it boasts the highest number of properties available for rent. While the website provides continuously updated data, the dataset utilized in this manuscript corresponds to a specific period from December 2022 up to February 2023. The website is publicly accessible and offers various parameters for selection, such as whether the property is for rent or sale, neighborhoods, and minimum and maximum values, among others. Given the available information, the following property characteristics will be used as predictor variables in our database: *Type* ("Apartment" or "House"), *Bedrooms* (number of bedrooms), *Bathrooms* (number of bathrooms), *Garages* (number of parking spaces), *Neighborhood* (property location), *Suites* (number of suites), and *Furnished* ("Yes" or "No", depending on whether the property is furnished or not). Once the information to be stored in our database is defined, web scraping is performed using the packages `requests` version 2.31.0 [23] and `Beautiful Soup` version 4.12.3 [24]. The former is a Python library that provides a simple and convenient way to make HyperText Transfer Protocol (HTTP) requests. It allows for sending HTTP requests to web servers and receiving responses, whether for retrieving information, sending data, or interacting with application

programming interfaces (APIs). Moreover, it simplifies complex tasks related to web service interactions, such as authentication, cookie handling, and parameter and header submission, among others. On the other hand, `Beautiful Soup` is a Python library used to extract information from HyperText Markup Language (HTML) and eXtensible Markup Language (XML) documents. It provides a simple and intuitive interface for parsing and manipulating the content of webpages, making it easier to extract specific data from a site. However, a basic understanding of Python is necessary to effectively utilize these capabilities. With `Beautiful Soup`, we can locate, filter, and access specific elements using Cascading Style Sheets (CSS) selectors and search methods. Once the desired data are found, they facilitate the extraction of text and attributes. After the web scraping is completed, the final data frame consists of 1166 rows and 8 columns, of which 3 are categorical (*Type*, *Neighborhood*, and *Furnished*), 4 are numerical (*Bedrooms*, *Bathrooms*, *Garages*, and *Suites*), and 1 is the output variable, represented in the column *Rent\_value*.

We implemented the ML models using the `ScikitLearn` library version 1.2.1 from Python [25], which provides a comprehensive suite of tools for data preprocessing, modeling, and evaluation. During data preparation, the dataset was randomly split into training and testing sets, with 80% allocated for training and 20% for testing. To encode categorical data into suitable numerical representations for ML algorithms, we created a customized transformation pipeline. This pipeline applied `OneHotEncoder` to categorical variables and `StandardScaler` to numerical variables, ensuring that both types of data were appropriately transformed for analysis.

For model selection, we defined commonly used regression models available in `ScikitLearn`, including `LinearRegression()`, `SVR()`, `RandomForestRegressor()`, `XGBRegressor()`, `KNeighborsRegressor()`, and `MLPRegressor()`. Each model was optimized using a grid search approach to identify the best hyperparameters. Specifically, we utilized `GridSearchCV`, which performs an exhaustive search over specified hyperparameter combinations using k-fold cross-validation (set to 5 folds in this study). The coefficient of determination ( $R^2$ ) was chosen as the scoring metric for evaluating model performance. Importantly, this grid search and cross-validation workflow was applied to the data after they had undergone the transformation pipeline, ensuring consistency and reproducibility. While `GridSearchCV` is a robust method for hyperparameter tuning, alternative approaches like randomized search or Bayesian optimization could also be employed depending on the specific requirements of the task.

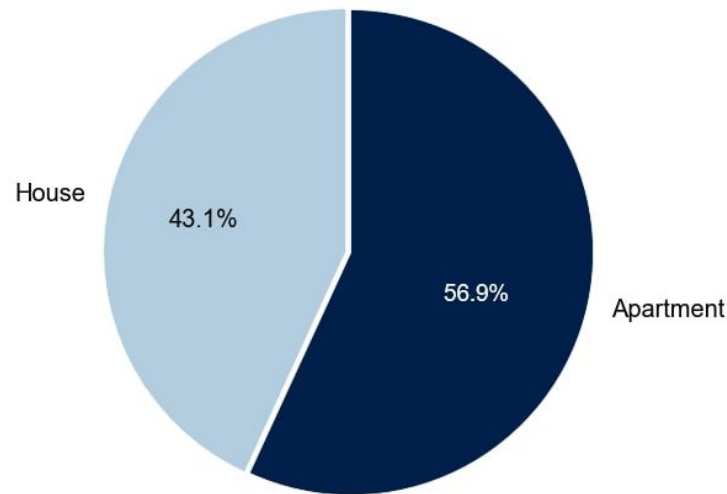
## 4. Results and Analysis

### 4.1. Exploratory Data Analysis

Exploratory data analysis plays a fundamental role in any data analysis project, regardless of the domain or the purpose of the study. It involves the initial exploration and understanding of data before conducting more advanced analyses or statistical modeling. In the current project, this analysis allows for the examination of potential patterns, trends, and relationships among the variables used and their connection to rental prices. It can be carried out using grouping functions, which enable the data to be divided and summarized based on specific categories or criteria to extract summarized information or perform specific calculations for distinct data groups. The results are plotted using the Python library `Matplotlib` [26].

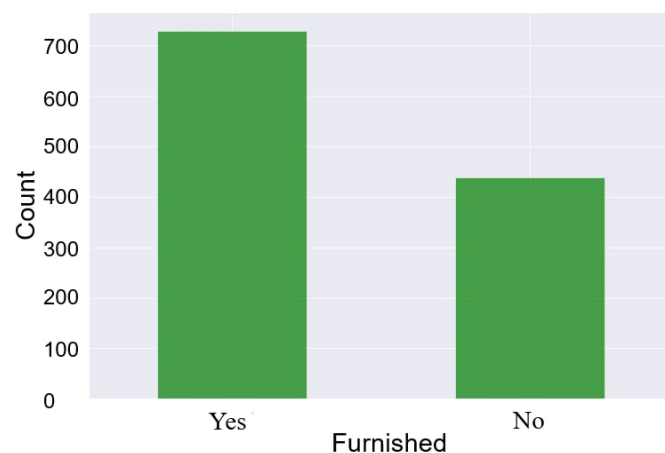
To understand the variables' individual characteristics, we began with a univariate analysis, followed by more advanced multivariate analyses to explore their interrelationships. Firstly, we examined the percentage of rental properties for each type. Figure 1 shows that the number of apartments available is greater than that of houses

(56.9% and 43.1%, respectively). This observation aligns with expectations for the city of São Carlos, considering the increased verticalization of the city [27].

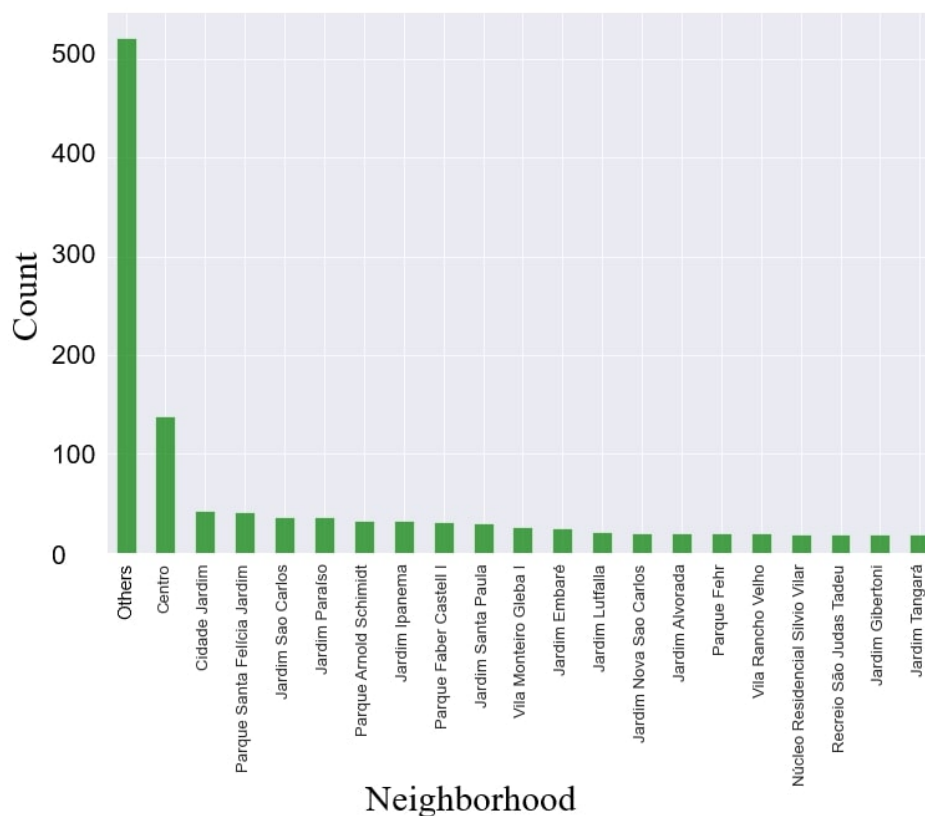


**Figure 1.** Percentage of each type of rental offered.

Regarding the other two categorical columns, namely *Furnished* and *Neighborhood*, Figure 2 illustrates a greater number of furnished properties (728). In terms of neighborhoods, Figure 3 presents the count for the top 20 neighborhoods with the highest number of properties out of a total of 151 neighborhoods. Those neighborhoods not included in the top 20 in terms of property count were grouped into the “Others” category solely for the purpose of this analysis, which was a necessary step to simplify analysis and avoid overfitting due to limited data for certain neighborhoods. While this approach may limit inference for specific areas, it ensures the model’s stability and predictive accuracy. The Centro (downtown) neighborhood has the most properties (138) for renting, indicating a higher demand for housing in central areas, possibly due to proximity to amenities, workplaces, or transportation hubs.



**Figure 2.** Count of “Yes” and “No” for the variable *Furnished*. The analysis reveals that the majority of properties (728) are categorized as “Yes” (furnished), while 438 are categorized as “No” (not furnished).

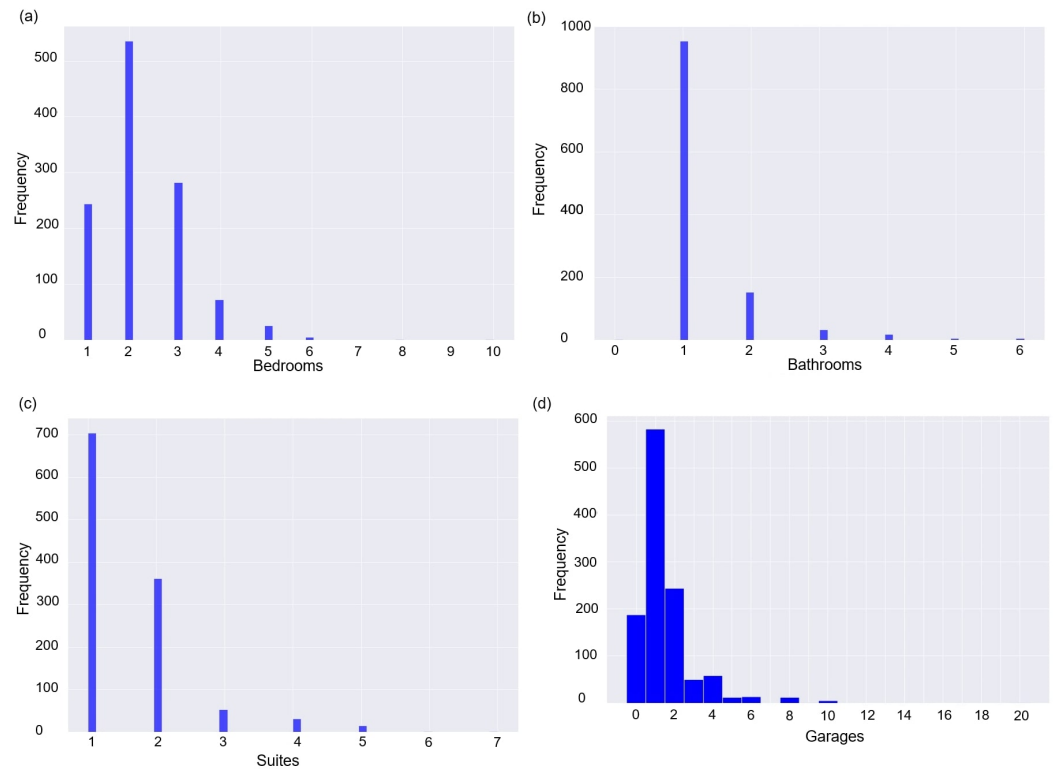


**Figure 3.** Count of variable *Neighborhood*, revealing the top 20 neighborhoods with the highest number of property offers.

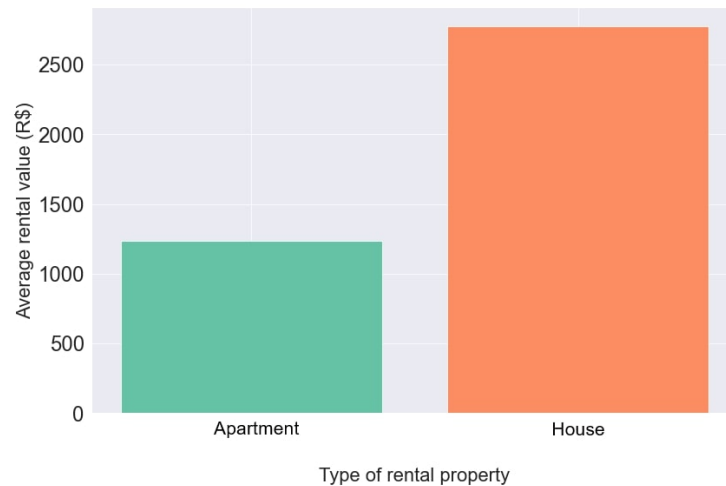
As for the numerical variables, Figure 4a presents the frequency for the *Bedrooms* variable and shows the prevalence of properties with 1, 2, and 3 bedrooms, with an average of 2.25 bedrooms per property, a minimum of 1 bedroom, and a maximum of 10. Figure 4b displays the frequency for the *Bathrooms* variable, with the majority of properties having 1 bathroom. The average is 1.26 bathrooms per property, a minimum of 0 bathrooms (studio apartments that are suites), and a maximum of 6 bathrooms. Regarding the *Suites* variable, Figure 4c shows that properties with 0 suites are predominant, with an average of 0.54 suites per property, a minimum of 0, and a maximum of 6 suites. Figure 4d illustrates the frequency for the *Garages* variable, with properties predominantly having 1 garage. The average number of garages per property is 1.52, with a minimum of 0 garages and a maximum of 20.

We now proceed to a statistical approach involving the simultaneous analysis of two inter-related variables, exploring the relationship between rental prices and predictor variables. Analyzing the average rental price versus the type of property (Figure 5), it is noted that houses have a higher value, often due to their larger total area. Additionally, the city has a significant supply of small apartments geared toward students, contributing to a lower average rental value for this type of property. Concerning the *Bedrooms* variable, it is expected that a higher number of bedrooms will correspond to a higher rental value. Figure 6a shows that this trend holds, but, for a very high number of bedrooms, such as 6 and 8, the trend is broken. However, it is observed that, in these cases, there are a few representative properties for these numbers (5 and 1 properties, respectively), which were not excluded.



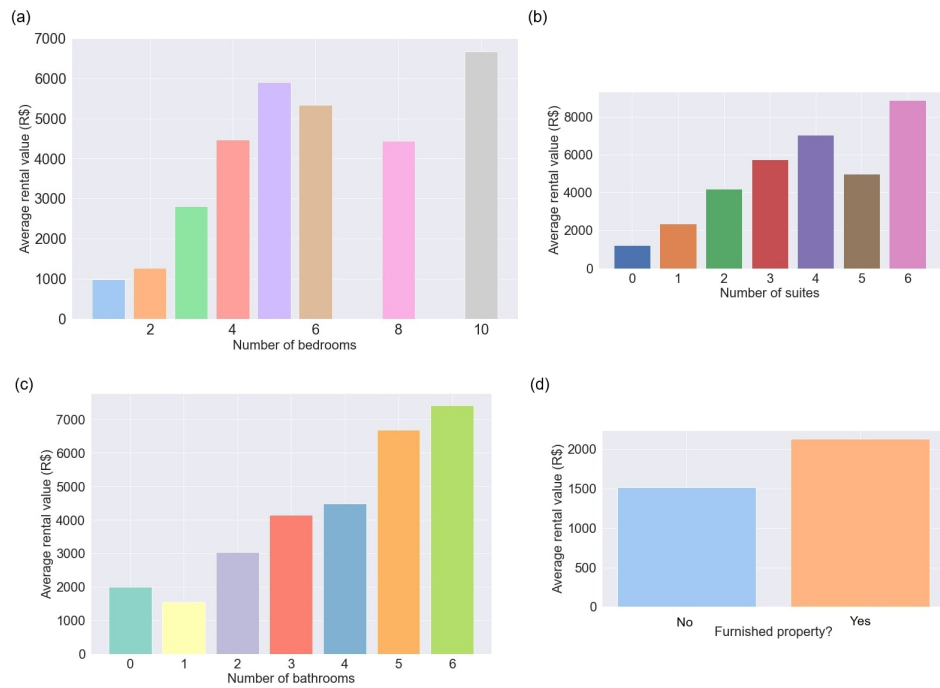


**Figure 4.** Histogram of variables: (a) *Bedrooms*, (b) *Bathrooms*, (c) *Suites*, and (d) *Garages*.



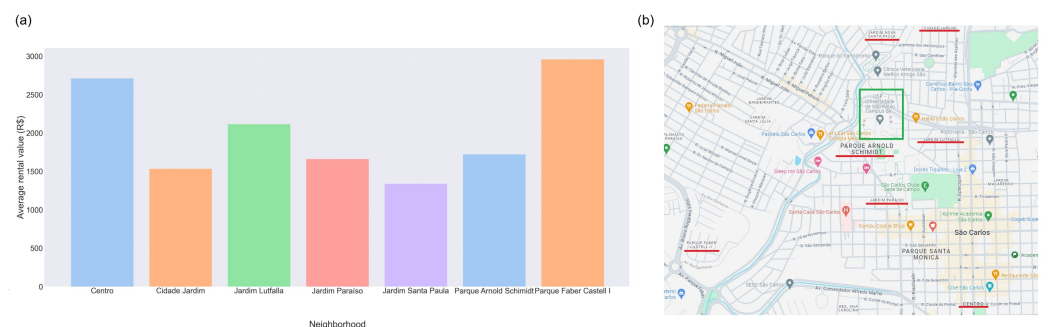
**Figure 5.** Average rental value in Brazilian reais (R\$) by property type: apartment (green) and house (orange).

For the *Suites* variable, the same positive correlation with the rental value is expected. Figure 6b confirms this trend, except for the value of 5 suites, where, again, there is only one property with this number of suites. As for the *Bathrooms* variable, the positive correlation is confirmed for all values, as shown in Figure 6c. The presence of properties with 0 bathrooms indicates that the apartment only contains a suite. Analyzing the average rental value for the *Furnished* variable, there is a difference, albeit not very large, between the value for furnished and unfurnished properties, with the former having a higher average value of BRL 2127.74, as shown in Figure 6d. This is also expected, as the inclusion of furniture enhances the property and increases the rental value.



**Figure 6.** Average rental value for the variables: (a) *Bedrooms*, (b) *Suites*, (c) *Bathrooms*, and (d) *Furnished*.

Another interesting observation that can be made is the relationship between the average rental value and the neighborhoods that stand out in the city, especially those around the USP campus and near the city center, as well as the city center itself, and new neighborhoods containing only houses, such as Parque Faber Castell [28]. Figure 7a shows how the average rental value varies across these neighborhoods. We can observe higher values in the city center (Centro neighborhood), a trend common in all cities, and in newer neighborhoods that exclusively have houses, such as Parque Faber Castell I, which exhibits the highest value since it is a well-located neighborhood with houses whose total area is larger than the average. Figure 7b shows a map centered around the USP campus (highlighted in green) and the concentration of high rental values in the neighborhoods near the campus, such as Cidade Jardim, Jardim Lutfalla, Jardim Paraíso, Jardim Santa Paula, and Parque Arnold Schimidt. In addition, all the highlighted neighborhoods are as far as up to 5 km from the USP campus.



**Figure 7.** (a) Average rental value for the variable *Neighborhood*, for neighborhoods with high counts of properties for rental and (b) their geographical distribution centered around the USP campus (green rectangle).

Now, if we turn our attention to the distribution of rental values, we can observe that the data are asymmetric and have a high variation, with values ranging from BRL

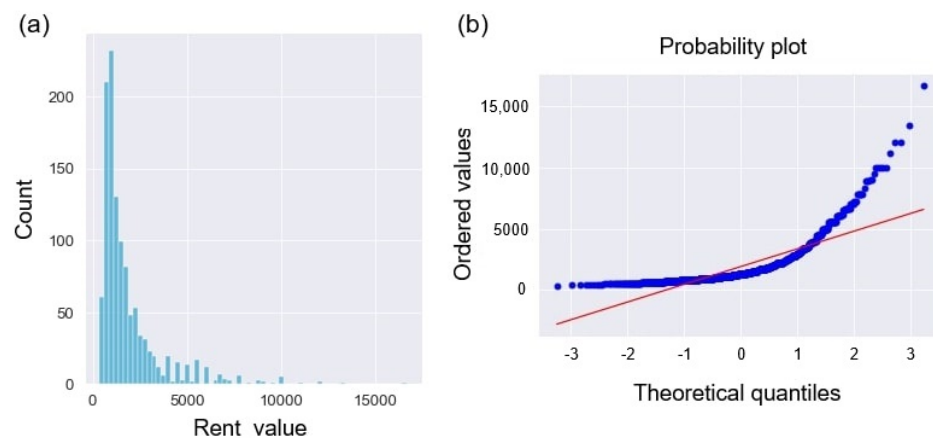


367 to BRL 16,667 (see Figure 8a). Additionally, the distribution deviates from the normal distribution, as shown in Figure 8b, where the red line represents the latter.

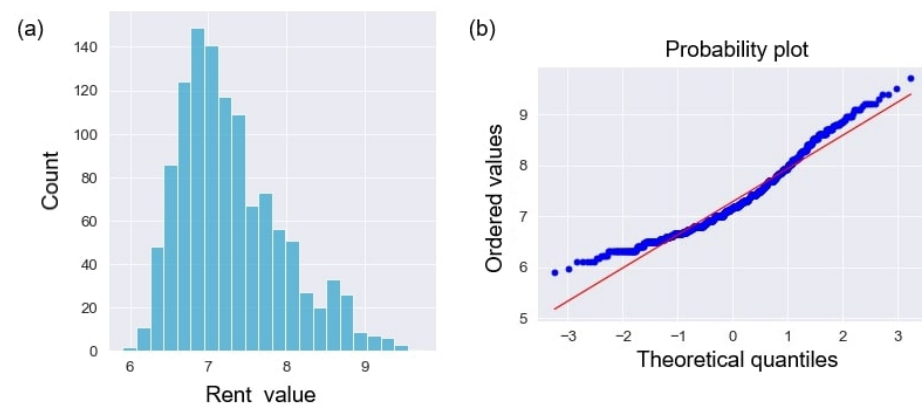
One way to reduce the range of values, decreasing the influence of extreme values, and to approach the distribution to a normal distribution is to apply a logarithmic transformation to the data, specifically the  $\log_{1p}$  transformation, given by

$$y = \log(1 + x), \quad (1)$$

where  $y$  is the transformed variable and  $x$  is the variable to be transformed. Applying this transformation to the data, we obtain Figure 9. It can be observed that the distribution is closer to normal, and the probability plot shows points closer to the red line. We will use these transformed data in the models, aiming to achieve better results.



**Figure 8.** (a) Rental value distribution and (b) probability plot, showing the asymmetry and high variation in values, indicating deviation from the normal distribution.

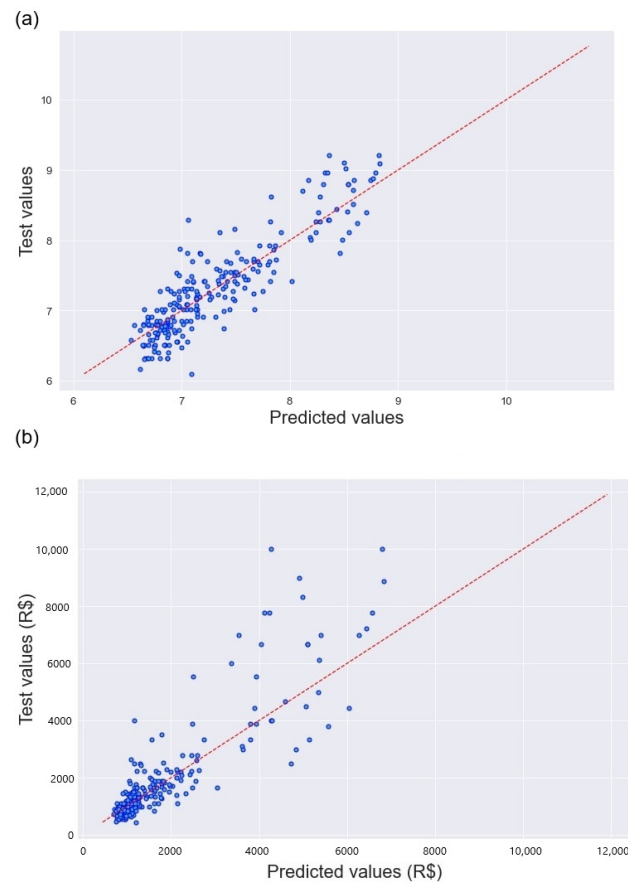


**Figure 9.** (a) Distribution of rental values transformed according to Equation (1), and (b) probability plot of the transformed data, indicating a closer approximation to the normal distribution.

#### 4.2. ML Models, Metrics, and Best Model

The XGBRegressor model achieved the best results given the metrics used, which were mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R2. These values are shown in Table 1, while Figure 10a shows the plot of predicted values versus actual values for the transformed data according to Equation (1), and Figure 10b shows the same plot but for the data with values in BRL (obtained by applying the inverse of Equation (1), which restores the original dataset). The R2 value indicates that 74% of the variability in the *Rent\_value* variable is explained by the predictor

variables, highlighting the model's effectiveness in capturing the relationships within the data.



**Figure 10.** (a) Distribution of rental values transformed according to Equation (1), and (b) test values of the original dataset *versus* the predicted values in reais (R\$).

**Table 1.** Metrics values calculated for the best model XGBRegressor, for the data transformed by Equation (1) and for the data in BRL.

Metric	Transformed Value	Value in BRL
MAE	0.24	517.77
MSE	0.10	910,429.47
RMSE	0.32	954.16
R2	0.79	0.74

XGBoost was selected due to its superior ability to handle non-linear relationships and interactions among variables, which are inherent in real estate datasets. Factors such as the number of bedrooms, proximity to amenities, and neighborhood characteristics often interact in complex ways, and these patterns are not adequately captured by simpler models like linear regression. XGBoost's gradient boosting framework allows it to iteratively improve predictive accuracy by combining weak learners and capturing intricate patterns that simpler approaches miss. Additionally, its regularization techniques help to mitigate the impact of outliers, making the model more robust in scenarios involving premium or unusually low-priced properties.

The logarithmic transformation applied to the rental value data was instrumental in addressing the skewness of the original dataset. This step normalized the distribution, stabilized the variance, and improved the model's ability to generate accurate predictions across a wide range of rental values. However, despite these improvements, the model

exhibited higher error rates for properties with exceptionally high rental values. These discrepancies are attributable to the limited representation of such properties in the dataset, as well as the inherent complexity of modeling extreme cases in real estate markets.

Despite these challenges, the calculated MAE of approximately BRL 518.00 is reasonable given the average rent value and the volatile nature of real estate markets. Economic conditions, seasonality, and localized market trends introduce variability that no model can fully eliminate. By leveraging XGBoost, we were able to provide a valuable decision-making tool that identifies critical property features that influence rental prices. While not perfectly accurate, the model's predictions offer actionable insights for tenants and landlords, enhancing their ability to make informed decisions in a dynamic and competitive market.

#### 4.3. Statistical Analysis Using OLS

The statistics provided by the ordinary least squares (OLS) method offer valuable insights into the relationships between variables in the dataset. OLS is a foundational statistical technique widely used in regression analysis, independent of any specific implementation or programming language. In this study, we utilized the `statsmodels` library (version 0.13) [29], a Python-based tool, to perform the OLS analysis. This library was chosen for its user-friendly interface and comprehensive suite of statistical features, which facilitated model fitting and result interpretation. While `statsmodels` was employed as the computational framework, the theoretical basis and statistical properties of the OLS method remain universally applicable, independent of the software used. Our analysis leverages this implementation to calculate key statistics, such as coefficients,  $p$ -values, and confidence intervals, which are crucial for understanding the significance and impact of predictor variables in the regression model.

Table 2 shows the result of the OLS method and provides a comprehensive overview of the model's performance in predicting rental values, presenting the most relevant results to analyze the performance of the method, focusing on the relative importance of the predictor variables. The model achieved an R-squared value of 0.711, which is very close to that obtained using the XGBRegressor model. The adjusted R-squared value, which accounts for the number of predictors in the model, is slightly lower at 0.666, suggesting a moderate degree of explanatory power after adjusting for potential overfitting. The F-statistic of 15.92, coupled with a highly significant  $p$ -value ( $\text{Prob} > \text{F-statistic} = 1.08 \times 10^{-186}$ ), confirms that the model is statistically significant, implying that at least one of the predictor variables is meaningfully contributing to the prediction of rental values. However, the relatively low log-likelihood value of  $-9635.1$ , along with AIC ( $1.958 \times 10^4$ ) and BIC ( $2.038 \times 10^4$ ) scores, indicates that while the model fits the data reasonably well, there may still be room for improvement, potentially through the inclusion of additional variables or the application of more sophisticated modeling techniques. Overall, these results suggest that the OLS model is a useful tool for understanding the factors influencing rental prices, but further refinement could enhance its predictive accuracy.

**Table 2.** OLS regression results.

<b>Dep. Variable:</b>	Rent_value	<b>R-squared:</b>	0.711
<b>Model:</b>	OLS	<b>Adj. R-squared:</b>	0.666
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	15.92
<b>No. Observations:</b>	1166	<b>Prob (F-statistic):</b>	$1.08 \times 10^{-186}$
<b>Df Residuals:</b>	1009	<b>Log-Likelihood:</b>	$-9635.1$
<b>Df Model:</b>	156	<b>AIC:</b>	$1.958 \times 10^4$
<b>Covariance Type:</b>	nonrobust	<b>BIC:</b>	$2.038 \times 10^4$

We can further evaluate the predictor variables by examining the OLS-estimated coefficients, as presented in Table 3 for variables such as *Furnished* (“Yes”), *Bedrooms*, *Bathrooms*, *Garages*, and *Suites*. The primary focus here is on the “ $P > |t|$ ” column, which reports the  $p$ -values, and the confidence intervals (“[0.025” and “0.975]”). The  $p$ -value indicates the statistical significance of each predictor in relation to the dependent variable, testing the null hypothesis that the coefficient of the variable is zero. In the case of the variables in Table 3, the  $p$ -values are effectively zero, underscoring their strong relevance in predicting rental values.

**Table 3.** OLS regression coefficients for the predictor variables *Furnished* (“Yes”), *Bedrooms*, *Bathrooms*, *Garages*, and *Suites*.

	coef	std err	t	P >  t	[0.025, 0.975]
Furnished[T.Sim]	258.4720	68.272	3.786	0.000	[124.501, 392.443]
Bedrooms	386.1371	44.553	8.667	0.000	[298.709, 473.565]
Bathrooms	415.9925	53.066	7.839	0.000	[311.860, 520.125]
Garages	160.3343	26.629	6.021	0.000	[108.079, 212.589]
Suites	522.3489	53.845	9.701	0.000	[416.689, 628.009]

The columns “[0.025” and “0.975]” indicate the confidence interval for the calculated coefficients. In this case, we notice that none of the intervals contain zero, confirming the importance of the variables in the model.

We also found that some neighborhoods, such as Vila Carmem and Vila Conceição (Table 4), do not contribute to price estimation as, per the OLS results, their  $p$ -values are high ( $>0.05$ ), and their confidence intervals are wide and include zero, indicating that they could be excluded from the model. However, it was found that excluding such neighborhoods did not significantly alter the metrics, and the decision was made to keep them so that, when using the website, the user has all the offered neighborhood options to choose from.

**Table 4.** OLS regression coefficients for neighborhoods.

	coef	std err	t	P >  t	[0.025, 0.975]
Neighborhood[T.Vila Carmem]	352.2406	680.464	0.518	0.605	[−983.045, 1687.527]
Neighborhood[T.Vila Celina]	1248.5893	566.906	2.202	0.028	[136.139, 2361.040]
Neighborhood[T.Vila Conceição]	162.8175	1106.468	0.147	0.883	[−2000.424, 2334.059]

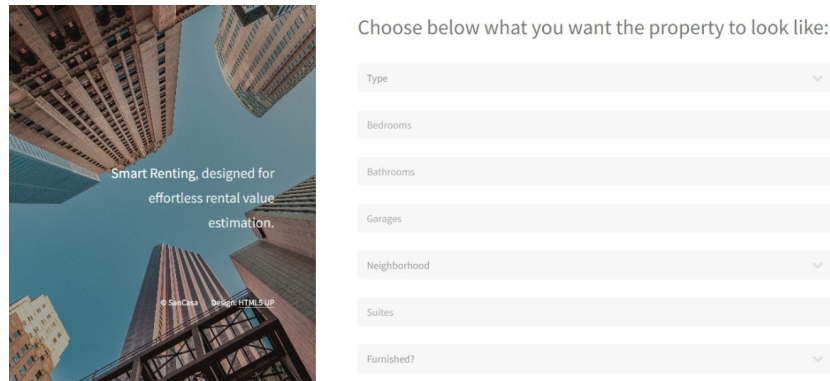
The XGBRegressor model demonstrated superior performance compared to the OLS model, especially when considering error metrics. The XGBRegressor achieved an R-squared value of 0.74, while the OLS model obtained an R-squared value of 0.711, which is slightly lower. The OLS model is also more susceptible to rigid linear assumptions, which may not capture all the nuances of complex real estate market data. Furthermore, while the OLS model is useful for identifying the statistical significance of predictor variables and providing clear statistical inferences, the XGBRegressor is more effective for practical predictions of rental values due to its ability to model nonlinear relationships and complex interactions between variables.

#### 4.4. Website Creation and Operation

In order to have a website that displays the predicted rental prices of properties, it is first necessary to build an API that facilitates the exchange of information between what the user selects on the website and what the model predicts for those variables. To create the API, we will use the Flask web framework version 2.2.3, which enables the efficient and customized development of web applications. The website code is hosted at the GitHub

webpage [https://github.com/sancasa/Smart\\_Renting](https://github.com/sancasa/Smart_Renting) (accessed on 20 January 2025) and can be run according to the instructions in the same page. A version of the website in Brazilian Portuguese is also available. Its operation is quite intuitive. The user is prompted to choose the property specifications, and the options include the following (see Figure 11):

- *Type*: choose between “House” and “Apartment”;
- *Bedrooms, Bathrooms, Garages, and Suites*: the user can input the desired number or click the up and down arrows until the desired number is reached;
- *Neighborhood*: choose among all the displayed neighborhoods;
- *Furnished*: choose between “Yes” and “No”.



Choose below what you want the property to look like:

Type

Bedrooms

Bathrooms

Garages

Neighborhood

Suites

Furnished?

**Figure 11.** Website displaying the options from which the user must select the property specifications.

All choices are mandatory, and if the user forgets or fails to fill in any of the fields, a message is displayed. In cases where a categorical variable is not filled, the message “Select an item from the list” appears, and, for a numerical variable, the message “Fill in this field” is displayed. Once the property characteristics are chosen, the user can click the “Calculate” button, and the website will provide the predicted value by the model, as shown in Figure 12. At any time, the user can choose to clear the selection by clicking the “Clear” button. The website also includes a Statistics section (see Figure 13), featuring some of the statistics shown above, which allows the user to obtain information about the average rental value based on different variables.

Choose below what you want the property to look like:

House

3

2

1

Downtown

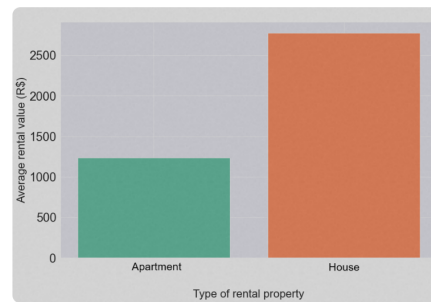
1

Yes

Calculated value: R\$3388

**Figure 12.** Website displaying the value based on the property characteristics chosen. We notice the agreement between the model prediction and the statistics, as the choice of a house with 3 bedrooms, 2 bathrooms, 1 garage, in the downtown neighborhood, with 1 suite, and furnished yields a high rent value.

## Statistics



### Average rental value x Type of property

Houses have a higher average rental value, due to the fact that, for the most part, they are properties with a larger area.



### Average rental value x Furnished properties

Furnished properties have a higher average rental price. Here, a property that has at least kitchen cabinets and/or wardrobes is considered furnished..

**Figure 13.** Statistics section of the website, where the user can access average rental values as functions of the variables.

## 5. Conclusions

The chosen ML model, XGBRegressor, is a model with a recognized performance, as shown in the “Results and Analysis” section. In the case of the data used here, the model resulted in metrics with good values, such as  $R^2 = 0.74$ , although the MAE was BRL 518.00, which is a reasonable error when informing the user about a potential rental value. Thus, an enrichment of this project would involve improving the metric values, either through a more in-depth search for the best hyperparameters of the models or by enhancing the database. The latter can be carried out in various ways, such as increasing the database by obtaining data from other real estate agencies in the city.

Beyond the technical achievements, this research offers a practical contribution through the development of a user-friendly web platform. By simplifying the rental decision-making process, the platform empowers tenants with reliable, data-driven insights and equips real estate agents with tools for transparent and informed pricing strategies. Further improvements, such as integrating interactive maps and conducting A/B testing, could optimize the platform’s usability and engagement, ensuring a seamless experience for users while addressing the needs of a competitive rental market.

This study provides a valuable methodological framework for applying ML in real estate, particularly in the underexplored rental sector. The integration of advanced predictive models and statistical analyses lays the foundation for future research, offering insights that extend beyond São Carlos-SP to other cities and regions. By bridging the gap between research and real-world application, the study advances academic discourse on predictive analytics while addressing tangible challenges in urban housing markets.

The platform proposed fosters transparency and efficiency, empowering tenants with actionable rental insights and aiding landlords in aligning their pricing strategies with market realities. As the platform scales to other urban contexts, it could serve as a transformative tool for fostering equitable rental practices and reducing information asymmetry between stakeholders.

Addressing the scope of this study, we acknowledge the limitations posed by the dataset, which focuses on a specific location and time period. Expanding the dataset to include additional locations and longer timeframes would undoubtedly enhance the generalizability and robustness of the results. However, logistical and temporal constraints have precluded such an expansion within this publication. Future research will aim to



address this limitation by incorporating broader datasets that capture more diverse urban dynamics, ensuring that the findings remain applicable across varying real estate markets.

Looking ahead, the potential for expanding this approach to other cities is vast. By tailoring ML models to the unique characteristics of local real estate markets, the platform could offer tenants a deeper understanding of rental trends, identify affordable housing options, and promote accessibility. For policymakers and urban planners, the aggregated insights from such platforms could inform housing policies and foster sustainable urban development.

This study underscores the transformative potential of integrating AI and ML into real estate analytics, bridging the divide between academic innovation and practical utility. By doing so, it paves the way for a more equitable, efficient, and transparent rental market, ultimately improving the rental experience for all stakeholders involved.

**Author Contributions:** Conceptualization, N.D.G. and P.H.F.; methodology, N.D.G. and P.H.F.; software, N.D.G., K.J.C.C.d.L. and P.H.F.; validation, N.D.G., K.J.C.C.d.L., P.H.F. and F.L.; formal analysis, N.D.G., P.H.F. and F.L.; investigation, N.D.G.; resources, P.H.F. and F.L.; data curation, N.D.G.; writing—original draft preparation, N.D.G.; writing—review and editing, N.D.G., K.J.C.C.d.L., P.H.F. and F.L.; visualization, N.D.G., K.J.C.C.d.L. and P.H.F.; supervision, P.H.F.; project administration, F.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The original data presented in the study are openly available in Smart\_Renting at [https://github.com/sancasa/Smart\\_Renting](https://github.com/sancasa/Smart_Renting) (accessed on 20 January 2025).

**Acknowledgments:** We are grateful to the professors and team that organize the MBA in Data Science offered by Instituto de Ciências Matemáticas e de Computação (ICMC) at University of São Paulo (USP) and the Centro de Ciências Matemáticas Aplicadas à Indústria (CeMEAI). We also thank the Fundação de Apoio à Física e à Química (FAFQ) for the scholarship offered.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

3D	Three Dimensions
API	Application Programming Interface
COVID-19	Corona Virus Disease 2019
CSS	Cascading Style Sheets
EUA	United States of America
GAMLSSs	Generalized Additive Models for Location, Scale and Shape
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
IBGE	Brazilian Institute of Geography and Statistics
JS	JavaScript
KNNs	K-Nearest Neighbors
MAE	Mean Absolute Error
MSE	Mean Squared Error
OLS	Ordinary Least Squares
ReLU	Rectified Linear Unit

RMSE	Root Mean Squared Error
SP	São Paulo
SVM	Support Vector Machine
SVR	Support Vector Regression
UFSCar	Federal University of São Carlos
URL	Uniform Resource Locator
USP	University of São Paulo
XGBoost	Extreme Gradient Boosting
XML	Extensible Markup Language

## References

- Nunes, J.M.; Longo, O.C.; Alcoforado, L.F.; Pinto, G.O. Análise dos impactos da covid-19 no mercado imobiliário brasileiro. *Res. Soc. Dev.* **2020**, *9*, e46891211317. [CrossRef]
- Tolentino, E. Como o Avanço Tecnológico Impacta o Mercado Imobiliário. 2023. Available online: <https://exame.com/bussola/como-o-avanco-tecnologico-impacta-o-mercado-imobiliario/> (accessed on 20 January 2025).
- Barr, J.R.; Ellis, E.A.; Kassab, A.; Redfearn, C.L.; Srinivasan, N.N.; Voris, K.B. Home price index: A machine learning methodology. *Int. J. Semant. Comput.* **2017**, *11*, 111–133. [CrossRef]
- Huang, Y. Predicting home value in California, United States via machine learning modeling. *Stat. Optim. Inf. Comput.* **2019**, *7*, 66–74. [CrossRef]
- Oike, V. Tendências do Mercado Imobiliário: O Que a Economia Nos Aponta? 2022. Available online: <https://conteudos.quintoandar.com.br/tendencias-mercado-imobiliario-2023/> (accessed on 20 January 2025).
- Geltner, D.; MacGregor, B.D.; Schwann, G.M. Appraisal smoothing and price discovery in real estate markets. *Urban Stud.* **2003**, *40*, 1047–1064. [CrossRef]
- Ghysels, E.; Plazzi, A.; Valkanov, R.; Torous, W. Forecasting real estate prices. *Handb. Econ. Forecast.* **2013**, *2*, 509–580.
- Theurillat, T.; Rérat, P.; Crevoisier, O. The real estate markets: Players, institutions and territories. *Urban Stud.* **2015**, *52*, 1414–1433. [CrossRef]
- Salzman, D.; Zwinkels, R.C. Behavioral real estate. *J. Real Estate Lit.* **2017**, *25*, 77–106. [CrossRef]
- Rolnik, R.; Pereira, A.L.d.S.; Moreira, F.A.; Royer, L.d.O.; Iacovini, R.F.G.; Nisida, V.C. O Programa Minha Casa Minha Vida nas regiões metropolitanas de São Paulo e Campinas: Aspectos socioespaciais e segregação. *Cad. Metrópole* **2015**, *17*, 127–154. [CrossRef]
- Matos, T. A Expansão do Mercado Imobiliário no Brasil: Um Paralelo Entre a Evolução dos Preços no Mercado Brasileiro e a Bolha Imobiliária Norte-americana. Dissertation. Pontifera Univ. Católica Do Rio De Janeiro, Dep. De Ciências Econômicas, Rio De Janeiro. v.24. 2017. Available online: [https://www.econ.puc-rio.br/uploads/adm/trabalhos/files/Thiago\\_Oliveira\\_Rio\\_Tinto\\_de\\_Matos.pdf](https://www.econ.puc-rio.br/uploads/adm/trabalhos/files/Thiago_Oliveira_Rio_Tinto_de_Matos.pdf) (accessed on 20 January 2025).
- Chau, K.W.; Chin, T. A critical review of literature on the hedonic price model. *Int. J. Hous. Sci. Its Appl.* **2003**, *27*, 145–165.
- Herath, S.; Maier, G. *The Hedonic Price Method in Real Estate and Housing Market Research: A Review of the Literature*; Institute for Regional Development and Environment, University of Economics and Business: Vienna, Austria, 2010; pp. 1–21.
- Valier, A. Who performs better? AVMs vs hedonic models. *J. Prop. Invest. Financ.* **2020**, *38*, 213–225. [CrossRef]
- Čeh, M.; Kilibarda, M.; Liseč, A.; Bajat, B. Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 168. [CrossRef]
- Fan, G.Z.; Ong, S.E.; Koh, H.C. Determinants of house price: A decision tree approach. *Urban Stud.* **2006**, *43*, 2301–2315. [CrossRef]
- Zulkifley, N.H.; Rahman, S.A.; Ubaidullah, N.H.; Ibrahim, I. House Price Prediction using a Machine Learning Model: A Survey of Literature. *Int. J. Mod. Educ. Comput. Sci.* **2020**, *12*, 46–54. [CrossRef]
- Estevam, A.C. *Modelagem Estatística Para Análise de Dados Imobiliários Completos e Com Censura à Esquerda*; Universidade Federal de São Carlos: São Carlos, Brazil, 2014.
- Ferreira, J.F.; Ferraudo, G.M.; Louzada-Neto, F. Inferência do Valor de Mercado de Lotes Urbanos. Estudo de Caso: Município de São Carlos, São Paulo, Brasil. Technical Report, Latin American Real Estate Society (LARES). 2008. Available online: [https://ideas.repec.org/p/lre/wpaper/lares\\_2008\\_artigo003-ferreira.html](https://ideas.repec.org/p/lre/wpaper/lares_2008_artigo003-ferreira.html) (accessed on 20 January 2025).
- Geraldo, G.P. *O Programa Minha Casa Minha Vida, o Mercado Imobiliário e o Direito à Cidade: Análise dos Impactos do Programa na Cidade de São Carlos-SP*; Universidade Estadual Paulista, Instituto de Geociências e Ciências Exatas: Rio Claro, Brazil, 2014.
- Petrucelli Neto, G. *A Influência das Universidades Públicas de São Carlos-SP na Dinâmica do Espaço Geográfico: A Importância da População Estudantil Para o Setor Imobiliário*; Universidade Estadual Paulista (Unesp): Rio Claro, Brazil, 2011.
- Cardinali. 2023. Available online: <https://www.cardinali.com.br/> (accessed on 20 January 2025).

23. Chandra, R.V.; Varanasi, B.S. *Python Requests Essentials*; Packt Publishing Ltd.: Birmingham, UK, 2015.
24. Richardson, L. Beautiful soup Documentation. 2007. Available online: <https://readthedocs.org/projects/beautiful-soup-4/downloads/pdf/latest/> (accessed on 20 January 2025).
25. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
26. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [[CrossRef](#)]
27. de Oliveira Ferreira, D.A. A produção e o consumo da habitação verticalizada em São Carlos–SP. *Geo UERJ* **2009**, *2*, 88–107.
28. Growww. Os Melhores Bairros Para Morar em São Carlos. 2020. Available online: <https://www.blog.cardinali.com.br/2020/12/12/os-melhores-bairros-para-morar-em-sao-carlos/> (accessed on 20 January 2025).
29. Seabold, S.; Perktold, J. statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.