Justus-Liebig Universität Gießen
Fachbereich Wirtschaftswissenschaften
Professur für Datenökonomie (VWL X)
Prof. Dr. Mirjam Stockburger

Housing Prices in Spain

A Regression Analysis

Term Paper

presented by

Lenon Ferreira
Mnr. 8105069
lenon.ferreira@wirtschaft.uni-giessen.de

Master of Science

Advisor:
Henrike Alm

Gießen, 23.06.2023

Contents

List of Figures III

List of Tables IV

1. Introduction 1

2. Data Description 2

2.1. Data Acquisition and their Sources 2

2.2. Data Encoding 3

3. Data Analysis 4

3.1. A Study of the Distribution of Prices 4

3.2. Regression Analysis 5

4. Conclusion 6

A. Appendix 8

A.1. Figures 8

A.2. Tables 10

List of Figures

1. Fotocasa.es 8
2. Kernel Distribution of Prices 8
3. Map of Median Prices by Autonomous Region 9
4. Box Plot and Bar Chart: Data Distribution across Regions. 9

List of Tables

1. Median Price per Autonomous Region 10
2. Descriptive Statistics 11
3. Linear Regression Variables 12
4. Linear Regression - Results 13

1. Introduction

The acquisition of houses as a means of portfolio diversification and protection against purchasing power deterioration has been extensively analyzed in various studies. As early as 1987, David Hartzell, Hekman, and Miles examined a portfolio comprising approximately 300 properties of diverse types from a Real Estate Investment Trust and concluded that its return serves as a robust defense against both expected and unexpected inflationary forces. Similar findings were reached by Salisu, Raheem, and Ndako 2020 when analyzing the hedging properties against inflation of stocks, gold, and real estate in the American market. However, the portfolios examined in their study were built by entities with economically well-trained agents and significant monetary liquidity.

During inflationary periods, one of the common measures implemented by central banks is raising interest rates, which makes housing less accessible for the average buyer who often requires third-party financing. This underscores the importance of conducting a study that elucidates the determining factors of property prices.

The objective of the present study is to identify the key factors influencing housing prices, where an statistical modeling approach is applied by estimating a function $y = f(X) + \varepsilon$, given covariates (X) that potentially relate to the price of the property (y). Although the data specifically focused on Spain, the cultural and economic similarities between Spain and other European countries suggest that the results obtained in this study may be generalized to other European nations.

Several variables are candidates to explain housing prices. Some obvious variables are inherently linked to the property's characteristics, such as its type, number of rooms, and size. Less obvious variables, are those associated with the property's surroundings, such as the income level in the municipality and its population density, as will be demonstrated at a later stage.

The following sections provide a detailed description of the data used in this study, including the acquisition method and sources, as well as the encoding applied to the data/variables (Section 2). Subsequently, an exploratory analysis of price distributions based on autonomous communities of Spain is presented and the results of the regression analysis are discussed (Section 3). Finally, Section 4 concludes and summarizes the findings. Among the 23 variables analyzed, 14 exhibited high explanatory power

on price formation.

2. Data Description

2.1. Data Acquisition and their Sources

Before starting the quantitative analysis, it is necessary to provide an explanation about the data used and their sources. Typically, real estate agency websites provide detailed information about their portfolio. Fotocasa is one of the largest real estate agencies in Spain and it offers around 1.5 million objects of various types, of which initially 13,015 were used in this study. The data was retrieved by means of a webscraper written in Python. Due to the large number of cities and properties a list of 1313 cities provided by wikipedia was used. From this list 329 cities were randomly selected, and the probability of city "i" appearing in the sample was computed using the percentage of the population of that city compared to the sum of the populations of all cities in the list. In other words, the probability of metropolitan cities appearing was higher than that of small cities, but as Spain has many more small town than large cities, the sample should be balanced. However, due to complications during the web scraping process, such as the IP address being blocked by fotocasa, data could only be obtained for 172 cities. These cities were distributed across 15 autonomous communities and 40 provinces. Fig. 1 illustrates the attributes that were retrieved from the website.

To take into account local factors that may have an effect on price formation, data on absolute crime rates and the level of unemployment at a province level, as well as the average income in the city where the object is located, were obtained from the National Institute of Statistics (INE) of Spain. Since larger cities usually have higher levels of crime, the absolute crime rates were converted to crime rates per 100,000 inhabitants to make the values comparable. Another important characteristic for determining property values is the level of demand in the location where it is situated. Wikipedia provides a table with information on the surface area in square kilometers of around 5,000 Spanish cities. As the population numbers were among the information available alongside the list of 1313 cities, they were used to calculate population density per square kilometer. For those cities not found in the list of 5000 cities, a small scraper was written in R to obtain the surface area information from the specific

city's Wikipedia page.

One of the most challenging tasks in the process was merging the five datasets, as the only common factor between them was the names of their cities or provinces. However, since Spain has five distinct languages, different names for the same city were often used in the different datasets. Thus, a simple string matching was not possible. To overcome this problem, the minimum string distance between, say, city "i" in dataset 1 and cities in dataset 2 was computed. However, the function used to calculate the minimum distance always returns a value, so visual inspection had to be applied to ensure accuracy.

2.2. Data Encoding

This section aims to clarify the transformations and encodings applied to the data used in the regression analysis. The variables pollution and energy consumption were combined into a single variable called "energy_class." In Spain, a classification system ranging from A to G is used, with A representing the best classification and G indicating the worst. According to the law, all objects must provide this information, along with the estimated annual energy usage in kWh/m² and the amount of CO_2 pollution. To convert these classifications into numerical values, the following method was employed:

- Firstly, each class was assigned a number from 1 to 7, denoted as $i = [(i|A) = 1, \ldots, (i|G) = 7]$. Since this classification applies to both pollution and energy consumption variables, the score was calculated using a weighted average as follows:

$$s_i = \frac{i_{energy}}{i_{energy} + j_{polluiton}} * \left(kWhm^2/año\right) + \frac{j_{pollution}}{i_{energy} + j_{polluiton}} * \left(kgCO_2m^2/año\right) \quad (1)$$

  For example, an object with the energy classification $C999kWhm^2/year$ and the pollution classification $G999kgCO_2m^2/year$ would have a score equal to $s_i = \frac{3}{3+7} * 999 + \frac{7}{3+7} * 999 = 999$.

- Subsequently, all scores were divided into seven intervals based on the score-distribution. Each interval's upper limit corresponds to the quantile: $\frac{100\%}{7} *$

interval. Units whose values fall in the first interval are assigned the value 0, being this the best classification. The classification ranges from 0 to 6.

Due to the absence of energy consumption information for 165 units, they were excluded from the dataset, resulting in a reduction from $13,015$ to $12,850$ observations. Regarding the availability of parking and lifts in the objects, as provided by fotocasa, these variables had a particularity: they were either marked as yes or left blank. Thus, when information is not provided, the values were replaced with 'no', which was subsequently encoded with 0 for no and 1 for yes. Moreover, given the significant number of missing data, the variables pertaining to object orientation, heating system, furniture, and hot water were excluded.

The variable 'floor,' indicating the object's location floor, was enumerated according to the object's floor level. It starts from -3 for objects below the basement, -2 for the basement, -1 for the mezzanine, 0 for the main floor, 1 for the first floor, and goes up to 15 for the $15th$ floor. Units located above the $15th$ floor are denoted as 16. The same coding system was applied to the variables age and condition. Regarding the former, fotocasa.es provides information in intervals, such that the variable was coded as $age_i = 0$ if the object is less than 5 years old, up to $age_i = 7$ if the object is over 100 years old. The coding for condition is 0 for poor, 1 for good, and 2 for very good condition or almost new.

The object type was encoded using dummy variables since it is strictly categorical. A binary variable was created for each type of object, with a value of 1 indicating that the object belongs to that type and 0 indicating otherwise. The types refer to whether the object is an apartment, a house, a loft, etc.

3. Data Analysis

3.1. A Study of the Distribution of Prices

When it comes to the distribution housing prices, the dataset exhibits a wide range of variations. The lowest price recorded was €$10,000$, which corresponded to two different properties—a house and an apartment located in Villanueva de Córdoba and Barcelona, respectively. On the other end, the highest price reached an astounding €$12,500,000$, involving a house situated in Barcelona.

Considering the dataset comprising 12,850 units, the average price amounted to €360, 380, while the median stood at €239, 000. Notably, the median value is lower than the mean, indicating a "right-skewed" distribution . This observation is further supported by a visual inspection of Fig. 2. Additionally, it is evident that the sample suffers from a significant outlier problem, which should be taken into account when analyzing the data (see Fig. 4).

Fig. 3 illustrates a map of Spain, divided by its autonomous communities. It is interesting to observe that coastal regions (03, 06, 16, 15, 10, 05) and densely populated areas like Madrid (13) and Catalonia (09) generally exhibit housing prices above the national median. Conversely, the interior regions of the country, which are less populated and developed, tend to have lower prices. However, it is crucial to be cautious when drawing general conclusions solely based on this figure. Some regions, such as Asturias, Navarra, Melilla, La Mancha, Cantabria, and Extremadura, are noticeably underrepresented, as indicated by Tab. 1 and Fig. 4. For the other regions, the number of observations is sufficiently large, so their median values can be considered a good approximation of their true population values.

3.2. Regression Analysis

To estimate the linear regression model, 23 covariates plus an intercept were included. To address the issue of outliers, both continuous and some discrete variables, such as the number of bathrooms and rooms, were transformed into logarithmic form. Housing units with prices below the 5th percentile and above the 95th percentile were also excluded from the sample. In total, the model was estimated using 2,847 observations.

As shown in Tab. 2, the average object in the sample is an apartment with a cost of €343, 483. It has 3 rooms, 2 bathrooms, covers an area of 107 square meters, is located on the third floor, is in relatively good condition, and is around 50 years old. It does not have a parking-space but has lift. Its energy efficiency is low. It is located in a city with a population of approximately 142, 475.15 inhabitants and has a high population density per square kilometer. The average salary in the city is €14, 825.75. The unemployment rate in the province is 11.5%, and the crime rate per 100, 000 inhabitants is 5658.161.

Tab. 3 contains the included variables along with an explanation of what they represent. The estimated results are found in Tab. 4. As observed, all variables directly

related to the object's characteristics, except for its type, are highly significant and different from zero. For example, the owner of an object with 2 bathrooms can expect a 4.2% higher price than a owner of an object with only 1 bathroom. A 1% increase in the size of the property leads to a 0.51% increase in its value. Having access to parking-space raises the average price by 7.61%, and the presence of a lift results in an appreciation of 20.95% in the value of the object. As expected, units with lower energy efficiency have a decrease in price of 1.3% for each worsening of their energy class.

Regarding the object's surroundings, a 1% increase in the city's population density raises the object's price by 0.06%. More populous cities also have higher prices, as a 1% increase in population corresponds to a 0.081% increase in price. A renovated apartment is worth 5.9% more than one in poor condition.

One variable that may initially cause surprise is the logarithmic value of the number of rooms variable. This indicates that a 1% increase in the number of rooms leads to a decrease in price of −0.83%. However, this result is reasonable since the effect of one variable on price assumes that all other variables remain constant, and dividing an area into smaller subareas does not necessarily result in an increase in the object's value. This fact is supported by the interaction term between the number of rooms and the area of the apartment, indicating that an increase in the number of rooms along with an increase in size leads to an increase in price.

One surprising finding is the positive correlation between prices and crime in the province where the object is located, which is likely due to some unmodelled latent variable. Nevertheless the model could explain approximately 67% of the price variation, which can be considered relatively good given its simplicity.

4. Conclusion

In summary, this article has shown that housing prices in Spain vary significantly, with coastal and densely populated regions having higher prices than inland regions, where the population and country's development are lower. Based on the regression analysis, the model indicates that certain characteristics inherently linked to the property, such as the number of bathrooms, size, floor level, and the availability of parking-space and elevators, have a positive effect on its price. The number of rooms only has a positive

effect when it increases along with the size of the property. As expected, the energy efficiency of the object has a negative correlation with its value. However, the type of property seems to have no relation to its price, possibly because apartments make up the majority of the data, and no interaction could be found between the condition and age variables. Nevertheless, separately, they contribute to value appreciation.

An important variable that couldn't be modeled due to lack of data is the expected average rent for each type of unit. Financial theory suggests that the value of a property is the present value of its future cash flows; therefore, once controlled for this variable, all else would become irrelevant for price formation. Such an analysis is an option for future studies.

A. Appendix

A.1. Figures

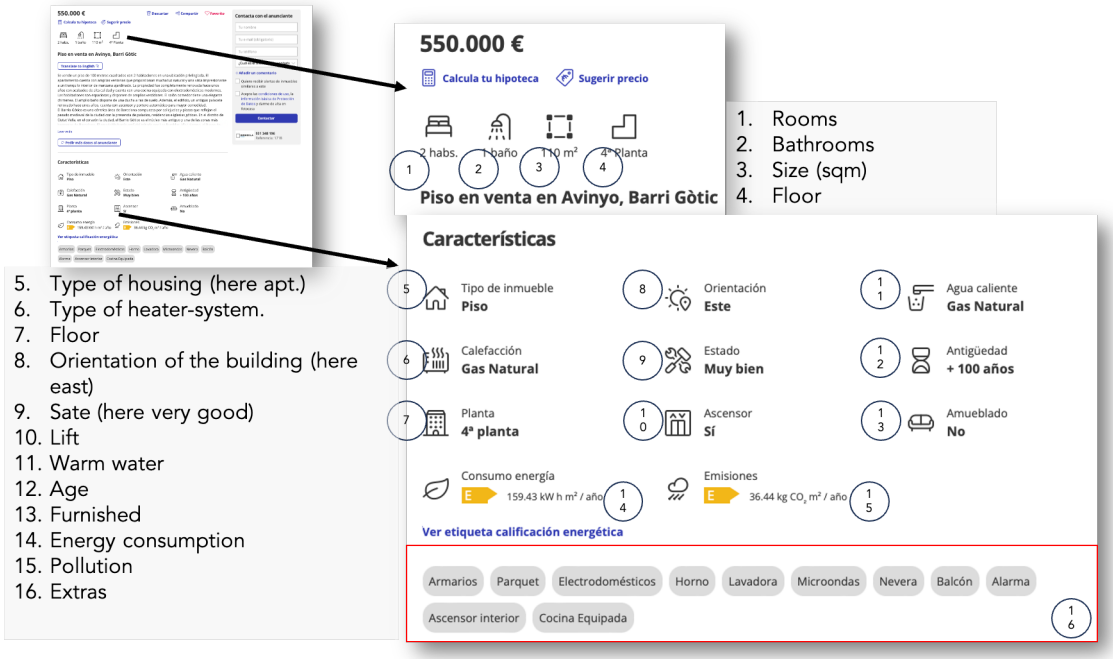

Figure 1: Fotocasa.es

Picture illustrates the attributes shown for one of the objects of Fotocasa.es. Information about the features from 1 to 15 were collected (when available). Because of the large inconsistency in the Extra items (16), those attributes were not collected.

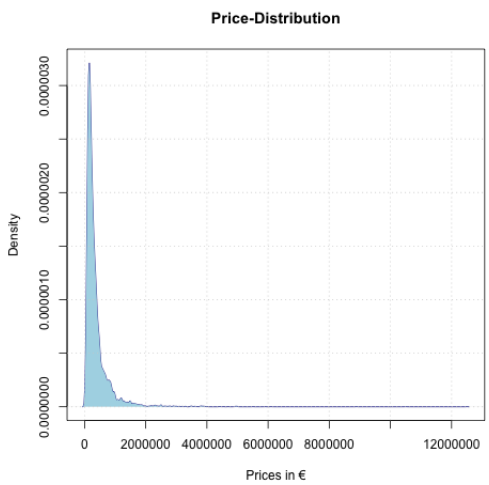

Figure 2: Kernel Distribution of Prices

Picture illustrate the a kernel-estimation of the price distribution using all data-points. Notice that the distribution resembles a χ^2 rather than a normal, since prices are always positive and theoretically with no right bounds.

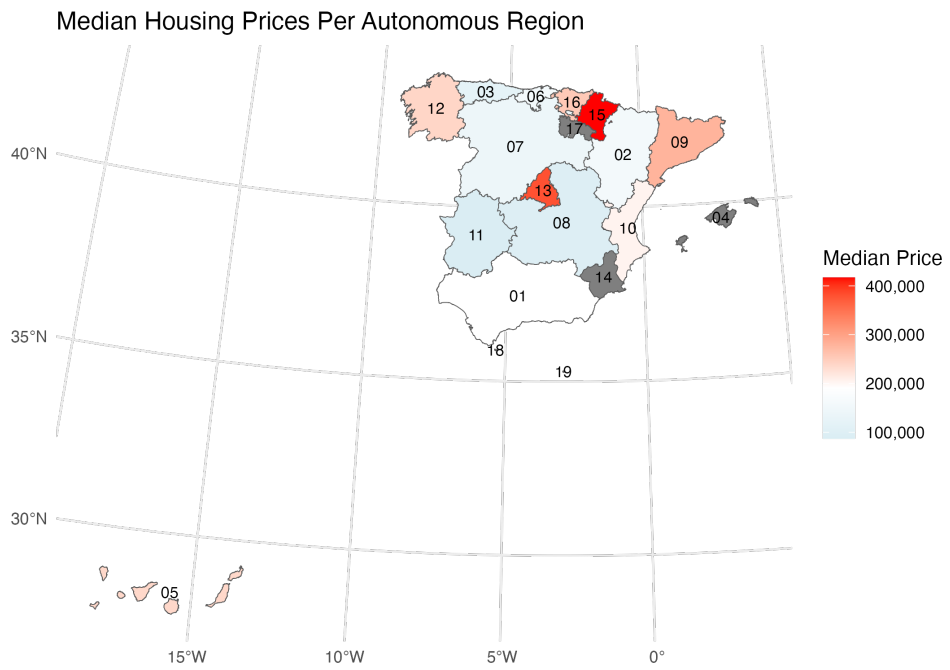

Figure 3: Map of Median Prices by Autonomous Region

Figure illustrates the median price of housing across the autonomous communities in Spain. Red represents values above the national median and blue those below. The codes of the regions are related to tab. 1

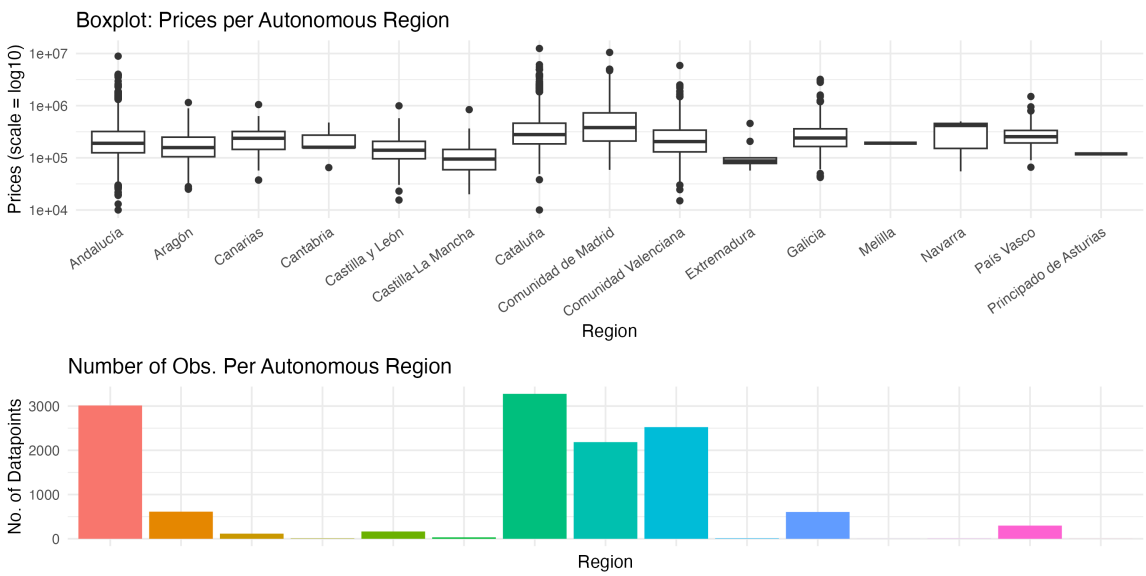

Figure 4: Box Plot and Bar Chart: Data Distribution across Regions.

The upper picture illustrates a boxplot of prices per Autonomous region of spain. Notice that for some regions, there are very few observations and that outliers are present for all regions. The bottom graph displays the number of observations per region.

A.2. Tables

Code	Provinces	Median Price in €	Rank-Median	No. Obs.
15	Comunidad Foral de Navarra	417000	1	3
13	Comunidad de Madrid	380000	2	2185
09	Cataluña	279000	3	3276
16	País Vasco	254500	4	298
12	Galicia	240000	5	606
05	Canarias	237000	6	116
10	Comunidad Valenciana	205000	7	2524
19	Melilla	191000	8	1
01	Andalucía	190000	9	3013
06	Cantabria	159500	10	8
02	Aragón	157172	11	613
07	Castilla y León	139900	12	165
03	Principado de Asturias	119000	13	1
08	Castilla - La Mancha	94750	14	32
11	Extremadura	87000	15	9
04	Islas Baleares	NA	NA	NA
14	Región de Murcia	NA	NA	NA
17	La Rioja	NA	NA	NA
18	Ceuta	NA	NA	NA

Table 1: Median Price per Autonomous Region

Table shows the median price per autonomous region in Spain. The regions are displayed in ascendant order according to their Rank-Median. The code provided in the first column are the same given in the map in Fig. 3

Variable	Min	Max	1st Qu	Median	Mean	3rd Qu
price	102000.00	1100000.00	185000.00	270000.00	343482.667	420000.00
room	1.00	8.00	2.00	3.00	2.980	4.00
bathroom	1.00	21.00	1.00	2.00	1.655	2.00
size_sqm	21.00	9618.00	72.00	92.00	107.282	118.00
floor	-3.00	16.00	1.00	2.00	2.962	4.00
condition	0.00	2.00	0.00	1.00	0.844	2.00
age	0.00	7.00	4.00	5.00	4.338	5.00
parking	0.00	1.00	0.00	0.00	0.262	1.00
lift	0.00	1.00	1.00	1.00	0.777	1.00
energy_class	0.00	6.00	2.00	5.00	3.779	5.00
density_skm	15.94	21406.80	4816.71	5688.69	9767.566	16508.86
mean_salary	9246.00	17059.00	12490.00	16750.00	14825.575	16750.00
unemp	0.06	0.26	0.10	0.11	0.115	0.12
crime_province	3200.89	6412.01	5339.74	5904.19	5658.161	6412.01
pop	7682.00	3280782.00	319515.00	1636193.00	1424755.152	1636193.00
sapt	0.00	1.00	0.00	0.00	0.019	0.00
apt	0.00	1.00	1.00	1.00	0.876	1.00
duplex	0.00	1.00	0.00	0.00	0.026	0.00
penth	0.00	1.00	0.00	0.00	0.056	0.00
rhouse	0.00	1.00	0.00	0.00	0.003	0.00
house	0.00	1.00	0.00	0.00	0.015	0.00
loft	0.00	1.00	0.00	0.00	0.003	0.00
ranch	0.00	1.00	0.00	0.00	0.002	0.00

Table 2: Descriptive Statistics

Table contains descriptive statistics about the variables used in the regression analysis before they were transformed to log-values.

i	x_i	Description
0	1	constant
1	lroom	log no. of rooms
2	lbathroom	log no. of bathrooms
3	lsize_sqm	log built-size of property in m^2
4	floor	floor where object is located
5	condition	condition of the object [poor(0), good(1), very good(2)]
6	age	age class of the object [0:<5 years to 7>100 years]
7	parking	parking space [1 if true]
8	lift	has lift [1 if true]
9	energy_class	energy class [0: best to 6:worst]
10	ldensity_skm	log pop. density per km^2
11	lmean_salary	log city mean salary
12	lunemp	log province unemployment rate
13	lcrime_province	log province criminality rate per 100,000 pop.
14	lpop	log city population
15	sapt	1 if small apartment
16	apt	1 if apartment
17	duplex	1 if duplex
18	penth	1 if penthouse
19	rhouse	1 if row-house
20	house	1 if house
21	loft	1 if loft
22	lroom * lsize_sqm	interaction between lsize and lroom
23	age * condition	interaction between age and cond. classes

Table 3: Linear Regression Variables

Table contains information about the covariates used to estimate the regression coefficients, where the dependent variable is the log-value of price.

	Coeff.	Std. Error	t-value	Pr(>|t|))
(Intercept)	3.8923***	0.6221	6.26	0.0000
lroom	-0.8343***	0.1435	-5.82	0.0000
lbathroom	0.4214***	0.0216	19.50	0.0000
lsize_sqm	0.5071***	0.0401	12.66	0.0000
floor	0.0106***	0.0027	3.87	0.0001
parking	0.0761***	0.0173	4.41	0.0000
lift	0.2095***	0.0171	12.24	0.0000
energy_class	-0.0130***	0.0039	-3.33	0.0009
ldensity_skm	0.0571***	0.0115	4.97	0.0000
lmean_salary	0.0000***	0.0000	6.21	0.0000
lunemp	0.0060	0.0494	0.12	0.9037
lcrime_province	0.4254***	0.0775	5.49	0.0000
sapt	0.1357	0.1593	0.85	0.3941
duplex	-0.0159	0.1566	-0.10	0.9189
penth	0.1440	0.1544	0.93	0.3512
rhouse	0.0144	0.1865	0.08	0.9385
apt	0.0032	0.1522	0.02	0.9833
house	-0.0351	0.1579	-0.22	0.8242
loft	-0.1997	0.1883	-1.06	0.2890
lpop	0.0806***	0.0167	4.82	0.0000
age	0.0338***	0.0071	4.73	0.0000
condition	0.0596***	0.0206	2.89	0.0038
lroom:lsize_sqm	0.1481***	0.0325	4.56	0.0000
age:condition	-0.0037	0.0045	-0.84	0.4024
Dependent Variable:	lprice	Adjusted R^2 :	0.668	
Notes:	***Significant at the 1 percent level			
	**Significant at the 5 percent level			
	*Significant at the 1 percent level			

Table 4: Linear Regression - Results

Literature

Hartzell, David, John S. Hekman, and Mike E. Miles (1987), Real Estate Returns and Inflation, *Real Estate Economics* 15.1, 617–637.

Salisu, Afees A., Ibrahim D. Raheem, and Umar B. Ndako (2020), The inflation hedging properties of gold, stocks and real estate: A comparative analysis, *Resources Policy* 66, 101605.

Eigenständigkeitserklärung

Ich versichere, dass ich die vorliegende Bachelor Thesis selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen.

Gießen, den 23.06.2023