https://doi.org/10.26219/heal.aueb.6011

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS POSTGRADUATE PROGRAM

Real Estate Analytics

By
Tseperis Spiridon

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfillment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
June 2024

https://doi.org/10.26219/heal.aueb.6011

https://doi.org/10.26219/heal.aueb.6011

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ
ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

Real Estate Analytics

ΤΣΕΠΕΡΗΣ ΣΠΥΡΙΔΩΝ

ΔΙΑΤΡΙΒΗ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Διπλώματος Μεταπτυχιακών Σπουδών στη Στατιστική

Αθήνα, Ελλάδα
Ιούνιος 2024

https://doi.org/10.26219/heal.aueb.6011

ABSTRACT

In this project, we analyze a dataset for a commercial website that contains properties located in Greece and are up for sale. One of the goals is to make predictive models for the asking price of properties given their characteristics, and especially how to use the location of the properties to get better predictions. The type of models we used are Standard Linear Regression, LASSO Regression, LAD Regression, and Random Forests. Also, in this project, we analyze the negotiation margin, and how it is affected by the location and other property characteristics. To understand how property characteristics influence the negotiation margin, we construct an explanatory linear regression model and interpret the beta coefficients. We do this for the whole country and then we specifically choose to focus on the three prefectures with the most properties, where do the same. Also, we provide heatmaps in order to find locations where the negotiation margin is larger than usual. Of course, before we do all the above, we present a thorough explanatory analysis of the dataset where we point out interesting facts and bizarre behaviors of the dataset.

https://doi.org/10.26219/heal.aueb.6011

 https://doi.org/10.26219/heal.aueb.6011

ΠΕΡΙΛΗΨΗ

Σε αυτή την εργασία θα αναλύουμε ένα σύνολο δεδομένων, που προέρχεται από έναν δημόσιο ιστότοπο, το οποίο περιλαμβάνει ιδιοκτησίες που βρίσκονται στην Ελλάδα και είναι προς πώληση. Ένας από τους στόχους μας είναι η δημιουργία προβλεπτικών μοντέλων για την ζητούμενη τιμή των ιδιοκτησιών, έχοντας ως δεδομένα τα χαρακτηριστικά τους. Επίσης εξετάζουμε πώς μπορούμε να εκμεταλλευτούμε την τοποθεσία για να βελτιώσουμε την πρόβλεψή μας. Οι τύποι μοντέλων που χρησιμοποιήσαμε είναι: Κανονική Γραμμική Παλινδρόμηση, Παλινδρόμηση LASSO, Παλινδρόμηση LAD και Random Forests. Επιπλέον, αναλύουμε το περιθώριο διαπραγμάτευσης στην τιμή τις ιδιοκτησίας και πώς επηρεάζεται από την τοποθεσία και άλλα χαρακτηριστικά της ιδιοκτησίας. Για να κατανοήσουμε πώς τα χαρακτηριστικά της ιδιοκτησίας επηρεάζουν το περιθώριο διαπραγμάτευσης, κατασκευάζουμε ένα ερμηνευτικό μοντέλο γραμμικής παλινδρόμησης και ερμηνεύουμε τους συντελεστές. Σε πρώτο στάδιο αυτό γίνετε για όλη την χώρα συνολικά και στη συνέχεια επιλέγουμε να επικεντρωθούμε στις τρεις νομαρχίες με τις περισσότερες ιδιοκτησίες. Ακόμα, παρέχουμε χάρτες θερμότητας ώστε να διακρίνουμε τοποθεσίες όπου το περιθώριο διαπραγμάτευσης εμφανίζετε να είναι μεγαλύτερο από το συνηθισμένο. Φυσικά, πριν κάνουμε όλα τα παραπάνω, παρουσιάζουμε μια λεπτομερή περιγραφική ανάλυση του συνόλου δεδομένων, όπου επισημαίνουμε ενδιαφέροντα γεγονότα και παράξενες συμπεριφορές που εντοπίσαμε.

https://doi.org/10.26219/heal.aueb.6011

 https://doi.org/10.26219/heal.aueb.6011

ACKNOWLEDGMENTS

I would like to thank my professor Dimitris Karlis, who helped me throughout my thesis and my academic career in general. Also, I would like to thank the whole department of Statistics, which guided me academically and professionally. Finally, I would like to thank my loving parents, my mother Adolyat Salahova and my father Tseperis Dimitris-Panagiotis, who believed in my skills and vision, and supported me with every possible way they could, in my academic journey.

https://doi.org/10.26219/heal.aueb.6011

https://doi.org/10.26219/heal.aueb.6011

CONTENTS

https://doi.org/10.26219/heal.aueb.6011

https://doi.org/10.26219/heal.aueb.6011

LIST OF ABBREVIATIONS

1. AVM: Automated Valuation Model
2. IVSC: International Valuation Standards Committee
3. NA: Not Available
4. SD: standard deviation
5. Skew: skewness
6. Kurt: kurtosis
7. OLS: Ordinary Least Squares
8. LAD: Least Absolute Deviations
9. RMSE: Root Mean Squared Error
10. BIC: Bayesian Information Criterion
11. LASSO: Least Absolute Shrinkage and Selection Operator

https://doi.org/10.26219/heal.aueb.6011

https://doi.org/10.26219/heal.aueb.6011

LIST OF TABLES AND FIGURES

https://doi.org/10.26219/heal.aueb.6011

https://doi.org/10.26219/heal.aueb.6011

https://doi.org/10.26219/heal.aueb.6011

https://doi.org/10.26219/heal.aueb.6011

1) INTRODUCTION

In this project, we will analyze a dataset that contains properties located in Greece that are up for sale. Our main goals are to find a model that can predict adequately the price of properties, understand the effect of location and how we can use it to make our predictions better, and finally search how the negotiation margin for the price is affected by location and other characteristics of the house. Although, before we begin to achieve the above goals, we will present some fundamental concepts of the real estate market, then we will deal with problems regarding the quality of the data and learn through descriptive analysis about our data and their characteristics.

The real estate market is a market that is generally regarded as a stable and reliable way for people to invest their money. Of course, everybody remembers the 2008 crisis that started when real estate prices and indices ceased to rise, first in America and then into the whole world, then rapidly "chaos" unleashed to all the markets globally. After that crisis, governments all around the world stepped in to stabilize the situation and prevent it from happening again by passing regulatory reforms like the Dodd-Frank Act in the United States and similar measures in other countries. Thus, we can understand that nowadays the real estate market is heavily regulated but also protected. This is where we come in and try to find a way to predict the price of a property, given its characteristics through an AVM, and try to understand the dynamics of different markets in Greece.

But why is it important to be able to predict the price of a property? The answer we get to this question heavily depends on who we ask it to. For example, a banker can use these predictions as the amount of collateral for a loan. A single investor is interested in looking at how the overall market is moving over time and decide on the optimal way to invest his money directly into the market by buying properties or indirectly by investing in construction companies. On the other hand, an investment company that is active in the property market is interested in portfolio valuation. The state can use those predictions to predict its future earnings by taxes. Economists are interested in studying the market to work on new economic theories or warn of imminent disasters in the market. Legal agents can use predictions to detect identification of fraudulent activities. Insurance companies need those predictions to calculate the premium or to compensate a customer for damages to his property. An owner who wants to sell his property is interested in having a prediction, to sell

https://doi.org/10.26219/heal.aueb.6011

it at a competitive market price, while a buyer can provide the characteristics of his desired property to the prediction model and use this prediction as a guide to start searching for good deals in the market. As we can see the prediction of a property's price is valuable for a huge array of uses and people, from the simple buyer who wants to buy a property to the whole government evaluating its future earnings from taxes. Now that we understand why predicting property prices is important, we need to explain how we get those predictions. Mainly there are two ways, the first one is through a certified appraiser and the second way is through an automated valuation model. There is a big debate in the community of property evaluation on who is more useful and reliable, for this reason, we have dedicated a small section where we discuss the advantages and disadvantages of each method. But before we continue it is important to recognize that we have two different kinds of approaches to the problem of evaluating the price of a property, signifying that people from two different fields of science try to "solve" the same problem, and often, work together towards this common goal. This further reinforces the significance of predicting property prices.

The dataset we will work on comes from a popular commercial site, where owners can post their properties and a series of characteristics to find potential buyers. The above already suggests that the data collected are given from the owners who want to sell their property, thus there might be bias to the information provided. Also, to use the site, you can do so for free for a limited time and when the time expires you have to pay a fee for the post to remain in the site. This fact can cause very serious problems because some of the owners might not want to pay the fee and thus, they constantly make new posts where they slightly change the characteristics of the house, which results in having the same house more than once in the dataset. Similarly, this problem is also caused when the owners reach out to one or more agencies that might also do a post on the site. Last but not least, the owner is not obligated to give every characteristic of a house, leaving us with "holes" (missing data) in our data.

In the first section, we will give some basic knowledge of the real estate market and provide some related work where the reader can look back to get more accustomed to real estate and AVMs. Then in the second section, we will deal with some of the data quality problems. In the third section, we will present the explanatory analysis of the data, to get more accustomed to them. In the fourth section, we will explore different kinds of models to predict the price of properties, the success of each model will be measured with indices. Then in the fifth section, we will analyze the

https://doi.org/10.26219/heal.aueb.6011

negotiation margins and how the location and other house characteristics affect them. Lastly, in the sixth section, we will have our conclusion and concisely present our results.

Finally, we have to clarify that the problem we are facing in this project is a data analysis problem and thus there is no completely right or wrong answer. The way someone approaches the problem may differ from someone else's, and for this reason, data analysis can be considered art by many, but we have to remember that it's also a science because you have to abide by some basic rules to produce results that make sense.

https://doi.org/10.26219/heal.aueb.6011

https://doi.org/10.26219/heal.aueb.6011

2) FUNDAMENTAL CONCEPTS OF REAL ESTATE AND RELATED WORK

In this section, our goal will be to familiarize ourselves with some fundamental concepts of the real estate market and present related work that can help the reader further understand the real estate market and the models used to make predictions.

The first concept we are going to explore is the definition of market value which is the most fundamental concept of understanding real estate markets. Although the definition of the market value has been throughout the years a subject of conflict, the IVSC defines it as: "the estimated amount for which an asset or liability should exchange on the valuation date between a willing buyer and a willing seller in an arm's length transaction, after proper marketing and where the parties had each acted knowledgeably, prudently and without compulsion" (IVSC 2003). The main points we should focus on are:

1. The market value is the exchange price in the marketplace at the date of valuation.
2. The transaction is an arm's length transaction, which means the parties involved are independent and have no pre-existing relationship or special circumstances that could potentially influence the terms of the deal.
3. Lastly, the transaction has to be made knowledgeably, prudently, and without compulsion.

Finally, the main assumption of the definition, is that the price is set under competitive market conditions.

According to Mooya (2017), the above definition does not fully resolve the underlying tensions and contradictions regarding the fundamental nature of market value, but it provides a common framework to guide the different valuation methods. Although there are different methods to define the market value, and those methods can be seen as an alternative, it does not mean that they are mutually exclusive.

https://doi.org/10.26219/heal.aueb.6011

Before moving on to the next concept regarding the AVMs, it is important to understand the common foundations of conventional market value theory, because all valuation methods, including AVMs, have the same objective of finding the market value. The theory of the market value has its roots in the neoclassical economic theory. The neoclassical economic theory is the combination of the 'marginal utility' school of economic thought and that of the 'classical' school. Alfred Marshall (1842–1924) was one of the greatest figures who supported the neoclassical school, and he was responsible for developing the famous Law of Demand: "The amount demanded increases with a fall in price, and diminishes with a rise in the price", by combining the idea of rational consumer choice with the law of diminishing marginal utility. Moreover, Alfred Marshall asked the question "What determines the market price", he answered this question by combining the view of the classical economists who supported that the market price is determined by the cost of production, and the view of marginalists who attributed the change to the demand. In fact, he supported that both supply and demand were responsible for establishing the market prices and explained that behind the cost lies the supply, and behind the utility lies the demand (Brue & Grant 2007). Finally, Alfred Marshall was one of the first economists who concerned himself with finding valuation techniques specifically for the real estate market.

The neoclassical theoretic approach is the theoretical foundation for most of the valuation approaches, including for AVMs, for real estate. Thus, it would be a crime to continue without briefly mentioning some key points. The main assumptions of the neoclassical theoretic approach are:

1. Rational Choice
2. Full Information
3. Equilibrium

The above key points are beautifully explained by Mooya (2017).

Rational choice is the theory that tries to explain how people make decisions. It is based on the assumption that people make logical decisions in order to maximize their gains and minimize their costs. According to Dyke (1981) the main points are:

1. Prefer more to less: This means that people always prefer more to less. As an example for our case, a buyer will always prefer the bigger house in contrast to the smaller one, while a seller will always prefer to sell at a higher price than a lower one.

https://doi.org/10.26219/heal.aueb.6011

2. Diminishing marginal utility: This means that the interest of a product diminishes as the quantity available increases.
3. Choice between alternatives: People have the ability of choice when presented with alternatives. This implies that they can express their preference or indifference between different options.
4. Transitive preferences: This means that the preferences of a person are consistent. As a typical example, when a person prefers A product to B product and B product to C product, then he has to prefer A product to C product.
5. Choosing the best alternative: People prefer always the choice which they think is better according to them.

In conclusion, the rational choice theory that people have at their disposal the required information and cognitive abilities to make rational decisions. Their decisions are based on maximizing their needs and desires but at the same time taking into consideration their limitations and the alternatives available to them. This theory is used to understand and model human behavior in economic and social contexts, such as buying, selling, investing, and other everyday decisions. Finally, as Lawson (2009) wrote: "Individuals exercise their own separate independent and invariable effect, whatever the context, thus guaranteeing that under some repeated conditions x, the same predictable outcome y will always follow".

Full information is one of the most important attributes in a market. Having full information is required if someone has to make a rational decision, and having full information means that everyone has information about the behavior and dispositions of other people and the attributes of the traded product. Therefore, we avoid any transaction costs that come as a byproduct of missing information, which means that the market price fully reflects all the attributes of the commodity in question thus limiting the potential for mispricing or divergence between prices and intrinsic worth or value.

Equilibrium: In our case, equilibrium is the state when the supply and demand are or tend to be in balance. Equilibrium is the result of both rational and fully informed decisions made by the people in the market. The concept of equilibrium can be understood in two ways:

1. Static equilibrium: This refers to a situation where the supply and demand for a product are equal at a given point in time.

https://doi.org/10.26219/heal.aueb.6011

2. This concept extends beyond a single market to consider the interrelations between multiple markets. It implies that the equilibrium in one market affects and is affected by equilibria in other markets.

At its core, the equilibrium suggests a state of rest, and the most important takeaway is that prices established by markets in equilibrium are held to be valid measures of market value.

To conclude, a market where people make rational and fully informed choices reaches equilibrium. According to the neoclassical theory, this market is the "perfect market" where the prices are meaningful representations of the value of a commodity. Finally, the equilibrium price/the market value set in such a market is both determinate and autonomous and the valuation is the process of attempting to discover this price, with 'valuation accuracy' as a measure of the degree of success.

The second key point we will discuss is the AMVs. First of all, AVM stands for Automated Valuation Model, and according to RICS (2013) "Automated Valuation Models use one or more mathematical techniques to provide an estimate of the value of a specified property at a specified date, accompanied by a measure of confidence in the accuracy of the result, without human intervention post-initiation". The main points of the above definition are:

1. AVMs are models that use mathematical techniques
2. The focus of AVMs is to provide an estimate of the value of a property
3. The resulting estimate is accompanied by a measure of confidence
4. After the initiation of the model, zero human interference is required

According to RICS (2013), the main variables that are included in AVMs are:

1. Architectural style
2. Property type
3. Age (or year built)
4. Location (e.g. geographic co-ordinates or allocation of a 'locality' identifier)
5. Floor area
6. Number of rooms
7. Number of bedrooms
8. Number of bathrooms
9. Plot size

https://doi.org/10.26219/heal.aueb.6011

10. Conservatory
11. Outbuildings
12. Parking/garaging
13. Quality (i.e., specification of the property)
14. Condition

The AVMs can be used for many different purposes RICS (2013):

1. By lenders for the loan origination process or subsequent revaluation for credit decision purposes: Lenders can use AVMs to check if the value is adequate, here the lenders avoid paying someone to do an onsite inspection, but they add risk by evaluating the price with an AVM. Also, lenders can use AVMs part-way through a mortgage term to check how a property value may be changing
2. In-arrears assessment and planning: The lenders can use AVMs to check the value of a security to ensure that the loan can be secured and if they can strike a deal with a customer who is in arrears.
3. In valuations: Valuers can request an evaluation from an AVM to serve as an audit of the original valuation
4. In mass appraisal, for local taxation purposes: sophisticated AVMs can be used to provide valuation estimates for thousands or millions of properties, cost-effectively
5. For the provision of valuation estimates for individual capital tax purposes
6. For the identification of fraudulent activities
7. For the valuation of whole portfolios
8. For estimating compensation payments to owners of residential property due to the effect of the use of new public works
9. For cost/benefit analyses for potential public expenditure
10. For lending: AVMs can be to a portfolio of properties subject to mortgage, to check how the portfolio is performing

According to Mooya (2017), all valuation methods are based on economic theories, and AVMs are no exception to this rule. In fact, AVMs are the most advanced phase of the application of economic theory to the valuation of real estate. AVMs are closely tied to market value by their reliance on neoclassical economic principles, particularly the ideas of equilibrium, supply and demand, and rational choice. These models use methods like regression to analyze large datasets in real estate and produce estimations of property values based on observed market behavior. By

https://doi.org/10.26219/heal.aueb.6011

considering several property characteristics as well as their marginal contributions towards total value, AVMs strive to be able to reflect true market value or equilibrium price where supply equals demand. This link ensures consistency with the theoretical foundation of market value hence, AVM-produced valuations are consistent, reliable, and aligned with the theoretical foundation ensuring they become important tools for overcoming inefficiencies in real estate markets.

The next fundamental concept we will discuss is the differences between AVMs and typical appraisers. AVMs are known for their speed, cost-effectiveness, requirement of minimal user interaction, consistency and objectivity over traditional methods. On the other hand, AVMs lack any insights provided on-site inspections, also AVMs assume only average conditions which can lead to misleading estimates in atypical situations, and finally, AVMs lack the "street level" judgment which is an attribute that human valuers "bring to the table" and is essential for the accurate prediction of a property. It is only logical at this point to ask the following question: Who is more accurate in their predictions? Mooya (2017) suggests that the correct question we should ask is: Is there something in the economic theory that predisposes one or the other method to greater accuracy? The answer he gave suggests that as the market becomes more competitive and homogeneous AVMs are better at valuation estimating prices but when the market is becoming less competitive, human valuers are more accurate with their predictions about the prices. Of course, in the real world, there is no perfectly competitive market or a market that is not competitive at all, but we can imagine the competitiveness of a market as a spectrum. In markets that tend to be more competitive, AVMs are more accurate predictors whereas in the less competitive markets, human valuers are more precise.

Before moving on it is essential to have a quick look at the most recent crisis that happened in 2008 and it was related to real estate. We will base our discussion on the paper D'Amato & Kauko (2017).

The global crisis of 2007-2008 served as a reminder that financial systems and real estate are closely connected, something that was at the time forgotten. At first, the property price indices had stopped rising and in fact, they took a downward turn for many countries (Martin (2011) and Kušar (2012)), and thus the belief in ever-growing property prices was finally put to rest together with the idea of property markets being the safest investment. According to Garton (2009), the main cause of the crisis was due to the gap in information regarding the institutional settings and

https://doi.org/10.26219/heal.aueb.6011

design of the 'shadow banking system' first seen in the early 2000s. The main gaps in information were:

1. Sensitivity of Securities to House Price Trends: The interconnected chain of securities was highly sensitive to fluctuations in house prices and thus when inevitably the real estate prices ceased to rise, the value of these securities plummeted
2. Creation of Symmetric Information via Complexity: The complexity of financial products and their underlying assets created an illusion of symmetric information. This means that investors believed that they understood the risks, but the complexity obscured the true nature of the securities.
3. Opaque Risk Distribution: The way risk was spread through various financial instruments was not transparent, and thus investors and institutions could not assess the risk accurately.
4. Trade in Asset-Backed Security Indices Linked to Subprime Bonds: The trading of asset-backed security indices tied to subprime mortgages worsened the crisis when the underlying assets began to fail.

Investment banks and financial institutions capitalized on the real estate "boom" and purchased more and more real estate mortgages, which were then securitized into financial products like Collateralized Debt Obligations (CDOs). These products bundled various types of debt, including mortgages that were highly rated by rating agencies. When the real estate market began to decline, these highly rated, but yet toxic products, led to significant financial hardships, and the investment banks and financial institutions that heavily invested in those properties faced enormous losses. Finally, did AVMs have anything to do with this crisis? When evaluating properties we can use the following (but we are not limited to) models, market-sales, income, and cost approach. In the market-sales approach the valuation happens by observing the price and the characteristics of comparable properties, Appraisal Institute (2011). The problem that emerged was that when a broker requested a valuation for a property, he also gave a target value for this property, and if the appraiser didn't meet the target value the broker could simply choose another appraiser that would give him the requested target value, (Bitner (2008), p. 92). This was possible because as said earlier the market-sales approach is based on comparable properties, thus the appraisers could easily choose to use only those comparable properties that would yield the target price. Finally, two models are based on the income approach, IVSC (2011), the first one is the direct capitalization, which is a process that transforms an infinitive group of rents into a value, and the second one is yield capitalization which

https://doi.org/10.26219/heal.aueb.6011

is a process that transforms a finite group of rents into a value, Appraisal Institute (2011). Those methods are reserved when no comparable properties are available. The main problems with the income methods are that they don't consider the market cycles and that there is an "information gap". Generally, real estate markets go through cycles of growth, stability, and decline and these cycles affect property values significantly. The "information gap" refers to the discrepancy between the information available to appraisers and the reality of market cycles, in fact this gap arises because of two reasons:

1. Appraisers often use historical data to project future income, which may not accurately reflect upcoming market changes.
2. The capitalization rates used in those methods are typically derived from recent market transactions, which might not account for impending market shifts.

The consequence of not addressing the market cycles when the market is on upswing is that the properties might be overvalued because the methods do not predict the eventual downturn, and on the other hand, when the market is in downturn, properties might be undervalued because the methods do not anticipate the recovery phase.

The last concept we will discuss is the definition of the negotiation margin. The negotiation margin in the real estate market refers to the difference between the asking price of a property and the final sale price agreed upon by the buyer and the seller. The negotiation takes the following form, a buyer makes their first offer (asking price) then a potential buyer appears, and the two parties begin the negotiation. The potential buyer having in mind the asking price makes a counteroffer to the seller, of course, the offer that the potential buyer makes is lower than the asking price. Next, the seller can either accept the price proposed by the potential buyer and strike a deal or suggest another price lower than his asking price but higher than the price given by the potential buyer. This back-and-forth repeats until both parties agree on a price or someone exits the negotiation. If the negotiation is successful and a deal is struck, we can measure the negotiation margin. For example, if a seller sets an asking price of 100000$ for a property, and then a potential buyer shows up and offers 50000$. The two parties begin the negotiations, and let's say that a deal is struck at the price of 75000$ for the property, the negotiation margin is 25000$ or 25% of the asking price.

https://doi.org/10.26219/heal.aueb.6011

Generally, real estate markets don't work exactly as we portrayed in the example above. Many times, a seller will set an asking price that is too high for any potential buyer, and after some time the seller will be forced to lower his asking price or leave the market.

The negotiation margin is subject to variations, the main factors that affect the following:

1. Market Conditions: If we have a market where supply exceeds demand, the negotiation margins are wider because sellers compete with each other to sell their property over a limited pool of buyers. On the other hand, if demand exceeds supply the negotiation margins are narrower because the sellers are tempted to find buyers who are willing to pay closer to the asking price. (Akin et al. (2014))

2. Property Characteristics: The characteristics of a property are one of the most important factors in negotiation margin.

3. Motivation: Sellers or buyers might be motivated to sell/buy a property for several reasons (ex. Need to move to another country). A motivated seller will offer a higher negotiation margin, while when there is a motivated buyer the negotiation margin becomes narrower.

Galesi et al. (2020) in his paper analyzed the negotiation margins of properties in Spain, he concluded that when there was a crisis during 2010-2014 the mean negotiation margin was 15.6%, and it peaked in 2012 when it reached 20.1%. After the restoration of the market crisis in 2014 the prices steadily rose and peaked in 2018 and at the time the mean negotiation margin was reduced to 9%. Finally, the paper presents the negotiation margin for all provinces of Spain. In the years of the crisis 2010-2014, thirty five out of 50 provinces had negotiation margins above 20%, with the highest negotiation margins being observed in Ourense at 27.5%, Murcia at 26.8%, and Leon at 25.7%. At the same time, the provinces of Madrid and Barcelona had negotiation margins of 13.8% and 18.5% respectively. On the other hand, during the years of 2015-2018 when the market recovered from the crisis the negotiation margins plummeted, exceptions were the provinces of Ourense with 22.9% and Lleida with 20.1% negotiation margins, while the rest had negotiation margins less than 20%. The lowest negotiation margins again appeared for Madrid and Barcelona with 5% and 8.6% respectively. To sum up we can understand that the market cycles and the location properties play important roles in determining the negotiation margin.

 https://doi.org/10.26219/heal.aueb.6011

Before we move on with our first touch with the data, it is essential to study some of the models used in the real estate market, some of which we will also use in this thesis.

The first model we will discuss is the Sales Comparison approach. Sales Comparison is a real estate valuation method that estimates the value of a property by comparing it to properties that are sufficiently similar (what counts as sufficiently similar is not clear and strictly defined (Eilers & Kunert (2017)), and recently sold in the same area. This approach is widely used for residential properties but can also be applied to commercial properties. The main point of this approach is that a buyer will not pay more for a property than the cost of acquiring a similar property with the same utility. The method of estimating the value of a subject property can be summarized with the following steps:

1. Find properties that are similar to the subject property in terms of size, location, condition, and features. Those properties should have been sold recently.
2. Examine the similarities and differences between the comparables and the subject property.
3. Make adjustments to the sale prices of the comparables to account for differences between them and the subject property. Adjustments can be positive or negative and can include factors such as: Location (neighborhood desirability), Size (square footage), Condition (age, maintenance level), and Features (number of bedrooms, bathrooms, presence of a garage or pool)
4. After making the necessary adjustments, determine the adjusted sale price for each comparable property.
5. Use the adjusted sales prices of the comparables to estimate the value of the subject property. This can be done either by simply averaging the adjusted prices of the comparables or by doing a weighted average of the comparables, where those that are more similar receive a bigger weight.

Before we move on, we will present a simple example.

Let's assume that we want to estimate a **subject property** with the following characteristics:

- Size: 150 square meters

https://doi.org/10.26219/heal.aueb.6011

- Bedrooms: 3
- Bathrooms: 2

And now let's assume we have the following comparables:

Property 1:

- Size: 125 square meters
- Bedrooms: 3
- Bathrooms: 1
- Sold: 480000 €

Property 2:

- Size: 165 square meters
- Bedrooms: 2
- Bathrooms: 2
- Sold: 500000 €

Property 3:

- Size: 150 square meters
- Bedrooms: 3
- Bathrooms: 2
- Sold: 490000 €

Now let's assume that from our market analysis, we find:

1. That each extra/less squared meter adds/subtracts 1000€ from the price
2. That each extra/less bedroom adds/subtracts 20000€ from the price
3. That each extra/less bathroom adds/subtracts 10000€ from the price

We calculate the Adjusted Sale Prices:

Property 1 Adjusted Sale Price = 480000 – 25×1000 – 10000 = 445000€

Property 2 Adjusted Sale Price = 500000 + 15×1000 -20000 = 495000€

Property 3 Adjusted Sale Price = 490000€ (exactly the same characteristics)

The Estimated Market Value of Subject Property is:

Estimated Value = (445000+495000+490000)/3 = **476666,67€**

https://doi.org/10.26219/heal.aueb.6011

The next model we will discuss is the Capitalized Income Value approach. Capitalized Income Value involves separately determining the value of the land and the improvements (buildings and other structures) on the land, and this approach is generally reserved for income-producing properties. In Germany, there is the Guideline Land Value (Bodenrichtwert) which is an average value per square meter for land in a specific location. These values are regularly updated and published by land valuation boards based on actual purchase prices. Those guideline values can be adjusted to reflect the unique characteristics of the property's site area (Eilers & Kunert (2017)). The Capitalized Income Value approach can be summarized by the following formula:

Capitalized Income Value = (Net income - land value × r) × V + land value

- Net income: This is the income the property generates from rent after deducting operating costs. Operating costs include maintenance, management, operation costs, and a provision for potential loss of rent.
- Land value: Average value per square meter for land in a specific location
- r: is the yield for annual return on land value
- V: V is a multiplier calculated as: $((1+r)^n-1)/((1+r)^n \times r)$
    - Where n is the remaining useful lifespan of the property in years.

Now let's see an example. Assume we have the following data:

- Guideline Land Value: €200 per square meter
- Site Area: 1000 square meters
- Net Rental Income: 50000€ per year
- Yield Rate (r): 5%
- Remaining Useful Lifespan (n): 20 years

We get the following:

Land Value = 1000 × 200 = 200000€

V = $((1+0.05)^{20} -1)/((1+0.05)^{20}\times 0.05)$ =12.46

Capitalized Income Value = (50000 - 200000 × 0.05) × 12.46 + 200000 = **698400€**

Next is the model of Depreciated Replacement Cost. Depreciated Replacement Cost is a valuation method commonly used in Germany, particularly for single-family and

https://doi.org/10.26219/heal.aueb.6011

two-family dwellings (Eilers & Kunert (2017)). This method combines the value of the land with the depreciated value of the improvements (buildings and external appurtenances). The land value is determined based on the site area and the guideline land value (Bodenrichtwert). We calculate the replacement cost of the improvements at the valuation date, this calculation is based either on typical building costs per squared meter gross external area (Bruttogrundfläche—BGF) (Eilers & Kunert (2017)). Then we add the value of the external appurtenances as a flat rate addition to the replacement cost of the improvements calculated before. After that, we reduce the previously calculated amount, with depreciation due to age and any reduction needed due to building defects or damage. Finally, we add this amount to the land value, to get the Depreciated Replacement Cost and we adjust it according to market conditions. If the achievable prices on the local market are above the replacement cost, we adjust by multiplying with a number larger than 1 otherwise, we adjust by multiplying with a number smaller than 1 (This multiplier is based on an empirical analysis of comparable properties). Thus, we end up with the following formula:

$$\text{Depreciated Replacement Cost} =$$
$$= (\text{Land Value} + \text{Replacement Cost of Improvements}) \times \text{Multiplier}$$

Let's assume the following example:

- Site Area: 500 square meters
- Guideline Land Value: 300€ per square meter
- Gross External Area: 150 square meters
- Replacement Cost per square meter of Gross External Area: 1500€
- External Appurtenances Value: 20000€
- Age Depreciation: 5% depreciation over 20 years
- Multiplier Adjustment: 1.2 (indicating the local market prices are 20% higher than replacement cost due to demand)

And we get:

Land Value= Site Area× Guideline Land Value = 500 × 300 =150000€

https://doi.org/10.26219/heal.aueb.6011

Replacement Cost of Improvements = Gross External Area × Replacement Cost of Gross External Area + External Appurtenances Value = 150 × 1500 + 20000 = 245000 €

Depreciated Value of Improvements = Replacement Cost of Improvements × (1 − Depreciation Rate)^{20} = 245000 × (1−0.05)^{20} = 87829€

Depreciated Replacement Cost=Land Value+Depreciated Value of Improvements = 87829 + 150000 = 237829€

Adjusted Depreciated Replacement Cost = Depreciated Replacement Cost × Multiplier = 237829 × 1.2 = **285395€**

Another very popular way to evaluate properties in real estate is the regression model. Regression models are statistical tools used to understand and quantify the relationship between a dependent variable and one or more independent variables. Regression models can be used for either explanatory purposes or prediction purposes. In our case, we focus on the predictive ability of those models, and our main goal is to predict the value of a property (dependent variable) based on its characteristics (independent variables) (ex. Number of bedrooms, number of bathrooms, type of property, and many more). There are many regression models, the most famous and basic model is the linear regression model. In the linear regression model, the core assumption is that this relationship between the dependent variable and the independent variable is linear, meaning that a change in the independent variable (property characteristics) corresponds to a proportional change in the dependent variable (value of a property) (Sopranzetti (2015)). In linear modeling, we try to minimize the difference between the predicted values from the model and the actual observed values. This method is called OLS (Ordinary Least Squares) and the target function we want to minimize is: $S = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$. The linear model can be modified to deal with problems found in the data. Some of its modifications include

1. Least Absolute Shrinkage and Selection Operator (LASSO)
2. Least Absolute Deviations (LAD)

LASSO is a variation of Linear regression, also known as L1 regularization. The main difference Linear regression is that instead of minimizing the OLS function we want to mimimize the following function: $L = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^{p}|b_j|$.The first part of the previous function is the same as the OLS, but we add an extra term. The λ is the regularization parameter, this parameter encourages sparsity in the

https://doi.org/10.26219/heal.aueb.6011

model by penalizing the absolute values of coefficients and this can lead some coefficients to be exactly zero, effectively performing variable selection and simplifying the model. LASSO is particularly useful when dealing with high-dimensional data or when there is a suspicion that some independent variables may not contribute significantly to the model. (Doumpos et al. (2021))

LAD is also a variation of Linear modeling and is also known as quantile regression. The main difference with Linear modeling is that LAD minimizes the function $L = \sum_{i=1}^{n} |Y_i - \hat{Y}_i|$. By minimizing absolute differences LAD becomes less sensitive to outliers thus a more robust method when dealing with datasets that contain outliers. Generally, LAD modeling can be used to estimate any quantile we want, but in this thesis, we will use LAD to estimate the median. (Coleman & Larsen (1991) and Narula et al. (2012))

Finally, in recent years there has been a development of new and more advanced methodologies like neural networks (Kauko et al. (2017)), that sadly we won't discuss in this thesis, and random forests (Antipov & Pokryshevskaya (2012)). Random forests are an ensemble learning method used for classification and regression tasks. Decision trees are predictive models built by recursively splitting the data into subsets based on the value of an attribute. As we said Random Forest is an ensemble method, which means it combines the predictions of multiple decision trees to produce a single result, this produces a more robust and accurate prediction than using a single decision tree. In our case, because we want to evaluate a property, we have a regression task, thus the final prediction of the random forest is the average of the predictions from all individual trees. Generally, Random forests are difficult to interpret, they are computationally intensive and time-consuming, especially with large datasets.

https://doi.org/10.26219/heal.aueb.6011

https://doi.org/10.26219/heal.aueb.6011

3) PRE-PROCESS OF DATA

In this section, we will have our first contact with the data. We will start by giving some general information about the data and then proceed to deal with the problem regarding the properties that appear more than once in our data.

The data we have to analyze came from a commercial site where owners post their properties, the characteristics, and the asking prices. It's crucial to understand that this price is not the market value of the property but the price that the owner would be "happy" to sell the property. Also, every characteristic is given to us by the owner, which means there might be bias or even worse false information about the property. The dataset contains 186590 lines which translates to properties, each line has 24 variables (the characteristics) which we present in Table 1. For each variable in the dataset, we give the name, an explanation, and the type of the variable. Finally, the data were collected for the first quarter of 2023.

NAME	EXPLANATION	TYPE
ID	A unique identification number given to each property	Numeric
TK	A 3-digit number based on the location of the property	Numeric
Price	The price of the house in euro	Numeric
Type	The type of the property	Categorical with four levels: 1. House 2. Flat 3. Duplex 4. Maisonette
Usable area	Usable residence area in square meters, excluding semi-covered and storage areas	Numeric

Table continues on the next page.

https://doi.org/10.26219/heal.aueb.6011

NAME	EXPLANATION	TYPE
Land area	The surface of the land (only for houses). This must be zero for all other types of property	Numeric
Year	Year the construction of the property was completed	Numeric
Floor	Highest floor of the **property** (0 for ground floor, -1 for first basement)	Numeric
Total number of floors	The highest floor of the **building**	Numeric
Parking	Existence of privately allocated parking spaces	Categorical with two levels: 1. YES 2. NO
Type of parking	Availability of at least one indoor private parking space	Categorical with two levels: 1. YES 2. NO
Quality	Quality of Construction	Categorical with four levels: 1. Bad 2. Medium 3. Good 4. Very good
Number of bedrooms	Number of bedrooms in the property	Numeric
Elevator	Availability of use of an elevator for the property.	Categorical with two levels: 1. YES 2. NO

Table continues on the next page.

https://doi.org/10.26219/heal.aueb.6011

NAME	EXPLANATION	TYPE
Number of baths	Number of bathrooms	Numeric
Condition	Condition of maintenance	Categorical with four levels: 1. Bad 2. Medium 3. Good 4. Very Good
Storage	Existence of own indoor storage area	Categorical with two levels: 1. YES 2. NO
Town plan	The property is located within the town plan.	Categorical with two levels: 1. YES 2. NO
Longitude	The x coordinates in the map	Numeric
Latitude	The y coordinates in the map	Numeric
Region name	The name of the region where the property is located	Categorical with 62 levels
Municipality name	The name of the municipality where the property is located	Categorical with 940 levels
Neighborhood name	The name of the neighborhood where the property is located	Categorical with 4611 levels

Table 1: Short presentation of the dataset variables

Before proceeding further with the exploratory analysis of our data we have to deal with the problem where some properties are represented more than once in our data. This happens for two reasons:

1. The site is only free for the first month and after that, the owner must pay a fee to keep the post on the site. So many owners might have multiple one-month posts.

https://doi.org/10.26219/heal.aueb.6011

2. An owner may speak with an agency to help him sell the house, so the same property may be posted by the owner and the agency.

The hard part of dealing with this problem is that we can never be 100% sure which of the houses are represented more than once. Also, there is no textbook solution for this problem, so we had to come up with our unique solution.

Step 1: Split the houses into groups based on the usable area according to the following rule. All the houses that have a usable area of 20-25 m^2 are in the first group, all the houses that have a usable area of 26-30 m^2 are in the second group, all the houses that have a usable area of 31-35 m^2 are in the third group, etc.

Step 2: For each group, we calculate the Gower distances based on the variables: Type, Year, Floor, Parking, Number of bedrooms, Number of baths, Region name, and Municipality name.

Step 3: We set a specific cutoff point, where we get those couples of properties with Gower distances below the cutoff point (based on the variables of step 2).

Step 4: We take the couples from Step 3, calculate their geometric distance, and keep those couples that have a distance of less than 300 meters.

Step 5: We set up a network and apply the Louvain algorithm to detect communities. (which houses are the same)

The Louvain algorithm is used in network analysis to identify communities and it was proposed by Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre in 2008. The algorithm optimizes the modularity function, which measures the quality of a partition of the network into communities. The algorithm is widely used in social network analysis, biology, and information retrieval.

Step 6: Then for the communities that have more than one property, we choose to represent them by selecting the property inside the community with the least NAs.

As said previously, this method is not a textbook solution and thus there might be better ways to tackle the problem. The advantage of this solution is that we combine logic and different but easy-to-understand methods to detect very similar properties. Finally, the original contained 186590 properties and after implementing our method the dataset has 179105 properties.

https://doi.org/10.26219/heal.aueb.6011

https://doi.org/10.26219/heal.aueb.6011

4) EXPLANATORY ANALYSIS

In this section, we will perform univariate exploratory analysis for each variable and examine bivariate relationships between each variable and the house prices. Also, in this section, we will address the issue of missing data in our variables.

The first variable we will explore is the type of the residence. It is a factor variable that has four levels, "HOUSE", "FLAT", "DUPLEX" and "MAISONETTE". Observing both Table 2 and Figure 1, we note that flats constitute more than 50% of our dataset, while houses are nearly one-fourth. Maisonettes account for 17.2% and finally, duplexes represent 0.8% of the dataset. We need to clarify that duplexes are maisonettes that were built recently.

INDEX	TYPE OF RESIDENCE				
	HOUSE	FLAT	DUPLEX	MAISONETTE	NA
Frequency	46816	100170	1363	30756	0
Percentage	26.1%	55.9%	0.8%	17.2%	0%

Table 2: Basic measures for the type of residence

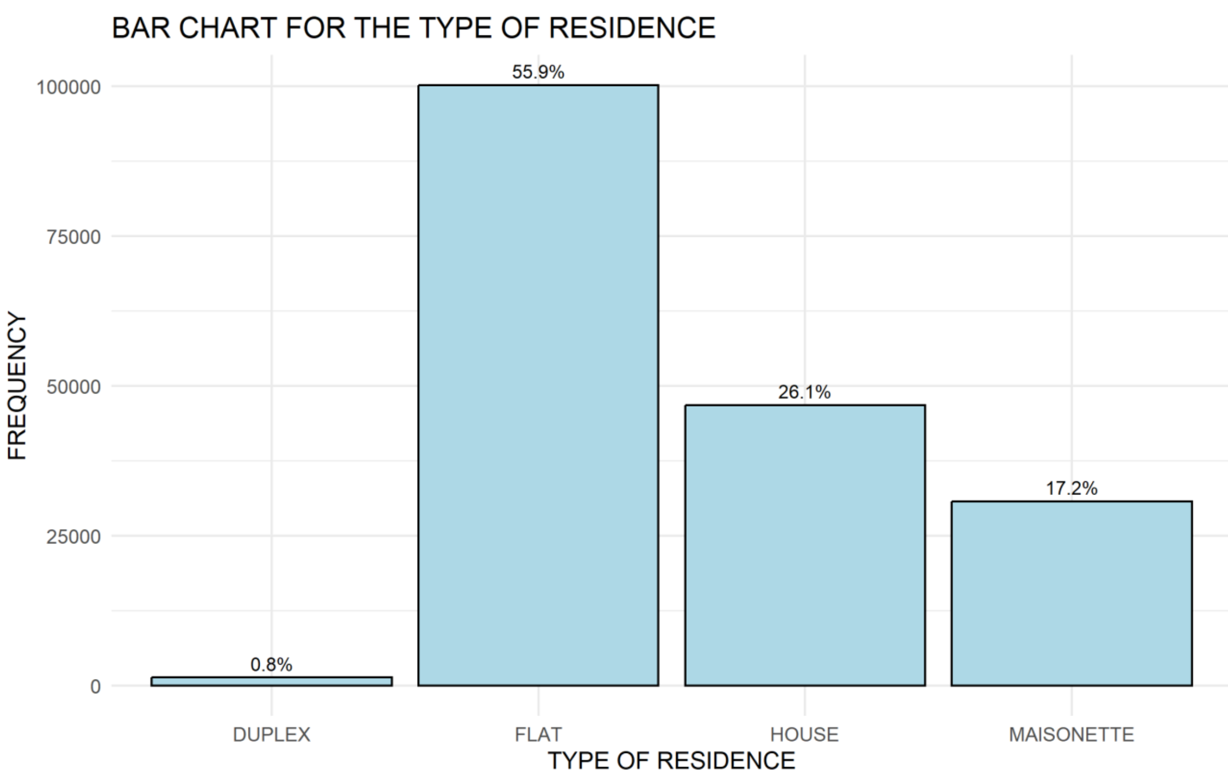

Figure 1: Bar-plot for the type of residence

https://doi.org/10.26219/heal.aueb.6011

The second variable we will explore is the price of the houses. This variable is continuous, and it is measured in euros. In Table 3, we observe a significant difference between the mean and the median, in fact the mean is much bigger than the median which suggests there is right skewness. This is confirmed by the skewness index which is much greater than 0. The minimum price in our data is 6600€, this is a very low price for any type of property we have in this dataset, which means that this might be some other type of property. On the other hand, the maximum price is 10000000€, this price is very high, and properties close to these prices belong to a more “premium market” which is probably not that competitive. Examining Figure 2 we confirm the presence of right skewness. Additionally, we observe some “spikes” in the prices. Upon closer examination, it becomes evident that although the price is a continuous variable, property owners tend to select rounded values. This is to be expected because the prices are inputted by people, and people tend to round numbers to make them easier to understand.

mean	median	SD	min	max	range	skew	kurt	NA%
3149336€	200000€	360216€	6600€	10000000€	9993400€	4.97	61.14	0%

Table 3: Basic measures for the price variable

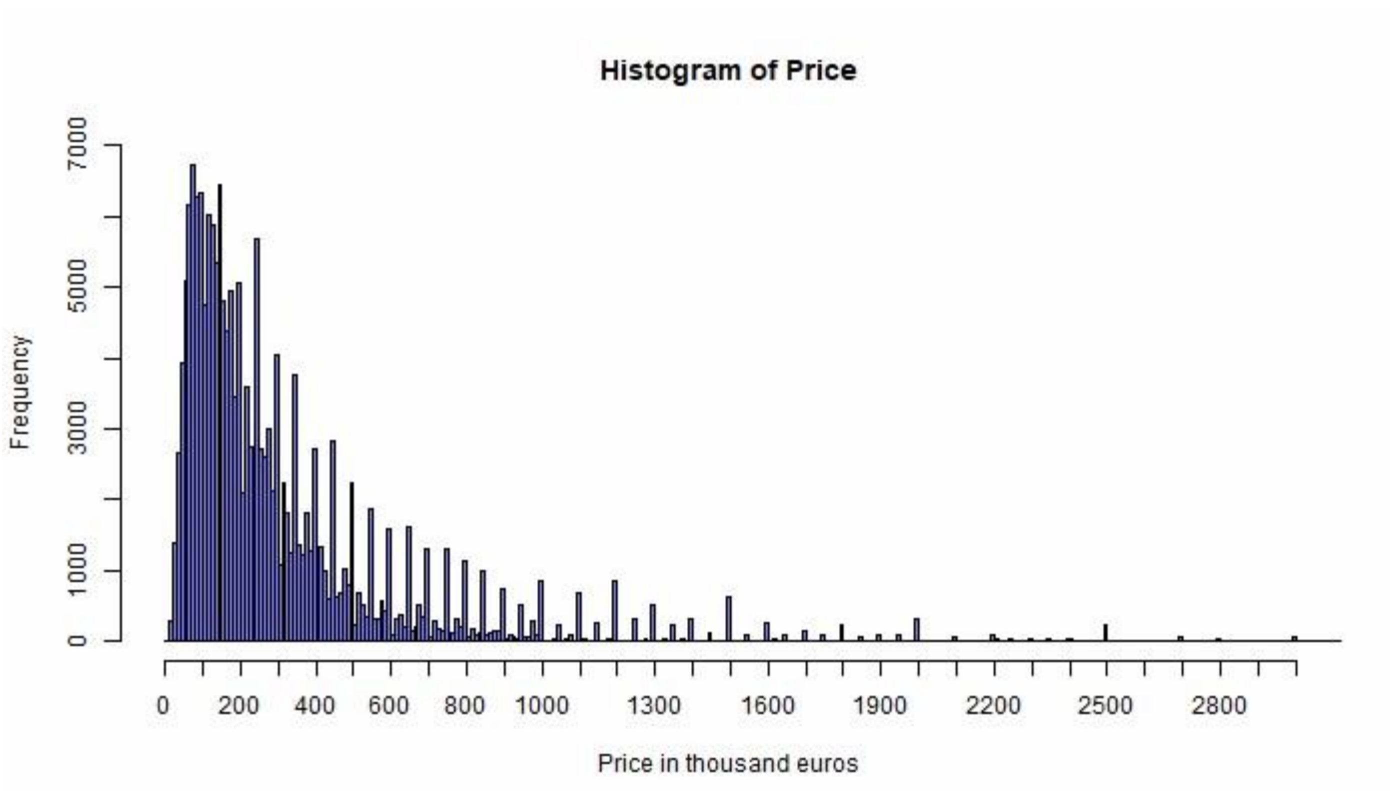

Figure 2: Histogram for Price

https://doi.org/10.26219/heal.aueb.6011

The next variable we will take a look at is the residence area of the property. This variable is also continuous, and it is measured in square meters. According to Table 4, the mean is bigger than the median which signals evidence of right skewness. Furthermore, the skewness index, which is larger than 0, once again suggests the presence of right skewness. Upon examining Figure 3 we conclude that there is right skewness. Additionally, we again detect "spikes" in the histogram. In this case, people tend to round the residence areas, for example, if a property is $39m^2$ the owner will probably say that the area is $40m^2$. By doing this they artificially "upgrade" their property. Finally, the presence of a minimum value of $20m^2$ reinforces our belief that not all properties conform to the types of residences observed earlier.

mean	median	SD	min	max	range	skew	kurt	NA%
$130.75m^2$	$106m^2$	$79.37m^2$	$20m^2$	$400m^2$	$380m^2$	1.31	1.26	0%

Table 4: Basic measures for the residence area

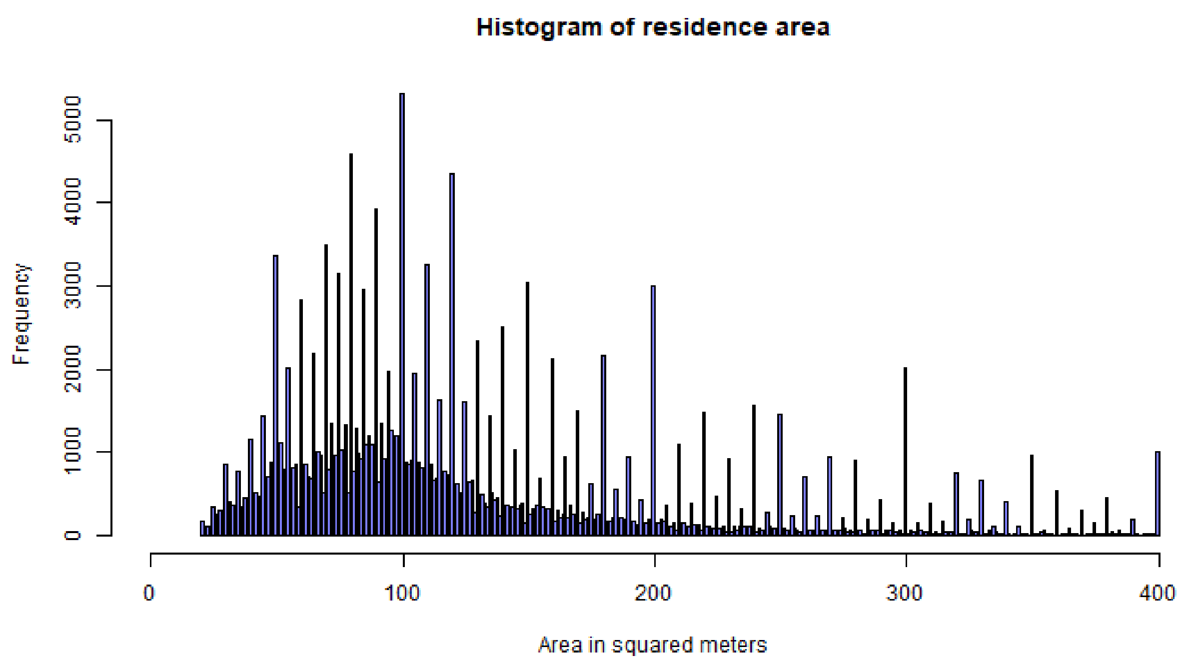

Figure 3: Histogram of residence area

 https://doi.org/10.26219/heal.aueb.6011

The year the property was constructed is a continuous variable, but in our dataset, takes many discrete values. Observing Table 5, we note that the mean and median are relatively close to each other. Also, the skewness index is close to 0. Upon examining Figure 4 we conclude that there is skewness, this we have left skewness. Once again, we detect "spikes", although their occurrence is less frequent on this occasion. These "spikes" notably appear in rounded years. The "spikes" appear in rounded years. Finally, upon closer inspection, we identify three distinctive peaks in the distribution, located near 1970, a steeper one around 2010, and a very steep, possibly incomplete peak near 2020.

mean	median	SD	min	max	range	skew	kurt	NA%
1988.56	1990	22.89	1900	2023	123	-0.71	0.98	0%

Table 5: Basic measures for the year the property was constructed.

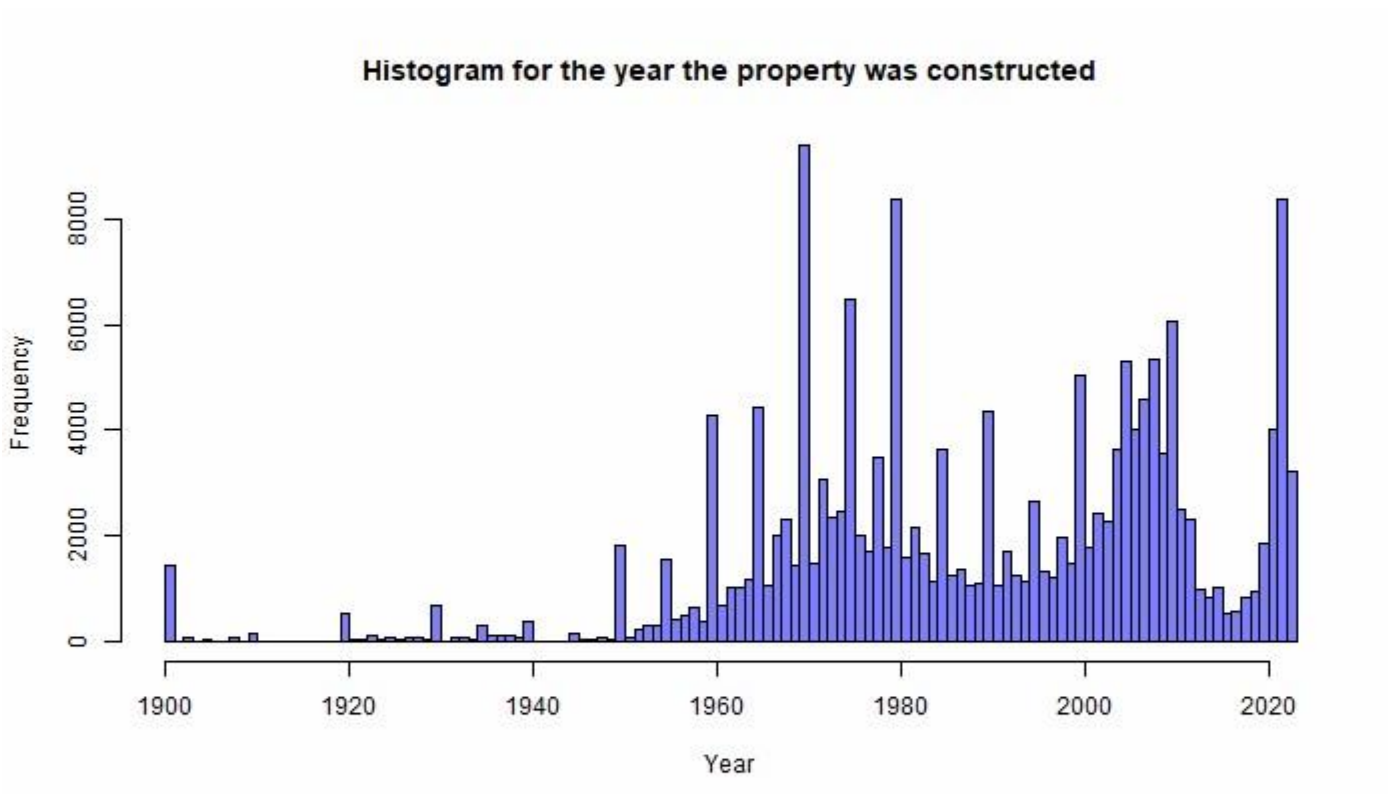

Figure 4: Histogram of the year the property was constructed.

https://doi.org/10.26219/heal.aueb.6011

The next variable we will examine is the floor of the residence. If the residence has multiple floors, then we keep the highest floor of the residence. This variable takes a few distinct numeric values, and for this reason, we will treat it as a factor. Observing both Figure 5 and Table 6 we see that the most common floor is the ground floor. Finally, we detect some NAs, in fact, 10.2% of the variable is missing.

Index	Floor of the residence										
	-1	0	1	2	3	4	5	6	7	8	NA
Frequency	4250	53497	36958	23027	17163	11361	8098	3835	1821	781	18314
Percentage	2.4%	29.9%	20.6%	12.9%	9.6%	6.3%	4.5%	2.1%	1.0%	0.4%	10.2%

Table 6: Basic measures for the floor of the residence

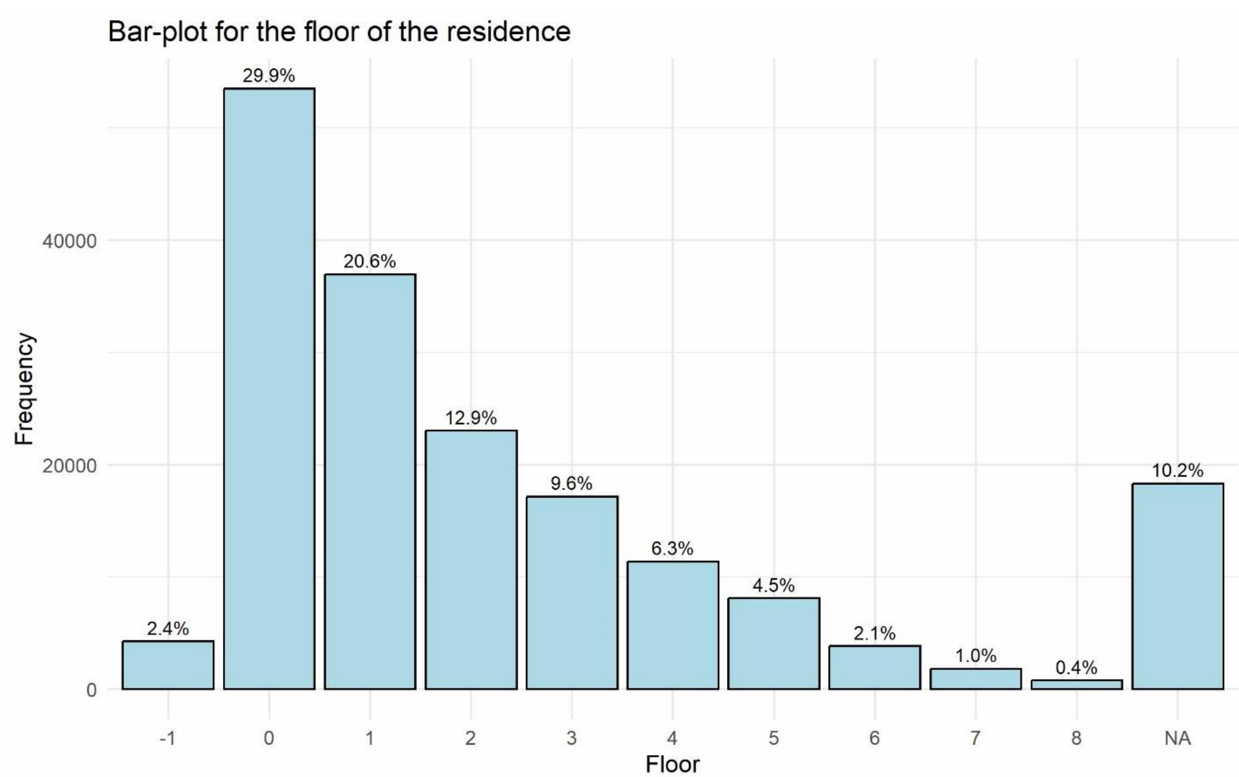

Figure 5: Bar-plot for the floor of the residence

https://doi.org/10.26219/heal.aueb.6011

Now we will have a look at the variable for the number of bedrooms of the residence. This variable takes a few distinct numeric values, and for this reason, we will treat it as a factor. By having a quick look at Table 7 or Figure 6, we conclude that most of the residences have two bedrooms. Additionally, it is noteworthy that 3% of the residences do not have a specified number of bedrooms.

Index	Number of bedrooms						
	1	2	3	4	5	6	NA
Frequency	29832	59864	52080	21806	7190	2963	5370
Percentage	16.7%	33.4%	29.1%	12.2%	4.0%	4.0%	3.0%

Table 7: Basic measures for the number of bedrooms in the residence

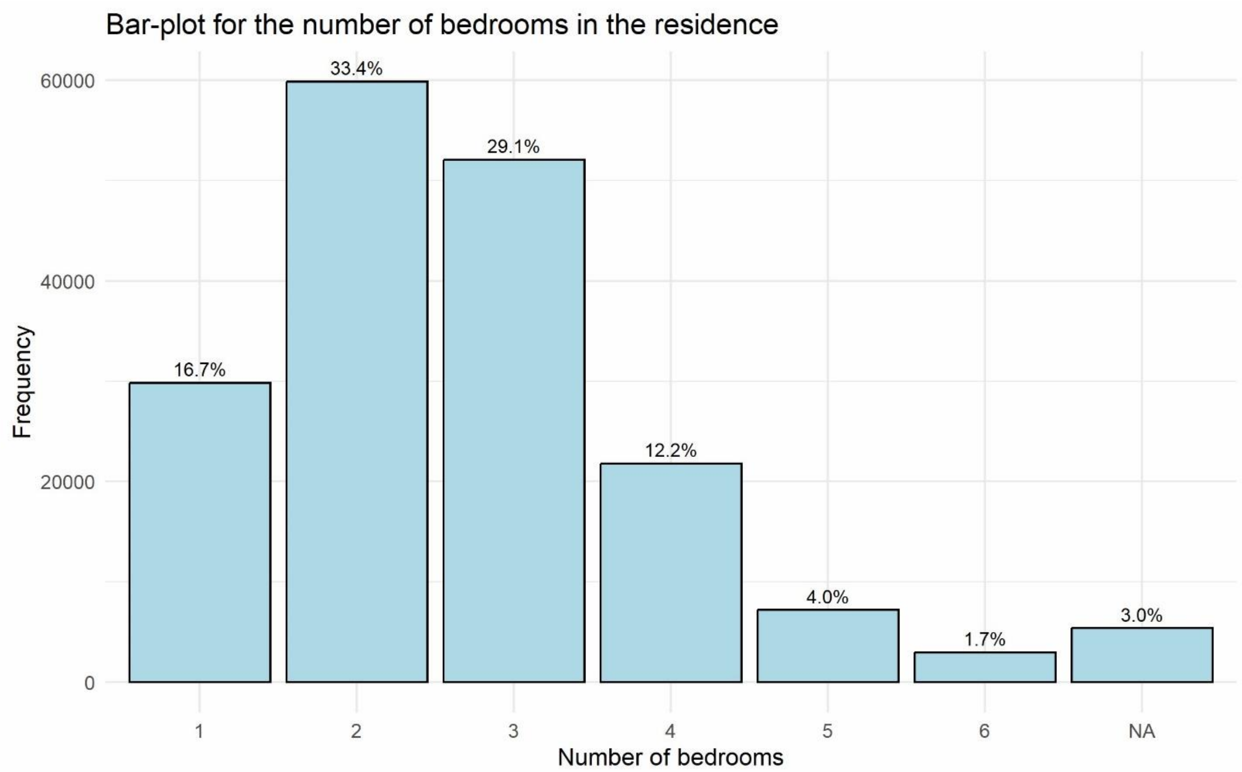

Figure 6: Bar-plot for the number of bedrooms in the residence

https://doi.org/10.26219/heal.aueb.6011

We continue with the number of bathrooms that the residence has. Again, the number of bathrooms is a variable that takes a few distinct numeric values, and for this reason, we will treat it as a factor. Upon consulting Figure 7 and Table 8, it becomes evident that the majority of houses have only one bathroom. Furthermore, as the number of bathrooms increases, the percentage of houses decreases. Finally, 6% of owners refrain from giving the number of bathrooms.

Index	Number of bathrooms									
	1	2	3	4	5	6	7	8	9	NA
Frequency	109374	38529	14161	4763	1052	370	32	11	8	10805
Percentage	61.07%	21.51%	7.91%	2.66%	0.59%	0.21%	0.02%	0.01%	0.01%	6.03%

Table 8: Basic measures for the number of bathrooms in the residence

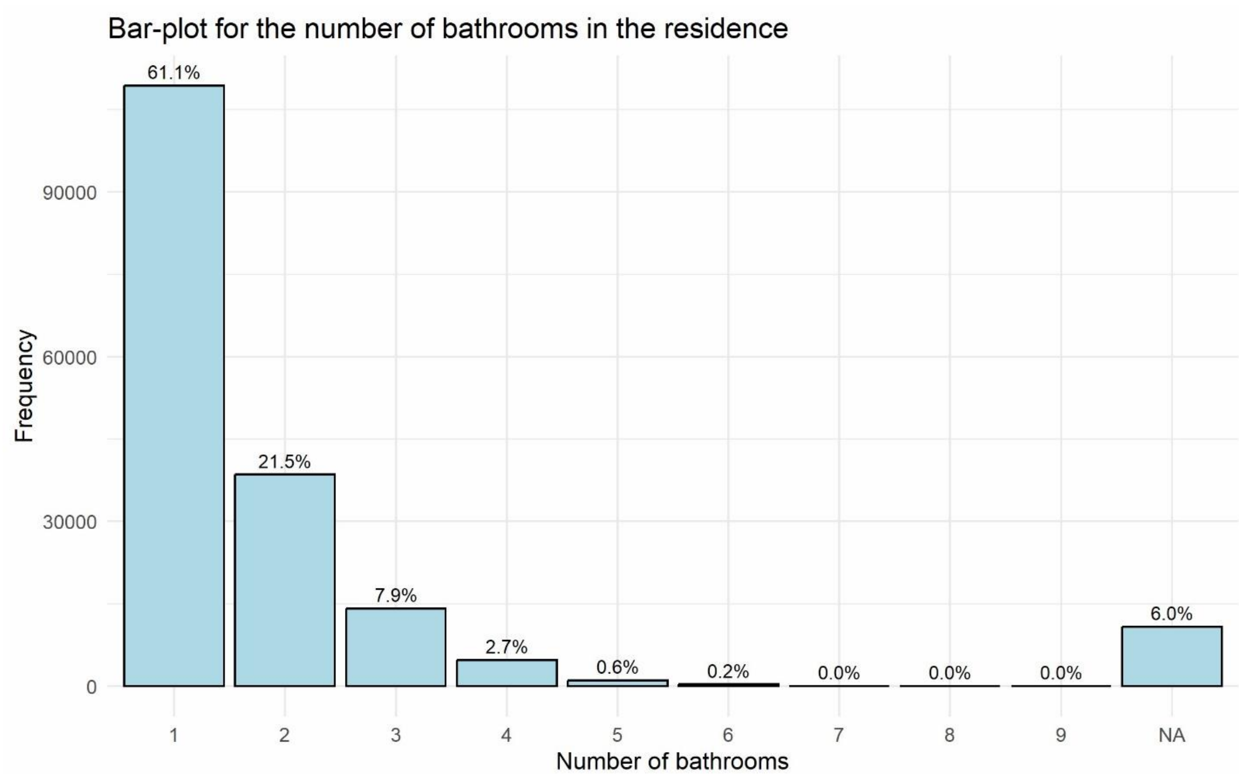

Figure 7: Bar-plot for the number of bathrooms in the residence

https://doi.org/10.26219/heal.aueb.6011

One very important variable included in our dataset is the quality of construction. This variable holds significant importance for potential buyers as it provides insights into the quality of construction, a crucial factor in making informed purchasing decisions. On the other hand, the owner of the property is the one who decides what value the variable will take. Here arises a significant problem: the inherent bias of property owners. In general, owners may tend to provide positive evaluations of their properties to secure a higher selling price. The variable has four levels: bad, medium, good, very good. Now, by having a look at Table 9 and Figure 8 we can spot two interesting points. The first one is that most of the owners didn't put a value at all, this shows us a negative predisposition to the variable. The second notable point is the absence of properties categorized as "medium" quality, which raises questions. Later on, we will discuss how will treat this variable.

Index	Quality of construction				
	Bad	Medium	Good	Very Good	NA
Frequency	12563	0	16316	31296	118930
Percentage	7%	0%	9.1%	17.5%	66.4%

Table 9: Basic measures for the quality of construction of the residence

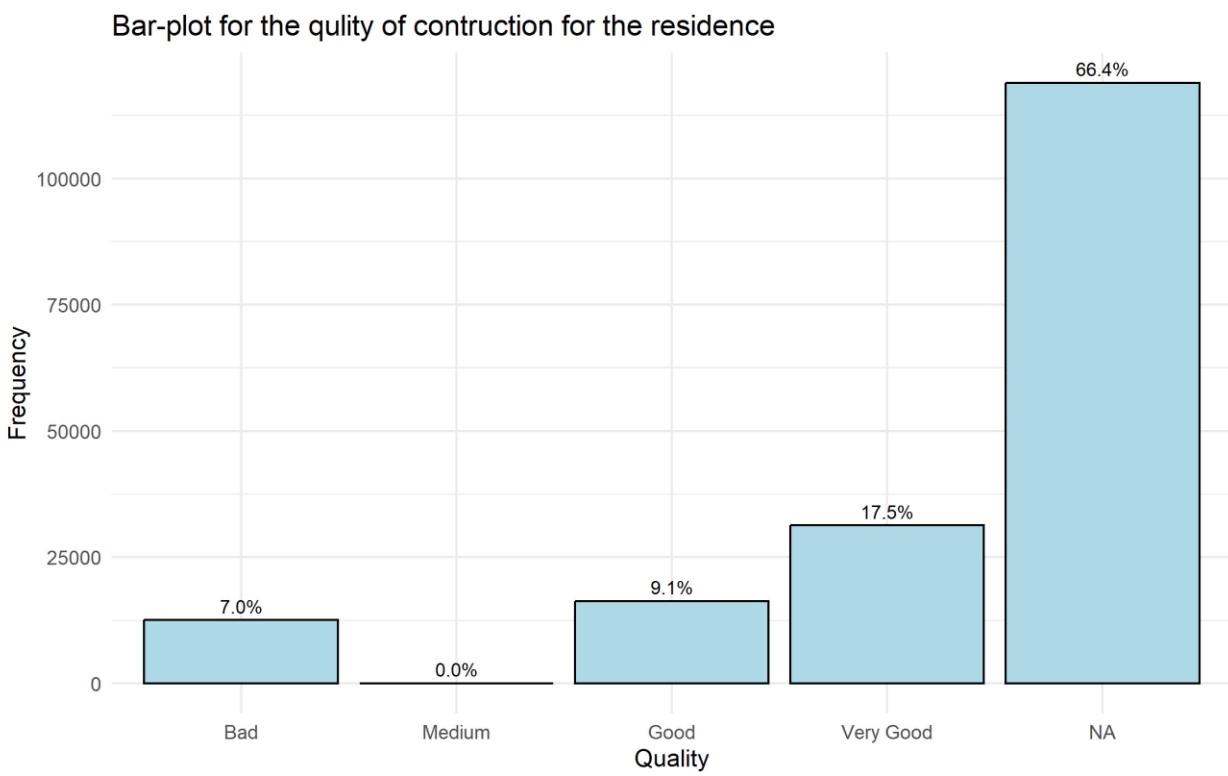

Figure 8: Bar-plot for the quality of construction of the residence

https://doi.org/10.26219/heal.aueb.6011

We detect similar problems in the condition of maintenance. Once again, the condition of maintenance offers valuable insights for potential buyers, providing a glimpse into the overall state of the property. However, as indicated by Figure 9 and Table 10, a significant number of the owners have not provided any information on the condition of maintenance. Moreover, none of the owners selected "medium" as the condition for their properties. Finally, it is important to note that this variable will undergo modifications in later analyses to deal with the problems we observed.

Index	Quality of construction				
	Bad	Medium	Good	Very Good	NA
Frequency	11759	0	55043	5971	106332
Percentage	6.6%	0%	30.7%	3.3%	59.4%

Table 10: Basic measures for the condition of maintenance of the residence

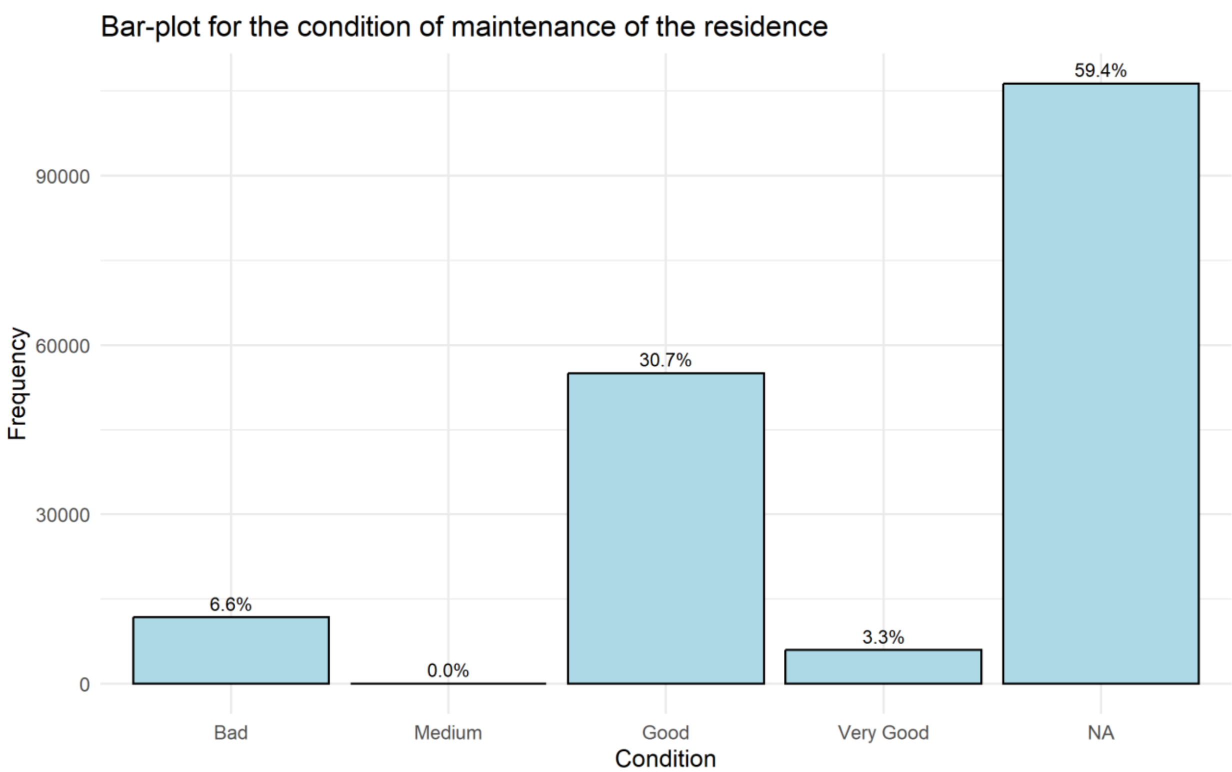

Figure 9: Bar-plot for the condition of maintenance of the residence

 https://doi.org/10.26219/heal.aueb.6011

Next, we will have a look at four yes/no variables. The first is for the existence of parking, the second for the existence of view, the third for the existence of elevator, and the last one for the existence of storage. As indicated by the first Bar-plot there are no residencies with "No" as an answer to the variable concerning parking. The absence of "No" responses to the parking variable suggests that owners without parking may have chosen to omit this variable from consideration. In the second Bar-plot, we spot that most of the houses have a view, although the meaning of having a view is very obscure, and owners "tend" to overestimate how good the residence is overall. Finally, the elevator variable exhibits a notable number of NA values, leading again to the conclusion that if a property lacks an elevator the variable is omitted.

	Yes%	Frequency Yes	No%	Frequency No	NA%	Frequency NA
Parking	23.5%	42104	0%	0	76.5%	137001
View	22.2%	39764	58%	103916	19.8%	35425
Elevator	36%	64448	30.6%	54757	33.4%	59900
Storage	27.9%	50039	54.4%	97432	17.7%	31634

Table 11: Basic measures for the existence of parking, view, elevator, and storage

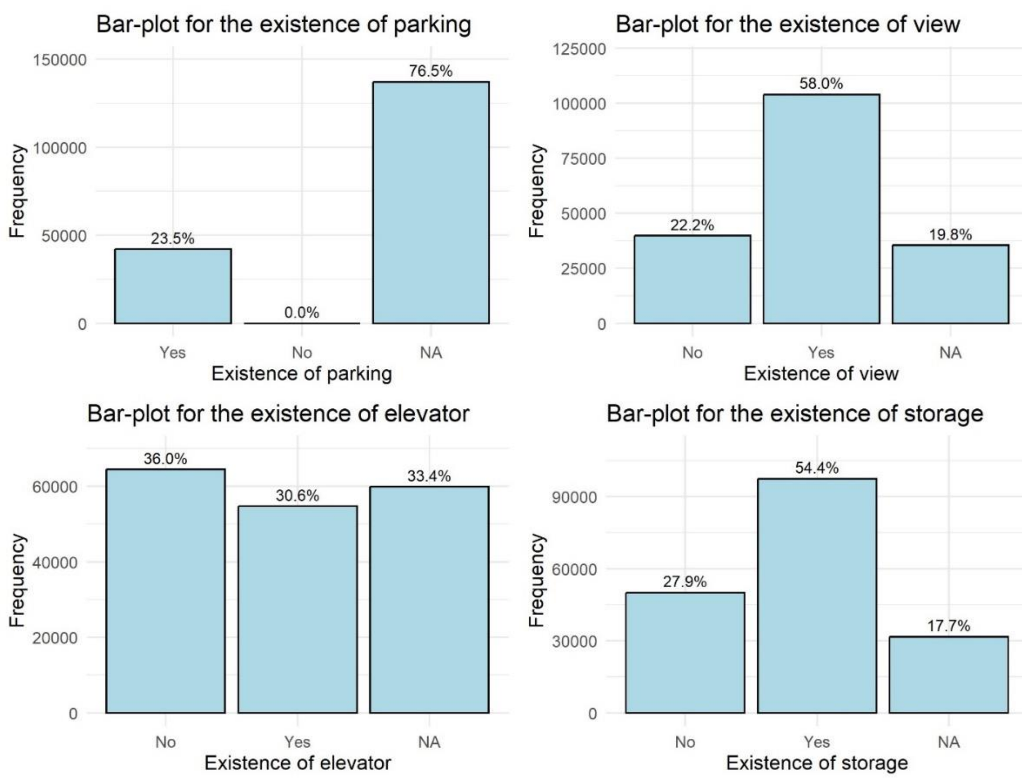

Figure 10: Bar-plot for the existence of parking, view, elevator, and storage

https://doi.org/10.26219/heal.aueb.6011

In our dataset, we are provided with the X and Y coordinates of each property. Of course, there is no meaning in looking at them separately or calculating measures like the mean and variance. Instead by utilizing the X and Y, we create a geographical representation of property locations on the map of Greece. Figure 11 is the map of Greece and we have added red dots for each property. Notably, a concentration of properties is observed in major cities such as Athens and Thessalonica.

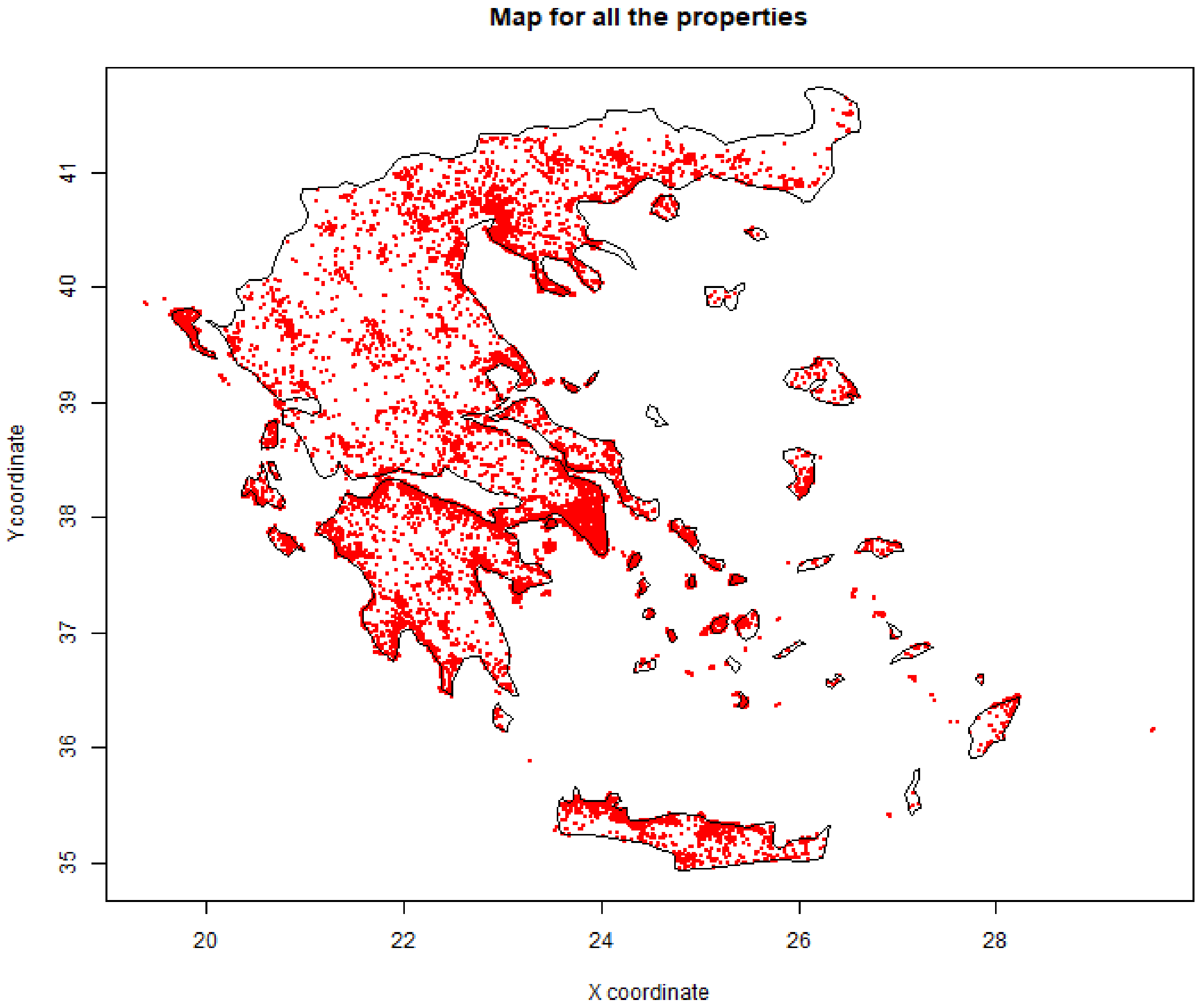

Figure 11: Map with the X and Y coordinates for each property

Finally, there are some variables we haven't talked about. Those are the region name, the municipality name, the neighborhood name, the 3-digit code referring to the location of the house, the ID, the land area, the total number of floors for the properties, the type of parking, and the town plan. The first four variables refer to the location of the house, but in our analysis, we will use only the region name and the 3-digit code. The most frequent region is "Athens Center", and the most frequent 3-digit code is "113". We have decided not to use the 'municipality name' and 'neighborhood name' variables due to their lack of reliability, which could potentially impact our models. Additionally, the rest of the variables we mentioned above, won't be used in our analysis because there is a significant number of missing information about them.

https://doi.org/10.26219/heal.aueb.6011

The next step is to explore the bivariate relationships between each variable and the property prices.

The first relationship we will look at is the house price with the residence area. Upon observing the scatterplots in Figure 12, we noticed a non-linear relationship between house prices and usable area. However, after applying a logarithmic transformation to both variables, a clear linear relationship emerged which is also indicated by the red line. The Pearson correlation for the logarithmic transformation of the variables is calculated as 0.72, surpassing the threshold of 0.7, where according to the empirical law of Chatfield and Collins (1980, p. 40-41), this suggests a strong linear relationship between house prices and residence area. Finally, the Pearson correlation index is positive which means that as the logarithm of the residence area increases, the logarithm of the price also increases.

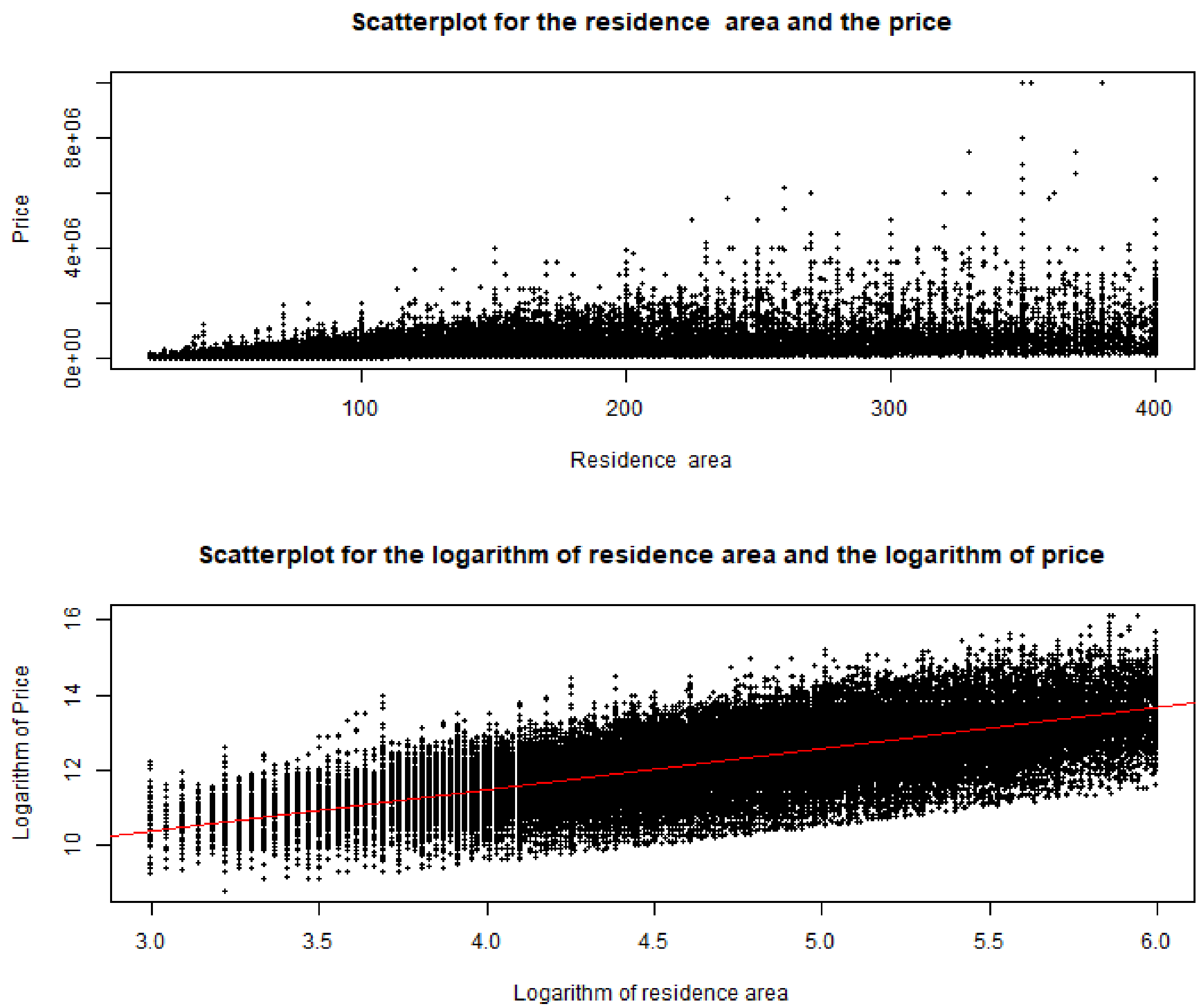

Figure 12: Scatterplots for residence area and the price

 https://doi.org/10.26219/heal.aueb.6011

The next relationship we will examine is between the year of construction and the property price. In Figure 13 we see that the relationship is not linear, but when we apply the logarithm function to the price there is a faint linear relationship. It is evident that the relationship between the year and price is not linear. However, applying the logarithm function to the price reveals a faint linear relationship, which is illustrated by the red line in the second graph. The Pearson correlation index of 0.36 indicates a weak positive linear relationship according to Chatfield and Collins (1980, p. 40-41), suggesting that as the year of construction increases, the logarithm of the price also increases.

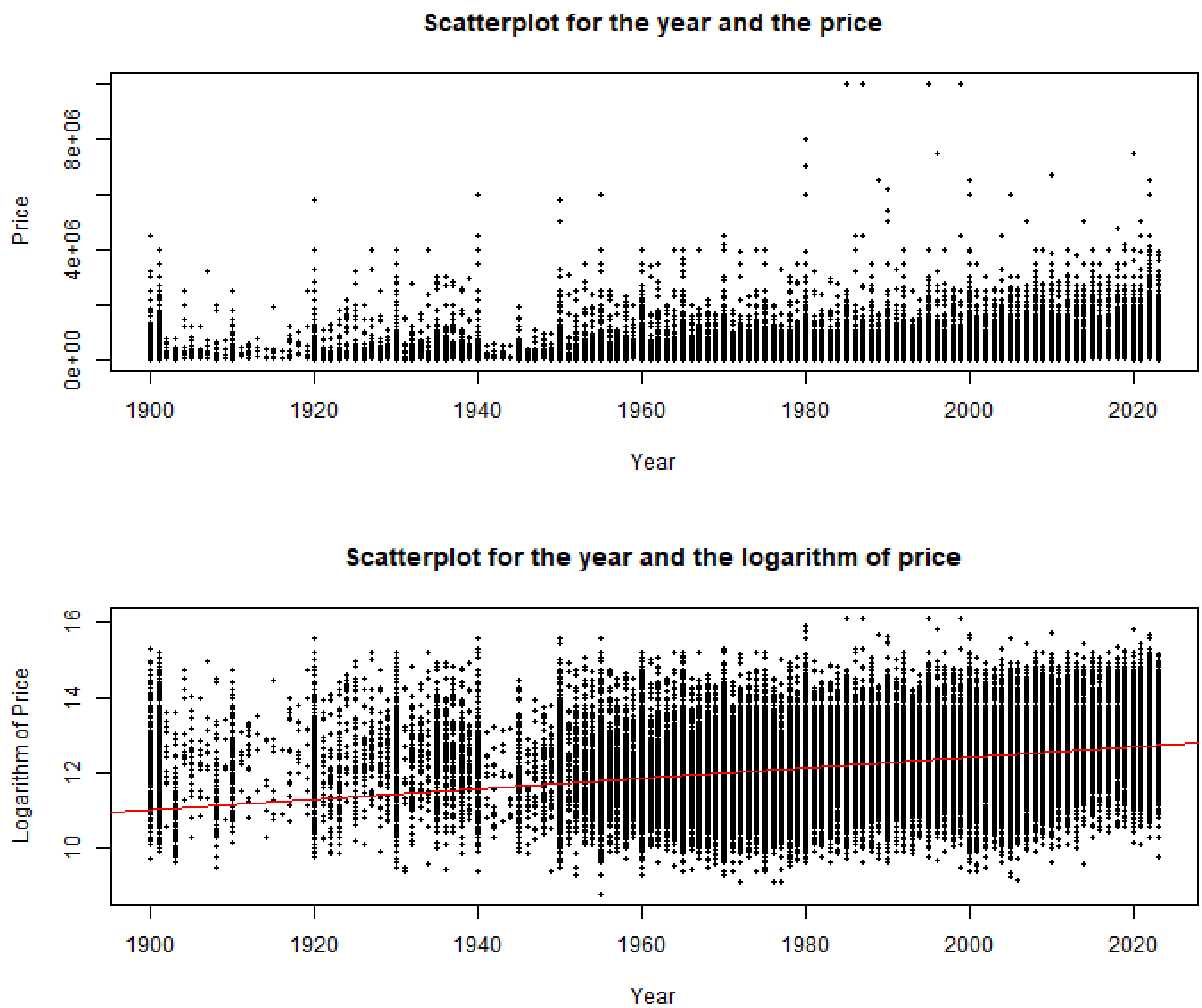

Figure 13: Scatterplots for the year of construction and the price

 https://doi.org/10.26219/heal.aueb.6011

The relationship between the price and the type of residence is illustrated in Figure 14 where we used the logarithm of the price instead of the original scale. We chose the logarithmic transformation because it enhances the interpretability of the figures by "compressing" the price values (this tactic will be used also for the rest of the variables). Table 12 provides basic measures regarding the box-plots of Figure 14 but in the original scale. The notches in Figure 14 are the 95% intervals for the median and in Table 12 we specify the upper and lower limits for this interval in the original scale. We can statistically assess differences in the median prices for the different types of residences by using the Kruskal-Wallis test. The test gives p-value=2.2e-16, this suggests strong evidence against the null hypothesis (for a=5%) that all medians are equal. On the other hand, it's crucial to consider the practical significance of these differences. By examining the medians in Table 12 we can see that Duplexes and Flats have similar prices, but they differ significantly from Houses and Maisonettes. Furthermore, Houses and Maisonettes also exhibit a notable difference. From the above, we can conclude that there is some meaningful difference in the median for the different types of residences. Finally, when we look at the minimum values, we spot very low prices. We made this point at the start of this section when we looked at the histogram of the prices and pointed out that there are some unexpectedly low prices, these low prices might be properties that don't correspond to any of the types we have in our dataset. This can simply happen because the site does not have options for other types of property, or those properties might be extreme outliers.

	Type of residence			
	Duplex	Flat	House	Maisonette
Minimum	12500€	9000€	6600€	20000€
1st Quantile	80000€	95000€	130000€	240000€
Median	145000€	158000€	250000€	375000€
3^{d} Quantile	259500€	269000€	480000€	600000€
Maximum	4000000€	8000000€	10000000€	10000000€
Lower Lim.	137878€	157181€	247626€	371917€
Upper Lim.	152489€	158823€	252396€	378108€

Table 12: Basic measures for the box-plot of type of residence and price

https://doi.org/10.26219/heal.aueb.6011

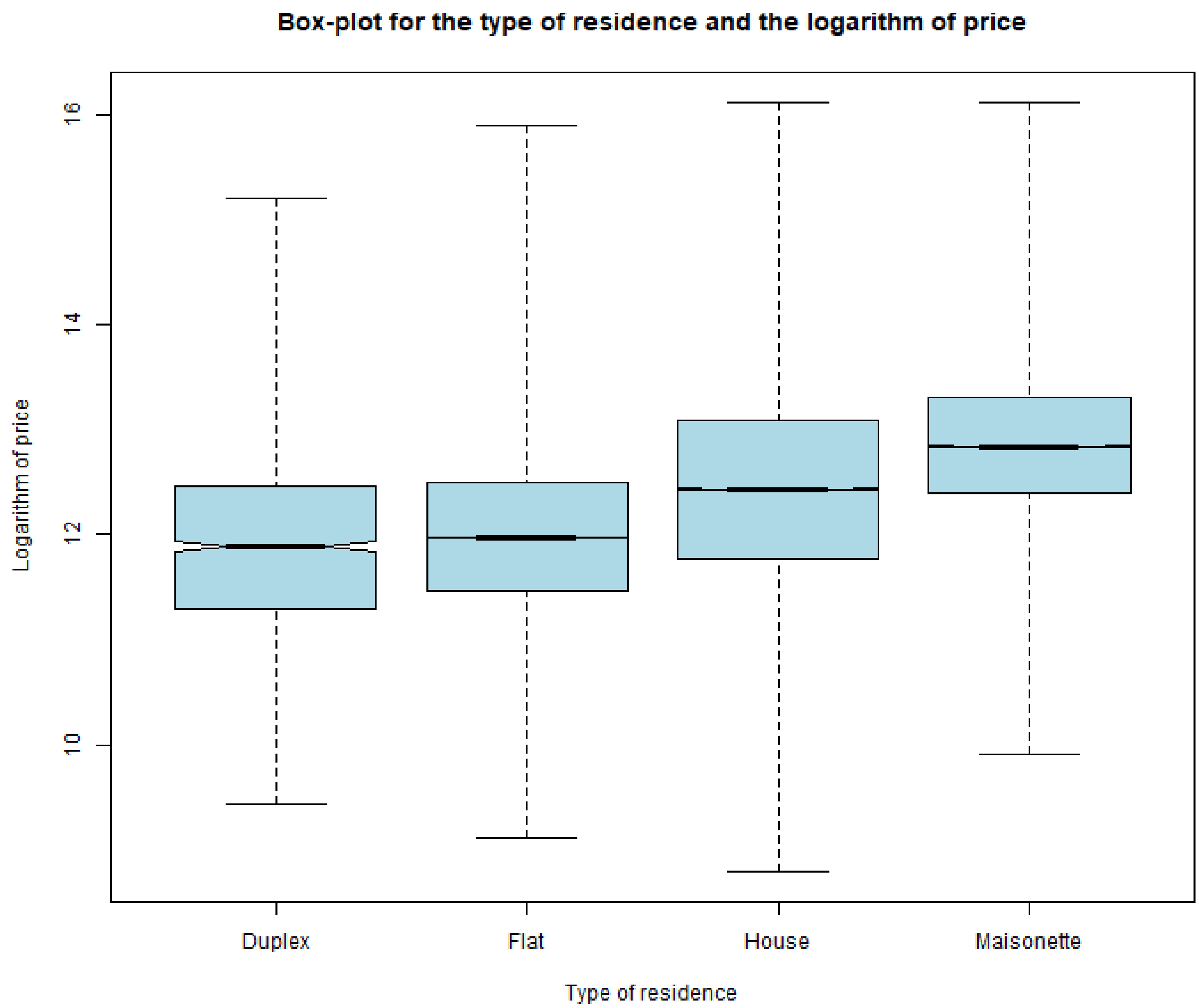

Figure 14: Box-plot for the type of residence and the logarithm of price

Next, we will have a look at the relationship between the floor and the price of the residence. When we examine Figure 15 for floors 1 to 8, we observe an upward trend in the median which also seems to be linear. Also, residences that are on floor -1 exhibit a vastly lower median compared to residences on other floors. On the other hand, residences on floor 0 show a very high median price. This can be attributed to the fact that most of the Houses are on floor 0 and as we saw in the previous figure Houses tend to have higher median prices. When we apply the Kruskal-Wallis test to check the null hypothesis that all the medians are statistically the same we get p-value=2.2e-16, which provides enough evidence against the null hypothesis (for a=5%). In Table 13 and for floors 1 to 8, we see a steady rise in the median and although neighboring floors might not have big differences, floors that are far from each other do have notable differences.

https://doi.org/10.26219/heal.aueb.6011

	Floor of residence									
In thousands of euros	-1	0	1	2	3	4	5	6	7	8
Minimum	9	6.6	9.5	14.3	15	15	17.6	21	14	27
1st Quantile	40	110	100	115	120	135	150	160	170	245
Median	65	230	160	185	200	220	250	270	300	350
3d Quantile	160	435	280	330	365	370	398	450	450	550
Maximum	10000	10000	10000	10000	10000	6200	6200	3900	2250	2600
Lower Lim.	62.8	227.8	158.6	182.9	197.3	216.7	245.7	262.9	289.3	334.3
Upper Lim.	67.2	232.1	161.3	187	202.7	223.3	254.3	277.2	311	366.3

Table 13: Basic measures for the box-plot of floor of residence and price in thousands of euros

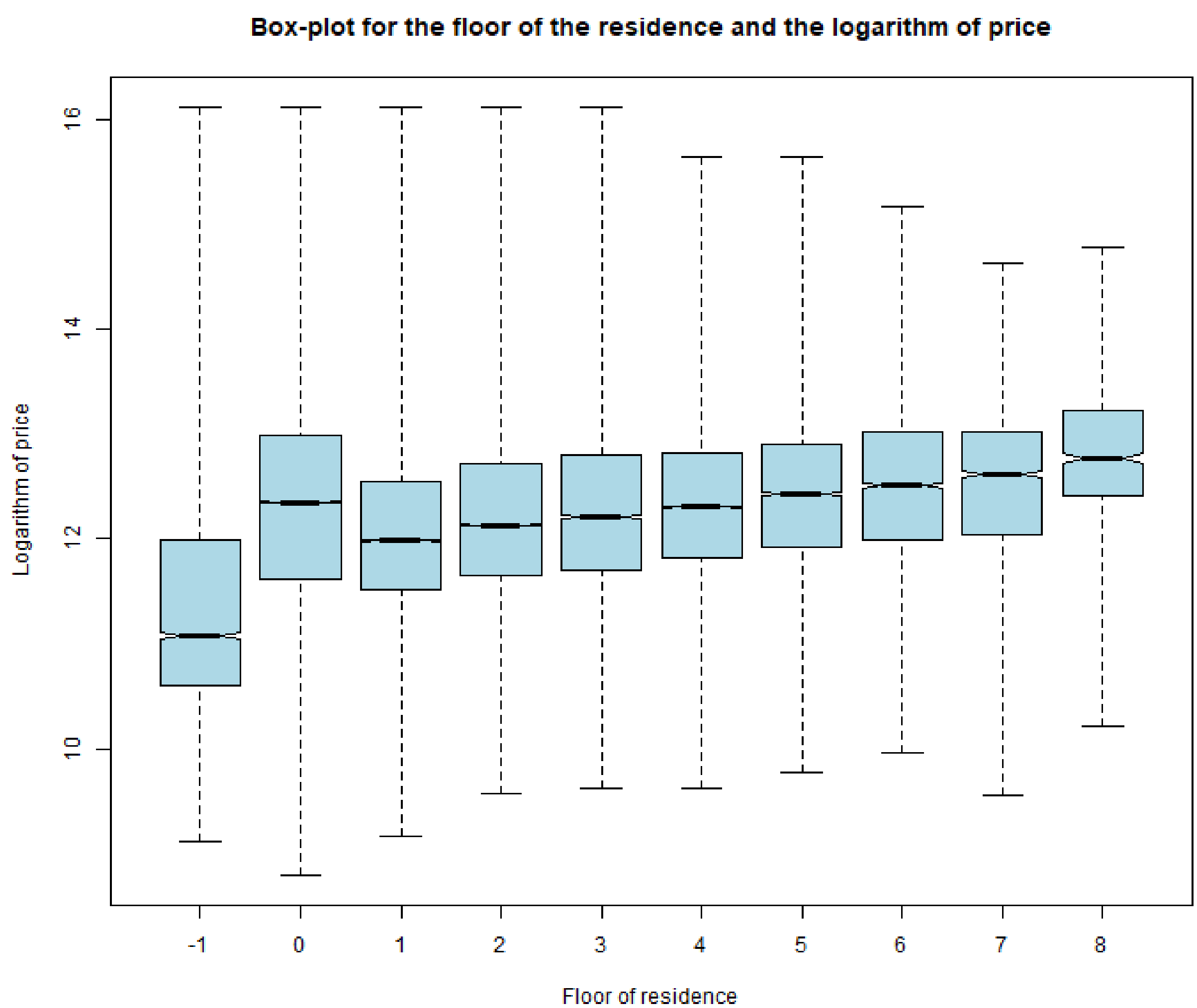

Figure 15: Box-plot for the type of floor and the logarithm of price

https://doi.org/10.26219/heal.aueb.6011

The next relationship is the one between the number of bedrooms of the residence and the price. From Figure 16 we conclude that as the number of bedrooms increases the median price increases too, but the relationship doesn't seem to be linear. The Kruskal Wallis test for the null hypothesis of equal medians gives p-value=2.2e-16, thus we have enough evidence (for a=5%) against the null hypothesis. When we examine the medians in Table 14 we confirm that the medians for the number of bedrooms seem to have notable differences.

	Number of bedrooms					
	1	2	3	4	5	6
Minimum	6600€	11000€	15000€	30000€	37000€	29900€
1st Quantile	62000€	100000€	180000€	260000€	340000€	340000€
Median	90000€	155000€	290000€	430000€	500000€	530000€
3^{d} Quantile	140000€	250000€	450000€	700000€	800000€	850000€
Maximum	2500000€	4000000€	7500000€	8000000€	10000000€	10000000€
Lower Lim.	89331€	154085€	288166€	425467€	492091€	516089€
Upper Lim.	90673€	155919€	291845€	434580€	508035€	544285€

Table 14: Basic measures for the box-plot of the number of bedrooms and price

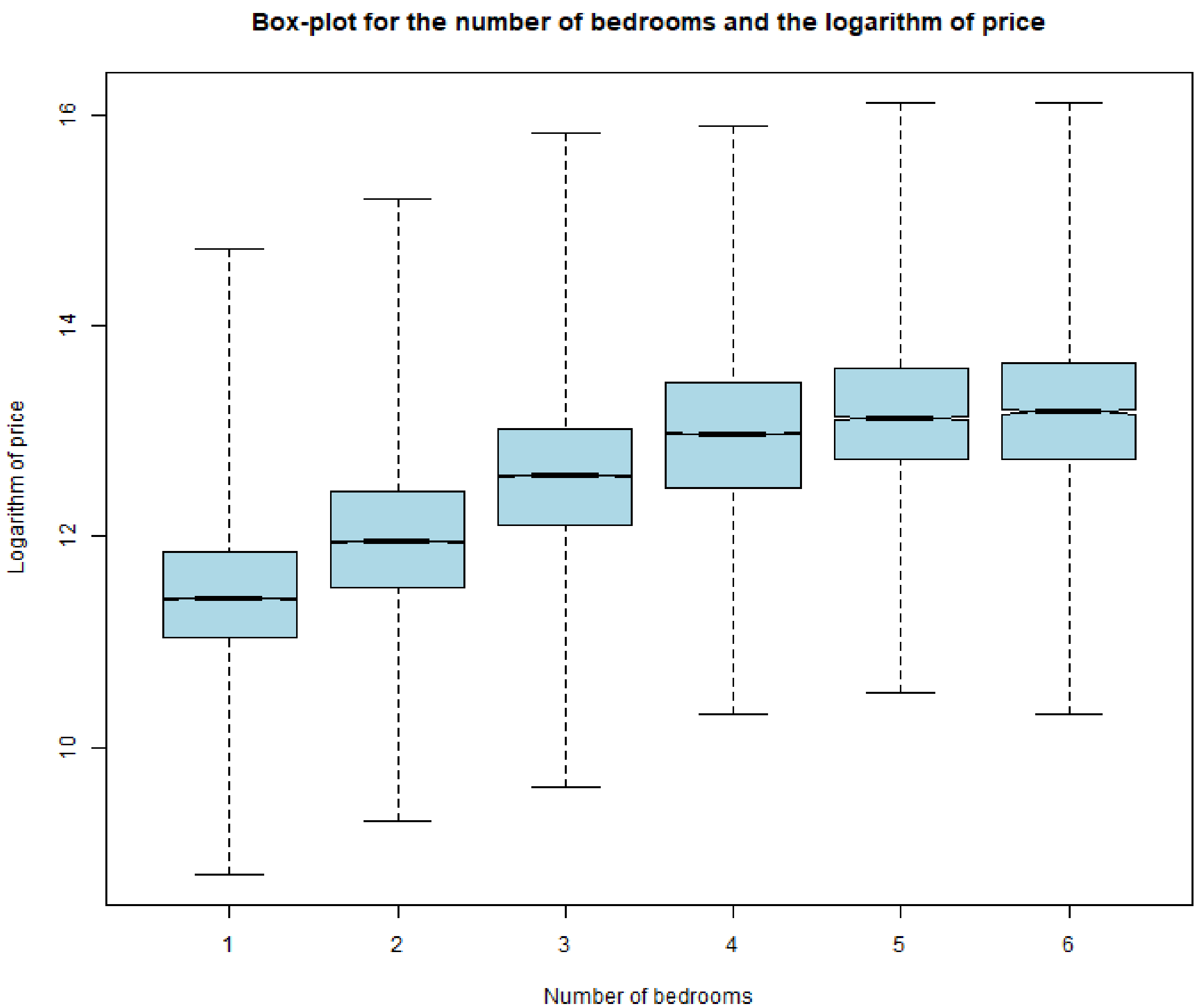

Figure 16: Box-plot for the number of bedrooms and the logarithm of the price

Another interesting relationship is the one between the number of bathrooms and the price of the residence. Upon observing Figure 17 we can see that there is an increase in the median price up to the point of 5 bathrooms. Beyond 5 bathrooms, there appears to be little change in the median price. This can be validated by Table 15 where the median values for residences with 5,6,7 and 8 bathrooms are nearly the same. The Kruskal-Wallis test for the null hypothesis of equal medians gives p-value=2.2e-16, so we can reject the null hypothesis (for a=5%) that all medians are the same. We have to be careful because the Kruskal-Wallis test looks if all the medians are the same, this means that if just two of the medians are different and all the others are statistically the same then the test will be rejected. In our case, we saw that for more than 5 bathrooms the median has a very small change. Also, by looking at the 95% intervals we can see that for residences with more than 5 bathrooms, the

https://doi.org/10.26219/heal.aueb.6011

intervals overlap with each other, which points to the conclusion that the medians are statistically the same. However, we must point out that the intervals for residences with 7, 8, and 9 bathrooms are not very trustworthy cause of the small sample. In conclusion, we saw some significant changes in the median price as the number of bathrooms increases but these differences become non-significant after a point.

	Number of bathrooms								
In thousands of euros	1	2	3	4	5	6	7	8	9
Minimum	9	22	33	40	70	80	50	26	115
1st Quantile	90	215	350	450	550	550	639.9	744.7	390
Median	150	340	550	700	850	900	894.4	900	670.8
3d Quantile	250	520	850	1100	1400	1400	1449.1	1140.1	1500
Maximum	10000	7500	7000	8000	10000	4500	3000	2100	1900
Lower Lim.	149.2	337.5	543.5	685.8	812.1	833513	711.8	734.7	316
Upper Lim.	150.7	342.4	556.5	714.4	889.5	971789	1123.8	1102.4	1423.6

Table 15: Basic measures for the box-plot of the number of bathrooms and price in thousands of euros

https://doi.org/10.26219/heal.aueb.6011

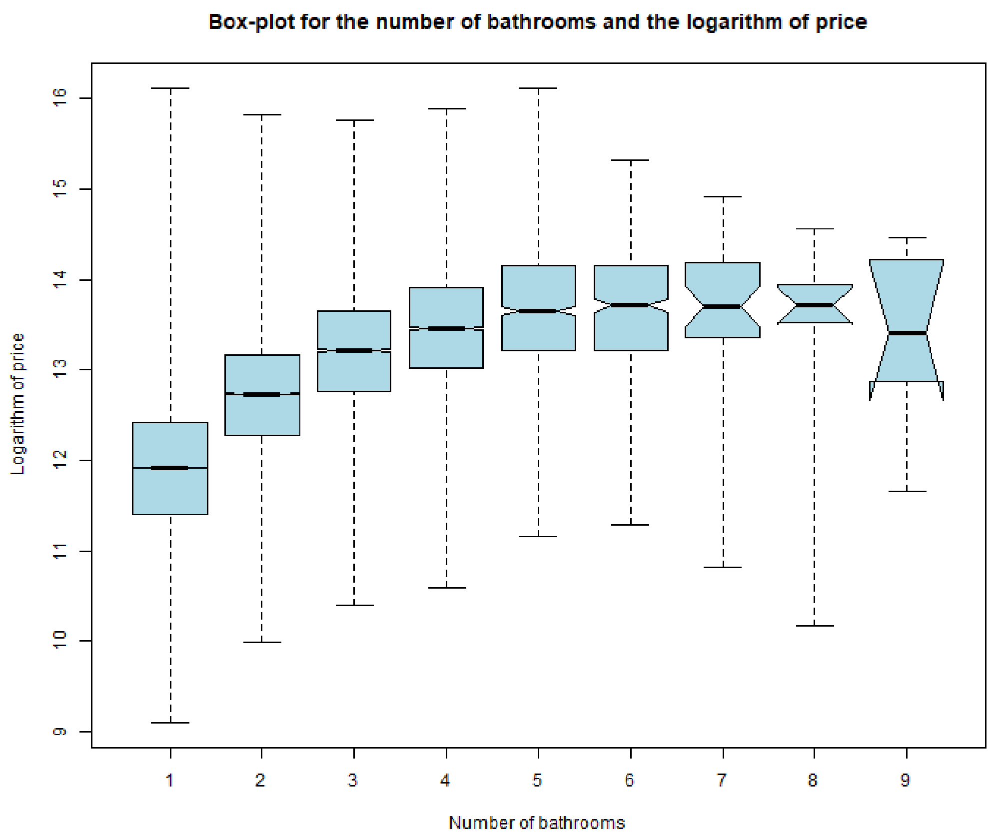

Figure 17: Box-plot for the number of bathrooms and the logarithm of the price

We continue with the relationship between the quality of construction and the price of the property. Figure 18 does not have a boxplot for the quality level “medium”, because as we said earlier owners didn’t use this level for any of the properties thus it’s not depicted here. Moreover, we see that as the quality of construction gets better the median price rises, when we do the Kruskal-Wallis test, where the null hypothesis is the equality of the medians for the different qualities of construction, we get p-value=2.2e-16. This provides us with enough evidence (for a=5%) to reject the null hypothesis. Furthermore, by having a look at the medians in Table 16 can conclude that there is a huge difference between the median price for each of the levels of quality.

https://doi.org/10.26219/heal.aueb.6011

	Quality of construction		
	Bad	Good	Very good
Minimum	6600€	12000€	10400€
1st Quantile	65000€	110000€	185000€
Median	105000€	200000€	310000€
3d Quantile	195000€	360000€	530000€
Maximum	4000000€	10000000€	6500000€
Lower Lim.	103386€	197088€	307099€
Upper Lim.	106638€	202954€	312927€

Table 16: Basic measures for the box-plot for the quality of construction and price

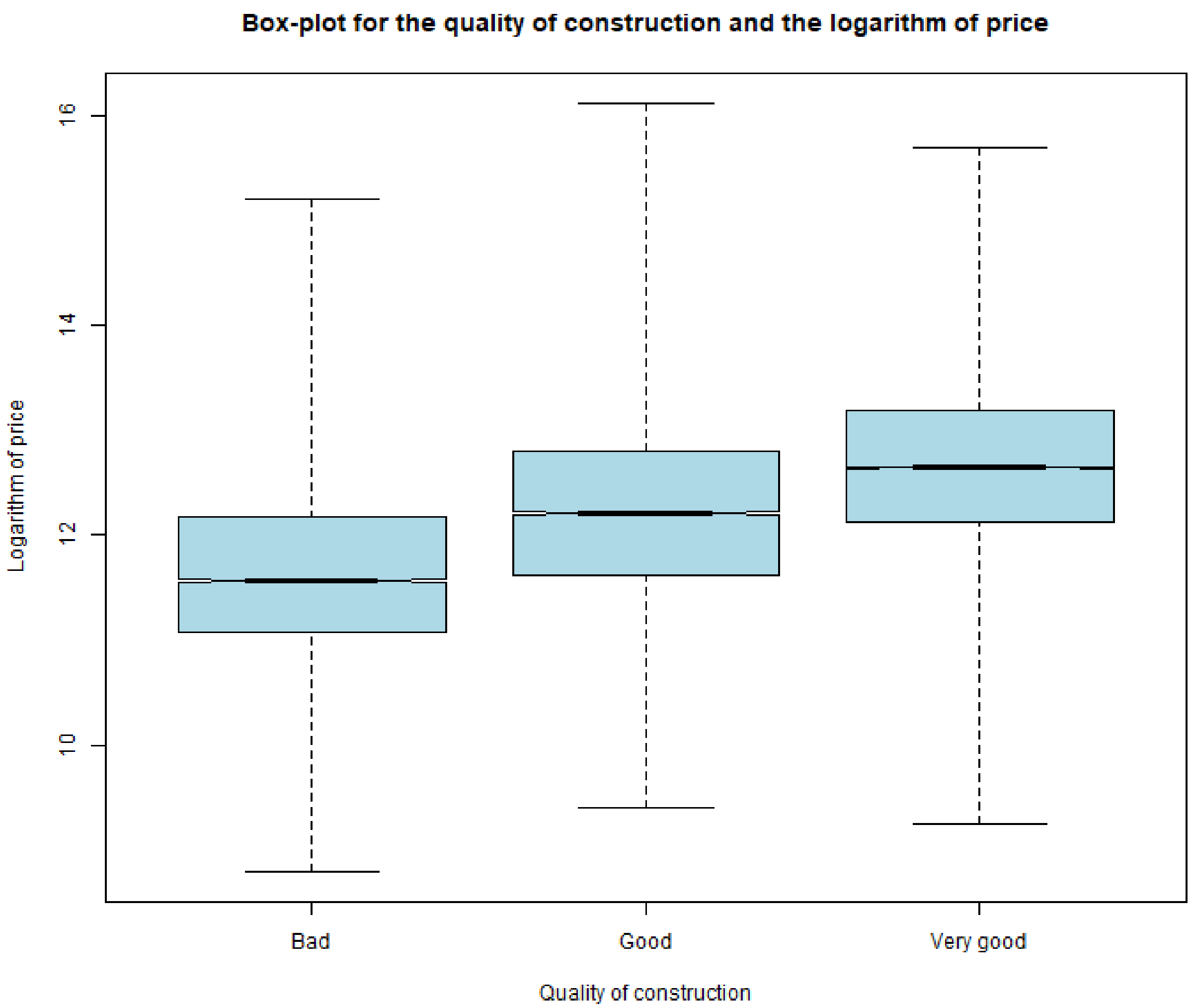

Figure 18: Box-plot for the quality of construction and the logarithm of price

https://doi.org/10.26219/heal.aueb.6011

Next is the relationship between the condition of maintenance and the price of the property. Similarly, with the quality of construction, the condition of maintenance is missing the category of "medium". In Figure 19 we see that as the condition of maintenance becomes better, the median price rises. Moreover, when we apply the Kruskal-Wallis test we get p-value=2.2e-16 and thus we have enough evidence against the null hypothesis that all the median prices are the same. Observing Table 17 we can confirm that there is also a practical difference between the median prices for the different conditions.

	Condition of maintenance		
	Bad	Good	Very good
Minimum	6600€	9000€	13500€
1st Quantile	64000€	110000€	160000€
Median	100000€	180000€	285000€
3d Quantile	190000€	320000€	500000€
Maximum	4000000€	1000000€	5000000€
Lower Lim.	98427€	178710€	278436€
Upper Lim.	101598€	181299€	291717€

Table 17: Basic measures for the box-plot for the condition of maintenance and price

https://doi.org/10.26219/heal.aueb.6011

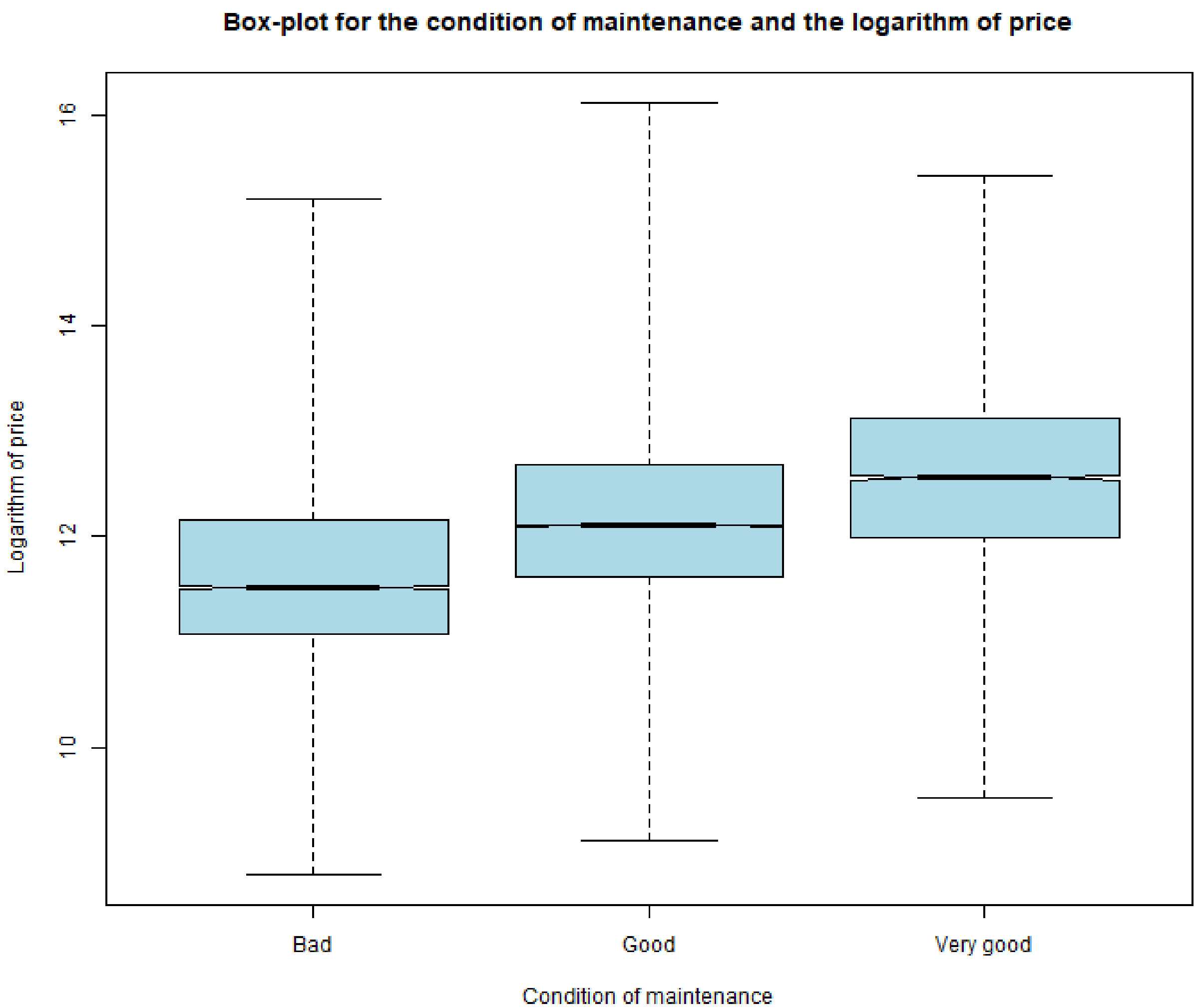

Figure 19: Box-plot for the condition of maintenance and the logarithm of price

The last relationships we will discuss are the existence of an elevator, view, and storage with the price. By having a quick look at Figure 20 we see that for all three variables when there is elevator, view, and storage the price is higher. According to the Kruskal-Wallis test, we get p-value=2.2e-16 for all the variables, which is enough evidence (for a=5%) to reject the null hypothesis of equal medians. Moreover, in Table 18 we confirm that there is an actual difference between the median price when there is an elevator, view, or storage and when there is not.

	Existence of elevator		Existence of view		Existence of storage	
	No	Yes	No	Yes	No	Yes
Minimum	6600€	14000€	9000€	11000€	6600€	13000€
1st Quantile	100000€	130000€	90000€	139000€	85000€	160000€
Median	180000€	225000€	157150€	250000€	140000€	280000€
3d Quantile	350000€	401826€	290000€	440000€	246000€	490000€
Maximum	10000000€	8000000€	5000000€	10000000€	7500000€	10000000€
Lower Lim.	178602€	223292€	155700€	248592€	138953€	278418€
Upper Lim.	181408€	226721€	158613€	251416€	141054€	281590€

Table 18: *Basic measures for the box-plots of the existence of elevator, view, and storage with the price*

https://doi.org/10.26219/heal.aueb.6011

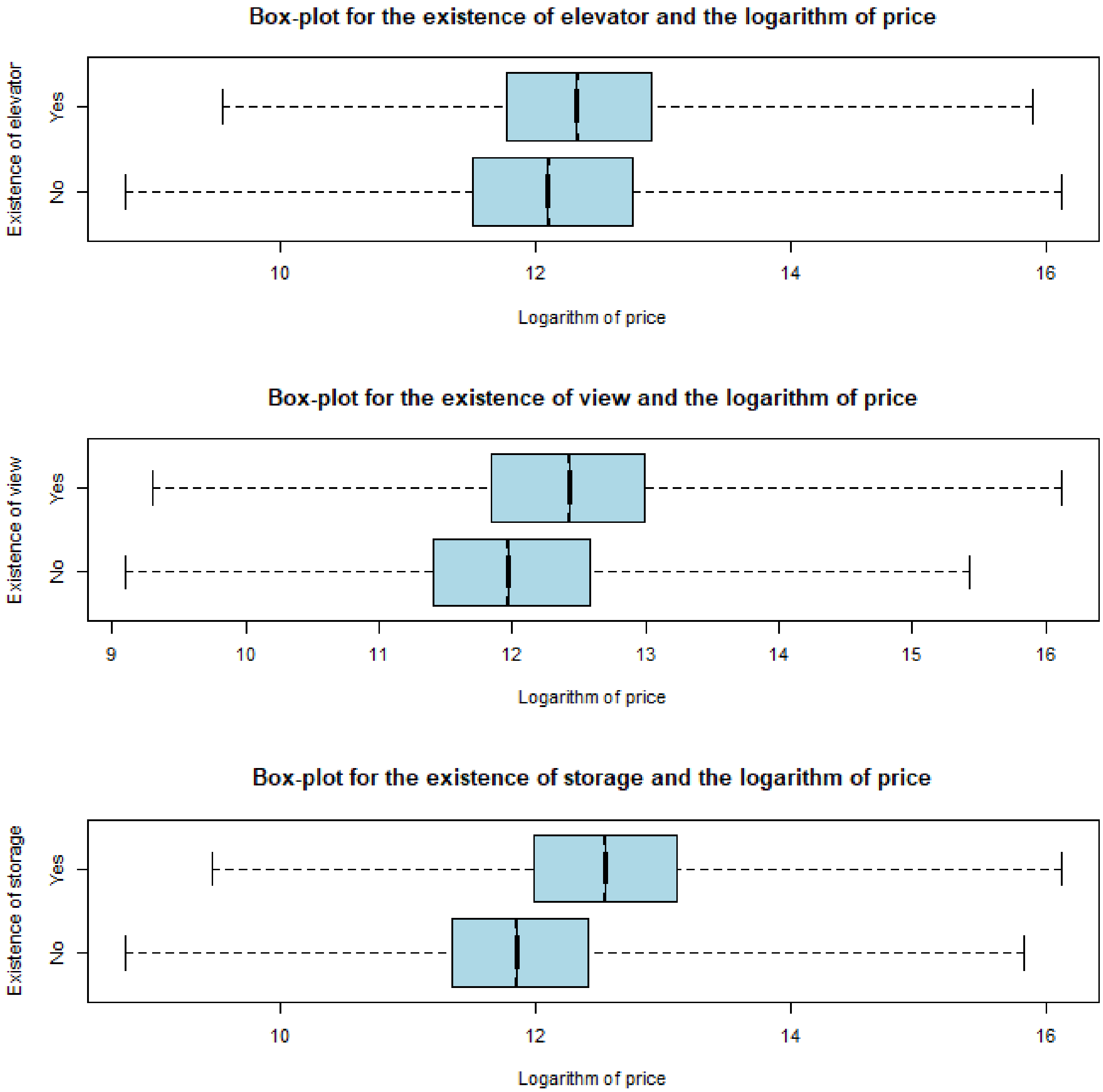

Figure 20: Box-plots for the existence of elevator, view, and storage with the logarithm of price

https://doi.org/10.26219/heal.aueb.6011

In this last part of the section, we will deal with the missing values we found in our variables. To be more specific, the variables we will fill in the missing values are:

1. The existence of elevator
2. The existence of parking
3. The existence of storage
4. The existence of view
5. The number of bedrooms
6. The number of bathrooms
7. The quality of construction
8. The condition of maintenance

For the first four variables (existence of elevator, parking, storage, and view) we filled the missing values with "No" which means the absence of the respective characteristic. This decision is grounded in the assumption that property owners usually would mention those characteristics if they are present to boost the price of the property.

The missing values for the number of bedrooms were filled in the following way:

1. Split properties in groups according to their usable residence area.
2. Find in each group how many and which properties have missing values for the number of bedrooms.
3. Sample with replacement as many times as needed in each group.
4. The sampled values fill in the missing values.

The missing values for the number of bathrooms were filled in the same way as the number of bedrooms.

Finally, for the quality of construction and the condition of maintenance, we decided that we will change the levels from "Bad", "Medium", "Good", and "Very Good" to "Bad" and "Good". The reasoning was that the level "Medium" was never mentioned in our data. The properties that were categorized as "Good" and "Very Good" changed to "Good" and the properties categorized as "Bad" didn't change. Then we proceed with filling in the missing values:

1. Split properties in groups according to their date of construction.
2. Find in each group how many and which properties have a missing value for the quality of construction/condition of maintenance.
3. Sample with replacement as many times as needed in each group.
4. The sampled values fill in the missing values.

https://doi.org/10.26219/heal.aueb.6011

https://doi.org/10.26219/heal.aueb.6011

5) PREDICTIVE MODELS

In this section, our focus will be to find the best prediction model for the price of properties. We will use different types of models and different ways to include information about the location of the property in the models. Our goal is to find the model which has the lowest out-of-sample. This means that the models will be "trained" in a part of our data and then we will test the efficiency of the models in a different part of the data (test data). Both the train and test data are subsets of the original dataset. Finally, we will not use all the data given to us, instead, we will only model the properties with a price of less than a million euros. The reasoning behind this decision is that properties with very high prices usually belong to a non-competitive market and thus the models are not suitable to predict the price of those properties.

The types of models we are going to use are:

1. Linear modeling
2. LASSO modeling
3. LAD modeling
4. Random Forests

For all the models the response variable is the logarithm of the price. We use the logarithm of the price because most of the types of models we are going to use are linear. As observed earlier, all the continuous variables have a linear relationship with the logarithm of the price. Thus, modeling the logarithm of the price will improve the predictive ability of the linear models.

Our variable selection process is guided by the earlier explanatory analysis. This decision is rooted in the sheer size of the dataset, where conventional selection methods like AIC might inadvertently include irrelevant variables. To mitigate this, we rely on insights gained from our explanatory analysis. Finally, the variables we used for all the models are:

1. The type of the property.
2. The logarithm of the usable area.
3. The year of construction of the property.
4. The floor the property is located.
5. The existence of parking.
6. The quality of construction.

https://doi.org/10.26219/heal.aueb.6011

7. The number of bedrooms (with categories 1,2,3,4+).
8. The number of bathrooms (with categories 1,2,3,4+).
9. The existence of elevator.
10. The existence of view.
11. The existence of storage.
12. A categorical variable that indicates if the property is at -1 floor or 0 floor or any other floor.

Also, we added the location variable to the models in four different ways:

1. Model without location variable (to act as the control).
2. The name of the region where the property is located. We changed the region variable by adding the "region" Other, which is composed of the ten regions with the lowest population of properties. The reason we did change this variable is because Random Forest models in R accept categorical variables with at most 53 levels.
3. The 3-digit number, which is based on the location of the property.
4. A model was constructed for each property in the test set, based on the one thousand nearest houses of each property (not for LASSO because of the high computational burden).

Finally, we used two methods to measure how well a model predicted the prices:

1. The mean of the absolute percentage difference between the price given in the dataset and the predicted price by the model.
2. The test RMSE.

In Table 19 we have the mean absolute percentage difference for each model. Upon examining each row individually, we note that LM, LASSO, and LAD exhibit similar performance, whereas Random Forest shows a notable improvement. However, this trend changes in the final row, where LM, LAD, and Random Forest give comparable results. Random Forest tends to outperform other models but at the expense of a significantly higher computational burden. While LM, LASSO, and LAD models take mere seconds to run, Random Forest requires almost three hours, this does not hold for the models where we used the closest 1000 properties.

Now, by focusing on each column independently, we notice a substantial decrease in the mean absolute percentage difference with the inclusion of the 'Region' variable. Furthermore, by substituting the "Region" with the 3-digit number, we are

https://doi.org/10.26219/heal.aueb.6011

led to a slight decrease in the index, while reducing the number of coefficients. By using the 'Closest 1000 properties' as the location variable, the result is a significant drop in the index, but at the cost of increased computational load. For LM, LASSO, and LAD models, the running time increases to approximately two hours, while Random Forest takes nearly three hours.

Finally, by taking the table as a whole, the LAD model with the 'Closest 1000 properties' seems to be the best-performing model based on the mean absolute percentage difference.

	LM	LASSO	LAD	Random Forest
No location	42.84%	43.03%	42.68%	40.45%
Region	33.32%	33.43%	32.53%	30.05%
3-digit number	33.2%	33.29%	32.46%	29.54%
Closest 1000	26.09%	-	25.94%	26.06%

Table 19: Mean absolute percentage difference for each model

In Table 20, we observe the RMSE for each model, expressed in euros. While presenting the error in a tangible currency like euro is intuitive, it's essential to note that RMSE can be misleading. For instance, an RMSE of 132860€ might seem high for properties priced at 150000€ but relatively low for properties priced at 900000€.

An intriguing finding is that when using Random Forest models, the RMSE is better when incorporating the 'Region' or '3-digit number' variables compared to using the 'Closest 1000 properties' which is different based on what we saw for the mean absolute percentage difference.

Finally, the model with the lowest RMSE is the LM model when utilizing the 'Closest 1000 properties' as the location variable. This suggests that, despite its simplicity, the LM model performs well in terms of minimizing prediction error based on RMSE.

	LM	LASSO	LAD	Random Forest
No location	132860€	132914€	132626€	125663€
Region	114414€	114271€	115213€	101425€
3-digit number	114315€	113955€	114635€	100840€
Closest 1000	92263€	-	93096€	119734€

Table 20: RMSE for each model

https://doi.org/10.26219/heal.aueb.6011

To summarize our findings, the most effective models utilize the 'Closest 1000 properties' as the location variable. It's crucial to note that this approach allows for a dynamic model that adjusts based on the location of the test property. While this versatility is advantageous, it comes with a computational cost, particularly when dealing with numerous test properties. However, this challenge can be addressed through hardware upgrades or optimization of coding techniques. Additionally, we have to keep in mind that although the results may not be optimal, the dataset we are provided has many problems as seen earlier. Moreover, we model the asking price, which means the owner is the one setting the price, which might be very far away from the real market value of the property.

6) NEGOTIATION MARGIN

In this section, our focus will be on studying the negotiation margin. Negotiation margin can be defined as the difference between the price the owner asks and the market value of the properties (later on we will refer to it as difference). A second definition we will use is the ratio between the price the owner asks, and market value (later on we will refer to it as ratio). However, a challenge we encounter is the absence of the actual market values for the properties. To address this issue, we collaborated with a bank to obtain predicted market values using an Automated Valuation Model (AVM). Firstly, we will analyze the negotiation margin for all properties across Greece, and then we will narrow our focus to the three prefectures with the highest property count according to our data and examine them individually.

When we study the first of the two definitions, we get that the mean is 58867€ and the median is 34030€, thus we expect right skewness which is confirmed by the skewness index and Figure 21. The advantage of presenting the negotiation margin as the difference between the asking price and the predicted market value is that we get all information translated into euros which is very easy to comprehend. On the other hand, this approach is flawed because, if for example a property is priced at 70000€ according to the median we can get 34030€ off the price, which is a considerable amount, and it might be far from the truth. For this reason, we will also use the ratio of the asking price with the predicted market value which essentially gives us the percentage discount. By looking at Table 21 we see the mean discount is 36% percent of the asking price while the median discount is 25% of the asking price. Again, according to Table 21 and Figure 22, we spot the right skewness.

	mean	median	SD	skew	kurt
Difference	58867€	34030€	102499€	1.79	6.24
Ratio	1.36	1.25	0.56	2.87	23.37

Table 21: Basic measures for Negotiation Margin

https://doi.org/10.26219/heal.aueb.6011

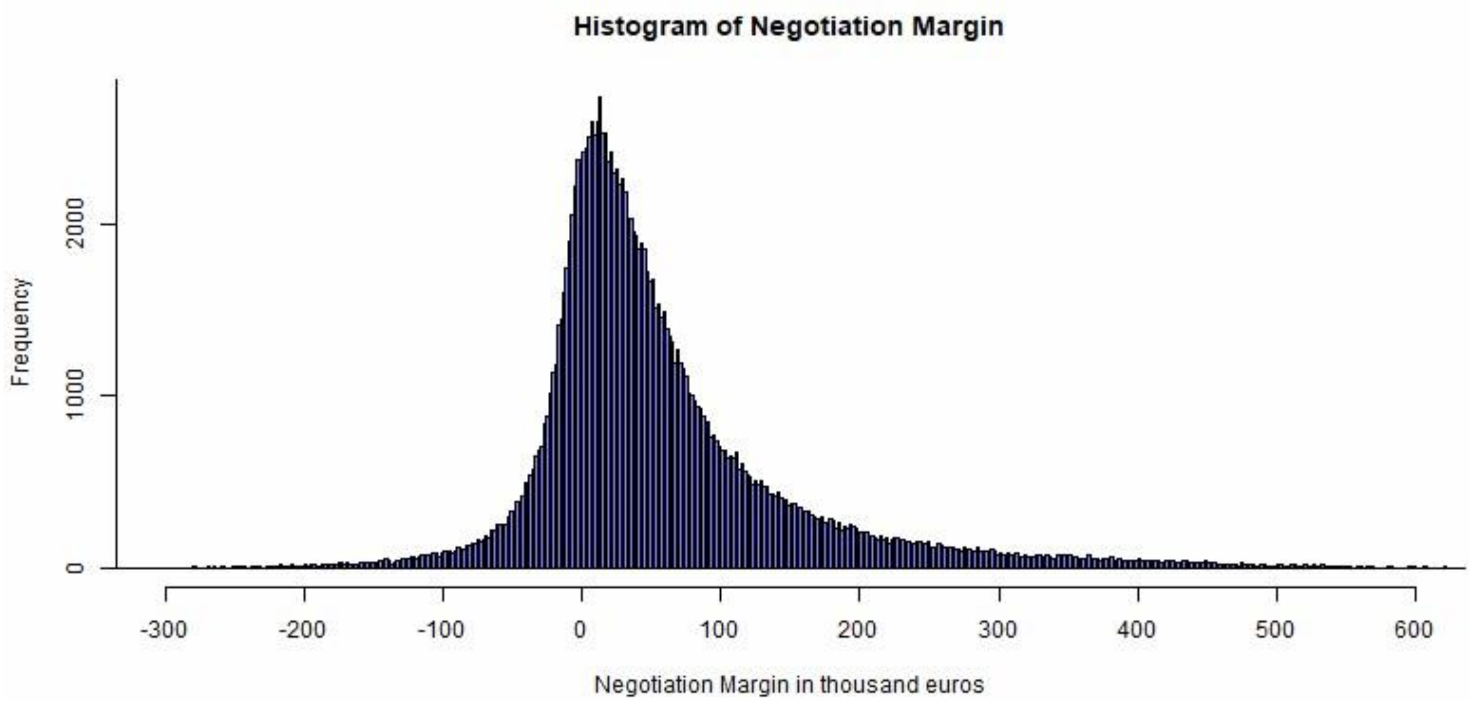

Figure 21: Histogram of Negotiation Margin (difference)

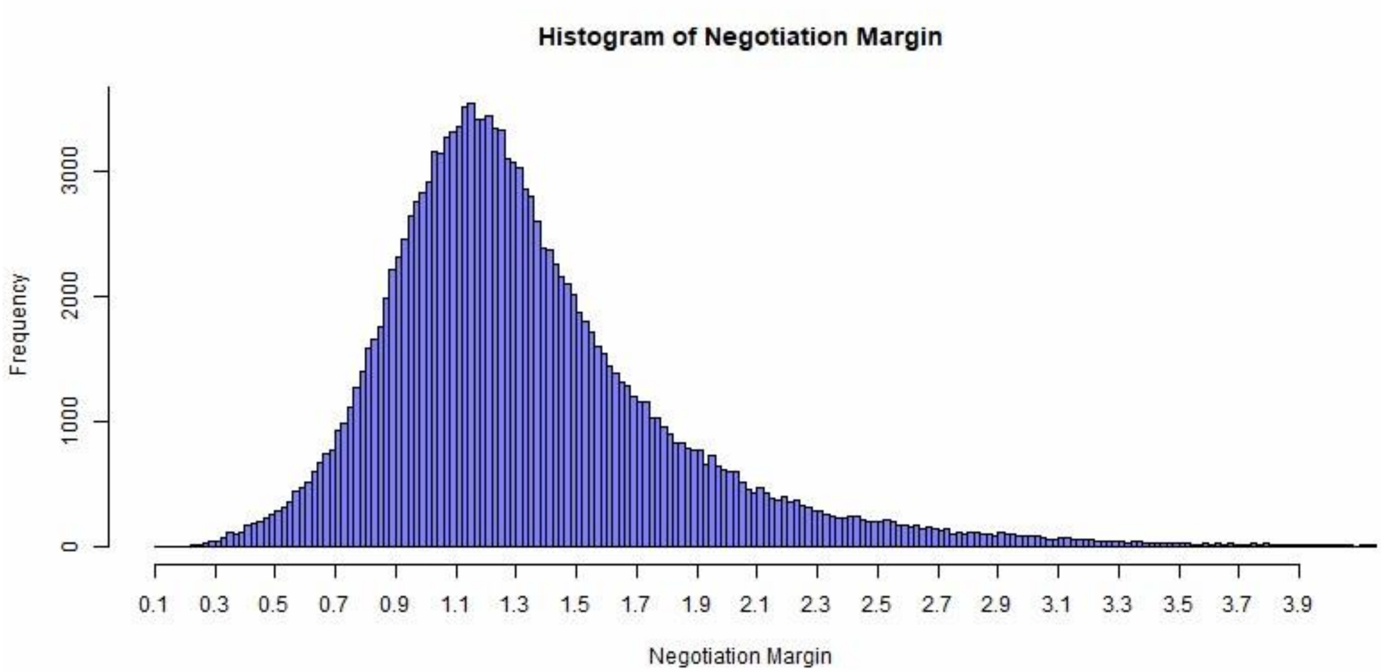

Figure 22: Histogram of Negotiation Margin (ratio)

https://doi.org/10.26219/heal.aueb.6011

Next, we will study the relationship between the negotiation margin and the type of property. In Figure 23 we present the logarithm of the negotiation margin (ratio) for each type of property. According to Table 22, the medians seem to be very close for all the types of properties. Although when we apply the Kruskal-Wallis test for the equality of the medians (null hypothesis), for a=5% we reject the null hypothesis, thus there is a significant statistical difference. But as we said earlier this does not necessarily imply that the differences are impactful. Again, by taking a closer look at Table 22 we can come to the conclusion that Houses tend to have bigger negotiation margins and Maisonettes smaller. Finally, according to Figure 23, we see that Flats have a bigger variance in the negotiation margin than the rest of the types.

	Type of property			
	Duplex	Flat	House	Maisonette
Median	1.246	1.247	1.268	1.23
Lower Lim.	1.225	1.244	1.260	1.225
Upper Lim.	1.267	1.249	1.276	1.235

Table 22: Basic measures of Negotiation Margin for each type of property

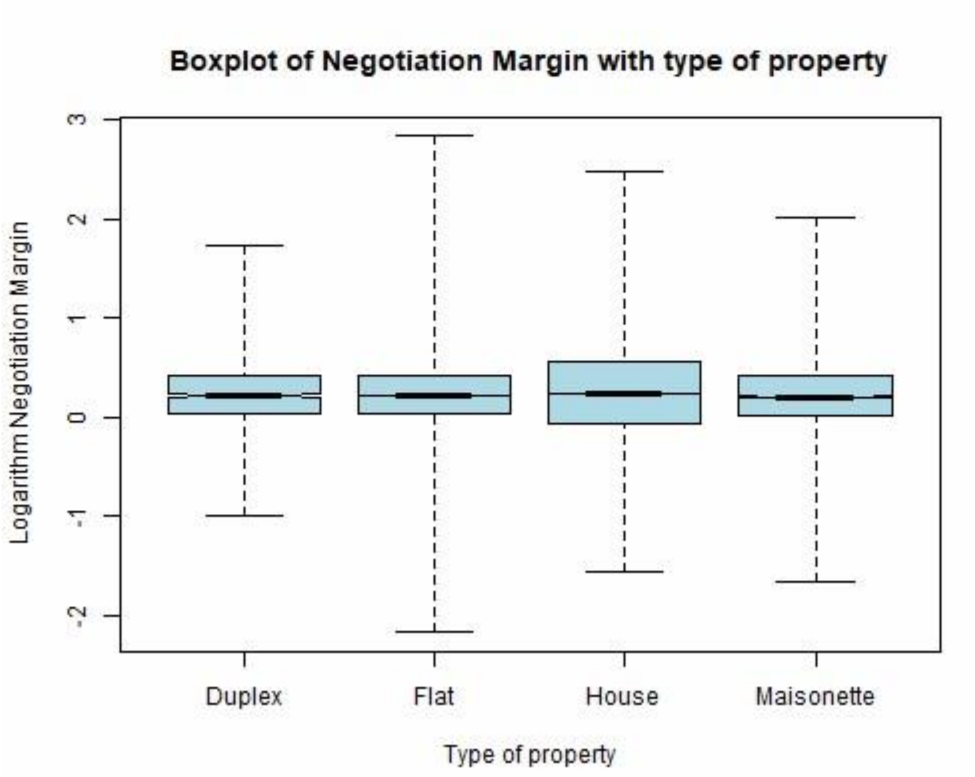

Figure 23:Boxplot for Negotiation Margin and type of property

Another interesting relationship we will study is the one concerning the negotiation margin and the year of construction. According to Table 23, the median negotiation margin becomes smaller as the year of construction is closer to the present. This can also be seen in the Figure 24. Applying the Kruskal Wallis test for the equality of the medians (null hypothesis), for a=5%, we reject the null hypothesis, thus there is a significant statistical difference. Finally, the variance seems to be the same for all the years.

	Year of construction				
	1920-1940	1941-1960	1961-1980	1981-2000	2001-2023
Median	1.327	1.267	1.278	1.253	1.219
Lower Lim.	1.298	1.253	1.274	1.248	1.216
Upper Lim.	1.356	1.281	1.283	1.258	1.221

Table 23: Basic measures of Negotiation Margin based on the year of construction

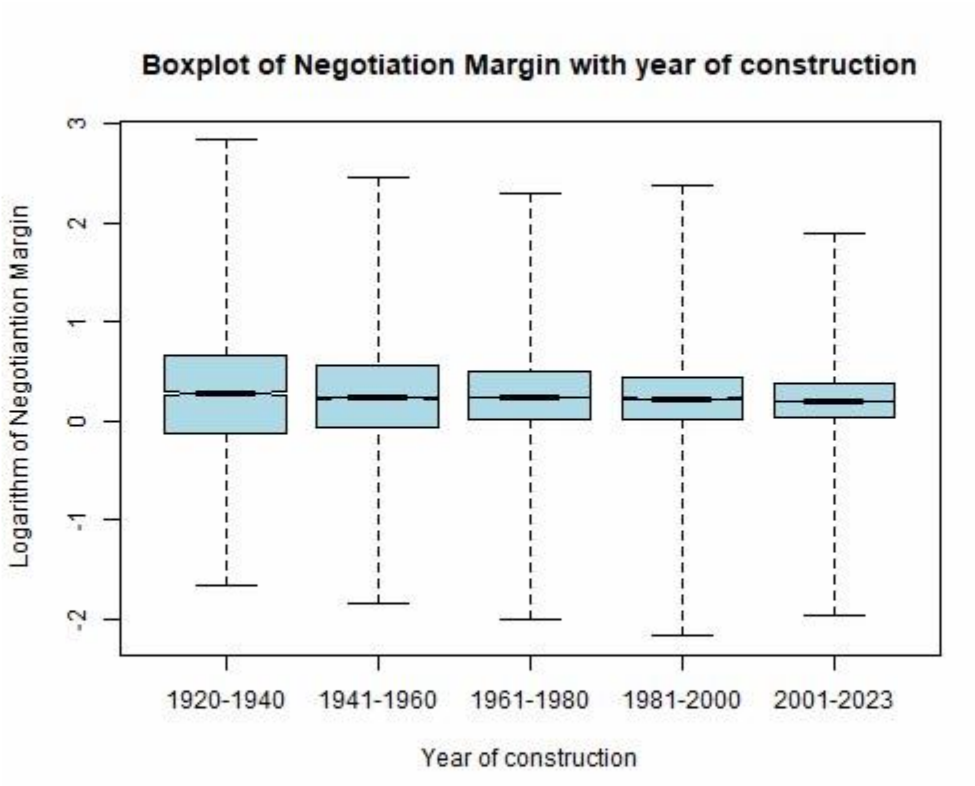

Figure 24:Boxplot for Negotiation Margin and year of construction

https://doi.org/10.26219/heal.aueb.6011

We will continue the analysis of the negotiation margin by constructing explanatory models. The difference between explanatory models and predictive models is that we don't want to optimize the prediction, but we want to construct a model to understand how the different characteristics of the property impact the negotiation margin. The response variable we are going to use is the logarithm of the ratio between the price the owner asks and the predicted market value. The reason we added the logarithm was because without it the constant variance assumption of the linear model is violated. Moreover, the normality assumption is violated thus we are unable the normal confidence intervals. For this reason, we constructed confidence intervals for the betas by using a bootstrap technique. In Table 24 we present predicted betas and the 95% confidence intervals. Finally, the way we choose which of the variables will be on the model is based on the BIC criterion.

Variable	**Prediction**	**Prediction (exponent)**	**Lower Limit (exponent)**	**Upper Limit (exponent)**
Intercept	0.063	1.065	1.051	1.08
Flat	-0.078	0.924	0.918	0.93
Duplex	-0.067	0.934	0.918	0.952
Maisonette	-0.03	0.969	0.963	0.976
Logarithm of usable area	-0.226	0.797	0.791	0.801
Year of construction	-0.002	0.997	0.996	0.998
Ground floor	0.165	1.179	1.165	1.193
Other floor	0.173	1.189	1.175	1.203
Parking	0.03	1.03	1.025	1.035
Quality	0.012	1.012	1.001	1.017
2 Bedrooms	0.039	1.04	1.034	1.046
3 Bedrooms	0.089	1.093	1.085	1.101
4+ Bedrooms	0.092	1.096	1.086	1.108
Elevator	0.059	1.061	1.057	1.065
View	0.082	1.085	1.081	1.09
2 Bathrooms	0.099	1.104	1.098	1.11
3 Bathrooms	0.137	1.147	1.137	1.157
4+ Bathrooms	0.146	1.157	1.142	1.173
Condition	0.158	1.171	1.164	1.177
Storage	0.031	1.032	1.028	1.036

Table 24: Exponents of the predicted betas and the 95% confidence intervals

https://doi.org/10.26219/heal.aueb.6011

Below we have the explanation for each of the predicted betas.

1. Intercept = 1.065: The expected negotiation margin for a property that is classified as House, the usable area is 70 square meters, the year of construction is 1920, it is located at -1 level, does not have parking, the quality of construction is bad, it has 1 bedroom, it has no elevator, it has no view, it contains one bathroom, the condition of maintenance is bad and has storage is 6.5% of the asking price.
2. Flat = 0.924: The expected negotiation margin for a property drops by 7.6% if it is classified as Flat instead of House, while the rest of the characteristics remain the same.
3. Duplex = 0.934: The expected negotiation margin for a property drops by 6.6% if it is classified as a Duplex instead of a House, while the rest of the characteristics remain the same.
4. Maisonette = 0.969: The expected negotiation margin for a property drops by 3.1% if it is classified as a Duplex instead of a House, while the rest of the characteristics remain the same.
5. Logarithm of usable area = $1.01^{-0.226}$= -0.22%: The expected negotiation margin for a property drops by 0.22% for a 1% increase in the usable area, while the rest of the characteristics remain the same.
6. Year of construction = 0.997: The expected negotiation margin for a property drops by 0.3% when the year of construction is raised by one.
7. Ground floor = 1.179: The expected negotiation margin for a property increases by 17.9% if is on the Ground floor instead of -1 level, while the rest of the characteristics remain the same.
8. Other floor = 1.189: The expected negotiation margin for a property increases by 18.9% if it is on any other floor except the ground floor and -1 level, while the rest of the characteristics remain the same.
9. Parking = 1.03: The expected negotiation margin for a property increases by 3% if parking is included, while the rest of the characteristics remain the same.
10. Quality = 1.012: The expected negotiation margin for a property increases by 1.2% if the quality of construction is good, while the rest of the characteristics remain the same.
11. 2 Bedrooms = 1.04: The expected negotiation margin for a property increases by 4% if it has two bedrooms instead of one, while the rest of the characteristics remain the same.

https://doi.org/10.26219/heal.aueb.6011

12. 3 Bedrooms = 1.096: The expected negotiation margin for a property increases by 9.3% if it has three bedrooms instead of one, while the rest of the characteristics remain the same.
13. 4+ Bedrooms = 1.093: The expected negotiation margin for a property increases by 9.3% if it has four or more bedrooms instead of one, while the rest of the characteristics remain the same.
14. Elevator = 1.061: The expected negotiation margin for a property increases by 6.1% if it is equipped with an elevator, while the rest of the characteristics remain the same.
15. View = 1.085: The expected negotiation margin for a property increases by 8.5% if it has a good view, while the rest of the characteristics remain the same.
16. 2 Bathrooms = 1.104: The expected negotiation margin for a property increases by 10.4% if it has two bathrooms instead of one, while the rest of the characteristics remain the same.
17. 3 Bathrooms = 1.147: The expected negotiation margin for a property increases by 14.7% if it has three bedrooms instead of one, while the rest of the characteristics remain the same.
18. 4+ Bathrooms = 1.093: The expected negotiation margin for a property increases by 9.3% if it has four or more bedrooms instead of one, while the rest of the characteristics remain the same.
19. Condition = 1.171: The expected negotiation margin for a property increases by 17.1% if the condition of maintenance is good, while the rest of the characteristics remain the same.
20. Storage = 1.032: The expected negotiation margin for a property increases by 3.2% if it is equipped with storage, while the rest of the characteristics remain the same.

https://doi.org/10.26219/heal.aueb.6011

Below we present a table for all the different prefectures of Greece and their negotiation margins according to the two different definitions we gave. The three prefectures with the biggest negotiation margins (based on the ratio) are, N. KEFALLONIAS with 51% or 58914€ next N. CHALKIDIKIS with 50% or 58847€, and lastly N. EVRYTANIAS with 39% or 30591€. The three prefectures with the most properties are N. ATHINON where the negotiation margin is 24% or 37950 €, the N. THESSALONIKIS with a negotiation margin of 26% or 26379 € and finally N. ANATOLIKIS ATTIKIS with a 23% negotiation margin or 53022€. We have to note that although a prefecture might have a higher negotiating margin than another prefecture, based on the ratio definition, it is not necessary that it also has a higher negotiation margin according to the difference definition. An example is the N. EVRYTANIAS and N. ATHINON where we have a 39% negotiation margin for the first and 24% for the second according to the ratio definition, but when we look at the negotiation margin according to the difference definition, we see that N. EVRYTANIAS has a negotiation margin of 30591€ while N. ATHINON has 37950€. Finally, there is only one prefecture that has a negative negotiation margin, this happened because the market value the AMV predicted was higher than the asking price.

https://doi.org/10.26219/heal.aueb.6011

Prefecture Name	Median Negotiation Margin (ratio)	Median Negotiation Margin (difference)
N. ACHAIAS	1.15	12590 €
N. ANATOLIKIS ATTIKIS	1.23	53022 €
N. ARGOLIDAS	1.14	19503 €
N. ARKADIAS	1.26	23174 €
N. ARTAS	1.11	7656 €
N. ATHINON	1.24	37950 €
N. CHALKIDIKIS	1.50	58847 €
N. CHANION	1.37	58675 €
N. CHIOU	1.15	17477 €
N. DODEKANISON	1.33	42475 €
N. DRAMAS	1.16	13122 €
N. DYTIKIS ATTIKIS	1.23	22353 €
N. ETOLOAKARNANIAS	1.18	14147 €
N. EVROU	1.31	27630 €
N. EVRYTANIAS	1.39	30591 €
N. EVVIAS	1.33	32332 €
N. FLORINAS	1.07	4721 €
N. FOKIDAS	1.26	30459 €
N. FTHIOTIDAS	1.11	7421 €
N. GREVENON	1.21	12655 €
N. ILIAS	1.20	15692 €
N. IMATHIAS	1.14	9071 €
N. IOANNINON	1.19	20265 €
N. IRAKLIOU	1.20	28931 €
N. KARDITSAS	1.02	1271 €
N. KASTORIAS	1.19	11977 €
N. KAVALAS	1.29	26583 €
N. KEFALLONIAS	1.51	58914 €
N. KERKYRAS	1.25	41189 €
N. KILKIS	0.91	-5829 €
N. KORINTHOU	1.33	34488 €
N. KOZANIS	1.11	7372 €
N. KYKLADON	1.33	71245 €
N. LAKONIAS	1.24	21280 €
N. LARISAS	1.17	14403 €

The table continues on the next page

https://doi.org/10.26219/heal.aueb.6011

Prefecture Name	Median Negotiation Margin (ratio)	Median Negotiation Margin (difference)
N. LASITHIOU	1.28	33083 €
N. LEFKADAS	1.38	51381 €
N. LESVOU	1.08	10170 €
N. MAGNISIAS	1.25	23598 €
N. MESSINIAS	1.36	43925 €
N. PELLAS	1.15	11139 €
N. PIERIAS	1.22	18149 €
N. PIREOS KE NISON	1.12	15139 €
N. PREVEZAS	1.33	24178 €
N. RETHYMNOU	1.16	19707 €
N. RODOPIS	1.18	14718 €
N. SAMOU	1.19	16060 €
N. SERRON	1.27	17064 €
N. THESPROTIAS	1.26	26379 €
N. THESSALONIKIS	1.33	35967 €
N. TRIKALON	1.09	6077 €
N. VIOTIAS	1.36	41571 €
N. XANTHIS	1.25	18458 €
N. ZAKYNTHOU	1.22	34416 €

Table 25: Median negotiation margin for every prefecture

https://doi.org/10.26219/heal.aueb.6011

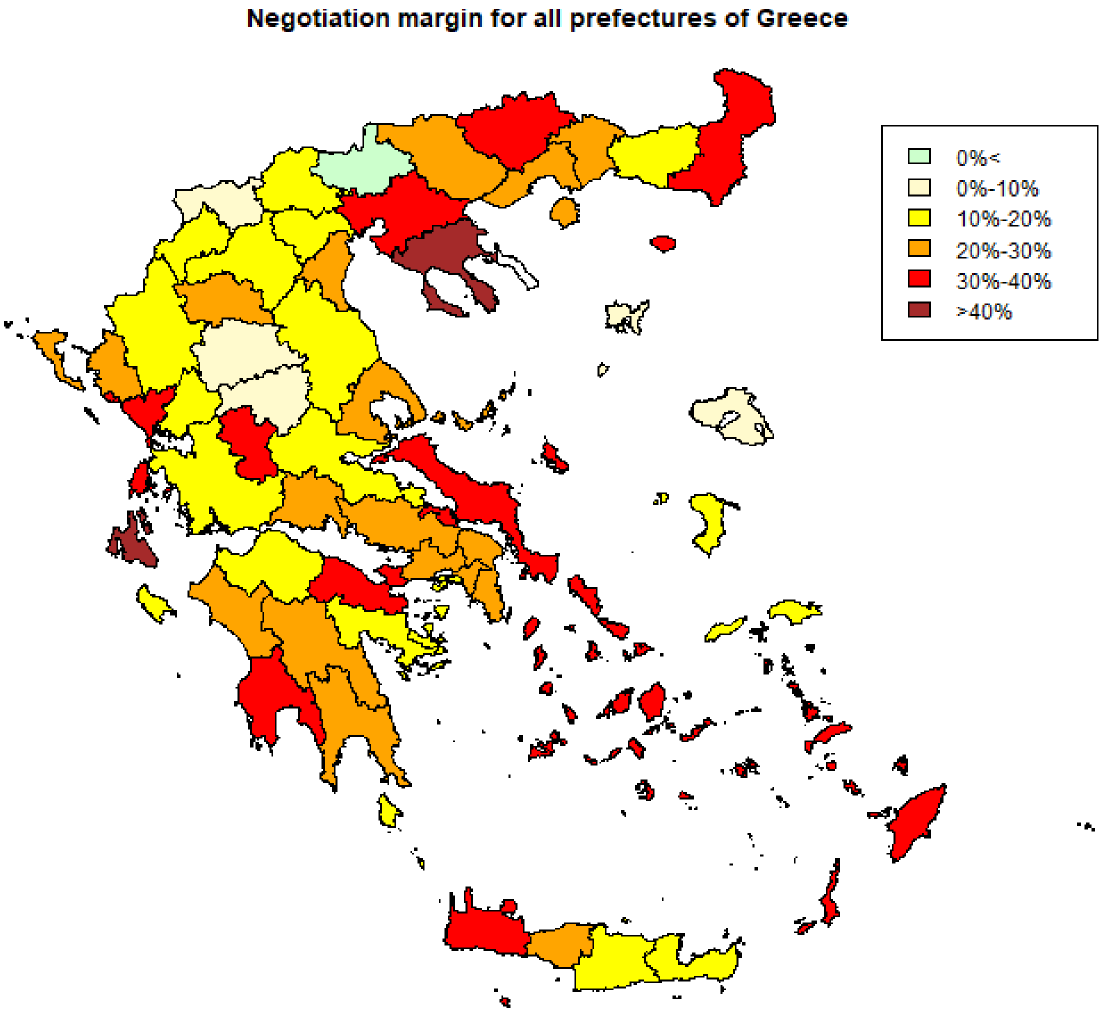

Figure 25: Heatmap for the negotiation margin for every prefecture

https://doi.org/10.26219/heal.aueb.6011

Our next step is to have a closer look at the three prefectures with the most properties.

The first prefecture we are going to analyze is N. ATHINON. N. ATHINON is the most populated prefecture of Greece and contains the city of Athens which is the capital of the country. Below we present a table with the basic measures for the two different definitions of negotiation margin and their respective histograms. For the difference, we see that the mean negotiation margin is 62266€ while the median is 37950€ this indicates right skewness, which is confirmed by the skewness index and Figure 26. Moreover, according to the kurtosis index, we conclude that the distribution of the negotiation margin is leptokurtic. Similar observations can be made for the negotiation margin when we use the ratio. The mean is 1.33 and the median is 1.24, indicating right skewness which is confirmed by the skewness index and Figure 27. Also, according to the kurtosis index the distribution can be classified as leptokurtic. Finally, for both definitions of the negotiation margin, the standard deviation appears to be very high.

	mean	median	SD	skew	kurt
Difference	62266€	37950€	98673€	1.69	5.57
Ratio	1.33	1.24	0.46	2.48	17.83

Table 26: Basic measures for Negotiation Margin for the prefecture of N. ATHINON

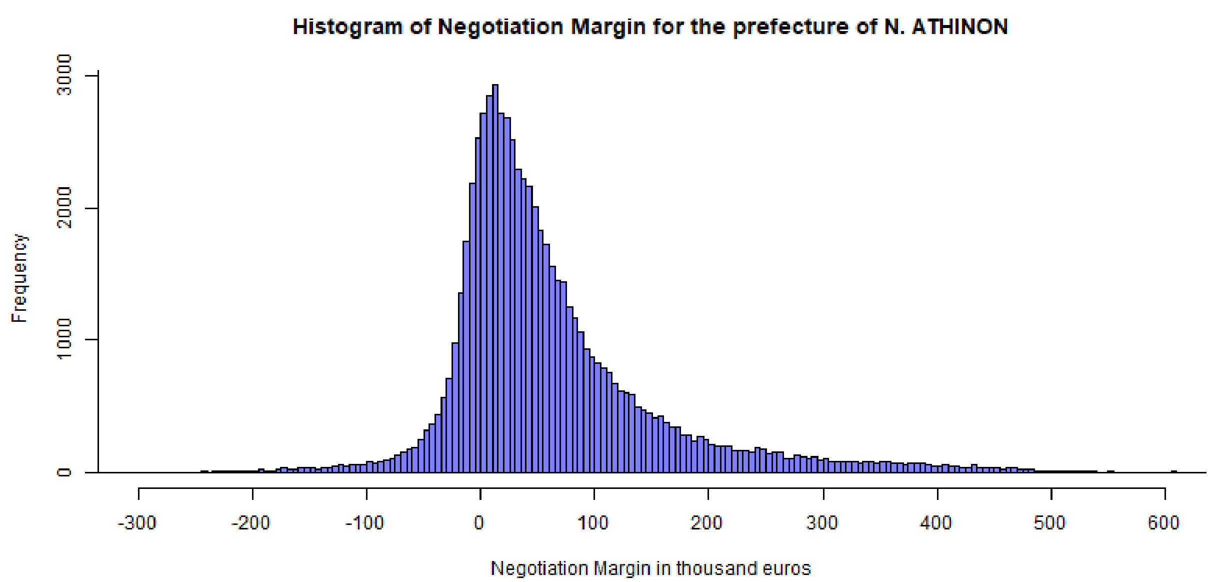

Figure 26: Histogram of Negotiation Margin (difference) for the prefecture of N. ATHINON

https://doi.org/10.26219/heal.aueb.6011

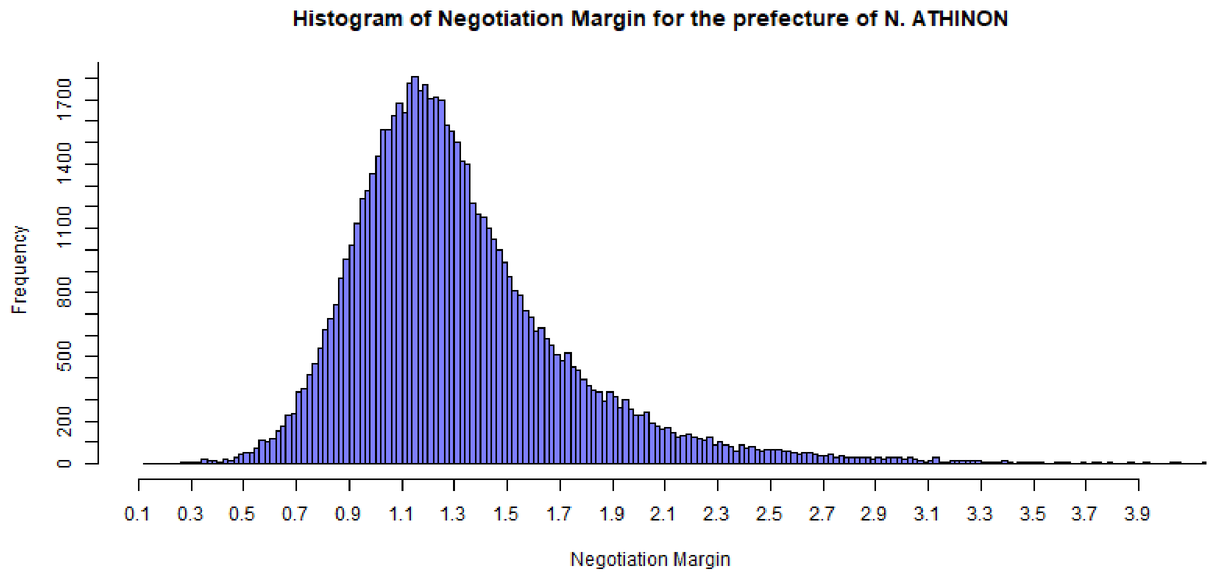

Figure 27: Histogram of Negotiation Margin (ratio) for the prefecture of N. ATHINON

We continue by studying the relationship between the negotiation margin and the type of property. In Figure 28 we present the logarithm of the negotiation margin (ratio) for each type of property. By looking at Figure 28 we cannot detect any important differences but according to Table 27 where we have calculated the medians and the 95% confidence intervals, Table 22 the medians do appear to be different. Specifically, Maisonettes appear to have the smallest negotiation margin, while Houses have the biggest. The Kruskal-Wallis test for the equality of the medians (null hypothesis), for a=5%, rejects the null hypothesis, thus there is a significant statistical difference. Finally, according to Figure 28, we see that Flats have a bigger variance compared to the rest of the types.

	Type of property			
	Duplex	Flat	House	Maisonette
Median	1.27 1	1.245	1.291	1.197
95% Lower Lim.	1.249	1.241	1.268	1.191
95% Upper Lim.	1.294	1.248	1.315	1.204

Table 27: Basic measures of Negotiation Margin for each type of property in the prefecture of N. ATHINON

https://doi.org/10.26219/heal.aueb.6011

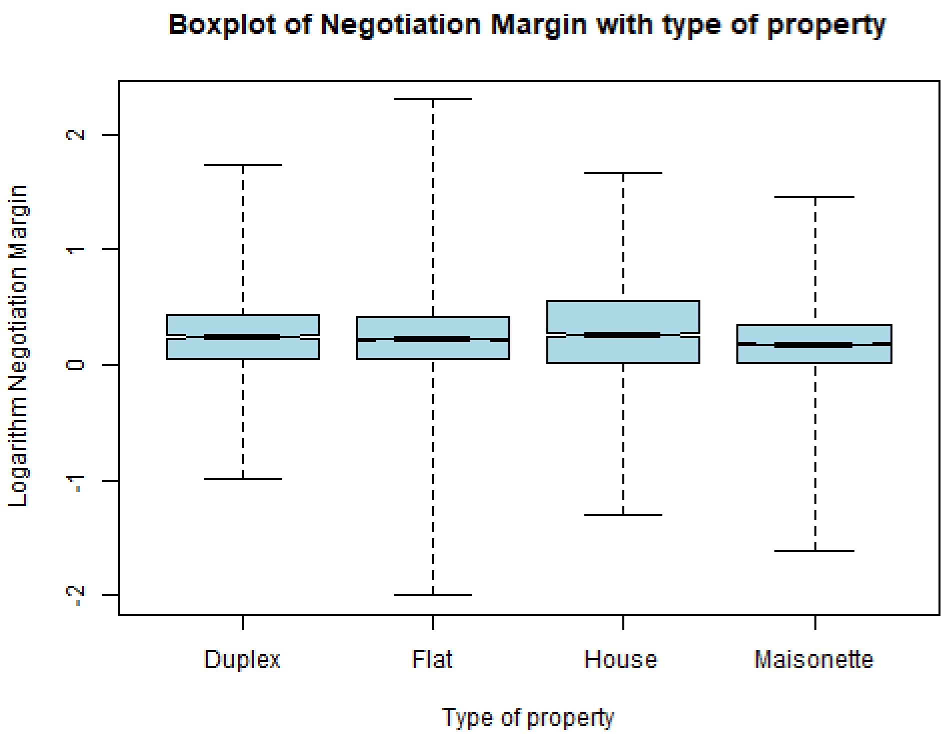

Figure 28:Boxplot for Negotiation Margin and type of property in the prefecture of N. ATHINON

Next, we will take a look at the relationship between the negotiation margin and the year of construction. According to Table 28, where we illustrate the medians and the 95% confidence intervals, we notice that the median negotiation margins become smaller as the year of construction is closer to the present. This can also be seen in Figure 29. Moreover, we see a very big drop in the negotiation margin from 1920-1940 to 1941-1960 meaning that owners of old properties tend to be more flexible in price. After applying the Kruskal-Wallis test for the equality of the medians (null hypothesis), for a=5%, we reject the null hypothesis, thus there is a significant statistical difference. Finally, properties constructed from 1961 to 1980 have the biggest variance in negotiation margin.

	Year of construction				
	1920-1940	1941-1960	1961-1980	1981-2000	2001-2023
Median	1.558	1.287	1.27	1.233	1.204
Lower Lim.	1.515	1.27	1.265	1.227	1.2
Upper Lim.	1.601	1.304	1.275	1.239	1.207

Table 28: Basic measures of Negotiation Margin based on the year of construction in the prefecture of N. ATHINON

https://doi.org/10.26219/heal.aueb.6011

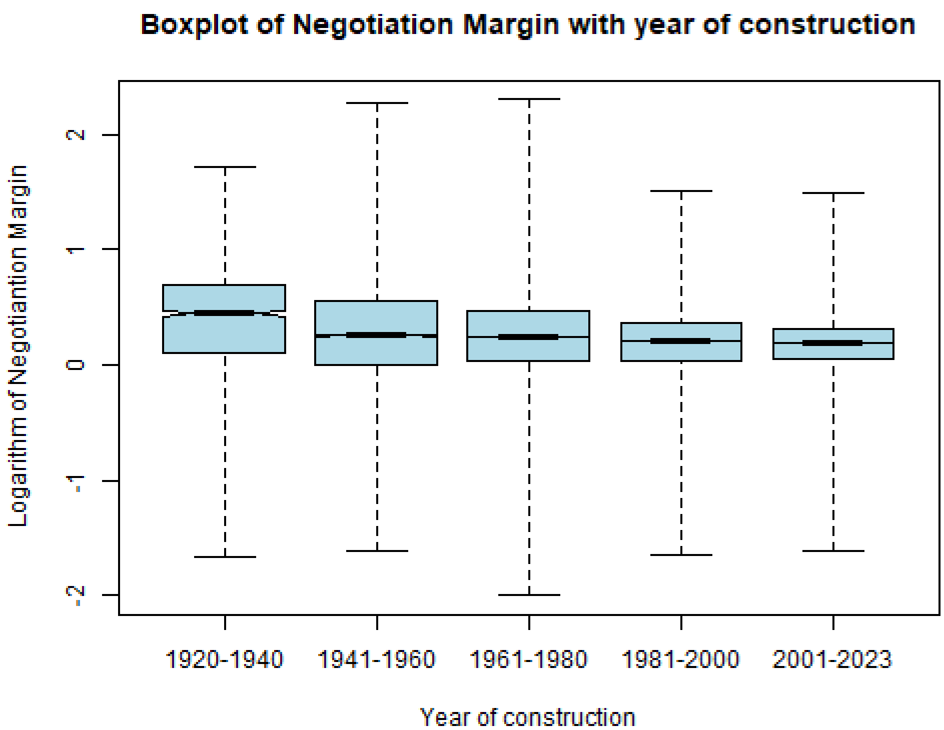

Figure 29:Boxplot for Negotiation Margin and year of construction in the prefecture of N. ATHINON

We will continue by fitting a model similar to what we did earlier but this time we will use only the properties located in the prefecture of N. ATHINON. Below we present the table with the predictions and the 95% bootstrap confidence intervals.

https://doi.org/10.26219/heal.aueb.6011

Variable	Prediction	Prediction (exponent)	Lower Limit (exponent)	Upper Limit (exponent)
Intercept	0.169	1.184	1.161	1.206
Flat	-0.135	0.873	0.862	0.885
Duplex	-0.104	0.9	0.881	0.921
Maisonette	-0.102	0.902	0.889	0.915
Logarithm of usable area	-0.155	0.856	0.848	0.863
Year of construction	-0.003	0.996	0.995	0.997
Ground floor	0.218	1.243	1.228	1.26
Other floor	0.252	1.287	1.272	1.304
Parking	0.061	1.063	1.056	1.07
2 Bedrooms	0.017	1.017	1.01	1.025
3 Bedrooms	0.032	1.033	1.023	1.042
4+ Bedrooms	-0.018	0.981	0.967	0.992
Elevator	0.043	1.044	1.039	1.049
View	0.049	1.05	1.046	1.055
2 Bathrooms	0.09	1.094	1.086	1.101
3 Bathrooms	0.106	1.111	1.098	1.126
4+ Bathrooms	0.064	1.066	1.044	1.092
Condition	0.128	1.137	1.128	1.145
Storage	0.037	1.038	1.033	1.044

Table 29: Exponents of the predicted betas and the 95% confidence intervals

Below we have the explanation for each of the predicted betas.

1. Intercept = 1.184: The expected negotiation margin for a property that is classified as House, the usable area is 70 square meters, the year of construction is 1920, it is located at -1 level, does not have parking, the quality of construction is bad, it has 1 bedroom, it has no elevator, it has no view, it contains one bathroom, the condition of maintenance is bad and has storage is 18.4% of the asking price.
2. Flat = 0.873: The expected negotiation margin for a property drops by 12.7% if it is classified as Flat instead of House, while the rest of the characteristics remain the same.
3. Duplex = 0.9: The expected negotiation margin for a property drops by 10% if it is classified as a Duplex instead of a House, while the rest of the characteristics remain the same.

https://doi.org/10.26219/heal.aueb.6011

4. Maisonette = 0. 902: The expected negotiation margin for a property drops by 9.8% if it is classified as a Duplex instead of a House, while the rest of the characteristics remain the same.
5. Logarithm of usable area = $1.01^{-0.226}$= -0.15%: The expected negotiation margin for a property drops by 0.15% for a 1% increase in the usable area, while the rest of the characteristics remain the same.
6. Year of construction = 0.996: The expected negotiation margin for a property drops by 0.4% when the year of construction is raised by one.
7. Ground floor = 1.243: The expected negotiation margin for a property increases by 24.3% if is on the Ground floor instead of -1 level, while the rest of the characteristics remain the same.
8. Other floor = 1.287: The expected negotiation margin for a property increases by 28.7% if it is on any other floor except the ground floor and -1 level, while the rest of the characteristics remain the same.
9. Parking = 1.063: The expected negotiation margin for a property increases by 6.3% if parking is included, while the rest of the characteristics remain the same.
10. 2 Bedrooms = 1.017: The expected negotiation margin for a property increases by 1.7% if it has two bedrooms instead of one, while the rest of the characteristics remain the same.
11. 3 Bedrooms = 1.033: The expected negotiation margin for a property increases by 3.3% if it has three bedrooms instead of one, while the rest of the characteristics remain the same.
12. 4+ Bedrooms = 0.981: The expected negotiation margin for a property decreases by 1.9% if it has four or more bedrooms instead of one, while the rest of the characteristics remain the same.
13. Elevator = 1.044: The expected negotiation margin for a property increases by 4.4% if it is equipped with an elevator, while the rest of the characteristics remain the same.
14. View = 1.05: The expected negotiation margin for a property increases by 5% if it has a good view, while the rest of the characteristics remain the same.
15. 2 Bathrooms = 1.094: The expected negotiation margin for a property increases by 9.4% if it has two bathrooms instead of one, while the rest of the characteristics remain the same.
16. 3 Bathrooms = 1.111: The expected negotiation margin for a property increases by 11.1% if it has three bedrooms instead of one, while the rest of the characteristics remain the same.

https://doi.org/10.26219/heal.aueb.6011

17. 4+ Bathrooms = 1.066: The expected negotiation margin for a property increases by 6.6% if it has four or more bedrooms instead of one, while the rest of the characteristics remain the same.
18. Condition = 1.137: The expected negotiation margin for a property increases by 13.7% if the condition of maintenance is good, while the rest of the characteristics remain the same.
19. Storage = 1.038: The expected negotiation margin for a property increases by 3.8% if it is equipped with storage, while the rest of the characteristics remain the same.

Below we have a heatmap depicting the negotiation margin for the prefecture of N. ATHINON. For this prefecture, we have a very big sample of properties and thus the depiction is very detailed. Also, as the negotiation margin rises the color goes closer to red, and as a result, we see which locations have great negotiation margins. According to Figure 30, the city of Athens (center) appears to have big negotiation margins. Moreover, we spot great negotiation margins along the coastline in the regions of Alimos and Glifada. Finally, there are a few locations where the negotiation margins are negative.

https://doi.org/10.26219/heal.aueb.6011

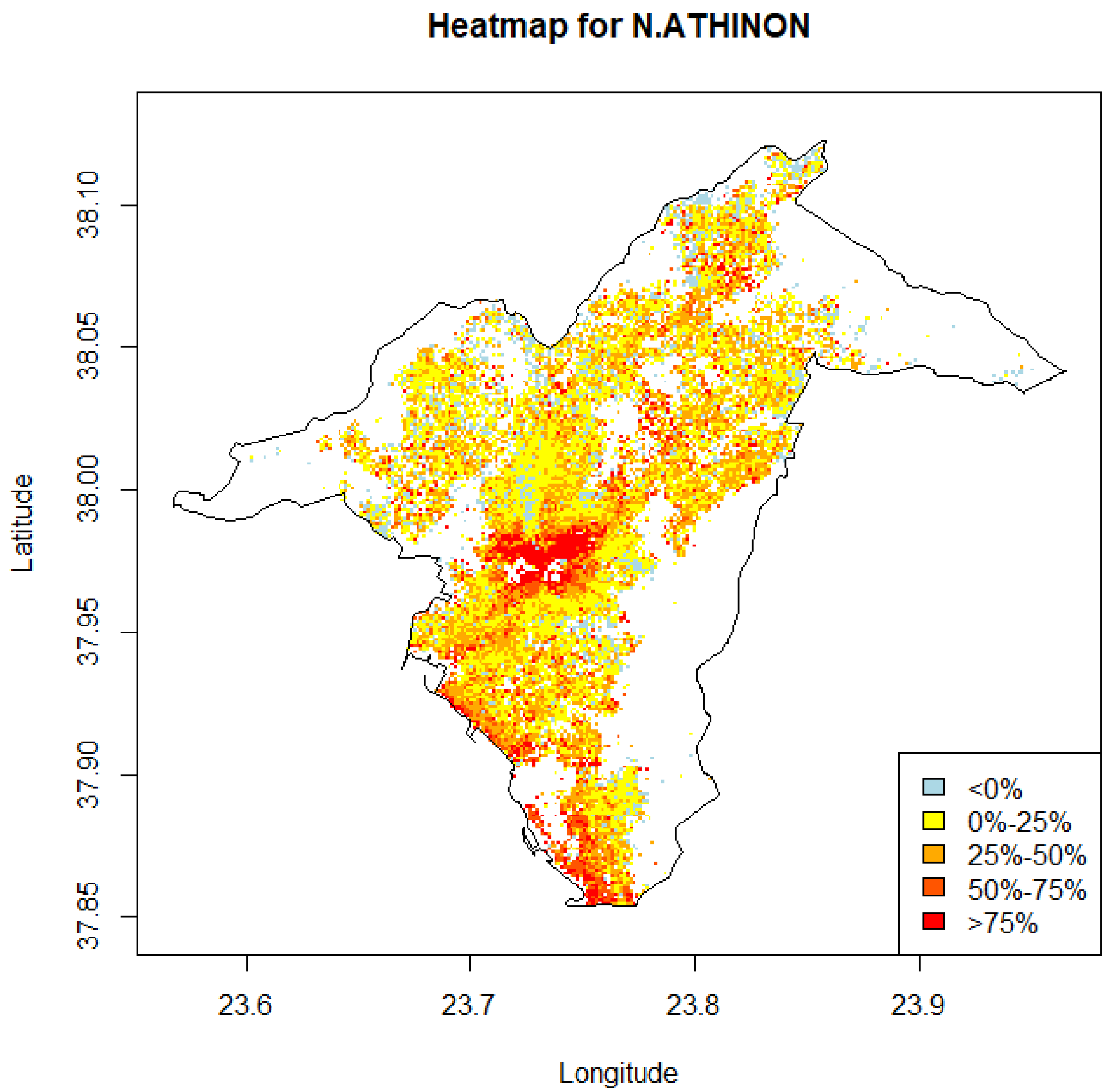

Figure 30: Heatmap for the negotiation margin for the prefecture of N. ATHINON

 https://doi.org/10.26219/heal.aueb.6011

The second prefecture we are going to analyze is N. THESSALONIKIS. N. THESSALONIKIS is the second most populated prefecture of Greece. It is located in the north part of the country and contains the city of Thessalonica which is the second largest city of the country. Here are the basic measures for two different definitions of negotiation margin, along with their respective histograms. The mean negotiation margin (for the difference) is calculated to be 52746€, with a median of 35966€, suggesting right skewness. This is supported by the skewness index and Figure 31, also according to the kurtosis index the distribution is leptokurtic. Similar observations are found for the negotiation margin using the ratio. The mean is 1.33 and the median is 1.24, indicating right skewness, which is also supported by the skewness index and Figure 32, additionally the kurtosis index suggests a leptokurtic distribution. For both definitions of the negotiation margin, we observe a high standard deviation.

	mean	median	SD	skew	kurt
Difference	52746€	35966€	76773€	2.84	14.28
Ratio	1.45	1.33	0.54	1.8	9.45

Table 30: Basic measures for Negotiation Margin for the prefecture of N. THESSALONIKIS

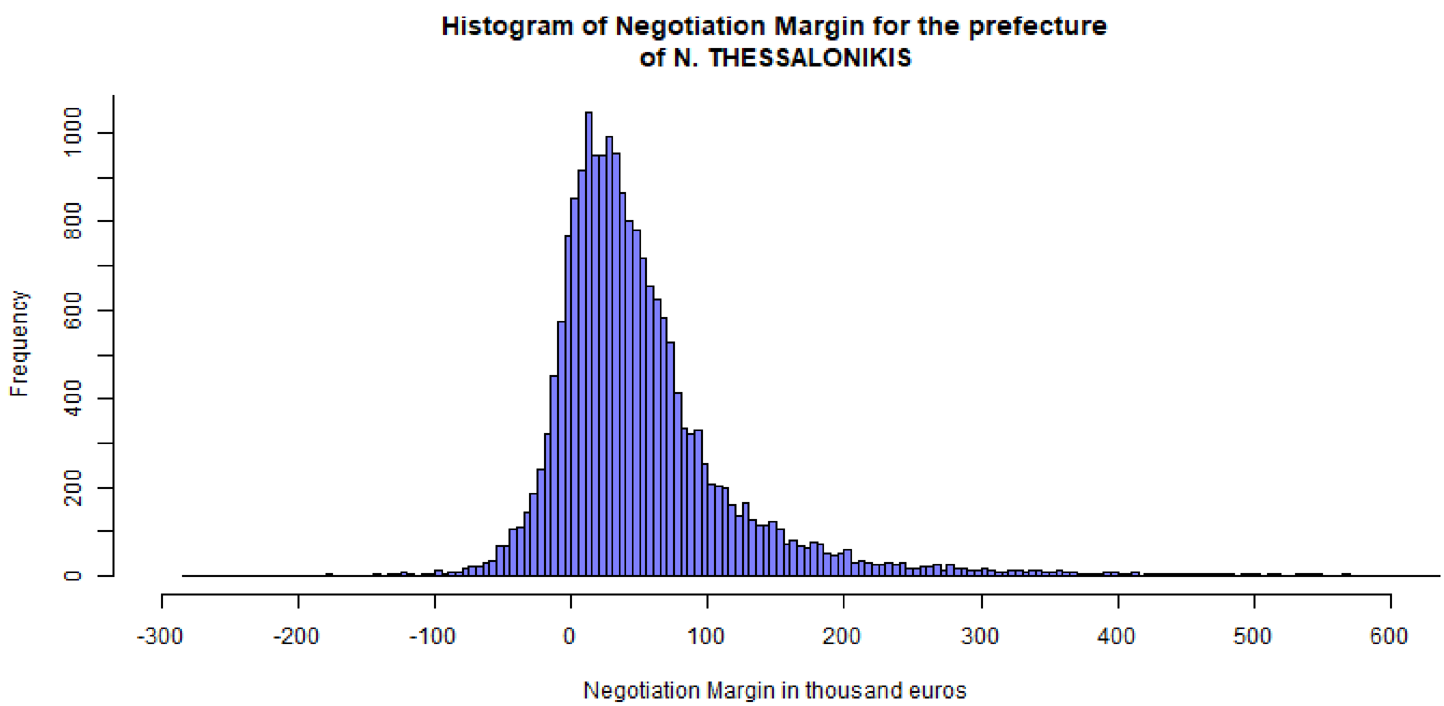

Figure 31: Histogram of Negotiation Margin (difference) for the prefecture of N. THESSALONIKIS

https://doi.org/10.26219/heal.aueb.6011

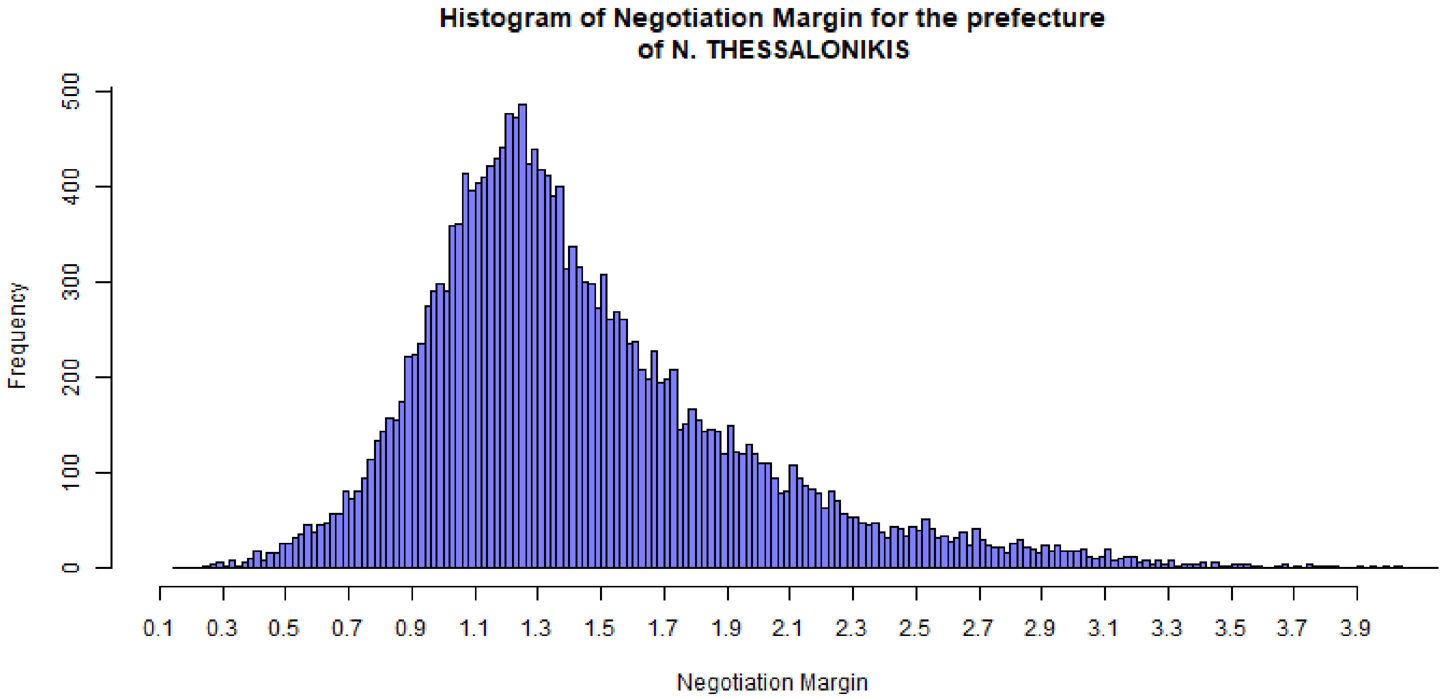

Figure 32: Histogram of Negotiation Margin (ratio) for the prefecture of N. THESSALONIKIS

Next, we will study the relationship between the negotiation margin and the type of property. In Figure 33 we present the logarithm of the negotiation margin (ratio) for each type of property. By examining Figure 33 we cannot detect any important differences although according to Table 31, we detect differences in the medians. Specifically, Duplexes appear to have the smallest negotiation margin, but also the largest confidence interval, on the other hand, Houses have the biggest median negotiation margin. The Kruskal-Wallis test for the equality of the medians (null hypothesis), for a=5%, rejects the null hypothesis, thus there is a significant statistical difference. Finally, according to Figure 33, we see that Flats have the biggest variance to the rest of the types.

	Type of property			
	Duplex	Flat	House	Maisonette
Median	1.244	1.332	1.386	1.306
95% Lower Lim.	1.156	1.324	1.358	1.294
95% Upper Lim.	1.338	1.339	1.414	1.318

Table 31: Basic measures of Negotiation Margin for each type of property in the prefecture of N. THESSALONIKIS

https://doi.org/10.26219/heal.aueb.6011

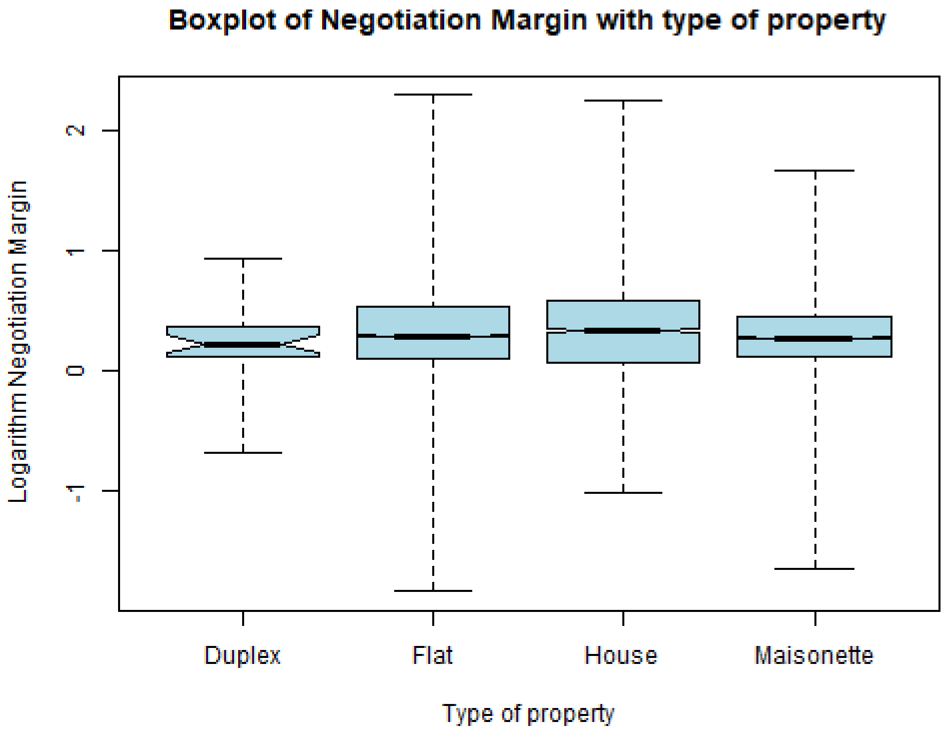

Figure 33:Boxplot for Negotiation Margin and type of property in the prefecture of N. THESSALONIKIS

One more relationship, we will examine is the one between the negotiation margin and the year of construction. According to Table 32, we notice that the median negotiation margins become smaller as the year of construction is closer to the present, this becomes more evident from Figure 34. This is not apparent for 1920-1940 and 1941-1960, where we see properties of 1920-1940 having a bigger median but the confidence interval is pretty big and contains the median of 1941-1960 which implies that the medians are the same statistically. After applying the Kruskal-Wallis test for the equality of all the medians (null hypothesis), for a=5%, we reject the null hypothesis, thus there is a significant statistical difference.

	Year of construction				
	1920-1940	1941-1960	1961-1980	1981-2000	2001-2023
Median	1.621	1.674	1.418	1.327	1.25
Lower Lim.	1.468	1.627	1.405	1.314	1.243
Upper Lim.	1.79	1.722	1.431	1.341	1.257

Table 32: Basic measures of Negotiation Margin based on the year of construction in the prefecture of N. THESSALONIKIS

https://doi.org/10.26219/heal.aueb.6011

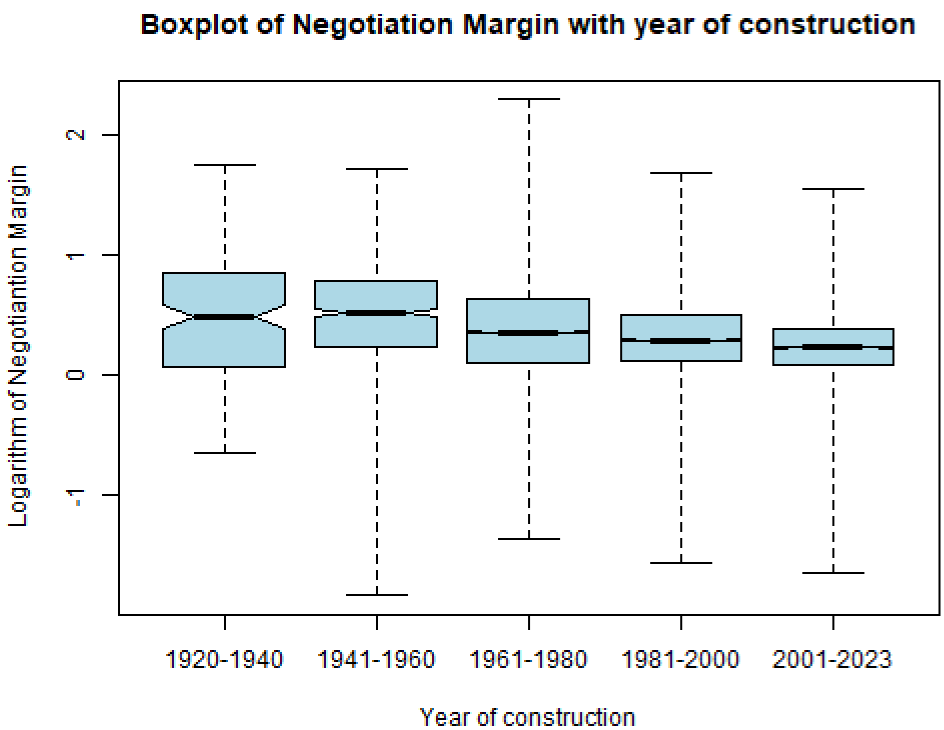

Figure 34:Boxplot for Negotiation Margin and year of construction in the prefecture of N. THESSALONIKIS

Finally, we fit a model for the prefecture of N. THESSALONIKIS and explain the betas. Below we present the table with the predictions and the 95% bootstrap confidence intervals.

 https://doi.org/10.26219/heal.aueb.6011

Variable	Prediction	Prediction (exponent)	Lower Limit (exponent)	Upper Limit (exponent)
Intercept	0.29	1.337	1.289	1.386
Flat	-0.169	0.843	0.825	0.862
Duplex	-0.198	0.82	0.739	0.919
Maisonette	-0.058	0.943	0.925	0.962
Logarithm of usable area	-0.186	0.83	0.815	0.843
Year of construction	-0.005	0.994	0.993	0.995
Ground floor	0.217	1.242	1.212	1.276
Other floor	0.25	1.284	1.253	1.318
Parking	0.051	1.053	1.035	1.071
Quality	0.022	1.022	1.01	1.034
2 Bedrooms	-0.062	0.939	0.926	0.953
3 Bedrooms	0.026	1.026	1.006	1.048
4+ Bedrooms	0.057	1.058	1.029	1.089
Elevator	0.132	1.141	1.13	1.152
View	0.068	1.07	1.06	1.081
2 Bathrooms	0.098	1.103	1.088	1.12
3 Bathrooms	0.178	1.194	1.163	1.226
4+ Bathrooms	0.201	1.223	1.16	1.288
Condition	0.205	1.227	1.21	1.245
Storage	0.017	1.018	1.007	1.029

Table 33: Exponents of the predicted betas and the 95% confidence intervals

Below we have the explanation for each of the predicted betas.

1. Intercept = 1.337: The expected negotiation margin for a property that is classified as House, the usable area is 70 square meters, the year of construction is 1920, it is located at -1 level, does not have parking, the quality of construction is bad, it has 1 bedroom, it has no elevator, it has no view, it contains one bathroom, the condition of maintenance is bad and has storage is 33.7% of the asking price.
2. Flat = 0.873: The expected negotiation margin for a property drops by 15.7% if it is classified as Flat instead of House, while the rest of the characteristics remain the same.

https://doi.org/10.26219/heal.aueb.6011

3. Duplex = 0.82: The expected negotiation margin for a property drops by 18% if it is classified as a Duplex instead of a House, while the rest of the characteristics remain the same.
4. Maisonette = 0. 943: The expected negotiation margin for a property drops by 5.7% if it is classified as a Duplex instead of a House, while the rest of the characteristics remain the same.
5. Logarithm of usable area = $1.01^{-0.226}$= -0.18%: The expected negotiation margin for a property drops by 0.18% for a 1% increase in the usable area, while the rest of the characteristics remain the same.
6. Year of construction = 0.994: The expected negotiation margin for a property drops by 0.6% when the year of construction is raised by one.
7. Ground floor = 1.242: The expected negotiation margin for a property increases by 24.2% if is on the Ground floor instead of -1 level, while the rest of the characteristics remain the same.
8. Other floor = 1.284: The expected negotiation margin for a property increases by 28.4% if it is on any other floor except the ground floor and -1 level, while the rest of the characteristics remain the same.
9. Parking = 1.053: The expected negotiation margin for a property increases by 5.3% if parking is included, while the rest of the characteristics remain the same.
10. Quality = 1.022: The expected negotiation margin for a property increases by 2.2% if the quality of construction is good, while the rest of the characteristics remain the same.
11. 2 Bedrooms = 0.939: The expected negotiation margin for a property decreases by 6.1% if it has two bedrooms instead of one, while the rest of the characteristics remain the same.
12. 3 Bedrooms = 1.026: The expected negotiation margin for a property increases by 2.6% if it has three bedrooms instead of one, while the rest of the characteristics remain the same.
13. 4+ Bedrooms = 1.058: The expected negotiation margin for a property increases by 5.8% if it has four or more bedrooms instead of one, while the rest of the characteristics remain the same.
14. Elevator = 1.141: The expected negotiation margin for a property increases by 14.1% if it is equipped with an elevator, while the rest of the characteristics remain the same.
15. View = 1.07: The expected negotiation margin for a property increases by 7% if it has a good view, while the rest of the characteristics remain the same.

https://doi.org/10.26219/heal.aueb.6011

16. 2 Bathrooms = 1.103: The expected negotiation margin for a property increases by 10.3% if it has two bathrooms instead of one, while the rest of the characteristics remain the same.
17. 3 Bathrooms = 1.194: The expected negotiation margin for a property increases by 19.4% if it has three bedrooms instead of one, while the rest of the characteristics remain the same.
18. 4+ Bathrooms = 1.223: The expected negotiation margin for a property increases by 22.3% if it has four or more bedrooms instead of one, while the rest of the characteristics remain the same.
19. Condition = 1.227: The expected negotiation margin for a property increases by 22.7% if the condition of maintenance is good, while the rest of the characteristics remain the same.
20. Storage = 1.018: The expected negotiation margin for a property increases by 1.8% if it is equipped with storage, while the rest of the characteristics remain the same.

Below we have a heatmap depicting the negotiation margin for the prefecture of N. THESALONIKIS. Although we have many properties in our dataset for this prefecture, we can see that most of the map is empty. Most of the properties are located in the center of the city of Thessalonica, and the closer to the city a property is the bigger the negotiation margin we have. Finally, Panorama and Thermi are two locations where we can spot great negotiation margins.

https://doi.org/10.26219/heal.aueb.6011

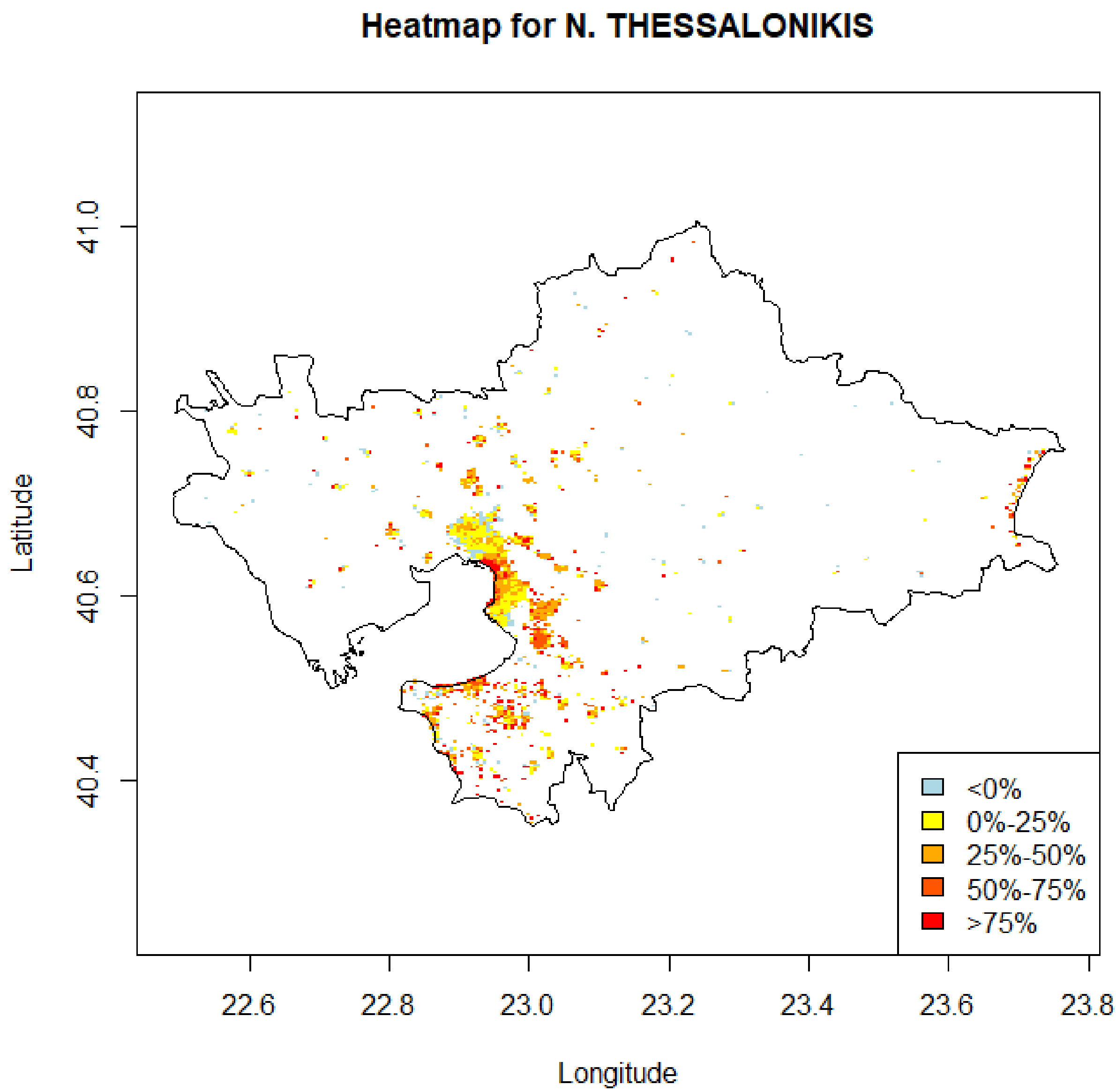

Figure 35: Heatmap for the negotiation margin for the prefecture of N. THESALONIKIS

https://doi.org/10.26219/heal.aueb.6011

The last prefecture we are going to analyze is N. ANATOLIKIS ATTIKIS. N. ANATOLIKIS ATTIKIS is a very big prefecture housing many people. It is located east of the prefecture of N. ATHINON. Here are the basic measures for two different definitions of negotiation margin, along with their respective histograms. The mean negotiation margin (for the difference) is calculated to be 75815€, with a median of 53022€, suggesting right skewness. This is supported by the skewness index in Table 34 and Figure 36, also according to the kurtosis index the distribution is slightly leptokurtic. Similar observations are found for the negotiation margin using the ratio. The mean is 1.32 and the median is 1.23, indicating right skewness, which is also supported by the skewness index and Figure 37, the kurtosis index suggests a leptokurtic distribution. For both definitions of the negotiation margin, we observe a high standard deviation.

	mean	median	SD	skew	kurt
Difference	75815€	53022€	134027€	0.76	1.75
Ratio	1.32	1.23	0.52	2.33	12.29

Table 34: Basic measures for Negotiation Margin for the prefecture of N. ANATOLIKIS ATTIKIS

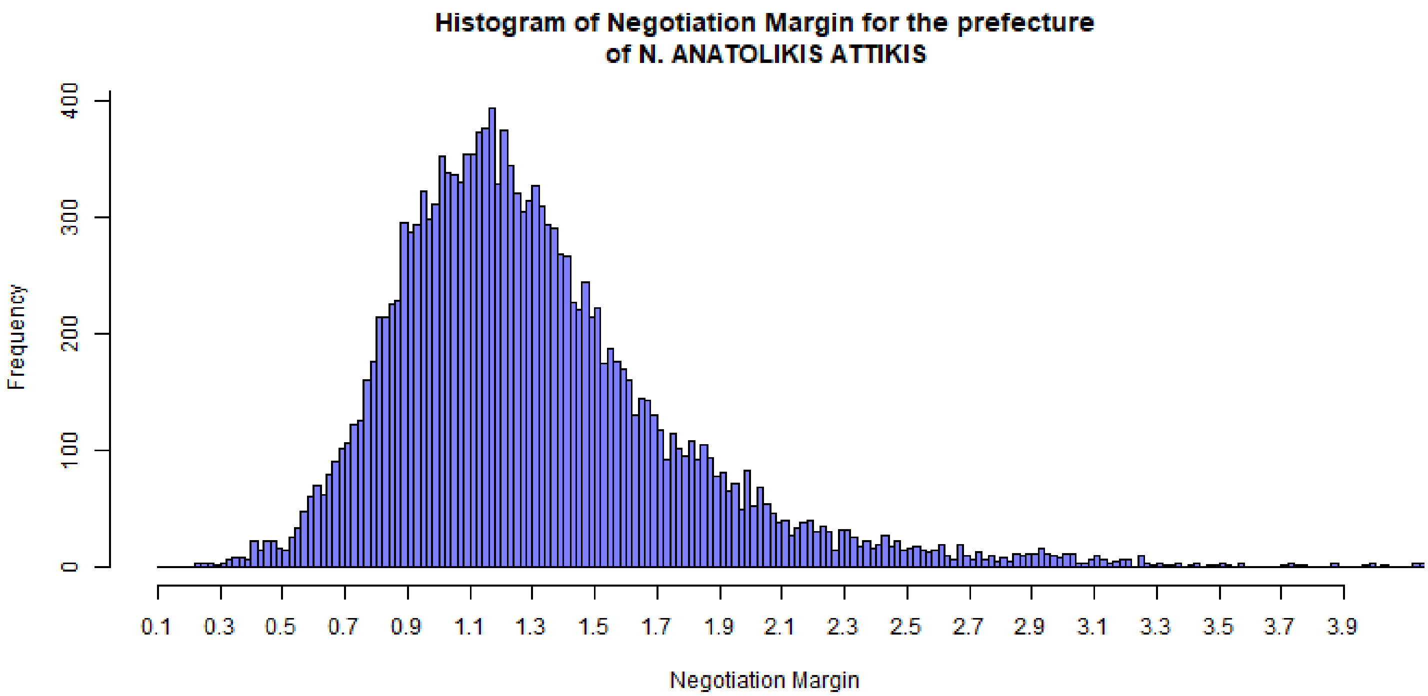

Figure 36: Histogram of Negotiation Margin (difference) for the prefecture of N. ANATOLIKIS ATTIKIS

https://doi.org/10.26219/heal.aueb.6011

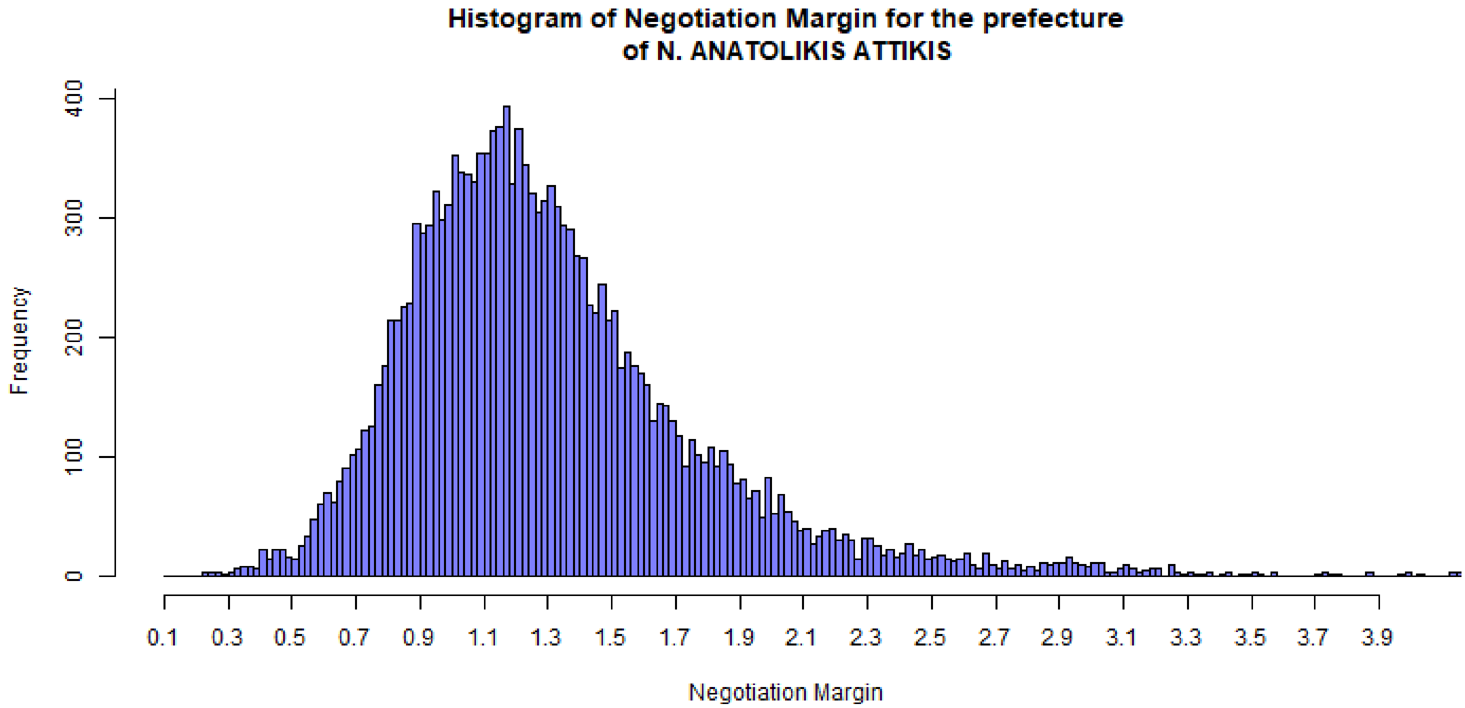

Figure 37: Histogram of Negotiation Margin (ratio) for the prefecture of N. ANATOLIKIS ATTIKIS

We continue by studying the relationship between the negotiation margin and the type of property. In Figure 38 we present the logarithm of the negotiation margin (ratio) for each type of property. By examining Figure 38 we detect that the median negotiation margin for a Maisonette is slightly lower than the other types of properties. According to Table 35, Flats have the largest median negotiation margin while Maisonettes have the smallest. The Kruskal-Wallis test for the equality of the medians (null hypothesis), for a=5%, rejects the null hypothesis, thus there is a significant statistical difference. Finally, according to Figure 38, we see that Flats have the biggest variance than the rest of the types.

	Type of property			
	Duplex	Flat	House	Maisonette
Median	1.255	1.328	1.237	1.137
95% Lower Lim.	1.177	1.317	1.222	1.128
95% Upper Lim.	1.338	1.339	1.252	1.146

Table 35: Basic measures of Negotiation Margin for each type of property in the prefecture of N. ANATOLIKIS ATTIKIS

https://doi.org/10.26219/heal.aueb.6011

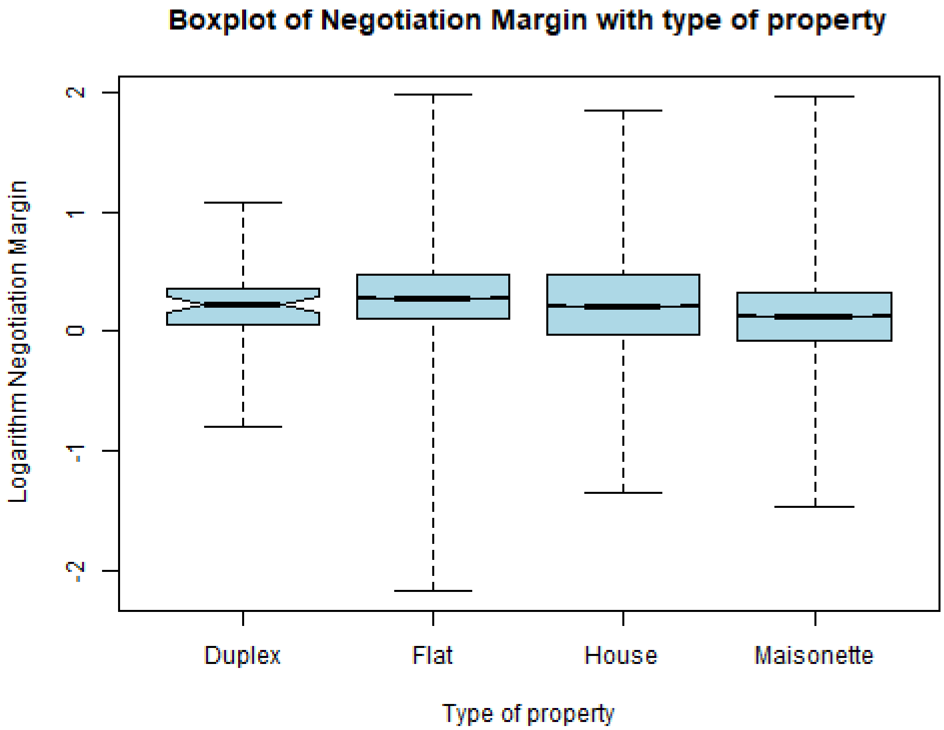

Figure 38:Boxplot for Negotiation Margin and type of property in the prefecture of N. ANATOLIKIS ATTIKIS

The last relationship we will examine is the one between the negotiation margin and the year of construction. According to Table 36, we notice that the median negotiation margins rise until 1961-1980, then for 1981-2000 drops, and for 2001-2023 drops even further. This behavior is different from what we saw in the other two prefectures. Finally, after applying the Kruskal-Wallis test for the equality of all the medians (null hypothesis), for a=5%, we reject the null hypothesis, thus there is a significant statistical difference.

	Year of construction				
	1920-1940	1941-1960	1961-1980	1981-2000	2001-2023
Median	1.458	1.496	1.533	1.237	1.18
Lower Lim.	1.347	1.35	1.505	1.223	1.173
Upper Lim.	1.578	1.658	1.562	1.251	1.187

Table 36: Basic measures of Negotiation Margin based on the year of construction in the prefecture of N. ANATOLIKIS ATTIKIS

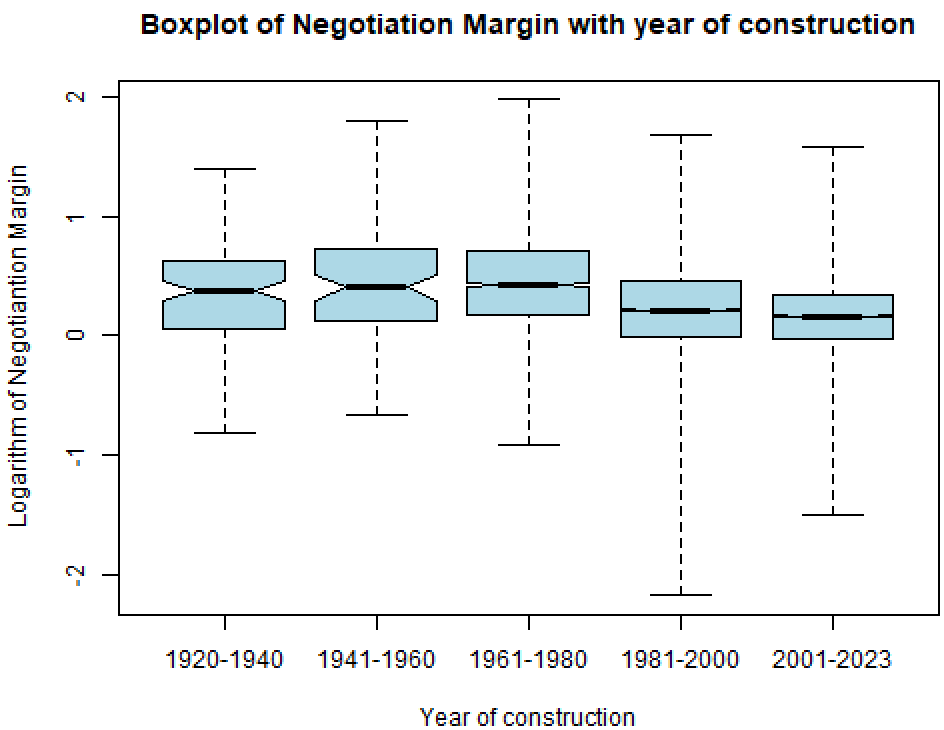

Figure 39:Boxplot for Negotiation Margin and year of construction in the prefecture of N. ANATOLIKIS ATTIKIS

https://doi.org/10.26219/heal.aueb.6011

Variable	Prediction	Prediction (exponent)	Lower Limit (exponent)	Upper Limit (exponent)
Intercept	0.497	1.645	1.593	1.695
Flat	-0.084	0.918	0.904	0.933
Duplex	-0.118	0.888	0.815	0.961
Maisonette	-0.061	0.94	0.928	0.952
Logarithm of usable area	-0.3	0.74	0.727	0.753
Year of construction	-0.003	0.996	0.995	0.996
Parking	0.026	1.027	1.015	1.037
2 Bedrooms	0.085	1.089	1.065	1.113
3 Bedrooms	0.104	1.11	1.083	1.139
4+ Bedrooms	0.089	1.093	1.061	1.128
Elevator	0.069	1.072	1.059	1.085
View	0.061	1.063	1.052	1.074
2 Bathrooms	0.054	1.056	1.041	1.07
3 Bathrooms	0.068	1.071	1.051	1.09
4+ Bathrooms	0.094	1.099	1.073	1.126
Condition	0.084	1.087	1.067	1.109
Storage	0.054	1.055	1.043	1.068

Table 37: Exponents of the predicted betas and the 95% confidence intervals

Below we have the explanation for each of the predicted betas.

1. Intercept = 1.645: The expected negotiation margin for a property that is classified as House, the usable area is 70 square meters, the year of construction is 1920, it is located at -1 level, does not have parking, the quality of construction is bad, it has 1 bedroom, it has no elevator, it has no view, it contains one bathroom, the condition of maintenance is bad and has storage is 64.5% of the asking price.
2. Flat = 0.918: The expected negotiation margin for a property drops by 8.2% if it is classified as Flat instead of House, while the rest of the characteristics remain the same.
3. Duplex = 0.888: The expected negotiation margin for a property drops by 11.2% if it is classified as a Duplex instead of a House, while the rest of the characteristics remain the same.

https://doi.org/10.26219/heal.aueb.6011

4. Maisonette = 0. 94: The expected negotiation margin for a property drops by 6% if it is classified as a Duplex instead of a House, while the rest of the characteristics remain the same.
5. Logarithm of usable area = $1.01^{-0.3}$= -0.29%: The expected negotiation margin for a property drops by 0.18% for a 1% increase in the usable area, while the rest of the characteristics remain the same.
6. Year of construction = 0.996: The expected negotiation margin for a property drops by 0.4% when the year of construction is raised by one.
7. Parking = 1.027: The expected negotiation margin for a property increases by 2.7% if parking is included, while the rest of the characteristics remain the same.
8. 2 Bedrooms = 1.089: The expected negotiation margin for a property increases by 8.9% if it has two bedrooms instead of one, while the rest of the characteristics remain the same.
9. 3 Bedrooms = 1.11: The expected negotiation margin for a property increases by 11% if it has three bedrooms instead of one, while the rest of the characteristics remain the same.
10. 4+ Bedrooms = 1.093: The expected negotiation margin for a property increases by 9.3% if it has four or more bedrooms instead of one, while the rest of the characteristics remain the same.
11. Elevator = 1.072: The expected negotiation margin for a property increases by 7.2% if it is equipped with an elevator, while the rest of the characteristics remain the same.
12. View = 1.063: The expected negotiation margin for a property increases by 6.3% if it has a good view, while the rest of the characteristics remain the same.
13. 2 Bathrooms = 1.056: The expected negotiation margin for a property increases by 5.6% if it has two bathrooms instead of one, while the rest of the characteristics remain the same.
14. 3 Bathrooms = 1.071: The expected negotiation margin for a property increases by 7.1% if it has three bedrooms instead of one, while the rest of the characteristics remain the same.
15. 4+ Bathrooms = 1.099: The expected negotiation margin for a property increases by 9.9% if it has four or more bedrooms instead of one, while the rest of the characteristics remain the same.

https://doi.org/10.26219/heal.aueb.6011

16. Condition = 1.087: The expected negotiation margin for a property increases by 8.7% if the condition of maintenance is good, while the rest of the characteristics remain the same.
17. Storage = 1.055: The expected negotiation margin for a property increases by 5.5% if it is equipped with storage, while the rest of the characteristics remain the same.

Finally, below we have a heatmap depicting the negotiation margin for the prefecture of N. ANATOLIKIS ATTIKIS. According to the heatmap, the locations where we can find the largest negotiation margins are around Voula, Vouliagmeni, and, Varkiza which are areas near popular beaches during summer.

https://doi.org/10.26219/heal.aueb.6011

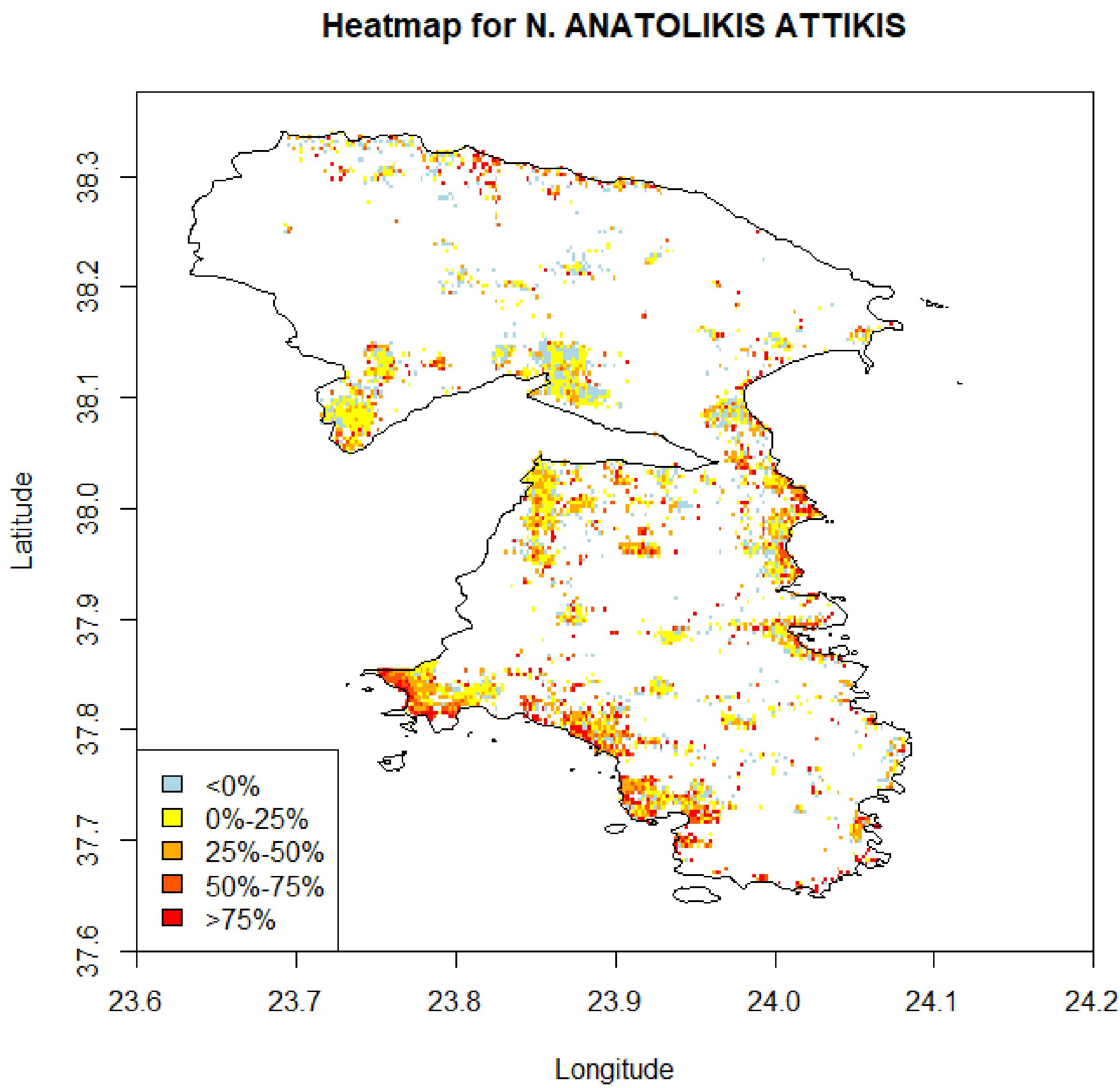

Figure 40: Heatmap for the negotiation margin for the prefecture of N. ANATOLIKIS ATTIKIS

https://doi.org/10.26219/heal.aueb.6011

https://doi.org/10.26219/heal.aueb.6011

7) CONCLUSION

In this project, we started by introducing the basic concepts of the real estate market and we presented related work that the reader can use to get more accustomed to the concepts and models used in the real estate market. Then we proceeded to clean the data by removing possible duplicate rows, this was achieved by using a combination of the GOWER distance and graph theory. Our next action was to do a thorough explanatory analysis of the variables contained in our dataset. We examined bizarre behaviors, like spikes in the distribution for the price, year of construction, and area of residence. Strange behaviors were also spotted for the categorical variables of quality of construction and condition of maintenance, where a whole level was not represented in our dataset. Those strange behaviors lead us to second guess the quality of data provided to us by the owners of the properties. We continued by doing a bivariate explanatory analysis for the variable of price with the rest of the variables, in order to better understand the relationship between those variables. We continued by creating predictive models for the asking price of a property, and we evaluated the performance of the models based on two indexes: 1) mean absolute percentage difference and 2) RMSE. According to the first index, the best predictive model was the LAD regression where we used the 1000 closest properties to our test observation to calculate the betas. On the other hand, according to the RMSE, the best model was the classical LM where again we used the 1000 closest properties to our test observation to calculate the betas, although the LAD regression scored very close to the LM. Very important is to note that Random Forest models and the models where we used the 1000 closest properties to our test observation were very time-consuming. In the last section, we gave two different definitions for the negotiation margin. The first one is the difference between the asking price and the predicted market value (according to the bank AVM) and the second one is the ratio between the asking price and the predicted market value. Next, we proceeded to the analysis of the negotiation margin for the whole country at first, and then we focused on the prefectures with the most properties. We found that the prefectures with the biggest negotiation margin according to the second definition were N. KEFALLONIAS, N. CHALKIDIKIS, and N. EVRYTANIAS, with 51%, 50%, and 39% negotiation margins. We also made an explanatory linear model to see how the different characteristics of the properties influence the negotiation margin. Finally, we presented detailed heatmaps for the prefectures with the most properties, where we could spot locations within the prefectures where the negotiation margin is higher.

https://doi.org/10.26219/heal.aueb.6011

In the future, we can continue this project by applying more predictive models for the asking price and experiment more with the location by dividing the prefectures into even smaller locations that are of interest.

https://doi.org/10.26219/heal.aueb.6011

8) REFERENCES

Akin, O., Montalvo, J. G., Villar, J. G., Peydró, J. L., & Raya, J. M. (2014). The real state and credit bubble: evidence from Spain. SERIEs, 5(2-3), 223-243.

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of random forest for valuation and a cart-based approach for model diagnostics. Expert Systems with Applications, 39(2), 1772–1778.

Appraisal Institute. (2011). The Appraisal of Real Estate (13th ed.).

Bitner, R. (2008). Confessions of a subprime lenders. An insider's tale of greed, fraud and ignorance. London: Wiley.

Brue, S. L., & Grant, R. R. (2007). The history of economic thought. Mason, OH: Thomsun South-Western.

Chatfield, C., & Collins, A.J. (1980). Preliminary data analysis. In: Introduction to Multivariate Analysis. Springer.

Coleman, J. W., & Larsen, J. E. (1991). Alternative estimation techniques for linear appraisal models. The Appraisal Journal, 59(4), 526–32.

D'Amato, M. (2017). A Brief Outline of AVM Models and Standards Evolutions. In: Advances in Automated Valuation Modeling. Springer.

D'Amato, M., & Kauko, T. (2017). Automated Valuation Models and Economic Theory. In: Advances in Automated Valuation Modeling. Springer.

https://doi.org/10.26219/heal.aueb.6011

Doumpos, M., Papastamos, D., Andritsos, D., & Zopounidis, C. (2021). Developing automated valuation models for estimating property values: a comparison of global and locally weighted approaches. Annals of Operations Research, 306, 415-433.

Dyke, C. E. (1981). Philosophy of economics. Englewood Cliffs, NJ: Prentice Hall.

Eilers, F., & Kunert, A. (2017).Automated Valuation Models for the Granting of Mortgage Loans in Germany. In: Advances in Automated Valuation Modeling. Springer.

Galesi, A., Mata, N., Rey, D., Schmitz, S., & Schuffels, J. (2020). Regional housing market conditions in Spain. Maastricht University Graduate School of Business and Economics Research Memoranda, No 29.

Sopranzetti, B.J. (2015). Hedonic Regression Models. In: Lee, CF., Lee, J. (eds) Handbook of Financial Econometrics and Statistics. Springer.

Garton, G. (2009). The subprime panic. European Financial Management, 15(1), 10–46.

IVSC. (2003). International valuation standards. London: IVSC.

IVSC. (2011). International Valuation Standards. International Valuation Standard Committee.

Kauko, T., Hooimeijer, P., & Hakfoort, J. (2017). Capturing housing market segmentation: An alternative approach based on neural network modelling. Housing Studies, 17(6), 875–894.

https://doi.org/10.26219/heal.aueb.6011

Kušar, S. (2012). Selected spatial effects of the global financial and economic crisis in Ljubljana, Slovenia. Urbani izziv, 23(2), 112–120.

Lawson, T. (2009). The current economic crisis: Its nature and the course of academic economics. Cambridge Journal of Economics, 33, 759–777.

Martin, R. (2011). The local geographies of the financial crisis: From the housing bubble to economic recession and beyond. Journal of Economic Geography, 11, 587–618.

Mooya, M. M. (2017). Automated Valuation Models and Economic Theory. In: Advances in Automated Valuation Modeling. Springer.

RICS. (2013). Automated valuation models (1st Ed.), RICS Information Paper.

Narula, S. C., Wellington, J. F., & Lewis, S. A. (2012). Valuating residential real estate using parametric programming. European journal of operational research, 217(1), 120–128.