Marta Pleń

Analysis of Factors Determining Housing Prices in the Secondary Market in Poland.

Warsaw, September 2024

LIST OF CONTENTS

1. Introduction

Housing prices are a multi-faceted topic that can be studied from both macro and microeconomic perspectives. While macroeconomic analyses often emphasize national trends and policy impacts, a microeconomic perspective focuses on individual determinants shaping property values. In this report, I will narrow my attention to the microeconomic level, aiming to explore the complex relationships between various factors and housing prices in the secondary market.

In the context of the growing housing supply gap in Poland in recent years, driven by government programs and external factors like the COVID-19 pandemic and the war in Ukraine, the dynamics of the real estate market have taken a significant place in the national discourse (PwC, 2021). The sharp increase in demand has resulted in a notable upward trend in housing prices, forcing consumers to adjust their housing preferences to economic realities. As the real estate market and economic conditions change in Poland, understanding the factors influencing housing prices becomes increasingly necessary for homeowners, investors, and researchers.

Many aspiring property owners find themselves in a situation where the dream of owning an ideal apartment often conflicts with the constraints of affordability. In response to rising prices, consumers are forced to recalibrate their housing preferences, opting for more pragmatic choices that align with their budget limitations.

The goal of this report is to unravel the intricate web of factors affecting housing prices in the secondary market in Poland's largest cities. The study aims to explain how property characteristics, amenities, and locational aspects contribute to price differences. Using a classic linear regression model, I will attempt to identify specific factors that may significantly impact housing prices.

2. Literature Review

The dynamics of residential real estate markets have often been a subject of analysis. The literature on housing prices is vast and diverse, reflecting the complexity of this phenomenon. A common method used in such research is the hedonic pricing model, which is based on Lancaster's consumer behavior theory (1966). He argued that it is not the good itself that creates utility, but the individual characteristics that define it. Moreover, housing is a multi-dimensional and heterogeneous good (Bourne, 1981) and its analysis can be divided into several parts: the characteristics of the dwelling itself and the characteristics of the location, including the neighborhood and accessibility. Due to the immobility of housing, the latter issue is particularly important (Wittowsky et al., 2020).

In his review of hedonic models of the housing market, Anthony Owusu-Ansah (2011) discusses the application of parametric models, including OLS and WLS, where the hedonic regression curve shows the relationship between the dependent variable (in this case, the price of a house) and the independent variable or explanatory factor (e.g., the number of bedrooms). He also analyzes other non-parametric and semi-parametric models. The logarithmic form of the dependent variable is widely used in hedonic pricing analysis, as it facilitates the interpretation of regression coefficients and helps minimize the problem of heteroskedasticity of residuals.

Dirk Wittowsky, Josje Hoekveld, Janina Welsch, and Michael Steier (2020) conducted an extensive analysis of housing prices in Dortmund, taking into account their characteristics, access to various public amenities, and neighborhood effects. They used OLS and spatial lag models, which proved to be more effective as they considered the influence of neighboring housing prices, preventing the overestimation of the impact of other variables.

Regarding housing characteristics, the condition of the property showed relatively high coefficients compared to other independent variables. The number of rooms was also significant, though the relationship was negative: more rooms were associated with lower prices, which may be a surprising result. The size of the living area was positively correlated with housing prices but was only significant for owner-occupied apartments.

Studies also show that services available in close proximity to the property are important. However, only two amenities were positively correlated with housing prices:

restaurants and parks. As predicted by the authors, the most important factors are directly related to the condition of the apartments, regardless of their type (Wittowsky et al., 2020).

The analysis of the housing market in Turin, presented by Luca D'Acci (2018), shows how the value of a home rises or falls depending on the area of the city, even among locations relatively close to each other. The results indicate that the value of a property decreases by 0.23% for every 1% increase in distance from the city center. The monetary costs of purchasing a home, time and transportation, and the quality benefits associated with the characteristics of a given location play a key role in household decision-making processes when choosing housing.

Similar conclusions are drawn by Emilia Tomczyk and Marta Widłak (2010) in their study based on data from transactions on the Warsaw secondary market, where they focus on constructing various hedonic price index models. The model estimated using the imputation method shows that the location of an apartment in a good district of Warsaw increases the price per square meter by about 29% compared to an apartment in an average location.

The authors also conclude that very large apartments are more expensive than small and medium-sized ones, indicating that size significantly impacts property prices. Furthermore, apartments with high-quality furnishings are, on average, 9% more expensive than those with standard furnishings, while a low standard lowers the total price by about 6% compared to the average.

Alastair Adair, Stanley McGreal, Austin Smyth, James Cooper, and Tim Ryley (2000) suggest that access to "central points" such as the Central Business District (CBD) has little significance in explaining house price variations across the entire city. Only at the submarket level, particularly in lower-income areas, does accessibility have a significant impact. Their article focuses on the factors influencing the structure of residential property prices in the urban area of Belfast, examining property characteristics, socio-economic factors, and accessibility. The analysis emphasizes the importance of research at the submarket level and draws conclusions about the complexity of relationships within the urban area.

A study by Katarzyna Małecka (2012) on the secondary housing market in Łódź confirmed the correlation between housing prices and their location and size. One district where apartments were priced above the average in Łódź was the Śródmieście district. Additionally,

the author notes that the more developed a real estate market is, the more attributes influence price formation.

As previous literature shows, access to public institutions such as schools or kindergartens significantly impacts property prices. Haizhen Wen, Yue Xiao, and Ling Zhang (2017) used housing data and educational institutions in Hangzhou, China, to develop hedonic and spatial price models to quantify the impact of educational institutions on housing prices.

Educational facilities have a positive impact on housing prices, with the results showing that property prices increase by an average of 2.737% or 0.904% when a home is located less than 1 km from a secondary school or university. However, different types of educational institutions have varying levels of impact on housing prices, and residents are willing to pay more for access to high-quality educational facilities. The proximity of nearby kindergartens, high schools, and universities increases property prices. However, the authors note that traditional hedonic models overestimate the positive effect of educational institutions, and using spatial econometric models helps eliminate this problem.

In an article by Joseph T.L. Ooi, Thao T.T. Le, and Nai-Jia Lee (2014), the authors examined the impact of construction quality on the sale price and value growth of new apartments. The Construction Quality Assessment System (CONQUAS) was used to measure construction quality in Singapore. Significant evidence was found that both sales prices and growth rates are strongly linked to the quality of new home construction. Interestingly, the "quality" premium exists in both the primary and secondary markets, and the impact of quality on resale markets is almost twice as large as in pre-sales markets. The authors state that buyers who pay high prices for high-quality homes can at least recover the premium for construction quality when reselling homes in the secondary market.

3. Hypotheses

To conduct an analysis of the factors that may influence housing prices in Polish cities, the following hypotheses were formulated. These are based on existing literature and my own assumptions.

H1: Apartment size is significant and positively correlated with price – larger apartments are generally more expensive than smaller ones.

H2: The number of rooms is significant and negatively correlated with price – apartments with more rooms are generally cheaper than those with fewer rooms.

H3: The condition of the apartment has a significant and positive impact on its price.

H4: Housing prices differ depending on access to key points of interest for residents, particularly educational institutions and restaurants.

H5: The ownership type of the apartment has a significant impact on its price.

H6: Apartments built with higher quality materials (e.g., brick) are generally more expensive than those built with other materials.

H7: Apartments on the lowest floor are the most expensive, and apartments on the highest floor are the cheapest.

H8: The presence of amenities such as parking spaces, elevators, balconies, security, or storage rooms has a significant positive effect on price.

H9: The age of the apartment is negatively correlated with price – the older the apartment, the cheaper it is on average.

H10: Apartments in the capital city are more expensive than those in other cities.

4. Data Description

The dataset used in this study was sourced from Kaggle.com and created by Krzysztof Jamroz, based on online apartment sale listings from 15 of the largest cities in Poland (Warsaw, Łódź, Kraków, Wrocław, Poznań, Gdańsk, Szczecin, Bydgoszcz, Lublin, Katowice, Białystok, Częstochowa). It also includes data from Open Street Map, detailing the distances of properties from points of interest. The collected observations cover the period from August to December 2023.

The dataset was cleaned of missing or unnecessary values, and some variables were transformed for the econometric model. The final sample used for model estimation included 2094 observations, with both continuous and binary variables.

Table 1. Variable Descriptions:

VARIABLE	DESCRIPTION	TYPE
id	Identification number for each apartment.	-
price	Price in PLN (Polish Zloty).	Continuous
squareMeters	Area in square meters.	Continuous
rooms: rooms1, rooms2, rooms3	Number of rooms (divided into three levels: 1-2 rooms, 3-4 rooms, 5-6 rooms).	Binary (dummy variables)
centreDistance	Distance to the city center in meters.	Continuous
schoolDistance	Distance to the nearest school in meters.	Continuous
clinicDistance	Distance to the nearest clinic in meters.	Continuous
kindergartenDistance	Distance to the nearest kindergarten in meters.	Continuous
restaurantDistance	Distance to the nearest restaurant in meters.	Continuous
collegeDistance	Distance to the nearest university in meters.	Continuous
pharmacyDistance	Distance to the nearest pharmacy in meters.	Continuous

VARIABLE	DESCRIPTION	TYPE
ownership	Apartment ownership: 1 = owned, 0 = cooperative	Binary
condition	Apartment condition: 1 = good, 0 = poor	Binary
buildingMaterial	Building material: 1 = brick, 0 = other	Binary
hasParkingSpace	Presence of a parking space: 1 = yes, 0 = no	Binary
hasBalcony	Presence of a balcony: 1 = yes, 0 = no	Binary
hasElevator	Presence of an elevator: 1 = yes, 0 = no	Binary
hasSecurity	Presence of security services: 1 = yes, 0 = no	Binary
hasStorageRoom	Presence of a storage room: 1 = yes, 0 = no	Binary
first_floor	Apartment on the first floor: 1 = yes, 0 = no	Binary
top_floor	Apartment on the top floor: 1 = yes, 0 = no	Binary
age	Age of the apartment in years.	Continuous
Building type: block	Block of flats: 1 – yes, 0 – no	Binary
Building type: apartmentB	Apartment building: 1 – yes, 0 – no	Binary
Building type: tenement	Tenement: 1 – yes, 0 – no	Binary
stolica	Flat in a capital city (Warsaw): 1 – yes, 0 – no	Binary

Table 1. List of variables in the database.

5. Statistical Analysis of Variables

5.1 Dependent Variable

The dependent variable in this study is *price*, which represents the housing price in PLN (Polish Zloty). The maximum price in the dataset was 3.2 million PLN, while the minimum price was 187 thousand PLN. The average housing price was 832,104 PLN. Additionally, 25% of the properties were priced below 536,500 PLN, and 75% below 999,000 PLN.

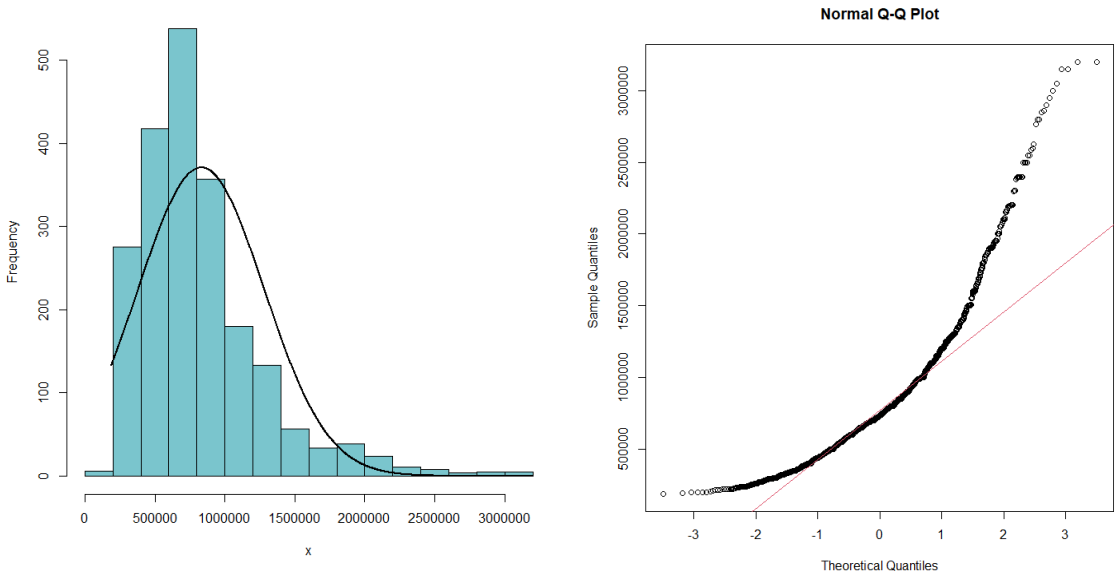
Minimum	1st Qu.	Median	Mean	3rd Qu.	Maximum
187000.00	536500.00	730000.00	832104.00	999000.00	3200000.00

Table 2. Basic statistics for the price variable.

The distribution of the *price* variable significantly deviates from a normal distribution, as evidenced by both the histogram with the normal distribution line and the Normal Q-Q Plot. A clear rightward skew of the distribution is observed. The result of the Jarque-Bera test also allows us to reject the null hypothesis of normality in the *price* distribution (p-value < 5%).

Based on literature that suggests using logarithmic transformations for housing prices in similar models, and my own analysis of the *price* variable, I decided to include the variable in its logarithmic form: *log(price)*. This significantly improved the histogram distribution, making it much closer to the normal distribution line, and the Normal Q-Q plot, where the sample quantiles now deviate less from the theoretical distribution line.

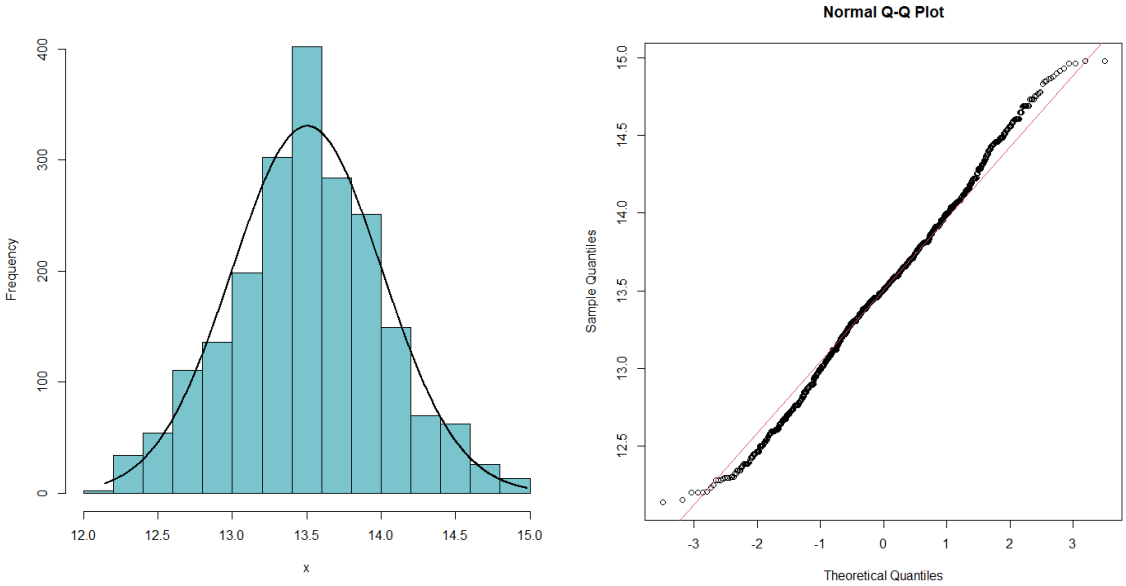
Distribution of variable *price*



```
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 2440.9719
  P VALUE:
    Asymptotic p Value: < 0.00000000000000022
```

Distribution of *log(price)*:



5.2 Continuous Independent Variables

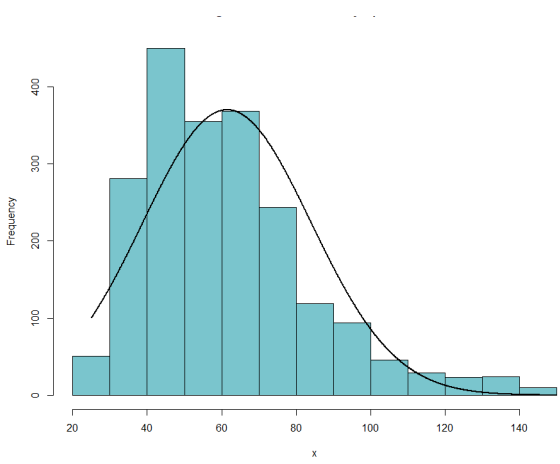
The first continuous independent variable analyzed is the apartment size in square meters, represented by the variable *squareMeters*. The smallest apartment in the dataset was only 25 m², while the largest was 150 m². The average size was 61.44 m², with 75% of the apartments being smaller than 73 m².

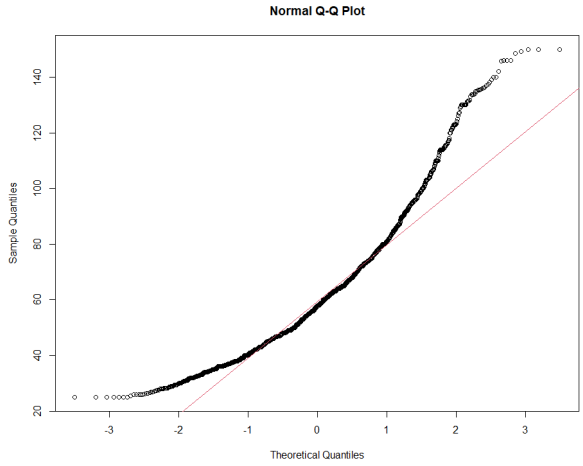
Minimum	1st Qu.	Median	Mean	3rd Qu.	Maximum
25.00	45.51	57.22	61.44	73.00	150.00

Table 3. Basic statistics for the squareMeters variable.

The apartment size distribution, as shown by the histogram and the Normal Q-Q plot, is right-skewed. This was further confirmed by the Jarque-Bera test result, with a p-value close to zero, allowing us to reject the hypothesis of normality at the 5% significance level.

Distribution of *squareMeters*:

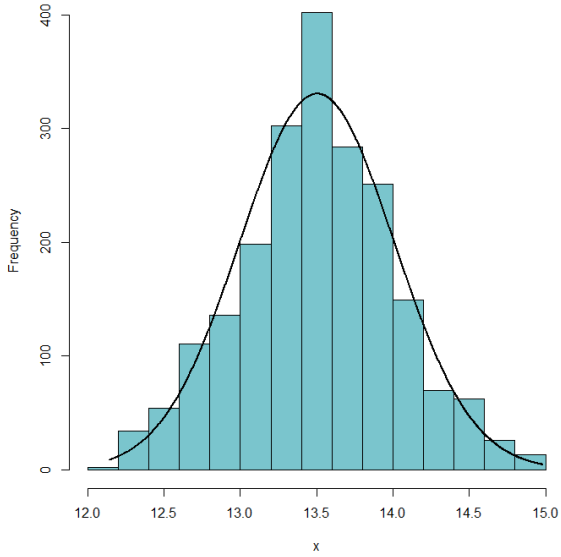


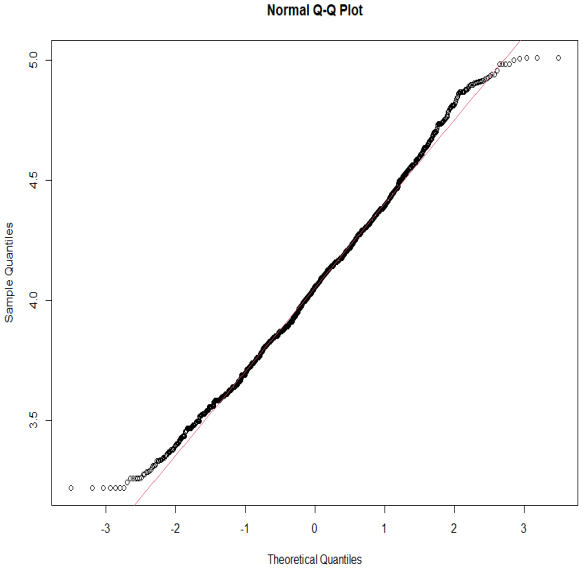



```
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 666.6859
  P VALUE:
    Asymptotic p Value: < 0.00000000000000022
```

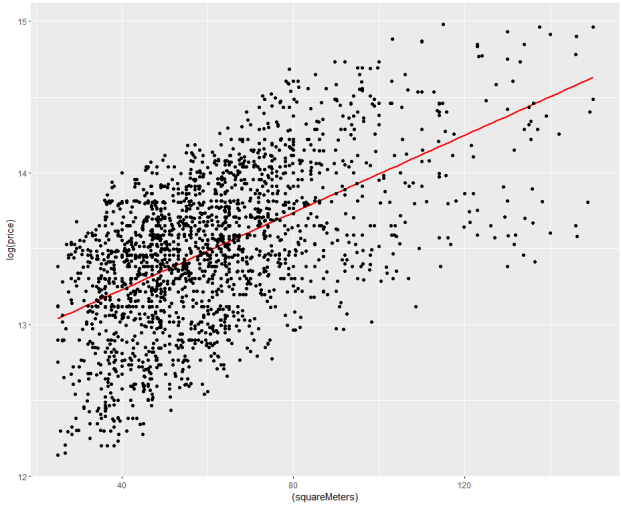
Distribution of *log(squareMeters)*:



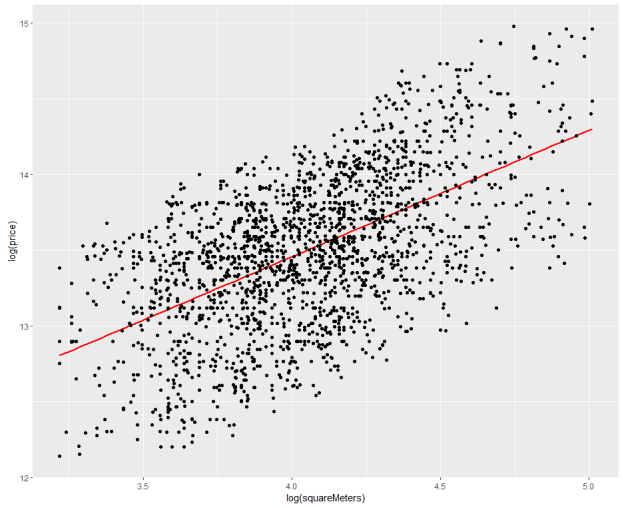


Scatterplots:

- between *log(price)* and *squareMeters*



- between *log(price)* and *log(squareMeters)*



Similar to the dependent variable, and based on scatterplots between log(price) and squareMeters, and between *log(price)* and *log(squareMeters)*, I decided to include the logarithmic form of *squareMeters* in the model. This significantly improved the distribution of the variable as well as the scatterplot, which may positively impact the functional form of the model.

In addition, variables representing the distances of apartments from points of interest, such as the city center, educational institutions, and restaurants, were treated as continuous variables and expressed in meters. Each of these variables was analyzed for distribution through

histograms, Normal Q-Q plots, and scatterplots with *log(price)*, as well as Jarque-Bera tests for normality. Most of these variables were included in their logarithmic form in the model, which, in most cases, significantly improved their distribution and scatterplots. The exceptions were *centreDistance* and *postOfficeDistance*, where logarithmic transformation did not make a significant difference but simplified the interpretation of the estimation results.

5.3 Binary Independent Variables

The model includes two property characteristics represented by dummy variables: building type and the number of rooms. To avoid the problem of linear dependence between categories, base levels were established for these variables. Typically, the least or most numerous groups are selected as the base level, and I chose the latter. The most numerous category for the *building type* variable was the group of apartments located in regular blocks of flats – 908 observations. For the number of rooms, the base category was apartments with 3-4 rooms, consisting of 1,074 observations.

Most of the surveyed apartments are owner-occupied and in good condition. Moreover, the majority are not located on the ground or top floors. Slightly more than half of the apartments have amenities such as an elevator or a balcony, while fewer (but still a substantial number) have a parking space. Only a small number of homes have security services, which is not surprising, especially in older housing estates. Notably, over 25% of the surveyed listings are from Warsaw.

VARIABLE	VALUE = 1	VALUE = 0
ownership	1802	292
condition	1367	727
hasParkingSpace	868	1226
hasBalcony	1245	849
hasElevator	1129	965
hasSecurity	95	1999
hasStorageRoom	1038	1056
first_floor	397	1697
top_floor	435	1659
stolica	631	1463
DUMMY VARIABLES		
building type	**block:**	**908**
	apartmentB:	697
	tenement:	489
rooms	rooms1:	933
	rooms2:	**1074**
	rooms3:	87

Tabela 5. Counts for binary variables.

6. Model Estimation

6.1 Initial Model

The analysis of the relationship between *price* and the explanatory variables began with refining the linear regression model to an appropriate functional form and checking for multicollinearity between factors. The assumed confidence level in the conducted tests was 5%.

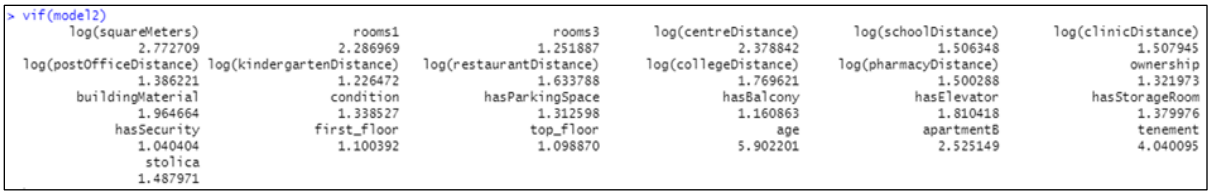
The first model (Model 1 in Table 6.) included almost all variables in the selected forms but without interactions or a separate category for apartments from the capital. The R^2 statistic was 64.3%, meaning this model explained 64.3% of the variability in housing prices in logarithmic form. The result of the RESET test, with a p-value much greater than 5%, allowed to accept the hypothesis of a linear functional form. Therefore, the first assumption of the Classical Linear Regression Model was fulfilled.

```
        RESET test

data:  model1
RESET = 0.14829, df1 = 2, df2 = 2067, p-value = 0.8622
```

Next, to check whether it was necessary to include the *stolica* (Warsaw) variable in the regression, I conducted the Chow test. This test compares whether model parameters change between two or more subgroups of data. The result was clear: the p-value close to zero indicated the need to reject the null hypothesis of stable parameters between the sample of apartments from the capital and those from outside Warsaw. I decided to include the *stolica* variable in the model (Model 2 in Table 6.), which also passed the RESET test.

Multicollinearity between independent variables is a key issue in linear regression models. This phenomenon occurs when two or more independent variables are highly correlated, leading to interpretational problems. Multicollinearity can make it difficult to assess the impact of individual variables, and parameter estimates become imprecise. The Variance Inflation Factor (VIF) was used to check multicollinearity. A VIF value higher than 5 indicates high multicollinearity, and a value above 10 suggests the need to remove correlated variables.



```
> vif(model2)
       log(squareMeters)                   rooms1                   rooms3      log(centreDistance)      log(schoolDistance)      log(clinicDistance)
                2.772709                 2.286969                 1.251887                 2.378842                 1.506348                 1.507945
 log(postOfficeDistance) log(kindergartenDistance)  log(restaurantDistance)      log(collegeDistance)     log(pharmacyDistance)                ownership
                1.386221                 1.226472                 1.633788                 1.769621                 1.500288                 1.321973
       buildingMaterial                condition           hasParkingSpace               hasBalcony              hasElevator          hasStorageRoom
                1.964664                 1.338527                 1.312598                 1.160863                 1.810418                 1.379976
           hasSecurity               first_floor                top_floor                      age               apartmentB                tenement
                1.040404                 1.100392                 1.098870                 5.902201                 2.525149                 4.040095
              stolica
                1.487971
```

Only the VIF for the *age* variable exceeded 5, indicating a relatively high correlation with other variables in the model. However, no variable qualified for removal. In the third model (Model 3 in Table 6.), several interactions that could influence the dependent variable and the estimation of coefficients were included:

- ***log(centreDistance) * log(squareMeters)*** – Apartments built in city centers tend to have smaller sizes than those in suburban areas, where population density is lower. Therefore, the relationship between "price and proximity to the center" may be affected by the size of the apartment.
- ***log(centreDistance) * hasParkingSpace*** – A parking space may be particularly important for people commuting to the city center by car. Thus, the effect of the *centreDistance* variable combined with *hasParkingSpace* could be amplified.
- ***log(schoolDistance) * hasParkingSpace*** – People driving their children to school may value having a parking space more, so the interaction between *schoolDistance* and *hasParkingSpace* could increase the willingness to pay for an apartment.
- ***ownership * tenement*** – Buyers of apartments in historic buildings may often undertake substantial renovations, and owners of privately owned apartments have more autonomy in making modernization decisions compared to those in cooperative housing.
- ***log(squareMeters) * condition*** – The condition of an apartment may be particularly important for buyers of larger properties. Larger apartments typically incur higher maintenance costs, and if the technical condition is good, the owner may avoid expensive repairs or upgrades in the future.
- ***log(centreDistance) * stolica*** – Especially in large metropolitan areas, the availability of building plots in the city center is limited, and apartments there are considered prestigious. Thus, the distance from the city center may have a more pronounced effect in Warsaw.

The third model showed the highest adjusted R^2 among the models tested so far, explaining 73.8% of the variability in the dependent variable (Table 6). In the next stage, additional

diagnostic tests were conducted to verify whether this regression model meets the assumptions of the Classical Linear Regression Model.

	Dependent variable:		
	log(price)		
	(1)	(2)	(3)
log(squareMeters)	0.791***	0.775***	0.431***
	(0.032)	(0.028)	(0.163)
rooms1	-0.010	-0.014	-0.008
	(0.020)	(0.018)	(0.017)
rooms3	0.025	0.019	0.006
	(0.037)	(0.033)	(0.032)
log(centreDistance)	0.157***	0.040***	-0.107
	(0.011)	(0.011)	(0.083)
log(schoolDistance)	-0.025**	-0.011	-0.022*
	(0.011)	(0.009)	(0.012)
log(clinicDistance)	-0.062***	-0.022***	-0.017**
	(0.009)	(0.008)	(0.008)
log(postOfficeDistance)	0.029***	0.015*	0.010
	(0.010)	(0.009)	(0.008)
log(kindergartenDistance)	-0.013	0.001	-0.001
	(0.010)	(0.009)	(0.008)
log(restaurantDistance)	-0.100***	-0.060***	-0.065***
	(0.009)	(0.008)	(0.007)
log(collegeDistance)	-0.056***	-0.049***	-0.048***
	(0.010)	(0.009)	(0.008)
log(pharmacyDistance)	-0.011	-0.011	-0.005
	(0.010)	(0.009)	(0.009)
ownership	-0.013	0.020	0.029
	(0.022)	(0.020)	(0.019)
buildingMaterial	0.121***	0.103***	0.094***
	(0.023)	(0.020)	(0.020)
condition	0.158***	0.168***	0.427***
	(0.016)	(0.014)	(0.144)
hasParkingSpace	0.026*	0.023*	0.315**
	(0.015)	(0.014)	(0.131)
hasBalcony	0.015	0.019	0.022*
	(0.015)	(0.013)	(0.013)
hasElevator	0.227***	0.153***	0.136***
	(0.018)	(0.016)	(0.017)
hasStorageRoom	-0.003	0.006	-0.003
	(0.016)	(0.014)	(0.013)
hasSecurity	0.065**	0.007	0.009
	(0.032)	(0.029)	(0.028)
first_floor	0.011	0.025	0.024
	(0.018)	(0.016)	(0.015)
top_floor	-0.031*	-0.028*	-0.033*
	(0.017)	(0.015)	(0.020)
age	-0.002***	-0.002***	-0.001***
	(0.0004)	(0.0004)	(0.0004)
apartmentB	0.139***	0.143***	0.147***
	(0.022)	(0.020)	(0.019)
tenement	0.196***	0.129***	0.245***
	(0.031)	(0.028)	(0.089)
stolica		0.364***	1.971***
		(0.016)	(0.149)
hasElevator:top_floor			0.021
			(0.029)
log(squareMeters):log(centreDistance)			0.053***
			(0.020)
log(centreDistance):hasParkingSpace			-0.064***
			(0.016)
log(schoolDistance):hasParkingSpace			0.040**
			(0.016)
ownership:tenement			-0.125
			(0.087)
log(squareMeters):condition			-0.063*
			(0.035)
log(centreDistance):stolica			-0.194***
			(0.018)
Constant	10.098***	10.419***	11.368***
	(0.170)	(0.152)	(0.670)
Observations	2,094	2,094	2,094
R2	0.644	0.717	0.738
Adjusted R2	0.640	0.714	0.734
Residual Std. Error	0.304 (df = 2069)	0.271 (df = 2068)	0.261 (df = 2061)
F Statistic	155.723*** (df = 24; 2069)	209.552*** (df = 25; 2068)	181.711*** (df = 32; 2061)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 6. Results of model 1, 2 and 3.

6.2 Diagnostic Tests

a) **Linearity of the Functional Form**

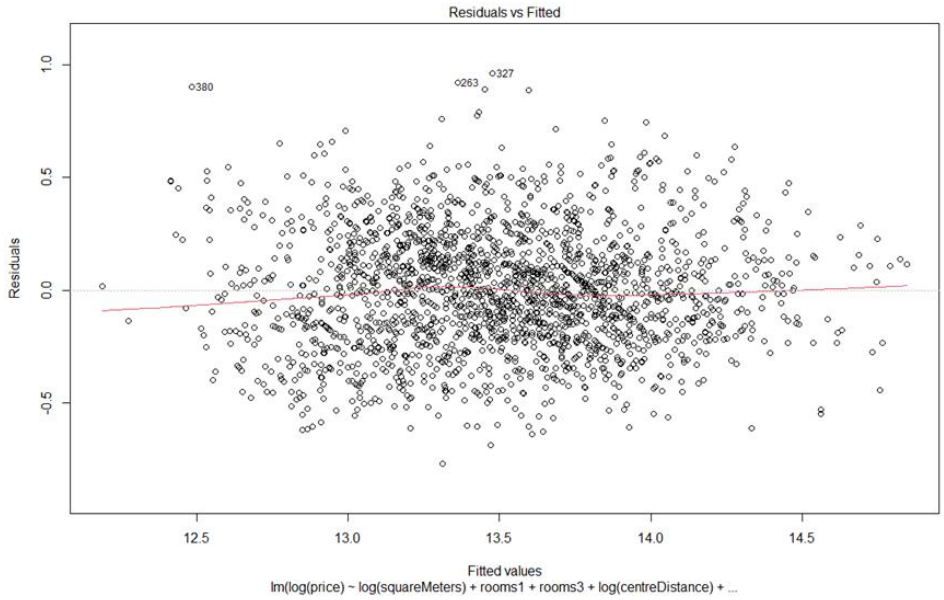
Model 3 was subjected to the RESET test, the result of which clearly allowed to accept the null hypothesis (H_0) of the correct functional form. The p-value was 0.8681 at the 5% significance level.

```
        RESET test

data:  model3
RESET = 0.14146, df1 = 2, df2 = 2059, p-value = 0.8681
```

b) **Homoscedasticity of Residuals**

Ensuring constant variance of residuals in the model is crucial because if this condition is violated, the standard errors of regression coefficients are incorrect, and consequently, the corresponding p-values are also incorrect. In such cases, the parameters of the model cannot be interpreted correctly because they are biased. I verified the assumption of homoscedasticity for Model 3 using the Residuals vs. Fitted plot and the Breusch-Pagan test.



The Residuals vs. Fitted plot allowed me to check whether there is a relationship between residuals and fitted values. Unfortunately, the plot indicated heteroscedasticity – the points tended to cluster in the middle of the plot, and the red line dropped downward at the beginning of the x-axis. The result of the Breusch-Pagan test confirmed this, with a p-value close to zero, leading to the rejection of the null hypothesis (H_0) of homoscedasticity.

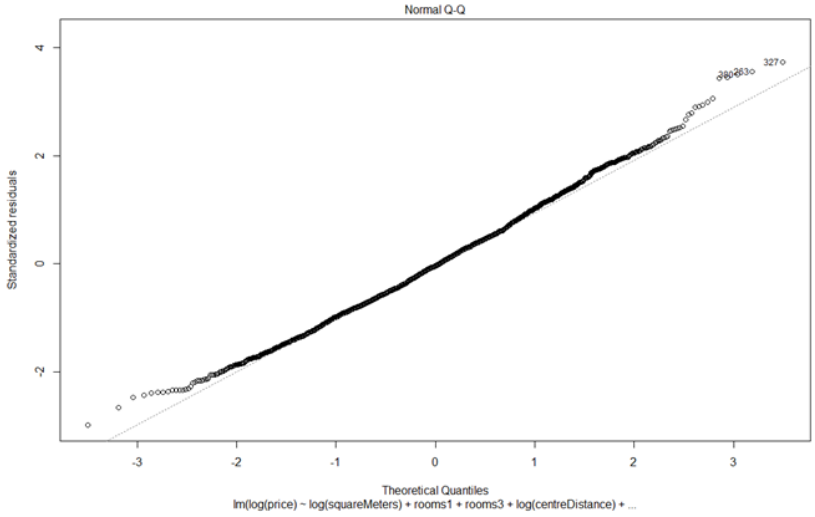
```
        Breusch-Pagan test

data:  model3
BP = 373.87, df = 32, p-value < 0.00000000000000022
```

c) **Normality of Residuals**

Testing the assumption of normally distributed residuals is especially important in cases with small samples. If this condition is not met, standard inference cannot be applied because the statistics do not follow t or F distributions.

To test this assumption for Model 3, I used the Normal Q-Q Plot and the Jarque-Bera test for normality.




```
Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 18.3239
  P VALUE:
    Asymptotic p Value: 0.000105
```

The Normal Q-Q Plot showed that most points lay on the dashed line, except for the points at the ends of the axes, which significantly disrupted the normal distribution. The result of the Jarque-Bera test (p-value = 0.000105) confirmed the need to reject the null hypothesis of normality.

However, this is not a concern because the dataset used in the study included 2094 observations, which can be considered a large sample. In this case, we can rely on standard inference due to the Law of Large Numbers and the Central Limit Theorem.

6.3 Final Model

To address the heteroscedasticity problem in the model and ensure proper interpretation of the parameters, I applied a robust variance-covariance matrix (White's correction). I chose this method because the heteroscedasticity likely resulted from the influence of several variables, which would exclude the use of Weighted Least Squares (WLS), and identifying them would be time-consuming (thus excluding the use of GLS).

Using the robust matrix affected the significance levels of the variables. Many of the variables included in the initial model turned out to be statistically insignificant at the 5% confidence level. To eliminate them from the model, I applied stepwise elimination, removing one variable at a time, starting with the one with the highest p-value.

The final model included 15 variables and 4 interactions (Table 7). It passed the RESET test for functional form correctness, and the lack of normality in the residuals is not a problem, as explained earlier. Heteroscedasticity was minimized using the robust variance-covariance matrix. This final model will now serve as the basis for interpreting the coefficients and verifying the hypotheses formulated at the beginning of the study.

Test RESET for final model:

```
        RESET test

data:  model16
RESET = 0.70578, df1 = 2, df2 = 2072, p-value = 0.4938
```

Test Breuscha – Pagana for final model:

```
        Breusch-Pagan test

data:  model16
BP = 338.48, df = 19, p-value < 0.00000000000000022
```

Test Jarque – Bera for final model:

```
Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 17.7344
  P VALUE:
    Asymptotic p Value: 0.0001409
```

	Dependent variable:	
	log(price)	
	OLS	coefficient test
	(1)	(2)
log(squareMeters)	0.430***	0.430**
	(0.160)	(0.184)
log(centreDistance)	-0.093	-0.093
	(0.082)	(0.091)
log(schoolDistance)	-0.021*	-0.021*
	(0.011)	(0.012)
log(clinicDistance)	-0.018**	-0.018**
	(0.008)	(0.008)
log(restaurantDistance)	-0.064***	-0.064***
	(0.007)	(0.007)
log(collegeDistance)	-0.049***	-0.049***
	(0.008)	(0.008)
buildingMaterial	0.101***	0.101***
	(0.018)	(0.017)
condition	0.171***	0.171***
	(0.014)	(0.014)
hasParkingSpace	0.289**	0.289**
	(0.131)	(0.143)
hasElevator	0.140***	0.140***
	(0.015)	(0.015)
top_floor	-0.031**	-0.031**
	(0.014)	(0.014)
age	-0.002***	-0.002***
	(0.0004)	(0.0004)
apartmentB	0.150***	0.150***
	(0.019)	(0.017)
tenement	0.132***	0.132***
	(0.027)	(0.028)
stolica	2.018***	2.018***
	(0.147)	(0.132)
log(squareMeters):log(centreDistance)	0.049**	0.049**
	(0.020)	(0.022)
log(centreDistance):hasParkingSpace	-0.062***	-0.062***
	(0.016)	(0.017)
log(schoolDistance):hasParkingSpace	0.042**	0.042***
	(0.016)	(0.016)
log(centreDistance):stolica	-0.199***	-0.199***
	(0.018)	(0.016)
Constant	11.414***	11.414***
	(0.652)	(0.748)
Observations	2,094	
R2	0.736	
Adjusted R2	0.734	
Residual Std. Error	0.261 (df = 2074)	
F Statistic	304.809*** (df = 19; 2074)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 7. Results of the final model before and after using the robust matrix.

7. Results

The model explains 73.6% of the variability in the dependent variable. Below is the interpretation of the estimated β coefficients for all variables and interactions. The constant was deliberately omitted, as its interpretation in this context does not make sense.

- **log(squareMeters)**: A 1% increase in apartment size results in an average price increase of 0.43%, holding other factors constant. This variable is statistically significant at the 5% confidence level.
- **log(centreDistance)**: The coefficient indicates a negative relationship between price and distance from the city center, but it is not statistically significant in this model.
- **log(schoolDistance)**: Although not statistically significant at the 5% level, the coefficient indicates a negative relationship between apartment price and distance from the nearest school. A 1% increase in distance translates into an average price decrease of 0.021%.
- **log(clinicDistance)**: A 1% increase in distance from the nearest clinic results in a price decrease of 0.018%. This variable is statistically significant at the 5% level.
- **log(restaurantDistance)**: A 1% increase in distance from the nearest restaurant results in a price decrease of 0.064%. This variable is statistically significant at the 1% level.
- **log(collegeDistance)**: A 1% increase in distance from the nearest university results in a price decrease of 0.049%. This variable is statistically significant at the 1% level.
- **buildingMaterial**: Apartments built with brick are, on average, 10.1% more expensive than those built with other materials, holding other factors constant. This variable is statistically significant at the 1% level.
- **condition**: Apartments in better condition are, on average, 17.1% more expensive than those in worse condition. This variable is statistically significant at the 1% level.
- **hasParkingSpace**: Apartments with a parking space are, on average, 28.9% more expensive than those without one. This variable is statistically significant at the 5% level.
- **hasElevator**: Apartments with an elevator in the building are, on average, 14% more expensive than those without one. This variable is statistically significant at the 1% level.
- **top_floor**: Apartments on the top floor are, on average, 3.1% cheaper than those on other floors. This variable is statistically significant at the 5% level.

- **age**: An apartment that is one year older is, on average, 0.2% cheaper than a newer apartment. This variable is statistically significant at the 1% level.
- **apartmentB**: Apartments in apartment buildings are, on average, 15% more expensive than those in regular blocks. This variable is statistically significant at the 1% level.
- **tenement**: Apartments in tenement houses are, on average, 13.2% more expensive than those in regular blocks. This variable is statistically significant at the 1% level.
- **stolica**: The coefficient for this variable indicates that apartments in Warsaw are, on average, 6.52 times more expensive than apartments in other cities. This variable is statistically significant at the 1% level.

8. Hypotheses Verification

Based on the final model, the hypotheses formulated at the beginning of the study were verified as follows:

H1 (Confirmed): The size of the apartment is significant and positively correlated with price. A 1% increase in size leads to a 0.43% increase in price.

H2 (Rejected): Contrary to expectations, the number of rooms was not statistically significant in the final model, and its relationship with the price was not confirmed. The hypothesis was rejected.

H3 (Confirmed): The condition of the apartment has a significant and positive impact on price. Apartments in better condition are 17.1% more expensive than those in worse condition.

H4 (Partially Confirmed): Proximity to key amenities, particularly clinics, restaurants, and universities, impacts housing prices. The distance to schools was not statistically significant, but the other amenities were. Therefore, the hypothesis is partially confirmed.

H5 (Rejected): The ownership type variable was not significant in the final model, so the impact of ownership on price was not confirmed. This hypothesis is rejected.

H6 (Confirmed): Apartments built with brick are more expensive than those made of other materials. The average premium is 10.1%.

H7 (Partially Confirmed): Apartments on the top floor are cheaper than those on other floors (by 3.1%), but the hypothesis that apartments on the first floor are the most expensive was not confirmed. Therefore, this hypothesis is partially confirmed.

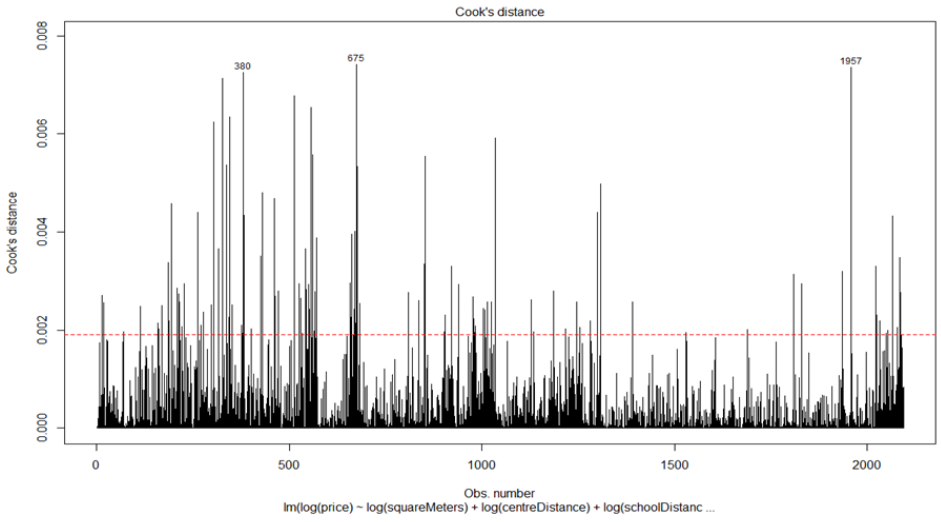
H8 (Partially Confirmed): Amenities such as parking spaces and elevators positively affect the price. However, variables such as balconies or security services did not show statistical significance in the final model. Therefore, the hypothesis is partially confirmed.

H9 (Confirmed): The age of the apartment is negatively correlated with price. For every additional year, the apartment's price decreases by 0.2%.

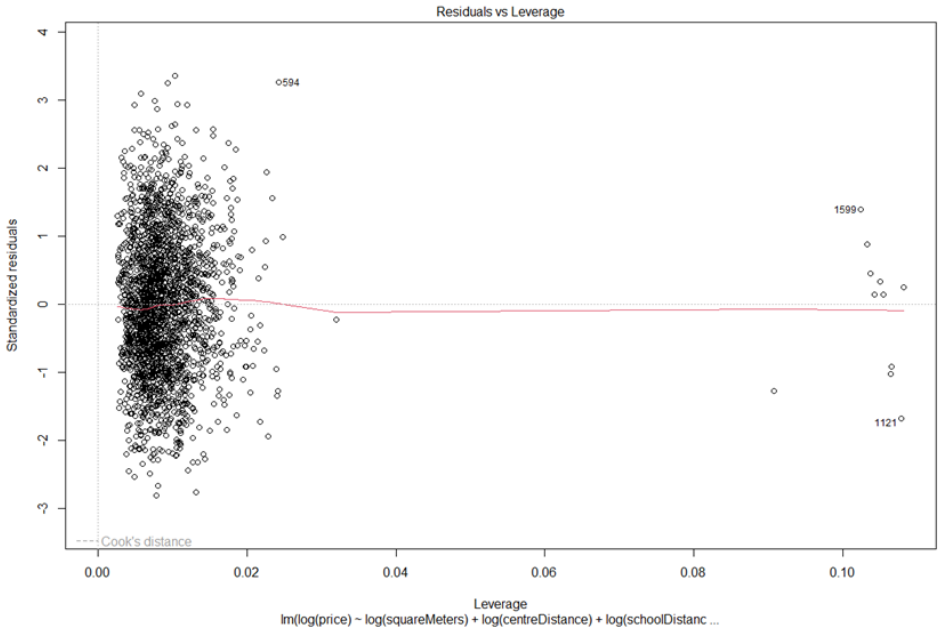
H10 (Confirmed): Apartments in the capital city, Warsaw, are significantly more expensive than those in other cities. The difference is substantial, with prices in Warsaw being 6.52 times higher on average.

9. Issues with Observations

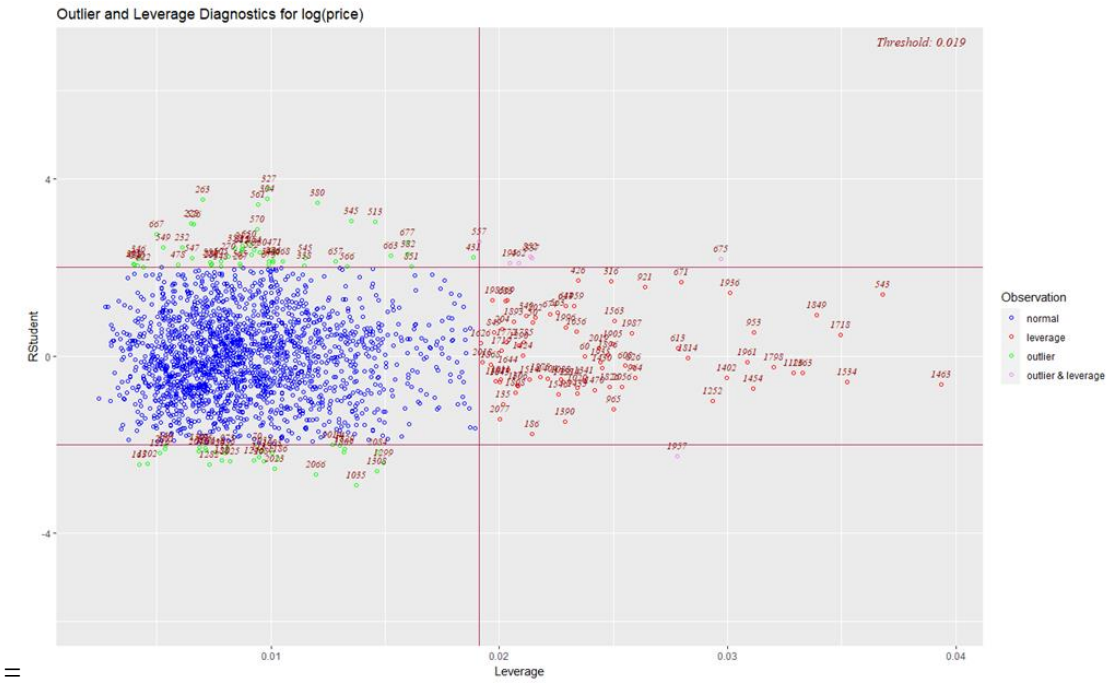
In econometric modeling, it is important to identify outliers or erroneous observations. The term "outlier" refers to observations that have unique characteristics compared to the rest of the dataset. Erroneous observations, on the other hand, are those whose occurrence cannot be explained within the framework of the theoretical model and typically arise from data entry or collection errors. Both types of observations can negatively affect model estimation. The impact of outliers can vary depending on their position relative to the regression line. To identify such variables in my model, I analyzed the residuals and leverage using plots and an appropriate indicator – Cook's Distance.



The observations with the highest Cook's Distances were numbers 675, 1957, and 380. This suggests that they may have had a significant impact on the regression.



Observations with particularly high leverage values included numbers 1121, 1463, 1500, and 543. These were located far from the cloud of other points but relatively close to the regression line. Therefore, they could be classified as **good leverage points**.



=

A significant portion of the observations in the model are normal/typical observations. However, some cases could be classified as **vertical outliers**, which are points that do not influence the slope of the regression curve but rather its height. Such observations include 327, 1035, and the previously mentioned 380. These were characterized by low leverage but high residuals.

Taking into account three conditions simultaneously: leverage > 2k/n, |std.errors| > 2, and Cook's Distance > 4/n, I identified six atypical observations that could be classified as **bad leverage points** (Table 8). These points are far from the cloud of other observations and significantly influenced parameter estimation, as they exhibit both high leverage and high residuals.

Observations 195, 337, 462, and 852 are very old apartments without additional amenities such as an elevator or parking space, yet they have large floor areas and very high, puzzling prices. However, these are tenement buildings, which may have significant historical value or be located in prestigious areas. It might be necessary to introduce another interaction term into the model: **tenement * age**, as an increase in age for tenement buildings may not necessarily lead to a price decrease—in fact, it may have the opposite effect.

Observation 675, with a price above 2 million PLN, stands out with an unusually high price compared to other apartments with similar characteristics. Likely, this is due to the fact that it is an apartment in a luxury building, perhaps located in a prestigious area. It would be worthwhile to examine the property prices in the neighborhood of this apartment.

The last atypical observation, number 1957, is an apartment in a block likely built over 40 years ago with lower-quality materials, lacking additional amenities, and with an average floor area. However, it has a relatively high price, which may be due to its location in an expensive district of Warsaw or proximity to the metro, factors that were not included in the model.

Nr	**195**	**337**	**462**	**675**	**852**	**1957**
price	1049000	1950000	1950000	2190000	720000	725000
centreDistance	140	590	710	270	180	630
squareMeters	72.13	150.00	150.00	95.89	46.59	67.60
schoolDistance	245	291	228	136	312	243
clinicDistance	1090	528	295	672	330	131
restaurantDistance	60	22	106	70	12	59
collegeDistance	307	424	695	531	301	229
buildingMaterial	1	1	1	1	1	0
Condition	1	0	0	1	0	1
hasParkingSpace	0	0	1	1	0	0
hasElevator	0	0	0	1	0	0
Top_floor	0	0	0	0	1	0
Age	86	95	95	1	126	44
apartmentB	0	0	0	1	0	0
Tenement	1	1	1	0	1	0
stolica	0	0	0	0	0	1

Tabela 8. Atypical observations.

10. Conclusions

This study provides valuable insights into the factors influencing housing prices in Poland's secondary market. The final model explains 73.6% of the variability in the dependent variable (logarithmic form of price), which is a satisfactory result for this type of analysis.

Key Findings:

- **Apartment size** and **condition** are the most important factors affecting prices. Larger apartments and those in better condition tend to be more expensive.
- **Proximity to amenities**, especially clinics, restaurants, and universities, significantly impacts prices. People are willing to pay more for properties closer to these locations.
- **Building material** and **age** of the property also play a role. Apartments made of brick are more expensive, while older apartments are cheaper.
- **Location** is a major determinant of price. Apartments in Warsaw are significantly more expensive than those in other cities.

Future Research:

While this study offers a comprehensive analysis of housing prices, future research could focus on several areas for further exploration:

- **City-specific analyses**: Separate models for individual cities could provide a more detailed understanding of local markets and allow for comparison across regions.
- **Public transport access**: Incorporating data on proximity to public transport and transit hubs could refine the analysis, as access to transportation may also influence housing prices.
- **Quality of surrounding infrastructure**: Future studies could examine how infrastructure quality (roads, utilities) affects prices.

BIBLIOGRAPHY

1. Bourne Larry Stuart. *The geography of housing*. Londyn, 1981.

2. D'Acci Luca. *Quality of urban area, distance from city centre, and housing value. Case study on real estate values in Turin. Cities*. Vol 91, 71-92, ISSN 0264-2751. 2019.

3. Lancaster Kelvin J. *A New Approach to Consumer Theory. Journal of Political Economy*, Vol. 74, No. 2, 132-157. 1966.

4. Owusu-Ansah Anthony. *A review of hedonic pricing models in housing research. Journal of International Real Estate and Construction Studies*. ISSN: 2153-6813. Vol. 1, Nr 1. 2011.

5. Pricewaterhouse Coopers. *Skąd ten boom? Zmiany na rynku mieszkaniowym w Polsce*. 2021.

6. Wittowsky Dirk, Hoekveld Josje, Welsch Janina, Steier Michael. *Residential housing prices: impact of housing characteristics, accessibility and neighbouring apartments – a case study of Dortmund, Germany. Urban, Planning and Transport Research*, 8:1, 44-70, DOI: 10.1080/21650020.2019.1704429. 2020.

7. Tomczyk Emilia, Widłak Marta. *Konstrukcja i własności hedonicznego indeksu cen mieszkań dla Warszawy. Bank i Kredyt* 41 (1), 99–128. 2010.

8. Adair Alastair, McGreal Stanley, Smyth Austin, Cooper James, Ryley Tim House *Prices and Accessibility: The Testing of Relationships within the Belfast Urban Area. Housing Studies*, 15:5, 699-716, DOI: 10.1080/02673030050134565. 2000.

9. Wen Haizhen, Xiao Yue, Zhang Ling. *School district, education quality, and housing price: Evidence from a natural experiment in Hangzhou, China. Cities,* Vol. 66, 72-80. 2017.

10. Małecka Katarzyna. *Badanie cen na rynku nieruchomościmieszkaniowych w Łodzi*. *Świat Nieruchomości,* 3(69), 52-55. 2009.