Econometrics - Project

Cross-sectional Analysis on Rental House Prices in Warsaw

Author: Michele Guderzo
(project developed in collaboration with a teammate)

Course Coordinator: Mgr Kateryna Zabarina

Warsaw, Academic Year 2021/22

Contents

1 Introduction **3**
1.1 Presentation . . . 3
1.2 Literature Review . . . 4
1.3 Hypothesis . . . 6

2 Data analysis **7**
2.1 Data description . . . 7
2.2 Statistical analysis . . . 8
2.2.1 Continuous variables . . . 9
2.2.2 Discrete variables . . . 14
2.2.3 Dummy variables . . . 14
2.2.4 Correlation analysis . . . 15

3 Model estimation **16**
3.1 Diagnostic . . . 18
3.1.1 RESET test . . . 18
3.1.2 Residuals Analysis . . . 19
3.1.3 Homoscedasticity . . . 20
3.1.4 Multicollinearity . . . 21
3.1.5 Diagnostic Conclusion . . . 22
3.2 Problems with Data . . . 22
3.2.1 Non-typical and wrong observations . . . 22
3.2.2 Collinearity . . . 24

4 Interpretation and Conclusion **25**
4.1 Conclusions . . . 26
4.2 Forecasting . . . 26

5 References **27**

1 Introduction

1.1 Presentation

The labour market, which is looking for people who are open to mobility, and the phenomenon of globalisation, which has broadened the horizons of individuals, have given impetus to a specific business in the real estate market that 40 years ago represented only a small share of it: the housing rental market. The market for buying and selling houses continues to suffer from severe contractions due to the mismatch between supply and demand that often occurs. Transaction volumes are decreasing and sales times are increasing due to more cautious buyers.

In addition, the sharp contraction in access to credit by credit institutions makes it even more difficult for an individual to buy a flat.

On the other hand, the housing rental market, being more liquid than the buying and selling market, has started to gain in prestige. Factors such as precarious conditions in the labour market and, consequently, the need for temporary accommodation and the availability of small periodic amounts of capital to pay for rent, have meant that it is now more common to rent a flat than to buy one.

For years, the international literature has focused mainly on the analysis of sales prices in the real estate market, paying little attention to the housing rental market.

The housing rental market has a significant size in the world, surpassing in recent years, in terms of number of transactions, the market of the sale of houses.

To give an idea of the size of the rental business, an analysis conducted by PwC, REAS and CMS entitled "Institutional Rental Market in Poland" revealed that in 2002 21% of households in Warsaw were rented rather than owned.

We, the authors of this paper, as Erasmus students in Warsaw, have already been through this topic and are therefore directly involved in it. Our temporary status as students at the University of Warsaw prompted us to look for temporary accommodation like many of our colleagues, paying particular attention to rented houses and specific student residences. Therefore, it is our interest to study and analyse the dynamics of rental prices in the city of Warsaw and, consequently, to find out which factors may be the most impactful.

However, knowing what most determines the price of renting a flat is a topic of interest to a heterogeneous group of people, including not only students planning to study in a city other than their city of residence, but also workers who find employment in others cities and even flat owners who intend to enter the rental business.

In the next section we will analyse some of the work in the literature on this topic so far, which will help us to formulate our hypotheses.

1.2 Literature Review

As mentioned in the previous paragraph, the older literature focused more on the selling price of flats. It is only in the last 10 to 15 years that the number of scientific researches on the phenomenon of rental housing and its prices has increased compared to studies on the purchase price of houses.

In the process of selecting scientific articles, we paid particular attention to which variables the researchers used rather than selecting studies with the same dependent variable, in some cases preferring the sale price of houses rather than the rental price.

Thus, Lishun Yuan in (2019) "A Regression Model of Single House Price in LA Constructing a Predicted Model for House Prices" (2019) attempted to construct a multiple linear regression model based on 140 observations that relate to house sales prices in seven cities in Los Angeles County.

The study found that, of the seven independent variables considered, the number of bathrooms in the flat is the aspect that has the greatest effect on its selling price, while the number of bedrooms has a negative effect on the price.

The author logarithmically transformed the dependent variable, thus, a one-unit increase in the number of bathrooms leads to a 9% increase in the sales price of flats, while a one-unit increase in the number of bedrooms leads to a 6.1% decrease in the sales price.

Other regressors were considered by M.Y. Aminah and I. Syuhaida, authors of the study (2012) "Multiple Regressions in Analysing House Price Variations".

The authors' intent was to explain the price variation for flats located in the city of Kuala Lumpur, Malaysia based on 1500 houses.

Two models were created, one for the year 2000 and one for 2007, and the explanatory power of both is quite high, with a R^2 of 77.2% and 83.6%.

From this study the variable *locality* turns out to be the characteristic that most influences the price to the detriment of variables such as the *age* of the house.

The price of house rental potentially has a number of factors that can influence it, ranging from economic to demographic and even geographical.

Abebaw Hailu Fikire in (2021) "Determinants of residential house rental price in Debre Berhan Town, North Shewa Zone, Amhara Region, Ethiopia" analyzed data collected through questionnaires for 385 residential houses in Debre Berhan Town, North Shewa Zone, Amhara Region, Ethiopia in the year 2019/2020.

Abebaw's study in this case starts from a demographic issue, as through the determination of the factors that most influence the rental price of houses, the author wanted to provide an additional insight into the problem of the rising rental price of flats in Debre Berhan, which was strongly correlated to the rapid demographic growth that is affecting this particular region of Central East Africa.

The study found that access to water and to bathroom are the main factors determining the rental price

of houses. In line with the results of Yuan, which considered the number of bathrooms, the presence of bathroom access increased the rental price by 36.4%, while the presence of water access increased it by 36.8%. Conversely, the year of the dwelling negatively affected the rental price.

Gustafsson and Wogenius' (2014) "Modelling Apartment Prices with the Multiple Linear Regression Model" study focuses more on the structural factors of an apartment that can influence its selling price. The thesis researches, through the multiple linear regression model, the characteristics that have a statistically significant effect on flat prices in central Stockholm. The constructed model is based on 8164 observations, i.e. the flats sold in central Stockholm for the years 2012 and 2013. The result of this research is that the final model predicts house prices by a percentage of 91%, and the regressors that are statistically significant for the sales price are *location*, *year of construction* and the variable *penthouse*, which indicates whether the house is on a penthouse or not.

While the variable *ground place*, which indicates whether the house is on the ground floor or not, is the variable that most negatively influences the sales price.

Cohen and Karpaviciute in (2016) "The analysis of the determinants of housing prices", on the other hand, preferred to place the analysis of the problem in a purely financial and macroeconomic context, selecting variables such as GDP, unemployment and interest rate.

This study evaluates the impact of these economic and financial factors on house prices in Lithuania in the period from 2001 to 2014 through a multiple linear regression analysis.

The authors, through the Granger causality test, which is used to determine the causality between variables in a model, eliminated *inflation*, *interest rate* and *emigration*. The final outcome of this research led to excellent results, with a R^2 of 98.76%.

The second most statistically significant variable is *GDP*, logarithmically transformed: a 1% increase in GDP leads to a 46.05% increase in house prices, also logarithmically transformed. *Unemployment*, on the other hand, is the variable that negatively affects sales prices, with a 13.67% decrease in the latter in response to a 1% increase in the former.

From the study, thus, it follows that macroeconomic factors are more influential than financial factors such as *credit conditions* and the *interest rate*, whose impact on the dependent variable is not yet clear.

While searching for past studies dealing with our topic, we came across numerous works employing hedonic regression. This type of regression has as dependent variable the price of a good and as independent variables the so-called attributes of the good. Hedonic regression uses ordinary least squares or more advanced regression techniques and is particularly used in the real estate market.

The last selected study uses precisely a hedonic regression to determine the impact of the various components of a house on its price. Darfo-Oduro, in (2020) "Determinants of Residential House Rental Prices in Accra Metropolis", determines the importance of neighbourhood attributes and structural attributes on residential rental prices in the city of Accra, Ghana, based on 310 questionnaires.

The author specified 4 models with different functional forms, as he found disagreements among scholars in the literature in the specification of the regression equation.
The results of this study are interesting because they tell us that the linear model has greater explanatory power than the logarithmic models. The distance to the place of worship is significant in all logarithmic models and is directly related to the monthly rent in the log-log and linear-log models and inversely related in the log-linear model.
In general, the characteristics of the neighbourhood and the distance of the house from the various centres of agglomeration and services were found to be more decisive than the structural characteristics of the flat, such as the size of the bedroom or the age of the building.
On the basis of the results obtained by the authors of these scientific articles we are now able to formulate the hypotheses for our study and check whether the subsequent statistical-quantitative results are consistent with our theoretical expectations.

1.3 Hypothesis

As we have seen above, the factors that can significantly determine the price of renting a house can belong to more or less different fields: from the economic field, to the financial field, to the demographic one. We tried to select the factors which, in our opinion, are a direct consequence of price determination; the main characteristics to which an individual pays most attention when looking for a house to rent. This consideration has led us to select more structural characteristics of the dwelling such as the year of construction of the building or house, the size of the flat in square metres, the number of rooms the dwelling has and the number of the floor on which the apartment is located.
In addition to these structural attributes, another variable, which in our opinion has a direct influence on the search for a flat, and which is therefore reflected in the rental price, is the location of the house. An individual looking for a rental house pays a lot of attention to the district to which the accommodation belongs. This regressor should, in our opinion, capture the qualitative characteristics of a district, since the further the district, and thus the house, is from the city centre, the lower the rental price of a flat. The latter observation is a simplification because the qualitative characteristics of a district are not limited to its distance from the centre, but include many other factors such as the presence of certain services in that district or how efficient the public transport service is within that district.
First of all, one of the most important elements to consider is the size of the accommodation. Since the price of a flat is paid per square metre, we consider it highly probable that the area can influence the rental price of a tenancy with direct proportionality: the larger the size, the higher the price that will be charged for the rental.
Another factor we have taken into consideration is the number of rooms. It is likely that this variable

tends to have a similar behavior to the one mentioned above. Specifically, we are inclined to think that the two variables are correlated with each other since, probably, a property with a higher number of rooms will also have a larger size. In the next sections, as we proceed with the study, we will be able to verify the truthfulness of this theory.

We do not think that, on the other hand, the floor can have such an impact. It is our opinion that the number of the floor on which the rented flat is located does not affect its price significantly, since, according to an individual, this characteristic does not alter his or her lifestyle in a decisive way, and is therefore considered as marginal when choosing a flat, even more so if the building is equipped with a lift.

The year of construction of the property was also taken into account. We believe that this variable cannot provide substantial variation in the response variable, as it tends to be of secondary importance to people looking for a rental.

Finally, we have considered the different districts which are present in the city of Warsaw. In general, it is well known that the cost of living, as well as the price of accommodation, is higher in central areas than in peripheral areas. Therefore, we believe that the geographical location of the districts reflects this trend: the highest prices will be distributed around the city centre, while as you move towards the periphery prices will tend to fall.

2 Data analysis

2.1 Data description

The initial dataset includes 85 variables and 3472 observations for as many flats in the city of Warsaw, Poland, in the year 2021. The dataset was provided by Kaggle.com. The variables selected from the dataset concern the number of square metres of the flat, the number of rooms, the number of the floor on which the property is located, the year of construction of the building and the district to which it belongs. All other variables in the dataset were not taken into account for the multiple linear regression analysis.

The original dataset contained 22 dummy variables for the districts, but by cross-checking on internet we saw that 4 districts in the dataset did not match, namely: Centrum, Metro Wilanowska, Warszawa and Mazowieckie. Hence, we proceeded to eliminate them and their observations that fell within the above-mentioned districts, narrowing the number to 18 total districts and the number of observations to 3347.

The analyses concerning the regression model, the estimation of coefficients and the discussion of the results were carried out with the software RStudio.

In our dataset we have the **dependent variable** y_i (*"gross_price"* in RStudio), continuos variable that identifies the rental price of the *i-th* flat in the city of Warsaw.

The **independent variables**, on the other hand, are the following:

- x_1 (*"area"*): continuous variable corresponding to the square footage of the lease;
- x_2 (*"room_num"*): discrete variable that counts how many rooms are in the flat;
- x_3 (*"floor"*): discrete variable indicating which floor the flat is located on;
- x_4 (*"year_built"*): continuous variable which corresponds to the year in which the house was ready to be lived in. It has structure "YYYY,MM", therefore the decimal part indicates fraction of one year;
- $x_5, ..., x_{22}$ (*"districts"*): dummy variables referring to the district to which the *i-th* observation belongs. They have the value '1' if the *i-th* flat falls within that specific district, '0' otherwise. In our model we include 18 districts: Bemowo, Białołeka, Bielany, Mokotów, Ochota, Praga-Południe, Praga-Północ, Rembertów, Targówek, Ursus, Ursynów, Wawer, Wesoła, Wilanów, Wola, Włochy, Śródmieście e Żoliborz.

2.2 Statistical analysis

In this section we performed a statistical analysis, providing the main statistical indicators, variable per variable.

First of all, as mentioned in the previous paragraph, we proceeded to eliminate from the initial dataset the variables that we will not consider in our linear regression analysis, reducing the number of independent variables. The remaining variables are: gross price (dependent variable), area of the flat, number of rooms, floor number, year of construction and all the districts.

Then, we converted the variable *year_built*, which indicated the year in which the apartment building was built, into the variable *age*, which instead indicates the age of the flat, changing the relationship with the dependent variable which will now be inversely proportional.

As last, we removed the 4 unmatched districts from the cross-checking on internet by deleting the respective observations which had 1 for these districts.

Our final dataset contains 22 independent variables and 3253 observations.

2.2.1 Continuous variables

The final dataset contains three continuous variable which are the dependent variable, *gross_price*, *area* and *age*. Below we presented the main statistical metrics of the dependent variable by implementing the functions *summary*, *sd*, *skewness* and *kurtosis* on *RStudio*.

Table 1: Gross Price - Summary Statistics

	Gross Price
Min	581.4
1st Qu.	2200.0
Median	2584.5
Mean	2747.8
3rd Qu.	3000.0
Max	11500.0
Std. Dev.	919.0068
Skewness	2.716067
Kurtosis	13.45812

A first, immediate indication that *gross_price* is not normally distributed are the values of Skewness and Kurtosis, respectively greater than 0 and greater than 3. With a value greater than zero, the skewness is positive, so the *gross_price* is skewed right, meaning that the right tails are longer than the left ones. The kurtosis value, which is greater than 3, indicates a leptokurtic distribution that in general has heavier tails and therefore a higher probability of extreme outlier values as we will see later with the boxplot. In addition to the numerical analysis, we used a graphical analysis to help us understand how *gross_price* distribution differs from the Normal distribution. To further help the reader, we overlapped the density function of the Normal distribution on the density of the *gross_price* variable.

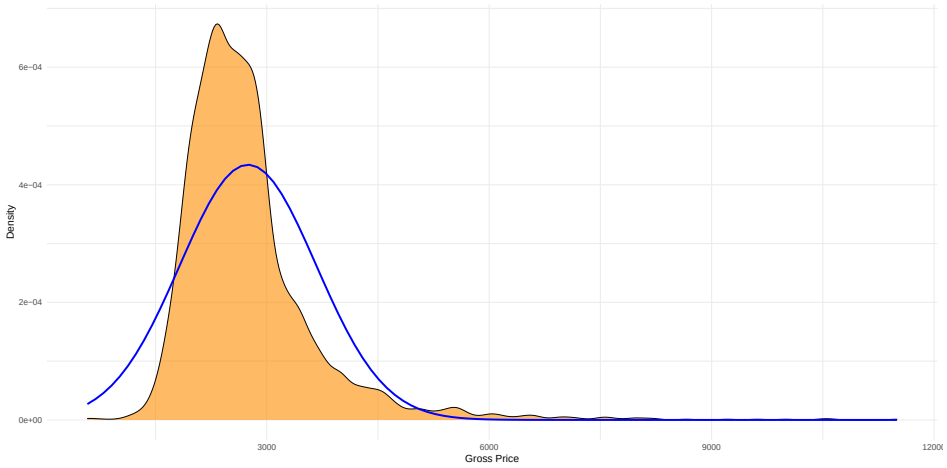

Figure 1: Gross Price vs Normal Density Distribution

A deviation of the *gross_price* density from the density function of the Normal is easily visible from the graphical comparison of the two curves. Most of the data is concentrated between the values 2000 PLN and 3000 PLN, therefore, we can interpret this saying that the most common price for renting a flat in Warsaw is around 2000/3000 PLN. This is confirmed by our previous numerical analysis, as the mean *gross_price* is 2747.8.

As already seen with the summary, the density chart shows us a pronounced asymmetry of the *gross_price* density function. More precisely, the data indicate a right-skewness, and therefore heavy tails on the right side, given a huge deviation from the rest of the set.

We further investigated this behaviour of our data with some statistical tests that are widely used to check whether and how much data density function differs from Normal one.

The tests for normality we considered are Shapiro-Wilk test and Jarque-Bera test.

Both tests have as null hypothesis that data are normally distributed. Following the graphical results, since both p-value tests are really low respect to the significance levele $\alpha = 0.05$, we reject the gaussian hypotesis.

The deviation from the Normal distribution can be checked with other two graphs: Quantile-Quantile plot and Boxplot.

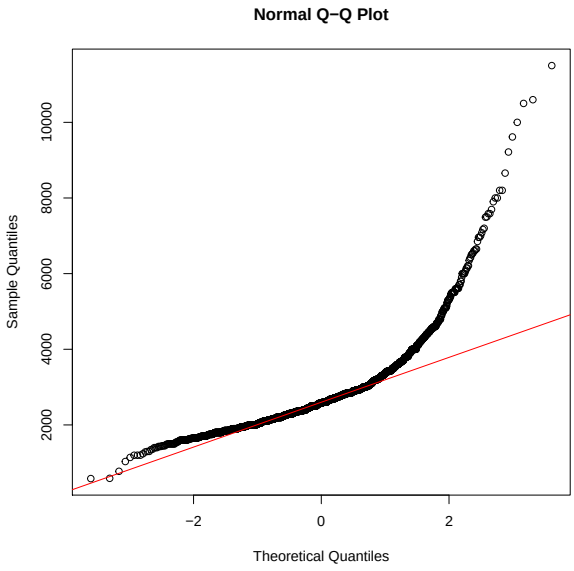

Figure 2: Gross Price Q-Q plot

The Q-Q plot is a plot of the quantiles of the *gross_price* data against the quantiles of the Normal distribution. By means of the Q-Q plot we can better see how the tails of the *gross_price* distribution deviate from those of the Normal distribution. In accordance with density graph and statistical tests, Q-Q plot presents right-side deviation from Normal distribution.

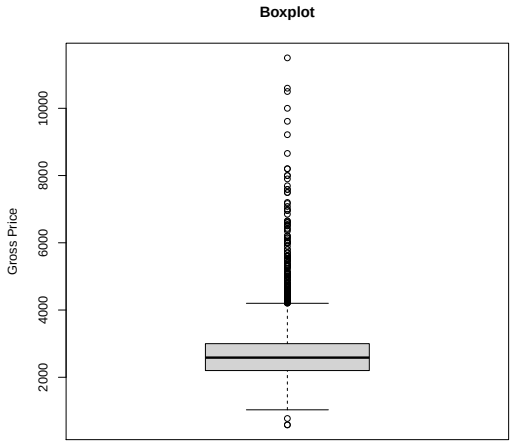

Figure 3: Gross Price Boxplot

The Boxplot, on the other hand, displays the *gross_price* distribution base on minimum value, first quartile, median, third quartile, and maximum value.
In this plot we visualize several outliers that indicate the non-Normal behavior of *gross_price* density function.

Now we proceed similarly for the variable *area*. We, then, reported and analysed the main statistics and graphs of the density, Q-Q plot and Boxplot.

Table 2: Area - Summary Statistics

	Area
Min	3.00
1st Qu.	35.00
Median	43.00
Mean	45.67
3rd Qu.	53.00
Max	220.00
Std. Dev.	17.60431
Skewness	2.224468
Kurtosis	14.11644

As with the dependent variable, the *area* variable does not follow the pattern of a Normally distributed variable. As the skewness index is greater than 0 and the kurtosis index is greater than 3, we have a

skewed right and keptokurtic distribution.

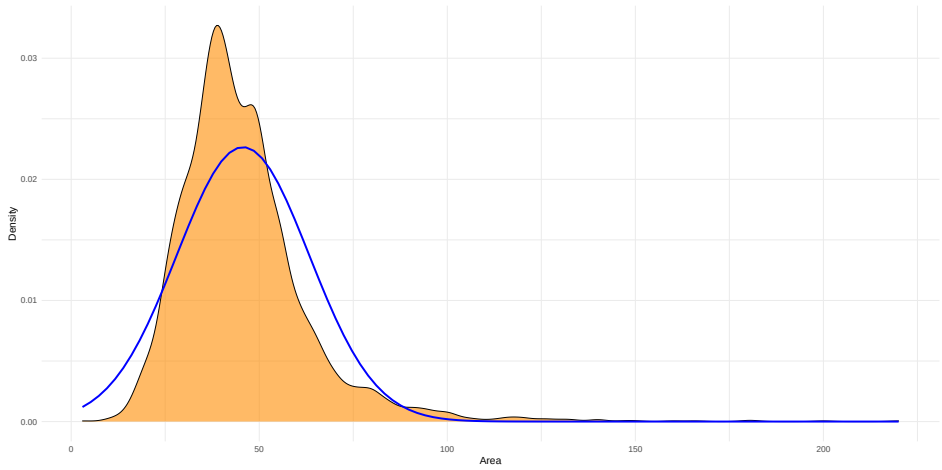

Figure 4: Area vs Normal Density Distribution

From Figure 4 it is immediately visible that the graph does not reflect the trend of the Normal distribution: as predicted in the table above, the distribution is right skewed. Moreover, performing the texts of Shapiro-Wilk and Jarque-Bera, both strongly reject the hypothesis of Normality.

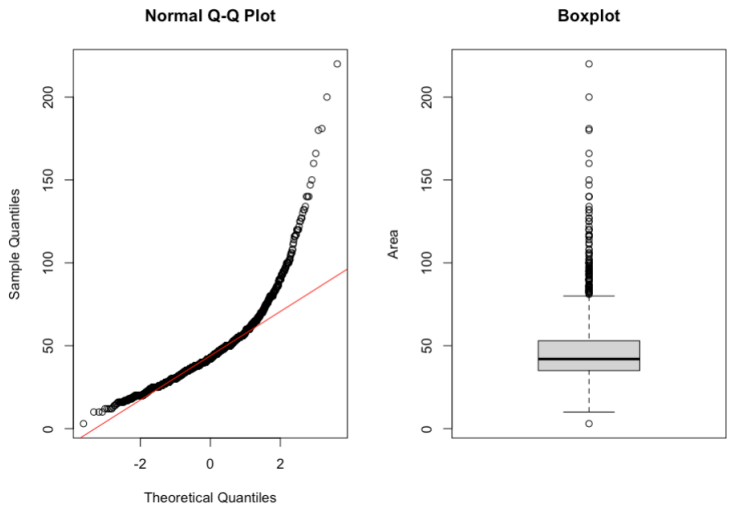

Figure 5: Area Q-Q plot and Boxplot

The Q-Q plot in Figure 5 shows us, in the right tail, a progressive detachment of the distribution of *area* from the Normal distribution.

As well as in the Q-Q plot, the Boxplot shows us a strong right skewed distribution of this variable, with a high number of outliers.

Finally, we analysed the last continuous variable in our dataset: *age*.

Table 3: Age - Summary Statistics

	Age
Min	1.00
1st Qu.	6.00
Median	18.00
Mean	23.55
3rd Qu.	30.44
Max	132.00
Std. Dev.	22.13497
Skewness	1.567788
Kurtosis	5.495752

Table 3 shows us an interesting result: although the skewness and kurtosis indices are once again greater than 0 and 3 respectively, the difference is not as marked as in the two previous variables.

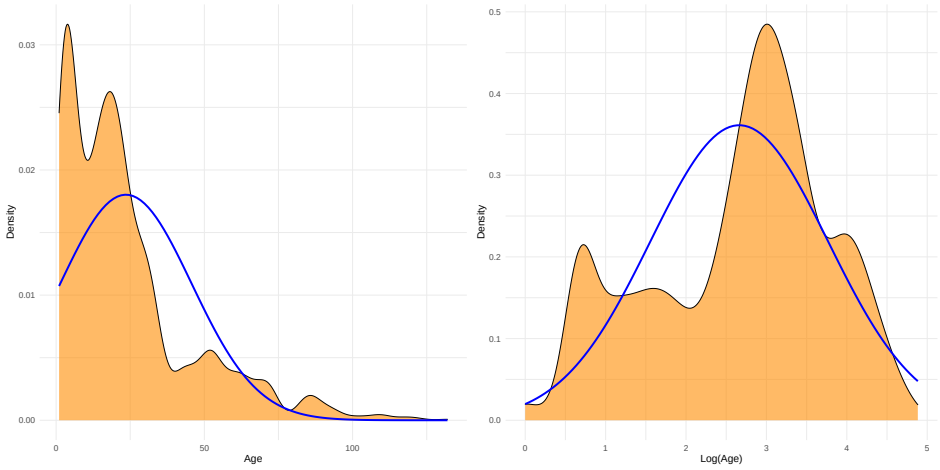

Figure 6: Age & Log(Age) vs Normal Density Distribution

The first graph in Figure 6 shows an atypical behaviour of the age variable. It is evident that the majority of the observations are between 0 and 25, this means that most of the rented flats are between 0 and 25 years old. We decided to place the graph of the density of *age* side by side with the logarithm of the density of the same variable, always comparing them with the Normal distribution. It can be seen that the density of *log(age)* fits better with the trend of the Normal distribution. Even though the Normality distribution hypothesis is rejected by the texts of Shapiro-Wilk and Jarque-Bera for both variables, in general we found that each continuous variable studied shows a closer behavior to Normal distribution once the logarithmic transformation is applied.

2.2.2 Discrete variables

In this section we analyzed the discrete variables in the model with frequency tables. Unlike statistics, these tables highlight the absolute and relative frequencies of each variable, thus allowing us to have a general overview of each variable. The discrete variables, specifically, are *room_num* and *floor*.

Table 4: Room_Num - Frequency Table

	Frequency	Percent	Cum. Percent
2	1874	57.6	57.6
1	762	23.4	81.0
3	507	15.6	96.6
4	94	2.9	99.5
5	14	0.4	99.9
7	1	0.0	100.0
6	1	0.0	100.0

The table 4 shows us that the most common number of rooms is 2, which alone accounts for more than half of the frequencies, at 57.6%. On the other hand, flats with six or seven rooms were counted the fewest number of times (1 and 1 respectively).

Table 5: Floor - Frequency Table

	Frequency	Percent	Cum. Percent
1	897	28	28
2	566	17	45
3	510	16	61
4	395	12	73
5	245	8	80
6	185	6	86
7	126	4	90
8	102	3	93
11	81	2	96
9	76	2	98
10	66	2	100
0	3	0	100
15	1	0	100

Here we have an interesting result. The first eight results in table 5 are made up neatly of floors 1 to 8, in descending order (the first floor being the most frequent, the eighth the least). We can also see that these floors account for 93% of the total frequencies. Strangely enough, the ground floor does not seem to be much considered.

2.2.3 Dummy variables

With regard to the dummy variables we also analysed the absolute and relative frequencies.
The table 6 shows that the districts of Mokotów, Wola and Śródmieście offer the highest availability for renting a flat. It can be seen that the Mokotów district alone accounts for almost 20% of all observations.

Table 6: Dummy Variables

District	Frequency	Percent
Bemowo	155	0.04764832
Białołeka	152	0.0467261
Bielany	158	0.04857055
Mokotów	635	0.1952044
Ochota	198	0.06086689
Praga-Południe	228	0.07008915
Praga-Północ	132	0.04057793
Rembertów	13	0.003996311
Targówek	106	0.03258531
Ursus	95	0.02920381
Ursynów	203	0.06240393
Wawer	40	0.01229634
Wesoła	13	0.003996311
Wilanów	123	0.03781125
Wola	492	0.151245
Włochy	75	0.02305564
Śródmieście	331	0.1017522
Żoliborz	104	0.03197049

On the other hand, the Rembertów and Wesoła districts are the ones where it is apparently most difficult to find a rental, with 13 flats each.

2.2.4 Correlation analysis

Lastly, we checked whether there were any correlations between the variables examined. Since our continuous variables are not Normally distributed, we could not use Pearson's method, so we proceeded to check for correlations using Spearman's method, which does not require the variables to be continuous and Normally distributed.

Below are the results obtained using the functions *cor()* in *RStudio*. The table below is the correlation matrix and shows the correlation coefficients for each pair of variables.

Table 7: Correlation Matrix

	Area	Room_num	Floor	Age	Gross_price
Area	1.00	0.79	-0.01	-0.08	0.73
Room_num	0.79	1.00	0.00	-0.07	0.64
Floor	-0.01	0.00	1.00	-0.01	0.04
Age	-0.08	-0.07	-0.01	1.00	-0.27
Gross_price	0.73	0.64	0.04	-0.27	1.00

In addition to the correlation coefficients, we also employed the function *cor.test()* which tests the correlation between paired samples, reporting both the correlation coefficient and the significance level of this correlation.

From the beginning we expected a high and positive correlation between the variables *area* and *room_num*. The correlation coefficient of these two is, in fact, the closest to 1, suggesting that the two variables measure something similar and it might be enough to include only one of them in the regression model.

However, the *cor.test()* result is insignificant as the p-value is $2.2e-16$.

On the other hand, we can consider as good the high correlation values between *area* and *gross_price*, *room_num* and *gross_price*, and the less high and negative correlation between *age* and *gross_price*, suggesting that these variables could be important regressors for our model.

In addition, we showed the scatterplot of the relationship between *gross_price* and *age*. From the graph it is immediately visible that the relationship is not linear, the cloud of points has the shape of a convex parabola, this leads us to think that the relationship between the two variables is more non-linear than linear.

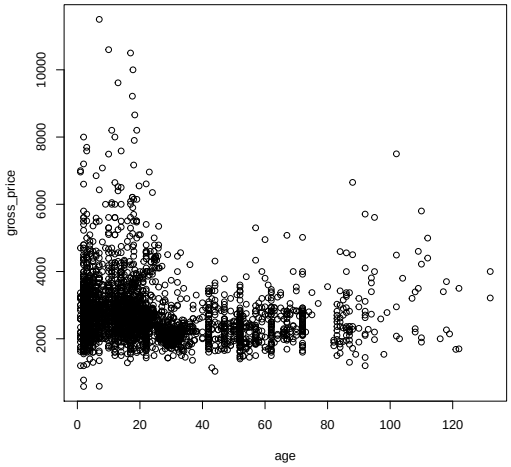

Figure 7: Scatterplot Age vs Gross Price

3 Model estimation

Since we have numerous (18) dummy variables indicating districts, we decided to group them into one macro group: we introduced the dummy variable *city_centre*, with the districts Mokotów, Ochota, Praga-Południe, Praga-Północ, Wola, Śródmieście and Żoliborz, which indicates all districts located approximately in the city centre of Warsaw. This dummy has the value 1 for districts falling within those listed above, 0 otherwise. We decided to create a single dummy for all districts for two reasons: firstly, for a purely interpretative reason, with such a large number of variables it would have been difficult for the reader to read the results provided by this work without getting confused; secondly, trying to build a model with all districts an error appeared, the Żoliborz district only produced *NAs* for coefficient, standard deviation, t value and p-value, this is most likely due to a multicollinearity problem.

Moving to the statistical analysis, we found it interesting to show mean, standard deviation and the other statistical indicators studied on *gross_price* conditional on the value of the variable *city_centre* (0 or 1). Below we showed the tables with the respective results.

Table 8: City_Centre = 0

	Obs	Mean	Median	Std. Dev.	Min	Max
gross_price	1133	7.831	7.824	0.2592208	6.651	9.259

Table 9: City_Centre = 1

	Obs	Mean	Median	Std. Dev.	Min	Max
gross_price	2120	7.900	7.871	0.2894492	6.365	9.350

As we can see from tables 8 and 9, the number of observations inherent in the districts located in the suburbs is about half that of the city centre districts. This data is interesting because the city centre, despite being smaller than the suburbs in terms of area, offers more rental availability and therefore has a higher density. Moreover, we note that the value of the average is not so different between the two values of the dummy variable. We expected average rental prices for flats in the city centre to be significantly higher than for flats in the suburbs, which was not surprisingly the case in the analysis.

After correcting our interpretation on the dummy variables, we proceeded to estimate the ideal model, starting with the simplest possible model, where no interaction between independent variables, no logarithmic transformation and no polynomial dependence is considered.
The coefficient estimators of the initial model are all significant at α=5%, so we performed the RESET test on *reg_1*, but found an error in the formulation of the functional form. The test has a p-value$<$ 5%, rejecting the null hypothesis of linearity of our functional form.

Table 10: Summary reg_1

Variables	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	681.2222	35.8628	18.995	$< 2e-16$***
Area	39.0141	0.8878	43.945	$< 2e-16$***
Room_num	54.9895	20.9101	2.630	0.00858**
Floor	8.9652	3.7185	2.411	0.01597*
Age	-4.5583	0.4422	-10.307	$< 2e-16$***
City_centre	384.8960	20.7732	18.528	$< 2e-16$***

This was to be expected given the non-Normality of the *gross_price*, *area* and *age* variables.
The function *summary()* on *RStudio* yields also two important indicators: the adjasted R^2, to assess the goodness of fit of the linear regression model; and the F-statistic, to verify whether the estimators of coefficients are jointly non-significant.
The adjasted R^2=0.6459, meaning that the 64.59% of the variability of *gross_price* is explained by the variability of the regressors.
F-statistic=1187 and its p-value$< 2.2e-16$, indicating that β estimators are jointly significant.
The second step, therefore, was to log-transform *gross_price*. Strangely, this transformation worsened the results of significance of the β estimators, in fact *floor* is significant only for α=10%, moreover the

RESET test produced the same results as the initial model.

Thus, we tried to consider the non-linear relationship between *gross_price* and *age*, inserting age^2. The coefficients are all significant for α=5%, but the RESET test still suggested changing the functional form.

For the fourth model we also took into account the high correlation between *area* and *room_num*, inserting the interaction between the two variables. The coefficients are all significant for α=5% and the RESET test results improve, but still reject the null hypothesis.

We tried to build other models, considering log-transformation of *area*, removing *floor* from the model, as it was very often non-significant even at α=10%, but for all these models the RESET test produced poor results.

The model that has all coefficients significant for α=5% and, at the same time, significantly improves the RESET test is our *reg_6*, whose RESET test p-value=3.784e-05, this result do not allow us to not reject the null hypothesis of the test, but we can consider this result positive when compared to other models. Moreover, we have an improvement in terms of adjasted R^2 and F-statistic.

The adjasted R^2=0.6619, meaning that the 66.19% of the variability of *gross_price* is explained by the variability of the regressors.

F-statistic=1062 and its p-value$< 2.2e-16$, indicating that β estimators are jointly significant. The table below shows the results of the summary of *reg_6*:

Table 11: Summary reg_6

Variables	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	7.1516863	0.0198378	360.507	$< 2e-16$***
Area	0.0150657	0.0004808	31.338	$< 2e-16$***
Room_num	0.1044970	0.0078955	13.235	$< 2e-16$***
Floor	0.0029231	0.0011099	2.634	0.00849**
log(Age)	-0.0462391	0.0026068	-17.738	$< 2e-16$***
City_centre	0.1224756	0.0061462	19.927	$< 2e-16$***
area:room_num	-0.0013767	0.0001249	-11.019	$< 2e-16$ ***

3.1 Diagnostic

3.1.1 RESET test

Diagnostic tests serve to check whether the CLRM assumptions are verified for the chosen model. The satisfaction or otherwise of the assumptions leads to different conclusions about the estimation of coefficients, their standard deviations and thus the reliability of the significance tests.

The RESET test is also part of the diagnostic tests. As we mentioned in the previous paragraph, the RESET test detects whether there has been a misspecification of the functional form during the construction of the model, suggesting, if appropriate, to improve its functional form.

The RESET test serves to verify the first of the four CLRM assumptions: that the model is linear in its parameters. Here is a table with the results of the RESET test for each model built:

Table 12: RESET test results

Model	RESET	df1	df2	p-value
reg_1	45.377	2	3245	$< 2.2e-16$
reg_2	51.744	2	3245	$< 2.2e-16$
reg_3	45.706	2	3244	$< 2.2e-16$
reg_4	11.923	2	3243	6.932e-06
reg_5	78.541	2	3245	$< 2.2e-16$
reg_6	10.214	2	3244	3.784e-05

3.1.2 Residuals Analysis

The second CLRM assumption to be verified is that the expected value of the error term is equal to zero: $E(\varepsilon_i) = 0$. This means that the distance between the fitted values and the observed values, which is called *residual*, on average, must be zero. To do this, we study the behaviour of the residuals, to see if they behave Normally.

CLRM assumption does not require residuals to be distributed as a Gaussian, but, from the OLS theory, we know that $\varepsilon \sim N\left(0, \sigma^2 I\right)$, and so by assessing whether the distribution of residuals is a N $(0, \sigma^2 I)$, we can tell for sure whether $E(\varepsilon_i) = 0$, and thus verify the assumption.

First, we proceed with a numerical analysis and then move on to a graphical analysis of the residuals.

Numerical analysis showed that the residuals have a slightly negative Skewness value, but very close to 0; while the kurtosis deviates more from 3, which represents the Normal value.

Table 13: Residual Analysis

	Residuals
Skewness	-0.2348391
Kurtosis	10.22449

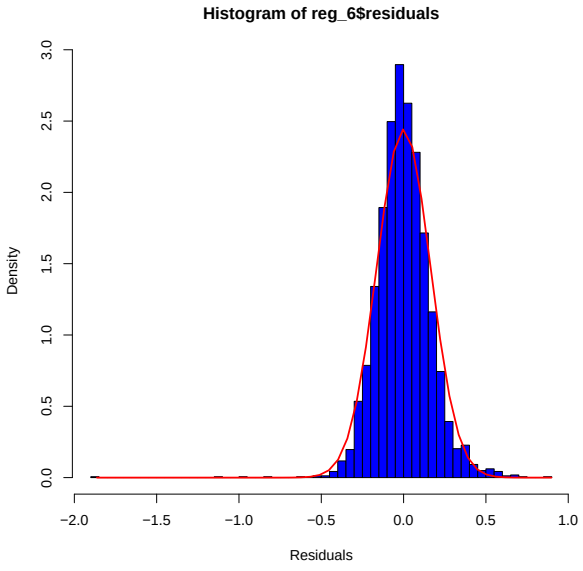

Figure 8: Residuals vs Normal Density Distribution

This subtle difference between the residuals and the Normal distribution can also be deduced by comparing the histogram of the residuals and the density function of the Normal. The two distributions are very similar, but not identical. One can see slight differences on the right-tail of the residuals and a left skewness on the maximum peak of the histogram. The deviation from the Normal distribution is more evident from the graphical representations in Figure 9. The Q-Q plot shows a particular deviation of both tails of the residuals from the Normal line, also indicating some observations as possible outliers.
The Boxplot also reveals some observations that deviate significantly from the others, but does not indicate which they are.

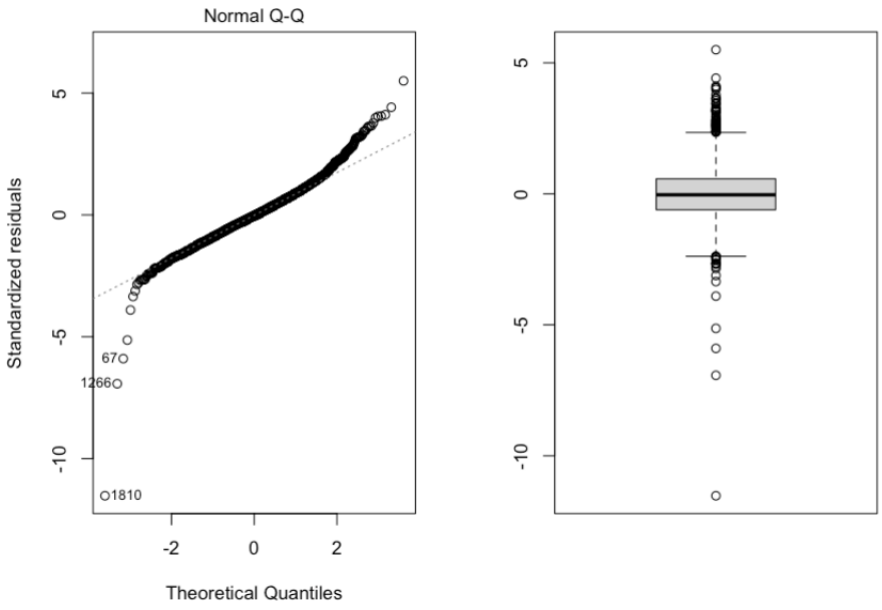

Figure 9: Residuals Q-Q plot and Boxplot

Finally we performed the usual Shapiro-Wilk test and Jarque-Bera test to test the Normality of our residuals. Despite the great graphical similarity in the histogram, the tests agreed with the differences accentuated by the other two plots. The p-value for both tests is less than 0.05, the tests reject the null hypothesis of Normality of the residuals.
The fact that both tests reject the Normality hypothesis for residuals should not worry us. In fact, having a large number of observations, Normality is guaranteed by the Law of Large Numbers.

3.1.3 Homoscedasticity

The third step is to check the homoschedasticity of residuals. Recall that homoschedasticity requires that the variance of the error term is constant for all observations, $\text{Var}(\varepsilon_i) = \sigma^2$ for $i = 1, 2, \ldots, N$. This assumption can also be verified either graphically or by performing specific tests.
The Figure 10 shows that the points are not randomly distributed, the pattern is very clear and there is greater volatility on the right-hand side of the graph. This leads us to say that the residuals are heteroschedastic.

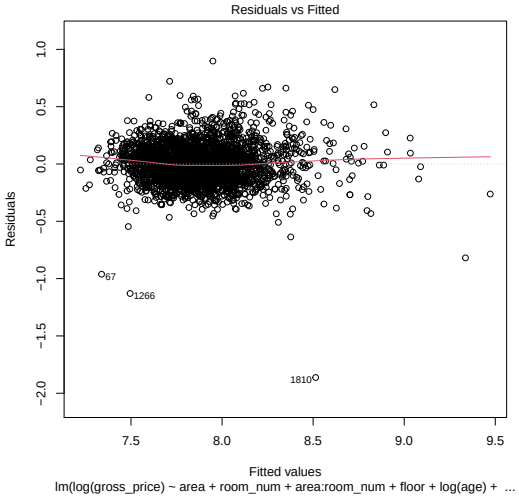

Figure 10: Residuals vs Fitted values

To be even more certain of such heteroschedasticity, we used the Breusch-Pagan test on our residuals, whose null hypothesis is the presence of homoschedasticity. Unfortunately, the test also confirms the heteroschedasticity of the residuals, reporting a p-value $< 2.2e-16$ and rejecting the null hypothesis. Heteroschedasticity makes biased the Covariance matrix of the β estimators, leading to erroneous conclusions. The best way to solve the problem of heteroschedasticity is to use the Robust Variance-Covariance matrix, which allows us to estimate the regression taking into account the heteroschedasticity of the variance.

Using the *rlm()* function on *RStudio* we obtain the standard deviations corrected for heteroskedasticity. Below we report the results of the summary:

Table 14: Robust Variance-Covariance matrix summary

Variable	Value	Std. Error	t value
(Intercept)	7.1841	0.0183	393.4429
Area	0.0148	0.0004	33.3367
Room_num	0.0894	0.0073	12.2953
Floor	0.0024	0.0010	2.3864
log(Age)	-0.0468	0.0024	-19.4948
City_centre	0.1139	0.0057	20.1402
Area:Room_num	-0.0012	0.0001	-10.4751

3.1.4 Multicollinearity

The Variance Inflation Factor (VIF) quantifies the severity of multicollinearity in an OLS regression analysis. It provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity.

The threshold indicating multicollinearity between the regressors is equal to 10. Thus, if for a variable in the estimated model the test produces a VIF>10, that variable should be eliminated from the

model because it does not provide any additional information that is not already provided by the other independent variables.

Table 15: VIF test

	VIF
Area	8.712994
Room_num	4.225159
Floor	1.024614
log(Age)	1.008503
City_centre	1.043374
Area:Room_num	13.452999

Fortunately, the VIF for all variables in our ideal model is less than 10, except for *Area:Room_num*, but this is normal since the interaction variable, by definition, depends on the variables from which it is formed.

3.1.5 Diagnostic Conclusion

To conclude, we tried to perform all the tests carried out during this paragraph again by replacing *reg_6* with *reg_robust*. When comparing the summary values of the two models, a difference is evident, even if subtle, but when performing the tests for linearity, homoschedasticity and normality of the residuals, the results are almost identical.

3.2 Problems with Data

In this section we explored two of the major problems that can be encountered when analysing a dataset.

3.2.1 Non-typical and wrong observations

Non-typical observations refer to those observations that have non-typical features, in comparison with other observations. On the other hand, wrong observations are more complicated: they usually appear as a result of a mistake, but sometimes they are real. The impact on regression is different: non-typical observations impacts positively on estimation precision and model fit, while wrong ones impacts them negatively. It all depends on how close the observations lie to regression line.

There are three measures we can use to detect not typical observations: leverage, standardized residuals and Cook's distance. The non-typical leverage occurs when its value is greater than 2K/N, where K is the number of explanations considering the constant and N is the number of observations. The standardized residuals, instead, show non-typical observations when its modulus is greater than two. If both values of leverage and standardized residuals are high, Cook's distance must also be considered, and its value should be greater than 4/N.

Below we proceeded to show the plots analysinig the results.

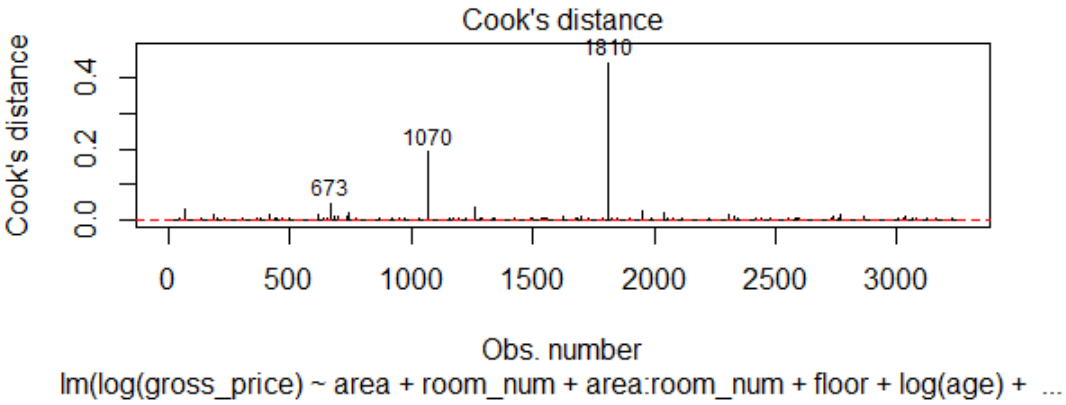

Figure 11: Cook's distance

The red horizontal line in Figure 11 shows the threshold of Cook's distance, so all the observations above that line are suspicious. Among these, we can see three observations that stand out the most: 673, 1070, 1810.

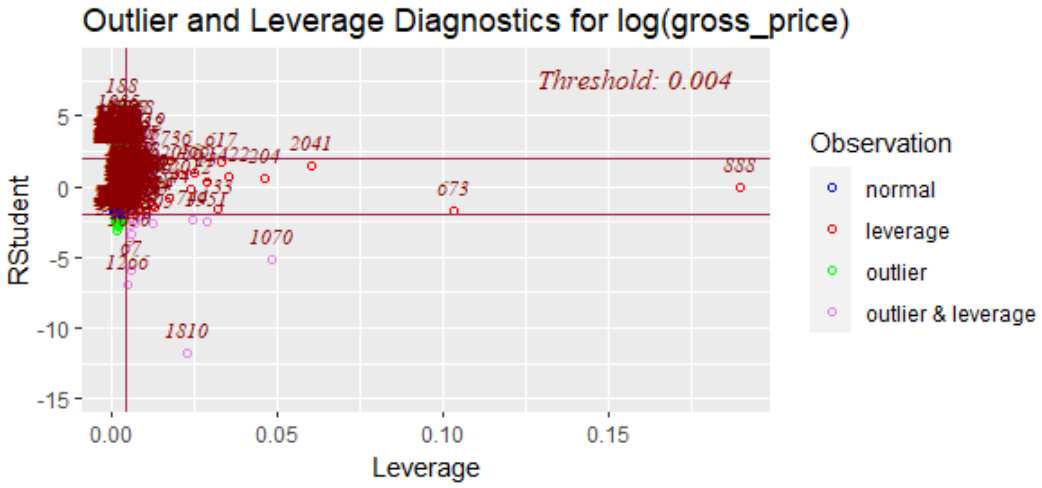

Figure 12: Outlier and Leverage Diagnostics

In Figure 12 the observations that need to be removed are the ones at the top-right and bottom-right corner. These observations indicate high leverage and high residuals in modulus. The most obvious non-typical observations are the same: Contrary to the previous graph, the observation 673 is not in one of the critical areas: it has a rather high leverage but the residuals are within the [-2,2] range. The same cannot be said for the other two. The observations 1070 and 1810 are within the critical values, so they must be deleted.

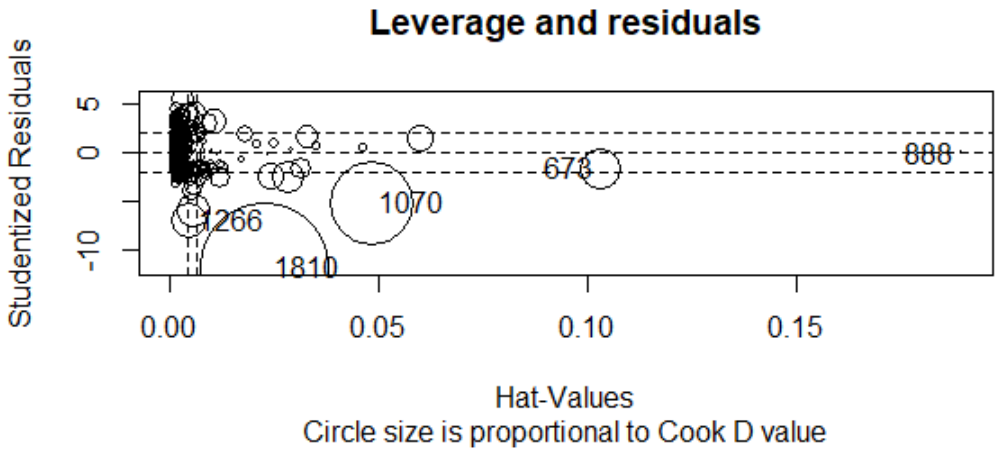

Figure 13: Leverage and Residuals

Finally, the Figure 13 reveals the outliers and leverage diagnostics together with Cook's distance: the larger the "bubble", the bigger the Cook's distance. As can be seen, the observations 1070 and 1810 are, once again, the most problematic ones.

Looking at these 3 plots, we can see that the main non-typical observations are, as we observed, 673, 1070 e 1810. These observations fall in the regions of high leverage and residuals and, therefore, must be removed from the model.

3.2.2 Collinearity

The collinearity is a correlation between predictor variables, it indicates that one independent variable can be linearly predicted from the others.

Here we showed our study of collinearity in the regression model.

Table 16: Collinearity

Variables	Tolerance	VIF
area	0.11477110	8.712994
room_num	0.23667748	4.225159
floor	0.97597723	1.024614
log(age)	0.99156843	1.008503
data$city_centre	0.95842905	1.043374
area:room_num	0.07433287	13.452999

From table 16 we can note that only the variables *area* and *area:room_num* have a quite high VIF value. However, the last variable mentioned has a VIF value greater than 10. This is accepted because that variable is, by construction, an interaction between *area* and *room_num*.

4 Interpretation and Conclusion

To sum up, we wanted to find the determinants that influenced the price of renting flats in Warsaw. We took into account the variables that we thought would have the most significant impact, while neglecting those with the least impact.

We estimated a model trying to respect all the assumptions that the OLS theory imposed on us. Not all of them have been met.

In the end, the model we propose has these variables and this functional form:

$$log(price) = 7.18 + 0.014area + 0.09room + 0.002floor - 0.05log(age) + 0.11city\_centre - 0.001area : room \quad (1)$$

This model is the one that in our opinion gives an idea about the effect of all variables on *gross_price* and at the same time has the functional form that takes into account all the characteristics of our data: from the non-Normality of *gross_price* and *age*, to the high correlation between *area* and *room_num*.

The model failed the RESET test and the Normality tests, but had no multicollinearity problems.

The interpretation of our coefficients is as follows.

Area: the effect of area is conditioned by the *room_num* in the flat. The higher the *room_num*, the smaller the effect of *area* on the dependent variable. If, for example, we had *room_num*=1, the effect of *area* on the rental price would be 0.2%, that is greater than 2-room flat.

Room_num: conversely, the effect of the number of rooms is affected by *area*. The larger the *area*, the smaller the effect of *room_num* on the rental price. For example, if *area*=10, the effect of *room_num* on *gross_price* would be 7.7%, that is greater than $20m^2$-flat.

Floor: if the flat goes up one floor, the rental price goes up by 0.24%.

log(age): if the age of the flat increases by 1%, the rental price decreases by 0.0468%.

City_centre: if the flat is located within the city centre, the rental price increases by 12.06%. The semi-elasticity effect calculated normally is slightly different from our interpretation percentage, since we have calculated it by using $100(exp(\beta) - 1)$.

4.1 Conclusions

Our initial assumptions were only partially fulfilled. Firstly, we expected a larger influence from all variables, the coefficients obtained are close to zero, indicating that large changes in the regressors lead to small changes in the rental price.

We expected a greater influence from the district variables and in fact *city_centre* is the variable with the highest coefficient, which means that the geographical location of the flat matters a lot for the determination of the rental price.

We expected a greater effect of *area* as well, but the coefficient is close to zero, which is a surprising result. The low significance of *floor* was respected, as well as the negative influence of *age*.

4.2 Forecasting

According to our model, if a person were looking for a $40m^2$, two-bedroom flat on the third floor of a building, 30 years old, which reflects the average age of flats in Warsaw, and located in the city centre, he would spend 8033.83 PLN. The result obtained is not in line with reality, the price for such a flat is very expensive. We can conclude that our model, while obtaining a good adjasted R^2 score, can be improved using alternative methods beyond OLS, or including other variables, or deleting non-typical observations that are influential.

5 References

- PwC, REAS & CMS. (n.d.). "Institutional Rental Market in Poland".
- Yuan, L. (2019). "A Regression Model of Single House Price in LA Constructing a Predicted Model for House Renting". Los Angeles, CA.
- Aminah, M.Y. & Syuhaida, Y. (2012). "Multiple Regression in Analysing House Price Variations". 383101, 9 pages. DOI: 10.5171/2012.383101. Johor, Malaysia.
- Abebaw, H.F. (2021). "Determinants of residential house rental price in Debre Berhan Town, North Shewa Zone, Amhara Region, Ethiopia". *Cogent Economics & Finance*, 9:1, 1904650, DOI: 10.1080/23322039.2021.1904650
- Gustafsson, A. & Wogenius, S. (2014). "Modeling Apartment Prices with the Multiple Linear Regression Model". Stockholm, Sweden.
- Cohen, V. & Karpaviciute, L. (2016). "The Analysis of the Determant in House Prices". *Independent Journal of Management & Production*, vol. 8, núm. 1, pp. 49-63. Avaré, Brasil.
- Darfo-Oduro, R. (2020). "Determinants of Residential House Rental Prices in Accra Metropolis". *SSRN Electronic Journal*.