UNIVERSITY OF WARSAW

Faculty of Economic Sciences

University of Warsaw

Faculty of Economic Sciences

Determinants of apartment prices in Warsaw

Xinyue Fang

No. 446197

Course: Econometrics

Prepared under the supervision of mgr Kateryna Zabarina

University of Warsaw, January 2023

Table of contents:

First Part

1.1 Introduction

Factors influencing apartment prices have long been a hot topic in economic research. The Polish real estate market has gone through several different periods, namely the period of relative stability in 2002-2005, the real estate boom in 2006-2008, and the slow return of the market to equilibrium from 2009 to 2015. (Leszczyński R, Olszewski K.,2017) What exactly is affecting apartment prices? Scholars have studied the mechanism of the impact of city housing prices at both the macro and micro levels. At the macro level, fluctuations in housing prices can be attributed to a large extent to changes in national income, real interest rates on housing loans and unemployment rates. In the case of the Warsaw housing market, demand for housing increases when incomes rise or interest rates fall. And for every 1% rise in unemployment, vacations fall by 1%.(Posedel P, Vizek M., 2009) However, in the short term, the main factors affecting housing prices within the city are micro-determinants. Distance to transportation infrastructure, distance to schools, age of the building, number of bedrooms or bathrooms, floor space, garage or kitchen size, etc. all have an impact on apartment prices.(Cui N, Gu H, Shen T, 2018)

In order to study the factors affecting apartment prices in Warsaw, this paper will use the characteristic price model to analyze them at the micro level.

In this paper, I will use data related to Warsaw apartment prices in November 2023 to analyze cross-sectional data on several factors such as apartment type, size, floor type, distance from the center, and so on.

1.2 Literature review

Characteristic price model (Hedonic Price Model, abbreviated as Hedonico model) is the earliest Kester's theory of consumption and Rosen's theoretical model evolved and developed, and then gradually applied to the price consistency analysis of goods.(Chau K W, Chin T L., 2003) After continuous development, there are more and more researches using Hedonico model to explore the implicit price of goods. The first to apply the characteristic price model to the study of factors affecting apartment prices was Ridker, who paid special attention to the impact of air pollution on apartment prices.(Ridker R G, Henning J A., 1967) Lancaster (Lancaster K J.,1966) and Rosen(Rosen S.,1974) also applied the characteristic price model to the research related to the real estate industry. In recent years, more and more empirical studies on real estate market use the characteristic price model to analyze, especially on the factors affecting the apartment price of a specific city in a specific country.In 2005, Wen Huizheng, Jia Shenghua and Ge Xiaoyu conducted an empirical study to analyze the characteristic price of urban housing in Hangzhou City, China.(Wen H Z, Sheng-hua J, Xiao-yu G.,2005) In 2012, Bhattacharjee A, Castro E, Marques J. conducted an empirical study on urban housing market in Aveiro, Portugal, analyzing the spatial

interactions in the feature pricing model. (Bhattacharjee A, Castro E, Marques J.,2012)

And there are many studies dealing with the determinants of apartment prices in Poland.In 2009, Posedel P and Vizek M. analyzed the determinants of apartment prices during the transition period in the EU-15 countries, exploring the impact of national income, housing loans and interest rates on the differences in real apartment prices, mainly at the macro level.(Posedel P, Vizek M.,2009) There are also empirical studies on the factors influencing the primary and secondary housing markets in Poland in both 2017 and 2019 studies, which were also analyzed mainly at the macro level in terms of unemployment rate, credit costs, etc. Some scholars have since conducted dynamic spatial and characterization of Warsaw apartment prices.(Olszewski K, Waszczuk J, Widłak M.,2017)

And what are the specific influencing factors of Warsaw apartment prices at a more micro level according to the November 2023 Warsaw apartment price data, I will empirically study and analyze in this paper.

1.3 Hypothesis

This paper focuses on the factors that influence the price of apartments in Warsaw. In order to carry out the research, I have formulated the following hypotheses based on the characteristics of apartments and referring to the literature.

We expect the type of apartment to affect its price. We categorize condominiums into three types, namely tenements, block of flats and apartment buildings. I'm assuming the tenement is the cheapest.

We expect a non-linear relationship between apartment size and price. We estimate that prices will rise rapidly as the size of the apartment increases.

Based on references and my life observations, I believe that there is also a positive relationship between the floor of an apartment and its price. For the purpose of this study, I categorized the floor types into low, medium and high. A floor that is shorter than one-third of the total floors in the building is a low floor, a floor that is higher than two-thirds of the total floors in the building is a high floor, and the rest are medium floors. I'm assuming the lower floors are the cheapest.

The fourth hypothesis is to check if apartments are cheaper the farther they are from the city center.

The fifth hypothesis is to check whether there is a positive relationship between the price of the apartment and the number of surrounding points of interest. Surrounding attractions include kindergartens, schools, universities, post offices, clinics, pharmacies, and restaurants, etc.

In addition, we also consider that there is a relationship between ownership relationships and apartment prices. So the sixth hypothesis is that there is a difference between the prices of apartments with condominium and cooperative ownership. I'm assuming the cooperative's appartment is cheaper.

Furthermore, we will also check whether apartments equipped with an elevator (test VII) and with a balcony (test VIII) will be more expensive.

Second Part

2.1 Short description of data

Data source: “Kaggle”
https://www.kaggle.com/datasets/krzysztofjamroz/apartment-prices-in-poland

The dataset contains apartment sell offers from the 15 largest cities in Poland (Warsaw, Lodz, Krakow, Wroclaw, Poznan, Gdansk, Szczecin, Bydgoszcz, Lublin, Katowice, Bialystok, Czestochowa). The data comes from local websites with apartments for sale. To fully capture the neighborhood of each apartment better, each offer was extended by data from the Open Street Map with distances to points of interest (POI). The data is collected monthly and covers timespan between August 2023 and January 2024.

I will use Warsaw apartment sell offers in November 2023 to analyze the factors affecting Warsaw apartment prices.

There are 2,692 offers in the data I'll be using.

My explained or dependent variable is apartment price.

In my model I will use the following explanatory variables: type dummy variables, size of the apartment, floor type dummy variables, distance to center, points of interest dummy variables, ownership dummy variables and balcony dummy variables, elevator dummy variables.

Type: a dummy variable comparing dummies covering three types of apartments, tenement, block of flats and apartment building. In data is named **type**.
Size: a variable that shows size of an apartment in square meters. In data is named **squareMeters**.
Floor type: is named **FloorType** in data. I compared the floor in an offer to the total number of floors, and it turns out that shorter than a third of the total floor is called low, higher than two thirds is called high, and the rest is called middle.
Center distance: a variable that shows distance from the city centre in km. In data is named **centreDistance**.
Points of interest: a dummy variable showing the number of nearby points of interest that are low, medium, or high. The number is collected by the points of interest in 500m range from the apartment (schools, clinics, post offices, kindergartens, restaurants, colleges, pharmacies). I define less than 30 points as small, more than 30as large. In the data is named **poiCount**.
Ownership: a dummy variable showing whether the apartment is a condominium or a cooperative. In the data is also named **ownership**.
Balcony: a dummy variable showing whether the apartment has balcony or not. In the data is named **hasBalcony**.
Elevator: a dummy variable showing whether the apartment has an elevator or not. In the data is named **hasElevator.**

2.2 Data Description and Statistical Analysis of Data

My analysis consists of 2692 observations.

2.2.1 Continuous variables

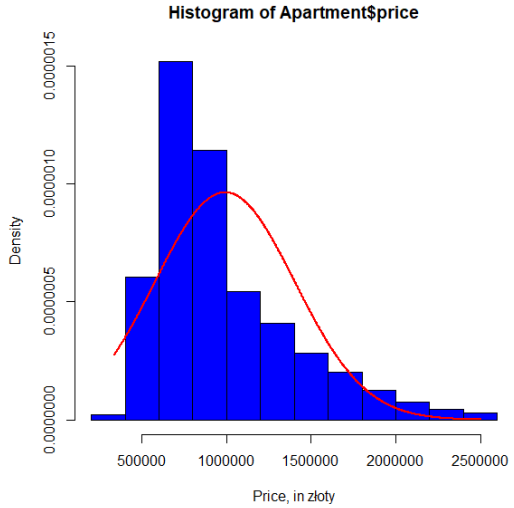
Price

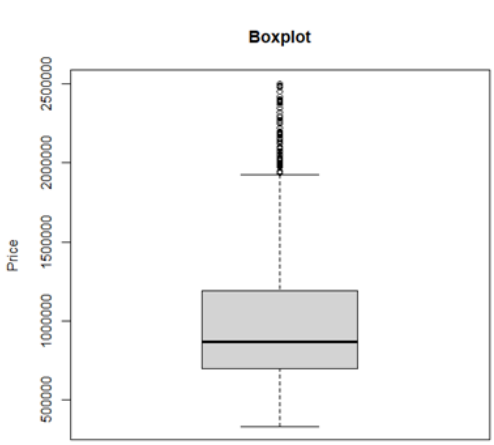
The continuous price variable has a distribution that is not normal, as can be seen from the density histogram, Q-Q plot and Boxplot. This problem can be solved later if we apply logarithm to our dependent variable, making it more normal. As we can see from those, there is a positive asymmetry that implies a right skewed distribution. That is a type of distribution in which most values are clustered around the left tail of the distribution while the right tail of the distribution is longer. Another confirmation of the right skewness of our data is the fact that the mean is larger than the median. This is due to the strong concentration of apartment prices in the range between 500000 and 1000000 PLN.

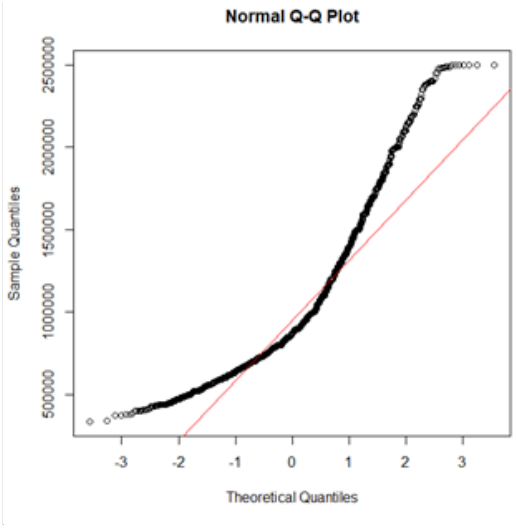
In particular, the non-normality is due to some outliers have a value less than 400000 and many of them are greater than 1700000, which give the right tail in the Histogram. The Q-Q plot graph is important for observing how the tails of our dependent variable (Price) diverges from the normal distribution. From the Boxplot we can also observe that there are many outliers have value greater than 1700000.

price

Min.	1stQu.	Median	Mean	3rdQu.	Max.
335000	699000	870000	990572	1191833	2500000





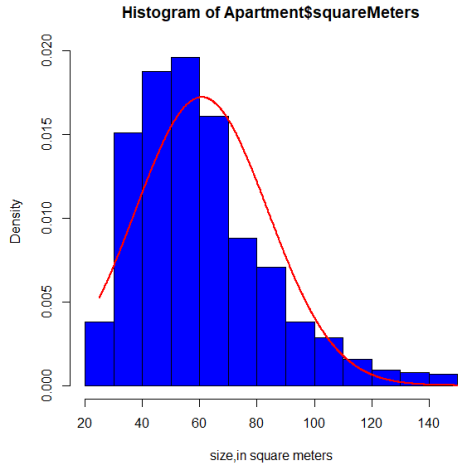


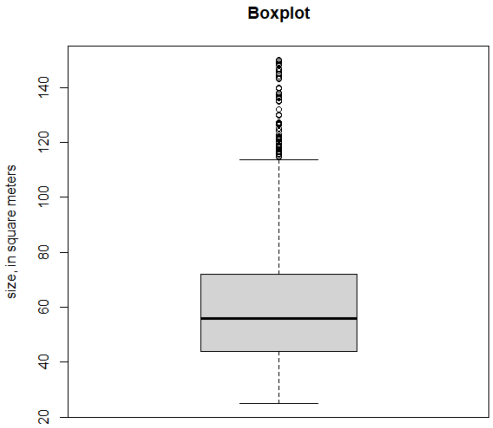
In addition to the graphs, I performed the Jacque-Bera test and the Shapiro-Wilk test to confirm the non-normality of the density distribution. These two tests strengthened my argument because the p-values were both very small and therefore I rejected the null hypothesis that the density distribution is normal.

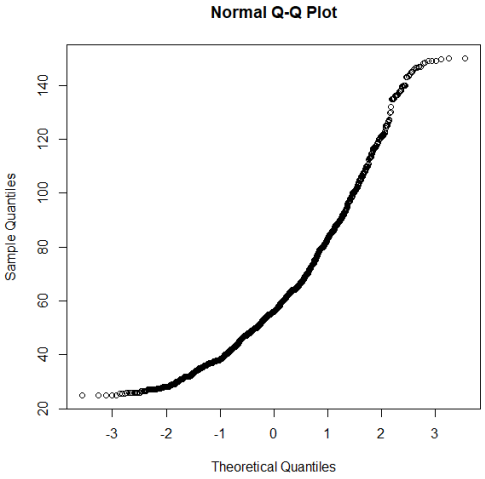
```
  Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 979.1482
  P VALUE:
    Asymptotic p Value: < 0.00000000000000022


        Shapiro-Wilk normality test

data:  Apartment$price
W = 0.888, p-value <0.0000000000000002
```

Size

We do the same to check the size. The continuous size variable has a distribution that is not normal, as can be seen from the density histogram, Q-Q plot and Boxplot. We also use the Jacque-Bera test and the Shapiro-Wilk test to confirm the non-normality of the density distribution. These two tests strengthened my argument because the p-values were both very small and therefore I rejected the null hypothesis that the density distribution is normal.







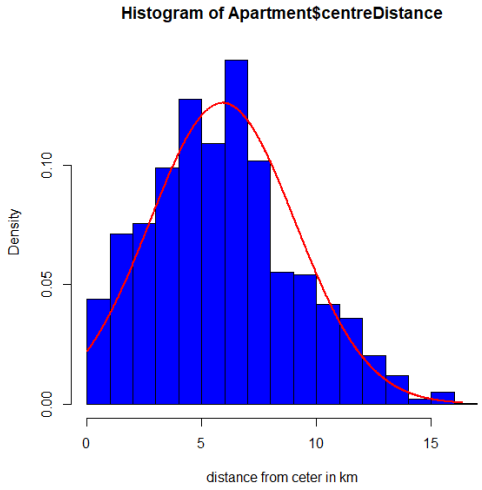

```
Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 784.6573
  P VALUE:
    Asymptotic p Value: < 0.00000000000000022
```

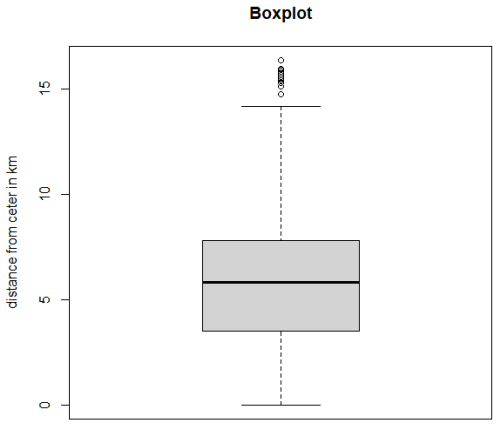
```
        Shapiro-Wilk normality test

data:  Apartment$squareMeters
W = 0.926, p-value <0.0000000000000002
```

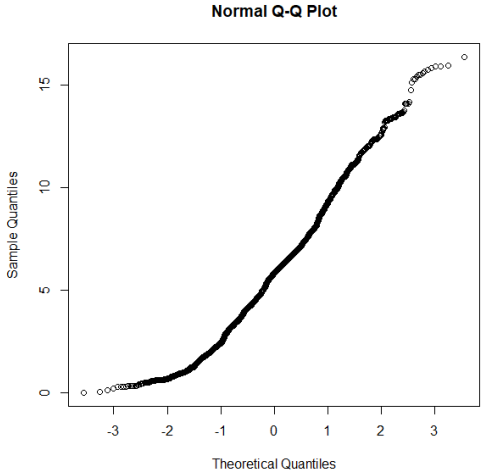
Distance from ceter

We do the same to check the distance from ceter. From the density histogram, Q-Q plot and Boxplot, it seems the continuous price variable has a distribution that is very closed to normal. But still we need to use the Jacque-Bera test and the Shapiro-Wilk test to check the normality of the density distribution. These two tests

strengthened my argument because the p-values were both very small and therefore I rejected the null hypothesis that the density distribution is normal.








```
Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 88.9936
  P VALUE:
    Asymptotic p Value: < 0.00000000000000022
```

```
        Shapiro-Wilk normality test

data:  Apartment$centreDistance
W = 0.979, p-value <0.0000000000000002
```

Correlation

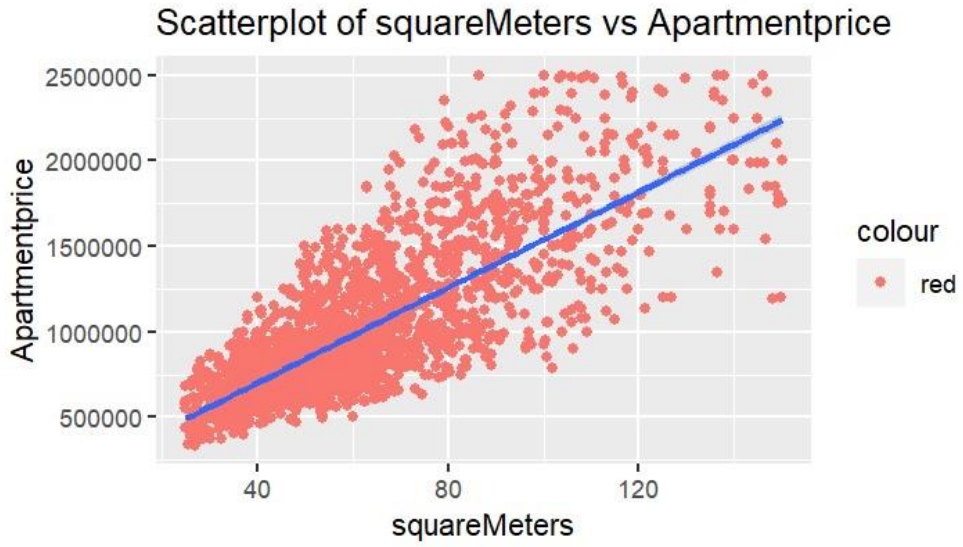
Non-normality of data means that we should not be using Pearson's method of measuring correlation but Spearman's method which does not assume normality.

1. Price VS Size

```
        Spearman's rank correlation rho

data:  Apartment$price and Apartment$squareMeters
S = 723550655, p-value <0.0000000000000002
alternative hypothesis: true rho is not equal to 0
sample estimates:
    rho
0.77747
```

```
        Kendall's rank correlation tau

data:  as.numeric(Apartment$price) and as.numeric(Apartment$squareMeters)
z = 45.3, p-value <0.0000000000000002
alternative hypothesis: true tau is not equal to 0
sample estimates:
    tau
0.58411
```

Correlation between price and size exists and equals 0.77747 according to Spearman's method, and 0.58411 according to Kendall's method.



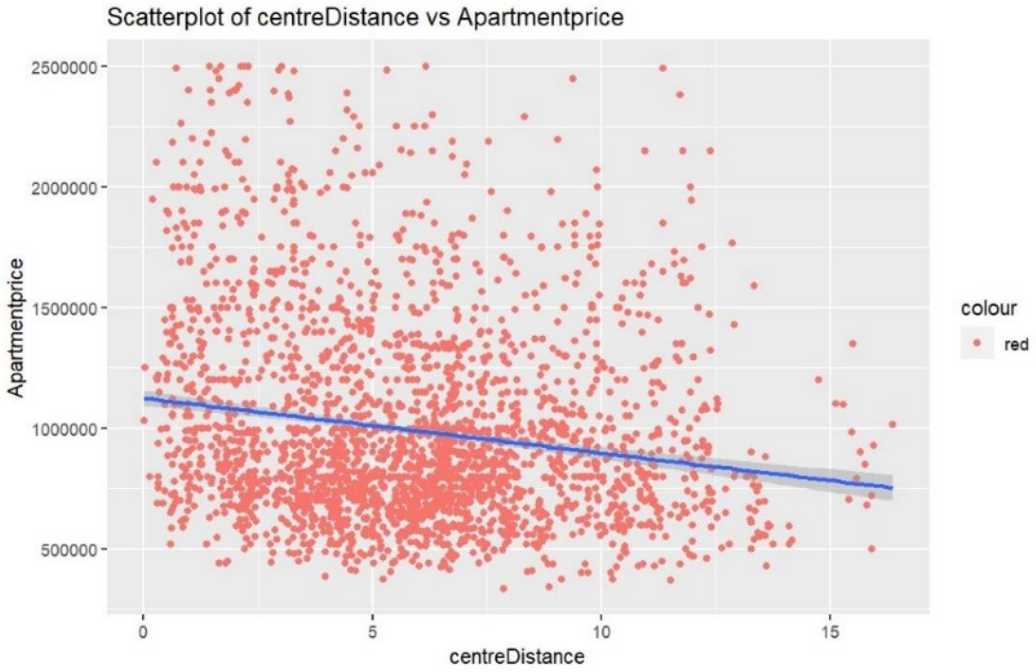
It can also be seen from this graph that there is a positive correlation between apartment prices and size.

2. Price VS Distance from center

```
        Spearman's rank correlation rho

data:  Apartment$price and Apartment$centreDistance
S = 3772538098, p-value <0.0000000000000002
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
-0.16027
```

```
        Kendall's rank correlation tau

data:  as.numeric(Apartment$price) and as.numeric(Apartment$centreDistance)
z = -8.25, p-value <0.0000000000000002
alternative hypothesis: true tau is not equal to 0
sample estimates:
     tau
-0.10628
```

Correlation between price and distance from center exists and equals -0.16027 according to Spearman's method, and -0.10628 according to Kendall's method. The correlation could also be seen from this graph.



2.2.2 Dummy variables

Now we are going to do an analysis on our dummy independent variables and provide a summary statistics of these variables.

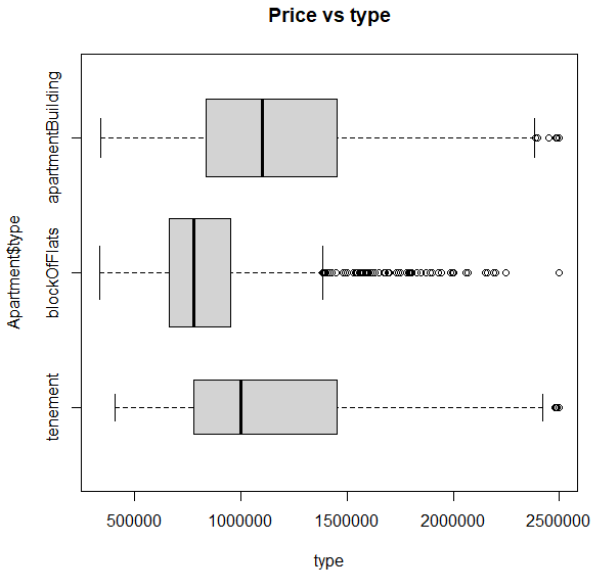
Type

In the data, 1 stands for tenement with 383 offers, 2 stands for block of flats with 1522 offers, 3 stands for apartment building with 787 offers.

To do this, we used through R the function by aggregate and table.

	type	xMin	x1stQu	xMedian	xMean	x3rdQu.	xMax.
1	tenement	405000	780000	998000	1154426.3	1455000	2500000
2	blockOfFlats	335000	660000	780000	852786.9	950000	2500000
3	Apartment Building	340800	833390	1100000	1177298.6	1455000	2500000

As we can see from the table and the Boxplot, all three types have the same max price. Block of Flats has the lowest min price. The average price of apartment building is the highest.



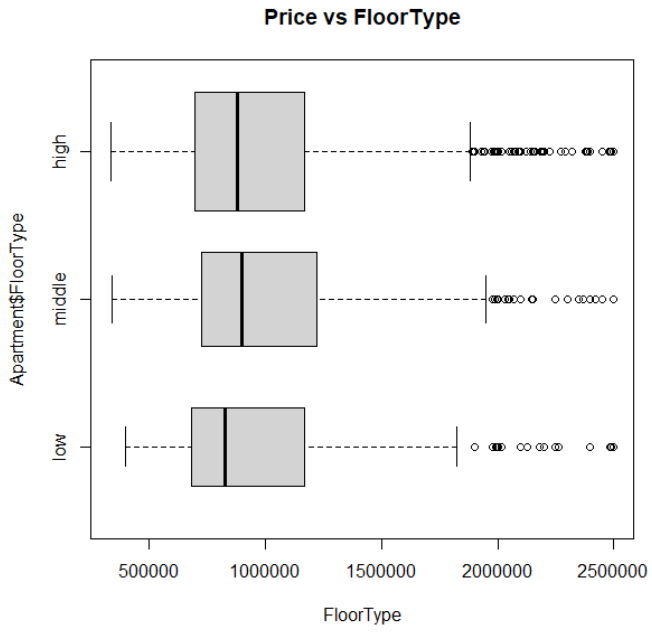
Floor type

I compared the floor in an offer to the total number of floors, and it turns out that shorter than a third of the total floor is called low, higher than two thirds is called high, and the rest is called middle.

According to the table summary, there are 570 offers have low floor type, 824 are middle type, 1298 are high type.

	FloorType	xMin	x1stQu	xMedian	xMean	x3rdQu	xMax.
1	low	399000	683500	828000	960780	1164250	2500000
2	middle	340800	723750	895888	1010439	1220000	2500000
3	high	335000	695000	881000	991043	1169750	2500000

Out of my expectation, the high floor type has the lowest min price. But we can see from the table and Boxplot, average price of the middle floor is the highest and the low floor is the lowest, which is expected.



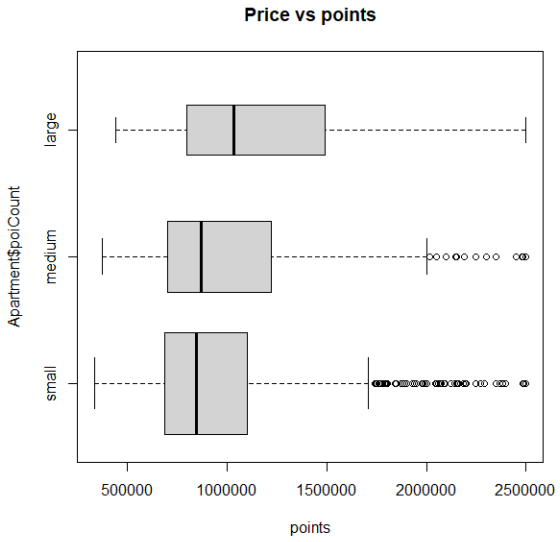
Points of interest

In the data, I divided the number of nearby points of interest into low or high these two types. I define less than 30 points as small, more than 30as large.

According to the table summary, there are 2089 offers have small number of nearby points of interest, 603 are large.

As I expected, the more points of interest near to the apartment, the higher price will be. We can see from the table and Boxplot that no matter min price or average, max price, larger number of points is higher.

	poiCount	x.Min.	x.1stQu.	x.Median	x.Mean	x.3rdQu.	x.Max.
1	small	335000	689000	846000	950225	1100000	2499999
2	large	375000	775000	980000	1130351	1426500	2500000

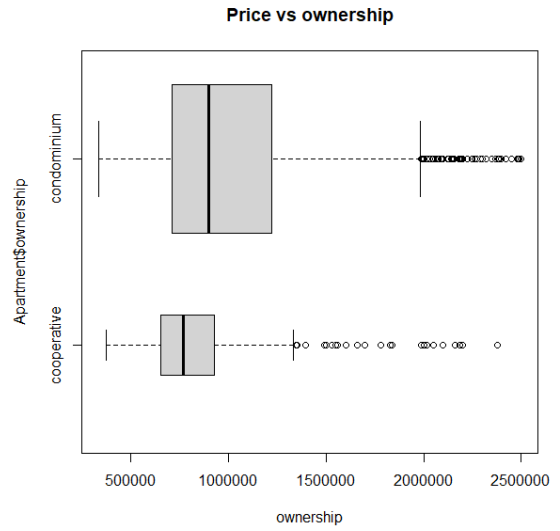


Ownership

In the data, o stands for cooperative with 381 offers, 1 stands for condominium with 2311 offers.

From the table and Boxplot we can tell that other than min price, the median, mean and max price of condominium are higher than cooperative.

	ownership	x.Min	x.1stQu.	x.Median	x.Mean	x.3rdQu.	x.Max.
1	cooperative	375000	650000	769000	851986	927000	2380000
2	condominium	335000	710000	897000	1013420	1220000	2500000



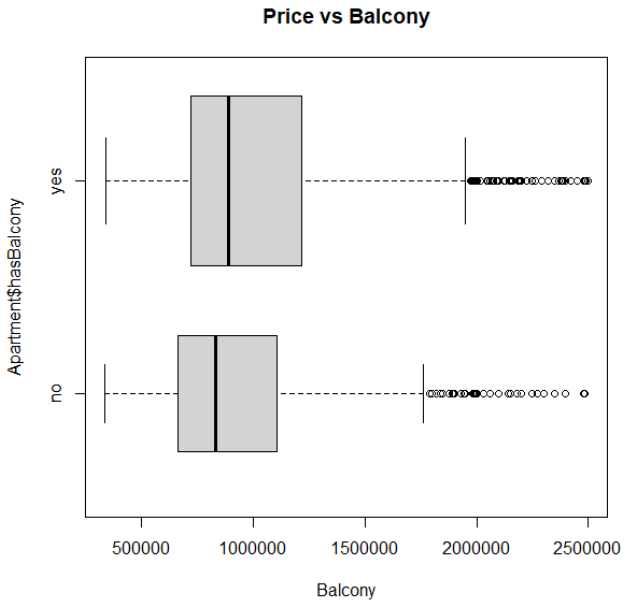
Balcony

In the data, o stands for the offer of apartment without balcony, 1 stands for with balcony.

There are 861 offers of apartment without balcony, 1831 with balcony.

As I expected, the apartment with balcony is more liked by people. The min, median, mean, max of apartment price with balcony is always higher.

	balcony	x.Min.	x.1stQu.	x.Median	x.Mean	x.3rdQu.	x.Max.
1	no	335000	660000	829000	942221	1104950	2485350
2	yes	340800	720000	890000	1013309	1217000	2500000



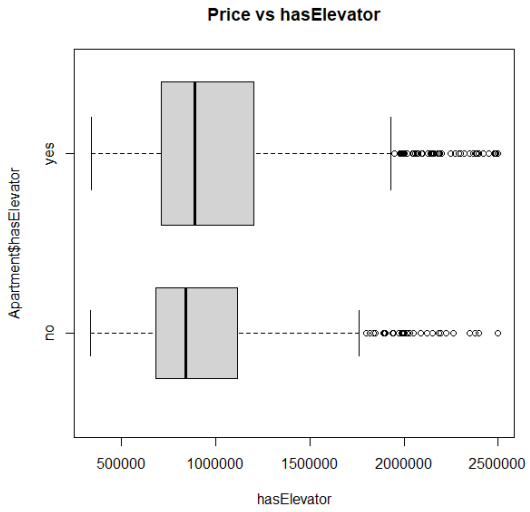
Elevator

In the data, o stands for the offer of apartment without elevator, 1 stands for with elevator.

There are 764 offers of apartment without elevator, 1928 with elevator.

Here we got almost the same result like balcony did, only difference is that the max price of apartment with elevator is the same as apartment without elevator.

	Elevator	x.Min.	x.1stQu.	x.Median	x.Mean	x.3rdQu.	x.Max.
1	no	335000	681500	840000	958455	1110250	2500000
2	yes	340800	708500	890000	1003299	1200000	2500000



Correlation

Let us now turn to the correlation analysis among our dependent and independent dummy variables. From the correlation matrix here we can tell the correlation between apartment price and apartment type, Floor types, type of points of interest, ownership, has balcony or not and has elevator or not.

The correlation between price and apartment types is 0.12.

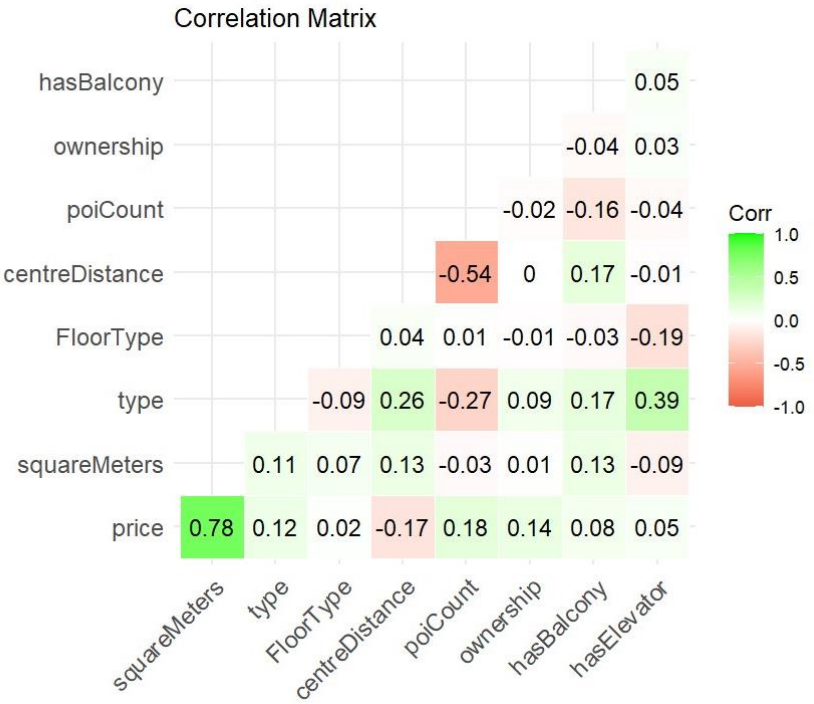
The correlation between price and floor types is 0.02.

The correlation between price and type of points of interest is 0.18.

The correlation between price and ownership is 0.14.

The correlation between price and has balcony or not is 0.08.

The correlation between price and has elevator or not is 0.05.



Third Part

In this section I am going to analyze through regression the relationship that is between the dependent variable (price) and the other explanatory variables. For the purpose of our regression, I decided to set the value of α= 1%

3.1 REGRESSION

We decided to analyse our initial model first, taking into account all the variables previously taken into consideration, without applying the log to the price and without considering any interaction.

To conduct our first regression, we decided to use a constant (intercept) to explain all phenomena that do not involve the independent variables. In order to avoid the dummy variable trap, we decided to use the tenement and low floor type, small number of interest points, cooperate ownership and no balcony, no elevator variables as the base, and thus to compare with the other categories for checking the affect of different categories on the apartment price.

```
Call:
lm(formula = price ~ type + squareMeters + FloorType + centreDistance +
    poiCount + ownership + hasBalcony + hasElevator, data = Apartment)

Residuals:
    Min      1Q  Median      3Q     Max
-804408 -129874  -18881  102919  893291

Coefficients:
                      Estimate Std. Error t value             Pr(>|t|)
(Intercept)             211809      23629    8.96 < 0.0000000000000002 ***
typeblockOfFlats       -173039      13871  -12.48 < 0.0000000000000002 ***
typeapartmentBuilding    -2040      15376   -0.13              0.89448
squareMeters             14038        178   78.89 < 0.0000000000000002 ***
FloorTypemiddle           9503      11283    0.84              0.39971
FloorTypehigh              989      10531    0.09              0.92516
centreDistance          -27578       1562  -17.65 < 0.0000000000000002 ***
poiCountmedium           40946       9608    4.26      0.0000209796925 ***
poiCountlarge            77293      14304    5.40      0.0000000710149 ***
ownershipcondominium     81382      11860    6.86      0.0000000000084 ***
hasBalconyyes            29068       8774    3.31              0.00093 ***
hasElevatoryes          100763       9885   10.19 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 205000 on 2680 degrees of freedom
Multiple R-squared:  0.757,     Adjusted R-squared:  0.756
F-statistic:  757 on 11 and 2680 DF,  p-value: <0.0000000000000002
```

As shown in the table above, there are three individually non-significant variable due to the fact that Pr(>|t|)> α= 1%. These non- significant variables are: type_apartment building, Floor type middle and floor type high. Considering that neither of the two dummy variables for floor type were significant, I decided to remove the variable for floor type.

However, all our variables result in joint significance, because the p-Value given by the F-statistic is less than α= 1%. Another important characteristic that is reported, is the value of R-squared and Adjusted R-squared. The adjusted $R^2 = 0.576$, meaning that the 57.6% of the variability of price is explained by the variability of the regressors.

By performing the RESET test on this model, no matter the RESET with powers of fitted values or RESET with powers of independent variables, the p-value turns out to be far less than α=1%. Consequently, the assumption of linearity is rejected. We thought that this is probably due to the presence of individually insignificant variables and to not normal distribution.

```
> resettest(regression1, power = 2:3, type = c("fitted"))

        RESET test

data:  regression1
RESET = 90.3, df1 = 2, df2 = 2678, p-value <0.0000000000000002
```

```
> resettest(regression1, power = 2:3, type = c("regressor"))

        RESET test

data:  regression1
RESET = 22.9, df1 = 4, df2 = 2676, p-value <0.0000000000000002
```

Consequently, in regression2 I log-transformed the price and some independent variable, to make our distribution more normal and I deleted the insignificant variable Floor type.

As I expected these changes have improved my model. This time the RESET test gave better results. Even though this time still we reject H0 for 'fitted', we don't reject H0 for 'regressor', where the p-value= 1.6%, which is larger than α=1%. So we don't reject the assumption of linearity.

```
Call:
lm(formula = log(price) ~ poly(squareMeters, 2, raw = T) + poly(centreDistance,
    3, raw = T) + type + type * squareMeters + hasBalcony + poiCount +
    hasElevator + ownership + hasElevator:poiCount + ownership +
    ownership * centreDistance, data = Apartment)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5724 -0.1180 -0.0076  0.1115  0.6063
```

```
> resettest(regression2, power = 2:3, type = c("fitted"))

        RESET test

data:  regression2
RESET = 17.2, df1 = 2, df2 = 2670, p-value = 0.000000036
```

```
> resettest(regression2, power = 2:3, type = c("regressor"))

        RESET test

data:  regression2
RESET = 1.98, df1 = 14, df2 = 2658, p-value = 0.016
```

	Dependent variable:
	log(price)
poly(squareMeters, 2, raw = T)1	0.027***
	(0.001)
poly(squareMeters, 2, raw = T)2	-0.0001***
	(0.00000)
poly(centreDistance, 3, raw = T)1	-0.060***
	(0.012)
poly(centreDistance, 3, raw = T)2	0.004**
	(0.002)
poly(centreDistance, 3, raw = T)3	-0.0001
	(0.0001)
typeblockOfFlats	-0.027
	(0.029)
typeapartmentBuilding	0.114***
	(0.031)
squareMeters	
hasBalconyyes	0.027***
	(0.007)
poiCountmedium	0.034**
	(0.015)
poiCountlarge	0.122***
	(0.021)
hasElevatoryes	0.096***
	(0.010)
ownershipcondominium	0.132***
	(0.023)
centreDistance	
typeblockOfFlats:squareMeters	-0.001***
	(0.0004)
typeapartmentBuilding:squareMeters	-0.001**
	(0.0004)
poiCountmedium:hasElevatoryes	0.003
	(0.017)
poiCountlarge:hasElevatoryes	-0.125***
	(0.022)
ownershipcondominium:centreDistance	-0.009**
	(0.004)
Constant	12.549***
	(0.047)

Observations	2,692
R2	0.796
Adjusted R2	0.795
Residual Std. Error	0.172
F Statistic	615.480***

==========

==

Note: *p<0.1; **p<0.05; ***p<0.01

Also, as we can see, after the changes, the Adjusted $R^2 = 0.795$ is higher than the original one, which means now rather than 57.6%, 79.5% of the variability of price is explained by the variability of the regressors. And these variables remain jointly significant. Only that we can see there is no value for squareMeters nor centreDistance. I think this is because I use the polynomials to the both variables and the interactions.

3.2 DIAGNOSTIC TEST

In this paragraph I'm going to check the CLRM assumptions that are: linearity, normality of residuals, homoscedasticity and multicollinearity. Because I have the cross-sectional data, there is no need to check autocorrelation.

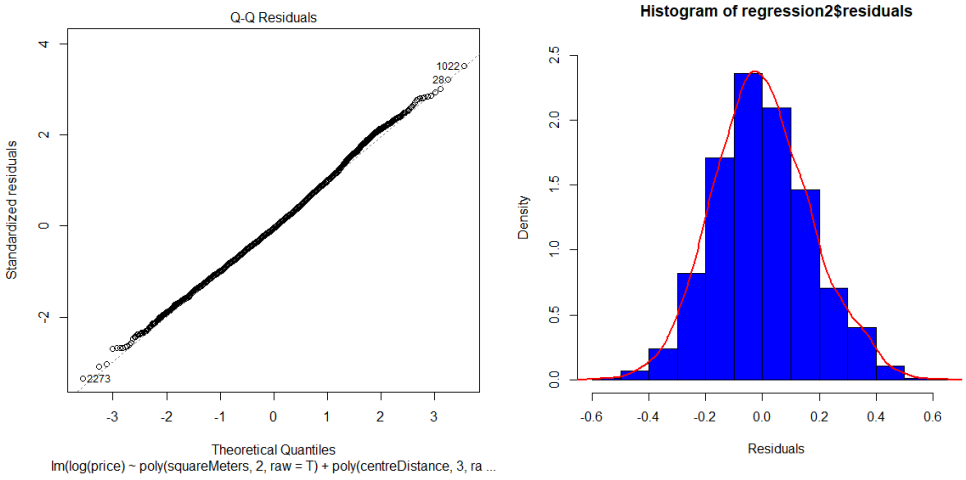
3.2.1 Linearity

The RESET test tends to check for: inappropriate choice of functional form (to solve this problem some or all of the variables in Y and X should be transformed into logs, polynomials) and/or possible omission of variables in the model and/or do interactions. As mentioned earlier after changes in regression2 the RESET test provided a p- value > α=1% which thus allowed us to accept the linearity assumption of our function.

```
> resettest(regression2, power = 2:3, type = c("regressor"))

        RESET test

data:  regression2
RESET = 1.98, df1 = 14, df2 = 2658, p-value = 0.016
```

3.2.2 Normality of residuals

Residual values in a regression analysis are the differences between the observed values in the dataset and the estimated values calculated with the regression equation. Residuals can be used to calculate the error in a regression equation. Normality is the assumption that the underlying residuals are normally distributed, it also implies that $E(\varepsilon_i) = 0$. To check graphically this assumption, we used other graphs that can be seen below. Both of them show how the distribution of our residuals may look normal.



Since the graphs seem like normal, we need to perform the Jarque-Bera test to confirm the normality. Unfortunately, the p-value in the test is smaller than 1%, so we reject the H0 hypothesis of normality of the residuals.

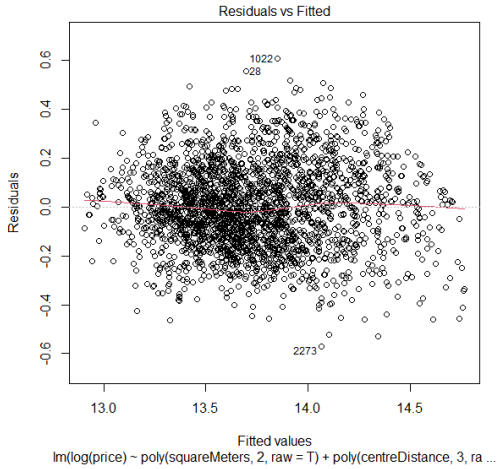
```
Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 14.4196
  P VALUE:
    Asymptotic p Value: 0.0007393
```

However, the fact that from the test the distribution of residuals is not normal is not a concern for us and we do not have to conduct further modifications, thanks to the law of large numbers. The law of large numbers, states that as a sample size grows, its mean gets closer to the average of the whole population. This is due to the sample being more representative of the population as the sample becomes larger, making the distribution normal. Our sample with 2692 observations can certainly be considered a sample large enough.

3.2.3 Homoscedasticity

Homoscedasticity refers to a condition in which the variance of the residual, in a regression model is constant Var (εi) = σ^2 for i = 1, 2, . . . , N.

It means that the error term does not vary much as the value of the predictor variable changes. To check the homoscedasticity I did a Residual vs Fitted plot.



As we can see in the graph, the red line is not completely horizontal. Then the average magnitude of the standardised residuals is changing as a function of the fitted values, but not that much. The problem is that the spread around the red line varies with the fitted values. For this reason, the variability of magnitudes varies as a function of the fitted values, creating a shaped figure and so heteroscedasticity.

This assumption is confirmed by Breusch-Pagan test, in which we reject the H0 that the regression is homoscedastic:

```
        Breusch-Pagan test

data:  regression2
BP = 103, df = 17, p-value = 0.000000000000024
```

To solve this problem we applied a robust variance-covariance matrix to our regression. As we can see from the summary below, Std. error changed and also estimates:

	Dependent variable:		
	log(price)		
	OLS	coefficient test	
	(1)	(2)	(3)
poly(squareMeters, 2, raw = T)1	0.027***	0.027***	0.027***
	(0.001)	(0.001)	(0.001)
poly(squareMeters, 2, raw = T)2	-0.0001***	-0.0001***	-0.0001***
	(0.00000)	(0.00000)	(0.00000)
poly(centreDistance, 3, raw = T)1	-0.060***	-0.060***	-0.060***
	(0.012)	(0.012)	(0.012)
poly(centreDistance, 3, raw = T)2	0.004**	0.004**	0.004**
	(0.002)	(0.002)	(0.002)
poly(centreDistance, 3, raw = T)3	-0.0001	-0.0001	-0.0001
	(0.0001)	(0.0001)	(0.0001)
typeblockOfFlats	-0.027	-0.027	-0.027
typeapartmentBuilding	0.114***	0.114***	0.114***
	(0.031)	(0.030)	(0.030)
squareMeters			
hasBalconyyes	0.027***	0.027***	0.027***
	(0.007)	(0.007)	(0.007)
poiCountmedium	0.034**	0.034**	0.034**
	(0.015)	(0.015)	(0.015)
poiCountlarge	0.122***	0.122***	0.122***
	(0.021)	(0.024)	(0.024)
hasElevatoryes	0.096***	0.096***	0.096***
	(0.010)	(0.010)	(0.010)
ownershipcondominium	0.132***	0.132***	0.132***
	(0.023)	(0.023)	(0.024)
centreDistance			
typeblockOfFlats:squareMeters	-0.001***	-0.001***	-0.001***
	(0.0004)	(0.0004)	(0.0005)
typeapartmentBuilding:squareMeters	-0.001**	-0.001**	-0.001**
	(0.0004)	(0.0005)	(0.0005)
poiCountmedium:hasElevatoryes	0.003	0.003	0.003
	(0.017)	(0.017)	(0.017)
poiCountlarge:hasElevatoryes	-0.125***	-0.125***	-0.125***
	(0.022)	(0.024)	(0.025)
ownershipcondominium:centreDistance	-0.009**	-0.009**	-0.009**
	(0.004)	(0.004)	(0.004)
Constant	12.549***	12.549***	12.549***
	(0.047)	(0.048)	(0.048)
Observations	2,692		
R2	0.796		
Adjusted R2	0.795		
Residual Std. Error	0.172 (df = 2674)		
F Statistic	615.480*** (df = 17; 2674)		
Note:			*p<0.1; **p<0.05; ***p<0.01

3.2.4 Multicollinearity

Multicollinearity is a statistical concept where several independent variables in a model are correlated. Multicollinearity reduces the precision of the estimated coefficients, which weakens the statistical power of the regression model.

To check the presence of multicollinearity in our model, we observed the VIF. As a rule of thumb, VIF should be smaller than 10, but because many of the variables that we will have to omit because of this are significant and are important for the RESET test, we remain the model unchanged.

	Variables	Tolerance	VIF
1	poly(squareMeters, 2, raw = T)1	0.0000000	Inf
2	poly(squareMeters, 2, raw = T)2	0.0464578	21.5249
3	poly(centreDistance, 3, raw = T)1	0.0000000	Inf
4	poly(centreDistance, 3, raw = T)2	0.0022368	447.0642
5	poly(centreDistance, 3, raw = T)3	0.0068016	147.0232
6	typeblockOfFlats	0.0539436	18.5379
7	typeapartmentBuilding	0.0553941	18.0525
8	squareMeters	0.0000000	Inf
9	hasBalconyyes	0.9127614	1.0956
10	poiCountmedium	0.2452342	4.0777
11	poiCountlarge	0.2097453	4.7677
12	hasElevatoryes	0.5040298	1.9840
13	ownershipcondominium	0.1652074	6.0530
14	centreDistance	0.0000000	Inf
15	typeblockOfFlats:squareMeters	0.0537836	18.5930
16	typeapartmentBuilding:squareMeters	0.0504040	19.8397
17	poiCountmedium:hasElevatoryes	0.2378780	4.2038
18	poiCountlarge:hasElevatoryes	0.2494211	4.0093
19	ownershipcondominium:centreDistance	0.0637874	15.6771

Fourth Part

4.1 Problems With Data

To detect non-typical observations (often referred to as outliers or influential points) in my data, I've chosen three measures: Leverage, Cook's Distance, and Standardized Residuals. By testing the analysis we can tell that there are 17 observations 12 variables in the dataset are non-typical.

Leverage points are observations that have unusual predictor (independent variable) values. High leverage points can have a disproportionate impact on the estimation of the regression model. To analyze leverage, we can plot leverage values and identify points that significantly stand out from others. A common rule of thumb is that an observation with a leverage value more than 2(k+1)/n (where k is the number of predictors and n is the sample size) is considered high. Through test, we get the threshold is 0.008915.

Standardized residuals are the residuals (the difference between observed and predicted values) divided by their standard deviation. They are useful for identifying outliers in the response variable. Observations with large absolute standardized residuals (typically greater than 2 or less than -2) are potential outliers.

Model Summary

R	0.870	RMSE	204776.565
R-Squared	0.757	Coef. Var	20.673
Adj. R-Squared	0.756	MSE	41933441583.079
Pred R-Squared	0.754	MAE	152331.399

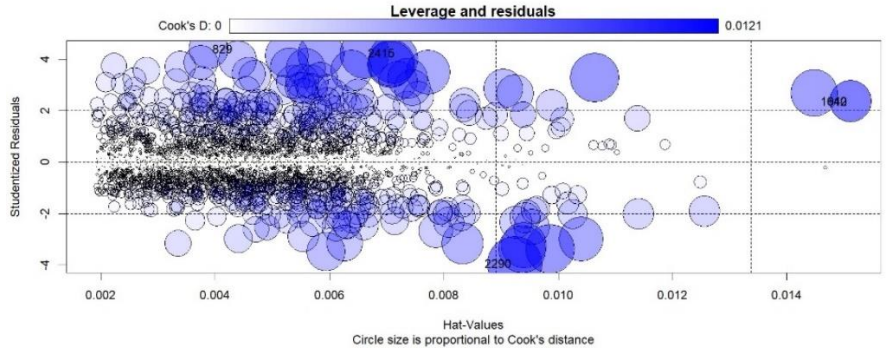
RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

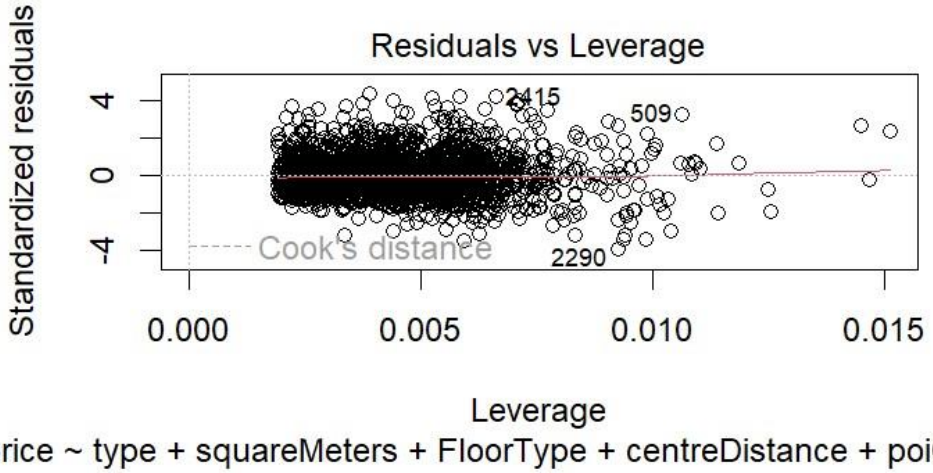
ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	349324021930985.875	11	31756729266453.262	757.313	0.0000
Residual	112381623442651.641	2680	41933441583.079		
Total	461705645373637.500	2691			

Parameter Estimates

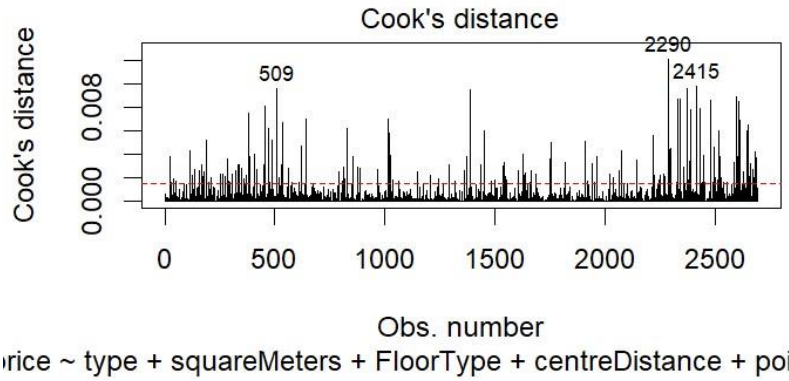
model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	211809.152	23628.550		8.964	0.000	165477.120	258141.183
typeblockOfFlats	-173039.208	13870.709	-0.207	-12.475	0.000	-200237.581	-145840.834
typeapartmentBuilding	-2039.575	15375.650	-0.002	-0.133	0.894	-32188.912	28109.762
squareMeters	14037.982	177.950	0.785	78.887	0.000	13689.049	14386.914
FloorTypemiddle	9502.985	11282.589	0.011	0.842	0.400	-12620.475	31626.445
FloorTypehigh	989.324	10531.166	0.001	0.094	0.925	-19660.708	21639.356
centreDistance	-27578.183	1562.309	-0.211	-17.652	0.000	-30641.636	-24514.731
poiCountmedium	40945.551	9607.591	0.045	4.262	0.000	22106.512	59784.591
poiCountlarge	77292.765	14303.623	0.065	5.404	0.000	49245.512	105340.018
ownershipcondominium	81381.644	11859.463	0.068	6.862	0.000	58127.021	104636.267
hasBalconyyes	29067.730	8773.569	0.033	3.313	0.001	11864.081	46271.378
hasElevatoryes	100763.100	9885.264	0.110	10.193	0.000	81379.585	120146.615





ırice ~ type + squareMeters + FloorType + centreDistance + poiCour

Cook's Distance measures the influence of each observation on the fitted values of the model. It combines information about both the leverage of the observation and its residual. To analyze Cook's Distance, look for observations with values significantly larger than the rest. A common heuristic is that observations with a Cook's Distance greater than 1 are considered highly influential. The threshold is 0.001486. From the graph below, we can tell that 509,2290,2415 are the outliers.



ırice ~ type + squareMeters + FloorType + centreDistance + poiCour

According to the three measures, I decide to delete the outliers 1012,1640,2290.

After removing atypical observations from the dataset, I have rerun the regressions on the most original and improved models and will recheck for linearity, normality, and homoskedasticity. In the most primitive models, the tests for linearity, normality and homoskedasticity still fail.

```
Call:
lm(formula = price ~ type + squareMeters + FloorType + centreDistance +
    poiCount + ownership + hasBalcony + hasElevator, data = Apartment)

Residuals:
    Min      1Q  Median      3Q     Max
-804408 -129874  -18881  102919  893291

Coefficients:
                      Estimate Std. Error t value             Pr(>|t|)
(Intercept)             211809      23629    8.96 < 0.0000000000000002 ***
typeblockOfFlats       -173039      13871  -12.48 < 0.0000000000000002 ***
typeapartmentBuilding    -2040      15376   -0.13              0.89448
squareMeters             14038        178   78.89 < 0.0000000000000002 ***
FloorTypemiddle           9503      11283    0.84              0.39971
FloorTypehigh              989      10531    0.09              0.92516
centreDistance          -27578       1562  -17.65 < 0.0000000000000002 ***
poiCountmedium           40946       9608    4.26      0.0000209796925 ***
poiCountlarge            77293      14304    5.40      0.0000000710149 ***
ownershipcondominium     81382      11860    6.86      0.0000000000084 ***
hasBalconyyes            29068       8774    3.31              0.00093 ***
hasElevatoryes          100763       9885   10.19 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 205000 on 2680 degrees of freedom
Multiple R-squared:  0.757,     Adjusted R-squared:  0.756
F-statistic:  757 on 11 and 2680 DF,  p-value: <0.0000000000000002
```

```
        RESET test

data:  regression_new
RESET = 90.3, df1 = 2, df2 = 2675, p-value <0.0000000000000002
```

```
        RESET test

data:  regression_new
RESET = 21.5, df1 = 4, df2 = 2673, p-value <0.0000000000000002
```

```
Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 473.5166
  P VALUE:
    Asymptotic p Value: < 0.00000000000000022
```

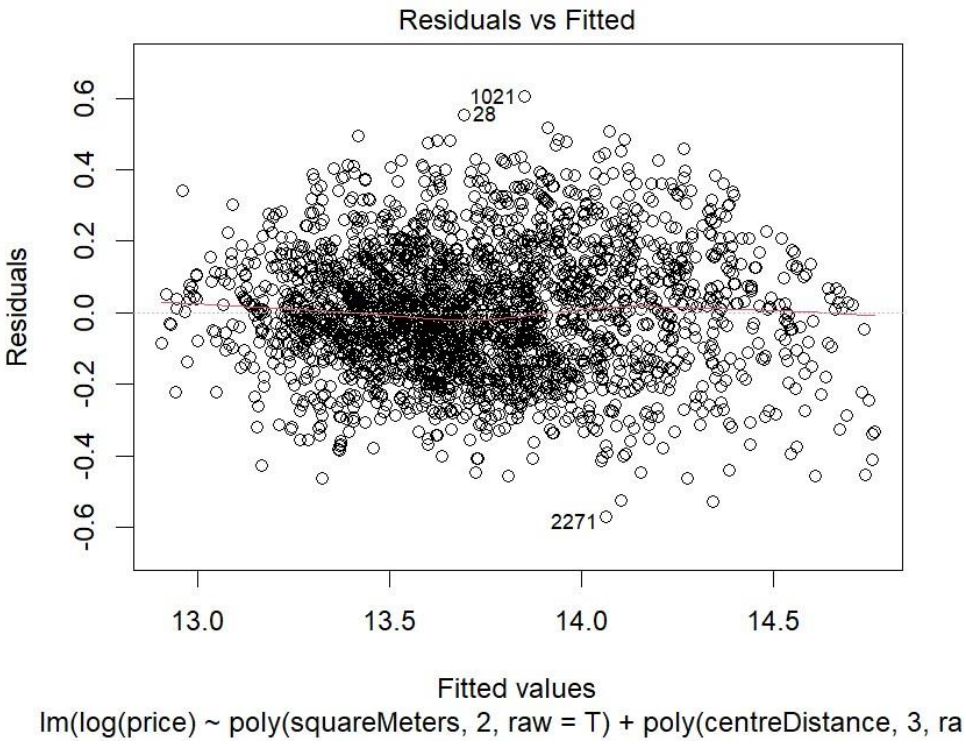
```
        studentized Breusch-Pagan test

data:  regression_new
BP = 494, df = 11, p-value <0.0000000000000002
```

But for the improved model, there is a significant improvement after removing the outliers. For linearity, the p_value increases from 0.16 to 0.17 in the reset test, and from 0.0007393 to 0.0007999 in the Jarque - Bera Normalality Test, only in the Breusch-Pagan homoskedasticity test does the p_value become slightly smaller. became slightly smaller. So the final model is an improved model after removing the outliers.

```
Call:
lm(formula = log(price) ~ poly(squareMeters, 2, raw = T) + poly(centreDistance,
    3, raw = T) + type + type * squareMeters + hasBalcony + poiCount +
    hasElevator + ownership + hasElevator:poiCount + ownership +
    ownership * centreDistance, data = Apartment_new)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5706 -0.1177 -0.0077  0.1115  0.6047

Coefficients: (2 not defined because of singularities)
                                    Estimate  Std. Error t value             Pr(>|t|)
(Intercept)                       12.55450438  0.04751416  264.23 < 0.0000000000000002 ***
poly(squareMeters, 2, raw = T)1    0.02726462  0.00077128   35.35 < 0.0000000000000002 ***
poly(squareMeters, 2, raw = T)2   -0.00009018  0.00000446  -20.24 < 0.0000000000000002 ***
poly(centreDistance, 3, raw = T)1 -0.06064096  0.01168038   -5.19          0.000000224 ***
poly(centreDistance, 3, raw = T)2  0.00410828  0.00161494    2.54              0.01102 *
poly(centreDistance, 3, raw = T)3 -0.00011273  0.00007149   -1.58              0.11494
typeblockOfFlats                  -0.02910656  0.02890761   -1.01              0.31408
typeapartmentBuilding              0.11530755  0.03105371    3.71              0.00021 ***
squareMeters                              NA          NA      NA                   NA
hasBalconyyes                      0.02632393  0.00746429    3.53              0.00043 ***
poiCountmedium                     0.03282712  0.01486593    2.21              0.02731 *
poiCountlarge                      0.11861946  0.02098059    5.65          0.000000017 ***
hasElevatoryes                     0.09537498  0.01038472    9.18 < 0.0000000000000002 ***
ownershipcondominium               0.12903053  0.02351754    5.49          0.000000045 ***
centreDistance                            NA          NA      NA                   NA
typeblockOfFlats:squareMeters     -0.00144566  0.00043946   -3.29              0.00102 **
typeapartmentBuilding:squareMeters -0.00105445  0.00044598   -2.36              0.01813 *
poiCountmedium:hasElevatoryes      0.00352051  0.01709001    0.21              0.83681
poiCountlarge:hasElevatoryes      -0.12241955  0.02246011   -5.45          0.000000055 ***
ownershipcondominium:centreDistance -0.00826687  0.00363408   -2.27              0.02300 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.172 on 2671 degrees of freedom
Multiple R-squared:  0.796,     Adjusted R-squared:  0.795
F-statistic:  614 on 17 and 2671 DF,  p-value: <0.0000000000000002
```

```
        RESET test

data:  regression2new
RESET = 1.96, df1 = 14, df2 = 2655, p-value = 0.017
```

```
Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 14.262
  P VALUE:
    Asymptotic p Value: 0.0007999


        studentized Breusch-Pagan test

data:  regression2new
BP = 104, df = 17, p-value = 0.000000000000017
```



Fifth Part

5.1 INTERPRETATION OF FINAL RESULTS

After all the modification, my final model is regression2new showing below.

log(price) = 12.55+0.027 square meters – 0.00009 squareMeters^2 -0.06centre distance + 0.0041 centre distance^2 - 0.0001 centre distance^3 - 0.029type block of flats +0.1153 type apartment building +0.026hasBalcony+0.033poiCountmedium + 0.1186poiCountlarge + 0.095hasElevator + 0.129ownership of condominium – 0.0014 type block of flats under same square meters – 0.0011 type apartment building under same square meters +0.0035 poiCountmedium under condition has elevator -0.1224 poiCountlarge under condition has elevator – 0.0083 ownership condominium under same centre distance.

Size

Semi-elasticity: size (square meters) has a coefficient of 0.027 and size squared of -0.00009, which means that as the size increases, the price of the apartment grows and then decreases. The derivation of price with respect to size shows that the point of maximum value is 151.67 square meters, so the price rises by 2.7% for each additional square meter up to 151.67 square meters, and after that, the price falls by 0.009% for each additional square meter.

Center Distance

Semi-elasticity: center distance has a coefficient of 0.06 and center distance squared of 0.0041, and the cube of distance of -.0001, which means the distance to the city center has gone through several stages of growth and decline. After the derivation of price with respect to distance we know that the extreme value point is -6 and the minimal value point is 33.3, and since the distance cannot be negative, we know that between 0 meters and 33.3 kilometers, the price decreases as the distance to the city center increases, and after it is greater than 33.3 kilometers, the price increases as the distance increases.

Type

Semi-elasticity: the coefficient of apartment type block of flats and price is -0.029, which means that block of flats has 2.9% lower price than tenement. The coefficient of apartment type apartment building and price is 0.1153, which means that apartment building has 11.53% higher price than tenement.

Has balcony

Semi-elasticity: coefficient 0.026 means that apartments with balcony have generally 2.6% higher prices than those without one.

Has elevator

Semi-elasticity: coefficient 0.095 means that apartments with elevator have generally 9.5% higher prices than those without one.

Type of points of interest:
Semi-elasticity: coefficient 0.033 means that apartments with medium type of interest points have generally 3.3% higher prices than those with small type. Coefficient 0.1186 means that apartments with large number of interest points have generally 11.86% higher prices than those with small type.

Ownership
Semi-elasticity: coefficient 0.129 means that apartments of condominium type ownership have generally 12.9% higher prices than those with cooperative type.

Interaction between type and size
Type is valued differently depending on the size of apartments. As such almost the reverse happens. Under the same square meters type block of flats has 0.14% lower price than tenement. Type apartment building under same square meters has 0.11% lower than tenement, complete opposite result than before.

Interaction between points type of interest and has elevator
Points of interest medium type number under condition has elevator have 0.35% higher prices than small type number of interest points. Points of interest larger type number under condition has elevator have 12.24% lower prices than small type number of interest points.

Interaction between ownership and center distance
With the same distance to city center, condominium type ownership have generally 0.83% lower prices than those with cooperative type.

5.2 HYPOTHESIS VERIFICATION

After analyzing the data, performing regressions, and solving the problems encountered, we can ultimately validate the correctness of our initial hypotheses by interpreting the estimates based on life experience and relevant literature.

The first hypothesis is rejected. Through regression analysis I found that tenement buildings are not the cheapest, flats are the cheapest, and apartment buildings are the most expensive.

The second hypothesis is also rejected There is indeed a non-linear relationship between apartment size and price, but instead of prices rising sharply as apartment size increases, they increase and then decrease.

Since the coefficient on the variable floor type was not significant, I removed it from the regression, so I have no evidence to reject the third hypothesis.

The fourth hypothesis is also rejected, as through regression we find a decreasing and then increasing relationship between distance to the city center and apartment prices.

There is no reason to reject the fifth hypothesis, and by regression I find that it is indeed true that the greater the number of points of interest within one kilometer of an apartment, the more expensive the apartment.

Similarly, there is no reason to reject the sixth, seventh and eighth hypotheses. We can see from the regression results that condominiums are more expensive than co-ops, those with elevators are also more expensive than those without, and those with balconies are indeed more expensive than those without.

Sixth Part

6.1 CONCLUSIONS AND SUMMARY

This study has two main targets: the first one is to test previously known determinants. The second one is to incorporate more important determinants in the characteristic price models. Through hypothesis testing and regression analysis, I found that the type of apartment floor among the original determinants does not have a significant effect on its price. We should remove it from the modeling of apartment price determinants. The size of the apartment and the distance from the apartment to the city center are also not simply linearly related. Moreover, different apartment sizes have a large impact on the relationship between the type of apartment and its price. The availability of an elevator also has a significant effect on the relationship between the number of points of interest in the surrounding area and price. Similarly, the distance from the apartment to the city center also affects the relationship between the apartment's ownership and the price.

The adjusted R-squared of our regression results is 0.795. this means that our variables explain 79.5% of the model. This allows us to use our regression variables to make predictions about prices, and therefore our model proves to be partially reliable. However, since there is still more than 20% that cannot be explained by my model, it would be particularly interesting to investigate this issue further, perhaps by considering other variables not analyzed in our model that could better explain the factors that affect apartment prices in Warsaw, or even the differences in apartment prices in other parts of the world.

References

Bhattacharjee A, Castro E, Marques J. Spatial interactions in hedonic pricing models: the urban housing market of Aveiro, Portugal[J]. Spatial Economic Analysis, 2012, 7(1): 133-167.

Chau K W, Chin T L. A critical review of literature on the hedonic price model[J]. International Journal for Housing Science and its applications, 2003, 27(2): 145-165.

Cui N, Gu H, Shen T, et al. The impact of micro-level influencing factors on home value: A housing price-rent comparison[J]. Sustainability, 2018, 10(12): 4343.

Leszczyński R, Olszewski K. An analysis of the primary and secondary housing market in Poland: evidence from the 17 largest cities[J]. Baltic Journal of Economics, 2017, 17(2): 136-151.

Lancaster K J. A new approach to consumer theory[J]. Journal of political economy, 1966, 74(2): 132-157.

Olszewski K, Waszczuk J, Widłak M. Spatial and hedonic analysis of house price dynamics in Warsaw, Poland[J]. Journal of Urban Planning and Development, 2017, 143(3): 04017009.

Posedel P, Vizek M. House price determinants in transition and EU-15 countries[J]. Post-Communist Economies, 2009, 21(3): 327-343.

Ridker R G, Henning J A. The determinants of residential property values with special reference to air pollution[J]. The review of Economics and Statistics, 1967: 246-257.

Rosen S. Hedonic prices and implicit markets: product differentiation in pure competition[J]. Journal of political economy, 1974, 82(1): 34-55.

Wen H Z, Sheng-hua J, Xiao-yu G. Hedonic price analysis of urban housing: An empirical research on Hangzhou, China[J]. Journal of Zhejiang University-Science A, 2005, 6(8): 907-914.