TU Dortmund

Introductory Case Studies

Project 3: Regression analysis

Lecturers:

Prof. Dr. Jörg Rahnenführer

Dr. Philipp Adämmer

Dr. Andrea Bommert

M. Sc. Julia Duda

M.Sc. Hendrick Dohme

Author: Ahmed Sameed

Group number: 11

Group members: Ahmed Sameed, Ojuri Moyinoluwa Samuel, Thua Phuc Nguyen Vo

July 9, 2021

Contents

1. Introduction

Several cities across Germany have their own 'rent index'. These indices provide tenants and landlords an overview of the housing market situation in a given area. A noteworthy example is that of Munich; which for the past 20 years, has been one of the most expensive cities for renters in the country (Pladson 2019). Munich's official city portal outlines its rent index (Wohnen und Migration München 2021). The website also features an online calculation program for the total rent of a property by prompting to insert its details such as its size, year of construction, the building type, etc. This estimation of the average net rent of a property from its characteristics constitutes a typical regression problem.

This report is closely related to the above-mentioned example of rental price evaluation. Here, the analysis is done on rental offers for the city of Dortmund. Rather than evaluating the total rent of a property, a linear model for estimating the rental price per square meter of the properties is built to check the factors which are affecting rental prices.

The statistical analysis of the data comprises of three parts. Firstly, the data is preprocessed to deal with missing values in some of the variables and to transform some of them by grouping their outcomes into new categories. In the second part, the univariate analysis of the variables is carried out to investigate the dataset's underlying structure. This is followed by analysing the trained linear model for the rental price per square meter, and its assumptions.

The goal of the report is to train and assess the aforementioned linear regression models for rent per square meter. Simple regression models are trained without employing any polynomial effects of the covariates or their interactions. Akaike information criterion and Bayesian information criterion has been deployed as a selection criterion to find the best predictor for rental price per square using the best subset selection. A best possible model is then estimated using the Akaike Information Criterion. Lastly, the predictive power of the linear model is judged by the adjusted R-squared coefficient.

Besides this Introduction, the report consists of four more sections. The Problem Statement section describes the variables in the dataset in detail and discusses the data quality. For instance, around 13.44% of the rental offers have missing values for the total rent, the variable used to compute the response variable for the linear model. The methods portion of the report provides formulas and assumptions for the linear regression models. It also outlines the process of best subset selection, the Akaike Information

Criterion and the coefficient of determination for the goodness of fit. The Statistical Analysis section describes the preprocessing of the dataset, provides univariate analysis of all variables and interprets the coefficients of the final linear regression model. Finally, the Summary portion concisely rehearses the most important results. It discusses potential improvements to the experiment, such as collecting more data, and suggests ways in which it can be extended.

2. Problem Statement

The data is collected from the German real estate web-portal Immobilienscout24. The website features rental offers as well as homes for sale. The dataset used in this report comprises of 12118 rental offers as of 20 February 2020. All properties are located in the state of North Rhine-Westphalia, Germany. The full dataset is available on `https://www.kaggle.com/corrieaar/apartment-rental-offers-in-germany`.
The dataset consists of 16 variables; ten of these being categorical and six numeric. There are five binary categorical variables which assume the values true or false. When the variable 'newlyConst' is true, this means that a property is newly constructed (i.e.constructed in 2019 or 2020), and vice versa. Similarly, when the variables 'balcony', 'hasKitchen', 'lift' or 'garden' are set to true, this implies that a property has a balcony, a kitchen, a lift or a garden and vice versa.
Three other nominal variables also have two levels. The variable 'condition' which indicates the condition in which the property is, takes on the values average and good. The variable 'lastRefurbish' specifies the time period in which a given property was last renovated, with levels last 5 years and over 5 years ago. The variable 'energyEfficiencyClass' indicates the energy efficiency class of the building and has the levels A+/A/B/C and D/E/F/G/H.
The two remaining categorical variables have more than 2 levels. The variable 'typeOfFlat' indicates the type of flat and has 10 levels: roof storey, apartment, ground floor, half basement, penthouse, terraced flat, maisonette, raised ground floor, loft and other. The other variable 'regio2' refers to the city in which a property is located and has 54 levels.
Out of the six numeric variables, four are discrete. The variable 'ID' is a unique arbitrary identification number assigned to each rental offer. It takes the values of the natural numbers from 1 to 12118. The variable 'yearConstructed' indicates the year of

construction of a property. The variable 'noParkSpaces' refers to the number of parking spaces provided with a rental offer. The variable 'floor' indicates the floor in which a property is located. For this variable, a value of -1 means that the property is located in a basement; a 0 indicates ground floor; a 1 indicates the first floor and so on.
The remaining two numeric variables are continuous. The variable 'totalRent' refers to the total rent of a property, which includes its base rent, service charges and heating costs. And finally, the variable 'livingSpace' indicates the size of a property in square meters.
The data quality suggests that preprocessing is required before training the linear regression models. For instance, many of the variables in the dataset have missing values. For 'totalRent' these amount to 2284; for 'noParkSpaces' 7931; for 'typeOfFlat' 714; for 'lastRefurbish' 8088; for 'condition' 2845; and for 'energyEfficiencyClass' 8457 in total. Hence, during the preprocessing step, these missing values have to be treated in a meaningful way.
The objectives of the report are, first of all, to preprocess the data and carry out a univariate analysis of the variables. Thereafter, a linear model is built for the rent per square meter of the properties as the response variable, and then, the predictive power of the models is judged using certain judgement criterion, their shortcomings are highlighted, and possible areas of improvement are identified.

3. Statistical Methods

The following statistical methods and mathematical formulas are used. The statistical software R (R Development Core Team 2020), version 4.0.3 has been used for analysis.

3.1. Multiple Linear Regression Model

Multiple linear regression models the effect of a vector of k independent variables or covariates, $x_1, ..., x_k$, on a dependent or response variable y. Here, the response variable is continuous while the covariates can be either continuous or appropriately coded categorical variables. The response variable is not a deterministic function $f(x_1, ..., x_k)$ of the covariates. Instead, this relationship shows random errors.

$$y_i = f(x_1, ..., x_k) + \varepsilon_i = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon_i$$

The linear function f is called the systematic component of the model while the error term ε is referred to as the random or stochastic component. To model a categorical covariate $x_2 \in (1, ..., c)$ with c categories, category c can be treated as a reference. Thereafter, c $-$ 1 dummy variables can be defined as follows and then included in the model.

$$x_{i1} = \begin{cases} 1 & x_i = 1 \\ 0 & \text{otherwise} \end{cases} \quad \dots \quad x_{i,c-1} = \begin{cases} 1 & x_i = c-1 \\ 0 & \text{otherwise} \end{cases}$$

The parameters $\beta_0, ..., \beta_k$ are unknown. By combining the covariates and the parameters into separate $p = k + 1$ dimensional vectors, $\boldsymbol{x} = (1, x_1, ..., x_k)'$ and $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)'$, the systematic component can be expressed as a vector product. Therefore,

$$y = f(\boldsymbol{x}) + \varepsilon = \boldsymbol{x'\beta} + \varepsilon$$

To estimate the parameters, data $(y_i, x_{i1}, ..., x_{ik})$ is collected where $i = 1, ...n$. The vectors $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ and the design matrix $\boldsymbol{X}$ are defined as follows:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}$$

Here the errors are normally distributed, with zero mean and a constant variance across them (i.e. homoscedastic errors), $\boldsymbol{\varepsilon} \sim (\boldsymbol{0}, \sigma^2\boldsymbol{I})$. The design matrix $\boldsymbol{X}$ is assumed to have full column rank, implying that all columns are linearly independent. Hence, n equations can be formed, as follows

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

The unknown parameters $\boldsymbol{\beta}$ are estimated using the method of least squares. Here the estimates $\hat{\boldsymbol{\beta}}$ are the minimizers of the sum of squared deviations.

$$\text{LS}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$$

Setting the derivative of the above expression with respect to $\boldsymbol{\beta}$ equal to zero, leads to the unique solution of the least squares estimator.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$$

(Fahrmeir et al. 2013, p. 74-107).

To test for the significance of a parameter β_j, the hypotheses are $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$. The estimated t-statistic is calculated as follows:

$$\hat{t}_j = \frac{\hat{\beta}_j}{\hat{se}_j},$$

where $\hat{se}_j = [\widehat{\text{Var}(\hat{\beta}_j)}]^{1/2}$ is the estimated standard error of $\hat{\beta}_j$. For a predefined significance level of α (in this report 0.05), the absolute value of the above statistic is compared to the $(1 - \alpha/2)$th quantile of the t-distribution with $n - p$ degrees of freedom. H_0 is rejected if:

$$|\hat{t}_j| > t_{1-\alpha/2}(n - p)$$

(Fahrmeir et al. 2013, p. 135).

3.2. Coefficient of Determination

For a linear regression model, the coefficient of determination R^2 is defined as

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})} = 1 - \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i{}^2}{\sum_{i=1}^{n}(y_i - \bar{y})}$$

where $\bar{y}$ is the mean value of the response variable and $\hat{y}_i$ for $i = 1, ..., n$ are its estimated values (Fahrmeir et al. 2013, p. 115).

R^2 has the range, $0 \leq R^2 \leq 1$. When R^2 is closer to 1, the residual sum of squares is smaller and the fit to the data is better. If R^2 is closer to 0, this sum is larger and the regression model is poorly fitted. R^2 is a measure of the proportion of the variance in the response variable that is predictable from the covariates.

The coefficient of determination is of limited value when it comes to model comparison. The corrected coefficient of determination mitigates its shortcomings by incorporating a correction term in the formula to account for the number of parameters in the model.

$$\bar{R}^2 = 1 - \frac{n-1}{n-p}(1 - R^2)$$

(Fahrmeir et al. 2013, p. 147-148)

3.3. Kernel Density Estimation

Kernel density estimation is useful for visualizing the shape of some data, as a kind of continuous replacement for the discrete histogram. It can also be used to generate points that look like they came from a certain dataset. Mathematically, KDE is given as

$$\widehat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

where $K(x)$ is called the kernel function that is generally a smooth, symmetric function such as a Gaussian and $h > 0$ is called the smoothing bandwidth that controls the amount of smoothing. Basically, the KDE smoothes each data point X_i into a small density bumps and then sum all these small bumps together to obtain the final density estimate. (Skrondal 2010, p. 232-233)

3.4. Best Subset Selection

Best subset selection is a method that aims to find the subset of independent variables X_i that best predict the outcome Y and it does so by considering all possible combinations of independent variables.It Works only for multiple linear regression models. The choice for model selection is given as

$$\begin{pmatrix} p \\ k \end{pmatrix} = p!/[k!(p-k)!]$$

Best subset selection ends up selecting 1 model which has highest adjusted-R^2 value from 2k possible models (Sergio Bacallado 2020).

3.5. Akaike Information Criterion

The AIC is an index used to to help choose among competing models. A smaller value of the index indicates the preferred model. The index is defined as follows:

$$\text{AIC} = 2k - 2\log(\hat{L})$$

where k is the number of parameters and $\hat{L}$ is the maximum value of the likelihood function. The first term in the equation above penalizes a larger number of parameters; the more the parameters, the higher the AIC value. The second one takes into account the statistical goodness of fit of the model; the better the fit, the lower the index value (Skrondal 2010, p. 10).

3.6. Bayesian Information Criterion

In statistics, the Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models. It is based, in part, on the likelihood function, and it is closely related to Akaike information criterion (AIC).
While fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. The BIC resolves this problem by introducing a penalty term for the number of parameters in the model. The penalty term is larger in BIC than in AIC. The formula for the BIC is as follows

$$-2 \cdot \ln p(x \mid k) \approx \text{BIC} = -2 \cdot \ln L + k \ln(n)$$

where

- x = the observed data;
- n = the sample size;
- k = is the number of regressors, including the intercept;
- $p(x \mid k)$ = the probability of the observed data given the number of parameters; or, the likelihood of the parameters given the dataset;
- L = the maximized value of the likelihood function for the estimated model

(Skrondal 2010, p. 37-38)

4. Statistical Analysis

4.1. Data Preparation

The given data set contains total of 12118 rental offers for properties located in the province of North Rhine-Westphalia. Since, our regression analysis is based only for the city of Dortmund, so a subset of values has been derived from original data set containing values for the city of Dortmund. Subset contains 558 observations for 16 variables. To deal with the missing values, a check has been done to see which variable has the most missing values. Figure-4 from Appendix shows that 75% of data is missing for the variable 'noParkSpaces', so we will drop this variable. Afterwards, each row of subset is searched for the presence of a single missing value and upon success that row is dropped. Lastly, the redundant variables 'ID' and 'regio2' are dropped from the subset as they will not have any role in further analysis. After the data cleaning, the subset contains 468 observations for 13 variables.
The response variable for the linear regression model is the rent per square meter for the properties. As this is not one of the 16 variables in the provided dataset, it has to be calculated by dividing each observation of the variable 'totalRent' by its corresponding observation of 'livingSpace' to form a new variable 'sqmPrice'.
The variable 'typeOfFlat' assumes 10 values. The value 'apartment' is made into its own category while the remaining 9 values are grouped into three more categories as follows: loft, maisonette, penthouse, terraced flat and 'other' are grouped to form the category 'luxurious_artistic_other'; ground floor and raised ground floor are grouped to form the category 'r_ground_floor' and roof storey and half basement are grouped to form the category 'roof_halfBasement'.

4.2. Univariate Analysis of the Variables

Table-1 shows the grouping of variable 'typeOfFlat' after the aforementioned transformations, suggesting that 76% of the properties in Dortmund are apartments. The figure-1 shows the distribution of price per square meter for the variable 'typeOfFlat". It can be seen that prices for 'Apartment' and 'Luxurious_artistic_other' are normally distributed with the mean of 10.6 and 12.4 respectively, whereas the 'r_ground_floor' and 'roof_half_basement' is right skewed with mean of 10.3 and 10.2 respectively.

Apartment	Luxurious/ artistic/other	Raised ground floor/ ground floor	Roof storey/ half basement
357	25	40	46

Table 1: Frequency distribution of the type of flat

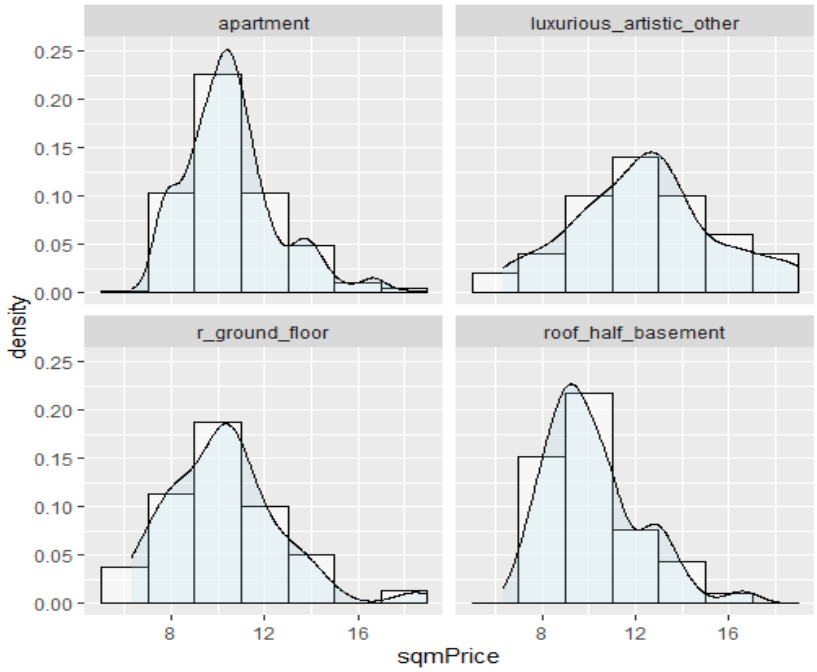


Figure 1: Histogram of Price per Square Meter for various types of flats

Figure-2 shows the correlation between the response variable 'price per square meter' and the explanatory variable 'Living space'. It can be seen that there is negative correlation between response and explanatory variable with coefficient of correlation equals to -0.101.

Box plot from figure-3(a) shows that there is an increase in price per square meter for the property which has kitchen included as compared to the property which does not has one. Boxplot from figure-3(b), suggests that there is a significant difference in the square meter price of newly constructed property as compared to the one which is old. Figures-5, 6, 7 and 8 from the Appendix, suggests that there is no significant change in price per square meter caused by the covariates "typeOfFlat", "condition", "balcony" and "lastRefubish" respectively.

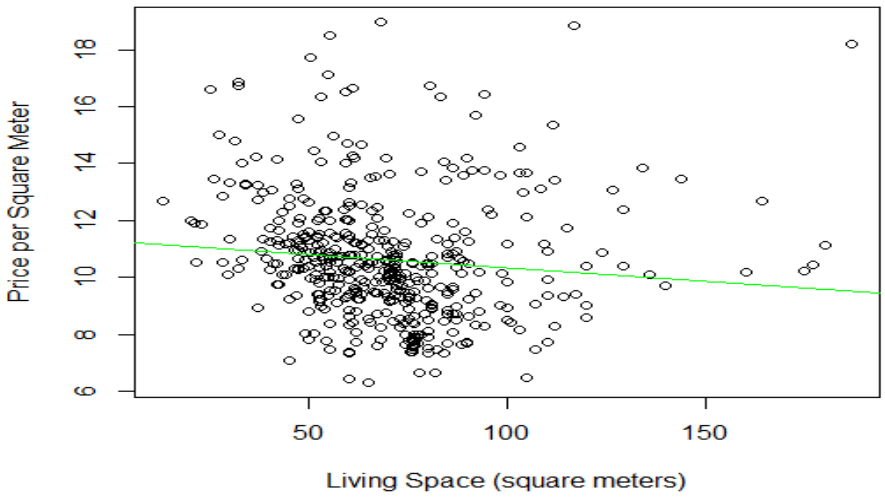
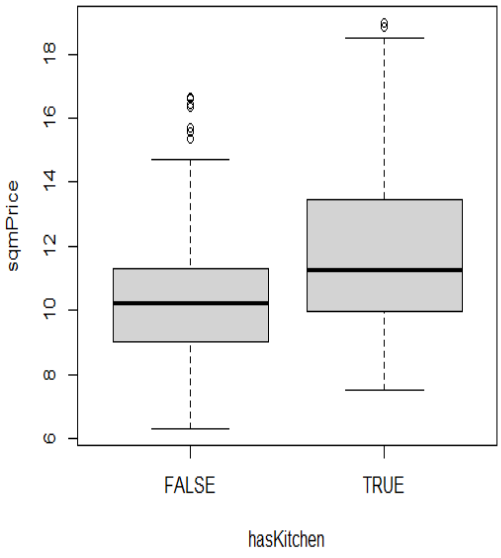
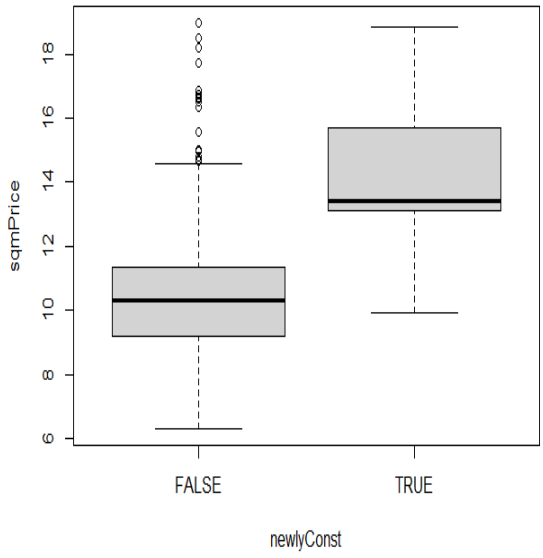


Figure 2: Price per square meter vs Living Space



(a) Price per square meter vs hasKitchen



(b) Price per square meter vs Newly Constructed

4.3. Linear Regression Analysis

For the linear regression model, the rent per square meter is modelled in terms of the remaining variables except the total rent. For regression analysis, the method of 'Best subset selection' is deployed as best predictor for response variable 'sqmPrice'. The "bestglm" function in 'R' is used for best subset selection. For dummy coding, all categorical variables are used while discrete variables are ignored. The first columns of are dummy variables are dropped as reference category to avoid dummy variable trap. For applying the "bestglm" function, the response variable is arranged in a way that it appears to be the right most variable of response-explanatory variable matrix.
After setting up the model as described above, best five models are selected using AIC and BIC as model selection criterion. Table-2 lists the best five models that are obtained using Akaike Information Criterion and Table-3 lists the five best models obtained using Bayesian Information Criterion. The best model that is predicted using Akaike Information Criterion is 'AICModel1' with AIC value equal to 1906.513, whereas for Bayesian Information Criterion, BICModel1 and BICModel3 has the same BIC value equals to 1950.062, hence any of the model can be used for further analysis. The models with lowest AIC and BIC values are chosen because smaller the AIC and BIC value, the better the model fit is.

Model	AIC
AICModel1	1906.513
AICModel2	1906.750
AICModel3	1907.021
AICModel4	1907.321
AICModel5	1907.755

Table 2: Subset of Best Models using AIC

Model	BIC
BICModel1	1950.062
BICModel2	1956.386
BICModel3	1950.062
BICModel4	1950.370
BICModel5	1951.810

Table 3: Subset of Best Models Using BIC

For $\alpha = 0.05$, the coefficients of the linear model are all statistically significant when tested for the hypothesis $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$.

The details for 'AICModel1' are given in the Table-4. The best model predicted by AIC has 13 covariates out of which 4 covariates have p-value greater than the significance level 0.05. Table-5 presents details for the 'BICModel1'. The best model predicted by BIC has 7 covariates and all of the covariates have p-values less than 0.05. The difference in number of covariates between AIC and BIC is because BIC penalizes models more for free parameters than AIC does.

Covariates	Estimated Parameter Value	(Pr> $\|t\|$
Intercept	-5.301504	0.48248
yearConstructed	0.008959	0.02003 *
livingSpace	-0.020474	1.00e-07 ***
floor	-0.130524	0.05172 .
newlyConst(TRUE)	1.624091	0.00187 **
hasKitchen(TRUE)	0.955459	7.14e-05***
lift(TRUE)	0.791034	0.00206 **
typeOfFlat(luxurious|art)	0.862907	0.03803 *
typeOfFlat(r|ground$_f loor$)	-0.643945	0.04989 *
typeOfFlat(roof|half_base)	-0.440715	0.14047
condition (good)	1.247873	1.12e-07 ***
energyEfficiencyClass (D/E/F/G/H)	-0.657118	0.06164 .
energyEfficiencyClass(NO INFORMATION)	-0.753162	0.00839 **

Table 4: Details for Best Model Using AIC

While interpreting any estimated coefficient for AICmodel1 from Table-4, the values for the other ones are assumed to be constant. According to the trained linear model, for a property that is newly constructed and has a fitted kitchen, the rent per square meter is typically 1.62 €/m^2, 0.95 €/m^2 higher, respectively, than for a property which is old and does not has a fitted kitchen. Moreover, good condition of property and having access to lift will also increase the value of property by 1.24 €/m^2 and 0.79 €/m^2 respectively. With increase in the living space and the allocated floor of rental property, there is a decrease of 0.02 €/m^2 and 0.13 €/m^2 in rent, respectively. So, if a size of property is increased and the floor number of property is also increased, then the price of property

Covariates	Estimated Parameter Value	(Pr> $\|t\|$
Intercept	11.25	2e-16 ***
livingSpace	-0.020297	1.54e-07 ***
newlyConst(TRUE)	2.215845	1.23e-05 ***
hasKitchen	0.980505	02e-05 **
lift(TRUE)	0.877235	0.000112 ***
typeOfFlat(luxurious|art)	1.046919	0.012078 *
condition(good)	1.268972	4.77e-08 ***

Table 5: Details for Best Model Using BIC

will recede. Hence the person living on fourth floor will pay less compared to the person living on ground floor. For luxurious and artistic flats, there is increase of 0.86 €/m^2 in price, whereas flats located on ground floor and in basement has decline of 0.64 €/m^2 and 0.44 €/m^2 in prices, respectively.

The adjusted R-squared value for the linear model is about 0.30. This means that our model can account for 30% of the variation in the rental price per square meter of the properties. Table-6 shows the p-values for each individual covariate for one tail test.

Covariates	2.5%	97.5%
(Intercept)	-20.123595432	9.5205884303
yearConstructed	0.001417579	0.0165208671
livingSpace	-0.027906193	-0.0130408850
floor	-0.267684985	0.0064375250
newlyConst_TRUE	0.603872524	2.6443098886
hasKitchen_TRUE	0.487019173	1.4238979417
lift_TRUE	0.289548150	1.2925196036
typeOfFlat_luxurious_artistic_other	0.047851267	1.6779633311
typeOfFlat_r_ground_floor	-1.287584751	-0.0003045903
typeOfFlat_roof_half_basement	-1.027253803	0.1458232209
condition_good	0.793034664	1.7027116617
energyEfficiencyClass_D.E.F.G.H	-1.346387837	0.0321524090
energyEfficiencyClass_NO_INFORMATION	-1.312243209	-0.1940814280

Table 6: Coefficient of best Model using AIC

5. Summary

This report concerns the rental price data for the province of North Rhine-Westphalia, Germany. 12118 rental offers are extracted from the real estate web-portal Immobilienscout24. For each offer, the following 16 variables are provided in the dataset: its ID; its total rent in Euros; its year of construction; the year when it was last renovated; the number of parking spaces provided with it; its size in square meters; whether or not it is newly constructed (i.e. constructed in the year 2019 or 2020); whether or not it has a balcony, a kitchen, a lift or a garden; its type of flat; the floor in which it is situated; the city in which it is located; its condition; and its energy efficiency class. The dataset is preprocessed to deal with the values for the city of Dortmund only. Variable with maximum number of missing values and the rows having single missing value have been removed afterwards. Next, the grouping of variables is done to group the selected variables together into categories and transform them for the ease of interpretation. From univariate analysis, it has been observed that, the properties which are newly constructed, has attached kitchen and a lift cost more, whereas there is a decline in price per meter square with increase in living space and the number of floor on which the property is located. Moreover, the type of property will also add value to its price per square meter.

After preprocessing and univariate analysis, a linear regression model is trained with the rent per square meter as response variable. Using best subset selection procedure and the AIC and BIC as the model selection criterion, the best model has been chosen. The variables ID and regio2 have been discarded from the model. The resulting model is able to explain 30% of the variation in the response variable.

To improve the quality of the results, the size of the dataset should be increased. Other cities should be taken under consideration to determine how the property value varies from one city to another. Polynomial effects can also be added to the model. Furthermore the variable "location within a city" can be added to see how the prices vary across different areas of the same city.

References

[1] Ludwig Fahrmeir et al. *Regression: Models, Methods and Applications*. Jan. 2013. ISBN: 978-3-642-34332-2. DOI: 10.1007/978-3-642-34333-9.

[2] Kristie Pladson. *Stuttgart unseats Munich as Germany's most expensive city for renters*. 2019. URL: https://www.dw.com/en/stuttgart-unseats-munich-as-germanys-most-expensive-city-for-renters/a-51374468.

[3] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2020.

[4] Jonathan Taylor Sergio Bacallado. *Best subset selection*. 2020. URL: http://web.stanford.edu/class/stats202/notes/Model-selection/Best-subset.html.

[5] B. S. Everitt | A. Skrondal. *The Cambridge Dictionary of Statistics, Fourth Edition*. Cambridge University Press, 2010.

[6] Amt für Wohnen und Migration München. *Mietspiegel für München*. 2021. URL: https://www.muenchen.de/rathaus/Stadtverwaltung/Sozialreferat/Wohnungsamt/Mietspiegel.html.

A. Additional figures

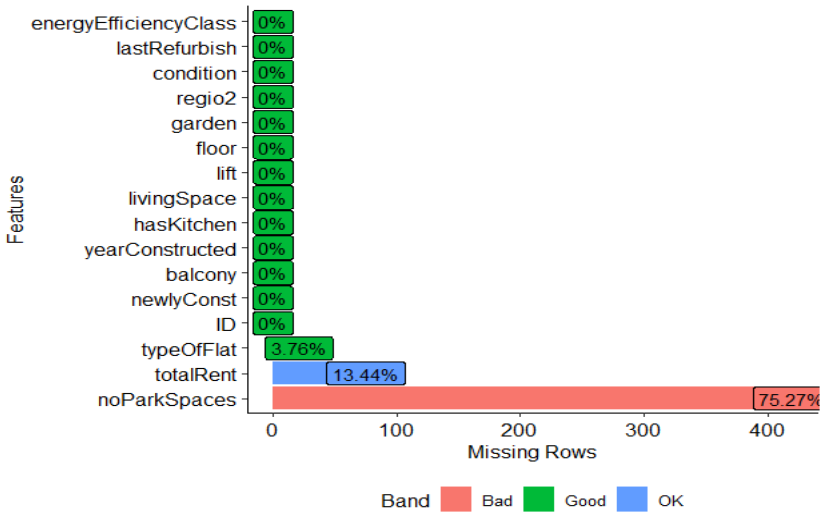


Figure 4: Variables with missing values

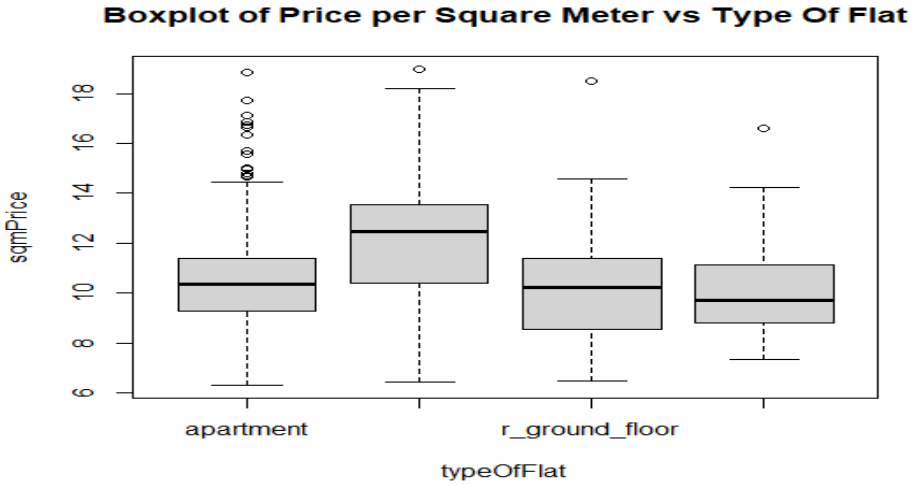


Figure 5: Price per square meter vs Type of flat

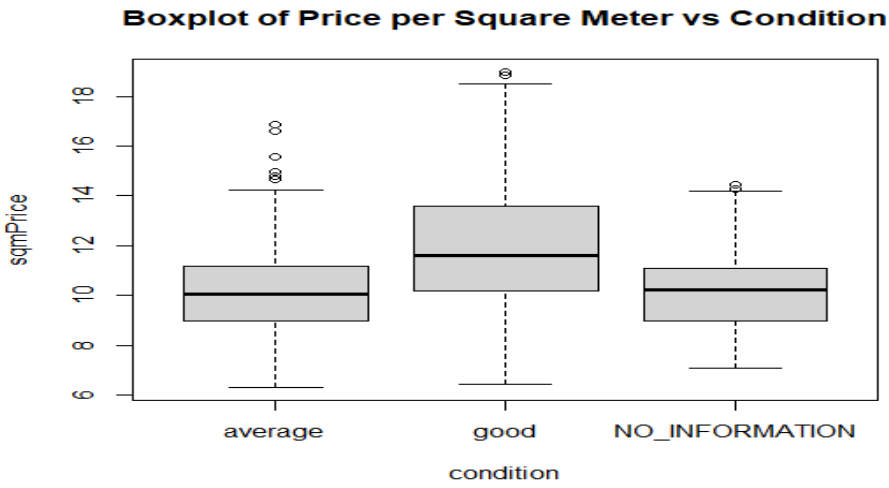


Figure 6: Price per square meter vs Condition

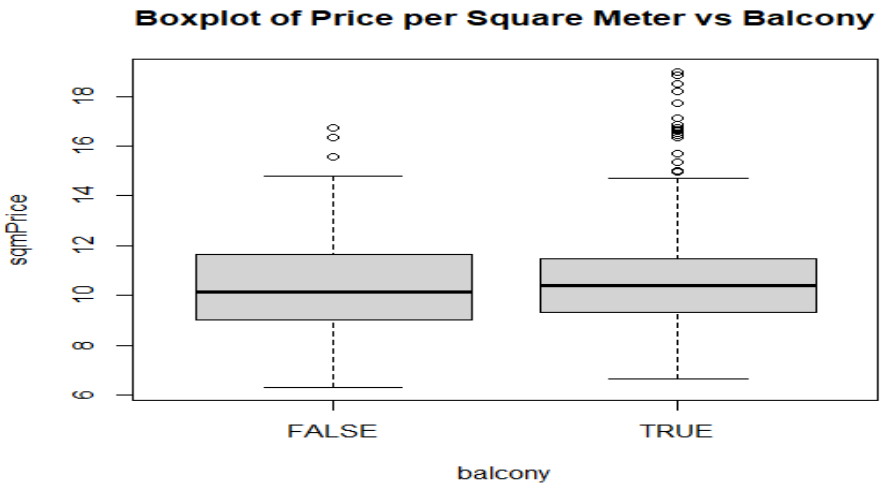


Figure 7: Price per square meter vs balcony

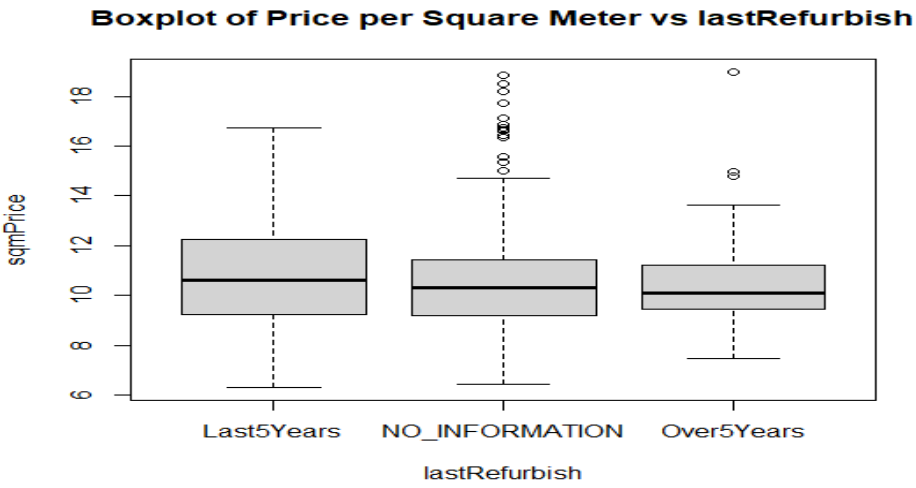


Figure 8: Price per square meter vs Last Refurbished