TU Dortmund

Introductory Case Studies

Project 2: Discrete covariates

Lecturers:

Dr. rer. nat. Maximilian Wechsung

Author: Ankita Sarkar

Group number: 22

Group members: Avisha Anilkumar Bhiryani, Janani Veeraraghavan , Kaushal Tajane, Shivam Shukla, Shubham Khochare

May 27, 2022

Contents

1 Introduction

The rent index is agreed between the landlord and the tenant in contract to give an overall market view of what the rent can be in any particular area in Germany. The analysis of this project primarily involves the net rent, living area and the quality of locations. The goal of the project is to check if the quality of the locations have any impact on the rent per square meter statistically. The possibilities of conducting parametric or nonparametric test are considered and the reasons behind choosing a nonparametric test are discussed in detail. In the next step, the pairwise test between each category of location and rent per square meter are performed to check for the significance. Multiple testing problem is taken into account while implementing the pairwise test.

Section 2 provides the description of the dataset involved. Section 3 gives a brief descriptions of the statistical methods such as normality assumptions check, Kruskal-Wallis test, Mann-Whitney-Wilcoxon test, Bonferroni correction etc which are used in this project. Section 4 presents the test results and discusses the interpretations. The final section summaries the test results and provides an outlook to further investigations.

2 Problem statement

2.1 Description of the Dataset

The rent index in Germany provides predictability and transparency in establishing the rent to the renter and the property owner. The rent index is calculated based on a number of factors one of which is the cost of living. It is especially advantageous to the tenants as the rent mutually agreed upon in contract can remain fixed for a period of time, also there are possibilities to reduce the rent depending on various factors. It typically helps with the market view of the rent of properties among different cities and communities in Germany. Thus the rent index law helps with the comparison of the rent in different living conditions (Fahrmeir et al., 2021, p. 5). This project makes use of a sample dataset containing the rent index of 1999 in Munich(Kneib et al., 1999).

The variables in the dataset are the living area in square meter, construction year from 1918 to 1997, quality of bathroom, kitchen, location and central heating. Three categories of location - 1: average, 2: good and 3: top are considered. A new variable so called rent per square meter is computed using the net rent and the living area. The

main variables of interest are net rent and the quality of locations. The net rent, rent per square meter, living area are numeric variable whereas the quality of location is measured on ordinal scale. The dataset contains the details of 3082 apartments in total out of which 1794 apartments are in average location, 1210 in good and 78 in top location. There are no missing values present in the dataset.

2.2 Project objectives

The purpose of this project is to check if the quality of location has any significant effects on the rent per square meter. To conduct this, normality assumptions of the data are verified first in order to decide which statistical test needs to be used. Normal distribution is checked using QQ normal plot and variance homogeneity using boxplots. The observations are assumed to be independent considering how the data are collected location wise. Based on the results from these checks, a nonparametric test Kruskal-Wallis test is performed on the dataset to check for significance. In the next step, the Kruskal-Wallis test is extended to the Mann-Whitney-Wilcoxon test for the pairwise differences between the locations and the rent per square meter. Lastly, the problem of multiple testing is taken into account and adjusted with the Boneferroni correction.

3 Statistical methods

In this section, various statistical methods and plots are described which are used to analyse the dataset. R programming language (R Development Core Team, 2020) version 4.1.0 is used.

3.1 Q-Q Normal Plots

QQ Normal plot compares the distribution of a sample against a standard normal distribution. To construct a normal probability QQ plot, a sample of n observations $y_1, y_2, ..., y_n$ is sorted in ascending order and is denoted as $y_{(1)}, y_{(2)}, ..., y_{(n)}$. These are referred to as observed quantile. Next, plotting (or probability) i points for $i = 1, .., n$

are computed using the below formula,

$$p_i = \begin{cases} (i - 3/8)/(n + 1/4) & \text{if } i < 10, \\ (i - 3/8)/n & \text{if } n > 10. \end{cases}$$

The plotting points p_i are used to compute theoretical quantile x_i corresponding to each sorted y_i with the formula $P(X < x_i) = p_i$ where $X \sim N(0, 1)$. QQ plots can be used to check for deviation from the normal distribution if present in any sample. We know that the normal curve is symmetric around it's mean. The deviation from normality can be observed with the concave trend of the plotted points on the plot. If the curve of the plotted points looks like a concave up, then it is an indication of existence of heavy right tail in the underlying data distribution or else if the curve looks like concave down, then the sample is likely to be skewed to the left or having heavy left tail.(Hay-Jahans, 2019, p. 146-153)

3.2 Hypotheses testing

The term hypothesis states about something that is supposed to be true. A hypothesis test also known as significance test is the process to decide whether the underlying data supports a particular hypothesis about a population or group. The hypothesis test involves two types of hypotheses - the null hypothesis (H_0) and the alternative hypothesis (H_A). The null hypothesis is a statement about a population parameter such as mean which is assumed to be true. The aim of the hypothesis testing is to check if the null hypothesis can be rejected in favour of the alternative hypothesis. Any hypothesis test involves four steps - state the hypotheses, set a criteria for the statistical decision, calculate the test statistic and finally make a decision. The null hypothesis is expressed as

$$H_0 : \mu = \mu_0,$$

where μ is the population mean and μ_0 is some number. As the hypothesis test aims at deciding if μ differs from μ_0, the alternative hypothesis can be formulated in three different ways. One is known as two-sided test if μ is different from μ_0. Left tailed test is when μ is less than μ_0 and right tailed test is when μ is greater than μ_0

$$H_A : \mu \neq \mu_0 \quad \text{and} \quad H_A : \mu < \mu_0 \quad \text{and} \quad H_A : \mu > \mu_0.$$

A hypothesis test is called one tailed test if it is either left tailed or right tailed. (Hartmann et al., 2018)

To set a criteria for the decision, a threshold known as significance level α is considered which is typically set at 5% or 1% and we also need to calculate a test statistic which provides the idea of how far the sample mean is from the population mean stated in the null hypothesis. The probability of obtaining the sample statistic is known as $p-$value and it is matched with the α value. If the p value is less than or equal to α, under the consideration that the null hypothesis is true, we reject the null hypothesis. When p value is greater than α, under the consideration that the null hypothesis is true, we do not reject the null hypothesis. In other words, if $p \leq \alpha$ reject H_0; otherwise if $p > \alpha$ do not reject H_0. Special note here we never say that H_0 is accepted in later case, we simply state that H_0 is not rejected as there is not sufficient evidence to support H_A.

3.3 One way ANOVA test

When the data comes from more than two populations and the data is quantitative, the Analysis of Variance (ANOVA) is performed. One way ANOVA checks if there is any difference between the populations means. To conduct the ANOVA test, the data has to follow the normal distribution and there is homogeneity of variances in the populations.

3.4 Nonparametric test

The traditional parametric hypothesis tests assume that the underlying data is generated from a well-defined distribution with certain set of assumptions which are often not so realistic. Such distributions are characterized by one or more unknown population parameters such as mean, median, variance, etc. Also the null hypothesis and the test statistic depend on such distribution assumptions. But some times the sample data departs from the particular set of assumptions and we can not be sure which distribution the underlying data is generated from. In such cases the tests are called nonparametric or distribution-free which are more flexible than their parametric counterparts though not powerful enough. (Bonnini et al., 2014, p. 2)

3.5 The Kruskal-Wallis test

The normality assumptions are not always possible to maintain. So one-way ANOVA test can not be applied to test the hypotheses. The Kruskal-Wallis test is used in such distribution-free cases.

Assumptions

- The sample is drawn from more than two populations.
- The observations are independent of each other.

Procedure

The null hypotheses can be formulated as

$$H_0 : \text{The distributions of the populations are same}$$

$$H_A : \text{The distributions of the populations are not same.}$$

For $j = 1, 2, ..., J$ and $i = 1, 2, ..., n_j$ let y_{ij} be the i^{th} observation in the j^{th} sample. The J samples are combined together and sorted in ascending order. Ranks r_{ij} are given to the observations and average ranks are assigned in case of ties. The rank sum for each sample $j = 1, 2, ..., J$ is calculated as below,

$$R_j = \sum_{i=1}^{n_j} r_{ij}$$

and the total number of observations is calculated as $n = \sum_{i=1}^{n_j} n_j$. The Kruskal-Wallis test statistic is given by

$$H^* = \frac{12}{n(n+1)} \sum_{j=1}^{J} (R_j^2/n_j) - 3(n+1).$$

Under the null hypothesis, the test statistic H approximates the χ^2-distribution with $J - 1$ degrees of freedom if either total groups $J = 3$ with total length of samples in each group $n_j \geq 6$, or $J > 3$ with $n_j > 5$. The null hypothesis is rejected if $H^* > \chi^2_{\alpha,j-1}$ obtained from the chi-square distribution table. (Hay-Jahans, 2019, p. 311-314)

3.6 The Wilcoxon(Mann-Whitney) test

Also known as Wilcoxon rank sum test, Mann-Whitney-Wilcoxon test comapares two independent samples while considering ordinal data and also when the population does not come from the normal distribution because of which parametric t-test can not be applied. Let us have two samples $X_1, ...X_{n_1}$ of length n_1 and $Y_1, ...Y_{n_2}$ of length n_2 and total $n = n_1 + n_2$ observations.

Assumptions

- The observations $X_1, ...X_{n_1}$ and $Y_1, ...Y_{n_2}$ each are two independent sample.
- The data is measured on ordinal scale.

Procedure

The null and alternative hypotheses are stated as

$$H_0 : \text{The distribution of group1} = \text{The distribution of group2}$$

$$H_A : \text{The distribution of group1} \neq \text{The distribution of group2}$$

The following description illustrates how the test statistic is computed. First the observations from the two samples are combined together and sorted from smallest value to the largest value. Then ranks are assigned to the sorted observations in the combined sample. If there are any ties present, median of the respective ranks are calculated and assigned to the ordered observations. The ranks in each individual sample are added up. The test statistic is

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \quad \text{and} \quad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

where R_1 and R_2 are the summed ranks in respective groups. Finally the smallest of the U_1 and U_2 are selected as the test statistic U (Witte and Witte, 2017, p. 387-391).

When the sample size becomes large ($n > 20$), the value of U approximates the normal distribution and the p value is matched against the critical value obtained from standard normal distribution with significance level set to α or can be calculated as below (Daniel

and Cross, 2018, p. 604)

$$Z = \frac{U - n_1 n_2/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}.$$

3.7 Multiple testing problem

When there are more than one hypotheses to be tested, the so called multiple testing problem arises. When we reject the null hypothsis even if it is true, the phenomenon is known as the Type I error. When a number of hypotheses are tested, the chance of Type I error occurring increases. The cost of rejecting the true null hypothesis is always compared with accepting the false null hypotheses (Shaffer, 1995, p. 562).

Let's suppose there are n hypotheses $H_1, ..., H_n$ with associated test statistics $T_1, ..., T_n$ and p-values $P_1, ...P_n$. The classical Bonferonni correction related to the test of overall hypothesis $H_0 = \cap(H_i : i = 1, ..., n)$ is at significance level α: sort the p-values from smallest to largest value. Reject H_0, if $P_{(1)} \leq \alpha/n$ where $P_{(1)}$ is the smallest p-value.(Hommel, 1988, p. 383). So the significance level α is adjusted to a lower value in this way to reduce the possibility of rejecting a true null hypothesis. Thus it prevents all test results from being significant when some are not significant.

4 Statistical analysis

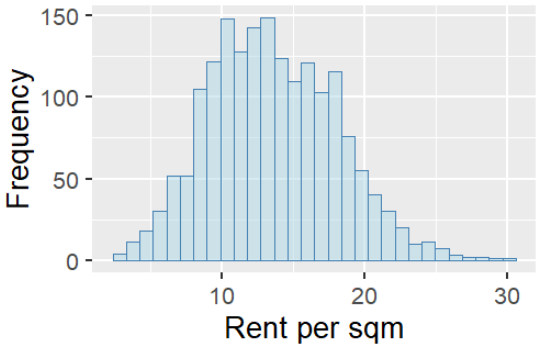
This section discusses the assumptions and the statistical tests performed on the given data set. It includes the interpretation of the results as well.

4.1 Normality assumptions check of the data distribution

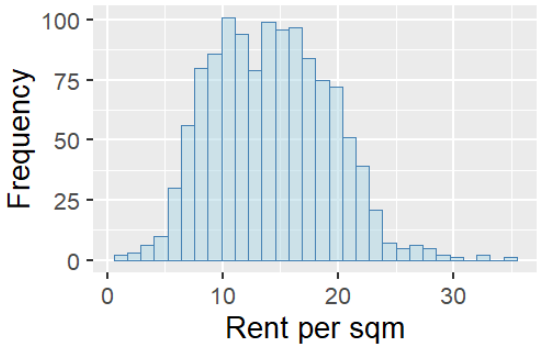
The dataset is analysed with respect to the variable rent per square meter computed from net rent and living area. Table 1 summaries the central tendencies of rent per square meter for the three types of locations. We observe that the mean rent per square meter is highest for the top location and lowest for the average location. But the good location has the maximum range for the rent per square meter out of three types of location.

Table 1: Summary table for the rent per square meter in different locations

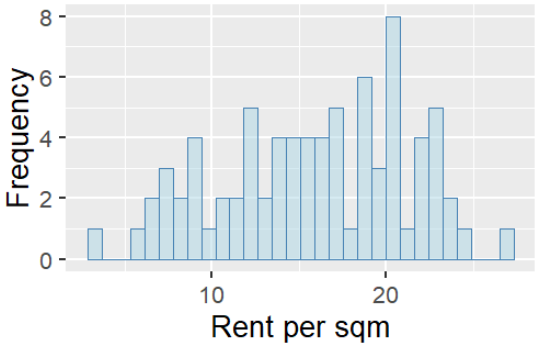
Location	Count	Min.	Q_1	Median	Mean	Q_3	Max.
Avg.	1794	2.76	10.24	13.26	13.56	16.68	30.08
Good	1210	0.81	10.22	14.10	14.18	17.81	34.56
Top	78	3.64	12.08	16.18	15.94	20.26	27.29



(a) Avg. location



(b) Good location

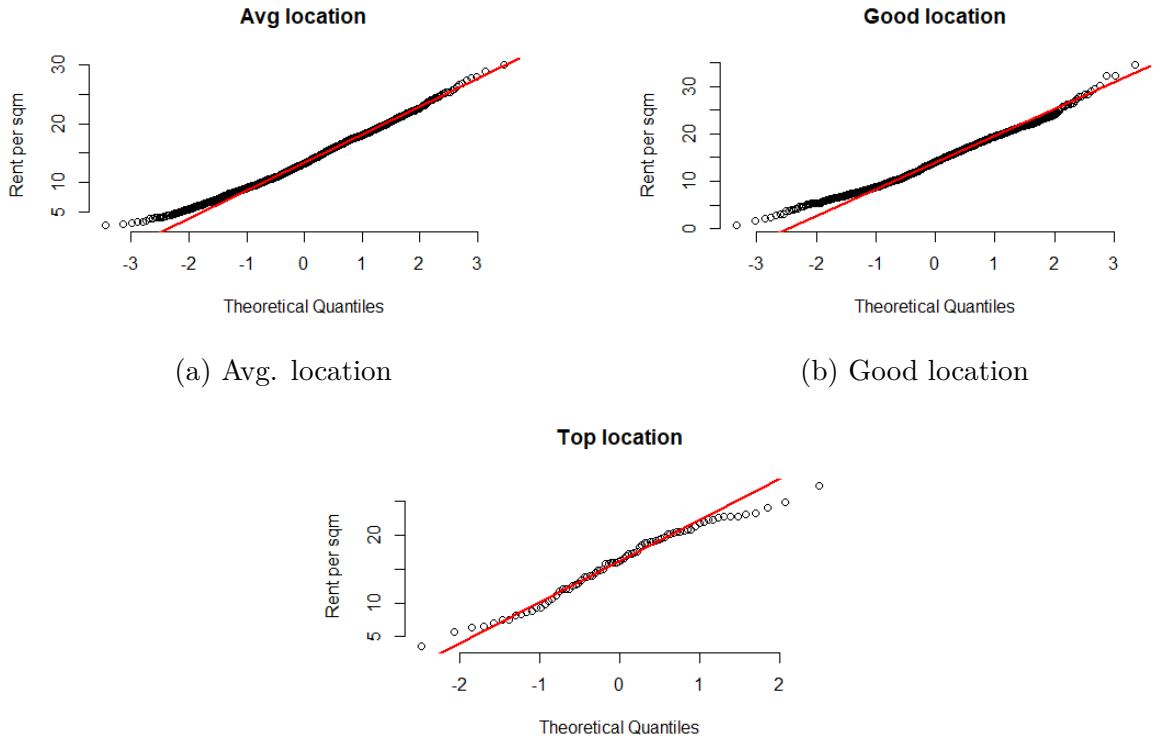


(c) Top location

Figure 1: Histogram for the rent per sqm in a) Avg. location, b) Good location and c) Top location.

Figure 1 displays the histogram distribution for each of the three locations. We can not make much conclusion just by looking at the histogram distribution though they shows us the general shape of the underlying data distributions. We check if the data distribution follows the normal distribution using the QQ plot. Figure 2 shows the QQ Normal plot for three of the locations. The upward concave shape of the graph for the rent per square meter in average and good location indicates heavy right tail in the data distribution whereas the graph for the rent per square meter in top location is both

concave up and down. Most notably all the plots deviate from the reference line at both ends which is the indication that the data distribution is not normal distribution.



(a) Avg. location

(b) Good location

(c) Top location

Figure 2: QQ plot for rent per sqm in a) Avg. location, b) Good location and c) Top location.

The homogeneity of variance for the rent per square meter is verified with boxplots in Figure 3. From the visual inspection of the IQR in each of the three boxplots, it can be said that the assumptions of the homogeneity of variance among the three groups are violated. We can also observe some extreme values in case of average and good locations though those do not add much value to the current analysis.

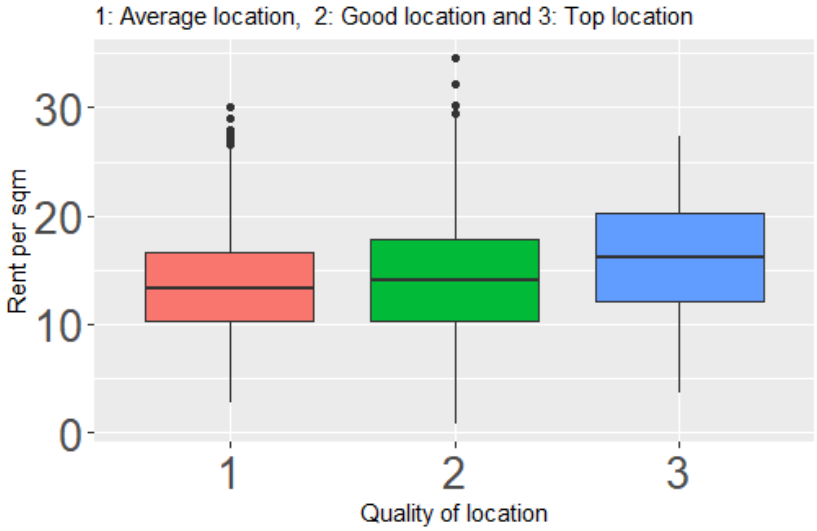


Figure 3: Comparison of the rent per sqm in each of the three locations

4.2 Nonparametric test - Kruskal-Wallis test

As the assumptions for the normal distribution are not maintained seen from Section 3, a general distribution-free approach is taken for the dataset to check whether the locations have significant effects on the rent per square meter. A nonparametric method Kruskal-Wallis test is used to check the significance. There are more than three groups present in the dataset. The quality of location is ordinal and rent per square meter is continuous. The observations are independent of each other. The variances among the groups are unequal and the groups are not of equal length. All of these assumptions indicate that the data do not follow the normal distribution and Kruskal-Wallis test can safely be employed. The null hypothesis is that all groups have the same distribution shape and the alternative hypothesis is that at least one group comes from a different distribution

$$H_0 : \text{The distributions of the three populations are same}$$

$$H_A : \text{At least one distribution of a population is different.}$$

The significance level α is set to 0.05. The Kruskal-Wallis test is performed on the dataset after confirming that the assumptions are met. The test results are summarized in the Table 2.

Table 2: Summary table for the Kruskal-Wallis test result

No of groups	Degrees of freedom	χ^2 test value	$\chi^2_{0.05,2}$	p-value	Reject H_0
3	2	23.451	5.99	<0.01	Yes

The p-value obtained from the test statistic is less than 0.01. Because the number of observations in each group are more than 6 (please see Table 1), the χ^2 test value is considered. It is greater than the $\chi^2_{0.05,2}$ critical value which is 5.99. So we reject the null hypothesis. Hence we conclude that at least one group comes from a different distribution. Thus the quality of locations have significant effects on the rent per square meter.

4.3 Pairwise differences between the rent per square meter and locations

The differences between the rent per square meter in any two different locations are measured pairwise using the Mann-Whitney-Wilcoxon test which is an extension of the Kruskal-Wallis test. First the assumptions are checked to conduct the test. It is already observed that each of the three locations are independent of each other and the variable of interest which is the rent per square meter is continuous and locations are measured on ordinal scale. After the assumptions are fulfilled, the null and alternative hypotheses are formulated as

H_0 : The distribution of population 1 = The distribution of population 2

H_A : The distribution of population 1 $\neq$ The distribution of population 2

H_0 : The distribution of population 1 = The distribution of population 3

H_A : The distribution of population 1 $\neq$ The distribution of population 3

H_0 : The distribution of population 2 = The distribution of population 3

H_A : The distribution of population 2 $\neq$ The distribution of population 3

where the population 1 is of the rent per square meter in average location, the population 2 in good location and the population 3 in top location. The significance level α is set to 0.05. The Mann-Whitney-Wilcoxon test is performed next. As there are testings of multiple null hypotheses involved, it increases the chance of Type I error, so Bonferonni correction is included to adjust the p-value obtained from the test statistic. The test results are summarized in Table 3 which compares the p-values before and after adjustment. The decisions of whether or not rejecting the null hypothesis are listed as well.

Table 3: Summary table for the Mann-Whitney-Wilcoxon test results

No	Group 1	Group 2	p-value	Reject H_0	Adj. p-value	Reject H_0
1	Avg.	Good	0.0021	Yes	0.0062	Yes
2	Avg.	Top	<0.0001	Yes	0.0001	Yes
3	Good	Top	0.0025	Yes	0.0074	Yes

According to the Bonferroni correction, the Adj. p-value is calculated as p-value $\times\, n$ where $n = 3$ tests. From the Table 3, we see that all the tests are deemed as significant as both the p-values and the adjusted p-values are lower than the significance level 0.05. So we reject the null hypotheses and conclude that the distribution of each location be it average, good or top is different from the others.

5 Summary

For the purpose of this project, a dataset containing the rent indices for the year of 1999 of Munich, Germany were used. The variables considered mainly were rent per square meter (net rent / living area) and the quality of locations - average, good and top. The aim of this project was to check if the quality of location had any significant effect on the rent per square meter. The underlying data distributions did not follow the normal distribution. QQ normal plots were used to show the deviation from the normality in the distribution. Also the assumptions of the homogeneity of variances were violated which was shown with boxplots comparison among the different locations. So one-way ANOVA could not be employed instead nonparametric Kruskal-Wallis test was conducted. The assumptions of the Kruskal-Wallis test were verified for the dataset. The assumption of independence were assumed depending on the rent in different locations.

Upon meeting the criterion, the null hypothesis was formulated as the distributions of the populations were same and the alternative hypothesis was set up as at least one population distribution was different. Next the test was performed and the p-value was checked. We could reject the null hypothesis and thus conclude that the distributions of rent per square meter for three of the locations were of different shape. But the Kruskal-Wallis test does not tell us exactly which population has different distribution. So the test was extended to the Mann-Whitney-Wilcoxon test which is a two-sample test and gives pairwise differences in populations. The assumptions for the Mann-Whitney-Wilcoxon test including independence of the samples, ordinal and continuous variables were satisfied in the dataset being considered in this project. So the test was performed by formulating the null and alternative hypotheses. After the test, the p-value was matched with z-value obtained from Z-table (because of large sample). Three set of null hypotheses were tested in this test - the distributions of any two pairs of the populations were same. Hence the problem of multiple testing may arise which falsely rejects the true null hypothesis. So Bonferroni correction was applied to the p-values obtained from the pairwise test to adjust them. Thus type I error were controlled. The null hypothses were rejected in both before and after the correction making all the tests as significant. So we concluded that each of the populations were generated from different distribution. The significance level was set to 0.05 in both the tests.

Though nonparametric tests employed in this project are flexible, we can better use the parametric approach which are more powerful. The unequal variances can be dealt with the transformation of the variables to make the variances homogeneous. Also the relationship between the rent per square meter and other variables such as living area can be considered for further analysis. More information can be included to the analysis if we consider the interaction between the independent variables.

Bibliography

Bonnini, S., Corain, L., Marozzi, M., and Salmaso, L. (2014). *Nonparametric hypothesis testing: rank and permutation methods with applications in R*. John Wiley & Sons.

Daniel, W. W. and Cross, C. L. (2018). *Biostatistics: a foundation for analysis in the health sciences*. John Wiley & Sons.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. D. (2021). *Regression - models, methods and applications*. Springer.

Hartmann, K., Krois, J., and Waske, B. (2018). *E-learning project SOGA: Statistics and geospatial data analysis*. Department of Earth Sciences, Freie Universitaet Berlin.

Hay-Jahans, C. (2019). *R Companion to Elementary Applied Statistics*. Chapman and Hall/CRC.

Hommel, G. (1988). *A stagewise rejective multiple test procedure based on a modified Bonferroni test - Biometrika*. Oxford University Press.

Kneib, T. et al. (1999). Munich rent index 99. URL: `https://www.uni-goettingen.de/de/551218.html/`.

R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shaffer, J. P. (1995). Multiple hypothesis testing.

Witte, R. S. and Witte, J. S. (2017). *Statistics*. John Wiley & Sons.