TU Dortmund

Introductory Case Studies

Project 2: Discrete Covariates

Lecturers:

Dr. rer. nat. Maximilian Wechsung

M. Sc. Hendrik Dohme

Author: Akshay Choudhary(224437)

Group number: 16

Group members: Alina Imtiaz, Vaibhav Grover, Vipul Chauhan, Nidhi Kiritbhai Patel

May 27, 2022

Contents

1 Introduction

Renting a property is always tricky as everyone has different requirements and opinions about renting the property. People consider many factors before renting the property, such as locality, transportation, and size of the apartment. According to the mentioned factors above, the city housing department also provides information about the registered property. The property price rises from standard to premium housing; for example, if someone wants private parking, they have to pay extra. Similarly, we also have to pay extra money for a premium kitchen and bathroom. In this project, the analysis through statistical inference on randomly collected data is performed, which contains the net rent per square meter according to the three locations: Average, Good and Top.

First, a decision has to be made whether to use a parametric (One way of anlysis of variance (ANOVA)) or non-parametric (Kruskal-Wallis test) method. After the successfull checking of the assumptions of ANOVA a decision has been made to use Kruskal-Wallis test. The assumptions must be met to perform Kruskal-Wallis test. The Kruskal-Wallis test results conducted on the data set rejects the null hypothesis, which suggests a significant difference in the net rent per square meter of different location groups.

Second, multiple testing is used to determine the rank sum of all properties in location pairings. After completing multiple testing using the Mann-Whitney U test, all the location groups might reject the null hypothesis, but additional testing increases the risk of a Type I error. As a result, the Bonferroni correction method is used to reduce this by either changing the value of significance level or by adjusting the p-value.

The data set is described in detail in Section 2 and the objectives of the report are also briefly mentioned. Section 3 explains the statistical measures used in the task analysis. Section 4 reflects the implementation of the approaches using the methods presented in Section 3. Finally, the Section 5 outlines the critical finding and address further studies that may be conducted on this dataset.

2 Problem statement

The following data set is provided by the professors of the course Introductory Case Studies at TU Dortmund University for the summer semester 2022 as a questionnaire. The data set is originally extract from rent index data for Munich, which was collected in

1999. The data contains net rent, six covariates, and 3082 observations, with no missing values.

2.1 Dataset description

The dataset contains the net rent and some attributes of the Munich apartments. There are seven columns in the dataset: net rent, living area, construction year, bathroom, kitchen, location quality, and central heating. The living area is calculated in square meters. For the analysis of the given task, a new variable rent per square meter is introduced, by dividing the net rent with the living area and the scale is numerically continuous. This newly introduced variable and the quality of the location are used for this project. The net rent of all the properties is measured according to the square per meter in Euros. The bathroom, kitchen, location quality, and central heating covariates are categorical variables, and construction year is a continuous variable.

2.2 Project objectives

The primary objective of this project is to infer the net rent per square meter of all the apartments from different locations which is achieved through statistical inference models. In this project, first, a decision has to be made to select a method that analyzes the significant effect of the rent per square meter on location quality. For this, either parametric(ANOVA) or non-parametric (rank-based) method is considered. The parametric method(ANOVA) required assumptions is checked and if it does not follow the assumptions then the non-parametric method(Kruskal-Wallis) test is performed. For the interpretation of tasks in this project, two hypotheses are assumed. One is the null hypothesis which indicates that the distribution of all the groups is the same. The other is the alternative hypothesis which is the complement of null hypothesis. The Kruskal-Wallis test is performed in initial task after meeting all the required assumptions, which evaluate the significant effect of the rent per square meter on location quality. The Kruskal-Wallis test rejects the null hypothesis if the p-value is smaller than the significance level (0.05).

Secondly, we compares the two groups of locations to see any pairwise difference in the net rent per square meter of apartments. It can be accomplished using a Mann-Whitney U test, in which the test statistic is compared to the critical value of the z-distribution. Suppose the value is less than the critical value. In that case, the null hypothesis is rejected, indicating that the hypothesis is statistically significant and there is no difference between the two group's distribution. As the number of comparisons grows, the likelihood of Type I error grows. The p-value is corrected by using the Bonferroni correction procedure. Furthermore, we may examine statistical differences using no adjustment and Bonferroni adjustment.

3 Statistical methods

In this report, we are using statistical inference methods. Statistical inference is inferring the distribution generated for a sample from a population.

3.1 Statistical hypothesis testing

A statistical hypothesis is a declaration of faith about a population parameter. This assumption may or may not be valid. The statement that the population means is ten is an example of a statistical hypothesis. A statistical experiment can be used to test the validity of this hypothesis. Statistical testing determines whether to accept or reject a hypothesis for an unknown parameter based on the distribution of a random variable. With the premise of random sampling, hypothesis testing is concerned with determining whether a parameter θ belongs to one subset of the parameter space or its counterpart (Levin, 2011, p.105).

3.1.1 Null/ alternative hypothesis

Consider a statistical situation with an unknown parameter whose value must lie within a defined parameter space Ω. Assume that Ω can is partitioned into two disjoint subsets, Ω_0 and Ω_1 and that the statistician wants to know if θ resides in Ω_0 or Ω_1. Here, H_0 denotes the null hypothesis and H_1 denotes the alternative hypothesis. If the result of the test for θ lies in Ω_1 then reject H_0. If not then do not reject H_0 as if θ lies in Ω_0(DeGroot and Schervish, 2012, p.531). Here,

$$H_0 : \theta \epsilon \ \Omega_0$$

$$H_1 : \theta \epsilon \ \Omega_1.$$

One-tailored and Two-tailored Hypothesis

Assume that θ is a single-dimensional parameter. Hypothesis of the following form:

Null Hypothesis: $H_0 : \theta \geq \theta_0$ or $H_0 : \theta \leq \theta_0$;

Alternative Hypothesis: $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$ is one-tailored or one-sided hypothesis whereas Hypothesis of the following form:

Null Hypothesis: H_0: $\theta = \theta_0$;

Alternative Hypothesis: H_1: $\theta \neq \theta_0$ is two-tailored or two-sided. For this project, we are using Two sided hypothesis (Levin, 2011, p.181).

3.2 Test statistics and critical region

Assume testing the following hypothesis:

H_0: $\theta \epsilon \Omega_0$, and H_1: $\theta \epsilon \Omega_1$.

Before determining which hypothesis to choose, consider a random sample $X = (X_1, ..., X_n)$ drawn from a distribution containing the unknown parameter. S stands for the sample space of the n-dimensional random vector $\boldsymbol{X}$. As a result, S is the set of all potential random sample values. Partitioning the sample space S into two subsets, with the values of $\boldsymbol{X}$ for rejecting H_0 in one subset which is S_1, and the values of $\boldsymbol{X}$ for not rejecting H_0 in the other subset S_0, describes a test technique. The test's critical area is defined by the set S_1 mentioned earlier,

$$S_1 = \{x : r(x) \, \epsilon R\}.$$

The critical region is commonly expressed as a statistic, such as T= r(X), where T denotes the test statistic and R denotes the rejection area of the test, with R being a subset of the real line.

For a test which is based on a test statistic T, the rejection area is usually a specified interval or the outside of a specified interval. If the test rejects H_0 when $T \geq c$, for example, the rejection area is the interval $[c, \infty)$.

The term parameter space and sample space are related, but they are not the similar. The parameter space and sample space are generally of different dimensions, therefore Ω_0 will invariably differ from S_0 (DeGroot and Schervish, 2012, p.533).

3.2.1 Power function

Let's assume that δ is a test process. The function $\pi(\theta|\delta)$ is known as the test's power function. The power function $\pi(\theta|\delta)$ is given by the relation $\pi(\theta|\delta) = Pr(X \in S_1|\theta)$ for $\theta \in \Omega$ if S_1 indicates the critical area of δ.

The power function for Ω is $\pi(\theta|\delta) = Pr(T \in R|\theta)$ for $\theta \in \Omega$ if it is specified in terms of a test statistic T and rejection region R (DeGroot and Schervish, 2012, p.534).

3.2.2 Type I/ II error

A Type I error, also known as a first-sort error, arises when a reasonable null hypothesis is rejected based on an inaccurate judgment. A Type II error, also known as a second-sort error, is a wrong decision not to reject a false null hypothesis.

For the above test procedure, if $\theta \in \Omega_0$, $\pi(\theta|\delta)$ is the likelihood that there will be a chance of commiting a Type I error, according to the power function. If $\theta \in \Omega_1$, 1-$\pi(\theta|\delta)$ is the chance of committing a Type II error (DeGroot and Schervish, 2012, p.535).

3.2.3 p-value and significance level (α)

It is the chance of seeing a test statistic value as extreme (or more extreme) than what was seen (under H_0). It has a value of 0 to 1. Whether the null hypothesis should be accepted or rejected, the p-value is always compared to the significance level (α). The level of significance α, which runs from 0 to 1, must be set before beginning the experiment and cannot be adjusted to reach the desired hypothesis. The lower the value of α, the more significant it is and the more comparable. The null hypothesis is rejected if the p-value is smaller than α, indicating a significant difference between specific demographic features. If the p-value is more than α, we will not be able to reject the null hypothesis, and there will be no difference in the population samples. The smaller the p-value, the more evidence that the null hypothesis should be rejected.

A statistically significant p-value is less than or equal to 0.05. It provides significant evidence against the null hypothesis, as it has a less than 5% chance of being correct. As a result, the null hypothesis is rejected, and the alternative hypothesis is accepted. The null hypothesis is rejected if the p-value is less than the significance threshold (usually $p<0.05$). However, this does not guarantee that the alternative hypothesis has a 95% chance of being true (Levin, 2011, p.187).

3.3 One-way analysis of variance (ANOVA)

The one-way analysis of variance (ANOVA) test identifies statistically significant evidence that the mean of more than two groups is significantly different. Still, it does not tell which group is statistically distinct from the others. If the group means are the same, the null hypothesis is not rejected; but, if any group mean differs from the others, the null hypothesis is rejected.

The null hypothesis, which states:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_p.$$

where µ = group mean and p = number of groups (Hay-Jahans, 2019, p.271)

There are three ANOVA assumptions:

1. Homogeneity of variance: Population variance in each group is same.
2. The one-way ANOVA works only for the groups which will satisfied the normality assumptions. If the sample sizes are large enough then the normality assumption is not needed.
3. The variables are i.i.d (independent and identically distributed).

For random variables $Y_1, Y_2, ...Y_p$, p is the population, where $j = 1, 2, ..., p$ and n_i is the observed response, where $i = 1, 2, ...n_j$, n_j denotes the sample size from the j^{th} population. So y_{ij} is the i^{th} observed response from the j^{th} population.

Let $\bar{y_j}$ be the sample mean of j^{th} sample and s_j^2 be the j^{th} sample variance.

The test statistic used to determine if the p means are all equal is:

$$F^* = \frac{\sum_{j=1}^{p} n_j(\bar{y_j} - \bar{y})^2/(p-1)}{\sum_{j=1}^{p}(n_j - 1)(s_j^2)/(p-1)} \sim F(p-1, n-p)$$

.

and p-value = $P(F \geq F^*)$.

The null hypothesis is rejected if the p-value of detecting these F-statistics under the null hypothesis is sufficiently low (i.e., less than significance level 0.05) (Hay-Jahans, 2019, p.271-272).

Degree of freedom: The number of independent sample points used to compute a statistic minus the number of parameters inferred from the sample points is referred to as degrees of freedom (df) (Hay-Jahans, 2019, p.270).

Table 1: The degrees of freedom for each sum of squares.

Sum of squares	Degrees of freedom
Between-groups	p-1
Within-groups	n-p
Total	n-1

3.4 Kruskal-Wallis test

The Kruskal-Wallis test also known as the "H-test", is a non-parametric test that works similarly as ANOVA and it is used when one of the ANOVA assumptions do not fulfill. It is performed on the independent groups from the population and test if their distributions differ. So, the hypothesis for the test are:

H_0: Distribution of independent groups are not different.

H_1: At least one of the independent group distribution is different.

There are three assumptions to perform Kruskal-Wallis test:

1. The dependent variable is at least ordinal or continuous scaled.
2. The independent variable must consist two or more categorical independent groups.
3. Collected observations must be random and independent from each other.

Generally, Mann-Whitney U test is performed for two independent groups. For more than two groups we use Kruskal-Wallis test.

In this test the actual value of the observations are not used. To test this method, ranks are defined to all groups in an order. For example, the smallest value gets the rank 1 and the largest value gets the rank n as there are total n number of observations.
The following steps are performed to calculate test statistic value:
1. Assign the rank to all the n observations.
2. Calculate the mean rank sum of each group: $\bar{R_k}$ where k is number of groups.
3. If there is no difference between the groups then the expected value of the ranks for each group is: $E_R = \frac{n+1}{2}$. The variance is calculated as: $\sigma_R^2 = \frac{n^2-1}{12}$.
4. The test statistic value H:

$$H = \frac{n-1}{n} \frac{\sum_{i=1}^{k} n_i(\bar{R}_i - E_R)}{\sigma_R^2}$$

To get the p-value from the calculated H statistic value, it is approximated with the χ^2 distribution with $k-1$ degree of freedom. We reject the null hypothesis if the p-value is less than the level of significance (0.05), which implicates that the group distributions differ from each other (Corder and Foreman, 2011, p.100).

3.5 Mann-Whitney U test

Mann-Whitney U test is also a non-parametric method which checks whether or not the two independent groups differ from each other. This is equivalent to the parametric t-test for independent samples. The assumptions of this test are also similar to the Kruskal-Wallis test, such as the observations are random and independent, the dependent variable must be ordinal or continuous, and the independent variable must have two categories.

So, the hypothesis for the test are:
H_0: The distribution of the two independent groups does not differ.

H_1: The distribution of the two independent groups differs.

Let two groups have random observations such as $X_1, X_2, ...X_{n_1}$ and $Y_1, Y_2, ...Y_{n_2}$. Here n_1 is the total number of observations in X group and n_2 is the total number of observations in Y group. The rank is given to all the observations from both groups in an order and rank sum $\sum R_1$ and $\sum R_2$ is calculated for both groups respectively. So, the test statistic for U is defined as:

$$U_i = n_1 n_2, + \frac{n_i(n_i+1)}{2} - \sum R_i.$$

where U_i is the test statistic of interested sample. n_i is the number of values from the interested sample, and $\sum R_i$ is the sum of ranks from the interested sample.

To calculate the significance of the test statistic U_i value is check via the z-standardizing which is defined as:

$$z = \frac{U_i - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}.$$

This z is the z-score for a normal approximation of the data and U¡ is the U statistic from the interested sample. This z-score is now checked for significance by comparing it to the critical value of the standard normal distribution for significance level 0.05. If the absolute value of the test statistics is higher than the critical value, the null hypothesis rejected (Corder and Foreman, 2011, p.58).

3.6 Bonferroni correction method

Multiple testing is simultaneous testing of more than one hypothesis at a single instance. For example, we want to know which coefficients do not affect the model in the multiple regression model. If multiple testing is not taken into account, then the probability of rejecting the true null hypothesis is very high. For instance, if 100 hypotheses are tested simultaneously with a given level of significance of 0.05, we assume that out of 100, 5 true null hypotheses are rejected. Assume that all the tests are independent of each other, then the probability that at least one true null hypothesis will be rejected is given by $1 - 0.95^{100} = 0.994$. When doing multiple hypothesis tests, the risk of producing one or more false discoveries, or Type I error, is known as the family-wise error rate (FWER) in statistics.

The methods which controls this Type I error or FWER is known as Bonferroni correction method. Bonferroni adjusts the level of significance α to a more rigorous value, reducing the chances of Type I Error. The α value is adjusted by dividing it by number of tests and get the adjusted α value. When the α value decreases after correction, the chances of obtaining at least one significant result decreases. Another way to adjust the α value is by multiplying the raw p-values with the number of tests i.e. n. In this multiple pairwise testing, we will compare the adjusted p-values to the raw p-values (Gordon and Glazko, 2007).

$$\alpha_{adjusted} = \frac{\alpha}{n} \quad \text{or} \quad p_{adjusted} = p \cdot n \quad \text{and} \quad FWER = 1 - (1 - \alpha)^n$$

.

- α is the significance level and p is the raw p-value .
- $\alpha_{adjusted}, p_{adjusted}$ are the adjusted α and p-values (Hay-Jahans, 2019, p.274).

4 Statistical analysis

In this section, The statistical methods explained above are applied to the provided data, and the results of the analysis are interpreted. For this report, the significance level α value is 0.05. For all calculations and visualizations the statistical software R (R Development Core Team, 2020), with packages for plots ggplot2 (Wickham, 2016), ggpubr (Kassambara, 2020) and, dplyr (Wickham and François, 2021) is used for analysis.

4.1 Descriptive analysis

First, a descriptive analysis is performed to give a general overview of the data collection. There are 3082 observations in the dataset. Table 2 comprises the values of count, mean, median, standard deviation (SD) and, variance (Var) of all locations.

Table 2: Table containing description of the data set.

Location	Count	Mean	Median	SD	Var
Average	1794	13.6	13.3	4.45	19.8
Good	1210	14.2	14.1	5.08	25.8
Top	78	15.9	16.2	5.38	29.0

4.2 Assumptions

First the ANOVA test assumptions were checked.
1. Homogeneity of variance: To perform the ANOVA test, variances of the distribution of the groups are equal. So, we assume σ_1^2, σ_2^2...., σ_k^2 are equal. Box plots can be used to understand the variability of the data, by examining the vertical size of each group. By observing Table 2 and Appendix Figure 1, we conclude that variance is not same for all the groups.

As the assumption of the ANOVA test does not satisfy, this test could not be performed on the given data set.

Now we check the assumptions for the Kruskal-Wallis test and Mann-Whitney U test.

1. There are two variables that we need to consider. One is net rent per square meter which is a continuous scaled variable, and second, location which has three categories.

So the assumption that the dependent variable should be continues, and the independent variable should be categorical is satisfied.

2. The data set is divided into a total of three categories i.e., Average location, Good location, and Top location. They are independent categories as there is no relation between them, satisfying the second assumption.

3. The data was acquired randomly, so we can assume that the assumption about the observations that the collected data is random. Since, the net rent of one apartment does not influence the net rent of another apartment, thus we assume it is independent of each other.

As all the required assumptions hold true, the Kruskal-Wallis test as well as Mann-Whitney U test can be performed.

4.3 Kruskal-Wallis test

In this subsection, the Kruskal-Wallis test is performed on the given dataset, and results are interpreted. The Kruskal-Wallis test checks whether or not the quality of the location has a significant effect on the rent per square meter. As mentioned in the Section 3 the null hypothesis and the alternative hypothesis for the Kruskal-Wallis test defined as:
H_0: The quality of the location does not have a significant effect on the rent per square meter.

H_1: The quality of the location has a significant effect on the rent per square meter.

The Table 3 shows the summary of the Kruskal-Wallis test performed in R on the given dataset. From the table, it is clear that the p-value for this test is very less than the level of significance (0.05). This implies that there is strong evidence against the null hypothesis and suggests that the quality of the location has a significant effect on the rent per square meter.

Table 3: Summary of Kruskal-Wallis test.

chi-squared	df	p-value
23.451	2	8.087e-06

4.4 Multiple testing

According to the Kruskal-Wallis test, the net rent per square meter differs significantly as per the quality of location. The Mann-Whitney U will be used to determine which group (or groups) are different from each other in order to explain the pairwise difference. The hypothesis for this test:

H_0: There is no pairwise difference between the rent per square meter for different locations.

H_1: There is a significant pairwise difference between the rent per square meter for different locations.

Table 4: Result of pairwise Mann-Whitney U test without and with Bonferroni adjustment with p-values.

Group 1	Group 2	p-value	p-value adjusted
Average	Good	0.002	0.006
Average	Top	0.00003	0.00010
Good	Top	0.002	0.007

Table 4 shows the result of the pairwise Mann-Whitney U test without and with Bonferroni adjustment. First, we observed the output of p-values without adjustment. According to the result in Table 4, there is strong evidence against all the paired groups because the p-value of all grouped pairs is less than the level of significance (0.05). This result suggests that we have to reject the null hypothesis for all group pairs. This explains that there is a significant pairwise difference between the rent per square meter for different location.

We have performed numerous tests which increase the chances of producing the Type I error as mentioned in Section 3. To prevent this as much as possible, Bonferroni correction method was considered. The adjusted p-value is also described in Table 4. The adjusted p-value increased three times more than the without adjusted p-value as there is a total of three pair of groups. Even after the adjustment there is sufficient evidence to reject the null hypothesis as all the paired p-values are less than the level of significance (0.05).

5 Summary

The rent index data for Munich, which was collected in the year 1999 is used to perform the analysis for this report. This data set contains a total of 3082 observations and net rent of the apartments with six covariates. There is no missing value in the data set. A new variable net rent per square meter was introduced by dividing the net rent with the square meter covariate. This newly introduced variable and quality of the location covariate were used for this report tasks.

This report intends to find the significance of the quality of the location over the net rent per square meter in the Munich area. The Kruskal-Wallis test was performed on the data set because the data set fails to fulfil all the assumptions of ANOVA test. The result of this test implicates that there is enough evidence to reject the null hypothesis because the p-value is very less than the level of significance (0.05), i.e., quality of the location has a significant effect on the net rent per square meter.

To better understand which pair of quality locations have significant influence over the net rent per square meter, the Mann-Whitney U test was performed. The result of this test provided the evidence against the null hypothesis and all location group pairs have less p-value than level of significance (0.05). Since there were multiple tests conducted, the chances of Type I error increased, and to control this we used Bonferroni correction method. Even after performing the Bonferroni correction method, there is no change in the end result and there is not sufficient evidence not to reject the null hypothesis. This implies that all the pairwise location have the significant effect on the net rent per square meter.

Hence, we can conclude that the net rent per square meter of the apartments in Munich differs according to the location. The further analysis can be more interesting if other attribute could also be considered to observe the influence of them on the net rent.

Bibliography

G.W. Corder and D.I. Foreman. *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. Wiley, 2011. ISBN 9781118211250.

M.H. DeGroot and M.J. Schervish. *Probability and Statistics*. Addison-Wesley, 2012. ISBN 9780321500465.

Alexander Gordon and Galina Glazko. Control of the mean number of false discoveries, bonferroni and stability of multiple testing. *The Annals of Applied Statistics*, 1(1): 179–190, 2007. ISSN 1932-6157. URL `https://www.jstor.org/stable/4537427`.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. Taylor and Francis Group, 2019. ISBN 9781138329164.

Alboukadel Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*, 2020. URL `https://CRAN.R-project.org/package=ggpubr`. R package version 0.4.0.

R.I. Levin. *Statistics for Management*. Pearson Education, 2011. ISBN 9788177585841.

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

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

Hadley Wickham and Romain François. *dplyr: A Grammar of Data Manipulation*, 2021. URL `https://dplyr.tidyverse.org/`. R package version 0.4.0.

Appendix

A Additional figures

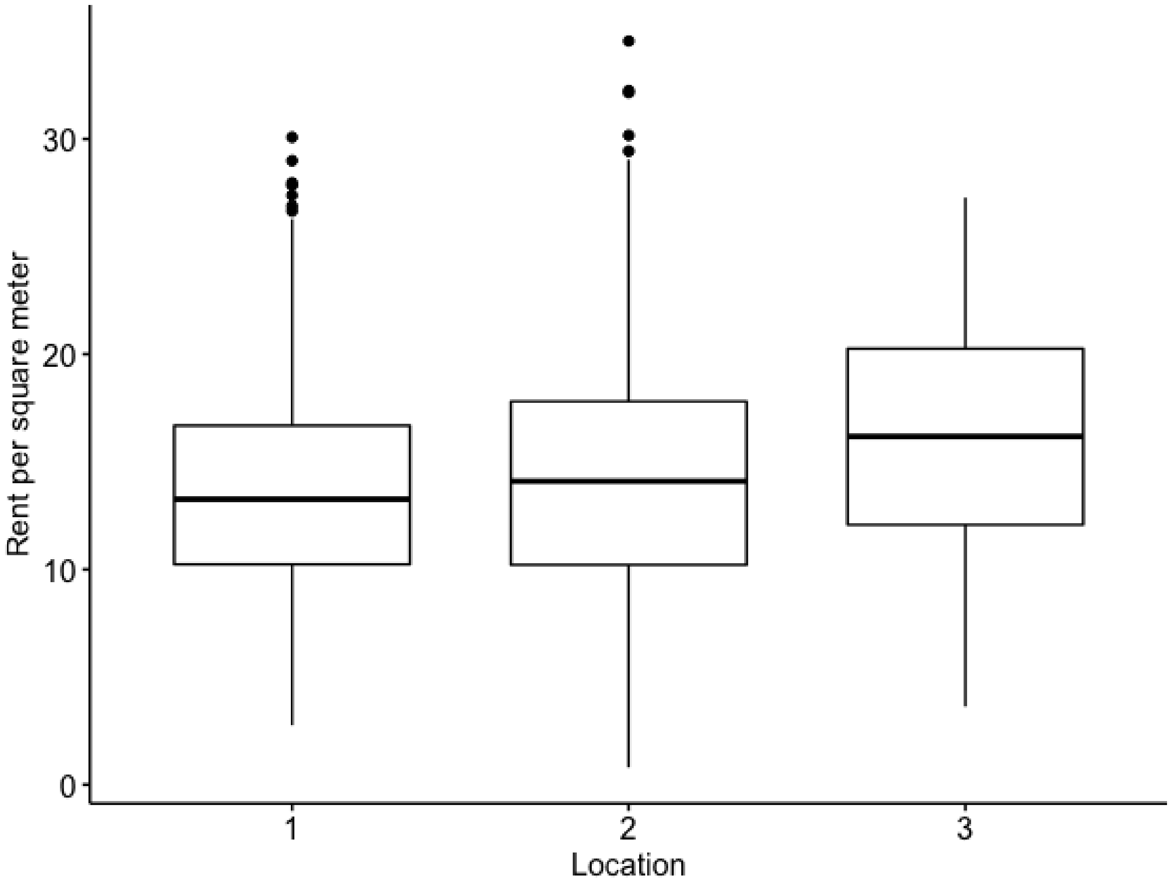

Figure 1: Homogenity of variance assumption analysis using Box Plot where location '1' Average, '2' Good and '3' Top.