# Contents

# 1 Introduction

The base rent for a property can be calculated based on a variety of criteria such as the year of construction, the size and location of the apartment, the amenities inside the apartment and the building, and other factors. Analyzing the rent index data of a particular city in the year 1999 will help us understand the costs of the above-mentioned factors before 23 years and it may also be useful to check how these costs have been changed from 1999 to 2022. In addition, it also helps to understand the real estate market and possibly provides tips for people planning to move into that particular city.

The aim of this report is to analyze the rental data of Munich from the year 1999 and determine whether the quality of location has an impact on the rent per square meter. First, we understand the provided dataset and check if the data set satisfies certain assumptions. Second, based on results from assumptions we determine whether to perform the test using the parametric (ANOVA) method or the nonparametric (rank-based) method to check whether the quality of location has a significant effect on rent per square meter. Third, if there exists a significant difference then we perform a test to determine which pairs of quality of location have a significant difference. Finally, we perform corrections to reduce the errors that may have occurred and make conclusions.

In Section 2, a brief description of the dataset is provided. In addition, the variable definitions and the project objectives are presented in more detail. In Section 3, different statistical methods relevant to this project are elaborated on and explained. In Section 4, we apply the statistical methods explained in the previous section to the project dataset and interpret the results. The final section (Section 5) will summarize our findings and ideas for further possible analysis of the given dataset.

# 2 Problem statement

## 2.1 Description of dataset

This project makes use of rent index data of Munich collected in the year 1999. The data set is provided by the lectures of the course Introductory to case studies at TU Dortmund University and it has a source from the book *Regression Modelle, Methoden und Anwendungen*. (Fahrmeir et al., 2007). The data set includes 3082 observations and 7 columns in total. Each observation corresponds to information about a particular

apartment in Munich. The explanatory variable *net rent* (numeric - continuous) refers to the base rent of a particular property in Munich. Further, the predictor variables like the *living area* (numeric - continuous) refer to the measured size of the living area of an apartment in square meters. *Construction year* (numeric - continuous) refers to the year in which the apartment is constructed. *Bathroom* refers to the quality of the bathroom in terms of categories (0 corresponds to standard and 1 corresponds to premium). *Kitchen* (categorical) refers to whether the apartment contains a standard *kitchen* or premium *kitchen* (0 = standard and 1 = premium). *Quality of location* (categorical) refers to how good the location of the apartment is based on several unknown factors (1 = average location, 2 = good location, 3 = top location). *Central heating* (categorical) refers to whether there is central heating in the apartment (0 = no, 1 = yes).

The quality of the data set is good because it is taken from a standard book, it contains a sufficient number of observations (3082 observations), and there are no missing values in the dataset.

## 2.2 Project objective

The project's purpose is to check whether the quality of location has a significant effect on the rent per square meter using statistical tests. To achieve this, first, a brief descriptive analysis is performed on the given data set and the summary of the results is presented in a table 3. Second, we check for the assumptions of normality, homogeneity of variance, and independence on the data set. Third, based on the result of the assumptions in the previous step, we choose to perform the parametric (ANOVA) or nonparametric (rank-based) method (Kruskal - Wallis test) to determine if there exists a difference in the rent per square meter for different categories of quality of location. Fourth, if there exists a difference, then we check for pairwise differences between all pairs of quality of location. Finally, we perform Bonferroni adjustments to control the type - I error that may occur due to multiple tests and we report the results in detail.

# 3 Statistical methods

The statistical methods essential for the analysis and testing of the data set are explained in detail, along with their graphs, formulas, and figures. All the following test results and graphs are created using the software R in version 4.0.5. (R Core Team, 2021).

## 3.1 Quantile - Quantile plot

For the data provided, a quantile-quantile plot can be used to determine whether or not the data corresponds to a specific distribution. In the case of a normality Q-Q plot, the data is checked to follow the normal distribution. If the data follows the normal distribution, then all the points of the data when plotted in the graph lie on the line x = y. The x-axis contains the range values of normal distribution, and the y - axis Contains the range of values in the dataset.

Assume a sample data with n observations $x_1, x_2, x_3, ..., x_n$.

Steps to construct Q-Q plot:

1. Sort the values of sample data in ascending order.
2. Calculate the i/(n+1)-quantile $q_i$ value of normal distribution for each $x_i$. Next, mark each pair of ordered $x_i$ and it's corresponding value $q_i$ in the graph.

(Dodge, 2008, p.437).

## 3.2 Inferential statistics

There are two statistical methods for analyzing data. The first one is descriptive statistics and the second one is inferential statistics. The method of inferential statistics uses the sample data, which represents the population and makes certain conclusions about the large population. In this subsection, the inferential statistical methods required for this report are described in detail.

**Null hypothesis and Alternative hypothesis**

A theory that a statistician wants to check if it is true or false can be called the null hypothesis ($\mathscr{H}_0$). The other possibility of the theories when a statistician assumes the null hypothesis is false can be termed as the alternative hypothesis ($\mathscr{H}_a$). It can be said that the null hypothesis is tested against the alternative hypothesis. If a statistician lacks enough evidence to reject the null hypothesis ($\mathscr{H}_0$), then the hypothesis is assumed to be valid. (Mood et al., 2017, p.405).

**Rejection and nonrejection region**

In the case of a statistical experiment, the results that don't fulfill the assumption of the null hypothesis are considered to be present in a particular region called the **rejection**

**region**, whereas, the results that fulfill the assumptions of the null hypothesis are considered to be in the opposite of rejection region and can be termed as the **nonrejection region**.(Black, 2019, p.297).

**Type - I and Type - II Errors**

The possibility of refuting the assumption of the null hypothesis ($\mathscr{H}_0$) even though it is true can be termed as the type - I error. Correspondingly, the possibility of not refuting the null hypothesis even though it is false can be termed as the type – II error. (Mood et al., 2017, p.405).

**Level of significance**

When conducting a statistical experiment, statisticians might restrict the possibility of making the type - I error to a certain level. This level can be termed as the level of significance. It is denoted by $\alpha$. In a certain experiment, if $\alpha$ is set to 0.02, then the probability to commit type - I error is 2%. (Black, 2019, p.298).

**Degree of freedom**

When estimating population parameters, the degrees of freedom determine the amount of information present in the selected sample. For example, In the case of the Kruskal Wallis test described in the subsection below, the degree of freedom is equal to $k - 1$.

where $k$ is count of the number of groups. (Sahai and Khurshid, 2001, p.77).

## 3.3 Hypothesis testing methods

### 3.3.1 Using the p-value to test the hypothesis

The p-value determines the minimum value of the level of significance that can be used to refute the assumption of the null hypothesis. The p-value can be termed as the *observed significance level*. In case of left tailed test if the obtained p-value from the test statistic is less than the level of significance then there is enough evidence to refute the assumption of the null hypothesis whereas if the p-value exceeds the value of ($\alpha$) then there is not enough evidence to refute the null hypothesis. (Black, 2019, p.302).

### 3.3.2 Kruskal Wallis test

The nonparametric method that can be used to decide whether the observations in different groups are similar or completely different from other groups can be called the

Kruskal Wallis test. It considers $k$ samples (greater than 2) are taken from $k$ different population groups to perform this test. Moreover, it also assumes that the groups are independent, the assumptions of normality and the homogeneity of variance are not satisfied.

The number of observations in each sample is denoted by $n_i$ where $i = 1, ..., k$.

$N$ represents the sum of the total number of observations in all the groups.

$N = \sum_{i=1}^{k} n_i$.

Further, to assign ranks, the $N$ values are sorted in ascending order irrespective of the samples. The ranks range from 1 to $N$, where the smaller values get lower ranks and the larger values get the higher ranks.

For instance, $X_{ij}$ denotes the observation $j$ of sample $i$ with $i = 1, ..., k$ and $j = 1, ...., n_i$. The rank for the $X_{ij}$ is represented as $R(X_{ij})$. If the same value is present in multiple observations, then the mean rank is assigned to them. The total of all ranks of observations in a particular sample group is represented by $R_i$.

$$R_i = \sum_{j=1}^{n_i} R(X_{ij}), \quad i = 1, ..., k,$$

The Kruskal Wallis test statistic for no mean ranks can be represented as follows:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \ ,$$

**Hypothesis**

The hypothesis to decide whether the observations of different groups are similar or entirely different from other groups can be stated as follows:

$H_0$: observations of k different groups are similar.

$H_a$: observations from at least one group are different from other groups.

**Decision rule**

Using the obtained test statistic and the degree of freedom $(k-1)$ the p-value can be extracted from the chi-square distribution table. If the p-value is less than the level of significance then there exists strong evidence to reject the null hypothesis $(H_0)$ else there is not enough evidence to reject the null hypothesis $(H_0)$. (Dodge, 2008, p.288).

### 3.3.3 Mann-Whitney U-test

A nonparametric statistical method that can be used to check whether two samples are related can be defined using the Mann-Whitney U-test. In the case of the nonparametric method, it is similar to the Wilcoxon rank sum test, and in the case of the parametric method, it is similar to the t-test for uncorrelated samples. When two samples are combined, the purpose of the test is to find out whether the observations of the two samples are ordered in a random fashion, or whether the values of each sample are almost completely separated in the opposite directions. Two samples are similar if they are ordered in a random fashion, and different if they can be grouped into separate samples. It assumes the sample data is independent.

**Hypothesis**

The null and alternate hypothesis for the Mann Whitney test can be stated as follows:

$H_0$: the distributions of values of all samples are equal.

$H_a$: at least one of the distribution of sample values is different.

For each of the two samples, the formula for the Mann-Whitney U-test statistic is as follows:

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

From the obtained values of two U-test statistics, the smaller value is considered.

Where:

$U_i$ - for a sample of interest it denotes the test statistic.

$n_i$ – count of values from our preferred sample.

$n_1$ – count of values in sample 1.

$n_2$ – count of values in sample 2.

$\sum R_i$ - addition of ranks from our preferred sample.

Use the Mann-Whitney critical value table to check the significance of the obtained U-test statistic. If the count of values in sample $n_i$ goes beyond the table values then another method is preferred to check significance, namely, large sample approximation. In this case, the z-score is calculated and the normal distribution table is used to find the critical region of the z-score.

In the case of large samples, the formulas to find the z-score of the U-test are mentioned below.

$$\bar{x}_U = \frac{n_1 n_2}{2},$$

Where:

$\bar{x}_U$ - represents mean.

$n_1$ – count of values in sample 1.

$n_2$ – count of values in sample 2.

$$s_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}},$$

Where $S_U$ represents the standard deviation.

$$z^* = \frac{U_i - \bar{x}_U}{s_u},$$

Where:

$z^*$ - z-score for normal estimation of data.

$U_i$ - for a sample of interest it denotes the test statistic.

(Corder and Foreman, 2009, p.58).

### 3.3.4 Bonferroni adjustments

When checking for a significant effect by performing multiple tests on the same variable, the chances to commit a type - I error increases. In order to control or decrease this error, Bonferroni adjustments can be used. It is the preferred method when performing a few concurrent tests (until 5 tests) as it gives appropriate results, whereas, in the case of more concurrent tests it is preferred to other procedures. If $m$ concurrent tests are performed on the same variable, in order to limit the type - I error to a certain level of significance ($\alpha$), we use the following formula $\alpha$/m.

Where:

$\alpha$ - level of significance.

$m$ - total number of tests.

(Everitt and Skrondal, 2010, p.58).

# 4 Statistical analysis

The statistical methods described in the previous section were applied on the dataset used in this report and the results are interpreted in this section.

Table 3 in the appendix gives information about the summary related to different categories of quality of location with respect to rent per square meter. The *Number of observations* refers to count of observations belonging to each category of quality of location. *Mean* refers to the average value of rent per square meter in each category. *sd* refers to the standard deviation of rent per square meter in each category. *IQR* refers to the length of the interquartile range of variable rent per square meter in each category.

All the data values in table 3 are rounded to 2 decimal places.

## 4.1 Analysis of assumptions

In order to choose between the parametric method (ANOVA) and nonparametric (rank-based) method, we first check for the following assumptions. If our data set satisfies these assumptions, then we perform the ANOVA test, otherwise, we perform the Kruskal Wallis (nonparametric rank-based) test to determine whether the quality of location has a significant effect on net rent per square meter.

**Assumption of independence**

In this part, we confirm that the observations of the sample taken from the population don't depend on each other. In the dataset used in this report, each observation is associated with a specific apartment in Munich, and these observations were collected at random. As values of one observation don't depend on the other, we can assume that the condition of independence is true within the scope of this report.

**Assumption of homogeneity of variance**

In this part, we check whether the samples drawn from the population contain equal variances (homogeneity) or not. The figure 1 visualizes the spread of the data using boxplots. From the figure and table 3 it is evident that the length of the interquartile

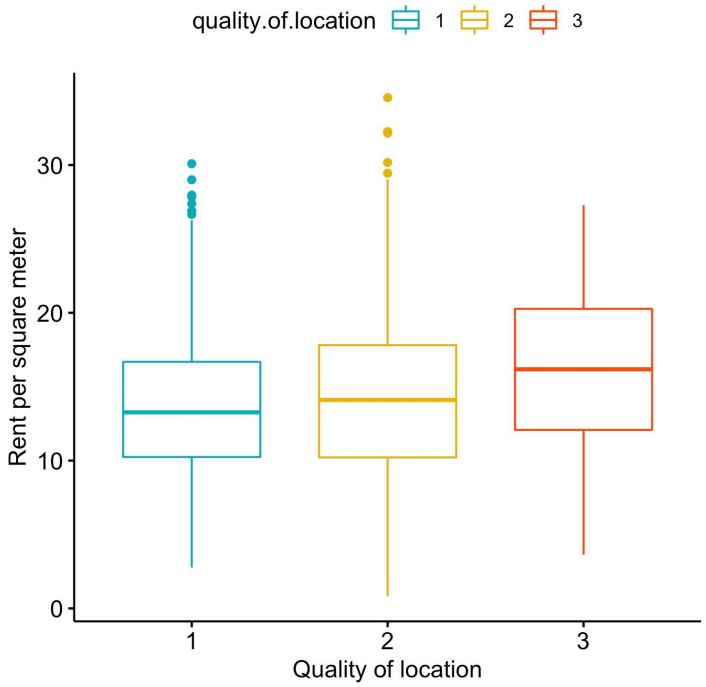


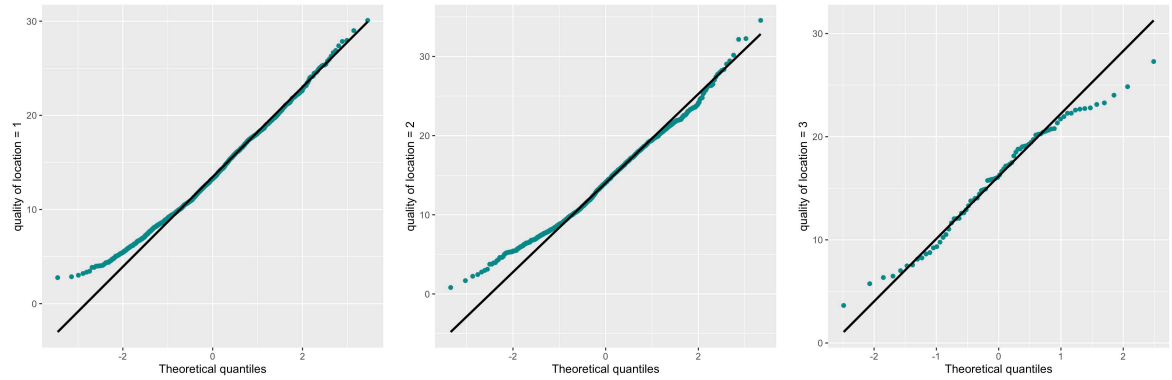

Figure 1: Box plot for each category of quality of location

range (IQR) is not the same for all the categories of quality of location. The quality of locations 1, 2, and 3 contain IQR values of 6.44, 7.59, and 8.19 respectively. Moreover, in categories of location quality 1  2, there exist more outliers. From the above statements, we can conclude that the data doesn't satisfy assumptions of homogeneity of variance.

**Assumption of normality**

In this part, we check whether the samples drawn from the population are normally distributed or not. The figure 2 visualizes the Q-Q plots for points grouped based on the quality of location. From the figure, we can notice that the data points of category 1 deviate a bit from the linear line for smaller values. In the case of category 2, the deviation can be observed for both smaller and larger values and in the case of category 3, the points have deviated a lot from the linear line. From these observations, we can conclude that the data doesn't satisfy the assumption of normality.

Finally, in this subsection, we can conclude that the data set used in this project satisfies the assumption of independence but it doesn't satisfy the assumption of normality and the assumption of homogeneity of variance. Therefore, we proceed with the Kruskal Wallis test.




(a) Quality of location = 1 (b) Quality of location = 2 (c) Quality of location = 3

Figure 2: Comparison of normality using Q-Q plots

## 4.2 Test for global hypothesis

In this part, we will perform the Kruskal Wallis test to check whether the rent per square meter varies for different categories of quality of location. The null and alternative hypotheses are described below. The test is performed by assuming the level of significance to be 5% or 0.05. This shows that the probability of falsely concluding that there exists a variation in net rent per square meter (Type - I error) is 5%.

**Null hypothesis** ($H_0$) : There exists no difference in rent per square meter for different categories of locations.

**Alternative hypothesis** ($H_a$) : At least one category of quality of location have values of rent per square meter different from other categories.

| Chi-squared value | Degrees of freedom | P-Value |
|---|---|---|
| 23.451 | 2 | 0.000 |

Table 1: Result of Kruskal-Wallis test

Table 1 shows the results of the Kruskal Wallis test. All the data values in table 1 are rounded to 2 decimal places. From the table, we can notice the degrees of freedom is 2. This is because the quality of location contains three categories in the quality of location and degrees of freedom is equal to the number of groups minus one. We can also notice the p-value is 0.000, which is less than the level of significance (0.05). This shows enough evidence to reject the null hypothesis. From this, we can conclude that there exists at

least one category of location that contain values of rent per square meter different from other categories.

## 4.3 Pairwise testing

In this part, our goal is to specify which pairs of qualities of location have pairwise differences between them with respect to the rent per square meter. We perform Mann Whitney test to determine the pairs with significant effects. Similar to the global test, the null and alternative hypotheses are illustrated below, and the significance level is specified at 5%. In this case, we can say that the probability to conclude there exists a pairwise difference when there is no pairwise difference is 5% (Type-I error). 3 unique pairwise tests are performed in total and the results are presented in Table 2.

**Null hypothesis** ($H_0$) : There is no pairwise difference between the categories of quality of location.

**Alternate hypothesis** ($H_a$) : There exists at least one pairwise difference between the categories of quality of location.

If the obtained p-value from the test is less than the level of significance then the null hypothesis is rejected. In this case when performing multiple tests on the same variable the possibility to commit type-1 error increases. To control this, we use the method of Bonferroni adjustments. The column p-value in table 2 represents the values before using Bonferroni adjustments, and the column Bonferroni adjusted p-value represents the values after performing the adjustments.

| Quality of location pair | p-value | Rejected | Bonferroni adjusted p-value | Reject adjusted |
|---|---|---|---|---|
| average-good | 0.002 | True | 0.006 | True |
| average-top | 0.000 | True | 0.000 | True |
| good-top | 0.002 | True | 0.007 | True |

Table 2: Pairwise test results with and without Bonferroni adjustments

We can notice from table 2, that the p-values of the quality of location pairs are less than the level of significance before applying Bonferroni adjustments. We can also notice that the adjusted p-values are also less than the level of significance. From these results, we can conclude that there exists a pairwise difference between the rent per square meter and all the pairs of quality of location.

# 5 Summary

In this report, tests are conducted to examine whether location quality has a significant impact on net rent per square meter, and the results are explained. The dataset used in this report was compiled by the instructors of the course Introductory to case studies at TU Dortmund University in the summer semester 2022. The data set was taken from the recognized book *Regression Modelle, Methoden und Anwendungen.* (Fahrmeir et al., 2007). The data set was related to rental index data of a specific city (Munich) for the year 1999. It contained a total of 3082 observations belonging to the year 1999 with *net rent* as the response variable and 6 other predictor variables namely *living area, construction year, bathroom, kitchen, quality of location,* and *central heating.*

To achieve the goals, First, descriptive analysis was performed on the relevant variables of the data set, and the summary of the results are included in the table 3. Second, the data set was checked against certain assumptions to decide between the parametric and nonparametric test methods. Here, we understood that our data satisfied the assumption of independence but failed to satisfy the assumptions of normality and homogeneity of variance. From this, a conclusion is made to proceed with a nonparametric test (Kruskal Wallis test) to check for significant effect. Third, From the Kruskal Wallis test, we understood that the quality of location had a significant effect on net rent per square meter. Fourth, to determine the pairwise differences between the rent per square meter and the different location qualities Mann Whitney test was performed, and also to control type-1 error Bonferroni adjustments was performed. Finally, from the results of the Mann Whitney test, we concluded that all the pairs of location qualities had differences with respect to rent per square meter even after the Bonferroni adjustments.

For further possible analysis, including observations from the current year may help to understand how the rent per square meter and other variables have changed over the years. Also, including observations from different cities can help to understand whether the quality of location has a significant effect on rent per square meter only in Munich, or if it has the same effect on other cities in general.

# Bibliography


Black, K. (2019), *Business statistics: for contemporary decision making.*, John Wiley.

Corder, G. W. and Foreman, D. I. (2009), *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*, John Wiley  Sons, Inc, Hoboken, New Jersey.

Dodge, Y. (2008), *The Concise Encyclopedia of Statistics*, Springer New York, New York, NY.

Everitt, B. and Skrondal, A. (2010), *The Cambridge Dictionary of Statistics*, Cambridge University Press.

Fahrmeir, L., Kneib, T. and Lang, S. (2007), *Regression Modelle, Methoden und Anwendungen*, Springer Berlin, Heidelberg.

Mood, A. M., Graybill, F. A. and Boes, D. C. (2017), *Introduction to the theory of statistics*, McGraw-Hill, Inc., USA.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Version: 4.0.5, Vienna, Austria.

Sahai, H. and Khurshid, A. (2001), *Pocket Dictionary of Statistics*, McGraw-Hill/Irwin.

# Appendix

## A Additional figures

## B Additional tables

| Quality of location | Number of observations | Mean | sd | IQR |
|---|---|---|---|---|
| 1 - average | 1794 | 13.60 | 4.45 | 6.44 |
| 2 - good | 1210 | 14.20 | 5.08 | 7.59 |
| 3 - top | 78 | 15.90 | 5.38 | 8.19 |

Table 3: Summary of quality of locations' rent per square meter