TU Dortmund

Introductory Case Studies

Project II: Discrete covariates

Lecturers:

Dr. rer. nat. Maximilian Wechsung

M. Sc. Hendrik Dohme

Author: Supritha Palguna

Matriculation number: 229593

Group number: 20

Group members: Kunal Kochar, Dhanunjaya Elluri, Harshini Eggoni, Naveen Kumar Bhageradhi, Ashish Saini

May 27, 2022

Contents

1 Introduction

A collection of data that displays the rental prices of similarly equipped residences in a certain area is known as a rent index. The rent index was created with the intention of protecting renters against excessive rents and rent increases. Landlords nowadays utilize it to justify a permissible increase in rent. A rent index is made up of different categories that describe the attributes of a property. These categories include location, construction year, equipment, energetic status, and so on. Given that no two rental properties are the same, it would be interesting to compare the rents of the properties based on categories and see how these categories affect the rental price.

The objective of this report is to use data obtained in 1999 to compare the quality of locations in Munich. First, the data is statistically analyzed using statistical metrics such as mean, variance, and standard deviation. Testing assumptions are checked to decide between the parametric and nonparametric methods of testing. The Kruskal test is then used to see if the location's quality has a substantial impact on the rent per square meter. Further, the Wilcoxon–Mann–Whitney (WMW) U-test is used to see if there are pairwise differences in rent per square meter for different location qualities. The Wilcox test results are adjusted using the Bonferroni method to eliminate Type I error, and the results are compared to the Wilcoxon test results.

Section 2 describes the dataset, which involves data collection technique, type and size of the sample data and description of variables. It also gives an overview of the objectives of the project. Section 3 provides a description of several statistical methods employed in this study. This comprises the Kruskal test, the Wilcox test, the Bonferroni correction method and so on. In section 4, results are presented with tables and graphs and also interpreted in relation to the problem statement. The last section includes a succinct summary of the important findings and an outlook on further analysis.

2 Problem statement

The dataset and project objective are discussed in detail in this section.

2.1 Dataset

The dataset examined for this project covers over 3082 apartments in Munich. It includes data on the net rent as well as six other covariates for the apartments. The total seven covariates in the dataset are described as follows: '*net rent*' is a continuous variable which corresponds to the net rent of an apartment, '*living area*' indicates the size of the living area of an apartment in square meters, '*construction year*', a discrete integer, is the year in which the apartment was constructed, '*bathroom*' is a binary variable which describes the quality of a bathroom, where '0' being standard and '1' being premium, '*kitchen*', a binary variable, indicates the quality of kitchen, where '0' indicates standard and '1' indicates premium, '*quality of location*' is a categorical variable that corresponds to the quality of an apartment's location, where '1' means an average location, '2' a good location and '3' a top location, '*central heating*' is a binary and nominal variable which tells whether the apartment has a central heating or not.

2.2 Project Objective

The goal of this project is to conduct hypothesis testing to determine the effect of quality of location on rent per square meter and to determine whether there are any pairwise differences between location qualities. Initially, using measures of central tendency and measures of dispersion, a descriptive analysis of the rent per square meter is performed across all three location qualities. Further, QQ plot and box plot are used to verify the assumptions of statistical tests. By proving that some of the assumptions are violated, a Kruskal-Wallis test is performed to determine if rent per square meter is effected by quality of location. Later, Wilcoxon–Mann–Whitney tests are conducted on three pairs of location qualities to check if there are any pairwise differences between them. To solve the multiple testing issue, the Wilcoxon–Mann–Whitney test results are adjusted using Bonferroni correction. Finally, the adjusted and unadjusted findings are compared.

3 Statistical methods

Various statistical measures and graphs used for analysis of the dataset are explained in this section. The statistical software R (R Development Core Team, 2020), version 4.0.3 is used for analysis. dplyr (Wickham et al., 2014), ggpubr (Kassambara, 2020), ggplot2 (Wickham, 2016) and gridExtra (Auguie and Antonov, 2017) are the R packages used.

3.1 Statistical Measures and Plots

In the last project, 'Descriptive analysis of demographic data,' statistical measures like mean, median, and variance, as well as graphs like scatter plot and box plot were discussed.

Quantile-Quantile Plot

QQ-plots help to compare distribution of a sample to that of the standard normal distribution. It is obtained by plotting quantiles of two variables against each other. The normality assumption is checked using QQ-plot by plotting the theoretical quantiles of a standard normal distribution against the quantiles from the standardized residuals. The y-axis represents the quantiles of the sample of data and the x-axis represents the theoretical quantiles. If the plotted points closely follow a linear trend then there is an evidence to say that normality assumption is met otherwise it is not (Hay-Jahans, 2019, p. 146-150), (Heumann et al., 2016, p. 44-45).

3.2 Hypothesis Testing

A research hypothesis is a prediction of what the researcher expects to happen in an experiment or study. A statistical hypothesis is a formal hypothesis structure that includes a null and alternative hypothesis for testing research hypotheses scientifically.

The research hypothesis which needs to be proved is considered as the statistical alternative hypothesis, represented as H_1. And, the opposite of the research hypothesis is articulated as the statistical null hypothesis, denoted by H_0.

A Type I error occurs when a correct null hypothesis is rejected. The null hypothesis is true in a Type I error, but decision is formed that it is false. A Type II error occurs when a false null hypothesis is not rejected. The null hypothesis is false in this scenario, but it is decided not to be rejected. These situations are clearly depicted in Table 1.

Table 1: Four possible outcomes when a hypothesis is tested.

	H_0 is true	H_0 is false
H_0 is not rejected	Correct decision	Type II error
H_0 is rejected	Type I error	Correct decision

After establishing null and alternative hypothesis, appropriate statistical tests need to be determined based on the assumptions of testing. The statistical tests are discussed in section 3.4 and 3.5. Further, α or level of significance needs to be determined. α is the probability of making a Type I error. The minimal value of alpha for which the null hypothesis can be rejected is defined by the p-value. Based on the comparison of the alpha and p-values, the null hypothesis is rejected or fails to be rejected. With the null hypothesis assumed to be valid, the p-value is the probability of generating outcomes at least as significant as the results obtained of a statistical hypothesis test. If the p-value for a test statistic is larger than the α value, the null hypothesis is not rejected; if it is less than or equal to the α value, the null hypothesis is rejected, indicating that the result is statistically significant (Black, 2014, p. 291-303), (Heumann et al., 2016, p. 211).

p-value $\leq \alpha$ indicates that the null hypothesis H_0 gets rejected

p-value $> \alpha$ indicates that the null hypothesis fails to get rejected.

3.3 Assumptions for testing

The data assumptions that must be checked before deciding whether to apply a parametric (ANOVA) or nonparametric (rank-based) technique are listed below.

- Sample Independence - The sample data has to be gathered independently of one another.
- Normality - Data is collected from populations that are normally distributed. A continuous probability distribution is the normal or Gaussian distribution. It is symmetrical around its mean and asymptotic to the horizontal axis, indicating that data close to the mean occur more frequently than those far from it. The area under the curve is 1, and it has mean and standard deviation as two parameters.(Black, 2014, p. 407-409).
- Homogeneity of variance - Different groups have the same population variances.

3.4 Kruskal-Wallis Test

The Kruskal-Wallis test, invented by William H. Kruskal and W. Allen Wallis in 1952, is a nonparametric alternative to the one-way analysis of variance. This test is used to see

if samples (≥ 3) are from the same population or from distinct populations. It's a tool for analyzing ordinal data and doesn't make any assumptions about population shape. The Kruskal-Wallis test assumes that the groups are independent and the individual items are chosen at random.

The hypothesis of Kruskal-Wallis test follows:

H_0: The populations are identical.

H_1: At least one of the populations is distinct from the others.

This test examines whether all of the groups are from the same or similar populations, or if at least one group is from a different one.

Kruskal-Wallis K statistic is computed as follows:

$$K = \frac{12}{n(n+1)} \left(\sum_{j=1}^{c} \frac{T_j^2}{n_j} \right) - 3(n+1)$$

where,

c = number of groups

n = total number of items

T_j = total number of ranks in a group

n_j = number of items in a group

$k \approx \chi^2$, with $df = c - 1$

With $c - 1$ degrees of freedom, the K value is approximately chi-square distributed (Black, 2014).

3.5 Wilcoxon–Mann–Whitney (WMW) U-Test

The Wilcoxon–Mann–Whitney (WMW) U-Test would be the suitable test to use when the ordinal data is provided or if the parametric assumptions are not met. It is used to compare two samples that are unrelated or independent. Both samples are pooled and rank ordered. The purpose of the method is to see if the values from the two samples are mixed at random in the rank ordering or if they cluster at opposite ends when integrated. A random rank order would imply that the two samples are unrelated, whereas a cluster

of one sample's values would suggest that they are (W.Corder and I.Foreman, 2009, p. 58-59).

Formula to compute a Wilcoxon–Mann–Whitney (WMW) U-Test statistic for each of the two samples is given by,

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

where,

U_i = test statistic for the sample of interest

n_i = number of values from the sample of interest

n_1 = number of values from the first sample

n_2 = number of values from the second sample

$\sum R_i$ = sum of the ranks from the sample of interest

The smaller of the two U statistics is the resultant value. The significance of the U statistic must be determined once it has been computed. Table of critical values may be used for this. In case, the number of values in each sample exceeds those accessible from the table, then a large sample approximation is to be conducted. Critical region of z-scores is obtained by computing z-score and using normal distribution table. To find z-score of a WMW U-test for large samples following formula can be used,

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

where, $\bar{x}_U$ is the mean. The standard deviation s_U is given by,

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

The z-score, z^* is given by,

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

3.6 Bonferroni Correction

The probability of making a type I error increases and surpasses the α value when numerous tests are run on the same sample of data. The Bonferroni correction is a conservative strategy for probability thresholding in multiple hypothesis testing to control the occurrence of false positives or type I error. The Bonferroni correction is applied to the p-value. By multiplying the reported p-value by the number of statistical analyses performed on the dataset, the new adjusted p-value is generated. The modified p-value is set to 1 if the new value is larger than 1. The modified p-value is then compared to the α value for hypothesis testing (Hay-Jahans, 2019, p. 274) (Herzog et al., 2019, p. 63).

4 Statistical analysis

The statistical analysis of rental prices per square meter across three different quality of locations is conducted in this section using the statistical methods outlined in Section 3.

4.1 Descriptive analysis

Table 2: Summary table of rent per square meter across the different quality of locations.

Quality of location	Count	Min	Mean	Variance	Max	IQR
1-average location	1794	2.76	13.56	19.79	30.08	6.44
2-best location	1210	0.81	14.18	25.79	34.56	7.59
3-top location	78	3.64	15.94	28.95	27.29	8.18

The statistical measures for rent per square meter for different location qualities are shown in Table 2. The number of apartments in the average location is 1794 which is the highest among all other location qualities. Good location has 1210 apartments, a bit lesser than the average location. The top location has the least number of apartments i.e, 78. The average rent per square meter is highest for the top location i.e., 15.94. and least for the average location (13.56). The highest and lowest rent per square meter among all location qualities can be found in average location which is 34.56 and 0.81 respectively.

4.2 Verification of Test Assumptions

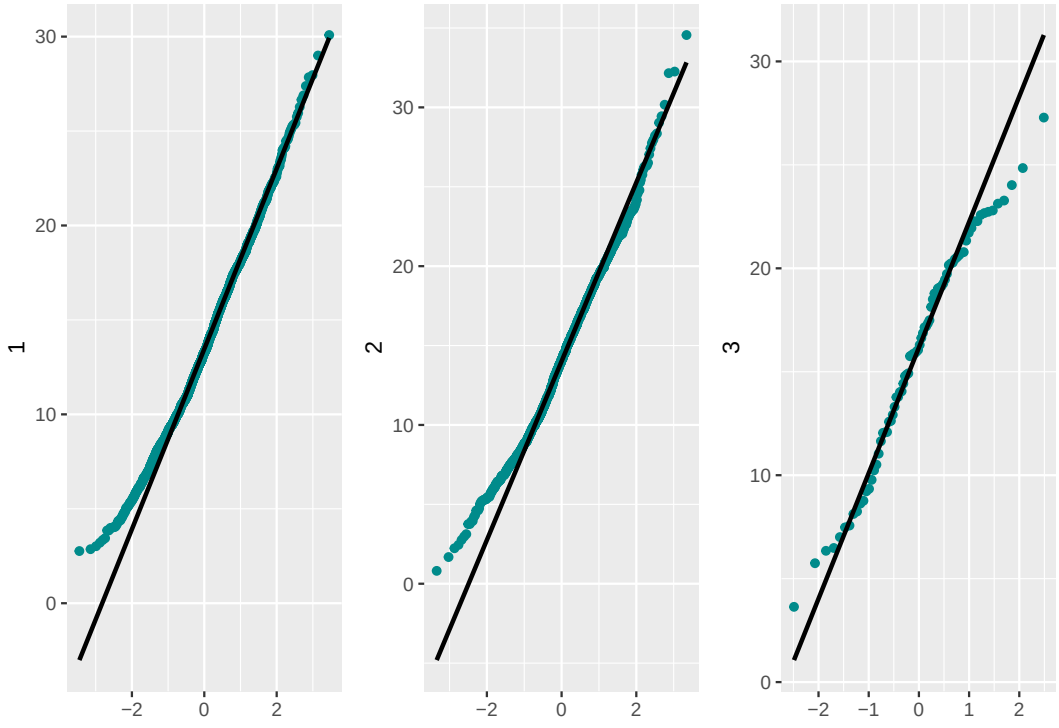

Figure 1: QQ Plots of three different location qualities.

The assumptions of the statistical tests indicated in Subsection 3.3 are evaluated on the dataset in this subsection. Since, the samples here are drawn from the dataset independently, the observations are independent. Hence, the independence assumption is not violated by the dataset.

A QQ plot is plotted for the normality assumption, as shown in Figure 1. The theoretical quantiles of the normal distribution are represented on the x-axis, while the quantiles of the rent per square meter of each location quality are represented on the y-axis. The points on the graph deviate away from the straight line for the top location. For the average and best locations, some points lie close to the straight line and some are far from the straight line. This signifies that the rent per square meter does not follow theoretical normal distribution. As a result, it may be concluded that the dataset violates the normality assumption.

The data is then evaluated for variance homogeneity across all the location qualities. Figure 2 shows a boxplot that is used for this purpose. The location qualities are represented on the x-axis, while rent per square meter is represented on the y-axis. The

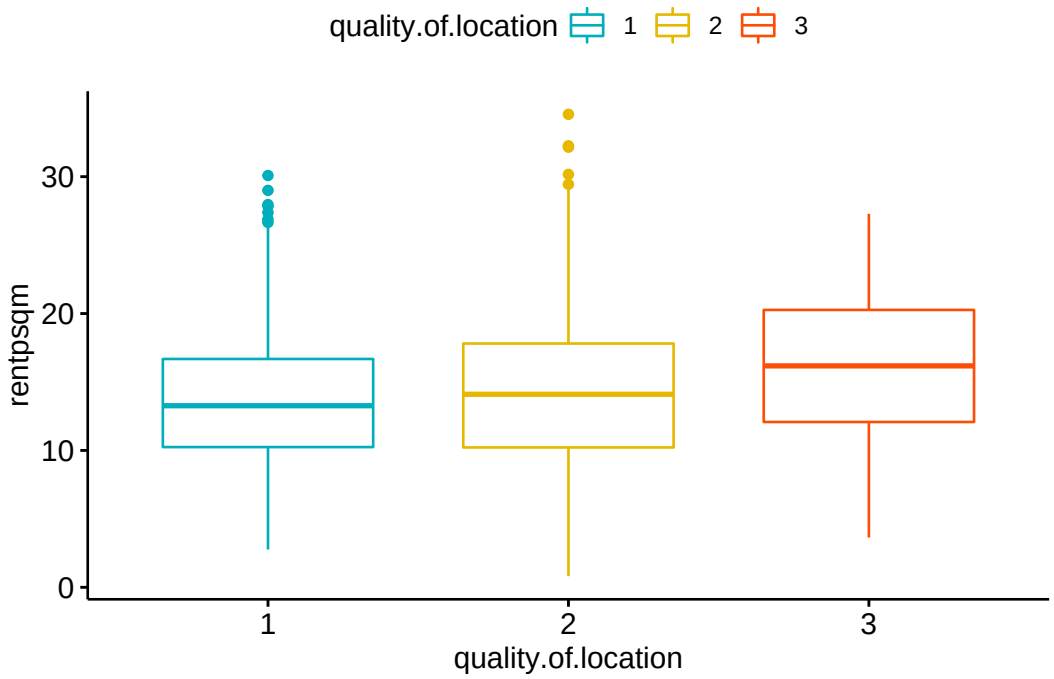

Figure 2: Boxplot representing the rent per square meter of different location qualities

inter-quartile range for the location qualities are quite different from each other. Also, from Table 2, it can be observed that the variance is highest for the top location i.e., 28.95 and the average location has the least variance of 19.79. There's a huge gap among the variances of different location qualities. Hence, it can be said that the homogeneity of variance is violated by the dataset.

Therefore, the dataset holds only one assumption of independence and violates the two assumptions of normality and homogeneity of variance. Hence, it would be appropriate to use Kruskal-Wallis and Wilcoxon–Mann–Whitney (WMW) tests.

4.3 Global test for relationship between rent and quality of location

In this subsection, the significant effect of the location quality on rent per square meter is discussed. Kruskal-Wallis Test is performed since the dataset contains categorical variables and violates normality and homogeneity of variance assumptions. The null and alternative hypotheses evaluated in this statistical test are listed below,

> H_0 : All three location qualities have same distribution in relation with rent per square meter.

H_1 : At least one of the location qualities' distribution in relation with rent per square meter is different from the others.

Table 3: Result of Kruskal-Wallis test

chi-squared	degrees of freedom	p-value
23.451	2	8.087e-06

The α is set to 0.05. The result of the Kruskal-Wallis test is shown in Table 3 . Here, the p-value is less than α (0.05). Hence, the null hypothesis can be rejected. Therefore, it can be said that there exists at least one location quality whose distribution based on rent per square meter is different from the rest and quality of location has a significant effect on the rent per square meter.

4.4 Pairwise comparison of location qualities

Kruskal-Wallis test lead to the conclusion that the distributions of quality of locations are not identical. To figure out which groups are different, all pairwise hypothesis are tested by Wilcoxon–Mann–Whitney test. Since there are three location qualities in the dataset, this test evaluates a total of three unique pairs of location qualities. A total of three WMW U-tests have been conducted.

Table 4: Summary of WMW U-test results for the three pairs of location qualities before and after the Bonferroni correction. ($\alpha = 0.05$).

		Before Correction		After Correction	
S.No	Quality of location	p-value	Reject Y/N	p-value	Reject Y/N
Test 1	1-average location and 2-best location	0.0021	Y	0.0062	Y
Test 2	1-average location and 3-top location	0	Y	0.0001	Y
Test 3	2-best location and 3-top location	0.0025	Y	0.0074	Y

The findings of the WMW U-test before and after the Bonferroni adjustment are summarized in Table 4. The column 'Quality of location' signifies the pair of location qualities. There are two sub-columns in 'Before Correction', which are 'p-value' referring to the p-value generated from the WMW U-test and 'Reject Y/N' representing the outcome of the hypothesis test before employing the Bonferroni correction. 'Y' in the column 'Reject Y/N' says that the null hypothesis is rejected for that test and 'N' describes

that the null hypothesis fails to get rejected. The α value is set to 0.05. By comparing the p-value to the α, the results of the hypothesis tests can be determined. The null hypothesis suggests that the rent per square meter for different location qualities does not differ. The alternate hypothesis, on the other hand, claims the opposite.

The null hypothesis is rejected for all three tests since the p-value is less than 0.05 for all the scenarios. Hence, there is enough evidence to says that there is a significant difference between the rent per square meter for location qualities average location - best location, average location - top location and best location - top location.

Bonferroni Correction

The p-values obtained from the WMW U-test are subjected to the Bonferroni correction. The result of the hypothesis test after the Bonferroni correction is shown in the column 'After correction' of Table 4. The α value used here is 0.05. It can be seen that p-values in all three cases are below 0.05 and hence the null hypothesis gets rejected for all three pairs of location qualities after Bonferroni correction. The p-values have increased slightly from before correction and the hypothesis result still remains the same.

5 Summary

The goal of this project is to see how the rent of 3082 apartments in Munich for different quality of location. The lecturers of the course "Introductory Case Studies" at TU Dortmund in the summer semester of 2022 have compiled and provided the dataset used in this project. A univariate analysis of rent per square meter was conducted first, followed by a global Kruskal-Wallis test to see if the rent per square meter is effect by location qualities. Then, for each of the three unique pairs of location qualities, a Wilcoxon–Mann–Whitney U-test was conducted to verify if are any pairwise difference the quality of location. To address the issue of multiple testing, the test results were then modified using Bonferroni correction. In addition, the corrected as well as the original findings were compared.

The average rent per square meter is the highest for top location quality which is 15.94 and has lowest number of apartments (78). Average location quality has the lowest rent per square meter but has the highest number of apartments (1794). Also, top location quality has the highest variance of 28.95. It was found that the variance of rents per

square meter for different quality of location is not homogeneous among each other. Normality assumption also failed to hold. Hence, Kruskal-Wallis was used as a global test. From this test it was concluded that there is a significant effect of quality of location on rent per square meter.

Further, Wilcoxon–Mann–Whitney U-tests were performed on the pairs of quality of location. It was found that there is a signification difference among each pair of location qualities. Later, the p-values were modified using the Bonferroni correction method to overcome the issue of multiple testing. It was seen that the null hypothesis was rejected for three test cases after Bonferroni correction confirming the existence of pairwise difference in location qualities.

For future analysis, variables such as location and construction year might be included and to see how these aspects differ between different locations by performing statistical tests. Controlling type II error can also be considered for future hypothesis testing, in addition to type I error.

Bibliography

Baptiste Auguie and Anton Antonov. *gridExtra: Miscellaneous Functions for "Grid" Graphics.* 2017. URL `https://cran.r-project.org/web/packages/gridExtra/index.html`. (Visited on 15th May 2022).

Ken Black. *Business statistics : for contemporary decision making.* Hoboken, New Jersey : Wiley, 2014.

Christopher Hay-Jahans. *R companion to elementary applied statistics.* Taylor Francis Group, 2019.

Michael Herzog, Gregory Francis, and Aaron Clarke. *Understanding Statistics and Experimental Design.* Spring St, New York, USA: Springer-Verlag New York, 2019.

Christian Heumann, Michael Schomaker, and Shalabh. *Introduction to Statistics and Data Analysis.* 2016. doi: 10.1007/978-3-319-46162-5.

Alboukadel Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots.* 2020. URL `https://rpkgs.datanovia.com/ggpubr/`. (Visited on 15th May 2022).

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

Gregory W.Corder and Dale I.Foreman. *Nonparametric statistics for non-statisticians.* A John Wilet & Sons, Inc., Publication, 2009.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. Spring St, New York, USA, 2016. URL `http://ggplot2.org`.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation.* 2014. URL `https://dplyr.tidyverse.org`. (Visited on 15th May 2022).