TU Dortmund

Introductory Case Studies

# Project 2: Comparison of multiple distributions

Lecturers:
Prof. Dr. Sonja Kuhnt
Dr. Birte Hellwig
Dr. Paul Wiemann
M. Sc. Hendrik Dohme

Author: Amirreza Khamehchin Khiabani

Group number: 1

Group members: Amirreza Khamehchin Khiabani, Rama Kassoumeh, Elchin Latifli, Shivam Dabral, Shubham Gupta

December 10, 2021

# Contents

# 1 Introduction

One of the three biggest real estate web portals in Germany is Immobilienscout24 (immoscout24.de). Lists of rental properties and homes for sale can be found on its website according to the place, the square meters per price of properties, and so on. In this report, the purpose is to examine the averages of rental price per square meters of four cities, Dortmund, Bochum, Duisburg, and Essen, to measure whether these values are equal to each other or not in the two tasks requested. Several statistical methods are applied for this examination, mostly focusing on statistical testing to help to understand the differences between the rental prices per square meter. The desire assumptions of these statistical methods firstly are examined. Since without checking, the results of all these different tests are sensitive to their correctness of suppositions.

In Section 2, a summary regarding the granted data set is provided and the quality of the data is discussed. Additionally, all tasks related to this report are discussed and determined the methods that would be suitable to use for understanding the differences. In the next section, the statistical methods, including verifying normality, assumption and formulation of one-way ANOVA, paired-samples t-test, and Bonferroni method are described besides representing some of the terminology adopted in these statistical tests. In the Statistical analysis section, all these statistical methods are practised to present the results of these tests and provide analysis on these results. Firstly, the assumptions of normality is checked with Q-Q plots. Then, the equality of variances are questioned by applying box plots. The equality of means of all cities are initially tested via ANOVA. Furthermore, paired sample t-test are conducted and results of this method are compared with and without applying Bonferroni correction. The final section includes a summary of the results, a review of what is examined and an outlook on extra analyses.

# 2 Problem statement

The data is provided by the website (kaggle.com), used for educational and research goals. This data is obtained from a large data set including rental offers of properties in Germany as of February 2020, containing the rental price per square meter for 200 properties placed in the four largest cities of the Ruhrgebiet (Ruhr area). The data is scrapped from Immoscout24 web page. This process was repeated three times (CorrieBar).

The original data set includes the essential properties, like the size of living area, the rent, both base rent as well as total rent (if applicable), the location (street and house number, if available, ZIP code and state), type of energy, and so on. The column of the date was included to provide the time of the scraping process. The data set of this report, on the other hand, can be obtained from an excel file (ImmoDataRuhr.csv), including three labels, ID, price per square meter, and the region with this value, Bochum, Dortmund, Duisburg, and Essen. There are 200 rows in this data set, 50 properties for each city. The ID and rent per square meter are numerical variables, discrete and continuous, respectively. The region or city variable is categorical in the form of nominal. The data quality is consistent. There is no missing value in the whole data set. Therefore, no statistical method is employed for this problem. These values represented samples of these four cities for conducting statistical tests for observing the differences between the average rental price per square meter. The data were sampled from these populations independently. Since the value of one sample does not reveal any information about the another one. For example, the value of rental price per square meter of Dortmund, provides no information for other cities.

The analytical aims are to solve three tasks regarding comparing the values of the rental per square meter. In the first, the main request is to implement a global test on the equality of average of these cities to conclude that these populations have approximately the same value for the average rental price. Secondly, a paired-samples t-test is conducted for examining pair to pair of each city to see the equality of mean statistically. Lastly, test results are adjusted via the Bonferroni method to discuss the problem of multiple testing. Then, the results of this task are compared with the previous one without adjustment. In each case, the assumptions of the tests are examined to assess the extent to which the outcomes of these tests can be relied on.

# 3 Statistical methods

The following statistical methods and mathematical formulas are used. The statistical software R (Version 4.1.0, R Development Core Team) has been used for analysis with applying two extra packages from R, including *ggplot2* (Hadley Wickham) and *dplyr* (Hadley Wickham, Romain François, Lionel Henry,and Kirill Müller).

## 3.1 Q-Q Plots

In this report, quantile-quantile or Q-Q plots are types of scatter plots that are applied to visually examine for the assumption of normally distributed samples. The ordered values of the sample, $y_1, y_2, ..., y_n$ are treated as experimental quantiles for the sample. These are plotted on the y-axis against the quantiles of a theoretical standard normal distribution on the x-axis. The theoretical quantile that corresponds to each ordered empirical quantile $y_i$ is determined as ( B. S. Everitt and A. Skrondal, p. 339)

$$\Phi^1[pi]$$

where

$$p_i = \frac{i-(1/2)}{n}$$

and

$$\Phi = \int_{\infty}^{x} \frac{1}{\sqrt{2\pi}} exp(\frac{-x^2}{2})dx.$$

The qqline() function loads a straight line on the plotted points and connects the 25th and 75th percentiles. If the points of a Q-Q plot lies roughly on the straight line, it confirms that the assumption quantiles originate from a normal distribution is true.

## 3.2 Terminology Employed in Statistical Tests

Statistical hypothesis is a declaration regarding the distribution of one or more random variables. A test for a specific hypothesis is a method adopted to decide whether the hypothesis needs to be rejected or not. The *null hypothesis*, denoted by $H_0$ is the hypothesis that is tested regarding some parameter $\theta$ of a distribution. The *alternative hypothesis* expressed by $H_1$ is a contrary statement concerning $\theta$. In brief, $H_0$ is tested against $H_1$ (Alexander McFarlane Mood, p. 402-403). The parameter space $\Theta$ is divided between $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \varnothing$.

Based on the results of a statistical test, we either reject $H_0$ or fail to reject it. A *type I error* occurs when $H_0$ is rejected even though it is true. In contrast, accepting $H_0$ when it is false is called a *type II error* (Alexander McFarlane Mood, p. 405). A critical value is compared with the numerical value of the test statistic to decide whether $H_0$ is considered to be rejected. The set of the values of the test statistic that induce $H_0$ to be rejected is called *the critical region* ( B. S. Everitt and A. Skrondal, p. 115). *The* *acceptance region* is the complement of the critical region or the rejection region. *The null Hypothesis* is rejected if the value of the test statistic is determined in the rejection region. Oppositely, the $H_0$ is accepted.

The probability at which $H_0$ is rejected is named $the\ significance\ level$ $\alpha$ $\in$(0,1). This is determined before doing the test and is conventionally matched to 0.05 or 0.01 ( B. S. Everitt and A. Skrondal, p. 393). $The\ confidence\ interval$ is the scale of values to which contains the true value of a parameter $\theta$ with a certain probability ( B. S. Everitt and A. Skrondal, p. 99). This probability is referred to as $the\ confidence\ level$ and additionally, the summation of $\alpha$ and the $the\ confidence\ level$ is necessarily 1. Hence, $\alpha$ equals 0.05 or 0.01 means which corresponds to a $confidence\ level$ of 0.95 or 95 percent and 0.99 or 99 percent, respectively. The probability that the test statistic takes on the calculated value, or an even more extreme value, under the assumption that $H_0$ is true, is defined as a $p - value$, calculated by the value of the test statistic( B. S. Everitt and A. Skrondal, p. 346). The null hypothesis is rejected if the p-value less than or equal to $\alpha$.

The $degrees\ of\ freedom$ is employed during calculation test statistics. Its value is the number of independent data points in a sample through the statistic's computation( B. S. Everitt and A. Skrondal, p. 127).

## 3.3 One-way ANOVA

The one-way analysis of variance test or ANOVA is a statistical method applied to examine the equality of the means of k samples in one variable. The value of k is 4 here in this report. The conditions of the test are as follows:

- The samples observed from the k populations are independent and randomly distributed for each case.
- The variances of each distribution of the populations are approximately equal to each other.
- The dependent variable should be roughly normally distributed for each group.

The two hypotheses are,

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k,$$

$$H_1 : \mu_i \neq \mu_j \text{ for at least one } (i, j), \text{ where } i \neq j \text{ and both } i, j \in 1, ..., k.$$

Let $X_{i1}$, ..., $X_{in_i}$ be a random sample of size $n_i$ from the ith normal population, which i = 1, ... , k. Consider that the ith population has mean $\mu_i$ and variance $\sigma^2$. Our object is to test the null hypothesis that all the population means are the same versus the alternative that not all the means are equal (Alexander McFarlane Mood, p. 436),

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{x_i} - \bar{x})/(k-1)}{\sum_{i}\sum_{j}(x_{ij} - \bar{x_i})^2/(n-k)} \geq \text{some constant c},$$

where $n_i$ is the sample size of the ith group, $\bar{x_i}$ is the sample mean of the ith group, $\bar{x}$ is the overall mean for all groups. we consider k as the number of groups, $x_{ij}$ is the jth observation of the ith group, and n is the total number of observations in all groups (Alexander McFarlane Mood, p. 437).

The ratio F is named the variance ratio, or F ratio in some textbooks. The constant c is determined so that the test will have size $\alpha$. The constant c is determined in the way that the test have size $\alpha$, i.e., $P[F \geq c \mid H_0] = \alpha$ . Since $X_i$ is independent of $(x_{ij} - \bar{x_i})^2$, numerator is independent of the denominator. The numerator divided by $\sigma^2$ has a chi-square distribution with (k - 1) degrees of freedom and the denominator divided by $\sigma^2$ has a chi-square distribution with (n - k) degrees of freedom. As a result, F from the equation above has an F distribution with (k - 1) and (n - k) degrees of freedom. We can conclude that the constant c is the (1 - $\alpha$)th quantile of the F distribution with (k - 1) and (n - k) degrees of freedom. We reject $H_0$ if $F \geq c$ (Alexander McFarlane Mood, p. 437).

## 3.4 Paired Samples t-test

Paired Samples t-test is modelized as the dependent samples t-test. This method is employed when the means of two variables are examined. Since it is paired samples or dependent samples test, both variables are sampled the same group of participants; i.e. if $X_1$, ..., $X_n$ and $Y_1$, ..., $Y_n$ are the two samples, then $X_i$ corresponds to $Y_i$. The assumptions of the paired samples t-test are listed below:

- The dependent variable needs to be continuous.
- The samples/groups are examined to be related, i.e. both groups have the same subjects.

- Approximately normally distributed of the differences between the paired values needs to be concerned.
- The differences between the paired values would not contain any outliers.

The two hypotheses are shown below, where $\mu_1$ is the population mean of the first variable and $\mu_2$ is the population mean of the second variable, i.e.,

$$H_0 : \mu_1 = \mu_2,$$

$$H_1 : \mu_1 \neq \mu_2.$$

The paired samples t-test is based on the T-statistic and it can be formalized as follow:

$$s_{\bar{x}} = \frac{s_{diff}}{\sqrt{n}},$$

$$t = \frac{\bar{x}_{diff}}{s_{\bar{x}}},$$

where $\bar{x}_{diff}$ is the sample mean of the differences of the paired values and n is the sample size for each group. it is assumed that $s_{diff}$ is the sample standard deviation of the differences and $s_{\bar{x}}$ is the estimated standard error of the mean. The calculated t-value is compared with the critical t-value with (n - 1) degrees of freedom at the selected significance level $\alpha$. If $t \geq$ critical value, the null hypothesis would be rejected (Kristin Yeager, Preya Bhattacharya, and Victoria Reynolds, 2021).

## 3.5 Bonferroni method

This method is applied with the same assumption of one-way ANOVA, i.e., the underlying random variables are independent and normally distributed with equal variance. The Bonferroni confidence intervals for the difference between two treatment means formulated as below for the m = p(p - 1)/2, i.e.,

$$(\bar{x_i} - \bar{x_j}) - t_{\alpha/(2m)} s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} < \mu_i - \mu_j < (\bar{x_i} - \bar{x_j}) + t_{\alpha/(2m)} s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}},$$

where $\bar{x_i}$ and $\bar{x_j}$ are sample means of random samples $X_1, ..., X_n$ from populations assumed.

As the definition, s is determined the previous standard error and $t_{\alpha/(2m)}$ such that p($t_{\alpha/(2m)}$) $= \alpha/(2m)$, where t has a t distribution with (n - p) degrees of freedom. Two population means, $\mu_i$ and $\mu_j$, are contemplated different at the own $\alpha/m$ significant level if zero does not belong to the blended confidence interval for $\mu_i - \mu_j$. The identical test statistic is as follows:

$$t^*_{ij} = \frac{\mu_i \mu_j}{s\sqrt{1/n_i + 1/n_j}},$$

where $t^*_{ij}$ has a t distribution with (n - p) degrees of freedom (R Development Core Team, p.274).

Due to conducting multiple examinations on the same variables, the chance of executing a type I error increases. For correction, a Bonferroni correction is conducted. Bonferroni correction is known as a conservative test. Although it protects the results from type I error, it is still powerless to type II errors (Statistics Solutions Team).

# 4 Statistical analysis

In this section, the statistical methods described in the previous part are applied to the granted data set to interpret the results.

## 4.1 Examine the independence and normality of the random samples

Initially, in the part of the problem statement, it is discussed that the method of sampling the data from the four cities, Dortmund, Duisburg, Bochum, and Essen, is completely independent, during the applying statistical method, the independence of the random samples is concerned.

Another critical assumption of the statistical test is that the data of these four cities samples are normally distributed. Figure 1 analyses this assumption using a Q-Q plot. This method is applied to simply check the acceptance of normality via its graph. For each figure, the quantiles of the one city are compared to the quantiles of a theoretical normal distribution. In the figure 1, it is shown by using R programming code for precise illustration. As the plot displays, the points adhere to the straight line connecting the 25th and 75th percentiles very strong. The points in the centre of the graphs lie close to

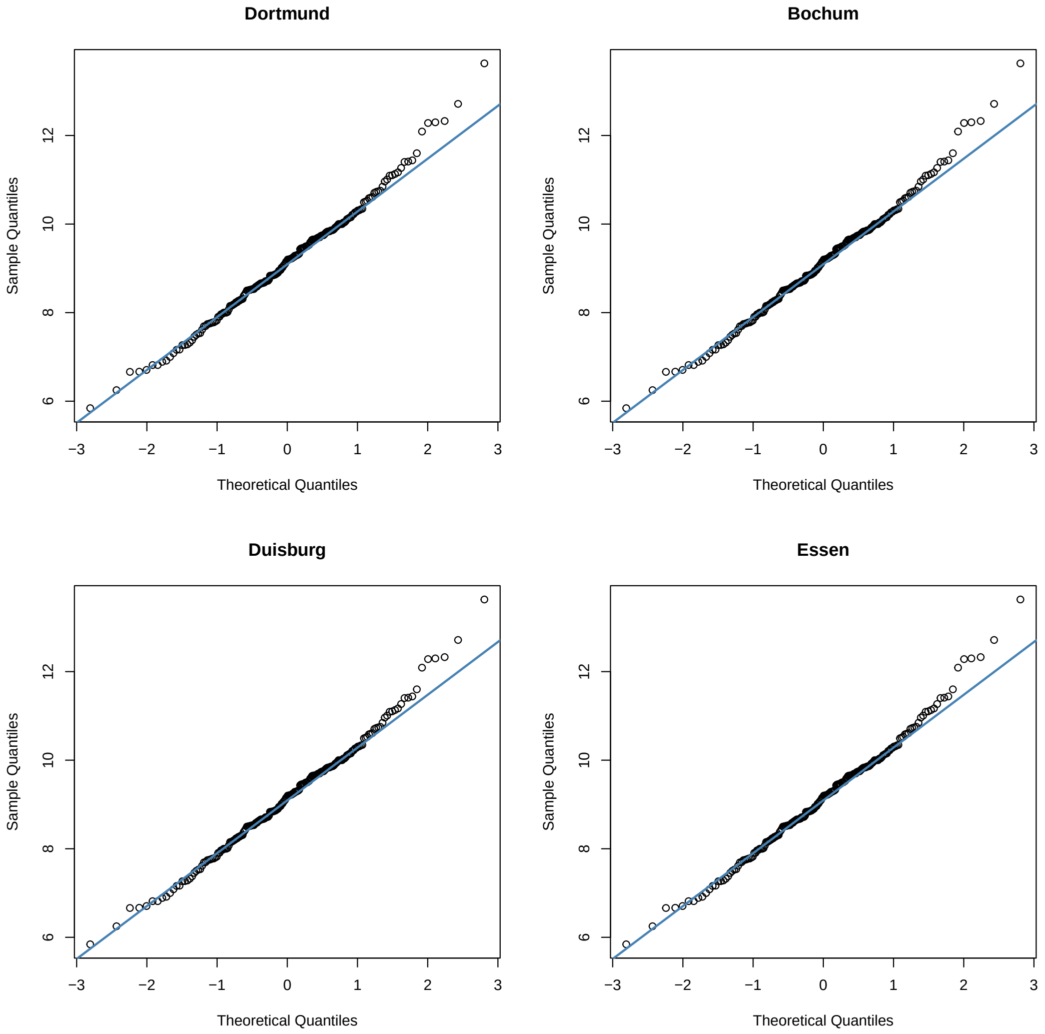

Figure 1: Q-Q plot of the quantiles of the differences of the relative deviations between four sample data of cities, versus normal theoretical quantiles

the line. Only for the last quantile, data points are divergent from the line of checking normality. It can be assumed that all four random samples are normally distributed according to the Q-Q plot, since approximately all figures display the same distribution for all points in the four different plots and their similarity is considerable.

## 4.2 Examine the equality of variances

The table 1 illustrates the values of descriptive distributions of four random samples of four cities, Bochum, Dortmund, Duisburg, and Essen. The result would be that the value amount of variances are remarkably close to each other, with values of 1.33 for Bochum, 1.36 For Dortmund, 1.14 for Duisburg, and 1.19 for Essen. Still, it would be necessary to check the assumption of the equality of variances statistically. Here, the method of box plots by checking the interquartile is implemented in order to check the interquartile range of each distribution.

The figure 2 shows four box plots of for different cities, where y-axis is the value of price per square meters. It is concluded that the value of the variance is approximately is the same following the fact that each box plot interquartile range is comparatively similar to other plots. Hence, applying statistical tests would accomplish in this report for each random sample. It can be represented from table 1 that the means of three variables approximately equal. However, the value of mean for Duisburg is not as similar as the other cities.

Table 1: Values of the descriptive distributions of variables

| Variables | count | mean | Sd | min | q 25 | median | q 75 | max |
|---|---|---|---|---|---|---|---|---|
| Bochum | 50 | 9.15 | 1.33 | 5.84 | 8.36 | 9.18 | 9.74 | 12.70 |
| Dortmund | 50 | 9.53 | 1.36 | 6.66 | 8.54 | 9.55 | 10.50 | 13.60 |
| Duisburg | 50 | 8.62 | 1.14 | 6.67 | 7.75 | 8.66 | 9.45 | 11.10 |
| Essen | 50 | 9.30 | 1.19 | 6.25 | 8.44 | 9.28 | 10.10 | 12.3 |

For Bochum and Dortmund, there are some outliers, which have their influence on the mean and the consequence of statistical tests. In this report, since data points and outliers are countably insignificant, no statistical method is applied to overcome this problem.

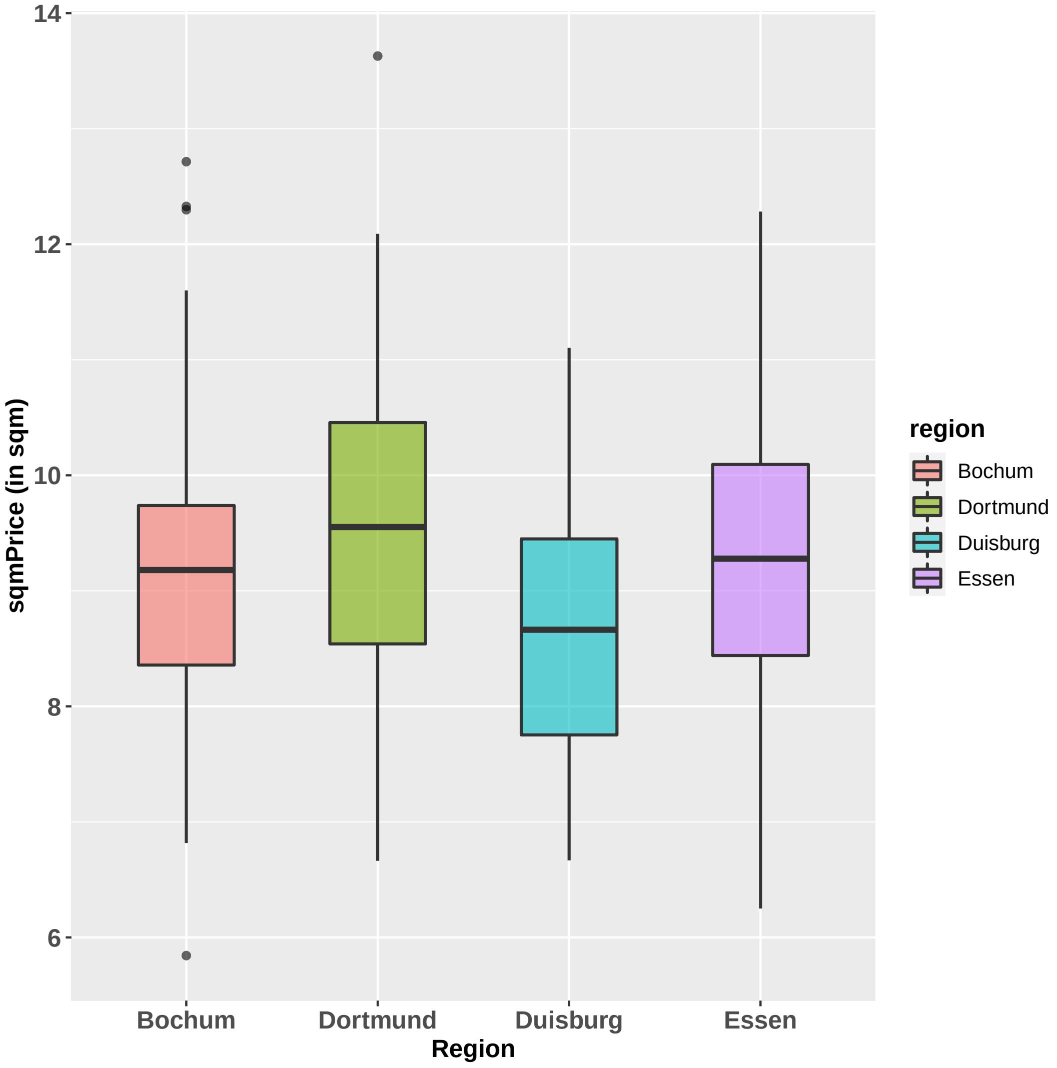

Figure 2: Box plot of the four cities for examining the different between variances

## 4.3 Compare rental price per square meter between the four cities

In this subsection, a one-way analysis of variance test is applied to resolve that if for all four cities, there is a difference in the quantity of rental price per square meter between them. Firstly, all assumptions are checked in the previous two subsections including normality and equality of variances. After these analyses, One-way ANOVA is conducted in order to statistically validate the equality of the average prices of all cities. As a consequence, all four samples have the same average and finally the populations have the equivalent amount of rental price as the samples. The null hypothesis here is determined as the equality of all city means together, i.e., $\mu_{Bochum} = \mu_{Dortmund} = \mu_{Duisburg} = \mu_{Essen}$

against altenative hypothesis that at least there are two cities that their means of rental price per square meter are not equal.

The table 2 illuminates that although the value of rental prices for each city in table 1 is barely close to each other, statistical analysis brings another resolution that since the p-value of 0.0035 is considerably smaller than the value of significance level (alpha = 0.05), consequently, the null hypothesis is rejected. As a whole, it is failed to declare that the global rental price per square meter is the same for all four cities according to the samples given.

Table 2: Results of the one-way ANOVA comparing the mean relative of four cities

| variable | F value | degrees of freedom (k - 1) | degrees of freedom (N - k) | p-value |
|---|---|---|---|---|
| value | 4.68 | 3 | 196 | 0.0035 |

## 4.4 Pairwise differences between rental price per square meter of cities

The pairwise comparisons can be applied with consideration of its assumptions. Firstly, from the problem statement part, since all variables are continuous and related to each other, and besides, the normality is held in the first part of statistical analysis, the pair sample t-test can be employed. Table 3 presents the values of each pairwise test for the given random samples. There is no adjustment method used in this task. The null hypothesis here is that for each pair of the cities, the average rental price per square meter is the same, e.g. $\mu_{Dortmund} = \mu_{Duisburg}$ for two particular cities. Consequently, the null hypothesis for the equality of the average rental price per square meter is accepted for these cities including, Bochum and Dortmund, Bochum and Essen, and Dortmund and Essen, which p-values are 0.1367, 0.5558, and 0.3670, respectively. Hence, All these values are greater than the value of $\alpha = 0.05$. This implies that the differences among the mean relative these cities are not statistically significant. However, the null hypothesis rejected or equivalently alternative hypothesis is accepted for pair of the city, Duisburg, with the other cities. Thus, the average rental price of this city is not equal to others with the result of the p-values 0.0364, 0.0004, 0.0076 for comparison of this city to Bochum, Dortmund, and Essen. Table 3 displays these values in more detail.

In the next part, the method of $Bonferroni$ is applied for solving the multiple testing problem. Firstly,the assumptions of normality and independent random variables in

Table 3: Results of paired Samples t-test comparing the mean of each pair cities without correction

| city | Bochum | Dortmund | Duisburg |
|---|---|---|---|
| Dortmund | 0.1367 | - | - |
| Duisburg | 0.0364 | 0.0004 | - |
| Essen | 0.5558 | 0.3670 | 0.0076 |

addition to equality of variances are checked. Since from the previous parts and using Q-Q plots, they are already examined, the Bonferroni correction is practiced. Here, the reader can recognize that from table 4 the accepting and rejecting the null hypothesis are almost the same from the result of the pair-sample t-test without applying any correction. Table 4 represents that even though the correction of the p-value of comparing Duisburg and Bochum average rental prices per square meter after the adjustment method is considerably different, there are still significant differences between Duisburg with Dortmund and Essen two by two with p-values of 0.0024 and 0.0456, respectively.

Table 4: Results of paired Samples t-test comparing the mean of each pair cities with Bonferroni correction

| city | Bochum | Dortmund | Duisburg |
|---|---|---|---|
| Dortmund | 0.8201 | - | - |
| Duisburg | 0.2182 | 0.0024 | - |
| Essen | 1.0000 | 1.0000 | 0.0456 |

For the other cities Bochum and Dortmund, Dortmund and Essen, and Dortmund and Essen, it can be conclude that it is more likely that the value of means of two populations statically are equal. Therefore, the null hypotheses are accepted. Table 4 shows more values of correction of Bonferroni method on comparison of p-values.

# 5 Summary

The goal of this report is to describe and employ statistical methods for the analysis of a dataset represented by the Immobilienscout24 for the four largest cities of the Ruhrgebiet (Ruhr area). The data includes three variables, including ID, price per square meter, and the region with the names of four cities. There are 200 rows in the data set, 50 properties for each city. Here, two main tasks are discussed for analysing the differences in rental price per square meter of four cities. Moreover, this value is compared between each pair of cities.

For answering these tasks, statistical methods are applied with their introduction and assumptions. For each test, they are partially met. The assumption of normality of samples is evaluated by an experimental graphic, where Q-Q plots are utilised to review for normality assumptions. Similarly, box plots are used to estimate the variation and central tendency of the observations. A descriptive table of values of the distribution of four samples is employed to analyze these assumptions.

By conducting one-way ANOVA test, it is concluded that the null hypothesis, equality of the mean of all cities, are rejected by calculation of p-value of this test. Furthermore, paired samples t-test is applied for the hypothesis that for each pair mean value of the cities, they are equal. It is discussed that although this value is the same for each pair cities of Bochum, Dortmund, and Essen, it is failed to conclude that the same result for Duisburg with other cities. Then, The method of Bonferroni is used for correction of error rate and different results are obtained. After using this method, the mean of rental price per square meter of Bochum and Duisburg, and also all other equal pair cities, are pair to pair equal. However, the rental mean of Duisburg compare to Dortmund and Essen are significantly different, respectively.

It can be concluded that it is helpful to conduct statistical tests for comparing different samples of populations for understanding the properties of populations. For further study, it is useful to have more number of samples and compare its results with this report for suitable results.

# Bibliography

[1] B. S. Everitt and A. Skrondal. *THE Cambridge Dictionary of Statistics.* Cambridge University Press, 2010.

[2] Alexander McFarlane Mood. *Introduction to the Theory of Statistics.* McGraw-Hill Inc., 1974.

[3] CorrieBar. Apartment rental offers in germany. URL `https://www.kaggle.com/corrieaar/apartment-rental-offers-in-germany/`. (visited on 12/04/2021).

[4] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*, 2016. URL `https://ggplot2.tidyverse.org`.

[5] Hadley Wickham, Romain François, Lionel Henry,and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2021. URL `https://CRAN.R-project.org/package=dplyr`. R package version 1.0.7.

[6] Kristin Yeager, Preya Bhattacharya, and Victoria Reynolds. Paired samples t test. URL `https://libguides.library.kent.edu/SPSS/PairedSamplestTest/`. (visited on 12/04/2021).

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

[8] Statistics Solutions Team. Bonferroni correction. URL `https://www.statisticssolutions.com/bonferroni-correction/`. (visited on 12/04/2021).