

TU DORTMUND

INTRODUCTORY CASE STUDIES

Project 2: Comparison of multiple distributions

Lecturers:

Prof. Dr. Sonja Kuhnt

Dr. Paul Wiemann

Dr. Birte Hellwig

M. Sc. Hendrik Dohme

Author: Aishwarya Dinni

Group number: 4

Group members: Vinay Kumarmath, Charishma Chinthakayala,
Gautham Prasad Kundapura, Hepsiba Komati

December 10, 2021

Contents

1	Introduction	1
2	Problem statement	1
2.1	Description of data set	1
2.2	Project objectives	2
3	Statistical methods	2
3.1	Assumptions for testing	3
3.2	Inferential statistics and hypothesis testing	4
3.3	Analysis of Variance - ANOVA method	6
3.4	Pairwise t-test	8
3.5	Multiple testing problem and Bonferroni correction	9
4	Statistical analysis	10
4.1	Analysis for assumptions	10
4.2	Global test for the relationship among means	12
4.3	Pair-wise comparison of means	13
5	Summary	14
	Bibliography	16
	Appendix	17
A	Additional figures	17
B	Additional tables	17

1 Introduction


In today's busy life it is hard to find a perfect place to purchase or to rent. In an ever-changing real estate market, it is crucial to gather information such as location, price, built-in area, and many other attributes to make a decision. The lack of housing overall affects everyone, but not everyone equally. There are many ways to search for a place, through agents or online websites such as ImmobilienScout24, Immowelt.de, Meinstadt.de, and many more.

As long as there is an investment, sales, there is always a boom in the scope of real estate. Thus, analysing the prices of different regions to increase success, can be of interest. A small set of data containing the rental price per square meter for 200 properties located in the four largest cities of Ruhrgebiet are examined. To understand the difference in the average prices between the cities, a comparison of multiple distributions across each city is put together. Three conditions are examined over each price per square meter: homogeneity using descriptive analysis and box plots, normality using Q-Q plots, and independence. Based upon these assumptions and using one-way ANOVA (Analysis of variance), a global test is done to establish the relationship between the prices of different cities. In addition to this, the price per square meter of each city is compared using a pairwise t-test. To address the multiple testing issue, the test values are tuned using the Bonferroni method. The obtained results suggest that there is a significant overall difference between the prices per square meter of each city.

Section 2 consists of an illustration of the data set and the key problems. Whereas section 3 includes information on the testing methods used, and section 4 shows the observations and inferences on analysing the given data. In the final section, central results are reviewed and summarised.

2 Problem statement

2.1 Description of data set

The given set of data comprises rental price per square meter for 200 properties located in the four largest cities of the Ruhrgebiet (Ruhr area). The mentioned data is acquired from the website ImmobilienScout24 ImmobilienScout developer. 

The mentioned data is stratified over four cities- Dortmund, Bochum, Essen, Duisburg. There is a total of 200 unique rows, with three variables-ID, sqmPrice, and regio2.

- ID (numeric - continuous): has a distinctive number for each row. As there is no data missing, all the values are included in the evaluation.
- The next variable is sqmPrice (numeric - continuous): the rental price per square meter for 200 properties located in one of the cities mentioned earlier.
- The last variable is the regio2 (categorical:) shows the city to which each sqmPrice belongs.

With the maximum cost of 13.628571 and the minimum, 5.842550, the mean of all the individual rental prices per square meter in the data is 9.148971; The table 4 in the Appendix shows the calculations on parameters - count, mean, standard deviation, minimum, maximum and others of the variable sqmPrice.

2.2 Project objectives

This project aims to compare the multiple distributions of the variable; sqmPrice. The data is analysed in detail using the statistical tests, measures, and graphical representations that are discussed in the succeeding sections. The given data is grouped by city. The initial analysis is done to check if the conditions for assumptions - homogeneity, normality, independence - are satisfied. Once the assumptions are set up, a global test called ANOVA is done to understand the overall difference in the price among the four cities. Later, considering all pairs of cities as two-sample t-tests the pairwise difference between the prices is tested. Furthermore, to address the multiple testing issues, the obtained test results are tuned with the Bonferroni method. At last, the results are deduced and a comparison between the outcomes with and without adjusting for the multiple testing is done.

3 Statistical methods

This section consists of several statistical tests and measures, which are later used for analysing the data set. To understand the variables and their properties, initial probing of the given data file-*ImmoDataRuhr.csv* is done using Microsoft Excel (Version 2010). For all calculations and visualisations, Python (Version 3.7.3) van Rossum (2021)

language is scripted with Jupyter Notebook (Version 6.0.0) launched using the software - Anaconda Navigator 3. The packages utilized are - pandas (<https://pandas.pydata.org/>), matplotlib (<https://matplotlib.org/>), seaborn (<http://seaborn.pydata.org/>), scipy (<https://www.scipy.org/>), statsmodels (<https://www.statsmodels.org/>), itertools (<https://docs.python.org/3/library/itertools.html>).

3.1 Assumptions for testing

The term homogeneity of variance refers to the fact that the level of variance for a particular variable is constant across the sample. The variance of the outcome variable(s) should be the same in each of the groups present in the collected groups of data. This is one of the important assumptions for hypothesis testing. When this assumption is not satisfied, the probability of falsely rejecting the null hypothesis increases. Homogeneity for groups against one another can be evaluated by examining the interquartile data in the box plots. This visualisation stipulate the spread rather than statistical confirmation of homogeneity



Normality: implies that the sample data follows a normal distribution. The normal/Gauss distribution is a type of continuous probability distribution for a real-valued random variable with parameters μ and σ . Where μ is the mean of the distribution while σ is the standard deviation. To plot a Q-Q plot, the data is ordered in ascending order for a sample size n , and their quantiles are calculated as sample quantiles. A normal distribution curve is drawn and divided into $n + 1$ segments or equally sized areas. Then using the z-table, the corresponding z-value is calculated for each of the segments. With this z-value, we can deduce how many standard deviations a particular segment is away from the mean.

The z-value can be calculated using $Z = \frac{x - \mu}{\sigma}$, where x is the observed value, μ is the mean of the sample and σ is the standard deviation. The theoretical quantiles are the z-values that correspond to the quantiles of an ideal normal distribution for a given μ and σ . With the theoretical quantiles along $x - axis$ and sample quantiles along $y - axis$, a graph is drawn, the instances are plotted against the axes, and the $x = y$ line is drawn. If the points strictly follow a linear trend with respect to the straight line, it can be assumed that the data follows a normal distribution. And if the points are scattered or follow a discrete trend, it indicates that there could be skewness and the distribution can be different.

3.2 Inferential statistics and hypothesis testing

Inferential statistics is a method where quantitative decisions are made based on observations in the given data. There are two types of statistical tests: parametric and non-parametric tests.

- Parametric tests: are those that make assumptions about the parameters of the population distribution from which the sample is drawn.
- Non-parametric tests: These tests are not based on any assumptions about the distribution. Thus it can be said that they are distribution-free and, as such, can be used for non-Normal variables

This project uses parametric tests. Hypothesis testing is done based on some assumptions which include sample size, shape/type of the population distribution, level of measurement of the variable, and method of sampling. A hypothesis is a proposed explanation for a phenomenon that is based on a parameter, θ . And then it is tested to determine if there is enough evidence to support the proposed hypothesis. The proposed statement to be tested is called as the Null Hypothesis, $H_0 : \theta \in \theta_0$. The statement contradicting to the null hypothesis is called the Alternate Hypothesis, $H_a : \theta \in \theta_a$ such that $\theta_0 \cap \theta_a = \phi$ and $\theta_0 \cup \theta_a = \theta$. Accepting the null hypothesis indicates that the null hypothesis is true or that there is not enough data to disprove/reject the null hypothesis (Hartmann, Krois and Waske, 2018, p. 67). Accepting or rejecting a hypothesis involves a criterion called Test statistic, and this is calculated using measures in the observation. A test statistic is a number calculated from the data set, which is obtained by measurements and observations.

It is a function $t(x)$ of the given sample values x_1, \dots, x_n , possibly in such a way that the difference between distribution $f(t | H_0)$ and distributions H_a are as large as possible. For a defined rejection region R , H_0 is rejected if $t(x) \in R$. If $t(x) \notin R$, we fail to reject H_0 , in favour of H_a (Hartmann, Krois and Waske, 2018, p. 67). Hypothesis tests are also divided into two types, one-tailed and two-tailed tests, illustrated in Figure 2 in Appendix.

- A one-tailed test determines if there is a difference between groups in a specific direction.
- The two-tailed test uses both the positive and negative tails of the distribution for the rejection region, testing for the possibility of positive or negative differences.

Rejection region is defined with a Critical value C . A null hypothesis is either rejected or accepted based on comparison of the test statistic $t(x)$ with the critical value C . For a one-sided test, the rejection region is either $R = (-\infty, C)$ or $R = (C, +\infty)$ and for a two-sided test, the rejection region is $R = (-\infty, C_1) \cup (C_2, +\infty)$, where C_1 and C_2 are the lower and upper limits of C .



An assumption made about a hypothesis with the given observations of the data can be wrong in the real world. Thus it can be said that there is some uncertainty in the decisions arrived at with hypothesis testing. There are two types of errors that are possible while performing a hypothesis test: Type I and Type II errors.

- Rejecting a true H_0 (false positive) leads to type I error. A type I error detects an effect that is not present. The probability of a type I error is called the Significance Level α of a hypothesis test.
- Failing to reject a false H_0 (false negative) leads to type II error. A type II error fails to detect an effect that is present. The probability of a type II error is denoted as β .

A test procedure can be constructed such that the risk of rejecting a true H_0 or the significance level α is small. Smaller the α , larger the β or the probability of not rejecting a false H_0 . The errors are compared in the table 2, in appendix.

If H_0 is rejected, the data provides sufficient evidence to support H_a . If H_0 is not rejected, the data does not provide sufficient evidence to support H_a . If the hypothesis test is performed at the significance level α , it means that the test results are statistically significant at the α level and vice versa. The observed test statistic $t(x)$ is compared to the critical value C . If $t(x)$ is more extreme than C , H_0 is rejected. If $t(x)$ is not as extreme as C , H_0 is not rejected.

C , calculated based on the given significance level α and the type of probability distribution of the idealized model, divides the area under the probability distribution curve into the rejection region(s) R and non-rejection region. In a two-tailed test, H_0 is rejected if $t(x)$ is either too small or too large as compared to the critical value C . For one-tailed tests, in a left-tailed test, H_0 is rejected if $t(x)$ is smaller than C and in a right-tailed test, H_0 is rejected if $t(x)$ is larger than C . Considering a 5% risk of concluding that a difference exists when there is no actual difference (95% confidence interval), the significance level α is set to 0.05 such that $\alpha \in (0, 1)$ (Hartmann, Krois and Waske, 2018, p. 68). Alternatively, another way to interpret the results of a hypothesis test is by

using a continuous parameter, P-value. The P-value p is the probability of observing the sample data at least as extreme as the obtained test statistic. Evidence against H_0 is strengthened with a smaller p .

If $p \leq \alpha$, H_0 can be rejected. Else, for $p > \alpha$, H_0 cannot be rejected. Assuming that the test statistic t falls between zero and infinity with a critical region ($t > t_C$) and the distribution under H_0 as $f_0(t)$, p as a function of t is calculated

$$p(t) = 1 - \int_0^t f_0(t') dt' = 1 - F_0(t), \quad (1)$$

where p is a unique monotonic function of t , p can be considered as a normalized test statistic equivalent to t (Bohm, Gerhard and Zech, Günter, 2010, p. 248).

3.3 Analysis of Variance - ANOVA method

The Analysis of variance (the ANOVA) method can be used to understand the statistical difference between the means of two or more independent groups. The mean of the continuous dependent variable is compared across the categories of the independent variable. This method works based on assumptions that:

- The samples are drawn is independent of each other.
- Each group sample is drawn from a normally distributed population.
- The variance is common among all the population (homogeneity)

The residuals ϵ_{ij} are the difference between each group mean and the grand mean, such that the grand mean $\bar{X}_{ij} = \mu_{ij} + \epsilon_{ij}$ for each group. ϵ_{ij} are assumed to be identically and independently distributed with mean μ and standard deviation σ . A one-way ANOVA technique considers a single factor and its effects on the sample are observed. Whereas, the two-way ANOVA is performed in cases when the given data set is classified under two different independent factors. This experiment uses the one-way ANOVA method (Black, 2019, p. 409).

The null hypothesis states that the mean is equal among all groups and the alternate hypothesis states that not all means are equal. For k groups, the hypotheses are defined as $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ and $H_a : \text{means are not all equal}$. Holding the above-discussed assumptions, independent random samples are taken from each group, sample means for each group are computed and a comparison is made for variation of sample means

between the groups to variation within the groups. Based on a test statistic, it can be decided if the means of the groups are all equal or not. Quantitative measures of variability are required here. So, the total variability is partitioned into two parts, to account for between-group variability and within-group variability. First, the measures of variability are calculated. The sum of squares total (SST) measures the total variability of the variable and is calculated as

$$SST = \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2)$$

where x_i is the observations in the samples and \bar{x} is the overall mean of all samples. The sum of squares groups (SSG) measures variability between groups and corresponds to squared deviation of group means from the overall mean, weighted by sample size. It is calculated as $SSG = \sum_{j=1}^n n_j (\bar{x}_j - \bar{x})^2$, where n_j is the sample size for group j and \bar{x}_j is the mean of group j . The sum of squares errors (SSE) measures the variability within groups, indicating variability that cannot be explained by group variables. The formula to calculate SSE is $SSE = \sum_{j=1}^n (n_j - 1) s_j^2$, where s_j^2 is the variance of group j . It can also be calculated as $SSE = SST - SSG$ (Black, 2019, p. 408).

Further, measures of mean variability are calculated. To get average variability, the calculated measures of variability are scaled by degrees of freedom (df), an expression of sample size n . The degrees of freedom are defined for each partition of variability.

- For total variability, degrees of freedom is calculated as $df_T = n - 1$, for overall sample size n .
- For the in between-group variability, degrees of freedom are calculated as $df_G = k - 1$, for k number of groups.
- For the within-group variability, degrees of freedom are calculated as $df_E = n - k$.

With the calculated measures of variability and degrees of freedom, the mean squares for the in between-group variability and the within-group variability are calculated. The average variability in between and within groups gives the total variability scaled by the associated degrees of freedom. The mean in between-group variability (mean squared variation due to group) is $MSG = \frac{SSG}{df_G}$. The mean within-group variability (mean square of error) is $MSE = \frac{SSE}{df_E}$ (Black, 2019, p. 408).

Finally, the mean variation between the groups (corresponding to the independent variable) - MSG is compared to the variation within the group (corresponding to the residuals) - MSE . The ratio of MSG and MSE is the F-statistic denoted by F .

$$F = \frac{MSG}{MSE} = \frac{\frac{1}{k-1} \sum_{i=1}^n n_j (\bar{x}_i - \bar{x})^2}{\frac{1}{n-k} \sum_{i=1}^n (n_j - 1) s_j^2} \quad (3)$$



The F-statistic follows F-distribution with $df = k - 1, n - k$. A high F value might suggest that the variation among the groups dominates over the variation within the groups, that is, there is strong evidence against H_0 . A low F value denotes that the variation among groups is smaller than variation within groups, suggesting little or weak evidence against H_0 (Hartmann, Krois and Waske, 2018, p. 91).

For the one-way ANOVA, the F-statistic can be converted into a P-value p . An F-table that gives the area to the right of F in the F-distribution curve corresponding to the degrees of freedom is used. With $k - 1$ in the numerator and $n - k$ in the denominator, the area of the curve corresponding to the upper and lower values between which F falls can be found. This gives the range within which p lies. To get a precise p , an F-distribution calculator can be used. When p is greater than the significance level α , there is no significant evidence against H_0 . For a small $p(p \leq \alpha : H_0$ can be rejected), the data provides convincing evidence that at least one pair of group means is different from each other. For a large $p(p > \alpha$: failing to reject H_0), the data does not provide convincing evidence that at least one pair of group means is different from each other.

3.4 Pairwise t-test

A function Pairwise t-test is defined to perform two-sample t-tests with every pair of a categorical group. The following assumptions are made: Data in each group must be obtained through a random sample from the population. Data in each group is normally distributed and their values are continuous. The variances for the two independent groups are equal. Assuming that standard deviations of the two populations in questions are equal but unknown, the means of the two populations are statistically tested. The two-sample t-test examines if there is a statistically significant difference between the means of two independent groups. For k groups, $M = \frac{k(k-1)}{2}$ paired tests are done.

The null hypothesis states that the difference between the two means is zero (means are equal). The alternative hypothesis states that the difference between the two means is not equal to zero (means are not equal). For two groups 1 and 2, the hypotheses are defined as $H_0 : \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$ and $H_a : \mu_1 \neq \mu_2$ or $\mu_1 - \mu_2 \neq 0$.

For M tests, H_0^1 vs H_a^1 , H_0^2 vs H_a^2 ..., H_0^M vs H_a^M

when the means are unequal, a mean can either be greater and lower than the other (the differences could go in either direction) and so, two-tailed t-tests are used (Hartmann, Krois and Waske, 2018, p. 73).

Considering the population means to be equal ($\mu_1 - \mu_2 = 0$) and assuming equal population variances (pooled) to use the common estimated sample standard deviation, t-statistic (t-value) t is calculated as

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{1}{n_1 + n_2 - 2}(S_1^2(n_1 - 1) + S_2^2(n_2 - 1))\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}} \quad (4)$$

where \bar{x}_1 and \bar{x}_2 are the sample means, n_1 and n_2 are the sample sizes, s_1 and s_2 are the standard deviations of group 1 and group 2, respectively (Black, 2019, p. 355). For this two-tailed test, with a significance level α fixed in prior, the critical value is calculated as $(-\frac{\alpha}{2}, \frac{\alpha}{2})$. If the calculated t-statistic t falls between $(-\frac{\alpha}{2}, \frac{\alpha}{2})$ - within the critical region - H_0 is failed to be rejected. If $t < -\frac{\alpha}{2}$ or $t > \frac{\alpha}{2}$ - outside the range of the critical region - H_0 is rejected (Hartmann, Krois and Waske, 2018, p. 73). Another way to interpret the results of this t-test is by calculating the p-values. Using t-distribution table, with degrees of freedom $df = n_1 + n_2 - 2$, the area of the curve corresponding the t-value can be found. Since this is a two-tailed test, the p-value p is twice the tail area obtained from the t-table. When $p \leq \alpha$, H_0 can be rejected. Otherwise, the studied data fails to reject H_0 .

3.5 Multiple testing problem and Bonferroni correction

When a problem examines multiple hypothesis tests, the probability of wrongly rejecting a null hypothesis can increase, which increases the type I error. The problem with multiple testing is that the more hypotheses are tested, the more likely it is to incorrectly reject the null hypothesis. Thus, to compensate for how many inferences are made, multiple comparison methods need a higher significance threshold α for each comparison. If k groups are tested in pairs, two at a time, $M = \frac{k(k-1)}{2}$ paired comparisons are possible.

For a family of M tests, the probability of not making a Type I error for the whole family is $(1 - \alpha)^M$. On the other hand, the probability of making one or more Type I errors on the family of tests, the family-wise error rate is $FWER = 1 - (1 - \alpha)^M$. $FWER$ is calculated to control the type I error rate. This problem can be accounted for using statistical methods like the Bonferroni correction method, Tukey multiple-comparison

method, Holm–Bonferroni method, Šidák correction. For normally distributed variables with equal variances, the Bonferroni adjustment is done in this study (Black, 2019, p. 418).

The Bonferroni correction rejects the H_0 for each $p_i \leq \frac{\alpha}{C}$, thereby controlling the *FWER* at $\leq \alpha$. It is done either by adjusting the p-values, multiplying all p with the number of tests M , or comparing the adjusted p-values against the set α . Another way is by dividing the significance level of each hypothesis by $M(\alpha/M)$ and comparing all the raw p-values with the corrected significance. The function in the software used rounds the p-values to 1.00 if the adjusted values are ≤ 1.00 (Hartmann, Krois and Waske, 2018, p. 93).

This correction method does not require any assumptions about dependence among the p-values or about how many of the null hypotheses are true. Bonferroni correction might reduce statistical power. Applying the Bonferroni adjustment increases the probability of producing false negatives (type II error).

4 Statistical analysis

Detailed analysis and comparison of multiple distributions in the given data set employing the methods discussed in the previous section are presented in this section.

4.1 Analysis for assumptions

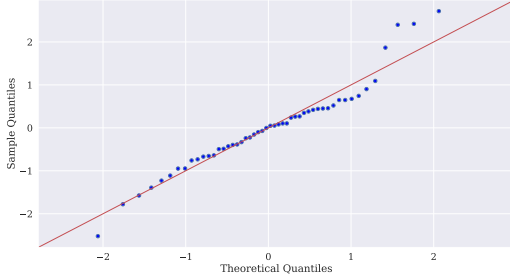
Analysing the data set, it was learned that the dependent variable, `sqmPrice` is continuous and the independent variable `regio2`, is categorical. `sqmPrice` depends on `regio2`. The mean for each group is also analysed with Table 5 in Appendix.



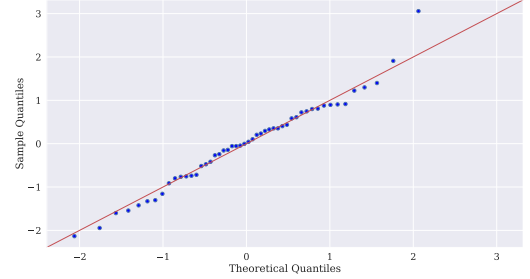
As it can be observed that there is no clear relationship between each rental property in the data set. Thus, within the scope of this study, the data is assumed to be independent, in spite of uncertainties. i.e. it is assumed that the samples are drawn independent of each other.

The normality is examined using a Q-Q plot and this is achieved by comparing the observed and predicted proportions. The calculation is done for theoretical and sample quantiles. Later these are plotted against each other in a straight line. The generated Q-Q plots for each city are shown in Figure 1. From this plot, it can be noticed that the

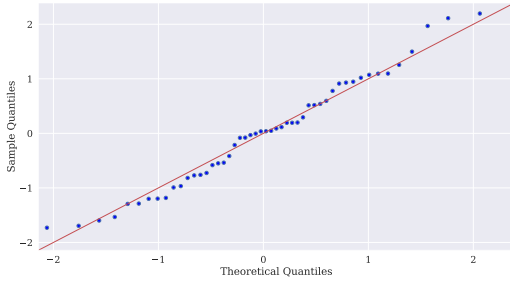
city, Duisburg tends to follow a less linear trend when compared to Q-Q plot for other cities. The data points in other cities- Dortmund, Essen, and Bochum largely fall on a straight line. In other words, it is assumed that the data is normally distributed within the scope of our study. That is, the assumption of normality is not violated.



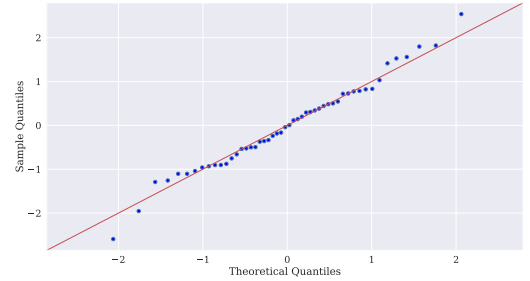
(a) Bochum



(b) Dortmund



(c) Duisburg



(d) Essen

Figure 1: Q-Q plots for each of the cities

The spread in the data is visualized using a box plot, and The homogeneity of variance is measured by comparing the empirical variances and standard deviations calculated for each group. Table 5 in Appendix exhibits the sample variances, standard deviations, and IQRs for each city. The box plots for each group or city are shown in Figure 2.

Based on the height of the boxes in this plot, the interquartile ranges (IQR) for each city are almost comparable, though not the same. The calculated variances are also close to each other on an overall scale. The sample IQR and standard deviation for Duisburg and Essen seem almost equal, although the median of Essen appears to be higher than that of Duisburg. The IQR for Bochum has the lowest value of 1.381 and on the other hand, Dortmund has the highest IQR value of 1.916. The sample standard deviation of Bochum and Dortmund seems to be similar. Duisburg has the lowest sample standard deviation of 1.141. While Dortmund has the highest sample standard deviation of 1.355. With these findings, it is indeed difficult to draw any definite conclusions. The

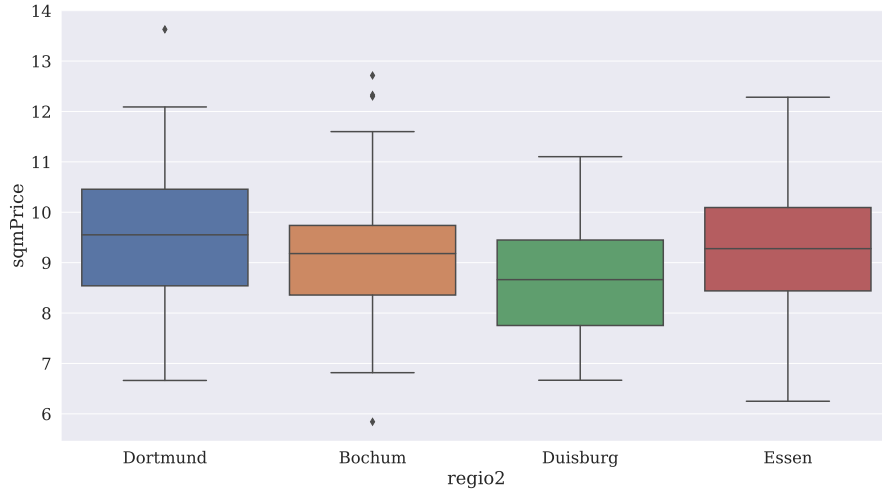


Figure 2: Box plot for each of the city

box plots do not contradict the homoscedasticity assumption. Although, because each of the groups has a small number of samples, it is assumed that the population variances are equal (homogeneous) for this study.

4.2 Global test for the relationship among means

In this sub-section, a global test for comparing the mean of each city with the underlying overall mean is carried out. The prices of each city are tested statistically to see if they are significantly different or not. The null hypothesis is postulated as the means of all the cities being equal. The alternate hypothesis is taken as not all the means are equal.

$$H_0 = \mu_{Do} = \mu_{Bo} = \mu_{Es} = \mu_{Du} \quad \text{and} \quad H_a : \exists \quad i \neq j \quad \mu_i \neq \mu_j$$

The significance level α is set to 0.05 indicating a 5 percent risk of concluding that a difference exists between the means when there is no actual difference. Presuming that all the assumptions made in sub-section 4.1 are true, and the original least-squares model is fit to the price per square meter of properties in each city. The sum of the errors or the residuals and the mean sum of squares is calculated for the ANOVA method. The mean sum of squares and the sum of errors/residuals are calculated for the ANOVA method. The results of the ANOVA are shown in Table 1, where sum_{sq} is the sum of squares, df is degrees of freedom, F is the F-statistic. The two-tailed F-statistic (4.6810872445745595) is converted into its one-way p-value ($p = 0.003$) using a function. This calculated p-value

is ~~significantly~~ smaller than $\alpha = 0.05$. In this test H_0 is rejected, as $p < \alpha$. The data shows strong evidence that roughly at least one pair of group means differs from each other at the 5% significance level. The response variable, sqmPrice, is influenced by the levels in the regio2 variable. The variation between the groups dominates over the variation within the groups.

Table 1: ANOVA output

	sum_sq	df	F	PR(>F)
regio2	22.145	3.000	4.681	0.004
Residual	309.077	196.000	NaN	NaN

4.3 Pair-wise comparison of means

In this sub-section, pairwise tests for comparing the means of each pair of cities are done. Considering all pairs of cities it is tested whether there exists a statistically significant pairwise difference between the mean price per square meter. A two-tailed two-sample t-test is conducted as the difference in the mean can tend to either a positive or a negative direction, when means are unequal. The null hypothesis is postulated as the means of the two city groups in question being equal. The alternate hypothesis is stated as the means of the two city groups not being equal. $M = 6$ tests are conducted for each unique pair of groups -

$$H_0^1 \text{ vs } H_a^1, H_0^2 \text{ vs } H_a^2 \dots, H_0^6 \text{ vs } H_a^6$$

$$H_0 : \mu_i = \mu_j \text{ and } H_a : \mu_i \neq \mu_j$$

The significance level α is set to 0.05 indicating a 5% risk of concluding that a difference exists between means when there is no actual difference. Holding all assumptions made in sub-section 4.1 true, pairs of cities are generated and a t-test is iterated over every pair, as a pairwise t-test.

The corresponding p-values are calculated and rounded off to three decimal places for each of the t-statistics and a decision is made for each pair on comparison with α . To address the multiple testing problem, the Bonferroni correction is made by multiplying all the p-values with $M = 6$. Decisions are made for rejecting H_0 based on these corrected p-values. When $p \leq \alpha$, the data has sufficient evidence against H_0 (reject H_0). When $p > \alpha$, there is weak or no evidence against H_0 with the given observations. The results with and without the Bonferroni correction are compared. The 6 pairs of cities,

their corresponding p-values, adjusted p-values, and decisions for rejection based on raw values and corrected values are all summarised in Table 3 in the appendix, ordered by increasing p. Before correction, for three t-tests out of six, $p > \alpha$: Dortmund and Bochum (0.165), Dortmund and Essen (0.375), Bochum and Essen (0.557).

For these three pairs of cities, there is not enough evidence to reject H_0 , meaning that the difference between their means may be significantly less, or that their means may be equal. After adjusting the p-values with the Bonferroni method, one other pair in addition to the aforementioned three pairs have their $p - values > \alpha$: Bochum and Duisburg (0.035). This pair had $p \leq \alpha$ before the Bonferroni correction. This adjustment prevented one out of the 3 tests from being wrongly rejected, controlling the increase of type I error. On the other hand, with this correction, the probability of type II error also increases.

The results without the correction (raw p-values) hint that for three pairs of cities, the mean could be equal. That is, this data gives sufficient evidence to suggest that there is no significant difference between the means in these three groups. For the other three groups, the data suggest that the means are not equal or that there exists a significant difference between the means. The results obtained after the Bonferroni correction (adjusted p-values) hint that for four pairs of cities, the mean could be equal. That is, this data gives sufficient evidence to suggest that there is no significant difference between the means in these four groups. For the other two groups, the data suggest that the means are not equal or that there exists a significant difference between the means.

5 Summary

Real estate is the most valued profession. Especially in Germany, finding a suitable house is quite tough. Every month with the search, purchase, financing, selling, renting, or moving into a new home over 20 million people go through their marketplace or web applications. The real estate investment market is considered a haven and is gaining momentum over the course of the [yearvon Erdély \(2021\)](#). Thus it can be said that real estate management is a booming field and can be improved by analysing certain specific related features. The main objective of this project was to examine the difference in prices per square meter of four different cities of the Ruhr area. Statistical

testing approaches were used to analyse and compare distributions and means, and then assumptions of homogeneity, independence, and normality were also examined.

The given data is a small slice of larger data that was posted in Immobilienscout24 as of February 20, 2020. This data consists of rental offers for properties throughout Germany. Thus it was difficult to arrive at firm conclusions. However, equality of variance was visualised using box plots and normality was checked with Q-Q plots, and the properties were assumed to be true, for all the tests. A basic descriptive analysis of data was performed. the significance level is fixed at 0.05. A global ANOVA test was done to investigate the difference between prices per square meter for four different cities. The test results indicate that there is an overall significant difference between mean prices for each city. Later, a pairwise t-test was conducted to study if the pairwise difference between sqmPrices of properties exists. A two-tailed t-test was conducted for each of the 6 sport team pairs.

The obtained results suggested that the possibility that the means are equal cannot be denied for that level of significance. Multiple testing problem was addressed and the Bonferroni correction was done. It was observed that with adjusting, four pairs of cities, including the one from the former (without correction), seemed to be equal. Therefore, by using Bonferroni correction type I error was controlled, but it could lead to an increase in type II error.

It is important to be aware of some common misinterpretations of the statistical tests used. If the assumptions for the tests do not hold, there could be an increased error rate. Rejecting the null hypothesis does not make the alternate hypothesis true, by default. It just means that there is not enough evidence favoring the null hypothesis. We are not 95% confident that the true parameter is found to be within an interval. Rather, 95% of such intervals obtained would capture the true value of the population mean. The p-value does not denote the probability of a hypothesis being true. It is the probability of obtaining an effect at least as extreme as the one in your sample data, for a true hypothesis. While using the Bonferroni correction to control type I error, the chance of type II error occurring should be considered Ovens (2018).

It is essential that the pricing of each property is analysed for every real-estate management system. A property with a price per square meter below/ above the average is influenced by city, size, location, and many other attributes. To boost a real estate managing company's potential and productivity, factors related to property search (such as SqmPrice) should be studied extensively and with large-scale data.

Bibliography

- Ken Black. *Business statistics: for contemporary decision making*. John Wiley & Sons, 2019.
- Bohm, Gerhard and Zech, Günter. *Introduction to statistics and data analysis for physicists*. Desy Hamburg, 2010.
- Hartmann, Krois and Waske. E-Learning Project SOGA: Statistics and Geospatial Data Analysis. <https://www.geo.fu-berlin.de/en/v/soga/index.html>, 2018.
- Immobilienscout developer. Immobilienscout24. <https://www.immobilienscout24.de/unternehmen/>. [Online; accessed 09-december-2021].
- Matthew Ovens. Common Misconceptions about Hypothesis Testing. <https://www.yourstatsguru.com/epar/our-publications/common-misconceptions-about-hypothesis-testing/>, 2018. [Online; accessed 09-december-2021].
- Guido van Rossum. Python: A dynamic, open source programming language. <https://www.python.org/>, 2021.
- Prof. Dr. Alexander von Erdély. CBRE GmbH. <https://www.cbre.de/en/research/Germany-Real-Estate-Market-Outlook-2021---Update-H1>, 2021. [Online; accessed 09-december-2021].

Appendix

A Additional figures


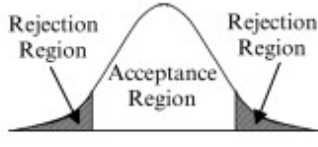

One-Tailed Test (Left Tail)	Two-Tailed Test	One-Tailed Test (Right Tail)
$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X < \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X \neq \mu_0$	$H_0 : \mu_X = \mu_0$ $H_1 : \mu_X > \mu_0$
		

Figure 3: Types of hypothesis tests

B Additional tables

Table 2: Type I and II errors

	when H_0 is true	when H_a is true
rejecting H_0	Type I error (α)	correct decision ($1 - \beta$)
failing to reject H_0	correct decision ($1 - \alpha$)	Type II error (β)

Table 3: Pairwise t-tests output

	City Pair	p-value	Adjusted p-value	Rejected	Rejected_Adjusted
0	Dortmund & Bochum	0.165	0.988	False	False
1	Dortmund & Duisburg	0.000	0.003	True	True
2	Dortmund & Essen	0.375	1.000	False	False
3	Bochum & Duisburg	0.035	0.209	True	False
4	Bochum & Essen	0.557	1.000	False	False
5	Duisburg & Essen	0.004	0.027	True	True

Table 4: Statistical description of variables

	ID	sqmPrice
count	200.000	200.000
mean	6219.545	9.149
std	3548.860	1.290
min	169.000	5.842
25%	3213.250	8.286
50%	6101.000	9.183
75%	9415.7500	9.896
max	12067.000	13.629

Table 5: Sample mean, variance, IQR, and standard deviation for each city

regio2	ID	Variance	sqmPrice	IQR	Standard deviation
Bochum	6199.76	1.756	9.150398	1.381138	1.325280
Dortmund	6596.06	1.838	9.525716	1.916166	1.355733
Duisburg	6000.52	1.302	8.621169	1.696234	1.141194
Essen	6081.84	1.411	9.298602	1.654103	1.187856