Exercise 6

a) We want to test whether the mean profits of pharmaceutical and computer companies differ statistically significantly. Therefore, this is a test of equality of two population means with independent samples. To use the parametric t-test with independent samples, the following assumptions must be met:

1. Our samples should be randomly selected.
2. There should be no extreme values in the sample data of each population exceeding 10%.
3. Each population should be adequately described by the normal distribution.

The first assumption relates to how we selected our samples, and it is satisfied.

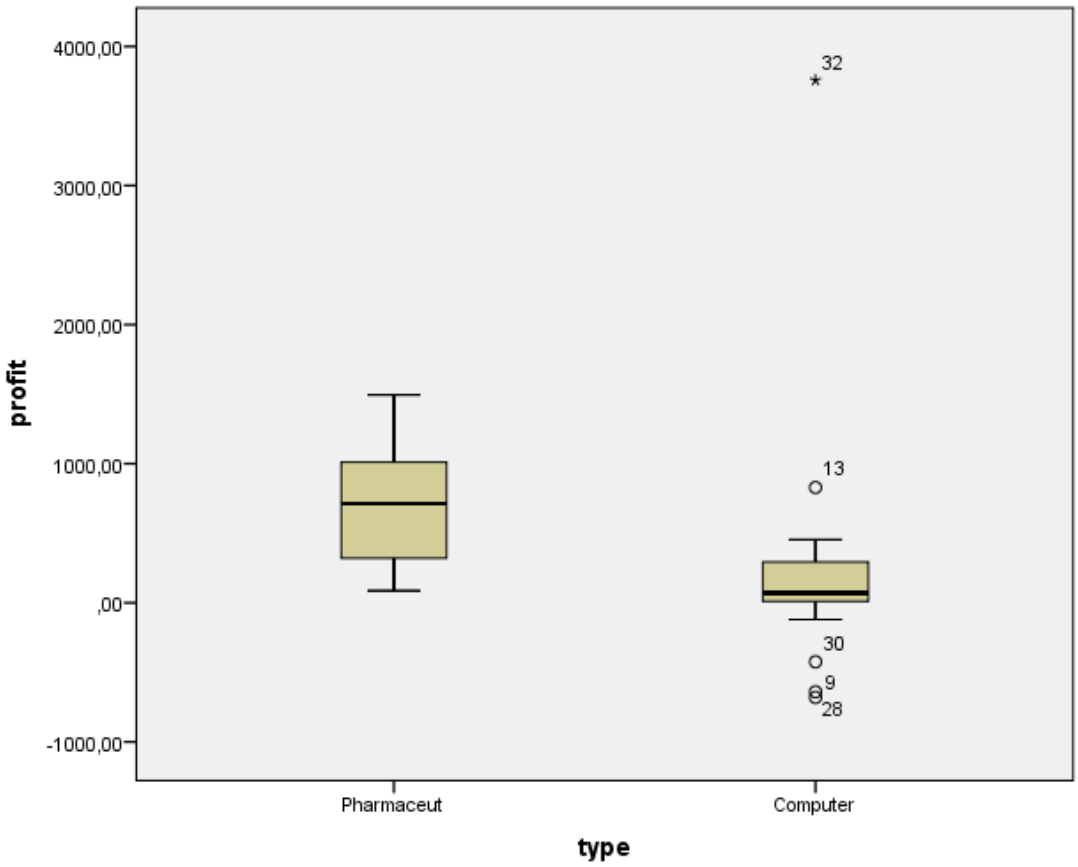
Initially, we check for the presence of extreme values in the sample observations that record the profits of the pharmaceutical and computer companies, respectively. The check for the presence of extreme values in the sample profit values of the 20 computer companies showed that there are at least 3 extreme values (percentage of extreme values > 10%), with observations numbered 32, 9, and 28, and profits of 3758, -680.40, and -639.30, respectively. However, the check for the presence of extreme values in the sample profit values of the 12 pharmaceutical companies showed that there are no extreme values (see boxplots 1, 2, 3, 4).

Then, since the percentage of extreme values exceeds 10%, we examine whether the transformation of the natural logarithm corrects the issue. The logarithm transformation does not correct the problem because the check for the presence of extreme values in the sample values of the natural logarithm of the profit of the 20 computer companies showed that there are at least 3 extreme values (percentage of extreme values > 10%), with observations numbered 9, 28, and 32. However, the check for the presence of extreme values in the sample values of the natural logarithm of the profit of the 12 pharmaceutical companies showed that there are no extreme values (see boxplots 4, 5, 6, 7).

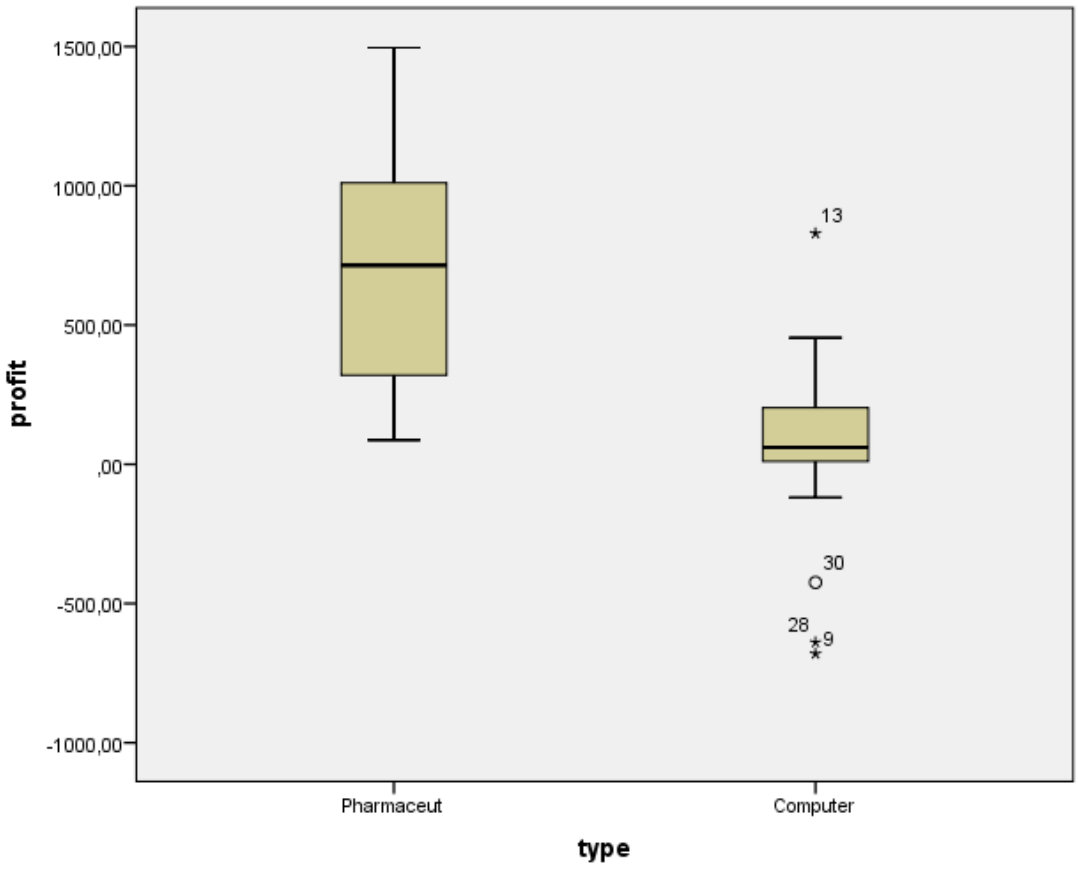
Therefore, since all the assumptions are not satisfied, we will use a non-parametric test to test the hypothesis of equality of population medians of profit between the two types of companies.

In conclusion, for a significance level of 5%, the median of profit for pharmaceutical companies does not differ from that of computer companies (Median test, p-value = 11, sample values of the median profit 714.05 and 70.150). To generalize the results to the population means, the sample mean and the corresponding median should be close. However, this is not the case, as for computer companies, the sample mean and median of profit are 240.8650 and 70.150, respectively, while the corresponding values for pharmaceutical companies are 690.075 and 714.05. Therefore, the results do not generalize to the population means.

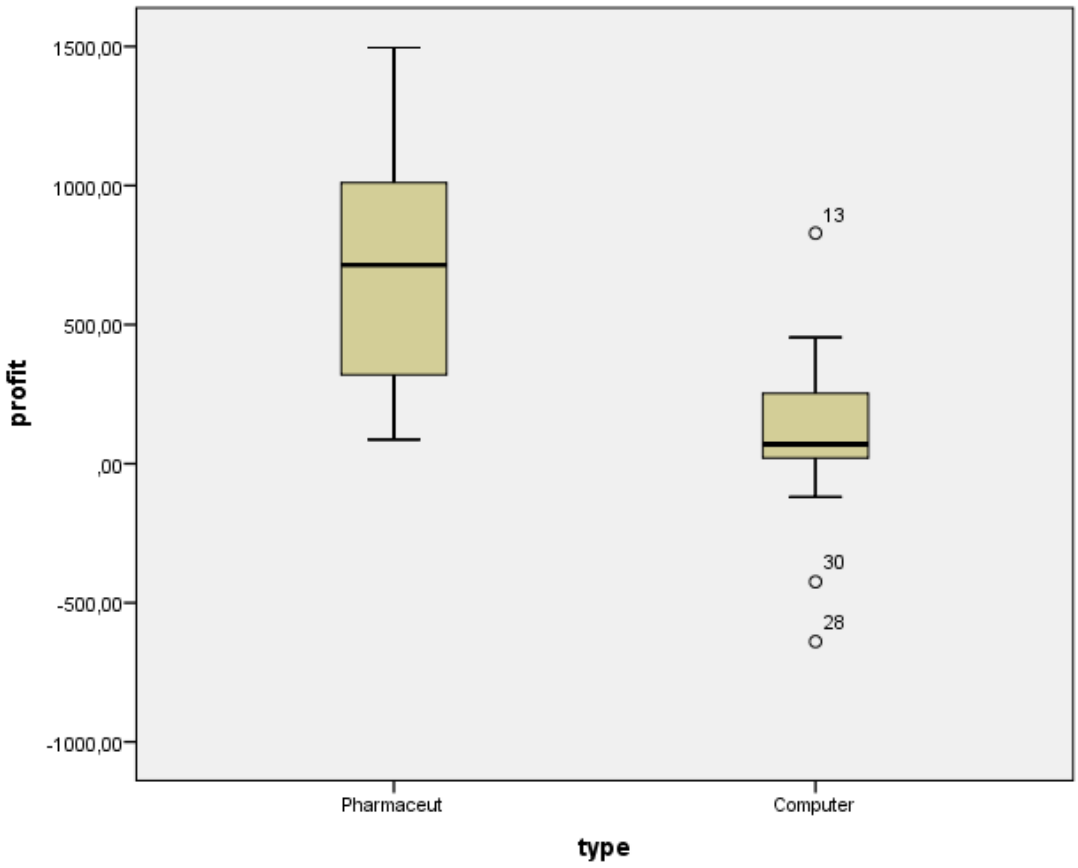
Boxplot 1


4000,00
3000,00
2000,00
1000,00
,00
-1000,00
profit
32
13
30
9
28
Pharmaceut
Computer
type

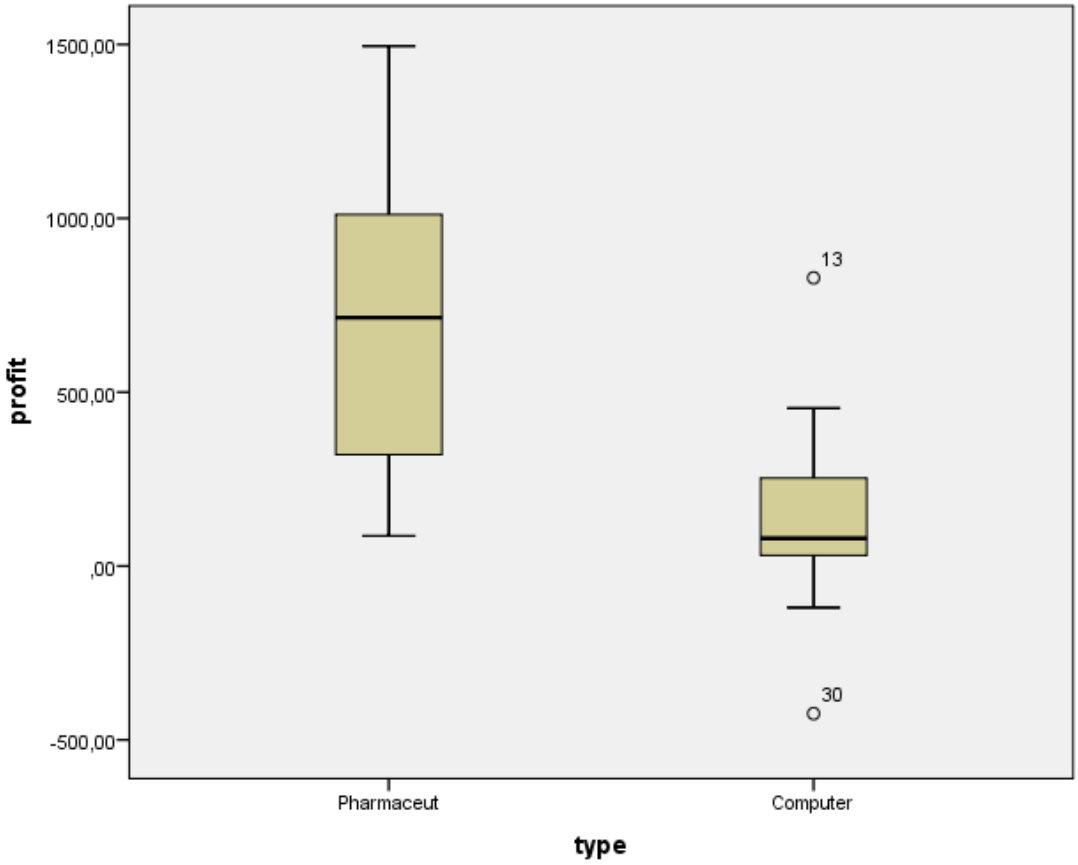
Boxplot 2


1500,00
1000,00
500,00
,00
-500,00
-1000,00
profit
13
30
28
9
Pharmaceut
Computer
type

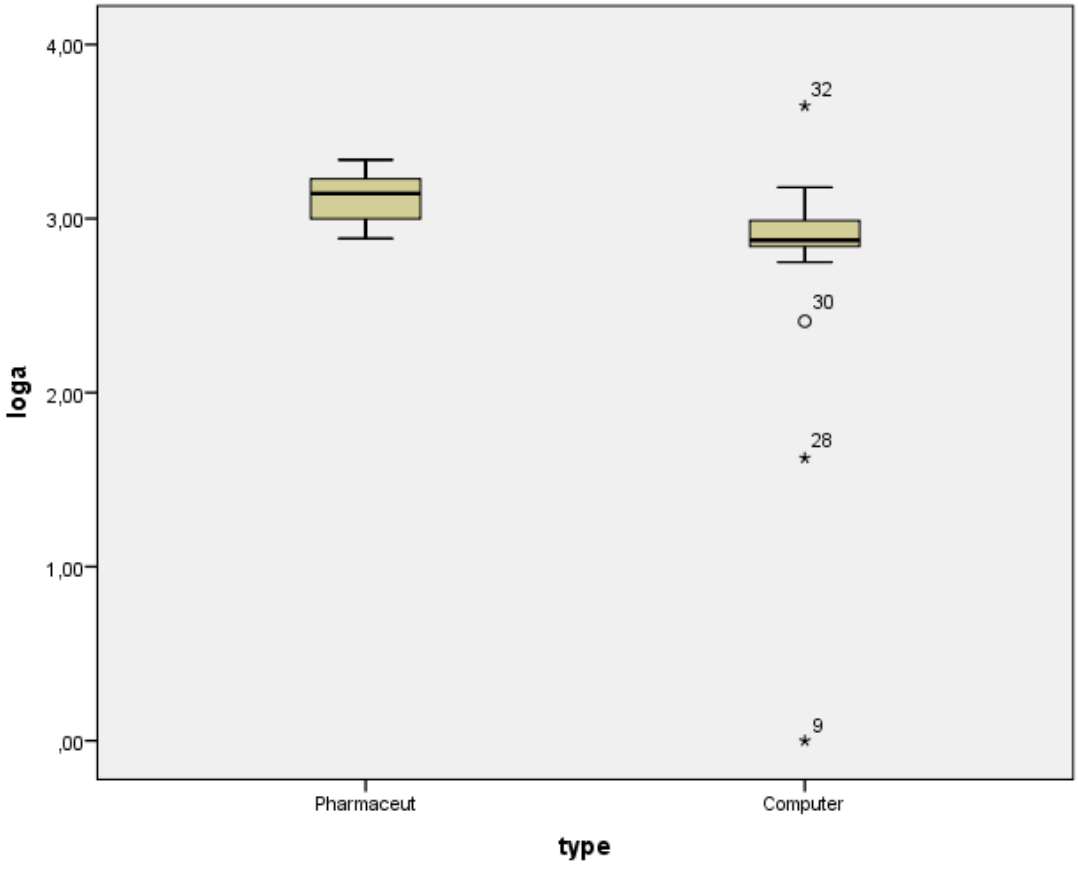
Boxplot 3



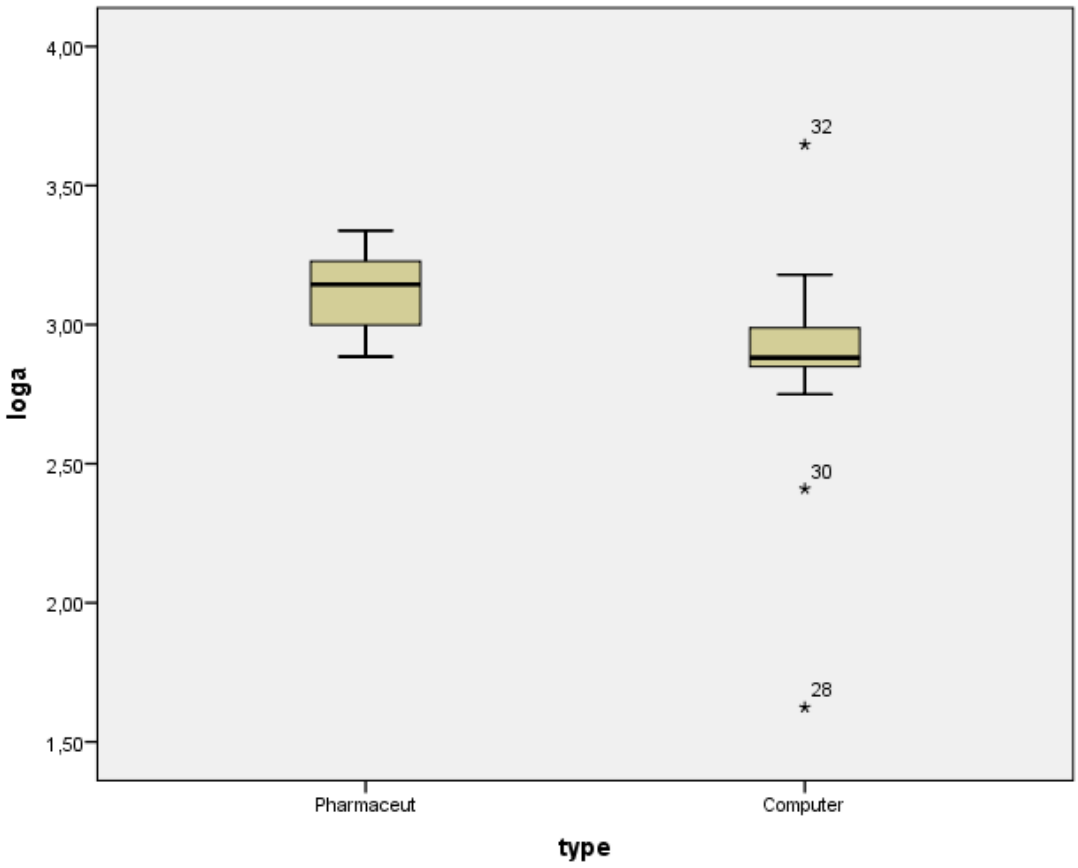
Boxplot 4


1500,00
1000,00
500,00
,00
-500,00
profit
13
30
Pharmaceut
Computer
type

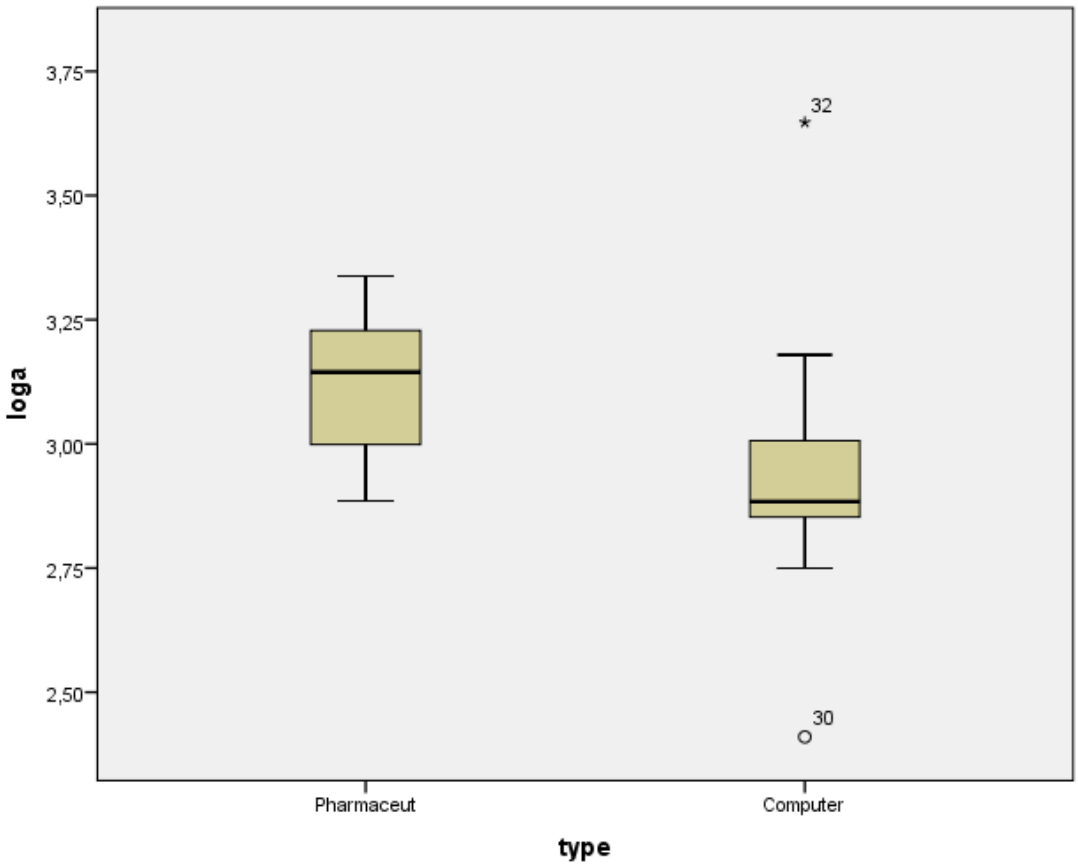
Boxplot 5


4,00
3,00
2,00
1,00
,00
loga
32
30
28
9
Pharmaceut
Computer
type

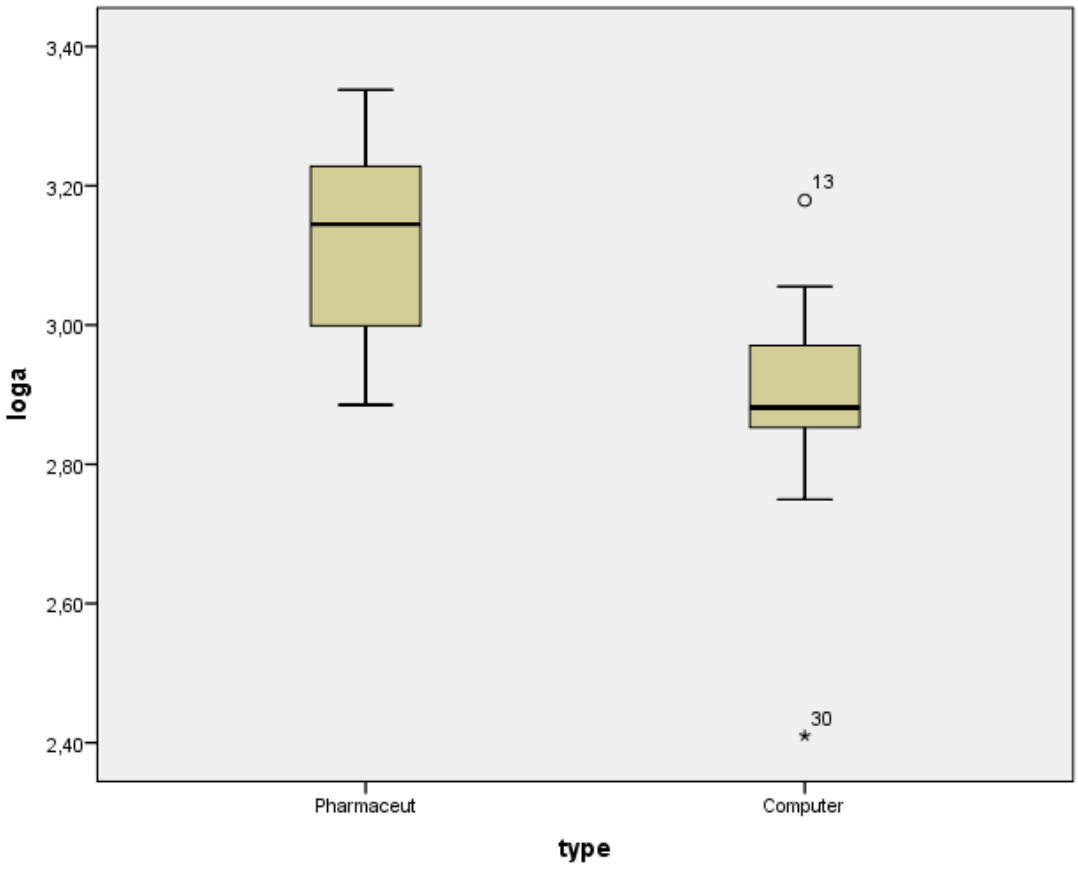
Boxplot 6



Boxplot 7


3,75
3,50
3,25
3,00
2,75
2,50
loga
32
30
Pharmaceut
Computer
type

Boxplot 8


3,40
3,20
3,00
2,80
2,60
2,40
loga
13
30
Pharmaceut
Computer
type

b) We want to test whether there is a statistically significant difference in sales among three types of companies based on their size. To use the parametric test, we need to meet the following assumptions:

1. Our samples should be randomly selected.
2. There should be no extreme values in the sample data of each population exceeding 10%.
3. Each population should be adequately described by the normal distribution.

The first assumption relates to how we selected our samples, and it is satisfied.

First, we check for the presence of extreme values in the sample observations that describe the sales of the Small category of companies. The check for the presence of extreme values in the sample sales values of the Small category showed that observation 5 with a value of 5284 is an extreme value (number of extreme observations <10%).

Next, the check for the presence of extreme values in the sample sales values of the Medium category showed that observation 17 with a value of 6698.40 is an extreme value (number of extreme observations >10%).

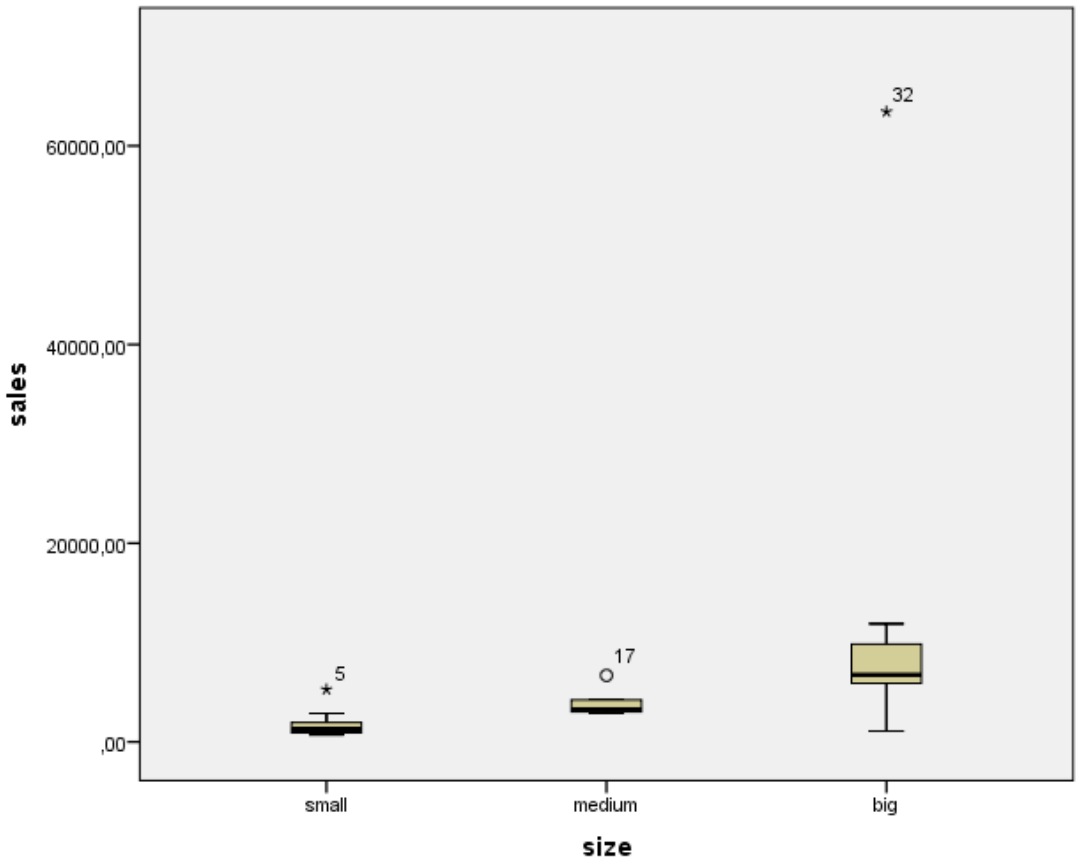
The check for the presence of extreme values in the sample sales values of the Large category showed that observation 32 with a value of 63438 is an extreme value (number of extreme observations >10%) (see boxplots 1, 2).

Since the percentage of extreme values exceeds 10%, we examine whether the transformation of the natural logarithm corrects the issue. The logarithm transformation does not correct the problem because the check for the presence of extreme values in the sample values of the natural logarithm of sales of the 9 Large companies showed that there is at least one extreme value (percentage of extreme values >10%), and the observation number is 32 (see boxplots 3, 4).

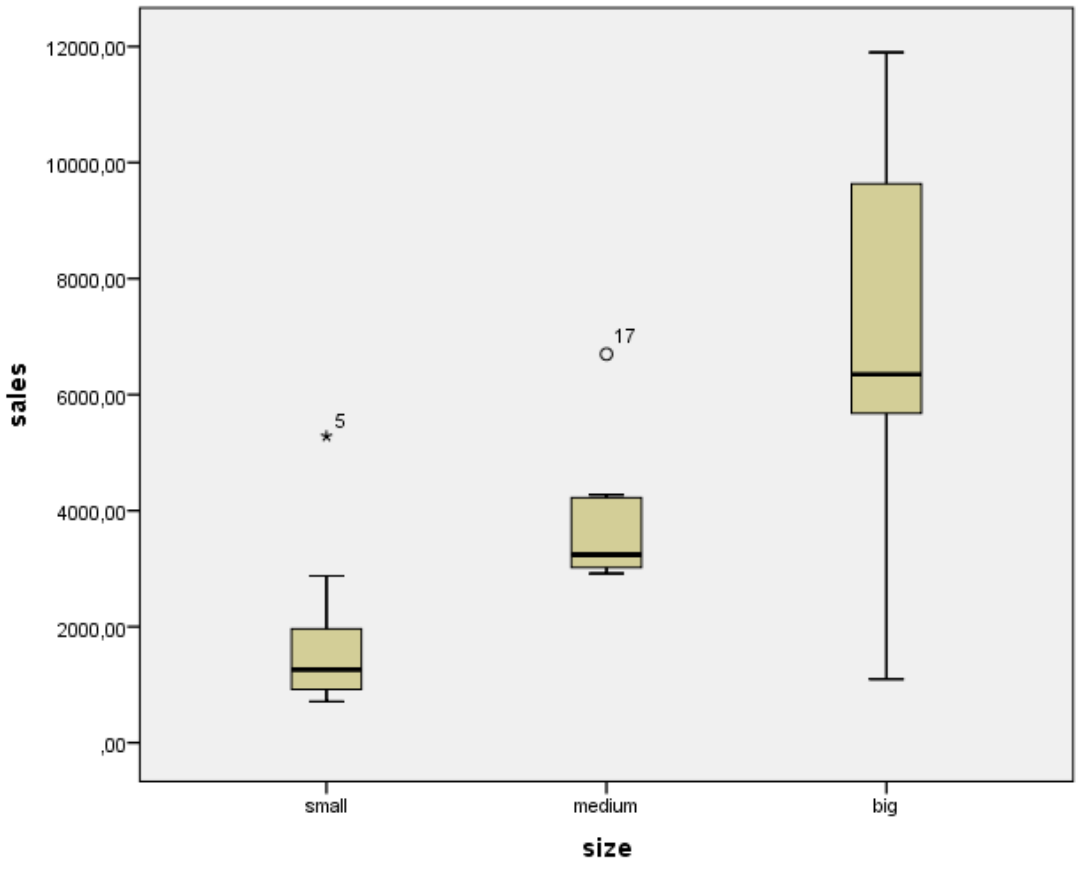
Therefore, since all the assumptions are not satisfied, we will use a non-parametric test to test the hypothesis of equality of population medians of sales among the three types of companies based on their size.

In conclusion, for a significance level of 5%, there is no statistically significant difference in the population medians of sales among the three types of companies (Independent Median test, p-value < 0.001). To generalize the results to the population means, the sample mean and the corresponding median should be close. However, this is not the case, as for Small category companies, the sample mean and median of sales are 1645.5125 and 1261.3000, respectively, while for Medium category companies, they are 3906.1429 and 3243, and for Large category companies, they are 13306.6778 and 6747. Therefore, the results do not generalize to the population means.

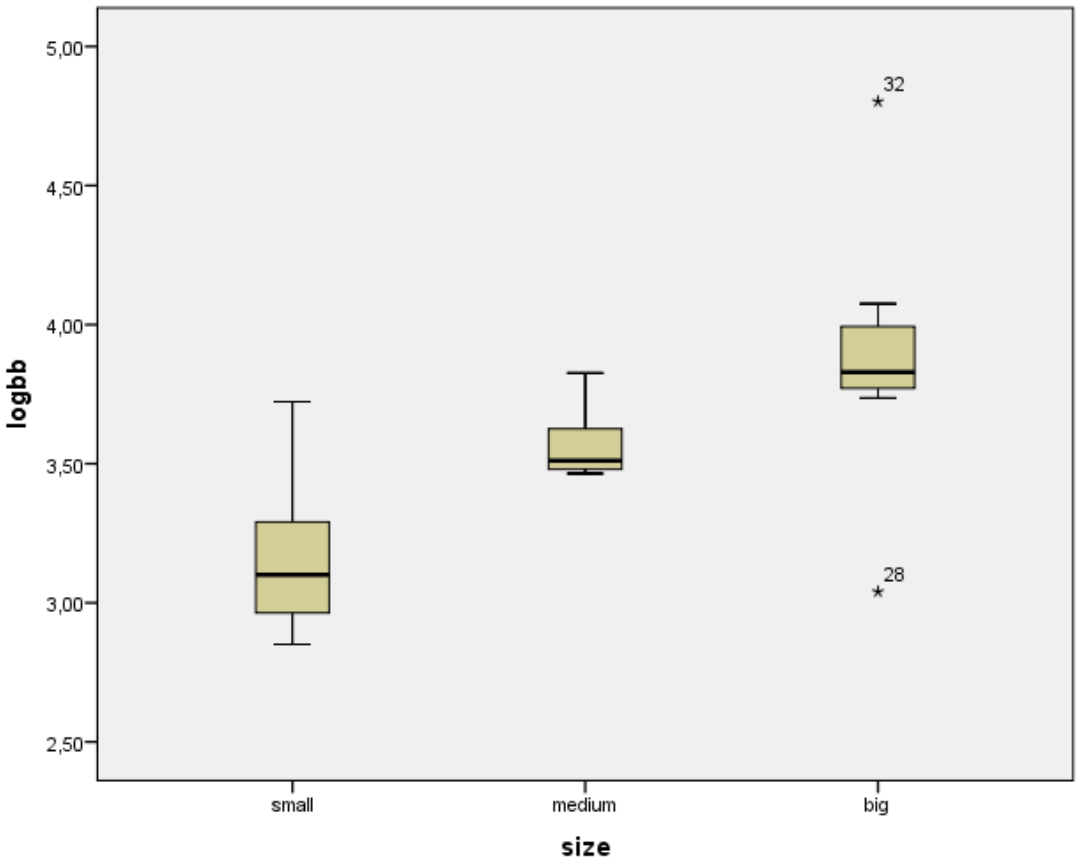
Boxplot 1


60000,00
40000,00
20000,00
,00
sales
32
17
5
small
medium
big
size

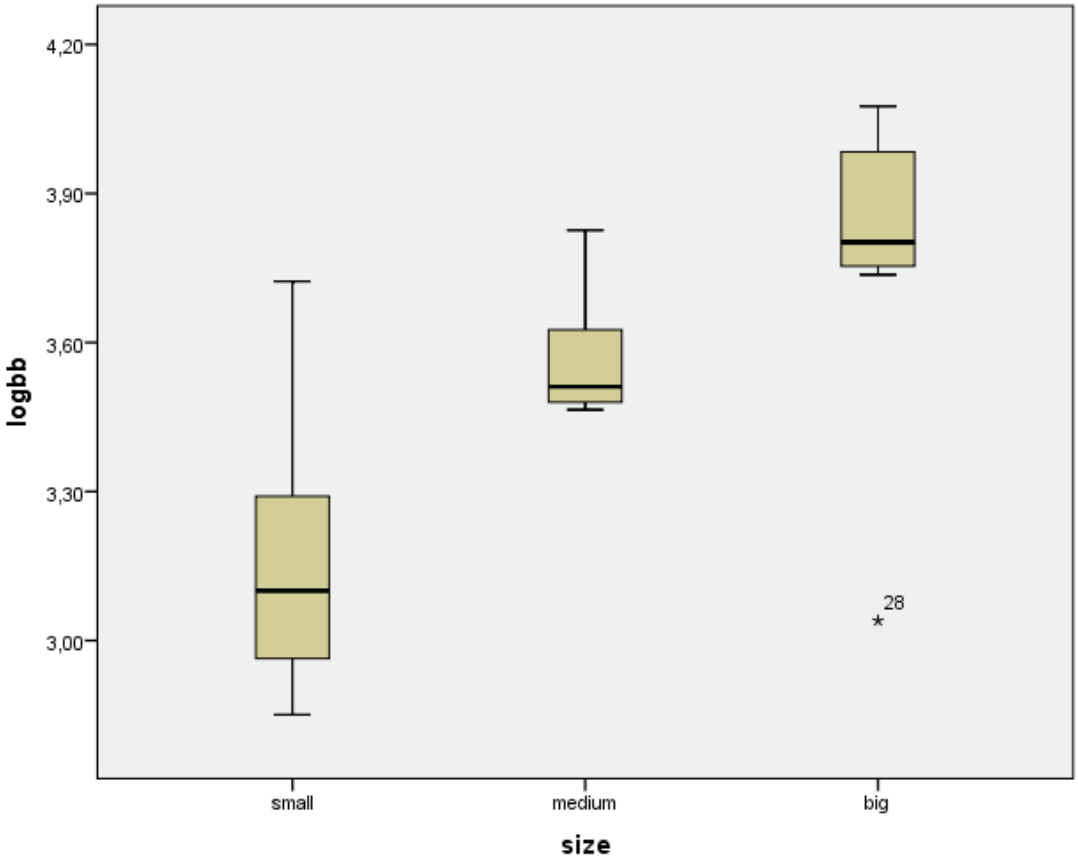
Boxplot 2


12000,00
10000,00
8000,00
6000,00
4000,00
2000,00
,00
sales
17
5
small
medium
big
size

Boxplot 3



Boxplot 4


4,20
3,90
3,60
3,30
3,00
logbb
28
small
medium
big
size