**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Statistical Techniques for Data Analysis |
| **Assessment Title:** | CA2 |
| **Lecturer Name:** | Kayoum Khbuli |
| **Student Full Name:** | Daniela Mariano Barreto |
| **Student Number:** | 2023278 |
| **Assessment Due Date:** | 21/12/2023 |
| **Date of Submission:** | 18/12/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Table of Contents

[**Task 1** 3](#_Toc153825298)

[**Task 2** 6](#_Toc153825299)

[**Task 3** 10](#_Toc153825300)

[**Task 4** 13](#_Toc153825301)

[**References** 16](#_Toc153825302)

# **Task 1**

*Load the dataset Q1.csv. It contains the exam scores (in percentages) of a sample of 50 students from a Dublin secondary school.*

*a. Find and comment on important summary statistics and produce an appropriate plot to summarise the dataset.*

Below contains the results of the summary statistics (Table 1). The results obtained show that, on average, the students got 68.74% of the score in the exam, which is lower than the national average (70%). Still, more investigation is necessary to confirm whether it is true.

The variance result indicated that the values in the dataset are more spread out than the mean. Comparing the mean with the median, we can see that the data are not too skewed, and we can confirm it by the function .skew(), which resulted in a value around zero, meaning that the distribution is approximately symmetric.

The kurtosis value can complement these results to understand the distribution shape better. In this case, the distribution has a lighter tail than a normal distribution, meaning fewer outliers in this curve region than the normal one.

**Table 1:** Results of the summary statistic.

A table of numbers with black text

Description automatically generated with medium confidence

The histogram below (Figure 1) represents the results obtained in the summary statistic, in which we can see an approximately normal distribution of the data scored by students.

The figure shows the data distribution with kernel density estimate (KDE) curve, which provides a more detailed view of the distribution shape, helping to identify peaks and trends. For example, in this case, it seems that many students achieved scores between 60% and 75%, and around 50% seem to have overcome the national average.

A graph of a distribution of exam scores

Description automatically generated

**Figure 1:** Distribution of exam scores with kernel density estimate (KDE) curve.

*b. One of the teachers is concerned about the performance of the students in the school. She suspects that their performance may be below the reported national average of 70%. Does the data show that her concerns are justified? Use a significance level of alpha = 0.05.*

I used the one-sample t-test in this question because it asked to compare a sample mean and a population mean. Although the sample size is 50, greater than 30 when I was supposed to use the z-test, this would not be possible because I do not have the population standard deviation, thus I chose the t-test.

1. State the hypothesis:

H0: µ = 70% 🡪 The mean score of the students is equal to the national average.

HA: µ < 70% 🡪 The mean score of the students is below the national average.

2. Find the critical value:

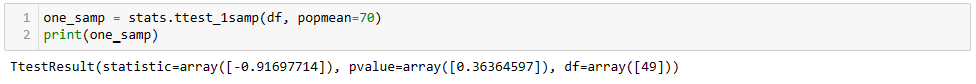
This question is related to a smaller (<) comparison, which means that this distribution has a left-tail and a significance level of 𝛼= 0.05, with a degree of freedom equal to 49. When I verified the t-table (link below), I found that the critical value is -1.6766 under this circumstance.

Degrees of freedom (df) 🡪 df = n – 1 = 50 – 1 = 49

t-table: <https://www.tutorialspoint.com/statistics/t_distribution_table.htm>

3. Calculate a test statistic based on the data:

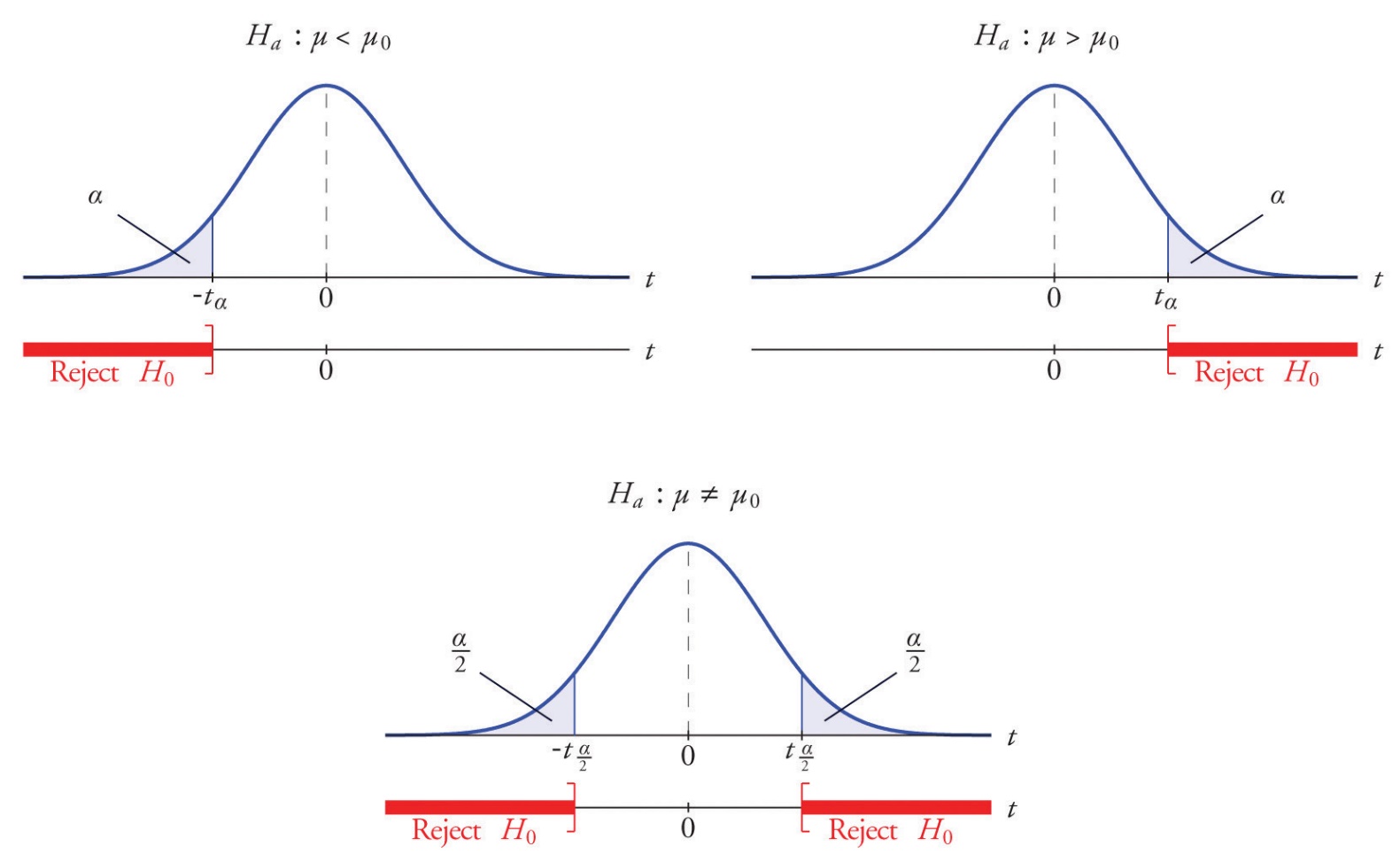
I used the scipy.stats library to calculate a one-sample t-test with the 'stats.ttest\_1samp' function. This function directly calculates the t-value and p-value, facilitating the comparison with the 𝛼 to find if I can reject or not H0. Below is the print screen of the code used (Figure 2).



**Figure 2:** Code used to calculate the t-test.

4. Make a decision:

As we can see, the p-value resulted in 0.3636, greater than the 𝛼; thus, we failed to reject H0, which means we do not have enough evidence to suggest a significant difference between the means (sample\_mean = 68.735 and population\_mean = 70). Figure 3 illustrates how I could decide if I would reject or not the H0.



**Figure 3:** Illustration of the critical value for 𝛼=0.05 of confidence level and t-value suggesting for do not reject H0.

5. Summarize the results:

The teacher was concerned that their student's performance may be below the national average (70%). However, with 95% confidence, there is not enough evidence to suggest a significant difference in the average exam score of the students compared to the reported national average. Therefore, her concern is not justified.

*c. Produce and comment on an appropriate plot to illustrate your findings.*

Figure 4 shows that the mean score of the students is below the national average; however, based on the statistics, a one-sample t-test with 95% confidence is not possible to make this inference because there is not enough evidence to support a significant difference between the means.

A graph of a bar graph

Description automatically generated

**Figure 4:** Chart with the frequency of exam scores of the 50 students. Red line: national average. Black dashed line: mean score of the 50 students.

# **Task 2**

*Load the diamonds dataset and print the first 5 rows. The color variable refers to the colour of the diamond, with categories from “D” to “J”. Colourless diamonds are considered better than diamonds with a yellow tint. Diamonds from “D” to “F” are considered colourless, and diamonds from colour “G” to “J” are not considered colourless (that is, they have a very faint colour).*

*a. Create a new binary variable in the dataframe called “colourless” which records 1 in rows with colourless diamonds and 0 otherwise.*

The new variable will be based on the color variable from the original dataset and called "colourless", with 1 for categories from 'D' to 'F' and 0 for categories from 'G' to 'J'. To do it I the method '.astype()' ensures that the 'color' column will be treated as a string before using the '.str.upper()' method that converts all characters to uppercase. Then I used >= 'D' & <=' F', resulting in a boolean series which returns 'True' = 1 for those elements between 'D' to 'F'; otherwise, will return 'False' = 0. The print screen of the code and the dataset are below (Figure 5).

A screen shot of a computer

Description automatically generated

**Figure 5:** Code and dataset whit the "colourless" variable.

*b. Perform an appropriate hypothesis test to determine whether there is any association between the clarity of a diamond and whether it is colourless or not. Use a significance level of alpha = 0.01.*

I will use the two-sample chi-square test because it is suitable for analysing two categorical variables to know if they are similar (Devore, 2012, p.617). In this question, I need to know if the variable 'clarity' and 'colourless' are associated.

1. State the hypothesis:

H0: There is no association between 'clarity' and 'colourless'. They are independent.

HA: There is an association between 'clarity' and 'colourless'. They are dependent.

The data is collected into a contingency table that counts diamonds based on the clarity and colourness (Figure 6).

A screen shot of a computer

Description automatically generated

**Figure 6:** The contingency table for 'clarity' and 'colourless'.

2. Find the critical value:

I did not find the critical value because the p-value was calculated below, thus I compared the 𝛼 with the p-value. In this case the p-value is less than 𝛼, therefore we can reject H0.

3. Calculate a test statistic based on the data:

I used the contingency Table with the scipy.stats library and the 'chi2\_contingency' function to calculate chi-square, p-value, and degrees of freedom. The resulted are below (Figure 7):

The Chi-Square compares the frequencies of observed and expected to assess whether there is a significant association or not. In this case, it resulted in more substantial evidence to reject H0 (small p-value), meaning that there is a significant association between 'clarity' and 'colourless’.

A screenshot of a computer program

Description automatically generated

**Figure 7:** Code with results for chi-square, p-value, degrees of freedom and the comparison between p-value and alpha.

4. Make a decision:

Considering the p-value smaller than 𝛼 = 0.01 and a high Chi-Square, I opted for reject H0.

5. Summarize the results:

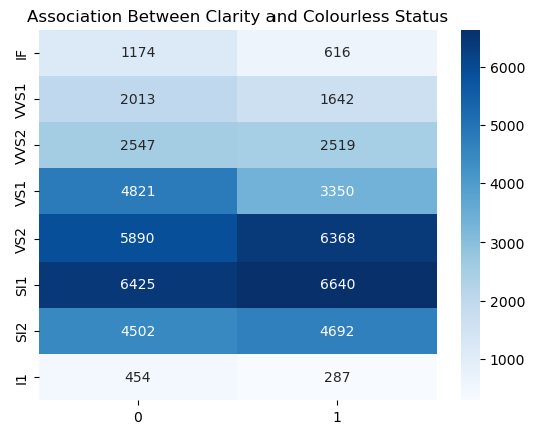
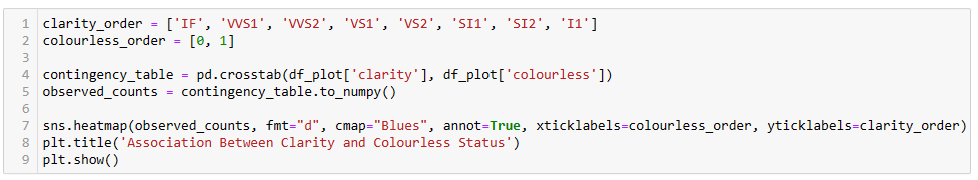
I rejected H0 based on the small p-value compared with 𝛼 = 0.01 and a high Chi-Square. That means I can accept HA, meaning there is a significant association between 'clarity' and 'colourless'. Therefore, 'clarity' and 'colourless' depend on each other.

*c. Produce and comment on an appropriate plot to illustrate your findings.*

Heatmaps are used to illustrate the relationship between two categorical variables, helping to understand how they are related in the dataset. I will use this plot to show the association between clarity and colourless that I found above. In Figure 8 is presented the code used and the heatmap.

In this plot, we can see in the x-axis the 'colourless' categories (0 and 1) and in the y-axis the 'clarity' categories ('IF', 'VVS1', 'VVS2', 'VS1', 'VS2', 'SI1', 'SI2', 'I1'). In addition, the intensity of the blue colour in each cell represents the count of observations for a specific combination of 'clarity' and 'colourless'.

For example, darker colours indicate higher counts that we can see in 'SI2', 'SI1', 'VS2', and 'VS1', meaning high dependencies of a particular 'colourless' category. In other words, in 'SI1', there is a strong dependence on the 'colourless' variable.



**Figure 8:** Association between ‘Clarity’ and ‘Colourless’ of diamonds and code used.

*d. Find and interpret 90% confidence intervals for both the mean price of colourless diamonds and the mean price of non-colourless diamonds.*

I will calculate the confidence intervals for the mean price of colourless and non-colourless using the formula below.

*A black and white text

Description automatically generated with medium confidence*

Where:

x̄: Sample Mean. The average price of the diamonds in each category.

σ: Standard Deviation. The measure of how spread out the prices are.

n: Sample Size. The number of diamonds in each category.

For 90% confidence interval (CI) the Z-score is equal to 1.645.

CI(colourless) = 3337.7594 ± (1.645 x (3527.0473/√26114)

CI(colourless) = 3337.7594 ± 35.90

CI(colourless)’ = 3301.8587

CI(colourless)’’ = 3373.6600

CI(colourless) = (3301.8587 , 3373.6600)

CI(non-colourless) = 4491.2300 ± (1.645 x (4305.0864/√27826)

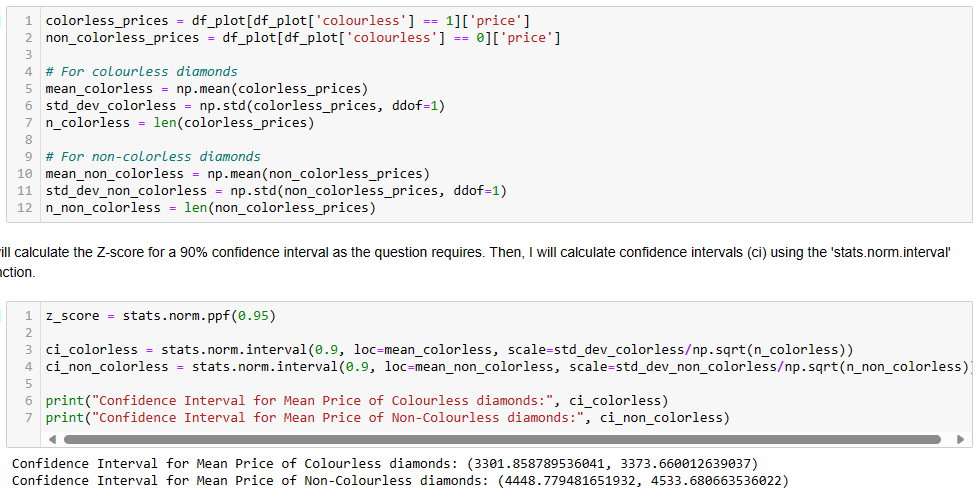
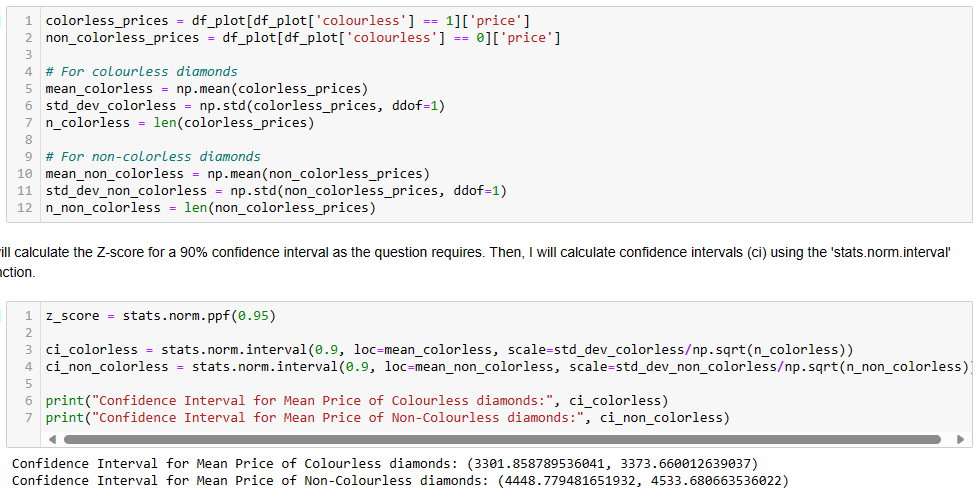
CI(non-colourless) = 4491.2300 ± 42.45

CI(non-colourless)’ = 4448.7794

CI(non-colourless)’’ = 4533.6806

CI(non-colourless) = (4448.7794 , 4533.6806)

I also calculated the confidence intervals in Python and the code are in Figure 9.



**Figure 9:** Calculation of confidence intervals for the mean price of colourless and non-colourless with 90% confidence intervals.

When we compare the manual calculation with the code result, we can verify that they are the same.

Considering that the confidence intervals provide a range of values within which we can estimate the real population mean. These results mean that I am 90% confident that, for the population, the mean price of a colourless diamond is between 3301.8587$ and 3373.66$, while the price of non-colourless is between 4448.7794$ and 4533.66$.

# **Task 3**

*Load the PlantGrowth dataset from the pydataset library. It contains the results of a small study comparing the yields of plants obtained under a control and under two different treatment conditions.*

*a. Find and comment on important summary statistics by treatment and produce an appropriate plot to summarise the dataset.*

The summary statistics results are in Table 2, showing a difference between means and medians, and suggesting a non-normal distribution. The skewness results confirmed that the groups (ctrl, trt1, trt2) do not have symmetric distributions.

To better understand the distributions' shape, we can use the kurtosis values, which, in this case, have lighter tails and fewer outliers than the normal distribution, which is relatively flat, as we can confirm in Figure 10.

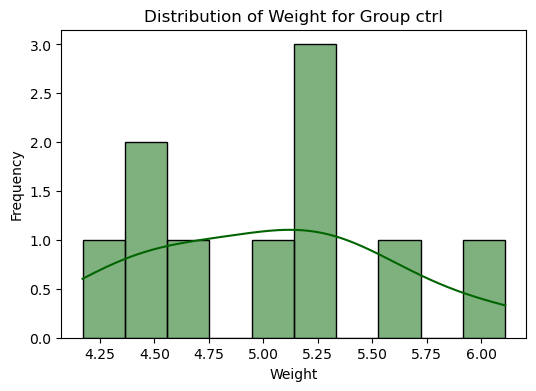
The variance is no different from the mean, suggesting that the values do not vary too much. With a superficial look at Figure 11, I can suppose that treatment 2 resulted in greater growth than treatment 1 and control, which seem equal. However, a statistical test will be necessary to verify it.

**Table 2:** Results of the summary statistic.

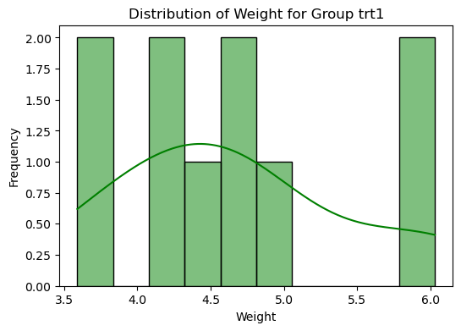
*A table of numbers and lines

Description automatically generated*

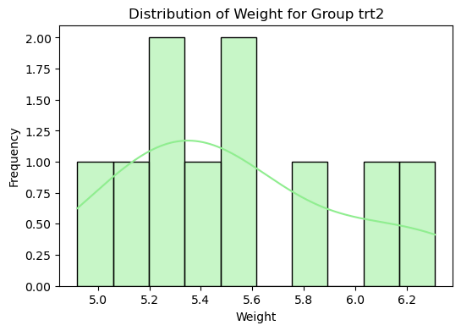
The histogram (Figure 10) represents the results obtained in the summary statistic, in which we can see the asymmetric and flat distribution for all treatments and controls. The kernel density estimate (KDE) curve provides a more detailed view of the distribution shape, helping to identify peaks and trends; for example, the treatments have a similar distribution, but the results are different, but again, it is necessary to perform statistic test to verify that.



a



b



c

**Figure 10:** Distribution of weight for group (ctrl, trt1, trt2) with kernel density estimate (KDE) curve. a) ctrl: control. b) trt1: treatment 1. c) trt2: treatment 2.

A graph of different colored rectangular shapes

Description automatically generated

**Figure 11:** Mean and standard deviation of plant growth for control and treatment 1 and 2.

*b. Conduct an appropriate hypothesis test to see if there is evidence of a difference between the three means (that is, the control and the two treatments). Use a significance level of alpha = 0.05.*

Considering the summary statistics results, the data are not normally distributed. Because of this, I will use a non-parametric test because it generates more reliable results in this case than in comparison with a parametric test. In this case, I will apply the Kruskal-Wallis test because it is appropriate when managing a non-normal distribution and small sample size (Devore, 2012, p.645), conditions that I encountered in this dataset.

As the data is skewed, the use of the median is more robust than the median; because of this, I will compare the medians and the Kruskal-Wallis, already considers the medians in its calculations (docs.scipy.org, n.d.).

1. State the hypothesis:

H0: There is no difference between the medians.

HA: There is difference between the medians.

2. Find the critical value:

I did not find the critical value because the p-value is calculated below. I will directly compare the 𝛼 = 0.05 with the p-value (0.01842). In this case, the p-value is less than 𝛼. Therefore the, we can reject H0.

3. Calculate a test statistic based on the data:

I used the scipy.stats library and Kruskal function to calculate the statistic test and the p-value (Figure 12). It resulted in:

H-Statistic: 7.988228749443715

P-Value: 0.01842375573147196

These resulted in enough evidence to reject H0 (p-value < 𝛼), meaning that there is a significant difference between the medians.

A close-up of a computer screen

Description automatically generated

**Figure 12:** Code with results for Kruskal-Wallis test and p-value.

4. Make a decision:

Considering the p-value smaller than 𝛼 = 0.05, I can reject H0.

5. Summarize the results:

I rejected H0 based on the small p-value compared with 𝛼 = 0.05. That means I can accept HA, meaning there is evidence to suggest differences among the groups (ctrl, trt1, trt2), and the differences are statistically significant at a significance level of 0.05.

*c. If there is evidence of a difference between the three means, find and comment on where this difference may be.*

As we can see above, a significant difference exists among at least one pair of treatments and control. Thus, I performed Dunn's test to find these differences, a non-parametric method for multiple comparisons usually used with the Kruskal-Wallis test (Terpilowski, 2019) (Figure 13).

A computer screen with a white background

Description automatically generated with medium confidence

**Figure 13:** Code with results for Dunn's test and p-value, with the comparison between each group. 1: ctrl, 2: trt1, and 3: trt2.

I found that treatments 1 and 2 are significantly different, considering the p-values less than 0.05, while treatment 1 is equal to the control group. Therefore, with 95% confidence, treatment 2 presented significant growth compared to the control and treatment 1 (Figure 11).

# **Task 4**

*Load the trees dataset from the pydataset library. It contains measurements of the diameter, height, and volume of timber in 31 felled black cherry trees. Note that the diameter (in inches) is labelled girth in the dataset. It is measured at 4 foot 6 inches above the ground.*

*a. Perform a correlation analysis between all numerical variables. Include and comment on the results of hypothesis tests for the population correlation coefficients between all three pairs of variables (you can use the pearsonr function from the scipy.stats library).*

Correlation analysis involves calculating correlation coefficients and quantifying the strength and direction of a linear relationship between two variables. The Pearson sample correlation coefficient is a common measure of the strength of the linear relationship (Devore, 2012, p.201) and will be used in this analysis.

I will use the pearsonr function from the scipy.stats library to obtain the correlation coefficients for each pair (Figure 14):

* Girth and Height
* Girth and Volume
* Height and Volume

I will also calculate the p-value for each pair shown above to perform the hypothesis test considering a confidence interval of 95% and a significance level 𝛼 = 0.05 that I can estimate the real population mean.

State the hypothesis:

* H0: There is no correlation between the variables.
* HA: There is a correlation between the variables.

Compare the p-value with 𝛼:

I will directly compare the 𝛼 = 0.05 with the p-value of each pair:

* If p-value < Reject H0 🡪 Therefore, there is a correlation between variables.
* If p-value > Do not Reject H0 🡪 Therefore, there is no correlation between the variables.

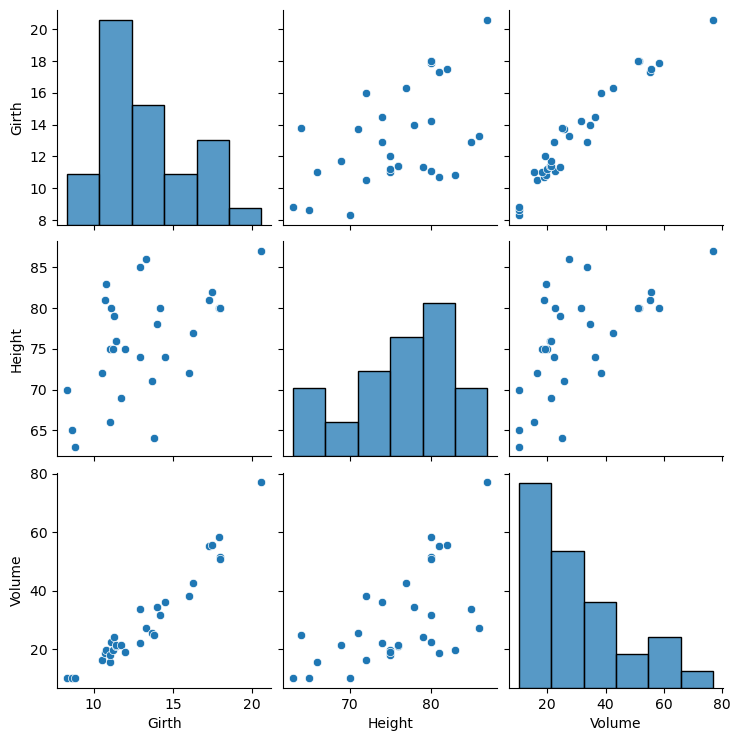
*A screenshot of a computer code

Description automatically generated*

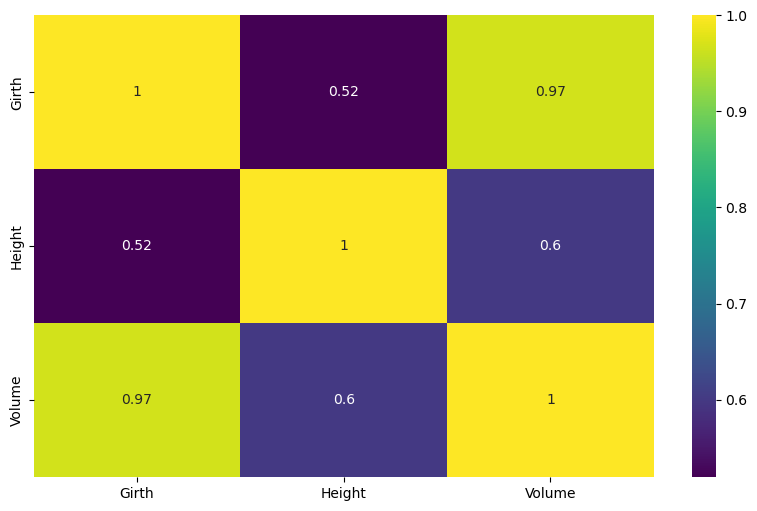
**Figure 14:** Code with Pearson’s correlation coefficient and p-value for each pair.

The results (Figure 14) mean that I am 95% confidence, for the population, that ‘Girth and Height' (r = 0.5193) and 'Height and Volume' (r = 0.5982) have a moderate linear positive correlation, and for both pairs, the p-value is less than 𝛼; therefore, there is a correlation between variables. These results suggest that in both pairs, on average, as Girth increases, Height increases, and for the second pair, when height increases, Volume increases; however, this correlation is not extremely strong, indicating that other factors may also influence tree height and Volume.

The pair 'Girth and Volume' (r = 0.9671) has a strong linear positive correlation with high statistical significance because the p-value is much less than 𝛼. This suggests that as girth increases, volumes also increase and are associated with a strong correlation. Therefore, Girth can be a good predictor of timber Volume in these black cherry trees. Figure 15 illustrate the correlation between the variables and their strength and direction of the relationship.



a



b

**Figure 15:** a) Pair plot and b) heatmap to show the correlation between the variables.

*b. There is interest in estimating the volume of timber from trees using either the girth or the height of the trees, or both. Perform a regression analysis to decide which of the three possible models you would recommend using. Interpret your results and provide a short conclusion of your findings.*

Regression analysis is part of statistics that investigates the relationship between independent and dependent variables (Devore, 2012, p.469). In this case, I will estimate the volume of timber (dependent variable) based on girth, height, or both.

Linear equation:

Y = m.X + c

Where:

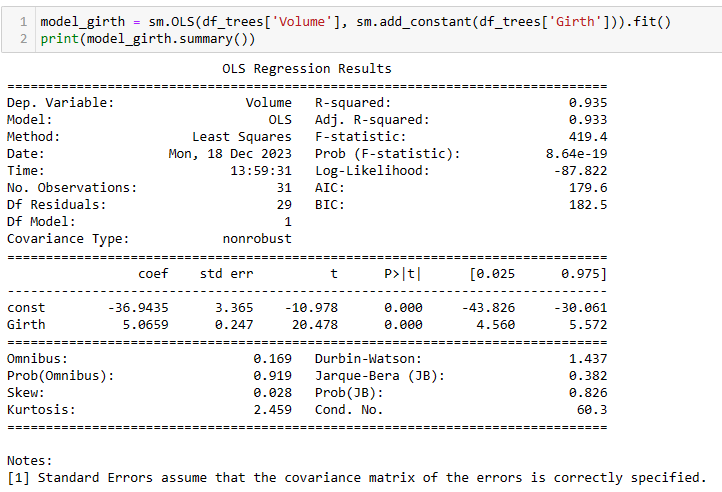
Y: dependent variable (target).

m: slope (coefficient), is the rate of change.

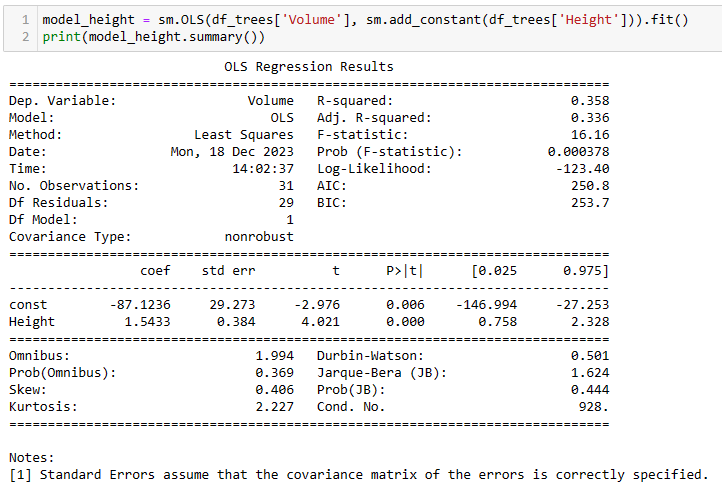
X: independent variable.

c: intercept, the point where the line intersects the y-axis.

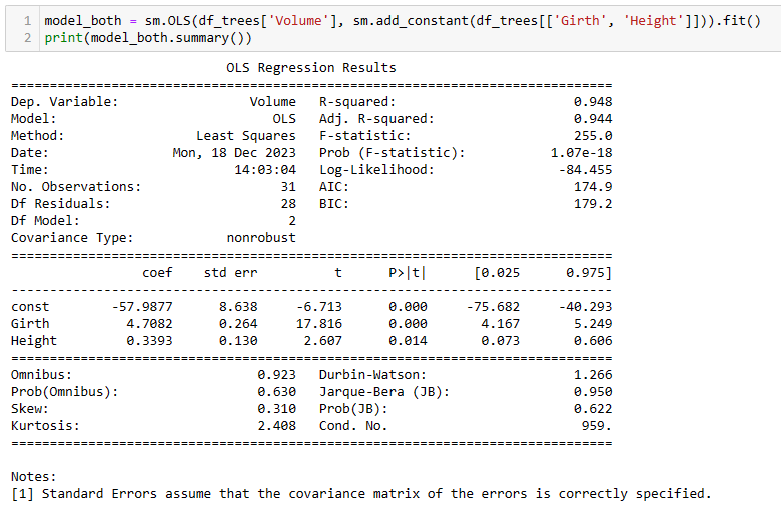
I used the 'statsmodels' library to access the 'sm.OLS' method to estimate the parameters of the linear regression, and I used the 'sm.add\_constant' to add the intercept to the independent variable ('Girth'). The .summary method printed all the summary statistics, mainly the R-squared value, p-value, and coefficients (Figure 16, a – c).



a



b



c

**Figure 16:** Code and summary statistics for each model. a) Model 1: Volume ~ Girth. b) Model 2: Volume ~ Height. c) Model 3: Volume ~ Girth + Height.

The higher the R-square value, the more successful the linear regression model in explaining y variation (Devore, 2012, p.485). Hence, I will analyse these values for each model to identify which one performed well. Secondly, I will check the p-value for the coefficients to verify if the results are statistically significant.

**Model Recommendation**

**Summary results:**

**- Model 1: Volume ~ Girth:** This model resulted in a high R-square, indicating that the Girth alone explains 93.5% of the variation in timber volume. Considering a confidence interval of 95% and a significance level 𝛼 = 0.05, the p-value less than 𝛼 suggests that Girth is a statistically significant predictor of timber volume.

**- Model 2: Volume ~ Height:** This model resulted in the worst R-square, indicating that only 35.8% of the variation in timber volume is explained by the Height. However, with a confidence interval of 95%, the p-value was less than 𝛼, suggesting that Height is a statistically significant predictor.

**- Model 3: Volume ~ Girth + Height:** This model presented the highest R-square, indicating that the combination of Girth and Height explains 94.8% of the variation in timber volume. In addition, considering a confidence interval of 95% and a significance level 𝛼 = 0.05, the p-value less than 𝛼 suggests that Girth and Height are statistically significant predictors of timber volume.

**Conclusion:**

Based on these results, models 1 and 3 performed very well in predicting the timber volume, with only a 1.3% difference between them; thus, both can be used confidently. However, considering that I have to recommend only one, I would choose the third model because the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) resulted in lower values than the first model. These metrics penalize models with more variables; lower values indicate a better-fitting model. Therefore, despite a slight difference, I would recommend model 3.

# **References**

Astanin, S. (2022). *tabulate: Pretty-print tabular data*. [online] PyPI. Available at: <https://pypi.org/project/tabulate/>. [Accessed 18 Dec. 2023].

Devore, J.L. (2012). *Probability and Statistics for Engineering and the Sciences*. 8th ed. Boston, MA: Brooks/Cole Pub Co.

docs.scipy.org. (n.d.). *SciPy documentation — SciPy v1.8.1 Manual*. [online] Available at: <https://docs.scipy.org/doc/scipy/index.html>. [Accessed 18 Dec. 2023].

docs.scipy.org. (n.d.). *scipy.stats.kruskal — SciPy v1.7.1 Manual*. [online] Available at: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html>. [Accessed 18 Dec. 2023].

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, [online] 9(3), pp.90–95. doi:https://doi.org/10.1109/mcse.2007.55. [Accessed 18 Dec. 2023]

Numpy (2009). *NumPy*. [online] Numpy.org. Available at: <https://numpy.org/>. [Accessed 18 Dec. 2023].

pandas.pydata.org. (n.d.). *User Guide — pandas 1.0.4 documentation*. [online] Available at: <https://pandas.pydata.org/docs/user_guide/index.html#user-guide>. [Accessed 18 Dec. 2023].

PyPI. (2020). *statsmodels*. [online] Available at: <https://pypi.org/project/statsmodels/>. [Accessed 18 Dec. 2023].

Python.org. (2019). *statistics — Mathematical statistics functions — Python 3.7.2 documentation*. [online] Available at: <https://docs.python.org/3/library/statistics.html>. [Accessed 18 Dec. 2023]

Terpilowski, M. (2019). scikit-posthocs: Pairwise multiple comparison tests in Python. *Journal of Open Source Software*, 4(36), p.1169. doi:https://doi.org/10.21105/joss.01169.

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, [online] 6(60), p.3021. doi:https://doi.org/10.21105/joss.03021. [Accessed 18 Dec. 2023].