**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | HDip in Science in Data Analytics for Business/ AI |
| **Assessment Title:** | Statistical Techniques for Data Analysis CA2 |
| **Lecturer Name:** | Kayoum Khbuli (kkhbuli@cct.ie) |
| **Student Full Name:** | Oluwatimileyin Oladipo Ayomide |
| **Student Number:** | 2023383 |
| **Assessment Due Date:** | 21st December 2023 |
| **Date of Submission:** | 21st December 2023 |

**Oladipo Ayomide. O**

**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. |

**Statistical Techniques for Data Analysis CA2**

By

Oluwatimileyin Oladipo Ayomide (Student ID: 2023383)

Higher Diploma in Science in Data Analytics for Business

Statistical Techniques

Kayoum Khbuli

CCT College

Dublin, Ireland

Contents

[INTRODUCTION 4](#_Toc154079589)

[Question 1: 5](#_Toc154079590)

[Question 2: 9](#_Toc154079591)

[Question 3: 14](#_Toc154079592)

[where the difference is? 17](#_Toc154079593)

[Question 4: 18](#_Toc154079594)

[**Girth and Height:** 19](#_Toc154079595)

[**Girth and Volume:** 20](#_Toc154079596)

[**Height and Volume:** 20](#_Toc154079597)

[Regression Analysis: 20](#_Toc154079598)

[Evaluating model performance: 21](#_Toc154079599)

[Comparing results: 22](#_Toc154079600)

[Conclusion: 22](#_Toc154079601)

[References 22](#_Toc154079602)

# INTRODUCTION

**What is data?**

Data is plain facts gathered within a defined context. Statisticians would refer to it as a set of observations, which contained within variables or columns whin columns possessing varying/unique characteristics. Data can also be referred to as a piece of information after it has been summarised, and when subjected to analysis, data can be referred to as evidence of a hypothesis. Before data can become information and ultimately evidence it must go through a process. Data analysis is the process in which raw data is ordered and organized, to be used in methods that help to explain the past and predict the future (Hector, 2013).

Data analysis requires good use of statistical techniques and probability in other to understand, summarise and efficiently and accurately manipulate data hence facilitating making reliable predicts to stakeholders of a domain. In addition to statistically summarising data to facilitate decision making, Hypothesis testing also plays a crucial role in the process (Biswal, 2023). Without hypothesis testing, we are at risk of making wrong conclusion about a population based on sample data hence wrong decision been made.

In this project we would be using we would use hypothesis testing to evaluate conclusions made on different datasets.

**What Is Hypothesis Testing?**

Hypothesis testing is a type of statistical analysis within the fields of Statistical inference (CCT, 2023). It involves making some assumptions about a population parameter and putting the assumptions to test to check if it is true or not. It is used to estimate the relationship between 2 statistical variables (Biswal, 2023). “A hypothesis test is the formal statistical test we use to reject or fail to reject a statistical hypothesis” (Zach, 2018).

**Types Of Hypothesis testing**

1. **Null-Hypothesis (*Ho*):** These types of statistical hypothesis states and assumes that there is no difference between a parameter and a specific value, or two parameters are not different (CCT, 2023).
2. **Alternate Hypothesis (*Ha*):** Hais the logical opposite of the null hypothesis. These types of statistical hypothesis states and assumes that there is difference between a parameter and a specific value, or difference exist between two parameters (CCT, 2023).

**Steps in Testing a Hypothesis**

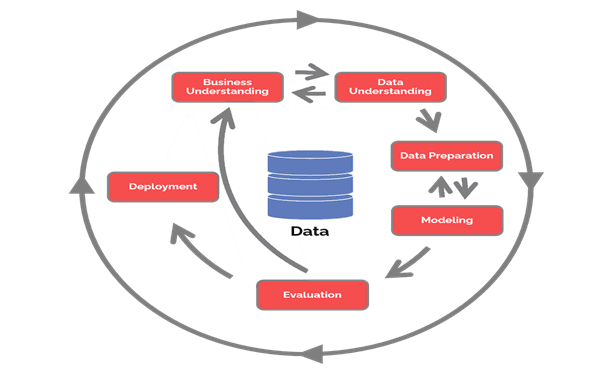
Fig.1 gives a summary of the steps involved in traditional hypothesis testing. The first step is to declare the null and alternate hypothesis, to be able to set limit for acceptance and define rejection zone, we find the critical value in step 2. Step 3 involves calculating the test statistics from the collected sample data. Step 4, we decide if we are to reject or fail to reject the null hypothesis in favour or not in favour of alternate hypothesis. Step 5 involves drawing conclusion on findings.

A screenshot of a test

Description automatically generated

*Figure 1: steps involved in traditional hypothesis testing (CCT,2023).*

In other to have a concise and methodical approach to the tasks, I will be using CRISP-DM project management system.



*Figure 2: Understanding CRISP-DM workflow.*

**Task:**

# Question 1:

**Task understanding:** Exam scores of 50 students from a Dublin secondary has been collected. We are to statistically summarise this data and make comments on important points. The secondary school also wants to know if the concern of one of their teachers is valid that the students has performed below the national average of 70%.

**Data Understanding:** Data was supplied in the task briefing, and it contains the test score of students from Dublin secondary school in percentage, stored in csv format. To facilitate data manipulation, I loaded the dataset into a pandas dataframe called ‘df’. Initial exploration of data shows there are 50 rows and 1 column in dataset, there were no null or duplicated values in the dataset, and all observations are in float dtype.

**Statistical summary**

Based on descriptive statistics table, result shows there were 50 student exam scores recorded in the sample collected. The overall average of the student results was 68.735600%, and there was no recurring exam score as shown by the mode result. The lowest and the highest exam-scores are 48.73000% and 87.360000% respectively. The standard deviation is a number that summarises how dispersed or close the observations are relative to the mean, the standard deviation of student result is 9.750143, this result suggests that the observations are very close to the average exam score as show in fig.4. The first and third interquartile range are 62.040000 and 74.630000 respectively. The median and 2nd interquartile range is 68.390000%, the median and the mean on a gaussian distribution should be the same, but the statistical results of student exam score show the distribution of the scores is slightly right skewed.

**A screenshot of a computer

Description automatically generated**

Figure 3: Summary table

**Visual exploration**

A graph with a blue line with Ryugyong Hotel in the background

Description automatically generatedA bar graph with a blue rectangle

Description automatically generated

Figure 4: Distribution plot for exam-scores

*Figure 5: Boxplot for exam-scores*

**Fig.4**, show a distribution plot of the Dublin secondary school student’s exam-scores. X-axis shows the range of score and Y-axis shows the count of each score. The diagram shows a somewhat gaussian distribution, it is however what noting that there is still a little difference between the mean and the median, suggesting the mean is pulled to the right due to larger exam-scores on the right, hence pulling it to the right, away from the median. The boxplot fig.5 is used to see the pattern of distribution of the observation in a sample. It shows if there are extreme value (outliers) and where exactly they are, on the lower or upper values in the distribution? Outliers are observations that are not in the same range as the rest of the distribution. They are values above the maximum value or below the minimum value. Fig.5 shows we don’t have any outliers in our sample, with the minimum and maximum values 48.73000% and 87.360000% respectively correctly represented in the plot. The middle line in the box is the median and the line above and below it is the third and first interquartile range respectively.

**Question 1b:**

**Defining parameters:**

1. Sample size(n)= 50

2. sample Mean: 68.735600

3. Standard deviation: 9.750143

4. Null-Hypothesis(H0): >= 70%

5. Alternate-Hypothesis (Ha): < 70%

6. alpha = 0.05

**Stating the hypothesis:**

1. Null-Hypothesis: Dublin secondary school Student average exam score is greater or equal to national average of (70%).
2. Alternate-Hypothesis: Dublin secondary school student average exam score is less than 70%.

**Type of test method and confidence-interval:**

This is a Left tail test because the alternate hypothesis assumes that the sample mean is less than the population mean (70%). Given the sample size is 50 and we have a known sample standard deviation, we are going to use the Z-test. Our confidence interval is 95% given a significance level (alpha) of 0.05 and our critical value for left-tail test at given alpha is -1.65.

**Test Statistics:**

A screenshot of a test

Description automatically generated

Figure 6: Z Test statistics formular

**Results:**

Fig.7 show the implementation of the formular in python.

**A screenshot of a computer

Description automatically generated**

Figure 7: Writing formular in Python.

**Making decision**

I used the alpha value to ‘make a decision’ about the exam-score hypothesis. Fig. 8 shows the guiding principle for using alpha to ‘make a decision’ using the court case analogy as given by (CCT, 2023). Result of the test statistics shown the following: {test statistics: -0.91697714, P-value: 0.36364597}. Using P-value to ‘make decision’, **we fail to reject the null-hypothesis.** This is because the P-value (0.36364597) result is not less than given alpha(0.05).

A screenshot of a computer

Description automatically generated

Figure 8: P-Value guide

**Conclusion:**

At 95% confidence interval, and 0.05 alpha, we fail to reject the null hypothesis. This is because the value of p-value (0.36364597) is not less than the alpha value (0.05).

This in conclusion suggests that we have not found enough evidence to substantiate the teaches claim of a less national average score for the student exam result. This however does not validate or confirm that the student’s exam score average above or at least the minimum national average of 70%, that is, this does not mean the Null-hypothesis is correct.

**Plot:**

A graph of a test results

Description automatically generated

Figure 9: Hist plot of the exam-score observations.

Fig.9 is a histogram plot of the Dublin student exam score, ‘bins’ parameter was used to divide the observations intervals into 7. In addition to the histogram plot, 2 ‘axvline’ was drawn to represent the sample mean and population mean. The population mean line is represented in red and the sample mean line is the dashed line. This plot was used to visualise the hypothesis. The results on the p-value lead to the conclusion that we do not have enough evidence to reject the null-hypothesis, it does not mean the hypothesis is right. The visualization showed that the student exam-score average is indeed lower that the population average which is the national average of 70%.

# Question 2:

**Data Understanding:**

Data was imported from the seaborn dataset library, and it contains characteristics of different diamonds. To facilitate data manipulation, I loaded the dataset into a pandas dataframe called ‘diamond’. Initial exploration of data shows there are 53940 rows and 10 columns in dataset, there were no null-values and 146 duplicate values in the dataset, there are 7 numerical variables and 3 object variables in the dataset.

**Statistical summary**

Statistical summary table was created using ‘describe’ method. Result shows there were 53940 diamond samples collected. All numerical variables except ‘price’ appears to be gaussian distribution with slight difference between the mean and median.

**A close-up of a white card

Description automatically generated**

Figure 10: Defining the get\_color variable.

The ‘colour’ variable has 7 unique object observations, [‘D’, ’E’, ’J’] are colourless diamonds, [‘G’, ‘H’, ‘I’, ‘J’] are tainted diamonds. To convert these columns to a binary categorical variable ‘Colourless’, I defined a function called ‘get\_color’ and using if statement, I filtered the colour variable to assigned 1 to list of unique observations in the colourless diamond, and 0 for observations not in the colourless diamond list. I then assigned this operation to a new variable called ‘Colourless’ hence creating a categorical variable of coloured or not- coloured diamonds.

**Hypothesis-testing:**

***Question interpretation:***

The question is asking us to perform a hypothesis testing on 2 variables, that in our diamond dataset happens to be categorical. Categorical variables are variables that categories observations into different categories/group/class. In this question there are two categorical variables in view:

1. Clarity: The clarity column is a categorical variable that groups diamonds into different categories presumably by their properties. A little exploration of the clarity variable using 'unique ()' return that there are 8 unique categories in it namely ['SI2', 'SI1', 'VS1', 'VS2', 'VVS2', 'VVS1', 'I1', 'IF'].

2. Colourless: The colourless column was engineered in question 2.c which now has 2 categories in it. Category ‘1’ are those diamonds that are colourless, and category ‘0’ are those diamonds that not colourless. Furthermore, to ensure colourless was treated as a categorical variable, it was converted from integer to categorical explicitly.

Bringing it all together, question 2 wants us to perform a hypothesis to check if there is any relationship or correlation between the clarity of a diamond, and weather it is going to be colourless or colourful. To do this the normal hypothesis testing methodology will be used.

**Stating the Hypothesis**

1. Null-Hypothesis (H0): Clarity and colourless variables are independent. There is no association between them.

2. Alternate-Hypothesis (Ha): Clarity and colourless variables are not independent. There is association between them.

**Rejection region:**

The P-value compared to the given alpha result would be used to determine if to reject or fail to reject the Null-Hypothesis. The given alpha is 0.01, if the P-value is less than alpha, the n we reject the null- hypothesis is favour of the alternate hypothesis. If the P-value is not less than alpha, then we fail to reject the Null-hypothesis.

**Calculate the test statistics:**

To perform the test statistics for this problem, we would be using the 'Chi-Square-Test of independence'. According to (Zach, 2020), the 'Chi-Square-Test of independence' is an omnibus test used to determine whether ‘there is a’ or ‘there is no’ significant association between two categorical variables.

A white background with black text

Description automatically generated

Figure 11: Chi-test of independence formular.

**Expected value = (row sum \* column sum) / table sum.**

**Steps to performing Chi-Test**

1. Create summary table.
2. Calculate the crosstab df.
3. Calculate Chi-square-Test of independence.
4. Make conclusion.

*Performing univariate analysis*

In other to perform univariate analysis, I installed the researchpy library and imported it as rp, which provides various statistical functions for data analysis in Python. The "rp.summary\_cat(df[["clarity", "colorless"]])" syntax was used to generate a summary table of the clarity and colourless categorical variables. "rp.summary\_cat()"is one of the functions in the researchpy library, and I used it to calculates the frequency and relative frequency of each unique value in clarity and colorless variables. The result from this code syntax as shown in *fig.12* is a summary table that displays the count of unique values in each variable, and the percentage they account for in the entire sample population. It was useful to have this summary table to understand the distribution and characteristics of clarity and colourless variables (code-mentor.ai, 2023).

A table with numbers and letters

Description automatically generated

Figure 12: Summary- table of clarity and colourless variables.

The summary table helps us to understand both variable at a glance and the proportion each category is to another in each variable [Colourless, Clarity].

***Creating contingency table*** *(PYTHON FOR DATA SCIENCE, 2020).*

The contingency table also referred to as crosstab table as shown in *fig.13* helps us to understand the relationship between the ‘clarity’ and ‘colourless’ categorical variables. In chi-square test of independence, the contingency table helps to determine is there is significant relationship between the clarity of a diamond, and it been colourless or not.

A screenshot of a computer

Description automatically generated

Figure 13: Contingency analysis result for clarity and colourless.

Result shows that ‘IF’ clarity has 1174 not colourless diamonds and 616 colourless diamonds, ‘VVS1’ clarity has 2013 not colourless diamonds and 1642 colourless diamonds, ‘VVS2’ clarity has 2547 not colourless diamonds and 2519 colourless diamonds, ‘VS1’ clarity has 4821 not colourless diamonds and 3350 colourless diamonds, ‘VS2’ clarity has 5890 not colourless diamonds and 6368 colourless diamonds, ‘SI1’ clarity has 6425 not colourless diamonds and 6640 colourless diamonds, ‘SI2’ clarity has 4502 not colourless diamonds and 4692 colourless diamonds, ‘I1’ clarity has 454 not colourless diamonds and 287 colourless diamonds.

**Perform Chi-Square Test of independence Statistics:**

Using statistics library, following “stats.chi2\_contingency(C\_t)” syntax, I performed the chi test, passing the contingency table as an argument. Result as shown in *fig.14* as the following components:

1. Chi-Square Test Statistics
2. P-value
3. Degrees of Freedom (dof)
4. Expected Frequencies (expected\_freq).

The **chi-square statistic test** result measures the level of discrepancy between the observed and expected frequencies (Zach, 2020). Our output shows the computed chi-square statistic is ≈486.47. The **p-value** result in this operation, indicates the probability of observing such an extreme (or more extreme) result if the variables were independent i.e null hypothesis is true. Here, the p-value is ≈6.48e101 (0.0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000064815871244549). As seen, it is a very small number close to zero.

The **Degrees of freedom** often denoted as ‘dof’ indicates how many numbers in the grid (contingency table) are independent. It is equal to "the number of rows minus one times the number of columns minus one: that is, (R-1) \* (C-1)"(Ling 300, 2008). In our result, the degrees of freedom is 7. The **expected frequencies** under the assumption of independence between the variables are the frequencies that would be expected in each cell of the contingency table if there were no association between the two variables. They are calculated based on the marginal totals and sample size. “The expected\_frequency array represents the expected frequencies for each combination of the two categorical variables"(code-mentor.ai, 2023).

**Making a decision**

**A white background with black text

Description automatically generated**

Figure 14: Chi-square test of independence result.

Based on result, we reject the Null-Hypothesis, this is because our p-value is close to zero and is smaller than the alpha which is 0.01.

**Conclusion**

In conclusion, it was found in our result that there is strong evidence against 'clarity' and 'colourless' variables being independent. There is some sort of dependency between the two categorical variables based on our result. Which is why we reject the Null-Hypothesis and accept the 'Alternative-hypothesis'. There seems to be a statistically significant association or relationship between ‘Clarity’ and ‘Colourless’ variable in the diamond dataset.

**Plot**

A graph showing different colored bars

Description automatically generated

Figure 15: side by side Count of the relationship between clarity and colourless variables.

The side-by-side count plot has been used to visualise the proportion of colourless and non-colourless diamonds in each clarity category. On the x-axis are all the unique categories in the clarity variable, and on the y-axis are the count of the diamonds. As shown in the plot, category 'I1' has the lowest number of overall diamonds count, the colourless diamond are relatively small compared to the coloured diamonds in this category. The 'SI1' category has the highest overall number of diamonds in it, the colourless diamonds in this category are more compared to the coloured diamonds.

**Disclaimer:** The steps and code I used for this question was inspired by **(PYTHON FOR DATA SCIENCE, 2020)** and I interpreted the results **myself** except where I explicitly referenced.

# Question 3:

**Data Understanding:**

Data was imported from the Pydataset library and contains the results of a small study comparing the yields of plants obtained under a control and under two different treatment conditions. To facilitate data manipulation, I loaded the dataset into a pandas dataframe called ‘plant’. Initial exploration of data shows there are 30 rows and 2 columns in dataset, there were no null and duplicated values in the dataset, there is 1 numerical variable and 1 categorical variable in the plant dataset.

**Statistical summary**

Statistical summary table was created using ‘describe’ method. Result shows there were 30 observations in plant df, and the numerical variables ‘weight’ appears to be gaussian distribution with slight difference between the mean and median. The standard deviation is small relative to the mean which means the observations are closer to the mean. The group column is a categorical column with ['ctrl', 'trt1', 'trt2'] unique group which represent each treatment in the experiment. Each experimental group has equal number of sample size (10).

A screenshot of a calculator

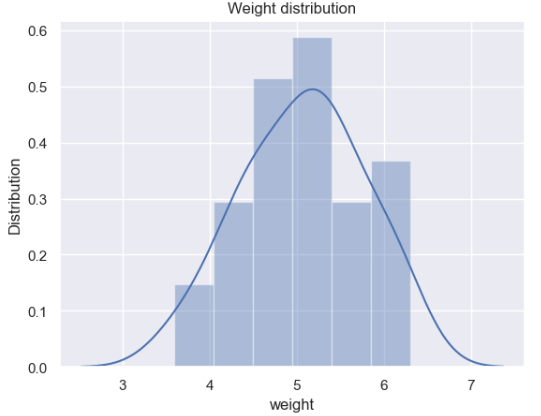
Description automatically generated

Figure 16: Statistical summary and histogram plot of plant weight.

**Question interpretation:**

The question is asking to conduct an appropriate hypothesis to see if there is difference between the means of the three groups ['ctrl', 'trt1', 'trt2'] in the categorical variable ‘group’ in plant df.

**Stating the Hypothesis**

1. Null-Hypothesis (H0): The means of the 3 groups ['ctrl', 'trt1', 'trt2'] are equal.
2. Alternate-Hypothesis (Ha): The means of the three groups are not equal.

**Rejection region:**

The P-value compared to the given alpha result would be used to determine if to reject or fail to reject the Null-Hypothesis. The given alpha is 0.05, if the P-value is less than alpha, then we reject the null- hypothesis is favour of the alternate hypothesis. If the P-value is not less than alpha, then we fail to reject the Null-hypothesis.

**Calculate the test statistics:**

To do this, I will use ANOVA testing. ANOVA testing stands for 'Analysis of Variance', it is an omnibus test, it tests the difference overall between all groups (PYTHON FOR DATA SCIENCE, 2020b).

**Types of ANOVA testing**

**One way ANOVA testing** also referred to as one factor ANOVA, is a parametric test compare that means of three or more independent groups in an independent variable. Since it is an omnibus test, it tests for a difference overall, i.e. at least one of the groups is statistically significantly different than the others. This is called one factor ANOVA, though there are 3 or more groups being tested is because the groups are in the same categorical variable, hence the name is because the analysis is in one variable. If there are two variables being compared it would in which case be called a **two-way, or two factor ANOVA test** if both variables are categorical (PYTHON FOR DATA SCIENCE, 2020b). (PYTHON FOR DATA SCIENCE, 2020b) reported before implementing ANOVA testing, it is of good practice to check Parametric test assumptions, which involved:

1. Population distributions are normal.
2. Samples have equal variances.
3. Independence.

**Population distribution**

To check the population distribution, we can used Q-Q plot (quantile-quantile) plot. (Varshney, 2023) reported that Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian distribution, uniform distribution, exponential distribution or even a Pareto distribution. Q-Q plot can be used to visualise the type of distribution in a variable.

A graph with a red line and blue dots

Description automatically generated A graph with a red line and blue dots

Description automatically generated

A graph with a red line and blue dots

Description automatically generated

Figure 17: Plot A B C are the Quantile-Quantile plot of the 3 treatment groups in the 'group' cat-variable and their weight.

Quantile is cut points dividing the range of a probability distribution into continuous intervals with equal number of observations. In our plot, the 3 different plots indicate the unique experiments in the group column in plant df. The x axis represents theoretical quantile, and the y axis represent the quantile from the categories in group variable of plant df. The data points are the intercept where the theoretical and data quantile intercept, as shown for example in ***fig.17***. Our graphical output shows that the groups do not follow a normal distribution because the intercepts do no line perfectly on the line.

**Performing ONE-WAY ANOVA (Test statistics)**

Before performing the test, the function 'rp. summary\_cont()' was used to was used to summarise the categorical variable 'group' column in the plant df. This table provided information about the central tendency, dispersion, and distribution of the 'weight' variable in the dataset (Codementor.ie, 2023). This gives us a visual summary of the categories in group individually in a glance.

A white background with black text

Description automatically generated

Figure 18: Formular for F-statistics (PYTHON FOR DATA SCIENCE, 2020b).

**Make a Decision:**

We want to check if we have enough evidence to support the null-hypothesis that states “the means of the 3 groups are equal” and reject the alternate hypothesis that they are not. This is why the ANOVA testing was conducted. After testing result show the ‘test statistic result’ is ≈ 4.8, and the P-Value is 0.0159099583256229 ≈ 0.02. Comparing the result of P-Value to the alpha, which is 0.05, we reject the Null-hypothesis.

**Conclusion:**

**A screenshot of a graph

Description automatically generated**

Figure 19: Statistical summary of the categories in 'group' from plant df.

Obviously, as shown in *fig.19*, the mean of the control group and two treatments are different. Significant difference exists between the treatment 1 and the rest of the groups. The difference between control and treatment 2 group is approximately 0.5 less than 1. Alo based on the P-value result which Is less than the alpha (0.05) we concluded that difference exist between mean of the categories in ‘group’ categorical variable in plant df.

# where the difference is?

To be able to find where the difference is among the groups, we can conduct a 'Post-Hoc Testing' which are various test that allows us to see which group(s) significantly differ from each other (PYTHON FOR DATA SCIENCE, 2020b). Tukey's Honestly Significant Difference (HSD) statistical testing is used to test differences among sample means for significance, among other statistical method, it is designed for this purpose and fully controls Type I error rate (Lane, 2010). To implement this, I imported the ‘multicomp’ library from the stasmodel.stats library ‘importstatsmodels.stats.multicomp as mc’ as shown in *fig.20.*

A screenshot of a computer

Description automatically generated

Figure 20: Multicomparission table.

**Visualising the difference.**

Using HSD as implemented above, we were able to compare the groups together, to see where the difference in the mean of the unique values in the group column is. Result shows the difference in mean exist between the 'trt1' and 'trt2', also visually confirmed using the "Difference plot of plant-growth group mean" plot.

A graph with black lines

Description automatically generated

Figure 21: Difference plot of plant-growth group mean.

In conclusion, we have been able to conduct ANOVA testing on the plant-growth dataset from Pydataset. The outcome of the test resulted in us proving there is sufficient result to reject the H0, which means there are some sort of difference in the 'means' of the unique groups in the 'group' column in plant dataframe. HSD helped us to narrow down where the difference is exactly, result of which revealed that the difference is between 'trt1' and 'trt2' group.

# Question 4:

**Understanding question:**

Question 4a Wants us to perform an analysis that checks the correlation of the 3 numerical variables in tree df [Girth, Height, Volume] and hypothesis testing of the numerical variables correlation coefficient using Pearson function. Before hypothesis can be performed on the correlation coefficient between the three numerical variables, we must create a correlation matrix.

**What is correlation? (Data science, 2023)**

Correlation is a method that can help us to quantify the relationship between on variable and the other. It is represented by correlation coefficient ranging from -1 to 1. When the correlation coefficient between a variable and the other is -1 to 1 this indicates the direction and strength of the linear relationship between the variable variables. A positive correlation means that as one variable increases, the other also increases. A negative correlation means that as one variable increases, the other decreases. A zero correlation means that there is no linear relationship between the variables**.** It is however what noting that correlation does not necessarily mean causation.

**Stating hypothesis:**

**Girth and Height:**

1. Null-Hypothesis: There is no correlation between Girth and Height numerical variable

2. Alternate Hypothesis: There is correlation between Girth and Height numerical variable

**Girth and Volume:**

1. Null-Hypothesis: There is no correlation between Girth and Volume numerical variable

2. Alternate Hypothesis: There is correlation between Girth and Volume numerical variable

**Height and Volume:**

1. Null-Hypothesis: There is no correlation between Height and Volume numerical variable

2. Alternate Hypothesis: There is correlation between Height and Volume numerical variable

**Rejection region:**

We would reject the null hypothesis and accept the alternate hypothesis if the Pearson Correlation Coefficient is less than the P-Value. If the Pearsonr Correlation Coefficient is more than the P-value, we accept the null-hypothesis and reject the alternate hypothesis.

**Correlation Matrix:**

In other to test the correlation hypothesis of variable in tree df using the pearsonr, imported the pearsonr function from scipy.stats library.

**A table with numbers and text

Description automatically generated**

As shown in correlation matrix table, the diagonal value which is '1. 00000..' is a perfect correlation values, this represents the correlation of [girth, height, volume] with themselves. Other correlation-coefficient values in 'corr\_matrix' df shows that all the variables have a moderate to strong positive relationship because their correlation values are close to 1.

**Test Statistics:**

A screenshot of a computer

Description automatically generated

**Girth and Height:**

Pearson Correlation Coefficient: 0.519 Interpretation: There is a moderate positive linear relationship between girth and height. P-value: 0.00276

Interpretation: The p-value is less than the significance level of 0.05, indicating strong evidence against the null hypothesis (no correlation). Thus, there is a statistically significant correlation between girth and height.

**Girth and Volume:**

Pearson Correlation Coefficient: 0.967 Interpretation: There is a strong positive linear relationship between girth and volume. P-value: 8.64×10−19.

Interpretation: The p-value is extremely close to zero, significantly below the significance level of 0.05. This indicates very strong evidence against the null hypothesis and confirms a highly significant correlation between girth and volume.

**Height and Volume:**

Pearson Correlation Coefficient: 0.598 Interpretation: There is a moderate positive linear relationship between height and volume. P-value: 0.00038.

Interpretation: The p-value is less than the significance level of 0.05, indicating strong evidence against the null hypothesis. Hence, there is a statistically significant correlation between height and volume.

**Make a Decision:**

We reject the null-hypothesis for all pairs because all pairs of variables (Girth and Height, Girth and Volume, Height, and Volume) exhibit statistically significant correlations based on their respective p-values being less than 0.05.

**Conclusion:**

In conclusion, all pairs of variables (Girth and Height, Girth and Volume, Height, and Volume) exhibit statistically significant correlations based on their respective p-values being less than 0.05. The strength of the correlations varies, with Girth and Volume showing a very strong positive correlation, while Girth and Height, as well as Height and Volume, demonstrate moderate positive correlations.

# Regression Analysis:

I used ‘linearregression ()’ model from Sklearn to predict the volume of timber from trees df. Three model was fitted:

**Both model:**

This model takes both Girth and height from the tree df to predict volume of tree at 20%test and 80%traing sizes.

**Girth model:**

This model uses Girth from the tree df only, to predict the volume of tree, at 20%test and 80%traing sizes.

**Height Model:**

This model uses Height from the tree df only, to predict the volume of tree, at 20%test and 80%traing sizes.

In other to be able to fit the Girth and Height independent variables into their models, I had to convert them from 1D array to 2D array.

# Evaluating model performance:

To evaluate the performance of the models I used four matrixes namely:

1. Mean absolute error (MAE): (Chugh, 2020) defined it has the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset after prediction.
2. Mean Square Error (MSE): Mean square error represents the average of the squared difference between the original and predicted values in the data set. It quantifies the variance of the residuals after prediction (Chugh, 2020).
3. RMSE: Root means square error is the square root of MSE, it measures the standard deviation of the error.
4. R-squared (R2): R2 also referred to as the coefficient of determination represents the proportion of the variance in the dependent variable which is explained by the linear regression model (Chugh, 2020).A screenshot of a computer error

   Description automatically generated

Figure 24: 'Both' model evaluation result.

A screenshot of a computer error

Description automatically generated

Figure 25: Height model evaluation result.

A computer error message

Description automatically generated

Figure 25: Grith model evaluation result.

# Comparing results:

Overall, Girth and Both models performed betted predicting volume compared to actual volume than Height model. The **MSE and MAE** of ‘Both’ and ‘Girth’ model was low 3.66 and 3.04 respectively, this indicates they both had better accuracy at predicting close to the actual Volume of tree, with the Girth model performing best in this regard. The **Height model** had a significantly high **MSE and MAE** suggesting a low accuracy at predicting volume of tree. The 3 models had the same **RMSE**. Girth model had the highest **R-squared** of ≈ 0.9, this based on earlier reported definition indicates that it explains the variance of volume better than **Height and Both** models. Next to Girth is the ‘Both model’, then the Height model both having a lower r2 values ≈0.94 and ≈0.27 respectively compared to Girth.

# Conclusion:

In conclusion, at 20% testing and 80% training, Girth and Both models performed better at predicting the volume of Tree compared to Height Model. The Girth model however performed significantly better than the other two model at prediction the volume of tree, based on the evaluation matrixes reported above. In addition to Girth model performing better are predicting volume of tree, I plotted a relationship plot between Gith and volume and as shown in *fig.26*, it has a more linear relationship with volume than height does as shown in *fig.27*.

A graph with blue stars and white squares

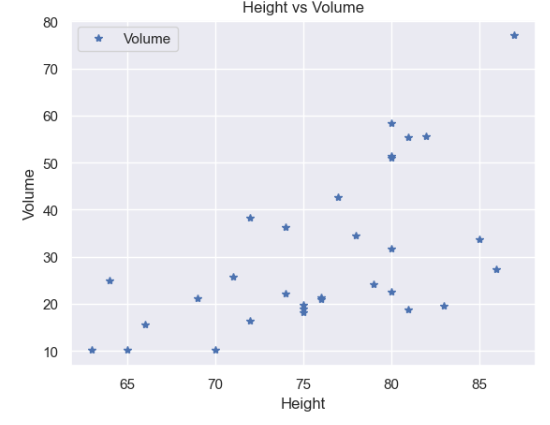
Description automatically generated

Figure 27: Height vs volume relationship plot.

Figure 26: Girth vs volume relationship plot.

# 

## References

Biswal, A. (2023). *What is Hypothesis Testing in Statistics? Types and Examples | Simplilearn*. [online] Simplilearn.com. Available at: https://www.simplilearn.com/tutorials/statistics-tutorial/hypothesis-testing-in-statistics [Accessed 19 Dec. 2023].

Casella, G., Fienberg, S. and Olkin, I. (2005). *A Modern Introduction to Probability and Statistics*. Springer Science+Business Media.

CCt (2023). *Statistical technique (Hypothesis testing 1)*. [online] moodle.cct.ie. Available at: https://moodle.cct.ie/mod/resource/view.php?id=147695 [Accessed 19 Dec. 2023].

code-mentor.ai. (2023). *AI Code Mentor: Best Tool for Code Optimization, Refactoring, and Review*. [online] Available at: https://code-mentor.ai/.

Codementor.ie (2023). *AI Code Mentor: Best Tool for Code Optimization, Refactoring, and Review*. [online] code-mentor.ai. Available at: https://code-mentor.ai/.

Data science (2023). *How can you determine correlation between variables in Python?* [online] www.linkedin.com. Available at: https://www.linkedin.com/advice/1/how-can-you-determine-correlation-between-variables#:~:text=corr()%20with%20Pearson%2C%20Spearman [Accessed 18 Dec. 2023].

Hayes, A. (2023). *Using the Variance Equation*. [online] Investopedia. Available at: https://www.investopedia.com/terms/v/variance.asp#:~:text=In%20statistics%2C%20variance%20measures%20variability.

Lane, D.M. (2010). *Tukey’s honestly significant difference (HSD)*. [online] Sagepub.com. Available at: https://methods.sagepub.com/reference/encyc-of-research-design/n478.xml [Accessed 12 Dec. 2023].

Ling 300 (2008). *Tutorial: Pearson’s Chi-square Test for Independence*. [online] Upenn.edu. Available at: https://www.ling.upenn.edu/~clight/chisquared.htm#:~:text=The%20degrees%20of%20freedom%20(often [Accessed 11 Dec. 2023].

Mahadevan, M. (2022). *Step-by-Step Exploratory Data Analysis (EDA) using Python -*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/.

Mulani, S. (2022). *Seaborn Distplot: A Comprehensive Guide | DigitalOcean*. [online] www.digitalocean.com. Available at: https://www.digitalocean.com/community/tutorials/seaborn-distplot [Accessed 8 Dec. 2023].

PYTHON FOR DATA SCIENCE (2020a). *Chi-square Test of Independence*. [online] www.pythonfordatascience.org. Available at: https://www.pythonfordatascience.org/chi-square-test-of-independence-python/#:~:text=Chi%2Dsquare%20test%20of%20independence%20with%20Scipy.&text=The%20method%20that%20needs%20to [Accessed 11 Dec. 2023].

PYTHON FOR DATA SCIENCE (2020b). *One-way ANOVA with Python*. [online] www.pythonfordatascience.org. Available at: https://www.pythonfordatascience.org/anova-python/ [Accessed 11 Dec. 2023].

simplilearn (2021). *What Is Descriptive Statistics? Definition, Applications, Examples & More | Simplilearn*. [online] Simplilearn.com. Available at: https://www.simplilearn.com/what-is-descriptive-statistics-article#types\_of\_descriptive\_statistics.

Statistics LibreTexts. (2015). *2.7: Skewness and the Mean, Median, and Mode*. [online] Available at: https://stats.libretexts.org/Bookshelves/Introductory\_Statistics/Introductory\_Statistics\_(OpenStax)/02%3A\_Descriptive\_Statistics/2.07%3A\_Skewness\_and\_the\_Mean\_Median\_and\_Mode#:~:text=Generally%2C%20if%20the%20distribution%20of.

Stojiljković, M. (2023). *Python Statistics Fundamentals: How to Describe Your Data – Real Python*. [online] realpython.com. Available at: https://realpython.com/python-statistics/#understanding-descriptive-statistics [Accessed 6 Nov. 2023].

Topre, G. (2022). *Bank Customer Churn Dataset*. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset.

Torres-Reyna, O. (2008). *Data Preparation & Descriptive Statistics*. [online] Available at: https://www.princeton.edu/~otorres/DataPrep101.pdf [Accessed 14 May 2022].

Varshney, P. (2023). *Q-Q Plots Explained | Built In*. [online] builtin.com. Available at: https://builtin.com/data-science/q-q-plot [Accessed 11 Dec. 2023].

Zach (2018). *Introduction to Hypothesis Testing*. [online] Statology. Available at: https://www.statology.org/hypothesis-testing/ [Accessed 19 Dec. 2023].

Zach (2020). *Chi-Square Test of Independence: Definition, Formula, and Example*. [online] Statology. Available at: https://www.statology.org/chi-square-test-of-independence/.