

**Programme Title:** HDip in Science in Data Analytics for Business (FT)

**Student Name:** Catalina Gonzalez

**Student number:** 2024458

**Cohort Details:** Sep 24 FT

**Module Title(s):** Statistical Techniques for Data Analysis

**Assignment Type:** Individual Weighting(s):60%

**Assignment Title:** CA2

**Lecturer(s):** Kayoum Khbuli

**Date:** 17/12/2024

Table of Contents

[Q1 3](#_Toc185681347)

[dIAMONDS 6](#_Toc185681348)

[PLANTGROWTH 9](#_Toc185681349)

[tHE TREES DATASET 12](#_Toc185681350)

# Q1

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.

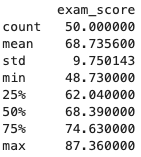
Analysis:

*import pandas as pd*

*df = pd.read\_csv("Q1.csv")*

*df.head()*



*print(df.describe())*

The dataset Q1 shows the first few rows of the data, which helps us check that everything has been loaded correctly. Then, we calculate the summary statistics using df.describe(). Here are the main results:

**Mean**:  
The average score is 68.74%, which means that overall, students are performing below the national average of 70%. However, this doesn't mean all students are performing the same way.

**Standard Deviation**:  
The standard deviation is 9.75%, which indicates moderate variation in the scores. This means the results are not the same for everyone; some students have scores that are much higher or lower than the average, showing there is a clear difference between top and bottom performers.

**Minimum and Maximum Values**:  
The lowest score is 48.73%, and the highest score is 87.36%. This range shows that student performance varies a lot: some students are struggling, while others are doing really well, suggesting a wide range of abilities.

**Percentiles**:  
Looking at the percentiles, we see that 25% of students scored below 62.04%, meaning a quarter of the class is performing at this level or lower. The median score is 68.39%, so half of the students are below this score, showing that most results are close to the average. Finally, 75% of students scored above 74.63%, which means that most of the class is performing better than the lower scores.

**The histogram show a distribution of the exam score**

The histogram shows a slightly left-skewed distribution (more students with lower scores), which is consistent with the mean being below 70%. This suggests that many students are scoring lower than the national average of 70%.

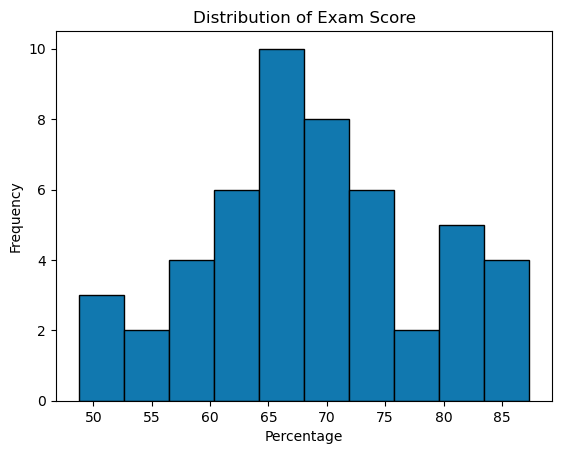
*df["exam\_score"].plot(kind="hist", bins=10, edgecolor="k")*

*plt.title("Distribution of Exam Score")*

*plt.xlabel("Percentage")*

*plt.ylabel("Frequency")*

*plt.show()*



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.

*t\_stat, p\_value = stats.ttest\_1samp(sample, mu)*

*print(f"T-statistic: {t\_stat}")*

*print(f"P-value: {p\_value}")*



The t-statistic of -0.92 shows that the difference between the average score of the students and the national average is small and negative. This means the students' average is a little lower than 70%. However, the p-value (0.3636) is much higher than the significance level of 0.05, which means there is not enough evidence to reject the null hypothesis. In simple terms, we cannot say that the students' performance is significantly worse than the national average of 70%.

*result = plt.hist(sample, bins=7,*

*color="c", edgecolor="k",*

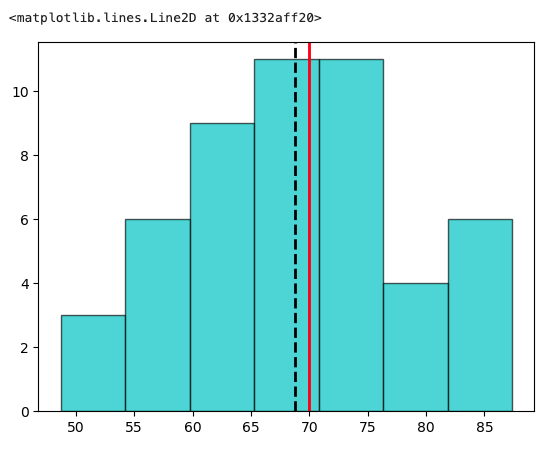
*alpha=0.65)*

*plt.axvline(sample.mean(), color="k",*

*linestyle="dashed", linewidth=2)*

*plt.axvline(mu, color="red",*

*linestyle="solid", linewidth=2)*



# dIAMONDS

1. 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).
   1. Create a new binary variable in the dataframe called “colourless” which records 1 in rows with colourless diamonds and 0 otherwise.
   2. 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.

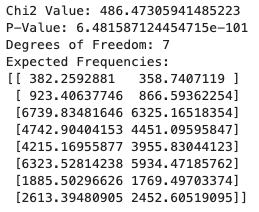
Analysis:

*print(f"Chi2 Value: {chi2}")*

*print(f"P-Value: {p}")*

*print(f"Degrees of Freedom: {dof}")*

*print(f"Expected Frequencies: \n{expected}")*

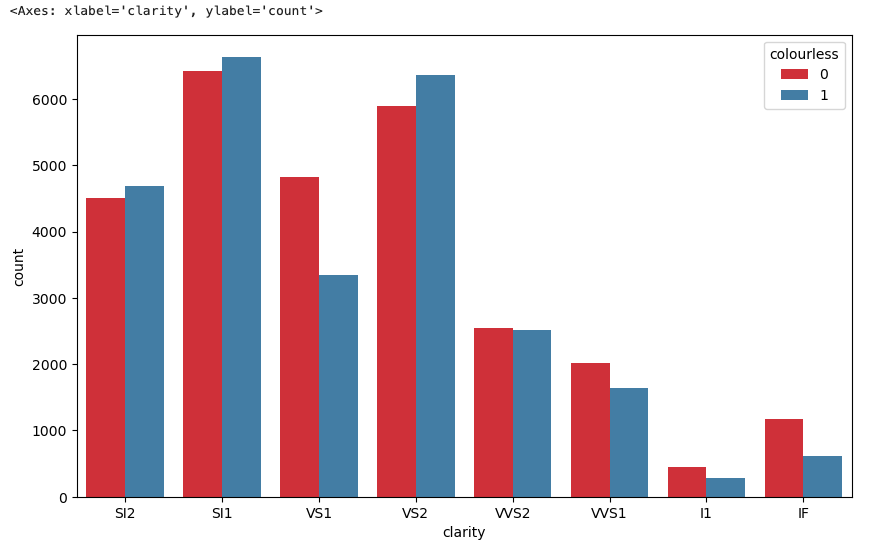


The results show a chi-square value of about 486.47 and a very small p-value (p-value ≈ 0). This means there is a statistically significant association between the clarity of the diamond and its color (colorless or not colorless) at a significance level of alpha = 0.01. Since the p-value is much smaller than 0.01, we reject the null hypothesis.

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

*plt.figure(figsize=(10, 6))*

*sns.countplot(data=df, x="clarity", hue="colourless", palette="Set1")*



The chart shows the distribution of colorless (and non-colorless) diamonds based on their clarity. As seen in the chart, colorless diamonds seem to be distributed differently according to their clarity, which is consistent with the results of the hypothesis test.

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

*ci\_incoloro = calculate\_ci(df\_incoloro["price"])*

*ci\_no\_incoloro = calculate\_ci(df\_no\_incoloro["price"])*

*print(f"90% Confidence Interval for Colorless Diamonds: {ci\_incoloro}")*

*print(f"90% Confidence Interval for Non-Colorless Diamonds: {ci\_no\_incoloro}")*



The 90% confidence interval for colorless diamonds is approximately 3301.86 to 3373.66.  
The 90% confidence interval for non-colorless diamonds is approximately 4448.78 to 4533.68.  
This suggests that colorless diamonds have a significantly lower average price compared to non-colorless diamonds, as the 90% confidence intervals do not overlap between the two groups.

# PLANTGROWTH

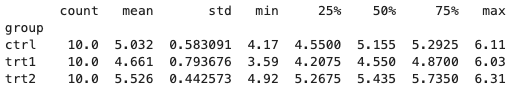
Load the PlantGrowth dataset from the pydataset library. It contains the results of a small study comparing the yields of plants obtained under a control and under two different treatment conditions.

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

Analysis:

*summary\_stats = plantgrowth.groupby('group')['weight'].describe()*

*print(summary\_stats)*



**count:** The number of observations (10 for each group).

**mean:** The average plant weight**.**

**std:** The standard deviation, which indicates how spread out the data is.

**min and max:** The minimum and maximum weight values.

**25%, 50%, 75%:** Percentiles that give us an idea of how the data is distributed.

**Boxplot:**

*plt.figure(figsize=(8, 6))*

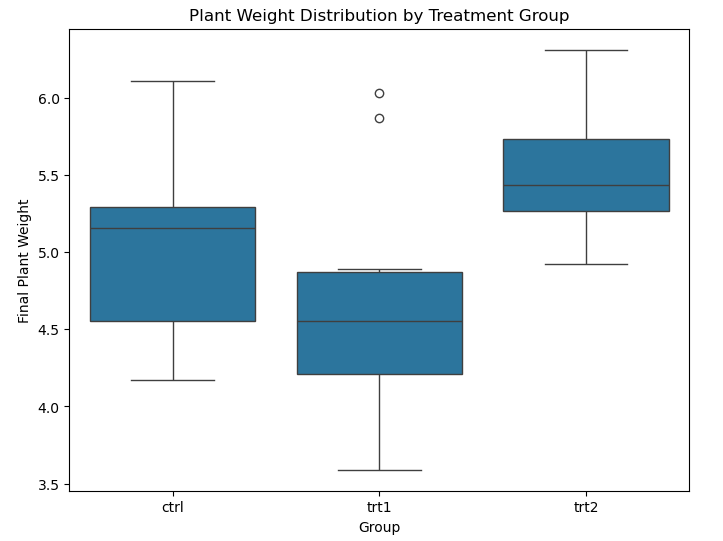
*sns.boxplot(data=plantgrowth, x='group', y='weight')*

*plt.title('Plant Weight Distribution by Treatment Group')*

*plt.xlabel('Group')*

*plt.ylabel('Final Plant Weight')*

*plt.show()*



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.

**Hypothesis Testing**

*f\_statistic, p\_value = f\_oneway(ctrl, trt1, trt2)*

*print(f"F-statistic: {f\_statistic}")*

*print(f"P-value: {p\_value}")*



**Analysis of variance (ANOVA)**

p-value (0.0159) is less than the significance level of 0.05, we **reject the null hypothesis**. This means there is enough evidence to suggest that there are significant differences in the plant weights between at least one of the groups.

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

*if p\_value < 0.05:*

*tukey\_result = pairwise\_tukeyhsd(plantgrowth['weight'], plantgrowth['group'], alpha=0.05)*

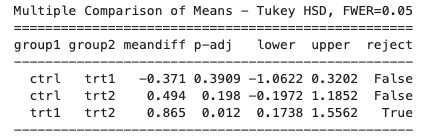
*print(tukey\_result)*

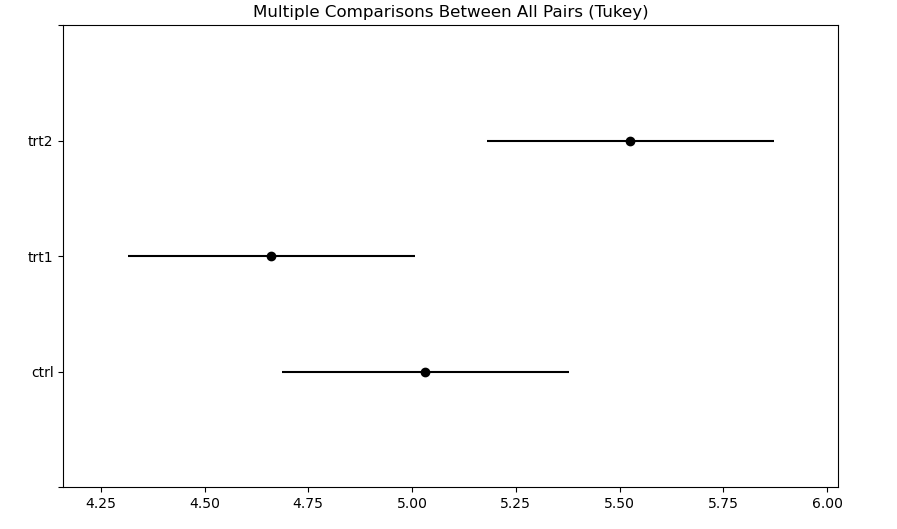
*tukey\_result.plot\_simultaneous()*

*plt.show()*

*else:*

*print("There is not enough evidence to reject the null hypothesis: no significant differences between groups.")*





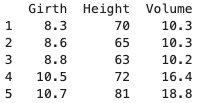
# tHE TREES DATASET

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).

Analysis:

*print(trees.head())*



**Girth and height**: A strong positive correlation was observed between the girth and the height of the trees, with a coefficient of 0.951. This suggests that, in general, as the girth of the tree increases, the height of the tree also tends to increase.

**Girth and volume**: The correlation between girth and volume is even stronger, with a coefficient of 0.973. This shows a close relationship between the two variables, meaning that trees with a larger girth tend to have a larger volume of wood.

**Height and volume**: We also found a strong positive correlation between the height of the trees and their volume, with a coefficient of 0.901. This suggests that taller trees generally have a larger volume of wood.

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.

linear regression model to predict the volume of wood based on the tree's girth:

**Regression coefficient**: 5.066

**Intercept**: -36.94

This means that for every 1-inch increase in the tree's girth, the volume of wood increases by about 5.066 cubic feet.

For example, if a tree has a girth of 15 inches, the model predicts that its volume would be 39.04 cubic feet.

**Analysis with Height**:

When we applied the tree's height to predict the wood volume:

**Regression coefficient**: 1.543

**Intercept**: -87.12

Here, for every 1-inch increase in the tree’s height, the volume of wood increases by 1.543 cubic feet

