**A****ssessment Cover Page**

**Student Full Name: Ilia Grishkin**

**Student Number: 2024139**

**Module Title: Statistical Techniques for Data Analysis**

**Assessment Title: CA2**

**Assessment Due Date: 19th-May-2024**

**Date of Submission: 18th-May-2024**

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.

Theory review 3

Z-test¶ 3

Hypotheses¶ 3

Types of test¶ 3

Test Statistic¶ 4

T-test¶ 4

Hypotheses¶ 4

Test Statistic¶ 4

Calculating metrics for the Testing¶ 5

Sample Mean¶ 5

Population Mean¶ 6

Alpha¶ 6

Critical Value¶ 6

Population Standard Deviation¶ 7

Sample Standard Deviation¶ 7

Sample Size¶ 7

Degree of Freedom¶ 8

Alpha 8

Critical Values¶ 8

Critical Values for Z-test¶ 8

Critical Values for T-test¶ 8

Degrees of Freedom for T-test¶ 9

Task 1 (Students' performance) 9

EDA 9

Distribution 10

Forming Hypothesis 10

Define metrics 11

Difine Test Type 12

Manual Calculation 12

Calculating using Scipy library 13

Calculating T-test using visual tools 13

Conclusion 14

Task 2 (Diamonds' clarity) 14

EDA 14

Associations between colour and clarity 15

Hypothesis 15

Manual calculation Using Chi-Squared test 15

Calculate frequency table 15

Calculate Expected Frequencies 16

Updated frequency tabled 19

Calculate the Chi-Squared Statistic 19

Calculating using Python libraries 23

Chi-squared 23

ANOVA 24

Tukey Kramer 24

Confidence intervals 27

Confidence intervals visualisation 28

Task 3 (PlantGrowth) 28

EDA 28

One-way ANOVA 29

Tukey's HSD test 29

Conclusion 30

Task 4 (Trees) 31

Pearson's correlation review 31

Correlations 32

Pearson's correlation coefficient 32

Scatter plot 33

Heatmap 33

Evaluation 34

Conclusion 35

References 36

Author 36

# Theory review

Z-test and T-test are commonly used approaches for Hypothesis testing. It was reviewed and explained by many different authors such as Sprinthall, R. (2011), John P. and Melvin L. (2006), Derrick, B. Toher, D. White, P. (2017).

## Z-test¶

The Z-test is used if the sample size is large enough ( n > 30 ) or the sample is normally distributed and the standard deviation is known.

### Hypotheses¶

Null Hypothesis (H₀): The population mean equals the hypothesized value.

Alternative Hypothesis (H₁): The population mean is not equal to the hypothesized value.

### Types of test¶

Two-tailed test

Left-tailed test

Right-tailed test

### Test Statistic¶

The Z-test statistic is calculated using the formula:

where:

**- X -** is the sample mean

**- -** is the hypothesized population mean

**- -** is the population standard deviation

**- n -**  is the sample size

## T-test¶

The T-test is used when the sample size is small (n<30) and the population variance is unknown.

### Hypotheses¶

Null Hypothesis (H₀): The population mean equals the hypothesized value.

Alternative Hypothesis (H₁): The population mean is not equal to the hypothesized value.

### Test Statistic¶

The T-test statistic is calculated using the formula:

where:

**- X -** is the sample mean

**- -** is the hypothesized population mean

**- s -** is the population standard deviation

**- n -**  is the sample size

**Comment:**

Both tests serve the same purpose which is the evaluation Null Hypothesis. However, there are a few conditions for applying each test. So, we need to make sure that we have all the metrics we need to complete the test. Additionally, important to understand which tail-type test we use.

## Calculating metrics for the Testing¶

### Sample Mean¶

The sample mean (X) is calculated as a summation of all values and dividing the sum by the number of records.

where:

**- n -** is the sample size

**- -** represents each individual sample value

### Population Mean¶

The population mean () is calculated identically to a sample mean, however, the number of records may be significantly larger.

where:

**- N -** is the population size

* **-** represents each individual population value

### Alpha¶

Alpha is a predefined level of significance.

Commonly used values:

alpha = 0.05 (5% significance level)

alpha = 0.01(1% significance level)

### Critical Value¶

The critical value depends on the type of test (Z-test or T-test), the significance level (alpha), and whether the test is one-tailed or two-tailed. The mentioned values help to define a critical value in the Z-table or T-table

### Population Standard Deviation¶

The population standard deviation (sigma) is calculated as the square root of the variance of the population. It shows how data is distributed

where:

- N - is the population size

**- -** represents each individual population value

**- -** is the population mean

### Sample Standard Deviation¶

The sample standard deviation (s) is calculated as the square root of the sample variance.

where:

**- n -** is the sample size

* **-** represents each individual sample value
* **-**  is the sample mean

### Sample Size¶

The sample size (n) is the number of records in the sample dataset.

### Degree of Freedom¶

The degrees of freedom (df) for a T-test are calculated differently depending on the type of T-test

## Alpha

The alpha value defines if the Null Hypothesis has to be rejected or not. The commonly used value is 0.05 which means that the hypothesis is applied to 95% of the sample. Each industry or science field defines its alpha. For example, in industries dealing with large datasets where many variables and noises such as betting, sales or retail, alpha should be strict. However, for industries such as physics, medicine, and chemistry alpha may be very small like (0.001) which means for 1000 observations alternative hypothesis must be approved 999 times.

## Critical Values¶

Critical values are predefined values which help to evaluate the hypothesis and make a decision to reject the Null hypothesis or not.

### Critical Values for Z-test¶

Critical value can be found in the z-table, where different z-scores depend on significance value (Alpha). Another approach is to use a scipy.stats library with package norm

Depending on the test type (One or Two-tailed) critical value may vary. For example:

One-tailed test with alpha = 00.5 the critical value is 1.645.

Two-tailed test with alpha = 00.5 the critical value is ±1.96.

### Critical Values for T-test¶

Similarly with Z-test, the critical value is defined by alpha using a T-table. However, there is an additional metric to define critical value for p-value such as degree of freedom (df). This value also can be retrieved using the same library scipy.stats but the package t

### Degrees of Freedom for T-test¶

The degrees of freedom (df) are used to slightly reduce a sample size. Natural data has some randomness. So, by excluding one observation we reduce the chances that an unrelated factor makes an impact on testing:

**One-sample T-test:**

**Independent two-sample T-test:**

**Paired (or dependent) two-sample T-test:**

# Task 1 (Students' performance)

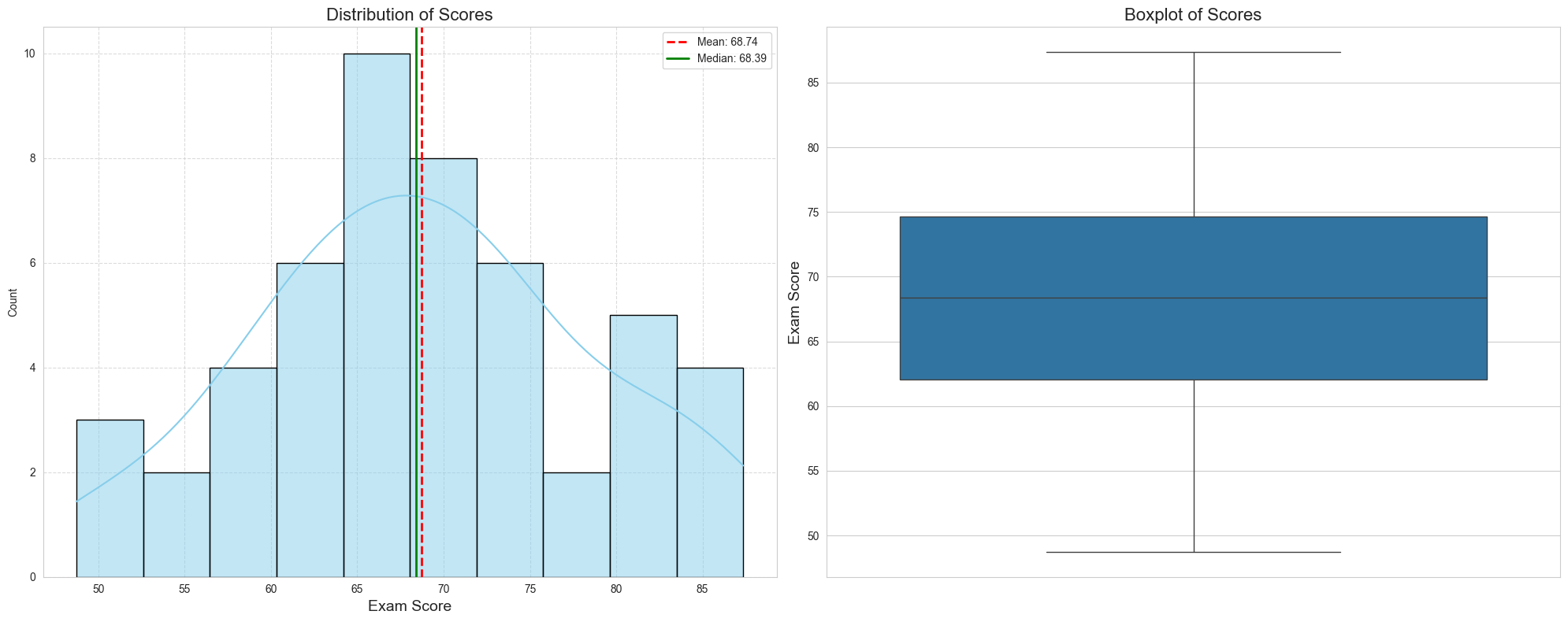
In this task, we are given a task to approve or reject teachers' suspicions about their students' exam scores.

We have been provided with a dataset of 50 exam results.

## EDA

* exam\_score  
  count 50.000000  
  mean 68.735600  
  std 9.750143  
  min 48.730000  
  25% 62.040000  
  50% 68.390000  
  75% 74.630000  
  max 87.360000

## Distribution



## Forming Hypothesis

In this case, the 50 students' dataset is a sample from the population of all students in the same conditions. The teacher's request can be formed as: "How accurately does the mean of the given dataset represent the population's mean value".

* **Null Hypothesis (H₀)**: There are no significant deviations for concluding that the sample does not represent the population. The sample's mean equals the population's mean.

* **Alternative Hypothesis (H₁)**: The sample's mean is significantly lower than the population's. Approving this hypothesis means that we should reject the Null Hypothesis.

## Define metrics

**Sample mean**

**Population mean**

**Alpha**

**Critical value**

Unknown yet, requires test-type.

**Population standard deviation**

Not given.

**Sample standard deviation**

**Sample size**

**Degree of freedom**

## Difine Test Type

Our sample is big enough to make a Z-test however we are missing a population's standard deviation. That is why in this case we are going to do a **T-test**.

Another parameter we need to define is the test tail-type based on the hypothesis we formed. It is going to be a **one/left-tail type**.

Now we can define critical value. According to the T-table for datasets with (**df = 49**) and (**alpha = 0.05**) **critical value is 1.677**

## Manual Calculation

To calculate the t-value we need to impute values we prepared into a T-test equation:

**T-value = -0.917**

**Critical value = - 1.677**

**P-value = 0.1829**

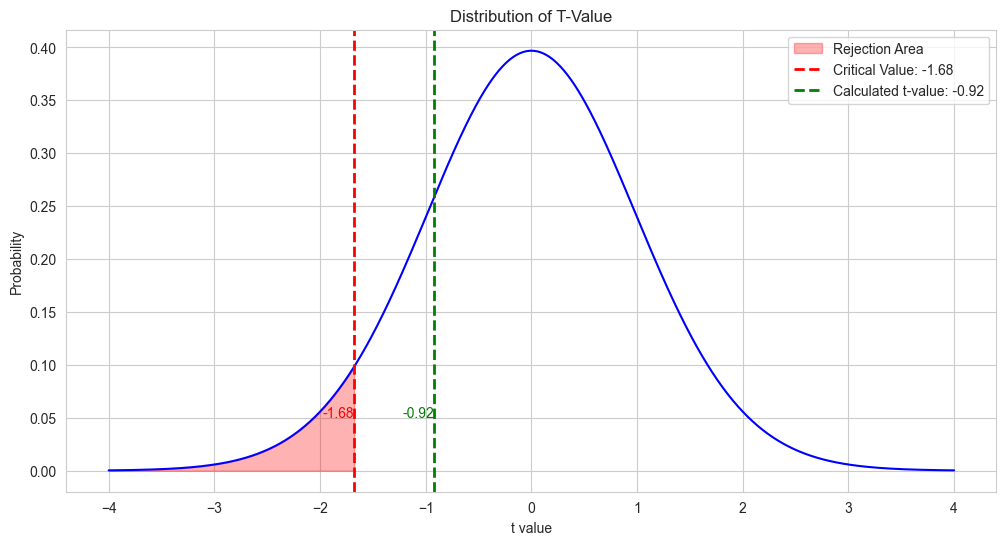
**-0.917 > - 1.677** -> T-value greater that Critical Value -> **Rejection the null hypothesis has failed**

## Calculating using Scipy library

sample\_mean = df['exam\_score'].mean()  
population\_mean = 70  
sample\_size = len(df)  
sample\_std = df['exam\_score'].std()  
alpha = 0.05  
t\_statistic = (sample\_mean - population\_mean) / (sample\_std / (sample\_size \*\* 0.5))  
df\_degrees\_of\_freedom = sample\_size - 1  
critical\_value = stats.t.ppf(alpha, df\_degrees\_of\_freedom)  
p\_value = stats.t.cdf(t\_statistic, df\_degrees\_of\_freedom)  
print(f"Sample Mean: {sample\_mean}")  
print(f"Population Mean: {population\_mean}")  
print(f"Sample Size: {sample\_size}")  
print(f"Sample Standard Deviation: {sample\_std}")  
print(f"T-statistic: {t\_statistic}")  
print(f"Critical Value: {critical\_value}")  
print(f"P-value: {p\_value}")  
**if** t\_statistic <= critical\_value:  
 print("Reject the null hypothesis.")  
**else**:  
 print("Rejection the null hypothesis has failed")

Sample Mean: 68.73559999999999  
Population Mean: 70  
Sample Size: 50  
Sample Standard Deviation: 9.75014288958613  
T-statistic: -0.9169771399834615  
Critical Value: -1.6765508926168542  
P-value: 0.18182298516031403  
Rejection the null hypothesis has failed

## Calculating T-test using visual tools



## Conclusion

Based on the data, statistical metrics and significance level provided we can conclude that there is not enough evidence to reject the Null Hypothesis. In other words, students' performance is not significantly lower than the average national score.

# Task 2 (Diamonds' clarity)

In this task, we need to test if there is any association between the two characteristics of diamonds.

We can perform a dependent two-sample test for proportions. We will compare the proportions of one variable across the levels of the other variable.

## EDA

**Statistics**

carat depth table price x \  
count 53940.000000 53940.000000 53940.000000 53940.000000 53940.000000   
mean 0.797940 61.749405 57.457184 3932.799722 5.731157   
std 0.474011 1.432621 2.234491 3989.439738 1.121761   
min 0.200000 43.000000 43.000000 326.000000 0.000000   
25% 0.400000 61.000000 56.000000 950.000000 4.710000   
50% 0.700000 61.800000 57.000000 2401.000000 5.700000   
75% 1.040000 62.500000 59.000000 5324.250000 6.540000   
max 5.010000 79.000000 95.000000 18823.000000 10.740000   
  
 y z   
count 53940.000000 53940.000000   
mean 5.734526 3.538734   
std 1.142135 0.705699   
min 0.000000 0.000000   
25% 4.720000 2.910000   
50% 5.710000 3.530000   
75% 6.540000 4.040000   
max 58.900000 31.800000

## Associations between colour and clarity

### Hypothesis

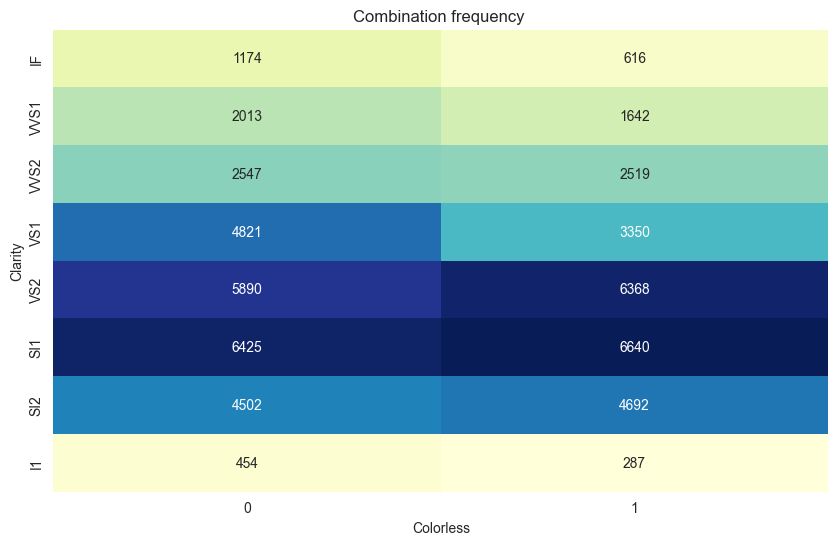
* **Null Hypothesis (H₀)**: There are no significant associations between diamonds colourless and clarity.
* **Chi-value lower than Critical value**
* **Alternative Hypothesis (H₁)**: There are no significant associations between diamonds' colourless and clarity.
* **Chi-value greater than Critical value**

### Manual calculation Using Chi-Squared test

For applying a Chi-Squared test we will follow the explanations and comments of authors Nikulin, M. (1996) and Fisher, R. (1922).

### Calculate frequency table

In this stage, we must determine how frequently each combination of two selected features occurs in the dataset. However, manual counting can be quite time-consuming because we have 53940 records. We can use **cross-tabulation** function which comes with pandas library to calculate combinations for us.



**Frequency table**

colourless 0 1 Total  
IF 1174 616 1790  
VVS1 2013 1642 3655  
VVS2 2547 2519 5066  
VS1 4821 3350 8171  
VS2 5890 6368 12258  
SI1 6425 6640 13065  
SI2 4502 4692 9194  
I1 454 287 741  
Total 27826 26114 53940

### Calculate Expected Frequencies

Where:

* **- -** is the total for row **i**
* **- -** is the total for column **j**
* - - is the total number of observations
* Expected frequency for cell (IF (Clarity cat.)), 0 (Not Colourless)):

* Expected frequency for cell (IF (Clarity cat.)), 1 (Colourless)):

* Expected frequency for cell (VVS1 (Clarity cat.)), 0 (Not Colourless)):

* Expected frequency for cell (VVS1 (Clarity cat.)), 1 (Colourless)):

* Expected frequency for cell (VVS2 (Clarity cat.)), 0 (Not Colourless)):

* Expected frequency for cell (VVS2 (Clarity cat.)), 1 (Colourless)):

* Expected frequency for cell (VS1 (Clarity cat.)), 0 (Not Colourless)):

* Expected frequency for cell (VS1 (Clarity cat.)), 1 (Colourless)):

* Expected frequency for cell (VS2 (Clarity cat.)), 0 (Not Colourless)):

* Expected frequency for cell (VS2 (Clarity cat.)), 1 (Colourless)):

* Expected frequency for cell (SI1 (Clarity cat.)), 0 (Not Colourless)):

* Expected frequency for cell (SI1 (Clarity cat.)), 1 (Colourless)):

* Expected frequency for cell (SI2 (Clarity cat.)), 0 (Not Colourless)):

* Expected frequency for cell (SI2 (Clarity cat.)), 1 (Colourless)):

* Expected frequency for cell (I1 (Clarity cat.)), 0 (Not Colourless)):

* Expected frequency for cell (I1 (Clarity cat.)), 1 (Colourless)):

### Updated frequency tabled

|  |  |  |  |
| --- | --- | --- | --- |
| Clarity | 0 "Not Colourless" (Expected freq.) | 1 "Colourless" (Expected freq.) | Total |
| IF | 1174 (923.4) | 616 (866.5) | 1790 |
| VVS1 | 2013 (1885.5) | 1642 (1769.4) | 3655 |
| VVS2 | 2547 (2613.3) | 2519 (2453.6) | 5066 |
| VS1 | 4821 (4215.1) | 3350 (3955.8) | 8171 |
| VS2 | 5890 (6323.5) | 6368 (5934.4) | 12258 |
| SI1 | 6425 (6739.8) | 6640 (6325.1) | 13065 |
| SI2 | 4502 (4742.9) | 4692 (4451) | 9194 |
| I1 | 454 (382.2) | 287 (358.7) | 741 |
| Total | 27826 | 26114 | 53940 |

### Calculate the Chi-Squared Statistic

Where:

**- -** is the observed frequency

**- -** is the expected frequency.

* Chi-Squared for cell (IF (Clarity cat.)), 0 (Not Colourless)):

* Chi-Squared for cell (IF (Clarity cat.)), 1 (Colourless)):

* Chi-Squared for cell (VVS1 (Clarity cat.)), 0 (Not Colourless)):

* Chi-Squared for cell (VVS1 (Clarity cat.)), 1 (Colourless)):

* Chi-Squared for cell (VVS2 (Clarity cat.)), 0 (Not Colourless)):

* Chi-Squared for cell (VVS2 (Clarity cat.)), 1 (Colourless)):

* Chi-Squared for cell (VS1 (Clarity cat.)), 0 (Not Colourless)):

* Chi-Squared for cell (VS1 (Clarity cat.)), 1 (Colourless)):

* Chi-Squared for cell (VS2 (Clarity cat.)), 0 (Not Colourless)):

* Chi-Squared for cell (VS2 (Clarity cat.)), 1 (Colourless)):

* Chi-Squared for cell (SI1 (Clarity cat.)), 0 (Not Colourless)):

* Chi-Squared for cell (SI1 (Clarity cat.)), 1 (Colourless)):

* Chi-Squared for cell (SI2 (Clarity cat.)), 0 (Not Colourless)):

* Chi-Squared for cell (SI2 (Clarity cat.)), 1 (Colourless)):

* Chi-Squared for cell (I1 (Clarity cat.)), 0 (Not Colourless)):

* Chi-Squared for cell (I1 (Clarity cat.)), 1 (Colourless)):

**Degree of freedom**

To calculate **Degree of freedom** in Chi-squared test we need to use the dimensions of the contingency table we created before.

df=(number of rows−1)×(number of columns−1)

**Critical value**

Based on the degree of freedom calculated and the significance level (alpha=0.01) provided we can determine the critical value **(18.48)**

**Decision**

**487.24 > 18.48**

**The Null Hypothesis at the 0.01 significance level is rejected.**

There is enough statistical evidence to declare that there is an association between "Clarity" and whether the diamond is colourless or not.

## Calculating using Python libraries

### Chi-squared

Chi-squared statistic: 486.4730594148522  
Significance level: 0.01  
Null Hypothesis rejected  
There is a significant association between the colour and clarity

**from** scipy.stats **import** chi2\_contingency  
df = sns.load\_dataset('diamonds')  
colourless\_values = []  
**for** index, row **in** df.iterrows():  
 **if** 'D' **in** row.values **or** 'E' **in** row.values **or** 'F' **in** row.values:  
 colourless\_values.append(1)  
 **else**:  
 colourless\_values.append(0)  
df['colourless'] = colourless\_values  
contingency\_table = pd.crosstab(df['clarity'], df['colourless'])  
chi2, p\_value, \_, \_ = chi2\_contingency(contingency\_table)  
alpha = 0.01  
print(f"Chi-squared statistic: {chi2}")  
print(f"Significance level: {alpha}")  
**if** p\_value < alpha:  
 print("Null Hypothesis rejected")  
 print("There is a significant association between the colour and clarity")  
**else**:  
 print("Rejection of Null Hypothesis has failed")  
 print("There is no significant association between the colour and clarity")

### ANOVA

F\_onewayResult(statistic=70.1182250986631, pvalue=2.2286936134194793e-101)

df = sns.load\_dataset('diamonds')  
colourless\_values = []  
**for** index, row **in** df.iterrows():  
 **if** 'D' **in** row.values **or** 'E' **in** row.values **or** 'F' **in** row.values:  
 colourless\_values.append(1)  
 **else**:  
 colourless\_values.append(0)  
df['colourless'] = colourless\_values  
  
group1 = df[(df.clarity == 'IF')]  
group2 = df[(df.clarity == 'VVS1')]  
group3 = df[(df.clarity == 'VVS2')]  
group4 = df[(df.clarity == 'VS1')]  
group5 = df[(df.clarity == 'VS2')]  
group6 = df[(df.clarity == 'SI1')]  
group7 = df[(df.clarity == 'SI2')]  
group8 = df[(df.clarity == 'I1')]  
  
stats.f\_oneway(group1['colourless'],  
 group2['colourless'],  
 group3['colourless'],  
 group4['colourless'],  
 group5['colourless'],  
 group6['colourless'],  
 group7['colourless'],  
 group8['colourless'])

**Comment:**

**F-statistic: 70.118**

A high F-statistic score indicates a greater degree of difference between the group's means relative to the variation within the groups.

**P-value: 2.2286936134194793e-101 (approx.0)**

The calculated p-value is much smaller than the given alpha level of 0.01, we can reject the null hypothesis.

**Conclusion:**

Since the p-value is extremely small, much less than the typical significance level of 0.01:

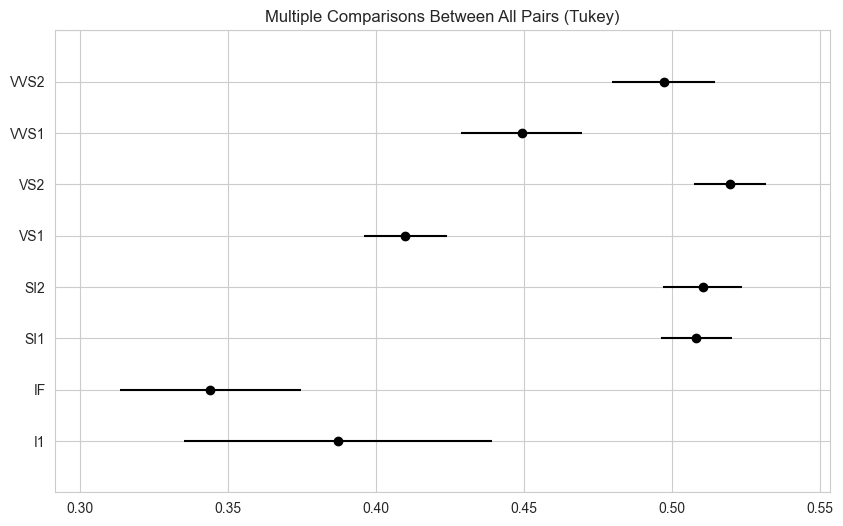
Reject the Null Hypothesis

Significant Differences Exist: There are significant differences in the proportions of colourless diamonds among the different clarity categories (IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1).

### Tukey Kramer

**from** statsmodels.stats.multicomp **import** pairwise\_tukeyhsd  
clarity\_levels = df['clarity']  
colourless\_values = df['colourless']  
tukey = pairwise\_tukeyhsd(endog=colourless\_values,   
 groups=clarity\_levels,   
 alpha=0.01)  
print(tukey)  
tukey.plot\_simultaneous()  
plt.show()

Multiple Comparison of Means - Tukey HSD, FWER=0.01   
====================================================  
group1 group2 meandiff p-adj lower upper reject  
----------------------------------------------------  
 I1 IF -0.0432 0.4908 -0.1198 0.0335 False  
 I1 SI1 0.1209 0.0 0.0547 0.1872 True  
 I1 SI2 0.123 0.0 0.056 0.19 True  
 I1 VS1 0.0227 0.9357 -0.0446 0.09 False  
 I1 VS2 0.1322 0.0 0.0658 0.1986 True  
 I1 VVS1 0.0619 0.042 -0.0088 0.1326 False  
 I1 VVS2 0.1099 0.0 0.0409 0.1789 True  
 IF SI1 0.1641 0.0 0.1199 0.2083 True  
 IF SI2 0.1662 0.0 0.1209 0.2115 True  
 IF VS1 0.0659 0.0 0.0201 0.1116 True  
 IF VS2 0.1754 0.0 0.131 0.2198 True  
 IF VVS1 0.1051 0.0 0.0545 0.1557 True  
 IF VVS2 0.1531 0.0 0.1049 0.2013 True  
 SI1 SI2 0.0021 1.0 -0.0218 0.026 False  
 SI1 VS1 -0.0982 0.0 -0.123 -0.0735 True  
 SI1 VS2 0.0113 0.6192 -0.0108 0.0333 False  
 SI1 VVS1 -0.059 0.0 -0.0918 -0.0261 True  
 SI1 VVS2 -0.011 0.8857 -0.04 0.018 False  
 SI2 VS1 -0.1003 0.0 -0.127 -0.0737 True  
 SI2 VS2 0.0092 0.8856 -0.015 0.0334 False  
 SI2 VVS1 -0.0611 0.0 -0.0954 -0.0268 True  
 SI2 VVS2 -0.0131 0.8054 -0.0438 0.0176 False  
 VS1 VS2 0.1095 0.0 0.0845 0.1346 True  
 VS1 VVS1 0.0393 0.0019 0.0043 0.0742 True  
 VS1 VVS2 0.0872 0.0 0.0559 0.1186 True  
 VS2 VVS1 -0.0702 0.0 -0.1033 -0.0372 True  
 VS2 VVS2 -0.0223 0.129 -0.0516 0.007 False  
 VVS1 VVS2 0.048 0.0002 0.0099 0.0861 True  
----------------------------------------------------



**Comment:**

The provided report compares the mean values of different clarity groups of diamonds, giving information about how each group differ based on a proportion of colourless.

**Significant difference**

If the reject column is True, there is a significant difference between the two groups at the significance level of 0.01.

If the reject column is False, there is no significant difference between the two groups.

**Direction of difference**

If meandiff is positive, the first group has a higher proportion of "colourless" diamonds.

If meandiff is negative, the second group has a higher proportion of "colourless" diamonds.

**Confidence intervals**

If the confidence interval (lower to upper) does not include 0, it confirms that the difference is statistically significant.

If the interval includes 0, it indicates that the difference is not statistically significant.

## Confidence intervals

df = sns.load\_dataset('diamonds')  
colourless\_values = []  
**for** index, row **in** df.iterrows():  
 **if** 'D' **in** row.values **or** 'E' **in** row.values **or** 'F' **in** row.values:  
 colourless\_values.append(1)  
 **else**:  
 colourless\_values.append(0)  
df['colourless'] = colourless\_values  
colourless = df[df['colourless'] == 1]  
non\_colourless = df[df['colourless'] == 0]  
**def** confidence\_interval(data, confidence=0.90):  
 n = len(data)  
 mean = np.mean(data)  
 se = stats.sem(data)  
 h = se \* stats.t.ppf((1 + confidence) / 2., n-1)  
 **return** mean, mean-h, mean+h  
colourless\_mean, colourless\_ci\_lower, colourless\_ci\_upper = confidence\_interval(colourless['price'])  
non\_colourless\_mean, non\_colourless\_ci\_lower, non\_colourless\_ci\_upper = confidence\_interval(non\_colourless['price'])  
print(f"Colourless Diamonds Mean Price: {colourless\_mean:.2f}")  
print(f"90% Confidence Interval: ({colourless\_ci\_lower:.2f}, {colourless\_ci\_upper:.2f})\n")  
print(f"Non-Colourless Diamonds Mean Price: {non\_colourless\_mean:.2f}")  
print(f"90% Confidence Interval: ({non\_colourless\_ci\_lower:.2f}, {non\_colourless\_ci\_upper:.2f})”)

Colourless Diamonds Mean Price: 3337.76  
90% Confidence Interval: (3301.86, 3373.66)  
  
Non-Colourless Diamonds Mean Price: 4491.23  
90% Confidence Interval: (4448.78, 4533.68)

**Mean prices:**

Colourless Diamonds Mean Price: 3337.76

Non-Colourless Diamonds Mean Price: 4491.23

This indicates that, on average, non-colourless diamonds are priced higher than colourless diamonds in the dataset.

**Confidence Intervals:**

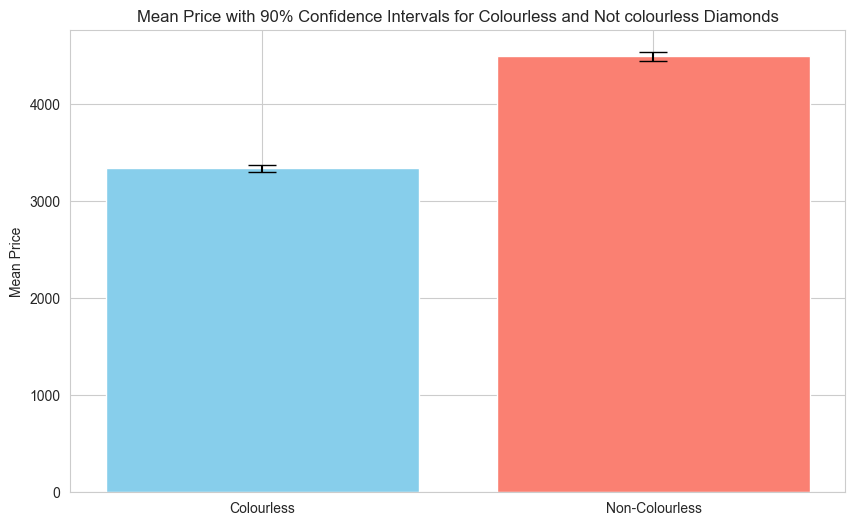
90% Confidence Interval for Colourless Diamonds: **3301.86 - 3373.66**

90% Confidence Interval for Non-Colourless Diamonds: **4448.78 - 4533.68**

The results mean that we can say with 90% that the mean values of diamonds will be placed within given intervals.

Additionally, we can conclude how colourless affects the price of diamonds.

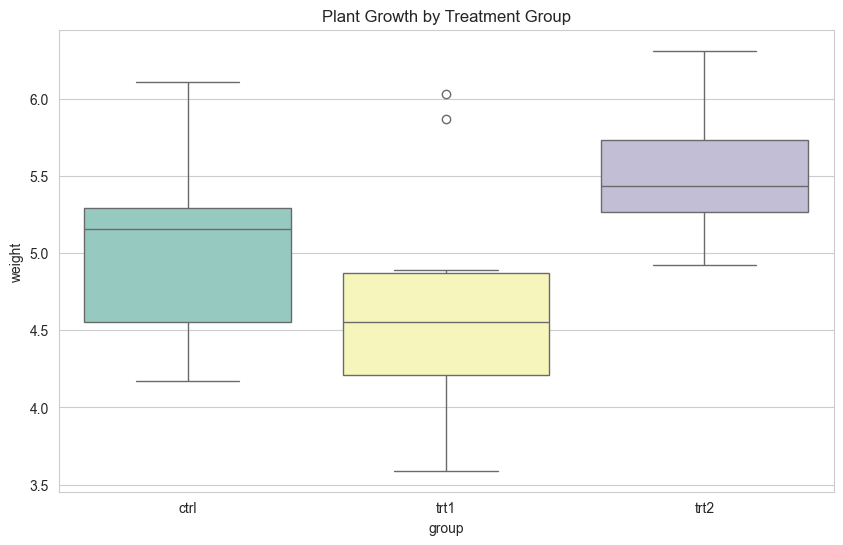
## Confidence intervals visualisation



# Task 3 (PlantGrowth)

### EDA

weight   
 count mean std min 25% 50% 75% max  
group   
ctrl 10.0 5.032 0.583091 4.17 4.5500 5.155 5.2925 6.11  
trt1 10.0 4.661 0.793676 3.59 4.2075 4.550 4.8700 6.03  
trt2 10.0 5.526 0.442573 4.92 5.2675 5.435 5.7350 6.31



### One-way ANOVA

control = df[df['group'] == 'ctrl']['weight']  
treatment1 = df[df['group'] == 'trt1']['weight']  
treatment2 = df[df['group'] == 'trt2']['weight']  
f\_stat, p\_value = f\_oneway(control, treatment1, treatment2)  
print(f'ANOVA F-statistic: {f\_stat:.4f}, p-value: {p\_value:.4f}’)

ANOVA F-statistic: 4.8461, p-value: 0.0159

### Tukey's HSD test

Because ANOVA testing performed significant differences between groups we can test our hypothesis further using Tukey's HSD test.

**if** p\_value < 0.05:  
 tukey = pairwise\_tukeyhsd(endog=df['weight'], groups=df['group'], alpha=0.05)  
 print(tukey)  
**else**:  
 print("No significant differences found between groups.”)

Multiple Comparison of Means - Tukey HSD, FWER=0.05  
===================================================  
group1 group2 meandiff p-adj lower upper reject  
---------------------------------------------------  
 ctrl trt1 -0.371 0.3909 -1.0622 0.3202 False  
 ctrl trt2 0.494 0.198 -0.1972 1.1852 False  
 trt1 trt2 0.865 0.012 0.1738 1.5562 True  
---------------------------------------------------

# Conclusion

Based on the p-value (0.0159) using the ANOVA test and significance level provided (alpha=0.05) we can conclude that the Null Hypotesis is rejected. In other words, there is a significant difference between different treatments.

Results of ANOVA testing justify a more detailed analysis using Tukey testing. Tukey testing compares the means of each group by using the following metrics:

* **group1, group2**
* The groups being compared.
* **meandiff**
* The difference in means between the two groups.
* **p-adj**
* The adjusted p-value for the comparison.
* **lower, upper**
* The lower and upper bounds of the 95% confidence interval for the mean difference.
* **reject**
* Indicates whether the null hypothesis (that there is no difference between the group means) is rejected.

**(ctrl) vs (trt1)**

|0.05 > 0.3909| -> |Alpha > P-value| -> |T-value greater that Critical Value| -> |**Rejection the null hypothesis has failed**|

**(ctrl) vs (trt2)**

|0.05 > 0.198| -> |Alpha > P-value| -> |T-value greater that Critical Value| -> |**Rejection the null hypothesis has failed**|

**(trt1) vs (trt2)**

|0.05 < 0.012| -> |Alpha < P-value| -> |T-value lower that Critical Value| -> |**Rejection the null hypothesis**|

Only this pair of treatment methods showed a significant difference in sample means.

# Task 4 (Trees)

# Pearson's correlation review

A detailed explanation of the use of Pearson's correlation equation and methods of implementation can be found in studies such as Pearson, K. (1900), Kendall, M. and Stuart, A. (1973).

The Pearson correlation coefficient (**r**) is calculated using the following formula:

Where:

- and - are the individual data points

- and - are the means of the **x** and **y** data sets, respectively.

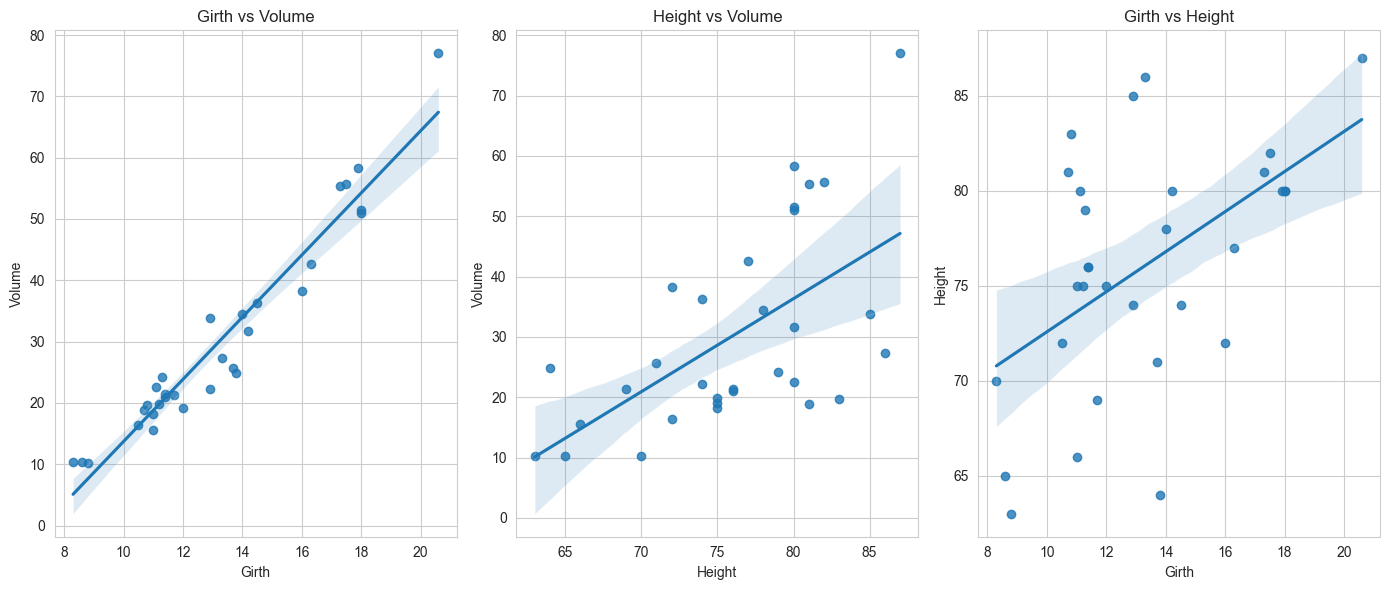
## Correlations

### Pearson's correlation coefficient

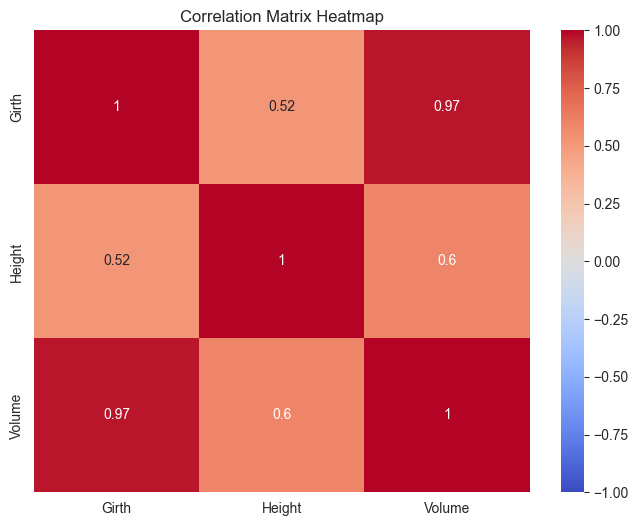
girth\_height\_corr, p\_value\_gh = pearsonr(df['Girth'], df['Height'])  
girth\_volume\_corr, p\_value\_gv = pearsonr(df['Girth'], df['Volume'])  
height\_volume\_corr, p\_value\_hv = pearsonr(df['Height'], df['Volume'])  
  
 print("No significant differences found between groups.”)

* Girth Height Volume  
  Girth 1.000000 0.51928 0.967119  
  Height 0.519280 1.00000 0.598250  
  Volume 0.967119 0.59825 1.000000  
  Correlation between Girth and Height: 0.5193, p-value: 0.0028  
  Correlation between Girth and Volume: 0.9671, p-value: 0.0000  
  Correlation between Height and Volume: 0.5982, p-value: 0.0004

## Scatter plot



### Heatmap



### Evaluation



## Conclusion

Comparing 3 combinations of independent features models performed the best accuracy with "Girth" and "Girth+Height" for predicting "Volume". However, the combination "Girth+Height" has shown less overall distance between training and testing R², which may indicate that the use of those features may prevent overfitting and inaccurate prediction in terms of new data.

Models such as **Linear Regression, Ridge Regression and Lasso Regression** performed approximately the same level of accuracy in training and testing sets, as well as an appropriate ratio between training and testing sets, which indicates that those models will perform similarly a new data.

Tree-based models such as **Decision Tree and Random Forest** performed well in the training set but in testing in part they could perform with the same accuracy. It means models are overfitted.

Models which rely on distances such as **SVR and KNN** gave different results, which may be related to scaling issues, which is very important for SVR.

Overall we can see that linear and vector-based models perform much more steadily than others, although they were not the best across all samples.

The best combination of features that may be used for prediction target "Volume" is "Grid" and "Grid+Height".

## References

Sprinthall, R. C. (2011). Basic statistical analysis (9th ed.). Pearson Education.

John P. K. and Melvin L. M. (2011). Survival Analysis (Hypothesis Testing) pp. 201–242.

Derrick, B. Toher, D. White, P. (2017). "How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017)" (PDF). The Quantitative Methods for Psychology. 13 (2): 120–126.

Nikulin, M. S. (1996), A guide to chi-squared testing, New York: Wiley.

Fisher R. A. (1922). "On the Interpretation of χ2 from Contingency Tables, and the Calculation of P". Journal of the Royal Statistical Society. 85 (1): 87–94.

Pearson K. (1900). "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". Philosophical Magazine. Series 5. 50 (302): 157–175.

Kendall, M. G. and Stuart, A. (1973) The Advanced Theory of Statistics, Volume 2: Inference and Relationship, Griffin.

### Author

**Ilia Grishkin**

<https://github.com/Ilia-Grishkin>