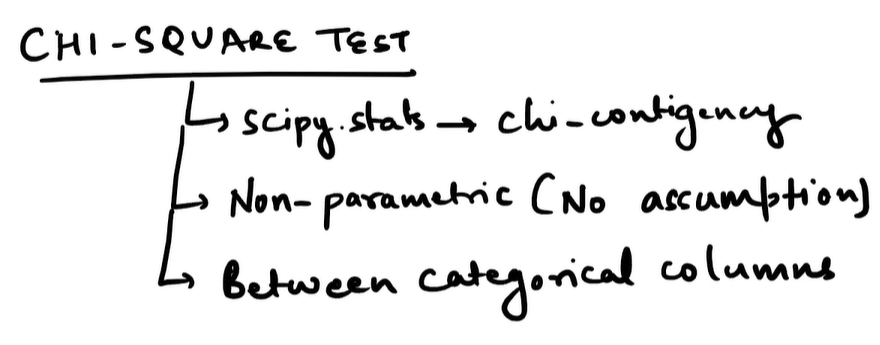
Day 10: Chi-Square Test:



**Chi-Square Test Overview**

The Chi-Square Test is a **non-parametric statistical test** used to determine if there's a significant association between **categorical variables**.

**The Chi-Square Test is applied when you want to check if two categorical variables are related or if observed data fits an expected distribution. It’s best used for frequency counts, not continuous data. For example, you can test whether gender is associated with voting preference or whether dice rolls follow a uniform distribution.**

Non-parametric tests are statistical methods that do not assume the data follows a normal distribution. Instead, they work directly with ranks, counts, or categories, making them ideal for skewed, ordinal, or categorical data. The Chi-Square test is one such non-parametric test.

 **Definition**: A statistical test that compares observed frequencies with expected frequencies under a null hypothesis.

 **Type of Data**: Works only with **categorical variables** (e.g., Male/Female, Yes/No, Red/Blue/Green).

 **Nature**: **Non-parametric** (no assumption of normal distribution).

 **Purpose**: To test **independence** between variables or **goodness-of-fit** to a theoretical distribution.

**📘 What Does "Non-Parametric" Mean?**

* **Parametric tests** (like t-tests, ANOVA) assume:
  + Data is normally distributed.
  + Variances are equal across groups.
  + Data is measured on an interval or ratio scale.
* **Non-parametric tests**:
  + Do **not** require normal distribution.
  + Can handle **ordinal** (ranked) or **categorical** data.
  + Are often called **distribution-free tests** because they don’t rely on strict assumptions about the population.

**🧪 Why Is Chi-Square Non-Parametric?**

* The **Chi-Square Test** compares **observed frequencies** with **expected frequencies**.
* It doesn’t require the data to be continuous or normally distributed.
* Works purely on **counts in categories** (e.g., Male/Female, Yes/No).
* **Assumption:** Only that the sample size is large enough (expected frequency ≥ 5 per cell).

**🧩 When to Apply Chi-Square**

| **Scenario** | **Why Use Chi-Square** | **Example** |
| --- | --- | --- |
| **Test of Independence** | To see if two categorical variables are related | Is smoking habit independent of gender? |
| **Goodness-of-Fit Test** | To check if observed data matches expected distribution | Do dice rolls follow a uniform distribution? |
| **Homogeneity Test** | To compare distributions across groups | Do different cities have similar voting patterns? |

**Conditions to apply:**

* Data must be in **frequency counts** (not percentages).
* Categories must be **mutually exclusive**.
* Expected frequency in each cell should be ≥ 5 for validity.

**⚖️ Comparison Table**

| **Feature** | **Parametric Test** | **Non-Parametric Test** |
| --- | --- | --- |
| Data Type | Continuous (interval/ratio) | Ordinal or categorical |
| Assumptions | Normal distribution, equal variance | No distribution assumption |
| Examples | t-test, ANOVA | Chi-Square, Mann–Whitney U, Kruskal–Wallis |
| Use Case | Comparing means | Comparing ranks or frequencies |

**🧪 Real-Time Example: Market Research**

Imagine a company wants to know if **product preference** depends on **age group**.

**Data Collected:**

| **Age Group** | **Prefers Product A** | **Prefers Product B** |
| --- | --- | --- |
| 18–30 | 40 | 20 |
| 31–50 | 35 | 25 |
| 51+ | 25 | 30 |

**Applying Chi-Square:**

1. Build a **contingency table**.
2. Run scipy.stats.chi2\_contingency() in Python.
3. Interpret:
   * **Null Hypothesis (H₀)**: Age group and product preference are independent.
   * **Alternative Hypothesis (H₁)**: Age group and product preference are related.
   * If **p-value < 0.05**, reject H₀ → Age influences product preference.

**🌍 Real-Life Applications**

* **Healthcare**: Is disease occurrence linked to lifestyle factors?
* **Education**: Are exam results associated with teaching methods?
* **Politics**: Is voting preference related to income level?
* **Retail**: Is product choice linked to customer demographics?

👉 So, whenever your data is **categorical, ordinal, skewed, or violates parametric assumptions**, non-parametric tests are the right choice.

**The Chi-Square Test is most often used to check if two categorical variables are related. Real-world examples include testing associations in healthcare, marketing, education, and social sciences.**

**🌍 Real-Life Examples of Chi-Square Tests**

**1. Healthcare**

* **Scenario**: A hospital wants to know if **smoking status** (smoker/non-smoker) is associated with **lung disease occurrence** (yes/no).
* **Chi-Square Use**: Build a contingency table of patients and test independence.
* **Outcome**: If p-value < 0.05, conclude smoking and lung disease are related.

**2. Marketing**

* **Scenario**: A retail company studies whether **gender** (male/female) influences **product preference** (Product A/Product B).
* **Chi-Square Use**: Compare observed purchase counts with expected counts.
* **Outcome**: Helps tailor marketing campaigns to specific demographics.

**3. Education**

* **Scenario**: A school investigates if **teaching method** (traditional/online) affects **pass/fail rates**.
* **Chi-Square Use**: Contingency table of teaching method vs exam outcome.
* **Outcome**: Reveals whether teaching style impacts student success.

**4. Politics**

* **Scenario**: A survey checks if **income level** (low/middle/high) is associated with **voting preference** (Party X/Party Y).
* **Chi-Square Use**: Test independence between income and voting choice.
* **Outcome**: Identifies socio-economic factors influencing elections.

**5. Retail & Consumer Behavior**

* **Scenario**: A supermarket tests if **day of the week** (weekday/weekend) affects **purchase of organic products** (yes/no).
* **Chi-Square Use**: Compare observed vs expected purchase frequencies.
* **Outcome**: Guides stocking and promotional strategies.

**📊 Why Chi-Square Works Here**

* All examples involve **categorical variables** (e.g., gender, product choice, teaching method).
* The test checks if **observed frequencies differ significantly from expected frequencies** under independence.
* It’s **non-parametric**, so no normal distribution assumption is needed.

**Key Points Explained**

1. scipy.stats → chi2\_contingency
   * This refers to the **Python implementation** using the scipy.stats module.
   * chi2\_contingency() is a function that performs the **Chi-Square test of independence** on a **contingency table** (a 2D array showing frequency counts).
   * Example usage:

Python

import numpy as np

from scipy.stats import chi2\_contingency

# Step 1: Create a contingency table

# Example: Gender vs Drink Preference

# Rows: Gender (Male, Female)

# Columns: Drink (Tea, Coffee)

data = np.array([[30, 20], [25, 35]])

# Step 2: Run the test

chi2, p, dof, expected = chi2\_contingency(data)

# Step 3: Interpret results

print("Chi2 Statistic:", chi2)

print("p-value:", p)

print("Degrees of Freedom:", dof)

print("Expected Frequencies:\n", expected)python

**Output Interpretation:**

* **Chi2 Statistic**: Measures how much observed values deviate from expected ones.
* **p-value**: If below significance level (e.g., 0.05), reject the null hypothesis.
* **Degrees of Freedom**:
* **Expected Frequencies**: What you'd expect if variables were independent.

**📘 Assumptions & Limitations**

* **Sample Size**: Expected frequency in each cell should be ≥ 5.
* **Independence**: Observations must be independent.
* **No continuous data**: You must bin continuous variables into categories.

**✅ Practical Use Cases**

* **Marketing**: Is product preference linked to age group?
* **Healthcare**: Is disease occurrence related to region?
* **Education**: Are test outcomes associated with teaching methods?

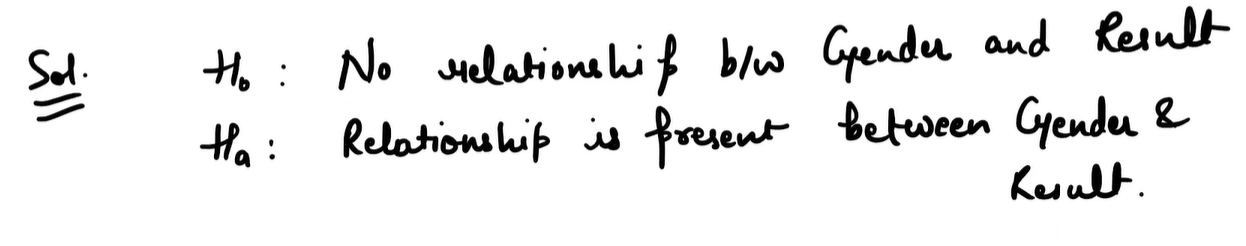
1. **Non-parametric (No assumption)**
   * The test **does not assume** a normal distribution.
   * It's suitable for **categorical data**, where parametric tests (like t-tests) aren't applicable.
   * It only assumes that the sample size is large enough for expected frequencies to be valid (usually ≥ 5 per cell).
2. **Between categorical columns**
   * The test is used to check **independence** between two categorical variables.
   * For example: Is gender independent of product preference? Is education level associated with voting behavior?

**What Is the Chi-Square Test?**

The **Chi-Square Test** evaluates whether observed frequencies differ significantly from expected frequencies under a specific hypothesis. It’s commonly used in two scenarios:

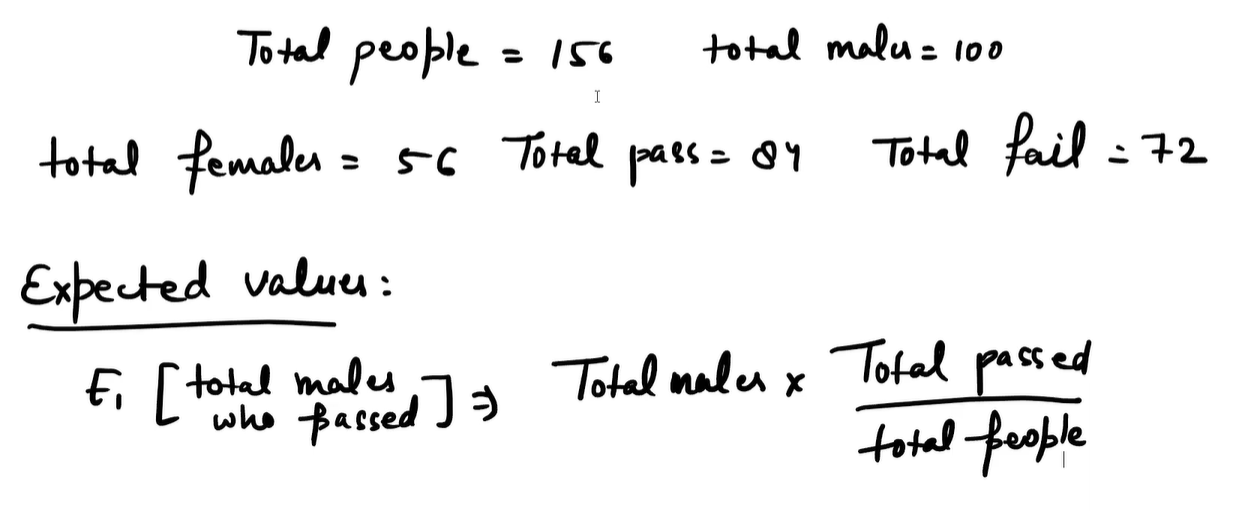
| **Type** | **Purpose** | **Example** |
| --- | --- | --- |
| **Test of Independence** | Checks if two categorical variables are related | Is gender associated with voting preference? |
| **Goodness-of-Fit Test** | Checks if a sample matches a theoretical distribution | Do dice rolls follow a uniform distribution? |





**🧠 Interpretation**

* **Null Hypothesis (H₀)**: Gender and result are independent.
* **Alternative Hypothesis (H₁)**: Gender and result are related.
* **p-value**: If less than 0.05, reject H₀ → conclude that gender affects result.



This is a classic setup for a **Chi-Square Test of Independence**, used to determine whether there's a statistically significant relationship between two categorical variables: **Gender** and **Result (Pass/Fail)**.

**📊 The Contingency Table**

| **Gender** | **Pass** | **Fail** | **Total** |
| --- | --- | --- | --- |
| Male | 60 | 40 | 100 |
| Female | 24 | 32 | 56 |
| Total | 84 | 72 | 156 |

This table shows how many males and females passed or failed. The question is: **Is the result (pass/fail) independent of gender?**

**📊 What Is a Contingency Table?**

* **It shows how the categories of one variable relate to the categories of another.**
* **Each cell in the table represents the count (or frequency) of observations that fall into that combination of categories.**
* **It’s the foundation for the Chi-Square Test of Independence.**

Cross-tabulation (often called a **crosstab**) is the process of creating a **contingency table** that summarizes the relationship between two categorical variables.

**🔍 What Cross-Tabulation Does**

* **Counts frequencies**: It shows how many observations fall into each combination of categories.
* **Organizes data**: Rows represent one variable (e.g., Gender), columns represent another (e.g., Result).
* **Prepares for analysis**: This table is the input for the **Chi-Square Test of Independence**, which checks if the two variables are related.

When you apply **cross-tabulation** (pd.crosstab in Python), it transforms the raw data into a summary table:

python

import pandas as pd

# Example DataFrame

df = pd.DataFrame({

'Gender': ['Male']\*100 + ['Female']\*56,

'Result': ['Pass']\*60 + ['Fail']\*40 + ['Pass']\*24 + ['Fail']\*32

})

# Crosstab

table = pd.crosstab(df['Gender'], df['Result'])

print(table)

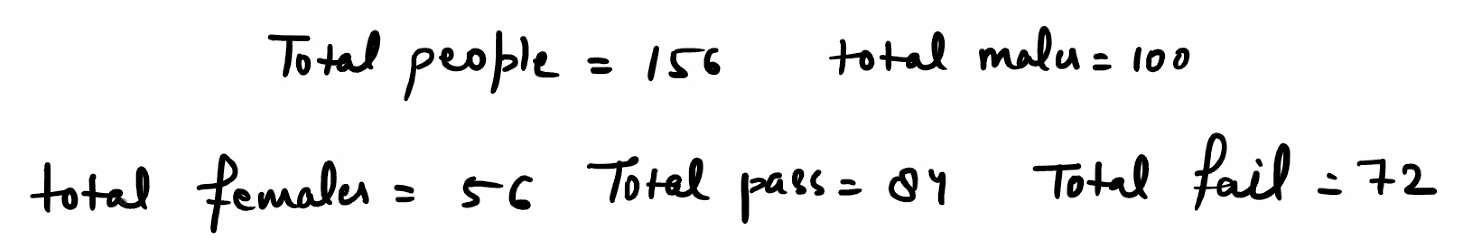
Output:

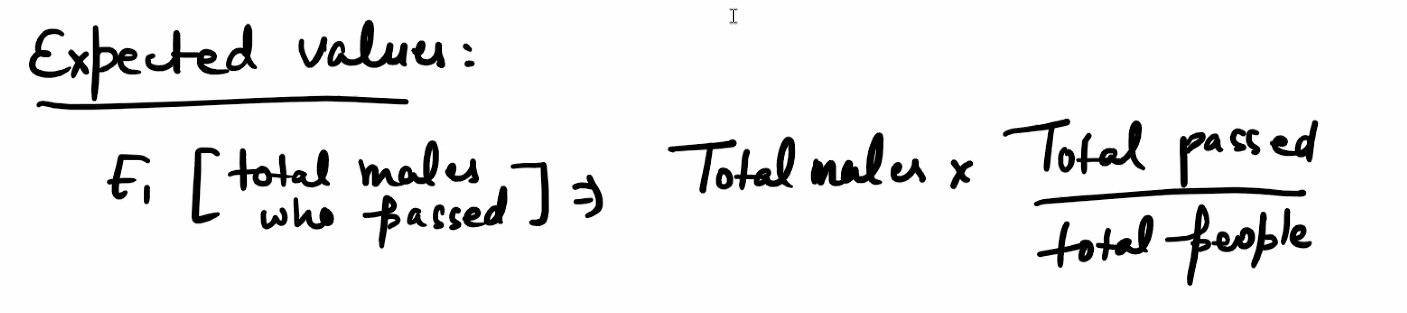
| **Result** | **Fail** | **Pass** |
| --- | --- | --- |
| Gender |  |  |
| Male | 40 | 60 |
| Female | 32 | 24 |

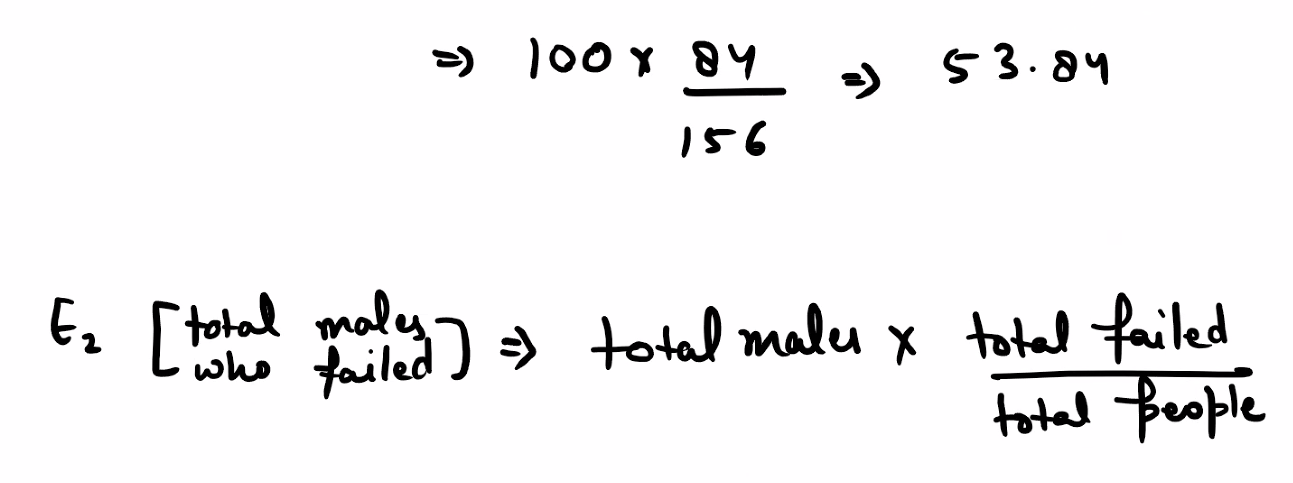
**🧠 Why It Matters**

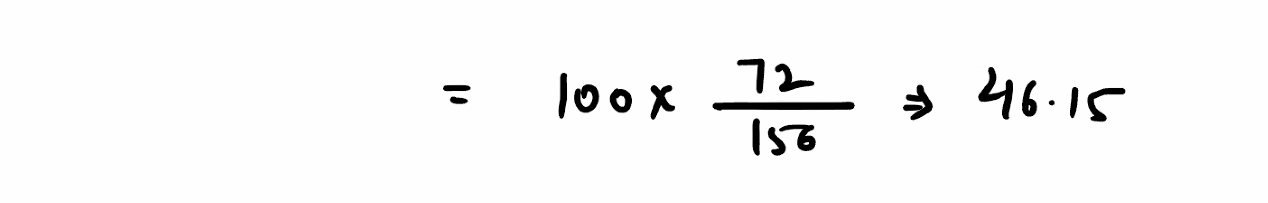
* **Simplifies analysis**: Instead of looking at hundreds of rows, you get a compact summary.
* **Foundation for Chi-Square**: The Chi-Square test uses this table to calculate expected frequencies and test independence.
* **Decision-making**: Helps you see patterns (e.g., more males passed, more females failed) before statistical testing.

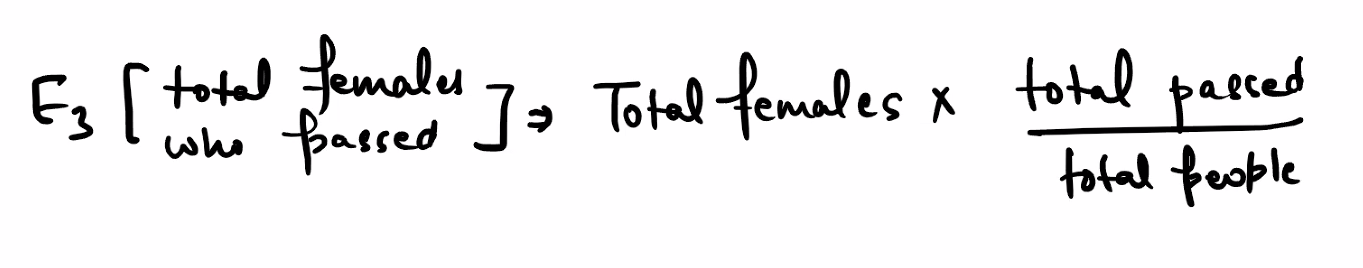
👉 In short: **Cross-tabulation converts raw categorical data into a structured contingency table, making it possible to run statistical tests like Chi-Square.**

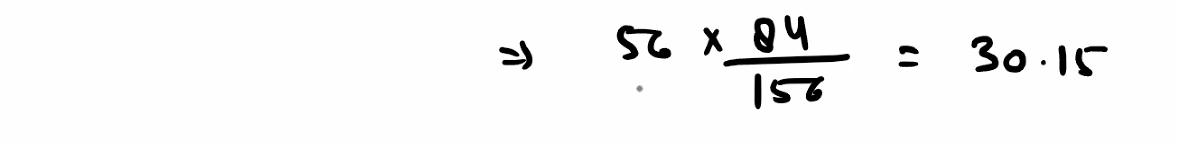


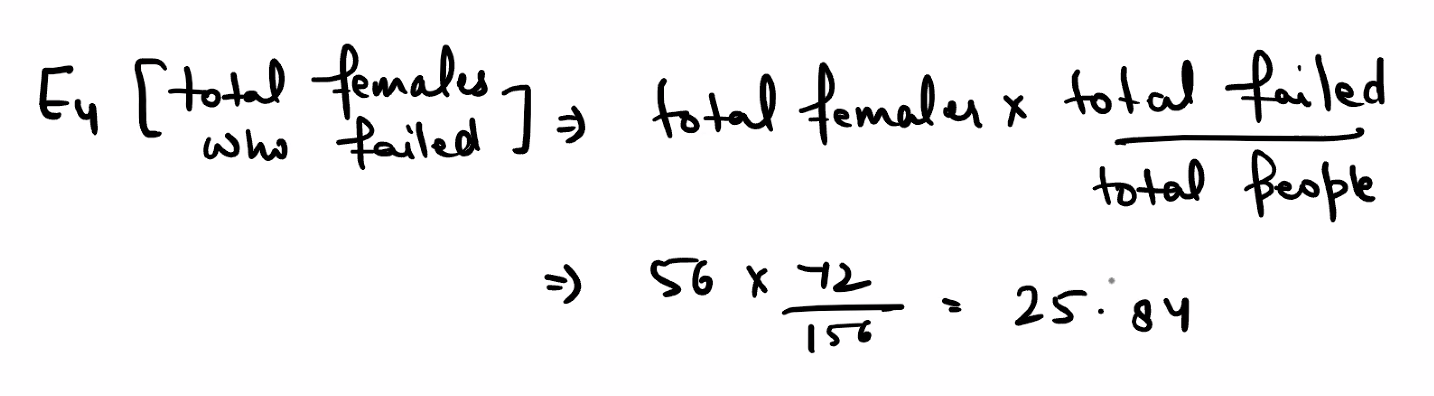












You want to know: *How many males would be expected to pass if gender and result were independent?*

**🧮 Expected Frequency Formula**

To compute the **expected value** for a cell in a contingency table:

In this case, you're calculating:

Substitute the values:

So, **if gender and result were independent**, you'd expect **53.85 males to pass**.

**🔍 Why This Matters**

* The **Chi-Square Test** compares this expected value with the **actual observed value** (which is 60 males passed).
* The difference between observed and expected values across all cells is used to compute the **Chi-Square statistic**.
* A large difference suggests a relationship between gender and result.

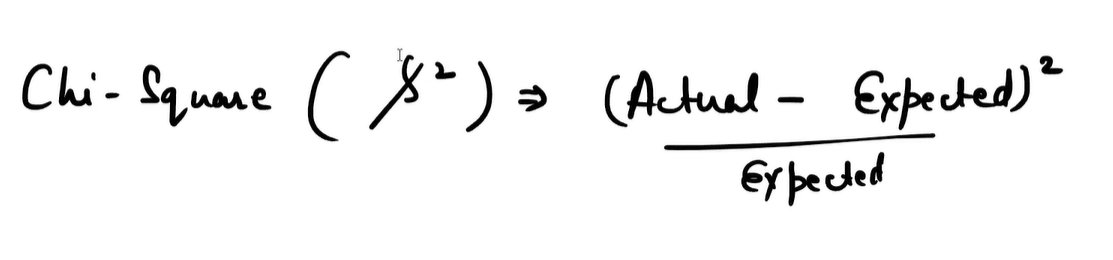
**✅ Next Steps in Chi-Square**

Repeat this calculation for all cells:

| **Gender** | **Pass (Expected)** | **Fail (Expected)** |
| --- | --- | --- |
| Male |  |  |
| Female |  |  |

Then use:

Where is observed and is expected.



**📘 Chi-Square Formula Explained**

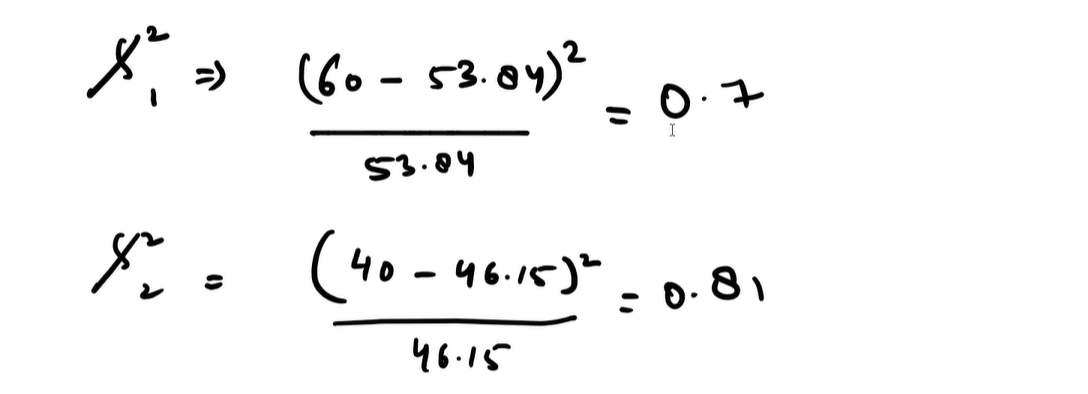
Where:

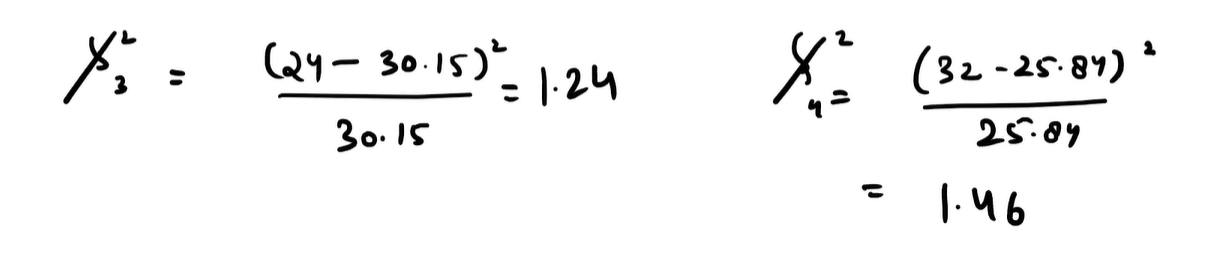
* = **Observed frequency** (actual count from data)
* = **Expected frequency** (calculated assuming independence)
* = Chi-Square statistic

This formula is applied to **each cell** in the contingency table, and the results are summed to get the total Chi-Square value.

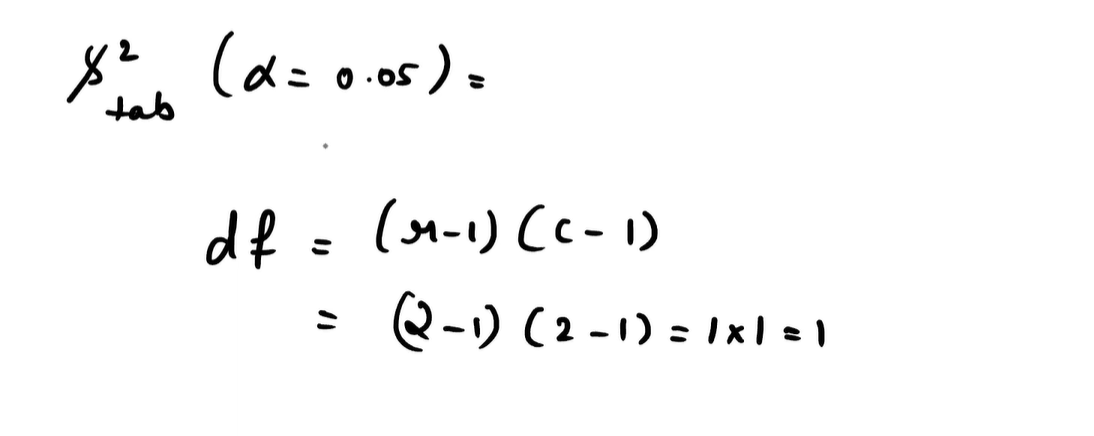
**🧠 What It Tells You**

* If the **observed values** are close to the **expected values**, the Chi-Square statistic will be **small** → likely **no relationship**.
* If the **observed values** differ significantly from the expected, the statistic will be **large** → possible **association** between variables.









You're now looking at the **final step of the Chi-Square Test of Independence** — calculating the total Chi-Square statistic by summing the individual contributions from each cell in the contingency table.

**🧮 Breakdown of the Formula**

For each cell in the table (e.g., Male-Pass, Male-Fail, Female-Pass, Female-Fail), the Chi-Square contribution is:

Where:

* = Observed frequency
* = Expected frequency

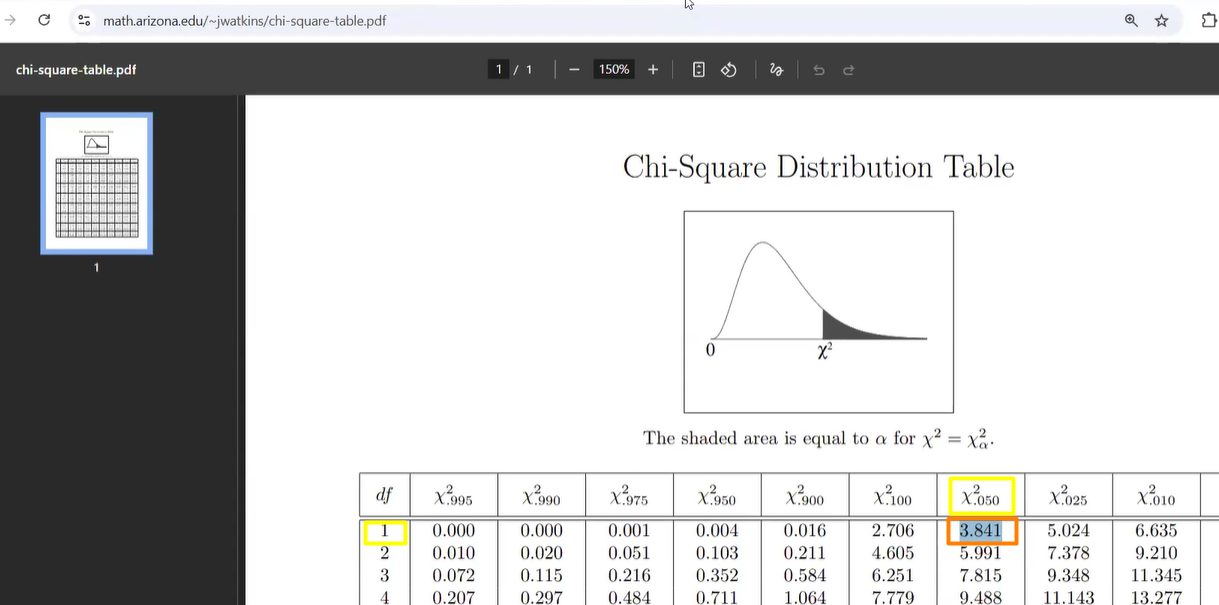
**📊 Your Calculations**

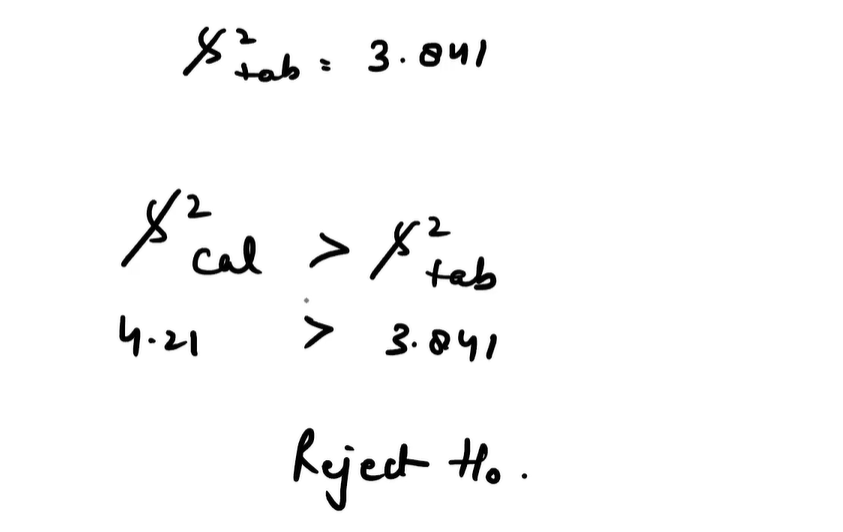
| **Cell** | **Observed (O)** | **Expected (E)** | **Contribution** |
| --- | --- | --- | --- |
| Male-Pass | 60 | 53.04 |  |
| Male-Fail | 40 | 46.15 |  |
| Female-Pass | 24 | 30.15 |  |
| Female-Fail | 32 | 25.84 |  |

**🔢 Final Chi-Square Value**

**🧠 What This Means**

* This total Chi-Square value (4.21) is compared against a **critical value** from the Chi-Square distribution table.
* Degrees of freedom (df) =
* At **df = 1**, the critical value at **α = 0.05** is **3.84**





**✅ Interpretation**

Since **4.21 > 3.84**, you **reject the null hypothesis**.

**Conclusion**: There is a statistically significant relationship between **Gender** and **Result** — they are **not independent**.

**🧑‍💻 Step-by-Step Python Code**

python

import pandas as pd

from scipy.stats import chi2\_contingency

# Step 1: Create raw data (Gender and Result for each person)

# In practice, this would come from a CSV or database

data = {

'Gender': ['Male']\*60 + ['Male']\*40 + ['Female']\*24 + ['Female']\*32,

'Result': ['Pass']\*60 + ['Fail']\*40 + ['Pass']\*24 + ['Fail']\*32

}

df = pd.DataFrame(data)

# Step 2: Build the contingency table using pd.crosstab

table = pd.crosstab(df['Gender'], df['Result'])

print("Contingency Table:\n", table)

# Step 3: Apply Chi-Square Test of Independence

chi2, p, dof, expected = chi2\_contingency(table)

# Step 4: Display results

print("\nChi-Square Statistic:", chi2)

print("Degrees of Freedom:", dof)

print("p-value:", p)

print("\nExpected Frequencies:\n", expected)

# Step 5: Interpretation

alpha = 0.05

if p < alpha:

print("\nConclusion: Reject Null Hypothesis → Gender and Result are related.")

else:

print("\nConclusion: Fail to Reject Null Hypothesis → No significant relationship between Gender and Result.")

**📊 What Happens Here**

1. **Raw Data** → Simulated with lists of "Male/Female" and "Pass/Fail".
2. pd.crosstab → Automatically builds the contingency table.

Code

Result Fail Pass

Gender

Female 32 24

Male 40 60

1. chi2\_contingency → Runs the Chi-Square test, giving:
   * **Chi-Square statistic** (≈ 3.58)
   * **Degrees of freedom** (1)
   * **p-value** (≈ 0.0584)
   * **Expected frequencies** (calculated under independence assumption)
2. **Interpretation** → Since p > 0.05, we fail to reject the null hypothesis.

**✅ Why This Is Useful**

* **Automates everything**: From raw survey-style data to statistical conclusion.
* **Scalable**: Works on larger datasets with multiple categories.
* **Flexible**: You can swap in any categorical variables (e.g., Age Group vs Product Preference).

**Degrees of freedom in chi-square tests**

Degrees of freedom describe how many values in your analysis are “free to vary” once certain constraints (like totals) are fixed. They determine which chi-square distribution you use to evaluate your test statistic and compute the p-value.

**Intuition with contingency tables**

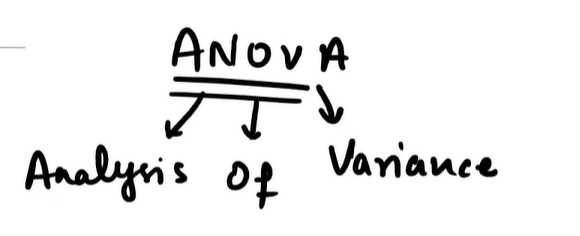
In an contingency table, once you fix all row totals and column totals, not every cell can vary independently—changing one cell forces others to adjust to keep the margins the same. The number of truly free cells is:

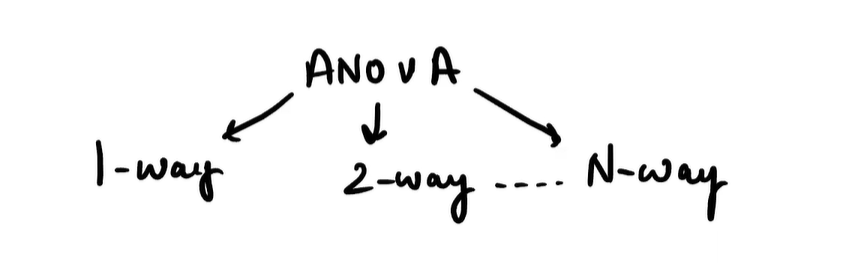
* **Why ?**
  + You can freely choose values in rows and columns.
  + The last row and last column are determined by the margins (they must make totals add up), so they’re not free.

**Example using your table (Gender × Result)**

Your table is (Male/Female × Pass/Fail).

* If you set three cells (say, Male–Pass, Male–Fail, Female–Pass) to any values consistent with totals, the fourth (Female–Fail) is forced by the row and column sums. That single “free choice” is why df = 1.





**ANOVA (Analysis of Variance) is a statistical method used to compare the means of multiple groups to determine if at least one group differs significantly. It's widely used in fields like marketing, healthcare, education, and manufacturing to analyze the impact of different factors.**

**📘 What Is ANOVA?**

* **ANOVA** stands for **Analysis of Variance**.
* It tests whether the **means of three or more groups** are significantly different.
* It works by comparing:
  + **Between-group variance**: How much group means differ.
  + **Within-group variance**: How much individuals within each group vary.

If between-group variance is large relative to within-group variance, the group means are likely different.

**🔀 Types of ANOVA**

| **Type** | **Description** | **Example** |
| --- | --- | --- |
| **One-Way ANOVA** | Tests effect of one factor | Does teaching method affect student scores? |
| **Two-Way ANOVA** | Tests two factors and their interaction | Does diet and exercise type affect weight loss? |
| **N-Way ANOVA** | Tests multiple factors | Does location, ad type, and time of day affect sales? |

**🌍 Real-World Applications**

**1. Marketing Campaign Analysis**

* **Scenario**: A company runs 3 ad campaigns (TV, Social Media, Email).
* **Goal**: Compare average sales generated by each campaign.
* **Method**: One-Way ANOVA tests if sales differ significantly across campaigns.
* **Outcome**: Helps allocate budget to the most effective channel.

**2. Healthcare Trials**

* **Scenario**: A pharmaceutical firm tests 4 drugs on blood pressure.
* **Goal**: Determine which drug lowers BP most effectively.
* **Method**: One-Way ANOVA compares mean BP reduction across groups.
* **Outcome**: Identifies the most promising treatment.

**3. 3-Way ANOVA Tests:**

* **Main effects: Does diet, exercise, or gender affect weight loss?**
* **Two-way interactions: Diet × Exercise, Diet × Gender, Exercise × Gender**
* **Three-way interaction: Diet × Exercise × Gender**

**Application: Helps personalize fitness plans based on combined effects.**

**📊 Summary Table**

| **Factor Count** | **Name** | **Use Case Example** | **Complexity** |
| --- | --- | --- | --- |
| 1 | One-Way | Teaching method vs test scores | Low |
| 2 | Two-Way | Diet & exercise vs weight loss | Moderate |
| 3 | Three-Way | Diet × Exercise × Gender vs weight loss | High |
| 4+ | N-Way | Machine × Operator × Material × Shift | Very High |



**📘 What Is the F-Statistic?**

* The **F-statistic** is a ratio used to compare **variances between groups** to **variances within groups**.
* It helps determine whether the **group means are significantly different**.
* Named after **Sir Ronald Fisher**, who developed the method.

**🧠 Why It Matters**

* If the **between-group variance** is much larger than the **within-group variance**, it suggests that the groups differ significantly.
* If the ratio is close to 1, it means the group differences are likely due to random variation.

**🧪 Real-World Example: One-Way ANOVA**

**Scenario:**

A teacher wants to compare test scores across three teaching methods:

* Lecture
* Online
* Flipped classroom

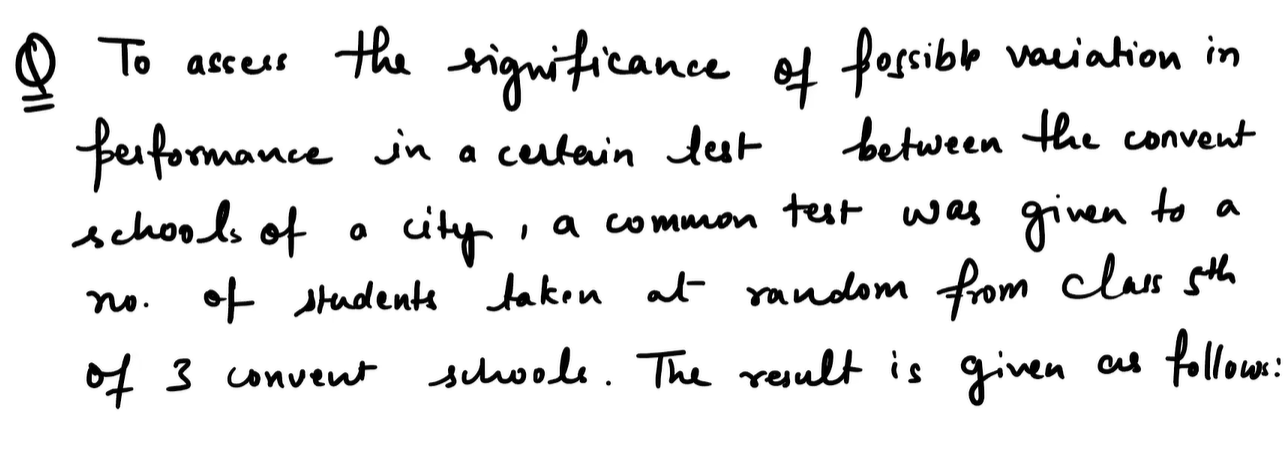
**Steps:**

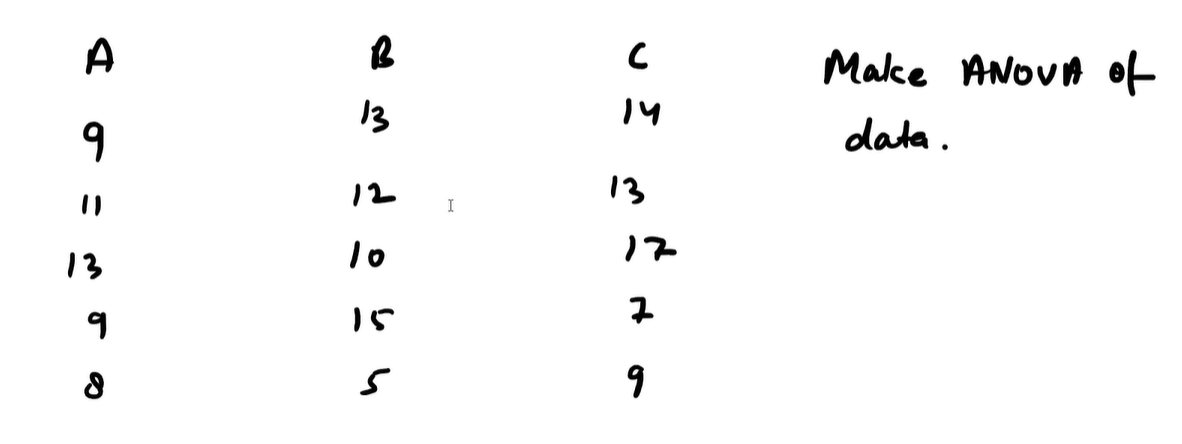
1. Calculate **mean score** for each group.
2. Compute **variance between groups** (how much group means differ).
3. Compute **variance within groups** (how much students vary within each method).
4. Use the F-statistic to test if the differences are statistically significant.

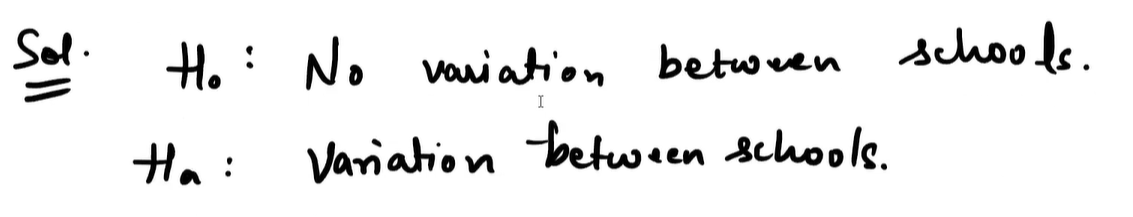
If the **F-value is large**, and the **p-value < 0.05**, the teacher concludes that **teaching method affects performance**.

**🔬 Where Else Is F-Statistic Used?**

* **Regression analysis**: To test if the overall model is significant.
* **Two-way and N-way ANOVA**: To test multiple factors and their interactions.
* **Experimental design**: To validate treatment effects across conditions.







This is a classic **One-Way ANOVA** scenario where you're testing whether the **mean test scores** of students from three different convent schools (A, B, and C) are significantly different.

**📘 Problem Setup**

* **Goal**: Assess if there's a significant variation in performance across schools.
* **Groups**:
  + School A: [9, 11, 13, 9, 8]
  + School B: [13, 12, 10, 15, 5]
  + School C: [14, 13, 17, 7, 9]
* **Assumptions**:
  + Students are randomly selected.
  + Scores are continuous and approximately normally distributed.
  + Variances across groups are roughly equal.

**🧪 What Is One-Way ANOVA?**

**One-Way ANOVA** compares the **means of three or more independent groups** to see if at least one group differs significantly.

**Formula for F-statistic:**

* **Between-group variance**: How much group means differ from the overall mean.
* **Within-group variance**: How much scores vary within each group.

**🔢 Python Code to Run This**

python

import pandas as pd

from scipy.stats import f\_oneway

# Step 1: Define scores for each school

school\_A = [9, 11, 13, 9, 8]

school\_B = [13, 12, 10, 15, 5]

school\_C = [14, 13, 17, 7, 9]

# Step 2: Run One-Way ANOVA

f\_stat, p\_value = f\_oneway(school\_A, school\_B, school\_C)

# Step 3: Output results

print("F-statistic:", f\_stat)

print("p-value:", p\_value)

# Step 4: Interpretation

alpha = 0.05

if p\_value < alpha:

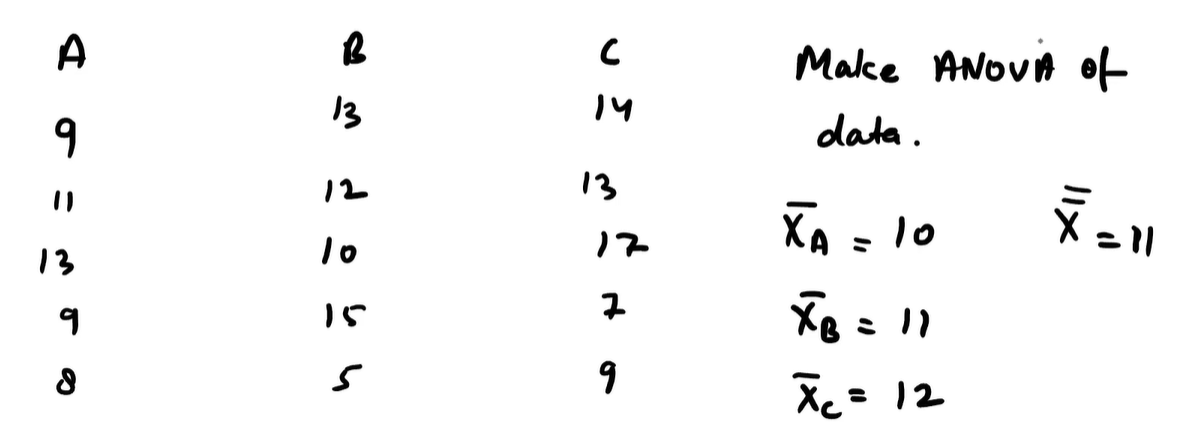
print("Conclusion: Significant difference in performance between schools.")

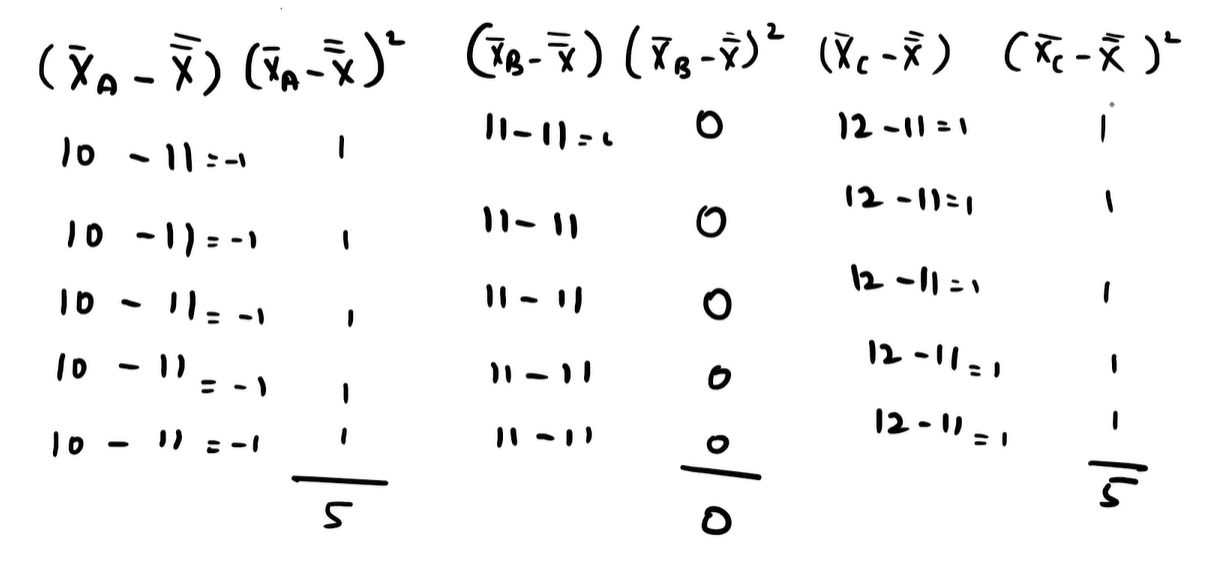
else:

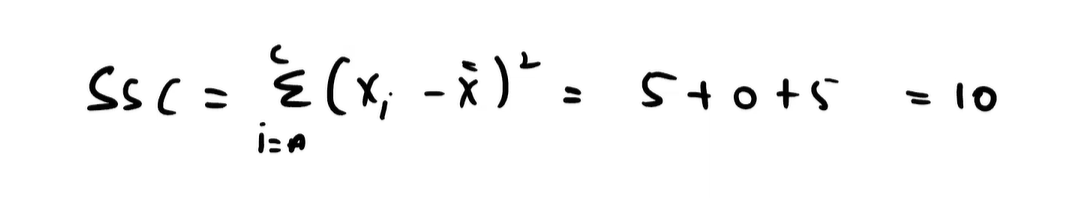
print("Conclusion: No significant difference in performance between schools.")

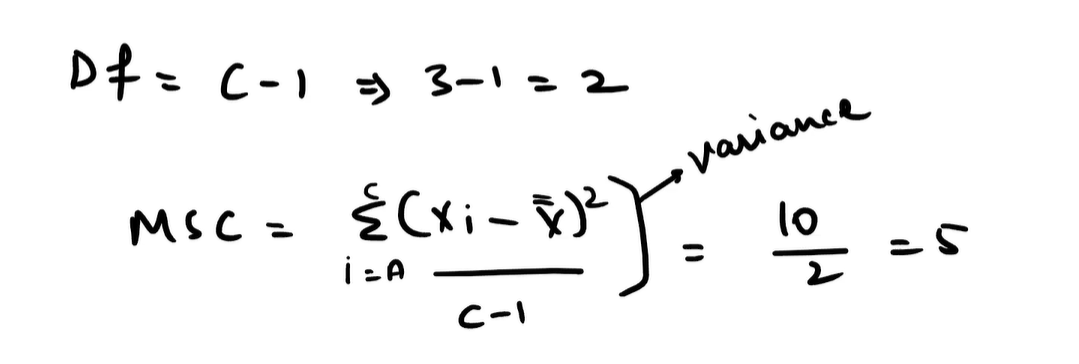
**🧠 Interpretation**

* **Null Hypothesis (H₀)**: All schools have the same mean score.
* **Alternative Hypothesis (H₁)**: At least one school has a different mean.
* If **p-value < 0.05**, reject H₀ → performance varies significantly across schools.









This image shows the **ANOVA calculation for Mean Square Between Groups (MSC)** and **Degrees of Freedom (df)** for comparing three groups (A, B, and C). Let’s break it down step by step:

**📘 What’s Being Calculated?**

**1. Degrees of Freedom (df) Between Groups**

Where:

* = number of groups (here, A, B, C → 3 groups)

So:

This tells us how many group means are free to vary once the overall mean is fixed.

**2. Mean Square Between Groups (MSC)**

This is the **variance between group means**, calculated as:

But in your image, it simplifies to:

Where:

* **SSC** = Sum of Squares Between Groups = 10 (from previous image)
* **df** = 2

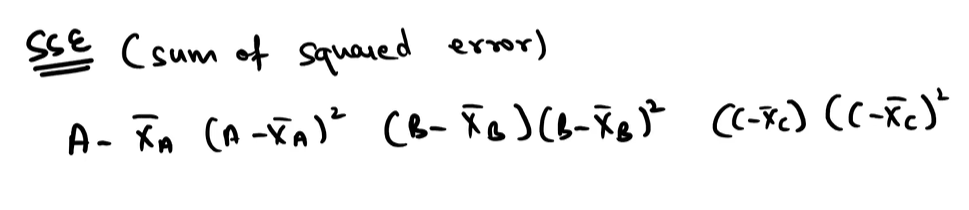
So the **mean square between groups** is 5.

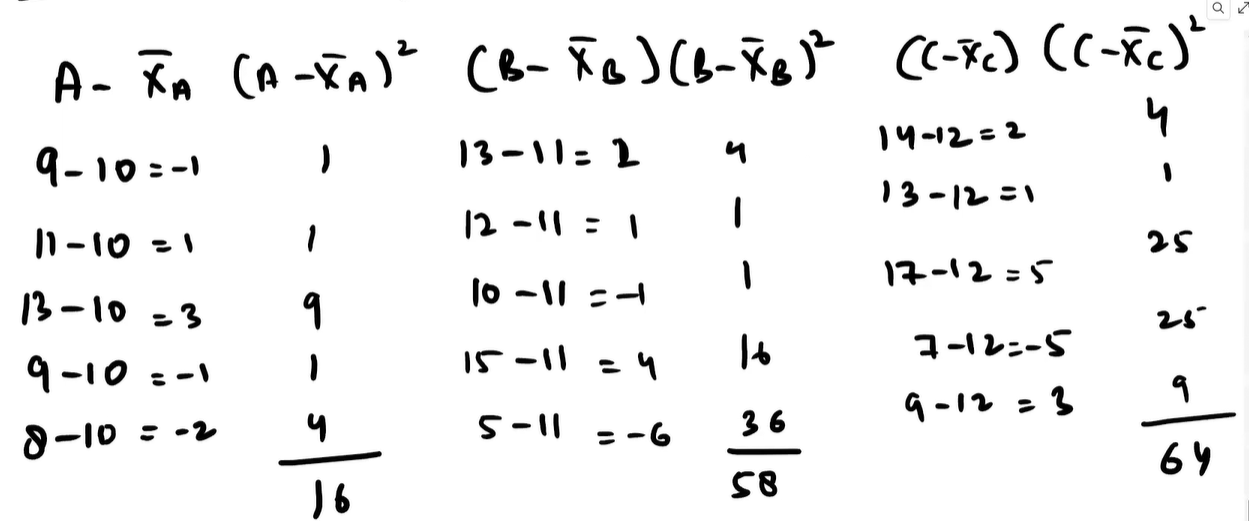
**🧠 Why This Matters in ANOVA**

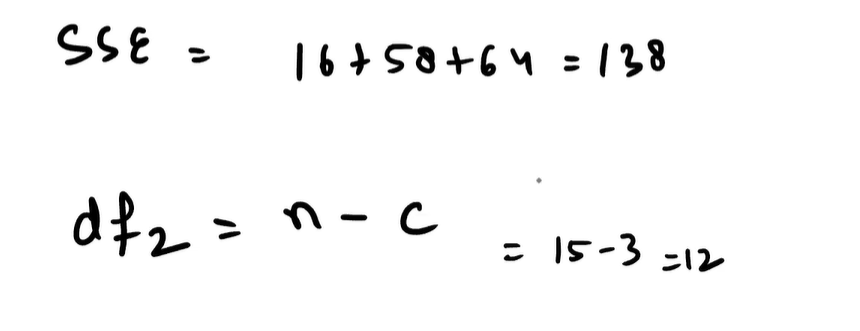
* **MSC** measures how much the group means differ from the overall mean.
* It’s compared to **MSW** (Mean Square Within Groups) to compute the **F-statistic**:
* A high F-value suggests that group means differ significantly.

**✅ Summary**

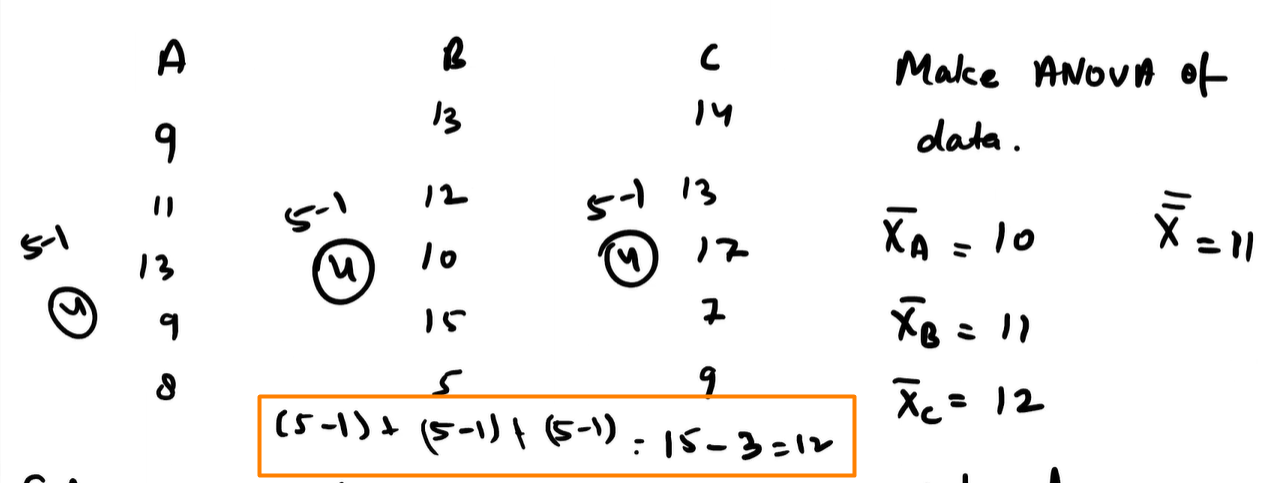
* You’ve calculated **df = 2** for 3 groups.
* You’ve found **MSC = 5**, which is the average squared deviation of group means from the overall mean.
* Next step: calculate **MSW (Mean Square Within)** and then compute the **F-statistic** to complete the ANOVA test.

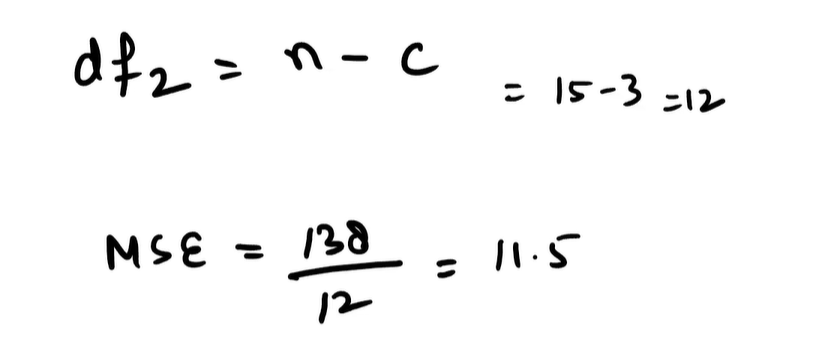






Degree of Freedom for columns





This image illustrates the concept of **Sum of Squared Errors (SSE)** in the context of **ANOVA**, specifically showing how to compute the **within-group variation** for each group (A, B, C).

**📘 What Is SSE?**

**SSE (Sum of Squared Errors)** measures how much individual scores deviate from their **group mean**. It’s also called **SSW (Sum of Squares Within)** in ANOVA.

Where:

* = individual score
* = mean of the group that score belongs to

**🧪 Breakdown by Group**

Let’s compute SSE for each group using your data:

**Group A: [9, 11, 13, 9, 8], Mean = 10**

**Group B: [13, 12, 10, 15, 5], Mean = 11**

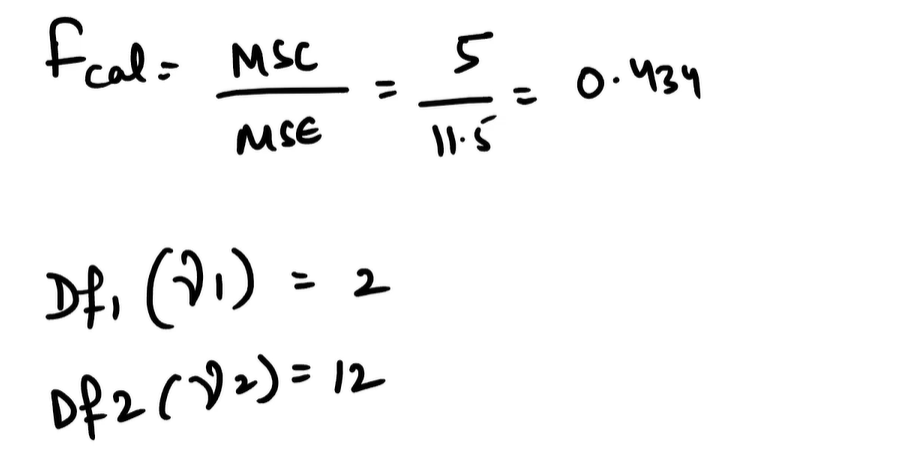
**Group C: [14, 13, 17, 7, 9], Mean = 12**

**🔢 Total SSE (SSW)**

This value is used to compute **Mean Square Within (MSW)**:

**🧠 Why It Matters**

* SSE captures **random variation** within each group.
* It’s compared to **MSC (Mean Square Between)** to compute the **F-statistic**:
* A high F-value suggests group means differ significantly.



This image shows the final step of your **ANOVA (Analysis of Variance)** calculation: computing the **F-statistic** and identifying the **degrees of freedom** for the test.

**📘 What’s Being Calculated?**

**🔢 F-Statistic Formula**

Where:

* **MSC** = Mean Square Between Groups (variation due to group differences)
* **MSE** = Mean Square Error (variation within groups)
* **F\_cal** = Calculated F-statistic

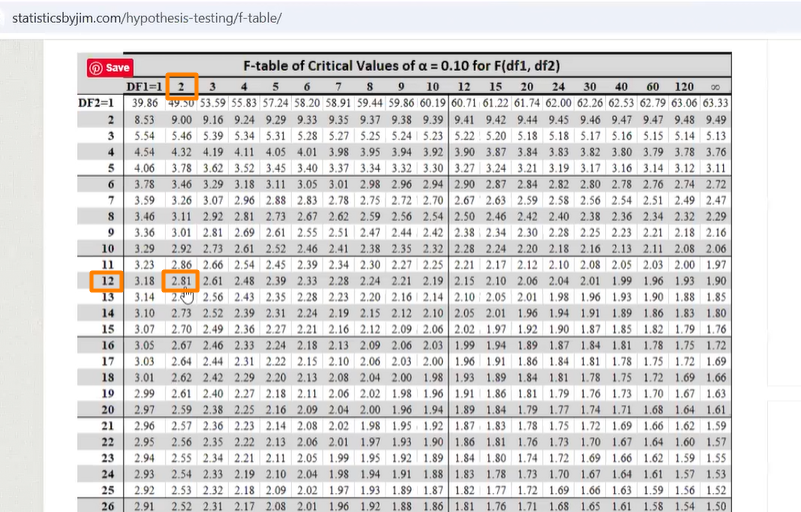
This ratio tells us whether the variation between group means is large relative to the variation within groups.

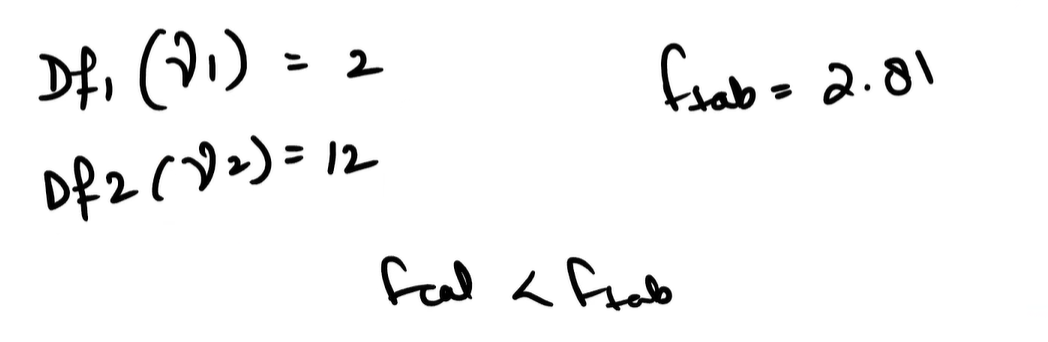
**🎯 Degrees of Freedom**

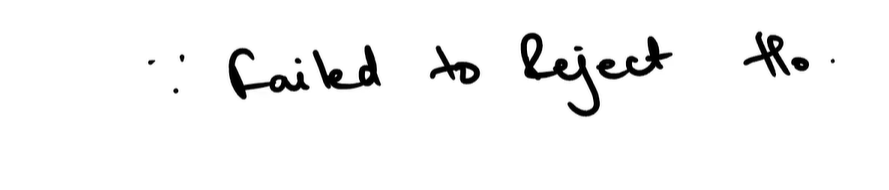
* **Df₁ (γ₁)** = **2** → for the numerator (between groups):
* **Df₂ (γ₂)** = **12** → for the denominator (within groups):

Where:

* = number of groups (A, B, C)
* = total number of observations (5 per group × 3 groups = 15)







This image shows the **final decision step in an ANOVA hypothesis test** using the **F-distribution table**. Let’s break it down clearly:

**📘 What’s Happening Here?**

You’ve calculated the **F-statistic** from your ANOVA analysis:

Now you're comparing this to the **critical value** from the **F-table** at a significance level of **α = 0.10**.

**🎯 Degrees of Freedom**

* **Numerator df (Df₁ or γ₁)** = 2 → from number of groups:
* **Denominator df (Df₂ or γ₂)** = 12 → from total observations:

**📈 Critical Value from F-table**

From the F-table for α = 0.10, df₁ = 2 and df₂ = 12:

**🧠 Decision Rule**

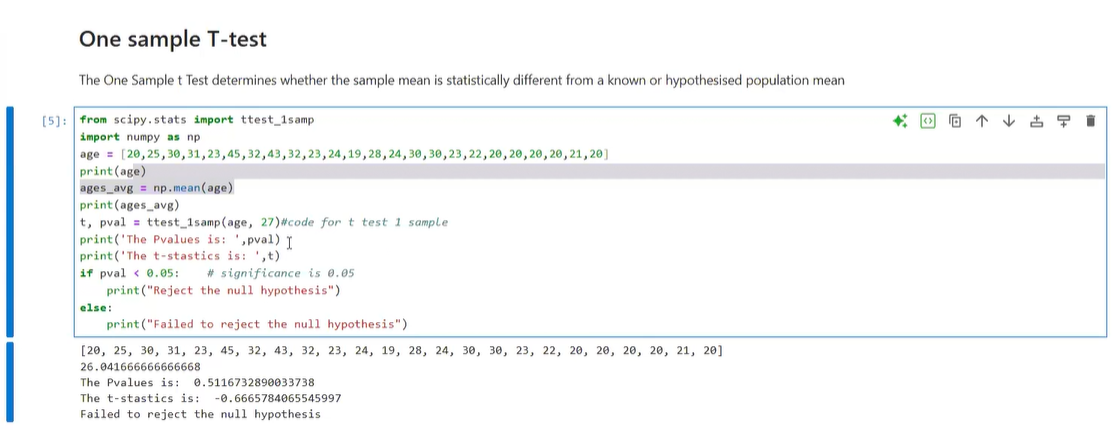
* If , **reject H₀**
* If , **fail to reject H₀**

Here:

So you **fail to reject the null hypothesis**.

**✅ Final Conclusion**

There is **no statistically significant difference** in test scores between the three schools at the 10% significance level.



This image shows a **one-sample t-test** performed in Python using scipy.stats.ttest\_1samp. The goal is to determine whether the **average age** in a sample differs significantly from a hypothesized population mean of **27 years**.

**🧪 Step-by-Step Breakdown**

**1. Sample Data**

python

age = [20, 25, 30, 31, 23, 45, 32, 43, 32, 23, 24, 19, 28, 24, 30, 30, 23, 22, 20, 20, 20, 20, 21, 20]

* 24 individuals
* Ages range from 19 to 45

**2. Sample Mean**

python

ages\_avg = np.mean(age)

* Result: **26.04**

**3. Hypothesis Test**

python

ttest\_1samp(age, 27)

* **Null Hypothesis (H₀)**: The true mean age is 27
* **Alternative Hypothesis (H₁)**: The true mean age is not 27

**📈 Output**

| **Metric** | **Value** |
| --- | --- |
| Sample Mean | 26.04 |
| t-statistic | -0.667 |
| p-value | 0.5117 |
| α (significance level) | 0.05 |
| Decision | **Fail to reject H₀** |

**🧠 Interpretation**

* The **p-value (0.5117)** is much greater than 0.05.
* This means the observed mean (26.04) is **not significantly different** from the hypothesized mean (27).
* So, we **do not have enough evidence** to say the average age is different from 27.

**What is meaning of 1 sample t-test here**

**A one-sample t-test is a statistical test used to determine whether the mean of a single sample is significantly different from a known or hypothesized population mean.**

📘 Definition

It answers the question:

“Is my sample average meaningfully different from a specific value I expect?”

🧪 Real-World Example

Let’s say you believe the average age of college students is 27 years. You collect a sample of 24 students and find their average age is 26.04.

You run a one-sample t-test to check:

* Null Hypothesis (H₀): The true mean age is 27
* Alternative Hypothesis (H₁): The true mean age is not 27

If the difference between 26.04 and 27 is statistically significant, you reject H₀.

🔢 Formula

Where:

* = sample mean
* = population mean (hypothesized)
* = sample standard deviation
* = sample size

🧠 Interpretation

* If p-value < α (e.g., 0.05), you reject H₀ → sample mean is significantly different.
* If p-value ≥ α, you fail to reject H₀ → no significant difference.

✅ When to Use It

* You have one group.
* You want to compare its mean to a specific value.
* Data is approximately normally distributed.

**How is ANOVA & T-test related :**

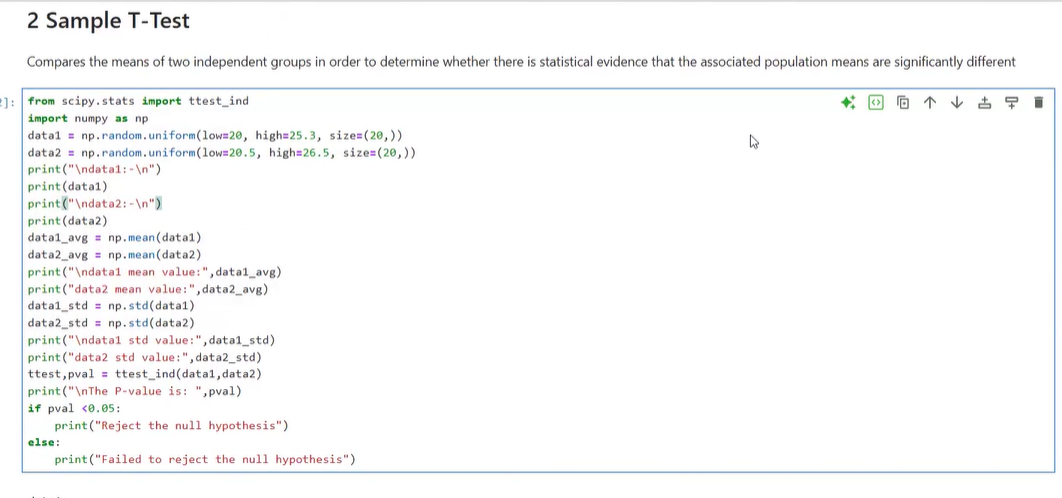
**the one-sample t-test and ANOVA are closely related, but they’re used in slightly different situations.**

**📘 Relationship Between One-Sample t-test and ANOVA**

* One-sample t-test: Compares the mean of a single sample against a known or hypothesized population mean. Example: Is the average age of students in your sample different from 27?
* ANOVA (Analysis of Variance): Compares the means of two or more groups simultaneously to see if at least one group differs. Example: Do students from School A, B, and C have different average test scores?

**🔗 How They Connect**

* **Mathematically, ANOVA is a generalization of the t-test:**
  + **A two-sample t-test (comparing two groups) is equivalent to a one-way ANOVA with two groups.**
  + **A one-sample t-test can be thought of as a special case where you’re comparing one group mean against a fixed value (instead of another group).**
* So, yes — the one-sample t-test is related to ANOVA because both rely on comparing variances to test hypotheses about means.
* 🧠 Intuition
* t-test: Focuses on differences in means using the t-distribution.
* ANOVA: Focuses on differences in means using the F-distribution.
* Both are variance-based methods, but ANOVA scales up to handle multiple groups at once.



**Two-sample t-test overview**

A two-sample t-test compares the means of two independent groups to determine whether the population means are significantly different. Your notebook generates two uniform random samples (data1 and data2), computes their means and standard deviations, and runs scipy.stats.ttest\_ind to obtain a t-statistic and p-value. Because the data are randomly generated, the exact numbers will vary each run—use a fixed random seed to make results reproducible.

**What the code is doing**

* **Generate data:**
  + data1 ~ Uniform(20, 25.3), n = 20
  + data2 ~ Uniform(20.5, 26.5), n = 20
* **Summaries:**
  + Prints sample means and standard deviations for both groups.
* **Hypotheses:**
  + H₀: The two population means are equal.
  + H₁: The two population means are different.
* **Test:**
  + ttest\_ind(data1, data2) computes the t-statistic and p-value (Welch’s test if equal\_var=False).

**How to interpret the output**

* **t-statistic:** Measures how many standard errors apart the sample means are. Larger absolute values suggest stronger evidence against H₀.
* **p-value:** Probability of seeing a difference at least as extreme as observed if H₀ were true.
  + If p < 0.05: Reject H₀ → the means are significantly different.
  + If p ≥ 0.05: Fail to reject H₀ → not enough evidence of a difference.

Because data1 is sampled from a slightly lower range than data2, you’ll often—but not always—see p < 0.05 indicating a difference. Randomness and sample size can push the result either way.



This is a **two-sample t-test** comparing two independent datasets (data1 and data2) to determine whether their **means are significantly different**.

**📊 Summary of the Data**

| **Metric** | **data1** | **data2** |
| --- | --- | --- |
| Mean | 23.22 | 23.66 |
| Std Dev | 1.39 | 1.68 |
| Sample Size | 20 | 20 |

**🧪 Hypothesis Test**

* **Null Hypothesis (H₀)**: The population means of data1 and data2 are equal.
* **Alternative Hypothesis (H₁)**: The population means are different.

**🔢 Test Result**

* **P-value**: 0.3877
* **Significance Level (α)**: 0.05
* **Decision**: Since **p > 0.05**, we **fail to reject the null hypothesis**.

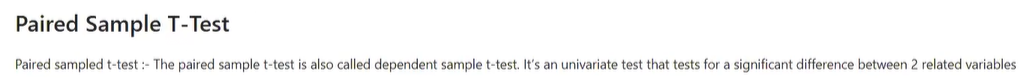
**🧠 Interpretation**

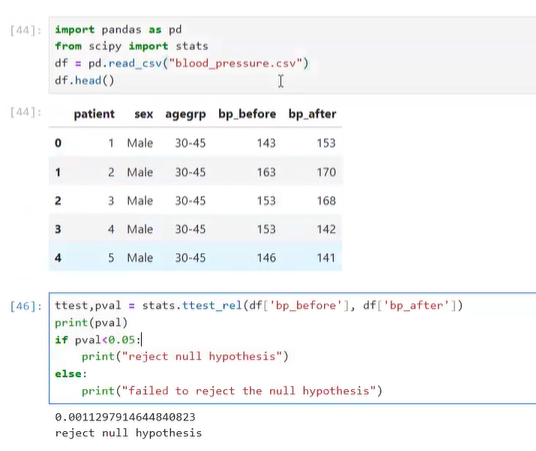
* The difference in means (23.22 vs 23.66) is **not statistically significant**.
* This suggests that the observed difference could be due to **random variation**, not a true difference in population means.

**✅ Conclusion**

There is **no strong evidence** to claim that data1 and data2 come from populations with different means.

**Paired Sample T-Test:**





This is a **paired sample t-test** applied to medical data to evaluate whether a treatment significantly changed patients' blood pressure. Let’s walk through it clearly:

**🧪 What Is a Paired Sample t-Test?**

It compares **two related measurements** from the **same subjects**—in this case, **blood pressure before and after treatment**.

**🔍 Why “paired”?**

Because each patient has:

* **bp\_before**: Blood pressure before treatment
* **bp\_after**: Blood pressure after treatment

We’re testing whether the **mean difference** between these paired values is statistically significant.

**📊 Data Snapshot**

| **Patient** | **bp\_before** | **bp\_after** |
| --- | --- | --- |
| 1 | 143 | 153 |
| 2 | 163 | 170 |
| 3 | 153 | 168 |
| 4 | 132 | 142 |
| 5 | 146 | 141 |

... and so on

**🧠 Hypothesis**

* **Null Hypothesis (H₀)**: The treatment has **no effect** → mean difference = 0
* **Alternative Hypothesis (H₁)**: The treatment **does affect** blood pressure → mean difference ≠ 0

**🔢 Result**

python

ttest, pval = stats.ttest\_rel(df['bp\_before'], df['bp\_after'])

* **p-value** = 0.00113
* Since **p < 0.05**, we **reject the null hypothesis**

**✅ Conclusion:**

There is a **statistically significant difference** in blood pressure before and after treatment. The treatment likely had an effect.

**What is Paired Sample T-Test :**

A **paired sample t-test** (also called a **dependent t-test**) is a statistical method used to determine whether the **mean difference between two sets of related observations** is significantly different from zero.

**🧪 When to Use a Paired Sample T-Test**

Use it when:

* You have **two measurements from the same subjects** (e.g., before and after treatment).
* You want to test whether the **change** is statistically significant.

**🔍 Examples:**

* Blood pressure **before and after** medication.
* Student scores **before and after** tutoring.
* Productivity **before and after** a new tool is introduced.

**📘 Hypotheses**

* **Null Hypothesis (H₀)**: The mean difference between paired observations is **zero**.
* **Alternative Hypothesis (H₁)**: The mean difference is **not zero**.

**🔢 Formula**

Where:

* = mean of the differences
* = standard deviation of the differences
* = number of pairs

**🧠 Interpretation**

* If **p-value < α** (e.g., 0.05), you **reject H₀** → the change is statistically significant.
* If **p-value ≥ α**, you **fail to reject H₀** → no significant change detected.

**✅ Assumptions**

* Data are **paired and dependent**.
* Differences are **approximately normally distributed**.
* Measurement scale is **interval or ratio**.

**What is T-test & What is Z-test, when to apply them, give me a practical example in detail**

**Quick Answer:** A **t-test** is used when the sample size is small or the population variance is unknown, while a **z-test** is used when the sample size is large and the population variance is known. Both are hypothesis tests comparing means, but they differ in assumptions and applicability.

**📘 T-test vs Z-test: Core Differences**

| **Feature** | **T-test** | **Z-test** |
| --- | --- | --- |
| **Population variance** | Unknown (estimated from sample) | Known |
| **Sample size** | Small (typically n < 30) | Large (typically n ≥ 30) |
| **Distribution used** | Student’s t-distribution | Standard normal (z) distribution |
| **Applications** | Comparing sample mean(s) when variance unknown | Comparing sample mean(s) when variance known |
| **Types** | One-sample, two-sample, paired | One-sample, two-sample, proportion tests |

Sources:

**🧪 When to Apply**

* **Use a T-test when:**
  + You don’t know the population variance.
  + You have a small sample size.
  + Example: Testing if the average exam score of 20 students differs from 50.
* **Use a Z-test when:**
  + You know the population variance (or have a very large sample so the sample variance approximates it).
  + You have a large sample size.
  + Example: Testing if the average height of 500 people differs from the national average of 170 cm (with known variance).

**🔍 Practical Example in Detail**

**Scenario: New Teaching Method**

A school claims that a new teaching method increases average test scores above the national average of **75**.

**Case 1: Small Sample (T-test)**

* Sample: 20 students
* Sample mean = 78
* Sample standard deviation = 5
* Population variance unknown

**Hypotheses:**

* H₀: μ = 75 (no improvement)
* H₁: μ > 75 (improvement)

**Test:**

Compare with critical t-value (df = 19). If p < 0.05, reject H₀ → teaching method works.

**Case 2: Large Sample (Z-test)**

* Sample: 200 students
* Sample mean = 76
* Population variance known = 16

**Hypotheses:**

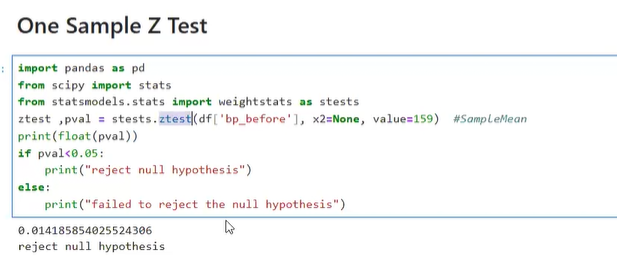
* H₀: μ = 75
* H₁: μ > 75

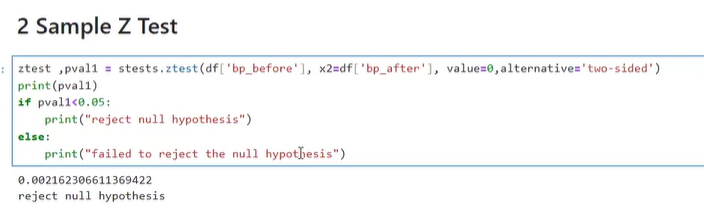
**Test:**

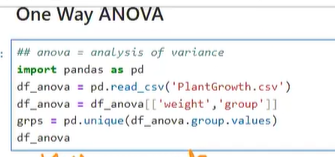
Compare with critical z-value (≈1.645 for one-tailed, α=0.05). Since 3.53 > 1.645, reject H₀ → teaching method works.

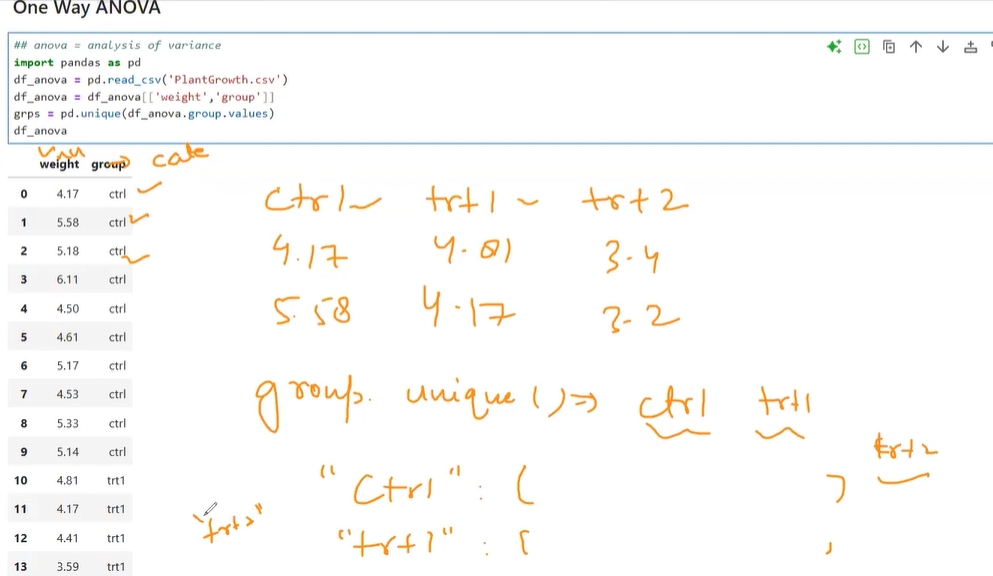
**🧠 Key Takeaway**

* **T-test**: Use when variance unknown or sample small.
* **Z-test**: Use when variance known or sample large.
* Both help determine if observed differences are statistically significant.









this image shows a **One-Way ANOVA setup in Python using the PlantGrowth dataset**, where you're comparing the **mean plant weights** across three treatment groups: ctrl, trt1, and trt2.

**🧪 What the Code Is Doing**

python

import pandas as pd

df\_anova = pd.read\_csv('PlantGrowth.csv')

df\_anova = df\_anova[['weight','group']]

grps = pd.unique(df\_anova.group.values)

**🔍 Breakdown:**

* **Reads the dataset** PlantGrowth.csv
* Selects only the relevant columns: weight (numeric) and group (categorical)
* Extracts **unique group labels**: ctrl, trt1, trt2

**📊 Data Structure**

| **weight** | **group** |
| --- | --- |
| 4.17 | ctrl |
| 5.58 | ctrl |
| 5.18 | ctrl |
| 6.11 | ctrl |
| 4.50 | ctrl |
| ... | ... |
| 4.17 | trt1 |
| 4.41 | trt1 |
| 3.59 | trt1 |

You’re preparing to test whether **mean plant weight differs significantly** across the three groups.

**🧠 What One-Way ANOVA Will Do**

* **Null Hypothesis (H₀)**: All groups have the same mean weight.
* **Alternative Hypothesis (H₁)**: At least one group differs.

**Steps:**

1. Calculate **group means** and **overall mean**.
2. Compute **between-group variance** (MSC) and **within-group variance** (MSW).
3. Calculate the **F-statistic**:
4. Compare with critical F-value or use p-value to decide.

**✅ Next Step**

You’re ready to run:

python

from scipy.stats import f\_oneway

ctrl = df\_anova[df\_anova.group == 'ctrl']['weight']

trt1 = df\_anova[df\_anova.group == 'trt1']['weight']

trt2 = df\_anova[df\_anova.group == 'trt2']['weight']

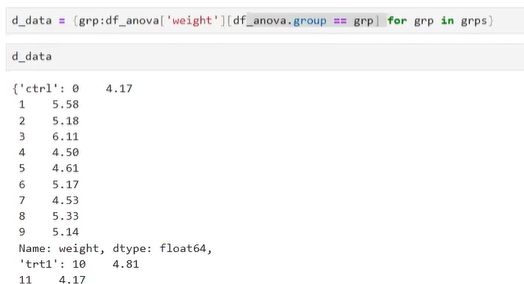
f\_stat, p\_val = f\_oneway(ctrl, trt1, trt2)

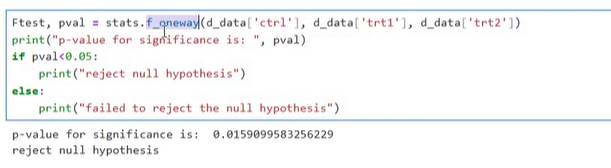
print("F-statistic:", f\_stat)

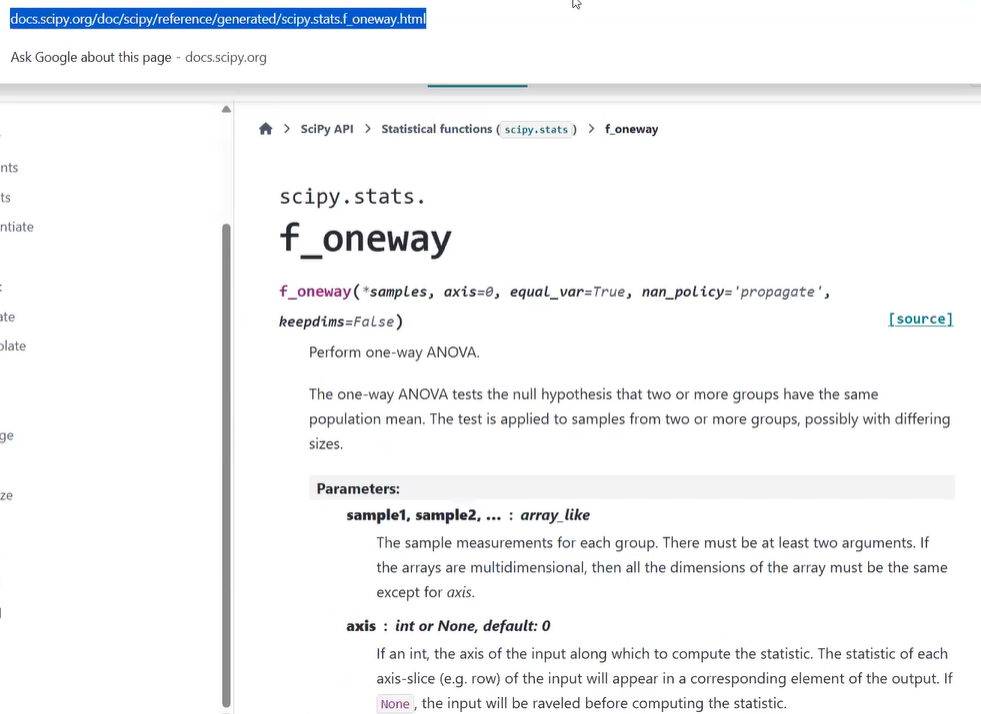
print("p-value:", p\_val)

If **p < 0.05**, you reject H₀ → at least one treatment affects plant growth significantly.

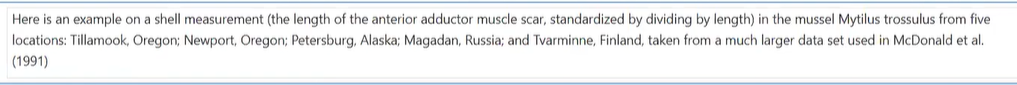
Dictionary Comprehentions:

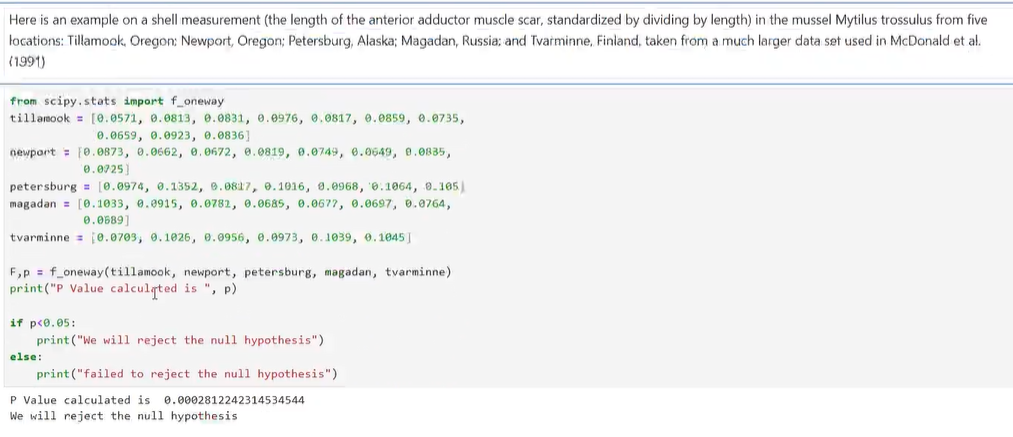




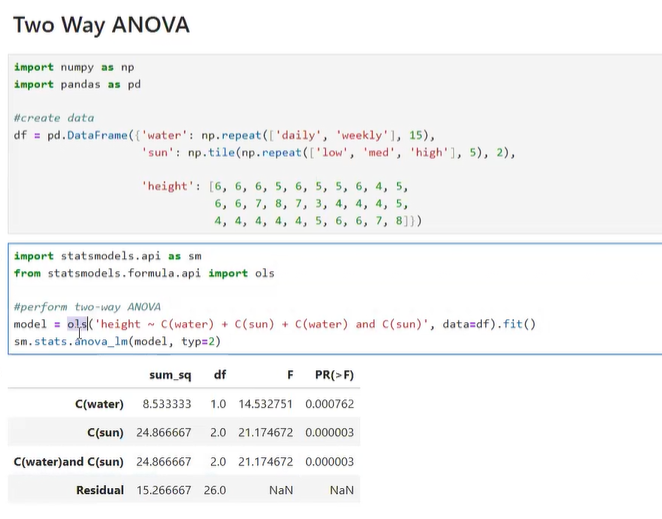


Another Example:





Two-Way ANOVA:



This is a **Two-Way ANOVA** analysis using Python’s statsmodels library. It evaluates how **two categorical factors**—water and sun—affect a continuous outcome: **plant height**, and whether there's an **interaction effect** between them.

**📘 What Is Two-Way ANOVA?**

Two-Way ANOVA tests:

1. **Main effect of Factor A** (e.g., water schedule)
2. **Main effect of Factor B** (e.g., sunlight level)
3. **Interaction effect** (does the effect of water depend on sunlight?)

**🧪 Dataset Overview**

* **Factors**:
  + water: daily vs weekly
  + sun: low, medium, high
* **Response**:
  + height: numeric plant height

**Sample Data Structure**

| **water** | **sun** | **height** |
| --- | --- | --- |
| daily | low | 6 |
| daily | med | 5 |
| weekly | high | 8 |
| ... | ... | ... |

**🔢 ANOVA Output Summary**

| **Source** | **Sum of Squares** | **df** | **F-value** | **p-value** |
| --- | --- | --- | --- | --- |
| **Water** | 8.53 | 1 | 14.53 | 0.00076 |
| **Sun** | 24.87 | 2 | 21.17 | 0.000003 |
| **Water × Sun** | 24.87 | 2 | 21.17 | 0.000003 |
| **Residual (Error)** | 15.27 | 26 | — | — |

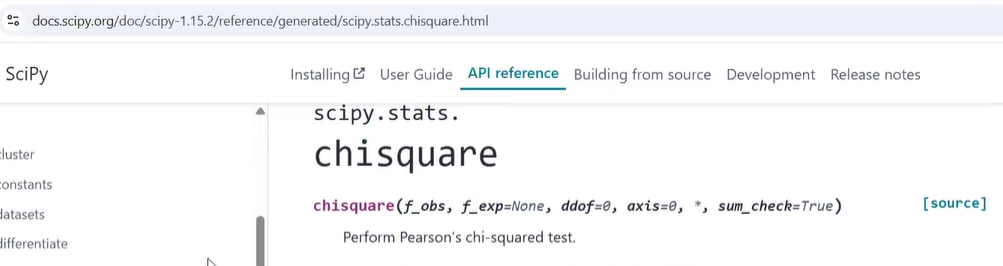
**🧠 Interpretation**

* **Water** has a significant effect on plant height (p < 0.001).
* **Sunlight** also has a significant effect (p < 0.001).
* **Interaction** between water and sunlight is significant → the effect of watering depends on sunlight level.

**✅ Conclusion:**

All three effects—**water**, **sun**, and their **interaction**—significantly influence plant height.

Read the document:



EDA – Python Code:

Explore the attached Exploratory Data Analysis Jupyter Notebook in the Folder.

