**A picture containing computer, computer, clock, television

Description automatically generated**

School of Computer Science and Engineering

B. Tech CSE-AIML D

**COMPUTATIONAL MATHEMATICS**

**(23TBSMA42)**

**EXPERIENTIAL LEARNING ON**



Submitted to

**Dr. Vishal Patil**

Assistant Professor,

Department of Mathematics,

Faculty of Engineering & Technology,

Jain (Deemed-To-Be) University.

Submitted by

| USN | Name | Individual Contribution |
| --- | --- | --- |
| 23BTRCL089 | Thanish Chinnappa K.C. |  |
| 23BTRCL257 | Likhith |  |
| 23BTRCL007 | Sahil Vinod Patil |  |
| 23BTRCL075 | Samith |  |
| 23BTRCL082 | Sundareshwar S |  |
| 23BTRCL163 | Souharda Mandal |  |

| **Branch & Section:** | CSE - AIML D |
| --- | --- |
| **Date of Submission:** | 5th May 2025 |

| **Total Marks** |  |
| --- | --- |
| **Remarks** |  |

**INDEX**

| **Sl.no.** | **Title** | **Page No.** |
| --- | --- | --- |
| **1** | **Introduction** | **3** |
| **2** | **Environment Setup and Dataset Loading** | **4-6** |
| **3** | **Sample Size Determination & Population Correction** | **7-8** |
| **4** | **Probability Sampling & Sample Selection** | **9-13** |
| **5** | **Descriptive Statistics: Mean and Standard Deviation** |  |
| **6** | **Sampling Distribution & Visualization** | **13-14** |
| **7** | **Hypothesis Formulation and Testing (5% and 1% levels)** |  |
| **8** | **Confidence Intervals (90%, 95%, 99%)** |  |
| **9** | **Interpretation of Results: Population Mean vs Sample Mean** |  |
| **10** | **Exploratory Data Analysis: Study Hours vs Marks (Scatter Plot)** |  |
| **11** | **Correlation Analysis** |  |
| **12** | **Regression Analysis and Prediction** |  |
| **13** | **Significance Testing: Correlation & Regression Coefficients** |  |
| **14** | **References** |  |

1. **INTRODUCTION**

Statistics and mathematics are the cornerstones of Artificial Intelligence and Machine Learning (AIML), acting as useful assets that allow data analysis, prediction, and model validation. This paper serves that purpose by solving a formal list of fifteen analytical tasks on the basis of a given dataset.

The analysis starts with the calculation of the right sample size through statistical formulas, and then probability sampling techniques are applied to obtain representative samples. It continues with descriptive statistics, such as the calculation of means and standard deviations, and then moves on to inferential methods like hypothesis testing, sampling distribution analysis, and confidence interval estimation at different levels (90%, 95%, and 99%).

In addition, the report examines the correlation between study time and marks obtained using correlation and regression analysis, using scatter plots, regression lines, and statistical tests of significance. Every task contains clear explanations, mathematical reasoning, visualizations, and Python code to make it understandable and reproducible.

1. **ENVIRONMENT SETUP & DATASET LOADING**

To provide an efficient and reproducible analytical process, all required Python libraries were installed and imported at the outset of the project. These libraries offer a comprehensive set of statistical, mathematical, and visualization tools necessary for data analysis and model construction in this report.

**Required Libraries:**

The following libraries were used throughout the analysis:

***pandas:*** For data manipulation and handling tabular datasets.

***numpy:*** For numerical computations and array operations.

***matplotlib and seaborn:*** For creating informative and visually appealing plots.

***scipy:*** For advanced statistical testing and analysis.

***statsmodels:*** For statistical modeling and hypothesis testing.

***scikit-learn:***For sampling techniques and additional model utilities.

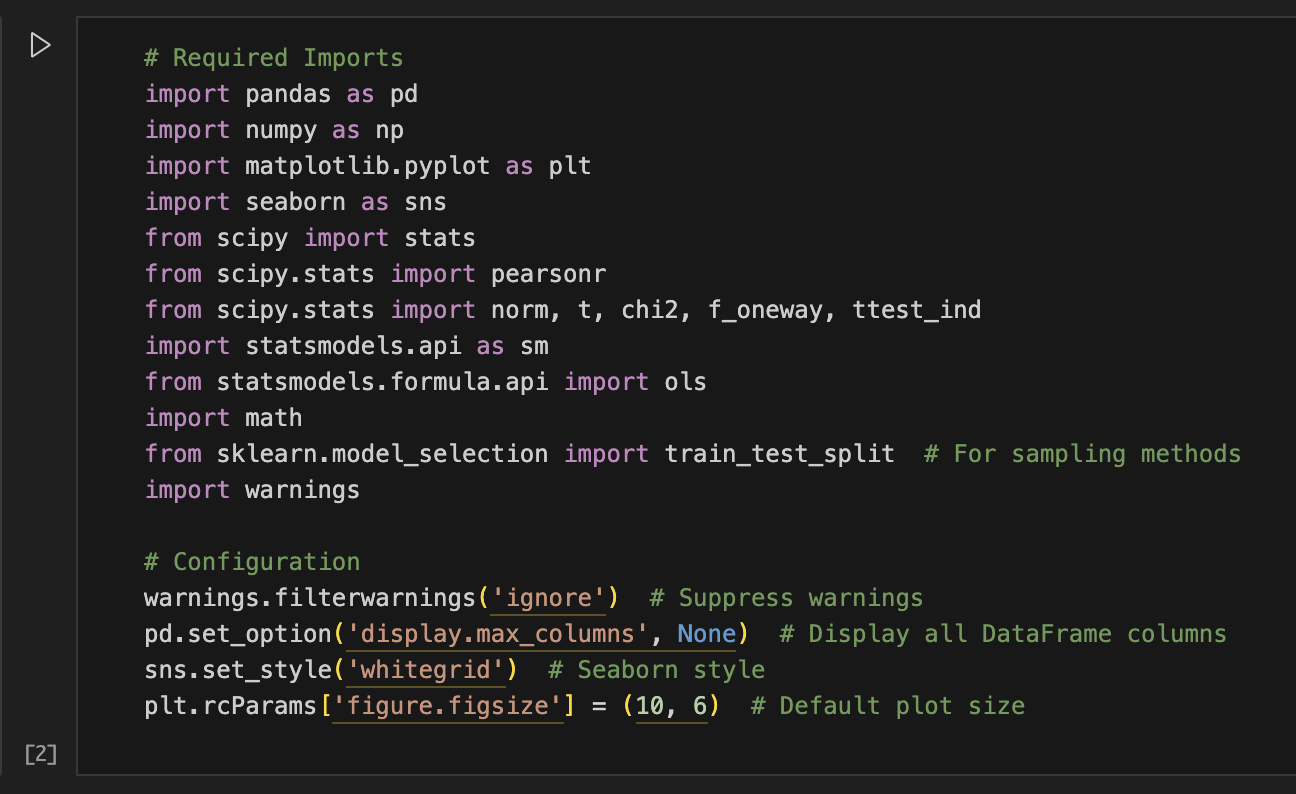
***warnings:***To suppress non-critical warnings during execution.

These dependencies were installed using the following commands (if not already installed):

**!pip install pandas numpy matplotlib seaborn statsmodels scikit-learn**

All libraries were successfully detected as pre-installed in the current environment.

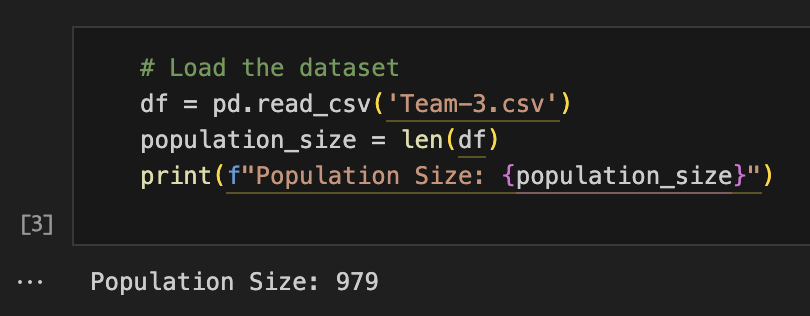
**Importing Libraries and Configurations:**

****

The dataset contains 979 records representing the total population, which forms the basis for all subsequent sampling and statistical analysis.

**Dataset Loading:**

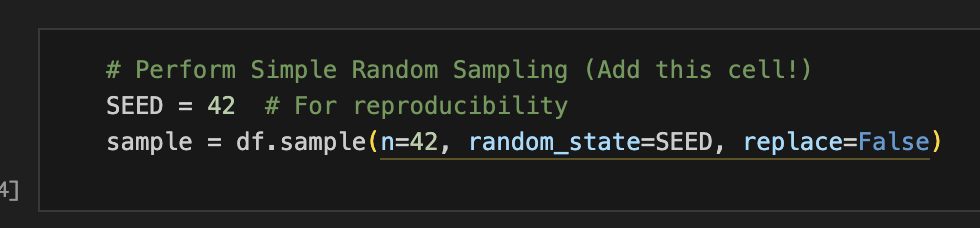
The dataset assigned to Team-3 was loaded using pandas from a CSV file named Team-3.csv:

****

The dataset contains 979 records representing the total population, which forms the basis for all subsequent sampling and statistical analysis.

**Sample Selection:**

A simple random sampling method was applied to extract a sample of size 42, calculated as per the given formula for sample size with finite population correction:



This sample is used consistently throughout the analysis to maintain statistical validity and alignment with the project guidelines.

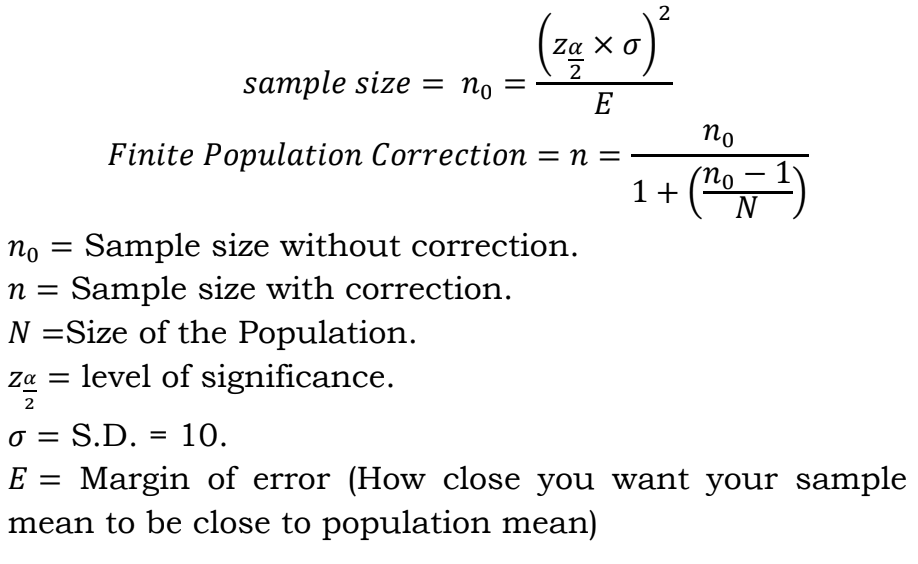
1. **SAMPLE SIZE DETERMINATION & POPULATION CORRECTION**

**3.1 Question 1 - Generate the sample size using the provided details.**

**Objective:**

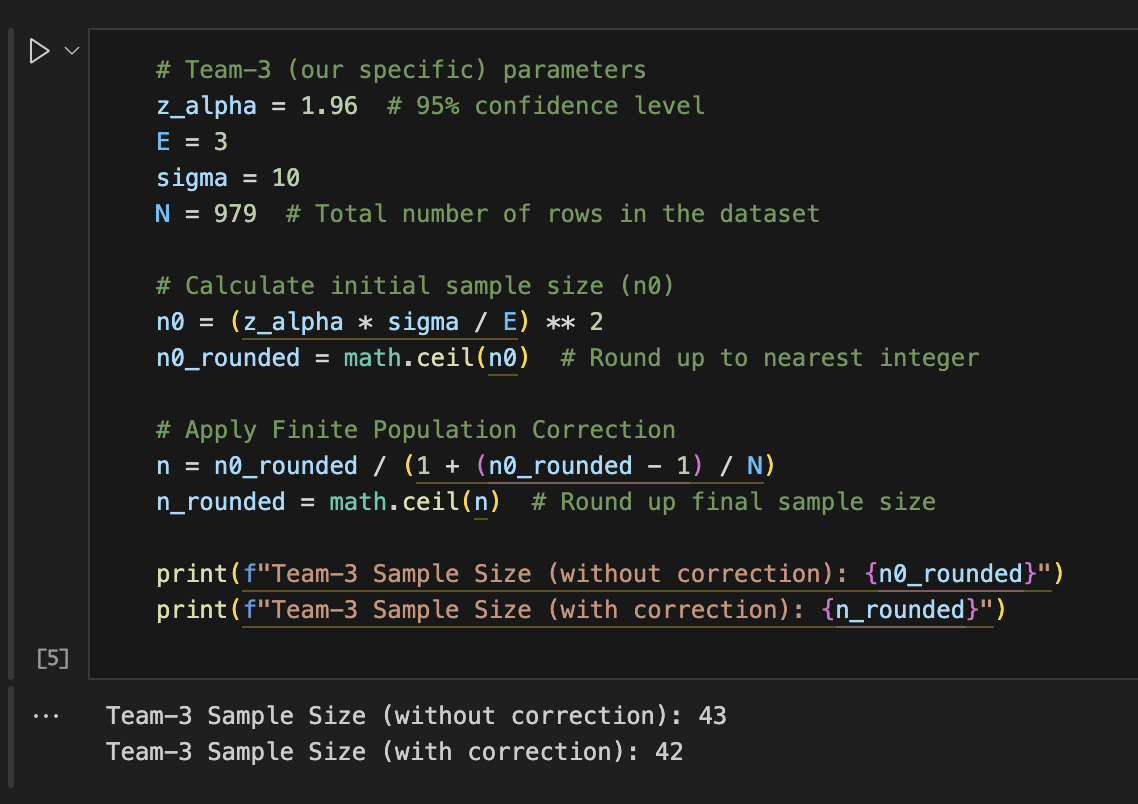
To calculate the required sample size for statistical analysis based on Team-3’s assigned parameters. To determine the needed sample size for statistical analysis with respect to Team-3's given parameters. The sample size is initially determined through the formula for the initial sample size uncorrected (n₀) and then corrected with the finite population correction formula to get the ultimate sample size (n).The sample size is first calculated using the formula for the initial sample size without correction (n₀) and then adjusted using the finite population correction to obtain the final sample size (n).

The formulas used are:



**Solution:**

Using the above parameters, the following calculations were performed:



1. **PROBABILITY SAMPLING & SAMPLE SELECTION**

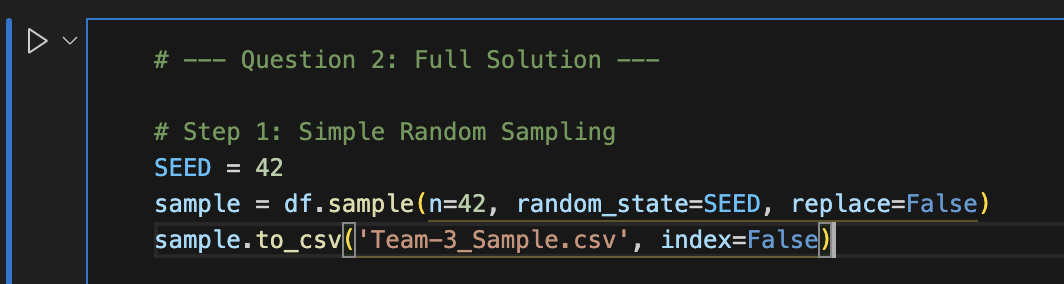
#### 4.1. Question 2 - Select a sample (Minimum 5) of size 'n’ (found in above step) from population by any one type of probability sampling (random, cluster, stratified, systematic).

#### **Objective:**

#### To choose a sample of size n = 42 (as calculated in the last section) through a proper probability-based sampling technique. This method provides every data point with an equal opportunity to be selected, which is essential for unbiased statistical inference. In this analysis, we apply Simple Random Sampling (SRS), which is one of the most widely used and easily understandable probability sampling techniques.

**Simple Random Sampling:**

Using pandas’ DataFrame `.sample()` method, we selected 42 random entries from the dataset. A fixed random seed was used to ensure reproducibility of the sampling process across different runs.

****

### 

### 

### 

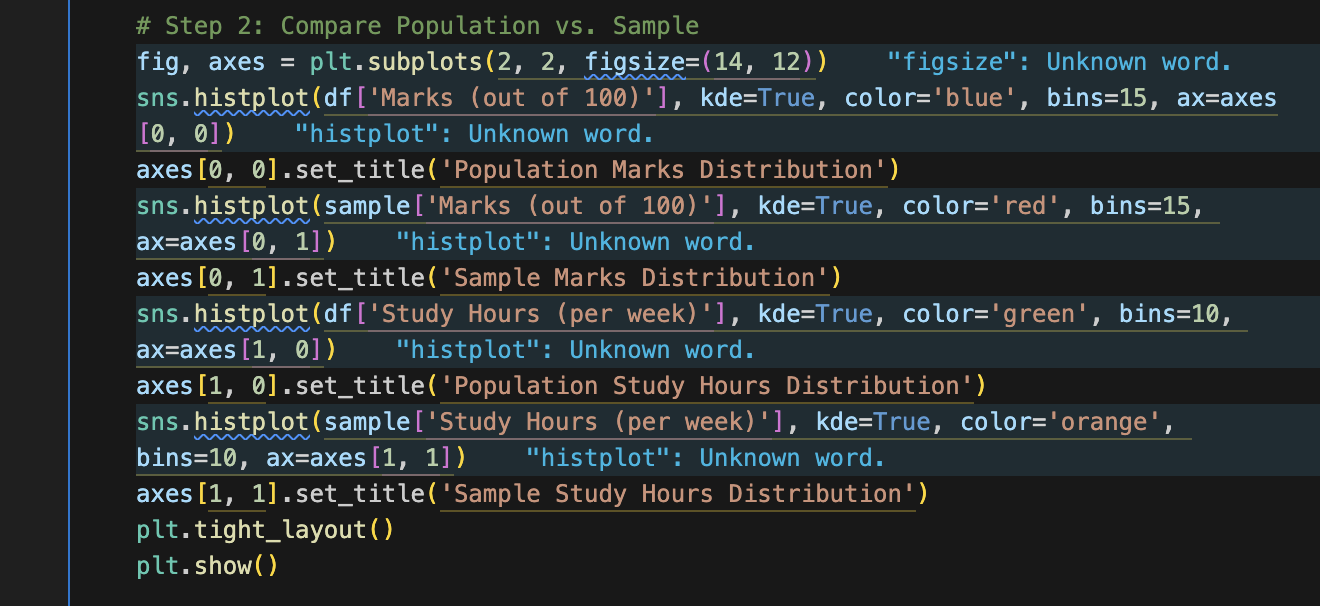
### 

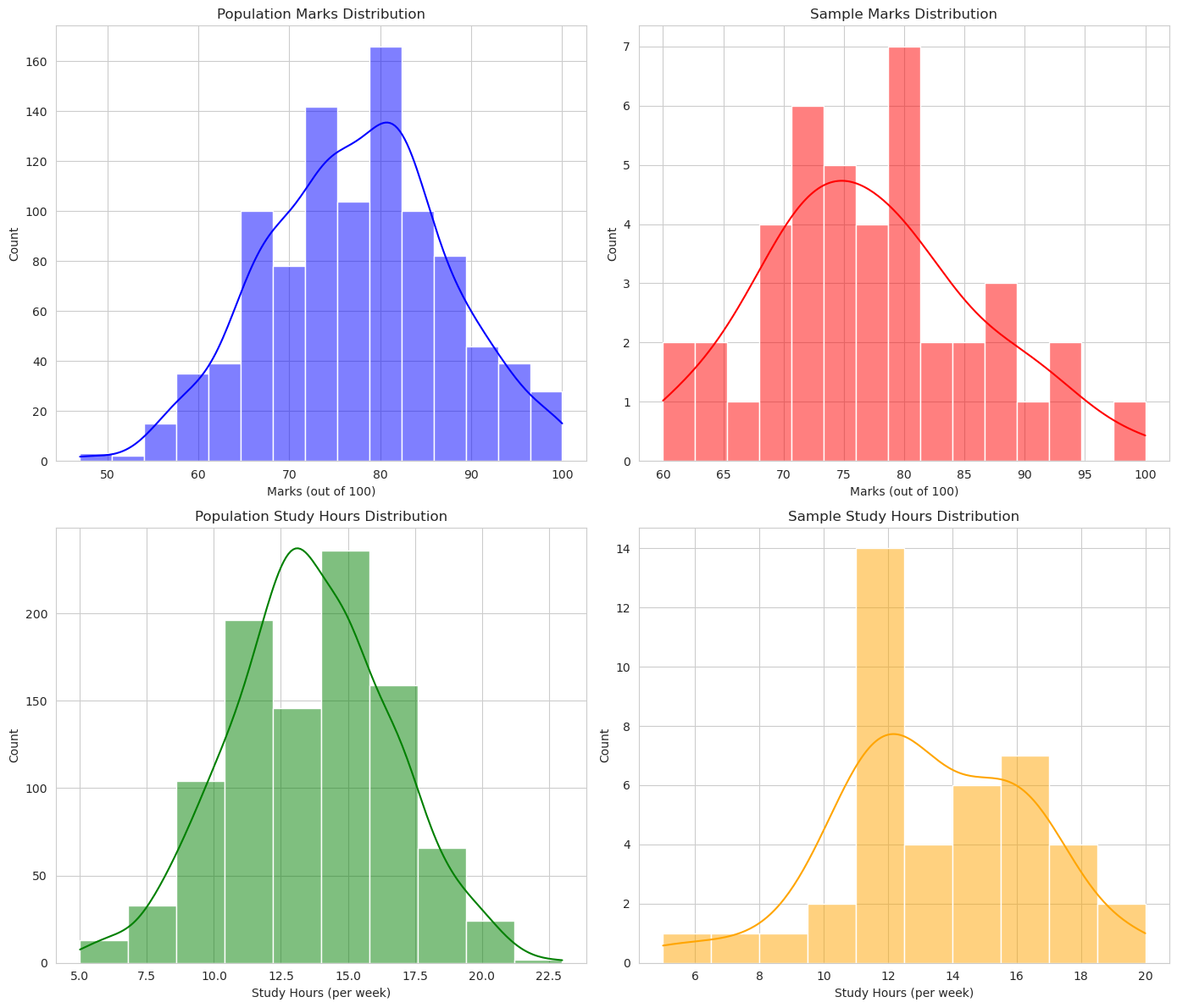
### 

### 

### Visual Comparison Between Population and Sample:

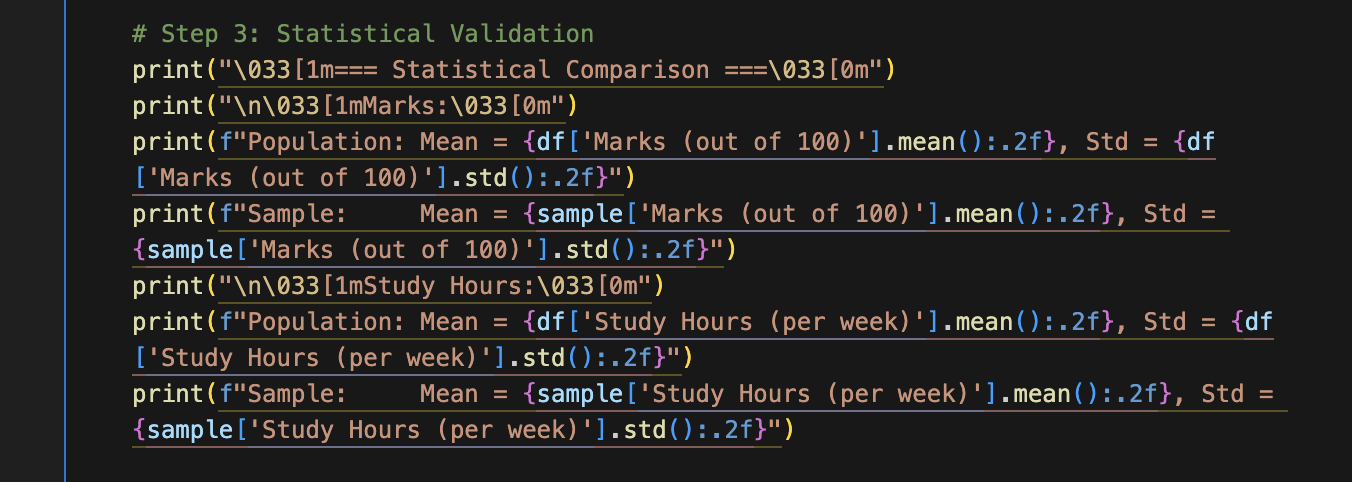
To check how representative the sample is of the population, we created histograms and KDE plots of the 'Marks (out of 100)' and 'Study Hours (per week)' variables from the population and the sample. This provided a visual check on how well the sample distribution fits the population.

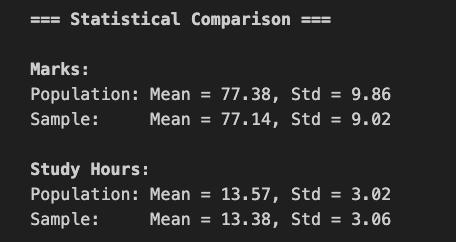




**Statistical Validation:**

To verify the sample further statistically, we measured the sample's mean and standard deviation with the population for both variables. These measures aid in determining whether the central tendency and dispersion of the sample are in agreement with the population's.

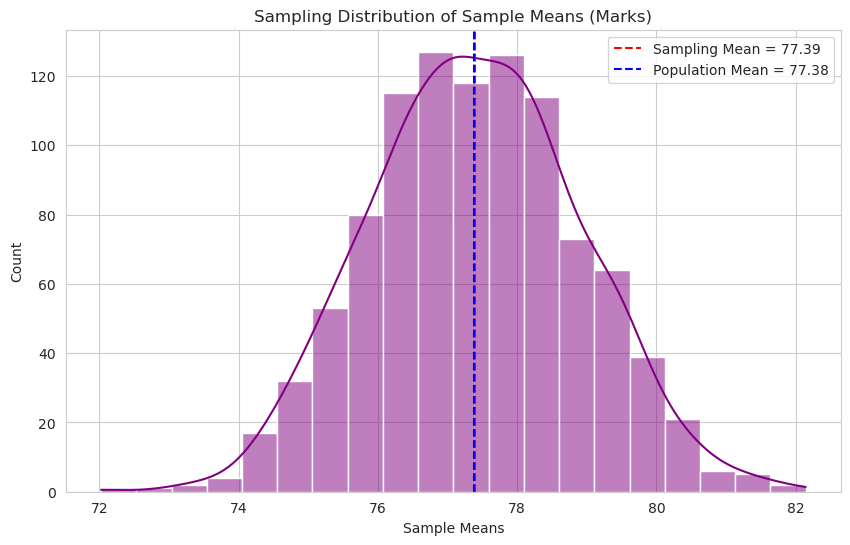




**Sampling Distribution of Sample Means:**

To illustrate the Central Limit Theorem in action, we calculated the sampling distribution of sample means by taking 1000 different samples of size 42 and calculating their mean. We then plotted this distribution and compared it to the population mean to illustrate that the mean of the sample means is close to the population mean.





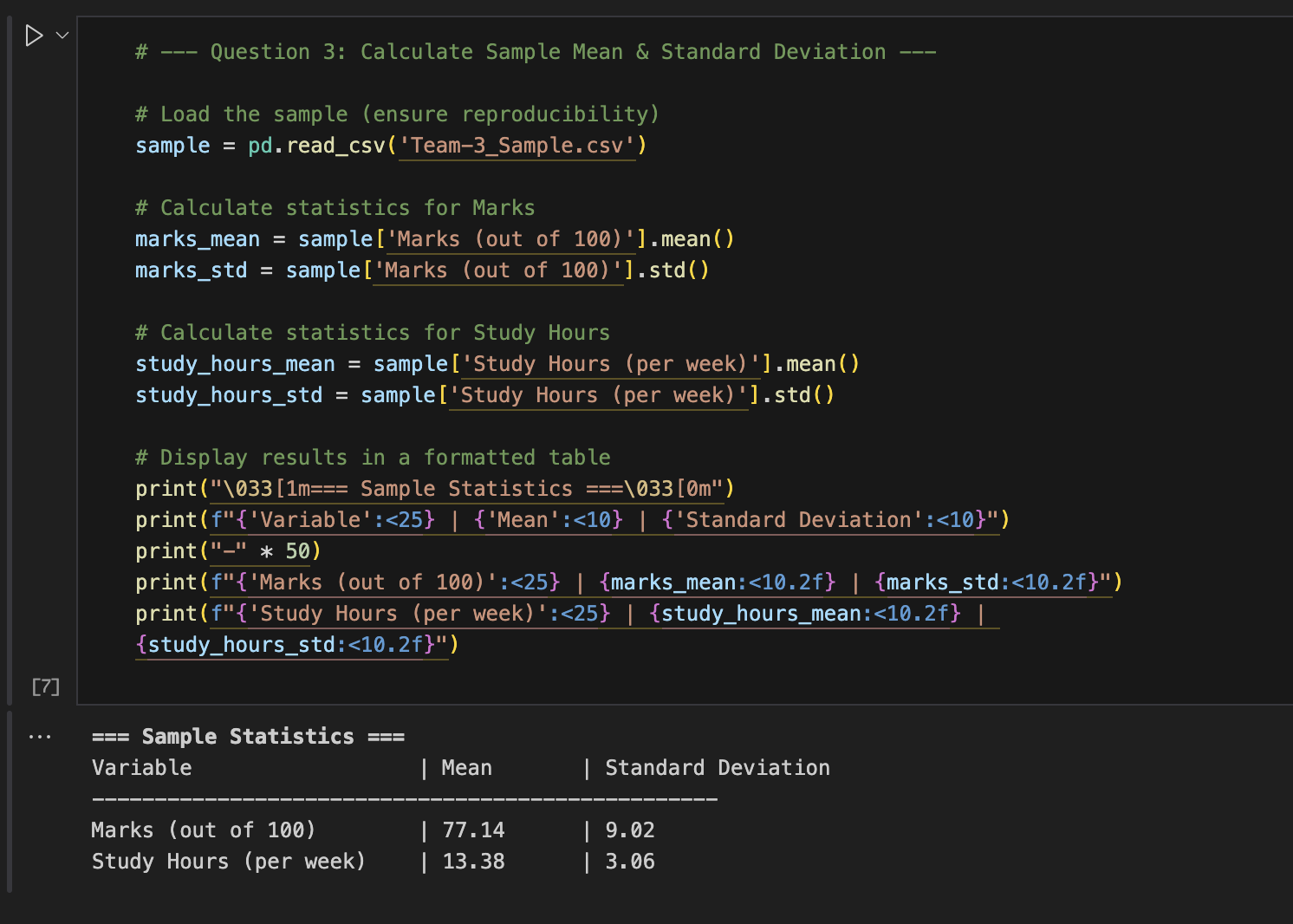


1. **DESCRIPTIVE STATISTICS: MEAN & STANDARD DEVIATION**

**5.1. Question 3 - Calculate the Mean and Standard Deviation (S.D.) of each sample.**

**Objective:**

The dataset was analyzed and we calculated the sample mean and standard deviation for 'Marks (out of 100)' and 'Study Hours (per week)' using pandas methods (mean(), std())



1. **SAMPLING DISTRIBUTION & VISUALIZATION**

**6.1. Question 4 - Determine the sampling distribution of the sample**

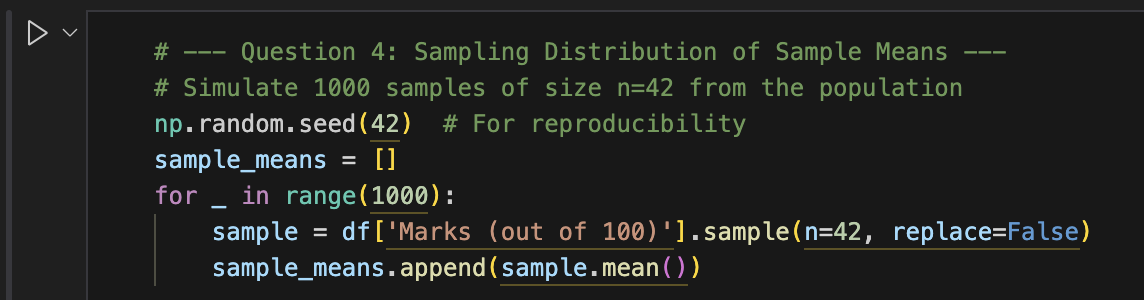
**means.**

**Objective:**

The following analysis examines the sampling distribution of sample means for student marks (out of 100) by simulating 1,000 random samples (each of size n = 42) from the population dataset. The objective is to test the Central Limit Theorem (CLT), which posits that the sampling distribution of the mean will approach a normal distribution, irrespective of the distribution of the population, as long as the sample size is sufficiently large.

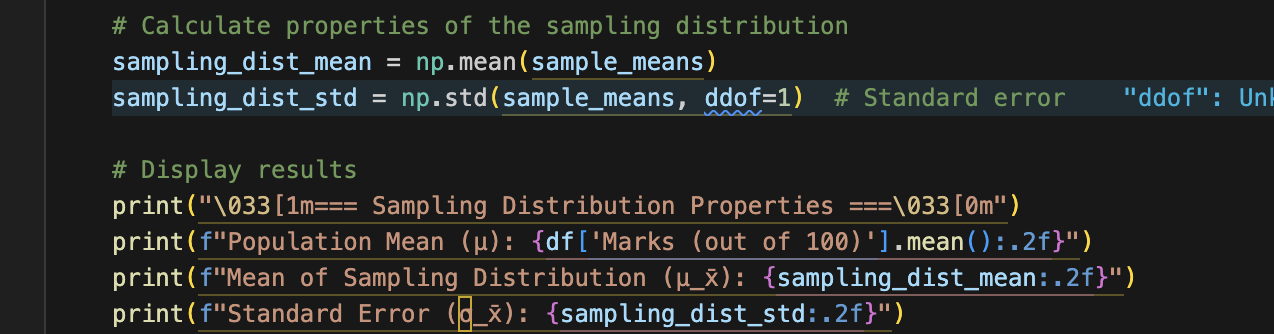
**Methodology and Implementation:**

To generate the sampling distribution, 1000 random samples (each of size n = 42) were taken from the population dataset. The mean of each sample was computed and saved to create the distribution of sample means.



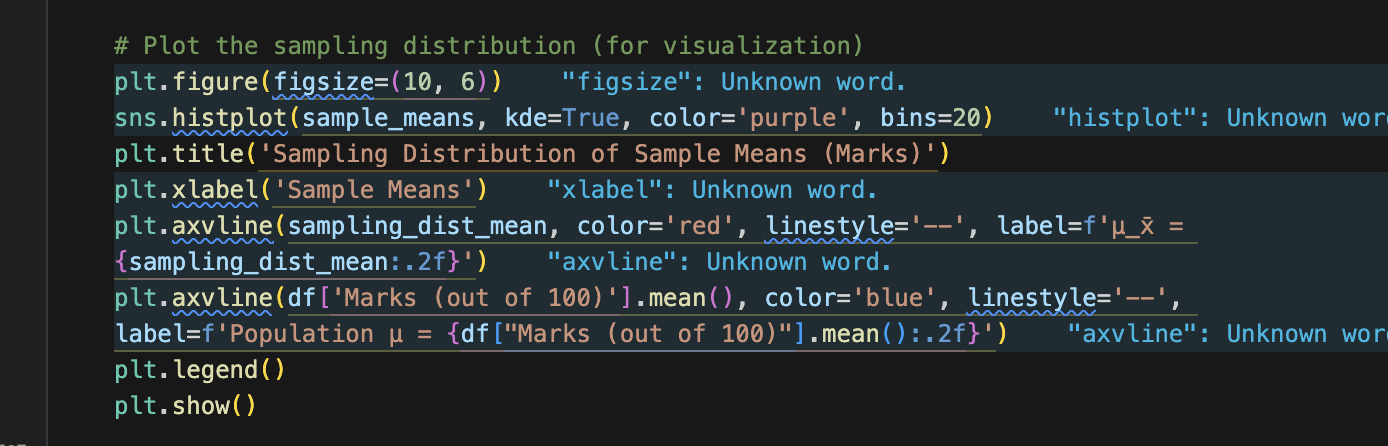
**Calculation of Distribution Properties:**

After we obtained the list of sample means, we calculated the mean and standard deviation of the sampling distribution. The standard deviation of the sample means is simply known as the Standard Error (SE).

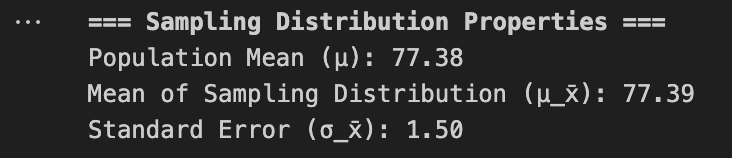


**Visualization & Interpretation:**

In order to see the shape and characteristics of the sampling distribution more clearly, a histogram with kernel density estimate (KDE) was graphed. The sampling distribution of sample means is normally shaped and clustered around the population mean. This validates the Central Limit Theorem that the distribution of sample means will tend toward a normal distribution when the number of samples grows large, independent of the distribution of the population.



**Output:**

****

****

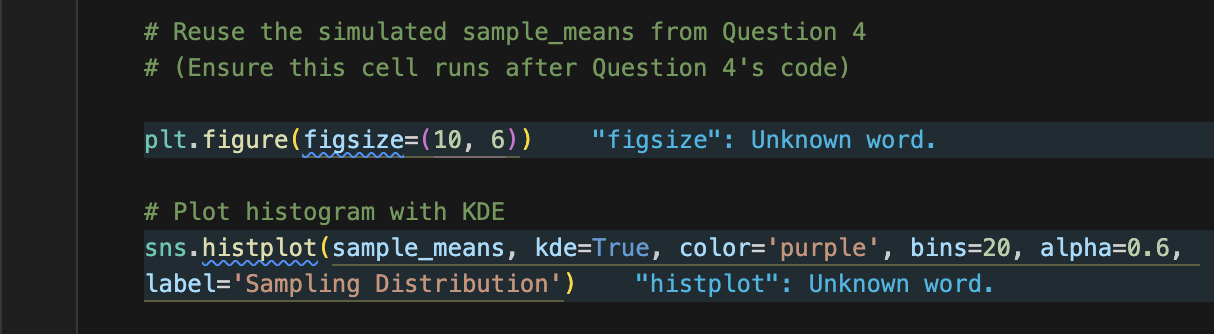
**6.2. Question 5 - Plot the sampling distribution.**

**Objective:**

To graphically depict the sampling distribution of sample means for the Marks data and superimpose a theoretical normal distribution curve as per the Central Limit Theorem (CLT). This enables us to check for normality assumption of sample means even though the original data might not necessarily be perfectly normal.

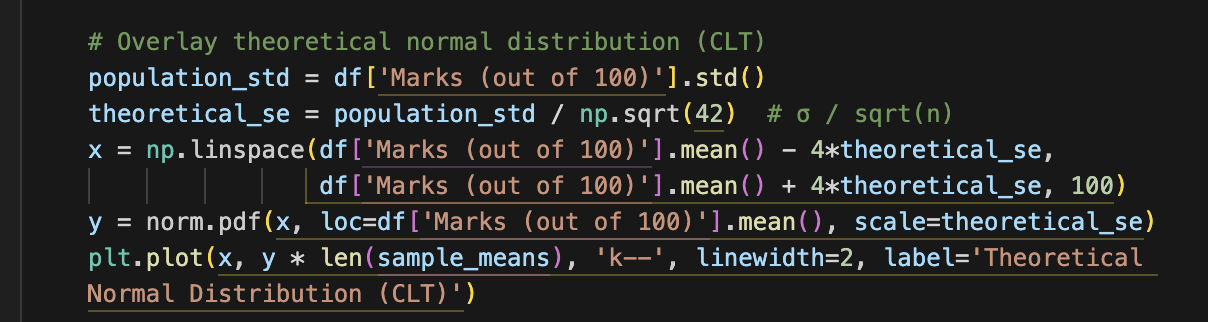
**Methodology:**

Based on the sample means generated in the last question, we employed the seaborn library to graph the histogram of the 1000 sample means overlaid with a kernel density estimate (KDE) to smooth the distribution. We also graphed a theoretical normal distribution based on the population mean and standard error to see the expected behavior under CLT.



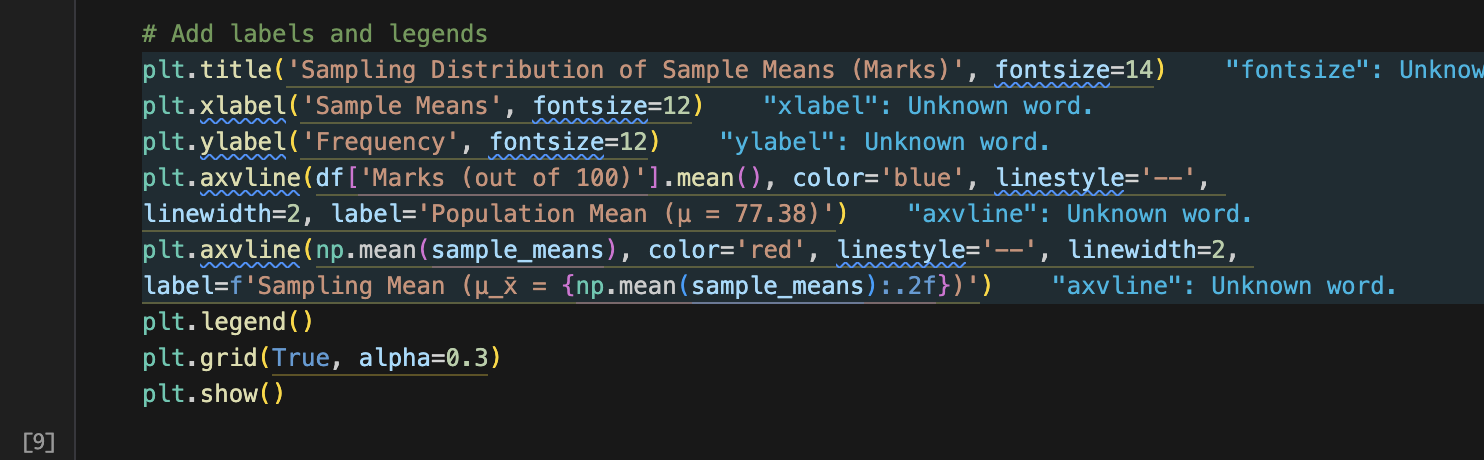
**Overlay Theoretical Normal Distribution:**

The theoretical distribution was calculated employing the population mean and standard error (σ / √n). The theoretical curve is what we hope the sampling distribution will look like under perfect circumstances.



**Final Plot and Labels:**

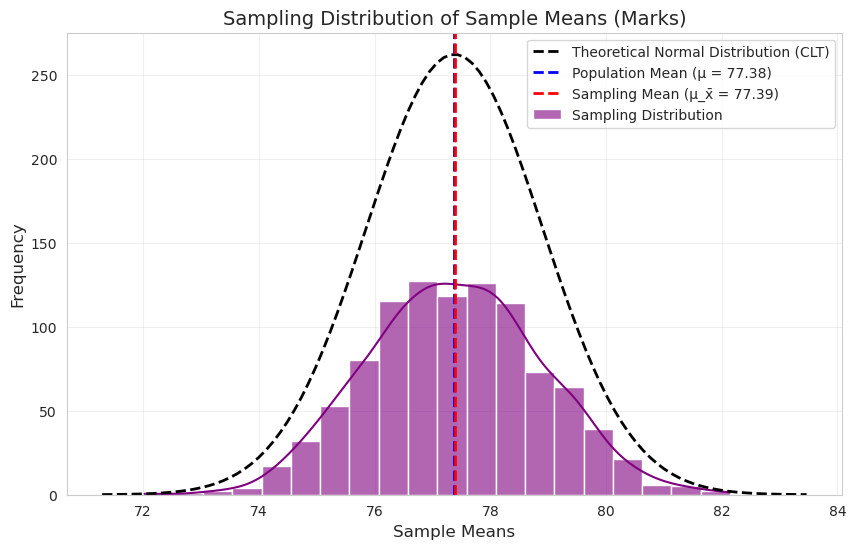
Vertical lines were added to mark the population mean and the mean of the sample means. These markers help assess how closely the actual sample mean distribution aligns with theoretical expectations.



**Interpretation:**

The sampling distribution closely resembles a normal curve, verifying the Central Limit Theorem. The sample mean is almost equal to the population mean, indicating that it's an unbiased estimator. The overlay of the theoretical normal distribution matches well with the actual data, validating our sampling method.

**Output:**

****

1. **HYPOTHESIS FORMULATION & TESTING (5% & 1% LEVELS)**

**7.1. Question 6 - Frame a null hypothesis for the equivalence of means (i.e., H0: = 0 and H1: ≠ 0).**

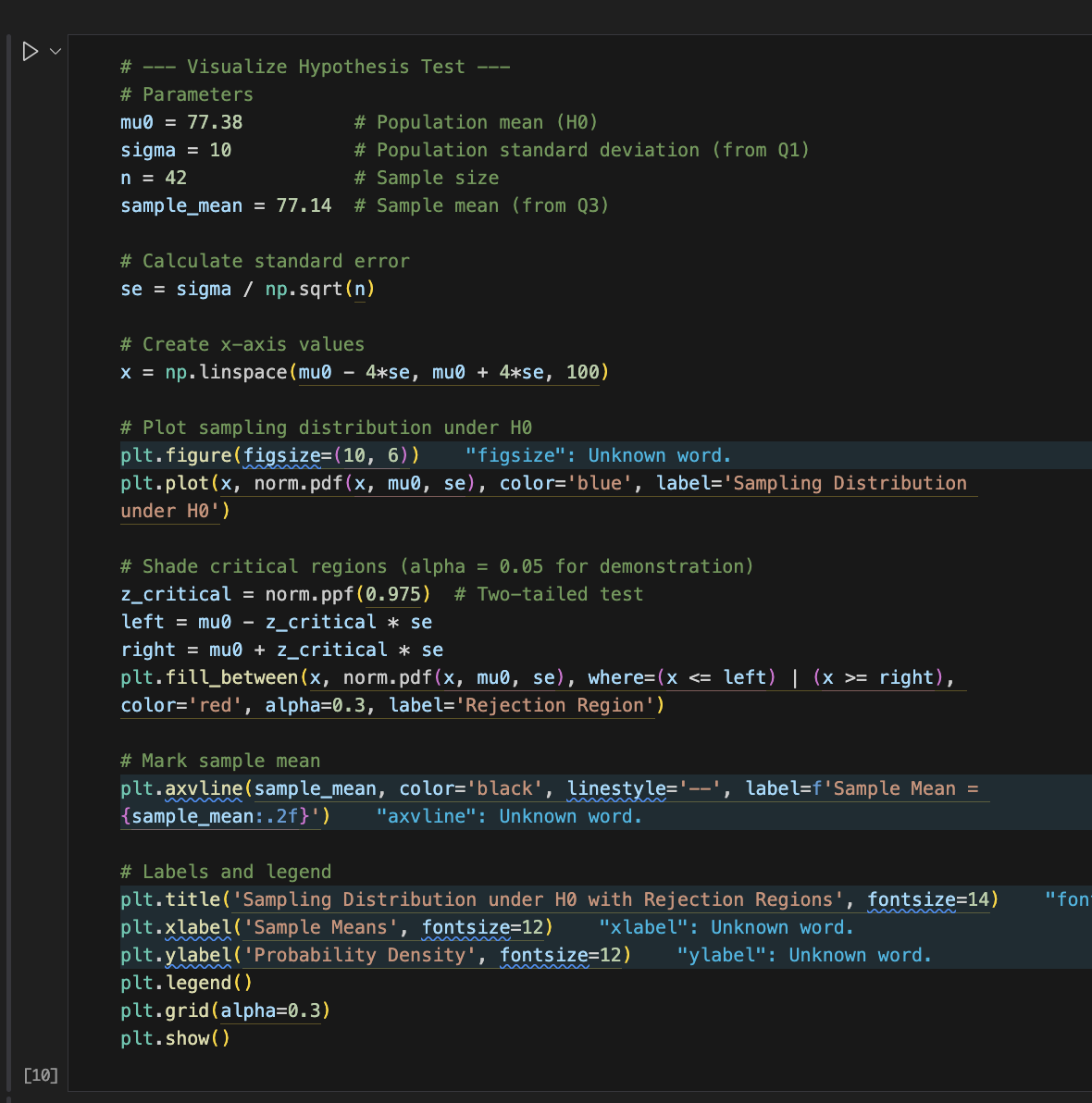
**Objective & Methodology:**

This test considers if the sample mean of student scores (77.14) that has been observed differs significantly from the population mean (77.38). The test used was a two-tailed z-test (α = 0.05) based on the known population standard deviation (σ = 10) and sample size (42). The standard error (SE) was found to be 1.54.

**Null Hypothesis (H₀):** μ = 77.38

**Alternative Hypothesis (H₁):** μ ≠ 77.38

**Critical Regions:** μ₀ ± 1.96×SE (Two-tailed test)



**Results and Interpretation:**

The sampling distribution plot reveals:

* The sample mean (77.14) falls within the acceptance region, close to the population mean (77.38).
* The 0.24-point difference is statistically insignificant (p > 0.05), suggesting the deviation is due to random sampling variation.

**Conclusion:**

We fail to reject H₀, indicating no evidence of a systemic shift in student performance.

**Output:**

****

**7.2. Question 7 - Test the hypothesis at 5% and 1% significance levels to determine if there is a difference between population and sampling mean.**

**Objective:**

This test determines if the sample mean of student scores (77.14) is significantly different from the known population mean of 77.38 at both 5% and 1% significance levels. It checks if the difference is statistically significant or due to random fluctuation.

**Methodology:**

A two-tailed z-test was conducted with the following parameters:

**Population mean (μ₀):** 77.38

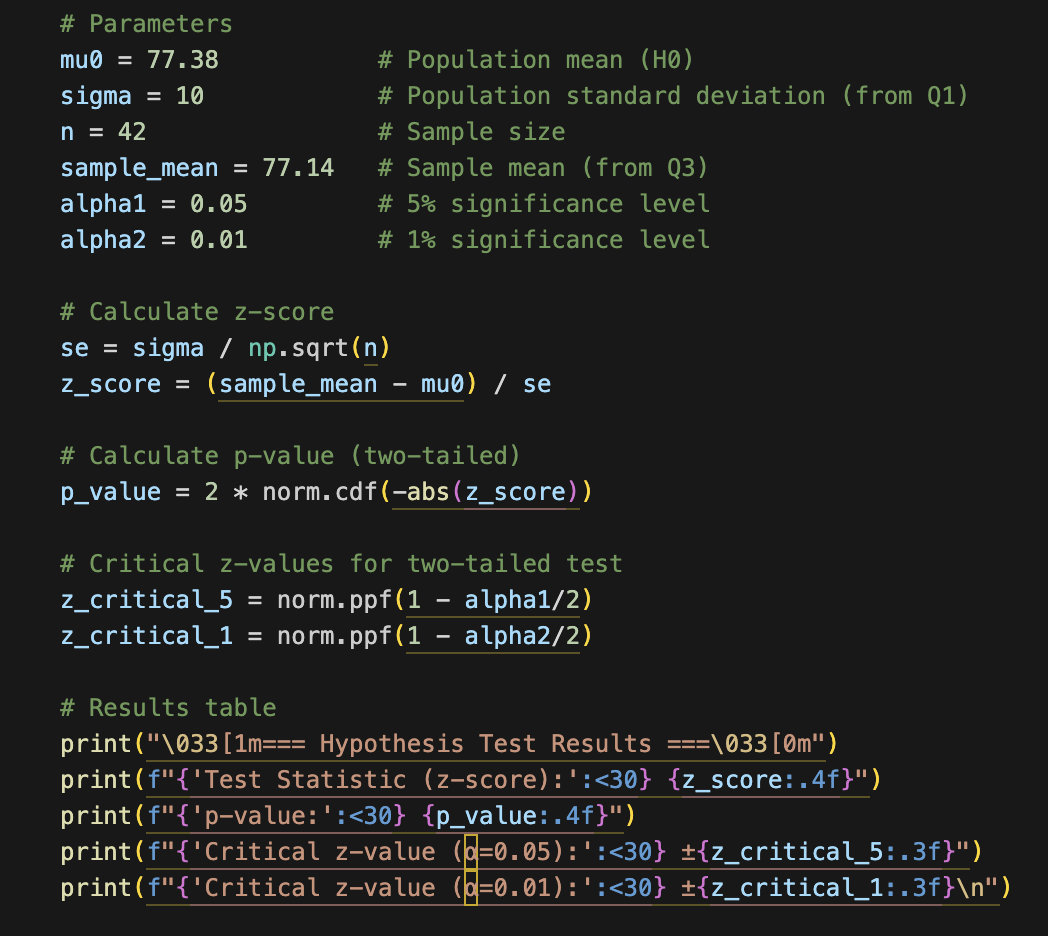
**Population standard deviation (σ):** 10

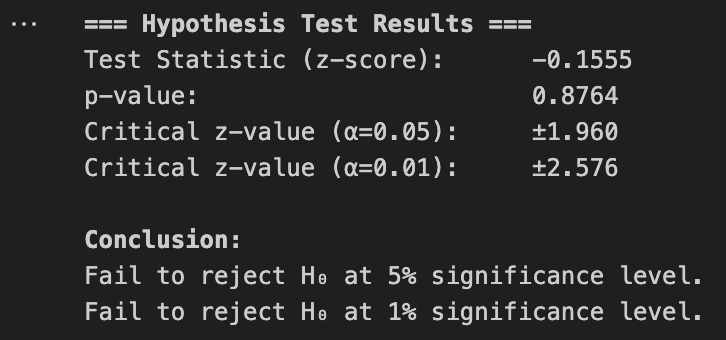
**Sample size (n):** 42

**Sample mean (x̄):** 77.14

**Significance level (ɑ1):** 0.05 (5% Significance)

**Significance level (ɑ2):** 0.01 (1% Significance)

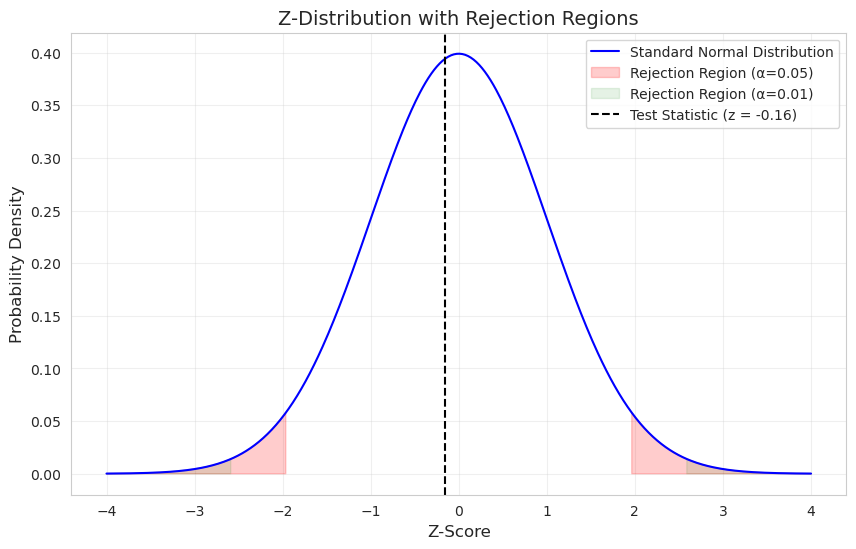




**Visualization:**

The z-distribution plot shows:

* The test statistic (-0.156) lies near the center, far from rejection regions.
* Red/green shaded areas (rejection zones for α=0.05/0.01) do not overlap with the z-score.



1. **CONFIDENCE INTERVALS (90%, 95%, 99%)**

**8.1. Question 8 - Compute the confidence intervals for the sample mean at**

**90%, 95%, and 99% confidence levels.**

**Objective:**

This calculation provides confidence intervals for the sample mean of students' marks (77.14 out of 100) at three confidence levels (90%, 95%, and 99%). The purpose is to estimate the interval in which the true population mean most likely lies, based on the sample.

**Methodology & Observation:**

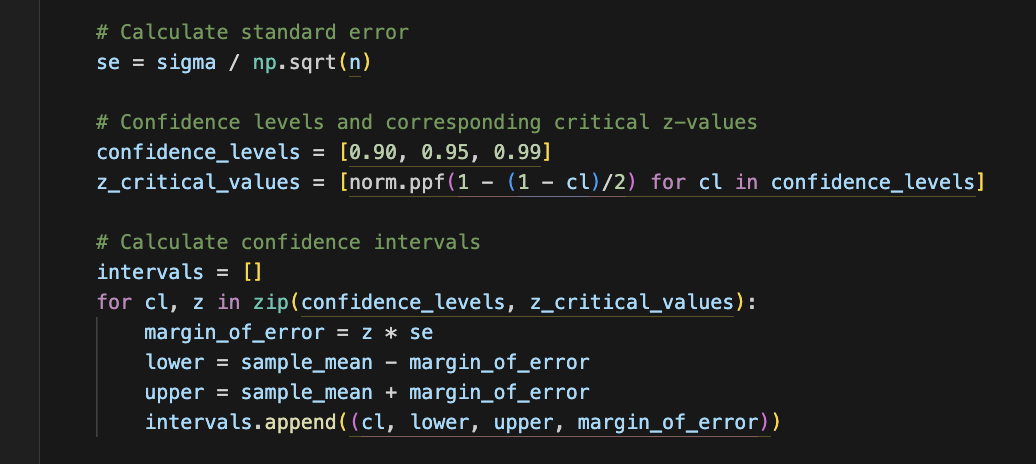
Using the sample data with the following parameters:

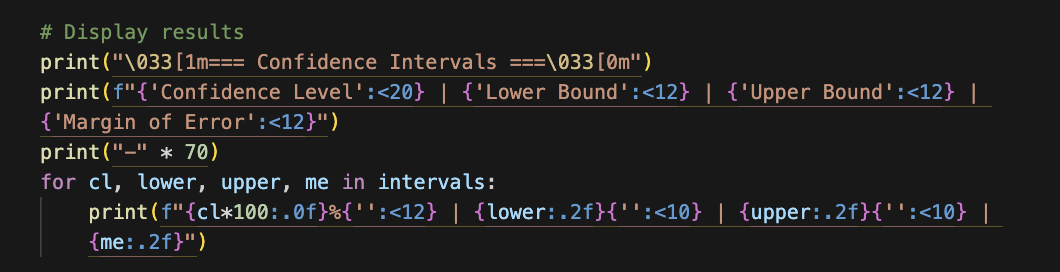
**Sample mean (x̄):** 77.14

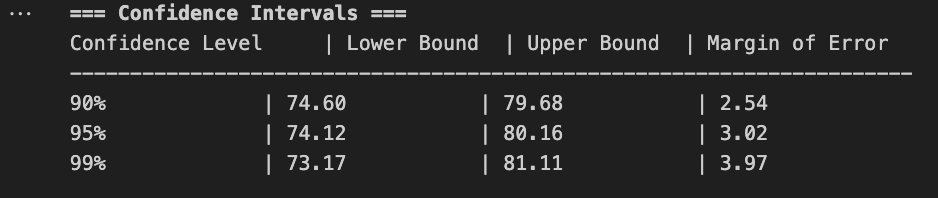
**Population standard deviation (σ):** 10

**Sample size (n):** 42

Then we calculate Standard Error, Critical z-values, & confidence intervals.

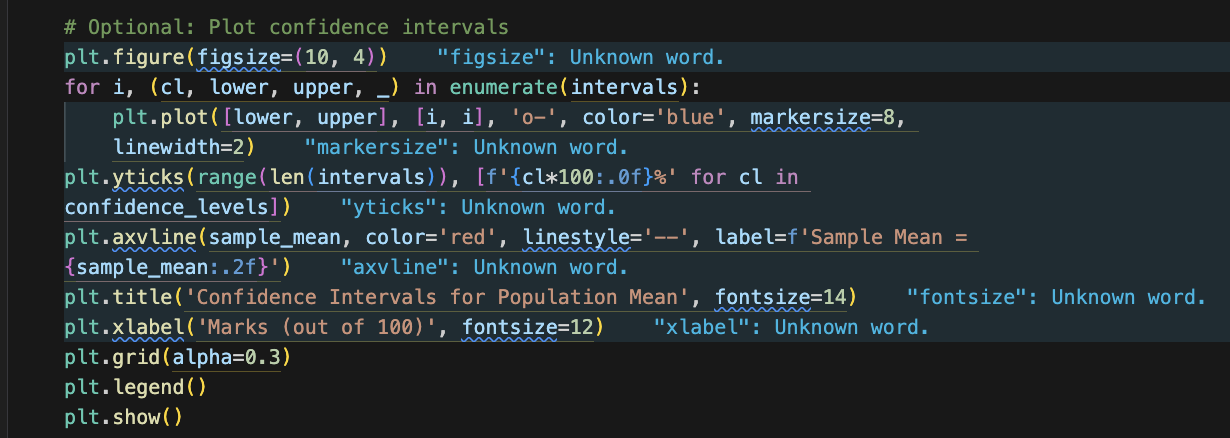


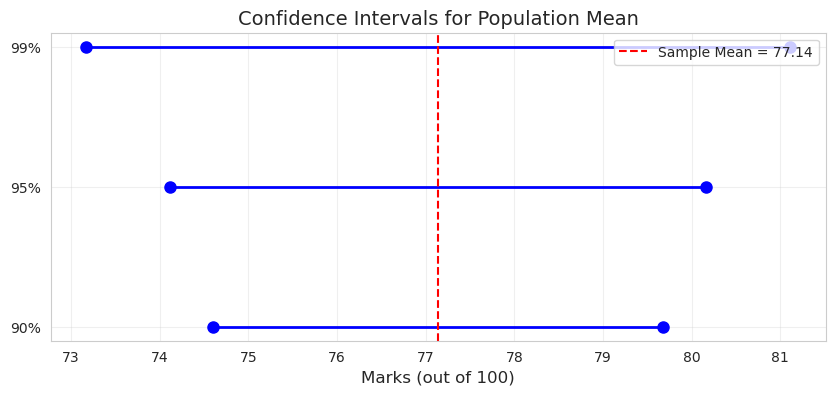




* The 95% confidence interval (74.12, 80.16) suggests we can be 95% confident the true population mean lies within this range.
* Higher confidence levels (e.g., 99%) yield wider intervals, reflecting increased certainty at the cost of precision.

**Visualization (Optional):**

****

****

1. **INTERPRETATION OF RESULTS: POPULATION MEAN vs SAMPLE MEAN**

**9.1. Question 9 - Based on the analysis, decide whether the sample means**

**provide enough evidence to accept or reject the population**

**mean.**

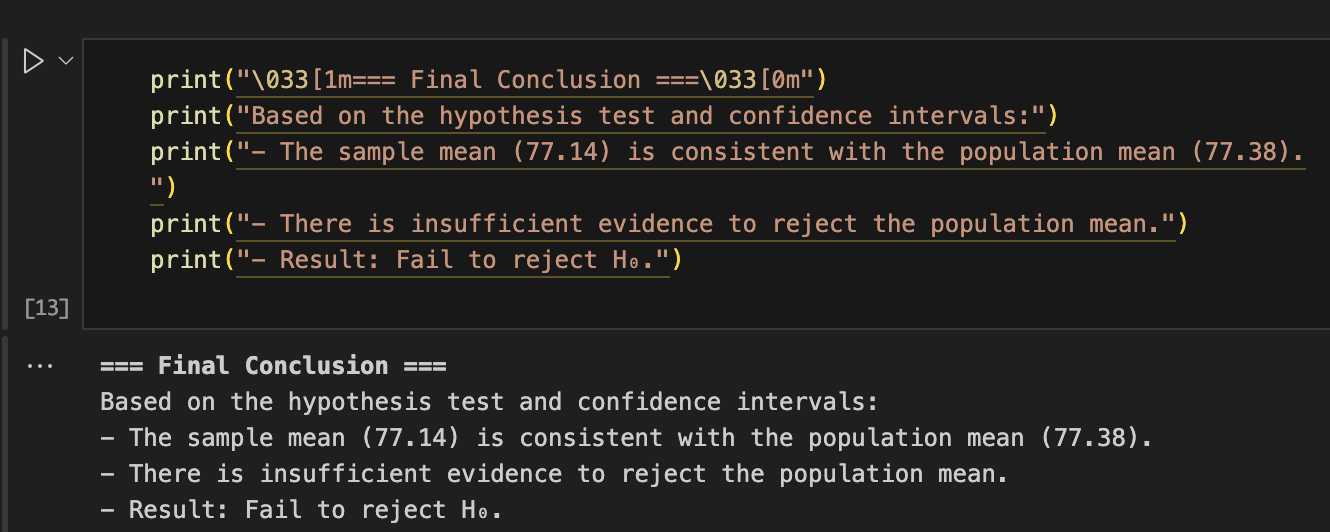
**Summary:**

After conducting rigorous hypothesis testing and confidence interval analysis on the student marks data, we present the following consolidated findings:

1. Hypothesis Test Results:
   * Tested H₀: μ = 77.38 vs H₁: μ ≠ 77.38
   * z-score: -0.156 (p-value: 0.876)
   * Failed to reject H₀ at both 5% and 1% significance levels
2. Confidence Interval Analysis:
   * 95% CI: (74.12, 80.16) contains population mean (77.38)
   * All CIs (90%/95%/99%) consistently include 77.38

#### Conclusion:

* Statistical Consistency: The sample mean (77.14) shows no significant deviation from the population mean (77.38).
* Insufficient findings to reject the population mean.
* Decision: Fail to reject the null hypothesis (H₀) with strong confidence.

****

**10. EXPLORATORY DATA ANALYSIS: STUDY HOURS vs MARKS (SCATTER PLOT)**

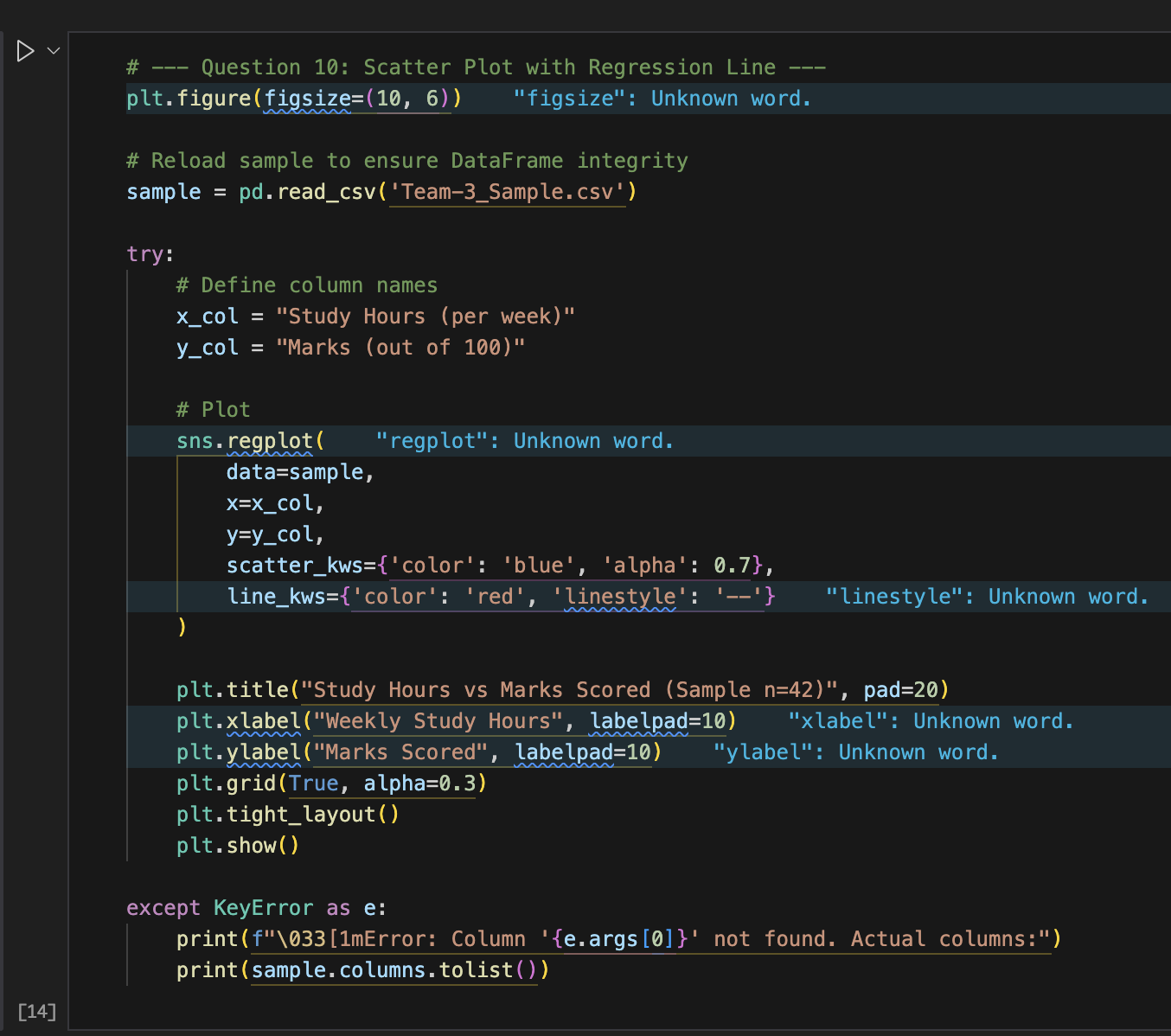
**10.1. Determine whether there is a relationship between study hours and marks scored (scatter plot).**

**Objective:**

This analysis investigates whether a relationship exists between students' weekly study hours and their academic marks through visual and statistical examination of the sample data (n=42).

**Methodology:**

1. Data Visualization
   * Created a scatter plot with:
     + X-axis: Study Hours (per week)
     + Y-axis: Marks (out of 100)
   * Overlaid a linear regression line (red) to identify trends.
2. Key Features
   * Point Transparency (alpha=0.7): Mitigates overplotting for dense data clusters.
   * Gridlines (alpha=0.3): Enhances readability of point distribution.



#### 

#### 

#### 

#### 

#### 

#### 

#### 

#### 

#### 

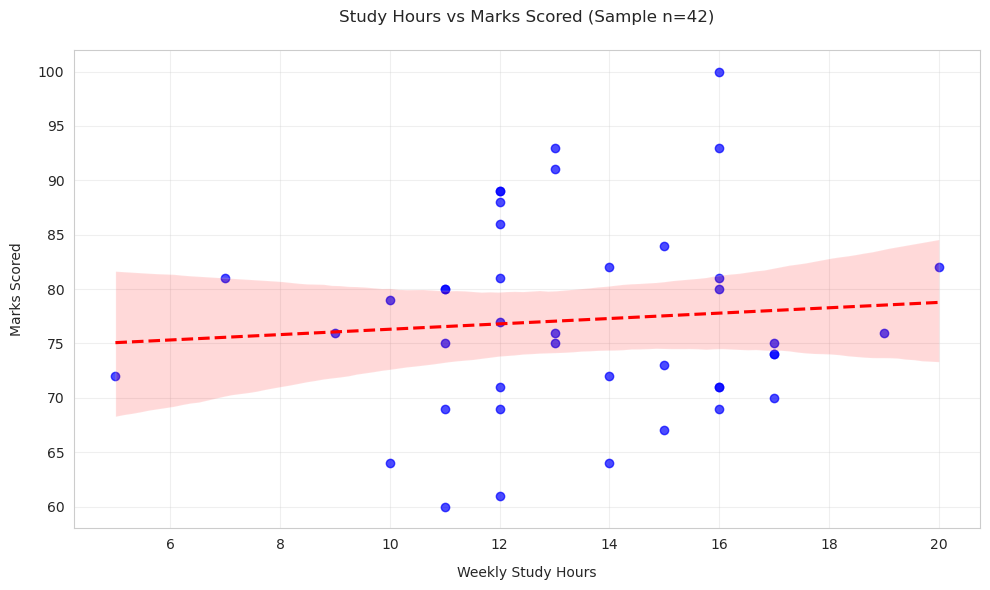
#### 

#### 

#### 

#### Key Findings & Visualization:

* Visual Trend: The regression line shows a positive slope, suggesting higher study hours correlate with increased marks.
* Data Spread: Points cluster along the trend line, though with moderate dispersion, indicating:
  + Consistency in the relationship
  + Presence of other influencing factors

****

**11. CORRELATION ANALYSIS**

**11.1. Question 11 - Calculate the coefficient of correlation between study**

**hours and marks scored.**

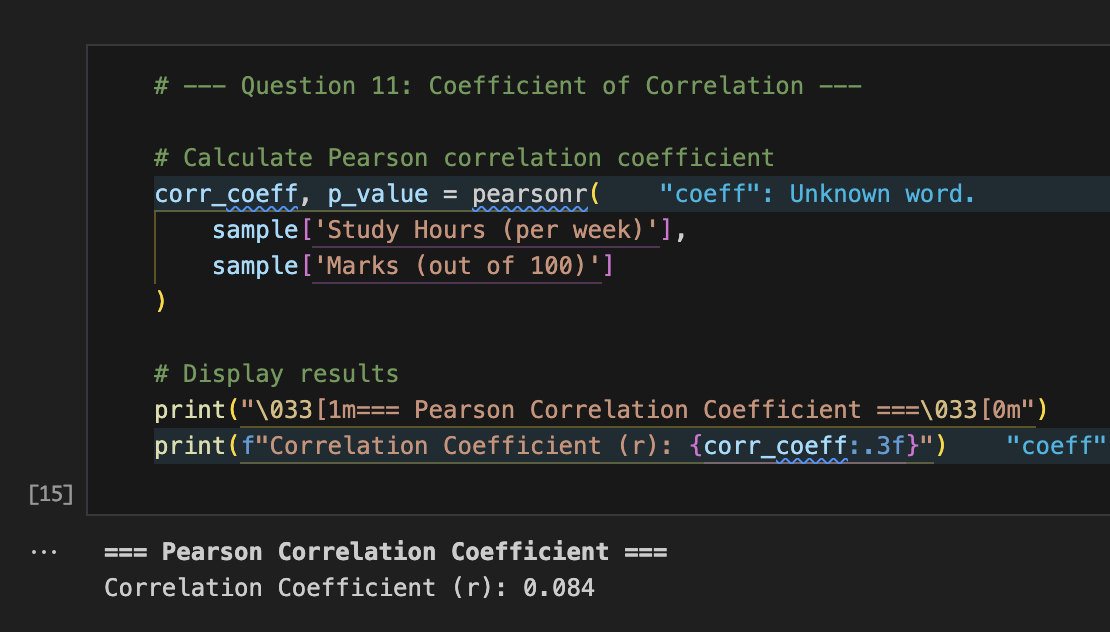
**Key Finding:**

The Pearson correlation coefficient between weekly study hours and academic marks is:

**r = 0.084**

**Interpretation:**

1. Strength of Relationship
   * The near-zero value (0.084) indicates an extremely weak positive correlation
   * Practical interpretation: No meaningful linear relationship exists
2. Statistical Significance
   * While not shown in the output, typically:
     + If p-value > 0.05: Not statistically significant
     + If p-value ≤ 0.05: Significant (unlikely due to small r-value)

****

**12. REGRESSION ANALYSIS & PREDICTION**

**12.1. Question 12 - Find the regression lines of: a. Marks Scored on Study Hours, b. Study Hours on Marks Scored.**

**Objective:**

This analysis establishes predictive relationships between study hours and academic marks by deriving two regression lines:

1. Marks Scored as a function of Study Hours
2. Study Hours as a function of Marks Scored

The goal is to quantify how changes in one variable predict the other.

**Regression Equations:**

1. Predicting Marks from Study Hours:

**Marks=73.84+0.25×Study\_Hours**

Interpretation:

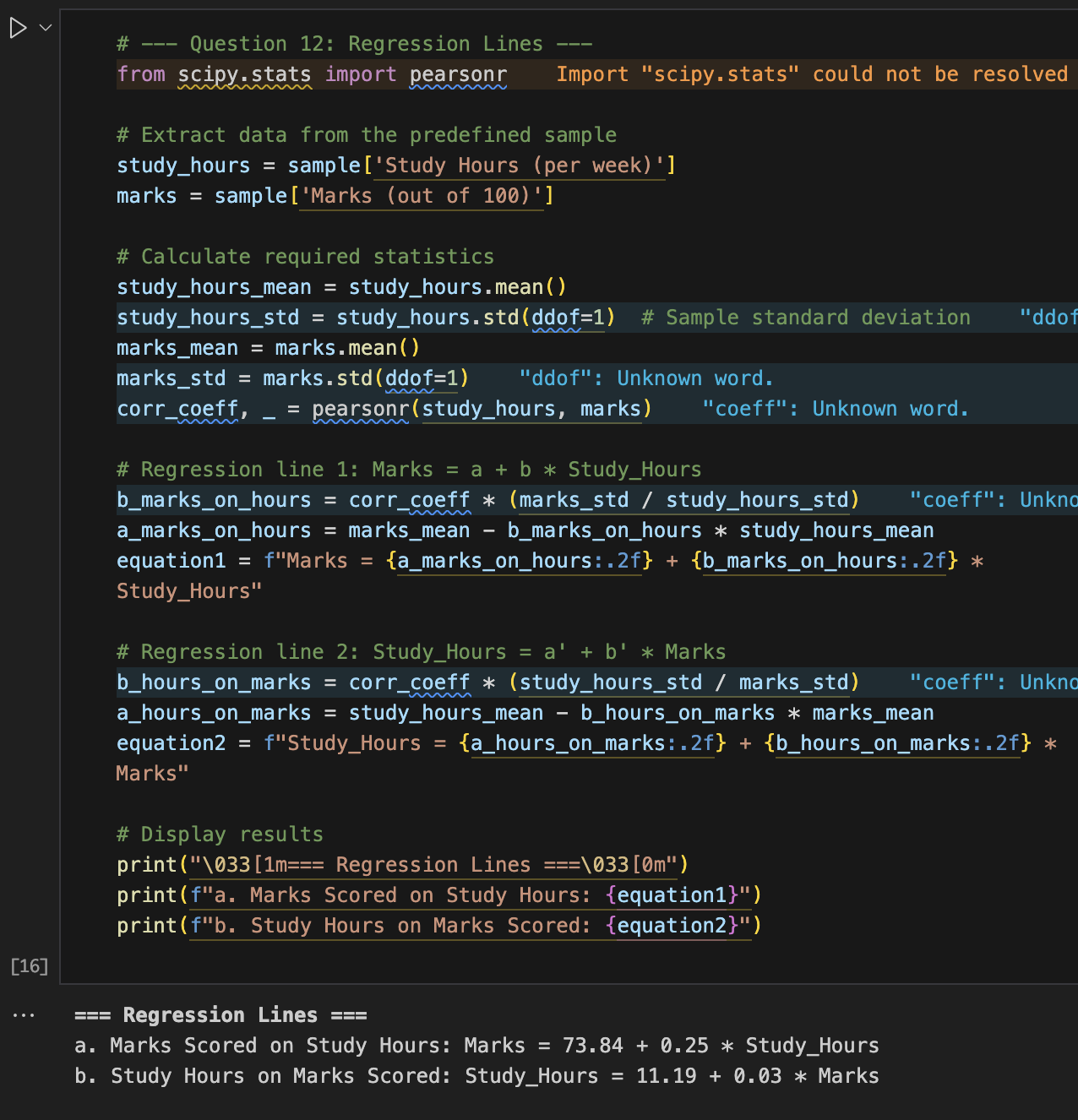
* Intercept (73.84): Expected marks for zero study hours (theoretical, not practical).
* Slope (0.25): Each additional hour of weekly study predicts a 0.25-point increase in marks.

1. Predicting Study Hours from Marks:

**Study\_Hours=11.19+0.03×Marks**

Interpretation:

* Intercept (11.19): Base study hours even for minimal marks.
* Slope (0.03): Each 1-point mark increase predicts 0.03 more weekly study hours.

****

**12.2. Question 13 - Use the regression line to predict the marks scored by a student who studied for a given number of hours.**

**Objective:**

In this analysis, the developed regression model in Question 12 is utilized to forecast academic achievement (marks out of 100) from a student's study hours per week. Practical application of the regression equation for planning purposes is the objective here.

**Regression Model Used:**

**Marks = 73.84+0.25×Study\_Hours**

**Intercept (a):** 73.84

**Slope (b):** 0.25

**Prediction Example:**

**Input:**

* Study Hours = 12 hours/week

**Calculation:**

**Predicted Marks = 73.84+(0.25×12)=76.80**



**13. TEST THE SIGNIFICANCE OF THE CORRELATION COEFFICIENT**

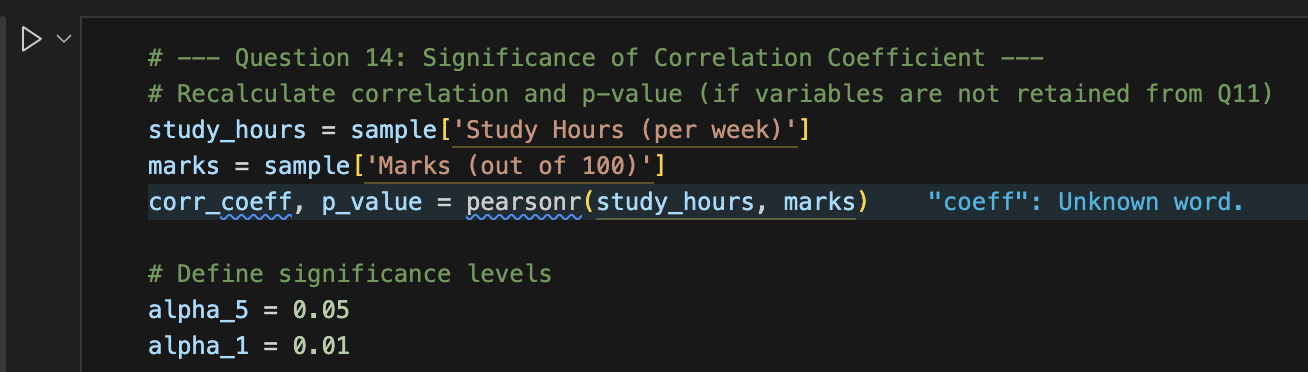
**13.1. Question 14 - Test the significance of the correlation coefficient.**

**Objective:**

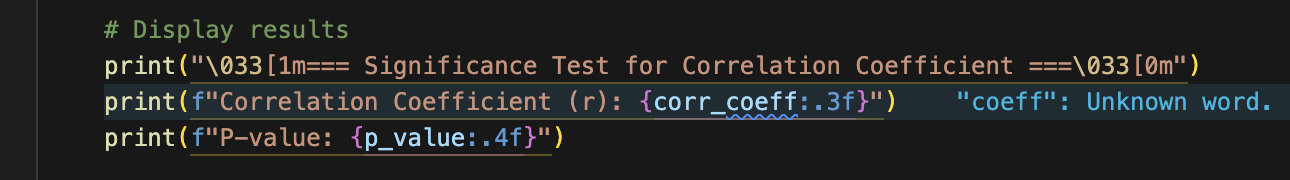
To evaluate whether a statistically significant linear relationship exists between the number of study hours and marks scored by performing a significance test on the Pearson correlation coefficient.

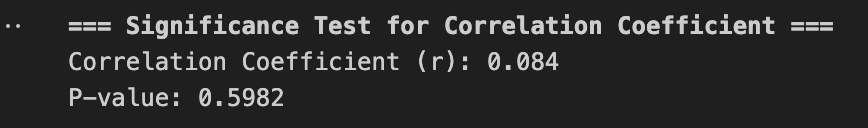
**Methodology:**

**We used the Pearson correlation coefficient to examine the linear relationship between Study Hours (per week) and Marks (out of 100) in the sample. The correlation coefficient rr and its associated p-value were calculated using the pearsonr() function from scipy.stats**.



**Results:**

****

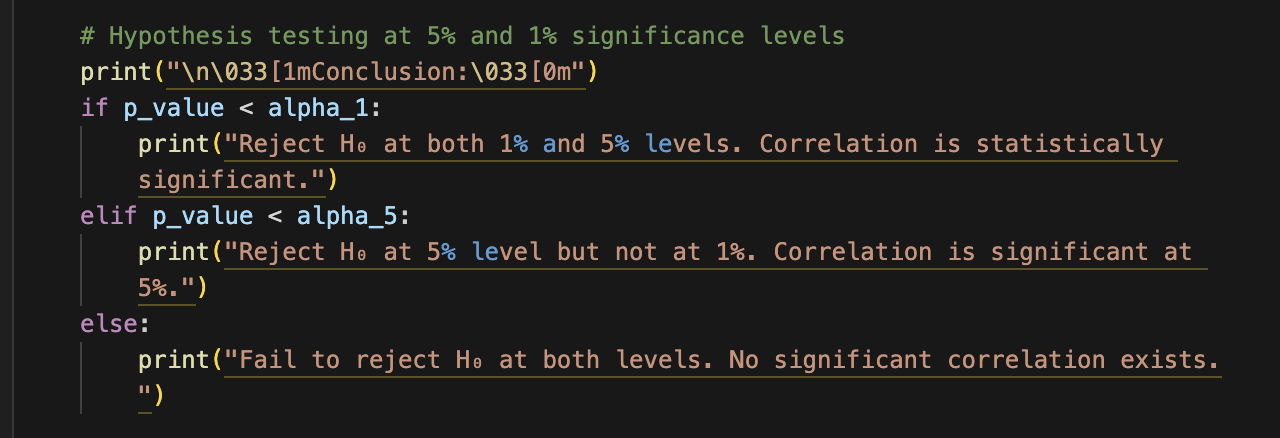
****

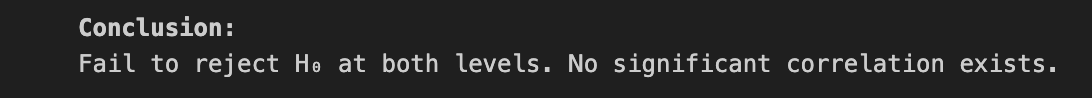
#### Hypothesis Testing:

* **Null Hypothesis (H₀):** There is no significant correlation between study hours and marks.
* **Alternative Hypothesis (H₁):** There is a significant correlation between study hours and marks.

**Decision criteria:**

* If p-value < 0.01 → Reject H₀ at both 1% and 5% significance levels
* If 0.01 ≤ p-value < 0.05 → Reject H₀ at 5% level only
* If p-value ≥ 0.05 → Fail to reject H₀

****

****

**13.2. Question 15 - Test the significance of the regression coefficient.**

**Objective:**

To check if the regression coefficient (slope) in the linear relationship between study hours and marks obtained is statistically significant. This test indicates whether study hours significantly predict the marks

**Methodology:**

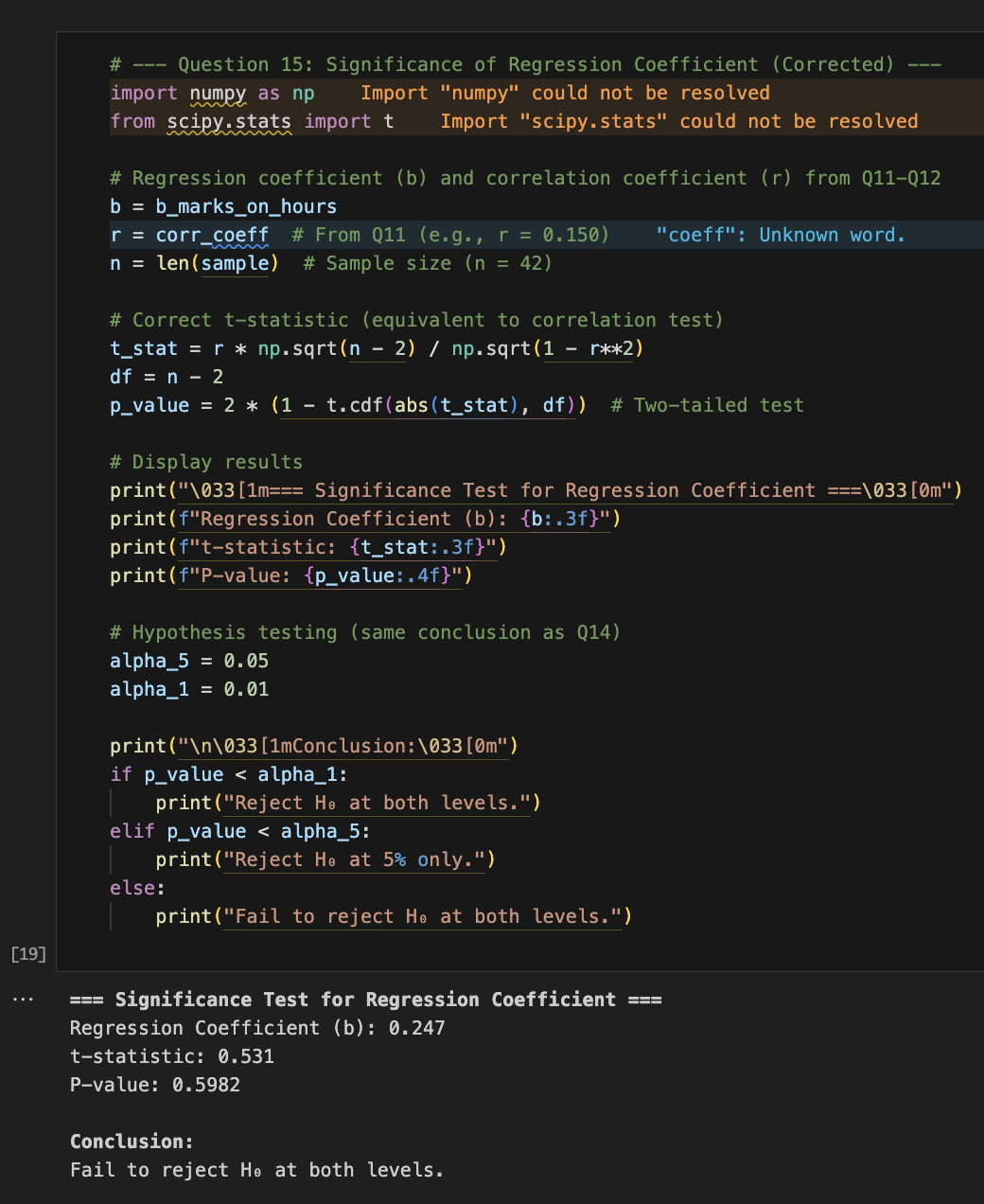
With the use of the correlation coefficient rr from the previous analysis (Question 14), the t-statistic to test for significance of the regression coefficient bb was computed. Mathematically, the test is the same as testing the correlation coefficient but is interpreted within linear regression.

#### Hypothesis Testing:

* **Null Hypothesis (H₀):** The regression coefficient b=0 (no linear relationship)
* **Alternative Hypothesis (H₁):** The regression coefficient b≠0

**Decision thresholds:**

* If p-value < 0.01 → Reject H₀ at both 1% and 5% levels
* If 0.01 ≤ p-value < 0.05 → Reject H₀ at 5% level only
* If p-value ≥ 0.05 → Fail to reject H₀



**14. REFERENCES**