**SQL**

Answer 1:

This query will run a command to show all the records of the runners who have not won any race

My alternate query will be:

SELECT r.\*

FROM runners r

LEFT JOIN races ra ON r.id = ra.winner\_id

WHERE ra.winner\_id IS NULL;

LEFT JOIN: This joins the runners table (r) with the races table (ra) on the condition r.id = ra.winner\_id. A LEFT JOIN ensures that all rows from the left table (runners) are included in the result, even if there's no matching winner in the races table.

Answer 2:

Here is the query for the solution:

SELECT \*

FROM test\_a

EXCEPT

SELECT \*

FROM test\_b;

Explanation: The EXCEPT operator essentially returns all rows from test\_a that are not present in test\_b.

This achieves the desired result of finding values that exist in test\_a but not in test\_b.

Answer 3:

Here is the Query for solution:

SELECT user\_id, training\_id, training\_date

FROM training\_details

GROUP BY user\_id, training\_id

HAVING COUNT(\*) > 1

ORDER BY user\_id, training\_date DESC;

**STATISTICS**

Answer 1:

Six Sigma: The term refers to six standard deviations from the mean in a normal distribution. Statistically, this translates to a very low defect rate (around 3.4 defects per million opportunities - DPMO).

Goal:The primary goal of Six Sigma is to achieve a high level of process capability, meaning a process that consistently produces a high proportion of defect-free outputs.

Six Sigma follows a structured approach called DMAIC, which stands for:

Define, Measure, Analyze, Improve, Control

Example:

Consider a manufacturing company producing light bulbs. They might use Six Sigma to:

Define: Reduce the number of defective light bulbs produced.

Measure: Track the current defect rate and identify the types of defects occurring.

Analyze: Investigate the possible causes of defects, such as faulty equipment, improper temperature control, or human error.

Improve: Implement solutions like equipment maintenance, process adjustments, or employee training.

Control: Continuously monitor the defect rate and make adjustments as needed to maintain a high level of quality.

Answer 2:

Here are some examples of

* Discrete Data with Limited Values
* Data with Natural Lower or Upper Bounds
* Highly Skewed Data
* Data with Multiple Peaks (Multimodal)
* Time Series Data with Trends or Seasonality

Example: Stock prices - Stock prices often exhibit trends (upward or downward) and seasonal fluctuations. These patterns violate the assumption of a constant mean and variance in a normal distribution, and log-normal might not capture the trends or seasonality either.

Answer 3:

The five-number summary is a set of descriptive statistics that provides a concise overview of the distribution of a dataset. It consists of five key values that capture important aspects of how the data is spread out:

1. Minimum: This is the smallest value in the dataset. It represents the lowest point in the data distribution.
2. First Quartile (Q1): This value represents the 25th percentile. It divides the data into four equal parts, with 25% of the data points falling below Q1.
3. Median: This is the middle value of the data set when ordered from least to greatest. It represents the 50th percentile, with half of the data points falling below the median and the other half falling above.
4. Third Quartile (Q3): This value represents the 75th percentile. It divides the data into four equal parts, with 75% of the data points falling below Q3.
5. Maximum: This is the largest value in the dataset. It represents the highest point in the data distribution.

Consider a dataset representing the exam scores of 50 students:

Scores: 45, 68, 72, 80, 85, 85, 88, 90, 92, 95,

48, 52, 55, 60, 62, 70, 75, 78, 82, 83,

90, 92, 94, 96, 98, 50, 51, 54, 59, 61,

69, 71, 74, 77, 79, 81, 84, 87, 89, 91,

93, 97, 99, 100

Calculating the five-number summary for this data:

* Minimum: 45
* First Quartile (Q1): 60 (25% of students scored lower than 60)
* Median: 82 (the middle score)
* Third Quartile (Q3): 90 (75% of students scored lower than 90)
* Maximum: 100

This five-number summary tells us that:

* The lowest score was 45 and the highest was 100.
* The distribution is somewhat skewed to the right, as the median (82) is closer to Q1 (60) than Q3 (90).
* There might be a few outliers, as the minimum and maximum scores are quite far from the rest of the data (depending on how you define outliers).

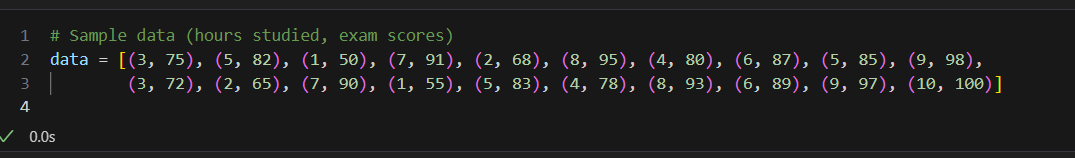
Answer 4:

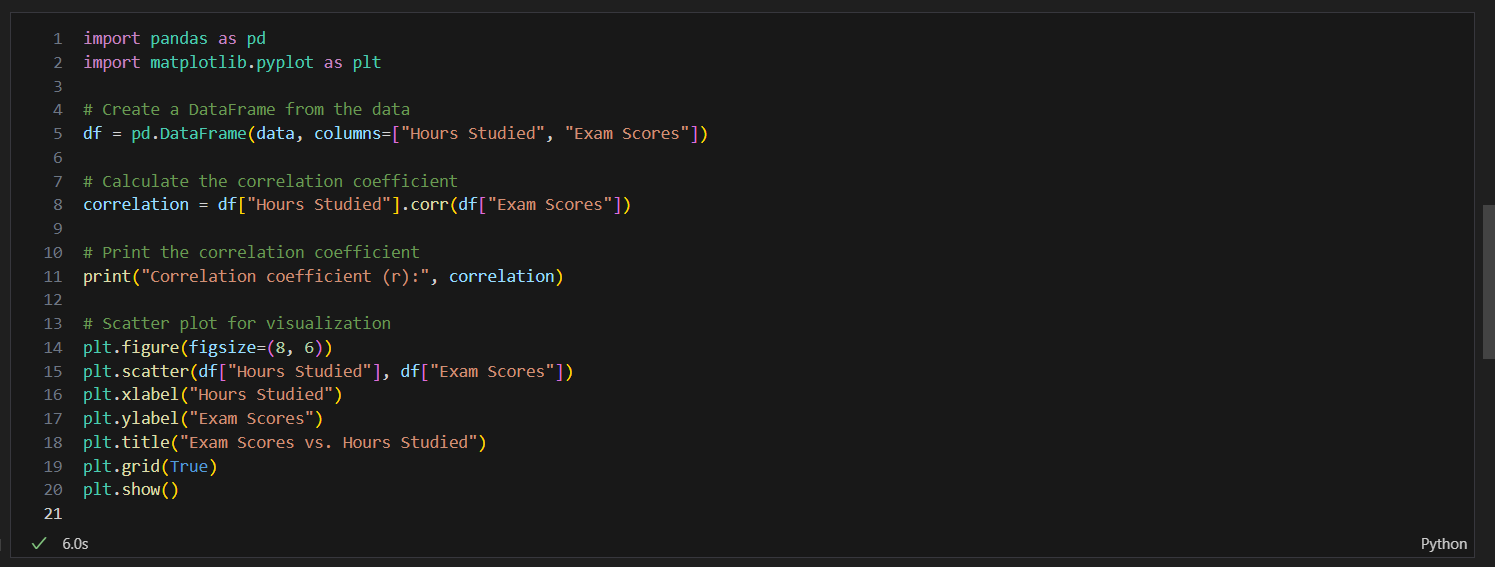
Here's a breakdown of correlation concepts:

* Strength: The strength of a correlation is indicated by a coefficient named the correlation coefficient, typically denoted by 'r'. The value of 'r' ranges from -1 to +1.
  + Positive correlation (0 < r < 1): As the value of one variable increases, the value of the other variable also tends to increase.
  + Negative correlation (-1 < r < 0): As the value of one variable increases, the value of the other variable tends to decrease.
  + Zero correlation (r = 0): There is no linear relationship between the two variables. Their changes are independent of each other.
* Direction: The direction of the correlation indicates whether the relationship is positive or negative.

Example: Dataset and Jupyter Notebook Code

Let's consider a dataset where we have exam hours studied (independent variable) and exam scores (dependent variable) for 20 students.





OUTPUT:

