ATTRIBUTES

- Data points or Samples are described by attributes.
- Attribute (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.
- Types
  - Nominal or Categorical
  - Ordinal
  - Binary
  - Numerical

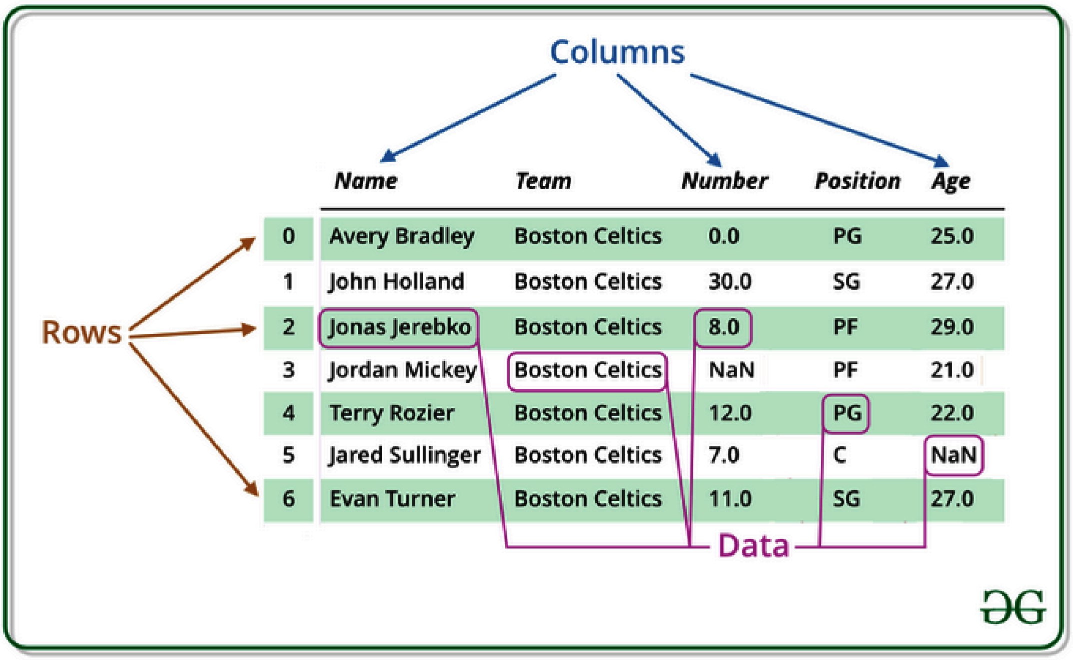
ATTRIBUTE TYPES

- Nominal: categories, states, or “names of things”
  - Hair color = {auburn, black, blond, brown, grey, red, white}
  - marital status, occupation, ID numbers, zip codes
- Ordinal: Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - Size = {small, medium, large}, grades, army rankings
- Binary: Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important, e.g., gender
  - Asymmetric binary: outcomes not equally important. e.g., medical test (positive vs. negative)
- Numeric: represents quantity (integer or real-valued)
  - Temperature, length, counts, grade point, CGPA, salary etc.

DISCRETE VS. CONTINUOUS ATTRIBUTES

- Discrete Attribute: has only a finite or countably infinite set of values
  - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- Continuous Attribute: has real numbers as attribute values
  - E.g., temperature, height, or weight
  - Continuous attributes are typically represented as floating-point variables

EXAMPLE: PANDAS DATAFRAME



	Name	Team	Number	Position	Age
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0

📊 Exploratory Data Analysis (EDA)

What is EDA?

Exploratory Data Analysis (EDA) is the process of analyzing and visualizing datasets to:

- Summarize key characteristics of the data
- Identify patterns and trends
- Detect anomalies and outliers
- Understand variable relationships
- Prepare the data for machine learning models

✔️ Why is EDA Important?

- Helps understand the **structure and distribution** of the data
- Detects **anomalies**, **outliers**, and **missing values**
- Identifies **relationships between variables** for feature selection
- Guides **feature engineering** and **model selection**
- Improves understanding of potential issues before model training

🔧 Common EDA Techniques

🔹 Visual Techniques

- **Histograms** – Understand distribution of numerical features
- **Box Plots** – Identify outliers
- **Bar Charts / Pie Charts** – Analyze categorical variables
- **Scatter Plots** – Explore relationships between variables
- **Heatmaps** – Show correlation between numerical variables

🔹 Statistical Techniques

- **Descriptive Statistics** – Mean, median, mode, standard deviation
- **Value Counts** – Frequency of unique values
- **Skewness & Kurtosis** – Distribution shape
- **Correlation Matrix** – Strength of linear relationships
- **Missing Value Analysis** – Locate null or NaN values

🔍 EDA vs. Data Preprocessing

Aspect	EDA (Exploratory Data Analysis)	Data Preprocessing
Purpose	Understand data patterns and relationships	Clean and prepare data for modeling
Techniques Used	Visualization, descriptive statistics, correlations	Handling missing values, scaling, encoding
Focus	Interpretation and discovery	Data cleaning and transformation
Outcome	Insights and hypotheses	A ready-to-use dataset for machine learning
Tools Commonly Used	`pandas`, `seaborn`, `matplotlib`	`pandas`, `sklearn.preprocessing`, `numpy`
When Performed	Before modeling, during exploration phase	Before model training, after EDA

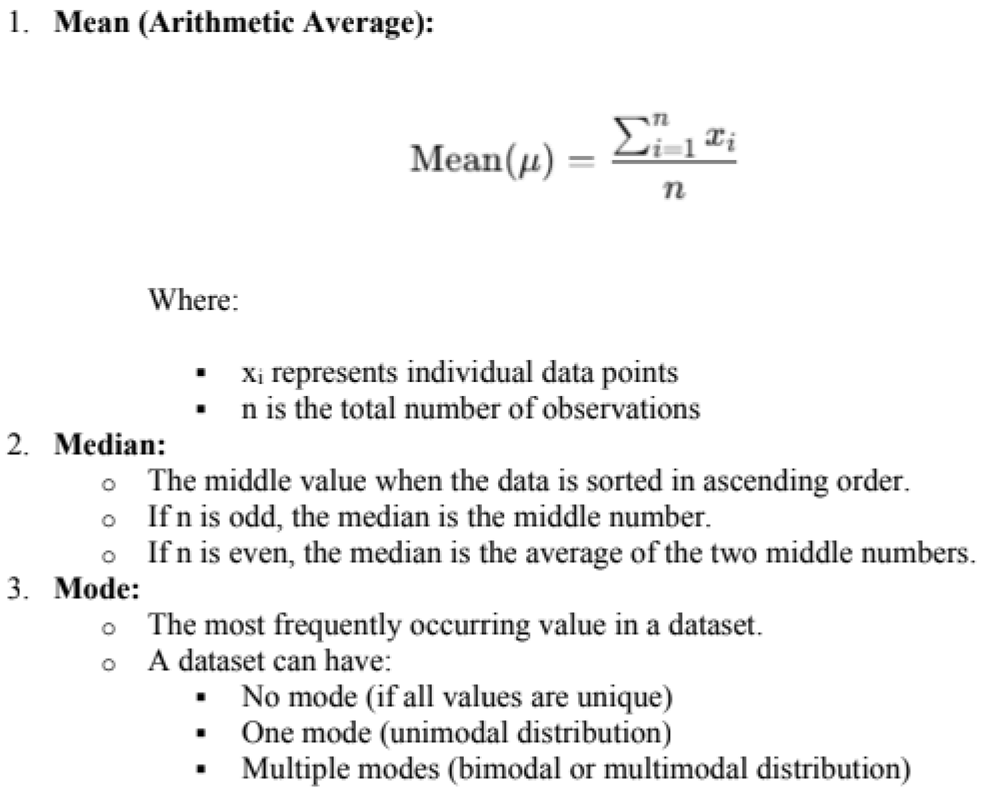
EDA and Data Preprocessing are **complementary steps** in the data science workflow:

- 🔹 **EDA** helps you understand **what the data is telling you**.
- 🔹 **Preprocessing** helps you **clean and shape** the data for models to understand it.

📊 Descriptive Statistics

📌 Measures of Central Tendency

Central tendency refers to values that represent the center or typical value of a dataset. The three main measures are:



1. **Mean (Arithmetic Average):**

$$\text{Mean}(\mu) = \frac{\sum_{i=1}^{n} x_i}{n}$$

Where:

   - x_i represents individual data points
   - n is the total number of observations

2. **Median:**
   - The middle value when the data is sorted in ascending order.
   - If n is odd, the median is the middle number.
   - If n is even, the median is the average of the two middle numbers.
3. **Mode:**
   - The most frequently occurring value in a dataset.
   - A dataset can have:
     - No mode (if all values are unique)
     - One mode (unimodal distribution)
     - Multiple modes (bimodal or multimodal distribution)

A dataset can have:

- **No mode** (if all values are unique)
- **One mode** → *unimodal distribution*
- **Two modes** → *bimodal distribution*
- **More than two modes** → *multimodal distribution*

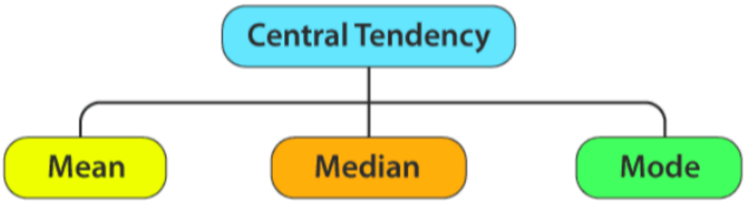
💡 Summary

Measure	Description	Sensitive to Outliers?
Mean	Arithmetic average	✔️ Yes
Median	Middle value	❌ No
Mode	Most frequent value	❌ No

CENTRAL TENDENCY

The central tendency is stated as the statistical measure that represents the single value of the entire distribution or a dataset.

These three values summarize the dataset using a single value.



2. Descriptive Statistics

Measures of Central Tendency

1. **Mean (Arithmetic Average):**

$$\text{Mean}(\mu) = \frac{\sum_{i=1}^{n} x_i}{n}$$

   Where:

   - x_i represents individual data points
   - n is the total number of observations

2. **Median:**
   - The middle value when the data is sorted in ascending order.
   - If n is odd, the median is the middle number.
   - If n is even, the median is the average of the two middle numbers.
3. **Mode:**
   - The most frequently occurring value in a dataset.
   - A dataset can have:
     - No mode (if all values are unique)
     - One mode (unimodal distribution)
     - Multiple modes (bimodal or multimodal distribution)

MEAN

The sum of all values divided by the number of values.

$$\text{Mean } = \overline{x} = \frac{\sum_{i}^{n} x_i}{n}$$

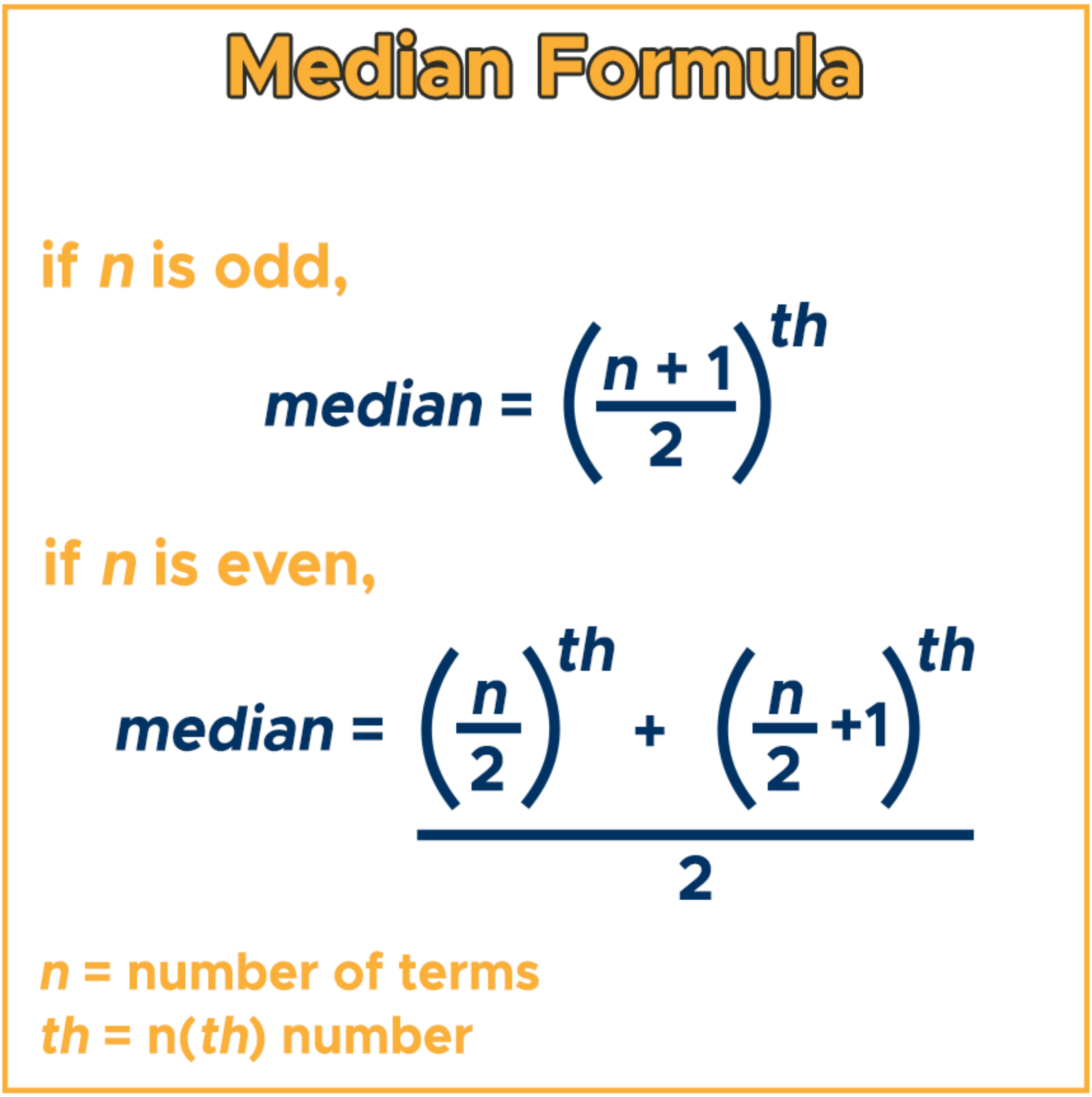
- Mean is influenced by extreme values (Outliers).
- An outlier is a value or an element of a dataset that shows higher deviation from the rest of the values.

TRIMMED MEAN

The average of all values after dropping a fixed number of extreme values from both ends.

$$\text{Trimmed mean } = \overline{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n-2p}$$

Preferable to use instead of ordinary mean as it can negate the effect of extreme values (Outliers).



Median Formula

if *n* is odd,

$$median = \left(\frac{n+1}{2}\right)^{th}$$

if *n* is even,

$$median = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$$

n = number of terms

th = n(*th*) number

PRACTICE PROBLEM - 1

A group of 10 students appeared in a test. Their obtained marks are given below.

60, 70, 80, 75, 65, 70, 80, 70, 65, 65

Sorted Order

60 65 65 65 70 70 70 75 80 80

Median = 70

Mean = 70

PRACTICE PROBLEM - 2

A group of 10 students appeared in a test. Their obtained marks are given below.

60, 70, 80, 75, 65, 70, 100, 70, 65, 65

Mean = 720/10 = 72
Without considering the 100, mean = 620/9 = 68.89

For Trimmed mean,
Sorting the dataset: ~~**60**~~ 65 65 65 70 70 70 75 80 ~~**100**~~
For p = 1 → 60 and 100 will be discarded.
Then trimmed mean = 70

Median = 70

PRACTICE PROBLEM - 3

A group of 10 students appeared in a test. Their obtained marks are given below.

60, 70, 80, 75, 65, 70, 10, 70, 65, 65

Mean = 630/10 = 63
Without considering the 10, mean = 620/9 = 68.89

For Trimmed mean,
Sorting the dataset: ~~**10**~~ 60 65 65 65 70 70 70 75 **80**
For p = 1 → 10 and 80 will be discarded.
Then trimmed mean = 67.5

Median = 67.5

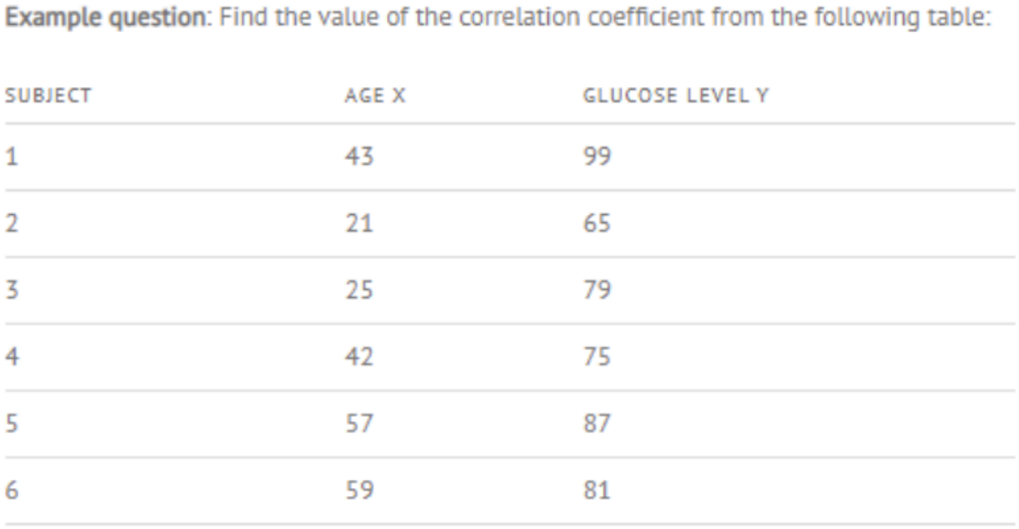
WEIGHTED MEAN

It is calculated by multiplying each data value x_i by a weight w_i and dividing their sum by the sum of the weights (w_i).

$$\text{Weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i}^{n} w_i}$$

Some values are intrinsically more variable than others, and highly variable observations are given a lower weight.

EXAMPLE



Example question: Find the value of the correlation coefficient from the following table:

SUBJECT	AGE X	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Measures of Dispersion

1. **Variance:**

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

   - Measures how much the data points deviate from the mean.

2. **Standard Deviation:**

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

   - Represents the spread of data points around the mean.

3. **Interquartile Range (IQR):**
   - Formula: IQR = Q3 – Q1
   - Where:
     - Q1 (First Quartile) is the 25th percentile.
     - Q3 (Third Quartile) is the 75th percentile.
   - IQR helps identify outliers in a dataset.

Skewness and Kurtosis

- **Skewness:** Measures the asymmetry of the data distribution.
  - A skewness of 0 indicates a perfectly symmetrical distribution.
  - Positive skew: Tail on the right (mean > median > mode).
  - Negative skew: Tail on the left (mean < median < mode).
- **Kurtosis:** Measures the tail-heaviness of the distribution.
  - High kurtosis: More outliers (heavy tails).
  - Low kurtosis: Fewer outliers (light tails).

3. Data Visualization

Univariate Analysis

- Used to analyze individual variables.
- Common plots:
  - **Histograms:** Show frequency distribution.
  - **Boxplots:** Detect outliers and spread.
  - **Density Plots:** Show probability density.

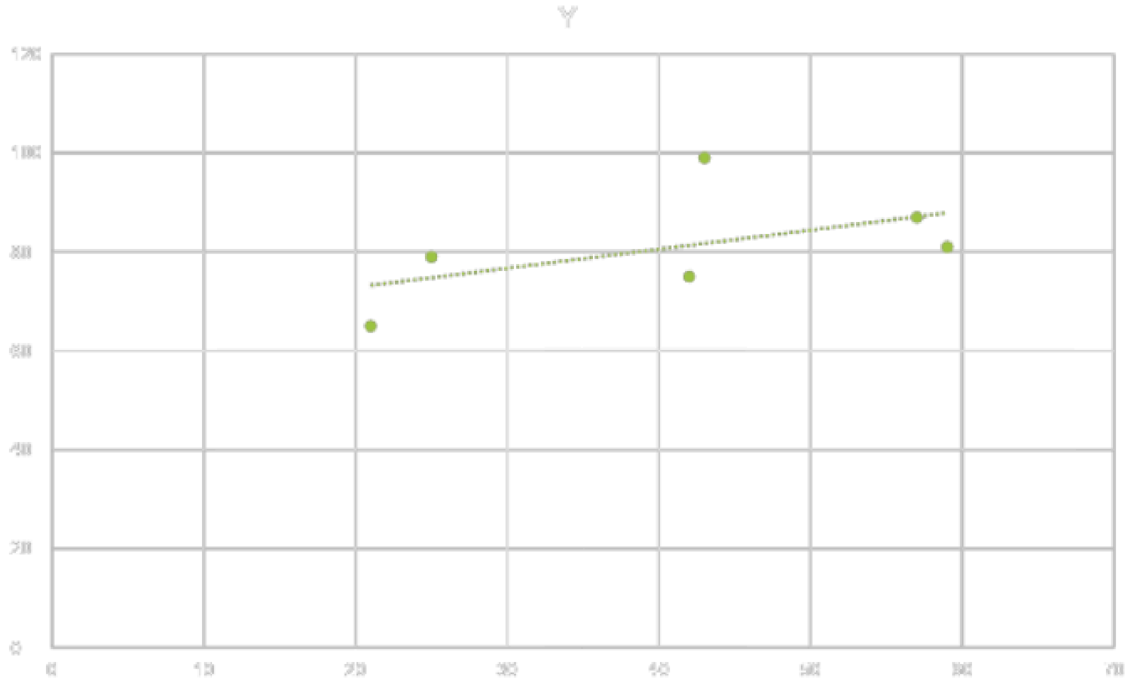
Bivariate and Multivariate Analysis

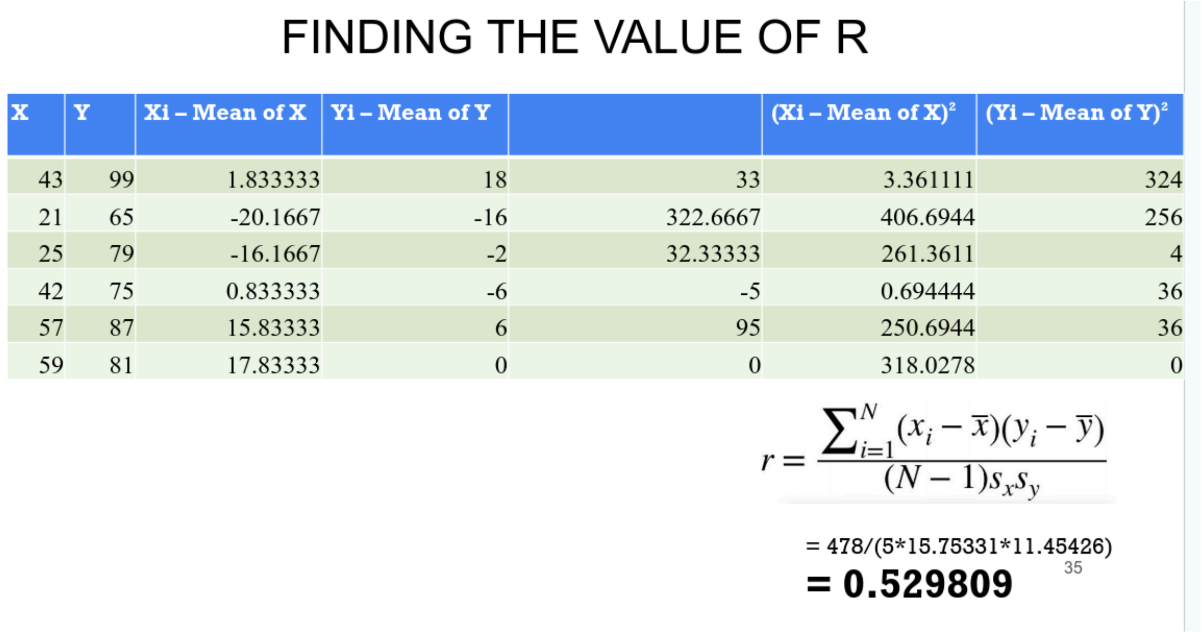
- **Scatter Plots:** Show relationships between two variables.
- **Pair Plots:** Visualize relationships across multiple numerical features.
- **Heatmaps:** Display correlations between numerical features.

Correlation and Relationships

- Pearson’s Correlation Coefficient measures the linear relationship between two variables.
- Spearman’s Rank Correlation is used for non-linear relationships.
- A correlation heatmap visually represents the strength of variable relationships.

SCATTER PLOT ON DATASET





FINDING THE VALUE OF R

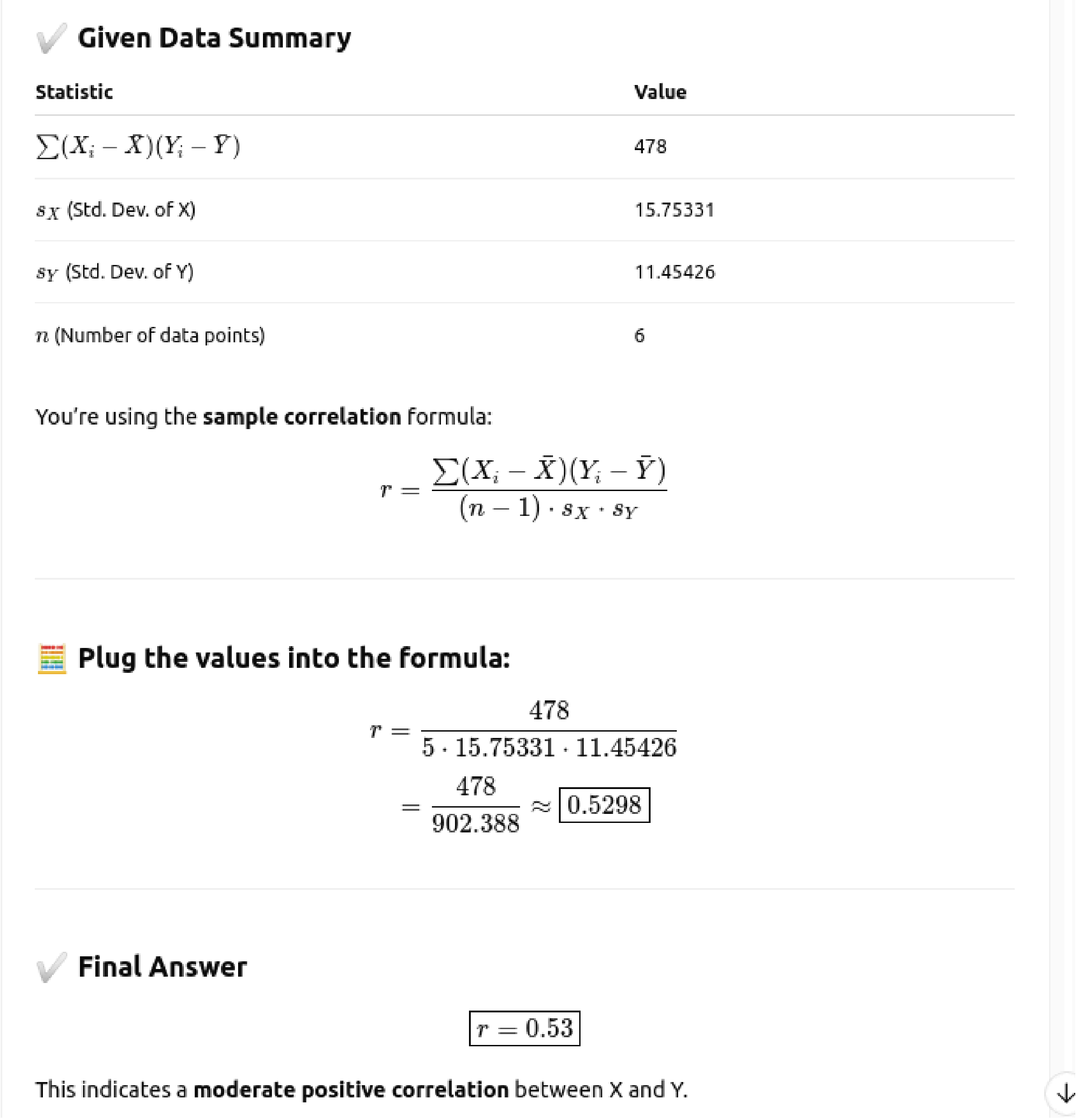
X	Y	Xi – Mean of X	Yi – Mean of Y		(Xi – Mean of X)²	(Yi – Mean of Y)²
43	99	1.833333	18	33	3.361111	324
21	65	-20.1667	-16	322.6667	406.6944	256
25	79	-16.1667	-2	32.33333	261.3611	4
42	75	0.833333	-6	-5	0.694444	36
57	87	15.83333	6	95	250.6944	36
59	81	17.83333	0	0	318.0278	0

$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{(N-1)s_x s_y}$$

= 478/(5*15.75331*11.45426)

35

= 0.529809



✓ Given Data Summary

Statistic	Value
$\sum(X_i - \bar{X})(Y_i - \bar{Y})$	478
s_X (Std. Dev. of X)	15.75331
s_Y (Std. Dev. of Y)	11.45426
n (Number of data points)	6

You're using the **sample correlation** formula:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)\cdot s_X \cdot s_Y}$$

🧮 Plug the values into the formula:

$$r = \frac{478}{5 \cdot 15.75331 \cdot 11.45426}$$
$$= \frac{478}{902.388} \approx \boxed{0.5298}$$

✓ Final Answer

$$\boxed{r = 0.53}$$

This indicates a **moderate positive correlation** between X and Y.