Data cleaning is a critical step in preparing datasets for analysis or decision-making. It ensures data accuracy, consistency, and reliability. This article provides a step-by-step guide to handling missing values using practical techniques and Excel functions.

Step 1: Check Data Types

The first step in cleaning data is to ensure that the data types for all features are consistent.

1. **Why Check Data Types?**
    - Inconsistent data types can lead to errors in analysis or visualizations.
    - For example, creating a chart based on dates will fail if the date column is stored as text.
2. **How to Check Data Types in Excel?**
    - Use the TYPE function in Excel to identify the type of data in a cell. This function returns:
        - 1 for numeric
        - 2 for text
        - 4 for boolean
        - 16 for errors
        - 64 for arrays
    - Example: If you have a column for Employee IDs, ensure all IDs are stored as text, even if they appear numeric (e.g., "12345").
3. **Steps to Fix Inconsistent Data Types:**
    - Identify inconsistent data types using the TYPE function.
    - Convert the data to the correct type (e.g., use Excel functions such as TEXT or VALUE).

Step 2: Handle Missing Values

Handling missing values is crucial for reliable analysis. The impact of missing values includes inaccurate results, incomplete datasets, and unreliable decision-making.

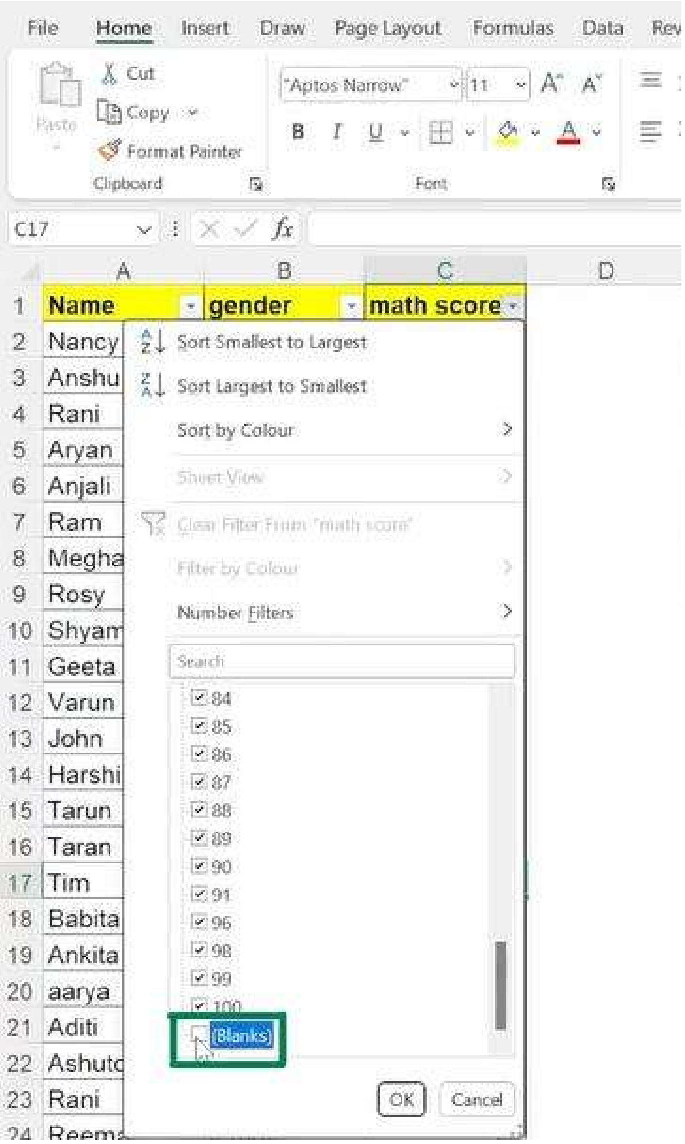
Impact of Missing Values

- Missing values can significantly affect:
    - **Accuracy**: Results derived from incomplete data might not be reliable.
    - **Completeness**: Missing values reduce the dataset's completeness.
    - **Decision-Making**: Incomplete data leads to suboptimal or incorrect decisions.
- Example: A dataset with 50% missing data would result in unreliable conclusions

Scenarios for Handling Missing Values

1. **Scenario 1: Large Dataset, Few Missing Values**

- If the percentage of missing values is low (e.g., <5%), use the filter option in Excel to remove rows with missing values.
- **Steps**:
  - Select the dataset. (You can use the following dataset)
  - Apply a filter.
  - Uncheck the blank option in the filter dropdown.
  - Analyze the cleaned data.



	A	B	C	D
1	Name	gender	math score	
2	Nancy			
3	Anshu			
4	Rani			
5	Aryan			
6	Anjali			
7	Ram			
8	Megha			
9	Rosy			
10	Shyam			
11	Geeta			
12	Varun			
13	John			
14	Harshi			
15	Tarun			
16	Taran			
17	Tim			
18	Babita			
19	Ankita			
20	aarya			
21	Aditi			
22	Ashutc			
23	Rani			
24	Reema			

☑ 84
☑ 85
☑ 86
☑ 87
☑ 88
☑ 89
☑ 90
☑ 91
☑ 96
☑ 98
☑ 99
☑ 100
□ (Blanks)

2. **Scenario 2: Limited Dataset, High Percentage of Missing Values**
   - If the dataset is small and contains a significant percentage of missing values, replace missing values using statistical techniques (mean, median, or mode).
   - **Steps**:

        - For numeric data without outliers, use the AVERAGE function to calculate the mean.
        - For numeric data with outliers, use the MEDIAN function.
        - For categorical data, use the most frequent value (mode).

Example:

    - Dataset: Scores of 5 students – 15, 16, 11, 14, 19.
    - Mean: (15+16+11+14+19) / 5 = 15}
    - Median: Sort the data [11,14,15,16,19][11, 14, 15, 16, 19][11,14,15,16,19]. Middle value = 15.
    - Mode: If the dataset contains categorical values like gender, use the most frequent category (e.g., "Female").

3. **Scenario 3: Preparing Data for Machine Learning**
    - Models cannot process missing values, so all missing data must be handled.
    - Use imputation methods (mean, median, mode) or advanced techniques such as predictive imputation.

Step 3: Use Excel Functions for Missing Values

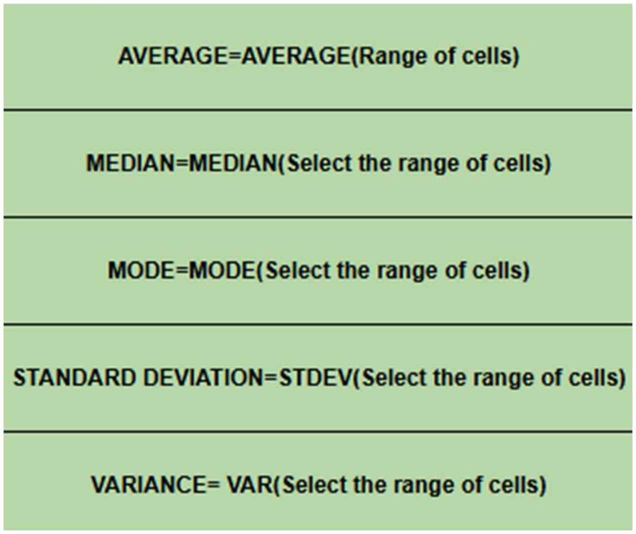
Excel provides powerful functions to handle missing values efficiently.

Functions in Excel for Handling Missing Values

- **AVERAGE**: Calculates the mean of a dataset.
- **MEDIAN**: Finds the middle value in a sorted dataset.
- **COUNTBLANK**: Counts the number of blank cells in a range.
- **COUNT**: Counts the number of non-blank cells.
- Example: If there are 100 entries and 15 are missing, COUNTBLANK will return 15, helping calculate the missing data percentage (15% in this case).

1. **COUNTBLANK Function**:
    - Identifies the number of blank cells in a range.
    - Example: =COUNTBLANK(A1:A100) counts all blank cells in the range.
2. **COUNT Function**:
    - Counts non-blank cells.
    - Example: If the total rows are 100 and COUNT(A1:A100) returns 85, the number of missing values is 100−85=15100 - 85 = 15100−85=15.

3. **AVERAGE, MEDIAN, and MODE Functions**:
    - Use these functions to calculate replacement values for missing data.
    - Example:
        - To calculate the mean of scores: =AVERAGE(A1:A5).
        - To calculate the median: =MEDIAN(A1:A5).



AVERAGE=AVERAGE(Range of cells)
MEDIAN=MEDIAN(Select the range of cells)
MODE=MODE(Select the range of cells)
STANDARD DEVIATION=STDEV(Select the range of cells)
VARIANCE= VAR(Select the range of cells)

Outliers are extreme values that significantly deviate from the rest of the data. They can skew analysis, lead to incorrect conclusions, and must be addressed efficiently. This article will provide a step-by-step guide on identifying and handling outliers, using the transcript and provided resources.

Step 1: Understanding Measures of Dispersion

Outliers are detected using statistical measures that quantify the spread of data. Below are the key measures:

1. **Standard Deviation (SD)**:
    - Measures the spread of data around the mean.
    - Formula:

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

    - Variance (σ^2) is the square of the standard deviation.
2. **Normal Distribution and Empirical Rule**:
    - A normal distribution is a bell-shaped curve where:
    - **Mean (μ)**, **Median**, and **Mode** are equal and located at the center.

   - The data is symmetrically distributed around the mean.

   - 50% of the data lies to the left of the mean, and 50% lies to the right.
   - Empirical Rule defines how data is distributed in a normal distribution:
      - 68% of data lies within $(\mu-\sigma,\mu+\sigma)$.
      - 95% of data lies within $(\mu-2\sigma,\mu+2\sigma)$.
      - 99.7% of data lies within $(\mu-3\sigma,\mu+3\sigma)$.
   - Values outside $(\mu-3\sigma,\mu+3\sigma)$ are considered potential outliers.
3. **Range**:
   - Difference between maximum and minimum values.
   - **Not robust to outliers**, as it depends on extreme values.
4. **Interquartile Range (IQR)**:
   - Difference between the 75th percentile (Q3) and the 25th percentile (Q1): $IQR=Q3-Q1$
   - Values outside:
      - $Q1-1.5\times IQR$
      - $Q3+1.5\times IQR$ are outliers.

Step 2: Detecting Outliers in Excel

Dataset Overview

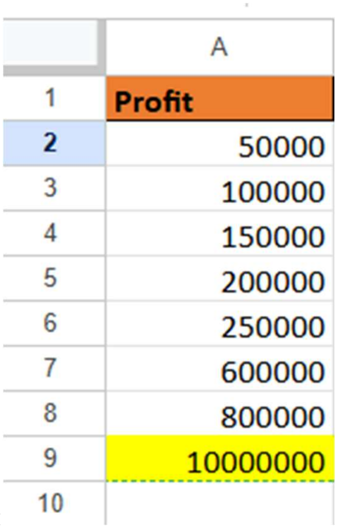
- The dataset contains profit values recorded over time, as follows:: ₹50,000, ₹1,00,000, ₹1,50,000, ₹2,00,000, ₹2,50,000, ₹6,00,000, ₹8,00,000, ₹1 crore.
- The aim is to analyze this data, detect outliers, and visualize it using Excel.

Step 2.1: Sort the Data

Sorting is essential to calculate quartiles correctly. Follow these steps:

1. Select the dataset (Column A, from A2:A9).
2. Go to the Data tab in Excel.
3. Click Sort → Select "Smallest to Largest".

  - After sorting, the dataset becomes:



	A
1	**Profit**
2	50000
3	100000
4	150000
5	200000
6	250000
7	600000
8	800000
9	10000000
10	

Step 2.2: Compute Quartiles

Use Excel's QUARTILE.EXC function to calculate quartiles (Q1, Q2, Q3):

1. **Quartile 1 (Q1) or 25th Percentile**:

    - Formula:

=QUARTILE(A2:A9, 1)

    - Result: **₹1,37,500**.

2. **Quartile 2 (Median or Q2) or 50th Percentile**:

- Formula:

=QUARTILE(A2:A9, 2)

- Result: **₹2,25,000**.

1. **Quartile 3 (Q3) or 75th Percentile**:

- Formula:

=QUARTILE(A2:A9, 3)

- Result: **₹6,50,000**.

Step 2.3: Calculate IQR

The Interquartile Range (IQR) is the difference between Q3 and Q1:

- Formula:

=Q3-Q1

Substitute the cell references where Q3 and Q1 are calculated:

=650000-1,37,500

- Result: **₹5,12,500**.

Step 2.4: Determine Upper and Lower Bounds

Using the IQR, calculate the bounds to identify outliers:

1. **Upper Bound**:
   - Formula:

=Q3 + 1.5 * IQR

Substitute:

=650000 + 1.5 * 512500

   - Result: **₹14,18,750**

2. **Lower Bound**:
   - Formula:

=Q1 - 1.5 * IQR

Substitute:

=137500 - 1.5 * 512500

   - Result: **₹-6,31,250** (negative bounds indicate there are no lower outliers).

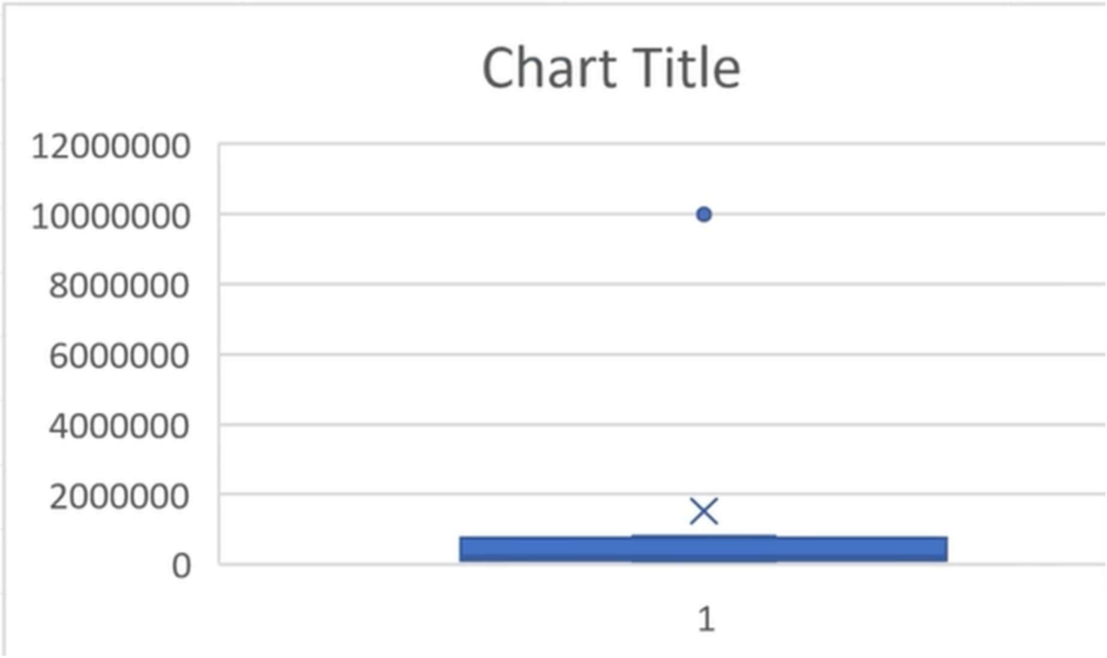
Step 2.5: Identify Outliers

Any value **above the Upper Bound** (₹14,18,750) or **below the Lower Bound** (₹-6,31,250) is an outlier.

1. Compare the dataset with these bounds:
   - ₹1 crore (**₹10,000,000**) exceeds the Upper Bound of ₹14,18,750 and is therefore an **outlier**.

Step 3: Visualize Using a Box Plot

1. Highlight the dataset range (A2:A9).
2. Go to Insert → Charts → Box and Whisker.
3. Excel will generate a box plot:
   - The **box** represents the IQR.
   - The **whiskers** extend to the minimum and maximum values within the bounds.
   - The **outlier** (₹1 crore) is displayed as a dot outside the whiskers.



In this article, we'll focus on **Univariate Analysis**, a fundamental method to analyze a single variable in a dataset. It covers the theoretical aspects, practical implementation in Excel, and visualization techniques. Both continuous and categorical variables are addressed with detailed steps.

What is Univariate Analysis?

- **"Uni" means one** and **"variate" refers to variable**.
- Univariate analysis deals with the analysis of a single variable at a time.
- It involves summarizing the data using statistical measures or visualizing its distribution.

Types of Univariate Analysis:

1. **Continuous Univariate Analysis**:
   - Analyzing numerical data (e.g., Age, Salary).
2. **Categorical Univariate Analysis**:
   - Analyzing categorical data (e.g., Gender, Marital Status).

Continuous Univariate Analysis

Statistical Measures

Continuous variables are analyzed using **Descriptive Statistics**. Key measures include:

1. Measures of Central Tendency:

- **Mean**: Average of all data points.
- **Median**: Middle value when data is sorted.
- **Mode**: Most frequently occurring value.

Excel Formulae

- Mean: =AVERAGE(range)
- Median: =MEDIAN(range)
- Mode: =MODE(range)

Example: Given an **Age** column:

- Mean: =AVERAGE(A2:A41) results in **55.19**.
- Median: =MEDIAN(A2:A41) provides the middle age.
- Mode: Use =MODE(A2:A41) for the most common age.

2. Measures of Dispersion:

- **Range**: Difference between the maximum and minimum values.
  Formula: =MAX(range) - MIN(range)
- **Variance**: Spread of data points from the mean.
  Formula: =VAR.P(range)
- **Standard Deviation**:
  Formula: =STDEV.P(range)
- **Interquartile Range (IQR)**:
  Formula: IQR=Q3−Q1
  Use Excel: =QUARTILE.EXC(range, 3) - QUARTILE.EXC(range, 1)

3. Trimmed Mean:

- Removes extreme values to calculate a robust mean.
- Formula: =TRIMMEAN(range, percentage)

Visualization Techniques

Continuous variables can be visualized using:

1. **Histogram**:
   - Provides a frequency distribution of data.
   - Excel Steps:
     - Highlight data, go to Insert → Histogram Chart.
2. **Box Plot**:
   - Shows median, quartiles, and potential outliers.
   - Excel Steps:
     - Highlight data, go to Insert → Box and Whisker Chart.
3. **Column Chart**:
   - Displays data frequency in custom bins.

Example:

  - Define custom bins (e.g., Age ranges: 0–10, 11–20).
  - Use the FREQUENCY function to calculate values for each bin.
  - Plot using a column chart for easy interpretation.

Example Dataset Implementation

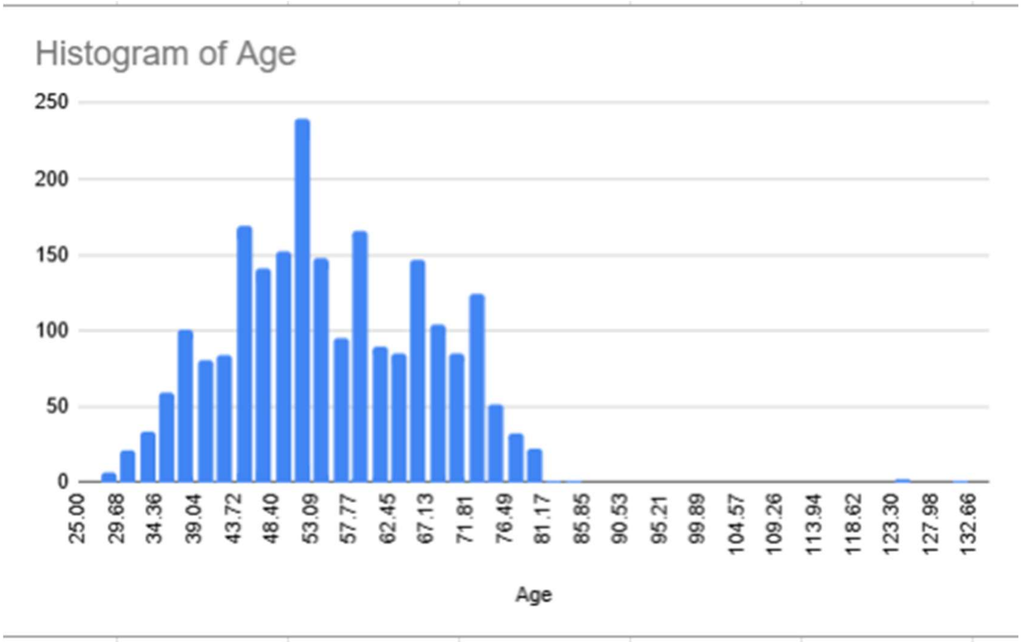
Dataset Overview

The dataset contains:

- **Age** (Continuous Variable): 28, 35, 40, 50, etc.
- **Marital Status** (Categorical Variable): Single, Married, Divorced.

Continuous Variable: Age

1. **Statistical Analysis**:
   - Mean: =AVERAGE(Age Range) → 55.19.
   - Median: =MEDIAN(Age Range) → Middle value.
   - Standard Deviation: =STDEV.P(Age Range) → Spread of ages.
2. **Visualization**:
   - **Histogram**:
     - Excel Steps:
       - Insert → Histogram Chart.



   - **Box Plot**:
     - Identifies outliers and shows quartile ranges.

Categorical Univariate Analysis

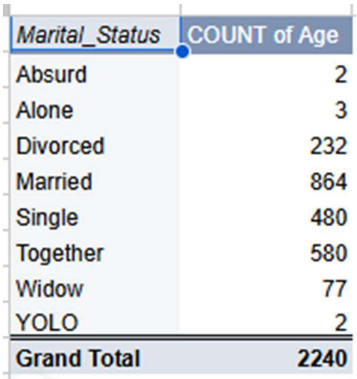
Statistical Measures

Categorical data, such as **Marital Status**, is analyzed using frequency distributions:

- Count how often each category appears.

Example: Given a column **Marital Status**:

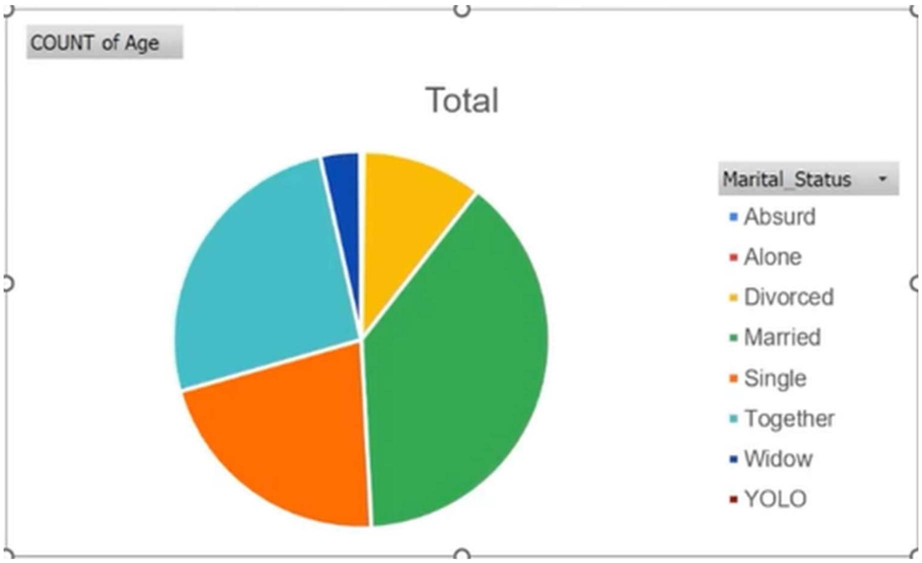
- Categories: Single, Married, Divorced.
- Create a **Pivot Table**:
    - Insert → PivotTable → Select "Marital Status" as rows and "Count of Age" as values.
    - Results:
        - Married: 864 employees.
        - Single: 480 employees.
        - Divorced: 232 employees.



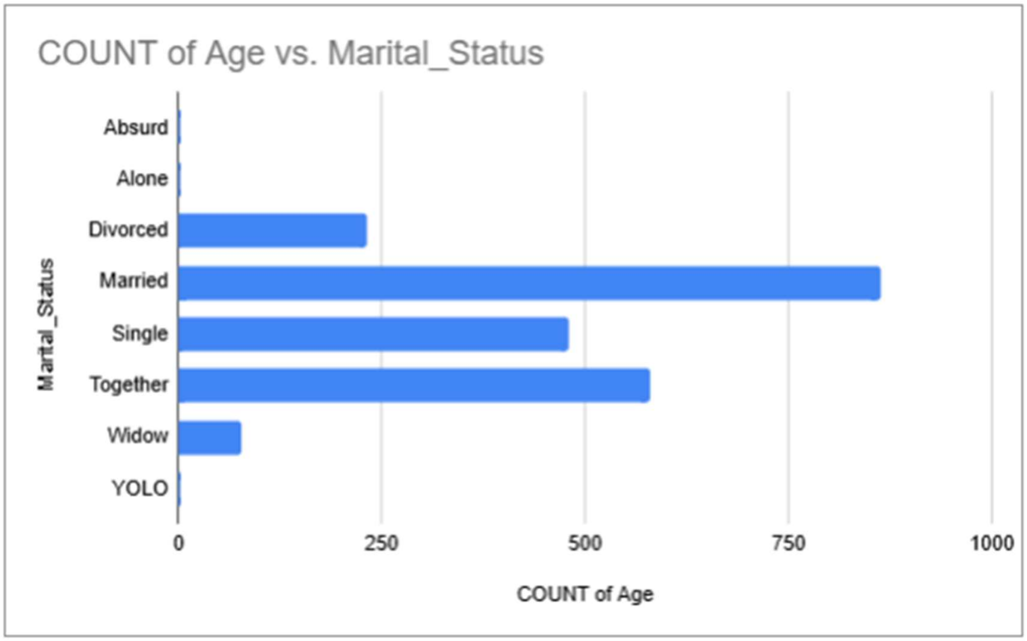
Marital_Status	COUNT of Age
Absurd	2
Alone	3
Divorced	232
Married	864
Single	480
Together	580
Widow	77
YOLO	2
Grand Total	**2240**

Visualization Techniques

1. **Pie Chart**:
    - Displays proportions of each category.
    - Excel Steps:
        - Highlight the Pivot Table, go to Insert → Pie Chart.



2. **Bar Chart**:
    - Compares frequencies across categories.
    - Excel Steps:
        - Highlight the Pivot Table, go to Insert → Bar Chart.



3. **Frequency Distribution Table**:
    - Provides a tabular summary of category counts.
    - Use Pivot Table for ease.

Data analysis is pivotal in uncovering patterns, relationships, and insights from datasets. Among the key types of analysis are **bivariate analysis**, which examines the relationship between two variables, and **multivariate analysis**, which extends this concept to multiple variables. This article will delve into these types of analysis, illustrating their applications with examples, visualization techniques, and practical implementation using the provided dataset.

Bivariate Analysis

What Is Bivariate Analysis?

Bivariate analysis focuses on understanding the relationship between two variables. It helps:

1. **Identify Relationships**: Determine whether a correlation exists between variables.
2. **Detect Patterns**: Discover trends, such as positive or negative correlations.
3. **Diagnostic Purpose**: Uncover unexpected relationships or outliers.

Types of Bivariate Data

1. **Continuous & Continuous**: Both variables are numerical, e.g., age vs. income.
2. **Categorical & Continuous**: One variable is categorical, and the other is numerical, e.g., age bracket vs. income.
3. **Categorical & Categorical**: Both variables are categorical, e.g., marital status vs. gender.

1. Continuous & Continuous

Analysis and Visualization

When both variables are numerical, two methods are commonly used:

1. **Correlation Coefficient**: Measures the strength and direction of the linear relationship.
   - A correlation value ranges between -1 and +1.
   - Magnitude interpretation:
     - Very Weak: 0.00–0.19
     - Weak: 0.20–0.39
     - Moderate: 0.40–0.59
     - Strong: 0.60–0.79
     - Very Strong: 0.80–1.00
   - Positive values indicate direct proportionality, while negative values imply inverse proportionality.
2. **Scatter Plot**: Visualizes the relationship between two variables.

Positive correlation occurs when an increase in one variable is associated with an increase in another variable. This means the two variables move in the same direction.

Example: Study Time vs. Exam Scores

- In the provided transcript, an example was given where **study time** (Variable X) and **exam scores** (Variable Y) are analyzed.
- The correlation shows that students who spend more time studying tend to achieve higher scores.

- This positive relationship can be summarized as:
    - More study time → Higher scores.
    - Less study time → Lower scores.

Negative correlation occurs when an increase in one variable is associated with a decrease in another variable. This means the two variables move in opposite directions.

Example: Time Spent Watching Reels vs. Exam Scores

- Another example in the transcript discusses **time spent watching reels on social media** (Variable X) and **exam scores** (Variable Y).
- The analysis shows that as students spend more time watching reels, their exam scores tend to decrease.
- This negative relationship can be summarized as:
    - More time on reels → Lower scores.
    - Less time on reels → Higher scores.

Interpreting the Correlation Coefficient

1. A **positive value (e.g., 0.6)** indicates a direct relationship, as seen in the Study Time vs. Exam Scores example.
2. A **negative value (e.g., -0.4)** indicates an inverse relationship, as seen in the Reels Watching vs. Exam Scores example.
3. A **value close to 0** indicates no correlation, meaning the variables are not related.

Implementation in Excel

Using the provided dataset:

- **Steps**:
    1. Select the Age and Income columns.
    2. Use Excel's **=CORREL(array1, array2)** to find the relationship between numerical variables.
    3. Generate a scatter plot to visualize the relationship.
- **Insights**:
    1. If the correlation coefficient is positive, older individuals might earn more.
    2. The scatter plot will reveal the trend (e.g., linear or non-linear).

2. Categorical & Continuous

Analysis and Visualization

For a combination of categorical and numerical variables:

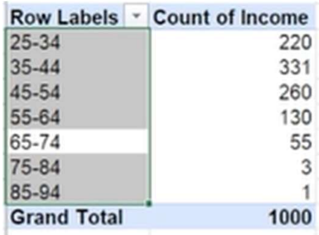
1. **Box Plot**: Displays the distribution of numerical data across categories.

2. **Bar Chart**: Compares numerical aggregates (e.g., average income) across categories.

Example: Income Across Age Brackets

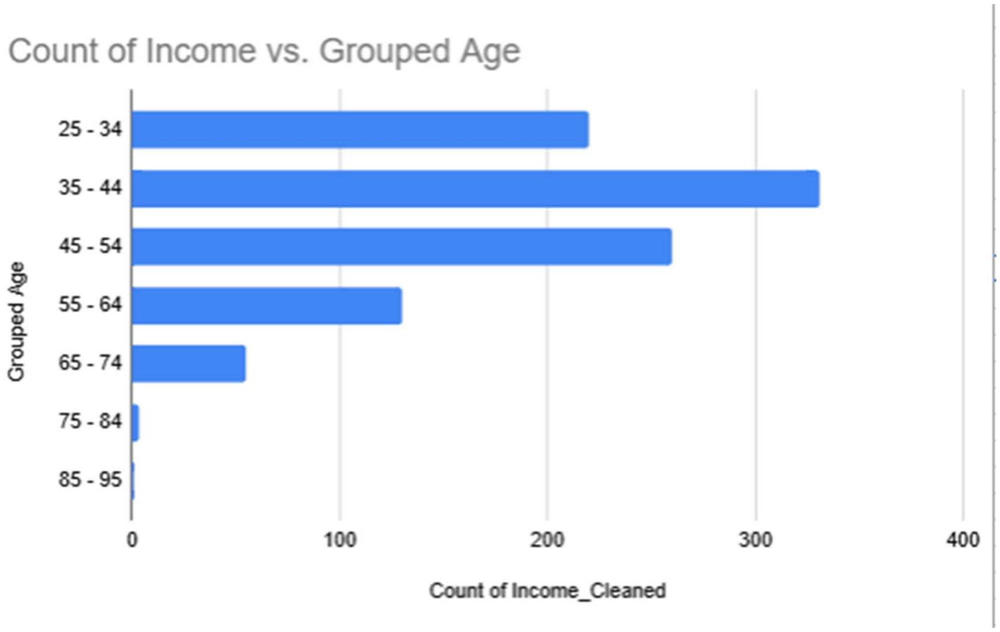
Using the dataset:

- **Steps**:
    1. Group ages into brackets (e.g., 25–34, 35–44).
    1. In Excel, use the Group function in a pivot table.
    2. In the pivot table:
        - Right-click on any age value in the **Rows** area.
        - Select **Group** from the context menu.
    3. In the **Grouping** dialog box:
        - Set the **Starting At** and **Ending At** values based on your data. For example:
            - Starting At: 25
            - Ending At: 95
        - Set **By** to 10 (this creates age brackets of 10 years: 25–34, 35–44, etc.).
    4. The Age column will now be grouped into age brackets.



Row Labels	Count of Income
25-34	220
35-44	331
45-54	260
55-64	130
65-74	55
75-84	3
85-94	1
Grand Total	**1000**

1. Summarize income within each bracket.
2. Create a box plot or bar chart for visualization.



- **Insights**:
    1. Younger age groups might have lower income compared to older brackets.
    2. Box plots can highlight outliers or variations within each group.

3. Categorical & Categorical

Analysis and Visualization

For two categorical variables:

1. **Frequency Plot**: Highlights the frequency of categories.
2. **Pivot Tables**: Tabulate relationships between categories.

Example: Gender and Marital Status

Using the dataset:

- **Steps**:
    1. Use a pivot table to summarize marital status (Single, Married) against gender (Male, Female).
    2. Add filters or visualizations like pie charts or bar charts.
- **Insights**:
    1. The distribution of marital status across genders can reveal societal trends or disparities.

Multivariate Analysis

What Is Multivariate Analysis?

This involves examining relationships between three or more variables simultaneously. It is especially useful for identifying complex patterns or combined effects.

Example: Income, Age, and Marital Status

- **Steps**:
    1. Use pivot tables to combine Income, Age Bracket, and Marital Status.
    2. Visualize using stacked bar charts or 3D plots.
- **Insights**:
    1. Younger married individuals might earn differently compared to single counterparts.