Data Analytics

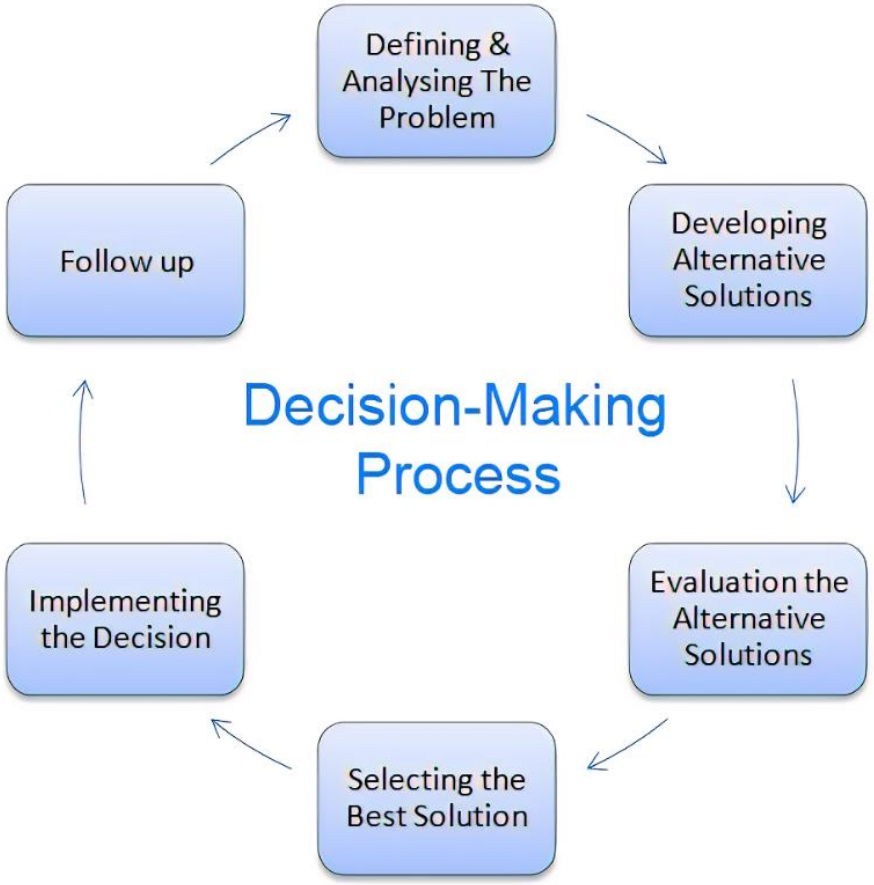
Introduction to Data Analytics

What is Data Analytics

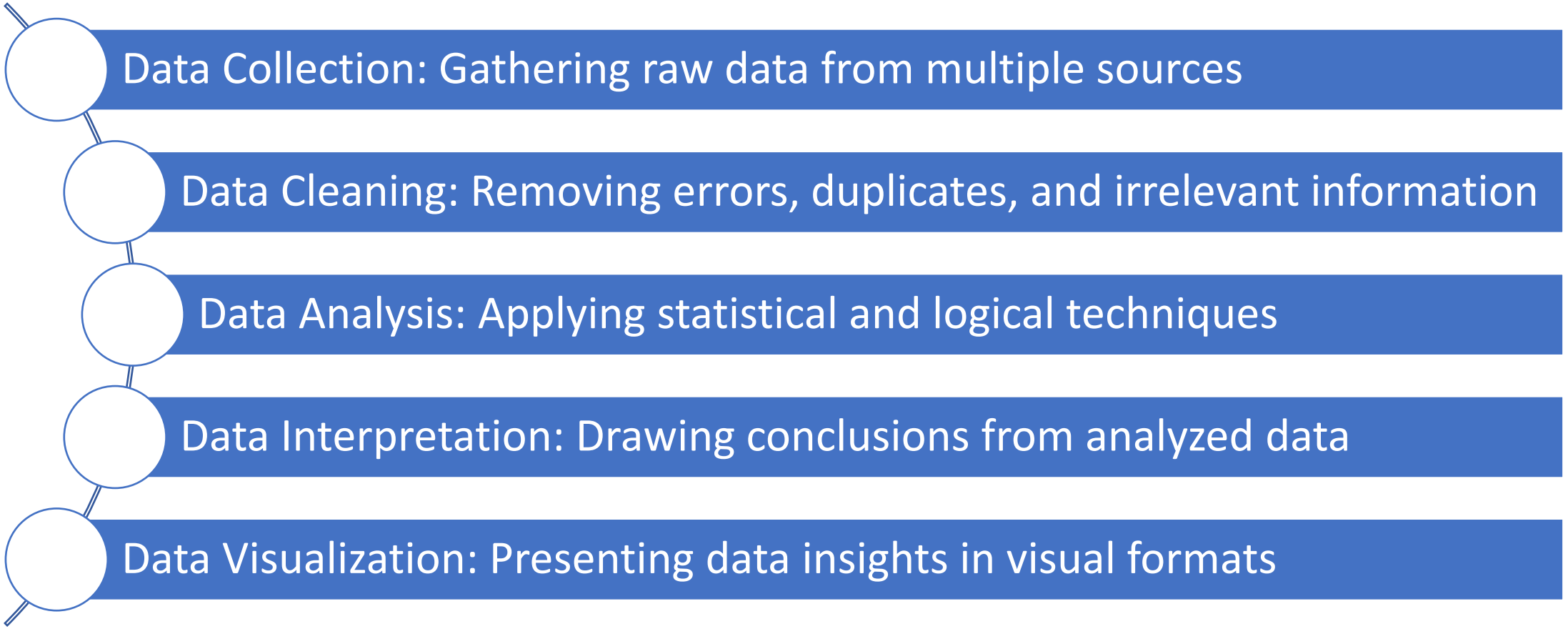
"Data analytics is currently one of the most buzzing terminologies. It involves the process of analyzing raw data to uncover trends and answer questions."

Why is Data Analytics Important?

- "Data analytics plays a crucial role in helping organizations make data-driven decisions."
- "It improves operational efficiency, enhances customer satisfaction, and helps businesses remain competitive."



The Steps of Data Analytics



Tools for Data Analytics

"Various tools are used in data analytics to process and analyze data.

These tools range from

- Programming Languages like Python and R
- visualization tools like Tableau and Power BI
- Microsoft Excel"

Data-Driven Decision making

"Data-driven decision making involves collecting, analyzing, and using data to guide business decisions and strategies."

"In today's data-rich environment, leveraging accurate data leads to more informed, objective, and successful outcomes."

Why is Data Crucial in Decision Making

"Data provides a factual basis for decision-making, eliminating guesswork and bias."

"It helps identify trends, measure performance, predict outcomes, and optimize strategies."

Types of Data for Decision Making

Descriptive Data: Provides information on what has happened (historical data).

Example: A retail company analyzing last year's sales data to understand which products sold the most during the holiday season.

Diagnostic Data: Explains why something has happened (cause-and-effect).

Example: A car manufacturer investigating a spike in customer complaints and finding that a faulty component caused increased product returns.

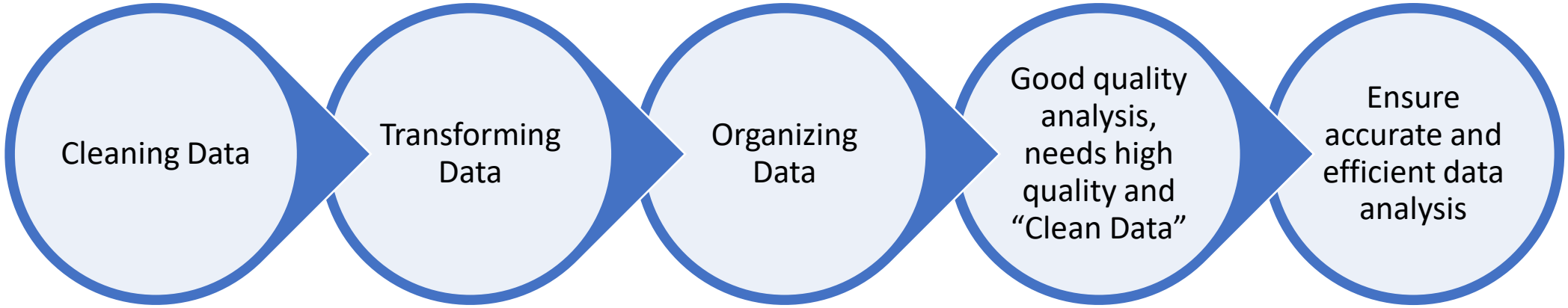
Predictive Data: Forecasts future trends and behaviors.

Example: An e-commerce company using customer browsing and purchasing history to predict what products each customer is likely to buy next month.

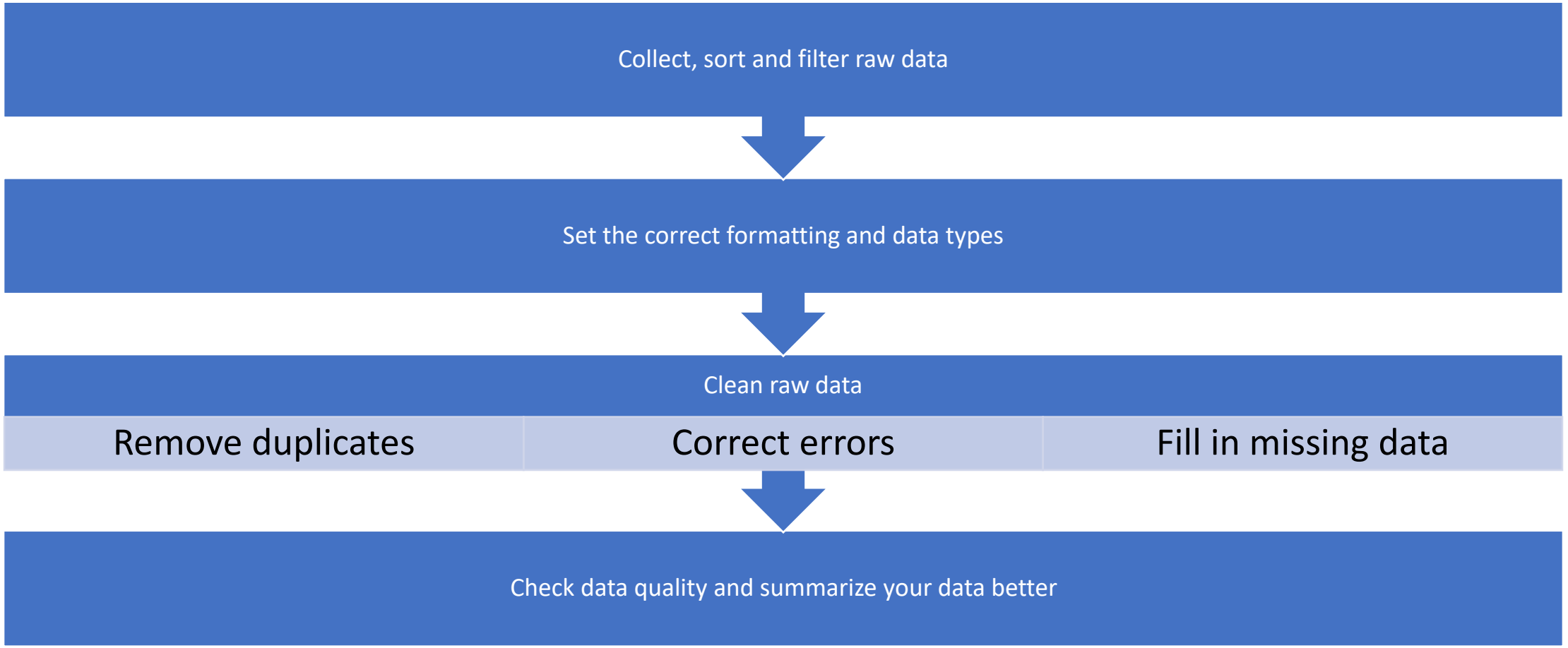
Prescriptive Data: Suggests actions based on the analysis.

Example: A logistics company using data from past deliveries and traffic patterns to recommend the most efficient routes for drivers to reduce delivery times.

Data preparation involves:



Doing data Preparation



Gathering raw data

- Collect the required raw data in Excel
- Databases like SQL Server can export data to Excel
- Enter data manually into Excel
  - Can be very time-consuming
- Import data from other files:
  - CSV files
  - Text files
  - Web (HTML) files

Removing duplicates

- Key step in data preparation process
- Excel has a “Remove Duplicates” feature
- Important to remove incorrectly recorded duplications
  - Keep values that are correctly repeated.

EXPLORATARY Data Analysis (EDA)

- The Importance of EDA
  - EDA is crucial for understanding data and generating insights.
  - Identifying patterns, anomalies, and relationships helps inform decision-making.
  - EDA is key to making data-driven decisions in today's world.

What is Exploratory Data Analysis (EDA)?

- **Exploratory Data Analysis (EDA)** is an iterative process used to:
- Summarize, visualize, and explore data.
- Find patterns, relationships, and anomalies.
- EDA helps build intuition before formal modeling or hypothesis testing.

Steps involved in EDA

Step 1: Understand the Problem and the Data

Step 2: Import and Inspect the Data

Step 3: Handling Missing Values

Step 4: Explore Data Characteristics

Step 5: Perform Data Transformation

Step 6: Visualize Data Relationships

Step 7: Handling Outliers

Step 8: Communicate Findings and Insights

Univariate Analysis

- **Definition**:
  - Univariate analysis involves analyzing a **single variable** to understand its distribution, central tendency, and spread. It does not involve any relationships with other variables.
- **Key Aspects**:
  - Distribution of data (normal, skewed)
  - Measures of central tendency (mean, median, mode)
  - Measures of spread (variance, standard deviation, range)
- **Example**:
  - Analyzing the **age distribution** of employees in a company.

Recommended charts for Univariate Analysis

1. **Recommended Charts**: **Histogram**: To visualize the frequency distribution of data.
    1. Example: Age distribution in a company.
2. **Box Plot**: To show the spread of the data and detect outliers.
    1. Example: Salary distribution in a company.
3. **Bar Chart**: For categorical data.
    1. Example: Count of employees based on departments.

Bivariate Analysis

- **Definition**:
  - Bivariate analysis explores the relationship between **two variables**. This can be used to identify correlations, dependencies, or trends between variables.
- **Key Aspects**:
  - Association between variables
  - Correlation (positive, negative, or none)
  - Causality (though this requires more than just correlation)
- **Example**:
  - Analyzing the relationship between **employee salary** and **years of experience**.

Recommended Charts: Bivariate Analysis

- **Scatter Plot**: To visualize the relationship between two continuous variables.
  - Example: Plotting years of experience vs. salary.
- **Line Chart**: To show trends over time.
  - Example: Monthly revenue vs. number of products sold over time.
- **Heatmap**: To visualize the strength of relationships between two categorical variables.
  - Example: Correlation matrix of exam scores between different subjects.
- **Side-by-Side Bar Chart**: To compare two categorical variables.
  - Example: Comparing gender distribution across departments.

Multivariate Analysis

- **Definition**:
- Multivariate analysis involves analyzing **more than two variables** simultaneously.
- This helps in understanding the complex relationships and interactions between multiple factors.
- **Key Aspects**:
- Understand how multiple variables interact together.
- Analyze the impact of one variable while controlling for others.
- **Example**:
- Analyzing how **age**, **years of experience**, and **education level** impact an employee's **salary**.

Recommended Charts for Multivariate Analysis

- **Pair Plot (or Scatterplot Matrix)**: To visualize pairwise relationships between all variables.
- Example: Age, salary, and years of experience.
- **3D Scatter Plot**: To show relationships between three continuous variables.
- Example: Visualizing salary based on years of experience and education level.
- **Bubble Chart**: To add a third dimension (size of bubbles) to a scatter plot.
- Example: Sales revenue vs. advertising budget, where the size of bubbles represents the number of products s
- **Stacked Bar Chart**: For analyzing more than two categorical variables.
- Example: Showing department, gender, and employee status in a stacked bar.

Relationships in Each Analysis:

- **Univariate Relationships**:
  Univariate analysis does not explore relationships between variables but focuses on summarizing a single variable.
- **Bivariate Relationships**:
  Looks for **correlations** or **associations** between two variables.
  Positive correlation: As one variable increases, the other increases (e.g., years of experience and salary).
  Negative correlation: As one variable increases, the other decreases (e.g., age and learning agility).
- **Multivariate Relationships**:
  Explores complex interactions between variables.
  Can involve controlling for one or more variables to see the effect of another (e.g., how education level affects salary while controlling for experience).
  Regression models and Principal Component Analysis (PCA) are often used in multivariate analysis to understand the effects of multiple factors.

Summary of Charts

Analysis Type	Purpose	Recommended Charts
Univariate	Understand the distribution of a variable	Histogram, Box Plot, Bar Chart
Bivariate	Examine relationship between two variables	Scatter Plot, Line Chart, Heatmap
Multivariate	Explore interaction between multiple variables	Pair Plot, 3D Scatter Plot, Bubble Chart, Stacked Bar Chart

Summary of Charts

Chart Type	Category	Explanation
Pie Chart	Univariate	Shows one variable distributed among different categories (e.g., market share).
Line Chart	Bivariate / Multivariate	Two variables (time and value), but can be extended to compare multiple variables using multiple lines.
Column/Bar Chart	Univariate / Bivariate / Multivariate	One variable (univariate) or comparison between two variables (bivariate) or multiple categories (multivariate).
Waterfall Chart	Univariate / Multivariate	Shows the sequential effect of one variable (e.g., profit) or can show multiple categories (multivariate).

Thank you