L02 – Understanding Data and Variable Type

Sarawoot Kongyoung, Ph.D.

sarawoot.kon@nstda.or.th

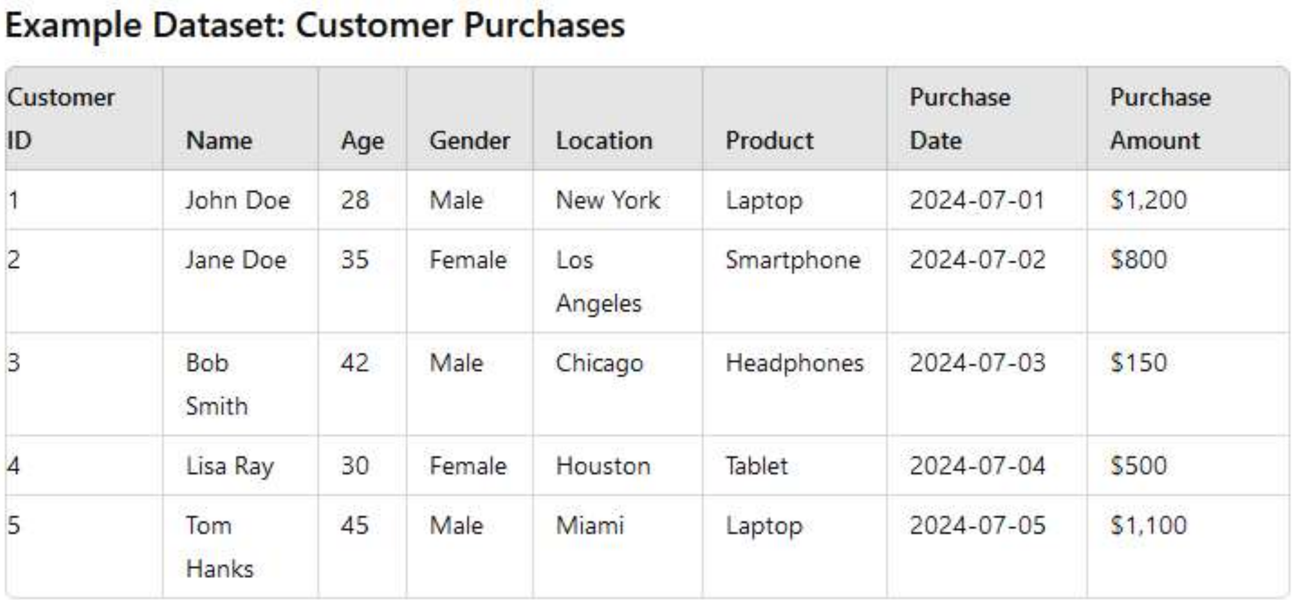
sarawoot.aj@siit.tu.ac.th

Last Lecture – Data Type

- Quantitative – Numeric data with natural ordering
  - Discrete Data
  - Continuous Data
- Qualitative – Qualities or characteristic
  - Categorical Data – Labelled
    - Nominal Data
    - Ordinal Data

Outline

- Data and relationship of data point
- Variables and variable types
- Data types conversion



Example Dataset: Customer Purchases

Customer ID	Name	Age	Gender	Location	Product	Purchase Date	Purchase Amount
1	John Doe	28	Male	New York	Laptop	2024-07-01	$1,200
2	Jane Doe	35	Female	Los Angeles	Smartphone	2024-07-02	$800
3	Bob Smith	42	Male	Chicago	Headphones	2024-07-03	$150
4	Lisa Ray	30	Female	Houston	Tablet	2024-07-04	$500
5	Tom Hanks	45	Male	Miami	Laptop	2024-07-05	$1,100

Data

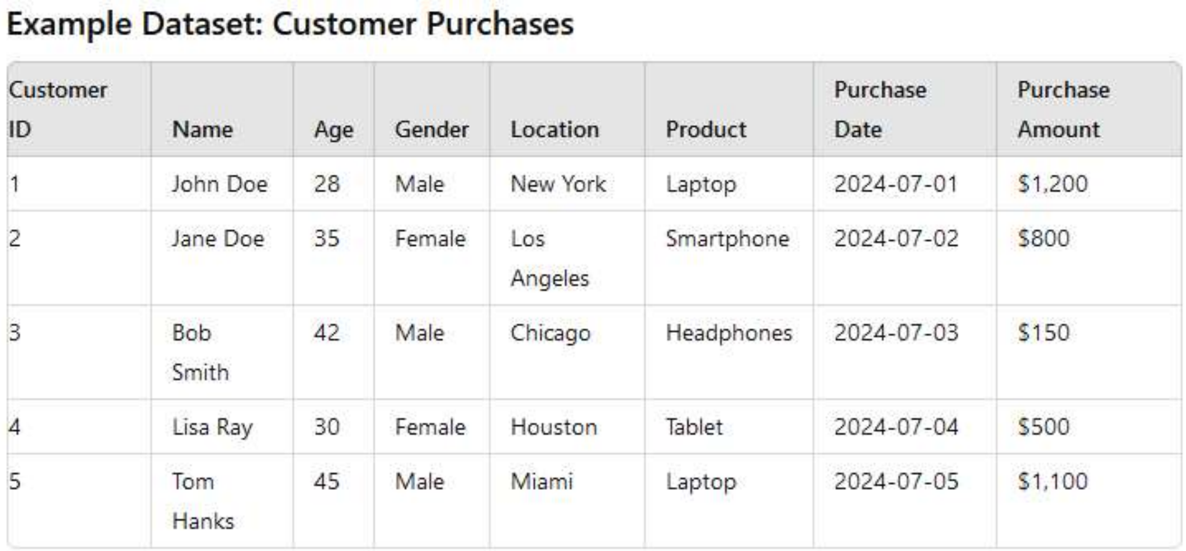
Data refers to information that is collected, stored, and analyzed to make decisions or gain insights.

Sources of Data

- **Internal Sources:**
  - Data generated within the organization, such as sales reports, customer databases, and financial statements.
- **External Sources:**
  - Data obtained from outside the organization, such as market research reports, industry statistics, and social media analytics.

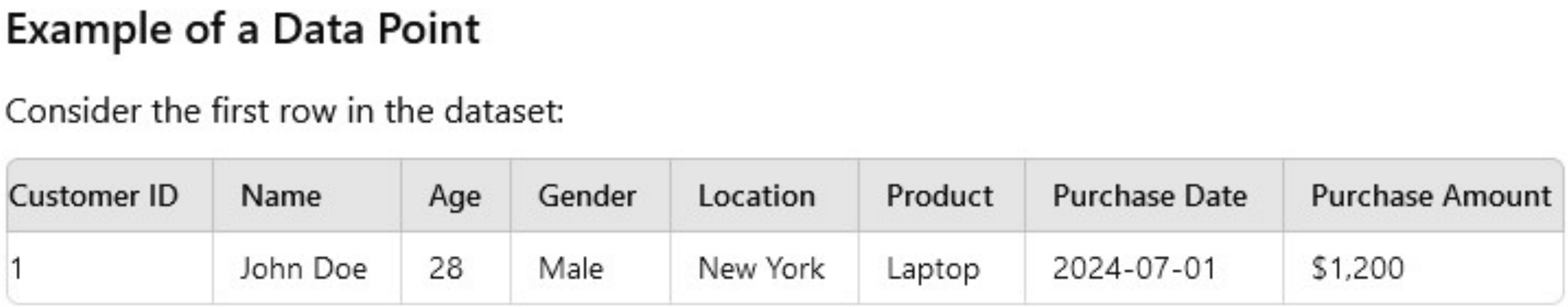
Data Point

- Refers to one row of the give data



Example Dataset: Customer Purchases

Customer ID	Name	Age	Gender	Location	Product	Purchase Date	Purchase Amount
1	John Doe	28	Male	New York	Laptop	2024-07-01	$1,200
2	Jane Doe	35	Female	Los Angeles	Smartphone	2024-07-02	$800
3	Bob Smith	42	Male	Chicago	Headphones	2024-07-03	$150
4	Lisa Ray	30	Female	Houston	Tablet	2024-07-04	$500
5	Tom Hanks	45	Male	Miami	Laptop	2024-07-05	$1,100



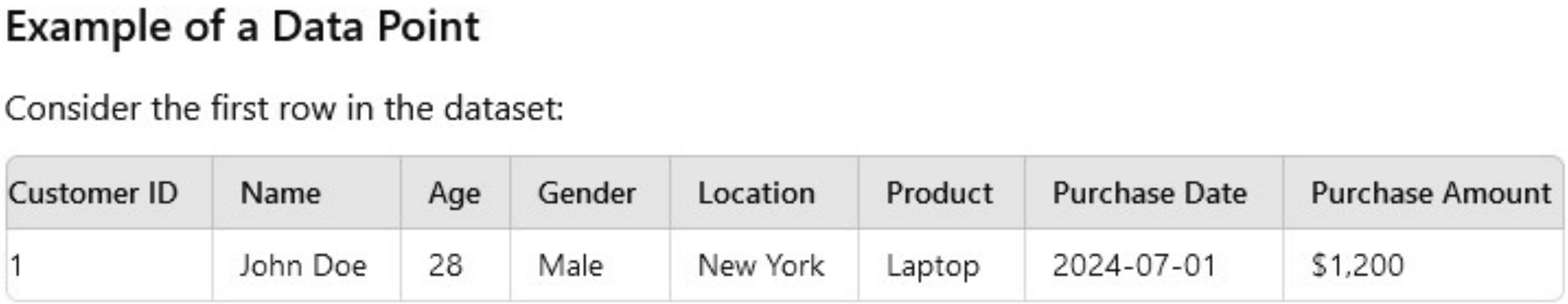
Example of a Data Point

Consider the first row in the dataset:

Customer ID	Name	Age	Gender	Location	Product	Purchase Date	Purchase Amount
1	John Doe	28	Male	New York	Laptop	2024-07-01	$1,200

Variable

- The columns that represent different attributes or characteristics of the data

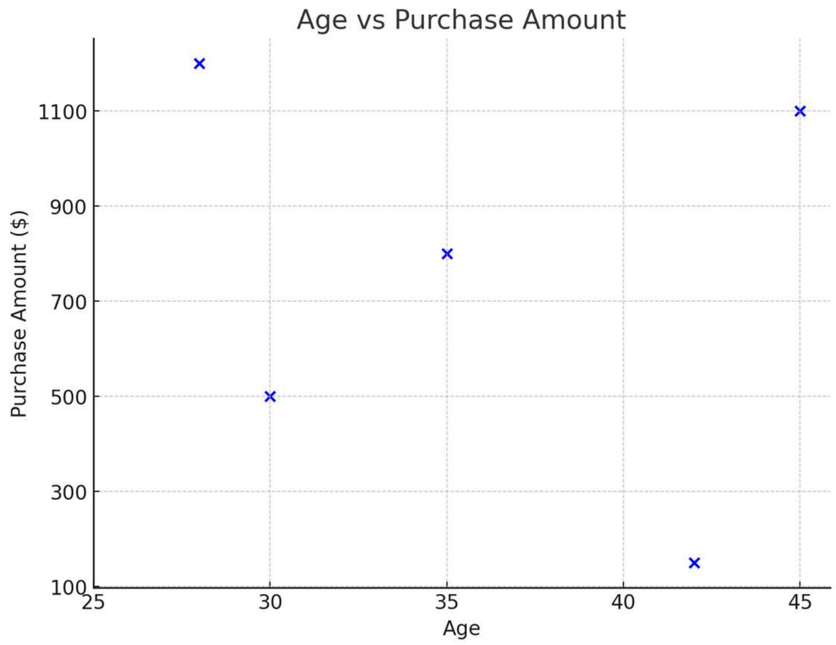


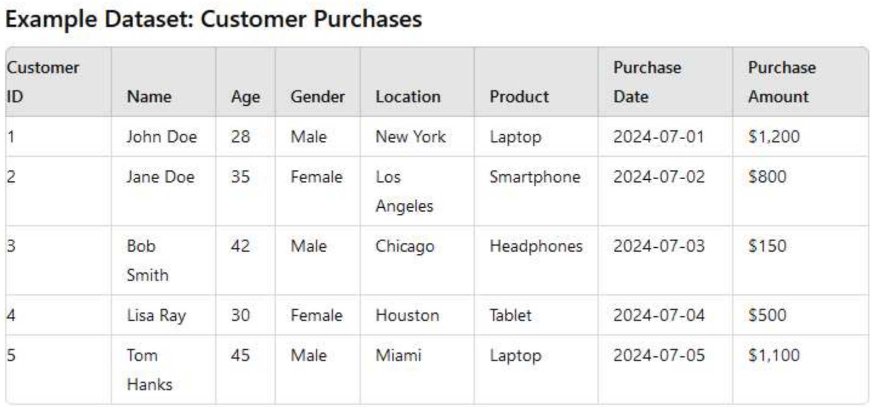
Example of a Data Point

Consider the first row in the dataset:

Customer ID	Name	Age	Gender	Location	Product	Purchase Date	Purchase Amount
1	John Doe	28	Male	New York	Laptop	2024-07-01	$1,200

- 5 variables in the above example:
  - Customer ID, Name, Age, Gender, Location, Product, Purchase Date, Purchase Amount

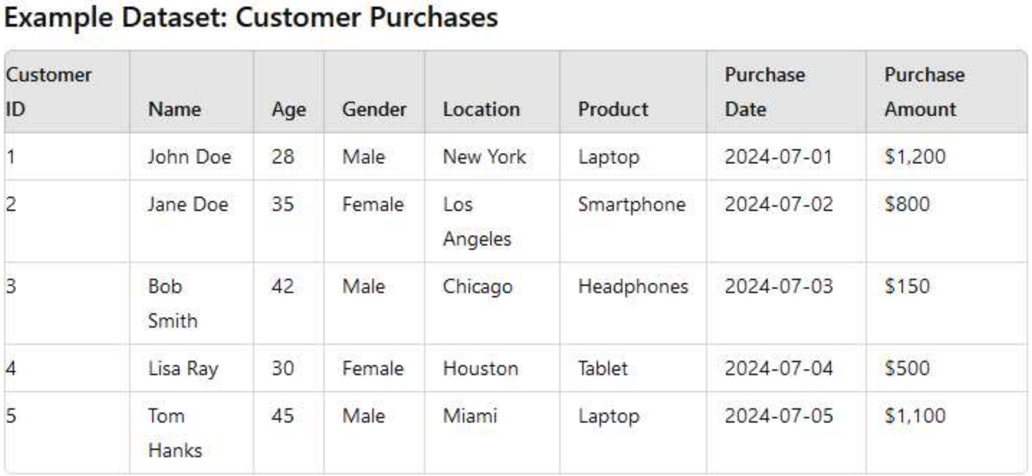




Example Dataset: Customer Purchases

Customer ID	Name	Age	Gender	Location	Product	Purchase Date	Purchase Amount
1	John Doe	28	Male	New York	Laptop	2024-07-01	$1,200
2	Jane Doe	35	Female	Los Angeles	Smartphone	2024-07-02	$800
3	Bob Smith	42	Male	Chicago	Headphones	2024-07-03	$150
4	Lisa Ray	30	Female	Houston	Tablet	2024-07-04	$500
5	Tom Hanks	45	Male	Miami	Laptop	2024-07-05	$1,100

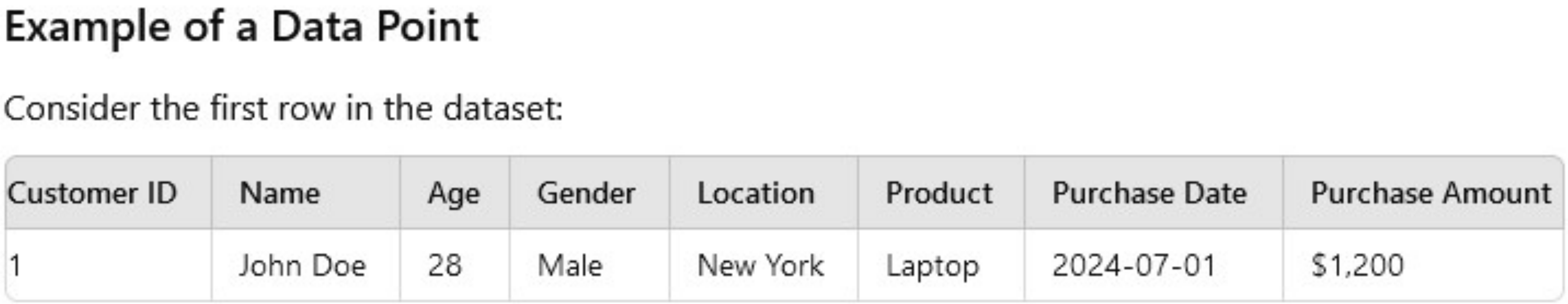
A scatter plot of Age and Purchase Amount



Example Dataset: Customer Purchases

Customer ID	Name	Age	Gender	Location	Product	Purchase Date	Purchase Amount
1	John Doe	28	Male	New York	Laptop	2024-07-01	$1,200
2	Jane Doe	35	Female	Los Angeles	Smartphone	2024-07-02	$800
3	Bob Smith	42	Male	Chicago	Headphones	2024-07-03	$150
4	Lisa Ray	30	Female	Houston	Tablet	2024-07-04	$500
5	Tom Hanks	45	Male	Miami	Laptop	2024-07-05	$1,100

Relationship of Data Point



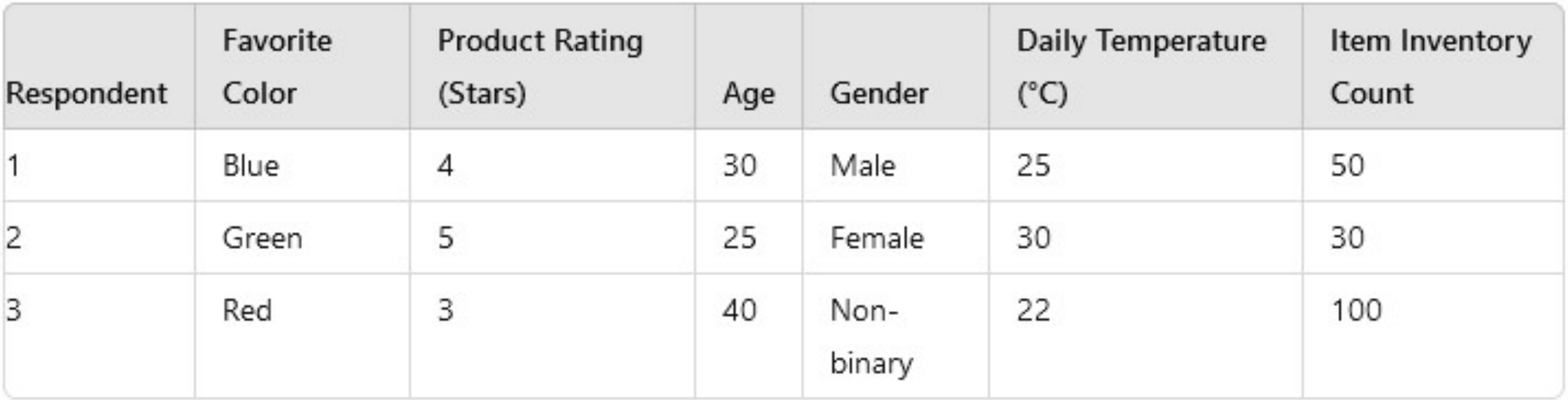
Example of a Data Point

Consider the first row in the dataset:

Customer ID	Name	Age	Gender	Location	Product	Purchase Date	Purchase Amount
1	John Doe	28	Male	New York	Laptop	2024-07-01	$1,200

Nondependency-Oriented Data

- Non-dependency-oriented data refers to data where the variables or data points **do not have a direct relationship among data points** or dependency on one another.



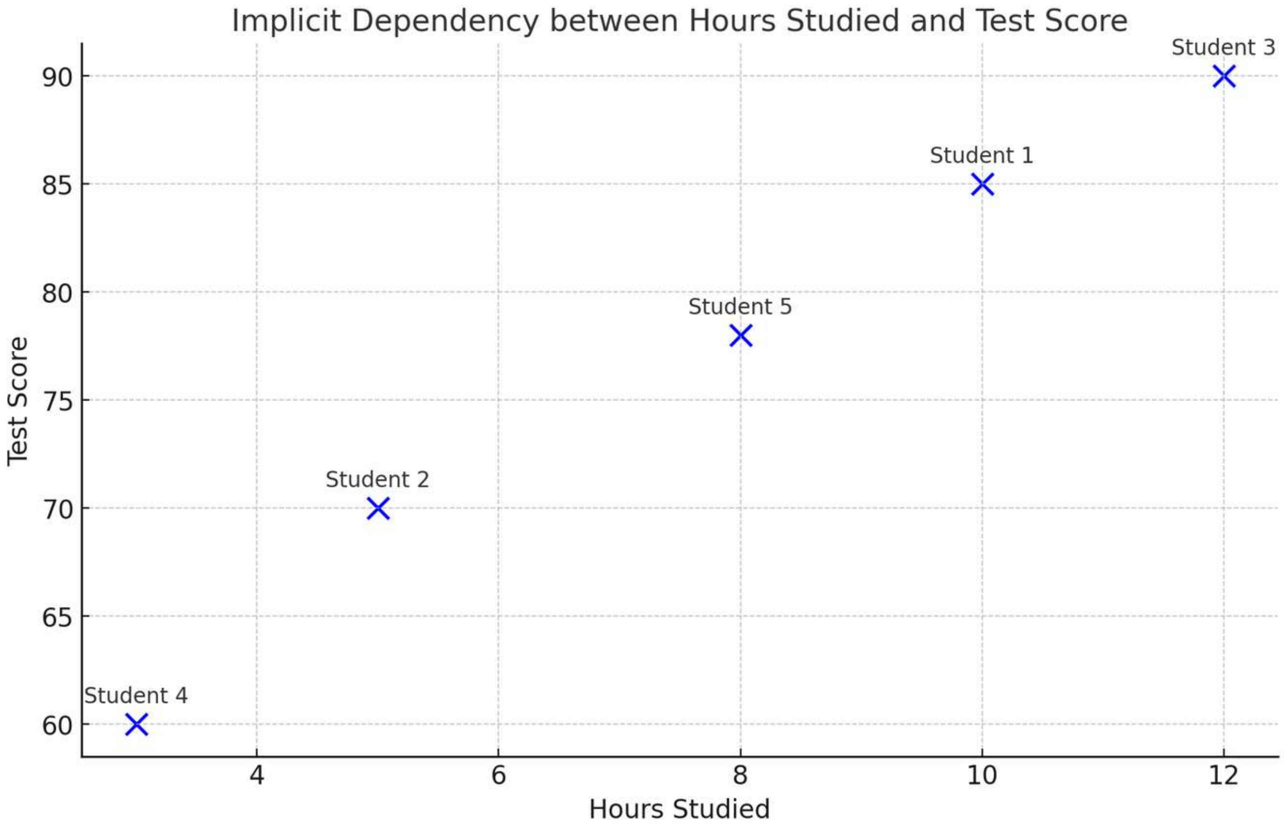
Respondent	Favorite Color	Product Rating (Stars)	Age	Gender	Daily Temperature (°C)	Item Inventory Count
1	Blue	4	30	Male	25	50
2	Green	5	25	Female	30	30
3	Red	3	40	Non-binary	22	100

Dependency-Oriented Data

- The data record may be implicitly or explicitly related to other data item
- **Implicit dependencies:**
  - The dependencies between data items are not explicitly defined or mandated but are inferred through analysis and observation.
- **Explicit dependencies:**
  - Clear, predefined relationships between variables
    - Financial model: Interest payment might be explicitly calculated as a percentage of the loan amount

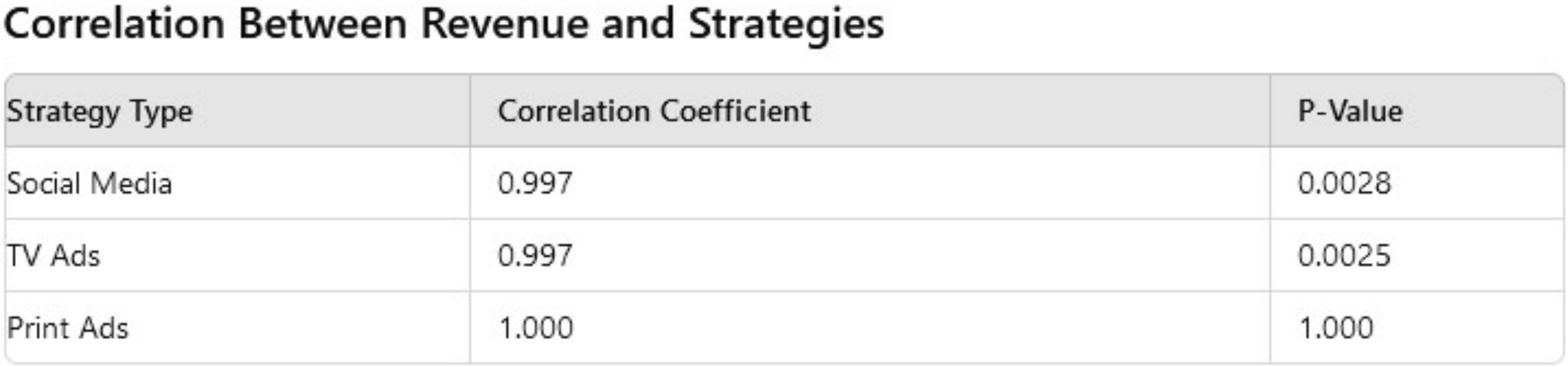
Implicit Dependencies

- **Observed Relationships:**
  - Based on observed patterns or trends



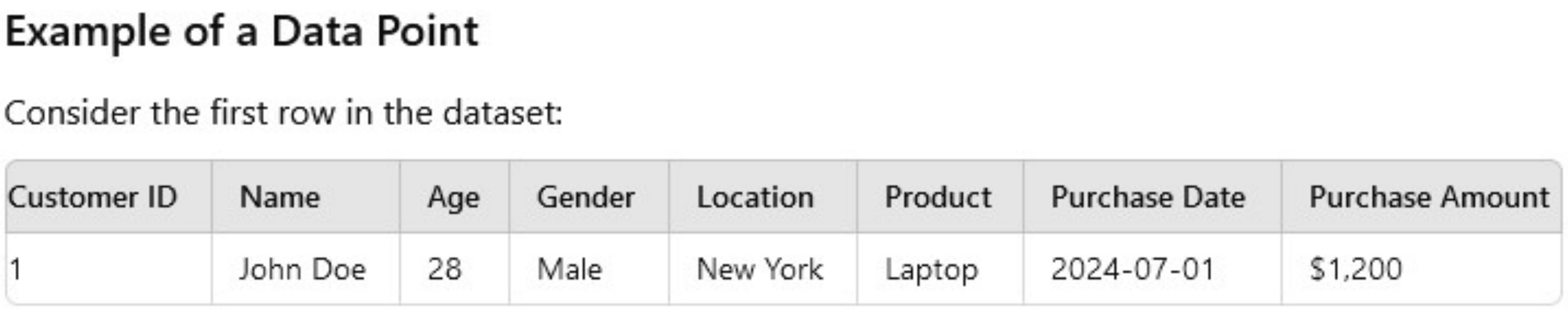
Implicit Dependencies

- **Statistical Inference:**
  - Use statistical methods or analysis are required to identify correlations or causal relationships between variables.



Correlation Between Revenue and Strategies

Strategy Type	Correlation Coefficient	P-Value
Social Media	0.997	0.0028
TV Ads	0.997	0.0025
Print Ads	1.000	1.000



Example of a Data Point

Consider the first row in the dataset:

Customer ID	Name	Age	Gender	Location	Product	Purchase Date	Purchase Amount
1	John Doe	28	Male	New York	Laptop	2024-07-01	$1,200

Types of Variable

Variable: Customer ID, Name, Age, Gender, Location, Product, Purchase Date, Purchase Amount

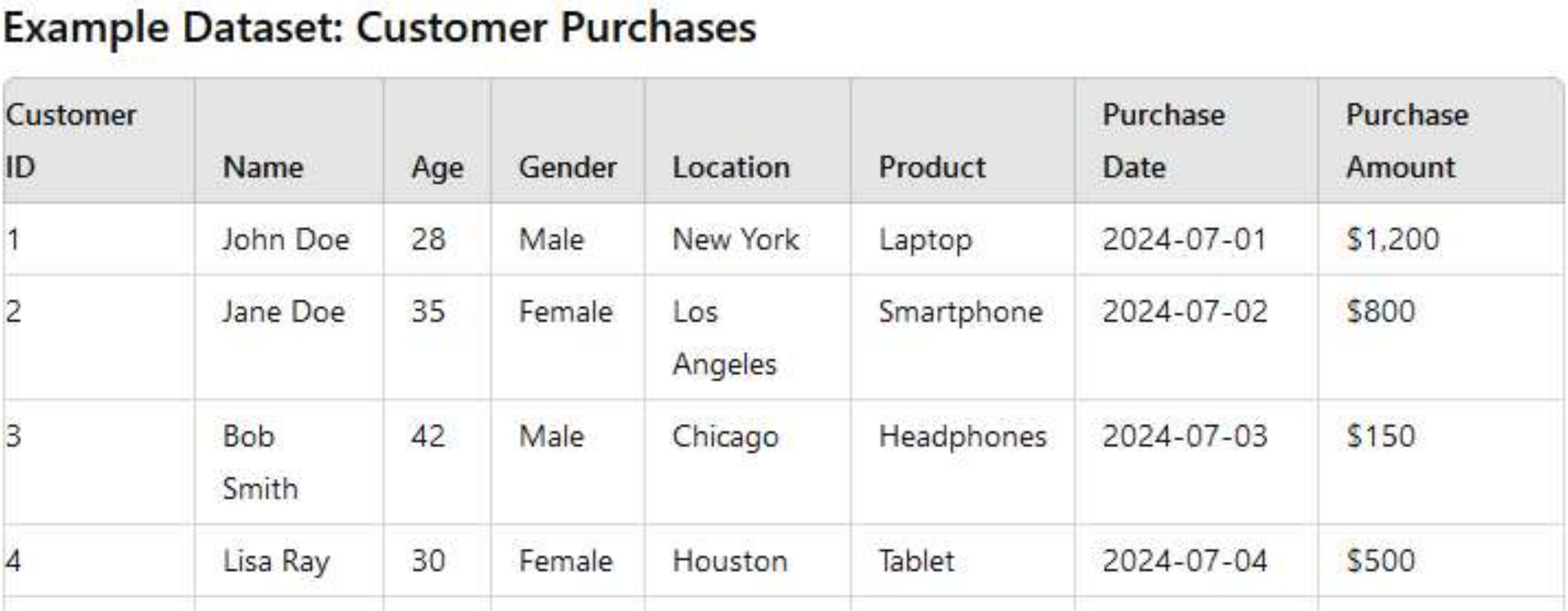
Type of Data - Last Lecture

Qualitative Data

- Numerical data have a natural ordering.
  - **Discrete Data:** Data refers to specific and distinct values or observations that can be counted.
  - **Continuous Data:** Data that can take any value in the range. It can be divided into smaller increments and measured with great precision.

Quantitative Data

- Non-numeric information that describes qualities or characteristics.
- **Categorical data:**
  - **Nominal Data:** Categories with no inherent order.
    - Example: Car brands (Toyota, Ford, BMW, Honda).
  - **Ordinal Data:** Categories with a meaningful order or ranking.
    - Example: Education level (High School, Bachelor's, Master's, Ph.D.).

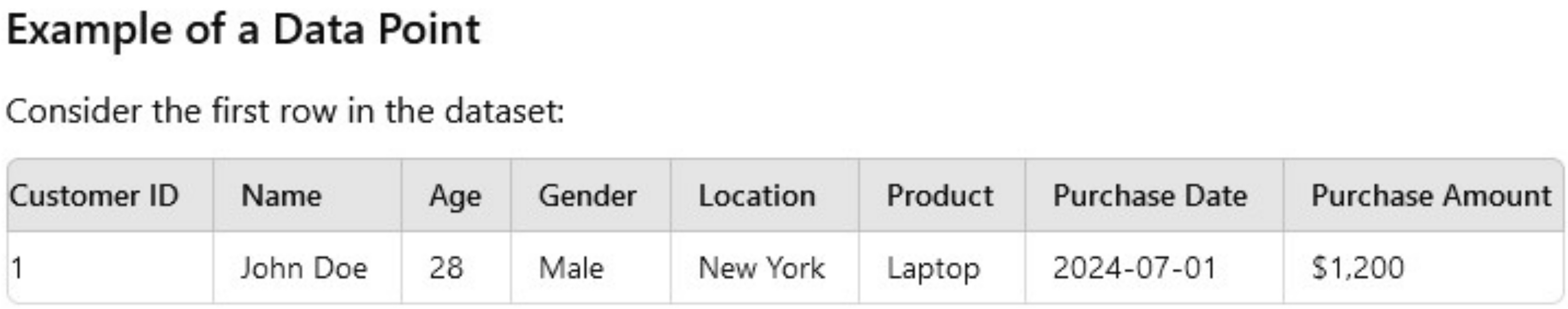


Example Dataset: Customer Purchases

Customer ID	Name	Age	Gender	Location	Product	Purchase Date	Purchase Amount
1	John Doe	28	Male	New York	Laptop	2024-07-01	$1,200
2	Jane Doe	35	Female	Los Angeles	Smartphone	2024-07-02	$800
3	Bob Smith	42	Male	Chicago	Headphones	2024-07-03	$150
4	Lisa Ray	30	Female	Houston	Tablet	2024-07-04	$500

Identify type of each variable (Quantitative vs. Qualitative)

Types of Variable



Example of a Data Point

Consider the first row in the dataset:

Customer ID	Name	Age	Gender	Location	Product	Purchase Date	Purchase Amount
1	John Doe	28	Male	New York	Laptop	2024-07-01	$1,200

- What is type of each variable in the given data point?
  - Numeric Variable
  - Categorical Variable
  - Binary Variable
  - Date/Time Variable

Numeric Variable

- Variables that represent quantities and are measured on a numeric scale.
  - **Types of Numeric Variable:**
    - **Continuous Variables**: Variable that can take any value within a range.
      - Examples: height, weight, and age.
    - **Discrete Variables**: Variable that can take only specific, separate values.
      - Examples: the number of customers, units sold, and days of the week.

Categorical Variables

- Variables that represent categories or groups and do not have a numeric value.
  - **Types of Categorical Variables:**
    - **Nominal Variables:** Categories with *no intrinsic order*.
      - Examples: gender (male/female), product type, and country.
    - **Ordinal Variables:** Categories with a *meaningful order* but the *intervals between the categories are not necessarily equal or quantifiable*.
      - Examples: customer satisfaction ratings (e.g., low, medium, high) and education level (e.g., high school, bachelor's, master's).

Binary Variable

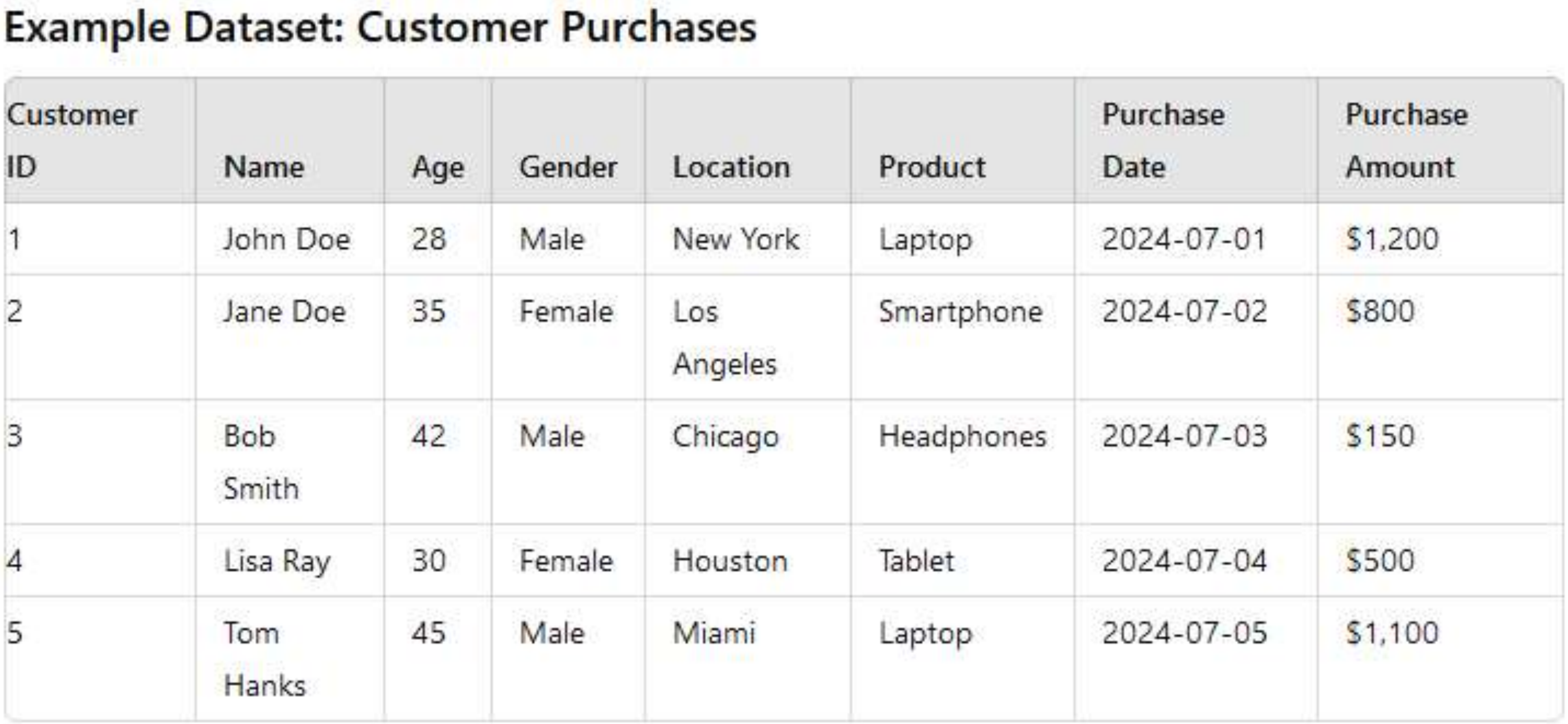
- A specific type of **categorical variable** with *only* ***two*** *possible values*.
  - Example of variables with value equals to yes/no, true/false, and male/female.

Date/Time Variable

- Variables that represent dates and times.
- It can be used to track temporal patterns, such as trends over time, seasonality, or event timing.
- Examples:
    - Transaction date and delivery time.

Text Variable (String)

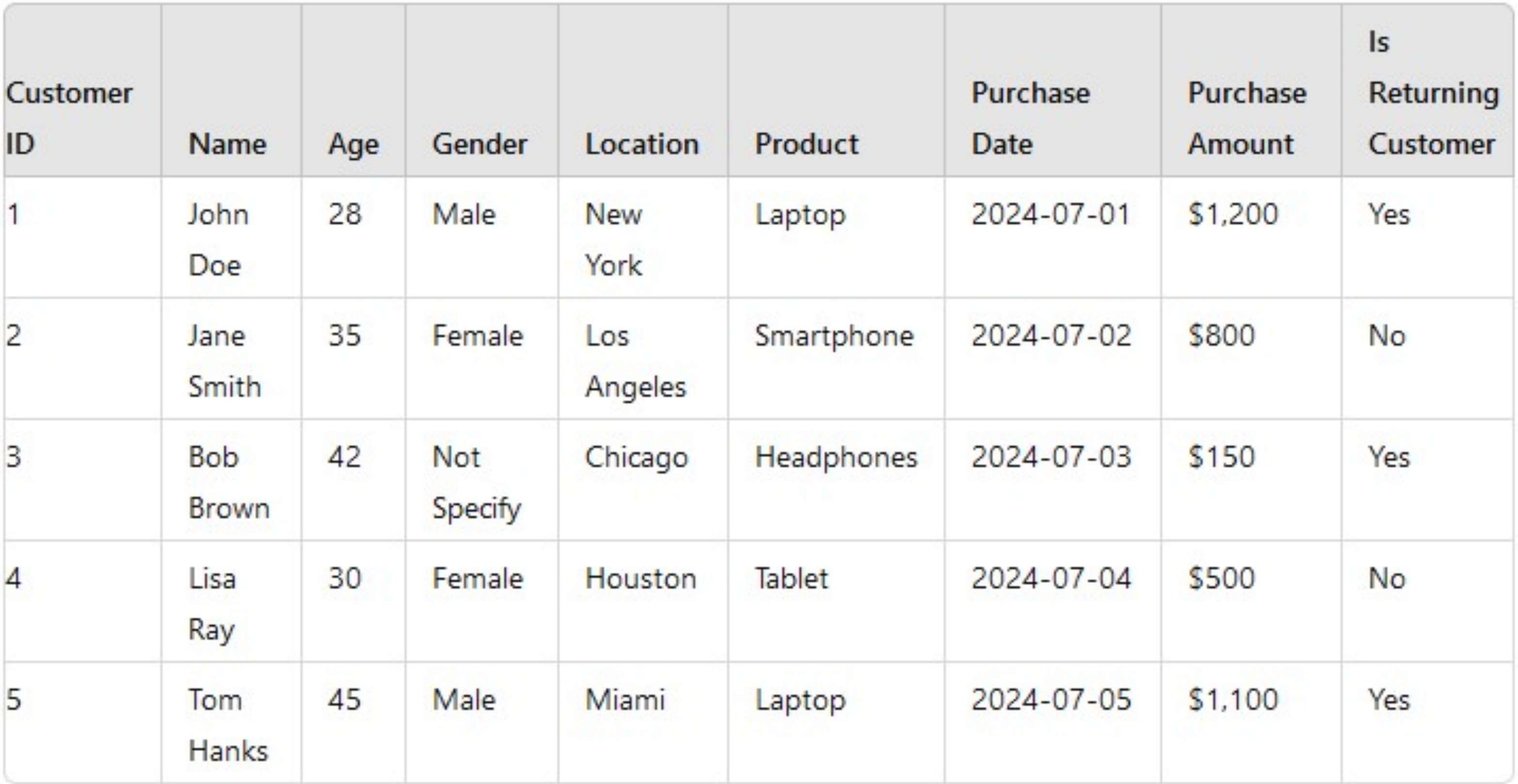
- Variable that stores sequences of characters.
  - Letters, numbers, symbols, spaces, and punctuation marks
- Example:
  - Simple Text: "Hello, world!"
  - Single Character: "A"
  - Numbers as Text: "12345" (Note: These are treated as text, not as numerical values.)
  - Special Characters: "@#$%^&*"



Example Dataset: Customer Purchases

Customer ID	Name	Age	Gender	Location	Product	Purchase Date	Purchase Amount
1	John Doe	28	Male	New York	Laptop	2024-07-01	$1,200
2	Jane Doe	35	Female	Los Angeles	Smartphone	2024-07-02	$800
3	Bob Smith	42	Male	Chicago	Headphones	2024-07-03	$150
4	Lisa Ray	30	Female	Houston	Tablet	2024-07-04	$500
5	Tom Hanks	45	Male	Miami	Laptop	2024-07-05	$1,100

Identify type of each variable



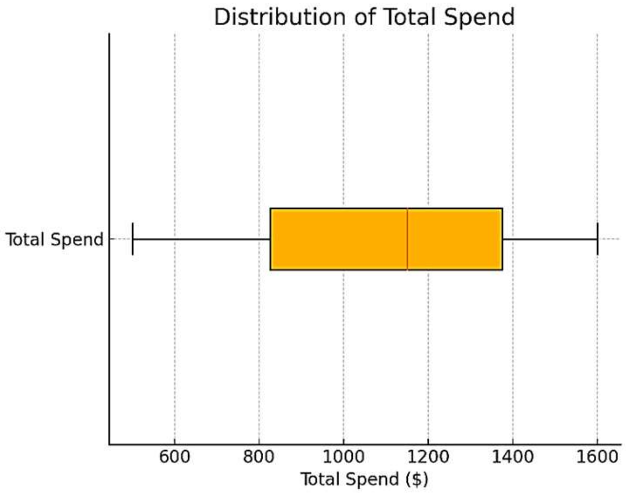
Customer ID	Name	Age	Gender	Location	Product	Purchase Date	Purchase Amount	Is Returning Customer
1	John Doe	28	Male	New York	Laptop	2024-07-01	$1,200	Yes
2	Jane Smith	35	Female	Los Angeles	Smartphone	2024-07-02	$800	No
3	Bob Brown	42	Not Specify	Chicago	Headphones	2024-07-03	$150	Yes
4	Lisa Ray	30	Female	Houston	Tablet	2024-07-04	$500	No
5	Tom Hanks	45	Male	Miami	Laptop	2024-07-05	$1,100	Yes

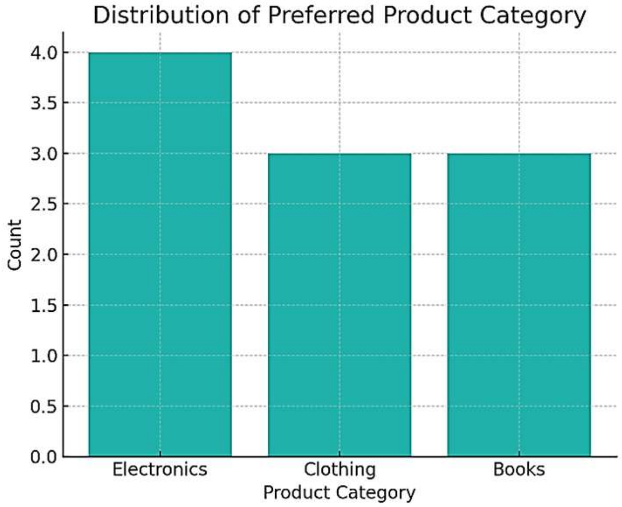
Identify type of “Gender” variable

Importance of Variable Types in Data Analysis

Descriptive Statistics

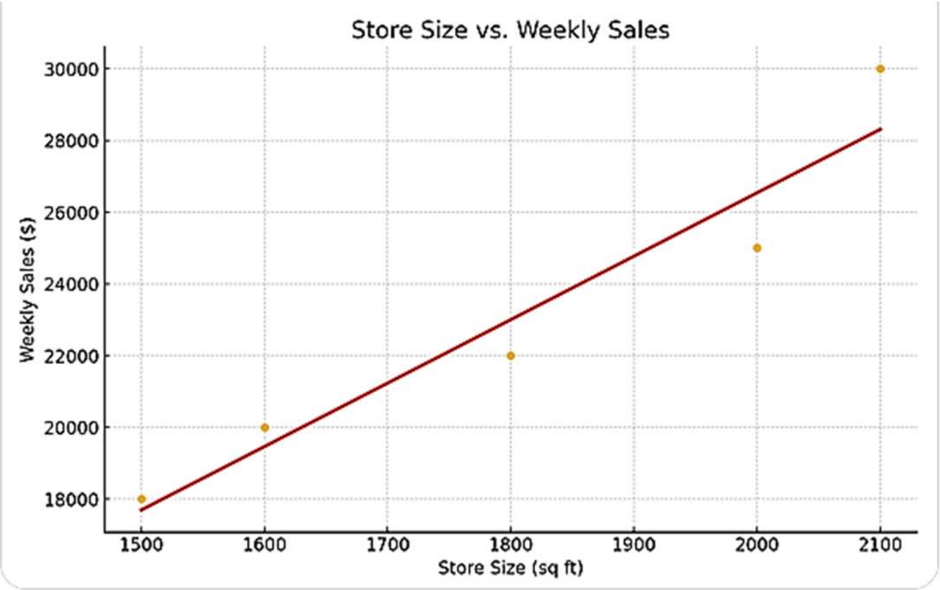
- **Numerical Variables:**
  - Calculate means, medians, and standard deviations for Age and Total Spend to understand average customer profiles and spending patterns.
- **Categorical Variables:**
  - Determine mode and frequency distribution for variables like Gender and Preferred Product Category to identify common traits.

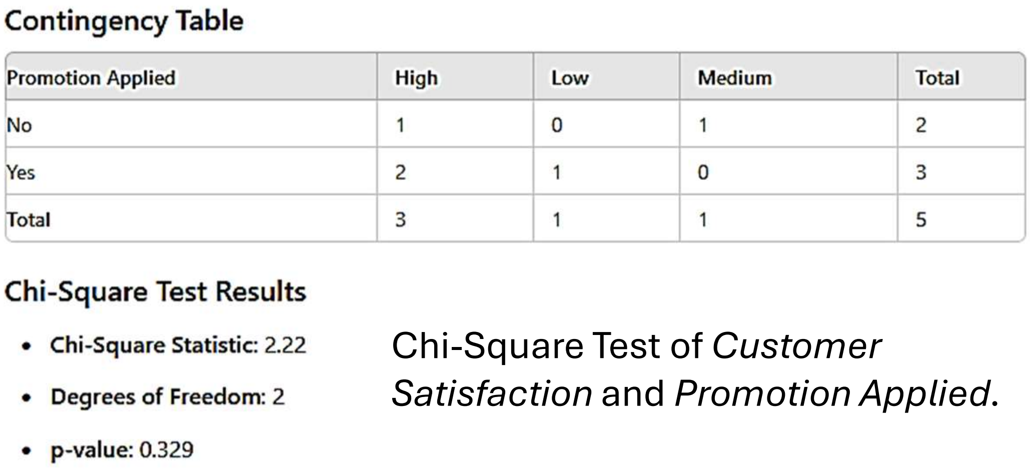




Statistical Analysis

- The statistical tests and models depends on the variable types
  - Example: regression for numerical variables, chi-square tests for categorical variables).





Contingency Table

Promotion Applied	High	Low	Medium	Total
No	1	0	1	2
Yes	2	1	0	3
Total	3	1	1	5

Chi-Square Test Results

- **Chi-Square Statistic:** 2.22
- **Degrees of Freedom:** 2
- **p-value:** 0.329

Chi-Square Test of *Customer Satisfaction* and *Promotion Applied*.

Data Cleaning and Preparation

- **Categorical Variables**:
  - Encode categorical variables for modeling (e.g., using one-hot encoding for Gender or Preferred Product Category).
- **Ordinal Variables**:
  - Ensure that the order is preserved during encoding (e.g., Membership Status: Bronze, Silver, Gold).

Location	Manager	Weekly Sales	Employee Count	Store Size (sq ft)	Promotion Applied	Customer Satisfaction	Gender	Membership
New York	Alice	25,000	10	2000	Yes	High	Female	Bronze
Los Angeles	Bob	18,000	8	1500	No	Medium	Male	Silver
Chicago	Carol	22,000	12	1800	Yes	Low	Female	Gold
Houston	David	20,000	9	1600	No	High	Male	Silver
Miami	Eva	30,000	11	2100	Yes	High	Female	Gold

Encoded categorical value

Location	Manager	Weekly Sales	Employee Count	Store Size (sq ft)	Promotion Applied	Customer Satisfaction	Gender_Male	Membership
New York	Alice	25,000	10	2000	Yes	High	0.0	0.0
Los Angeles	Bob	18,000	8	1500	No	Medium	1.0	1.0
Chicago	Carol	22,000	12	1800	Yes	Low	0.0	2.0
Houston	David	20,000	9	1600	No	High	1.0	1.0
Miami	Eva	30,000	11	2100	Yes	High	0.0	2.0

Importance of Variable Types in Data Analysis

- The type of variable determines which summary statistics are appropriate (e.g., mean for numerical variables, mode for categorical variables).

- The choice of statistical tests and models depends on the variable types (e.g., regression for numerical variables, chi-square tests for categorical variables).

- Variable types influence feature engineering and model selection, impacting the effectiveness of predictive models.

Location	Manager	Weekly Sales	Employee Count	Store Size (sq ft)	Promotion Applied	Customer Satisfaction	Gender	Membership
New York	Alice	25,000	10	2000	Yes	High	Female	Bronze
Los Angeles	Bob	18,000	8	1500	No	Medium	Male	Silver
Chicago	Carol	22,000	12	1800	Yes	Low	Female	Gold
Houston	David	20,000	9	1600	No	High	Male	Silver
Miami	Eva	30,000	11	2100	Yes	High	Female	Gold

- Give example of data point in the given data.
- List variables of each data point.
- List type of variables of each data point.

Lab

- Familiarization with data analytics tools and software
- Basic data exploration exercises
- Hands-on practice with data types conversion