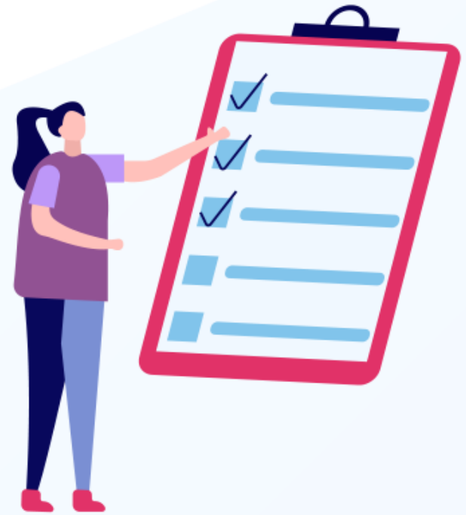
Applied Data Science with Python

Data Analysis

Learning Objectives

By the end of this lesson, you will be able to:

- Explain data and its lifecycle
- Classify the different types of data
- List the various steps involved in working with data
- Analyze how data can be imported and exported
- Analyze how data can be imported and exported



Business Scenario

ABC is an international retail company. It aims to optimize its supply chain and inventory management system by analyzing data from various sources such as store transactions, warehouse records, and online sales.

In this scenario, the company will use data science techniques to process and analyze collected data. They will wrangle, manipulate, and visualize the data using tools like spreadsheets, Tabula, dplyr, Matplotlib, Plotly, and Seaborn.

The insights gained will help the company make informed decisions to optimize its processes, reduce costs, and improve overall efficiency.

Understanding Data

Discussion: Understanding Data

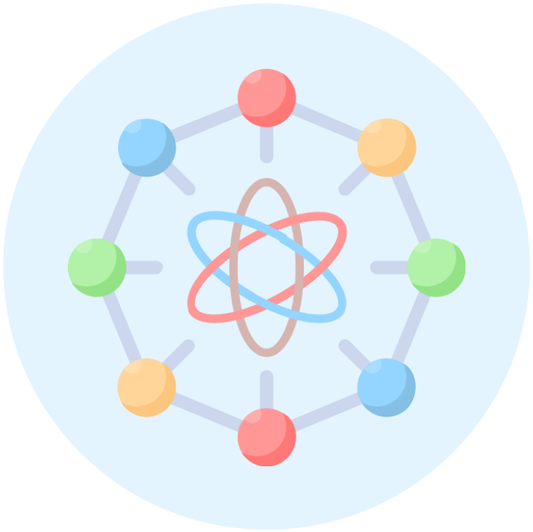
Duration: 10 minutes

- What are the factors that contributed to the emergence of data science and data analytics?
- What is the difference between qualitative and quantitative data?

Introduction

The fields of data science and data analytics have evolved as a result of the current data boom.

The following factors contributed to the emergence of data science and data analytics:

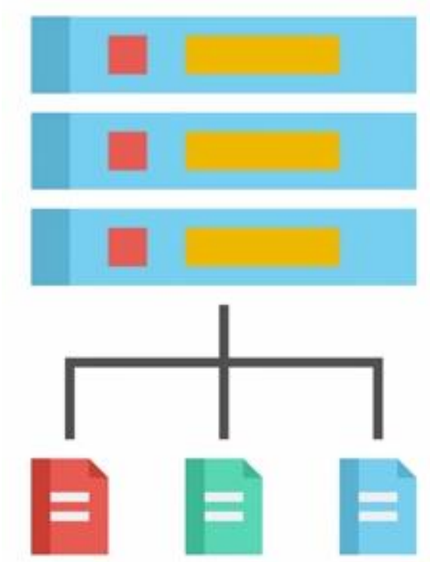


- Evolution of digital computers
- Evolution of the Internet and its wide-scale adoption by diverse entities

Data is at the heart of data science.

What Is Data?

According to TechTarget, data is information that has been translated into a form that is efficient for movement or processing.



It is converted into a binary digital form.

It can be used as a singular or plural subject.

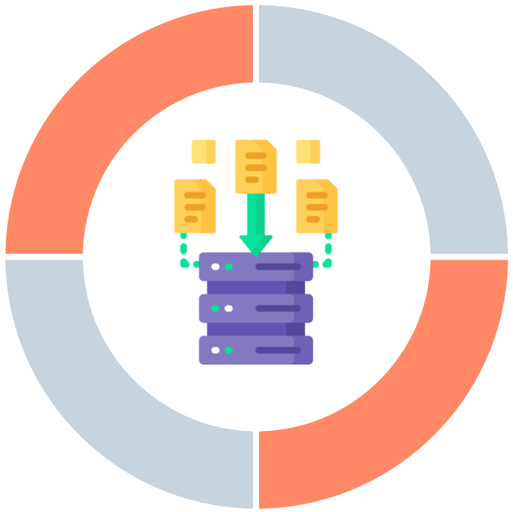
Raw data is information in its most fundamental digital form.

What Is Data?

Unless stated otherwise, data is always stored digitally. It is amenable to:

Analysis

Visualization



Processing

Interpretation

Any digitally stored data is subject to a definitive lifecycle.

Data Lifecycle

Data lifecycle is also known as:

Information lifecycle

Information supply chain lifecycle

Information resource lifecycle

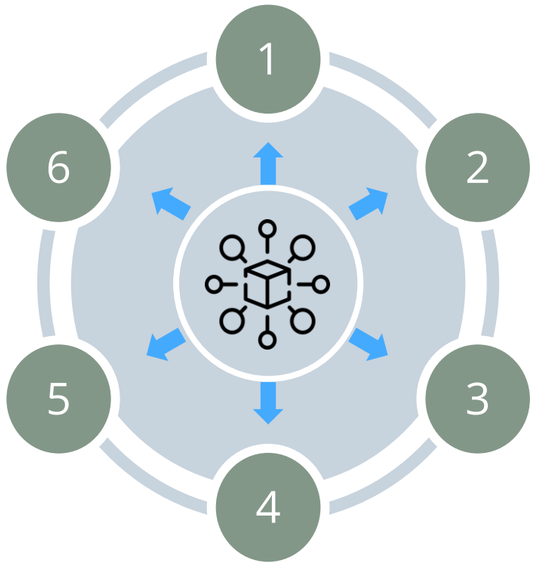
Data goes through six stages of information lifecycle, referred to as POSMAD.

What Is Data Lifecycle?

The distinct POSMAD phases are:

Plan:

Prepares for the resource



Obtain:

Allows acquiring of the resource

Store or share:

Holds information about the resource and makes it accessible

Maintain:

Ensures that the resource continues to work properly

Apply:

Uses resources to support and address business needs

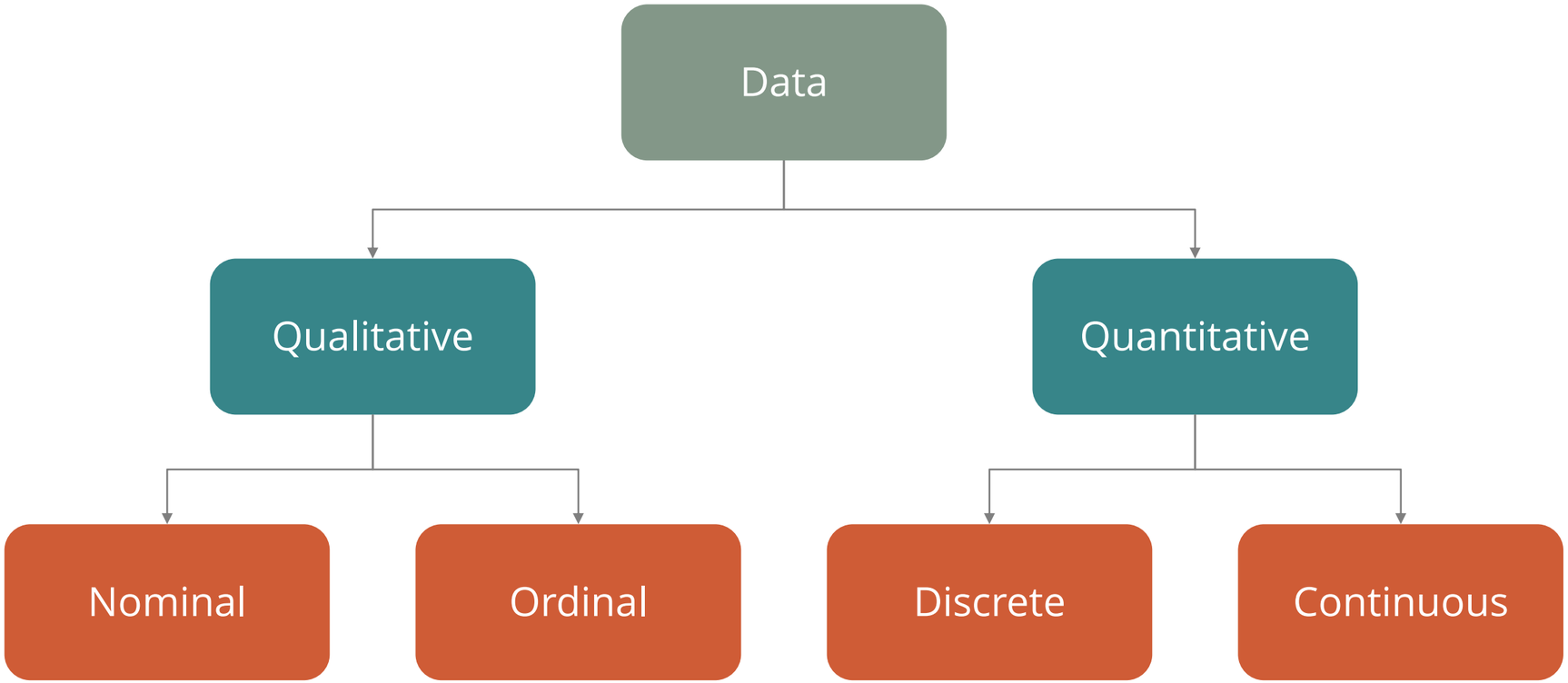
Dispose:

Deletes or discards a resource when it is no longer useful

Types of Data

Types of Data

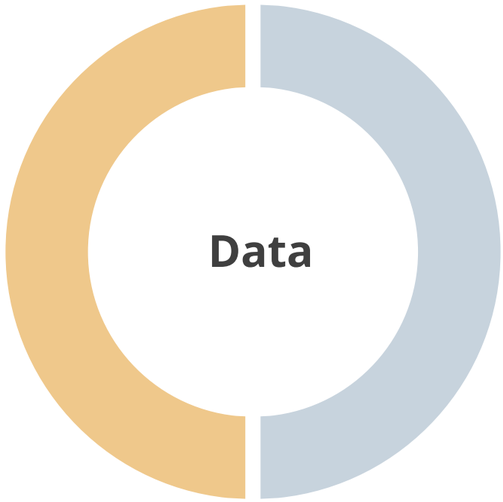
Data is broadly categorized as:



Types of Data

Qualitative

It describes qualities or characteristics. Example: Colors, sounds, words, symbols, images, or videos



Quantitative

It represents numbers and is associated with certain units. Example: Employee count, speed of a car, etc.

Types of Data: Qualitative

Nominal data

These are labels and they differ for different data types.

Example:
Gender includes male and female
Results include pass or fail

Ordinal data

These indicate numbers having natural, ordered categories. The distance between categories are unknown.

Example:
Service quality rating could be poor, average, good, excellent, or outstanding

Types of Data: Quantitative

Discrete data

It represents a whole number.

Example:
Population of a town, number of students who applied for scholarship in a particular year, etc.

Continuous data

It can be represented on a continuum and indicates a precise number.

It is represented by a floating-point number with a unit indicating what it stands for.

Example:
Volume of a tank (in cubic meters), average speed of wind (miles per hour), etc.

Types of Data

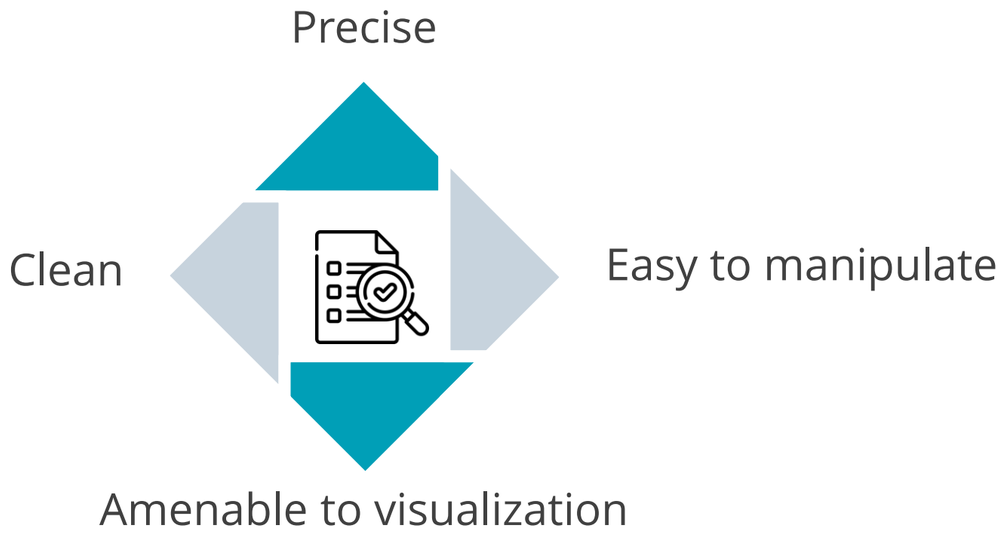
Programming languages used in data science include:

They can efficiently handle the different types of data.

Working with Data

Working with Data

Data scientists must ensure that the information they use is:



It simplifies data interpretation and produces clear and exact findings.

Working with Data

There are numerous steps involved when working with data. They are:

Data wrangling

Data manipulation

Data collection

Data visualization

Data Collection

It is the process of gathering and measuring information on targeted variables in an established system, that enables one to answer relevant questions and evaluate outcomes.

Data is gathered from various sources.

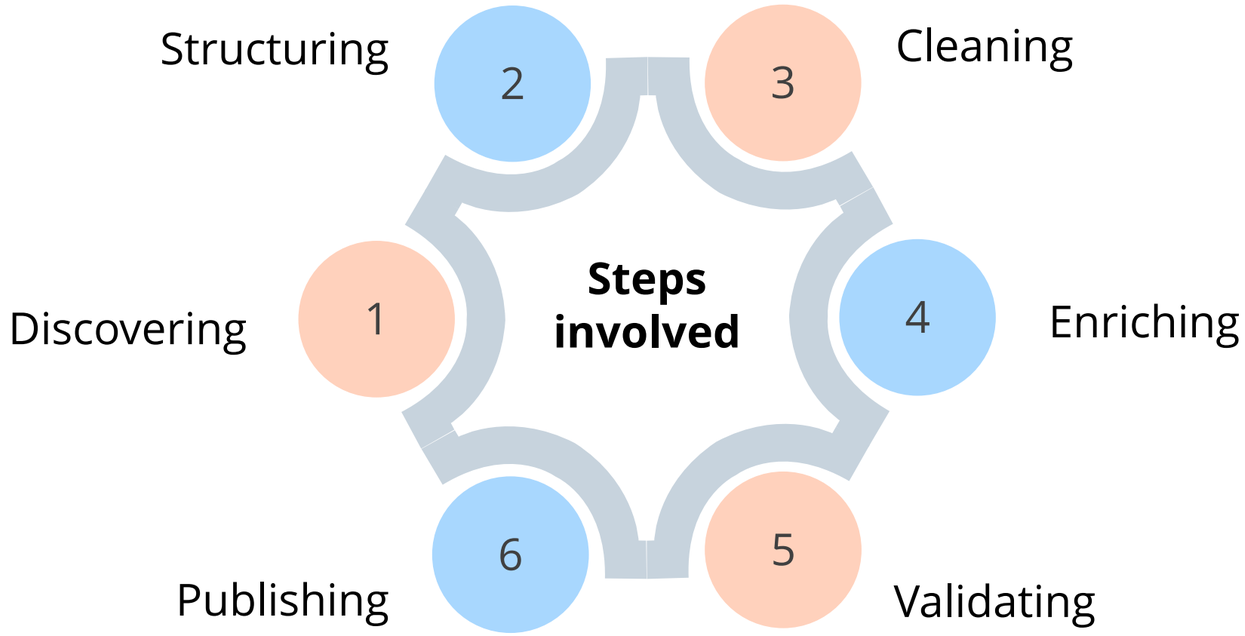
It can either be entered manually or automatically into a digital system.

Raw data cannot be used immediately for analytics.

Data Wrangling

In this, data is converted from one raw format to another to prepare it for down-stream analytics.

It involves considering the data's context and quality, and involves the following steps:



Data Wrangling

Python pandas alternatives

Spreadsheets

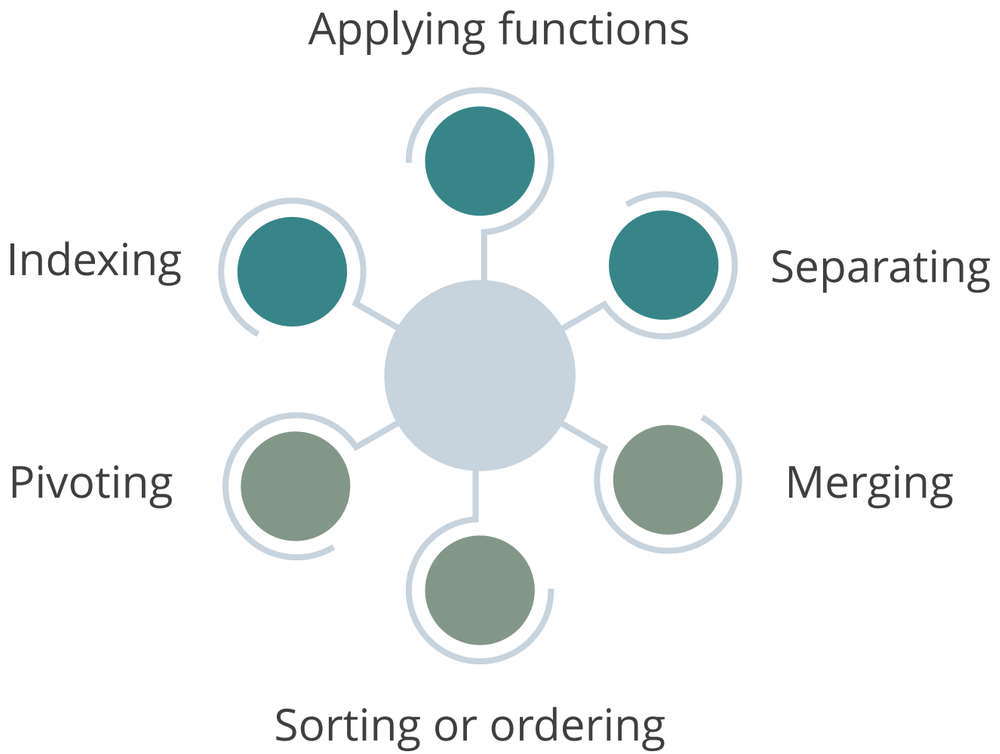
Tabula

Dplyr and Purr in R environment

Data Manipulation

Data can be jumbled, massive, complex, and inconsistent.

Here are some data manipulation techniques to address that:



Data Visualization

It is one of the most important components of working with data and gives a quick overview of the data, including textual data visualization.

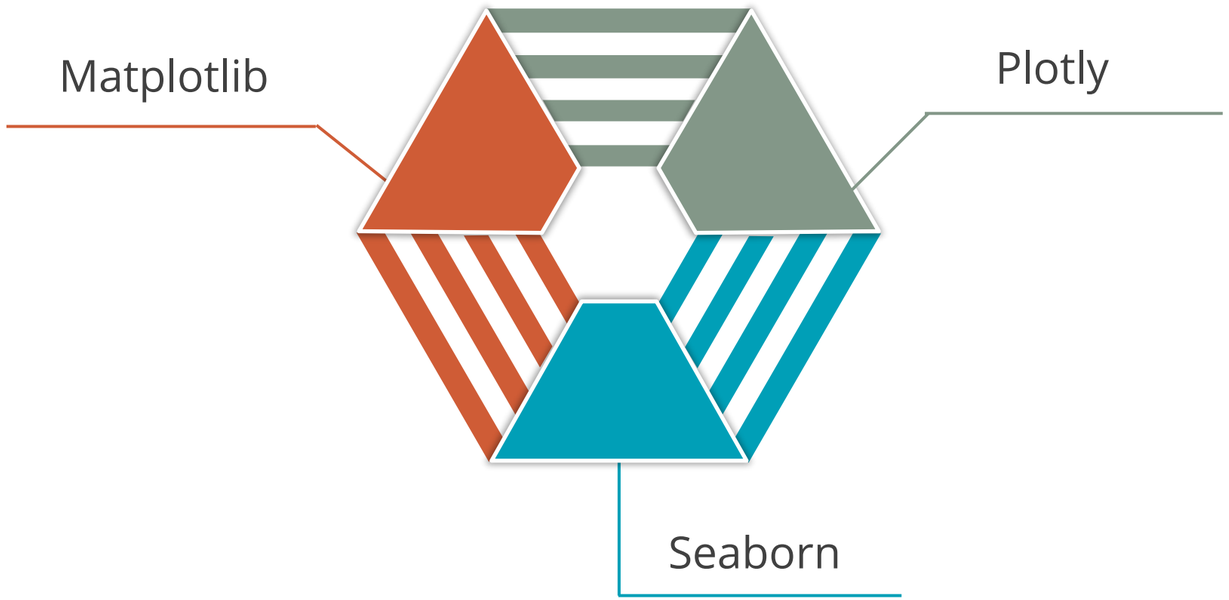
Example

Pandas' DataFrame provides head() and tail() functions to visualize the top and bottom parts of tabulated data.

Plotting technique is used for a graphical visualization experience of data and provides a first-hand impression, based on the graphical data output.

Data Visualization

Some of the graphical libraries used in the Python environment are:



Discussion: Understanding Data

Duration: 10 minutes

- What are the factors that contributed to the emergence of data science and data analytics?

Answer: The factors that contributed to the emergence of data science and data analytics are:

  - Evolution of digital computers
  - Evolution of the Internet and its wide-scale adoption by diverse entities

- What is the difference between qualitative and quantitative data?

Answer: Qualitative data describes qualities or characteristics, classifying observations into distinct categories or groups. On the other hand, quantitative data is numerical and measurable, often associated with specific units of measurement.

Assisted Practices

Let’s understand the topics below using Jupyter Notebooks.

- 10.5_Importing and Exporting Data in Python
- 10.6_Introduction to Regular Expressions
- 10.7_Manipulating Text with Regular Expressions
- 10.8_Accessing Databases Using Python

Note: Please download the pdf files for each topic mentioned above from the Reference Material section.

ASSISTED PRACTICE

Key Takeaways

- Information that is converted into a format that is useful for movement or processing is known as data.
- Data collection is the process of gathering and measuring information on targeted variables in an established system, which then enables one to answer relevant questions and evaluate outcomes.
- To get data ready for downstream analytics, data is transformed from one raw format to another.
- The wrangled data may be enormous, complicated, and contain missing values.

Knowledge Check

Knowledge Check 1

Which of the following phase of the data lifecycle holds information about the resource and makes it accessible for use through some distribution method?

A. Plan

B. Obtain

C. Store or Share

D. Apply

Knowledge Check 1

Which of the following phase of the data lifecycle holds information about the resource and makes it accessible for use through some distribution method?

A. Plan

B. Obtain

C. Store or Share

D. Apply

The correct answer is **C**

The Store or Share phase holds information about the resource and makes it accessible for use through some distribution method.

Knowledge Check 2

Which of the following is the process of gathering and measuring information on targeted variables in an established system?

A. Data collection

B. Data wrangling

C. Data manipulation

D. Data visualization

Knowledge Check 2

Which of the following is the process of gathering and measuring information on targeted variables in an established system?

A. Data collection

B. Data wrangling

C. Data manipulation

D. Data visualization

The correct answer is **A**

Data collection is the process of gathering and measuring information on targeted variables in an established system.

Knowledge Check

3

How is qualitative data categorized?

A. Discrete and continuous

B. Nominal and ordinal

C. Descriptive and inferential

D. None of the above

Knowledge Check 3

How is qualitative data categorized?

A. Discrete and continuous

B. Nominal and ordinal

C. Descriptive and inferential

D. None of the above

The correct answer is **B**

Qualitative data is categorized into nominal and ordinal.

Thank You