Data Preprocessing-Data Wrangling

Data Wrangling: The Process of Transforming Raw Data into a Clean and Useful Format

Data wrangling, also known as data munging or data cleaning, is the process of transforming raw data from various sources into a clean, consistent, and usable format. This process involves several steps to ensure that the data is accurate, complete, and in a suitable form for analysis.

Steps Involved in Data Wrangling:

1. **Data Collection**: Gathering data from various sources such as files, databases, or APIs.
2. **Data Cleaning**: Identifying and correcting errors, inconsistencies, and missing values in the data.
3. **Data Transformation**: Converting the data into a suitable format for analysis, such as converting date formats or aggregating data.
4. **Data Integration**: Combining data from multiple sources into a single dataset.

Example:

Suppose we have a dataset of customer information with the following structure:

Customer ID	Name	Email	Phone
1	John Smith	john.smith@example.com	123-4567
2	Jane Doe	jane.doe@example.com	234-5678

The data is collected from various sources, but it contains some errors and inconsistencies:

- Missing values in the "Email" column
- Incorrect phone number format (e.g., 123-4567 instead of 1234567)
- Typos in the "Name" column

Data Cleaning:

We need to identify and correct these errors. For example, we can use data validation techniques to check for missing values, incorrect formats, or typos.

Customer ID	Name	Email	Phone
1	John Smith	john.smith@example.com	1234567
2	Jane Doe	jane.doe@example.com	2345678

Data Transformation:

We need to convert the phone number format from string to numeric.

Customer ID	Name	Email	Phone (numeric)
1	John Smith	john.smith@example.com	1234567
2	Jane Doe	jane.doe@example.com	2345678

Data Integration:

We need to combine this dataset with another dataset containing customer order information.

Customer ID	Order Date	Product Name
1	2022-01-01	iPhone
1	2022-02-01	Laptop
2	2022-03-01	Tablet

After data wrangling, we have a clean and integrated dataset:

Customer ID	Name	Email	Phone (numeric)	Order Date	Product Name
1	John Smith	john.smith@example.com	1234567	2022-01-01	iPhone
1	John Smith	john.smith@example.com	1234567	2022-02-01	Laptop
2	Jane Doe	jane.doe@example.com	2345678	2022-03-01	Tablet

The data is now ready for analysis, and we can perform various statistical and machine learning tasks to gain insights from the data.

Curated by Brajesh Kumar