Welcome to

INTERNSHIP STUDIO

Module 04 | Lesson 04

Data Manipulation

Data Cleaning & Preparation with Pandas

Introduction to Data Cleaning

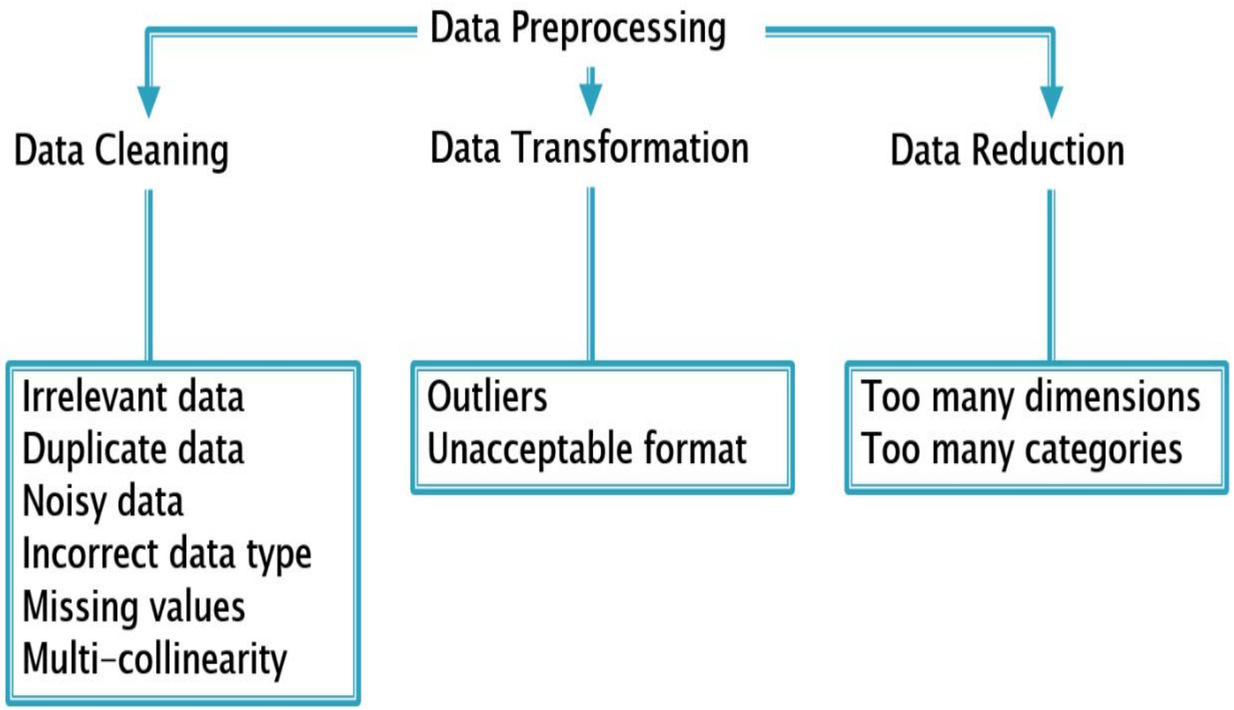
- Data cleaning and preparation are crucial steps in the data analysis process.
- Pandas is a powerful Python library that provides efficient tools for data cleaning and manipulation.

Why Data Cleaning is important?

- Data can be messy, inconsistent, or contain errors.
- Cleaning data ensures accuracy and reliability in analysis.
- Improves data quality for better decision-making.

Common Data Cleaning Tasks

1. Handling missing data
2. Removing duplicates
3. Correcting inconsistent data
4. Formatting data types
5. Handling outliers



Handling Missing Data

- Identify missing values using Pandas' functions like `isnull()` or `isna()`
- Options for handling missing data:

1. Remove rows or columns with missing values using `dropna()`
2. Fill missing values using `fillna()` with methods like mean, median and forward/backward filling

internship studio

Removing Duplicates

- Identify duplicates using **duplicated()** or **drop_duplicates()** functions.

- Remove duplicates using **drop_duplicates()** function with appropriate parameters.

Correct Inconsistent Data

- Standardize data formats using Pandas' string manipulation functions (**str.lower()**, **str.upper()**, **str.replace()**).

- Correct inconsistent data using functions like **replace()** or regular expressions (**re** module).

Formatting Data types

- Convert data types using **astype()** or **to_numeric()** functions.

- Ensure appropriate data types for each column (e.g., date, numeric, categorical).

Handling Outliers

- Detect outliers using statistical methods (e.g., z-score, IQR).
- Decide on appropriate actions:
  - Remove outliers if they are erroneous data points.
  - Keep outliers if they represent valuable information.
  - Replace outliers with a suitable value (e.g., mean, median).

SUMMARY

You got this

- Data cleaning and preparation are essential for accurate analysis.
- Pandas provides a wide range of functions for handling common data cleaning tasks.
- Understanding and applying these techniques improves data quality and analysis outcomes.

Next session

Merging data