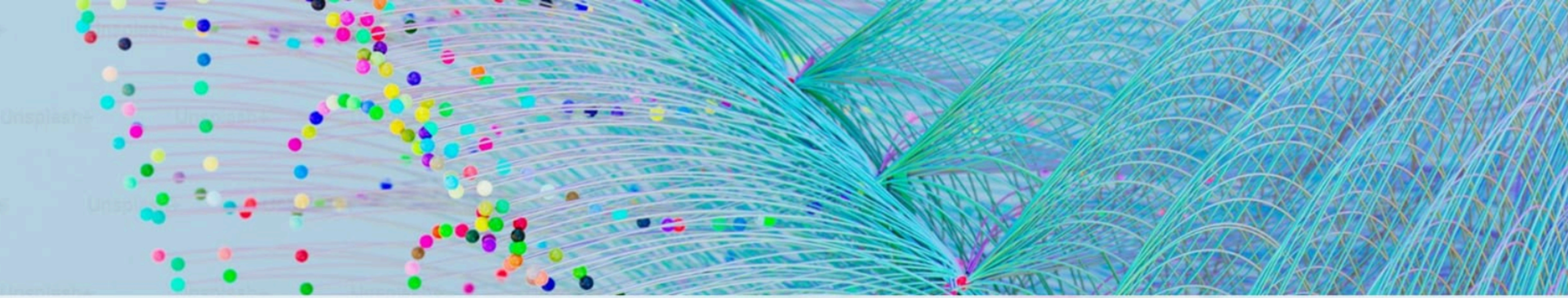


Pandas Data Cleaning

Mastering essential techniques for preparing and cleaning data using Python's pandas library for accurate analysis

by R&R Team

IBM Students

Contents

What is Pandas?

Powerful Data Analysis Library

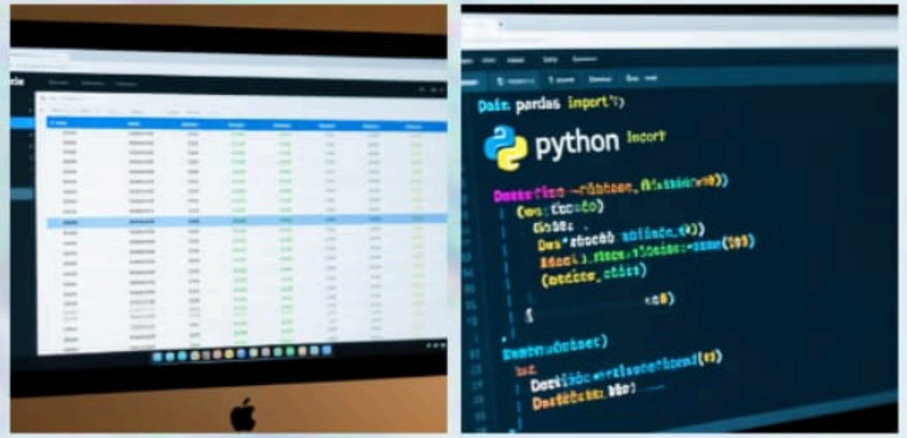
Pandas is Python's premier library for data manipulation and analysis, providing high-performance data structures and tools for working with structured data

DataFrame Objects

Tabular data structure with labeled rows and columns, similar to SQL tables or Excel spreadsheets

Data Cleaning Tools

Comprehensive methods for handling missing data, duplicates, and inconsistencies in datasets



Why Data Cleaning Matters

Strengths of Clean Data

Accurate analysis, reliable insights, better decision-making, and improved model performance

S

W

Risks of Dirty Data

Biased results, incorrect conclusions, wasted resources, and damaged credibility

Common Threats

Missing values, duplicate records, inconsistent formats, and outlier contamination

T

O

Cleaning Opportunities

Feature engineering, data enrichment, quality improvement, and process optimization

Data Cleaning Workflow

Data Inspection

Initial exploration using head(), info(), describe() to understand data structure and quality

Issue Identification

Detect missing values with isnull(), find duplicates with duplicated(), identify outliers

Cleaning Operations

Handle missing data, remove duplicates, correct data types, and standardize formats

Validation & Testing

Verify data integrity, check for remaining issues, and validate cleaning results

Handling Missing Values

Detection Methods

Identify missing data patterns and understand their impact on analysis

- df.isnull().sum() - Count missing values per column
- df.info() - Overview of non-null counts and data types
- Missingno library - Visualize missing data patterns

Treatment Strategies

Choose appropriate methods based on data type and missing pattern

- df.dropna() - Remove rows/columns with missing values
- df.fillna() - Replace missing values with mean, median, or mode
- df.interpolate() - Fill missing values using interpolation

Removing Duplicates & Outliers

Duplicate Detection

Identify and remove duplicate records to maintain data integrity

- df.duplicated() - Find duplicate rows
- df.drop_duplicates() - Remove duplicate entries
- subset parameter - Specify columns for duplicate checking

Outlier Management

Detect and handle extreme values that may skew analysis

- Z-score method - Statistical outlier detection
- IQR method - Interquartile range filtering
- Winsorization - Cap extreme values at percentiles

Data Type & Format Standardization

Type Conversion

Ensure consistent data types for accurate analysis and processing

- pd.to_datetime() - Convert strings to datetime objects
- pd.to_numeric() - Convert strings to numeric types
- astype() - Explicit type casting for columns

String Operations

Clean and standardize text data for better analysis

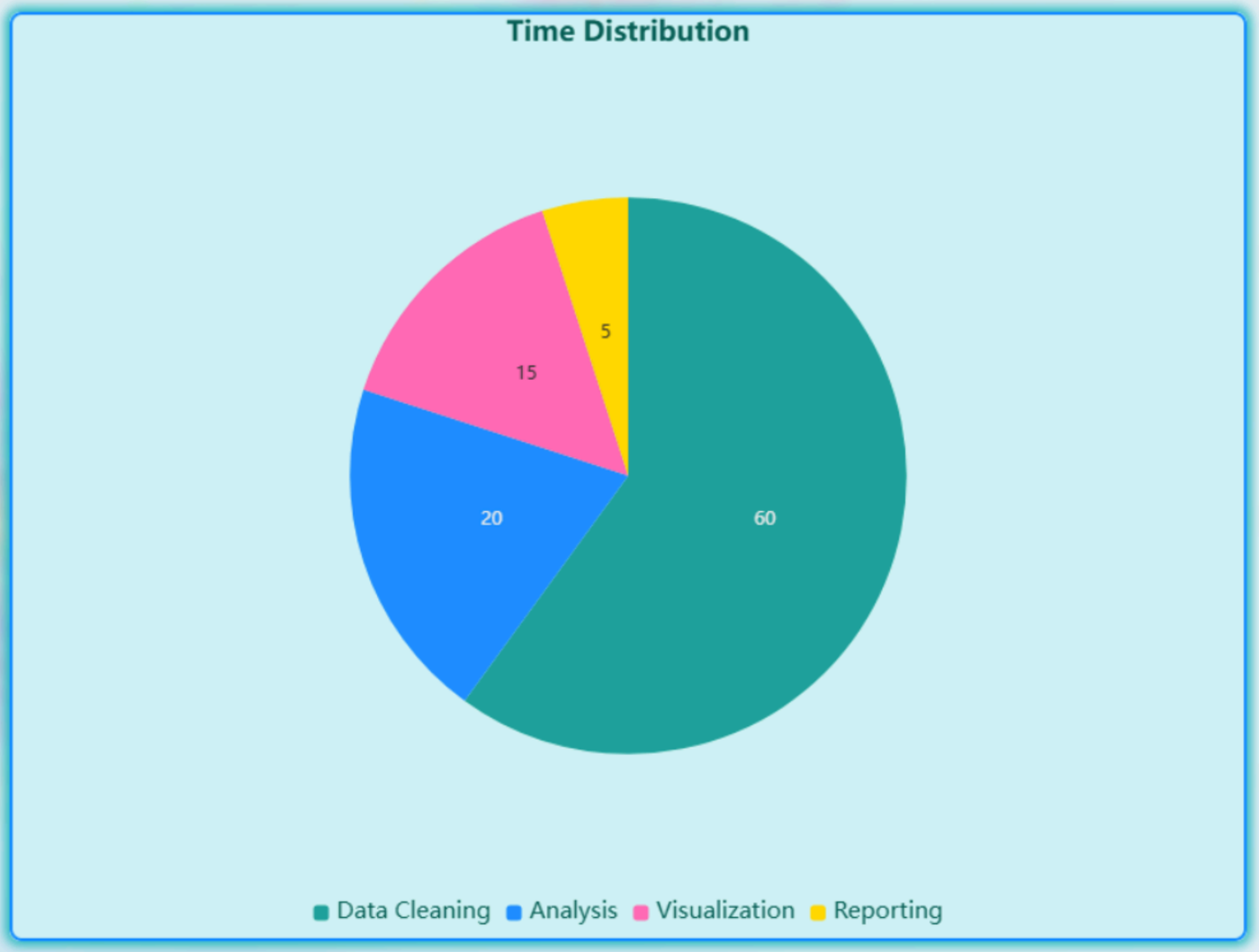
- str.strip() - Remove leading/trailing whitespace
- str.lower() - Convert to lowercase for consistency
- str.replace() - Standardize text formats and patterns

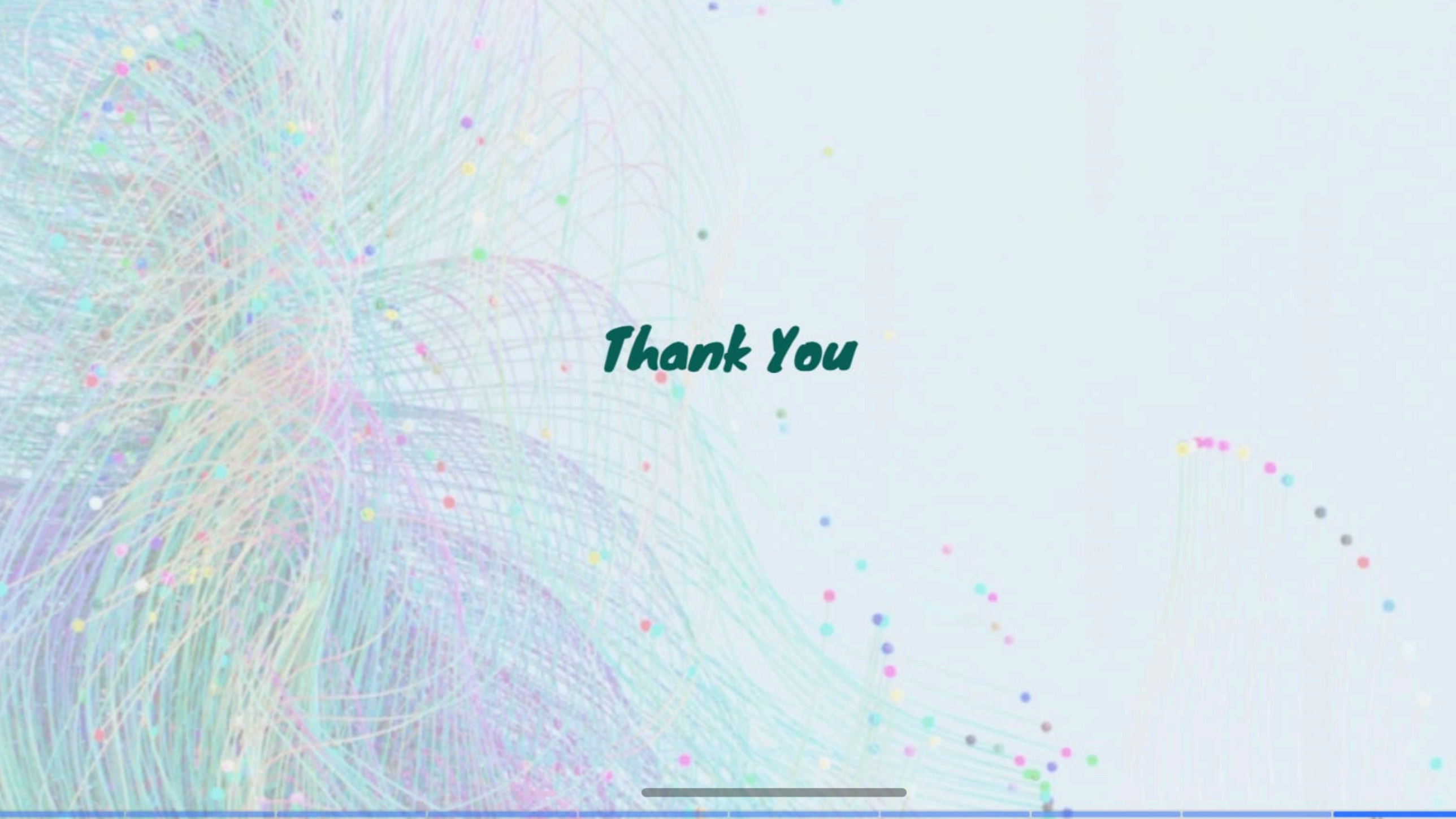
Data Cleaning Impact

Quality Improvement

Effective data cleaning significantly improves analysis accuracy and reliability. Clean datasets lead to more trustworthy insights and better business decisions.

> Studies show that data cleaning can improve model accuracy by up to 30% and reduce analysis time by 40%




Thank You