Master Data Cleaning with Python

Streamline your data preparation process

Pooja Jain

Remove Duplicates

Duplicates in a dataset can skew analysis results. You can remove duplicates using Pandas' drop_duplicates() method.

Pooja Jain

Handle Missing Values

Missing values can impact the accuracy of analysis. You can identify and handle missing values using Pandas' isnull(), dropna(), or fillna() methods.

Pooja Jain

Convert Data Types

Standardizing data formats ensures consistency across the dataset. You can use Pandas' to_datetime() function or custom functions to convert data into a specific format.

Pooja Jain

Handle Data Inconsistencies

Data inconsistencies such as typos or variations in formatting can be corrected using string methods or custom functions.

Pooja Jain

Removing Outliers

Outliers can significantly affect statistical analysis. You can identify and remove outliers using statistical techniques like z-score or IQR method.

Pooja Jain

Data Validation

Validating data against predefined rules or constraints helps ensure data integrity. You can use custom functions or libraries like pandas-schema for data validation.

Pooja Jain

Handling categorical variables

For machine learning models, categorical variables often need to be encoded. You can use Pandas' get_dummies() method for one-hot encoding.

Pooja Jain

Scaling Numerical Features

Some machine learning algorithms require scaled features. You can use scikit-learn's preprocessing module for scaling.

Pooja Jain

In a nutshell

Clean data by removing duplicates, handling missing values, converting data types, and standardizing text.

Pooja Jain

Ready to level up your data cleaning game?

Share your favorite data cleaning tips in the comments below!

Pooja Jain