Introduction to Pandas

Pandas is a powerful Python library for data analysis and manipulation. Pandas provides data structures and functions designed for working with tabular data, making it a popular choice for data scientists and analysts.

by Adamya Neupane

Creating DataFrames

From Lists

Create a DataFrame from a list of lists, where each list represents a row.

From Dictionaries

Create a DataFrame from a dictionary, where keys become column names and values become the data.

Using the pd.DataFrame() function

Create a DataFrame from various data structures like lists, dictionaries, or NumPy arrays.

Reading Data from Files

1

CSV Files

Use the pd.read_csv() function to read data from comma-separated value files.

2

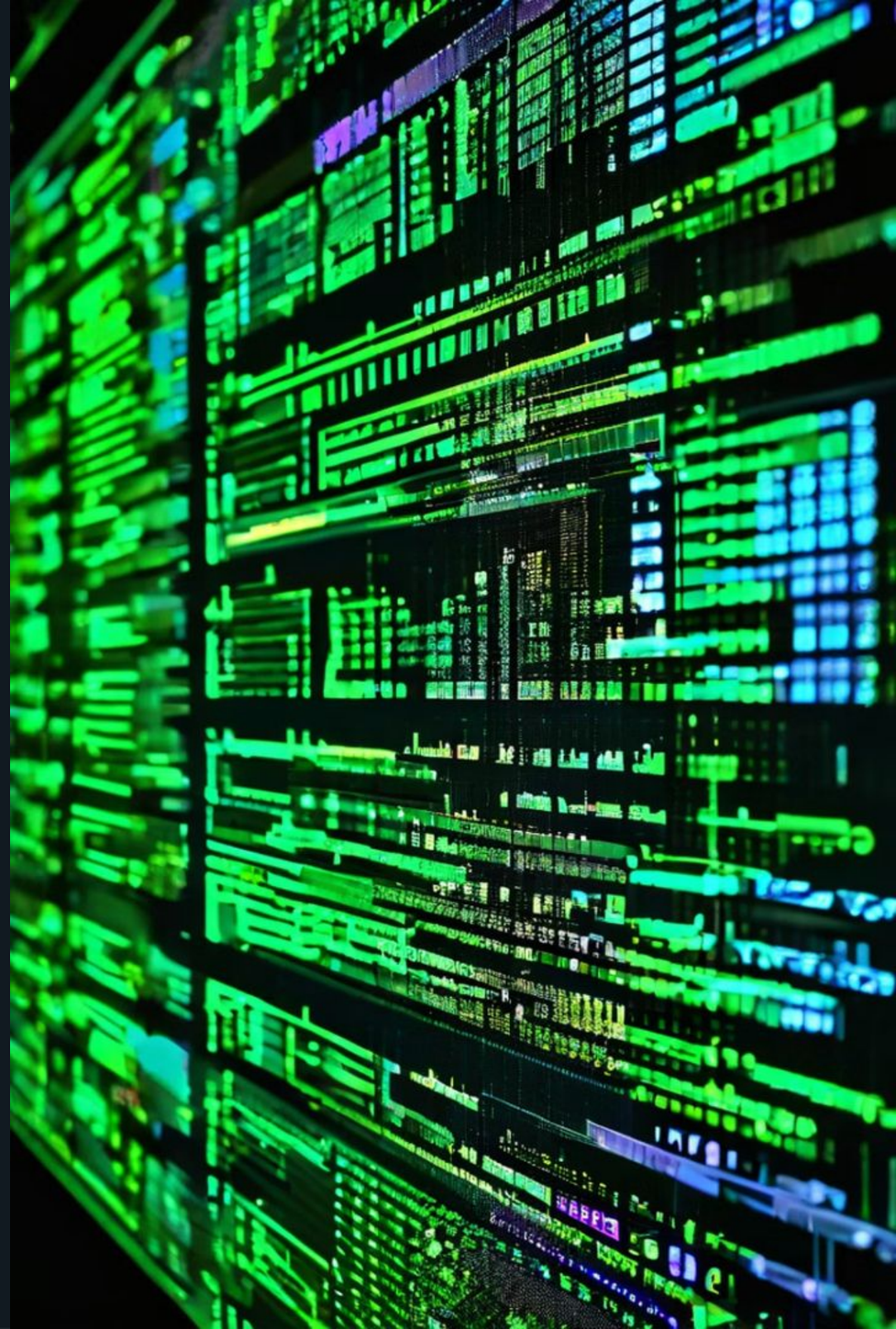
Excel Files

Use the pd.read_excel() function to read data from Excel spreadsheets.

3

Other File Formats

Pandas supports reading data from various file formats, including JSON, HTML, and SQL databases.



DataFrame Operations

1

Selecting Columns

Use square brackets [] to select specific columns by their name.

2

Filtering Rows

Use Boolean indexing to select rows based on specific conditions.

3

Applying Functions

Apply functions to individual columns or the entire DataFrame.

Handling Missing Data

Identifying Missing Values

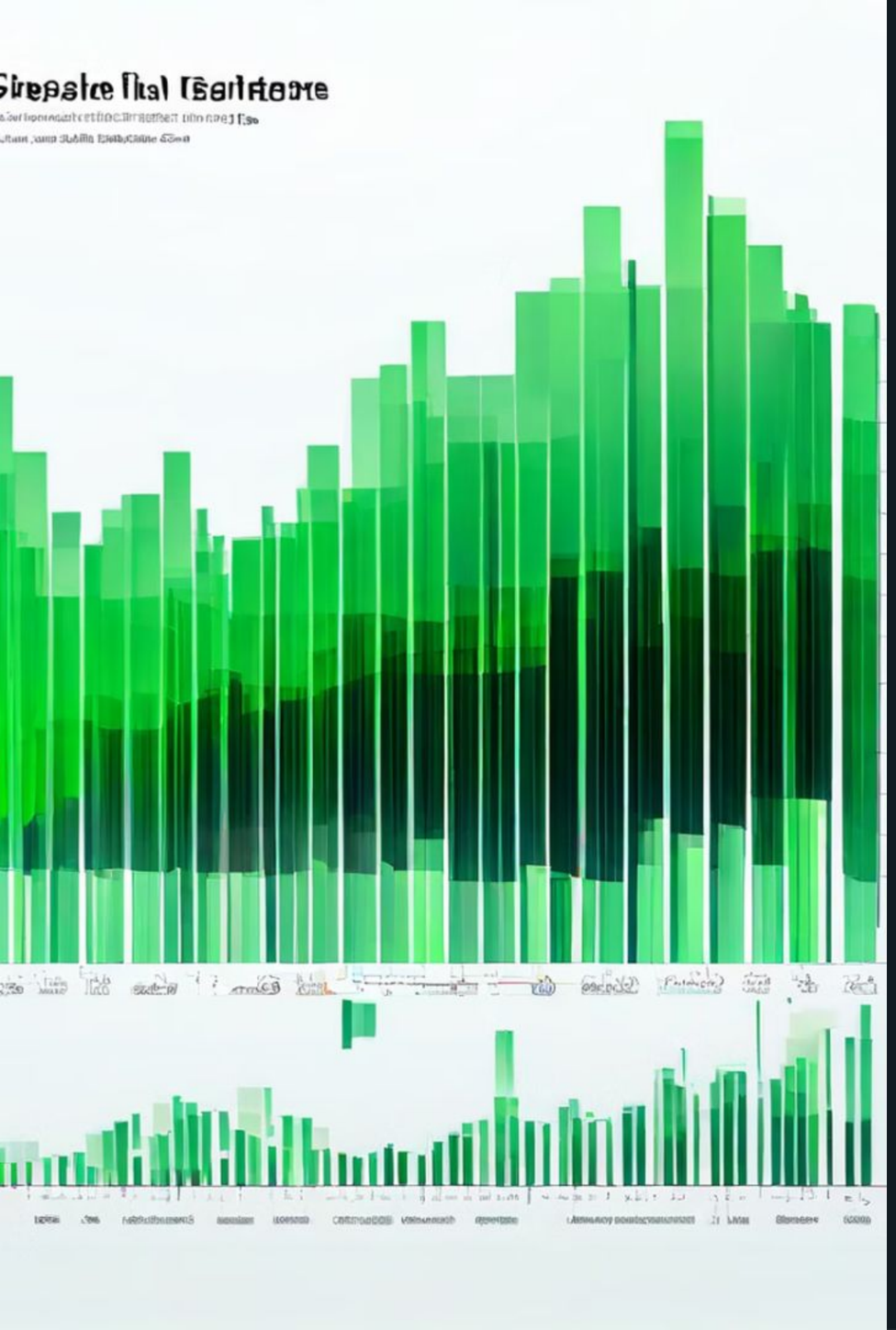
Use methods like .isnull() and .isna() to identify missing values.

Dropping Missing Values

Use the .dropna() method to remove rows or columns containing missing values.

Filling Missing Values

Use the .fillna() method to replace missing values with a specific value or by using interpolation.



Grouping Data

1

Group by Column

Use the .groupby() method to group data based on one or multiple columns.

2

Aggregate Data

Apply aggregation functions like mean, sum, or count to the grouped data.

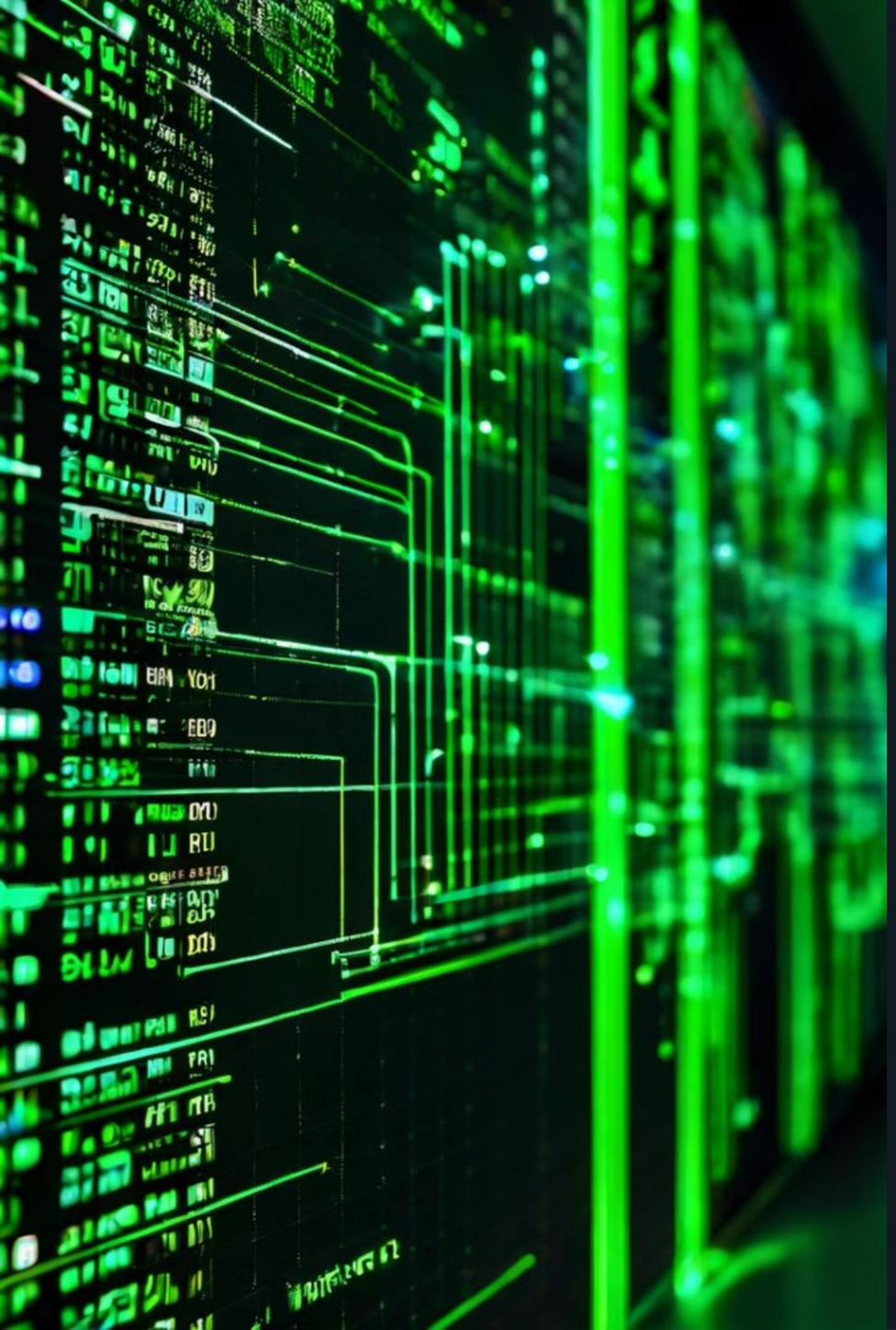
3

Analyze Grouped Data

Explore and visualize the aggregated results for insights and comparisons.

Merging and Joining DataFrames

Method	Description
pd.merge()	Combines DataFrames based on a shared column or index.
pd.concat()	Concatenates DataFrames along rows or columns.
pd.join()	Merges DataFrames on their indices.



Sorting Data

Ascending Order

Use the .sort_values() method with ascending=True to sort data in ascending order.

Descending Order

Use the .sort_values() method with ascending=False to sort data in descending order.

Multiple Columns

Sort by multiple columns by providing a list of column names to the .sort_values() method.

Pivot Tables

1 Multi-Dimensional Summaries

Create a pivot table using the pd.pivot_table() function to summarize data across multiple dimensions.

2 Aggregating Data

Apply aggregation functions like sum, mean, or count to the pivot table data.

3 Data Visualization

Pivot tables provide a structured way to analyze data and can be easily visualized for insights.

Conclusion

Pandas is an essential tool for data scientists and analysts, enabling them to clean, manipulate, analyze, and visualize data efficiently.