DATA SCIENCE PROJECT

Online Retail Dataset: Data Preparation & Preprocessing

Transforming raw transactional data in to a clean, structured, and analysis-ready formatformachinelearning and businessanalytics.

OVERVIEW

Project Workflow

This pipeline follows industry-standard preprocessing steps to prepare the *Online Retail* dataset for downstream analysis.

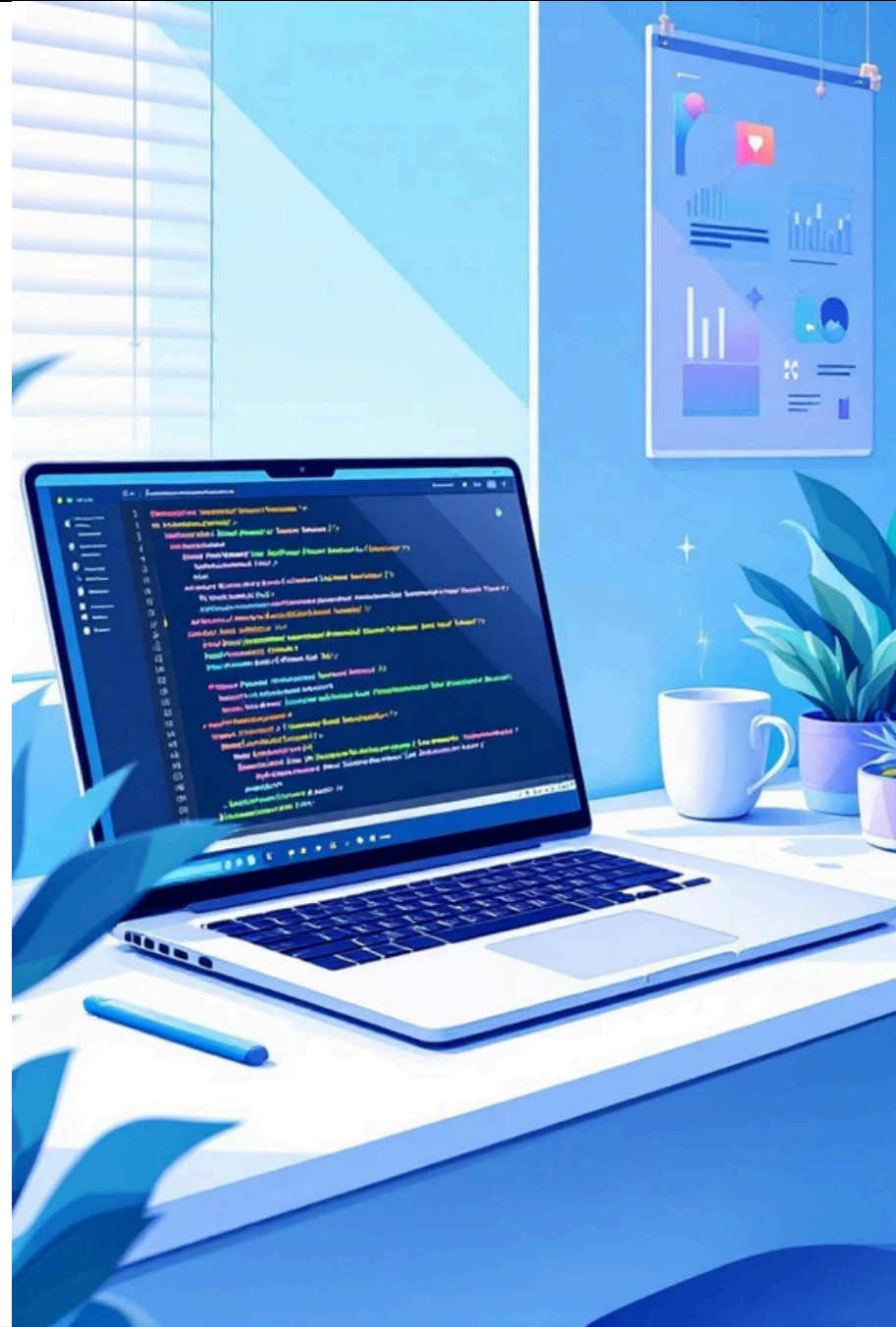
Each stage builds on the previous, ensuring data quality and analytical readiness at every step.

STEP 1

Data Loading

The data set was loaded from an Excel file using the **Pandas** library. Proper verification ensured successful loading before proceeding with preprocessing.

> Verification at this stage prevents cascading errors in later pipeline steps.



STEP 2

Data Understanding & Exploration

Initial exploratory data analysis (EDA)was conducted to understand the dataset's structure and quality.

head()

Displayed first few records

info()

Inspected dataset structure

describe()

Generated statistical summariesIdentified missing values

isnull().sum()

This step revealed inconsistencies, missing values, and required data type corrections.

STEP 3

Data Cleaning

Cleaning Operations

- Removed records with missing `CustomerID`
- Filled missing values where appropriate
- Removed duplicate entries
- Converted `InvoiceDate` to datetime format
- Corrected numerical data types for accurate computation

These steps ensured improved **data quality** and **consistency**.

STEP 4

Data Transformation

Preparing the data set for machine learning required scaling and encoding.

Standardization

Applied `StandardScaler` to normalize numerical features

One-Hot Encoding

Encoded categorical variables into binary columns

Model-Ready

Structured features for unbiased model input

STEP 5

Feature Engineering

1

TotalAmount

Quantity × U nitP rice

2

Date Features

Extracted Year, Month, Day, and Weekday from InvoiceDate

3

Return Indicator

Created a binary feature flagging returned transactions

Feature engineering improved the dataset's **predictive and analytical capability**.

STEP 6

Final Output

Export Details

The cleaned andtransformed dataset was exported as:

`cleaned_normalized_data_fixed.csv`

- **Free from duplicates**
- **Properly formatted**
- **Scaled and encoded**
- **Ready for ML or business analysis**

DOWNSTREAM APPLICATIONS

What This Dataset Enables

Clustering

Customer segmentation and behavioural grouping

Classification

Predict customer churn or purchase behavior

Forecasting

Revenue and demand trend prediction

Dashboard Analytics

Interactive business intelligence reporting

Conclusion

This project demonstrates a **complete data preprocessing pipeline** aligned with real-world data science workflows. The final dataset is structured, reliable, and ready for downstream tasks.

Python & Pandas

Practical implementation

Industry-Standard

Best-practice pipeline

Analysis-Ready

Clean, scaled, encoded