

Data Quality Report After PreProcessing

Initial Data Shape

The data initially had 536641 rows and 9 columns.

Missing Data

InvoiceNo	0
StockCode	0
Description	1454
Quantity	0
InvoiceDate	0
UnitPrice	0
CustomerID	135037
Country	0
Active	0

dtype: int64

As we have a huge dataset, instead of treating missing data, we are directly dropping the rows containing missing values.

Inconsistent Data

We removed 35425 rows where the 'StockCode' contained alphabets or special characters to ensure only numeric values remain.

Incorrect Data

Number of CustomerIDs marked as both Active and Inactive: 1878

We removed 272993 rows where CustomerIDs had conflicting active/inactive statuses.

Data Quality Report After PreProcessing

Preprocessed Data Shape

After all preprocessing, the dataset now has 93186 rows and 9 columns.