

Team Neura

1. What is the source of your dataset?

(Provide website/API/source link)

→ Source of the Dataset

- Website/Source Link: [UCI Machine Learning Repository - Online Retail II](#).
- Alternative Reference: The dataset is also available via the official `ucimlrepo` Python library (ID: 352).
- Donated by: Dr. Daqing Chen, University of Portsmouth.

2. Why did you choose this dataset for your problem statement?

(Relevance and clarity)

→

This dataset was selected for its high relevance to the Grey Market & Illegal Product Sales Detection problem statement due to the following reasons:

- Real-World Transactional Data: It contains 541,909 real transactions from a UK-based non-store online retail firm between 2010 and 2011.
- Multinational Scope: The data includes sales across multiple countries, which is essential for identifying "Parallel Imports"—genuine products diverted from authorized regions to unauthorized markets.

- Price and Quantity Granularity: With specific **UnitPrice** and **Quantity** for every transaction, the dataset allows us to calculate Sales Price Variance. This is a primary indicator of grey market activity, where unauthorized sellers often undercut official prices by 30-60%.
- Wholesale Patterns: A significant portion of the customers are wholesalers. Analyzing these bulk transactions helps identify unauthorized redistribution networks, a key driver of grey market leakage.

3. How was the data collected?

(Web scraping, API, open dataset, etc.)

→

Collection Method: The dataset is an Open Dataset provided for non-commercial research purposes.

Original Acquisition: It was originally collected as a real-time transactional data log from a registered UK non-store online retailer.

Extraction for Project: For this event, the data was retrieved using the UCI Machine Learning Repository API via the **ucimlrepo** library, ensuring a clean and direct ingestion process that adheres to the competition's "No Kaggle" rule.