📘 Technical Documentation

Table of Contents

Project Title: Customer and Returns Analytics for Online Retail (2009–2011)

Tools Used: Python (Pandas, Numpy,Matplotlib, Seaborn), Excel, Tableau Public

🔍 Project Objective

To explore, clean, and analyze an online retail dataset to uncover:
- Revenue trends
- Best-performing products and customers
- Country-wise revenue distribution
- Return behavior patterns
- Top returning customers and return rates

📁 Dataset Description

Source: Online Retail transactional dataset (Excel format)
Period Covered: December 2009 – December 2011
Features: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country
Size: ~500K transactions

🔧 Tools & Libraries

Python:
- pandas, numpy for data cleaning and aggregation
- matplotlib, seaborn for exploratory plots

Excel:
- Export clean summary tables for Tableau

Tableau Public:
- Final visualization and dashboards

🖌️ Step-by-Step Process

1️⃣ Data Loading & Merging

- Loaded two Excel sheets, parsed and merged
- ✅ Challenge: Inconsistent date formats
- 🔧 Solution: Used pd.to_datetime(..., errors='coerce') to fix parsing errors.

2️⃣ Negative Transactions = Returns

- Created a subset df_negative with Quantity < 0 and computed:
  - Top 10 returned products
  - Returns by country
  - Monthly return volumes
  - Top returning customers (by frequency)
- ❗ Important Finding: The dataset includes many rows where the 'Description' field is not a valid product but rather a return reason.
  These include entries such as 'given away', 'ebay sales', 'printing smudges/thrown away', 'unsaleable, destroyed', 'check', 'missing', '?', and blank strings.
  These were not excluded initially but are important to highlight during visualization and interpretation.
  For accuracy, such rows can either be filtered or visualizations renamed as 'Top Returned Products / Return Reasons'.
- ✅ Challenge: Mixed return reasons and unclear products
- 🔧 Solution: Retained and flagged those entries, documented for consideration in analysis (Potential fraud analysis/ anomaly activities analysis)

3️⃣ Data Cleaning

- Removed:
  - Missing customer IDs
  - Negative/zero prices or quantities
- Added:
  - TotalPrice column (Quantity × Price)

4️⃣ Core Exploratory Data Analysis

- Calculated:
  - Monthly revenue
  - Top 10 products by volume
  - Top 10 customers by spending
  - Revenue by country
- ✅ Challenge: Outliers and currency formatting
- 🔧 Solution: Formatted Euro amounts with mtick.StrMethodFormatter('€{x:,.0f}')

5️⃣ Advanced Return Metrics

- Combined purchase and return behaviors to calculate:
  - Return rate by units
  - Return rate by value
- Created new customer-level metrics:
  - ReturnRateUnits = abs(TotalUnitsReturned) / (TotalUnitsPurchased + 1)

📊 Tableau Dashboards

[EDA visualization | Tableau Public]

✅ Dashboard 1: Returns Analysis

- Top Returning Customers (bar chart)
- Most Returned Products / Reasons (bar chart)
- Monthly Returns Volume (bar chart)
- Highest Return Rate Customers (bar chart)
- 💡 Refinements:
  - Used orange color palette for returns
  - Added tooltips for product descriptions
  - Renamed Customer ID field to resolve linking issues

1. KPI – Top Returning Customer

Displays:
- Customer ID with the highest total return count
- Return volume (units)

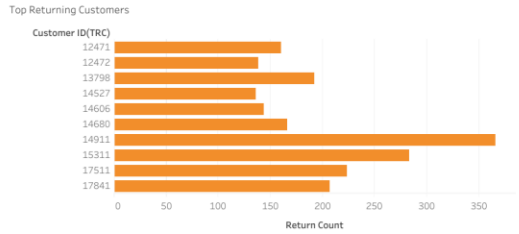
This is implemented using a calculated Boolean field (Is Top Customer) and a text label that dynamically pulls the top customer and their return volume using a fixed filter.

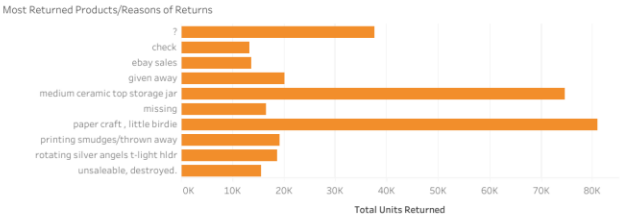
2. KPI – Total Units Returned

Displays:
- Aggregate return count across the dataset
- Serves as a global snapshot of product return behavior

These KPIs provide immediate reference points and summary insights without scrolling

through full visualizations, increasing executive readability and usability for decision-makers.



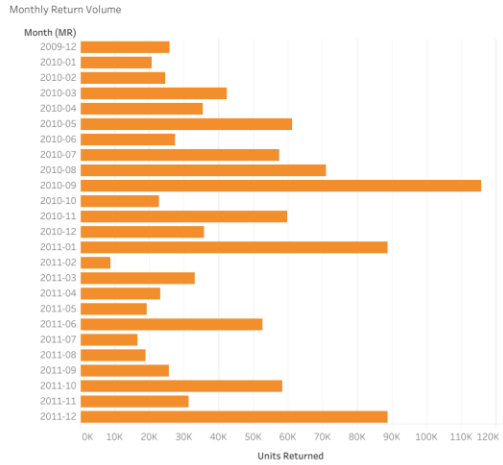


Go to next

TOP RETURNING CUSTOMER
Customer ID: 14911
Returns: 366 Units

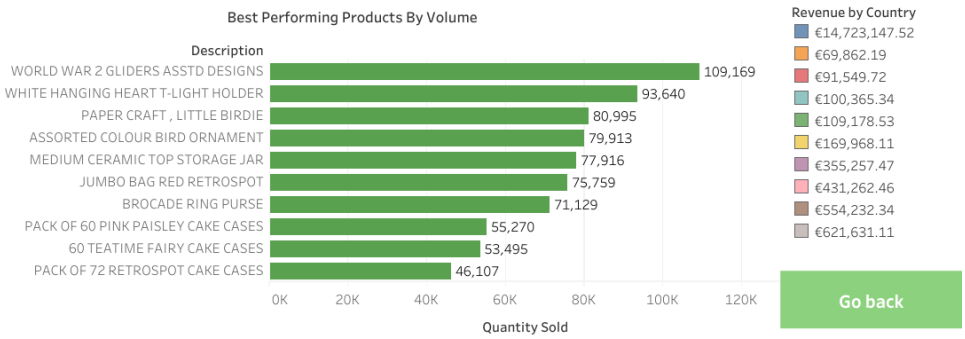
KPI - Total Units Returned

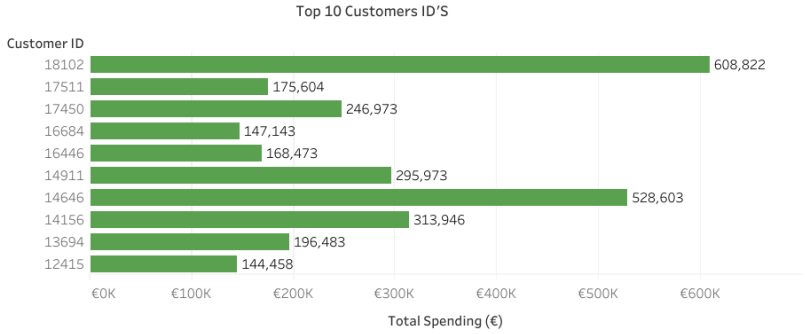
310,150

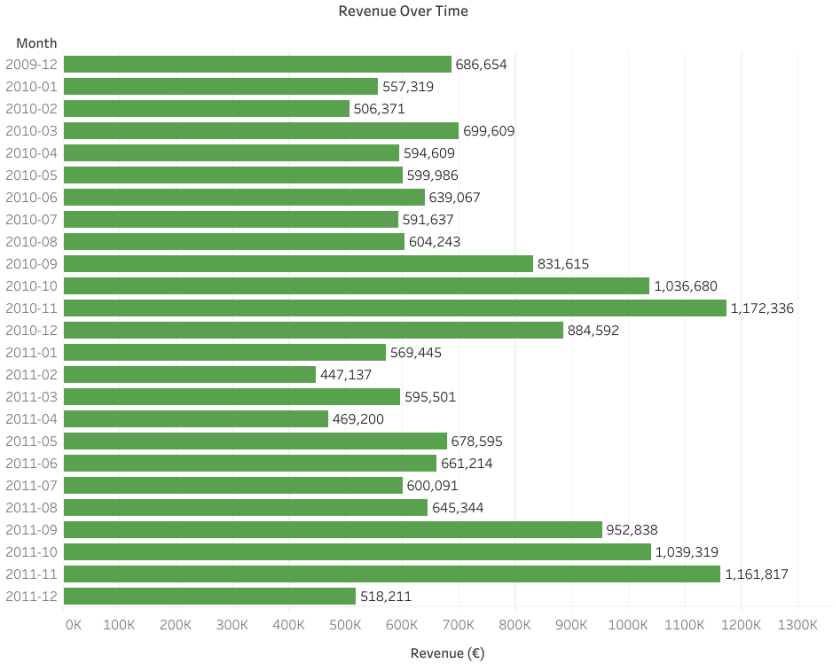


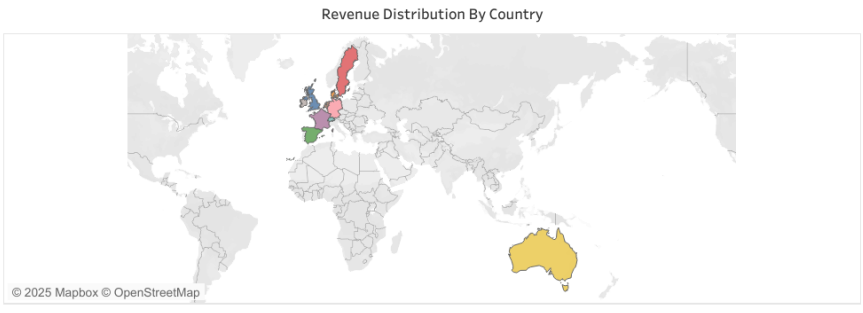
✅ Dashboard 2: Revenue & Sales Insights

- Top 10 Customers by Spending (bar chart)
- Top 10 Products Sold (bar chart)
- Monthly Revenue Trend (bar chart)
- Revenue by Country (map)
- 💡 Design Enhancements:
  - Currency formatting (€#,##0K)
  - Title consistency
  - Country name aliasing for accurate geo-mapping


Best Performing Products By Volume
Description
WORLD WAR 2 GLIDERS ASSTD DESIGNS 109,169
WHITE HANGING HEART T-LIGHT HOLDER 93,640
PAPER CRAFT , LITTLE BIRDIE 80,995
ASSORTED COLOUR BIRD ORNAMENT 79,913
MEDIUM CERAMIC TOP STORAGE JAR 77,916
JUMBO BAG RED RETROSPOT 75,759
BROCADE RING PURSE 71,129
PACK OF 60 PINK PAISLEY CAKE CASES 55,270
60 TEATIME FAIRY CAKE CASES 53,495
PACK OF 72 RETROSPOT CAKE CASES 46,107
0K 20K 40K 60K 80K 100K 120K
Quantity Sold
Revenue by Country
€14,723,147.52
€69,862.19
€91,549.72
€100,365.34
€109,178.53
€169,968.11
€355,257.47
€431,262.46
€554,232.34
€621,631.11
Go back


Top 10 Customers ID'S
Customer ID
18102 608,822
17511 175,604
17450 246,973
16684 147,143
16446 168,473
14911 295,973
14646 528,603
14156 313,946
13694 196,483
12415 144,458
€0K €100K €200K €300K €400K €500K €600K
Total Spending (€)


Revenue Over Time
Month
2009-12 686,654
2010-01 557,319
2010-02 506,371
2010-03 699,609
2010-04 594,609
2010-05 599,986
2010-06 639,067
2010-07 591,637
2010-08 604,243
2010-09 831,615
2010-10 1,036,680
2010-11 1,172,336
2010-12 884,592
2011-01 569,445
2011-02 447,137
2011-03 595,501
2011-04 469,200
2011-05 678,595
2011-06 661,214
2011-07 600,091
2011-08 645,344
2011-09 952,838
2011-10 1,039,319
2011-11 1,161,817
2011-12 518,211
0K 100K 200K 300K 400K 500K 600K 700K 800K 900K 1000K 1100K 1200K 1300K
Revenue (€)


Revenue Distribution By Country
© 2025 Mapbox © OpenStreetMap

-

⚠ Challenges Encountered

- Data Quality: Missing values, junk text, null invoices — cleaned with dropna() and filters
- Tableau Linking: Some columns weren't linked correctly — renamed and verified field types
- Currency Format: Excel-exported data had currency symbols — converted to float before export
- Chart Errors: Bar chart wouldn't render — reclassified fields as measures
- Design Balance: Scrollbars, clipped axes — adjusted number of top entries and layout sizes
- Description Integrity: Return reasons mixed with product names — flagged and described for clarity

☑ Final Deliverables

- 2 Professional Dashboards:

  - 🔁 Returns Performance Analysis
  - 📊 Sales & Revenue Overview
- Cleaned and transformed dataset in Excel format
- Full code base in Jupyter Notebook
- Technical documentation (this file)

📊 Project Summary: Top 3 Insights from the Online Retail Analysis

1. GB High Revenue Concentration in the UK:
   - The United Kingdom generated over €14.7 million in revenue — the vast majority of total sales.
   - This highlights the importance of prioritizing UK-based marketing and logistics strategies.

2. 🎯 Few Customers Drive Majority of Revenue:
   - A small group of customers contributes disproportionately to revenue (e.g. Customer ID 18102 spent €608K+).
   - This supports developing personalized offers and loyalty strategies for top buyers.

3. 🔁 High Return Volume vs. High Return Rate:
   - High returners by quantity (e.g., ID 14911) differ from customers with high return *rates* (e.g., ID 14213).
   - Calls for nuanced return management policies to distinguish volume buyers from risky returners.

💬 Bonus Observations

- 'PAPER CRAFT, LITTLE BIRDIE' and 'MEDIUM CERAMIC TOP STORAGE JAR' were both bestsellers and top-returned items — possibly indicating product issues.
- Several return entries like '?', 'check', 'given away', and 'missing' indicate unclear reasons

or poor labeling in raw data.
- Return spikes in late 2010 and 2011 may reflect seasonal effects or fulfillment inconsistencies.

P.S . **Lessons Learned**

1. **Data Quality Can Mask Hidden Patterns**
   Early in the project, several entries in the product Description field turned out to be return *reasons* (e.g., "check", "missing", "given away") rather than valid product names. Initially treating these as products would have led to misleading visualizations.
   ➤ *Lesson:* Always validate categorical fields for semantic integrity, especially when building dashboards based on them.

2. **Tableau Field Types and Data Structure Matter**
   Some columns (like Total Spent) appeared as strings instead of measures due to currency formatting, which blocked chart generation. Also, Customer ID sometimes defaulted to strings, breaking filters.
   ➤ *Lesson:* Before visualizing, verify data types are correctly interpreted by Tableau. Format numbers, remove currency symbols, and avoid leading/trailing spaces in Excel.

3. **Clear Filtering Enhances Storytelling**
   In early versions, dashboards used full datasets, which made results noisy and scroll-heavy. By limiting to *Top N* customers/products and adding KPI snapshots, the dashboards became far more digestible and executive-friendly.
   ➤ *Lesson:* Curate content purposefully — good analysis isn't just about showing everything, but showing what matters most.

4. **Return Behavior is Multi-Dimensional**
   A single return metric doesn't explain all behavior. This project revealed a difference between customers with *high return volume* vs. *high return rate*.
   ➤ *Lesson:* Understand the business context and use multiple metrics when analyzing customer behavior.

5. **Documentation Matters (A Lot)**
   Keeping a running log of technical and analytical decisions (like whether to exclude unclear return reasons) made it easier to communicate results clearly — especially across Python, Excel, and Tableau.
   ➤ *Lesson:* Maintain structured documentation throughout — it boosts project reproducibility and team collaboration.