Project 2

Online Retail II: Sales Analysis & Customer Segmentation

A SQL and Python project analyzing e-commerce sales, customer behavior, and segmentation strategies

Ironhack Data Science and Machine Learning Bootcamp

Author: Ginosca Alejandro Dávila
Date: December 20, 2024

Project Overview

- **Project Goals:**
  - Analyze sales trends, customer behavior, and product performance using real-world e-commerce data from a UK-based online retailer
  - Build a clean, relational database and use SQL to answer business questions
  - Segment customers using RFM metrics for marketing strategy

- **Project Scope:**
  - Data cleaning and transformation
  - Exploratory data analysis (EDA)
  - SQL-based business insights
  - Customer segmentation

- **Tools & Technologies:**
  - Python (Pandas, Matplotlib, Seaborn, SQLite, dotenv)
  - SQL (MySQL)
  - Tableau
  - Google Colab
  - Jupyter
  - Anaconda Prompt
  - Github

Dataset Summary

🧾 Dataset

- Online Retail II - UCI Machine Learning Repository
- Over 1 million e-commerce transactions from UK-based online retailer

📁 Source File:

- `online_retail_II.xlsx`
- Sheets used:
  - `Year 2009-2010`
  - `Year 2010-2011`
- Date range: 2009-12-01 to 2011-12-09

🔑 Key Variables

- `InvoiceNo`: Transaction ID (prefix ‘C’ = cancellation)
- `StockCode`: Product ID
- `Description`: Product name
- `Quantity`: Number of items purchased per transaction
- `InvoiceDate`: Date and time of the transaction
- `UnitPrice`: Price per unit (GBP)
- `CustomerID`: Unique customer identifier
- `Country`: Country where the transaction occurred

Data Quality Assessment

- **Initial Issues Identified:**
  - ❌ Missing values in `CustomerID` and `Description`
  - ⛔ Canceled invoice (prefix “C” in `InvoiceNo`)
  - ⚠️ Negative or zero values in `Quantity` and `UnitPrice`
  - 🔁 Fully duplicated rows in raw dataset
  - ❗ Non-product `StockCode` (e.g., POST, DOT, BANK CHARGES)
  - 🔄 `StockCode-Description` inconsistencies
  - Conflicting `CustomerID-Country` and `InvoiceNo-InvoiceDate` mappings
  - Duplicate line items within invoices (same `InvoiceNo` and `StockCode`)

- **Assessment Approach:**
  - 🧪 Validated foreign key candidates
  - 📅 Checked data range consistency (2009-2011)
  - 🔍 Verified identifiers and invoice integrity
  - 🧹 Used Python (`pandas`) in `1_data_cleaning_online_retail_ii.ipynb` to inspect , validate, and understand data issues

Data Cleaning Summary

- **Cleaning Steps Performed:**
  - 🔤 Standardized column names (`snake_case`) and categorical values
  - ➖ Removed rows with negative or zero values in `quantity` and `unit_price`
  - ⛔ Dropped canceled invoices (`invoice_no` starting with "C")
  - 🧹 Removed rows with missing `customer_id` and `description`
  - 🆔 Cast identifiers like `invoice_no` and `stock_code` to strings
  - 🔁 Removed fully duplicates rows
  - 💶 Created `line_revenue` as `quantity×unit_price`
  - 🧾 Removed non-product `stock_code` values
  - 🧼 Cleaned and resolved `stock_code-description` conflicts
  - 🌍 Standardized `customer_id-country` mappings
  - 📦 Cleaned invoice metadata: enforced one-to-one mapping between `invoice_no`, `customer_id`, and `invoice_date`
  - 📊 Aggregated repeated invoice items (same `invoice_no` and `stock_code`)
  - 🧱 Normalized into 4 relational tables: `customers.csv`, products.csv, invoices.csv, invoice_items.csv
  - 🔗 Resolved ID conflicts and ensured referential integrity

- **Output Files:**
  - ✅ `cleaned_online_retail_II.csv` (flat file)
  - ✅ 4 normalized relational tables (for MySQL): `customers.csv`, `products.csv`, `invoices.csv` and `invoice_tems.csv`

SQL Project Structure

- **Business Questions (12 total):** Grouped into 4 sections based on analytical focus:

🛒 Sales Performance

- Q1. Monthly Revenue Trend
- Q2. Top 10 Products by Revenue
- Q3. Top 10 Invoices by Transaction Value

🌍 Country and Regional Insights

- Q4a. Revenue by Country (incl. UK)
- Q4b. Revenue by Country (excl. UK)
- Q5. Customer Behavior by Country

👥 Customer Insights

- Q6. One-Time vs. Repeat Customers
- Q7. Top 10 Customers by Avg. Order Value
- Q8. Top 10 Customers by Total Spend

📊 RFM Analysis

- Q9. Recency (Days Since Last Purchase)
- Q10. Frequency (Number of Purchases)
- Q11. Monetary(Total Spend)
- Q12. RFM-Based Customer Segmentation

Project Workflow

- 📊 **EDA Phase:** Python + Pandas on the flat cleaned dataset in `2_eda_online_retail_ii.ipynb`
- 🧱 **SQL Queries in SQLite:** SQL executed via Python in `3_sql_analysis_sales_performance_online_retail_ii.ipynb` on normalized tables
- 🛠️ **MySQL Validation & Final SQL Execution:** Real SQL environment for final results in `1_validate_online_retail_ii.sql` and `2_business_questions_online_retail_ii.sql`

MySQL Environment Setup & Validation

- 🛠️ **Notebook:** `4_mysql_real_env_setup_online_retail_ii.ipynb`
  - 🗃️ Created retail_sales schema (4 relational tables: `customers`, `products`, `invoices`, and `invoice_items`)
  - 📥 **Loaded cleaned CSVs** into MySQL from cleaned relational tables using Python
  - 🔐 **Secure credentials** handled via `.env` + `dotenv`
- 🧪 Initial Validation in Python:
  - ✅ Row counts checks after insertion
  - ✅ Foreign key relationship tests
  - ✅ Checked for:
    - Orphan invoice items
    - Invoice with no customer
    - Invoice with no items
    - Products never sold

Final SQL Validation & Business Questions Execution in MySQL

📄 **SQL Scripts:**

- 🧪 `1_validate_online_retail_ii.sql`
  - ✅ Confirmed shema structure (4 tables: `customers`, `products`, `invoices`, and `invoice_items`)
  - ✅ Verified row counts, primary keys, and foreign keys
  - ✅ Verified invoice date ranges and data consistency
  - ✅ Ran sanity checks on quantities, unit prices, and line revenue
  - ✅ Invoice date within expected range (2009-2011)
- 📊 `2_business_questions_online_retail_ii.sql`
  - ✅ Answered all 12 business questions using MySQL queries
  - ✅ Results matched EDA and SQLite outputs
  - ✅ Simulated real SQL production environment

Business Insights - Sales Performance

📈 Monthly Revenue Trend

- **Revenue peaked in November** of both **2010 (£1.16M)** and **2011 (£1.14M)** reflecting strong **pre-holiday shopping activity**.
- A consistent **post-holiday dip** is visible in **January** and **February**, aligning with typical retail seasonality.
- The **Q4 surge** followed by **Q1 slowdown** highlights clear annual seasonality in consumer behavior.

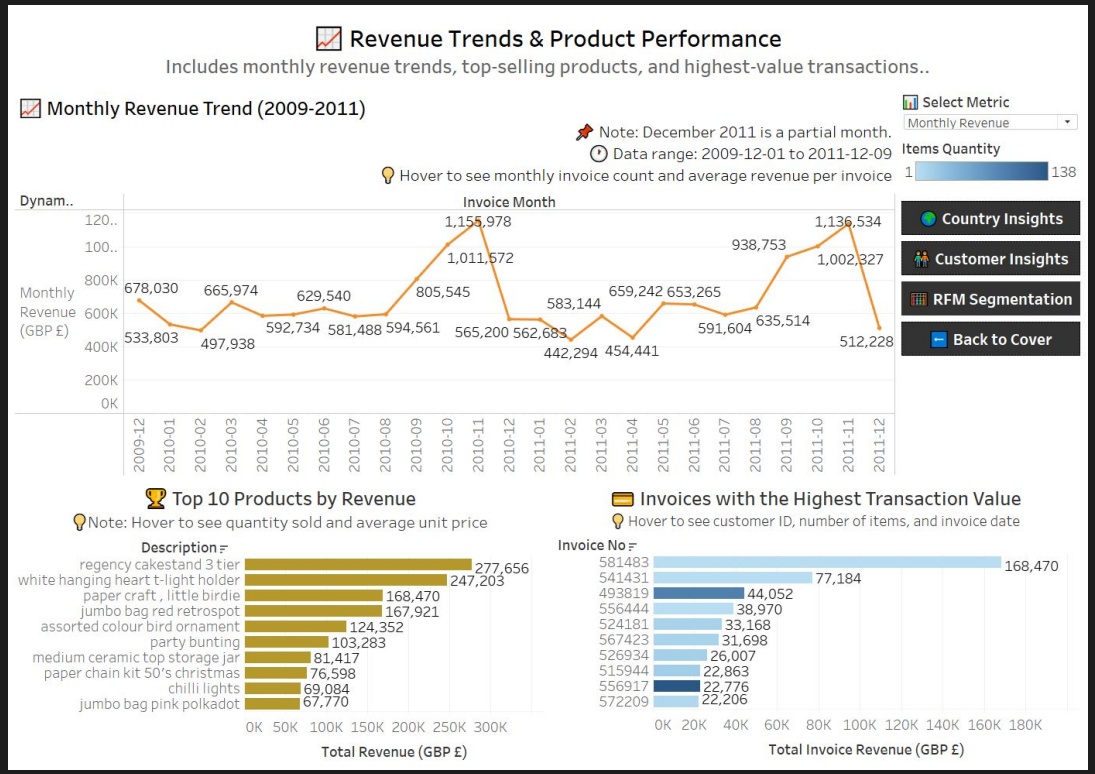
🏆 Top Products by Revenue

- The top product, ***regency cakestand 3 tier***, generated **£277,656.25** in revenue from **24,124 units**, with an average unit price of **£12.46**.
- In second place, ***white hanging heart t-light holder*** earned **£247,048.01** from **91,757 units** - a **low-cost, high-revenue product** that sold exceptionally well.
- Other high-revenue products were **affordable, decorative items** such as: *paper craft, little birdie*; *assorted colour bird ornament* and; *jumbo bag red retrospot*.

💰 Highest Revenue Invoices

- Invoice 581483 generated **£168,469.60** from a single item - an extreme outlier likely a bulk order or possible data anomaly
- Top 10 invoices range from **£22,206.00** to **£168,469.60** in total revenue. Most fall between **£22k** and **£77k**.
- Item counts vary widely - some invoices contain over 130 items, while others list only 1-2 items.
- Customers 18102 and 17450 appear multiple times, suggesting loyal, high value clients with repeated large purchases.

Dashboard - Sales Performance



Business Insights - Country & Regional Insights

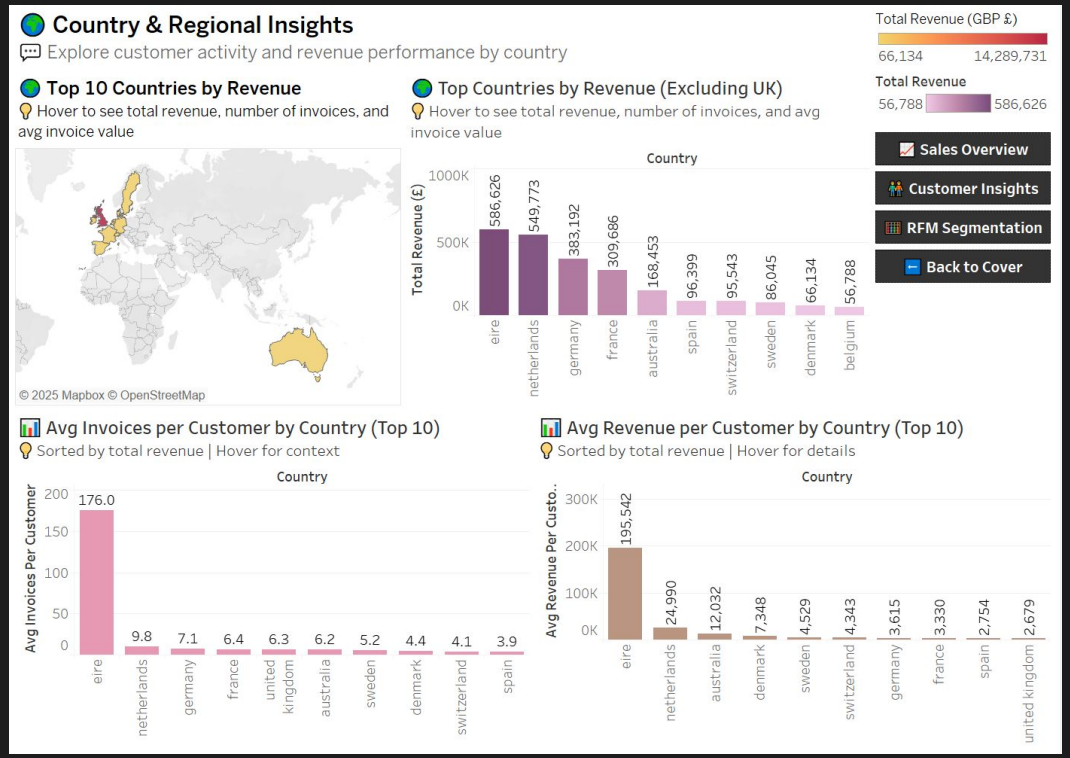
🌍 Revenue by Country

- **UK** generated **£14.29M** across **33,374 invoices** - accounting for **~91%** of all revenue.
- **Ireland (Eire)** leads among non-UK countries, with approximately **£586k** in revenue and a high average invoice value of **~£1,111.**
- The **Netherlands** tops in average invoice value at **~ £2,545**, despite a moderate volume of 216 invoices.
- Countries like **Australia**, **Switzerland**, and **Denmark** generate significant revenue from **fewer than 100 invoices** each.
- **Germany** and **France** show a healthy balance of volume and value, indicating mature and stable market performance.
- Markets such as **Netherlands**, **Australia**, and **Denmark** are strong candidates for **B2B or premium expansion** due to high-value, low-frequency trends.

👥 Customer Behavior by Country

- The **UK** has **5,334 customers** with an average of **6.26 invoices** and **£2,679 revenue per customer** - our largest and most balanced group.
- With only **3 customers**, Ireland shows **exceptionally high values** (**~£195 revenue** and **176 invoices per customer**) - likely due to **atypical client behavior** such as bulk purchasing client, internal use account, or testing/demo user.
- **Netherlands** has strong potential market with **£24.9K per customer** and **~10 invoices/customer** - indicating high engagement and value.
- **France and Germany** show a **balance customer base** and solid per-customer revenue (**~£3.3K-£3.6K**)
- **Sweden, Switzerland, Denmark**, and **Australia** show **low-frequency but high value** purchasing behavior - ideal for **premium offerings** or **B2B targeting**.

Dashboard - Country & Regional Insights



Business Insights - Customer Behavior

🔁 One-Time vs Repeat

- **Nearly 3 out of 4 customers** made **multiple purchases,** suggesting **decent customer retention.**
- Around **28% of users** only purchased **once**, which presents an **opportunity for re-engagement campaigns**.

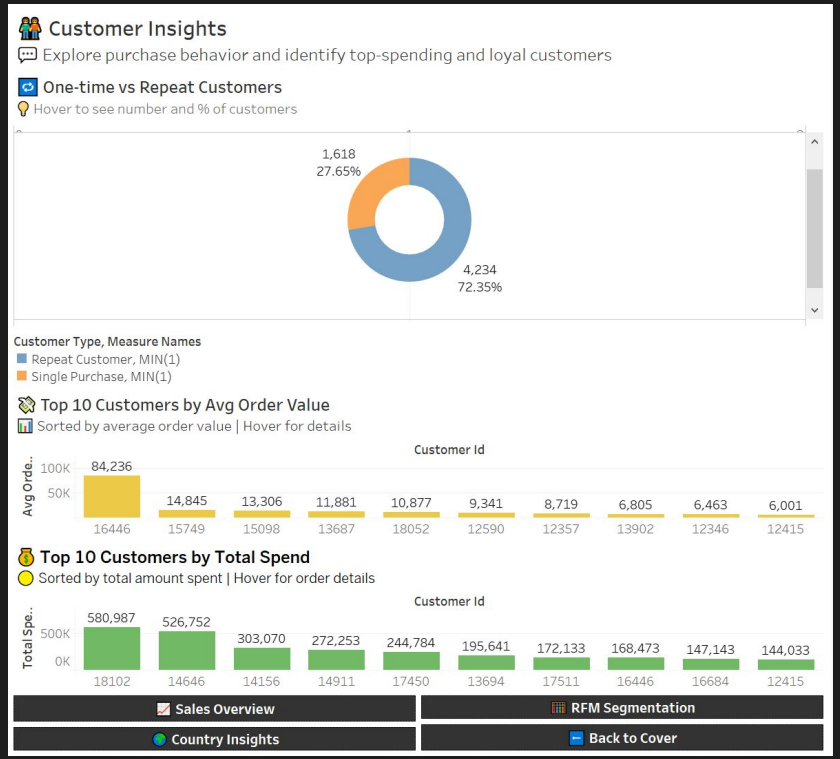
👥 Top Customers by AOV

- Customer **16446** stands out with an average order value of **£84,236.25**, placing just two orders - a clear outlier likely reflecting bulk or business purchases..
- Other high-ranking customers, such as **15098** and **15749**, average between **£13.3K** and **£14.8K** per order.
- While these values are impressive, **99.8% of customers have an AOV under £5,600**, showing a strongly right-skewed distribution with a few extreme cases.

💰 Top Customers by Total Spend

- **Customer 18102** tops the list with **£580,987** across 145 orders.
- Several others (e.g. 14646, 14156) show high revenue from **consistent repeat ordering**.
- Customer **16446** stands out with just 2 orders but **£168K total spend** - likely a **bulk B2B transaction or anomaly**.
- These customers warrant **special attention** for retention, loyalty programs or custom offers.

Dashboard - Customer Behavior



Business Insights - RFM Metrics

📅 Recency (R)

- Measures days since a customer's last purchase (relative to 2011-12-09)
- Most customers made their most recent purchase within the **first 100 days**
- Recency spans from 0 to 738 days.
- Some customers haven't purchased in over 2 years - indicating churn risk
- Distribution is **right-skewed**, with many recent buyers and a few long-inactive users.

🔁 Frequency (F)

- Number of unique purchases per customer (invoice count).
- Most users placed **1 to 8 orders**, with some outliers exceeding **100+ orders**.
- The spread reveals a core of casual shoppers and a small group of highly engaged customers.

💰 Monetary (M)

- Total revenue generated per customer.
- Over 99% of customers spent less than **£20K**.
- A few outliers - including one at **£580K** - generate a large share of total revenue.
- Distribution is highly right-skewed, emphasizing top-value clients.

Business Insights - RFM Segmentation

📊 RFM Metrics

- Each customer was scored from 1 (low) to 4 (high) on:
  - **Recency** → Days since last purchase (as of 2011-12-09)
  - **Frequency** → Number of unique purchases
  - **Monetary** → Total spending

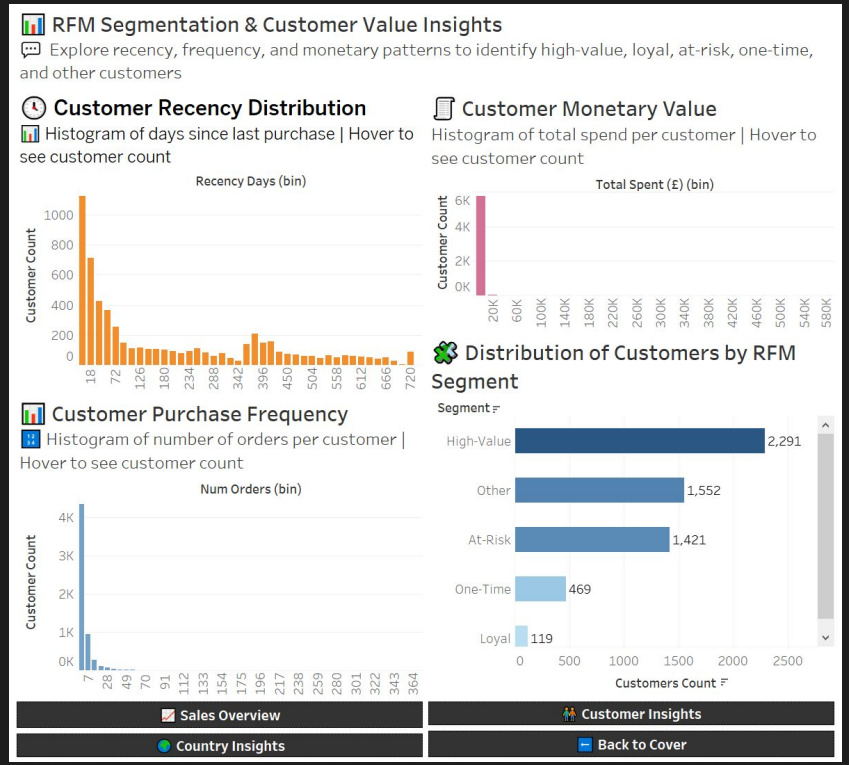
🧩 Segment Breakdown

- Customers were grouped into key behavioral segments:
  - 🟡 High Value → Big Spenders with frequent orders
  - 🟢 Loyal → Recent and frequent buyers
  - 🔴 At risk → Long inactive but previously valuable
  - 🔵 One-Time → Single-purchase, low-engagement users
  - ⚪ Other → Doesn't meet specific criteria for grouping

📌Key Insights

- **High-Value** segment is largest - major revenue driver.
- **Other** and **At-Risk** segments are the next largest groups.
- **Loyal** and **One-Time** customers form smaller but strategically important segments.
- This distribution reveals a **diverse customer base** with distinct behavioral patterns.

Dashboard - RFM Analysis



Business Recommendations

📌 Key actions based on Python and SQL analysis:

- **Prioritize High-Value Customers**
  - Reward top spenders with VIP perks, early access, or loyalty programs to boost retention.
- **Retain Loyal & Re-Engage At-Risk Buyers**
  - Use targeted email campaigns, offers, and incentives to keep loyal users engaged and bring back long-inactive customers.
- **Convert One-Time Buyers**
  - Design onboarding flows and personalized discounts to turn first-time shoppers into repeat customers.
- **Capitalize on Seasonal Peaks**
  - Strengthen marketing and inventory strategies ahead of Q4, especially October-November, to maximize holiday revenue.
- **Grow in High-Value Markets**
  - Expand marketing in Ireland, Netherlands, Germany, and explore premium strategies for Switzerland and Australia
- **Optimize Product Strategy**
  - Double down on bestsellers and seasonal items. Test bundling and complementary offers to increase average order value.

Future Steps

🧭 Strategic Next Steps

- **Customer Lifetime Value (CLV) Modeling**
  - Predict long-term value per segment to guide budget allocation and retention strategies.
- **Cohort & Churn Analysis**
  - Track behavior of customer groups over time and identify early signs of churn for proactive.
- **A/B Testing & Campaign Tracking**
  - Run experiments on emails, offers, and loyalty perks to optimize conversions and ROI.
- **Product Affinity & Bundling**
  - Analyze purchase patterns to identify frequently bundled items and improve cross-selling opportunities
- **Personalized Marketing**
  - Use RFM segments and buying behavior to tailor messaging, promotions, and product recommendations.

Limitations

- **Data is Historical (2011)**
  - The dataset ends in December 2011, so customer behavior trends may not reflect current realities.
- **Lack of Demographic Information**
  - No age, gender, or region data - limiting personalization and advanced customer segmentation.
- **Unclear Customer Type (B2B vs B2C)**
  - The dataset does not distinguish between business and individual buyers, which could affect segmentation logic.
- **Limited Product Classification**
  - No structured product categories - insights are based on item descriptions and stock codes only
- **RFM Threshold Assumptions**
  - Segment labels are based on quartile logic, which may not fully capture business-specific nuances.

Thank You!

Questions?

Contact:

👩‍💻Ginosca Alejandro Dávila

📍 Data Analyst & Educator
🔗 linkedin.com/in/g-alejandro