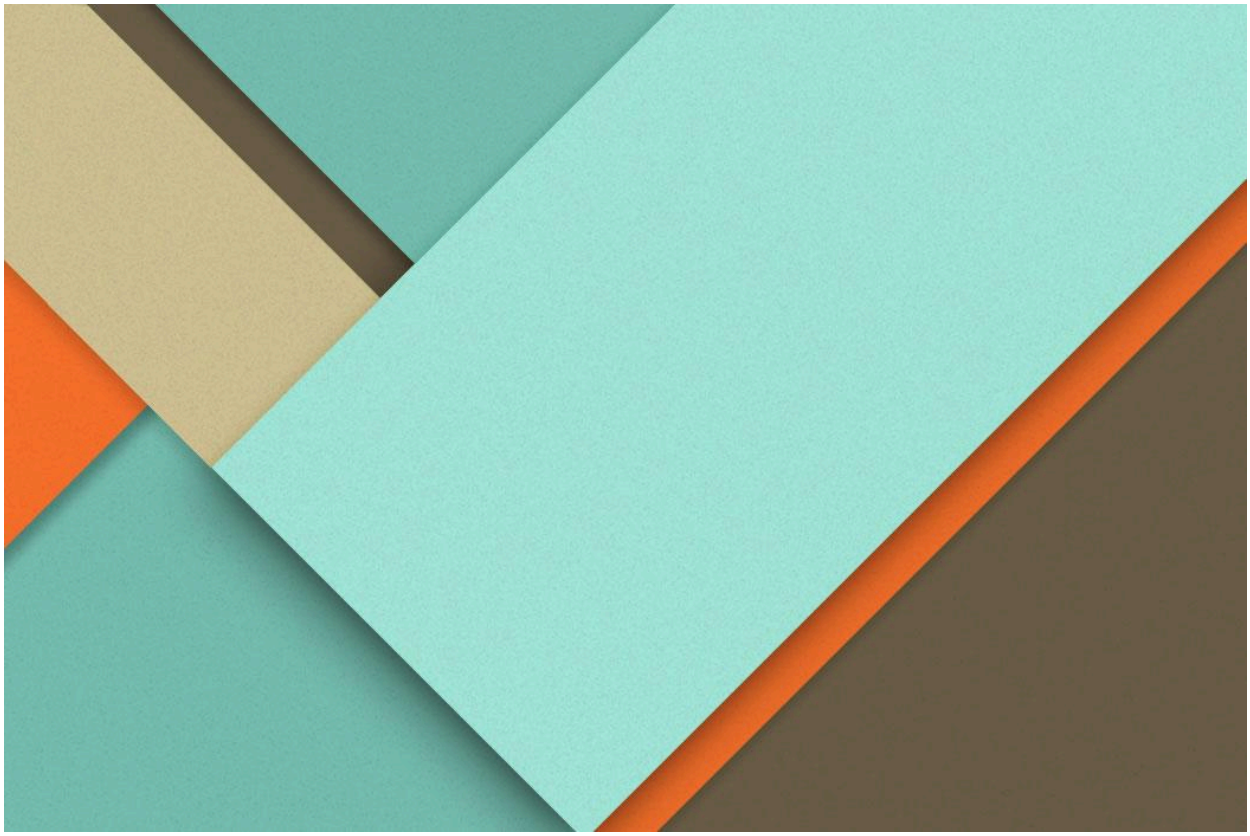

Online Retail Transaction Analysis

Data Analytics with AI Individual Formative Assignment One: Python ETL and Visualisation

Data Extraction, Transformation, and Loading (ETL), Data Visualisation, and Agile Project Management.

Helen Bliss

11.11.25

Executive Summary

This project analyses a year of wholesale retail data to understand sales performance, product behaviour and customer dynamics.

Revenue trends show steady growth throughout the year, with a sharp seasonal peak in the final quarter driven by Christmas trade purchasing. Product analysis reveals a strong reliance on low-priced, fast-moving items, with most products falling under £5 and only a small number contributing disproportionately to total sales.

Customer behaviour is highly varied: most accounts place occasional low-value orders, while a smaller group of trade customers purchase frequently and contribute the majority of revenue.

An RFM segmentation model was developed to classify customers into six groups, from Top Customers to Dormant Accounts, highlighting clear differences in recency, frequency and monetary value.

These insights provide a foundation for targeted strategies, including prioritising top accounts, nurturing high-value and regular buyers, stabilising steady buyers, and implementing win-back campaigns for at-risk or dormant customers.

Overview

Data source:
https://www.kaggle.com/datasets/abhishekrp1517/online-retail-transactions-dataset

The "Online Retail Transaction" dataset contains information on customer transactions made through an online retail platform. It includes data on products purchased, quantities, transaction dates and times, prices, customer identifiers, and customer locations. This dataset can be used to analyse customer behaviour and preferences, identify popular products, and optimise pricing and marketing strategies.

Goals

Data analysis goals:

- Analyse online retail transaction data to understand customer behaviour
- Provide insights into customer behaviour for marketing output.

Personal goals:

- Learn how to clean a large dataset of 55,000 records
- Complete my first coding project
- Learn from trial and error

Required Outcomes

- Design and implement an ETL pipeline using Python.
- Visualise data using Matplotlib, Seaborn, and Plotly.
- Adhere to the key milestones and deliverables.
- Maximise future maintainability through documentation, code structure, and organisation.
- Document and present the project process and outcomes.
- Demonstrate and document the development process through a version control system such as GitHub.

Project overview requirement

ETL Pipeline:

1. Extract: Load data from the provided dataset.
2. Transform: Clean the data, handle missing values, encode categorical variables, and create new features such as total transaction value.
3. Load: Store the transformed data in a format suitable for analysis (e.g. a cleaned DataFrame).

Data Visualisation:

1. Descriptive Statistics: Display basic statistics such as total sales, average transaction value, and number of unique customers.
2. Trend Analysis: Visualise sales trends over time.
3. Customer Segmentation: Segment customers based on purchase behaviour and visualise their characteristics.
4. Product Analysis: Identify and visualise the most popular products.

Project checklist

Assessment Criteria	Minimum Score	Maximum Score
Scoring	0.9 (1 for custom project)	4.5 (5 for custom project)
ETL extraction	No data	All CSV, Excel or API data was extracted correctly.
ETL data load	No data loaded	Data loaded into a suitable format for Python.

ETL transformation	No data transformation	All data is logically cleaned, preprocessed and transformed into the correct format for analysis
ETL data quality and integrity	No improvement in data quality	Data is of high quality/integrity and has been validated
ETL data documentation	ETL process not documented	Documentation for the ETL process, including data sources and transformation logic.
Static visualisation	It is not clear what is visualised	Clear, easy-to-view basic visualisation such as histograms, bar charts, and line plots using Matplotlib.
Advanced visualisations	It is not clear what is visualised	More complex visualisations like pair plots, heatmaps, and violin plots using Seaborn.
Visualisation interactivity	No purpose for visualisation	Interactive Plotly visualisations
Visualisation customisation	Only standard chart type	Customise plots with titles, labels, legends, and annotations for better clarity.
Visualisation relevance	Visualisation not relevant to the report's purpose	Each visualisation is important to the report
Project planning	No plan	Project task process documented using project management tool (Kanban board)
Project documented	No documented tasks	Project tasks and user stories are documented within the project management tool (Kanban board)
Project adaptation	No flexibility	Evidence of adjusting the project plan based on feedback and changing requirements.
Project review	No review	Evidence of conducting reviews to assess progress and make necessary adjustments.
Project reflection	No reflection on progress	Evidence of performing retrospectives to identify lessons learned and areas for improvement.

README file	No README	The README explains the steps taken to create the visualisation and how to use them. It also explains the project, its objectives, and how to reproduce it.
Presentation of process	No mention of how the learner got from data to the visualisations	Clear presentation of the deliverables and milestones on the visualisation creation process. Evidence of version control.
Project naming	Poorly named files and functions.	Name files consistently and descriptively, without spaces or capitalisation, to allow for cross-platform compatibility. Organise Jupyter Notebook code into well-defined and commented sections
Code quality	Poorly written code	Write code that meets at least minimum standards for readability (consistent indentation, blank lines only appear individually or, at most, in pairs). Ensure all code from external sources is properly attributed and referenced.
Presentation of findings	No conclusions on the data.	Present the story of what key findings were made with the data.

ETL Pipeline planning

1. Basic cleaning:

- Clean the text in each column to remove any spaces etc
- Tidy title casing
- Review data types and convert as needed (e.g. change InvoiceDate column to 'datetime')
- Review initial missing data across the dataset (NaN).

2. Product validation:

- Handle invalid description/stock codes - remove when validated as invalid.
- Address the stock codes with multiple product descriptions (x650 at this stage). De-dupe records /replace text as required.

3. Add sales features:

- Create new column 'Sales Total' (quantity * unit price)
- Add columns for sales month and year, create basket value column
- Identify rows with zero 'Unit Price' values and remove (complete after sales total column is created).

4. Final validation checks.

5. Data enrichment and customer features:

- Aggregate customer records
- Customer segmentation to enable analysis of lapsed customers, repeat customers, multibuy customers.

Data Visualisation approach

REVENUE ANALYSIS

What do I want to find out?

- How has revenue changed over time?
- When are the peak and low sales periods?
- How do basket values vary across orders?

Rationale:

- Supports forecasting
- Helps identify seasonal patterns
- Reveals shopper spending behaviour

Insight	Visual	Library	Chart Type	Why it's useful
Revenue trend over time	Daily/Monthly revenue line chart	**Matplotlib**	Line Plot	Shows seasonality & performance trends
Distribution of basket values	BasketValue histogram	**Matplotlib**	Histogram	Reveals typical order sizes & spending spread
Smooth spend distribution	BasketValue KDE	**Seaborn**	KDE	Shows density and skew of order values

PRODUCT PERFORMANCE

What do I want to find out?

- Which products drive the most revenue?
- How do unit price ranges segment into budget bins?
- Which products perform best in each country?

Rationale:

- Highlights key revenue drivers
- Supports pricing strategy
- Shows product popularity by region

Insight	Visual	Library	Chart Type	Why it's useful
Top revenue-generating products	Interactive top 10 bar chart	**Plotly**	Bar Chart	Easy to explore & compare products
Product performance by country	Country → product treemap	**Plotly**	Treemap / Sunburst	Visual hierarchical insight
Product price bands	Price bin countplot	**Seaborn**	Countplot	Clear segmentation of product pricing

Price vs quantity relationship	Scatter plot	**Seaborn**	Scatter	Shows elasticity / buying patterns

CUSTOMERS & BUYER BEHAVIOUR

What do I want to find out?

- What segments do customers fall into (Active, Repeat, Lapsed, MultiBuy)?
- How frequently do different customer segments purchase?
- Who are the high-value customers?
- What does recency look like (lapsed intervals)?

Rationale:

- Supports targeted marketing
- Helps retention strategies
- Identifies CLV opportunities

Insight	Visual	Library	Chart Type	Why it's useful
Segment distribution	Segment count bar chart	**Seaborn**	Bar Plot	Shows structure of customer base
Spend differences per segment	BasketValue by segment	**Seaborn**	Boxplot	Highlights behavioural variation
Recency (lapsed intervals)	Days_since histogram	**Matplotlib**	Histogram	Shows how recently customers purchased
Recency distribution	Recency KDE	**Seaborn**	KDE	Shows smooth recency trends
Top spending customers	Top 10 customer bar chart	**Matplotlib**	Bar Chart	Classic CLV identification

Customer value distribution	TotalSpend KDE	**Seaborn**	KDE	Shows high-value vs low-value profiles

Data analysis findings

1. Sales Performance Overview

The sales analysis shows a strong overall performance, with **global revenue increasing steadily over time** and clear seasonal patterns. Average revenue per day also rises in line with total revenue, indicating consistent customer demand rather than a small number of extreme days driving performance.

Looking at the UK specifically — the largest and most active market — revenue follows a similar trend to global sales. UK demand builds through the year before rising sharply in **October, November and early December**, highlighting the importance of Christmas trading in this wholesale business.

The **quantity vs revenue scatter plot** confirms a typical wholesale pattern: most transactions involve **high quantities of low-priced items**, while a smaller number of orders generate much higher revenue through bulk purchasing.

Seasonality analysis reinforces this trend, with a substantial increase in revenue and order count in the final quarter of the year. The heatmap and monthly line charts make it clear that **Q4 is the key trading period**, while early-year months tend to be subdued.

Basket-level analysis shows that **most orders are low-value**, but there is a clear **long-tail of higher-value baskets**, reflecting occasional large trade purchases. The capped basket-value distribution reveals a highly skewed pattern, consistent with wholesale behaviour.

Overall, the sales summary points to a stable business with predictable peaks, heavily driven by trade purchasing cycles and seasonal restocking activity.

---

2. Product Insights

The **Top 10 UK products** highlight a strong reliance on a small group of fast-moving, low-cost gift and homeware lines. These items drive a significant share of revenue and order volume, indicating a classic "core range" pattern in wholesale retail.

The **unit price band analysis** shows a clear structure in the product catalogue:

- most products fall under **£5**,
- a meaningful group sit between **£5–£20**,
- and only a very small number exceed **£20**.

This pricing distribution supports high-volume, low-unit-value purchasing typical of wholesale giftware. It also explains the long-tail effect visible in the basket analysis.

---

3. Customer Behaviour and Segmentation

Spend and Customer Contribution

Customer spend analysis shows a pronounced long-tail pattern. A small number of trade accounts contribute a disproportionately large share of revenue, while many smaller accounts place occasional, low-value orders. The Top 10 UK customers demonstrate this clearly, with the highest-value accounts significantly ahead of the rest.

Order Patterns

The analysis of **order value by day of week** reveals a strong weekday focus, with no Saturday trading and extremely limited Sunday activity — consistent with B2B purchasing behaviour.

The **frequency vs orders** and **best-to-worst frequency ranking** charts show that customer order behaviour varies widely: while most customers place only a handful of orders, a minority place regular and repeated transactions throughout the year.

Customer Segmentation & RFM Behaviour

The segment distribution chart shows a wide spread of customers across the lifecycle, with clear distinctions between:

- **Top Customers** — recent, frequent, and high-spending
- **High-Value Customers** — strong monetary value with semi-regular activity
- **Regular Buyers** — consistent mid-level behaviour
- **Steady Buyers** — mixed engagement and moderate volumes
- **At-Risk Accounts** — low frequency and long recency
- **Dormant Customers** — inactive trade accounts

The recency and frequency boxplots highlight these behavioural differences clearly. Top and High-Value accounts show strong recency and frequency, while At-Risk and Dormant customers cluster towards long recency and minimal orders.

Monetary analysis reinforces these patterns, showing that spend is heavily concentrated among the top tiers.

Overall, customer behaviour follows a typical wholesale lifecycle: **a few highly engaged trade accounts**, a middle tier of reliable buyers, and a long tail of occasional or lapsed customers.

Recommended Actions by Segment

Top Customers

Prioritise with enhanced service, early access to new ranges, and proactive account management. Protecting these accounts is commercially vital.

High-Value Customers

Encourage regularity through restock prompts, range updates, and seasonal buying guidance. These customers can easily move into Top status with light nurturing.

Regular Buyers

Promote growth through cross-selling, add-on suggestions, and multi-buy incentives. This segment is the most promising for revenue development.

Steady Buyers

Identify early signs of disengagement. Offer trial packs and re-engagement messaging to stabilise buying patterns and prevent decline.

At-Risk Accounts

Use win-back campaigns, personal outreach, and promotional triggers. Some can be recovered with targeted prompts.

Dormant Customers

Light-touch seasonal reactivation may work for a small proportion, particularly Christmas-only buyers. Most should be deprioritised to maintain database quality.

Did I achieve my goals?

Data analysis goals:

- I analysed online retail transaction data to understand customer behaviour
- I provided insights into customer behaviour for marketing output.

Personal goals:

- I learnt how to clean a large dataset of 55,000 records
- I completed my first coding project
- I learnt from trial and error

Yes, I achieved my goals.

Project reflections

Targeting

If I'm just going to look at the UK market, don't spend 1.5 days cleaning a database of 55,000 global records.

Outliers

Don't remove outliers before you've understood the business model. I excluded an outlier in error due to extremely high volume and sales total whereas it was likely to have been an actual true transaction.

Tenacity

I've learnt that I'm very diligent at manually checking large datasets as well as coding. I've also learned that I'm tenacious at trying something I've not done before. I struggled with a coding problem for 2 days. I came out on top, eventually.

New challenges uncovered

I truly stretched myself in this project. However, there were more areas that I wanted to cover if I had time. E.g. converting 38 countries data across 38 different currencies. Next time.....