It includes:
✓ Project Objective
✓ Stakeholder Questions
✓ Data Cleaning & Preparation
✓ Data Analysis
✓ Dashboard Summary
✓ Key Insights
✓ Final Conclusion
✓ Methods Used

Online Retail – Exploratory Data Analysis (EDA) Project

Project Objective

The objective of this Online Retail EDA project was to analyze transactional data from an international retail store to understand overall sales performance, customer purchasing behavior, product trends, and seasonal sales patterns.
The goal was to uncover insights about top products, top customers, major countries contributing to revenue, and time-based sales behavior, ultimately helping the business improve strategic decisions, revenue forecasting, and inventory planning.

Stakeholder Questions

The dataset contained nearly 540K+ rows and multiple features such as InvoiceNo, StockCode, Description, Quantity, UnitPrice, Country, and CustomerID. The project aimed to answer key business questions, including:

- How are Quantity and UnitPrice distributed? Are there negative or extreme values?
- Which products generated the highest revenue?
- Who are the top customers based on total purchase value?
- How many unique customers are there?
- Which countries contribute the most to revenue?
- Which months have the highest sales and revenue trends?
- Are there any seasonal, monthly, or weekday patterns in sales?
- What are the major data quality issues (missing CustomerID, negative Quantity, duplicates)?
- How do sales differ across quantity segments (1–5, 6–20, 100+)?

Data Cleaning and Preparation

Before analysis, the dataset was thoroughly cleaned and made ready for accurate insights:

Data Cleaning Steps

- Removed duplicate records from the dataset.
- Converted *InvoiceDate* into a proper datetime format.
- Converted *CustomerID* into string format and replaced missing values with **"Guest"** when needed.
- Removed transactions with negative **Quantity** or **UnitPrice** when performing valid-sales analysis (negative values treated separately as returns).
- Standardized Description values and removed missing descriptions by replacing them with "None".

Feature Engineering

New columns were created for deeper analysis:

- **Revenue:** Quantity × UnitPrice
- **InvoiceYear:** Extracted using datetime
- **InvoiceMonth:** Extracted for monthly trend analysis
- **MonthYear:** Aggregated month-year combinations
- **InvoiceWeekday:** Used to analyze weekday sales patterns
- **QuantityBin:** Categorized into 1–5, 6–20, 21–50, 51–100, 100+
- **TopProduct:** Labeled top-selling products

These new features allowed revenue segmentation, trend analysis, and customer profiling.

Data Analysis

The analysis was performed using Python (Pandas, NumPy) and Power BI for dashboarding. Multiple visualizations were used to understand trends and patterns.

Sales Trends

- **Monthly Revenue Trend:** November had the highest revenue (~2.9M), with October as the runner-up (~2.2M). Strong **Q4 seasonality** was observed.
- The dataset shows clear monthly fluctuations, indicating potential seasonal demand around holidays.

Product Performance

- **Top Selling Products:**
    - REGENCY CAKESTAND
    - DOTCOM POSTAGE
    - WHITE HANGING HEART
    - PARTY BUNTING

- JUMBO BAG RED
  These products consistently appeared in the top revenue list and drove a major share of total sales.

Customer Analysis

- Women not relevant here; instead:
- **Unique Customers:** ~5,851
- **Top Customers:** Customer IDs like *10, 18102, 14646* generated the highest revenue.
- A small group of customers contributed a disproportionate share of revenue (Pareto behavior).

Geographic Analysis

- **Top Countries by Revenue:**
  - The United Kingdom accounted for **85%+** of the total revenue.
  - Other contributing countries included the Netherlands, Germany, France, and Ireland.

Order Status

- Most transactions were sales; negative quantities indicated returns.
- Returns were minimal compared to completed sales but showed the importance of outlier handling.

Quantity & Price Patterns

- Most purchased quantities were **between 1–5 units**, but orders with **100+ units** accounted for **15%+** of revenue—indicating strong wholesale/bulk behavior.
- UnitPrice was mostly low, but outliers existed due to bulk or special items.

Weekday Analysis

- Revenue was highest on **Wednesday, Thursday, Friday**.
- Lowest on **Saturday**, indicating order drop during weekends.

Dashboard Creation

An interactive Power BI dashboard was built using:

- Cards for KPIs (Total Revenue, Avg Sales, Customer Count, Invoice Count)
- Bar Charts for Top 10 Customers and Top Products
- Donut Chart for QuantityBins
- World Map for Revenue by Country
- Line/Area Chart for Weekday Sales

- Monthly Revenue Trend Chart
- Stacked Bar Charts for Yearly Revenue Summary

Slicers included:

- Month
- Country
- CustomerID
- QuantityBin

This allowed dynamic filtering to explore sales patterns.

Key Insights

- **Total Revenue:** ~19.2M
- **UK Dominance:** UK contributed over **85%** of total revenue.
- **High-Value Customers:** Top 5 customers generated a major share of the overall business.
- **Seasonal Trends:** November and October saw peak sales; strong Q4 performance.
- **Quantity Patterns:** Small quantities (1–5 units) dominated orders, but bulk orders (100+) contributed significantly to revenue.
- **Top Products:** A small group of products consistently drove sales across months.
- **Weekday Pattern:** Wednesday–Friday were peak sales days; weekend drop observed.
- **Returns:** Few return transactions identified through negative quantities.

Final Conclusion

To improve future performance, the business should:

- Focus on **highest-revenue products** through targeted promotions.
- Strengthen relationships with **top customers** using loyalty benefits.
- Plan inventory for **Q4**, the strongest performing season.
- Strategically promote international sales outside the UK to grow market share.
- Separate **retail vs bulk customer** marketing, as both show distinct buying patterns.
- Improve data quality by tracking missing CustomerIDs and validating extreme UnitPrice values.

These strategies can increase revenue, improve customer retention, and optimize inventory and supply chain decisions.

Methods Used

Tools & Technologies

- Python (Pandas, NumPy, Matplotlib, Seaborn)
- Power BI (Dashboard creation)

Techniques

- Data Cleaning
- Feature Engineering
- Descriptive Statistics
- Pivot-style Aggregations
- Time-Series Analysis
- Customer & Product Segmentation

Functions & Methods (Python)

- groupby(), value_counts(), agg(),
- dt.month, dt.year, apply(),
- describe(), corr()

Visualization Types

- Bar Charts
- Line Charts
- Donut Chart
- World Map Chart
- KPI Cards
- Pairplots
- Heatmaps
- Boxplots & Histograms

Analysis Focus

- Sales trends
- Customer segmentation
- Product performance
- Time-based patterns
- Revenue contribution
- Outlier behavior