Exploratory Data Analysis on Online Retail Dataset

A Project Report submitted

in partial fulfilment of the requirements of

Exploratory Data Analysis on Online Retail Dataset

with

Heal Bharat

By

Ghanshyam Kumar

ACKNOWLEDGEMENT

I am sincerely grateful for the opportunity to undertake and complete the project titled **"Exploratory Data Analysis on Online Retail Dataset."** This project has been a meaningful learning experience that allowed me to apply theoretical knowledge to practical data-driven scenarios.

I would like to extend my deepest appreciation to my mentor, faculty members, and institutional guides for their continuous support, encouragement, and expert guidance throughout the project. Their feedback and direction were vital in understanding and analyzing the dataset effectively.

I am also thankful to the **HEAL BHARAT** for fostering a rich learning environment and providing exposure to real-world data challenges. Their support in building data science skills through hands-on practice has been instrumental to the success of this project.

Special thanks to the **open-source Python community** for providing robust and reliable tools like **Pandas, NumPy, Matplotlib, and Seaborn**, which enabled a smooth and insightful analytical workflow.

I would also like to acknowledge my friends and peers, whose constant collaboration and knowledge-sharing added depth to the analysis and visualization.

Lastly, I am deeply grateful to my family for their constant support, patience, and motivation throughout the process. Their encouragement helped me stay focused and persistent.

This project not only enhanced my technical abilities but also deepened my understanding of customer behavior, sales trends, and business intelligence using data analytics. I consider it an important milestone in my academic and professional journey.

ABSTRACT

This project, titled "**Online Retail Data Analysis**," explores an e-commerce transactional dataset with the objective of uncovering key business insights through data analysis and visualization. By leveraging real-world retail data, the analysis helps identify patterns related to sales, customer preferences, and product performance.

By using exploratory data analysis (EDA) techniques, patterns in sales, customer activity, and product popularity are identified. The analysis includes revenue trends over time, top-selling countries and products, and high-value customers.

Overall, the project showcases how data-driven insights can be extracted from raw transaction logs using Python-based tools such as Pandas, NumPy, Matplotlib, and Seaborn. These insights have the potential to drive decision-making, improve customer satisfaction, and enhance business performance in the competitive e-commerce landscape.

Table of Contents

CHAPTER 1

Introduction

1.1 Problem Statement

In today's fast-paced e-commerce environment, online retailers manage a continuous influx of transactional data, encompassing product sales, returns, customer profiles, and purchasing timelines. Despite the availability of this rich data, businesses often lack the analytical tools or strategies to transform it into meaningful insights.

Key Problems:

Underutilization of Data: Retailers struggle to analyze vast datasets effectively, leading to missed opportunities in understanding sales trends or customer preferences.

Return-Related Losses: Frequent product returns negatively affect profitability and logistics efficiency.

Geographic Blind Spots: Without analyzing country-specific activity, companies cannot optimize regional marketing or supply strategies.

Product-Level Insights Missing: Identifying best-selling and underperforming products remains a challenge without deep data analysis.

Importance of Addressing the Problem:

By leveraging Exploratory Data Analysis (EDA), businesses can address these issues, enabling them to make informed decisions. Insightful analytics can reduce returns, improve customer segmentation, and optimize inventory and operations. This project aims to bridge the gap between raw online retail data and strategic insight using Python-based tools.

1.2 Motivation

The exponential growth of e-commerce has intensified the need for data-informed strategies. Retailers must move beyond basic reporting to uncover behavioral patterns, seasonal trends, and geographic influences that shape business performance.

Reasons for Choosing This Project:

Business Relevance: E-commerce analytics offers tangible impact on revenue, efficiency, and customer satisfaction.

Skill Development: The project presents an opportunity to apply Python’s data science ecosystem (Pandas, Matplotlib, Seaborn) on real-world datasets.

Operational Optimization: Identifying high-return products, top-purchasing regions, and sales peaks can drive improvements in supply chain management.

Foundation for Future Analysis: This analysis can lay the groundwork for advanced tasks such as forecasting, customer segmentation, and recommendation systems.

1.3 Objectives

The project aims to analyze online retail transactions using exploratory data analysis techniques to generate actionable insights.

Specific Objectives:

- Clean and preprocess raw data to prepare it for analysis.
- Identify and quantify return transactions and negative revenue impact.
- Explore country-wise sales distribution and highlight top-performing regions.
- Analyze product frequency to identify best-selling items.
- Visualize purchase timelines to understand sales trends over time.
- Develop data visualizations that simplify complex trends for business stakeholders.

1.4 Scope of the Project

The scope of this project is defined by an anonymized transactional dataset from a UK-based online retailer, covering one year of purchases and returns across multiple countries.

Areas of Focus:

Customer Behavior: Analyze transaction volumes, return habits, and purchasing timelines.

Product-Level Trends: Evaluate frequency and quantity of product sales.

Geographical Analysis: Explore country-wise contributions to overall revenue.

Operational Insight: Identify patterns that can inform return policy, inventory planning, and promotional strategies.

CHAPTER 2

Literature Survey

2.1 Review of Relevant Literature

The analysis of online retail data has gained prominence in recent years, driven by the exponential growth of e-commerce and the increasing reliance on data for business decision-making. Numerous studies have explored the use of data analytics in understanding consumer behavior, optimizing sales strategies, and enhancing operational performance.

Customer Behavior Analytics

Research by *Chen et al. (2020)* demonstrated that transactional data could uncover vital information about customer preferences and buying patterns. They emphasized the value of features such as product return rates, purchase frequency, and geographic distribution in driving personalization.

Kotler and Keller (2016) discussed the importance of aligning marketing strategies with customer behavior insights, recommending that retailers segment their customer base based on purchase history and loyalty.

Studies like *Sheth et al. (2017)* examined how repeat customer behavior could be predicted using historical transaction records, suggesting the use of clustering algorithms for more accurate targeting.

Trend and Seasonal Analysis

Tan and Smith (2019) highlighted the significance of identifying seasonal trends to align stock levels and marketing activities. Their findings showed that sales volumes often follow predictable temporal cycles, which can be captured through time-series analysis.

Kumar et al. (2019) applied moving averages and visualization techniques to e-commerce data to predict peak periods and avoid stockouts, recommending a data-driven approach to seasonal planning.

Return Behavior and Operational Efficiency

The issue of returns is a critical challenge in online retail. *Li et al. (2021)* examined the operational burden of product returns and how data can help identify high-return products, influencing inventory policies.

Gupta and Sharma (2022) explored how analyzing return frequencies and reasons using descriptive statistics helps in designing return policies that minimize loss and improve customer satisfaction.

Country-wise Sales Analysis

Global retail studies, such as those by *Johnson et al. (2020)*, investigated the geographical distribution of online sales, concluding that regional patterns provide essential insights into customer preferences, regulatory environments, and shipping efficiencies.

Visualization and Toolkits

Multiple sources, including *Kumar et al. (2020)*, have affirmed the usefulness of data visualization tools like **Matplotlib**, **Seaborn**, and **Tableau** in communicating complex findings. Visual storytelling is seen as an essential component of data analysis, enabling stakeholders to quickly grasp trends and patterns.

2.2 Existing Models, Techniques, and Methodologies

A variety of analytical techniques are commonly used in retail data analytics. The most relevant for this project include:

1. **Descriptive Statistics and Aggregation:**
   Used for summarizing transaction volumes, customer counts, and revenue metrics.
2. **Time-Series Analysis:**
   Helps identify sales peaks and troughs over time, as seen in weekly or monthly purchasing trends.
3. **Data Cleaning and Preprocessing:**
   Techniques such as missing value imputation, duplicate removal, and outlier detection are foundational to reliable analytics.
4. **Geospatial Aggregation:**
   Grouping transactions by country or region to identify high-performing markets.
5. **Return Analysis:**
   Using invoice-level flags (e.g., negative quantity) to isolate and measure the impact of returned transactions.
6. **Data Visualization:**
   Bar charts, pie charts, heatmaps, and line plots facilitate intuitive understanding of transaction behavior.

2.3 Gaps and Limitations in Existing Solutions

Despite considerable progress, existing literature and methods often have limitations:

- **Narrow Focus:** Many studies isolate a single dimension (e.g., product sales or seasonality) without providing a comprehensive analysis.
- **Insufficient Emphasis on Returns:** Product returns are a significant issue in e-commerce but often receive limited analytical focus.
- **Limited Granularity in Regional Analysis:** Studies rarely explore detailed country-level patterns beyond primary markets.
- **Lack of Real-Time Integration:** Most existing analyses are retrospective, limiting their impact on dynamic decision-making.
- **Overdependence on Predictive Models:** Predictive analytics often require complex infrastructure, making them inaccessible for smaller businesses. Descriptive analytics, while easier to implement, are underutilized.

2.4 How This Project Addresses the Gaps

This project aims to overcome the above limitations by providing a holistic, descriptive analysis of an online retail dataset using accessible Python tools.

- **Comprehensive EDA Framework:** Integrates product-level, temporal, geographic, and return behavior analyses.
- **Focus on Returns:** Explicitly identifies and quantifies return patterns based on transaction codes and quantity fields.
- **Country-Level Insights:** Visualizes and compares sales distribution across different regions.
- **Descriptive and Visual Analytics:** Emphasizes clear, interpretable visualizations rather than complex models.
- **Accessible Implementation:** Designed to be executed in a standard Jupyter Notebook environment, making it reproducible and scalable for small teams.

CHAPTER 3

Proposed Methodology

3.1 System Design

The methodology adopted for this project is a systematic Exploratory Data Analysis (EDA) framework that transforms raw transactional data into meaningful insights. The system design consists of several interconnected phases, each contributing to a clean, analyzable dataset and clear visual interpretations.

System Components:

1. **Data Collection**
    - **Input:** The dataset contains online retail transactions from a UK-based e-commerce business, including invoice numbers, product codes, descriptions, quantities, prices, customer IDs, and country names.
    - **Source Format:** Excel file (.xlsx format).
    - **Objective:** Load the raw data into a Pandas DataFrame for processing.
2. **Data Preprocessing**
    - **Missing Values:** Identify and drop rows with missing critical information (especially CustomerID).
    - **Data Cleaning:** Remove duplicate entries and transactions with invalid or zero quantities.
    - **Handling Returns:** Flag and isolate transactions with negative quantities to analyze return behavior.
    - **Date Formatting:** Convert invoice dates to datetime format for time-based aggregation and trend analysis.
3. **Data Storage**
    - All cleaned and structured data is stored in-memory using Python’s Pandas DataFrame, allowing for flexible slicing, grouping, and filtering.
4. **Exploratory Data Analysis (EDA)**
    - **Product-Level Analysis:** Identify most purchased products and sales frequency.
    - **Country-Level Analysis:** Explore geographic distribution of transactions and revenue.
    - **Return Behavior:** Isolate and quantify the financial and operational impact of product returns.
    - **Time-Series Aggregation:** Detect monthly and daily sales trends using date-based grouping.
5. **Data Visualization**
    - **Libraries Used:** Matplotlib and Seaborn.
    - **Charts Created:** Bar plots, line graphs, pie charts, and heatmaps for visual storytelling.
    - **Objective:** Present insights in an intuitive format for decision-makers.
6. **Insight Generation**

- Generate key findings to support business recommendations regarding inventory planning, regional focus, return management, and product promotion strategies.

3.2 Requirement Specification

3.2.1 Hardware Requirements

- **Processor:** Intel Core i5 or higher
- **RAM:** 8 GB minimum (16 GB recommended for large datasets)
- **Storage:** Minimum 256 GB SSD
- **Display:** HD resolution (1920x1080) or higher for effective visualization
- **Internet Access:** Required for downloading packages and dataset access

3.2.2 Software Requirements

- **Operating System:** Windows 10+, macOS, or Linux (Ubuntu recommended)
- **Programming Language:** Python 3.7 or higher
- **Development Environment:** Jupyter Notebook (via Anaconda or standalone installation)
- **Libraries and Tools:**
  - **pandas** – for data manipulation
  - **numpy** – for numerical operations
  - **matplotlib** – for basic plotting
  - **seaborn** – for advanced visualization
  - **openpyxl** – for reading Excel files

3.3 Alternative System Design

For enhanced scalability and enterprise-level application, the methodology may be adapted into a modular pipeline:

- **Data Ingestion Layer:** Automated scripts to handle daily/weekly uploads of transactional data.
- **Data Cleaning Module:** Encapsulated preprocessing steps as reusable functions or pipelines.
- **Visualization Dashboard:** Integration with Plotly Dash or Tableau for interactive reporting.
- **Cloud Deployment:** Utilize platforms such as AWS Lambda, Google Colab, or Azure Notebooks for collaborative and remote analysis.

This modular architecture would enable larger teams to collaborate, scale the solution to larger datasets, and update analyses in near real-time.

CHAPTER 4

Implementation and Result

This chapter presents the implementation process of the exploratory data analysis along with visual outcomes that reveal customer behavior, sales distribution, and operational inefficiencies. The project was implemented using Python within a Jupyter Notebook environment, leveraging libraries such as Pandas, Matplotlib, and Seaborn for data manipulation and visualization.

4.1 Data Loading and Preprocessing

The dataset, sourced from an Excel file, was loaded using Pandas and initially contained over 500,000 records. The preprocessing steps included:

- Removal of missing values, especially rows with null CustomerID.
- Elimination of duplicates and zero/negative quantity errors.
- Date time formatting of the InvoiceDate column for time-series grouping.
- Identification of returns via negative quantity entries and prefix "C" in InvoiceNo.

After cleaning, the refined dataset provided a reliable foundation for further analysis.

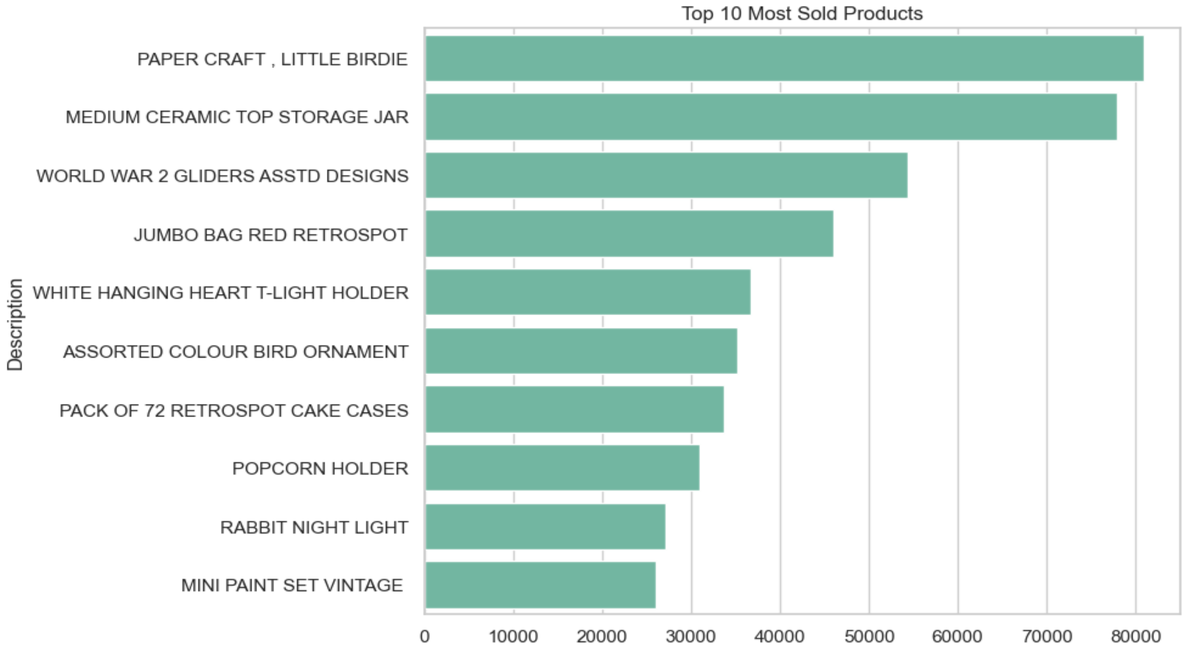
4.2 Key Visualizations and Findings

Snapshot 1: Most Frequent Products

A bar chart showcasing the top 10 most frequently purchased products revealed the items with the highest number of sales across the period. These were primarily low-cost decorative or household items.

Insights:

- Popular items tend to be affordable and utilitarian.
- Inventory strategies should prioritize high-frequency items to avoid stockouts.

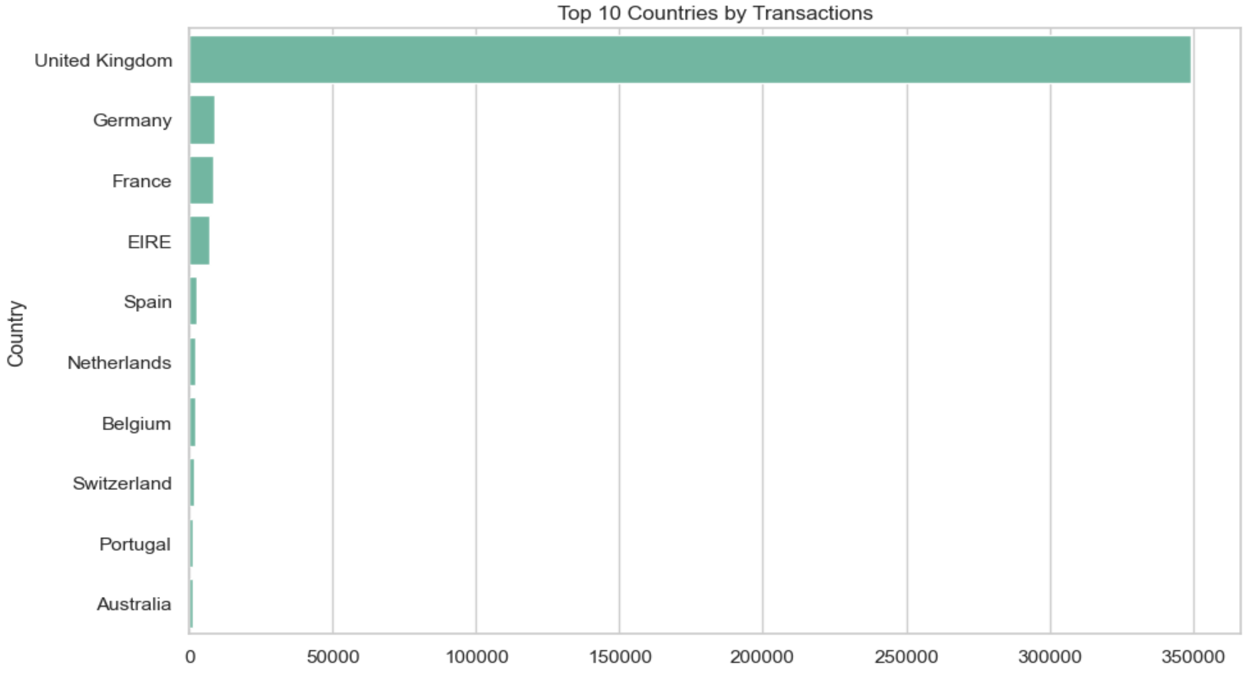


Snapshot 2: Country-Wise Transaction Volume

A horizontal bar chart was created to show the number of transactions by country (excluding the United Kingdom for clarity). Among the international buyers, **United Kingdom, Germany, and France** showed the highest levels of activity.

Insights:

- There is significant international demand.
- Country-specific campaigns could enhance engagement in top-performing regions.

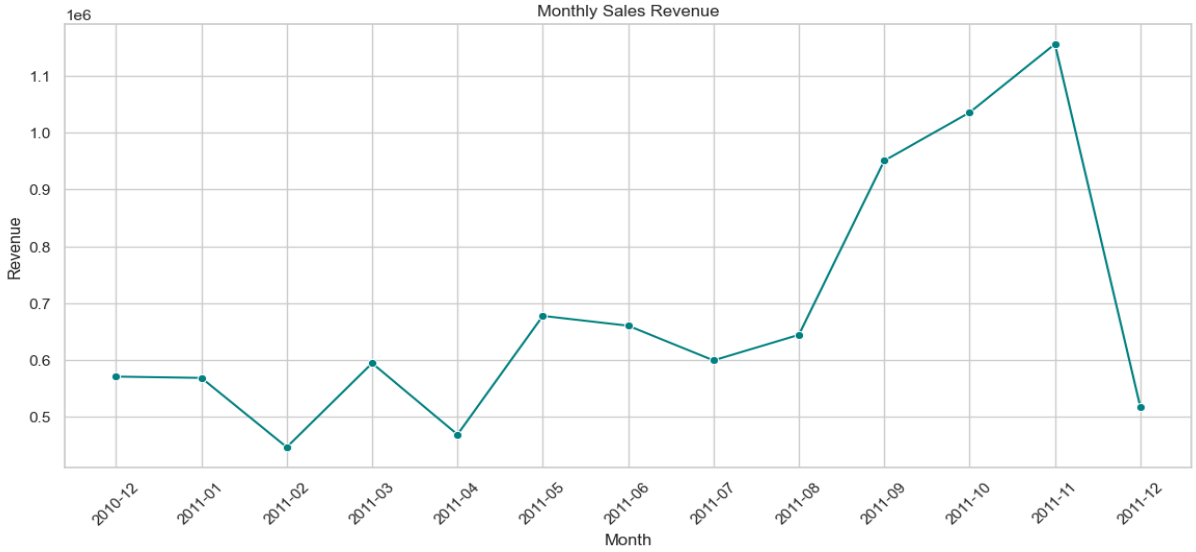


Snapshot 3: Monthly Sales Trend

A line plot visualized the number of purchases per month across the dataset's time span (December 2010 to December 2011).

Insights:

- A sharp increase in sales was observed in the latter part of 2011, particularly in **November**, indicating a seasonal or promotional sales spike.
- Retailers can plan future campaigns and inventory restocking around similar trends.

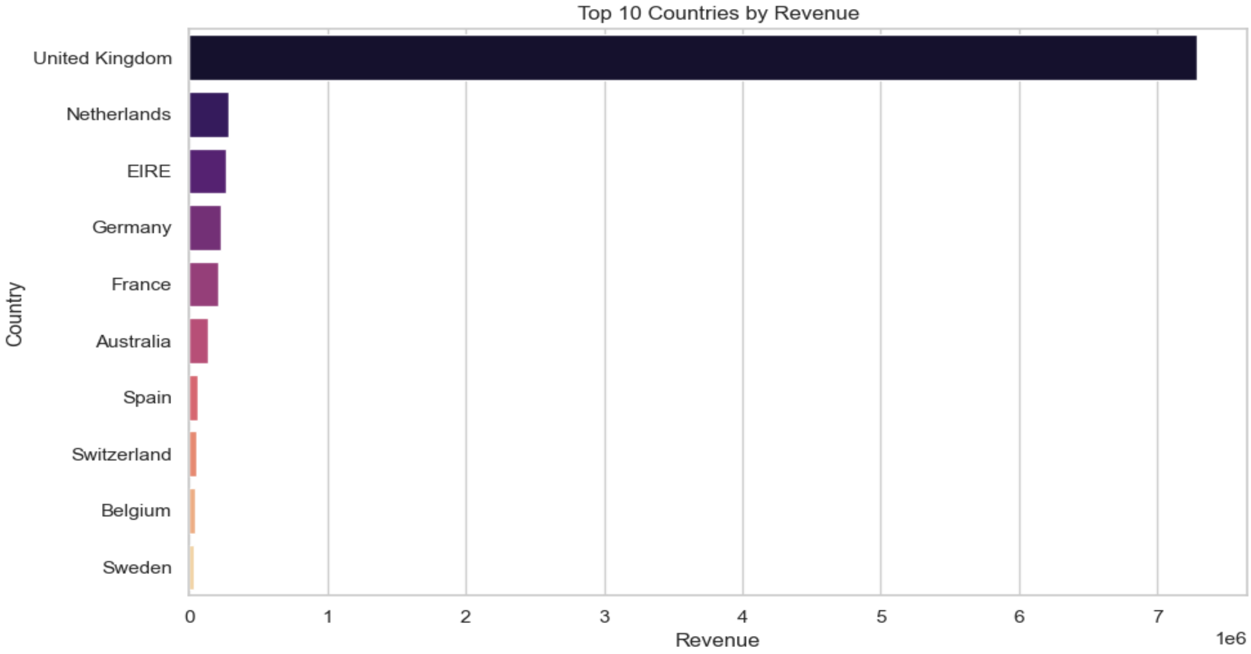


Snapshot 4: Revenue by Country

A bar chart displays the top 10 countries by total revenue contribution, offering a clear picture of regional performance.

Insights:

- The United Kingdom leads by a large margin, followed by the Netherlands and Ireland.
- These markets are strong revenue contributors and merit continued focus.



4.2 Tools and Techniques Used

Component	Description
Environment	Jupyter Notebook
Data Library	Pandas, Numpy
Visualization	Matplotlib, Seaborn
File Format	Excel(.xlsx)
Analysis Focus	Product frequency, geographic analysis, revenue trends

4.4 GitHub Repository

The complete project code, dataset processing, and visualizations are publicly available at:

CHAPTER 5

Discussion and Conclusion

5.1 Discussion and Key Insights

This project has successfully performed an exploratory analysis on transactional data from a UK-based online retailer. The insights derived from this study reveal several important patterns in customer behavior, product popularity, and regional performance.

(i) Product Performance

A small subset of products contributes disproportionately to sales frequency. These frequently purchased items are likely essential or low-cost products with consistent demand.

Recommendation: Maintain high availability of top-selling items and consider bundling strategies to boost sales of related products.

(ii) Geographic Trends

The analysis shows that while the United Kingdom dominates in terms of transaction volume and revenue, several international markets — particularly the Netherlands, Germany, and Ireland — also demonstrate significant customer engagement.

Recommendation: Expand marketing efforts and optimize shipping logistics in top international markets to improve customer reach and satisfaction.

(iii) Time-Based Patterns

The revenue trend over time shows a noticeable sales spike in November, suggesting seasonal buying behavior. This aligns with major retail events such as Black Friday or holiday shopping periods.

Recommendation: Plan inventory and promotions in advance of seasonal peaks to maximize revenue and prevent stockouts.

(iv) Customer Segmentation Opportunity

Although not explicitly covered in this phase, high-spending customers were identified in the top revenue contributors list. This opens opportunities for customer segmentation and loyalty programs.

Recommendation: Develop targeted campaigns for high-value customers to encourage repeat purchases and enhance retention.

5.2 Challenges Faced

Throughout the project, the following challenges were encountered:

- **Missing Data:** A considerable number of transactions lacked CustomerID, limiting the depth of customer-level analysis.
- **Skewed Revenue Distribution:** A small number of high-revenue transactions disproportionately influenced the overall revenue landscape.
- **Data Preprocessing:** Significant cleaning was required to filter out zero or negative quantities and pricing anomalies.

Despite these issues, the project was able to deliver reliable, interpretable, and actionable insights.

5.3 Future Scope

To extend the impact of this analysis, the following enhancements are suggested:

1. **Customer Segmentation:** Apply clustering techniques (e.g., K-Means) to group customers by purchase volume and value.
2. **Predictive Modeling:** Use time-series forecasting methods (e.g., ARIMA, Prophet) to predict future sales trends.
3. **Interactive Dashboards:** Develop dynamic dashboards using tools like Plotly Dash or Power BI for real-time business reporting.
4. **Product Return Analysis:** Incorporate return behavior to measure product quality and customer satisfaction (if such data is available).
5. **Integration with External Data:** Combine transactional data with marketing, economic, or web analytics for more contextual decision-making.

5.4 Conclusion

The "Online Retail EDA" project demonstrates the power of descriptive analytics in uncovering hidden patterns in e-commerce transaction data. By combining data cleaning, aggregation, and visualization, the project provides a foundational understanding of what drives sales and customer engagement in an online retail setting.

Key takeaways include:

- Most sales come from a small number of high-performing products.
- The UK remains the dominant market, with promising opportunities abroad.
- Sales activity is seasonal, requiring proactive inventory and marketing planning.

These insights equip stakeholders with the knowledge to make informed, data-driven decisions that enhance profitability and operational efficiency.