NOVA IMS Information Management School

MDSAA

Master's Degree Program in

Data Science and Advanced Analytics

Big Data Analytics Project

The Gift Whisperers: Knowing Your Customers, Predicting Their Next Move

Hassan Bhatti

Moeko Mitani

Oumayma Ben Hfaiedh

Ricardo Pereira

Group 77

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

0. AREAS THAT CHANGED AFTER THE DEFENSE

- We decided not to write an article, instead we will provide our insights and conclusions in this report.
- For the forecasting component, we initially presented our solution using a Linear Regression model during defense. Afterward, we extended our analysis by testing additional models, including Gradient Boosting Regressor, Random Forest Regressor, and SARIMAX. To improve the RMSE scores, we applied hyperparameter tuning using a randomized search strategy.
- We also explored the possibility of extending our models to forecast quantities for multiple products simultaneously. However, due to the limitations of the Databricks Community Edition (which causes the runtime to break each time it exceeds one hour), we were unable to implement multi-series time forecasting at scale.
- We have created a graph visualization for clusters that we identified in our customers.

1 INTRODUCTION AND BACKGROUND

In today's highly competitive retail landscape, data-driven decision-making has become a critical factor in achieving business success and maintaining a competitive edge. For our Big Data Analytics project, we chose to focus on the retail industry due to its dynamic nature and the vast volume of data it generates daily. Specifically, we worked with a dataset from a company that specializes in sourcing and distributing a wide range of high-quality products.

The core problem we aim to address through big data analytics involves extracting actionable insights from large and complex datasets. Our project focuses on two key objectives:

1. **Customer Segmentation:** By creating data-driven clusters of customers based on purchasing behavior and other attributes, we aim to help the company enhance its marketing strategy through targeted promotions and personalized offerings.
2. **Sales Forecasting:** Using predictive analytics, we aim to forecast future product demand, enabling the company to optimize inventory levels, improve supply chain efficiency, and allocate resources more effectively over the coming months.

Throughout this project, we applied techniques such as data preprocessing, feature engineering, unsupervised clustering, and supervised machine learning for time series forecasting. By addressing real world business problems using these methods helped us develop hands-on skills and see firsthand how large-scale data can be turned into valuable insights for smarter business decisions.

2 PROJECT MOTIVATION

Our project was driven by the growing need for the company to leverage big data for smarter customer understanding and demand planning. By combining customer segmentation with sales forecasting, we aimed to provide a practical solution that drives both marketing personalization and operational efficiency in the fast-paced retail industry. Future plans also include the implementation of a recommendation system using customer segmentation.

3 DATA COLLECTION & PREPROCESSINGDATA SOURCES

Our data source is a public dataset from Kaggle. We were able to find a dataset that has 1 million rows and columns. This dataset represents the transactions of consumers over a 3-year period. The data closely aligns with real-world retail scenarios, offering both volume and variety for meaningful analysis.

4 DATA CHARACTERISTICS

The dataset is **structured**, with each row representing an individual transaction. There are eight key features, including customer information, product information, and transaction information (see Appendix A).

5 DATA CLEANING AND PREPROCESSING

Before conducting any analysis, we performed several essential data cleaning and preprocessing steps including:

- **Handling Missing Values:** A significant number of entries were missing values in the *CustomerID*. Since customer-level data is essential for clustering and behavior analysis, we chose not to discard these rows outright. Instead, for transactions sharing the same invoice number, we generated a placeholder for the *CustomerID* to preserve the transactional context and maintain data integrity for analysis.
- **Removing Duplicates:** Duplicate records were identified and removed to prevent double-counting and ensure accurate statistical summaries.
- **Data Type Conversion:** The *Quantity*, *InvoiceDate*, *Price*, and *ID* were converted into proper data types (to Integer, Timestamp, Decimal, and Integer sequentially). The step of changing the *InvoiceDate* to Timestamp was essential for time-based analysis, such as identifying seasonal trends and forecasting future sales.
- **Resolving Data Inconsistencies:** We found several transactions with a unit price of 0, which are likely data entry errors or incomplete records. These entries were removed, as they do not contribute meaningful insights and could distort the analysis.

We have created new datasets based on transformations on the original dataset for different purposes:

For Clustering:
We transformed the raw transactional data into a **customer-centric dataset** by aggregating transactions at the customer level. The *CustomerID* became the index, and we engineered new features (see Appendix B). New features allow us to capture behavioral patterns for each customer, making the dataset suitable for segmentation through clustering algorithms.

For Sales Forecasting:
We prepared a **product-centric dataset**, where the *StockCode* served as the index. We aggregated sales data over time to create a time series format and engineered features to improve model accuracy (see Appendix B). These engineered features were essential in aligning the raw data with the specific goals of our project and provided a solid foundation for building effective clustering and forecasting models.

6 METHODOLOGY & TOOLS

6.1 CLUSTERING

For the clustering component of our analysis, we chose to use the **K-Means** algorithm due to its simplicity and effectiveness in segmenting data into meaningful groups, considering the limitation of Databricks Community Edition.

To determine the optimal number of clusters (K), we used the **Elbow Method**, which helped us visualize the point where adding more clusters no longer significantly reduces the within-cluster sum of squares. Based on this analysis, we identified a noticeable “elbow” at **K=3**. Therefore, we decided to segment the data into three distinct groups.

We then profiled each segment based on behavioral metrics:

- **Cluster 1:** High-value loyal customers with frequent purchases and high total spending.
- **Cluster 2:** Occasional buyers with moderate activity and potential for re-engagement.
- **Cluster 3:** Low engagement customers with infrequent, low value purchases.

Cluster profiling was supported by distribution plots, helping us build actionable personas for marketing purposes (see Appendix C).

6.2 SALES FORECASTING

We approached sales forecasting as a supervised regression problem, aimed at predicting monthly product demand, specifically the *total_quantity* sold per product. To maintain temporal integrity and simulate real world forecasting conditions, we split the dataset chronologically:

- **Training set**: June 2023 - June 2024 (1 year)
- **Validation set**: July 2024 - December 2024 (6 months)
- **Test set: January** 2025 - June 2025 (6 months)

As mentioned before; to enhance model accuracy, we engineered a set of time sensitive features to capture seasonality, trends, and volatility in sales behavior (see Appendix B).

We evaluated a range of models, including **Linear Regression**, **Gradient Boosting Regressor**, **Random Forest Regressor**, and **SARIMAX**. Model performance was measured using **Root Mean Squared Error (RMSE)** on the validation set. The best results were achieved using a **Random Forest Regressor**, which was further optimized through hyperparameter tuning, using **RandomizedSearchCV**, a cross-validated randomized search strategy. The following hyperparameters were tuned:

- n_estimators: Number of trees in the forest
- max_depth: Maximum depth of each tree
- min_samples_split: Minimum number of samples required to split an internal node
- min_samples_leaf: Minimum number of samples required to be at a leaf node
- max_features: Number of features to consider when looking for the best split

The search was performed over 20 iterations using 3-fold cross-validation and RMSE as the scoring metric. The final model included the following configuration:

- n_estimators = 300
- max_depth = 20
- min_samples_split = 10
- min_samples_leaf = 2
- max_features = 'sqrt'

The hyperparameter tuning process resulted in a notable reduction in RMSE from **79.07 to 70.79**, reflecting an approximate **10.47%** improvement in predictive accuracy. This enhancement highlights the model's strong predictive capability and its ability to generalize effectively to unseen data. Forecasts were visualized alongside actual values, revealing that the model successfully captured key underlying patterns, including seasonal trends and sales volatility. It is worth noting that these results are based on the forecast for only one product. The limitations of Databricks Community Edition, particularly its tendency to crash during resource-intensive model evaluations, prevented us from forecasting the entire product range in the dataset.

6.3 GRAPH

To better understand the structure of our customer segments, we created a graph-based visualization using **NetworkX**. In this graph, each node represents a customer, and nodes are color-coded based on their assigned cluster from the K-Means algorithm. Customers in the same cluster are connected to show the cohesion of each group. This visual approach helped us identify the density and distribution of clusters, revealing how distinct or overlapping the customer segments are (see Appendix D).

7 POWER BI DASHBOARD

For the visualization phase of our project, we chose Power BI as our primary tool. Since the Databricks Community Edition does not support direct integration with Power BI, we worked around this limitation by exporting two key datasets: the preprocessed transactional data and the clustering results with assigned labels.

We then designed a multi-page Power BI dashboard to bring the cleaned retail data to life. We created two reports: **EDA Dashboard** and **Clustering Dashboard**. Each page in both reports was dedicated to a specific focus area, allowing users to explore insights across different dimensions of the business in a clear and structured way (see Appendix E):

- **EDA Dashboard**
  - **Sales Overview:** Total sales, returns, and trends over time
  - **Customer Insights**: Top customers, average spend, and behavior
  - **Product Performance**: Bestsellers and most returned items
  - **Time Trends**: Sales by day, hour, and month
  - **Geographical View**: Sales and returns by country on an interactive map
- Clustering Dashboard
  - **Clusters Dashboard**: Summary of each cluster

We used calculated columns and DAX measures like *TotalSales* and *ReturnRate* to power dynamic visuals and filters. The result is a user-friendly dashboard that turns raw data into insights.

8 CONCLUSION

This project demonstrated how big data analytics can transform raw retail data into actionable business insights. By combining customer segmentation and sales forecasting, we provided a practical solution to enhance both marketing effectiveness and operational planning. Using tools like Databricks for processing big data and Power BI for visualization, we were able to uncover patterns in customer behavior, predict future demand, and present our findings through an interactive dashboard. Overall, the project not only addressed real-world business challenges but also strengthened our ability to work with large datasets, apply machine learning techniques, and deliver data-driven solutions.

APPENDIX

Appendix A
Features in The Dataset

Features	Description
Invoice Number (*Invoice*)	Unique identifier for each transaction.
Stock Code (*StockCode)*	Product identifier.
Description (*Description*)	Textual description of the products.
Quantity (*Quantity*)	Number of units purchased.
Invoice Date (*InvoiceDate)*	Timestamp of the transactions.
Unit Price (*Price*)	Price per unit of the products.
Customer ID (*ID*)	Unique customer identifier.
Country (*Country*)	Country of the invoices.

Appendix B
New Features in New DataFrame for Clustering and Sales Forecasting

For Clustering:

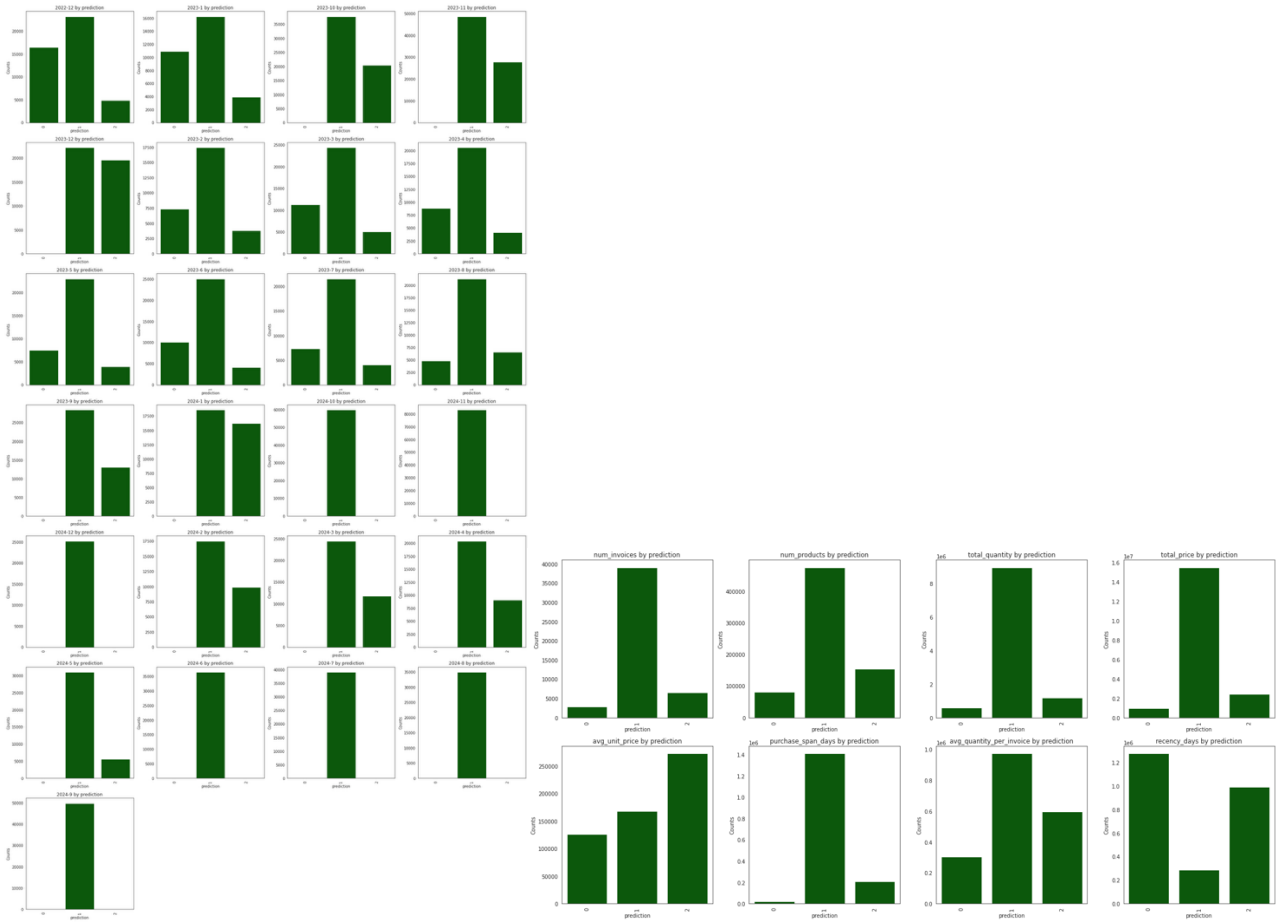
New Feature	Description
Total Price (*TotalPrice*)	Total price of each transaction line.
Number of Products (*num_products*)	Number of unique products purchased by the customer.
Total Quantity (*total_quantity*)	Total number of items purchased by the customer.
Total Price (*total_price*)	Total amount spent by the customer.
Average Unit Price (*avg_unit_price*)	Average price per unit item purchased.
First Purchase Date (*first_purchase_date*)	Date of customer's first recorded purchase.
Last Purchase Date (*last_purchase_date*)	Date of customer's most recent purchase.
Purchase Span (*purchase_span_days*)	Time between first and last purchase.
Average Quantity per Invoice (*avg_quantity_per_invoice*)	Average number of items per invoice.
Recency (*recency_days*)	Days since the customer's last purchase (as of 09/12/2024).
Monthly Purchase Counts	Count of purchases per month (pivoted).

For Sales Forecasting:

New Feature	Description
Year (*year*)	The year extracted from the date (e.g., 2023, 2024).
Month (*month*)	The year and month in "YYYY-MM" format (e.g., 2024-01).
Date (*date*)	The actual calendar date, typically set to the 1st of each month (e.g., 2024-01-01).
Total Quantity (*total_quantity*)	Total number of items sold (or transacted) for a product (*StockCode*) in a given month.
Number of Months (*month_num*)	A numeric representation of the month (e.g., 202401 for Jan 2024) used to sort time chronologically.
Month Sin (*month_sin*)	A cyclical transformation of the month using a sine function, helping models capture seasonality patterns.
Month Cos (*month_cos*)	Another cyclical transformation using cosine to pair with *month_sin* for better seasonal modeling.
6 Month Lag Quantity (*lag_6m_quantity*)	The total quantity value from exactly 6 months earlier for the same product.
6 Month Rolling Average (*rolling_avg_6m*)	The average quantity sold over the last 6 months, including the current one, helps track recent trends.
6 Month Standard Deviation Average (*rolling_std_6m*)	The standard deviation of quantities over the last 6 months, measures sales volatility or stability.

Appendix C
Customer Profiling

1. **Profiling Visualization**



2. **Profiling Details**

Cluster 0: Inactive or Lost Customers (Size: 2,109 customers)

Key Characteristics:

- Lowest in invoices, products, quantity, revenue, and average unit price.
- Very long recency - they haven't purchased in a long time.
- Short customer lifetime (Purchase Span Days).
- Lowest average quantity per invoice.
- Most recent purchases occurred mainly in 2022-2023 summer.
- No purchases at all in late 2023 or 2024.

Profile Summary:

These are inactive or churned customers. They used to engage at a low level but have now completely stopped purchasing. This group may have switched to competitors, closed their business, or found less value in our offerings.

Actionable Strategy:

- Restart Campaigns: Special reactivation offers or targeted email campaigns.
- Feedback collection: Try to learn why they stopped ordering.
- Consider removing them from active targeting if unresponsive after multiple campaigns.

Cluster 1: High-Value Loyal Customers (Size: 4,524 customers)

Key Characteristics:

- Highest across all transactional metrics: invoices, products, quantity, revenue.
- Consistent across all months and years, highly active and reliable.
- Frequent purchases with large average quantities per invoice.
- Recent and ongoing engagement.
- Long customer lifetime (long span days).

Profile Summary:
These customers are our top-tier wholesale customers who are loyal, high-frequency, and high-value buyers. They show strong ongoing engagement and form the backbone of our business.

Actionable Strategy:

- Loyalty Programs: Reward frequent buyers such as discounts, early access, bulk deals.
- Develop recommendation system: Tailor product recommendations.

Cluster 2: Occasional or Premium Buyers (Size: 2,810 customers)

Key Characteristics:

- Middle in invoice count, product variety, quantity, and revenue.
- Highest average unit price which means they buy fewer, but more expensive items.
- Somewhat consistent but limited engagement through 2023 and early 2024, disappeared the middle of 2024.
- Moderate recency tells that they are not fully inactive, but not as recent as Cluster 1.
- Average quantity per invoice is moderate.

Profile Summary:
These are likely small boutique resellers or premium buyers. Occasional purchases, preference for luxury and selected gift items, and small, high-margin orders. Their pattern shows selective engagement.

Actionable Strategy:

- Premium Product Bundles: Offer curated or seasonal premium boxes.
- Targeted product launches (high-margin or exclusive items).
- Encourage regular use with small loyalty rewards or “limited time” offers.

3. **Profiling Summary**

Cluster	Size	Activity Level	Customer Type	Recency	Spending	Strategy
0	2,109	Inactive	Churned/Lost	High	Low	Restart Campaign, survey
1	4,524	Active	High-value/Loyal	Low	High	Loyalty, Upsell
2	2,810	Moderate	Occasional/Premium	Medium	Medium	Premium-focused

Appendix D
Graph

Observation

1. Cluster Distribution
The colors (nodes) are spread out across the graph, but it can be still seen that groups of same colored nodes sticking together. This suggests that the clustering algorithm found meaningful patterns, even if the graph is dense.

2. Cluster Separation
Red, green, and blue dominate in some areas, which means that the model has found customers with distinct behaviors. However, because the graph is connected at many edges, clusters may be interrelated or slightly overlapping, which could be normal in customer behavior.

3. Outliers and Bridges
Nodes that are isolated or on the edges could be outliers (unusual customer behavior) or “Bridge customers” connecting the two clusters (may be transitional or seasonal behavior).

4. Graph Density
The middle is very dense, meaning there are a lot of similar customers. It could indicate many customers with similar behavior.

Appendix E

Power BI Dashboard

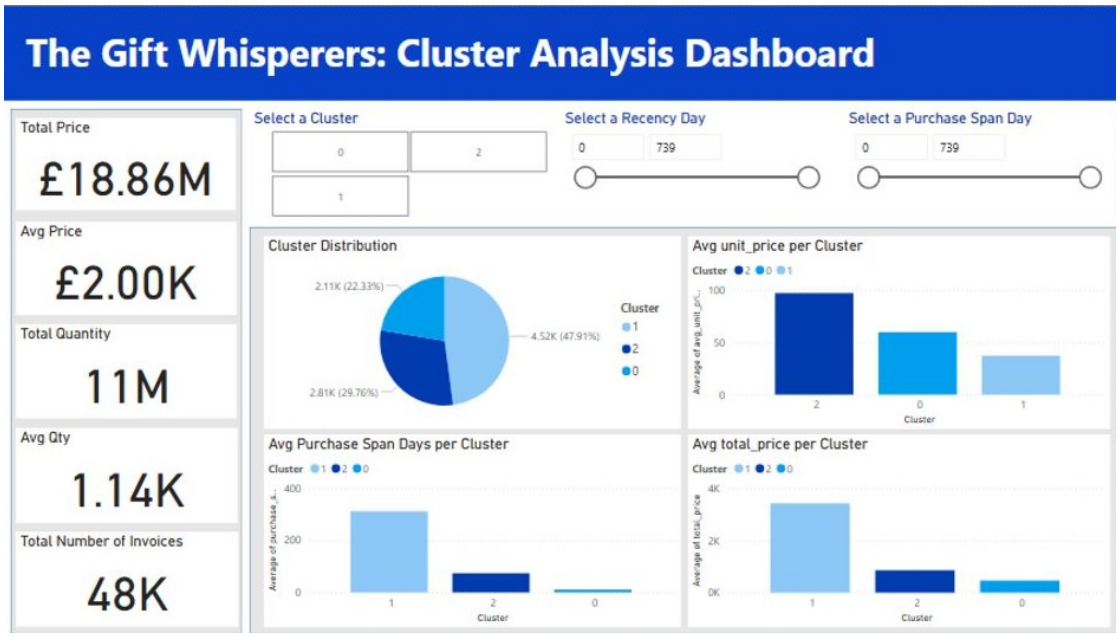


Figure 1: Clustering Dashboard – Clusters Dashboard

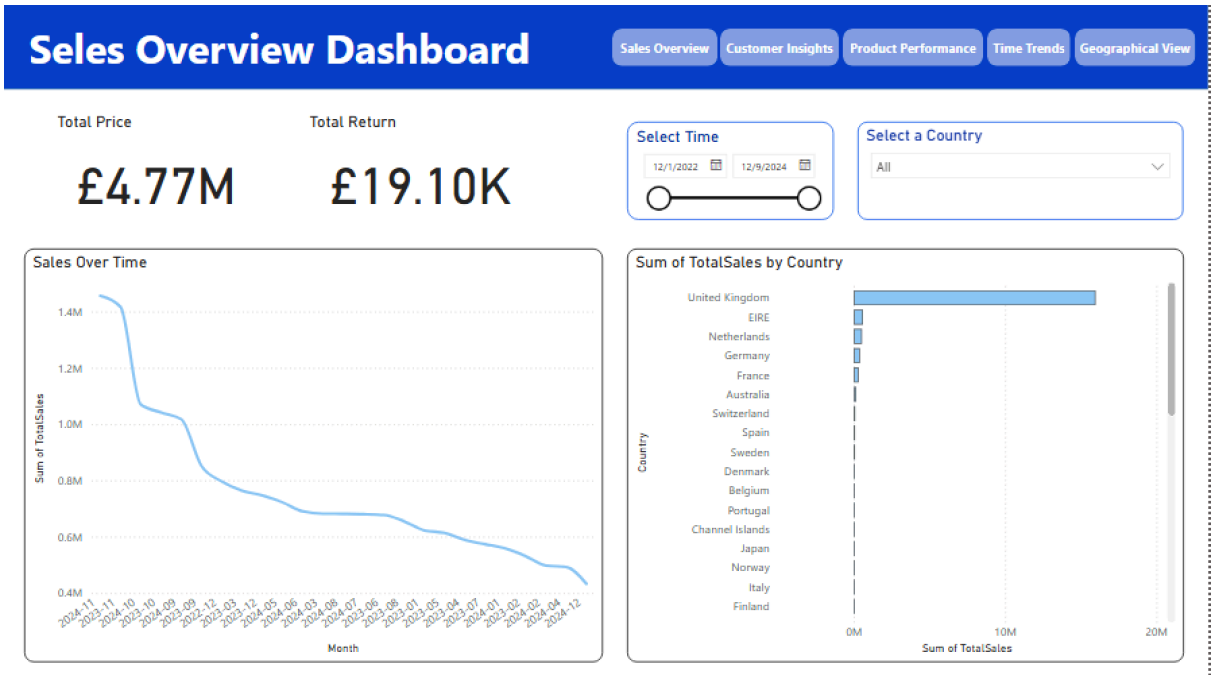


Figure 2: EDA Dashboard – Sales Overview Dashboard

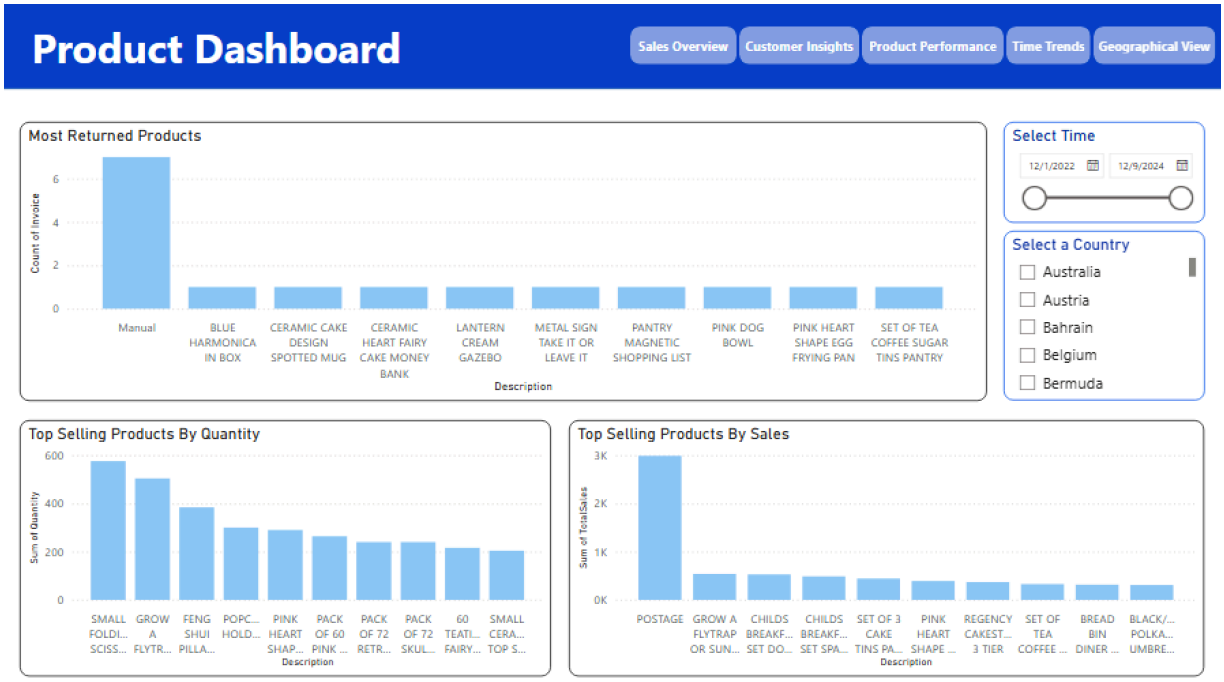


Figure 3: EDA Dashboard – Product Dashboard

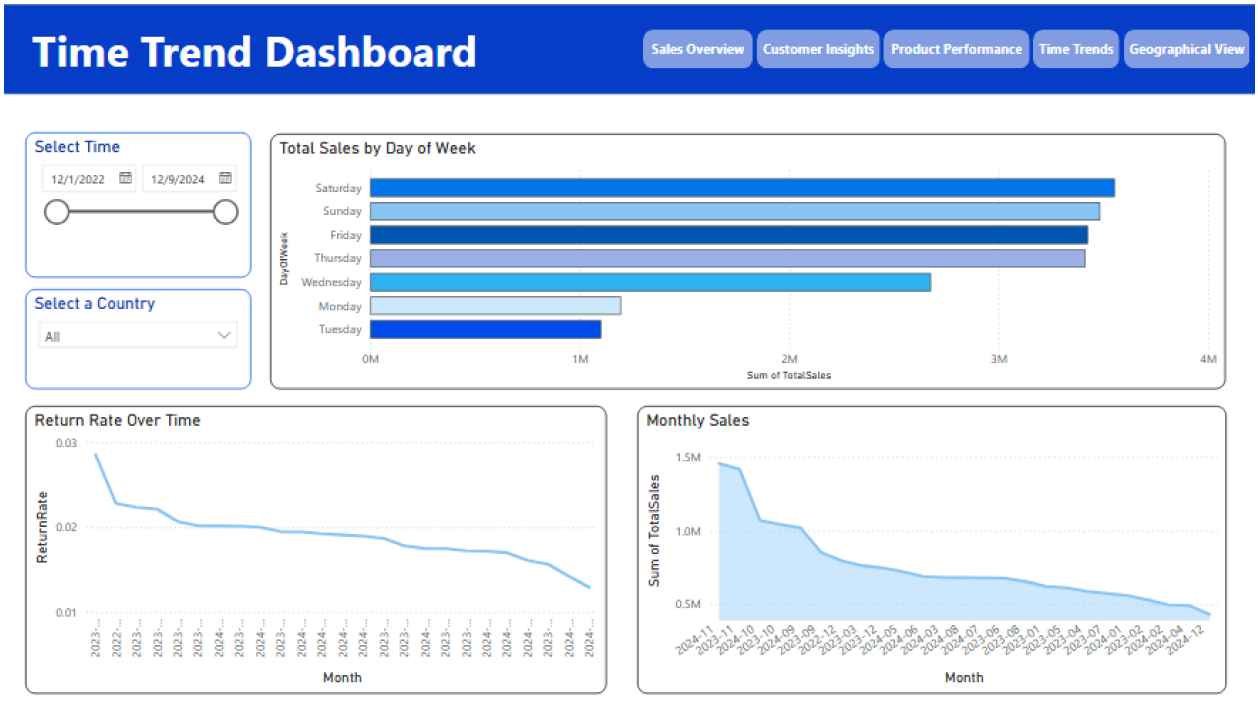


Figure 4: EDA Dashboard –Time Trend Dashboard

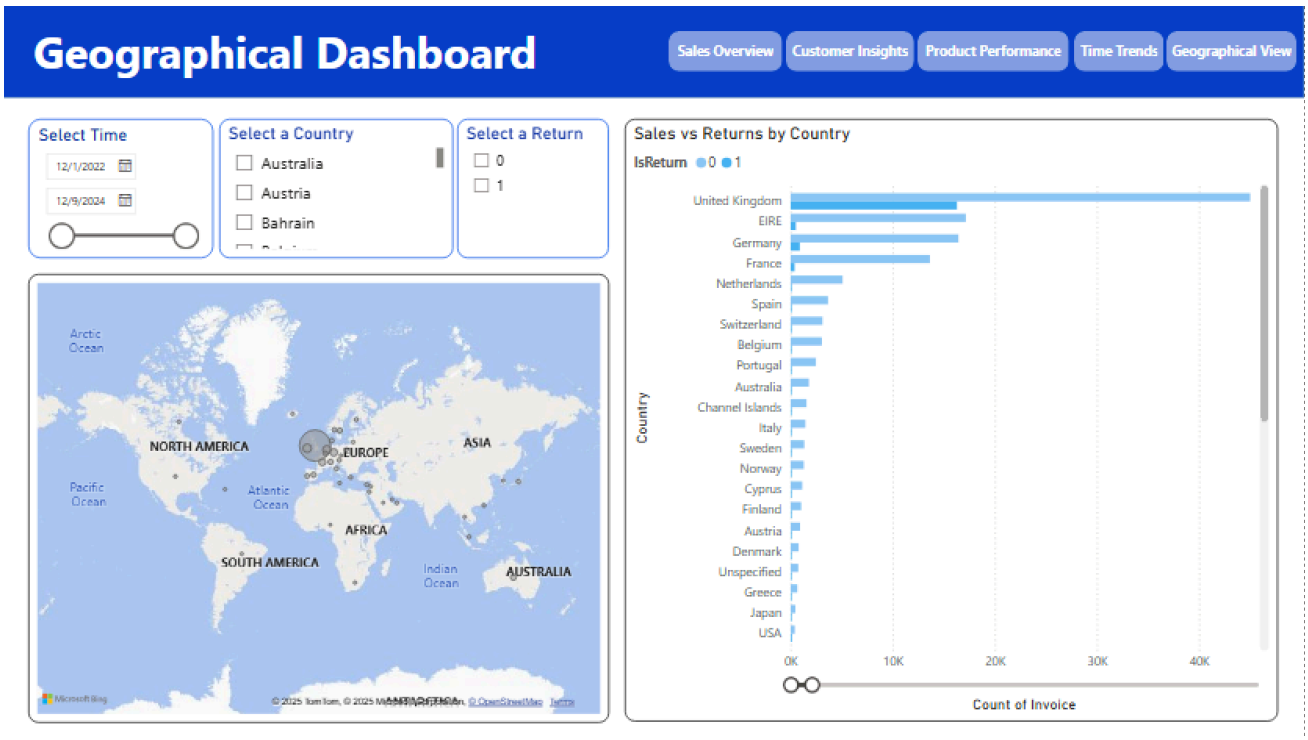


Figure5: EDA Dashboard – Geographical Dashboard

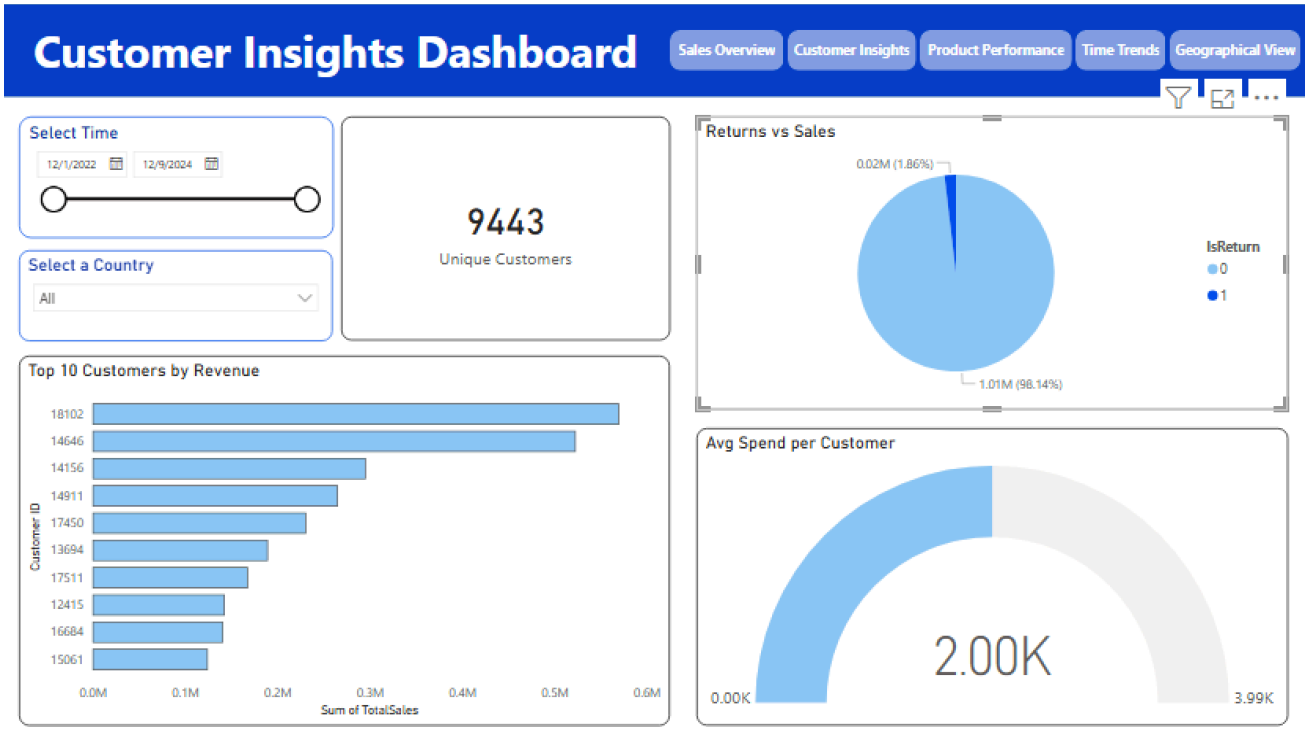


Figure 6: EDA Dashboard – Customer Insights Dashboard