**Course:** BIA 652-WS – Multivariate Data Analysis

**Instructor:** Prof. Thomas F. Brantle

**Team 3 :** Sanjeeth Rao Gona Venkata, Anita Aifesehi, Hariharan Gopikrishnan

Puja Dey, Rohini Mallikarjunaiah

**Semester:** Spring 2025

**Project Title: Understanding Customer Behavior and Churn Risk in E-Commerce Using Multivariate Techniques**

**Table of Contents**

[1. Executive Summary: 2](#_Toc197895040)

[2. Introduction: 2](#_Toc197895041)

[3. Dataset Summary: 2](#_Toc197895042)

[4. Data Preparation: 3](#_Toc197895043)

[5. Descriptive Statistics & Normality: 4](#_Toc197895044)

[6. Factor Analysis (EFA): 8](#_Toc197895045)

[7. Multiple Regression: 10](#_Toc197895046)

[8. Logistic Regression (Churn Prediction): 13](#_Toc197895047)

[9. Cluster Analysis: 14](#_Toc197895048)

[10. MANOVA (Country Comparison): 15](#_Toc197895049)

[11. Insights & Recommendations: 17](#_Toc197895050)

[Appendix: 18](#_Toc197895051)

# 1. Executive Summary:

This project aimed to explore customer purchasing behavior and churn risk using the UCI Online Retail dataset. By leveraging transactional-level data and converting it into customer-level features, the project applied a comprehensive suite of multivariate statistical techniques in SAS. These techniques included exploratory factor analysis to uncover latent behavioral dimensions, multiple regression to predict customer value, logistic regression for churn classification, cluster analysis for behavioral segmentation, and MANOVA to identify regional behavioral differences.

Key behavioral indicators such as Recency, Average Order Value (AOV), Return Rate, and Purchase Frequency were engineered and used across models to analyze patterns and generate actionable insights. Findings revealed clear segmentation between high-value loyal customers and churn-prone segments, as well as significant differences in behavior by country. These insights are intended to support targeted retention efforts, optimized customer marketing strategies, and data-driven decision-making in an e-commerce setting.

# 2. Introduction:

Online retailers face increasing pressure to understand customer behavior to remain competitive in a data-driven environment. One of the primary challenges is predicting customer churn, which can significantly impact revenue and customer lifetime value. In parallel, businesses aim to identify and retain high-value customers, tailor marketing strategies to behavioral patterns, and uncover hidden trends that traditional analysis may overlook.

This project leverages multivariate statistical methods taught in the BIA 652 course to address these challenges. By using customer-level features derived from transactional data, the study applies factor analysis, regression modeling, logistic classification, clustering, and MANOVA to analyze patterns in purchasing, returns, and customer engagement across geographies.

# 3. Dataset Summary:

The dataset used in this study comes from the UCI Online Retail repository and has been processed to support advanced analytics. The resulting insights are designed to empower decision-makers to segment customers, anticipate churn risk, and develop targeted strategies that improve retention and customer value.

The dataset used in this project was obtained from the [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Online+Retail). It contains transactional data from a UK-based online retailer spanning the period from December 2010 to December 2011. Each transaction includes information on the quantity and unit price of products purchased, customer identification, invoice date, and country of origin.

The raw dataset includes over 500,000 rows of transaction-level data. For the purposes of this project, the data was preprocessed and aggregated to the customer level, resulting in a cleaned dataset comprising 4,339 unique customer-level observations.

Key derived features include:

* **Quantity**: Total units purchased
* **Unit Price**: Average unit price of items
* **Total Spend**: Calculated as Quantity × Unit Price
* **Recency**: Number of days since the customer's last purchase
* **ReturnRate**: Proportion of returned items per customer
* **Churn**: Binary flag indicating whether the customer was a one-time buyer
* **Country**: Geographic origin of the customer

A screenshot of a spreadsheet

AI-generated content may be incorrect.

***Figure 1. Original Dataset***

# 4. Data Preparation:

The initial raw dataset contained transaction-level records. Using Python, the data was preprocessed to remove missing values, negative quantities, and irrelevant rows. The dataset was then aggregated to a customer level to prepare for multivariate analysis.

Key derived features were calculated from the base fields as follows, with justifications for their use:

* → This measures the total revenue contribution of each customer, directly reflecting their monetary value.
* → Helps identify how recently a customer engaged, which is critical for churn prediction and lifecycle marketing.
* → Provides insight into the average value of each customer transaction, useful for differentiating between high- and low-ticket buyers.
* if the customer made only one purchase (one invoice), otherwise 0 → Creates a binary outcome for logistic modeling to distinguish repeat customers from one-timers.
* → Captures how often a customer returns items, which may indicate dissatisfaction or risky buying behavior.

These variables were engineered to convert transaction-level behavior into meaningful customer-level metrics. They serve as essential inputs for multivariate analysis by allowing grouping (clustering), prediction (regression), and pattern discovery (factor analysis). Once derived, these metrics were imported into SAS for structured, statistically rigorous analysis aligned with the business objective.

A screenshot of a spreadsheet

AI-generated content may be incorrect.

***Figure 2. Final Dataset for Multivariate Analysis***

# ****5****. Descriptive Statistics & Normality:

**Descriptive** analysis was conducted using PROC MEANS and PROC UNIVARIATE in SAS to examine the central tendencies and distributional properties of key variables such as Quantity, TotalSpend, and ReturnRate.

* **Spend**: Mean = 2053.79, Std Dev = 8988.25
* **Quantity**: Mean = 508.77

This widespread, particularly in TotalSpend, is indicative of skewed purchase behavior, which was confirmed by the normality tests.

* **TotalSpend and ReturnRate** showed extremely high skewness and kurtosis values (skewness > 19 and kurtosis > 400), which indicate significant right-skew and heavy tails.

These results suggest that the data does not follow a normal distribution, necessitating the use of robust multivariate techniques that do not rely on normality assumptions.

A screenshot of a computer screen

AI-generated content may be incorrect.

***Figure 3. Descriptive Statistics Summary Table*  
A graph and a diagram

AI-generated content may be incorrect.**

***Figure 4. Normality Plot for Recency***

**A graph of a distribution plot

AI-generated content may be incorrect.**

***Figure 5. Distribution of TotalSpend (Q-Q Plot)***

***A graph of a distribution and a number of graphs

AI-generated content may be incorrect.***

***Figure 6. Histogram of ReturnRate***

# 6. Factor Analysis (EFA):

Factor analysis was conducted to identify underlying behavioral dimensions among the selected customer-level variables. The method used was Principal Component Analysis (PCA) with Varimax rotation, which simplifies the interpretation by maximizing the variance of factor loadings.

The input variables included:

* Recency (X1)
* Purchase Frequency (X2)
* Total Spend (X3)
* Average Order Value (X4)
* Average Unit Price (X5)
* Return Rate (X9)

The analysis yielded two meaningful components:

* **Factor 1: Spend Behavior** – characterized by high loadings from TotalSpend, PurchaseFrequency, and AvgOrderValue. This factor captures a customer's monetary value and buying intensity.
* **Factor 2: Risk Indicators** – dominated by Recency and ReturnRate. This reflects how recently customers transacted and their tendency to return items, both of which may signal potential churn.

These components provide interpretable constructs that were later used for customer segmentation and profiling.

***A graph of the same type of plot

AI-generated content may be incorrect.***

***Figure 7. Scree Plot of Eigenvalues***

*A screenshot of a graph

AI-generated content may be incorrect.*

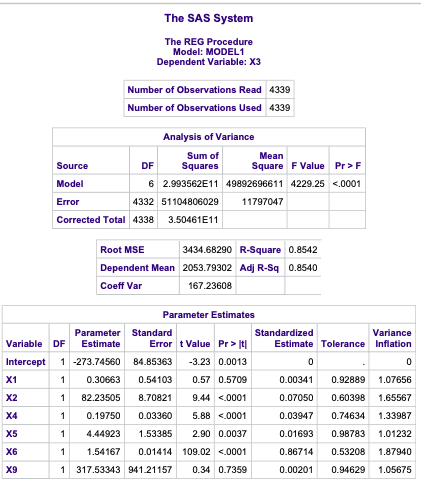
***Figure 8. Rotated Factor Pattern – Varimax Loadings***

# 7. Multiple Regression:

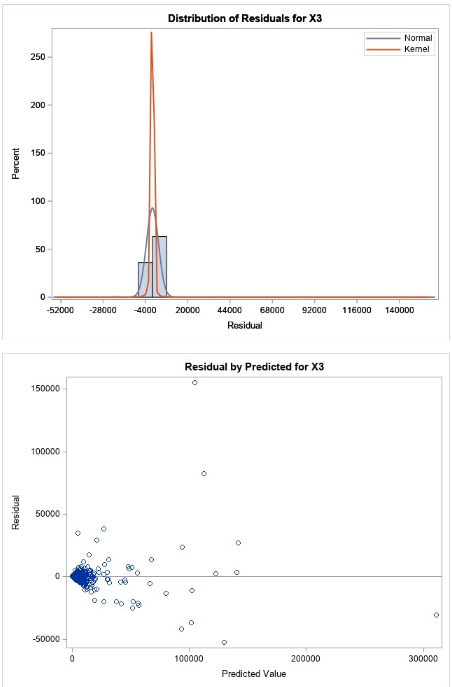
A multiple linear regression model was developed to predict the total customer spend based on key behavioral metrics. The dependent variable was **TotalSpend**, and the selected predictors included **Recency**, **Purchase Frequency**, **Average Order Value (AOV)**, **Unit Price**, and **Return Rate**.

* **R² = 0.854**, indicating that the model explains approximately 85% of the variance in customer spending.
* The **most significant predictors** were **Purchase Frequency** and **AOV**, both of which positively influence total spend.
* **Multicollinearity** was assessed using **Variance Inflation Factor (VIF)** values, and no serious issues were detected.
* Model residuals were checked for normality and randomness using residual plots and univariate analysis.

This regression model serves as a reliable tool for estimating customer value based on observed behavioral variables and contributes to identifying high-value segments.

******

***Figure 9. Regression Coefficients and Fit Statistics***

***Figure 10. Residual Plot and Normality of Residuals***

# 8. Logistic Regression (Churn Prediction):

To predict customer churn, a logistic regression model was developed using Recency, Average Order Value (AOV), and Return Rate as independent variables. Initially, TotalSpend was included; however, it introduced quasi-complete separation, a condition where the model could almost perfectly predict the outcome, leading to unstable estimates. To ensure statistical robustness, TotalSpend was removed, and the model was refit.

The final model showed:

* **Recency (X1)**: Customers with higher recency values (less recent purchases) had higher odds of churn, but the predictor was not statistically significant at p > 0.05.
* **AOV (X4)**: A statistically significant predictor (p < 0.001); lower AOV increased churn likelihood. For every unit increase in AOV, the odds of churn decreased.
* **ReturnRate (X9)**: Though included, this variable was not statistically significant (p > 0.7), but its positive coefficient aligned with expectations that customers with higher return behavior are more churn prone.

The overall model accuracy improved after adjustment, and diagnostics confirmed the fit as reasonable. The model supports strategic decisions to flag high-risk customers based on purchase behavior and guide proactive retention efforts.

A screenshot of a logistic procedure

AI-generated content may be incorrect.

***Figure 11. Logistic Regression Model Summary and Fit Statistics***

# 9. Cluster Analysis:

Cluster analysis was used to segment customers based on behavioral metrics. The method applied was K-means clustering via PROC FASTCLUS in SAS. The input variables included Recency, TotalSpend, Average Order Value (AOV), Unit Price, and ReturnRate. These variables were first standardized to ensure uniform scaling and eliminate bias from differences in units.

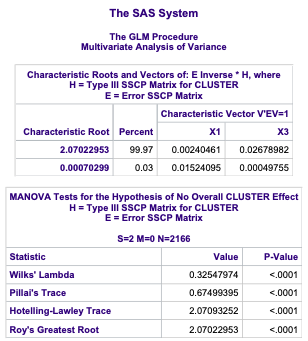
The algorithm grouped the 4,339 customers into **four distinct clusters**, each representing a unique customer profile:

* **Cluster 1: High-Value Loyal** – Customers with high spend, high frequency, low return rate, and recent engagement.
* **Cluster 2: One-Time Buyers** – Customers with low spend and frequency, usually purchasing only once and showing signs of churn.
* **Cluster 3: High Return Rate Segment** – Customers with higher than average returns, moderate engagement, and unpredictable spending.
* **Cluster 4: Moderate Spenders with Recent Activity** – Customers with average spend and frequency but lower recency, indicating they are still active.

# Validation using ANOVA for X1, X3, X4, X5, and X9:

* The ANOVA validation for cluster analysis in this study was performed to statistically confirm whether the identified customer segments (clusters) differ significantly across key behavioral variables. Using the K-means clustering method on standardized variables—Recency (X1), Total Spend (X3), Average Order Value (X4), Average Unit Price (X5), and Return Rate (X9)—the team generated four clusters. To validate these clusters, one-way ANOVA tests were applied individually for each variable against the cluster membership. The GLM procedure in SAS showed that for each variable, the differences among cluster means were statistically significant, as indicated by low p-values in the F-tests. Additionally, Tukey's post-hoc tests were used to identify which clusters differed from each other. This approach ensures that the clusters are not only numerically distinct but also behaviorally meaningful, supporting their use in targeted e-commerce strategies.

# Cluster profiles were validated using MANOVA again to confirm that the differences between clusters were statistically significant.



***Figure 12. MANOVA Results Validating Differences Between Clusters***

A screenshot of a graph

AI-generated content may be incorrect.

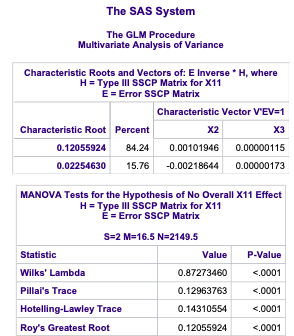
***Figure 13. Distribution of Recency (X1) Across Customer Clusters***

# 10. MANOVA (Country Comparison):

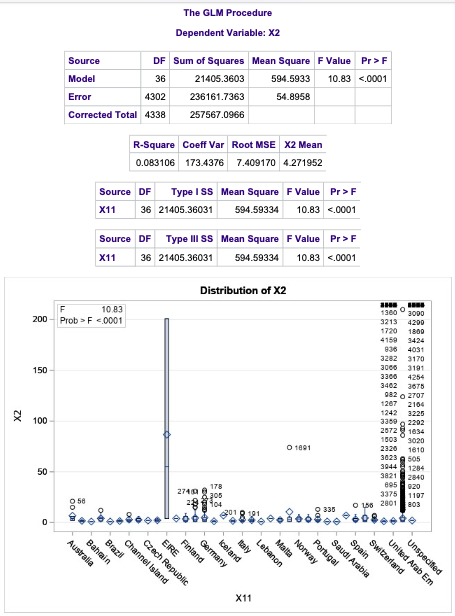
To determine whether customer behavior differs significantly across countries, a Multivariate Analysis of Variance (MANOVA) was conducted using TotalSpend and Purchase Frequency as dependent variables and Country (X11) as the grouping factor. The MANOVA test allows simultaneous comparison of group means across multiple continuous variables, controlling for Type I error.

The results were highly significant, with all multivariate test statistics (Wilks’ Lambda, Pillai’s Trace, Hotelling–Lawley Trace, and Roy’s Greatest Root) yielding p-values less than 0.0001. This confirms that there are statistically significant differences in spending and purchase behavior between countries.

These findings support the need for region-specific marketing strategies and suggest that customer behavior is influenced by geographic factors.



***Figure 14. MANOVA Results for Country-Based Differences***



***Figure 15. Distribution of Purchase Frequency (X2) by Country***

# 11. Insights & Recommendations:

The findings from this multivariate analysis provide several strategic opportunities for improving customer engagement and value generation:

* **Prioritize high AOV and low return segments for loyalty campaigns**: Customers who demonstrate consistent purchasing behavior with high average order values and minimal return activity represent high-value, low-risk segments. These individuals should be targeted with loyalty rewards, exclusive offers, and retention-focused outreach to preserve long-term revenue.
* **Monitor and re-engage recent one-time buyers**: Logistic regression highlighted that customers who made only one purchase are highly susceptible to churn. Engagement strategies such as personalized follow-ups, incentives for second purchases, or onboarding flows could help convert them into repeat buyers.
* **Customize marketing efforts by region**: MANOVA results showed statistically significant differences in purchasing behavior by country. Marketing messages, promotions, and product offerings can be tailored by geography to reflect regional preferences and purchasing power.
* **Use factor-derived behavioral traits to define personas**: Factor analysis revealed two clear behavioral themes: spend behavior and churn risk. These insights can be used to construct behavioral personas (e.g., “value-driven loyalist” or “risk-prone opportunist”) and guide targeted email campaigns, segmentation rules, and UX design for different customer types.

**12. Limitations:**

While the analysis yielded strong insights, there are few limitations to consider:

* **Time Scope**: The dataset includes only one year of transactions (Dec 2010 – Dec 2011), limiting long-term behavioral insights.
* **Return Estimation**: Return behavior was inferred from negative quantities, which may not fully capture actual return dynamics.
* **Missing Attributes**: Key customer attributes like demographics and marketing channel data were not available, limiting personalization potential.

# Appendix:

**Appendix A: Supporting Project Components**

* **Project Proposal** (submitted and approved)
* **Datasource (**[UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Online+Retail))
* **Cleaned Dataset** (Python-processed and SAS-ready)
* **SAS Code Files** (based on course template: Descriptive, EFA, REG, LOGISTIC, MANOVA, CLUSTER)
* **SAS Output** (used for figures, diagnostics, and validation)
* **Professor’s Course Concepts** (including Shapiro-Wilk, VIF, Scree Plot, MANOVA,