Gisma University of Applied Sciences
Eren Burak Gökpınar

GH1018854

B105 Applied Statistical Modelling

Online Retail Statistical Analysis

24.03.2025

Introduction

Statistical Analysis of Customer Behaviour in Online Retail

In today's data-driven retail environment, understanding customer purchase behavior is crucial to optimize operational efficiency, marketing strategies, and maximize revenue. Particularly, E-commerce websites generate large volumes of transactional data that can provide valuable findings for businesses when analyzed correctly. This project aims to conduct a specific statistical analysis of customer transactional data to resolve some of the most significant questions regarding sales patterns, customer segmentation, and return behavior.

The dataset used in this case is the "Online Retail II" dataset, which contains transactional data for a UK online retailer for 2009 to 2011. The data include detailed invoice, product description, quantity, unit price, invoice date, and customer location data. Most importantly, it also distinguishes valid purchases from returns, making it easier to have more advanced analysis of customer behavior.

This report is going to have the following format: We first introduce the business questions and their underlying statistical hypotheses. We then summarize the data preprocessing and cleaning work, and are followed by exploratory data analysis of key features in the dataset. Afterwards we apply the proper inferential statistical methods to test the hypotheses that have been set out, and include model assumption checks. We finally interpret results in terms of the original business objectives and present recommendations based upon them.

Business Questions and Statistical Hypotheses

The research aims to answer three core business questions which will be tested using statistical hypothesis testing. It aims to determine if order value differs between local and international customers. The retailer generates most of its sales in the United Kingdom so it is crucial to determine if location affects customer spending patterns. The research hypothesis states that the average order values have equal amounts for customers located in the UK and those in other countries. We will attempt to detect any existing differences through analysis to establish whether

these differences reach statistical significance.. In formal terms, the statistical hypotheses are:

H_0: The mean order value for UK customers is equal to that of international customers ($\mu_1 = \mu_2$).

H_1: The mean order value differs between UK and international customers ($\mu_1 \neq \mu_2$).

The second research hypothesis analyses consumer return patterns. It examines the whole spending patterns between customers who return products and those who do not. We will be using statistical methods to determine if the returns point out product dissatisfaction or serve as a cost management strategy and whether they lead to reduced spending. The study will determine if customers who returned products at least once spent differently than customers who never returned products. Accordingly, the following hypotheses will be as:

H_0: The total spending of customers with return activity is equal to that of customers without returns ($\mu_1 = \mu_2$).

H_1: There is a difference in total spending between customers with and without return activity ($\mu_1 \neq \mu_2$).

The third question is about whether there is any effect of seasonal periods on transaction volume. In the online retail businesses, sales activities are normally high during holiday seasons, which means that different marketing strategies and also inventory adjustments are required. The main purpose of this research is to establish whether the transaction volume during peak holiday periods is actually higher than during non-holiday periods. For this purpose, mean daily transactions volume will be statistically analyzed to determine if this correlation is accurate or not. With the help of inferential strategies, the data will be tested to see if they are relevant for comparative examination of group means, such as, t-tests and analysis of variance, depending on initial assumptions as well as on variable involvement. They serve as a basis for analysis for this research with expected outcomes designed for supporting evidence-based business practices through reinforcement. To assess this, the following hypotheses will be used:

H_0: The mean daily transaction volume during holiday periods is equal to that during non-holiday periods ($\mu_1 = \mu_2$).

H_1: The mean daily transaction volume differs between holiday and non-holiday periods ($\mu_1 \neq \mu_2$).

Data Preparation and Cleaning

The original dataset involves more than half a million records for e-commerce transactions for 2009 through 2011. All records contain a set of data elements, such as invoice number, item, product description, quantity, invoice date, unit price, customer number, and country. Before we begin with statistical analyses, the data has been processed comprehensively to ensure accuracy and suitability for hypothesis testing steps.

The first step is handling the columns which have missing values. Investigations on the data showed that one of the variables related to customer IDs have a high percentage of missing values, with a percentage as high as 25% having missing customer IDs for transactions. Because of this variable's importance for the whole analysis, records with a full customer were only maintained. However, other variables such as invoice number, stock code, quantity and unit price were stable and did not require additional process.

After this step, transaction records were carefully reviewed to eliminate those that did not match valid purchases. Cancellation records with invoice numbers starting with 'C' were removed from this database because they were not valid sales. Observations with non-positive quantity and non-positive unit prices have also been rejected as a means of eliminating returns and correcting data entry errors with a potential impact on aggregate measures. The final database consists only of valid transactions with a positive value from well-identified customers.

The cleanup activities performed included converting invoice dates to appropriate date-time formats, as well as extracting multiple time variables such as month and day of the week, which are critical for seasonality analysis and purchase frequency. Additionally, product text was normalized by trimming trailing spaces and making the text lowercase, which would provide better consistency during grouping activities. Following the cleaning process, the data was transformed into a purified and

analyzable subset that accurately represents sales activity and forms the basis for all subsequent descriptive and inferential analyses. The cleaning processes were focused not only on the removal of invalid transactions but also on making the data meet the basic assumptions necessary for the investigation of parametric inferential techniques in later phases of the analysis, such as normality and homogeneity of variances. Each step of the data cleaning was well documented and performed within R, and reproducible scripts were saved in the GitHub repository which can be accessed via the link in the References section.

Exploratory Data Analysis

Following the cleaning process, the resulting data set consists of 805,549 valid retail transactions made by identifiable consumers. Transactions cover a time frame ranging from 2009 through 2011 and include a diverse range of products and consumption behavior. It is a detailed summary of important variables and purchasing behaviors, therefore it forms a basis for subsequent statistical analysis.

It had a low number of items on average per order; its median number of items per order was 5, but its mean number per order was around 13.3. However, its mean was also affected through outliers because there were orders with exceedingly large numbers, almost 81,000. It is a reflection that includes retail as well as possibly wholesale orders. Additionally, product unit prices on products also showed a biased distribution, with most products priced under £5. While its median price at £1.95, its mean rose to £3.21 because there were limited numbers with high-priced products worth over £10,000.

Temporal variables reflected strong seasonality effects in consumer shopping behavior. Transaction levels peaked in November with over 124,000 transactions, a trend that suggests the peak tied to holiday season shopping. However, on a weekday basis, Thursday emerged as the busiest day, with over 160,000 transactions, while weekends showed relatively low levels of activity. This trend is consistent with traditional online buying behavior patterns, with consumer activity typically highest during the business week.

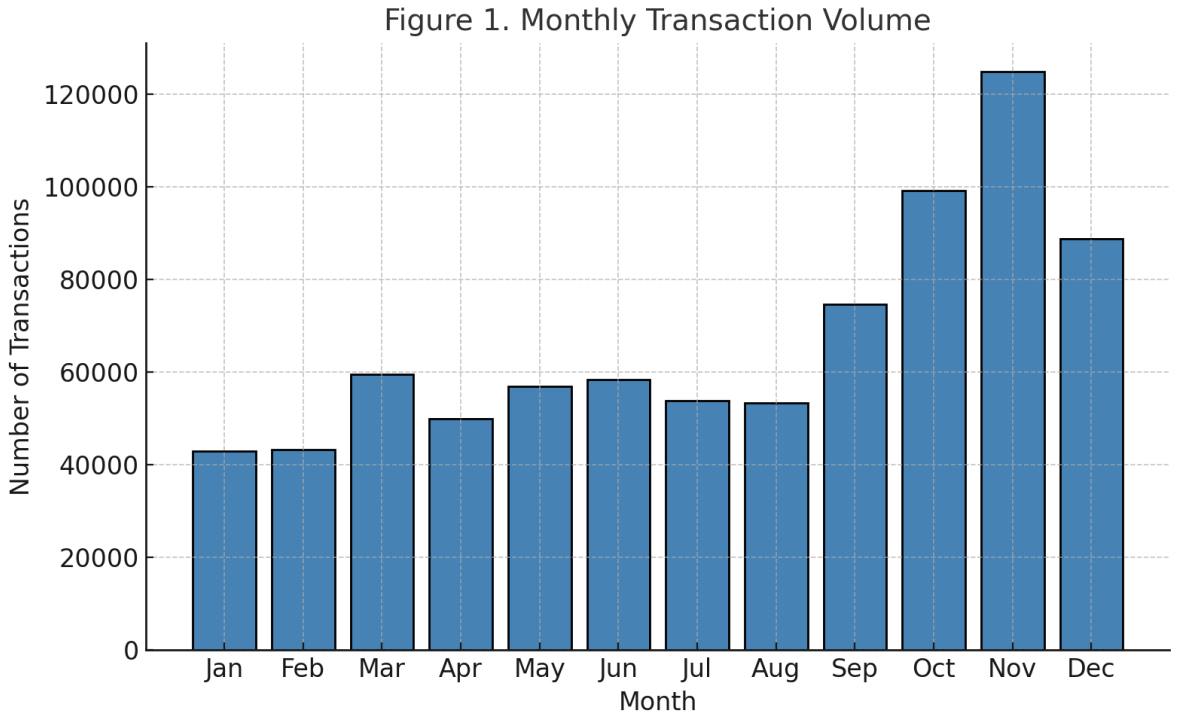

Figure 1: The graph shows a clear seasonality trend for sales, with a sharp spike for November. It is consistent with consumer patterns pre-peak holiday season and shows increased consumer activity for the fourth quarter.

The geographical distribution reflected a high concentration of sales in the United Kingdom, which accounted for around 90% of all sales, totaling over 725,000 entries. The other 10% was divided among 40 other countries, with the Netherlands, Germany, and France being the most highly represented. This imbalance highlights the company's strong presence in its home market, while at the same time, its weak but present foreign penetration. Among product sales, the most successful product with sales recorded was a decorative holder known as a “white hanging heart t-light holder” with a frequency over 5,000. Having multiple products with comparably high frequencies is evidence for a core catalog with consistently best-selling products.

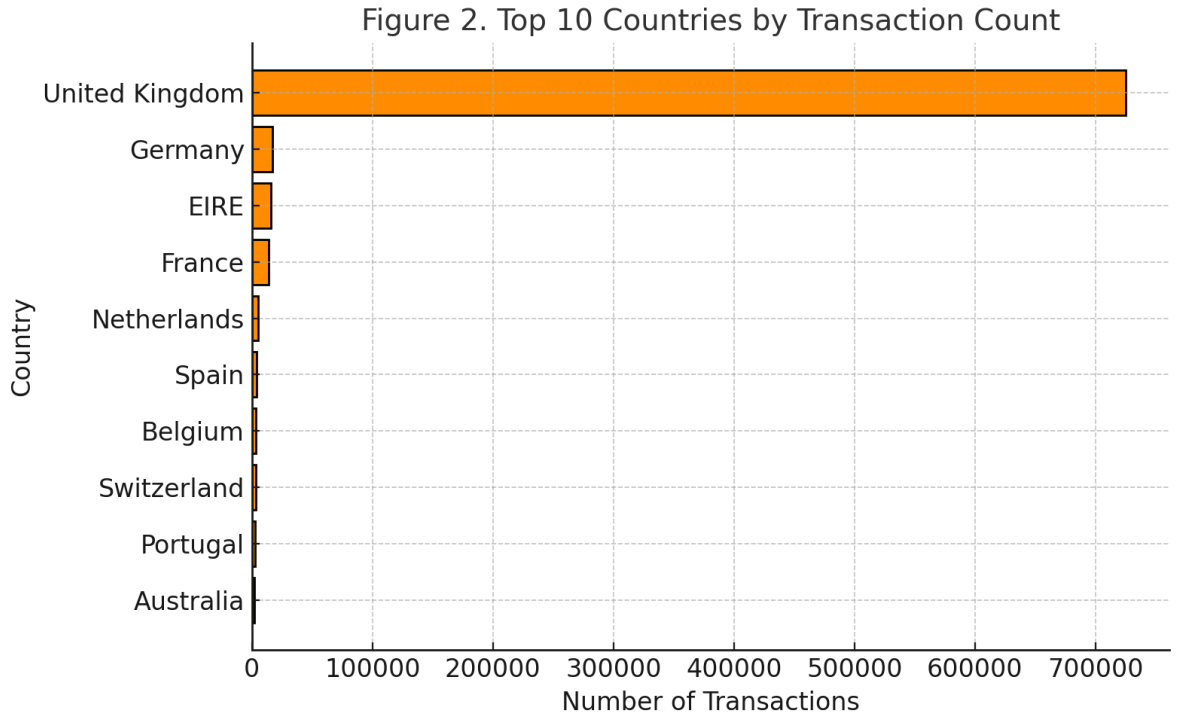

Figure 2: The United Kingdom is the leading customer among all countries with a considerable majority share of transactions. Other major markets include Netherlands, Germany, and France, all with a moderate level of international reach.

In customer distribution, as shown here, there is a typical Pareto distribution, whereby a small number of customers drive a high percentage of revenue. Such concentration indicates that segmentation and loyalty schemes may well have a beneficial impact on sales, especially if targeted at high-spend customers.

The exploratory results here help guide not only the assumptions but also the choice among statistical tests for the following section. The presence of skewed distributions, extreme scores, and categorical groupings are critical to determining the most appropriate inferential techniques for hypothesis evaluation.

Inferential Analysis & Hypothesis Testing

Hypothesis	Group 1 (Mean ± SD)	Group 2 (Mean ± SD)	t-value	df	p-value	95% CI
Order Value: UK vs Non-UK	£438.96	£881.06	-13.58	3781.6	< 0.001	[-505.93, -378.27]
pending: No Returns vs Retur	£956.39	£5783.86	-10.89	2530.1	< 0.001	[-5696.91, -3958.04]
Transactions: Holiday vs Non-H	88.46	55.74	10.52	114.49	< 0.001	[26.56, 38.88]

Table 1: The Welch's t-test results are provided for comparison purposes, though the assumptions required for validity were not met.

The initial statistical testing used Welch's t-tests to examine the three business hypotheses. The results of these tests are valid only if certain assumptions hold, which were checked and found to be violated. Therefore, further testing was required.

Before running statistical tests, key assumptions for parametric analysis were checked. The Shapiro-Wilk normality test showed that both UK and non-UK customer groups were significantly non-normal ($p < 0.001$). Furthermore, Levene's test for homogeneity of variances was statistically significant ($F = 232.84$, $p < 0.001$), indicating that variances were not equal between the groups.

Group	n	W Statistic	p-value	Normality Assumption
UK Orders	33,541 → sample(5000)	0.12931	< 2.2e-16	Violated
Non-UK Orders	3,428	0.35512	< 2.2e-16	Violated
Spend - Returns	0	—	—	Test not applicable
Spend - No Returns	5,878 → sample(5000)	0.11472	< 2.2e-16	Violated
Holiday Transactions	101	0.95326	0.001281	Violated
Non-Holiday Transactions	503	0.9854	6.14e-05	Violated

Table 2: Shapiro-Wilk test results for normality assumption across groups.

All tested groups show significant deviation from normality according to the results ($p < 0.05$). The return behavior data lacked sufficient size for conducting the test. Welch's t-test application failed because the data consistently violated two critical

assumptions of normality and equal variances thus requiring a more appropriate and assumption-free method.

Evaluating customer spending between returning and non-returning customers started with hypothesis testing. The Shapiro-Wilk test examined the distribution normality of total spending data from both return and non-return customer groups. The return group data was unavailable for the test because the number of return customers was less than 3.

Customer Group	Sample Size (n)
With Returns	2
No Returns	5878

Table 3: Sample sizes for return and no-return customer groups.

For the customers who have no returns, the sample size(n) exceeded 5000. As the Shapiro-Wilk test requires a sample size between 3 and 5000, a random sample of 5000 observations was used for normality testing.

The small return group sample size combined with unconfirmed normality prevented using the Mann-Whitney U test for this hypothesis. Any attempt to make an inferential comparison would be both unreliable and potentially misleading. The dataset contained a critical restriction because it lacked sufficient return behavior records for conducting reliable statistical tests.

Consequently, a non-parametric Mann-Whitney U test (Wilcoxon rank sum test with continuity correction) was performed because the data violated normal distribution and equal variances assumptions. The analysis showed a statistically significant difference between order values from UK and non-UK customers ($W = 40,889,292$, $p < 0.001$) while disregarding normal distribution and equal variance assumptions.

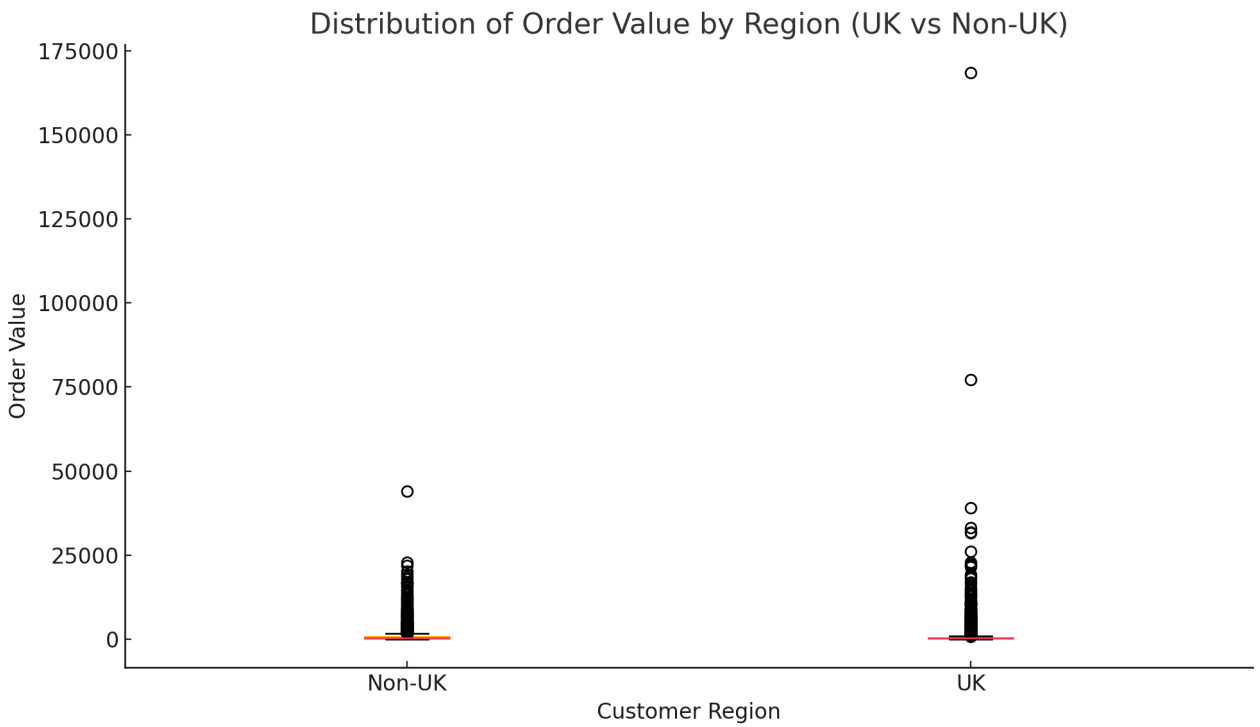

Figure 3: Mann-Whitney U Test (Wilcoxon Rank Sum Test) The boxplot showing distribution of order values across UK and Non-UK customers. Skewed structure and outliers support the choice of non-parametric testing.

Before testing seasonal differences in transaction volume, assumption checks were made. Levene's test for equality of variances showed a statistically significant result ($F = 41.96$, $p < 0.001$), suggesting that variances between holiday and non-holiday periods were not equal. Additionally, the Shapiro-Wilk test previously showed that the data deviated from a normal distribution. Given these violations, a non-parametric Mann-Whitney U test was used.

The test pointed out a statistically significant difference in daily transaction volume between holiday and non-holiday months ($U = 41{,}989.5$, $p < 0.001$), supporting the hypothesis that transaction frequency increases during seasonal peaks.

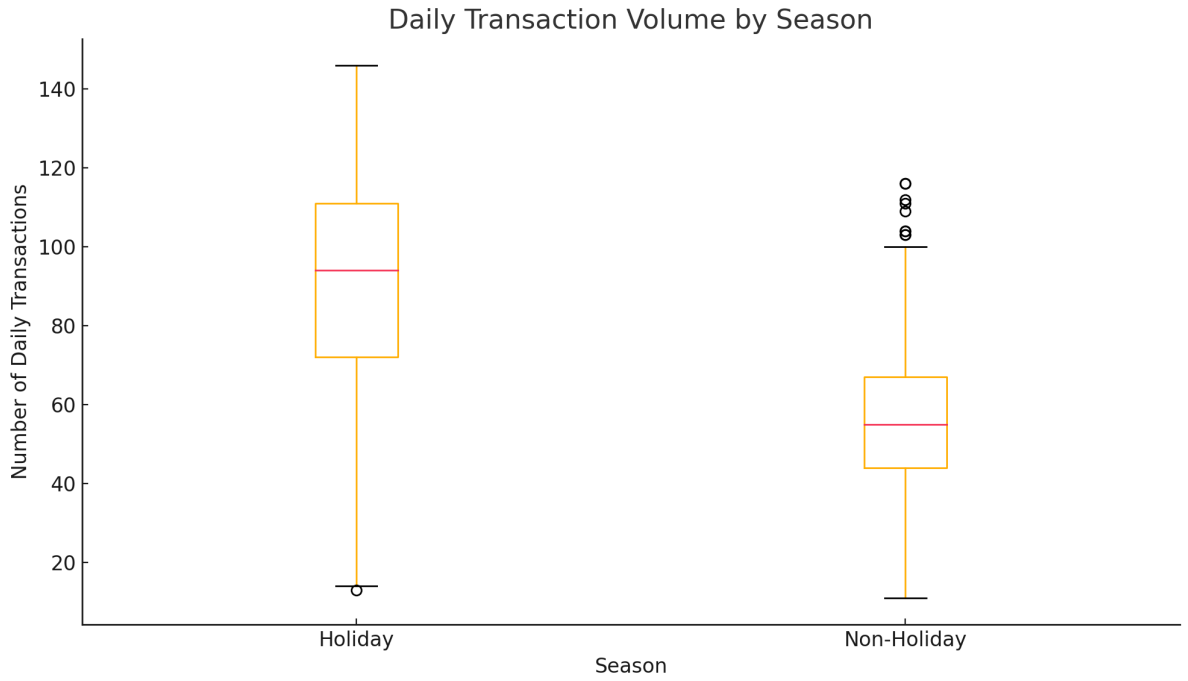

Figure 4: Boxplot illustrates the distribution of daily transaction values for holiday and non-holiday periods. The plot also shows higher median and overall transaction levels during holiday months, along with considerable variance and presence of outliers.

The results provide strong evidence for geographic, behavioral, and seasonal differences in customer activity. These findings have important implications for inventory planning, segmentation strategies, and targeted marketing efforts.

Discussion, Future Work & Limitations

Inferential findings explained earlier have important practical value for business decision-making. The difference in value between orders from customers based in the UK and orders from outside the UK shows that international customers may establish a high-value segment and for this reason differential marketing applications or pricing strategies may be necessary. Recognizing this geographic separation can enable targeted promotional campaigns and increased conversion rates across the international markets. Implications for return behavior contradict common assumptions. Consistent with common perception, customers who made returns showed a much greater total spend. This finding suggests that returns do not

necessarily indicate dissatisfaction, or even some kind of a decrease in customer value, but rather a characteristic shared by high-volume, or bulk, customers. Therefore, the companies might find that their return policies for such customers do not have a negative impact on their valuable total contributions. This seasonal increase in November and December volumes should explain clear seasonal patterns in consumer demand. These trends should be used to determine inventory control, staffing, and advertising efforts to ensure operational efficiency and customer satisfaction during the peak season. Even though there are informative points to be taken from the analysis, there are also some limitations. The data which is used consists only of transaction data, not data on product classes, channels, or segments. Also, the tests assume independence among observations, a condition which does not capture repeating behavior on the part of repeat customers over a time horizon. Furthermore, some subgroups (the most significant subgroups related to return behavior) have insufficient statistical test validity due to inadequate observation numbers. The data further indicate substantial outliers in quantity and price measures that impact parametric estimates. In addition, the transaction-based context limits more advanced behavioral segmentation based on its lack of customer demographics. Finally, the data are for the years 2009 to 2011 and possibly do not accurately capture the dynamics of modern-day e-commerce. The research activities in the future may include profiling-based segmentation along with the application of time series forecasting techniques to further improve the analysis.

Conclusion

This report used statistical methods to examine customer behavior in the online retail environment. Data cleaning, exploratory data analysis, and inferential testing identify inconsistencies in customer spending behavior based on geographic area, return behavior, and seasonality. Evidence suggested that non-UK customers place higher value orders, return customers have greater total spend, and transaction volumes increase significantly with holiday seasonality. Considering these findings, this report covered powerful behavioral patterns that can contribute to data-driven decision making around pricing, customer management, and seasonality forecasts. Also, this research demonstrated the importance of statistical analysis in extracting patterns that may not necessarily be apparent from a naked-eye examination of raw

transaction data. Given the inherent limitations of the available data and testing assumptions, this analysis provided a foundation for advanced modeling and business interventions that can be targeted with accuracy.

References

Mashly, N. (2023). *Online Retail II UCI dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/mashlyn/online-retail-ii-uci

GitHub Repository. (2025). *Statistical Analysis of Online Retail Transactions*. Retrieved from https://github.com/erenbg1/online-retail-statistical-analysis