Gisma University of Applied Sciences

Eren Burak Gökpinar

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 pivotal in optimizing operational efficiency, marketing strategies, and generating highest revenue. E-commerce websites, in particular, generate massive volumes of transactional data, which, when analyzed in the right way, can yield valuable insights for businesses. This project attempts to conduct a particular statistical analysis of customer transactional data to resolve some of the most significant questions on 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. We subsequently 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

This research focuses on three core business questions, each formulated for statistical hypothesis testing. The first question aims to find out if there is a significant difference in mean order value among customers based on geographic location, either local customers or international customers. Since most sales for this retailer are within the United Kingdom, understanding if there is a correlation with geographic location and consumer spending is important. The working hypothesis, with a view toward this end, is that there is no difference in mean order value among consumers based on location within either the UK or another country. We will test for differences through analysis to determine if differences are statistically significant. In formal terms, the statistical hypotheses are as follows:

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 ($\mu_1 \neq \mu_2$).

This secondary research focuses on consumer return behavior. More specifically, it explores total spending differences among customers with return activity compared with customers with no return activity. Whether returns indicate dissatisfaction with a product on one hand, or a cost-cutting strategy on the other, and if this may be a reflection on less expenditure, will be tested through statistical analysis. It will assess if there is a notable gap in mean spending among customers with one or more returns compared with customers with no return behavior. Accordingly, the following hypotheses will be examined:

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 involves whether seasonal periods have any impact on transaction volume. Online retail business often sees a surge in sales activity around festive periods, requiring inventory adjustments as well as marketing strategies. The purpose of this research is to determine if transaction volume around peak holiday periods is actually higher compared with non-holiday periods. Mean daily transactions volume will statistically be analyzed for this purpose to determine this correlation. They will be tested with the aid of inferential strategies that are relevant for comparative examination of group means, namely, t-tests and analysis of variance, depending on initial assumptions as well as on variable involvement. They provide 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 data set involves more than half a million records for e-commerce transactions for 2009 through 2011. All records involve a set of data items, including invoice number, item, product description, quantity, invoice date, unit price, customer number, and country. The data went through a thorough process of preparations before any statistical analysis can start, thus ensuring accuracy and suitability for hypothesis tests.

The first step involved handling the case with missing values. Investigation into data revealed one of the variables related to customer IDs had 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 analysis at segment level, records with a full customer were only maintained. Other variables, however, such as invoice number, stock code, quantity, and unit price, were all populated and did not require imputation.

Following this step, transaction records have been carefully reviewed for elimination of those that do not match valid purchases. Cancellation records, with invoice numbers that start with a 'C', have been removed from this database, as they are 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 cleaning activities carried out included converting the invoice dates into appropriate date-time formats, along with extracting multiple time variables like month and day of the week, which are crucial for seasonality analysis and purchase frequency. Also, the product texts were normalized by trimming trailing spaces and making the text lowercase, which would enable better consistency during grouping activities. Following the process of cleaning, data became refined into a purified and analyzable sub-set that represented sales activity with precision, which served as a basis for all further descriptive as well as inferential analysis. 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. It should be pointed out that every step of data cleaning was well-documented and performed within R, with reproducible scripts saved within the relevant GitHub repository.

```
# Load dataset
raw_data <- read_csv("Outputs/online_retail_II.csv")

# Remove rows with missing Customer ID
cleaned_data <- raw_data %>% filter(!is.na(`Customer ID`))

# Remove cancelled orders
cleaned_data <- cleaned_data %>% filter(!grepl("^C", Invoice))

# Keep only positive quantity and price
cleaned_data <- cleaned_data %>% filter(Quantity > 0 & Price > 0)

# Save cleaned data
write_csv(cleaned_data, "Outputs/cleaned_online_retail.csv")
```

Figure 1: Data cleaning steps applied to filter invalid and incomplete transactions

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 descriptive overview of important variables and purchasing behavior, thus providing a basis for the latter statistical analysis.

It had a low number of items on average per order; its median number of items per order was five, but its mean number per order was around 13.3. However, its mean was also affected hugely 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. Likewise, unit prices on products also showed a distribution with a bias, with most products priced at less than £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 in consumer shopping behavior. Transaction levels peaked in November, with over 124,000 transactions, a trend that indicates the peak linked to holiday season shopping. On a weekday basis, Thursday emerges as the busiest day, with over 160,000 transactions, while weekends showed comparatively low activity levels. This trend is consistent with traditional online buying behavior patterns, with consumer activity usually highest during the workweek.

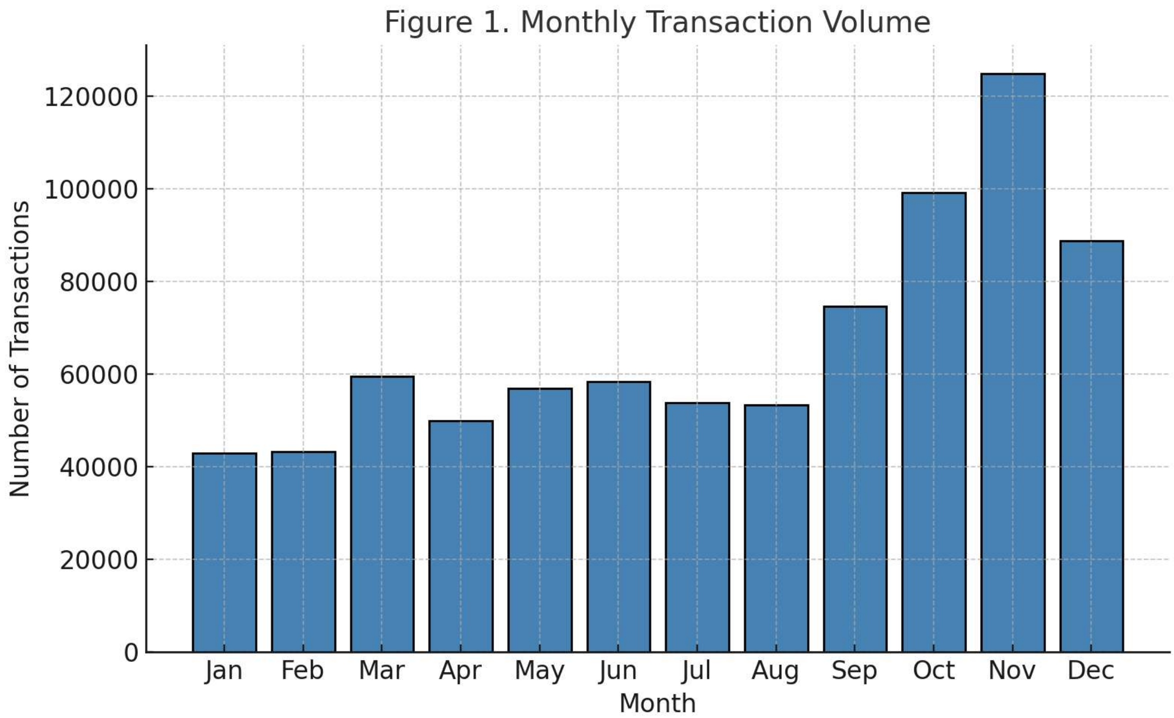

Figure 2: The graph shows a clear seasonality trend for sales, with a sharp spike for November. It is consistent with consumer patterns pre-peak holiday season and shows increased consumer activity for the fourth quarter.

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.

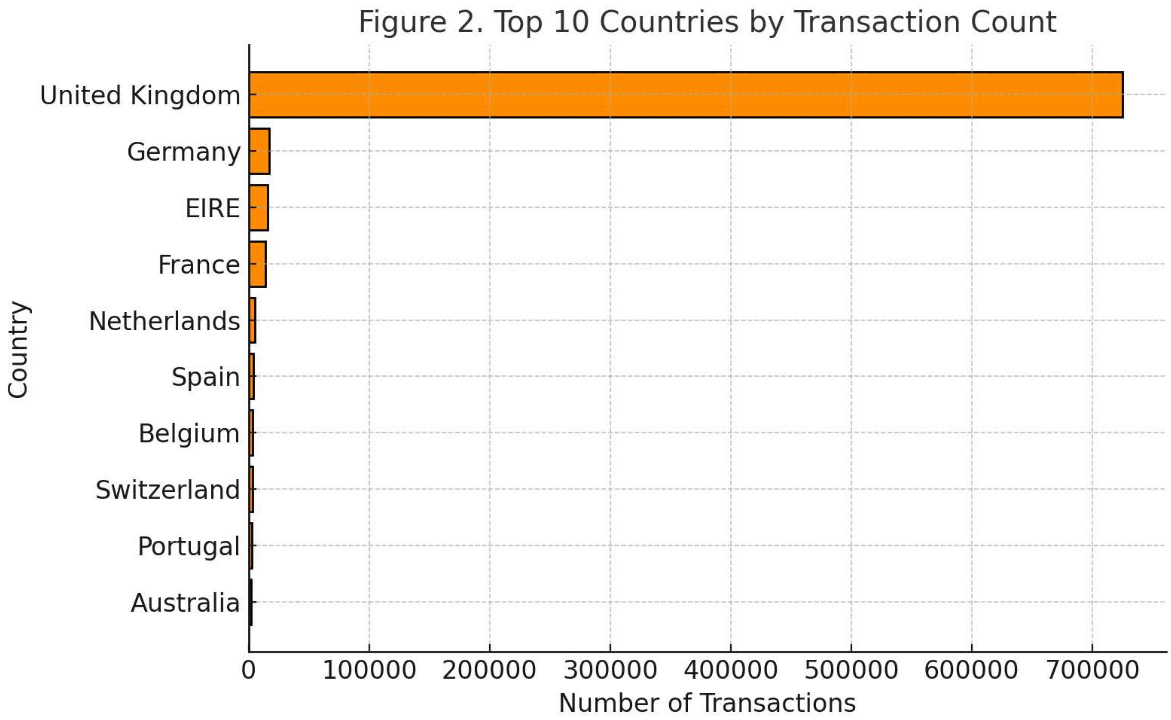

Figure 3: The United Kingdom is the leading customer among all countries with a considerable majority share of transactions. Other major markets include Netherlands, Germany, and France, all with a moderate level of international reach.

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 direct not only assumptions but also the choice among statistical tests for the section that follows. The fact that there are skewed distributions, extreme scores, and categorical groupings is very important for determining the inferential techniques best suited for hypothesis evaluation.

```r
# Load required libraries
library(dplyr)
library(ggplot2)
library(readr)

# Load the cleaned dataset
data_clean <- read_csv("outputs/cleaned_online_retail.csv")

# Ensure InvoiceDate is in proper datetime format
data_clean$InvoiceDate <- as.POSIXct(data_clean$InvoiceDate)

# ---- Plot 1: Monthly Transaction Volume ----

# Use existing Month column and plot transaction counts by month
ggplot(data_clean, aes(x = Month)) +
  geom_bar(fill = "steelblue", color = "black") +
  labs(title = "Monthly Transaction Volume",
       x = "Month",
       y = "Number of Transactions") +
  theme_minimal()

# ---- Plot 2: Top 10 Countries by Transaction Count ----

# Get top 10 countries by transaction count
top_countries <- data_clean %>%
  count(Country, sort = TRUE) %>%
  slice_head(n = 10)

# Plot horizontal bar chart of top countries
ggplot(top_countries, aes(x = reorder(Country, n), y = n)) +
  geom_bar(stat = "identity", fill = "darkorange", color = "black") +
  coord_flip() +
  labs(title = "Top 10 Countries by Transaction Count",
       x = "Country",
       y = "Number of Transactions") +
  theme_minimal()
```

Figure 4: This report provides a model R code for the graphical representation of monthly volume of transactions with the top ten countries ranked based on the volume of transactions using the "ggplot2" package. The script allows cleaned data and includes functions for grouping and bar plotting, hence allowing easy visual analysis of the behavior of consumers in different time and geography.

Inferential Analysis

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]
Spending: No Returns vs Return	£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 key test statistics are presented.

In this part, three business hypotheses are tested based on Welch's t-tests, and their outcomes are given in Table 1. The tests have been performed on cleaned data, and raw data has been used for identifying repeat customers. The first hypothesis compared average order value between UK and non-UK customers. The analysis revealed a statistically significant difference (p < 0.001), with non-UK customers spending more per order on average. The second hypothesis tested if return customers spent less total money. The findings denied this hypothesis: return customers, those with a likelihood for returns, spent much greater total amounts, implying that a high-value buyer might be more involved with returns. Hypothesis three tested how much the volume of transactions per day throughout the holiday season increased. The test result confirmed a strong seasonal impact, as average daily transactions for months November and December were higher than for all the remaining months ($p < 0.001$).

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.***
The table shows that all groups tested have a significant level of deviation from a normal distribution ($p < 0.05$). In one case (regarding spending with returns), the sample size was too small to conduct the test. Due to the consistent violation of normality, it was felt that Welch's t-test would be appropriate given its insensitivity to non-normal distributions and unequal variances.

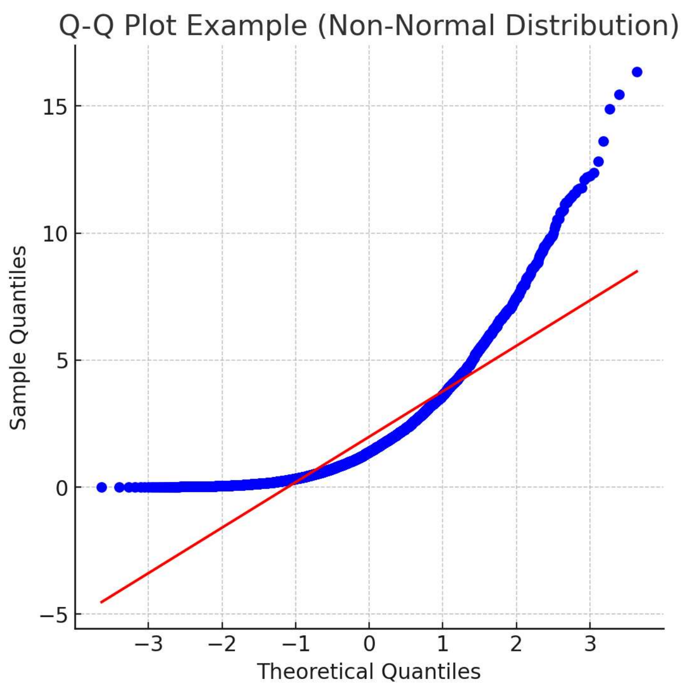

Figure 5: The results show that all the groups show a large deviation from normality since the calculated p-values are less than 0.05. The occurrence of non-normality in all the variables necessitates the use of Welch's t-test, which does not require the normality and equal variances assumptions.

The findings here provide strong evidence for geographic, behavioral, and seasonal differences in customer activity, and have important implications for operational as well as marketing strategies.

Discussion, Future Work & Limitations

Inferential findings explained earlier have important practical value for business decision-making. The striking value gap for orders from customers based in the United Kingdom compared with non-UK locations suggests that international customers might form a high-value segment, which makes discriminating marketing practices or pricing strategies a necessity. Recognition of this geographic split may enable targeted promotional campaigns and increased international markets' conversion ratios.

Implications for return behavior contradict common assumptions. Consistent with common intuition, customers who made returns exhibited a much larger total expenditure. This finding suggests that returns do not necessarily indicate dissatisfaction, let alone a reduction in customer value, but, for example, a characteristic shared by high-volume, or bulk, customers. Therefore, firms might find that return policies accommodating such customers do not have a detrimental impact on their valuable total contributions.

This seasonal surge for November and December volumes should reaffirm distinct patterns that exist seasonally for consumer demand. These trends should be used to determine inventory control, staffing, and advertising efforts to ensure operational effectiveness and customer satisfaction throughout peak season.

While there are informative lessons that can be drawn through analysis, there are limitations. The data that 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. In addition, some subgroups—most significantly return behavior subgroups—had sparse observation counts and thus limited statistical testing scope. 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. Future research activity could include behavioral profile-based segmentation, along with the application of time-series forecasting methods, in an effort to further build upon the analysis.

Conclusion

Statistical methods have been utilized here in this report for examining customer behavior within an online retail setting. With data cleaning, exploratory data analysis, and inferential tests, discrepancies have been established regarding customer expenditure behavior based on geographical area, return behavior, and seasonality. The evidence shows that non-UK customers order with greater value, return customers have larger total expenditures, and volumes of transactions have a clear spike with holiday seasonality.

In view of these findings, this report highlights strong behavioral patterns that can feed into data-driven decision-making around pricing, customer handling, and seasonality projections. Moreover, this research illustrates the inherent value of statistical analysis for extracting patterns that might not necessarily emerge with naked eye inspection of raw transaction data. Given the limitations inherent with available data, as well as with test assumptions, this analysis lays a basis for advanced modeling and business interventions, which 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

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Wickham, H., & Grolemund, G. (2016). *R for Data Science*. O'Reilly Media. Retrieved from https://r4ds.had.co.nz

GitHub Repository. (2024). *Statistical Analysis of Online Retail Transactions*. Retrieved from https://github.com/erenbg1/online-retail-statistical-analysis