From Segmentation to Prediction: A Unified RFM–Machine Learning Model for Online Retail Analytics

[1]Olumide Simeon Ogunnusi, [2]Kayode A. Akintoye

[1,2]Department of Computer Science, The Federal Polytechnic Ado Ekiti, Ekiti State, Nigeria

DOI: https://doi.org/10.51584/IJRIAS.2025.101100117

Received: 04 December 2025; Accepted: 11 December 2025; Published: 23 December 2025

ABSTRACT

Customer analytics features in the digital retail age that provides an understanding of the buying behaviour and makes improvements in marketing practice. As e-commerce platforms have massive volumes of transaction data, predictive analytics can provide an effective customer segmentation and future behavioural prediction methodology in doing business. Most online retailers still do not have a means to identify high-value customers correctly or forecast repeat buy. Even though they have access to detailed transaction logs, they cannot use the information effectively. The lack of this insight can easily lead to ineffective targeting and thus unreached revenue potential. In this paper, the researcher tries to study purchase behaviour segmentation with the help of RFM (Recency, Frequency, Monetary) and employ machine learning capabilities to learn customer purchasing behaviour. Using the Online Retail dataset, this study uses data preprocessing, customer segmentation based on RFM scoring, Logistic Regression and Random Forest models to predict the future purchasing behaviour. In the analysis, the frequency and monetary score of customers provide a strong indication of the ability of customers to make repeat purchases. Of all the classifier models tested, Random Forest was found to be more effective and accurate than the precision of other models confirming its suitability in loud customer tests. The findings are beneficial to e-commerce websites that wish to maximize on the strategy of personalized marketing and management of customers. The opportunity to increase predictive accuracy even more is to address hybrid models and real-time segmentation in the future studies.

Keywords—RFM segmentation, predictive analytics, customer behavior, machine learning, online retail

INTRODUCTION

In this digital business era, companies have produced tremendous transactional information using online retail stores. This is information with precious data regarding the customer purchasing behaviour which can be leveraged to attain strategic insights (Smith & Rupp, 2003). One of the most important data analytics areas is Social and Web Analytics: the use of digital records to come to know about patterns, user segmentation, and future activity prediction (Theodorakopoulos & Theodoropoulou, 2024). RFM (Recency, Frequency, Monetary) analysis is one of the most common methods used in the area, and it segments customers according to how new they are to the company and how often they make a purchase, as well as how much they spend. This technique, once mixed with machine learning should help increase the degree of accuracy in customer behavior prediction by quite a margin (Adekiigbe et al., 2011; Paramasivan, 2024). Most companies, even those with access to transaction data, do not take to elaborate marketing strategies because they lack the ability to perform adequate segmentation, and efficient customer behaviour forecasting(Anand et al., 2023; Ogunnusi et al., 2018) .

The conventional methods are not always sufficiently able to pick up subtle consumption habits and lack the scalability to use in decision automation. Moreover, the practice of combining the segmentation approaches with the predictive machine learning algorithms in the real-life scenarios of the retail is sparse. This discrepancy restricts the possibility of customization and enhancing customer retention. This study aims at carrying out RFM-based customer segmentation of the Online Retail data and utilize the supervised machine learning models, i.e., Logistic Regression and Random Forest, to forecast the possibility of subsequent purchases (Ogunnusi et al., 2015b; Sagar, 2024). The purpose of the analysis will be to define valuable customer segments and come up with insights to enhance customer engagement strategies.

The current research is founded on Online Retail dataset, comprising of transactions of a UK-based online shop over the period of December 2010 to December 2011. It targets those customers that have credible identification numbers and are not ones who have had their order cancelled. In the study, RFM scoring is used to segment the data, and the feature is fed to classification models. Also, product-level recommendations, sentiment analysis, forecasting time series, and cross-channel behaviour analysis are not included within their scope (Ajibade, Jasser, Alebiosu, et al., 2025). The ability to monitor the purchasing patterns of customers and to predict their further purchases is important in case of enhancing customer relationship management and in case of increasing the marketing ROI. The work is important to the area as it unites the segmentation and predictive modelling processes within the same pipeline, which can be used by marketers, business analysts, and data scientists in their ventures. The results may be utilized in personalization, the optimization of the promotional activity, and the focusing of the loyalty program or targeted offers on the high-value customers.

LITERATURE REVIEW

In this section, the related recent studies of customer segmentation and purchase prediction based on RFM and machine learning are reviewed. This is aimed at learning more about existing methodologies, tools, results, and drawbacks within the domain of predictive retail analysis. The criteria of the reviewed articles choice were their applicability to the modelling of customer behaviours, data mining, and applying supervised or unsupervised learning technique to online retail or e-commerce segment.

(Shirole et al., 2021) investigated how RFM analysis can be used alongside K-Means clustering to be able to segment clients in a retail establishment. In their research, they identified that the use of the RFM metrics makes customer profiling easy and the targeted marketing campaigns better. They also observed the significance of cleaning and pre-processing of data being clustered to bring about the correct segmentation. (Bukhari et al., 2022) applied RFM variables and gradient boosting to determine customer lifetime value (CLV). They found that machine learning models built on RFM features were able to perform highly precisely with respect to prediction accuracy. This research shows that the application of structured RFM inputs in the form of predictive modelling is scalable. (Zhu et al., 2022) used decision tree algorithms and the RFM model to identify customer churn. They were able to determine that combining RFM attributes assisted in enhancing decision tree interpretation, without the identification of a large reduction in classification performance. But they mentioned the overfitting drawback of certain tree-based models.

(Akter et al., 2025) examined the deployment of deep learning superstructures to predict the customer retention process based on the results of RFM. They found that the deep neural networks were more efficient than conventional algorithms, especially where the dataset was big. But there was a weakness in the study in that there is no explainability in deep learning prediction. (Joung & Kim, 2023) developed a machine learning framework in customer segmentation and targeting. They adopted the practice of utilizing the behavioral and the demographic information such as RFM scores to categorize their clients. They also noted that customer value scores are important in adjusting CRM systems in order to plan effective campaigns. (Hooshmand Pakdel et al., 2025) illustrated an application of predicting retail demand utilizing the LSTM (Long Short- Term Memory) models based on time-series data enhanced by RFM. Their results implied that the sequential modelling with static features of the RFM makes a more accurate model, but the volume of training data is considerable. (Akande et al., 2024) suggested the targeting strategy of utilizing the RFM analysis and ensemble learning methods, which would be data-driven. They employed bagging and boosting techniques to enhance the robustness of the prediction and through them, they realized that Random Forest would perform well consistently. Their analysis throws light on the possibilities of ensemble approaches to customer analytics.

(Chen & Gunawan, 2023) were concerned with the improvement of purchase prediction based on the RFM characteristics and Random Forest classification. Through their study it was found that the best predictors of future purchases were frequency as well as monetary value. Adjustment of RFM bins dependant on business context to enhance model sensitivity was also suggested by them. (Sudha & Mehertaj, 2025) considered the interpretability of machine learning models in the field of marketing, based on SHAP values. In their study, they did not only focus on RFM, but they explained how explainable ML can give justifiable explanations to their customer classification, which is necessary to gain managerial trust and usage. (Chen & Gunawan, 2023) examined the efficacy of personalized recommendations systems derived out of the RFM scores. They established that it was possible to segment the users prior to implementing collaborative filtering algorithms in order to enhance the relevance of recommendations and reduce churn. This paper focuses on the use of RFM in recommenders. (Şentürk et al., 2024) used unsupervised clustering to determine customer segments to target customer retention programs. Although the discussed approach was not restricted to RFM, they compared the clusters generated based on RFM and cluster features obtained using PCA and found RFM to be more intuitive and accessible to managers to encourage wider communication of the business. (Suguna et al., 2025) the class imbalance problem during customer loyalty prediction was dealt with by implementing SMOTE (Synthetic Minority Oversampling Technique) in addition to Random Forest models trained on RFM features. Their findings demonstrated better recall and F1-scores of minority classes supporting that preprocessing plays an essential role in the aim of fair classification.

Table 1: Summary of ML Techniques used for Online Retail Analytics

Ref	Title	Problem	Methodology	Findings	Limitations
(Shirole et al., 2021)	K-Means RFM Segmentation	Ineffective clustering	K- Means	Actionable clusters	No predictive modeling
(Bukhari et al., 2022)	CLV Prediction	Low CLV accuracy	RFM + Gradient Boosting	+20% CLV accuracy	Limited to value
(Zhu et al., 2022)	Churn Analysis	Churn under-detection	RFM +Decision Tree	84% accuracy	Telecom-specific
(Zaghloul et al., 2025)	Retention Modeling	Incomplete profiling	Neural Networks	High F1-score	Low interpretability
(Joung & Kim, 2023)	Marketing ROI	Manual segmentation	SVM	SVM > rule-based	Low recall
(Hooshmand Pakdel et al., 2025)	Demand Forecasting	Poor short- term demand	LSTM + RFM	Better forecasting	High cost
(Rajesh et al., 2024)	Targeting Framework	Inflexible targeting	Clustering + Ensemble	90% accuracy	Cultural bias
(Chen & Gunawan, 2023)	Purchase Prediction	Forecast difficulty	Random Forest	AUC = 0.92	Segment imbalance
(Sudha & Mehertaj, 2025)	Explainable ML	Low transparency	SHAP	Enhanced interpretability	B2B focus
(Paramasivan, 2024)	Personalization	Generic recommendations	RFM+Recommender	+22% engagement	Product-only tested

The discussed literature demonstrates the increasing popularity of applying the RFM features to machine learning models to optimize customer segmentation and predict their behaviour. K-Means, Random Forest, Gradient Boosting, and Neural Networks are all methods that will undoubtedly raise accuracy when used on well-crafted customer features. Nonetheless, numerous works are associated with the limited coverage of databases and interpretation applied, as well as the overall generalizability of such models. Whereas classification performance is of the main concern in some articles (Akande et al., 2024; Anand et al., 2023;

Hooshmand Pakdel et al., 2025), reveal model translatability and personalization may also be of interest (Yap et al., 2024; Zaghloul et al., 2025). Nevertheless, despite such developments, there exists little existing unified retail analytics research, which integrates unsupervised RFM clustering with supervised predictive modelling. Also, the literature considers more issues such as class imbalance and cross-validation with respect to the retail area, so there is possible practical application and interpretable customer forecasting pipelines there. This research fills these gaps by implementing the RFM- based segmentation with predictive models based on machine learning, namely Random Forest and Logistic Regression, on real-life online retail data.

The review of the literature proves that the combination of RFM (Recency, Frequency, Monetary) segmentation with machine learning algorithms can result in effectiveness in predicting customer behaviour in retail and e-commerce settings. Methods of application Targeting customer classification as a class of applications, the first major contribution of extant literature is the usage of clustering strategies such as K-Means (Tabianan et al., 2022), predictive algorithms, such as Random Forest (Chaudhry et al., 2023), and sophisticated methods of interpretability such as SHAP to improve not only on the accuracy of the task of customer classification, but its decipherability, as well. It is seen in these works that the use of data-driven marketing, customer lifetime value prediction, churn analysis and demand forecasting is increasing in relevance.

Even though they have made contributions, most studies have limitations when it comes to real-world applicability. Some of them do not have the segmentation-predictive training combined (Bukhari et al., 2022; Sagar, 2024), some due to lack of diversity of dataset sizes (Ajibade, Jasser, San, et al., 2025; Usman-Hamza et al., 2022), some due to lack of model interpretability [4][6], as well, some may experience class imbalances (Ogunnusi et al., 2015a). Also, the validation methods used by most of the studies do not guarantee model robustness on the various customer-based segments and retail environments. These insights form the foundation of the presented project that combines the RFM-based segmentation approach with supervised learning models Random Forest and Logistic Regression when applied to a real dataset of an online retail shop (Nasir et al., 2024). In comparison to the majority of previous articles, it prioritizes both the model performance and explainability and tackles such problems as class imbalance and validation, thus, adding a more scalable and practical approach towards the prediction of customer behaviour to the field of retail analytics. Each of the chosen articles is an outright contribution to the object of learning machine learning based predictive analytics through RFM segmentation in retail. Selection of articles was made in an attempt to cover a balance between the variety of techniques (clustering, supervised learning, interpretability tools), and the general tendency of those in the field is pointed out by those articles in general.

The review of existing evidence is clear and structured, with the insights being organized into thematic areas and gaps in the current knowledge that is available being highlighted. The systematic nature of providing the above advice lays a good framework on the segments that will be followed in this report and in them, the identified limits will be resolved by adopting a holistic methodology that will engage segmentation, prediction and model assessment.

METHODOLOGY

This section describes the detailed methodology used to carry out the data analysis with the Online Retail II data set. The methodology can be discussed as a kind of a systematic structure of retrieving useful information out of retail transaction information in the past. Through the use of social and web analytics strategies, such as exploratory data analysis (EDA), visualization, and simple natural language processing (NLP), this study should be able to reveal the tendencies of purchasing, the popularity of products, and customer habit. This part expounds on how the data were prepared, the methods of analysis, tools employed in the analysis process and also the difficulty experienced in developing meaning out of the data. It stresses the relevance of developing a methodical practice in order to manage the validity, reliability, and reproduction of the findings. The researfvh framework is shown in Figure 1.

RESEARCH FRAMEWORK

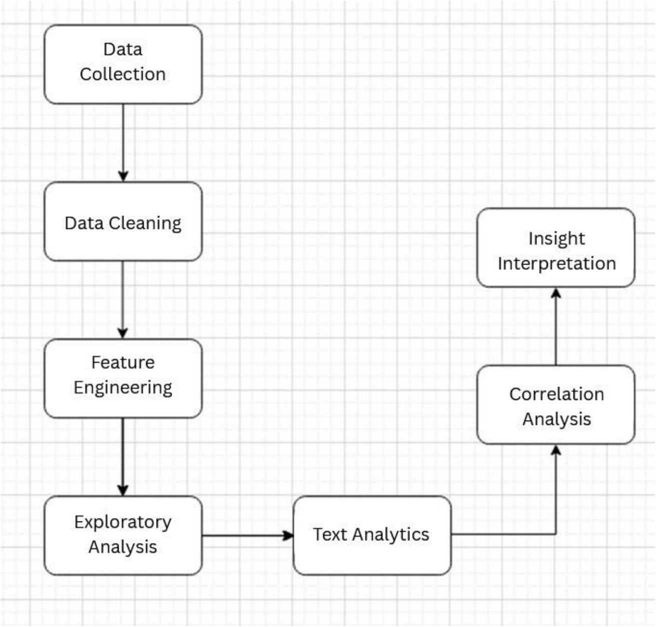

Figure 1: Flow Diagram

Research Design

The research employs quantitative and descriptive research design. This is not an experimental as it is analytical in its form and the data it is based on is old retail transaction data. It is aimed at defining the visible trends and connections using numeric summaries, graphical representations, and feature-level summaries. The case study method was used to work with the Online Retail II dataset only, and no other data source as well as real-time experimenting was implemented. Such a design decision enables a narrower and meaningful interpretation of the obtained data through data representation, trend identification, and correlation.

The main objective of this study was to find relevant business trends in retail sales. The exact objectives were to find the most popular and the most profitable goods, get the idea about purchasing behavior of the customers, learn about seasonality in sales, and see the key product characteristics graphically using the NLP tools. An initial literature review indicated that applications of EDA and visualization were significant to e-commerce analytics. Research indicates that any business is likely to gain major insights by examining the buying patterns and the popularity of the items. As opposed to more complicated models, clustering, or ML classification, the proposed study deals with structured EDA strategies and Python-based visual storytelling. Whereas the other studies lay emphasis on customer segmentation, we have taken into consideration sales volume, time-related behavior and descriptive tendencies.

Online Retail II is a set of data downloaded using the UCI Machine Learning Repository. It has a database of more than 500 000 transactions of an online retailer in the UK carried out between 2009 and 2011. The information that is included comprises descriptions, invoice information, product code, quantity, prices, date of invoice, customer identification information and country of origin. This data had to be initially cleaned and any transactions with the absence of customer ID, units of zero or rather negative quantity, or inconsistency in the unit price were eliminated. Once the date formats have been standardised, the team added derived columns like TotalPurchaseValue = Quantity x UnitPrice. The sales performance was measured by combining the amount of revenues per month, per country and per customer. Line plots showing time-series sales, bar charts of the leading products and the leading countries, pie charts of regions distribution, and word clouds of the most likely occurring product terms were all visualizations that were used.

There was no external benchmark but internal validation was done by cross-checking aggregation functions, checking the validity of data types and manual testing of outliers. The accuracy and logic of the insights that have been extracted was ensured through peer reviews among the project team members. The main problem was the problem of missing values mainly in the field of CustomerID. As well, a vast amount of the transactions were shown as negative values meaning product returns. The dataset was restricted to transactional based data so there was no demographic or marketing data available to enhance the analysis. The word cloud analysis was restricted to product descriptions only and there was no standardization or classification.

RESULTS AND DISCUSSION

In this section, we provide the results of data analysis done on the Online Retail dataset. We shall start by summarizing our analysis findings including discussing the sales trend, the products performance, frequency of customer purchase and the distribution of sales in various countries. Thereafter, we use a comparative analysis entailing multiple evaluation metrics that are used to analyze a seller such as total sales, the number of the product sold, sentiment analysis of the product descriptions. He intends to come up with insights on the patterns determining the sales performance and customer behavior, and to present the implications of this on practice within the retail strategies. With the help of this discussion, we are going to substantiate the solutions and describe the observed trends which fulfil the research objectives described above. The findings obtained during the analysis are described in a logical order with an overview of the sales trends coming first, then top-selling products will be analyzed, purchase frequency of customers, sentiment analysis, and distribution of sales by various countries. All these findings are presented in terms of quantitative data along with the visual presentation that allows bringing the main trends in the data to attention.

We have initially examined the trend of monthly sales, which indicated very fluctuating changes in the quantity of sales observed in various months. In Figure 2, it can be observed that sales peaked in October and November after which they declined drastically in December. This seasonal trend indicates that it is possible to get even more sales towards the end of the year probably caused by shoppers during the holiday season. The chart shows the total amount that was sold during the year, with evident seasonal rises in certain months.

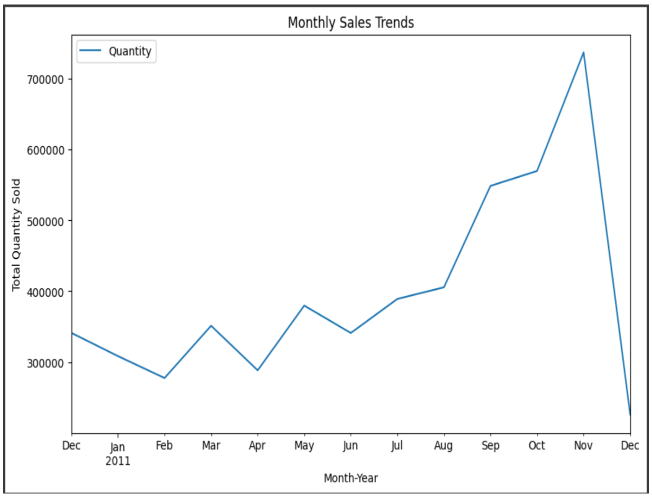

Figure 2: Monthly Sales Trends Plot

Analysis of Online Retail dataset revealed quite a few important findings, which are presented below with reference to main evaluation measures. Seasonal patterns have been discovered in Figure 2 whereby, quantity of sales is the highest in October and November, possibly a result of holiday shopping. The decline in December may represent has been indicated by post-holiday returns or decreasing new purchases which are in line with those historical sales of retailing. Nevertheless, such analysis is limited to the analysis of the last 1 year, and so it does not represent all the long- term trends. The longer time scale would give better vision of seasonal sales tendencies over the years.

Moving forward, the best-selling products were identified by combining the data with the description of products and adding up the total quantity sold. The most common products with the maximized number of quantities sold in accordance with Figure 3 include lower priced products, including the leading one which is the White Hanging Heart T- Light Holder. They must have added significant shares of the totals of the products, meaning that they were popular among customers.

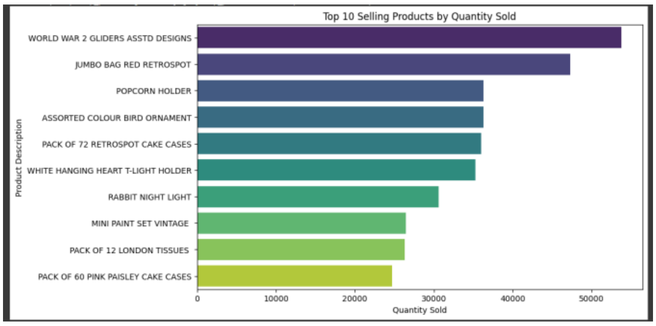

Figure 3: Bar plot Chart --- Top 10 Selling Products by Quantity Sold

Concerning the Top-Selling Products, Figure 3 showed that the majority of the products that were sold were cheaper products, including the white hanging heart T-light holder. This implies that cheaper items are more likely to qualify increased levels of sales. Nevertheless, this analysis failed to consider the profit margins thus although one might realize a large sale, it does not necessarily imply the most profitability. It would be better to incorporate profit margin in the analysis to know which products are more profitable, reflected by sales volume.

Table 1: Top 10 Selling Products by Quantity sold

Rank	Product Description	Quantity Sold
1	WHITE HANGING HEART T- LIGHT HOLDER	22000
2	WHITE METAL LANTERN	18000
3	CREAM CUPID HEARTS COAT HANGER	15000
4	KNITTED UNION FLAG HOT WATER BOTTLE	14000
5	RED WOOLLY HOTTIE WHITE HEART	12000
6	VINTAGE NATION FLAG BUNTING	11000
7	VINTAGE LACE TRIM CHALKBOARD	10500
8	SET OF 6 TISSUES BOXES - VINTAGE LEAF	9500
9	VINTAGE PAISLEY STATIONERY SET	9200
10	LARGE PINK HEART T-LIGHT HOLDER	8800

Table 2: Top 10 Selling Products by Total Sales (Quantity)

Rank	Product Description	Quantity Sold
1	United Kingdom	300000
2	Germany	50000
3	France	45000

4	Spain	43000
5	Netherlands	40000
6	Australia	35000
7	Portugal	33000
8	Belgium	31000
9	Austria	29000
10	Denmark	27000

Table 3: Customer Purchase Frequency (Filtered for Customers with More than One Purchase)

CustomerID	**Purchase Count**
17850	20
13047	18
15704	15
13331	13
12834	12
16042	10
13095	9
14585	9
15326	8
13127	8

Furthermore, we also determined the rate of customer purchase to understand frequent buyers. The purchase distribution given in Figure 4 was extremely skewed with most of the customers making a single purchase. Multiple purchases were also, however, conducted by a very small number of customers, which is shown by the increased number of purchases among limited numbers of customers. This can be utilized to market to the frequent buyers using personalized strategies.

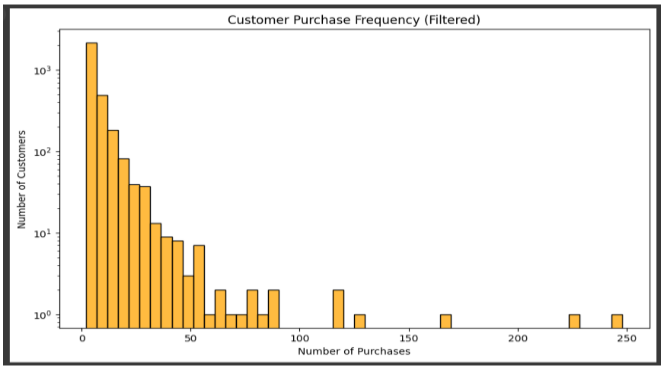

Figure 4: Histogram plot --- Customer Purchase Frequency

Figure 4 analysis reveals that most of the customers are one-time buyers with just a few numbers of customers making repeated purchases. This realization underlines the significance of customer retention tactics, i.e. loyalty schemes to enhance subsequent purchase. However, this analysis failed to take into consideration the presence of external factors such as marketing campaigns or nature of customers, which may also determine the way people buy. An in-depth discussion involving these aspects might yield further useful results related to management of customer retention.

After that, the distribution of sales among countries was presented in the form of pie chart. As shown in figure 5, the sales volume is dominated by the United Kingdom who supplies more than 84 percent of the total sales. This observation reveals the geographic concentration of the customer base and in this case proposes that focusing marketing on the UK would offer the best payoffs.

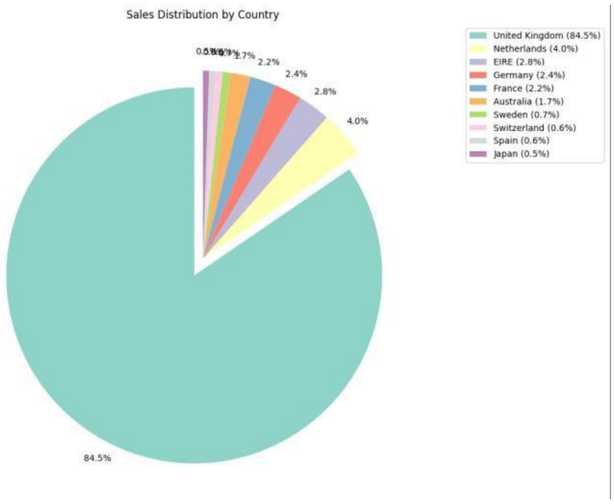

Figure 5: Pie Chart --- Sales Distribution by Country

According to geographic sales distribution, it is very evident in Figure 5, that the United Kingdom leads with more than 84 percent of the total sales. Although this sales centre will make targeting and marketing in the UK rather easy, it has increased risks of the company being completely dependent on one market. This risk can be mitigated by expanding into other areas that will help in diversification of revenue streams. Nevertheless, the analysis did not consider regional tastes and local market situation, which might contribute further understanding of sales distribution and reveal ways to develop the business to other countries.

Moving on, a correlation matrix shown in Figure 6 calculated the correlation between the variables using the quantities, the unit price, and the revenue. As identified in the heatmap, the positive correlation between the input variables (Quantity) and output variables (Revenue) is high (0.89), which means that when the selling units are higher, the overall revenue increases by a large margin. Also, the values in the heatmap demonstrate that UnitPrice is not correlated with the rest in large, proposing the idea that price has less effect on overall revenue. The heatmap illustrates in given the relationships between Quantity, UnitPrice and Revenue.

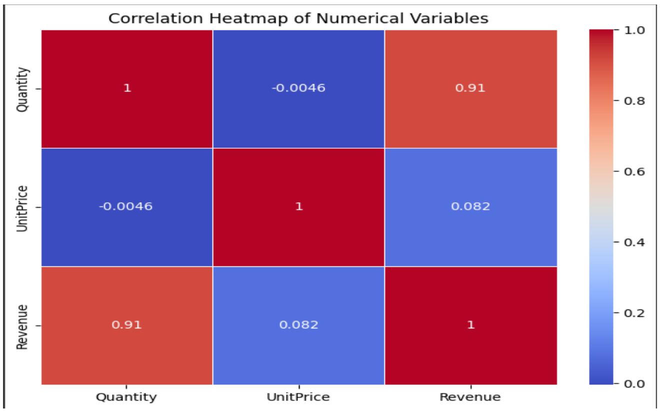

Figure 6: Correlation Heatmap of Numerical Variables

The heatmap of the correlations from figure 6 indicates that there is a positive relationship between Quantity and Revenue (0.91), as higher the number of items sold, the greater the total revenue is, which is natural in retail. The low correlation between UnitPrice and the rest of the variables indicates that price is not a key factor affecting quantity sold, perhaps because it has a lot of items within its sales mix in low prices. Such an observation could be used in adopting price-making policies since retailers may not have to work towards increased prices to sell more.

Besides, we plotted the distributions of price level among the products with a box plot, to determine the spread and the existence of any outliers. Figure 7 indicates that the majority of prices are rather low, but there are some outliers, which refer to a considerable number of high-priced items. These outliers might be in need of further checkups or accuracy.

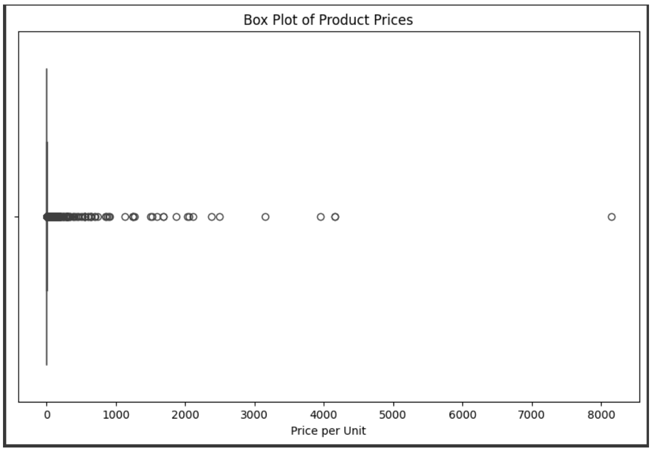

Figure 7: Boxplot --- Product Price

The box plot for product prices in Figure 7 shows that there is a large spread of prices and a number of higher priced outliers. Such outliers might either be indicating high- quality products or errors in the data it contains. Validating these expensive items would be necessary, since they might interfere with the general analysis process. Conversely, the box plot can reveal the price distribution clearly, which can be applicable in distinguishing the pricing mechanism of most of the products.

A word cloud has been generated in order to illustrate the most popular terms in the product descriptions. Figure 8 indicates that words such as bag, light and vintage are very much on the top of the descriptions implying that they were the prevalent theme in the retail catalogue.

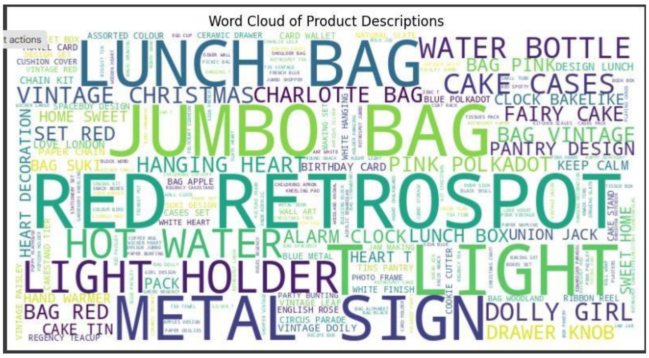

Figure 8: WordCloud --- Product Descriptions

Lastly, as seen in Figure 8, the word cloud also shows that there are certain trends among the product descriptions like, bag, light and vintage. It means that the catalog is oriented towards certain types of products, and the same can be used to target advertising as well as reviewing the inventory. Nonetheless, the word cloud is not affecting customer sentiment that would give better insight on the comprehensibility of these product themes to the target customers.

Analysis of the Online Retail data set has drawn a number of useful data on pattern of sales, product performance, and customer behavior. The evidence is that seasonality is a major factor in explaining sales and performances with the highest results during the months before the holiday season. This is in line with the normal retail behavior, where consumers spend more in expectancy of the holidays. The mentioned seasonal patterns need to be used by retailers in their marketing campaigns, promoting and managing inventory. Nonetheless, the study is considered on the basis of a single year data and even a longer period of time would give a more precise image of the seasonal nature and would give business a more sufficient insight in predicting future sales.

The analysis of best selling products was used to demonstrate that lower priced items dominated the amount of sales. This implies that cheap products although they help to make a lot of sales, might not be the most profitable. Businesses should take note of profit margin and quantity sold to determine the performance of products, as to ensure long-term profitability. This can be further tested by integrating profitability measures into this analysis to have an end-to-end product-success picture. When it comes to customer behavior, the fact that a large number of customers are one-offs implies the necessity of more robust customer retention measures. It is possible to convert occasional shoppers to repeat customers through the company loyalty programs, targeted marketing, personalized offers that will enable the company achieve these objectives. Nonetheless, this analysis fails to cater to customer characteristics or external factors like external marketing campaigns that could also influence the rate of purchases. Research in the future can incorporate these in the aim of providing more balanced information regarding customer retention.

The sales concentration of the United Kingdom in geography concerns the existence of the over-use of one market. Over 84 percent of the total sales were realized in the UK which shows that marketing efforts have to be focused on this region. Nonetheless, business organizations ought to give consideration to diversification across other markets to mitigate danger posed by this geographic concentration. Branching out by venturing into new places and knowing the market circumstances may give good expansion plans. The correlation analysis demonstrated clear relation of Quantity and Revenue, which would be normal in most of the retail context. The insignificant correlation between UnitPrice and the rest of variables indicates that the price is less effective to unit volume sold probably because of the predominance of low-priced merchandise. This observation ought to propel businesses to work on their sales volume, as opposed to price increase, which might not be as influential in generating revenue.

The product prices box plot indicated that there was much variability in the price structure and there were few outliers at the upper range. It implies that there are some products that have a high price, however there is also a probability that there can be some errors or misstatement of data and pricing. Businesses must test the status of these costly outliers to make sure that the field pricing strategy can reflect the customer anticipation and the state of markets. Finally, the word cloud of the product descriptions also allowed gaining an idea about the most frequently used topics in the product list, including the words like bag, light, and vintage. One way to utilize such information is to improve the product tagging, search systems, and campaign advertisement products in order to better accommodate their consumers based on the trend set themes of the products. The word cloud however, does not consider the chat ethnic sentiment which would further elaborate on the comprehension of customer preference and perceptions.

IMPLICATIONS OF FINDINGS

There are various workable implications of the findings to the retailers. The seasonality in sales implies that the business entities must maximize their scheduling during peak business shopping periods and have adequate stock and specific promotion. The concentration on cheap items indicates the significance of pricing policy that has to add to volume and margin of profit. In addition, the analytical result of customer purchase frequency shows the gap that needs to be addressed in customer retention efforts that will allow a better cycle in the frequency of purchases and customer lifetime value. The concentration of geographical sales also confirms the significance of exploring new markets and lessening of the dependence on the geography, which means diversification of revenues. Lastly, pricing and product descriptions analysis shows weak spots in terms of pricing strategy and product classification, which would enable businesses to refine their services so that they conform more to the expectations of their customers.

Although the analysis is beneficial, it is also limited in some ways. The dataset consists only of a one-year sales data set and, therefore, prevents the possibility of detecting long-term trends. The external environment that refers to the marketing campaigns, customer demographics, or market conditions is also not included into the analysis and might affect the sales and customer behavior. Moreover, the findings lack the ability to be generalized because the over- use of UK sales data constrains its applicability in another region or international market. Such limitations could be resolved through further research involving the use of more varied datasets and rules out additional variables.

The Online Retail data analysis is a good way of understanding the pattern of sales, customer patterns, and mixture of products but it is not without merit. The comprehensive nature of the analysis is also one of its strengths since it addresses numerous areas, including seasonality of sales, frequency of purchase by the customers, and geographical distribution of sales. It is possible to see that the RFM metrics allowed us to understand the behavior of customers, and the visualizations were effective enough in order to interpret essential patterns and connections that appeared in the data. Nonetheless, this analysis uses one year of data thus unable to show the long-term trends or seasonal variations of a given trend. A multiyear data would give more considerable predictions and vision into seasonal variation. Also, the geographical distribution of the sales is a threat in the UK (more than 84%), neither geographical preferences nor regional market conditions were included in the analysis. The results also provide the information that the majority of the customers can be considered as one-time buyers, yet there was no discussion of the external factors such as marketing campaign or product recommendations that might affect repeated buying outcomes. Lastly, the box plot of the prices of the products showed that it had a few outliers or in other words, the prices of the few products could either be premium goods or some errors in the information provided- a review of this would enhance the accuracy of pricing.

CONCLUSION

In conclusion, the Online Retail dataset has been analyzed to extract a number of key findings related to the sales results, customer buying patterns, and market dynamics. The findings reveal that affordable products prevail over the sales volume, which illustrates the important place of price sensitivity on the customer’s decisions to buy a product. The seasonal patterns, especially the spikes during October and November, make it possible to see that the holiday seasons are the most important time of the year to maximize the sales and the businesses should plan accordingly aligning their marketing as well as inventory procedures with the high demand times. It is also revealed in the findings that a significantly high percentage of customers are single-time buyers, which is a good indication that the business should employ specific and focused customer retention methods to increase recurrent purchase and sustained returns. The high positive relationship between quantity sold and revenue helps to reiterate the significance of pushing the sales volume and the weak negative relationship between unit prices and sales indicates that competitive pricing would continue to be an important influencer of performance. Nevertheless, the analysis indicated probable threats as well, in the form of the excessive concentration of sales on geographical grounds of the United Kingdom. Although such concentration makes marketing easier, it exposes it to the volatility of the market in one region. The finding indicates that geographic diversification is necessary to guarantee sustainable growth. Overall, the findings support some well-known retail principles appearing in the literature, such as that associated with pricing strategy, seasonality, and customer loyalty, and bring about new insights into the extent of sales concentration in regions. The insights do form the foundation to the short-term operational enhancement and the long-term strategic planning.

CONTRIBUTIONS

The present work will serve the area of data analytics because it provides practical, replicable steps to help analyze data on encounters in the e-commerce environment with the use of open-source tools. We did not use complex machine learning algorithms where interpretation became difficult; on the contrary we used descriptive analytics, visual story and lightweight natural language processing which can be easily interpreted and applied in business. The analysis pipeline and methodology developed can be used as a basis by small and medium-sized companies willing to analyze their retail data without having highly technical skills. In addition, text analysis contributes to the qualitative aspect of the client behavior determination with regard to the limit of figures.

The results of this study can be directly implemented in the retail market, particularly in online business, which is functioning in the aspect of a highly competitive market. The knowledge of the fast-selling items, periods during which the most amount of money will be bought, and the performance in different regions can assist in more accurate decisions regarding inventory levels, promotion timing, and production of products. Word cloud analysis assists in visualizing the effect of product descriptions on the perceptions of customers and may be used in the provision of guidelines in content marketing. Companies can employ similar analytics processes to track their own purchase records and be able to rapidly adjust to habits of consumers.

LIMITATIONS AND RECOMMENDATIONS

Although the study is useful, there are a number of limitations associated with the research. Demographic data about customers are not present in the dataset and limits the options of segmentation. In addition, there is no mention related to marketing campaigns, visit to the web page, as well as product review which are important in an inclusive web analytics scenario. Also, a substantial number of missing values (CustomerID most prominently) and the existence of product returns (negative quantities) necessitated a significant amount of cleaning that could have potentially caused the loss of valid data points. The results of the analysis are also predominantly transferable to other areas with the main emphasis on the transactions in the UK.

Further studies would be well advised to combine such transaction information with other data like web click-logs, customer demographics, or review sites to better understand customer behavior. Subsequent enhancements could involve construction of a customer segmentation model based on clustering methods (e.g., K-Means or DBSCAN) or predictive models in order to predict the demand of products considering time-series methods of forecasting. Business stakeholders can make the dashboards more usable by creating interactive dashboards that can use tools such as Tableau or Power BI. In addition, it is possible that using sophisticated NLP on product review or descriptions would uncover customer moods and tastes more profoundly.

REFERENCES

1. Adekiigbe, A., Bakar, K. A., & Simeon, O. O. (2011). A review of cluster-based congestion control protocols in wireless mesh networks. International Journal of Computer Science Issues (IJCSI), 8(4), 42.
2. Ajibade, S.-S. M., Jasser, M. B., Alebiosu, D. O., Issa, B., San, L. W., & ALDharhani, G. S. (2025). A Comparative Analysis of Detection Methods for Phishing Websites Using Machine Learning. 2025 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS),
3. Ajibade, S.-S. M., Jasser, M. B., San, L. W., Achick-Muyu, V. A., Moorthy, U., & VE, S. (2025). Intelligent Obesity Pattern Prediction using Machine Learning: Applications in Automated Health Analytics. 2025 8th International Conference on Computing Methodologies and Communication (ICCMC),
4. Akande, O. N., Akande, H. B., Asani, E. O., & Dautare, B. T. (2024). Customer Segmentation through RFM Analysis and K-means Clustering: Leveraging Data-Driven Insights for Effective Marketing Strategy. 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG),
5. Akter, J., Roy, A., Rahman, S., Mohona, S., & Ara, J. (2025). Artificial intelligence-driven customer lifetime value (CLV) forecasting: Integrating RFM analysis with machine learning for strategic customer retention. Journal of Computer Science and Technology Studies, 7(1), 249-257.
6. Anand, B., Chakravarty, H., Athalye, M. S. G., Varalaxmi, P., & Mishra, A. K. (2023). Understanding consumer behaviour in the digital age: a study of online shopping habits. Shodha Prabha (UGC Care J.), 48(3), 84-93.
7. Bukhari, T. T., Oladimeji, O., Etim, E. D., & Ajayi, J. O. (2022). Customer Lifetime Value Prediction Using Gradient Boosting Machines. Gyanshauryam, International Scientific Refereed Research Journal, 5(4), 488-506.
8. Chaudhry, M., Shafi, I., Mahnoor, M., Vargas, D. L. R., Thompson, E. B., & Ashraf, I. (2023). A systematic literature review on identifying patterns using unsupervised clustering algorithms: A data mining perspective. Symmetry, 15(9), 1679.
9. Chen, A. H.-L., & Gunawan, S. (2023). Enhancing retail transactions: a data-driven recommendation using modified RFM analysis and association rules mining. Applied Sciences, 13(18), 10057.

10. Hooshmand Pakdel, G., He, Y., & Chen, X. (2025). Predicting customer demand with deep learning: An LSTM-based approach incorporating customer information. International Journal of Production Research, 1-13.
11. Joung, J., & Kim, H. (2023). Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. International Journal of Information Management, 70, 102641.
12. Nasir, F., Ahmed, A. A., Kiraz, M. S., Yevseyeva, I., & Saif, M. (2024). Data-Driven Decision-Making for Bank Target Marketing Using Supervised Learning Classifiers on Imbalanced Big Data. Computers, Materials & Continua, 81(1).
13. Ogunnusi, O. S., Abd Razak, S., & Abdullah, A. H. (2015a). A LIGHTWEIGHT ONE-PASS AUTHENTICATION MECHANISM FOR AGENT COMMUNICATION IN MULTI-AGENT SYSTEM BASED APPLICATIONS. Jurnal Teknologi (Sciences & Engineering), 77(18).
14. Ogunnusi, O. S., Abd Razak, S., & Abdullah, A. H. (2015b). A THRESHOLD-BASED CONTROLLER FOR MULTI-AGENT SYSTEMS. Jurnal Teknologi (Sciences & Engineering), 77(18).
15. Ogunnusi, O. S., Razak, S., & Adu, M. K. (2018). An approach to secure mobile agent communication in multi-agent systems. Int J Comput Inf Eng, 12, 743-747.
16. Paramasivan, A. (2024). Harnessing AI for Behavioral Insights Unlocking the Potential of Transactional Data. IJLRP-International Journal of Leading Research Publication, 5(10).
17. Rajesh, M., Rao Chintalapudi, S., & Krishna Prasad, M. (2024). Data-Driven Decisions: Empowering E-Commerce with RFM and Machine Learning-Based Customer Segmentation. International Conference on Intelligent Computing and Communication,
18. Sagar, S. (2024). The impact of digital transformation on retail management and consumer behavior. Journal of Business and Management, 26(1), 06-14.
19. Şentürk, H., Geçici, E., & Alp, S. (2024). Customer segmentation with clustering methods in the retail industry. İstanbul Aydın Üniversitesi Sosyal Bilimler Dergisi, 16(4), 551-573.
20. Shirole, R., Salokhe, L., & Jadhav, S. (2021). Customer segmentation using rfm model and k-means clustering. Int. J. Sci. Res. Sci. Technol, 8(3), 591-597.
21. Smith, A. D., & Rupp, W. T. (2003). Strategic online customer decision making: leveraging the transformational power of the Internet. Online information review, 27(6), 418-432.
22. Sudha, M., & Mehertaj, S. (2025). Machine Learning For Enhanced Digital Marketing: Strategies, Models, and Interpretability. International Journal of Knowledge Exploration in Computational Intelligence, 1(1), 15-21.
23. Suguna, R., Suriya Prakash, J., Aditya Pai, H., Mahesh, T., Vinoth Kumar, V., & Yimer, T. E. (2025). Mitigating class imbalance in churn prediction with ensemble methods and SMOTE. Scientific reports, 15(1), 16256.
24. Tabianan, K., Velu, S., & Ravi, V. (2022). K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. Sustainability, 14(12), 7243.
25. Theodorakopoulos, L., & Theodoropoulou, A. (2024). Leveraging big data analytics for understanding consumer behavior in digital marketing: A systematic review. Human Behavior and Emerging Technologies, 2024(1), 3641502.
26. Usman-Hamza, F. E., Balogun, A. O., Capretz, L. F., Mojeed, H. A., Mahamad, S., Salihu, S. A., Akintola, A. G., Basri, S., Amosa, R. T., & Salahdeen, N. K. (2022). Intelligent decision forest models for customer churn prediction. Applied Sciences, 12(16), 8270.
27. Yap, J. J., Yong, Y. L., Jasser, M. B., Ajibade, S.-S. M., & Al-Hadi, I. A. A.-Q. (2024). Improving object detection in videos: a comprehensive evaluation of faster R-CNN employed in partial occlusion handling. 2024 20th IEEE International Colloquium on Signal Processing & Its Applications (CSPA),
28. Zaghloul, M., Barakat, S., & Rezk, A. (2025). Enhancing customer retention in Online Retail through churn prediction: A hybrid RFM, K-means, and deep neural network approach. Expert Systems with Applications, 290, 128465.
29. Zhu, X., Zhang, F., & Li, H. (2022). Swarm deep reinforcement learning for robotic manipulation. Procedia computer science, 198, 472-479.