QBUS6600 Project 1 : L'Oréal Dermatological Beauty - Predicting Customer Transaction Value

530050470

Table of Contents

1. Introduction

The main objective of this report is to explore and analyze the dataset provided by L'Oréal Dermatological Beauty (LDB), uncovering key behaviors and variables that influence customers' transaction value over the next six months. Additionally, it aims to build a machine learning model to predict future transaction value, providing business personnel with insights to prioritize customers and optimize customer management. This will enable them to implement targeted marketing strategies for different customer segments, thereby enhancing customer retention and increasing sales.

1.1 Business Context and Problem Statement

LDB is a division of L'Oréal, specializing in providing dermatological solutions for various skin concerns, such as acne, eczema, and so on. Currently, LDB's products are primarily sold through pharmacies. The division aims to expand its online presence and sales by expanding its brand-dedicated online stores. Online platforms enable constant communication with customers, timely notifications of new product launches and promotions, and the ability to match products with individual skin concerns and preferences, providing a personalized experience that enhances customer loyalty. Therefore, to enhance personalized services, LDB is committed to analyzing customers' historical purchase data, identifying their behaviors and preferences, and more accurately predicting transaction values for the next six months starting from November 20, 2023, to achieve targeted marketing.

2. Data Processing

2.1 Data Description

The subjects of the study are customers who made transactions within the six months following November 20, 2023. The observations in the dataset consist of all their historical purchase records prior to November 20, 2023. The dataset used in this report was provided by LDB and contains a total of 6.4k rows and 62 columns. All of columns are numerical variables, broadly categorized into the following types:

- Customer personal information variables: such as CustomerID (see Appendix A for details).
- Historical spending amount and frequency variables: spending in the past one month, past three months, past six months, past nine months, past twelve months, and total cumulative spending (see Appendix A for details).
- Skin concern variables: such as 'Skin Concern_Acne-Prone Skin'. (see Appendix A for details).
- Product function description variables: such as 'Class Description_Anti-Acne' (see Appendix A for details).
- Product category and subcategory variables: such as 'Category_Body Care' (see Appendix A for details).
- Specific product variables: such as 'EAN_HyaluB5Serum30ml' (see Appendix A for details).

2.2 Data Processing

The first step in the data analysis involves identifying any irregular patterns in the dataset that could cause misinterpretations. Afterward, these irregular variables should be adjusted accordingly to resolve data issues and enable more accurate analysis.

2.2.1 Duplicate Variables & Missing Values & Irrelevant Variables

To ensure the feasibility of further data analysis, a quick check for potential data quality issues was conducted. It was found that there are duplicate columns named '*Category_Face Care*' in the dataset. Based on the official response, we will combine the data from these two columns by summing them into the correct column '*Category_Face Care*'.

When examining the dataset for missing values, only one variable ('*Post Code*') contains missing. It has a significant amount of missing data and thus it will be removed in the analysis. Additionally, according to Data Dictionary, the '*CustomerID*' and '*Has_Transaction_Nov23_May24*' should be removed as well.

2.2.2 Outliers

We primarily examined whether there are outliers in certain key variables. The quick screening criterion for key variables are based on the subsequent 'Correlation Analysis' section. Since the objective of the analysis is to predict customers' future transaction value, variables related to transaction amount and transaction frequency are important influencing factors. Therefore, we mainly focused on identifying outliers in variables related to transaction amount and frequency. First, based on the quartiles of the given columns, we calculated the upper and lower bounds for each key variable using 1.5 times the IQR to filter out observations that exceed these bounds for further analysis. Using histograms and boxplots to analyze observations exceeding the bounds (as shown in Appendix B and Appendix C).

Overall, the identified "anomalies" can be categorized into two main situations. The first situation involves customers who had low spending amounts or frequencies in the past but have higher-than-expected spending in the following six months. The second situation involves customers who had very high spending amounts or frequencies in the past but have lower-than-expected spending in the following six months. Based on business understanding, these observations should not be considered outliers but rather significant changes and trends in customer behavior, which hold important business value. Thus, we keep these observations.

3. Train-Test Split

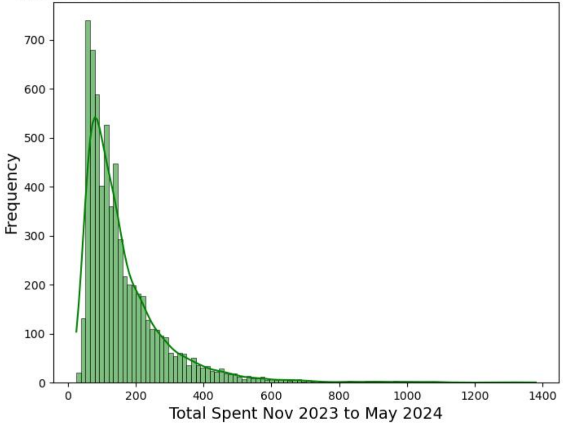
Before conducting EDA, we will perform a train-test split to ensure that data leakage is avoided during data analysis and modeling, and to maintain the fairness of model evaluation. Therefore, EDA analyses will be based on the training set.

4. EDA Analysis

4.1 Target Variable

'*Total_Spent_Nov23_May24*' is the target variable which means the total value of all transactions in the 6-month window starting on November 20, 2023. **Figure 1** shows that the dependent variable '*Total_Spent_Nov23_May24*' also exhibits positive skewness (kurtosis value of 2.71) and a large standard deviation, reflecting differences in customer spending behavior during this period, with some customers spending far more than the average. Therefore, it is necessary to apply quantile transformation to the total spending variable to better address skewness in the data distribution and make it more symmetrical.

Figure 1. Distribution of Total Spent from Nov 2023 to May 2024



4.2 Correlation

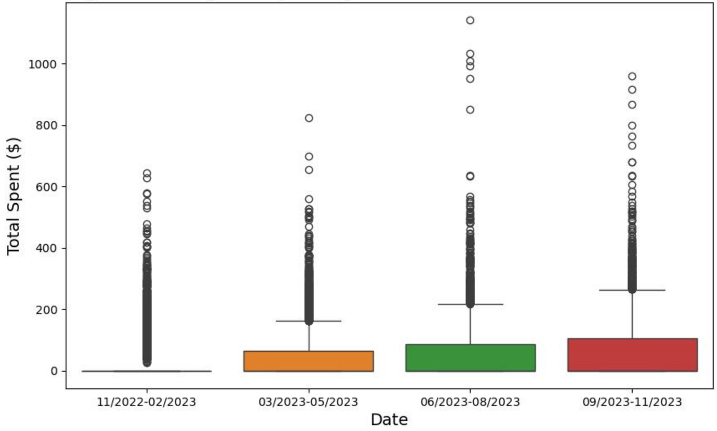
According to Appendix D, historical spending shows a moderate positive correlation with total spending over the next six months, with correlation coefficients ranging from 0.35 to 0.39. Additionally, the number of transactions over different time periods also has a positive correlation with future spending, with correlation coefficients between 0.14 and 0.26. Transaction counts over longer time windows (e.g., 12 months) are more strongly correlated with future spending than those over shorter windows (e.g., 1 month). Spending on facial care products, especially facial serums, significantly contributes to total spending over the next six months, showing moderate correlation. Furthermore, anti-aging-related products (such as ‘Hyalu B5’ and ‘Retinol LRP’) also exhibit a positive correlation with future spending, with correlation coefficients between 0.17 and 0.27. In contrast, sunscreen products (such as ‘Anthelios’) have a minimal impact on total spending in the next six months, with low correlation. Additionally, there is multicollinearity in the data, particularly between spending amounts, purchase frequency, and their interrelations. Therefore, in subsequent modeling, it is recommended to remove highly correlated variables or use methods like ridge regression to mitigate the impact of multicollinearity.

4.3 Time Pattern Analysis

4.3.1 Total Consumption in Different Seasons

By calculating the spending differences, we aim to analyze time patterns and identify whether there are specific periods when customer spending intent is higher. Since the data from one year ago are too dated to be useful for reference, we only consider and analyze the consumption in the past 12 months. Figure 2 below illustrates that Overall, there is a noticeable seasonal purchasing pattern: customers tend to have significantly higher purchase frequency and spending amounts in winter and spring compared to summer and autumn, which is also further supported by the descriptive analysis (Appendix E). This may require more aggressive marketing strategies during these periods to maximize revenue.

Figure 2. Changes in Spending Behavior Over Different Time Periods



4.3.2 Consumption of Different Value Users in Different Seasons

Therefore, we want to further explore the spending patterns of different customers across different seasons. The division between high-value and regular customers in this section is based on the average spending per transaction in each season. Customers whose average spending per transaction exceeds that of 80% of the population are classified as high-value customers for that season, while those whose spending is below the 80% threshold are classified as regular customers. As a result, we obtain high-value and regular customers for each season, and analyze the purchasing habits of these groups from the perspective of different product categories.

According to Appendix F, overall, regardless of the season, all customers show high demand for the Anthelios, Effaclar, and Toleriane series, and the purchase of facial moisturizers and serums is significantly higher than other product categories. High-value customers have a greater demand for brightening and anti-aging facial serums in the autumn and winter seasons, specifically products like Hyalu B5 Serum 30ml, Retinol B3 Serum 30ml, and Vitamin C10 Serum 30ml. In contrast, during summer, their demand shifts toward tinted sunscreen products, such as Anthelios Invisible Sunscreen 50ml.

4.4 Consumer Segmentation Analysis by using RFM model

The RFM model is an effective method for identifying customer groups that require special attention (Safari, Safari and Montazer, 2016). By using the RFM model, LDB's operations team can identify specific customer segments and deliver relevant messages and marketing activities that generate higher response rates, improve loyalty, and increase customer lifetime value.

4.4.1 Scoring Metrics and Define Customer Classification

The RFM scoring system divides Recency (R), Frequency (F), and Monetary (M) metrics into five ranking groups, where higher scores indicate customer behaviors preferred by LDB (Anitha and Patil, 2019). The specific scoring criteria are detailed in Appendix G. Recency represents the number of months since a customer's last order, with higher values indicating a greater likelihood of churn. Frequency measures the number of purchases over a 9-month period, selected for its ability to capture seasonal and promotional cycles. Monetary reflects the average amount spent per purchase, and average spending is used instead of total spending to avoid high correlation between the Frequency and Monetary metrics, ensuring more distinct customer segmentation.

For customer classification, the R_Score, F_Score, and M_Score each range from 1 to 5. To simplify the segmentation, these scores are compared to the average, with values above the average assigned 1 and those below assigned 0. This binary system creates 8 possible customer segments based on whether their Recency, Frequency, and Monetary values are above or below average (Chen, Sain and Guo, 2012). The "*RFM*" column concatenates these values, providing a clear and simplified method for identifying different customer segments. We created a new feature variable '*Customer_Segment*', to assign corresponding labels to customers. Different label categories correspond to different customer segments, as detailed in Appendix H (Buckland, 2020).

4.4.2 Group Features Analysis

To better understand the characteristics of these eight customer segments, we will analyze from three dimensions: proportion of different customer types, contribution to spending by each customer type, and behavioral analysis of different customer types. This will provide valuable insights for subsequent marketing strategies.

Analysis of Customer Distribution and Spending Contribution

Figure 3 shows that Churned Customers, Best Customers, High-Value Reactivation Customers, and Big Spenders make up a significant portion of the total customer base, accounting for 87%. On the other hand, High-Potential Customers, New Customers, High-Value at Risk Customers, and Regular Customers constitute only 13% of the total customer base. This indicates that new and potential customers are relatively few in LDB's user base. This composition suggests that while LDB has a considerable number of high-value customers, it also faces a high risk of customer churn. At the same time, the potential for expanding the new customer base is relatively limited, which may require LDB to strengthen its customer acquisition and retention strategies.

Figure 3. Proportion of Different Customer Segments

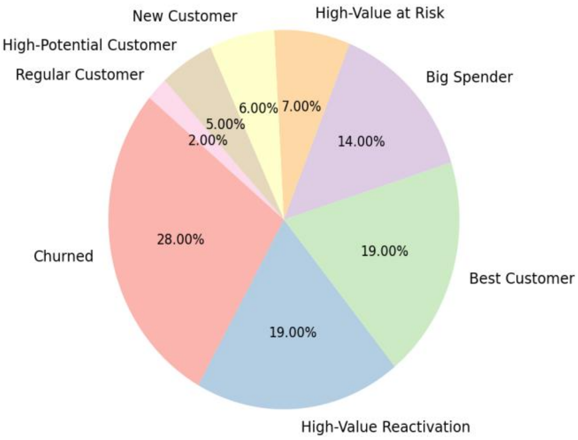


Figure 4. Spending Proportion by Customer Segment

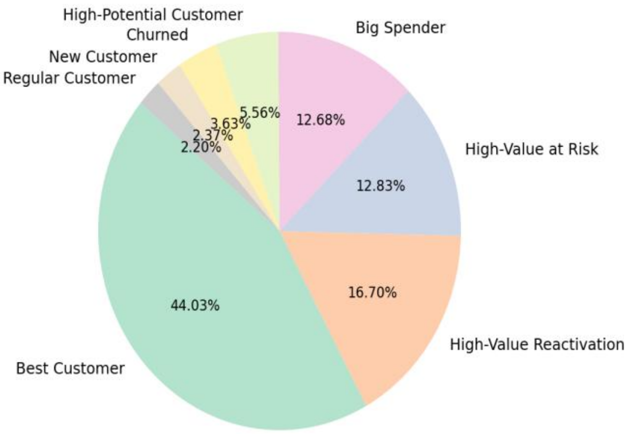


Figure 4 shows that in general, 80% of LDB's revenue is primarily derived from a smaller segment of high-value customers, including Best Customers (44.03%), High-Value Reactivation Customers (16.70%), High-Value at Risk (12.83%), and Big Spenders (12.68%). All of them are considered as high-value customers based on RFM criteria segmentation. On the other hand, High-Potential Customers, New Customers, Regular Customers, and New Customers only contribute 20% of revenue. Notably, churned customers contribute only 3.63% of the revenue, yet they make up 28% of LDB's customer base. How to convert the consumption behavior of these users, who "only take advantage of promotions," into retained customers is an issue worth considering for LDB.

Product Preference Analysis for Different Customer Segments

Understanding the product preferences of different customer segments is crucial for LDB to push products that these customer groups are more interested in, thereby enhancing customer loyalty. We will delve into understanding customers' behaviors from the perspectives of brand, product functionality, skin concerns, and specific products.

In terms of brands, according to Figure 5, we can see that Tolerian, Effaclar, and Anthelios brands are favored by all customer segments. In contrast, functional brands such as Niacinamide and Retinol LRP have significantly lower purchase quantities compared to other brands, possibly reflecting that the audience for these brands is more specialized.

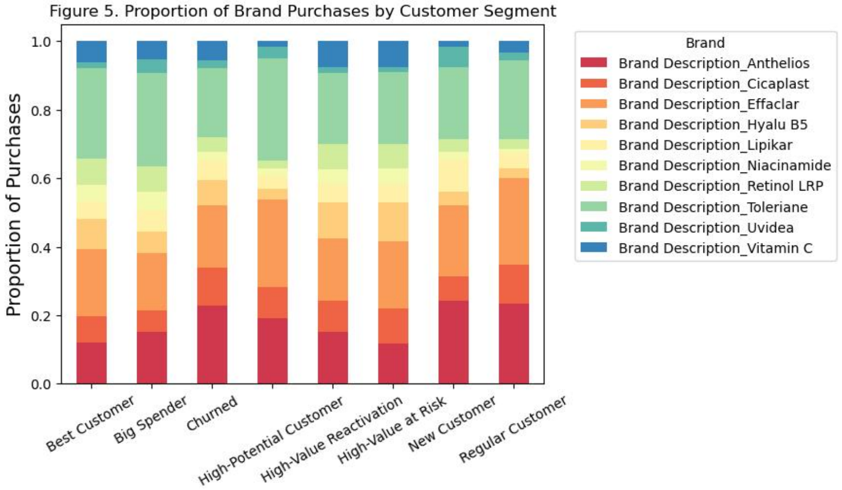


Figure 5. Proportion of Brand Purchases by Customer Segment

Figure 6 and Figure 7 show that facial care products are the most popular across all customer segments, especially among Best Customers. For high-value customers, anti-aging products are the second most in-demand after sensitive skin products, with High-Value Reactivation Customers showing a particularly strong demand for anti-aging products, indicating that new product launches or personalized recommendations could be effective in re-engaging them. For regular customers, body care products are their second most in-demand, particularly among Churned Customers, who may have stopped purchasing due to unmet expectations. Additionally, products suitable for sensitive skin are highly popular across all segments, especially among Best Customers, which aligns with La Roche-Posay's focus on sensitive skin care. It is also worth noting that demand for products targeting pigmentation and dark spots is generally low across all customer segments.

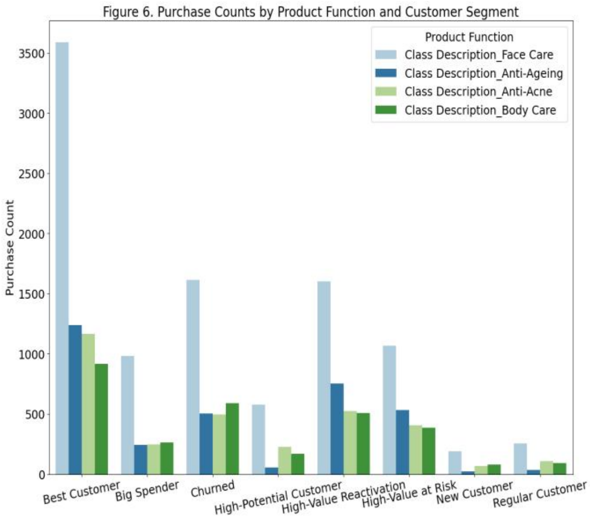


Figure 6. Purchase Counts by Product Function and Customer Segment

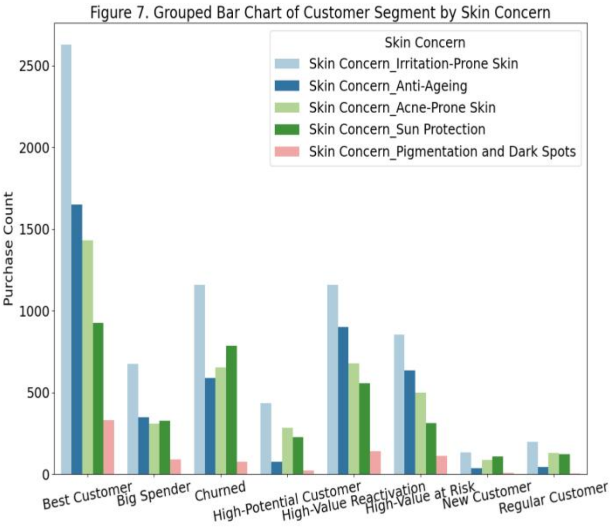
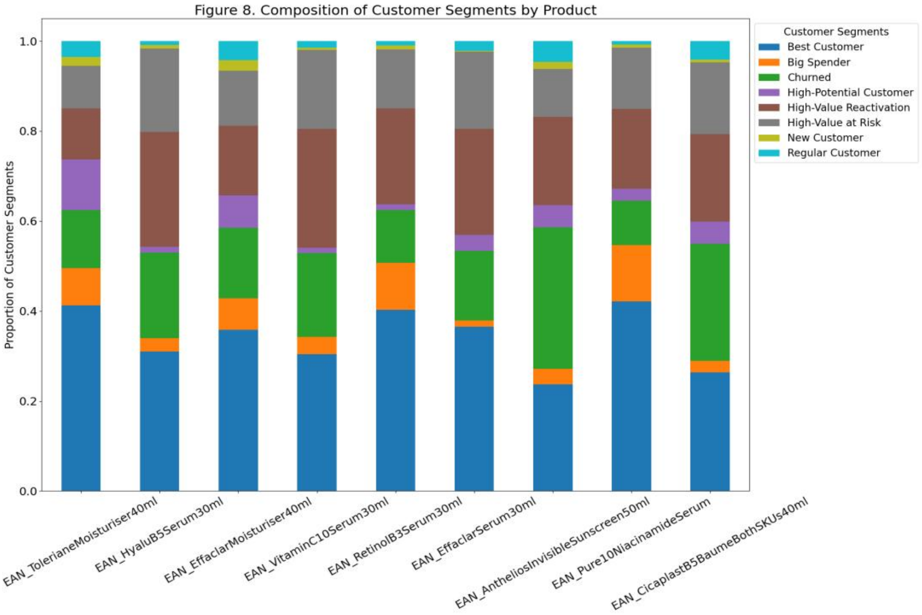


Figure 7. Grouped Bar Chart of Customer Segment by Skin Concern

Connected with the heatmap in Appendix I, Figure 8 shows the purchasing habits of different customer segments. It is evident that Best Customers have high purchase volumes across all product types, indicating broad consumer demand. Additionally, for regular customers (including churned customers), they tend to favor anti-acne and sunscreen products such as Anthelios Invisible Sunscreen 50ml and Effaclar Moisturiser 40ml. Notably, churned customers have purchased Anthelios Invisible Sunscreen 50ml even more frequently than Best Customers. For high-spending segments like High-Value Reactivation Customers and High-Value at Risk Customers, their demand for antioxidant and anti-aging products is higher than that of other segments.

Figure 8. Composition of Customer Segments by Product



5. Feature Engineering

5.1 Target Variable with Quantile Transformation

Based on EDA analysis, the target variable shows the right-skewed distribution and thus the log transformation is used in the target variable. While applying log transformation effectively mitigates the right skewness, the target variable still exhibits a pronounced bimodal distribution as Figure 10. Therefore, Quantile-Transform method is used to address the issue as Figure 9.

Figure 9. Distribution of Total Spent from Nov 2023 to May 2024 (Quantile Transformation)

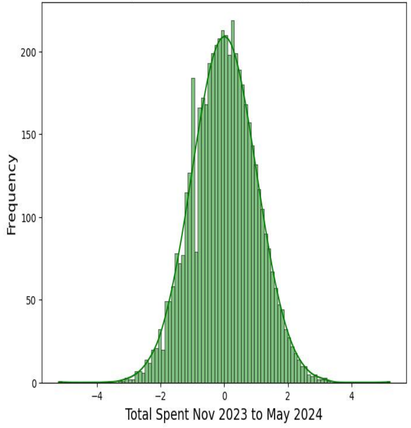
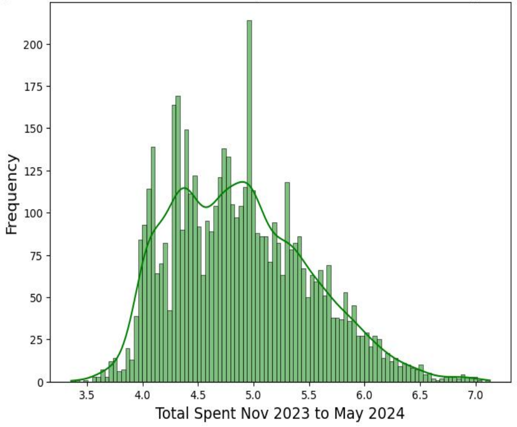


Figure 10. Distribution of Total Spent from Nov 2023 to May 2024 (log)



5.2 New Features

New features with product

According to the heatmap (in Appendix D), many features exhibit multicollinearity, such as 'Category_Sun Care' and 'Brand Description_Anthelios'. Thus, to mitigate the multicollinearity, we combined features related to brands, products, and categories that share the same functionality based on business understanding, thereby creating seven new features, as shown in Table 1 below. We delete all sub-category variables to avoid data being double-counted since the sum of the sub-category variables equal to Category variable. Besides, since facial care products are the most popular and have a variety of functions, they were separately created as a new independent feature 'Face_Care'.

Table 1 New Features with Product Created

New Feature Name	Merged Columns
'Sunscreen'	'Brand Description_Anthelios', 'Brand Description_Uvid', 'Category_Sun Care', and 'Skin Concern_Sun Protection'
'Basic_Cleaning_ Moisturizing'	'Brand Description_Cicaplast', 'Brand Description_Lipikar', 'Brand Description_Eau Thermale', 'Brand Description_Serozinc', 'Brand Description_Toleriane', 'Class Description_Body Care', 'Category_Body Care', and 'Skin Concern_Irritation-Prone Skin'.
'Anti_Ageing_Antioxidation'	'Brand Description_Hyalu B5', 'Brand Description_Retinol LRP', 'Brand Description_Vitamin C', 'Class Description_Anti-Ageing', and 'Skin Concern_Anti-Ageing'
'Anti_Acne'	'Brand Description_Effaclar', 'Class Description_Anti-Acne', and 'Skin Concern_Acne-Prone Skin'
'Spot_Correction _Pigmentation'	'Brand Description_Niacinamide' and 'Skin Concern_Pigmentation and Dark Spots'
'Bundle'	'Brand Description_Bundle' and 'Class Description_Bundle'
'Face_Care'	'Class Description_Face Care', 'Category_Face Care'

New features with time-related variables

Based on the EDA analysis, there exists the seasonal purchasing behaviour. In order to better capture the seasonality factors in the model, eight new features are created by taking the difference of the existing variables to represent the consumption amount and frequency for the four quarters, which is shown in Table 2 below.

Table 2 New Features with Time-related Variables Created

New Feature Name	New Feature Name
'spent_12plus'	'count_12plus'
'spent_9to12'	'count_9to12'
'spent_6to9'	'count_6to9'
'spent_3to6'	'count_3to6'

For example, 'spent_9to12' represents the total spending of the customer from December 2022 to February 2023, and it also reflects the customer's spending during the summer season. 'count_9to12' represents total number of transactions made by the customer from December 2022 to February 2023.

5.3 Feature Scaling

The range of values for the independent variables varies significantly, with some representing purchase frequency and others representing monetary amounts. Therefore, to improve model performance, the Z-score standardization method was applied to independent variables.

6. Model Building

The analysis includes three predictive models: regression models, decision tree models, and XGBoost models. The objective is to identify the most effective model to make recommendations as the final model. Evaluation is based on Root Mean Squared Error (RMSE), which measures prediction accuracy—lower RMSE indicates higher accuracy, making it more favorable. All of these models aim to predict the same response variable 'Total_Spent_Nov23_May24'. Detailed discussions of each model will be provided in the subsequent sections of this report.

6.1 Model 1: Linear Model

6.1.1 OLS/LASSO/Elastic Net

The linear regression model assumes a linear relationship between the target variable and the predictors. A total of 26 features were selected to fit four types of linear regression models, including Ordinary Least Squares (OLS), Least Absolute Shrinkage and Selection Operator (LASSO), and Elastic Net regression.

The OLS method fits the model by minimizing the residual sum of squares between the observed and predicted values of the target variable, while Lasso and Elastic Net introduce a penalty term, alpha, to regularize the model's complexity and prevent overfitting. When alpha equals zero, LASSO regression is equivalent to OLS. As alpha increases, in the case of Lasso, some coefficients are shrunk to zero, thus performing feature selection. Elastic Net combines Lasso and Ridge by fine-tuning the hyperparameter alpha, allowing the model to approach Lasso as alpha increases to 1 or approach Ridge as alpha decreases to 0.

Table 3 shows the comparison of test RMSE among the three models. The RMSE on the transformed data reflects how the model performs on the transformed scale, but it cannot be directly used to assess the model's predictive accuracy on the original data scale. Therefore, an inverse transformation should be applied to the transformed target variable to accurately represent the model's predictive performance on the original data. The LASSO model has the lowest RMSE on the test set, outperforming the other two models. Consequently, LASSO model is selected as the baseline model.

Table 3: Linear Model Result

Models	Test RMSE (after inverse)
LASSO	**124.347**
Elastic Net	125.403
Linear	126.328

According to J, the top 20 β coefficients in the LASSO model are presented. It can be clearly seen that features with larger positive coefficients are all related to past spending amounts, such as "Total_Spent_3M" (0.27), indicating that these features are positively correlated with the customer's future spending amount. An increase in the values of these features is associated with an increase in the target sales units. Specifically, for every unit increase in the total spending over the past three months, the customer's future spending amount is, on average, expected to increase by 0.27 units, holding other variables constant. In contrast, some independent variables have a negative impact on the target variable, such as EAN_AntheliosInvisibleSunscreen50ml. When the value of this feature increases, the predicted future spending amount may decrease.

In addition, we also studied and analyzed the assumptions of linear regression, focusing mainly on the absence of perfect multicollinearity between independent variables, as well as the normality and homoscedasticity of errors. First, we used the VIF (Variance Inflation Factor) method to calculate the VIF value for each independent variable (Daoud, 2017). If the VIF value is high, it indicates that there is a strong linear relationship between the independent variable and other independent variables. According to Appendix K, we can observe creating new features has effectively mitigated multicollinearity between the independent variables. In addition, The Q-Q plot (Appendix K) is used to check the normality of the residuals. The results show that the sample almost aligns with the red line representing the normal distribution, indicating that most of the residuals follow a normal distribution, though some deviation exists in the tails. The residual plot (Appendix K) is used to check homoscedasticity, and the results show a clear funnel-shaped distribution, which suggests the presence of heteroscedasticity. This indicates that as the predicted values increase, the errors also increase. Therefore, further work may be required to introduce more complex models to better capture the nonlinear relationships and finer details in the higher transaction value ranges.

6.2 Model 2: Nonlinear Model

Based on the previous correlation analysis, it was found that the linear correlation between features and the target variable is weak. Traditional linear regression analysis cannot handle complex and nonlinear datasets, so it is necessary to explore models capable of capturing the nonlinear relationships between features and the target, such as tree-based models. Moreover, in building tree models, nonlinear models are inherently better at handling the original non-normal distribution of data. Additionally, the extreme values of transaction amounts (such as large transactions) contain important information for prediction. Therefore, we ultimately decided to use the original dependent variable rather than the transformed one with distribution changes.

6.2.1 Decision Tree

The key hyperparameters of the decision tree model include max_depth, min_samples_leaf, min_samples_split, and ccp_alpha. max_depth determines the depth of the tree, with deeper trees potentially leading to overfitting; min_samples_leaf controls the minimum number of samples in leaf nodes, and smaller values result in deeper trees, increasing the risk of overfitting; min_samples_split sets the minimum number of samples required to split a node, where smaller values make the tree more complex and may also lead to overfitting; ccp_alpha controls pruning, with larger values resulting in more pruning and reduced tree complexity. By optimizing these parameters through grid search, overfitting and underfitting can be effectively avoided.

To calculate the criterion of each node, we use mean squared error and the greedy algorithm that we require the mean squared error of each feature split have to be smallest which uses the formula:

$$\mathrm{MSE_{total}} = \frac{\mathrm{N_{left}}}{\mathrm{N_{total}}} \times \mathrm{MSE_{left}} + \frac{\mathrm{N_{right}}}{\mathrm{N_{total}}} \times \mathrm{MSE_{right}}$$

where N refers to the number of samples in left and right split and the total split. This method selects the best splits however does not look ahead to optimise the future splits.

After grid search, the best performed hyperparameters are set as following: ‘max_depth’ as 6, ‘min_samples_leaf’ as 44, ‘min_samples_split’ as 2 and ‘ccp_alpha’ as 0. The result table 4 below shows the comparison results between using original dependent variable and using quantile transferred dependent variable under the best hyperparameters on test set, as mentioned, the original dependent variable has better performance on the model, possibly because that the quantile transformation has changed the scale and reduced the performance when making a split.

Table 4: Comparison of Decision Tree under different conditions	
Model	**Test RMSE**
Decision Tree (Original)	123.734
Decision Tree (Inverse Transformed)	128.680

The decision tree regression in Appendix L prioritizes features that minimize the squared error of the target variable, meaning the splits are ordered by feature importance. By inspection, the variable 'Total_Spent_3M' emerges as the most significant as it is the top node of the tree, consistent with the earlier EDA findings that most recent quarterly spending significantly impact the spending patterns. The first split shows that customers spending $167.8 or less have an average predicted target value of $150.989, while those spending more have an average predicted value of $250.15. The second split involves spending-related variables from the nearest two quarters as well, 'Total_Spent_3M' and 'spent_3to6', reinforcing that recent quarterly spending patterns are important in predicting customer behavior. Among the features, only 'Face_Care', 'Anti_Ageing_Antioxidation', and 'Basic_Cleaning_Moisturizing' are product-specific, suggesting these are the most important product-related features. On the other hand, previous quarterly spending provides more accuracy in predictions as they occupy most nodes, indicating the importance of historical spending data. This insight can guide LDB's product marketing strategies and improve customer loyalty classification , ultimately enhancing overall customer value.

6.2.2 XGBoost

The XGBoost model is widely praised for its efficiency and effectiveness in handling complex datasets (Chen & Guestrin, 2016). Compared to a single decision tree, XGBoost enhances accuracy and reduces the risk of overfitting by iteratively training multiple weak learners, where each tree improves upon the errors of the previous tree (Bentéjac et al., 2020).

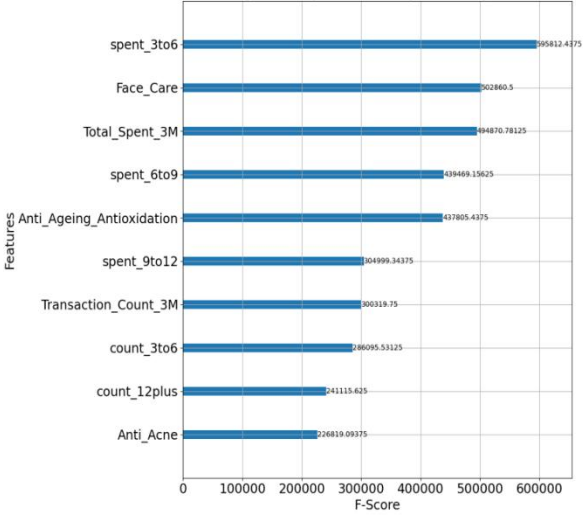
First, a baseline XGBoost model was constructed and trained, achieving an RMSE of 122.414 on the test set. To further improve model performance, a staged approach was used to tune the hyperparameters of the XGBoost model. This method is both reasonable and effective, as it improves tuning efficiency and model performance (Bartz-Beielstein et al., 2023). Specifically, the approach involves first tuning the number of trees (n_estimators), followed by tuning parameters that directly control model complexity (such as max_depth and min_child_weight), then adjusting sampling rates and regularization parameters, and finally fine-tuning the learning rate, which controls the model's convergence speed. This tuning sequence effectively reduces the search space, enhances tuning efficiency, avoids interference between parameters, and reduces the risk of overfitting (Chen & Guestrin, 2016). Furthermore, GridSearchCV was employed for each hyperparameter tuning stage. This technique explores various hyperparameter combinations, including max_depth, n_estimators, learning_rate, subsample, and colsample_bytree. By training and cross-validating models on each combination, it identifies the optimal hyperparameters, resulting in a well-tuned XGBoost model that achieved an RMSE of 120.620 on the test data. The hyperparameter tuning process is crucial as it allows the model to better align with the characteristics of the specific data, thus enhancing prediction accuracy. The following table 5 show the results of using the transformed and original dependent variables in the model. As can be seen, even after hyperparameter tuning, the model with the transformed dependent variable has a higher RMSE on the test set compared to the baseline model with the original dependent variable. This further illustrates that while Quantile Transformation can make the distribution of the dependent variable closer to a normal distribution, it does not always improve the model's predictive performance.

Table 5: Comparison of XGBoost model under different conditions

	Test RMSE (transformed y)	**Test RMSE (original y)**
Baseline XGBoost	126.798	122.414
Tuned XGBoost	125.283	**120.620**

Figure 11 shows that the features ‘Face Care’, ‘Total_Spent_3M’, and ‘spent_3to6’ have a significant impact on predicting future transaction values, highlighting the importance of facial care products as well as consumer behavior during the winter and spring months. Additionally, the historical purchase frequency of anti-aging and antioxidant products also significantly affects future transaction values.

Figure 11. Top 10 Feature Importances by Gain



Appendix M illustrates XGBoost model tends to underestimate when actual future transaction values increase. This pattern indicates that while the XGBoost model captures the overall trend, its accuracy at higher transaction value levels is insufficient. In future analyses, it may be necessary to introduce more features related to high spending amounts to enable the model to better capture the behavior patterns of high-spending customers, improving its ability to predict high transaction amounts.

6.3 Model Evaluation

Table 6 presents the RMSE values for various models tested on the dataset, with the XGBoost model showing the lowest RMSE. The comparative analysis highlights the trade-off between model complexity and performance, with advanced models like Decision Trees and XGBoost outperforming simpler models like OLS. While more complex models demonstrate better performance and predictive capabilities, they also come with drawbacks such as reduced interpretability and increased computational requirements. Other challenges of complex models include sensitivity to hyperparameters and the tendency to overfit, whereas simpler models may oversimplify predictions.

Considering the critical role of customer transaction value in the success of L'Oréal, it is preferable to prioritize and invest in complex models that can provide both accurate predictions and interpretability. Therefore, the Decision Tree model has been selected as the most suitable model for predicting future customer transaction value for L'Oréal, as it strikes the optimal balance between accuracy and interpretability.

Table 6. RMSE Results Table

Models	**Test RMSE (transformed y)**	**Test RMSE (original y)**
OLS	126.328	N/A
Lasso	124.437	N/A
Elastic Net	125.403	N/A
Decision Tree	128.863	123.734
XGBoost	126.124	120.620

7. Conclusions and Recommendations

To help LDB optimize customer management and boost sales performance, our model estimates that the following two strategies can encourage customers to spend more over the next six months: Strategy A, promoting the tier upgrade of certain high-value customers; Strategy B, enhancing AI technology to identify more skin issues. The strategy is primarily focused on high-value customers because they contribute to 86% of LDB's revenue. High-value customers are defined as customers with an M indicator of 1, including Best Customer, High-Value Reactivation, High-Value at Risk, and Big Spender. High-Value Reactivation, High-Value at Risk, and Big Spender customers still have significant marketing potential, so strategies are primarily focused on these three segments.

A. Tiered Marketing Strategy for High-Value Users Based on the RFM Model

The goals of strategy A include 1) convert ‘High-Value at Risk’ to ‘Best Customers’: Improve the ‘R-Score’ by increasing the purchase frequency in the recent period. 2) Convert ‘High-Value Reactivation’ to ‘Big Spenders’: Improve the ‘R-Score’ by increasing the purchase frequency in the recent period. 3) Convert ‘Big Spenders’ to ‘Best Customers’: Improve the ‘F-Score’ by increasing the total purchase frequency.

In goal 1 and goal 2, "High-Value at Risk" and "High-Value Reactivation" customers are the target groups. These customers have an average purchase amount of over $70 per transaction, but their purchase frequency has recently declined. To re-engage these customers and encourage them to purchase LDB products again, the strategy will be implemented based on their main purchasing preferences, which include anti-aging and anti-acne products in the face care category, such as EAN_HyaluB5 Serum 30ml, EAN_VitaminC10 Serum 30ml, and EAN_Effaclar Moisturiser 40ml. The marketing strategy consists of limited-time discounts and a loyalty reward program. First, offer limited-time discounts based on users' product preferences, such as special discounts on EAN_HyaluB5 Serum 30ml and EAN_Effaclar Moisturiser 40ml, and notify users via email and SMS. Second, implement a loyalty reward program where the more customers buy, the more discounts and exclusive benefits they receive. These benefits could include access to exclusive products, early pre-orders, or higher-tier membership perks. For example, customers who spend over $70 per transaction could enjoy a discount and the opportunity to purchase exclusive items.

The target group for goal 3 consists of "Big Spenders" customers who already have high spending amounts but insufficient purchase frequency. This indicates that these customers tend to prefer high-end skincare products. Our goal is to increase the number of transactions among these high-value customers. These customers primarily purchase anti-aging, anti-acne, and anti-spot products in the face care category, specifically including EAN_Effaclar Moisturiser 40ml, EAN_Retinol B3 Serum

30ml, and EAN_Pure10 Niacinamide Serum. Their main skin concerns are similar to other high-value customers, focusing on irritation-prone skin, skin aging, and acne. Therefore, the marketing strategy includes offering value-added services and product enhancement. First, increase customer loyalty by providing value-added services. For example, LDB can introduce a skincare subscription service that offers customized product sets every month or quarter, such as a combination of Toleriane cleansing and moisturizing products with anti-aging serums. This ensures that customers receive timely replenishments when their regular purchase interval is about to end, helping to establish their usage cycle. Second, in terms of product enhancement, LDB can introduce limited edition high-end products, such as an upgraded formula of Retinol B3 Serum 30ml. The scarcity of limited editions can stimulate the purchasing desire of Big Spenders and encourage them to purchase more frequently.

B. Use AI to develop skin diagnosis tools

Since core users pay much attention to irritation-Prone and anti-aging issues, LDB can develop AI tools for skin problem diagnosis and provide personalized product recommendations based on skin problems. At present, LDB's AI tool - SPOTSCAN+ can only diagnose Acne-Prone Skin problems, so LDB can have a further development for it to diagnose more skin problems on this basis.

Revenue and Cost Estimation

Since 'Total_Spent_3M' and 'Total_Spent_Nov23_May24' exhibit the strongest positive correlation, the amount spent by customers in the last three months significantly drives their spending in the following six months. Therefore, our strategy primarily focuses on influencing customers' spending behavior in the recent three months. We assume that once the strategy is implemented, the final effect will be reflected in changes to 'Total_Spent_3M'.

For Strategy A, we first identify all rows corresponding to 'High-Value at Risk' customers in the test dataset. Then, we create a new column, 'New_Total_Spent_3M,' to represent the 'Total_Spent_3M' under the influence of the strategy. Since these customers are expected to transition into the 'Best Customer' segment, we use the median 'Total_Spent_3M' of 'Best Customers' in the original test dataset as the value for the 'New_Total_Spent_3M' column. The same approach is applied to 'High-Value Reactivation' and 'Big Spender' customers, updating their corresponding 'New_Total_Spent_3M' values based on the median values of the customer segments they are expected to transition into under the strategy's influence. This will result in a new test dataset containing only these three customer segments.

Next, we build a new decision tree model, training it with the original training dataset that includes only these three customer segments. We then predict the future six-month transaction values for these three customer segments in both the original test dataset and the new test dataset, obtaining 'Spent_Nov23_May24' and 'After_Strategy_Spent_Nov23_May24,' respectively. By calculating the difference between the two and summing the results, we can determine the revenue increase attributed to the marketing strategy. Ultimately, the model calculates an increase in revenue of $3306 for the test dataset (385 samples). Since the total number of these three high-value customer segments in the dataset is 2445, we estimate that the implementation of the strategy will increase LDB's revenue by 20,995.25 AUD from them. Additionally, for the Dermatological Beauty division, the operating profit margins for 2022 and 2023 were 21.9% and 22.2%, respectively (L'Oréal Finance, 2024). By taking the weighted average, we calculate an average operating profit margin of 22.05% for these two years, and thus, deduce the marketing cost to be $16,365.8, as shown in the following figure.

Estimated Revenue and Costs of Strategy Execution			
Revenue	**Operating profit margin**	**Operating profit**	**Cost**
$20995.25	22.05%	$4629.45	$16365.80

8. Executive Briefing

1) Background & challenges & strategy

As a division of L'Oréal, LDB is specializing in dermatological solutions, and is expanding its online presence. In this paper, LDB uses historical customer transaction data to predict LDB's future transaction amounts and analyze user behavior to provide its "patsumers" a more personalised shopping experience.

2) Key Findings

The analysis reveals that the total spending of customers over the past six months, especially the total spending in the most recent three months, significantly stimulates customers' spending in the following six months. Additionally, customer spending behavior shows seasonal patterns. Winter spending is much higher than in summer, with the most significant fluctuations occurring in the last three months, from September to November. Furthermore, customers who prefer specific product categories, particularly those favoring face care and anti-aging products, are expected to significantly increase their spending in the next six months. This group of customers is often identified as LDB's high-value customers. Thus, the subsequent marketing strategies will target high-value customers.

3) Model Implementation and Results

In this paper, we used three types of models, regression models, decision tree models, and XGBoost models. By comparing the Root Mean Squared Error (RMSE) results of these three types of models, we found that the RMSE of the XGBoost model has the smallest value, followed by the decision tree model. However, considering the requirements for both accurate predictions and interpretability, we decided to finally choose the decision tree model as the optimal model (RMSE value is 123.73). Through the feature importance of the model, the variable 'Total_Spent_3M' has the most significant impact on the prediction results of the model, followed by 'spent_3to6' and 'Face_care'.

4) Recommendations

Based on the results of our above analysis, we provide two marketing strategies, one is Tiered Marketing Strategy for High-Value Users Based on the RFM Model, which aims to target key segments of high-value users with tailored offers and services to elevate their customer status and boost their engagement with LDB products. The other is expanding its current AI tool, SPOTSCAN+, to diagnose a wider range of skin issues, so that to increase customer loyalty.

5) Limitations

Since the data covers only one year, the conclusions on seasonal consumption behavior may be inaccurate. Multi-year data analysis is needed to confirm these patterns. Additionally, our current model assumes all customers will engage in transaction behavior. Besides, in reality, it may be necessary to build a two-stage model, that is, to first build a predictive classification model to evaluate whether these customers will make transactions, and then to predict the transaction amount.

Reference List

Anitha, P. and Patil, M.M. (2019). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, 34(5). doi:https://doi.org/10.1016/j.jksuci.2019.12.011

Bartz-Beielstein, T., Chandrasekaran, S., & Rehbach, F. (2023). Case Study II: Tuning of Gradient Boosting (xgboost). *Hyperparameter Tuning for Machine and Deep Learning with R*, 221–234. https://doi.org/10.1007/978-981-19-5170-1_9

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3). https://doi.org/10.1007/s10462-020-09896-5

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification And Regression Trees. Routledge. https://doi.org/10.1201/9781315139470

Buckland, T. (2020). *RFM Segments using RFM Analysis [In-Depth Guide]*. [online] MoEngage Blog. Available at: https://www.moengage.com/blog/rfm-analysis-using-rfm-segments/

Chen, T., & Guestrin, C. (2016). XGBoost: a Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 785–794

Daoud, J. I. (2017). Multicollinearity and regression analysis. In *Journal of Physics: Conference Series* (Vol. 949, No. 1, p. 012009). IOP Publishing.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. In Springer Series in Statistics. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

L'Oréal finance. (February 8, 2024). 2023 Annual Results. https://www.loreal-finance.com/eng/news-release/2023-annual-results

Safari, F., Safari, N. and Montazer, G.A. (2016). Customer lifetime value determination based on RFM model. Marketing Intelligence & Planning, 34(4), pp.446–461. doi:https://doi.org/10.1108/mip-03-2015-0060

Appendix

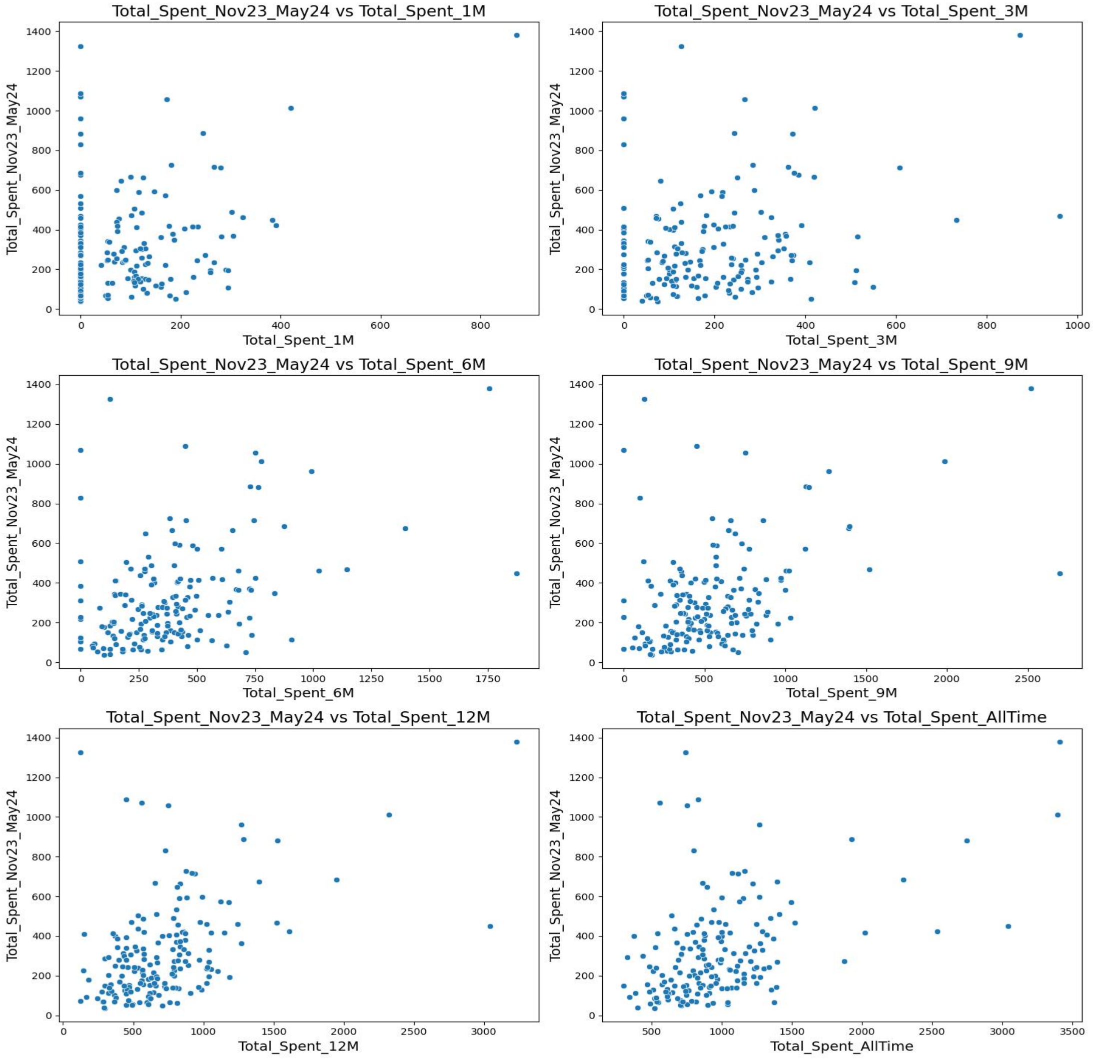
Appendix A

Summary Statistics					
Variable	**Non-Null Count**	**Dtype**	**Variable**	**Non-Null Count**	**Dtype**
CustomerID	6400	object	Post Code	1074	Float64
Has_Transaction_Nov23_May24	6400	Int64	Total_Spent_1M	6400	Float64
Total_Spent_Nov23_May24	6400	Float64	Transaction_Count_1M	6400	Int64
Total_Spent_3M	6400	Float64	Total_Spent_6M	6400	Float64
Transaction_Count_3M	6400	Int64	Transaction_Count_6M	6400	Int64
Total_Spent_9M	6400	Float64	Total_Spent_12M	6400	Float64
Transaction_Count_9M	6400	Int64	Transaction_Count_12M	6400	Int64
Total_Spent_AllTime	6400	Float64	Brand Description_Vitamin C	6400	Int64
Transaction_Count_AllTime	6400	Int64	Brand Description_Cicaplast	6400	Int64
Brand Description_Anthelios	6400	Int64	Brand Description_Eau Thermale	6400	Int64
Brand Description_Bundle	6400	Int64	Brand Description_Hyalu B5	6400	Int64
Brand Description_Effaclar	6400	Int64	Brand Description_Niacinamide	6400	Int64
Brand Description_Lipikar	6400	Int64	Brand Description_Serozinc	6400	Int64
Brand Description_Retinol LRP	6400	Int64	Brand Description_Uvidea	6400	Int64
Brand Description_Toleriane	6400	Int64	Category_Body Care	6400	Int64
Category_Face Care	6400	Int64	Category_Sun Care	6400	Int64
Sub-Category_Body Moisturiser	6400	Int64	Sub-Category_Body Wash	6400	Int64
Sub-Category_Eye Cream	6400	Int64	Sub-Category_Face Cleanser	6400	Int64
Sub-Category_Face Mask	6400	Int64	Sub-Category_Face Moisturiser	6400	Int64
Sub-Category_Face Serum	6400	Int64	Sub-Category_Sunscreen	6400	Int64
Sub-Category_Tinted Sunscreen	6400	Int64	Sub-Category_Toner & Mist	6400	Int64
Skin Concern_Acne-Prone Skin	6400	Int64	Skin Concern_Anti-Ageing	6400	Int64

Skin Concern_Irritation-Prone Skin Skin Concern_Sun Protection	6400	Int64	Skin Concern_Pigmentation and Dark Spots EAN_TolerianeMoisturiser40ml	6400	Int64
EAN_HyaluB5Serum30ml	6400	Int64	EAN_EffaclarMoisturiser40ml	6400	Int64
EAN_VitaminC10Serum30ml	6400	Int64	EAN_RetinolB3Serum30ml	6400	Int64
EAN_EffaclarSerum30ml	6400	Int64	EAN_AntheliosInvisibleSunscreen50ml	6400	Int64
EAN_Pure10NiacinamideSerum	6400	Int64	EAN_CicaplastB5BaumeBothSKUs40ml	6400	Int64

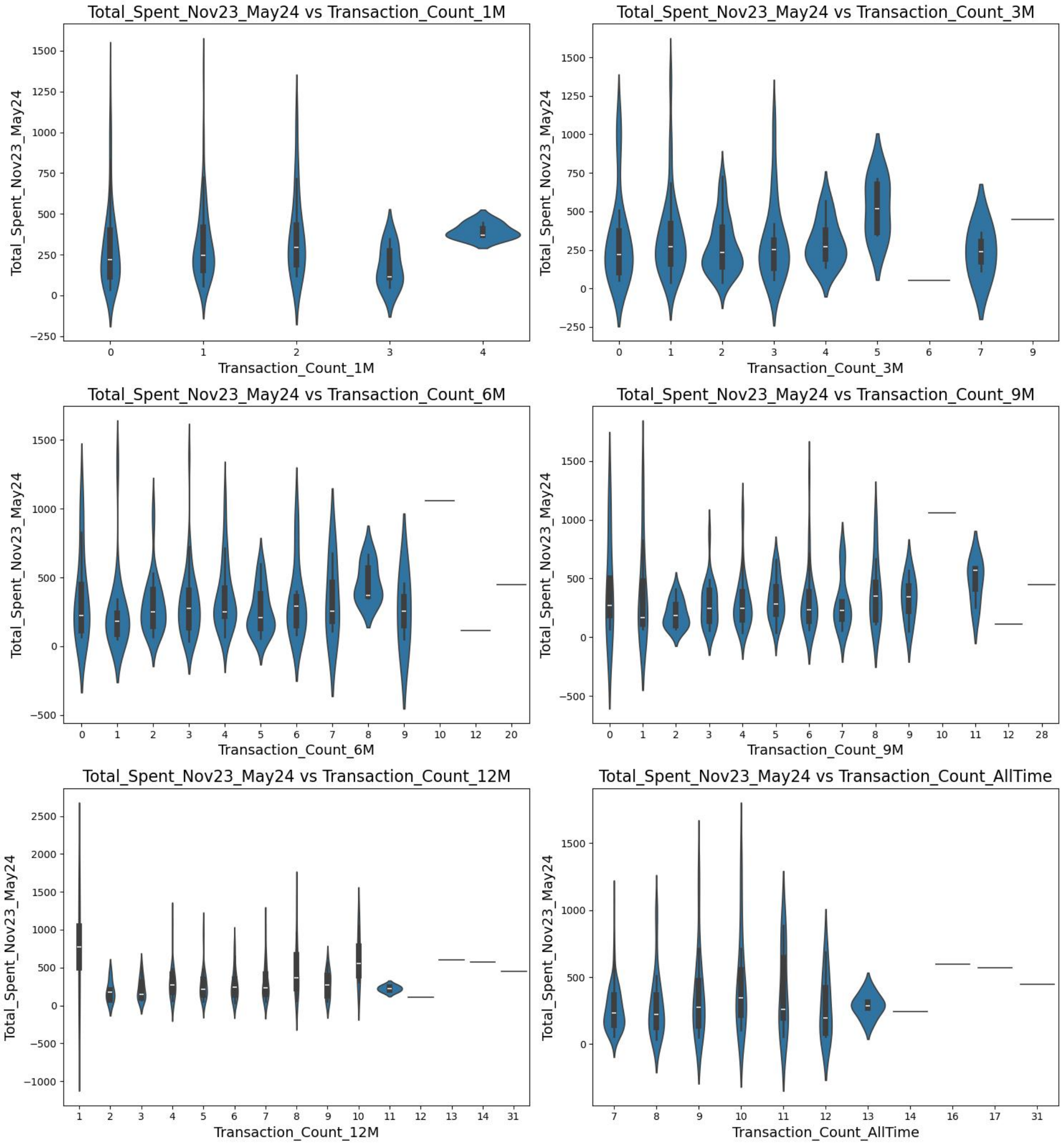
Appendix B

Appendix B. Analysis of Observations in Spent Variable Beyond Thresholds

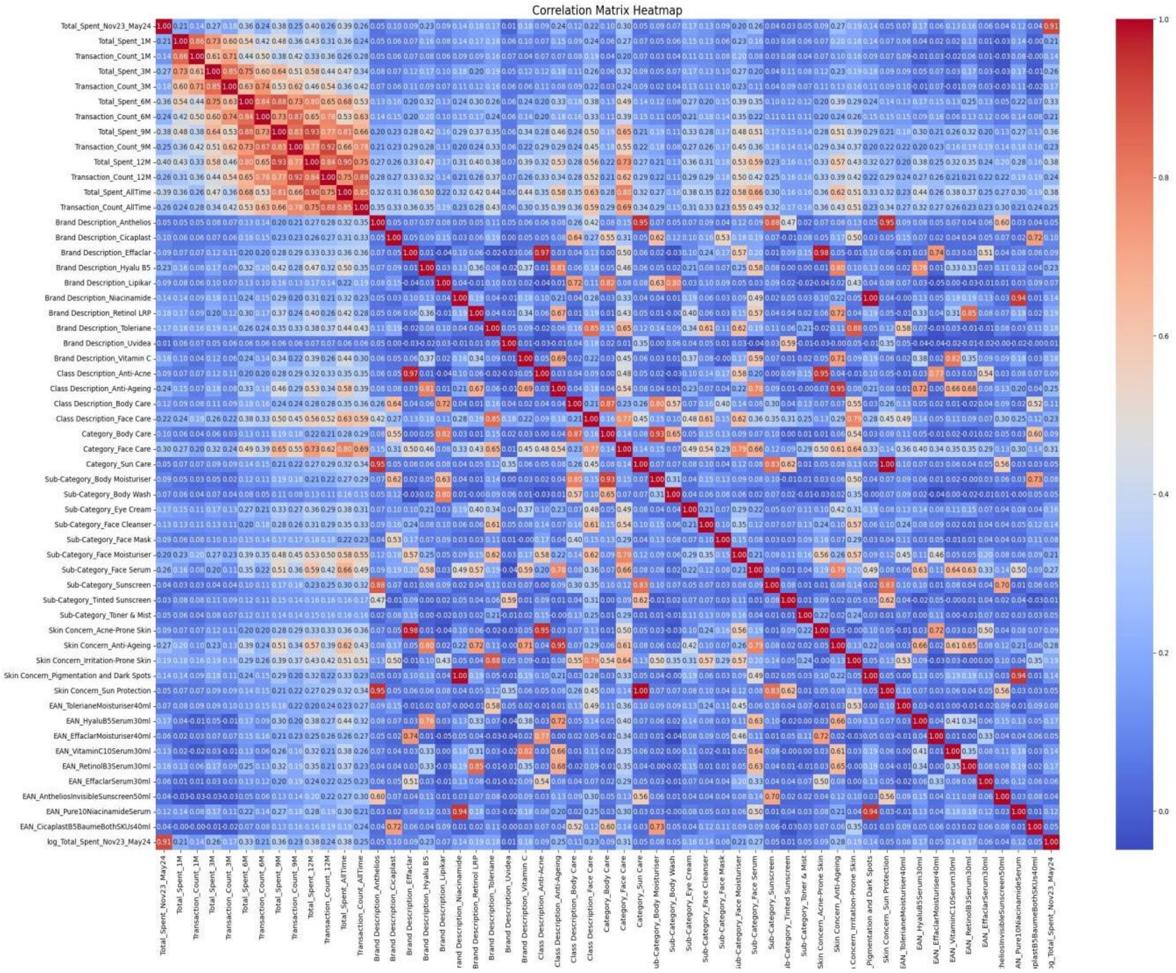


Appendix C

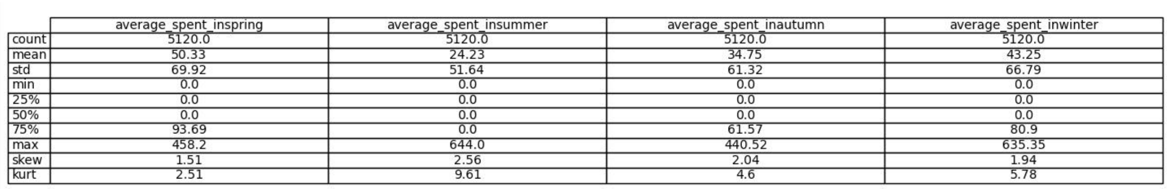
Appendix C. Analysis of Observations in Count Variable Beyond Thresholds



Appendix D

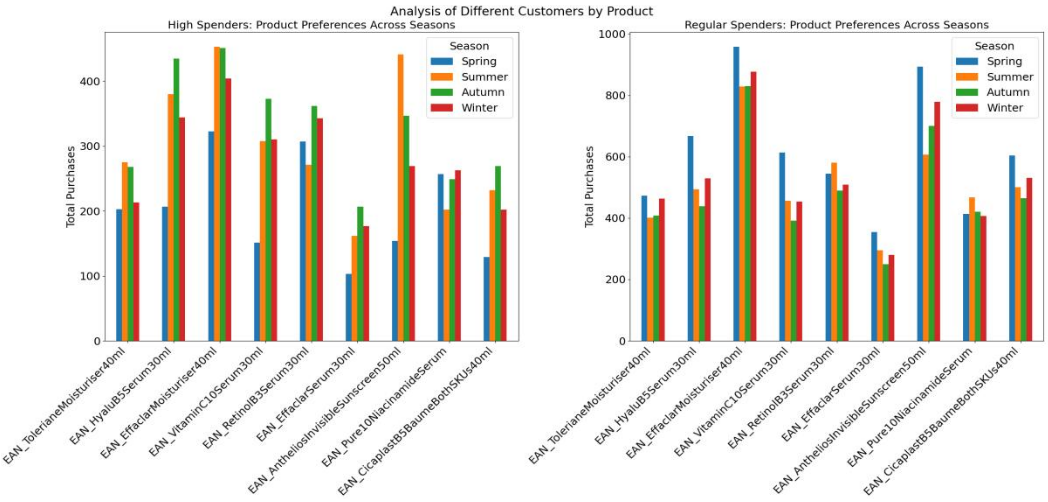


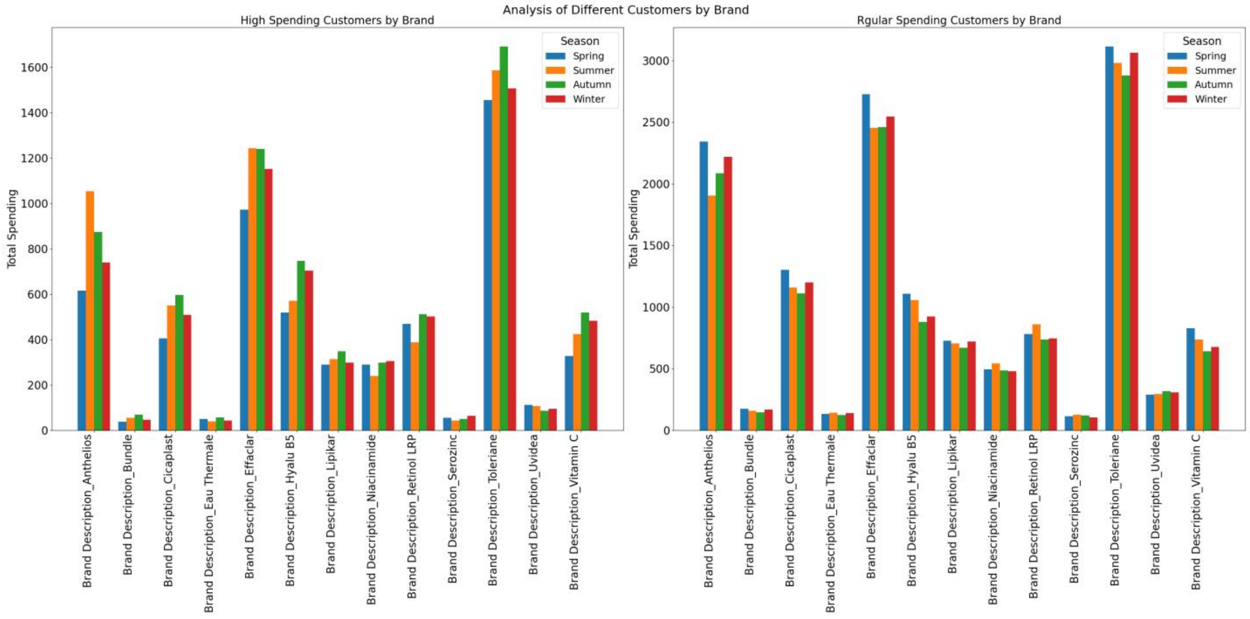
Appendix E

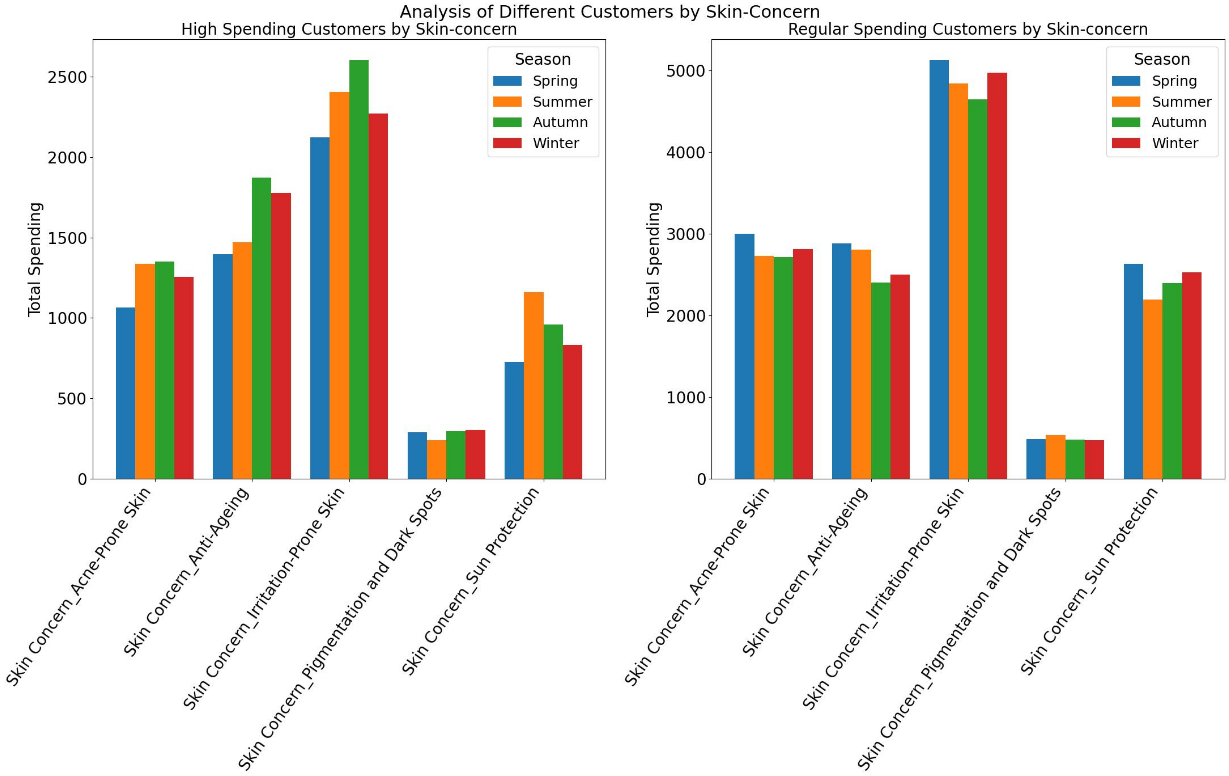


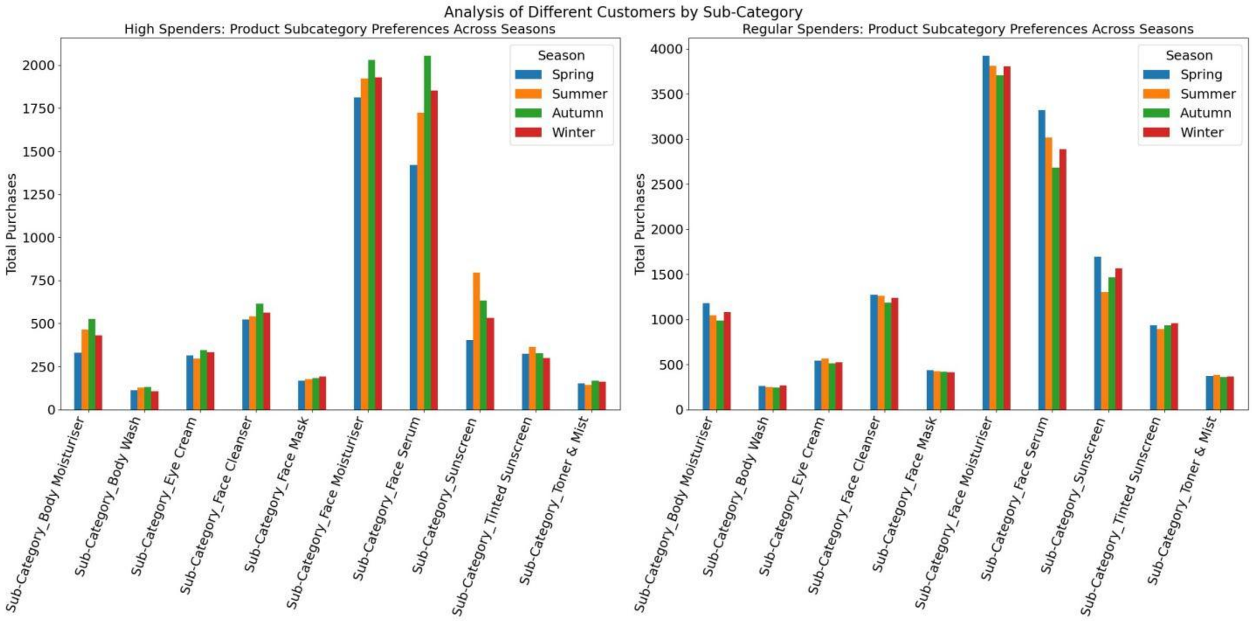
	average_spent_inspring	average_spent_insummer	average_spent_inautumn	average_spent_inwinter
count	5120.0	5120.0	5120.0	5120.0
mean	50.33	24.23	34.75	43.25
std	69.92	51.64	61.32	66.79
min	0.0	0.0	0.0	0.0
25%	0.0	0.0	0.0	0.0
50%	0.0	0.0	0.0	0.0
75%	93.69	0.0	61.57	80.9
max	458.2	644.0	440.52	635.35
skew	1.51	2.56	2.04	1.94
kurt	2.51	9.61	4.6	5.78

Appendix F




Analysis of Different Customers by Brand
High Spending Customers by Brand
Rgular Spending Customers by Brand
Total Spending
Season
Spring
Summer
Autumn
Winter
0
200
400
600
800
1000
1200
1400
1600
500
1500
2000
2500
3000
Brand Description_Anthelios
Brand Description_Bundle
Brand Description_Cicaplast
Brand Description_Eau Thermale
Brand Description_Effaclar
Brand Description_Hyalu B5
Brand Description_Lipikar
Brand Description_Niacinamide
Brand Description_Retinol LRP
Brand Description_Serozinc
Brand Description_Toleriane
Brand Description_Uvidea
Brand Description_Vitamin C


Analysis of Different Customers by Skin-Concern
High Spending Customers by Skin-concern
Regular Spending Customers by Skin-concern
Total Spending
Season
Spring
Summer
Autumn
Winter
0
500
1000
1500
2000
2500
3000
4000
5000
Skin Concern_Acne-Prone Skin
Skin Concern_Anti-Ageing
Skin Concern_Irritation-Prone Skin
Skin Concern_Pigmentation and Dark Spots
Skin Concern_Sun Protection



Appendix G

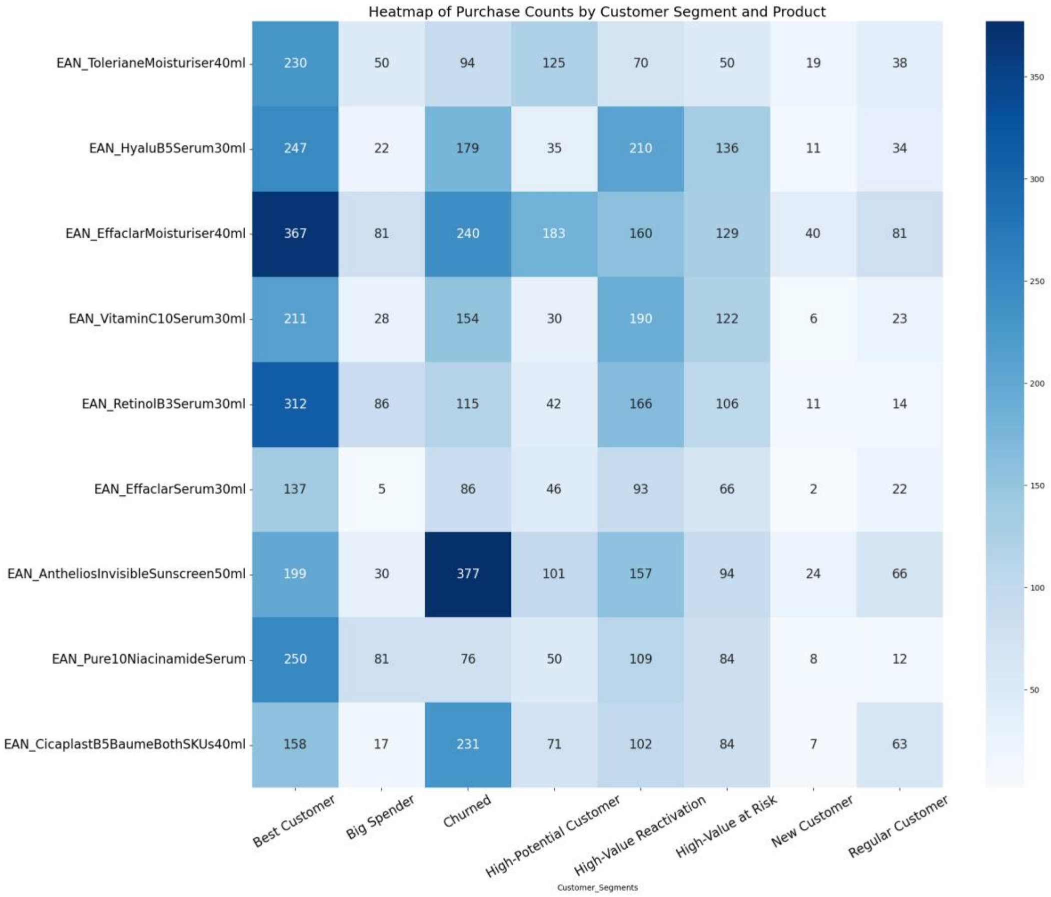
Appendix G. The Scoring Metrics of RFM		
R_Score	**R**	**Definition**
1	[9, ∞]	Last purchase was more than 9 months ago.
2	[6, 9)	Last purchase was 6-9 months ago.
3	[3, 6)	Last purchase was 3-6 months ago.
4	[1, 3)	Last purchase was 1-3 months ago.
5	[0, 1)	Last purchase was within one month.
F_Score	**F**	**Definition**
1	0	Purchased 0 times in the past nine months.
2	1	Purchased 1 times in the past nine months.
3	2	Purchased 1-2 times in the past nine months.
4	[3,5)	Purchased 3-4 times in the past nine months.
5	[5,∞）	More than five purchases in the past nine months.

M_Score	**M**	**Definition**
1	[0,50)	Average payment per purchase in the past nine months is below $50 (exclusive).
2	[50,75)	Average payment per purchase in the past nine months is between 50 (inclusive) and 75 (exclusive).
3	[75,100)	Average payment per purchase in the past nine months is between $75 (inclusive) and $100 (exclusive).
4	[100,140)	Average payment per purchase in the past nine months is between $100 (inclusive) and $140 (exclusive).
5	[140,∞)	Average payment per purchase in the past nine months is above $140.

Appendix H

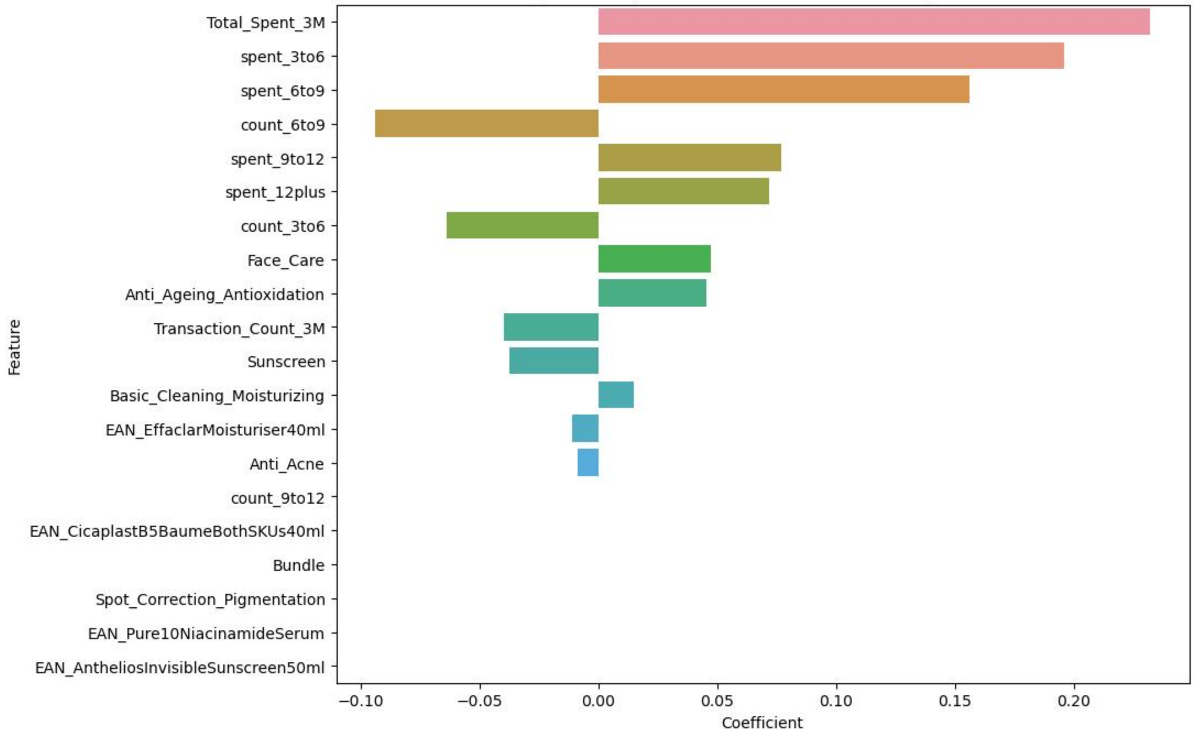
Appendix H. Customer Segmentation		
RFM	**Customer_Segments**	**Definition**
111	Best Customers	Recent purchase, high frequency, high spending
110	High-Potential Customers	Recent purchase, high frequency, low spending
101	Big Spenders	Recent purchase, low frequency, high spending
100	New Customers	Recent purchase, low frequency, low spending
11	High-Value at Risk Customers	Recent inactive, high frequency, high spending
10	Regular Customer	Recent inactive, high frequency, low spending
1	High-Value Reactivation Customers	Recent inactive, low frequency, high spending
0	Churned Customers	Recent inactive, low frequency, low spending

Appendix I

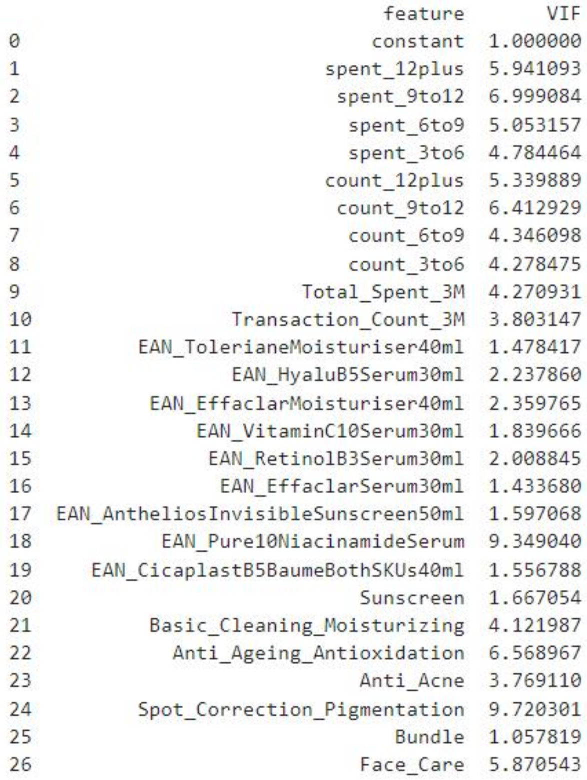


Appendix J

Figure 11 Top 20 Coefficients from Lasso Regression



Appendix K



```
                                 feature       VIF
0                               constant  1.000000
1                            spent_12plus  5.941093
2                             spent_9to12  6.999084
3                              spent_6to9  5.053157
4                              spent_3to6  4.784464
5                            count_12plus  5.339889
6                             count_9to12  6.412929
7                              count_6to9  4.346098
8                              count_3to6  4.278475
9                          Total_Spent_3M  4.270931
10                    Transaction_Count_3M  3.803147
11            EAN_TolerianeMoisturiser40ml  1.478417
12                    EAN_HyaluB5Serum30ml  2.237860
13             EAN_EffaclarMoisturiser40ml  2.359765
14                 EAN_VitaminC10Serum30ml  1.839666
15                  EAN_RetinolB3Serum30ml  2.008845
16                   EAN_EffaclarSerum30ml  1.433680
17   EAN_AntheliosInvisibleSunscreen50ml  1.597068
18              EAN_Pure10NiacinamideSerum  9.349040
19       EAN_CicaplastB5BaumeBothSKUs40ml  1.556788
20                               Sunscreen  1.667054
21             Basic_Cleaning_Moisturizing  4.121987
22               Anti_Ageing_Antioxidation  6.568967
23                               Anti_Acne  3.769110
24             Spot_Correction_Pigmentation  9.720301
25                                  Bundle  1.057819
26                               Face_Care  5.870543
```

Figure 12 Q-Q Plot of Residuals

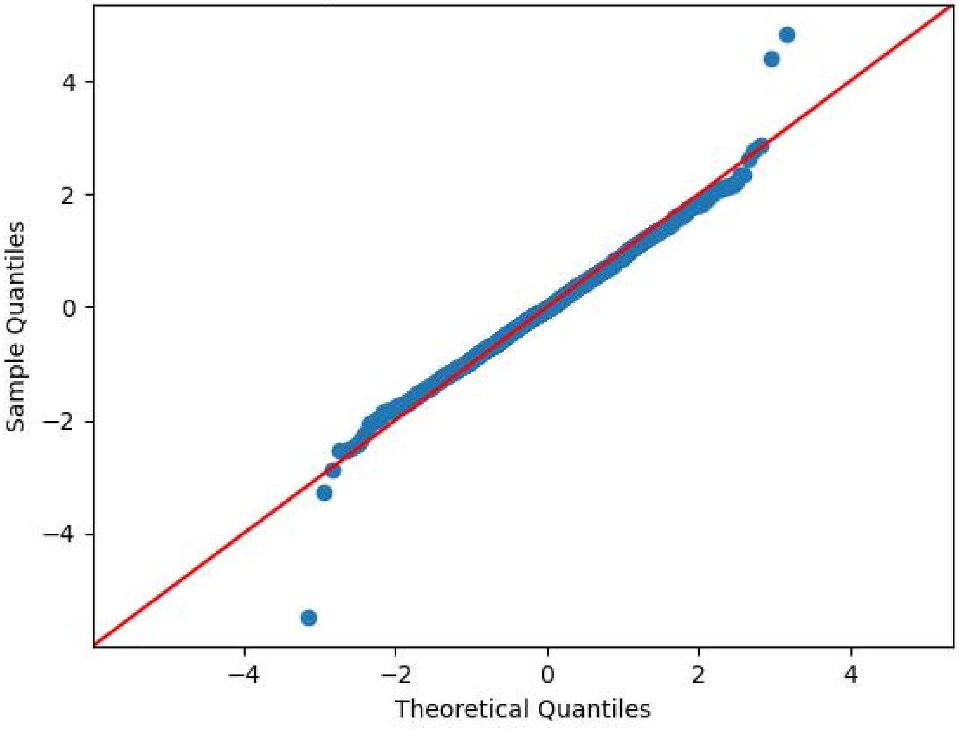
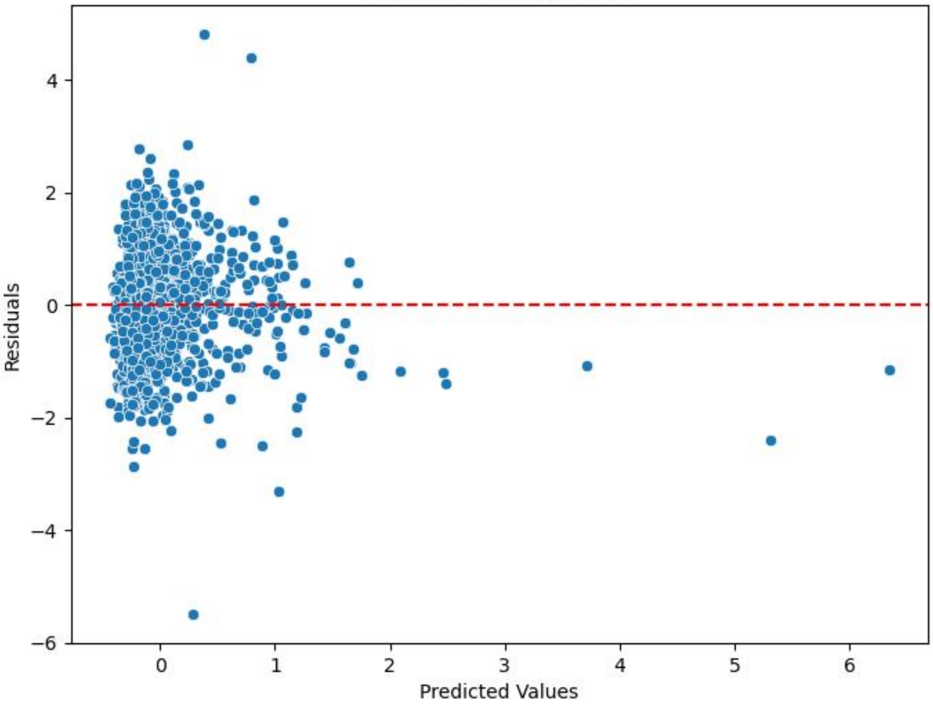
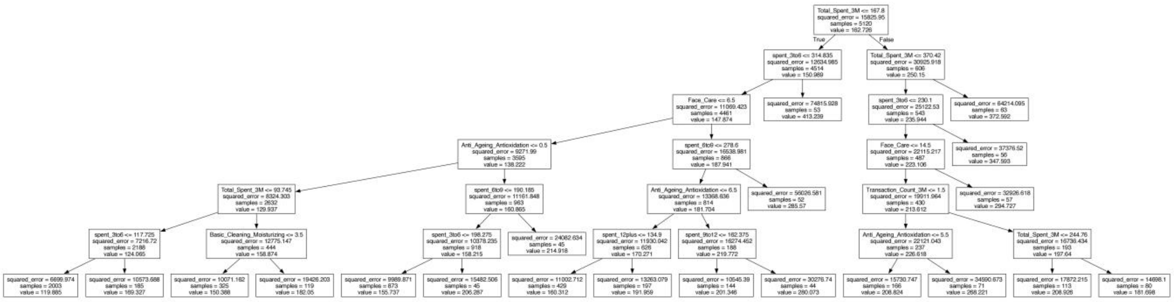


Figure 13: Residual Plot

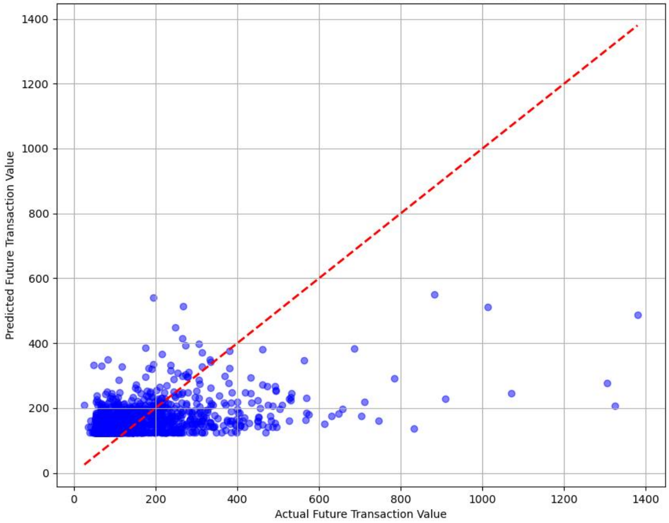


Appendix L



Appendix M

Figure 12: XGBoost Model: Comparsion of Actual and Predicted future transaction value


Predicted Future Transaction Value
Actual Future Transaction Value