QBUS3600 Business Analytics Capstone

(S2 2024)

Table of contents

1. Executive briefing..... 2
1.1 Background..... 2
1.2 Problem Description, challenges, and key questions:..... 2
1.3 Key Findings:..... 3
1.4 Suggestions for improvement:..... 3
1.5 Areas for Further Investigation:..... 3
2. Data preprocessing..... 4
2.1. Data cleaning..... 4
2.2. Data type treatment..... 4
2.3. Data transformation..... 4
2.4. Discrete data and quartile encoding..... 4
3. Descriptive statistics..... 5
3.1. Key spending variables..... 5
3.2. Outliers and Distribution Patterns:..... 5
3.3. Low, medium and high spenders..... 6
3.4. Identifying Key Product characteristics..... 6
3.5. Identifying Key Product Combinations..... 7
3.6. RFM score..... 8
4. Exploring potential relationships..... 8
4.1 Correlations..... 8
4.2 Scatter plot vs Total_Spent_6M..... 10
4.3 Box plot and relationship with brands..... 11
4.4 Box plot and relationship with categories:..... 11
4.5 Box plot and relationship with sub-categories:..... 12
4.6 Box plot and relationship with skin-concerns:..... 12
4.7 Interaction with Spender Category..... 13
4.8 Relationship with RFM..... 14
5. Conclusion..... 15
Customer Spending Patterns..... 15
Brand Loyalty..... 15
Subcategory and Category Trends..... 15
RFM Score as a Predictor..... 15
Skin Concern Influence..... 15
Spender Categories..... 15
Recommendations:..... 16
6. Reference..... 16

1. Executive briefing

1.1 Background

L'Oréal Dermatological Beauty (LDB) operates in a competitive skincare industry, offering products designed to address specific skin concerns such as acne, anti-ageing, and irritation. As a division of L'Oréal, LDB has built a reputation through trusted brands like La Roche-Posay, CeraVe, and SkinCeuticals. However, with increasing consumer preferences for online shopping and personalized skincare solutions, LDB is facing challenges in maximizing its online sales potential while maintaining customer loyalty.

1.2 Problem Description, challenges, and key questions:

The business problem at the core of this investigation involves understanding customer behavior to drive sales growth through effective product bundling, targeted marketing, and customer retention strategies. This is more than a data problem—LDB needs to align its product offerings with customer needs, provide personalized recommendations, and optimize the customer journey.

Key Challenges:

- Shifting Consumer Behavior: Customers are increasingly seeking personalized experiences, expecting brands to provide tailored product recommendations that align with their skin concerns and preferences. Failing to meet these expectations can lead to lost sales and diminished loyalty.
- Sales Growth through Bundling: LDB is looking to increase the average order value by identifying products that are frequently purchased together. Bundling popular products based on customer behavior offers an opportunity to streamline the shopping experience and boost sales.
- Customer Retention: Maintaining long-term customer relationships is critical in the skincare market, where repeat purchases (e.g., moisturizers, serums) drive a significant portion of revenue. Identifying high-value customers and providing personalized offers is crucial to ensuring their continued engagement with the brand.
- Optimizing Marketing Strategies: LDB needs to improve the efficiency of its marketing by targeting the right customers with the right products. Understanding which customers are most likely to respond to promotions, based on their purchasing history and preferences, can help the company reduce marketing costs and improve conversion rates.

Key questions:

- Which product combinations are most frequently purchased together?
- Which customer segments (based on recency, frequency, and monetary spending) should be prioritized for personalized marketing?
- How can LDB use historical transaction data to create effective product bundles that drive sales?
- What do the patterns in customer spending reveal about opportunities for targeted promotions and retention strategies?
- Which brands, categories, subcategories … are driving the most customer spending?

1.3 Key Findings:

- **Popular Product Combinations**: Our analysis shows that certain products are frequently purchased together: (Anthelios, Effaclar, Toleriane), (Anthelios, Cicaplast, Toleriane), and (Hyalu B5, Retinol LRP, Vitamin C)
- **RFM (Recency, Frequency, Monetary) Analysis:** Customers who spent more in the last six months showed a tendency to continue purchasing from the same brands. High-value customers (with high recency, frequency, and spending) are prime targets for personalized promotions and loyalty incentives to drive repeat purchases.

1.4 Suggestions for improvement:

- Introduce Product Bundles: Based on popular combinations, LDB should offer bundled products (e.g., Anthelios, Effaclar, Toleriane) that cater to common skin care needs. This can streamline the purchasing process for customers and increase the average order value.
- Target High-Value Customers: Focus marketing campaigns and exclusive offers on customers identified as high-value based on their RFM scores. Personalized incentives can improve customer retention and boost loyalty.

1.5 Areas for Further Investigation:

- Customer Retention: Further analyze customer segments that have reduced spending to identify patterns and address retention challenges with tailored offers.
- Seasonal Trends: Explore how seasonal factors (e.g., sun care products in summer) impact purchasing behavior and adjust product offerings and promotions accordingly.

2. Data preprocessing

2.1. Data cleaning

The original dataset contains 6,400 entries.

- The dataset only had missing values in postcode column. However, since we are not interested in understanding relationships between post codes and total purchases in the report. Thus, no action was required.
- There was duplicate rows for Category_Face Care. I summed both columns into one and dropped the duplicate rows.

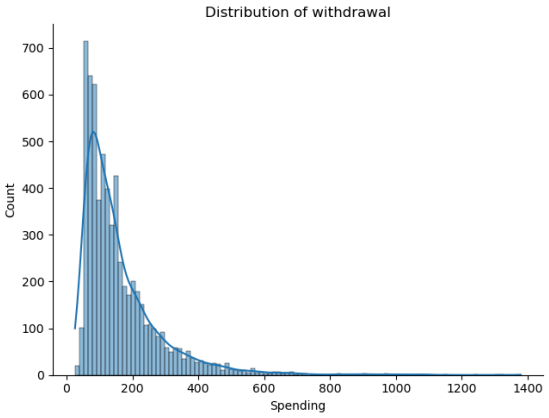
The final data set now contains 6,400 rows and 69 columns.

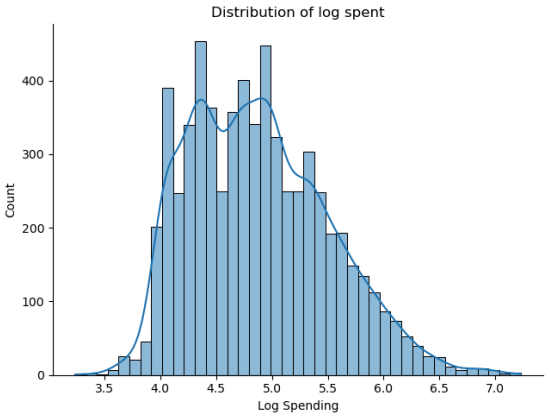
2.2. Data type treatment

The data types of the categorical variables are in binary and numerical variables are in discrete values and seemingly in correct presentation. Thus, there was no need for data type treatment.

2.3. Data transformation

Log transformation on the "Total_Spent_Nov23_May24" exhibits a more symmetric, bell-shaped distribution, reducing skewness and making the data closer to a normal distribution. This transformation mitigates the effect of outliers and makes it easier to apply statistical models that assume normality.





2.4. Discrete data and quartile encoding

The brand, category, subcategory, classes, and skin concern columns in the dataset indicate how often customers purchased specific products before November 2023. To better understand customer behavior, I applied quartile encoding using pd.qcut(), transforming these purchase frequencies into four groups (Q1 to Q4), representing different levels of purchasing activity. This helps with exploring data later on.

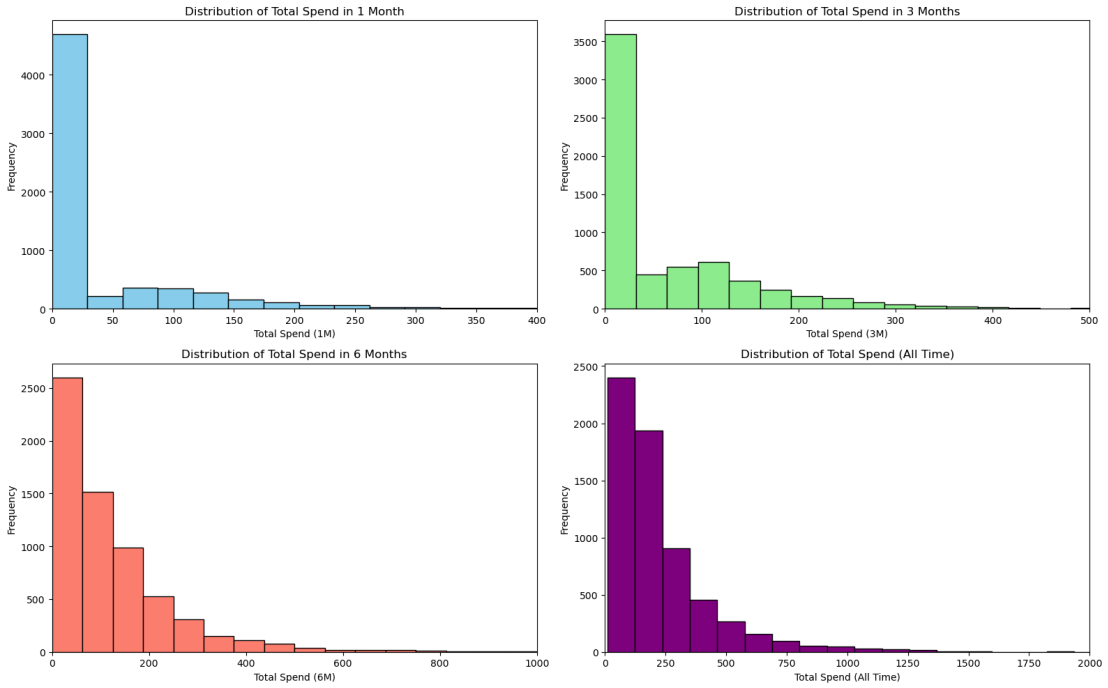
Q1: Bottom 25% (fewest purchases); Q2: Next 25% (moderate purchases); Q3: Higher purchase frequency; Q4: Top 25% (most frequent buyers)

3. Descriptive statistics

3.1. Key spending variables

In this part of the analysis, I generated descriptive statistics for key spending variables, including Total_Spent_Nov23_May24, Total_Spent_1M, Total_Spent_3M, Total_Spent_6M, Total_Spent_9M, Total_Spent_12M, and Total_Spent_AllTime. I also zoomed into the histogram of the key variables.

	Total_Spent_Nov23_May24	Total_Spent_1M	Total_Spent_3M	Total_Spent_6M	Total_Spent_9M	Total_Spent_12M	Total_Spent_AllTime
Skewness	2.87	2.79	2.35	2.68	3.18	3.45	3.62
Kurtosis	13.3	12.3	9.33	15.09	23.49	28.47	26.87



Key Insights:

1. **Skewed Spending Distribution**: All spending variables show positive skewness, indicating that most customers spend less, while a small group of high spenders significantly drives the average up. This suggests the presence of outliers or high-value customers.
2. **High Kurtosis**: The high kurtosis values across all periods indicate that spending is concentrated among a few customers, reinforcing the idea of a small number of outliers who spend considerably more than the rest.

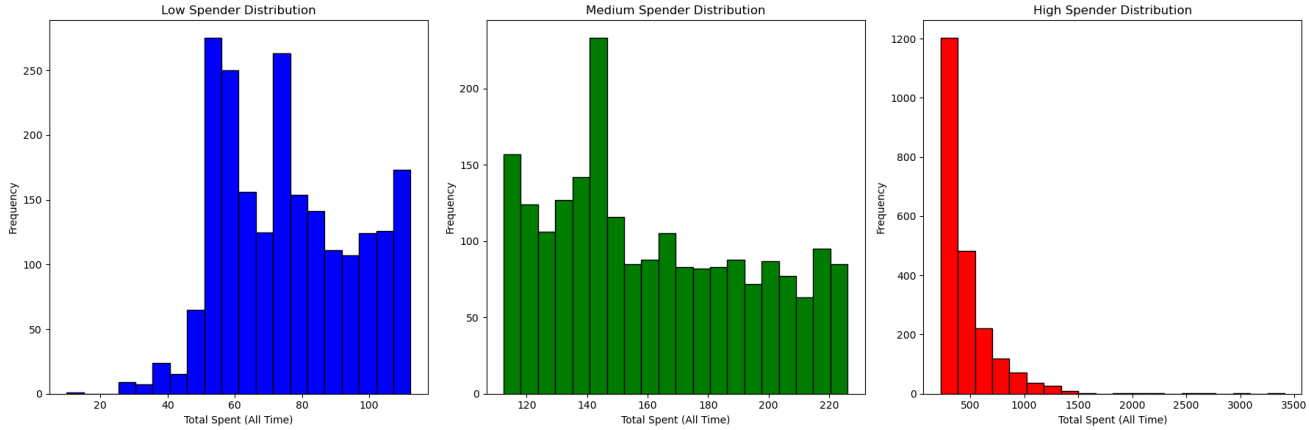
3.2. Outliers and Distribution Patterns:

In the analysis, outliers were identified in the spending data across all periods, as evidenced by the high **skewness** and **kurtosis** values. These outliers represent customers who have significantly higher spending compared to the majority of the customer base. Rather than

omitting these outliers, they were retained in the analysis because they provide valuable insights into high-value customers who contribute disproportionately to overall sales. These customers are likely driving much of the revenue and may require targeted marketing strategies to maintain and enhance their engagement.

3.3. Low, medium and high spenders

Let's look at the distribution between low, medium and high spenders for all time.



Observations:

- Low spenders have a broad spending range but tend to peak around 50–60.
- Medium spenders are more concentrated with a peak around 140, suggesting more uniform behavior.
- High spenders are skewed, with a majority spending around 500 , but a few making much larger purchases, creating a long tail in the distribution.

3.4. Identifying Key Product characteristics

In this analysis, I aimed to identify the products, categories, and skin concerns that drive the most transactions across our customer base. To achieve this, I summed the transaction counts across various product groupings—brands, sub-categories, product classes, and categories. This approach allows us to pinpoint the top-performing areas and provides valuable insights for shaping business strategies.

The brands with the highest transaction counts are as follows:

- Toleriane: 5,715 transactions

- Effaclar: 4,538 transactions

- Anthelios: 3,712 transactions

Sub-Categories:

- Face Moisturiser: 7,121 transactions

- Face Serum: 5,941 transactions

- Sunscreen: 2,616 transactions

Class Descriptions:

- Face Care: 12,274 transactions

- Anti-Ageing: 4,269 transactions

- Anti-Acne: 3,982 transactions

Categories:

- Face Care: 16,910 transactions

- Sun Care: 4,196 transactions

- Body Care: 2,415 transactions

Skin Concerns:

- Irritation-Prone Skin: 9,105 transactions

- Anti-Ageing: 5,370 transactions

- Acne-Prone Skin: 4,977 transactions

3.5. Identifying Key Product Combinations

Objective:

The purpose of this analysis was to identify the top three combinations of brand products that customers frequently purchase together. By analyzing brand combinations, we aim to create product bundles that are likely to sell well and align with customer purchasing behavior.

Methodology:

- Combination Generation: I generated all possible combinations of three items from a selection of key brands.
- Purchase Frequency Calculation: For each combination, I calculated the total number of times all three brands were purchased together by customers.
- Sorting: The combinations were then sorted by their total purchase counts to determine which combinations drive the highest sales.

Results:

The analysis revealed the following top three product combinations based on the highest purchase frequencies:

- Combination 1: [Brand Description_Anthelios, Brand Description_Effaclar, Brand Description_Toleriane] – Purchased together 306 times.
- Combination 2: [Brand Description_Anthelios, Brand Description_Cicaplast, Brand Description_Toleriane] – Purchased together 291 times.
- Combination 3: [Brand Description_Hyalu B5, Brand Description_Retinol LRP, Brand Description_Vitamin C] – Purchased together 289 times.

3.6. RFM score

RFM (Recency, Frequency, Monetary) analysis is a proven customer segmentation method that helps businesses identify their most valuable customers based on their purchasing behaviors (KABASAKAL, 2020).

I calculated the **RFM score** by analyzing three key customer behaviors: recency, frequency, and monetary value. Here's how it was done:

1. **Recency**: I defined a function that assigns higher scores for more recent transactions, with a higher score given to customers with recent activity (within 1 month) and lower scores for older transactions.
2. **Frequency**: I calculated the total transaction count over a 12-month period and ranked customers into quintiles, assigning higher scores to customers with more frequent purchases.
3. **Monetary**: I ranked customers based on their total spending and divided them into quintiles, with higher scores for those who spent more.

Finally, I combined these three metrics (recency, frequency, and monetary) by summing the scores to create a final **RFM Score**. This score helps categorize customers based on their engagement, allowing targeted marketing strategies for different customer segments.

4. Exploring potential relationships

4.1 Correlations

In this analysis, I calculated the correlation matrix to identify which variables are most closely associated with log_Total_Spent_Nov23_May24 (the log-transformed total spending from November 2023 to May 2024). The goal was to uncover potential relationships between this spending variable and other key features, both spending-related and categorical.

The top variables correlated with **log_Total_Spent_Nov23_May24** are:

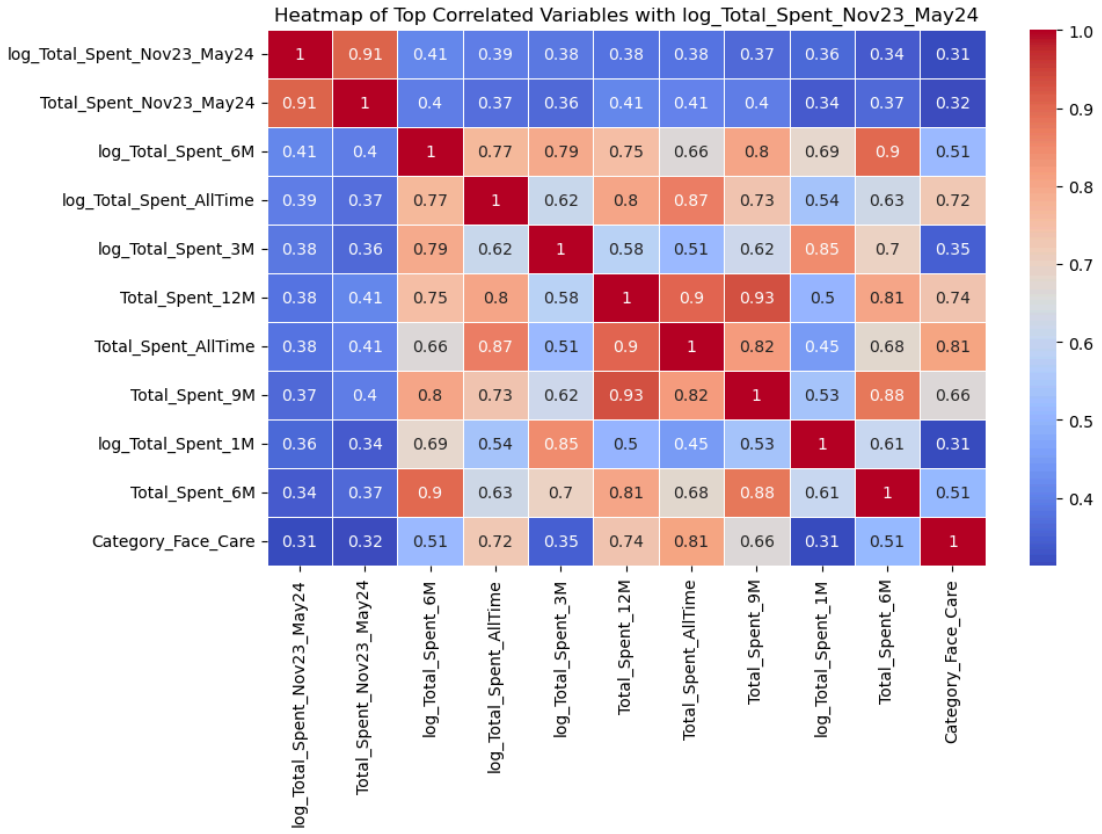
- **log_Total_Spent_6M**: 0.41
- **log_Total_Spent_AllTime**: 0.39
- **log_Total_Spent_3M**: 0.38

Key Insights:

- Spending in recent periods, particularly the last 6 months, has the strongest correlation with spending in the November-May window.
- Cumulative spending over longer periods (12 months, All Time) also shows significant correlation, indicating that more frequent or high spenders tend to consistently spend more across all periods.
- The category Face Care is notable, suggesting that customers who purchase face care products may also be contributing to higher total spending in the specified period.

To better understand the relationships, I created a heatmap visualizing the correlations among the top 10 variables and **log_Total_Spent_Nov23_May24**. This heatmap highlights

how closely related these variables are, helping to identify potential patterns or trends in customer spending behavior.



I also looked the significance of the top variables, the results yielded

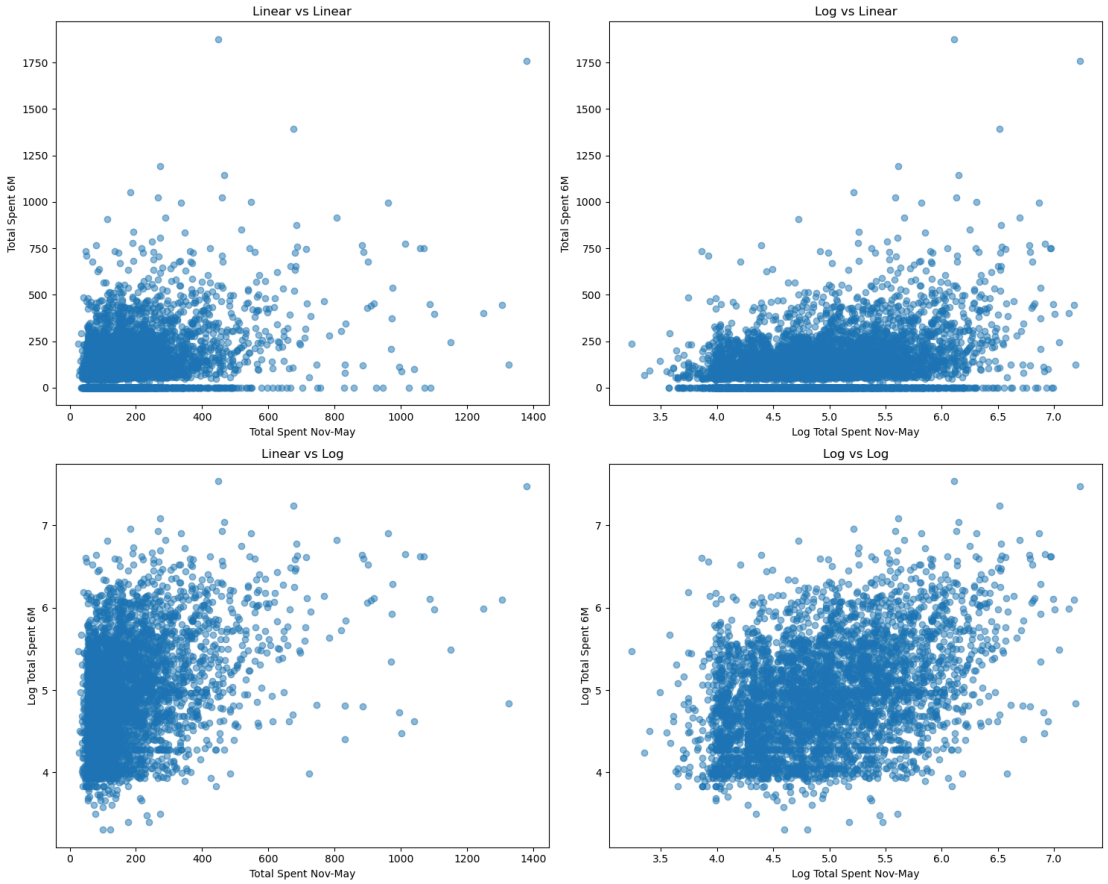
	log_Total_Spent_6M	log_Total_Spent_AllTime	log_Total_Spent_3M	Total_Spent_12M	Total_Spent_AllTime	Total_Spent_9M	log_Total_Spent_1M	Total_Spent_6M	Category_Face_Care
Value	3.231119	4.615683	1.530454	-0.55895	-3.01344	0.775982	0.246212	-3.91011	1.887285

Conclusion for the 10 most correlated variables:

- Significant variables: Constant, log_Total_Spent_6M, log_Total_Spent_3M, and Brand Description_Toleriane.
- Not significant variables: log_Total_Spent_AllTime, Total_Spent_12M, Total_Spent_9M, log_Total_Spent_1M, Transaction_Count_12M, Transaction_Count_9M, and Transaction_Count_6M.

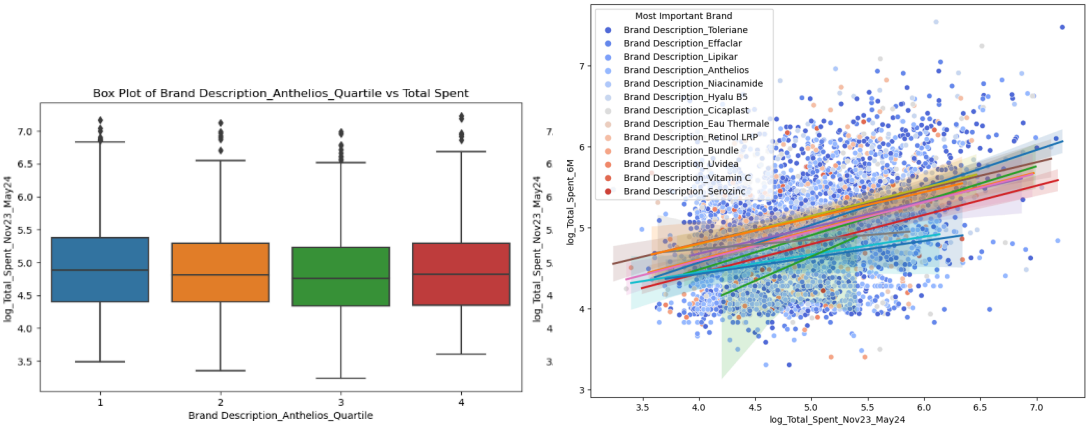
4.2 Scatter plot vs Total_Spent_6M

In the analysis, I focused on log-transformed values of Total_Spent_6M and Total_Spent_Nov23_May24 because their correlations were most significant and strongest.



The **Log vs Log** transformation is the most useful in revealing the underlying relationship between customer spending during different time periods. The log transformation effectively reduces skewness, allowing for a clearer interpretation of trends, and is more suited for building predictive models compared to the raw data.

4.3 Box plot and relationship with brands

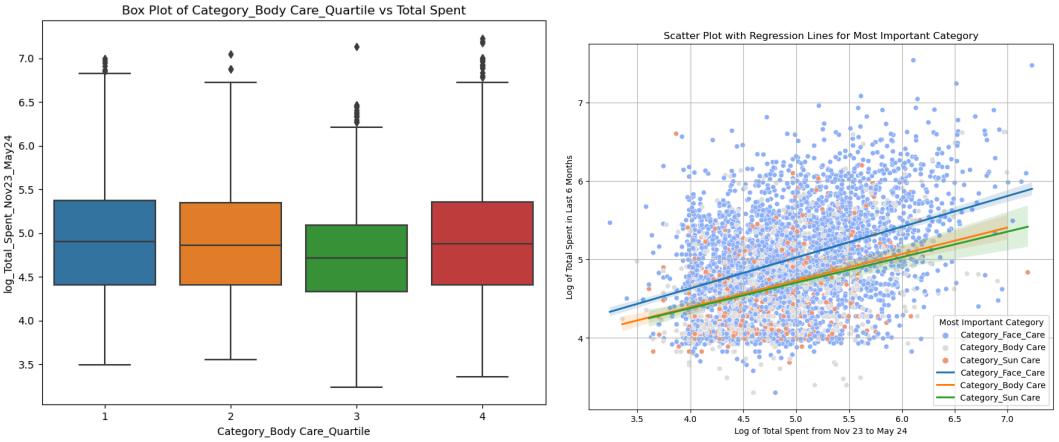


Key observations

Based on the analysis, there appears to be little noticeable differentiation in customer spending patterns across brand quartiles. An example boxplot is given above. The spending behavior for most brands remained stable across different quartiles, suggesting that quartile segmentation, in this case, may not be a strong indicator of variation in customer spending.

The scatter plot visualizes the interaction between log_Total_Spent_Nov23_May24 and log_Total_Spent_6M, segmented by the most important brand for each customer. This indicates the presence of interaction terms between spending behavior and specific brands. The varying slopes for different brands suggest that brand loyalty influences how spending evolves over time.

4.4 Box plot and relationship with categories:

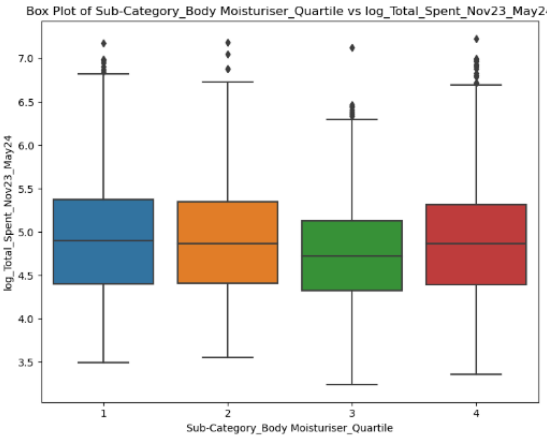


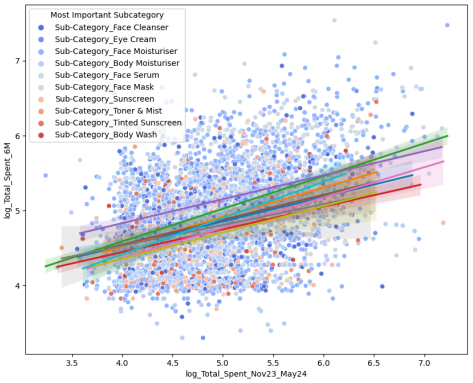
Key observations

The log-transformed spending for Body Care and Face Care remains consistent across quartiles. This further confirms that customers spend relatively similar amounts within these categories, regardless of their quartile. The Sun Care category shows slightly more variation, particularly in the higher quartiles (Q3 and Q4), where log-transformed spending is marginally higher. This suggests that higher quartile customers may spend more on Sun Care, although the difference is still moderate.

This scatter plot illustrates the relationship between log_Total_Spent_Nov23_May24 and log_Total_Spent_6M, segmented by the most important category. Categories such as Face Care show steeper slopes, suggesting a stronger relationship between spending over time. The presence of interaction terms indicates that customers' spending behavior is influenced by the category they prioritize, with Face Care customers showing the most consistent spending across periods.

4.5 Box plot and relationship with sub-categories:



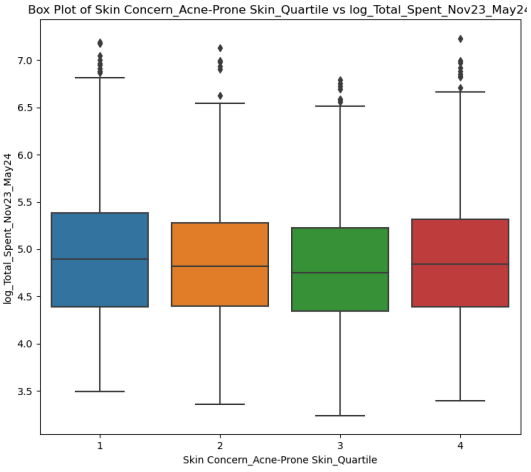


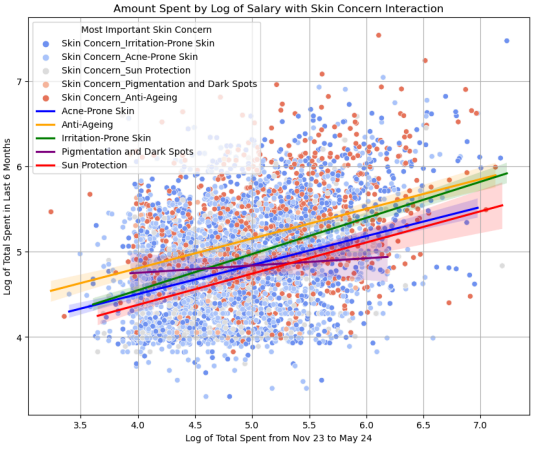
Key observations:

The spending patterns across the sub-categories show only slight variations, with higher quartiles occasionally reflecting marginally higher spending, particularly for facial products and sunscreen. However, overall spending behavior remains relatively consistent across quartiles for most sub-categories.

The scatter plot visualizes the interaction between **log_Total_Spent_Nov23_May24** and **log_Total_Spent_6M**, segmented by the most important subcategory for each customer. The varying slopes for each subcategory suggest that specific product subcategories, such as **Face Moisturiser** and **Sunscreen**, influence spending behavior over time. These variations indicate that customers loyal to certain subcategories tend to exhibit different spending patterns.

4.6 Box plot and relationship with skin-concerns:





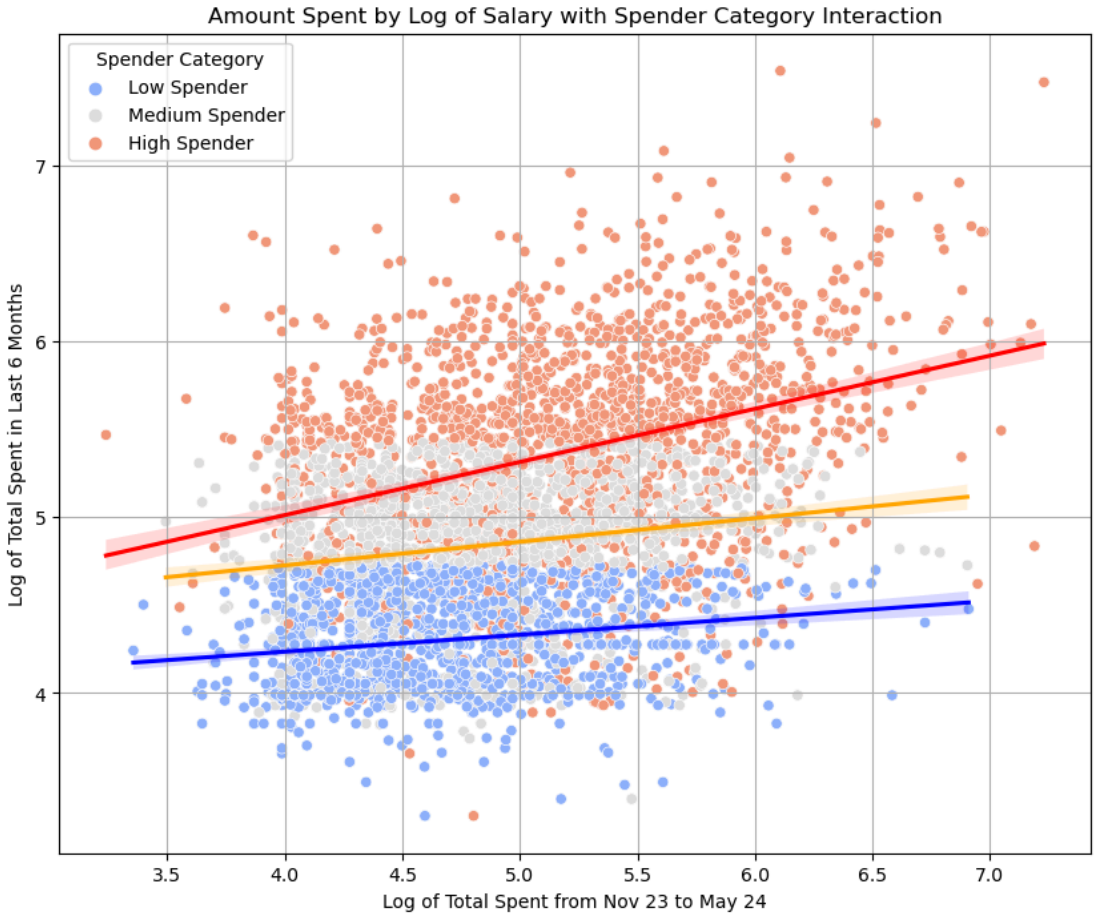
Key observations:

The spending patterns across the skin concerns are generally stable, with only slight increases in higher quartiles for some categories, such as Anti-Ageing and Sun Protection. Overall, quartile segmentation does not seem to have a significant impact on spending behavior in these skin concerns.

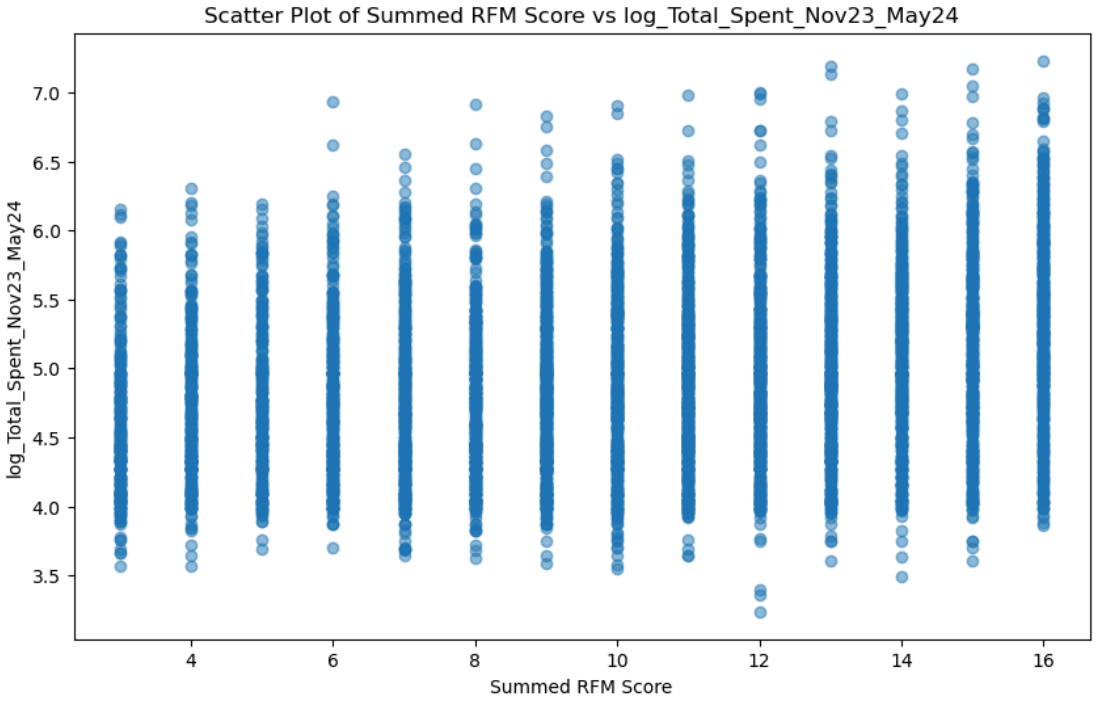
The scatter plot demonstrates the interaction between **log_Total_Spent_Nov23_May24** and **log_Total_Spent_6M**, segmented by the most important skin concern for each customer. The interaction terms suggest that specific concerns, such as **Anti-Ageing** and **Sun Protection**, drive more consistent spending over time, while others like **Acne-Prone Skin** exhibit flatter slopes, indicating weaker spending consistency.

4.7 Interaction with Spender Category

This scatter plot visualizes the interaction between log_Total_Spent_Nov23_May24 and log_Total_Spent_6M, segmented by spender category (Low, Medium, High). The steep slope for High Spenders highlights a strong correlation between spending over time, while Low Spenders have a flatter slope, indicating less consistent spending behavior. The presence of interaction terms shows that spending patterns vary significantly based on the spending category.



4.8 Relationship with RFM



OLS Regression Results

Dep. Variable:	log_Total_Spent_Nov23_May24	R-squared:	0.077
Model:	OLS	Adj. R-squared:	0.077
Method:	Least Squares	F-statistic:	532.0
Date:	Wed, 11 Sep 2024	Prob (F-statistic):	3.86e-113
Time:	15:23:55	Log-Likelihood:	-5781.3
No. Observations:	6400	AIC:	1.157e+04
Df Residuals:	6398	BIC:	1.158e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.4198	0.021	206.861	0.000	4.378	4.462
RFM_Score_Sum	0.0458	0.002	23.065	0.000	0.042	0.050

Key Observations:

Weak Positive Correlation: The calculated correlation between the Summed RFM Score and log_Total_Spent_Nov23_May24 is 0.277, indicating a weak positive correlation. This suggests that higher RFM scores are somewhat associated with higher spending, but the relationship is not very strong.

Scatter Plot Interpretation: The scatter plot shows that as the RFM score increases, there is a slight upward trend in log_Total_Spent_Nov23_May24. However, the spread of points across all RFM score values suggests considerable variability, indicating that the RFM score alone may not fully explain customer spending patterns.

RFM Score as a Predictor: While the RFM score can provide insights into customer behavior, the weak correlation implies that it may need to be combined with other variables or enhanced to improve its predictive power for future spending.

5. Conclusion

The analysis has provided critical insights into customer spending patterns, brand loyalty, and factors influencing repeat purchases for L'Oréal Dermatological Beauty (LDB). Through exploratory data analysis, RFM scoring, correlation analysis, and visualizations, the following key business insights were identified.

Customer Spending Patterns

The correlation between log_Total_Spent_6M and log_Total_Spent_Nov23_May24 revealed that spending in the past six months is the strongest predictor of recent spending. While notable, this relationship shows that historical data alone cannot fully predict future spending, suggesting the need to incorporate other variables such as promotions or customer preferences for improved predictions.

Brand Loyalty

Brands like Toleriane, Effaclar, and Lipikar exhibit stronger customer loyalty, with steeper slopes indicating more consistent spending over time. This highlights the importance of focusing on these brands for marketing campaigns, loyalty programs, and targeted offers to drive increased customer lifetime value (CLV).

Subcategory and Category Trends

Customers who prioritize Face Care and Sun Care products show stronger and more consistent spending patterns. This suggests opportunities for LDB to design targeted marketing campaigns and product bundles within these high-performing categories to boost retention and customer engagement.

RFM Score as a Predictor

The Summed RFM Score and log_Total_Spent_Nov23_May24 showed a weak correlation (0.277), indicating that while RFM is useful for segmentation, it is not sufficient as a standalone predictor of future spending. LDB should incorporate additional data such as browsing behavior and product preferences to improve predictive models.

Skin Concern Influence

Customers focused on Anti-Ageing and Sun Protection concerns exhibit more consistent spending. LDB can capitalize on this by creating skin concern-specific campaigns, launching new products, or offering subscription services to drive higher engagement and retention.

Spender Categories

High Spenders showed the strongest growth in spending over time. LDB should target these customers with exclusive offers and personalized communication. For Low and Medium Spenders, nudges such as discounts or cross-sell opportunities may help boost spending.

Recommendations:

1. Focus on High-Value Brands: Prioritize Toleriane, Effaclar, and Lipikar in marketing efforts to enhance customer loyalty and retention.

2. Target Key Categories: Promote Face Care and Sun Care products through personalized offers and bundling strategies.

3. Expand Predictive Models: Include additional data beyond RFM to improve future spending predictions.

4. Retain High Spenders: Develop exclusive incentives for high spenders and encourage growth in lower-spending segments.

5. Skin Concern Campaigns: Launch targeted campaigns for customers with concerns like Anti-Ageing and Sun Protection to deepen engagement.

6. Reference

KABASAKAL, İ. (2020). Güncellik Sıklık Parasallık Modeline Dayalı Müşteri Bölümlendirme: E-Perakende Sektöründe Bir Uygulama. *Bilişim Teknolojileri Dergisi*, pp.47–56. doi:https://doi.org/10.17671/gazibtd.570866.