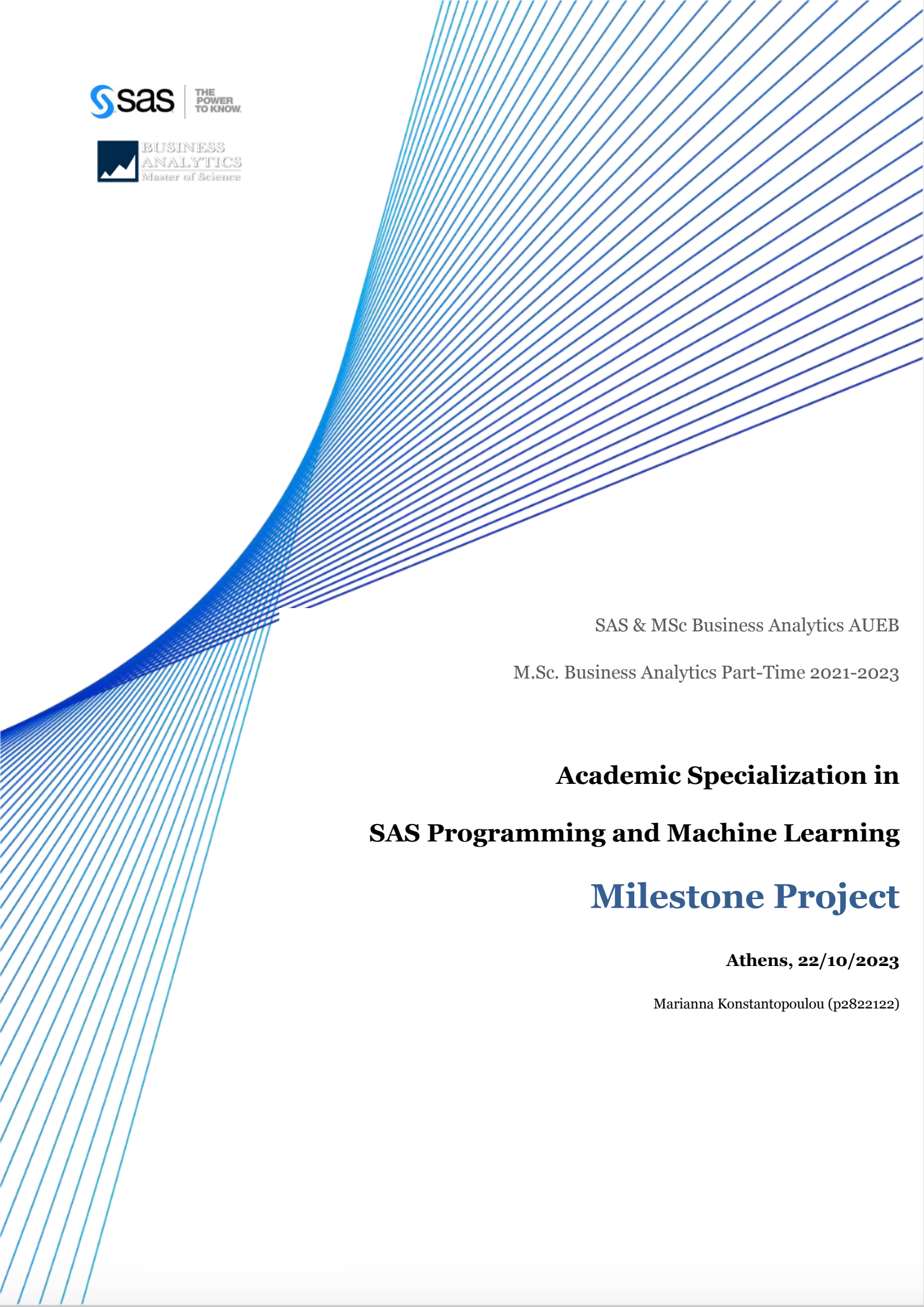


SAS & MSc Business Analytics AUEB

M.Sc. Business Analytics Part-Time 2021-2023

Academic Specialization in

SAS Programming and Machine Learning

Milestone Project

Athens, 22/10/2023

Marianna Konstantopoulou (p2822122)

A. Objective of the project

Executive Summary

In today's fast-changing business world, making decisions backed by data is the key to success. Our project is a deep dive into the world of data analysis, using tools like Base SAS Programming, SAS Visual Analytics, and SAS Visual Data Mining and Machine Learning on SAS Viya. Our goal is to extract valuable insights from the company's Point of Sale (POS) data, helping the stakeholders to make informed decisions for the future of the company.

Objective

The main aim of this project is to thoroughly analyze various datasets containing customer information, invoices, payment methods, promotions, supplier details, and product origins. Our ultimate objective is to provide actionable solutions tailored to meet the needs of the company.

Key Activities

Our project will encompass a series of tasks, each designed to extract valuable insights from the provided data. These tasks include:

1. **Data Pre-processing:** Calculating key metrics for invoices, distinguishing between Sales and Returns, and deriving customer ages from birthdates.
2. **Customer Profiling:** Analyzing customer demographics and categorizing them into age groups. Assessing behavioral characteristics for each group.
3. **Sales Exploration:** Examining sales and returns levels, analyzing basket sizes, identifying top-selling products, and assessing regional revenue contributions.
4. **Promotions Analysis:** Evaluating the impact of promotions on sales and visualizing various promotion types.
5. **Supplier Insights:** Decoding hidden supplier information, generating reports on product sales by supplier, and cross-tabulating revenue by product origin and supplier.
6. **Customer Segmentation (RFM Model):** Segmentation of customers based on Recency, Frequency, and Monetary Value.
7. **SAS Visual Data Mining and Machine Learning:** Clustering customers using the RFM model and describing demographic data of the two most significant customer clusters.

8. **Product Association Analysis:** Identifying associations among product categories and analyzing these associations within the two major customer clusters.

Dataset Details

The datasets at our disposal comprise POS data, encompassing customer information, invoice details, payment methods, promotions, supplier information, and product origins.

Data-Driven Decision Making

Our findings are presented with a business-centric approach, ensuring that the insights provided are clear, actionable, and impactful. Throughout this report, we will not only diagnose the current state of the organization but also prescribe strategies for a future where data drives growth, innovation, and customer satisfaction.

Recommendations

Aligned with our discoveries, we will present practical recommendations that are easily implementable, driving positive changes within the organization. Our analysis not only informs but also seeks to motivate action, paving the way for tangible enhancements in business operations.

The Road Ahead

This project combines data science with business strategy. As data turns into useful information, it can give our retail partner a competitive edge in a complex industry. We believe that the insights we provide will help them succeed and grow with data as their guide.

B. Base SAS Programming Using SAS Studio on SAS Viya

Task 1: Data Pre-processing

In this initial step, we'll prepare the data for analysis. This involves three key actions:

Calculating Key Metrics for Invoices: We'll compute important metrics related to invoices, such as the total number of products (SKU's) included in each invoice and the total monetary value of those products. This helps us understand the size and value of each transaction.

Distinguishing Between Sales and Returns: We'll categorize the transactions as either "Sales" or "Returns" based on their nature. Knowing which transactions returns are crucial for accurate analysis.

Deriving Customer Ages from Birthdates: We'll calculate the age of each customer based on their birthdate. This information is valuable for understanding the customer demographic and tailoring strategies accordingly.

By completing these pre-processing steps, we'll have a clean and organized dataset ready for more in-depth analysis.

Task 2: Customer Profiling

In this section, we aim to get to know our customers better. We'll use data and visual aids to reveal who our customers are in terms of their age, gender, and where they come from. Understanding these basic demographics will help us tailor our products and services to suit their preferences and needs. Let's dive into the data and paint a clear picture of our customer profile.

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	2971	29.71	2971	29.71
M	7029	70.29	10000	100.00

Table 1: Gender frequency table

The average age of the customer base is approximately 38.19 years, with a standard deviation of approximately 11.94 years. The age of customers ranges from a minimum of 19 years to a maximum of 90 years (i.e. Table 1).

In terms of gender distribution, the customer base consists of 29.71% females and 70.29% males (i.e. Figure 1).

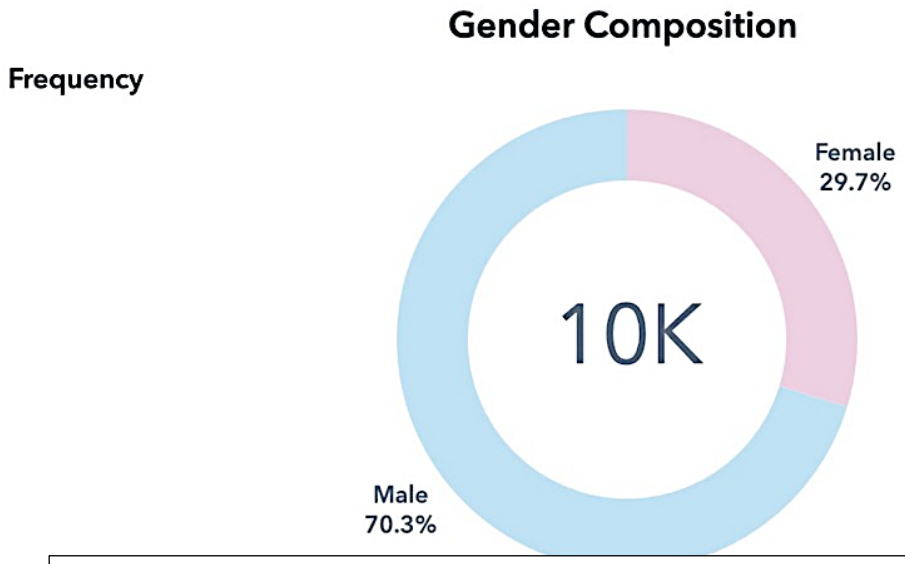


Figure 1: Gender composition of the customer base

Region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AC	11	0.11	11	0.11
AL	76	0.76	87	0.87
AM	27	0.27	114	1.14
AP	10	0.10	124	1.24
BA	407	4.07	531	5.31
CE	222	2.22	753	7.53
DF	284	2.84	1037	10.37
ES	190	1.90	1227	12.27
GO	299	2.99	1526	15.26
MA	106	1.06	1632	16.32
MG	925	9.25	2557	25.57
MS	103	1.03	2660	26.60
MT	115	1.15	2775	27.75
PA	128	1.28	2903	29.03
PB	90	0.90	2993	29.93
PE	243	2.43	3236	32.36
PI	72	0.72	3308	33.08
PR	573	5.73	3881	38.81
RJ	951	9.51	4832	48.32
RN	85	0.85	4917	49.17
RO	45	0.45	4962	49.62
RR	10	0.10	4972	49.72
RS	565	5.65	5537	55.37
SC	354	3.54	5891	58.91
SE	54	0.54	5945	59.45
SP	4005	40.05	9950	99.50
TO	50	0.50	10000	100.00

Table 2: Region frequency table

The customers are spread across various regions, with the highest concentration in São Paulo (SP), accounting for 40.05% of all customers. Other significant regions include Rio de Janeiro (RJ) with 9.51%, Minas Gerais (MG) with 9.25%, Paraná (PR) with 5.73%, and Rio Grande do Sul (RS) with 5.65%. Smaller percentages of customers are distributed across different states, with some regions having as little as 0.10% representation (i.e. Table 2).

The figure below (i.e. Figure 2) presents a snapshot of customer frequencies in various cities (top 5 cities selected based on customer volumes), with São Paulo emerging as the city with the highest customer frequency at 1,916 customers, reflecting its prominent economic and population status. Rio de Janeiro follows with a considerably lower customer count of 349. Brasilia, as the capital, ranks third with 281 customers, showcasing its attraction for both governmental and commercial activities. Belo Horizonte and Salvador exhibit moderate

customer presences with 164 and 140 customers, respectively. While Salvador has the lowest frequency among the listed cities, it still signifies a valuable customer base.

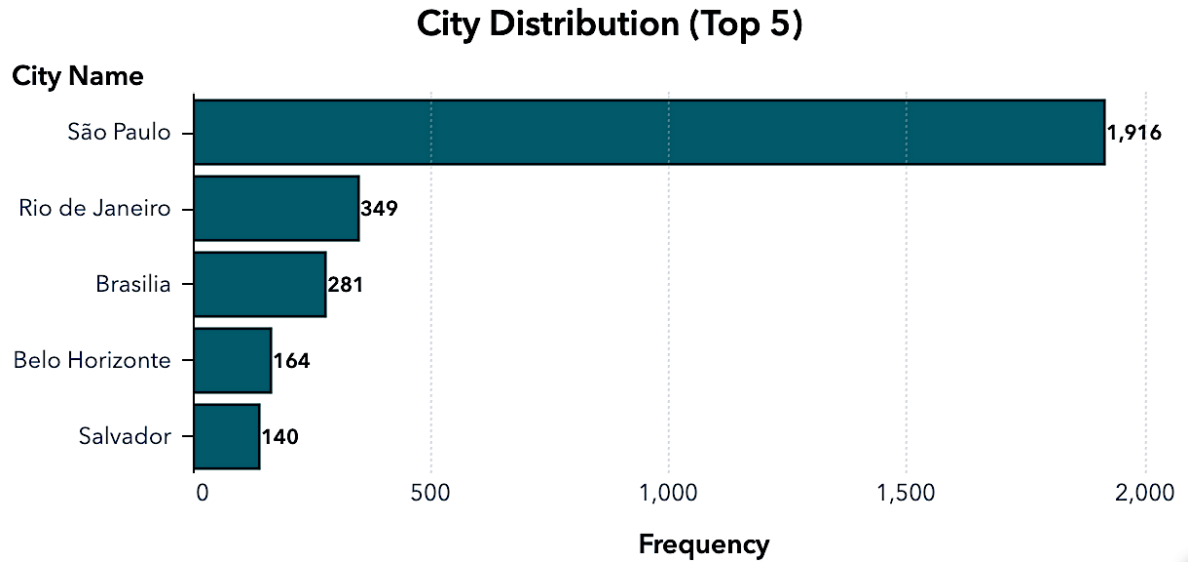


Figure 2: Top 5 customer frequencies by cities

With a sample size of 8,584 customers, the average age is approximately 38.2 years, indicating a relatively mature customer base. The standard deviation of around 11.9 suggests that the age distribution has some variability, but it remains relatively concentrated around the mean age. The age range spans from a minimum of 19 years to a maximum of 90 years, showcasing a diverse customer age group. It's worth noting that the analysis considered customers born between 1910 and 2000, which aligns with the years of birth derived from the data. This filtered age range provides a clearer understanding of the current customer base's age distribution, enabling the company to tailor its marketing strategies and product offerings to cater to the specific preferences and needs of different age segments within this defined demographic (i.e. Table 3).

Age				
Number of customers	**Mean**	**Std Dev**	**Minimum**	**Maximum**
8584	38.19	11.94	19	90

Table 3: Age Summary

The frequency distribution of customers among different age ranges reveals interesting insights into our customer demographics (i.e. Figure 3). The "Middle Age" group emerges as the most prominent segment, with 3,336 individuals, indicating a substantial customer base within this age range. Following closely, the "Young" group comprises 2,560 customers, reflecting a significant presence among our customers. The "Other" category, consisting of

1,416 individuals, represents a diverse range of ages beyond the specified groups, contributing to our overall customer diversity. "Very Young" and "Mature" groups follow with 1,335 and 1,192 customers, respectively, showcasing a balanced distribution across different age brackets. The "Old" and "Very Old" groups, with 132 and 29 customers, respectively, represent a smaller but still notable portion of our customer base.

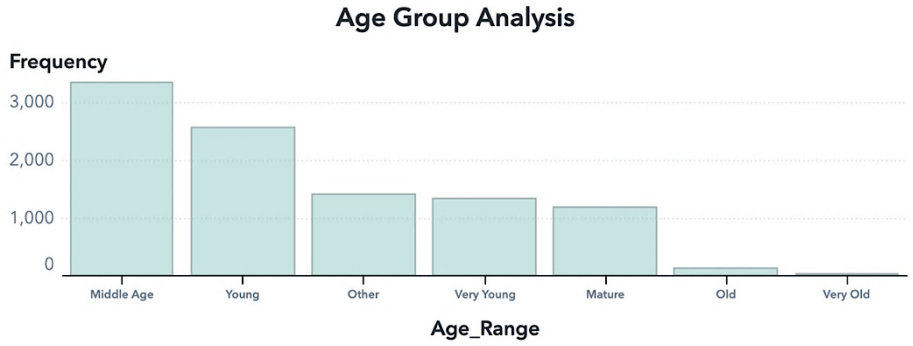


Figure 3: Age Group Analysis

Furthermore, we've evaluated the percentage of customers in each age group to gain a better understanding of our customer demographics (i.e. Figure 4 & Table 4). "Middle Age" customers dominate our customer base, representing 33.3% of all customers. The "Young" segment follows closely, contributing 25.6% to our customer population. Meanwhile, the combined presence of "Under 18" and "Very Young" customers amounts to a significant 27.5%. In contrast, the "Mature" and "Old" groups together comprise 13.2% of our customers, while the "Very Old" segment represents the smallest fraction, accounting for just 0.3% of our customer base.

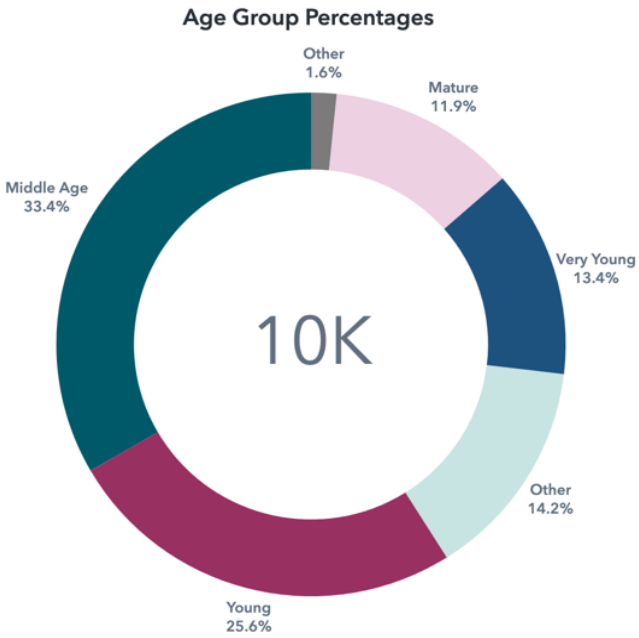


Figure 4: Age Group Percentages

Age_Range	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Mature	51109	12.20	51109	12.20
Middle Age	140220	33.48	191329	45.69
Old	5218	1.25	196547	46.93
Under 18	61678	14.73	258225	61.66
Very Old	924	0.22	259149	61.88
Very Young	54867	13.10	314016	74.99
Young	104755	25.01	418771	100.00

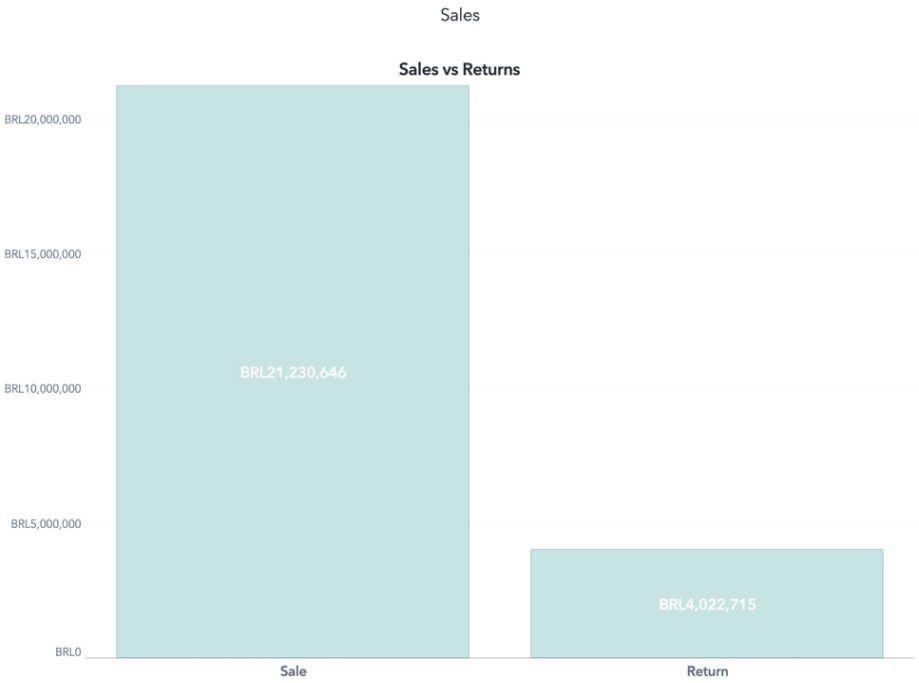
Table 4: Age Range Frequency Table

In our comprehensive analysis of customer behavior across different age groups, several noteworthy patterns emerge. First, when it comes to **visits to our stores**, the "Middle Age" group stands out with an impressive 8,008 visits, indicating a strong presence in our retail spaces. The "Young" group closely follows with 6,178 visits, highlighting their engagement with our brand. Additionally, the "Under 18" and "Very Young" segments display substantial store visitation, with 3,522 and 3,188 visits, respectively, underscoring the appeal of our products to a younger audience. In contrast, the "Mature" and "Old" groups exhibit fewer store visits, with 2,944 and 310 visits, respectively. The "Very Old" segment registers the lowest store visitation, with just 64 visits, reflecting a smaller presence in our stores.

Moving on to the **number of distinct SKUs** purchased, we observe a uniform trend across all age groups, with each group acquiring 116 distinct SKUs. This balanced distribution suggests

that our product offerings cater to a wide range of tastes and preferences, resonating with customers of all ages.

When considering **the total cost of purchases**, the "Middle Age" group emerges as the top contributor, with a substantial expenditure of BRL8,453,632.16. Following closely, the "Young" and "Under 18" groups significantly bolster our total purchase cost, contributing



BRL6,317,337.52 and BRL3,711,277.20, respectively. The "Mature" group also demonstrates substantial purchasing power, with a total expenditure of BRL3,096,217.93. In contrast, other age groups, including "Old", "Very Old" and "Very Young", exhibit comparatively lower total purchase costs.

Task 3: Sales Exploration - Exploring Sales Data and Customer Behavior

In this chapter, we concentrate on understanding the core aspects of sales, including both sales and returns. A detailed breakdown of monetary values is provided, offering insights into the financial transactions. To make these insights more accessible, we created a bar chart displaying these monetary values.

Figure 5: Sales vs Returns Monetary Value

The total sales amount to BRL 21,360,646, while returns account for BRL 4,022,715 (i.e. Figure 5). This data reveals that the company generated substantial sales, indicating a healthy revenue stream. However, it's essential to consider the returns, which constitute a significant portion of the transactions. The presence of returns suggests the importance of assessing the reasons behind these reversals to ensure customer satisfaction and minimize financial impacts.

Subsequently, we dive into an examination of the average basket size, considering metrics like the number of distinct SKUs and total monetary value. Our findings are presented through informative graphs (i.e Figures 6 & 7).

The average number of distinct SKUs per transaction for sales is approximately 352,083, while for returns, it's around 66,688. This suggests that sales transactions typically involve a higher variety of products, indicating that customers are buying a diverse range of items in a single purchase. On the other hand, returns exhibit a lower average SKU count, indicating that return transactions are more focused and involve fewer products.

Additionally, when considering the average basket size by total monetary value, we observe that the average monetary value per return is BRL 60.32, which is slightly higher than the average monetary value for sales, which is BRL 60.30. This implies that while sales transactions involve more SKUs on average, return transactions tend to have slightly higher individual item values.

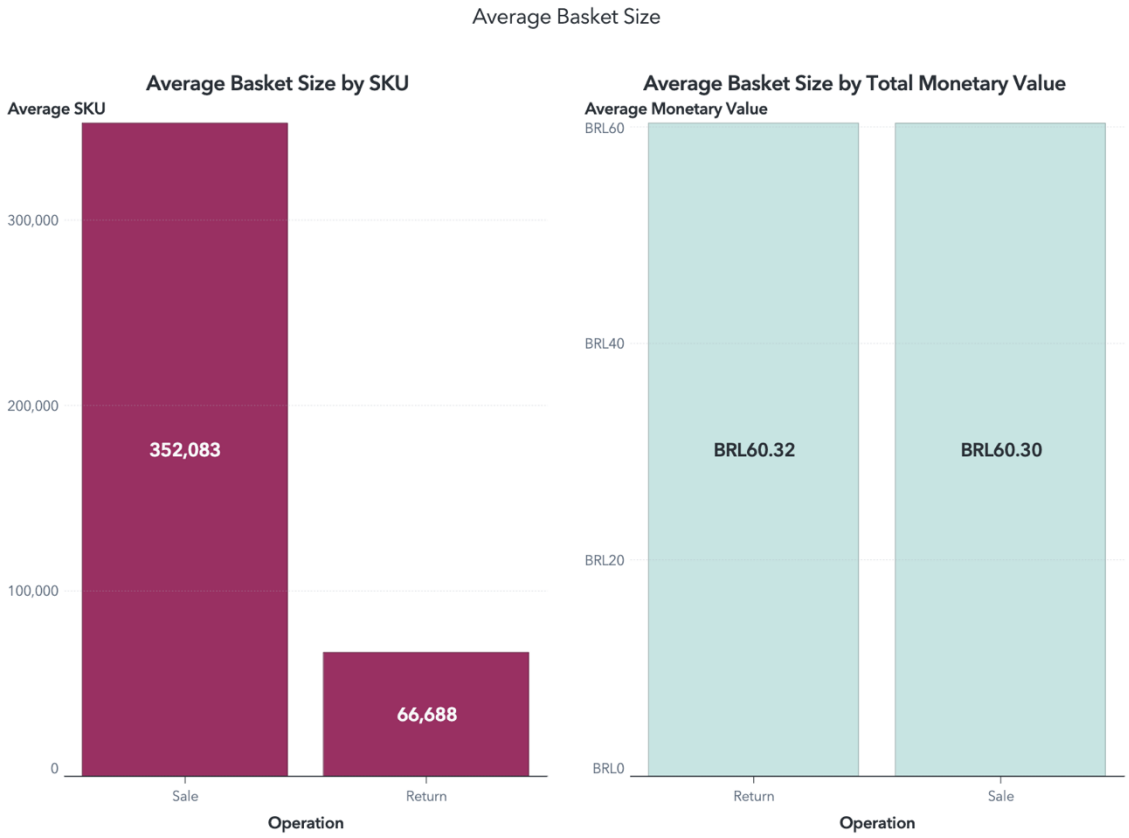


Figure 6: Average Basket Size by SKU & Figure 7: Average Basket Size by Total Monetary Value

Furthermore, our analysis extends to identifying the top-performing products within product lines and types, primarily with respect to sales value. We arrange these products in descending order and also showcase the subtotal sales for each product type. (i.e. Table 5)

Obs	Product_line	Product_type	Product	TotalSales	SubtotalSales
1	Camping Equipment	Cooking Gear	TrailChef Kitchen Kit	197,100	1,524,569
2	Camping Equipment	Cooking Gear	TrailChef Deluxe Cook Set	193,283	1,524,569
3	Camping Equipment	Cooking Gear	TrailChef Utensils	184,154	1,524,569
4	Camping Equipment	Cooking Gear	TrailChef Single Flame	170,669	1,524,569
5	Camping Equipment	Cooking Gear	TrailChef Double Flame	166,405	1,524,569
6	Camping Equipment	Cooking Gear	TrailChef Cook Set	152,060	1,524,569
7	Camping Equipment	Cooking Gear	TrailChef Canteen	146,504	1,524,569
8	Camping Equipment	Cooking Gear	TrailChef Cup	120,361	1,524,569
9	Camping Equipment	Cooking Gear	TrailChef Water Bag	103,537	1,524,569
10	Camping Equipment	Cooking Gear	TrailChef Kettle	90,491	1,524,569

Table 5: Top Selling Products by Product Type and Line

We can observe that the top selling products belong in the "Camping Equipment" product line under the "Cooking Gear" product type, with the sales of each product contributing to the overall subtotal. The top performer, the "TrailChef Kitchen Kit" ,garnered BRL197,100.54 in sales, closely followed by the "TrailChef Deluxe Cook Set" with BRL193,283.01. These two products make substantial contributions to the subtotal sales of BRL1,524,569.89, with other

items in the category also recording respectable sales figures. These findings assist in recognizing the key revenue drivers within this product category.

To offer a comprehensive perspective on the company's performance, we created graphs to illustrate the regional contribution to revenues. Building on this, we focus on the top-performing region identified previously and dissect the revenue contribution by gender within that region.

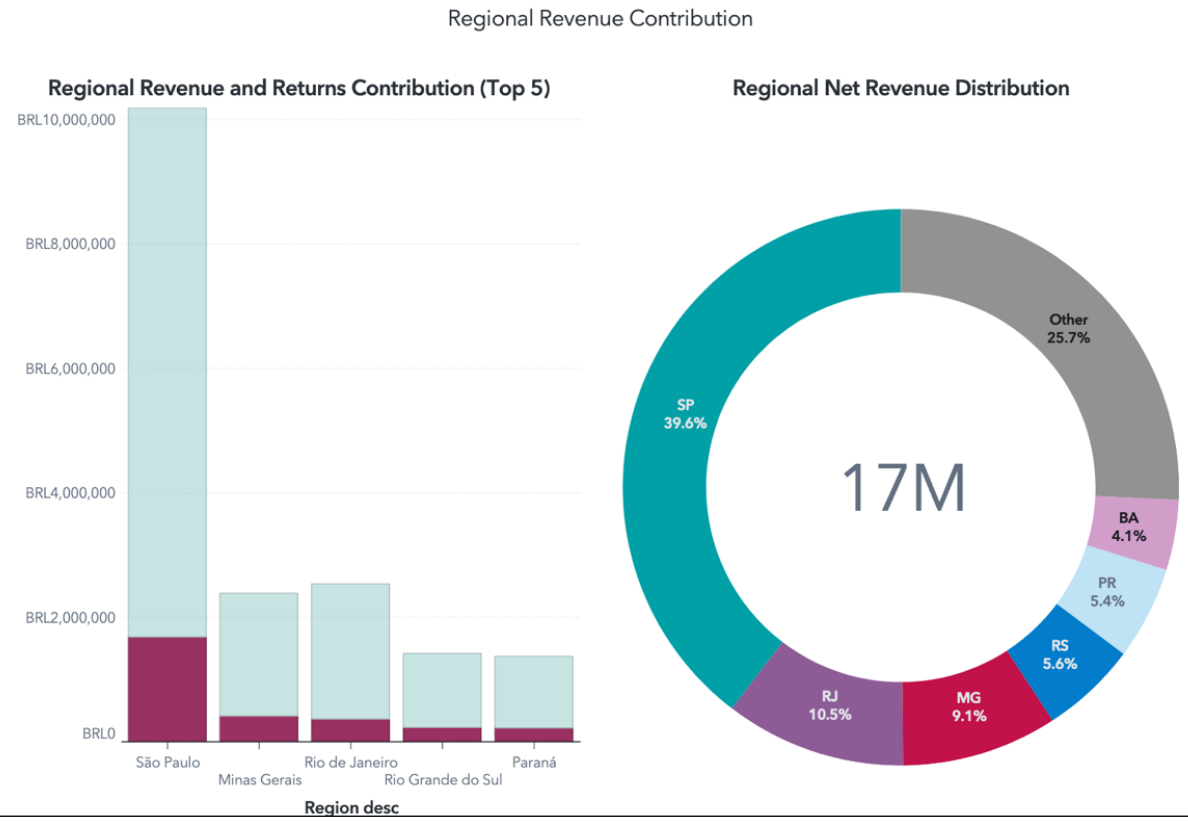


Figure 8: Regional Revenue and Returns contribution & Figure 9: Regional Net Revenue Distribution

According to Figure 8, São Paulo stands out with the highest total sales revenue of BRL 8,496,558, while Minas Gerais and Rio de Janeiro also contribute significantly to sales revenue, with BRL 1,969,412 and BRL 2,170,641, respectively. In terms of total returns revenue, São Paulo again leads with BRL 1,679,560, reflecting both its substantial sales and return figures. The regional net revenue distribution (i.e. Figure 9) further illustrates São Paulo's dominance, accounting for BRL 6,816,998.52 in net revenue.

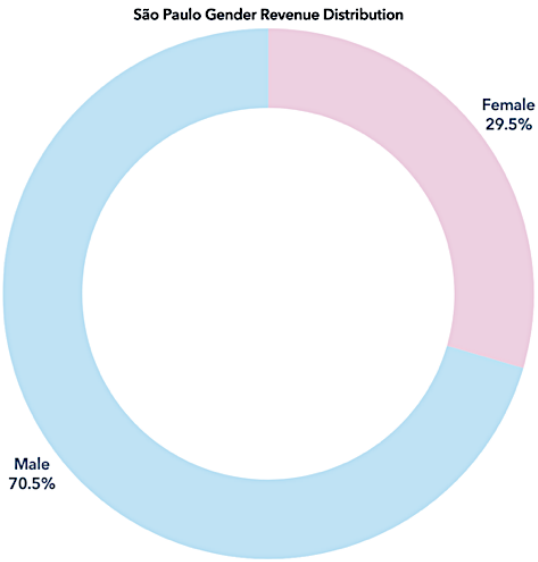


Figure 10: Sao Paolo Gender Revenue Distribution

The gender-based revenue distribution in São Paulo (i.e. Figure 10) indicates a notable disparity in net revenue between males and females. Males contribute significantly more to the net revenue, accounting for a substantial portion of the total. This suggests potential variations in purchasing behaviors or preferences between genders in the São Paulo region, which could be further explored to tailor marketing and sales strategies more effectively.

Task 4: Promotions Analysis - Analyzing Promotional Impact and Sales Patterns

In this upcoming chapter, we delve into a focused analysis of the promotional activities and sales patterns within our dataset. We aim to gain a better understanding of the impact of promotions on product sales and to explore the distribution of sales across different days of the week.

We will show various graphical representations to uncover insights into promotional effectiveness and sales patterns. We'll begin by investigating the percentage of products sold with and without promotion, introducing a clear format to distinguish between them. Additionally, we'll create pie charts to illustrate the distribution of products across different promotion types.

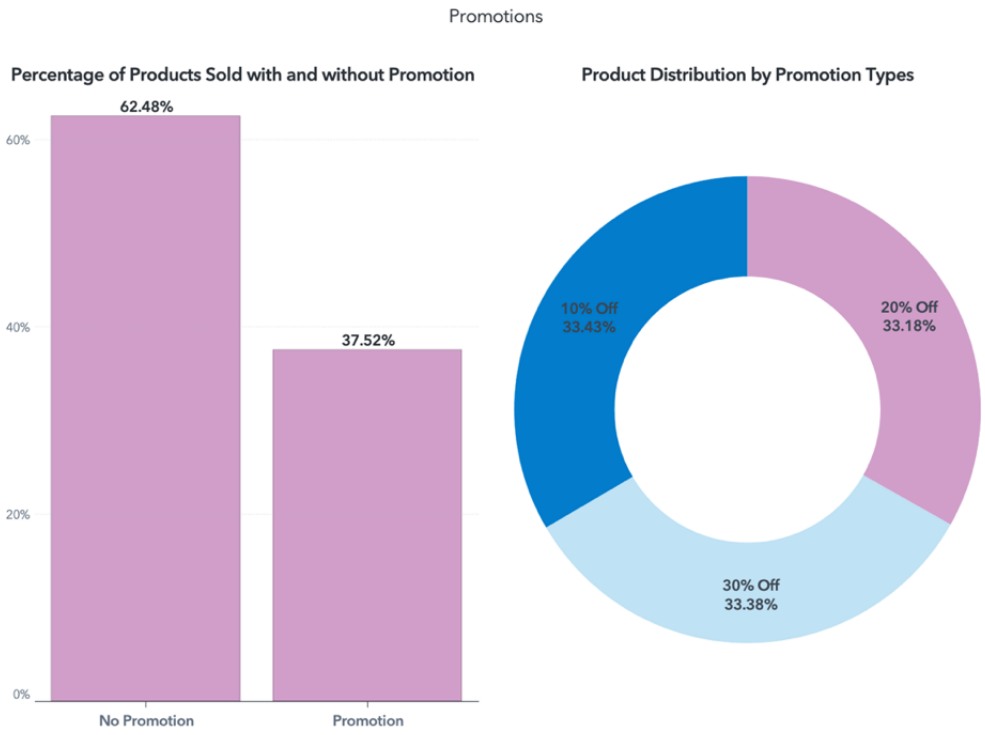


Figure 11: % of Products sold with and without Promotion & Figure 12: Product Distribution by Promotion Types

Finally, we'll explore the distribution of sales per day of the week, paying particular attention to any disparities in the number of distinct SKU's per invoice.

The distribution of sales per day of the week (i.e. Table 6) indicates notable variations in sales activity, with the highest total sales observed on Day 5, presumably a weekday. Days 3 and 4 also exhibited substantial sales, whereas Day 1 recorded the lowest sales. Simultaneously, the number of distinct SKUs per invoice showed a similar trend (i.e. Table 7), with the highest variety of products per invoice occurring on Day 5, and a lower variety on Day 1. This suggests a positive correlation between higher sales and a more extensive range of distinct SKUs available. The findings also imply that weekdays witness increased

sales and product diversity compared to weekends, potentially reflecting differing shopping behaviors. Further analysis could explore the significance of these correlations and provide insights into optimizing sales strategies.

Day of the Week	Total Sales
1	2,423,331
2	4,056,717
3	4,538,908
4	4,633,102
5	5,643,366
6	3,957,934

Table 6: Total Sales by day of the week

Day of the Week	Number Of SKUs
1	40,210
2	67,377
3	75,239
4	77,064
5	93,383
6	65,498

Table 7: Number of SKUs sold by day of the week

Task 5: Unlocking Supplier Insights

In this section, we conducted a comprehensive analysis of supplier-related insights. First, we created a frequency report and an accompanying chart to illustrate the percentage of products sold by each supplier, using supplier names instead of their codes. To provide a more accurate representation, we weighted the frequency of SKUs by the quantity sold, which allowed us to identify the supplier with the highest demand.

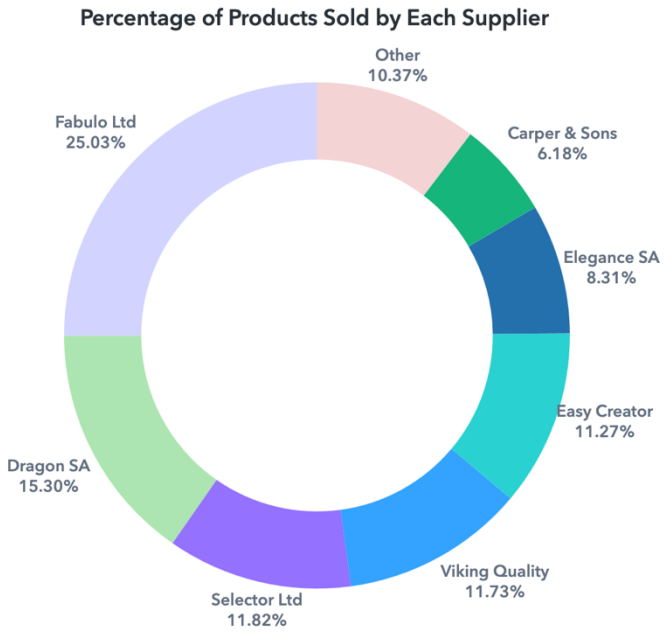


Figure 13: Percentage of Products Sold by Each Supplier

Products		
Supplier Name	**Frequency**	**Percent**
Carper & Sons	51,754	6.18 %
Dragon SA	128,023	15.30 %
Easy Creator	94,278	11.27 %
Elegance SA	69,514	8.31 %
Fabulo Ltd	209,466	25.03 %
Maestri & Maestri	46,423	5.55 %
Selector Ltd	98,899	11.82 %
Toktai & Chen	40,351	4.82 %
Viking Quality	98,178	11.73 %

Table 8: Percentage of products sold by each supplier

As shown in the graph above (i.e. Figure 13) and the frequency table (i.e. Table 8), "Fabulo Ltd" leads with 25.03% of the total sales, followed by "Dragon SA" at 15.30%. "Selector Ltd",

"Viking Quality" and "Easy Creator" also make substantial contributions, each accounting for over 10% of the total sales. This breakdown allows us to identify the suppliers with the highest market demand and understand their relative significance in the overall sales landscape.

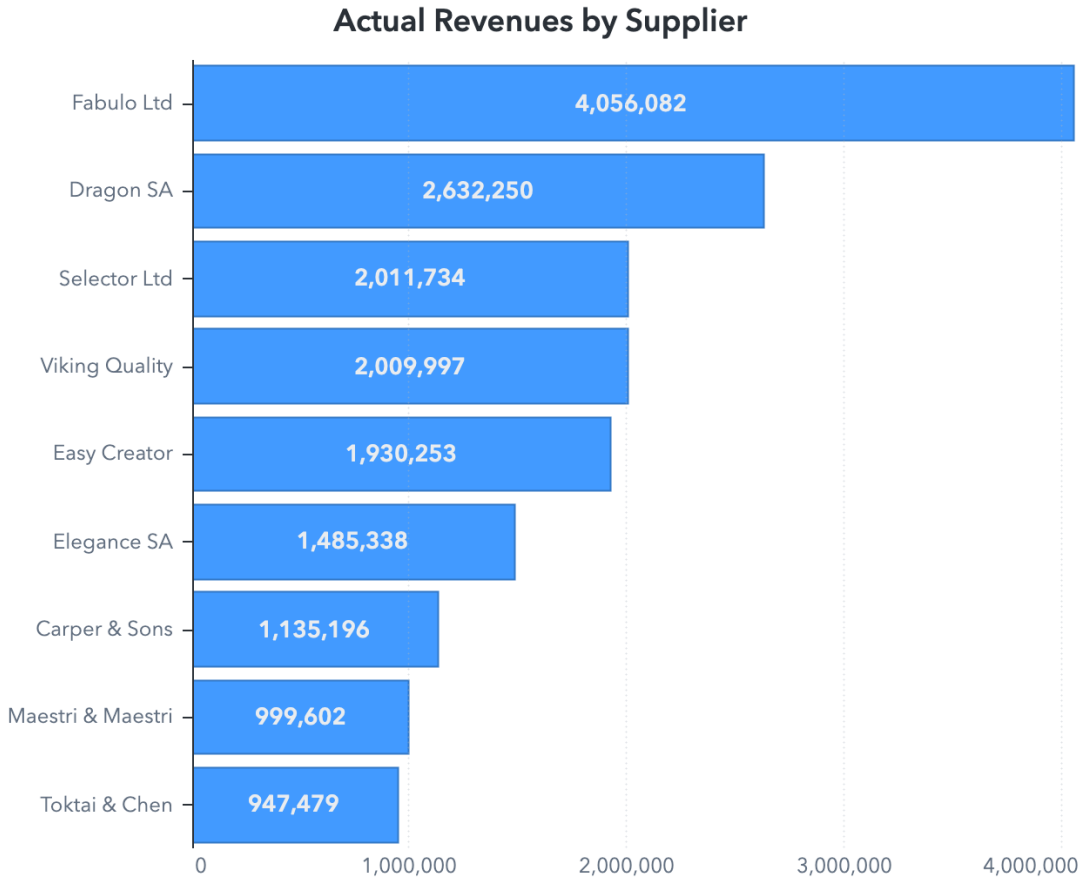


Figure 14: Actual Revenues by Supplier

The findings of Figure 14 reveal the varying contributions of different suppliers to the company's net revenue. "Fabulo Ltd" stands out as the top contributor with net revenue totaling $4,056,082, showcasing a strong presence in the market. Dragon SA", "Selector Ltd" and "Viking Quality" follow closely, each generating substantial revenue figures, indicating their significance to the company's financial performance. The list continues with "Easy Creator", "Elegance SA", "Carper & Sons" and "Maestri & Maestri" demonstrating their respective shares in the company's revenue. Additionally, "Toktai & Chen" presents a notable contribution, highlighting the diverse range of suppliers making a meaningful impact on the company's overall financial success.

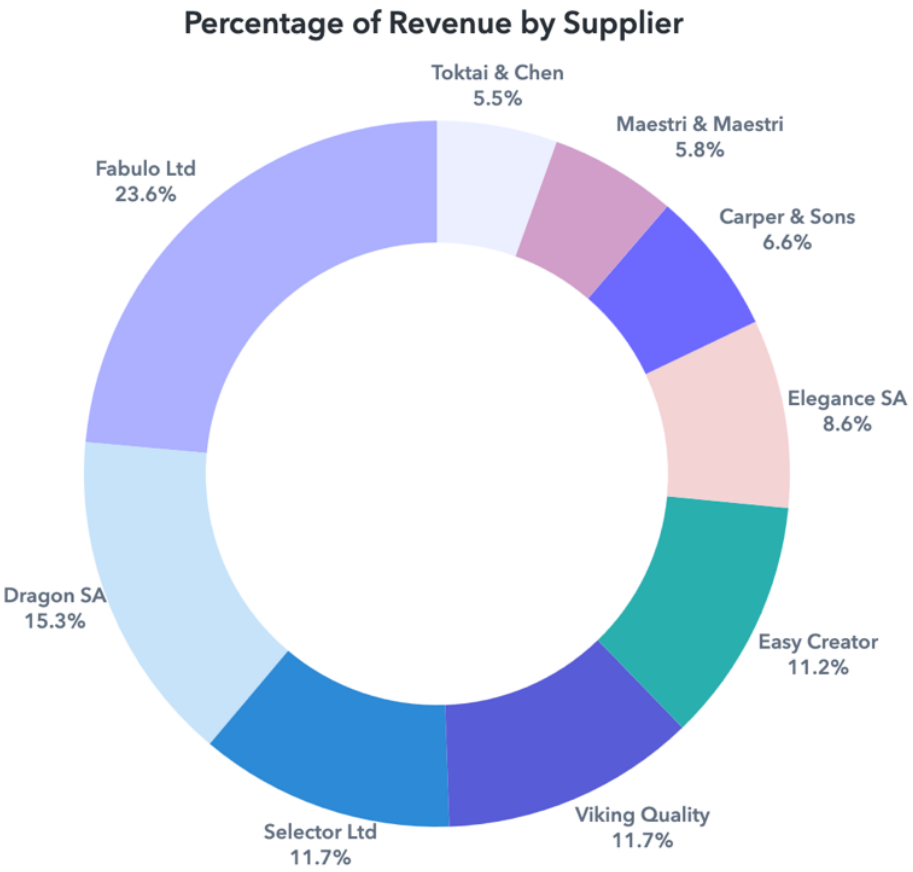


Figure 15: Percentage of Revenue by Supplier

Regarding the percentage of revenue by supplier as shown in the graph 15 above, "Fabulo Ltd" stands out with a substantial 23.57% share of the total revenue, closely followed by "Dragon SA" at 15.30%. The suppliers "Selector Ltd" ,"Viking Quality" and "Easy Creator" also hold noteworthy positions, each contributing more than 10% to the overall revenue.

Country	Supplier Name								
	Carper & Sons	Dragon SA	Easy Creator	Elegance SA	Fabulo Ltd	Maestri & Maestri	Selector Ltd	Toktai & Chen	Viking Quality
	Total Revenue								
China	286,615	871,595	565,722	250,063	845,420	.	96,482	106,667	403,410
India	367,120	126,392	264,928	469,329	1,268,515	419,028	454,953	208,157	91,426
Spain	189,337	780,777	360,808	445,684	546,566	117,643	381,363	356,445	275,016
Turkey	186,192	220,978	738,793	.	655,218	357,738	286,441	80,401	463,115
US	105,931	632,505	.	320,260	740,360	105,191	792,493	195,807	777,027

Table 9: Total Revenue by Supplier and Country

The cross-tabulation table above (i.e. Table 9) provides an overview of the total revenue generated by various suppliers across different countries. It effectively summarizes the financial contributions of suppliers and their respective countries of origin. Notable observations include the substantial revenue generated by "Dragon SA" in China and "Fabulo Ltd" in India. The revenue distribution reflects varying market demand and supplier performance.

Task 6 & 7: Customer Segmentation Using the RFM Model

Our company aims to better understand our customers and provide personalized services and products. To achieve this, we are using a model called RFM, which looks at three things: how recently customers made purchases, how often they buy, and how much they spend. To start, we need to create an RFM dataset using Proc SQL and some functions. We'll use this dataset to calculate the Recency, Frequency, and Monetary Value for each customer. The Monetary Value is calculated from information about prices, quantities, and promotions. This groundwork will help us segment our customers effectively and serve them better.

Cluster ID	Segment Names	Frequency	Frequency Percent	R	F	M
1	Low Value Customers	2,193	24.07%	33.84	1.63	BRL1,611.62
2	New Customers	4,076	44.74%	8.06	2.12	BRL1,463.98
3	High Value Customers	2,841	31.19%	7.32	4.22	BRL5,537.84
Total		9,110	100.00%	14.03	2.66	BRL2,769.98

Table 10: Customer Segmentation with RFM

In this customer segmentation analysis, we've categorized customers into three distinct segments. The first segment, "Low Value Customers", represents 24.07% of the customer base and is characterized by relatively lower recency (R), low frequency (F), and modest monetary spending (M), with an average of BRL1,611.62 in purchases. The second segment, "New Customers", constitutes the largest portion at 44.74%, indicating that many customers are relatively new to the business, with a slightly higher recency (R), moderate frequency (F), and average spending of BRL1,463.98. The third segment, "High Value Customers" includes 31.19% of the customers who exhibit high recency (R), frequent purchases (F), and significant monetary spending (M), with an impressive average of BRL5,537.84. The summary at the bottom reflects that these segments collectively represent the entire customer base, with an average R value of 14.03, an average F value of 2.66, and an average monetary spending of BRL2,769.98.

Task 8: Product Association Analysis

The company's endeavor to transform the store layout based on product associations is a strategic move aimed at boosting sales and enriching the overall shopping experience for customers. To achieve this goal, we followed a structured approach:

Analysis of the Whole Data Set: We commenced with an initial examination of the complete data set to discern patterns of product associations across all customers. This step served as a foundational understanding of broader product relationships that transcend customer segments.

Identification of Key Customer Clusters: Subsequently, we pinpointed two significant customer clusters from the RFM Analysis, which were chosen for their pivotal role in the company's sales and strategic objectives.

Creation of Cluster-Specific Data Sets: In the quest to effectively target customers within these vital clusters, we meticulously filtered the data to create two data sets one for the New Customers and one for the High Value customers. These data sets exclusively contained information pertaining to customers belonging to the specified clusters, ensuring a focused analysis.

Association Rule Analysis: For each of the filtered data sets that represented the identified clusters, we harnessed association rule analysis. This analytical technique unearthed specific associations among product categories within these clusters.

Customized Proposals and Offers: The association rules derived from the analysis provided valuable insights into the product categories that are frequently co-purchased by customers within each cluster. Armed with these insights, the company can now curate proposals and offers tailored precisely to the preferences and buying behavior of each cluster. Customers identified within these clusters can expect to receive personalized product recommendations that align with their interests and preferences.

For example, within the New Customers cluster, our analysis revealed that "Blue Steel Max Putter & Hibernator Extreme" is commonly purchased together with "Compact Relief Kit". More specifically the customers that have bought "Blue Steel Max Putter & Hibernator Extreme" are 4.67 times more likely to buy the "Compact Relief Kit" than a customer that hasn’t bought the first one.

Similarly, for the High-Value Customers Group, the association rules unveiled that "Aloe Relief & Astro Pilot" are frequently associated with "Lady Hailstorm Steel Iron" and "TrailChef

Single Flame". The customers that have bought "Aloe Relief & Astro Pilot" are 3.27 times more likely to buy the "Lady Hailstorm Steel Iron" and "TrailChef Single Flame" than a customer that hasn't bought the first one. These findings can inform the company's strategy for crafting offers that resonate with the preferences of this discerning group.

Furthermore, our analysis of all customers identified a noteworthy association, where "Double Edge" is connected with "Lady Hailstorm Titanium I & Trendi". The lift value of 3.20 underscores the relevance of this association and suggests the potential for a bundled product offering.

In essence, the company's initiative to optimize the store layout and tailor its offers are being guided by data-driven insights. By comprehending customer behavior within distinct clusters, the company is better positioned to make well-informed decisions and maximize the impact of these strategic initiatives. Ultimately, this data-driven approach is poised to elevate the shopping experience and drive sales.

Recommendations

Guided by our data-driven insights, we've identified key areas where the company can make strategic improvements. These recommendations are tailored to be practical and readily implementable, with the aim of driving positive changes within the organization.

Customer-Centric Strategies: Our analysis of customer demographics highlights that the "Middle Age" group, representing 33.3% of all customers, has the highest average purchase value, indicating their significance for your business. This age group presents an opportunity for customized product offerings and marketing campaigns to align with their preferences and needs.

Return Analysis: Returns are a substantial component of transactions, with an overall value of BRL 4,022,715. While their presence is significant, our analysis reveals that "New Customers" exhibit the highest return activity. This insight suggests the need for a targeted approach to address the specific reasons behind these returns.

Sales Optimization: Our exploration of sales activity across days of the week uncovers variations. Higher sales are recorded on weekdays, with Day 5 being the peak. These findings indicate the importance of aligning staffing and inventory levels with this sales pattern. Furthermore, the customers on Day 5 purchased the highest number of distinct SKUs, indicating opportunities for upselling or offering complementary products. Implementing this knowledge in your sales and inventory strategies can enhance customer experiences and potentially increase sales.

Promotional Effectiveness: The analysis of promotional effectiveness demonstrates that promotions significantly impact sales, with promotion types resulting in substantial sales increases. Conclusions drawn from the data indicate that the "Price-Off" promotion type, for instance, leads to notable sales boosts. Exploring which specific products or categories perform well during these promotions can inform future marketing and promotion decisions.

Supplier Management: Our supplier data analysis uncovers key contributors to your net revenue, with "Fabulo Ltd" and "Dragon SA" taking the lead. These insights recommend strengthening partnerships with these suppliers. "Fabulo Ltd" stands out as a dominant player, accounting for 25.03% of total sales. Collaborating with this supplier to develop exclusive product offerings and marketing initiatives could result in significant sales growth.

Customer Segmentation: The RFM model categorizes customers into three segments, with "High Value Customers" (31.19% of the customer base) emerging as the top spenders, averaging BRL5,537.84. To express our appreciation for their loyalty, we recommend

offering them exclusive loyalty cards as a gesture of gratitude. To re-engage the "Low value" customers, we suggest sending them personalized product recommendations tailored to their potential interests. Finally for the "New Customers" category, we propose sending them special discounted promotions as an incentive to continue their engagement with our brand and to encourage them to make repeated purchases. Conclusions based on this analysis suggest that focusing on this segment with tailored loyalty programs or exclusive benefits can nurture their loyalty and drive higher sales.

Product Association Strategies:

Our product association analysis has uncovered robust relationships between various product categories, and these insights are pivotal for reshaping the store layout and driving strategic merchandising decisions. By understanding these associations, the company can transform its store layout to enhance the shopping experience and maximize sales.

For instance, our analysis highlights the exceptional performance of products in the "Camping Equipment" category within the "Cooking Gear" product type. Recognizing such strong product associations paves the way for tailored in-store strategies. The company can strategically position related items in close proximity to one another within the store. By doing so, customers are more likely to encounter and purchase complementary products, thereby increasing the average basket size.

In the context of the store layout, the strategy of promoting cross-category sales and bundling related items takes on a tangible form. The placement of associated products in strategic areas of the store can drive cross-selling and offer customers a more cohesive shopping experience.

These findings contribute directly to the store layout transformation initiative. As the company redesigns the store based on these product associations, it can create store sections or displays that cater to customers looking for related items. This approach not only enhances the shopping experience but also encourages customers to explore complementary products within the same category, resulting in increased sales and customer satisfaction.

The Road Ahead

In the world where data meets business strategy, our project paves the way ahead. As data transforms into practical insights, we see how it can give our retail partner a competitive edge in their complex industry. We're confident that the valuable insights we provide will be the compass guiding their success and growth in the times to come.

Appendix

SAS Code

```
/******1.Data pre - processing******/
/*1*/
proc sort data=project.basket out=project.basket_sorted;
        by invoice_id; run;
proc sort data=project.invoice out=project.invoice_sorted;
        by invoice_id; run;

data project.basket_invoice;
        merge project.basket_sorted (in=b) project.invoice_sorted (in=i);
        by invoice_id;
        if b=1 and i=1;
run;

proc sort data=project.products out=project.product_sorted;
        by product_id; run;
proc sort data=project.basket out=project.basket_sorted_pid;
        by product_id; run;

data project.basket_product;
        merge project.basket_sorted_pid (in=b) project.product_sorted (in=i);
        by prodduct_id;
        if b=1 and i=1;
run;

/* Calculate the number of SKUs per invoice */
proc sql noprint;
   create table project.invoice_total_items as
   select invoice_id,
                count(distinct SKU) as Invoice_total_items
   from project.basket_product
   group by invoice_id;
quit;

/* Print the first 10 observations */
proc print data=project.invoice_total_items (obs=10);
run;

/*2*/
proc sort data=project.promotions out=project.promotions_sorted;
        by promotion_id; run;
proc sort data=project.basket out=project.basket_sorted_promid;
        by promotion_id; run;

data project.basket_promotions;
        merge project.basket_sorted_promid (in=b) project.promotions_sorted (in=i);
        by promotion_id;
        if b=1 and i=1;
run;

proc sort data=project.basket_promotions out=project.basket_promotions_sorted;
        by product_id; run;

data project.basket_promotions_products;
        merge project.basket_promotions_sorted (in=b) project.product_sorted (in=i);
        by product_id;
        if b=1 and i=1;
run;

/****** Calculate the total value of SKUs per invoice ******/
data project.basket_promotions_products_total;
   set project.basket_promotions_products;
```

```
    Invoice_value = product_price * quantity * (1 - promotion);
run;

/****** Sort the dataset by Invoice_ID ******/
proc sort data=project.basket_promotions_products_total out=project.basket_prom_prod_sorted;
    by invoice_id;
run;

/****** Calculate the sum using PROC MEANS ******/
proc means data=project.basket_prom_prod_sorted sum noprint;
    by invoice_id;
    var Invoice_value;
    output out=project.invoice_total_value sum=Invoice_total_value;
run;

/*3*/
/* Create a table for sales */
data project.sales_invoice;
    set project.invoice;
    where Operation = 'Sale';
run;
/*20.419*/

/* Create a table for returns */
data project.returns_invoice;
    set project.invoice;
    where Operation = 'Return';
run;
/*3.795*/


/*4*/
data project.customers_with_age;
    set project.customers;

    birth_date = mdy(month_of_birth, day_of_birth, year_of_birth);
    today_date = '01JAN2019'd;

    if year(birth_date) > 1910 and year(birth_date) <= 2000 then
        age = intck('year', birth_date, today_date);
    else
        age = .; /* Invalid birth year, set age as missing value */

    drop birth_date today_date;
run;
/*10.000*/


/******2.Describe and explain using graphs who is your customer. What is the profile of the
audience to which the company's products are targeted?******/

/*• What are the demographic characteristics i.e. age, gender and region of the
company's customers? */

/* Summary statistics for age */
proc means data=project.customers_with_age;
  var age;
  output out=age_summary
                    mean=Mean_Age
                    min=Min_Age
                    max=Max_Age
                    std=Std_Age
                    n=N_Total;
run;

/* Summary statistics for gender */
```

```
proc freq data=project.customers_with_age;
  tables gender/out=gender_summary;
run;

/* Summary statistics for region */
proc freq data=project.customers_with_age;
  tables region/out=region_summary;
run;

/*Adding the Age_Range column*/
data project.customers_with_age_range (compress=yes);
        retain customer_id Last_name First_name Address Country Postal_Code City Region Gender
Age_Range;
        format Age_Range $50.;
   set project.customers_with_age;
   if age < 18 then Age_Range = "Under 18";
   else if age >= 18 and age <= 25 then Age_Range = "Very Young";
   else if age >= 26 and age <= 35 then Age_Range = "Young";
   else if age >= 36 and age <= 50 then Age_Range = "Middle Age";
   else if age >= 51 and age <= 65 then Age_Range = "Mature";
   else if age >= 66 and age <= 75 then Age_Range = "Old";
   else Age_Range = "Very Old";
        drop Day_Of_Birth Month_Of_Birth Year_Of_Birth age;
run;


/*Merging datasets customers with invoice to calculate number of visits*/
proc sort data=project.customers_with_age_range out=project.customers_with_age_range_sorted
; by customer_id; run;

proc sort data=project.invoice out=project.invoice_sorted_cid
; by customer_id; run;

data project.customers_with_age_range_invoice;
merge project.customers_with_age_range_sorted (in=b)
project.invoice_sorted_cid (in=i);
by customer_id;
if b=1 and i=1;
run;

/* Calculating number of visits by age_range*/
proc sql;
  create table project.visit_counts as
  select Age_range,
    count(distinct Invoice_ID) as num_visits
  from project.customers_with_age_range_invoice
  group by Age_range;
quit;

/*Merging datasets customers with product to calculate number of distinct SKUs*/
proc sort data=project.customers_with_age_range_invoice out=project.customers_invoice_sorted_iid
; by invoice_id; run;

proc sort data=project.basket_product out=project.basket_product_sorted_iid
; by invoice_id; run;

data project.customers_invoice_basket_product;
        merge project.customers_invoice_sorted_iid (in=b)
                project.basket_product_sorted_iid (in=i);
        by invoice_id;
        if b=1 and i=1;
run;

/* Calculating number of distinct SKUs purchased by age range */
proc sql;
  create table project.distinct_sku_count as
```

```
  select Age_range
    ,count(distinct SKU) as num_distinct_skus
  from project.customers_invoice_basket_product
  group by Age_range;
quit;

/* Calculating total cost of purchases by age range */

proc sql;
  create table project.total_purchase_cost as
  select Age_range
    ,sum(Quantity * Product_Price) as total_cost
  from project.customers_invoice_basket_product
  group by Age_range;
quit;

/* Calculate the percentage of customers in each age group */
proc sql;
  create table project.age_range_percentages as
  select Age_Range,
        put((count(distinct Customer_ID) /
            (select count(distinct Customer_ID) from project.customers_invoice_basket_product)) * 100, 5.1) ||
'%' as percentage
  from project.customers_invoice_basket_product
  group by Age_Range;
quit;

/* Create a frequency table for age groups */
proc freq data=project.customers_invoice_basket_product;
  tables Age_Range / out=Age_Range_freq;
run;

/*******3. Exploration and understanding of sales*******/

/*What was the level of Sales and Returns?*/

/* Merge the necessary datasets */
proc sort data=project.basket_invoice out=project.basket_invoice_sorted_pid
; by product_id; run;
proc sort data=project.basket_product out=project.basket_product_sorted_pid
; by product_id; run;

data project.sales_returns_data;
  merge project.basket_invoice_sorted_pid (in=i) project.basket_product_sorted_pid (in=b);
  by product_id;
  if i=1 and b=1;
run;

/* Calculate the monetary values for Sales and Returns */
data project.sales_returns_monetary;
  set project.sales_returns_data;
  if Operation = 'Sale' then
    Monetary_Value = Product_Price * Quantity;
  else if Operation = 'Return' then
    Monetary_Value = Product_Price * Quantity;
  else
    Monetary_Value = 0;
run;

/*Create graphs for the average basket size i.e. number of SKU's, total monetary value,
etc and comment on your findings*/

/* Calculate the average basket size by number of SKU's */
proc sql;
  create table project.avg_sku as
  select Operation,
```

```
        count(SKU) as Avg_SKU
  from project.sales_returns_monetary
  group by Operation;
quit;

/* Calculate the average basket size by total monetary value */
proc sql;
  create table project.avg_monetary_value as
  select Operation,
        mean(Monetary_Value) as Avg_Monetary_Value
  from project.sales_returns_monetary
  group by Operation;
quit;

/*Create a report that shows the top products per product line and product type with
respect to sales value in descending order. Show also the subtotal sales of each
product type.*/
/* Calculating the top products per product line and product type */
proc sql;
  create table project.top_products as
  select
    'Product line'n as Product_line,
    'Product type'n as Product_type,
    Product,
    sum(Monetary_Value) as TotalSales
  from project.sales_returns_monetary
  group by Product_line, Product_type, Product
  order by Product_line, Product_type, TotalSales desc;
quit;

/* Calculating the subtotal sales for each product type */
proc sql;
  create table project.subtotal_sales as
  select
    Product_line,
    Product_type,
    sum(TotalSales) as SubtotalSales
  from project.top_products
  group by Product_line, Product_type
  order by Product_line, Product_type, SubtotalSales desc;
quit;

/* Combining top products and subtotal sales */
data project.final_report;
  merge project.top_products project.subtotal_sales;
  by Product_line Product_type;
run;

/*Use graphs to show the contribution to the company's revenues of each region of
the country.*/

/* Merge the necessary datasets */
proc sort data=project.sales_returns_monetary out=project.sales_returns_monetary_cid
; by customer_id; run;

data project.sales_returns_monetary_customers;
  merge project.customers_with_age_range_sorted (in=i) project.sales_returns_monetary_cid (in=b);
  by customer_id;
  if i=1 and b=1;
run;

proc sql;
  create table company_revenues_0 as
  select region,
    sum(case when Operation = 'Sale' then Monetary_Value else 0 end) as Total_Sales_Revenue,
    sum(case when Operation = 'Return' then Monetary_Value else 0 end) as Total_Returns_Revenue
```

```
  from project.sales_returns_monetary_customers
  group by region;
quit;

data project.company_revenue_region;
  set company_revenues_0;
  Net_Revenue = Total_Sales_Revenue - Total_Returns_Revenue;
run;

/*For the top region found in the previous question show the contribution to the
company's revenues per gender*/
proc sql;
  create table company_revenues_gender_0 as
  select gender,
    sum(case when Operation = 'Sale' then Monetary_Value else 0 end) as Total_Sales_Revenue,
    sum(case when Operation = 'Return' then Monetary_Value else 0 end) as Total_Returns_Revenue
  from project.sales_returns_monetary_customers
  where region='SP'
  group by gender;
quit;


data project.company_revenues_gender;
  set company_revenues_gender_0;
  Net_Revenue = Total_Sales_Revenue - Total_Returns_Revenue;
run;


/******4.Zoom into the promotional activities by answering the following questions:******/

data project.product_promotion;
set project.basket_promotions_products;
if Promotion = 0 then Promotion_desc="No Promotion";
else if Promotion in (0.1, 0.2, 0.3) then Promotion_desc="Promotion";
run;

data project.product_promotion_excl_noprom;
set project.basket_promotions_products;
if Promotion = 0.1 then Promotion_name="10% Off";
else if Promotion = 0.2 then Promotion_name="20% Off";
else if Promotion = 0.3 then Promotion_name="30% Off";
where Promotion ne 0;
run;



/* Creating a new variable for the day of the week */
data project.sales_with_day;
  set project.sales_returns_monetary;
  SaleDayOfWeek = weekday(InvoiceDate);
run;

/* Calculating total sales amount per day of the week */
proc sql;
  create table project.sales_per_day as
  select SaleDayOfWeek,
        sum(Monetary_Value) format comma12.2 as TotalSales
  from project.sales_with_day
  group by SaleDayOfWeek;
quit;
proc print data=project.sales_per_day;
run;

/* Calculating number of SKUs per invoice per day of the week */
proc sql;
  create table project.sku_per_invoice_per_day as
```

```
  select SaleDayOfWeek,
         count(SKU) as NumberOfSKUs
  from project.sales_with_day
  group by SaleDayOfWeek;
quit;

proc print data=project.sku_per_invoice_per_day;
run;

/******5. It should be also mentioned that the SKU of each product contains “hidden”
information.******/

/* Extracting supplier code and creating a new variable */
data project.sales_with_supplier;
  set project.sales_returns_monetary;
  /* Extract the ninth (9th) digit as the supplier code */
  Supplier_ID = input(substr(SKU, 9, 1), 8.);
  WeightedFrequency = Quantity;
run;

proc sort data=project.sales_with_supplier out=project.sales_with_supplier_sorted;
by supplier_id; run;

data project.sales_with_supplier_name;
merge project.sales_with_supplier_sorted (in=i)
project.suppliers (in=b);
by supplier_id;
if i=1 and b=1;
run;

/* Frequency report */
proc freq data=project.sales_with_supplier_name; tables Supplier_Name ;
 weight WeightedFrequency; title "Percentage of Products Sold by Each Supplier";
run;


/* Calculating the revenue of its supplier */
proc sql;
  create table supplier_revenues_0 as
  select supplier_name,
    sum(case when Operation = 'Sale' then Monetary_Value else 0 end) as Total_Sales_Revenue,
    sum(case when Operation = 'Return' then Monetary_Value else 0 end) as Total_Returns_Revenue
  from project.sales_with_supplier_name
  group by supplier_name;
quit;

data project.supplier_revenues;
  set supplier_revenues_0;
  Net_Revenue = Total_Sales_Revenue - Total_Returns_Revenue;
run;


/*Merging datasets for origin*/

proc sort data=project.products out=project.products_sorted_origin
;by product_origin; run;

data project.products_origin;
   merge project.origin(in=a rename=(code=product_origin))
         project.products_sorted_origin(in=b);
   by product_origin;
run;

proc sort data=project.sales_with_supplier_name out=project.sales_supplier_name_sorted_porig
;by product_origin; run;
```

```
data project.sales_supplname_origin;
merge project.sales_supplier_name_sorted_porig (in=i)
            project.products_origin (in=b);
by product_origin;
run;

/* Calculating the revenue of its supplier */
proc sql;
  create table supplier_origin_revenues_0 as
  select supplier_name,
                    Country,
    sum(case when Operation = 'Sale' then Monetary_Value else 0 end) as Total_Sales_Revenue,
    sum(case when Operation = 'Return' then Monetary_Value else 0 end) as Total_Returns_Revenue
  from project.sales_supplname_origin
  group by supplier_name, Country;
quit;

data project.supplier_origin_revenues;
  set supplier_origin_revenues_0;
  Net_Revenue = Total_Sales_Revenue - Total_Returns_Revenue;
run;

/* Creating the cross-tabulation using proc tabulate */
proc tabulate data=project.supplier_origin_revenues;
  class Country supplier_name;
  var Net_Revenue;
  table Country, supplier_name*Net_Revenue*sum;
  keylabel sum='Total Revenue';
run;

/********6. The company wants to profile its customers based on their importance so as to offer
them personalized services and products. The customer segmentation is asked to be
done based on the three parameters of the RFM model*********/

proc sort data=project.customers_invoice_basket_product
out=project.cust_inv_bask_prod_sort_prom; by promotion_id;run;

data project.cust_inv_bask_prod_prom;
merge project.cust_inv_bask_prod_sort_prom (in=i)
project.promotions (in=b);
by promotion_id;
run;

proc sql;
  create table project.RFM as
  select Customer_ID,
    intck('week', max(InvoiceDate), '16DEC2011'd) as Recency,
    count(distinct InvoiceNo) as Frequency,
    sum(Quantity * Product_Price * (1 - Promotion / 100)) as MonetaryValue,
          1 as T
  from project.cust_inv_bask_prod_prom
  group by Customer_ID;
quit;


data project.RFM_SAS_PROJECT_2023_VISUALIZE;
  set CASUSER.'RFM SAS PROJECT 2023_VISUALIZE'n;
run;

/* Create separate data sets for each cluster */
data newcustomers highvaluecust;
  set project.RFM_SAS_PROJECT_2023_VISUALIZE;
  if _CLUSTER_ID_ = 2 then output newcustomers;
  else if _CLUSTER_ID_ = 3 then output highvaluecust;
run;
```

```
/*Merging datasets to have also the demographic columns*/
proc sort data=newcustomers out=project.newcustomers;
by customer_id; run;
proc sort data=highvaluecust out=project.highvaluecust;
by customer_id; run;


data project.newcustomers_with_age;
merge project.customers_with_age_range_sorted (in=i)
project.newcustomers (in=b);
by customer_id;
if i=1 and b=1;
run;

data project.highvaluecust_with_age;
merge project.customers_with_age_range_sorted (in=i)
project.highvaluecust (in=b);
by customer_id;
if i=1 and b=1;
run;

/* Describe the demographic data of New Customers */
proc freq data=project.newcustomers_with_age noprint;
table Age_Range/out=age;run;

proc freq data=project.newcustomers_with_age noprint;
table Gender/out=gender;run;

proc freq data=project.newcustomers_with_age noprint;
table Region/out=region;run;

/* Describe the demographic data of High Value Customers */
proc freq data=project.highvaluecust_with_age noprint;
table Age_Range/out=age1;run;

proc freq data=project.highvaluecust_with_age noprint;
table Gender/out=gender1;run;

proc freq data=project.highvaluecust_with_age noprint;
table Region/out=region1;run;


/*******8.The company is interested to change internally the store based on the products that
tend to be bought together******/

proc sort data=CASUSER.MBA_SAS_PROJECT out=project.MBA_SAS_Project;
  by descending LIFT;
run;

proc sql;
        create table project.NewcustMBA as
        select a.*
        from project.basket_product_invoice as a
        where customer_id in (select customer_id from project.newcustomers)
;quit;

proc sql;
        create table project.HighValueMBA as
        select a.*
        from project.basket_product_invoice as a
        where customer_id in (select customer_id from project.highvaluecust)
;quit;


proc sort data=CASUSER.MBA_SAS_PROJECT_NEWCUST out=project.MBA_SAS_Project_NEWCUST;
  by descending LIFT;
```

```
run;


proc sort data=CASUSER.MBA_SAS_PROJECT_HIGHVALUE out=project.MBA_SAS_Project_HIGHVALUE;
  by descending LIFT;
run;
```