

Contents

Executive Summary	1
Data and initial cleansing	3
Data pre-processing.....	3
Customer Profile.....	4
Demographic characteristics	4
Behavioral characteristics	6
Sales Exploration	11
Level of sales and returns	11
Graphs for the average basket size by gender.....	15
Product categories' performance	18
Customer Preferences.....	22
Size distribution of women's blouses	30
Promotional Activities	32
Product Launch.....	40
Customer Segmentation	42
Customer Profiling.....	43
Product Associations	49

Executive Summary

The purpose of this case study was to conduct an analysis of a clothing company in order to provide suggestions on how the organization should act in order to operate in a more efficient and effective way.

After making the necessary pre-processing of the data and performed a basic exploratory analysis, we addressed various issues such as the sales and returns transactions, the colors and sizes mostly preferred and the existence of seasonality in sales. We also performed customer segmentation using RFM analysis and we implemented market basket analysis by using the association rules among the product categories for all the customers as well as for the customers that we deemed as the best and valuable ones (based on the comparison of their Recency, Frequency and Monetary values from the RFM analysis with those of the average customer of the store).

In order to perform our analysis, we used the SAS Enterprise Guide and SAS Enterprise Miner. We had a set of POS data in our disposal that were related to sales of women's clothes and were provided by the store and contained information regarding the customers, the issued receipt and the baskets. For the RFM and market basket analysis we used a dataset that we created based on the given data. This dataset contained some outliers which could seriously affect our results, so we removed them (the resulted RFM dataset contained 41,983 observations instead of 46,739).

Our main analysis contained, among others, an exploration of the products returned (e.g., most-returned product categories, country of origin of the returned products, etc.), a gender-wise and age-wise analysis (e.g., preferred product categories and colors, average basket size, etc.), an analysis by day of the week (e.g., evaluation of the success of advertising activities, distribution of purchases, etc.) and an identification of the product categories that were most popular, were influenced by promotional activities and resulted in higher revenues or higher returns.

The main findings of the study were that black was the most bought color, closely followed by 'denimltstone' from men, while the women had a wider variety of options as their 2nd choice. We also saw that Monday was, by far, the day with the most purchases and the best day for successful promotional activities to take place. After we dive deeper into our analysis, we observed that depending on the product the store wants to promote, the most-suitable day changes (i.e., the day in which the increase in sales, after promotion, reaches its peak differs for each product). Additionally, before organizing promotional activities or before placing an order to the suppliers, the company should also take into account other factors such as the age (e.g., younger customers mainly buy products on 'Monday') and the season (e.g., from March till June, 'damen-blusen' was the most sold product category). An interesting finding was also that 'damen-blusen' was bought only in small, medium and large sizes (mainly in the latter two).

Furthermore, from the best and valuable customers (i.e., the two most important clusters created), we observed that over 85% of them were women and also that the people who are between 36 to 75 years old correspond to almost 92% of these types of the company's clients. The main difference among them regarding the customers age, was that the middle-aged customers that were identified as the best ones constituted more than 4% of the total customers compared to those identified as valuable, while the valuable customers aged between 66 to 75 years were 3%

more compared to the best ones. This knowledge can help the store to better target its customer's types and use appropriate marketing strategies based on their age and gender.

Finally, we saw that what should be considered from the store as the best or most suitable proposal differs based on the product category and the customer type. For instance, if the company wants to recommend a product to one of its customers, then its first choice should be to identify the customers that have bought the 'MyOwn' product and offer them the 'ViaCortesaDOB' product category, as they are very likely to purchase the latter product also (vice versa). Similarly, if the store decides to target those considered as its best customers, then it should start by promoting the 'Herren-Shirt/Sweat' product category to those who buy 'Jeanswear' (vice versa). Also, if the valuable customers are to be targeted, then the recommendations to this customer type should begin by selecting those who have bought the 'ViaCortesaDOB' product category and suggesting them the 'MyOwn' product.

Data and initial cleansing

The main data entities contained in the POS data were related to the customers (i.e., their ID, date of birth and gender), the issued receipt entitled as document table (i.e., ID of the receipt, date when the receipt was issued, whether the receipt was sale or return and the customer ID that is related with the receipt/basket) and data related to each basket (i.e., ID of the receipt, code of the promotional activity as well as the SKU¹ and the price of the product). Additionally, there are some other data that are used as lookup tables which contain the coding (i.e., code and description) of the product categories and the colors. Each data entity was a .csv format file and the data were analyzed with SAS Enterprise Guide and SAS Enterprise Miner.

The initial step was loading the different csv files in SAS Enterprise Guide for further processing via Base SAS programming. After removing all the missing values (i.e., 6,368 observations from the customer table) and duplicate rows (i.e., 39 rows from the basket table) the customer table contained 46,758 observations, the document table 139,758 observations, the basket table 513,103 observations, the product category table 42 observations and the color table 65 observations.

Additionally, we removed all the spaces for all character strings in each table and regarding the two lookup tables we set added a leading zero for the first 9 observations of the ID column in order to be properly sorted.

Data pre-processing

Before continuing with our analysis, there was some necessary pre-processing that needed to be done. At first, we created two new datasets regarding the number and value of SKUs for each basket. Specifically, we computed the total number of SKU's for every basket where we observed that a typical basket contains approximately four items, the smallest basket one item and the largest basket 378 items. We also calculated the total value of the SKU's for every basket which was equal to 23,115,305 euros. We observed that the basket total values range from 10 to 20,163 euros with a mean value of 165 euros and that the basket with the highest value (i.e., 20,163 Euro) is the one with the most items (i.e., 378 total items).

Furthermore, we created some new tables. The issued receipt of a purchase related to either sale or return of product(s), thus we created from the document table (139,758 observations) two new tables, one for the sales transactions (90,710 observations) and one for the returns transactions (49,048 observations). Additionally, for the customers that data about their birth date existed, we calculated each customer's age based for today's date (i.e., 01/01/2018). For validity of the dates purposed, we considered only customers that were born after 1920 and were adults.

¹ A stock-keeping unit (SKU) is a scannable bar code composed of an alphanumeric combination of eight-or-so characters, which allows vendors to automatically track the movement of inventory.

Customer Profile

Demographic characteristics

We started our analysis by reviewing some demographic characteristics of the company's customers. Regarding their distribution based on gender, we observed from Figure 1 that approximately 17% were men and 83% were women, which was expected given that the store sells women's clothes. As for the customers' age, we observed from Figure 2 that the age variable follows a normal distribution, which means that the main customer base for the company are the middle-aged customers.

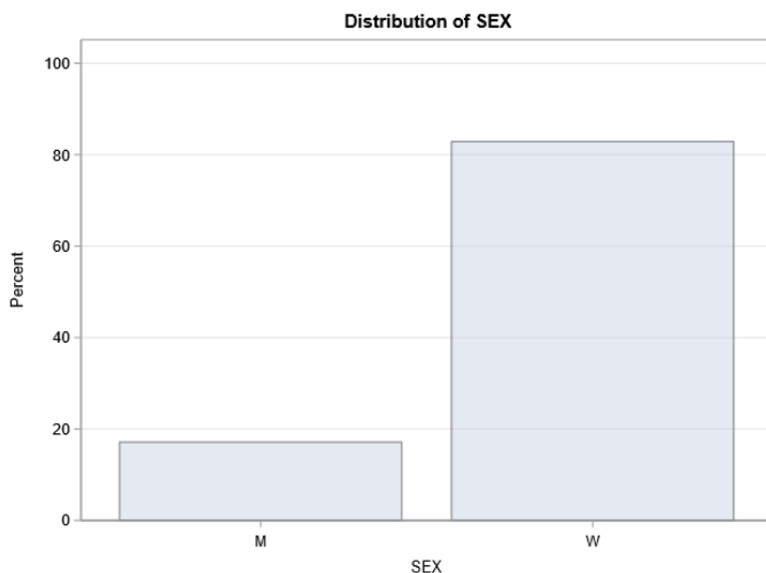


Figure 1: Bar Chart for the distribution of customers based on gender



Figure 2: Histogram for the distribution of customers based on age

In order to be more accurate regarding the age distribution of the customers, we created some basic categories based on the customers' age. Specifically, 18-25 age group were labeled as "very young", 26-35 as "young", 36-50 as "middle age", 51-65 as "mature", 66-75 as "old" and the customers who were older than 75 years old were labeled as "very old".

By reviewing now the distribution based on the age groups, we observed from Figure 3 that almost 52% of the customers were between 51 to 65 years old (i.e. mature), 24% were 26 to 35 years old, 15% 66 to 75 years old, 4% 26 to 35 years old, 4% older than 75 years and 0.4% were 18 to 25 years old.

So, we see that the people who are between 36 to 75 years old correspond to almost 92% of the company's clients - with the main customer base (i.e. more than 50%) consists of 51 to 65 years old customers - and that the younger age groups (i.e. 18 to 35 years old) corresponded to less than 5% of the company's customers.

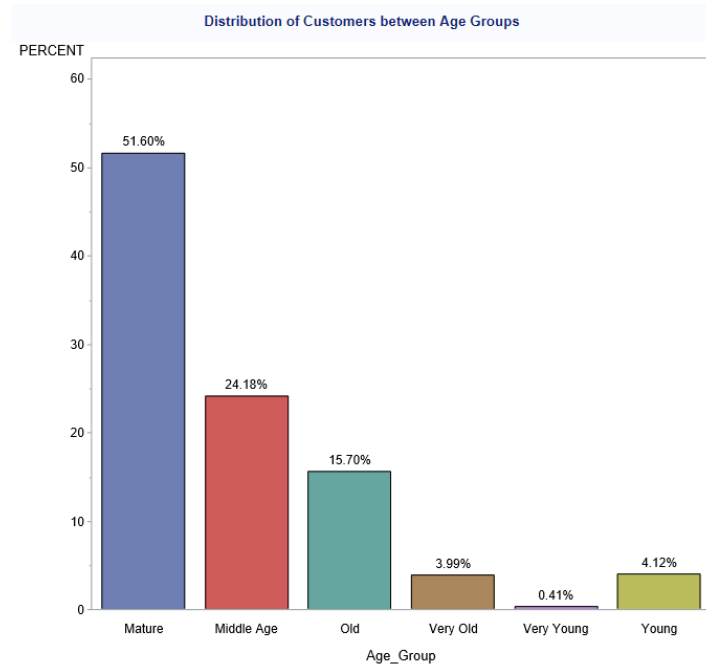


Figure 3: Distribution of customers between age groups

Later, we decided to study the behavioral characteristics of each age group. Firstly, we created a new dataset which combined the information contained in many different tables. This was the reason, why we observed in table 1 and figure 4 some discrepancies in the customers' percentages that belong to each age group compared to figure 3.

Specifically, we observed that about 53% of the customers are between 51 and 65 years old, which combined with the middle aged (i.e. 36 to 50 years old) customers (circa 28%) and old (i.e. 66 to 75 years old) customers correspond to 93% of the total customers. The rest refer to the young (i.e. 26 to 35) years old and older than 76 years old customers, along with a few of very young (i.e. 18 to 25 years old) customers (0.3%).

Pie Chart of customers for each age group

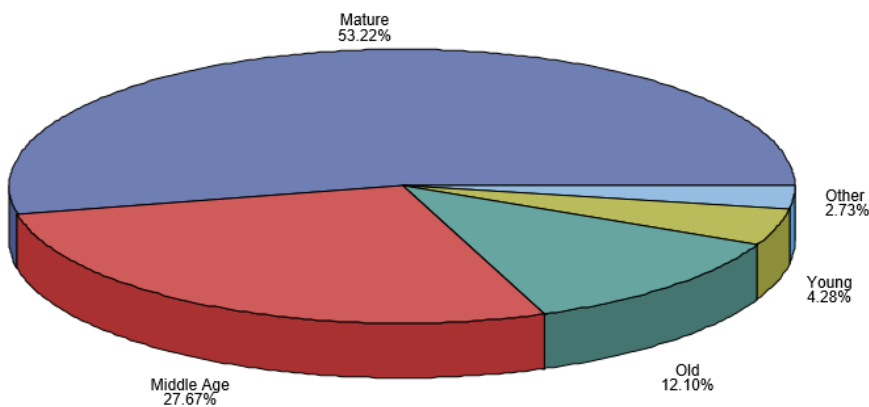


Figure 4: Customer distribution for each age group

Table 1: One-Way frequency table for customer distribution for each age group

Age_Group	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Mature	230041	53.22	230041	53.22
Middle Age	119586	27.67	349627	80.89
Old	52300	12.10	401927	92.99
Very Old	10473	2.42	412400	95.41
Very Young	1332	0.31	413732	95.72
Young	18503	4.28	432235	100.00

Behavioral characteristics

From Table 2 we observed that there were 432,235 visits of customers to the stores, from which the people aged between 26 and 65 were the ones with the most visits to the stores on average while the customers that were older than 76 years old did not visit the stores as much.

Specifically, we saw that the 'middle-aged' group was the one with the most visits to the stores with 10.6 on average, followed by the 'young' and 'mature' age groups with 9.6 and 9.5 average visits to the stores correspondingly. The 'old' group had 7.1 average visits to the stores, 0.2 more than the 'very young' age group, while the 'very old' age group only averaged about 5.6 visits to the stores.

Table 2: Average and total visits to the stores for each group

Age_Group	tot_visits	Avg_visits	Visits_pct
Middle Age	119586	10.6	27.7%
Young	18503	9.6	4.3%
Mature	230041	9.5	53.2%
Old	52300	7.1	12.1%
Very Young	1332	6.9	0.3%
Very Old	10473	5.6	2.4%

As for the 42 different product categories that exist in the data, we observed that the customers' age groups are interested in all the available categories except from the very young customers. Specifically, the five product categories that the very young customers do not buy are the "Aktionsware", "Herren-Sakkos", "Krawatten", "SchiesserHerren- Wäsche" and "Triumph".

The product category that was mostly preferred by the middle age and mature groups (i.e. from 36 to 65 years old) was the "Damen-Blusen" while for the younger and older customers (i.e. from 18 to 35 and older than 66) it was "Jeanswear".

In order to get a better understanding of the products preferred by each age group, we present in the table 3 the results in descending order of preference. From there, we noticed that the most bought product categories for all ages were the "Jeanswear" and the "Damen-Blusen".

Additionally, we observed that the younger (i.e. 18 to 35 years old) customers has exactly the same preferences (first the "Jeanswear", followed by the "Damen-Blusen" and "MyOwn"). The same happened for the older (i.e. older than 66 years old) customers (first the "Jeanswear", followed by the "Damen-Blusen" and "Damen-Hosen").

Table 3: Three most frequent product categories per age group

Age Group	Product description
Mature	1. Damen-Blusen 2. Jeanswear 3. Damen-Hosen
Middle Age	1. Damen-Blusen 2. Jeanswear 3. MyOwn
Old & Very Old	1. Jeanswear 2. Damen-Blusen 3. Damen-Hosen
Very Young & Young	1. Jeanswear 2. Damen-Blusen 3. MyOwn

The information on the product categories preferred by each age is presented on its utmost detail in figures 5-10, on which our previous conclusions were based.

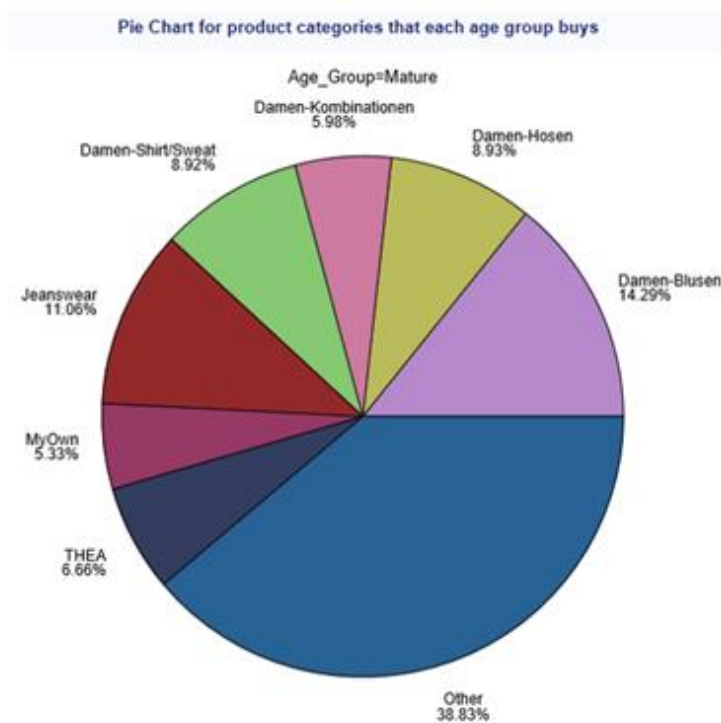


Figure 5: Pie Chart for product categories bought by 'Mature' age group

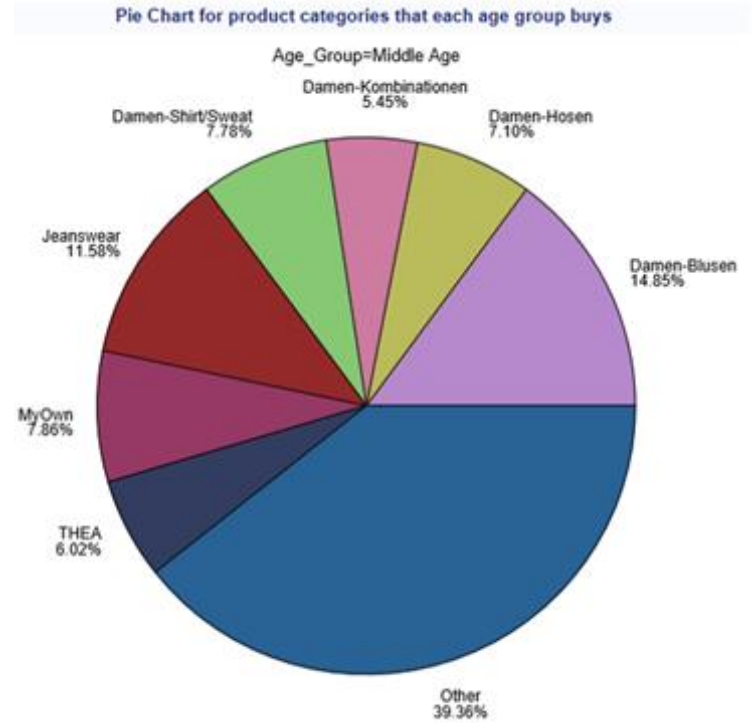


Figure 6: Pie Chart for product categories bought by 'Middle Age' age group

Pie Chart for product categories that each age group buys

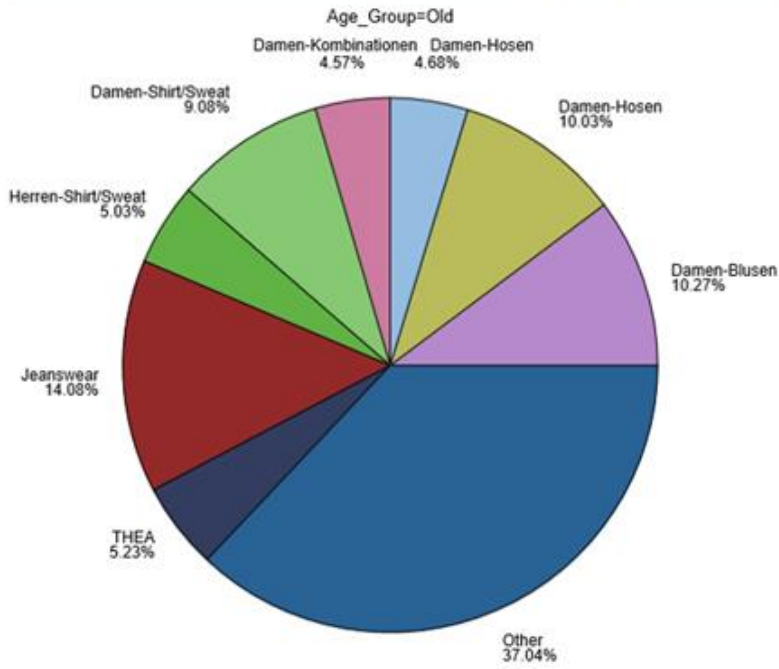


Figure 7: Pie Chart for product categories bought by 'Old' age group

Pie Chart for product categories that each age group buys

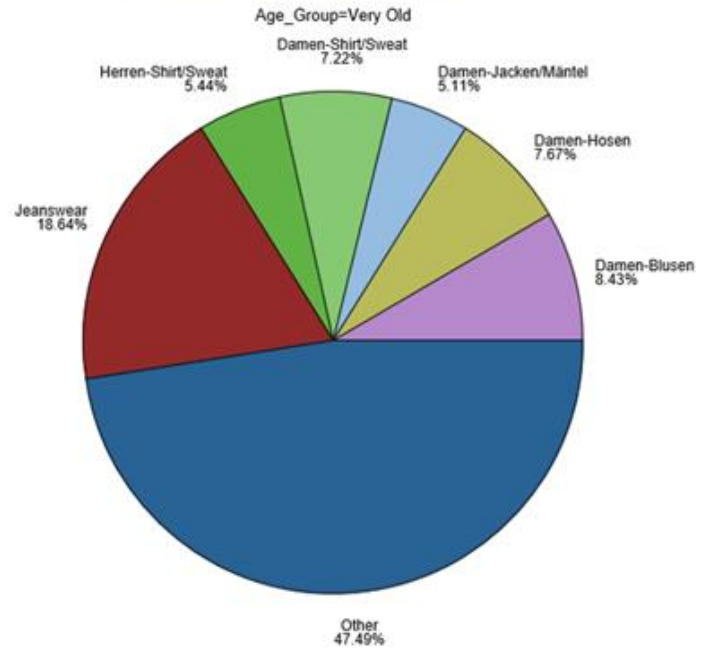


Figure 8: Pie Chart for product categories bought by 'Very Old' age group

Pie Chart for product categories that each age group buys

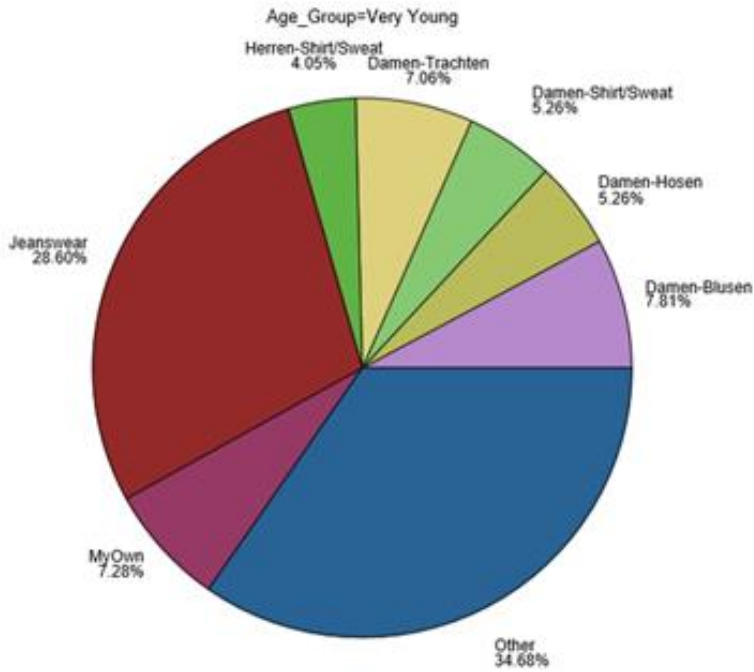


Figure 9: Pie Chart for product categories bought by 'Very Young' age group

Pie Chart for product categories that each age group buys

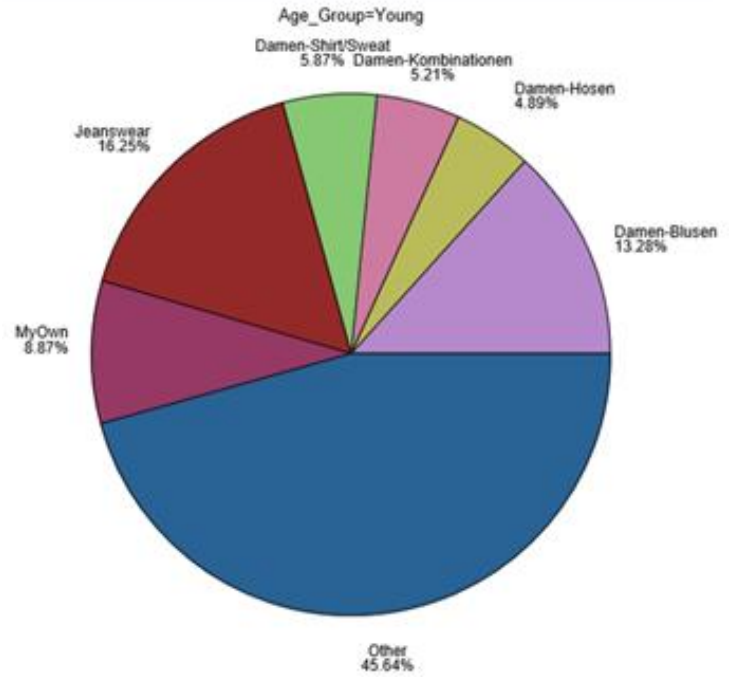


Figure 10: Pie Chart for product categories bought by 'Young' age group

Regarding the cost of purchases, we calculated some other statistics for each age group (see Table 4). At first, we saw that a high price of a product did not affected its purchases in specific customers as all age groups bought products that their price ranged from 10 to 80 euros.

In terms of total cost of purchases, the customers labeled as mature had spent over 10 million euros, the middle aged more than 5 million, the old over than 2 million, the young about 800 thousand euros, the very old about half a million and the very young just about 60 thousand euros. Nevertheless, this was expected as the spending's of each age group depend on the number of visitors per age group (i.e. more customers, more spending's).

Thus, by looking at the average cost we observed that the differences are almost insignificant and in contrast with the total cost, the very young customers are the ones with the more spending's, but also the one's with the bigger variances in their consumer behavior, while the other age groups are most stable leading by the customers who are older than 76 years old who exhibit the most stable behavior among all. This difference between the total and average cost of purchases for each age group can also be seen by Figures 11, 12 below.

Table 4: Summary statistics for cost of purchases

Age_Group	Total_cost	Average_cost	sd	Minimum_cost	Maximum_cost	Range	Number_of_obs
Mature	10350991	45.00	20.60	10	80	70	230041
Middle Age	5397872	45.14	20.62	10	80	70	119586
Old	2332511	44.60	20.74	10	80	70	52300
Very Old	462933	44.20	19.94	10	80	70	10473
Very Young	62137	46.65	22.92	10	80	70	1332
Young	837204	45.25	20.25	10	80	70	18503

Pie Chart for the total cost of purchases of each age group

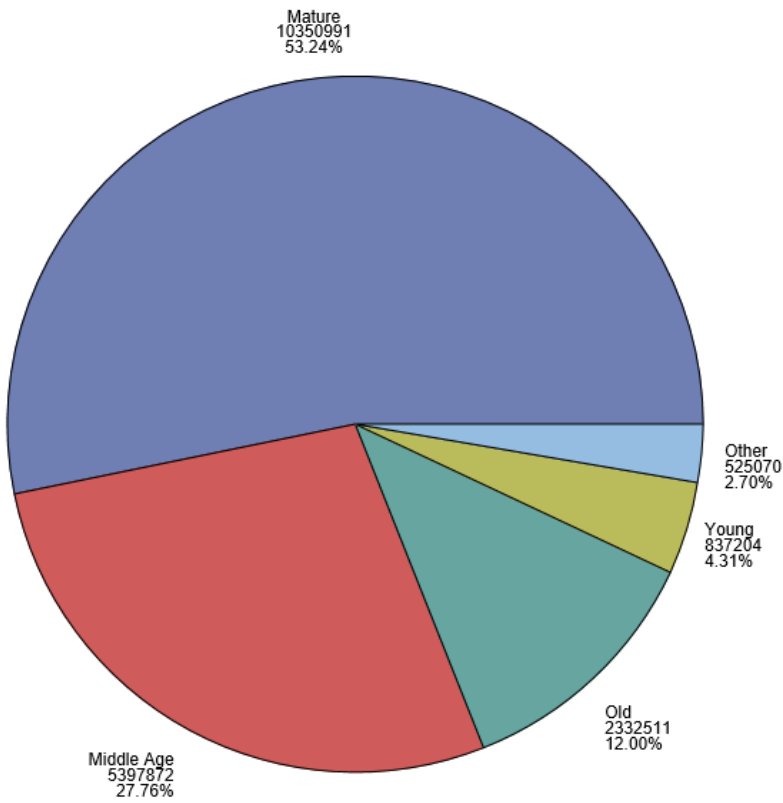


Figure 11: Pie Chart for the total cost of purchases of each age group

Pie Chart for the average cost of purchases of each age group

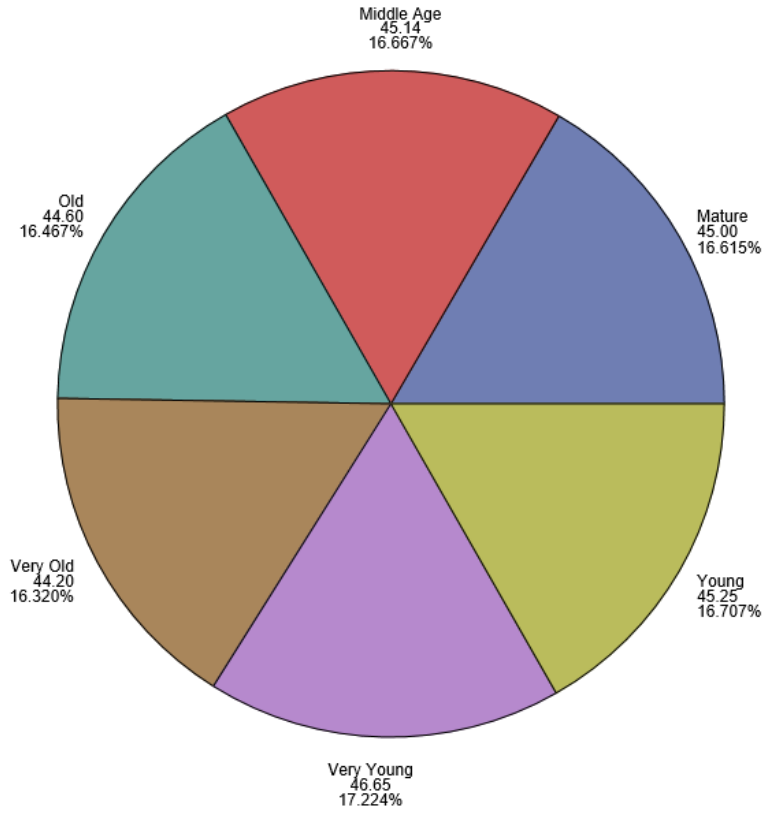


Figure 12: Pie Chart for the average cost of purchases of each age group

Sales Exploration

Level of sales and returns

First of all, we observed that the value of the returned products is more than 6.7 million dollars, that's why we decided to further investigate the sales and returns transactions. Specifically, from Table 5 we noticed that the sales of clothes correspond to about 70% of the total transactions of the customers, while the returns of products (i.e. cancellation of product orders) corresponds to about 30% of those the total customers' transactions.

To get a better understanding of these percentages, we created a bar chart with the monetary values for each of the two different transactions shown in the issued receipt (see Figure 13). In order to be as accurate as possible we wanted to have the most available data values in our disposal, so this chart was created by using only the tables that contained the necessary information.

From this chart we observed that the percentage that corresponds to the sales transaction (i.e. 70.6%) translates into about 17 million earnings for the shop, while the return of a product (i.e. 29.4%) translates into a little more than 6 million euros.

The percentage of the returns of the products is really big considering that it would normally allow the shop earn 6 more million euros. Thus, we decided to dive further into the returns transactions in order to identify potential problems with the clothes returned. In order to do this, we combined all the information that exists in the data. After doing this, we observed that we lost some observations, but from Table 6 and Figure 14 we saw that the distribution of sales and return transactions were almost the same (i.e. 0.04% different) and as expected the total monetary value for each transaction is lower (i.e. about 3 million lower for sales and 1 million lower for returns), so the reduction of the dataset will cause a problem in the inferences that will be made.

Table 5: Sales and Returns distribution

Frequency table for sales and returns

The FREQ Procedure

MOVEMENT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Sale	362225	70.59	362225	70.59
Cancellation	150878	29.41	513103	100.00

Table 6: Sales and Returns distribution (full-merged dataset)

Frequency table for sales and returns

The FREQ Procedure

MOVEMENT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Sale	304957	70.55	304957	70.55
Cancellation	127278	29.45	432235	100.00

Bar Chart with the monetary values for the issued receipt

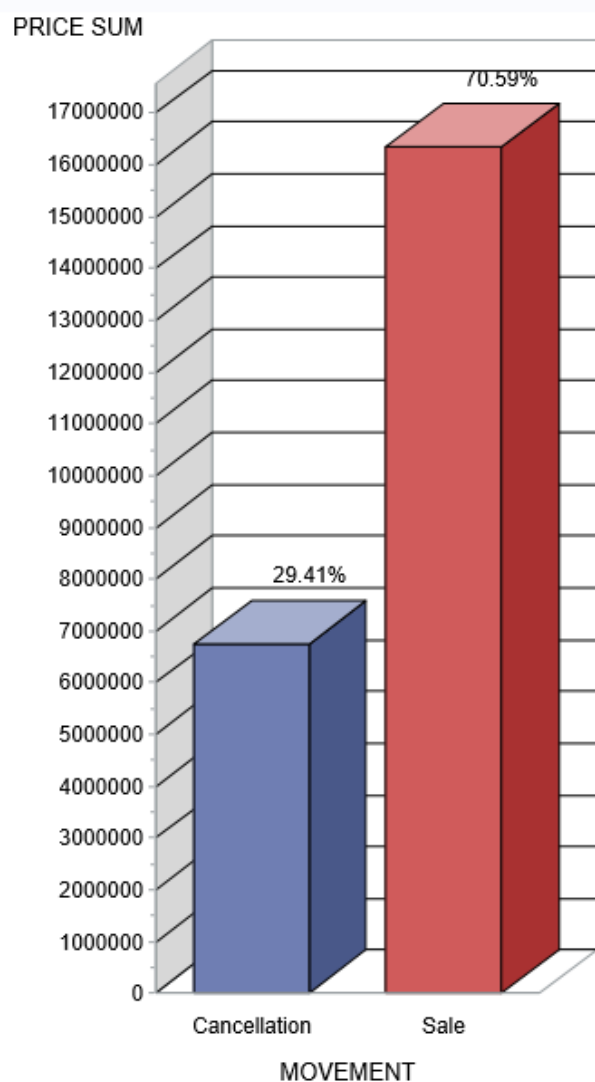


Figure 13: Bar Chart with the monetary values for the issued receipt

Bar Chart with the monetary values for the issued receipt

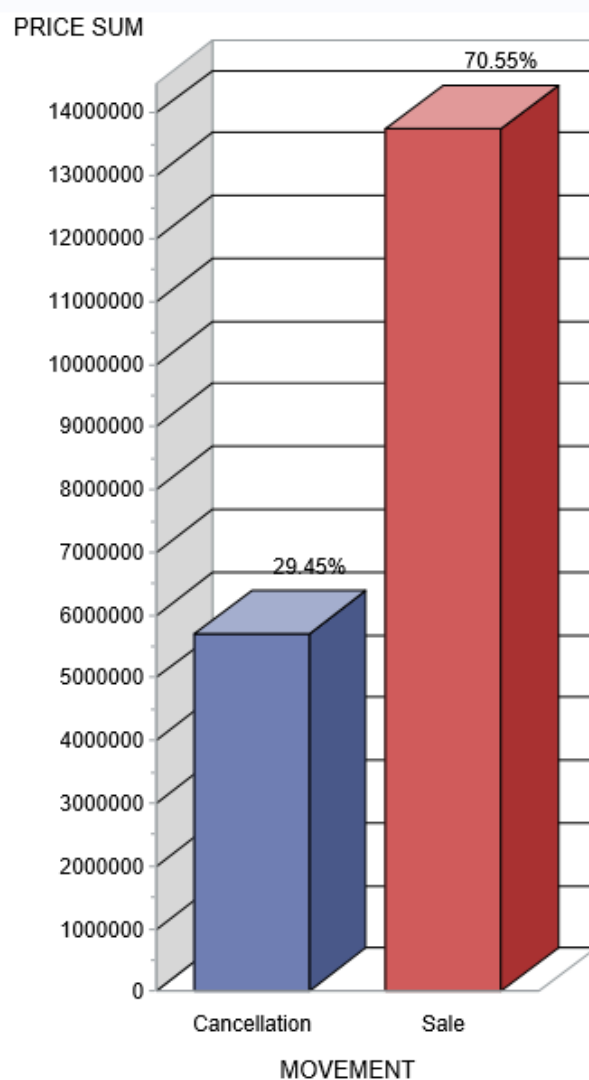


Figure 14: Bar Chart with the monetary values for the issued receipt (full-merged dataset)

At first, we focused on the product categories that are returned. Specifically, we observed that the “Damen-Blusen” was the one returned most, followed by the ‘Jeanswear”, “Damen-Hosen”, “Damen-Shirt/Sweat” and “Damen-Kombinationen”. A table with the ten most frequent product categories returned (see Table 7) and a pie chart for product categories that are most returned (see Figure 15) are presented below.

Table 7: Ten most frequent product categories returned

PRODUCT_DESCRIPTION	COUNT	PERCENT
Damen-Blusen	19657	15.4441
Jeanswear	13410	10.5360
Damen-Hosen	11457	9.0016
Damen-Shirt/Sweat	9700	7.6211
Damen-Kombinationen	9056	7.1151
MyOwn	8745	6.8708
THEA	8524	6.6972
Damen-Jacken/Mäntel	5084	3.9944
MarkenshopsDOBmodisch	3745	2.9424
Damen-Strick	3374	2.6509

Pie Chart for product categories that are most returned

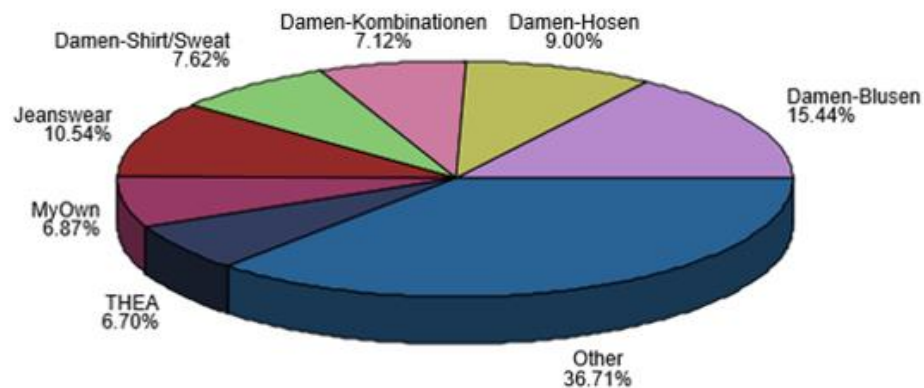


Figure 15: Pie Chart for product categories that are most returned

Then, we focused on the days in which the products are returned. From Figure 16 (or from Table 8), we observed that the days with the most returns are Tuesday and Wednesday, where there are more than 25% and 20% of the returns made correspondingly. Additionally, we saw that about the same amount of products is returned during Thursday and Friday, while Monday is the day with the fewest returns made. In Saturday there was not a return made by any of the customers.

Pie Chart for days at which products are returned

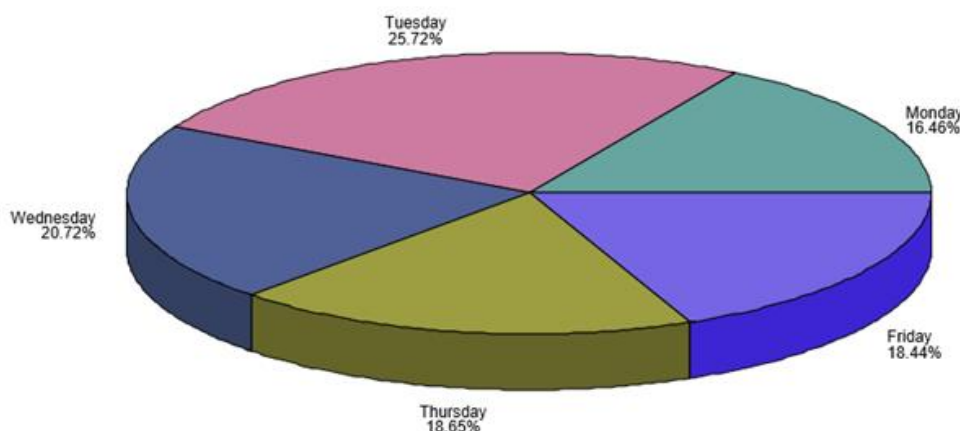


Table 8: Distribution of products returned during each day

The FREQ Procedure

Day_of_week	Frequency	Percent
Tuesday	32734	25.72
Wednesday	26374	20.72
Thursday	23743	18.65
Friday	23476	18.44
Monday	20951	16.46

Figure 16: Pie Chart for distribution of products returned during each day

Furthermore, we analyzed the returned products based on the country in which the company is based (by using the first three digits of the SKU). From Figure 17, we saw that almost 91% of the products that are returned refer to companies that are based in Germany and more than 8% to companies that are based on Austria.

The companies that were based in the rest of the countries (i.e. Denmark, Faroe Islands, Netherlands, Spain and Andorra, Belgium and Luxembourg, Switzerland and Liechte, Hong Kong, Bookland (ISBN), United Kingdom, Restricted distribution) had almost no returns. The exact percentages of returns for all countries can be seen in Table 9.

Table 9: Distribution of returned products based on country of origin

Pie Chart for country of origin of returned products

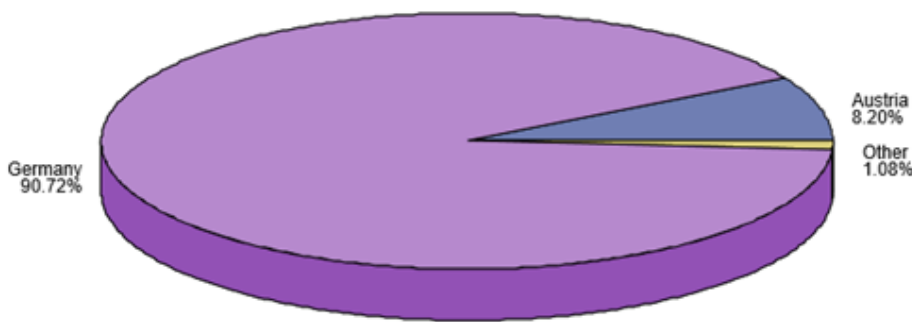


Figure 17: Pie Chart for country of origin of returned products

The FREQ Procedure

Country	Frequency	Percent
Germany	115482	90.72
Austria	10440	8.20
Denmark, Faroe Islands	497	0.39
Netherlands	419	0.33
Spain and Andorra	334	0.26
Belgium and Luxembourg	54	0.04
Switzerland and Liechte	37	0.03
Hong Kong	27	0.02
Bookland(ISBN)	4	0.00
United Kingdom	3	0.00
Restricted distribution	1	0.00

Graphs for the average basket size by gender

Next, we created graphs for the average basket size based on the gender of the customers. From Tables 10, 11 and Figure 18, we observed that 83% of the customers are female and only 17% are male and that the average basket size for men and women contained about 4.3 and 3.5 SKUs correspondingly. We also saw from Figure 19 and Table 12 that females spend more than 16 million in buying clothes from the shop, while the male spend approximately 3 million euros.

Regarding the product categories, we noticed that all different categories are preferred by both genders. To be more thorough, we created the Figures 20, 21 and Tables 13, 14 from which we drawn some interesting conclusions. As for men, we observed that they mostly prefer buying “Jeanswear” (35% of their purchases), while also liking the “Herren-Shirt/Sweat” and “GroßenGrößen” categories (each about 8% their purchases).

The women on the other hand, do not have one major preferred product category that far surpasses all the others. The “Damen-Blusen” is their most popular choices (15%), followed by “Damen-Shirt/Sweat” and “Damen-Hosen” (each about 10% their purchases). As a result, we observe that women have a wider variety of choices compared to men and a different taste in style. For instance, men vastly preferred the “Jeanswear” (35% of their purchases) which potentially thought as a “safe” and easy choice to make, while only 7% of the women bought “Jeanswear”.

Pie Chart for number of SKU's by gender

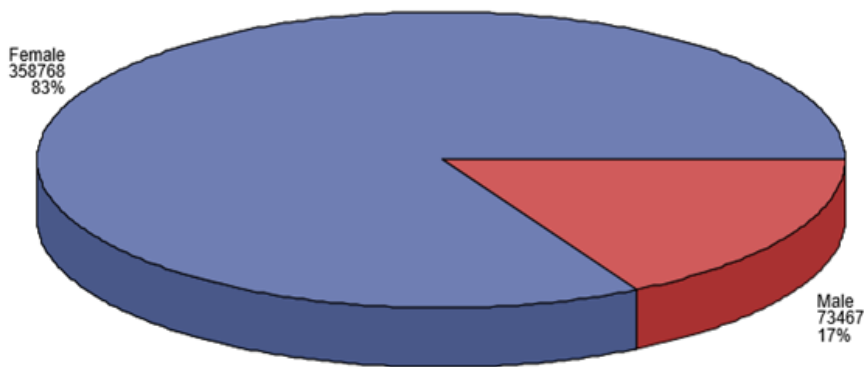


Figure 18: Pie Chart for number of SKU's by gender

Table 10: Average Basket Size per Gender

SEX	Basket_count	Avg_Basket_size
Men	73467	4.3
Women	358768	3.5

Table 11: Number of SKU's by gender

The FREQ Procedure

SEX	Frequency	Percent
Female	358768	83.00
Male	73467	17.00

Pie Chart for total monetary value spent by gender

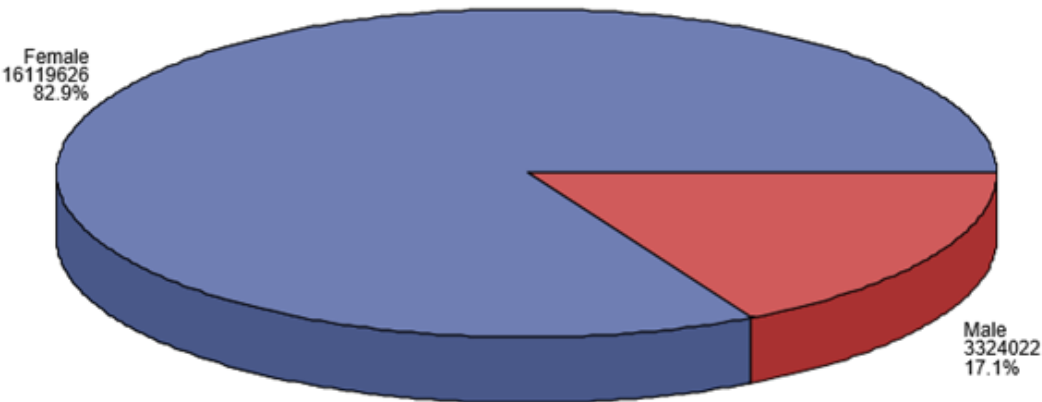


Table 12: Total monetary value spent by gender

SEX	Total_monetary_value
Male	3324022
Female	16119626

Figure 19: Pie Chart for total monetary value spent by gender

Pie chart for product categories preferred by customer gender

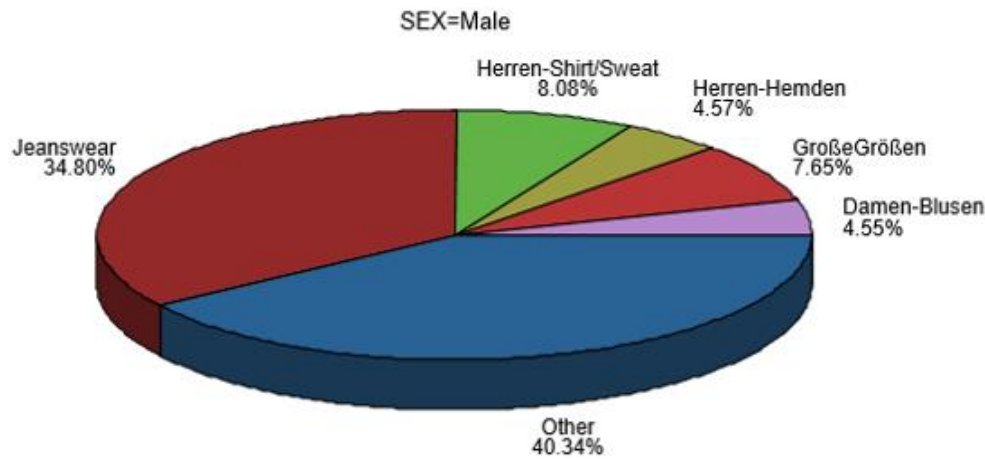


Figure 20: Pie chart for product categories preferred by Men

Table 13: Top-10 categories bought by Men

Top-10 products categories men buy

PRODUCT_DESCRIPTION	COUNT	PERCENT
Jeanswear	25568	34.8020
Herren-Shirt/Sweat	5939	8.0839
GroßeGrößen	5618	7.6470
Herren-Hemden	3361	4.5748
Damen-Blusen	3344	4.5517
Herren-Hosen	2829	3.8507
Herren-Wäsche	2485	3.3825
Damen-Shirt/Sweat	2275	3.0966
Damen-Hosen	2149	2.9251
ViaCortesaHAKA	1397	1.9015

Pie chart for product categories preferred by customer gender

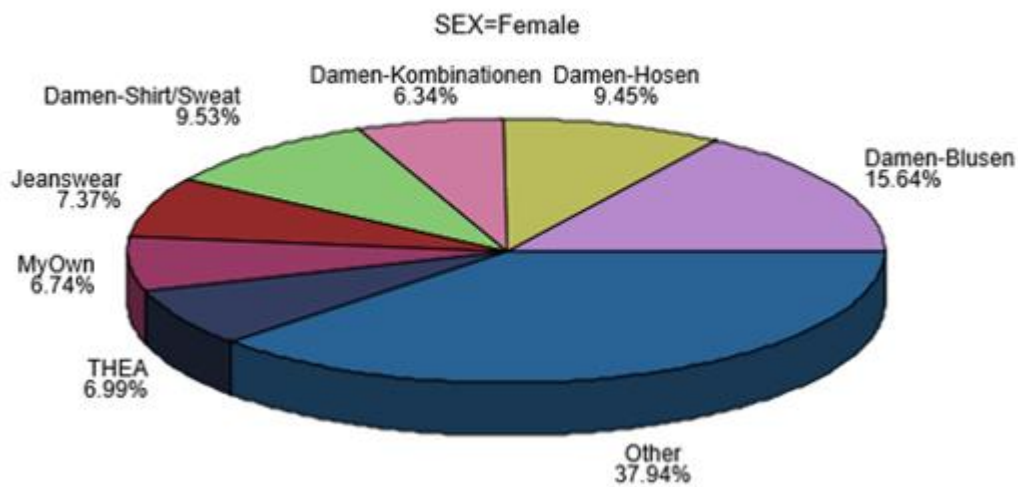


Figure 21: Pie chart for product categories preferred by Women

Table 14: Top-10 categories bought by Women

Top-10 products categories women buy

PRODUCT_DESCRIPTION	COUNT	PERCENT
Damen-Blusen	56103	15.6377
Damen-Shirt/Sweat	34198	9.5321
Damen-Hosen	33897	9.4482
Jeanswear	26430	7.3669
THEA	25074	6.9889
MyOwn	24187	6.7417
Damen-Kombinationen	22751	6.3414
Damen-Jacken/Mäntel	14305	3.9873
Damen-Wäsche	11367	3.1683
Damen-Strick	10459	2.9153

Product categories' performance

To get a better understanding regarding the product categories, we firstly created pie charts for the categories based on the issued receipt (i.e. sale or return). From Figures 22 & 23, we noticed that despite the “Herren-Shirt/Sweat” that is one of the main categories of products sold but not of products returned, the other main product categories are the same with little discrepancy on their results (i.e. a variation of 1-2%). For instance, the product proportion on the sales and returns transactions (in descending order of proportions) for the “Damen-Blusen” are 13% and 15% correspondingly, for the “Jeanswear” 12% and 10% correspondingly and for the “Damen-Hosen” 8% and 9% correspondingly.

Pie Chart for the product categories based on the issued receipt

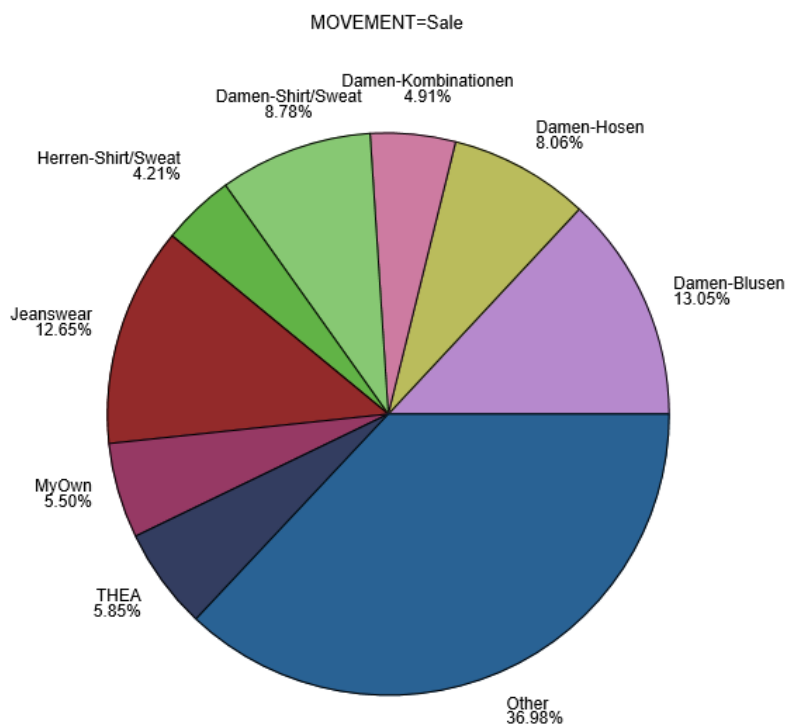


Figure 22: Pie Chart for the product categories for Sale

Pie Chart for the product categories based on the issued receipt

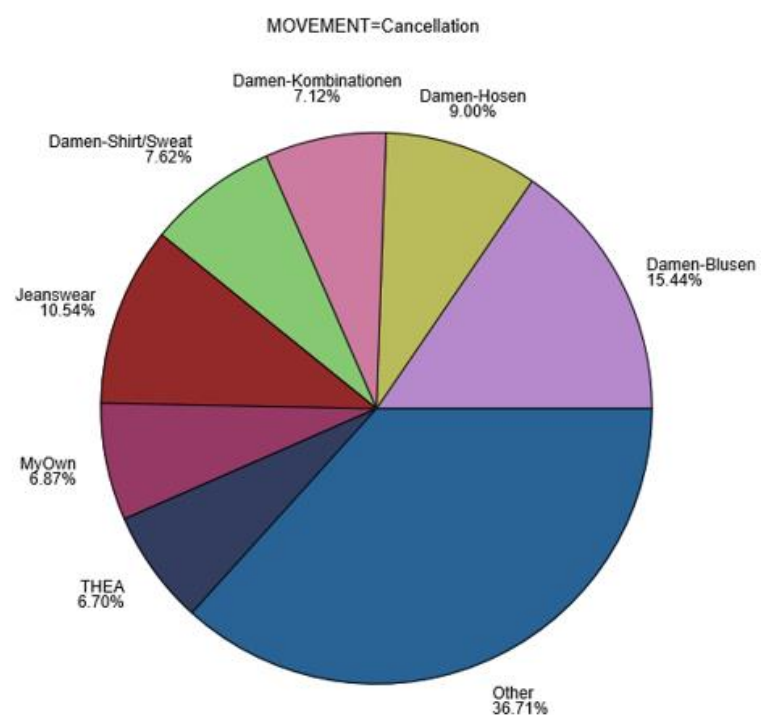


Figure 23: Pie Chart for the product categories for Return

We will now describe the performance of the various product categories that are sold in order to understand the product categories in which the company should focus and the ones that will be the least important.

From Figure 24 and Table 15, we saw that the most popular categories (i.e. the ones that are most bought by the customers) were the “Dame-Blusen” and the “Jeanswear” (both about 13% of the customers’ preferences), followed by the “Damen-Shirt/Sweat” and “Damen-Hosen” (both about 8% of the customers’ preferences).

Pie Chart for most popular categories sold

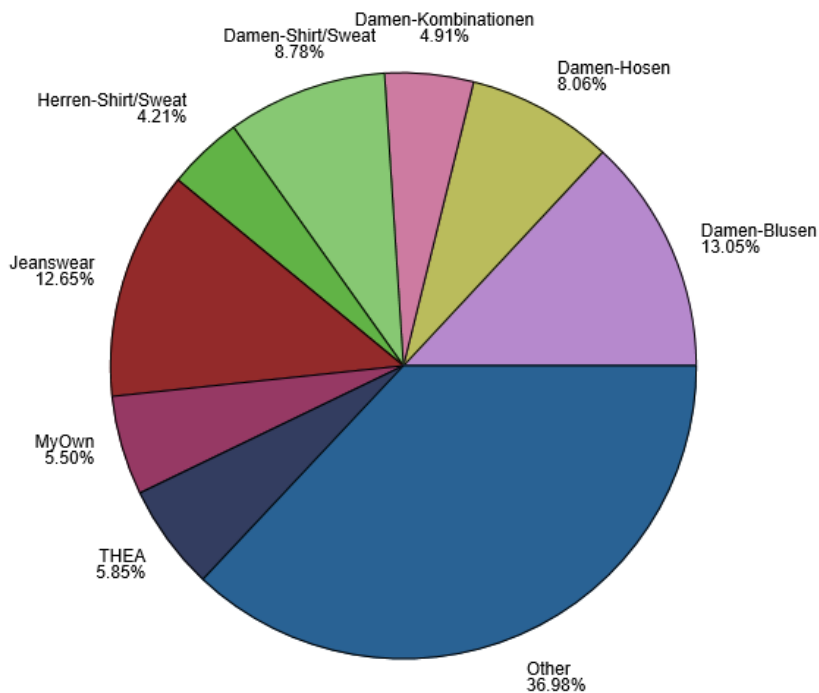


Figure 24: Pie Chart for most popular categories sold

Table 15: Ten most popular product categories sold

PRODUCT_DESCRIPTION	COUNT	PERCENT
Damen-Blusen	39790	13.0477
Jeanswear	38588	12.6536
Damen-Shirt/Sweat	26773	8.7793
Damen-Hosen	24589	8.0631
THEA	17843	5.8510
MyOwn	16781	5.5027
Damen-Kombinationen	14971	4.9092
Herren-Shirt/Sweat	12850	4.2137
Damen-Jacken/Mäntel	10318	3.3834
Damen-Wäsche	9421	3.0893

As for the product categories that brought the biggest earnings to the company, we saw from Figure 25 that those were in accordance with the most popular mentioned before, with a little discrepancy in the percentages. For example, the “Jeanswear” (second most popular category) is bought by 12.65% of the customers, while it corresponds to 12.96% of the company’s earnings.

Pie Chart for product categories with the most earnings

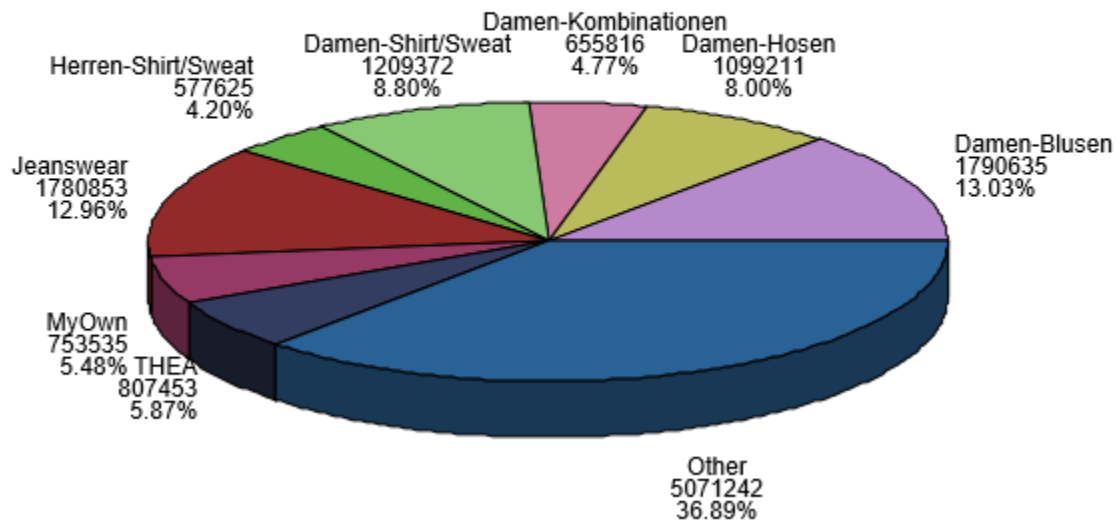


Figure 25: Pie Chart for product categories with the most earnings

We also noticed that from the 42 different product categories, more or less all categories were influenced by promotional activities in the sales transactions except from 'Krawatten' and 'Steilmann-Shop'.

Among these 40 categories, we observed from Table 16 that "Jeanswear" is the one influenced the most (20% of the promotional activities target this product) which might explain the reason why men in general do not know a lot about women's fashion, prefer buying this product. Then the "Damen-Blusen" and "Damen-Hosen" are influenced a lot by promotions (about 14% and 11% correspondingly).

Table 16: Top-10 categories that are influenced the most from promotional activities

PRODUCT_DESCRIPTION	Frequency	Percent
Jeanswear	36152	20.19
Damen-Blusen	25507	14.25
Damen-Hosen	20082	11.22
Damen-Shirt/Sweat	17686	9.88
Herren-Shirt/Sweat	10330	5.77
Damen-Jacken/Mäntel	7477	4.18
Damen-Strick	6204	3.47
Damen-Kombinationen	5617	3.14
MyOwn	4984	2.78
Damen-Wäsche	4818	2.69

Furthermore, we used Figure 26 and Table 17 in order to see the product categories that customers return most. We saw that the "Damen-Blusen" was the one with the most returns (about 15% of the returns correspond to this product). Next, were the "Jeanswear" (about 10% of the products returned referred to this product) followed by the "Damen-Hosen" (9% of the returns).

For further information, we present a table with the top-10 categories with the highest returns as well as a pie-chart depicting the product categories that were returned the most by the customers.

Pie Chart for categories with highest returns

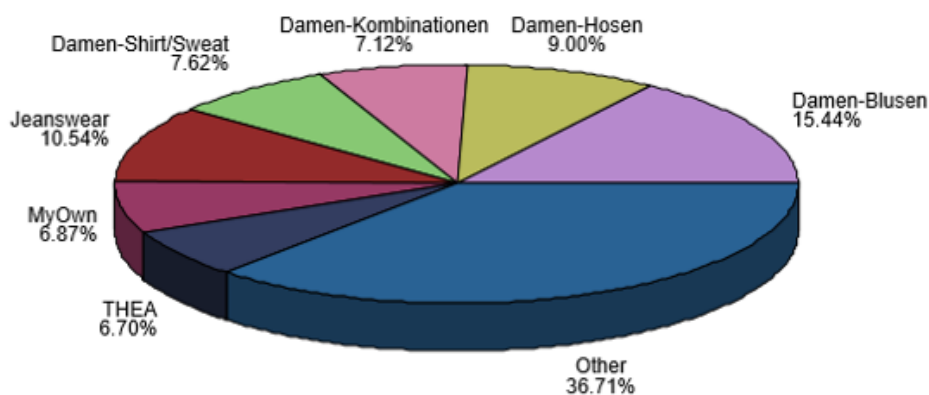


Figure 26: Pie Chart for categories with highest returns

Table 17: Top-10 categories with highest returns

PRODUCT_DESCRIPTION	COUNT	PERCENT
Damen-Blusen	19657	15.4441
Jeanswear	13410	10.5360
Damen-Hosen	11457	9.0016
Damen-Shirt/Sweat	9700	7.6211
Damen-Kombinationen	9056	7.1151
MyOwn	8745	6.8708
THEA	8524	6.6972
Damen-Jacken/Mäntel	5084	3.9944
MarkenshopsDOBmodisch	3745	2.9424
Damen-Strick	3374	2.6509

Customer Preferences

From Figure 27 and Table 18, we observed that the color that customers mostly bought was “Schwarz” (more than 17% of the clothes bought were black), followed by “Weiss” (White) and “Blaumittel-dunkel” (Medium-dark blue).

Pie Chart for the colors preferred by customers

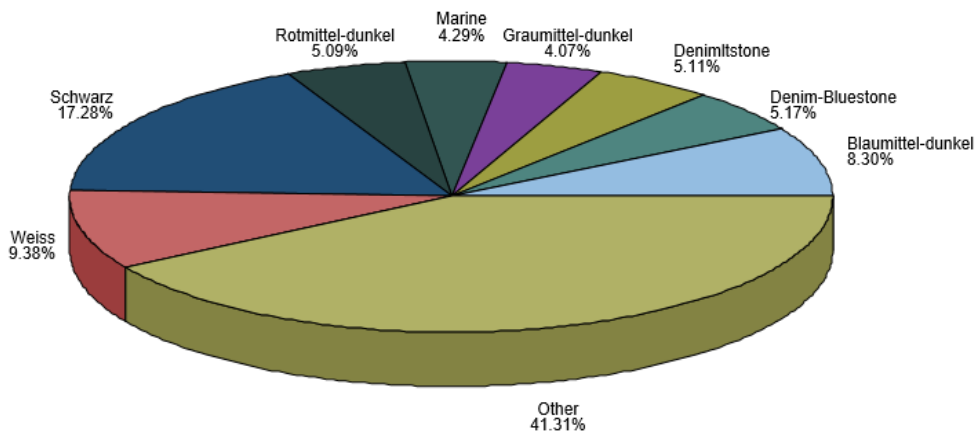


Figure 27: Pie Chart for most preferred colors in clothes by the customers

Table 18: Top-10 colors customers prefer buying

Top-10 colours customers prefer to buy

COLOUR_DESCRIPTION	COUNT	PERCENT
Schwarz	52690	17.2778
Weiss	28598	9.3777
Blaumittel-dunkel	25325	8.3044
Denim-Bluestone	15779	5.1742
Denimltstone	15574	5.1089
Rotmittel-dunkel	15511	5.0883
Marine	13090	4.2924
Graumittel-dunkel	12423	4.0737
GrA?nmittel-dunkel	11088	3.6359
Modelfarben	8672	2.8437

By reviewing the pie charts for the preferred colors based on the issued receipt (see Figures 28, 29) we observed that the colors bought and returned by the customers were approximately the same percentage, which means that the color did not affect as much whether a product will be returned or not. But, in order to be more accurate due to the fact that there were some discrepancies, we will use only the data from the sales transactions for our further analysis.

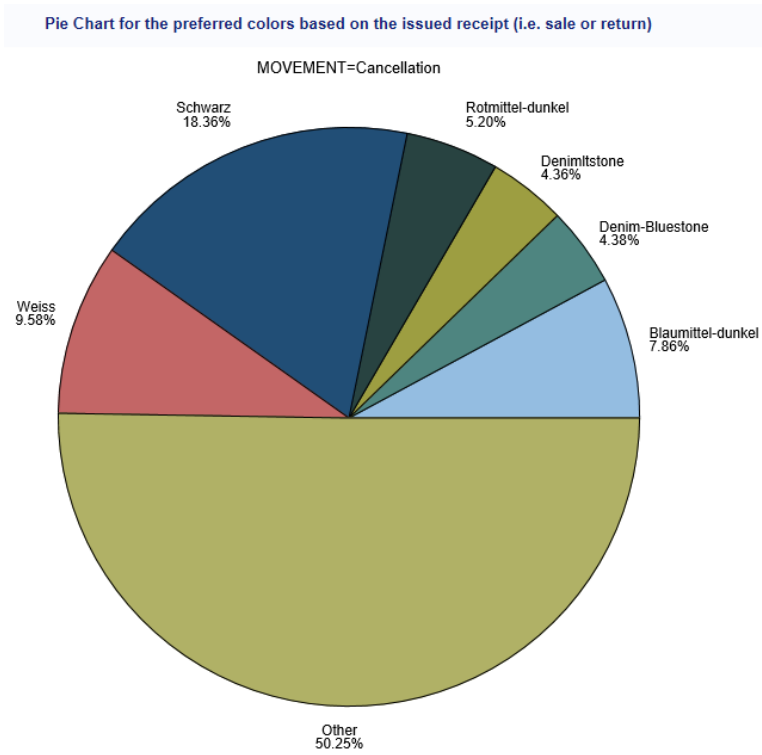


Figure 28: Pie Chart for the preferred colors for products sold

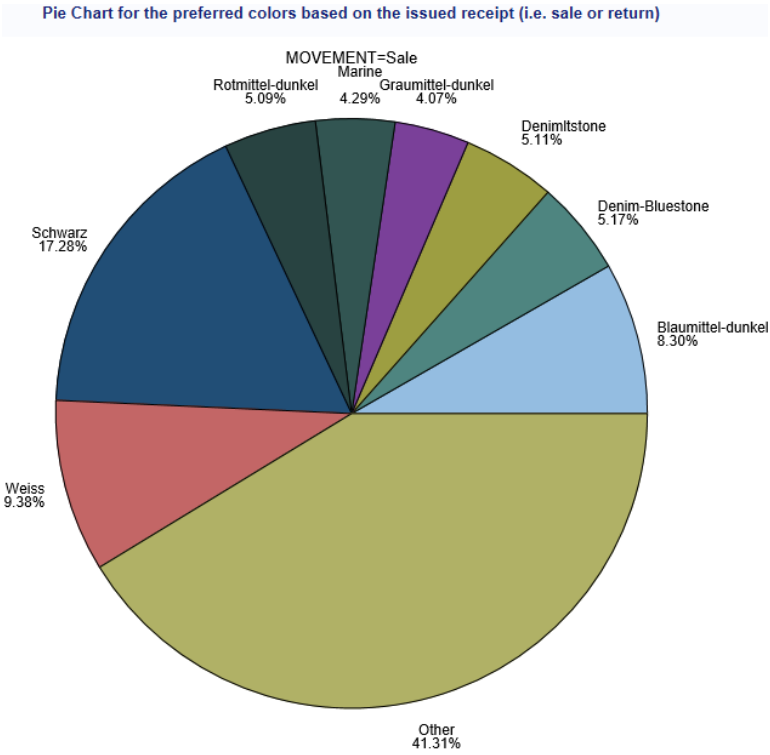


Figure 29: Pie Chart for the preferred colors for products returned

We continued our analysis by reviewing the colors that customers prefer to buy based on their sex and age group.

Regarding the sex, we observed from Figures 30, 31 and Tables 19, 20 that the most men bought “schwartz” and “denimiltstone” clothes (about 32% of them). Other popular choices were the “denim-bluestone” and “blaumitter-dunkel” (each about 9% of the customers’ preferences) and worth mentioning was that about 5% of the men bought either “weiss” (white), “rotmittel-dunkel” (medium-dark red) or “marine” colors. The rest 35% of their choices were distributed among different colors.

As for the women, the color that they mostly preferred was also “schwartz” (about 18%), but the next popular choice was “Weiss” (White) (about 10%), followed by “Blaumitter-dunkel” (8%). After that approximately 5% of their choices were either “Rotmittel-dunkel”, “Marine” or “Graumittel-dunkel”. This aforementioned colors counted about half of the colors mostly bought by the women. The other half of their choices, was distributed among other colors.

As a result, we can claim that men have two main colors they buy on a regular basis and seven colors account for 65% of their choices, while the clothes women prefer for their wardrobe consists about one fourth (1/4) of black and white clothes, along with a variety of a lot different clothes in order to combine them all together.

Additionally, we noticed that among the top choices of men and women, men have in general a good idea about women preferences but not perfect as for example the 2nd and 3rd colors most men buy are purchased at a much lower rate by women.

From the total percentages regarding the color preferences of customers presented in Table 21 we observed that men’s purchases correspond a lot less to the total revenues of the store (about 18% in total). For example, regarding the most popular color which was the black and corresponds to about 18% of the customers purchases, men only contribute to about 3% of these purchases. One value that we identified and deemed as ‘problematic’ came from the “denimiltstone” color for which both men and women buy at the same percentage (2.4% of their purchases). The fact that this color of women’s clothes is the second most bought by men but one of the least preferable by the women, depicts the different fashion tastes of men and women.

So, if we were to make proposals to the store we would suggest that it take more into account the women’s preferences. As a result, the product categories ordered from the suppliers should consist mainly of black clothes along with an adequate number of white and “blaumitter-dunkel” clothes. Now, if the store wants to also satisfy men’s choices, it should add to the above order a higher number of “denimiltstone” clothes as well as “denim-bluestone” and “blaumittel-dunkel” clothes.

Pie Chart for the preferred colors based on sex

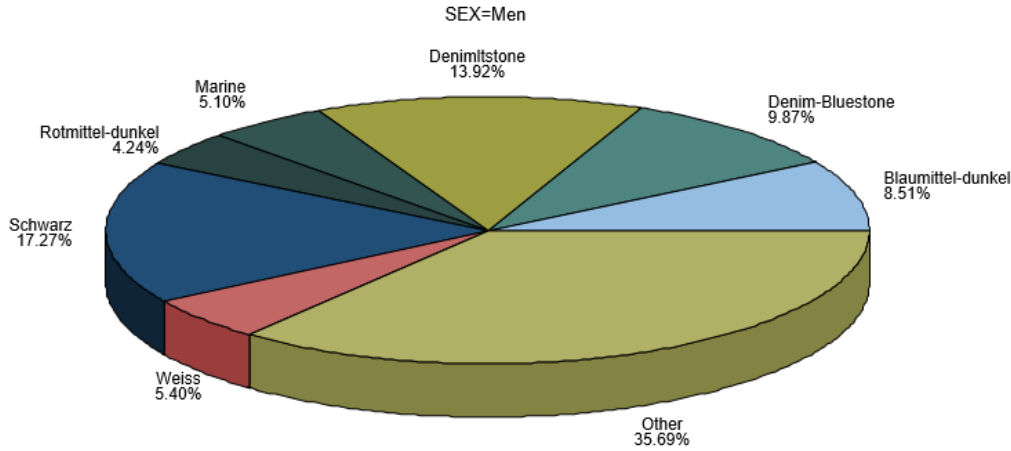


Figure 30: Pie Chart for the preferred colors for Men

Table 19: Top colors bought by Men

Top-10 colours men prefer buying

COLOUR_DESCRIPTION	COUNT	PERCENT
Schwarz	9828	17.2697
Denimltstone	7923	13.9222
Denim-Bluestone	5618	9.8719
Blaumittel-dunkel	4845	8.5136
Weiss	3071	5.3963
Marine	2900	5.0959
Rotmittel-dunkel	2412	4.2383
Anthrazit	2129	3.7411
Graumittel-dunkel	2021	3.5513
GrA?nmittel-dunkel	1510	2.6534

Pie Chart for the preferred colors based on sex

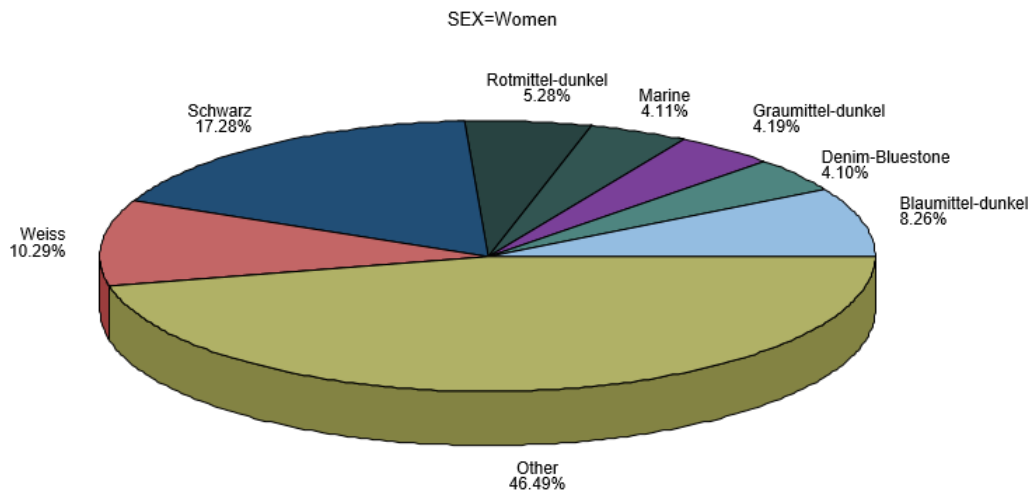


Figure 31: Pie Chart for the preferred colors for Women

Table 20: Top colors bought by Women

Top-10 colours women prefer buying

COLOUR_DESCRIPTION	COUNT	PERCENT
Schwarz	42862	17.2797
Weiss	25527	10.2912
Blaumittel-dunkel	20480	8.2565
Rotmittel-dunkel	13099	5.2808
Graumittel-dunkel	10402	4.1935
Marine	10190	4.1081
Denim-Bluestone	10161	4.0964
GrA?nmittel-dunkel	9578	3.8613
Denimltstone	7651	3.0845
Modifarben	7589	3.0595

Table 21: Part of the cross-tabulation table for color and sex

Table of COLOUR_DESCRIPTION by SEX			
COLOUR_DESCRIPTION	SEX		
	Women	Men	Total
Schwarz	42862 14.06	9828 3.22	52690 17.28
Weiss	25527 8.37	3071 1.01	28598 9.38
Blaumittel-dunkel	20480 6.72	4845 1.59	25325 8.30
Denim-Bluestone	10161 3.33	5618 1.84	15779 5.17
Denimltstone	7651 2.51	7923 2.60	15574 5.11
Rotmittel-dunkel	13099 4.30	2412 0.79	15511 5.09
Marine	10190 3.34	2900 0.95	13090 4.29
Graumittel-dunkel	10402 3.41	2021 0.66	12423 4.07
Natur+Ecrú	4119 1.35	407 0.13	4526 1.48
Denimdkstone	3216 1.05	1224 0.40	4440 1.46
Beige	6941 2.28	1460 0.48	8401 2.75
Total	248048 81.34	56909 18.66	304957 100.00

By looking at Figures 32-37, we drew many useful conclusions regarding the colors that are mostly preferred by the customers based on their age group that they belong. Specifically, for the black color we saw that the senior (i.e. older than 66 years old) customers are the ones that prefer it the least (about 14% of their preferences) and that the younger the customer the more purchases of black clothes will be done. It is indicative that about 1/4th of the adults that are under 25 years old buy black clothes.

Although white clothes is one of the most bought product categories, we noticed that among the age groups we have separated the customers, the younger ones (i.e. under 25 years old) are those who prefer the least this color which is a choice that older people (over 36 years old) typically prefer.

The “denimilstone” was a product category bought in higher frequency (circa 10% of their purchases) by the very young (i.e. under 25 years old) and very old (i.e. over 76 years old) adults and in between these age groups it was just one of the main categories bought by the customers (i.e. about 5% of their purchases).

The “blauittel-dunkel” is another of the main categories bought by the customers (about 8% of their preferences) and the ones that bought it the most were customers aged between 26 to 50 years old. As for the “rotmittel-dunkel” and the “marine”, they were two other main color categories which in general were preferred by the customers in the same amount (about 5%) regardless of their age.

Furthermore, we observed that the “beige” color could be labeled as a color for the elders as it is was bought in a higher frequency (more than 4%) only by the customers older than 76 years old. Similarly, the “graumittel-dunkel” is a color mostly bought by adults under 25 years old and –in a lesser frequency- by middle-aged customers (i.e. between 36 to 50 years old).

Finally, we can claim that customers that are under 50 years old have more standard choices regarding their colors as about 60% of their purchases is covered by seven main colors, while the customers over 50 want a little more variety of different colors in the clothes they buy as only 54% of their purchases refers to seven main colors.

To conclude with, we observed that if the company decided to focus its orders from the suppliers only to the “shcwarz”, “weiss”, “denimilstone”, “denim-bluestone” and “blauittel-dunkel” colors it would have covered over 50% of the demand of the younger (under 35 years old) customers and about 45% of the demand for the rest of them. Among the aforementioned five categories, the most popular one was the “shcwarz” color, especially for the younger customers who for every four clothes they buy one of them is black. So, this is the color it should focus on having the proper stock of products.

Afterwards and depending on the age group that the company wants to target it can easily understand of the preferences explained above what proportions of each color should order from its suppliers. For example, if the customers from 26 to 75 years old are targeted, then the two most bought colors behind black, are the white and “blauittel-dunkel” which account for over 8% of those customers preferences.

Pie Chart for the preferred colors bought based on age group

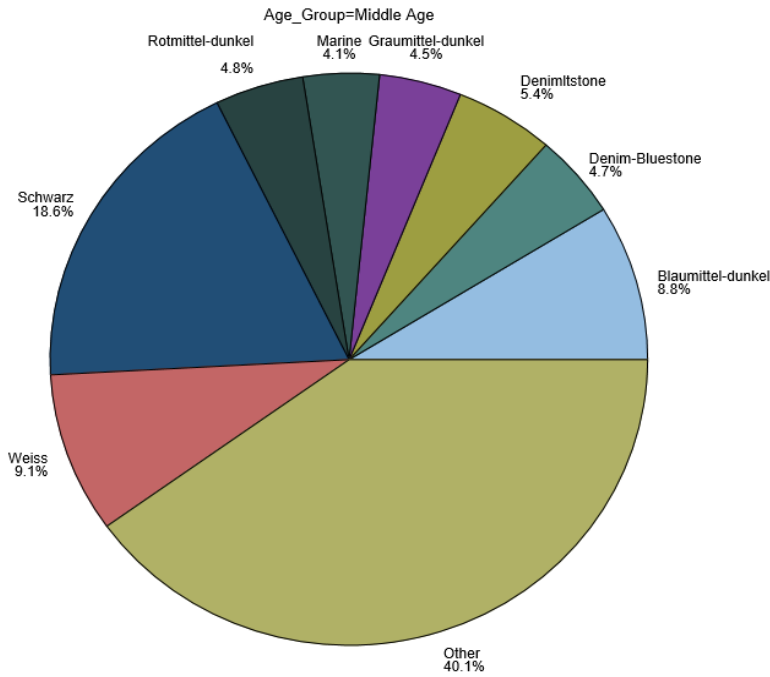


Figure 32: Pie Chart for the preferred colors for 'Middle Age' group

Pie Chart for the preferred colors bought based on age group

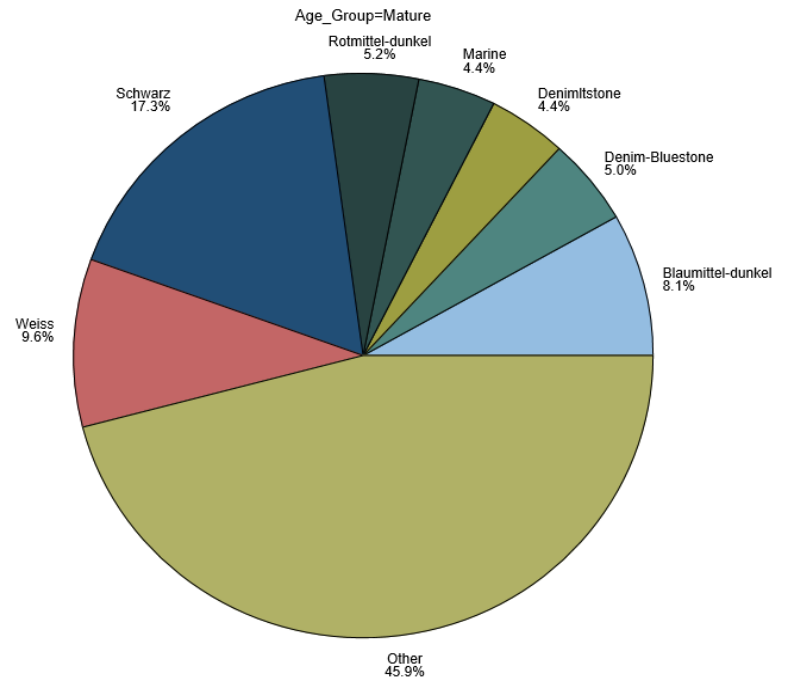


Figure 33: Pie Chart for the preferred colors for 'Mature' group

Pie Chart for the preferred colors bought based on age group

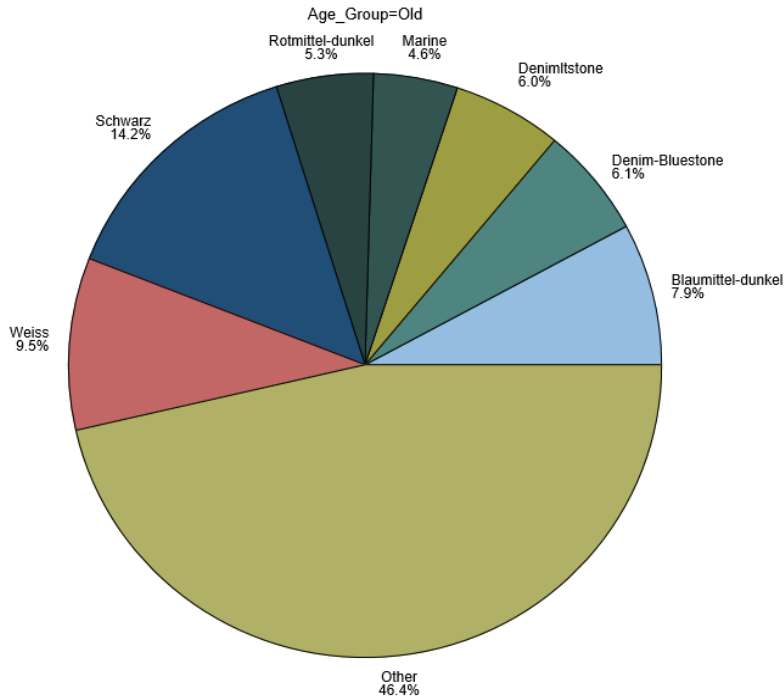


Figure 34: Pie Chart for the preferred colors for 'Old' group

Pie Chart for the preferred colors bought based on age group

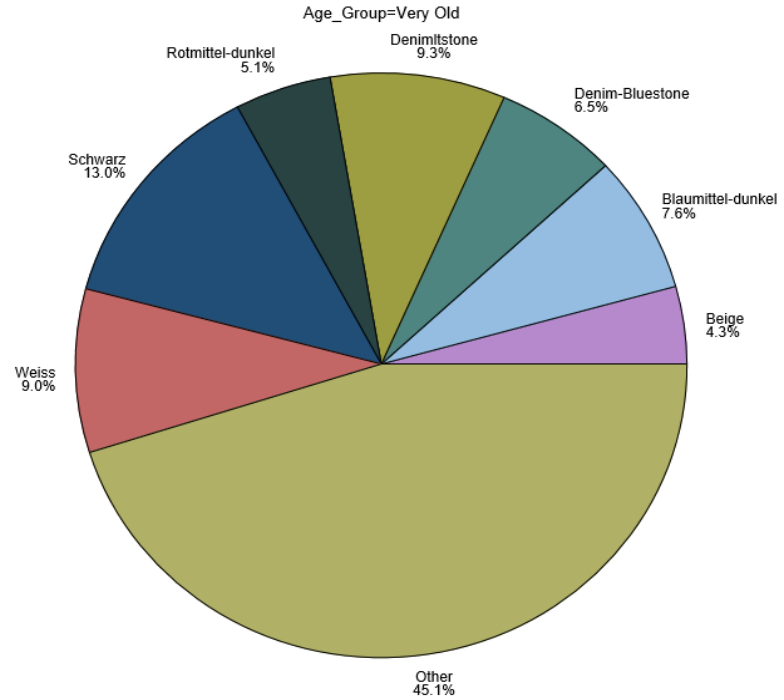


Figure 35: Pie Chart for the preferred colors for 'Very Old' group

Pie Chart for the preferred colors bought based on age group

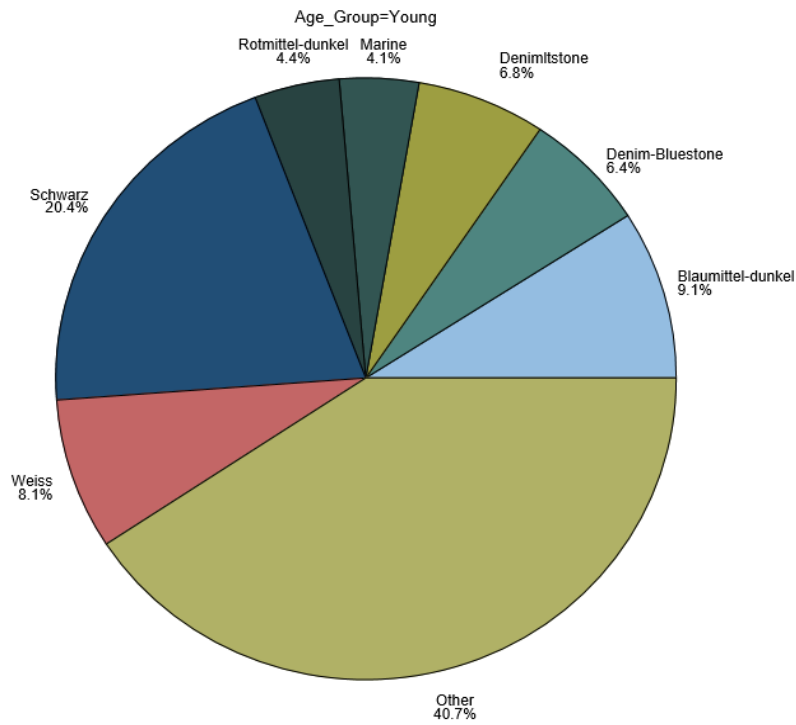


Figure 36: Pie Chart for the preferred colors for 'Young' group

Pie Chart for the preferred colors bought based on age group

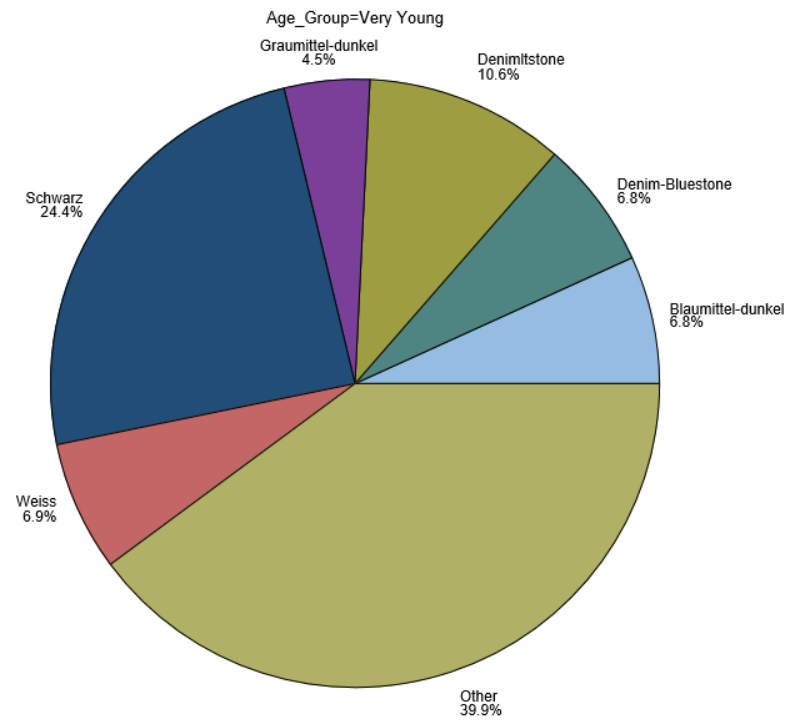


Figure 37: Pie Chart for the preferred colors for 'Very Young' group

Size distribution of women's blouses

Regarding the “damen-blusen”, we mentioned earlier that it was one of the top two most bought product categories and we also observed that for the middle age and mature groups (i.e. from 36 to 65 years old) it was found to be their most preferable category. Specifically, “damen-blusen” was the most popular choice among women, which constitute the main customer base of the store (i.e. 83% of the total customers). Nevertheless, we observed that this category was the one returned most by the customers. This, along with the fact that “damen-blusen” was the second most favored product category shows that it is an important asset for the store, which means that it a better exploitation of this products could lead to increased sales.

So, the main area we decided to focus regarding this category was the distribution of the sizes of the women's “damen-blusen” sold.

By computing some summary statistics, we noticed that the size values range from 2 to 11 with an average value of 7, which is why we decided to divide these values into the three following groups: small (sizes 2-4), medium (sizes 5-7) and large (sizes 8-11).

As shown in Figures 38 & 39, we created a pie chart and a bar chart that shown the distribution of the sizes of the women's blouses for the “damen-blusen” product category that are being sold (Note: same output, different representation). These charts, revealed the frequency and percentage of sales of the aforementioned product that belong to each size category. The same results, are also depicted in Table 22.

From these visualizations, we observed that the small, medium and large sizes accounted for about 14%, 45% and 41% respectively of the “damen-blusen” sales. We noticed that the medium and large size categories combined, represent about 86% of total sales. So, the store could ask their vendors to supply them with only these three sizes for this product, because these are the only ones that are being sold. In addition, it would be in the best interest of the store to order a lot more medium and large sizes due to the fact that these were found to be the ones in greatest demand.

Pie Chart for the distribution of the sizes sold (frequency and percent)

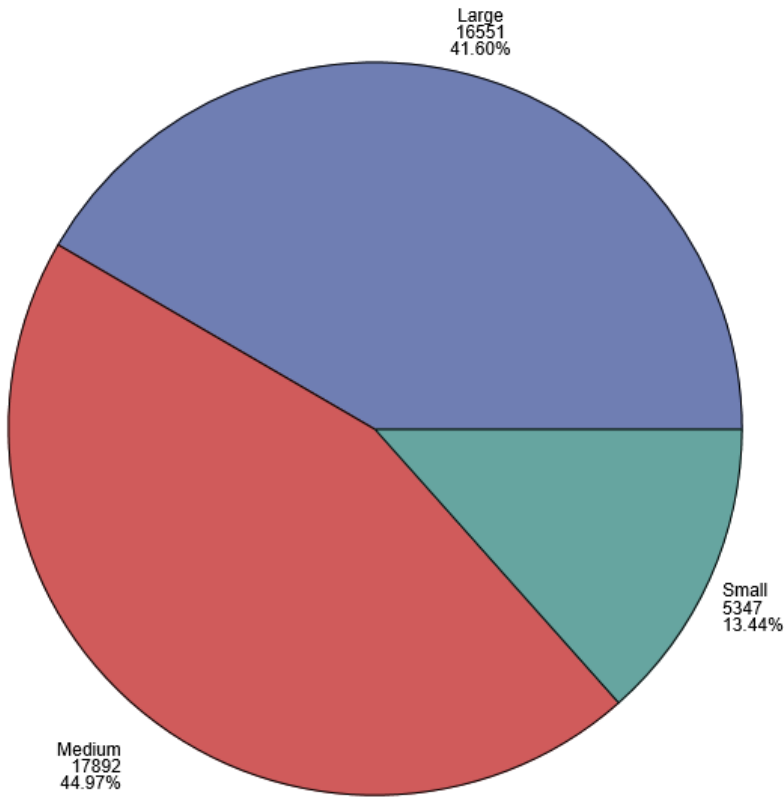


Figure 38: Pie Chart for the distribution of the sizes sold of "Damen-Blusen" (frequency and percent)

Bar Chart for the distribution of the sizes sold (frequency and percent)

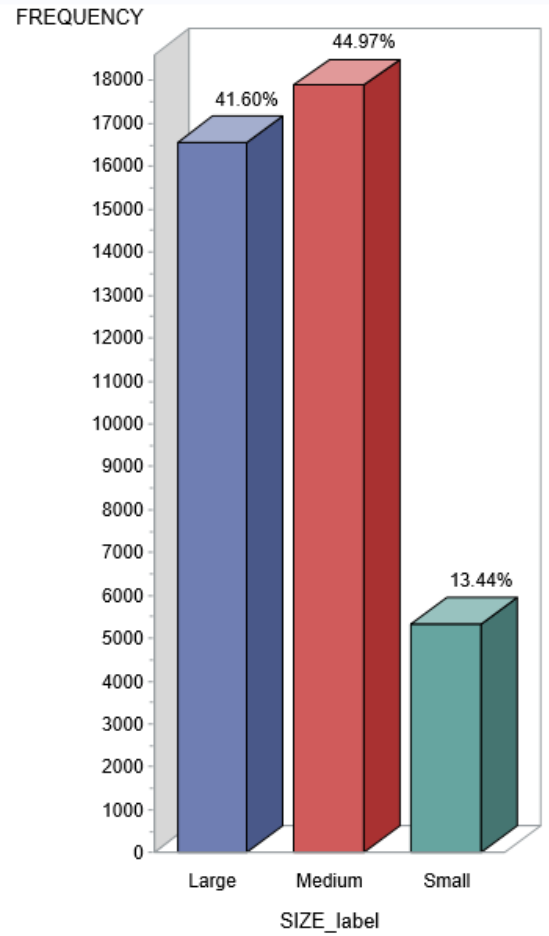


Figure 39: Bar Chart for the distribution of the sizes sold of "Damen-Blusen" (frequency and percent)

Table 22: Distribution of the sizes sold of "Damen-Blusen" (frequency and percent)

SIZE	Frequency	Percent
Small	5347	13.44
Medium	17892	44.97
Large	16551	41.60

Promotional Activities

In order to identify the day(s) at which it would be at the store's best interest to conduct promotional activities, we created a pie chart and a table that depict the number and percentage of sales during each day of the week.

From Figure 40 and Table 23, we observed that Monday was by far the day with the most sales, specifically over 32% of the customers' purchases happened in Monday. The next day with the higher sales was Tuesday (over 15% of the total sales), followed by Wednesday and Thursday (both days contributed to about 13% of the total sales). As for the days with the least sales these were Friday (~12%) and Saturday (~11%).

Before continuing with our analysis, we checked to see whether the conduction of promotional activities really affected the sales and at what extent. From Table 24, we observed that the purchases that happened from Saturday till Tuesday were influenced by the promotional activities that took place and increased the number of sales, especially during Monday which was found to be the most effective day for this kind of activities. In contrast, the sales from Wednesday to Friday were actually lower after the conducting promotional activities. We also noticed that that as the weekdays pass by, the promotional activities are less effective, as the sales percentage drops - although at a different rate - moving from Monday to Saturday.

As a result, if the store wants to increase the sales of certain products it seems that Monday is the best day to organize promotional activities as about one out of three of the store's customer base shop during Monday and considering that it was the day with the highest increase in sales after the promotion was done.

On the other hand, if the shop wants to increase the productivity of its sales during each day of the week, it has to take into account the fact that the promotion was found to be unsuccessful Wednesday till Friday. So, in order for the shop to have increased sales during every day of the week and have an effective promotional campaign, then the promotional activities should be conducted only from Saturday till Tuesday and avoided from Wednesday till Friday. We also believe that the store should start by conducting promotional activities in Saturday, which is the day of the week with the fewest sales overall, in order to motivate potential customers to visit the store on a non-working day.

Sales % per day of the week

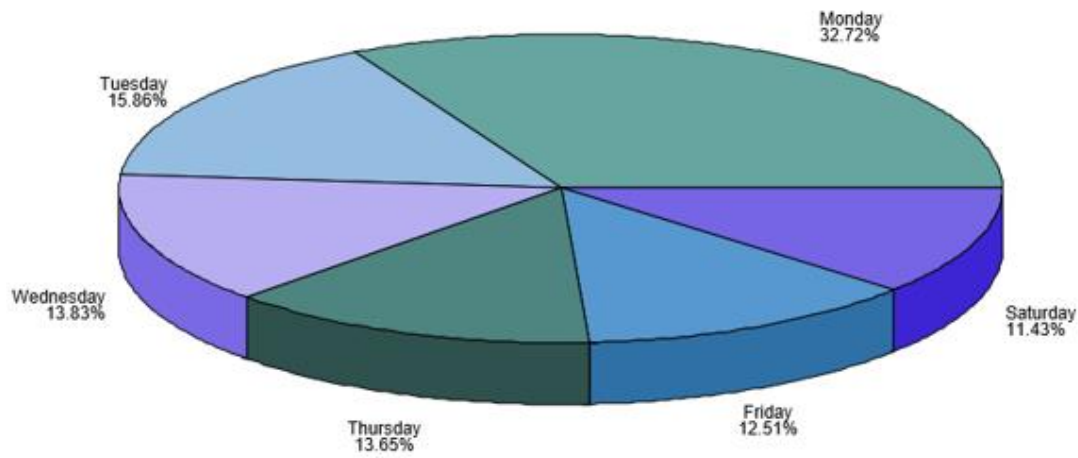








Table 23: Sales distribution per day of the week

Day_of_week	Frequency	Percent
Monday	99789	32.72
Tuesday	48379	15.86
Wednesday	42188	13.83
Thursday	41612	13.65
Friday	38164	12.51
Saturday	34845	11.43

Figure 40: Sales percentage per day of the week

Table 24: Sales distribution per day of the week based on the existence of promotional activities

Day of week	Sales % without promotion	Sales % with promotion
Monday	31.74 %	33.41 % 
Tuesday	15.12 %	16.39 % 
Wednesday	14.44 %	13.41 % 
Tuesday	14.29 %	13.19 % 
Friday	13.17 %	12.05 % 
Saturday	11.25 %	11.55 % 

After recognizing the days that are more suitable for promotional activities to take place and those that are not, we decided to see the products that are mostly bought during these days in order to analyze whether the sales of a product depends on the day of the week.

First of all, we observed from Table 25 that in total numbers, about 25,000 baskets are sold each Monday with the number keep decreasing as days of the week pass by. Specifically, during Tuesday the number dropped to about 11,500 baskets, from Wednesday to Thursday it ranged at around 10,700 baskets before dropping 1,000 more in Friday until it reached a week low of 9,000 in Saturday.

We also observed that despite the fact that the frequency of purchases decreases significantly as the week ends, the size of the basket does not have large fluctuations, with an average basket size of about 4 products.

Table 25: Number of Baskets sold and Average Basket Size per day of the week

Day_of_week	Baskets_count	Basket_items	Avg_Basket_size
Monday	25012	99769	4
Tuesday	11620	48379	4.2
Wednesday	10890	42188	3.9
Thursday	10626	41612	3.9
Friday	9618	38164	4
Saturday	9005	34845	3.9

We also created a table showing the three most frequent product categories purchased per day of week, along with pie charts depicting the most bought products per day of the week. So, from Table 26 we observed that during all weekdays the “Damen-Blusen” and “Jeanswear” were the two most preferred products by the customers, where the most units sold for the first product was during Thursday and for the second product during Tuesday (more than 4% sales compared to the other days).

Table 26: Three most frequent product categories sold per day of week

Day_of_week	PRODUCT_DESCRIPTION	COUNT
Monday	Damen-Blusen	12406
Monday	Jeanswear	11486
Monday	Damen-Shirt/Sweat	9293
Tuesday	Jeanswear	7841
Tuesday	Damen-Blusen	6192
Tuesday	Damen-Shirt/Sweat	3974
Wednesday	Damen-Blusen	5854
Wednesday	Jeanswear	5022
Wednesday	Damen-Shirt/Sweat	3634
Thursday	Damen-Blusen	5880
Thursday	Jeanswear	5075
Thursday	Damen-Hosen	3472
Friday	Jeanswear	5129
Friday	Damen-Blusen	4966
Friday	Damen-Shirt/Sweat	3096
Saturday	Damen-Blusen	4492
Saturday	Jeanswear	4035
Saturday	Damen-Shirt/Sweat	3368

In order to be more accurate and draw more useful conclusions, apart from Table 26, we also created some pie charts depicting the most bought products per day of the week.

From Figures 41-46, we observed that except from “Herren-Shirt/Sweat” that was bought most during Monday, Friday and Saturday and corresponds to about 4% of the total sales per day, the main categories bought – more than 60% of the total sales - were exactly the same for each day of the week (i.e. “Damen-Blusen”, “Damen-Shirt/Sweat”, “Jeanswear”, “Damen-Hosen”, “Damen-Kombinationen”, “Herren-Shirt/Sweat”, “MyOwn” and “THEA”).

The two most-bought product categories during all days of the week were found to be the “Damen-Blusen” and “Jeanswear” (about 13.2% and 12.8% of the total sales per each day). It was also noticed that the “Jeanswear” reached its peak in sales during Tuesday (over 16% of the total sales per day), while the sales for the other product categories were more stable and did not seem to be affected as much by the day of the week.

Furthermore, the two most bought products during all days were the “Damen-Shirt/Sweat” and the “Damen-Hosen” that corresponded to approximately 8.7% and 8% of the sales per day correspondingly. We also noticed that the first one was most bought during Monday and Saturday, while the sales of the “Damen-Hosen” was in general equally distributed along the days, reaching its peak in Thursday.

As a result we understand that there are days that favor the sale of a product, so depending on the product the store wants to promote it could pick the day to boost its sales even more (e.g. promote ‘Jeanswear’ any day of the week except from Tuesday) or the day where the increase in sales reaches its peak.

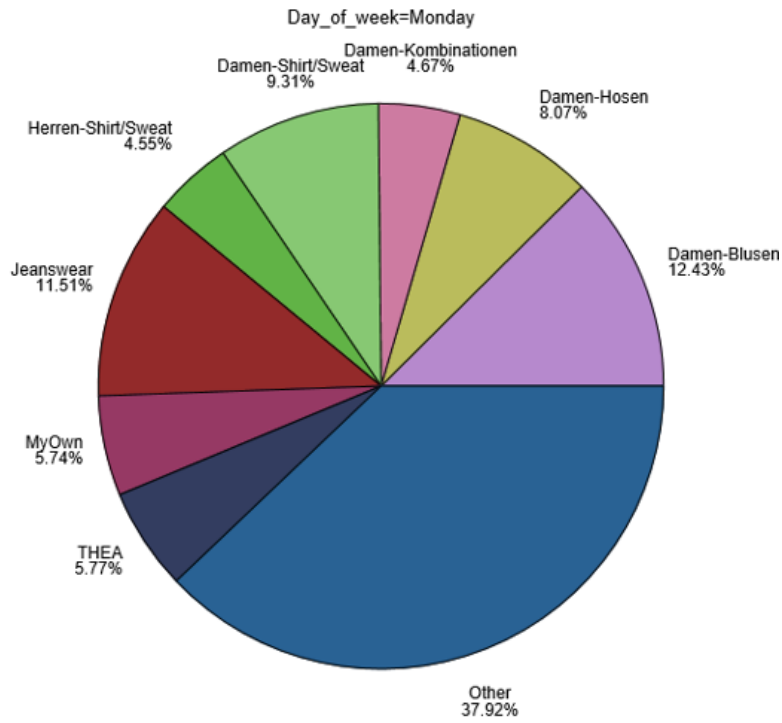


Figure 41: Most bought products during Monday

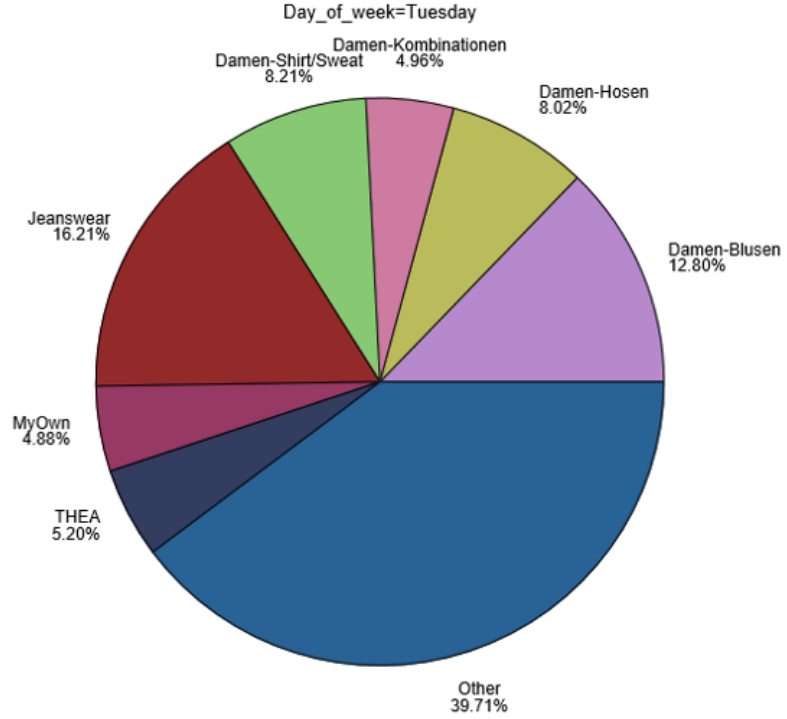


Figure 42: Most bought products during Tuesday

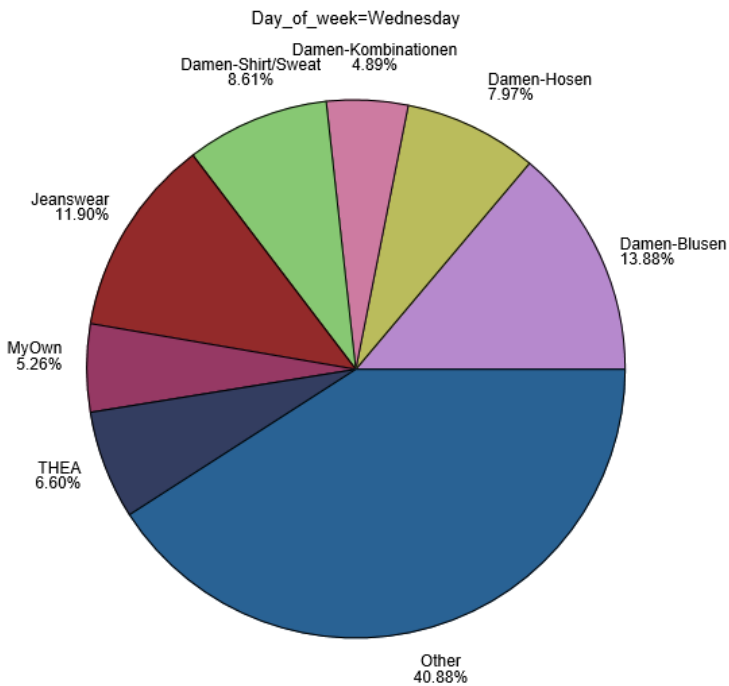


Figure 43: Most bought products during Wednesday

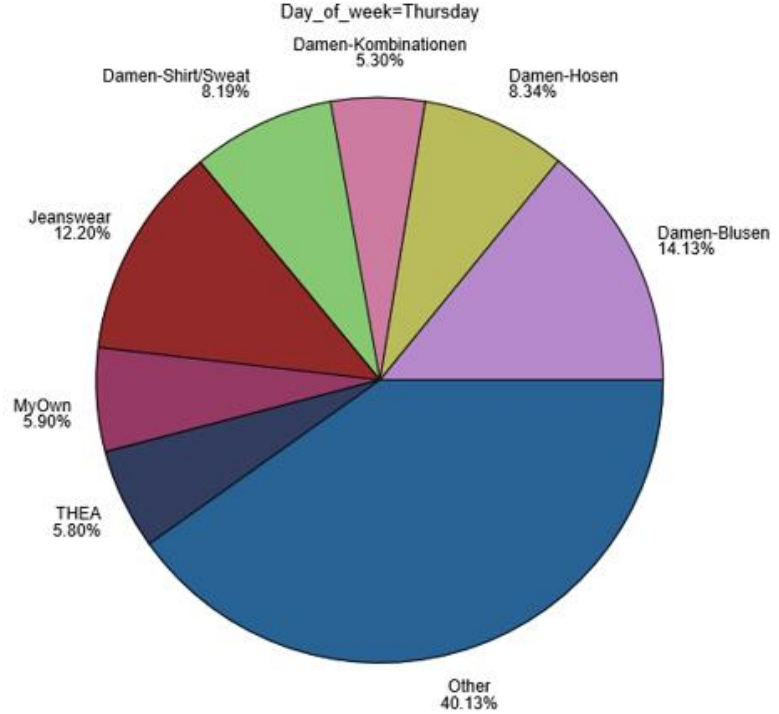


Figure 44: Most bought products during Thursday

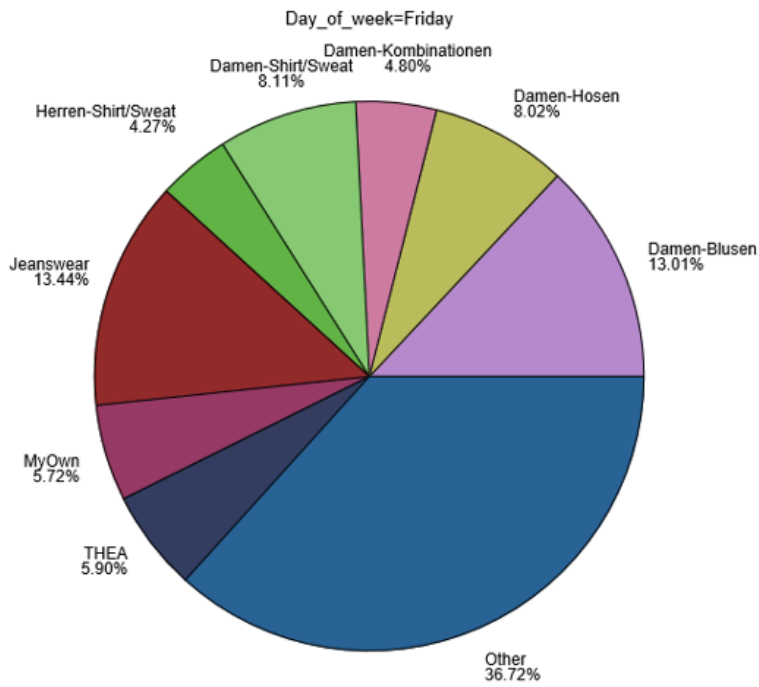


Figure 45: Most bought products during Friday

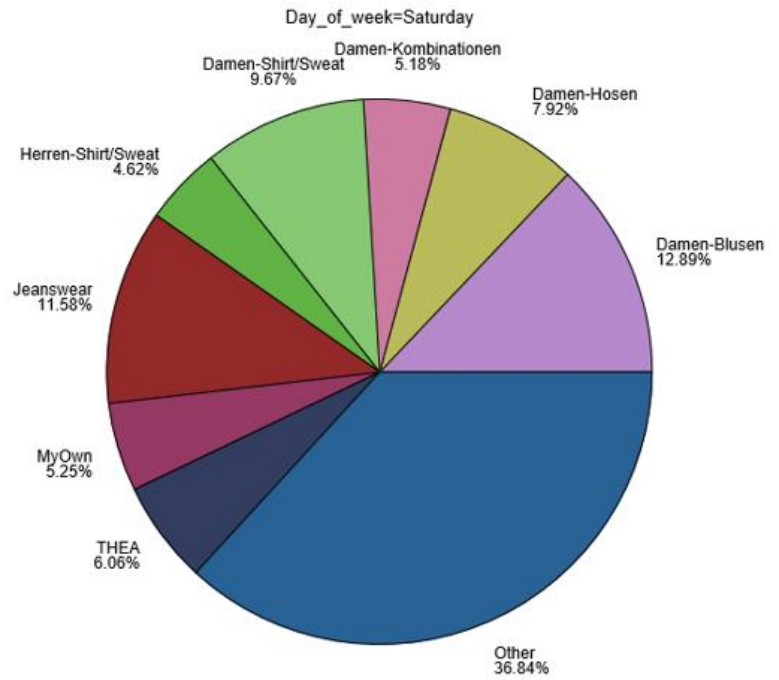


Figure 46: Most bought products during Saturday

Furthermore, we created some cross tabulation tables for each day of the week based on sex and on age group of the customers to see if there were any trends worth uncovering.

From the cross-tabulation table 27 that depicts the sales for each day of the week based on sex, we observed that the distribution of purchases of each gender is approximately the same during all days of the week. So, we do not see a clear correlation between the days that each gender decides to shop. Apart from that, there was nothing more worth mentioning from this table that has not already been discussed.

As already discussed, Monday is the day where over 32% of the total purchases are performed, followed by Tuesday where about half of Monday's sales are done and the rest of the days during which the sales are almost equally distributed. It would be interesting to see which age groups are the ones that drive the sales the most during each day. From the cross-tabulation table 28 that depicts the customers from each age group for each day of the week, we observed that for Monday for all ages the sales of products were about 30% of the total purchases, except for the adults who were under 25 years old for which the same percentage dropped about to 20%. This shows that the younger customers do not buy in the same high frequency on Monday as the other ages. In contrast, most of their purchases (over 30%) were during Thursday which was over 20% more compared to the other ages and the least products they bought was during Wednesday and Saturday (6% of their purchases) which was more than half of what the other customers have bought during the same days.

So we understand that the under 25 years old customers, buy more than 50% of their products during Monday and mainly Thursday, while avoid shopping so much during Wednesday and Saturday.

As for the other ages, we noticed that they follow similar patterns with the general population. Worth mentioning is the behavior of the older than 66 years old people, who are the ones that buy more products during Saturdays compared to the other customers. Similarly to the under 25 year old group, the customers who are older than 76 years old buy a lot more products during Tuesday compared to those aged from 26 to 75 years old.

So, if the company wants to promote clothes which are mostly chosen by the under 25 years old customers, it should perform its activities during Thursday, while for the other groups Monday is the most wise choice. Now, if the store wants to promote products more than one day of the week, the appropriate choice for the younger customers is Monday and for the other customers Tuesday would be a really good choice.

Table 27: Cross- tabulation table for sales
per day of the week based on sex

Frequency Percent Row Pct Col Pct	Table of Day_of_week by SEX			
	Day_of_week	SEX		
		W	M	Total
Monday		80761	19008	99769
		26.48	6.23	32.72
		80.95	19.05	
		32.56	33.40	
Tuesday		38513	9866	48379
		12.63	3.24	15.86
		79.61	20.39	
		15.53	17.34	
Wednesday		34721	7467	42188
		11.39	2.45	13.83
		82.30	17.70	
		14.00	13.12	
Thursday		34248	7364	41612
		11.23	2.41	13.65
		82.30	17.70	
		13.81	12.94	
Friday		31198	6966	38164
		10.23	2.28	12.51
		81.75	18.25	
		12.58	12.24	
Saturday		28607	6238	34845
		9.38	2.05	11.43
		82.10	17.90	
		11.53	10.96	
Total		248048	56909	304957
		81.34	18.66	100.00

Table 28: Cross- tabulation table for customers
from each age group for each day of the week

Frequency Percent Row Pct Col Pct	Table of Day_of_week by Age_Group							
	Day_of_week	Age_Group						Total
		Mature	Middle Age	Old	Young	Very Old	Very Young	
Monday		54421	27393	11517	3615	2606	217	99769
		17.85	8.98	3.78	1.19	0.85	0.07	32.72
		54.55	27.46	11.54	3.62	2.61	0.22	
		33.47	33.64	29.56	28.74	31.12	21.38	
Tuesday		25670	12544	6129	2202	1642	192	48379
		8.42	4.11	2.01	0.72	0.54	0.06	15.86
		53.06	25.93	12.67	4.55	3.39	0.40	
		15.79	15.41	15.73	17.51	19.61	18.92	
Wednesday		22427	11168	5419	2198	906	70	42188
		7.35	3.66	1.78	0.72	0.30	0.02	13.83
		53.16	26.47	12.84	5.21	2.15	0.17	
		13.79	13.72	13.91	17.47	10.82	6.90	
Thursday		21706	11143	5631	1894	923	315	41612
		7.12	3.65	1.85	0.62	0.30	0.10	13.65
		52.16	26.78	13.53	4.55	2.22	0.76	
		13.35	13.69	14.45	15.06	11.02	31.03	
Friday		20194	10246	5034	1284	1251	155	38164
		6.62	3.36	1.65	0.42	0.41	0.05	12.51
		52.91	26.85	13.19	3.36	3.28	0.41	
		12.42	12.58	12.92	10.21	14.94	15.27	
Saturday		18185	8926	5238	1385	1045	66	34845
		5.96	2.93	1.72	0.45	0.34	0.02	11.43
		52.19	25.62	15.03	3.97	3.00	0.19	
		11.18	10.96	13.44	11.01	12.48	6.50	
Total		162603	81420	38968	12578	8373	1015	304957
		53.32	26.70	12.78	4.12	2.75	0.33	100.00

Product Launch

In order to check whether there is seasonality in the sales of the product categories and thus be able to propose the product categories that would be worth to be launched for the autumn - winter 2018 season, we created pie charts that describe the product categories sold every quarter of the year (see Figures 47-50).

First of all, we observed from these figures that the most-bought product during the first and fourth quarter of the year was "Jeanswear" (~14% of sales) and that during the 2nd quarter it was "Damen-Blusen" (~15% of sales), just like in the 3rd quarter (closely followed by "Jeanswear"). Specifically, these two categories were among the three most bought product categories during all quarters so they should be the first to be launched for the new season.

However, the quarter of the season each product is launched should be carefully selected. For "Jeanswear" we observed that from April through June the percentage of its sales dropped more than 3%, in contrast with the other quarters where the discrepancies between sales were very small. During this period (i.e. 2nd quarter), the product that reached its peak was "Damen-Shirt/Sweat" (more than 6% increased sales compared to the other quarters). As for "Damen-Blusen", its sales decreased more than 5% during the October-March period compared to April through September.

So, a potential strategy for the company would be to make sure that it has enough stock of "Jeanswear" during the first and the last quarter of the year and from March till September it should concentrate on the sales of "Damen-Blusen". Additionally, from March till June, the company should also further promote "Damen-Shirt/Sweat", "Damen-Hosen" and "Damen-Kombinationen" products that reach their peak in sales during this period (about 13%, 9% and 7% correspondingly) and from July to September focus more on "Jeanswear".

Furthermore, if the company wants to promote some other products alongside the previously mentioned, it should notice that the "MyOwn" and "Damen-Jacket/Mantel" product categories are mostly bought (more than 7% of the total sales) during the winter (i.e. October to December). As, for the "Herren-Shirt/Sweat" and "THEA" products, we observed that are preferred all year round by customers as their sales have very few differences throughout the year.

Pie Chart for product categories sold per quarter

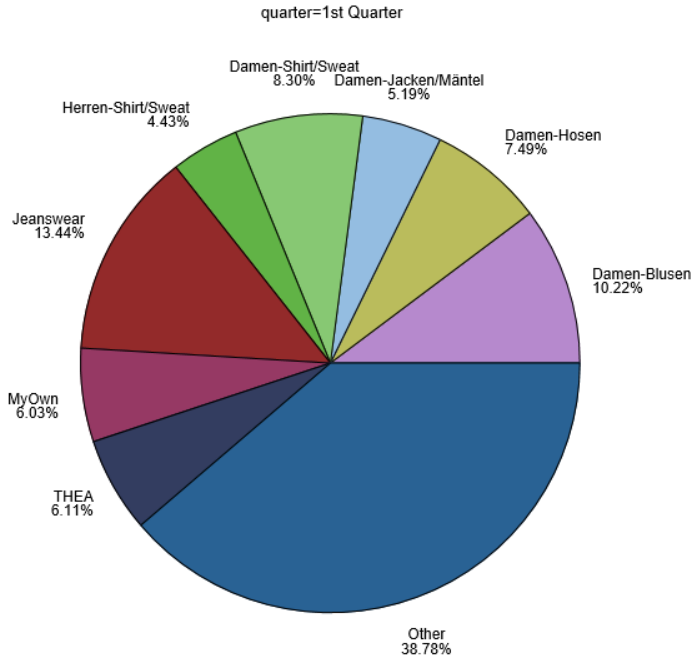


Figure 47: Pie Chart for product categories sold for 1st Quarter of the year

Pie Chart for product categories sold per quarter

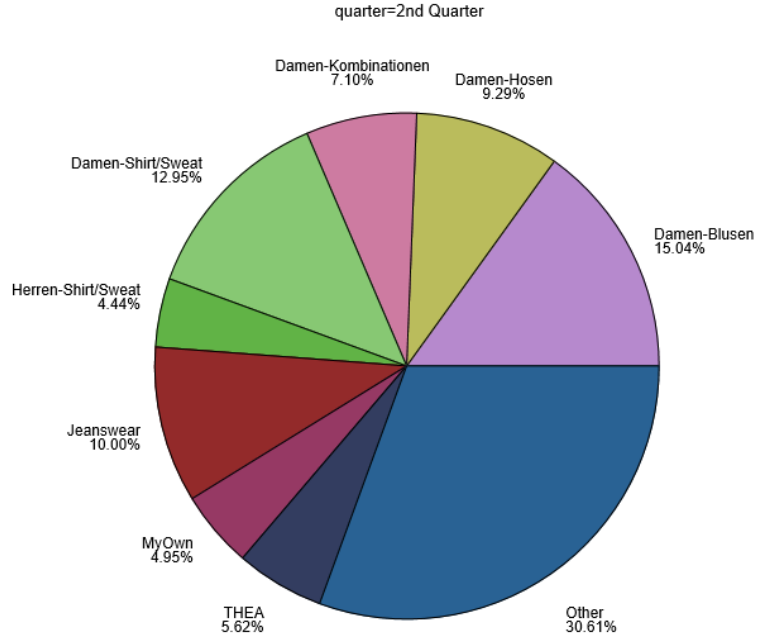


Figure 48: Pie Chart for product categories sold for 2nd Quarter of the year

Pie Chart for product categories sold per quarter

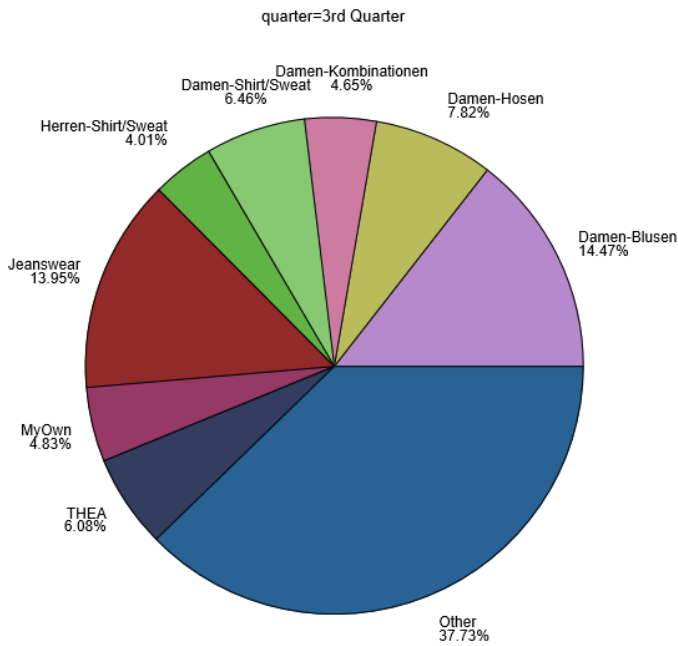


Figure 49: Pie Chart for product categories sold for 3rd Quarter of the year

Pie Chart for product categories sold per quarter

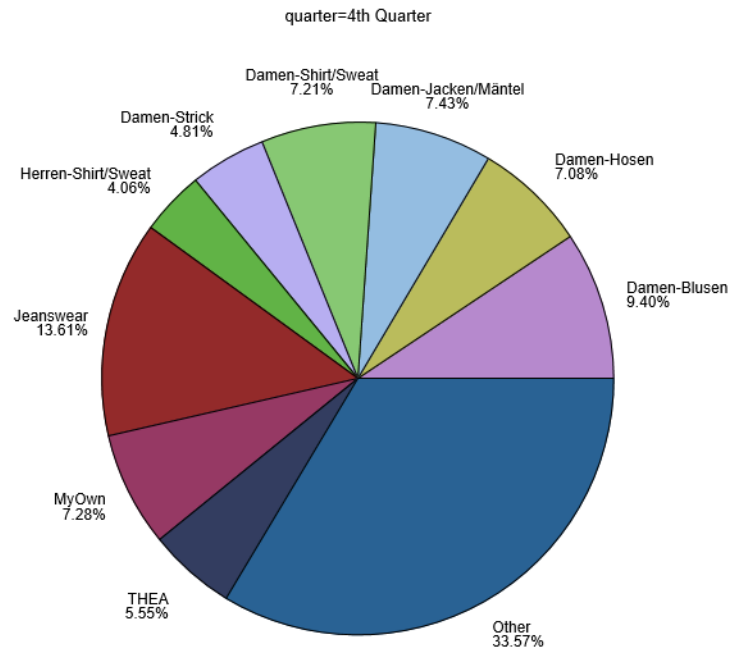


Figure 50: Pie Chart for product categories sold for 4th Quarter of the year





Customer Segmentation

The management team of the organization suggested an RFM model to perform the segmentation of the customers. RFM (Recency, Frequency, Monetary) analysis is a customer segmentation technique that uses past purchase behavior to divide customers into groups. These RFM metrics are important indicators of a customer's behavior because frequency and monetary value affects a customer's lifetime value, and recency affects retention, a measure of engagement. In order to implement the RFM analysis, we had to first create the RFM dataset by calculating those three parameters for each customer.

At first, we calculated the total number of purchases for each customer to get the 'Frequency' value and then we summed the amount of money each customer spent for his/her purchases to get the 'Monetary' value. Finally, we identified the most recent date that each customer issued a receipt from the store and then we used this date to calculate the months passed from the most recent transaction date till '01Jan2018' in order to find the 'Recency' value for each customer.

Before starting the analysis of the RFM dataset, we used the interquartile range via the 1.5 IQR rule of thumb in order to detect potential outliers². After implementing this rule, we identified 4,756 outliers which we removed from the data. As a result, the RFM dataset that we have now in our disposal contains 41,983 observations instead of 46,739. In Table 29, a part of them newly created RFM data set is presented.

Table 29: First 5 rows of the RFM data set created

 ID	 M	 F	 R
10231824	11	1	35
10239214	81	3	37
10239691	13	1	37
10241396	54	1	30
10242027	412	9	31










² An observation with fitted value away from the actual.

Customer Profiling

The management team of the organization wants to exploit the available data to profile its customers based on their importance in order to offer them personalized services and products. The profiling of the customers was done by using two algorithms. Firstly, Ward algorithm was used to identify the optimal number of clusters for the customers based on the RFM variables and then k-means was implemented in order to place the customers into these clusters.

By executing the Ward algorithm, we found that the optimal number of clusters was equal to three. Afterwards, we applied the k-means algorithm in order to create the three segments and then identify the cluster that each customer belongs to. The results are depicted in the table below, where we see a typical customer that belongs to each segment and the typical customer of the whole customer base (i.e. the averages for the characteristics of the customers that belong to each cluster and overall).

Table 30: Comparative table of the segments

Segment	Recency (Average)	Frequency (Average)	Monetary (Average)	Segment Description	Number of Customers in Segment	Segment Size
1	32.4 	13.5 	616.6 	Best	5,899	14.05%
2	31.9 	6.4 	288.2 	Valuable	13,149	31.32%
3	35.4 	2.4 	103.8 	Worst	22,935	54.63%
Whole Customer Base	33.8	5.2	229.6	-	41,983	100%

From Table 30, we see that the three segments created, which describe essentially three different customer types, share common characteristics among the customers that belong to them. Specifically, a customer who belongs to the first segment is expected to spend on average 616.6 €, buy 13.5 times from the store and its last purchase was about 2 years and 8 months ago (i.e. recency=32.4). Similarly, we understand the purchasing behavior of the customers that belong to the other two segments.

One detail that should be elaborated is that in the table there is an extra row that is declared as overall, which shows the average numbers of all the customers. So, we observed that a typical customer of the store spends on average 229.6 €, buys 5.2 times from the store and its last purchase was about 2 years and 10 months ago (i.e. recency=33.8).

Inside Table 30, we also placed some green and red arrows inside the Recency, Frequency and Monetary columns, which reflect the comparison of average customer that belongs to each segment with the typical customer overall. We used these arrows as indications that will allows us to have a broader understanding of the characteristics of each segment. This will help us in attempting to give a proper description (see 'Segment Description' column) and in suggesting

suitable strategies for each one of the segments. In general, it would be preferable for the store to have customers with a small recency value and high frequency and monetary values.

The rationale behind the placement of the arrows was the following: If a segment has better characteristics than those of the average customer of the whole customer base we use a green upwards arrow in that column, but if the segment has worsened characteristics compared to the average we put a red downwards arrow in that column.

From the table we observed that the customers that belong to the second segment spend more (288.2>229.6), more frequent (6.4>2.4) and more recently (31.9<33.8) compared to the average customer of the whole customer base. The first segment seems have the same behavior as the second. Nevertheless, despite the fact that there is a slight discrepancy in their recency status, the other two features are much better which was the reason why we labeled the customers that belong to the second segment as 'valuable' and considered those that belong to the first segment as the best ones.

As for the customers described by the third segment we deemed them as the worst customer type, because we observed that they spend a small amount of money in their purchases, buy rarely and a lot of time has been passed since their last purchase.

The reason why the segmentation of the customers is useful is due to the fact that we are now able to suggest (potential) strategies that would be appropriate for each customer type and could prove very useful for the organization.

Specifically, we observed that the **valuable customers** are very active. As a result, we can claim that their proper utilization can generate more revenue, which is why the store should make sure to keep these kind of customers as satisfied as possible. Examples of potential marketing actions that could be conducted by the store are a personal interaction from the store representatives with them and engagement activities such as subscription to the newsletter of the store or asking them for product reviews.

As for the **worst** customers, which could also be considered as lost customers, there is not much that can be done. These customers generate only a tiny amount of revenue. Our personal belief is that the company should not spend any resources on them as it is highly unlikely to win them back, but even if that happens they will not spend any significant amount of money for their purchases or buy on a regular basis.

Finally, there are those we considered as the **best** customers of the store, who corresponded to a smaller proportion of the company's clientele (about 14% of the total customers). We observed that they had the highest frequency and monetary scores as well as very high recency, thus it is of pivotal importance for the store to put a major amount of effort into keeping them satisfied in order to make as sure as possible that their current purchasing behavior is preserved.

Possible customer retention strategies that the store could implement would be to provide this type of customers with extra benefits compared to the regular ones such as access to limited product codes, special discounts to make them feel valuable or even use them as ambassadors to promote the website. In addition, they could also be asked for regular feedback, as they may have a better view regarding some products even from store managers.

Furthermore, we wanted to describe the age and city demographic data regarding the customers that belong to the two most important clusters previously created (i.e., 'best', 'valuable' segments).

Regarding the distribution of the company’s customers, we observed that from the total 79,835 transactions performed by those that belong to the “best” cluster approximately 13.8% came from men and 86.2% came from women. Similarly, from the customers identified as ‘valuable’, men performed about 14.8% of the total 84,576 transactions while women were responsible for the rest 85.2%. A visual representation of these percentages is presented in Figures 51 & 52 below.

Distribution of Sex in most important cluster

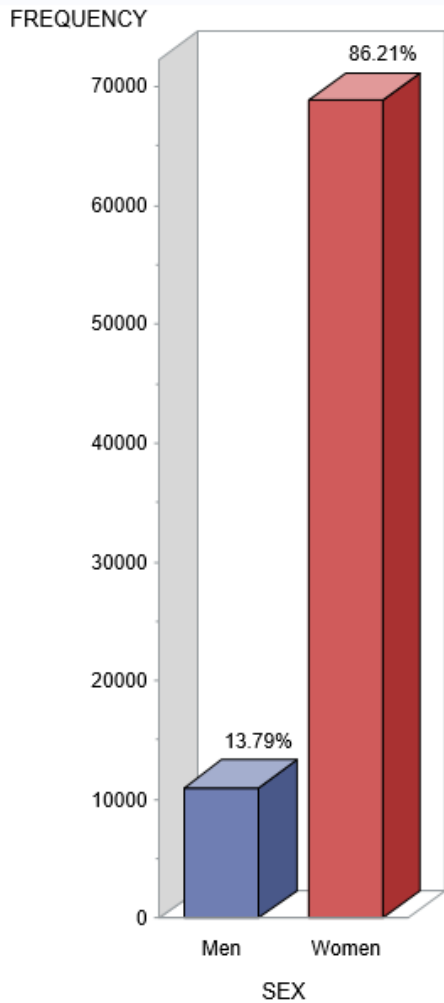


Figure 51: Bar chart for the distribution of sex in the ‘best’ cluster

Distribution of Sex in 2nd most important cluster

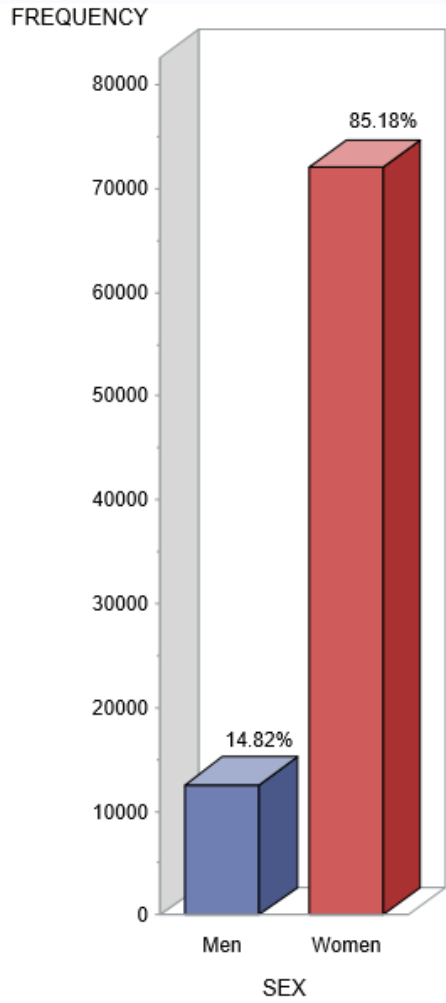


Figure 52: Bar chart for the distribution of sex in the ‘valuable’ cluster



Figure 53: Histogram showing the distribution of sex in the 'best' cluster

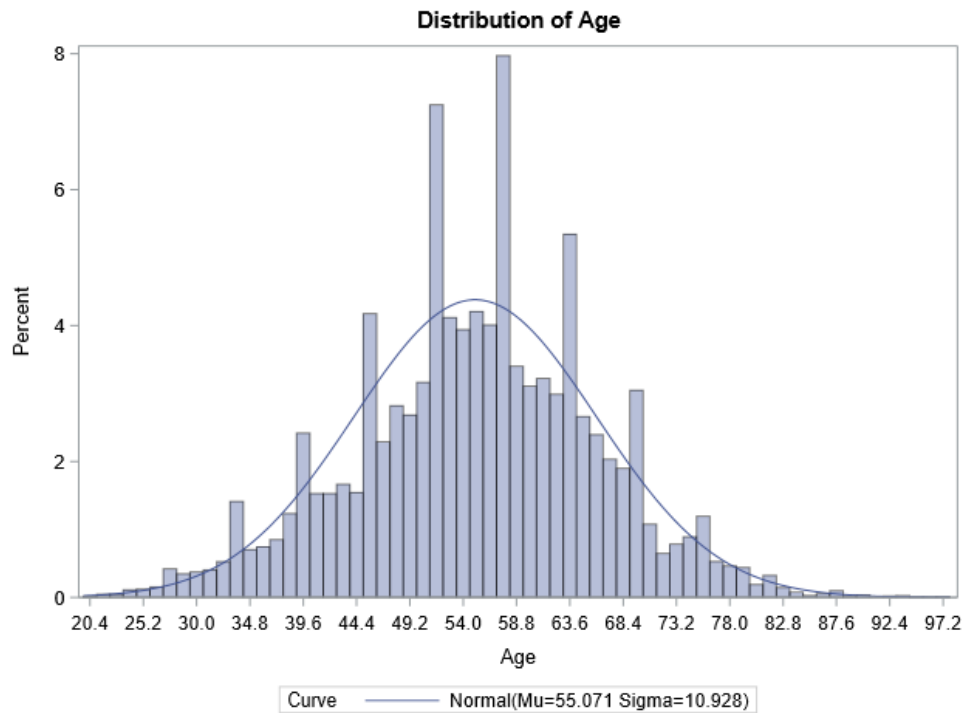


Figure 54: Histogram showing the distribution of sex in the 'valuable' cluster

As for the customers' age, we observed from Figures 53 & 54 that for the two most important clusters the age variable seems to follow a normal distribution, which means that the main customer base for the company are the mature customers (i.e. aged from 51 to 65 years old). We also saw that the mean age from the 'best' customers was about 53.9 (with a standard deviation of 10.6 years), while for the 'valuable' customers the mean age was equal to 55 years old (with a standard deviation of 10.9 years).

In order to grasp a better idea and for easier comprehension, we also reviewed the distribution based on the age groups. From Figures 55 & 56, we observed that the people who are between 36 to 75 years old correspond to almost 92% of these types of the company's clients - with the main customer base (i.e. more than 50%) consists of 51 to 65 years old customers - and that the younger age groups (i.e. 18 to 35 years old) corresponded to less than 5% of the company's customers.

The main difference among the two most important clusters was found regarding the customers between 36 to 50 years old and the older customers (i.e., 66 to 75 years old). Specifically, the middle-aged customers that belonged to the 'best' cluster constituted more than 4% of the total customers compared to those from the 'valuable' segment, while the valuable customers aged between 66 to 75 years were 3% more compared to the best ones.

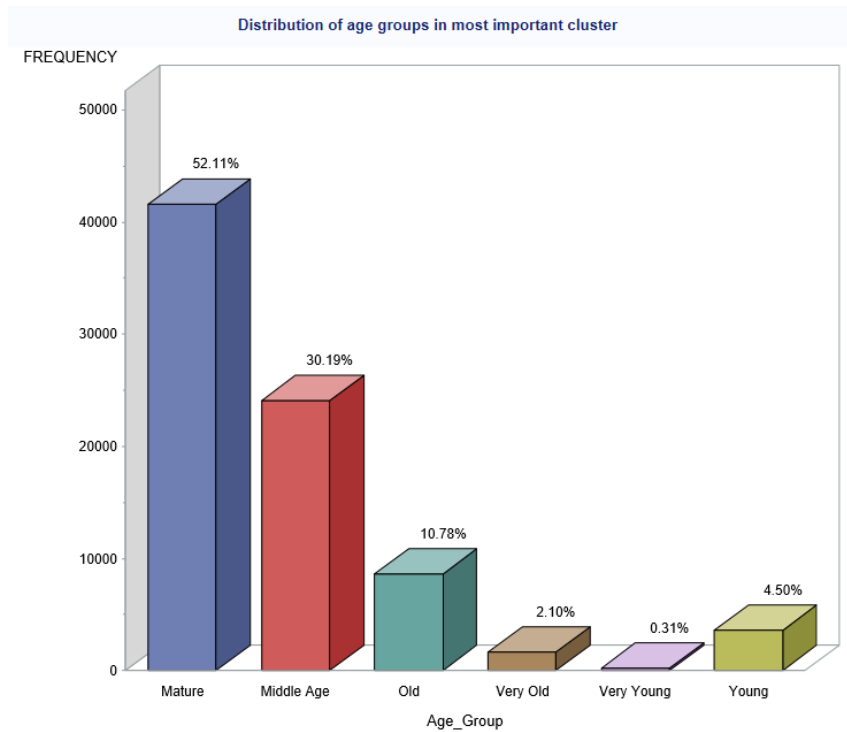


Figure 55: Bar plot for the distribution of age groups in the 'best' cluster

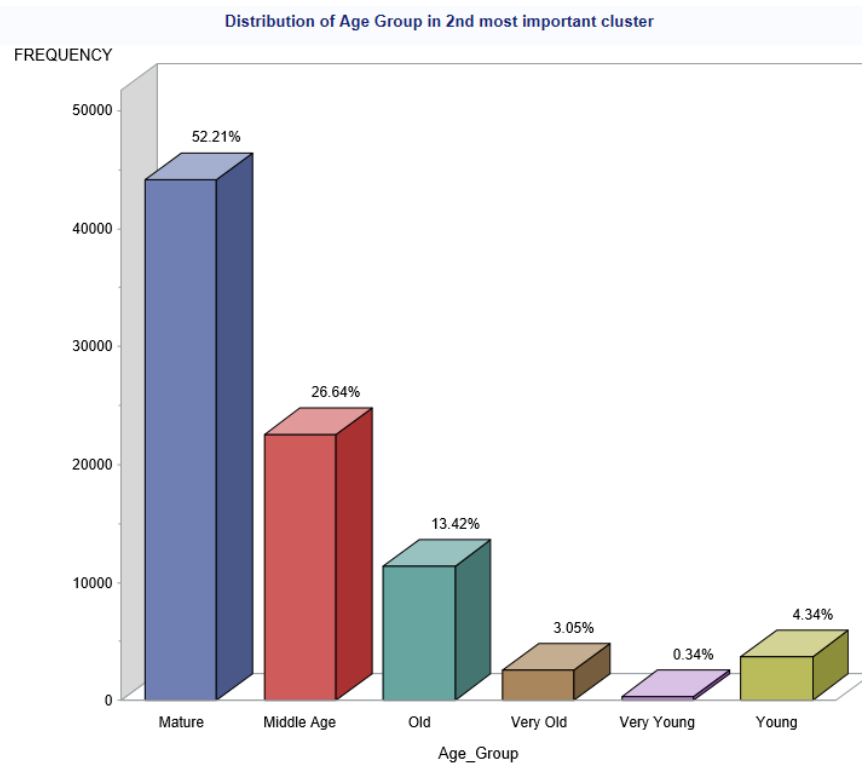


Figure 56: Bar plot for the distribution of age groups in the 'valuable' cluster

Product Associations

The company wanted to change internally the store based on the products that tend to be bought together. In order to help the company in this regard and identify the product categories that are bought together (i.e. association of product categories), we implemented a market basket analysis by using association rules, which is a very popular mining method.

The associations among product categories were firstly identified based on the whole dataset and then in regard with the two most important clusters. After running the model with the association's rules, we compared the calculated metrics (i.e., support, confidence and lift) in order to determine which products should be suggested to each customer.

We observed that the whole dataset contained information about 220,144 past transactions that referred to 41,983 different customers and to all of the 42 available product categories. From these transactions, we found that 79,835 belong to the 'best' cluster which describes 5,899 different customers and 84,576 past transactions came from the customers that belong to the 'valuable' cluster to which 13,149 customers belong. We also saw that both these clusters contain information about all the 42 different product categories.

After creating the association rules via the SAS Enterprise Miner, we sorted the results in descending order based on the lift metric in order to refer to product categories that are more correlated than expected. The lift measure shows the strength of the association and is computed as the ratio of the confidence of the rule and the expected confidence of the rule. The confidence value is defined as the ratio of the support of the joined rule body and rule head divided by the support of the rule body, while the expected confidence of a rule is defined as the product of the support values of the ruling body and the rule head divided by the support of the ruling body.

The lift takes values between 0 and infinity, where a lift value near 1 indicates that there is no relationship between the compared products, a lift smaller than 1 implies that there is a negative relationship between the compared products (i.e., they occur together less often than random) and a lift value greater than 1 indicates that there is a positive relationship between the compared products (i.e., they occur together more often than random).

So, it is clear that in order to make the best proposals to the store we have to use the associations with the higher lift values. Also, it has to be noted that the support and lift are symmetric measures, while confidence is asymmetric. As a result, we will present the lift value for the unique product categories contained in the association rules.

Below, we presented a table that contains the association rules with the top-3 higher lift values in descending order and was created based on the whole customer base of the store.

Table 31: Association rules for the top-3 higher lift values (created based on the whole customer base)

RULE	LIFT
ViaCortesaDOB → MyOwn	2.53
MarkenshopsDobklassisch → Damen-Kombinationen	2.24
MyOwn & Damen-Blusen → Damen-Shirt/Sweat	2.12

From Table 31, we understand that the customers who bought 'ViaCortesaDOB' product category are 2.53 times more likely to purchase the 'MyOwn' product compared to a random customer and via versa. Similarly, for the customers who have purchased the 'MarkenshopsDobklassisch' product it is 2.24 times more likely to also purchase 'Damen-Kombinationen' than a random customer. Also, for the customers purchased 'MyOwn' and 'Damen-Blusen' product categories the likelihood of also buying 'Damen-Shirt/Sweat' is 2.12 times higher than the one of a random customers that visits the store.

As a result, we can claim that if the store wants to promote the 'MyOwn' product, then it should check to see which customers have bought the 'ViaCortesaDOB' product category and then promote it to these customers. In the same rationale, for the customers who have bought 'MarkenshopsDobklassisch' the recommendation should relate to the 'Damen-Kombinationen' product category and for the customers who have purchased both 'MyOwn' and 'Damen-Blusen' product categories the most suitable suggestion would be the 'Damen-Shirt/Sweat'.

As an extra note, we observed that there were 716 transactions (.3%) related to the purchase of both 'ViaCortesaDOB' and 'MyOwn' product categories (21st most bought combination of products), 543 transactions for the customers that have bought both 'MarkenshopsDobklassisch' and 'Damen-Kombinationen' and 647 transactions for those who have bought 'MyOwn', 'Damen-Blusen' and 'Damen-Shirt/Sweat' product categories.

Now, we will examine the association rules created regarding the customers that were identified as the best ones. The association rules with the top-3 higher lift values can be seen in the following table.

Table 32: Association rules for the top-3 higher lift values (created based on the best customers)

RULE	LIFT
Jeanswear → Herren-Shirt/Sweat	3.12
ViaCortesaDOB → MyOwn & Damen-Shirt/Sweat	2.24
Damen-Kombinationen → MarkenshopsDobklassisch	2.22

From Table 32, we observed that the most important association of product categories among the best customers concerns the purchase of jeans. Specifically, the best customers who have bought 'Jeanswear' are 3.12 times more likely to also purchase the Herren-Shirt/Sweat product category in comparison with a random customer and via versa. We also saw that, from the best customers who have purchased the 'ViaCortesaDOB' product it is 2.24 times more likely to also purchase both 'MyOwn' and 'Damen-Shirt/Sweat' product categories compared to a random customer. Furthermore, one of the best customers that has purchased the 'Damen-Kombinationen' product has 2.22 higher likelihood to buy also 'MarkenshopsDobklassisch' than a random customer that visits the store.

As a result, we understand that if the store wants to target the best customers, it should start with those who buy 'Jeanswear' and promote to them the 'Herren-Shirt/Sweat' product category. Furthermore, other good suggestions would be to select the customers that have bought 'ViaCortesaDOB' and for them the recommendations provided should relate to the 'MyOwn' and

'Damen-Shirt/Sweat' product categories as well as for the customers who have purchased 'Damen-Kombinationen' the most suitable suggestion would be the 'MarkenshopsDobklassisch'.

Finally, we will examine the association rules created regarding the customers that were identified as valuables for the store. The association rules with the top-3 higher lift values can be seen in the following table.

Table 33: Association rules for the top-3 higher lift values (created based on the valuable customers)

RULE	LIFT
ViaCortesaDOB → MyOwn	2.05
MarkenshopsDobklassisch → Damen-Kombinationen	1.81
MyOwn → MarkenshopsDOBmodisch	1.76

First of all, we observed from Table 33 that the two most important associations of product categories among the valuable customers were the same as those we identified in the whole customer base, but with a decrease in the lift value. So, we may say that for some product categories the purchasing behavior of the valuable customers seems to reflect the behavior of a typical customer of the store.

More specifically, we saw that the valuable customers who have bought 'ViaCortesaDOB' product category are 2.05 times more likely to purchase the 'MyOwn' product compared to a random customer and via versa. Similarly, for the valuable customers who have purchased the 'MarkenshopsDobklassisch' product it is 1.81 times more likely to also purchase 'Damen-Kombinationen' than a random customer. Also, for valuable customers that have purchased the 'MyOwn' product have 1.76 higher likelihood to buy also 'MarkenshopsDOBmodisch' than a random customer that visits the store.

As a result, we understand that if the store wants to target the valuable customers, it should start with those who have bought the 'ViaCortesaDOB' product category and promote to them the 'MyOwn' product category. In the same rationale, for the valuable customers who have bought 'MarkenshopsDobklassisch' the recommendations provided should relate to the 'Damen-Kombinationen' product category and for the valuable customers who have purchased 'MyOwn' product category the most suitable suggestion would be the 'MarkenshopsDOBmodisch'.

The full results regarding the association rules we created via the SAS Enterprise Miner can be seen in the Excel files we attach below.

Rules table created based on the whole customer base



Rules_Table_CustomerBase.csv

Rules table created based on the most important cluster (i.e. 'best')



Rules_Table_BestCluster.csv

Rules table created based on the second most important cluster (i.e. 'valuable')



Rules_Table_NextBestCluster.csv