Mai 4th , 2022

PY-Sales-Optimizer

Benjamin Eberwein

Henry Ye

Michael Kroschel

Xiuli Dong

Table of contents

Introduction

It is a well-known fact that customer targeting and personalized communication are playing an increasingly important role in successful marketing strategies. Especially in the e-commerce sector where companies usually compete against a vast number of competitors, it is of vital importance to approach the right customer with the right message at the right point of time.

To achieve this goal, leading companies continuously collect customer information and apply different segmentation techniques to separate and approach clients according to their interests and needs. One of those famous segmentation methods is known as RFM segmentation. RFM stands for recency, frequency and monetary and refers to the classification of customers based on their historic purchase behaviour.

With the present document we aim to give a practice-oriented introduction to the RFM concept and provide profound analyses of outcomes and usability of different RFM-based segmentation methods. A special focus of the work is dedicated to the application of Machine Learning algorithms and how they support and improve the “classical” clustering framework. Based on these insights we develop a marketing strategy for customer targeting which aims to minimize costs and maximize conversion of a potential marketing campaign.

The document utilizes a public available dataset collected from various Pakistan e-commerce merchants and is structured in the following way:

- The first part gives a detailed description of the dataset and its variables. An additional focus is set on the evaluation of data quality aspects and first pre-processing steps.
- The second chapter further explores the characteristics and dependencies of selected variables which are specifically relevant for the RFM-segmentation.
- The third section applies different RFM-segmentation methods with a strong focus on Machine Learning concepts. It provides a deep-dive analyses on the outcomes and explores the usability for developing a marketing strategy.
- The forth part creates a dedicated proposal for a marketing campaign based on the ”best” model derived from chapter three. It includes the description of the customer target groups, the corresponding individualized offer strategy and the best moments for activation.

1 Data description

The underlying dataset, used for the following analyses, consist of online transactions from several Pakistani e-commerce merchants. The transactions were conducted between July, 2016 and August, 2018.

The dataset is publicly available under the following link:
https://www.kaggle.com/datasets/zusmani/pakistans-largest-ecommerce-dataset

The data is organized in 1.048.575 rows and 26 columns where each row relates to a single transaction. The table below summarizes the corresponding attributes and provides a short description:

#	Name	Type	Description
1	item_id	float64	Unique identifier related to individual transactions
2	status	object	Status of transaction (eg. completed, cancelled)
3	created_at	object	Date of transaction
4	sku	object	Product identifier
5	price	float64	Price of product*
6	qty_ordered	float64	Amount of products per transaction
7	grand_total	float64	Total price of transactions minus discounts (per order)*
8	increment_id	object	Unique order identifier (order may consist of several transactions)
9	category_name_1	object	Name of product category
10	sales_commission_code	object	Commission identifier
11	discount_amount	float64	Discount granted (per transaction)*
12	payment_method	object	Type of payment (eg. cob, internetbanking)
13	Working Date	object	Equals “created_at”
14	BI Status	object	Grouping of “status” in contextual clusters
15	MV	object	Total price of transaction (“price“ times ”qty_ordered”)*
16	Year	float64	Year in which transaction took place
17	Month	float64	Month in which transaction took place
18	Customer Since	object	Year and Month of first transaction (per customer)
19	M-Y	object	Month and year of transaction in “M-Y” format
20	FY	object	Refers to transactions between second-half of last year and first-half of current year for given year-value
21	Customer ID	float64	Unique customer identifier
22	Unnamed: 21	float64	Not applicable
23	Unnamed: 22	float64	Not applicable
24	Unnamed: 23	float64	Not applicable
25	Unnamed: 24	float64	Not applicable
26	Unnamed: 25	float64	Not applicable

**in Pakistani Rupee*

A first evaluation of data quality aspects reveal the existence of empty rows and empty columns in the dataset, which result from additional separator signs in the raw data. Another accumulation of missing values can be identified in the "sales_commission_code" column. Apart from that, the amounts of missing values per column are negligible (<0.03%). The following graphic provides a visualization of the situation:

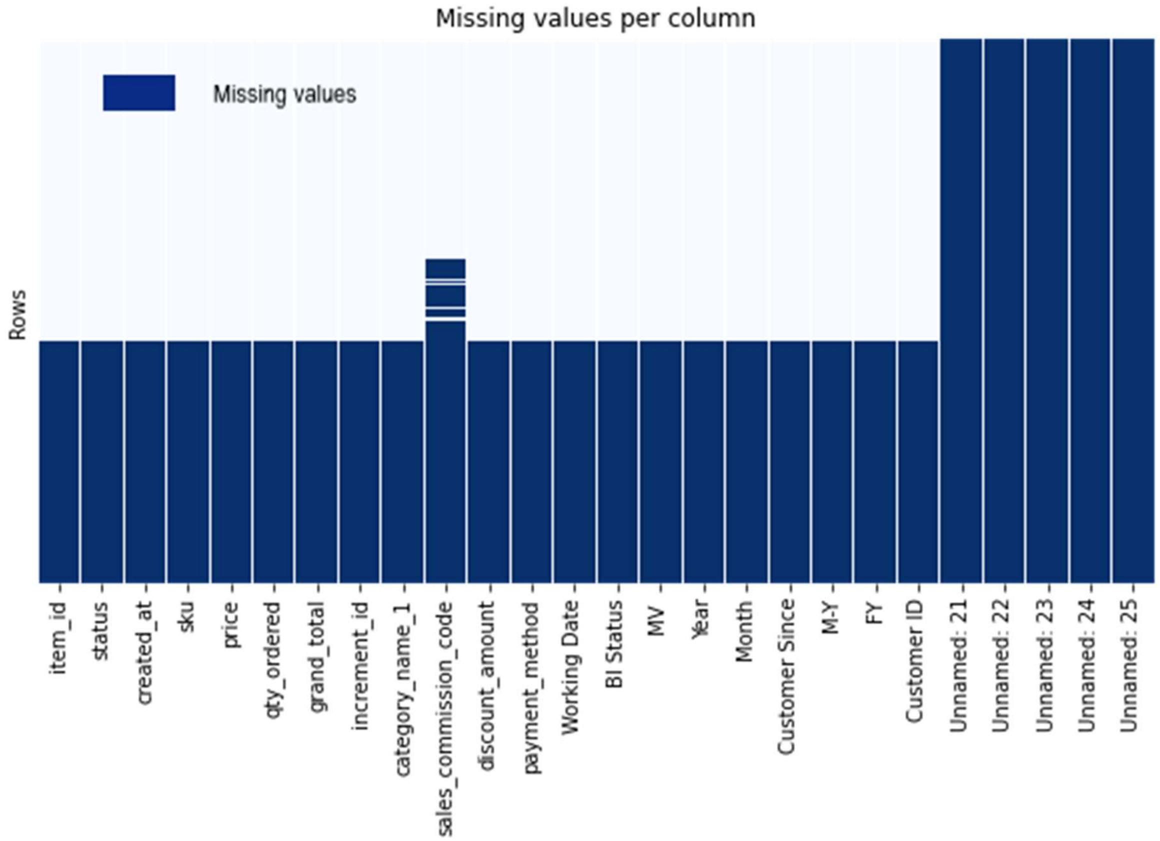


Fig 1.1: Missing values per column (empty fields marked in dark blue)

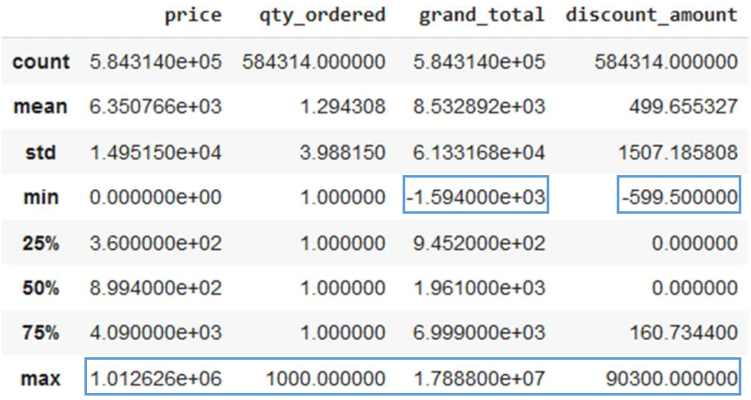
A closer look into "sales_commission_code" shows that 76% of the existing data points are not further specified and are labelled by "\N". With that, the variable only presents information of limited value for the subsequent classification analyses.

Based on these findings we can comfortably suggest:

- to delete the identified empty rows and empty columns
- to exclude "sales_commission_code" from our analyses
- to reduce the data set by dropping rows with remaining missing values

After conducting the proposed pre-processing steps, the data is of dimension (584314, 20) and does not include any duplicates.

Additional attention must be paid to extreme values and outliers which can be observed for the numeric variables:



	price	qty_ordered	grand_total	discount_amount
count	5.843140e+05	584314.000000	5.843140e+05	584314.000000
mean	6.350766e+03	1.294308	8.532892e+03	499.655327
std	1.495150e+04	3.988150	6.133168e+04	1507.185808
min	0.000000e+00	1.000000	-1.594000e+03	-599.500000
25%	3.600000e+02	1.000000	9.452000e+02	0.000000
50%	8.994000e+02	1.000000	1.961000e+03	0.000000
75%	4.090000e+03	1.000000	6.999000e+03	160.734400
max	1.012626e+06	1000.000000	1.788800e+07	90300.000000

Fig 1.2: Distribution characteristics of numeric variables (“grand_total” only indicative)

As highlighted in figure 1.2, the dataset includes orders with a negative price and also transactions with negative discount levels. Furthermore, all presented variables have a maximum value which is multiple times higher than its corresponding 75%-quantile. Both phenomes indicate the existence of extreme values or eventually incorrect data points and require further investigation. The following chapter is dedicated to this task and conducts a deeper analysis of the variables which are relevant for the later applied segmentation methods.

2 Data exploration & pre-processing

To prepare the dataset adequately for the RFM segmentation, it is of crucial importance to understand the characteristics and the dependencies between the individual variables. Herby the report is focusing on selected attributes and findings which are particularly relevant for the subsequent clustering concepts and the development of a marketing strategy.

2.1 Transactions, orders & status

As indicated in the first chapter, each row in the dataset is related to a single transaction. Nevertheless several transactions can belong to the same *order.* Strictly speaking, the 584.314 remaining transactions are resulting from 408.659 single orders. This result is based on the circumstance that 80% of orders only contain one transaction. However, this should not hide the fact that orders with up to 72 transactions exist:

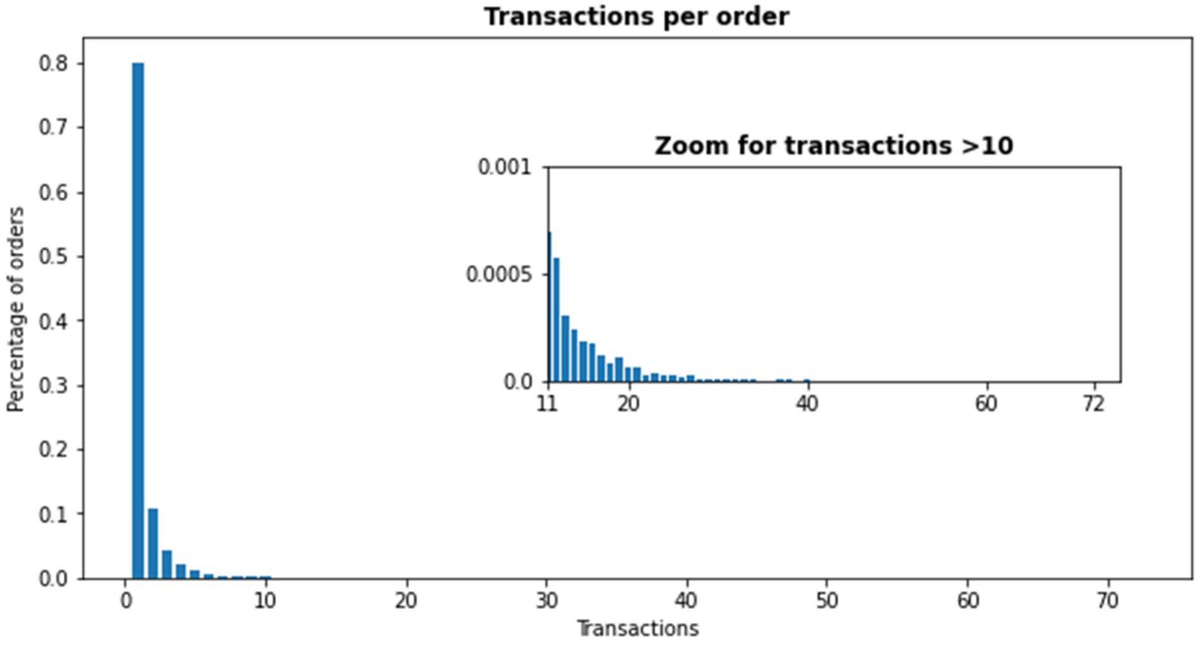


Fig 2.1.1: Transactions per order

Another relevant aspect is the period of time in which transactions (resp. orders) take place. This information is especially useful to identify the right moment of activation when rolling-out a marketing campaign. In this context, March, May and November clearly stand out (see graph below) which will be further investigated in chapter 4.

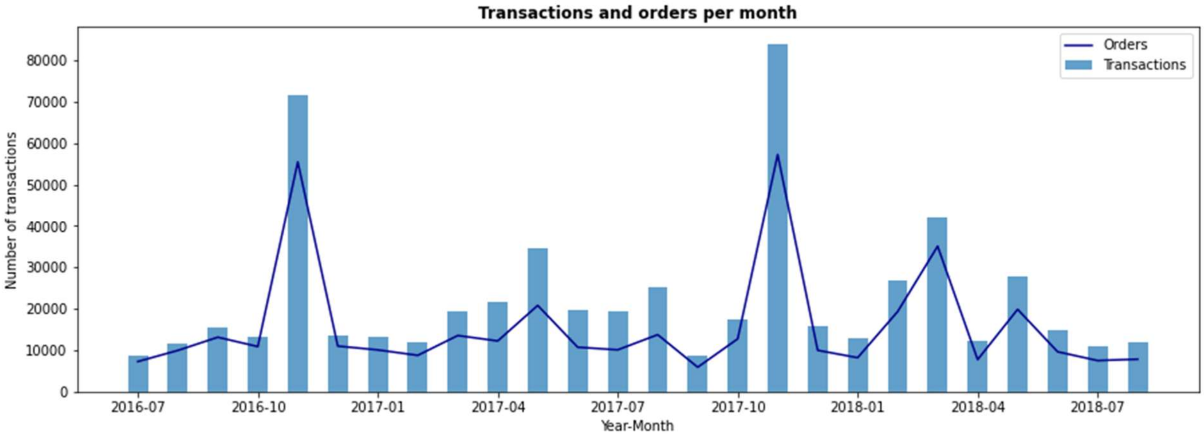


Fig 2.1.2: Transactions and orders per month

Also the *status* of a transaction is of special interest as we primary aim to increase successful product sales. As shown in the following graph, the vast majority of transactions (74%) are either classified as *complete* (40%) or *cancelled* (34%).

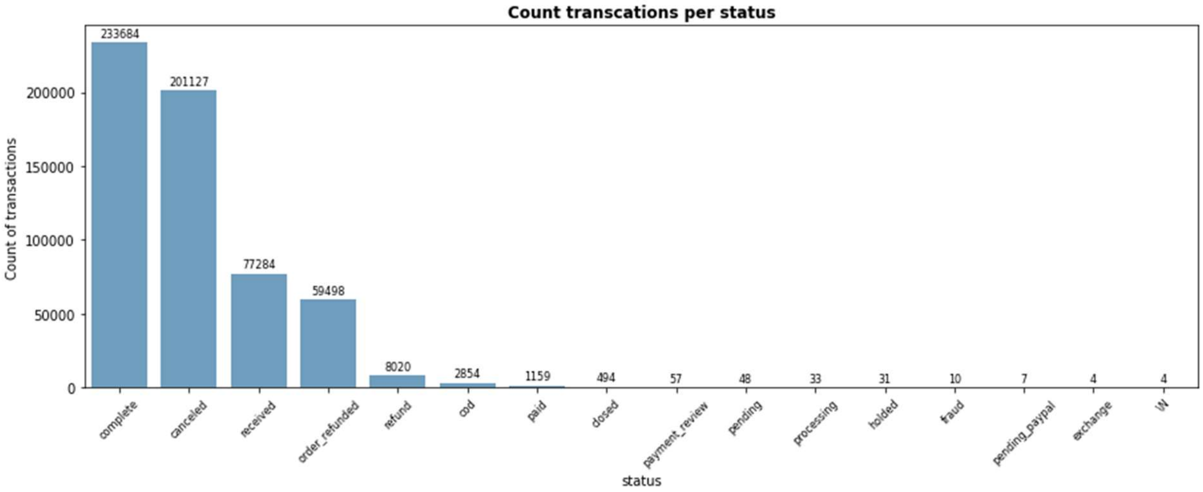


Fig 2.1.3: Status overview

A closer look into the different modalities indicate similarities regarding the names, for example *order_refunded* and *refund*, but also in terms of meaning, like *pending* and *holded*. For simplification we grouped the status into six contextual clusters:

refunded	completed	processing	pending	cancelled	fraud
- order_refunded *- refund*	*- complete* *- closed*	*- paid* *- received* *- cod* *- exchange* *- processing*	*- holded* *- pending_paypal* *- payment review* *-pending*	*- cancelled*	*- fraud*

2.2 Price, discount amount & grand total

To identify patterns in the purchase behaviour of the customers, it is important to understand the relationship between the financial variables “price”, “discount_amount” and “grand_total”. While the first two items are available on transaction basis, “grand_total” refers to the order to which the transaction belongs. The relationship can be described as follows:

$$\boxed{grand\_total_{order\ i} = \sum_{transactions\ k,i} price_k * qty\_ordered_k - discount\_amount_k}$$

At this point, our dataset reveals weaknesses regarding data quality. Not only that we observed zero or negative values for “price”, “grand_total” and “discount_amount” (see chapter 1), the described formula is only valid in 80% of cases. As these issues occur also for *completed* orders and especially due to the relevance of the underlying attributes, we decided:

- to exclude transactions with a “price” of zero
- to exclude transactions with a negative “discount_amount”
- to delete transactions which belong to an order without positive “grand_total”
- to reduce the dataset to orders where the formula above holds true

Additionally, to improve the interpretability of the variable “discount_amount”, we calculate the “discount_rate” per transaction:

$$\boxed{discount\_rate \coloneqq \frac{discount\_amount}{price * qty\_ordered}}$$

As we can see in Fig. 2.2.1, the calculated indicator is significantly correlated with the share of successful product sales as it increases the likelihood for completion and reduces the likelihood for cancellation of a transaction. This insight will be further analysed in chapter 4.

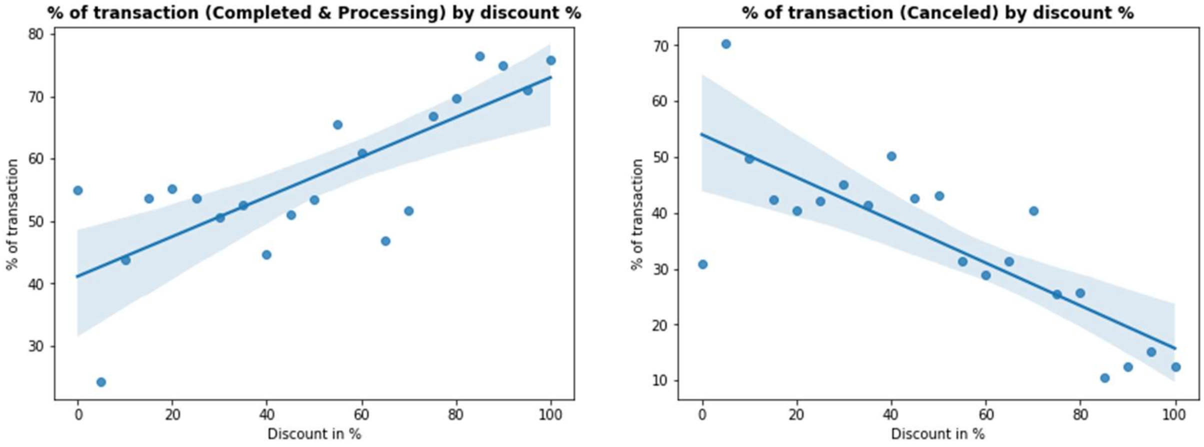


Fig 2.2.1: Impact of discount on transactions

Looking now at the resulting distributions (see Fig. 2.2.2 below) of the pre-discussed variables, it is immediately apparent that all histograms are strongly right-skewed. Underlined by the added boxplots, it is noticeable that the vast majority of transactions (resp. orders) are concentrated in a relatively small range. Nevertheless extreme high values are present in all four dimensions. For instance, we see that 86% of transactions relate to one single item but on the other hand transactions with up to 1000 items exist. In the next chapter we introduce and apply a technique to identify and exclude those outliers from the analyses.

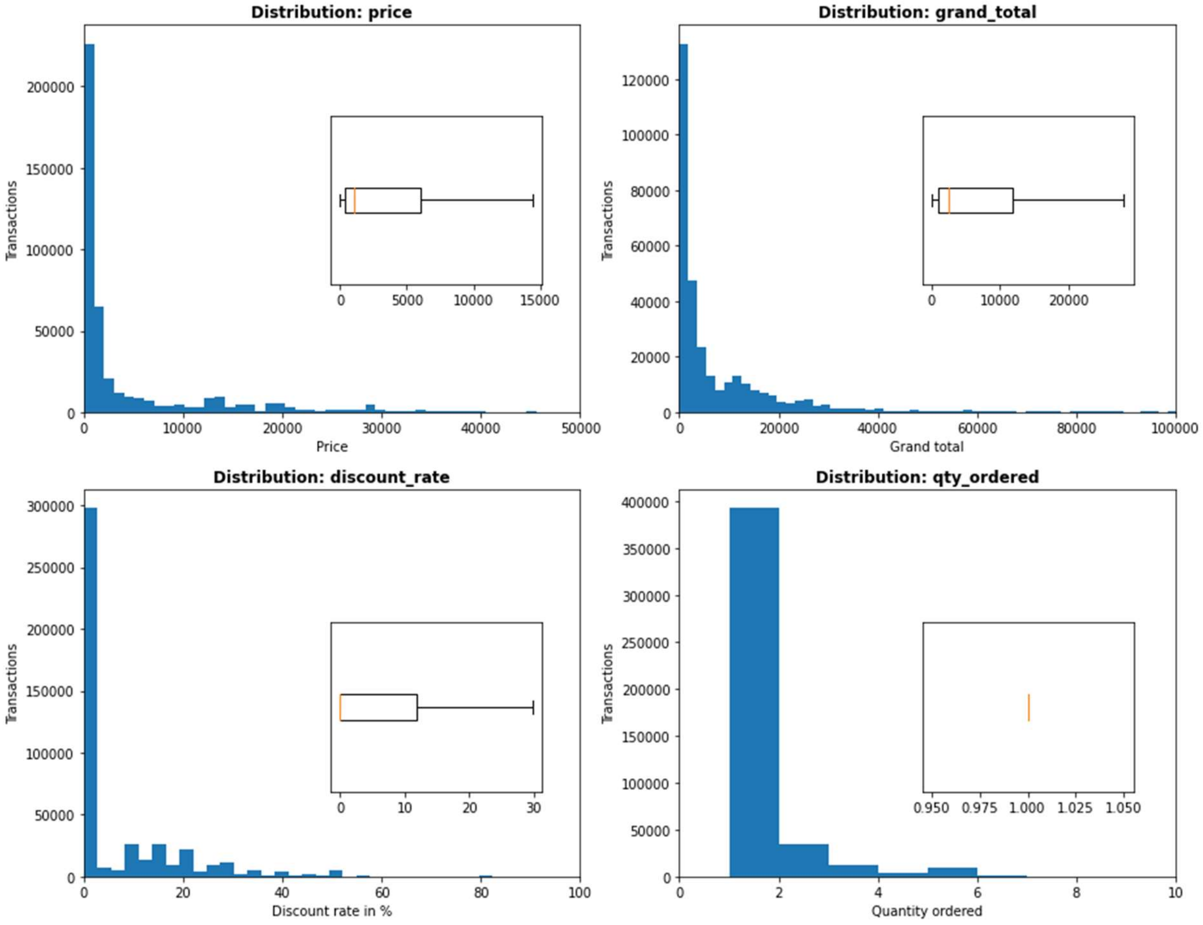


Fig 2.2.2: Distribution plots

2.3 Consumption categories

In the last part of this chapter we analyse the consumption categories which belong to the individual transactions and evaluate their top-line contribution. As summarized in Fig 2.3.1, there are 14 specified and one *Other* category to which the ordered products are assigned to.

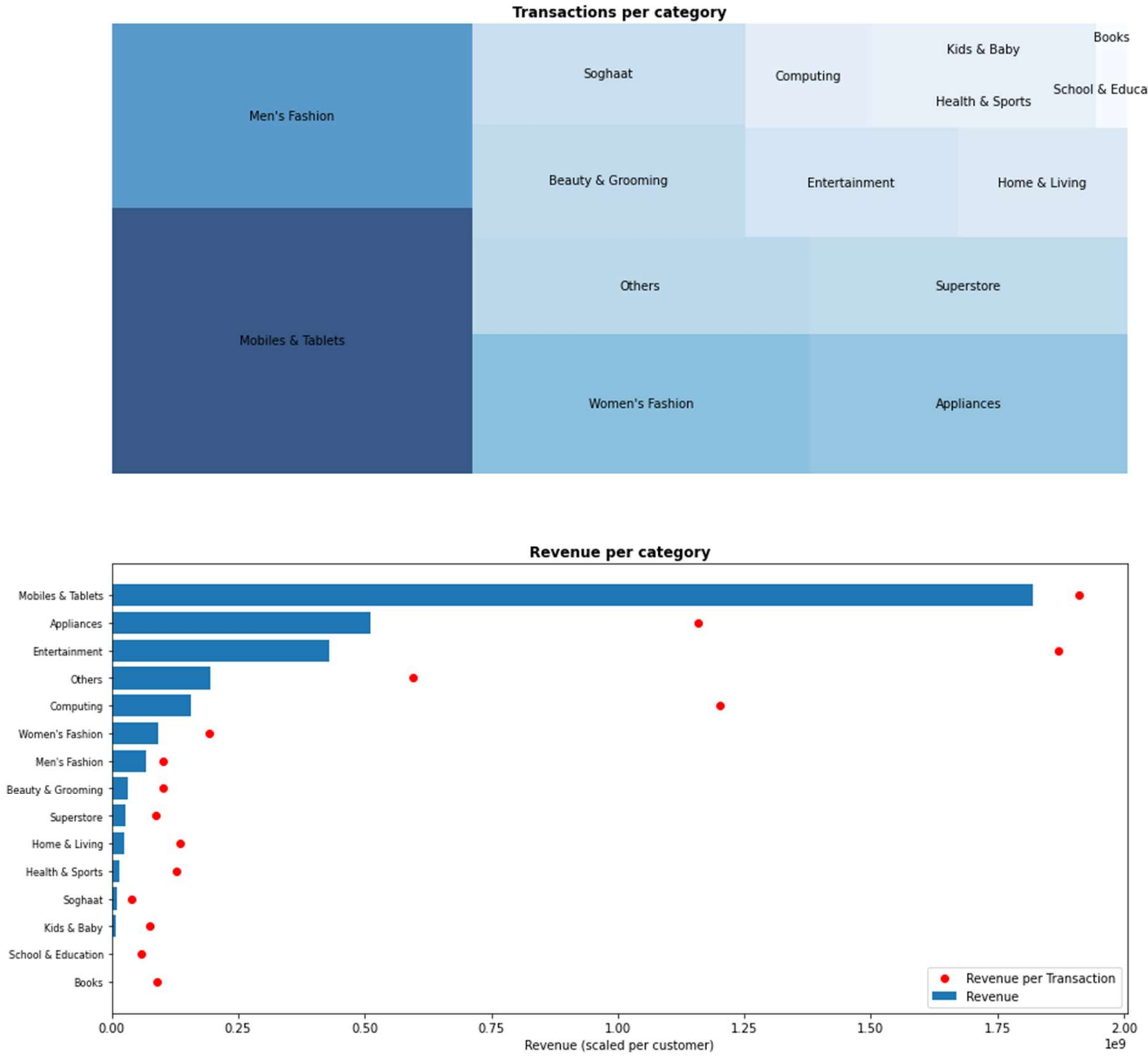


Fig 2.3.1: consumption categories and revenue

By aggregating *Men's fashion* and *Women's fashion* into one cluster, fashion products form the biggest group and contains 25% of all transactions. Nevertheless *Mobiles & Tablets*, *Appliances* and *Entertainment* products deserve as well special attention. These products are frequently demanded (see bucket size) - and due to their nature - show the strongest revenue impact and also highest revenues per transaction. These findings will build the basis for the development of a targeted offer strategy in chapter 4.

To close this section, we take an additional look into customer's purchase behaviour. Indicated by one scatter point for each single customer (see Fig 2.3.1 below), we conclude that the majority of customers conduct a rather small number of transactions and show a tendency to stick to the same categories. The blue box in the graphic contains already 85% of customers.

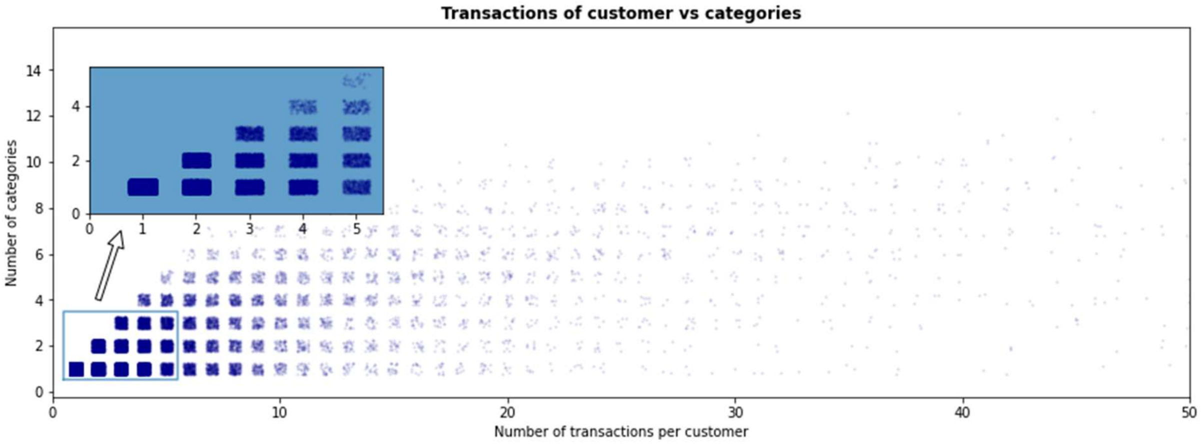


Fig 2.3.2: Completed transactions per customer and categories

3 RFM segmentation

After being familiar with the variables and having prepared the dataset, we are now able to approach the central part of this report – the RFM segmentation. RFM refers to **R**ecency, **F**requency, **M**onetary and indicates the grouping of customers according to the characteristics of their recent transactions. With that, the concept should help to develop targeted marketing campaigns. In the following chapters we give a short introduction into this approach and perform and evaluate different RFM segmentation methods.

3.1 Introduction to RFM segmentation

In this section we guide through the basic elements of an RFM segmentation. Hereby we are loosely following the article *"RFM analyses (recency, frequency, monetary)"* from Gavin Wright (https:/www.techtarget.com/searchdatamanagement/definition/RFM-analysis).

In a first step, each customer is scored based on three different dimensions:

"

- ***Recency**: How recent was the customer's last purchase? Customers who recently made a purchase will still have the product on their mind and are more likely to purchase or use the product again […].*
- ***Frequency**: How often did this customer make a purchase in a given period? Customers who purchased once are often are more likely to purchase again […].*
- ***Monetary**: How much money did the customer spend in a given period? Customers who spend a lot of money are more likely to spend money in the future and have a high value to a business "*[1]

After calculating these scores for all customer, the total bandwidth of each indicator is typically split in up-counting sections. For instance, customers with a comparable *"low frequency"* are grouped in frequency-cluster 1, while those with "highest frequency" find themselves for example in frequency-cluster 5. Therefore the customer is characterized - and in the same time segmented - in a three-dimensional space. To reduce the number of potential segments (here 125), several adjacent clusters are combined in a final step. It is important to mention that the process aims to generate groups which are homogenous in the same cluster but heterogenic between them. With that, marketing campaigns can be tailored to specific customer types and should help to reduce marketing spends or maximize the return of investments.

The table below visualizes a potential result of a RFM segmentation and builds the background for the development of a baseline model in the next section:

[1] (https:/www.techtarget.com/searchdatamanagement/definition/RFM-analysis)

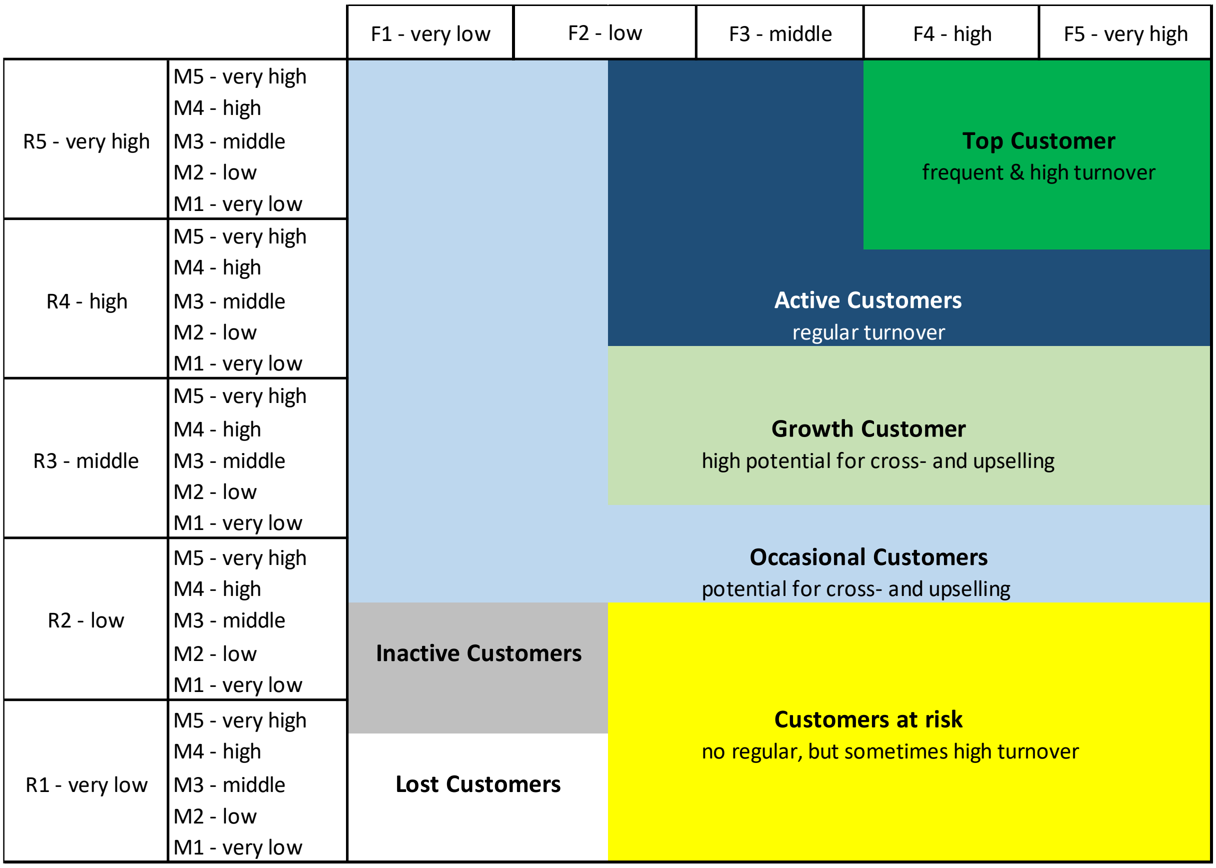


Fig 3.1.1: RFM segmentation (based on translated version from datatsolut)[2]

3.2 RFM scores & baseline model

Before we explore the advantageous of machine learning algorithms to perform the clustering tasks, we develop the underlying RFM-tables and explore the applicability of the baseline approach described above.

3.2.1 Customer scoring

In line with the presented interpretations, we calculated the RFM scores based on the following definitions:

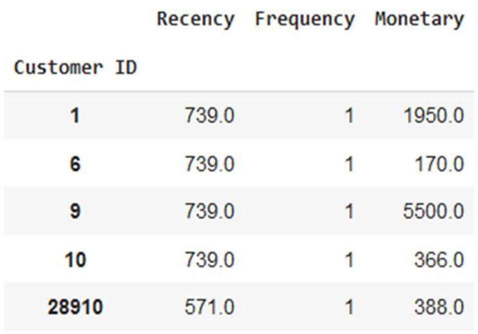
- ***Recency****: Time-difference in days between the last transaction of a customer and the most recent available date in the dataset (16.08.2018)*
- ***Frequency:*** *Number of orders of a customer*
- ***Relative Frequency:*** *Frequency divided by the time-difference in days between customers first transaction and the most recent available date in the dataset (16.08.2018)*
- ***Monetary****: Total amount of spending for all transactions of a customer*

[2] (https://datasolut.com/rfm-analyse/)

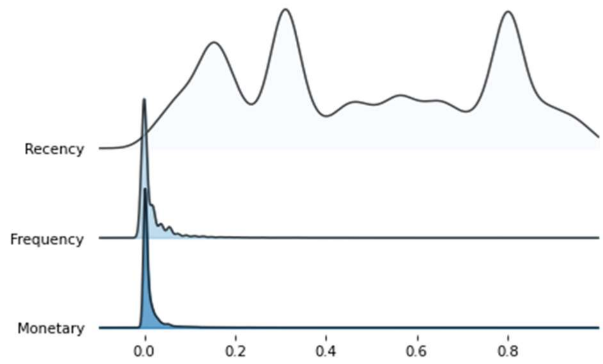
Herby we only focus on successful (*completed*) or at least promising (*processing*) transactions as we would otherwise strongly bias the customer scores and eventually maximize *cancelled* orders. Additionally we applied the z-score transformation, a standard method[3] to identify and exclude outliers. This means we only consider orders where the values of relevant input parameters “price”, “qty_ordered”, “discount_amount” and the orders per customer lie in the interval:

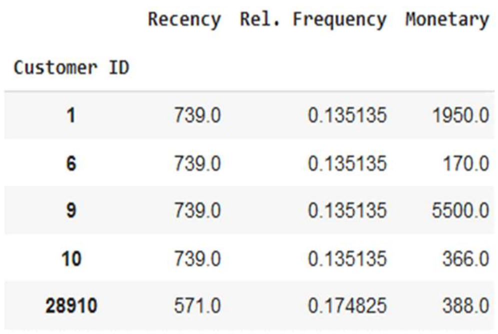
$$\boxed{[mean - 3 * std,\ mean + 3 * std]}$$

As a final step we calculate the RFM tables. In the upper part we present the actual derived values with “Frequency” and “Relative Frequency” on the left and the related distributions on the right-hand side. Beneath, we provide the corresponding scores translated in numbers between one and five. For this calculation, the values of each variable are segmented in five equal-sized groups. The only exception is “Frequency” where we clustered all customers with only one order in the same logical bucket, with two orders in a second bucket, with three in a third one, with five or six in the forth and the remaining ones in a fifth cluster. With that, we distribute the values “as good as possible” in an equal manner. Please be aware that customers with highest “Recency” must receive the lowest Recency-scores. These tables include 60.911 customers and build the basis for the analyses conducted in chapter 3.3.

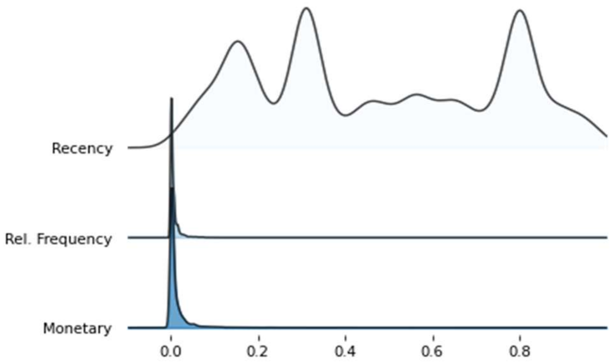


	Recency	Frequency	Monetary
Customer ID			
1	739.0	1	1950.0
6	739.0	1	170.0
9	739.0	1	5500.0
10	739.0	1	366.0
28910	571.0	1	388.0

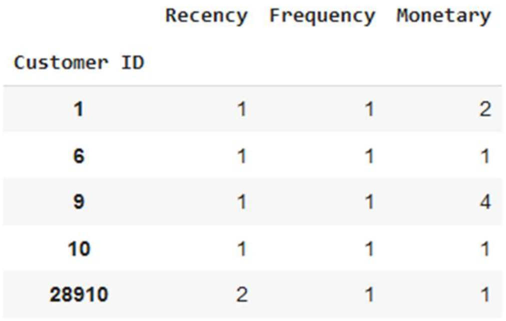




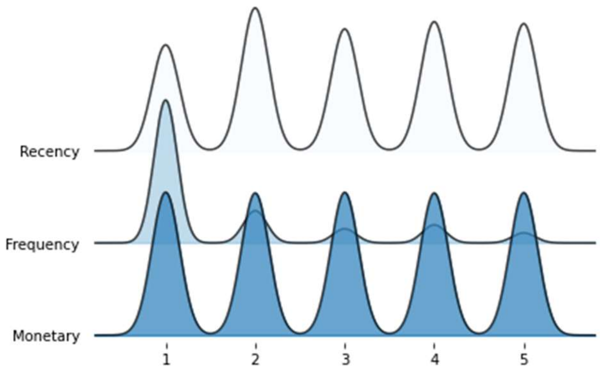
	Recency	Rel. Frequency	Monetary
Customer ID			
1	739.0	0.135135	1950.0
6	739.0	0.135135	170.0
9	739.0	0.135135	5500.0
10	739.0	0.135135	366.0
28910	571.0	0.174825	388.0

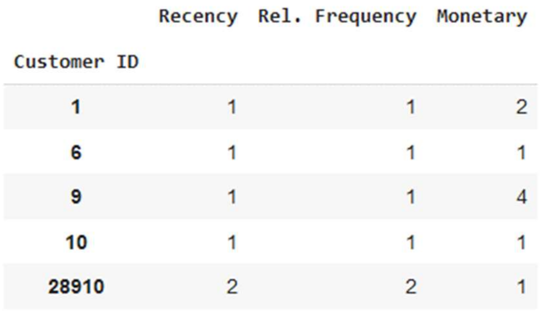


[3] (https://www.geeksforgeeks.org/z-score-for-outlier-detection-python/)



Customer ID	Recency	Frequency	Monetary
1	1	1	2
6	1	1	1
9	1	1	4
10	1	1	1
28910	2	1	1





Customer ID	Recency	Rel. Frequency	Monetary
1	1	1	2
6	1	1	1
9	1	1	4
10	1	1	1
28910	2	2	1

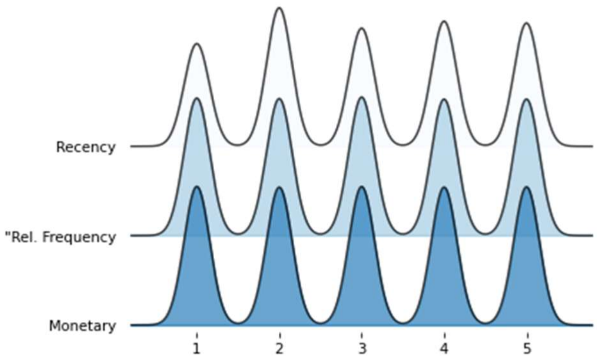
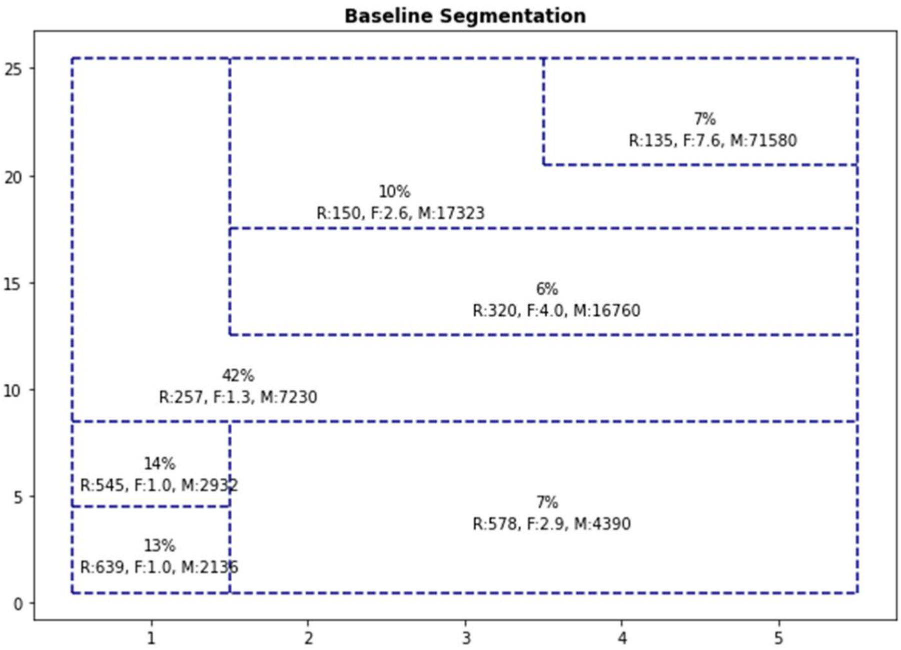


Fig 3.2.1: RFM tables for five selected customers (left) and distributions of variables. Values MinMax-scaled for visualisation reasons

Analysing the distributions, it is immediately noticeable that the absolute values of “Frequency”, “Rel. Frequency” and “Monetary” are still strongly right-skewed. This is due to the fact that 66% of the customers just bought once and 73% spend below 10.000 Rupee. Only “Recency” is well distributed over the whole range. By design, the clusters in table three and four have the same size. Only “Frequency” stands out, as explained above. For the following baseline segmentation we continue according to the approach described in 3.1 and with that, apply the third RFM-table for grouping the customers.

3.2.2 Baseline segmentation

After assigning the customers to their respective buckets, we are able to describe and evaluate the segmentation based on its usability for developing a targeted marketing strategy.



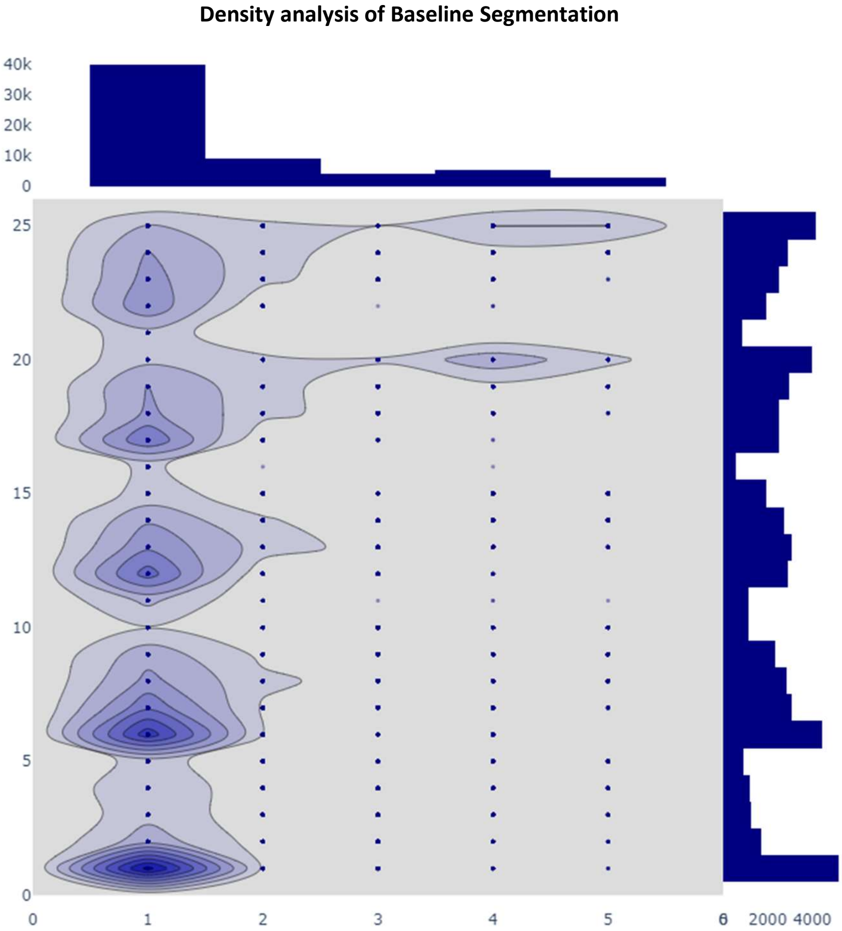


Fig 3.2.2.1: Baseline segmentation and density (R: Recency, F: Frequency, M: Moneta

According to Fig. 3.2.2.1 the resulting groups are of rather small size when compared with "Occasional customers". This segment dominates and encompasses already 42% of the clients. Nevertheless it is noteworthy that its average Frequency is only 1.3, which results from a large number of customers with just one order. This is also indicated by the density plot which reveals an overweight of one-time clients with rather low spending levels. Another accumulation of individuals can be observed in the "Top-" and "Active-group" with high monetary values. Especially the "Top-group" stands out with a monetary average more than four times higher than the next best segment.

As described above, the core objective of the segmentation is to identify sub-groups which are as homogenous as possible in order to approach them by an individualized messaging. On the other hand, these groups must show high heterogeneity between each other to sharply separate the customers. With the silhouette-score we examine how well the applied baseline model supports this goal. For equal treatment of the three dimensions, we used a Min-Max-scaling before calculating the score:

Silhouette-score: -0.17

The negative score implies a poor performance of our model. Therefore we will explore alternative concepts by applying machine learning techniques and variations in the next section.

3.3 Clustering based on machine learning

One of the reasons for the unsatisfactory results we received in the last chapter is induced from applying a rule-based approach that divides customers in equal-sized groups without considering relations (in sense of geometric distance) between them. To overcome these limitations we leverage clustering concepts that search for center points in the data.

For the following, we tested a wide range of K-Mean models. This includes the application of different scaling methods, introduction of new variables like “Relative Frequency” (see above) or logarithmic transformations of the input data. Nevertheless, as we considered the same optimization and evaluation criteria, we provide explanations and descriptions representative for one model and refer for the details of the remaining ones to our code documentation and the summary table below. For this initial model we use the upper RFM-table from Fig. 3.2.1 and conduct the K-Mean clustering to identify best separating customer segments.

Model development. In a first step, the data is transformed with a Min-Max-Scaler to ensure independency of outcomes from initial scales. To find the best k-value (number of clusters), we leverage the “elbow”-methodology and verify the findings with silhouette plots:

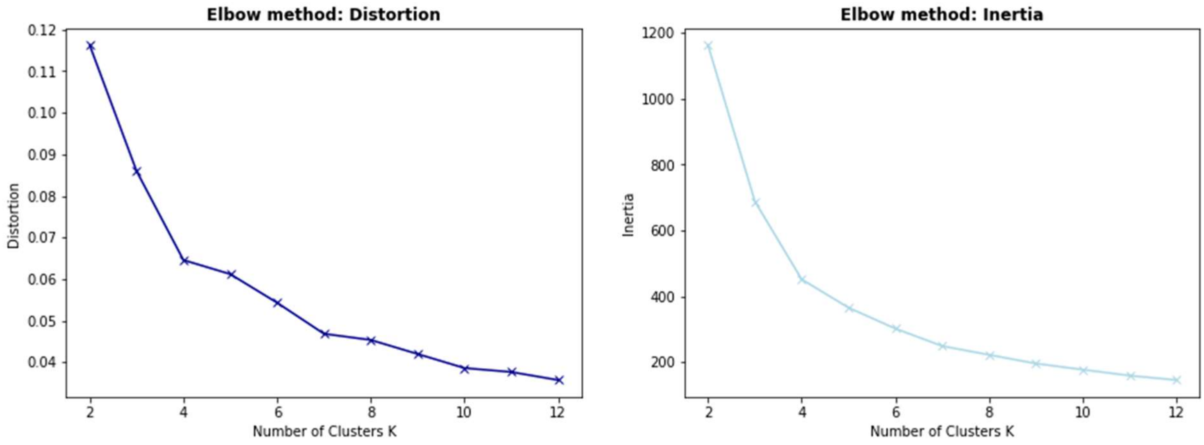


Fig. 3.3.1: Elbow-charts for distortion and inertia

As the resulting curves in Fig 3.3.1 decrease sharply and slow down after k = 4, we consider four clusters as the optimal number to be used in the K-Mean algorithm. This is confirmed by Fig 3.3.2 through the uniform thickness of the clusters, with silhouette-scores above average and a very small share of “miss-classified” data points (see top-left).

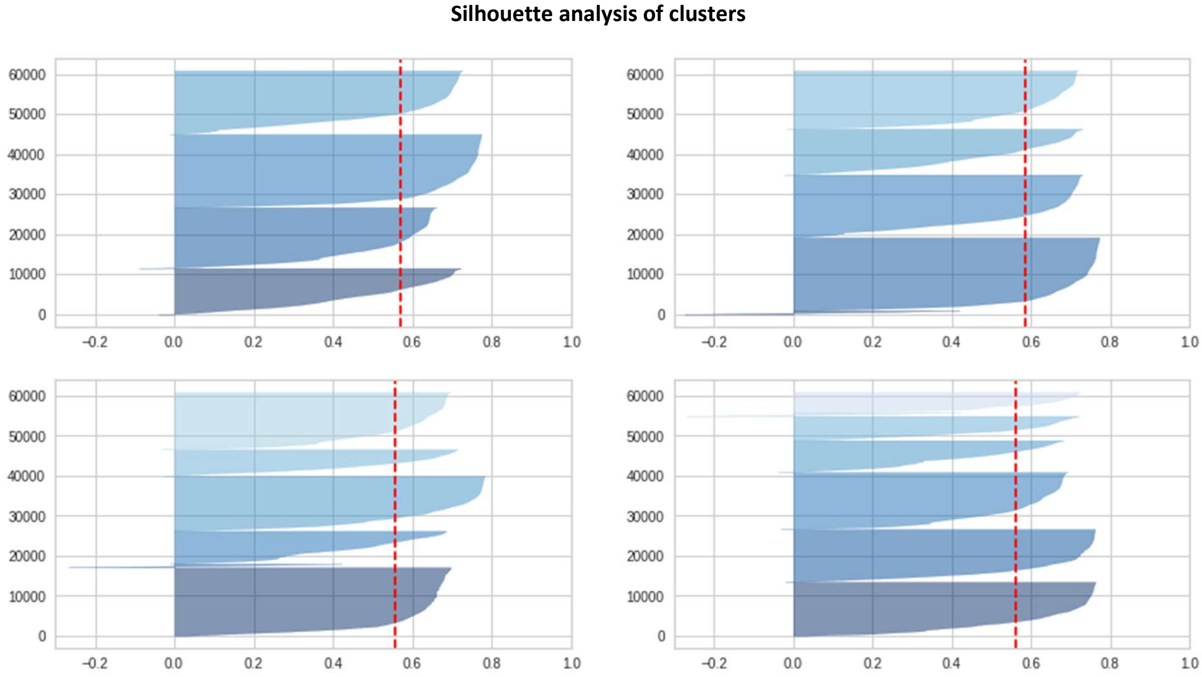


Fig. 3.3.2: Silhouette plots for k=4,5,6,7

Exploration. A 3d-representation of the developed segmentation helps to understand and interpret the outcome of the model. As visualized in Fig. 3.3.3, the algorithm divides the dataset in four separated areas including respectively 25%, 26%, 19% and 30% (bottom to top) of the total customer base.

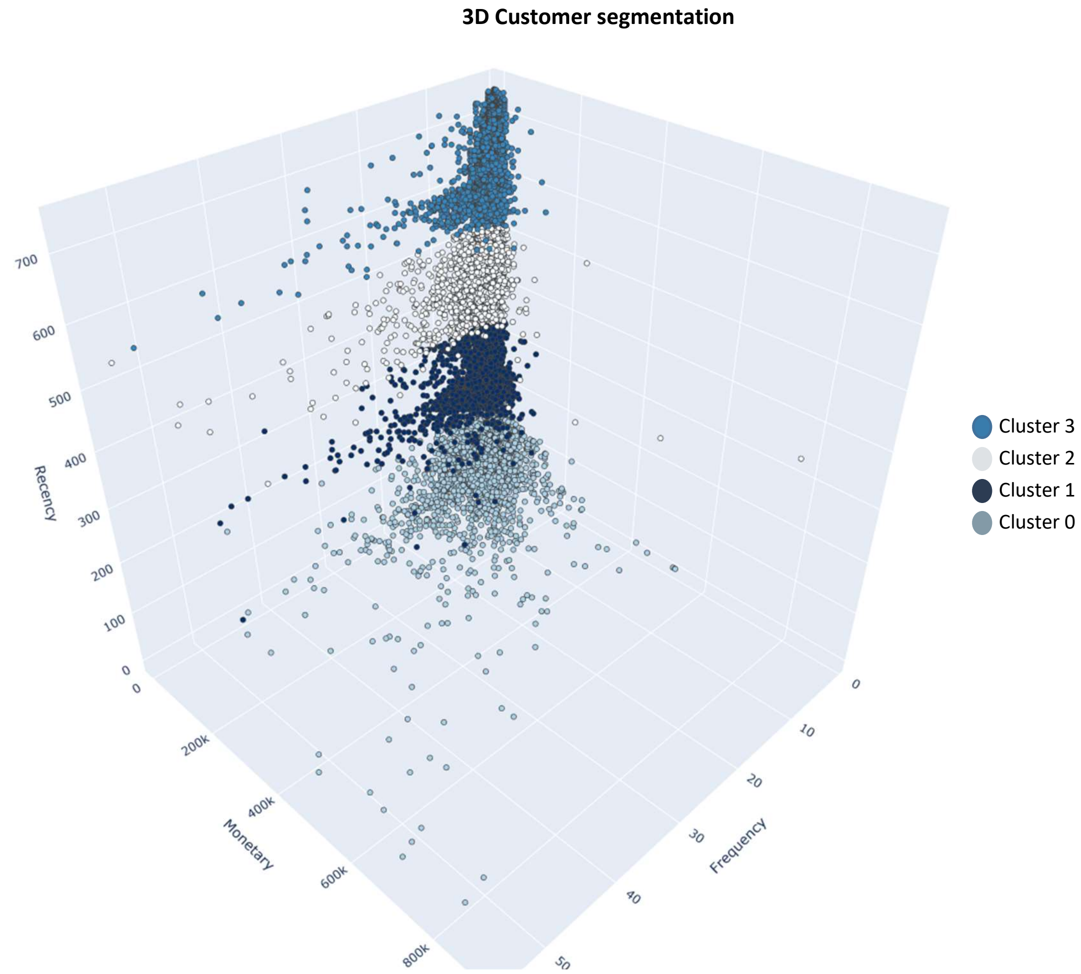


Fig. 3.3.3: Cluster visualisation

It can be clearly observed that clusters are primary selected by their differences in the "Recency"-values. "Monetary" aspects and "Frequency" seem to have a subordinate role. This is even stronger underlined in the boxplot-graphs below:

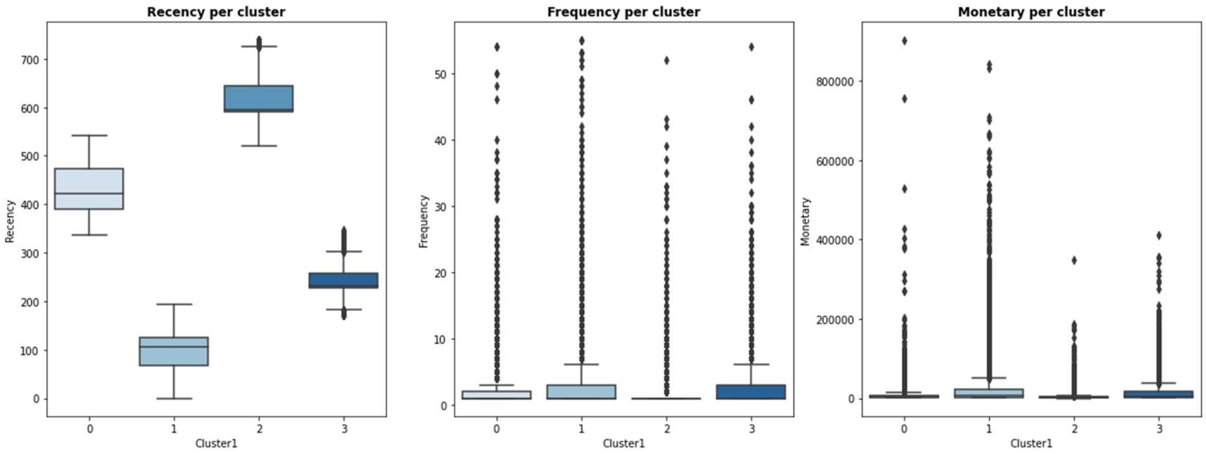


Fig. 3.3.4: Boxplots per cluster and dimension

As already discussed in chapter 3.2.1, "Recency is fairly spread over the whole value range. In contrast, the data points related to "Frequency-" and "Monetary"- observations are compressed on the left side of the distributions. Therefore, the distance-based concept of K-Mean primary optimizes towards "Recency" and also transfers the described properties into the clusters. This can be well observed in the boxplot representation – and even stronger – in the snake-plot below (see Fig.3.3.4). The graph reduces the clusters to their scaled mean-values of dimensions and a difference in "Frequency" and "Monetary" is basically not observable.

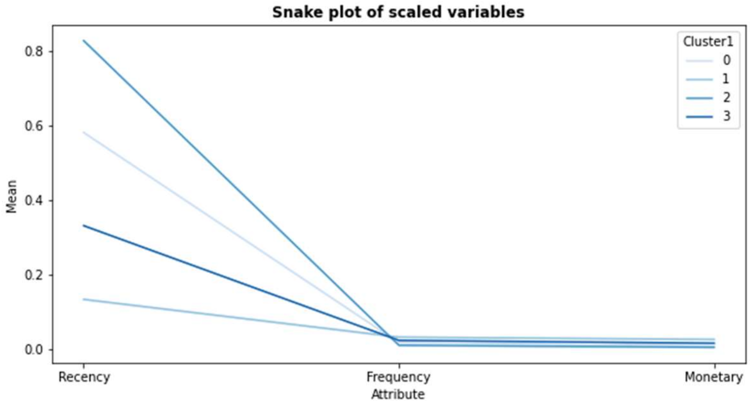


Fig. 3.3.4: Snake-plot for scaled variables: Mean of attributes per cluster

Evaluation. As introduced in 3.2.2, we are primary using the Silhouette-score to quantify the success of our clustering approach. From this point of view, we can be quite satisfied, as we receive a score of 0.71. Nevertheless, we do not consider it suitable to base a targeted marketing strategy purely on the time since customer's last purchase and therefore reject the model.

Due to that, we further experiment with variations of the described K-Mean model which are summarized in the table below. For details we refer to the provided code documentation.

Model #	Concept	Reasoning	"Elbow"-Clusters (k)	Range: Customers per cluster	Silhouette-score	Mean-Variation of clusters (Snake)
1	Initial model (see above)	Baseline approach	4	19% to 30%	0.71	Only for "Recency"
2	RFM-table with Relative Frequency (see above)	Evaluates amount of orders per time (see above)	4	19% to 30%	0.77	Only for "Recency"

3	Initial model, standard scaled	Relates extreme values to the distribution	3	3% to 51%	0.66	2 similar clusters per dimension
4	Model 2, standard scaled	Relates extreme values to the distribution	6	0% to 37%	0.60	Significant different clusters
5	Model 1 with logarithmic transformation of input data	Creates more symmetric distributions	3	19% to 50%	0.34	Significant different clusters
6	Add "Lifetime"-variable to input data of model 1	Considers time since customer exist	3	25% to 42%	0.71	Only for "Recency" and "Lifetime"
7	Add "Lifetime"-variable to input data of model 5	Considers time since customer exist	3	18% to 50%	0.49	Significant different clusters

Even we received from a pure "technical" point of view well acceptable silhouette-scores (see model 2) and segments with significant different clusters (see model 4 and 5) we are not satisfied with the applicability of the results for deriving a targeted marketing campaign. This is due to a certain lack of a clear interpretability of the resulting segments. Therefore we combine quantitative and qualitative aspects in the next section to derive the "final model".

3.4 Final model & segmentation

From a marketing point of view it is a key difference if a customer has already made several purchases with the company or whether she bought only once. For returning customers we can assume that a previous experience was in a certain sense satisfactory for the client and use this information when approaching the customer again. Keeping this qualitative argument in mind, we split our dataset for the next analyses in two conceptual different parts: Customer with one- and customers with several orders.

In this sense, the dataset for the first part of the analyses – single purchase – consist of 39.652 customers. As "Frequency" is not a distinguish feature anymore, we can exclude the variable and transfer the clustering task into two dimensions. Furthermore we conduct a Mean-Shift clustering in order to not determine the cluster number beforehand. Nevertheless, due to the long run-time, we avoid the application of a grid-search for identifying the optimal bandwidth. We rather conduct a manual step-wise optimization by calculating the silhouette-score for different scenarios. With that, we receive a segmentation with seven different clusters (see Fig 3.4.1 below).

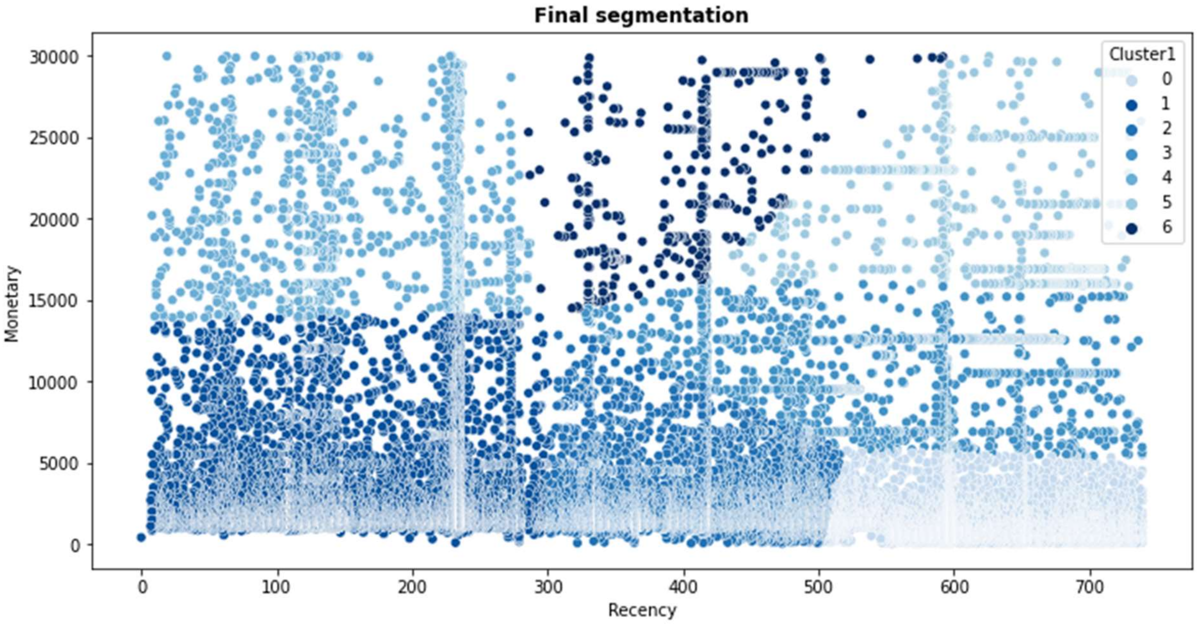
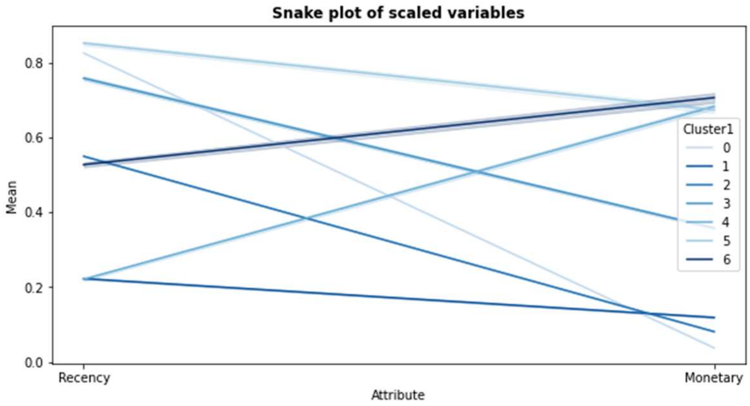


Fig. 3.4.1: Scatter-plot: customers and their associated clusters

Base on the described approach, we achieve a satisfactory silhouette-sore of 0.67. Compared to the K-Mean models explored above, the number of clusters seem to be relatively high. Nevertheless they have a clear interpretability. “Recency” is split in three sections (low, medium, high). Each section is sub-divided based on the “Monetary” values along a “straight line”. Additionally, as seen in Fig. 3.4.2, the clusters are significantly different as indicated by the mean-values of the two dimensions. On the other hand, it seems recommendable to aggregate smaller segments before using them for marketing purposes. A more detailed description of the clusters and their characteristics, as well as a sub-selection for a targeted campaign is provided in chapter 4.



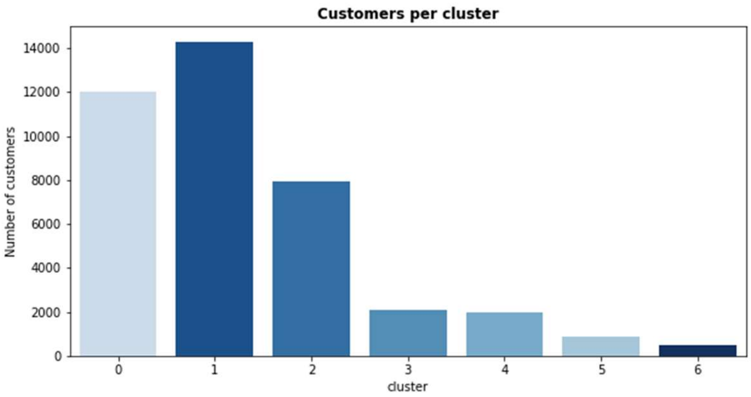


Fig. 3.4.2:Snake-plot (above): Mean of dimensions per cluster. Histogram (below): Cluster-size

For the second step of the analyses, we follow the argumentation above and consider only customers with more than one order. This embraces 20.539 customers. Herby we apply a K-Mean approach again and select the optimal cluster number through the “elbow-method” (see details above). Based on this approach we developed a segmentation with five clusters.

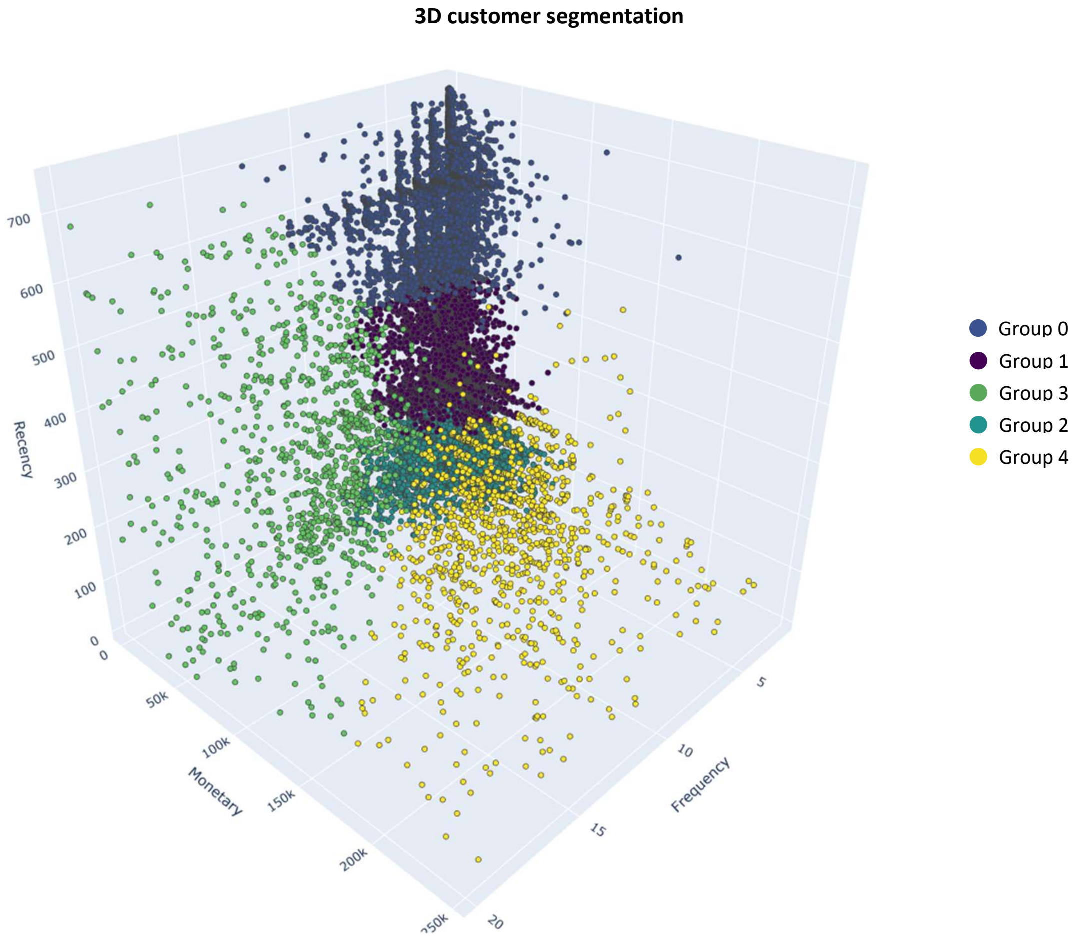


Fig. 3.4.1: Scatter-plot: customers and their associated clusters

Compared to the models described in 3.3, the silhouette-score (0.58) is considerably lower. Nevertheless the usability and interpretability of the derived segments is more convincing. Two clusters stand out, one with relatively high “Monetary”-values and the other with specifically high “Frequencies”. The remaining customers are separated according to their “Recency” in three buckets (low, medium, high), as already seen in the Mean-Shift model above. Again, we recognize significant differences between the clusters with regard to their mean-values but also related to the number of customers they include.

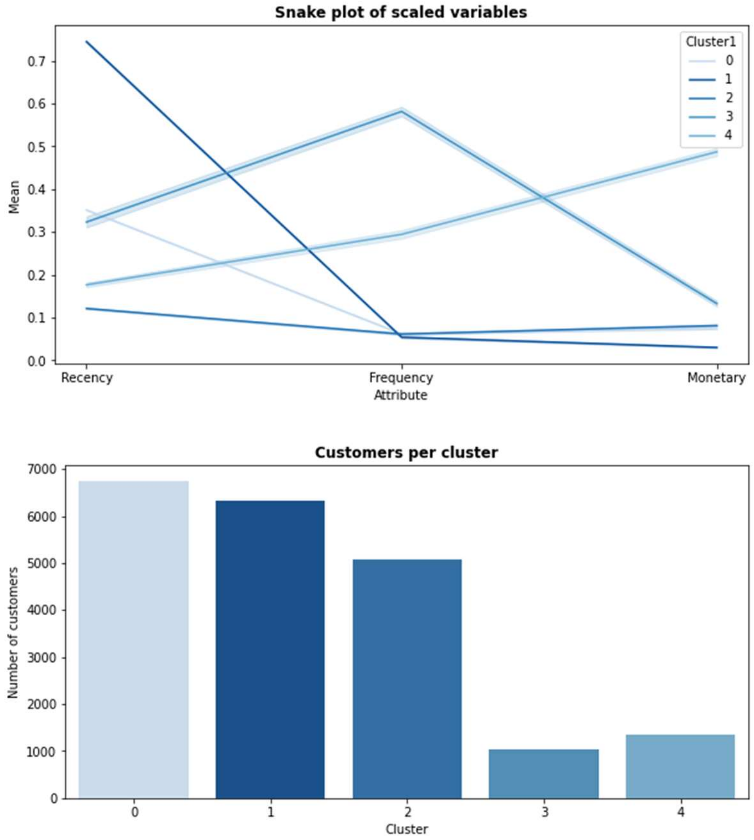


Fig. 3.4.4: Snake-plot (above): Mean of dimensions per cluster. Histogram (below): Cluster-size

Due to the good interpretability of the results, we are confident to develop a marketing strategy based on the presented approach. In the next chapter we describe these customer segments in more detail and deep-dive into their characteristics. This builds the basis for an optimized and individualized customer targeting.

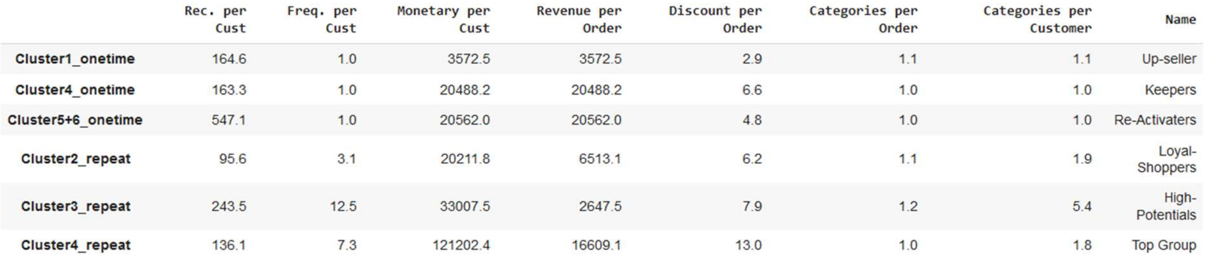
4 Marketing Strategy

The objectives of a marketing strategy and the associated campaigns can be manifold. Sometimes the main purpose is simply to promote the brand and increase the awareness in the market. Others wishing to strengthen the reputation of the company or to build up a positive image in a specific target group. In the context of this report we are looking into a third group of communication initiatives that aim to support company's sales by approaching customers with a tailor-made product offer. This should increase the sales rate by minimizing costs of customer contacts. To make this successful, it is crucial to approach the right customer with the right offer at the right point in time.

4.1 The right customer

Chapter 3 provides a concept which helps to cluster the customer base in twelve significant different segments (seven for the "one-time clients" and five for the recurring ones). Now it is usually the exclusive task of the marketing department to choose the respective groups that show the highest likelihood to respond positively to a campaign. Here we follow the argumentation from 3.1 and focus on those with low "Recency-", high "Frequency-" and/or high "Monetary-" values. This is concretely cluster 1, cluster 4, cluster 5 and cluster 6 for the "one-time clients" (while cluster 5 and 6 will be combined due to size) and cluster 2, cluster 3 and cluster 4 for the recurring customers.

The table below provides an overview of these segments and analyses relevant KPIs for a customized targeting. Additionally we assign names to the clusters and discuss their characteristics below.



	Rec. per Cust	Freq. per Cust	Monetary per Cust	Revenue per Order	Discount per Order	Categories per Order	Categories per Customer	Name
Cluster1_onetime	164.6	1.0	3572.5	3572.5	2.9	1.1	1.1	Up-seller
Cluster4_onetime	163.3	1.0	20488.2	20488.2	6.6	1.0	1.0	Keepers
Cluster5+6_onetime	547.1	1.0	20562.0	20562.0	4.8	1.0	1.0	Re-Activaters
Cluster2_repeat	95.6	3.1	20211.8	6513.1	6.2	1.1	1.9	Loyal-Shoppers
Cluster3_repeat	243.5	12.5	33007.5	2647.5	7.9	1.2	5.4	High-Potentials
Cluster4_repeat	136.1	7.3	121202.4	16609.1	13.0	1.0	1.8	Top Group

A glance at the figures reveal major differences between the individual customer groups and their purchase behaviour. In the following we examine those differences and describe the reasoning for including the group in the campaign:

- ***Up-seller**: These are "one-time customers" who bought recently a product from the company. It can be assumed that the purchase is still positively present in their mind.*

Additionally, the customers rather bought products on a lower price level. Consequently, a potential for further up-selling exists when offering the right product.

- ***Keepers**: These customers belong as-well to the "one-time segment" who recently purchased a product. Nevertheless, they show a high average revenue. This indicates an interest in the rather high-priced product categories (see 4.2). As these customers are not strongly connected to the company yet, it is relevant to invest in keeping them.*

- ***Re-Activators**: This group shows similar values like the "Keepers" which we discussed above. The difference is the higher "Recency". Nevertheless, it makes sense to also invest in these customers for re-activation. The main reason lies in the categories the customers are interested in (see 4.2). As the last order happened in average more than one year ago, we can assume that there is a next generation of the mobile or tablet available which the customer bought last time.*

- ***Loyal-Shoppers**: "Loyal-Shoppers" represents the customer cluster with the lowest "Recency". These customers already bought several times from the company and show therefore a solid total revenue. For a marketing strategy, it is important to keep in mind that these customers tend to stick to the same product categories.*

- ***High-Potentials**: This customer-group is of specific interest due to the very high average "Frequency". Additionally, the customers buy from several different categories. It is very likely that these customers respond positively to a marketing campaign and with that, show the highest potential.*

- ***Top-Group**: This segment is most attractive for the company, due to the low "Recency", the high "Frequency" and the outstanding "Monetary"-values the customers show in average. Nevertheless, like for the "Loyal-Shoppers", it is important to consider the limited number of categories these customers use.*

4.2 The right offer

After analysing and describing the six different target groups, we look into the right messaging to attract the individual customers. As already mentioned above, most customers show a tendency to buy products from only one or two categories. Only the “High-Potentials” form an exception in this sense. To respond adequately to this behaviour, we tailor the offer to the dominating product-category in each group:

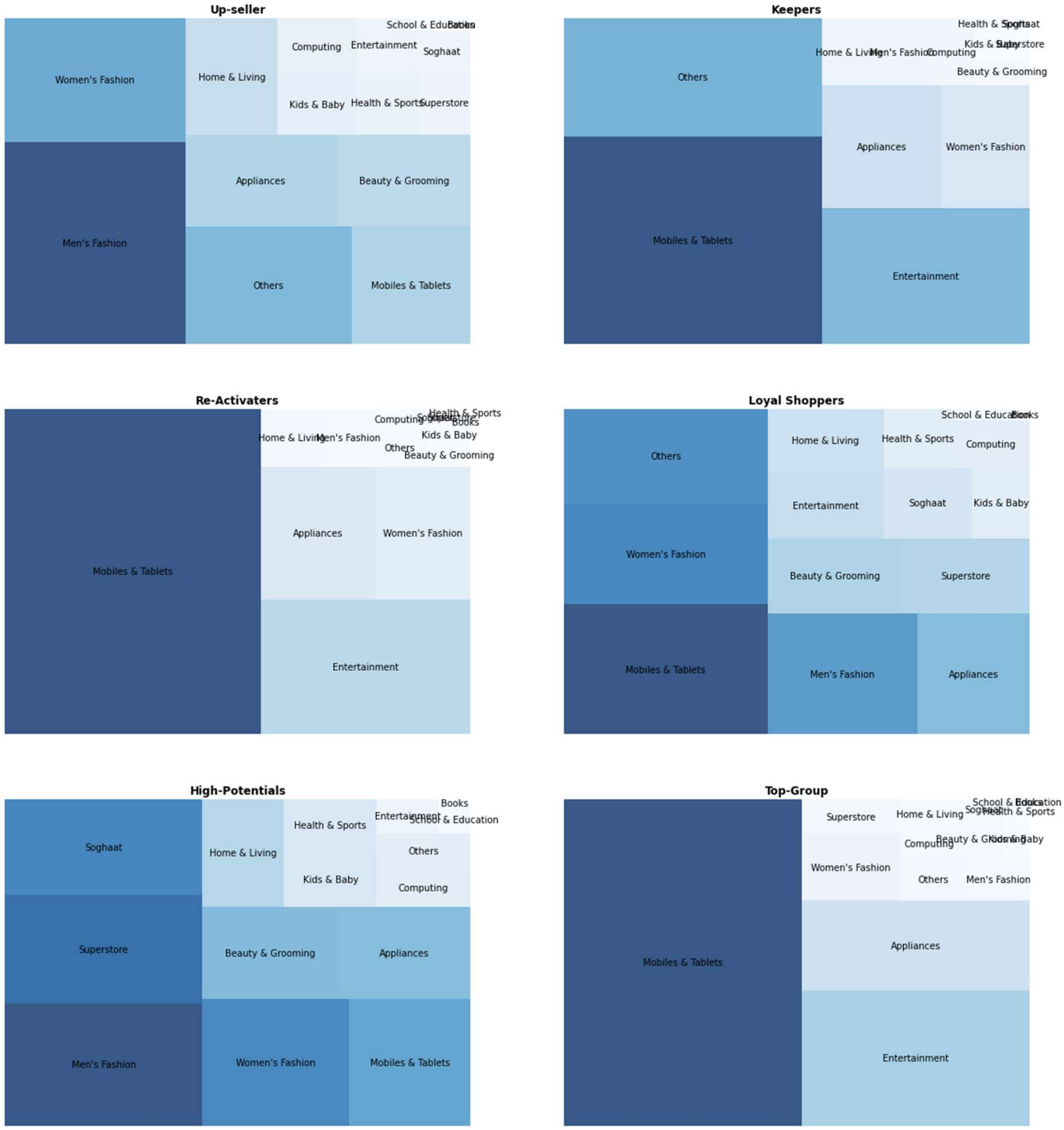


Fig.4.2.1: Categories of transactions per target-group

Fashion products build the biggest category when counting the number of transactions for the "Up-seller". In contrast, "Keepers" and "Re-activators" are described by the considerable share of *Mobiles & Tablets* and *Entertainment* products. This immediately explains the significant differences we have seen in the average "Monetary-"amounts of the groups. A very different picture we observe for the "Loyal-Shoppers" and the "High Potentials". Both clusters cover a wider spectrum of categories which includes for example *Fashion, Mobiles & Tablets, Beauty & Grooming* or *Appliances.* Therefore these segments can - but also need to - be approached by a broader messaging which, as an exception, does not relate to just a single product type. For the "Top-Group", *Mobiles & Tablets* but also *Entertainment* products stand-out. Also this group primary focus on one category, which needs to be considered when approaching the customers.

A second aspect of the offer strategy relates to the discount amount we grant our customers. As seen in Fig. 2.2.1, the share of completed orders significantly increases with the discount rate. Nevertheless, the effectiveness of discounts may vary considerably depending on the related product category. The following visualization analyses these effects:

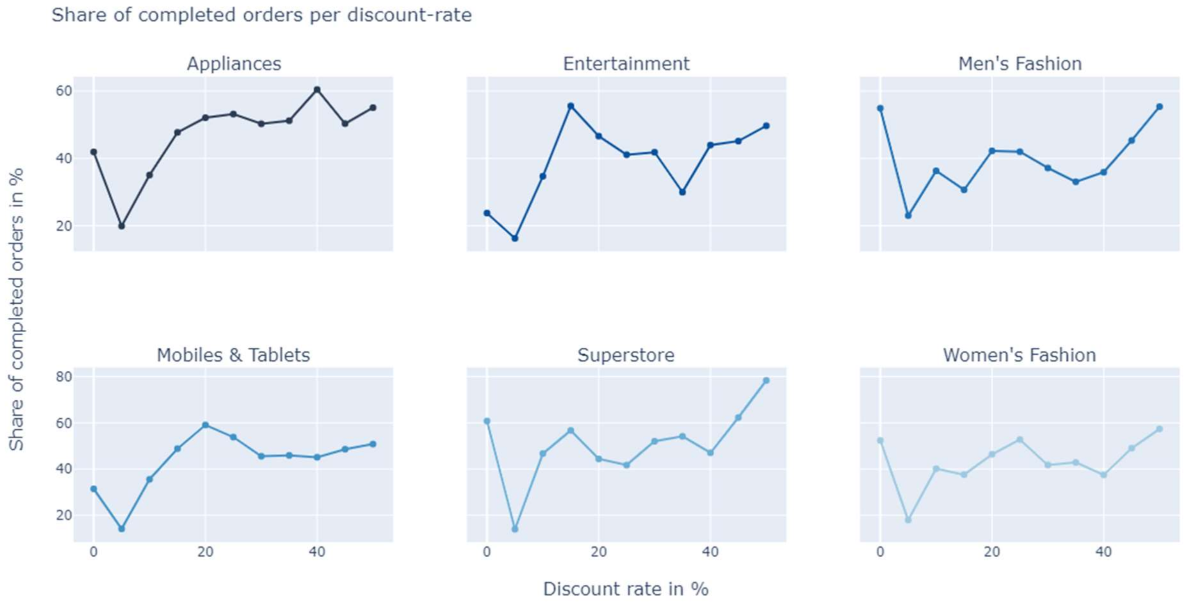
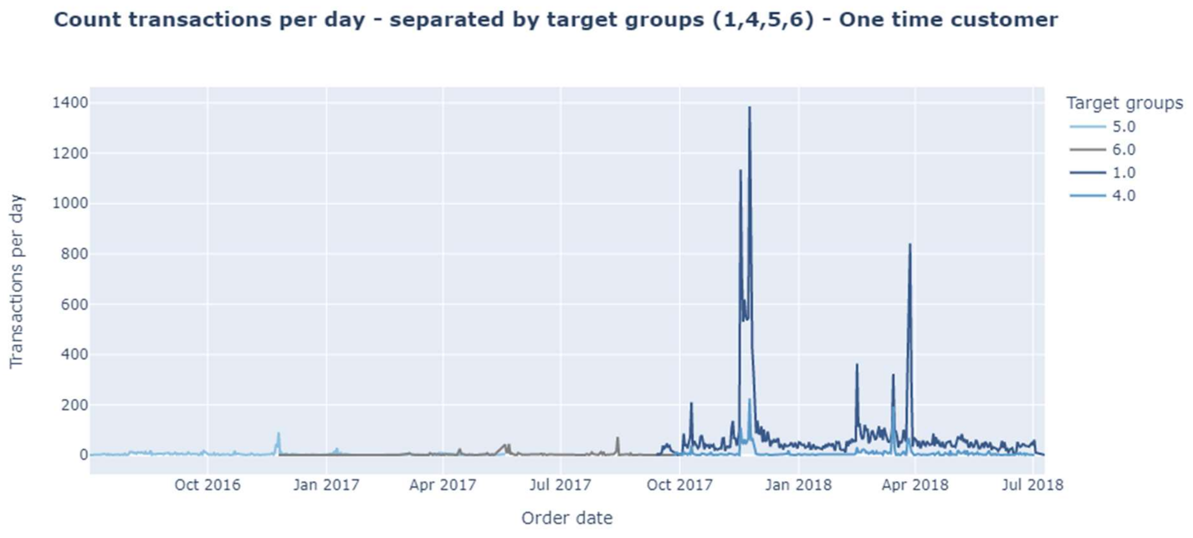


Fig. 4.2.2: Share of completed orders per discount-rate for main categories discussed above

With regard to the offer strategy, we consider an amount between 5% and 20% of the product price reasonable for a marketing campaign. In this range, the share of completed orders indicate an upward trend when increasing the discount-rate. Nevertheless, the individual optimum point per category needs to be further evaluated based on the cost structure and the profit margins per category.

4.3 The right time

For every marketing campaign the moment of activation is a crucial element that cannot be underestimated. It is of almost relevance to appeal to customers at a time when they are paying most attention. This is typically related to specific events like holidays, seasonal-closings or "Black-Friday-weeks". To reveal a suitable timeframe for our marketing campaign, we analyse patterns in the existing purchase behaviour of our target groups:



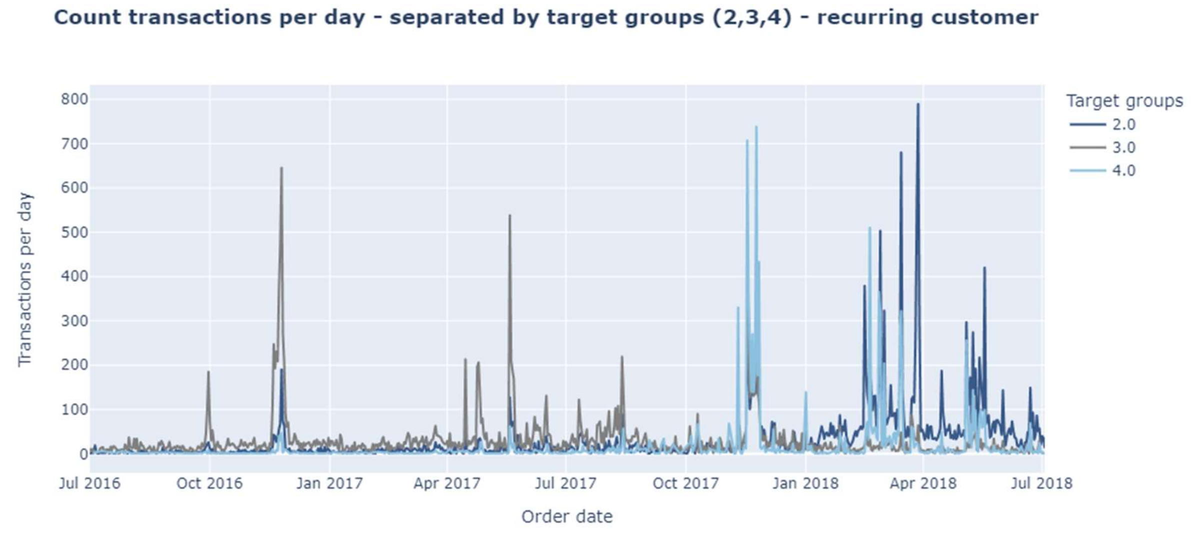


Fig. 4.3: Number of transactions per day per target group

Looking at the amount of transactions, we recognize a repeating peak in November of each year. This is related to the "Black Friday"-week in Pakistan and offers a great opportunity for placing our campaign. Another strong increase can be identified end of March where Pakistan celebrates a historic national day (23.3.). This is traditionally connected to shopping activities in the entire country and gives us the opportunity to be present with special offers.

Conclusion

One of the main objectives of this theses was the introduction and implementation of an RFM-segmentation and to demonstrate the advantageous of machine learning algorithms for the clustering task. For that we started with a detailed analyses of the available dataset, its variables and relationships between them. Based on these results we pre-processed the data to increase quality but also to prepare the dataset suitable for the RFM approach. In a next step we calculated the “Recency-”, “Frequency-” and “Monetary-“ values per customer, which form the fundamental concept for assigning them into homogenous groups. Now we followed a standard approach frequently described in related literature. Hereby we divided the customers equally per those dimensions. Nevertheless, the analyses of the outcome revealed unsatisfactory differentiations of customers between groups. To overcome this limitations, we applied several unsupervised clustering algorithms to select and segment the clients based on their euclidean distance. The results increased fundamentally but showed room for improvement with regards to interpretability. With the combination of qualitative considerations and incorporation of another machine learning technique we finally succeeded in developing a convincing and actionable segmentation.

A second goal was the development of a recommended marketing strategy. Herby we focused on the three core pillars of each campaign: “The right person”, “the right message” and “the right point of time”. To identify the right target groups, we analysed the RFM-segments base on their likelihood to respond positively to a campaign. In a next step, we mapped out the most appealing categories for each sub-group to individualize the product offer and to develop a targeted messaging. As a last step, we conducted a time-series analyses which showed recurring patterns and time windows with high potential for campaigns to be recognized by the customers.

With that we were successful to develop a targeted marketing strategy based on a machine learning driven RFM-segmentation. Nevertheless, we do not consider the explained approach as ideal in a professional settings. The RFM segmentation bases the clustering purely on the historic purchase behaviour. This means, that other relevant – or even more important – characteristics are not considered when optimizing our campaigns. This is for the example the age, the gender, the income or other socioeconomic or geographical factors. There we suggest to extend the RFM approach by available data points and explore their usability.