CAS Applied Data Science, University of Bern

Łukasz Maciaś

lukasz.macias.01@gmail.com

Data Science Project

Client segmentation for financial institutions specializing in distribution of Structured Investment Products using RFM and K-Means clustering model

Conceptual Design Report

6 October 2024

Abstract

This report outlines a conceptual project design for implementing the RFM (Recency, Frequency, Monetary) client segmentation model, enhanced with K-Means clustering, within financial institutions that specialize in distribution of structured investment products. The methodology described in this paper aims to provide a cost-effective and efficient data-driven approach for analyzing the company's customer base. By segmenting clients based on their transaction history, investment banks can identify their key customer groups, target clients at risk of losing and optimize the allocation of resources for marketing and sales efforts. The application of the K-Means clustering algorithm on top of the classical RFM client segmentation model allows for more refined client segmentation, better tailored to the specific characteristic of the analyzed customer base.

Table of Contents

Abstract	**1**
Table of Contents	**1**
1 Project Objectives	**2**
2 Structured Investment Products	**3**
3 Methods	**5**
3.1 Infrastructure	**5**
3.2 Software libraries and tools	**5**
3.3 Analytical Methods	**5**
3.3.1 RFM client segmentation Model	**5**
3.3.2 RFM client segmentation model implementation	**7**
3.3.3 K-Means clustering RFM client segmentation model	**8**
3.3.4 K-Means clustering RFM client segmentation model implementation	**11**
4 Data and Metadata	**14**
5 Data Flow	**15**
6 Data Quality	**16**
7 Data Model	**17**
8 Risks	**18**
9 Conclusions	**18**
Statement	**19**
References and Bibliography	**20**

1 Project Objectives

The objective of this conceptual project design report is to present a straightforward and cost-effective client segmentation methodology, based on the RFM (Recency, Frequency, Monetary) client segmentation model, specifically tailored to the structured investment products business. There are numerous firms in the Swiss and international financial institutions landscape that specialize in the issuance and distribution of structured investment products. The most well known Swiss companies leading in this sector include: Leonteq, Vontobel, GenTwo, CAT Financial Products, Swissquote etc. Given the transactional nature of the business conducted by those companies, where clients regularly buy and sell structured investment products, using the RFM (Recency, Frequency, Monetary) model, enhanced by K-Means clustering, is a compelling approach for conducting comprehensive and efficient analysis of the client base. Cost-effectiveness of this methodology manifests itself in a relative ease of implementation (at least in its classical form) and its ability to quickly deliver actionable insights into the customer base for the company's management. The goal of this report is to present how this methodology can be utilized for this specific use case and to guide the reader through the segmentation process.

2 Structured Investment Products

Important for the understanding of the project is a brief introduction into the basic workings of the Structured Investment Products. According to SSPA (Swiss Structured Products Asociacion): *"Structured Investment Products are instruments that combine, for instance, bonds or shares with derivatives, usually options, securitized in a standalone commercial paper. The repayment value of Structured Products depends, among other things, on the movements of one or more underlying assets, and/or the non-occurrence of a credit event on the part of the respective reference issuer."*[1] Structured Investment Products differ from traditional investment assets like shares, cryptocurrencies or commodities in that they offer a clearly defined payoff profile, established at the time of issuance. Additionally, most of them provide coupon payments throughout the product's lifetime. [2]

[1] Swiss Structured Products Association (SSPA). *Swiss Map of Structured Products 2024.* SSPA, December 2023. Available at: https://sspa.ch/wp-content/uploads/2023/12/en_sspaswissmap_2024.pdf

[2] *Die Welt der Strukturierten Produkte: Das Standardwerk für Strukturierte Produkte.* Swiss Structured Products Association (SSPA), 2023.

The following example of one of the most popular product categories—Barrier Reverse Convertible—will help to illustrate how Structured Investment Products work. The specific product in this example is the "15.80% p.a. Barrier Reverse Convertible on Nvidia".[3]

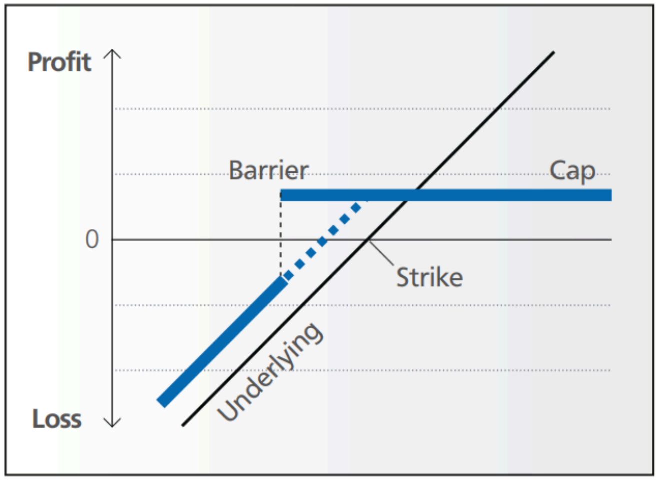


Figure 1: Illustration of a payoff profile of a Barrier Reverse Convertible product.[1]

Key characteristics of the product include:

- **Defined Duration:** The product has a fixed duration of 1 year.
- **Tradability:** It is tradable like any other security, with liquidity provided by the issuer.
- **Underlying Asset:** The product is based on a single underlying asset—NVIDIA.
- **Early Call or Liquidation:**
    - The investor can sell the product back to the issuer at any time at the current market price.
    - The issuer can call the product on predefined autocall dates or the product can reach its maturity, returning 100% of the notional amount to the investor.
    - The product can be liquidated if the underlying asset hits a barrier level, set at 55% for this product. On the Initial Fixing Date, the product's price is set at 100%. If NVIDIA's share price drops below 55%, the barrier event occurs and the product is terminated. Investor receives NVIDIA shares worth 55% or less of the invested notional, resulting in a loss.
- **Coupon Payment:** Regardless of the underlying asset's performance, during its lifetime, the product offers a fixed quarterly coupon payment, totaling 15.8% of the invested notional over the year.

[3] Payoff.ch. *Term Sheet for Structured Product CH1369852335.* 21 August 2024. Available at: https://data.payoff.ch/termsheets/CH1369852335_de_20240821_090251.pdf

3 Methods

3.1 Infrastructure

Due to data confidentiality, typical in the financial industry, the analysis will be conducted by an analyst with authorized access to the necessary data sources, entirely on a company-provided virtual desktop environment.

3.2 Software libraries and tools

As a minimum requirement, the analyst should have access to Excel and a Python distribution, such as Anaconda. The following Python libraries will be utilized:

- **cx_Oracle** or another suitable library to connect to the company's data warehouse and extract the necessary raw data for the analysis.
- **Pandas** for data manipulation and traditional RFM segmentation.
- **scikit-learn** to perform RFM segmentation using K-Means clustering.
- **Matplotlib.pyplot** for data visualization.

Additionally, to effectively publish and share the results, the analyst would need at least PowerPoint. The ideal approach for sharing insights would be through a data analytics tool like Microsoft Power BI or Tableau.

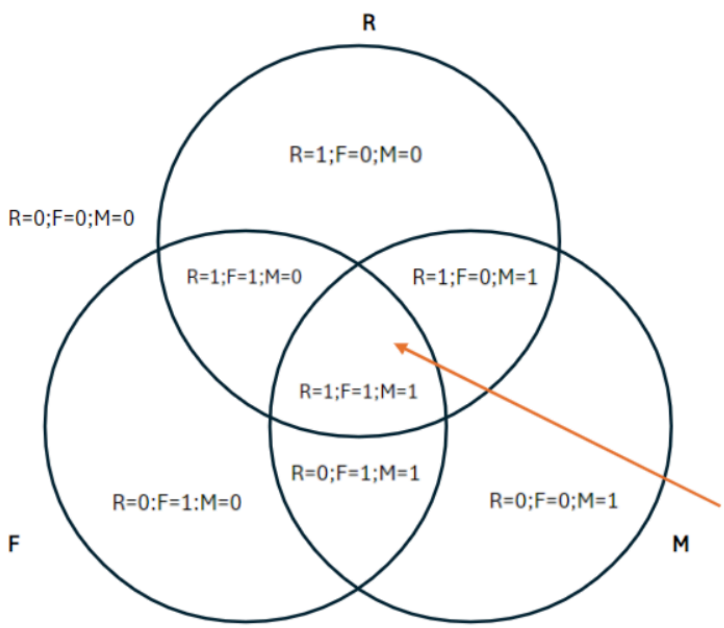
3.3 Analytical Methods

3.3.1 RFM client segmentation Model

Customer segmentation involves dividing a company's customer base into distinct groups based on shared characteristics, enabling the company to target each group more effectively and tailor its marketing strategies appropriately.[4] The classical RFM (Recency, Frequency, Monetary) model, first discussed in Robert C. Blattberg's 1994 book "Database Marketing," was initially used in direct mail marketing by companies like American Express. It categorizes customers based on transaction history: Recency (how recently a purchase was made), Frequency (how often), and Monetary (how much was spent) in a predefined observation period. This model is ideal for transactional businesses, therefore it suits very well the business model of the investment banks offering structured products.

[4] Diaz Ruiz, Carlos A. (2024). *Elgar Encyclopedia of Consumer Behavior*, Edward Elgar Publishing

In the standard RFM segmentation model, each client is assigned a score in each of the 3 categories based on whether its Recency, Frequency, or Monetary value falls above or below a specified percentile level in the investigated population - usually set at 75th percentile. For instance, if a given client's number of transactions is above the 75th percentile of the entire population, it receives 1 point for Frequency; similarly, scoring for Recency and Monetary value follows the same method. The percentile thresholds and the time frame for transactions must be determined by the analyst, who is a subject matter expert.[5] The segments are hierarchical, with Segment 1 containing the most valuable and engaged clients and Segment 8 the most disengaged ones. In the author's view, the client group with R=0, F=0, M=1 should rank higher in the hierarchy, despite having a lower overall RFM score than the group with R=1, F=1, M=0, as they are typically considered more valuable.



Segment	Score by R;F;M	RFM Score sum
1	R=1;F=1;M=1	3
2	R=1;F=0;M=1	2
3	R=0;F=1;M=1	2
4*	R=0;F=0;M=1	1
5*	R=1;F=1;M=0	2
6	R=1;F=0;M=0	1
7	R=0;F=1;M=0	1
8	R=0;F=0;M=0	0

Figure 2: RFM Client segmentation (own work)

[5] Blattberg, Robert C. (1994). *Database Marketing: Analyzing and Managing Customers.* Springer Science & Business Media

3.3.2 RFM client segmentation model implementation

Observation period.

As a first step, the analyst must define a time frame for the analysis. Structured Investment Products typically trade over a fixed, predefined period known as the duration, which spans from the issue date to the product's (early) redemption date. For most of the products this duration averages around one year. [6] Typically, after product termination, the redeemed capital is reinvested in a process known as rolling. To capture the entire product lifecycle, including this rolling phase, a time frame of 2 years seems to be optimal.

RFM Parameters

Recency - refers to the time elapsed since a client's most recent trade in primary market (issuance of a new product) or secondary market (increasing a position in an existing product).

Frequency - measures the number of trades executed by a client within the defined analysis time frame. Recently and Frequency are measures of the client's engagement.

Monetary - represents the revenues generated by a given client from fees associated with the sale of Structured Investment Products during the analysis time period.

Determining the percentile scores thresholds

Based on the literature and an exploration of method implementations available online, the 75th percentile is the most commonly used threshold for quantile scoring. In the author's opinion, a more effective approach for determining this threshold is to analyze each dataset individually and examine the distribution of each parameter while taking into account the skewness of the distribution and overall business characteristics.

Segmentation

The scores derived from this calculation allow us to segment clients into eight distinct categories (see Figure 3), which can be ranked hierarchically from the most valuable and engaged clients for which R, F and M values are all above the 75th percentile, to the least for

[6] Bärlocher, Jürg. *Experience with the Collection and Publication of Data on Structured Products in Switzerland.* BIS (Bank for International Settlements), IFC Bulletin No. 31, 1 September 2009. Available at: https://www.bis.org/ifc/publ/ifcb31l.pdf

which all of them are below that threshold. To facilitate interpretation, each client segment should be assigned a meaningful and descriptive name:

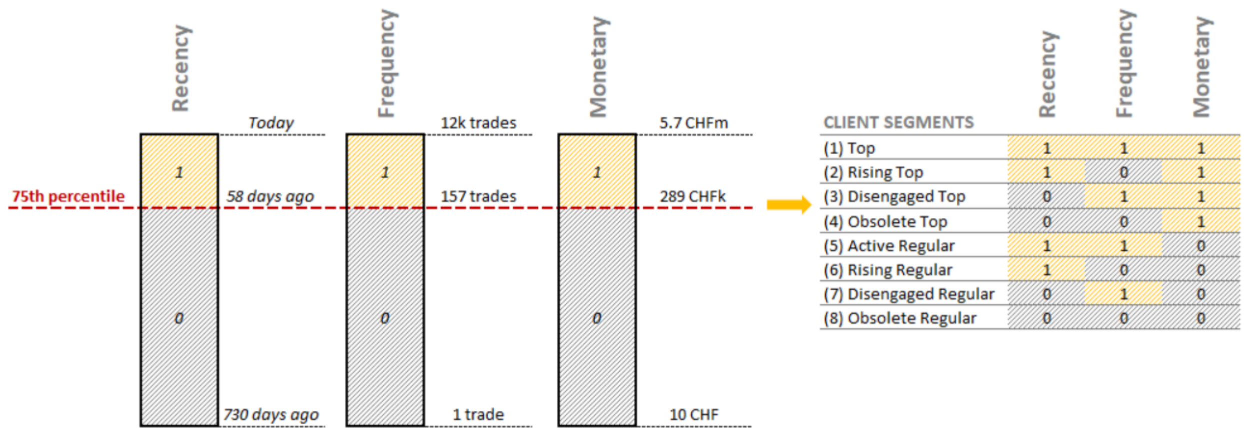


Figure 3: Implementation of the RFM client segmentation for institutions distributing Structured Investment Products (own work)

Constraints and advantages

Advantages of the RFM method include its simplicity, cost-effectiveness, and ease of implementation, making it accessible even to less experienced analysts. It provides actionable insights for optimizing client retention, prioritizing resources, and improving marketing. The analysis can be scaled to any dataset size and offers quantitative, data-driven insights.

However, disadvantages include a fixed number of segments, which can lead to redundant or overpopulated segments, and arbitrary selection of the percentile scoring threshold, where analyst decisions can introduce bias and affect accuracy.

3.3.3 K-Means clustering RFM client segmentation model

K-Means clustering - unsupervised machine learning model

To address the disadvantages of the above mentioned segmentation model and enhance the analysis the k-Means clustering algorithm can be applied to the same dataset used for the classical RFM segmentation. This algorithm classifies data by grouping similar data points together to uncover underlying patterns. It analyzes the properties of the data to determine the optimal way to divide the dataset, assigning discrete labels to each data point. The k-Means algorithm seeks to identify a fixed number of clusters (k) within the dataset. A cluster represents a group of data points that the algorithm assigns the same label, based on their similarities. Data points within the same cluster are supposed to be more similar to each other than to any data point in a different cluster.[7]

Data Preprocessing

To effectively apply the k-Means algorithm to the dataset, it must first be preprocessed to meet specific assumptions. **(1) Symmetric Distribution:** the data should have a symmetric (non-skewed) distribution. This can be achieved by applying a logarithmic transformation to the dataset, which helps normalize the distribution. **(2) Centering** involves adjusting the data so that the mean of each attribute becomes zero. Centering is important because k-Means relies on the Euclidean distance between points, and features with different means could bias the clustering process. **(3) Scaling** involves adjusting the range of features so that they all have the same variance. This is typically done by dividing each feature by its standard deviation after centering. Scaling is important because k-Means treats all features equally when calculating distances; features with larger ranges could dominate the clustering process if not scaled. The goal of these preprocessing steps is to ensure that all data points contribute equally to the clustering process, thereby preventing any feature from disproportionately influencing the k-Means algorithm. These steps—logarithmic transformation, centering, and scaling—can be conveniently carried out using the *StandardScaler()* method from the *sklearn.preprocessing* library.[8]

Determining the number of clusters

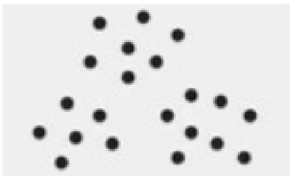
One of the benefits of performing client segmentation based on k-Means clustering algorithm is that we are not bound to a fixed number of segments. We can actually select the *right* number of segments, tailored to the characteristics of our dataset and the client base that we are analyzing. The general rule is that the segments should be characterized by distinct properties and should differ from each other. One approach to determining the optimal number of clusters involves running the clustering algorithm multiple times, for example in the range from k=1 segment to k=10 segments, and calculating the SSR (Sum of Squared Residuals) score for each number of segments. By plotting the SSR scores against the number of clusters, the analyst can identify the point on the graph, where increasing the number of clusters further leads to diminishing returns in reducing the SSR score. The point where the curve starts to flatten, indicates that adding more clusters doesn't significantly improve the clustering performance, and hence increasing the number of clusters is pointless. [7, 8]

[7] Van der Plas, J. (2017). *Python Data Science Handbook*. O'Reilly Media.
[8] DataCamp (2024). *Cluster Analysis in Python* [Online course]. Retrieved from https://www.datacamp.com

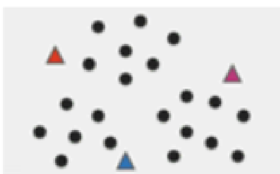
K-Means Algorithm

The algorithm operates as follows:

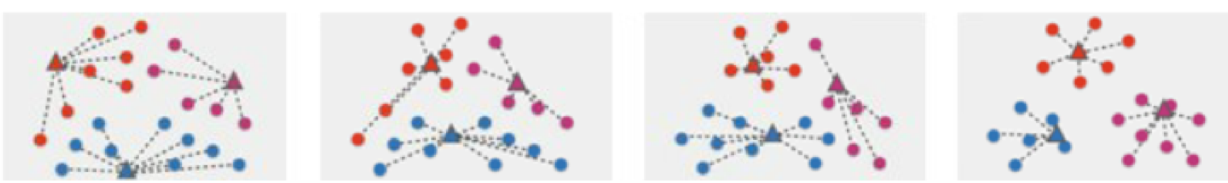
1. After preprocessing the data we obtain a set of unlabelled data points



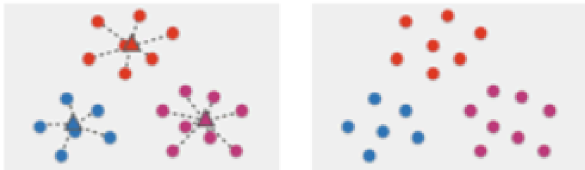
2. Next, we randomly initialize in the dataset the chosen k number of "centoids" - which are the cluster centers



3. The algorithm works using a concept called Expectation-Maximization (E-M). First, each data point is assigned to the closest centroid based on the Sum of Squared Residuals (SSR), which measures the sum of the distance from each point to its cluster center. Then, with each iteration of the algorithm the centroids "move around" and adjust their positions to minimize the SSR score and label the "nearest" data points. [7, 8]



4. The iteration process stops once all centroids reach their optimal position, meaning:
   - The SSR score is as low as possible.
   - Each data point is assigned to its nearest centroid.
   - The clusters are as concise and well-defined as possible [7]



Figure(s) 4: Schematic Illustration of the K-Means Clustering Process [7]

Naming convention and hierarchy.

After labeling all data points, the analyst's task is to assign meaningful segment names and rank the segments in order, from the highest-value clients to the lowest.

3.3.4 K-Means clustering RFM client segmentation model implementation

Due to very strict data secrecy policies in the financial industry the author cannot use the real data in this report, therefore this chapter presents the implementation of a K-Means clustering RFM segmentation model on a randomly generated dataset, simulating a typical customer base of an investment bank specializing in distribution of Structured Investment Products.

Figure 5: Determining the optimal number of clusters in the sample dataset (own work)

The first step in the analysis involves determining the optimal number of clusters. In this modeled dataset, the SSR score levels off at around the fourth cluster, which was selected for segmentation. Choosing the number of clusters is an arbitrary decision of the analyst and therefore can be considered as one of the weaknesses of the analysis. Ideally the decision regarding the number of clusters should be based not only on the calculating the SSR scores but also based on the general knowledge about the business that the analyst has. Figure 6 shows the average Recency, Frequency, and Monetary values for each cluster, along with the number of clients in each group. This dataset used in this example simulates a typical investment bank client base, characterized by significant disparities in trading frequency and volume across clients. Clients in the highest-value segment traded, on average, 14 days prior to the analysis execution day, completing 256 trades and generating an average of $1.7 million in revenue during the observation period. In contrast, clients in the lowest-value segment (cluster 4) traded for the last time over a year and three months before the analysis execution day, averaging just two trades and generating around $8,000 in revenue.

Cluster	R	F	M	Clients
1	14	256	1’702’945	118
2	25	31	139’704	134
3	246	20	216’476	186
4	473	2	8’327	188
			TOTAL	**626**

Figure 6: Average values of Recency, Frequency, and Monetary attributes in the simulated client dataset and corresponding client counts (own work)

The 3D scatter plot below visualizes the unlabelled dataset, which will be used for the segmentation analysis, 1 dot represents 1 client:

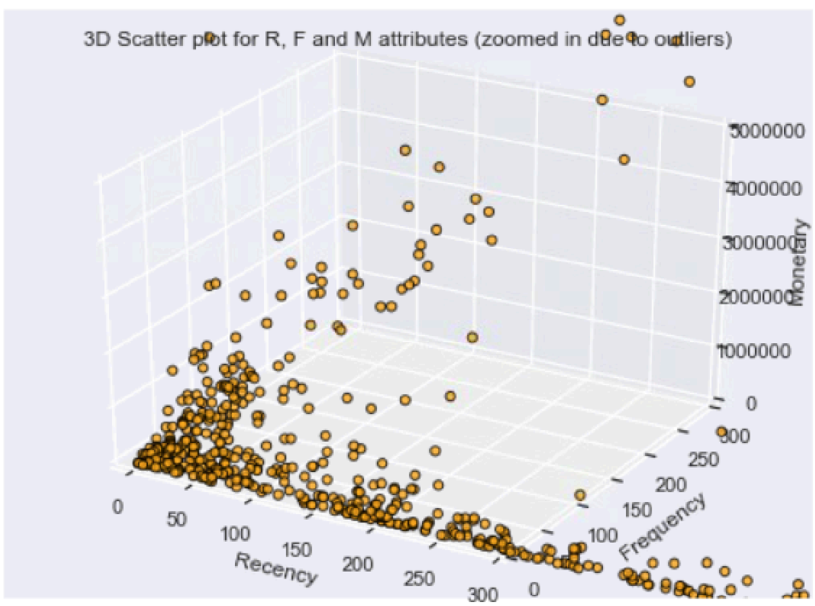


Figure 7: Scatter Plot of the RFM Attributes of the clients in the simulated dataset (own work)

Next the dataset has to be preprocessed to meet the requirements of the clustering algorithm. This preprocessing includes scaling, centering, and ensuring a symmetric distribution. The effect of the preprocessing is presented in Figure 8.

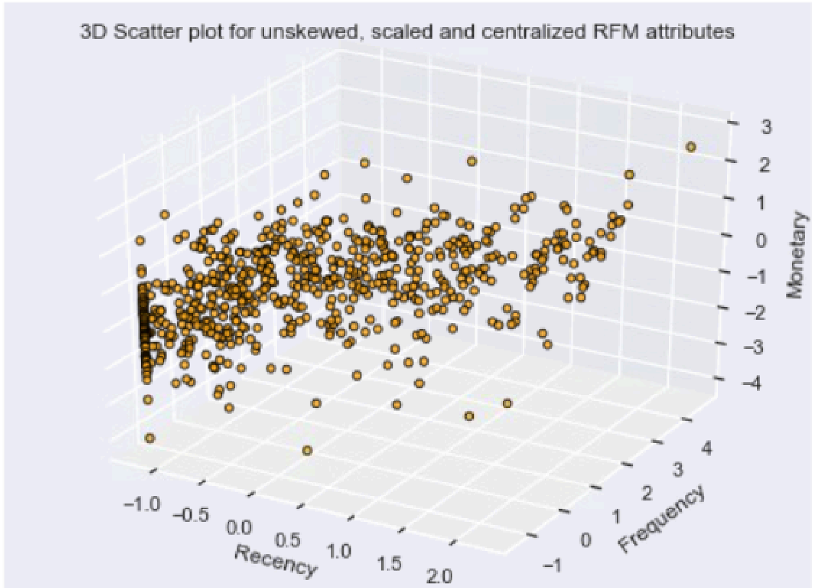


Figure 8: Scatter Plot of the preprocessed RFM Attributes of the clients in the simulated dataset (own work)

The dataset is now prepared for the application of the K-Means clustering algorithm. The analyst selects the chosen number of clusters and initiates the algorithm. As a result, each data point is assigned to one of four labels, as illustrated below.

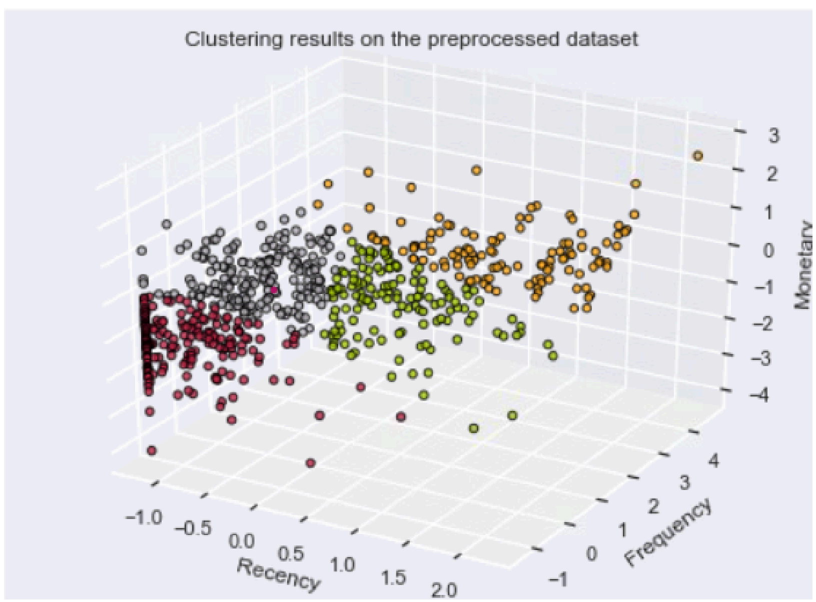


Figure 9: K-Means clustering results in the preprocessed dataset (own work)

Finally, the cluster labels are applied to the original client dataset. The scatter plot below presents the final result of the clustering. In the final step, the analyst ranks the clusters hierarchically, from the most valuable clients to the least, and assigns meaningful names to each cluster. The segmentation is finished and ready for interpreting the results.

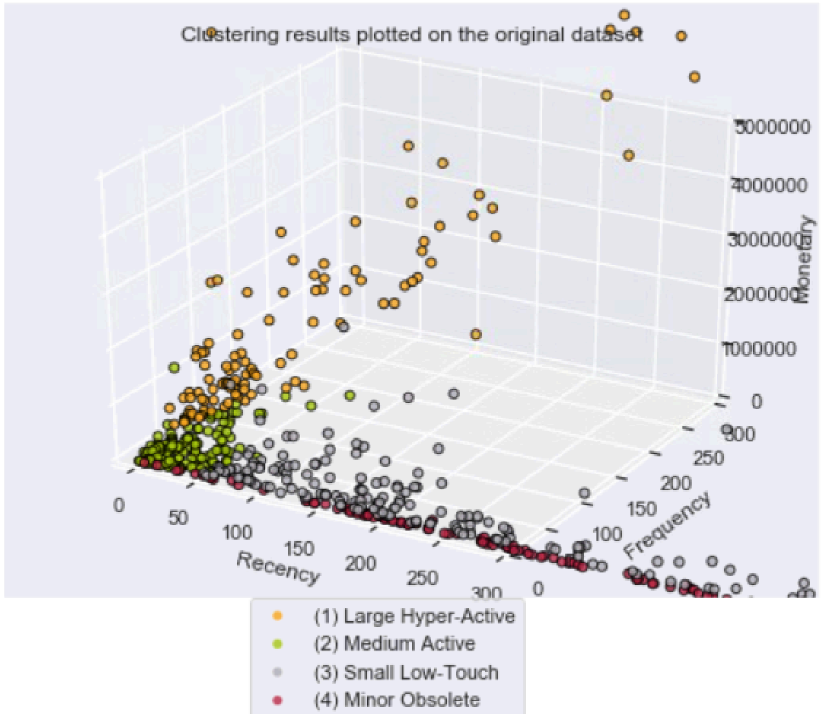


Figure 10: K-Means clustering segmentations results (own work)

4 Data and Metadata

The analysis presented in this paper will utilize the company's transactional data sourced from a category of applications known as "Order Management Systems." The minimum data requirements are quite rudimentary. The dataset for this analysis should include the following four key attributes: unique trade identifier, trade date, client identifier, and fee amount. Typically, this data is collected from an enterprise-level data warehouse, necessitating minimal preprocessing and cleaning as described further in the "Data Flow" and "Data Quality" chapters. To enhance the analysis and provide deeper insights, the analyst may consider adding additional attributes to the segmented dataset, such as client owner salesperson, client domicile, client type, client onboarding date etc.

Due to data confidentiality, the analysis will be conducted entirely on a company-provided virtual desktop environment. Depending on the needs or a chosen method of publishing the results of the analysis of the target audience of the report. The data can be stored in excel format, and for more extensive sizes of data in CSV or pandans pickle format.

5 Data Flow

At the very beginning of this chapter, it is important to emphasize that the data flow described here offers a simplified and highly generalized picture of the actual data flow. Additionally, it focuses solely on aspects and potential data quality issues relevant to the analysis at hand. Providing a comprehensive overview of the entire IT landscape would not be feasible within the scope of this paper.

Clients buy and sell Structured Investment Products using applications referred to in Figure 11 as "Order Entry Applications." These applications can function as self-service platforms, allowing clients to structure and trade products independently, which is common among clients who conduct a large number of low-volume transactions. Alternatively, products can be structured and trades executed by a salesperson on behalf of the client. This approach is more typical for higher-volume trades with larger fees, where the transaction is preceded by extensive advisory discussions and negotiations of the product's parameters and fees. Additionally, retail clients can purchase pre-structured products in the primary market or trade the products on the secondary market through their brokerage accounts.

Once a deal is settled and the trade is released from the "Order Entry Application," it is recorded in the "Order Management System." This system stores all the essential trade details, including product specifics, invested notional, fees, trade date, and client information. At this stage, the trade undergoes a final review by the trading department before receiving final approval.

Once final approval is granted, the trade is executed. At this point, all underlying assets of the product are purchased on the market, hedge positions are opened, and all components of the product are booked in the "Position Keeping System." With the product now actively trading, its price is quoted, and events such as coupon payments, barrier events, or redemptions are processed as they occur.

Daily ETL processes extract, transform, and load data from the company's applications into a centralized database. The data extracted from the source application is typically preprocessed, enriched, and structured into a data warehouse or a data mart. Order data, for instance, is often organized into a star schema within a data warehouse. The fact table in such a schema might include key metrics like order ID, trade date, invested amount, and fees, while the dimension tables might include product, client, salesperson, and order type (e.g., buy, sell, primary, secondary market).

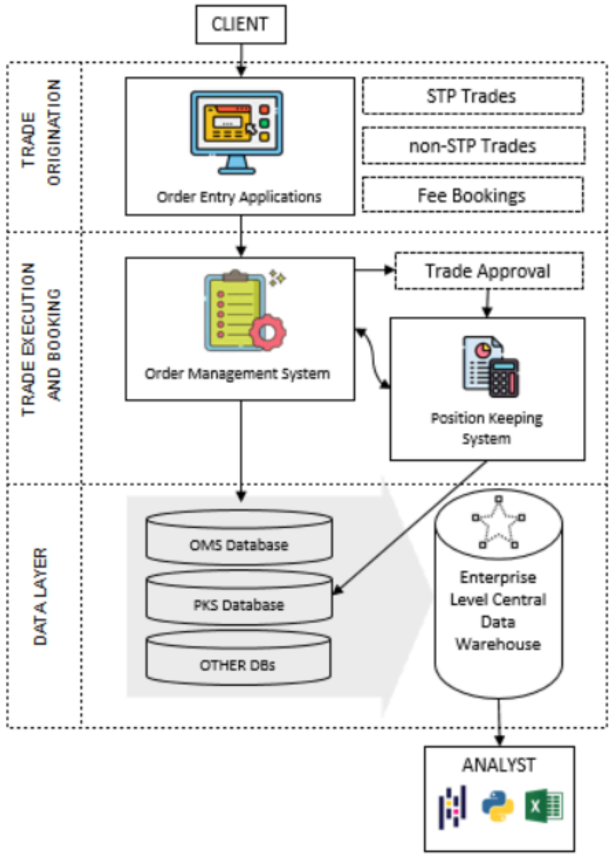


Figure 11: Simplified order data flow in an investment Institution distributing structured investment products (own work)

Having access to this structured, preprocessed, and normalized dataset is a significant advantage for the analyst, allowing them to focus entirely on conducting the analysis without needing to spend time on data preparation tasks.

6 Data Quality

Since this data analysis is conducted using a single company's trading record dataset, there is no concern about selecting a representative sample; the analysis is performed on the entire history of orders within the observation period. The size of the dataset can vary depending on the company's size, client base, and trading activity, ranging from a few thousand transactions to potentially several million. Given the rigorous controls, approvals, and validations inherent in the banking industry, the data is expected to be highly accurate, with missing values being a rare exception.

Tanking aside recurring fee bookings orders can be broadly classified into 2 categories:

1. **STP Trades (Straight Through Processing)**: These trades are fully automated, with no manual intervention required from the Front, Middle, or Back Office. The automated nature of STP trades includes built-in safeguards throughout the booking process, which significantly reduces the likelihood of data quality issues.
2. **Non-STP Trades**: These trades require manual intervention by the operations team, making them more susceptible to errors. Examples of data quality errors that might influence the RFM Segmentation might be:
    a. Incorrect trade currency, an extreme example would be booking a trade in KRW (Korean Won) instead of CHF (Swiss Franc), which could lead to significant discrepancies due to the exchange rate difference (1:1577). Such issues can often be identified by performing descriptive statistics on the dataset and closely examining extreme outliers.
    b. Another common issue is missing customer identifiers, which usually occurs when a secondary market trade is executed via an exchange, making it initially impossible to identify the client behind the trade. The operations team must investigate each case to confirm the client. Occasionally, this essential information for client segmentation might be temporarily unavailable (null). To address this, the analyst can isolate the null values, review past trades involving the same product, and identify the client based on historical data.

7 Data Model

At the conceptual level the primary goal of this analysis is to provide a reliable and easy to use tool for companies distributing structured investment products, enabling them to gain valuable insights into their client database. This analysis seeks to identify high-value clients, clients at risk of churn, clients who have completely disengaged, and clients with significant growth potential. By delivering these insights, the analysis can assist sales and marketing teams in precisely targeting their marketing and retention efforts. Additionally, it can help optimize the sales force's approach to effectively serve the customer base and objectively measure their success in developing customer relationships.

At the logical level, the minimal requirements regarding the attributes necessary to conduct the analysis are:

- **Unique Client Identifier** (numeric or string)
- **Trade Date** (date)
- **Client Identifier** (numeric or string)

- **Fee Amount** (float)

At the physical level, the infrastructure required for this analysis is quite rudimentary and should typically be available in any mature organization engaged in this type of business. The analyst should be authorized to have access to a database storing company's transactional data, the minimal requirements to conduct the analysis in its classical form is just an excel spreadsheet and for the K-Mean clustering segmentation additionally Python with the necessary libraries installed.

8 Risks

The risks associated with this project are usually low. RFM client segmentation, whether in its classical form or utilizing the k-Means algorithm, is relatively straightforward to implement. It relies on readily available data at most companies that deal with Structured Investment Products, making it an efficient endeavor in terms of the analyst's time commitment. Moreover, given the nature of the industry and its high compliance standards, the likelihood of data quality issues should be rare and exceptional.

However, there are potential bottlenecks in this analysis. The greatest risk lies in the segmentation process itself, particularly in determining the percentile score levels for classic RFM and selecting the number of segments for k-Means RFM analysis. These parameters are often chosen arbitrarily by the analyst, which can lead to incorrect decisions and can introduce a bias. If the parameters are misselected, it could skew the results of the analysis. The analyst may lack sufficient insights into the customer base, or possess inadequate business knowledge, all of which could hinder optimal parameter selection.

9 Conclusions

In summary, this conceptual project design report presents the application of the RFM client segmentation model, both in its classical form and enhanced with the K-Means clustering algorithm, for financial institutions distributing structured investment products. This approach provides an efficient, data-driven method for analyzing client bases. The insights gained from this analysis enable financial institutions to identify high-value clients, target those at risk of disengagement, optimize resource allocation, and improve marketing strategies. By leveraging these tools, investment banks can enhance customer engagement and drive more informed business decisions.

Statement

The following part is mandatory and must be signed by the author or authors.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date: Signature(s):

02.10.2024

References and Bibliography

Van der Plas, J. (2017). *Python Data Science Handbook*. O'Reilly Media.

DataCamp (2024). *Cluster Analysis in Python* [Online course]. Retrieved from https://www.datacamp.com

Blattberg, Robert C. (1994). *Database Marketing: Analyzing and Managing Customers.* Springer Science & Business Media

Bärlocher, Jürg. *Experience with the Collection and Publication of Data on Structured Products in Switzerland.* BIS (Bank for International Settlements), IFC Bulletin No. 31, 1 September 2009. Available at: https://www.bis.org/ifc/publ/ifcb31l.pdf

Swiss Structured Products Association (SSPA). *The World of Structured Products.* SSPA, 2023. Available at: https://sspa.ch/en/book/

Die Welt der Strukturierten Produkte: Das Standardwerk für Strukturierte Produkte. Swiss Structured Products Association (SSPA), 2023.

Swiss Structured Products Association (SSPA). *Swiss Map of Structured Products 2024.* SSPA, December 2023. Available at: https://sspa.ch/wp-content/uploads/2023/12/en_sspaswissmap_2024.pdf

Payoff.ch. *Term Sheet for Structured Product CH1369852335.* 21 August 2024. Available at: https://data.payoff.ch/termsheets/CH1369852335_de_20240821_090251.pdf

Diaz Ruiz, Carlos A. (2024). *Elgar Encyclopedia of Consumer Behavior*, Edward Elgar Publishing