RFM Analysis Methodology

Contents

```r
# Load necessary libraries
library(ggplot2)
library(factoextra)
library(cluster)

library(plotly)
library(dplyr)


# Step 1: Create Sample RFM Data

####---Data Uploading---###

# Upload observed data using a .txt file
# Substitute "K-Means_Data.txt" with your Data file name.
# Remember to keep your Data file in the same folder as your
    Code file
# Setting: Session -> Set Working Directory -> To Source File
    Location

rfm_data = read.table("K-Means_Data_RFM_.txt",
                      sep = "\t",
                      header = TRUE,
                      fill = FALSE)

# Step 2: Scale the RFM Variables
###---Data Preparation---###

# We create a table with only relevant variables for the
    clusterisation
rfm_points_data_cluster = subset(
  rfm_data,
  select = c(CustomerID, R_points, F_points, M_points)
)

# Selecting only RFM points
rfm_points <- rfm_points_data_cluster[, c("R_points",
    "F_points", "M_points")]

# Step 3: Apply K-Means Clustering
###---Analysis---###
```

```r
set.seed(123)   # For reproducibility
kmeans_result_points <- kmeans(rfm_points, centers = 4, nstart
    = 25)

# Elbow Method for K selection
fviz_nbclust(rfm_points, kmeans, method = "wss")

# Step 4: Add Cluster Assignments to Original Data
rfm_data$Cluster_points <-
    as.factor(kmeans_result_points$cluster)

# Step 5: Visualize the Clusters
###---MOST RELEVANT FEATURES---###

fviz_cluster(kmeans_result_points,
             data = rfm_points,
             geom = "point",
             ellipse.type = "norm",
             ggtheme = theme_minimal(),
             main = "K-Means Clustering of RFM Data Points")

# PCA Analysis
pca_res <- prcomp(rfm_points, scale. = TRUE)
summary(pca_res)

fviz_contrib(pca_res, choice = "var", axes = 1)
fviz_contrib(pca_res, choice = "var", axes = 2)

# Visual exploration
ggplot(rfm_data,
       aes(x = R_points,
           y = F_points,
           color = Cluster_points)) +
  geom_point(size = 4) +
  scale_color_manual(values = c("brown1",
                                "darkolivegreen2",
                                "cadetblue2",
                                "mediumorchid1")) +
  labs(title = "Customer Segmentation based on RFM Scores",
       x = "R_points",
```

```r
         y = "F_points") +
   theme_minimal ()

ggplot(rfm_data,
        aes(x = R_points,
             y = M_points,
             color = Cluster_points)) +
   geom_point(size = 4) +
   scale_color_manual(values = c("brown1",
                                 "darkolivegreen2",
                                 "cadetblue2",
                                 "mediumorchid1")) +
   labs(title = "Customer Segmentation based on RFM Scores",
        x = "R_points",
        y = "M_points") +
   theme_minimal ()

# 3D Visualization
p <- plot_ly(rfm_data,
             x = ~R_points,
             y = ~F_points,
             z = ~M_points,
             color = ~Cluster_points,
             colors = c("brown1",
                        "darkolivegreen2",
                        "cadetblue2",
                        "mediumorchid1")) %>%
   add_markers(size = 1.5)

print(p)
```

Listing 1: Data simulation in R

1 Overview and timeframe

We decided to conduct two separate RFM analyses over an 18-month timeframe (between 01/01/2024 and 30/06/2025) , assuming we launched our new service on January 1st, 2024 :

- **RFM for private customers**

- **RFM for families**

The choice of an 18-month timeframe is driven by the dynamic nature of our service, which is built around anniversaries, new film releases, and seasonal holidays. An 18-month observation window allows us to:

- **monitor purchase frequency across two different seasonal cycles**: this prevents the analysis from being skewed by seasonal anomalies or the outlier success of a single major event;

- **distinguish between "heavy users" and "dormant users"**: for those holding an Experience Card, 18 months provides the necessary time to see if they exhaust their credits quickly or remain inactive, allowing us to calculate real frequency rather than partial data;

- **identify post-launch stabilization**: as a new service, an initial peak in patronage is expected due to the "novelty effect." This extended timeframe helps us distinguish truly loyal customers from those who only engaged with the service because of the initial launch hype;

- **assess family engagement with event format**: families typically plan their attendance at immersive events—such as karaoke or costume parties—around major holidays and blockbuster animation releases. Since the lead time between a film's announcement (trailer release) and its actual theatrical debut often exceeds one year, a 12-month window would be insufficient. Therefore, an 18-month period is essential to verify if a family unit genuinely valued the experience enough to return for a subsequent animated release.

Private customers include consumers who pay for a single ticket, those who purchase a Life Moment pass, and those who purchase an Experience card pass. It should be noted that if the same person attends both individually and as part of a family group, they are assigned different IDs, as they belong to two distinct analytical populations and therefore to two separate RFM analyses. Instead, a private individual who initially buys single tickets and then decides to subscribe is identified with the same ID.

Schools are also among our clients; however, they are treated as business customers. For reasons of methodological simplicity, and to avoid highly complex scenarios involving public tenders and institutional funding, schools are excluded from the RFM analysis.

For both analyses, we consider the following variables:

- **Recency (R)**: number of days from the date of the last purchase made compared to the reference date (snapshot date).

- **Frequency (F)**: frequency of service consumption.

- **Monetary (M)**: total expenditure incurred by each customer.

2 RFM analysis of private customers

For the RFM analysis of private customers we considered a sample of 1000 consumers (table photo).

The table is structured into 9 columns, each representing a key variable of the analysis:

- **ID**: unique codes associated with each consumer;

- **Last purchase**: date of the most recent purchase;

- **Recency**;

- **Monetary**;

- **Frequency**;

- **R_Points**: score assigned to the Recency variable on a scale of 1 to 5, calculated based on the quantiles of the distribution;

- **M_points**: score assigned to the Monetary variable on a scale from 1 to 5, calculated based on the quantiles of the distribution;

- **F_Points**: score assigned to the Frequency variable on a scale from 1 to 5, calculated based on the quantiles of the distribution;

- **RFM_Points**: overall score obtained by combining the R, F, and M scores.

2.1 Calculation of the Monetary range for random data generation

In order to generate realistic random data for the RFM analysis, a value range for the Monetary (M) variable was defined between €40 and €1,500. The determination of these bounds was not arbitrary but based on assumptions consistent with plausible consumer behavior.

The **upper bound** of the range was estimated by considering a realistic maximum case, represented by a highly engaged customer who decides to renew the Experience Card—the most expensive subscription option—at least four times over the 18 months analyzed. This assumption implies participation in at least one third of the events organized in each four-month period. In this scenario, the expenditure associated with subscriptions alone would amount to approximately €300 per renewal, for a total of €1,200 over the entire observation period.

However, taking into account the possibility that the same consumer made additional purchases, such as single tickets, gadgets or catering services, it was considered appropriate to extend the upper limit of the range up to €1,500.

The **lower bound** of the range instead reflects the opposite extreme scenario, namely that of a casual consumer who attended only a single event and decided not to repeat the experience. In that case, the overall expenditure is reduced to the purchase of one single-entry ticket priced at €40, which therefore constitutes the lower bound of the Monetary variable.

2.2   Calculation of frequency range for random data generation

To define the range of the **Frequency (F)** variable for the generation of random data in the RFM analysis, we considered values between 1 and 50. As in previous cases, the choice of range limits is based on realistic assumptions regarding the frequency of service use.

The **upper bound** of the range was chosen considering the frequency with which we organize events reserved for private individuals and families, that is, approximately eight times a month (two events a week). Throughout the timeframe, we will organize a total of:

$$8 \text{ events/month} \times 18 \text{ months} = 144 \text{ events.}$$

Consistent with the assumptions adopted for the Monetary variable, the maximum realistic scenario is represented by a highly engaged customer who renews the Experience Card—the most expensive subscription option—at least four times over the 18-month period (every four months). This type of customer uses our service approximately 40 times in the timeframe.

In order to introduce a safety margin and account for potential variability in customer behavior, we consider a margin of another 10 times, bringing the upper bound to 50 times.

The **lower bound** of the range reflects the worst scenario, in which a consumer attends a single event and subsequently decides not to repeat the experience (so its frequency will be equal to 1).

2.3   Calculation of the range of last purchase dates for random data generation

For the generation of random dates corresponding to the last purchase, we considered the start and end dates of the observation timeframe, namely 01/01/2024 and 30/06/2025.

2.4 Explanation of how the random data for Frequency and Monetary were generated

We have set ourselves the problem of ensuring a certain proportionality between the frequency of consumption and the monetary value of each consumer.

Given the specific characteristics of our service, which relies on fixed pricing schemes where a single-event ticket costs €40, the Experience Card subscription (valid for 10 events) costs €300, and the Life Moment subscription (valid for 4 events) costs €120—it is unlikely that a consumer who attended only one event (F=1) spent a lot (e.g., M=€500) or another who is a frequent visitor (with very high attendance values) spent relatively little.

In order to more or less ensure compliance with this proportionality, we decided to link the two variables (Frequency and Monetary) through a third "fictitious" variable, which we named **Reference**. As you can see from the image on the side, this third variable was generated using the Excel function `CASUALE()`, which generates random values on the range $[0, 1]$.

Reference
0,70
0,94
0,29
0,06
0,10
0,06
0,44
0,82
0,34
0,51
0,75
0,82
0,19
0,52
0,09
0,05
0,09
0,40
0,51
0,73
0,15
0,19
0,15

After that, for each Customer ID, we calculated Frequency and Monetary as follows:

$$\text{Frequency} = \texttt{ARROTONDA}\Big(\text{Reference} \cdot (\text{max\_frequency} - \text{min\_frequency}) + \text{min\_frequency}; 0\Big), \tag{1}$$

$$\text{Monetary} = \texttt{ARROTONDA}\Big(\text{Reference} \cdot (\text{max\_monetary} - \text{min\_monetary}) + \text{min\_monetary}; 0\Big). \tag{2}$$

In our case, given the selected ranges:

$$\text{Frequency} = \texttt{ARROTONDA}(\text{Reference} \cdot (50 - 1) + 1; 0), \tag{3}$$

$$\text{Monetary} = \texttt{ARROTONDA}(\text{Reference} \cdot (1500 - 40) + 40; 0). \tag{4}$$

=ARROTONDA((R2)*(50-1)+1;0)

monetary	frequency	Rpoint	Fpoint	Mpoint	RFM point
817	27	4	4	4	444
192	6	1	1	1	111
1007	33	4	4	4	444
532	18	2	2	2	222
1049	35	4	4	4	444
941	31	4	4	4	444
832	28	4	4	4	444
339	11	1	1	1	111
278	9	1	1	1	111

=ARROTONDA((R2)*(1500-40)+40;0)

monetary	frequency	Rpoint	Fpoint	Mpoint	RFM point
817	27	4	4	4	444
192	6	1	1	1	111
1007	33	4	4	4	444
532	18	2	2	2	222
1049	35	4	4	4	444
941	31	4	4	4	444
832	28	4	4	4	444
339	11	1	1	1	111
278	9	1	1	1	111

with R2 = Reference for CUST001

However, the correlation between the frequency and monetary columns turns out to be too perfect, since by using the same control variable for both the R and F variables. It has in fact generated almost perfectly overlapping variables, which can be assimilated to gods "clones" from a mathematical point of view.

Instead, in a realistic RFM analysis, there is a need for correlation, but not perfect proportionality: there must be a trend (those who come often tend to spend more), but with exceptions (e.g., those who come little but buy the €300 Experience Card). For this reason, we decided to introduce an additional variable representing a "disturbance factor" variable called **Storm**,which introduces random noise, with a weight of 30 with respect to the variable "Reference". The values assumed by this variable range within the interval [0, 1] and are generated using the Excel the `CASUALE()` function.

Storm
0,96
0,11
0,76
0,83
0,72
0,33

As a result, the previous formulas are modified as follows:

$$\text{Frequency} = \texttt{ARROTONDA}\big(\text{Reference} \cdot (\text{max\_frequency} - \text{min\_frequency}) \qquad (5)$$
$$+ \text{min\_frequency}; 0\big)$$

$$\text{Monetary} = \texttt{ARROTONDA}\big(\text{Reference} \cdot (\text{max\_monetary} - \text{min\_monetary}) \qquad (6)$$
$$+ \text{min\_monetary}; 0\big)$$

In our specific case:

$$\text{Frequency} = \texttt{ARROTONDA}((0.7 \cdot \text{Reference} + 0.3 \cdot \text{Storm}) \cdot (50 - 1) + 1; 0), \qquad (7)$$
$$\text{Monetary} = \texttt{ARROTONDA}((0.7 \cdot \text{Reference} + 0.3 \cdot \text{Storm}) \cdot (1500 - 40) + 40; 0). \qquad (8)$$

=ARROTONDA((R2*0,7+Q2*0,3)*(50-1)+1;0)

	D	E	F	G	H	I
	monetary	frequency	Rpoint	Fpoint	Mpoint	RFM point
	1005	33	1	4	4	144
	196	6	5	1	1	511
	1048	35	5	4	5	545
	747	25	5	2	2	522
	1063	35	5	4	5	545
	817	27	4	4	4	444
	659	22	2	2	2	222
	295	10	5	1	1	511

=ARROTONDA((R2*0,7+Q2*0,3)*(1500-40)+40;0)

	D	E	F	G	H	I
	monetary	frequency	Rpoint	Fpoint	Mpoint	RFM point
	1005	33	1	4	4	144
	196	6	5	1	1	511
	1048	35	5	4	5	545
	747	25	5	2	2	522
	1063	35	5	4	5	545
	817	27	4	4	4	444
	659	22	2	2	2	222
	295	10	5	1	1	511

with R2 = Reference for CUST001 and Q2 = Storm for CUST001.

2.5 Explanation of how the random data for Recency were generated

To calculate Recency, we generated random dates corresponding to the last purchase of each customer within the interval [01/01/2024, 30/06/2025] (Last Purchase), using the Excel function `CASUALE.TRA(45292;45838)`, where 45292 is the numerical representation of 01/01/2024 and 45838 corresponds to 30/06/2025.

Subsequently, **Recency** was computed as the difference between the snapshot date (30/06/2025) and the last purchase date for each customer, thereby obtaining the number of days elapsed since the most recent transaction.
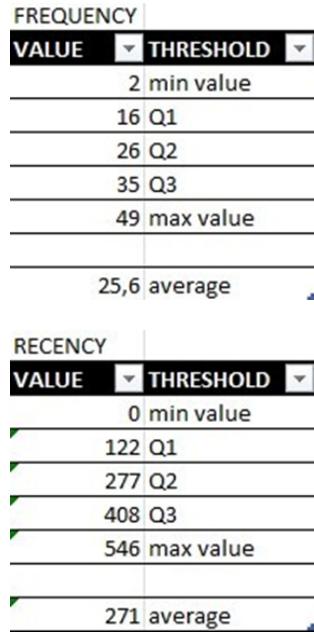
2.6 Calculation of R_Points, M_Points, and F_Points

For each variable of the RFM analysis we calculated the quantities that characterize the distribution of their values, which are:

- **minimum value**,

- **first quartile (Q1)**,

- **second quartile (Q2)**,

- **third quartile (Q3)**,

- **maximum value**,

- **average**.

Below we report the tables (also present in the Excel attachment), each relating one of the three variables, in which we have written the corresponding values found for each of these quantities:

MONETARY VALUE	THRESHOLD
57	min value
500	Q1
780	Q2
1046	Q3
1461	max value
773,339	average

12

FREQUENCY		
VALUE	**THRESHOLD**	
2	min value	
16	Q1	
26	Q2	
35	Q3	
49	max value	
25,6	average	

RECENCY		
VALUE	**THRESHOLD**	
0	min value	
122	Q1	
277	Q2	
408	Q3	
546	max value	
271	average	

2.6.1 Score assignment for Monetary

The score assignment (from 1 to 5) for the Monetary variable follows the scheme below:

- $M\_points = 1$ if Monetary $\leq Q1$

- $M\_points = 2$ if $Q1 <$ Monetary $\leq$ average

- $M\_points = 3$ if average $<$ Monetary $\leq Q2$

- $M\_points = 4$ if $Q2 <$ Monetary $\leq Q3$

- $M\_points = 5$ if $Q3 <$ Monetary $\leq$ max value

2.6.2 Score assignment for Frequency

The score assignment (from 1 to 5) for the Frequency variable follows the scheme below:

- $F\_points = 1$ if Frequency $\leq Q1$

- $F\_points = 2$ if $Q1 <$ Frequency $\leq$ average

- $F\_points = 3$ if average $<$ Frequency $\leq Q2$

- $F\_points = 4$ if $Q2 <$ Frequency $\leq Q3$

- $F\_points = 5$ if $Q3 <$ Frequency $\leq$ max value

2.6.3 Score assignment for Recency

The score assignment (from 1 to 5) for the Recency variable follows the pattern below:

- $R\_points = 5$ if Recency $\leq Q1$

- $R\_points = 4$ if $Q1 < \text{Recency} \leq \text{average}$

- $R\_points = 3$ if $\text{average} < \text{Recency} \leq Q2$

- $R\_points = 2$ if $Q2 < \text{Recency} \leq Q3$

- $R\_points = 1$ if $Q3 < \text{Recency} \leq \text{max value}$

Finally we created the box plots for each distribution of the variables (Monetary, Frequency, Recency) and associated the corresponding `RFM_points` with each Customer ID.
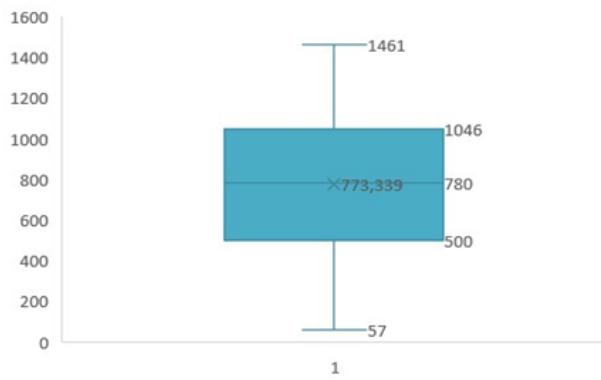

Figure 1: Box plot – Monetary (private customers).
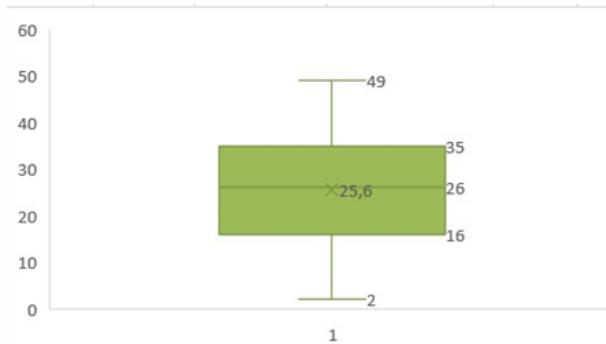
Figure 2: Box plot – Frequency (private customers).

14
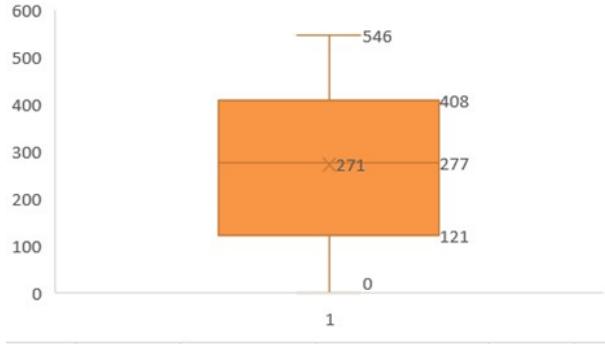
Figure 3: Box plot – Recency (private customers).

3 RFM analysis of families customers

For the RFM analysis of family customers we considered a sample of 1000 consumers (table photo).

The table is structured into 9 columns, each representing a key variable of the analysis:

- **ID**: unique codes associated with each consumer;

- **Last purchase**: date of the most recent purchase;

- **Recency**;

- **Monetary**;

- **Frequency**;

- **R_Points**: score assigned to the Recency variable on a scale of 1 to 5, calculated based on the quantiles of the distribution;

- **M_points**: score assigned to the Monetary variable on a scale from 1 to 5, calculated based on the quantiles of the distribution;

- **F_Points**: score assigned to the Frequency variable on a scale from 1 to 5, calculated based on the quantiles of the distribution;

- **RFM_Points**: overall score obtained by combining the R, F, and M scores.

3.1 Calculation of the Monetary range for random data generation

For the calculation of the Monetary (M) range, aimed at generating random data for our RFM analysis, we considered a value interval between €40 and €4,000.

Our family package provides that each family receives one free ticket: for example, a family of four people will pay the equivalent of three tickets. Therefore, instead of paying

€40 × 4 = €160, they will pay €40 × 3 = €120. The package applies starting from family units composed of at least two people, provided that at least one child is present.

The calculation of the **upper bound** of the range was not arbitrary. We considered as the maximum realistic case a family composed of six people, which would spend approximately €200 per event. Assuming that such a family is particularly loyal, it would attend about one event per month; therefore, considering our time frame, it would participate in approximately 18 events in total. Consequently, a family of this type would spend around €200 × 18 = €3,600 over the entire time frame considered.

However, since the family may have purchased merchandise or used catering services, and in order to account for more extreme situations (such as larger families), we decided to raise the upper limit of the range to €4,000.

As for the **lower bound** of the range, it corresponds to the opposite extreme case, namely a family composed of only two people (for example, in the case of divorced families with one child, who attends the events with either the mother or the father). In this situation, the family would pay for only one ticket (€40) and would have attended just one event, which they did not enjoy, thus deciding not to return.

3.2 Calculation of frequency range for random data generation

To define the range of the **Frequency (F)** variable for the generation of random data in the RFM analysis, we considered values between 1 and 20. As in previous cases, the choice of range limits is based on realistic assumptions regarding the frequency of service use.

The **upper bound** of the range was chosen considering the frequency with which we organize events reserved for private individuals and families, that is, approximately eight times a month (two events a week). Throughout the timeframe, we will organize a total of:

$$8 \text{ events/month} \times 18 \text{ months} = 144 \text{ events}.$$

As in the case of **Frequency**, we assumed as the maximum realistic scenario a family composed of six people. Assuming that this family is particularly loyal, it would attend approximately one event per month; therefore, considering our time frame, it would take part in about 18 events in total. In order to introduce a safety margin and account for possible behavioral variations, we considered an additional margin of two events, thus setting the upper limit at 20.

As for the **lower bound** of the range, we considered the worst-case scenario, namely a family composed of two people, at least one of whom is under 14 years old, that attends only one event and subsequently decides not to repeat the experience (therefore, its frequency equals 1).

3.3 Calculation of the range of last purchase dates for random data generation

For the generation of random dates corresponding to the last purchase, we considered the start and end dates of the observation timeframe, namely 01/01/2024 and 30/06/2025.

3.4 Explanation of how the random data for Frequency and Monetary were generated

We addressed the issue of ensuring a certain degree of proportionality between the frequency of consumption and the monetary value for each family. Given the specific characteristics of our family package, it is unrealistic that a family that attended our events only once (F = 1) would have spent a large amount (e.g., M = €500), or that another family that is a very frequent attendee (with very high frequency values) would have spent a relatively low amount.

In order to more or less ensure compliance with this proportionality, we decided to link the two variables (Frequency and Monetary) through a third "fictitious" variable, which we named **Reference**.

Reference
0,70
0,94
0,29
0,06
0,10
0,06
0,44
0,82
0,34
0,51
0,75
0,82
0,19
0,52
0,09
0,05
0,09
0,40

As you can see from the image on the side, this third variable was generated using the Excel function `CASUALE()`, which generates random values on the range $[0, 1]$.

After that, for each Customer ID, we calculated the frequency and monetary, respectively as follows: (ranges: Frequency $[1, 20]$, Monetary $[40, 4000]$):

$$\text{Frequency} = \texttt{ARROTONDA}(\text{Reference} \cdot (20 - 1) + 1; 0), \tag{9}$$

$$\text{Monetary} = \texttt{ARROTONDA}(\text{Reference} \cdot (4000 - 40) + 40; 0). \tag{10}$$

However, the correlation between the frequency and monetary columns turns out to be too perfect. In a realistic RFM analysis, however, correlation is required, but not perfect proportionality: there must be a general tendency (those who attend more frequently tend to spend more), while also taking into account the number of members in a family, which has a significant impact on total expenditure. In fact, it may occur that a very large family, even if it is not a frequent user of our service, still incurs a very high level of spending, which is therefore linked not so much to its consumption frequency as to the number of people composing the family.

For this reason, we introduced a disturbance factor variable called **Storm** (random in $[0, 1]$ using `CASUALE()`), with a weight of 50% with respect to **Reference**. The formulas become:

Storm
0,47
0,33
0,81
0,75
0,67
0,33
0,96

$$\text{Frequency} = \text{ARROTONDA}((0.5 \cdot \text{Reference} + 0.5 \cdot \text{Storm}) \cdot (20 - 1) + 1; 0), \tag{11}$$

$$\text{Monetary} = \text{ARROTONDA}((0.5 \cdot \text{Reference} + 0.5 \cdot \text{Storm}) \cdot (4000 - 40) + 40; 0). \tag{12}$$

=ARROTONDA((R2*0,5+0,5*Q2)*(4000-40)+40;0)

D Monetary	E Frequency	F R_points	G F_points	H M_points	I RFM_points
2363	12	1	4	4	144
2555	13	1	4	4	144
2218	11	4	3	4	434
1654	9	5	2	2	522
1567	8	1	1	2	112
811	5	2	1	1	211
2814	14	4	5	5	455
1847	10	4	2	2	422
1519	8	5	1	2	512
1471	8	2	1	2	212
2303	12	1	4	4	144
3634	18	1	5	5	155
918	5	2	1	1	211
1792	9	1	2	2	122

with R2 = Reference for CUST001 and Q2 = Storm for CUST001.

=ARROTONDA((R2*0,5+Q2*0,5)*(20-1)+1;0)

D	E	F	G	H	I
Monetary	Frequency	R_points	F_points	M_points	RFM_points
2363	12	1	4	4	144
2555	13	1	4	4	144
2218	11	4	3	4	434
1654	9	5	2	2	522
1567	8	1	1	2	112
811	5	2	1	1	211
2814	14	4	5	5	455
1847	10	4	2	2	422
1519	8	5	1	2	512
1471	8	2	1	2	212
2303	12	1	4	4	144
3634	18	1	5	5	155
918	5	2	1	1	211
1792	9	1	2	2	122

3.5 Explanation of how the random data for Recency were generated

To calculate **Recency**, we generated random dates corresponding to the last purchase of each customer within the interval [01/01/2024, 30/06/2025] (Last Purchase), using the Excel function `CASUALE.TRA(45292;45838)`, where 45292 is the numerical representation of 01/01/2024 and 45838 corresponds to 30/06/2025.

Subsequently, Recency was computed as the difference between the snapshot date (30/06/2025) and the last purchase date for each customer, thereby obtaining the number of days elapsed since the most recent transaction.

3.6 Calculation of R_Points, M_Points, and F_Points

For each variable of the RFM analysis we calculated the quantities that characterize the distribution of their values, which are:

- **minimum value**,

- **first quartile (Q1)**,

- **second quartile (Q2)**,

- **third quartile (Q3)**,

- **maximum value**,

- **average**.

Below we report the tables (also present in the Excel attachment), each relating one of the three variables, in which we have written the corresponding values found for each of these quantities:

MONETARY	
VALUE	THRESHOLD
135	min value
1453,25	Q1
2050	Q2
2593,25	Q3
3869	max value
2028,545	average

FREQUENCY	
VALUE	THRESHOLD
1	min value
8	Q1
11	Q2
13	Q3
19	max value
10,56	average

3.6.1 Score assignment for Monetary

The score assignment (from 1 to 5) for the Monetary variable follows the scheme below:

- $M\_points = 1$ if Monetary $\leq Q1$

- $M\_points = 2$ if $Q1 <$ Monetary $\leq$ average

- $M\_points = 3$ if average $<$ Monetary $\leq Q2$

- $M\_points = 4$ if $Q2 <$ Monetary $\leq Q3$

- $M\_points = 5$ if $Q3 <$ Monetary $\leq$ max value

3.6.2 Score assignment for Frequency

The score assignment (from 1 to 5) for the Frequency variable follows the scheme below:

- $F\_points = 1$ if Frequency $\leq Q1$

- $F\_points = 2$ if $Q1 <$ Frequency $\leq$ average

- $F\_points = 3$ if average $<$ Frequency $\leq Q2$

- $F\_points = 4$ if $Q2 <$ Frequency $\leq Q3$

- $F\_points = 5$ if $Q3 <$ Frequency $\leq$ max value

3.6.3 Score assignment for Recency

The score assignment (from 1 to 5) for the Recency variable follows the pattern below:

- $R\_points = 5$ if Recency $\leq Q1$

RECENCY	
VALUE	**THRESHOLD**
2	min value
147,75	Q1
267	Q2
411,25	Q3
545	max value
272,0275	average

- $R\_points = 4$ if $Q1 < \text{Recency} \leq \text{average}$

- $R\_points = 3$ if $\text{average} < \text{Recency} \leq Q2$

- $R\_points = 2$ if $Q2 < \text{Recency} \leq Q3$

- $R\_points = 1$ if $Q3 < \text{Recency} \leq \text{max value}$

Finally we created the box plots for each distribution of the variables (Monetary, Frequency, Recency) and associated the corresponding `RFM_points` with each Customer ID.
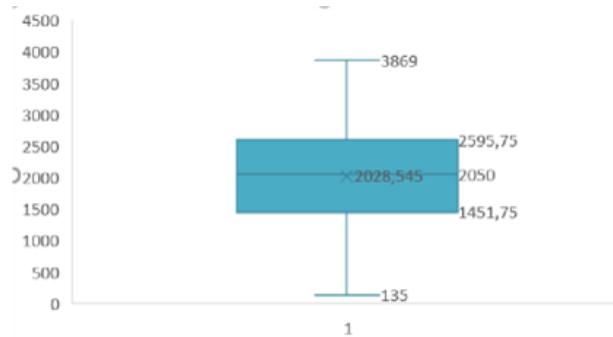
Figure 4: Box plot – Monetary (private customers).
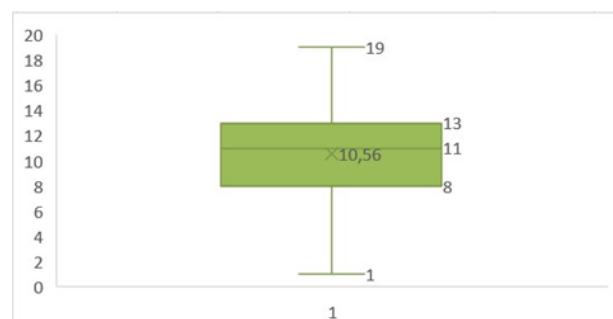
Figure 5: Box plot – Frequency (private customers).
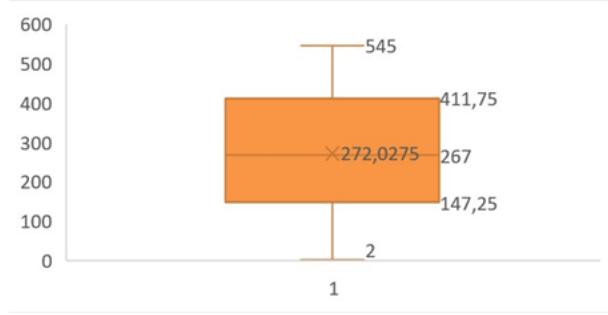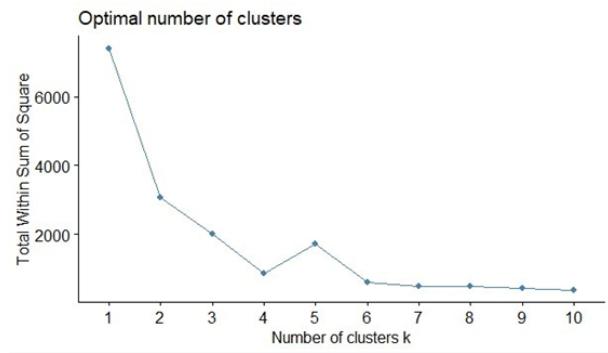
Figure 6: Box plot – Recency (private customers).

4 RFM and K-means Analysis

We decided to apply **K-means analysis** to the results obtained from the RFM analysis conducted on both individual customers and families, thus combining the two quantitative analysis techniques. K-means was applied to the *RFM_points* scores obtained from the RFM analysis carried out on private customers.

4.1 K-means on Private Customers
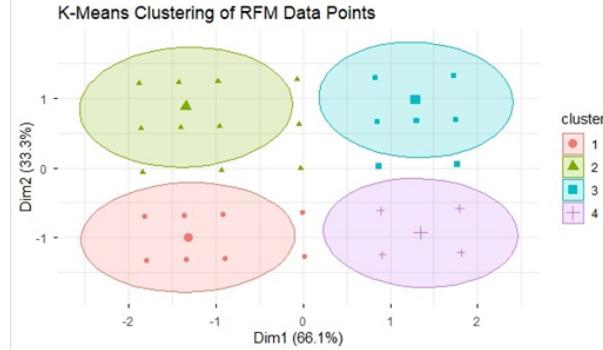
To apply K-means analysis to the *RFM_points* scores obtained from the RFM analysis on private customers, it is first necessary to perform the Elbow Method in order to identify the optimal number of clusters. As shown in the graph, the horizontal axis represents the number of clusters (k), while the vertical axis shows the Total Within Sum of Squares (TWSS), that is, the sum of the squared distances of the data points from the centroid of their respective clusters, which is indicative of within-group variability.
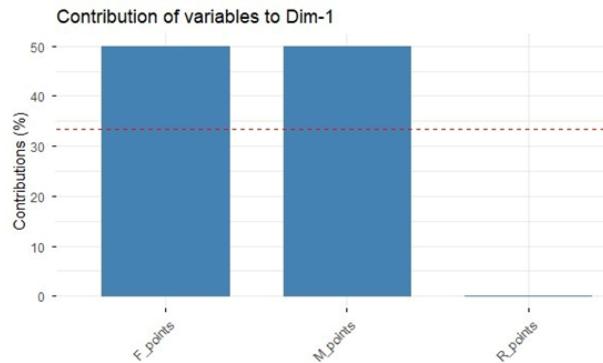
From the graph, there is a clear decrease in TWSS when moving from $k = 1$ to $k = 2$, and this reduction continues up to $k = 4$. At $k = 5$, small fluctuations appear, but they do not change the overall shape of the curve, which still shows a clearly identifiable elbow point.

In conclusion, the graph suggests that $k = 4$ represents the optimal number of clusters. This choice allows for a sufficiently informative and stable segmentation, avoiding excessive data fragmentation that would make the interpretation of the results more complex.
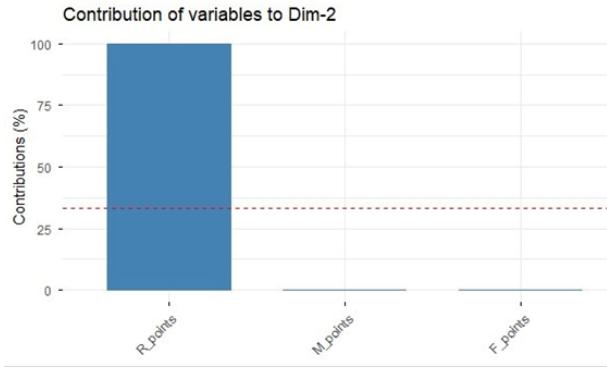
Moreover, it ensures that the clusters are not too similar to one another, while maintaining a high level of homogeneity within each cluster.
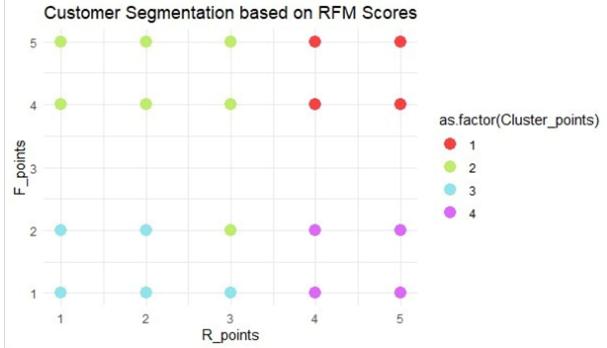

From the graph, four clearly distinct and well-separated clusters can be observed, each represented by an ellipse showing how the data points are distributed within the cluster. This clear separation indicates that the model effectively distinguishes between different customer behavior profiles based on the RFM variables.

The graph shows the percentage contribution of the variables from the RFM analysis to the first dimension of the clustering. The horizontal dashed line represents the expected average contribution. The analysis reveals that the Frequency (*F_points*) and Monetary (*M_points*) variables provide a contribution well above the average line, appearing equivalent and dominant in defining the first dimension. This outcome is attributable to the dummy variable *Reference*, introduced to ensure a correlation between Monetary and Frequency during the random data generation process. n contrast, the Recency (*R_points*) variable contributes very little to the first dimension, remaining well below the average value.

Contribution of variables to Dim-2
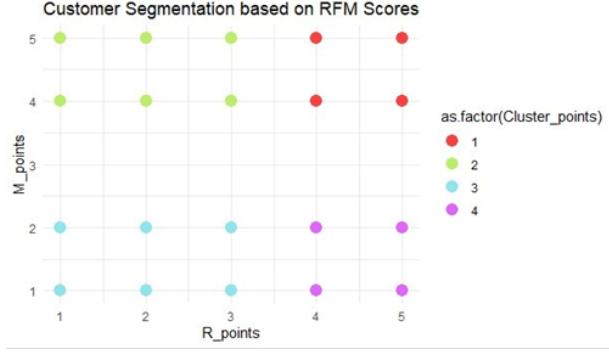
The graph representing the second dimension shows that the Recency variable (R_points) is the main contributor, positioning itself well above the average level. In contrast, the Monetary (M_points) and Frequency (F_points) variables show contributions close to zero. Indeed, the values assumed by the Recency variable were calculated based on those related to the "Last Purchase" variable, which were generated without any correlation with Monetary and Frequency. Therefore, it is reasonable that the second dimension mainly depends on Recency, since it is independent from the other two variables.


Customer Segmentation based on RFM Scores
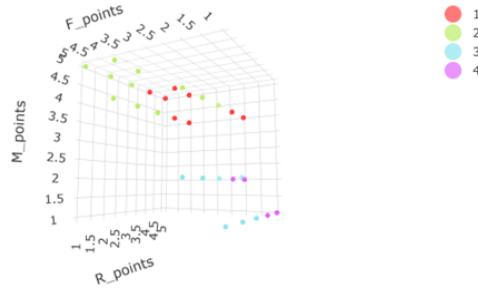
The graph represents the segmentation of customers based on their RFM scores, with Recency (R_points) on the horizontal axis and Frequency (F_points) on the vertical axis. The points are colored according to their cluster membership, as determined by the K-means clustering algorithm. The analysis reveals a clear separation of clusters along both dimensions considered. The clusters located in the upper part of the graph, characterized by high Frequency values, correspond to regular customers. Conversely, the clusters positioned in the lower area display lower Frequency levels, which can be associated with more sporadic consumption behaviors. Similarly, the distribution along the Recency axis makes it possible to distinguish more recent customers, with high R_points values, from less active ones.

The graph represents customer segmentation based on RFM scores, with Recency (R_points) on the horizontal axis and Monetary (M_points) on the vertical axis. Here as well, a clear separation of clusters is observed along both dimensions considered. The clusters located in the upper part of the graph represent customers with high monetary

24

Customer Segmentation based on RFM Scores

value, whereas those in the lower part identify customers with lower spending levels. The distribution along the Recency axis also makes it possible to distinguish more recent and active customers from less recent ones.
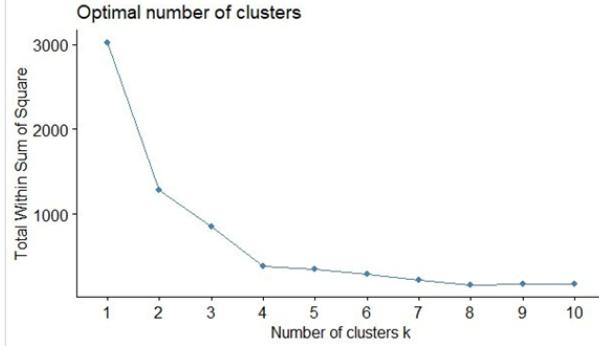

This is the three-dimensional graph representing customer segmentation based on the RFM analysis. We also attach a dynamic representation of this graph, available at the following link. `https://aleimola.github.io/ddm_graph.html`

Considering the three dimensions jointly, we can distinguish the following:

- **Red cluster** represents customers with high recency, high frequency, and high monetary value. These are "champion" customers: loyal, highly engaged, and characterized by high economic value. To enhance their value, Neovision adopts strategies based on exclusivity, experience personalization, cultural engagement, and the creation of an elite community. Specifically, we offer early and exclusive access to the most popular events and prestigious locations, meetings with directors, actors, and critics, exclusive discussions after the events, participation in choosing films and formats through voting, personalized extra content via smart glasses (such as backstage scenes, expert comments, and archive material), as well as personalized and limited-edition gadgets.

- **Purple cluster** represents customers with high recency, low frequency, and low monetary value. These are new customers who have recently shown interest in our service but do not yet display regular purchasing behavior or high economic value. The objective is to transform them into habitual customers by guiding them toward

25
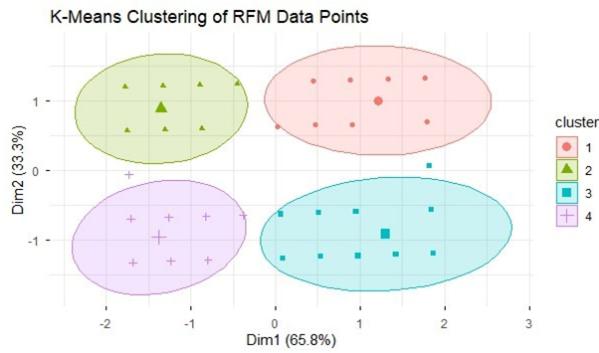
a stable relationship with the brand. An initial operational strategy consists of incentivizing the second purchase by facilitating customer return through dedicated temporary offers, vouchers usable for subsequent events, and by encouraging the sharing of the experience with friends or family. To further encourage customers to return, Neovision plans to personalize the offer by inviting them to events that match their interests, such as films of the same genre or by the same director, or set in historical contexts similar to those they have already experienced. In addition, through post-event communication, storytelling, and extra content, the company can increase the perceived value of the experience and highlight the uniqueness of the Neovision service.

- **Blue cluster** represents customers with low recency, low frequency, and low monetary value. These are "hibernating" customers who have not made purchases for a long time, buy infrequently, and spend little, and are therefore currently inactive and weakly engaged. They can be reactivated through targeted strategic actions, such as invitations to particularly attractive events—cinematic anniversaries or tributes to renowned actors and directors—focusing on the cultural and emotional value of the experience rather than on price. Furthermore, for hibernating customers it is important to communicate how the service works, highlighting technological improvements, new locations, or renewed formats, in order to convey the idea of an evolved experience compared to the past.

- **Green cluster** represents customers with low recency, high frequency, and high monetary value. These are "at-risk" customers who have not purchased for some time but who previously purchased frequently and spent significantly. As they are valuable customers, timely intervention is necessary to stimulate engagement and prevent imminent churn. Neovision should recognize the importance of these customers through personalized messages, referring to their past participation and their contribution to the growth of the experience. The company can invite at-risk customers to selected events, previews, or special evenings in prestigious locations, highlighting the unique and exclusive nature of the Neovision experience. We also consider it important to offer high-value personalized content, such as extra materials, cultural insights, or advanced narrative paths that recall the interests previously demonstrated by the customer. The goal is not just to bring the customer back once, but to rebuild a long-term relationship and restore a level of engagement that reflects their high past value.

Optimal number of clusters
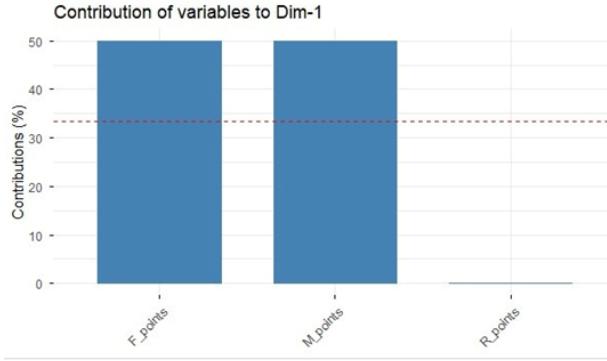
4.2 K-means on Families

To apply K-means analysis to the RFM_points scores obtained from the RFM analysis on families, it is first necessary to perform the Elbow Method in order to identify the optimal number of clusters. As shown in the graph, the horizontal axis represents the number of clusters (k), while the vertical axis shows the Total Within Sum of Squares (TWSS), that is, the sum of the squared distances of the data points from the centroid of their respective clusters, which indicates the variability within the groups. From the graph, a significant reduction in TWSS can be observed when moving from k = 1 to k = 2, and this reduction continues up to k = 4. In conclusion, the graph suggests that k = 4 represents the optimal number of clusters. This choice allows for a sufficiently informative and stable segmentation, avoiding excessive data fragmentation that would make the interpretation of the results more complex. Moreover, it ensures that the clusters are not too similar to one another or positioned too close together, while maintaining a high level of homogeneity within each cluster.


K-Means Clustering of RFM Data Points
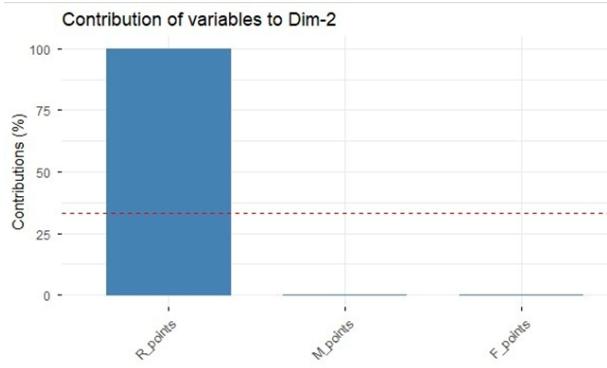
From the graph, four distinct and well-separated clusters emerge, each identified by an ellipse representing its internal dispersion. The clear separation suggests that the model is able to effectively distinguish between different customer behavior profiles based on the RFM variables.

The graph shows the percentage contribution of the RFM analysis variables to the first dimension of the clustering. The horizontal dashed line represents the expected average contribution. The analysis shows that the Frequency (*F_points*) and Monetary (*M_points*)
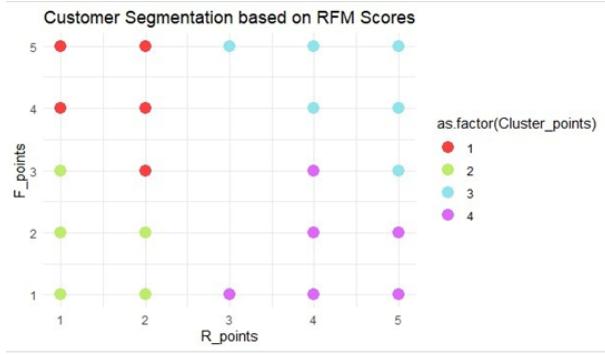
Contribution of variables to Dim-1

variables provide a contribution well above the average line, appearing equivalent and dominant in defining the first dimension. This result is attributable to the dummy variable *Reference*, which was introduced to ensure a correlation between Monetary and Frequency values during the random data generation process for the RFM analysis. In contrast, the Recency (*R_points*) variable shows a negligible contribution to the first dimension, remaining well below the average value.


Contribution of variables to Dim-2
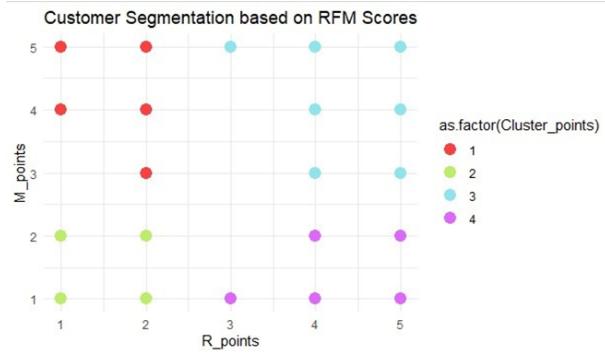
The graph represents the percentage contribution of the RFM variables to the second dimension. The horizontal dashed line indicates the expected average contribution. From the graph, it clearly emerges that the Recency variable (*R_points*) provides a largely dominant contribution to the second dimension, positioning itself well above the average line. In contrast, the Monetary (*M_points*) and Frequency (*F_points*) variables show contributions close to zero. In fact, the values of the Recency variable were calculated based on those of the "Last Purchase" variable, which were generated without any correlation with Monetary and Frequency. Therefore, it is reasonable that the second dimension depends mainly on Recency, given its independence from the other two variables.
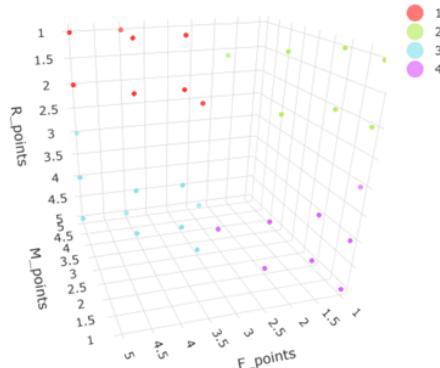
The graph shows customer segmentation based on RFM scores, with Recency (*R_points*) on the horizontal axis and Frequency (*F_points*) on the vertical axis. The points are colored according to the cluster identified by the K-means algorithm. The analysis shows a clear separation of clusters along both dimensions. Clusters in the upper part of the graph, with high Frequency values, represent regular customers. In contrast, clusters in the lower part show lower Frequency levels and reflect more occasional purchasing behavior.

Customer Segmentation based on RFM Scores

Similarly, the distribution along the Recency axis makes it possible to distinguish more recent customers, with high R_points values, from less active ones.

Customer Segmentation based on RFM Scores

A further graph represents customer segmentation with Recency (*R_points*) on the horizontal axis and Monetary (*M_points*) on the vertical axis. Here as well, clusters are clearly separated. Customers in the upper area of the graph show higher monetary value, whereas those in the lower area are characterized by lower spending levels. Recency again makes it possible to distinguish more recent and active customers from less recent ones.

This is the three-dimensional graph representing customer segmentation based on the RFM analysis. We also attach a dynamic representation of this graph, available at the

following link. `https://aleimola.github.io/ddm_graph.html` Considering the three dimensions jointly, we can distinguish the following:

- **Blue cluster** represents families with high recency, high frequency, and high monetary value. Champion families are characterized by frequent participation, significant spending, and recent interaction with the Neovision service. For this segment, the goal is to strengthen the long-term relationship. To do this, Neovision can use strategies to encourage regular family participation, such as offering premium family subscriptions that include several events for children and exclusive benefits for adults. To further strengthen brand loyalty, we also plan to create age-based experiences that support children as they grow.

- **Purple cluster** represents families with high recency, low frequency, and low monetary value. These are recent families who have only recently tried the Neovision service and have not yet developed regular participation. In this case, the marketing strategy should aim to encourage repeat attendance after the first experience, for example by inviting families to screenings of genres similar to those recently viewed or to participate in a themed film cycle. Communication also plays a key role in presenting the Neovision service as a shared educational and cultural experience for both parents and children.

- **Green cluster** represents families with low recency, low frequency, and low monetary value. These hibernating families currently provide limited economic value and have not participated for a long time. The objective is to reactivate this segment without investing excessive resources. To this end, we propose low-commitment events, both economically and organizationally, emphasizing the playful and engaging nature of the experience for children. In addition, collaborations with schools may represent an effective channel to reconnect families with the service. With the aim of reintroducing Neovision, communication with this segment should adopt a simple and positive tone, focused on fun and the quality of shared family time.

- **Red cluster** represents families with low recency, high frequency, and high monetary value. These at-risk families are the most critical for the company: they have not purchased for a long time but were frequent and high-spending customers in the past. To re-establish interaction, Neovision should first investigate the reasons behind the decline in participation, which may be related to time constraints, changes in children's needs or preferences, or the birth of new children, all of which may reduce attendance at events. The company can then offer more flexible events and screenings in terms of scheduling and duration, as well as packages specifically designed for families with young children, including parallel entertainment activities alongside film screenings intended for a more adult audience. It is also essential for Neovision

to acknowledge, through personalized communications, the contribution made by these families, referring to their past participation.