Group assignment – Financial Programming

MBD 2024

Written by: Yaqing HU
Johnny CHREIM
Edouard LENOIR
Somkenechukwu MBAMALI

Introduction

This project focuses on helping a bank improve its understanding of customer behavior to enhance service offerings and mitigate risks. The primary objective is to predict whether a client will be granted a loan or issued a credit card in 1997 based on their financial activities and demographic information from 1996. By analyzing client data and associated account activities, the bank can make more informed decisions, offering additional services to reliable customers and identifying potential risks.

As a data analytics consultant, we derived several meaningful independent variables from the dataset, combining client characteristics, transactional data, and district-level demographics to achieve this goal. Key features include variables like LOR (length of relationship), gender, age_group to represent individual customer attributes. Economic and transactional indicators such as dist_unemploy_rate, dist_salary_avg, urban_inhabitant_ratio and log_avg_balance were calculated to capture financial and regional dynamics. Additionally, behavioral metrics like RFM_Score (recency, frequency, and monetary score), Credit_F_Score, and Withdrawal_F_Score were designed to quantify customer financial patterns and engagement.

The analysis begins with data cleaning and value transformation to ensure accuracy and consistency. Following this, the analysis focuses on visualizing and interpreting independent variables and the target variables—granted_loan and card_issued. This includes exploring trends, correlations, and distributions to identify significant patterns that inform customer behaviors and bank decision-making.

Through detailed exploration and visualization, this project provides actionable insights into how various factors influence the likelihood of a client being granted a loan or credit card. These findings will guide future predictive modelling and strategic decision-making for the bank.

Variable Creation

1. Creation of independent variables

- **Length of Relationship (LOR):** Derived by taking 1996 as the benchmark year for client-account duration.
- **Client Demographics:** Variables such as gender, age, and age group were extracted.
- **Recency, Frequency, and Monetary (RFM):** Metrics for evaluating client transaction behaviour, including recency (time since the last transaction), frequency (number of transactions), and monetary (total amount spent in 1996).
- **Transaction Details:** Variables like the average and standard deviation of transaction amounts (avg_amount_trans_96, trans_std_96) and account balances (avg_balance_96, std_balance_96) for 1996 were calculated.
- **Frequency Issuance Statistics (freq iss. stats):** Indicates how frequently account statements are issued to clients.
- **Average District Salary, Unemployment Rate and Urban Inhabitant Ratio:** Captures key socioeconomic conditions within the client's district.
- **Other Variables:** Includes total credit per client, total withdrawal per client, and the frequency of credit and withdrawal transactions for each client.

2. Creation of dependent variables

Two dependent variables were defined to measure client outcomes in 1997:

- **Granted Loan:** Whether a loan was granted to the client (Binary: 0/1).
- **Card Issued:** Whether a credit card was issued to the client (Binary: 0/1).

Data Correction and Value Transformation

1. Data correction

1.1 Handling missing values

client_id	0
freq iss. stats	0
LOR	0
gender	0
age_group	0
total_credit	1
total_withdrawal	0
credit_frequency	1
withdrawal_frequency	0
avg_amount_trans_96	0
transaction_std	0
avg_balance	0
balance_std	0
recency_days	0
frequency	0
monetary	0
urban_inhabitant_ratio	0
avg_dist_salary	0
dist_unemploy_rate	0

dtype: int64

Figure 1: Information about the missing value

Missing values were found in total_credit and credit_frequency. Based on their distributions (shown in Figures 2, and 3):

- Total Credit: Skewed distribution—replaced missing values with the median.
- Credit Frequency: Normal distribution—replaced missing values with the median.

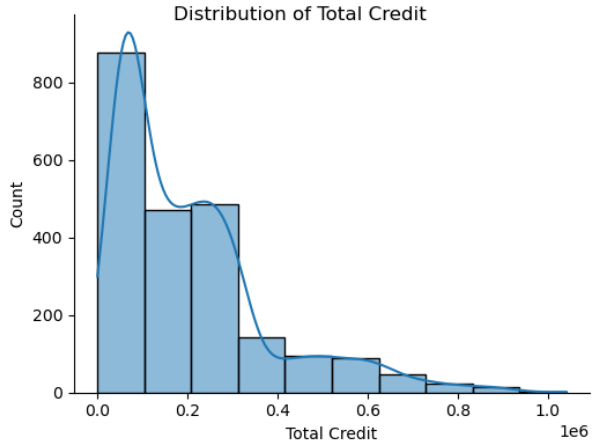

Figure 2: Distribution of Total Credit

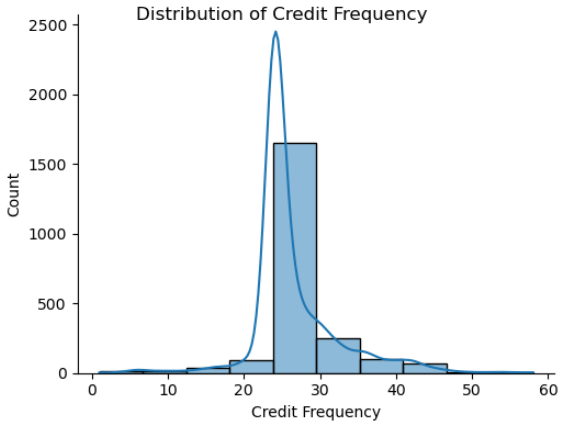

Figure 3: Distribution of Credit Frequency

1.2 Handling outliers

Outliers were identified in most numerical variables and categorized into three outlier groups based on value ranges. Credit_frequency has the highest number of outliers and is unique as it includes both small and large-value outliers, indicating significant variability in client transaction frequency. Most other

variables with outliers (e.g., total_credit, total_withdrawal, monetary) only have large-value outliers, reflecting clients with exceptionally high values. In contrast, variables like urban_inhabitant_ratio and dist_unemploy_rate are stable with no outliers, showing uniform distribution within expected ranges.

```
total_credit [ float64 ] 38 outlier(s)
total_withdrawal [ float64 ] 37 outlier(s)
credit_frequency [ float64 ] 50 outlier(s)
withdrawal_frequency [ float64 ] 12 outlier(s)
avg_amount_trans_96 [ float64 ] 25 outlier(s)
transaction_std [ float64 ] 7 outlier(s)
avg_balance [ float64 ] 4 outlier(s)
balance_std [ float64 ] 7 outlier(s)
recency_days [ int64 ] 17 outlier(s)
frequency [ int64 ] 30 outlier(s)
monetary [ float64 ] 38 outlier(s)
urban_inhabitant_ratio [no outliers]
avg_dist_salary [no outliers]
dist_unemploy_rate [no outliers]
```

Figure 4: Information about the outliers

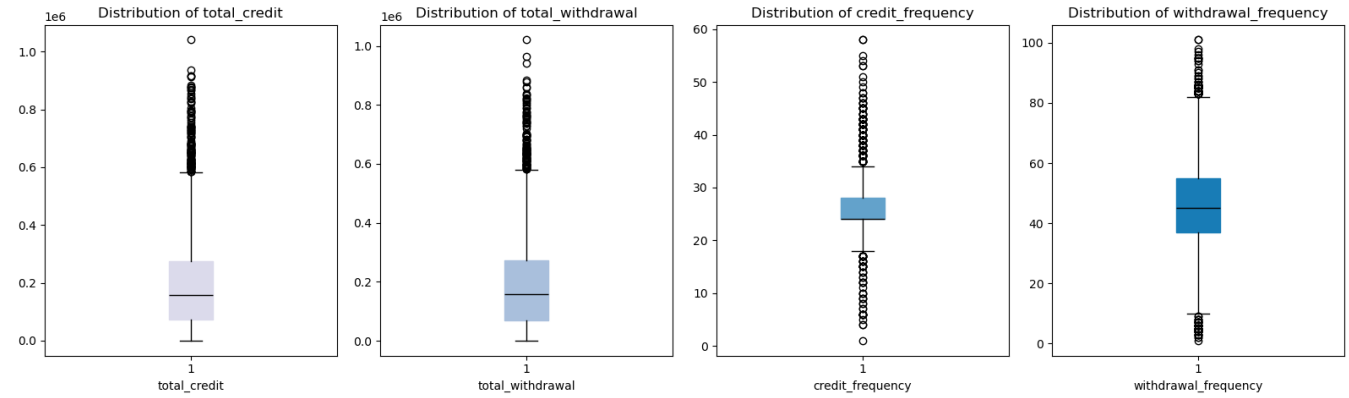

Figure 5: Boxplots of variable set 1

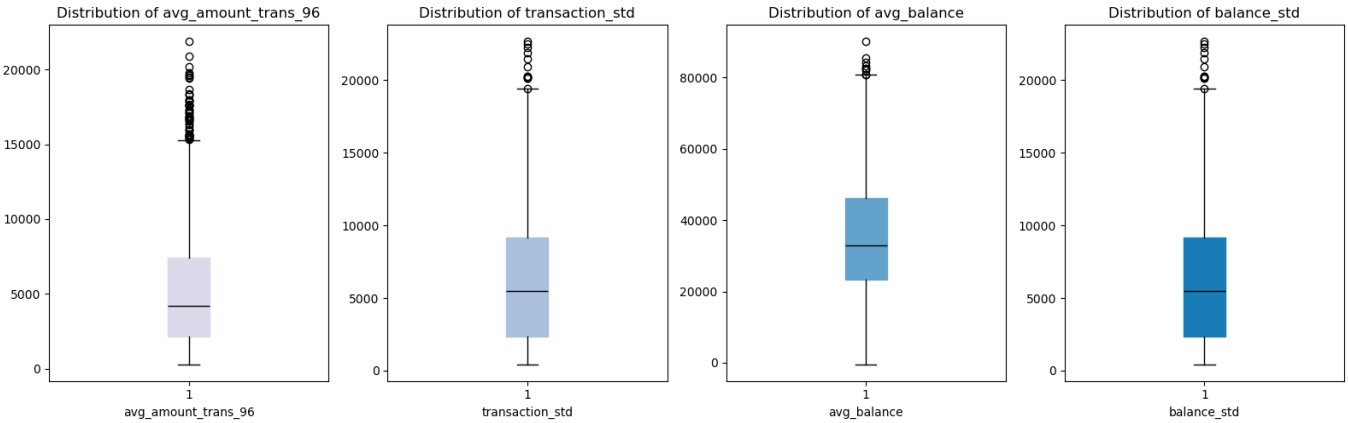

Figure 6: Boxplots of variable set 2

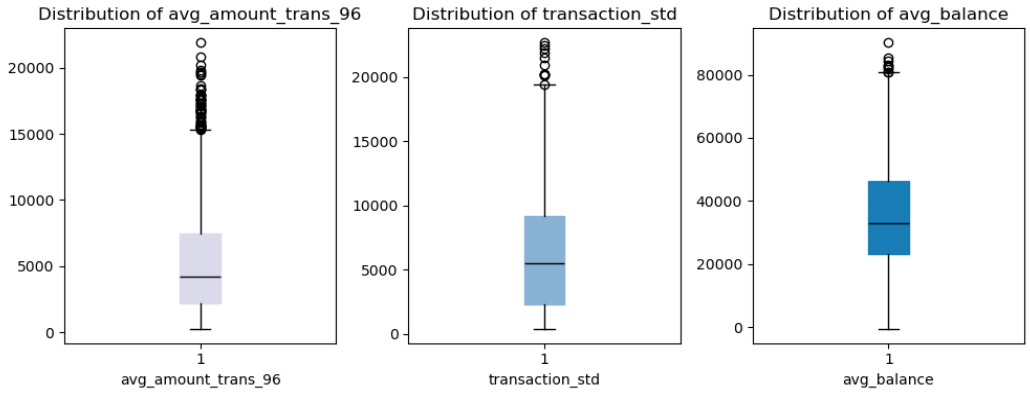

Figure 7: Boxplots of variable set 3

- **Treatment:** The outliers were treated by applying a method that replaces them with the upper and lower boundaries, calculated as +/-3 standard deviations (SD) from the mean.

2. Value transformation

2.1 Discretization

To simplify analysis and enhance interpretability, some continuous variables were divided into categorical bins, enabling better customer data segmentation. This transformation allows for easier comparison and understanding of customer profiles. For example,

- R_Score: Recency score, ranging from 1 to 4, where 4 represents longer recency (lower recency) and 1 indicates a more recent transaction.
- F_Score: Frequency score, ranging from 1 to 4, where 4 indicates higher frequency of transactions.
- M_Score: Monetary score, ranging from 1 to 4, where 4 indicates higher monetary value (larger transaction amounts).
- RFM_Score: A composite score that combines the R, F, and M scores to reflect overall customer engagement and transaction behavior.

Similarly, the variables credit_frequency and withdrawal_frequency were discretized into:

- Credit_F_Score: Scores range from 1 to 4, where 4 indicates higher credit frequency.
- Withdrawal_F_Score: Scores range from 1 to 4, where 4 indicates higher withdrawal frequency.

2.2 Dummy coding

For categorical variables, the freq_iss_encoded variable, which indicates the frequency of account statement issuance, was dummy-coded into two separate binary variables:

- stat_issued_after_trans: A variable indicating if the statement was issued after a transaction.
- stat_issued_weekly: A variable indicating if the statement was issued on a weekly basis.

• stat_issued_monthly: This variable is represented when both stat_issued_after_trans and stat_issued_weekly are 0, indicating that the statement was issued monthly.

2.3 Log transformation

Skewed variables with large ranges were log-transformed to reduce skewness and stabilize variance. Distribution changes are shown in Figures 8 and 9. As we can see from these histograms, after applying the log transformation, there is improved symmetry in the distributions. The values are now more evenly spread and appear closer to a normal distribution. This indicates that the transformation effectively mitigated the influence of extreme values.

The log transformation not only stabilized the variance but also made the data more suitable for analysis and modeling, ensuring that patterns and relationships could be more accurately identified.

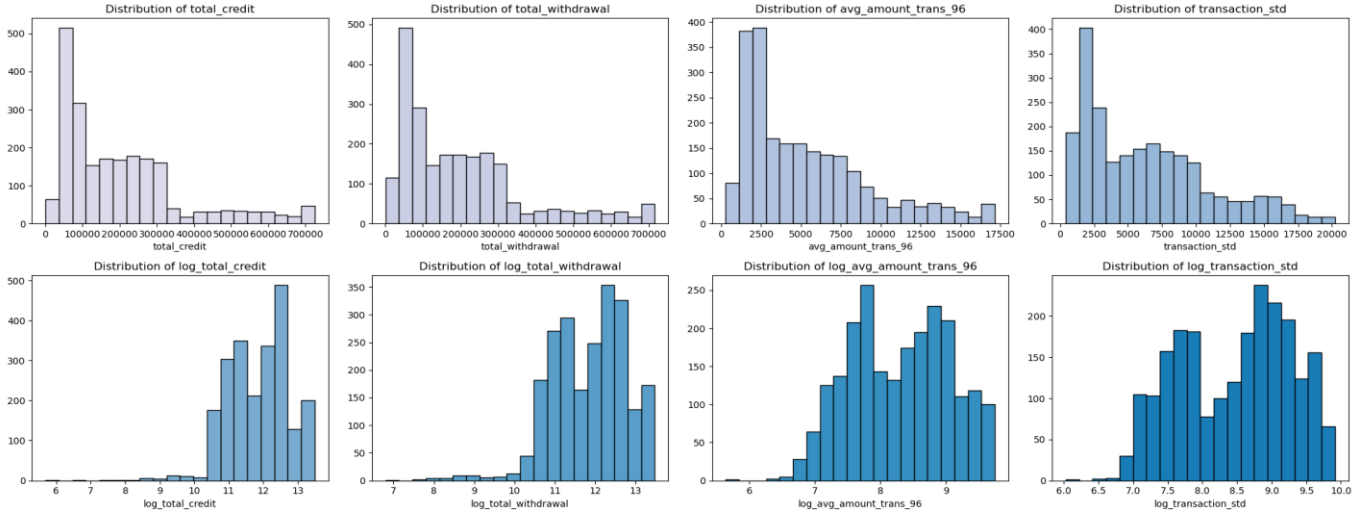

Figure 8: Histograms of log transform variable set 1

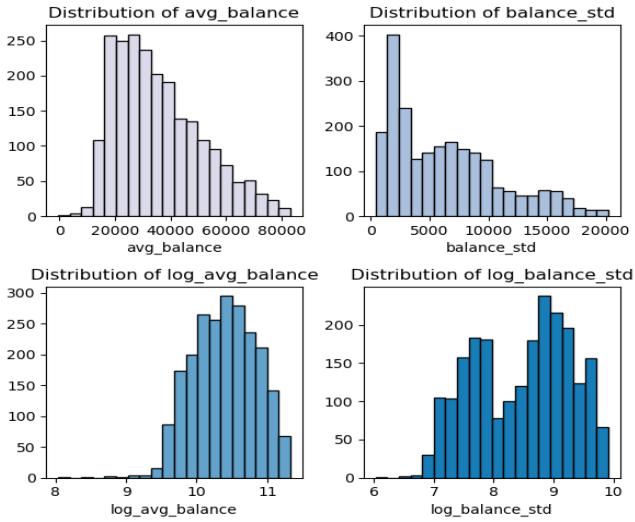

Figure 9: Histograms of log transform variable set 2

2.4 Normalization

To ensure comparability and improve the analysis of features with different scales, variables like urban_inhabitant_ratio and dist_unemploy_rate were normalized. This step brought these features to the same scale, ensuring that they contribute equally to the analysis and avoiding the dominance of larger-scale variables in modelling.

Variable Description and Visualization

1. LOR and client demographics

1.1 LOR

LOR represents the length of the relationship between the client and the bank. It is calculated by extracting the year of the account opening from the accounts table. We subtract this extracted year from the year 1996 to get the LOR value of each client. If we check the distribution of the variable (in the below figure) , we can see that the distribution is left-skewed; most of the accounts were created in 1993.

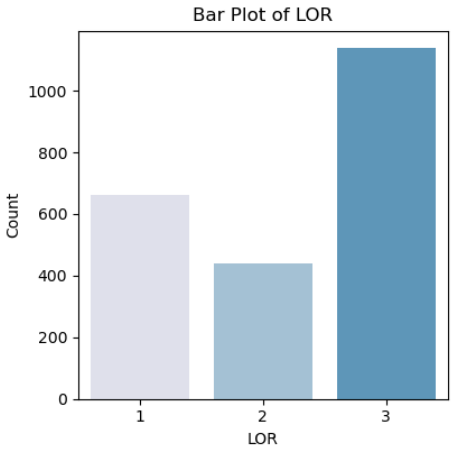

Figure 10: Barplot of LOR

1.2 Client demographics

For the client gender, the bar plot illustrates a near-equal distribution of genders, with the M category having a slightly higher count than F. This balanced distribution ensures both genders are well-represented in the dataset, allowing for equitable analysis.

When we visualize the distribution of the age groups, we can see that the majority of clients belong to the age groups of 20, 30, 40, and 50, with relatively equal representation across these categories. However, the youngest (10) and oldest (60 and 70) age groups have noticeably lower counts. This is logical as those age groups are the most active people in society. This can also show that lots of clients opened their bank accounts at a relatively old age at the bank; since we had a maximum LOR of 3 and many clients are now more than 30 years old.

```
gender description and key numbers:
count     2239
unique       2
top          M
freq      1155
Name: gender, dtype: object
```

```
age_group description and key numbers:
count    2239.000000
mean       37.405092
std        17.354316
min        10.000000
25%        20.000000
50%        40.000000
75%        50.000000
max        70.000000
Name: age_group, dtype: float64
```

Figure 11: Description of gender and age_group

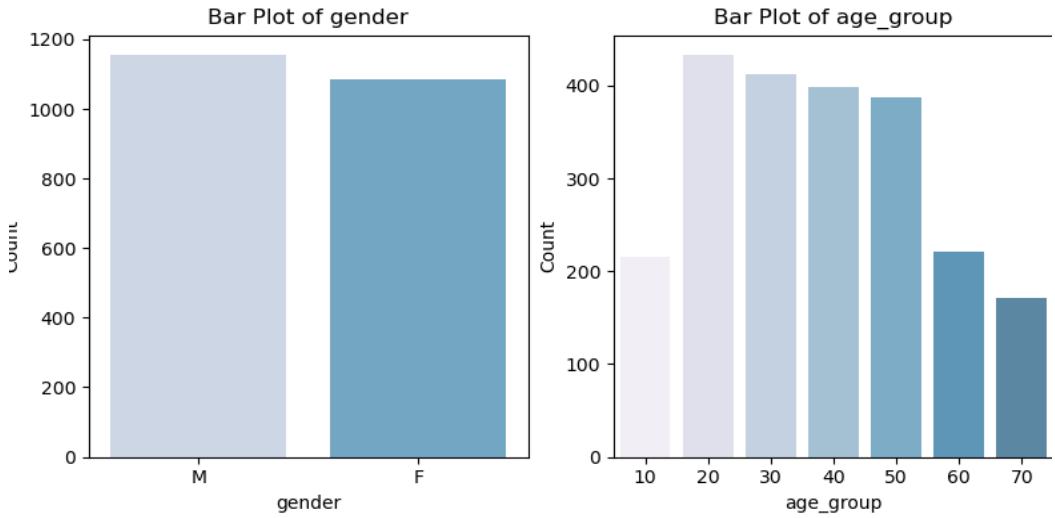

Figure 12: Barplots of gender and age_group

2. R Score, F Score, M Score and RFM Score

R_Score:

The distribution of R_Score is highly skewed, with the majority (2,218) of customers having a score of 4. Lower scores (1 and 2) are rare, with only 18 and 3 customers, respectively, and no customers having a score of 3. This indicates that most customers have a high recency score, suggesting recent interactions or purchases.

F_Score:

The F_Score distribution is concentrated around scores of 2 and 3, representing 1,160 and 823 customers, respectively. Fewer customers have scores of 4 (170) and 1 (86), indicating that while some customers purchase frequently, there are outliers with minimal or very frequent purchases.

M_Score:

The M_Score distribution shows a heavy skew toward a score of 1, with 1,218 customers. Scores of 2 (720) and 3 (147) have moderate representation, while score 4 (154) is the least frequent. This suggests that most customers have lower monetary contributions, with a small segment contributing more significantly.

RFM_Score:

A higher RFM score (closer to 12) indicates highly valuable customers who are recent, frequent buyers with significant spending. A lower RFM score (closer to 3) indicates customers who are less engaged and spend less.

The RFM_Score shows a broader distribution, with scores of 7 (777 customers) and 8 (650 customers) being the most common. Higher scores like 9 (398) and 10 (170) are less frequent, while scores above 10 (e.g., 11 and 12) and below 6 are rare. This reflects variability in customer overall RFM segmentation, with most customers falling into mid-range scores.

Insights for Modeling:

R_Score should be carefully considered due to its skewness while F_Score, M_Score, and RFM_Score offer more evenly distributed data, making them strong predictors for segmentation or predictive analytics.

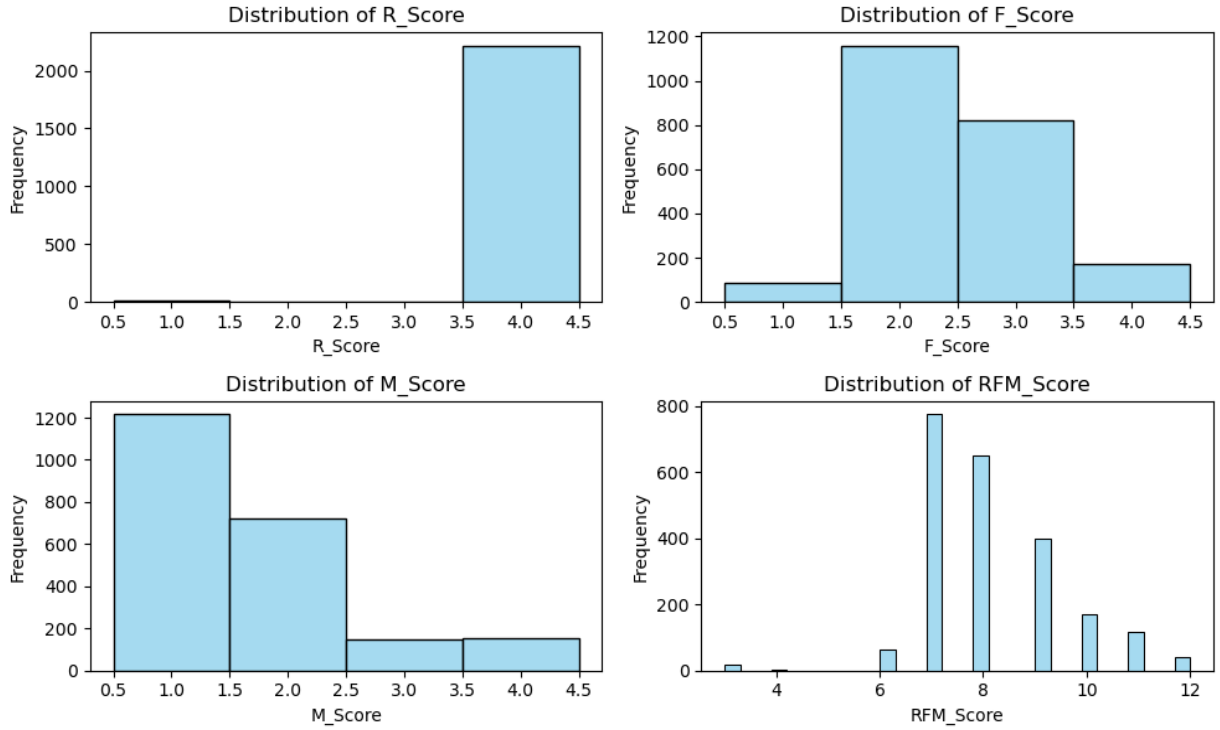

Figure 13: Distribution of R_Score, F_Score, M_Score and RFM_Score

Correlation:

As we can see from the correlation matrix, there are strong correlations between RFM_Score and F_Score (0.79), as well as M_Score (0.85), while other variables exhibit weak to moderate correlations with one another. To address potential multicollinearity and redundancy, it is important to carefully select features during model building to ensure they are not highly correlated, as this can impact the model's stability and interpretability.

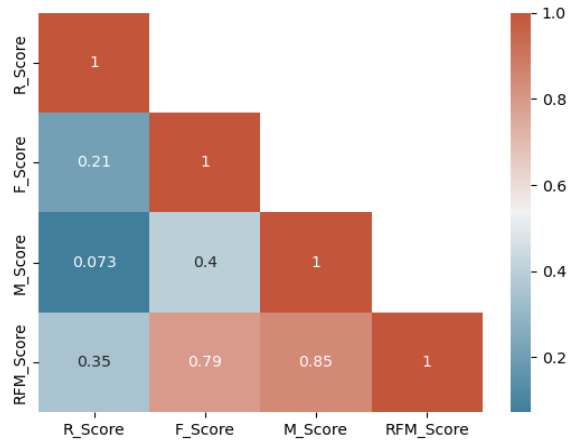

Figure 14: Correlation matrix of R_Score, F_Score, M_Score and RFM_Score

3. Other independent variables

Stat_issued_after_trans, **stat_issued_weekly**:

These two variables represent whether a statement was issued after a transaction or weekly respectively. The majority of customers (2,069) issue a monthly statement without being tied to transactions which shows client preferences for communication frequency and methods, particularly contrasting weekly vs. monthly patterns.

```
stat_issued_after_trans  stat_issued_weekly
False                    False                2069
                         True                  122
True                     False                  48
Name: count, dtype: int64
```

Figure 15: Description of Stat_issued_after_trans and stat_issued_weekly

Dist_salary_avg:

Most clients are from districts categorized as Low salary, indicating that a significant portion of the population in the dataset resides in areas with lower average income.

The Medium-salary and High salary districts are less represented, offering insights into regional economic disparities.

This variable can help in understanding client behaviour and needs based on their district's economic profile, enabling localized targeting or analysis.

```
dist_salary_avg
Low salary       1436
Medium salary     497
High salary       306
Name: count, dtype: int64
```

Figure 16: Description of dist_salary_avg

Withdrawal_F_Score, Credit_F_Score:

The high counts are observed for the pairs (2,2) with 814 occurrences and (3,3) with 545 occurrences. This indicates a strong correlation between higher credit and withdrawal scores, suggesting that customers with higher scores in one category (Credit_F_Score) tend to have similar scores in the other category (Withdrawal_F_Score).

The lower and upper extreme values for both scores (e.g., Credit_F_Score = 1 and Withdrawal_F_Score = 4, or Credit_F_Score = 4 and Withdrawal_F_Score = 1) show significantly fewer occurrences, suggesting that customers with either very low or very high scores in both categories are rare.

Withdrawal_F_Score	1	2	3	4
Credit_F_Score				
1	48	6	3	0
2	22	814	545	131
3	5	231	188	37
4	1	90	97	21

Figure 17: Withdrawal_F_Score with different Credit_F_Score

4. Dependent variables

4.1 Basic statistics and distribution

The granted_loan variable indicates whether a customer was granted a loan. Most customers did not receive a loan, with a total of 2208 occurrences (2093 with no card and 115 with a card). The card_issued variable indicates whether a customer was issued a card. Most customers did not receive a card, with a total of 2119 occurrences (2093 with no loan and 26 with a loan). A smaller number of customers received a card, with only 120 occurrences (115 without a loan and 5 with a loan), and an even smaller subset of customers received both a loan and a card.

card_issued	0	1
granted_loan		
0	2093	115
1	26	5

Figure 18: Granted_loan with different card_issued

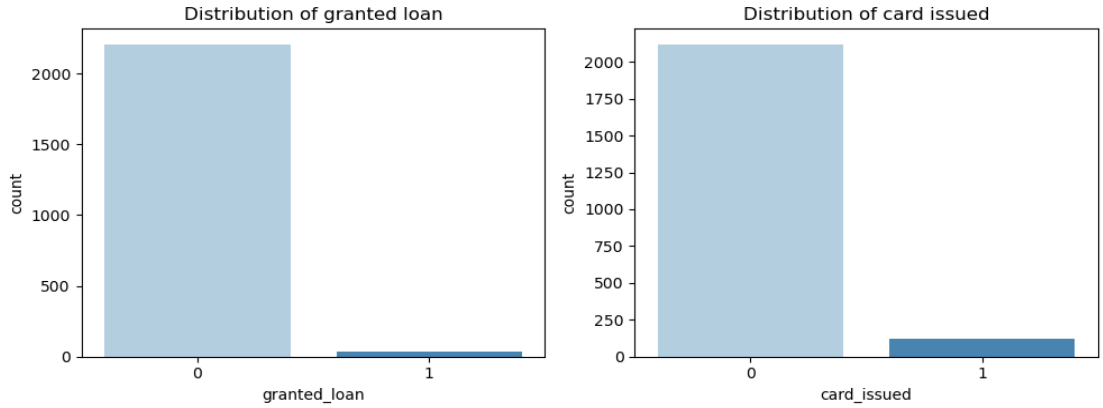

Figure 19: Distribution of granted_loan and card_issued

4.2 Relationship between independent variables and target variables

Most of these orange dots, representing customers who were granted a loan or issued a card, are clustered in the middle of the scatter plots. This suggests that a higher proportion of customers who received a loan or card tend to have a medium urban inhabitant ratio and medium unemployment rate, as their values fall around the middle of their respective ranges.

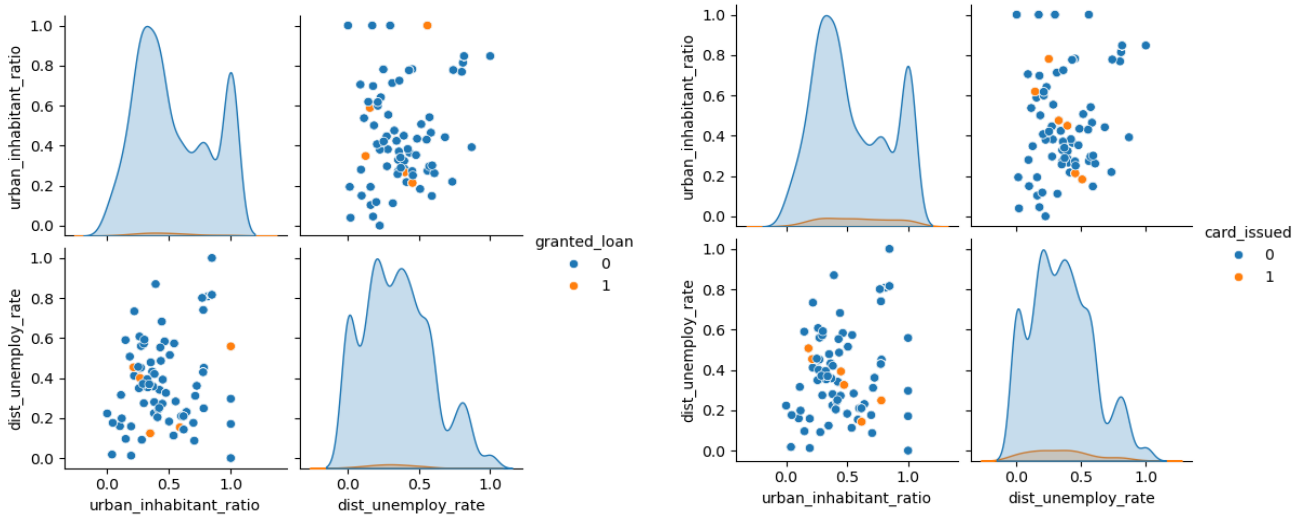

Figure 20: Distribution of R_Score, F_Score, M_Score and RFM_Score

Due to the high dimensionality of the dataset, using all variables can be inefficient and may introduce redundancy. Some variables may convey overlapping or redundant information, leading to potential multicollinearity issues. Selecting an appropriate subset of variables is crucial to enhance model accuracy and interpretability.

By analyzing the correlations among variables, we can identify whether independent variables included in the same regression model are sufficiently independent of each other. For instance, RFM_Score, F_Score,

and M_Score are highly correlated, suggesting redundancy. Likewise, log_total_credit and log_total_withdrawal exhibit strong positive correlations, emphasizing the importance of removing certain variables to avoid inefficiencies.

This step, along with other feature selection methods, is a vital part of the model-building process to ensure efficiency and accuracy in predicting the dependent variables.

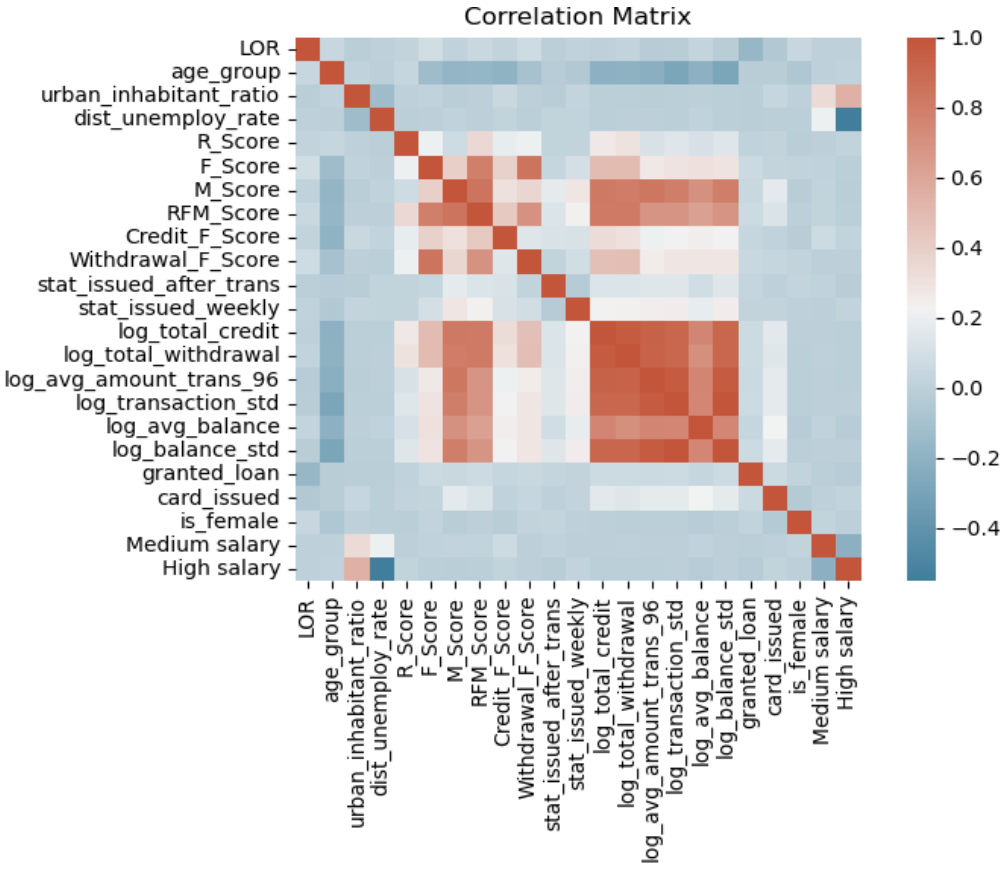

Figure 21: Correlation matrix of all the variables