CAPSTONE PROJECT

PROJECT NOTES 1 REPORT

BANKING PROBABILITY OF DEFAULT

By: PRIYA DUTTA

PGP DATA SCIENCE AND BUSINESS ANALYTICS

PGPDSBA.O.SEP23.C

CONTENTS:

1) **Introduction of the business problem ……………………………………... 2**

a) Problem Statement **2**

b) Need of the Project **3**

c) Understanding business/social opportunity **4**

2) **Data Report ………………………………………………………………... 5**

a) Understanding how data was collected in terms of time, frequency and methodology **5**

b) Visual inspection of data (rows, columns, descriptive details) **5**

c) Understanding of attributes (variable info, renaming if required) **8**

3) **Exploratory Data Analysis…………………………………………………9**

a) Univariate Analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones) **9**

b) Bivariate Analysis (relationship between different variables, correlations) **18**

c) Multivariate Analysis **20**

d) Removal of unwanted variables (if applicable) **22**

e) Missing Value treatment (if applicable) **23**

f) Outlier treatment (if required) **26**

g) Variable transformation (if applicable) **28**

4) **Business insights from EDA ………………………………………….. 29**

a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business **29**

b) Any business insights using clustering (if applicable) **29**

c) Any other business insights **30**

BANKING PROBABILITY OF DEFAULT

1] INTRODUCTION OF THE BUSINESS PROBLEM

A. PROBLEM STATEMENT

This business problem is a supervised learning example for a credit card company. The objective is to predict the probability of default (whether the customer will pay the credit card bill or not) based on the variables provided. There are multiple variables on the credit card account, purchase and delinquency information which can be used in the modelling. PD modelling problems are meant for understanding the riskiness of the customers and how much credit is at stake in case the customer defaults. This is an extremely critical part in any organization that lends money [both secured and unsecured loans].

OBJECTIVE

The primary objective is to develop a predictive model that estimates the probability of a credit card customer defaulting on their payments. This model will help the credit card company assess the riskiness of its customers and manage the potential credit risk exposure. The outcome will be used to guide credit approval processes, determine credit limits, and manage customer portfolios effectively.

B. NEED OF THE STUDY / PROJECT

The primary need for this study is driven by the significant financial risk that credit card companies face due to customer defaults. When customers fail to repay their credit card bills, the company not only loses the owed amount but also incurs additional costs related to debt recovery. A proactive approach to predicting and managing these risks can:

- **Reduce Financial Losses:** By accurately predicting which customers are likely to default, the company can take preventive measures, such as adjusting credit limits, offering tailored payment plans, or increasing debt collection efforts.
- **Improve Credit Risk Management:** Understanding the factors that contribute to default allows the company to refine its credit approval processes and optimize its risk assessment strategies.
- **Enhance Customer Relationship Management:** Identifying high-risk customers early on can enable the company to engage with them proactively, offering support or interventions that may prevent default, thus preserving the customer relationship.
- **Regulatory Compliance:** Credit card companies are subject to regulatory requirements that mandate prudent risk management practices. Developing a robust predictive model aligns with these requirements and helps the company avoid regulatory penalties.

C. UNDERSTANDING BUSINESS / SOCIAL OPPORTUNITY

Business Opportunity:

- **Market Competitiveness:** In a competitive financial market, companies that effectively manage credit risk can offer more attractive credit products with lower interest rates, thus attracting more customers while maintaining profitability.
- **Customer Segmentation and Personalization:** By using predictive models, the company can better segment its customer base and offer personalized financial products and services. For example, low-risk customers might receive higher credit limits or promotional offers, while high-risk customers could be offered financial counselling or secured credit cards.
- **Profitability Optimization:** By minimizing defaults, the company can maximize its profits. Additionally, by understanding the risk profiles of its customers, the company can better allocate resources to customer acquisition, retention, and risk management strategies.

Social Opportunity:

- **Financial Inclusion:** The model can help extend credit to underserved populations by accurately assessing their risk levels. For individuals with thin credit files or unconventional income sources, a well-calibrated model might offer the opportunity to access credit, fostering greater financial inclusion.
- **Promoting Responsible Borrowing:** By identifying customers at risk of default, the company can offer educational resources or interventions that promote responsible borrowing and financial literacy, ultimately contributing to improved financial health for individuals.
- **Economic Stability:** On a larger scale, reducing defaults contributes to the overall stability of the financial system, which is beneficial for both the economy and society. When credit card companies manage risk effectively, it reduces the likelihood of systemic financial crises.

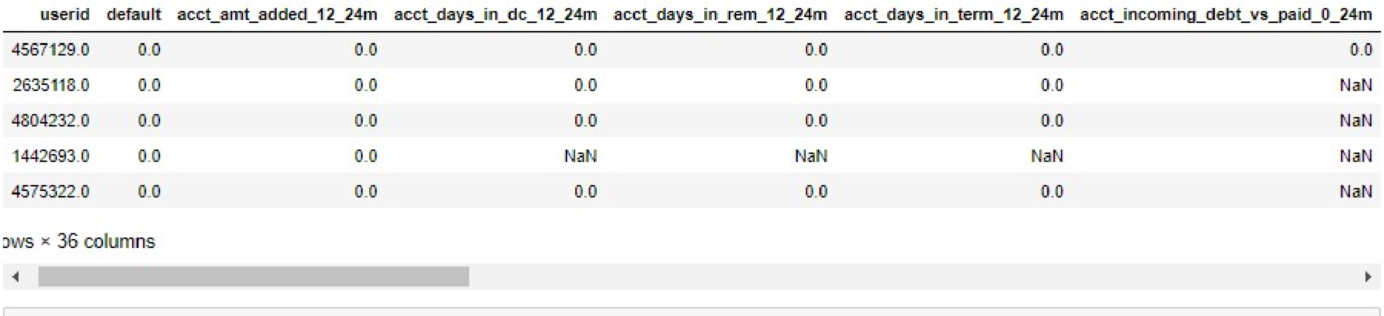
2] DATA REPORT

A. UNDERSTANDING HOW DATA WAS COLLECTED

- **Time Frame**: The dataset was collected over a period of time, though the exact duration isn't specified in the dataset itself. The columns such as acct_days_in_dc_12_24m and acct_worst_status_12_24m suggest that data spans multiple months, likely covering both short-term (3-6 months) and long-term (12-24 months) periods.
- **Frequency**: The frequency of data collection likely varies across different columns. For instance:
  - **Transaction Data**: Columns like acct_amt_added_12_24m and sum_paid_inv_0_12m suggest aggregation over 12-24 months.
  - **Account Status**: Status-related columns (e.g., acct_status, acct_worst_status_*) might be updated on a monthly or bi-monthly basis.
- **Methodology**: The data was collected from financial records, most likely from credit card transactions and account status reports. The methodology would involve regular tracking of customer transactions, payment behaviours, and account statuses across various time frames.

B. Visual Inspection of Data

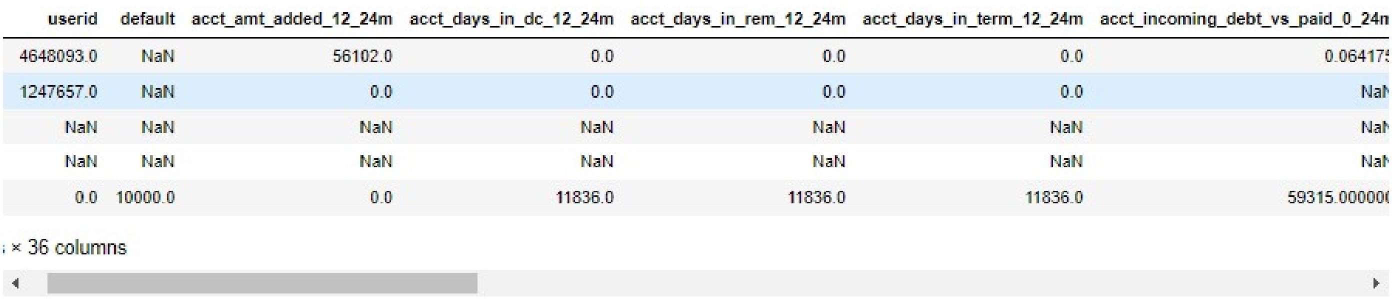
- **Rows and Columns**:
  - The dataset has 99,979 rows and 35 columns.
  - The columns include both numerical (float64) and categorical (object) data types.



userid	default	acct_amt_added_12_24m	acct_days_in_dc_12_24m	acct_days_in_rem_12_24m	acct_days_in_term_12_24m	acct_incoming_debt_vs_paid_0_24m
4567129.0	0.0	0.0	0.0	0.0	0.0	0.0
2635118.0	0.0	0.0	0.0	0.0	0.0	NaN
4804232.0	0.0	0.0	0.0	0.0	0.0	NaN
1442693.0	0.0	0.0	NaN	NaN	NaN	NaN
4575322.0	0.0	0.0	0.0	0.0	0.0	NaN

ows × 36 columns

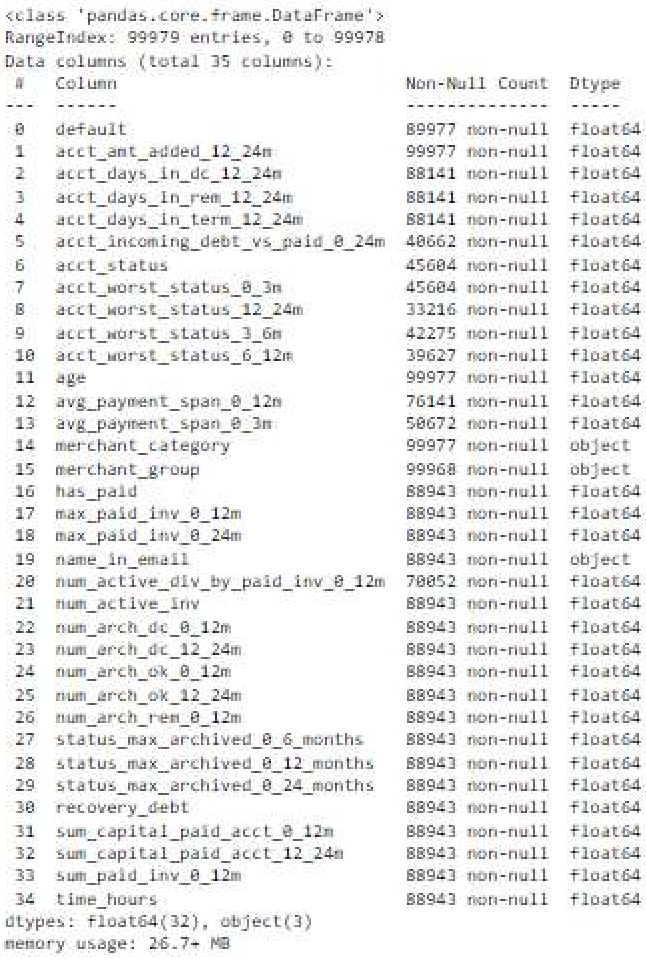
Table 1: Top 5 rows of the dataset



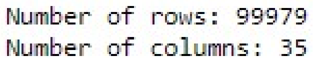
userid	default	acct_amt_added_12_24m	acct_days_in_dc_12_24m	acct_days_in_rem_12_24m	acct_days_in_term_12_24m	acct_incoming_debt_vs_paid_0_24n
4648093.0	NaN	56102.0	0.0	0.0	0.0	0.06417
1247657.0	NaN	0.0	0.0	0.0	0.0	Na
NaN	NaN	NaN	NaN	NaN	NaN	Na
NaN	NaN	NaN	NaN	NaN	NaN	Na
0.0	10000.0	0.0	11836.0	11836.0	11836.0	59315.00000

× 36 columns

Table 2: Last 5 rows of the dataset



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99979 entries, 0 to 99978
Data columns (total 35 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   default                            89977 non-null  float64
 1   acct_amt_added_12_24m              99977 non-null  float64
 2   acct_days_in_dc_12_24m             88141 non-null  float64
 3   acct_days_in_rem_12_24m            88141 non-null  float64
 4   acct_days_in_term_12_24m           88141 non-null  float64
 5   acct_incoming_debt_vs_paid_0_24m   40662 non-null  float64
 6   acct_status                        45604 non-null  float64
 7   acct_worst_status_0_3m             45604 non-null  float64
 8   acct_worst_status_12_24m           33216 non-null  float64
 9   acct_worst_status_3_6m             42275 non-null  float64
 10  acct_worst_status_6_12m            39627 non-null  float64
 11  age                                99977 non-null  float64
 12  avg_payment_span_0_12m             76141 non-null  float64
 13  avg_payment_span_0_3m              50672 non-null  float64
 14  merchant_category                  99977 non-null  object
 15  merchant_group                     99968 non-null  object
 16  has_paid                           88943 non-null  float64
 17  max_paid_inv_0_12m                 88943 non-null  float64
 18  max_paid_inv_0_24m                 88943 non-null  float64
 19  name_in_email                      88943 non-null  object
 20  num_active_div_by_paid_inv_0_12m   70052 non-null  float64
 21  num_active_inv                     88943 non-null  float64
 22  num_arch_dc_0_12m                  88943 non-null  float64
 23  num_arch_dc_12_24m                 88943 non-null  float64
 24  num_arch_ok_0_12m                  88943 non-null  float64
 25  num_arch_ok_12_24m                 88943 non-null  float64
 26  num_arch_rem_0_12m                 88943 non-null  float64
 27  status_max_archived_0_6_months     88943 non-null  float64
 28  status_max_archived_0_12_months    88943 non-null  float64
 29  status_max_archived_0_24_months    88943 non-null  float64
 30  recovery_debt                      88943 non-null  float64
 31  sum_capital_paid_acct_0_12m        88943 non-null  float64
 32  sum_capital_paid_acct_12_24m       88943 non-null  float64
 33  sum_paid_inv_0_12m                 88943 non-null  float64
 34  time_hours                         88943 non-null  float64
dtypes: float64(32), object(3)
memory usage: 26.7+ MB
```

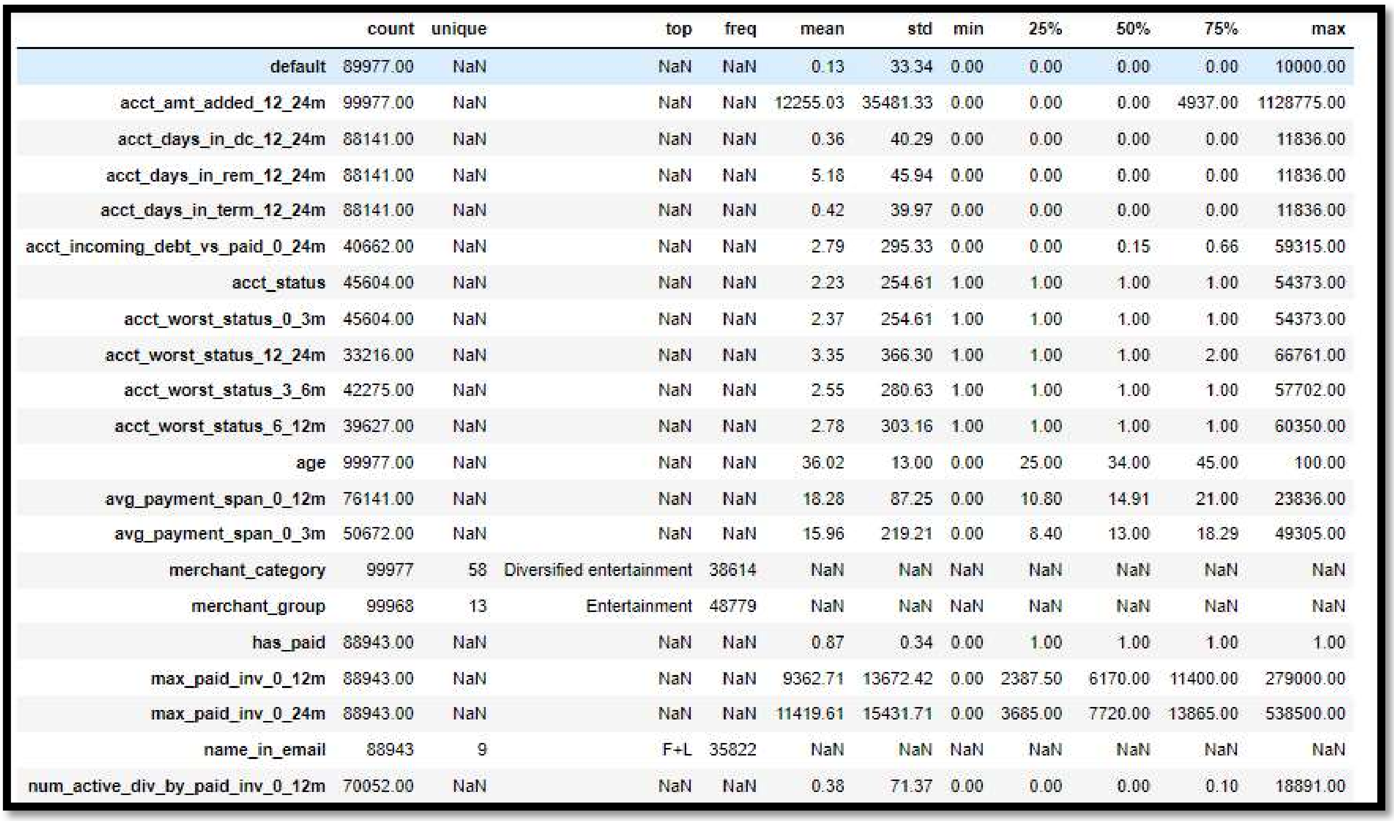


```
Number of rows: 99979
Number of columns: 35
```

Shape of the Data

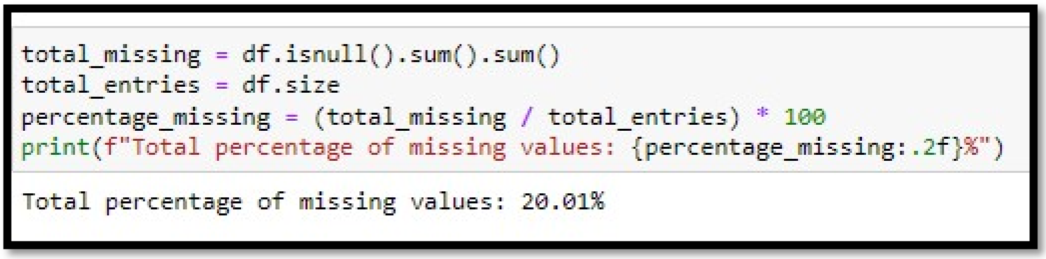
Table 3: Basic information of the dataset

- **Descriptive Details**:
  - **Numerical Data**: Includes transaction amounts, account statuses, number of days in various statuses, and other financial metrics.
  - **Categorical Data**: Includes merchant_category, merchant_group, and name_in_email.



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
default	89977.00	NaN	NaN	NaN	0.13	33.34	0.00	0.00	0.00	0.00	10000.00
acct_amt_added_12_24m	99977.00	NaN	NaN	NaN	12255.03	35481.33	0.00	0.00	0.00	4937.00	1128775.00
acct_days_in_dc_12_24m	88141.00	NaN	NaN	NaN	0.36	40.29	0.00	0.00	0.00	0.00	11836.00
acct_days_in_rem_12_24m	88141.00	NaN	NaN	NaN	5.18	45.94	0.00	0.00	0.00	0.00	11836.00
acct_days_in_term_12_24m	88141.00	NaN	NaN	NaN	0.42	39.97	0.00	0.00	0.00	0.00	11836.00
acct_incoming_debt_vs_paid_0_24m	40662.00	NaN	NaN	NaN	2.79	295.33	0.00	0.00	0.15	0.66	59315.00
acct_status	45604.00	NaN	NaN	NaN	2.23	254.61	1.00	1.00	1.00	1.00	54373.00
acct_worst_status_0_3m	45604.00	NaN	NaN	NaN	2.37	254.61	1.00	1.00	1.00	1.00	54373.00
acct_worst_status_12_24m	33216.00	NaN	NaN	NaN	3.35	366.30	1.00	1.00	1.00	2.00	66761.00
acct_worst_status_3_6m	42275.00	NaN	NaN	NaN	2.55	280.63	1.00	1.00	1.00	1.00	57702.00
acct_worst_status_6_12m	39627.00	NaN	NaN	NaN	2.78	303.16	1.00	1.00	1.00	1.00	60350.00
age	99977.00	NaN	NaN	NaN	36.02	13.00	0.00	25.00	34.00	45.00	100.00
avg_payment_span_0_12m	76141.00	NaN	NaN	NaN	18.28	87.25	0.00	10.80	14.91	21.00	23836.00
avg_payment_span_0_3m	50672.00	NaN	NaN	NaN	15.96	219.21	0.00	8.40	13.00	18.29	49305.00
merchant_category	99977	58	Diversified entertainment	38614	NaN	NaN	NaN	NaN	NaN	NaN	NaN
merchant_group	99968	13	Entertainment	48779	NaN	NaN	NaN	NaN	NaN	NaN	NaN
has_paid	88943.00	NaN	NaN	NaN	0.87	0.34	0.00	1.00	1.00	1.00	1.00
max_paid_inv_0_12m	88943.00	NaN	NaN	NaN	9362.71	13672.42	0.00	2387.50	6170.00	11400.00	279000.00
max_paid_inv_0_24m	88943.00	NaN	NaN	NaN	11419.61	15431.71	0.00	3685.00	7720.00	13865.00	538500.00
name_in_email	88943	9	F+L	35822	NaN	NaN	NaN	NaN	NaN	NaN	NaN
num_active_div_by_paid_inv_0_12m	70052.00	NaN	NaN	NaN	0.38	71.37	0.00	0.00	0.00	0.10	18891.00

- **Missing Data**:
  - Some columns have missing values, such as acct_worst_status_12_24m with only 33,216 non-null entries out of 99,979, indicating significant missing data in some areas.



```
total_missing = df.isnull().sum().sum()
total_entries = df.size
percentage_missing = (total_missing / total_entries) * 100
print(f"Total percentage of missing values: {percentage_missing:.2f}%")
```

```
Total percentage of missing values: 20.01%
```

C. Understanding of Attributes

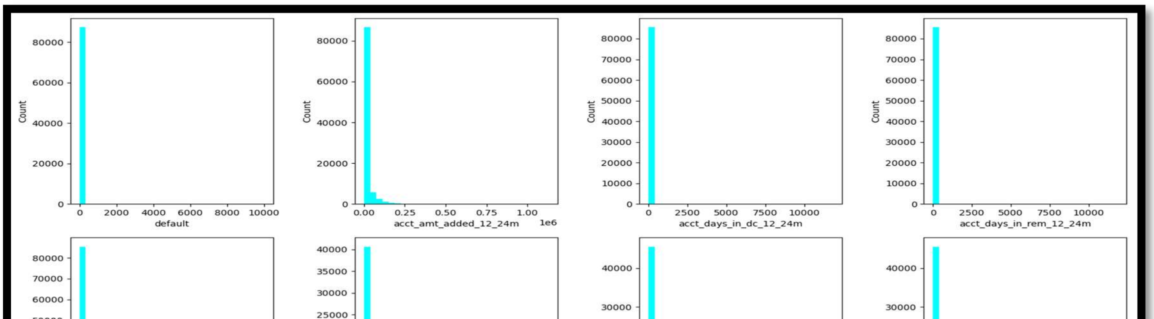
1. **Variable Information**:
   - **default**: The target variable, indicating whether a customer defaulted on their payments.

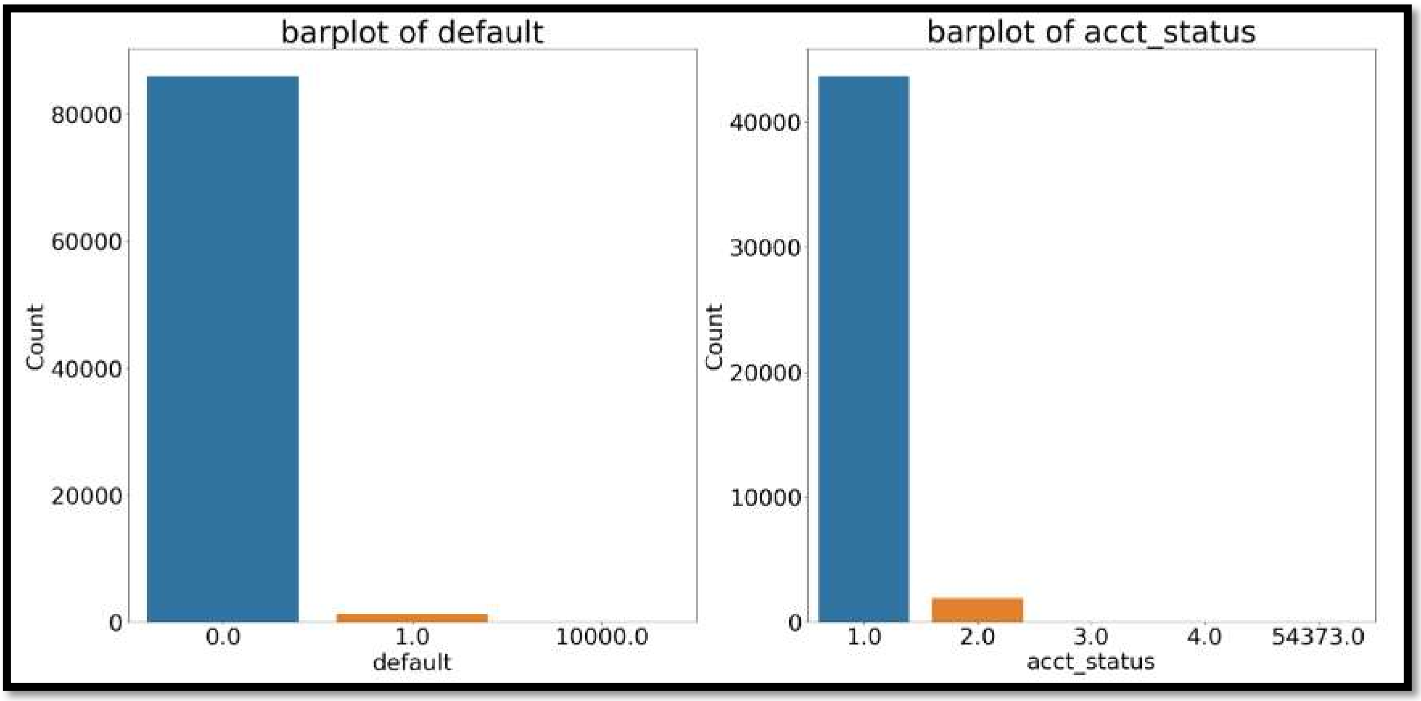
- **acct_amt_added_12_24m**: Total amount of purchases made in the last 12-24 months.
- **age**: The age of the customer.
- **merchant_category**: The category of merchants with which transactions were made.
- **num_active_inv**: Number of active invoices (unpaid bills).

2. **Renaming**:
   - Some columns could be renamed for clarity:
     - acct_amt_added_12_24m → total_purchase_amount_12_24m
     - acct_days_in_dc_12_24m → days_in_debt_collection_12_24m
     - acct_worst_status_* → worst_status_last_*m
     - num_active_div_by_paid_inv_0_12m → unpaid_vs_paid_invoices_0_12m
     - name_in_email → email_name_indicator

3] EXPLORATORY DATA ANALYSIS

A. UNIVARIATE ANALYSIS





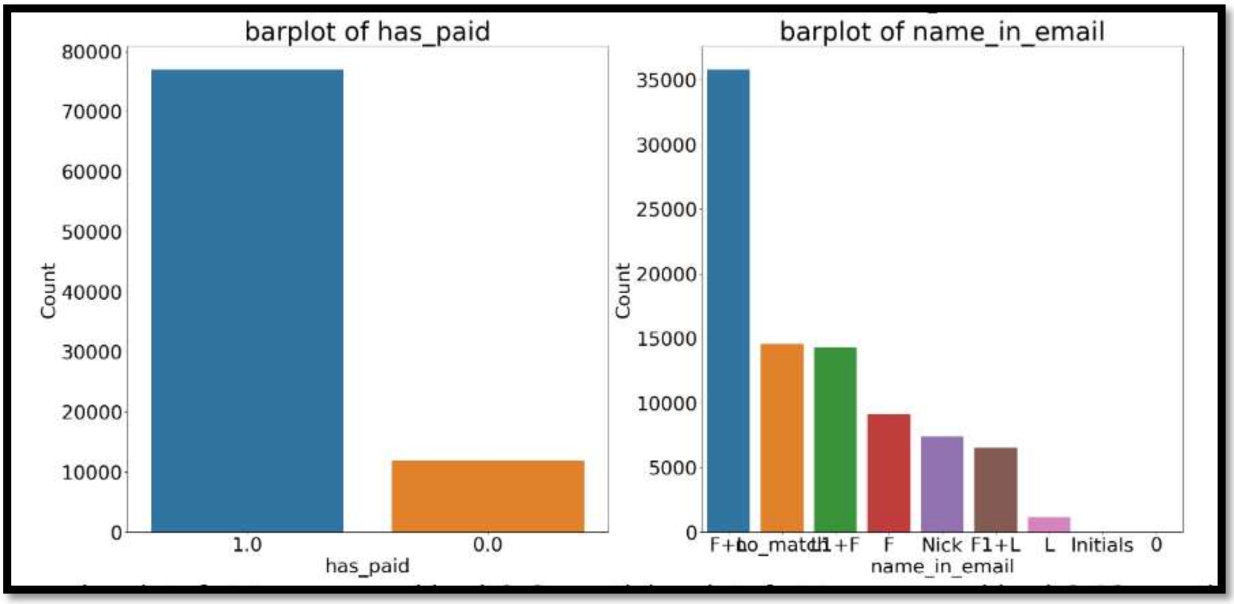
1. **Default:** This is the target variable.

Imbalance: The default variable is highly imbalanced with only 1,288 defaulters (1.29% of the total dataset).

Implication: This imbalance suggests the need for careful handling during model training, potentially requiring techniques like resampling (oversampling the minority class or undersampling the majority class) or using algorithms that account for class imbalance.

2. **acct_status (Account Status)**

- **Expected Categories**: The data dictionary suggests binary values 0 and 1, representing inactive and active account statuses, respectively.
- **Observed Categories**: The dataset contains four categories (1, 2, 3, 4).
  - **Category 1**: Constitutes 96% of the data, likely representing active accounts.
  - **Category 2**: Represents 4% of the data.
  - **Categories 3 & 4**: Have negligible values.
- **Action**: Further investigation is needed to clarify the nature of categories 2, 3, and 4, possibly indicating data entry errors.



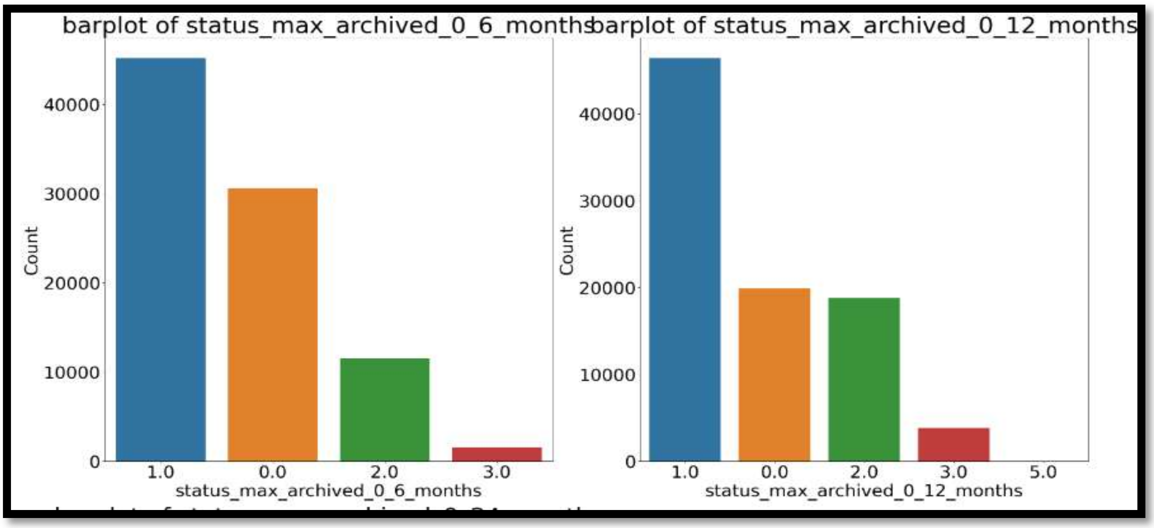
3. **has_paid (Bill Payment Status)**

- **Distribution**:
  - 87% of users have paid their current credit card bill.
  - 13% of users have not paid.
- **Implication**: This variable is useful in understanding the payment behavior of users, which might correlate with default risk.

4. **name_in_email**

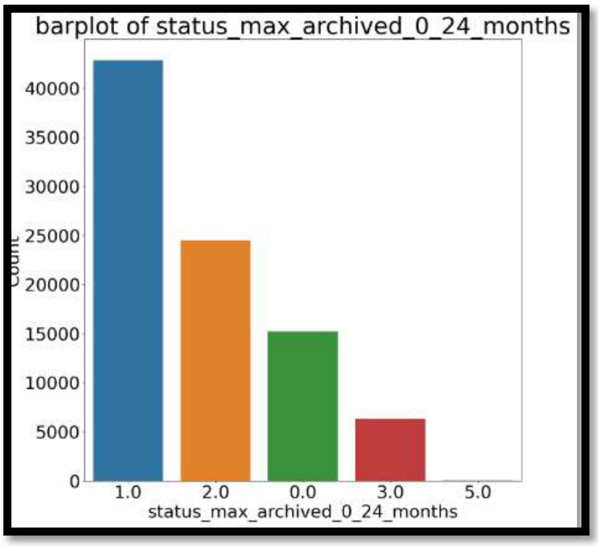
   Anomaly Detected: The bar plot reveals an anomaly, with values labeled as 'F' and 'L' instead of actual names.

   Action: Since this variable doesn't make logical sense and might be the result of a data entry error, it is recommended to drop this variable before proceeding to model building.



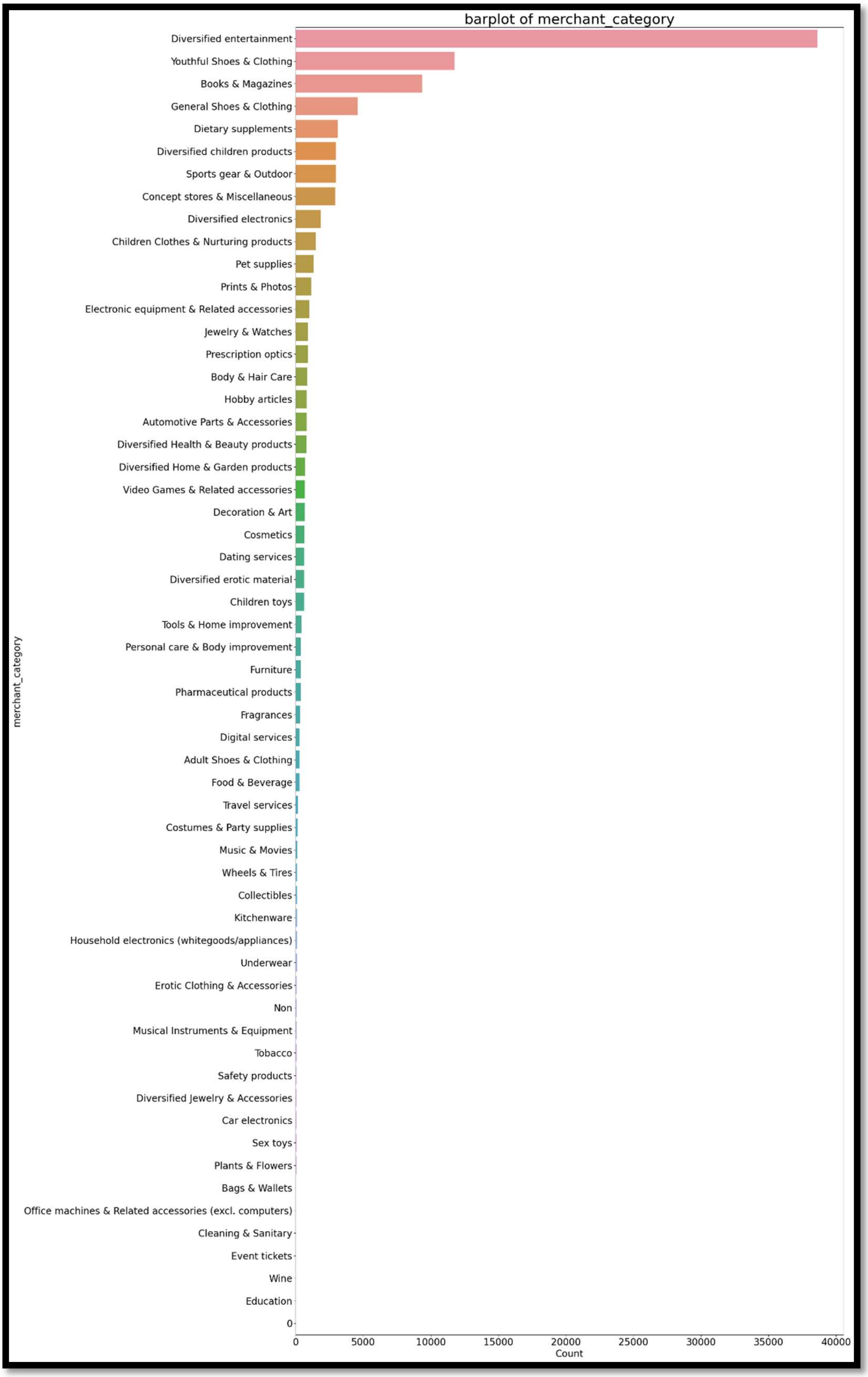
5. status_max_archived_0_24_months

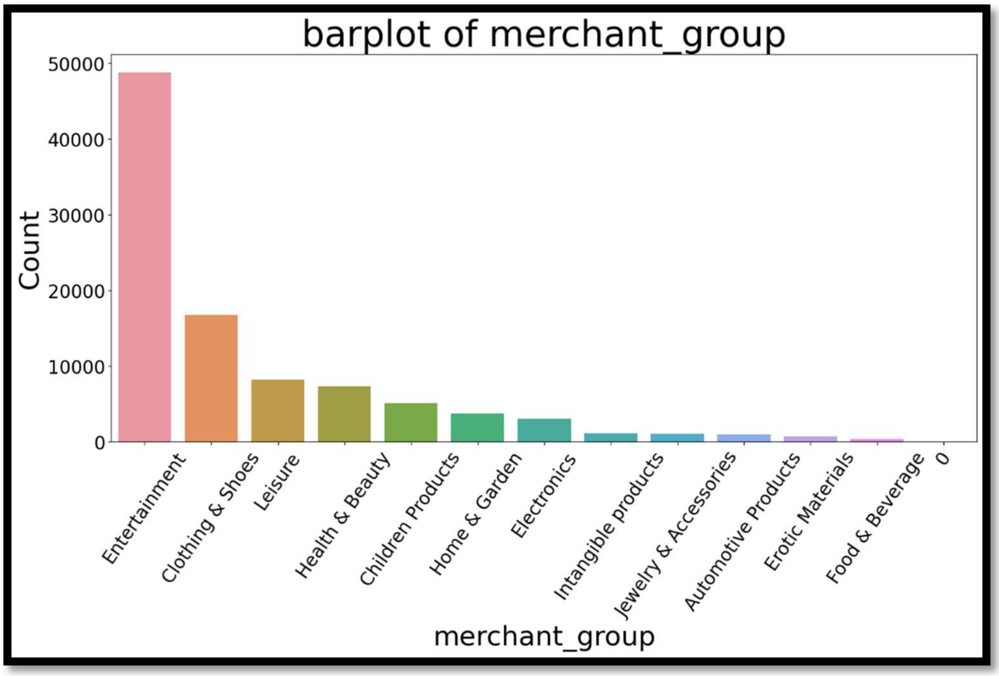
- **Categories**:
  - **48%** of users have been in archived status at least once in the past 24 months.
  - **17%** have never been archived.
  - **28%** have been archived twice.
  - **7%** have been archived three times.
  - **0.02% (16 users)** have been archived five times.
- **Implication**: This variable shows varying degrees of financial activity and risk, which might be useful for predicting default.



6. merchant_category and merchant_group

- **Assumed Meaning**: These variables likely represent the categories and groups where users have transacted the most.
  - **Most Frequent Merchant Group**: "Entertainment" followed by "Clothing & Shoes."
  - **Most Frequent Merchant Category**: "Diversified Entertainment" followed by "Youthful Shoes and Clothing."
- **Implication**: Understanding where users spend the most can provide insights into spending habits and potential risk factors.


barplot of merchant_category
Diversified entertainment
Youthful Shoes & Clothing
Books & Magazines
General Shoes & Clothing
Dietary supplements
Diversified children products
Sports gear & Outdoor
Concept stores & Miscellaneous
Diversified electronics
Children Clothes & Nurturing products
Pet supplies
Prints & Photos
Electronic equipment & Related accessories
Jewelry & Watches
Prescription optics
Body & Hair Care
Hobby articles
Automotive Parts & Accessories
Diversified Health & Beauty products
Diversified Home & Garden products
Video Games & Related accessories
Decoration & Art
Cosmetics
Dating services
Diversified erotic material
Children toys
Tools & Home improvement
Personal care & Body improvement
Furniture
Pharmaceutical products
Fragrances
Digital services
Adult Shoes & Clothing
Food & Beverage
Travel services
Costumes & Party supplies
Music & Movies
Wheels & Tires
Collectibles
Kitchenware
Household electronics (whitegoods/appliances)
Underwear
Erotic Clothing & Accessories
Non
Musical Instruments & Equipment
Tobacco
Safety products
Diversified Jewelry & Accessories
Car electronics
Sex toys
Plants & Flowers
Bags & Wallets
Office machines & Related accessories (excl. computers)
Cleaning & Sanitary
Event tickets
Wine
Education
0
merchant_category
0
5000
10000
15000
20000
25000
30000
35000
40000
Count



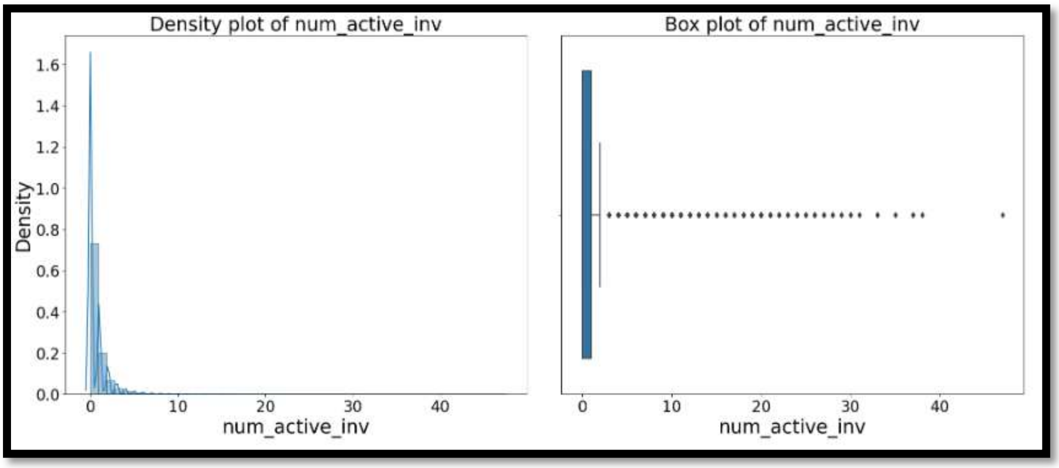
7. **Distribution of Continuous Variables**

- **Visualization**:
  - **Density Plot**: Preferred over histograms due to better execution time and improved visual representation.
  - **Box Plot**: Used alongside density plots to highlight the spread and outliers in the data. It is shown in the Jupyter notebook for reference.
- **Selected Variables**: Only the most important continuous variables are highlighted, with detailed analysis available in the accompanying Jupyter notebook.

Plot Combinations and Inferences

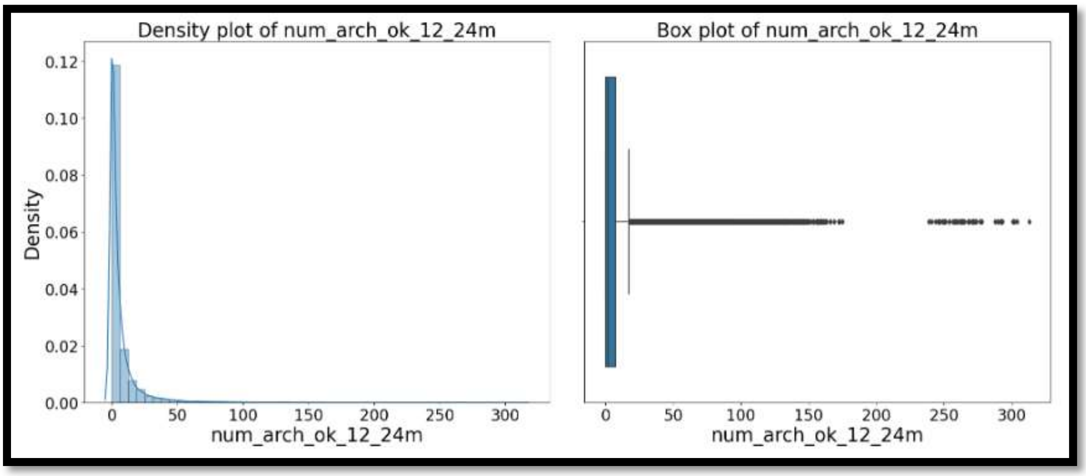
1. **num_active_inv (Total Number of Active Invoices per User)**

- **Definition**: This variable represents the total number of unpaid bills (active invoices) per user.
- **Distribution**:
  - **69% of users** have no active invoices, which indicates a good payment discipline among the majority of users.
  - **Right-Skewed Distribution**: The majority of users have between 1 to 5 active invoices, but the distribution has a long right tail with some users having over 30 active invoices.
  - **Outliers**: The presence of users with more than 30 active invoices indicates potential high-risk individuals who may struggle to pay off their debts.



2. num_arch_ok_12_24m (Number of Archived Purchases Paid in the Last 12 to 24 Months)

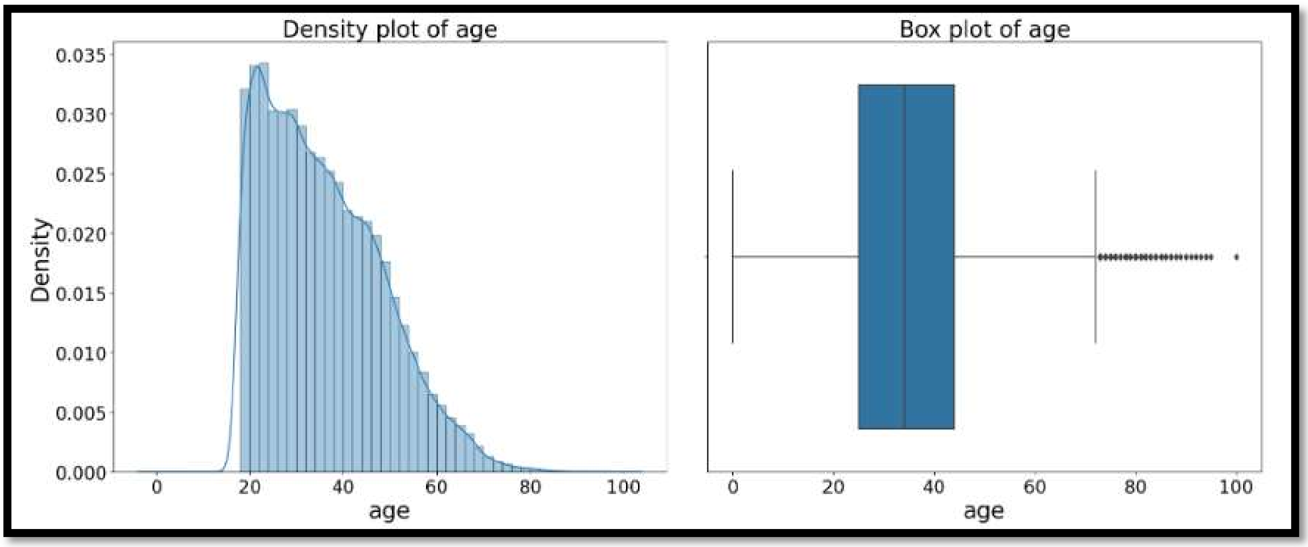
- **Definition**: This variable indicates the number of archived purchases that were paid between 12 and 24 months ago.
- **Distribution**:
  - **Right-Skewed Distribution**: The distribution is heavily right-skewed, with most users having a relatively small number of archived purchases.
  - **Outliers**: Some users have an extremely high number of archived purchases (up to 300), which again indicates potential outliers and possibly different spending or payment patterns compared to the majority.



3. Age

- **Distribution**:
  - **Majority Age Range**: Approximately 95% of users fall within the age range of 20 to 80 years.

- **Outliers**: While outliers exist in the dataset, the age distribution is relatively normal compared to other variables, suggesting a more stable demographic profile of the users.



The significant number of outliers suggests a need for further investigation. These outliers could either be high-risk users or data entry errors.

Data Imbalance:

- For variables like num_active_inv, where the majority of users have no active invoices, the imbalance could influence model performance. Techniques such as stratified sampling or specialized algorithms may be needed.

Age Distribution:

- The relatively normal distribution of age indicates that age could be a stable predictor in the model. However, outliers should still be checked to ensure they are legitimate values.

Univariate Analysis Inferences

1. **Missing Values:**
    - Rows with excessive missing values should be dropped to maintain the integrity of the dataset. Ensure to document the extent and reason for missing data.
2. **Unexpected Values:**
    - **acct_status:** This variable should ideally only contain 0 and 1, but it has values 1, 2, 3, and 4. Investigate the source or context for these values. If clarification cannot be obtained, consider dropping the variable.
3. **Irrelevant Variables:**

- **userid** and **name_in_email:** These do not contribute to model building and should be removed.

4. **Skewness and Outliers:**
   - Many continuous variables exhibit high skewness and outliers. Consider transformations (e.g., log transformation) and outlier treatment methods (e.g., capping) to normalize distributions.
5. **Scaling:**
   - Variables are on different scales, so scaling using StandardScaler is appropriate to ensure all features contribute equally to the model.

Recommendations

1. **Handling Missing Values:**
   - Drop rows with excessive missing values or consider imputation methods based on the nature of the missing data (mean, median, mode, or predictive imputation).
2. **Investigate Anomalous Values:**
   - Reach out to the data source or consult domain experts to understand unexpected values, especially for acct_status.
3. **Outlier Detection and Treatment:**
   - Identify and treat outliers using methods such as IQR (Interquartile Range) or Z-scores. Document how outliers are handled for transparency.
4. **Data Transformation:**
   - Apply transformations to skewed distributions, such as logarithmic transformations, to reduce skewness.
5. **Data Scaling:**
   - Use StandardScaler to standardize the features so that they all have a mean of 0 and a standard deviation of 1.
6. **Exploratory Data Analysis (EDA):**
   - Continue with EDA to uncover patterns or relationships that could inform feature selection or engineering.

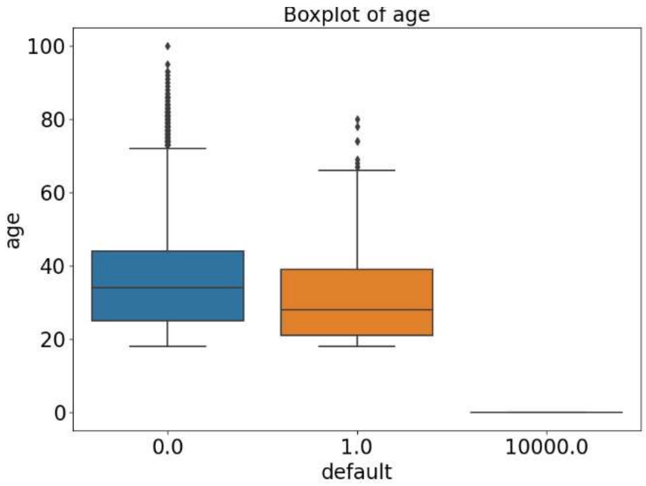
B. BIVARIATE ANALYSIS

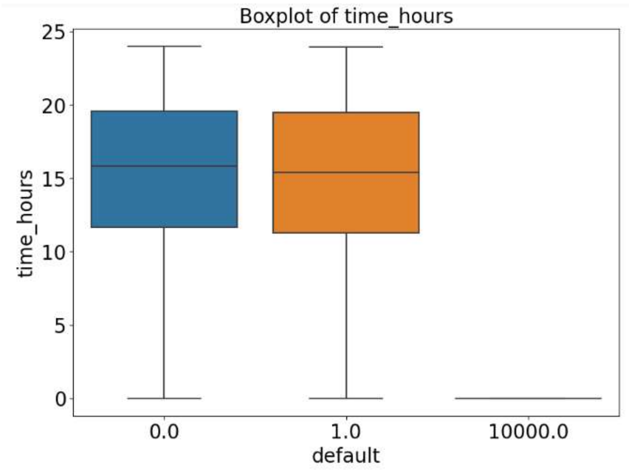
1. Bivariate Analysis with the Target Variable

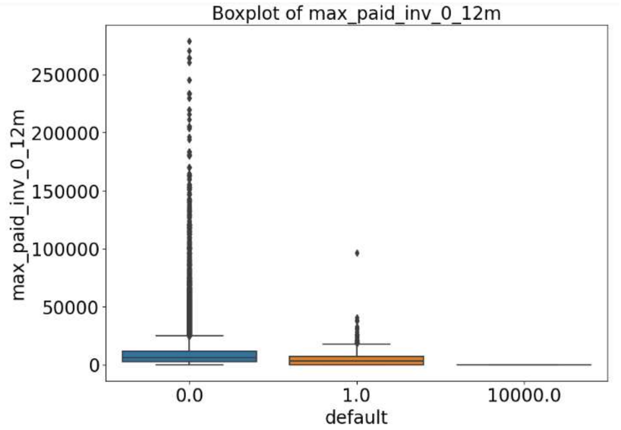
Given the extreme imbalance in the target variable (default), visualizing the relationship between categorical variables and the target variable using bar plots may not be very informative. Instead, consider the following approaches:

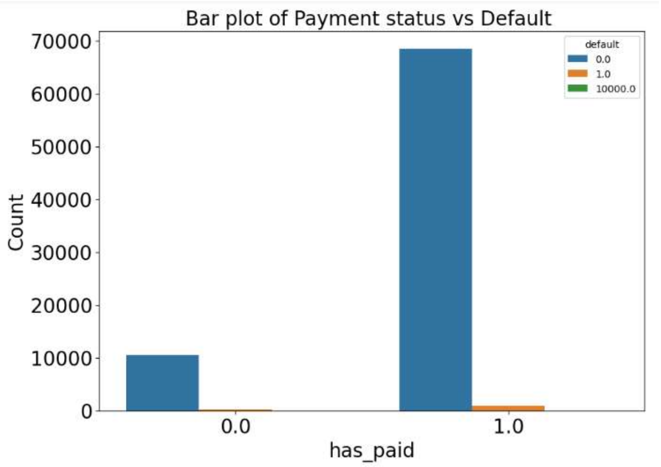
- **Categorical Variables:**
  - **Bar Plots with Proportions:** Compare proportions of defaulters and non-defaulters within each category. This helps in understanding if certain categories are more likely to default.
  - **Stacked Bar Charts:** Use stacked bar charts to show the distribution of the target variable within each category.

- **Continuous Variables:**
  - **Box Plots:** Compare the distribution of continuous variables for defaulters and non-defaulters using box plots. This can help in understanding how the distributions differ between the two classes.









Insights from Bivariate Analysis with the Target Variable

1. **Age vs. Default:**

- **Observation:** Defaulters tend to be younger, with an average age of around 30 years, compared to non-defaulters whose average age is closer to 40 years.
- **Insight:** Younger users, particularly those in the 20-30 age range, are more likely to default. This may suggest that age could be an important predictor of default risk.

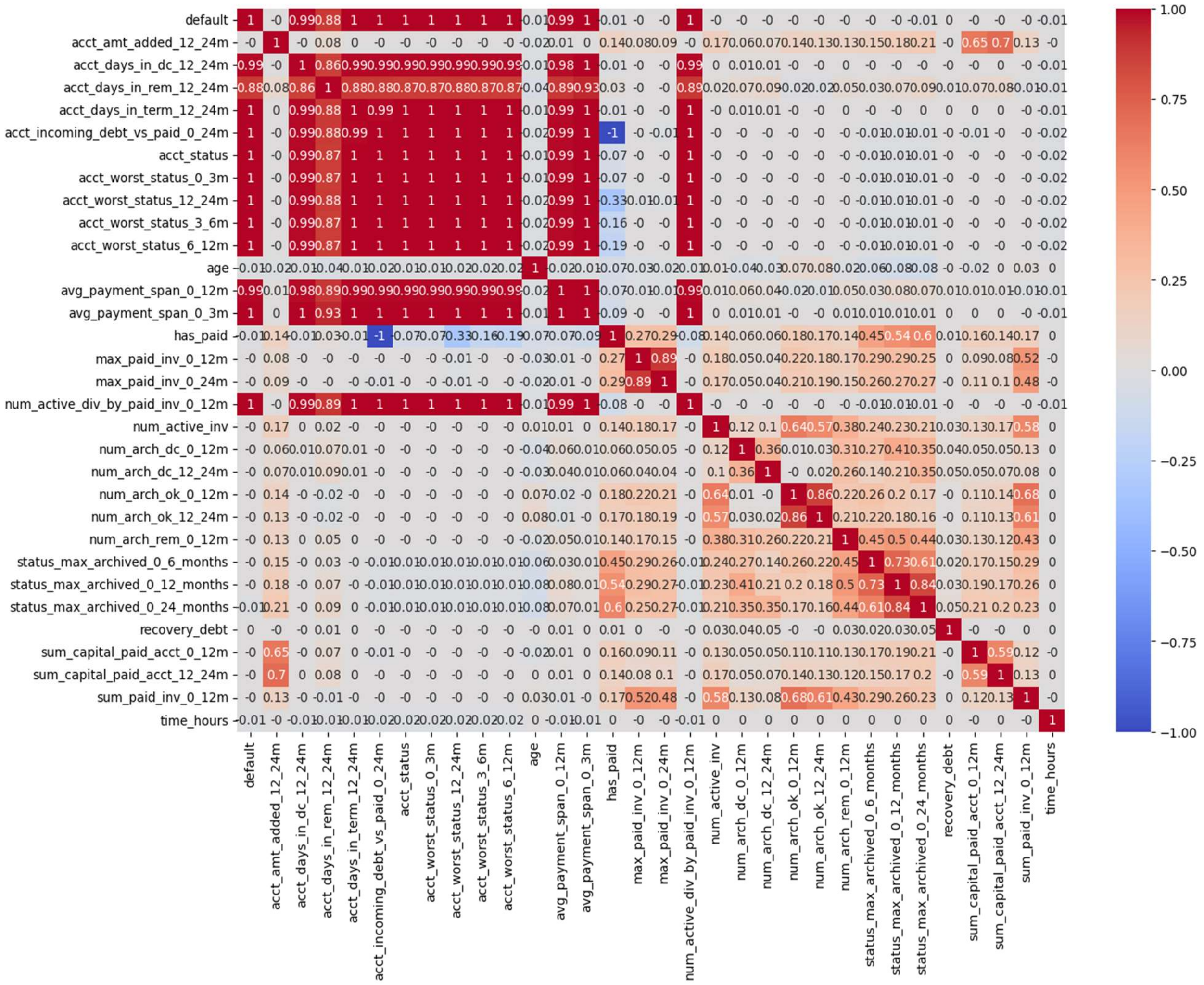
2. **Average Payment Span in Last 3 Months vs. Default:**
   - **Observation:** Defaulters generally take longer to make payments (average of 20 days) compared to non-defaulters (average of 15 days). There are significant outliers in non-defaulters, with some taking up to 80 days.
   - **Insight:** Users who take more than 15 days to make a payment are more likely to default. This indicates that late payments are a risk factor, though outliers should be considered carefully as they may distort the overall analysis.
3. **Maximum Bill Amount Payment in Last 12 Months vs. Default:**
   - **Observation:** There is minimal difference between defaulters and non-defaulters in terms of maximum bill amount payments, though defaulters have fewer outliers compared to non-defaulters.
   - **Insight:** While maximum bill amounts are similar between defaulters and non-defaulters, the presence of many outliers among non-defaulters suggests that spending behavior may differ. Non-defaulters might have more financial flexibility, while defaulters might have tighter financial constraints.
4. **Current Payment Status vs. Default:**
   - **Observation:** There is little difference in default rates between users who have paid their current credit card bill and those who haven't.
   - **Insight:** The current payment status alone may not be a strong predictor of default. This suggests that other factors, such as payment history and financial behavior, might be more indicative of default risk.

INFERENCES

- **Age:** Younger users are at higher risk of default, suggesting age could be a valuable feature in predicting default risk.
- **Payment Span:** Users with longer payment spans are more likely to default, but be cautious of outliers.
- **Bill Payment Amount:** Although the maximum bill amount doesn’t show significant differences, the outlier presence indicates differing financial behaviors.
- **Current Payment Status:** Current payment status is less indicative of default, highlighting the need for a comprehensive analysis of payment history and other financial behaviors.

C. MULTIVARIATE ANALYSIS

The correlation heatmap provides a visual representation of the relationships between different variables in the dataset. The color intensity and direction of the color (red or blue) indicate the strength and type of correlation.



Inferences from Multivariate Analysis

1. **Correlation Insights:**
   - **High Correlation Pairs:**
     - **avg_payment_span_0_12m and avg_payment_span_0_3m:** These variables measure the average payment span over different time periods, and their high correlation suggests that they provide overlapping information about payment behavior.
     - **max_paid_inv_0_12m and max_paid_inv_0_24m:** Similarly, these variables measure the maximum paid invoice amounts over different time frames, indicating that they may capture redundant information.
     - **acct_days_in_dc_12_24m, acct_days_in_rem_12_24m, acct_days_in_term_12_24m:** These variables appear to be highly correlated with each other, suggesting that they might be measuring similar aspects of account activity.

        - **num_arch_dc_0_12m, num_arch_dc_12_24m, num_arch_ok_0_12m, num_arch_ok_12_24m, num_arch_rem_0_12m:** These variables are also strongly correlated, indicating a relationship between the different types of account archives.

2. **Moderate Positive Correlations:**

    - **acct_amt_added_12_24m with sum_capital_paid_acct_0_12m and sum_capital_paid_acct_12_24m:** This suggests that accounts with higher amounts added are more likely to have higher capital payments.
    - **age with time_hours:** Older accounts might have been active for a longer duration.

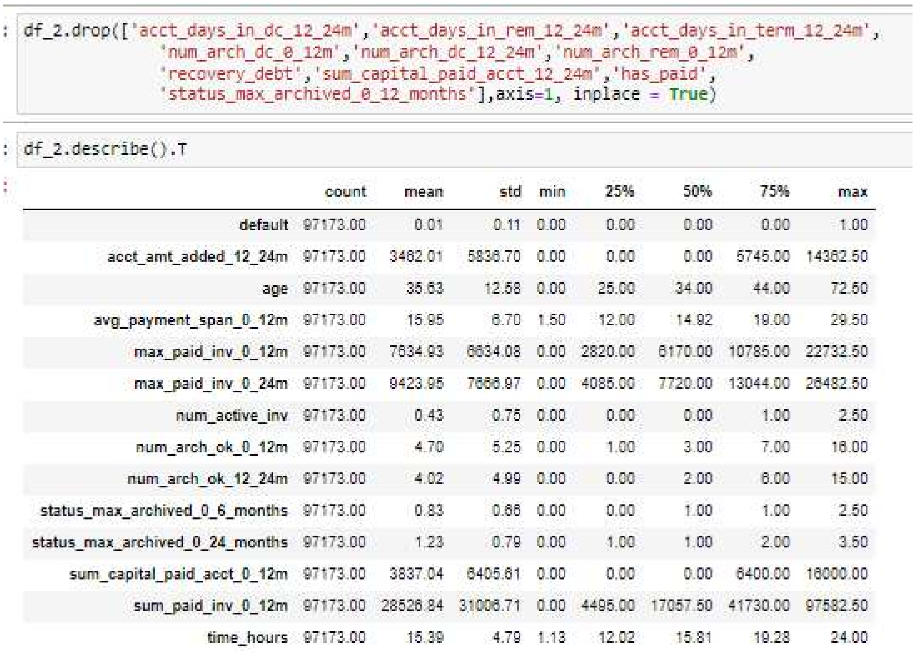
3. **Negative Correlations:**

    - **default with acct_amt_added_12_24m, sum_capital_paid_acct_0_12m, sum_capital_paid_acct_12_24m, sum_paid_inv_0_12m, time_hours:** These negative correlations suggest that accounts with higher activity or payments are less likely to default.

4. **Multicollinearity:**
    - The presence of high correlations among these variables indicates multicollinearity in the dataset. Multicollinearity can lead to redundancy and can adversely affect the performance of regression models by making the coefficient estimates unstable and difficult to interpret.

5. **Mitigation Strategy:**
    - **Variance Inflation Factor (VIF):** To address multicollinearity, you will use the VIF technique. VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors.
        - **Calculation:** For each variable, VIF is calculated by running a regression of that variable against all other variables and then computing the ratio of the variance of this regression to the variance of the original variable.
        - **Threshold:** A common threshold for high multicollinearity is a VIF value greater than 10. Variables with high VIF values can be considered for removal or transformation.

D. REMOVAL OF UNWANTED VARIABLES



```python
df_2.drop(['acct_days_in_dc_12_24m','acct_days_in_rem_12_24m','acct_days_in_term_12_24m',
           'num_arch_dc_0_12m','num_arch_dc_12_24m','num_arch_rem_0_12m',
           'recovery_debt','sum_capital_paid_acct_12_24m','has_paid',
           'status_max_archived_0_12_months'],axis=1, inplace = True)
```

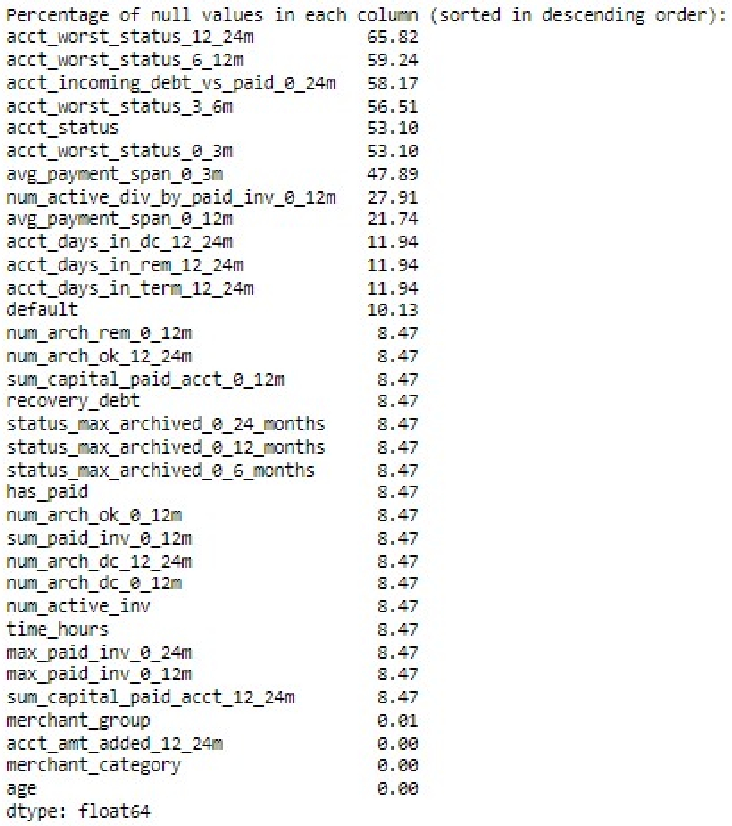
```python
df_2.describe().T
```

	count	mean	std	min	25%	50%	75%	max
default	97173.00	0.01	0.11	0.00	0.00	0.00	0.00	1.00
acct_amt_added_12_24m	97173.00	3482.01	5836.70	0.00	0.00	0.00	5745.00	14362.50
age	97173.00	35.63	12.58	0.00	25.00	34.00	44.00	72.50
avg_payment_span_0_12m	97173.00	15.95	6.70	1.50	12.00	14.92	19.00	29.50
max_paid_inv_0_12m	97173.00	7634.93	6634.08	0.00	2820.00	6170.00	10785.00	22732.50
max_paid_inv_0_24m	97173.00	9423.95	7666.97	0.00	4085.00	7720.00	13044.00	26482.50
num_active_inv	97173.00	0.43	0.75	0.00	0.00	0.00	1.00	2.50
num_arch_ok_0_12m	97173.00	4.70	5.25	0.00	1.00	3.00	7.00	16.00
num_arch_ok_12_24m	97173.00	4.02	4.99	0.00	0.00	2.00	6.00	15.00
status_max_archived_0_6_months	97173.00	0.83	0.86	0.00	0.00	1.00	1.00	2.50
status_max_archived_0_24_months	97173.00	1.23	0.79	0.00	1.00	1.00	2.00	3.50
sum_capital_paid_acct_0_12m	97173.00	3837.04	6405.61	0.00	0.00	0.00	6400.00	16000.00
sum_paid_inv_0_12m	97173.00	28526.84	31006.71	0.00	4495.00	17057.50	41730.00	97582.50
time_hours	97173.00	15.39	4.79	1.13	12.02	15.81	19.28	24.00

To streamline the dataset for model building, the following variables will be removed:

1. **Redundant Variables:**
   - **userid** and **name_in_email:** These variables are considered irrelevant for the modeling process as they do not contribute meaningful information for predicting credit card default.
2. **Categorical Variables with High Cardinality:**
   - **merchant_category** and **merchant_group:** Both of these categorical variables exhibit high cardinality, with **merchant_category** containing 58 unique categories and **merchant_group** containing 13 unique categories. Due to their high number of unique categories, encoding these variables (e.g., through one-hot encoding or label encoding) would not be effective and could lead to an explosion in feature space without adding significant value to the model. Consequently, these variables will be excluded from the analysis.

E. MISSING VALUE TREATMENT



```
Percentage of null values in each column (sorted in descending order):
acct_worst_status_12_24m          65.82
acct_worst_status_6_12m           59.24
acct_incoming_debt_vs_paid_0_24m  58.17
acct_worst_status_3_6m            56.51
acct_status                       53.10
acct_worst_status_0_3m            53.10
avg_payment_span_0_3m             47.89
num_active_div_by_paid_inv_0_12m  27.91
avg_payment_span_0_12m            21.74
acct_days_in_dc_12_24m            11.94
acct_days_in_rem_12_24m           11.94
acct_days_in_term_12_24m          11.94
default                           10.13
num_arch_rem_0_12m                 8.47
num_arch_ok_12_24m                 8.47
sum_capital_paid_acct_0_12m        8.47
recovery_debt                      8.47
status_max_archived_0_24_months    8.47
status_max_archived_0_12_months    8.47
status_max_archived_0_6_months     8.47
has_paid                           8.47
num_arch_ok_0_12m                  8.47
sum_paid_inv_0_12m                 8.47
num_arch_dc_12_24m                 8.47
num_arch_dc_0_12m                  8.47
num_active_inv                     8.47
time_hours                         8.47
max_paid_inv_0_24m                 8.47
max_paid_inv_0_12m                 8.47
sum_capital_paid_acct_12_24m       8.47
merchant_group                     0.01
acct_amt_added_12_24m              0.00
merchant_category                  0.00
age                                0.00
dtype: float64
```

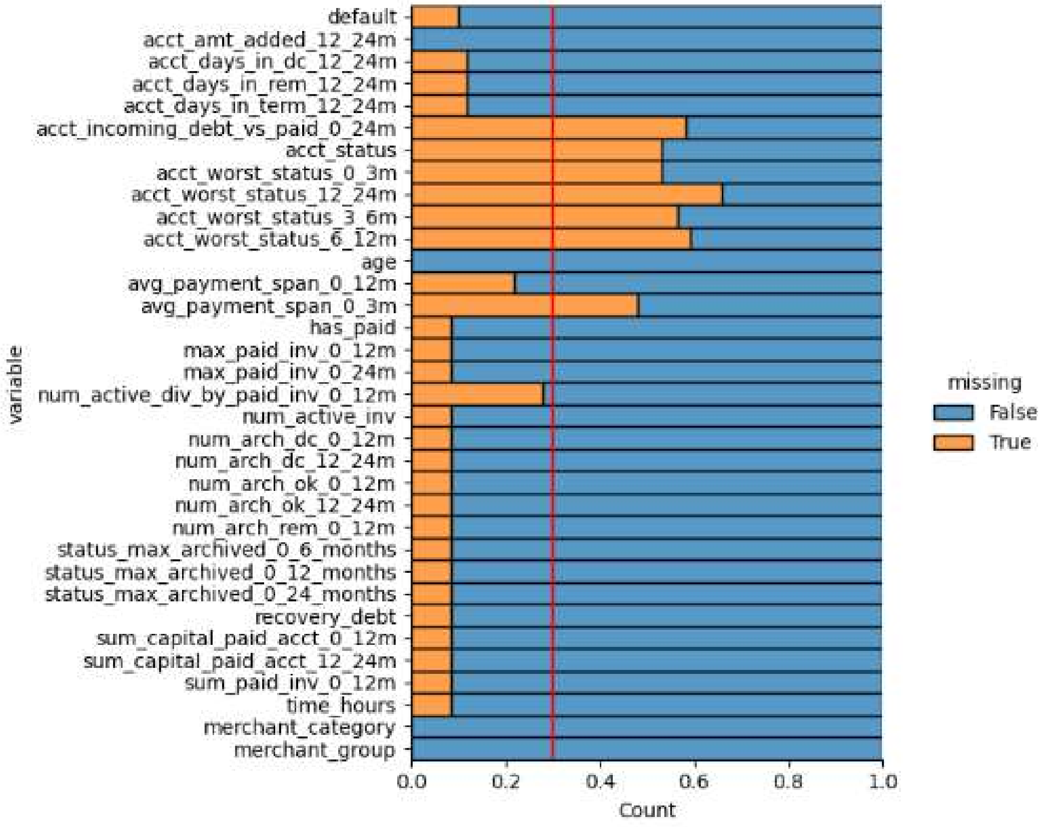
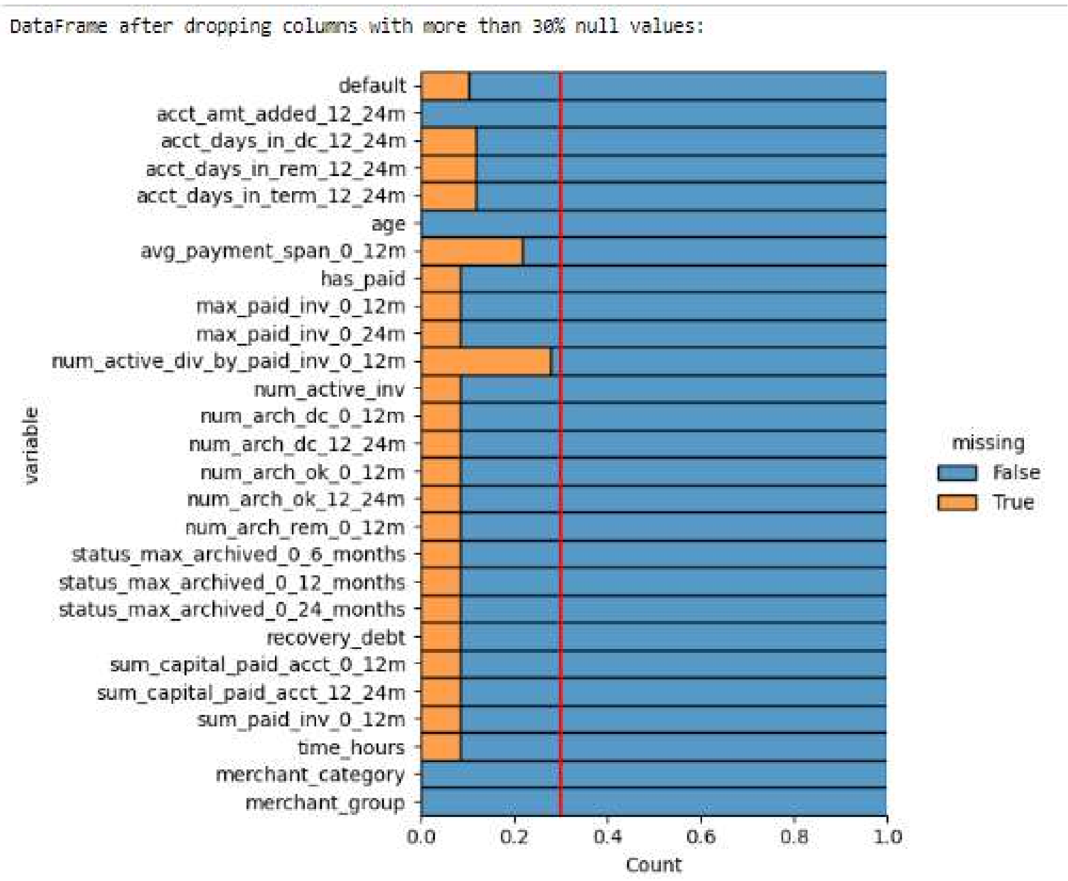
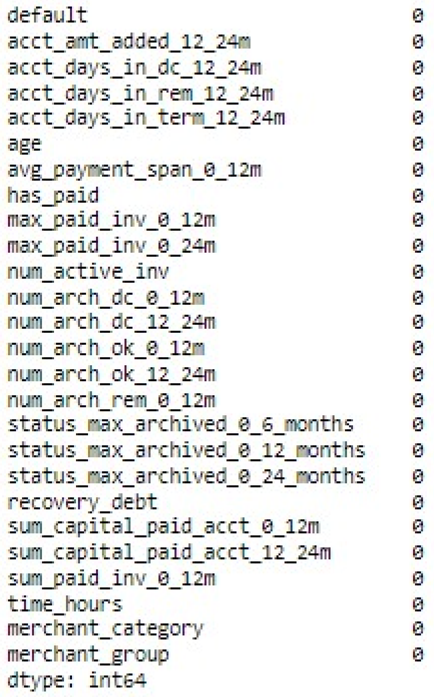


Figure: Before Dropping Columns with 30% Null Values



```
DataFrame after dropping columns with more than 30% null values:
```

Figure: After Dropping Columns with 30% Null Values



```
default                            0
acct_amt_added_12_24m              0
acct_days_in_dc_12_24m             0
acct_days_in_rem_12_24m            0
acct_days_in_term_12_24m           0
age                                0
avg_payment_span_0_12m             0
has_paid                           0
max_paid_inv_0_12m                 0
max_paid_inv_0_24m                 0
num_active_inv                     0
num_arch_dc_0_12m                  0
num_arch_dc_12_24m                 0
num_arch_ok_0_12m                  0
num_arch_ok_12_24m                 0
num_arch_rem_0_12m                 0
status_max_archived_0_6_months     0
status_max_archived_0_12_months    0
status_max_archived_0_24_months    0
recovery_debt                      0
sum_capital_paid_acct_0_12m        0
sum_capital_paid_acct_12_24m       0
sum_paid_inv_0_12m                 0
time_hours                         0
merchant_category                  0
merchant_group                     0
dtype: int64
```

Figure: After Missing Value Treatment

Visual Inspection:

- A heatmap of the dataset shows the distribution of missing values, with white bars indicating their presence. The dataset contains approximately 20% missing values, necessitating a careful approach to treatment.

Handling Missing Values:

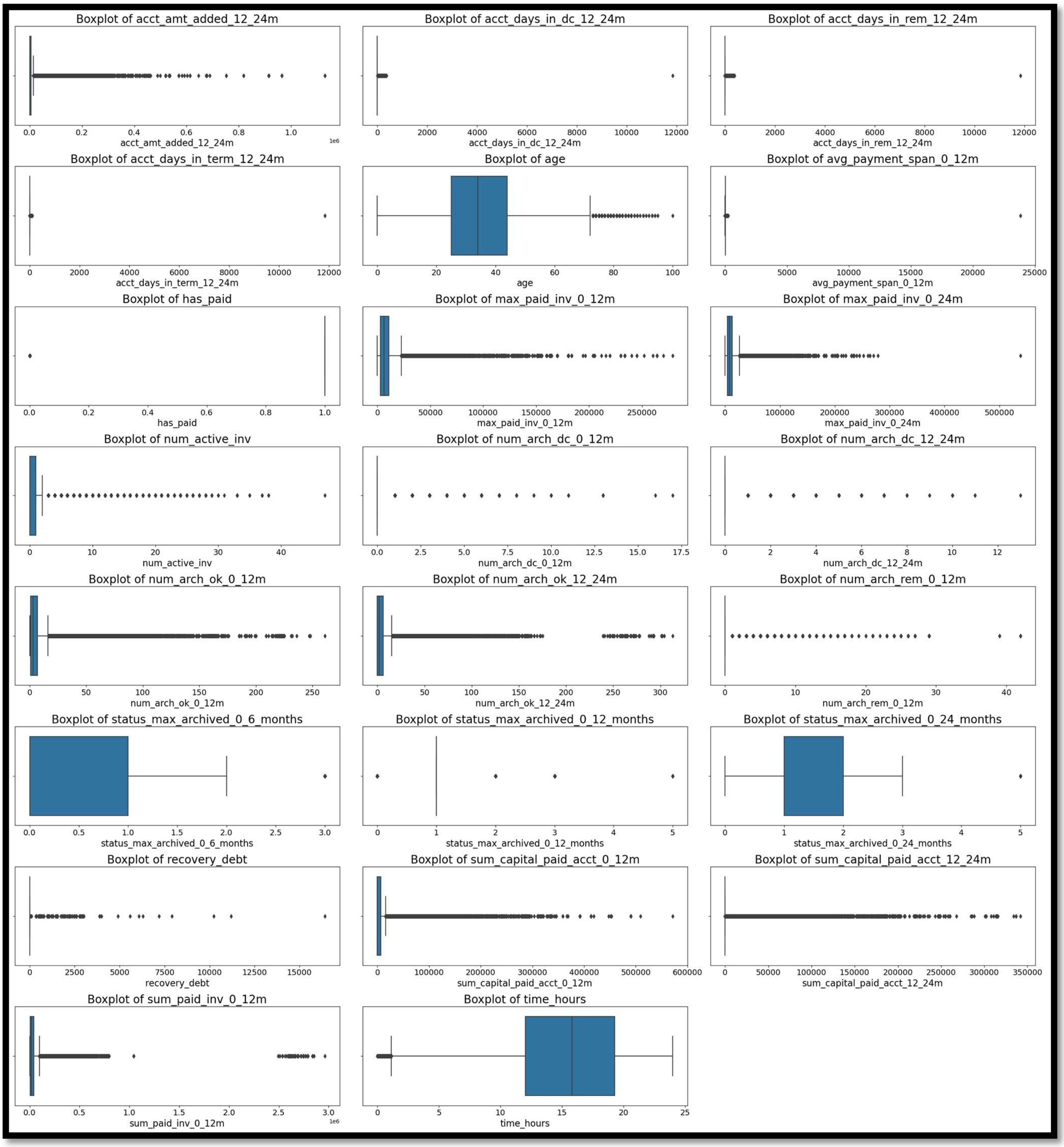
1. **Dropping Variables:**
   - Variables with more than 30% missing values will be dropped. Imputing such a large proportion of missing values could introduce analyst bias, as different imputation methods may lead to varying results. After removing these high-missing variables, we retain 24 independent variables.
2. **Imputing Remaining Missing Values:**
   - **merchant_category** and **merchant_group:** Due to the high proportion of missing values in these categorical variables, the missing values will be dropped rather than imputed.
   - **Target Variable (default):** With approximately 10% missing values, we will impute the missing values using the mode (0, indicating non-default). Following imputation, the dataset will be resampled using the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance.
   - **Continuous Variables:** Missing values will be imputed with the median value to mitigate the impact of outliers.

Key Observations:

- **High Proportion of Missing Values:** Several variables have a significant number of missing values. Notable examples include acct_incoming_debt_vs_paid_0_24m, acct_status, acct_worst_status_*, and avg_payment_span_*.
- **Consistent Missingness:** Some variables, such as acct_status, acct_worst_status_*, and avg_payment_span_*, seem to have a consistent pattern of missingness. This might suggest a common underlying reason for the missing data.
- **Low Proportion of Missing Values:** A few variables, such as default, acct_amt_added_12_24m, and age, have very few missing values.

F. OUTLIER TREATMENT

Checking the outliers per column using boxplot:



Observation:

- All variables except **time_hours** contain outliers. The extent of outliers varies across variables—some have only a few, while others have a large number.
- **Presence of Outliers:** Several variables exhibit outliers, as indicated by the individual data points outside the whiskers of the boxplots.

- **Skewness:** The distribution of some variables is skewed, which can contribute to the presence of outliers.

Outlier Treatment:

- The Inter-Quartile Range (IQR) method was employed to treat outliers across all variables. The IQR method is effective for identifying and handling outliers by capping or removing extreme values outside the acceptable range defined by the first and third quartiles.
  - **Removal:** If outliers are deemed to be errors or anomalies, they can be removed from the dataset. However, this should be done cautiously to avoid losing valuable information.

Post-Treatment Analysis:

- After applying the IQR method, some variables were left with only zeros or ones, rendering them non-informative for further analysis. These variables were subsequently dropped. Methods that are less sensitive to outliers, such as median and interquartile range, can be used for summary statistics.

Final Dataset:

- Following the outlier treatment and removal of non-informative variables, the dataset was reduced to 14 variables.

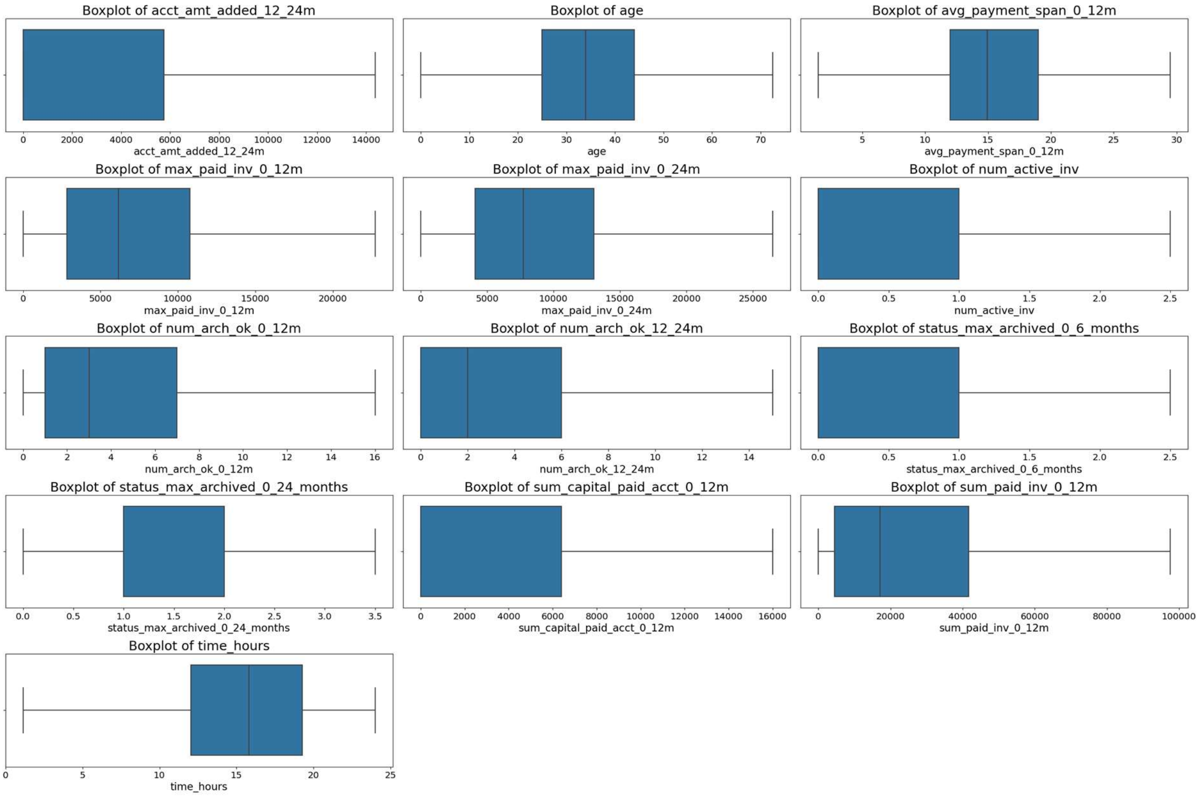
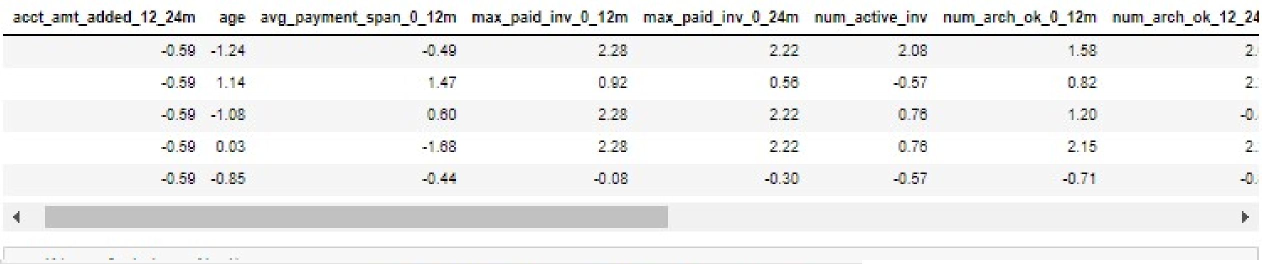
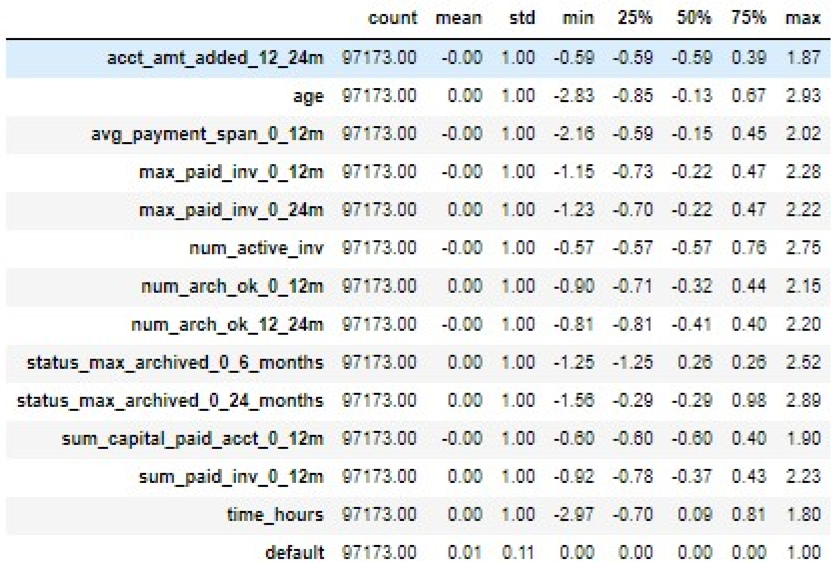


Figure: After Treatment of Outliers

G. VARIABLE TRANSFORMATION



acct_amt_added_12_24m	age	avg_payment_span_0_12m	max_paid_inv_0_12m	max_paid_inv_0_24m	num_active_inv	num_arch_ok_0_12m	num_arch_ok_12_24
-0.59	-1.24	-0.49	2.28	2.22	2.08	1.58	2.
-0.59	1.14	1.47	0.92	0.56	-0.57	0.82	2.
-0.59	-1.08	0.60	2.28	2.22	0.76	1.20	-0.
-0.59	0.03	-1.68	2.28	2.22	0.76	2.15	2.
-0.59	-0.85	-0.44	-0.08	-0.30	-0.57	-0.71	-0.



	count	mean	std	min	25%	50%	75%	max
acct_amt_added_12_24m	97173.00	-0.00	1.00	-0.59	-0.59	-0.59	0.39	1.87
age	97173.00	0.00	1.00	-2.83	-0.85	-0.13	0.67	2.93
avg_payment_span_0_12m	97173.00	-0.00	1.00	-2.16	-0.59	-0.15	0.45	2.02
max_paid_inv_0_12m	97173.00	-0.00	1.00	-1.15	-0.73	-0.22	0.47	2.28
max_paid_inv_0_24m	97173.00	0.00	1.00	-1.23	-0.70	-0.22	0.47	2.22
num_active_inv	97173.00	-0.00	1.00	-0.57	-0.57	-0.57	0.76	2.75
num_arch_ok_0_12m	97173.00	0.00	1.00	-0.90	-0.71	-0.32	0.44	2.15
num_arch_ok_12_24m	97173.00	-0.00	1.00	-0.81	-0.81	-0.41	0.40	2.20
status_max_archived_0_6_months	97173.00	0.00	1.00	-1.25	-1.25	0.26	0.26	2.52
status_max_archived_0_24_months	97173.00	0.00	1.00	-1.56	-0.29	-0.29	0.98	2.89
sum_capital_paid_acct_0_12m	97173.00	-0.00	1.00	-0.60	-0.60	-0.60	0.40	1.90
sum_paid_inv_0_12m	97173.00	0.00	1.00	-0.92	-0.78	-0.37	0.43	2.23
time_hours	97173.00	0.00	1.00	-2.97	-0.70	0.09	0.81	1.80
default	97173.00	0.01	0.11	0.00	0.00	0.00	0.00	1.00

Objective:

- The dataset contains variables on different scales, which can affect the performance of distance-based algorithms such as K-means Clustering. Therefore, it was necessary to standardize these variables to bring them onto a common scale.

Approach:

- We applied **StandardScaler** from the **scikit-learn** library in Python to perform **Z-score normalization** on the dataset.
- StandardScaler transforms the data by adjusting each variable to have a **mean of 0** and a **standard deviation of 1**.

Benefits:

- Standardizing the variables ensures that each feature contributes equally to the distance metric in the clustering algorithm, preventing any bias from variables with larger scales.
- This transformation is crucial for maintaining the effectiveness and accuracy of K-means Clustering and other machine learning algorithms that rely on distance calculations.

By standardizing the dataset, we ensure that all variables are on the same scale, improving the performance of our clustering and other machine learning models.

4] BUSINESS INSIGHTS FROM EDA

A. Is the data unbalanced? If so, what can be done? Please explain in the context of the business

Data Imbalance:

- **Observation**: The dataset is highly imbalanced, with only 1% of the 99,979 users being defaulters.
- **Impact**: This imbalance can lead to biased model predictions, where the model may favor the majority class (non-defaulters) and fail to accurately identify defaulters.
- **Solution**: To address this, techniques such as **SMOTE (Synthetic Minority Over-sampling Technique)** or **undersampling the majority class** can be employed to balance the target class “Defaults”. These methods help balance the dataset by either generating synthetic data for the minority class or reducing the size of the majority class, leading to more accurate model predictions.
- **Business Context**: In the context of credit risk management, accurately predicting defaulters is crucial. By addressing the imbalance, the model can better identify potential defaulters, enabling the business to take proactive measures, such as early intervention or more stringent credit checks, thus minimizing financial losses.

B. Any business insights using Clustering

Clustering Insights:

- **Observation**: Clustering can reveal hidden patterns in the data, such as grouping users based on spending behavior, payment habits, or account status.
- **Example**: By performing **K-means clustering**, users can be segmented into different clusters based on variables like payment span, max paid invoice, and account status.
- **Business Insight**: For instance, one cluster might represent users who consistently pay on time and have a high credit limit, while another cluster might represent users who frequently delay payments and have a higher likelihood of defaulting.
- **Application**: These insights can be used to tailor marketing strategies, offer personalized credit products, or prioritize collections efforts based on the risk profile of each cluster.

C. Any Other Business Insights

- **Age Factor**: Younger users (aged 20-30) are more likely to default compared to older users. This could indicate that younger users might have less financial stability or are more prone to risky financial behavior. The business could consider offering financial education or designing credit products that cater specifically to younger users to reduce default rates.
- **Payment Behavior**: Users who take longer than 15 days to pay their credit card bills are more likely to default. The business could implement stricter payment terms or offer incentives for early payment to encourage timely bill settlements.
- **Spending Habits**: There is no significant difference in the maximum bill amounts between defaulters and non-defaulters, but non-defaulters have more outliers in spending. This suggests that while defaulters may be more conservative in spending, they may also be at a tipping point financially. Offering targeted financial advice or monitoring spending patterns could help in identifying and supporting users at risk of defaulting.