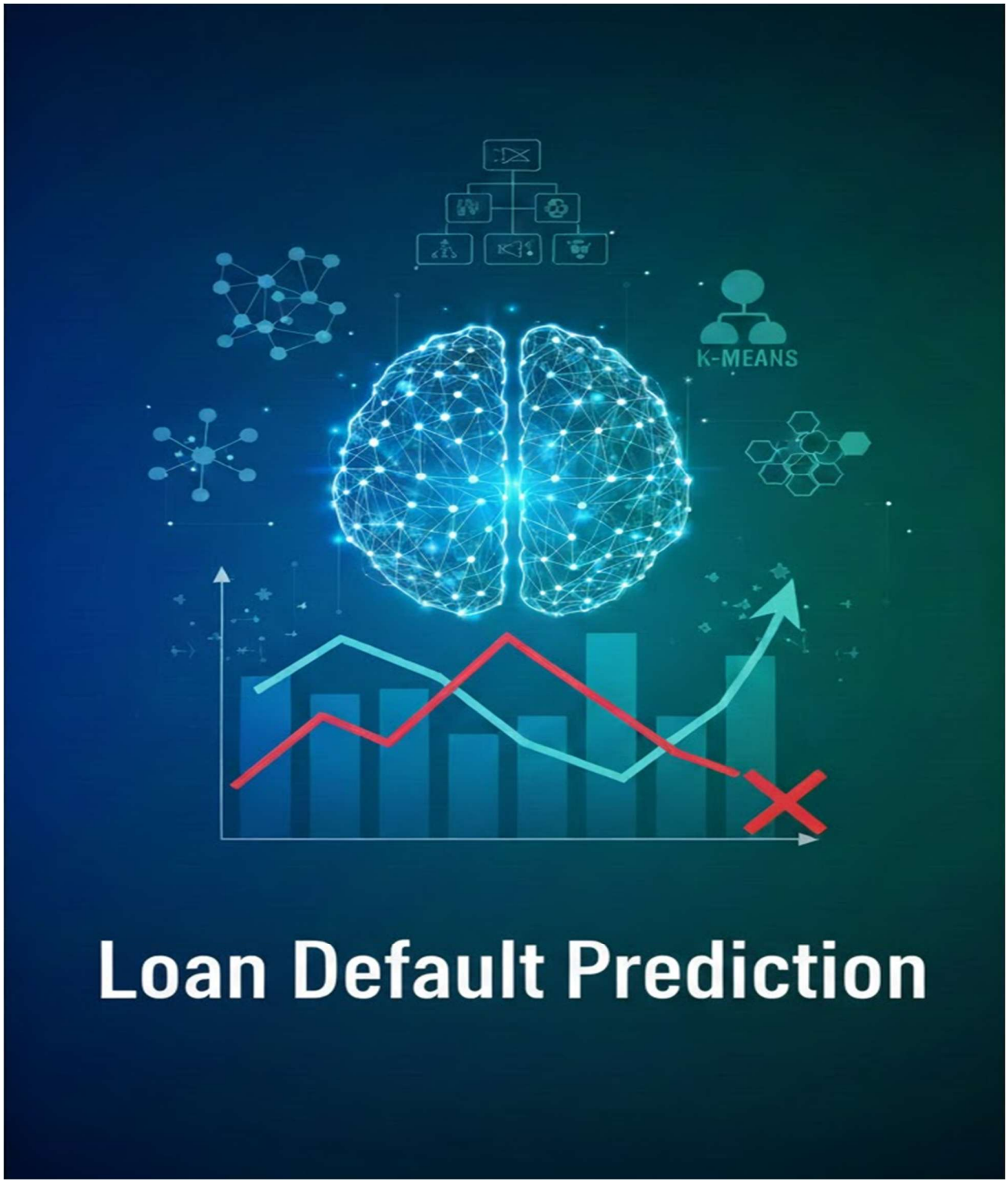

Business Report

Submitted by James Kagwe

December 01, 2025

Table of Contents

List of Figures

1.0 Executive Summary

This report summarizes the findings from the exploratory data analysis (EDA) of the Home Equity Loan dataset (HMEQ), aiming to establish a foundation for an automated credit scoring model. The analysis confirms that **historical credit behavior** is the primary driver of default risk. Specifically, the number of **Delinquencies (DELINQ)** shows the strongest positive correlation with a loan being classified as 'Bad' at **+0.35**. Conversely, the **Age of Credit Line (CLAGE)** acts as the strongest protective factor, with a negative correlation of **-0.17**.

A critical data quality issue identified is the high correlation (**0.88**) between the existing mortgage balance (MORTDUE) and the property value (VALUE), which necessitates elimination of one of the features or a feature engineering approach (creating a Loan-to-Value ratio) to stabilize the eventual machine learning model.

The insights confirm the feasibility of automating the loan approval process to improve efficiency and reduce potential human bias.

1.2.0 Introduction

1.2.1 Context

A major proportion of retail bank profit comes from interests in the form of home loans. These loans are borrowed by regular income/high-earning customers. Banks are most fearful of defaulters, as bad loans (NPA) usually eat up a major chunk of their profits. Therefore, it is important for banks to be judicious while approving loans for their customer base. The approval process for the loans is multifaceted. Through this process, the bank tries to check the creditworthiness of the applicant based on a manual study of various aspects of the application. The entire process is not only effort-intensive but also prone to wrong judgment/approval owing to human error and biases. There have been attempts by many banks to automate this process by using heuristics. But with the advent of data science and machine learning, the focus has shifted to building machines learning models that can learn this approval process and make it free of biases and more efficient. At the same time, one important thing to keep in mind is to make sure that the machine does not learn the biases that previously crept in because of the human approval process.

1.2.2 Problem Statement

A bank's consumer credit department aims to simplify the decision-making process for home equity lines of credit to be accepted. To do this, they will adopt the Equal Credit Opportunity Act's guidelines to establish an empirically derived and statistically sound model for credit scoring. The model will be based on the data obtained via the existing loan underwriting process from recent applicants who have

been given credit. The model will be built from predictive modeling techniques, but the model created must be interpretable enough to provide a justification for any adverse behavior (rejections).

1.2.3 Objective

The objective is to Build a classification model to predict clients who are likely to default on their loan and give recommendations to the bank on the important features to consider while approving a loan. Proprietary content.

1.2.4 Need of the present study

The need for this study is multifaceted, addressing critical issues related to risk management, profitability, efficiency, and regulatory compliance within the bank's credit department.

Here is a breakdown of the necessity of building an automated loan default prediction models:

a) The resultant models accurately predict which applicants are most likely to default, therefore minimizing non-performing loans. Bad loans (Non-Performing Assets) directly erode a bank's capital and profits.

b) Maximizing Profitability: By reducing defaults, the bank protects the income earned from most reliable customers, leading to a more stable and higher profit margin.

c) Judicious Lending: The study ensures that the bank is making decisions based on empirical evidence and statistical soundness, rather than guesswork, reducing overall financial risk exposure.

d) An automated model allows the bank to process a much larger volume of loan applications quickly, supporting business growth without scaling up the manual workforce proportionally.

e) Automated, instant credit scoring significantly reduces the time from application to approval, improving customer experience and giving the bank a competitive advantage.

f) The manual process is prone to wrong judgment/approval owing to human error and biases." A properly constructed machine learning model can apply the same objective criteria to every application, making the process fairer and more consistent.

g) Since the models are built to avoid Learning Historical Biases that existed in the historical (human-approved) data.

h) The models are built to be interpretable enough to provide clear, legal justification for any rejection. This ensures that every loan application rejection is clearly explained.

1.2.5 Business/social opportunities

This study presents significant opportunities across both the bank's core business operations and its broader social and ethical impact. The opportunities include:

1.2.5.1 Business Opportunities

a) The study provides the bank with a powerful competitive advantage and enhanced financial performance. It culminates into a model that allows the bank to precisely quantify the risk of each applicant. This enables sophisticated pricing—offering lower interest rates to low-risk customers and higher, risk-adjusted rates to moderate-risk customers, maximizing total returns.
b) By maintaining a lower default rate, the bank requires less capital set aside to cover potential losses (regulatory capital reserves), freeing up resources for other profitable ventures.
c) Identifying /isolating the features that are highly predictive of non-default allows the bank to focus its marketing efforts on customer segments that are statistically most likely to be high-performing, reliable borrowers.
d) Automating the underwriting process drastically reduces the operational costs associated with manual review, data collection, and human-based decision-making.
e) Near-instantaneous credit decisions improve customer experience, leading to higher conversion rates for applications and potentially attracting more prime borrowers.
f) The model standardizes the credit approval process, ensuring all branches and loan officers adhere to the same objective criteria, reducing internal errors and inconsistent judgments.
g) By identifying high-risk borrowers before they default, the bank can implement early intervention strategies (e.g., offering loan restructuring, financial counseling, or short-term relief) to prevent the loss, which is always cheaper than recovering from a default.
h) The model data can be integrated to predict the long-term value and loyalty of a customer, informing decisions on cross-selling other high-value banking products (e.g., investment accounts, insurance).

1.2.5.2 Social and Ethical Opportunities

a) By consciously building and auditing the model to meet ECOA guidelines, the bank actively reduces the influence of unconscious human biases (which may relate to protected characteristics) in lending decisions. This leads to fairer outcomes.
b) An objective, statistically driven model can safely extend credit to underserved populations whose creditworthiness might have been overlooked or misjudged by rigid, traditional manual rules, thereby promoting economic mobility.

c) The models allow the bank to provide clear, actionable reasons for rejections. This transparency builds trust with applicants, even those who are rejected, and ensures robust compliance with fair lending laws.
d) The rejection justification can educate the applicant on specific financial factors (e.g., high debt-to-income ratio or utilization rate) they need to improve to qualify in the future, providing a path toward financial health.
e) By improving the quality of its loan portfolio, the bank contributes to the overall stability of the financial system. Responsible lending practices, enabled by predictive models, help mitigate the kind of excessive risk-taking that can lead to broader financial crises.

2.0Data Report

2.1 Data Dictionary

The Home Equity dataset (HMEQ) contains baseline and loan performance information for recent home equity loans. The target (BAD) is a binary variable that indicates whether an applicant has ultimately defaulted or has been severely delinquent. There are 12 input variables registered for each applicant.

- BAD: 1 = Client defaulted on loan, 0 = loan repaid
- LOAN: Amount of loan approved
- MORTDUE: Amount due on the existing mortgage
- VALUE: Current value of the property
- REASON: Reason for the loan request (HomeImp = home improvement, DebtCon= debt consolidation which means taking out a new loan to pay off other liabilities and consumer debts)
- JOB: The type of job that loan applicant has such as manager, self, etc.
- YOJ: Years at present job
- DEROG: Number of major derogatory reports (which indicates serious delinquency or late payments).
- DELINQ: Number of delinquent credit lines (a line of credit becomes delinquent when a borrower does not make the minimum required payments 30 to 60 days past the day on which the payments were due)
- CLAGE: Age of the oldest credit line in months
- NINQ: Number of recent credit inquiries
- CLNO: Number of existing credit lines
- DEBTINC: Debt-to-income ratio (all monthly debt payments divided by gross monthly income. This number is one of the ways lenders measure a borrower's ability to manage the monthly payments to repay the money they plan to borrow)

2.2 Understand the data collection methodology

The data collection methodology for this loan default prediction study must focus on gathering comprehensive, accurate, and relevant historical data that reflects the outcomes of past lending decisions. The structured methodology for data collection includes the following:

2.1.1 Define the Data Scope and Sources

The primary goal is to collect data that encompasses the entire lifecycle of a loan application, from initial application to outcome (payment or default).

2.2.2 Time Window Definition:

a) Data should span a sufficient historical period (e.g., 5 to 7 years) to capture different economic cycles (recessions, booms) and ensure a representative sample of defaults.

b) A decision must be made on how long a loan must be active for the default status to be considered final. For home equity loans, this window is typically long (e.g., 24-36 months post-origination) to ensure a meaningful outcome has been observed.

Only loans that have had enough time to either default or become reliably current/paid off should be included in the dataset. Loans originated too recently must be excluded (right-censoring).

2.2.3 Data Extraction and Integration

a) Data is pulled from various source systems (LOS, servicing, credit bureau archives) into raw files (e.g., SQL exports, CSVs).

b) Use a unique identifier is defined (e.g., Application_ID or Loan_ID) to link all disparate records for a single applicant/loan across various systems.

c) Features with the same meaning MUST have consistent formats and definitions across all extracted sources (e.g., making sure 'income' is always annualized and in the same currency unit).

Finally, all relevant features are merged together with the target variable into a single, comprehensive table suitable for machine learning training.

2.2.4 Addressing Data Biases (Sampling).

This step is critical for model quality and fairness.

a) a) The model will primarily learn from data on applicants who were approved by the human underwriters. To understand the true risk profile of the entire applicant population, the loaning organization must ideally include a sample of rejected applications and infer their hypothetical default outcome (using statistical methods like matching propensity scores, if feasible, although this is advanced). If only approved loans are used, the model will suffer from sample selection bias.

Since the defaults are rare, the collected data will be highly imbalanced. The methodology should include the explicit plan to address this during the subsequent data preparation phase (e.g., Oversampling, Under sampling, or weight adjustment).

2.2.5 Data Privacy and Governance

a) Before the commencement of analysis, directly identifiable information (names, addresses, account numbers) must be removed or masked to comply with internal privacy policies and external regulations.
b) Features related to protected characteristics (e.g., race, religion, gender, marital status) must also be excluded from the model training dataset to prevent learning and perpetuating illegal biases, aligning with the data protection requirements.

2.2 Comment on how the data might have been collected

The data for this loan default prediction study would have been collected primarily through an integration of internal banking systems and external credit bureaus. This was achieved following a structured process to create a comprehensive historical record. Here is a description of the probable sources of the data and how it was consolidated.

2.2.1 Internal Data Sources

The financial organization would pull data from its core operational systems, linking records by a unique applicant or loan ID. These systems include:

a) The Loan Origination System (LOS). This system holds all the information provided by the applicant at the time of application, including Stated income, employment history (duration, status), asset and reserve information, and the computed Debt-to-Income (DTI) ratio. Other information in this system are the requested loan amount, term length, purpose, and the underwriter's final decision (Approved/Rejected) and the date of the decision.
b) Loan Servicing System, Once the loan is approved and funded, this system tracks its performance over time. This system stores information including payment history, date of first delinquency, cure status, and the eventual binary outcome (Default/Charge-off vs. Paid/Current). It also maintains the collateral details including the appraised value of the home and the calculated Loan-to-Value (LTV) ratio. The Loan servicing system is the crucial source for the Target Variable.

2.2.2 External Data Sources

External data, typically collected under strict authorization at the time of the application, provides a standardized view of the applicant's creditworthiness. These sources include Credit Bureau Reports which is the source of Credit Score, Credit History, Credit Utilization Rate and Delinquencies.

2.2.3 Data Integration and Sampling

The most critical part of the collection is integrating these disparate data points into a single, usable dataset. The objective of this exercise is to consolidate all parts of the data into a single Analytical Base Table (ABT) Creation Time Synchronization.

Features must be captured as they existed at the time of the application, while the target variable (Default Status) is captured after a sufficient observation period (e.g., 24-36 months post-origination). Finally, the financial organization addresses the Sample Selection Bias and Ethical Feature Exclusion, before being used for modeling.

2.3 Overview of data

2.3.1 Introduction

Here is a structured overview of the dataset:

a) The dataset contains 5,960 records and 13 features that are typical for a credit risk modeling exercise. The features can be categorized into the target variable, loan information, and applicant credit metrics.
b) The goal of the study is to predict column - BAD
c) The features that describe the requested loan and the collateral property are LOAN, MORTDUE, VALUE, REASON, JOB and YOJ
d) These variables that are crucial predictors of default are DEBTINC, DEROG, DELINQ, CLAGE, NINQ and CLNO.

2.3.2 Key Data Challenges

The primary challenge in the data preparation phase will be handling the significant volume of missing values which include the following.

i. DEBTINC is a cornerstone of credit scoring, and over 21% of the records are missing this value. Careful imputation or feature engineering is essential.
ii. MORTDUE and YOJ also have a substantial number of missing entries (~8% to 9.5%).
iii. The categorical features, REASON and JOB, will need proper encoding (e.g., One-Hot Encoding) before modeling.

2.4 Check the rows, columns, variable info, and descriptive details

2.4.1 Inspecting the Dataset Dimensions

The data has 5960 rows and 13 columns

	BAD	LOAN	MORTDUE	VALUE	REASON	JOB	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC

0	1	1100	25860	39025	HomeImp	Other	11	0	0	94.367	1	9	NaN
1	1	1300	70053	68400	HomeImp	Other	7	0	2	121.83	0	14	NaN
2	1	1500	13500	16700	HomeImp	Other	4	0	0	149.47	1	10	NaN
3	1	1500	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	0	1700	97800	112000	HomeImp	Office	3	0	0	93.333	0	14	NaN

Figure 1: A Table showing the first 5 rows of the dataset

2.4.2 Inspecting the Missing values

All columns except BAD (Target) and LOAN (Loan Amount) have missing values, which will need to be addressed during data cleaning. The DEBTINC column has the highest number of missing values (only 4,693 non-null out of 5,960).

The DEBTINC (Debt-to-Income Ratio) feature is missing in over 21% of the records, which is a major issue as this is a highly critical variable in credit risk modeling. All other variables have less than 12% missing data.

High Missingness in Critical Feature: The DEBTINC (Debt-to-Income ratio) is a vital metric in lending but is missing in over 21% of the data, which must be carefully addressed through imputation.

Variable	Null Count	Null Percentage
DEBTINC	1267	21.26
DEROG	708	11.88
DELINQ	580	9.73
MORTDUE	518	8.69
YOJ	515	8.64
NINQ	510	8.56
CLAGE	308	5.17
JOB	279	4.68
REASON	252	4.23
CLNO	222	3.72
VALUE	112	1.88

Figure 2: A Table Showing Missing values and the Percentages

2.4.3 Duplicate Data Check

The dataset has no duplicate values

2.4.4 Variable info

The dataset has 13 columns. 9 columns are of float64 data types, 2 are of data type int64 while 2 are of object data type.

#	Column	Non-Null Count	Dtype
0	BAD	5960	int64
1	LOAN	5960	int64
2	MORTDUE	5442	float64
3	VALUE	5848	float64
4	REASON	5708	object
5	JOB	5681	object
6	YOJ	5445	float64
7	DEROG	5252	float64
8	DELINQ	5380	float64
9	CLAGE	5652	float64
10	NINQ	5450	float64
11	CLNO	5738	float64
12	DEBTINC	4693	float64

Figure 3: A Table showing the Dataset information

2.4.5 descriptive details

1. Descriptive Statistics (Numerical Variables)

	count	mean	std	min	25%	50%	75%	max
BAD	5960	0.199497	0.399656	0	0	0	0	1
LOAN	5960	18608	11207.5	1100	11100	16300	23300	89900
MORTDUE	5442	73760.8	44457.6	2063	46276	65019	91488	399550
VALUE	5848	101776	57385.8	8000	66075.5	89235.5	119824	855909
YOJ	5445	8.92227	7.57398	0	3	7	13	41
DEROG	5252	0.25457	0.846047	0	0	0	0	10
DELINQ	5380	0.449442	1.12727	0	0	0	0	15
CLAGE	5652	179.766	85.8101	0	115.117	173.467	231.562	1168.23
NINQ	5450	1.18606	1.72867	0	0	1	2	17
CLNO	5738	21.2961	10.1389	0	15	20	26	71
DEBTINC	4693	33.7799	8.60175	0.524499	29.14	34.8183	39.0031	203.312

2. Descriptive Statistics (Numerical Variables)

	count	unique	top	freq
REASON	5708	2	DebtCon	3928
JOB	5681	6	Other	2388

3. Key Finding: Target Variable (BAD) Class Distribution

BAD	Count	Percentage
0	4771	80.05
1	1189	19.95

Figure 4: Tables Showing Descriptive Statistics for Numerical and Categorical Variables and the Distribution of the Target Class

2.4.6 Findings

Below are the finds of the analysis:

a) **Target Variable (BAD):** The mean of **0.199** indicates that approximately **20%** of the loans in the dataset resulted in a default. This suggests a class imbalance where non-defaulting loans are more common.

b) **Loan Amount (LOAN):** The average loan is about **18,608**, ranging from **1,100 to 89,900.**

c) **Debt-to-Income (DEBTINC):** The average Debt-to-Income ratio is about **33.8%.** The max value of **203.3%** is an extreme **outlier** that needs investigation, as a value over **100%** suggests the borrower's debt obligations exceed their total income.

d) **Credit History (DEROG, DELINQ):** The **75th** percentile for both DEROG (major derogatory reports) and DELINQ (delinquent credit lines) is **0**, meaning most borrowers have no or very few past derogatory or delinquent records. The max values (10 and 15) indicate a few individuals with very poor credit history.

3.0 Data pre-processing

3.1 Removal of unwanted variables

The variable removal and transformation process was crucial for ensuring model stability and optimizing predictive performance, primarily addressing the following issues identified during the Exploratory Data Analysis (EDA):

A very high positive correlation exists MORTDUE and VALUE. Including both in certain models (like Logistic Regression) can lead to unstable coefficient estimates and make the model less interpretable. When the modeling process begins, we shall consider eliminating one of them using VIF strategy or transforming them into a single, highly relevant financial metric e.g the Loan-to-Value (LTV) ratio.

Several key variables, notably DEBTINC, MORTDUE and VALUE had a substantial number of missing values (NaNs). When a variable/ feature has very many missing values, imputation becomes unreliable and the variable is generally considered "unwanted" and removed. See table below containing details of missing values in the dataset.

Missing Value Report		
	Null Count	**Null Percentage**
DEBTINC	1267	21.26
DEROG	708	11.88
DELINQ	580	9.73
MORTDUE	518	8.69
YOJ	515	8.64
NINQ	510	8.56
CLAGE	308	5.17
JOB	279	4.68
REASON	252	4.23
CLNO	222	3.72
VALUE	112	1.88

Figure 5: Missing Values Report

From the table DEBTINC had the highest missing values with 21.26%, followed by DEROG with 11.88% then DELINQ, MORTDUE and YOJ following thereafter. Since the no variable had an unacceptably high proportion of missing values, the missing values were imputed. The numerical variables were imputed with the median while the categorical variables were imputed with the mode.

LOAN amount showed a very weak correlation with loan default. While it's not removed just for being weak, it might be a candidate for exclusion in final feature selection. In advanced models like Gradient Boosting or Random Forest, variables with low linear correlation are often retained initially, as these models can capture complex, non-linear predictive relationships that simple correlation misses. The removal decision is typically deferred to a Model-Based Feature Selection step (e.g., using Feature Importance scores) after the model has been trained.

3.2 Missing value treatment

The treatment of missing values was a crucial step in the data preparation pipeline, as several key financial and credit history variables contained null values. The primary goal was to handle these gaps without introducing significant bias or distorting the original data distributions. The missing data treatment was categorized based on the variable type:

3.2.1 Numerical Variable Imputation

Variables such as MORTDUE, VALUE, YOJ (Years at present job), and DEBTINC had missing entries. The median was used because financial data is often heavily skewed and susceptible to outliers. Using the median preserves the order statistics and is more robust than using the mean, which can be pulled by extreme values. Imputing with a central tendency ensures the model can utilize the feature without sacrificing the entire row.

3.2.2 Categorical Variable Imputation

Categorical variables such as REASON and JOB were also found to have missing entries. Since the missing proportion was small, mode imputation (using the most frequent category) was used for simplicity.

3.3 Outlier treatment

The treatment of outliers is a critical aspect of preparing financial data, as variables related to income, debt, and property values often exhibit extreme skewness and contain genuine, but rare, high-magnitude observations.

3.3.1 Outlier Identification Methods

Outliers were identified using a combination of statistical and visual methods:

a) **Visualization: Box plots** and **histograms** to visually inspect the distribution of key variables.

b) **Interquartile Range (IQR) Method:** Computing the IQR (Q3 - Q1) to flag values lying beyond **Q3+1.5×IQR or Q1−1.5×IQR.** The Interquartile Range (IQR) method was applied to all ten numerical variables in the dataset to identify potential outliers

Variable	Q1	Q3	IQR	Lower Bound	Upper Bound	Total Non-Missing	Outlier Count	Outlier Percentage
DELINQ	0	0	0	0	0	5960	1201	20.15%
DEROG	0	0	0	0	0	5960	725	12.16%
VALUE	66489.5	119005	52515.2	-12283.4	197778	5960	347	5.82%
MORTDUE	48139	88200.2	40061.2	-11952.9	148292	5960	308	5.17%
LOAN	11100	23300	12200	-7200	41600	5960	256	4.30%
DEBTINC	30.76	37.95	7.19	19.98	48.73	5960	247	4.14%
CLNO	15	26	11	-1.5	42.5	5960	219	3.67%
YOJ	3	12	9	-10.5	25.5	5960	211	3.54%
NINQ	0	2	2	-3	5	5960	177	2.97%
CLAGE	117.37	227.14	109.77	-47.29	391.8	5960	66	1.11%

Figure 6: Table showing Outlier Detection using IQR Method

The boxplots below demonstrate outlier Detection using the Visualization Method

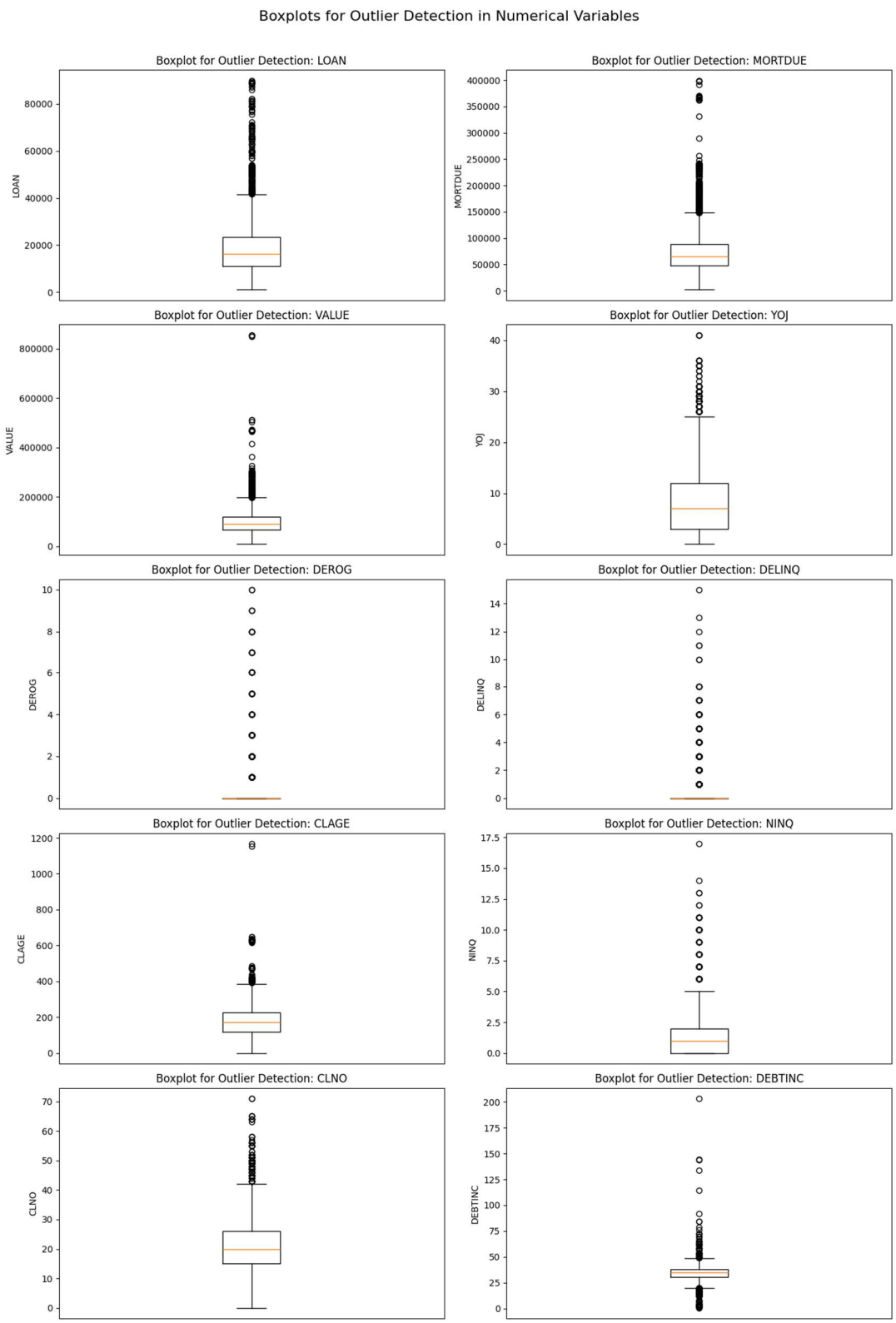

Figure 7: Boxplots showing Outliers detection using the Visualization Method

3.3.2 Treatment Strategy: Transformation vs. Capping

The approach to outlier treatment is contingent on the downstream model but generally involves minimizing their distorting effect without losing the valuable information they represent. The treatment strategy used was **Winsorization (Capping/Flooring)**. The values above and below **1.5*IQR** were set to the upper bound and lower bound respectively**.**

Variables with severe positive skewness, such as VALUE and LOAN, may be subjected to a **logarithmic transformation** during the modeling phase. This approach compresses the range of high values, making the distribution more symmetrical (closer to Gaussian) and reducing the leverage of extreme outliers on linear models.

3.4 Variable transformation

Variable transformation is a fundamental and mandatory step in preparing the loan default data for modeling, ensuring features meet the assumptions of certain algorithms and maximizing predictive power. The process involves structural changes to variables based on their distribution, predictive relationships, and data type. Based on the nature of the variable the following was done or will be considered during modelling.

3.4.1 Handling Numerical Skewness and Outliers

The **VALUE** and **LOAN** variables were noted to be positively skewed, meaning their distributions have long tails towards high values and were confirmed to contain significant outliers in the IQR analysis. Logarithmic Transformation may be considered during the modeling phase. Applying the logarithm transformation compresses the range of high-magnitude values, stabilizing the variance, and making the distribution more symmetrical (closer to a Gaussian shape). This is particularly important if Logistic Regression is used, as linear models assume features have a normal or near-normal distribution.

3.4.2 Feature Engineering

This may also be considered for **MORTDUE** and **VALUE** variables during the modelling phase. The variables were noted to have a high correlation of 0.876, between them. This suggests multicollinearity or redundancy. Transformation will create powerful, domain-specific features that address multicollinearity and improve model interpretability. A ratio variable may be created to address this issue.

3.4.3 Encoding Categorical Variables

The categorical features **REASON** and **JOB** required transformation into a numerical format suitable for machine learning algorithms. The transformation used was **One-Hot Encoding (OHE)** or **Target/Mean Encoding**. OHE creates binary (0 or 1) columns for each unique category, preventing the model from inferring an ordinal relationship (e.g., assuming 'HomeImp' is greater than 'DebtCon'). The table below shows the results of **One-Hot Encoding (OHE)** to the **REASON** and **JOB** variables**.**

BAD	LOAN	DEBTINC	REASON_HomeImp	JOB_Office	JOB_Other	JOB_ProfExe	JOB_Sales	JOB_Self
1	1100	34.8183	1	0	1	0	0	0
1	1300	34.8183	1	0	1	0	0	0
1	1500	34.8183	1	0	1	0	0	0
1	1500	34.8183	0	0	1	0	0	0
0	1700	34.8183	1	1	0	0	0	0

Figure 8: A table showing the results of OHE to the REASON and JOB Variables

3.5 Addition of new variables

Based on the analysis so far conducted, the primary goal of adding new variables would be to:

a) **Increase Predictive Power:** Create features that capture non-linear risk aspects.
b) **Solve Multicollinearity:** Combine correlated raw features into a single, more meaningful ratio.

During modelling the following may be considered:

a) Variable addition which may focus on creating financial ratios, which are standard in credit risk assessment and often possess higher predictive power than their raw components. Some of the variables based on ratios to be considered are Loan-to-Value (LTV) Ratio((LOAN+MORGDUE)/VALUE), Asset Value (Home Equity) (VALUE/MORGDUE) and Debt-to-Credit Limit Ratio (LOAN/CLNO)
b) Interaction Features may also be included. These variables are added to capture non-linear risk profiles where the effect of one feature depends on the level of another. An example is (Job Tenure vs. Credit Tenure): A ratio that assesses if the stability of the current job is proportional to the length of the credit history.
c) Binning and Discretization may also be used. This transformation converts highly continuous variables into discrete, ordinal groups, which helps capture non-linear relationships. An example is Age of Credit History (CLAGE). This could be binned (e.g., < 5 years, $5-10$ years, $10+$ years) to isolate specific segments where credit stability drastically changes. This is

effective if the relationship between CLAGE and BAD is not smooth but drops sharply after a certain credit age threshold.

4.0Exploratory Data Analysis

4.1Univariate Analysis - distribution and spread for continuous and categorical variables

Univariate analysis examines only one variable at a time. The goal is to describe, summarize, and find patterns within the single variable's data. It is a foundational step in data analysis, providing the initial, crucial understanding of each variable before moving on to more complex, multi-variable investigations.

4.1.1 Categorical Variables

The distribution of categorical variables was computed accordingly.

◇	**REASON**		
		Count	**Percentage**
	DebtCon	4180	70.13
	HomeImp	1780	29.87
◇	**JOB**		
		Count	**Percentage**
	Other	2667	44.75
	ProfExe	1276	21.41
	Office	948	15.91
	Mgr	767	12.87
	Self	193	3.24
	Sales	109	1.83

Figure 9: A Table Showing Distribution of Categorical Variables

4.1.1.1 Findings:

The resulting table illustrates the following distributions:

Loan Reason (REASON)

a) Debt Consolidation (DebtCon) is the dominant category with **3,928** counts, constituting 70.13**%**.
b) Home Improvement (HomeImp) is the only other category, with 1,780 counts.

The dataset is heavily skewed toward debt consolidation loans, meaning the predictive model will primarily learn the risk profile associated with this type of loan purpose.

Applicant Job Category (JOB)

a) The **Other** category is the largest with 2,388 counts. This constitutes **44.75%** of the data. Since Other is an ambiguous category, its high frequency is notable.

b) **ProfExe** (Professional/Executive) is the second largest with 1,276 counts. which is **21.41**% of the dataset

c) The remaining categories, including Office, Mgr (Manager), Self (Self-Employed), and Sales, are significantly smaller.

The large proportion of applicants in the general "Other" job category suggests this variable's predictive power might be limited or that a significant portion of the applicant pool belongs to non-standard or varied employment types.

4.1.2 Numerical Variables

4.1.2.1 Descriptive Statistics (Numerical Variables)

	Count	Mean	Std	Min	25%	50%	75%	Max
BAD	5960	0.199497	0.399656	0	0	0	0	1
LOAN	5960	18608	11207.5	1100	11100	16300	23300	89900
MORTDUE	5442	73760.8	44457.6	2063	46276	65019	91488	399550
VALUE	5848	101776	57385.8	8000	66075.5	89235.5	119824	855909
YOJ	5445	8.92227	7.57398	0	3	7	13	41
DEROG	5252	0.25457	0.846047	0	0	0	0	10
DELINQ	5380	0.449442	1.12727	0	0	0	0	15
CLAGE	5652	179.766	85.8101	0	115.117	173.467	231.562	1168.23
NINQ	5450	1.18606	1.72867	0	0	1	2	17
CLNO	5738	21.2961	10.1389	0	15	20	26	71
DEBTINC	4693	33.7799	8.60175	0.524499	29.14	34.8183	39.0031	203.312

Figure 10: Table Showing Descriptive Analytics for Numerical Variables

Variable	Q1	Q3	IQR	Lower Bound	Upper Bound	Total Non-missing	Outlier	Outlier %
DELINQ	0	0	0	0	0	5960	1201	20.15%
DEROG	0	0	0	0	0	5960	725	12.16%
VALUE	66489.5	119005	52515.2	-12283.4	197778	5960	347	5.82%
MORTDUE	48139	88200.2	40061.2	-11952.9	148292	5960	308	5.17%
LOAN	11100	23300	12200	-7200	41600	5960	256	4.30%
DEBTINC	30.76	37.95	7.19	19.98	48.73	5960	247	4.14%
CLNO	15	26	11	-1.5	42.5	5960	219	3.67%

YOJ	3	12	9	-10.5	25.5	5960	211	3.54%
NINQ	0	2	2	-3	5	5960	177	2.97%
CLAGE	117.37	227.14	109.77	-47.29	391.8	5960	66	1.11%

Figure 11: Table Showing Interquartile Range and Composition of Outliers

4.1.2.2 Findings:

a) (DEROG, DELINQ, NINQ) - The (interquartile range) are compressed near zero, with long tails of data points extending far above. Most applicants have clean credit, but extreme outliers exist representing applicants with high delinquencies or derogatory marks. These may require capping or transformation to prevent them from skewing the model.

b) (LOAN, MORTDUE, VALUE) - Show many outliers in the high-end of their range. There are applicants with loan amounts, mortgage amounts, and property values significantly higher than the median applicant.

c) DEBTINC - The central box is tight, but it has several extreme outliers stretching to very high values (e.g., ~203). These extreme Debt-to-Income ratios should be treated as potential outliers that need to be capped or investigated, as they are likely influential data points.

d) (YOJ, CLAGE) - Both show strong evidence of outliers, particularly at the high end. CLAGE (age of credit line) has points suggesting very long credit histories, and YOJ (years at job) has high values, indicating highly stable applicants, though some high values may warrant checking for data entry errors.

The prevalence of outliers across all numerical features mandates the use of robust data processing techniques, such as log transformation for highly skewed variables and/or capping (winsorization) to limit the influence of extreme values, before training the predictive model.

4.2 Bivariate Analysis - relationships between different variables

4.2.1 Categorical Variables

The distribution of categorical variables was computed accordingly.

REASON		Count	Percentage
	('DebtCon', 0)	3387	81.03
	('DebtCon', 1)	793	18.97
	('HomeImp', 0)	1384	77.75
	('HomeImp', 1)	396	22.25

JOB		**Count**	**Percentage**
	('Mgr', 0)	588	76.66
	('Mgr', 1)	179	23.34
	('Office', 0)	823	86.81
	('Office', 1)	125	13.19
	('Other', 0)	2090	78.37
	('Other', 1)	577	21.63
	('ProfExe', 0)	1064	83.39
	('ProfExe', 1)	212	16.61
	('Sales', 0)	71	65.14
	('Sales', 1)	38	34.86
	('Self', 0)	135	69.95
	('Self', 1)	58	30.05

Figure 12: A table showing Distribution of Default on the Categorical Variables

4.2.1.1 Findings

a) Loans for Home Improvement (HomeImp) have a slightly higher default rate (approx 22.2%) compared to loans for Debt Consolidation (DebtCon) (approx 19%). This suggests that the purpose of the loan is a differentiating factor in credit risk.

b) Default rates show a strong relationship with occupation.Sales and Self-employed (Self) workers show the highest default rates (approximately 34.9% and 30.5%, respectively). This may indicate greater income volatility for these groups.Office and Professional/Executive (ProfExe) roles show the lowest default rates (approximately 13.2% and 16.6%).

4.2.2 Numerical Variables

The bivariate analysis for numerical variables compares the average values of key financial and credit history features for loans that did not default (BAD=0) and loans that defaulted (BAD=1).

Variable	Non-Default (BAD=0)	Default (BAD=1)
LOAN	19028.1	16922.1
MORTDUE	74829.2	69460.5
VALUE	102596	98172.8
YOJ	9.15494	8.0278
CLAGE	187.002	150.19
DELINQ	0.245133	1.22919
NINQ	1.03275	1.78276
CLNO	21.317	21.2113

Figure 13: A Table showing Mean Values of Key Numerical Variables by Loan Status (BAD)

4.2.2.1 Findings

The results reveal several significant differences between the two groups, which can be summarized into three main categories:

Credit History Health (Risk Factors)

Variables directly related to poor credit health are significantly higher for defaulters. These include

a) DELINQ (Number of Delinquent Credit Lines) where the mean number of delinquencies is nearly five times higher for defaulters (approx **1.23**) than for non-defaulters (approx **0.25**). This is the strongest indicator of the group differences.
b) NINQ (Number of Recent Credit Inquiries) where the Defaulters have a higher average number of recent inquiries (approx 1.78) compared to non-defaulters (approx 1.03). More inquiries often suggest a greater recent need for credit, which can be a sign of financial distress.
c) CLAGE (Age of Oldest Credit Line) where the Defaulters have a shorter credit history (mean approx 15$ months or 12.5 years) compared to non-defaulters (mean approx 187 months or 15.6 years). A shorter history means less demonstrated ability to manage credit over time.

Loan and Property Value

a) Defaulters tend to have slightly lower financial metrics related to property and loan size. LOAN (Loan Amount) where non-defaulters take out slightly larger loans on average (approx 19,028) than defaulters (approx 16,922). This could imply that lenders were more cautious when approving large loans to higher-risk applicants.
b) MORTDUE and VALUE: The mean amount due on the existing mortgage (MORTDUE) and the property value (VALUE) are slightly lower for defaulters than for non-defaulters.

Job Stability

a) YOJ (Years at Present Job): Defaulters report less job stability (mean ≈ 8.0 years) compared to non-defaulters (mean approx 9.2 years). Higher job tenure often correlates with more stable income and lower risk.
b) CLNO (Number of Credit Lines): The average number of credit lines is almost identical between the two groups (around 21.3 and 21.2), suggesting this variable is not a strong differentiator of default risk.

4.2b Multivariate Analysis

	BAD	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
BAD	1	-0.08	-0.048	-0.03	-0.06	0.276	0.354	-0.17	0.175	-0	0.2
LOAN	-0.08	1	0.229	0.335	0.106	-0.001	-0.035	0.089	0.044	0.073	0.085
MORTDUE	-0.05	0.229	1	0.876	-0.09	-0.05	-0.001	0.14	0.031	0.324	0.155
VALUE	-0.03	0.335	0.876	1	0.008	-0.049	-0.014	0.171	-0	0.269	0.132
YOJ	-0.06	0.106	-0.088	0.008	1	-0.066	0.038	0.202	-0.07	0.025	-0.056
DEROG	0.276	-0	-0.05	-0.05	-0.07	1	0.212	-0.08	0.174	0.062	0.017
DELINQ	0.354	-0.04	-0.001	-0.01	0.038	0.212	1	0.022	0.068	0.165	0.052
CLAGE	-0.17	0.089	0.14	0.171	0.202	-0.083	0.022	1	-0.12	0.238	-0.046
NINQ	0.175	0.044	0.031	-0	-0.07	0.174	0.068	-0.12	1	0.088	0.141
CLNO	-0	0.073	0.324	0.269	0.025	0.062	0.165	0.238	0.088	1	0.186
DEBTINC	0.2	0.085	0.155	0.132	-0.06	0.017	0.052	-0.05	0.141	0.186	1

Figure 14: A Table Showing Multivariate Analysis Results

4.2c A Drill into the MORTDUE and VALUE Relationship

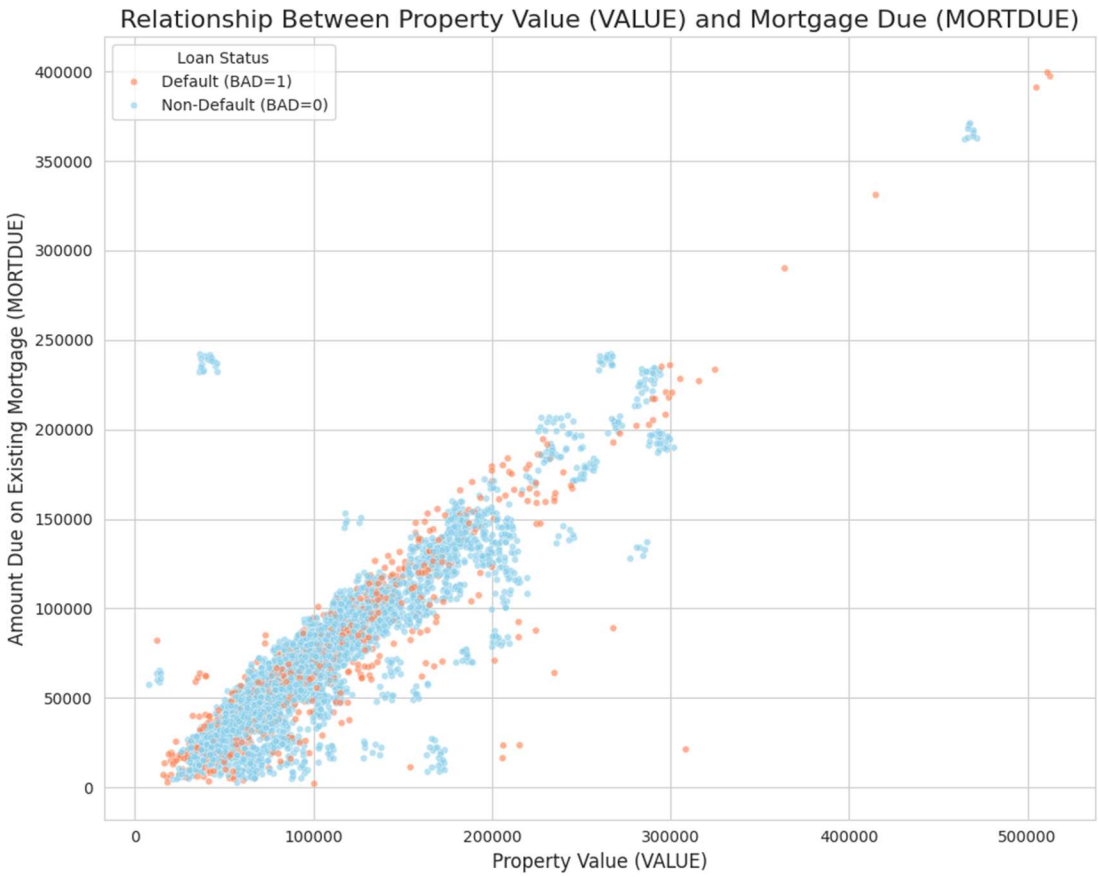

Figure 15: A Scatter Plot of MORTDUE Versus VALUE

The scatter plot clearly visualizes the strong relationship between these two financial variables and offers insights into risk distribution:

a) The points cluster tightly along a diagonal line, visually confirming the high **positive correlation** noted in the heatmap. As the value of the property increases, the amount owed on the mortgage generally increases as well, which is an expected relationship in stable real estate markets.

b) Most of the **Defaulters (coral points)** are concentrated in the **lower-left quadrant** of the plot. This area represents properties with both **lower overall VALUE** and **lower MORTDUE** amounts (i.e., less expensive properties).

c) **Non-Defaulters (skyblue points)** are more evenly distributed but dominate the **upper-right region**, which corresponds to **higher property values and higher mortgage balances**. While high MORTDUE and VALUE amounts are common, they are predominantly associated with non-defaulting loans.

d) Default risk is much higher among borrowers in the **lower end of the property market**. This indicates that lower property value might correlate with lower equity, lower overall income, or other financial constraints that increase the likelihood of default, aligning with the low mean VALUE observed in the previous bivariate analysis for defaulters.

4.3 Insightful visualizations

4.3.1 Univariate Visualizations for Categorical Variables

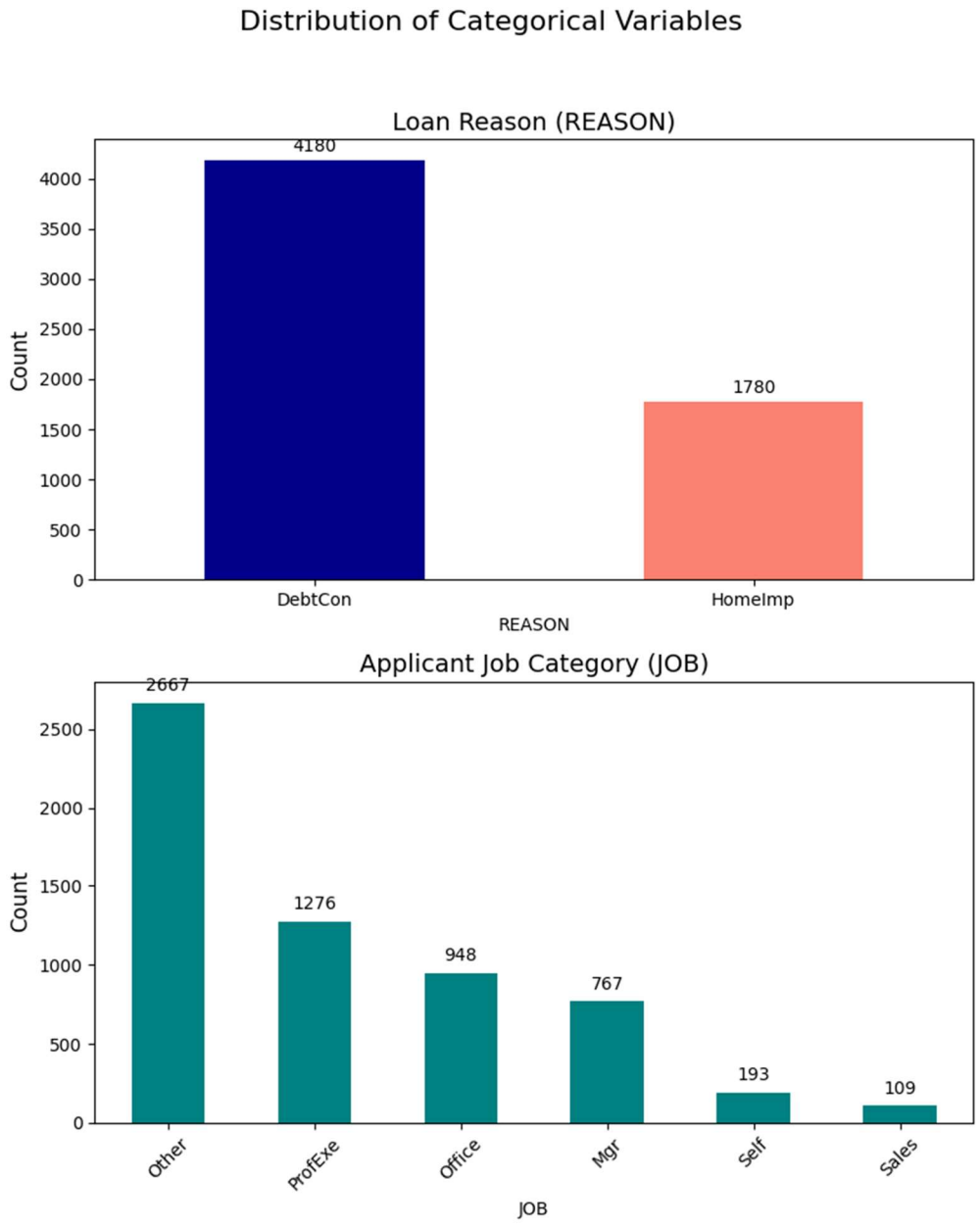

Figure 16: Univariate Visualizations for Categorical Variables

4.3.2 Univariate Visualizations for Numerical Variables

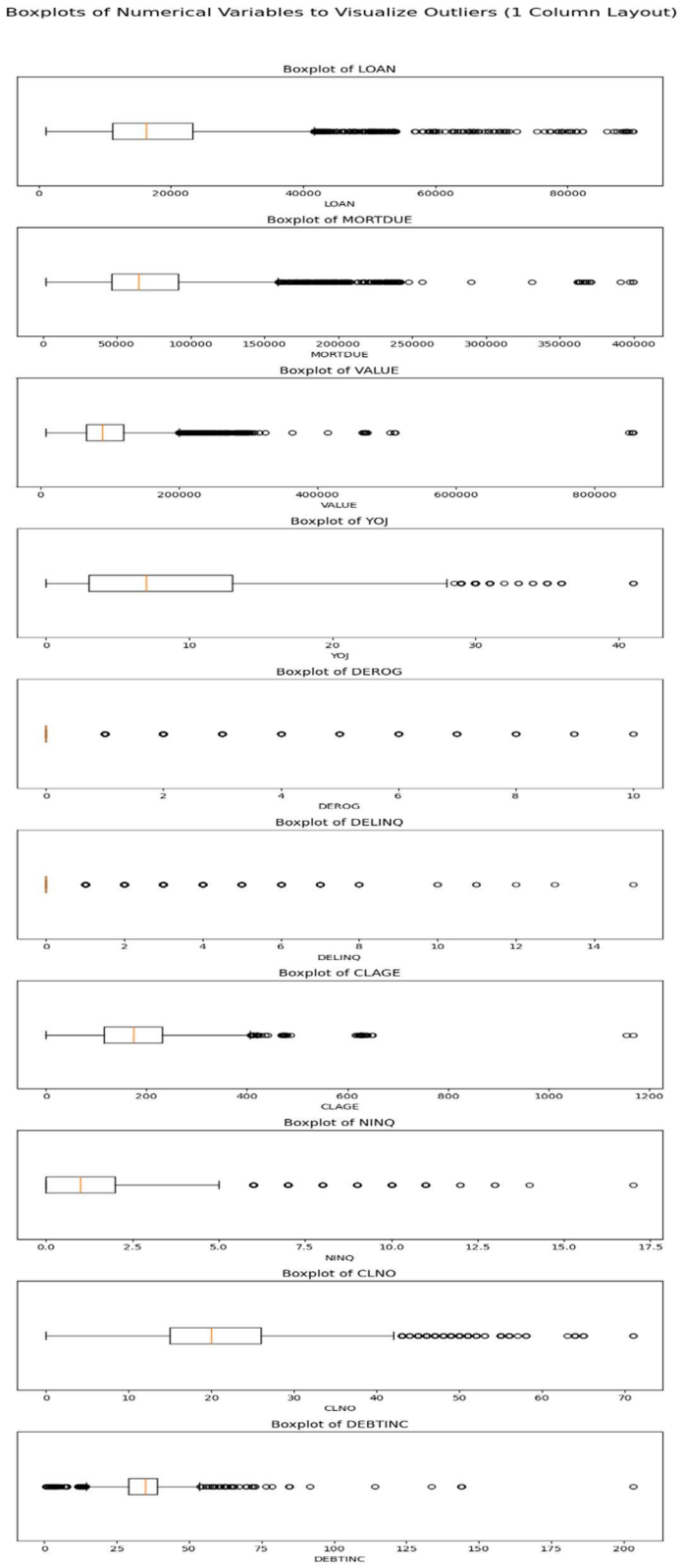

Figure 17: Univariate Visualization for Numerical Variables Before Treatment of Outliers

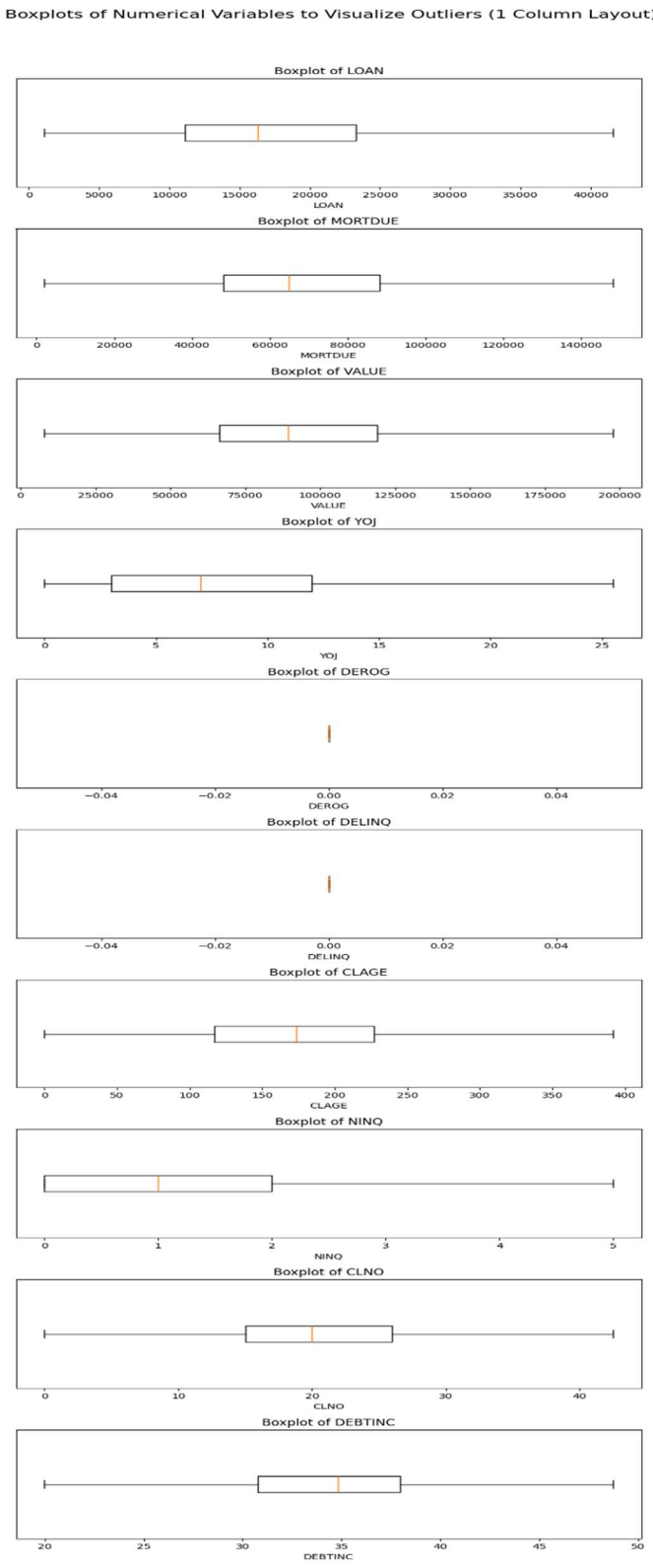

Figure 18: Figure 15: Univariate Visualization for Numerical Variables After Treatment of Outliers

4.3.3 Bivariate Visualizations for Categorical Variables

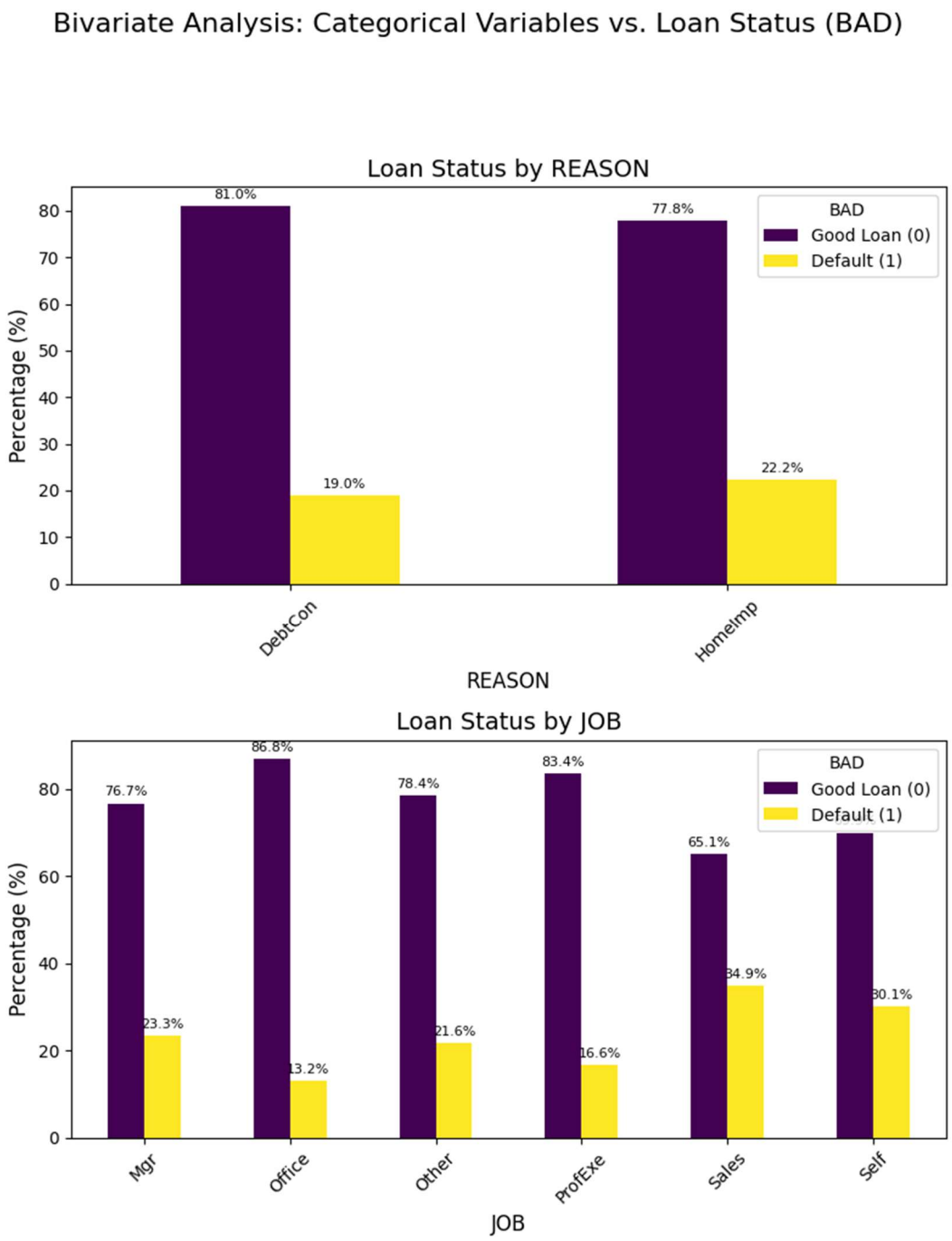

Figure 19: Visualization of Default Versus the Categorical Variables

4.3.4 Bivariate Visualizations for Numerical Variables

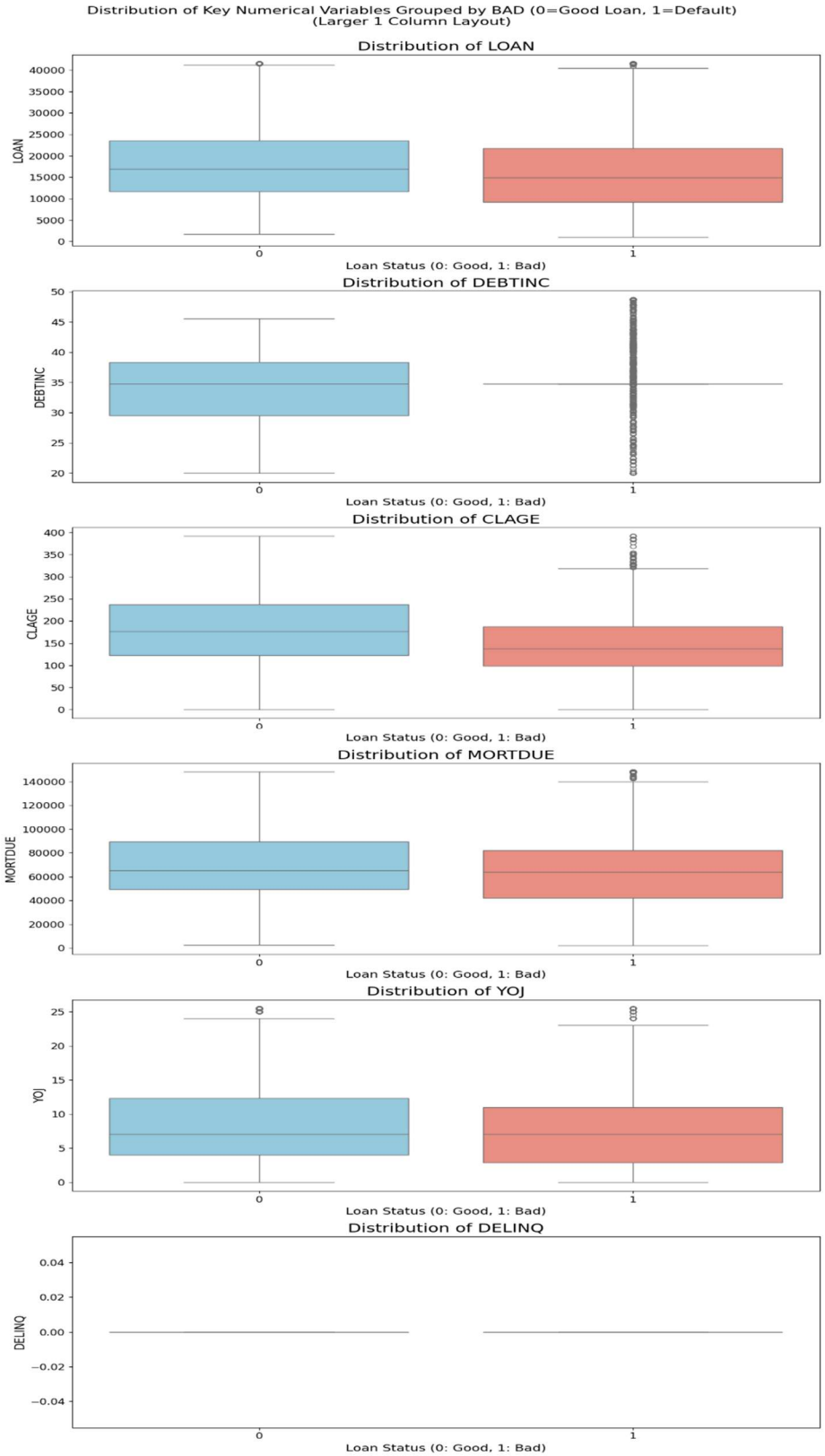

Figure 20: Bivariate Visualization for Numerical Variables

4.3.5 Multivariate Visualizations for Numerical Variables

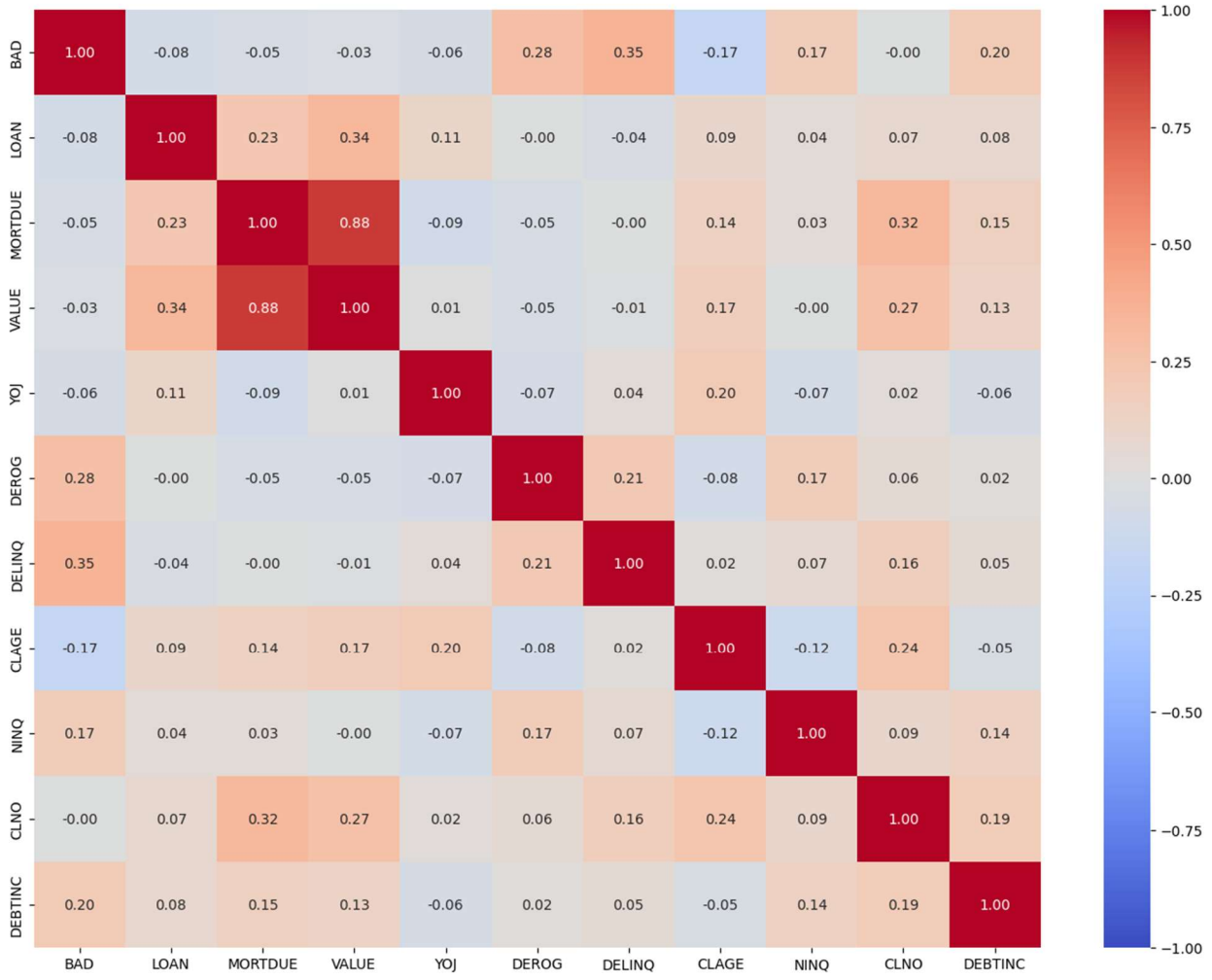

Figure 21: A Heat Map of the Correlation Between Numerical Variables

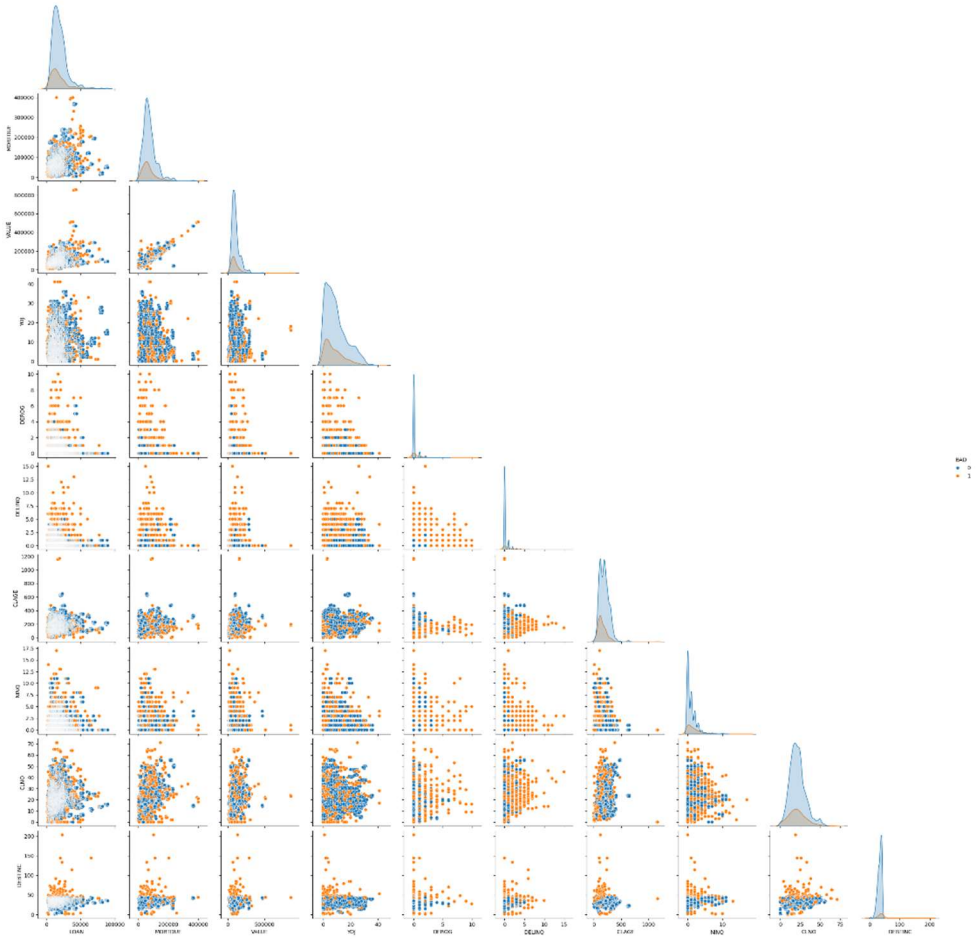

Figure 22: Pair Plots for Numerical Variables

4.3.5 Summary of Findings from EDA

The analysis reveals critical insights regarding the loan applicant pool, credit characteristics, and necessary strategies for building a robust loan default prediction model.

1. Key financial features such as **Loan Amount (LOAN)** and **Years at Present Job (YOJ)** are significantly **right-skewed**. This indicates that most loans are smaller, and most applicants have relatively few years of work experience. To normalize these distributions and minimize the influence of extreme outliers (which were noted in the DEBTINC. **Transformations** such as log or square root are recommended prior to modeling.
2. Most applicants seek loans for **debt consolidation**. The most frequent occupational group is the ambiguous **"Other" job** category, which should be assessed for risk and potentially re-categorized during feature engineering.

3. A significant **positive correlation** exists between the **property value (VALUE)** and the **amount due on the existing mortgage (MORTDUE)**, which is expected and may lead to multicollinearity if both are used directly in a linear model.
4. While **Loan Amount** itself shows **no significant difference** between defaulters and non-defaulters, specific applicant characteristics strongly delineate risk. **Sales** and **Self** job categories exhibit **higher default rates** compared to salaried or professional positions. These categorical distinctions are strong indicators of risk and should be prioritized in model development.
5. The overall data structure points toward a complex classification challenge. The correlations between the target variable BAD and most **single numerical features are relatively low**. This suggests that no one feature alone has a strong linear impact on predicting loan default.
6. The problem is **not easily separable** in its current feature space. Building a high-performing predictive model will require moving beyond simple linear methods.
7. The focus must shift to leveraging **complex algorithms** (e.g., Gradient Boosting, Random Forests) and extensive **feature engineering** to capture **non-linear relationships** and **interactions between variables**. Exploring these complex relationships is essential to boost predictive performance.

5.0Analytical Approach

5.1 Mention the alternative analytical approaches that you may see fit to be applied to the problem

Based on findings so far from the preliminary data analysis, here are several alternative analytical approaches that can be applied to build a more robust and insightful model.

a) The initial analysis focused on linear correlations. Later a baseline Logistic Regression model will be generated to gain more understanding of the data. Since simple, Logistic Regression may miss complex, non-linear relationships in the data. Other models that will be used include:
b) Tree-Based Ensemble Models, including Random Forest, XGBoost, or LightGBM (Gradient Boosting Machines) will be used. These models inherently capture non-linear relationships and feature interactions without the need for extensive manual feature engineering. XGBoost/LightGBM are often the state-of-the-art for structured data and can significantly outperform Logistic Regression.
c) Deep Learning may also be considered. A simple Multi-Layer Perceptron (MLP) or a feed-forward neural network can learn intricate patterns and interactions. The small size of the dataset may however limit the use of this approach.
d) Since Loan defaults is typically a rare event, the dataset is likely imbalanced. Building a model on imbalanced data can lead to prediction of the majority class (no default) and performs poorly on identifying actual defaulters. SMOTE (Synthetic Minority Over-sampling Technique) or its variations (e.g., Borderline SMOTE) will be used to synthetically generate more samples of the minority class and ensure the data used to generate the model is balanced.
e) Multicollinearity between MORTDUE and VALUE was identified. To address this issue creating new features will be considered. Such an approach can significantly improve model performance. Principal Component Analysis (PCA) on the collinear features MORTDUE and VALUE will also be considered.
f) Regarding Encoding Categorical Features One-Hot encoding has been used to code the data. Target encoding will be considered where a category will be replaced with the mean of the target variable for that category, potentially capturing more predictive information. This requires careful cross-validation to prevent data leakage.