BANK LOAN CASE STUDY

- ISWARIYA S

Agenda

01 Project Description

02 Approach

03 Tech-Stack Used

04 Insights

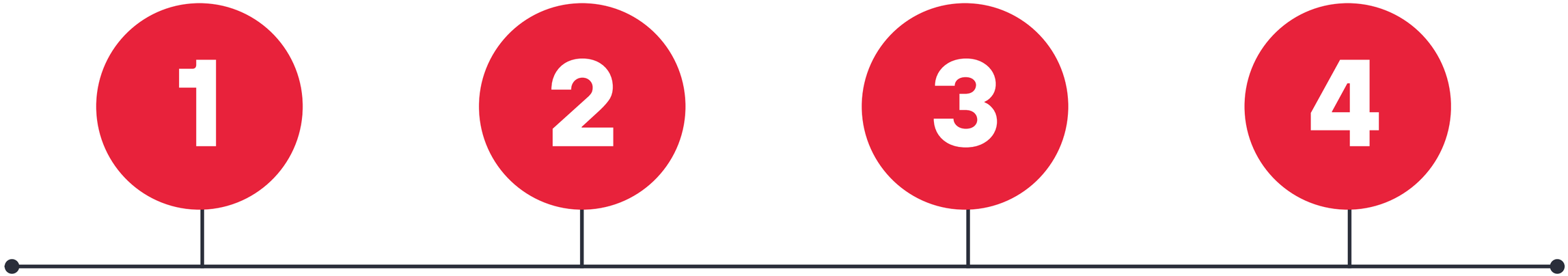
05 Result

06 Conclusion

INTRODUCTION

This project immerses participants in the practical application of Exploratory Data Analysis (EDA) within the banking sector, offering a hands-on exploration of risk analytics. It goes beyond conventional EDA methods to illuminate the intricate factors influencing lending decisions. By examining variables such as credit scores and income levels, participants gain insights into how data analysis reduces the risk of financial losses in consumer lending. The case study not only hones EDA skills but also imparts a foundational understanding of risk analytics in financial services. It serves as a practical bridge between theory and application, shedding light on the pivotal role data plays in shaping strategic decisions within the dynamic landscape of banking.

APPROACH



Data Cleaning

Streamlined two extensive datasets, removing irrelevant columns and addressing blank data to optimize relevance for risk assessments.

Outlier Management

Identified and eliminated outliers to enhance dataset accuracy, ensuring a reliable foundation for subsequent analysis.

Analysis Techniques

Employed univariate and bivariate analysis with pivot tables and charts, unveiling crucial insights into risk assessment variables.

Iterative Refinement

Iteratively refined the dataset, ensuring the robustness of insights and contributing to a comprehensive understanding of risk factors.

TECH-STACK USED

•Microsoft PowerPoint:

For presentation of the project

•Microsoft Excel:

To be the leading provider of transformative technologies, revolutionizing industries and enriching lives through our commitment to excellence and social responsibility.

UNDERSTANDING DATA

The given dataset contains 3 files having information about loan applications. It includes two types of scenarios:

1. Customers with payment difficulties: These are customers who had a late payment of more than X days on at least one of the first Y installments of the loan.
2. All other cases: These are cases where the payment was made on time.

01. application_data.csv: encompasses comprehensive client details during the application, focusing on identifying whether a client is facing payment difficulties.

02. previous_application.csv: details the client's historical loan data, indicating the approval status of the prior application—whether it was Approved, Cancelled, Refused, or received an Unused offer.

03. columns_description.csv: serves as a data dictionary, elucidating the meanings and descriptions of variables within the dataset.

INSIGHTS

FILE 1: APPLICATION DATA

Columns – 122	Rows – 307511	Columns with null - 64

A. Identify Missing Data and Deal with it Appropriately:

- Observing that 41 columns exhibit a missing value rate surpassing 50%, we intend to eliminate these columns, particularly those predominantly associated with client residential information.
- Upon a closer examination, it is apparent that FLAG_MOBIL consistently holds the value 1, with only a solitary instance of 0. Considering the negligible variation in this feature, we have chosen to discard it.

COLUMNS TO DROP :

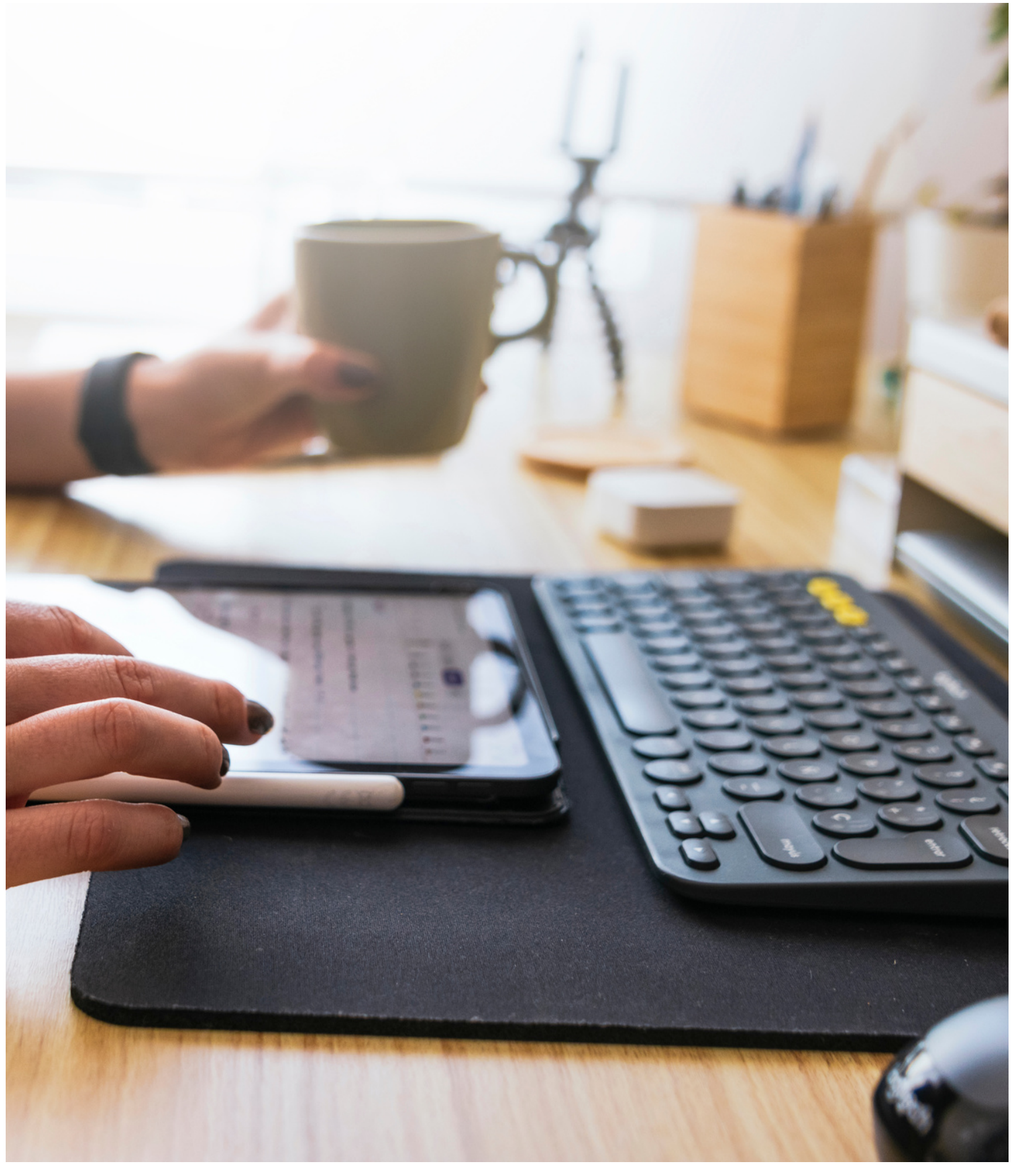
	HOUSETYPE_MODE	WAL LSMATERIAL_MODE	BASEMENTAREA_MEDI	FLOORSMIN_MEDI
	LIVINGAREA_AVG	ELEVATORS_AVG	LANDAREA_AVG	LIVINGAPARTMENTS_AVG
	LIVINGAREA_MODE	ELEVATORS_MODE	LANDAREA_MODE	LIVINGAPARTMENTS_MODE
	LIVINGAREA_MEDI	ELEVATORS_MEDI	LANDAREA_MEDI	LIVINGAPARTMENTS_MEDI
	ENTRANCES_AVG	NONLIVINGAREA_AVG	OWN_CAR_AGE	FONDKAPREMONT_MODE
	ENTRANCES_MODE	NONLIVINGAREA_MODE	YEARS_BUILD_AVG	NONLIVINGAPARTMENTS_AVG
	ENTRANCES_MEDI	NONLIVINGAREA_MEDI	EARS_BUILD_MODE	NONLIVINGAPARTMENTS_MODE
	APARTMENTS_AVG	EXT_SOURCE_1	YEARS_BUILD_MEDI	NONLIVINGAPARTMENTS_MEDI
	APARTMENTS_MODE	BASEMENTAREA_AVG	FLOORSMIN_AVG	COMMONAREA_AVG
	APARTMENTS_MEDI	BASEMENTAREA_MODE	FLOORSMIN_MODE	COMMONAREA_MODE
	COMMONAREA_MEDI			

- We have 26 columns with less than 50% missing values. We'll assess these for any less significant ones that could be removed. If none are identified, we'll address missing values in continuous features by imputing them with the mean or median, considering the impact of outliers. Categorical features with missing values will be filled using the mode.

AMT_ANNUITY	YEARS_BEGINEXPLUATATION_MODE	DEF_60_CNT_SOCIAL_CIRCLE
AMT_GOODS_PRICE	FLOORSMAX_MODE	DAYS_LAST_PHONE_CHANGE
NAME_TYPE_SUITE	YEARS_BEGINEXPLUATATION_MEDI	AMT_REQ_CREDIT_BUREAU_HOUR
OCCUPATION_TYPE	FLOORSMAX_MEDI	AMT_REQ_CREDIT_BUREAU_DAY
CNT_FAM_MEMBERS	TOTALAREA_MODE	AMT_REQ_CREDIT_BUREAU_WEEK
EXT_SOURCE_2	EMERGENCYSTATE_MODE	AMT_REQ_CREDIT_BUREAU_MON
EXT_SOURCE_3	OBS_30_CNT_SOCIAL_CIRCLE	AMT_REQ_CREDIT_BUREAU_QRT
YEARS_BEGINEXPLUATATION_AVG	DEF_30_CNT_SOCIAL_CIRCLE	AMT_REQ_CREDIT_BUREAU_YEAR
FLOORSMAX_AVG	OBS_60_CNT_SOCIAL_CIRCLE	

FEATURES TO DROP
FLOORSMAX_AVG
FLOORSMAX_MODE
FLOORSMAX_MEDI
EXT_SOURCE_2
YEARS_BEGINEXPLUATATION_AVG
YEARS_BEGINEXPLUATATION_MODE
YEARS_BEGINEXPLUATATION_MEDI
TOTALAREA_MODE
EXT_SOURCE_3
EMERGENCYSTATE_MODE

FEATURES TO KEEP	
OCCUPATION_TYPE	OBS_30_CNT_SOCIAL_CIRCLE
AMT_REQ_CREDIT_BUREAU_HOUR	DEF_30_CNT_SOCIAL_CIRCLE
AMT_REQ_CREDIT_BUREAU_DAY	OBS_60_CNT_SOCIAL_CIRCLE
AMT_REQ_CREDIT_BUREAU_WEEK	DEF_60_CNT_SOCIAL_CIRCLE
AMT_REQ_CREDIT_BUREAU_MON	DAYS_LAST_PHONE_CHANGE
AMT_REQ_CREDIT_BUREAU_QRT	AMT_GOODS_PRICE
AMT_REQ_CREDIT_BUREAU_YEAR	AMT_ANNUITY
NAME_TYPE_SUITE	CNT_FAM_MEMBERS



DROPPING THE NULL ROWS

As the dataset comprises only a few instances (less than 15 out of 307,512) with NULL values in the listed features, we've opted to remove these NULLs specifically from the following attributes:

1.AMT_ANNUITY

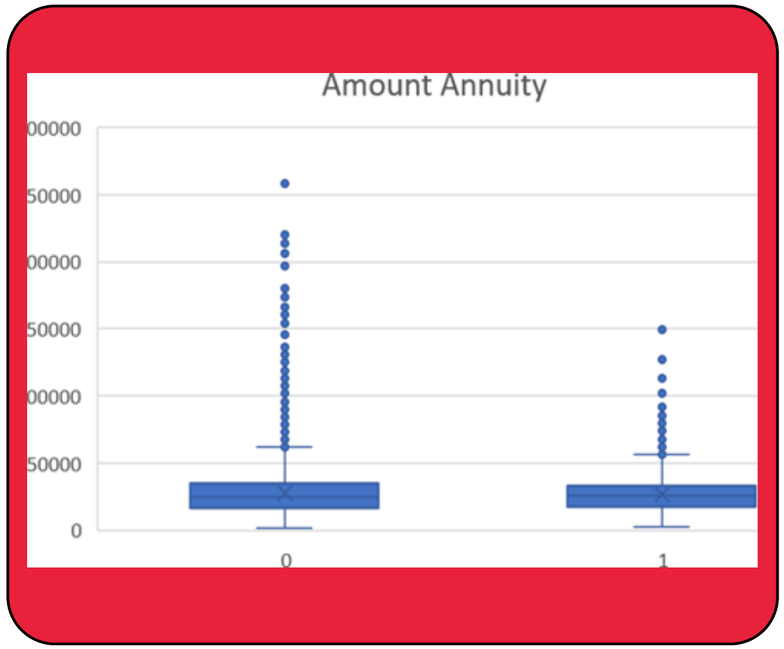
2.CNT_FAM_MEMBERS

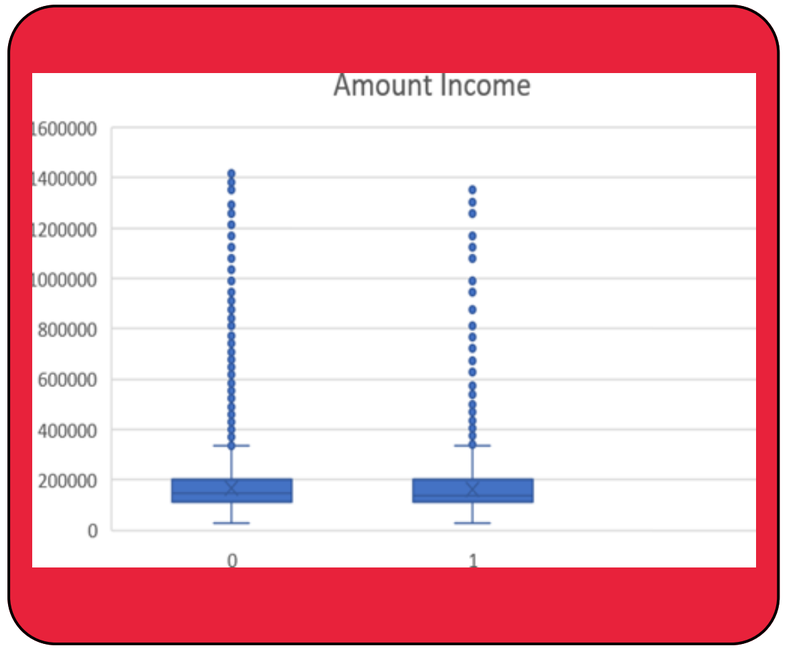
3.DAYS_LAST_PHONE_CHANGE

B. Identify Outliers in the Dataset :

Detecting outliers relies on the presence of numeric variables. To explore the association with the target column, we will utilize box plots for the following variables:

1. Credit amount
2. Income amount
3. Annuity amount



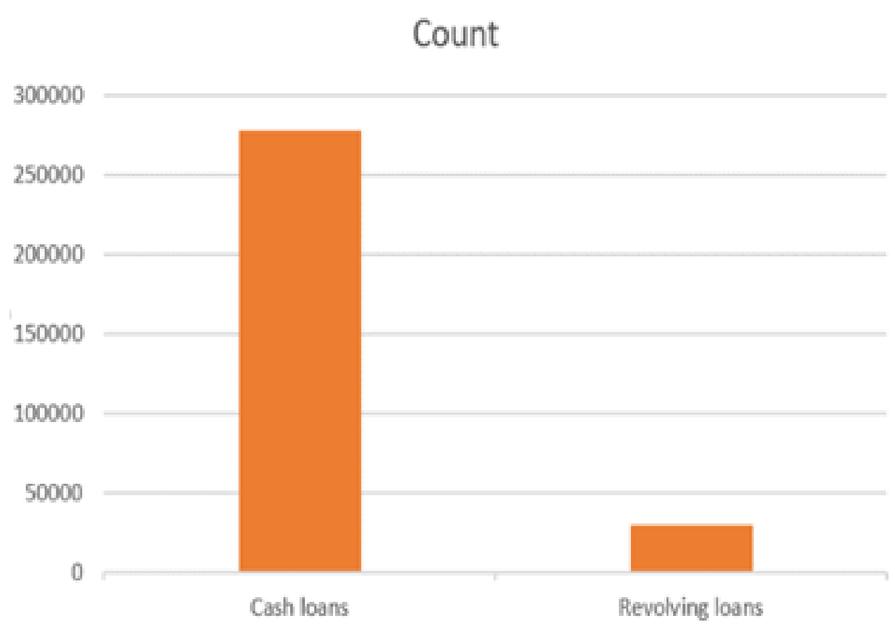


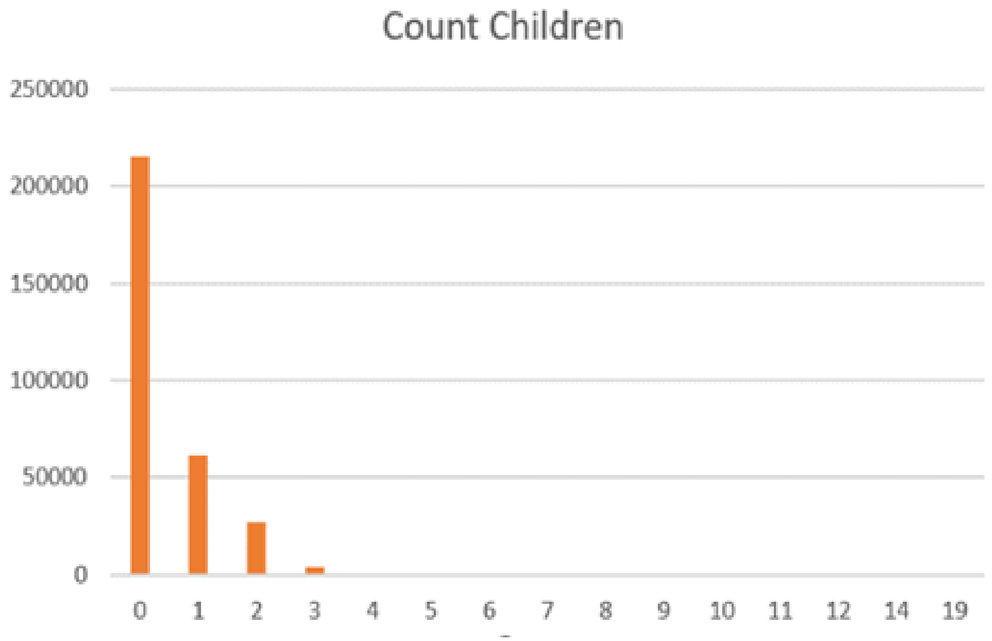
- Outliers detected in AMT_INCOME_TOTAL, with one instance reaching an unusually high salary of 117,000,000, raising accuracy concerns.
- Outliers identified in DAYS_EMPLOYED, indicating individuals supposedly employed for over 1000 years, which is implausible.
- AMT_GOODS_PRICE and AMT_CREDIT exhibit outliers, reflecting amounts significantly higher than the norm.

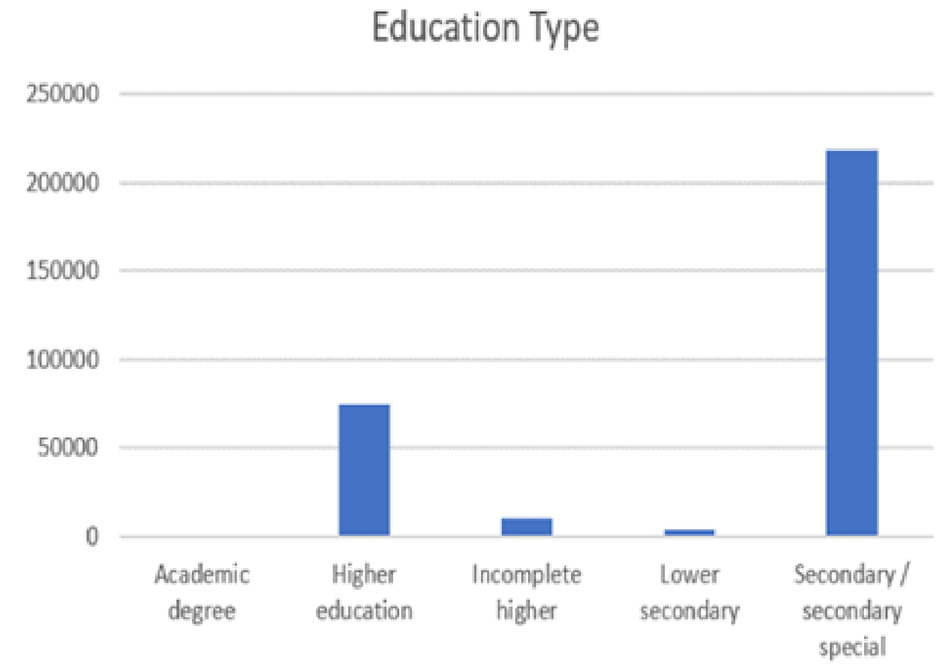
- No outliers found in DAYS_BIRTH.
- Outliers in DAYS_LAST_PHONE_CHANGE suggest data anomalies, such as individuals using the same phone for nearly 12 years, seemingly improbable with technological advancements.
- Outliers in CNT_CHILDREN indicate instances where individuals purportedly have 19 children, a situation deemed unrealistic in today's context.

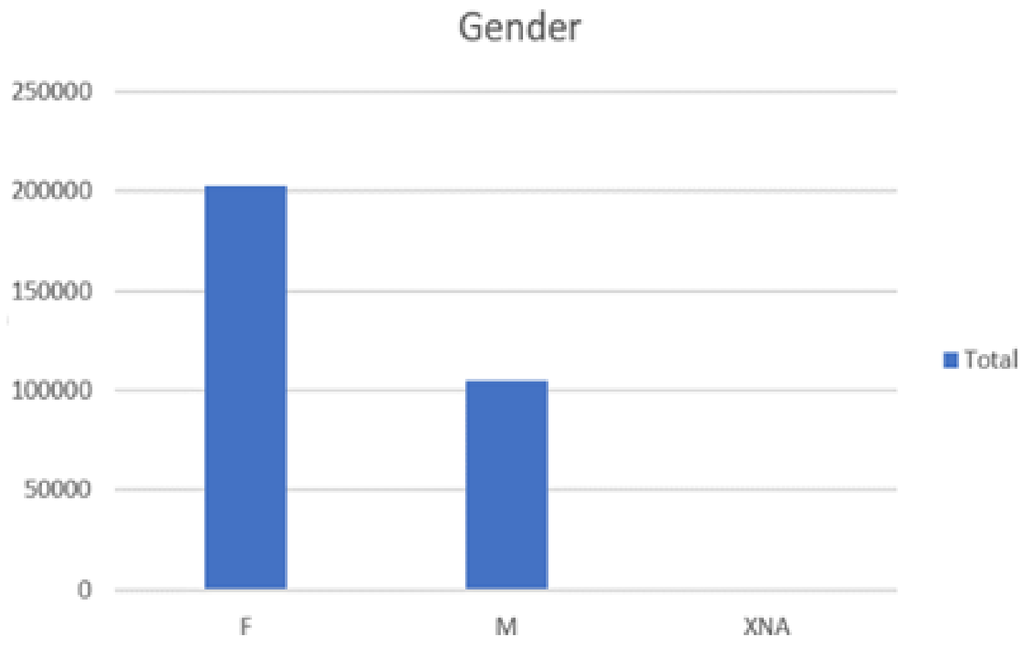
C. Analyze Data Imbalance:

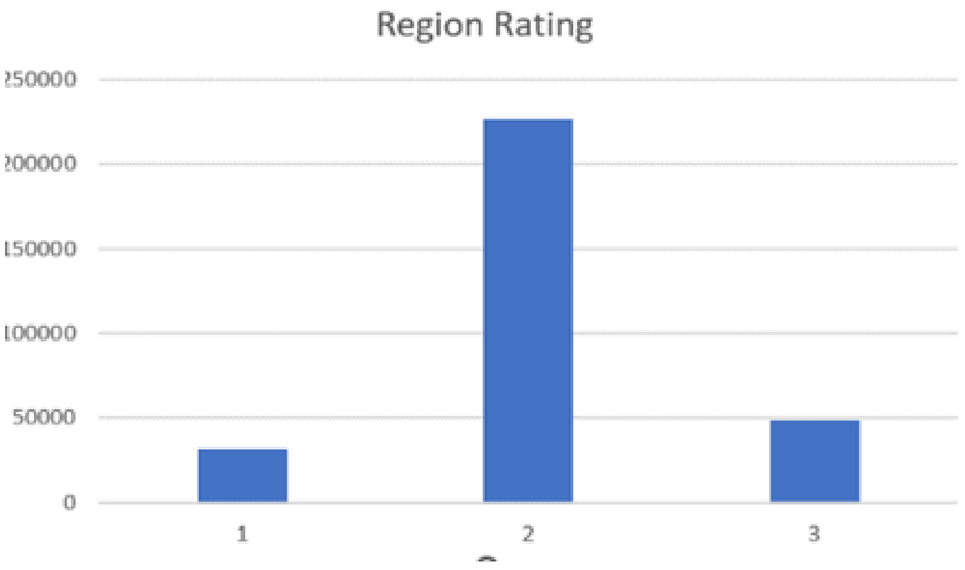
Data imbalance poses a critical challenge in binary classification, affecting the accuracy of analysis. In such scenarios, understanding the data distribution becomes crucial for constructing reliable models, as imbalances can lead to skewed predictions, particularly favoring the majority class. Addressing these imbalances through strategic techniques is essential to ensure accurate and fair predictions across all classes in the dataset.

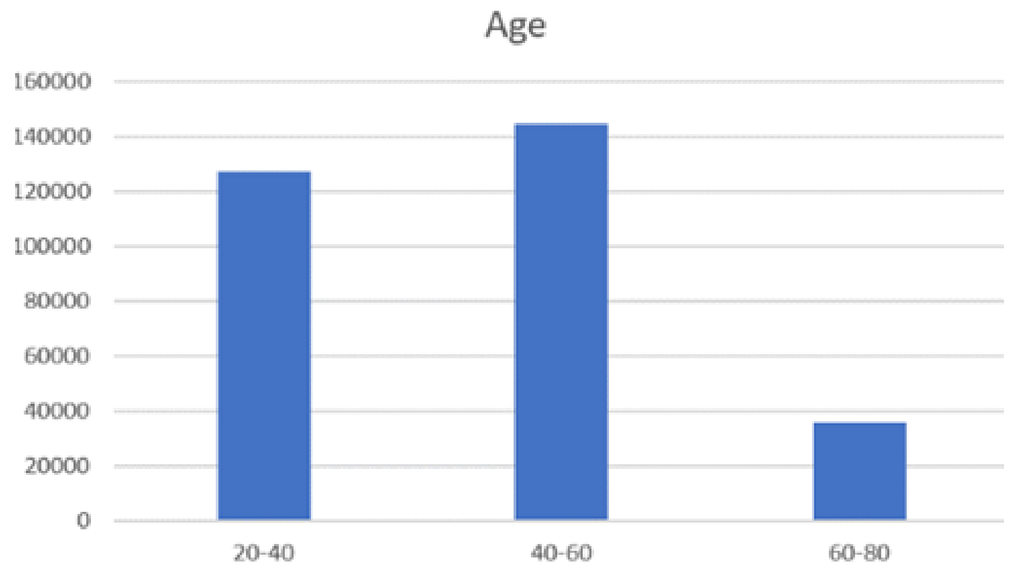





Education Type
250000
200000
150000
100000
50000
0
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special


Gender
250000
200000
150000
100000
50000
0
F
M
XNA
Total

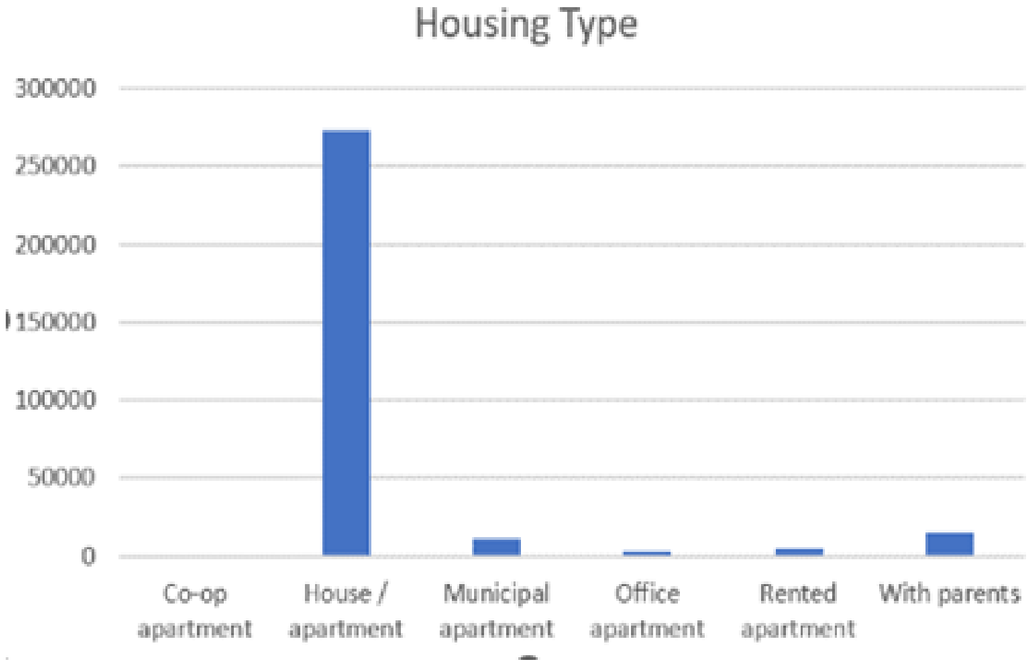

Region Rating
250000
200000
150000
100000
50000
0
1
2
3

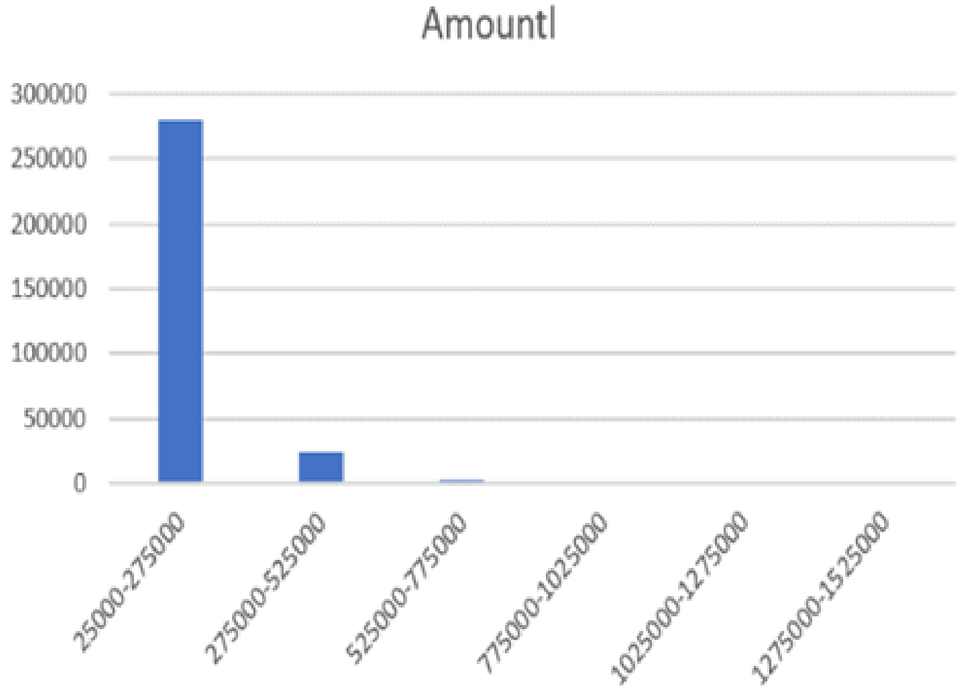

Age
160000
140000
120000
100000
80000
60000
40000
20000
0
20-40
40-60
60-80

D. Perform Univariate, Segmented Univariate, and Bivariate Analysis:

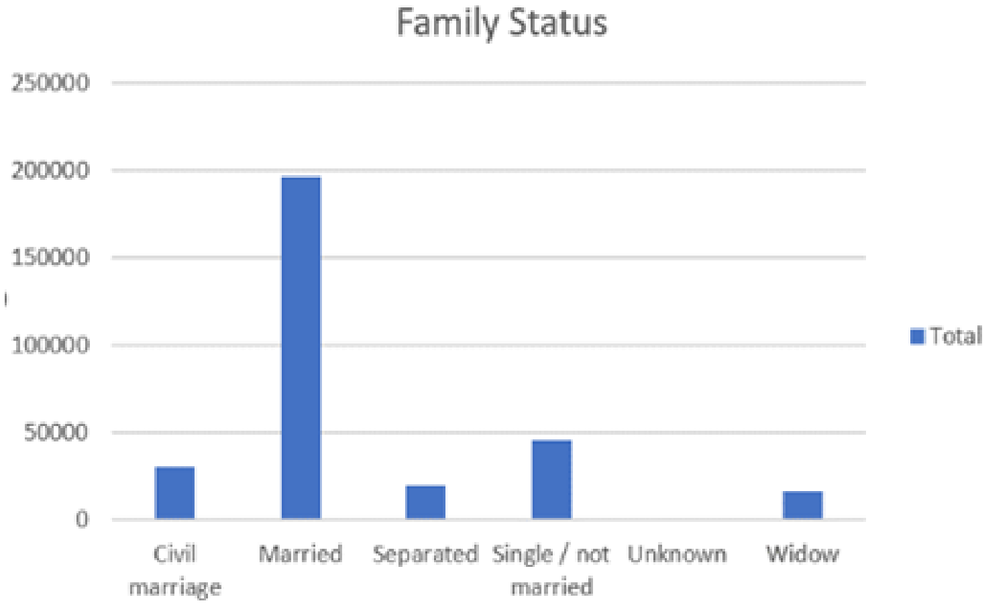
Univariate Analysis:

- Higher-income individuals show less interest in loans.
- Typical bank loan amounts range from 45,000 to 1,045,000.
- Most loan applicants are aged 35 to 50.
- Those with 0 to 8 years of work experience are more likely to seek loans.
- Homeownership and marriage correlate with increased loan applications.
- Employed individuals and unaccompanied minors request loans more frequently.

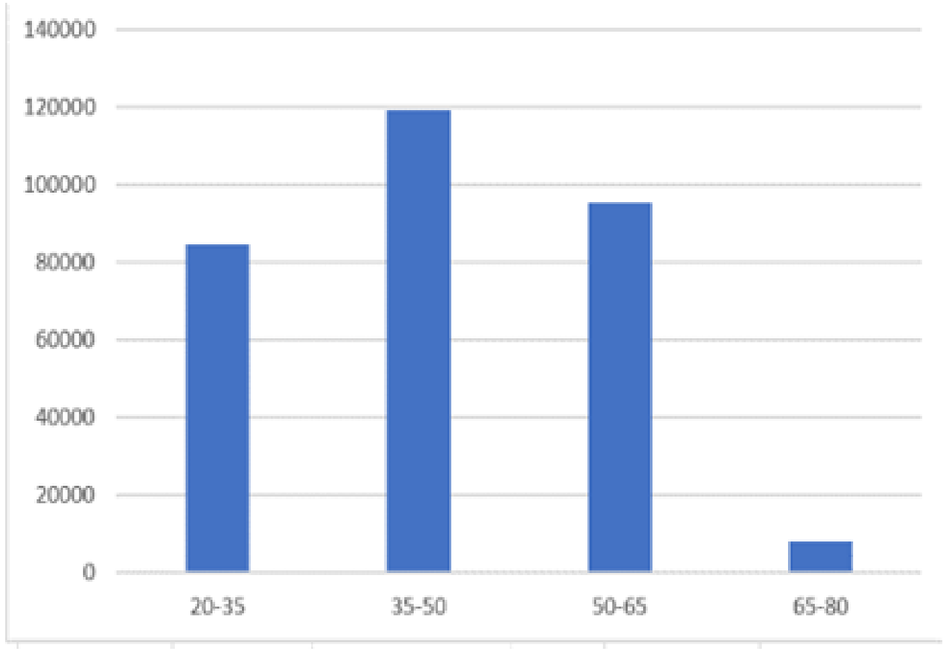

Housing Type
300000
250000
200000
150000
100000
50000
0
Co-op apartment
House / apartment
Municipal apartment
Office apartment
Rented apartment
With parents


Amountl
300000
250000
200000
150000
100000
50000
0
25000-275000
275000-525000
525000-775000
775000-1025000
1025000-1275000
1275000-1525000

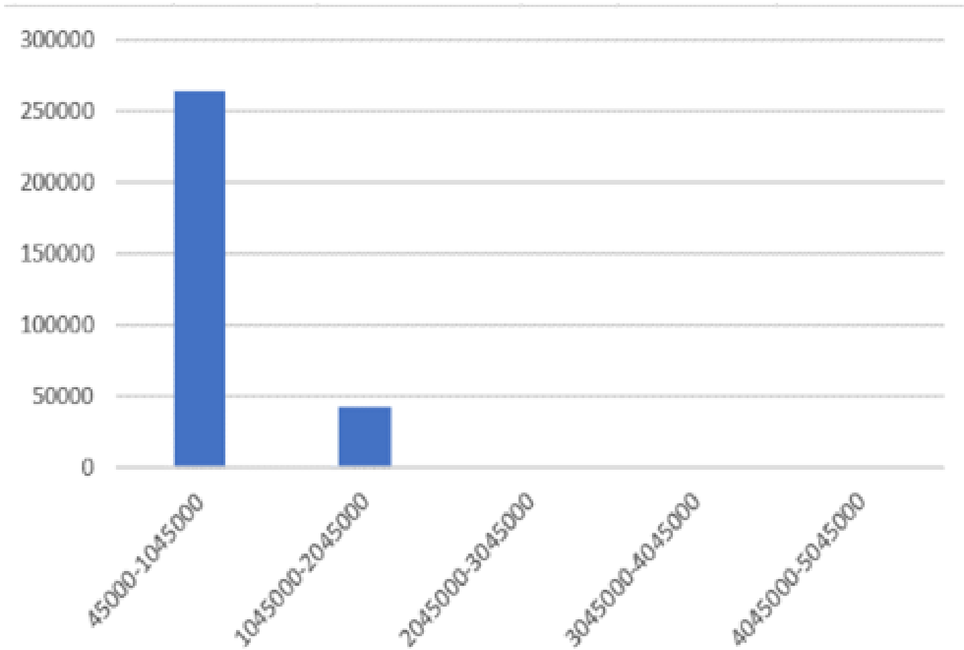

Months Employed
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
<0
0-100
100-200
200-300
300-400
400-500
500-600


Family Status
250000
200000
150000
100000
50000
0
Civil marriage
Married
Separated
Single / not married
Unknown
Widow
Total

Age



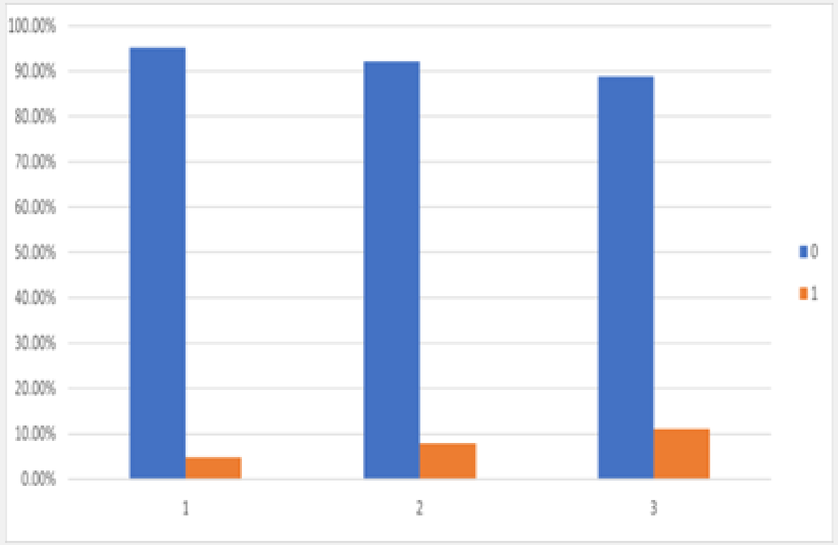
Amount Credit



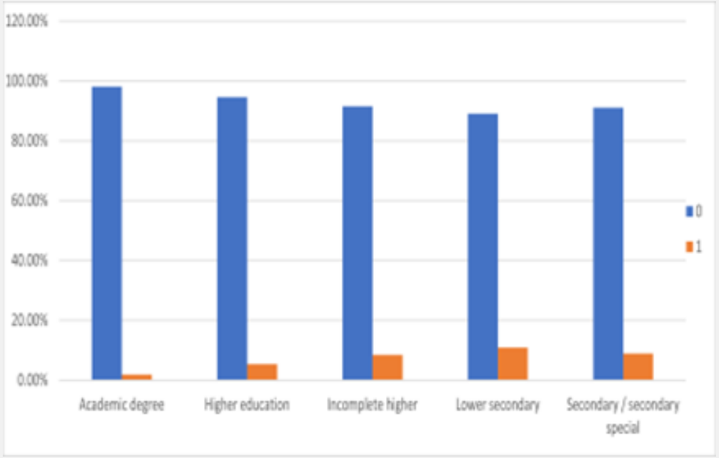
Bivariate Analysis:

- Residents in low-rating areas face higher default rates.
- Lower income correlates with increased default likelihood.
- Younger individuals are more prone to defaults, declining with age.
- Females are less likely to default compared to males.
- Maternity leave and unemployment elevate default predictions.
- Families with over five members are more likely to default.
- Lower education and limited work experience heighten default chances.

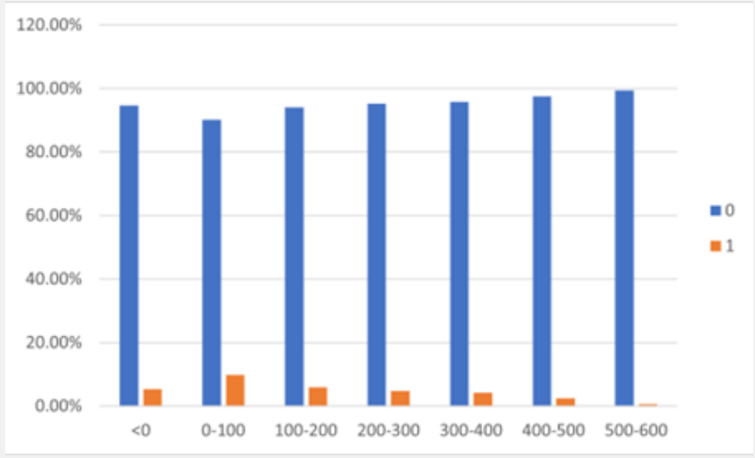
REGION RATING CLIENT VS TARGET



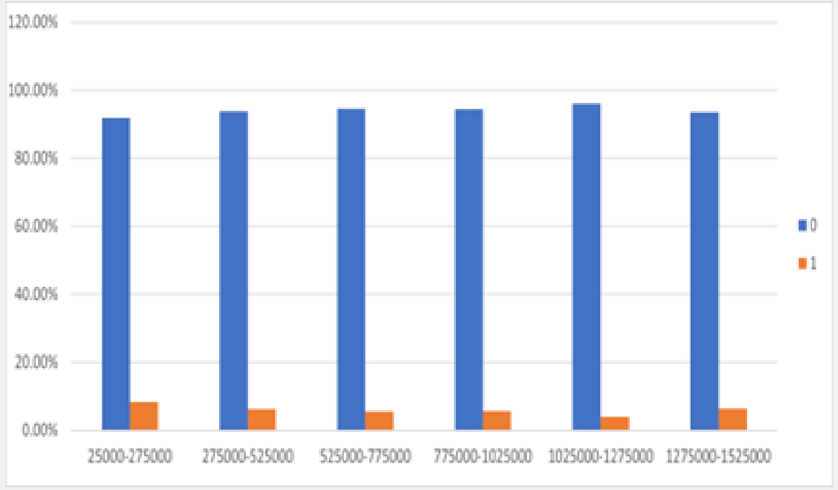
EDUCATION TYPE VS TARGET



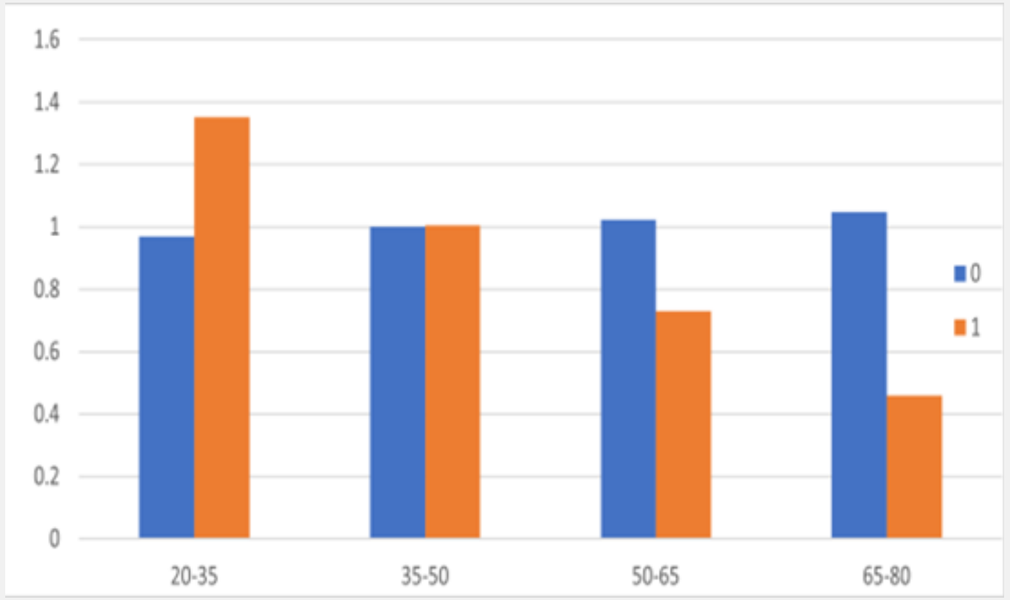
MONTHS EMPLOYED VS TARGET



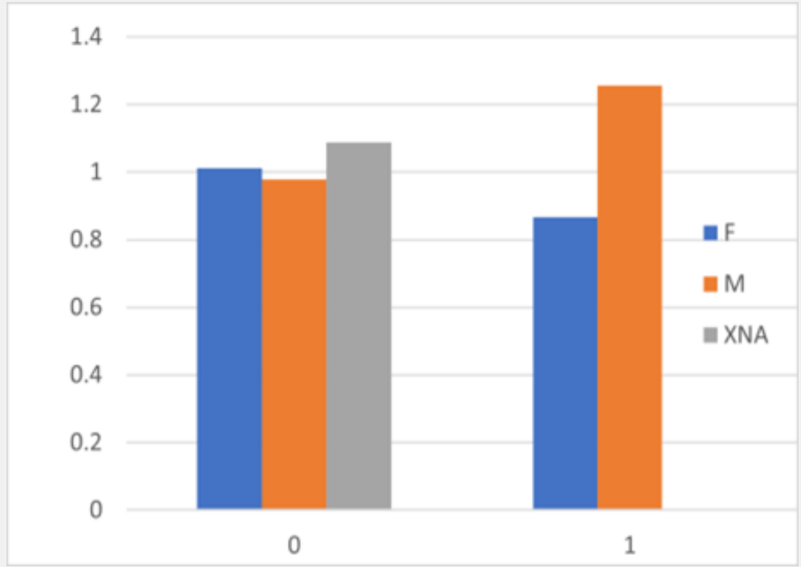
AMOUNT INCOME VS TARGET



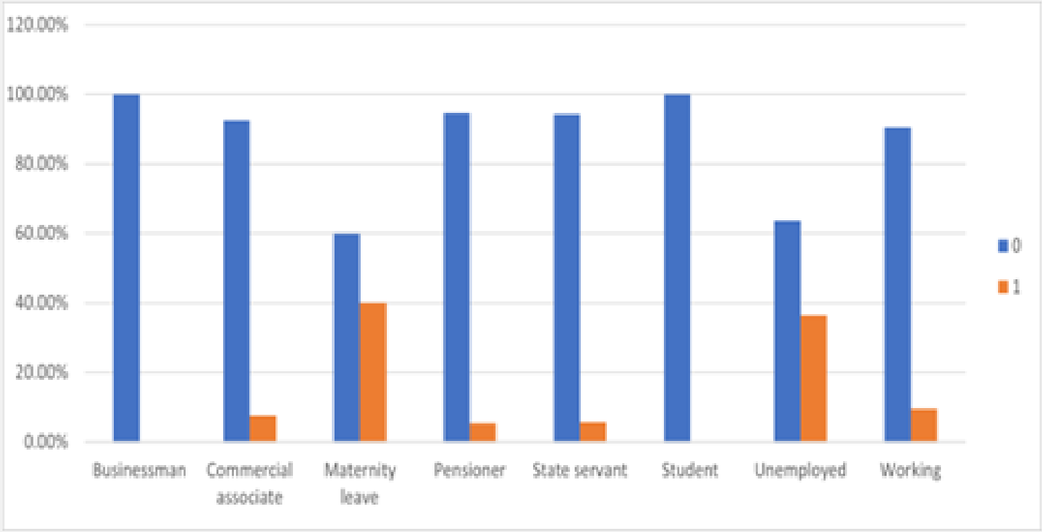
AGE VS TARGET



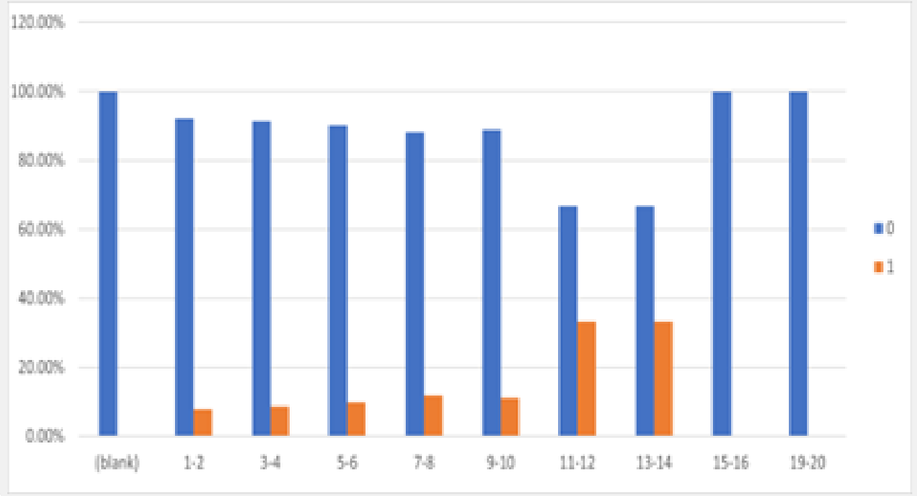
GENDER VS TARGET



INCOME TYPE VS TARGET



FAMILY MEMBER VS TARGET



E. Identify Top Correlations for Different Scenarios:

Understanding the correlation between variables and the target variable can provide insights into strong indicators of loan default.

Top 10 driving factors in current application.csv

1	Income type
2	Count of Family Members
3	Children count
4	External source
5	Region rating of client
6	Age
7	Months Employed
8	Amount credit
9	Amount Goods Price
10	Amount total income

FILE 2 : PREVIOUS APPLICATION DATA

A. Identify Missing Data and Deal with it Appropriately:

- The dataset comprises 1,670,214 rows, exceeding Excel's maximum limit of 1,048,576 rows. However, in adherence to project requirements, analysis will be constrained to the use of Excel, accommodating a maximum of 1,048,576 rows.

Columns – 37	Rows - 1670214	Columns with null – 15

- Observing that four columns have missing values exceeding 50%, we will eliminate these columns.
- Upon additional examination, we deduce that the mentioned columns exert no impact on loans or their repayments. Consequently, we opt to exclude them.
- Given the negligible data loss of 0.02% in PRODUCT_COMBINATION and the surplus data that exceeds Excel's loading capacity, we choose to eliminate rows with NULL values.

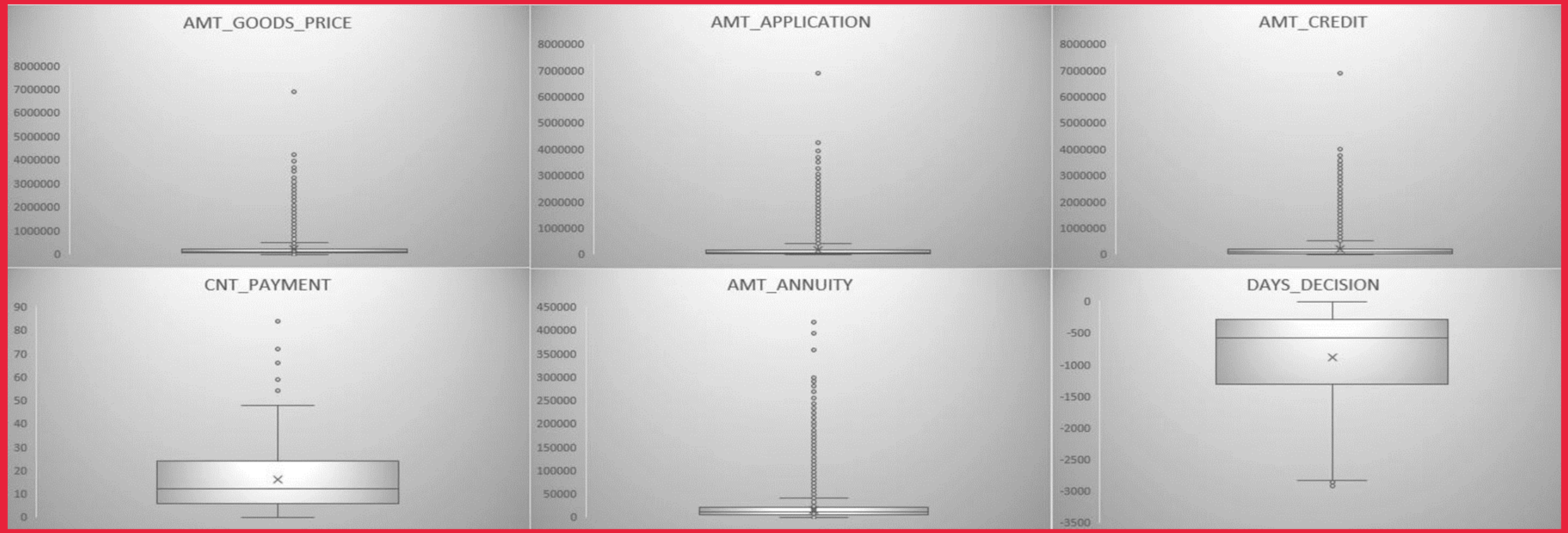
COLUMNS TO DROP

RATE_INTEREST_PRIMARY
RATE_INTEREST_PRIVILEGED
AMT_DOWN_PAYMENT
RATE_DOWN_PAYMENT
NAME_TYPE_SUITE
WEEKDAY_APPR_PROCESS_START
HOUR_APPR_PROCESS_START
FLAG_LAST_APPL_PER_CONTRACT
NFLAG_LAST_APPL_IN_DAY

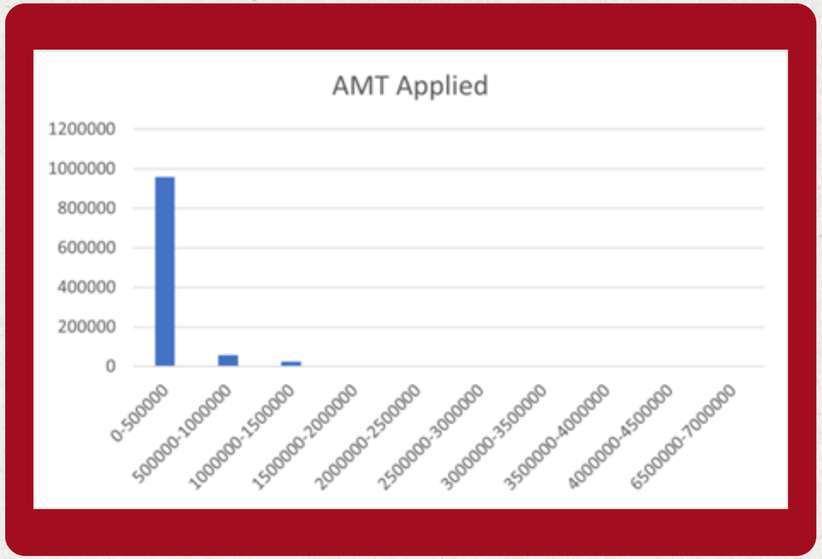
DROPPING NULL ROWS

PRODUCT_COMBINATION

B. Identify Outliers in the Dataset :

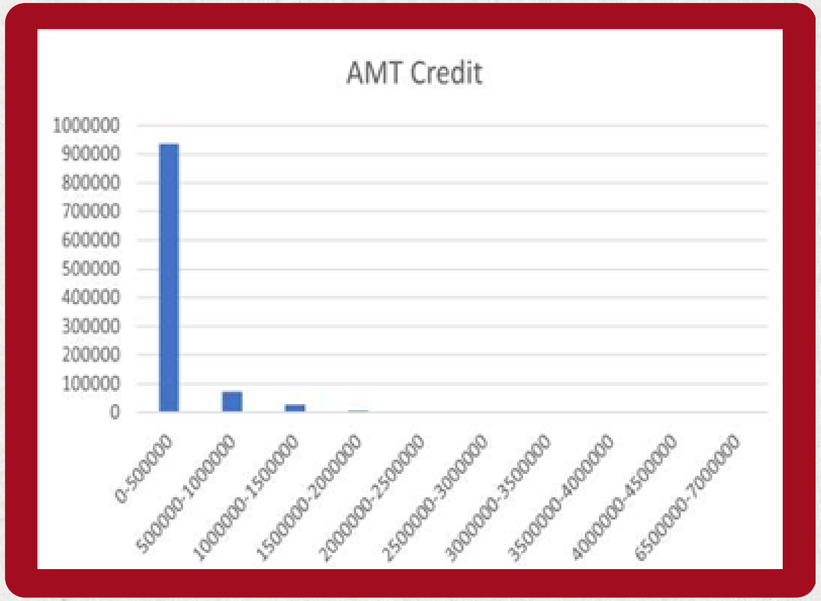


- Outliers in DAYS_DECISION indicate prolonged decision times, a potential concern for business operations.
- A significant number of outliers are observed in AMT_GOODS_PRICE, AMT_APPLICATION, AMT_CREDIT, and AMT_ANNUITY.
- Outliers are also present in CNT_PAYMENT.



C. Analyze Data Imbalance:

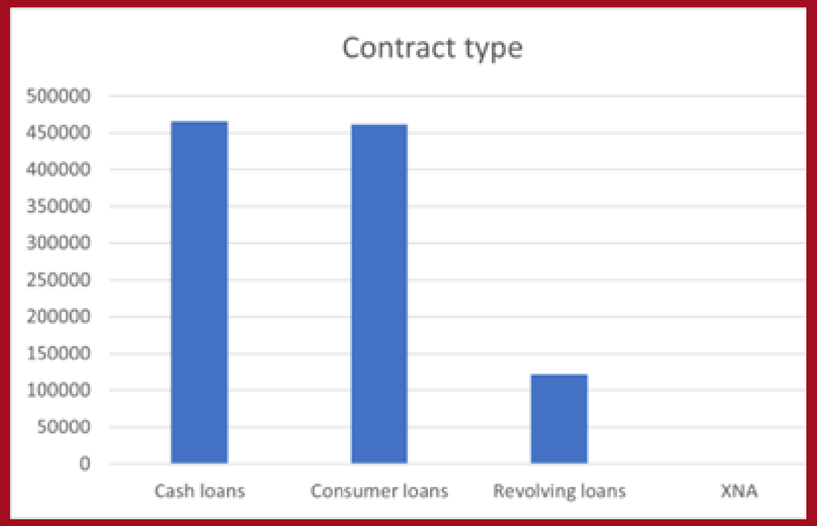
Data imbalance can affect the accuracy of the analysis, especially for binary classification problems. Understanding the data distribution is crucial for building reliable models. The columns where data is unevenly distributed are :

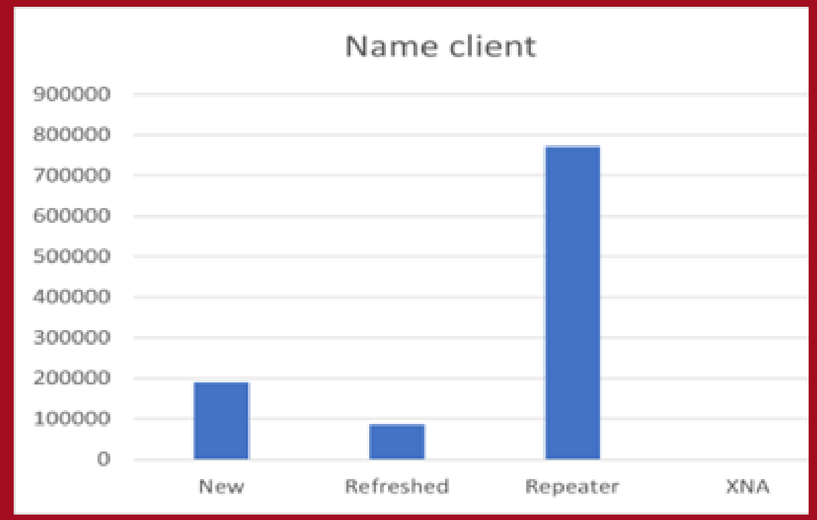


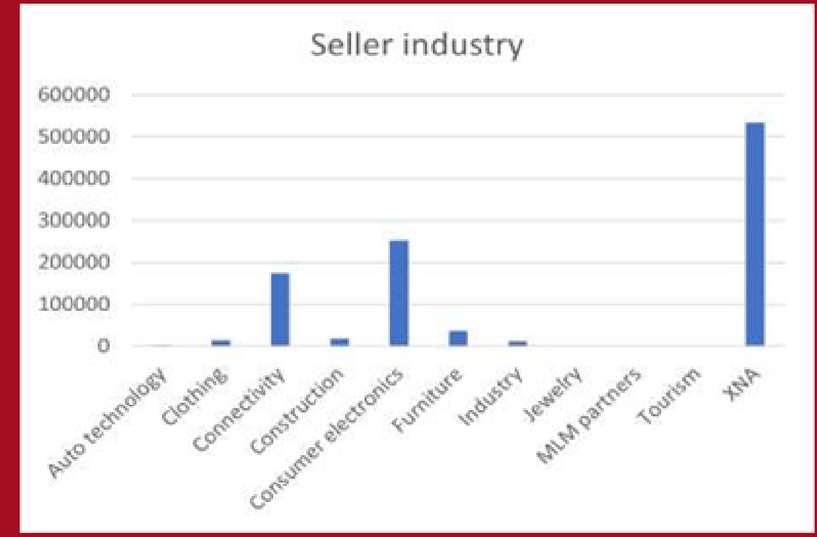
D. Perform Univariate, Segmented Univariate, and Bivariate Analysis:

Univariate Analysis:

Customers predominantly opt for cash and consumer loans, with a substantial portion being repeat clients. The majority of current loan applicants are individuals who sought loans within the past ten months, and there is an increased demand for loans related to consumer gadgets.

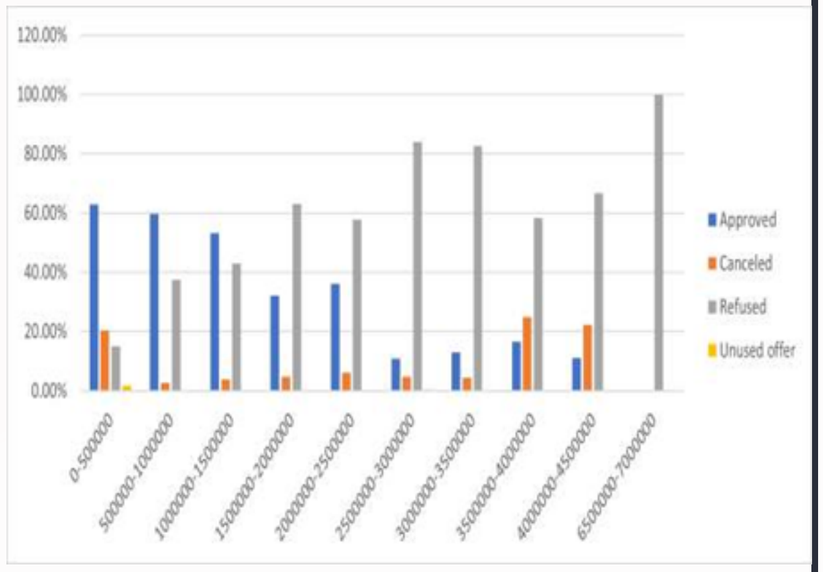


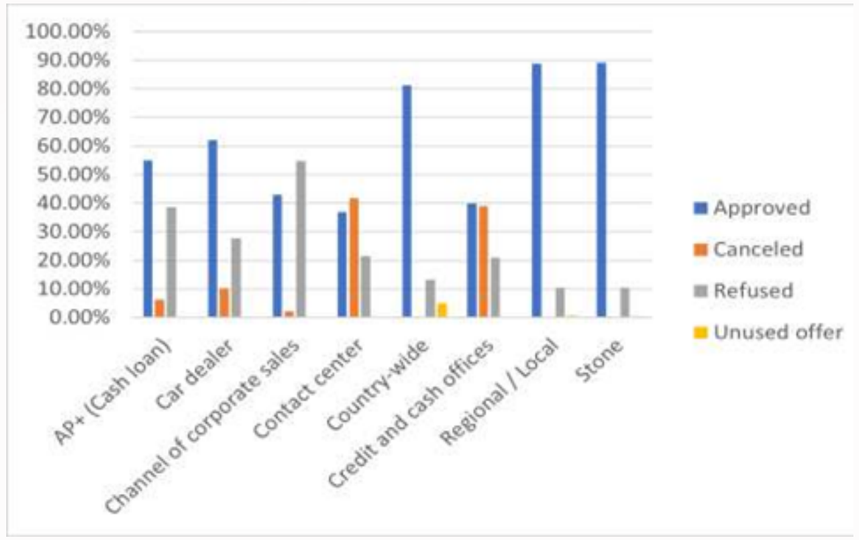


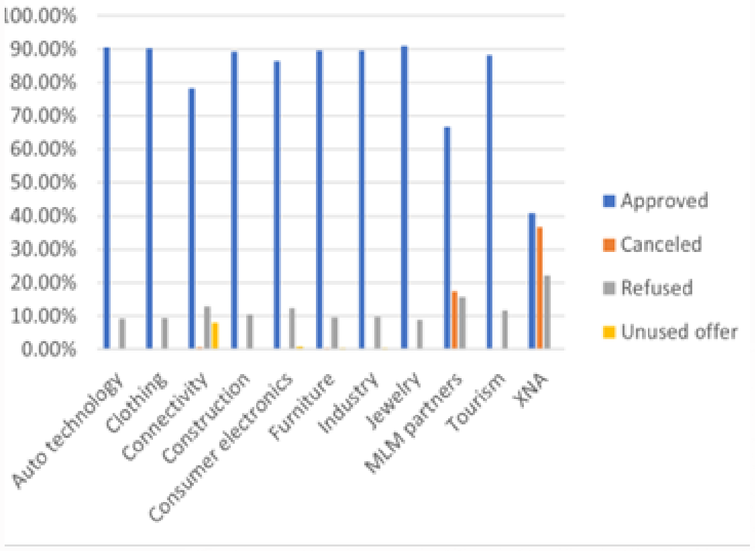


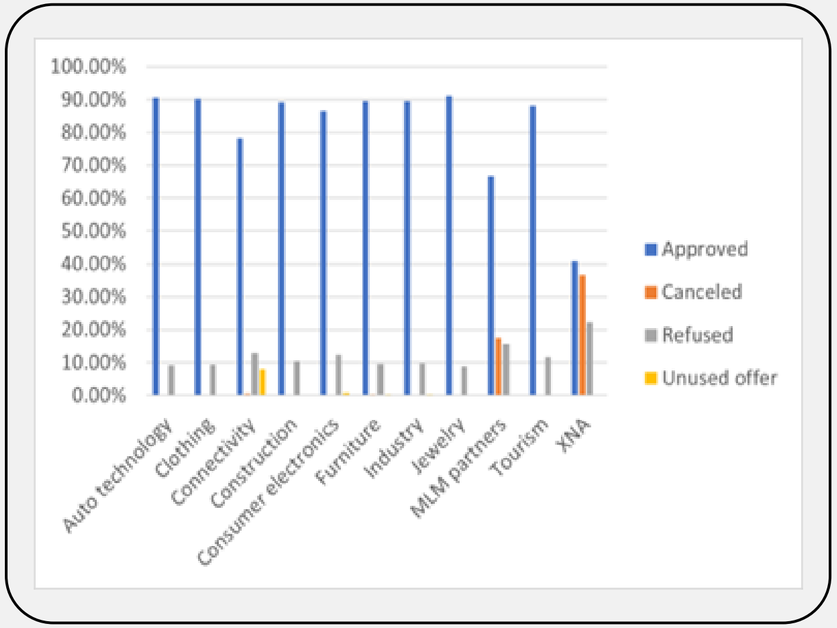
Bivariate Analysis:

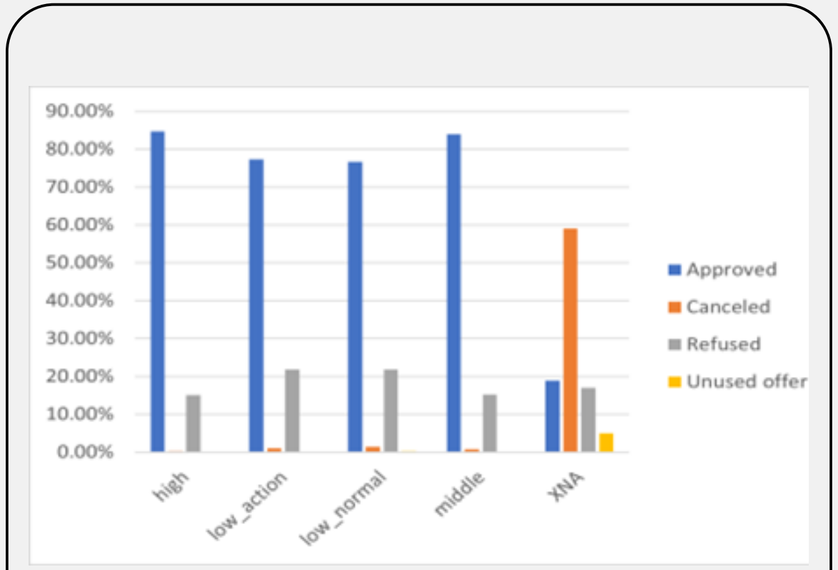
Clients requesting amounts exceeding Rs. 350,000 are likely to encounter denials. The majority of loans initiated through Credit and Cash agencies end up being cancelled. New clients experience high approval rates, particularly for consumer loans, whereas car loans consistently face denials. Loans extended to MLM partner clients are prone to cancellations, accounting for nearly 80% of authorized loans. Consumer loans demonstrate minimal cancellations and the highest approval rates. Conversely, loans for the first Selling place area group witness several cancellations. Clients reapplying for a loan within 10 months of their previous one are more susceptible to cancellations. Walk-in loans exhibit a heightened rate of refusals.

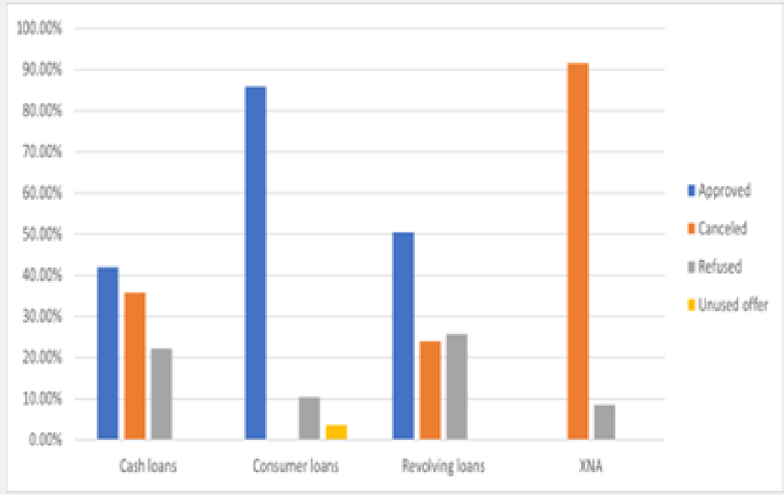


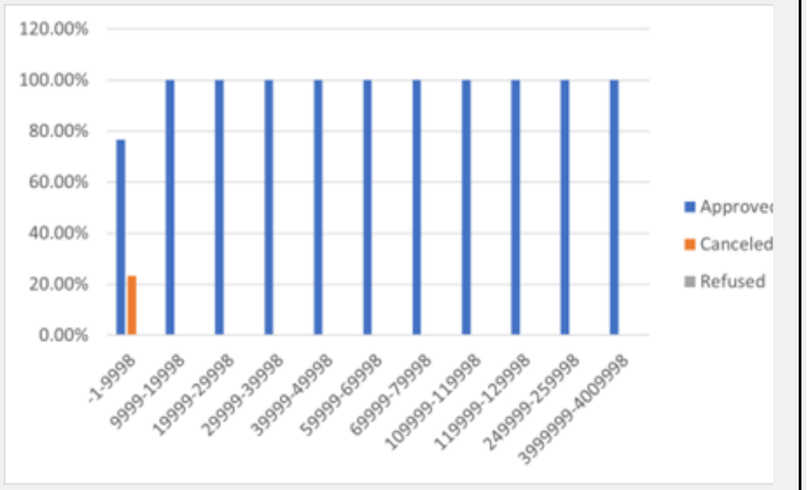


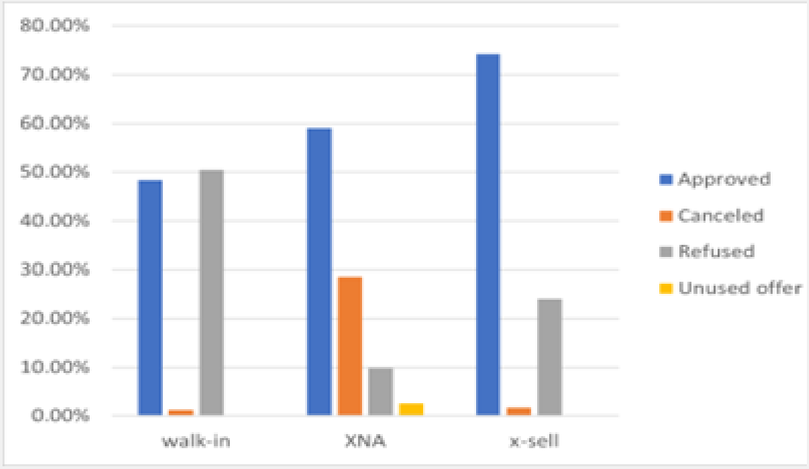


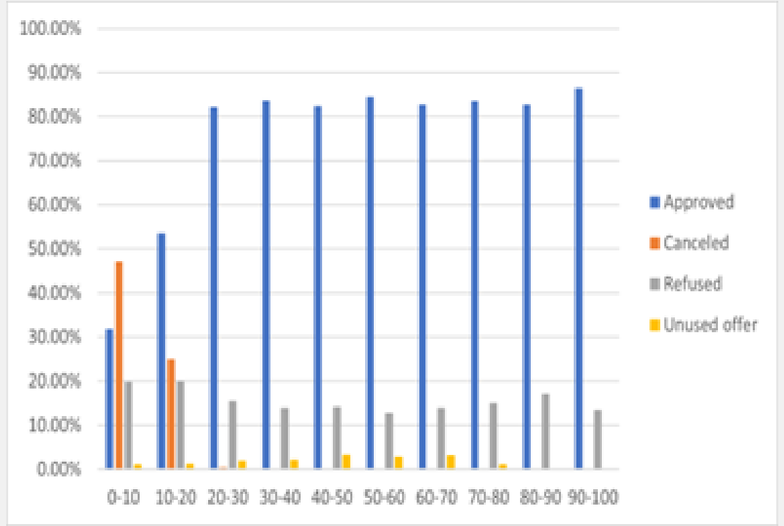

100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
Auto technology
Clothing
Connectivity
Construction
Consumer electronics
Furniture
Industry
Jewelry
MLM partners
Tourism
XNA
Approved
Canceled
Refused
Unused offer


90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
high
low_action
low_normal
middle
XNA
Approved
Canceled
Refused
Unused offer


100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
Cash loans
Consumer loans
Revolving loans
XNA
Approved
Canceled
Refused
Unused offer


120.00%
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
-1-9998
9999-19998
19999-29998
29999-39998
39999-49998
59999-69998
69999-79998
109999-119998
119999-129998
249999-259998
3999999-4009998
Approved
Canceled
Refused


80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
walk-in
XNA
x-sell
Approved
Canceled
Refused
Unused offer


100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
Approved
Canceled
Refused
Unused offer

E. Identify Top Correlations for Different Scenarios:

(Finding Correlations): Uncovering the top ten reasons for loan cancellation and refusal involves a comprehensive analysis of factors like credit scores, income levels, employment history, and loan amounts. By utilizing statistical tools and data visualization, we aim to discern significant relationships and dependencies within the dataset. The objective is to provide actionable insights for stakeholders to enhance the loan approval process.

1	Amount Application
2	Cash loan Purpose
3	Goods Category
4	Product Combination
5	Product type
6	Channel type
7	Months Decision
8	Contract type
9	Client type
10	Payment type

RESULT

Recommended Loan Groups:	High-Risk Loan Groups:
1. Clients with approved previous applications 2. Married individuals 3. Senior citizens 4. Clients with higher educational qualifications 5. Individuals with high income levels 6. Clients with significant external sources of income 7. Female applicants 8. Customers with substantial work experience	1. Unemployed clients 2. Young clientele 3. Customers with a history of denied prior applications 4. Low-income individuals 5. Clients lacking sufficient external sources of income 6. Customers with limited work experience 7. Clients on Maternity Leave 8. Individuals with a larger number of family members

CONCLUSION

Through this project, we aimed to identify patterns predicting customer payment challenges, aiding decisions on loan approval, denial, or adjustment. The insights gained enhance the company's ability to make informed decisions, particularly in understanding key factors influencing loan defaults. The project's success lies in providing actionable insights, contributing to more effective risk assessment and improved decision-making in the Bank Loan Case Study.