Bank loan Case study

TRAINITY

ATTULAYA KUMAR SINGH

attulaya@gmail.com

PROJECT DESCRIPTION:

This case study demonstrates the practical application of EDA in a real business environment. It also provides insights into risk analytics in the banking and financial services industry and how data is utilized to minimize lending risks and financial losses.

APPROACH:

In this case study, there are two large datasets: the current application dataset and the previous application dataset. These datasets contain unnecessary columns that are not relevant for risk assessments, as well as numerous missing values. To evaluate this extensive dataset, the first step was to clean the data by removing irrelevant columns and handling missing data. Outliers were identified and eliminated from the dataset. Subsequently, univariate and bivariate analysis were conducted using pivot tables and charts to gain insights and understand the relationships between variables.

THE FOLLOWING TECHNOLOGY STACK WAS USED:

- MySQL Workbench 8.0 CE
- Microsoft Excel office 365

RESULTS:

1. OVERALL METHOD TO ANALYSIS:

The bank's objective is to identify the primary factors contributing to bank loan defaults, which will aid in risk assessment within the organization. We have been provided with two extensive datasets for this purpose:

1. 'application_data.csv': This dataset comprises detailed information about clients at the time of their loan application, including indicators of financial distress.

2. 'previous_application.csv': This dataset contains data related to the clients' previous loan applications, indicating whether those applications were Accepted, Cancelled, Refused, or Unused.

Both datasets contained numerous unnecessary columns that are irrelevant to risk analytics, as well as several missing values. Therefore, the initial step involved data cleaning and preprocessing to eliminate these undesired columns and handle missing data effectively.

Following the data cleaning procedure, I split columns in the dataset based on two categories of variables.

1) Categorical variables

2) Numerical variables

Categorical variables (non-numerical variables)- person's occupation, education status.

Numerical variables - income, credit etc.,

The following are some of the categorical and numerical variables from the provided data set.

Categorical variables	Numeric variables
Gender	Age
Name contract type	Days employed
Income type	Amount Income
Education	Amount Annuity
Housing type	Amount Credit

I completed full EDA on the present application and then on the previous application. Then, in this report, I summarised the results of both applications and provided business insights.

CURRENT APPLICATION.CSV

TASK 2 (FIND MISSING DATA):

The existing application sheet included 161 columns.

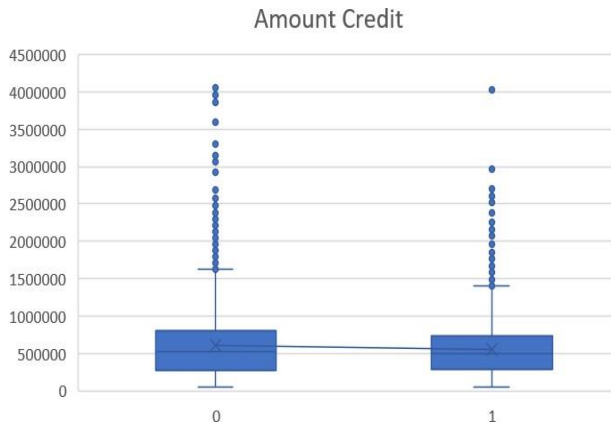
I deleted columns with more than 5% blank data.

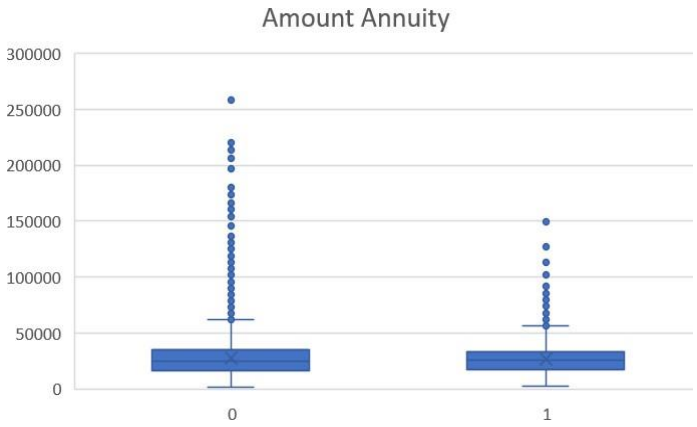
Task 3:

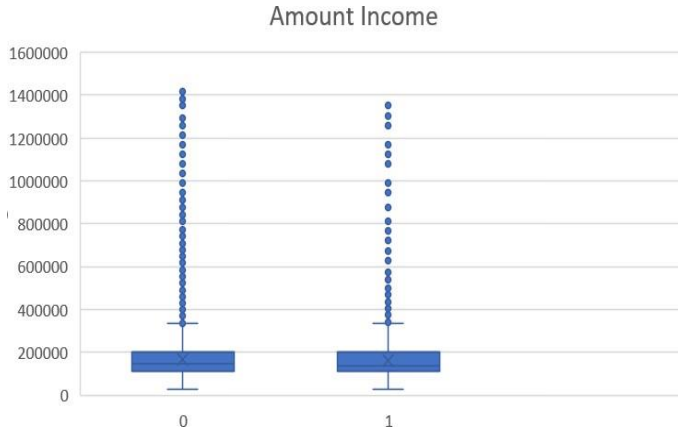
Outliers can only be identified on Numeric variables.

Box plotted Target column vs

1) Amount credit
2) Amount Income
3) Amount Annuity



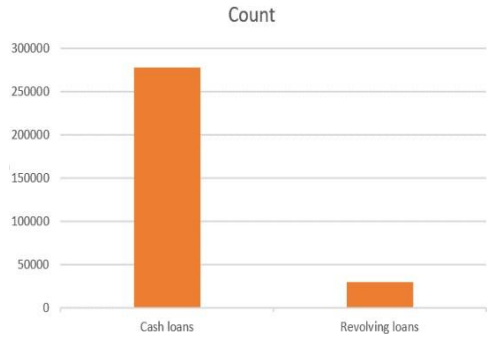


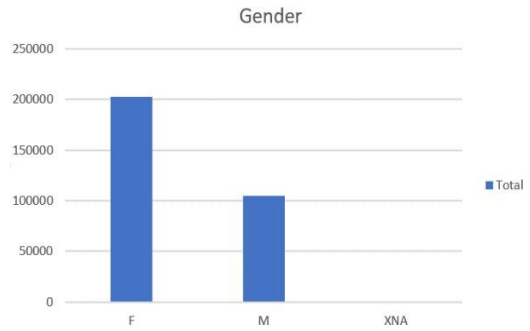


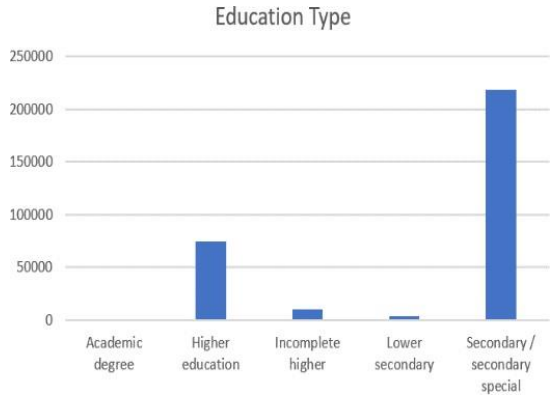
TASK (DATA IMBALANCE):

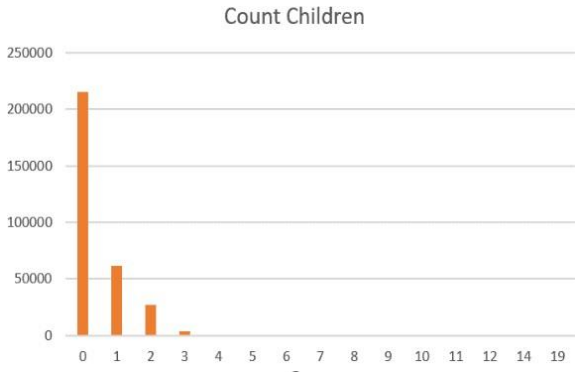
Data imbalance occurs when data is disseminated in an unequal manner. I plotted data imbalance using Pivot charts.

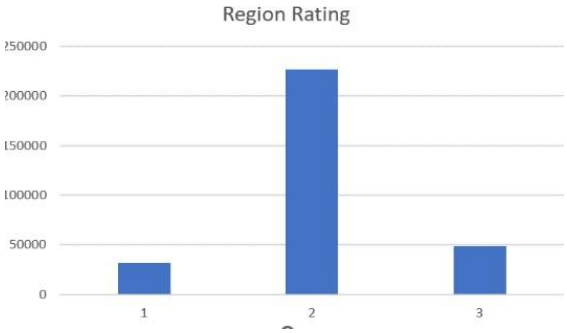
NAME CONTRACT TYPE

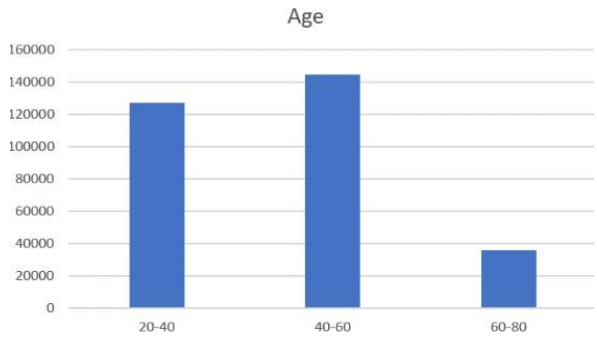












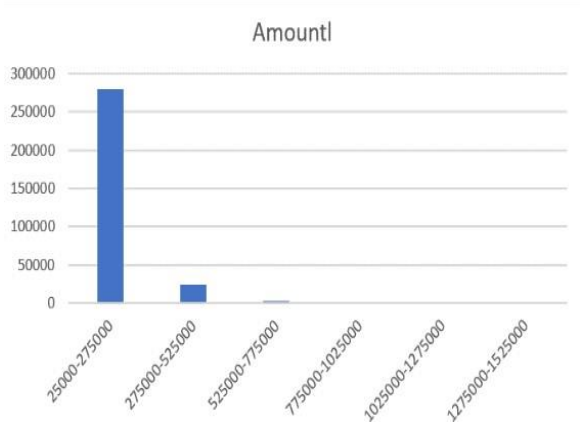
TASK 5 (EDA):

Univariate Analysis:

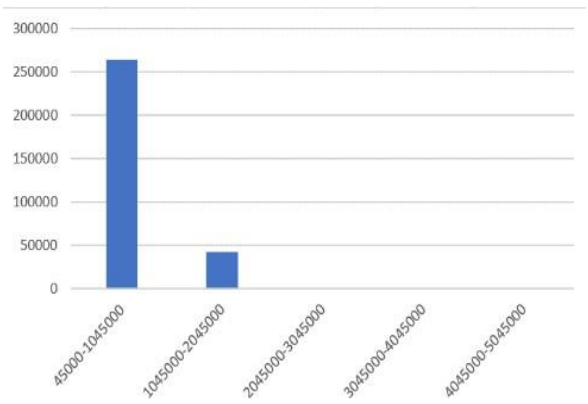
INFERENCE

Individuals with higher incomes are less likely to apply for loans. The credit amount of a bank loan is typically in the range of 45000 to 1045000. The majority of loan applications have come from people between the ages of 35 and 50. Those with 0 to 8 years of work experience are the most likely to seek for loans. Individuals who own homes are more likely to apply for loans than others. Those who are married have taken out more loans. More loans have been requested by working people. Unaccompanied minors have requested for extra loans.

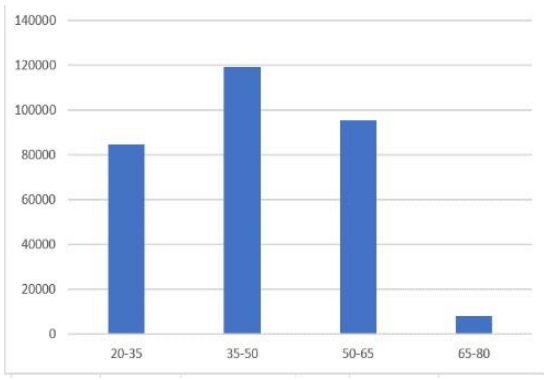
Amount Income

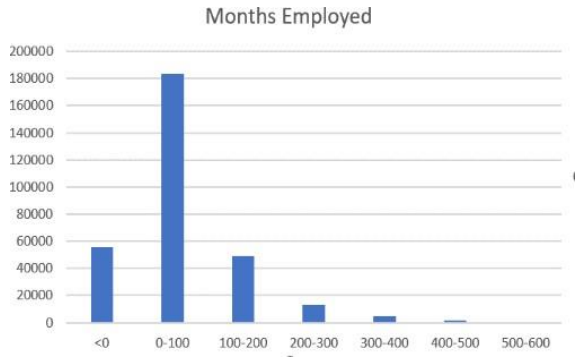


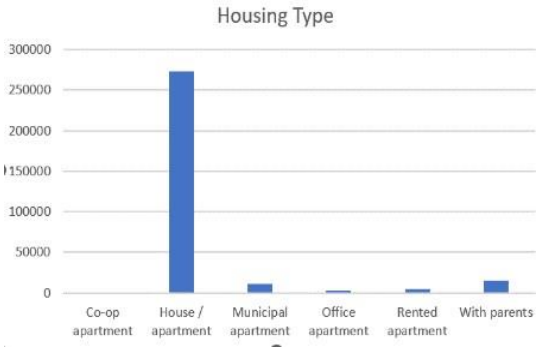
Amount Credit

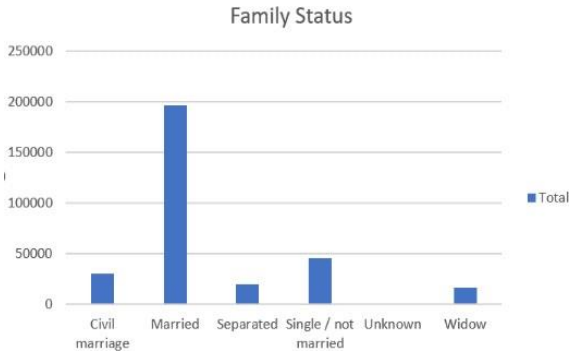


AGE

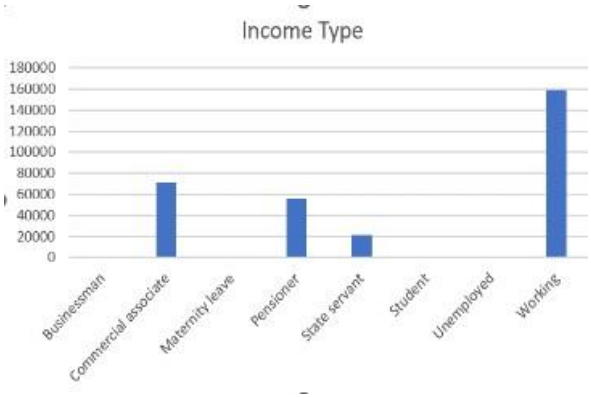


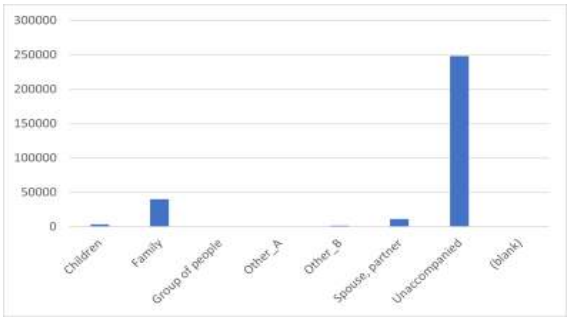






Name suite type





BIVARIATE ANALYSIS:

INFERENCE

Based on the analysis conducted, the following patterns and trends were observed regarding bank loan defaults:

1. Customers residing in areas with lower ratings are more prone to defaults.

2. Individuals with lower incomes have a higher likelihood of defaulting on their loans.

3. Younger individuals exhibit a higher probability of default, while the default rate tends to decrease with increasing age.

4. Females are less likely to default compared to males.

5. Defaults are more prevalent among customers on maternity leave or facing unemployment.

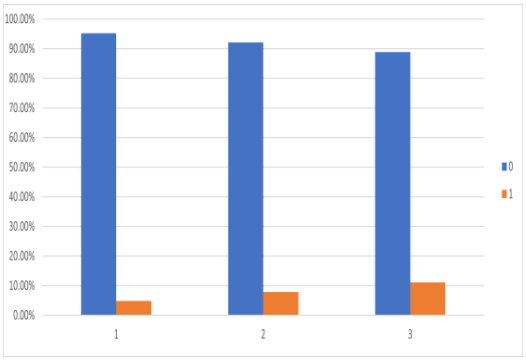
6. Customers with larger families (more than five members) are more likely to default on their bank loans.

7. Customers with lower levels of education have a higher likelihood of defaulting.

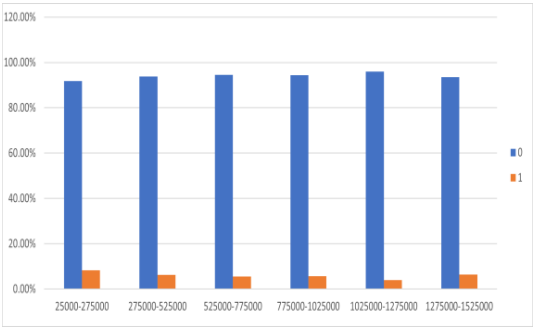
8. Individuals with limited work experience are more prone to loan defaults.

These insights provide valuable information for risk assessment in identifying the major causes of bank loan defaults.

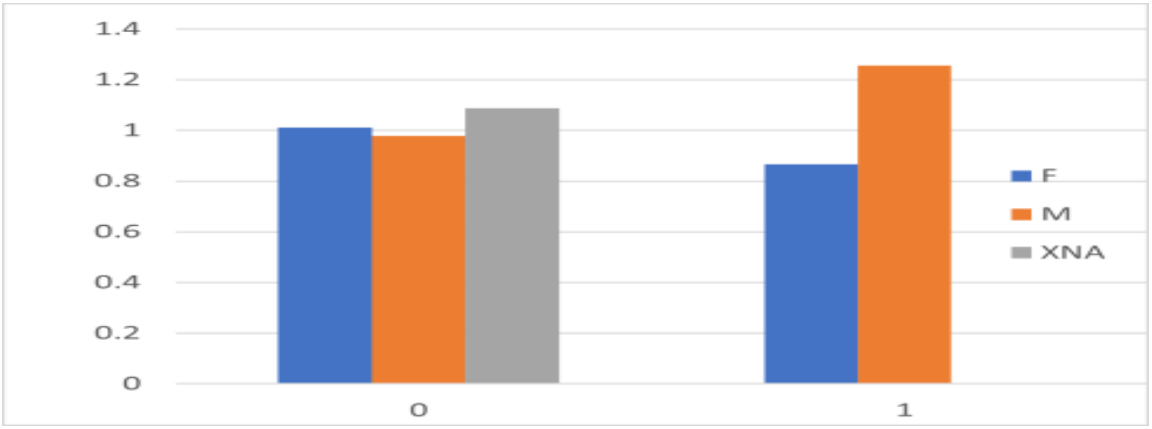
Region Rating Client vs Target



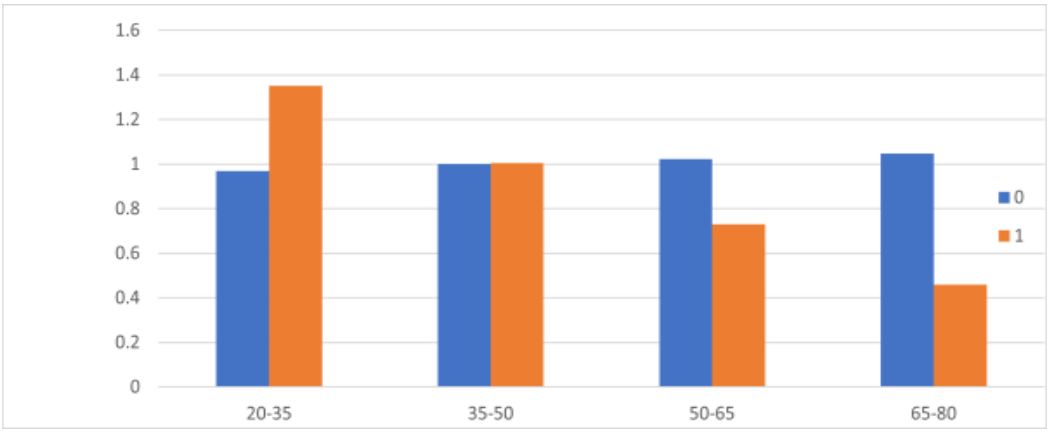
Amount Income vs Target



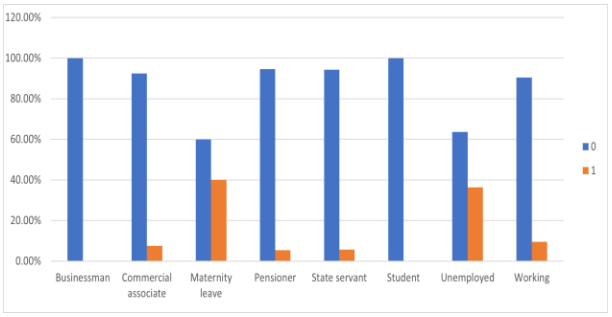
GENDER VS TARGET



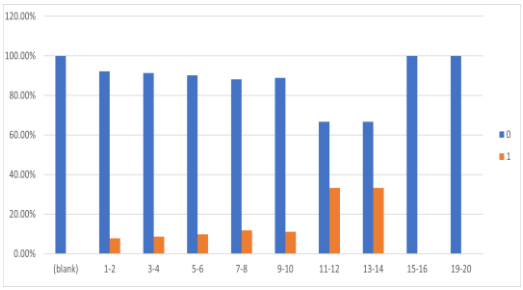
Age vs Target



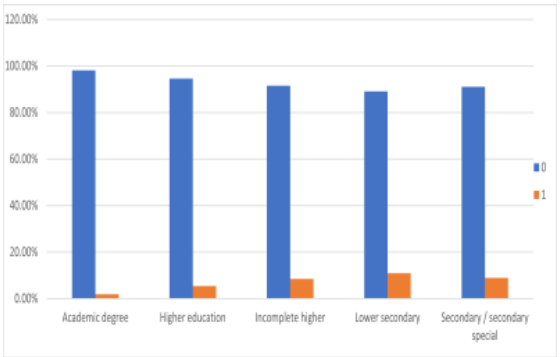
INCOME TYPE VS TARGET



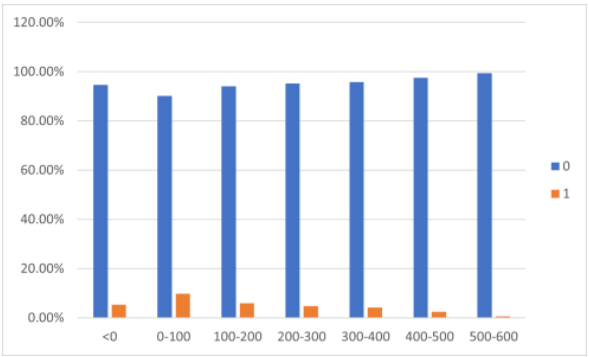
FAMILY MEMBER VS TARGET



EDUCATION TYPE VS TARGET



MONTHS EMPLOYED VS TARGET



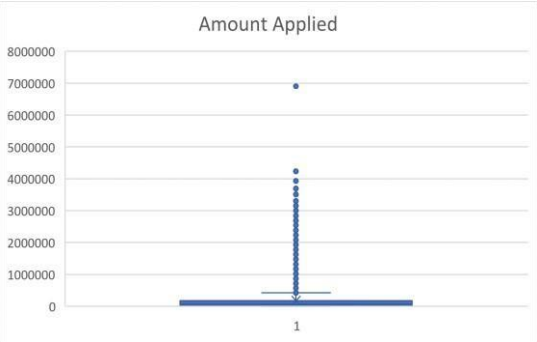
TASK 6 (FINDING TOP 10 CORRELATIONS):

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

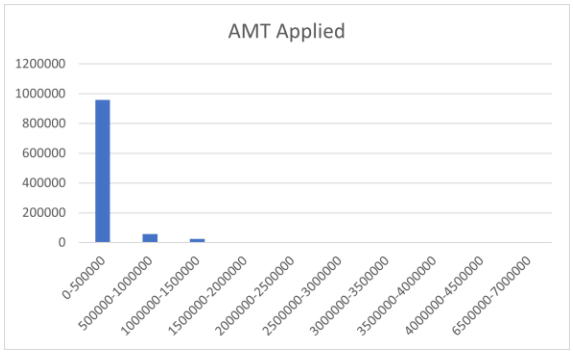
EARLIER APPLICATION.CSV

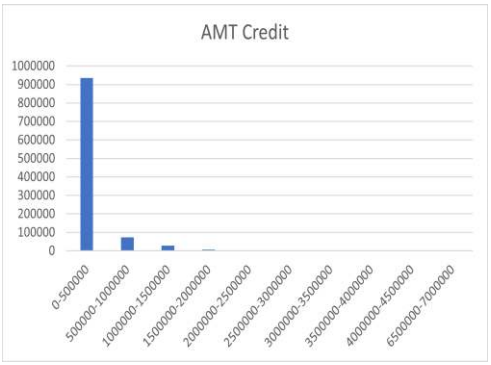
TASK 3 (FINDING OUTLIERS):



TASK 4(DATA IMBALANCE):

Below are the columns where data is unevenly distributed





TASK 5 (EDA):

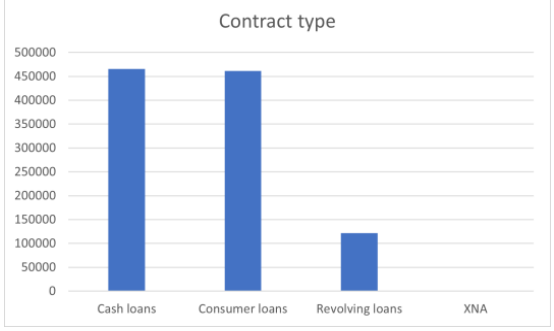
Univariate Analysis:

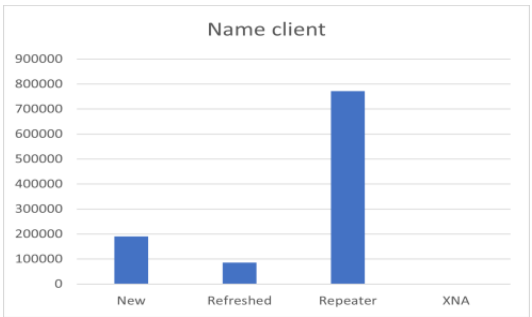
Inference

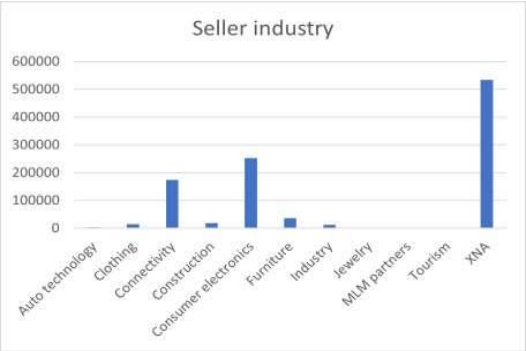
Based on the data analysis, the following observations were made regarding the loan preferences of customers:

1. Cash and consumer loans are the most popular choices among customers.
2. A significant portion of the customer base consists of repeat borrowers.
3. The majority of current loan applicants had previously applied for loans within the past ten months.
4. There is a higher demand for loans related to consumer gadgets.

These findings highlight the trends in customer loan preferences and can assist in understanding the loan portfolio composition.







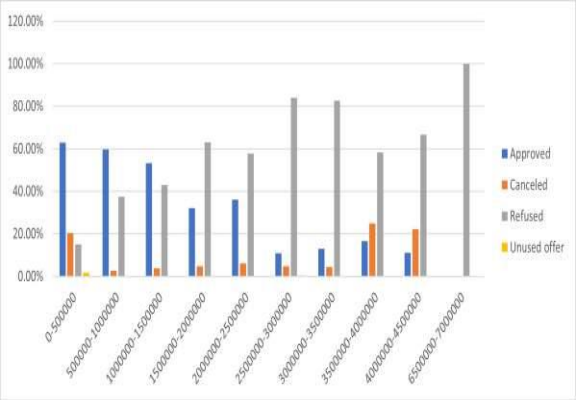
BIVARIATE ANALYSIS:

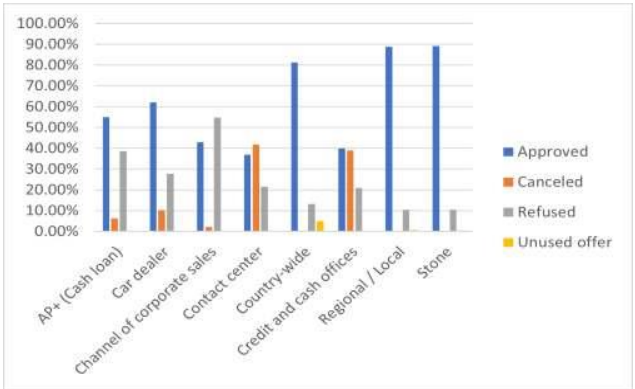
INFERENCE

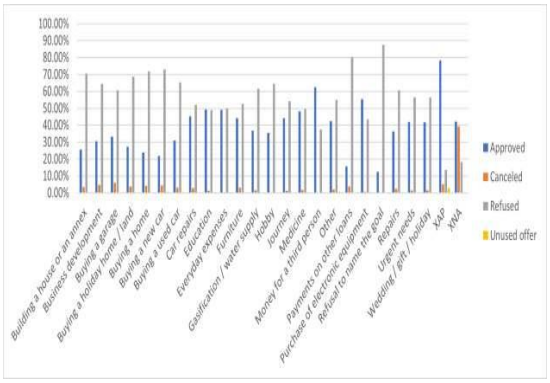
Based on the analysis, the following patterns and trends were observed in relation to loan approvals and denials:

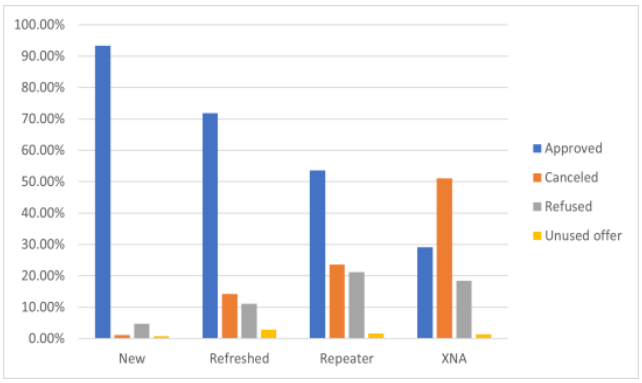
1. Loan applications for amounts exceeding Rs. 350,000 are more likely to be denied.
2. Loans sought through Credit and Cash agencies have a higher cancellation rate.
3. New clients have a higher approval rate for their loan applications.
4. Car loans have predominantly been denied thus far.
5. Loans granted to MLM partner clients are more likely to be cancelled.
6. Approximately 80% of the loans analysed were approved, with a consistent number of rejections.
7. Consumer loans have a very low cancellation rate and the highest approval rate.
8. Loans for the first Selling place area group experienced several cancellations.
9. Clients who apply for another loan within 10 months of their previous loan are more likely to have it cancelled.
10. Walk-in loans have a higher rate of refusal.

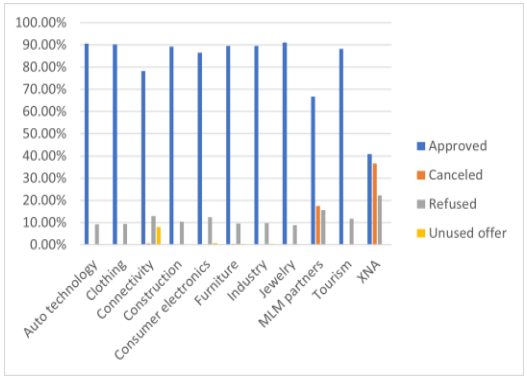
These insights provide valuable information on the loan approval and denial trends, which can be utilized for risk assessment and decision-making processes in the lending business.

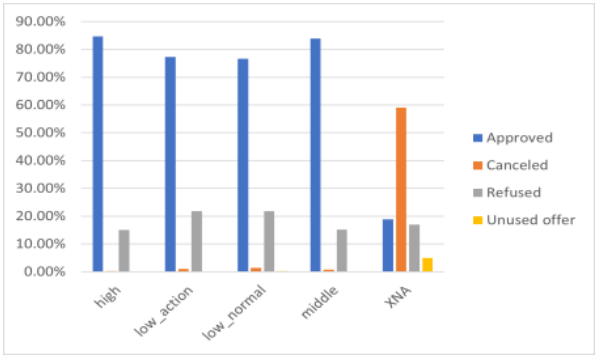


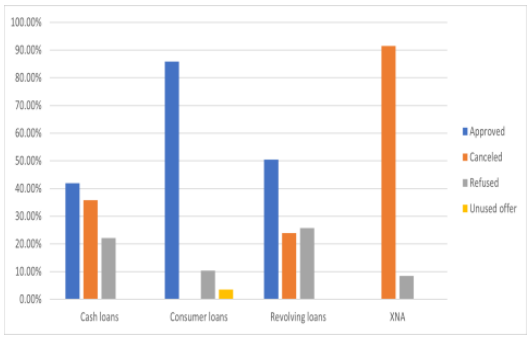


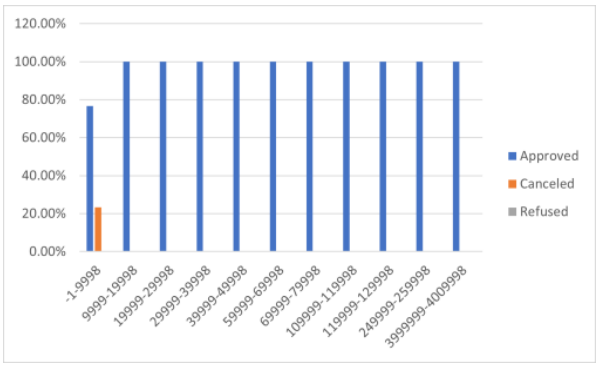

100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
Building a house or an annex
Business development
Buying a garage
Buying a holiday home / land
Buying a home
Buying a new car
Buying a used car
Car repairs
Education
Everyday expenses
Furniture
Gasification / water supply
Hobby
Journey
Medicine
Money for a third person
Other
Payments on other loans
Purchase of electronic equipment
Refusal to name the goal
Repairs
Urgent needs
Wedding / gift / holiday
XAP
XNA
Approved
Canceled
Refused
Unused offer

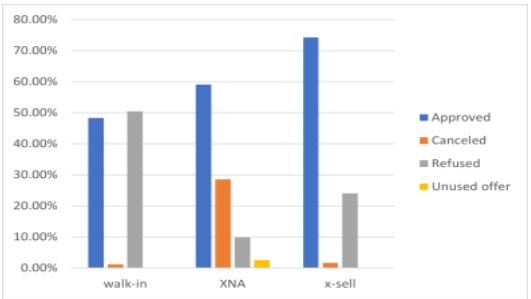

100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
New
Refreshed
Repeater
XNA
Approved
Canceled
Refused
Unused offer

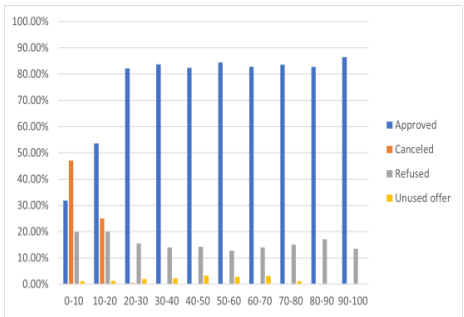

100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
Auto technology
Clothing
Connectivity
Construction
Consumer electronics
Furniture
Industry
Jewelry
MLM partners
Tourism
XNA
Approved
Canceled
Refused
Unused offer


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

TASK 6 (FINDING CORRELATIONS):

Top ten reasons for loan cancellation and refusal

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

TASK 7 (COMBINING TWO SHEETS):

I used MySQL to join the table

QUERY:

```
SELECT
TARGET,
SK_ID_CURR,
NAME_CONTRACT_TYPE,
AMT_APPLICATION,
NAME_CASH_LOAN_PURPOSE,
NAME_CONTRACT_STATUS,
NAME_CLIENT_TYPE,
DAYS_DECISION,
CODE_REJECT_REASON,
NAME_SELLER_INDUSTRY,
NAME_PORTFOLIO,
NAME_PRODUCT_TYPE,
```

CHANNEL_TYPE,

SELLERPLACE_AREA,

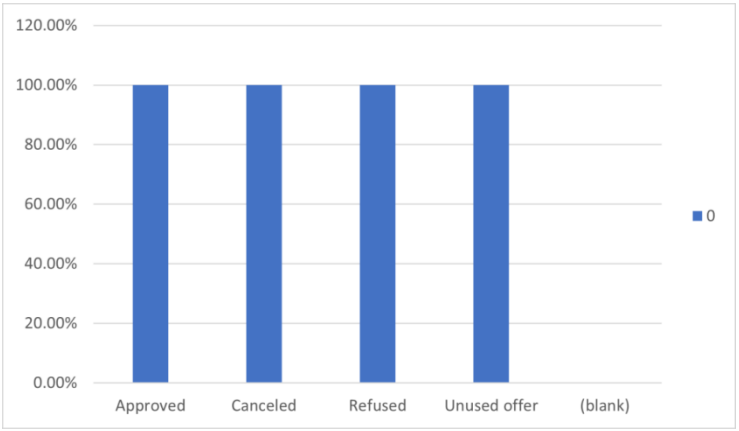
NAME_YIELD_GROUP,

PRODUCT_COMBINATION

FROM APPLICATION_DATA

JOIN PREVIOUS_APPLICATION ON SK_ID_CURR;

PIVOT TABLE ANALYSIS



Clients who have applied for previous loans have no defaults in current loans

SUMMARY:

LOAN- HIGHLY RECOMMENDED GROUPS	LOAN- HIGH RISK GROUPS
1. Previous application approved clients 2. Married clients 3. Senior clients 4. More educated clients 5. Customers with a High Income 6.Clients with a greater external source 7. Females 8.Customers with strong work experience	1. Clients that are unemployed 2. Youth clients 3. Customers whose prior applications were denied 4. Low-income clientele 5. Clients with insufficient external sources 6.Customers with little work experience 7.Customers on Maternity Leave 8. Clients with a larger number of family members