DA Project 2:
EDA Credit Case Study

Swornim Shrestha (28/05/2023)

Table of Contents

Problem Statement

Understanding the Dataset

Analysis Approach:

Analysis Findings:

Conclusion:

Problem Statement:

- Risk of Loan Defaults: Insufficient credit history poses challenges for loan providers, increasing the risk of loan defaults.
- Importance of Minimizing Risk: Minimizing defaults is crucial for the profitability and sustainability of lending operations.
- Objective: Perform EDA on loan application data to identify the driving factors behind loan defaults.
- Actionable Insights: EDA provides actionable insights for informed decision-making, enabling adjustments in loan amounts, interest rates, and loan approvals to minimize defaults and maximize profitability.

Understanding the Dataset

DATASET USED FOR THE ANALYSIS WERE: APPLICATION_DATA.CSV AND PREVIOUS_APPLICATION.CSV

TO UNDERSTAND THE VARIABLE , DATA DICTIONARY FILE COLUMNS_DESCRIPTION.CSV FILE WAS USED.

Analysis Approach:

- Data was Cleaned and missing values were handled.
- Errors in data types was also handled.
- Outliers were analysed.

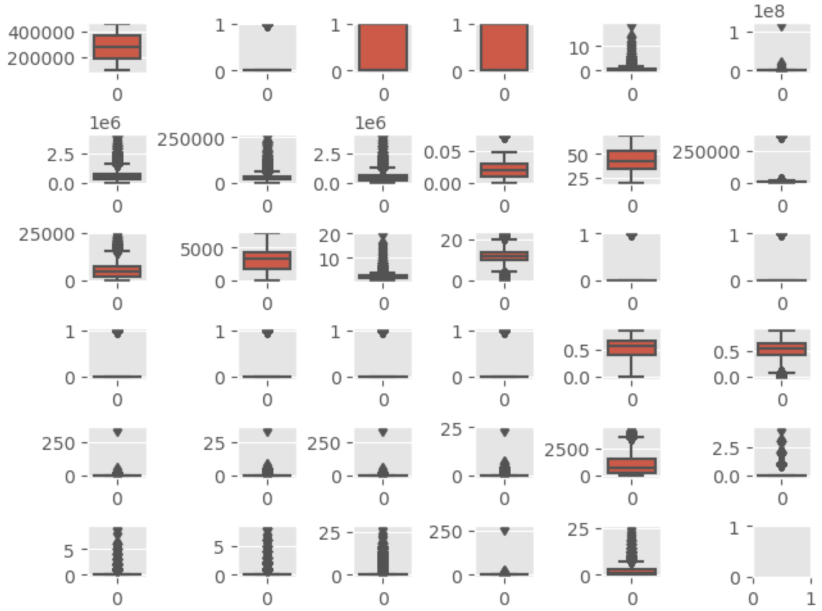


Fig: Outliers analysis.

- Data Imbalance:It was done based on target(1 and 0) which are defaulter(having payment difficulties) and non-defaulter(other difficulties).

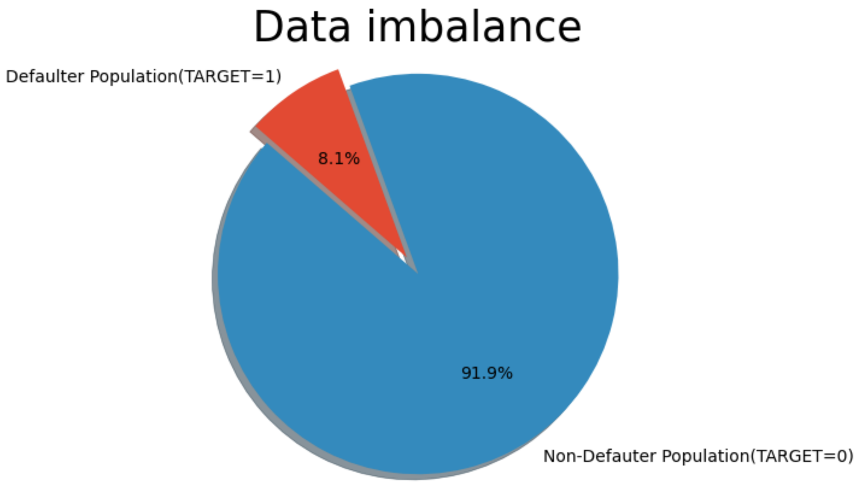
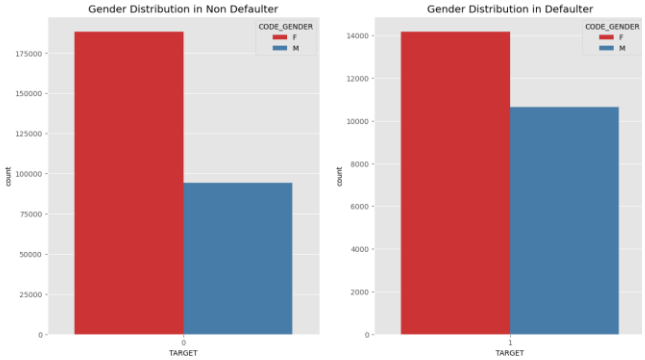


Fig: Target0 vs Target1

Univariate Analysis: Categorical:

Based on Gender



Gender Distribution in Non Defaulter

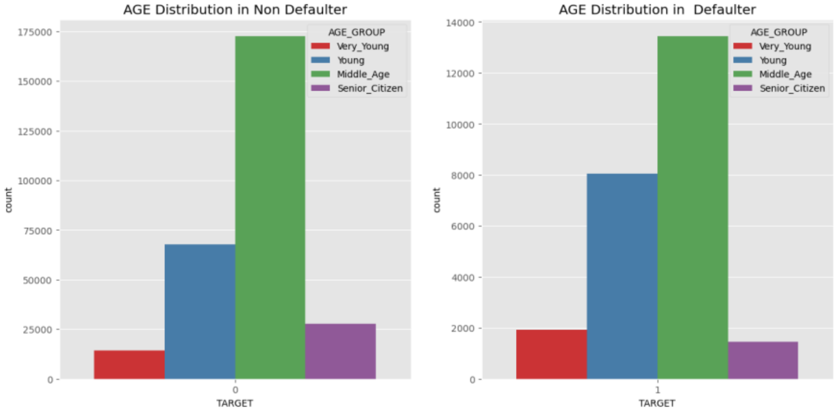
Gender Distribution in Defaulter

- It seems like Female clients applied higher than male clients for loan.

- 66.6% Female clients and 33.4% male clients are payment difficulties.

- 57% Female clients and 42% male clients are with payment difficulties.

Based on Age



AGE Distribution in Non Defaulter

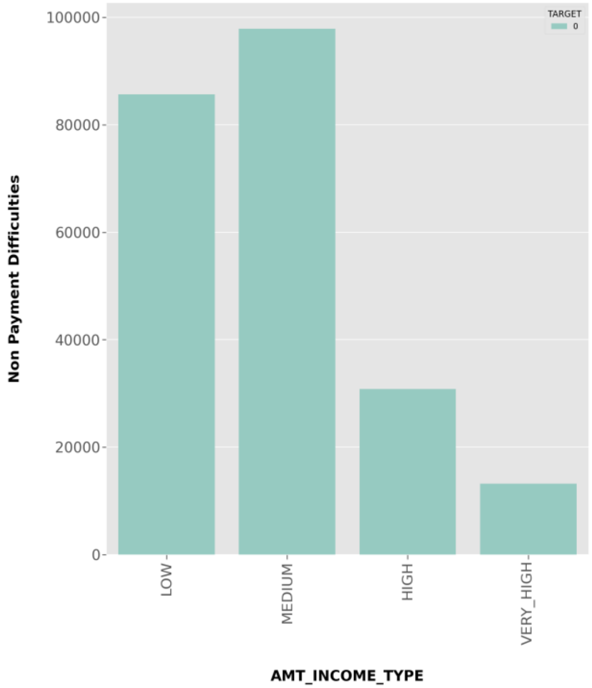
AGE Distribution in Defaulter

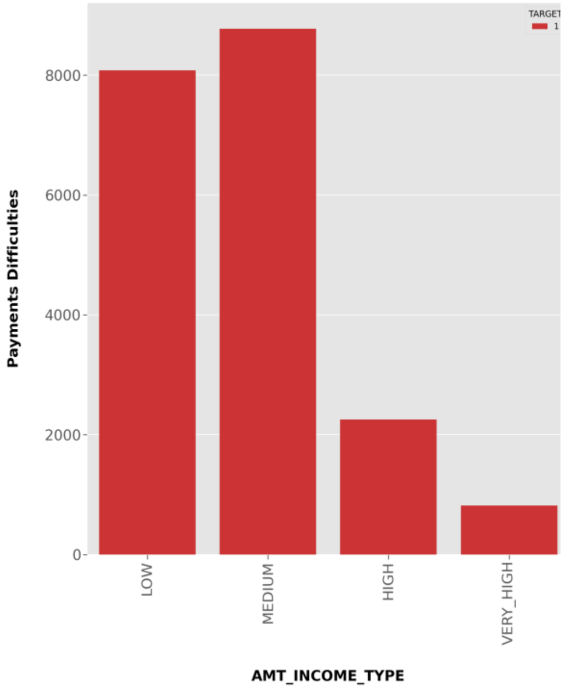
- Middle-age group dominated in the number of applying loans for both cases of defaulters and non-defaulters.

- This domination shows the payment difficulties faced by middle age group and young people group rank in second with more than 8000 of loans with payment difficulties.

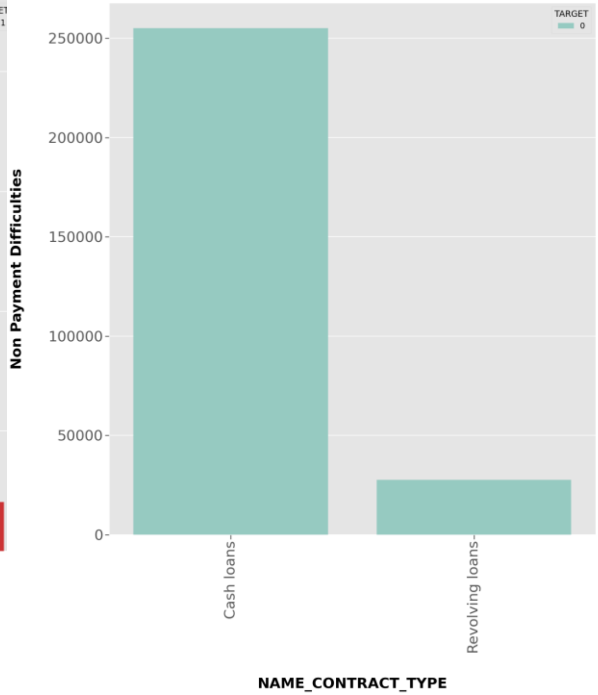
Contract/Income Type with respect to Target variables

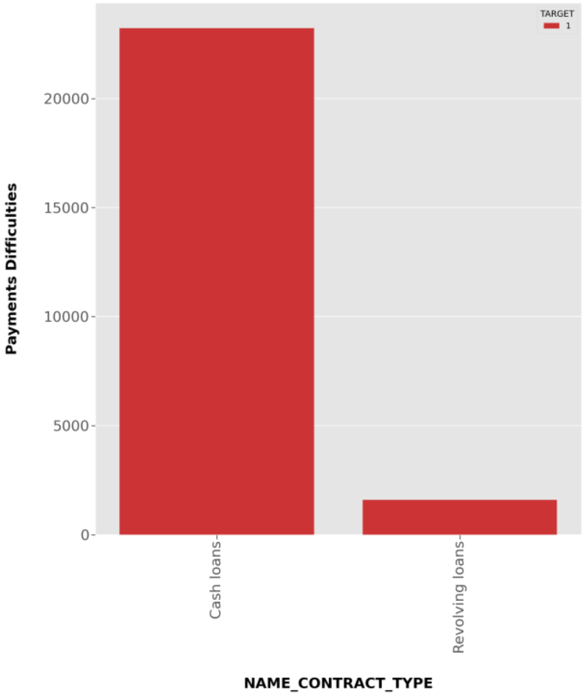
Based on Income type





Based on Contract type

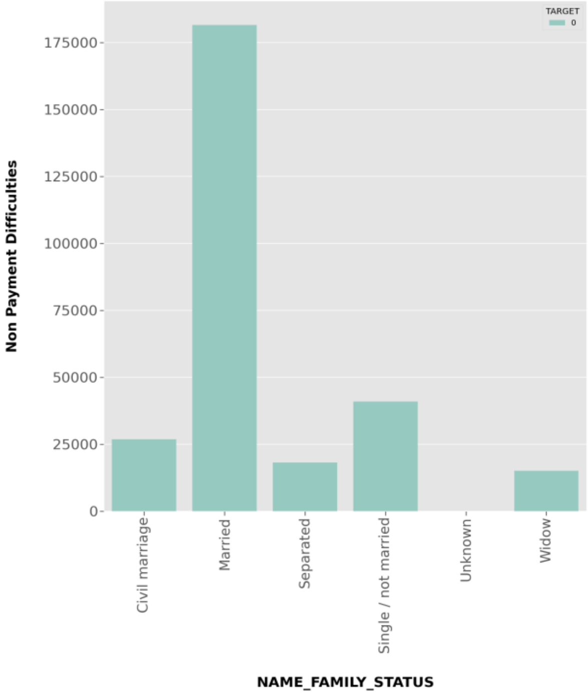


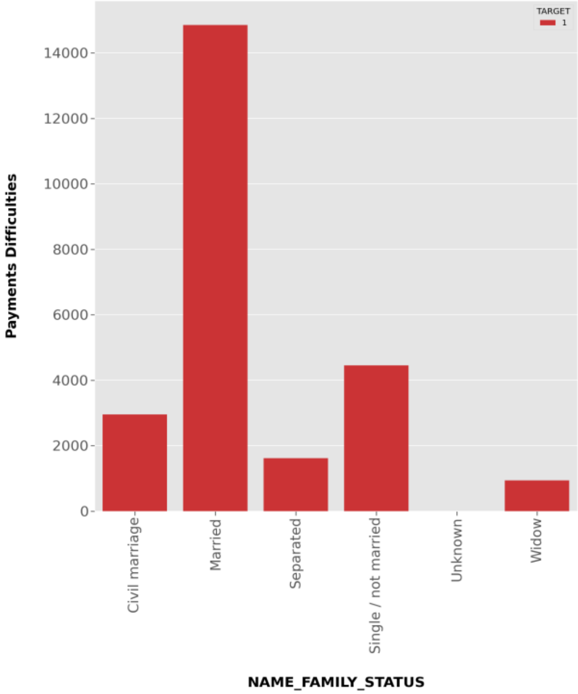


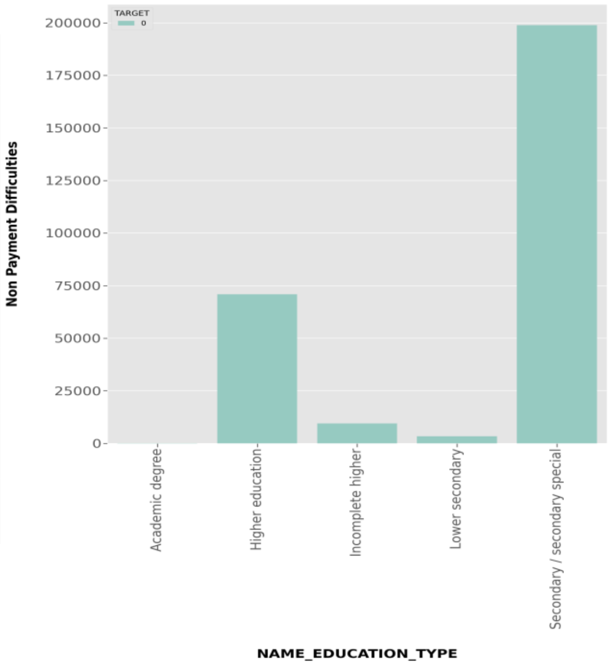
- Low earners and medium earners have applied for the greatest number of loans with or without payment difficulties and as expected there is lowest number of very high earners.

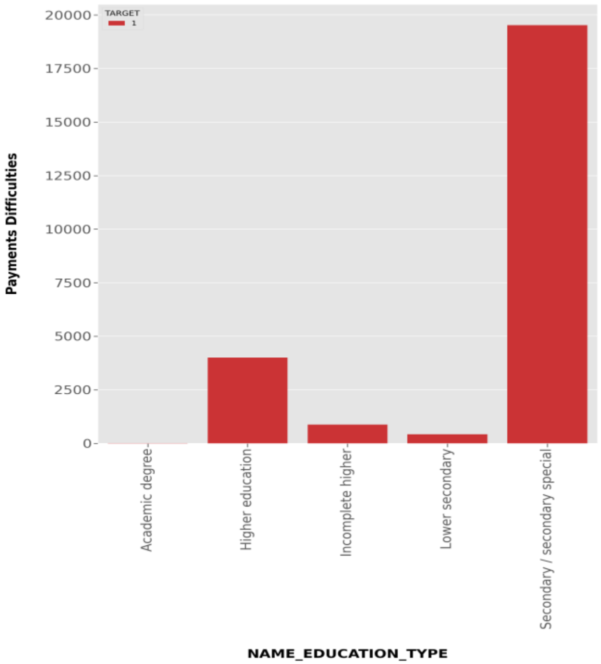
- We can see the bulk of cash loans and small number of Resolving loan for both cases.

Education/Family type with respect to Target Variable:





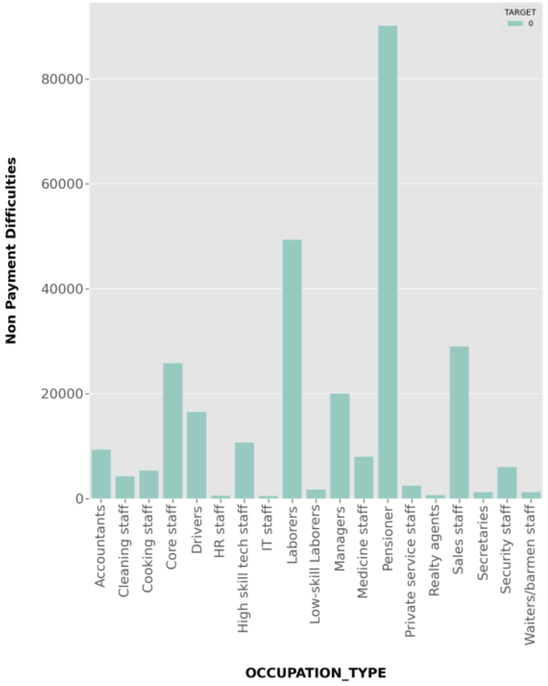


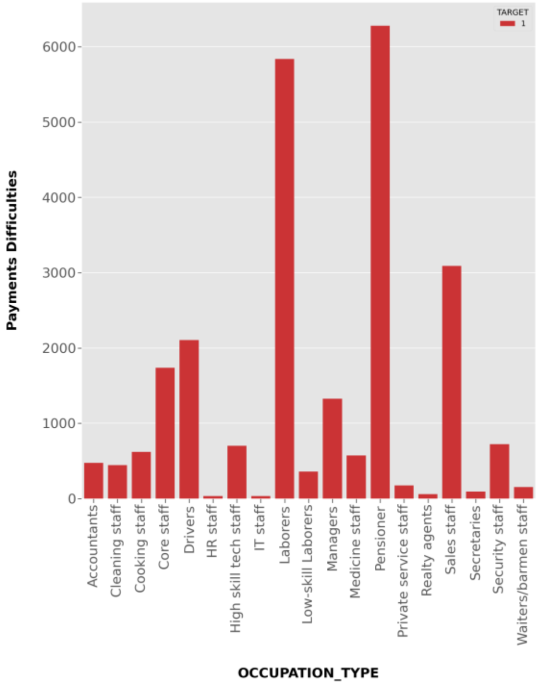


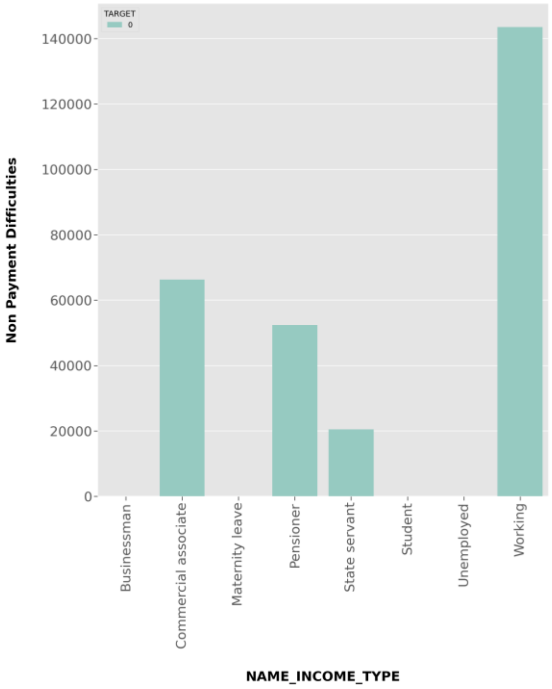
- Married people seems to have applied for higher number of loans both due to payment difficulties and other cases whereas single peoples have struggled with payment difficulties resulting in applying loans.

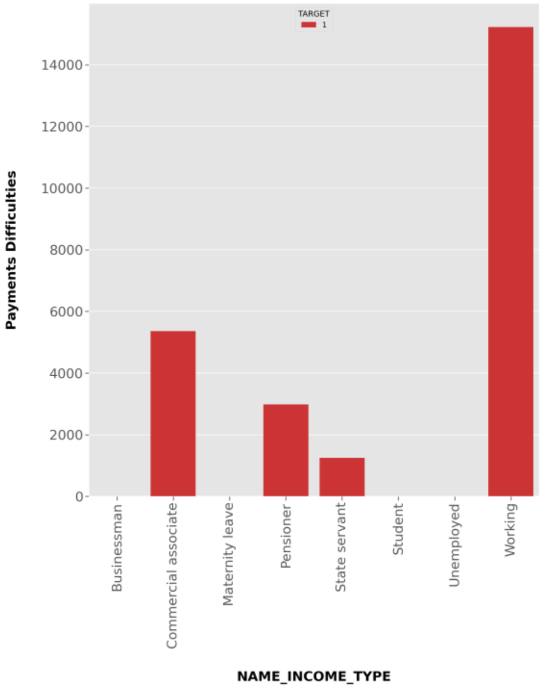
- Likewise, for both cases, people with secondary/secondary special level of education has applied for the loans with people around 75 thousand applying for loan with non-payment difficulties for higher educated people and less than 5 thousand of higher educated people applying for loans due to payment difficulties.

Occupation/Income Range with respect to Target variables





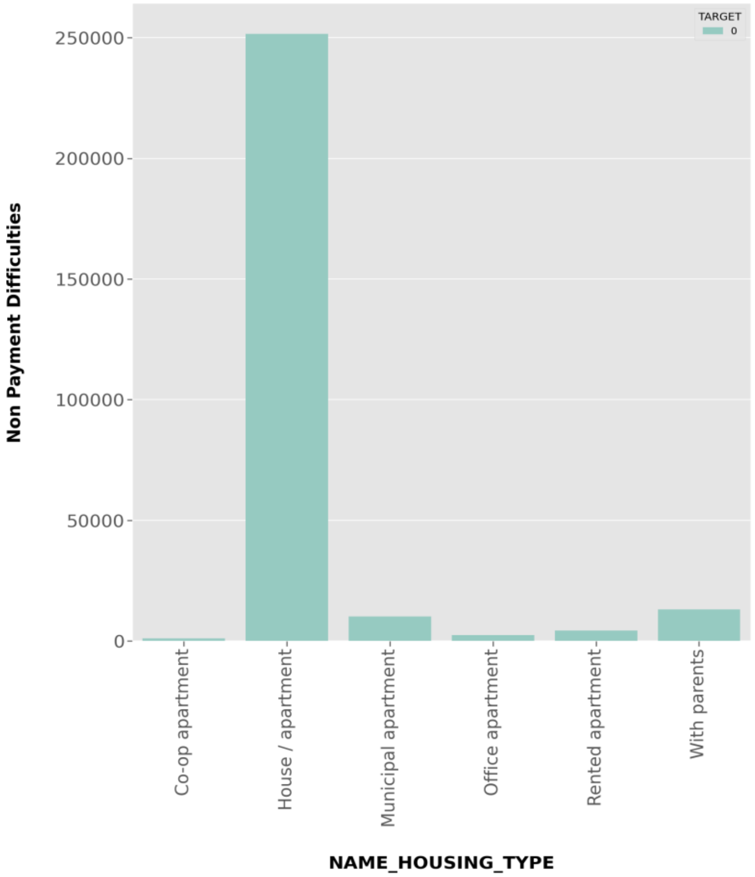


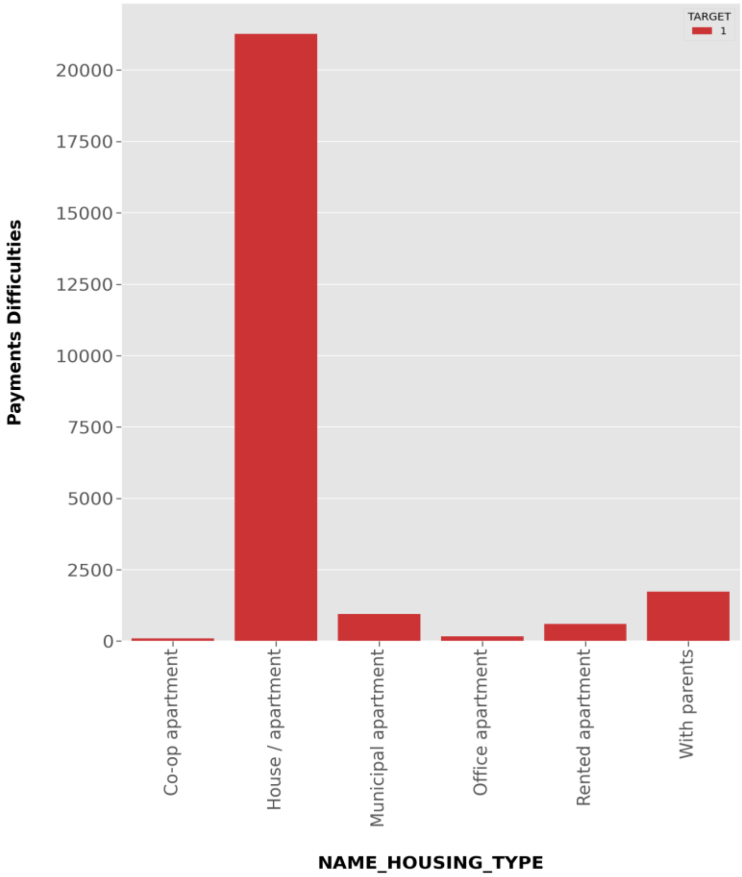


Pensioner seems to have loans more than 80 thousand due to non-payment difficulties whereas laborer fall on second whereas for payment difficulty, we can see sales staff applying for more with least number from IT staff.

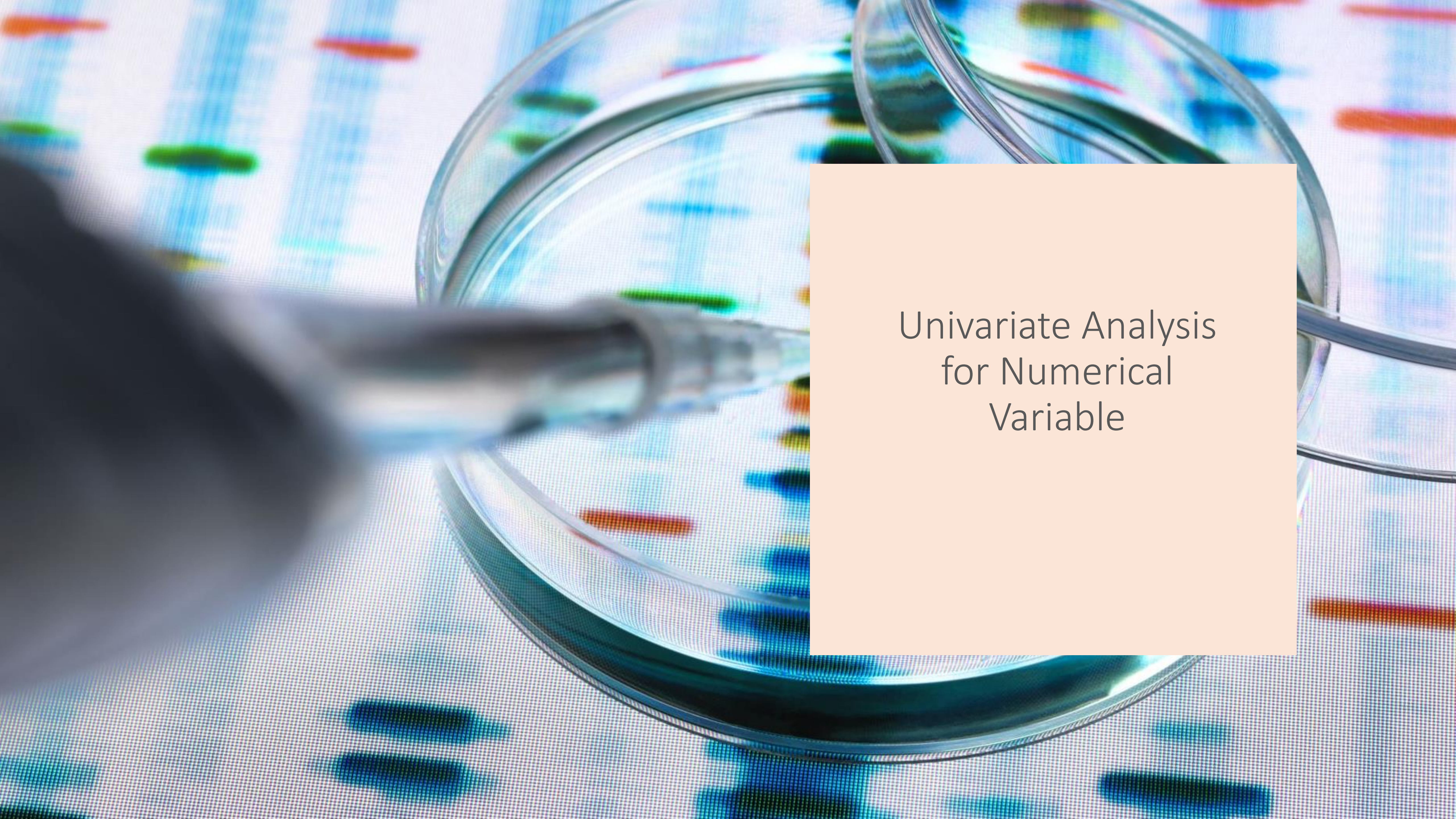
For both defaulters and non-defaulters, we can see the working class having more number of loans and second one being commercial associate but there is huge gap between working class and commercial associate whereas there is no loan for students and unemployed.

Housing Type/Credit Range with respect to Target variables

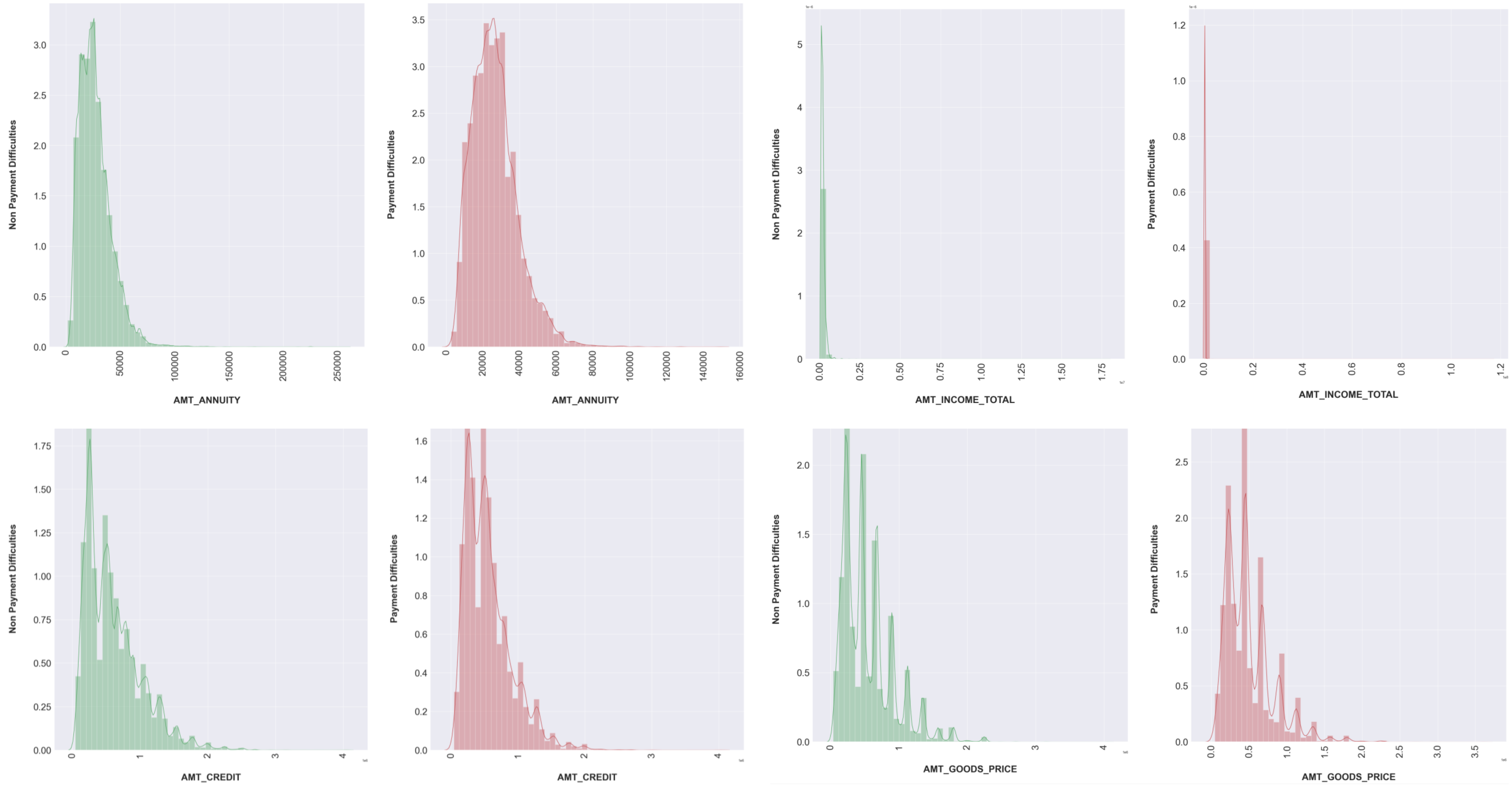




- As of housing/apartment prices rising every year, it is particularly an easy guess on this category being in major number for applying loans due to both cases.



Univariate Analysis for Numerical Variable



AMT_Annulity vs Income total vs Credit vs Goods price based on Target Variable

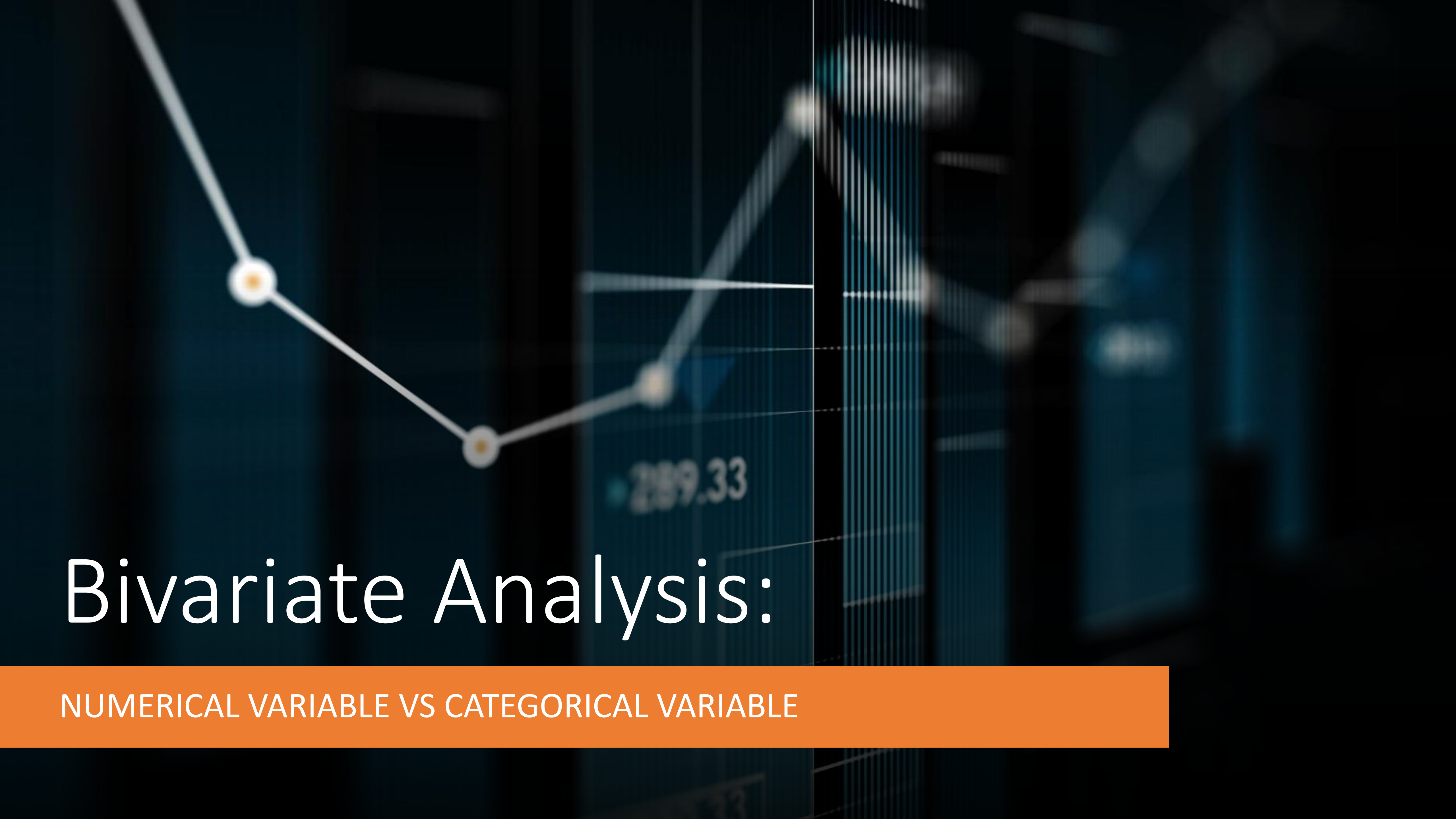
Findings from Univariate Analysis:

The distribution plots reveal that individuals with Target 1, indicating payment difficulties, exhibit a more varied income distribution compared to those with Target 0.

The distribution shapes of Income Total, Annuity, Credit, and Good Price are similar for Target 0, while they are also similar for Target 1.

Additionally, the plots emphasize that individuals facing loan repayment difficulties are associated with their income level, loan amount, price of goods purchased with the loan, and annuity.

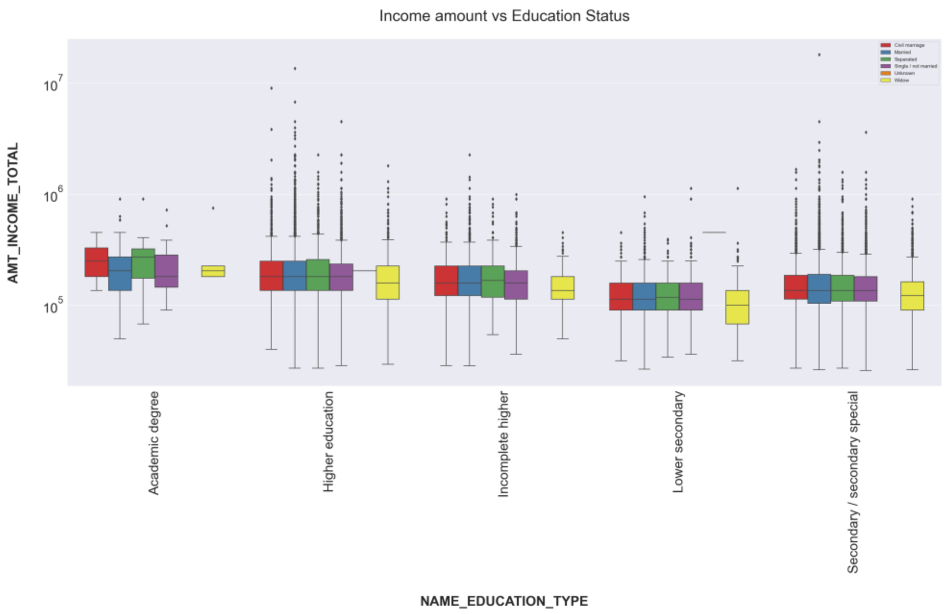
Furthermore, the distribution plots indicate that the curve shape is broader for Target 1, indicating payment difficulties, whereas it is narrower with well-defined edges for Target 0.



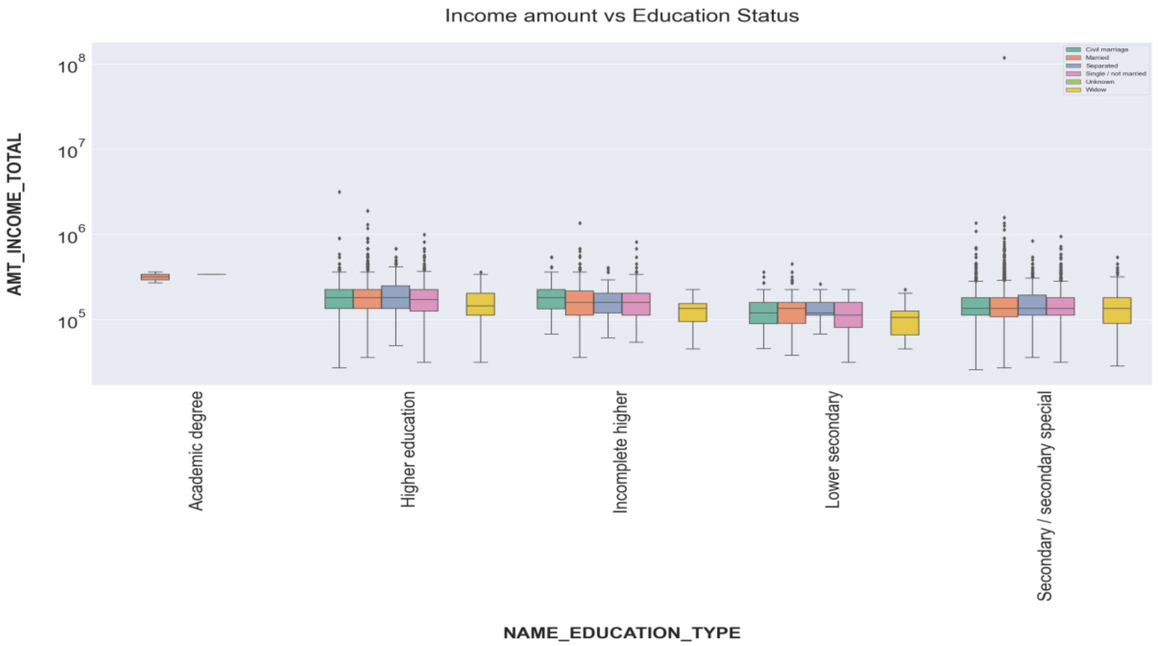
Bivariate Analysis:

NUMERICAL VARIABLE VS CATEGORICAL VARIABLE

Income_Amount Vs Education_Status Vs FAMILY_Status among Target Variables



- Income of the clients with all type of family status having rest of the education type lie Below the First quartile i.e., 25%
- we can see the highest number of outliers for clients with secondary/secondary level of education.
- Higher education background has resulted in high income earner with exception from education background of secondary/secondary special.
- Married clients with good educational background are dominating the high-income earner market.



- we can see the domination of the married client with education background(higher to secondary education) earning more income than others.
- comparing the defaulters, we can clearly see the income flow difference between them.

Credit Vs Education_Status Vs FAMILY_Status among Target Variables

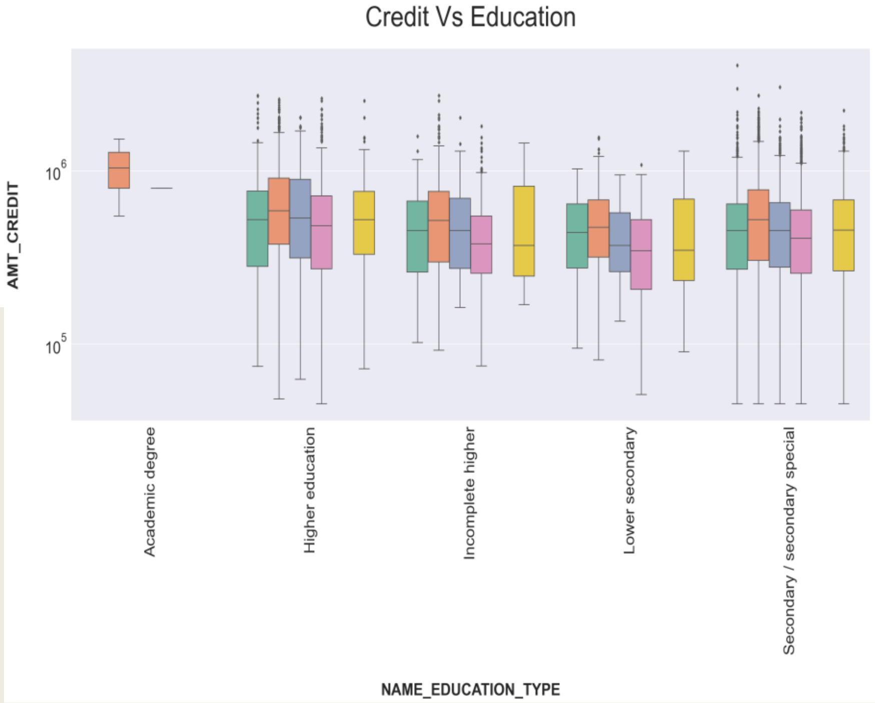


Fig: Target0

- Some of the clients with Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special Education are more likely to take high amount of credit loan.

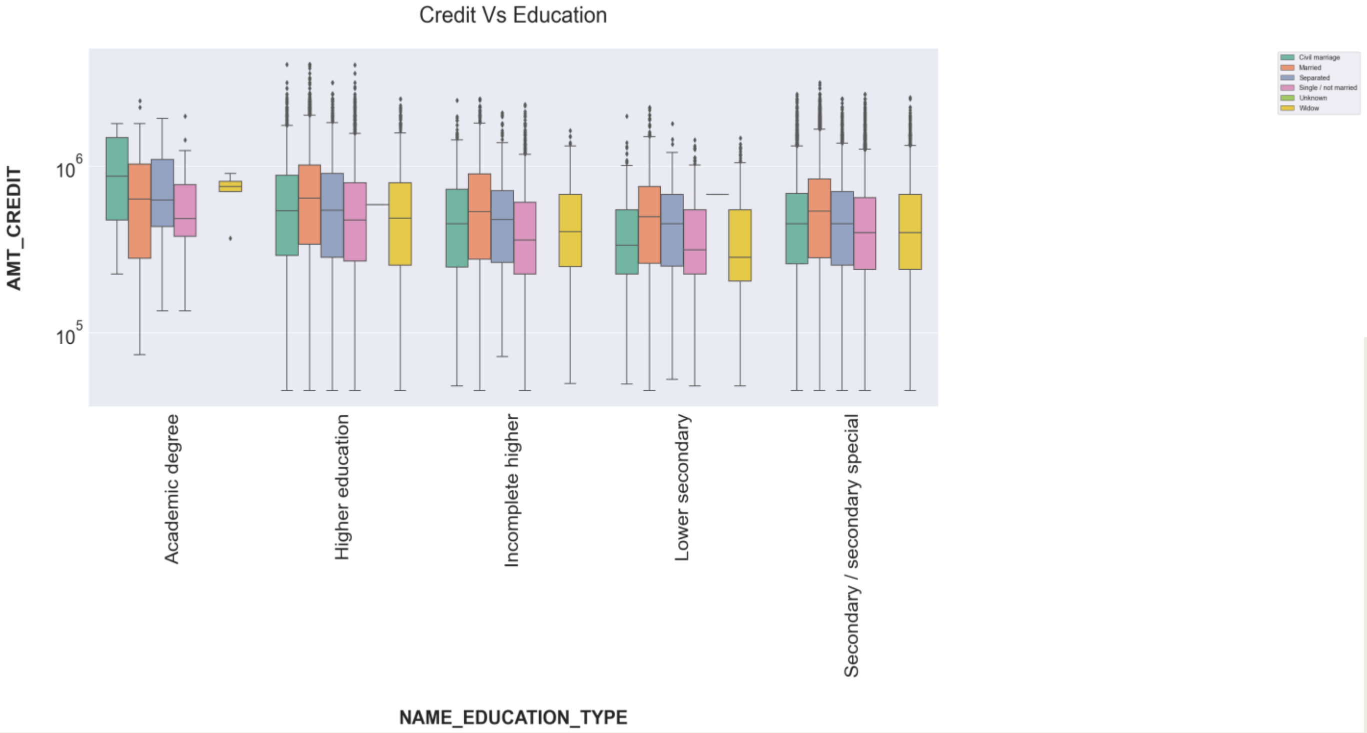


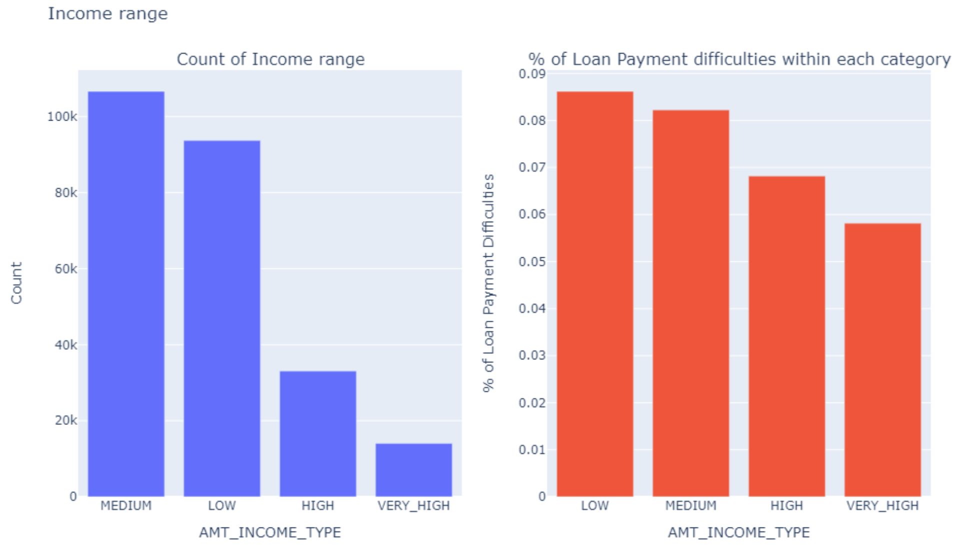
Fig: Target 1

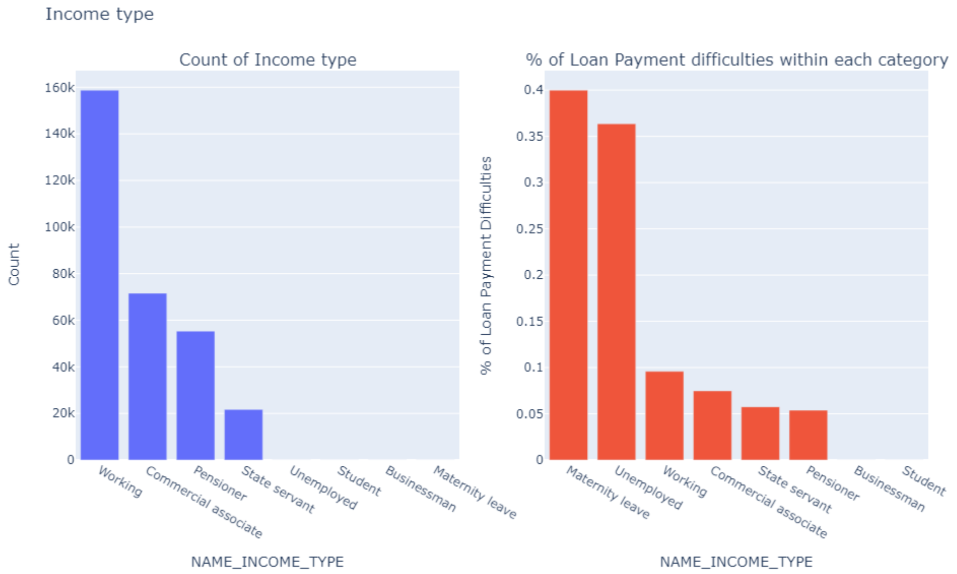
- Clients with higher education have large number of outliers and majority of credit amount lies below 25%.
- Married clients again are on the top list of taking loans regardless less of their education background.

Bivariate Analysis:

Categorical vs Categorical Variable

Categories having maximum % of Risks of default using biplot.



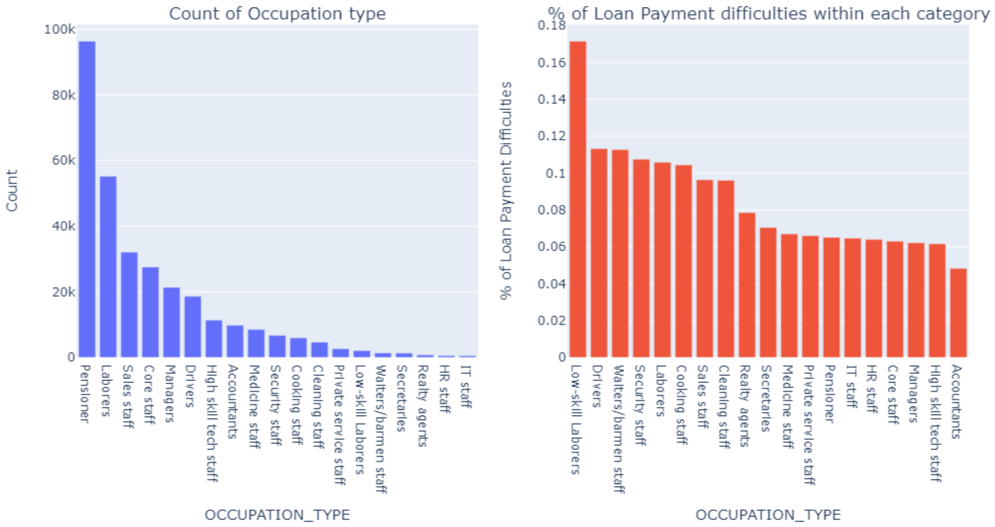


Low and Middle earning class has struggled heavily with paying the payment of the low and can see max number of those people applying for loan.

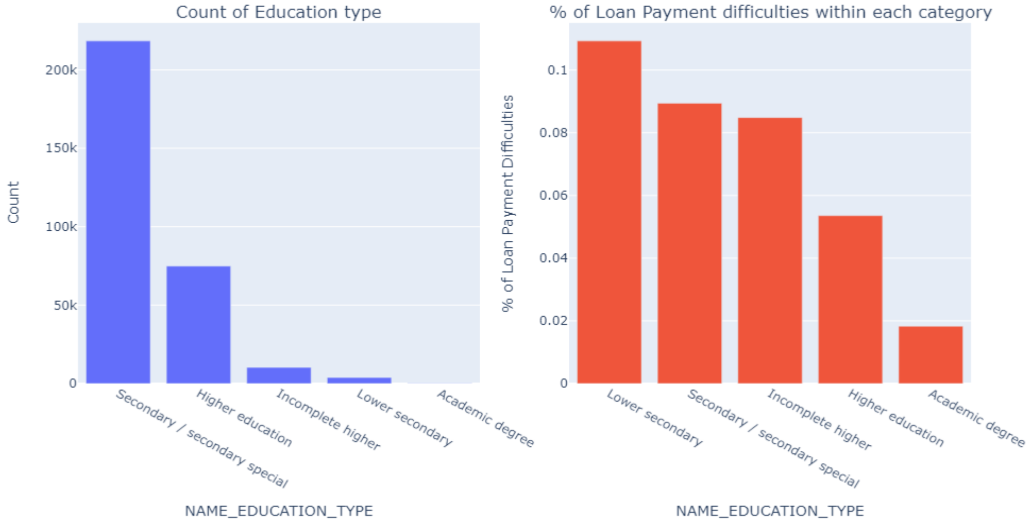
Even though Working- class people having highest number of loans, they are comparatively low on the payment of the loan.

Maternity leave people are at high level of risk along with unemployed people though can see a vey less amount of loans for this group.

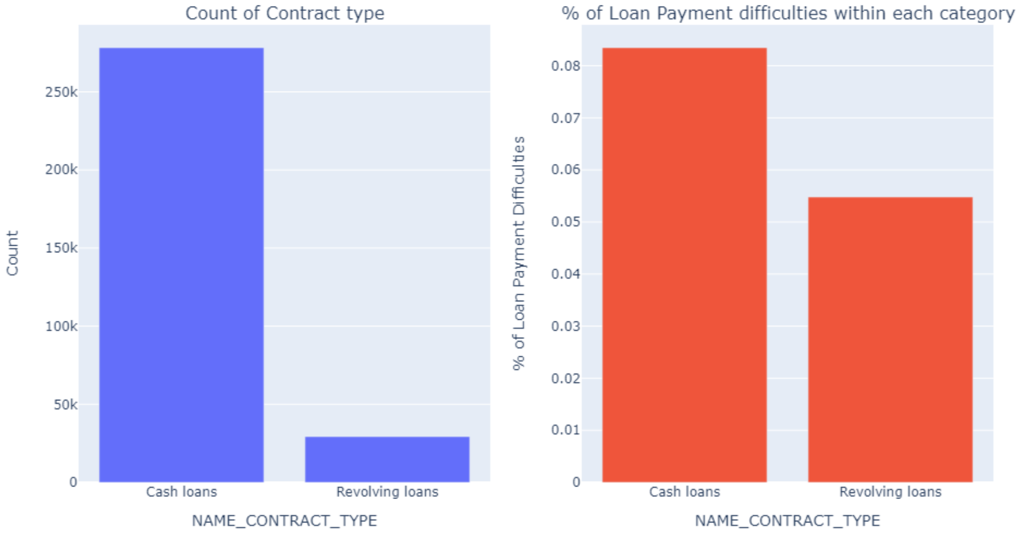
Occupation type



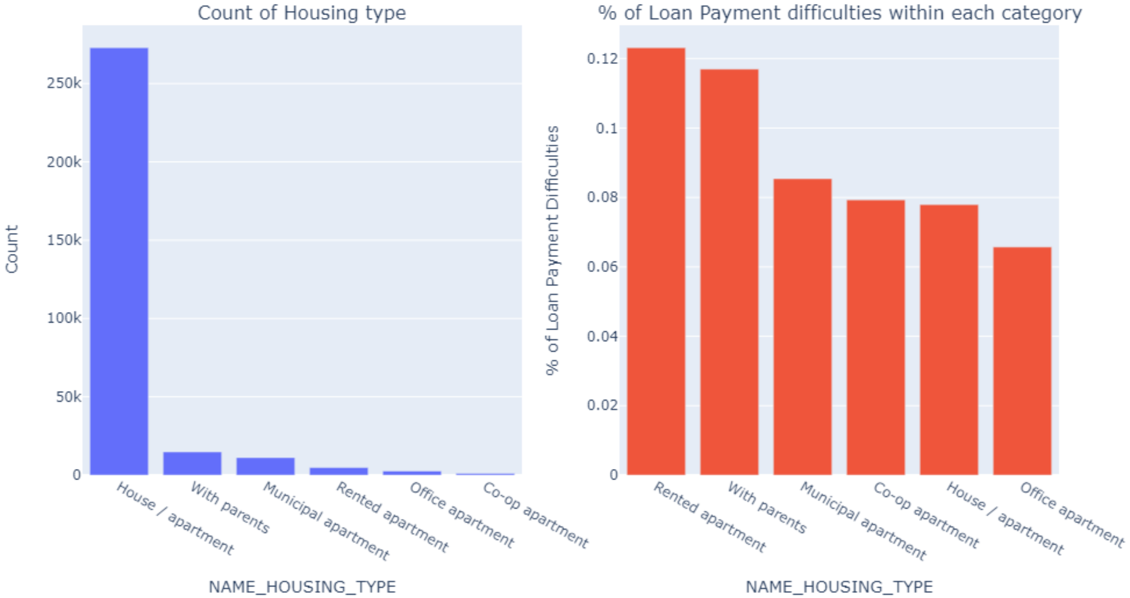
Education type



Contract type



Housing type



Findings from Analysis:

- Female clients with Academic degree and high-income type have higher risk to default whereas Male clients with Secondary/Secondary Special Education having all types of salaries have higher risk to default.
- Cash loan are higher in number and also has a highest percentage of risk compared to resolving loans.
- This shows us the difficulty for lower secondary level of education clients which is higher than secondary/secondary specail level.
- Clients having Academic Degree and higher Education have lower risk to default.
- Clients having Lower Secondary , Secondary/Secondary Special Education have very high risk to default.

Correlation between numerical variables using Heat Maps.

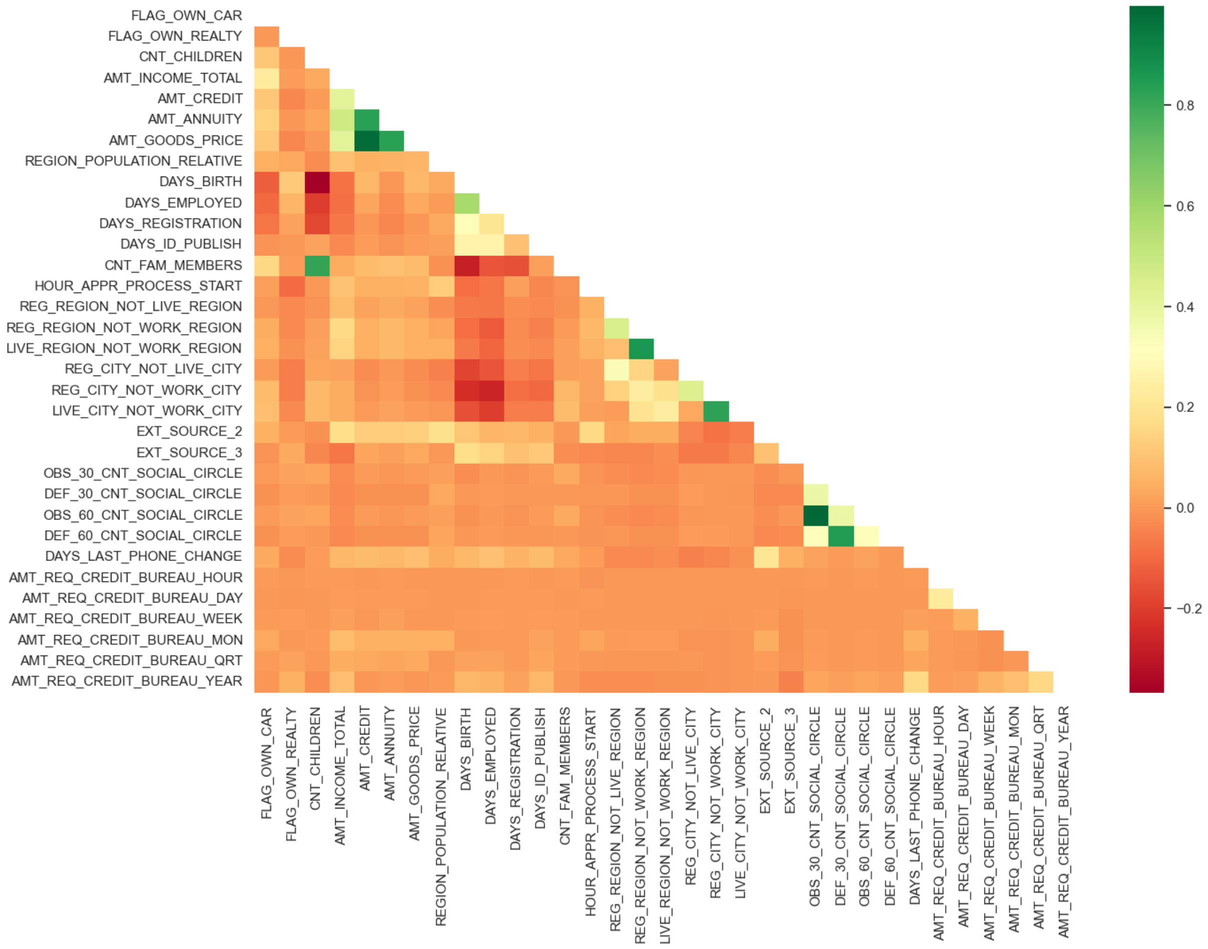


Fig: For target 0

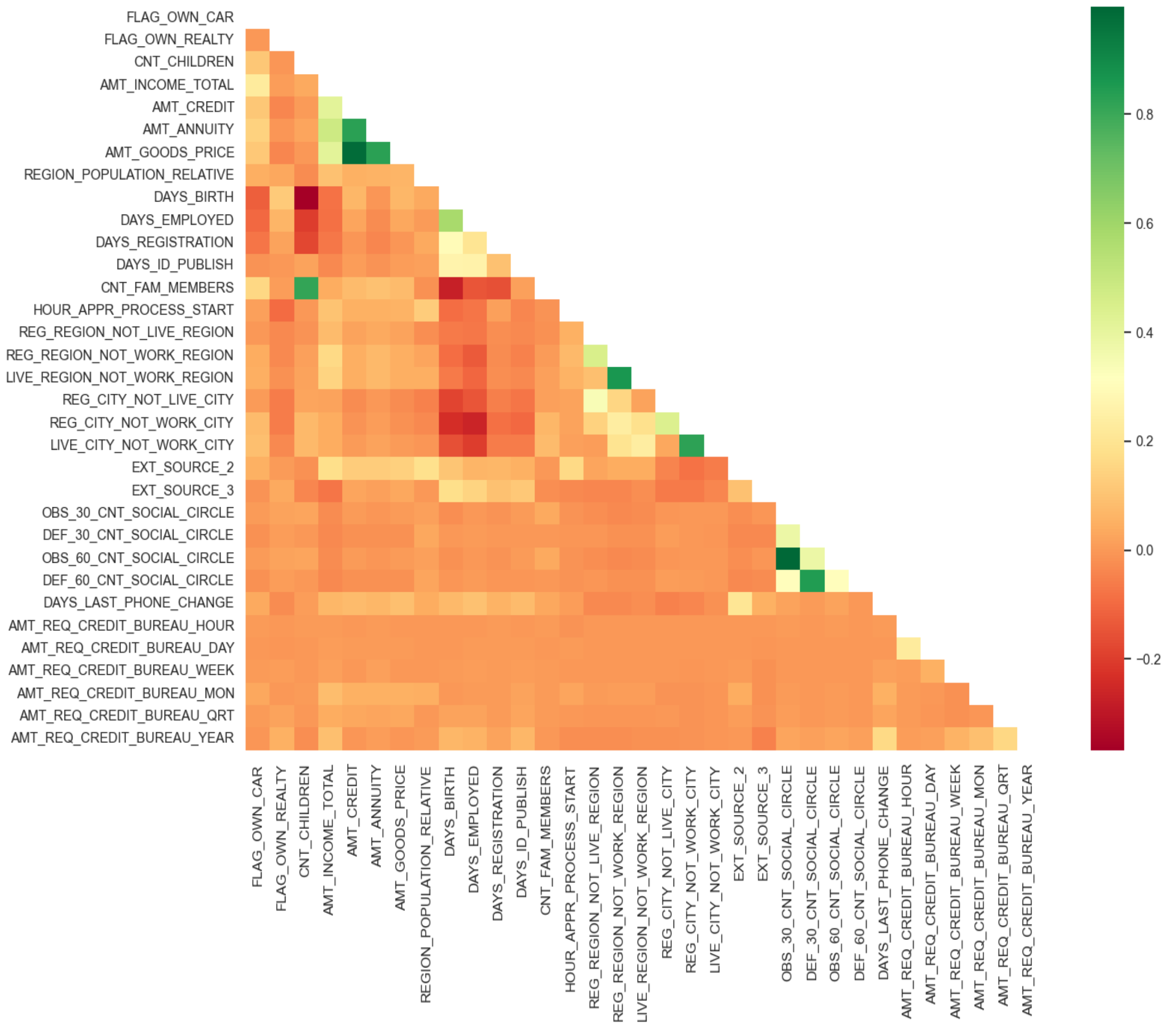


Fig: for Target 1

Insights on Heat Maps

For Target0

€ AMT_CREDIT is higher to densely populated area and AMT_INCOME_TOTAL is also higher in densely populated area.

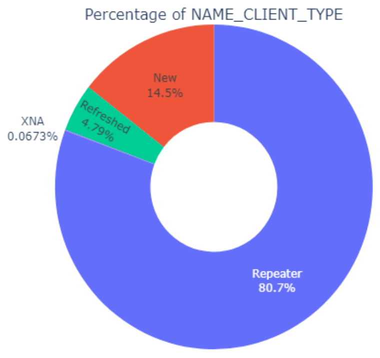
AMT_CREDIT is inversely proportional to DAYS_BIRTH and CNT_CHILDREN.

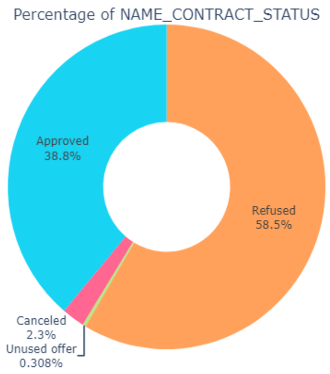
For Target1

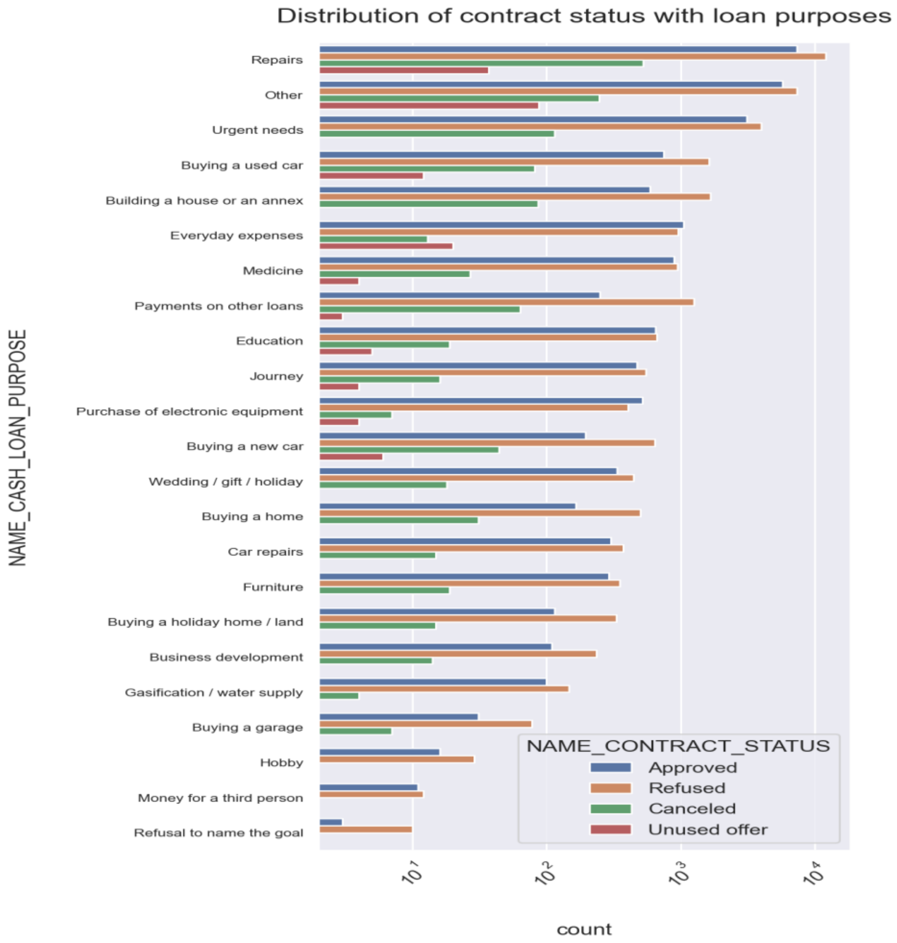
- There is an inverse relationship between AMT_CREDIT and DAYS_BIRTH, indicating that individuals in the lower age group tend to have higher credit amounts, and vice versa.
- AMT_CREDIT is inversely proportional to CNT_CHILDREN, meaning that clients with fewer children have higher credit amounts, while clients with more children have lower credit amounts.
- AMT_INCOME_TOTAL is inversely proportional to CNT_CHILDREN, indicating that clients with fewer children have higher income, while clients with more children have lower income.
- Clients in densely populated areas tend to have fewer children.
- AMT_CREDIT is higher in densely populated areas, suggesting that individuals in these areas are taking higher credit amounts.
- AMT_INCOME_TOTAL is also higher in densely populated areas, indicating that individuals in these areas have higher incomes.

Loan Distribution and Purposes

Percentage of `NAME_CONTRACT_STATUS` and `NAME_CLIENT_TYPE` and contract status vs Loan Purposes







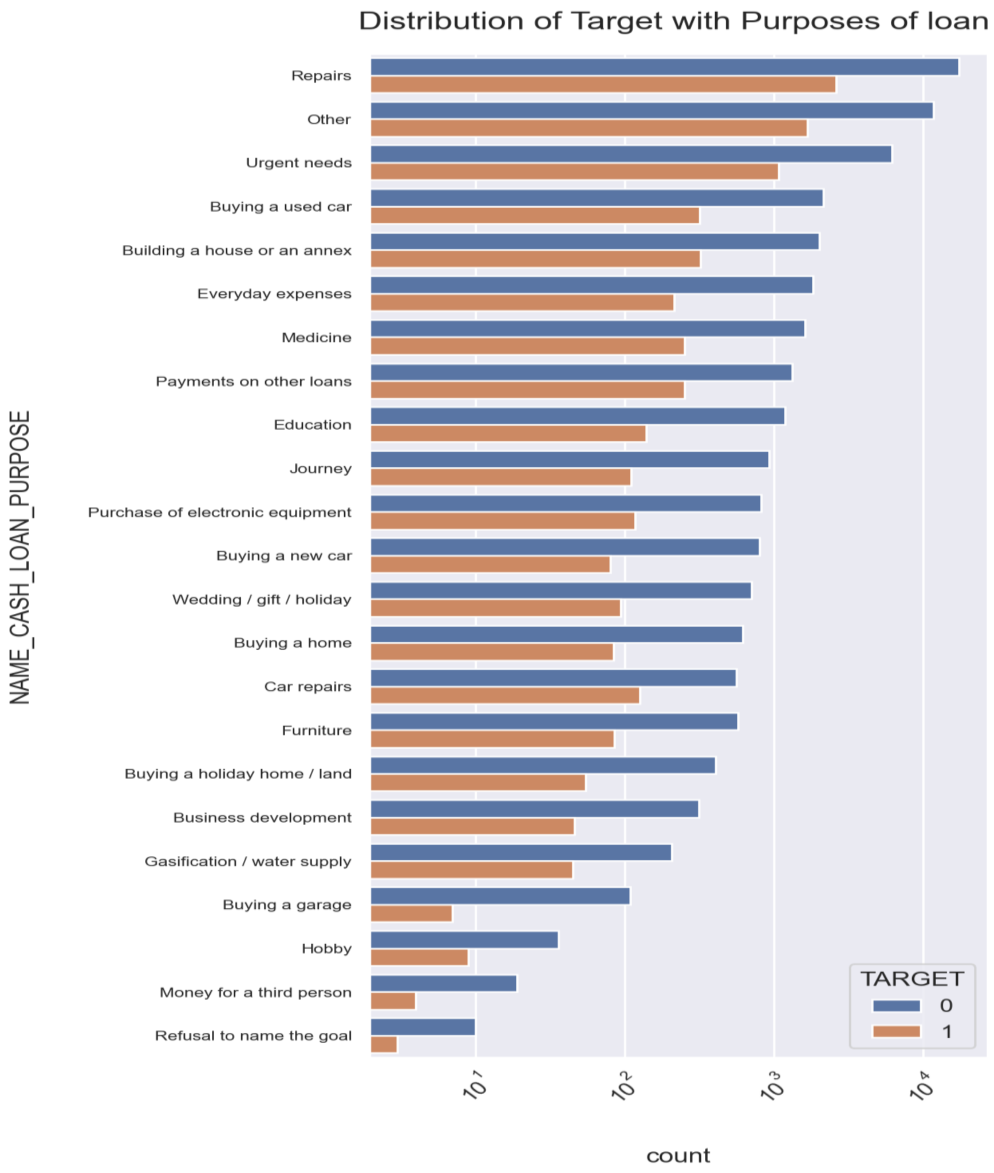
NAME_CLIENT_TYPE: around 80% were repeaters while 14.5% were new and for NAME_CLIENT_STATUS: 58% was refused and 2.3 % was cancelled.

Most rejection has come from Repair purposes but also has higher number of loans approved.

Buying house or a used car had also resulted in higher number of refusal.

Distribution of TARGET Vs LOAN_PURPOSE

- Variation for Repairs is the highest for both cases and we can see people using loans for Medicine had struggled to payback.



Key Findings:

Payment difficulties: A decrease in the percentage of payment difficulties among pensioners, while an increase among working individuals.

Loan payment difficulties: Decrease in the percentage of married and widowed individuals, increase in single and civil married individuals.

Educational qualifications: Increase in payment difficulties for secondary/secondary special qualifications and decrease for higher education qualifications.

"Low skilled Laborers": Relatively low count, but highest percentage of payment difficulties (17%).

"Lower Secondary" education type: Relatively low count, but highest percentage of payment difficulties (11%).

Contract types: Banks should focus more on clients with contract types Student, pensioner, and businessman.

Housing types: Target clients with housing types other than Co-op apartment or Office apartment for successful payments.

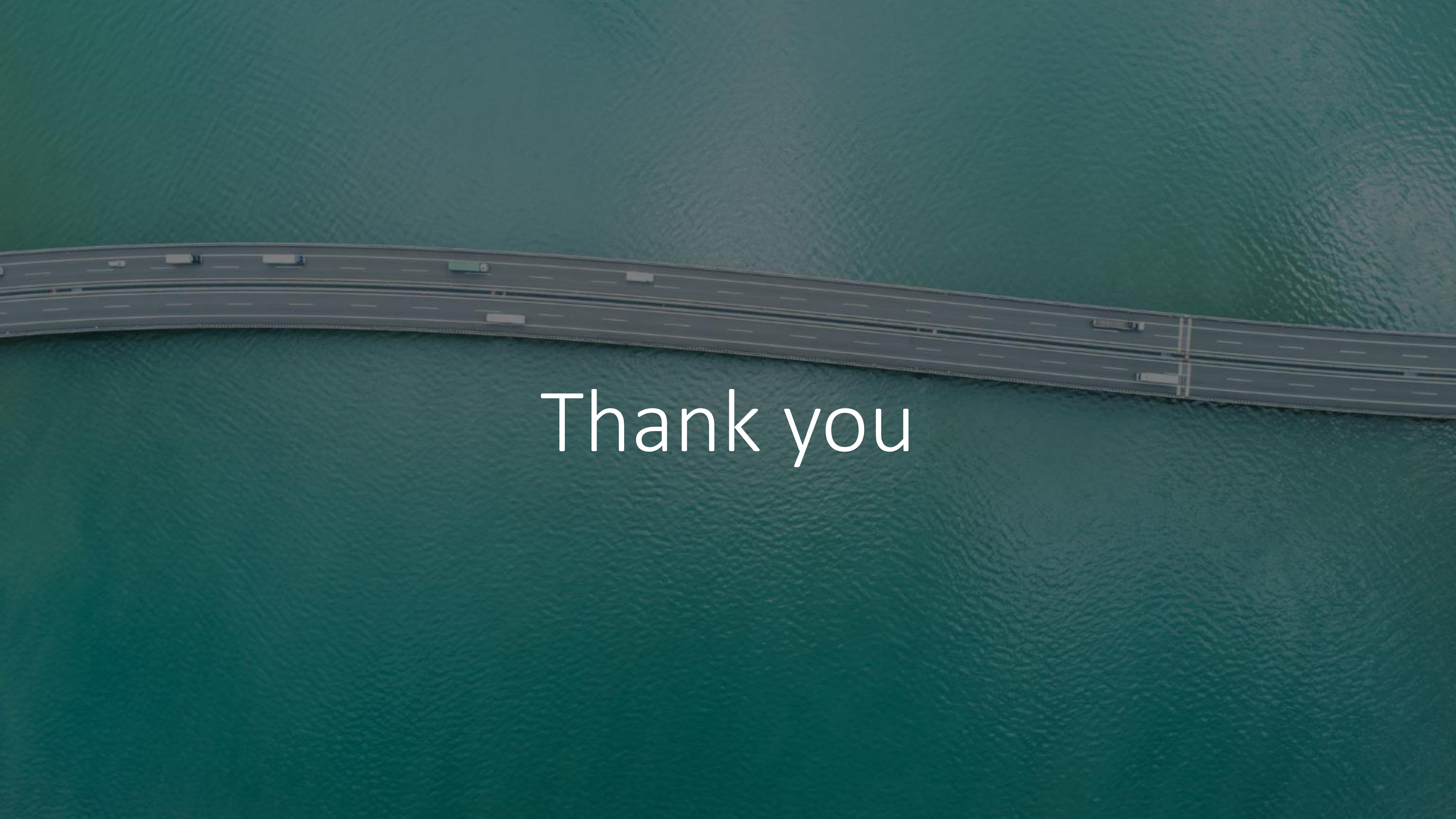
Income type: Less attention should be given to clients with income type Working.

Loan purposes: Clients with loan purposes related to repairs have a higher number of unsuccessful payments.

Housing type: Banks should target clients from housing type With parents, as they have the least number of unsuccessful payments.

Conclusion:

- Our analysis reveals important insights into the relationships between demographic factors and credit characteristics.
- Understanding the inverse relationships between age, family size, and credit amounts can help in assessing loan risks and determining appropriate lending strategies.
- Additionally, the findings suggest that densely populated areas exhibit distinct credit and income patterns, which can inform targeted marketing and risk assessment strategies.
- By leveraging these insights, loan providing companies can make more informed decisions, minimize risk, and enhance their lending processes.



Thank you