EDA Case Study

Credit Risk Analysis

Presented By
Vatsal Gohel

Introduction

1. Objective

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

2. Purpose

Credit Risk Analysis will allow a company to decide on whether to accept or reject a loan application based on the applicant's profile. As a result, the company is able to minimize business losses and avoid financial loss.

3. Data Set used

- current_data : ‘application_data.csv’
- previous_data : ‘previous_application.csv’

4. Steps Performed

- Loading Data into Jupyter Notebook
- Inspecting the Data
- Data Cleaning
  - Missing Value Check
  - Duplicate Value Check
  - Data Type Correction
  - Outliers Detection
  - Data Preparation
- Data Analysis
  - Univariate Analysis
  - Bivariate Analysis
  - Correlation
- Merging ‘previous_application’ to ‘application_data’
- Conclusion

EDA Analysis

- Load the Excel file into the Jupyter Notebook

  1. Load Dataset application_data.csv
  2. Load Dataset previous_application.csv

- Inspecting the Data set

  - Understanding the data

    - Checking the data set using function such as .info(), .shape(), .describe()

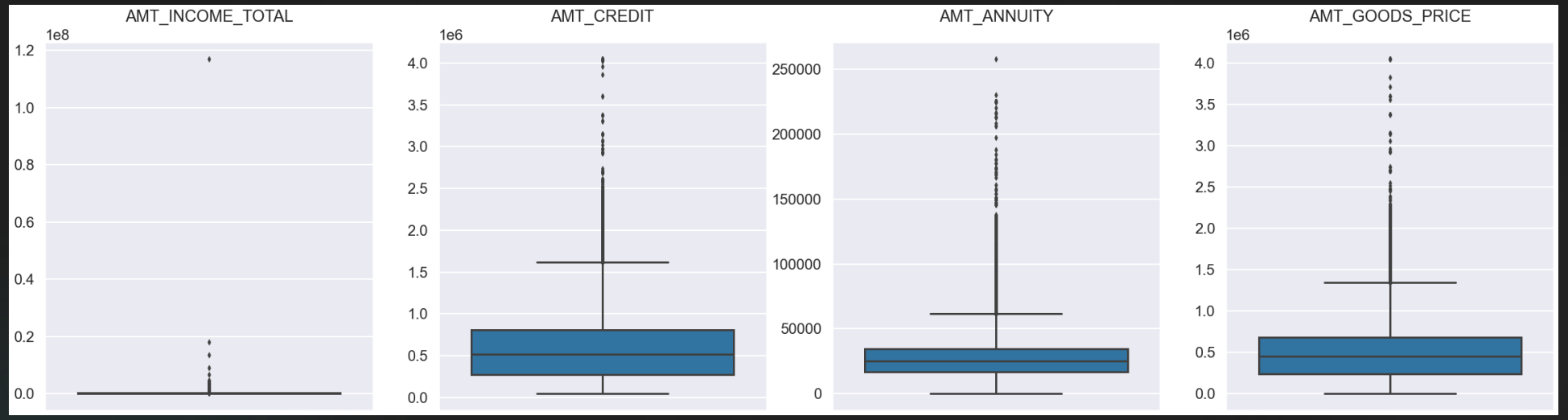
  - Data Cleaning

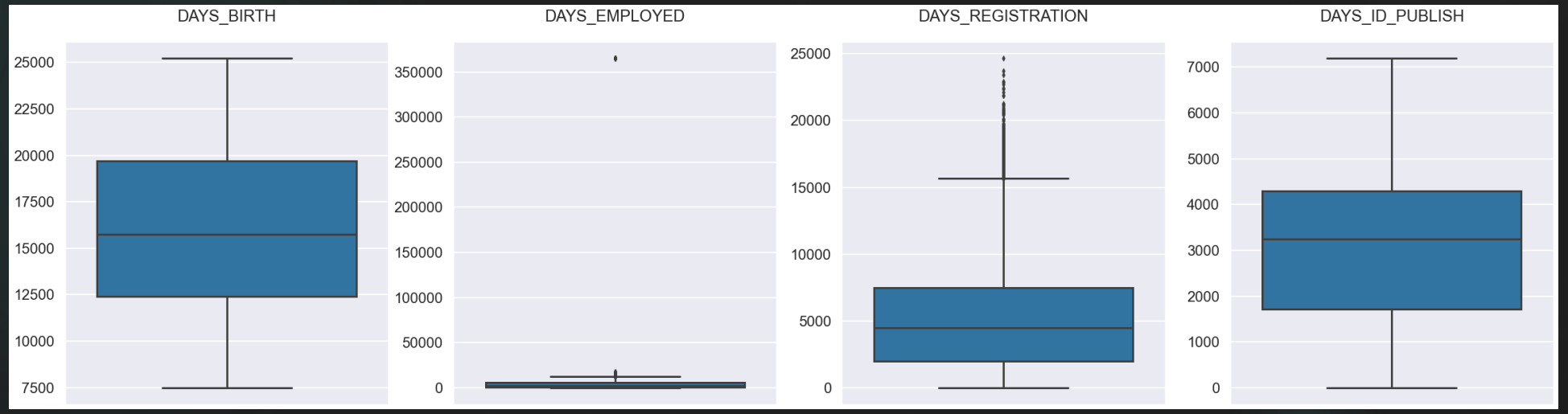
    - Checking the percentage of missing value in the data set.
    - Counting the number of null columns.
    - Dropped the column having null values greater than or equal to 40%.
    - Imputing the column with the mean, mode, and median for the numerical column, and for the categorical column, see what category you can use for the null values.

EDA Analysis

- We observe in the data set that the columns contain negative or mixed values, so we have imputed those values into absolute values for analysis.
- Looking for the outliers in various Numerical Columns.
- Grouping a number of Continuous Variables into smaller group for better analysis.
- For better analysis, there are some columns that have different data types, which are converted to their appropriate types.
- There are so many columns in the data set. We will remove the extra columns which we don’t need for further analysis.

Finding Outliers

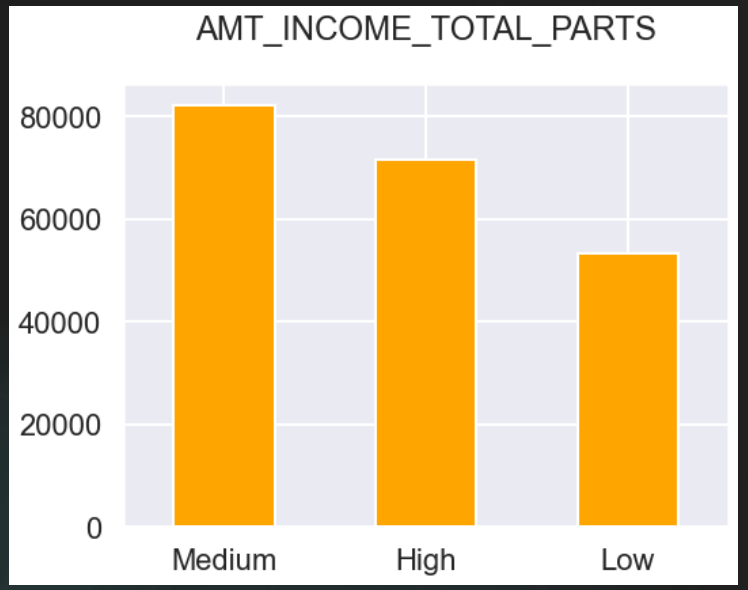


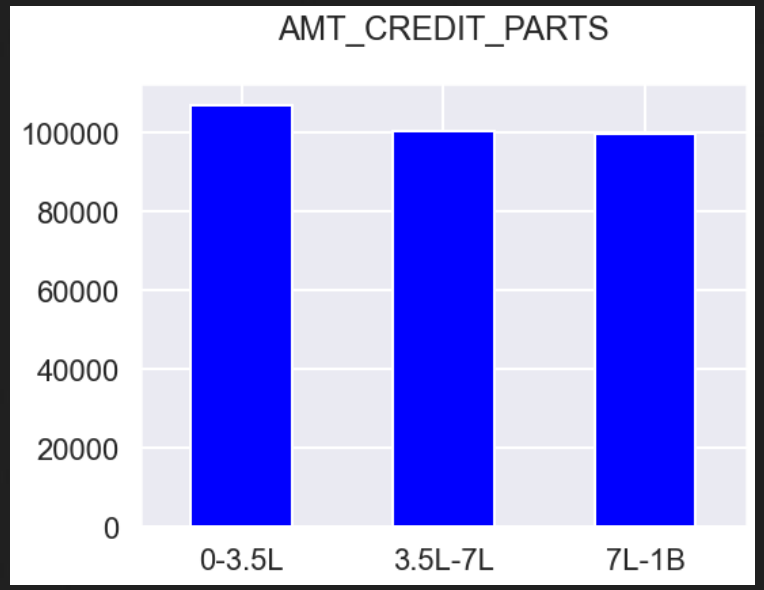


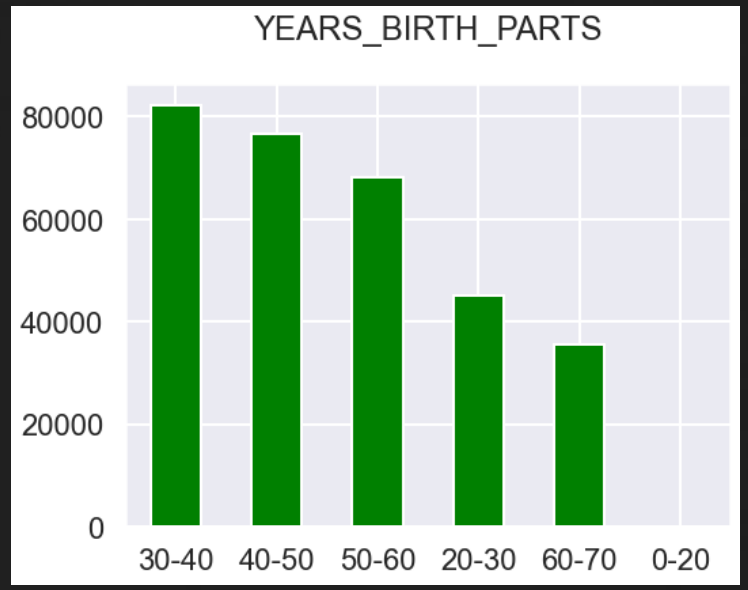
Finding Outliers

- Based on the Box Plot and Whisker Calculation, the lower whisker value for 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'AMT_INCOME_TOTAL', and 'AMT_CREDIT' is negative. As there can't be negative amounts, we look for outliers that are above the upper whisker value.
- From the graph above, it is clear that there are no outliers for 'DAYS_BIRTH' and 'DAYS_ID_PUBLISH'.
- When viewing the graph and data for 'DAYS_EMPLOYED', it is clear that the maximum value is 360000+ as it states an employee is working that many days which is not possible, so it is an outlier.
- From the graph and calculation of the 'DAYS_REGISTRATION' data, it can be seen that there are outliers in the data.

Grouping Continuous Numerical Variable





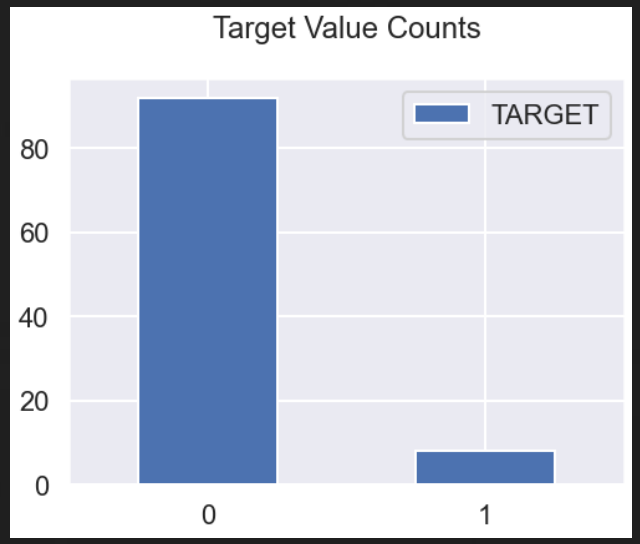


- **'AMT_INCOME_TOTAL' graph indicates that more clients are earning a medium-income than high and low income.**
- **'AMT_CREDIT_PARTS' graph indicates all bins have almost equal credit amounts.**
- **'YEARS_BIRTH_PARTS' graph indicates that the age range of 30-40 represents the majority of clients.**

Checking the Target Variable

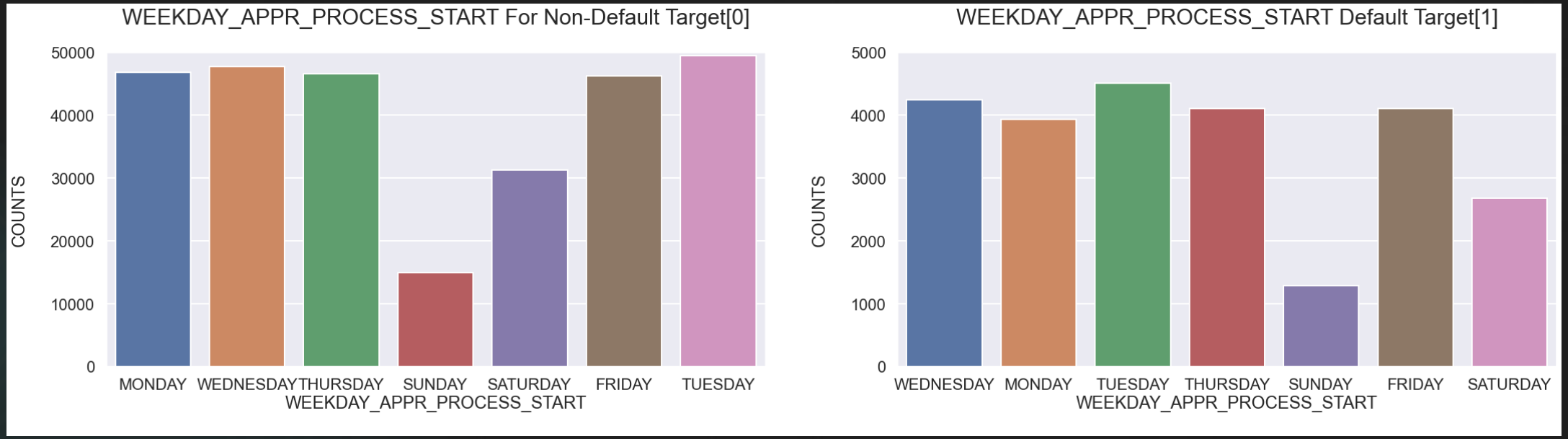
- We can see that there this an imbalance in the data set.
- The default rate is approximately 8.07% among customers who have difficulty making payments; on the other hand, 91.92% of customers do not face payment difficulties.
- Target imbalance Ratio: '11.39'.
- There is a huge imbalance between Target variables 0 and 1, so it makes more sense to divide the data into two sub-data sets and perform the analysis.

- The Data sets is divided as follows:
  1. Target 0 (current_data_0): Clients without Payment Difficulties, Non-Defaulters
  2. Target 1 (current_data_1): Clients with Payment Difficulties, Defaulters



Univariate Analysis on the basis of Target Variables

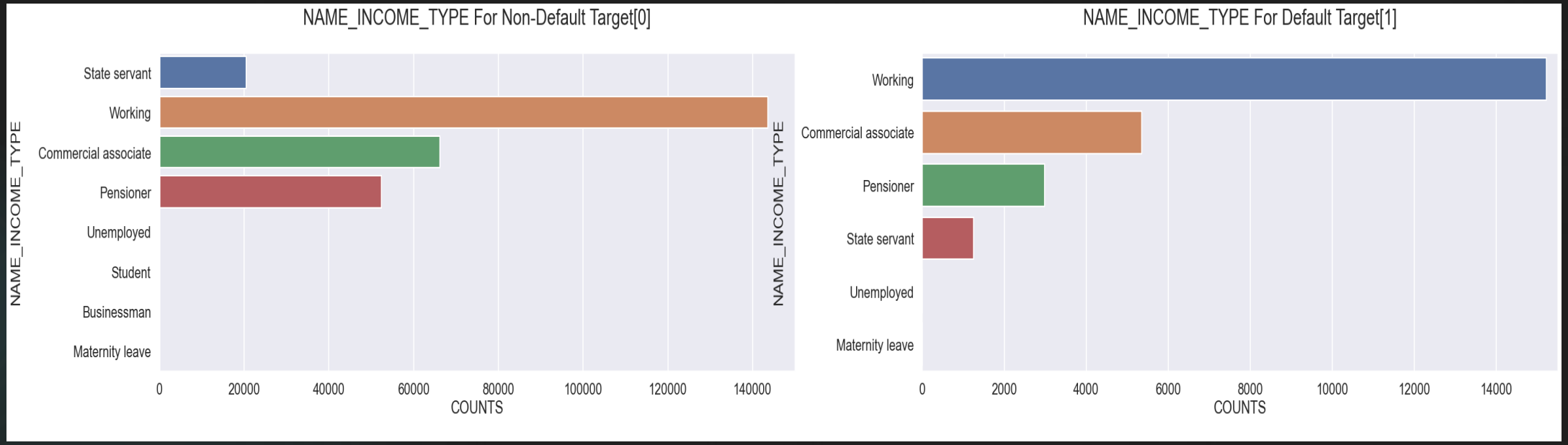
‘WEEKDAY_APPR_PROCESS_START’ Vs Target Variable



- **On the basis of the above graph, we can conclude that for defaulters and non-defaulters, application processes usually begin on Tuesdays and Wednesdays, and less often on weekends.**

Univariate Analysis on the basis of Target Variables

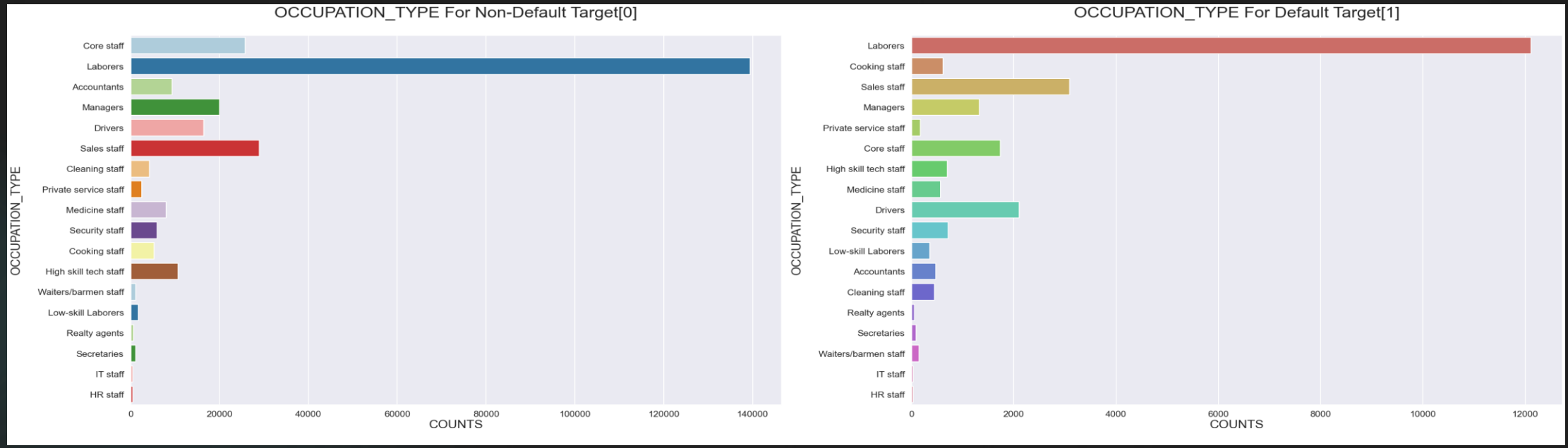
‘NAME_INCOME_TYPE’ Vs Target Variables:



- From the graph, we can conclude that the majority of defaulters are working, but working people are also bringing in a lot of money. The number of non-defaulters is also higher for State employees, pensioners, and commercial associates.
- Giving a loan should therefore be given to working clients first, then commercial associates, pensioners, and public servants.

Univariate Analysis on the basis of Target Variables

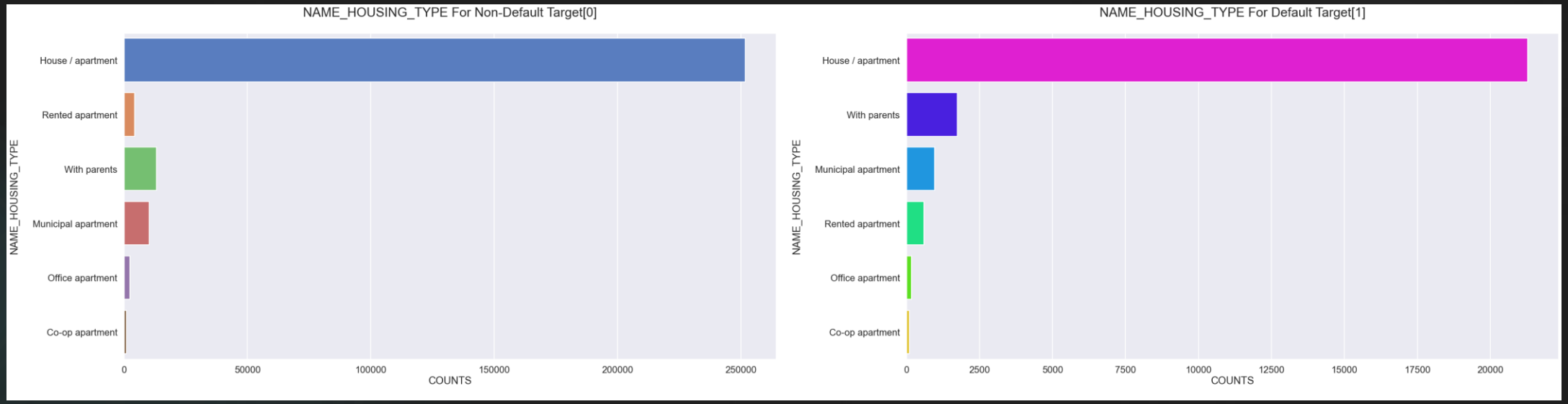
‘OCCUPATION_TYPE’ Vs Target Variable



- Observing the graph, we can conclude that non-defaulters with an occupation of Laborer make their payments on time, followed by those with an occupation of Sales staff, Core Staff, and Managers. Conversely, defaulters are more likely to hold the occupation of laborer followed by Sales staff, Drivers, and Core staff.
- Laborers, Sales staff, Core Staff, Managers, and Drivers should be provided loans according to this preference ranking. Loans beyond the limitations should not be provided as it would likely be difficult to recover.

Univariate Analysis on the basis of Target Variables

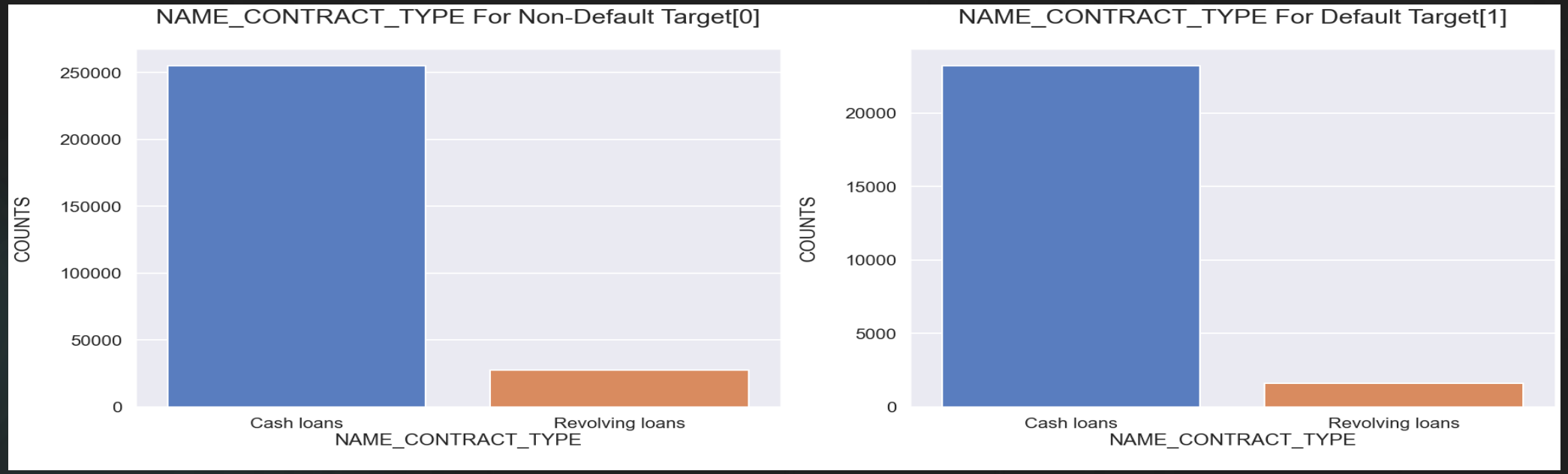
‘NAME_HOUSING_TYPE’ Vs Target Variable



- We can conclude from the graph that non-defaulters who own their house and live with their parents are most likely to make payments. On the other hand, defaulters also own a house, live with their parents, and may also live in a municipal apartment fails to make payment.
- Banks should provide loan to the Customer who owns their house and avoid giving loans to other.

Univariate Analysis on the basis of Target Variables

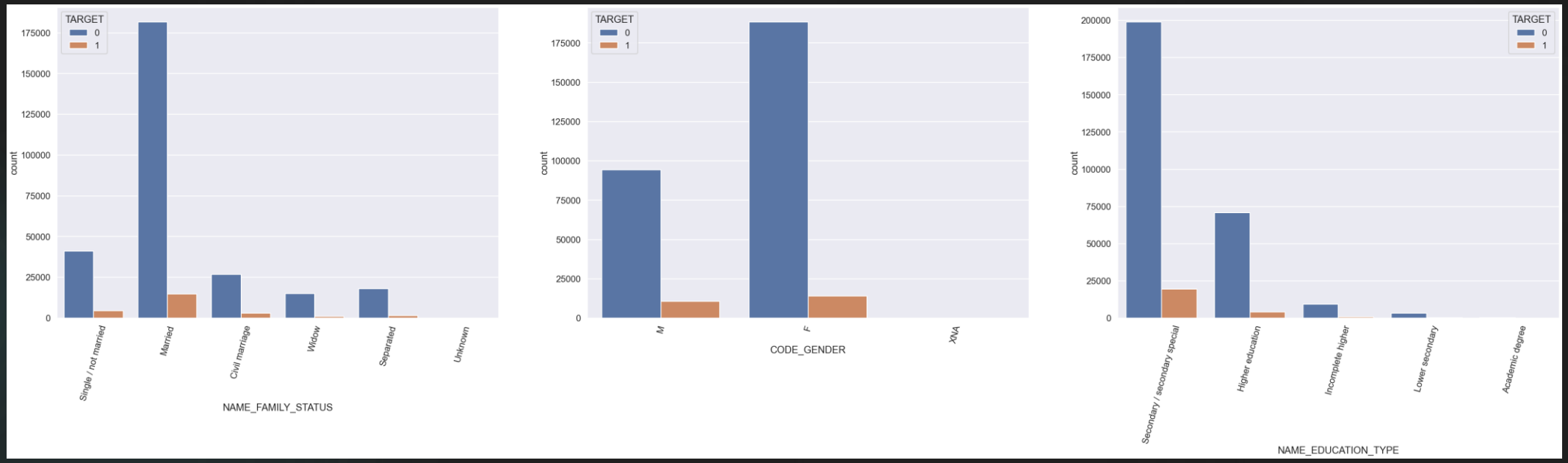
‘NAME_CONTRACT_TYPE’ Vs Target Variable



- Based on the graph, we can conclude that non-defaulters have issued more Cash-loans and also paid the installment on time. Defaulters also have more cash loans, but they often have difficulty paying installments.
- Banks should offer more Cash loans over Revolving loans.

Univariate Analysis on the basis of Target Variables

‘Multiple Categorical Columns’ Vs Target Variable



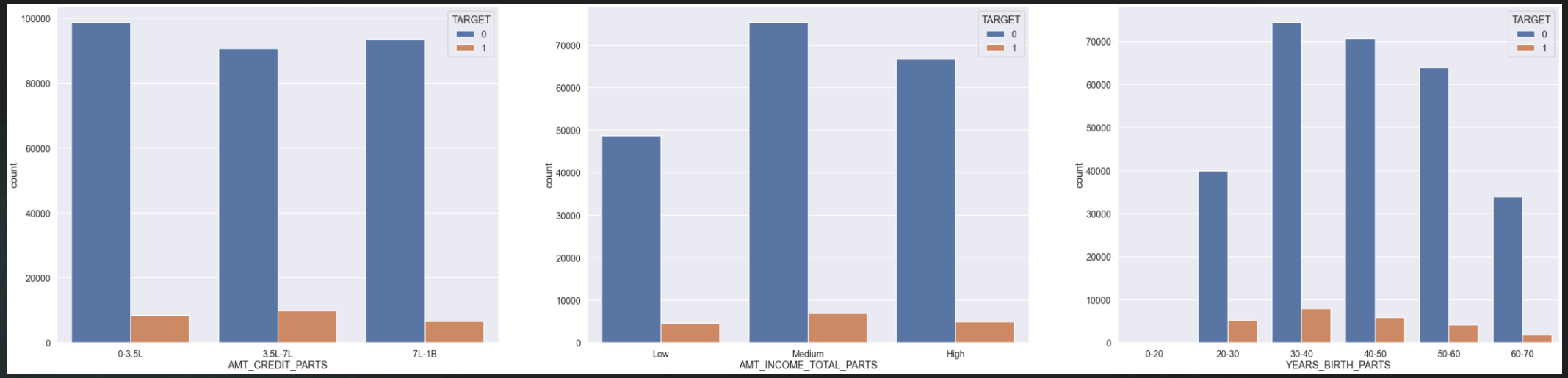
Univariate Analysis on the basis of Target Variables

‘Multiple Categorical Columns’ Vs Target Variable

- From the graph "NAME_FAMILY_STATUS", we can conclude that married people tend to take more Loan and are likely to make payment on time as compared to other categories.
- From the graph "CODE_GENDER", Females tend to take more loans compared to males. defaulters count is slightly higher for females compared to males.
- From the graph "NAME_EDUCATION_TYPE", we can conclude that Secondary/secondary special education and Higher Education people are applying for loans in higher numbers and are most likely to make payments on time.
- Women, married people, and those with a secondary or higher education should be given preference by the bank in terms of loans.

Continuous Univariate Variables Analysis

‘AMT_CREDIT’, ‘AMT_INCOME_TOTAL’, ‘DAYS_BIRTH’ Vs Target Variable



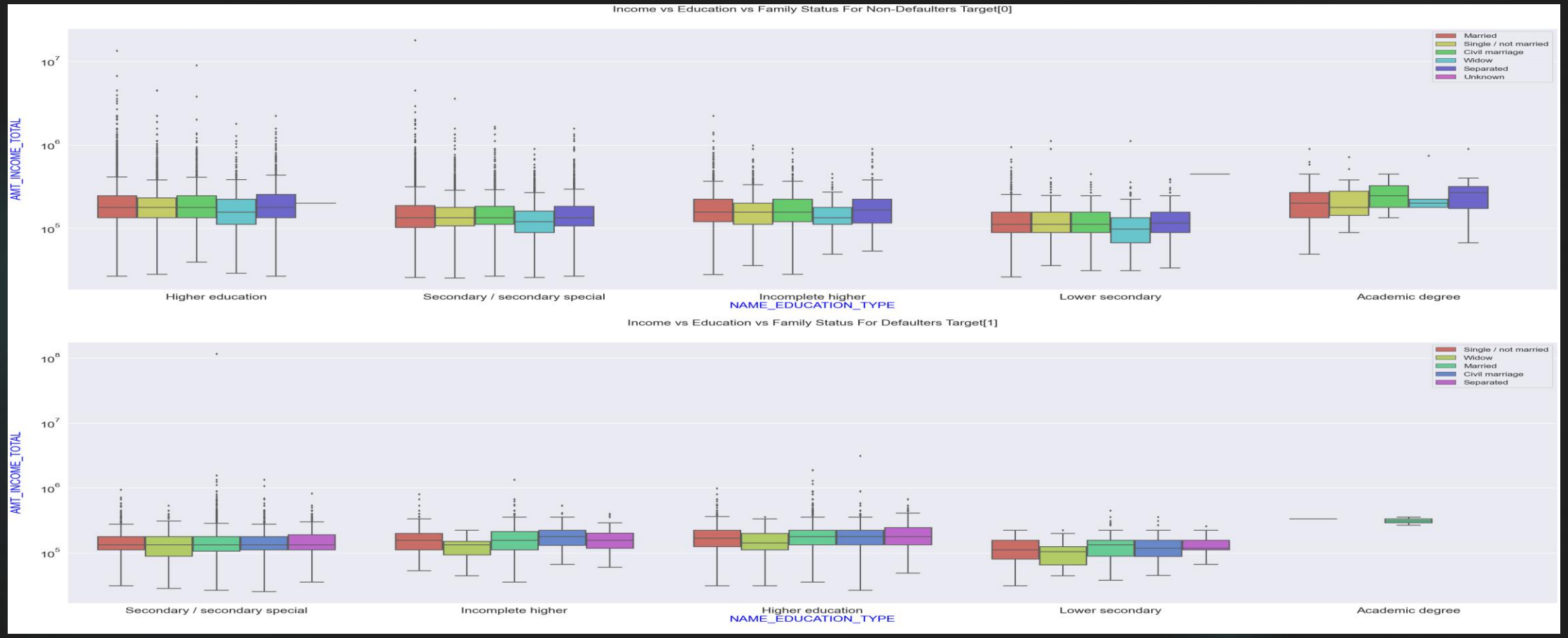
Continuous Univariate Variables Analysis

‘AMT_CREDIT’, ‘AMT_INCOME_TOTAL’, ‘DAYS_BIRTH’ Vs Target Variable

- From the graph "AMT_CREDIT_PARTS", we can conclude that clients are taking a loan between 0-3.5L and are more likely to make their payments on time. A client with a credit limit between 3.5-7L is most likely to default.
- From the graph "AMT_INCOME_TOTAL_PARTS", we can conclude that clients having a medium-income are more likely to default.
- From the graph "YEARS_BIRTH_PARTS", we can conclude that clients age range between 30-50 takes more loan and are more likely to make their payments on time.
- It would be beneficial for the bank to give preference to clients with medium incomes, a credit limit of 0-3.5L, and a range of 30-50 years of age.

Bivariate Analysis

‘AMT_INCOME_TOTAL’ Vs ‘NAME_EDUCATION_TYPE’ Vs ‘NAME_FAMILY_STATUS’ Among Target Variable



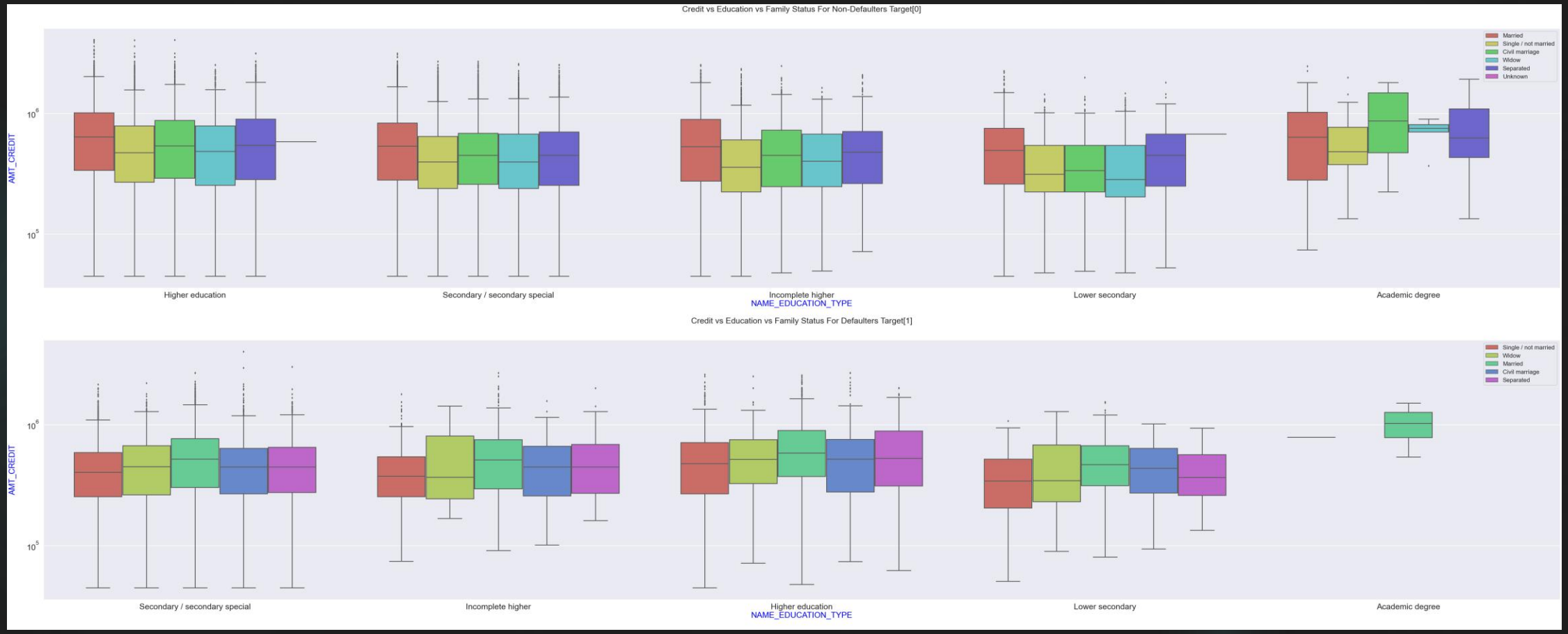
Bivariate Analysis

‘AMT_INCOME_TOTAL’ Vs ‘NAME_EDUCATION_TYPE’ Vs ‘NAME_FAMILY_STATUS’ Among Target Variable

- Non-Defaulters - Target [0]
  - Clients with all types of academic degrees have very few outliers, and of those widows, the client does not fall in the First and Third Quartile.
  - A higher number of outliers can be seen in Higher Education, Secondary Special, Incomplete Higher, and Lower Secondary.
  - Married clients with a High Education or Secondary Specialized tend to earn more money.
- Defaulters – Target [1]
  - The income of Married clients with an academic degree is much less than that of others.
  - Comparing Defaulters to Non-Defaulters, their incomes are relatively lower.

Bivariate Analysis

‘AMT_CREDIT’ Vs ‘NAME_EDUCATION_TYPE’ Vs ‘NAME_FAMILY_STATUS’ Among Target Variable



Bivariate Analysis

‘AMT_INCOME_TOTAL’ Vs ‘NAME_EDUCATION_TYPE’ Vs ‘NAME_FAMILY_STATUS’ Among Target Variable

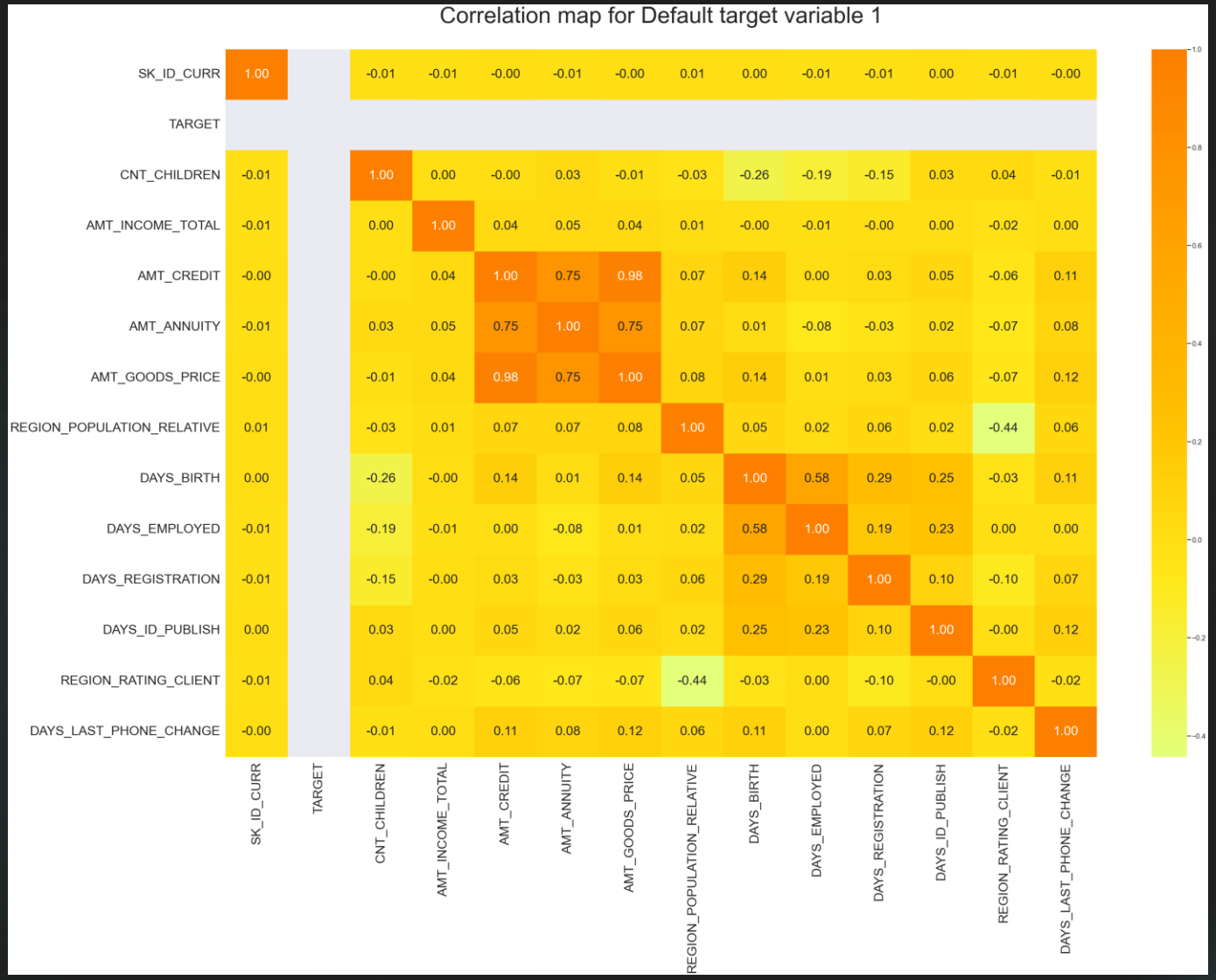
- Non-Defaulters - Target [0]
  - Outliers are more common among clients with any type of education except academic degrees.
  - Clients who are married and with different educational backgrounds tend to have a wide variety of credit loan options available to them.
- Defaulters – Target [1]
  - An academically qualified married client applied for a higher credit loan and had no outliers.
  - Clients with Higher Education, Incomplete Higher Education, Lower Secondary Education, and Secondary/Secondary Special Education might take out a high amount of debts.

Correlations for the Target variables

Target 0

Correlations for the Target variables

Target 1



Correlations for the Target variables

Target 0

	VARIABLE1	VARIABLE2	Correlation	Corr_abs
88	AMT_GOODS_PRICE	AMT_CREDIT	0.986966	0.986966
89	AMT_GOODS_PRICE	AMT_ANNUITY	0.776624	0.776624
74	AMT_ANNUITY	AMT_CREDIT	0.771248	0.771248
134	DAYS_EMPLOYED	DAYS_BIRTH	0.626114	0.626114
175	REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	-0.539005	0.539005
73	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418906	0.418906
87	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349473	0.349473
59	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799	0.342799
114	DAYS_BIRTH	CNT_CHILDREN	-0.336966	0.336966
148	DAYS_REGISTRATION	DAYS_BIRTH	0.333151	0.333151

Vs

Target 1

	VARIABLE1	VARIABLE2	Correlation	Corr_abs
88	AMT_GOODS_PRICE	AMT_CREDIT	0.982854	0.982854
89	AMT_GOODS_PRICE	AMT_ANNUITY	0.752891	0.752891
74	AMT_ANNUITY	AMT_CREDIT	0.752195	0.752195
134	DAYS_EMPLOYED	DAYS_BIRTH	0.582185	0.582185
175	REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	-0.443236	0.443236
148	DAYS_REGISTRATION	DAYS_BIRTH	0.289114	0.289114
114	DAYS_BIRTH	CNT_CHILDREN	-0.259109	0.259109
162	DAYS_ID_PUBLISH	DAYS_BIRTH	0.252863	0.252863
163	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.229090	0.229090
128	DAYS_EMPLOYED	CNT_CHILDREN	-0.192864	0.192864

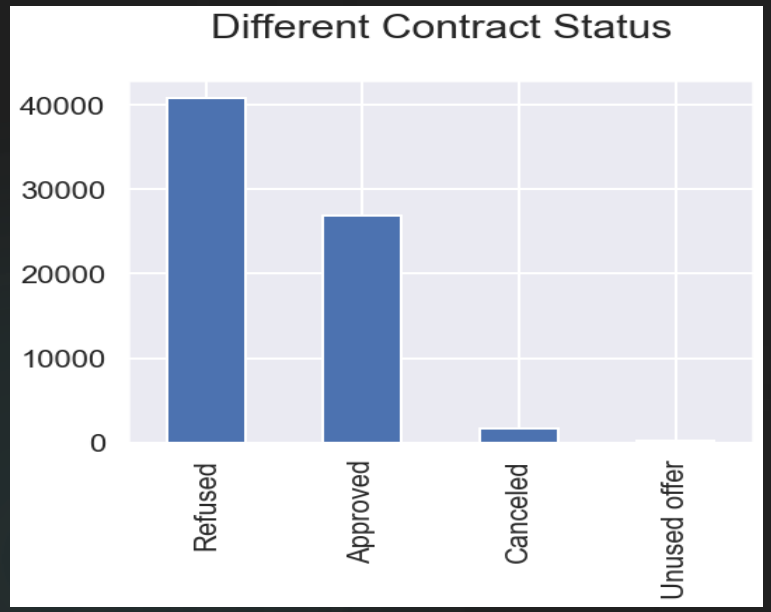
Correlations for the Target variables

- For Non-Default Target [0] Variables
  - From the heat map and table we get our Top 10 Correlations.
  - AMT_GOODS_PRICE is highly correlated with AMT_CREDIT, which has a value of 0.98, and with AMT_ANNUITY, which has a value of 0.77.
- For Default Target [1] Variables
  - From the heat map and table we get our Top 10 Correlations.
  - AMT_GOODS_PRICE is highly correlated with AMT_CREDIT, which has a value of 0.98, and with AMT_ANNUITY, which has a value of 0.75.

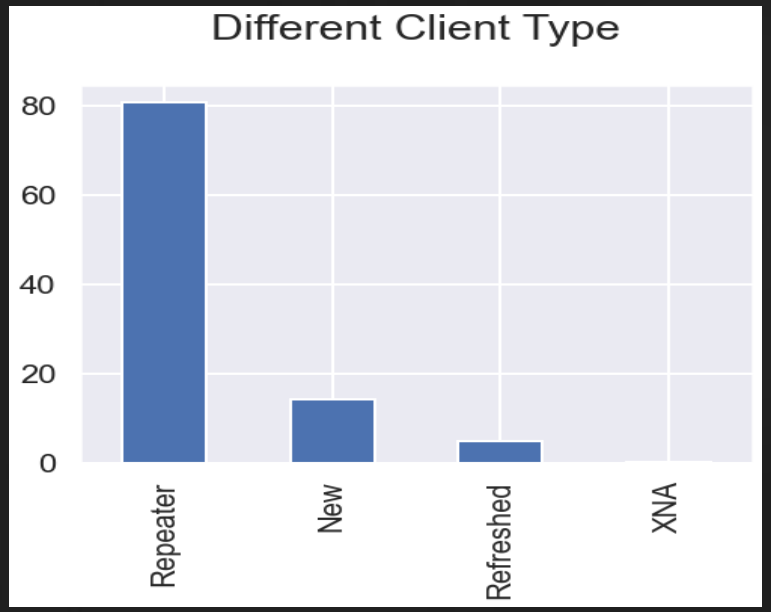
➢ The Credit amount is the most important factor to Non-Default and Default customers, as the clients get the credit amount they request based on the Goods price they are putting up as security. Another factor is the punctuality of the client to pay the debt on time.

Analysis From Previous application Data Set

NAME_CONTRACT_STATUS



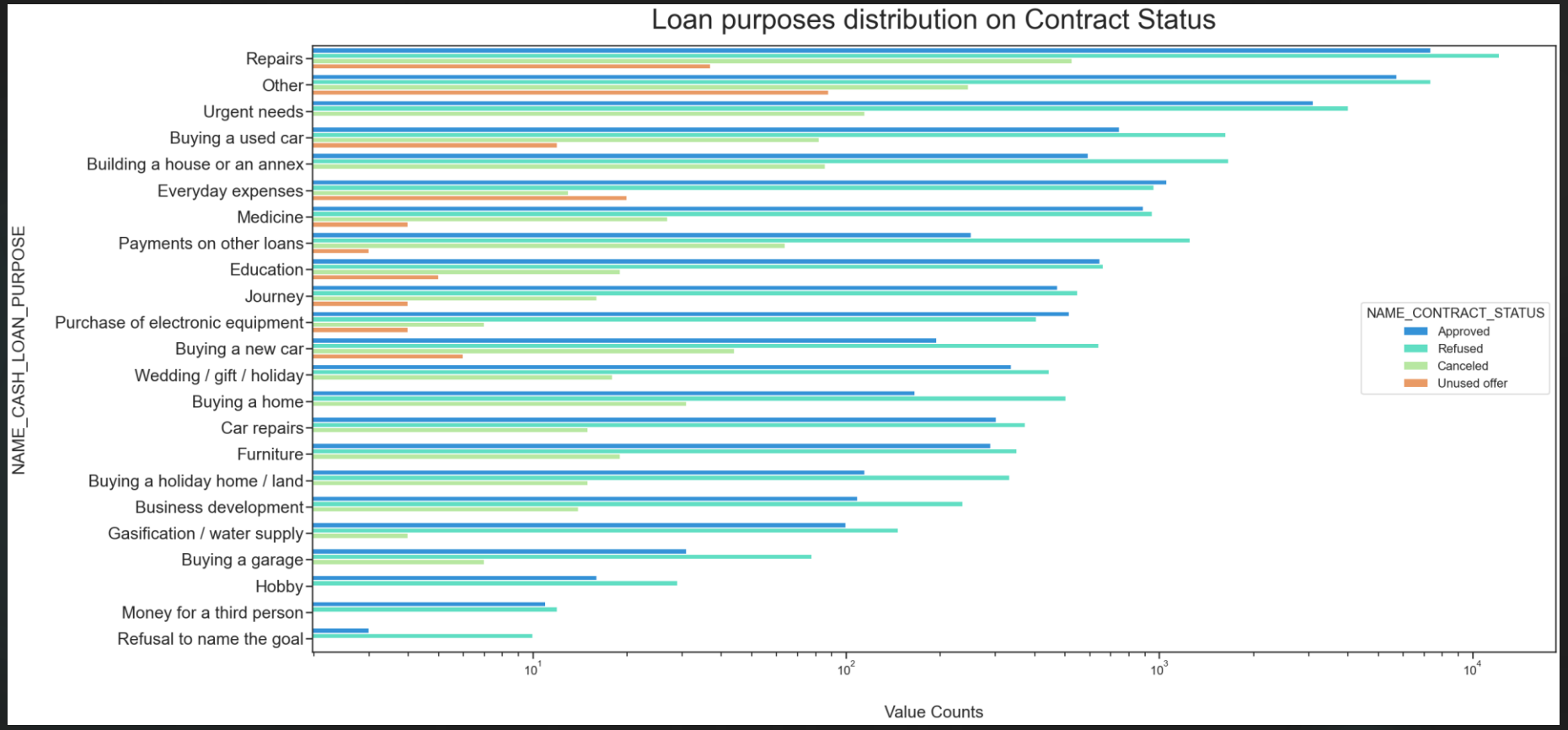
NAME_CLIENT_TYPE



- From ‘NAME_CONTRACT_STATUS’, we found that the number of denied applications is around 40000 and the number of approved applications is about 27000.
- From ‘NAME_CLIENT_TYPE’, we found that 80.7% of the clients were repeated clients who applied for loans and 14.3% of the clients were new to the process.

Loan purposes distribution on Contract Status

‘NAME_CONTRACT_STATUS’ Vs ‘NAME_CASH_LOAN_PURPOSE’

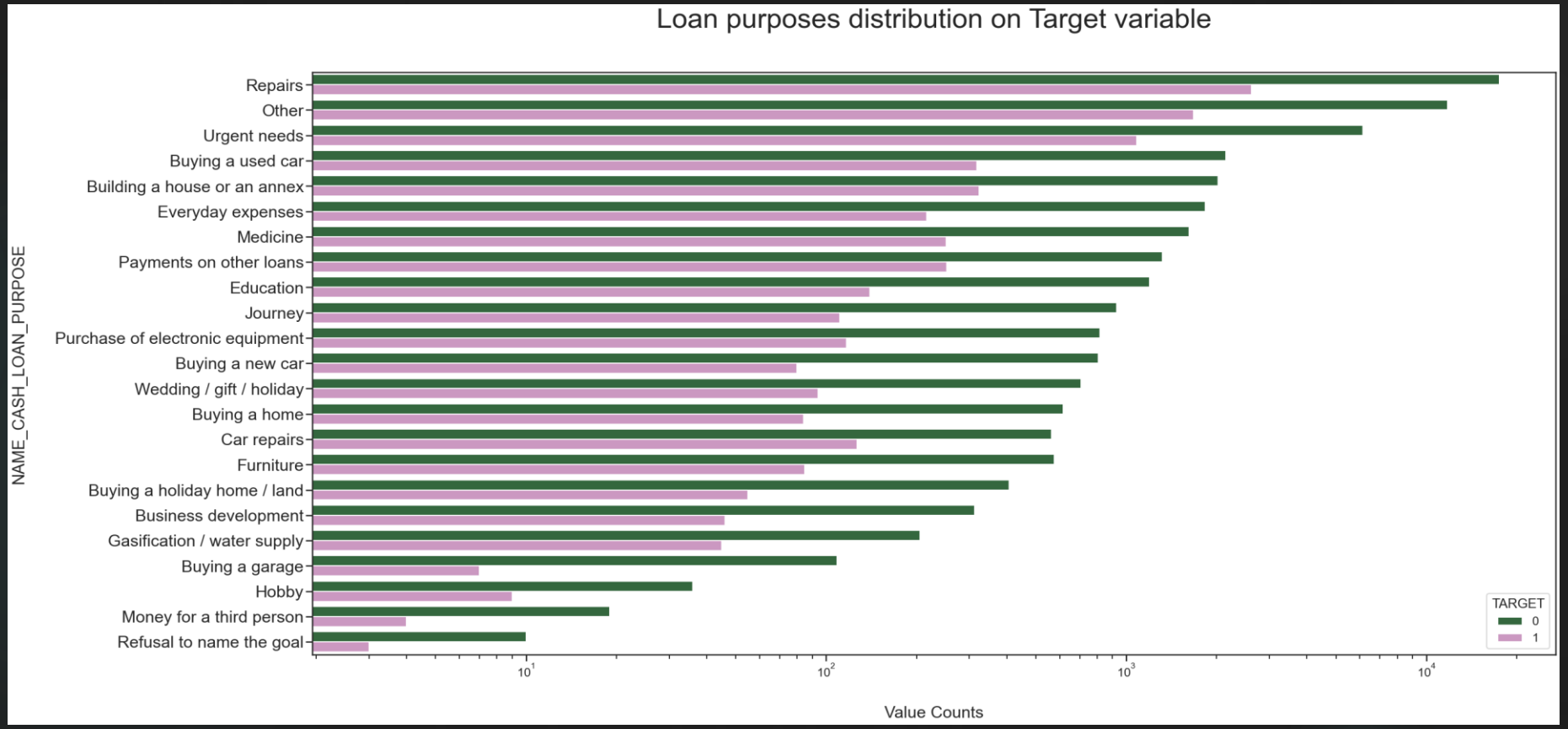


Loan purposes distribution on Contract Status

- On the Basis of Contract Status.
  - Despite the fact that there are a greater number of rejections for loans given for repair purposes, the number of approvals is higher as well.
  - The rejection rate for repayments on other loans, buying a car, and buying a home or vacation home is higher than approval rates.
  - The number of approvals and rejections for loans taken for educational and medicinal purposes is equal.
  - We can clearly see that there are a significant number of unused loan offers in the other loan purpose, since the clients might not have a reason to take a loan.

Loan purposes distribution on Target variable

‘TARGET’ Vs ‘NAME_CASH_LOAN_PURPOSE’

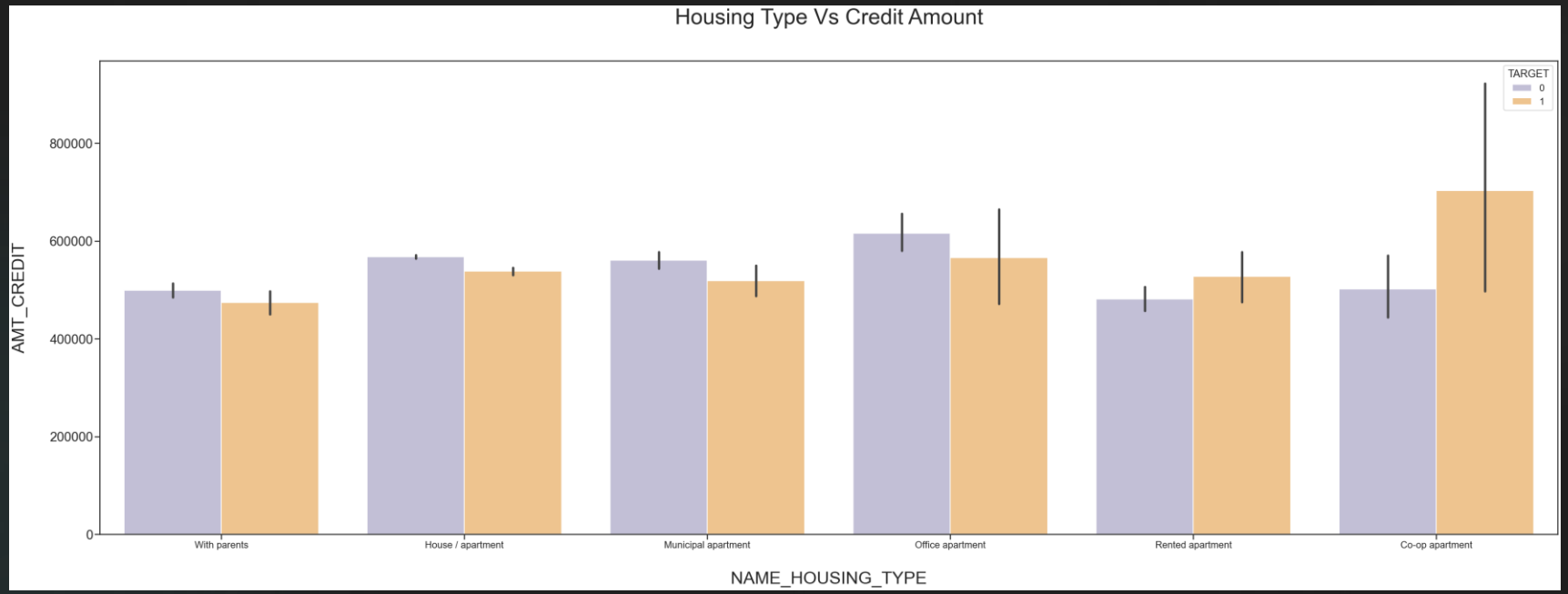


Loan purposes distribution on Target variable

- On the Basis of Target Variable.
  - Although the loan given for the purposes of repair is more likely to be paid on time or have no payment difficulties, the count of defaulters is also higher for the same.
  - Those who took loans to buy a new car and to buy a home have the same problems paying their installments.
  - Comparing Education & Medicine loan purpose, there is a high number of clients experiencing difficulties paying off their loans when compared to "Education“.
  - Most of the clients who take out a loan on an urgent need or for some other purpose make the payment on time.

Housing Type Vs Credit Amount

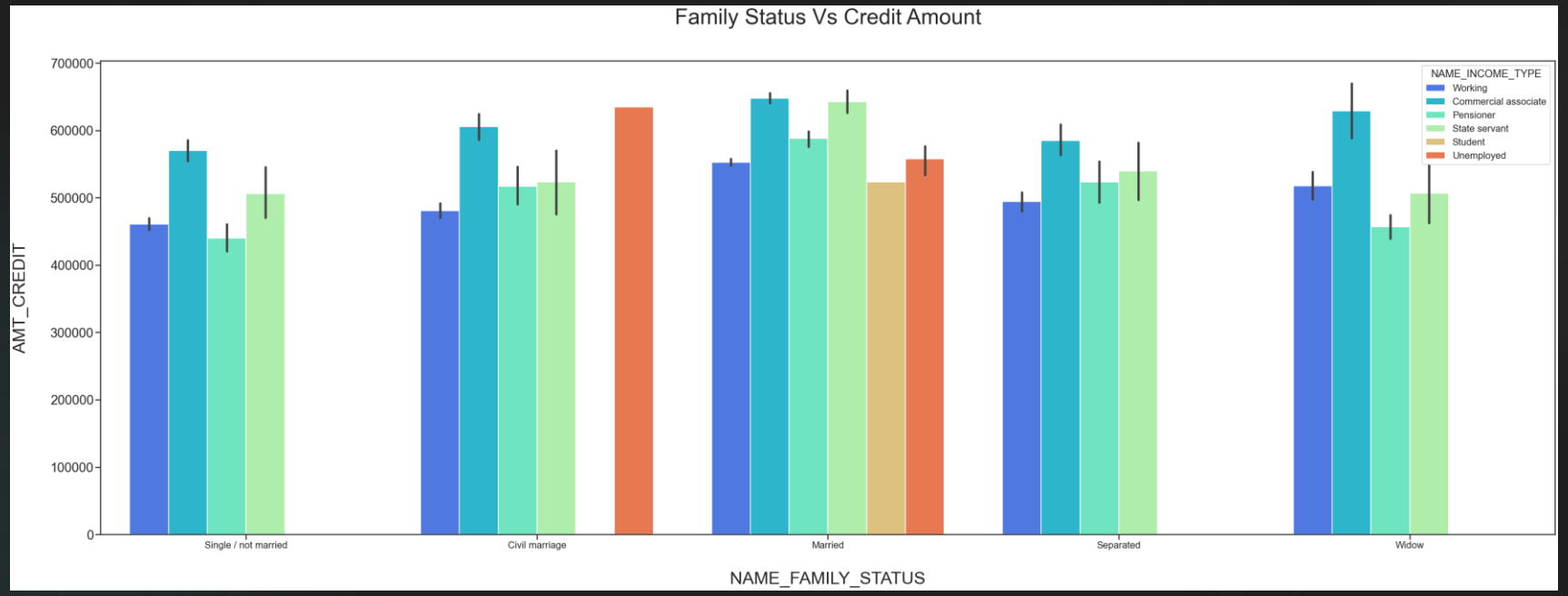
‘NAME_HOUSING_TYPE’ Vs ‘AMT_CREDIT’



- The above graph shows that the Client with a co-op apartment has a higher credit amount for Target 1 whereas the Client with an office apartment has a higher credit amount for Target 0.
- Banks should avoid giving loans to housing types like Co-op apartments since they experience payment difficulties, whereas they should focus on housing types like office apartments, municipal apartments, or houses.

Family Status Vs Credit Amount

‘NAME_FAMILY_STATUS’ Vs ‘AMT_CREDIT’



- The above graph shows the Client with a family status of married and an income type of Commercial Associate and State Servant has a higher credit limit.
- We can see that there are clients who are married and have student income and have a good credit limit.
- Clients who have done civil marriages and are unemployed tend to have higher credit limits.

Final Conclusion

- To minimize risk, loan applications should be reviewed for the following factors.

1. CODE_GENDER
2. AMT_INCOME_TOTAL
3. AMT_CREDIT
4. AMT_ANNUITY
5. NAME_FAMILY_STATUS
6. NAME_HOUSING_TYPE
7. NAME_EDUCATION_TYPE
8. NAME_CASH_LOAN_PURPOSE

Final Conclusion

- Recommended groups to be consider

1. An applicant whose application has been approved in a previous application.
2. Married applicant compared to other family status.
3. females are more favorable than males.
4. Other than Co-op apartment, housing type can be consider.
5. Client with a well-recognized job and are highly educated.
6. Client who are more intrested in Cash loans over Revolving loans.

- Recommended groups not to be consider

1. Unemployed clients.
2. Low income people groups.
3. Customers who previously refused, cancelled or did not use an offer.
4. client with younger age and having a lower or secondary education.

Final Conclusion

- To ensure successful payments, banks should pay more attention to contracts type for students, pensioners, and businessmen with housing types other than co-op apartments or office apartments.
- There is an increase in payment difficulties among those who work and a decrease in payment difficulties among those who are pensioners while comparing both default payment and non-default payment.
- Female clients and married clients are more likely to pay their outstanding debts on time.
- Clients with higher education face fewer payment difficulties.
- Those with low income should be less targeted because the majority of the time they have trouble paying their debt on time.
- Obtain as many clients as possible from housing types with parents, as they experience the smallest number of unsuccessful payments.

THANK YOU!!!