Case Study on Loan Applications

By Abhijeet Srivastava and Gilla Saiteja

Objectives

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.
- 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.
- The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:
  - The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, (Target 1)
  - All other cases: All other cases when the payment is paid on time. (Target 0)
- When a client applies for a loan, there are four types of decisions that could be taken by the client/company):
  - Approved: The Company has approved loan Application
  - Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
  - Refused: The company had rejected the loan (because the client does not meet their requirements etc.).
  - Unused offer: Loan has been cancelled by the client but on different stages of the process.
- In this case study, we will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Approach used

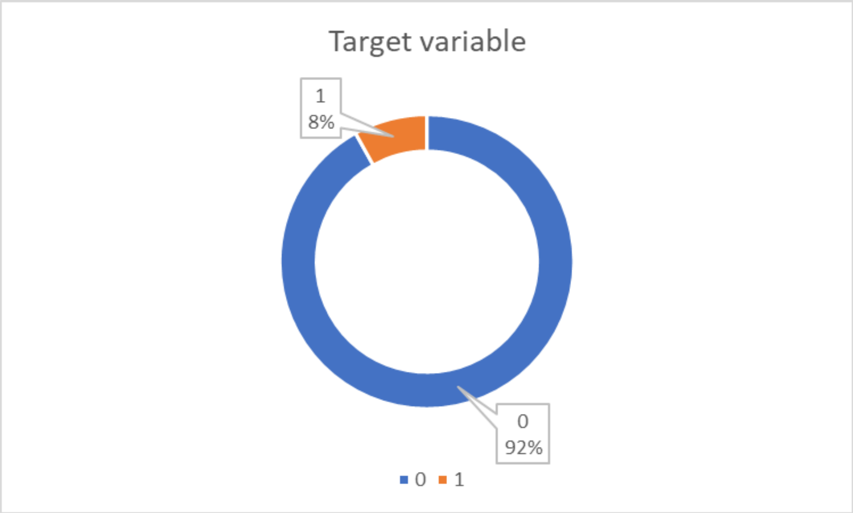
- This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.
- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.
- First, we explored and cleaned the data in both files application_data and previous_application
- We decided to drop all columns with more than 40% data in both the files as replacing the missing values in those columns would have resulted in skewing the data
- For columns with less than 40% missing values the following approach was used:
  - Categorical columns: adding a new category 'Unknown'
  - Numerical columns: based on the statistics, either median/mode
  - For numerical columns with discrete values mode was used

Approach used continued

- After cleaning the data, on analysing the Target column, we noticed a data imbalance with 8% in defaulters and 92% in repayers in the 3.1 lakh entries
- We split the application_data data frame by target and ran univariate and bivariate analyses to get some insight into the overall patterns and trends (more details in the following slides)
- We then used a correlation matrix to find the top 10 correlations for both Target 0 and 1
- After this we merged both the data frames using SK_CURR_ID to find patterns in accepting/rejecting loans based on
    - Gender, Age, Marital status, education level, employment status, family status, credit score (add graph in ipynb)
    - Owning a house/car
    - Income of client vs credit amt of loan vs loan annuity vs price of goods for which loan is give
    - Status of application based on reason for loan
    - Days taken to process application

Univariate Analysis on application_data

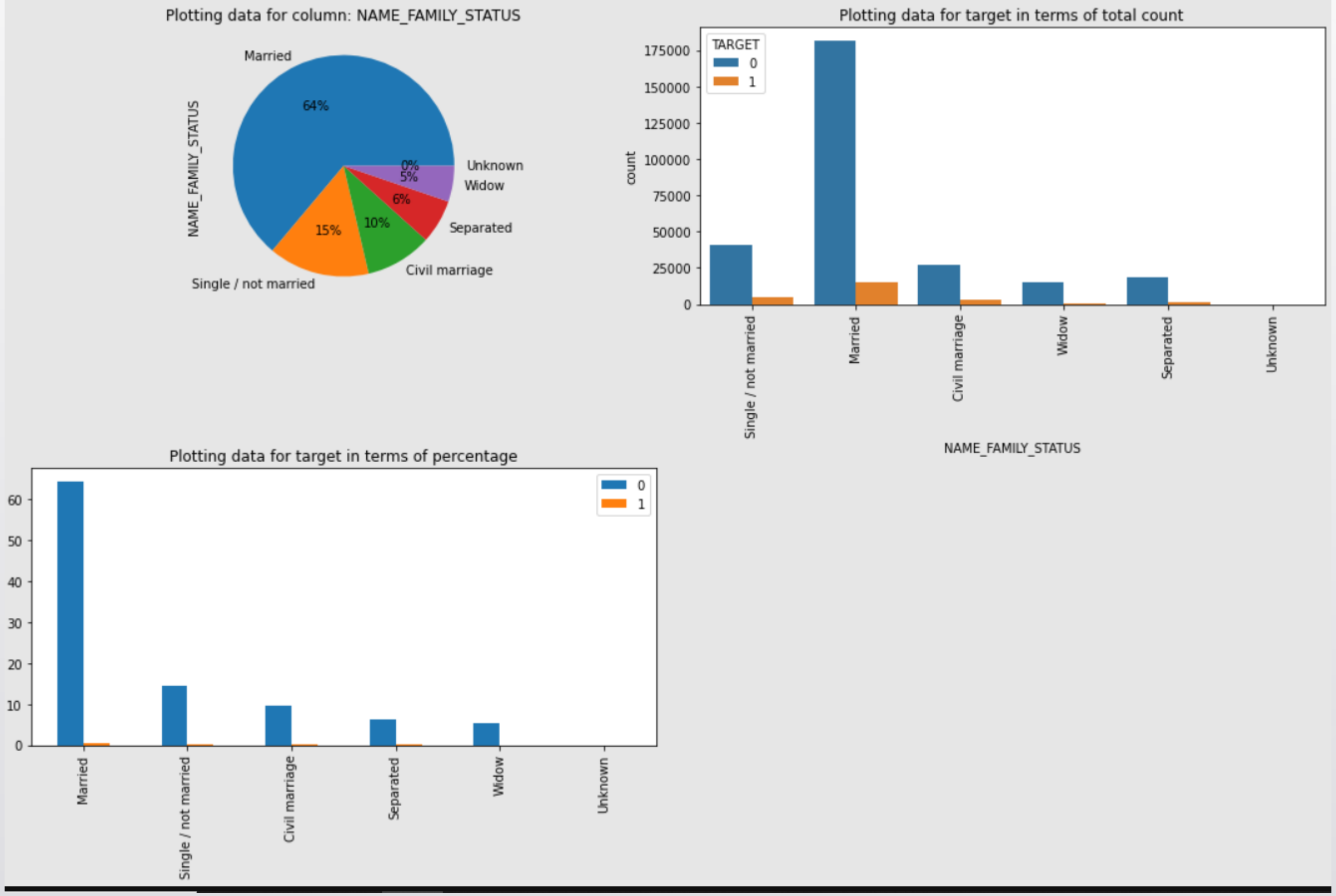
Target	Count	Percentage
0	282686	92%
1	24825	8%
Total	307511	



Insights:

- Class imbalance
- DF was split up into train_0 and train_1 to get insights for correlation

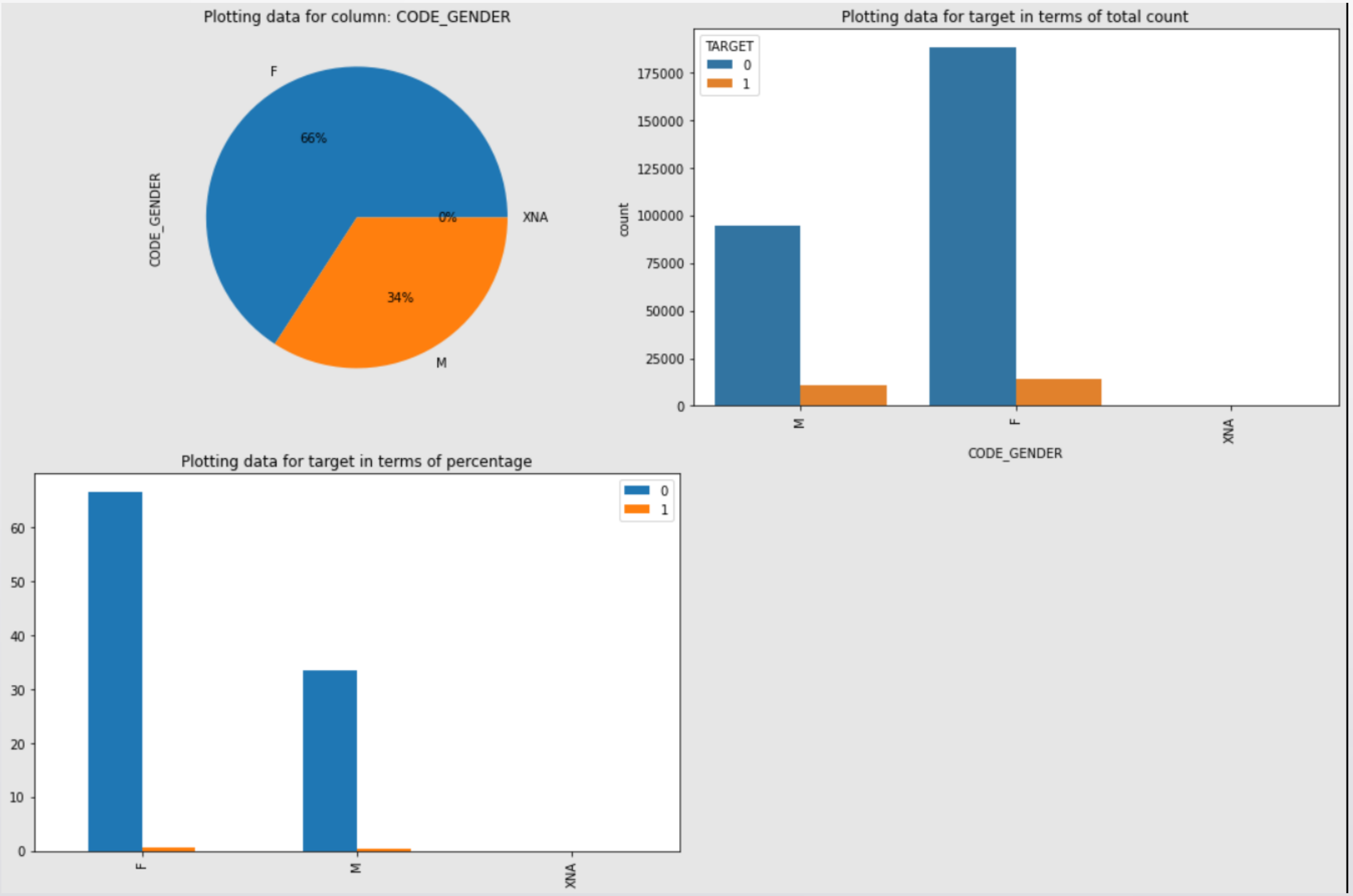
Univariate Analysis for marital status



Insight:

- Married people apply for more loans (64%) and have very few defaulters

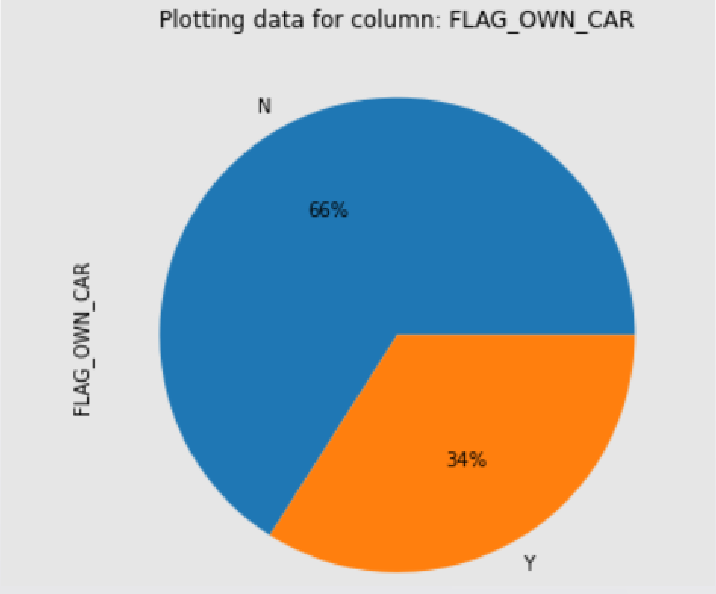
Univariate Analysis for Gender

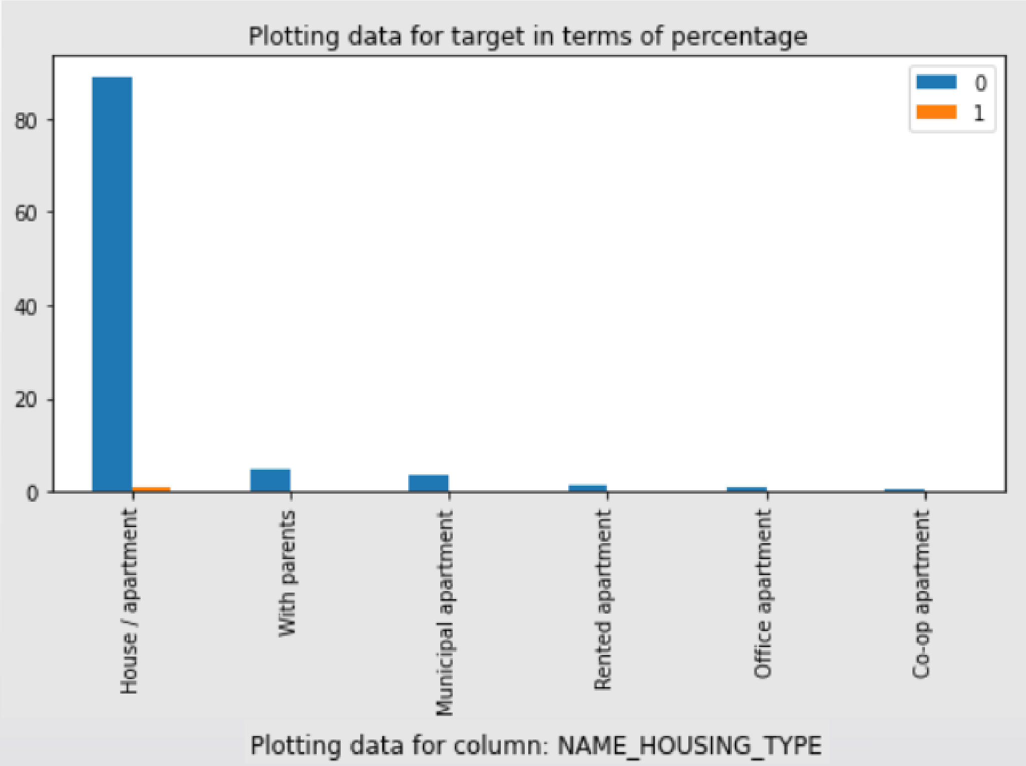


Insight:

- Females apply for more loans than males and default less than males

Univariate Analysis for owning car and housing

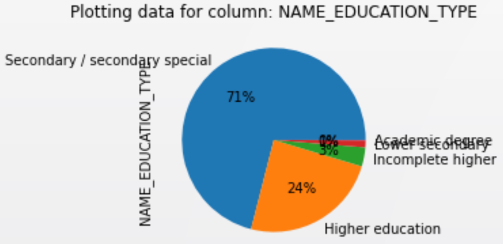


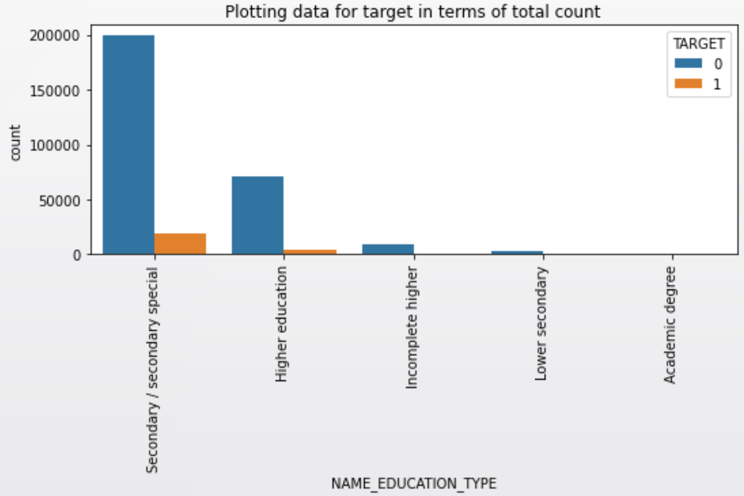


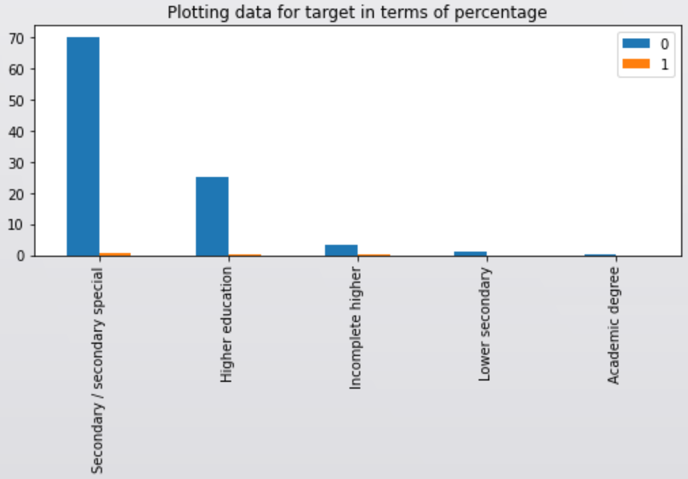
Insight:

- 34% defaulters own a car, 66% don't. Maybe the 66% apply for loans for cars
- 89% of repayers live in a house or apartment, implying that house owners are not risky in terms of repaying loans

Univariate Analysis for education level



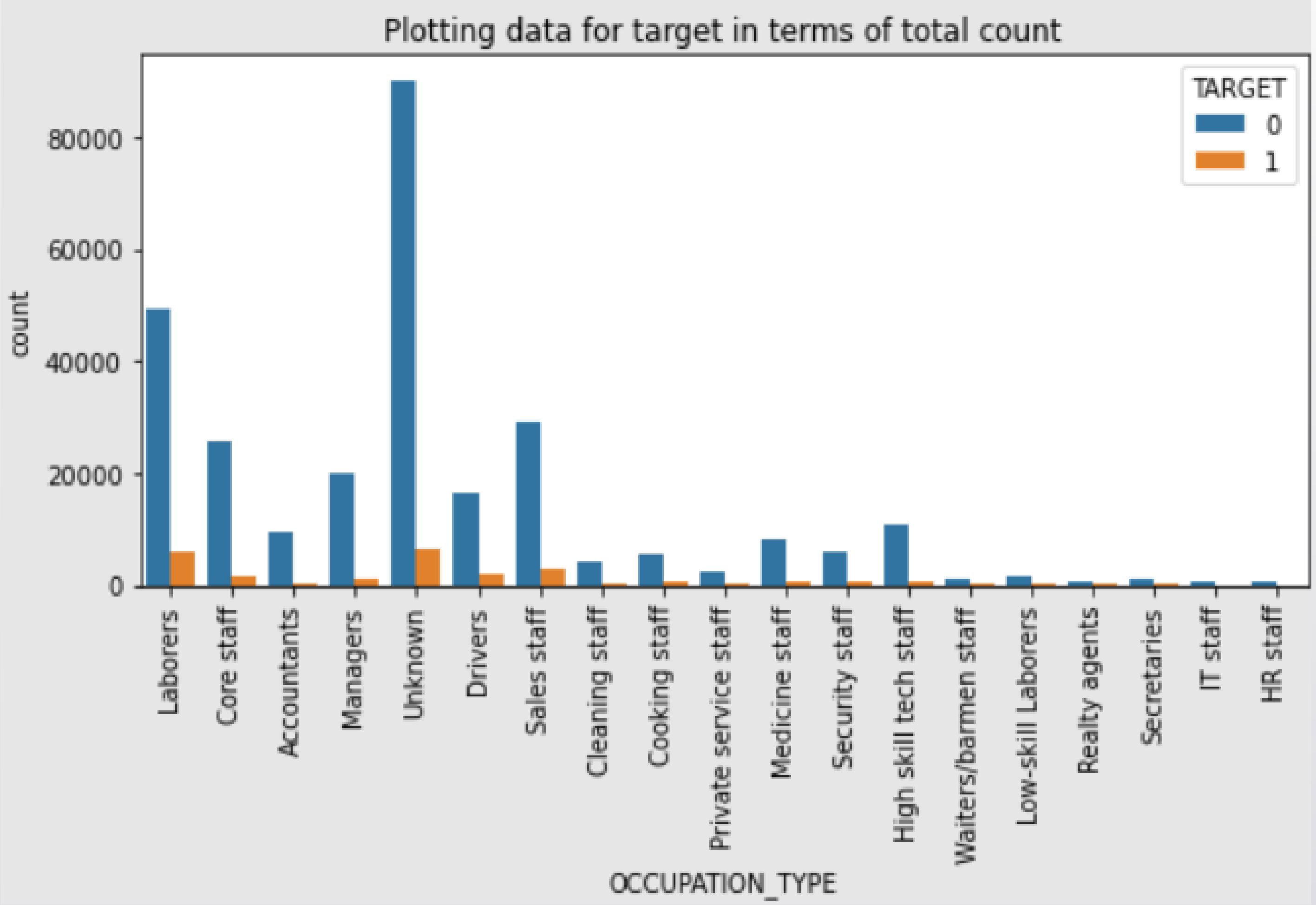




Insight:

- People who's highest education level is Secondary education have applied more loans and have relatively lower defaulters.
- Graphs also show that higher the education level, lower the defaulting rate

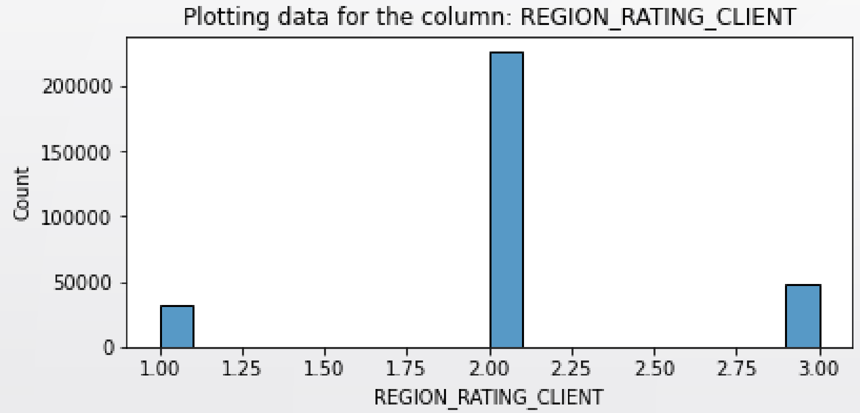
Univariate Analysis for Occupation Type

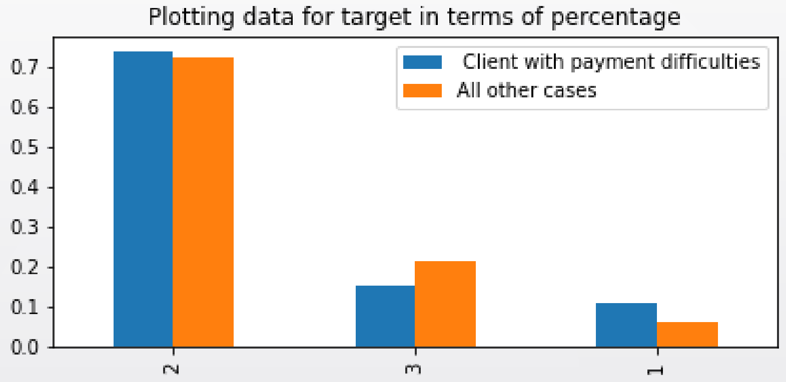


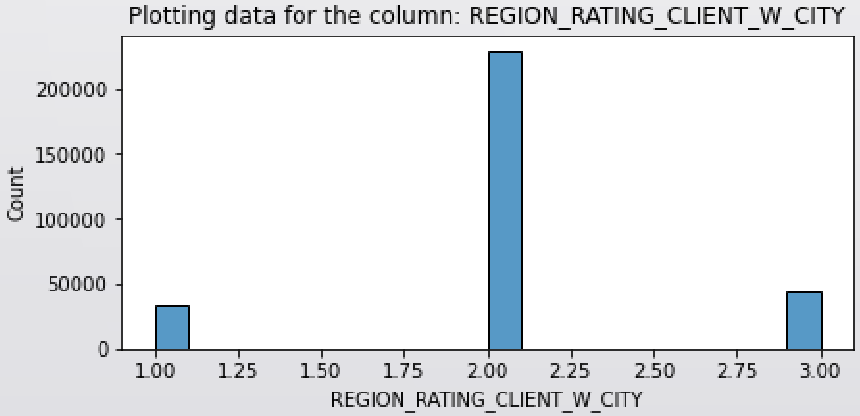
Insight:

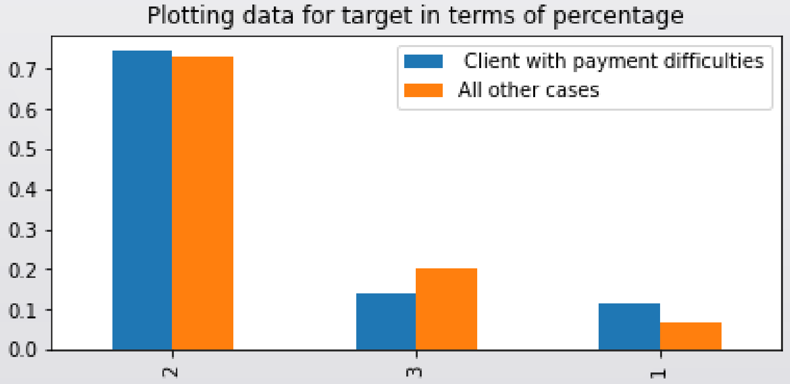
- Ignoring the unknown category, Laborers are the highest occupation type with loans

Univariate Analysis for Region





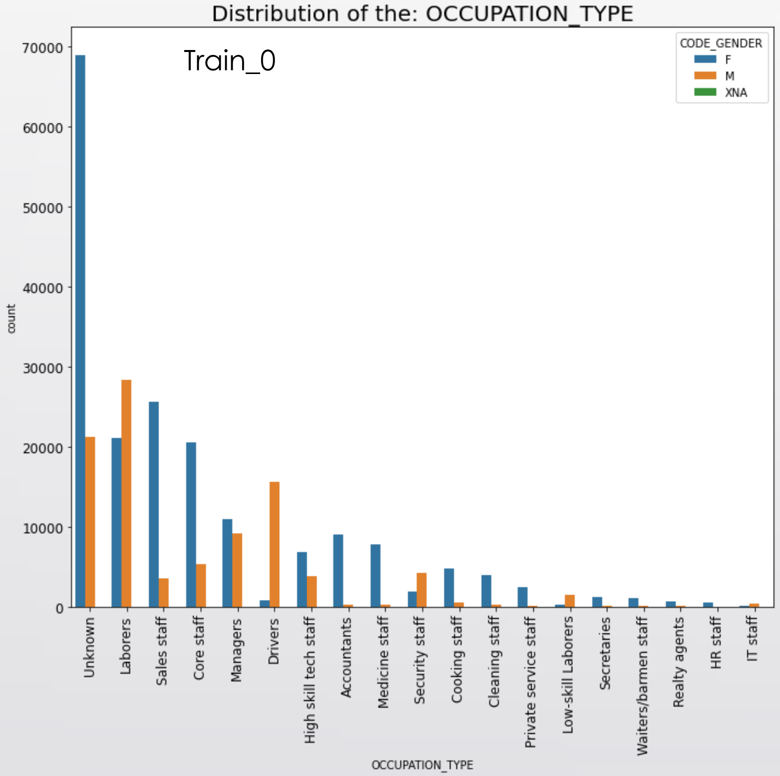


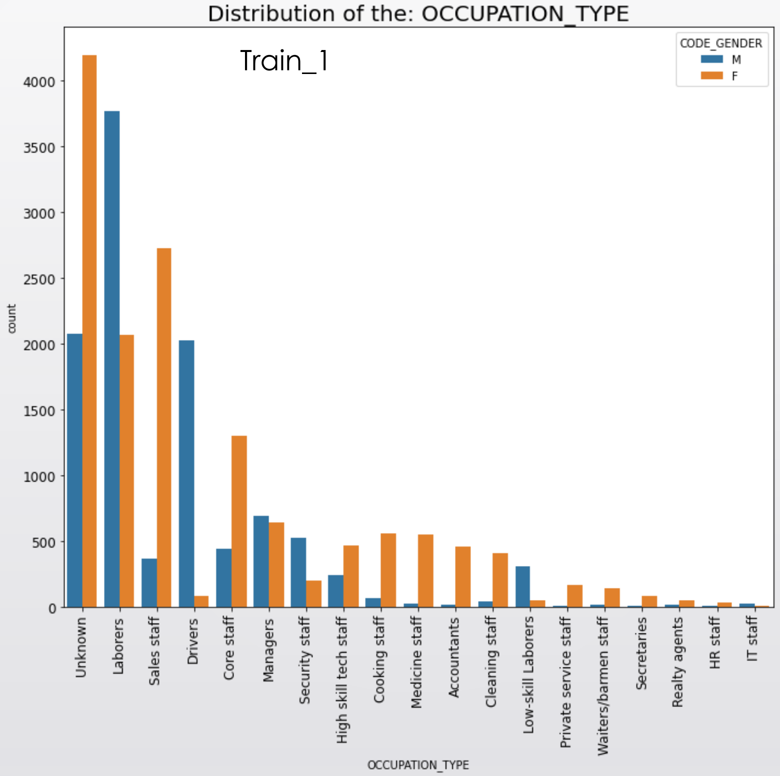


Insight:

- Region 2 has the most applications
- Region 3 has more defaulters
- Region 1 has more repayers

Bivariate Analysis for Occupation vs Gender

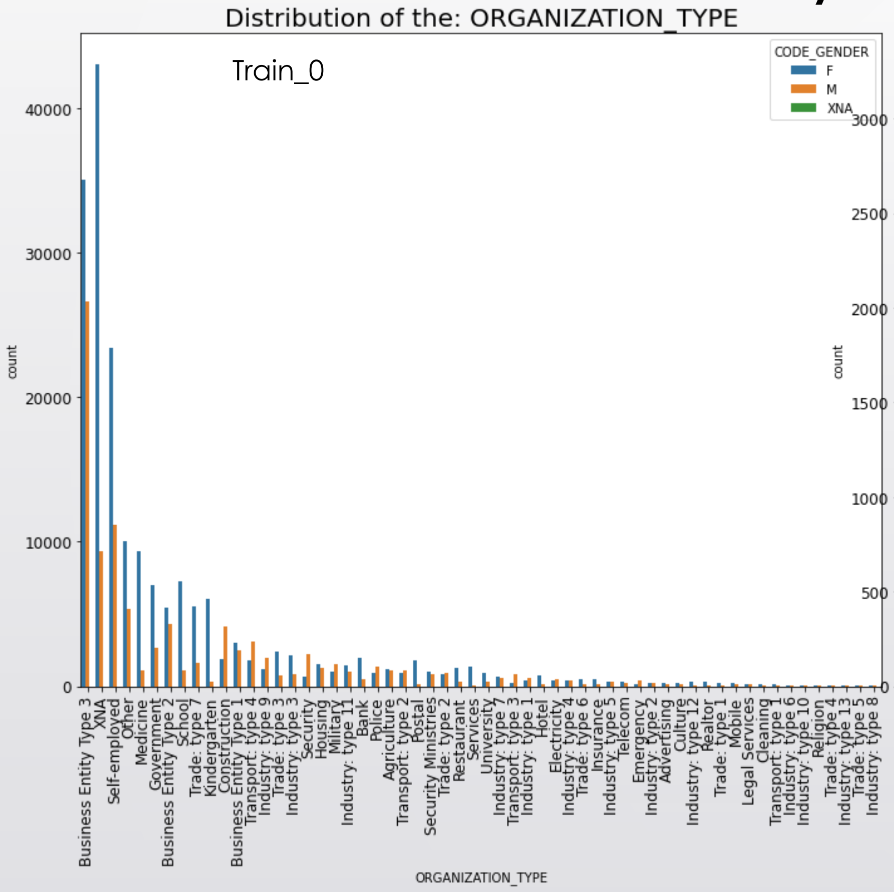


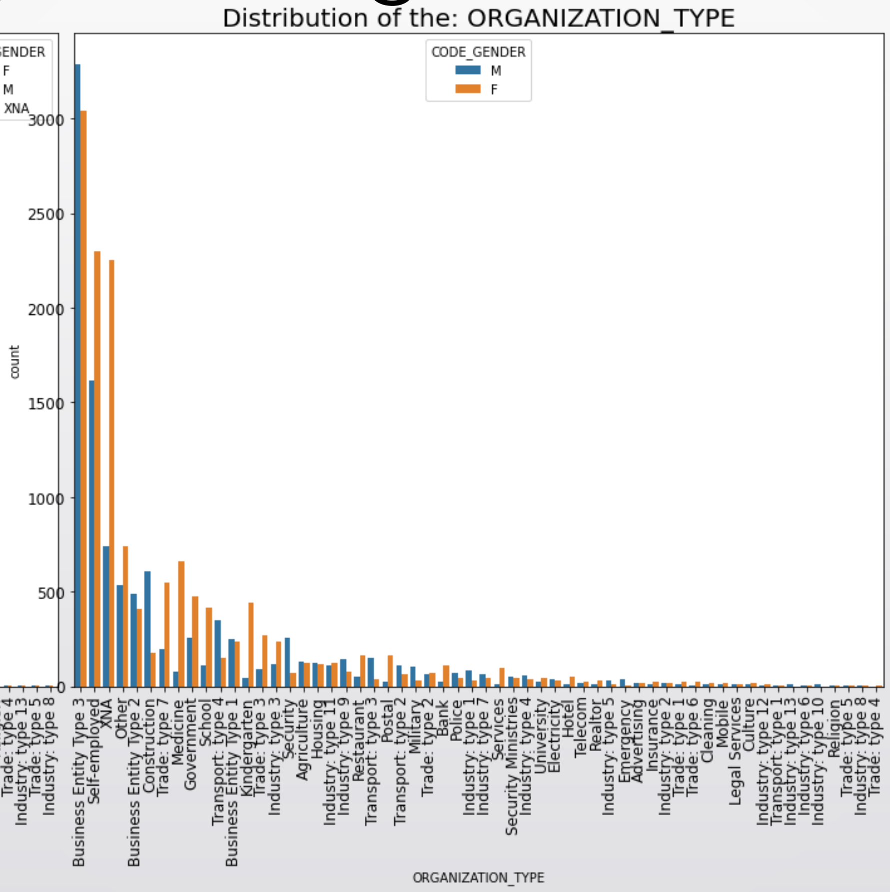


Insights:

- Male laborers have the highest loans and highest defaulters
- More skilled jobs have lower defaulters

Bivariate Analysis for Organization vs Gender

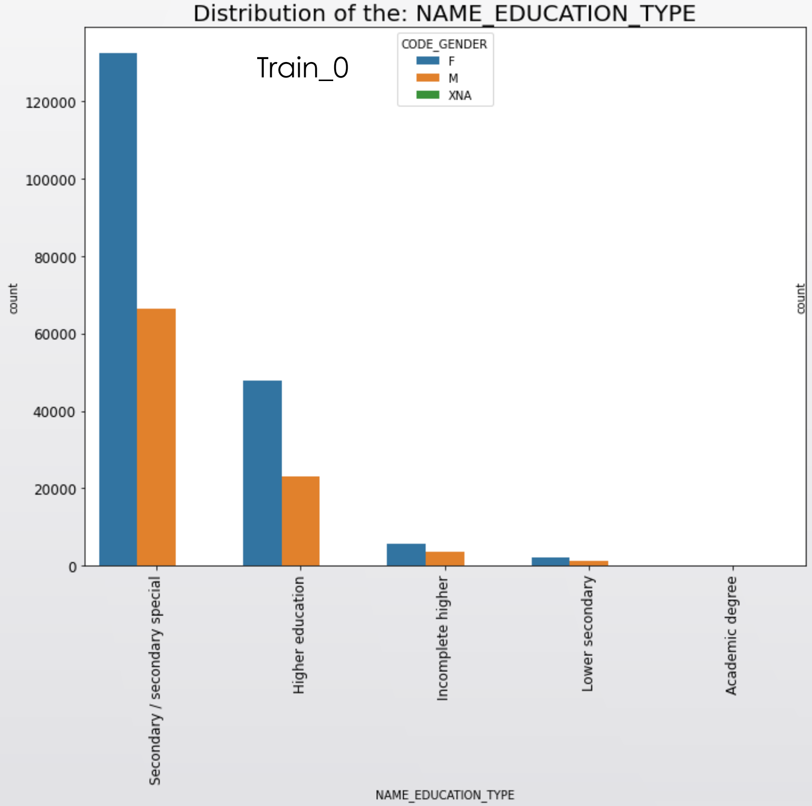


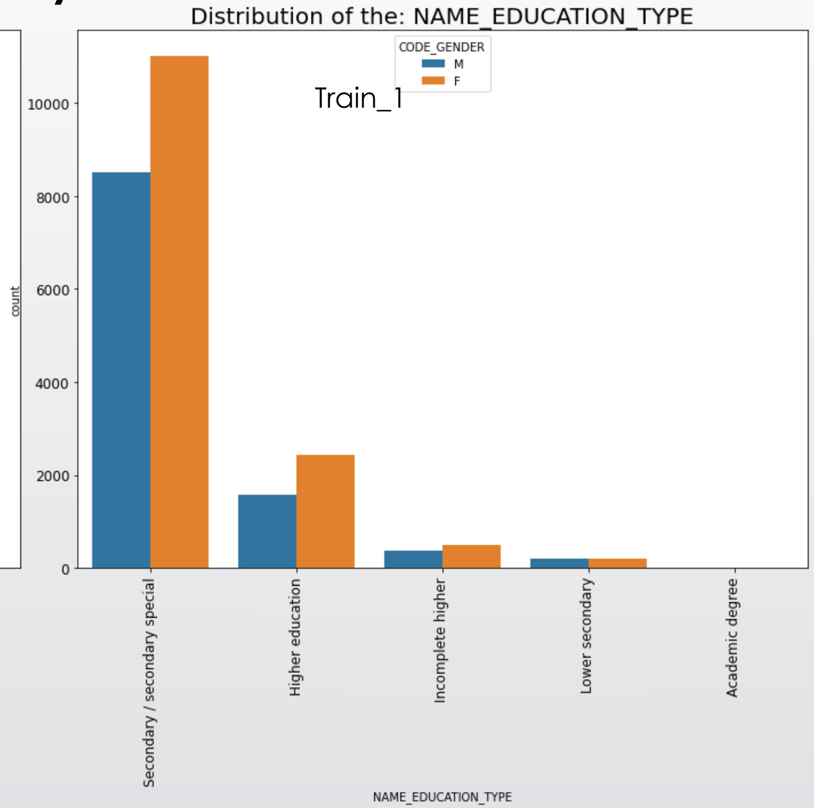


Insights:

- Females are better repayers than males
- Loan applicants who applied for loans majorly belong to the organization type ‘Business entity Type 3’ , ‘Self employed’ , ‘Other’ , ‘Medicine’ and ‘Government’.
- Payment defaulters are the most in 'Business Entity Type 3', 'Self employed', 'other' categories.

Bivariate Analysis for Education vs Gender

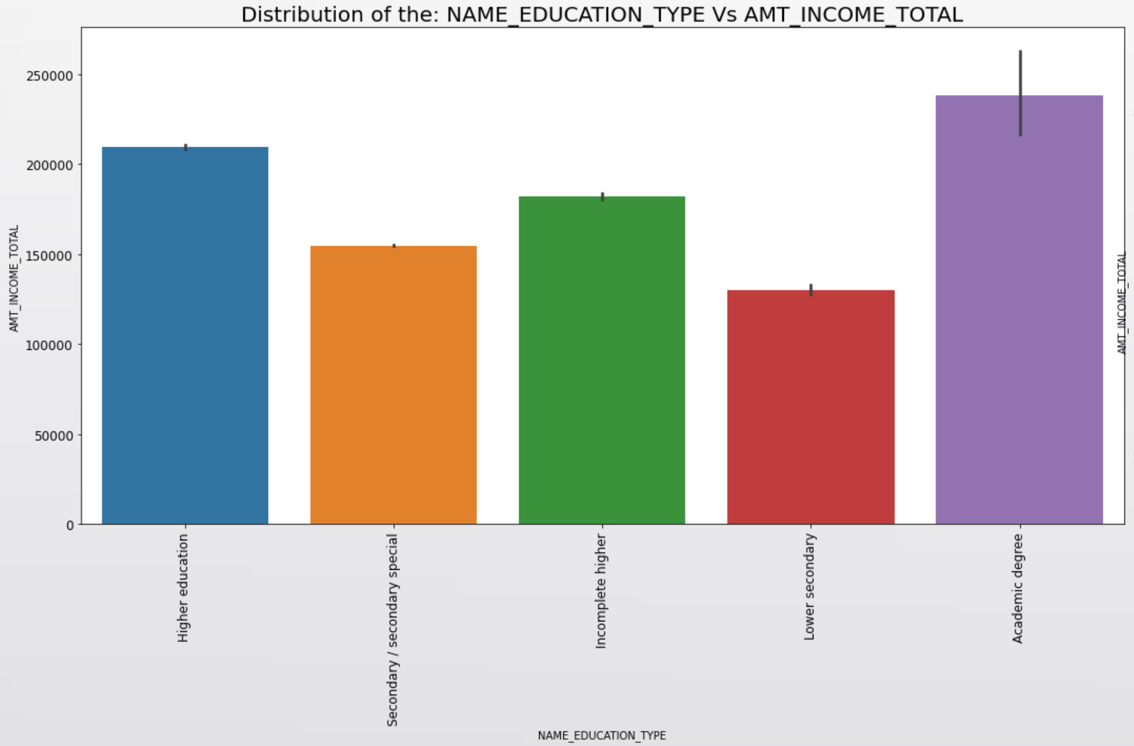


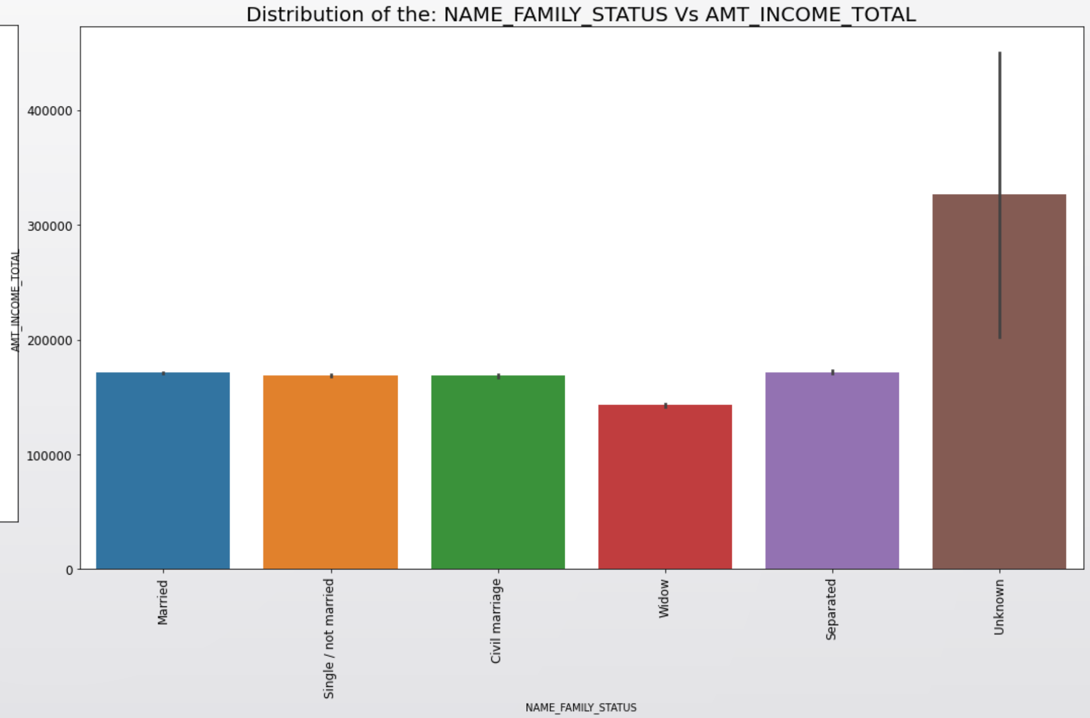


Insights:

- Females are better repayers than males
- People with Secondary education apply for the loans the most in both genders.

Bivariate Analysis Income vs education & family status





Insights:

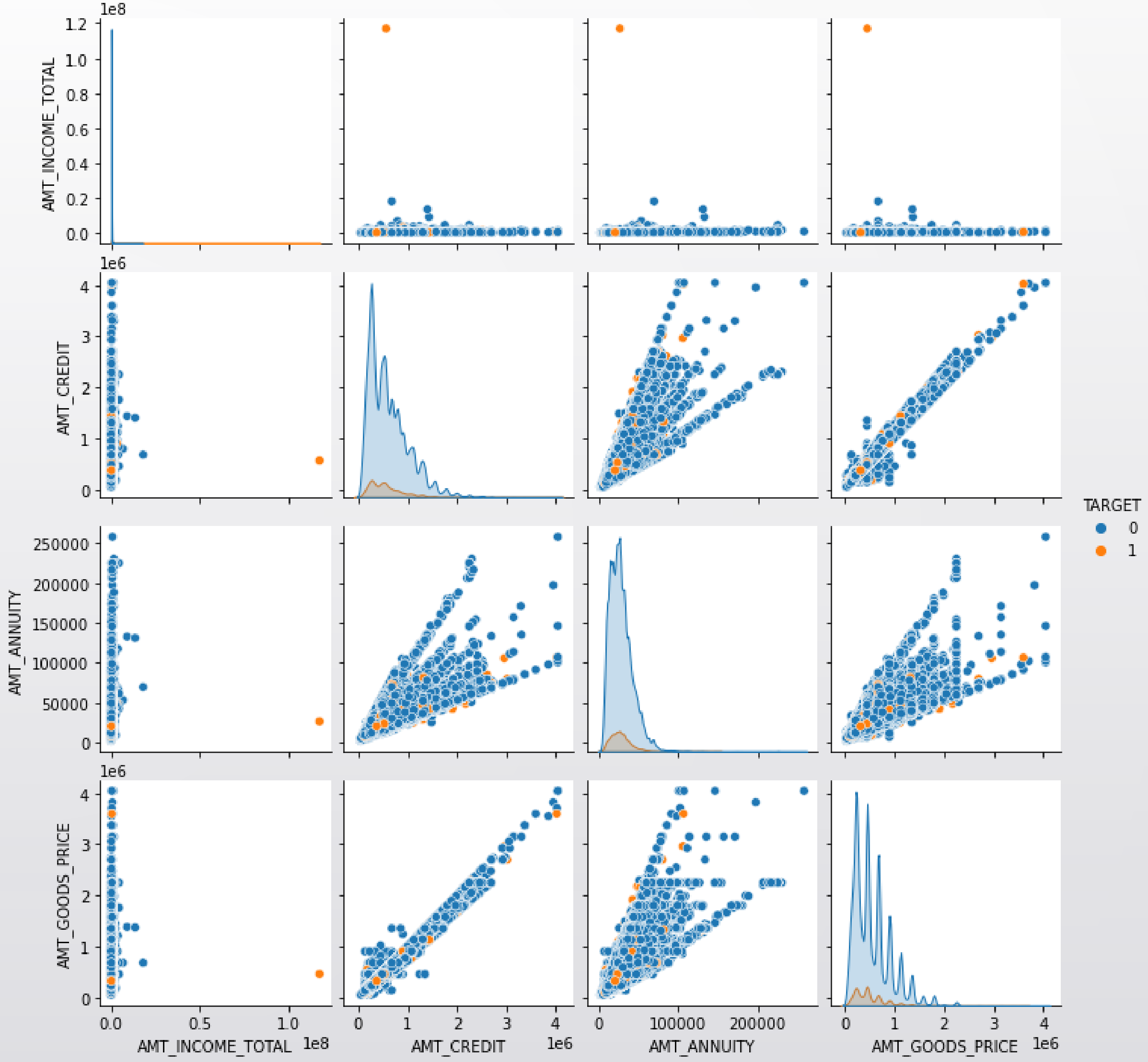
- Academic degree has the highest income
- Barring unknown family status, the income levels are comparable except for Widows

Pair Plot for all AMT cols

AMT_INCOME_TOTAL - Income of the client
AMT_CREDIT - Credit amount of the loan
AMT_ANNUITY - Loan annuity
AMT_GOODS_PRICE - For consumer loans it is the price of the goods for which the loan is given

Insights:

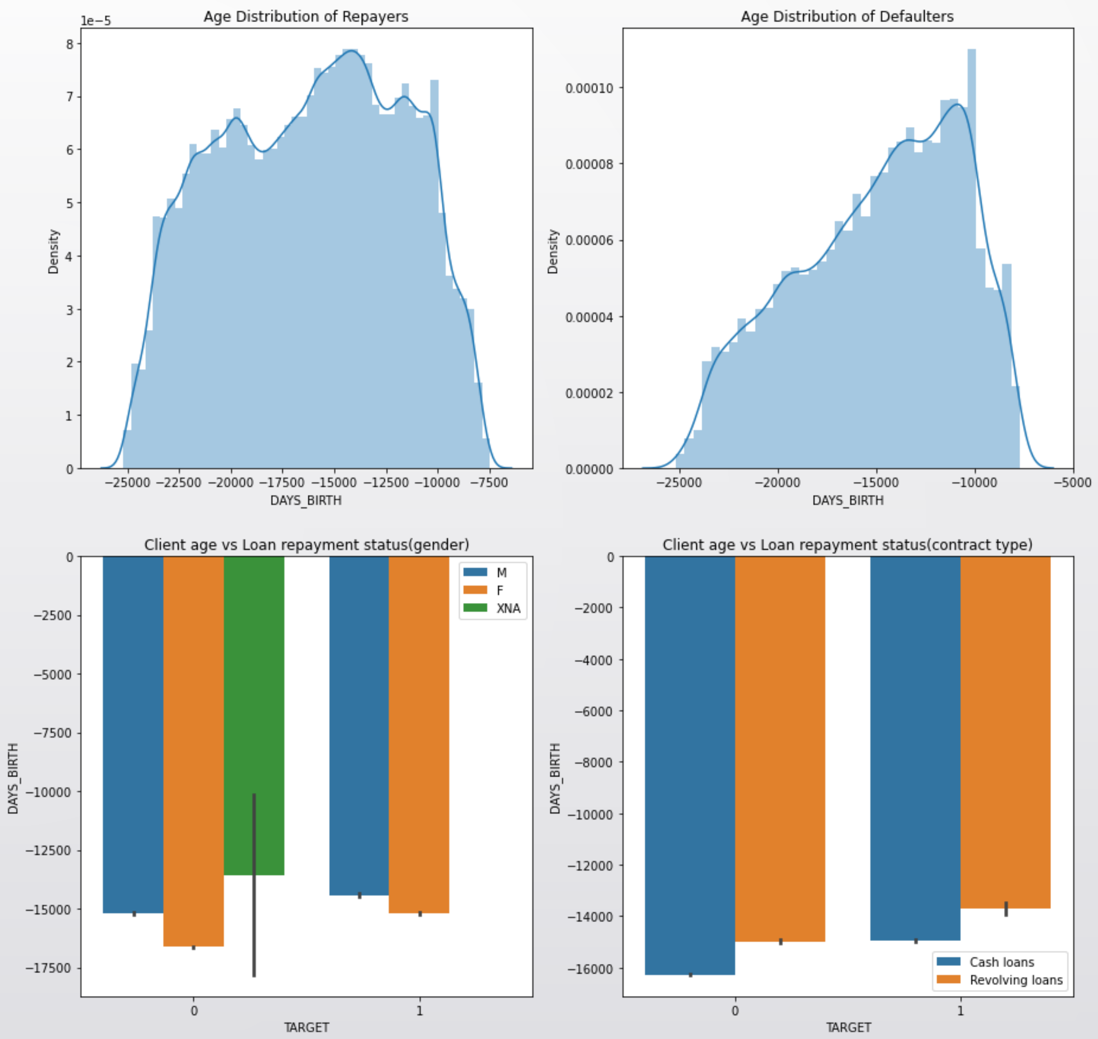
- Strong Positive correlation between AMT_CREDIT and AMT_GOODS_PRICE
- Positive correlation between AMT_ANNUITY and AMT_CREDIT



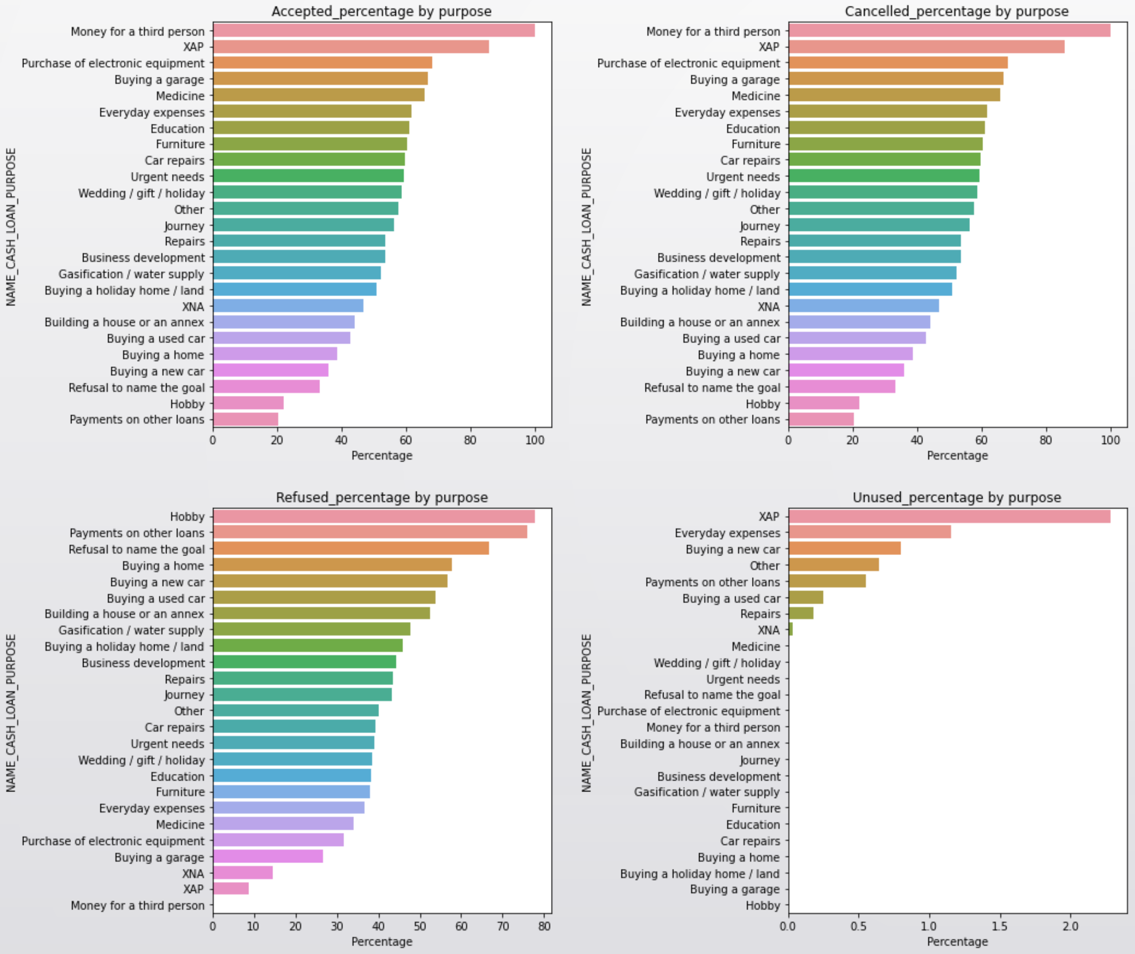
Age vs Loan Repayment

Insights:

- Defaulters are younger than repayers



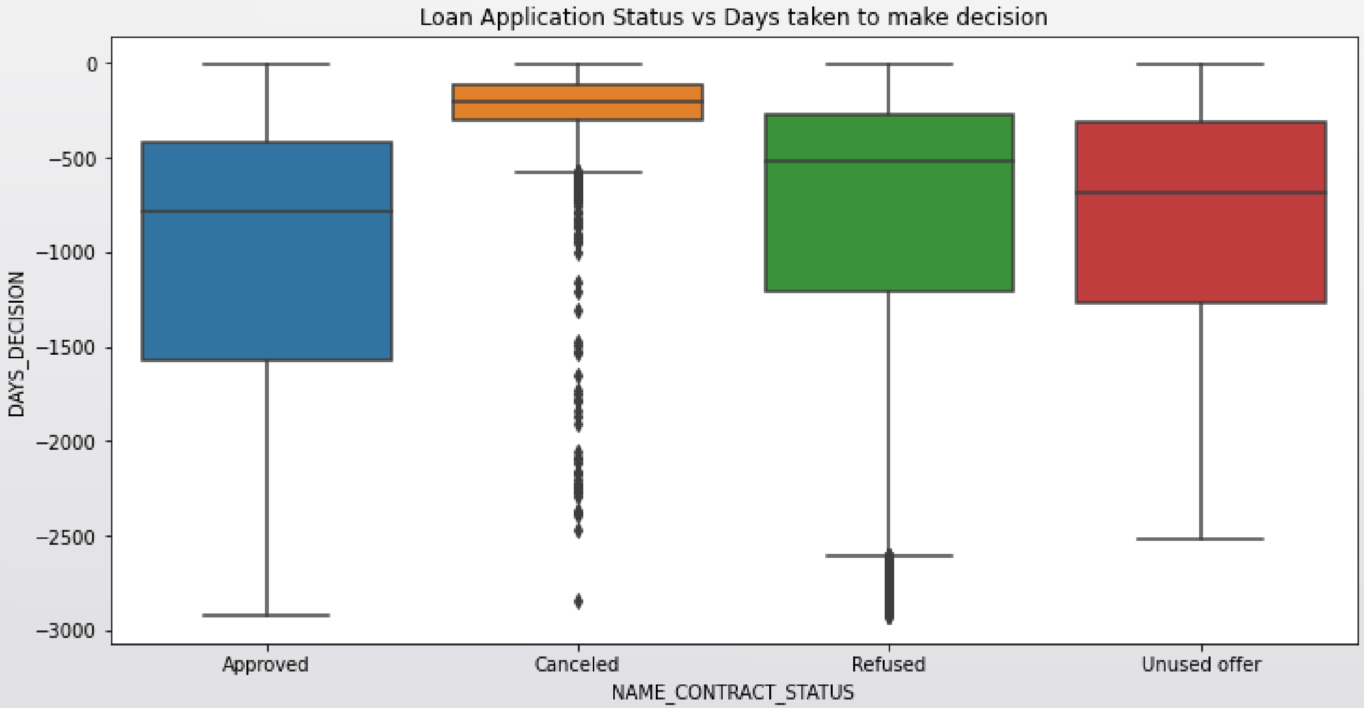
Bivariate Analysis purpose of loan vs Status



Insights

- Money to a third person, XAP , purchase of electronic equipment ,medicine, every day expenses and education have higher loan acceptance.
- 37.5% of XNA purpose loans are cancelled.
- Loan puporses like Hobby, Payment of other loans ,Refusal to name goal, Buying new home or car have higher rejections.
- XAP has has highest unused percentage

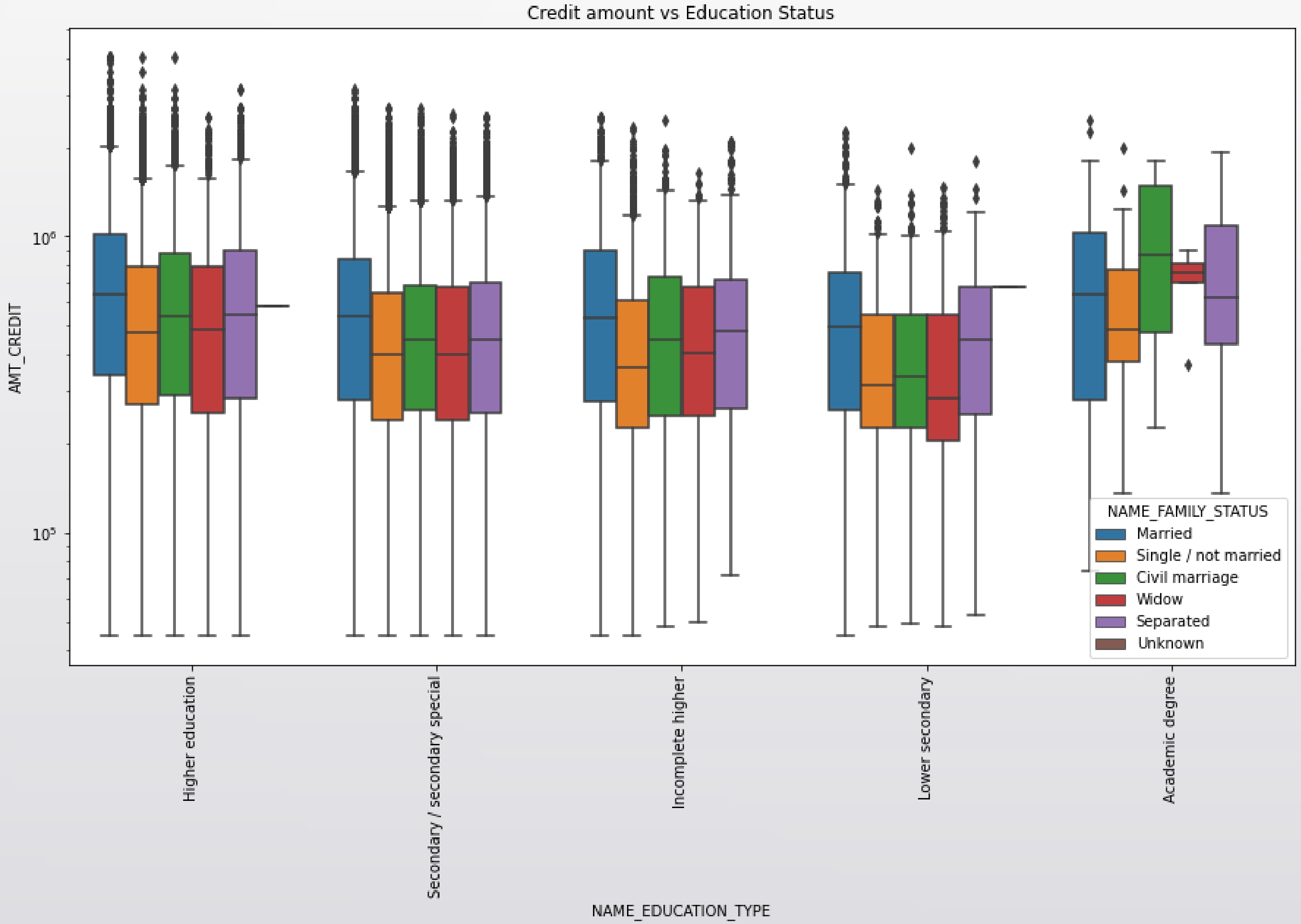
Bivariate analysis: Application status relative to decision made about previous application.



Insights:

- It is observed that on average approved applications have higher number of decision days compared to cancelled, refused offer applications.
- Cancelled applications have a significant number of outliers

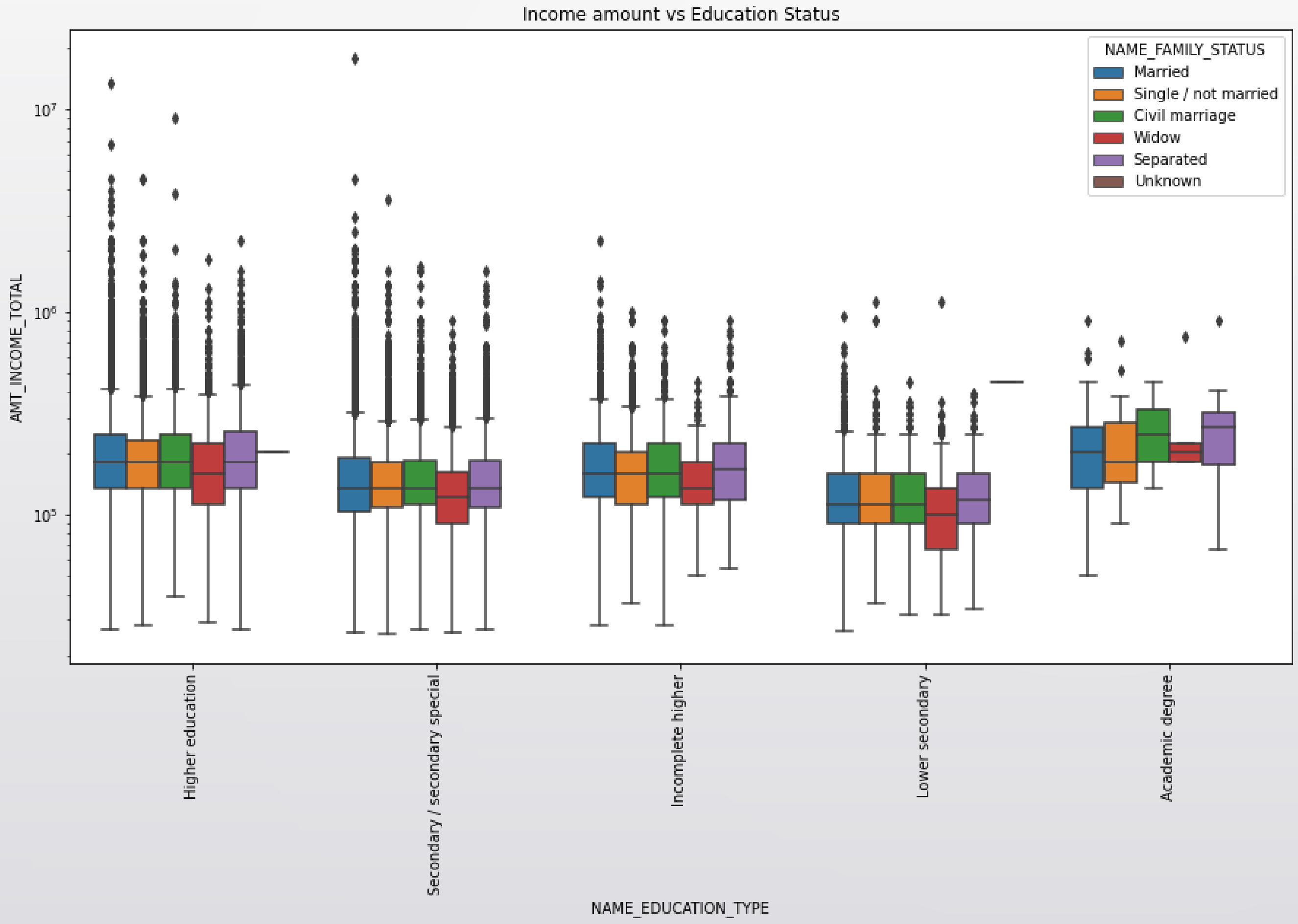
Bivariate Analysis Education vs Credit Amount



Insights:

- From the above box plot we can conclude that Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.
- Also, higher education of family status of 'marriage', 'single' and 'civil marriage' are having more outliers. Civil marriage for Academic degree is having most of the credits in the third quartile.

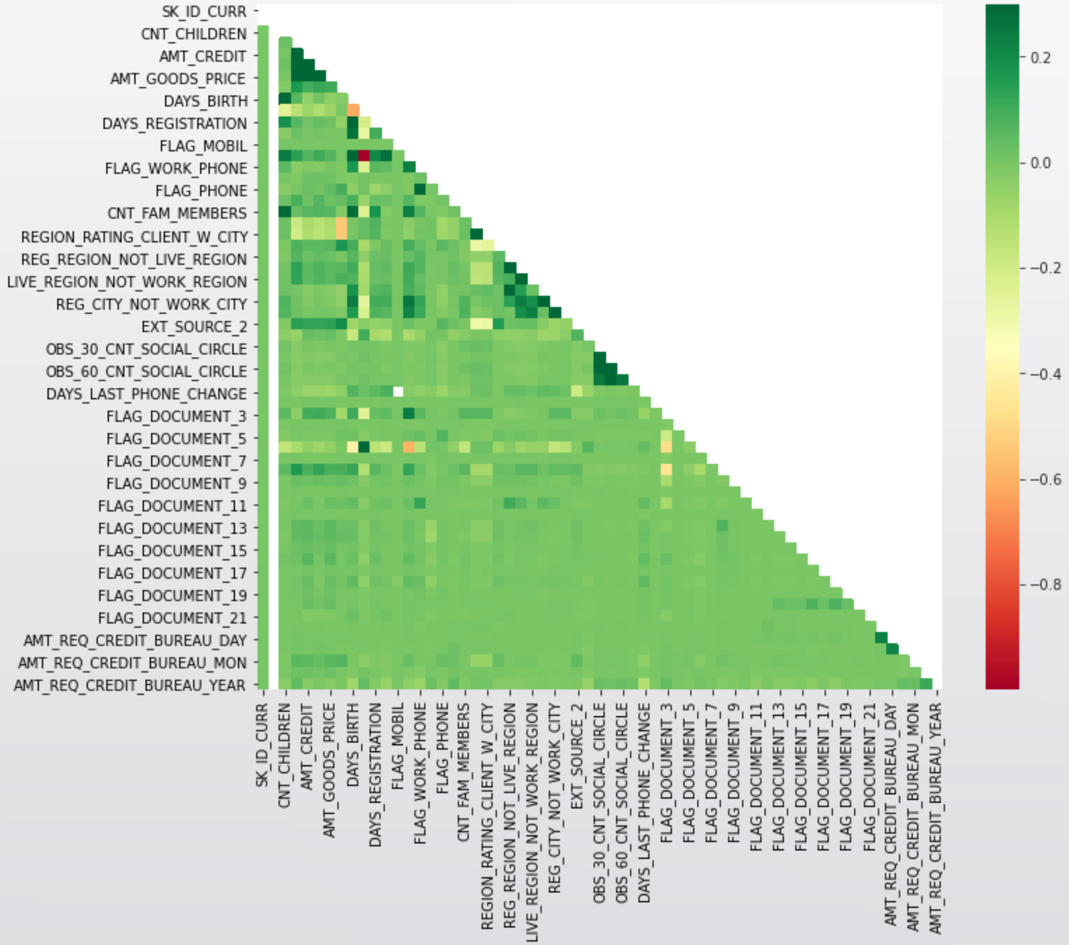
Bivariate Analysis Income vs Education



Insights:

- From above boxplot for Education type 'Higher education' the income amount is mostly equal with family status, though there are many outliers.
- Fewer outliers for Academic degree but the income amount is little higher that Higher education.
- Lower secondary of civil marriage family status are have less income amount than others.

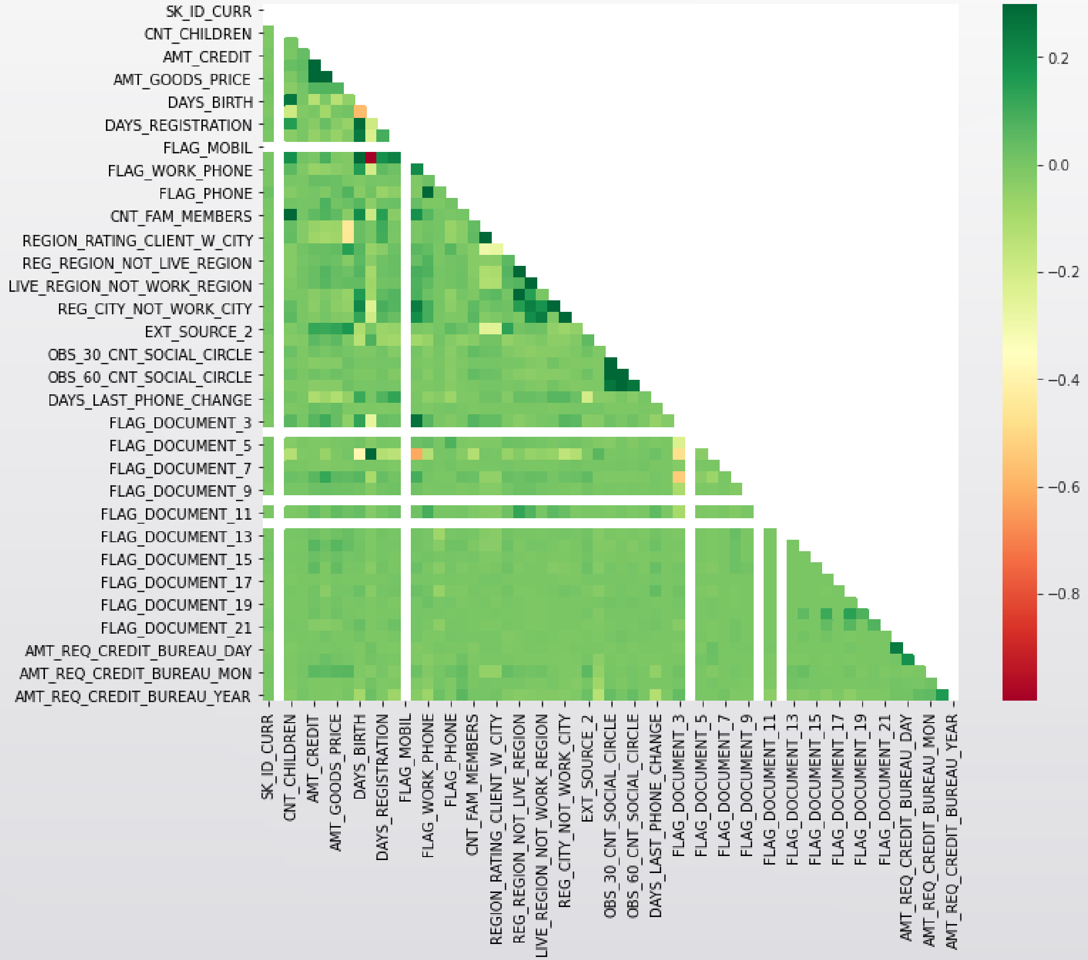
Correlation analysis on train_0



Top 10 correlations:

CNT_CHILDREN	CNT_FAM_MEMBERS	0.878571
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950149
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149
AMT_CREDIT	AMT_GOODS_PRICE	0.987022
AMT_GOODS_PRICE	AMT_CREDIT	0.987022
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998510
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998510
DAYS_EMPLOYED	FLAG_EMP_PHONE	0.999758
FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999758

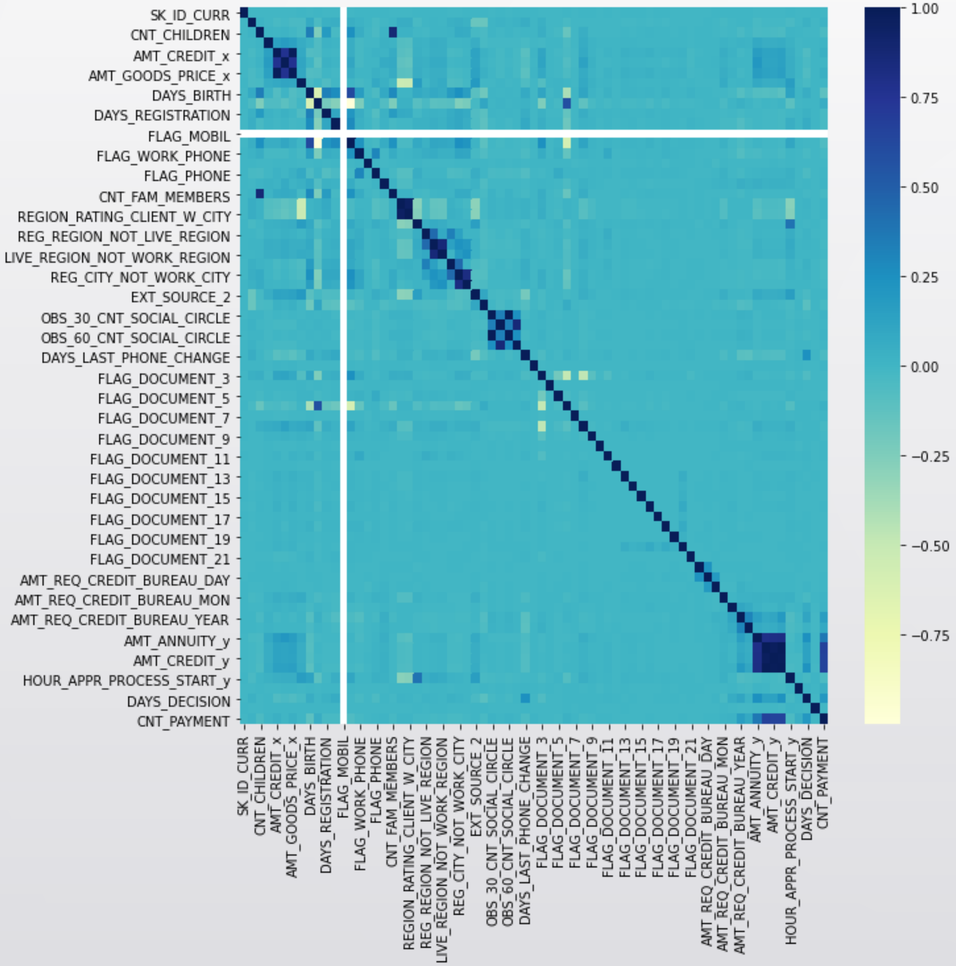
Correlation analysis on train_1



Top 10 correlations:

CNT_CHILDREN	CNT_FAM_MEMBERS	0.885484
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956637
AMT_CREDIT	AMT_GOODS_PRICE	0.982783
AMT_GOODS_PRICE	AMT_CREDIT	0.982783
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998270
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998270
FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999702
DAYS_EMPLOYED	FLAG_EMP_PHONE	0.999702

Correlation after merging



Top 10 correlations

AMT_CREDIT_x	AMT_GOODS_PRICE_x	0.986397
AMT_GOODS_PRICE_x	AMT_CREDIT_x	0.986397
AMT_CREDIT_y	AMT_GOODS_PRICE_y	0.992128
AMT_GOODS_PRICE_y	AMT_CREDIT_y	0.992128
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998503
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998503
FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999772
DAYS_EMPLOYED	FLAG_EMP_PHONE	0.999772
AMT_GOODS_PRICE_y	AMT_APPLICATION	0.999940
AMT_APPLICATION	AMT_GOODS_PRICE_y	0.999940

Conclusions:

Since, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The following conclusions were observed:

1. It is observed that people with higher levels of education apply for more loans and the default rate is very low, the company can should use this as a metric for risk of defaulting
2. Females apply for loans more than Males and have a lower defaulting rate
3. Owning a car has an inverse relationship with defaulting on loans whereas housing type of home has a direct relationship
4. Region rating is a possible indicator of defaulting
    - Region 2 has the most applications
    - Region 3 has more defaulters
    - Region 1 has more repayers
5. More skilled jobs have lower defaulters Managers are getting high salary and Laborers are getting neither high nor low, to satisfy their family needs more laborers are taking loans.
6. Higher Income and Educational levels have lower loans/default rates
7. Age is also a strong indicator of defaulting. Lower the age, higher the chances of deafaulting
8. Reason/Purpose for loan also shows a pattern: Education, medicine, equiment purchase have higher acceptance rates, whereas Hobby, Payment of other loans ,Refusal to name goal, Buying new home or car have higher rejections.
9. Days taken to approve loan are higher than refusal, cancels relative to previous application
10. There is a very strong correlation between
    - Amt_credit vs Amt_Goods_Price 0.986397
    - Amt_Application vs Amt_Goods_Price 0.999940

The company can use the above mentioned variables to lower risk of defaulters and increase loan approvals to increase business!