

EDA CASE STUDY (Loan Defaulter)

Case Study (Partners):

Siddhesh Sanjay Borge and Sarath Chandra Sobhira



Problem Statement

➤ Introduction

- This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers

➤ Business Understanding

- 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 specializes in lending various types of loans to urban customers. You have to use EDA to analyze 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 instalments of the loan in our sample,
 - All other cases: All other cases when the payment is paid on time.
- When a client applies for a loan, there are four types of decisions that could be taken by the client/company):
 1. **Approved:** The Company has approved loan Application
 2. **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.
 3. **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
 4. **Unused offer:** Loan has been cancelled by the client but on different stages of the process.
- In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Problem Statement

➤ Business Objectives

- 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.
- To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

➤ Data Understanding

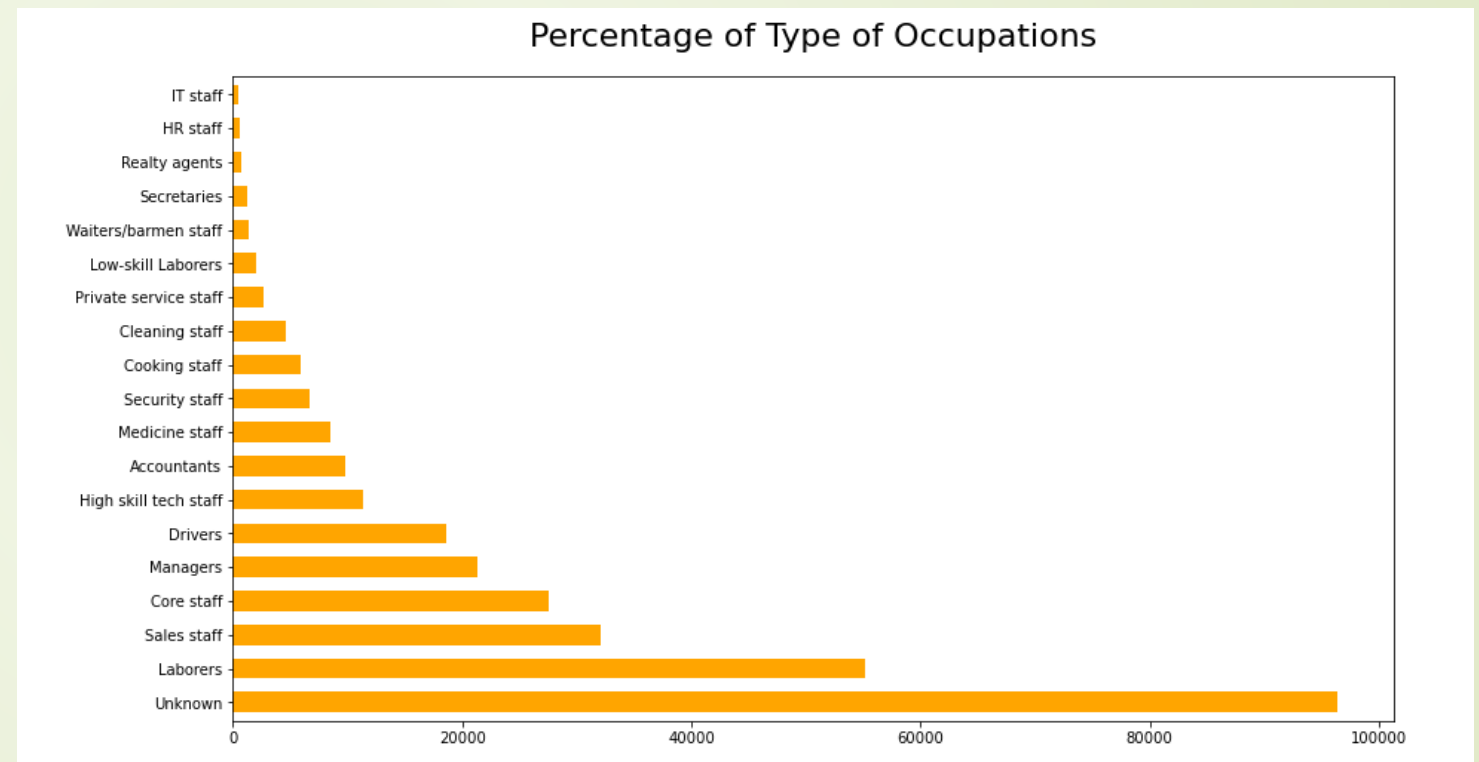
- This dataset has 3 files as explained below:
- 1. '*application_data.csv*' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
- 2. '*previous_application.csv*' contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
- 3. '*columns_description.csv*' is data dictionary which describes the meaning of the variables.



Current Application DATA

Imputing the Value Counts of Occupation Type

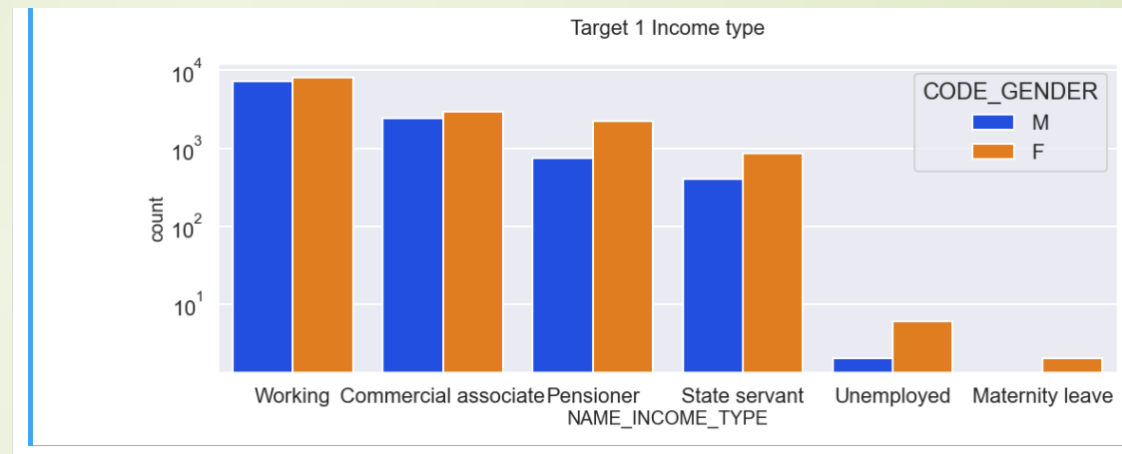
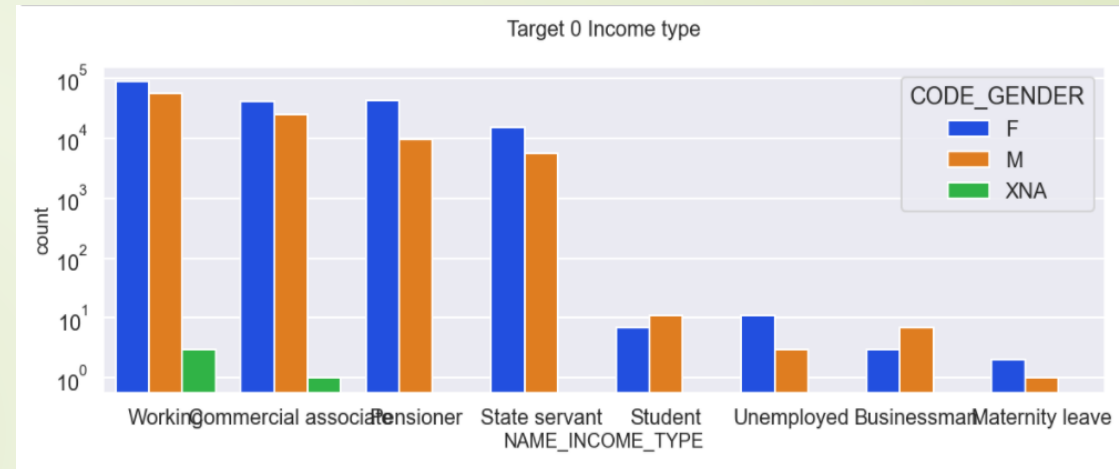
- So from here we can clearly see that the Highest value counts in Occupation type belongs to the Unknown and Laborers which generally comprises of low-income group.
- While the lowest value counts are IT staff and then HR staff of higher income ones.



Univariate Analysis for 'Income Type'

Income Type Analysis of '**Target 0**' and '**Target 1**' :

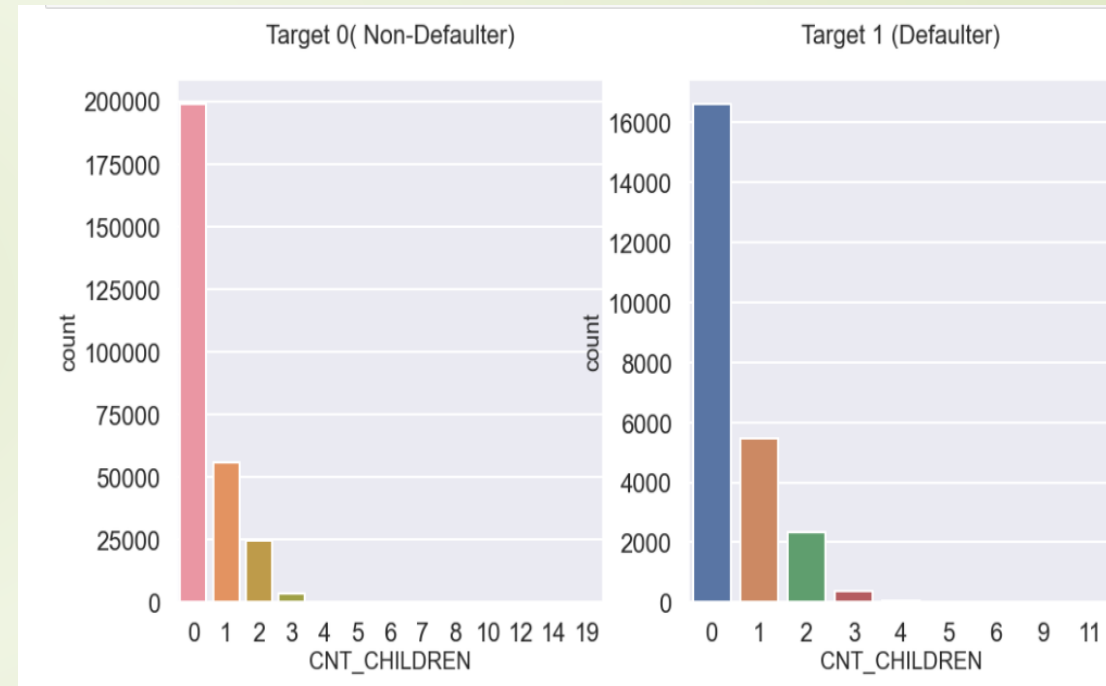
- For 'Target 0 Income Type', the low number of credit is for student , unemployed , businessman and maternity leave, while the high number of credit is for working, commercial associate, pensioner and state servant.
- For 'Target 1 Income Type' , there is a higher number of credit for income type such as Working, Commercial associate, Pensioner, State Servant, while there is a number of lower credit for income type for the unemployed and maternity leave.



Univariate Analysis for 'CNT Children'

Analysis of **Target 0** (Non-Defaulter) and **Target1** (Defaulter) :

- From the graph we cannot explore much as the low child counts maximizes of chance being a defaulter(Target 1) and Non-defaulter (Target 2).



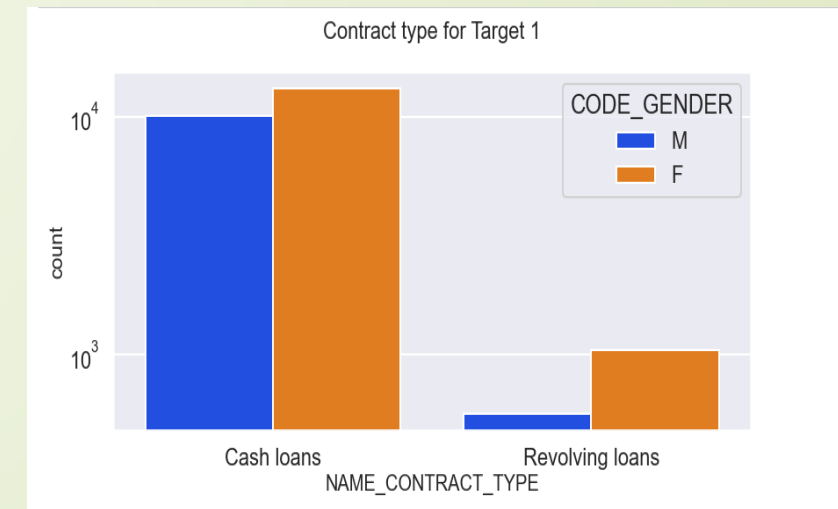
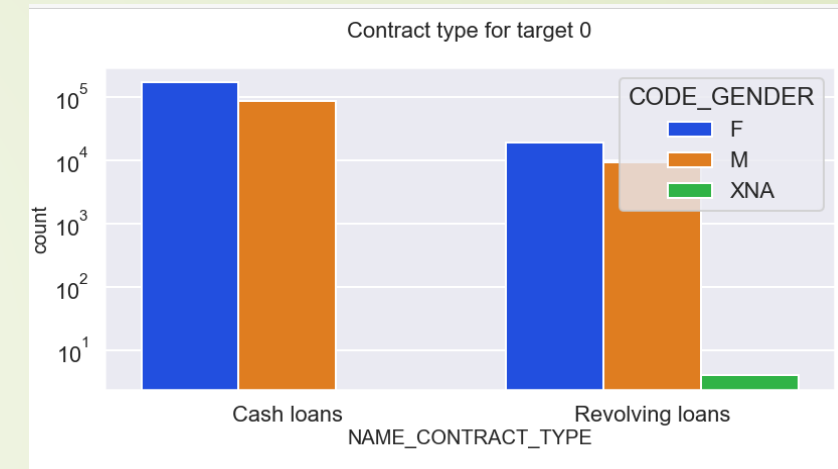
Univariate Analysis for 'Contract Type'

Analysis for 'Contract type' for **Target 0** (Non-defaulter):

- Here the count of Females are more than males.
- The Cash loans contracts have relatively higher contracts from revolving loans.

Analysis for 'Contract type' for **Target 1** (Defaulter):

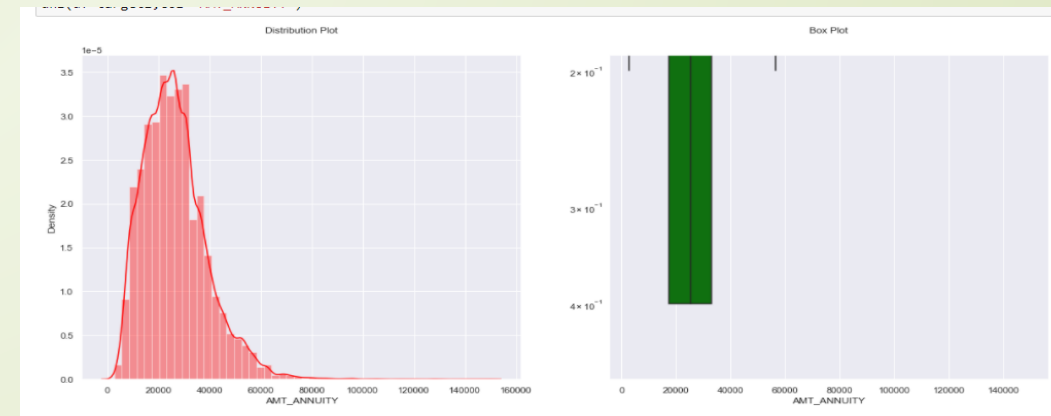
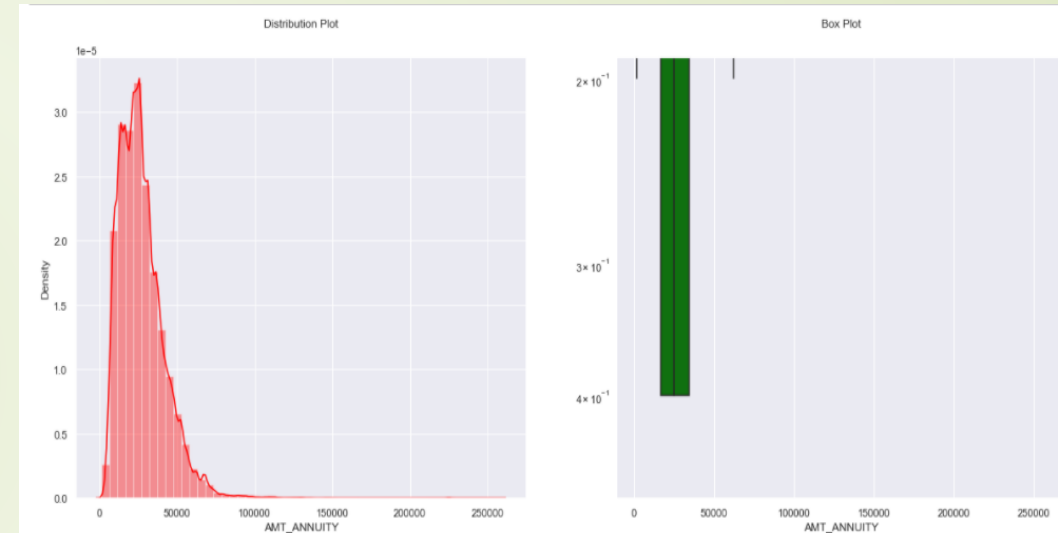
- The cash loan contracts have higher number of credit than revolving loan contracts.
- In terms of Revolving loans contract, the female count is much more higher than the males.



Univariate Analysis for Numerical Variable of 'Target Variable'

Analysis of 'Target 0' and 'Target 1' of 'AMT_ANNUITY':

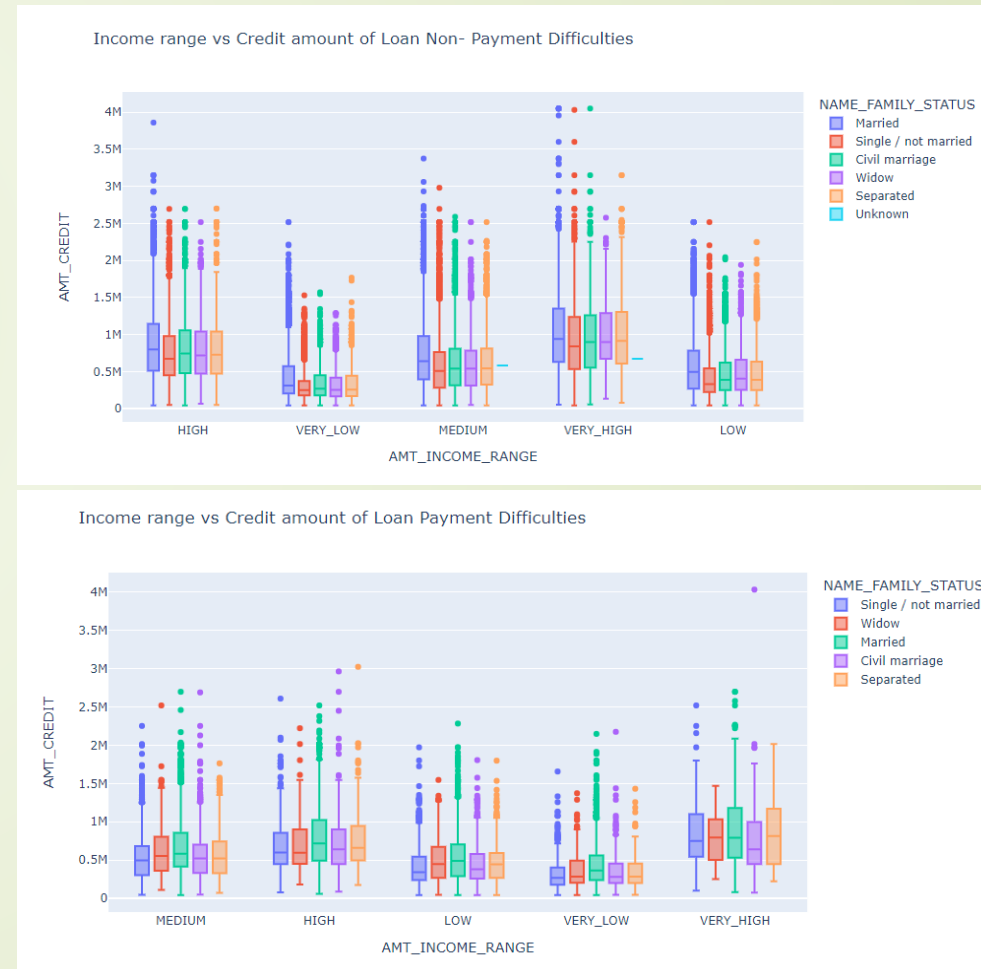
- We can clearly see some outliers, that the third quartile is smaller than the first quartile.
- Which proves to be relatively bigger and most of the Annuity client are in the first quartile.



Bivariate Analysis (Analysis 1) : 'AMT_INCOME_RANGE' and 'AMT_CREDIT'

Analysis 1 :

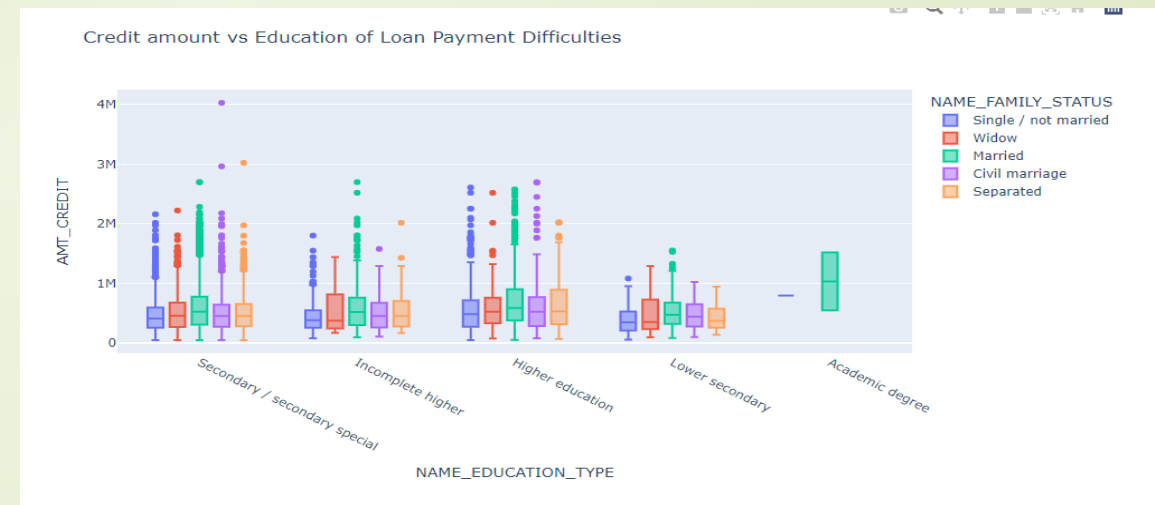
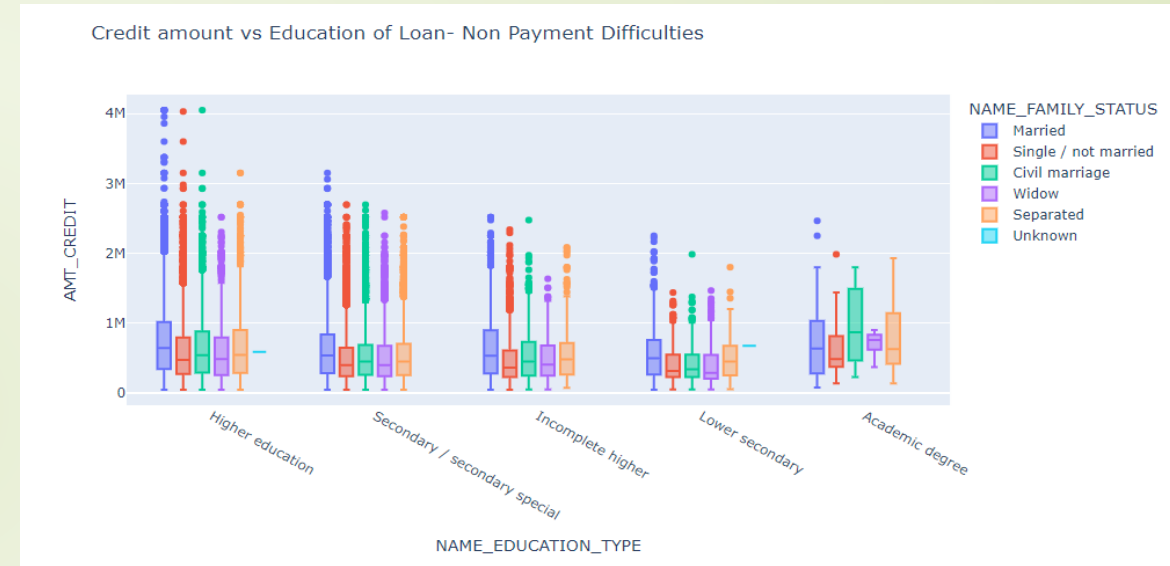
- It is been quite interesting that the graph of both the payment and non-payment difficulties are found to be similar.
- The family status of single , married and separated income range are very high and the credit range is also higher then others.





Bivariate Analysis (Analysis 2) : 'NAME_EDUCATION_TYPE' and 'AMT_CREDIT'

Analysis 2 :

- Most of the outliers are from Education type Higher education and Secondary.
- The higher number of credits is been observed to family status of civil marriage, marriage and separated of Academic degree education.
- Civil marriage for Academic degree is having most of the credits in the third quartile.



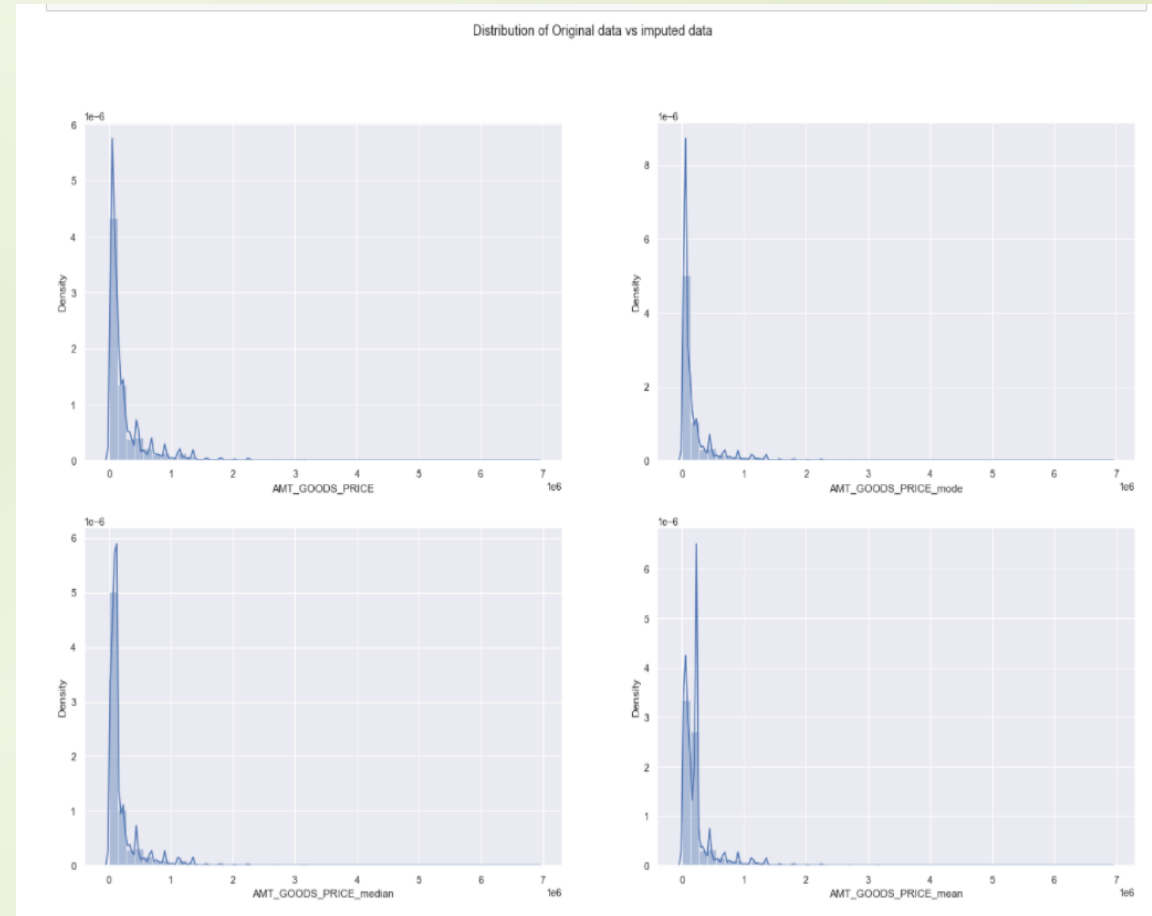


DATA ANALYSIS ON PREVIOUS APPLICATION DATA

Analysis on the Continuous Variable

Analysis:

- Here we have did the distribution of original data of 'AMT_GOODS_PRICE'.
- We have imputed the original distribution data with the Mean, Median and Mode.
- The imputed distribution data is very much closer with the original distribution.



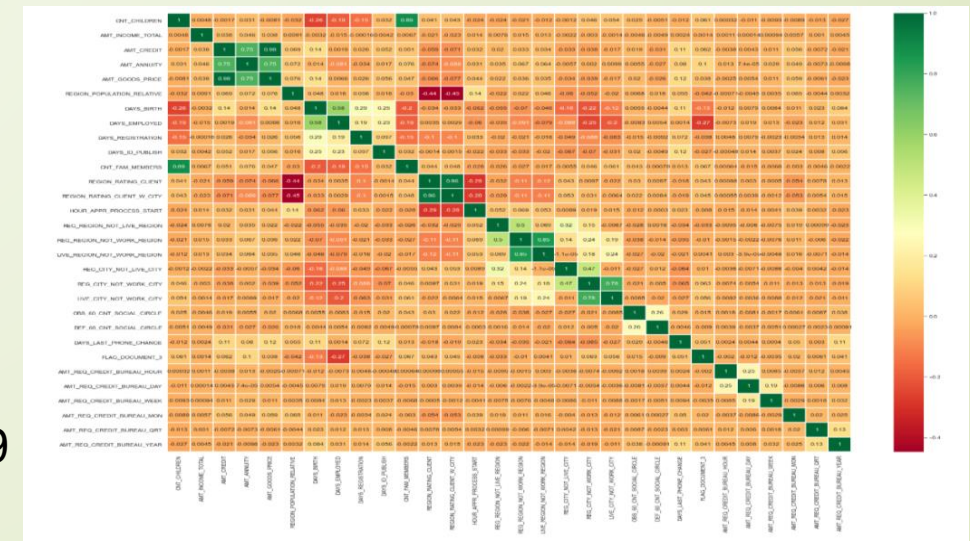
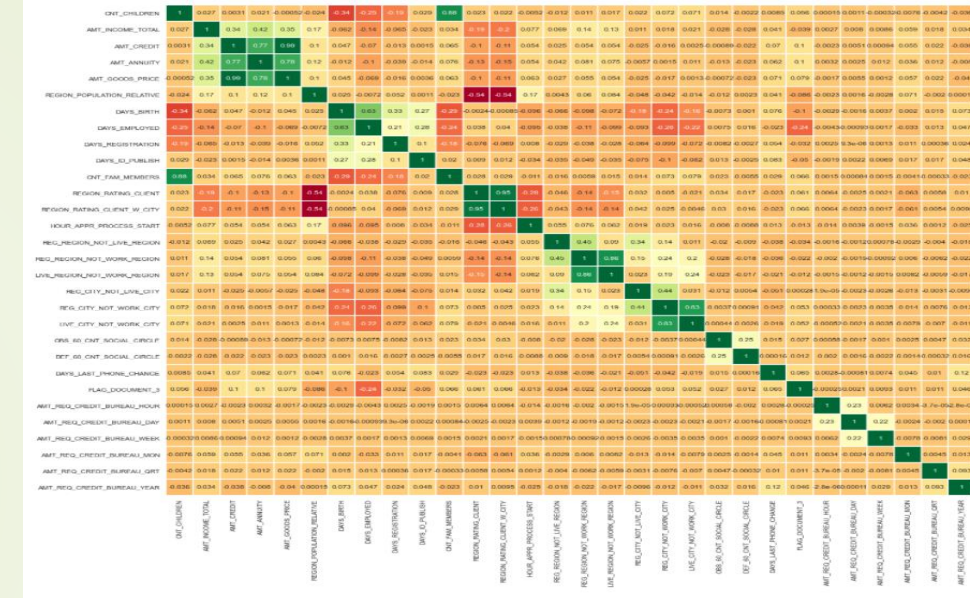
Correlation and the Heat Map

Correlation of Repayers Data:

- We can see that the repayers have a high correlation in the number of days employed.
- Credit amount is highly co-related with Good Price amount, Loans Annuity and Total income.

Correlation of Defaulters Data:

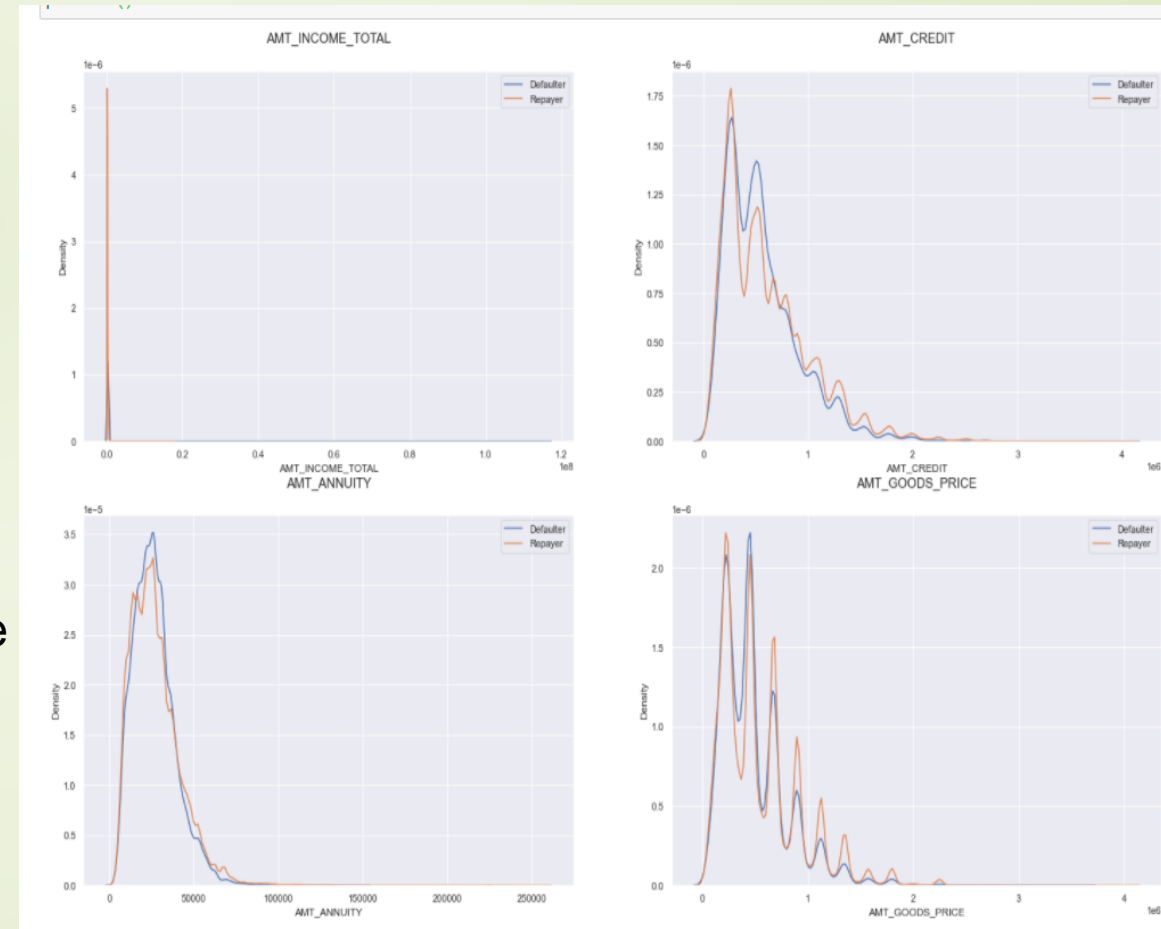
- There is a slight increase in defaulted to observed count in social circle among defaulters(0.264) when compared to repayers(0.254)
- Credit amount is highly correlated with good price amount.
- Defaulters have a less number of co-relation with number of days which is 0.58 while with the repayers it is 0.62 which is high.
- Loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayers(0.77)
- Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in repayers.



Numerical Univariate Analysis : 'AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUIITY', 'AMT_GOODS_PRICE'

Analysis:

- Most people pay annuity below 50K for the credit loan.
- Many loans are given at good prices below 10 Lakhs.
- Credit amount of the loan is mostly less then 10 lakhs.
- The repayers and defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision.



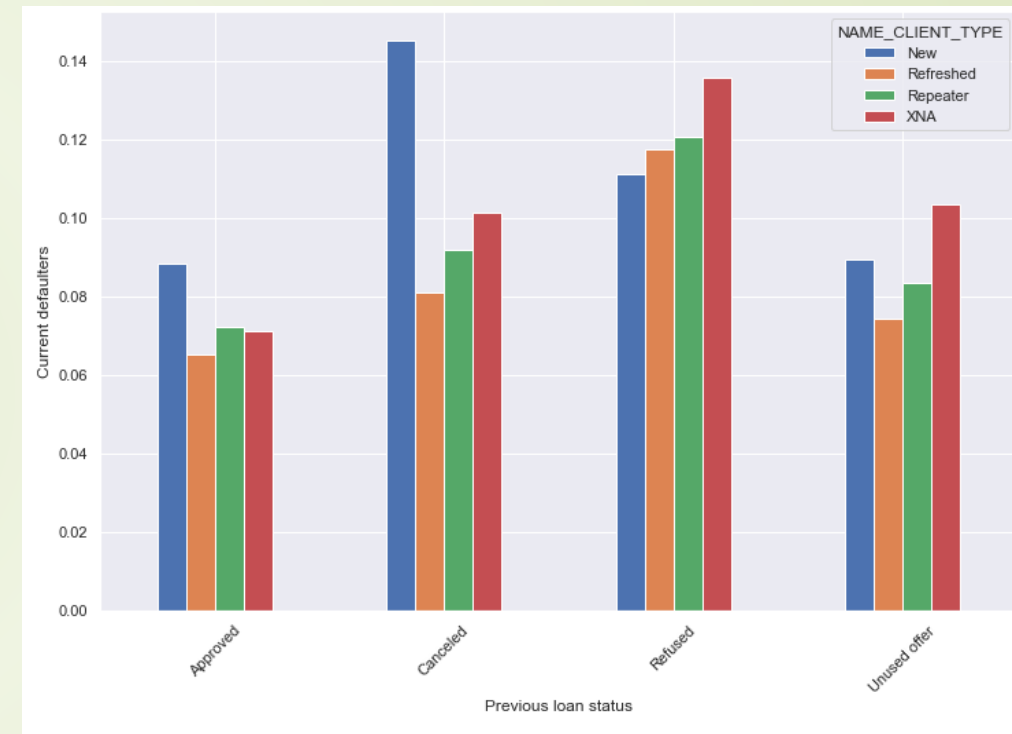


Merging Data Frames of Current application data and previous application

Data Frame of current defaulters with previous loan status and the 'Name_Client_Type'

Analysis:

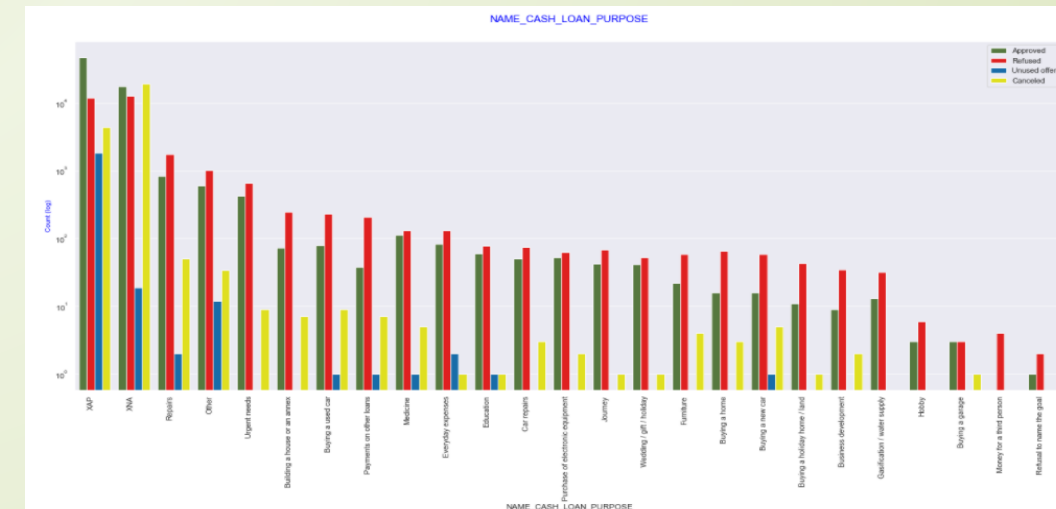
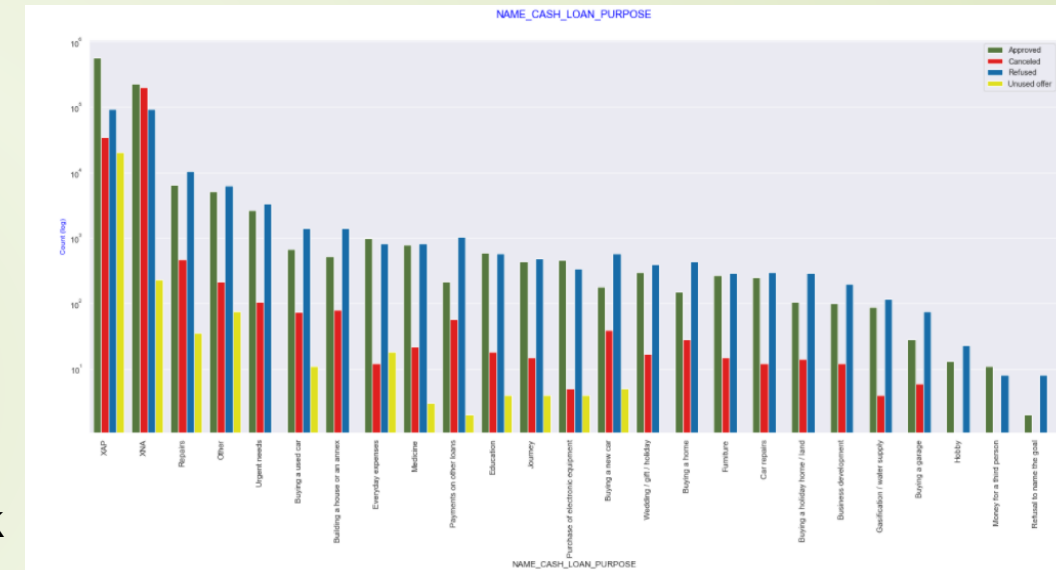
- The previously Approved status the New clients were more defaulted followed by Repeater.
- For previously Cancelled applicants the Defaulters are more New clients.
- We can see that the Defaulters are more for previously Unused offers loan status clients, who were found to be new.
- Previously Refused applicants and the defaulters are more refreshed clients.



Representation of Univariate merged of Name_Cash_loan_Purpose and Name_Contract_Status

Analysis of '**Name_Cash_loan_Purpose**' and '**Name_Contract_Status**' :

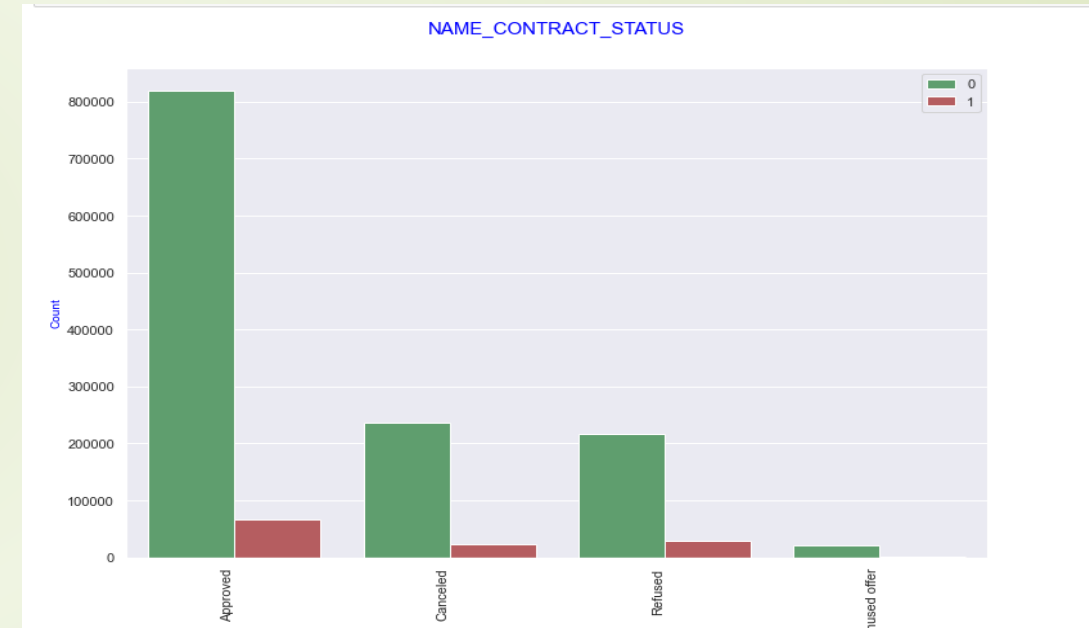
- A very high number application have been rejected by bank or refused by client which has purpose as repair or other. This shows that purpose repair is taken as high risk by bank and either they are rejected or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the loan.
- Loan which is taken for the purpose of Repairs seems to have highest default rate.
- Loan purpose has high number of unknown values.



Univariate Merged of Contract Status based on loan repayment status

Analysis:

- 88% of the clients who have been previously refused a loan has payed back the loan in current case.
- 90% of the previously cancelled client have actually repayed the loan. So, revisiting the interest rates would increase business opportunity for these clients.
- The refusal of loan by the client should be done for further analysis to create a potential repayer customer.



Conclusion of Case Study

From the above analysis we can clearly define who are the defaulters and repayers:

--Reapayers--

- NAME_EDUCATION_TYPE: Academic degree has less defaults.
- NAME_INCOME_TYPE: Student and Businessmen have no defaults.
- AMT_INCOME_TOTAL: Applicant with Income more than 700,000 are less likely to default
- NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, Buying garage are being repayed mostly.
- CNT_CHILDREN: People with zero to two children tend to repay the loans.

--Defaulters--

- CODE_GENDER: Men are at relatively higher default rate
- NAME_FAMILY_STATUS : People who have civil marriage or who are single default a lot.
- NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education
- NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot.



Solutions to make bank more profitable and less Debt

- There are around 90% of the previously cancelled client have actually repayed the loan ,record the reason for cancellation which might help the bank to determine and negotiate terms with these repaying customers in future for increase business opportunity.
- 88% of the clients who were refused by bank for loan earlier have now turned into a repaying client. Hence , documenting the reason for rejection could lead the business loss and these clients could be contacted for further loans which help the bank to make profits.
- It is also recommended to verify thoroughly before giving credit , for instance in terms of Laborer the default value is very high as the bank have not authenticated correctly which may lead to bank losses and debtness and thus should be avoided.
- It is risky to grant loans to the young lower income group section which must be taken care of.