CREDIT RISK ANALYSIS

EDA CASE STUDY

BY:
ANUP KUMAR

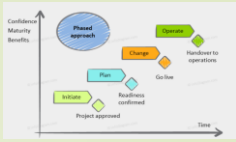
PROBLEM STATEMENT

- Understanding driver variables behind loan default and finding out which variables are strong indicators of a future default.

- This is achieved by using EDA (Exploratory Data Analysis) on the bank data, wherein we analyze the pattern present in data and help bank in mitigating two types of risk associated with loan approvals:

  - 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

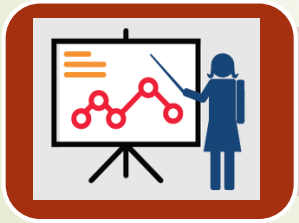
APPROACH



- Importing and cleaning of provided data

- Formatting or Grouping for an effective analysis

- Performing Univariate & Bivariate analysis on categorical and Numerical fields



- Draw useful Insights

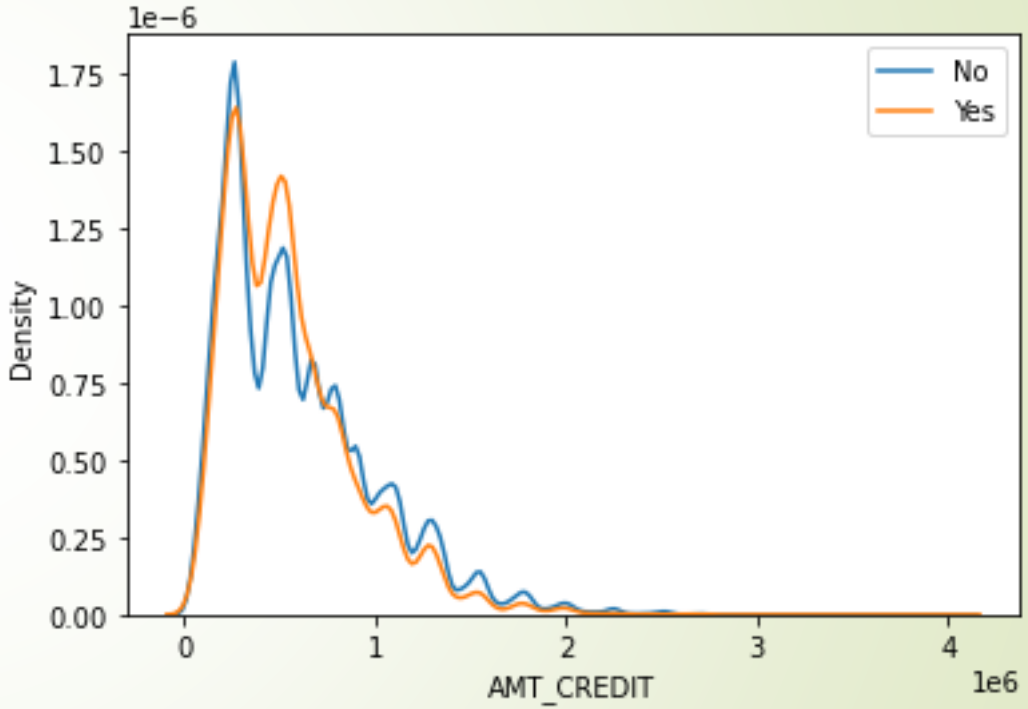
DATA CLEANING

- Dropping all columns with 50% and above null values
- Further dropping columns with 15% and above null values if they are not important with reference to out analysis
- Checking for outliers in important numerical columns
- Based on data dictionary further identifying the columns which are not required for our analysis and hence dropping them
- DAYS_EMPLOYEED have a very high positive value which is not possible and hence replaced by null
- DAYS columns have negative values and hence all have changed to absolute number
- Few columns have 'XNA' or 'XAP' and hence need to be imputed

APPLICATION DATA

Contains information of the client at the time of loan application. The data captures whether a client has payment difficulties

Almost 92% of the clients don't have any payment difficulties

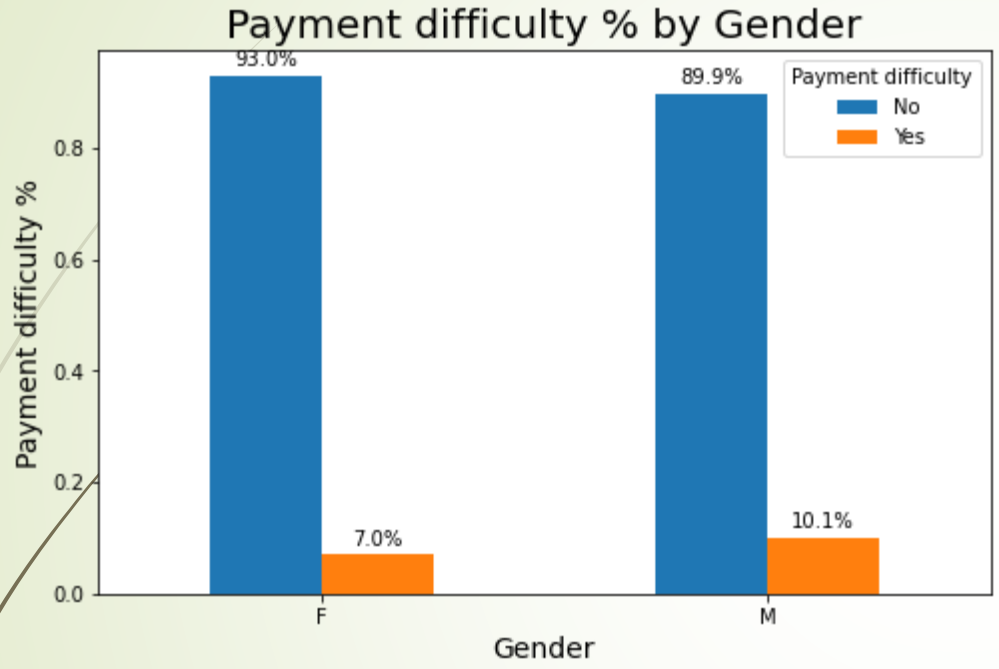


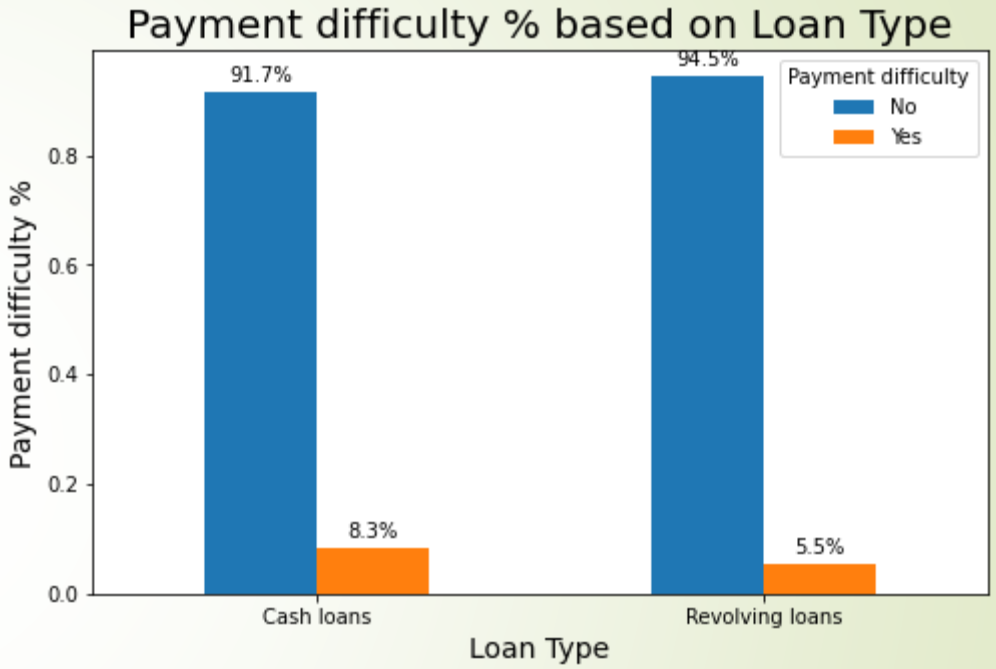
As the loan amount increases, the number of people returning to the bank also increases. Hence, bigger loans seems to be safer for banks

DATA BINNING & FORMATING

- Columns 'AMT_INCOME_TOTAL' and 'AMT_CREDIT', which are continuous variable. New categorical columns have been created 'AMT_INCOME_RANGE' and 'AMT_CREDIT_RANGE' using quantiles for better analysis
- Column 'DAYS_BIRTH' has been converted in years and then a new categorical column has been created using it for better analysis

UNIVARIATE ANALYSIS -1

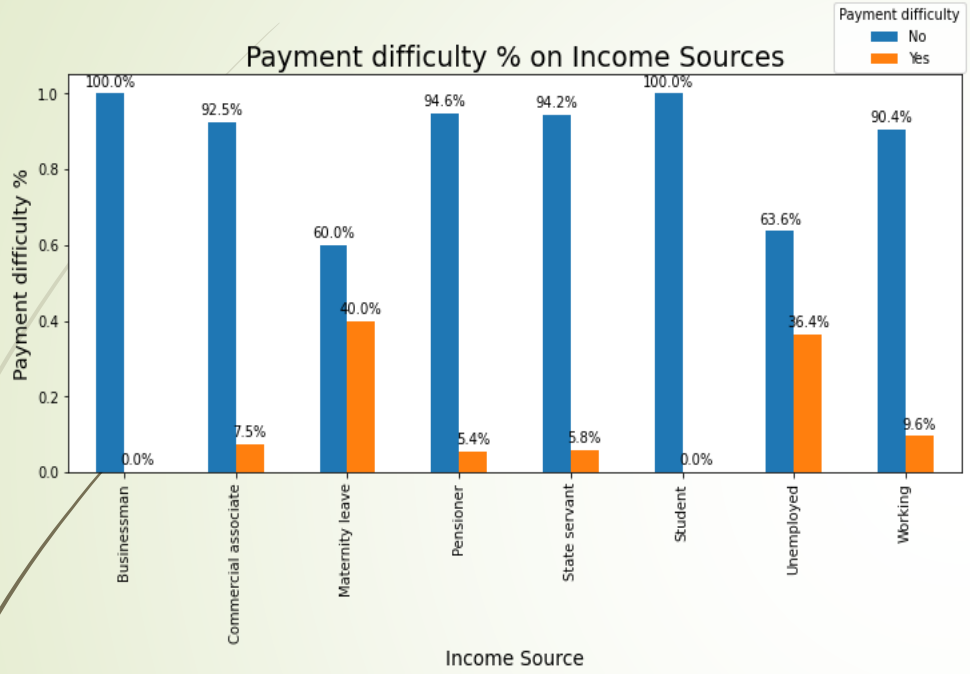




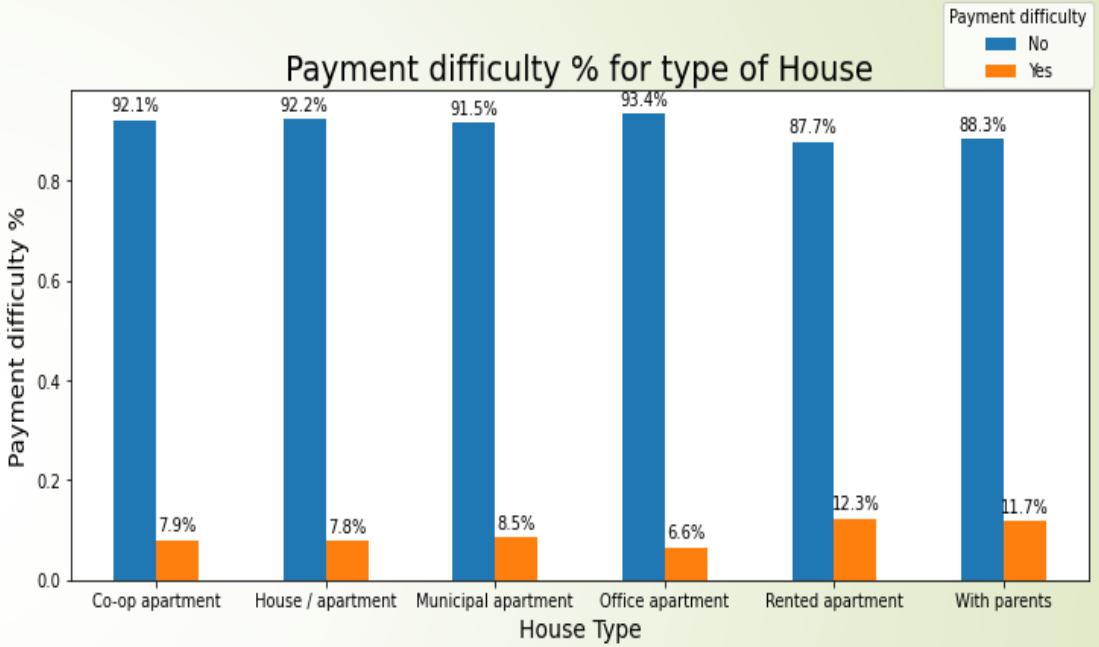
Comparing the Payment Difficulties and Non Payment Difficulties on the basis of Gender, we observe that Females are the majority in both the cases although there is an increase in the percentage in Male from Non-Payment Difficulties

We can observe that cash loans are preferred by both Loan Payment Difficulties and Loan-Non Payment Difficulties although there is a decrease in the percentage of Payment Difficulties who opt for revolving loans.

UNIVARIATE ANALYSIS -2

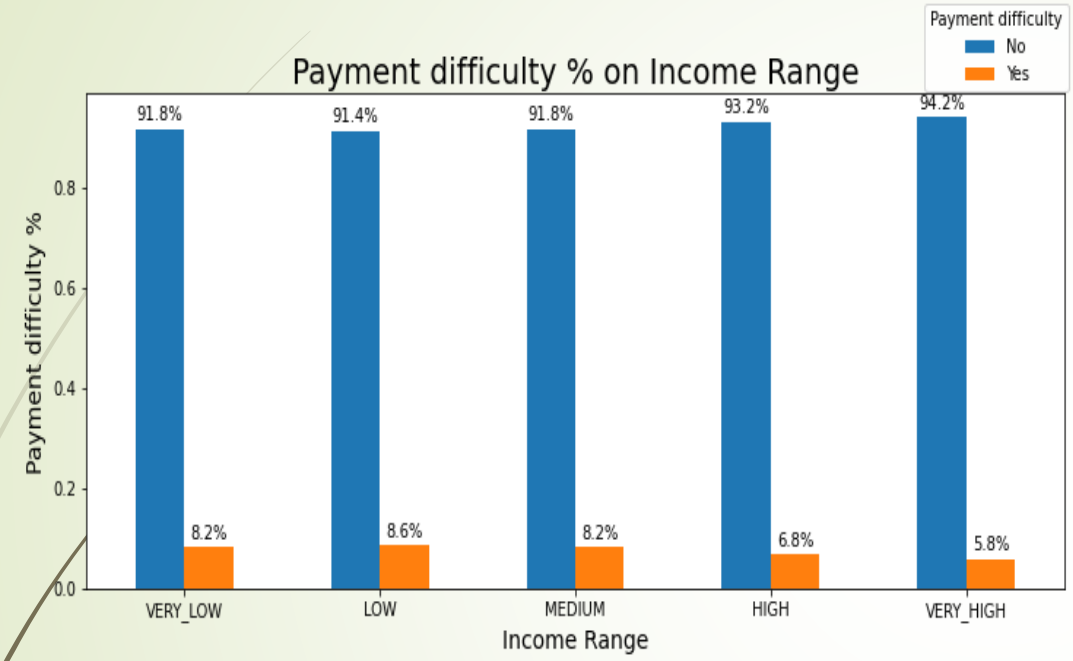


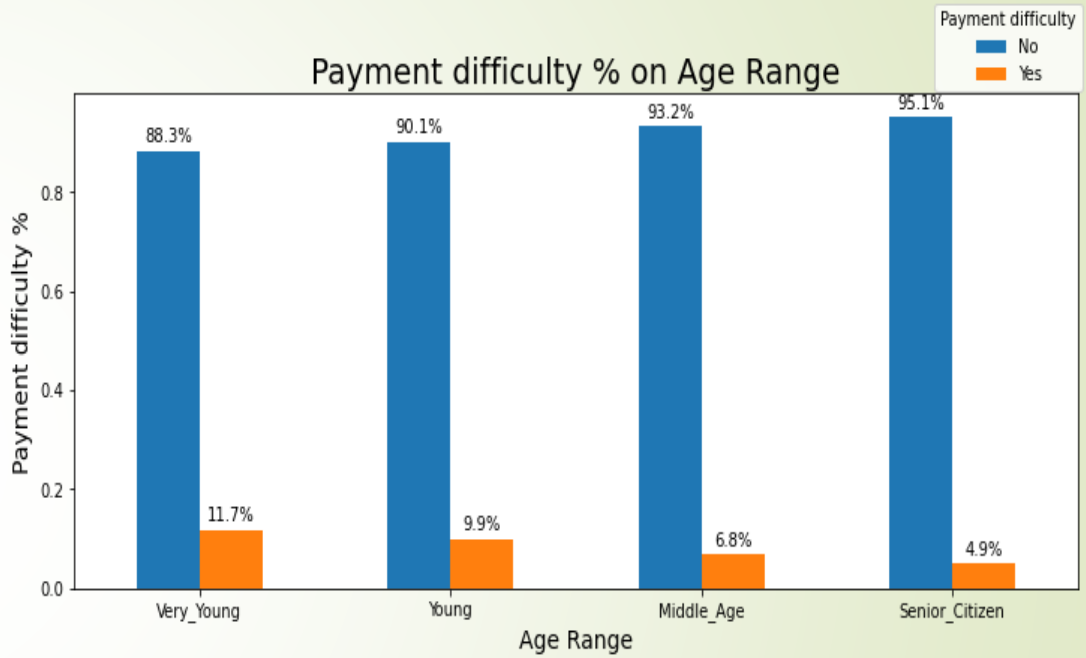
We observe that Businessman and Student have no payment difficulties where as Maternity leave and Unemployed have highest % of payment difficulties. Also, decrease in the percentage of Payment Difficulties for pensioner has been observed as compared to applicants who are working.



We observe an increased percentage of Loan Payment Difficulties who live in Rented apartment and with parents and a decreased percentage of Loan Payment Difficulties who live in Office apartment.

UNIVARIATE ANALYSIS -3

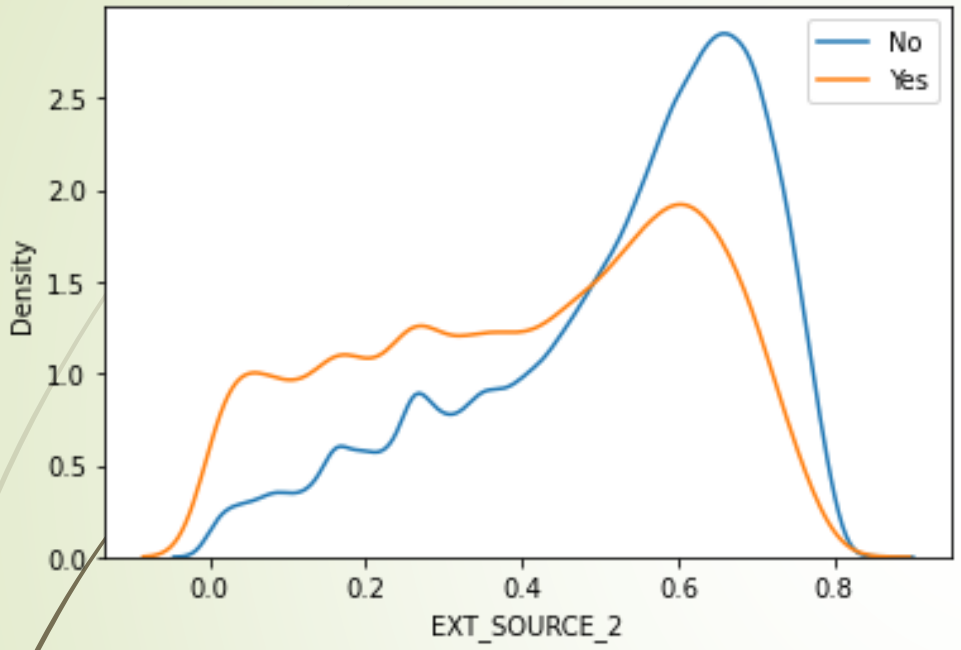




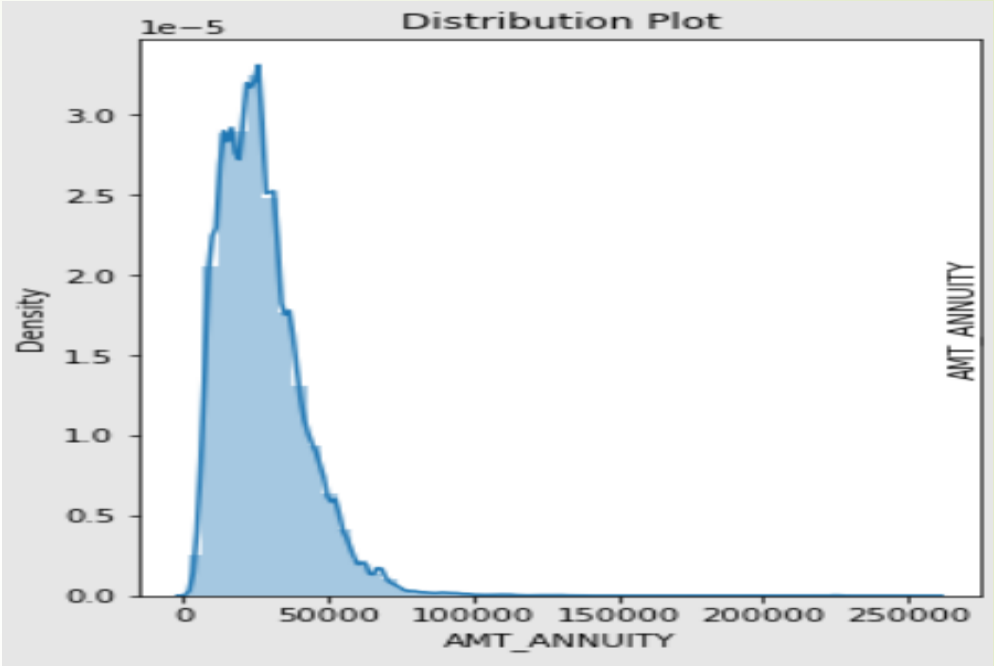
We observe an increased percentage of Loan Payment Difficulties whose income is very low and medium

We observe that there is an increase in the percentage of Loan Payment Difficulties who are young in age. In another word there is indirect correlation between age and payment difficulty %.

UNIVARIATE ANALYSIS -4

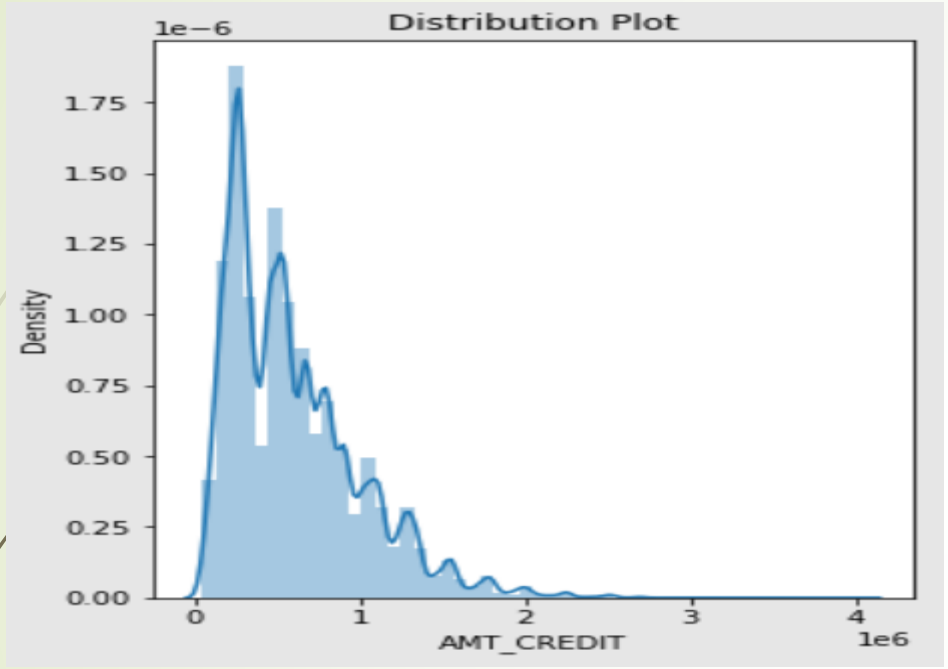


The better the external score of the applicant, the lesser is the payment difficulties.

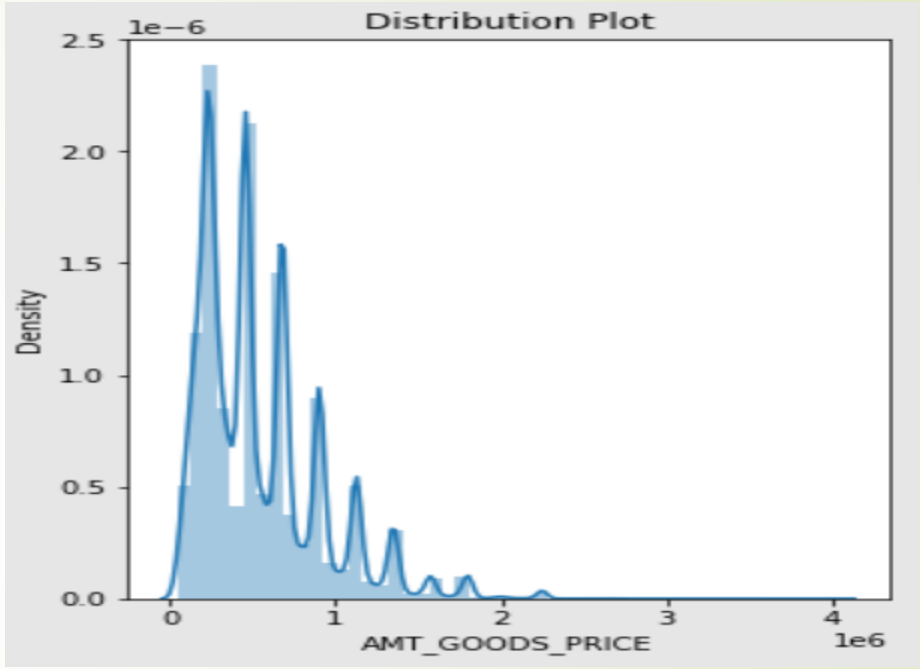


Most people pay annuity below 50K for the loans.

UNIVARIATE ANALYSIS -5



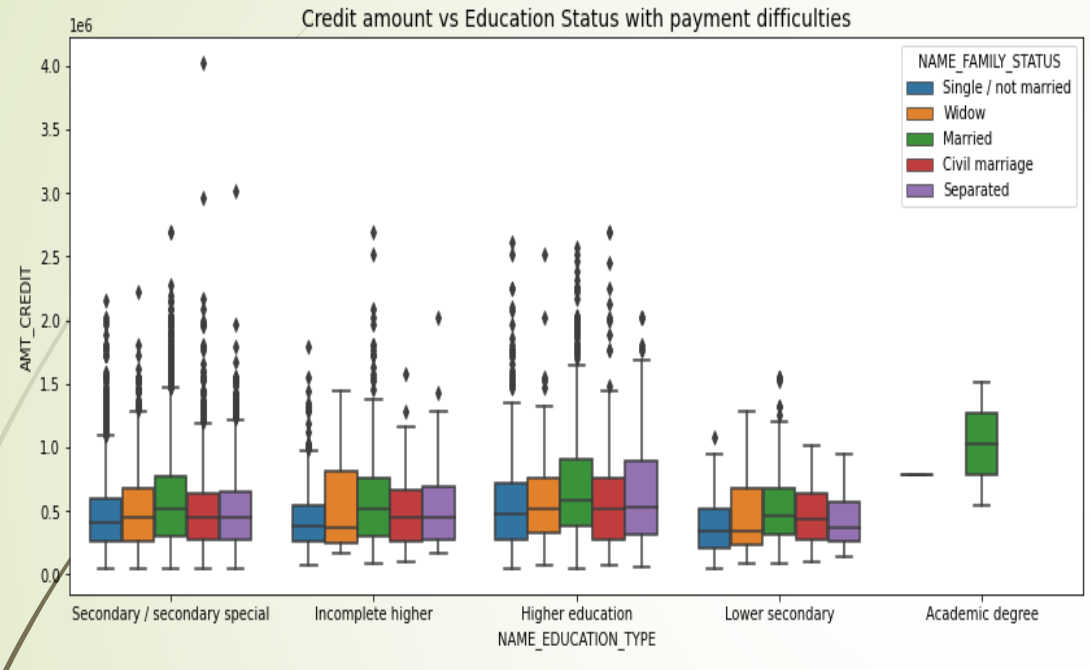
Credit amount of the loan is mostly less then 10 lakhs



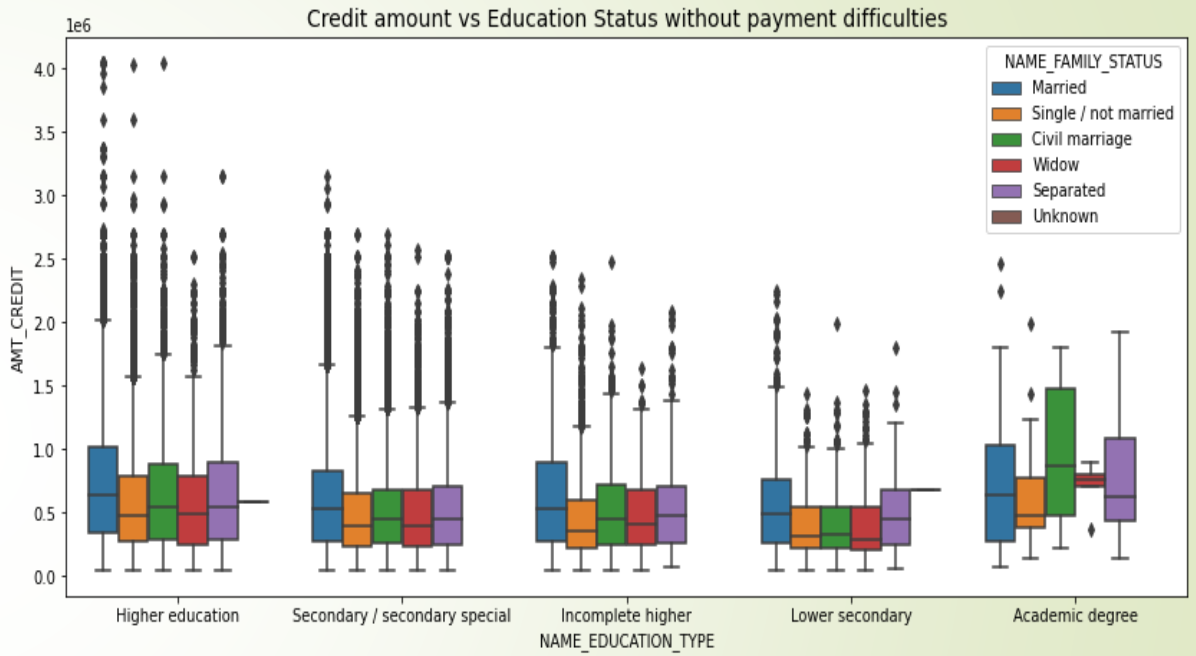
Most number of loans are given for goods price below 10 lakhs.

BIVARIATE ANALYSIS -1

Applicants without payment difficulties



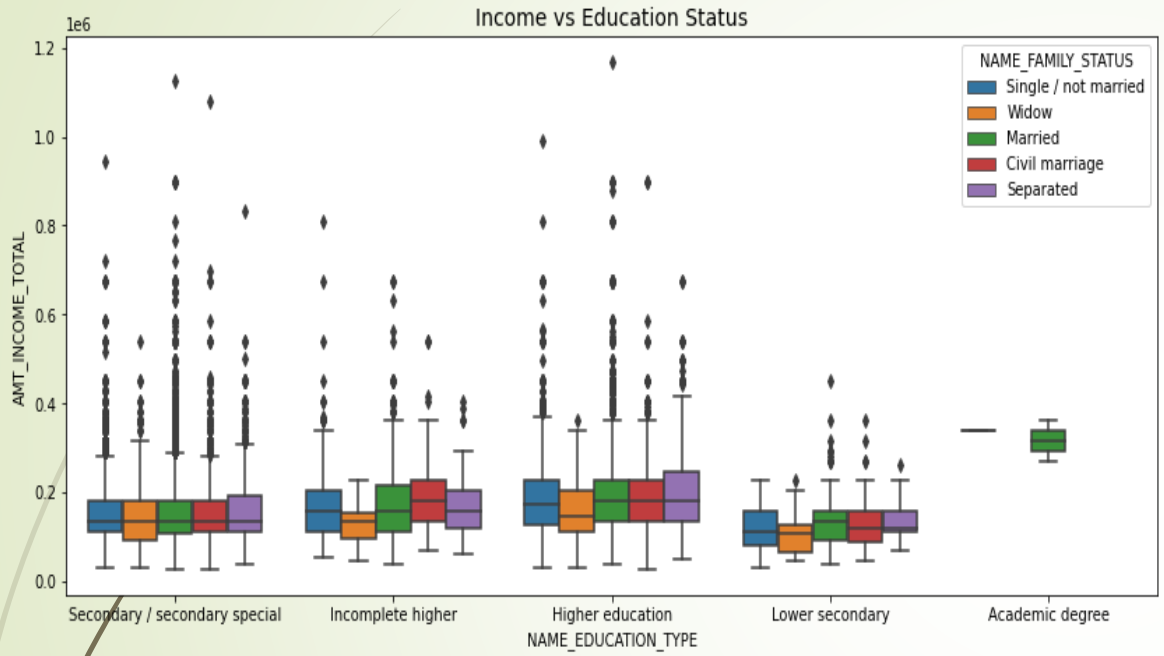
Applicants with payment difficulties



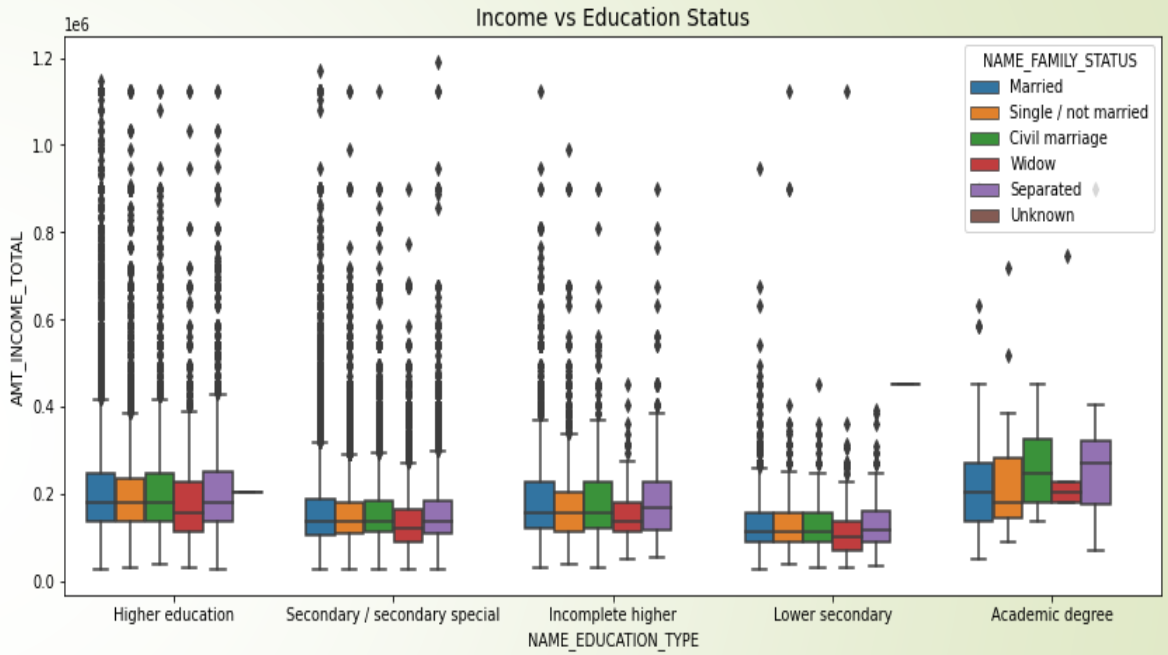
- Applicants with Academic degree have no outliers and only Married people face payment difficulties
- In Married category, people with higher loan amount are less likely to have payment difficulties

BIVARIATE ANALYSIS -2

Applicants without payment difficulties



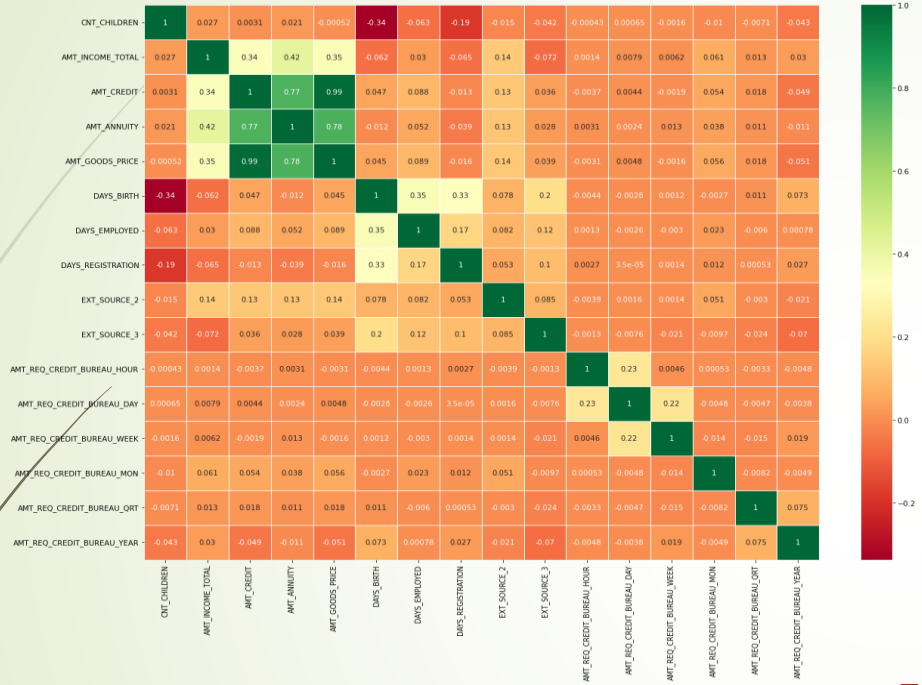
Applicants with payment difficulties



- Have some similarity with no payment difficulties, From above boxplot for Education type 'Higher education' the income amount is mostly equal with family status.
- Less outlier are having for Academic degree but there income amount is little higher that Higher education.
- Lower secondary are have less income amount than others.

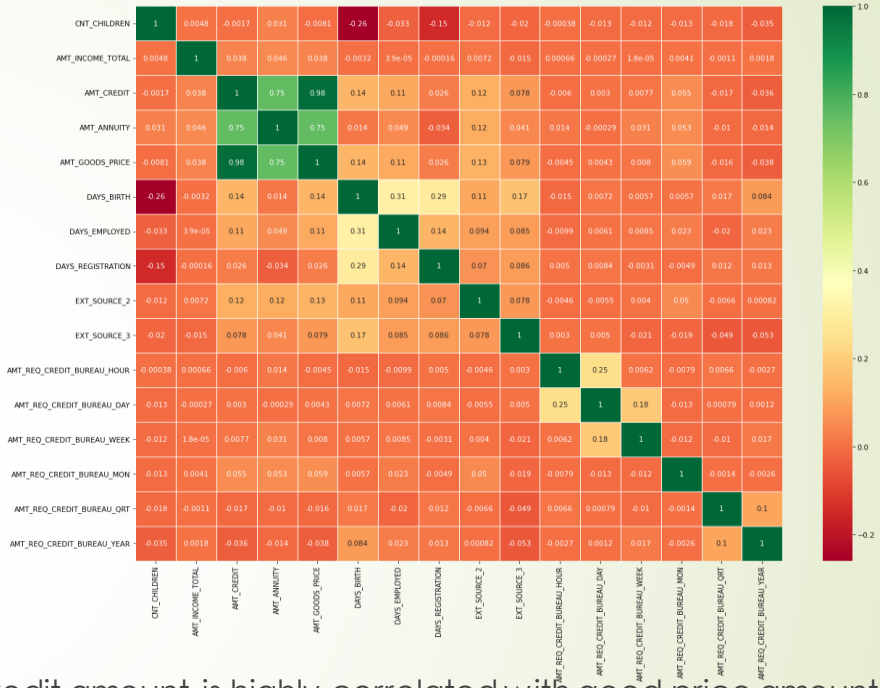
CORRELATION HEATMAP

Applicants without payment difficulties



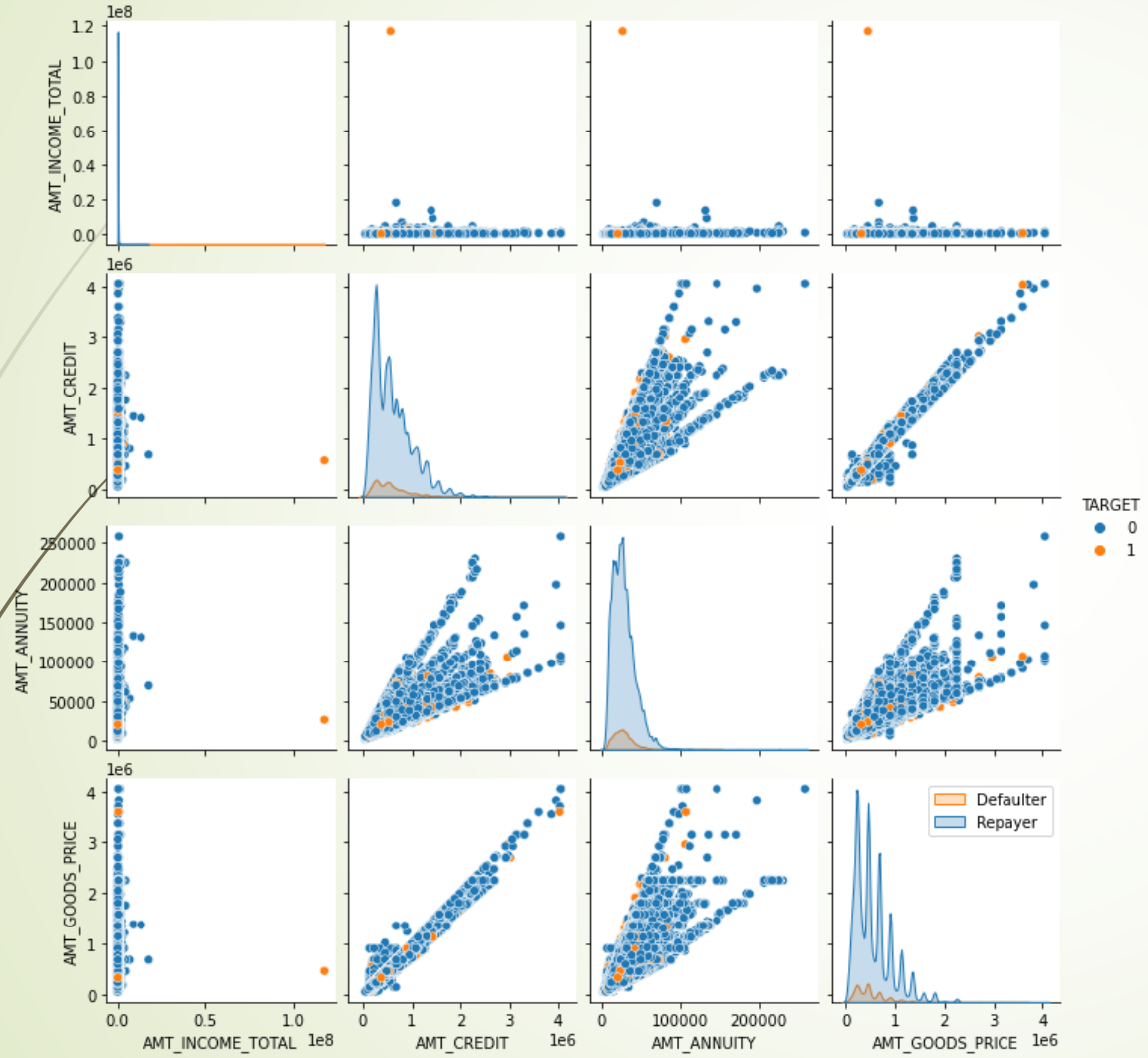
- Credit amount is highly correlated with:
  - Goods Price Amount
  - Loan Annuity
  - Total Income
- We can also see that re-payers have high correlation in number of days employed.

Applicants with payment difficulties



- Credit amount is highly correlated with good price amount which is same as re-payers.
- Loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to re-payers(0.77)
- We can also see that re-payers age have high correlation in number of days employed(0.35) when compared to defaulters(0.31).
- The correlation between total income of the client and the credit amount(0.33) amongst defaulters whereas it is (0.37) among re-payers.
- Days_birth and number of children correlation has reduced to 0.26 in defaulters when compared to 0.34 in re-payers.

PAIRPLOT AGAINST LOAN REPAYMENT STATUS



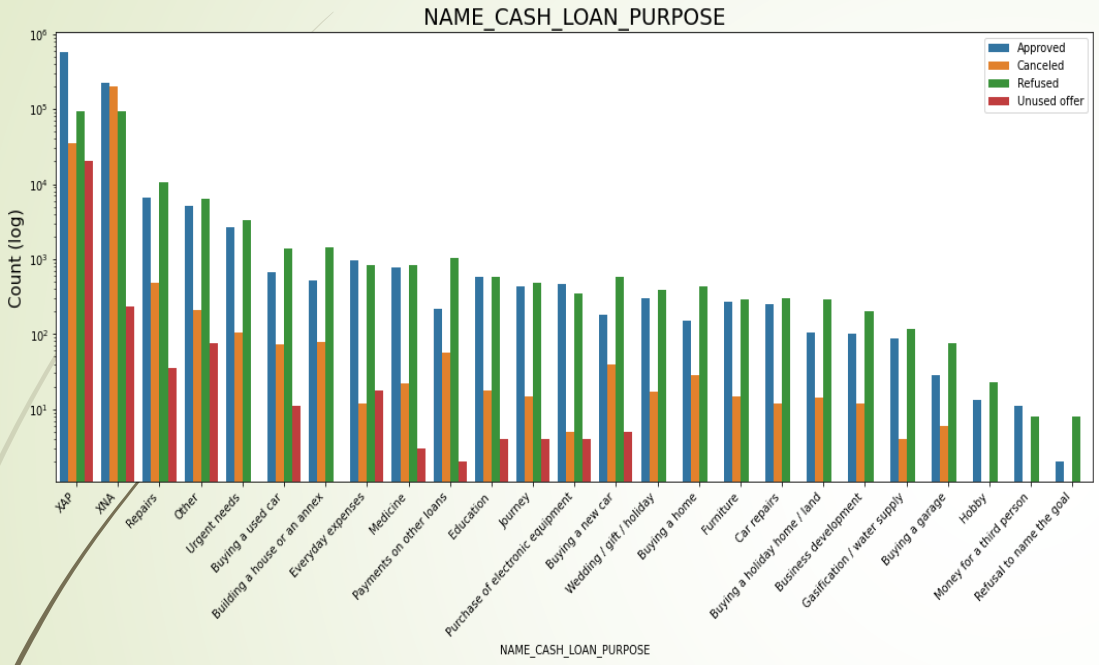
- When Annuity Amount > 15K and Good Price Amount > 20 Lakhs, there is a lesser chance of defaulters
- Loan Amount(AMT_CREDIT) and Goods price(AMT_GOODS_PRICE) are highly correlated as based on the scatterplot where most of the data are consolidated in form of a line
- There are very less defaulters for AMT_CREDIT >20 Lakhs

MERGED DATAFRAME ANALYSIS

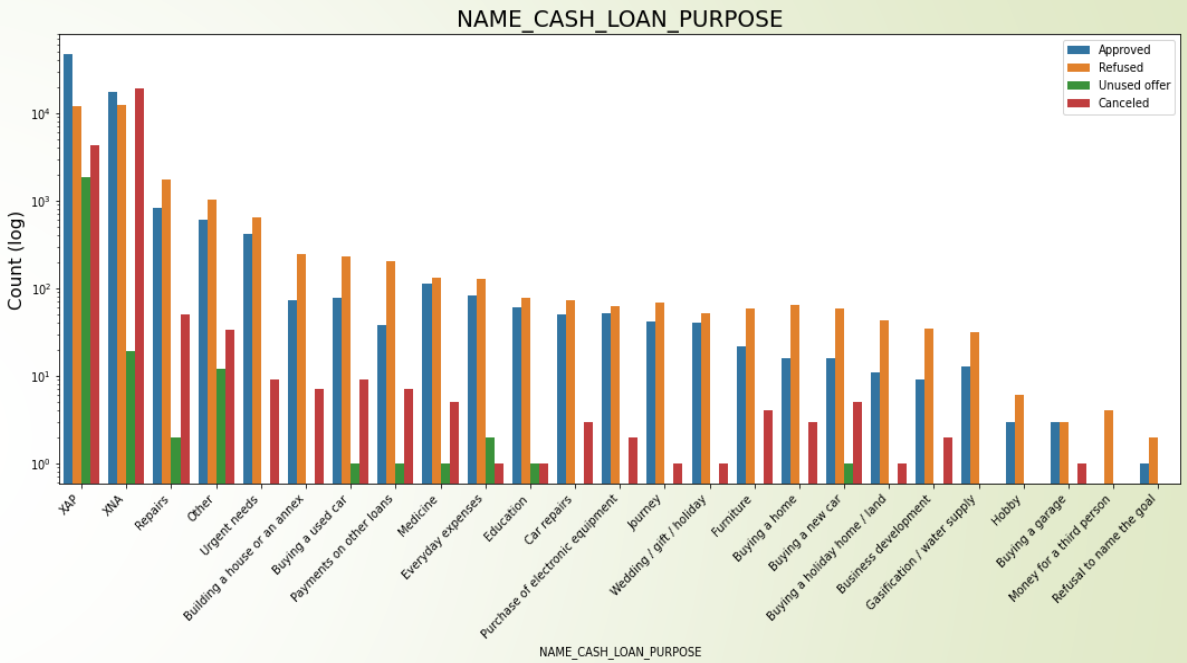
Merged both dataframes on SK_ID_CURR with Inner Joins

PLOTING CONTRACT STATUS vs PURPOSE OF LOAN

Applicants without payment difficulties

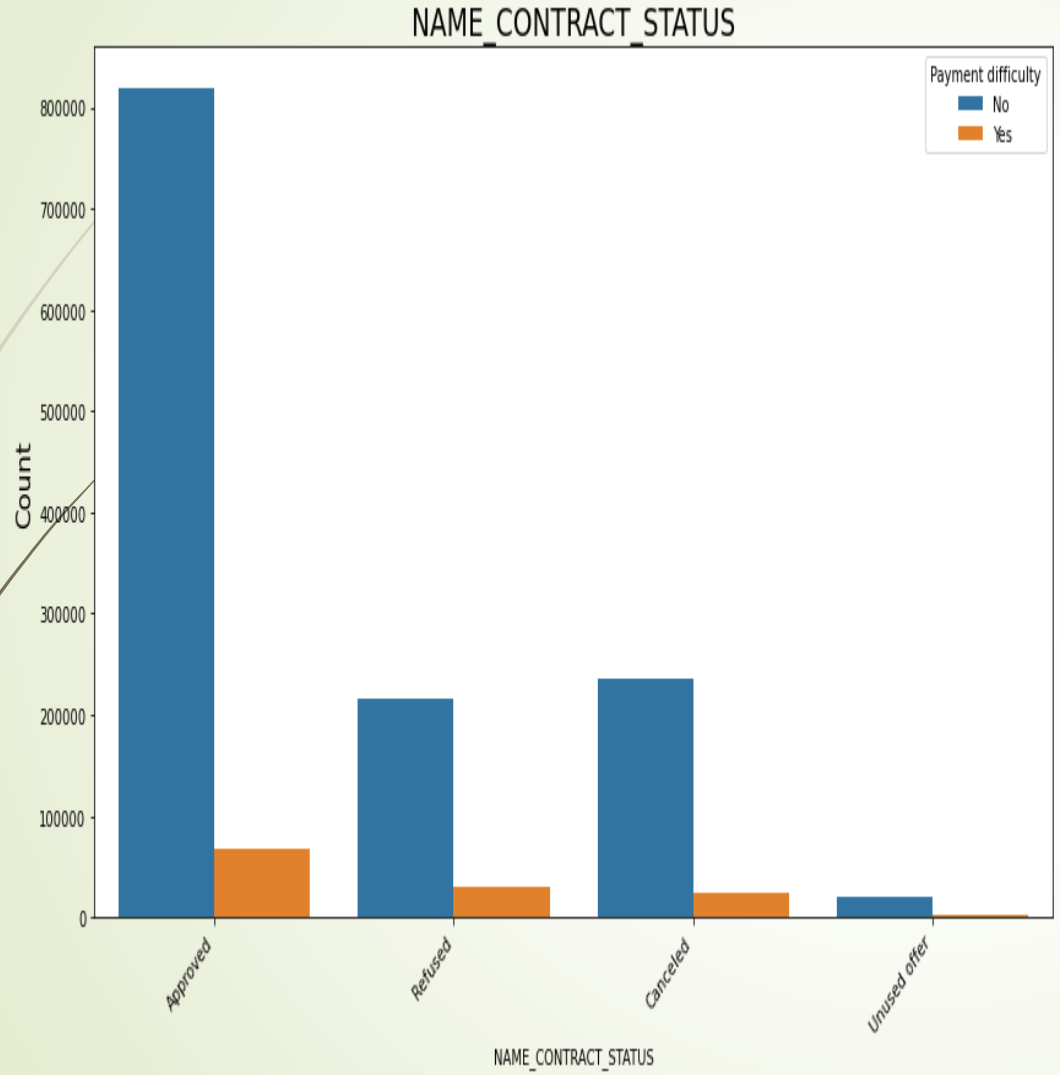


Applicants with payment difficulties



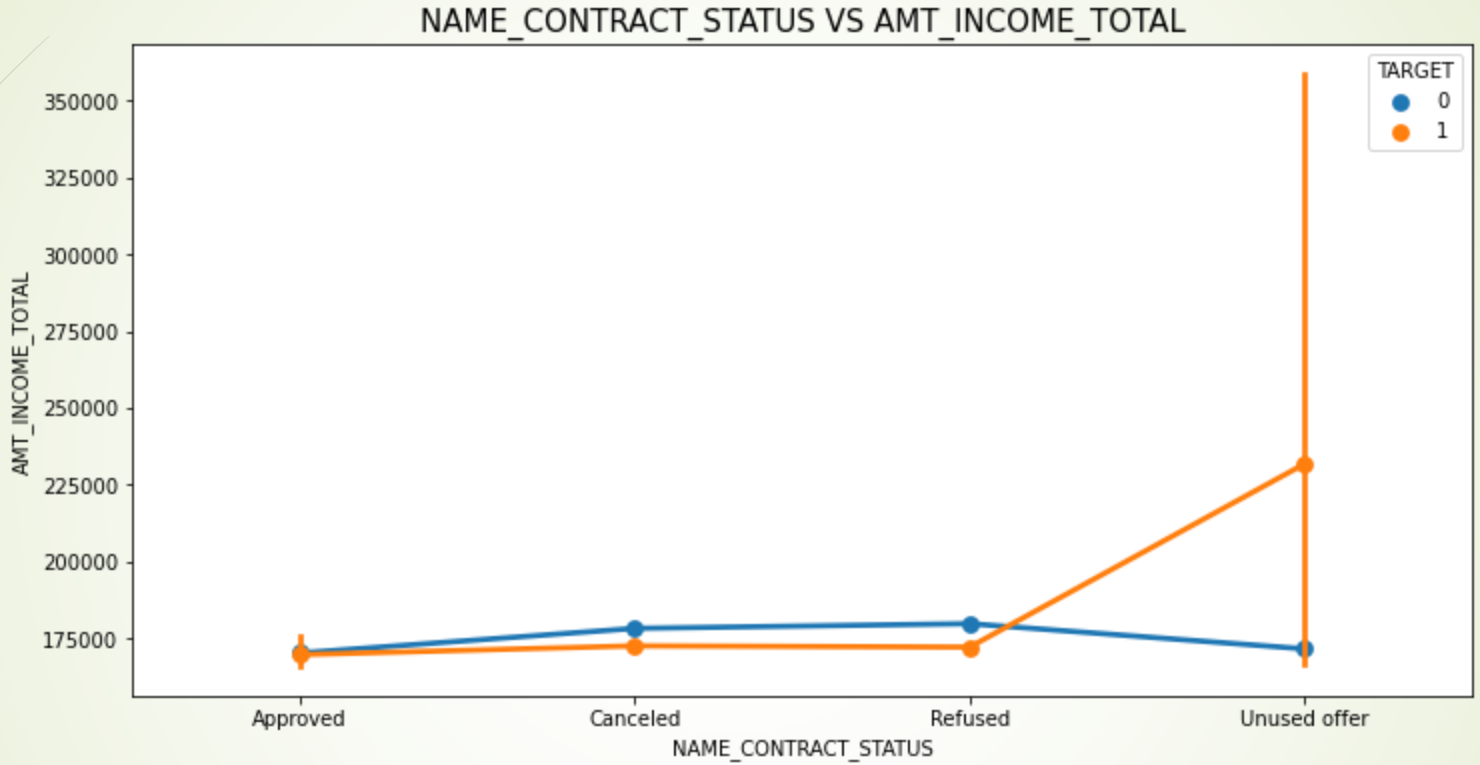
- Loan purpose has high number of unknown values (XAP, XNA)
- Loan taken for the purpose of Repairs looks to have highest default rate
- Huge number application have been rejected by bank or refused by client which are applied for Repair or Other. from this we can infer that repair is considered high risk by bank. Also, either they are rejected or bank offers loan on high interest rate which is not feasible by the clients and they refuse the loan.

CONTRACT STATUS BASED ON LOAN REPAYMENT STATUS



- 90% of the previously cancelled client have actually repayed the loan. Revising the interest rates would increase business opportunity for these clients
- 88% of the clients who have been previously refused a loan has payed back the loan in current case.
- Refusal reason should be recorded for further analysis as these clients could turn into potential repaying customer.

CONTRACT STATUS BASED ON INCOME



- The point plot show that the people who have not used offer earlier have defaulted even when there average income is higher than others

CONCLUSIONS -1

After analyzing the datasets, there are few attributes of a client with which the bank would be able to identify if they will repay the loan or not. The analysis is consisted as below with the contributing factors and categorization:

- **A. Decisive Factor whether an applicant will be Repayer:**
  - NAME_EDUCATION_TYPE: Academic degree has less defaults.
  - NAME_INCOME_TYPE: Student and Businessmen have no defaults.
  - ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%
  - DAYS_BIRTH: People above age of 50 have low probability of defaulting
  - DAYS_EMPLOYED: Clients with 40+ year experience having less than 1% default rate
  - AMT_INCOME_TOTAL:Applicant with Income more than 700,000 are less likely to default
- **B. Decisive Factor whether an applicant will be Defaulter:**
  - 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.
  - OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as their default rate is huge.
  - ORGANIZATION_TYPE: Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed people have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.
  - DAYS_BIRTH: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting
  - DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.
  - AMT_GOODS_PRICE: When the credit amount goes beyond 3lakhs, there is an increase in defaulters.

CONCLUSIONS -2

- **C. Factors that Loan can be given on Condition of High Interest rate to mitigate any default risk leading to business loss:**
  - NAME_HOUSING_TYPE: High number of loan applications are from the category of people who live in Rented apartments & living with parents and hence offering the loan would mitigate the loss if any of those default.
  - AMT_CREDIT: People who get loan for 3-6 Lakhs tend to default more than others and hence having higher interest specifically for this credit range would be ideal.
  - AMT_INCOME: Since 90% of the applications have Income total less than 3Lakhs and they have high probability of defaulting, they could be offered loan with higher interest compared to other income category.
- **D. Suggestions:**
  - 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 mitigate the business loss and these clients could be contacted for further loans.


THANK
YOU!
ctrl
alt