EXPLORATORY DATA ANALYSIS OF LOAN DEFAULTERS CASE STUDY

PREPARED BY

Manika Goyal

PROBLEM STATEMENT

Introduction-

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, 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:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
2. 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.

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.

STRATEGY FOR DATA CLEANING :

- Null value Calculation(percentage of missing values in ‘application_data’ and ‘previous_application’
- Analyze & Delete Unnecessary Columns. (Drop the columns which have more than 40% null values also drop columns which are not important to business perspective)
- Null Value Data Imputation.
- Identifying the outliers.

METHOD TO IMPUTE THE MISSING VALUES

- To impute null values in categorical variables which has lower null percentage, mode() is used to impute the most frequent items.

- To impute null values in categorical variables which has higher null percentage, a new category is created.

- To impute null values in numerical variables which has lower null percentage, median() is used

- To impute null values in continuous variables, we plotted the distribution of the columns and used :
    median if the distribution is skewed.
    mode if the distribution pattern is preserved

IMBALANCE ANALYSIS OF THE TARGETVARIABLE

Insights:

According to TARGET column, 91% people faced no difficulty whereas 8% people faced difficulty. i.e Repayor is greater than defaulter and in terms of domain its relevant imbalance

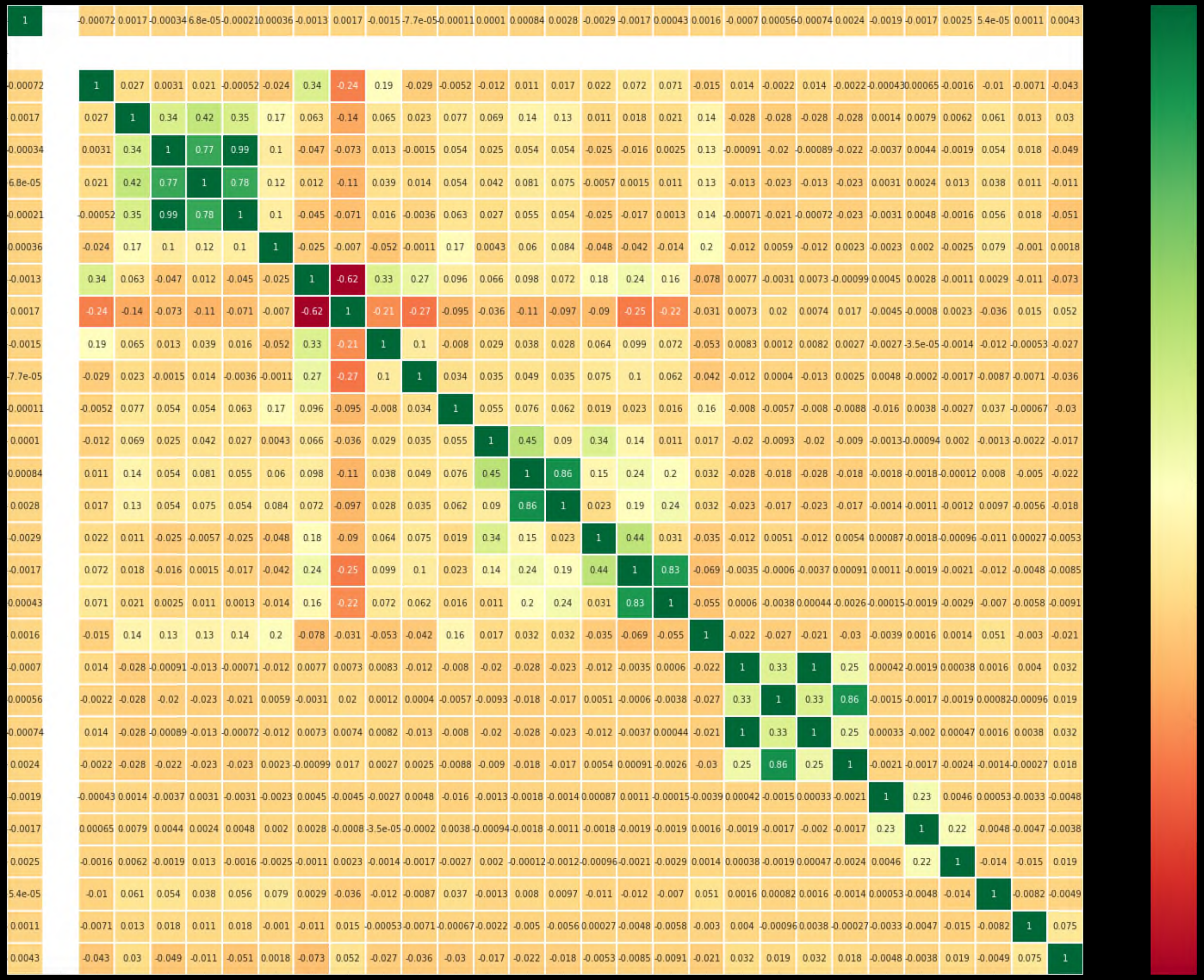
CORRELATION BETWEEN NUMERIC VARIABLE

Top 10 correlation of Repayors :

- We can see that the repayors 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. REGION_RATING_CLIENT_W_CITY &
- REGION_RATING_CLIENT are correlated. DAYS_EMPLOYED DAYS_BIRTH
- ,FLAG_EMP_PHONE DAYS_BIRTH gives the relation of the age group . FLAG_EMP_PHONE & DAYS_EMPLOYED high correlation which we can get information whether the client is still employed.

	VAR1	VAR2	Correlation
680	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508
184	AMT_GOODS_PRICE	AMT_CREDIT	0.987250
464	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
711	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859332
557	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830381
185	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
154	AMT_ANNUITY	AMT_CREDIT	0.771309
278	DAYS_EMPLOYED	DAYS_BIRTH	0.618048
433	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.446101
526	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.435514

CORRELATION BETWEEN NUMERIC VARIABLE



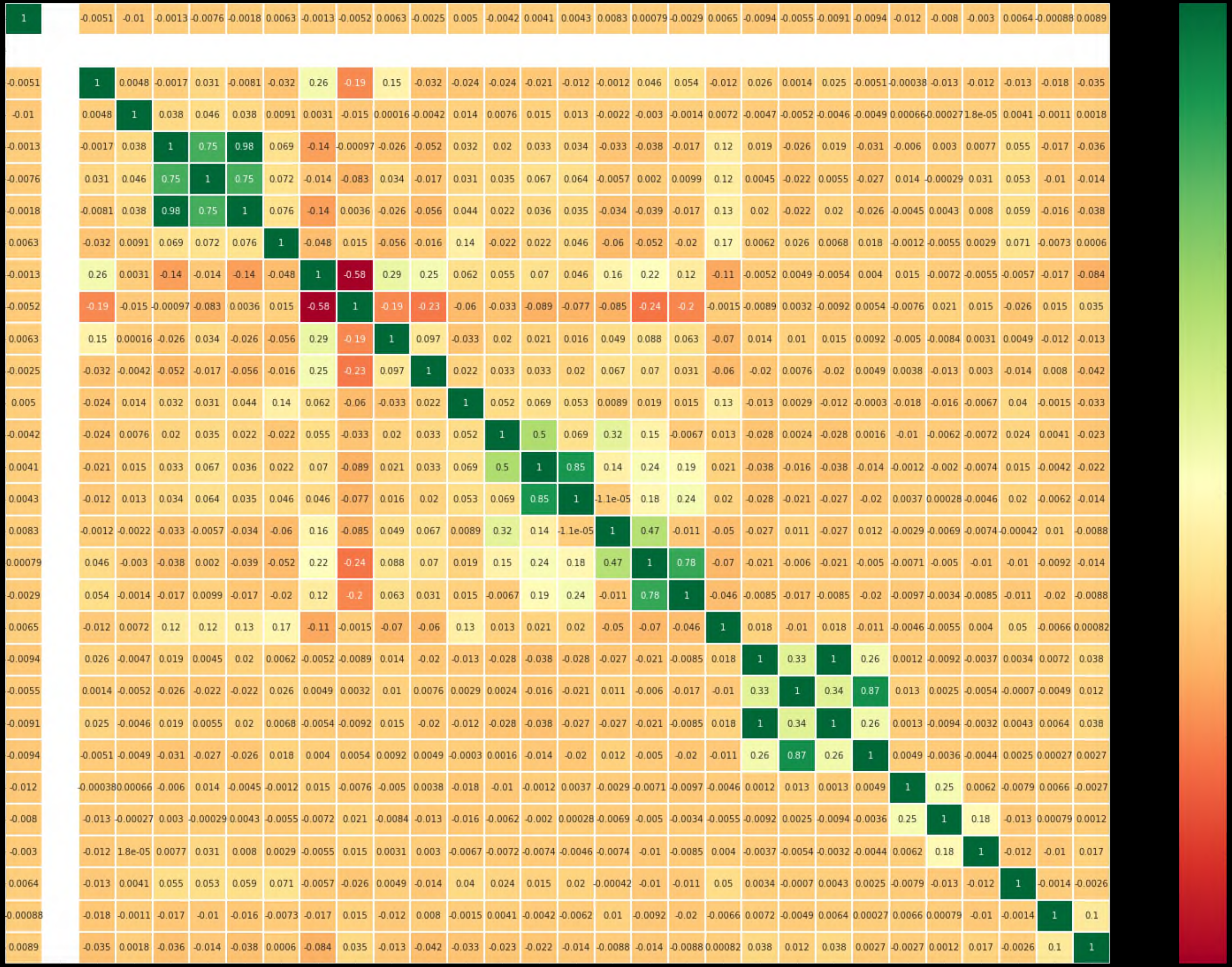
CORRELATION BETWEEN NUMERIC VARIABLE

Top 10 correlation of defaulters:

- Insights :
- 1. FLAG_EMP_PHONE & DAYS_EMPLOYED are highly corelated as it gives the information of the
- 2. Credit amount is highly correlated with good price amount.
- 3. Loan annuity correlation with credit amount has slightly reduced from (.77 to .75)in defaulters when compared to repayers
- 4. Age of the client and the no. of days employed give the insight about the working profile.(such as student or experienced employee)

	VAR1	VAR2	Correlation
680	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269
184	AMT_GOODS_PRICE	AMT_CREDIT	0.983103
711	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994
464	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885
557	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540
185	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
154	AMT_ANNUITY	AMT_CREDIT	0.752195
278	DAYS_EMPLOYED	DAYS_BIRTH	0.575097
433	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.497937
526	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.472052

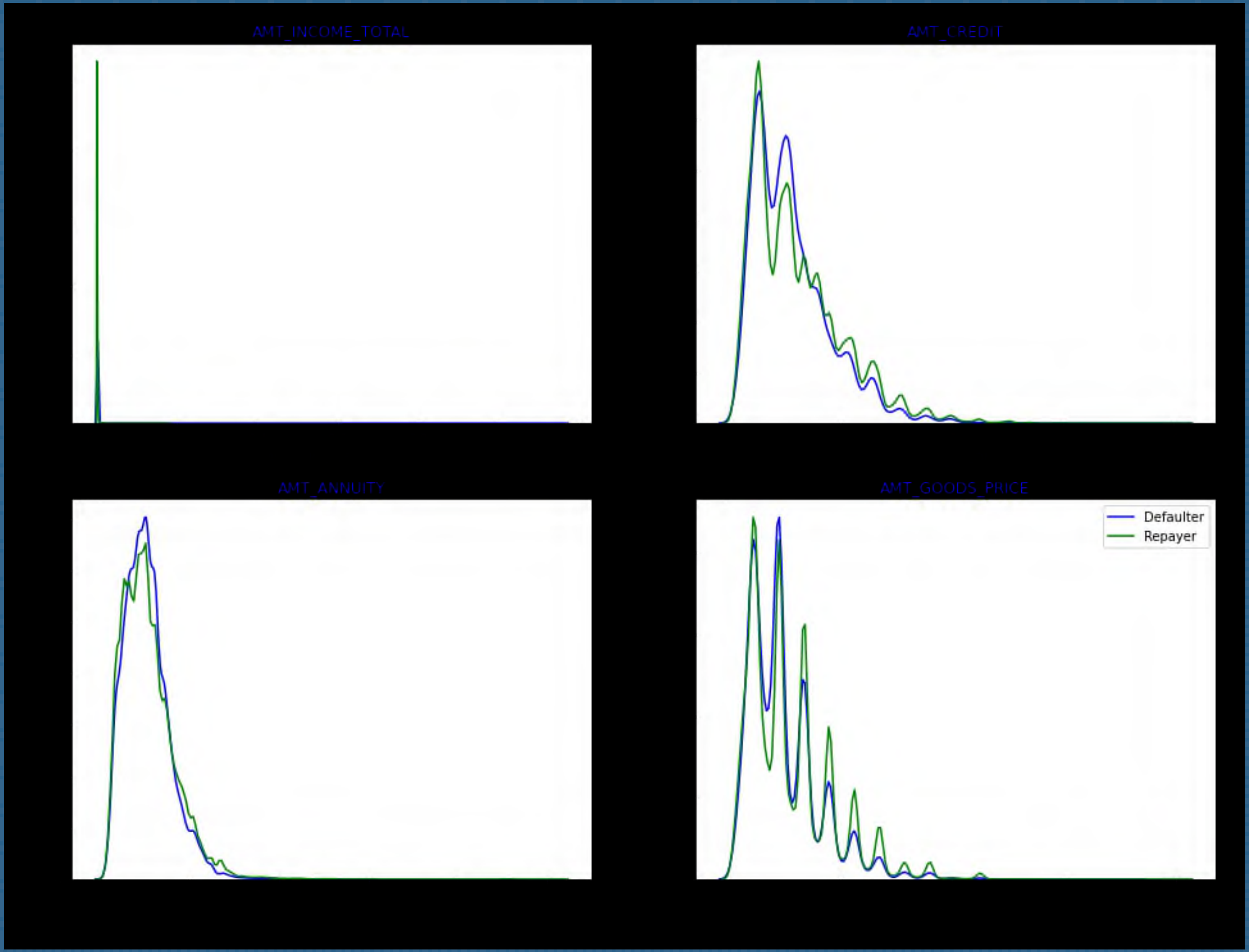
CORRELATION BETWEEN NUMERIC VARIABLE



UNIVARIATE ANALYSIS OF NUMERICAL VARIABLE

Insights:

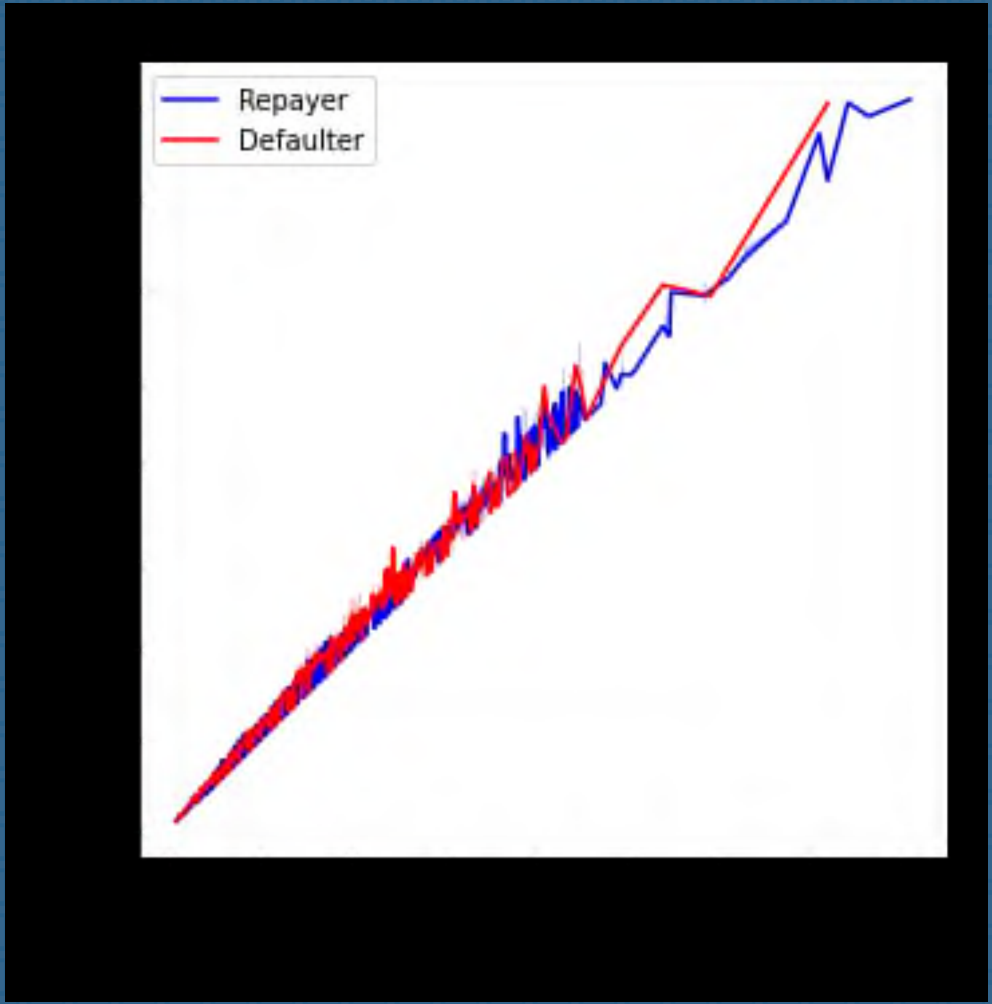
- In AMT_CREDIT,most of the loans given under 10 lakhs.
- Most of the people pay Annuity below 50000.
- Credit amount for goods was mostly under 10 lakhs.
- AMT_INCOME_TOTAL got overlapped so we will not use this further.



BIVARIATE ANALYSIS OF NUMERICAL VARIABLES

Insights:

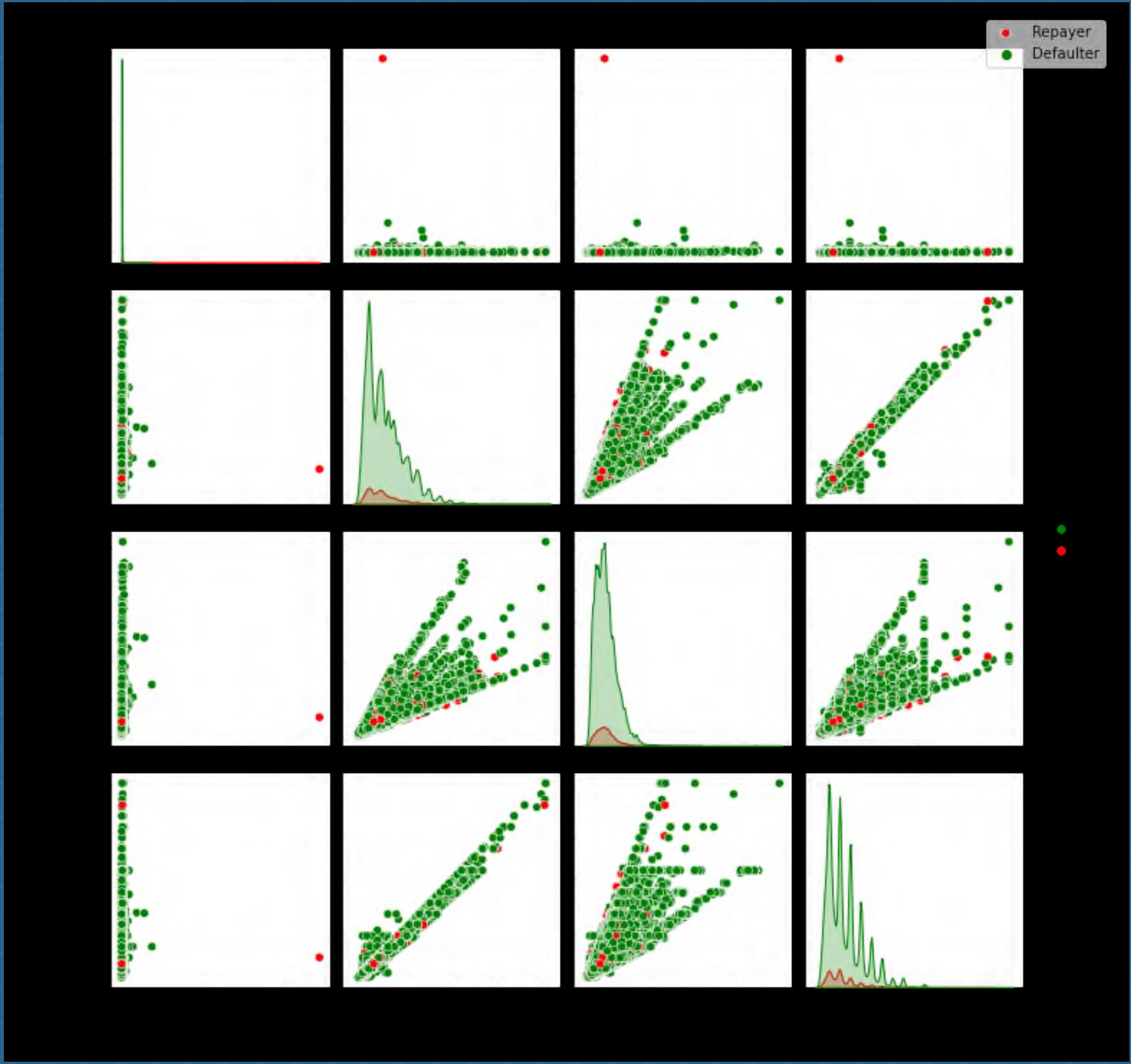
When the credit_amount goes up above 3M than there is increase in defaulters.



BIVARIATE ANALYSIS OF NUMERICAL VARIABLES

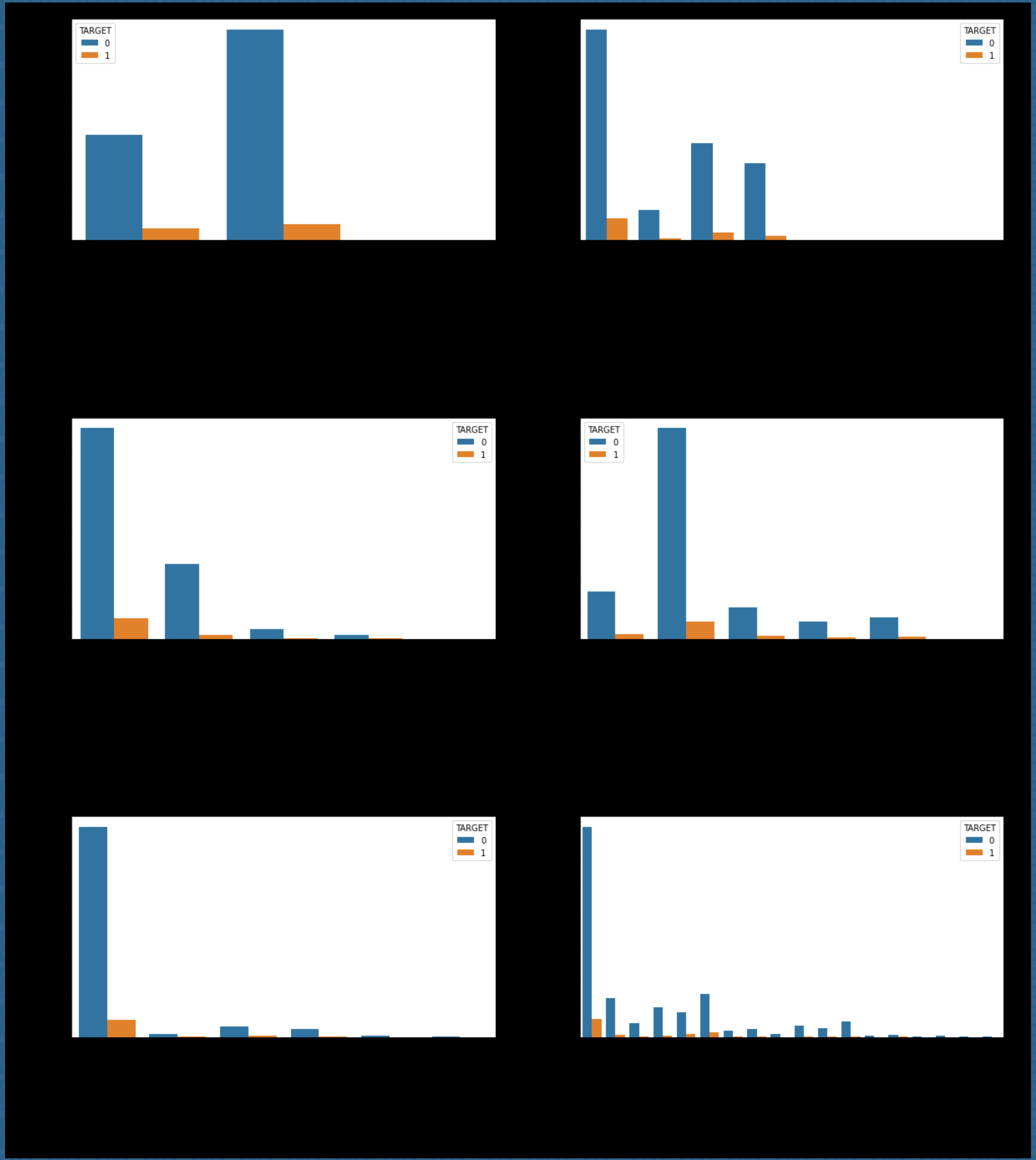
Insights:

- When amt_annuity >15000 amt_goods_price> 3M, there is a lesser chance of defaulters.
- AMT_CREDIT and AMT_GOODS_PRICE are highly correlated as based on the scatterplot where most of the data are consolidated in form of a link.
- There are very less defaulters for AMT_CREDIT >3M.
- Inferences related to distribution plot has been already mentioned in previous distplot graphs inferences section



UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLES

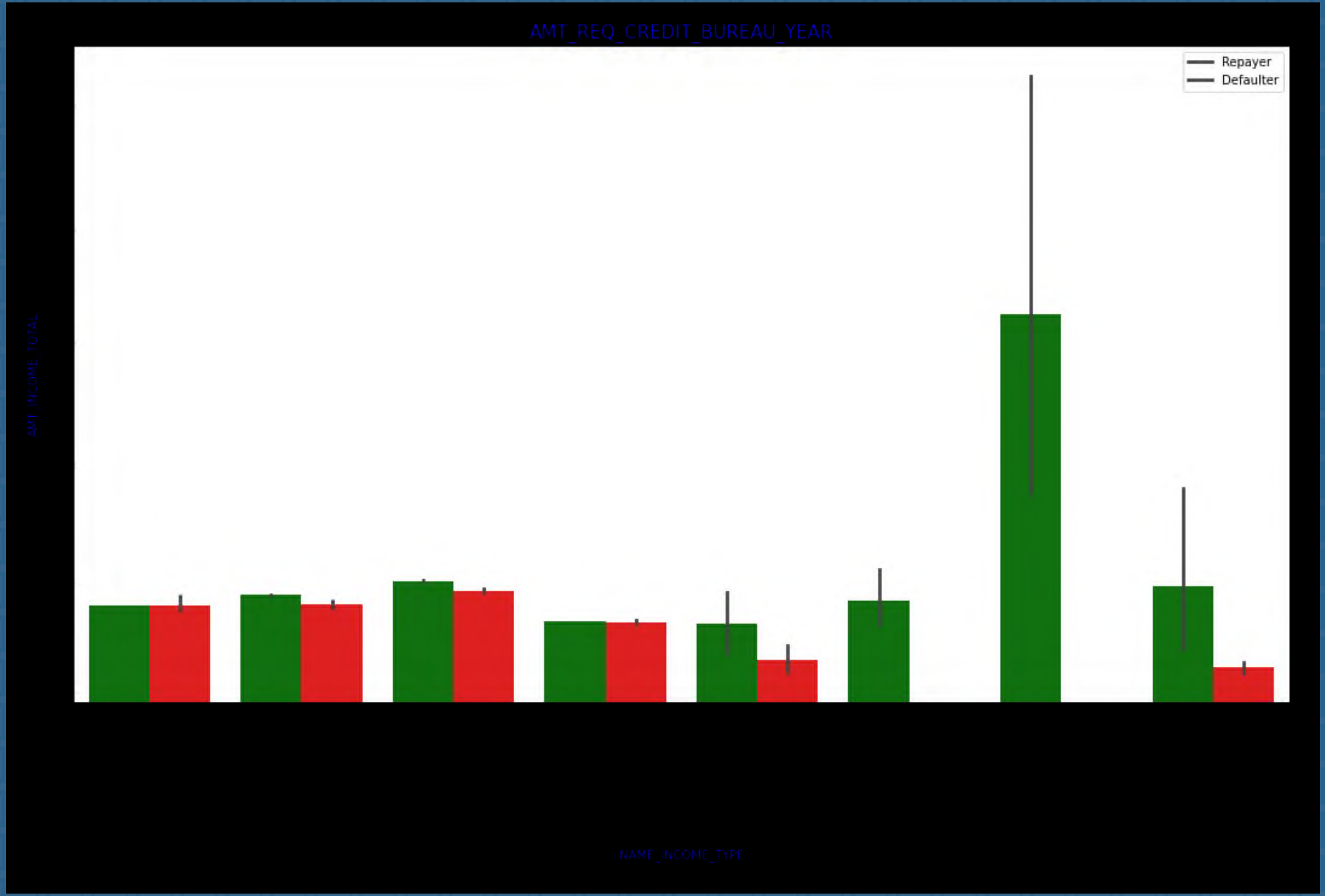
- There are higher chances in male for not returning the loan.
- Students and businessman categories do not have any default record so these 2 categories are safe to provide loan.
- Working clients can be targeted to provide loans and there are higher chances of replaying the loans in them.
- Clients with Secondary education are more repayable.
- Academic degree have less default rate and lower secondary have more chances of default.
- Most of the loans are taken by married people and most of the defaulters are belong to Civil marriage people.
- People living in office apartment and co-ap apartment have lowest default rate and people with rented apartment have higher probability of default.
- Category with highest percentage of not returning the loans are low skilled laborers followed by drivers,waiters/barmen, security staff, laborers and cooking staff.



BI/MULTIVARIATE ANALYSIS OF CATEGORICAL VARIABLES

Insights :

It can be seen It can be seen that business man's income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a business man could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs

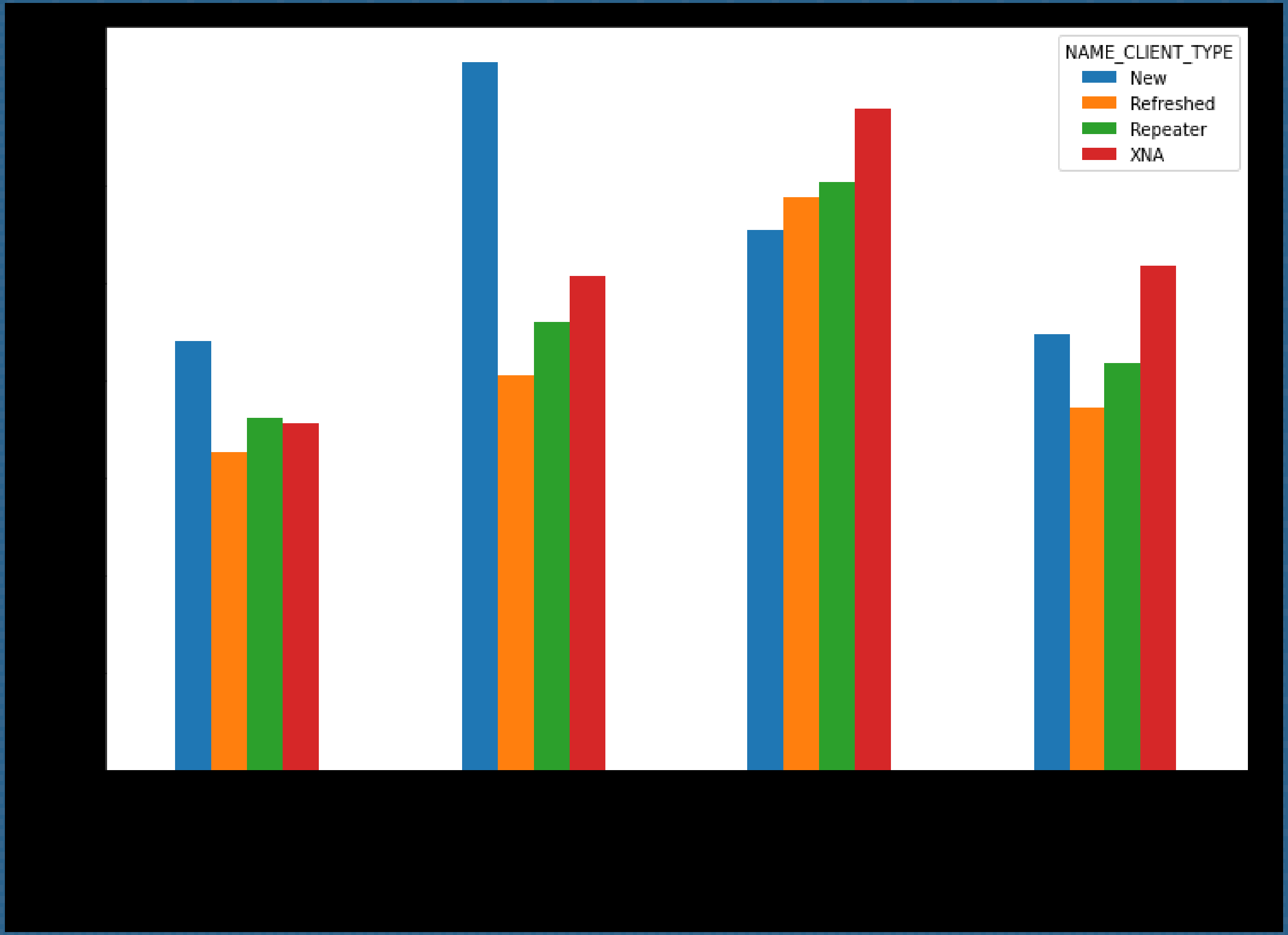


Merged Data frames Analysis

DATA FRAME OF CURRENT DEFAULTERS WITH PREVIOUS LOAN STATUS AND THE NAME_CLIENT_TYPE

Insights :

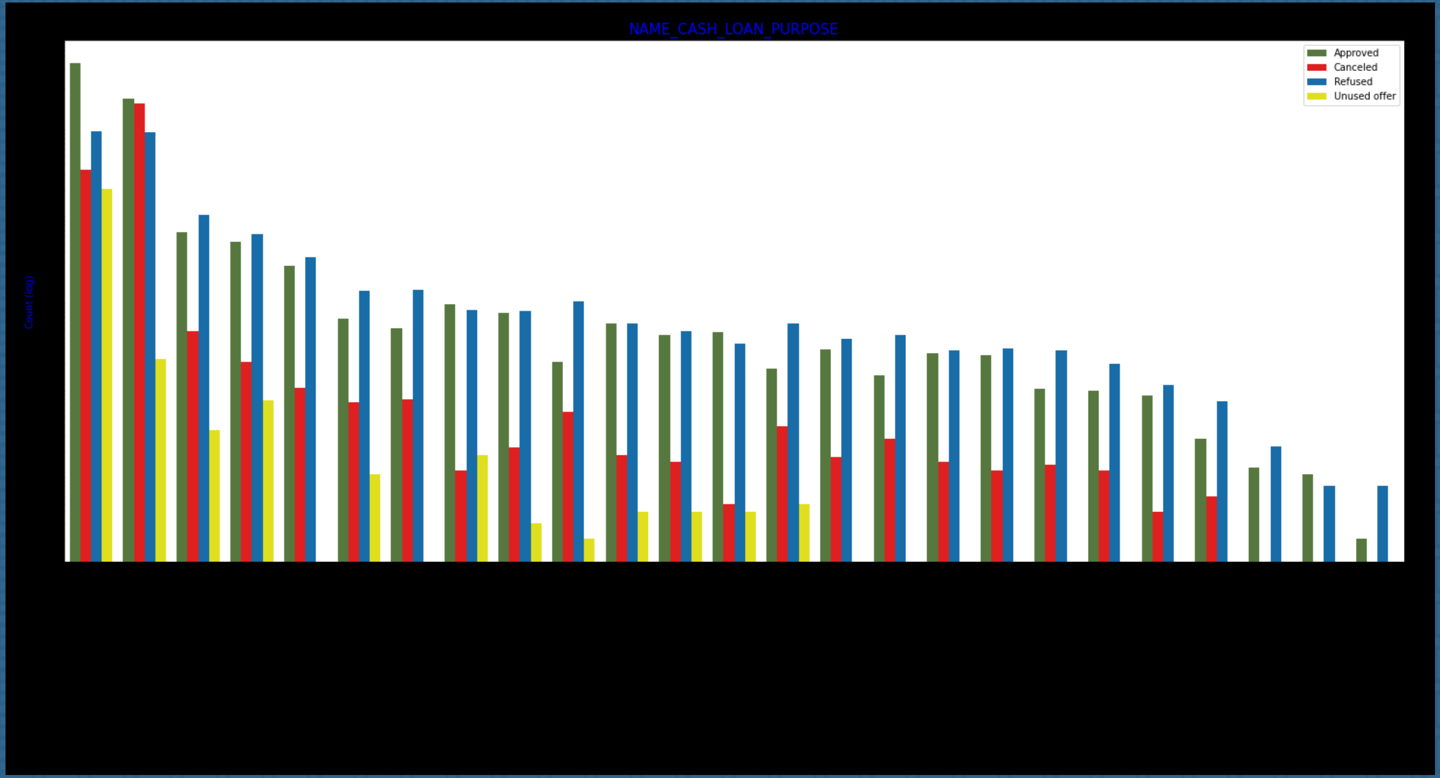
- The previously Approved status the New clients were more defaulted followed by Repeater.
- For previously Canceled applicants the Defaulters are more New clients.
- We can see that the Defaulters are more for previously Unused offers loan status clients, who were found to be new.
- Previously Refused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



REPRESENTATION OF UNIVARIATE MERGED OF NAME_CASH_LOAN_PURPOSE AND NAME_CONTRACT_STATUS

Insights :

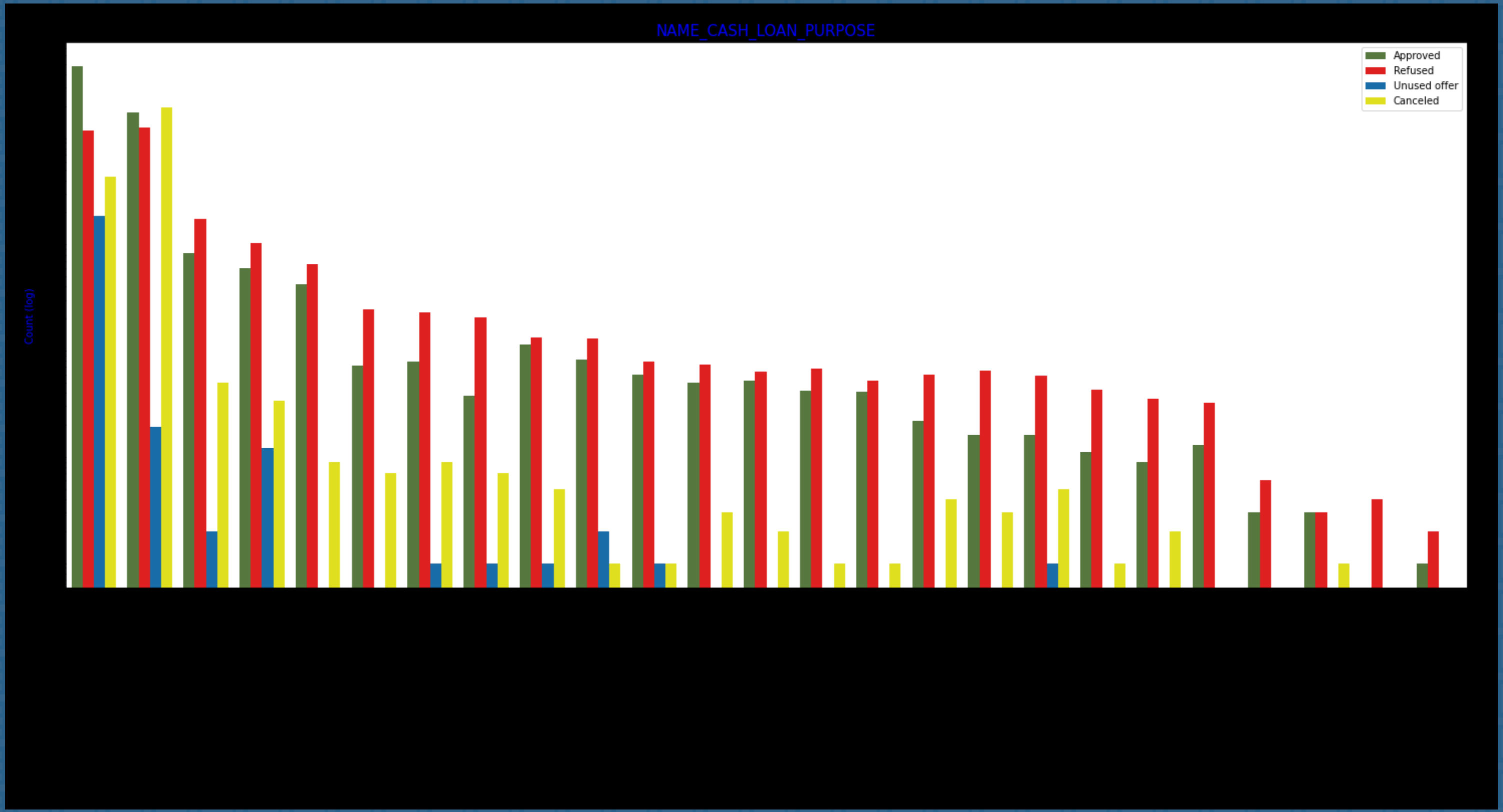
- A very high number application have been rejected by bank or refused by client which has purpose as repair or other. This shows that purpose repair is taken as high risk by bank and either they are rejected or bank offers very high loan interest rate which is not feasible by the clients, thus they 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.



REPRESENTATION OF UNIVARIATE MERGED OF NAME_CASH_LOAN_PURPOSE AND NAME_CONTRACT_STATUS

Insights :

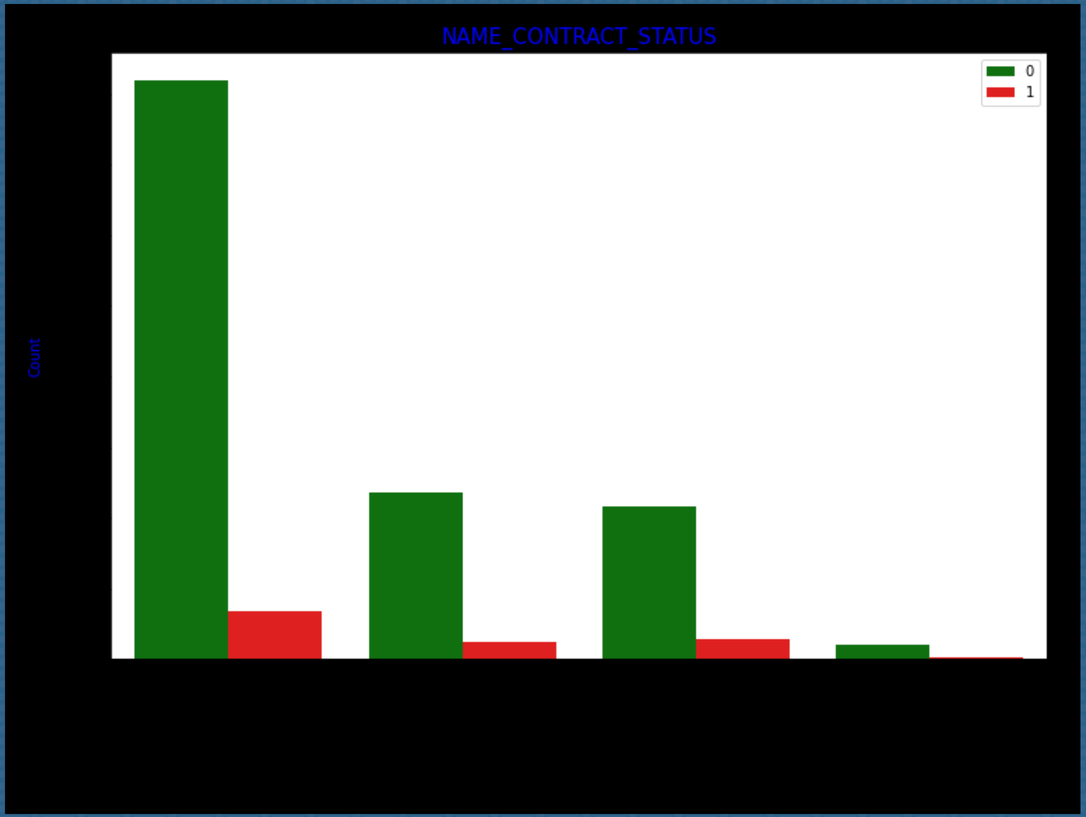
- A very high number application have been rejected by bank or refused by client which has purpose as repair or other. This shows that purpose repair is taken as high risk by bank and either they are rejected or bank offers very high loan interest rate which is not feasible by the clients, thus they 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

Insights :

- 88% of the clients who have been previously refused a loan has paid back the loan in current case.
- 90% of the previously cancelled client have actually repaid 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 repayor customer.

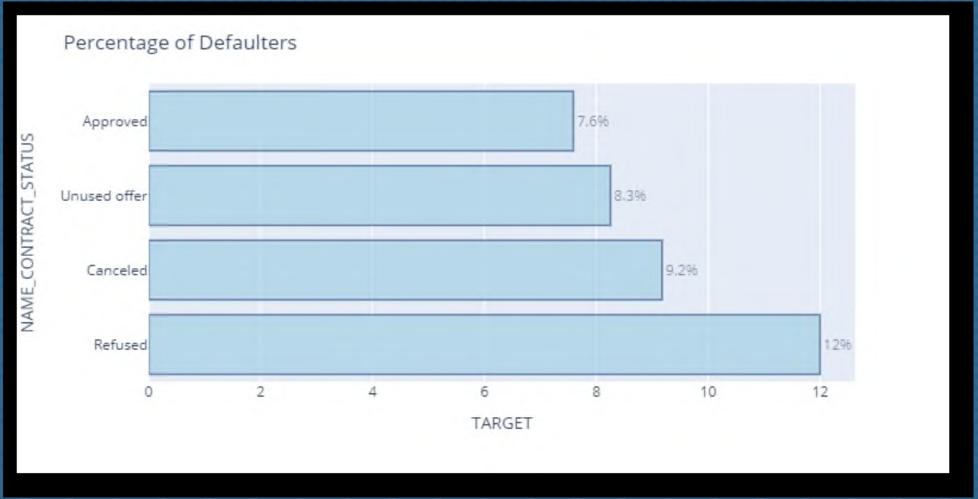


NAME_CONTRACT_STATUS	TARGET		
Approved	0	818856	92.41%
	1	67243	7.59%
Canceled	0	235641	90.83%
	1	23800	9.17%
Refused	0	215952	88.0%
	1	29438	12.0%
Unused offer	0	20892	91.75%
	1	1879	8.25%

PERCENTAGE OF DEFAULTERS OF MERGED DATAFRAME BASED ON PREVIOUS CONTRACT STATUS ANDCLIENTS

Insights:

- Large number of new applicants approved with revolving loans are defaulters.
- 12% clients were previously refused are defaulters.



OBSERVATION FROM THE DATA

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 the 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.

  AMT_GOODS_PRICE: When the credit amount goes beyond 3M, there is an increase in defaulters.

OBSERVATION FROM THE DATA

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_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

CONCLUSION

- Bank should more focus on the collecting information and investigation of whether the submitted docs are legit or not.
- Pay more attention to credit behavior of peer groups of clients
- Continue to prefer High education, High Income segments as profit segments,
- A new product can developed for lower income groups according to income cap.
- Social information about (Guarantor) should be specified it helps to identify in defaulter or re-payer