Credit-Default EDA

BY:

MANNE ASHRIT KUMAR.

SUJATA SWAIN.

About EDA

Problem Statement: 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 minimise the risk of losing money while lending to customers.

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 utilise this knowledge for its portfolio and risk assessment.

About Dataset

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.

Steps in EDA

Understanding the Data

- Observe the data frame.
- Checking Shape of data frame.
- Checking Datatypes of columns/Attributes of dataset.
- Observing Descriptive statistics of numerical Data of dataset.

Data Cleaning

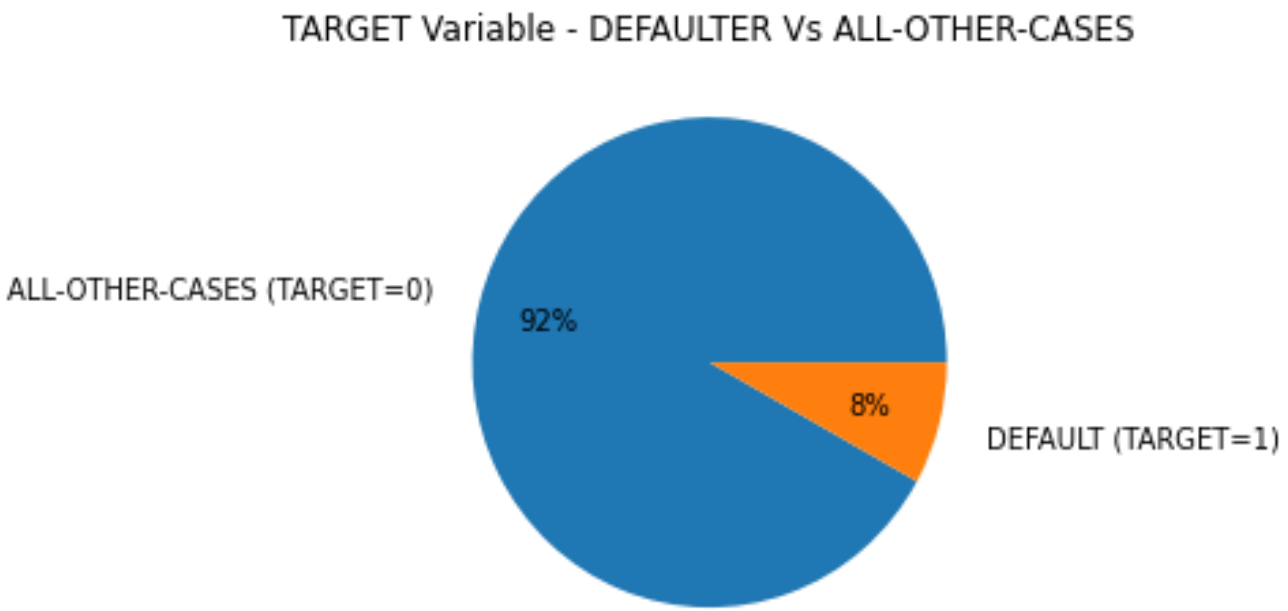
- Handling Missing values by dropping columns(high % of NULL VALUES), Imputation.
- Standardizing variables.
- Fixing Invalid Data-types and Filter Data(Correcting Data-types, Quality Checks.)
- Handling Outliers.

Data Analysis

- Univariate Analysis.
- Segmented Univariate Analysis.
- Bivariate Analysis and Multivariate Analysis.
- Data visualization of above Analysis.
- Checking Correlation.
- Drawing Insights from Analysis.

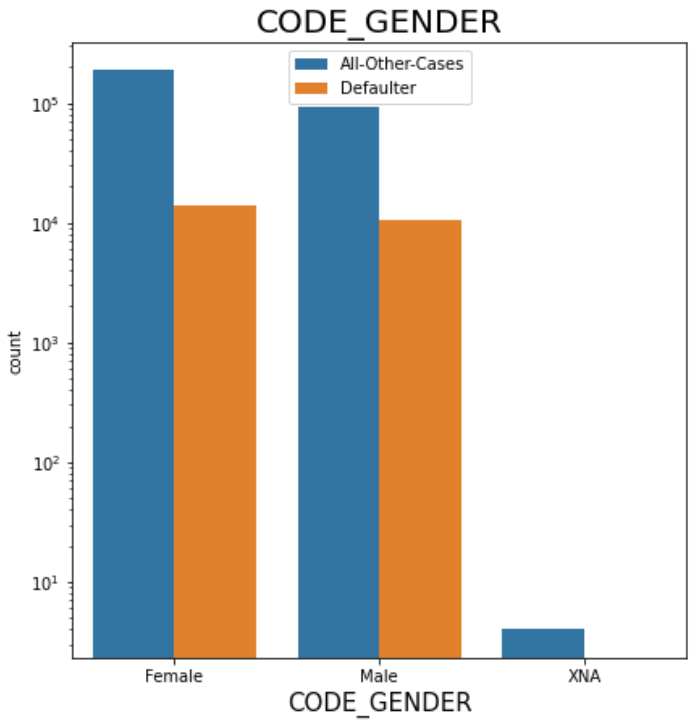
Application Dataset.

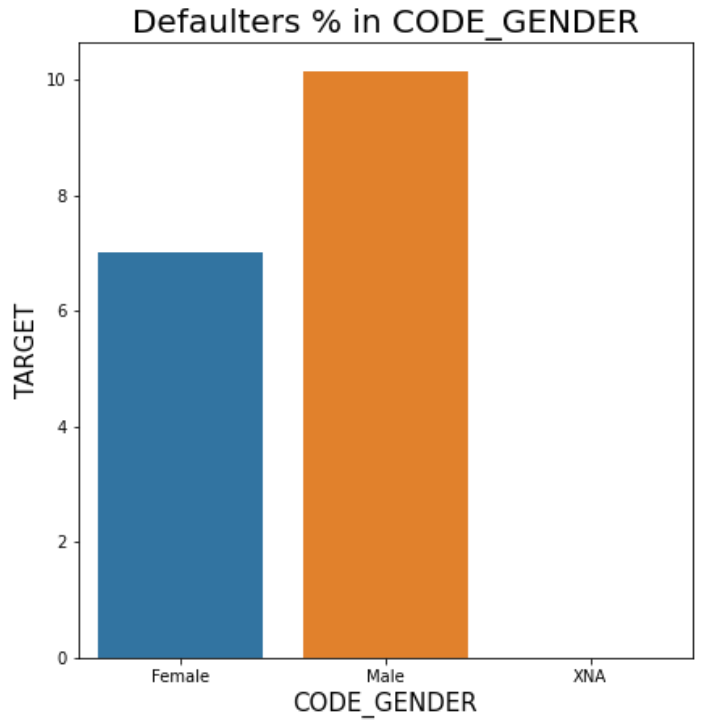
Imbalance between people who defaulted and who didn't default. More than 92% of people didn't default as opposed to 8% who defaulted.



The number of female clients is almost double the number of male clients.

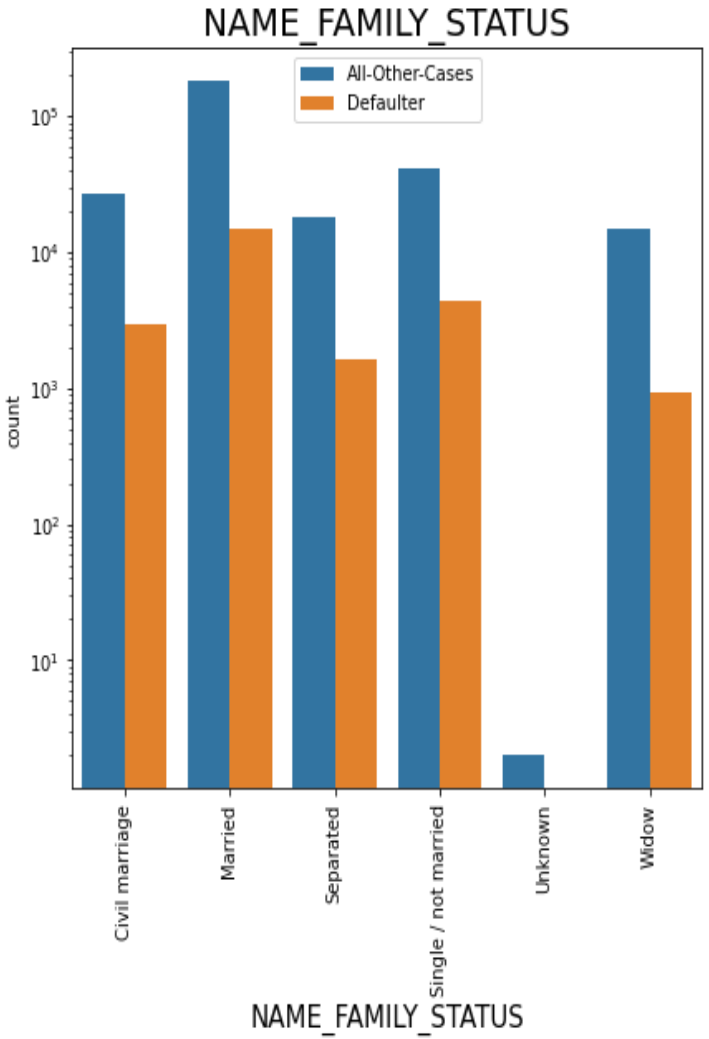
Based on the percentage of defaulted applicants, males have a higher chance of defaulting, comparing with women.

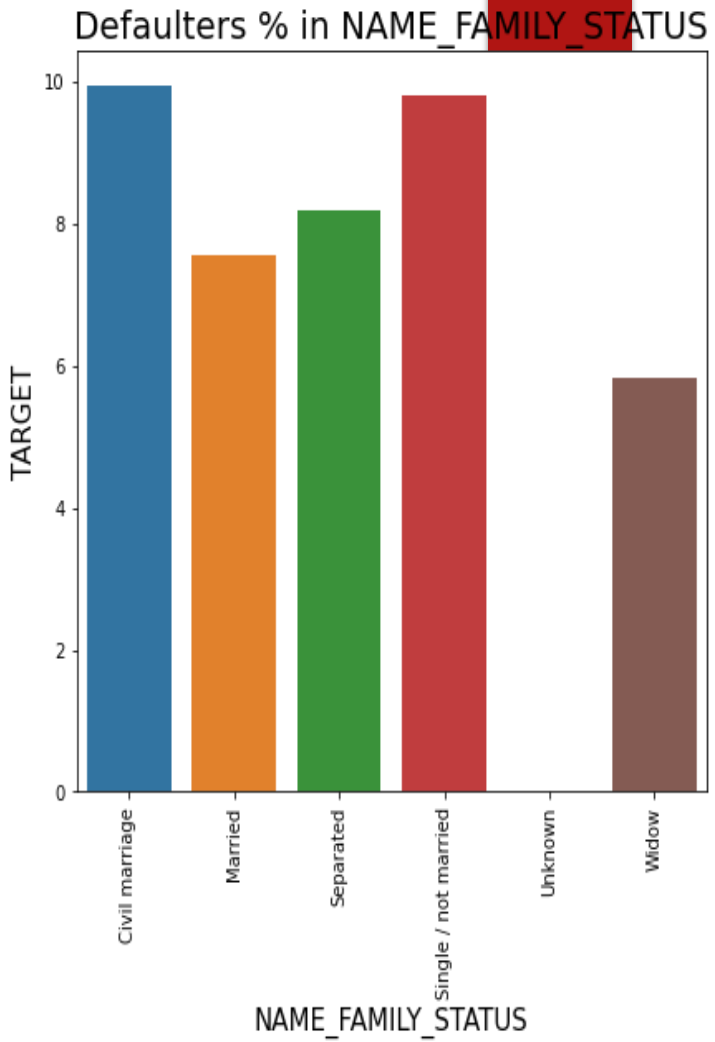




Most of the people who have taken loan are married, followed by Single/not married and civil marriage

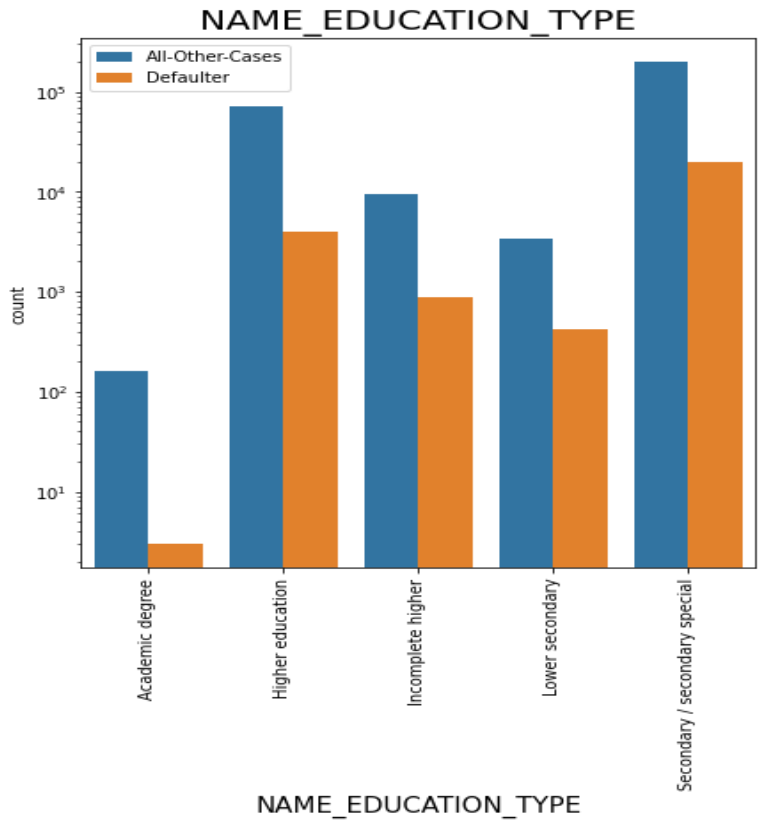
In Percentage of defaulters, Civil Marriage and Single have the highest percent around and widow has the lowest.

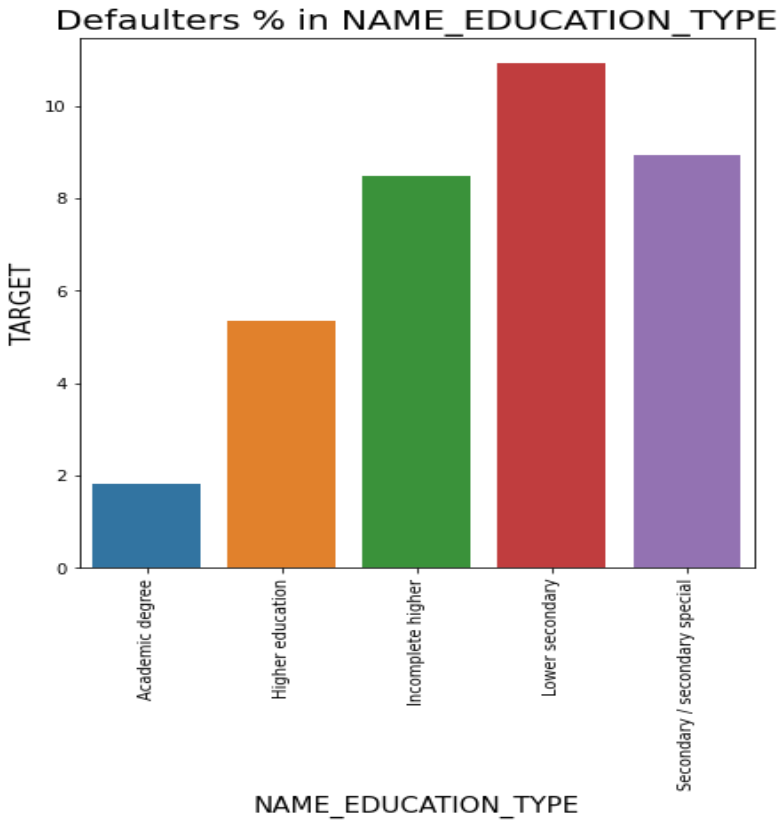




Lower secondary category have highest rate of defaulting.

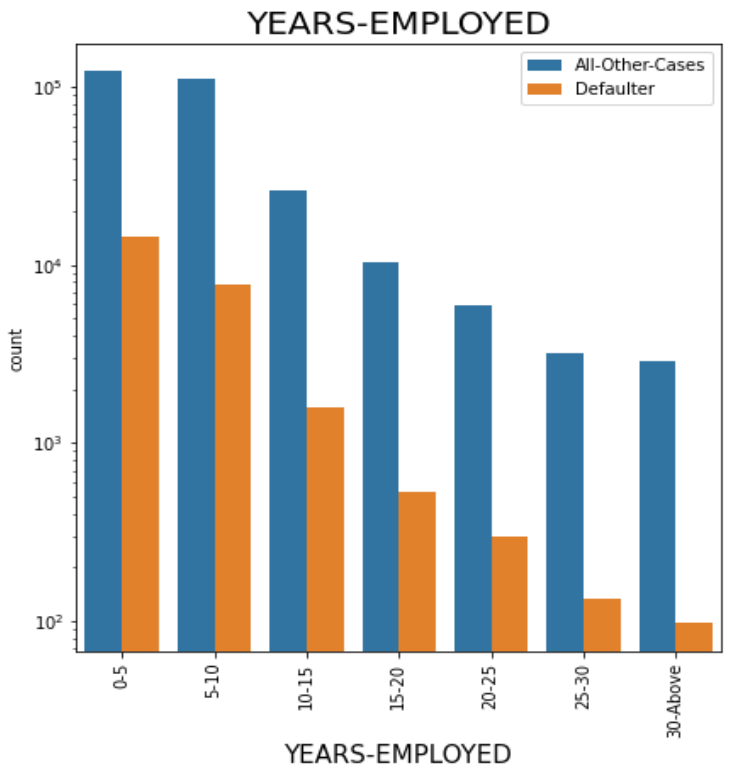
People with Academic degree are least likely to default.

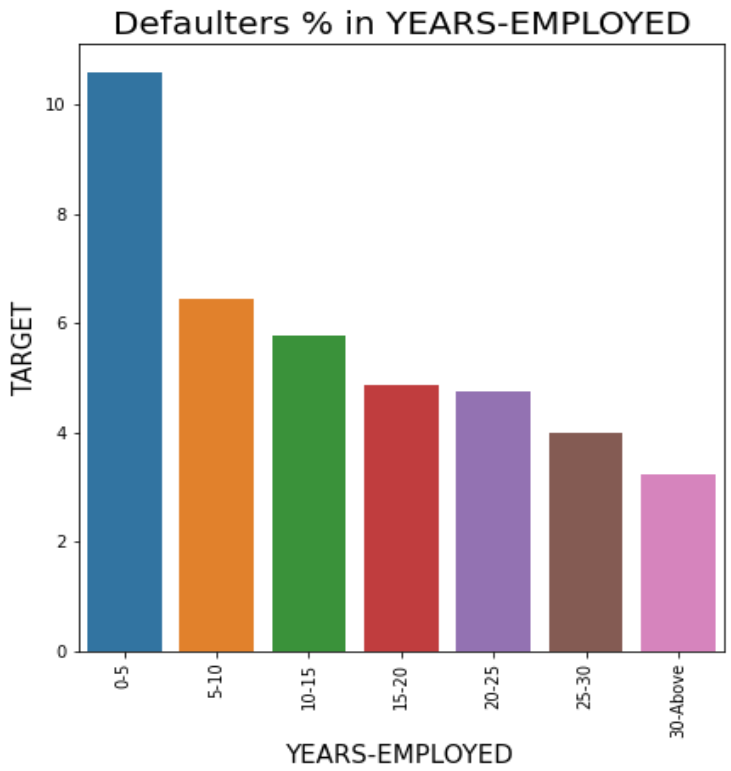




Majority of the applicants having working experience between 0-5 years are defaulters.

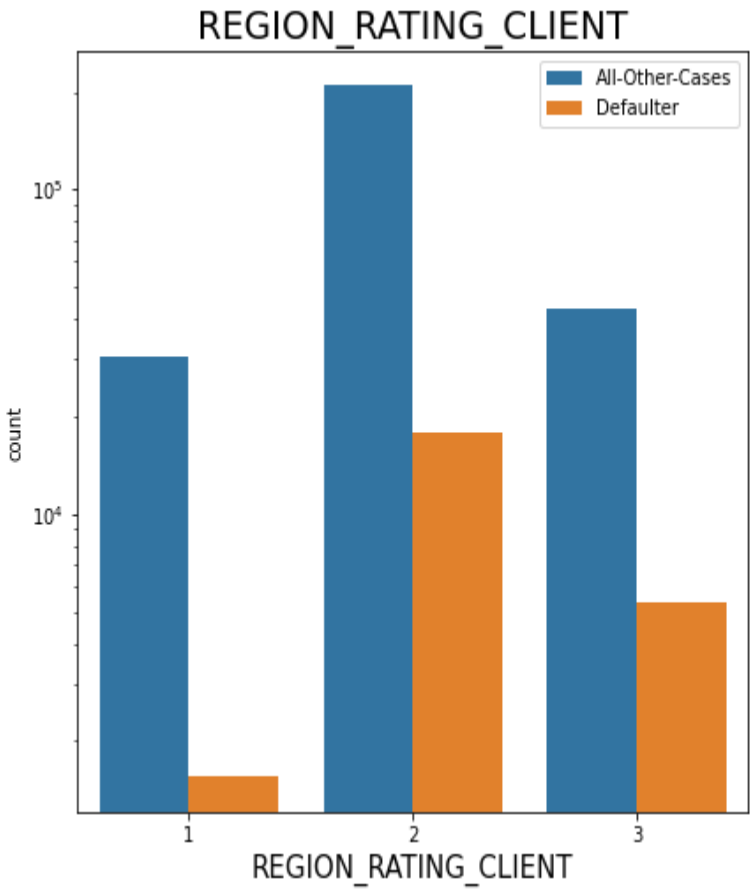
With increase of employment year, defaulting rate is gradually decreasing.

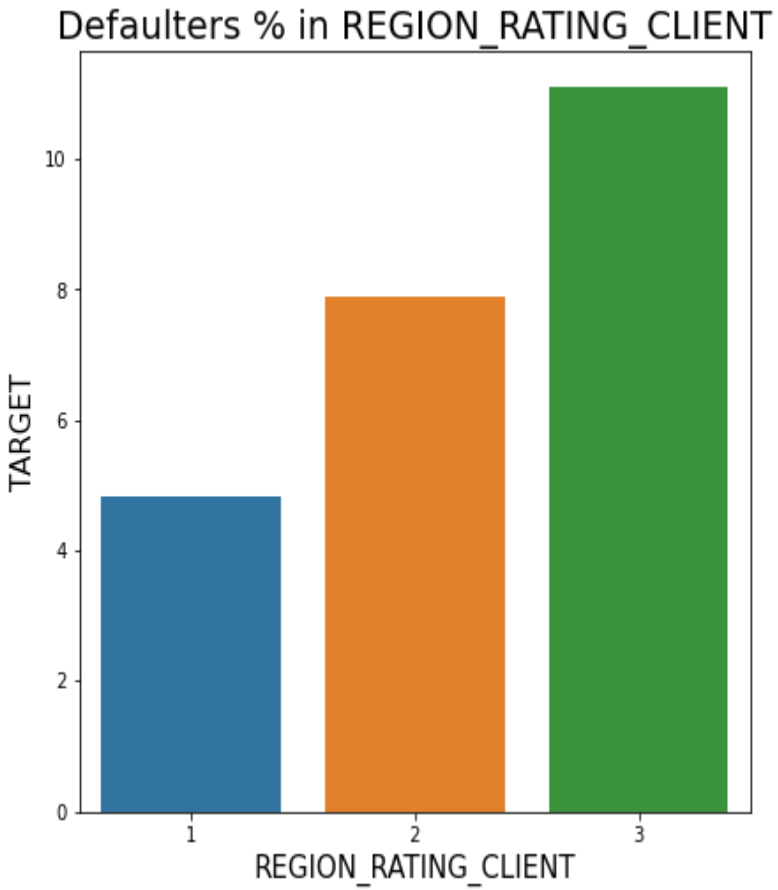




More people from second tier regions tend to apply for loans.

We can infer that people living in better areas(Rating 3) tend contribute more to the defaulters by their weightage. People living in 1 rated areas

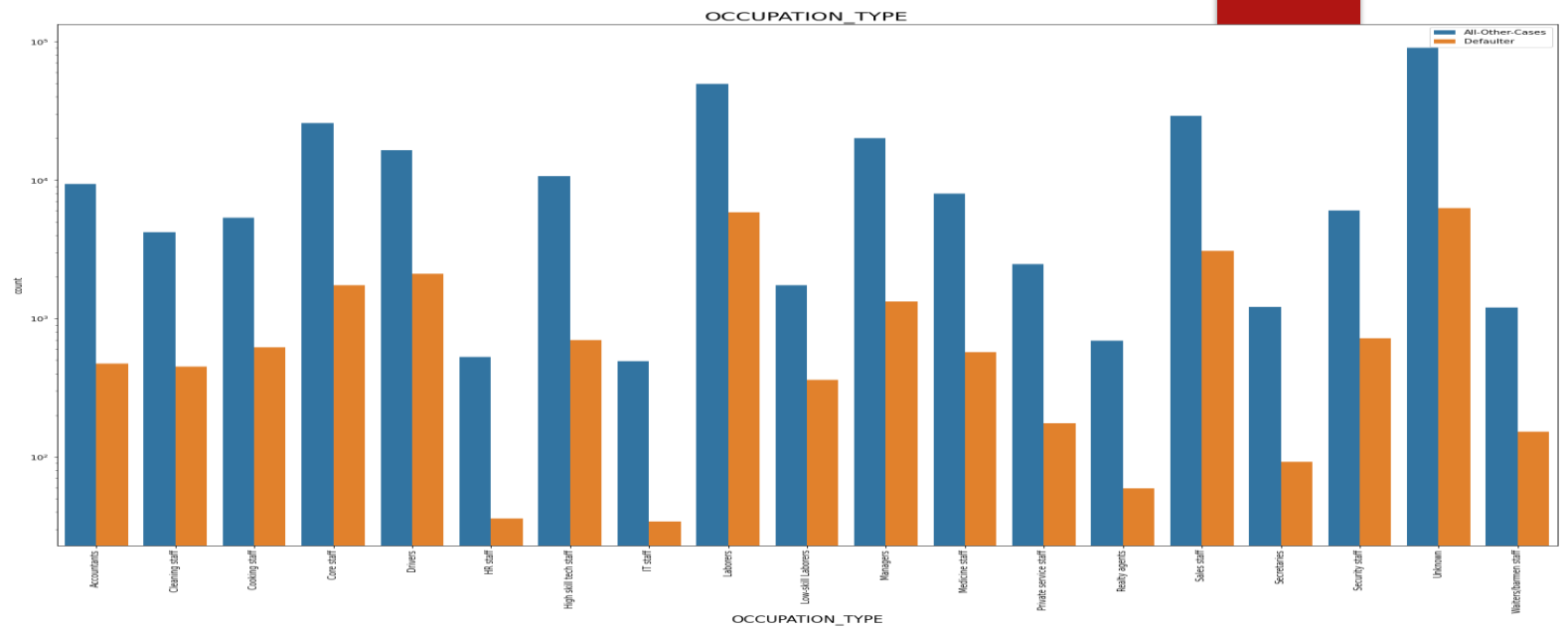


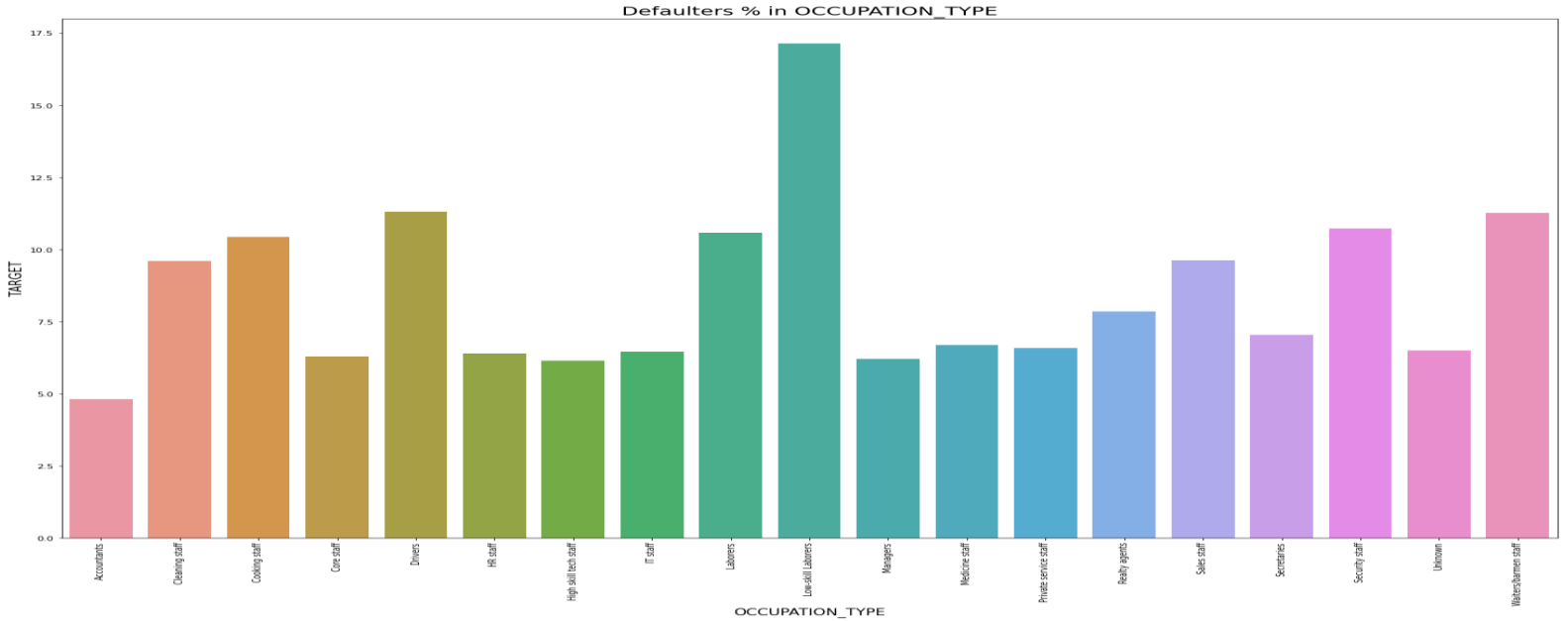


Most of the loans are taken by Laborers, followed by Sales staff.

IT staff are less likely to apply for Loan.

Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as their default rate is huge.





Top 10 Correlation for Defaulters

	Var-X	Var-Y	Correlation	Abs_Correlation
64	AMT_GOODS_PRICE	AMT_CREDIT	0.982783	0.982783
65	AMT_GOODS_PRICE	AMT_ANNUITY	0.752295	0.752295
43	AMT_ANNUITY	AMT_CREDIT	0.752195	0.752195
131	DAYS_EMPLOYED	DAYS_BIRTH	0.582185	0.582185
152	DAYS_REGISTRATION	DAYS_BIRTH	0.289114	0.289114
300	FLAG_DOCUMENT_3	DAYS_EMPLOYED	-0.272169	0.272169
263	DEF_60_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.264159	0.264159
173	DAYS_ID_PUBLISH	DAYS_BIRTH	0.252863	0.252863
351	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_HOUR	0.247511	0.247511
174	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.229090	0.229090

Top 10 Correlation for Non-Defaulters

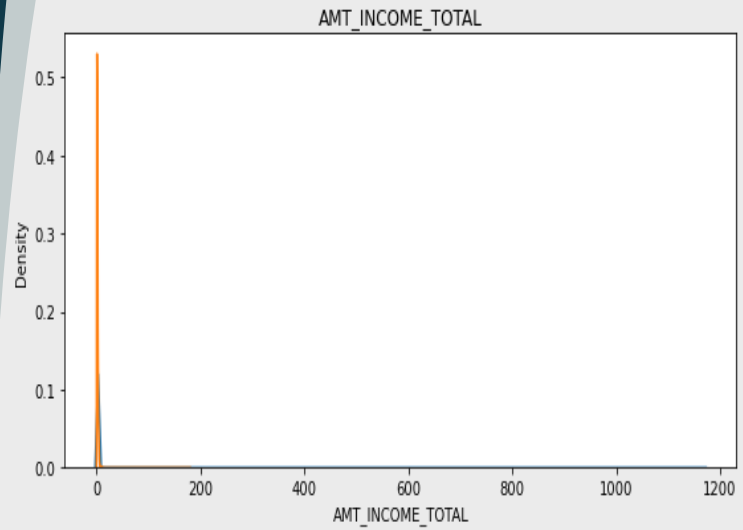
	Var-X	Var-Y	Correlation	Abs_Correlation
64	AMT_GOODS_PRICE	AMT_CREDIT	0.987022	0.987022
65	AMT_GOODS_PRICE	AMT_ANNUITY	0.776433	0.776433
43	AMT_ANNUITY	AMT_CREDIT	0.771309	0.771309
131	DAYS_EMPLOYED	DAYS_BIRTH	0.626114	0.626114
42	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953	0.418953
63	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349426	0.349426
21	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799	0.342799
152	DAYS_REGISTRATION	DAYS_BIRTH	0.333151	0.333151
174	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.276663	0.276663
173	DAYS_ID_PUBLISH	DAYS_BIRTH	0.271314	0.271314

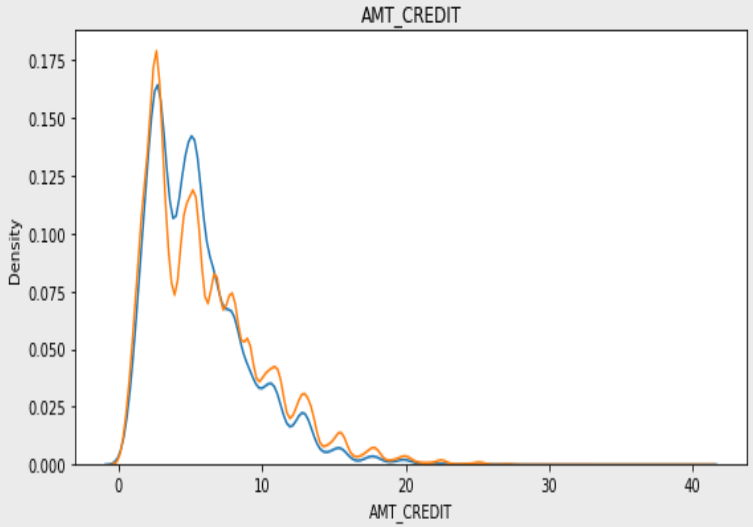
Most no of loans are given for goods price below 10 lakhs

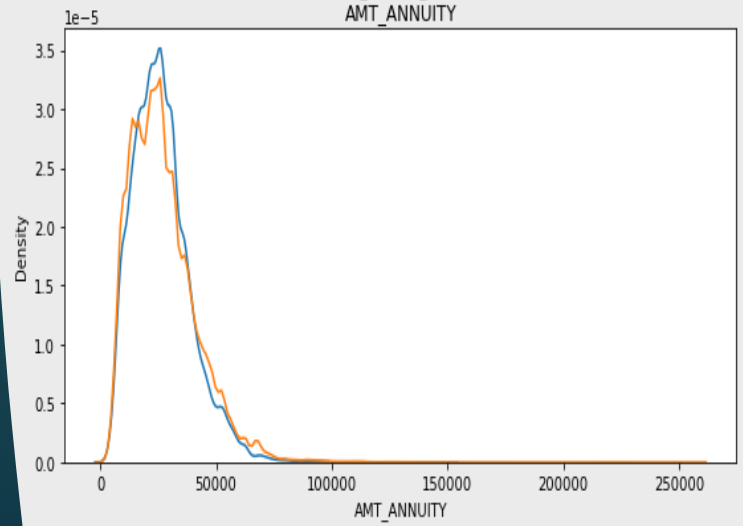
Most people pay annuity below 50K for the credit loan

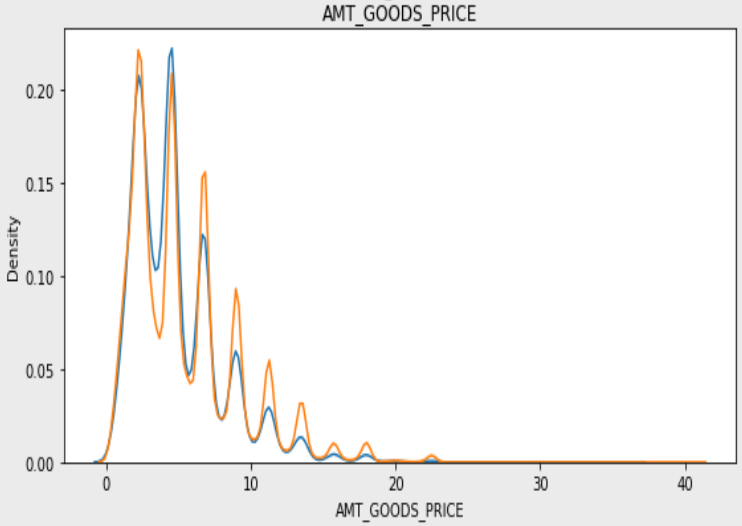
Credit amount of the loan is mostly less then 10 lakhs

The non-defaulters and defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision

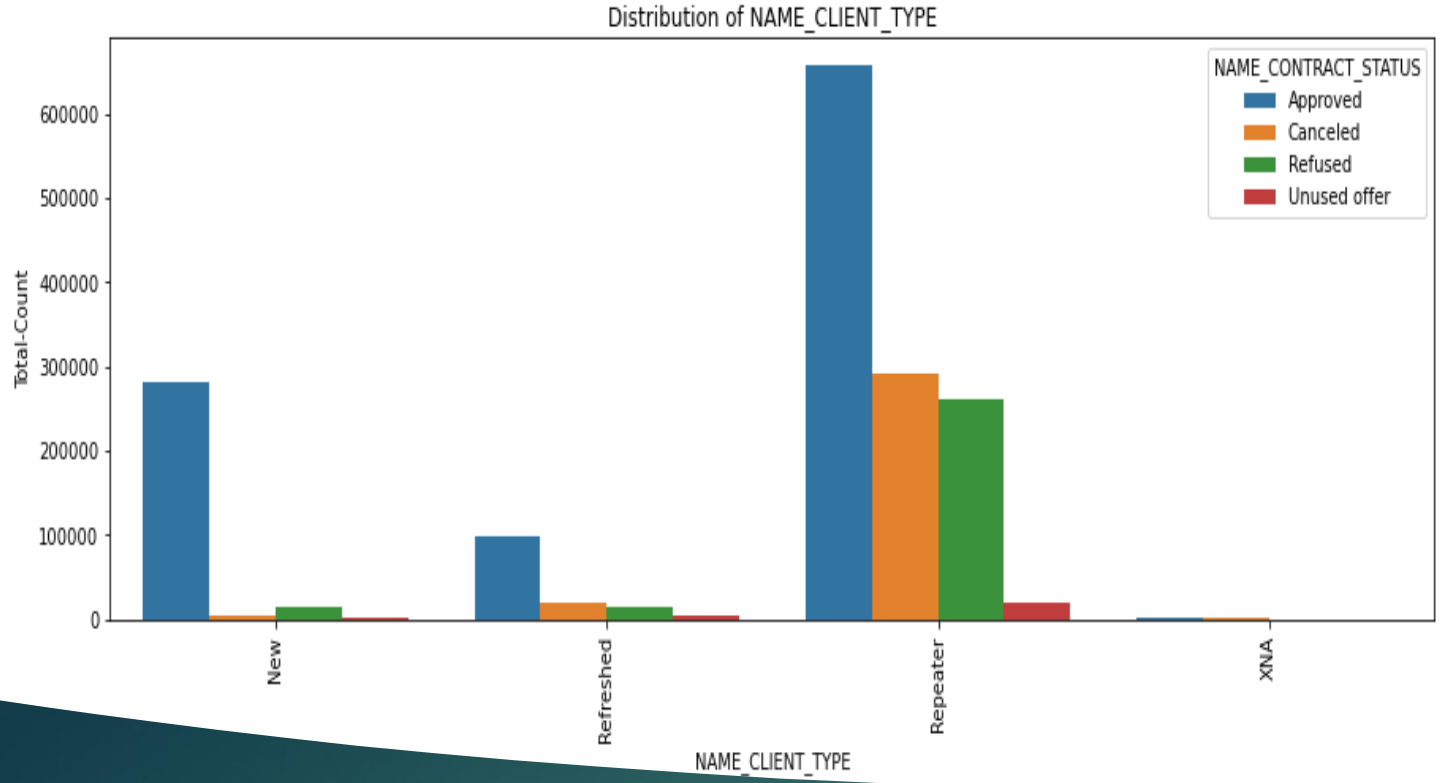




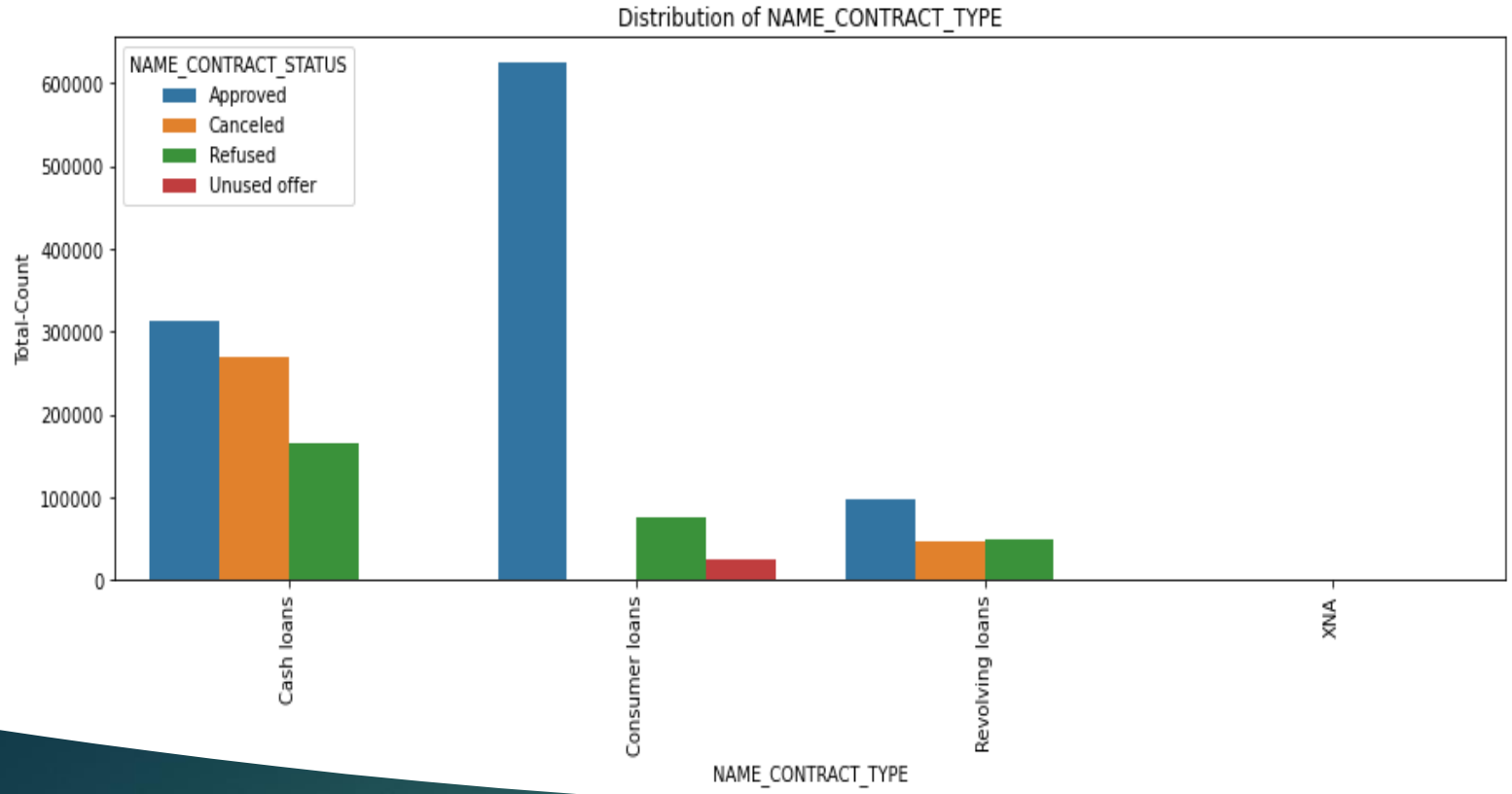




Previous Application DATA

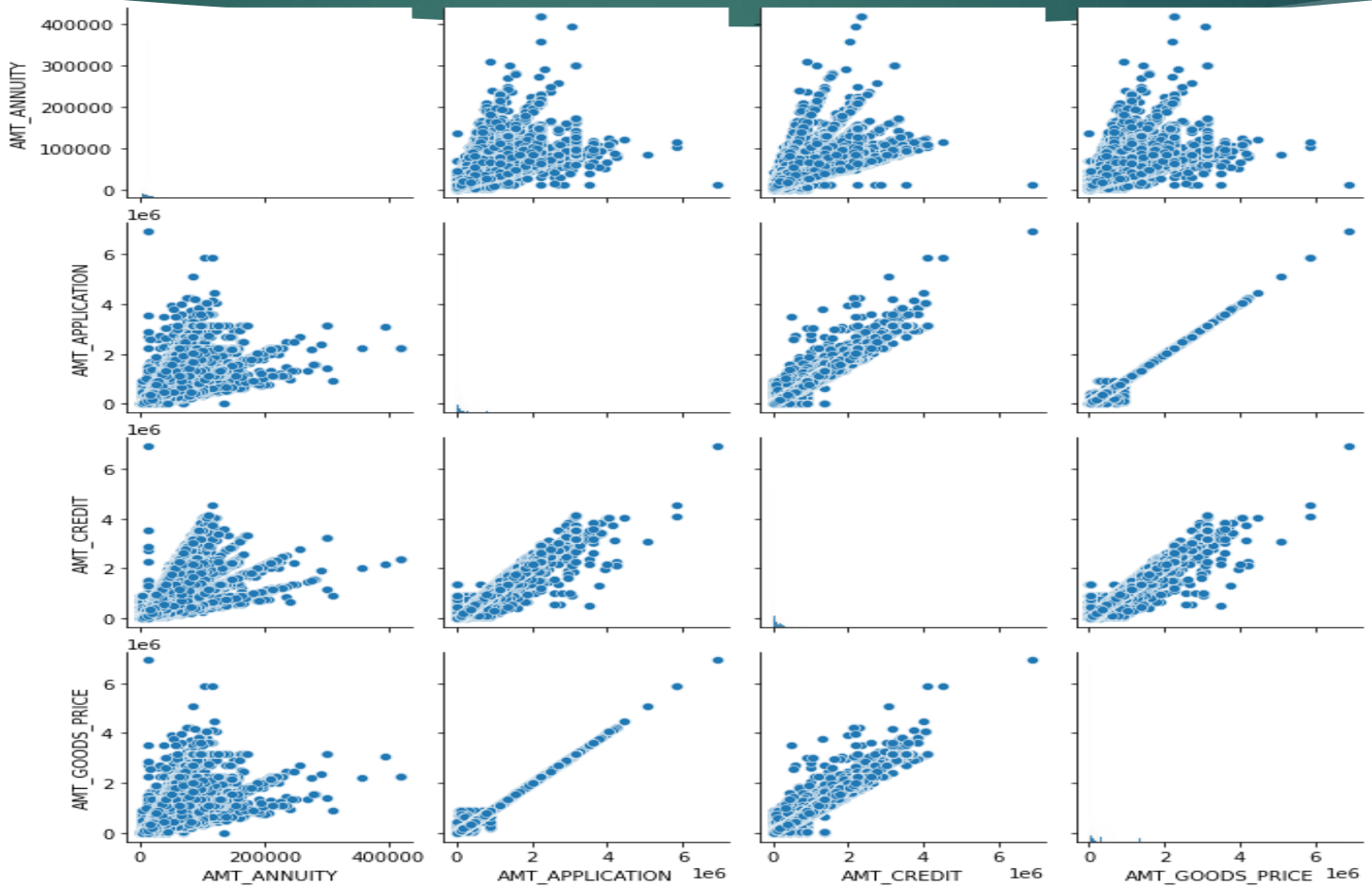


Banks are inclined to give loans to repeat customers



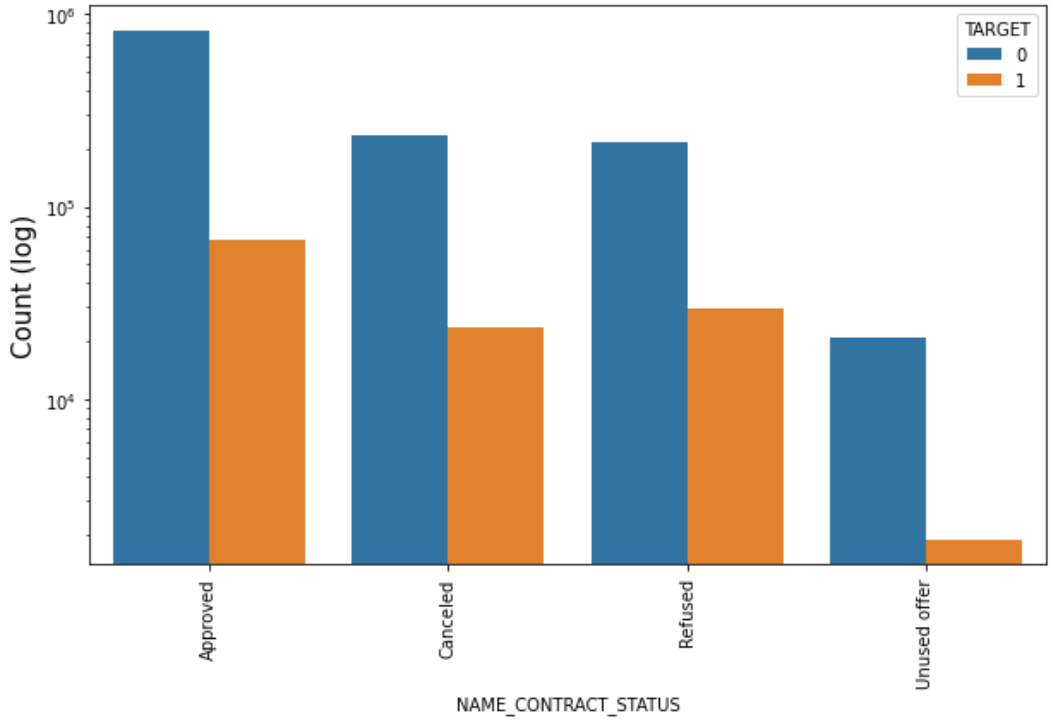
From above we can see that, Bank approval rate for consumer loan is very high. May be because many types of consumer loans such as car loans, student loans, mortgage loans come with some type of collateral.

Application Amount, Credit, Goods Price are highly related.

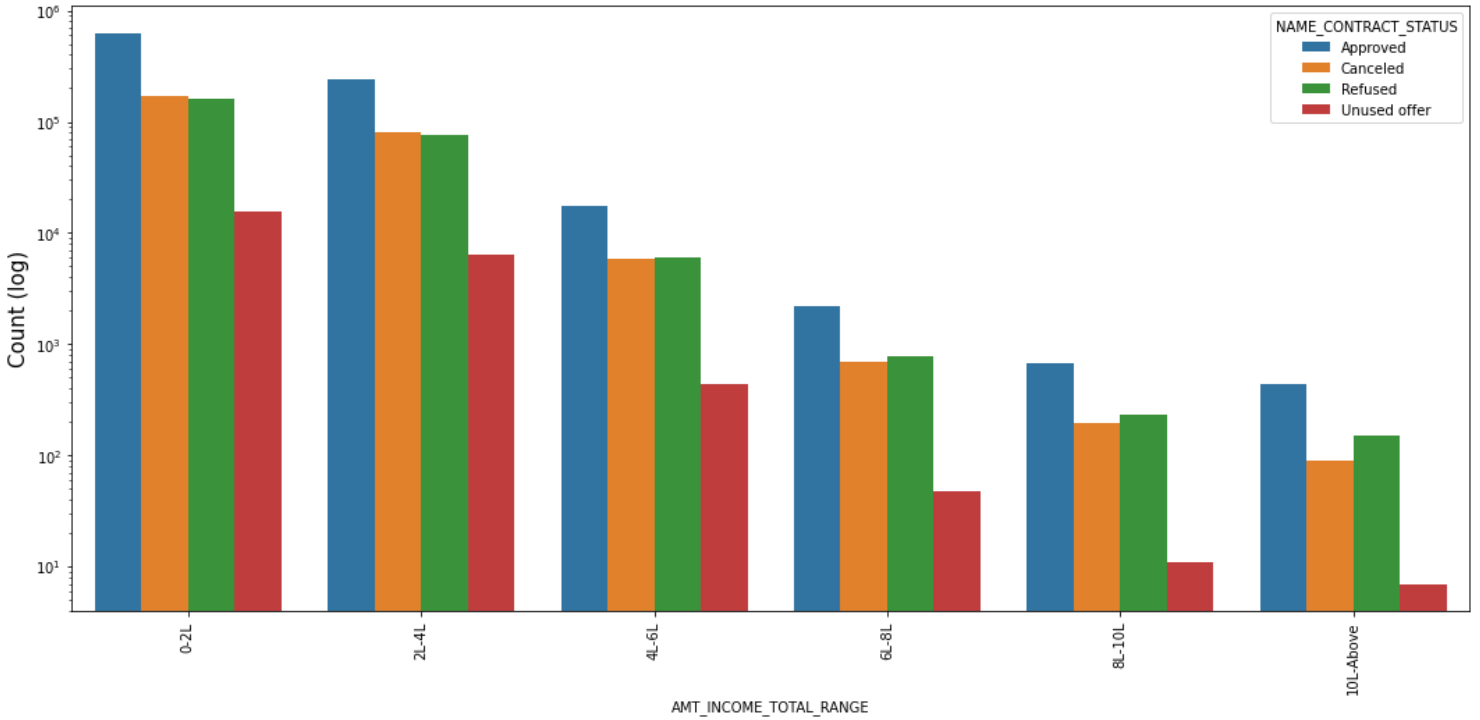


Merged data analysis

Most of the previously cancelled client have actually repayed the loan. Revising the interest rates would increase business opportunity for these clients. Most of the clients who have been previously refused a loan has payed back the loan in current case.



Clients with high incomes have low refusal rates.



Conclusions--Defaulters

Factors to Determine Defaulting behavior among Clients:

- Gender—Men.
- Family Status—Civil Marriage and Single.
- Education—Lower Secondary and Secondary.
- Occupation—Low-skill laborers, Waiters/Barsmen, Drivers, Security Staff, Loborers.
- Region Rating– Regions with rating 3.
- Employment Experience—less than 5 years riskier.

Conclusions--Non-Defaulters

Factors to Determine No-Defaulting behavior among Clients:

- Education—Academic Degree holder.
- Income-Type—Student, Businessmen.
- Employment Experience--40+
- Region Rating—Regions with rating 1 are safer.