Credit EDA Case Study

by Rancy Chadha

Problem Statement – Objective - Approach

Problem Statement:

Financial companies often find it difficult to decide loan approval or rejection for an applicant. It may lead to following:

* Rejecting a loan to an applicant who is capable of pay loan is a loss of business to the companies

* On the contrary Approving a loan to an applicant who is likely to default, then approving the loan may lead to a financial loss for the company.

Objective of the case study is to perform Exploratory Data Analysis on financial datasets provided and find variables that are deciding factors between Defaulters and Payers

---

Approach

Importing &
Data Cleaning

Data Formatting &
Standardization

Perform Univariate/Bivariate
Analysis

Draw Inferences

Approach for Missing Values and finding outlier

Strategy to work with Missing Values:

Dropped columns that had greater than 40% null values
Imputed missing values of categorical columns with MODE
Imputed missing values of continuous columns with MEDIAN after checking box plots
Checked columns correlation with Target before deciding to drop them or keep them

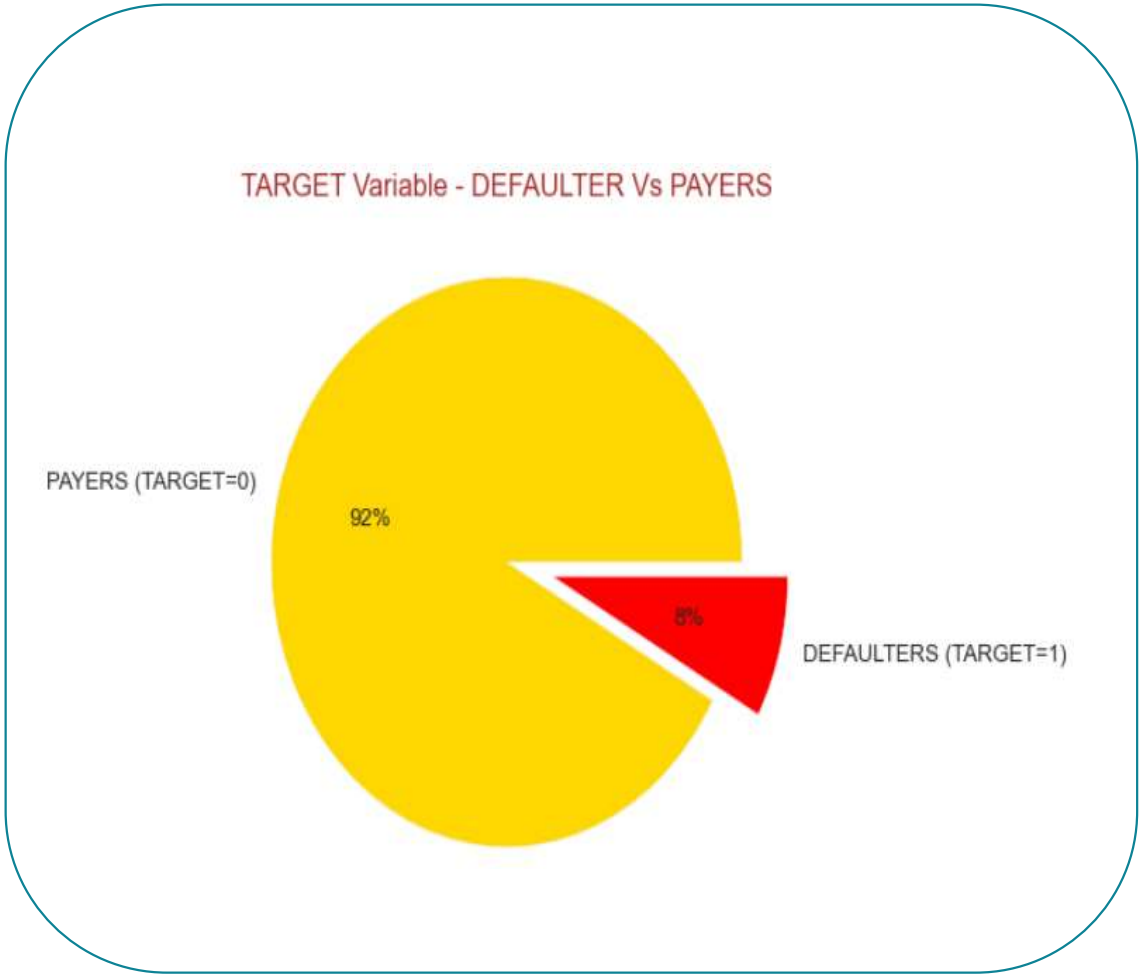
Strategy to identify Outliers:

For both the datasets used boxplots to identify outliers

Data Imbalance

It is evident that there is data imbalance in the target variable

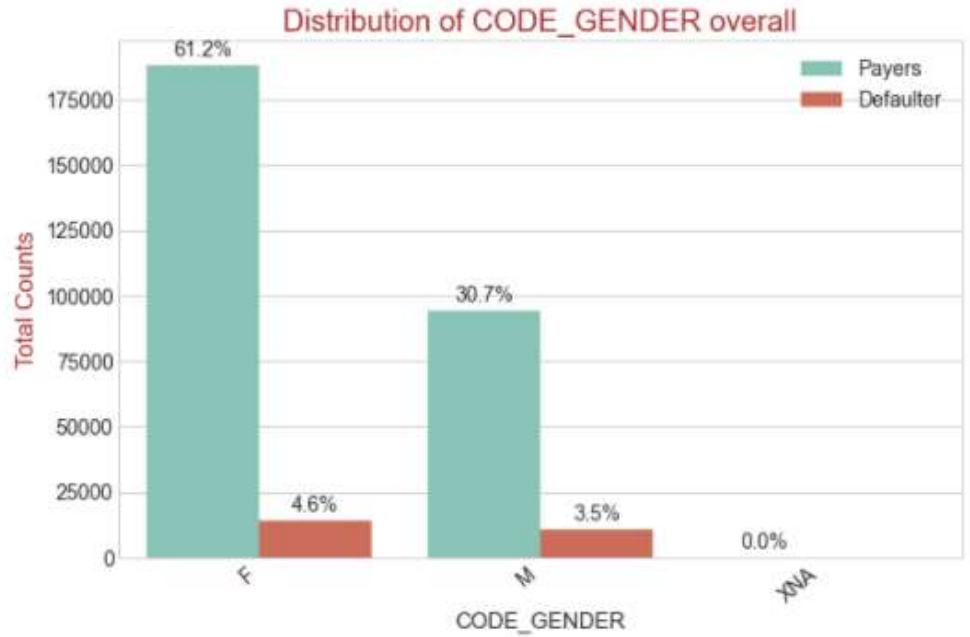
The ratio of imbalance is 9:1 between Payers and Defaulters respectively



Univariate Analysis - Gender

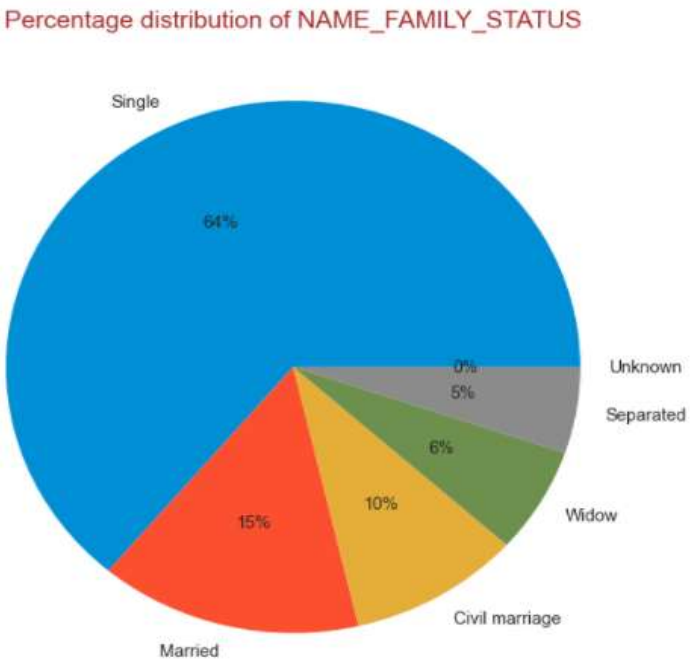
Women applicant are double than their male counterparts by almost double

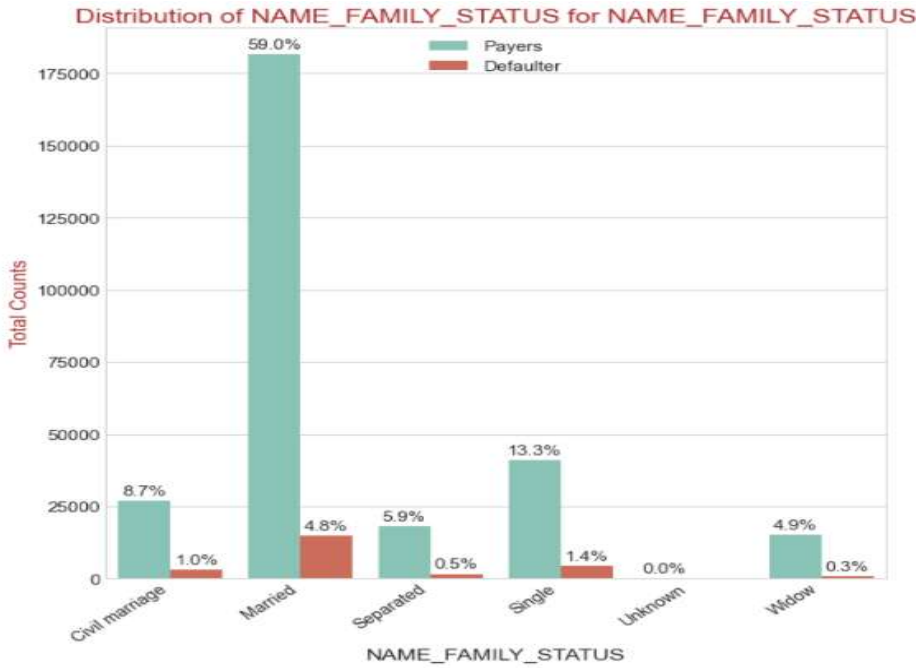
Defaulter percentage of females is comparatively lesser than males even though it shows higher, since the applicant % is double than males



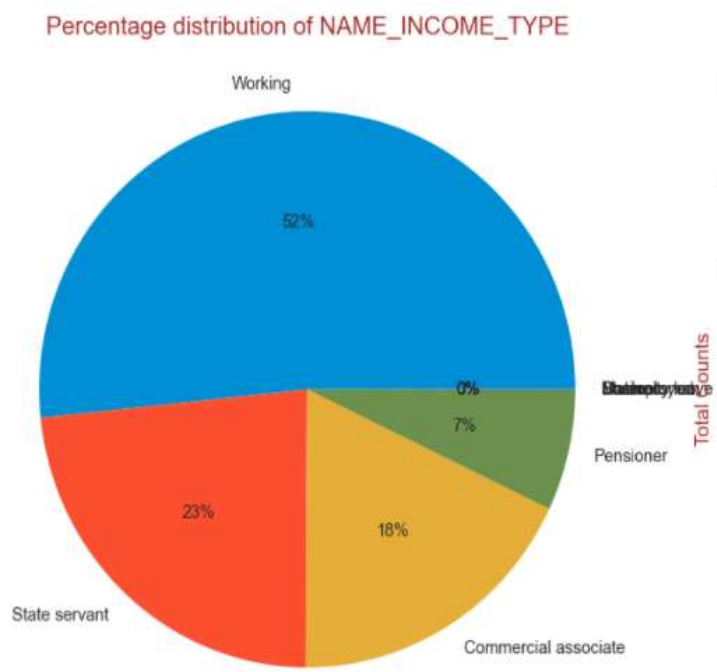
Univariate Analysis – Family Status

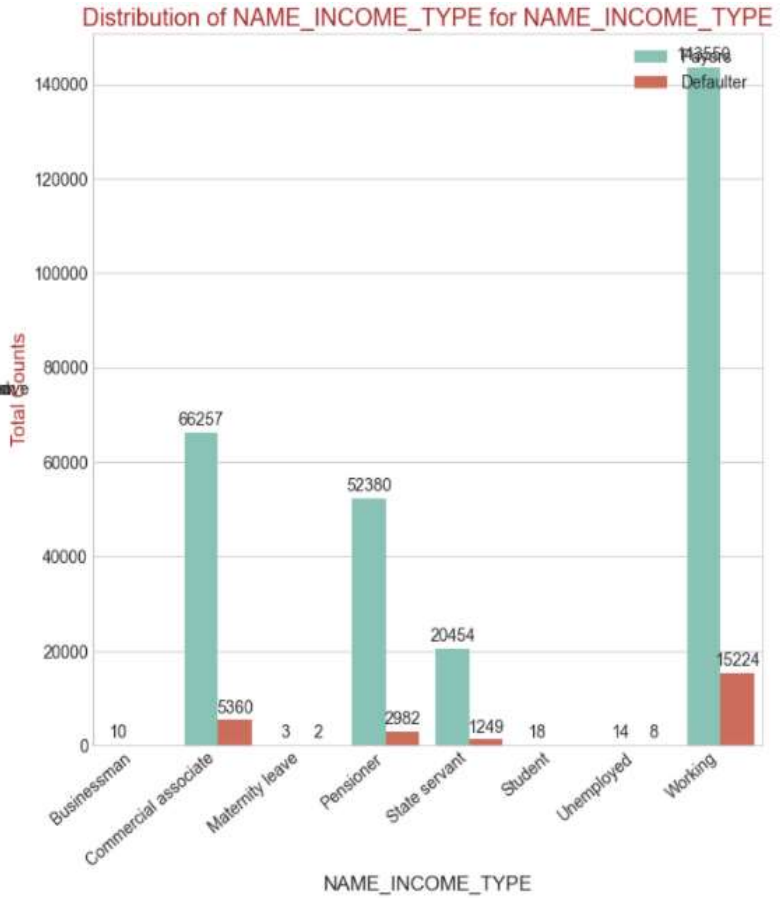
Ratio of Single persons applying for loan is double to the married people
However Married and Civil Married people cumulatively tend to default much more than the single/widow/separated category





Univariate Analysis – Income Type



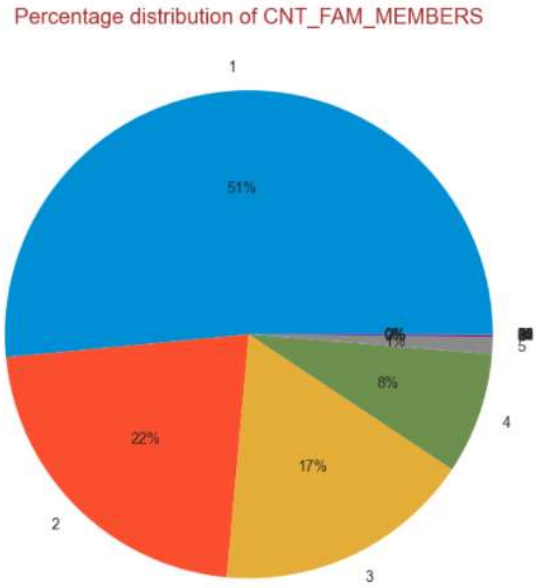


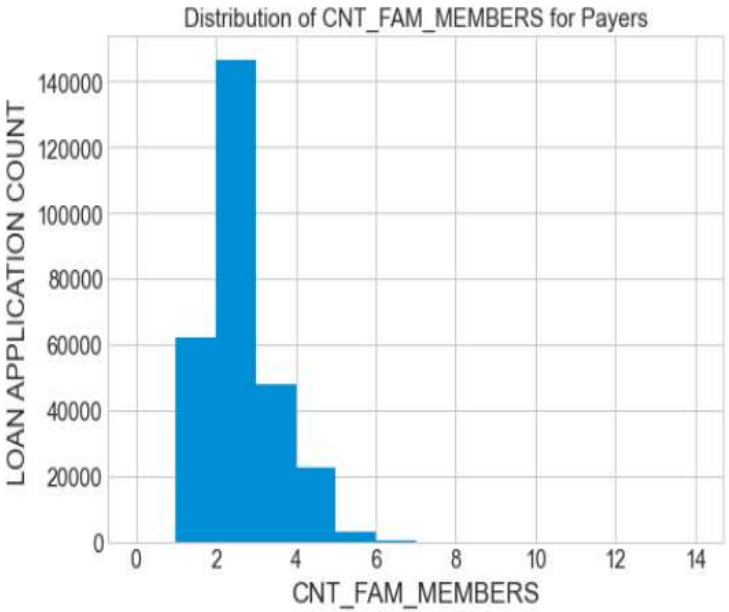
Most of the applicants are From Working Category & Commercial Associate and Least are Businessman and Student

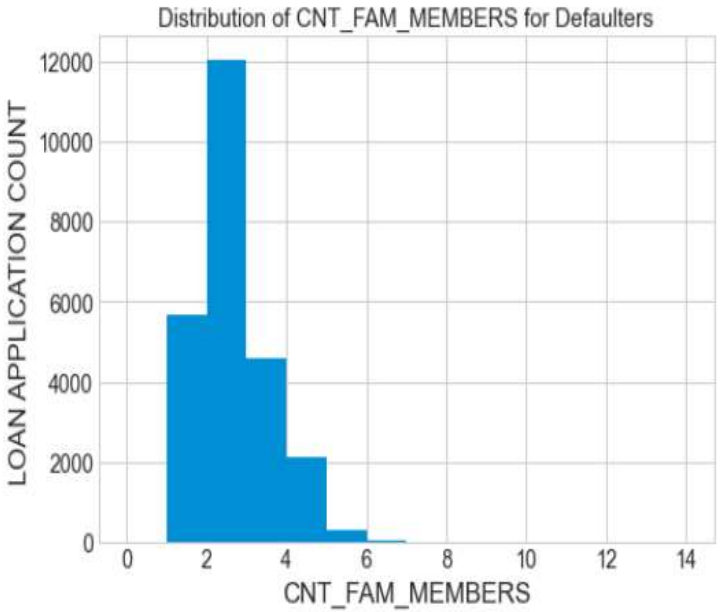
Defaulter percentage is quite high for women on Maternity Leave and Unemployed applicant group.

For unemployed they do not have income which result in higher default rate.
As for women onmaternity leave it could be they are on without pay

Univariate Analysis – Family Members



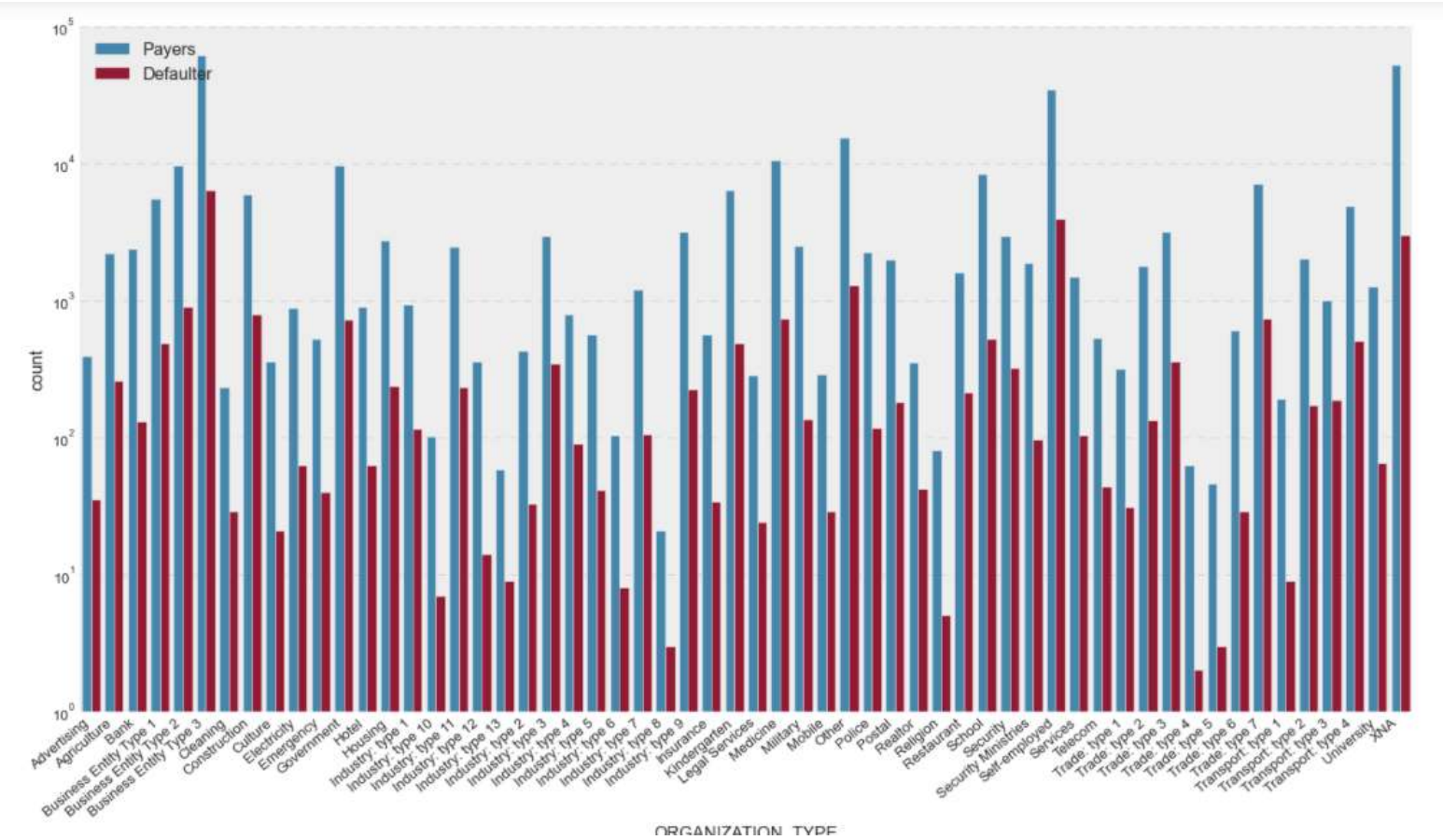




While family of 1 and 2 formed major chunk of applicants but their percentage of defaulting were also low

Additionally, it appears that family of 3 applies loan more often than the other families.

Univariate Analysis – Organization Type



Most of the Applicants are from Business Entity Type 3 , Missing (XNA) and Self Employed

Self Employed people seemingly have pretty high defaulting rate

Financial institutes need to scrutinize self employed people and to be on safer side they can offer loan at higher interest rate

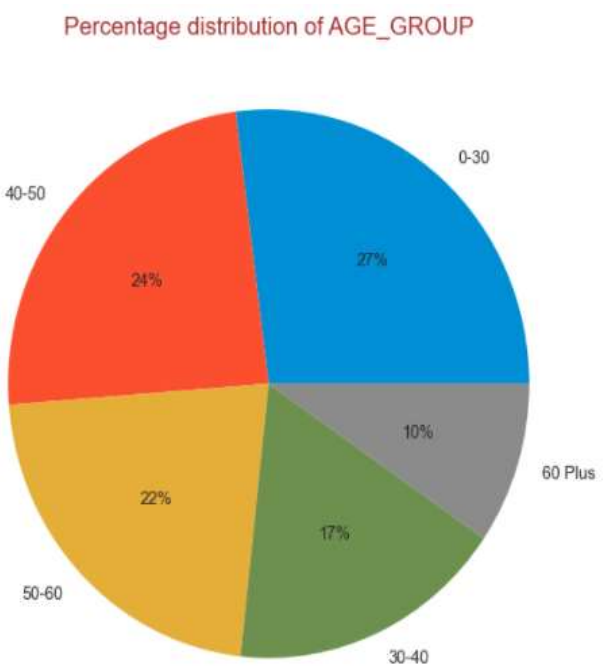
Univariate Analysis – Age Group

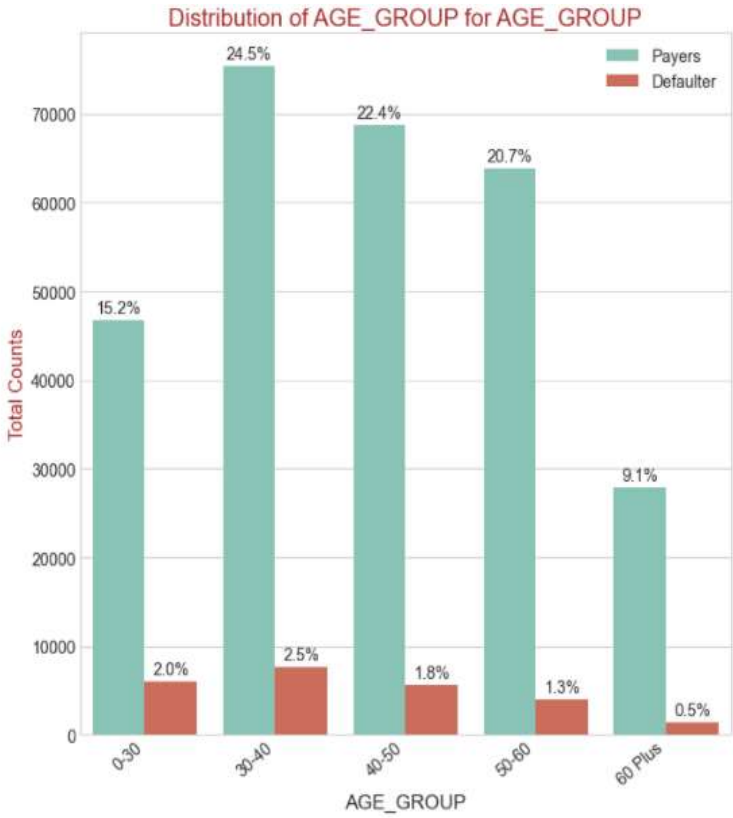
0-30 and 40-60 age group tend to default more often. So they are the riskiest people to loan to.

With increasing age group, people tend to default less starting from the age 40.

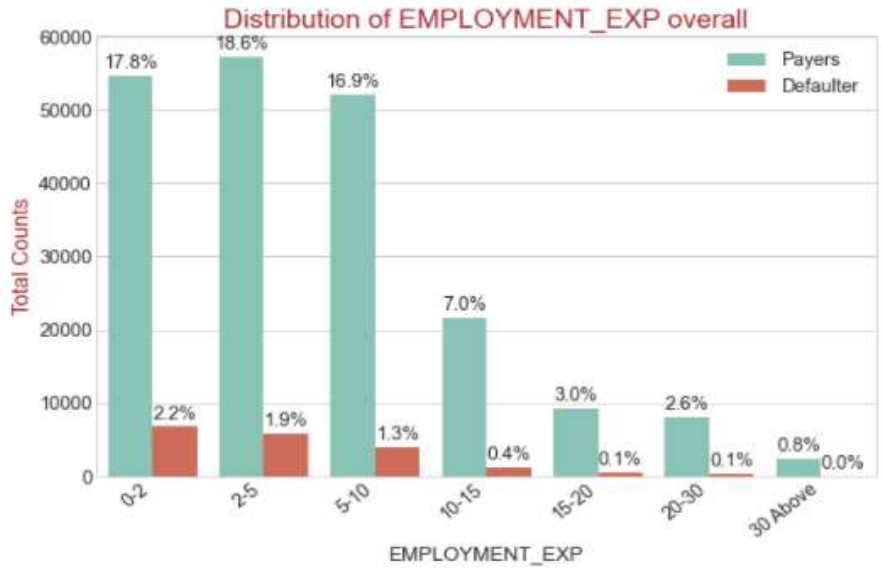
One of the reasons could be that they have sufficient salary and saving to pay back loans

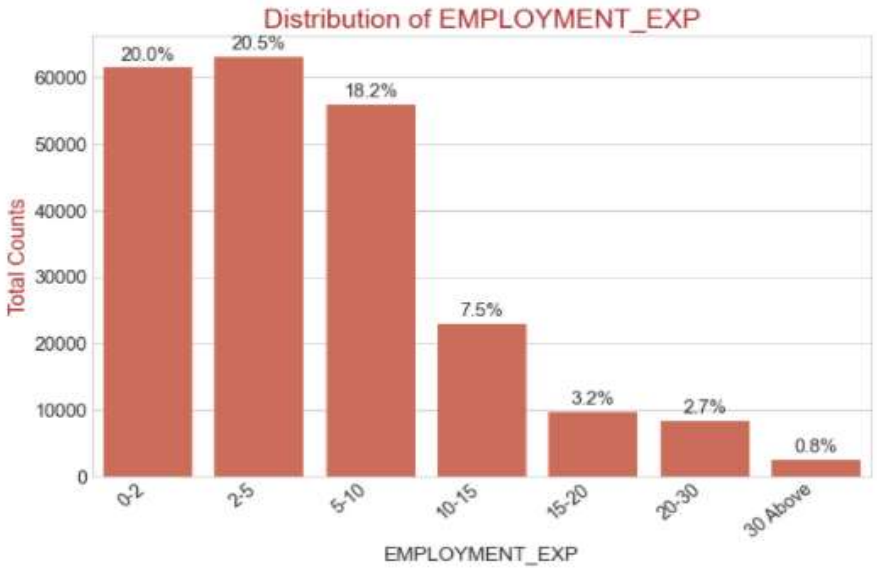
.





Univariate Analysis – Employment Experience

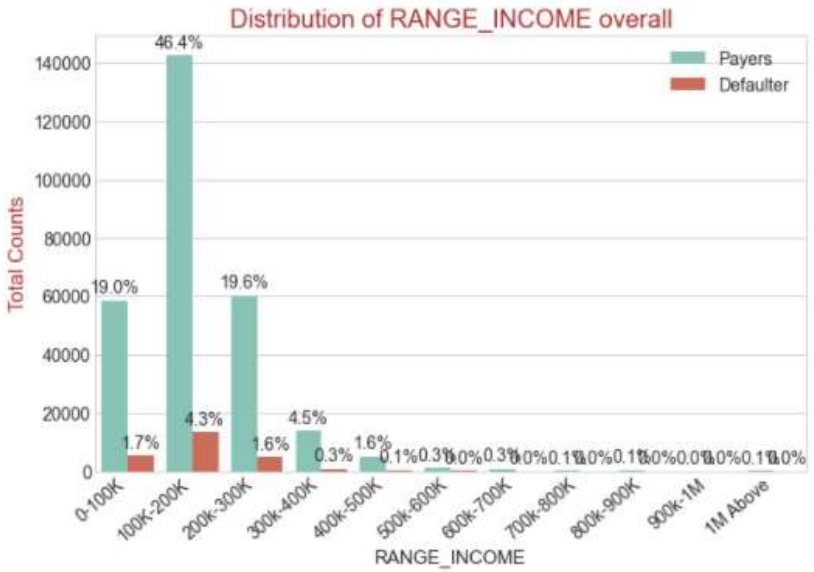


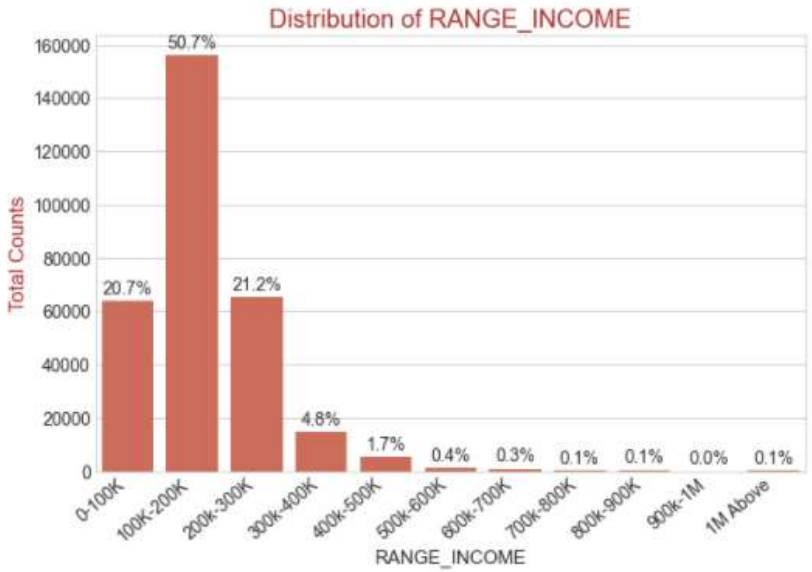


Majority of the applicants have been employed in between 0-5 years. The defaulting % of this group is also the highest

With increase of employment year, defaulting rate is gradually decreasing with people having 30+ year experience

Univariate Analysis – Income Range



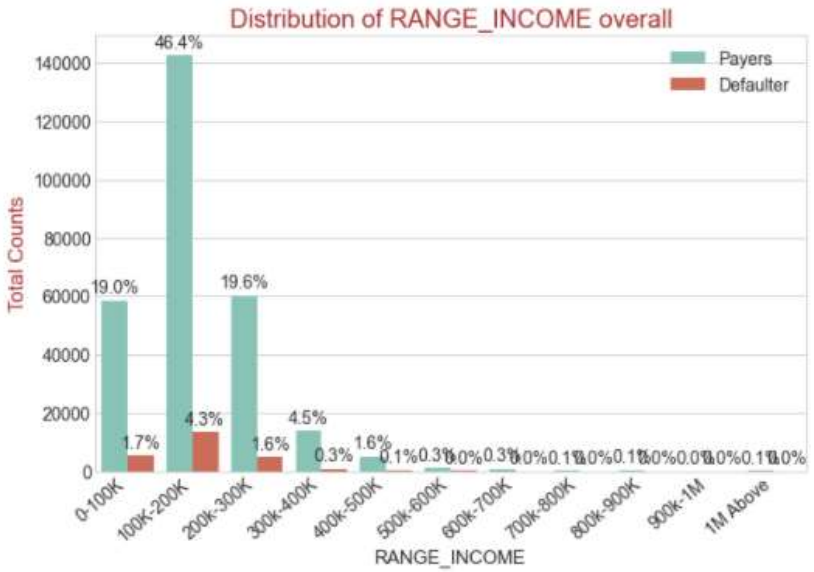


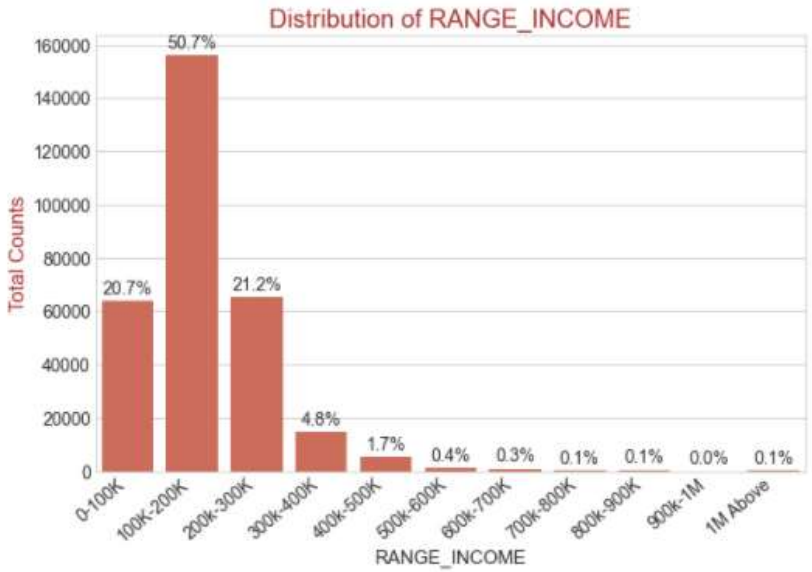
Almost 90% of applicant are in less than 300K income range

Defaulting percentage is higher among income in bracket 0-200K

Applicants with greater than 600K income are quite less likely to default

Univariate Analysis – Income Range



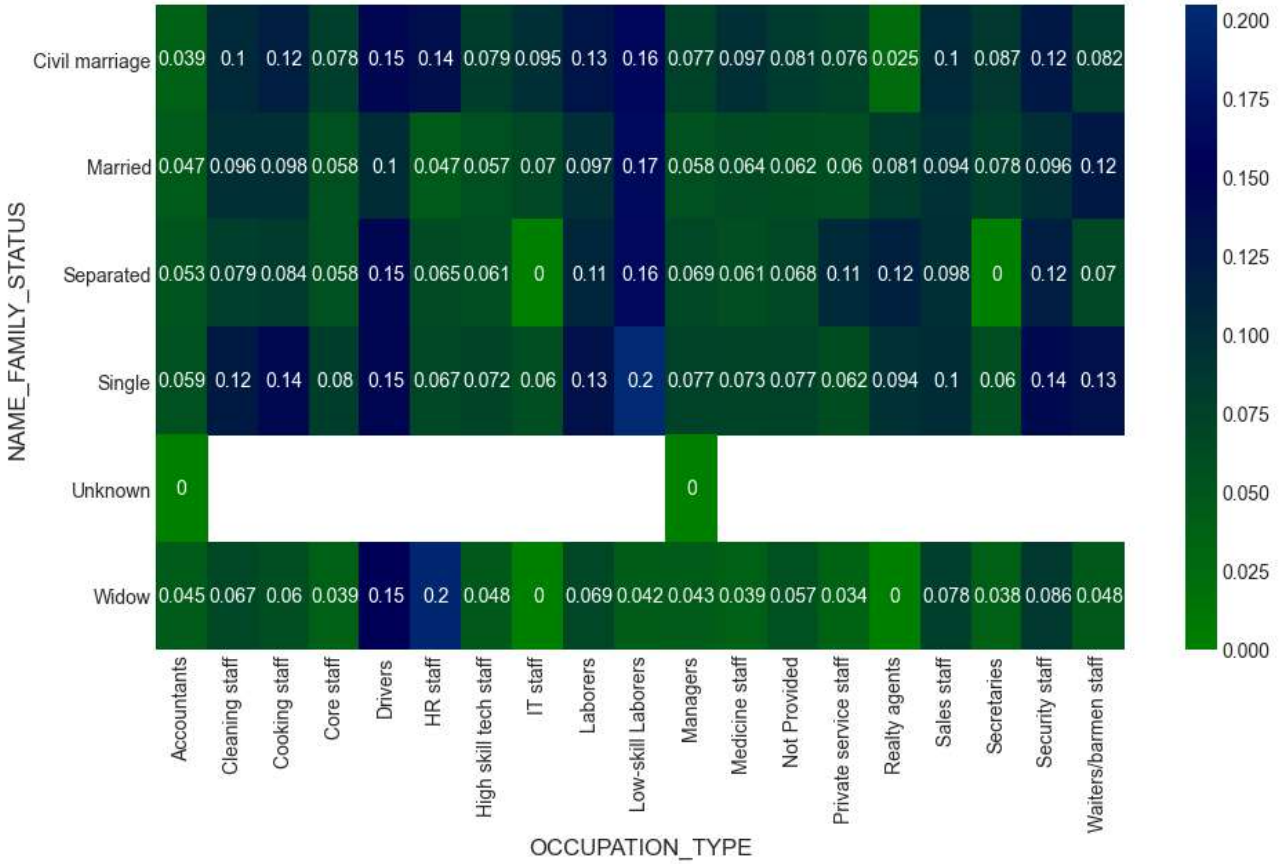


Almost 90% of applicant are in less than 300K income range

Defaulting percentage is higher among income in bracket 0-200K

Applicants with greater than 600K income are quite less likely to default

Bivariate Analysis – OCCUPATION_TYPE vs NAME_FAMILY_STATUS



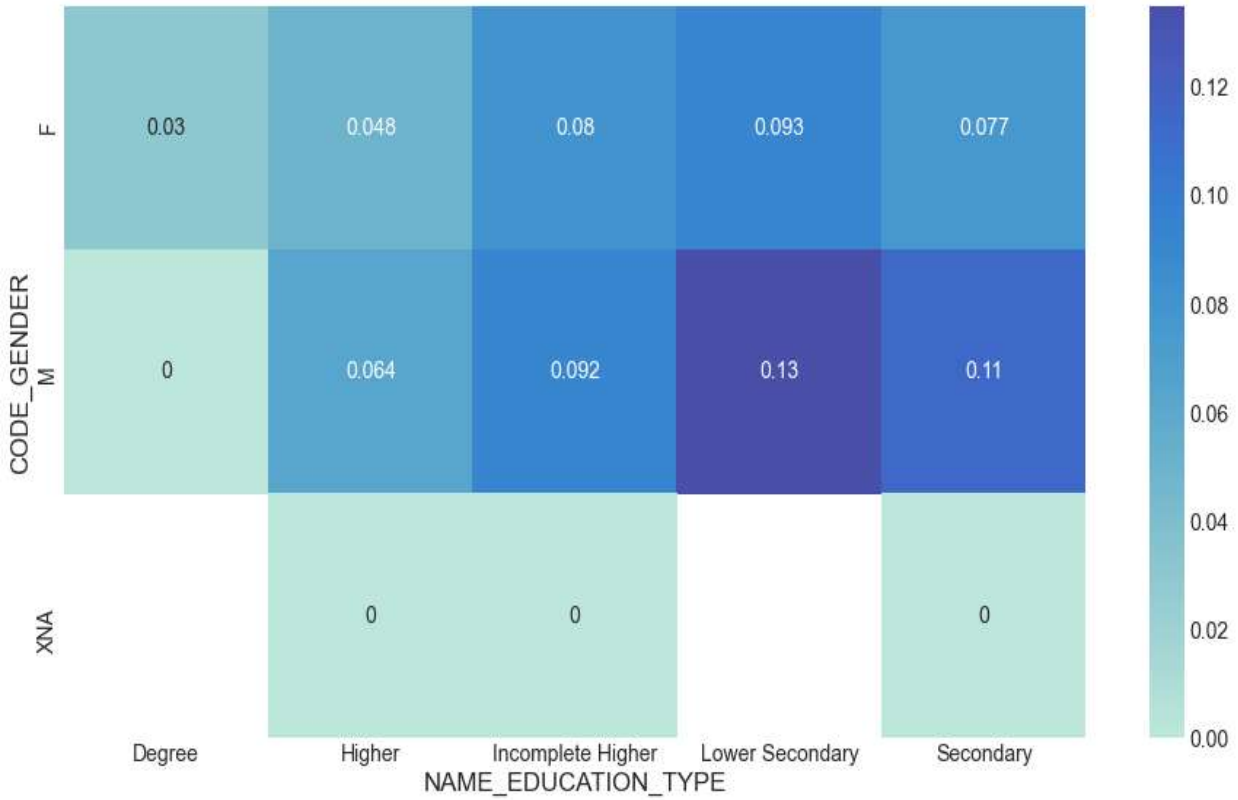
Drivers and Low-skill laborers irrespective of their family status are highest defaulters

While Widows tend to be payers but once in occupation of HR staff and driving tend to default more than others

On the contrary Civil married status default in mulitple occupation categories than the Married ones.

Married family status have strong tendency to be Payers

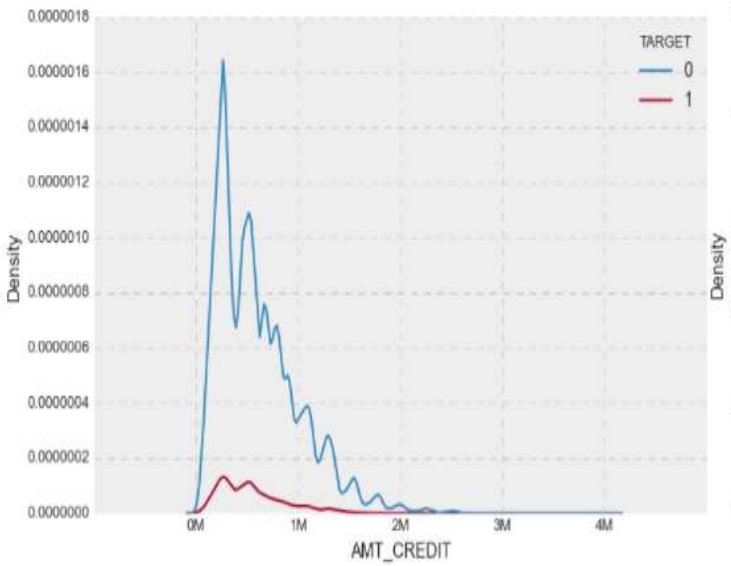
Bivariate Analysis – NAME_EDUCATION_TYPE vs CODE_GENDER

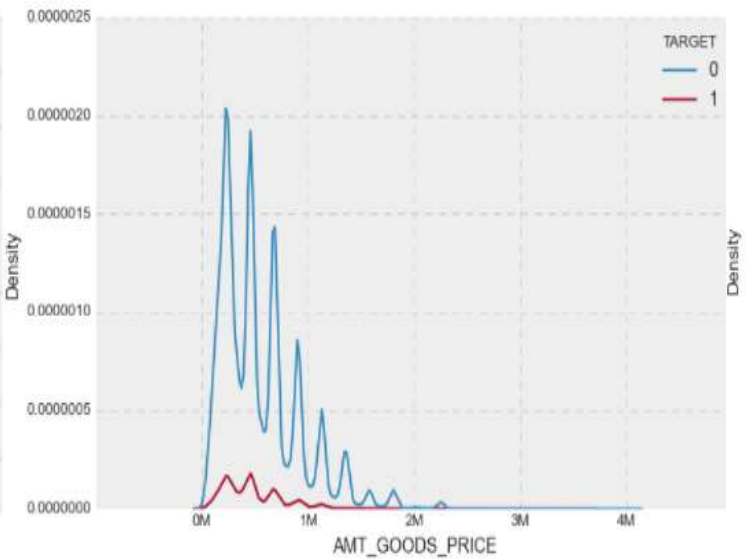


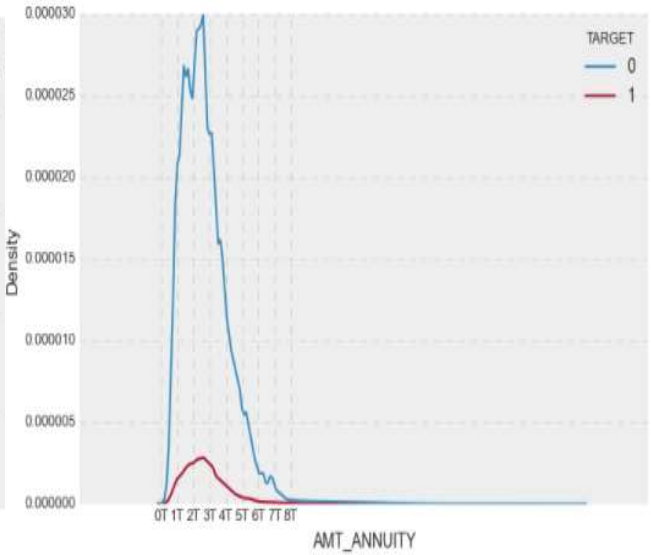
Evidently males who are Lower Secondary and Secondary qualification have tendency to default than others

Both males and females with Degree are Payers and do not default

Numeric Univariate Analysis – Amounts

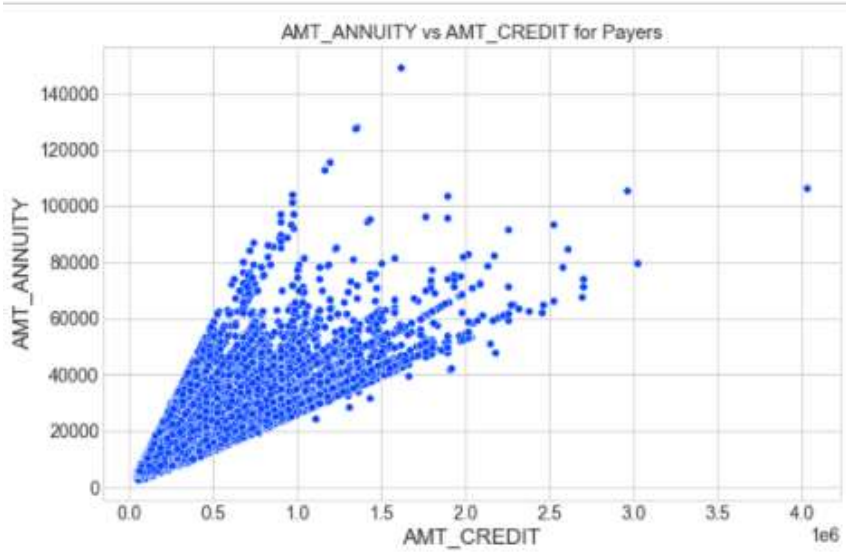


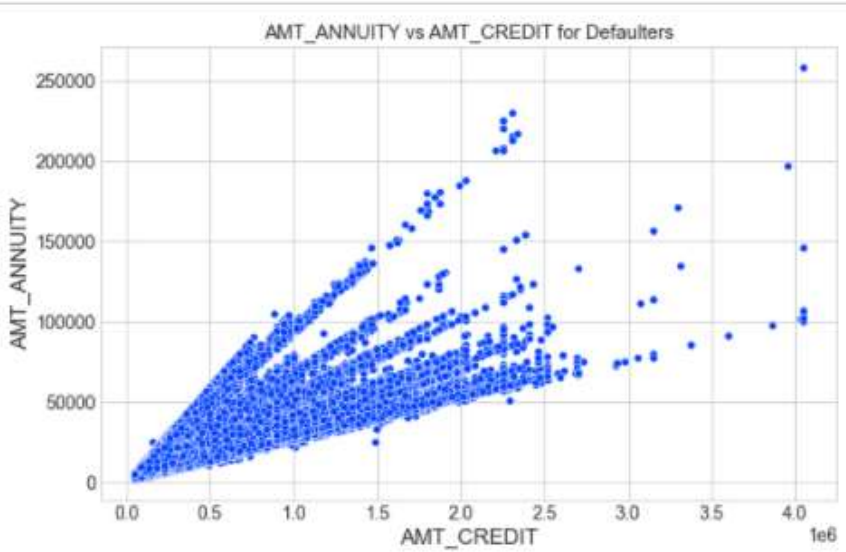




Maximum number of Credits are below 1M or 10L
We see similar trend with GOODs price, most goods price are below 10L
While defaulter density is less for both CREDIT and GOOD Price but their trend looks alike, so we can conclude that these amounts do not have any influence on DEFAULT rate

Numeric Bivariate Analysis – Amounts





We can conclude that when the AMT_CREDIT cores beyond 2M we see rise in defaulters.

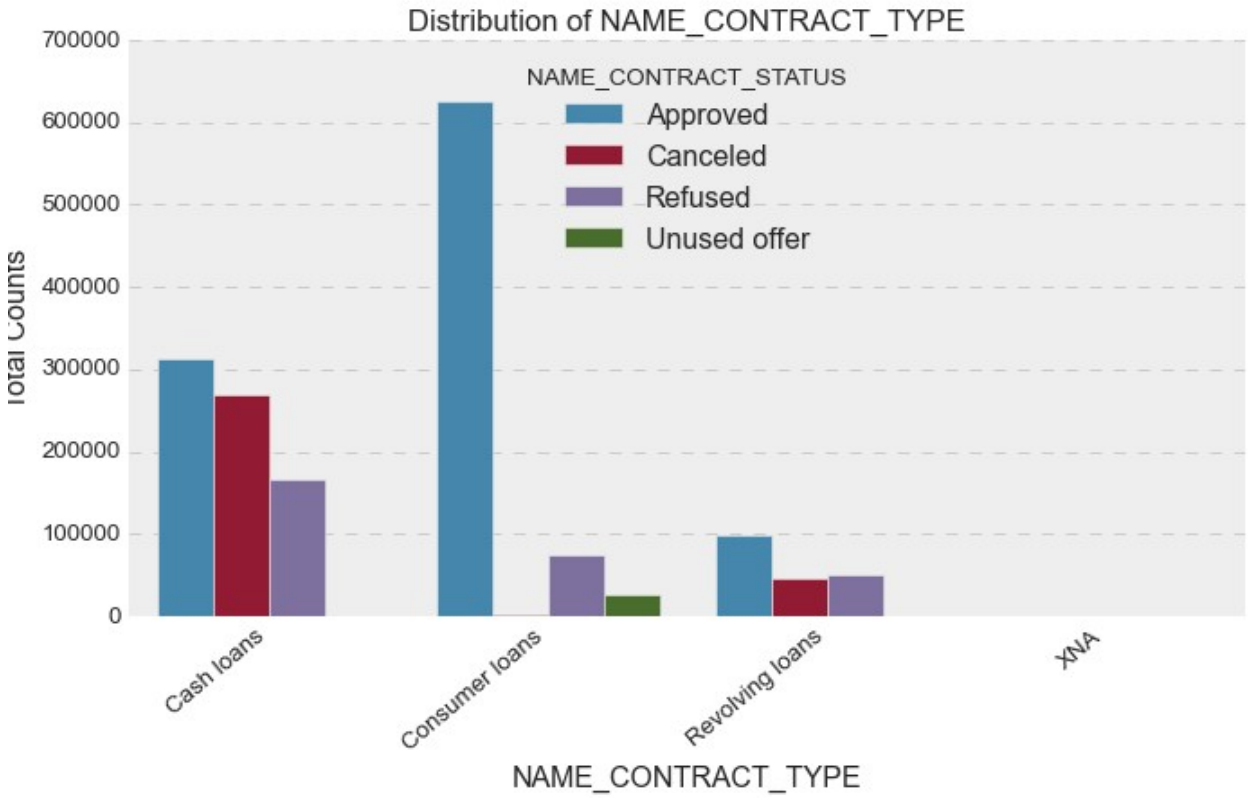
It is evident that AMT_CREDIT is directly related to AMT_ANUUITY

Univariate Analysis – Previous_application.csv

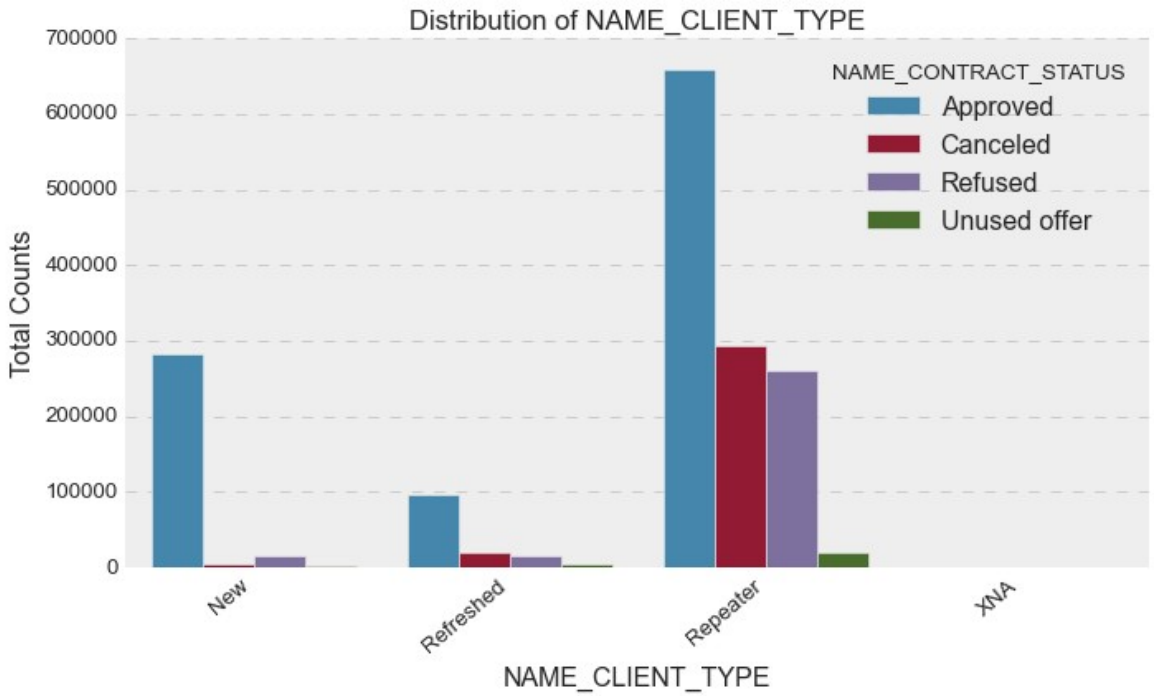
Clearly Consumer loans and Cashloans were most requested type of loans

While we see that Consumer loans has 100% approval rate but we see that Cashloan does have rejections

.



Univariate Analysis – Previous_application.csv



Repeater count is the highest among the type of clients, it amounts to almost 70%

New applications come next to the Repeaters

We see highest refusal rate amount Repeaters

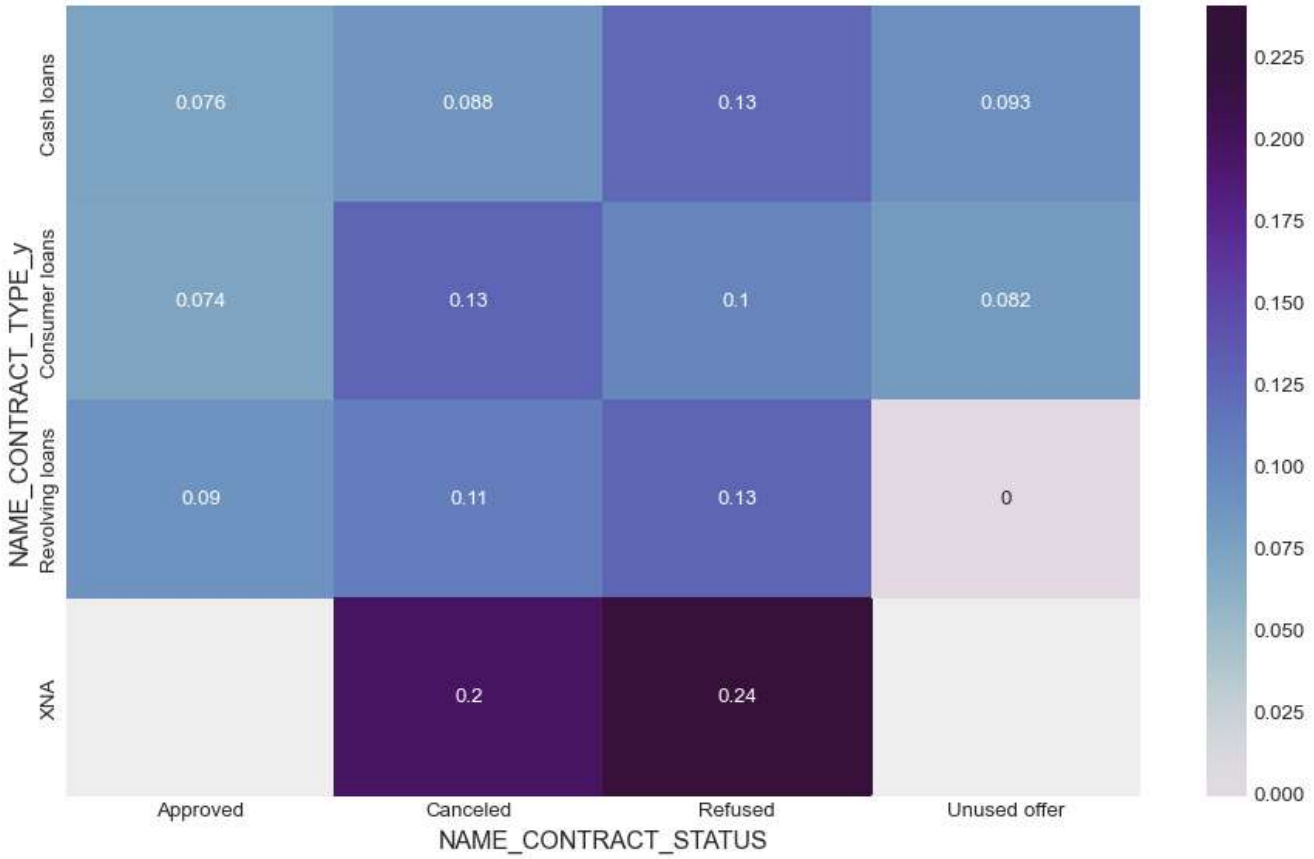
.

Combined Dataset Analysis

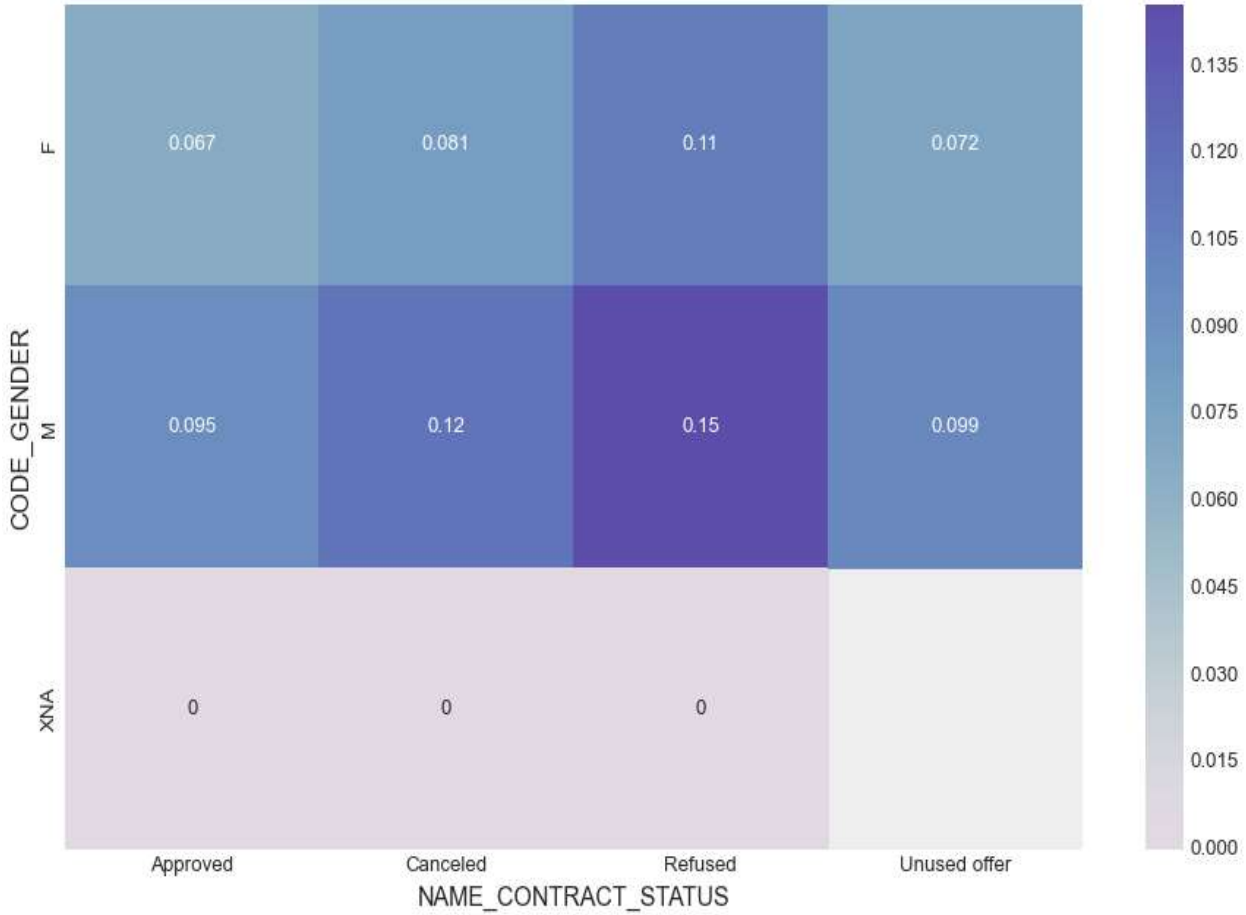
Correlation between NAME_CONTRACT_TYPE and NAME_CONTRACT_STATUS

Defaulting % is higher for people who were refused revolving or cash loans

Overall, we see that people for whom loans were refused have higher chances to default



Combined Dataset Analysis



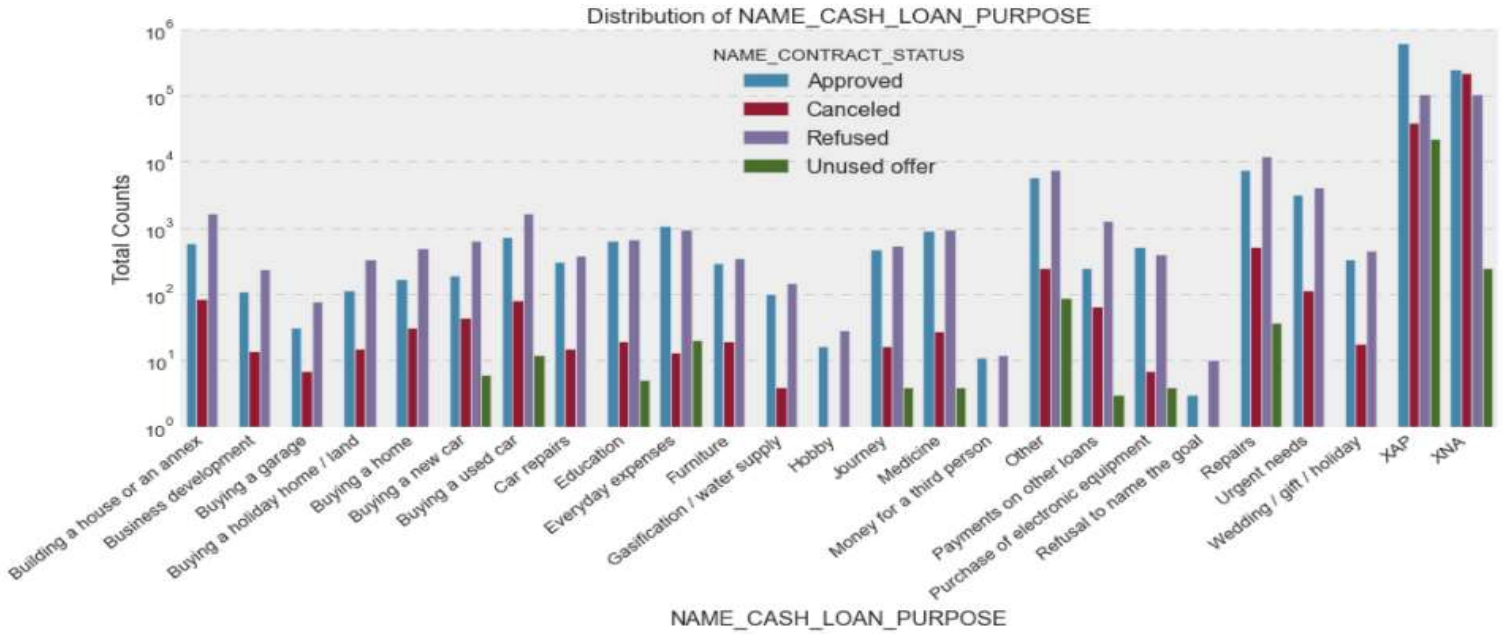
Correlation between GENDER and NAME_CONTRACT_STATUS

As seen earlier that female have lesser chances of default compared to males. This reflects even from the previous application.

We do not see any bias from financial institutes based on gender when it comes to approval or rejection of loans

Like seen earlier rejected applicants in past have chances to default

Combined Dataset Analysis



Clearly Loan purpose are considerable high for unknown values like XAP and SNA

There is large number of rejections for loans taken against Repairs and Other Category apart from XAP and XNA

It appears that financial institutes find high risk in lending loans for REPAIR purposes

TOP 10 Correlations in Defaulters and Payers

- Top 10 correlation variables for defaulters and payers are in almost same range 0.29 to 0.99
- When comparing defaulters with payers, we see correlation between total income and the credit amount significantly dropped from 0.64 for Payers to 0.0028 for defaulters
- For both Defaulters and Payers Credit Amount and Good Prices are highly correlated
- DAYS_BIRTH and CNT_CHILDREN correlation dropped for defaulters to 0.59 from Payers which was at 0.65
- We see possible correlation between Social Surroundings variables for both defaulters and Payers

Variable1	Variable 2	Defaulter - Correlation	Payer- Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.99827	0.99827
AMT_GOODS_PRICE	AMT_CREDIT	0.982783	0.982783
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484	0.885484
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.869016	0.869016
AMT_GOODS_PRICE	AMT_ANNUITY	0.752295	0.752295
AMT_ANNUITY	AMT_CREDIT	0.752195	0.752195
DAYS_EMPLOYED	DAYS_BIRTH	0.582185	0.582185
OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.337389	0.337389
DEF_30_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.334035	0.334035
DAYS_REGISTRATION	DAYS_BIRTH	0.289114	0.289114

Top Inferences from EDA

PAYERS

NAME_FAMILY_STATUS: Single Family status persons are better Payers than others
CNT_FAM_MEMBERS: Families with 3 or less members are
EMPLOYMENT_EXP: People with greater than 15 years experience have the least defaulting percentage
RANGE_INCOME: Applicants with greater than 600K income are quite less likely to default
NAME_EDUCATION_TYPE: Applicant who have DEGREE are best Payers

DEFAULTERS

CODE_GENDER: Men tend to default more than women
FLAG_OWN_REALTY: Defaulting rate seemed to be higher in people with NO REALTY
ORGANIZATION_TYPE: Self Employed people seemingly have pretty high defaulting rate
OCCUPATION_TYPE: Laborers group that has defaulted higher in comparison to others
NAME_EDUCATION_TYPE: Lower Secondary educated customers have highest defaulting percentage
AMT_CREDIT: When the AMT_CREDIT cores beyond 2M we see rise in defaulters

Thank You