Credit Case Study EDA

Loan Defaulters

Performed by:
- Mohammed Hussain Chitapulla
- Sweta Singh

Problem Statement

The purpose of this case study is to provide a comprehensive research and identify patterns which indicate if a client has difficulty paying their installments, which may be further 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 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.

Analysis Approach

Data Preparation and Data Cleaning

- Data Inspection
- Missing value analysis
- Outlier Analysis
- Standardisation-sanity/quality checks
- Binning/Bucketing

Data Analysis

- Data Segmentation
- Univariate Analysis
- Bivariate Analysis
- Correlation Matrix
- Necessary mergings

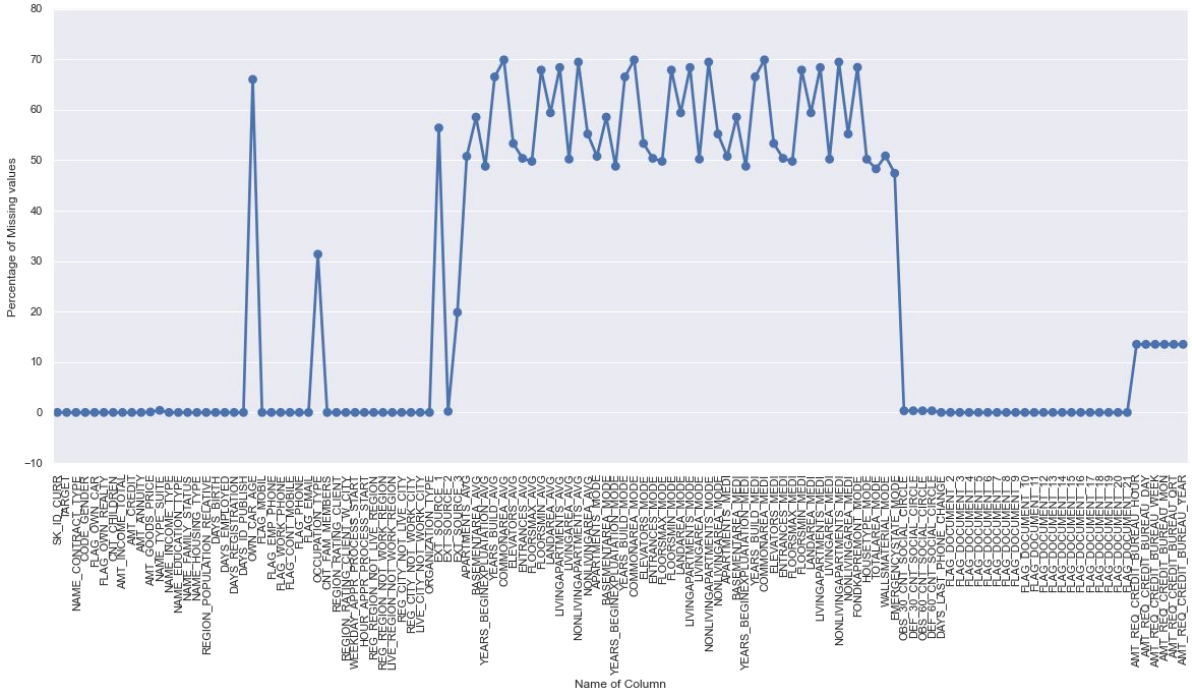
Final Inferences

- Inferences from univariate analysis
- Inferences from bivariate analysis
- Top correlations
- Combined Inferences

Data Cleaning

Missing Values

It can be observed from the graph below that a significant number of columns contain missing values.



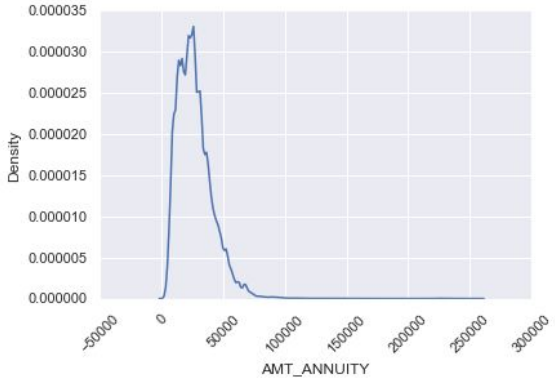
There are many columns in the dataset with greater than 50% of missing values.

Strategy used for imputation of Missing Values

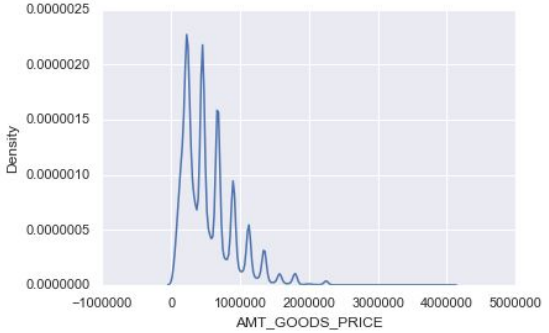
1. Dropped the columns with **more than 50%** missing values

2. **Less than 13% missing values** - Imputed the values
   2.1 Numerical Columns: imputed based on mean, median mode

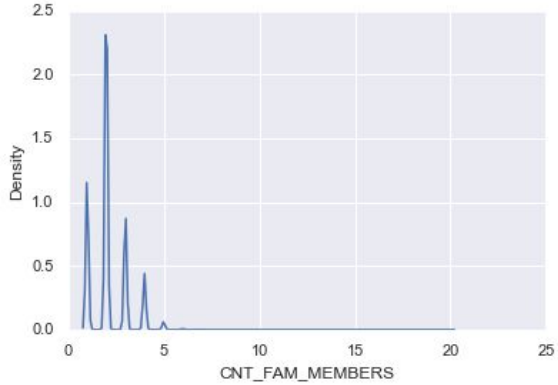
AMT_ANNUITY : The density distribution is skewed. Median method has been considered to impute the values



AMT_GOODS_PRICE : Mean method has been considered to impute the values



CNT_FAM_MEMBERS : Multiple peaks. Mode method has been used



2.2 Categorical Columns:

NAME_TYPE_SUITE : Mode method has been used to impute null values

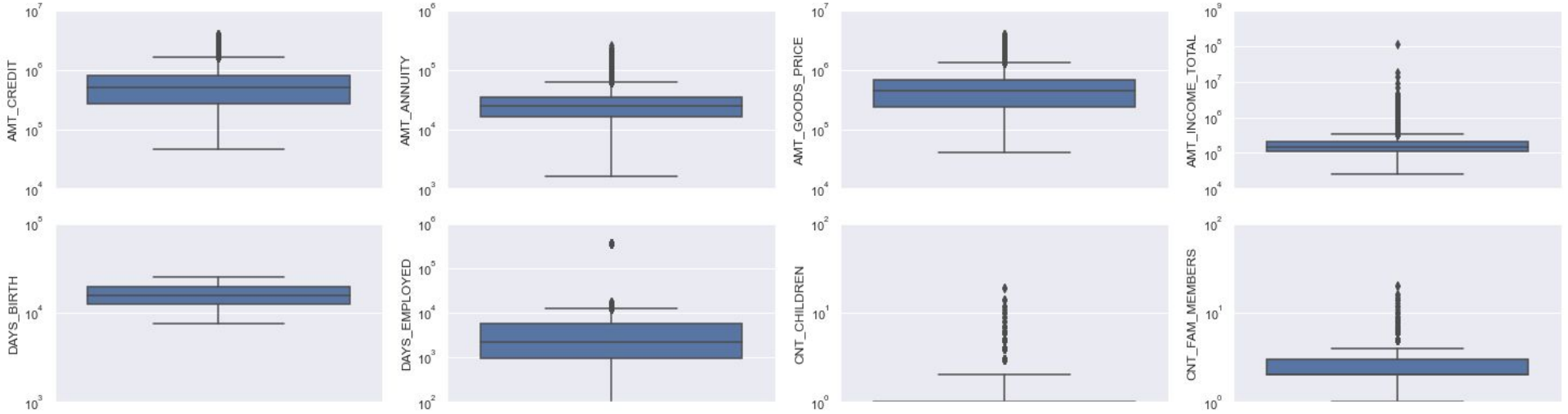
3. Missing values between 13% to 50%

OCCUPATION_TYPE. : consists of 31% missing values : New category “Unknown” created for missing values.

Outlier Analysis

Outlier analysis is performed on numerical columns using the box plots.

'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'AMT_INCOME_TOTAL', 'DAYS_BIRTH', 'DAYS_EMPLOYED','CNT_CHILDREN', 'CNT_FAM_MEMBERS'



Inferences from Outlier box plots

1. 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE' have almost continuous distribution.

2. 'DAYS_BIRTH' has no outliers gives the perfect information.

2. 'CNT_CHILDREN' and 'CNT_FAM_MEMBERS' have few outliers.

4. AMT_INCOME_TOTAL clearly have outliers which shows some loan applicants have incomes in high brackets as compared to others.

5. DAYS_EMPLOYED have exceptionally high value i.e. 365243 days which is around 1000 years. This is an impossible value and need to be imputed through capping.

Data Quality issues/ Sanity Checks

1. Hidden missing values, columns containing '**XNA**'

   1.1. CODE_GENDER - .0013% rows with XNA values : Imputed with Mode

   1.2. ORGANIZATION_TYPE - 18% rows with XNA : assigned to new category "Not Specified"1.

2. Negative values for days columns : DAYS_BIRTH, DAYS_EMPLOYED

   There can be Two scenario 1) +ve days refer to client employed in a organization -ve days refer to days the client left the organization 2) It could be issue at the data level

   In our case let's consider its a data issue lets replace all negative values to positive.

3. Derived "Age" from DAYS_BIRTH, 'Employment In Years' from 'DAYS_EMPLOYED'

Binning/bucketing of continuous variables

1. **Age_Group**: converted to ordered categorical by dividing to the age group bins '<18','18-25','25-35','35-45','45-60','60+"
2. **Work_Group**: converted to ordered categorical by dividing to the age group bins '<18','18-25','25-35','35-45','45-60','60+"
3. **AMT_CREDIT** continuous variable categorized into buckets

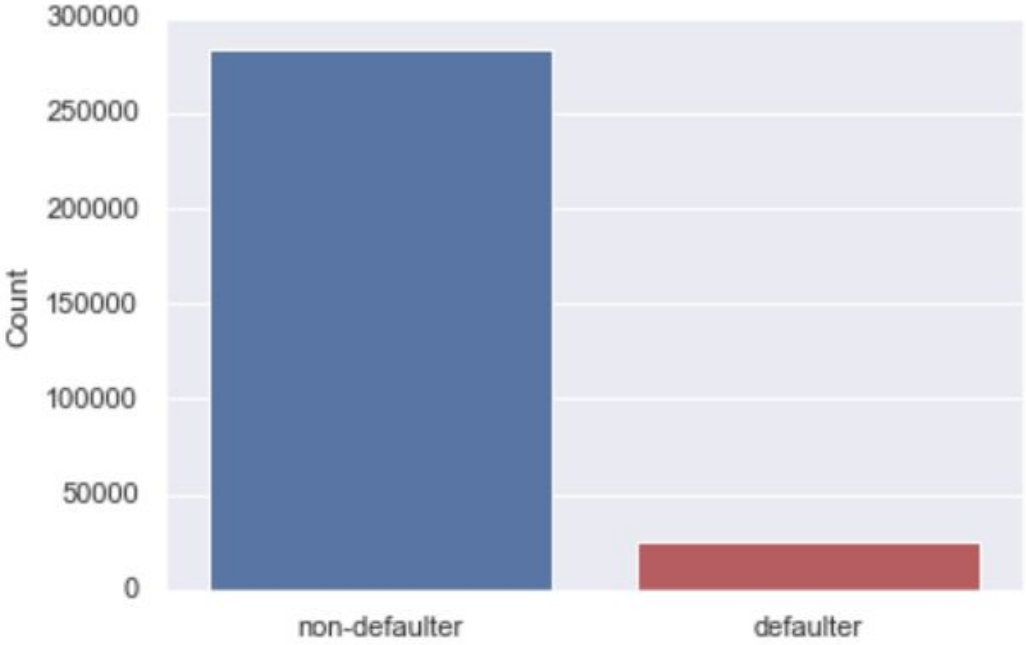
   '<25K', '25K-50K', '50K-1L' ,'1L-2.5L' ,'2.5L-5L' ,'5L-10L' ,'10L-25L' ,'25L-50L' ,'50L<'

4. **AMT_INCOME_TOTAL:** divided into income brackets

   '<25K','25K-50K','50K-1L','1L-2.5L','2.5L-5L','5L-10L','10L-25L','25L-50L','50L<'

Imbalance Percentage

Data imbalance is checked based on the “TARGET” column, where “1” means the person is defaulter unable to repay the loan, 0 is non-defaulter.



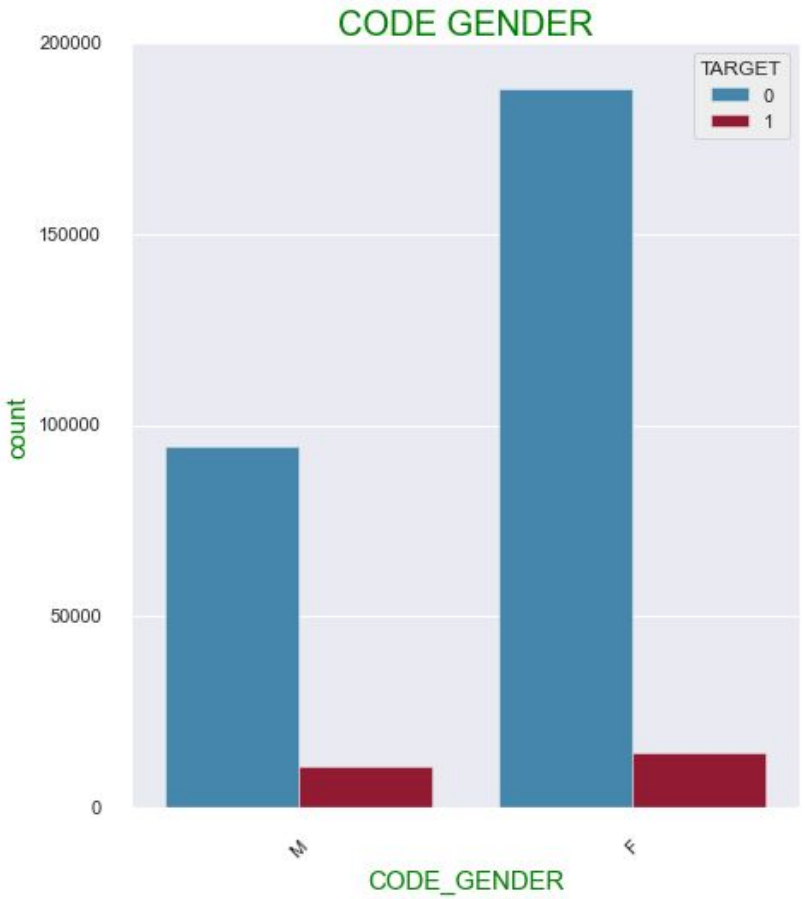
1. Ratios of imbalance in percentage with respect to non-defaulter and Defaulter datas are: **91.93%** and **8.07%**
2. Ratios of imbalance in relative with respect to non-defaulter and Defaulter datas is **11.39** (approx)
3. Data is divided into two datasets for “TARGET = 1” and “TARGET=0”

Data Analysis

Univariate Analysis on Segmented Data

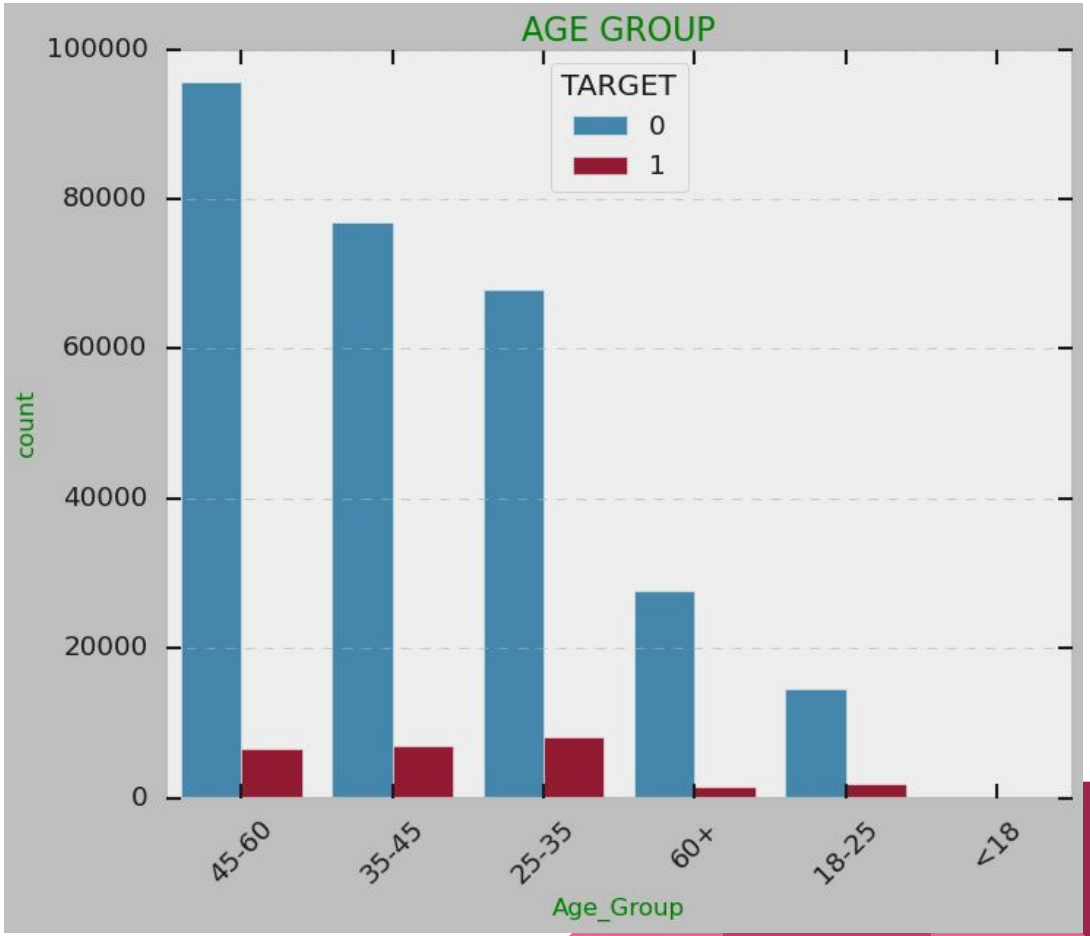
GENDER

1. The number of female clients is almost double the number of male clients.
2. Male clients have higher chances of defaulting than females



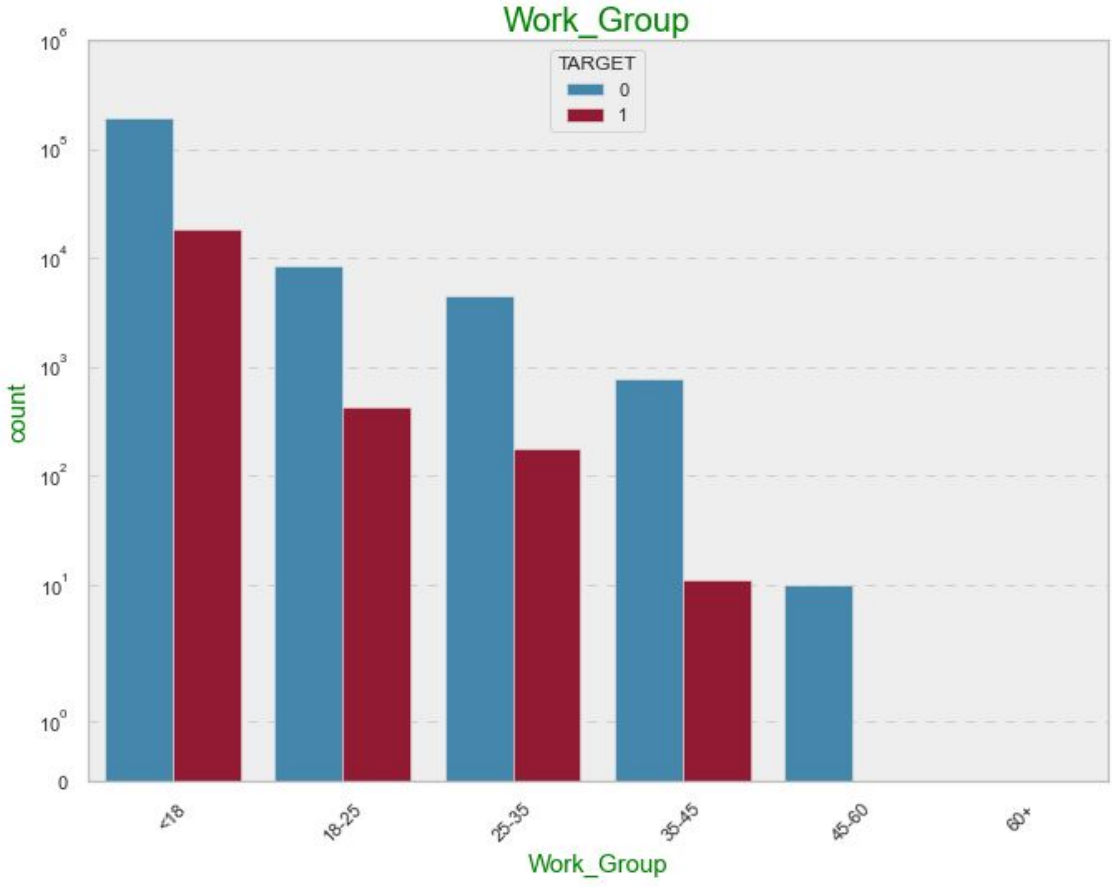
AGE GROUP

1. Most loan takers seems to be of age group 45 to 60.
2. People of age 25-45, have high chances of defaulting.

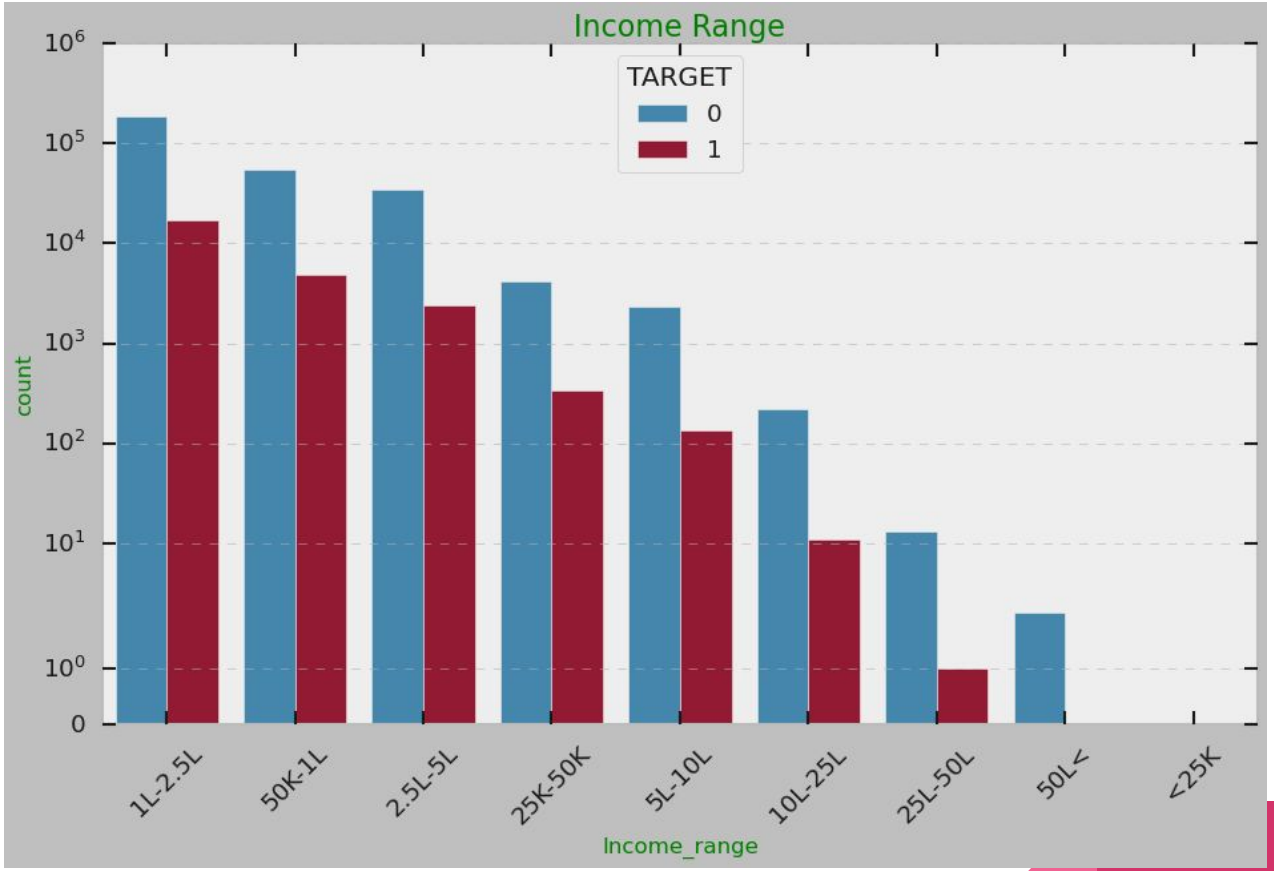


Employment In Years

1. Customer with more than 45 years of experience has no defaulters
2. Customer with less work experience has high defaulters

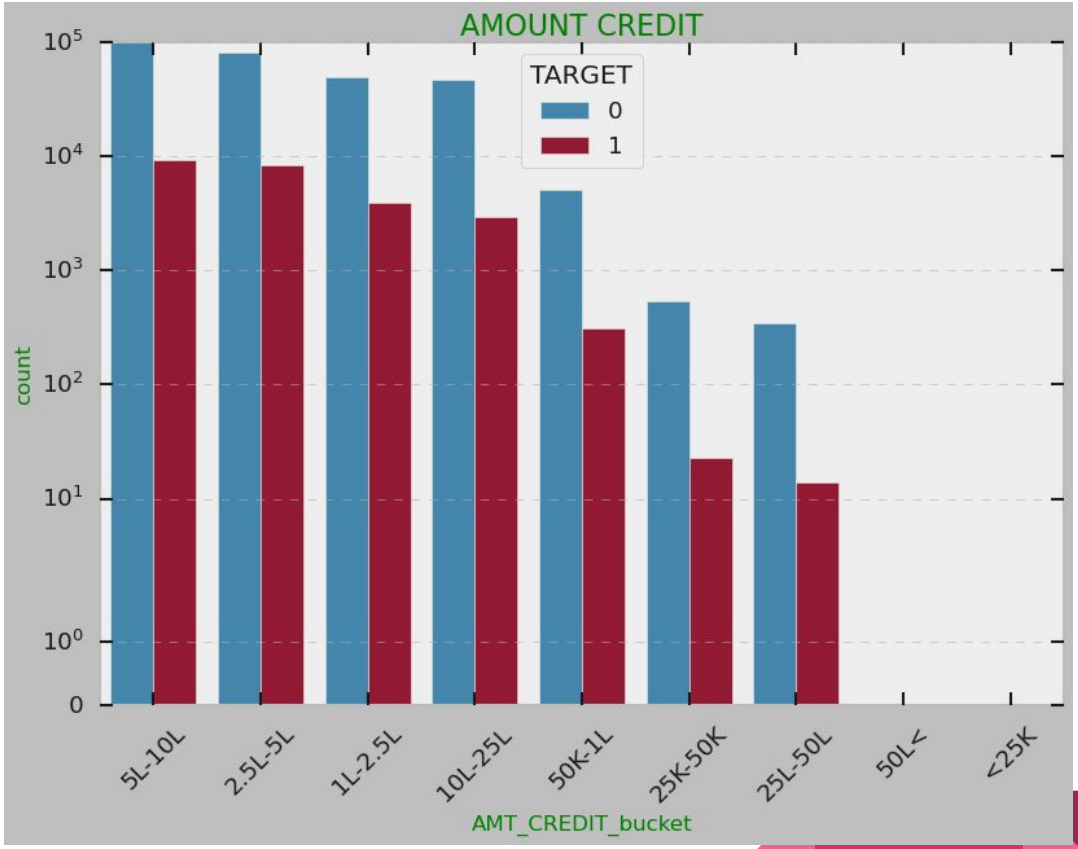


Income Range



1. Maximum number of applications have income less than 2.5 lakh.
2. People with income less than 2.5 lakh have more chances of defaulting.
3. Applicant with income more than 5 lakhs less likely to default.

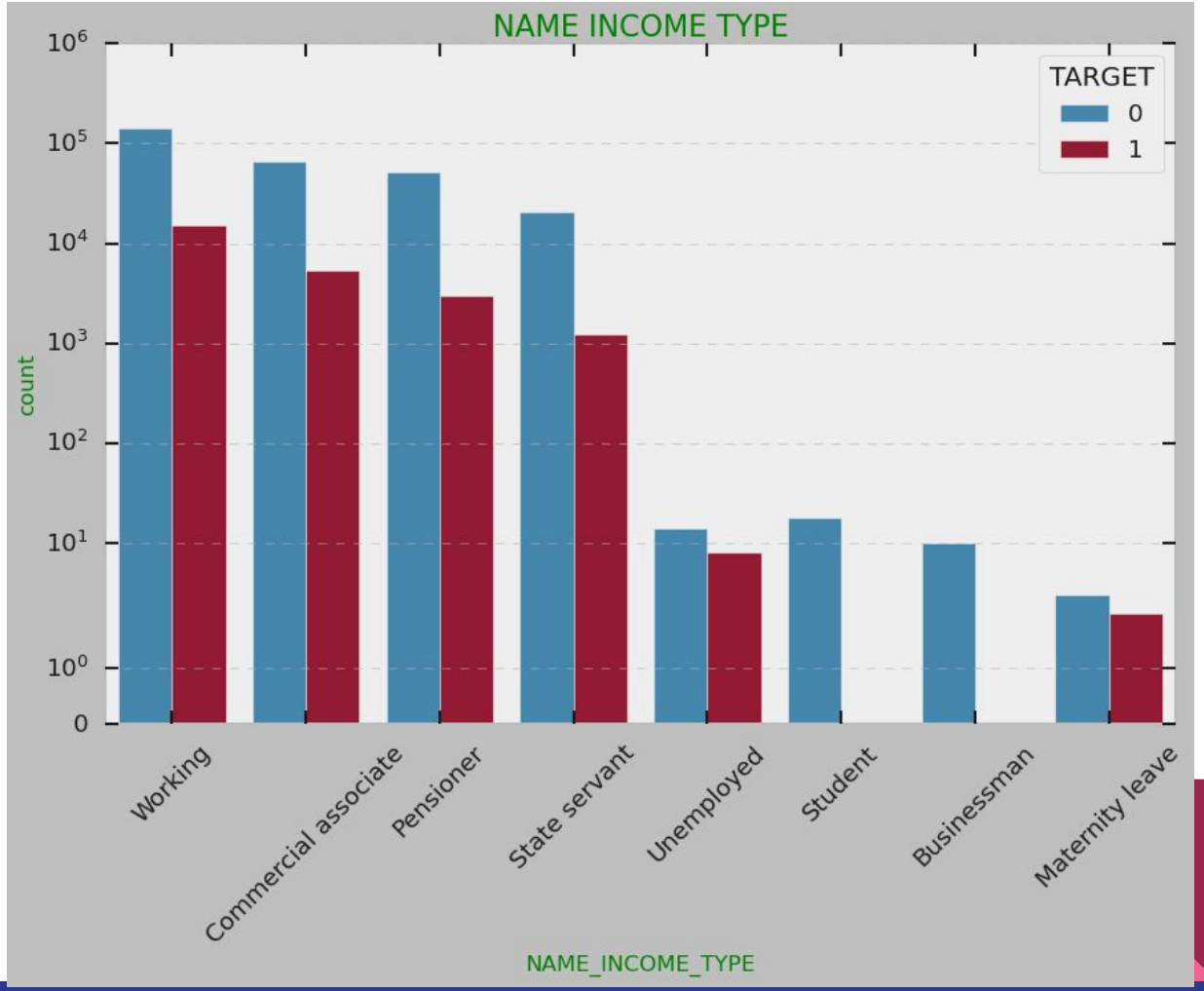
Account Credit Bucket



1. Most of the applicants got credit of less than 10 lakhs.
2. People with in credit range 2.5 lakhs to 10 lakhs tends to default more.

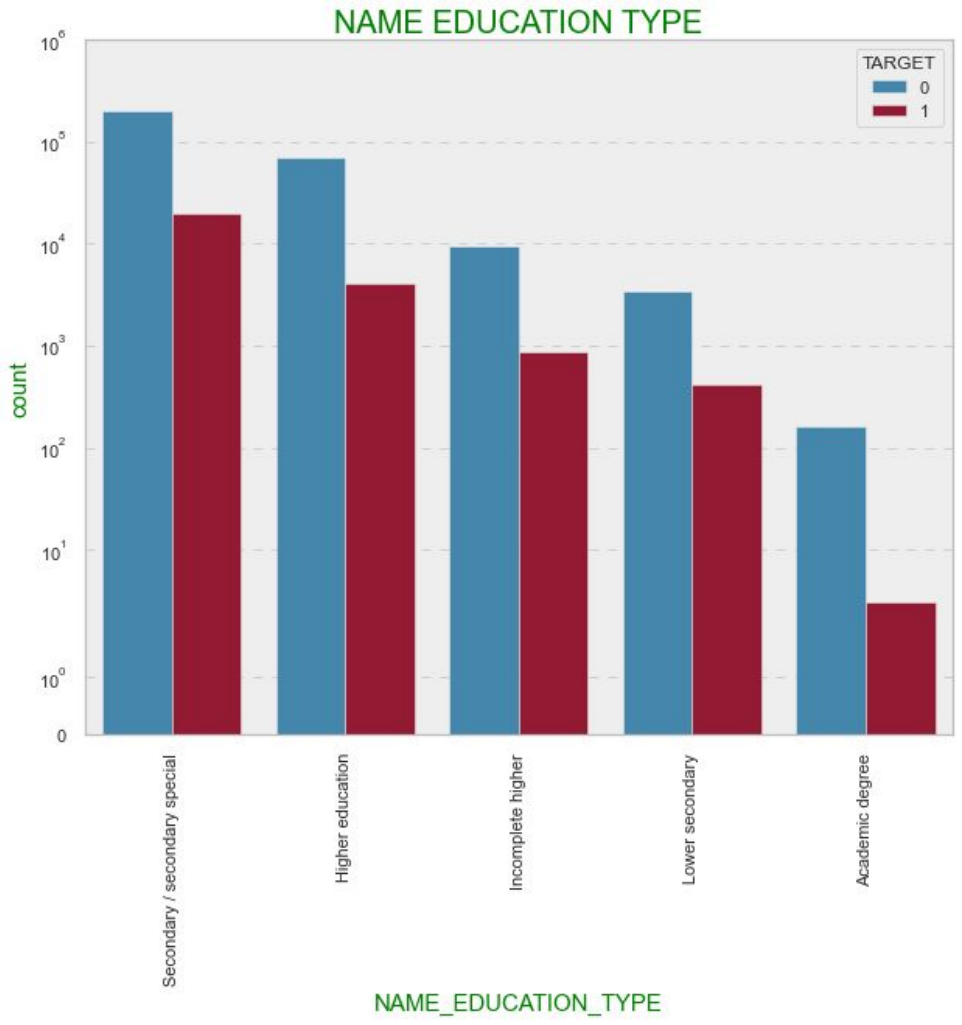
Income Type

1. Most of the applicants of loan have income type Working followed by Commercial associates then Prisoner and State servant.
2. Student and Businessman seems to be the safest categories with no default rates.
3. People in category Unemployed and Maternity leave have the highest chances of defaulting.



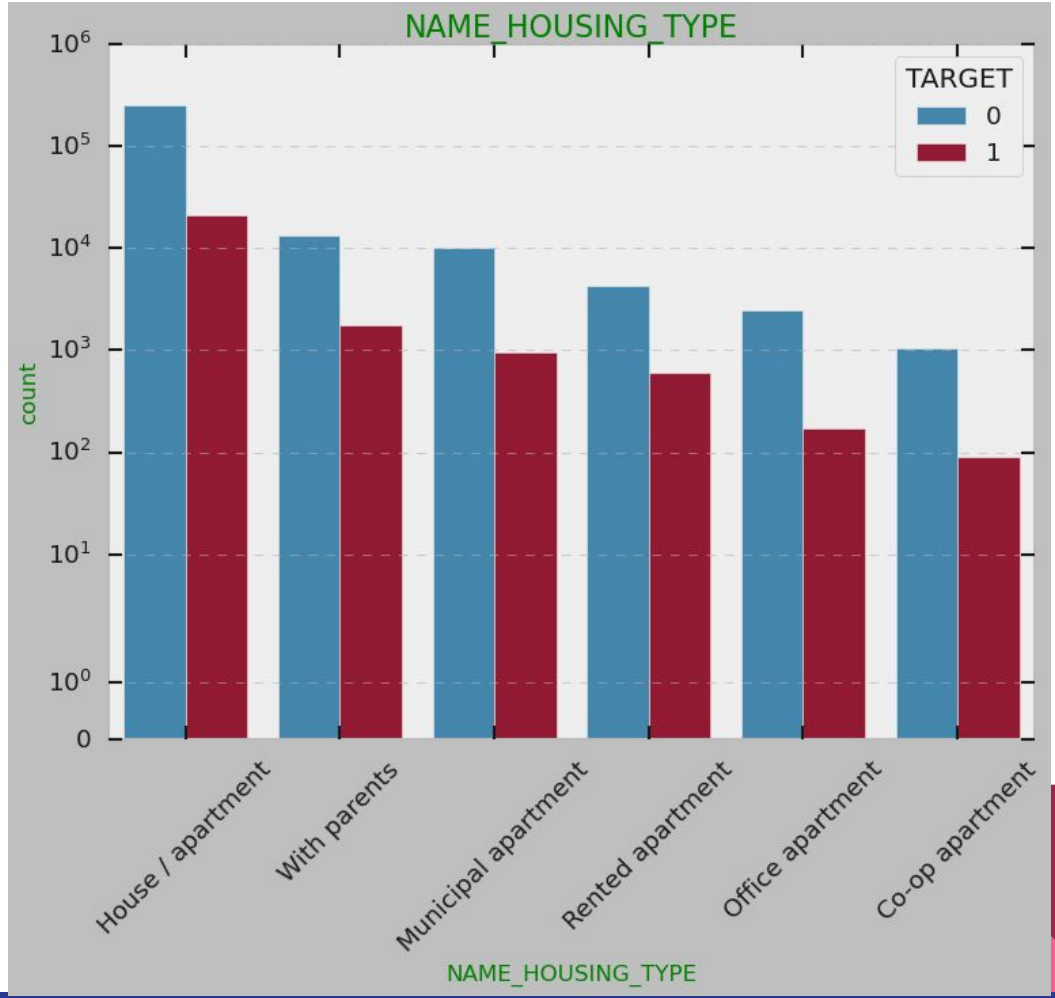
Education Type

1. Most of the applicants comes from category Secondary/ secondary special and higher education.
2. Academic degree though have less number of clients but have relatively low number of defaults.



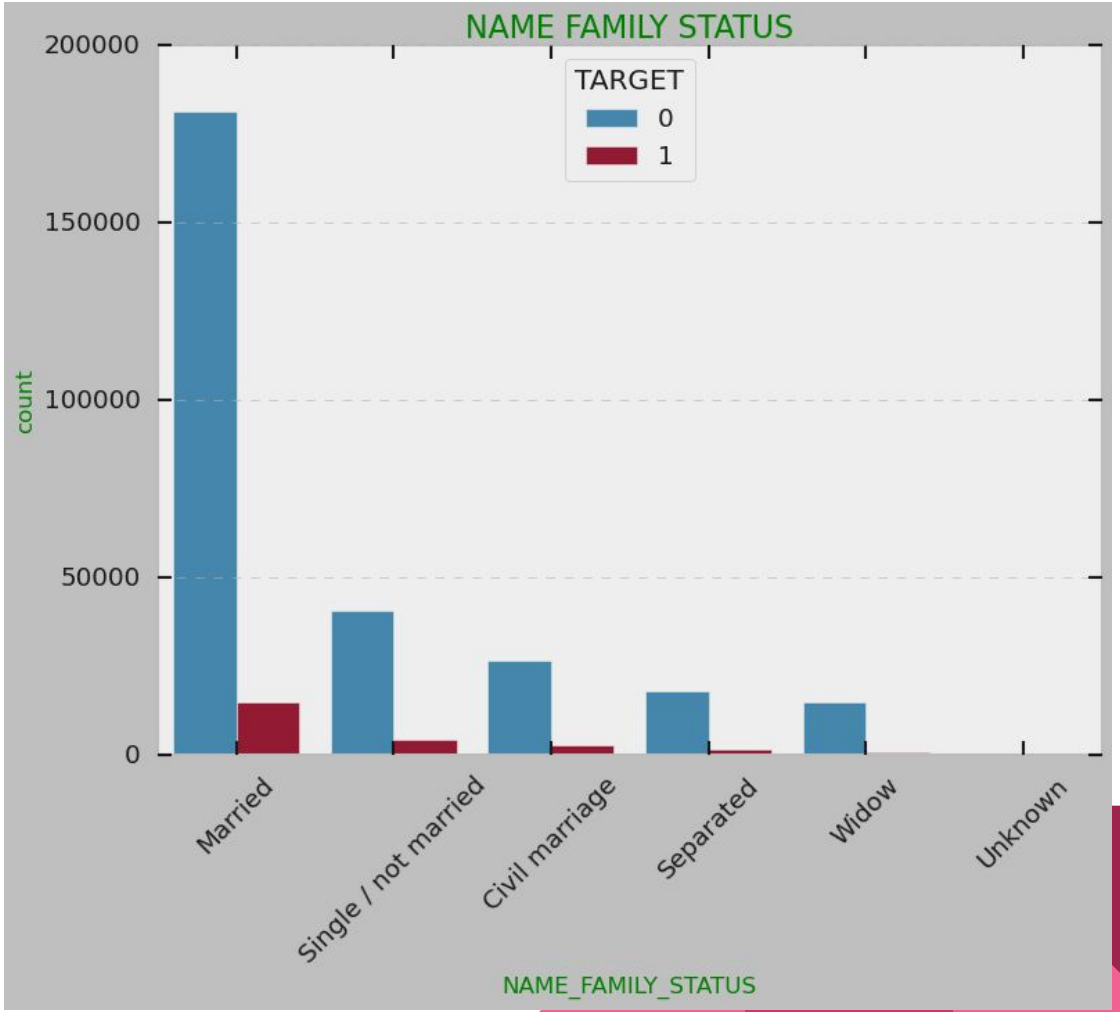
Housing Type

1. Most of the applicants comes from category Secondary/ secondary special and higher education.
2. Academic degree though have less number of clients but have relatively low number of defaults.

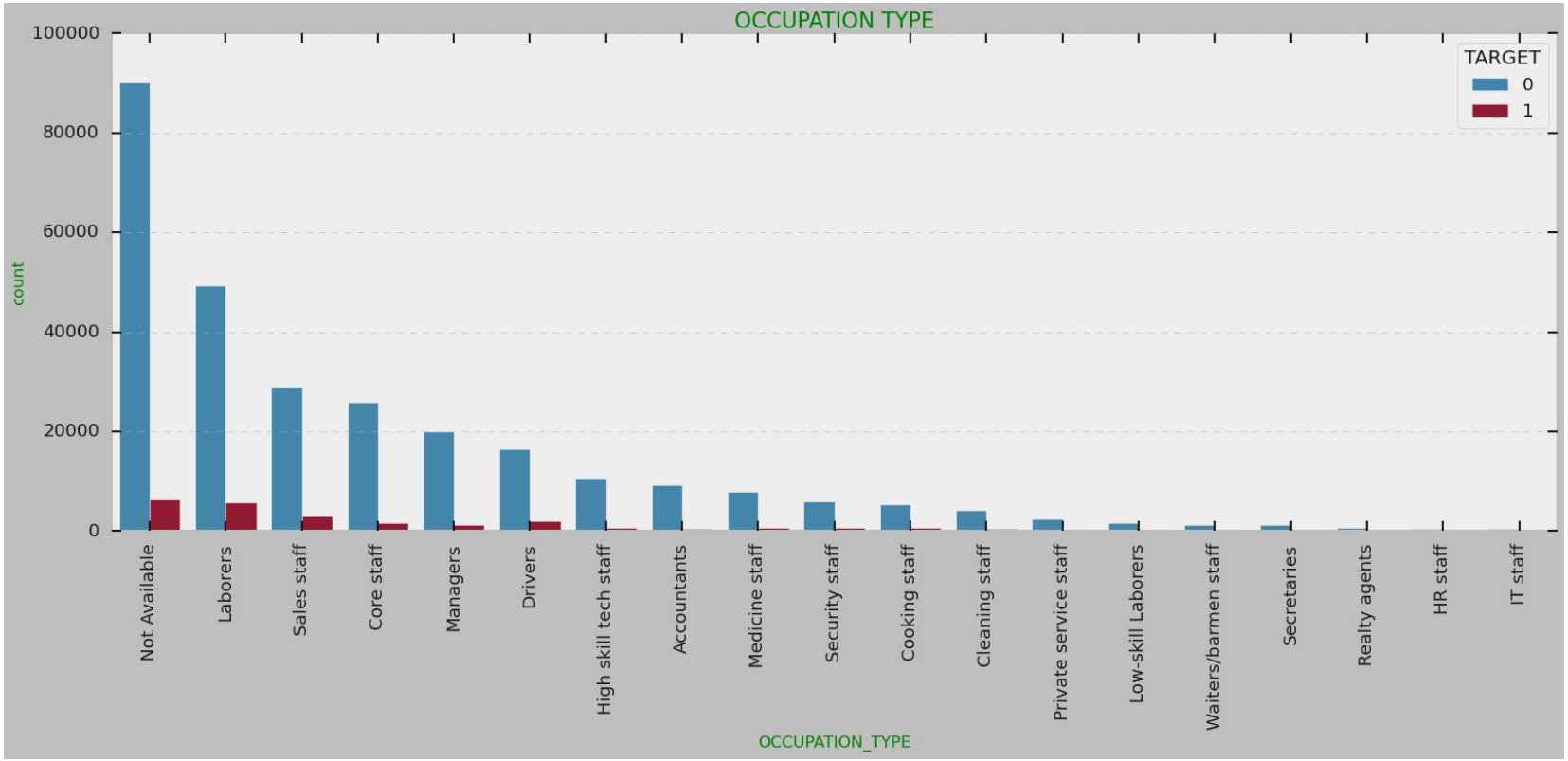


Family Status

1. Most of the loan applicants are married followed by single/not married.
2. Married and widow seems less likely to default.

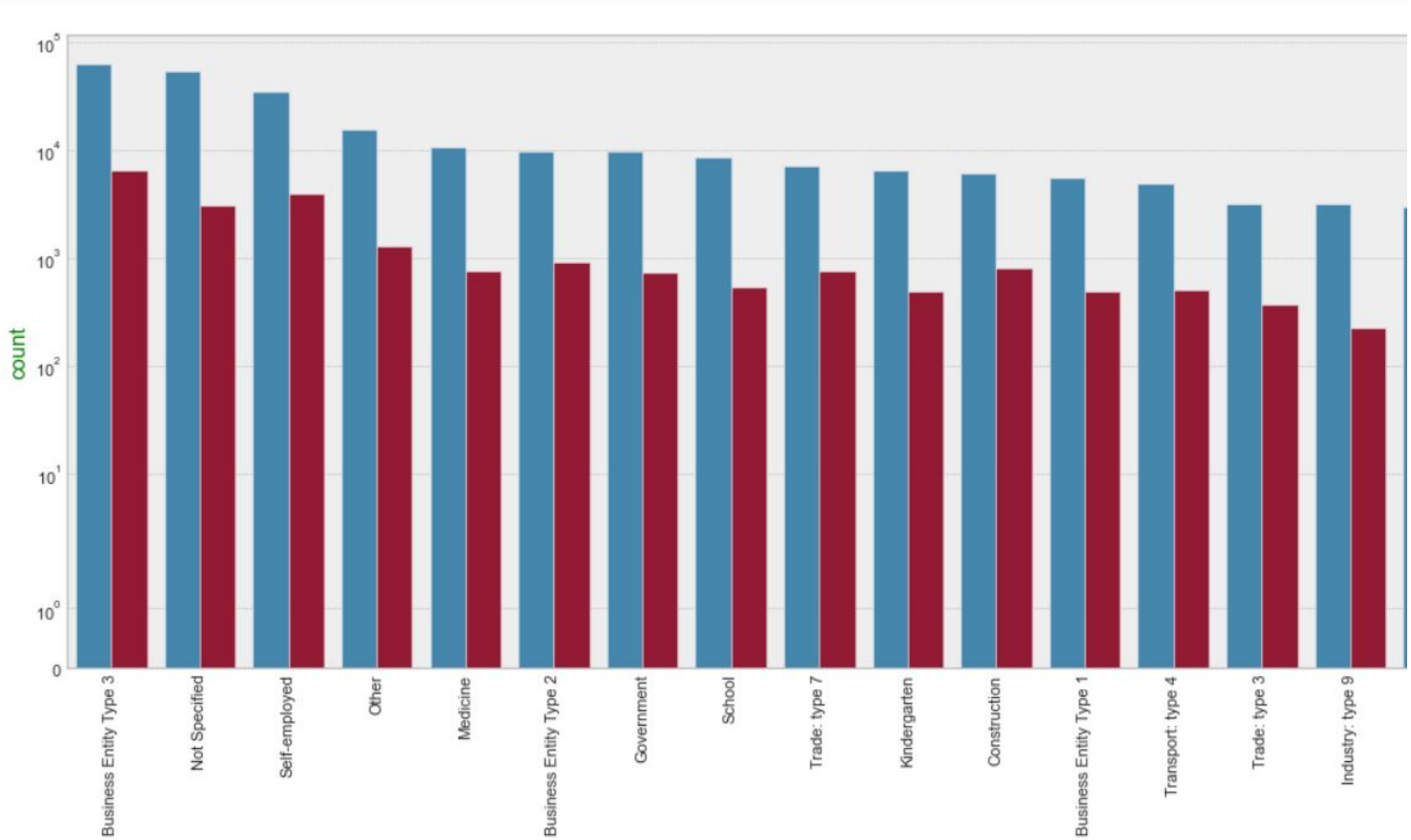


Occupation Type

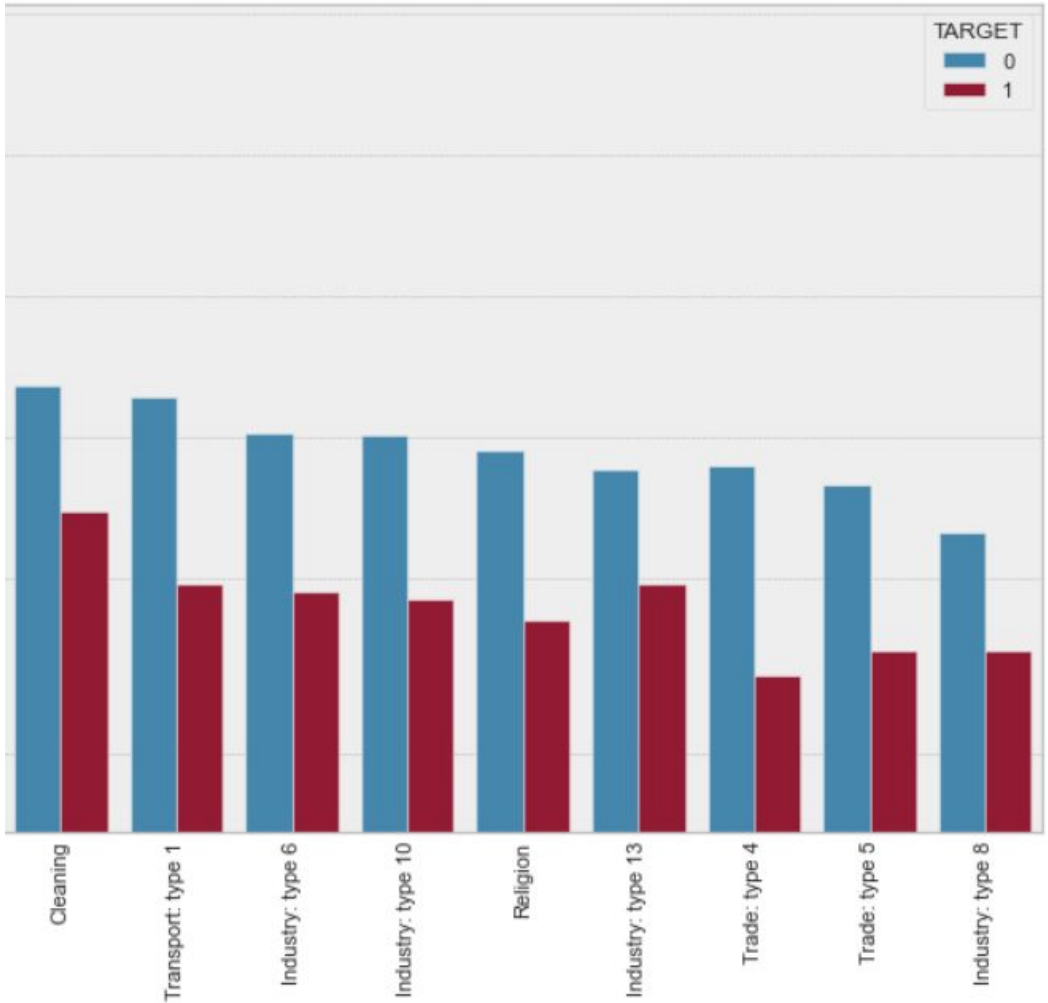


Most of the loan taking clients are laborers followed by sales staff then core staff. IT staff and HR staff seems to take lowest amount of loans.

Organization Type



1. Most of clients are from category **Business Entity Type 3**

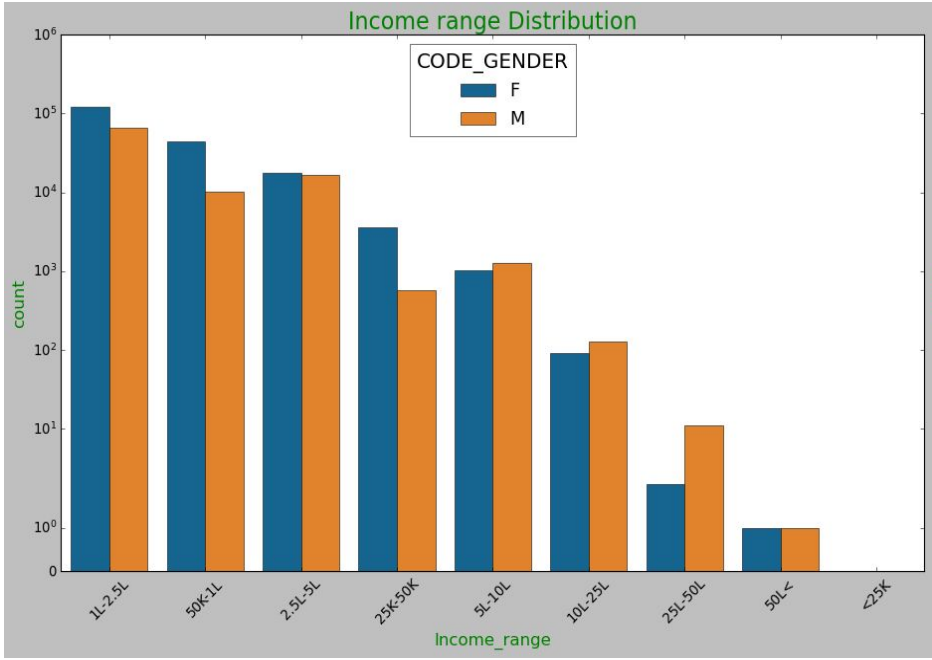


2. Trade : type 4, Trade: type 5 seems to have relatively less defaulters.

Bi/Multivariate Analysis

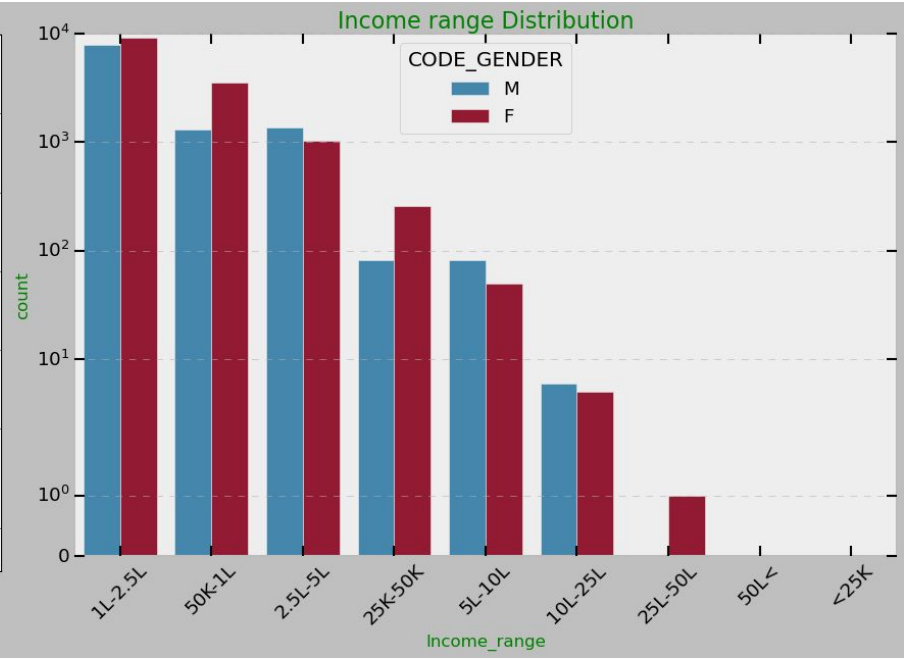
Categorical-Categorical: **Income bucket with CODE_GENDER**

Target 0: Non defaulter



1. Maximum of customer have an income range of 25K to 5L
2. Females count is higher when compared to male in the income range of 25K to 5L
3. In higher Income bracket females counts is lower than the males

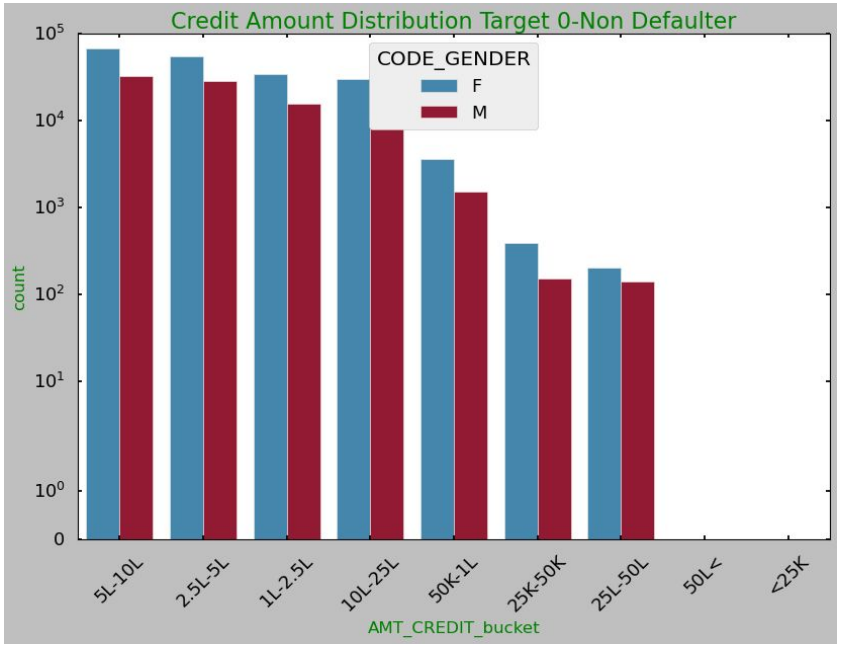
Target 1: defaulter



1. Income range count for female is higher in the 25k to 2.5L
2. Income range count for males are higher in the range from 2.5L to 25L
3. In the come bucket 25L -50 L there are no male Defaulter

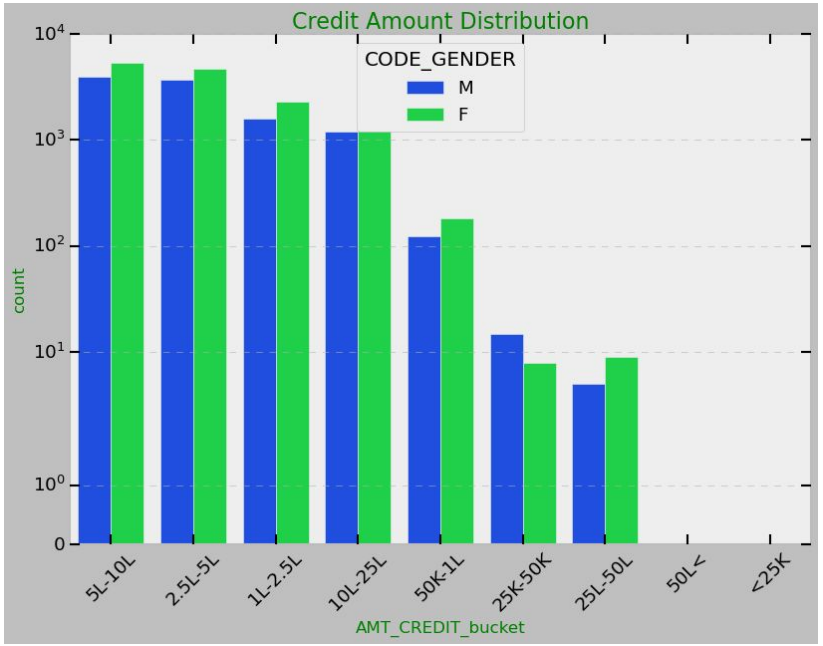
Categorical-Categorical: **Credit Amount Distribution with gender**

Target 0: Non defaulter



1. From the plot it is clear that females prefer credit more than males
2. Credit bucket 5L-10L has the highest no of applicant

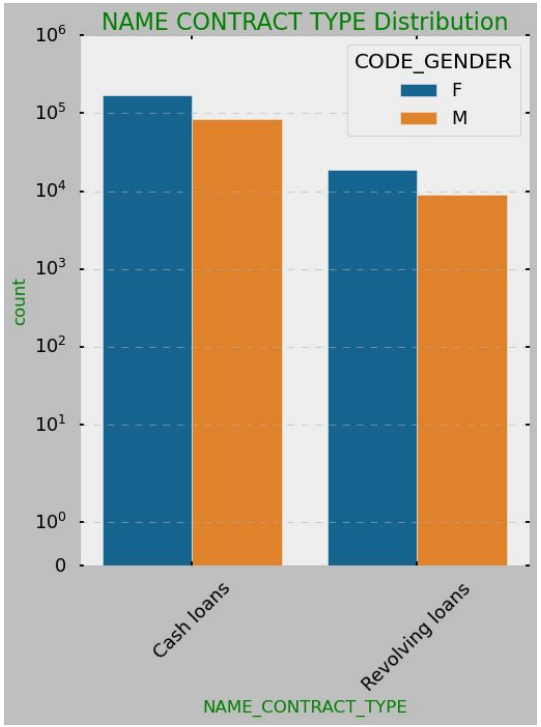
Target 1: defaulter



1. Credit count is higher for males in the lower bracket 25K to 50K
2. In all other bucket Credit count of females is higher than the males
3. Credit Range of 5L-10L has the highest number of credit defaulter

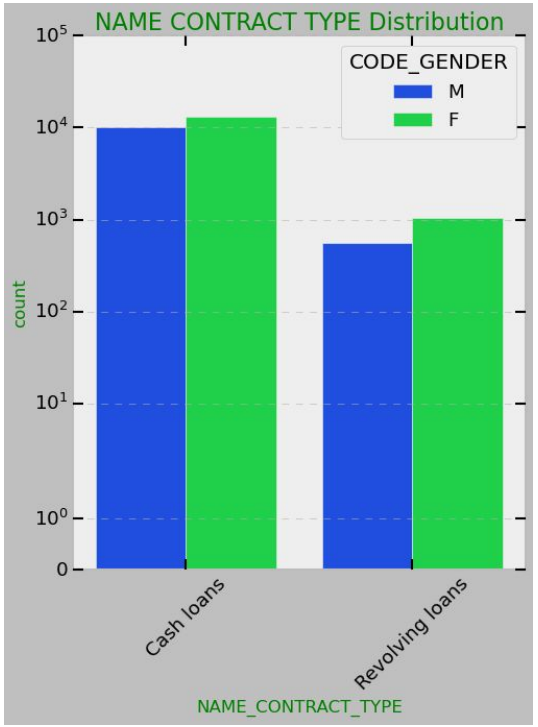
Categorical-Categorical: **Name Contract Type with gender**

Target 0: Non defaulter



1. Cash loan is more popular than the Revolving loans
2. Females count is higher in both the loan types

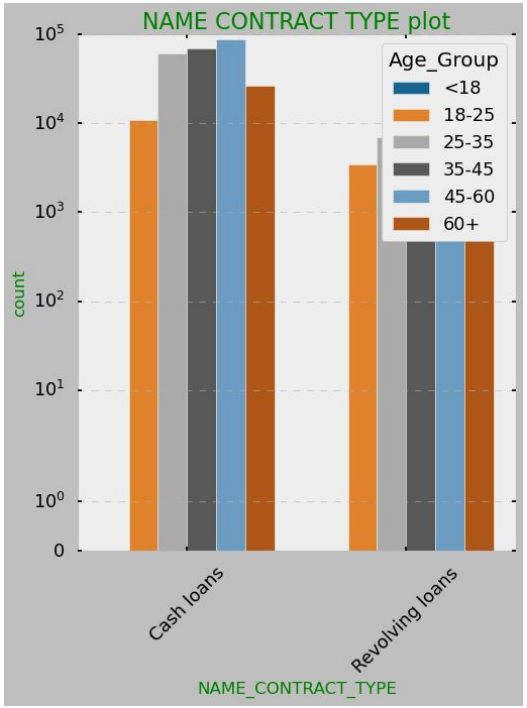
Target 1: defaulter



1. Cash loan has highest number of defaulter then Revolving loans
2. Count of females Defaulter are highest in both the contract type

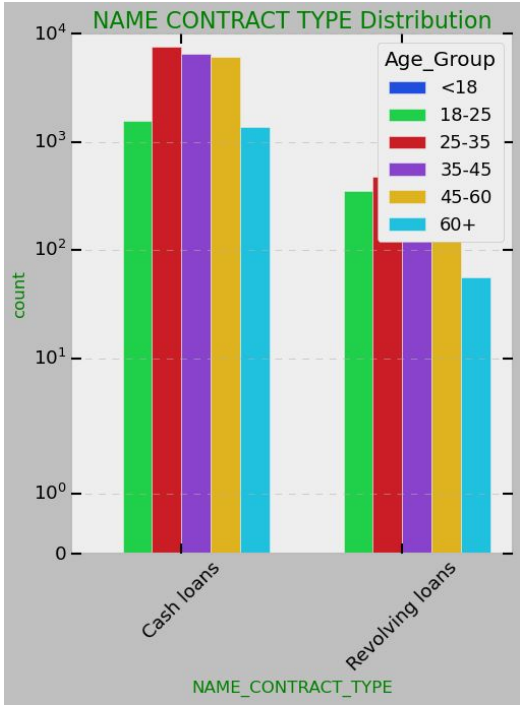
Categorical-Categorical: **Name Contract Type with age group**

Target 0: Non defaulter



1. Cash loan is most prefered by customer in the age group of 45-60
2. Age group 18-25 have the least count in both the types

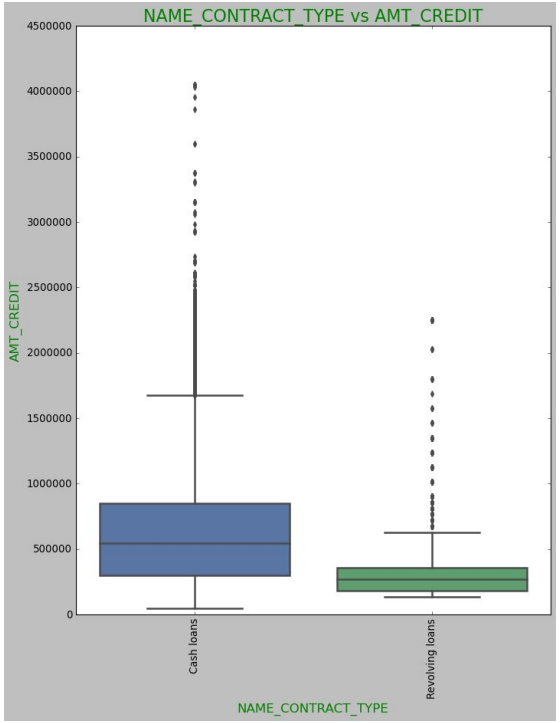
Target 1: defaulter



1. Age Group 25-35 have the highest defaulter Cash Loan Category
2. Age group 60+ have the least no of defaulters

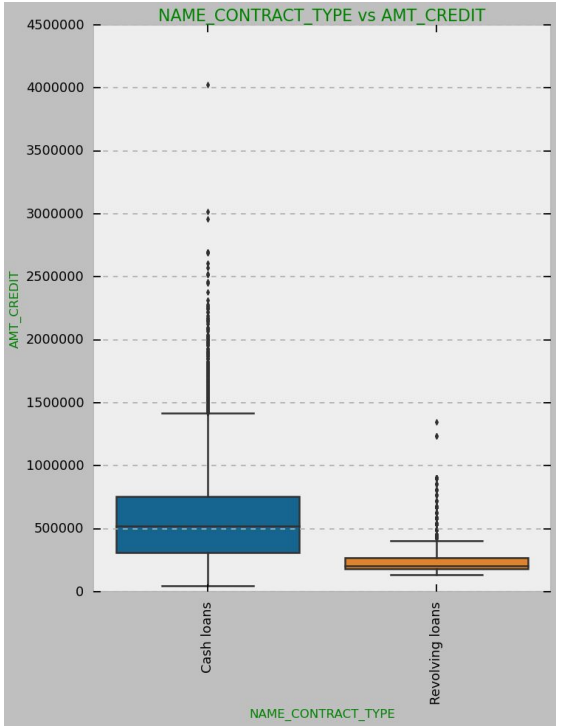
Categorical-Numerical: **Name Contract Type with AMT CREDIT**

Target 0: Non defaulter



1. Median of cash loan is higher than the Revolving loans

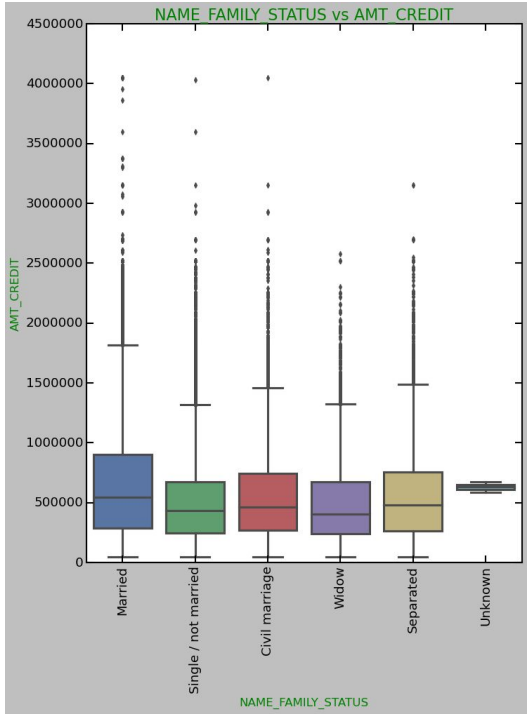
Target 1: defaulter



1. Median of the Cash loan is higher than the 100 Percentile of Revolving Loans

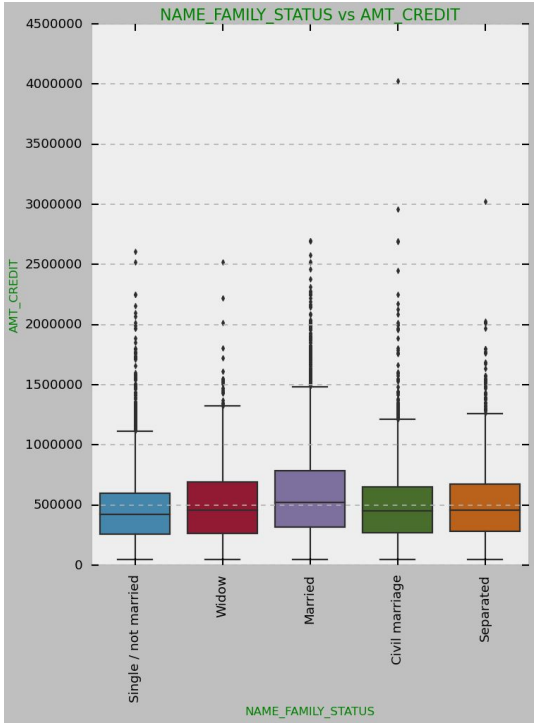
Categorical-Numerical: **Name Family Status with AMT CREDIT**

Target 0: Non defaulter



1. Median of Married Customers is higher than any other Family status
2. Widow customer have the least credit median.

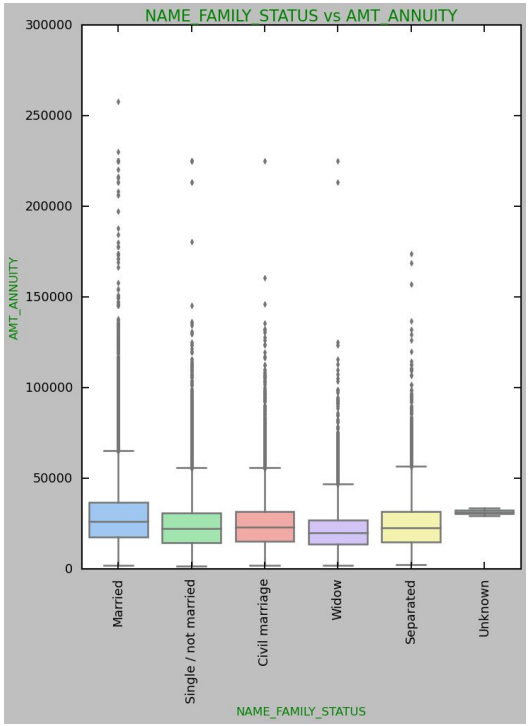
Target 1: defaulter



Married Customer Defaulter have the highest mean than any other Family status

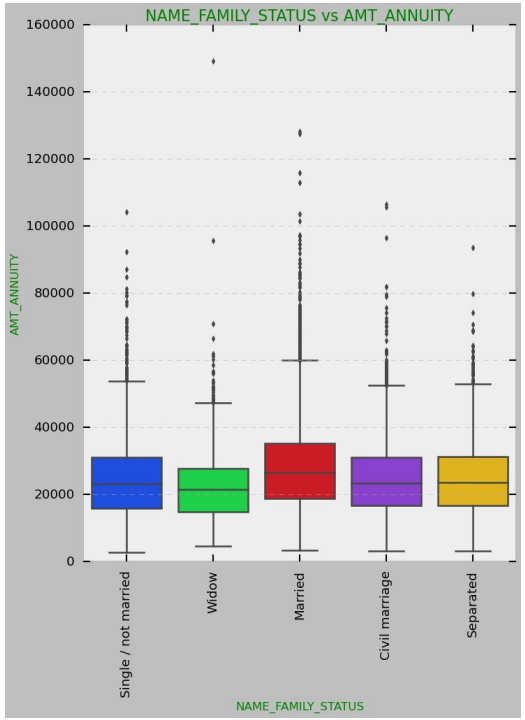
Categorical-Numerical: **NAME_FAMILY_STATUS vs AMT_ANNUITY**

Target 0: Non defaulter



1. Median and 3rd quartile for Married Customer have the highest annuity amt
2. Median of Widow Customer have the least annuity amt

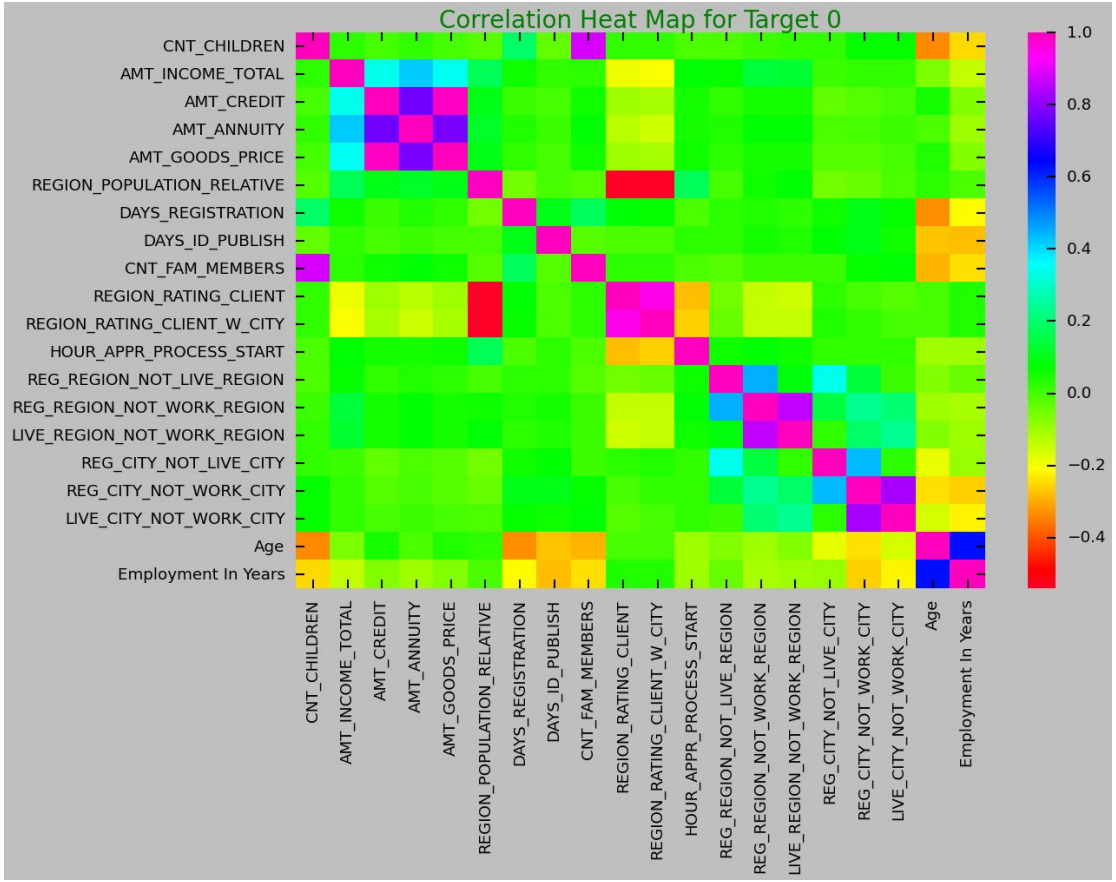
Target 1: defaulter



Married Customer Defaulter have the highest mean than any other Family status

Correlation Matrices

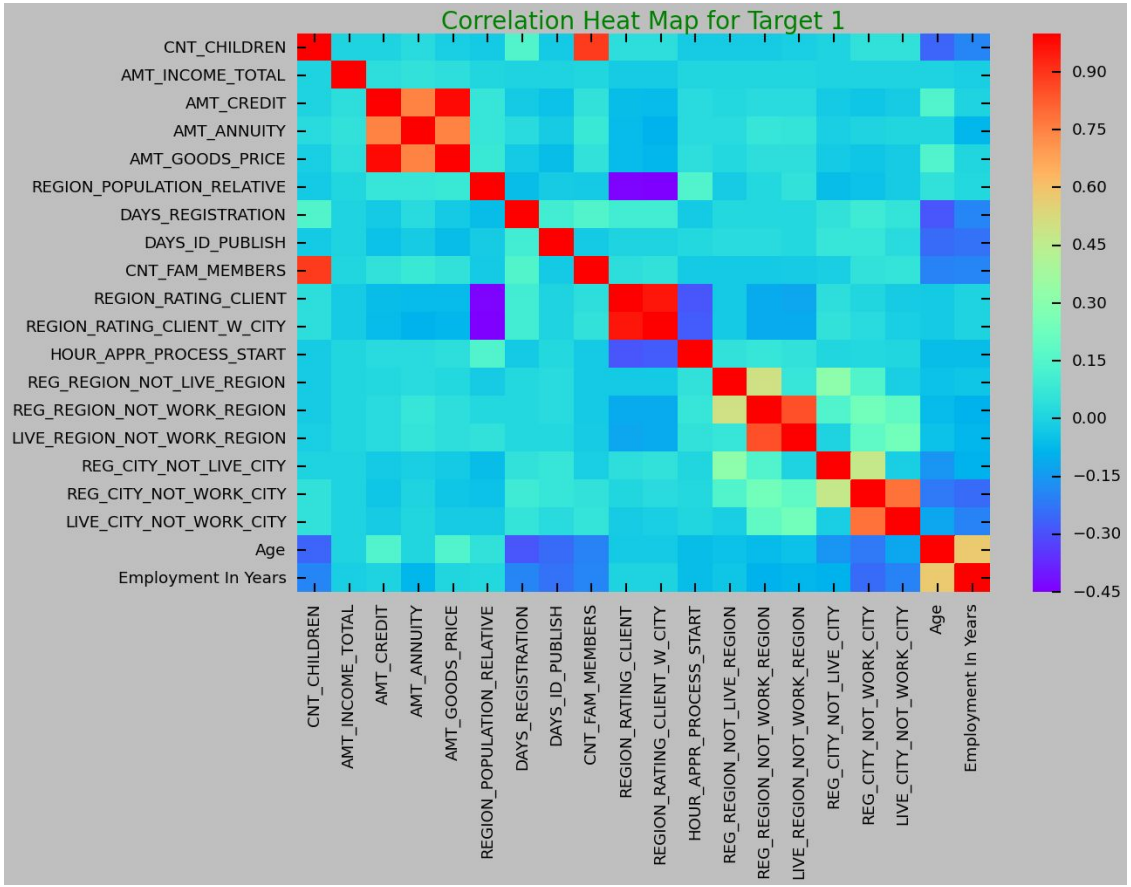
Numeric-Numeric: **Correlation HeatMap for Non-Defaulters (TARGET=0)**



Inferences from above plot

1. Income amt is directly related to AMT_ANNUITY,AMT_GOODS_PRICE and AMT_CREDIT ,customer with higher income have higher credit,investment and spending capacity
2. REGION_POPULATION_RELATIVE is inversely related with REGION_RATING_CLIENT and REGION_RATING_CLIENT_W_CITY, which implies customer with high region rating lives in less populated areas
3. Age and employment years of a customer are directly corelated as and when the customer age increases its employment age also increase
4. REG_CITY_NOT_LIVE_CITY and REG_CITY_NOT_WORK_CITY are directly corelated which implies customer with different permanent address have same live location and work location

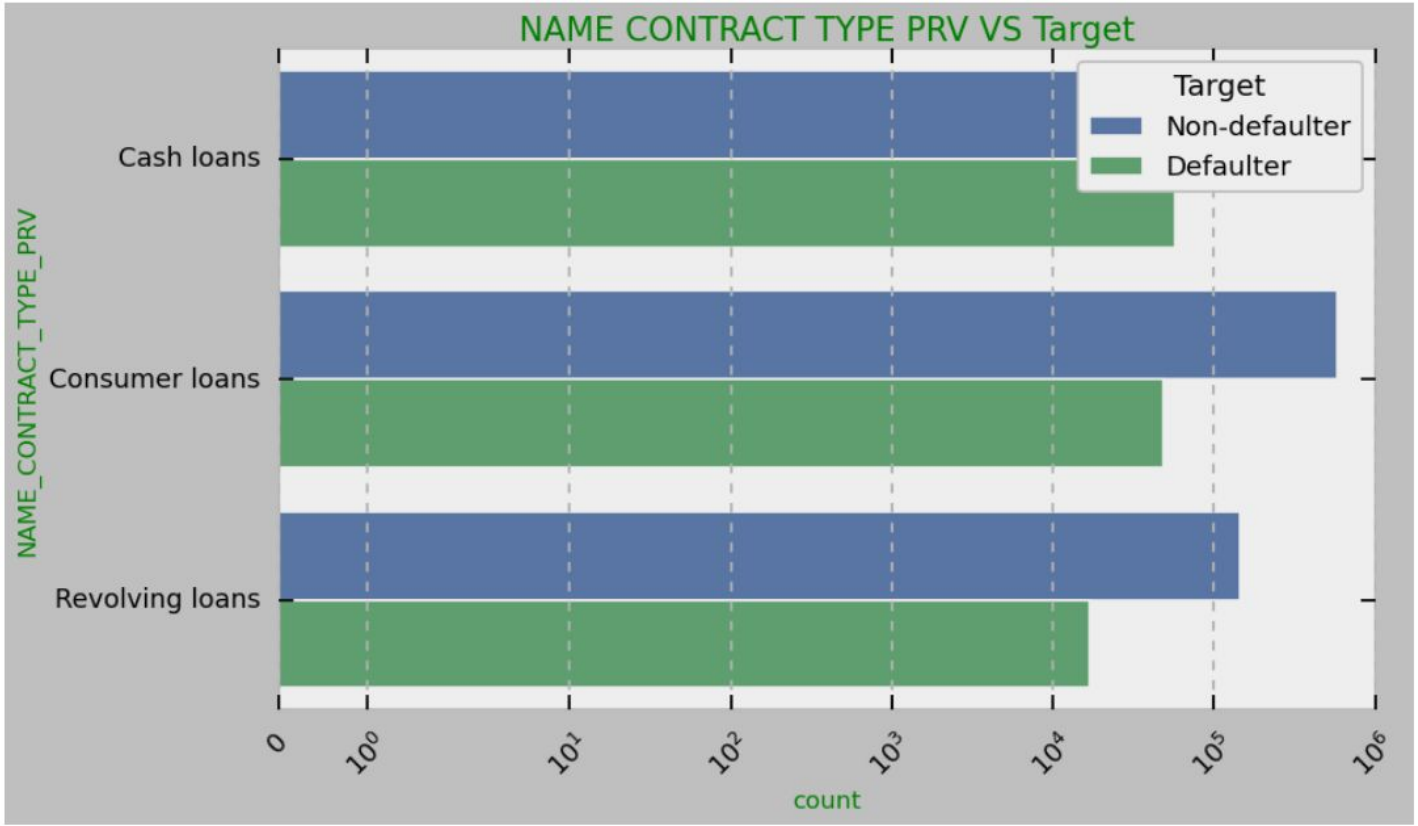
Numeric-Numeric: **Correlation HeatMap for Defaulters (TARGET=1)**



Inferences from above plot

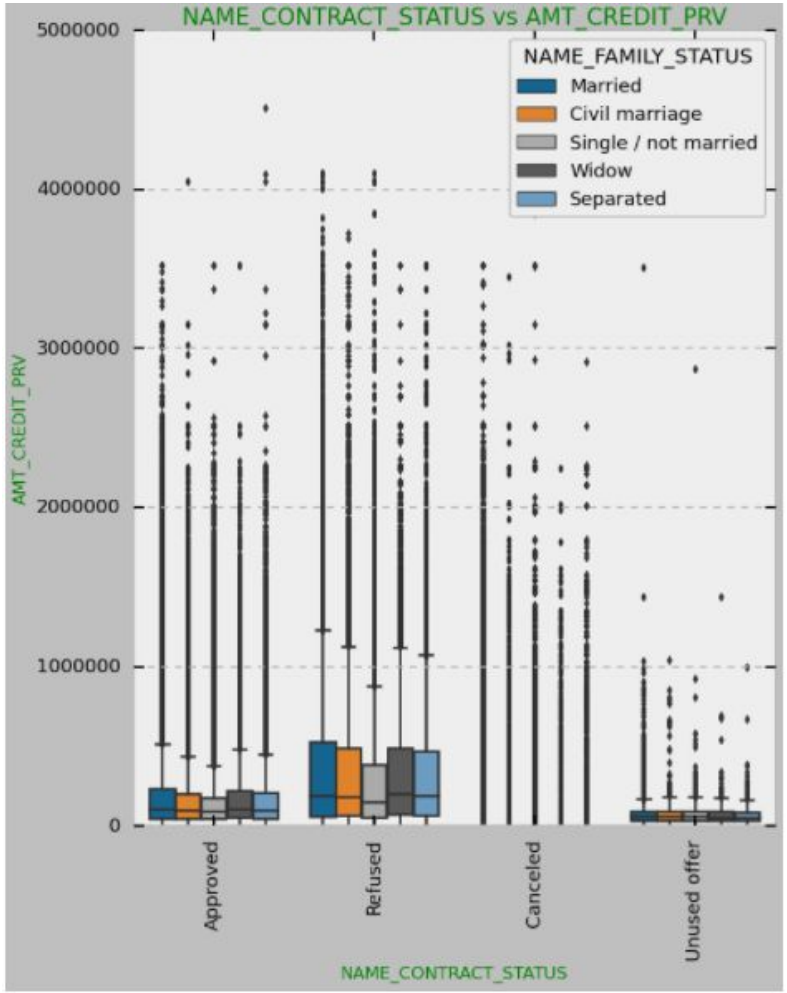
1. Income amt is directly related to AMT_ANNUITY,AMT_GOODS_PRICE and AMT_CREDIT ,customer with higher income have higher credit,investment and spending capacity
2. REGION_POPULATION_RELATIVE is inversely related with REGION_RATING_CLIENT and REGION_RATING_CLIENT_W_CITY, which implies customer with high region rating lives in less populated areas
3. Age and employment years of a customer are directly correlated as and when the customer age increases its employment age also increase
4. REG_CITY_NOT_LIVE_CITY and REG_CITY_NOT_WORK_CITY are directly correlated which implies customer with different permanent address have same live location and work location
5. Correlation for defaulters is similar to non defaulters

Analysis for Client facing Payment Difficulties



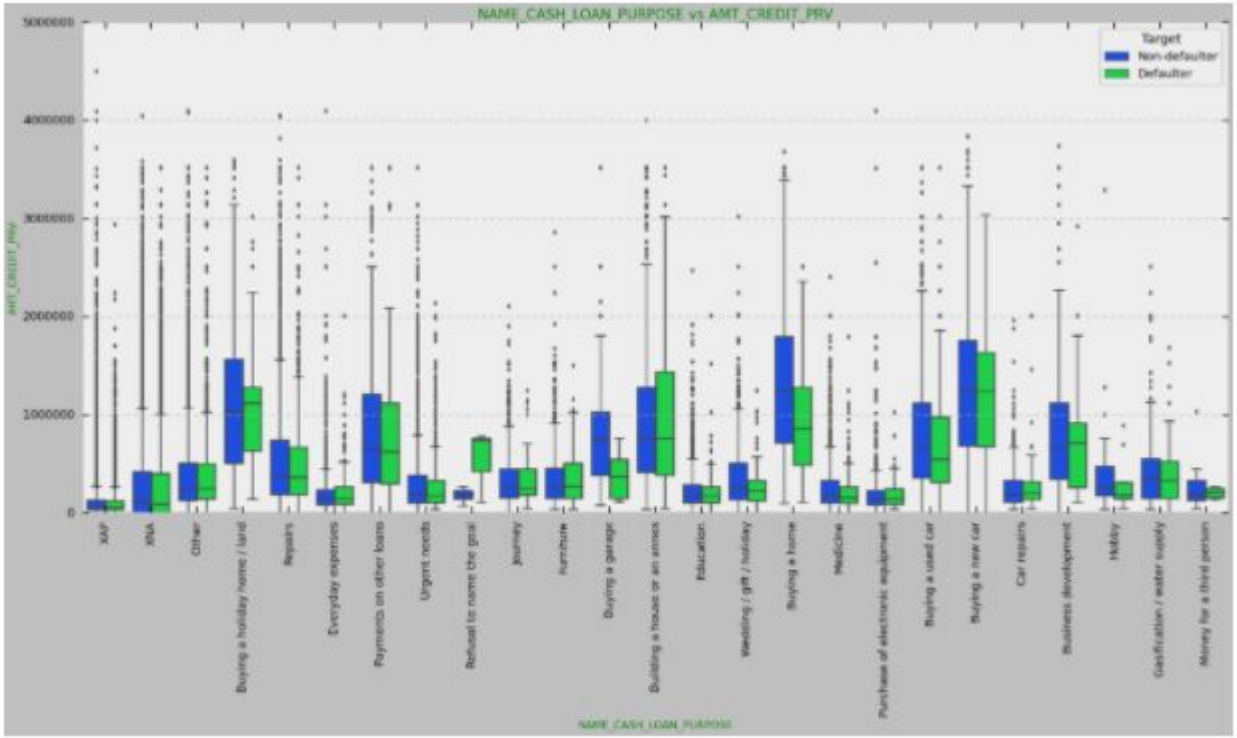
CONTRACT TYPE VS TARGET

1. Cash Loan and Consumer loan have the highest no of defaulter.
2. Revolving loan has least number of defaulter



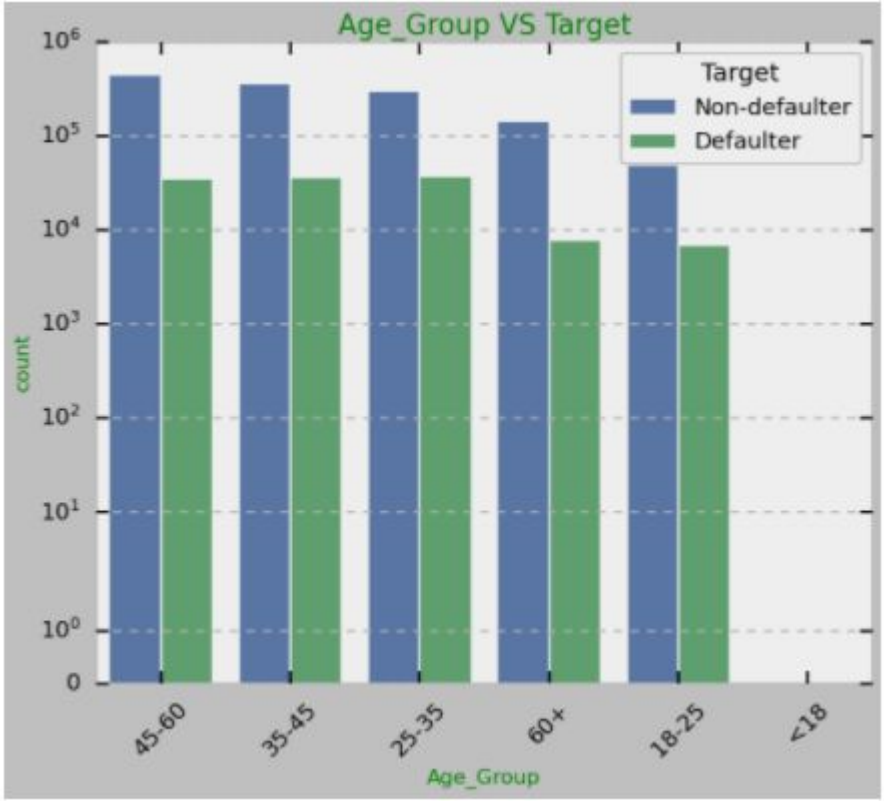
CASH LOAN PURPOSE vs AMT CREDIT PRV WITH FAMILY STATUS

1. For cancelled status since the amt was not credited, the credit amt for cancelled cases is 0, There are outliers in cancelled cases where loan got cancelled post disbursement.
2. Outliers exist in all status Type.
3. We can observe outliers are continous in Married Family status with Approved loan status upto 25.5L



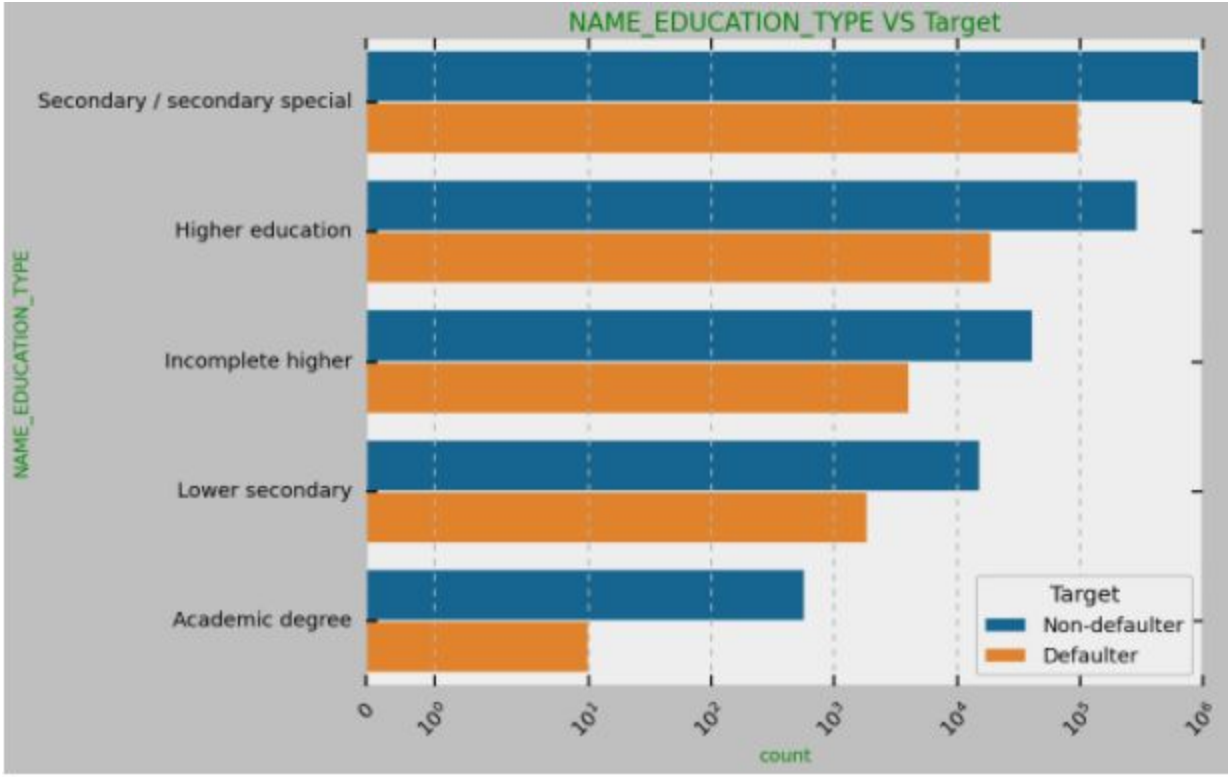
CASH LOAN PURPOSE vs AMT CREDIT PRV WITH TARGET

1. Defaulter Customer whose loan purpose is refused to make a goal has the highest Median for Credit Amt in defaulter
2. Defaulter Customer with loan purpose for Building a house for annex has 3rd quartile higher than its corresponding Non defaulter 3rd quartile range.
3. Non Defaulter customer with loan purpose for Buying a home has the highest median across all loan purposes.



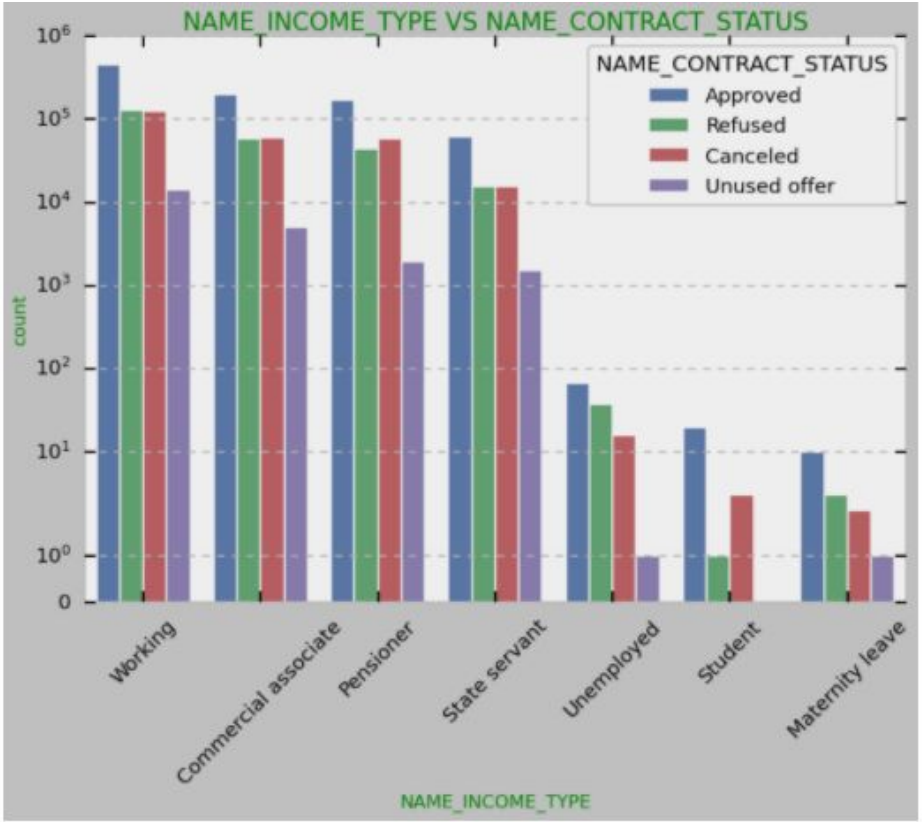
AGE GROUP VS TARGET

1. Customer of age group 25-35 ,35-45 and 45-60 have the highest no of defaulter
2. Customer of age group 60 + and 18-25 group have the least number of defaulter



EDUCATION TYPE VS TARGET

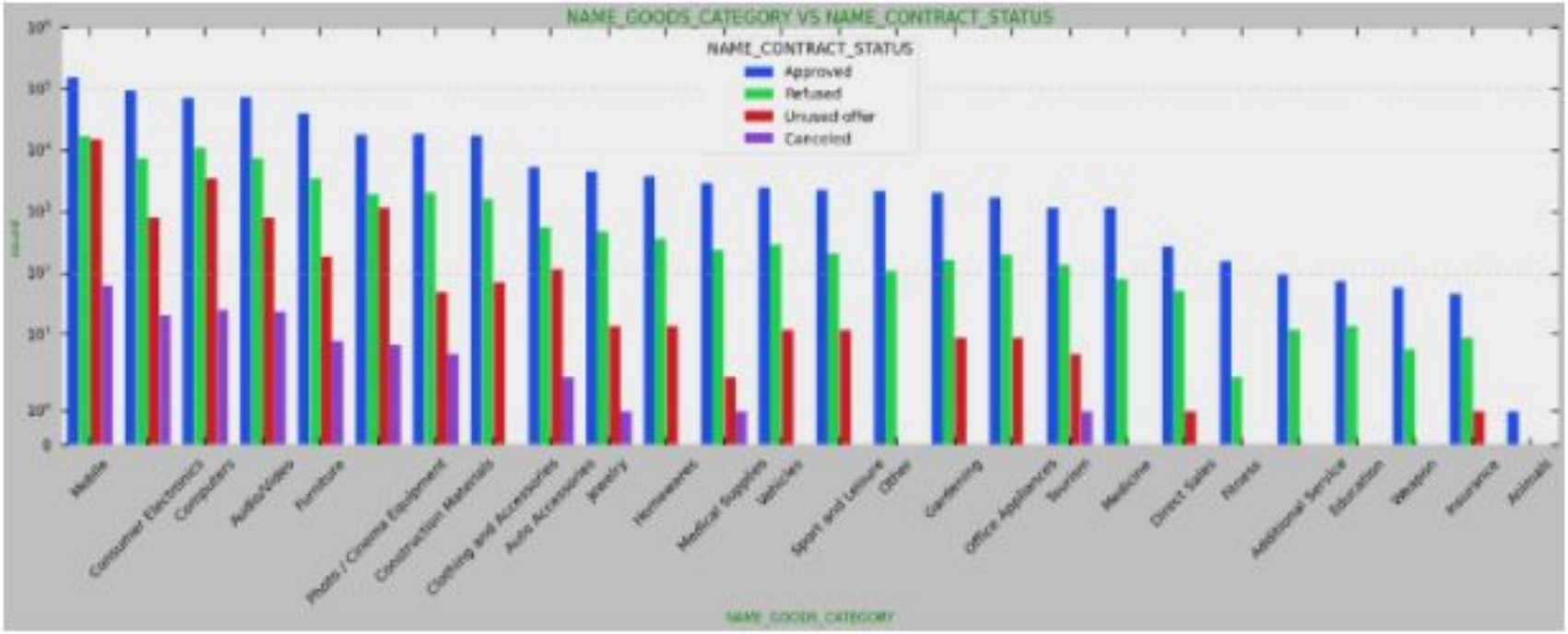
1. Academic degree holder have the least no of defaulter
2. Defaulters are higher for Secondary and Higher education qualification



INCOME TYPE VS CONTRACT STATUS

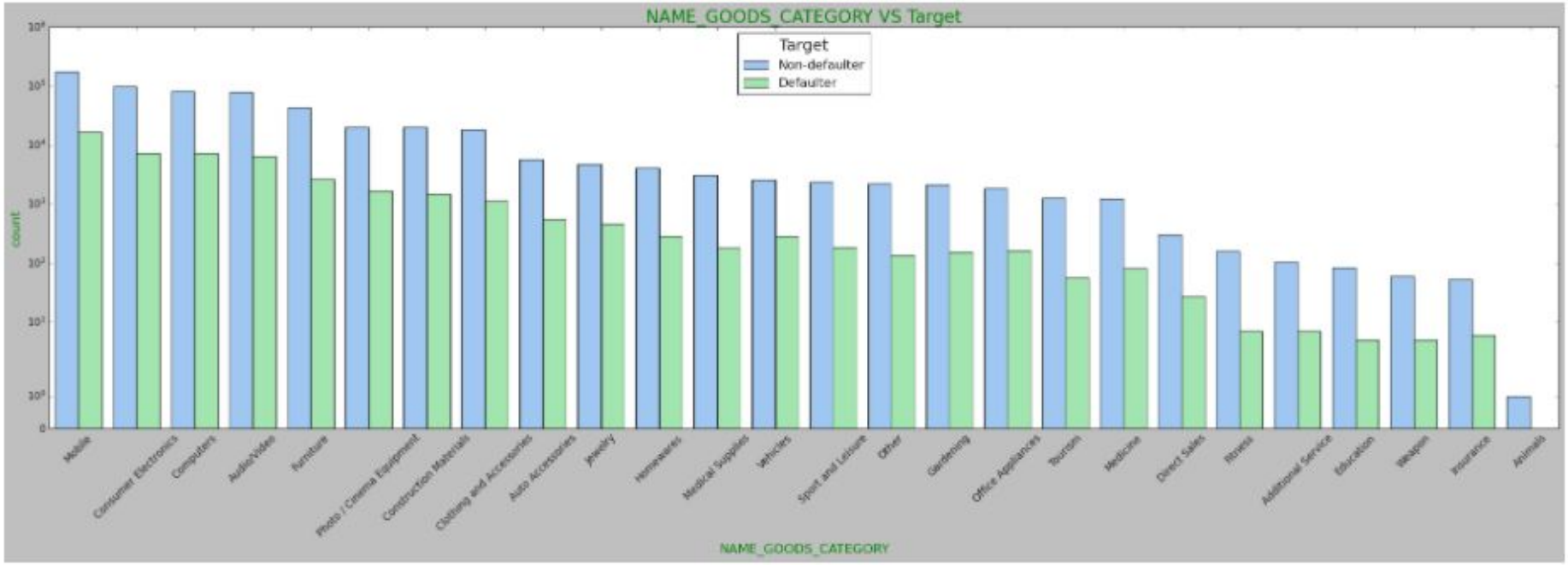
1. Student have no unused offer
2. Approval is higher for Working income type
3. For customer on maternity leave has the least loan approval

GOODS TYPE VS CONTRACT STATUS

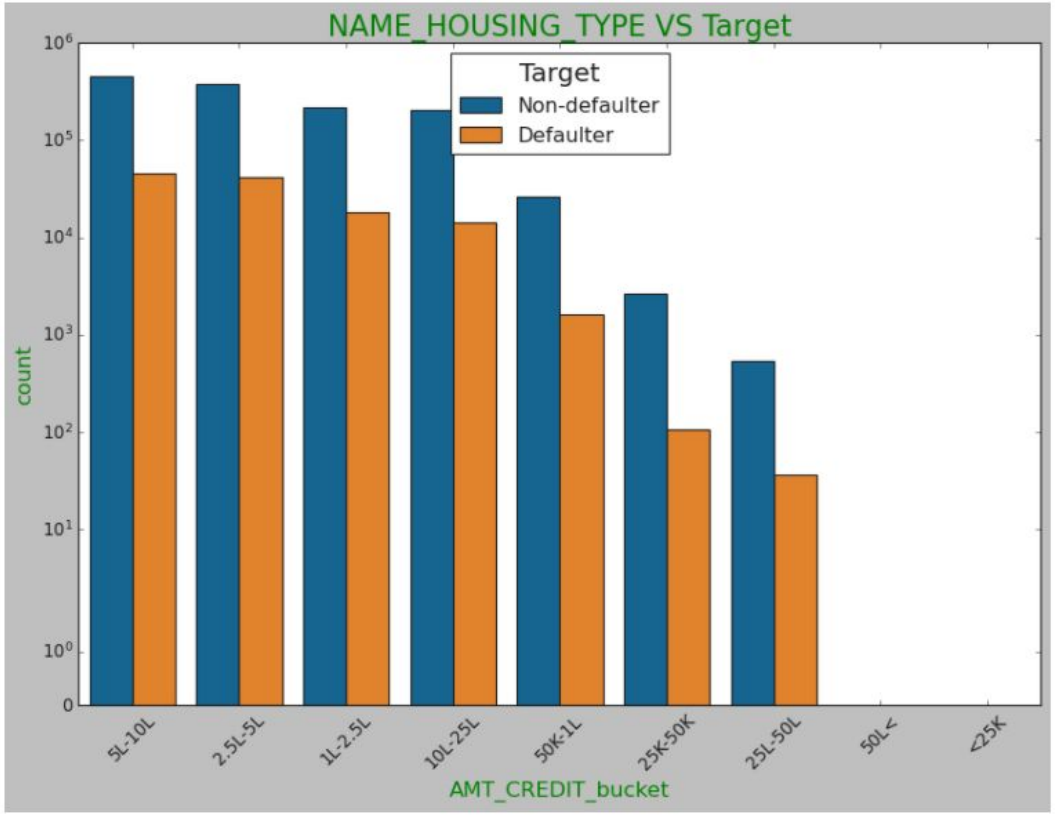


1. Customer with goods type as mobile has highest loan approval
2. Customer for goods type Others, Medicine, Fitness, additional service education ,weapon and animals has zero loan cancellation.

GOODS TYPE VS TARGET



1. Customer with goods type fitness, Additional service, education, weapon ,insurance and Animals has the least number of defaulters
2. Defaulters are higher for Mobile and Computer electronics goods



HOUSING TYPE VS TARGET

1. Customer with housing type Co-op apartment and office appartment has the least number of defaulter
2. Customer staying in House/ Apartment have the highest number of defaulter

Conclusions

1. **NAME_EDUCATION_TYPE:** Banks Should issue loan Customer with Academic degree as they have least defaulter.

2. **NAME_HOUSING_TYPE:** Loans should be issued to customer staying in Co-op apartment and office appartment because they have less chances of defaulting

3. AGE_GROUP: Customer with age group 18-25 and customer with age 60+ have less chances of defaulting

4. Customer with loan purpose as refused to make goal should be avoided as they have high chances of defaulting

5. NAME_INCOME_TYPE: Customer with income type Student and business should be targeted as they are less likely to default.

6. NAME_GOODS_CATEGORY: Customer with goods type fitness, Additional service, education, weapon ,insurance and Animals has the least number of defaulters

7. CODE_GENDER: Females have higher income than males and prefer credit more than males. Males tends to default more.

8. WORK_GROUP: Customer with work experience of more than 45 years of age have less chances to default.

Conclusions Cont.

9. AMT_INCOME_TOTAL: Customer with income higher than 5L are less likely to default.

10. AMT_CREDIT: Customer with credit limit less than 10 have high chances to default