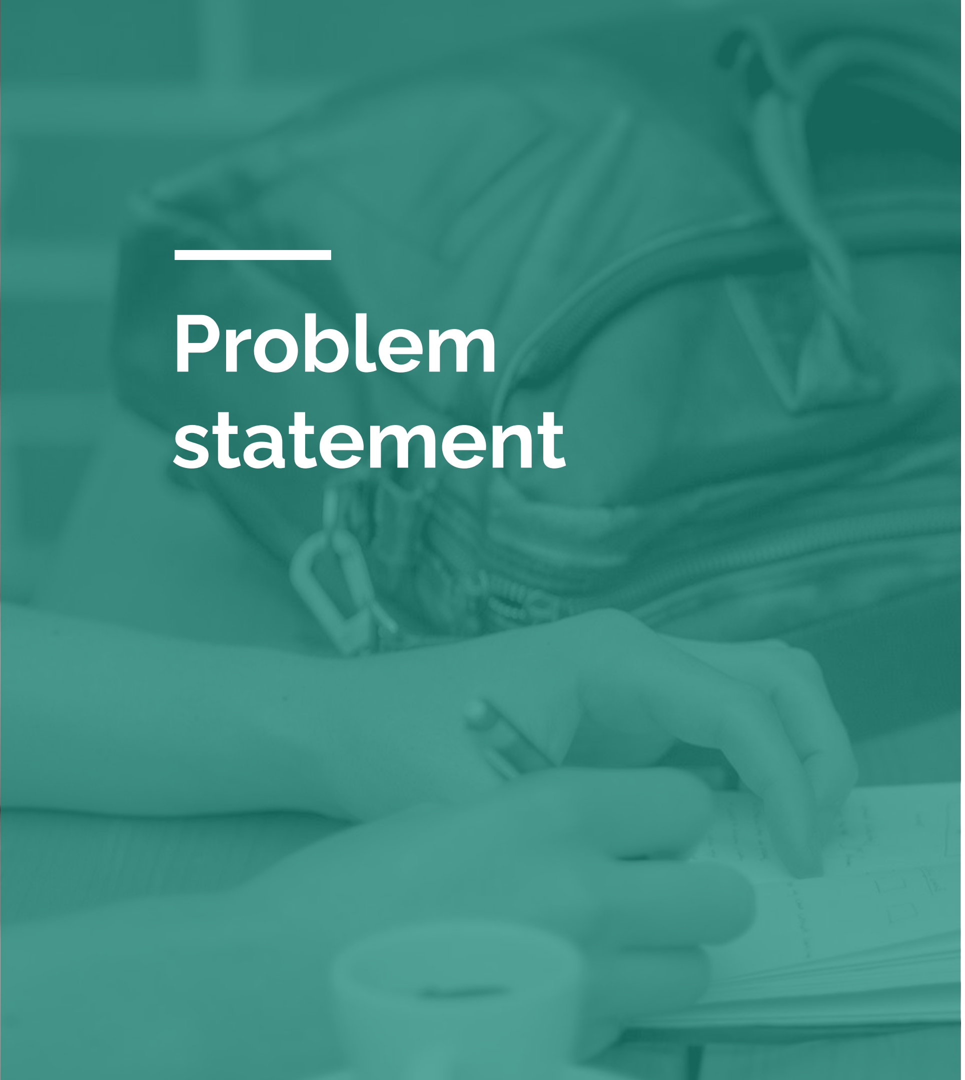
Exploratory Data Analysis: Bank Loan Default

- Bharathy Arunachalam

Purpose of the Case Study

This case study will provide a understanding of risk analysis in bank and understand how the data is used to minimise the risk of losing the money when lending it to the customers.



Problem statement

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. 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:

- 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.

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.

Datasets used for EDA Case study

Below are the two datasets used for which EDA has been carried out:

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.

Data Analysis Approach

Steps to perform Data Analysis for Application and Previous application datasets

Step 1: Data Sourcing

- Load the application data set.

Step 1: Data Cleaning

- Fixing rows and columns.
- Identifying the missing values and use appropriate method to deal with it (remove/replace).
- Identify the incorrect data types and fix it.
- Handling Outliers.
- Standardizing the values.

Step 2: Perform Univariate Analysis

- Univariate numerical analysis
- Correlation between numerical values
- Univariate Unordered categorical analysis
- Univariate Ordered categorical analysis

- *Continued*

Data Analysis Approach

Step 3: Perform Bivariate Analysis

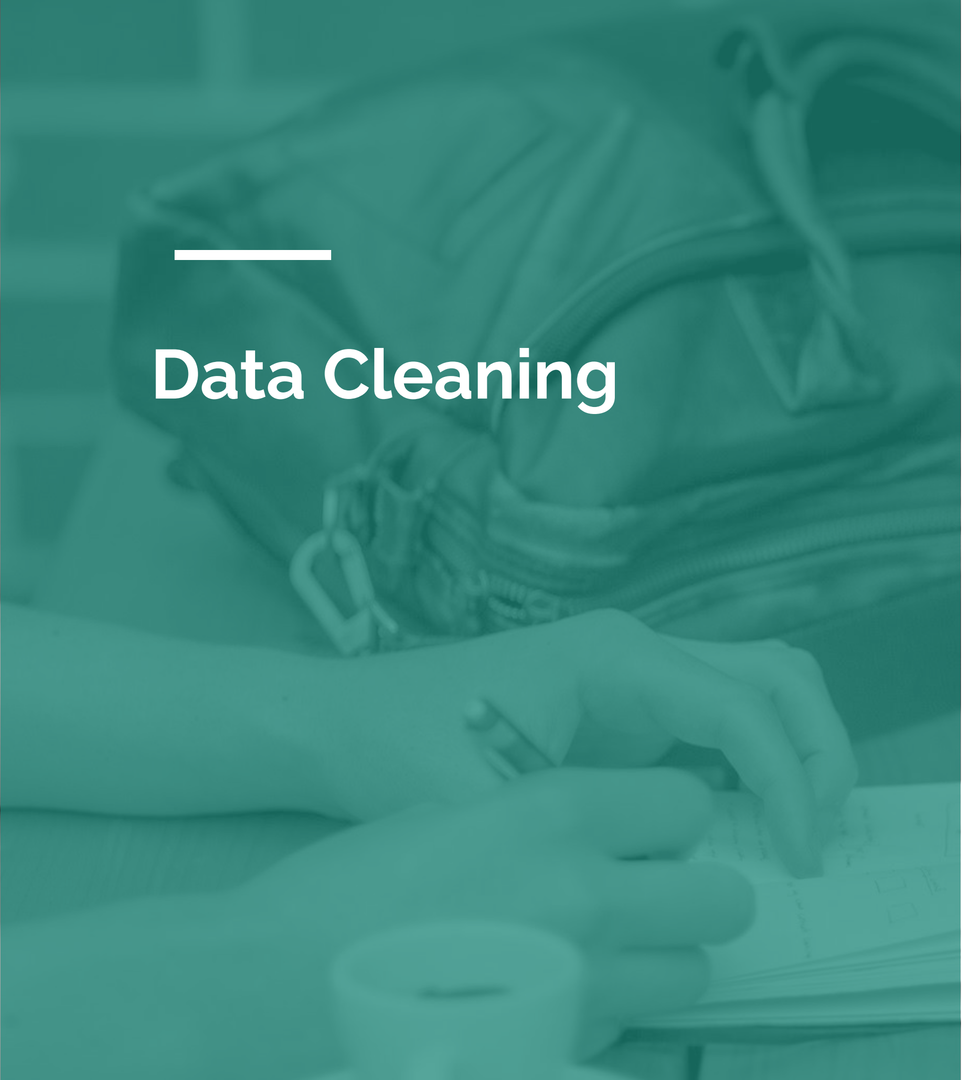
- Bivariate numeric - numeric analysis
- Bivariate numeric - categorical analysis
- Bivariate categorical - categorical analysis

Step 4: Perform Multivariate analysis

Step 5: Data analysis of previous application dataset.

Step 6: Merging of both the data frames and provide the insights.

Step 7: Conclusion.



Data Cleaning

The data cleaning has been performed in the following ways:

Missing value Imputation:

- The missing values are identified and the columns having more than 40% data missing are dropped from the data frame.
- Few of the missing data were replaced with mean, median and mode.

Data types:

- The incorrect data types have been identified and replaced with correct one.

Handling Outliers:

- The outliers were identified and replaced with median, median and for few of the columns the outliers were capped.

Standardizing the values.

- Handling Negative Values
- Handling XNA values

Data Analysis

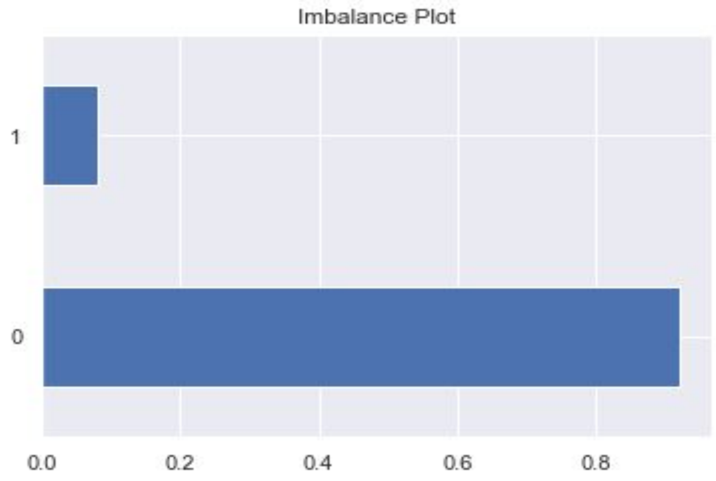
Below are the Analysis methods followed:

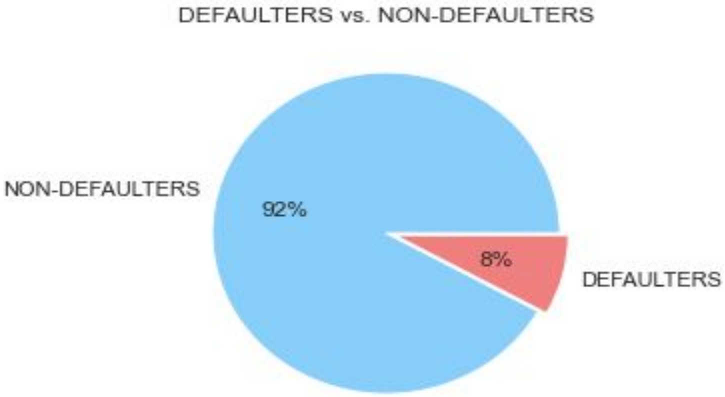
- Imbalance Analysis
- Univariate Analysis
- Bivariate Analysis
- Multivariate analysis

Imbalance Analysis:

Non-Defaulters and Defaulters

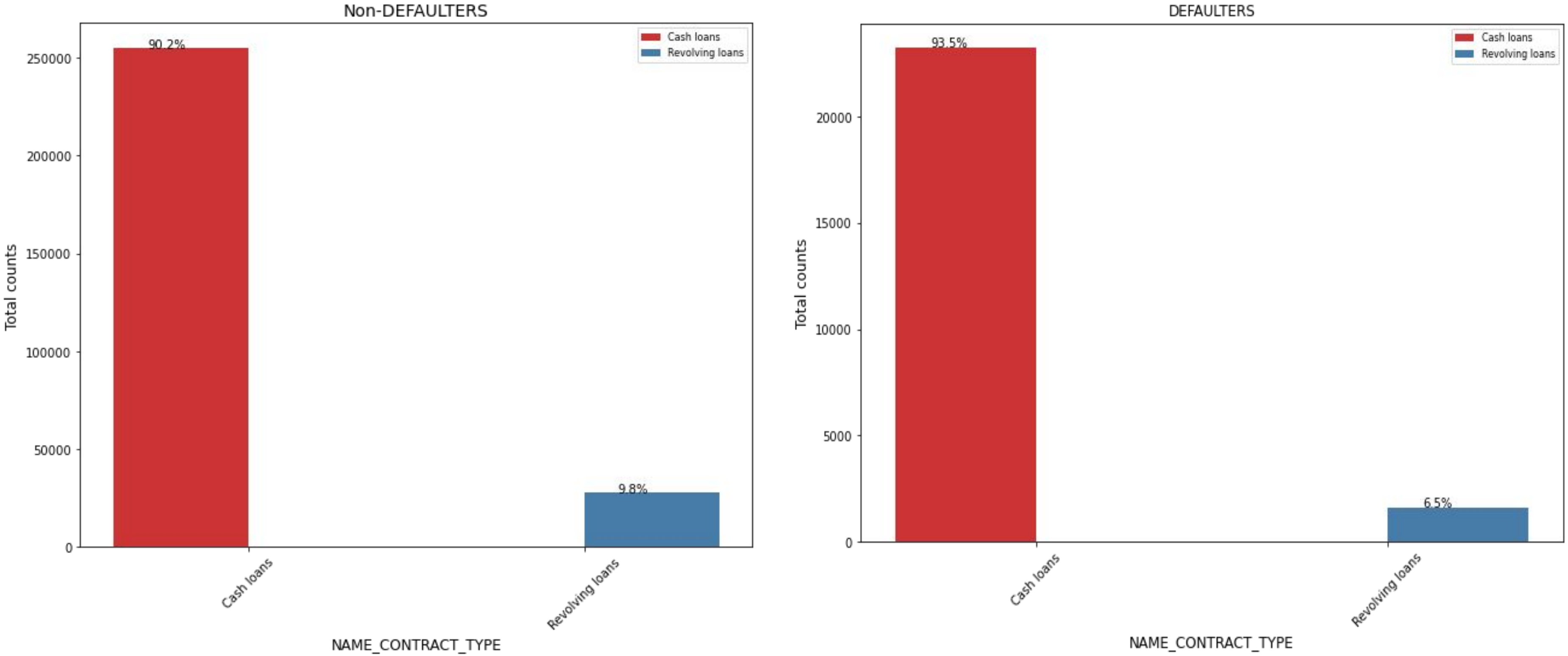
The TARGET column has 8.07% of 1's which means there are 8.07% defaulters who have difficulty in paying the loan amount and 91.9% are non defaulters who does not have any difficulty in repaying the amount. The imbalance ratio is 11.39



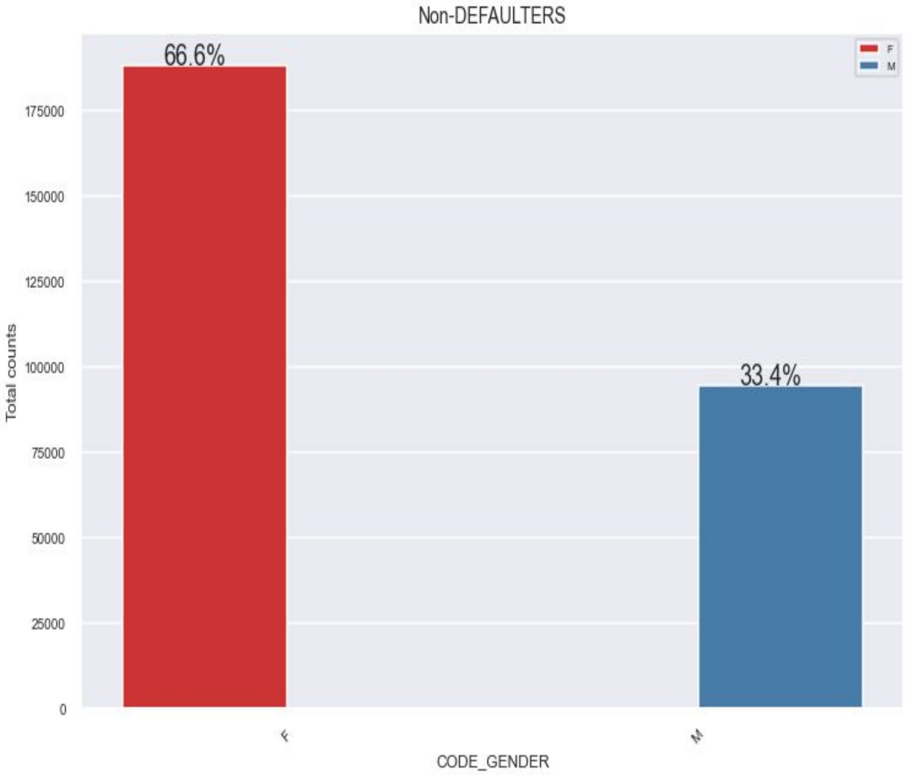


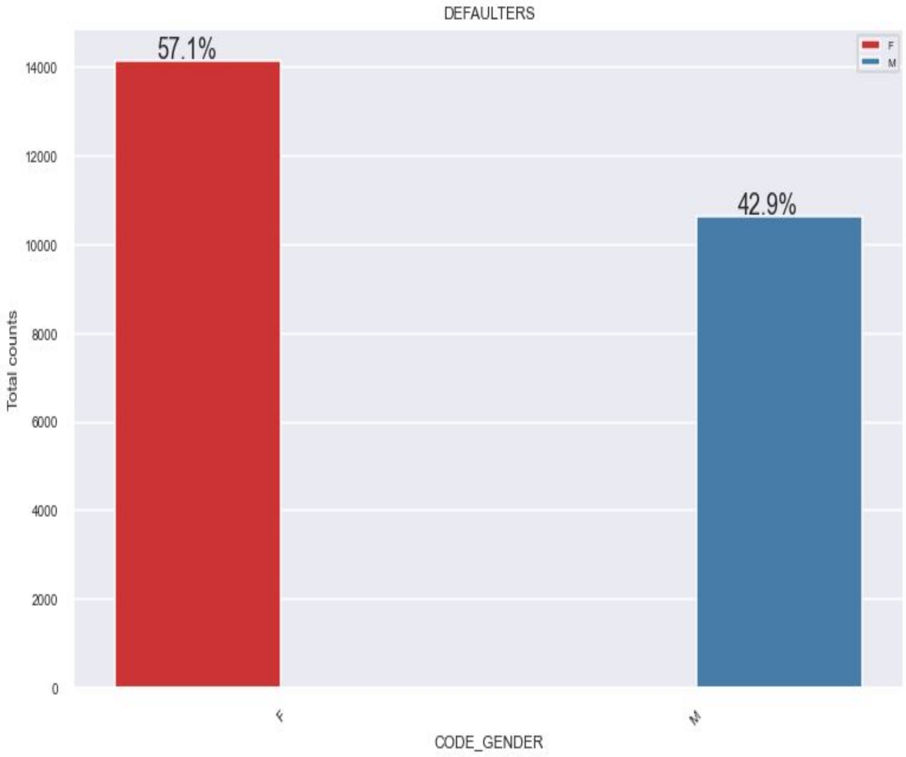
Univariate Categorical UnOrdered Variables

NAME_CONTRACT_TYPE - Non-Defaulters vs. Defaulters

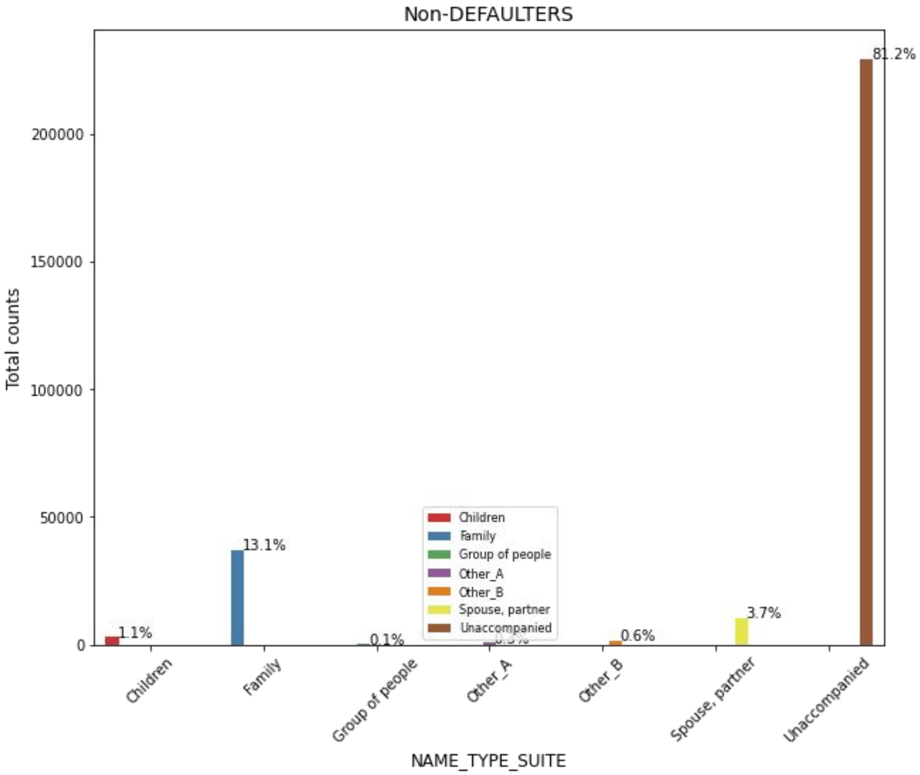


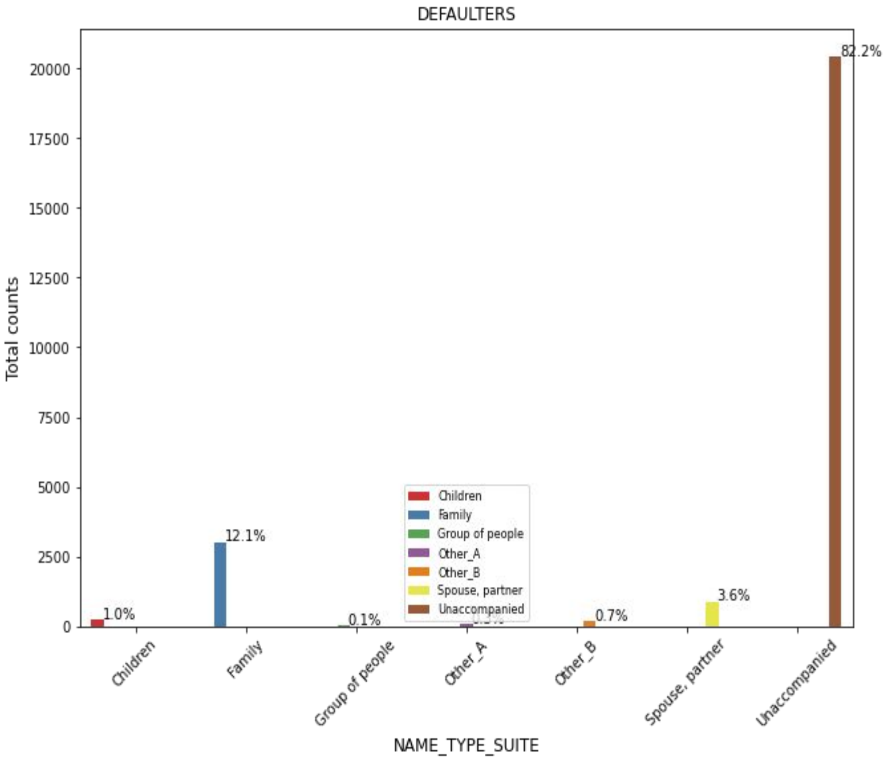
CODE_GENDER - Non-Defaulters vs. Defaulters



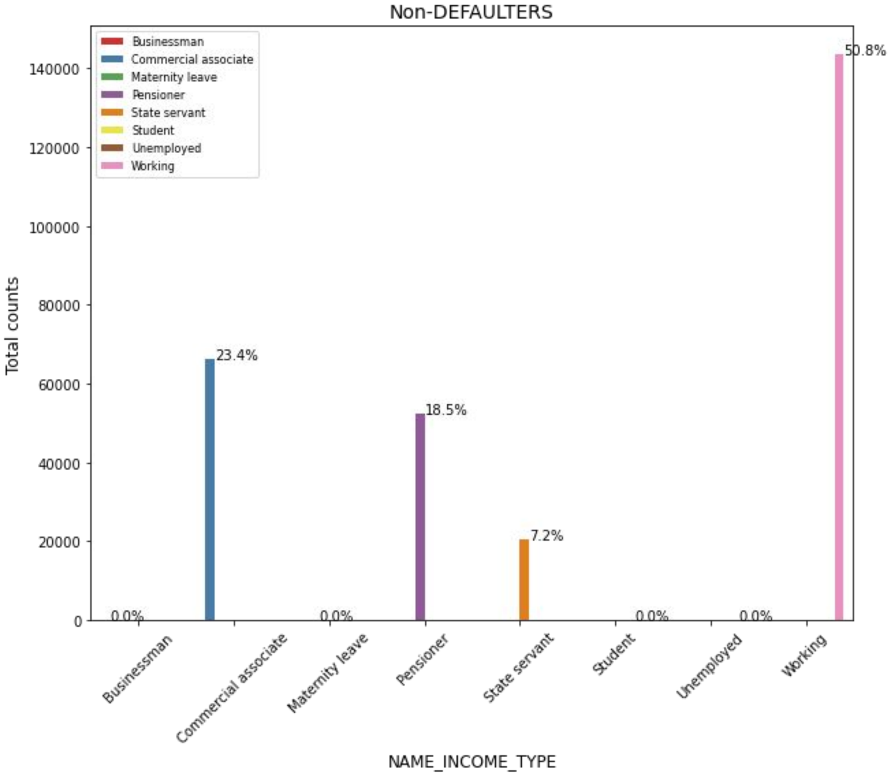


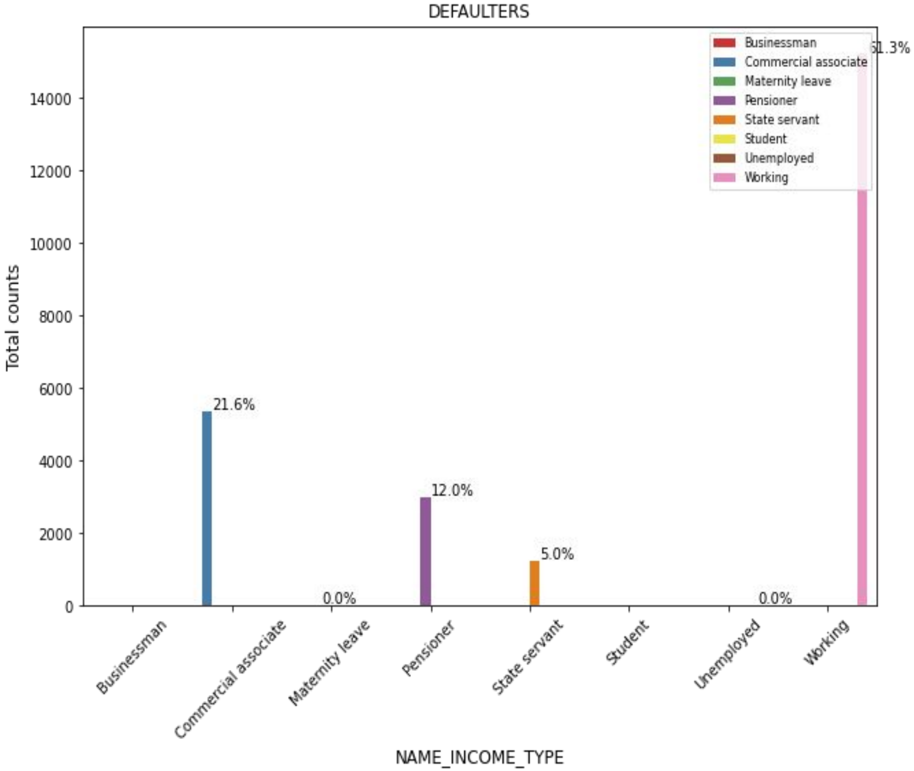
NAME_TYPE_SUITE - Non-Defaulters vs. Defaulters



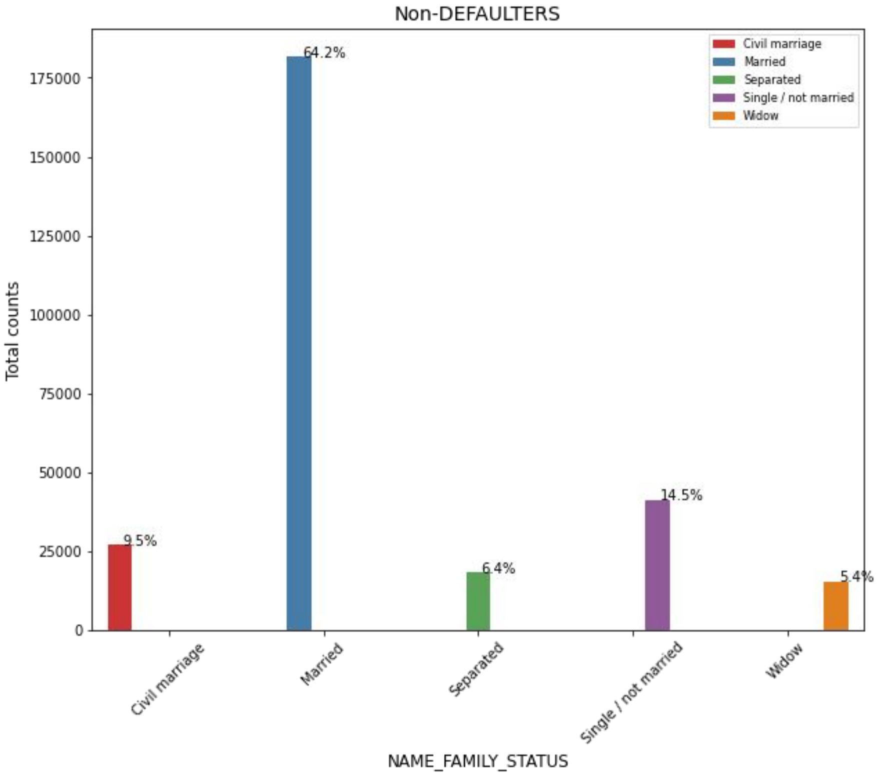


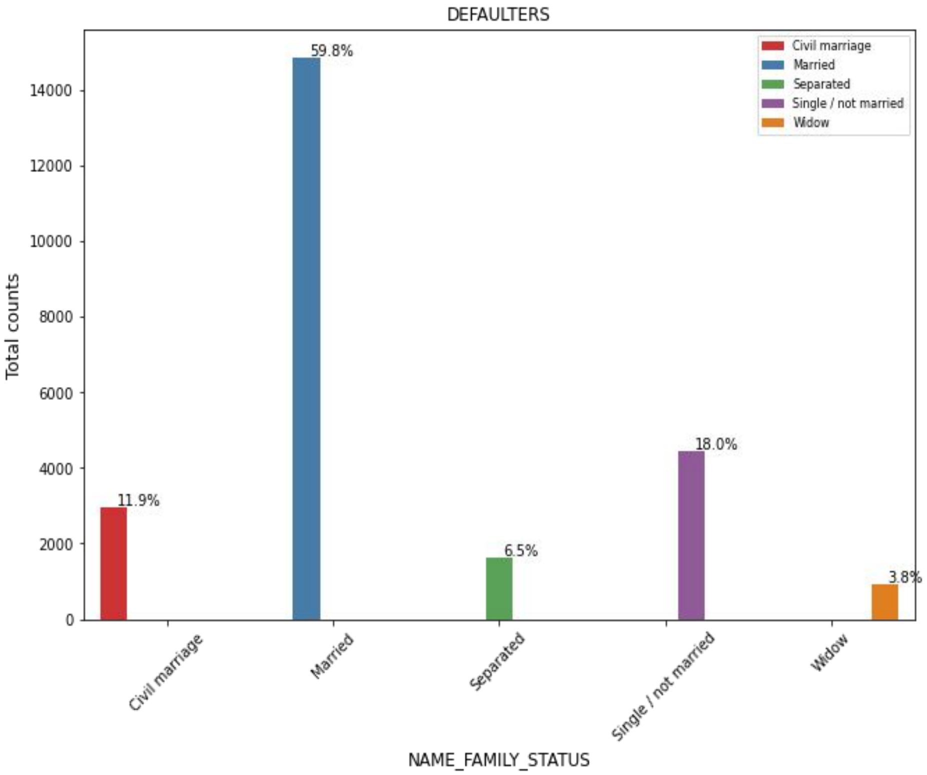
NAME_INCOME_TYPE - Non-Defaulters vs. Defaulters



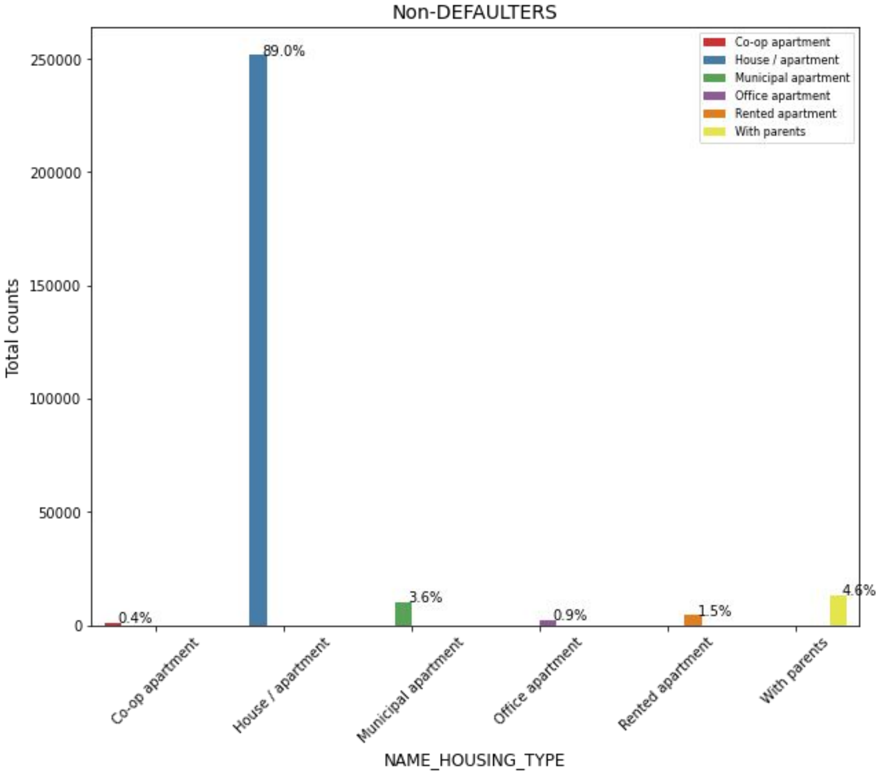


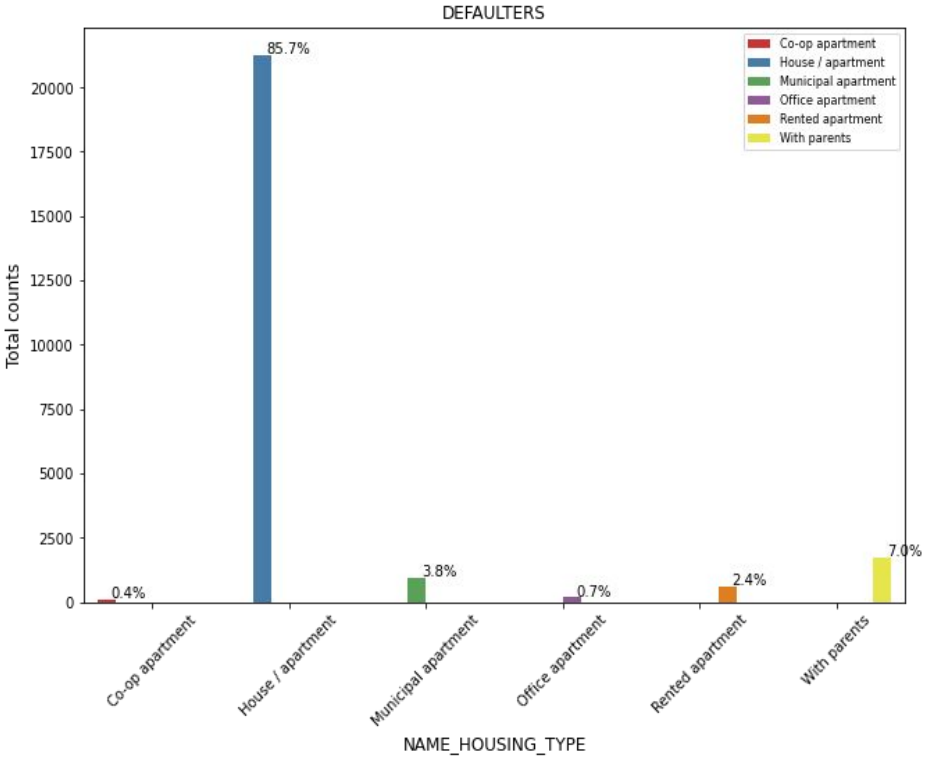
NAME_FAMILY_STATUS - Non-Defaulters vs. Defaulters



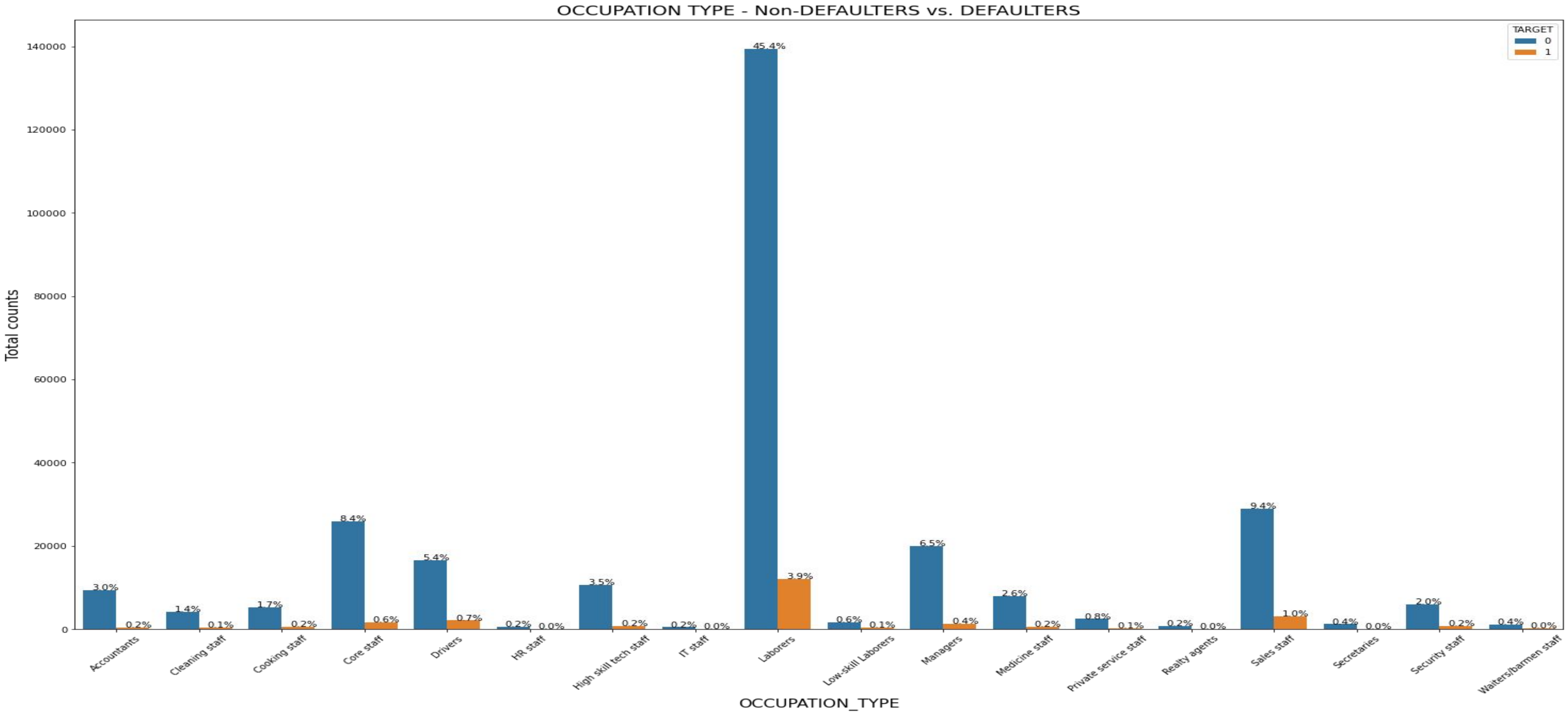


NAME_HOUSING_TYPE - Non-Defaulters vs. Defaulters

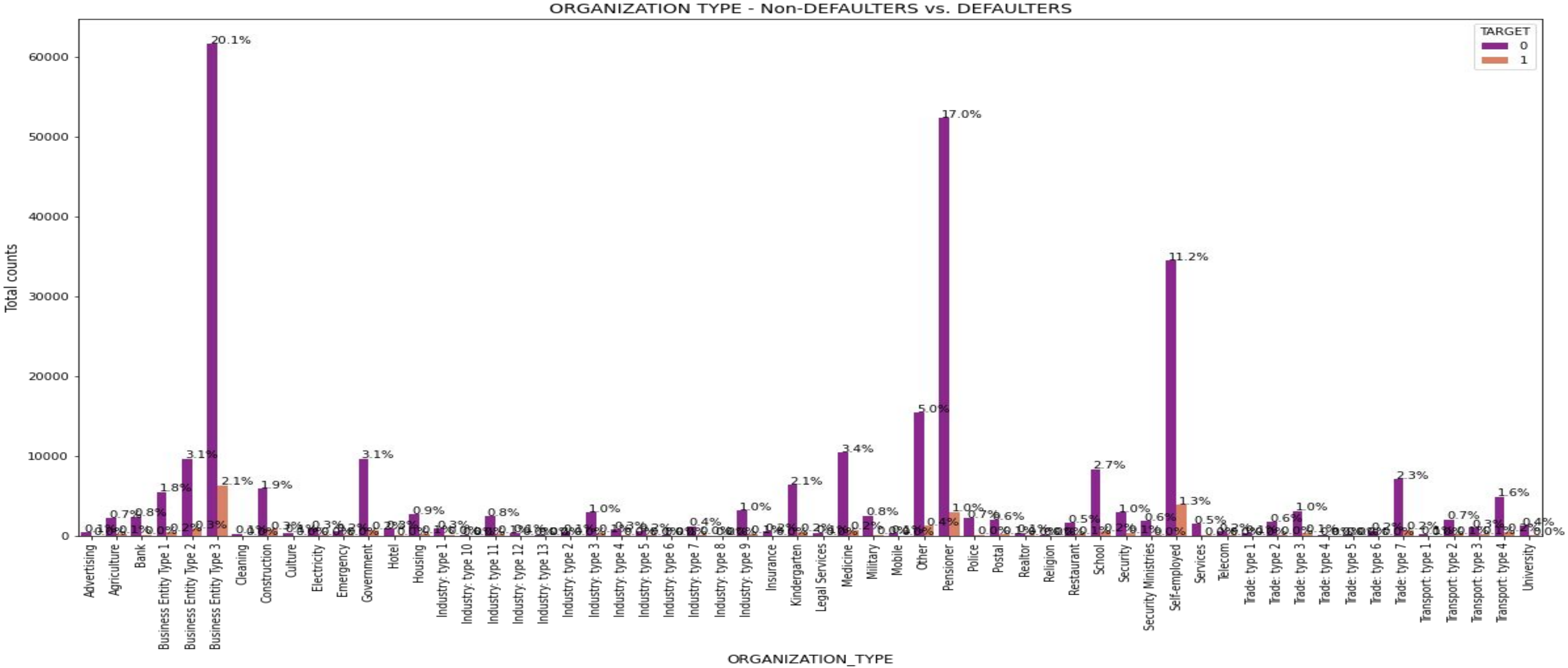




OCCUPATION TYPE - Non-Defaulters vs. Defaulters



ORGANIZATION TYPE - Non-Defaulters vs. Defaulters



Univariate Categorical UnOrdered Variables

Insights:

- NAME_CONTRACT_TYPE: The revolving loans are applied less when compared to cash loans. Also the percentage of defaulters are less for revolving loans.

- CODE_GENDER: Female clients are the ones who apply loan when compared to male. 57% Female clients are defaulters while 42% male clients are defaulters. The percentage of male defaulters are more when compared to male non-defaulters.

- NAME_TYPE_SUITE: The family has the second highest percentage in accompanying the client. The percentage is almost same for both defaulters and non-defaulters. The first highest is that the client is unaccompanied while applying the loan.

- NAME_INCOME_TYPE: The income type working are the ones who apply for the loan in both Non-defaulters and defaulters. Business, Student, unemployed are less likely to apply loan.

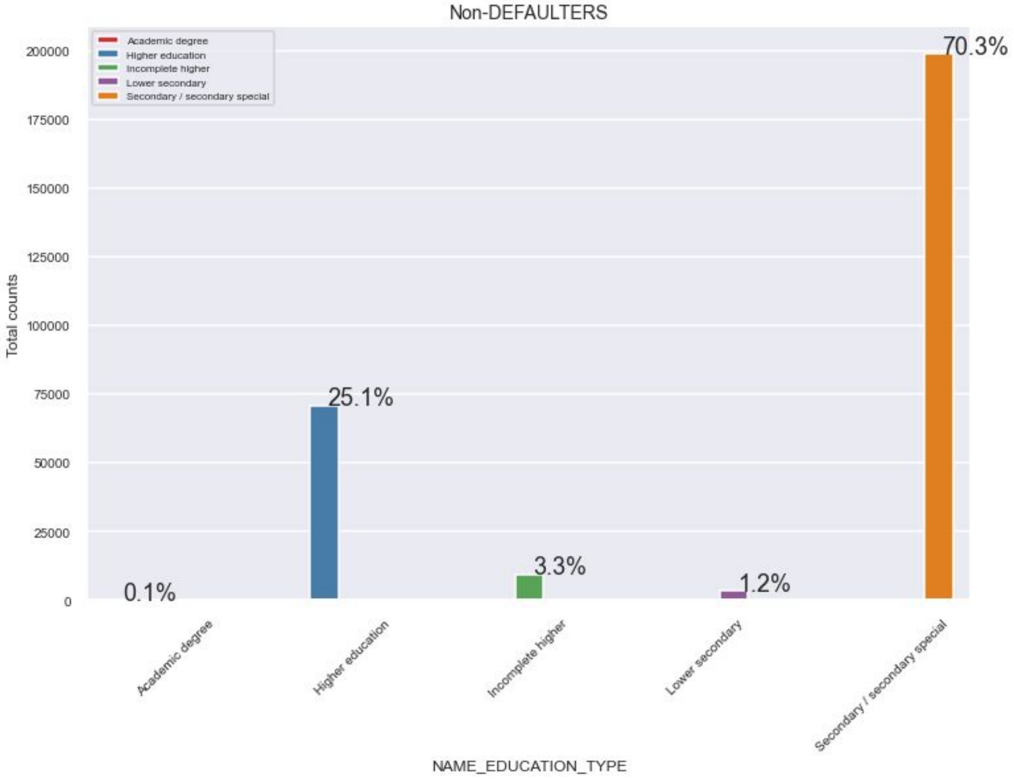
- NAME_HOUSING_TYPE: House / apartment has highest percentage of defaulters and non defaulters and it is clear that most of the clients has a own house or apartment.

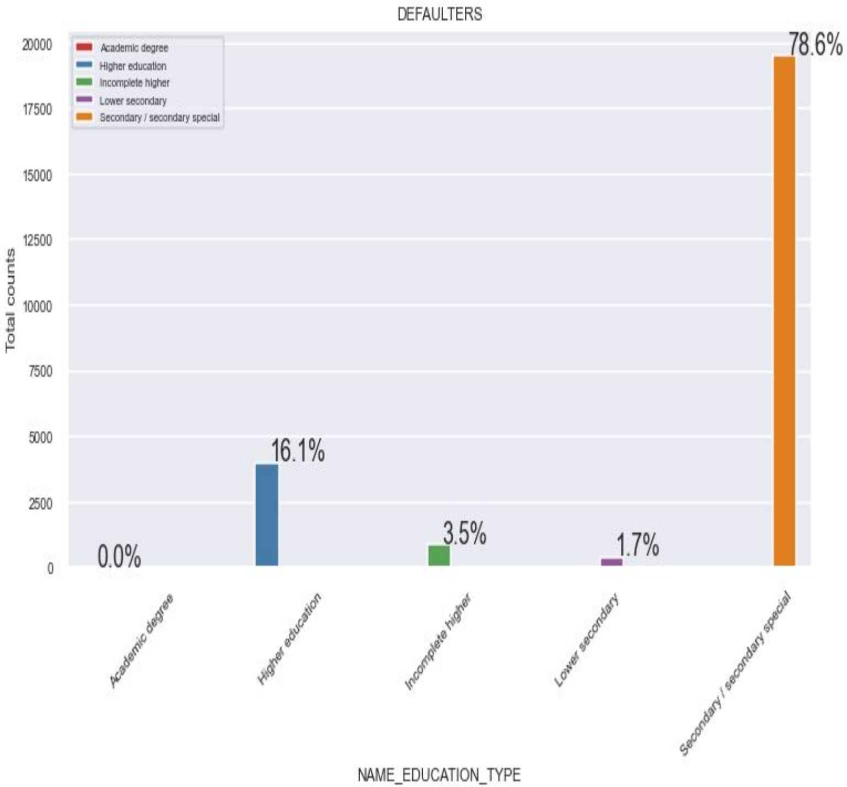
- OCCUPATION_TYPE: The laborers have highest percentage of defaulters and the highest percentage of non-defaulters are also laborers.

- ORGANIZATION_TYPE: Clients who have applied loan are mostly from ‘Business entity Type 3’ , ‘Self employed’ , ‘Other’ , ‘Medicine’ and ‘Government’. The more defaulters are from Business entity Type 3 and second is from Self employed.

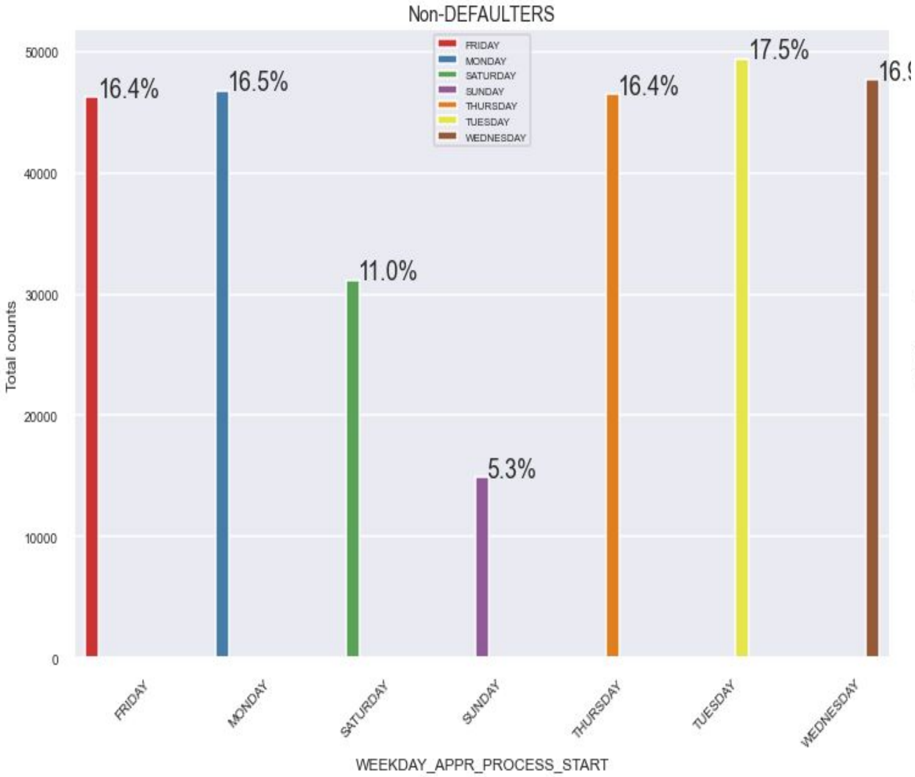
Univariate Categorical Ordered Variables

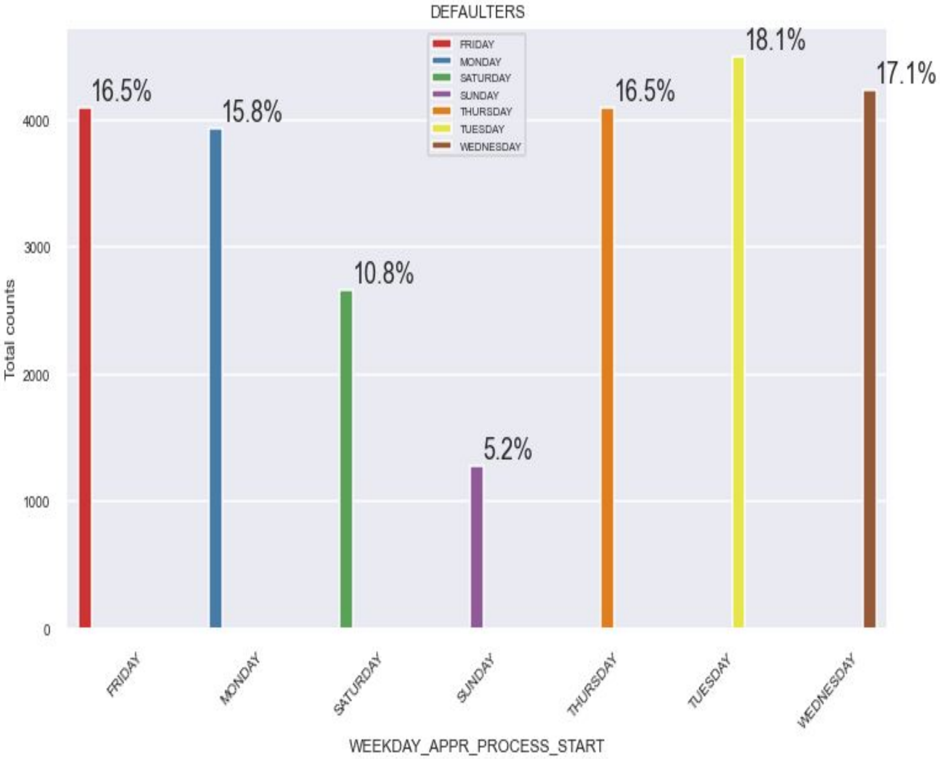
NAME_EDUCATION_TYPE- Non-Defaulters vs. Defaulters



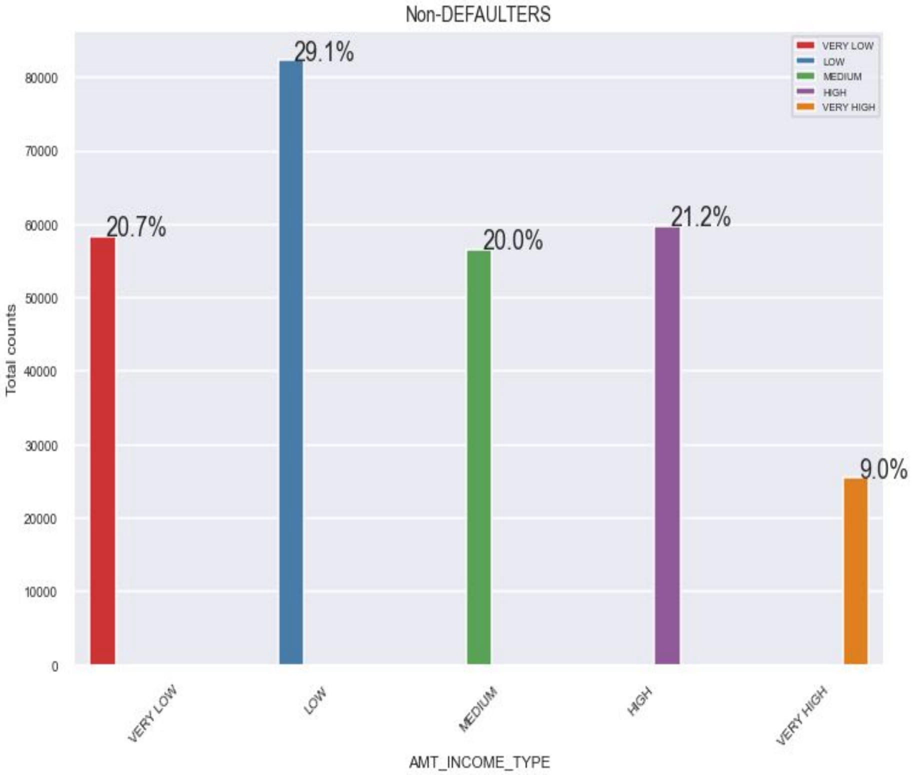


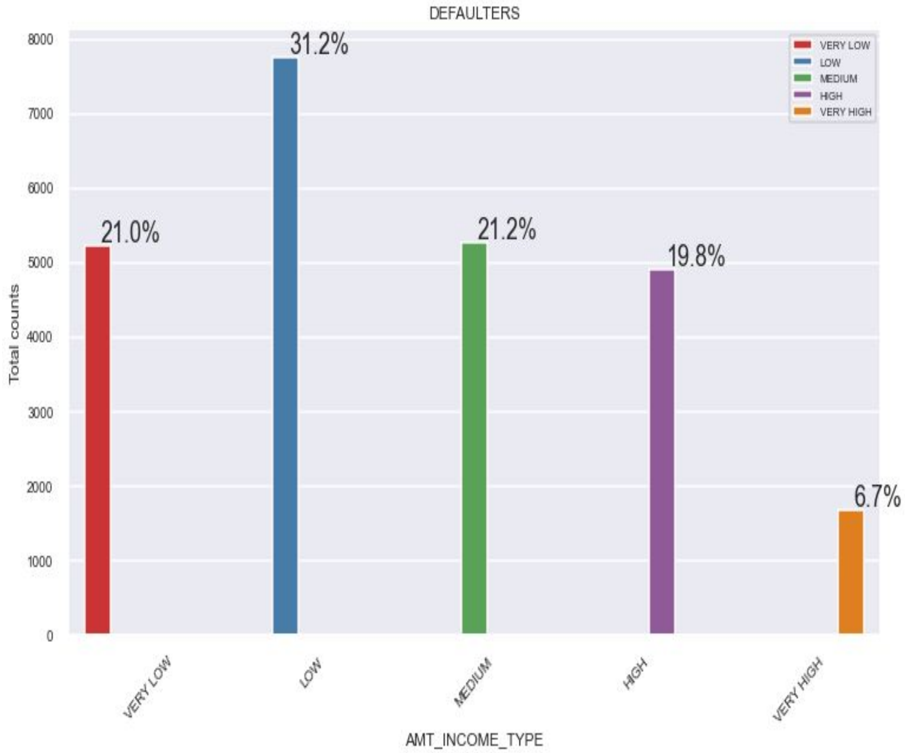
WEEKDAY_APPR_PROCESS_START- Non-Defaulters vs. Defaulters



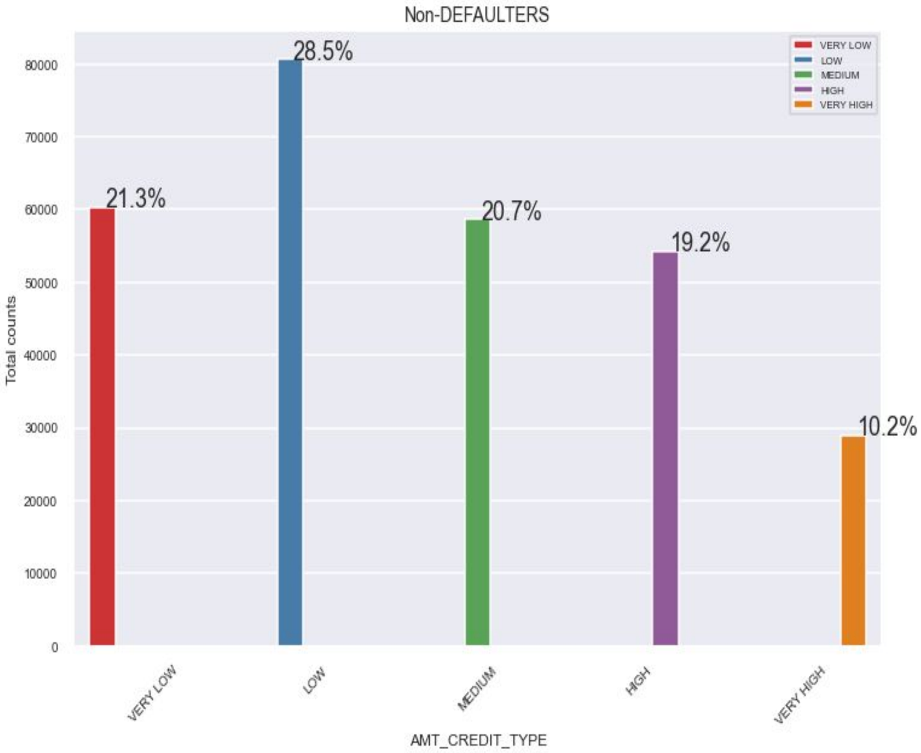


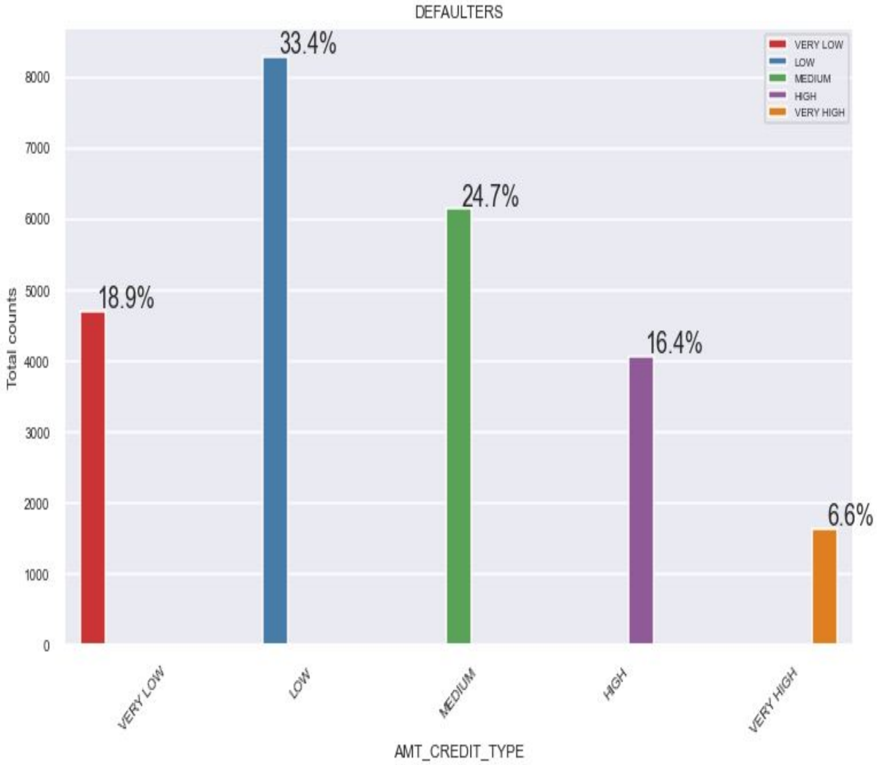
AMT_INCOME_TYPE - Non-Defaulters vs. Defaulters



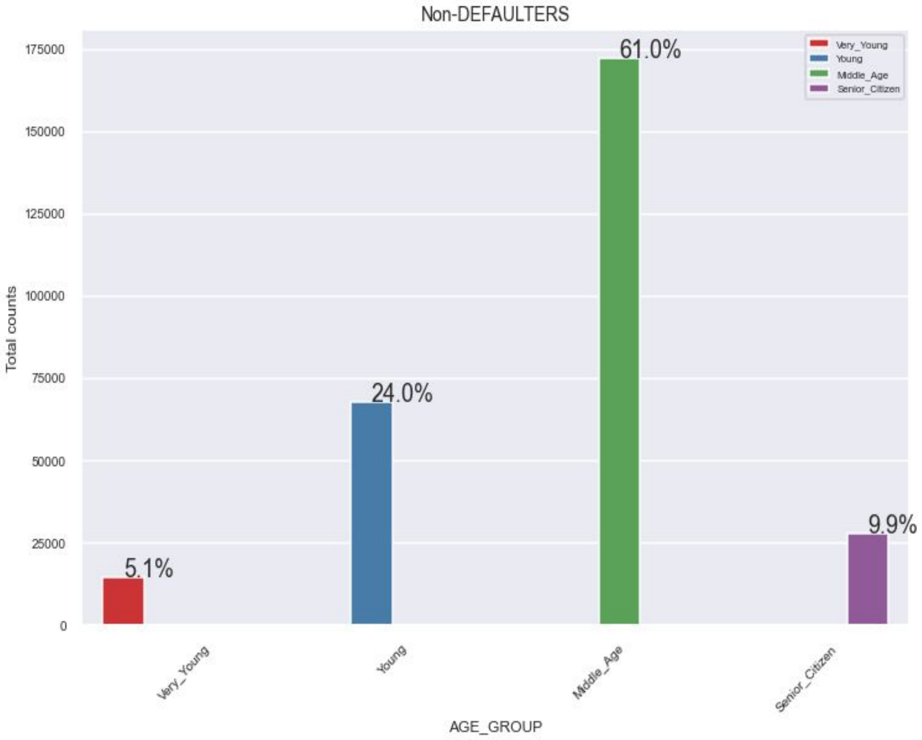


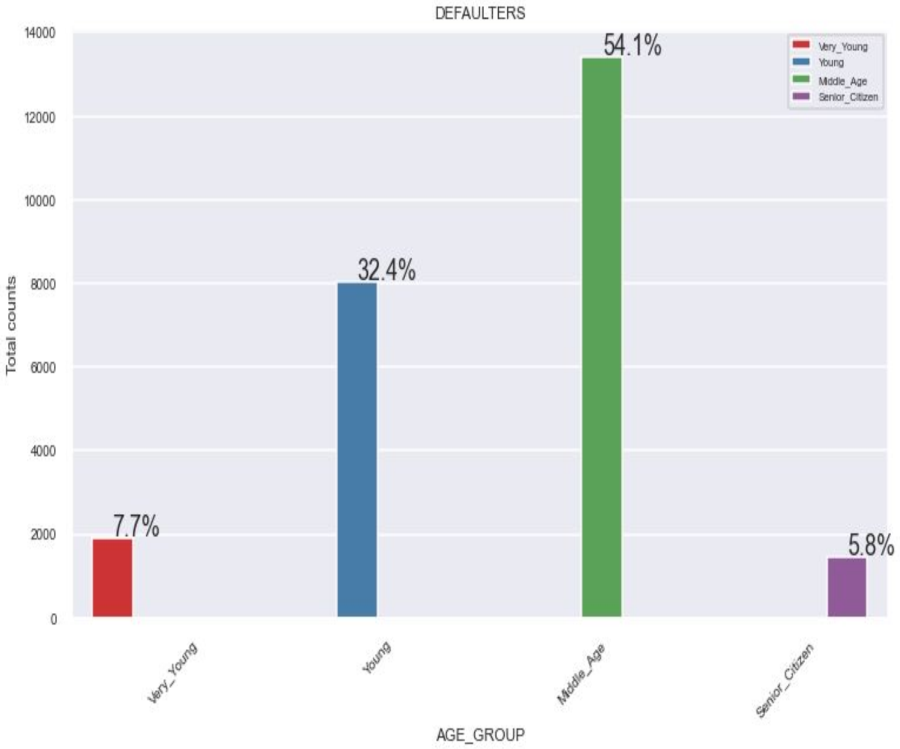
AMT_CREDIT_TYPE - Non-Defaulters vs. Defaulters





AGE_GROUP - Non-Defaulters vs. Defaulters





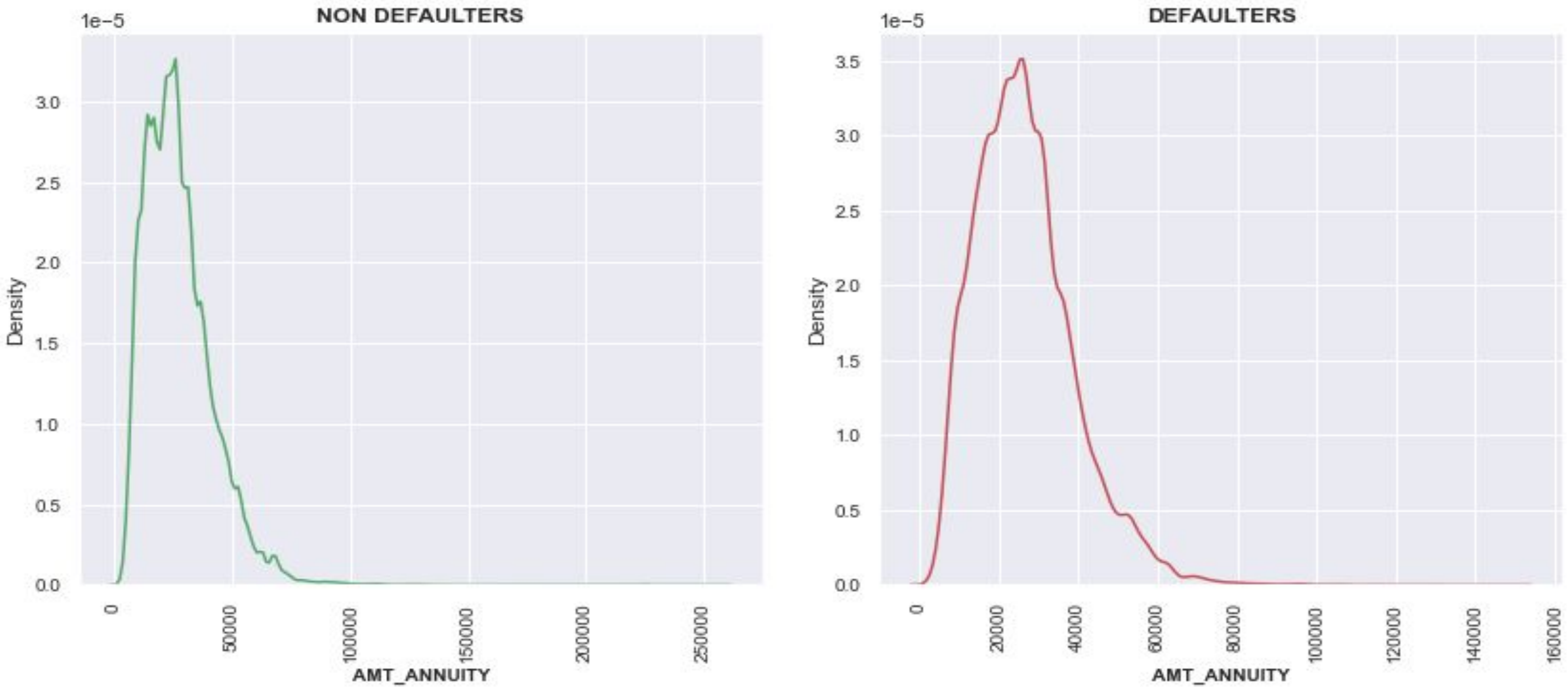
Univariate Categorical Ordered Variables

Insights:

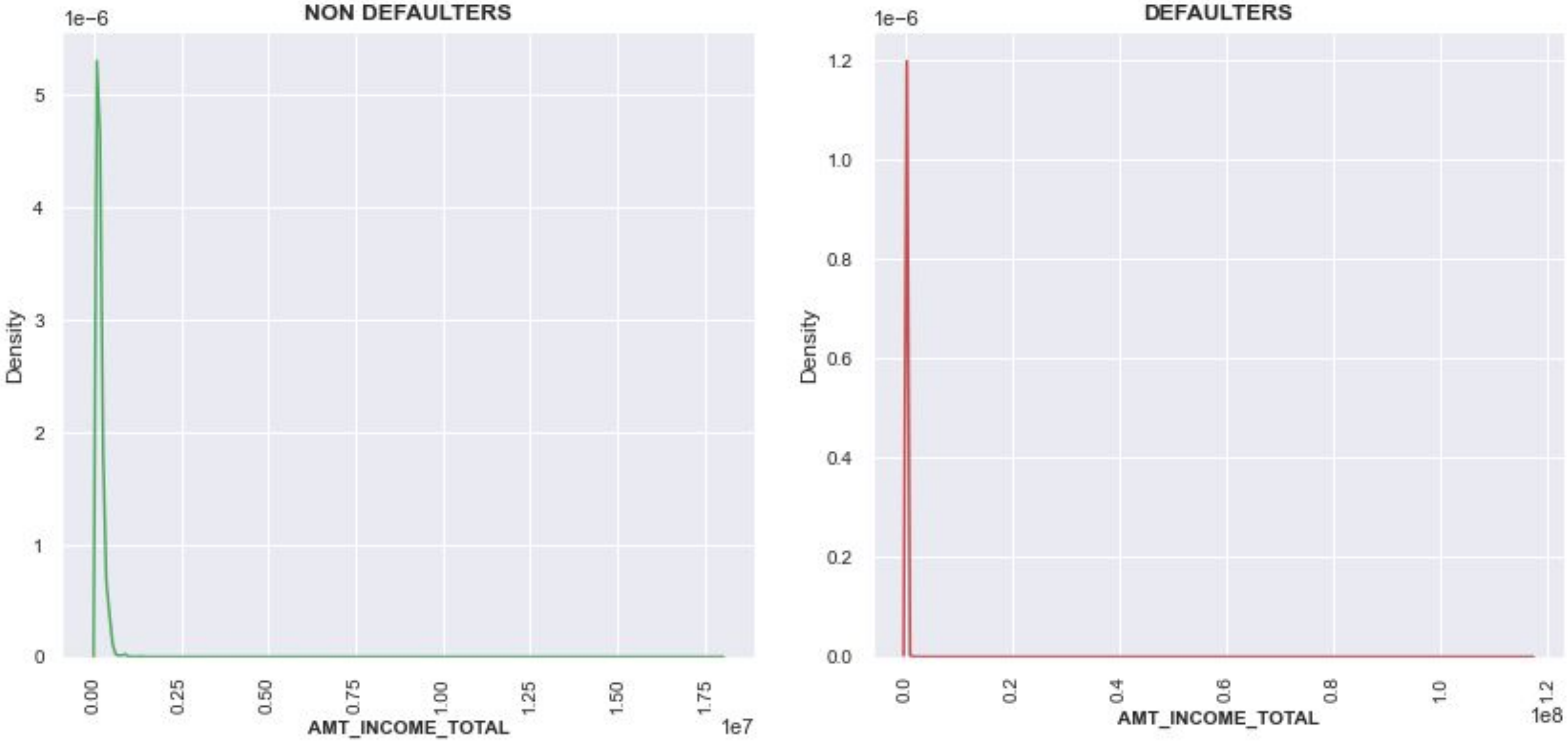
- NAME_EDUCATION_TYPE: Higher education and Secondary/Secondary Special are more likely to apply loan and they are the ones who have high risk to default. Academic degree and lower education are less likely to default.
- WEEKDAY_APPR_PROCESS_START: There is no major difference in days for both defaulters and non-defaulters.
- AMT_INCOME_TYPE: Client having low income are at high risk to default.
- AMT_CREDIT_TYPE: Clients having very low, low and high credit amount are having more default percentage.
- AGE_GROUP: The middle age group tend to default more. The senior citizens are less likey to default.

Univariate Numerical Variable Analysis

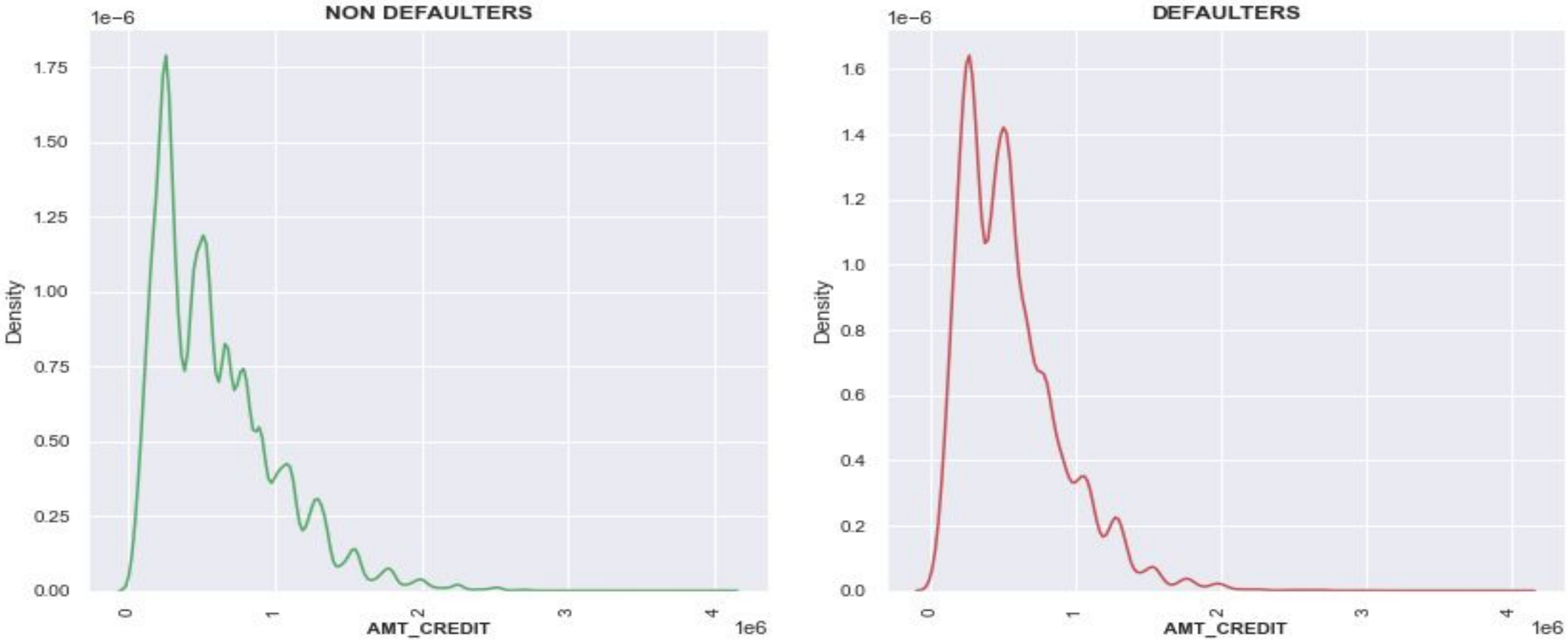
AMT_ANNUITY - Non-Defaulters vs. Defaulters



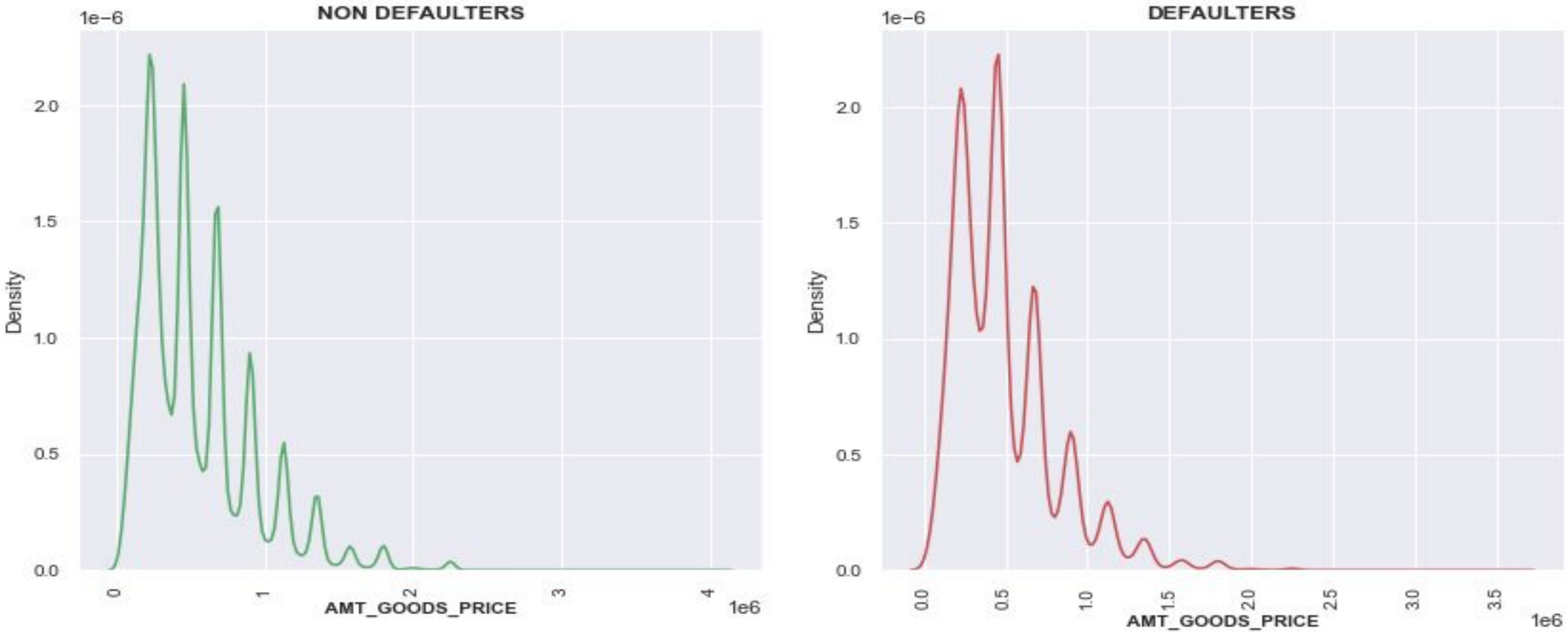
AMT_INCOME TOTAL - Non-Defaulters vs. Defaulters



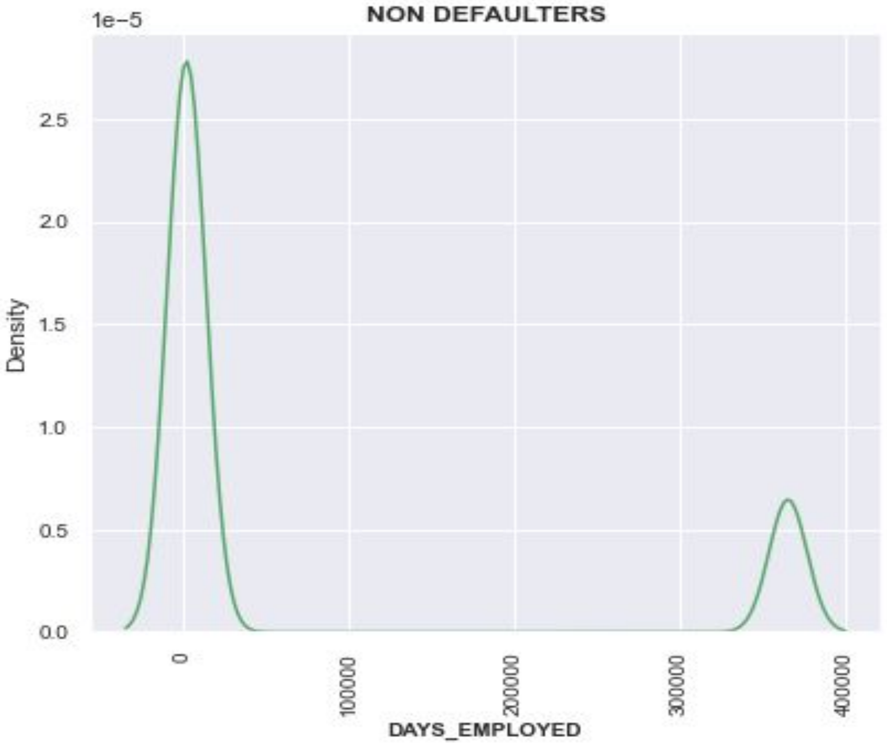
AMT_CREDIT - Non-Defaulters vs. Defaulters

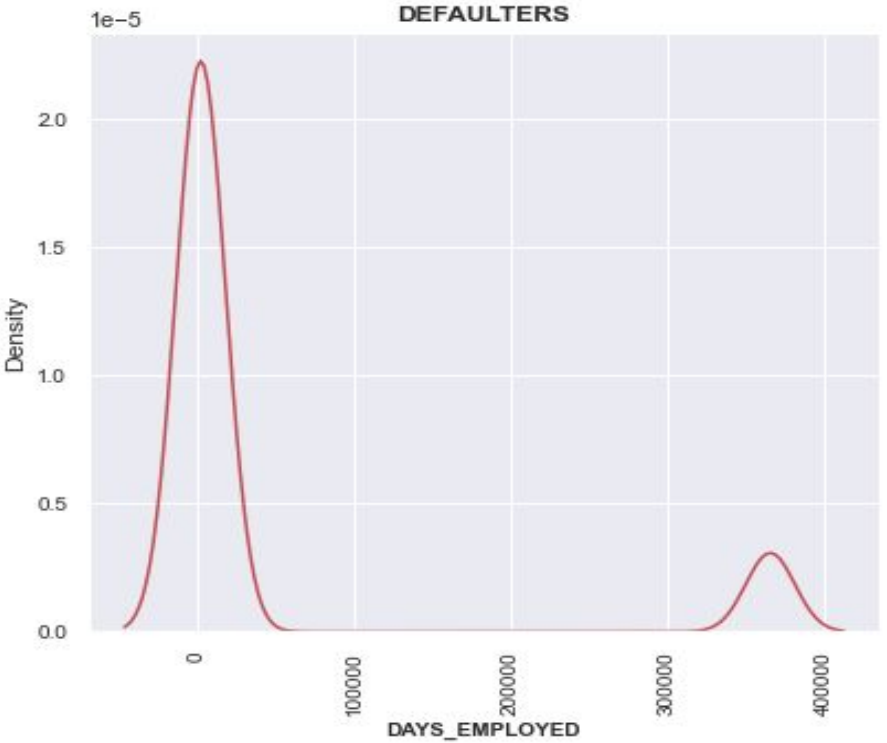


AMT_GOODS_PRICE - Non-Defaulters vs. Defaulters

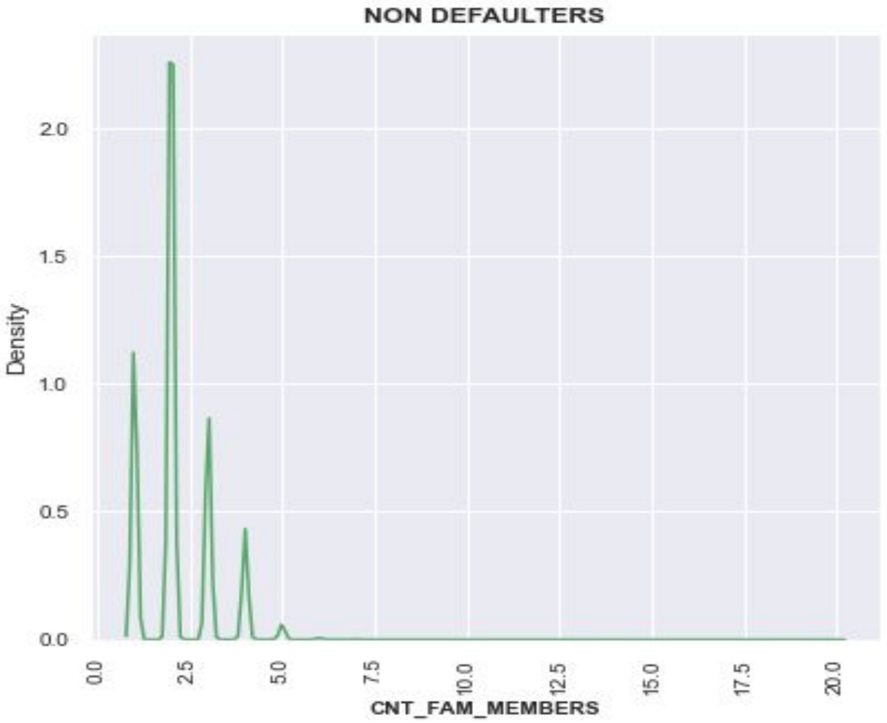


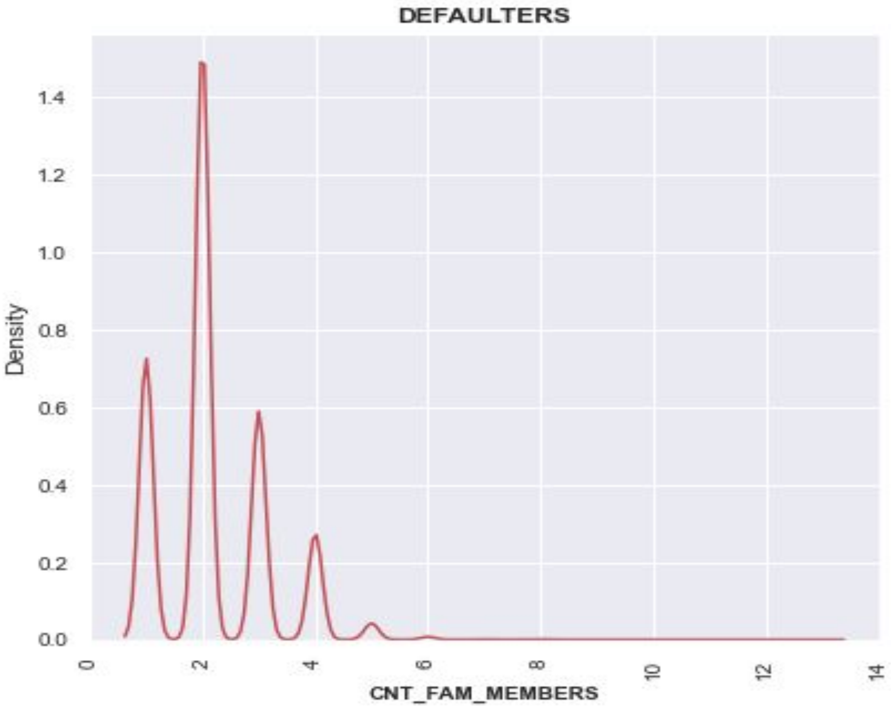
DAYS_EMPLOYED - Non-Defaulters vs. Defaulters

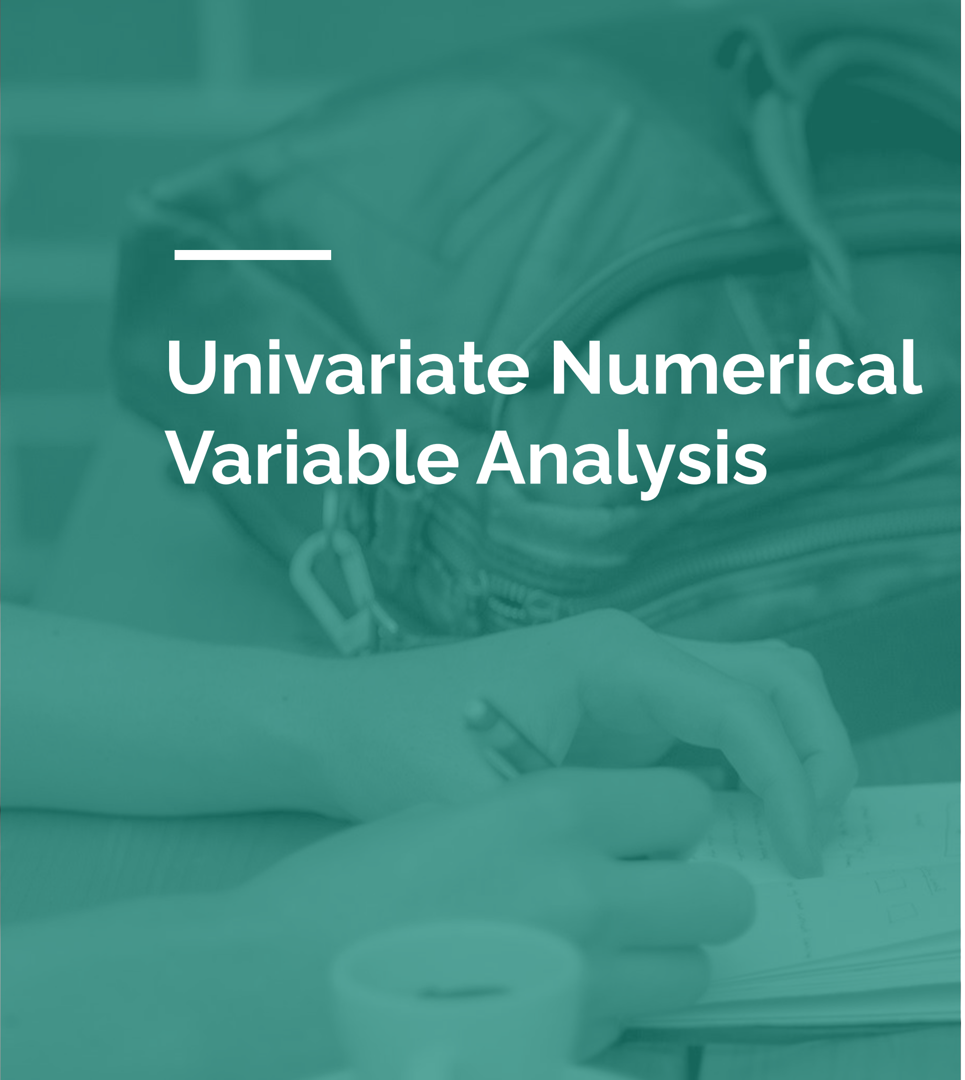




CNT_FAM_MEMBERS - Non-Defaulters vs. Defaulters







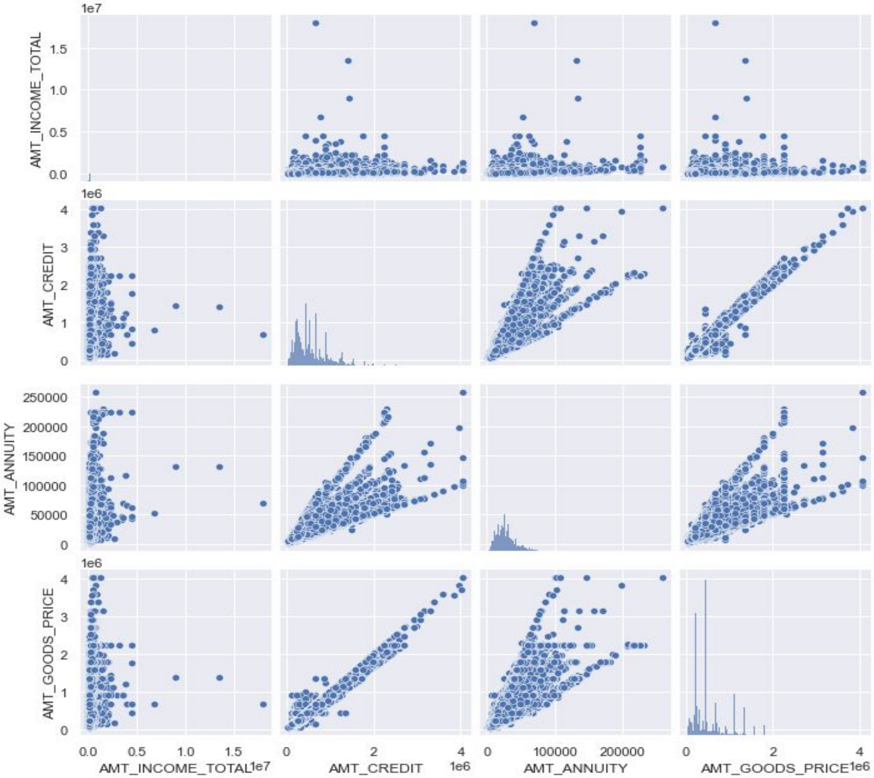
Univariate Numerical Variable Analysis

Insights:

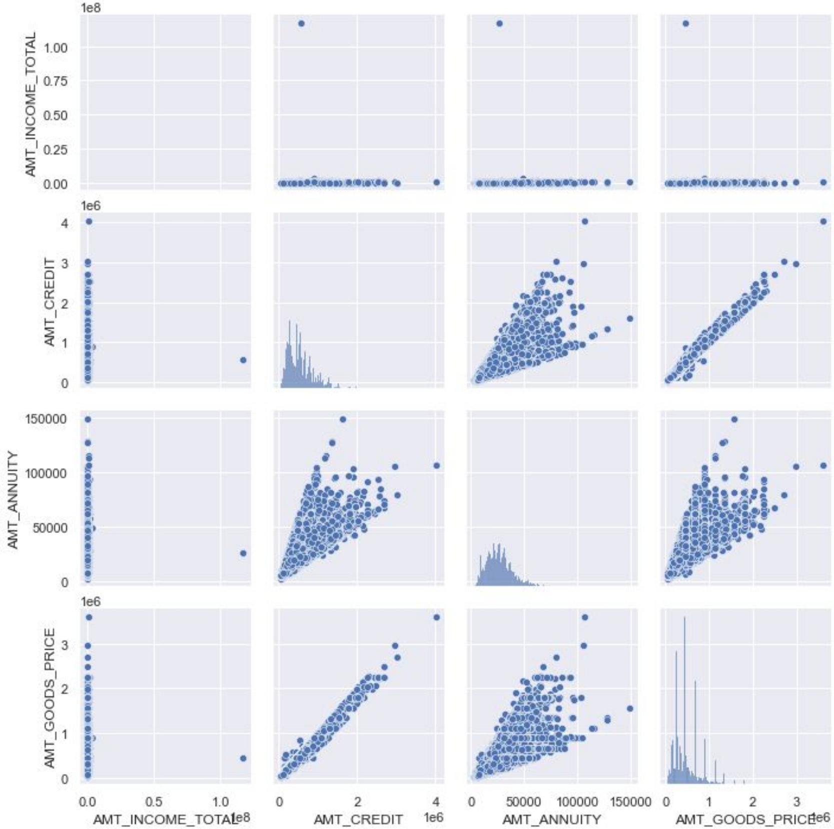
- Clients having more family members are at high risk to default.
- Also clients having more than 4 children are more likely to default.
- Most of the clients have no children

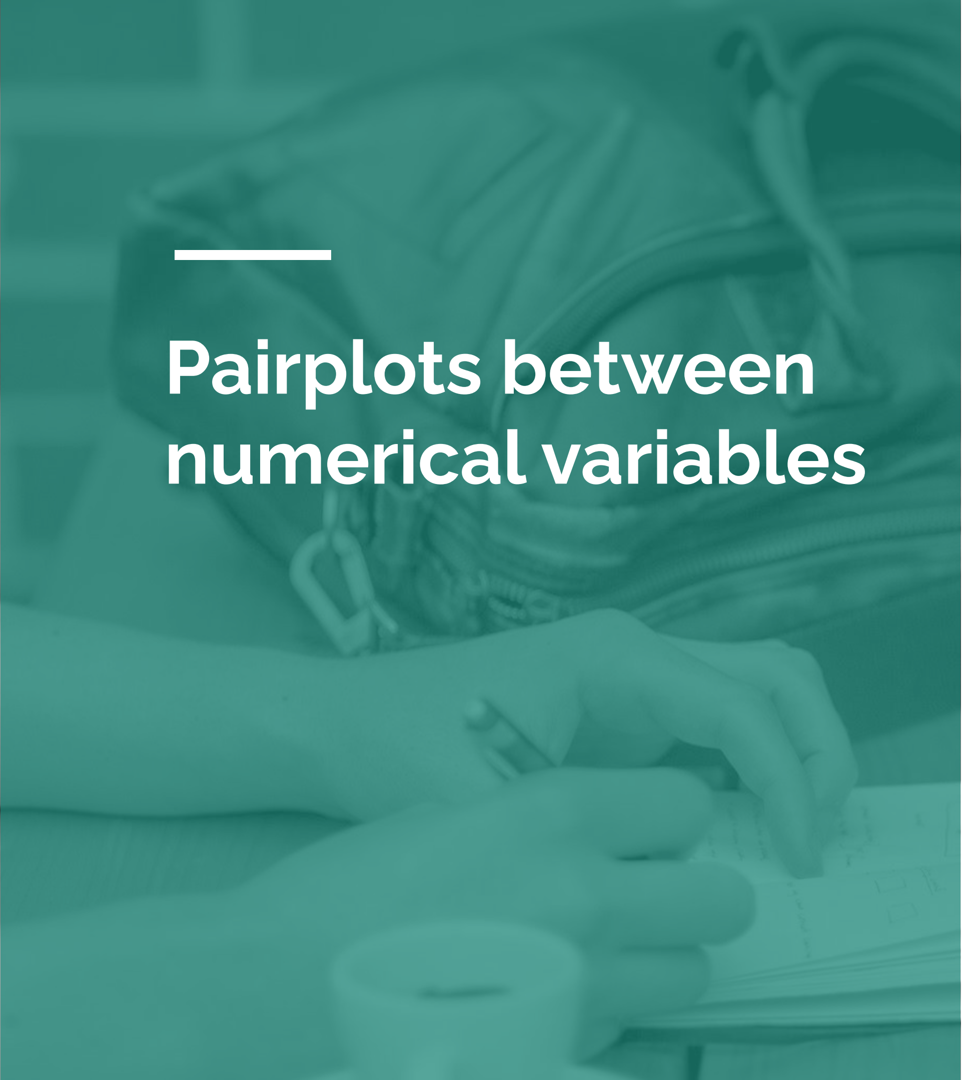
Pairplots between numerical variables

NON-DEFAULTERS



DEFAULTERS


1e8
AMT_INCOME_TOTAL
1.00
0.75
0.50
0.25
0.00
1e6
AMT_CREDIT
4
3
2
1
0
AMT_ANNUITY
150000
100000
50000
0
1e6
AMT_GOODS_PRICE
3
2
1
0
0.0
0.5
1.0
AMT_INCOME_TOTAL
1e8
0
2
4
AMT_CREDIT
1e6
0
50000
100000
150000
AMT_ANNUITY
0
1
2
3
AMT_GOODS_PRICE
1e6



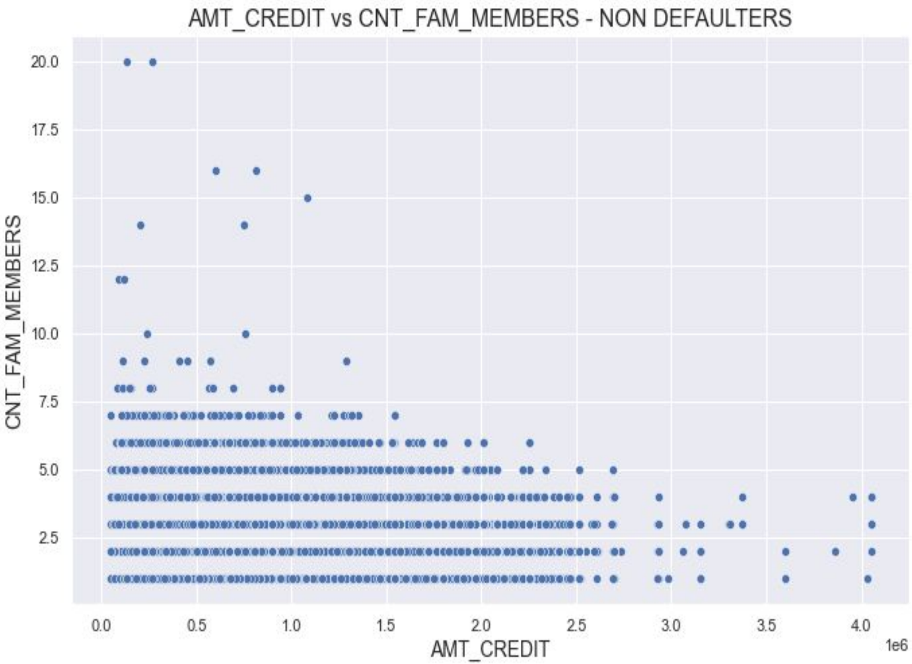
Pairplots between numerical variables

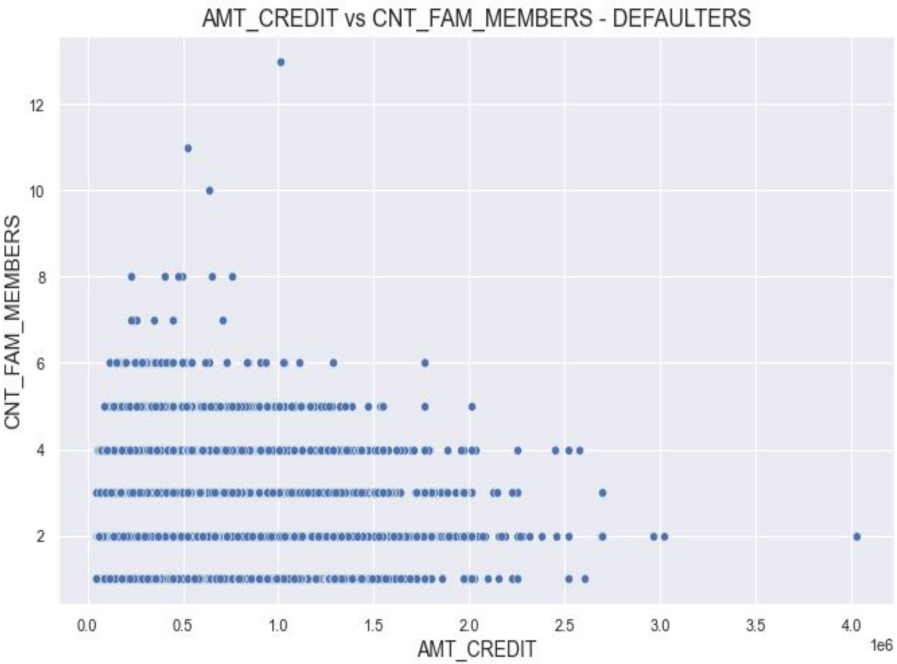
Insights:

- AMT_CREDIT and AMT_GOODS_PRICE are highly correlated variables for both defaulters and non–defaulters and they show almost a line in the plot.
- AMT_CREDIT and AMT_ANNUITY are also highly correlated variables for both defaulters and non–defaulters

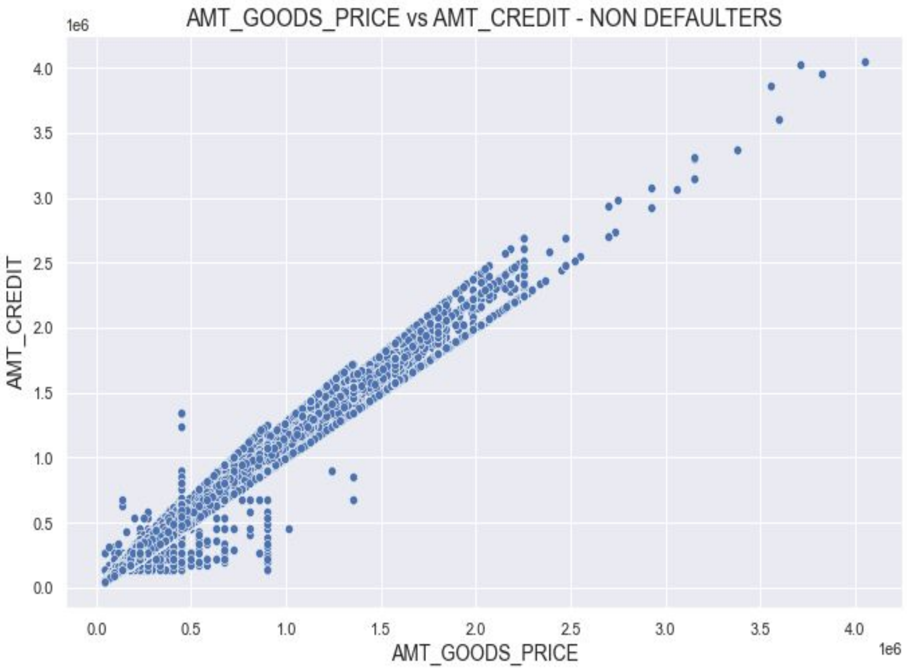
Bivariate Numeric - Numeric Analysis

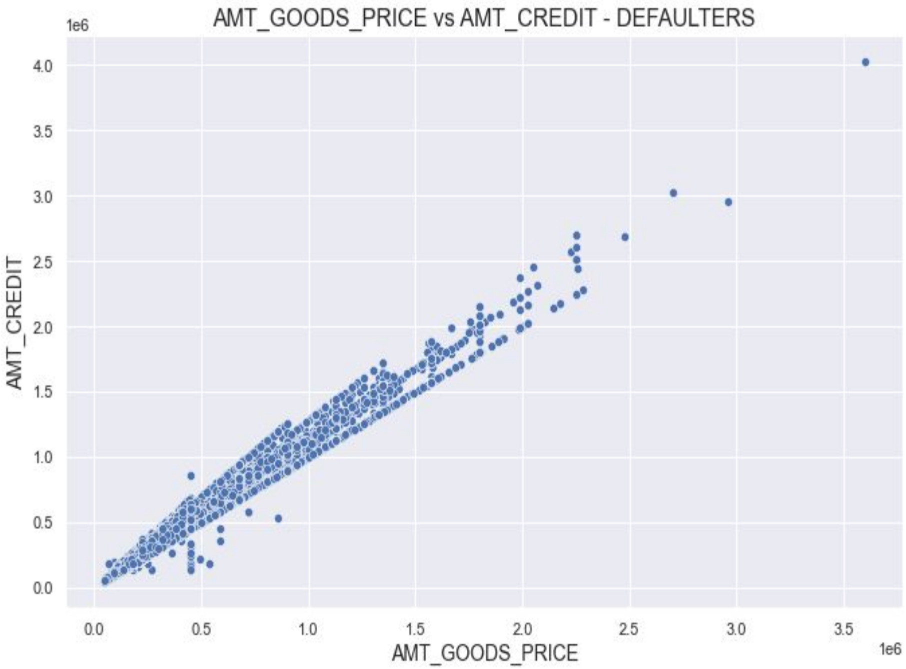
NAME_EDUCATION_TYPE vs. CNT_FAM_MEMBERS - Non-Defaulters and Defaulters



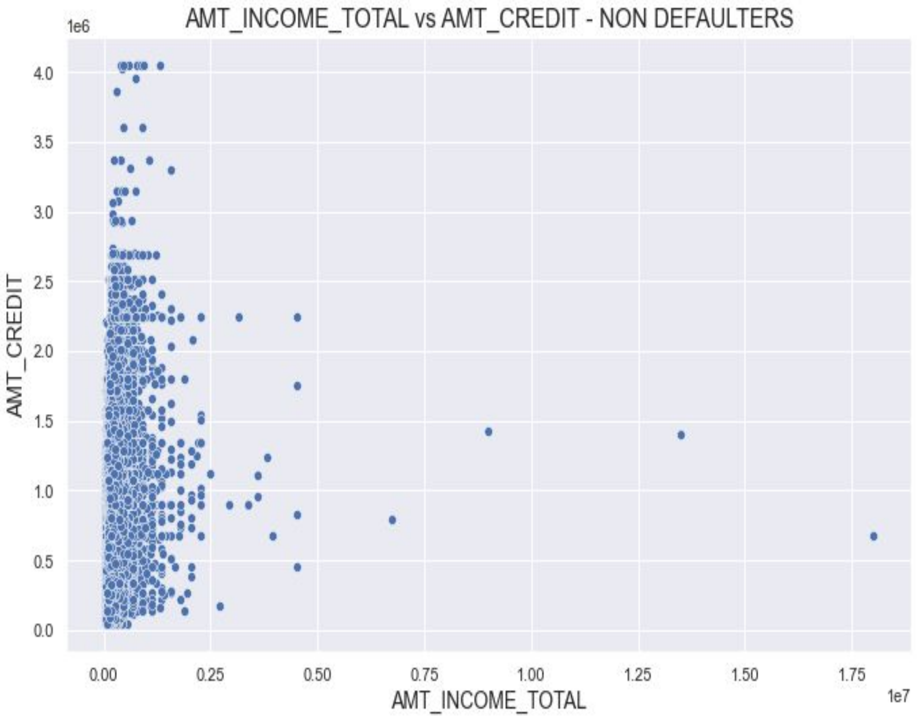


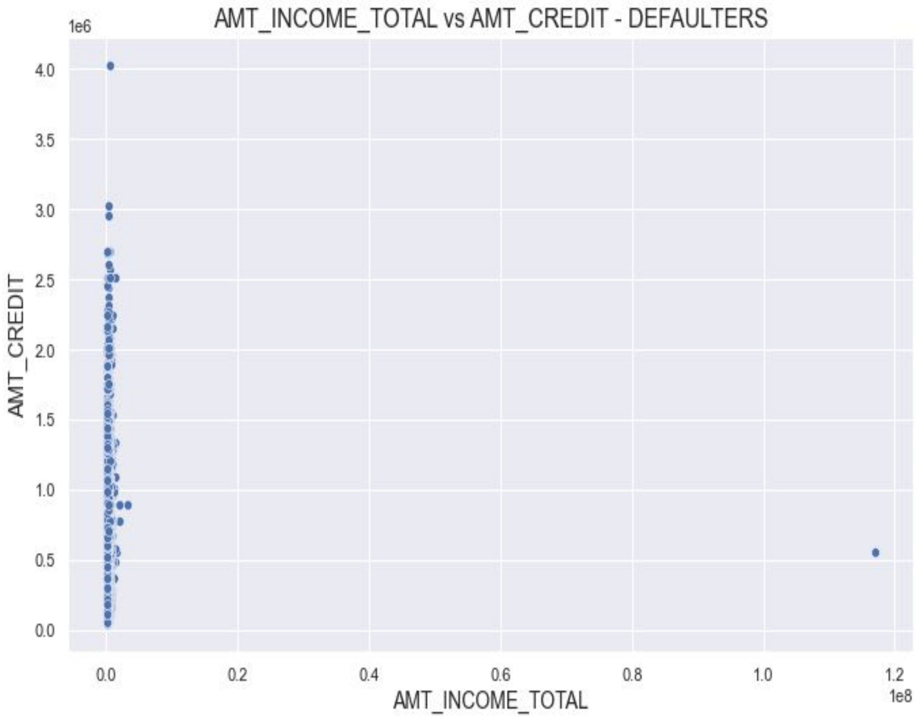
AMT_GOODS_PRICE vs. AMT_CREDIT - Non-Defaulters and Defaulters

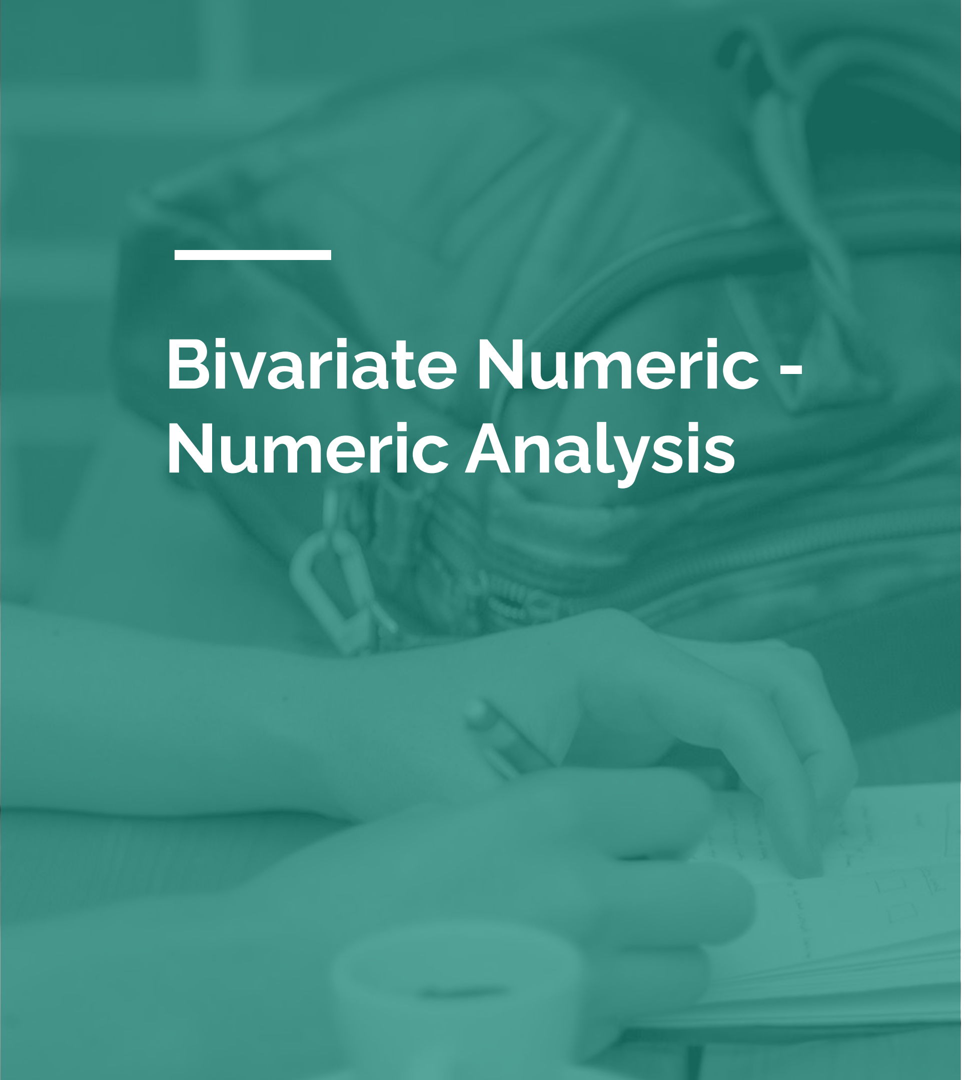




AMT_INCOME_TOTAL vs. AMT_CREDIT - Non-Defaulters and Defaulters







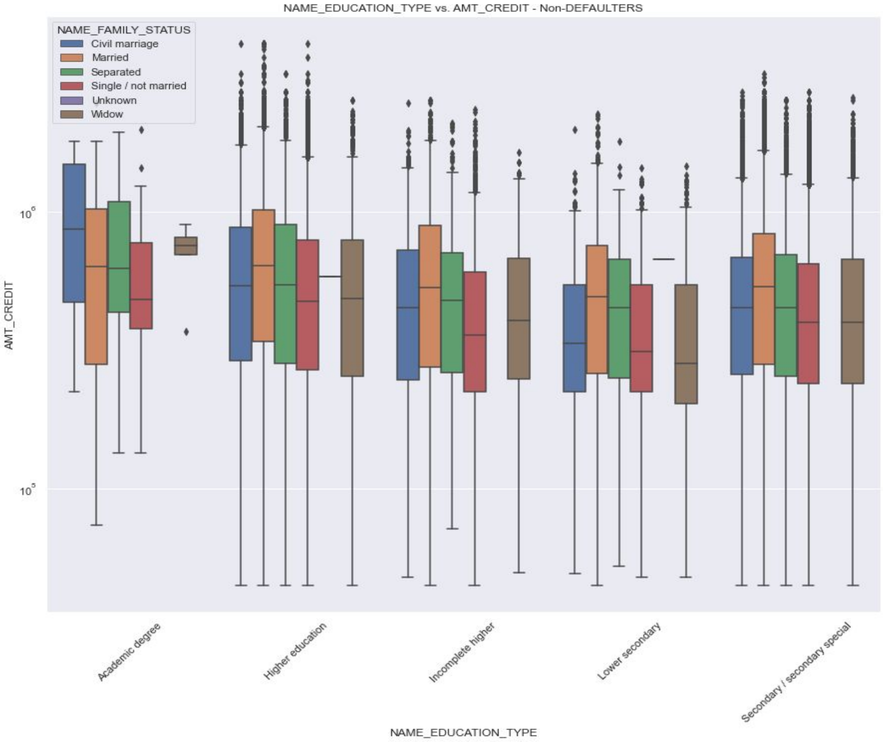
Bivariate Numeric - Numeric Analysis

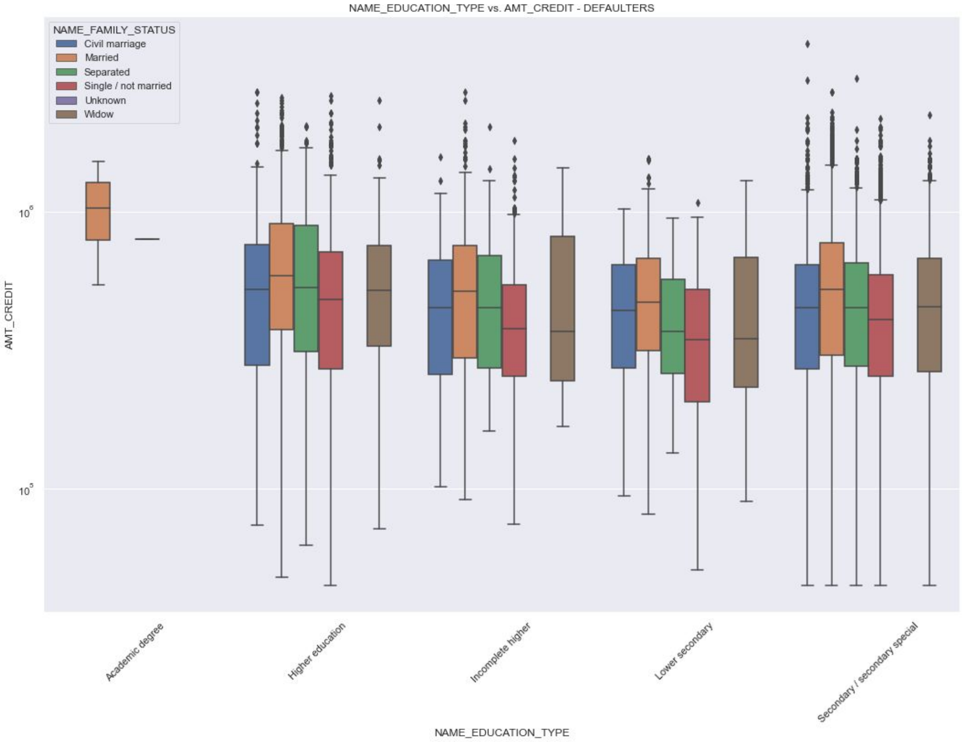
Insights:

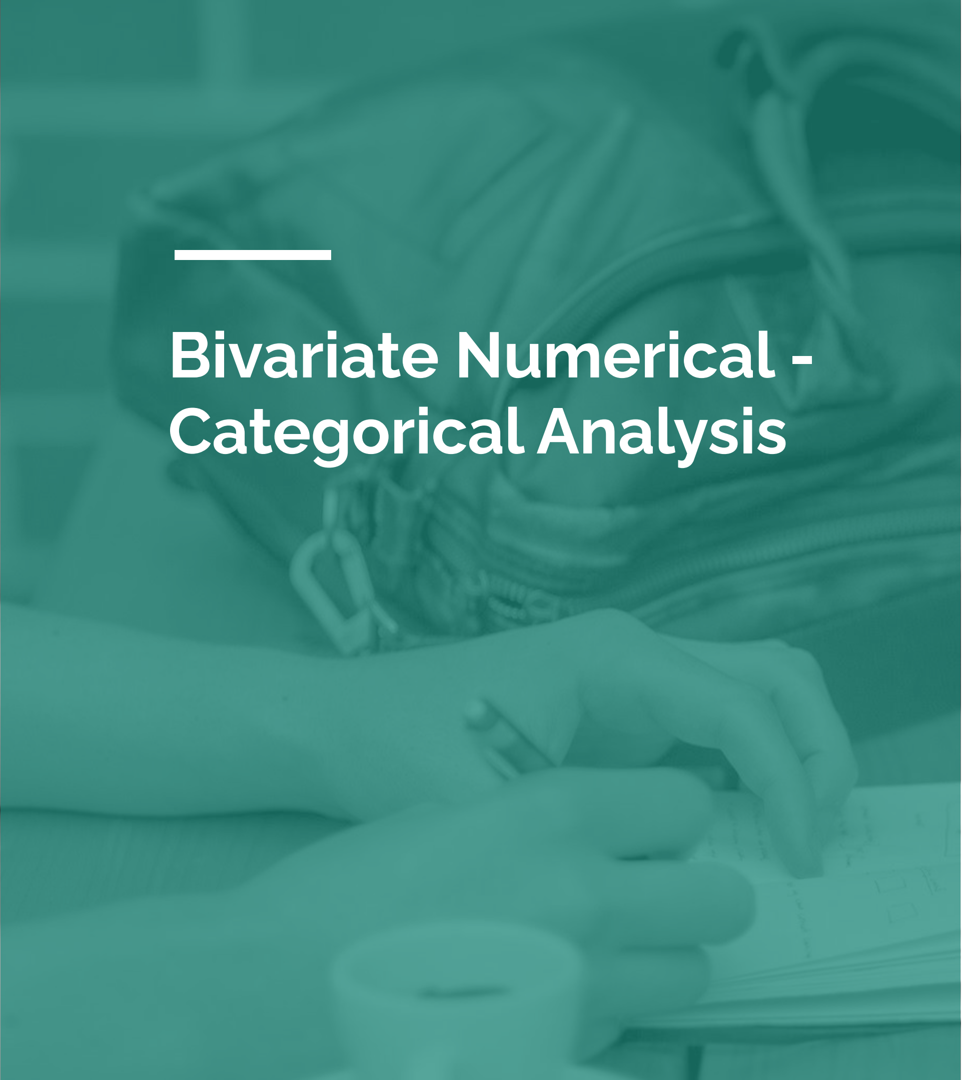
- The non-defaulters are more likely to get higher credits than defaulters.
- Non-defaulters who have higher goods price have higher credits than those with higher goods price but didnt pay loan.
- The density in the lower left corner is same in both non-defaulters and defaulters. Hence the people are equally likely to default if the family is small and the AMT_CREDIT is low.
- Also larger families and people with larger AMT_CREDIT default less often.

Bivariate Numerical - Categorical Analysis

NAME_EDUCATION_TYPE vs. AMT_CREDIT - Non-Defaulters and Defaulters







Bivariate Numerical - Categorical Analysis

NAME_EDUCATION_TYPE vs. AMT_CREDIT - Non-Defaulters and Defaulters

Insights:

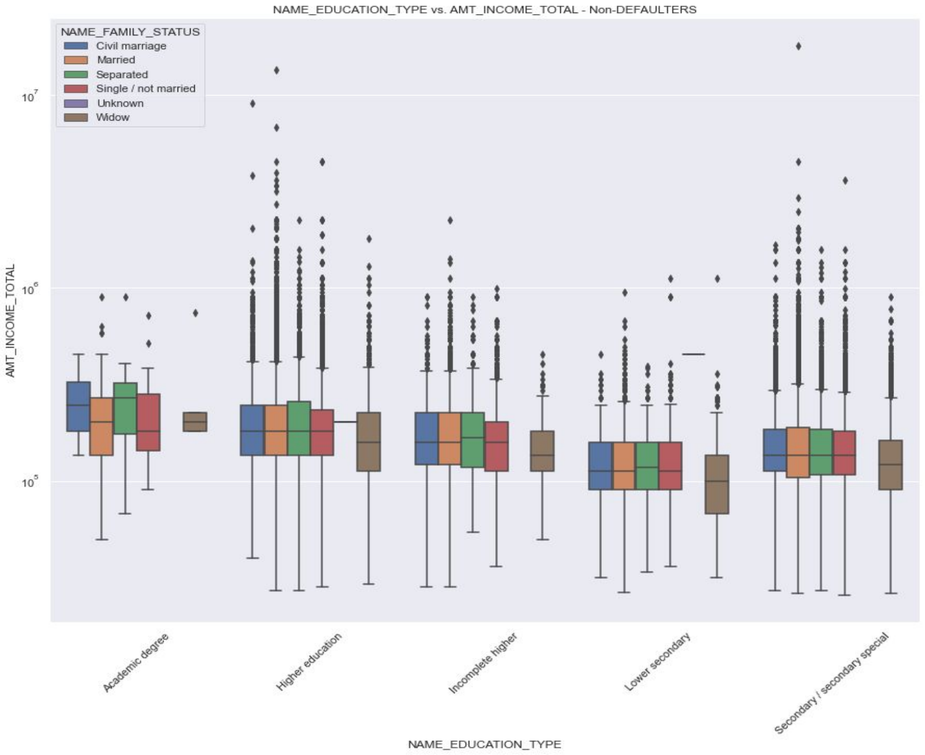
Non-defaulters:

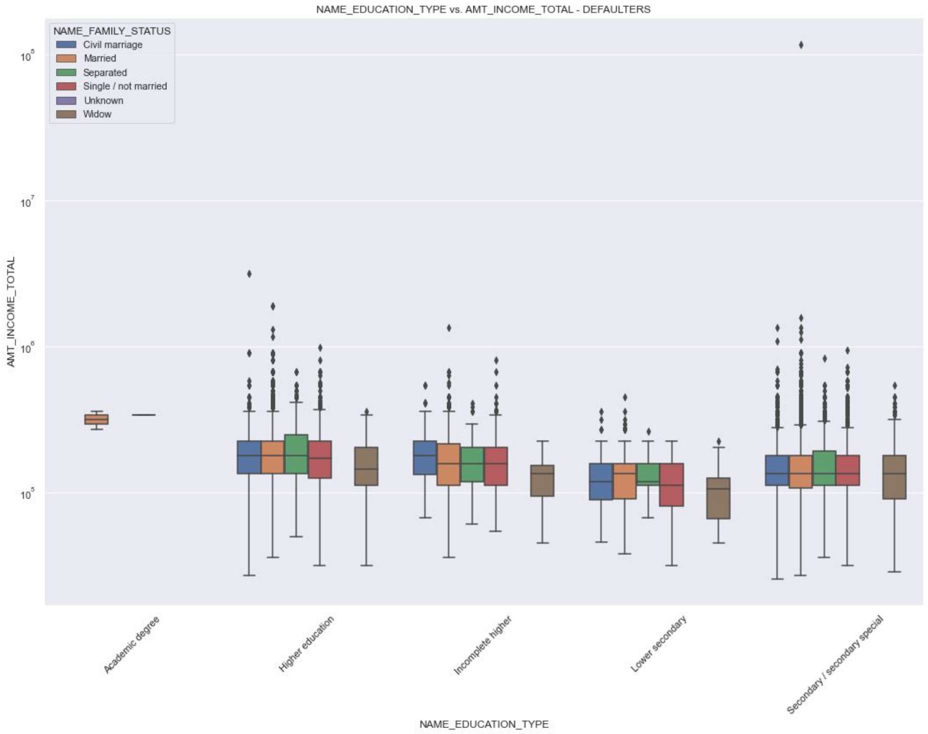
- Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.
- Also the higher education of civil marriage, married and single are having more outliers.
- Customers with less education have lower credit limits, with widows having the lowest.

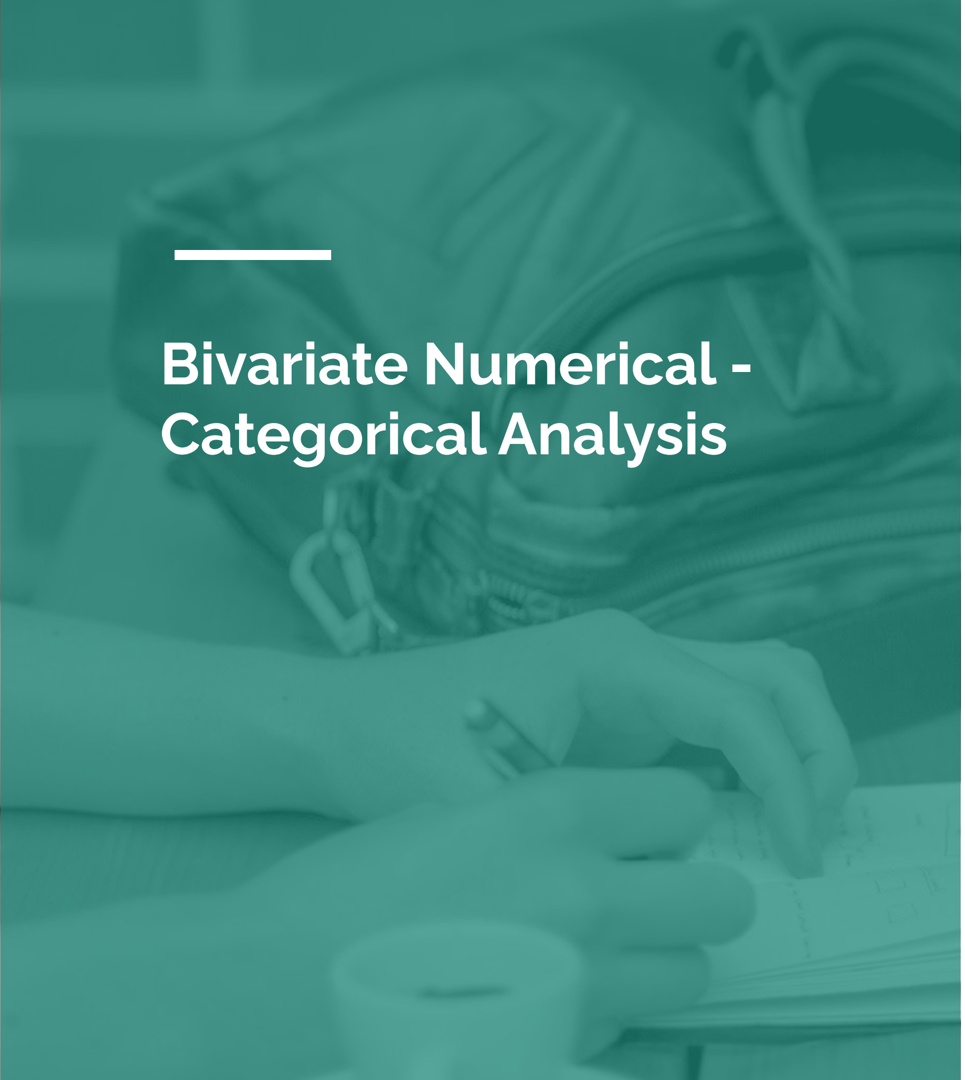
Defaulters:

- Customer with Academic degree education are having higher number of credits than others.
- Family status of 'civil marriage', 'marriage' and 'separated' of Secondary/Secondary special are having more outliers.
- Customers with less education have lower credit limits.

NAME_EDUCATION_TYPE vs. AMT_INCOME_TOTAL - Non-Defaulters and Defaulters







Bivariate Numerical - Categorical Analysis

NAME_EDUCATION_TYPE vs. AMT_INCOME_TOTAL - Non-Defaulters and Defaulters

Insights:

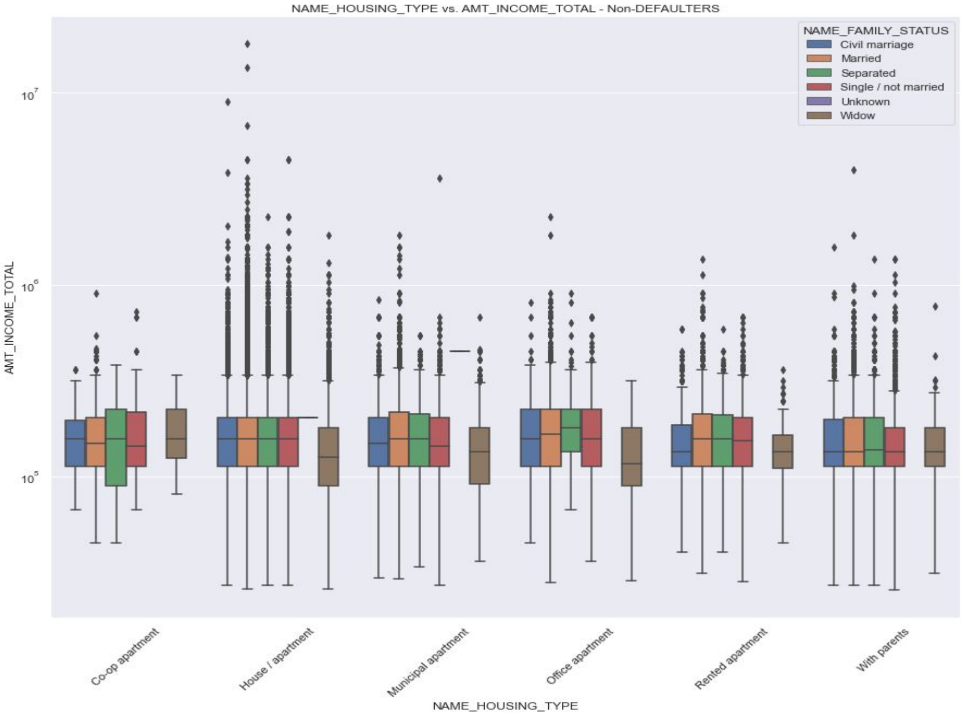
Non-defaulters:

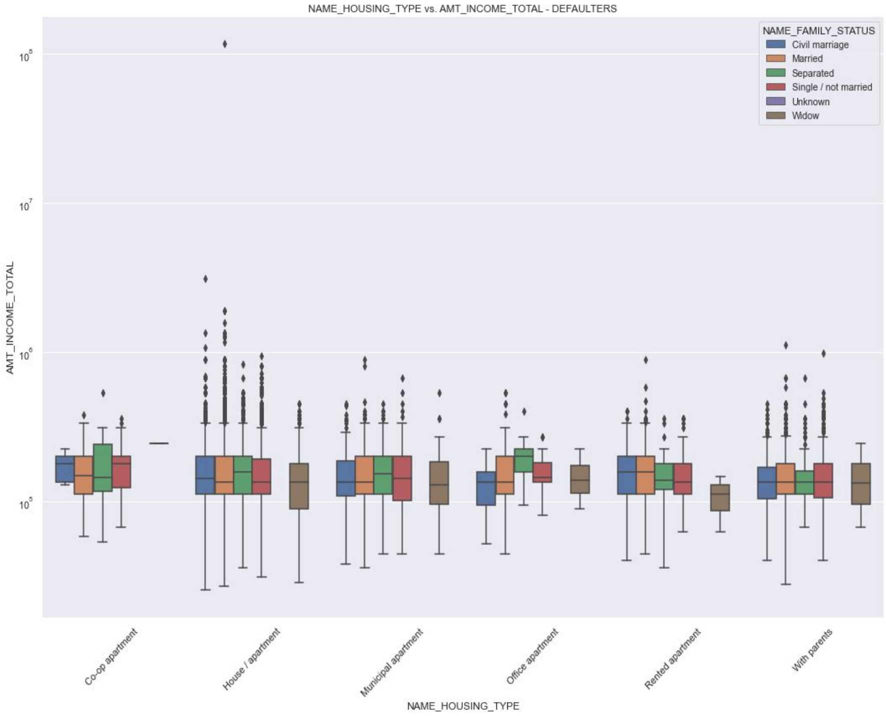
- Family status of civil marriage and separated of Academic degree education are having higher income than others.
- Also the higher education and Secondary/Secondary special of married are having highest outliers.
- Customers with less education have lower income, with widows having the lowest.

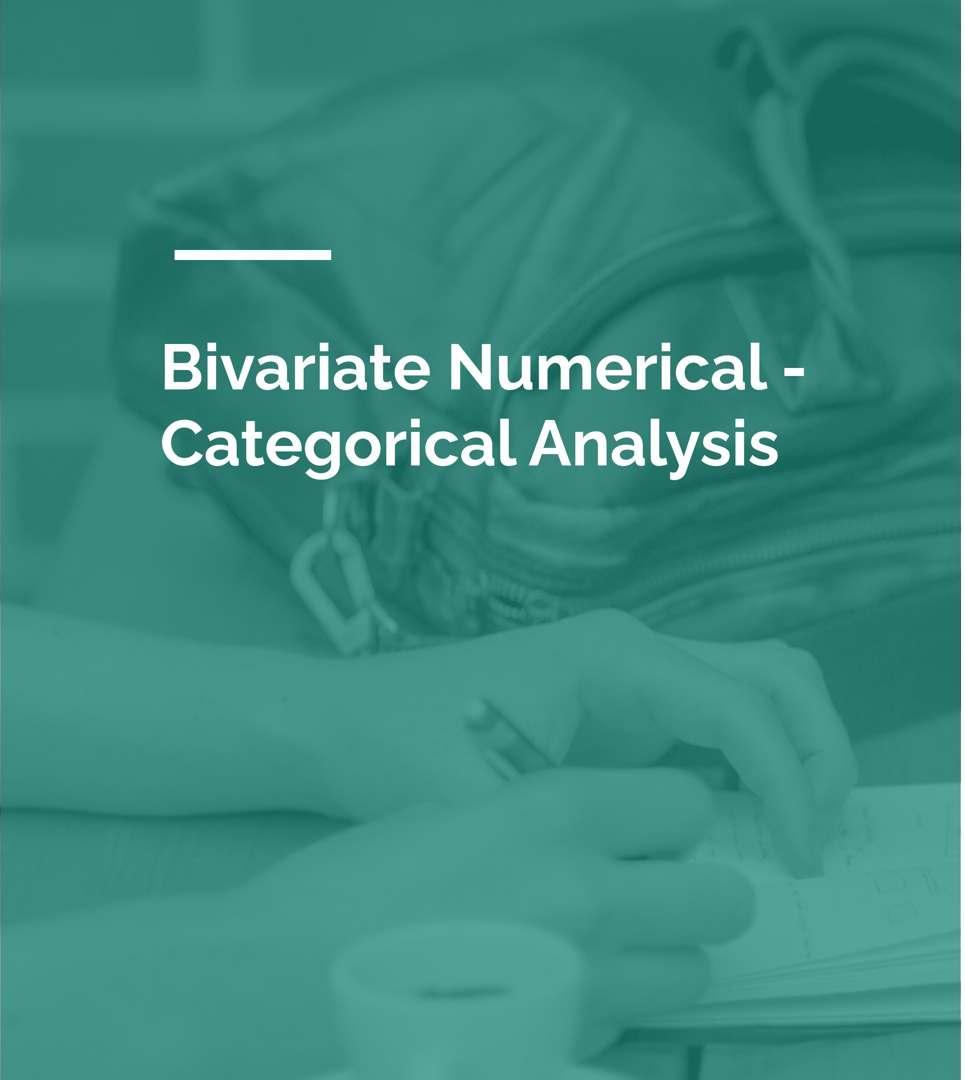
Defaulters:

- Customer with Academic degree education are having higher income than others.
- Family status of marriage of Secondary/Secondary special is having highest outliers.
- Customers with less education have lower income, with widows having the lowest.

NAME_HOUSING_TYPE vs. AMT_INCOME_TOTAL - Non-Defaulters and Defaulters







Bivariate Numerical - Categorical Analysis

NAME_HOUSING_TYPE vs. AMT_INCOME_TOTAL - Non-Defaulters and Defaulters

Insights:

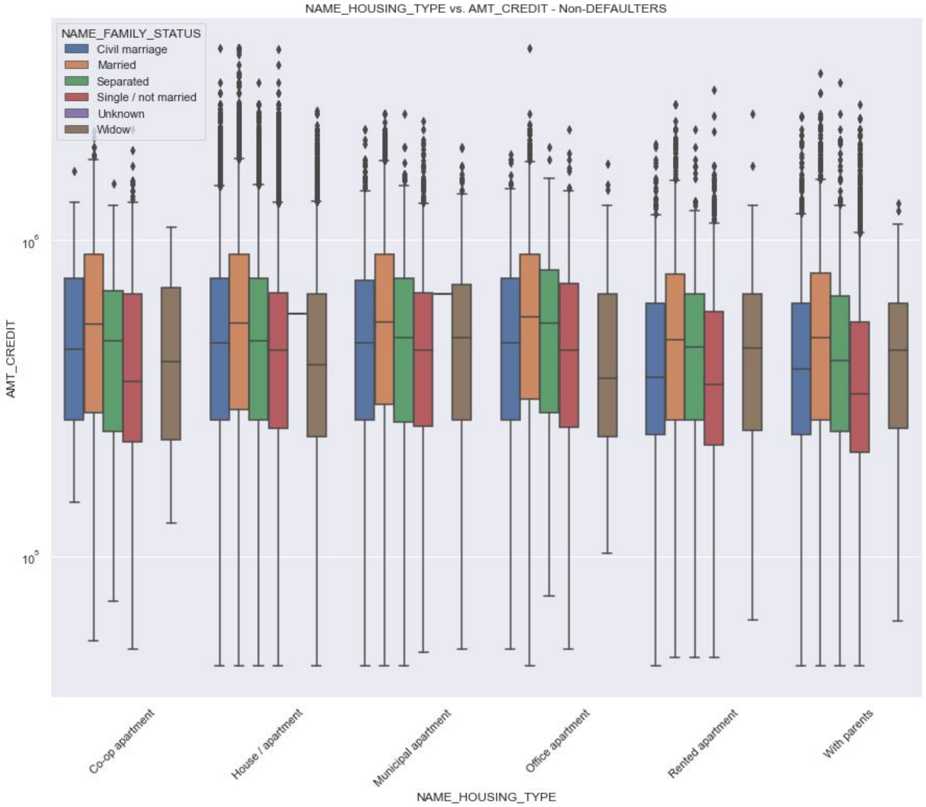
Non-defaulters:

- Family status of separated of co-op apartment are having higher income than others.
- Also the house/apartment of married are having highest outliers.
- Rented apartment customer especially widow have lower income.

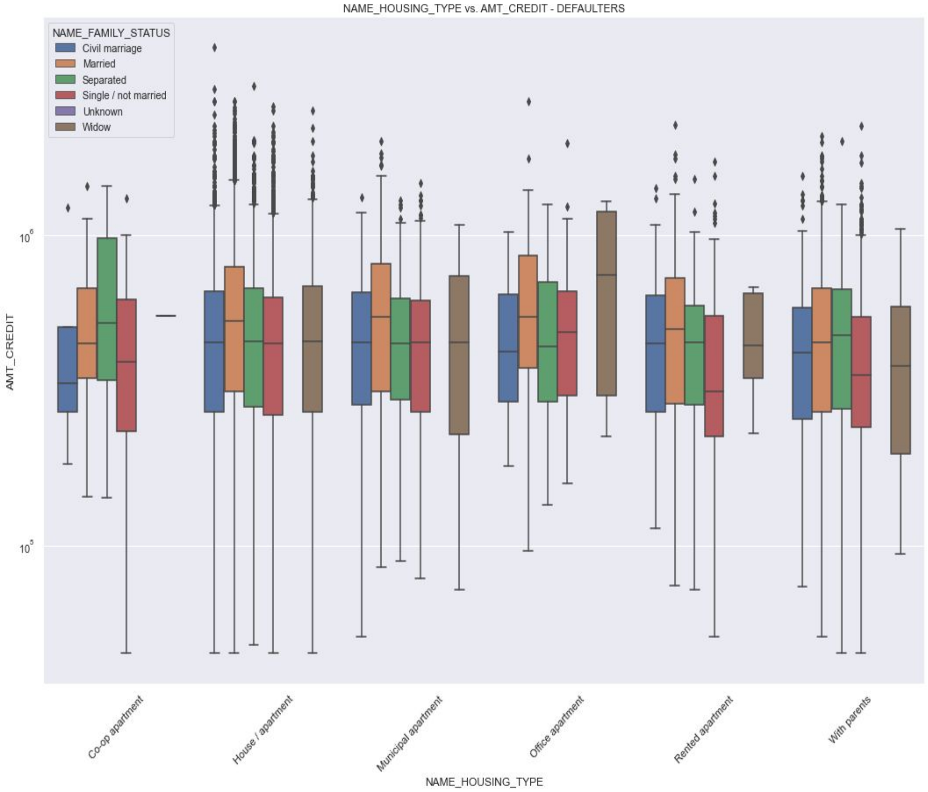
Defaulters:

- Family status of separated of co-op apartment are having higher income than others.
- Also the house/apartment of married are having highest outliers.
- Rented apartment customer especially widow have lower income.

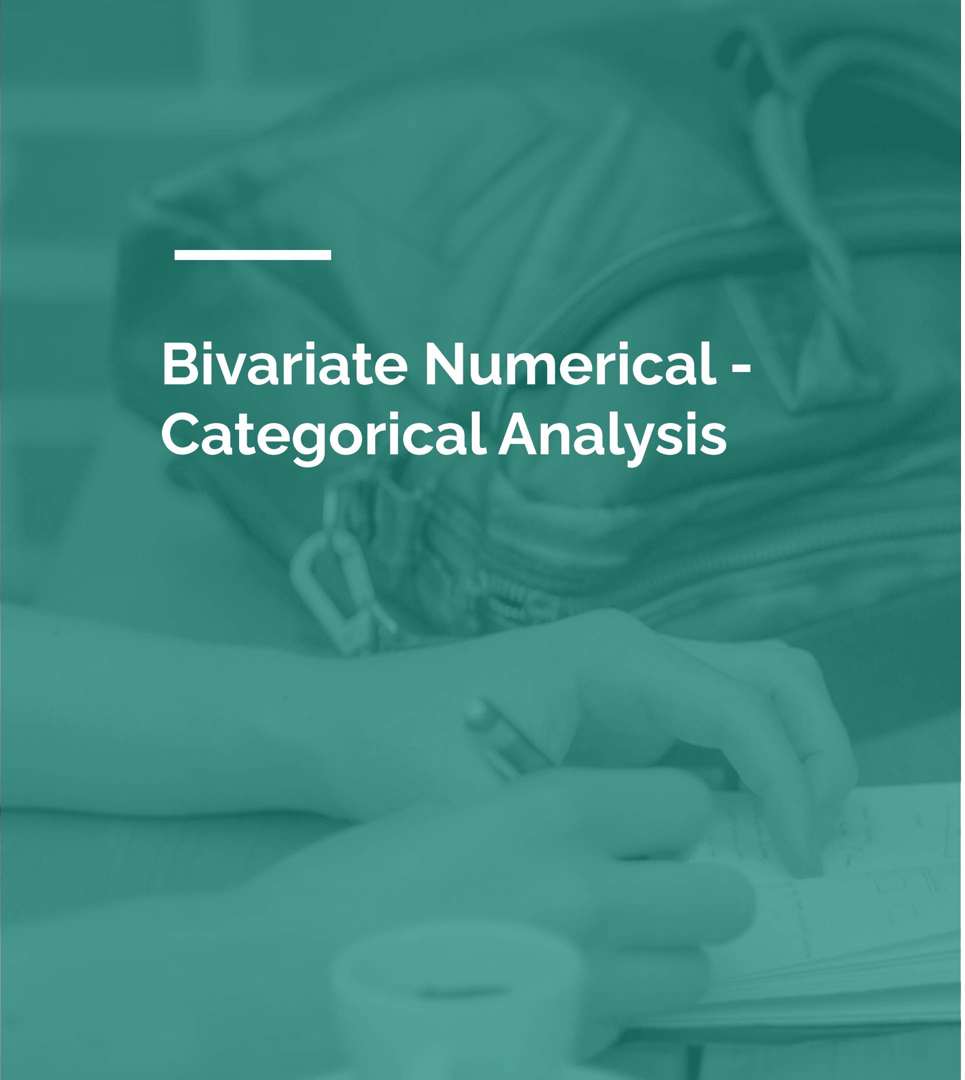
NAME_HOUSING_TYPE vs. AMT_CREDIT - Non-Defaulters and Defaulters



NAME_HOUSING_TYPE vs. AMT_CREDIT - Non-DEFAULTERS



NAME_HOUSING_TYPE vs. AMT_CREDIT - DEFAULTERS



Bivariate Numerical - Categorical Analysis

NAME_HOUSING_TYPE vs. AMT_CREDIT - Non-Defaulters and Defaulters

Insights:

Non-defaulters:

- Family status of married in office apartment are having higher credit limit.
- Also the house/apartment of civil marriage, married and single are having highest outliers.
- Single with parents and rented apartments have low credit amount.

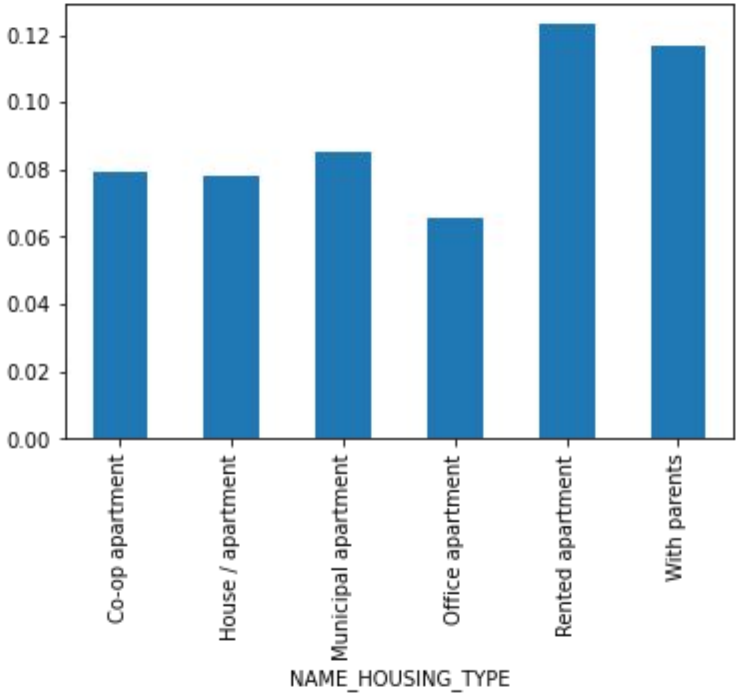
Defaulters:

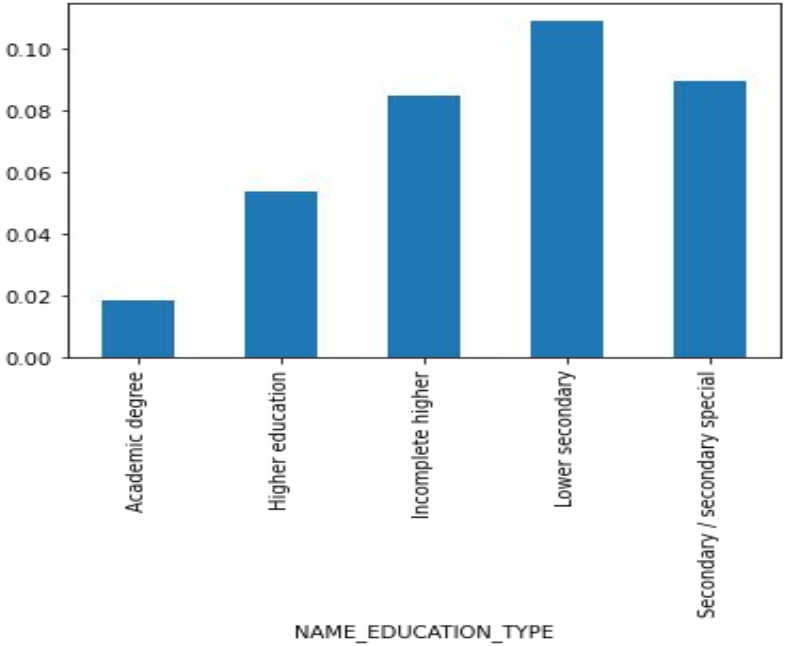
- Family status of widow of office apartment is having higher credit amount than others.
- Also the house/apartment of civil marriage is having highest outliers.
- single in Rented apartment have lower credit amount.

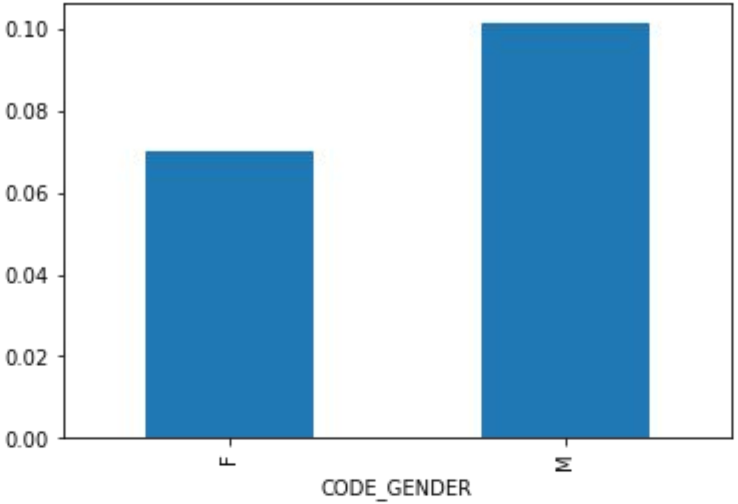
Bivariate Categorical - Categorical Analysis

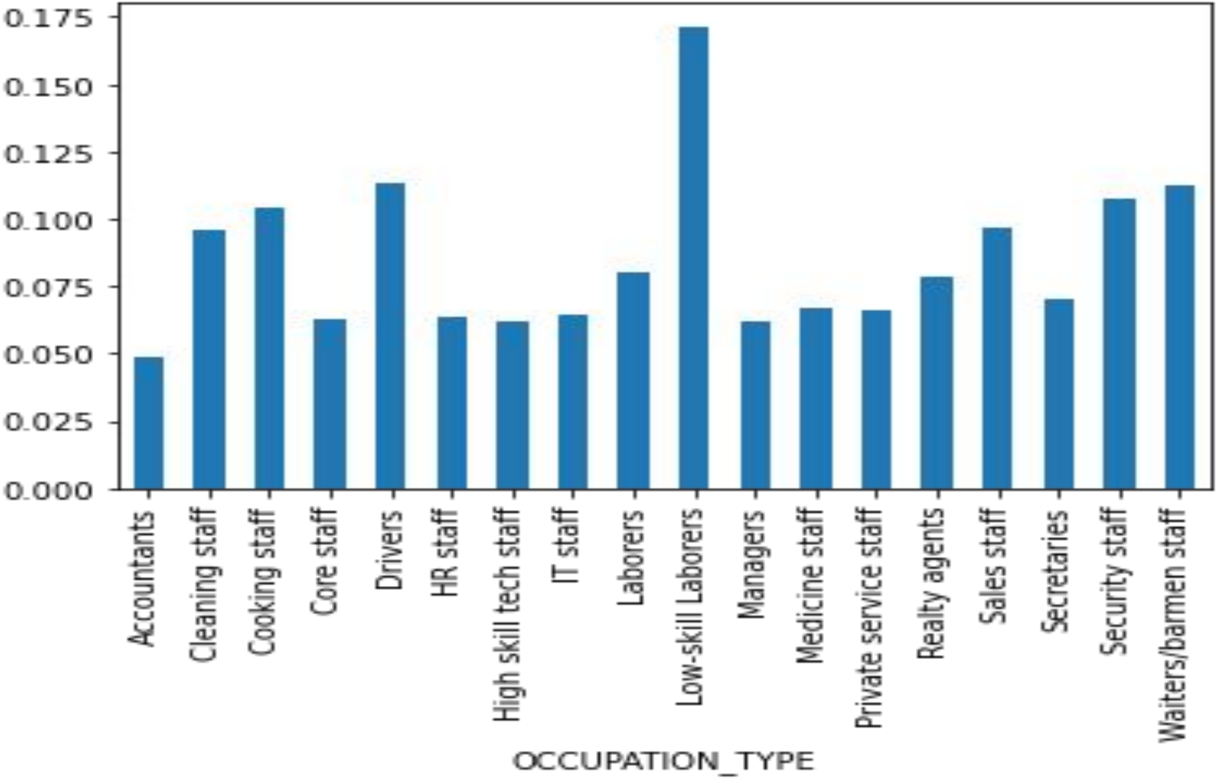
Insights:

- The Rented apartment, Male client, Low-skill labourers and lower secondary are highly at risk to default.




0.10
0.08
0.06
0.04
0.02
0.00
Academic degree
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
NAME_EDUCATION_TYPE

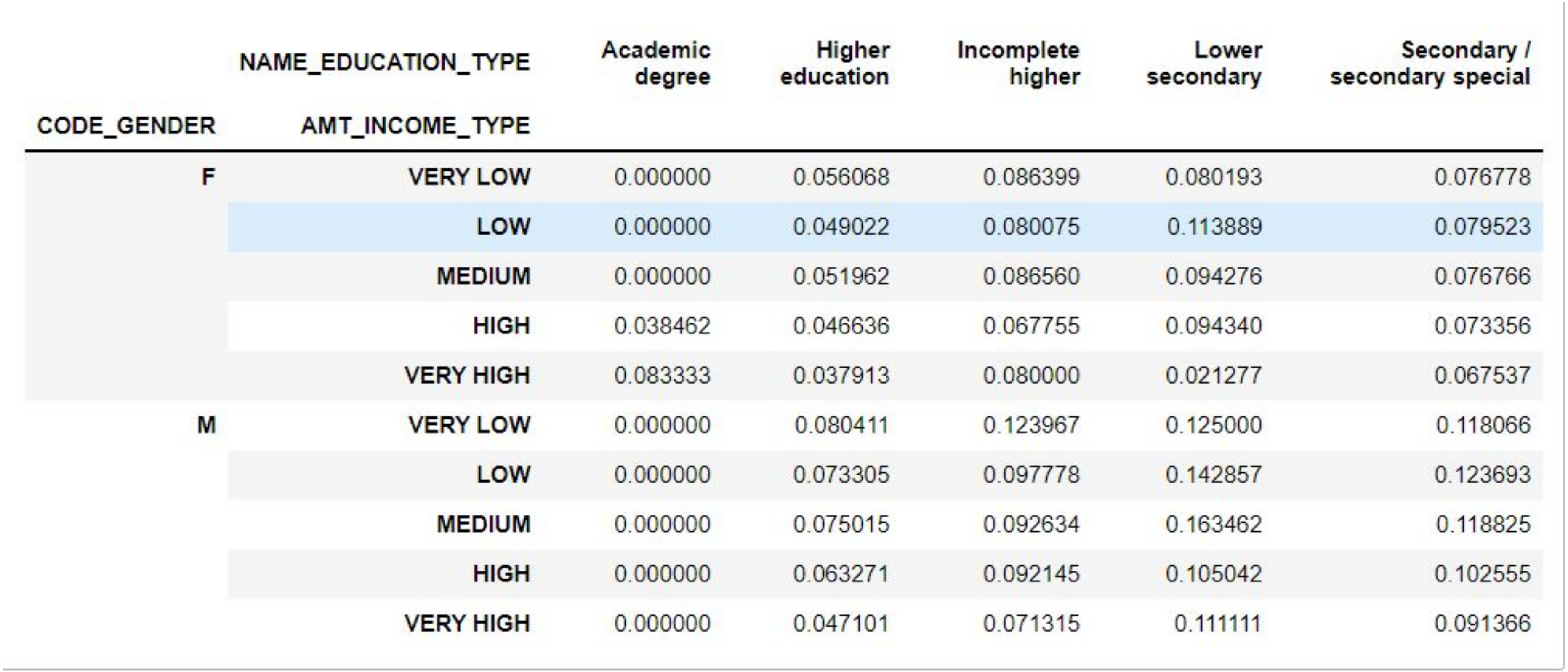

0.10
0.08
0.06
0.04
0.02
0.00
F
M
CODE_GENDER


0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Accountants
Cleaning staff
Cooking staff
Core staff
Drivers
HR staff
High skill tech staff
IT staff
Laborers
Low-skill Laborers
Managers
Medicine staff
Private service staff
Realty agents
Sales staff
Secretaries
Security staff
Waiters/barmen staff
OCCUPATION_TYPE

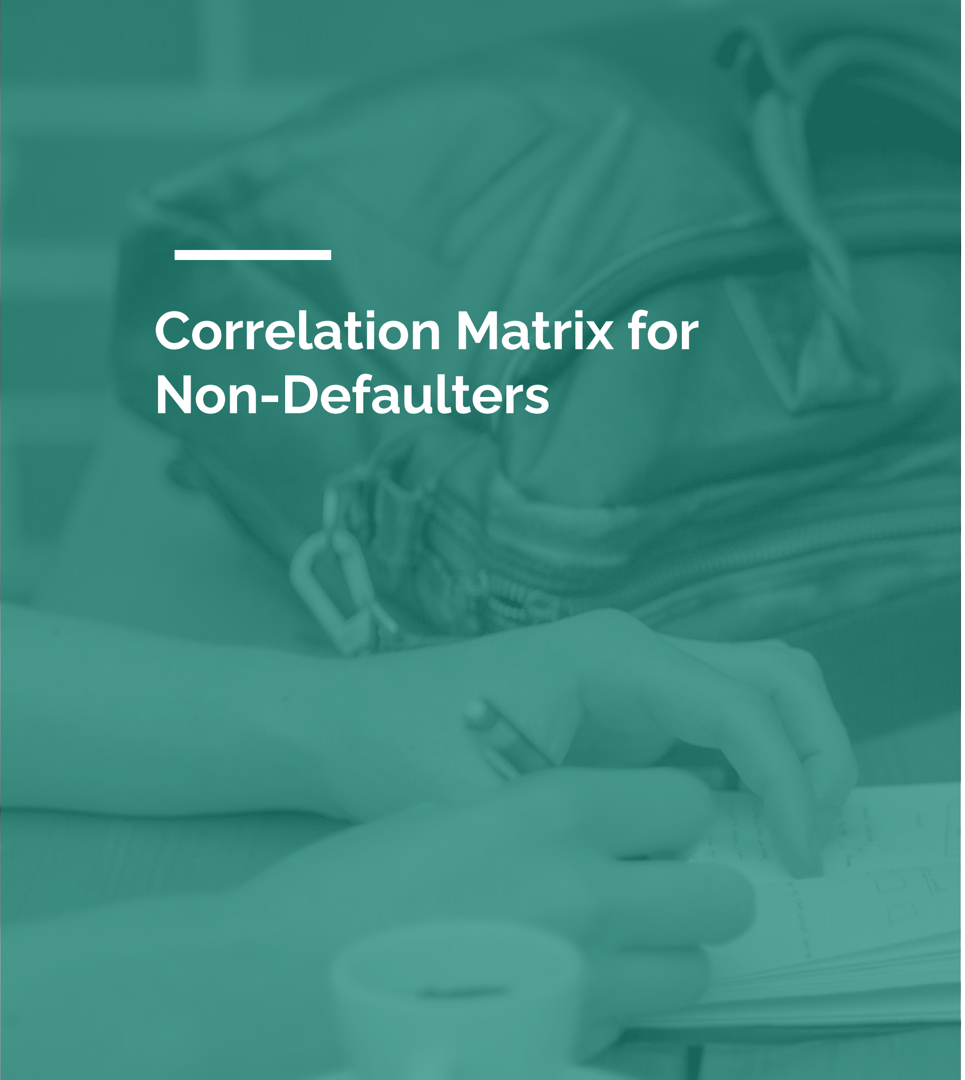
Multivariate Analysis

Insights:

- Female clients with an Academic degree and high-income type have a higher risk of default.
- Female clients with Low income and Lower education have a higher risk of default.
- Male clients with Incomplete Education having very low salaries have a high risk of default.
- Male Clients with Lower Secondary Education having having all types of salaries have a high risk to default.
- Male clients with Secondary/Secondary Special Education having almost all types of salaries have a higher risk of default.



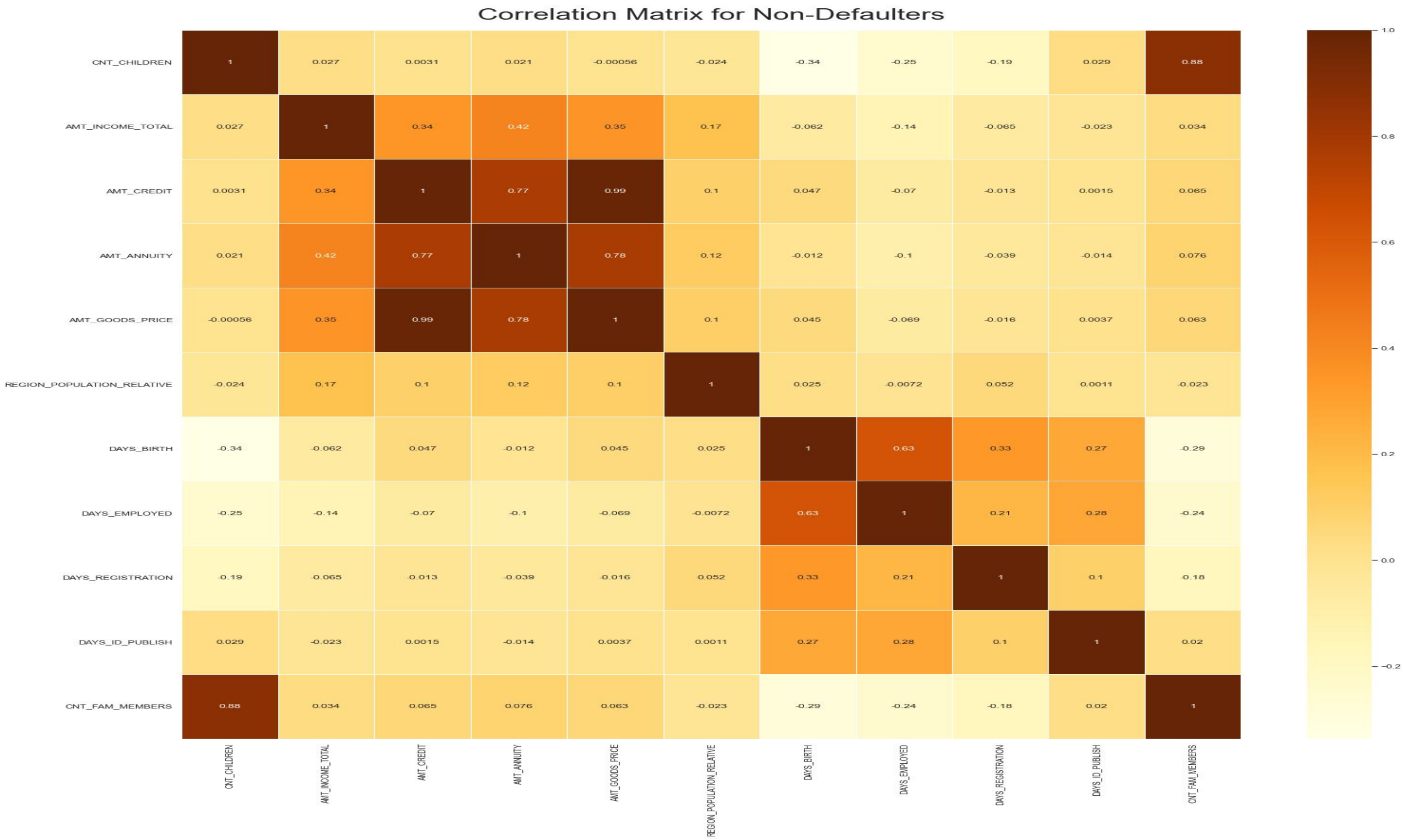
	NAME_EDUCATION_TYPE	Academic degree	Higher education	Incomplete higher	Lower secondary	Secondary / secondary special
CODE_GENDER	AMT_INCOME_TYPE					
F	VERY LOW	0.000000	0.056068	0.086399	0.080193	0.076778
	LOW	0.000000	0.049022	0.080075	0.113889	0.079523
	MEDIUM	0.000000	0.051962	0.086560	0.094276	0.076766
	HIGH	0.038462	0.046636	0.067755	0.094340	0.073356
	VERY HIGH	0.083333	0.037913	0.080000	0.021277	0.067537
M	VERY LOW	0.000000	0.080411	0.123967	0.125000	0.118066
	LOW	0.000000	0.073305	0.097778	0.142857	0.123693
	MEDIUM	0.000000	0.075015	0.092634	0.163462	0.118825
	HIGH	0.000000	0.063271	0.092145	0.105042	0.102555
	VERY HIGH	0.000000	0.047101	0.071315	0.111111	0.091366



Correlation Matrix for Non-Defaulters

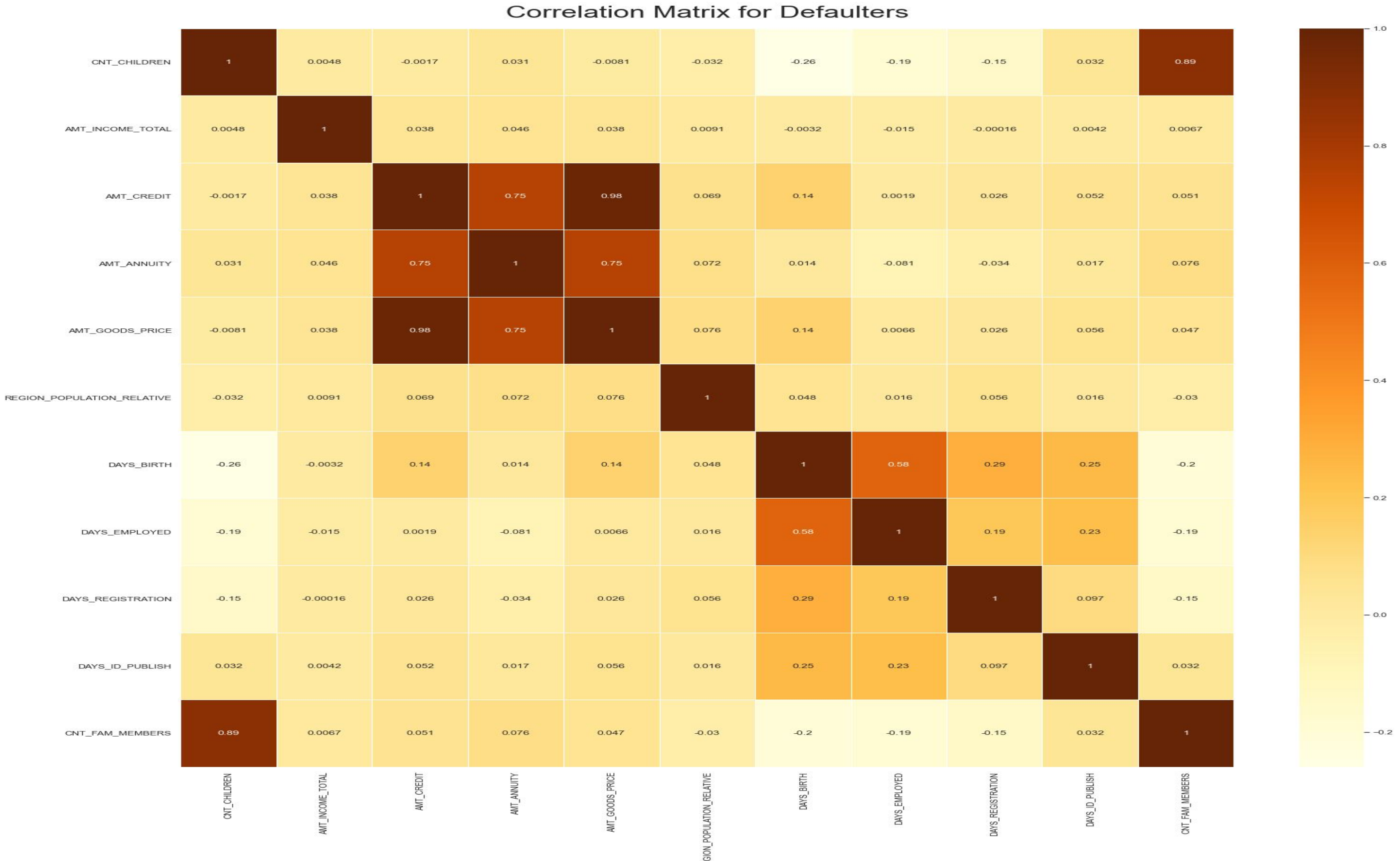
Insights:

- AMT_CREDIT is highly correlated with AMT_INCOME_TOTAL, AMT_ANNUITY and AMT_GOODS_PRICE for both defaulters and non-defaulters.
- There is a severe drop in the correlation between total income of the client and the credit amount(0.038) amongst defaulters.
- Days_birth and number of children correlation has reduced in defaulters when compared to non-defaulters.
- But the AMT Annuity correlation with AMT credit has slightly reduced in defaulters when compared to non-defaulters.



Correlation Matrix for Non-Defaulters

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS
CNT_CHILDREN	1	0.027	0.0031	0.021	-0.00056	-0.024	-0.34	-0.25	-0.19	0.029	0.88
AMT_INCOME_TOTAL	0.027	1	0.34	0.42	0.35	0.17	-0.062	-0.14	-0.065	-0.023	0.034
AMT_CREDIT	0.0031	0.34	1	0.77	0.99	0.1	0.047	-0.07	-0.013	0.0015	0.065
AMT_ANNUITY	0.021	0.42	0.77	1	0.78	0.12	-0.012	-0.1	-0.039	-0.014	0.076
AMT_GOODS_PRICE	-0.00056	0.35	0.99	0.78	1	0.1	0.045	-0.069	-0.016	0.0037	0.063
REGION_POPULATION_RELATIVE	-0.024	0.17	0.1	0.12	0.1	1	0.025	-0.0072	0.052	0.0011	-0.023
DAYS_BIRTH	-0.34	-0.062	0.047	-0.012	0.045	0.025	1	0.63	0.33	0.27	-0.29
DAYS_EMPLOYED	-0.25	-0.14	-0.07	-0.1	-0.069	-0.0072	0.63	1	0.21	0.28	-0.24
DAYS_REGISTRATION	-0.19	-0.065	-0.013	-0.039	-0.016	0.052	0.33	0.21	1	0.1	-0.18
DAYS_ID_PUBLISH	0.029	-0.023	0.0015	-0.014	0.0037	0.0011	0.27	0.28	0.1	1	0.02
CNT_FAM_MEMBERS	0.88	0.034	0.065	0.076	0.063	-0.023	-0.29	-0.24	-0.18	0.02	1



Correlation Matrix for Defaulters

	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	GION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS
CNT_CHILDREN	1	0.0048	-0.0017	0.031	-0.0081	-0.032	-0.26	-0.19	-0.15	0.032	0.89
AMT_INCOME_TOTAL	0.0048	1	0.038	0.046	0.038	0.0091	-0.0032	-0.015	-0.00016	0.0042	0.0067
AMT_CREDIT	-0.0017	0.038	1	0.75	0.98	0.069	0.14	0.0019	0.026	0.052	0.051
AMT_ANNUITY	0.031	0.046	0.75	1	0.75	0.072	0.014	-0.081	-0.034	0.017	0.076
AMT_GOODS_PRICE	-0.0081	0.038	0.98	0.75	1	0.076	0.14	0.0066	0.026	0.056	0.047
REGION_POPULATION_RELATIVE	-0.032	0.0091	0.069	0.072	0.076	1	0.048	0.016	0.056	0.016	-0.03
DAYS_BIRTH	-0.26	-0.0032	0.14	0.014	0.14	0.048	1	0.58	0.29	0.25	-0.2
DAYS_EMPLOYED	-0.19	-0.015	0.0019	-0.081	0.0066	0.016	0.58	1	0.19	0.23	-0.19
DAYS_REGISTRATION	-0.15	-0.00016	0.026	-0.034	0.026	0.056	0.29	0.19	1	0.097	-0.15
DAYS_ID_PUBLISH	0.032	0.0042	0.052	0.017	0.056	0.016	0.25	0.23	0.097	1	0.032
CNT_FAM_MEMBERS	0.89	0.0067	0.051	0.076	0.047	-0.03	-0.2	-0.19	-0.15	0.032	1

Previous Application Dataset

Analysis:

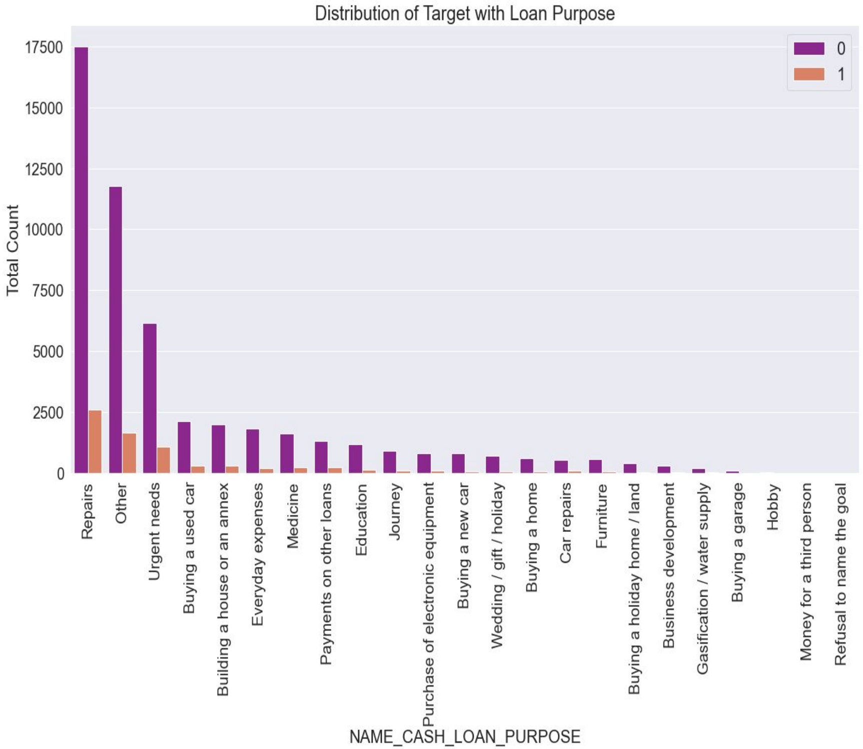
The same steps of analysis done for Application dataset has to be followed for previous application data set also.

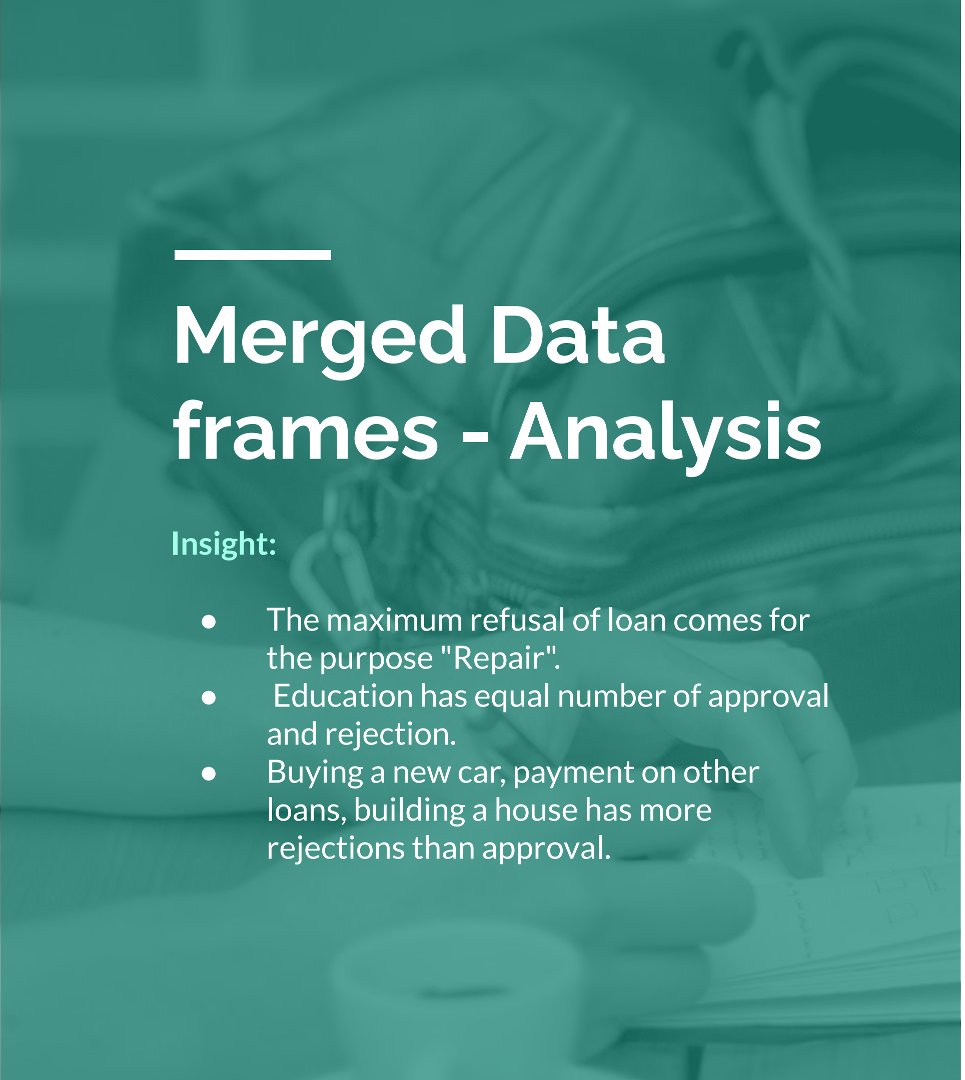
Merged Data frames - Analysis

Insights:

- Loan purposes with 'Repairs' are having more difficulties in paying on time when compared to others.
- There is high number of clients having difficulties for repaying the loan amount under 'Medicine' when compared to "Education".
- 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car' and 'Education' are showing less paying difficulties. Hence Bank can focus on these areas.

TARGET with NAME_CASH_LOAN_PURPOSE



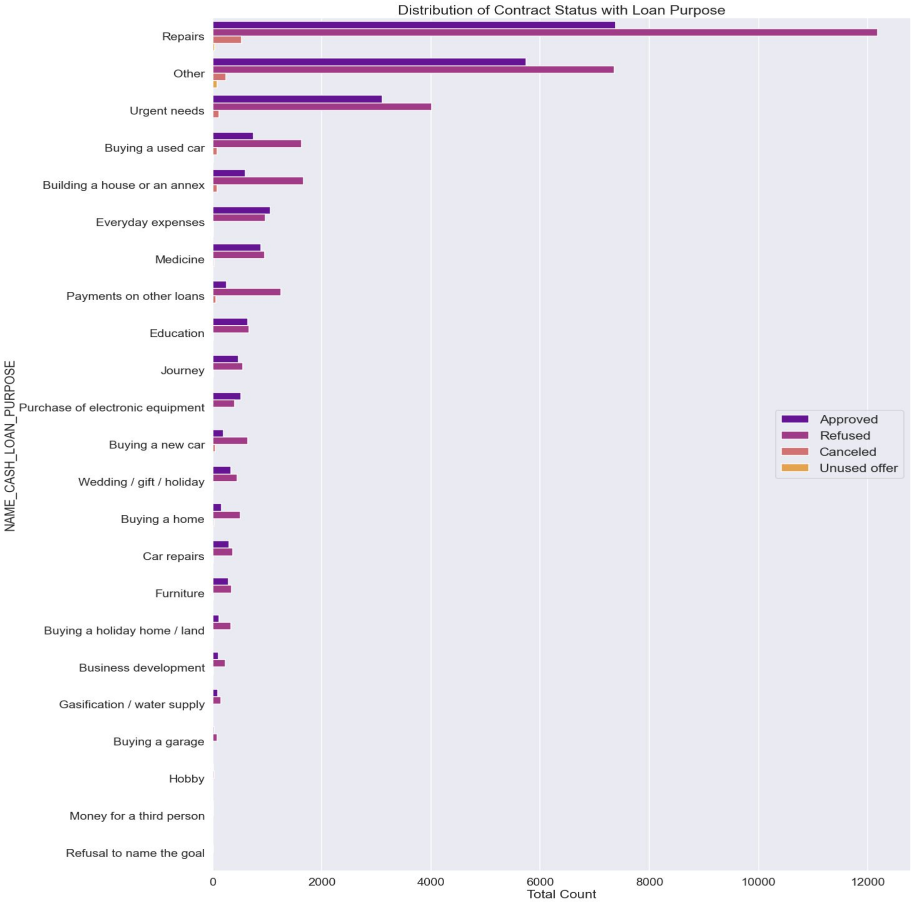


Merged Data frames - Analysis

Insight:

- The maximum refusal of loan comes for the purpose "Repair".
- Education has equal number of approval and rejection.
- Buying a new car, payment on other loans, building a house has more rejections than approval.

NAME_CONTRACT_STATUS with NAME_CASH_LOAN_PURPOSE

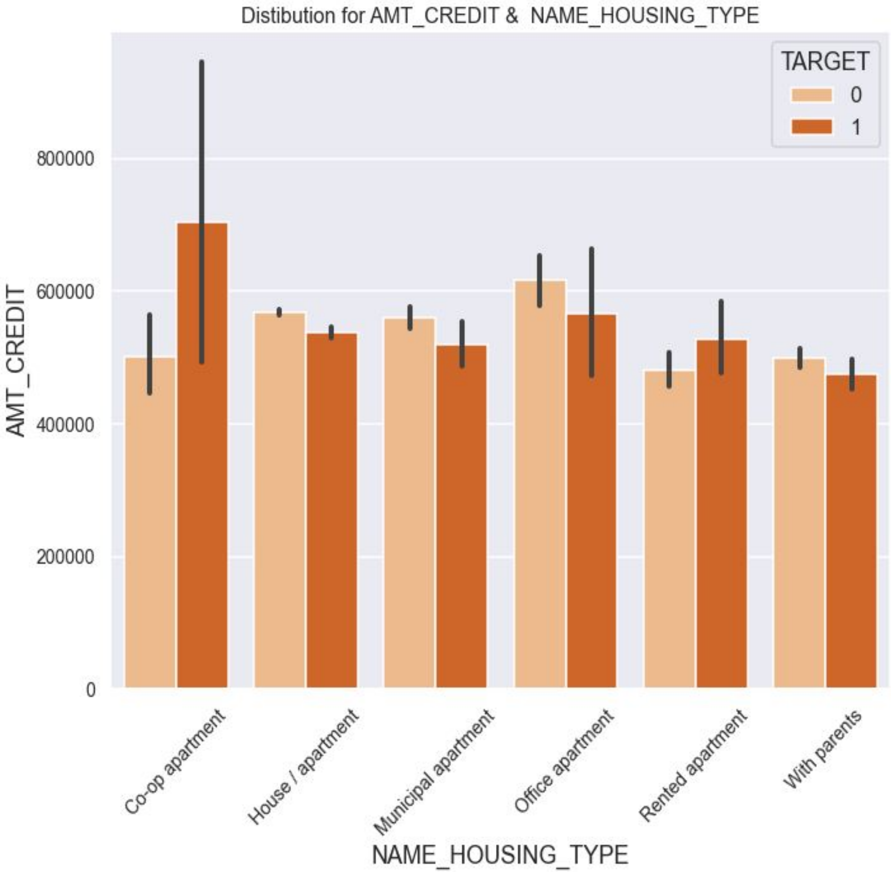


Merged Data frames - Analysis

Insight:

- The housing type 'Co-op apartment' has more clients having payment difficulties. So bank avoid giving loans to this housing type.
- Also bank should give loans for the housing type category with parents or Municipal apartment as the clients having payment difficulties are less.

AMT_CREDIT and NAME_HOUSING_TYPE

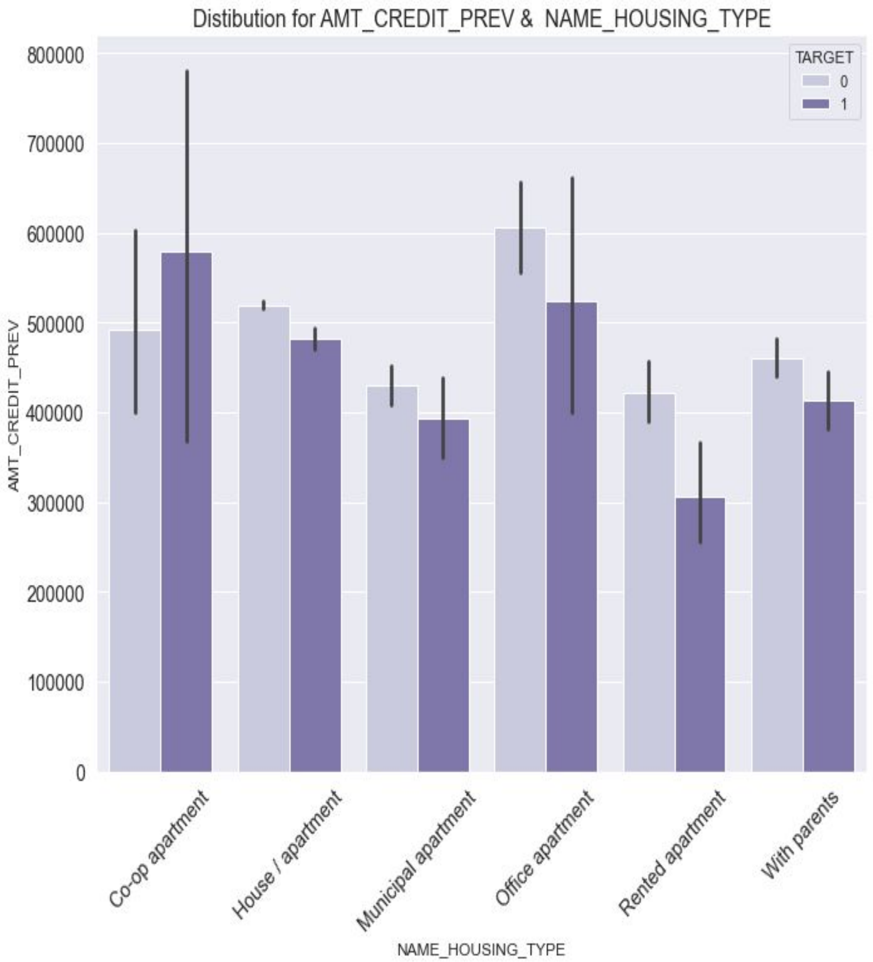


Merged Data frames - Analysis

Insight:

- In the previous application the housing type 'Co-op apartment' had more clients having payment difficulties and same as current application.
- Also people who are in Rented apartment are the ones having less payment difficulties.
- The highest people who did not had payment difficulty come under the housing type Office apartment.

AMT_CREDIT_PREV and NAME_HOUSING_TYPE



Conclusion

- NAME_EDUCATION_TYPE: Academic degree has less defaults. People with Lower Secondary & Secondary education more to default.
- NAME_INCOME_TYPE: Student and Businessmen have no defaults. Bank should avoid giving loans to income type Working as they are the one having most unsuccessful payments.
- CNT_CHILDREN: People with zero to two children are able to repay the loans. Client who have children equal to or more than 9 have high default rate.
- AMT_INCOME_TOTAL: Applicant with Income more than 700,000 are less likely to default.
- DAYS_BIRTH: Very young and people above age of 50 has less default. Avoid young people who are in age group of 20-40 as they have higher probability of defaulting

- OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as the default rate is huge.
- NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed are at high risk to default.
- ORGANIZATION_TYPE: Self-employed people have relative high defaulting rate
- NAME_HOUSING_TYPE: Bank should approach the client who have the housing type as With parents.
- CODE_GENDER: Men are at high default rate.
- NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, Buying garage are being repayed mostly. Since the Loan purpose of 'Repair' has the highest number of payment difficulties bank should avoid providing loans to this category.

- NAME_FAMILY_STATUS : Widows, seperated default a lot.
- DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.