CREDIT RISK ANALYSIS

EDA Case Study

By: Parinita Dwivedi

CONTENTS

PROBLEM STATEMENT

Understanding driver variables behind loan default and finding out which variables are strong indicators of a future default.

This is achieved by using EDA (Exploratory Data Analysis) on the bank data, wherein we analyze the patterns present in data and help the bank in mitigating two types of risk associated with loan approvals:

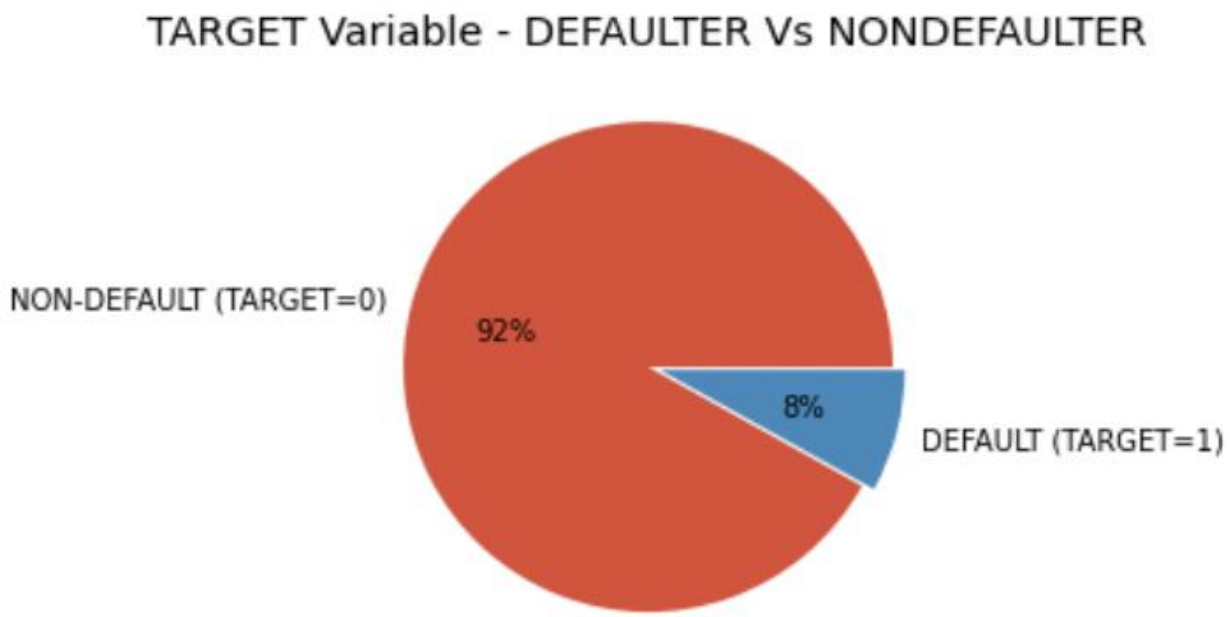
- If the applicant is likely to repay the loan, then not approving the loan will result in loss of business for the bank.

- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may result into financial loss for the company.

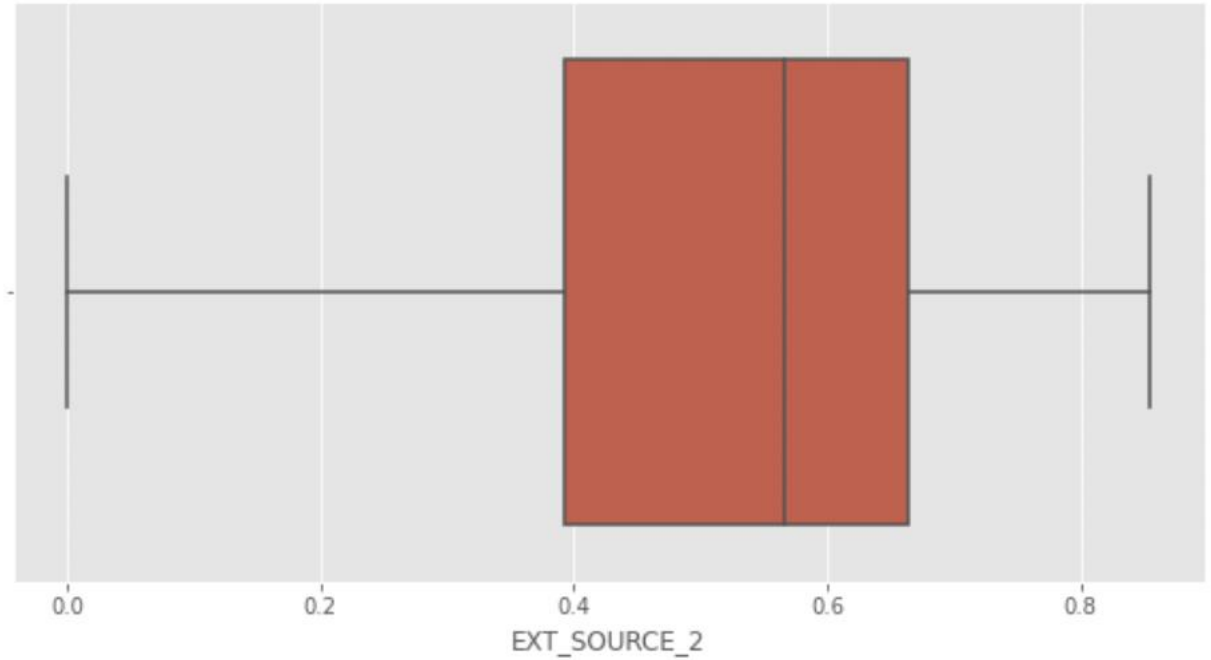
OVERALL APPROACH / STEPS

1. *Data Understanding and Sourcing*
2. *Data Loading and basic sanity checks*
3. *Check for data quality issues like: null values, outliers, etc*
4. *Check for data imbalance*
5. *Identifying important variables using univariate, bivariate and segmented univariate analysis of application data*
6. *Identify top correlations*
7. *Identifying patterns using univariate and bivariate analysis of previous application data*
8. *Merging of application data with previous application data*
9. *Analysis on merged data*
10. *Recommendations and Risks based on the data analysis done*

Checking Imbalance in Target

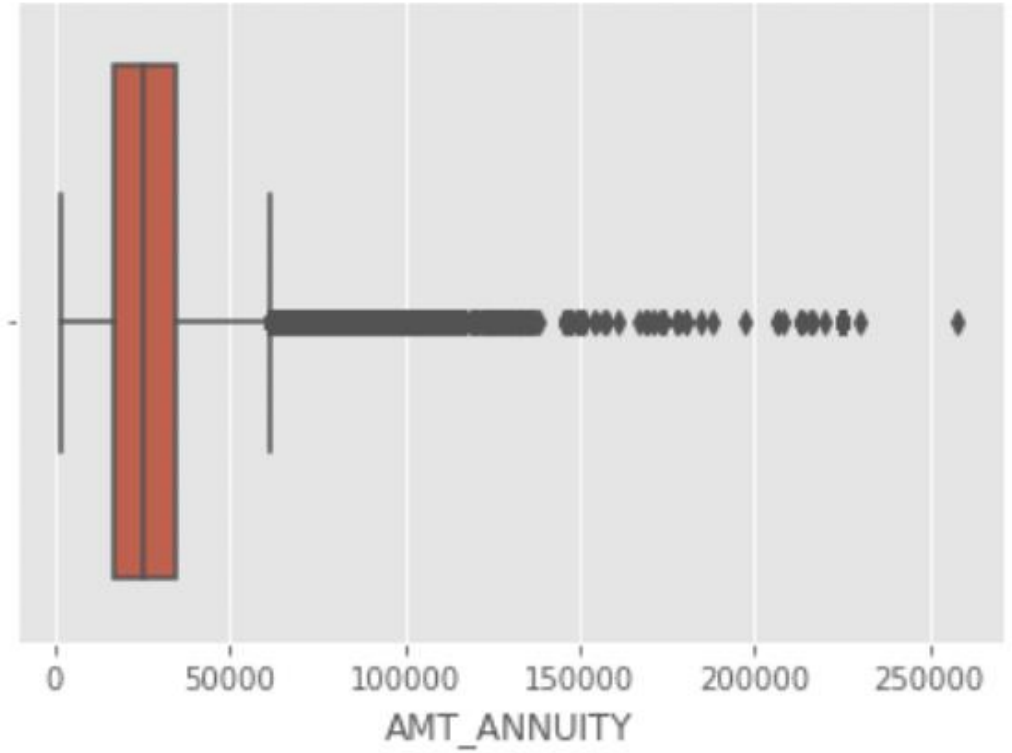


Checking for outliers in Source_2 to impute



Since EXT_SOURCE_2 has no outlier, the column can be imputed using the mean of the column i.e. 0.51

Checking for outliers in AMT_ANNUITY to impute



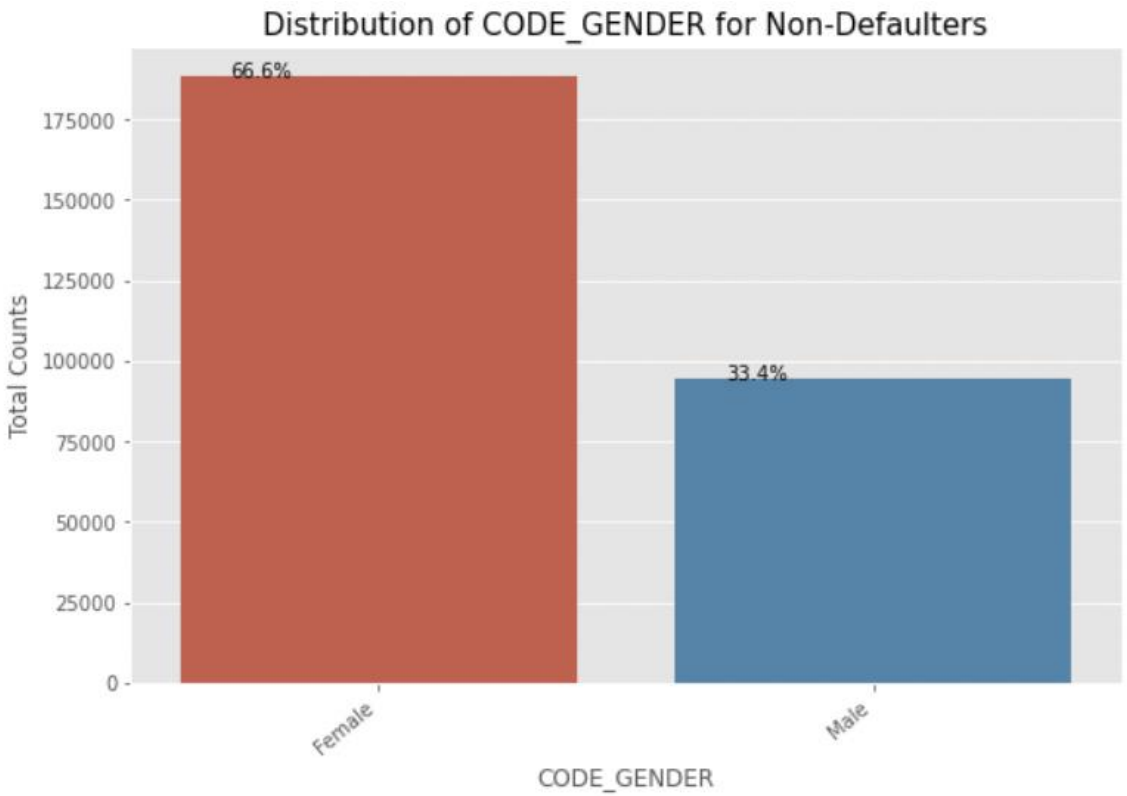
Since AMT_ANNUITY has outliers, the column can be imputed using the median of the column i.e. 24903.0

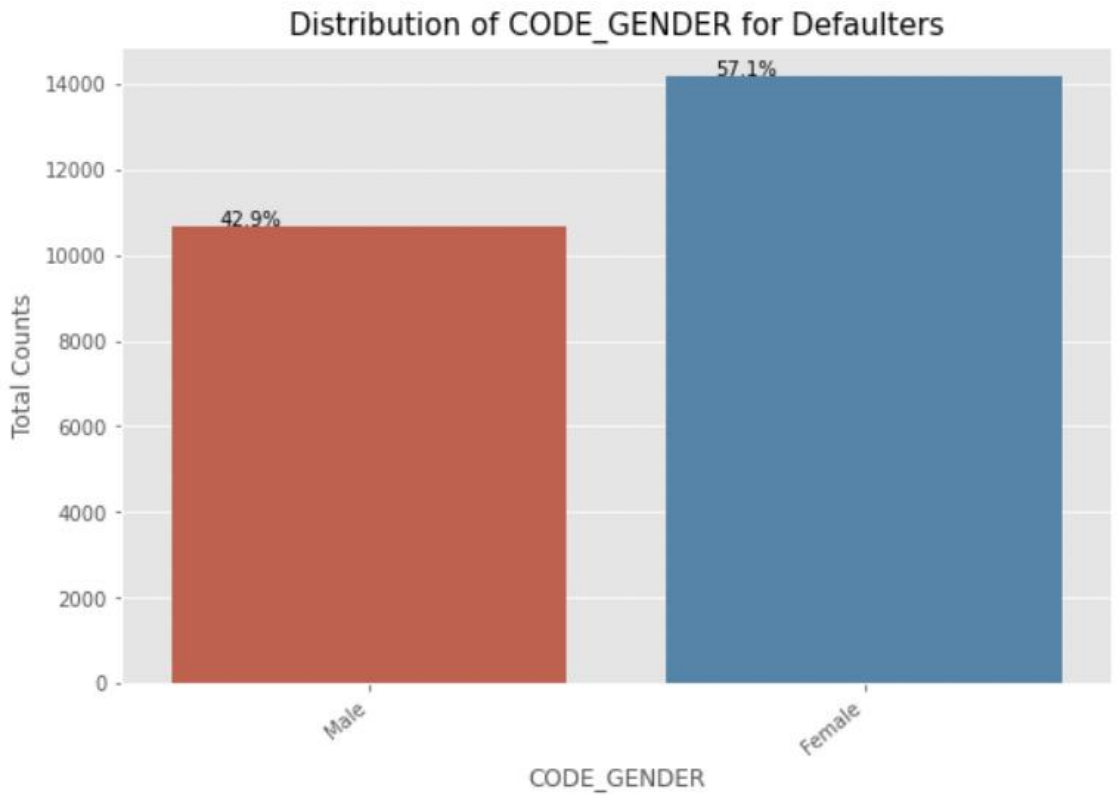
NAME_TYPE_SUITE imputation

```
Unaccompanied      248526
Family              40149
Spouse, partner     11370
Children             3267
Other_B              1770
Other_A               866
Group of people       271
Name: NAME_TYPE_SUITE, dtype: int64
```

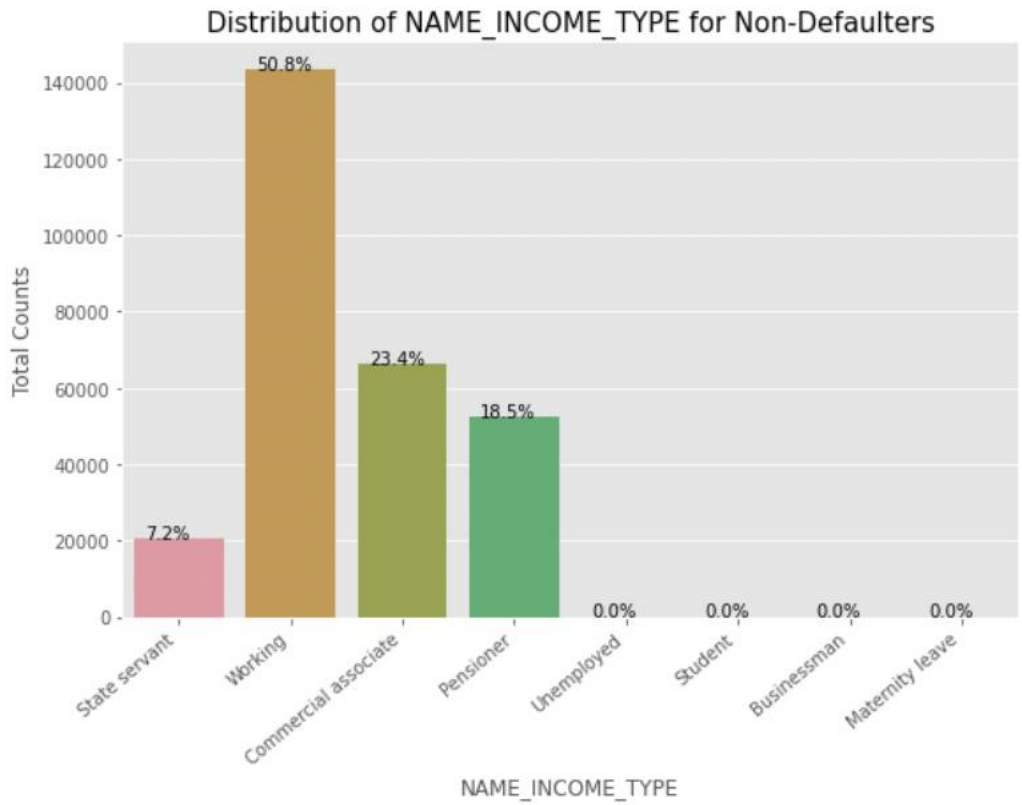
Clearly the column NAME_TYPE_SUITE is a categorical column. So this column can be imputed using the mode of the column i.e Unaccompanied

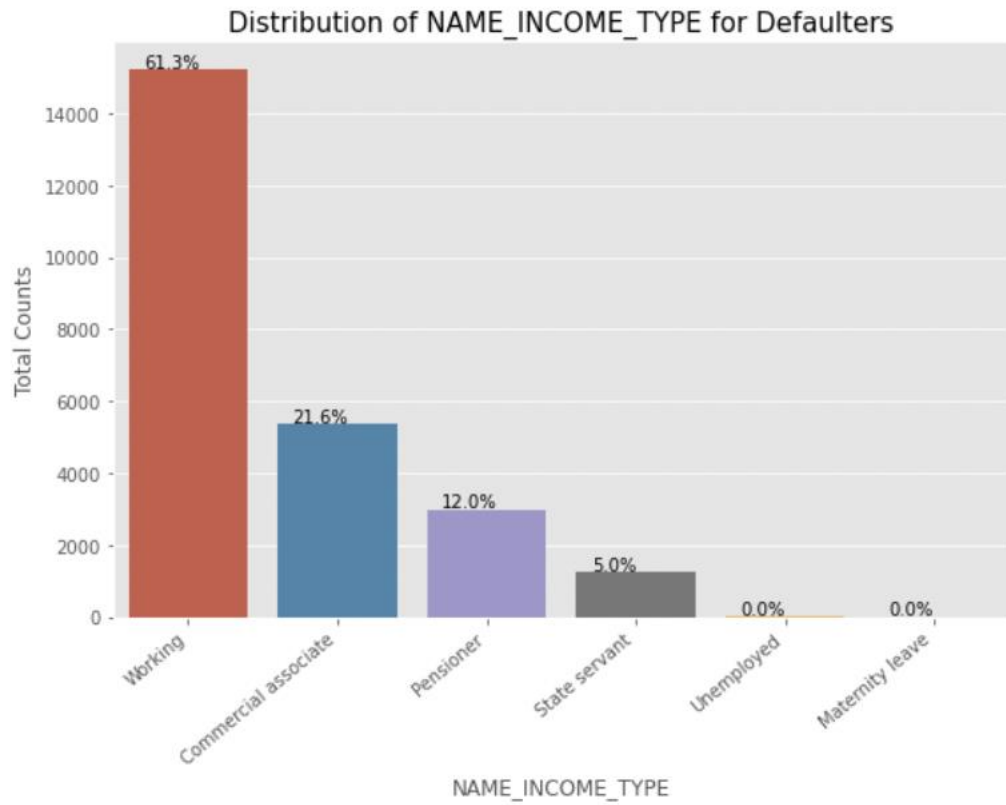
Univariate Analysis on Application Data



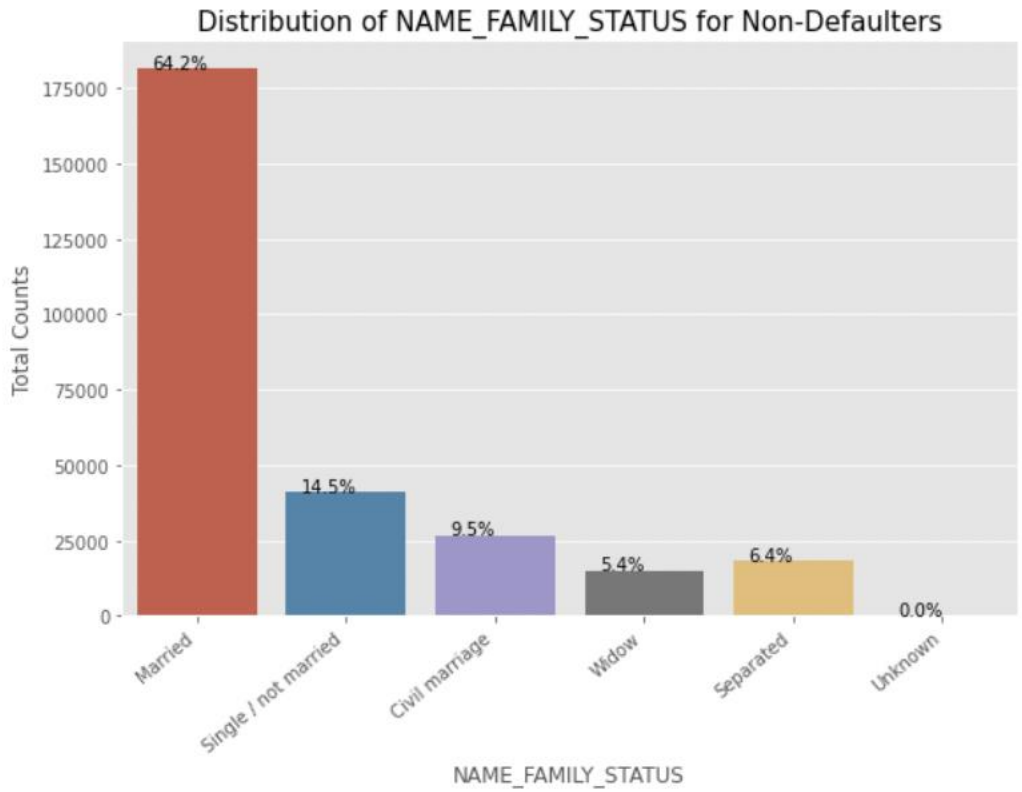


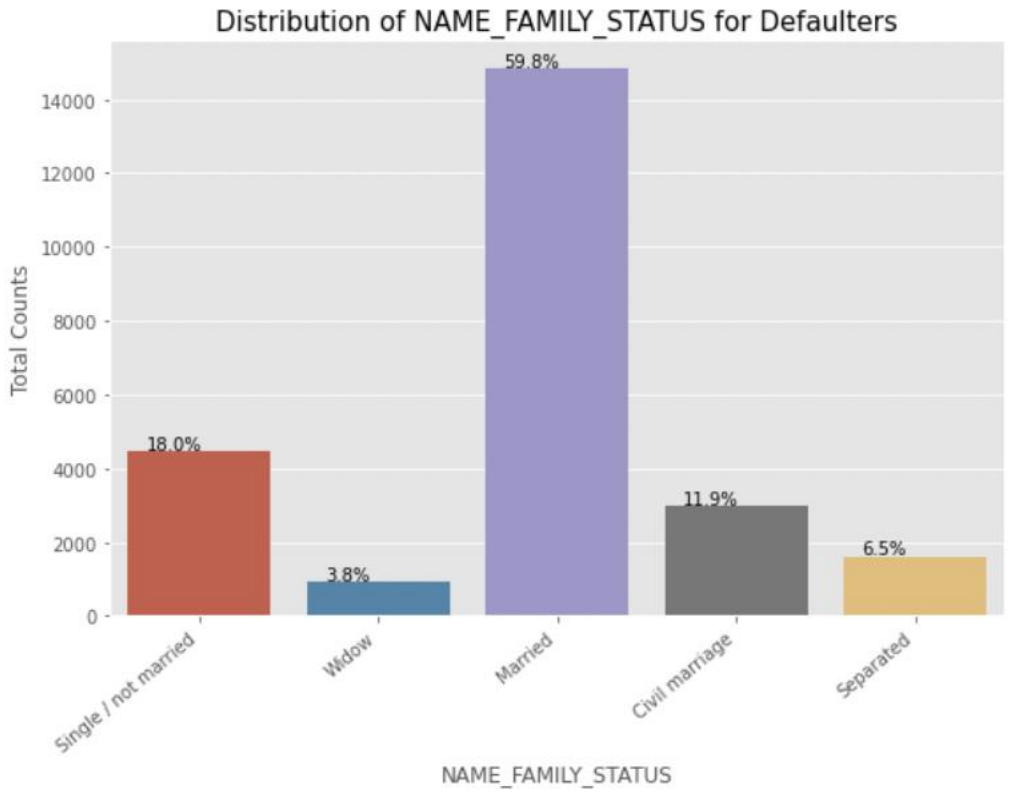
We can see that Female contribute 67% to the non-defaulters while 57% to the defaulters. We can conclude that We see more female applying for loans than males and hence the more number of female defaulters as well.
But the rate of defaulting of FEMALE is much lower compared to their MALE counterparts.



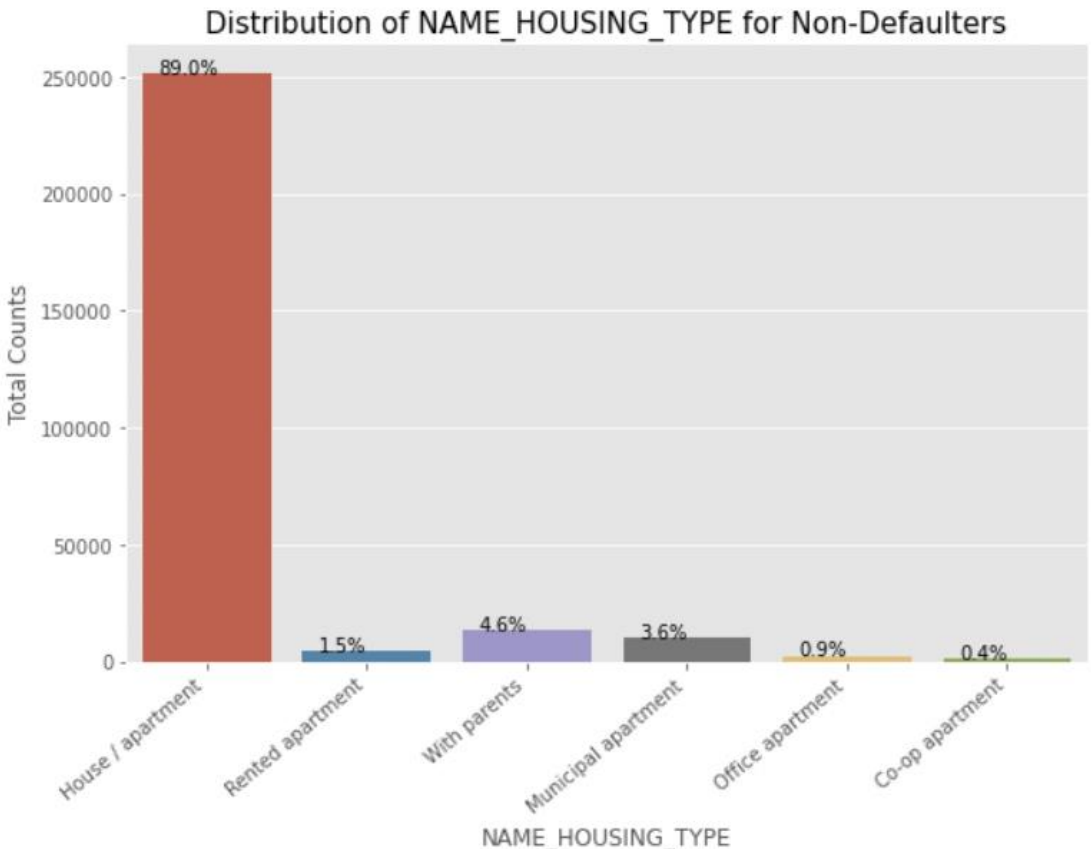


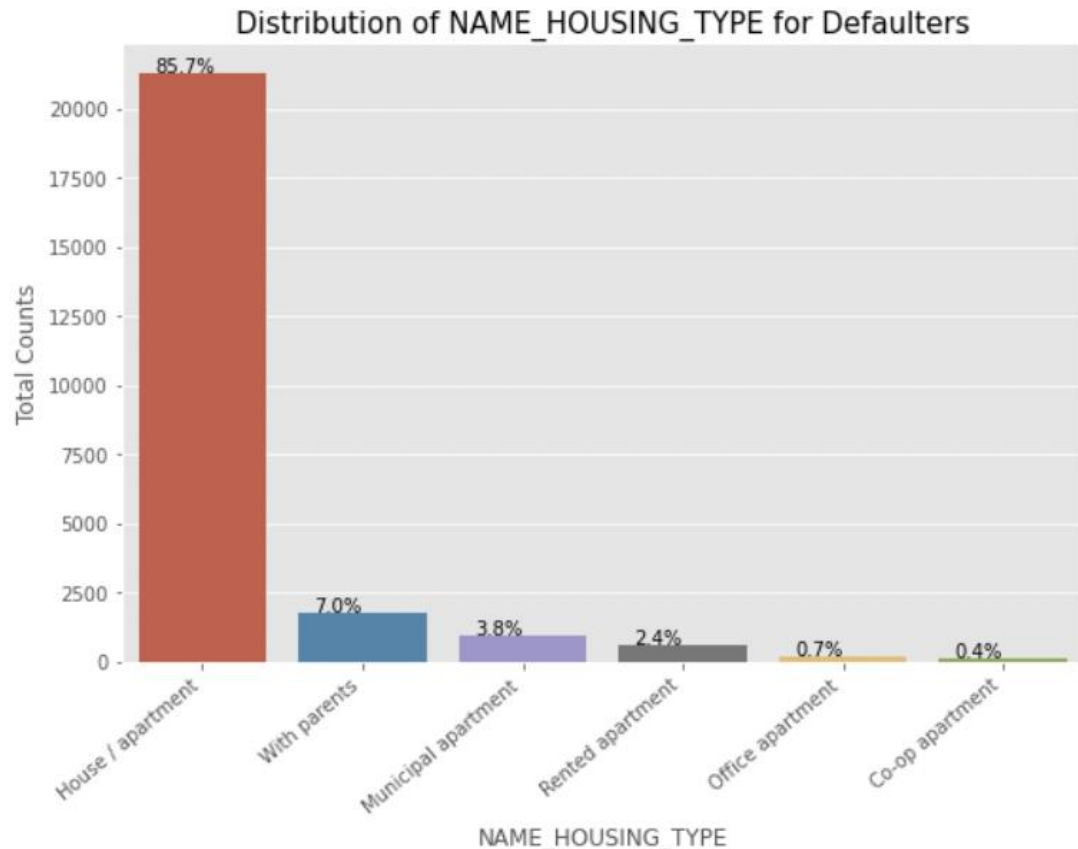
> We can notice that the students don't default. The reason could be they are not required to pay during the time they are students.
> We can also see that the BusinessMen never default.
> Most of the loans are distributed to working class people
> We also see that working class people contribute 51% to non defaulters while they contribute to 61% of the defaulters. Clearly, the chances of defaulting are more in their case.



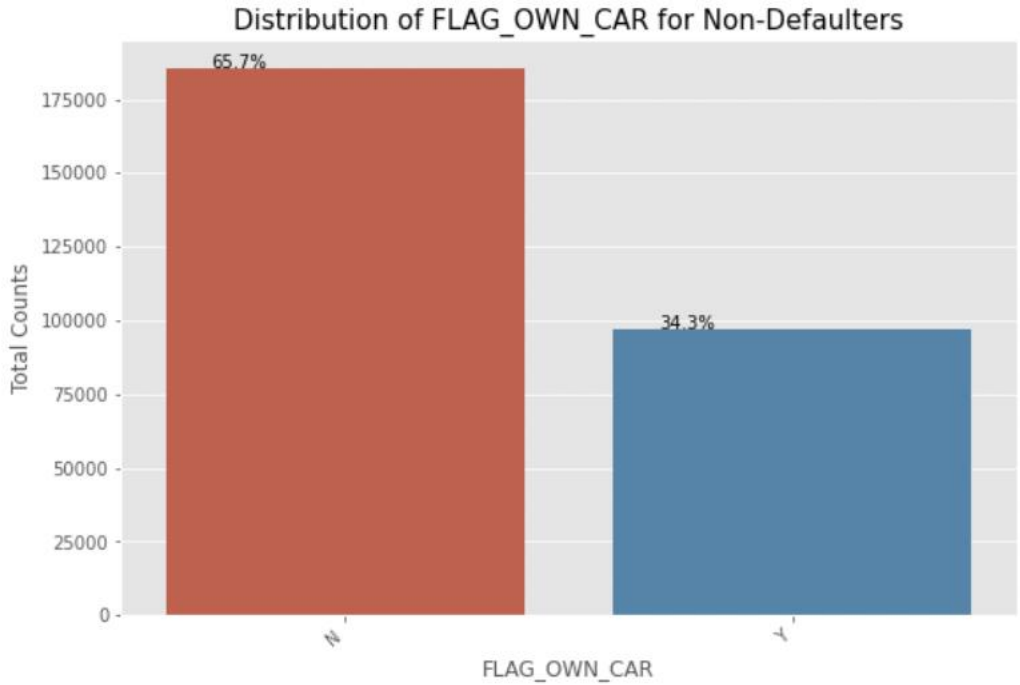


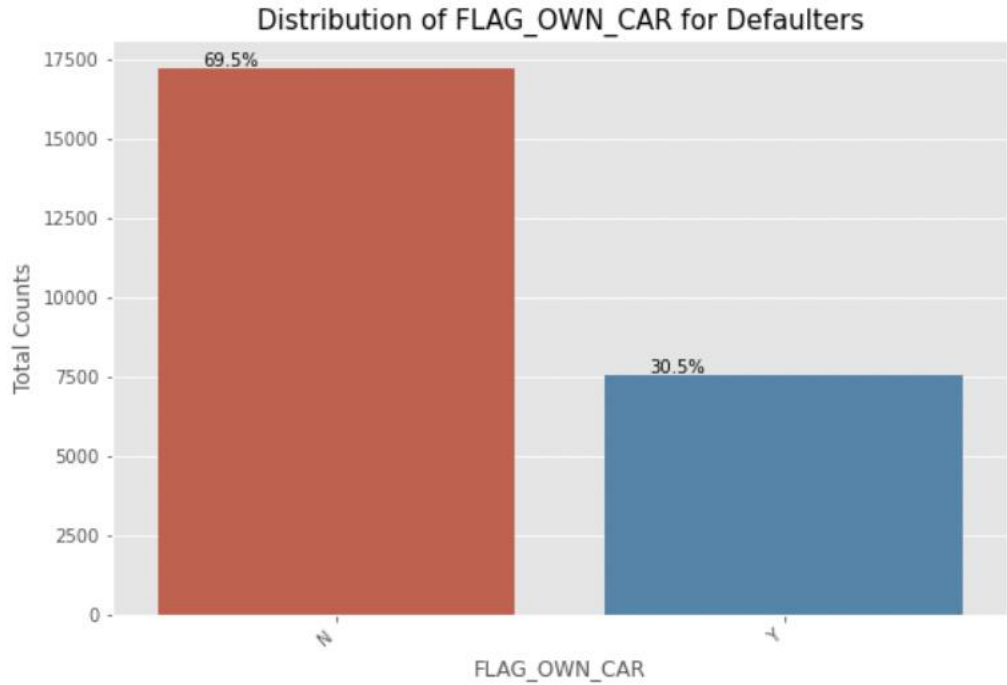
> Married people tend to apply for more loans comparatively.
> But from the graph we see that Single/non Married people contribute 14.5% to Non Defaulters and 18% to the defaulters. So there is more risk associated with them.



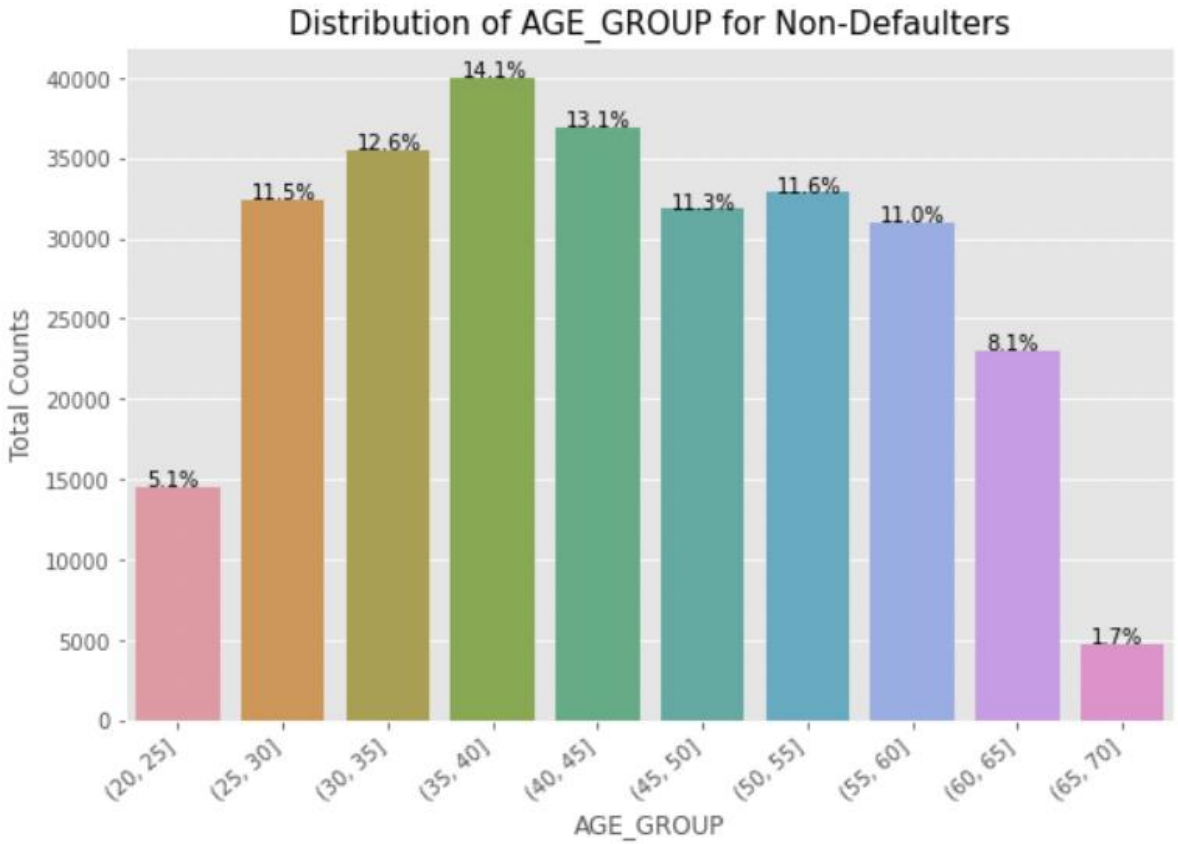


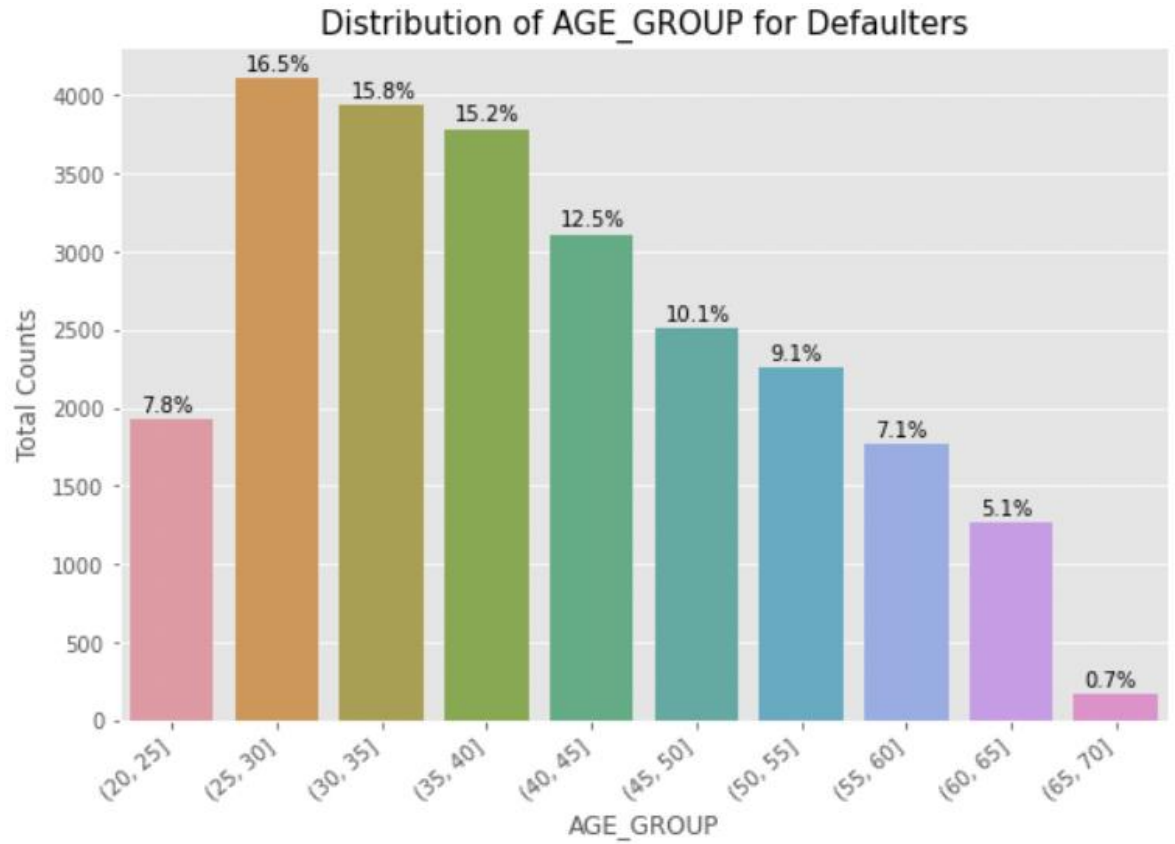
> It is clear from the graph that people who have House/Appartment, tend to apply for more loans.
> People living with parents tend to default more often when compared with others.The reason could be their living expenses are more due to their parents living with them.



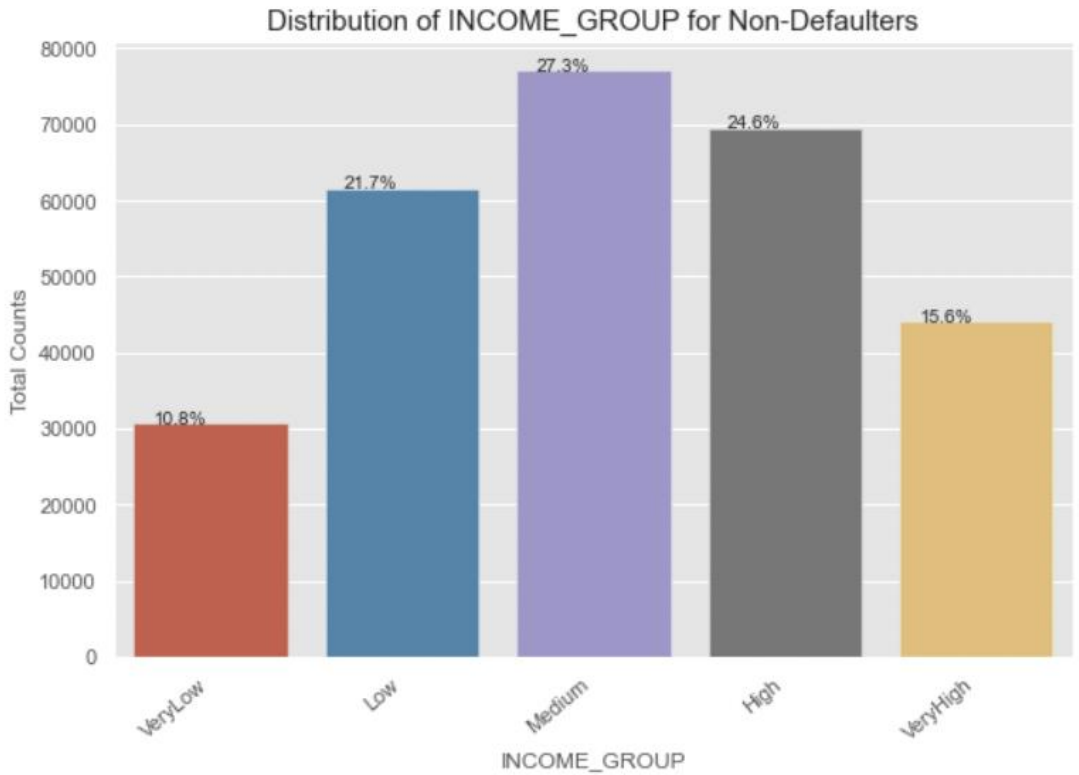


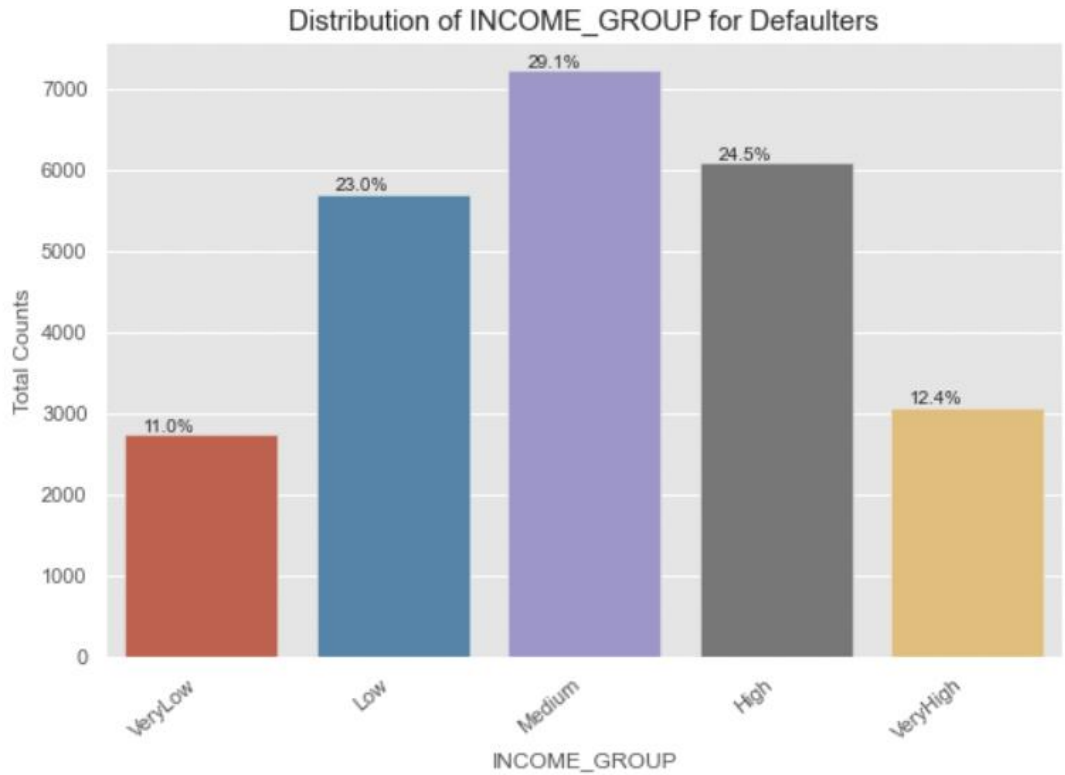
> We can see that people with cars contribute 65.7% to the non-defaulters while 69.5% to the defaulters. We can conclude that
> While people who have car default more often, the reason could be there are simply more people without cars
> Looking at the percentages in both the charts, we can conclude that the rate of default of people having car is low compared to people who don't.



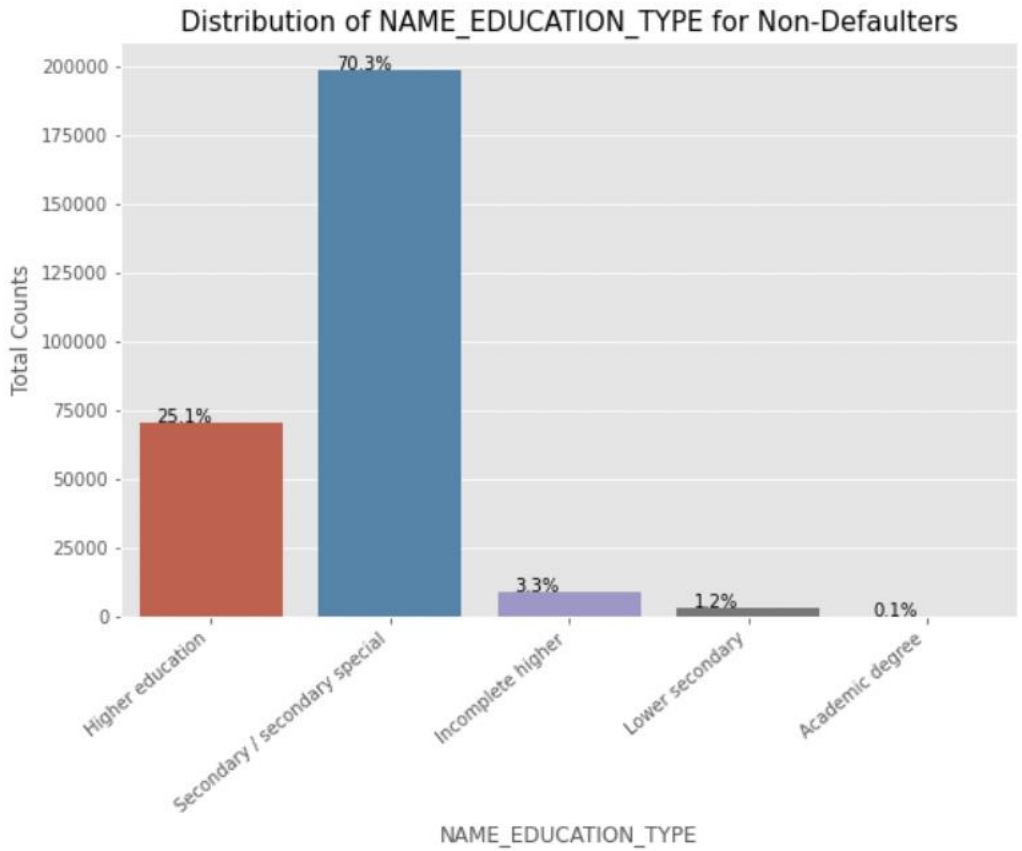


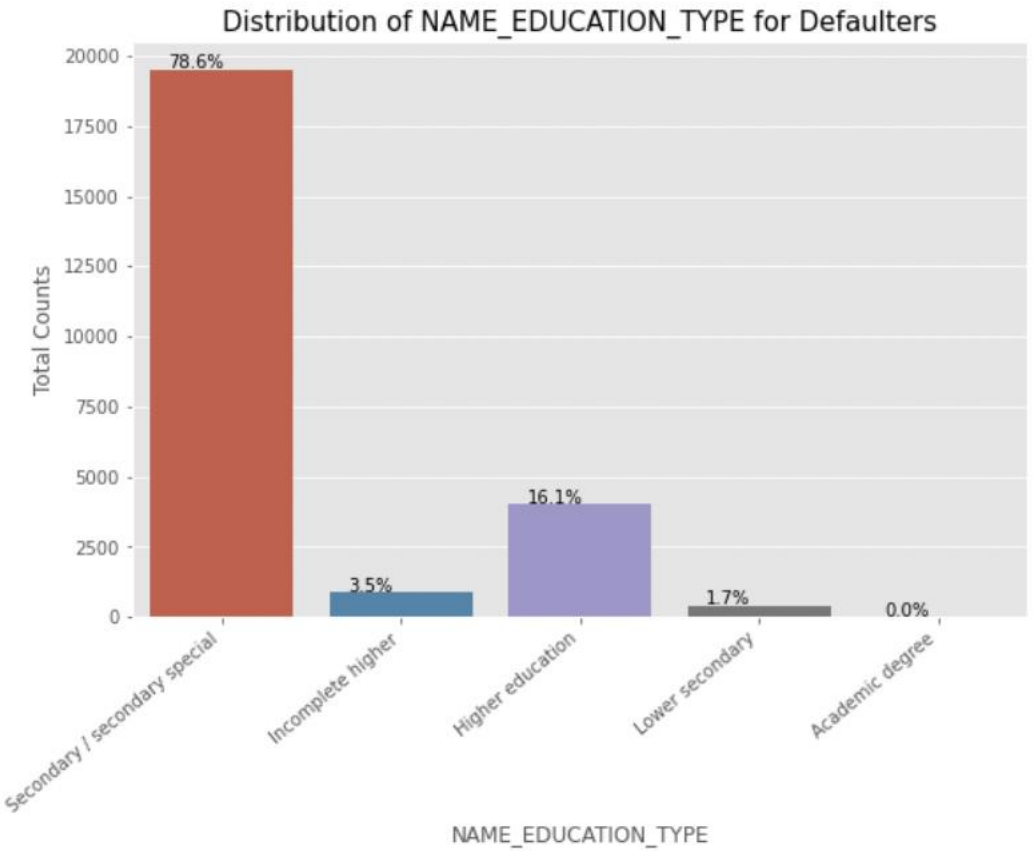
> We see that (25,30] age group tend to default more often. So they are the riskiest people to loan to.
> With increasing age group, people tend to default less starting from the age 25. One of the reasons could be they get employed around that age and with increasing age, their salary also increases.



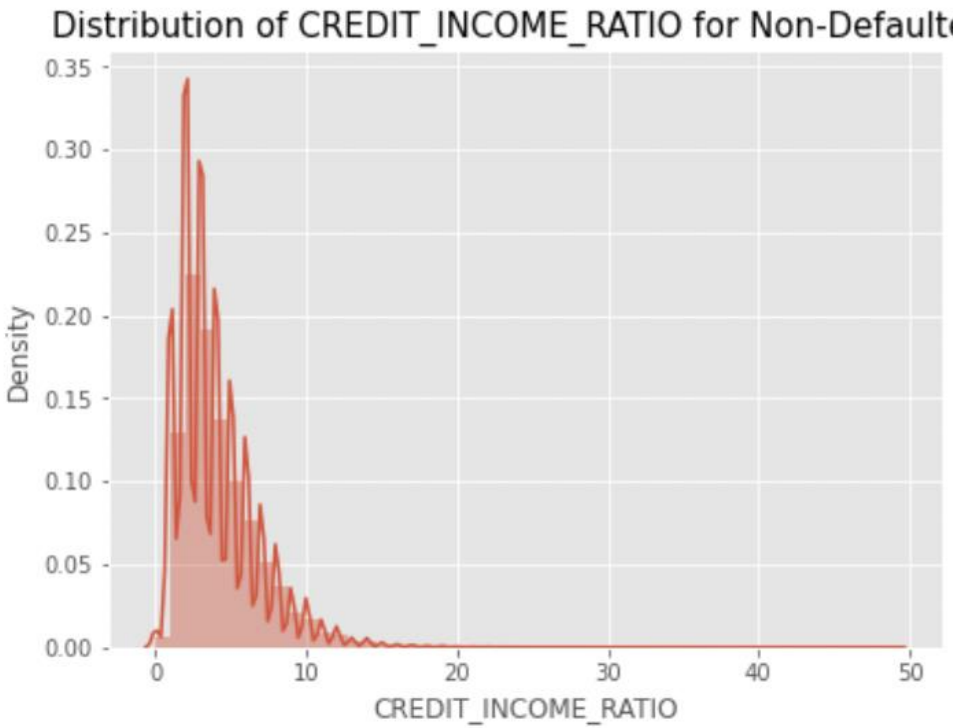


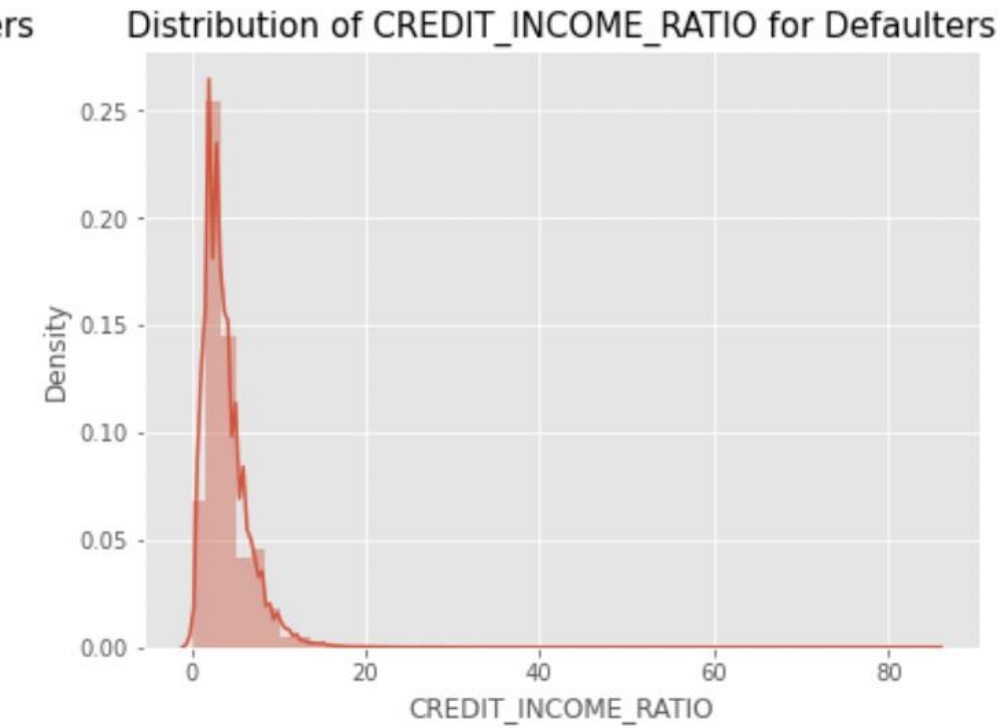
> The Very High income group tend to default less often. They contribute 12.4% to the total number of defaulters, while they contribute 15.6% to the Non-Defaulters.



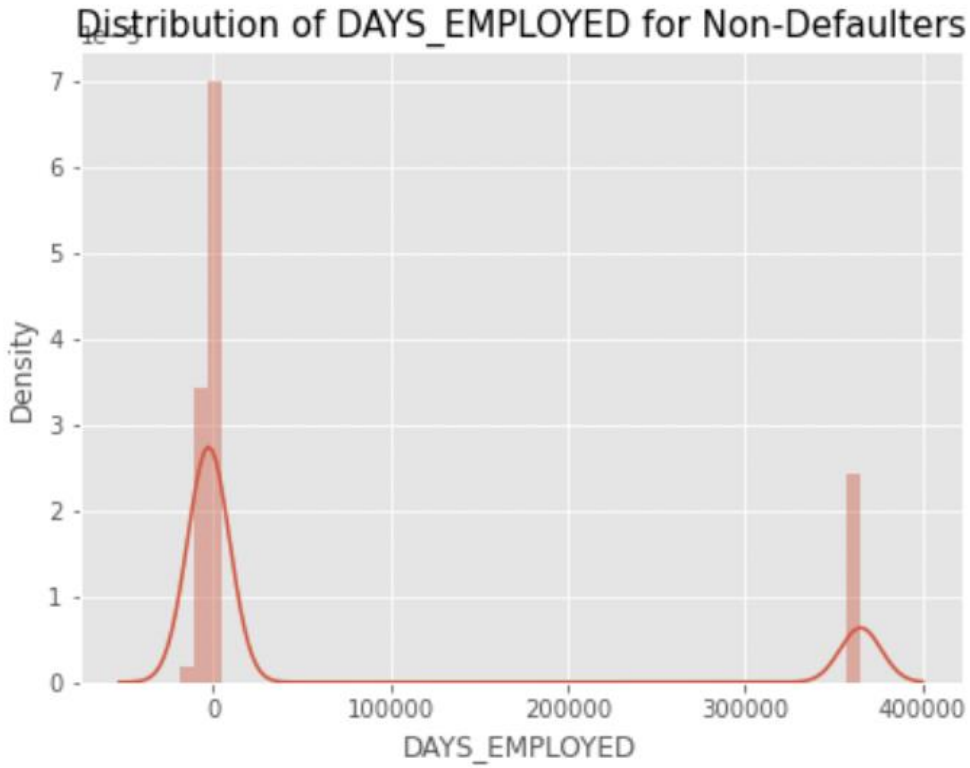


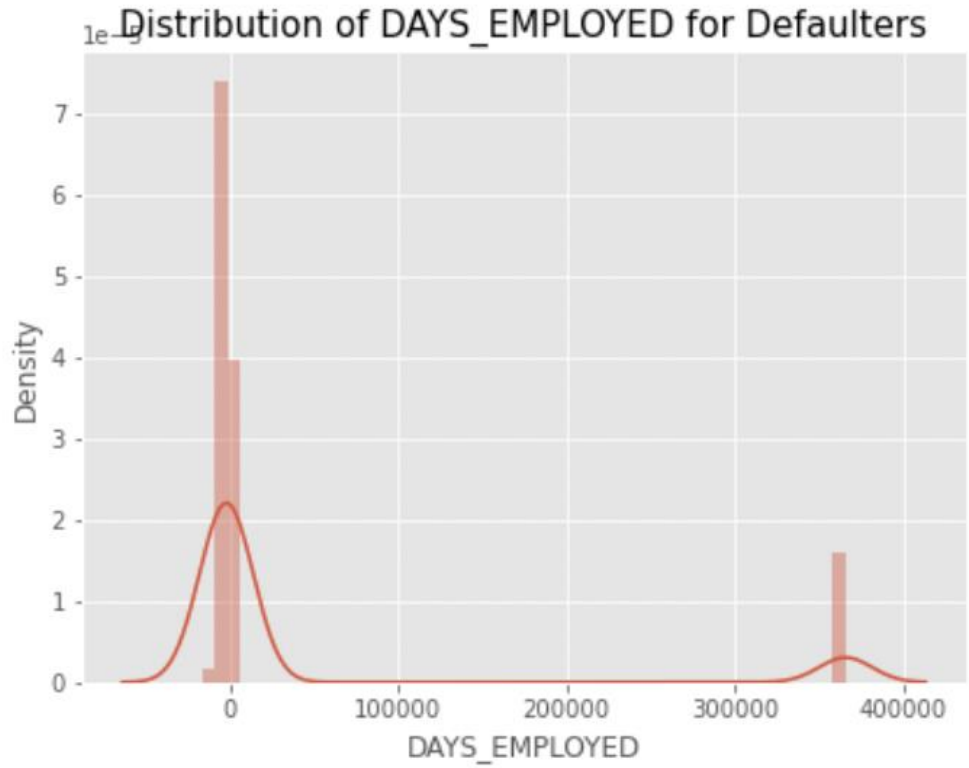
> Almost all of the Education categories are equally likely to default except for the higher educated ones who are less likely to default and secondary educated people are more likely to default

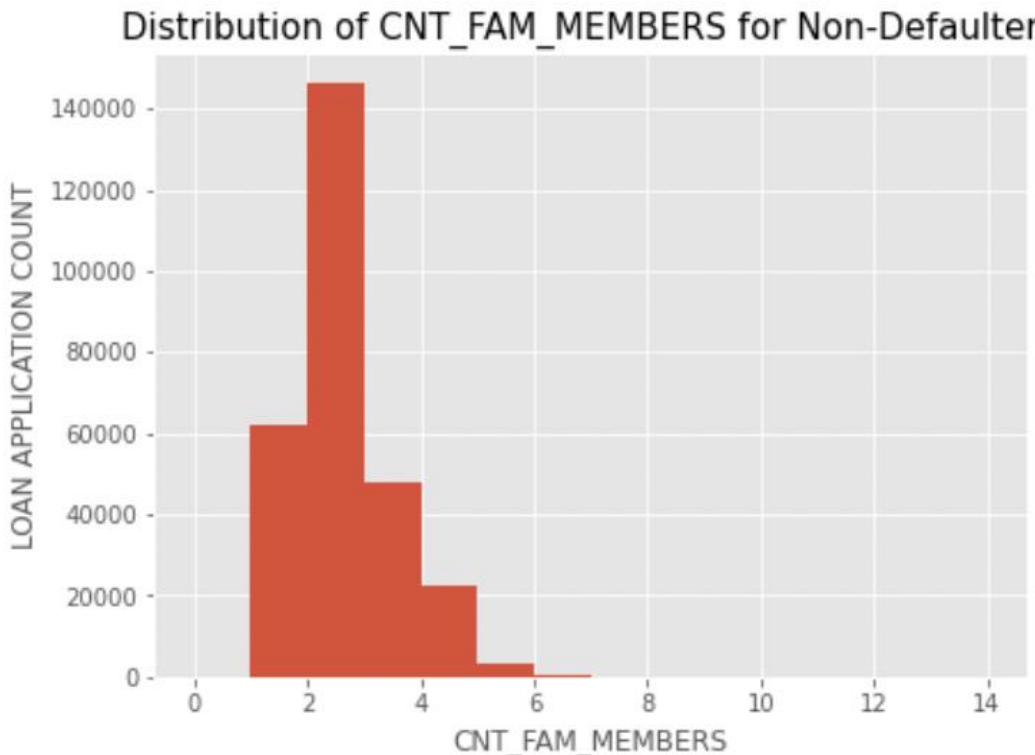


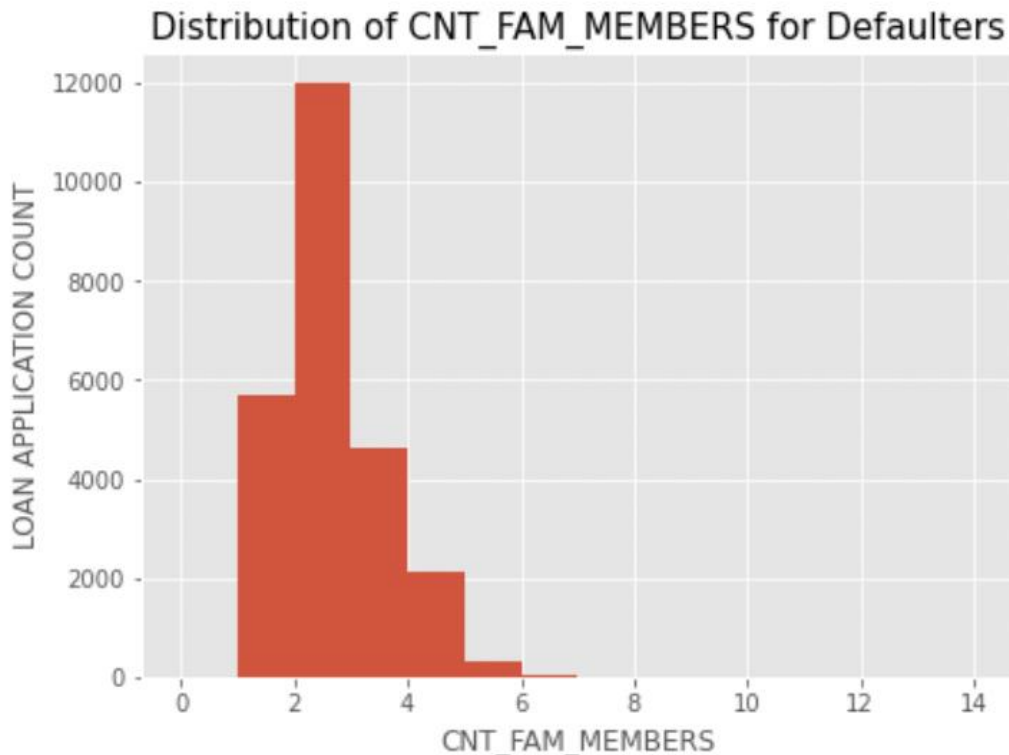


> Credit income ratio the ratio of AMT_CREDIT/AMT_INCOME_TOTAL.
> Although there doesn't seem to be a clear distiguish between the group which defaulted vs the group which didn't when compared using the ratio, we can see that when the CREDIT_INCOME_RATIO is more than 50, people default.


Distribution of DAYS_EMPLOYED for Non-Defaulters
1e-5
7
6
5
4
3
2
1
0
Density
0
100000
200000
300000
400000
DAYS_EMPLOYED


Distribution of DAYS_EMPLOYED for Defaulters
1e-5
7
6
5
4
3
2
1
0
Density
0
100000
200000
300000
400000
DAYS_EMPLOYED





> We can see that a family of 3 applies loan more often than the other families

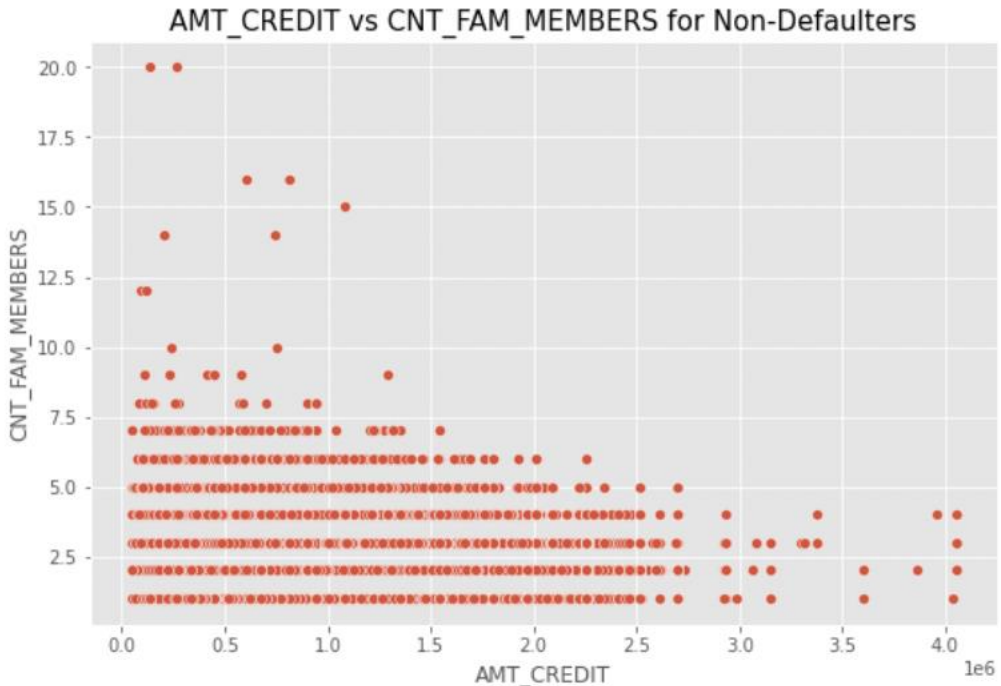
TOP 10 CORRELATIONS FOR DATA WHERE TARGET = 0

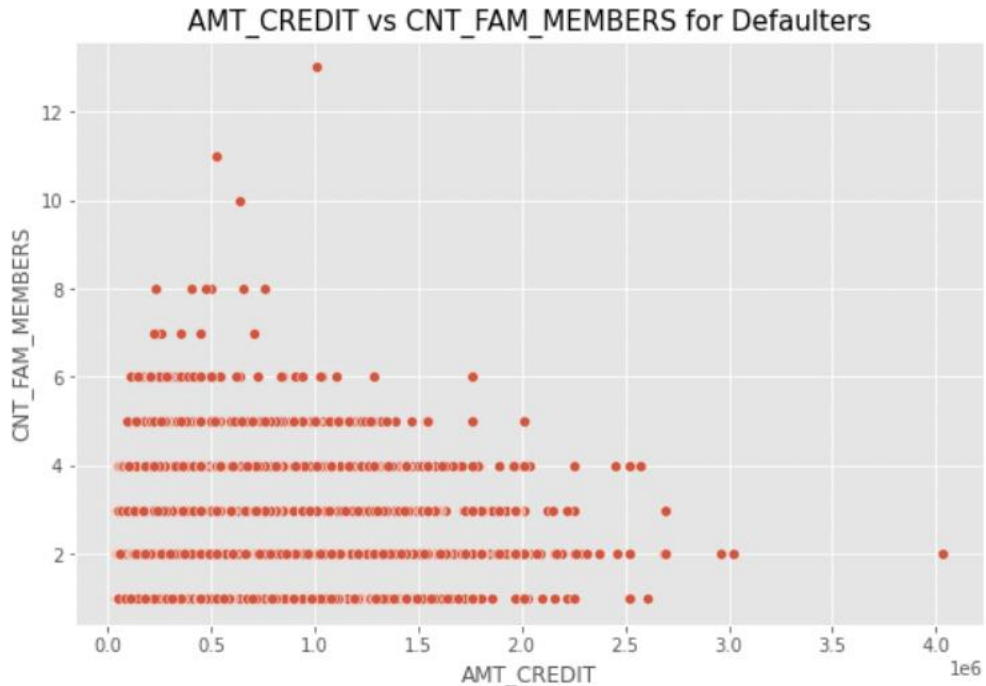
	Column1	Column2	Correlation	Abs_Correlation
308	AMT_GOODS_PRICE	AMT_CREDIT	0.987253	0.987253
297	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950148	0.950148
208	SOCIAL_CIRCLE_60_DAYS_DEF_PERC	SOCIAL_CIRCLE_30_DAYS_DEF_PERC	0.873003	0.873003
321	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686	0.776686
272	AMT_ANNUITY	AMT_CREDIT	0.771308	0.771308
74	CREDIT_INCOME_RATIO	AMT_CREDIT	0.648589	0.648589
310	AMT_GOODS_PRICE	CREDIT_INCOME_RATIO	0.628749	0.628749
273	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418954	0.418954
274	AMT_ANNUITY	CREDIT_INCOME_RATIO	0.391499	0.391499
309	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349461	0.349461

TOP 10 CORRELATIONS FOR DATA WHERE TARGET = 1

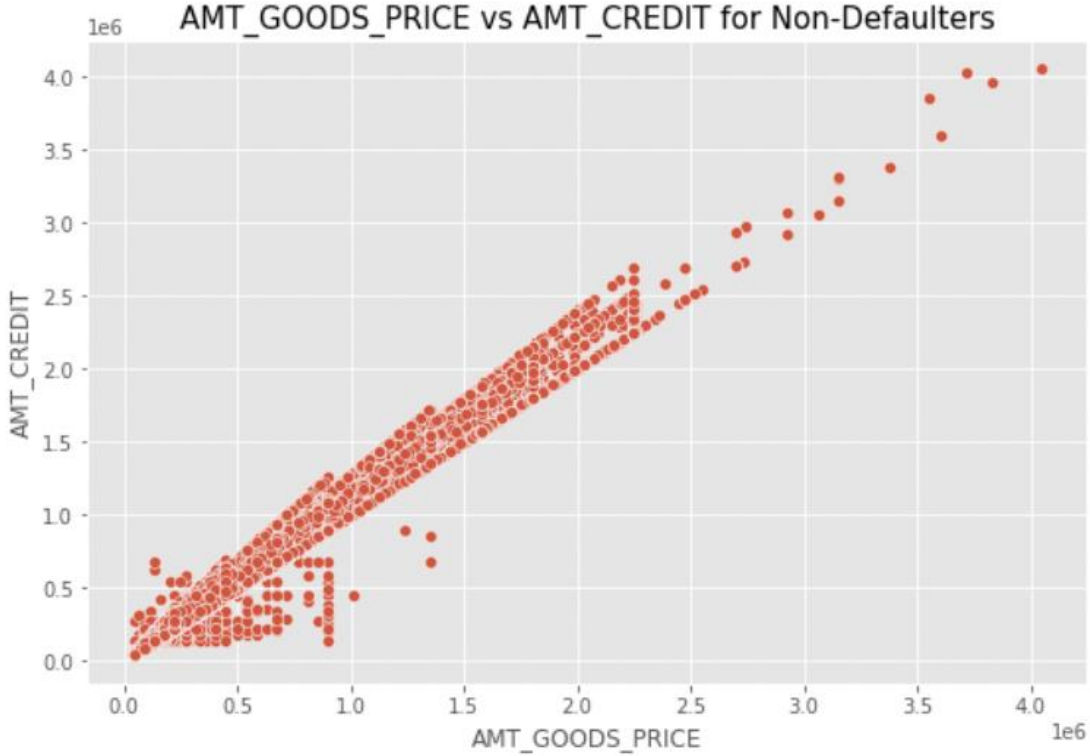
	Column1	Column2	Correlation	Abs_Correlation
308	AMT_GOODS_PRICE	AMT_CREDIT	0.983103	0.983103
297	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956637	0.956637
208	SOCIAL_CIRCLE_60_DAYS_DEF_PERC	SOCIAL_CIRCLE_30_DAYS_DEF_PERC	0.874562	0.874562
321	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699	0.752699
272	AMT_ANNUITY	AMT_CREDIT	0.752195	0.752195
74	CREDIT_INCOME_RATIO	AMT_CREDIT	0.639744	0.639744
310	AMT_GOODS_PRICE	CREDIT_INCOME_RATIO	0.623163	0.623163
274	AMT_ANNUITY	CREDIT_INCOME_RATIO	0.381298	0.381298
113	DAYS_REGISTRATION	DAYS_EMPLOYED	-0.188929	0.188929
149	CNT_FAM_MEMBERS	DAYS_EMPLOYED	-0.186561	0.186561

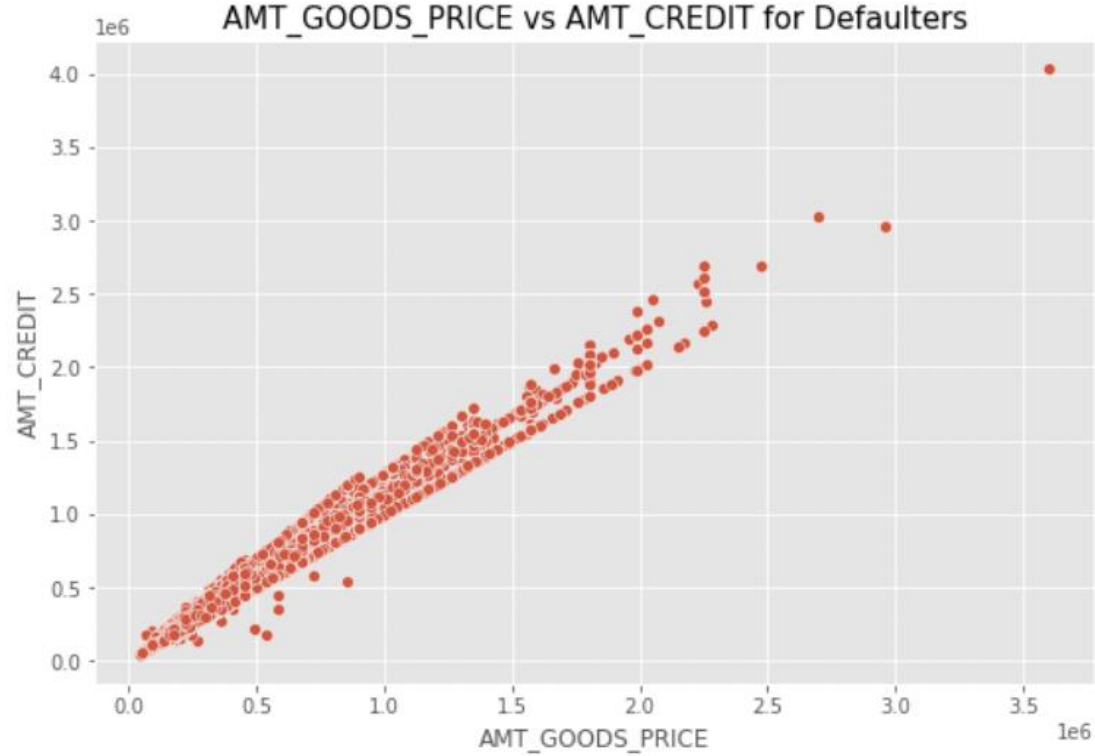
Bivariate Analysis on Application Data



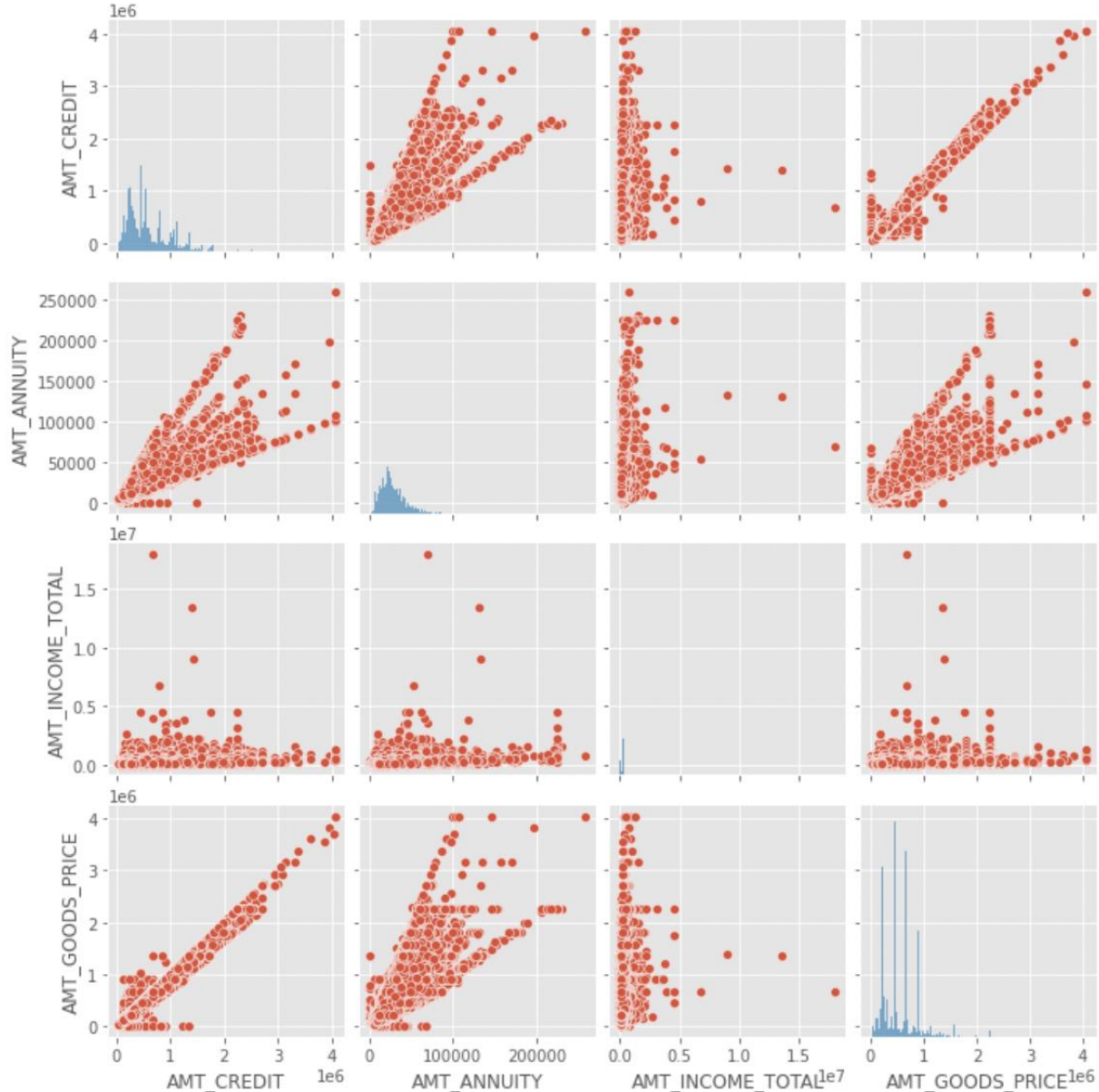


> We can see that the density in the lower left corner is similar in both the case, so the people are equally likely to default if the family is small and the AMT_CREDIT is low. We can observe that larger families and people with larger AMT_CREDIT default less often

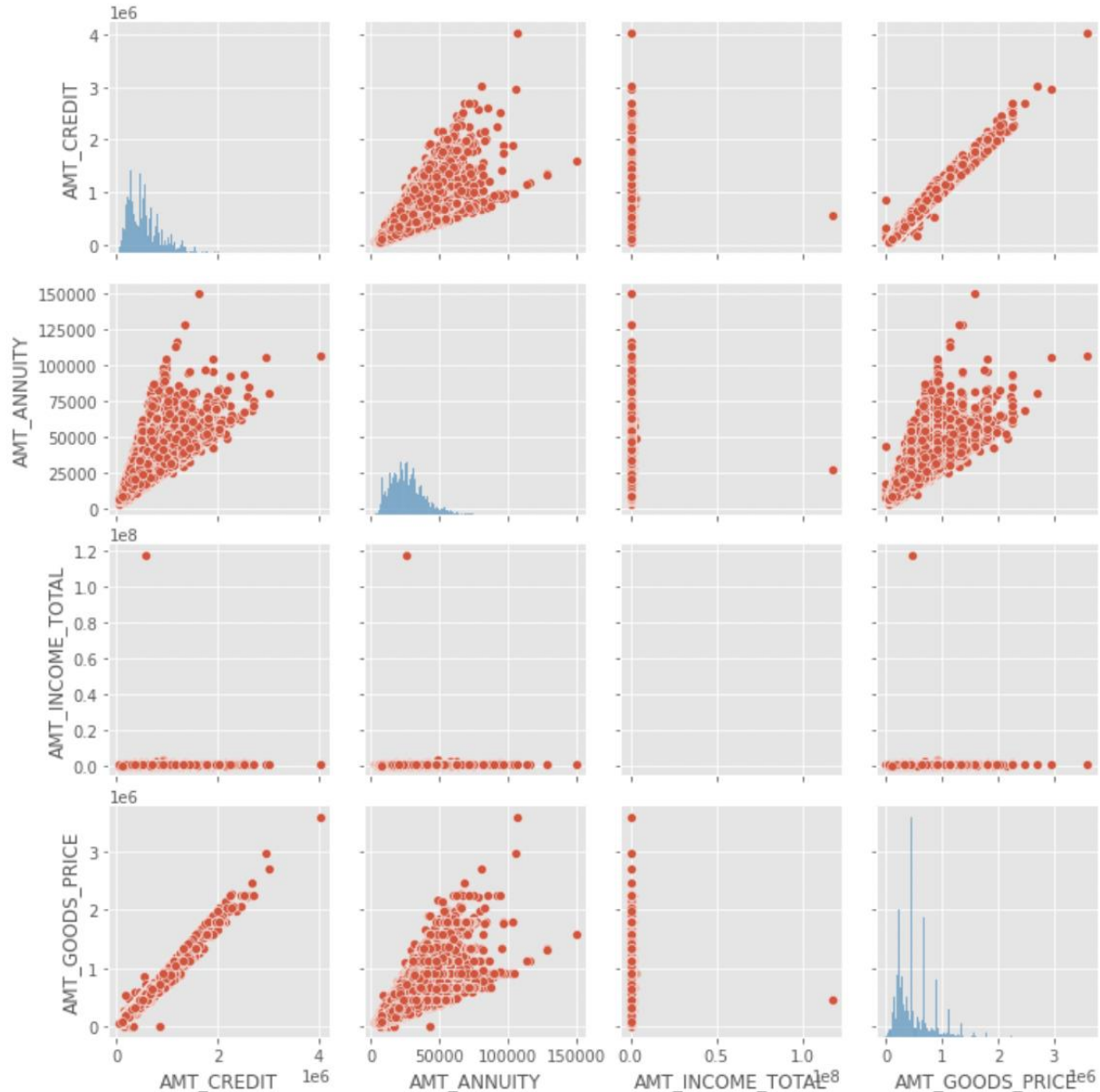

AMT_GOODS_PRICE vs AMT_CREDIT for Non-Defaulters
1e6
AMT_CREDIT
AMT_GOODS_PRICE
1e6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


AMT_GOODS_PRICE vs AMT_CREDIT for Defaulters
1e6
AMT_CREDIT
AMT_GOODS_PRICE
1e6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

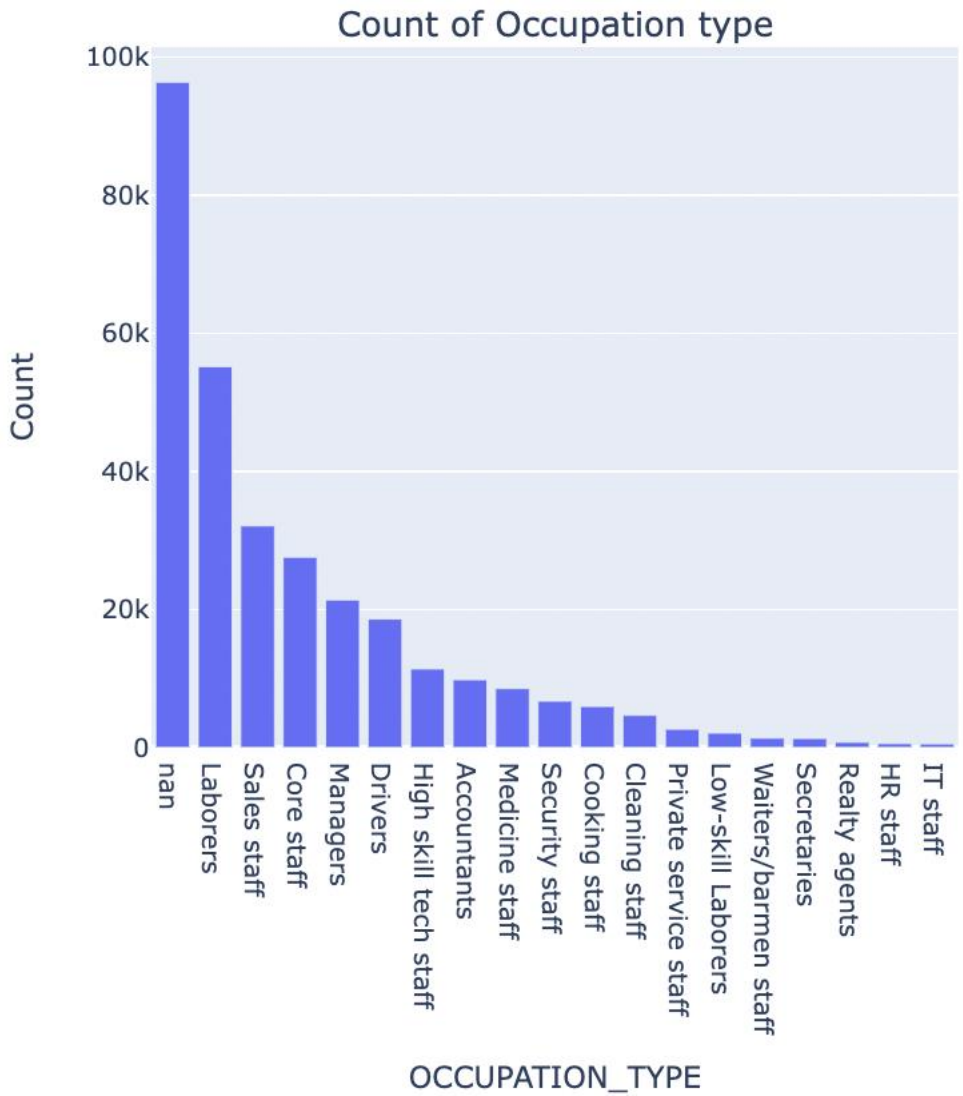
PAIR PLOT FOR TARGET 0

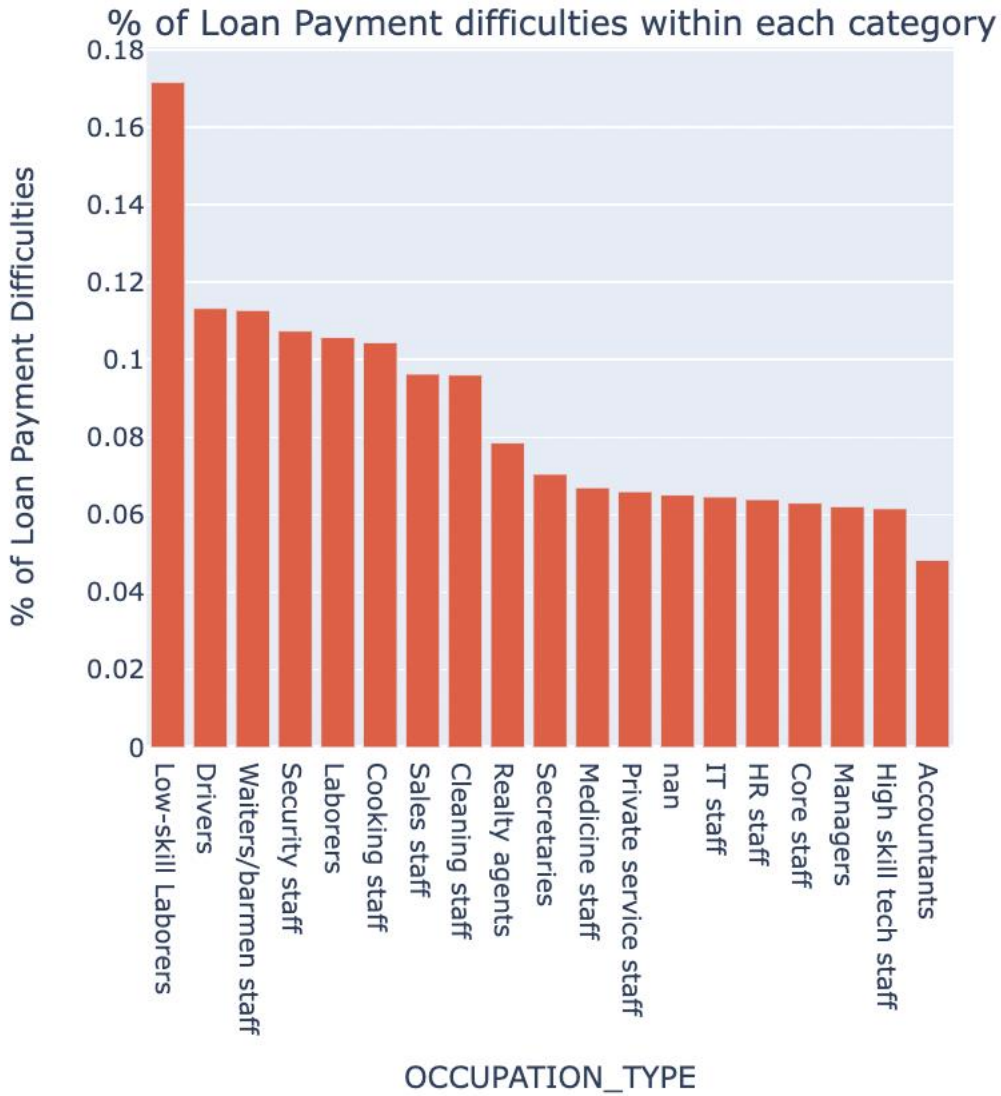


PAIR PLOT FOR TARGET 1

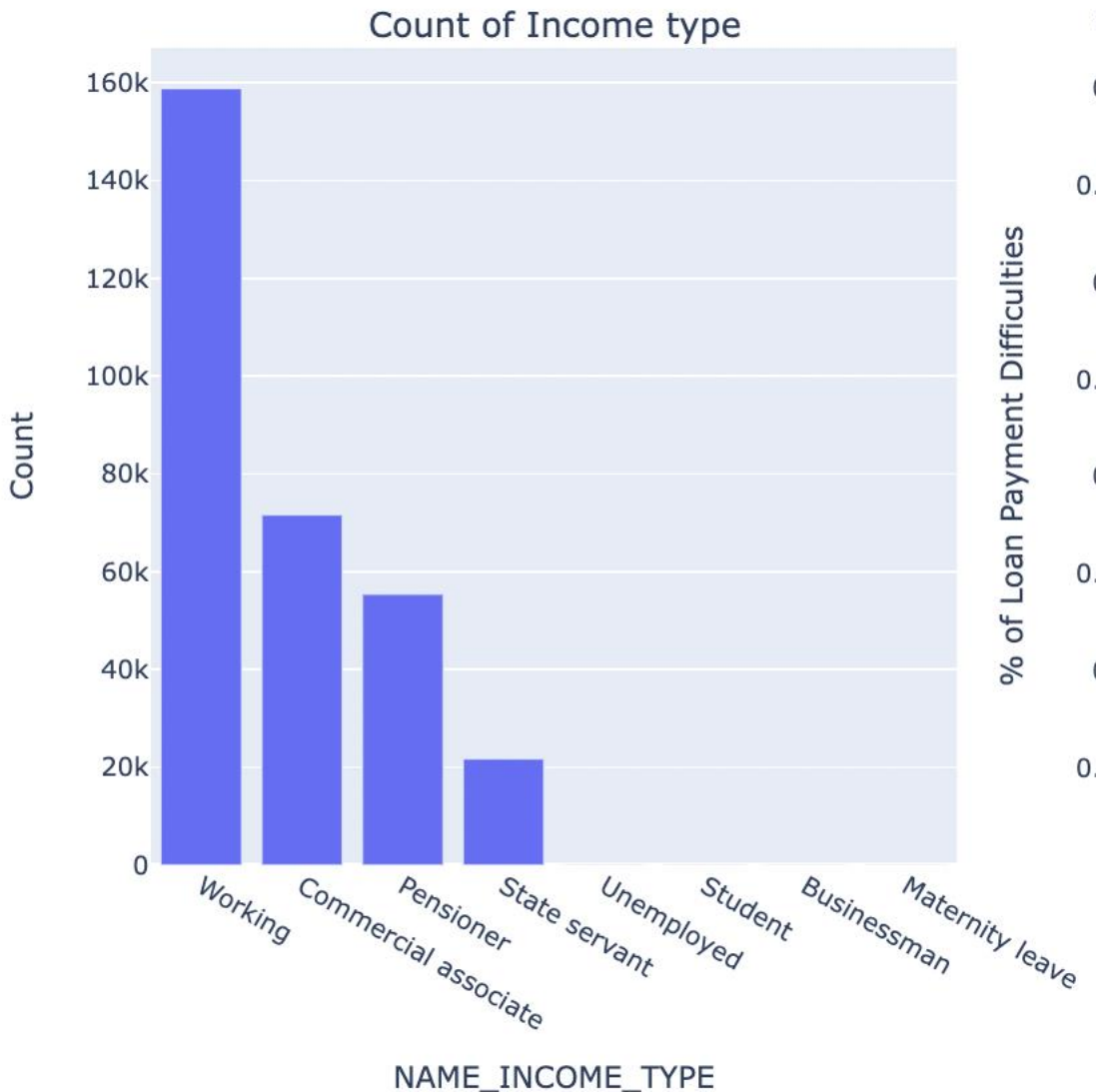


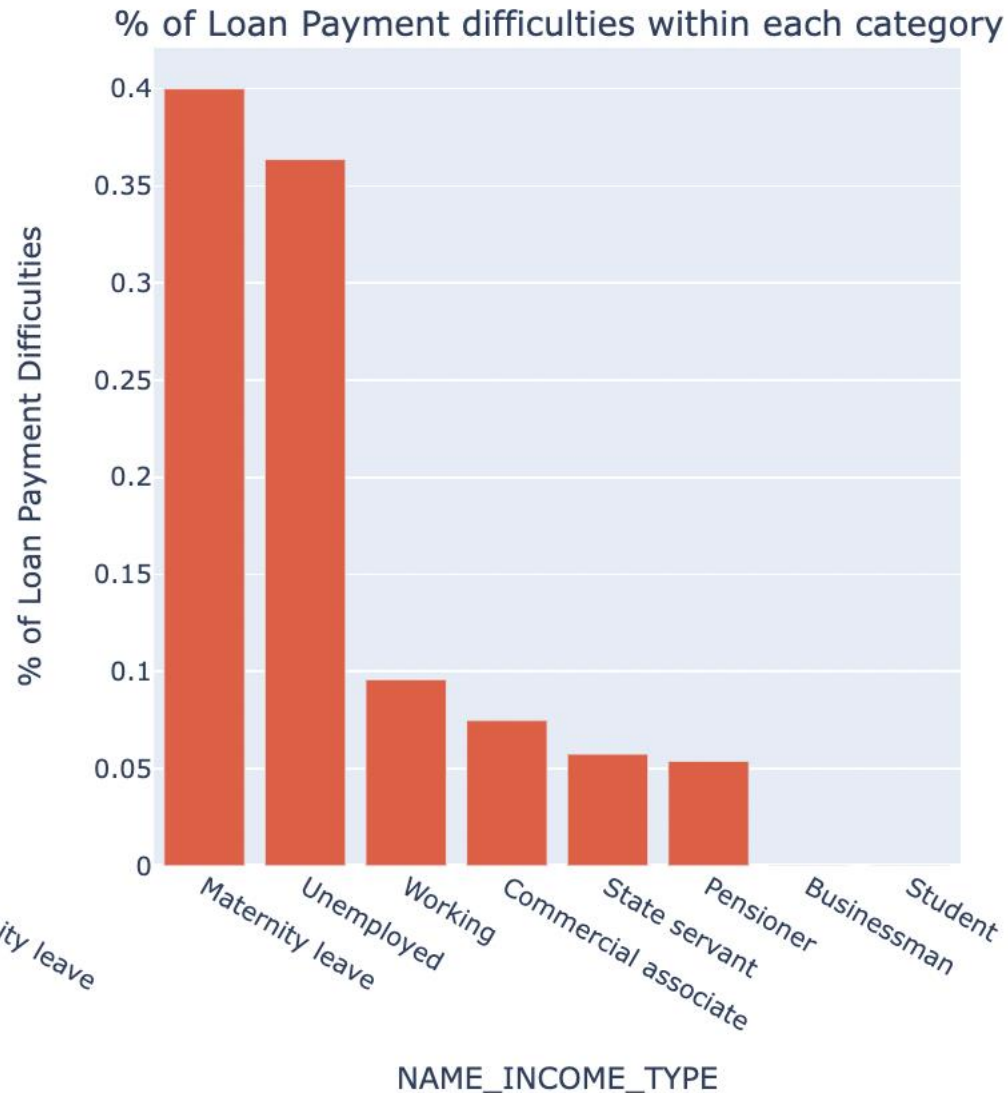
Occupation type



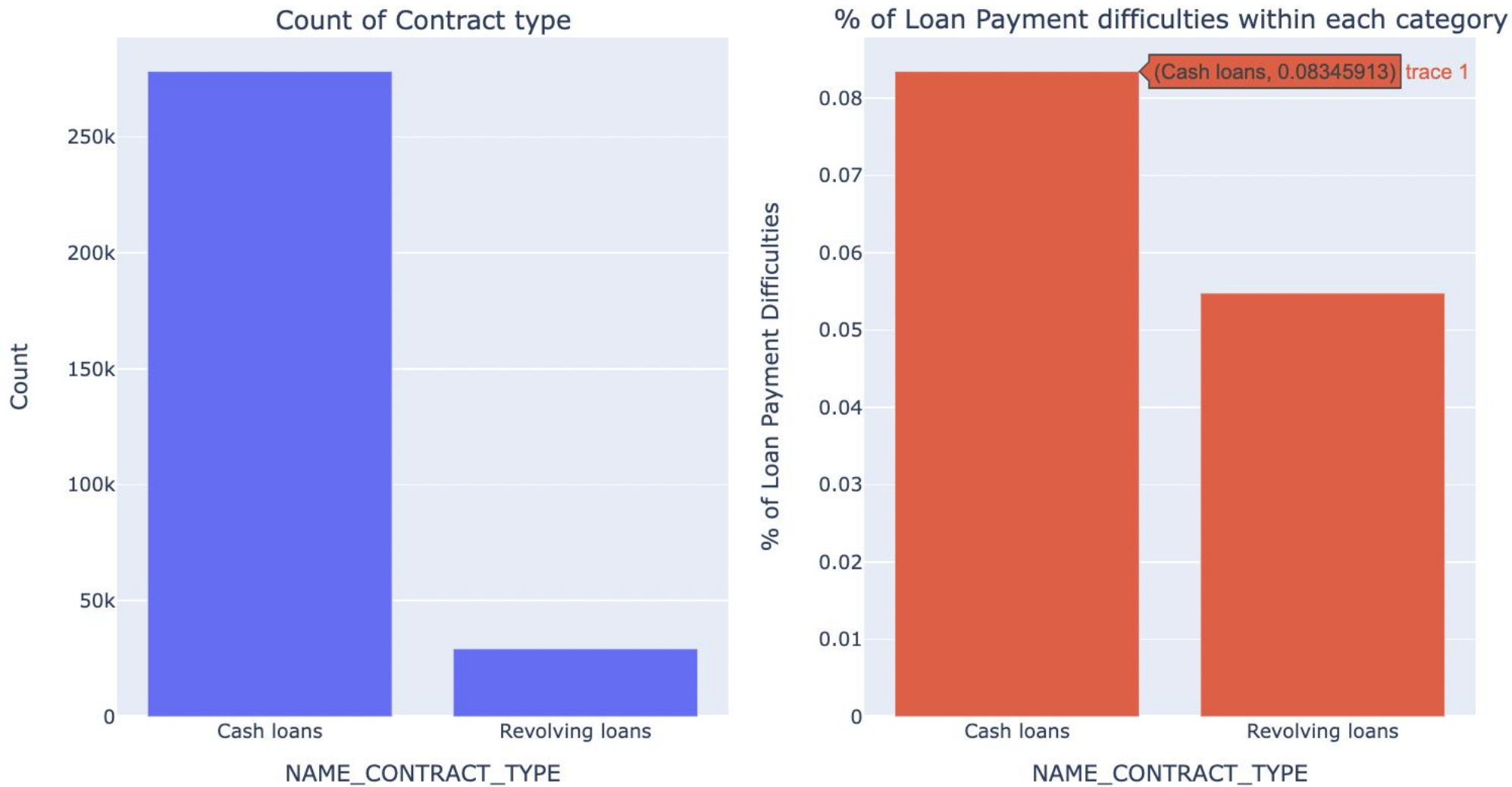


Income type

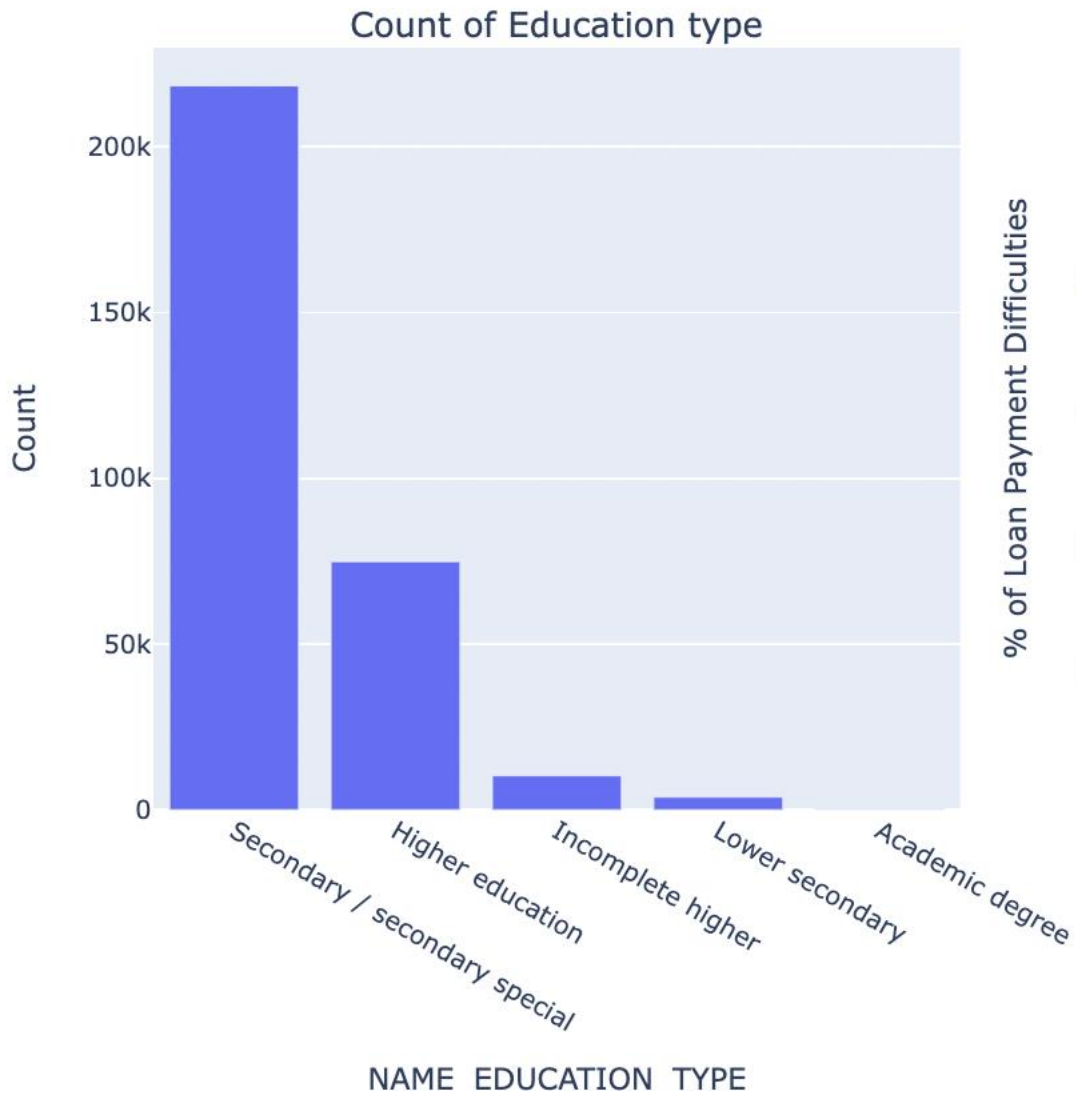

Count of Income type
Count
160k
140k
120k
100k
80k
60k
40k
20k
0
Working
Commercial associate
Pensioner
State servant
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE

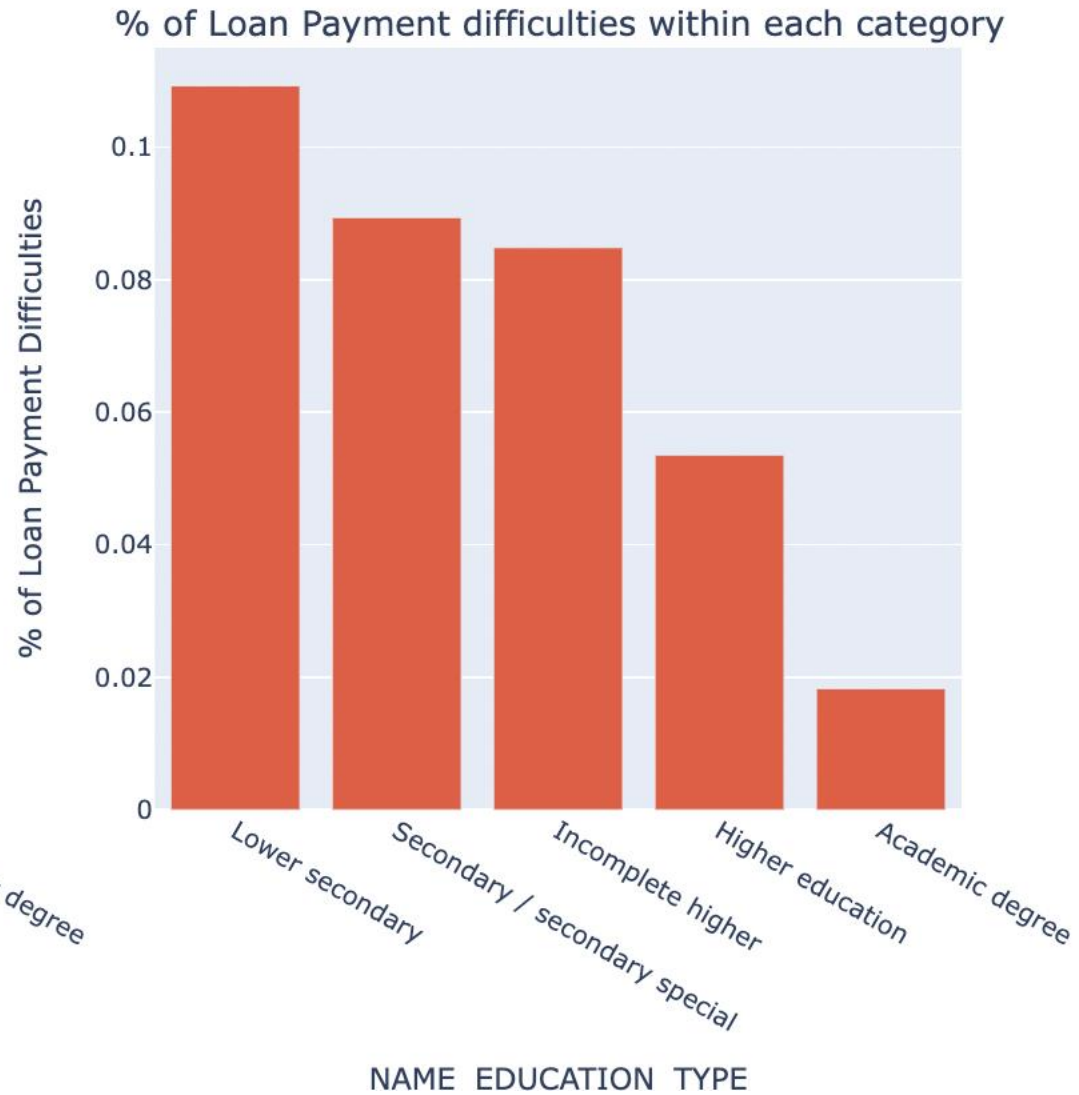

% of Loan Payment difficulties within each category
% of Loan Payment Difficulties
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Maternity leave
Unemployed
Working
Commercial associate
State servant
Pensioner
Businessman
Student
NAME_INCOME_TYPE

Contract type

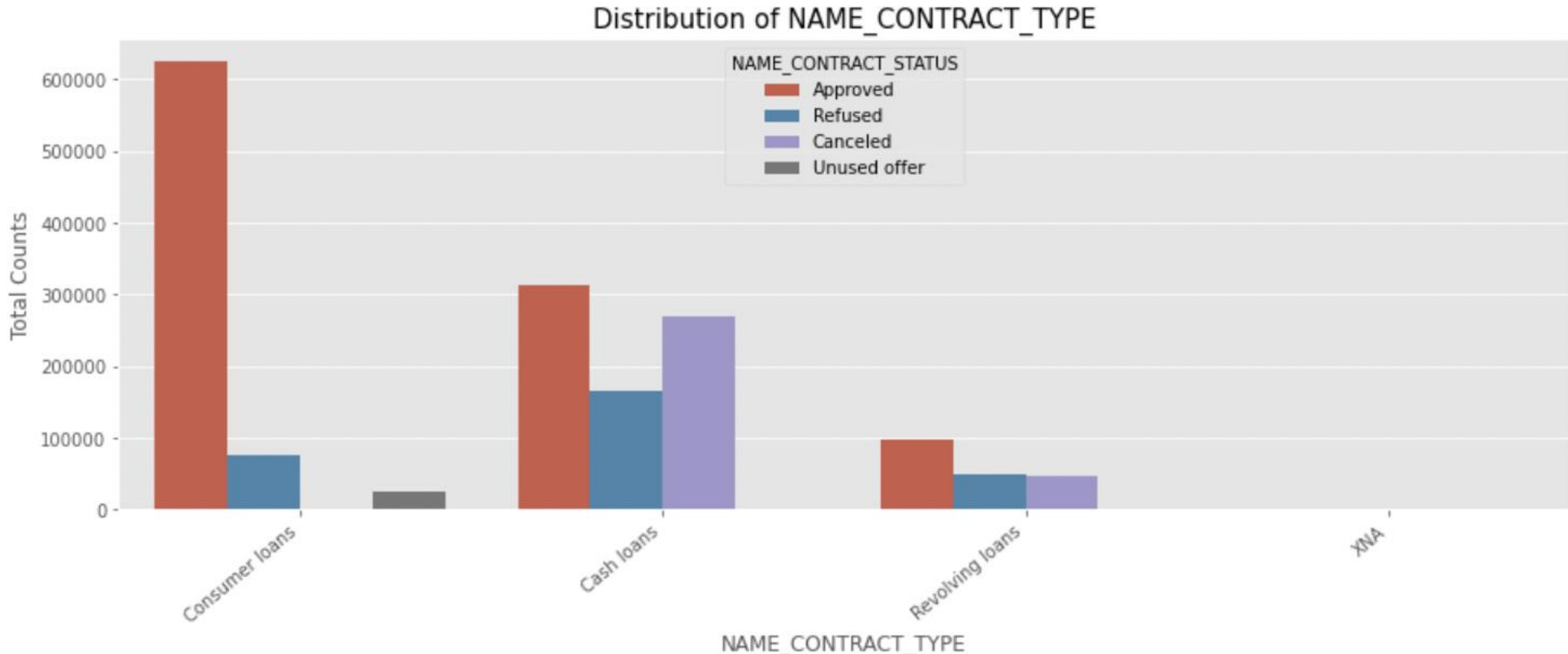


Education type

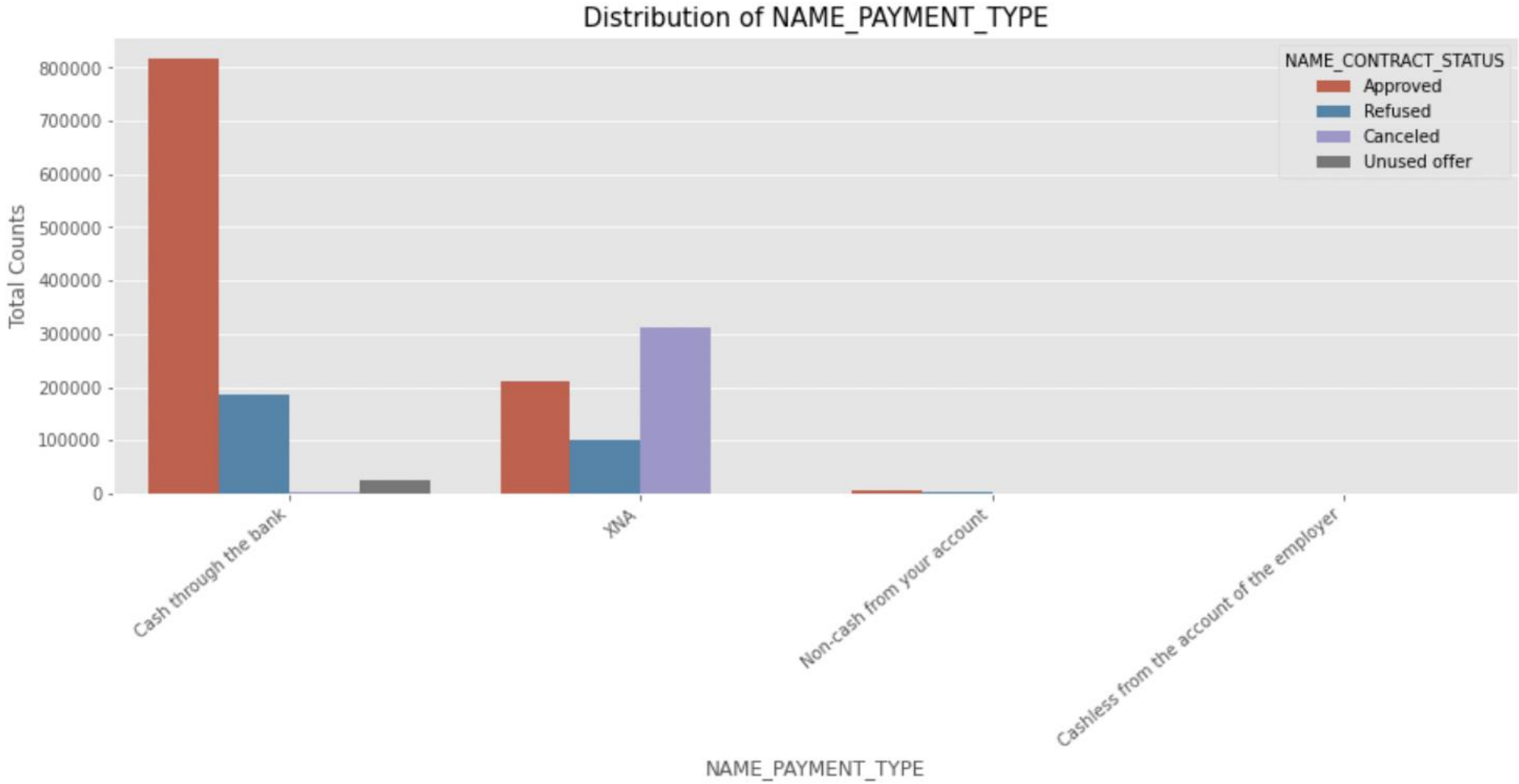




Univariate Analysis on Previous Application Data



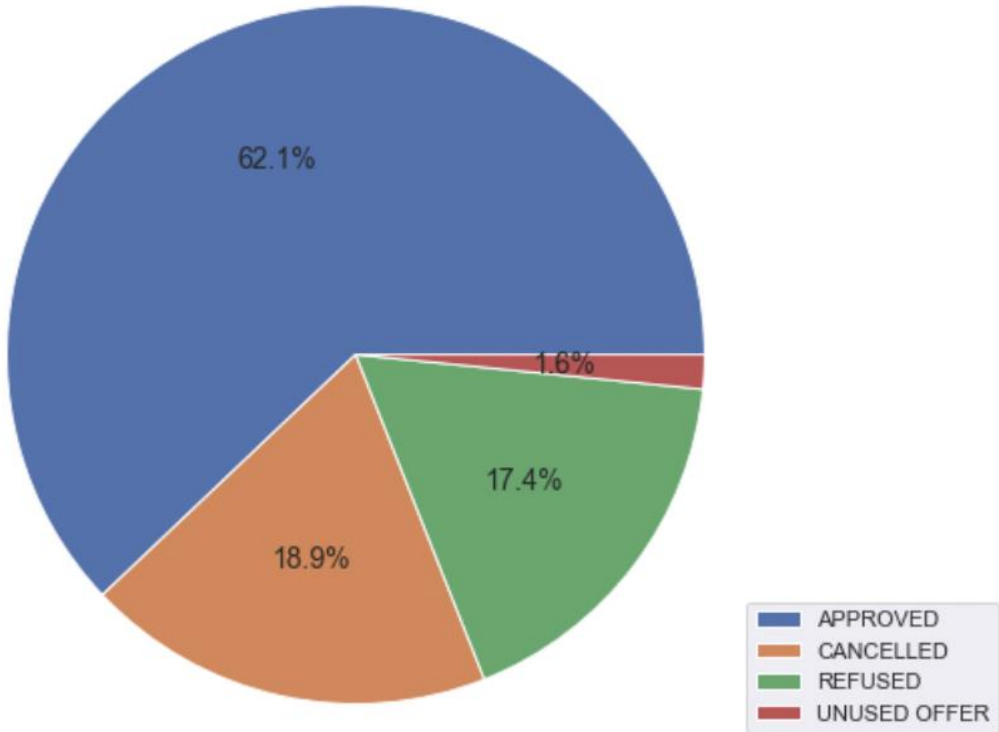
> From the above chart, we can infer that, most of the applications are for 'Cash loan' and 'Consumer loan'. Although the cash loans are refused more often than others.



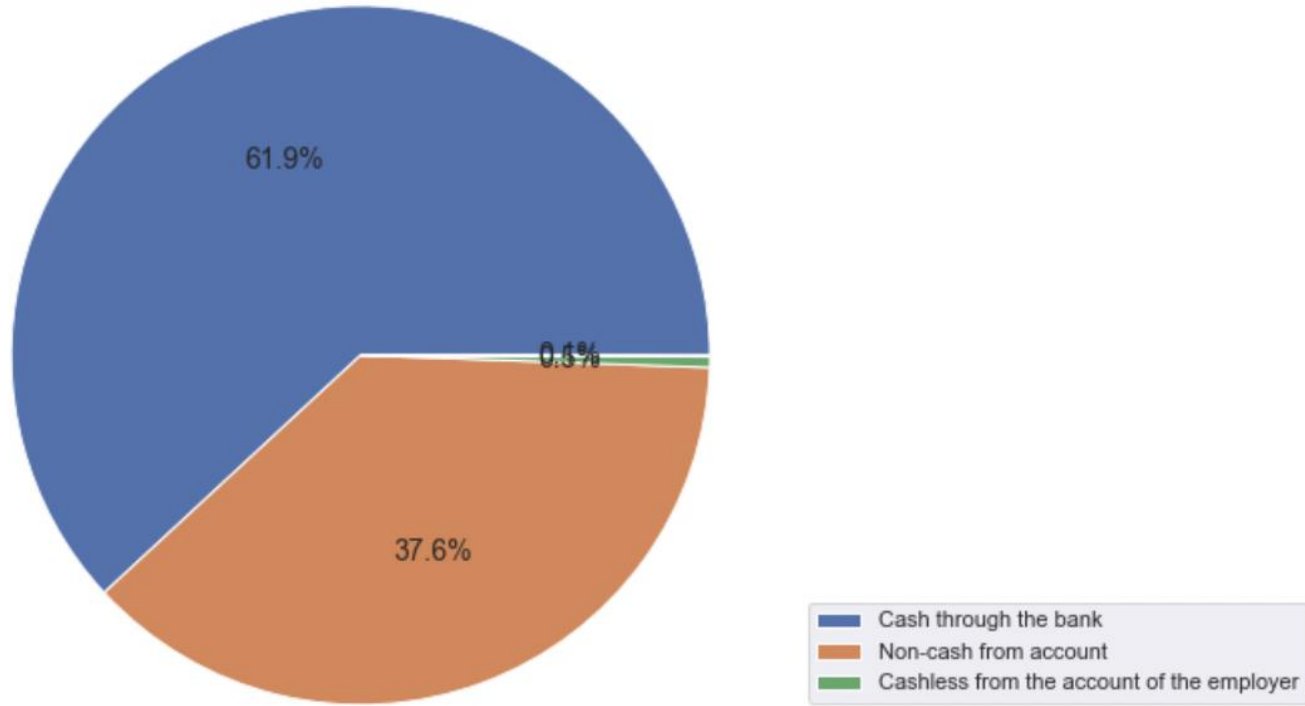
From the above chart, we can infer that most of the clients chose to repay the loan using the 'Cash through the bank' option
We can also see that 'Non-Cash from your account' & 'Cashless from the account of the employee' options are not at all popular in terms of loan repayment amongst the customers.

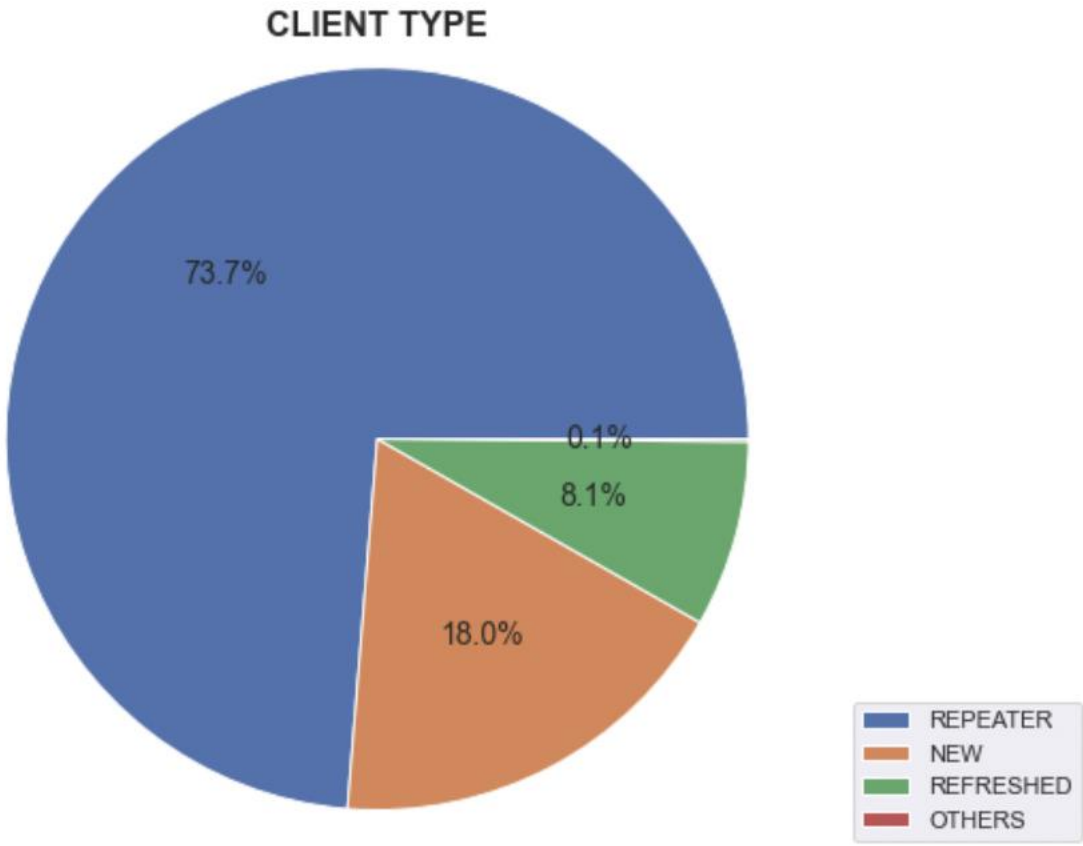
> Most of the loan applications are from repeat customers, out of the total applications 70% of customers are repeaters. They also get refused most often.

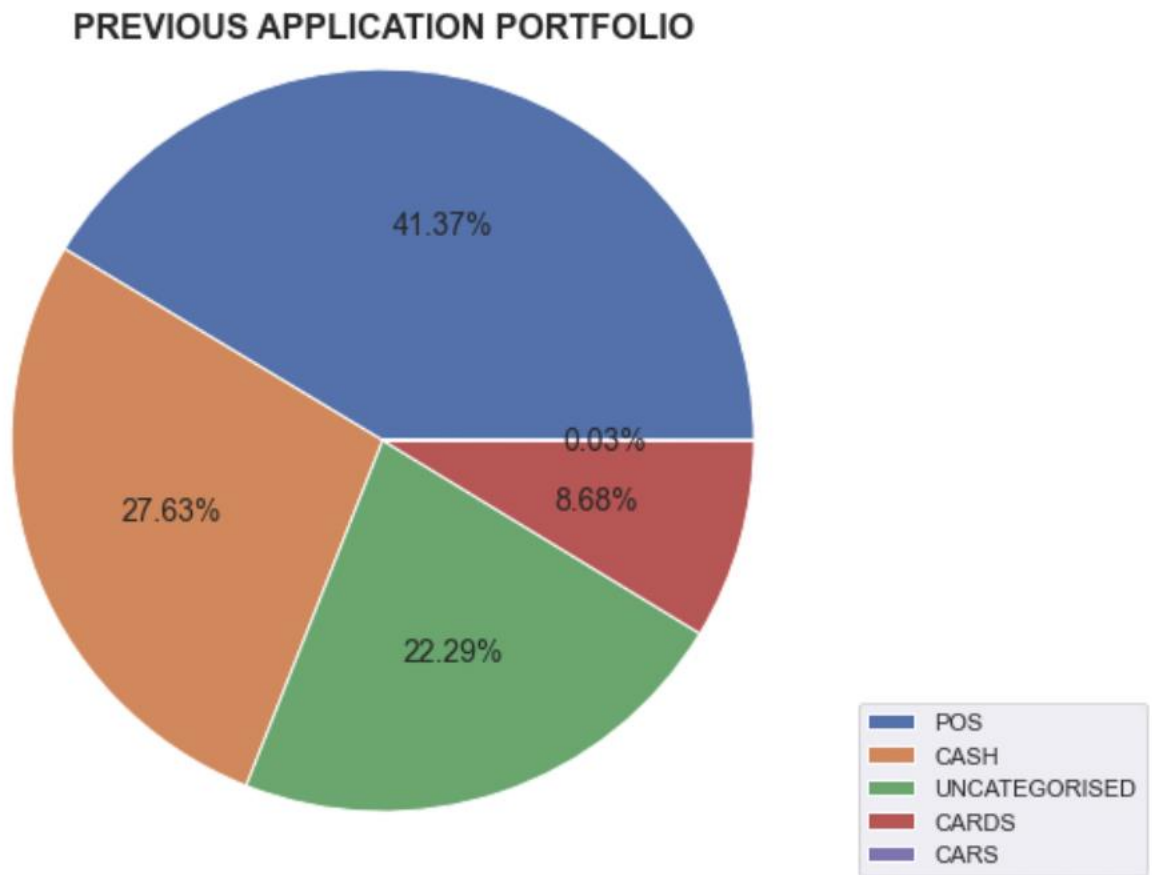
CONTRACT STATUS OF PREVIOUS APPLICATION



PREFFERED PAYMENT METHOD



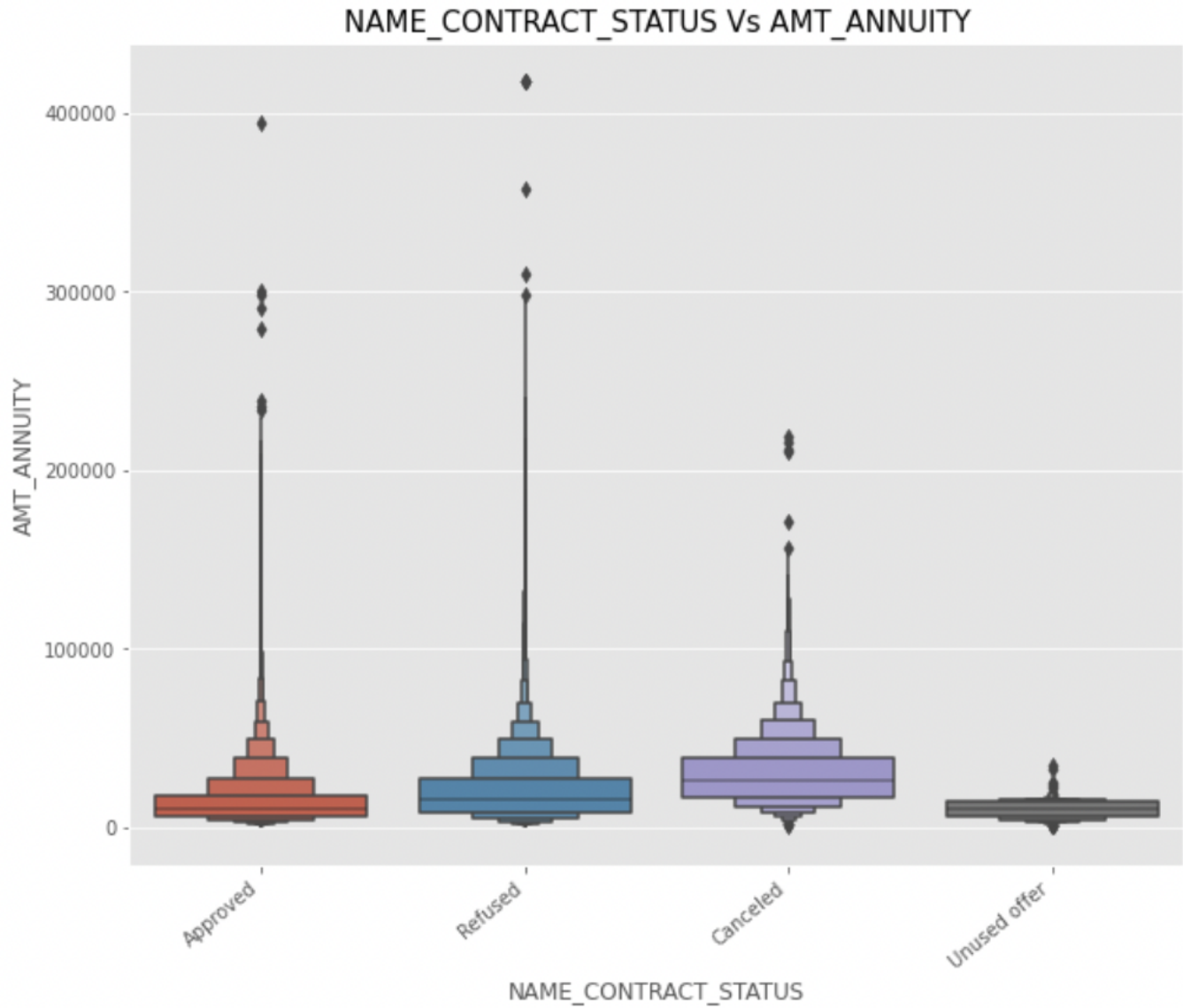

CLIENT TYPE
73.7%
0.1%
8.1%
18.0%
REPEATER
NEW
REFRESHED
OTHERS


PREVIOUS APPLICATION PORTFOLIO
41.37%
0.03%
8.68%
27.63%
22.29%
POS
CASH
UNCATEGORISED
CARDS
CARS

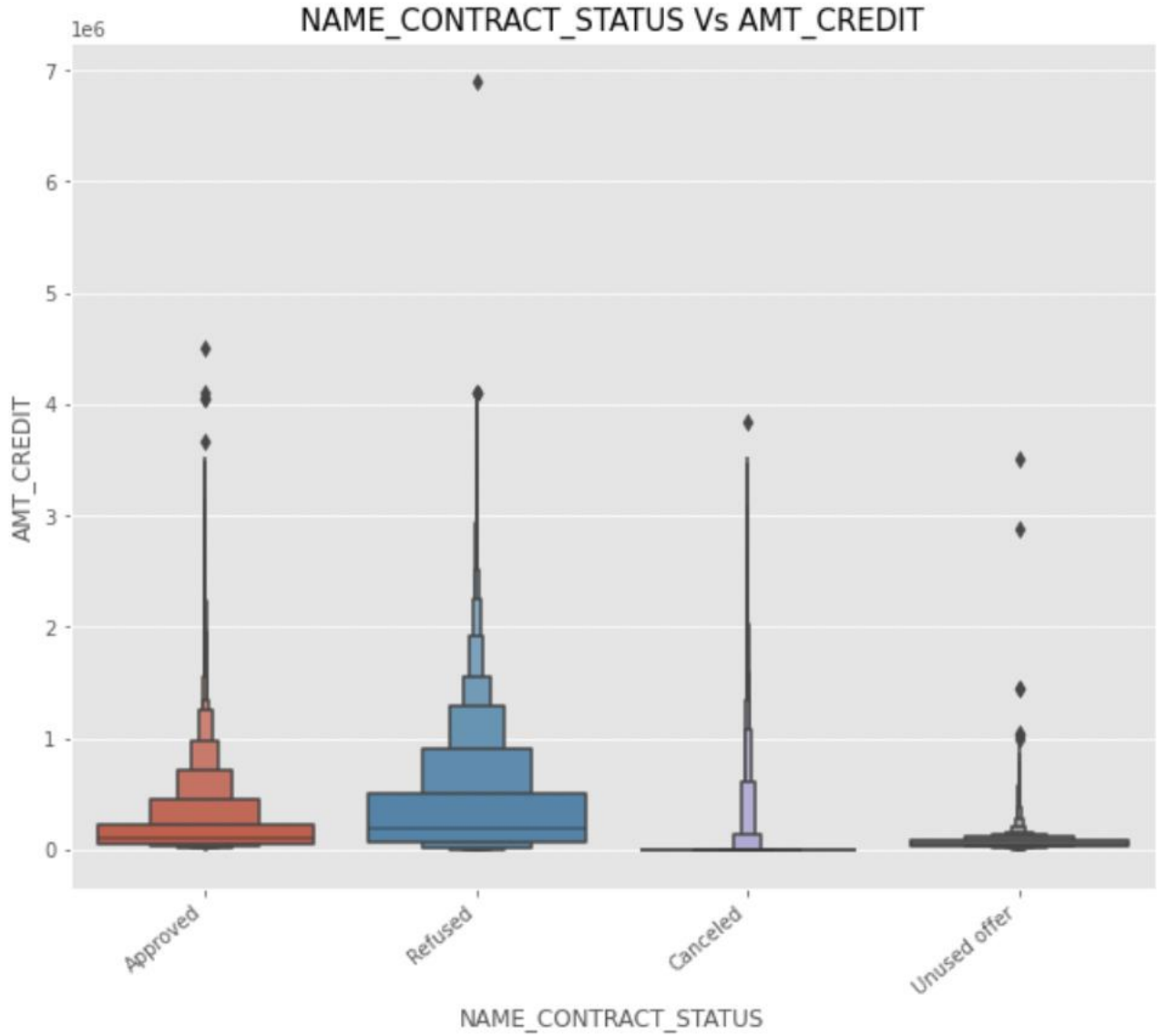
TOP 10 CORRELATIONS IN PREVIOUS APPLICATION DATA

	Column1	Column2	Correlation	Abs_Correlation
88	AMT_GOODS_PRICE	AMT_APPLICATION	0.999884	0.999884
89	AMT_GOODS_PRICE	AMT_CREDIT	0.993087	0.993087
71	AMT_CREDIT	AMT_APPLICATION	0.975824	0.975824
269	DAYS_TERMINATION	DAYS_LAST_DUE	0.927990	0.927990
87	AMT_GOODS_PRICE	AMT_ANNUITY	0.820895	0.820895
70	AMT_CREDIT	AMT_ANNUITY	0.816429	0.816429
53	AMT_APPLICATION	AMT_ANNUITY	0.808872	0.808872
232	DAYS_LAST_DUE_1ST_VERSION	DAYS_FIRST_DRAWING	-0.803494	0.803494
173	CNT_PAYMENT	AMT_APPLICATION	0.680630	0.680630
174	CNT_PAYMENT	AMT_CREDIT	0.674278	0.674278

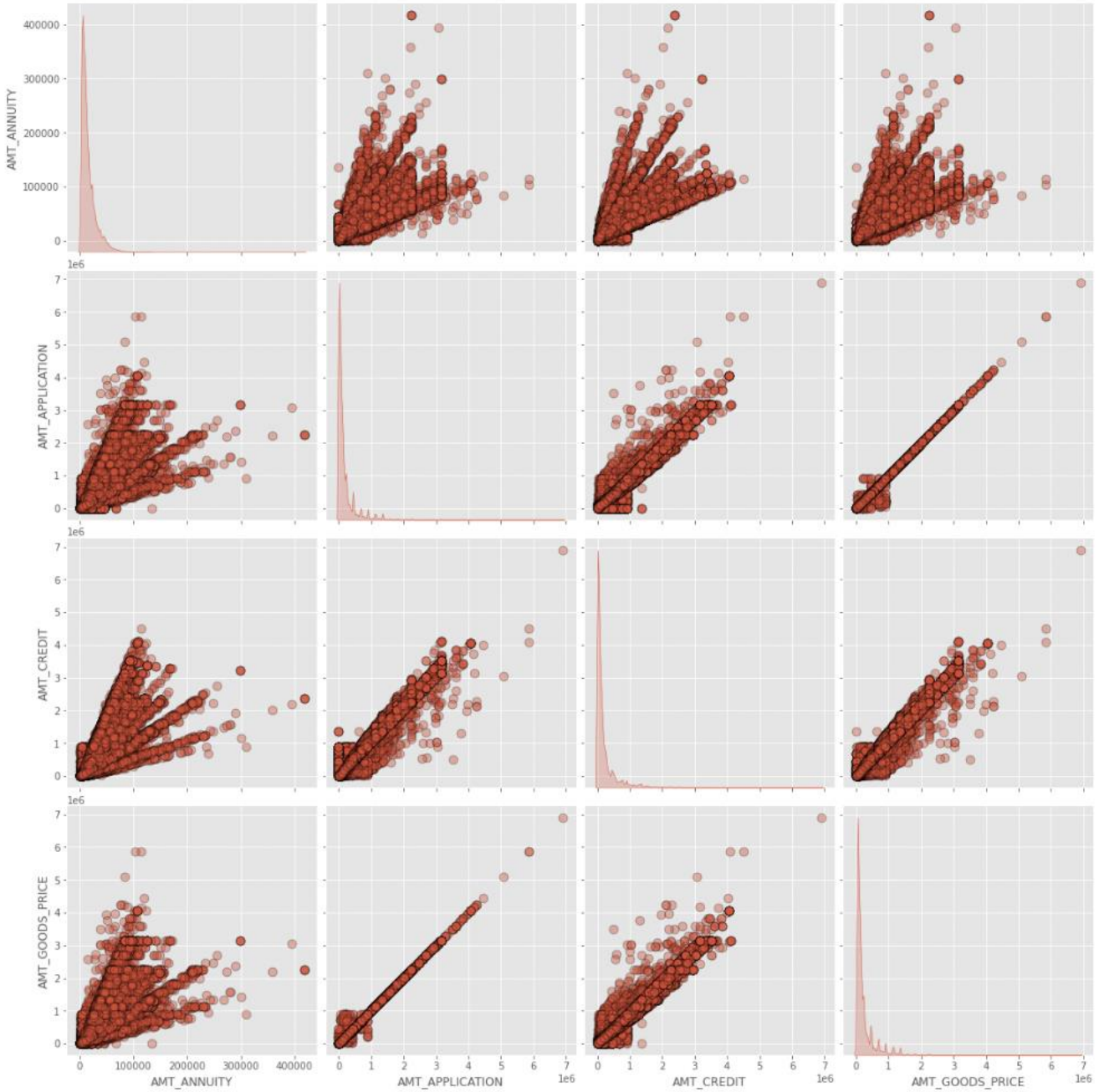
Bivariate Analysis on Previous Application Data



> From the above plot we can see that loan application for people with lower AMT_ANNUITY gets canceled or Unused most of the time.
> We also see that applications with too high AMT ANNUITY also got refused more often than others.



> We can infer that when the AMT_CREDIT is too low, it get's cancelled/unused most of the time.



Annuity of previous application has a very high and positive influence over: (Increase of annuity increases below factors)
(1) How much credit did client asked on the previous application
(2)Final credit amount on the previous application that was approved by the bank
(3) Goods price of good that client asked for on the previous application.

For how much credit did client ask on the previous application is highly influenced by the Goods price of good that client has asked for on the previous application

Final credit amount disbursed to the customer previously, after approval is highly influence by the application amount and also the goods price of good that client asked for on the previous application.

Analysis on Merged Application Data

Target variable (0 - Non Defaulter 1 - Defaulter)

We can see that the people who were approved for a loan earlier, defaulted less often where as people who were refused a loan earlier have higher chances of defaulting.

INFERENCES

After the EDA we have come to the following conclusions:-

Bank should give out more revolving loans as people with revolving loans are less likely to default.

Consumer loans are mostly likely to get approved by bank. However, a lot of people are cancelling and leaving their loans unused. This reason should be further investigated.

Females are less likely to default than males and also present a greater market base for the bank.

Education level of a customer is a good indicator of a persons capability to pay back. Higher the education level more reliable is the customer.

Age is also an important factor. Older people are less likely to default, however that doesn't mean young people should be excluded out of customer base, they should be properly educated on the entire loan cycle.

Repeated business is good for bank. For large share in the total population currently their default percentage is low.

There is a high correlation between goods price and amount credited in both current and previous application. Thus people are getting the required amount of loan.