Risk Analytics in Banking and Financial services

- Through Exploratory Data Analysis

By:

Rahul Raj & Shivangi Mishra

Business Understanding

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter.

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.

Business Understanding: Data Segregation

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample.
- **All other cases:** All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

- **Approved:** The Company has approved loan Application
- **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
- **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
- **Unused offer:** Loan has been cancelled by the client but on different stages of the process.

Business objectives

- Case study aims to identify patterns which indicates if client has potential to become a defaulter or not.

- The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default

Analysis: Data Available

- *'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.**
- *'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.**
- *'columns_description.csv'* is data dictionary which describes the meaning of the variables.
- Size of Available data:

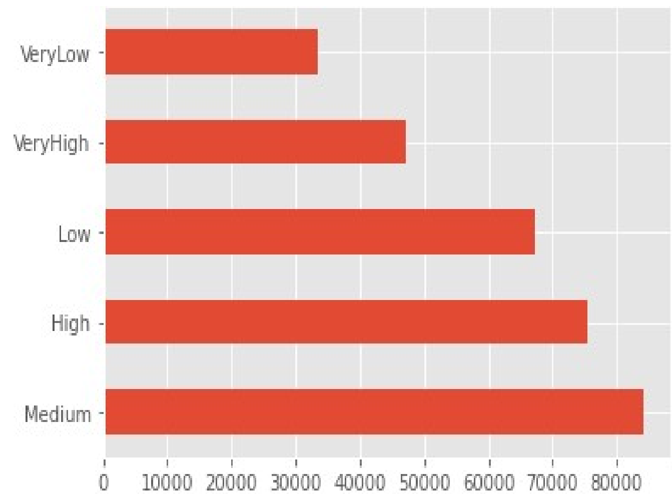
```
Shape of application DataFrame :  (307511, 122)

Shape of prev_application DataFrame :  (1670214, 37)
```

Defining Income levels for Analysis

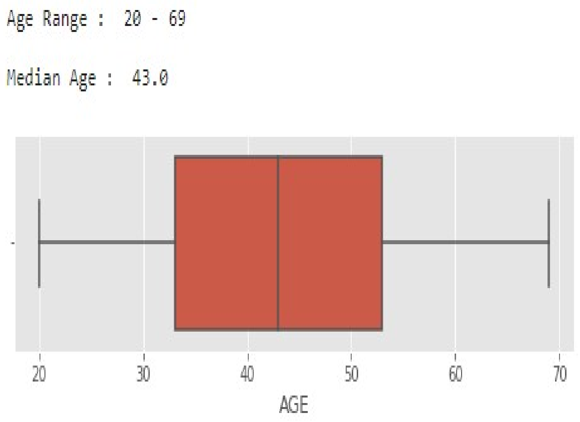
Variable for Binning:

1).Income into the following levels bases on quantiles

- Very Low - [0.0 – 10%]
- Low - [10% - 30%]
- Medium - [30% - 60%]
- High - [60%- 80%]
- Very High - [80% - 100%]

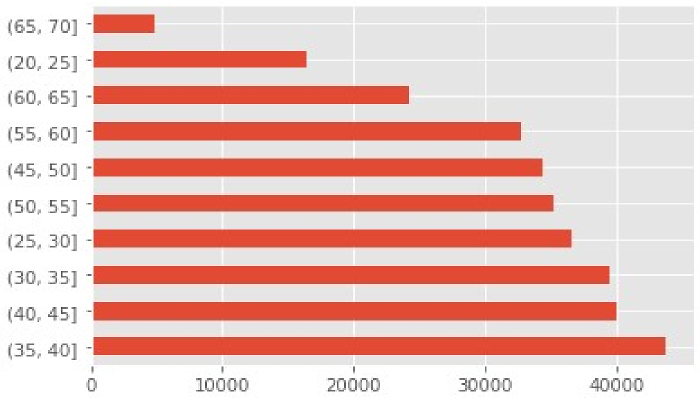


Defining Age Group for Analysis



- 2). Age into groups with an interval of 5 years

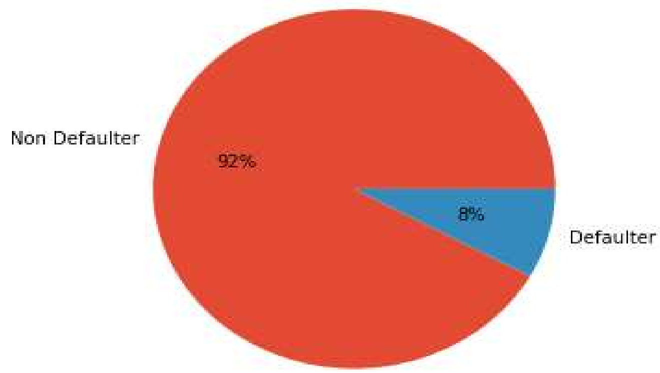
- Binning the age in years to Age groups



Analysis: Understanding the Target Areas

- Here is the distribution of target areas, as shown below:

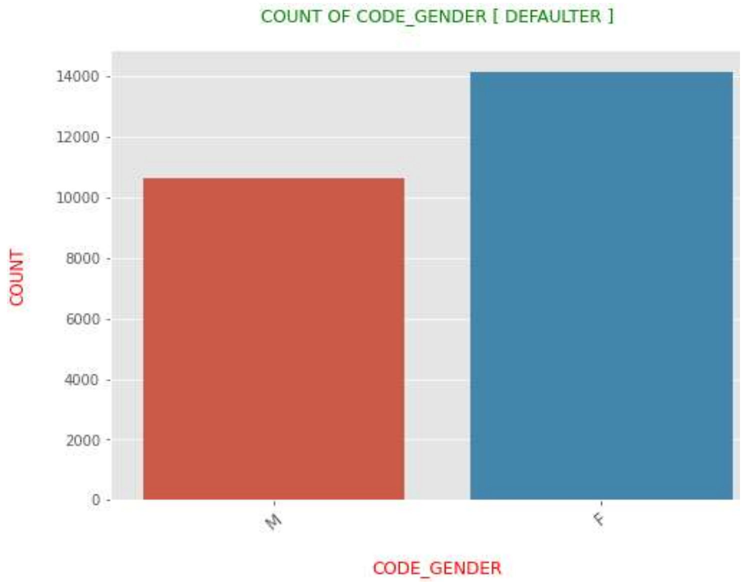
	Count	% Distribution
Non Defaulter	282682	91.93
Defaulter	24825	8.07



- Clearly 8% people has got defaulted on their loans on the other hand 92% people are good at the payments.

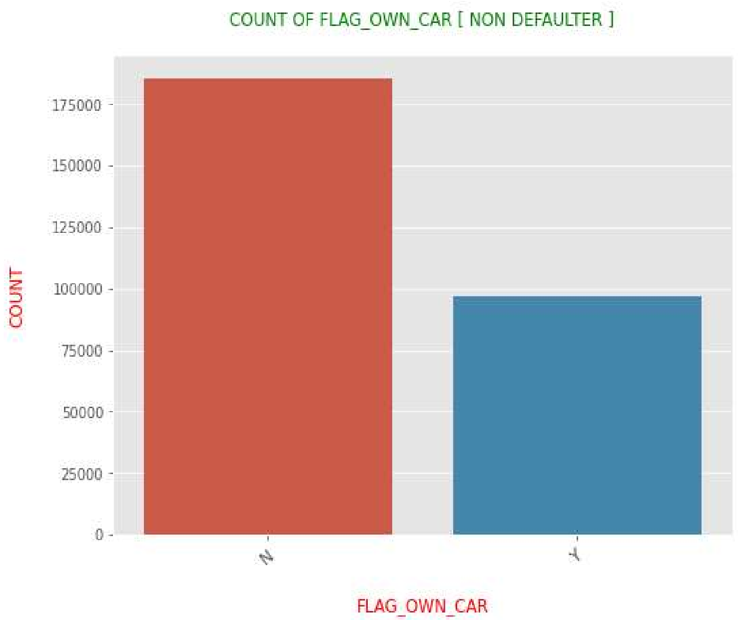
Analysis: Targeting areas on the basis of “Gender”

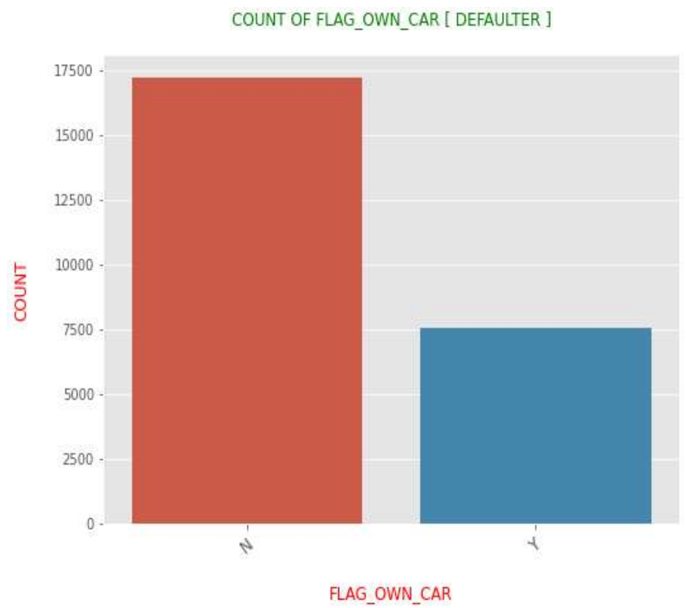
- Let’s plot graph on the basis of “Gender”, to get extensive view.



- Female Applicants contribute about 67% in the pool of non-defaulters application whereas male applicants contributes about 33 %
- Since the numbers of female applicants are significantly high, they tent to contribute 57% of the defaulters contribution
- It is evident that the rate of default is lower in female applicants in comparison with male applicants

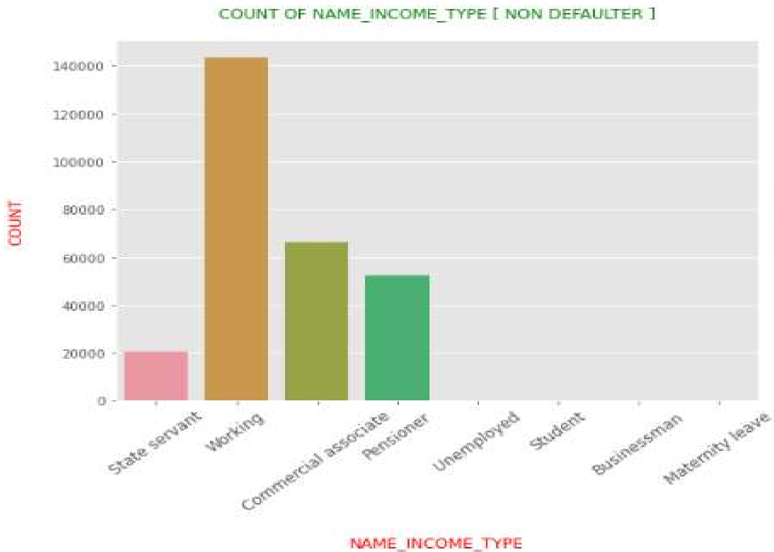
Analysis: Targeting areas on the basis of “Car Ownership”

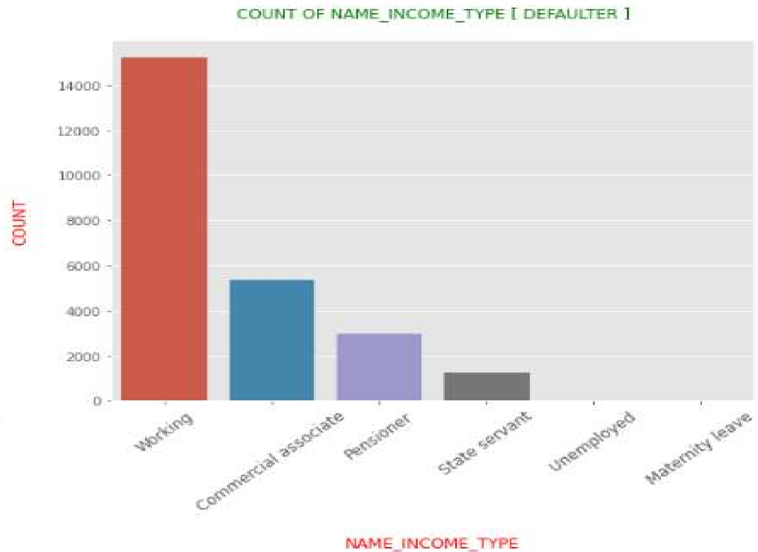




1. People with cars contribute about 66% to the non-defaulters while 69 % to the defaulters.
2. People who have car tend to default more often than those people don’t have cars.
3. Rate of defaulting is low for people having car compared to people who don't.

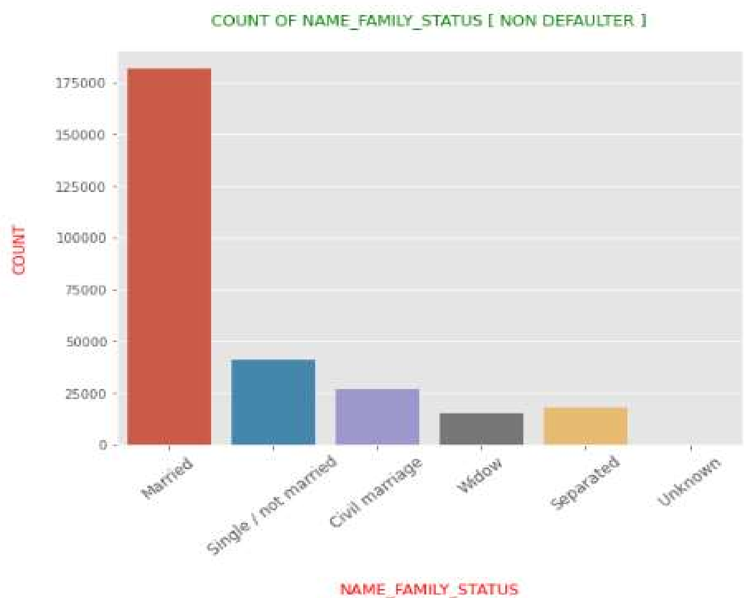
Analysis: Targeting areas on the basis of “Income Type”

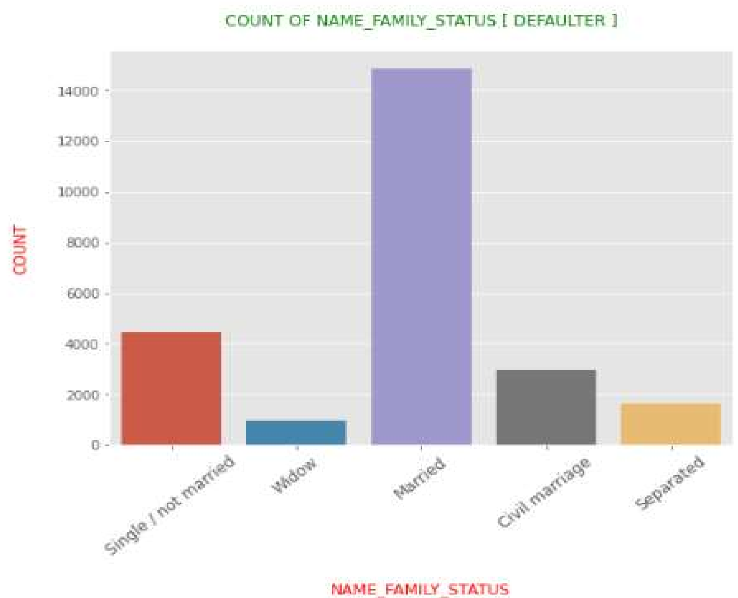




- Students & Businessmen tend to be Non-Defaulters
- Most of the loans are distributed to working class people
- Working class people contribute 51% to non defaulters while they contribute to 61% of the defaulters.
- Chance of getting defaulted is high for working class people

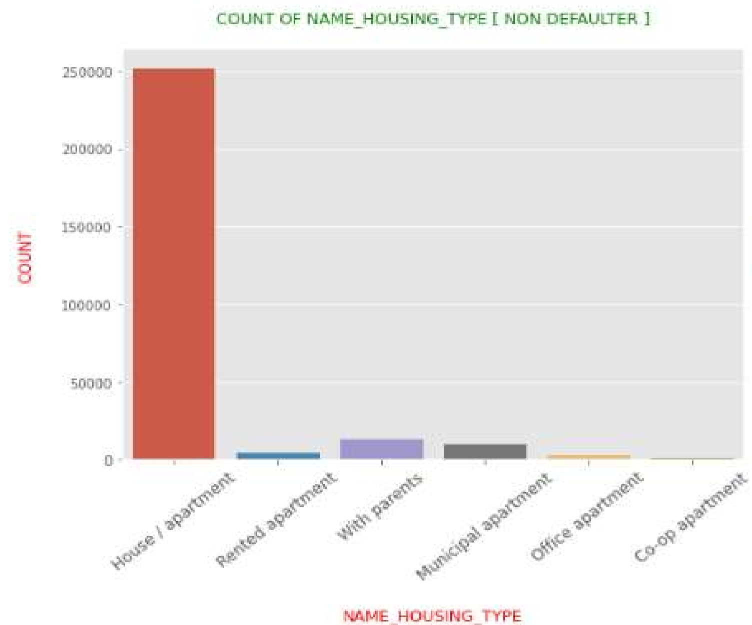
Analysis: Targeting areas on the basis of “Family Status”

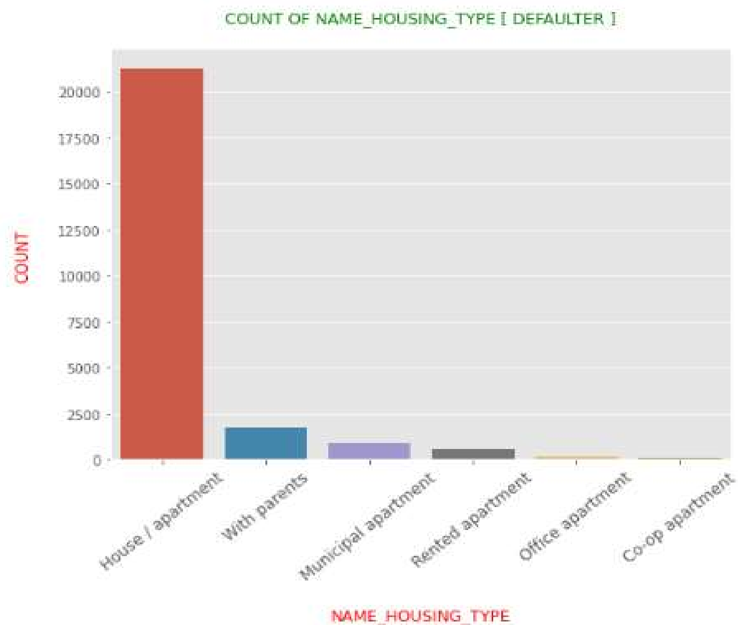




•Married people tend to apply for more loans comparatively.
•Single/non Married people contribute 14% to Non Defaulters and 18% to the defaulters, Hence there is more risk associated with them.

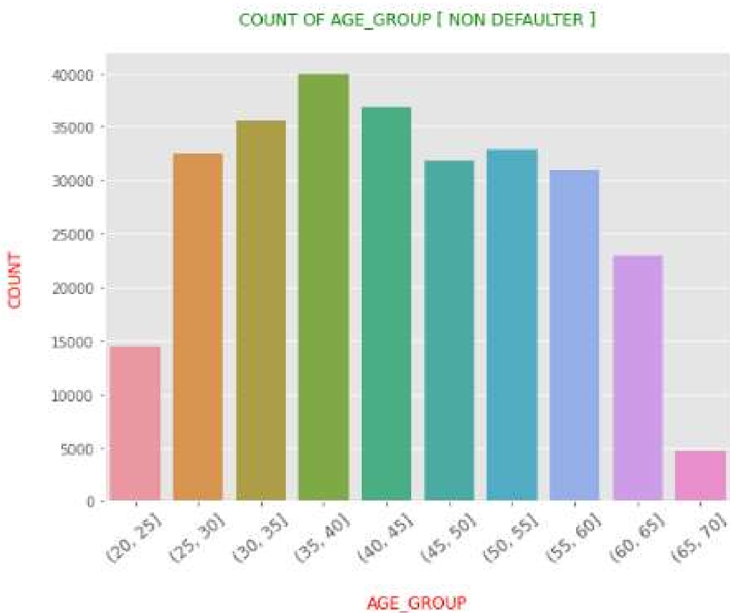
Analysis: Targeting areas on the basis of “Housing type”

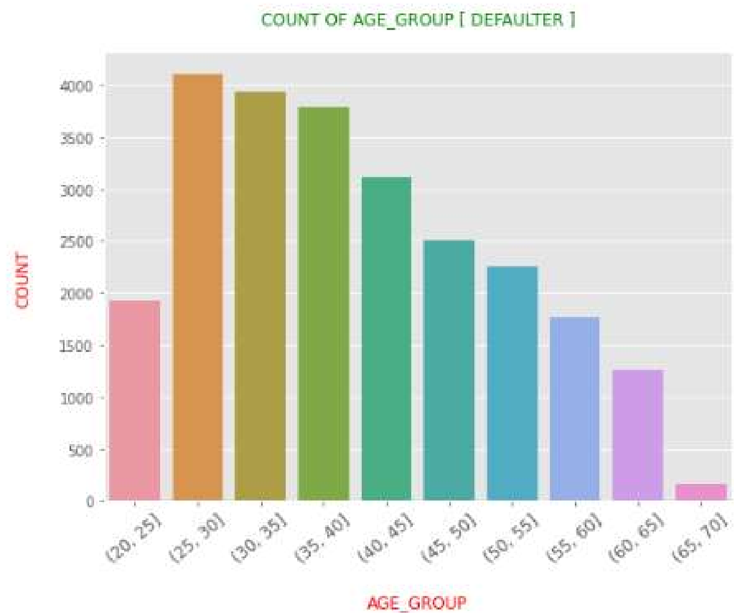




- People who have House/Apartment, tend to apply for more loans.
- People living with parents tend to default more often when compared with others.

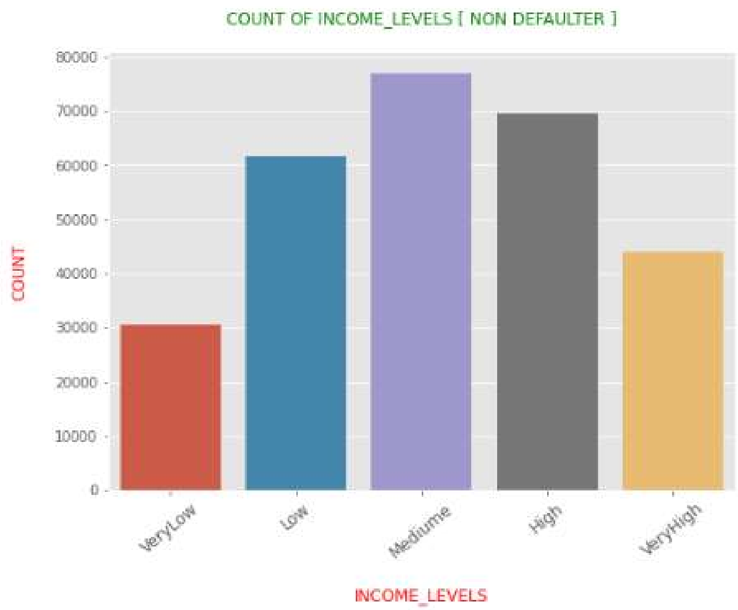
Analysis: Targeting areas on the basis of “Age Group”

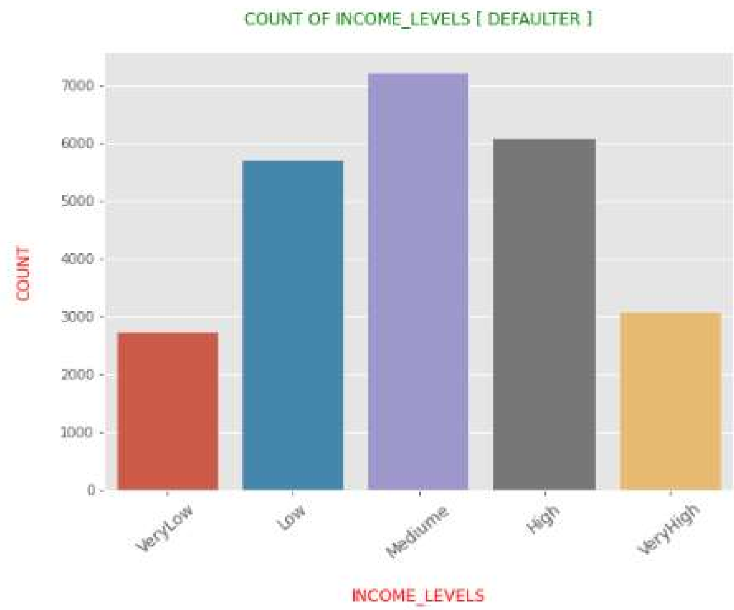




- Age group - (25-30) tend to default more often. hence they are the riskiest people to loan to.
- With increasing age group, people tend to default less starting from the age of 25.

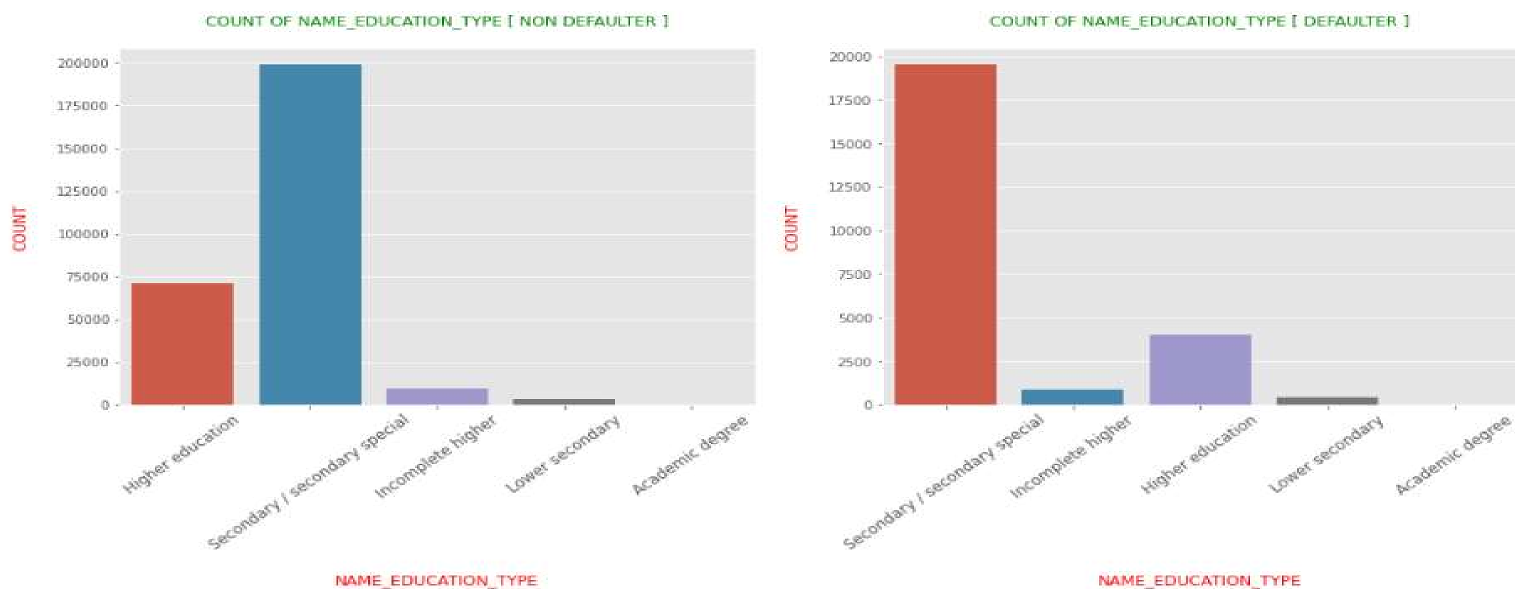
Analysis: Targeting areas on the basis of “Income Levels”





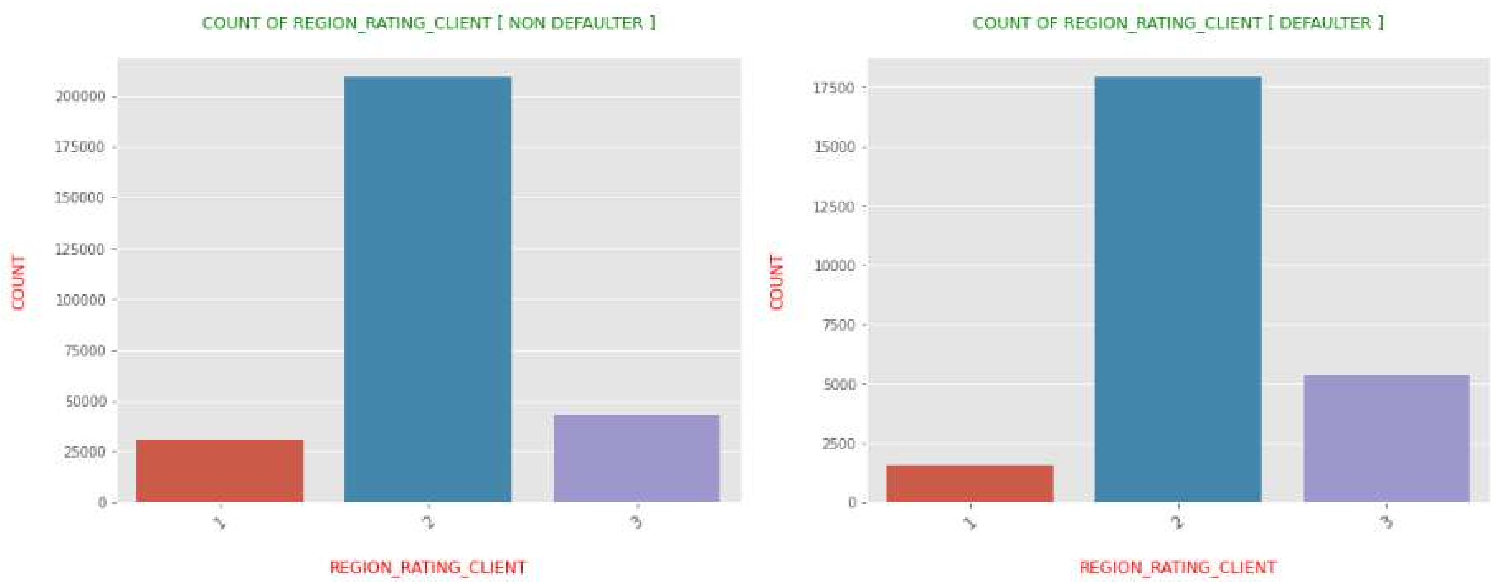
•Medium income levels tend to apply more loans
•The Very High income group tend to default less often.
•Very High income level contribute 12 % to the total number of defaulters, while they contribute 16 % to the Non-Defaulters.

Analysis: Targeting areas on the basis of "Education Type"



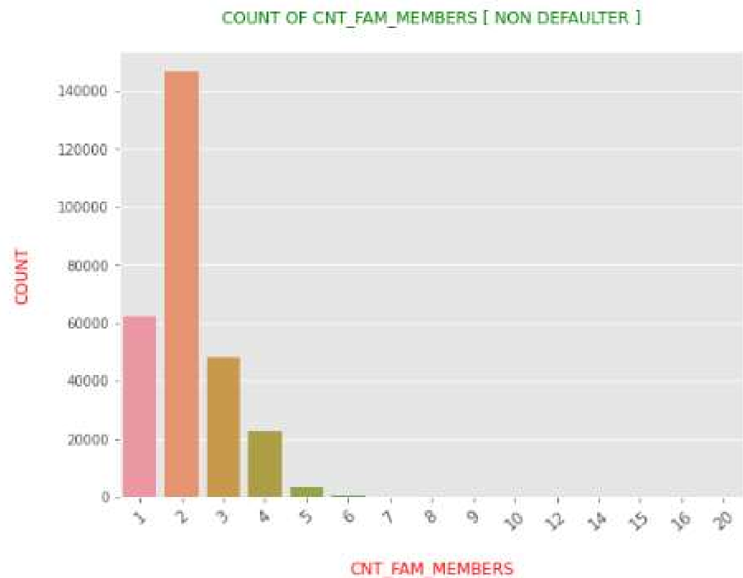
- Secondary Educated people are more likely to default
- Higher educated ones who are less likely to default
- All others all of the Education categories are equally likely to default

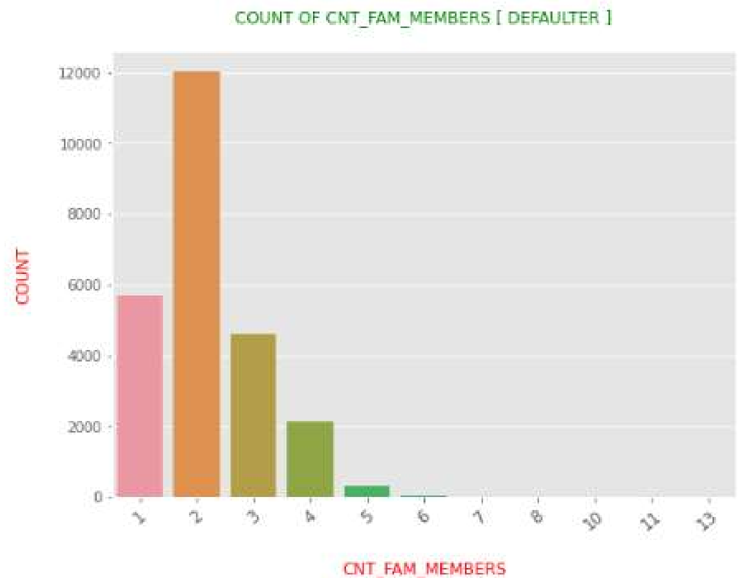
Analysis: Targeting areas on the basis of "Region Rating"



- More people from second tier regions tend to apply for loans.
- People living in 1 rated areas is less likely default

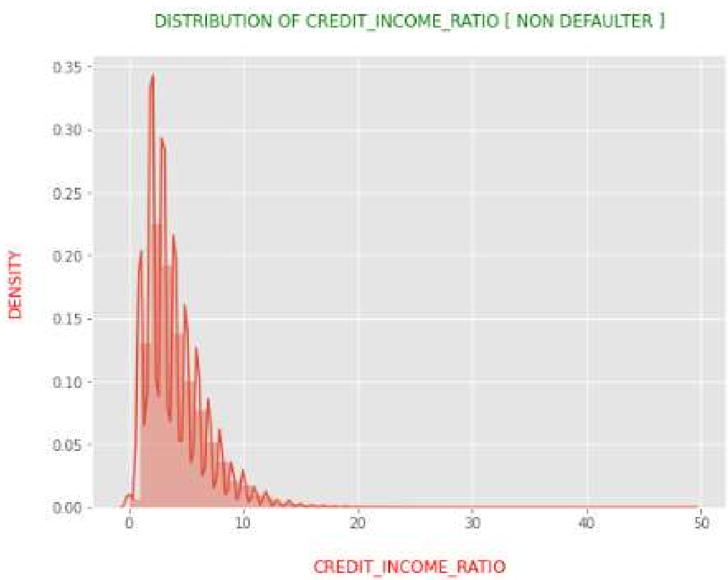
Analysis: Targeting areas on the basis of “Family members count”

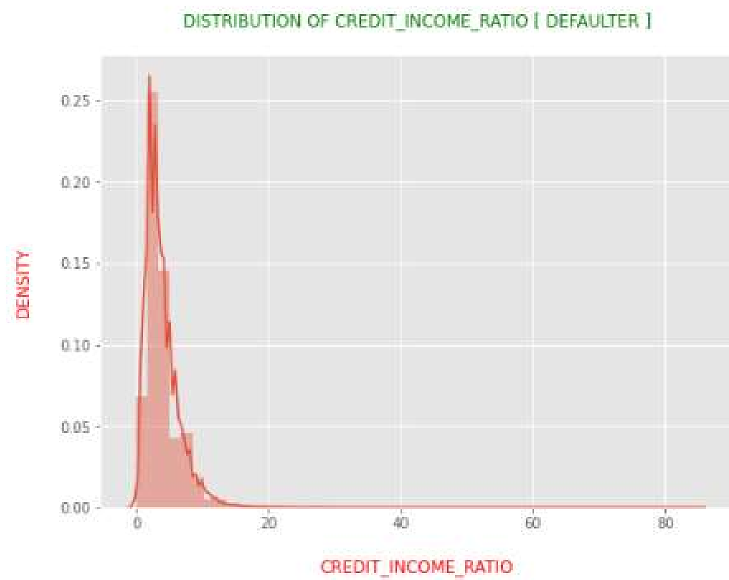




•Families of 2 apply for loans more often than the other families

Analysis: Distribution of Credit Income Ratio





- 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

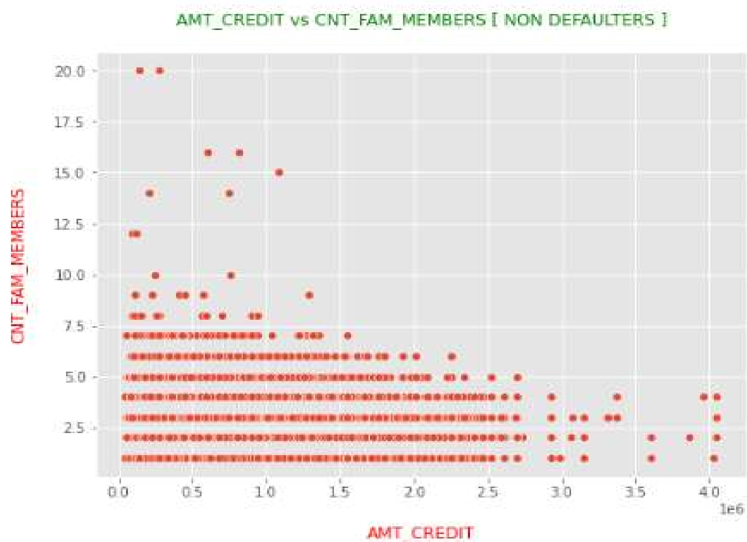
Analysis: Correlation analysis Between Variables on Non Defaulters

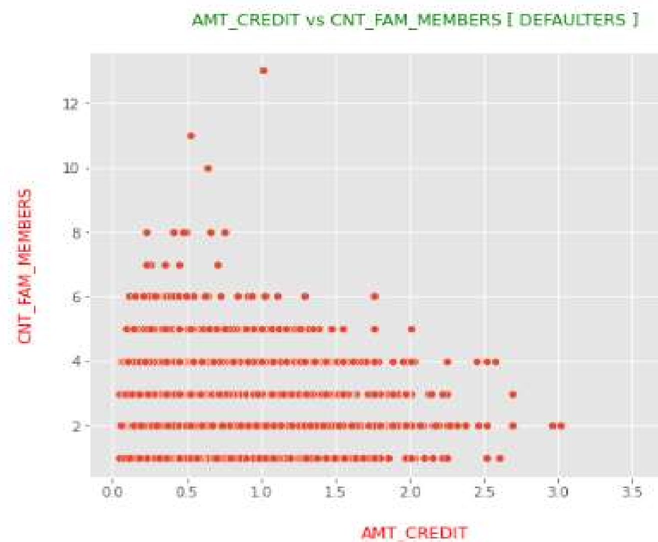
	VAR1	VAR2	CORR	VAR_SET	TYPE_CORR	ABS_CORR
53	AMT_GOODS_PRICE	AMT_CREDIT	0.987024	[AMT_CREDIT, AMT_GOODS_PRICE]	Positive	0.987024
178	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950148	[REGION_RATING_CLIENT, REGION_RATING_CLIENT_W_...	Positive	0.950148
191	SOCIAL_CIRCLE_60_DAYS_DEFLT_PERC	SOCIAL_CIRCLE_30_DAYS_DEFLT_PERC	0.873003	[SOCIAL_CIRCLE_30_DAYS_DEFLT_PERC, SOCIAL_CIRC...	Positive	0.873003
287	AMT_GOODS_PRICE	AMT_ANNUITY	0.776421	[AMT_ANNUITY, AMT_GOODS_PRICE]	Positive	0.776421
51	AMT_ANNUITY	AMT_CREDIT	0.771296	[AMT_ANNUITY, AMT_CREDIT]	Positive	0.771296
40	CREDIT_INCOME_RATIO	AMT_CREDIT	0.648589	[AMT_CREDIT, CREDIT_INCOME_RATIO]	Positive	0.648589
39	AMT_GOODS_PRICE	CREDIT_INCOME_RATIO	0.628732	[AMT_GOODS_PRICE, CREDIT_INCOME_RATIO]	Positive	0.628732
59	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418950	[AMT_ANNUITY, AMT_INCOME_TOTAL]	Positive	0.418950
37	AMT_ANNUITY	CREDIT_INCOME_RATIO	0.391498	[AMT_ANNUITY, CREDIT_INCOME_RATIO]	Positive	0.391498
71	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349425	[AMT_GOODS_PRICE, AMT_INCOME_TOTAL]	Positive	0.349425

Analysis: Correlation analysis Between Variables on Defaulters

	VAR1	VAR2	CORR	VAR_SET	TYPE_CORR	ABS_CORR
53	AMT_GOODS_PRICE	AMT_CREDIT	0.982783	[AMT_CREDIT, AMT_GOODS_PRICE]	Positive	0.982783
178	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956637	[REGION_RATING_CLIENT, REGION_RATING_CLIENT_W_...	Positive	0.956637
191	SOCIAL_CIRCLE_60_DAYS_DEFLT_PERC	SOCIAL_CIRCLE_30_DAYS_DEFLT_PERC	0.874562	[SOCIAL_CIRCLE_30_DAYS_DEFLT_PERC, SOCIAL_CIRC...	Positive	0.874562
287	AMT_GOODS_PRICE	AMT_ANNUITY	0.752296	[AMT_ANNUITY, AMT_GOODS_PRICE]	Positive	0.752296
51	AMT_ANNUITY	AMT_CREDIT	0.752195	[AMT_ANNUITY, AMT_CREDIT]	Positive	0.752195
40	CREDIT_INCOME_RATIO	AMT_CREDIT	0.639744	[AMT_CREDIT, CREDIT_INCOME_RATIO]	Positive	0.639744
89	AMT_GOODS_PRICE	CREDIT_INCOME_RATIO	0.623100	[AMT_GOODS_PRICE, CREDIT_INCOME_RATIO]	Positive	0.623100
87	AMT_ANNUITY	CREDIT_INCOME_RATIO	0.381298	[AMT_ANNUITY, CREDIT_INCOME_RATIO]	Positive	0.381298
96	DAYS_REGISTRATION	DAYS_EMPLOYED	-0.188929	[DAYS_EMPLOYED, DAYS_REGISTRATION]	Negative	0.188929
98	CNT_FAM_MEMBERS	DAYS_EMPLOYED	-0.186561	[CNT_FAM_MEMBERS, DAYS_EMPLOYED]	Negative	0.186561

Analysis: Bivariate Analysis of numerical variables

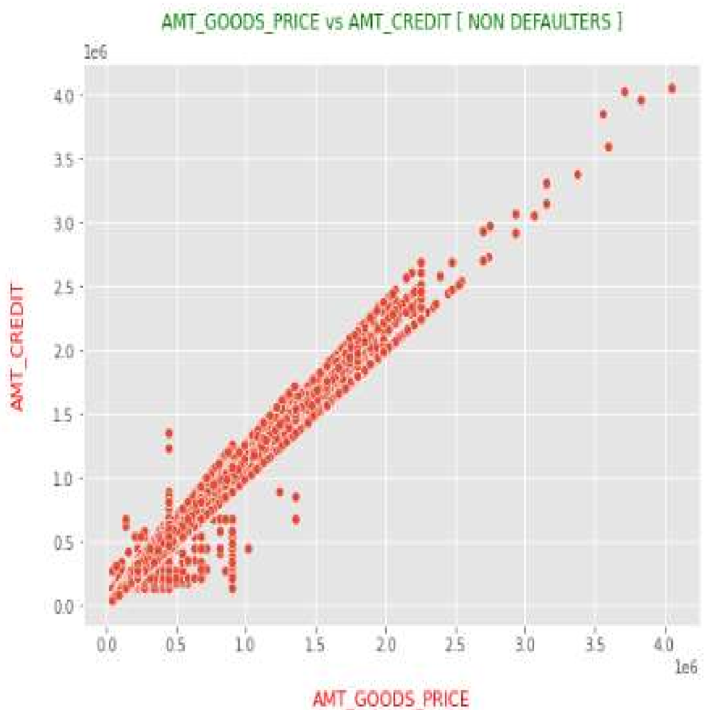


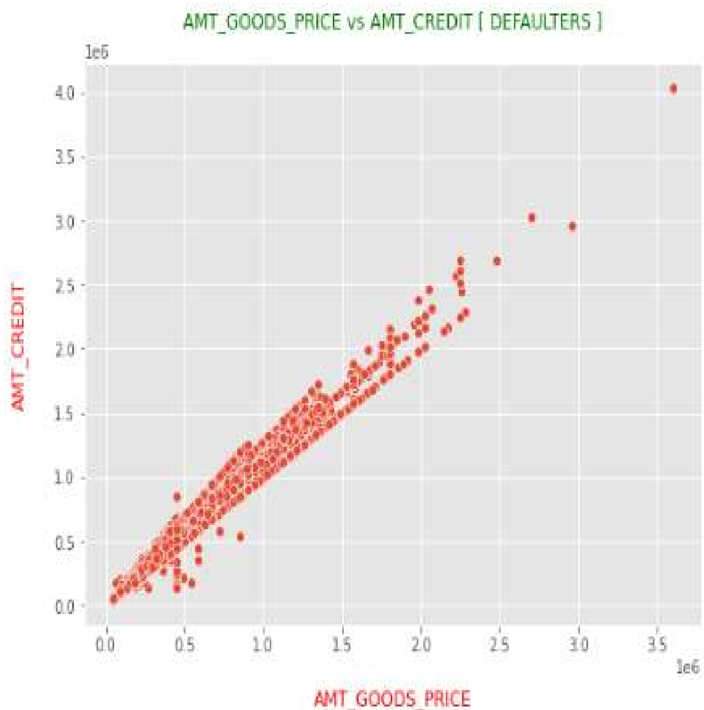


•The density in the lower left corner is similar in both the case, Hence the people are equally likely to default if the family is small and the AMT_CREDIT is low.

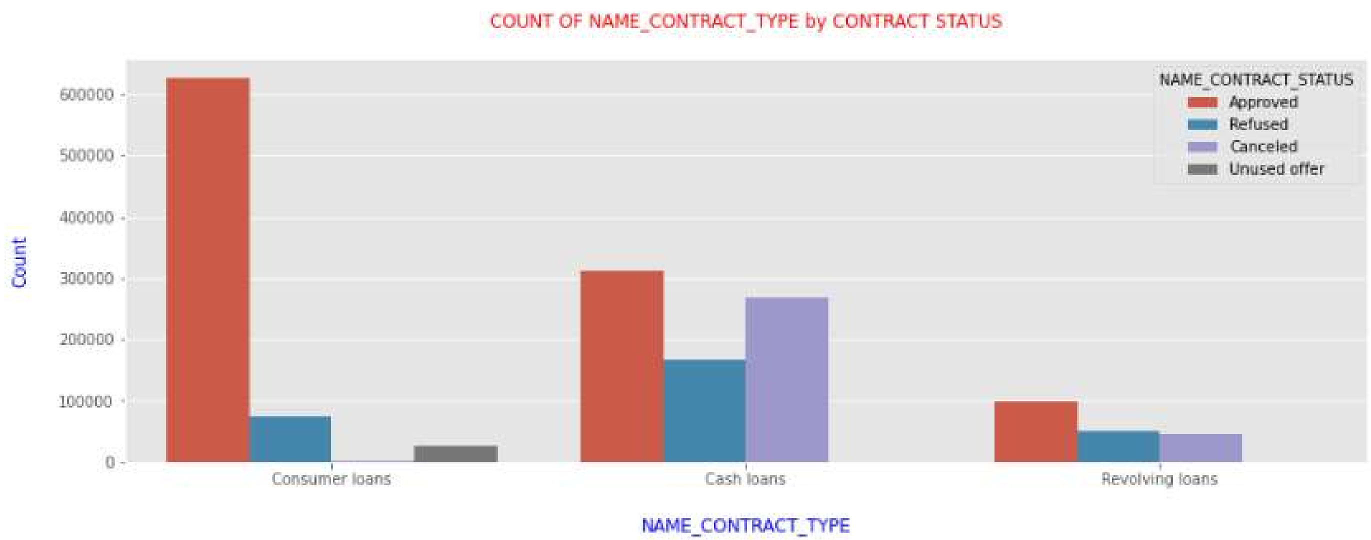
•We can observe that larger families and people with larger AMT_CREDIT default less often

Analysis: Bivariate Analysis of numerical variables





Analysis: Count of Name Contract Type by Contract Status



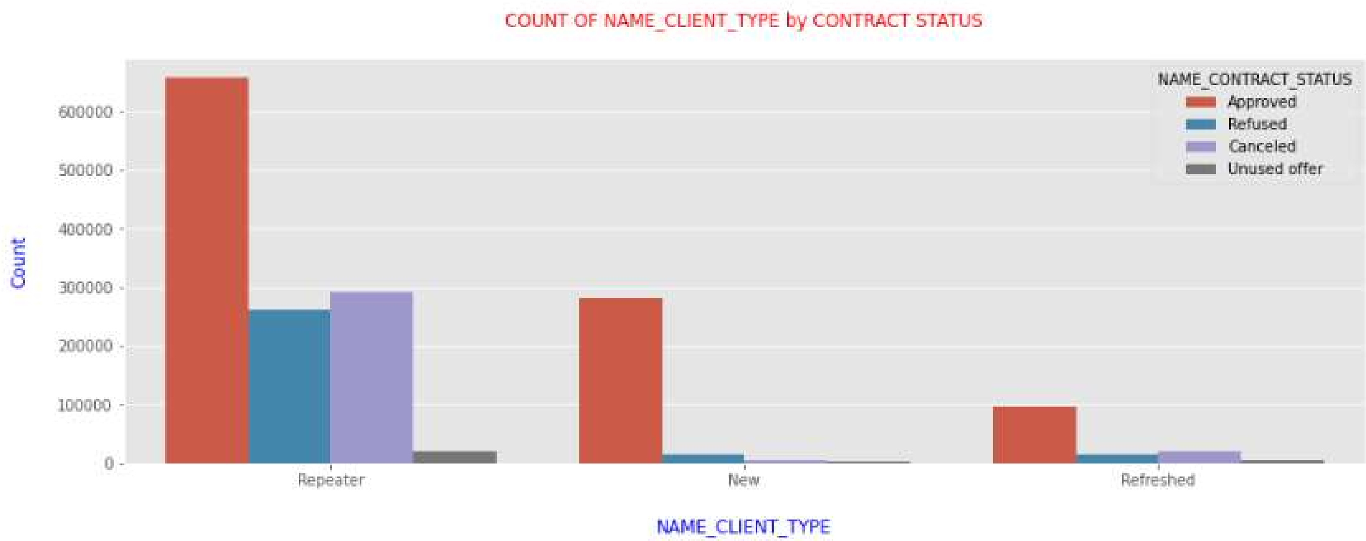
- Most of the applications are for 'Cash loan' and 'Consumer loan'.
- Although the cash loans are refused more often than others.

Analysis: Count of Name Payment Type by Contract Status

•Most of the clients chose to repay the loan using the 'Cash through the bank' option

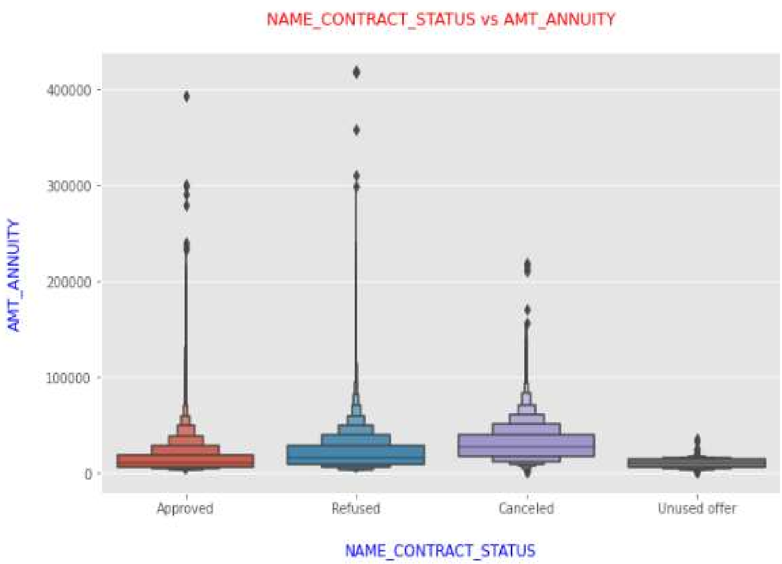
•Non-Cash from your account' & 'Cashless from the account of the employee' options are not at all popular in terms of loan repayment amongst the customers

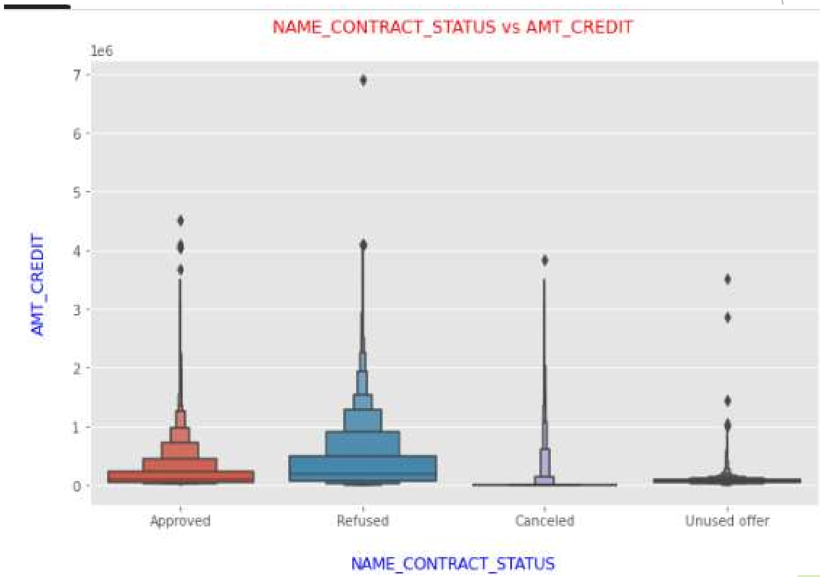
Analysis: Count of Name Client Type by Contract Status



•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.

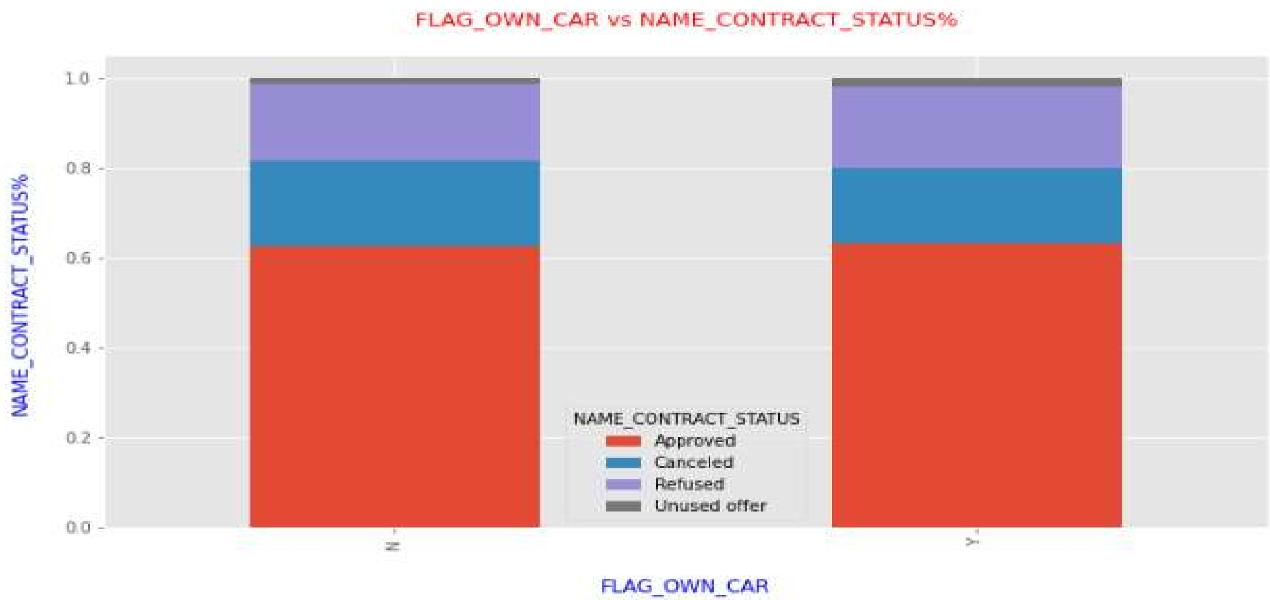
Analysis: Contract Status vs Annuity





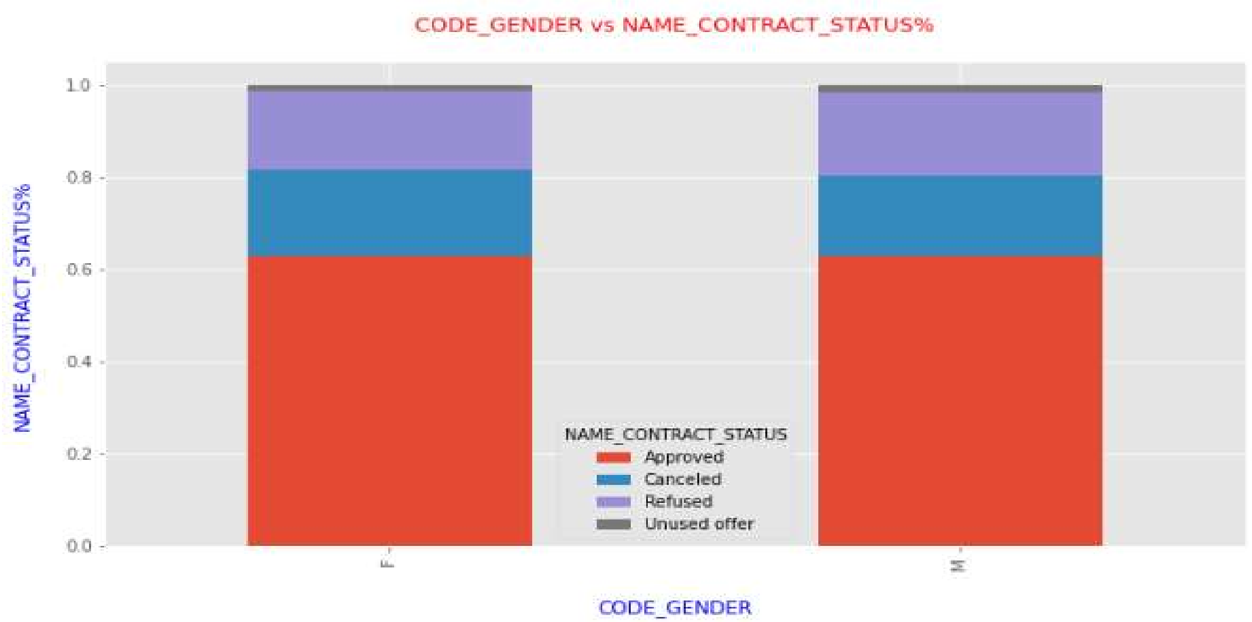
•When the AMT_CREDIT is too low, it get's cancelled/unused most of the time.

Analysis: Flag Own vs Name Contract Status%



- Code gender doesn't have any effect on application approval or rejection.
- But we saw earlier that female have lesser chances of default compared to males.
- The bank can add more weightage to female while approving a loan amount.

Analysis: Code Gender vs Name Contract Status%



- Code gender doesn't have any effect on application approval or rejection.
- But we saw earlier that female have lesser chances of default compared to males.
- The bank can add more weightage to female while approving a loan amount.

Analysis: Target Name vs Contract Status

•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.