CREDIT EDA CASE STUDY

PROBLEM STATEMENT

- **When the company receives a loan application, the company must 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.

Using EDA to understand how consumer attributes and loan attributes influence the tendency of default.

ANALYSIS

- Analysis of the datasets has been done in the Python (**Jupyter notebook**)
  - Attached the notebook for the step-by-step procedure that we have done as part of this case study.
  - What have we done ??
    - Starting by importing the application_data.csv
    - Check the structure of the data (Its normal routine check) (Info, Describe, Shape, etc..).
    - Data quality check and missing values
      - Finding percentage of missing values of all the columns.
      - Remove columns with high missing values.
      - Explained best metric to impute the missing values.
      - Checked the datatypes of the column and converted to correct data types.
      - Checked outliers for the numerical columns,
      - Binning of continuous variables

ANALYSIS CONT…

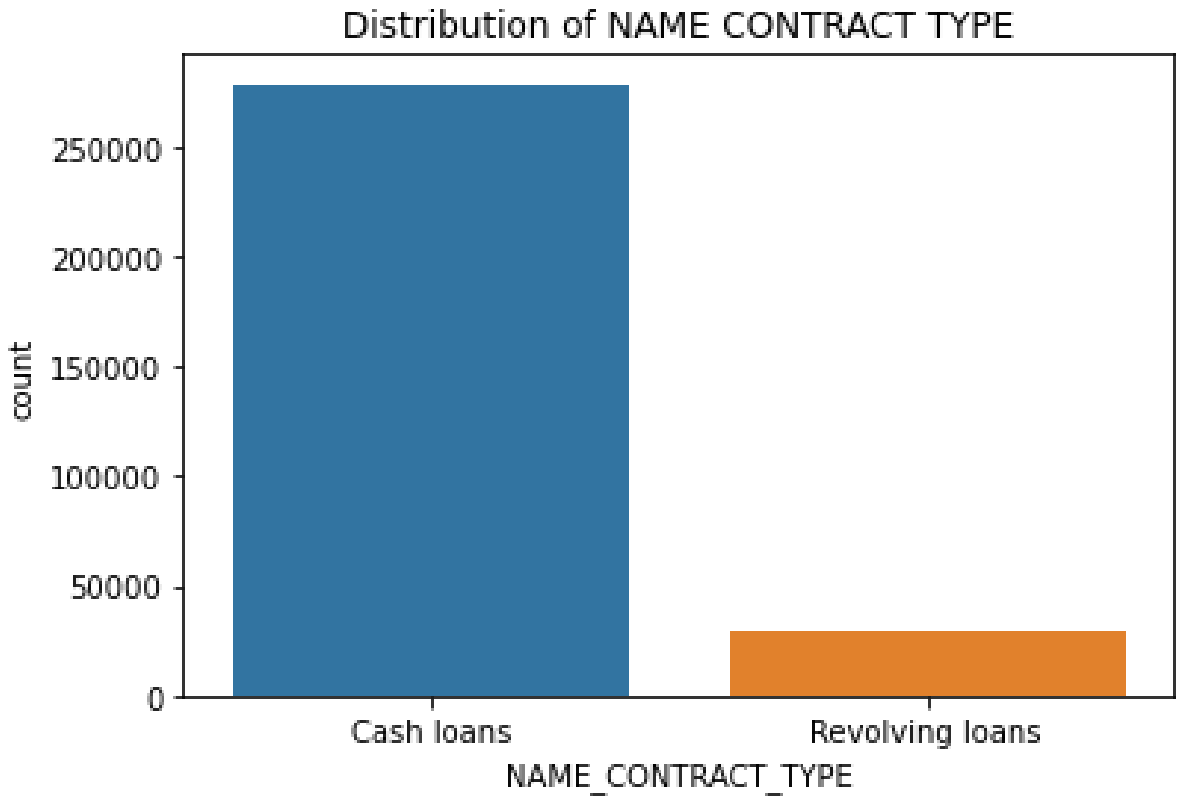
- Complete Analysis
  - Check the imbalance percentage.
  - Dividing the dataset into 2 data frames (defaulters and non-defaulters).
  - Performed correlations. (for both defaulters and non-defaulters)
    - Verified the highest correlation in both the data frames and added insights
  - Performed Univariate Analysis (for both defaulters and non-defaulters
    - Numerical Columns.
    - Categorical Columns.
  - Performed Bivariate Analysis (for both defaulters and non-defaulters)
    - Numeric – Numeric (Continuous – Continuous)
    - Categorical – Numerical
    - Categorical - Categorical
- There by started analysis on previous application data
  - Analyzed it individually
  - Merged with application data set and perform all the analysis.

USED LIBRARIES

DISTRIBUTION OF COLUMNS

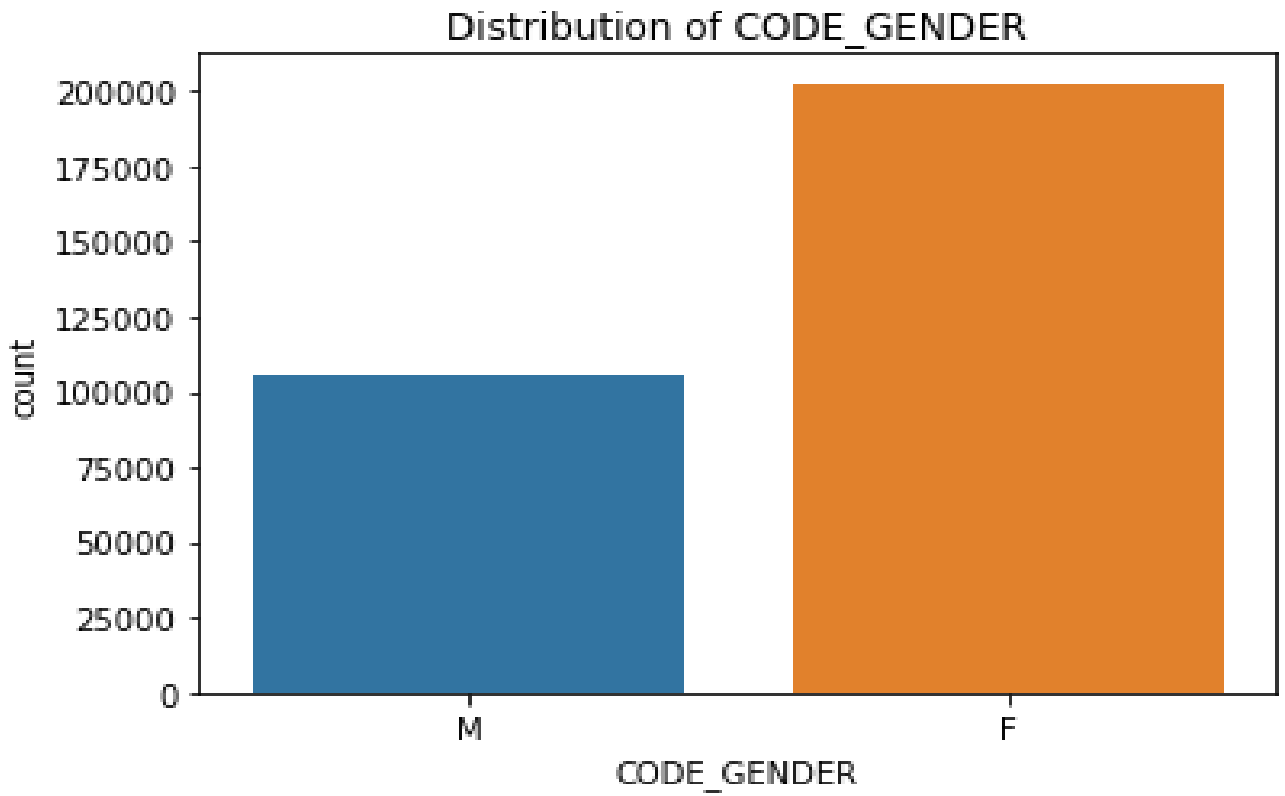
APPLICATION DATA

NAME_CONTRACT_TYPE



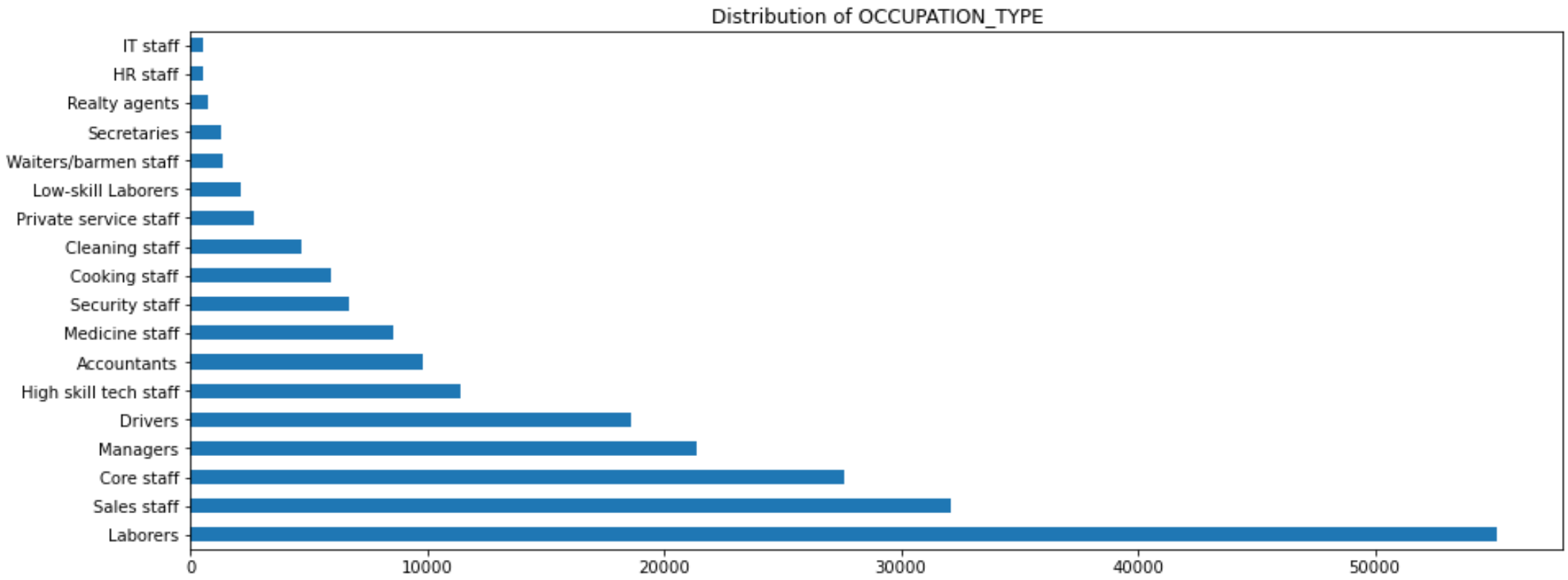
Clearly cash loans are more in number.

CODE_GENDER



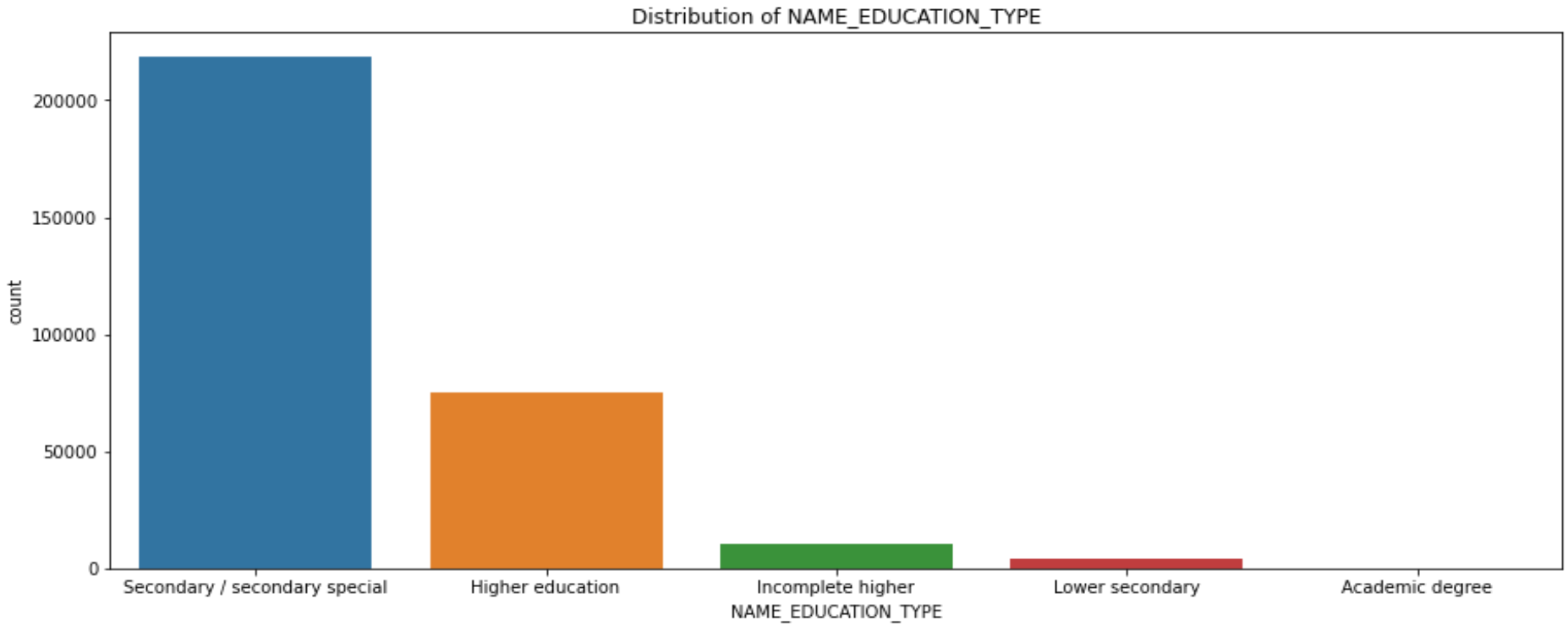
More number of applicants are Females. (As per the application data)

OCCUPATION_TYPE



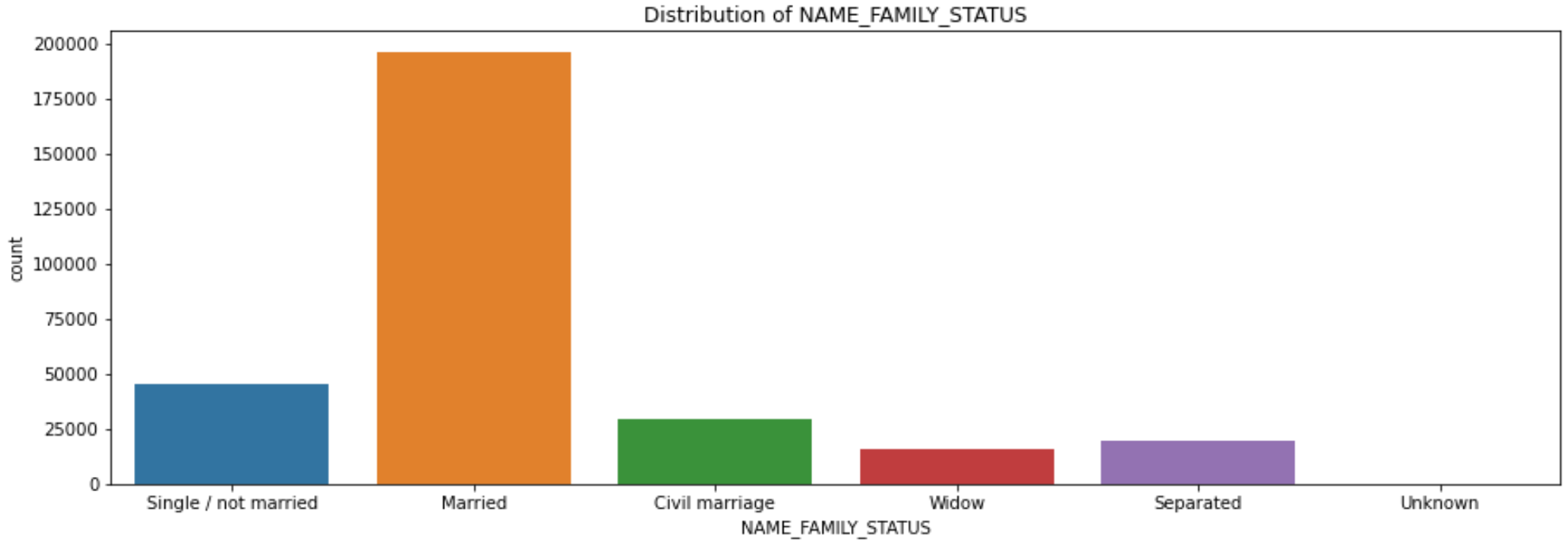
We can see that laborers, sales staff constitute the majority.

NAME_EDUCATION_TYPE



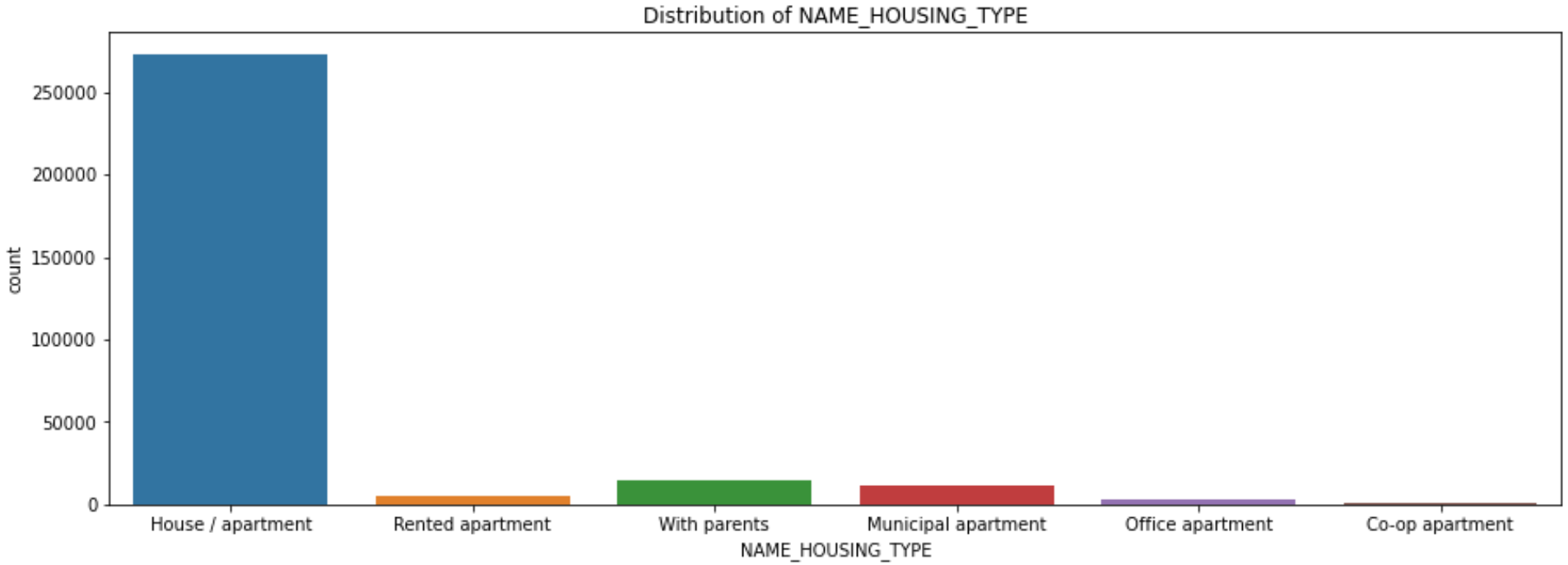
By this, we can say that a greater number of applicants are from the secondary/secondary special education type.

NAME_FAMILY_STATUS



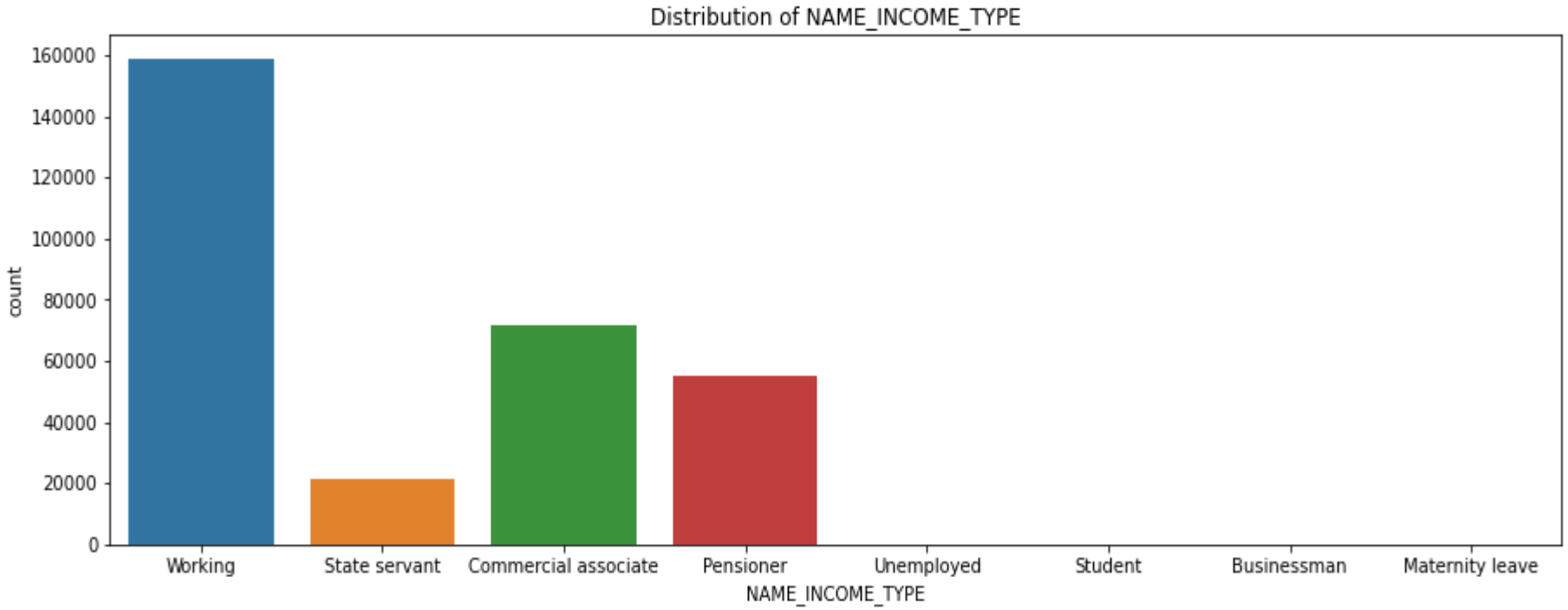
By this, we can say that a greater number of applicants are of married

NAME_HOUSING_TYPE



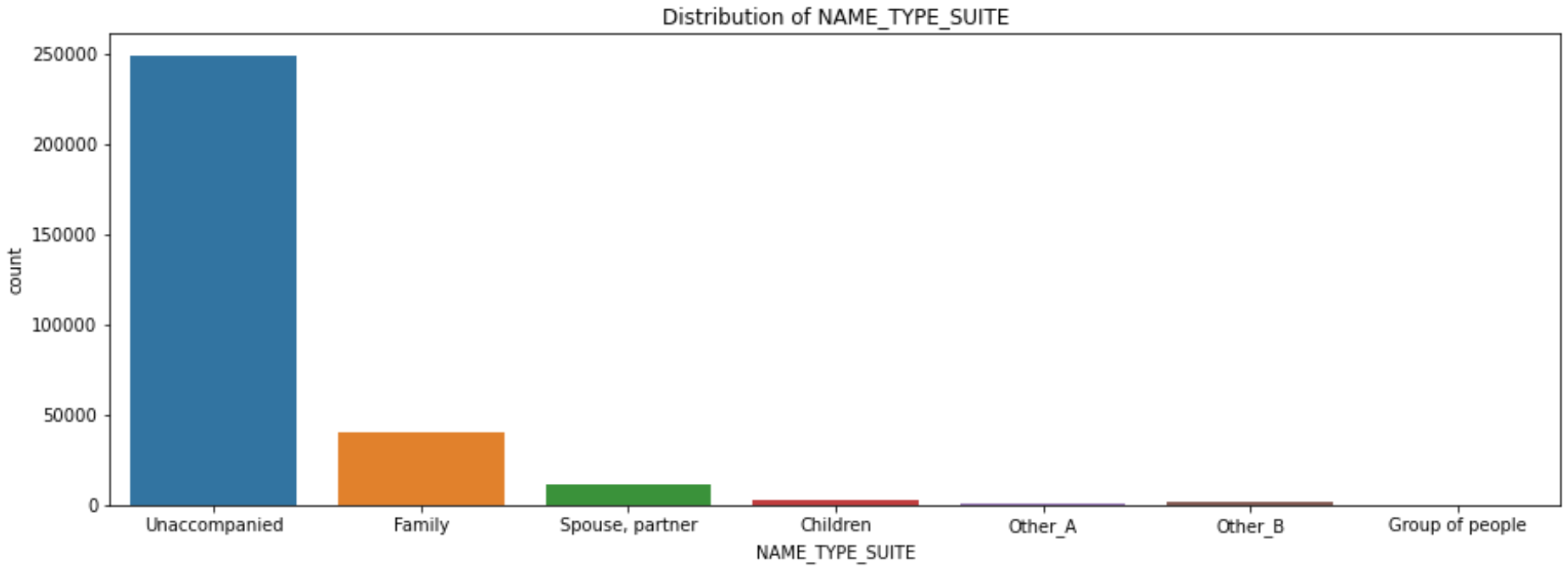
By this, we can say that a greater number of applicants housing situation is House/apartment

NAME_INCOME_TYPE



By this, we can say that a greater number of applicants income type is by `working`

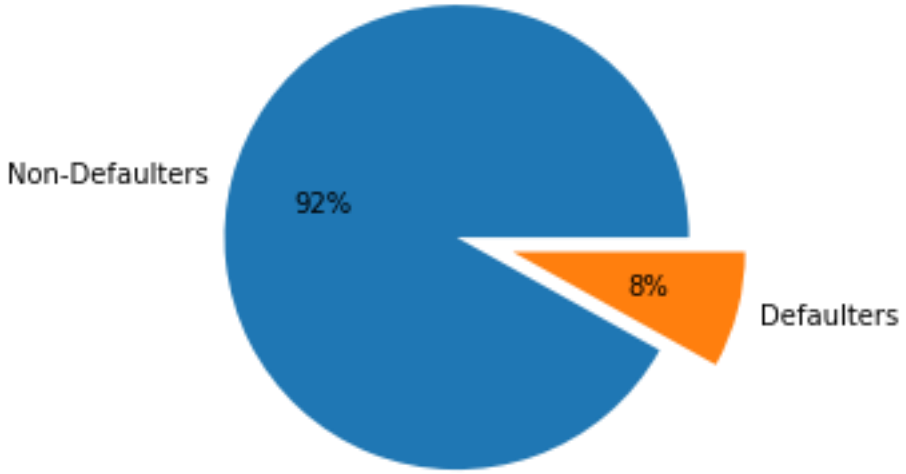
NAME_TYPE_SUITE



By this, we can say that a greater number of applicants are `Unaccompanied`

IMBALANCE RATIO

Percentage of Defaulters vs Non-Defaulters



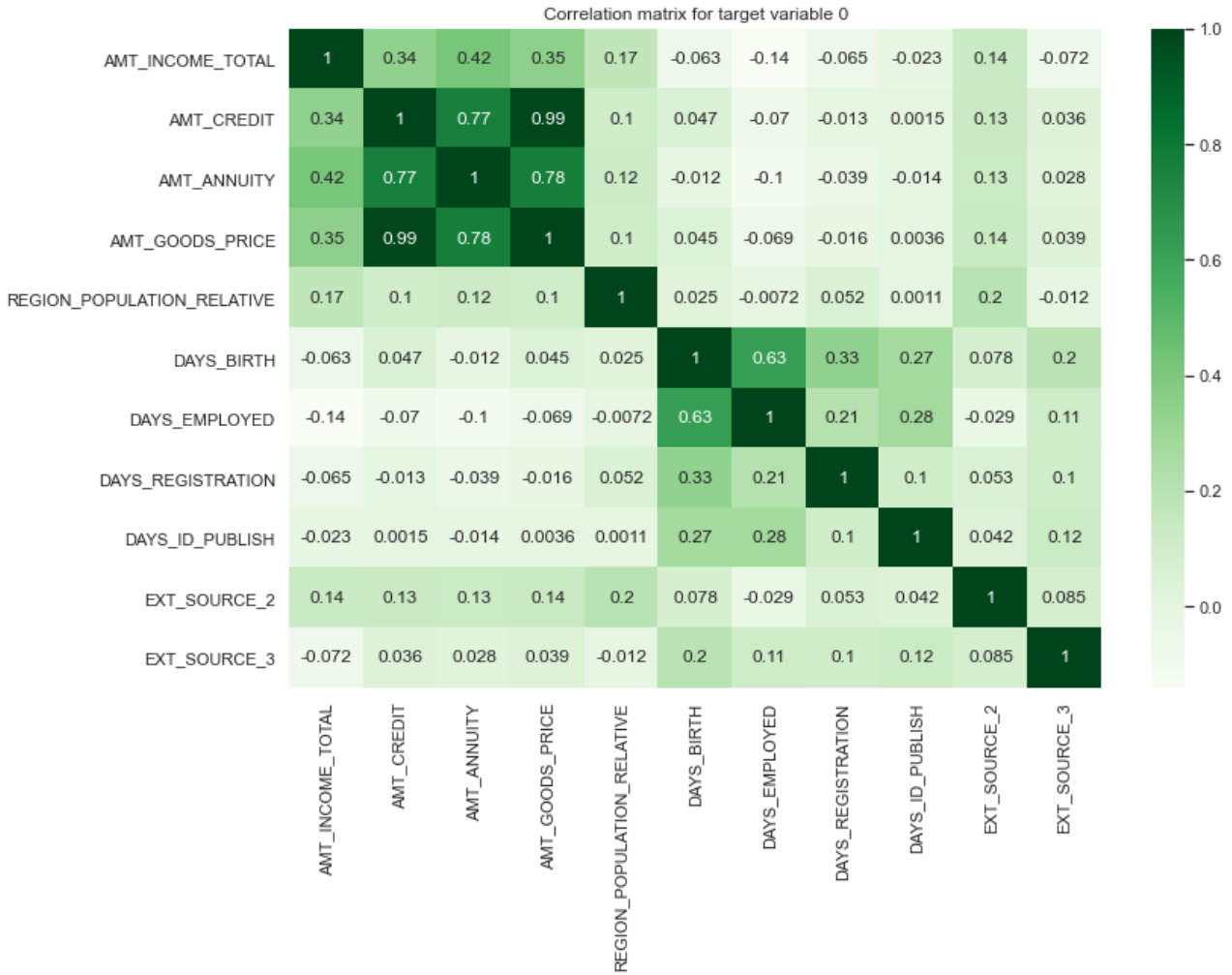
Imbalance Ratio:
92: 8

Clearly, We can see that there is a lot of imbalance in the dataset.

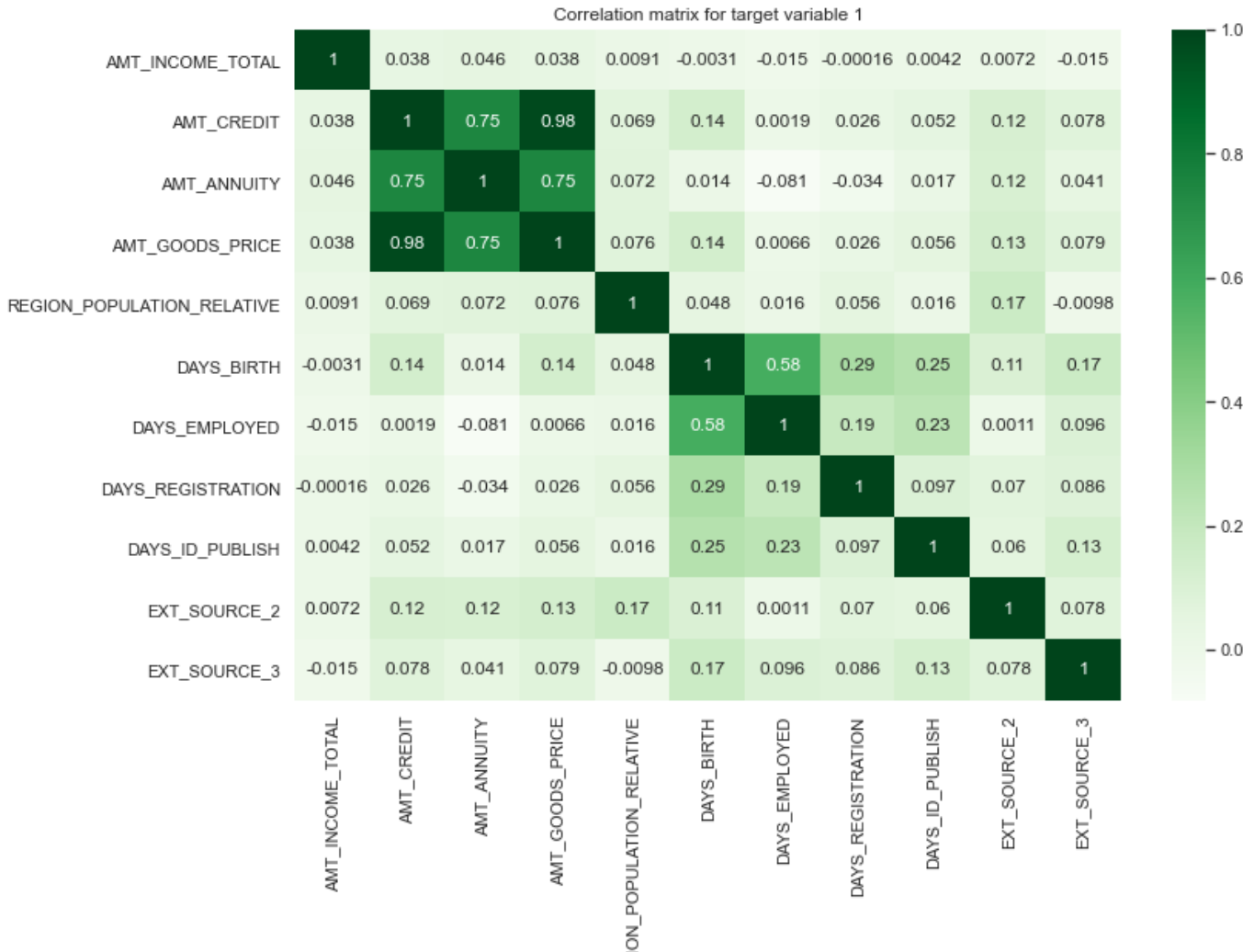
CORRELATION

FOR BOTH DEFAULTERS AND NON-DEFAULTERS

CORRELATION MATRIX FOR TARGET VARIABLE 0 (NON-DEFAULTERS)



CORRELATION FOR TARGET VARIABLE 1 (DEFAULTERS)



TOP 10 CORRELATIONS FOR DEFAULTERS AND NON DEAULTERS

NON-DEFAULTERS

	VAR1	VAR2	Correlation_Value	Corr_abs
34	AMT_GOODS_PRICE	AMT_CREDIT	0.987250	0.987250
35	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686	0.776686
23	AMT_ANNUITY	AMT_CREDIT	0.771309	0.771309
71	DAYS_EMPLOYED	DAYS_BIRTH	0.626114	0.626114
22	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953	0.418953
33	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349462	0.349462
11	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799	0.342799
82	DAYS_REGISTRATION	DAYS_BIRTH	0.333151	0.333151
94	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.276663	0.276663
93	DAYS_ID_PUBLISH	DAYS_BIRTH	0.271314	0.271314

DEFAULTERS

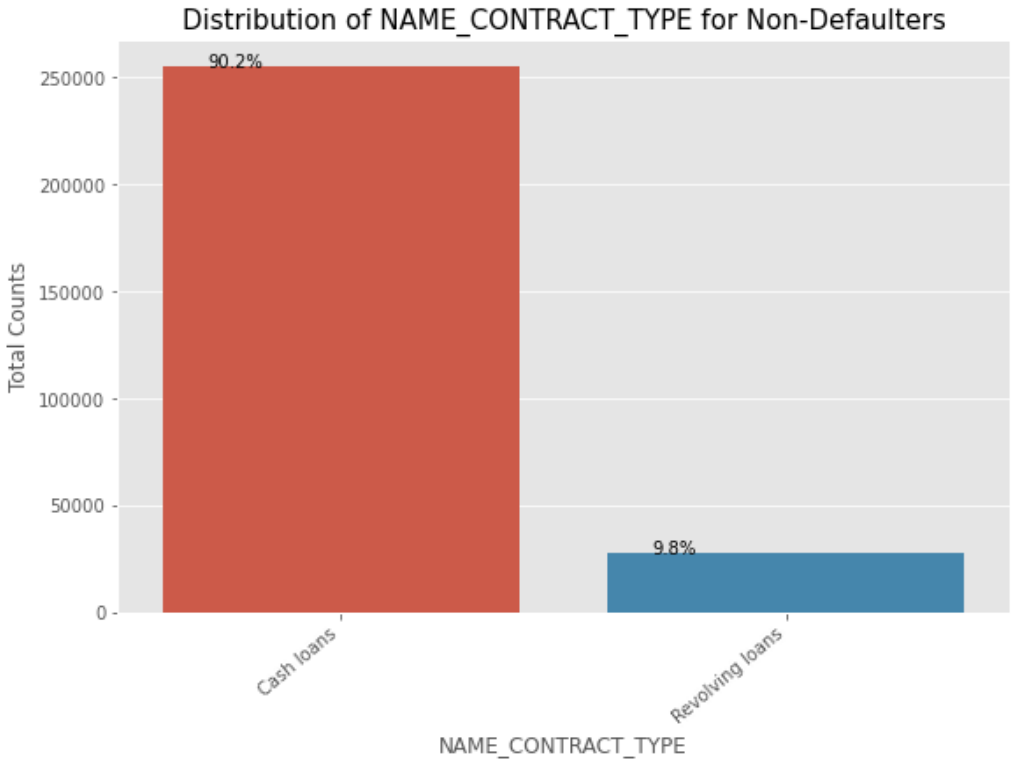
	VAR1	VAR2	Correlation_Value	Corr_abs
34	AMT_GOODS_PRICE	AMT_CREDIT	0.983103	0.983103
35	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699	0.752699
23	AMT_ANNUITY	AMT_CREDIT	0.752195	0.752195
71	DAYS_EMPLOYED	DAYS_BIRTH	0.582185	0.582185
82	DAYS_REGISTRATION	DAYS_BIRTH	0.289114	0.289114
93	DAYS_ID_PUBLISH	DAYS_BIRTH	0.252863	0.252863
94	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.229090	0.229090
83	DAYS_REGISTRATION	DAYS_EMPLOYED	0.192455	0.192455
115	EXT_SOURCE_3	DAYS_BIRTH	0.171621	0.171621
103	EXT_SOURCE_2	REGION_POPULATION_RELATIVE	0.169751	0.169751

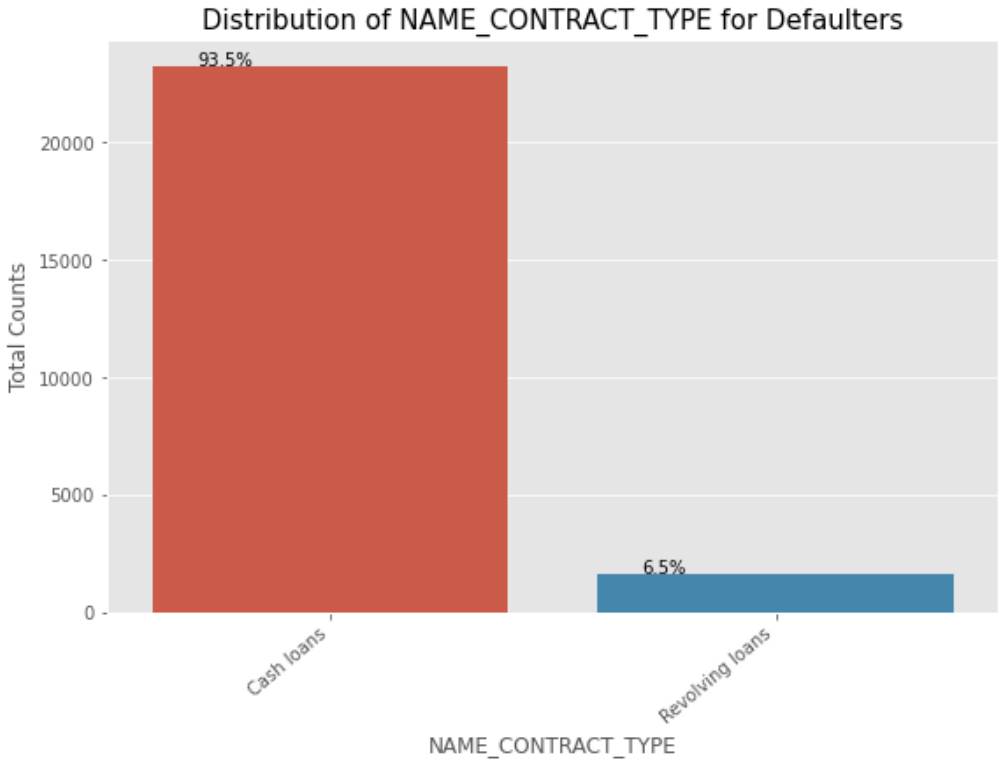
We can see from the top 10 correlations from both the data frames, the top 4 for both are similar.

UNIVARIATE ANALYSIS

CATEGORICAL COLUMNS

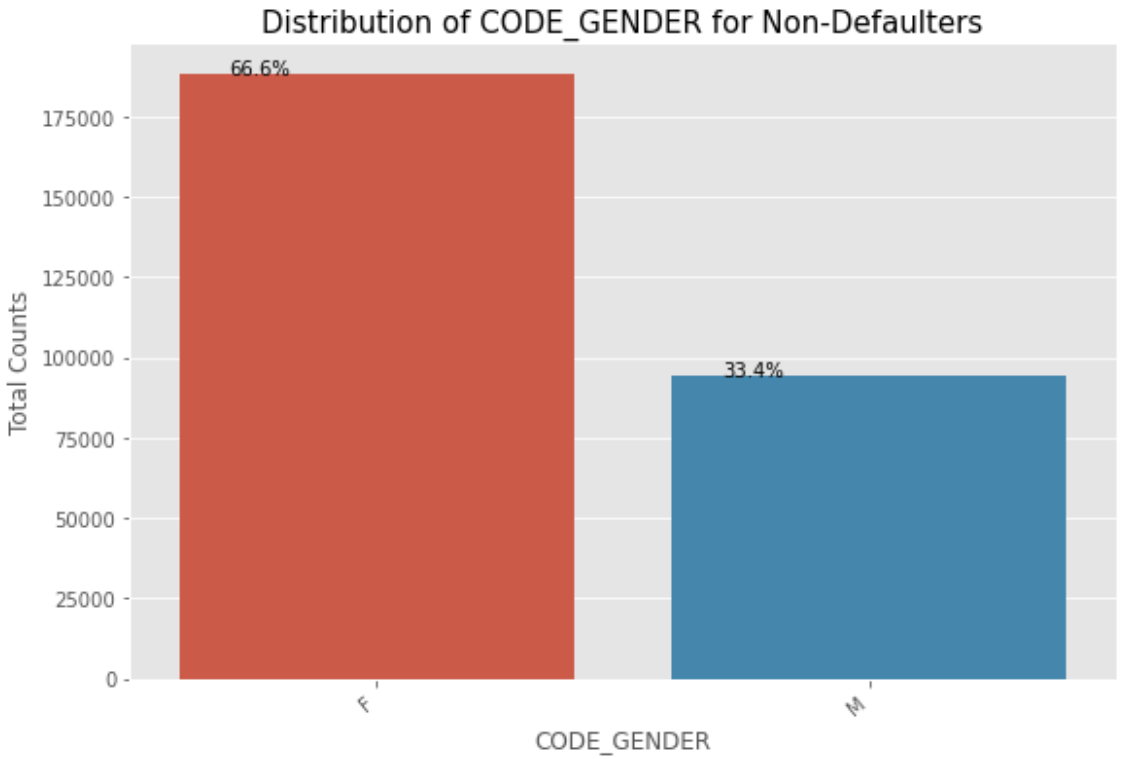
NAME_CONTRACT_TYPE

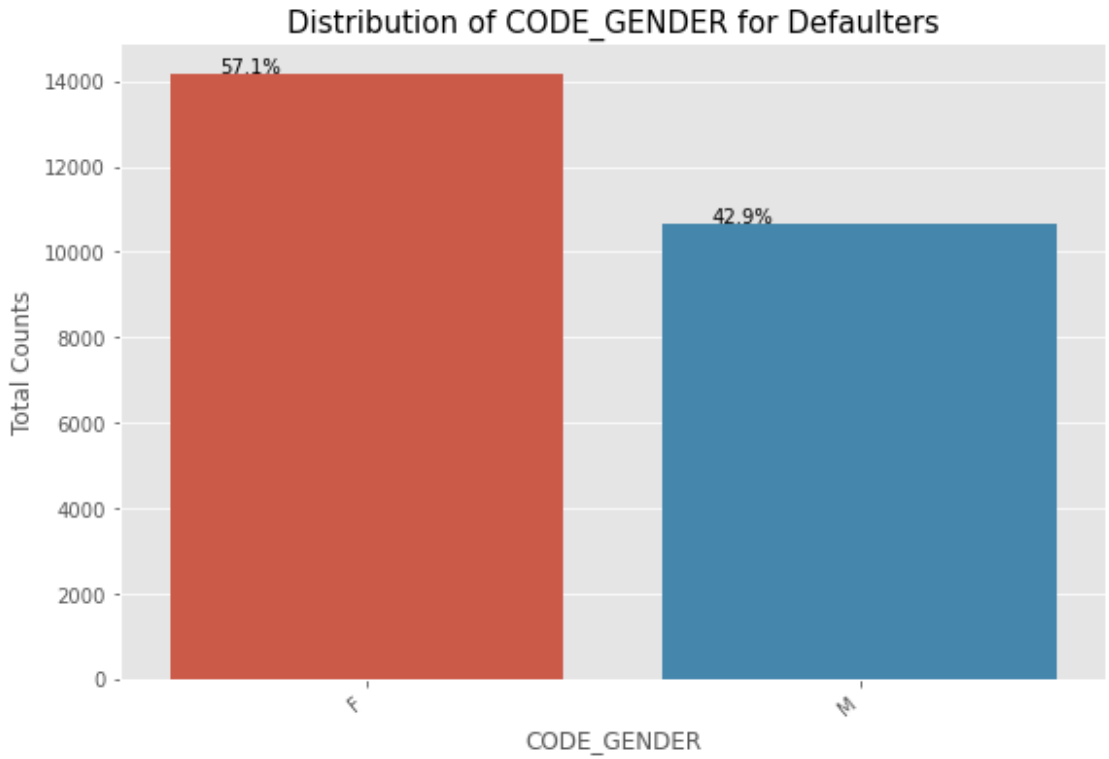




- For cash loans, the percentage of defaulters are more than the non-defaulters.
- We can see that the defaulters and non-defaulters both are taking maximum of cash loans.

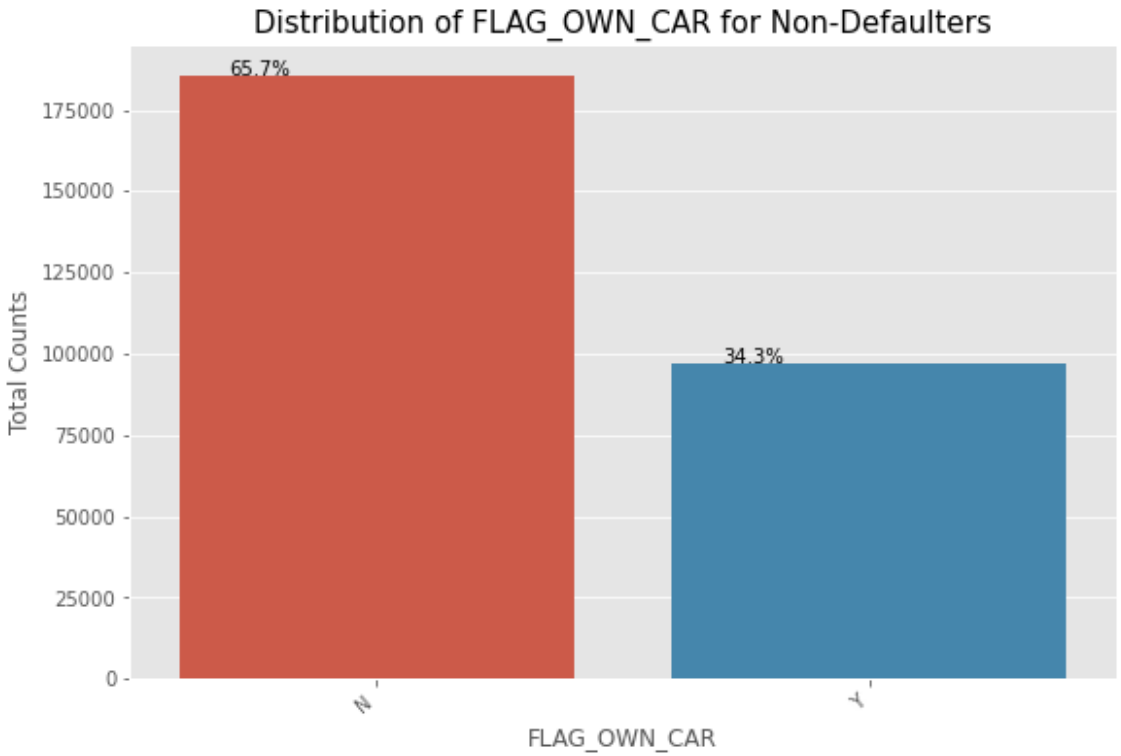
CODE_GENDER

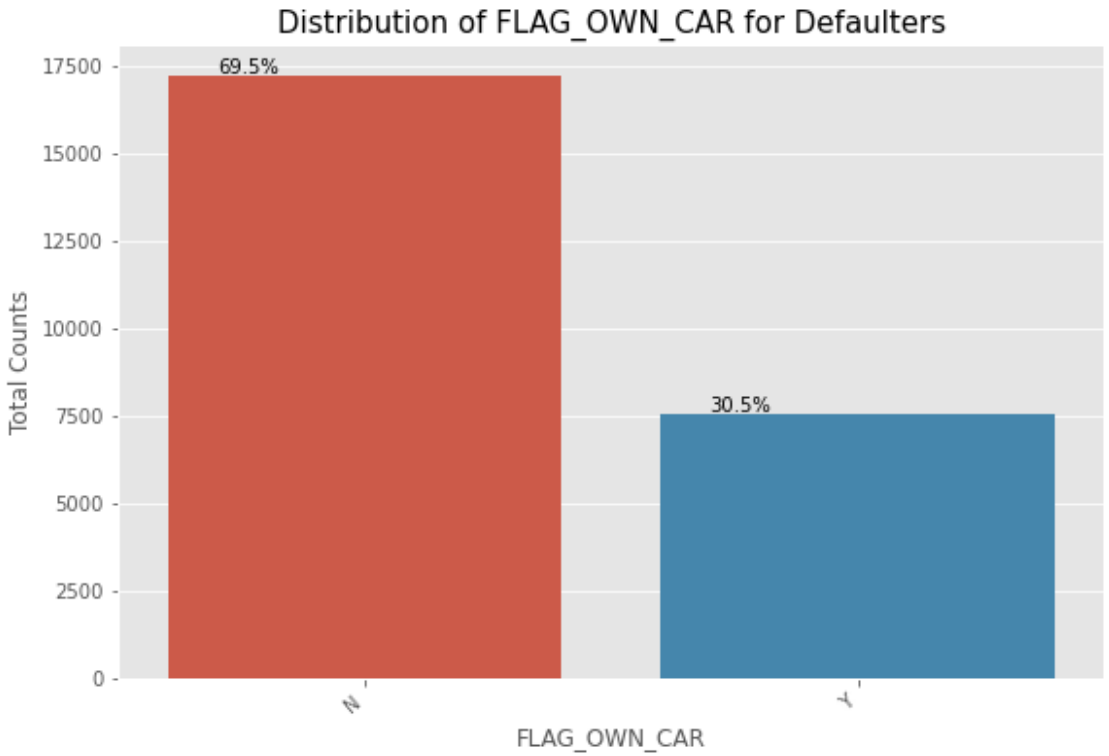




- We can see that Female contribute 67% to the non-defaulters while 57% to the defaulters.
- We see more female applying for loans than males.
- But the percentage of defaulting of females is much lower compared to males.

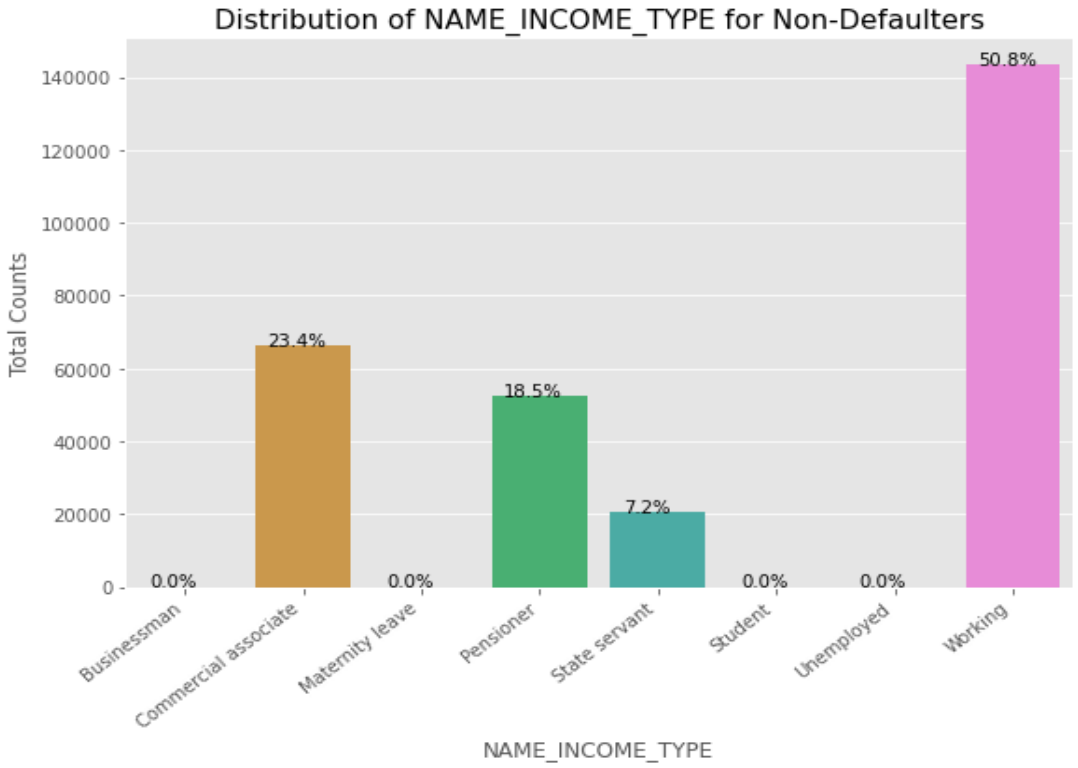
FLAG_OWN_CAR

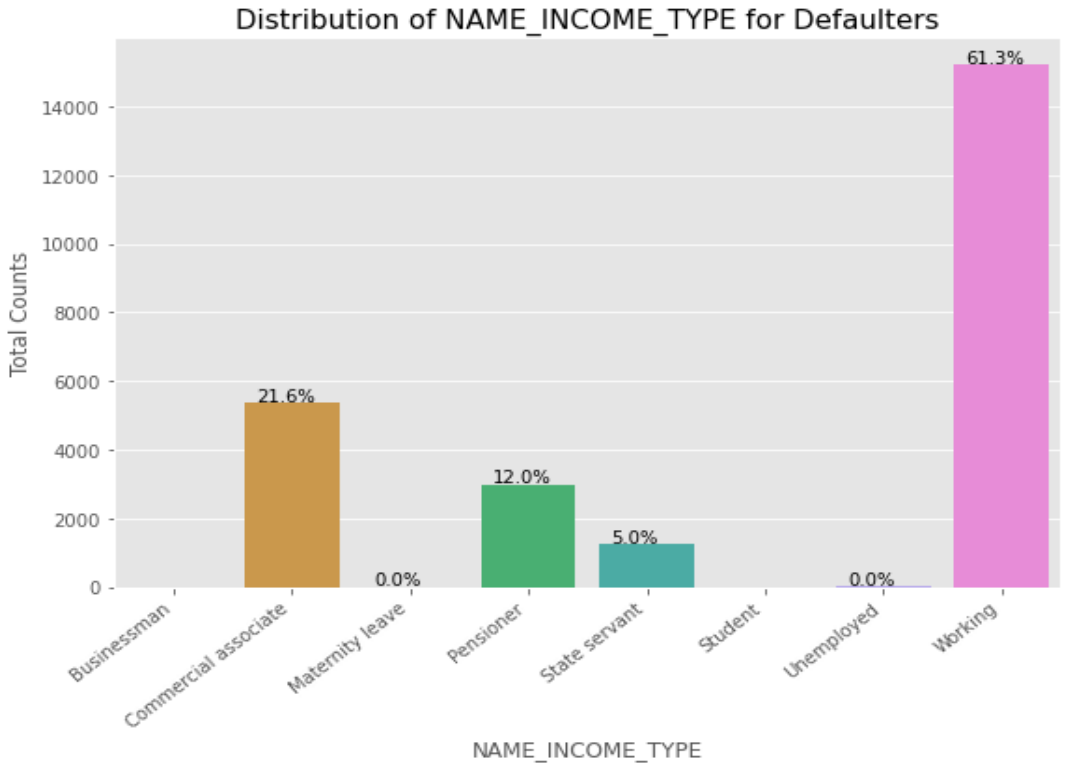




- We can see that people with cars contribute 65.7% to the non-defaulters while 69.5% to the defaulters
- Most of the people don't have car.
- We can conclude that the percentage of defaulters having car is low compared to people who don't. (In both the cases)

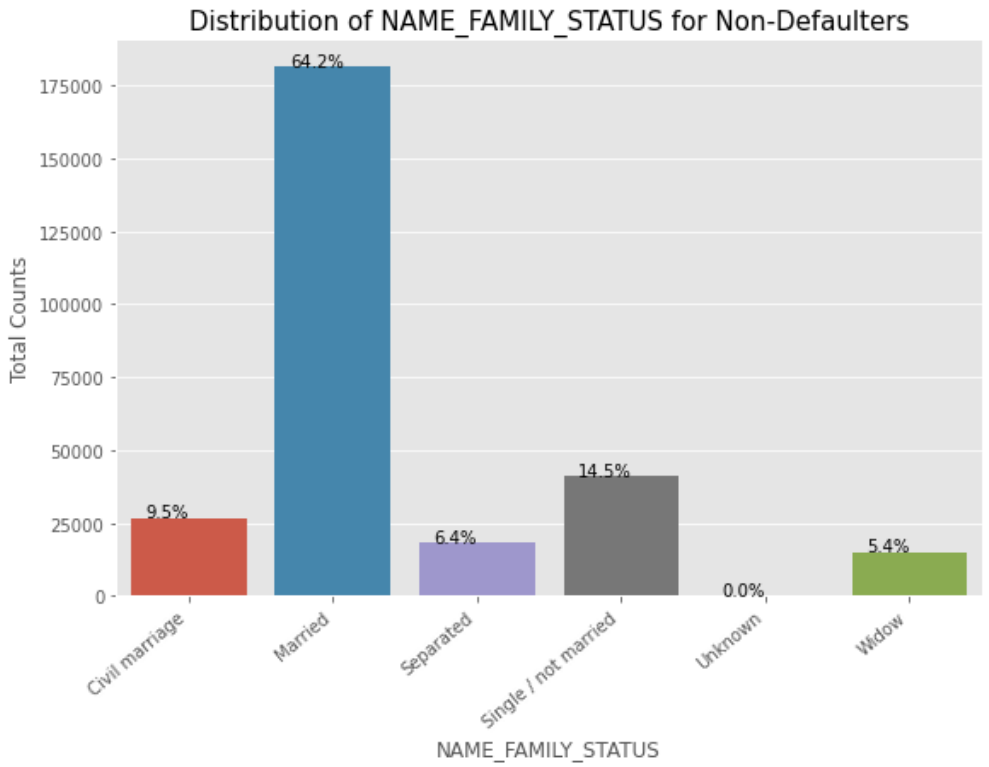
NAME_INCOME_TYPE

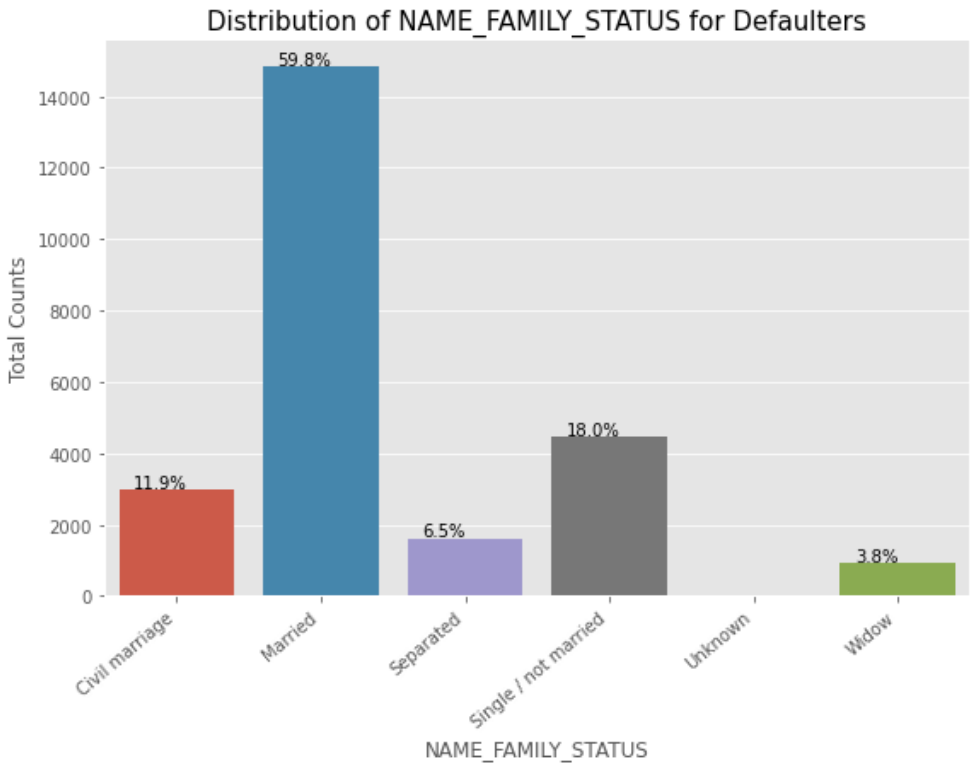




- Most of the loans are distributed to working category.
- We also see that working class people contribute 51% to non defaulters while they contribute to 61% of the defaulters.
- Clearly, the chances of defaulting are more in their case.
- Students are not defaulters (Reason could be they are not yet employed)

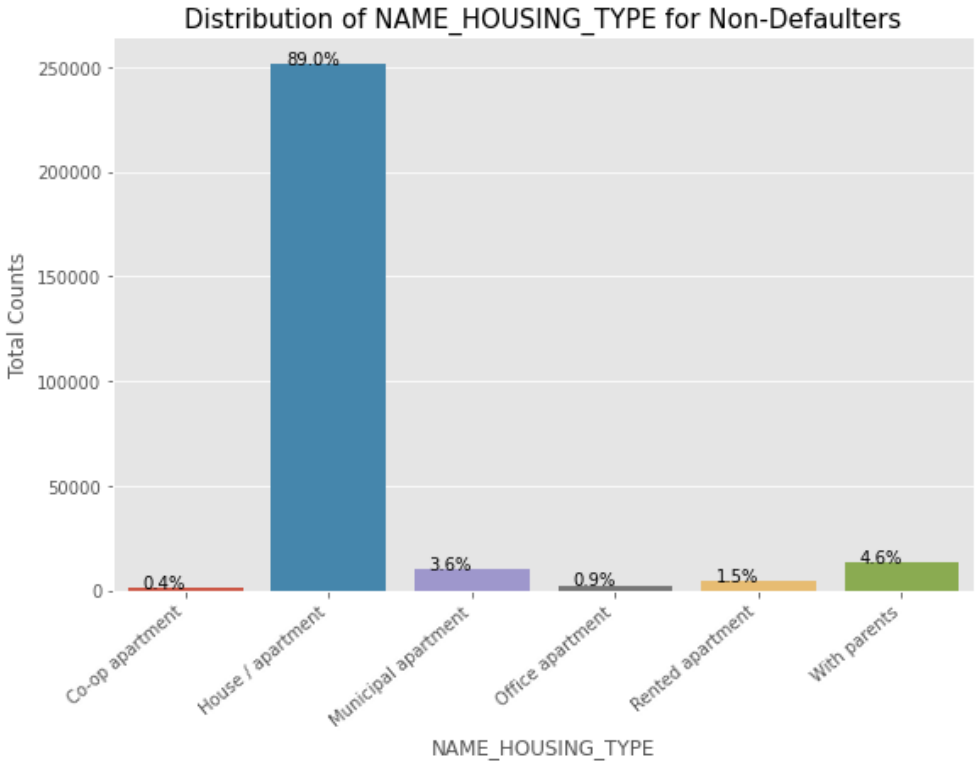
NAME_FAMILY_STATUS

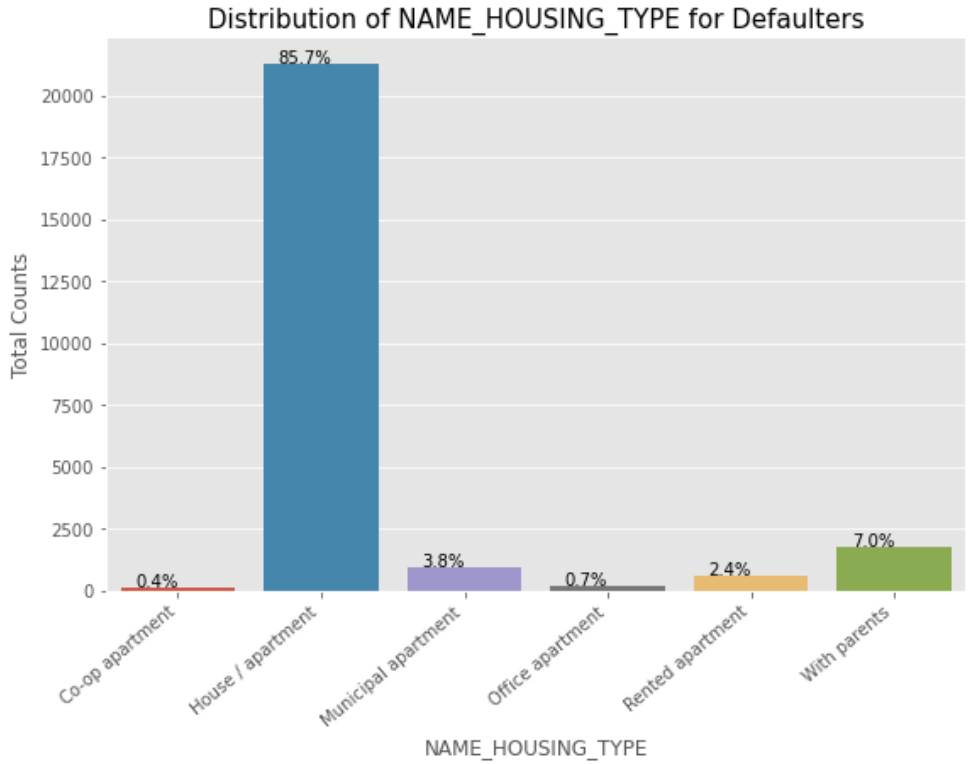




- Married people tend to apply for more loans comparatively.
- We see that Single/non-married people contribute 14.5% to non-defaulters and 18% to the defaulters. So, there is more risk associated with them.
- We see that Civil marriage people contribute 11.9% to Defaulters and 9.5% to the non-defaulters. So, there is more risk associated with them.

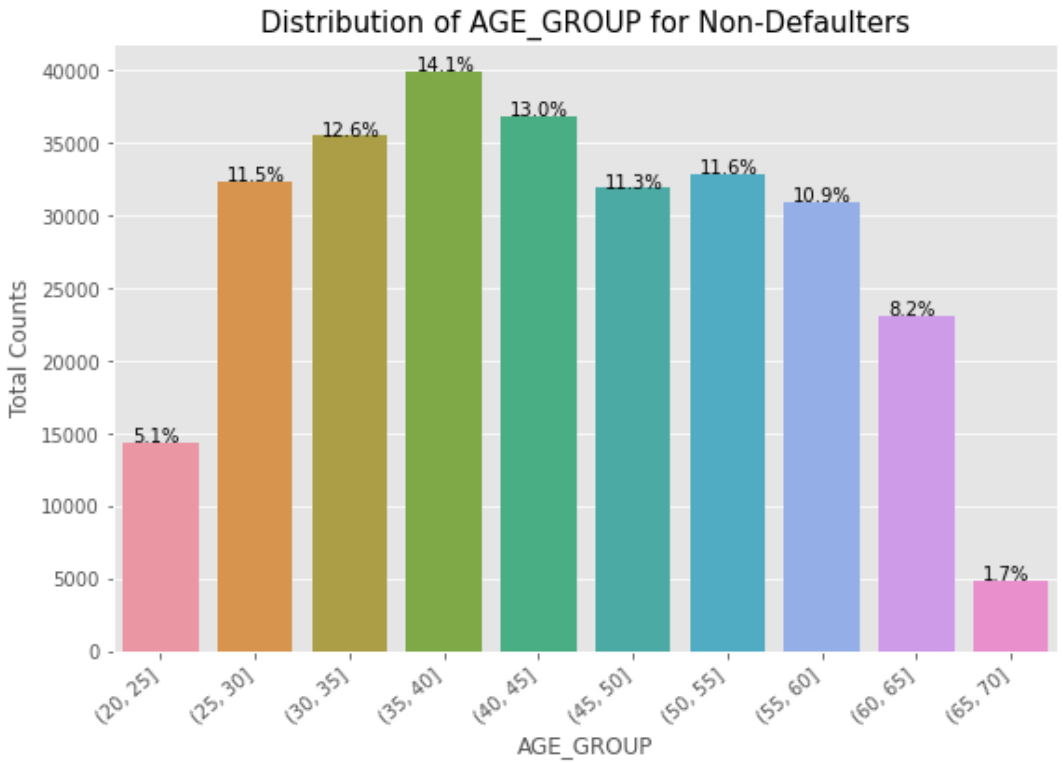
NAME_HOUSING_TYPE

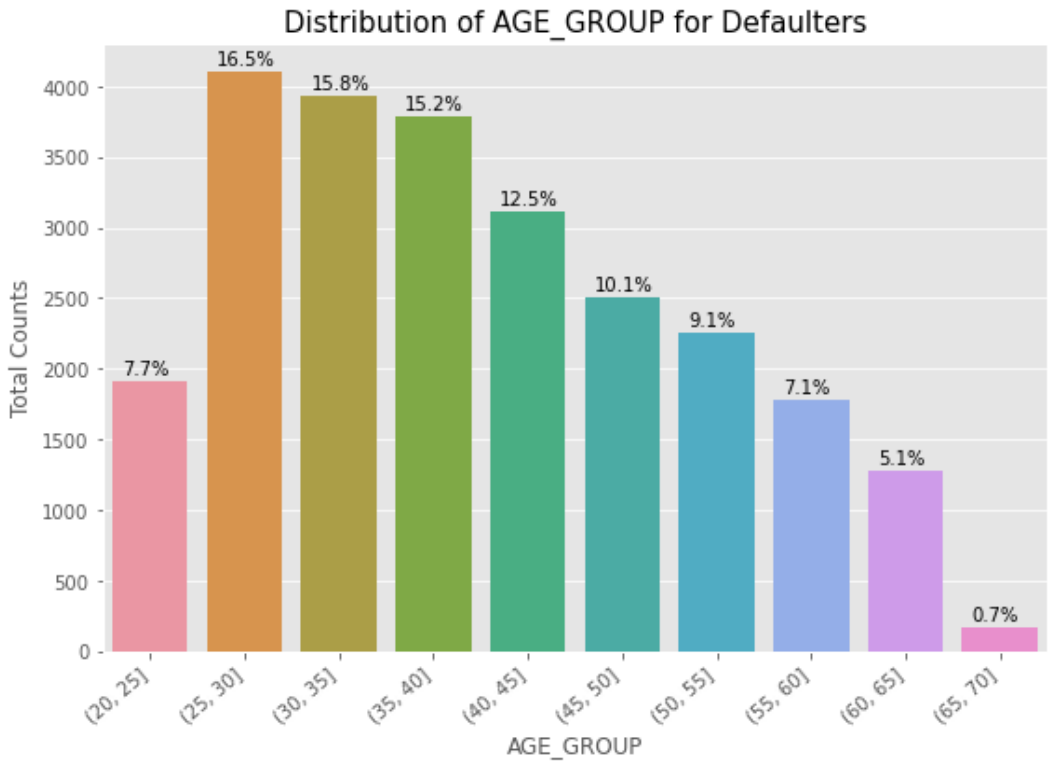




- It is clear from the graph that people who have House/Apartments, tend to apply for more loans and even we see 85.7% defaulters and 89% non-defaulters.
- People living with parents tend to default more often when compared with others.

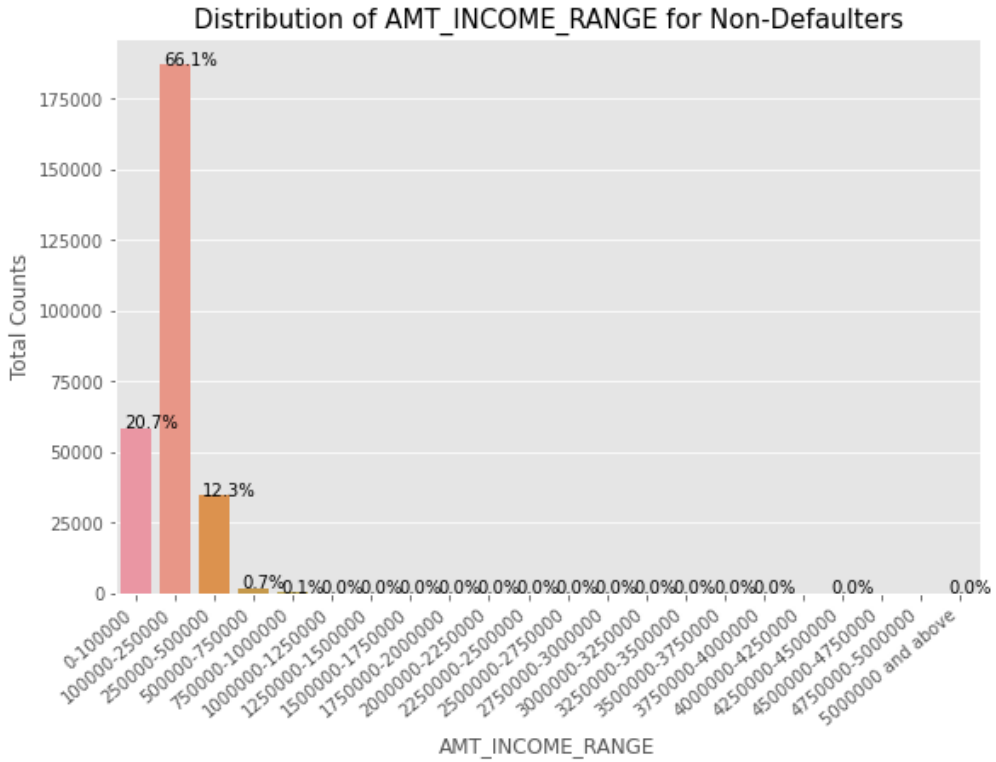
AGE_GROUP

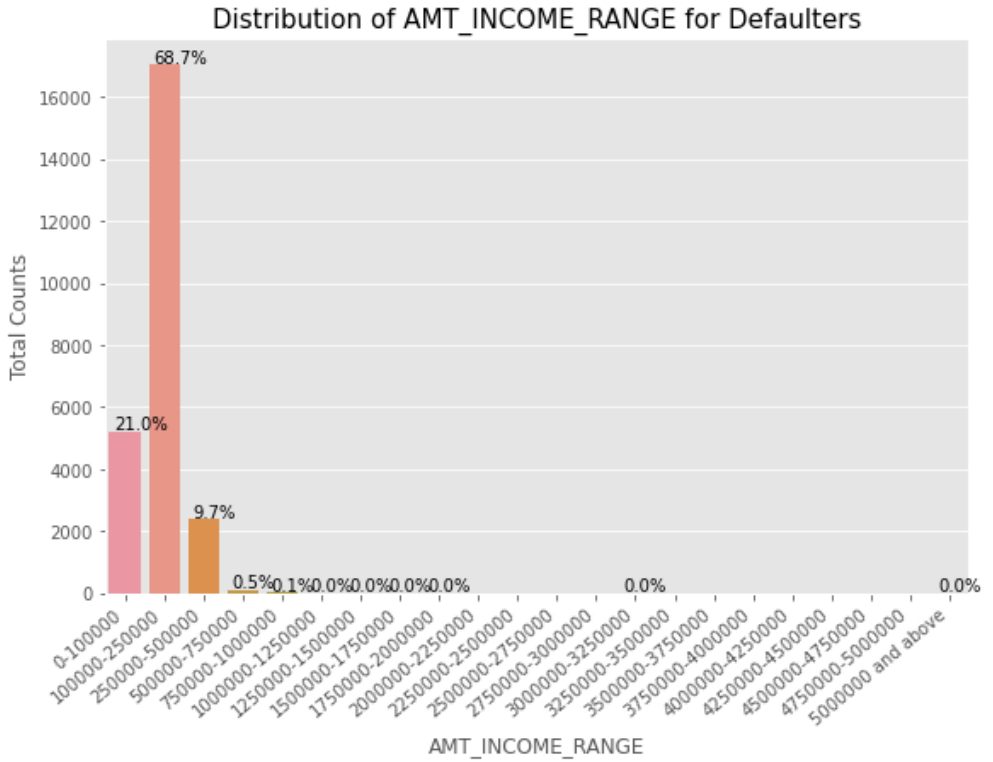




- We see that (25,30] age group tend to default more often. So, the people with age group 20 to 35 are risky.
- With increasing age group, people tend to default less starting from the age 25.
- The reason could be they get employed around that age. (With increasing age, their salary might also increase).

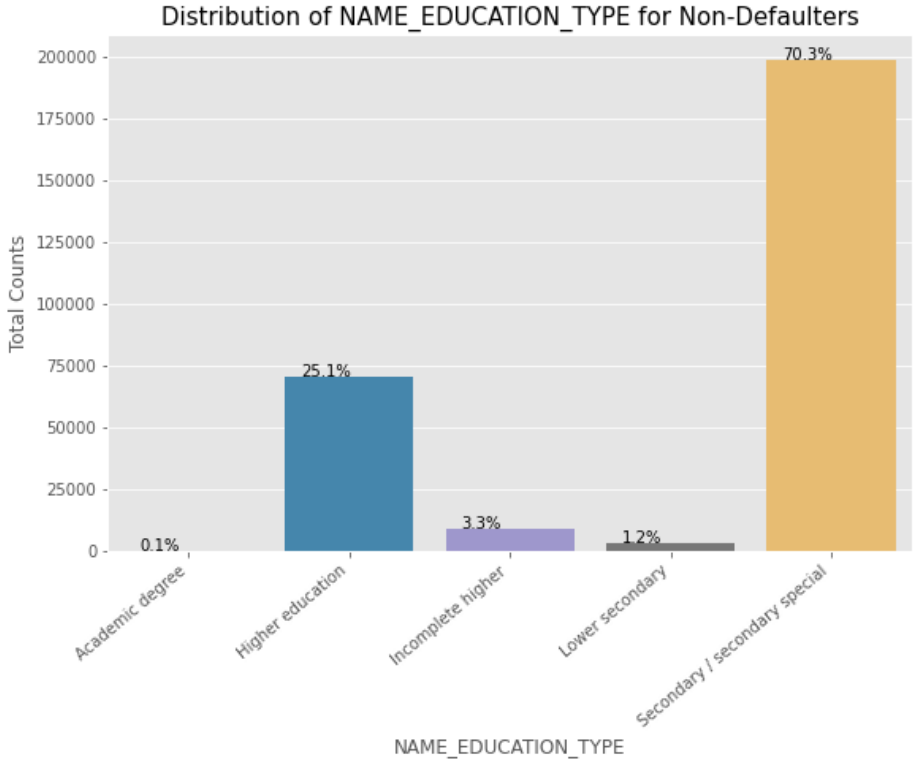
AMT_INCOME_RANGE

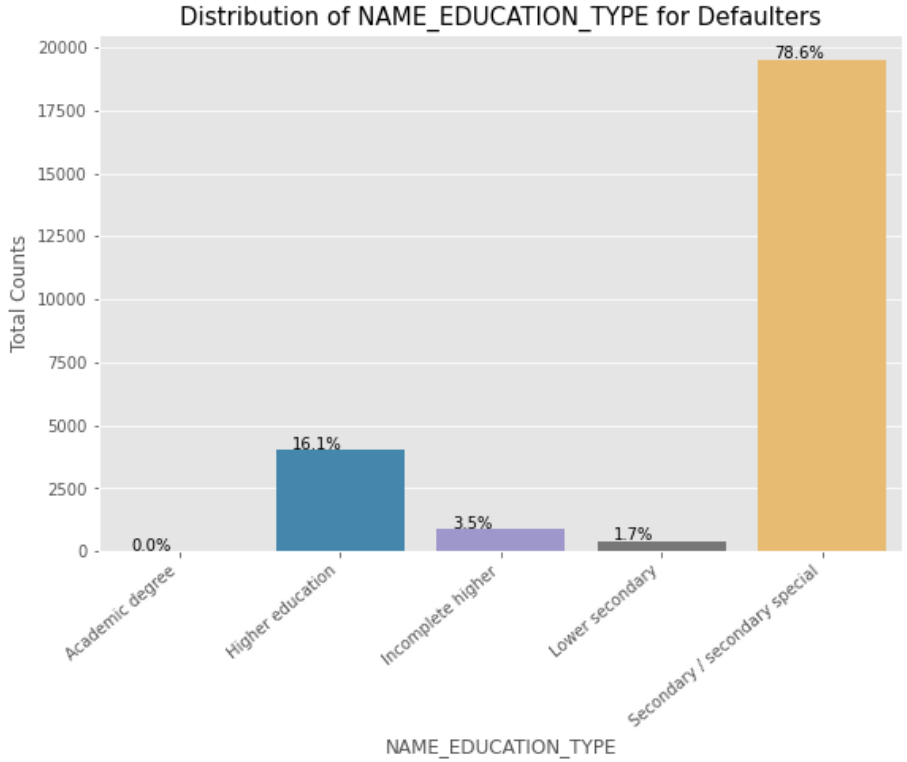




We see that people with income range 100000-250000 contribute 68.7% to defaulters and 66.1% to the non-defaulters. So, there is some what risk associated with them.

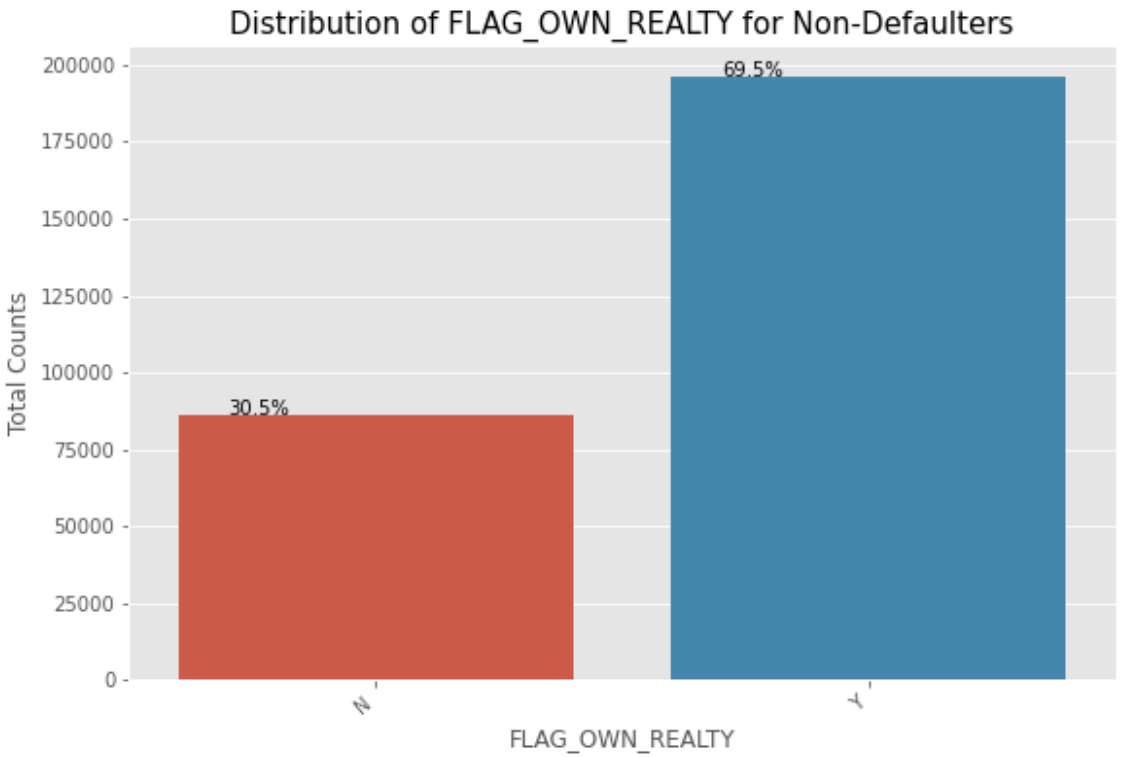
NAME_EDUCATION_TYPE

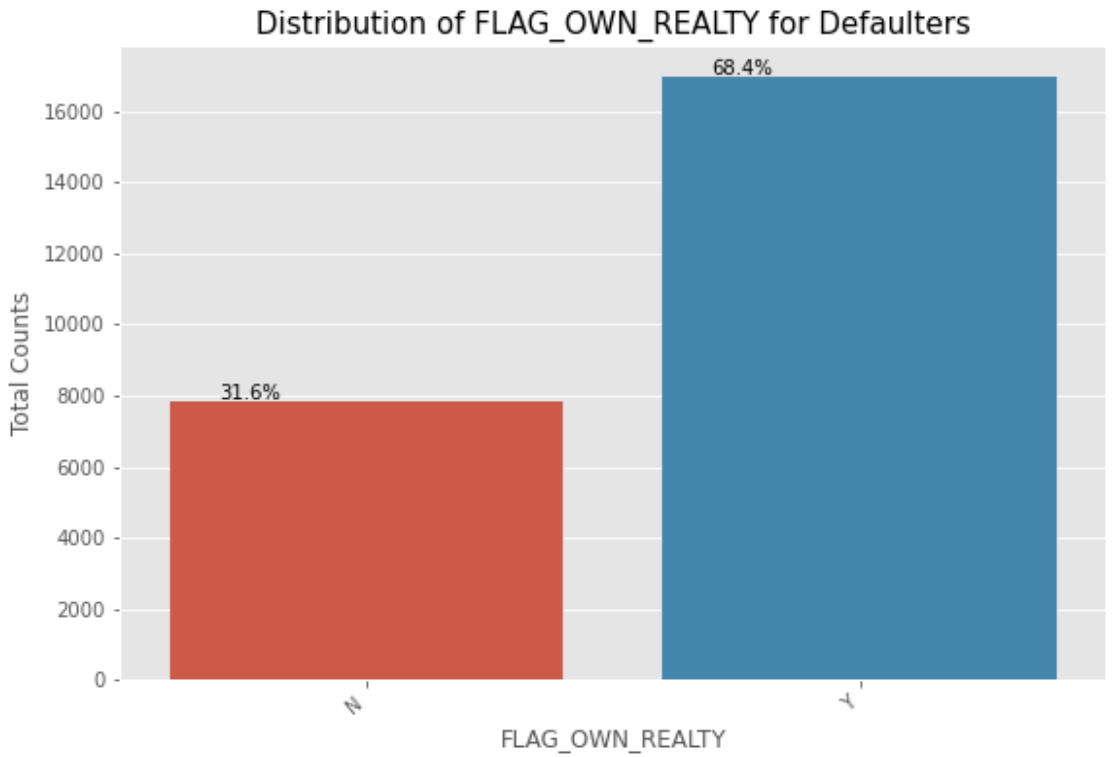




- Almost all the Education categories are equally likely to default except for the higher educated ones who are less likely to default, and secondary educated people are more likely to default

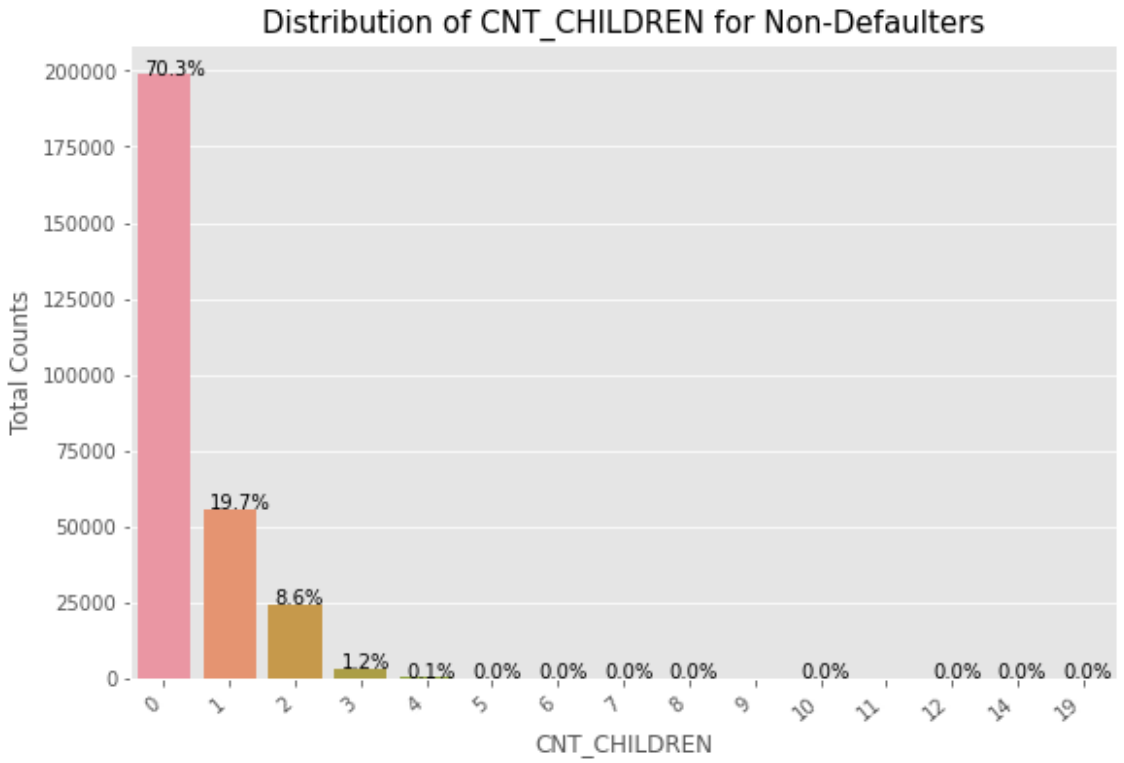
FLAG_OWN_REALTY

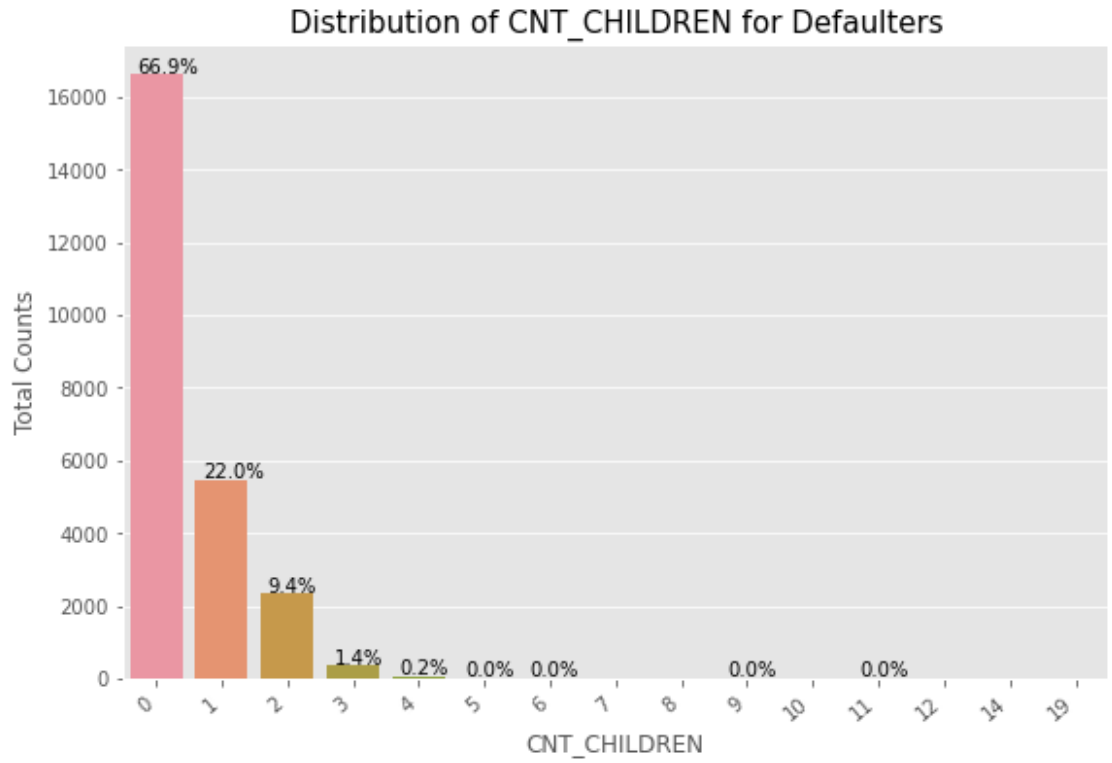




- Most of the applicants owns a house or flat.
- There are more defaulters when compared to non-defaulters if, the applicants has no house or flat.

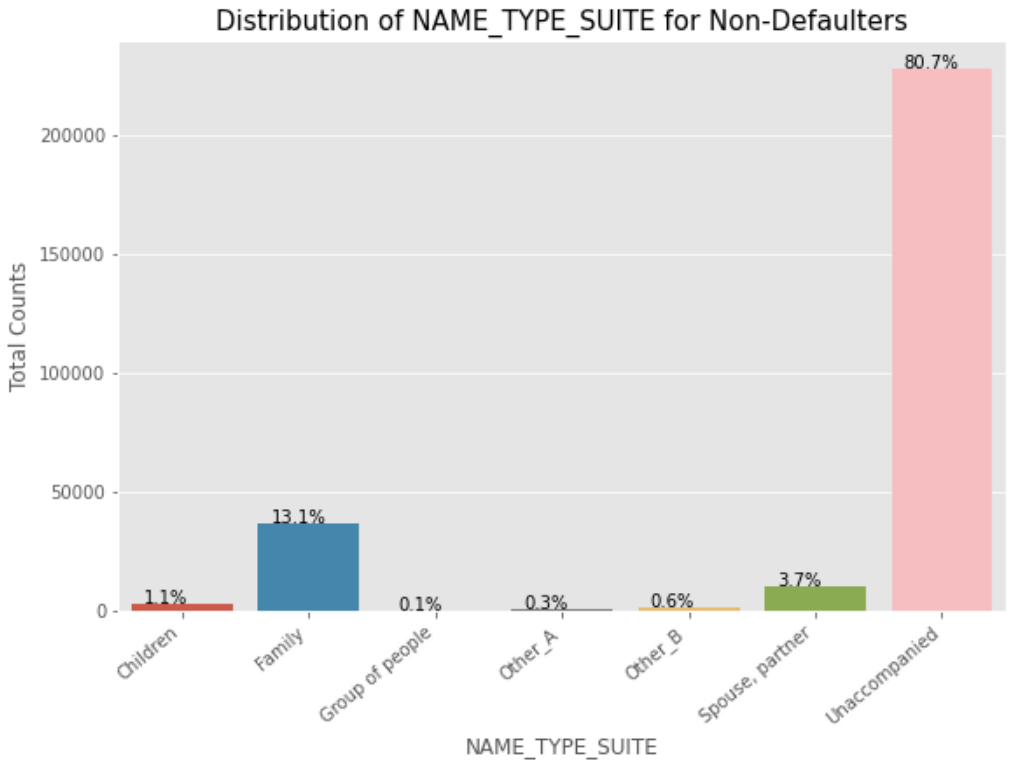
CNT_CHILDREN

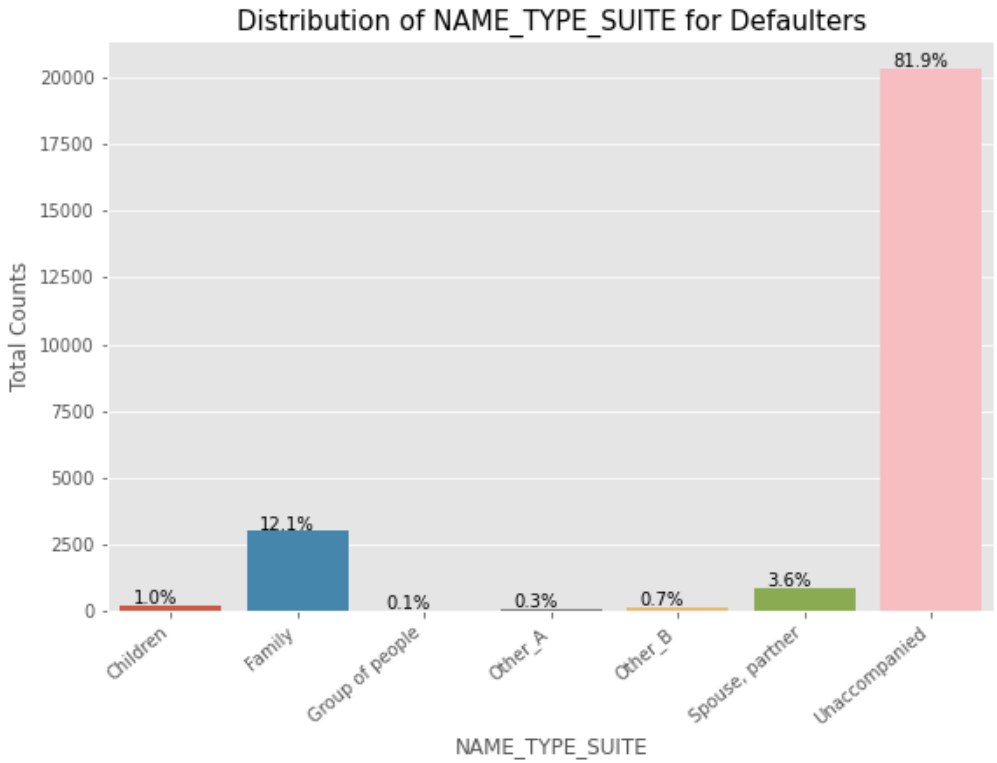




- Most number of applicants have max of 3 children.
- Applicants having 1 child are most likely to default. (But its non conclusive)

NAME_TYPE_SUITE



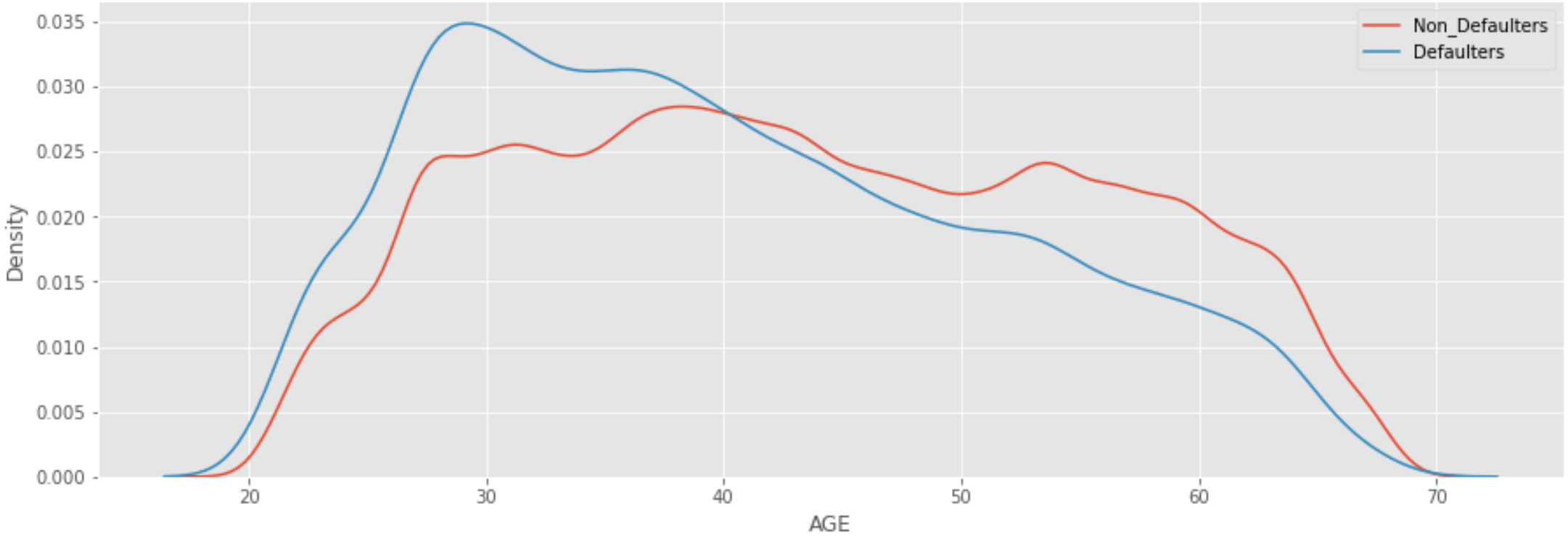


Most of the applicants are unaccompanied

UNIVARIATE ANALYSIS

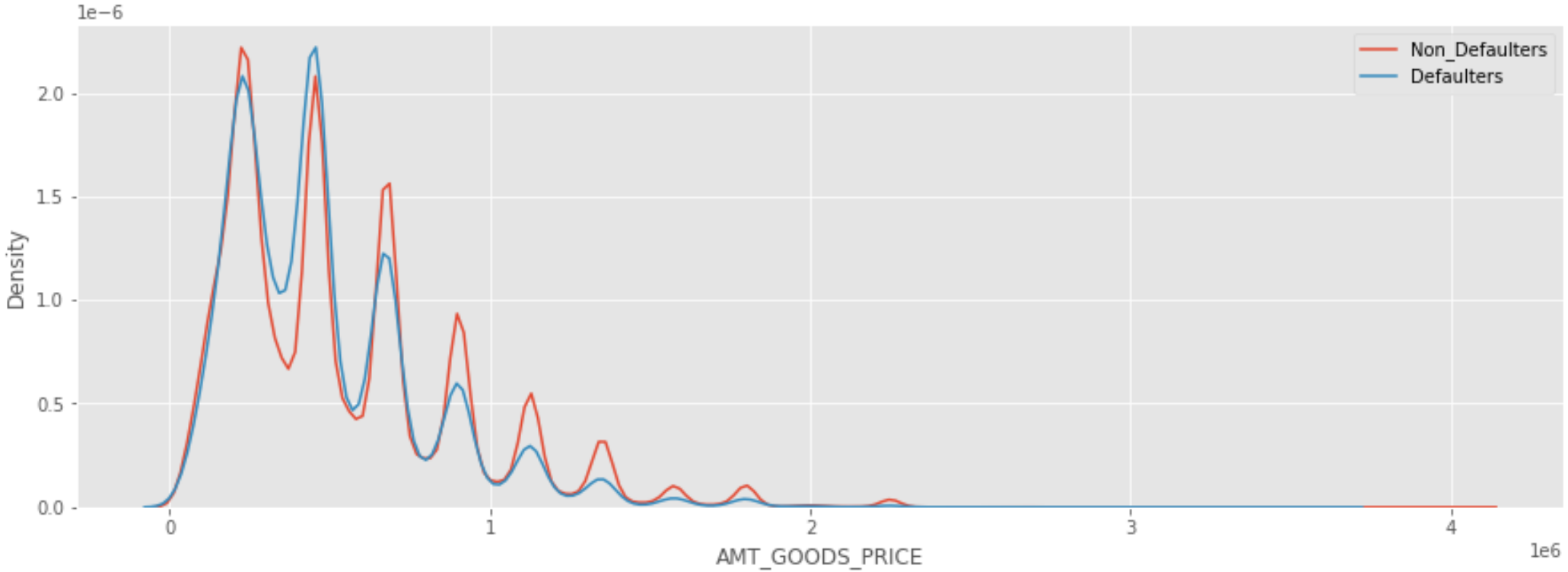
NUMERICAL COLUMNS

DAYS_BIRTH



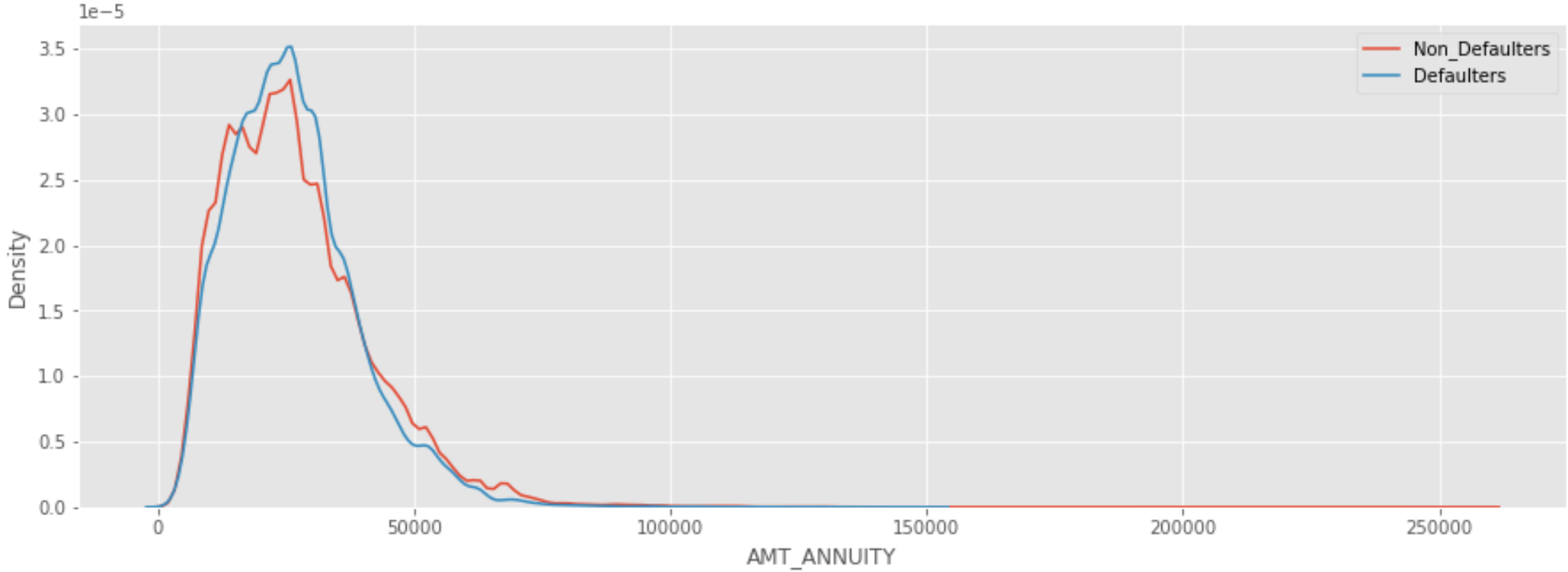
- By this plot, we can say that...The greatest number of defaulters are in the range 28 to 30.
- If we see one intersection in the plot, After the age of 40, we can clearly see that the number of defaulters are less compared to the non-defaulters.
- As age increases, the number of defaulters decreases...One of the reason could be as age increases, people will get employed

AMT_GOODS_PRICE



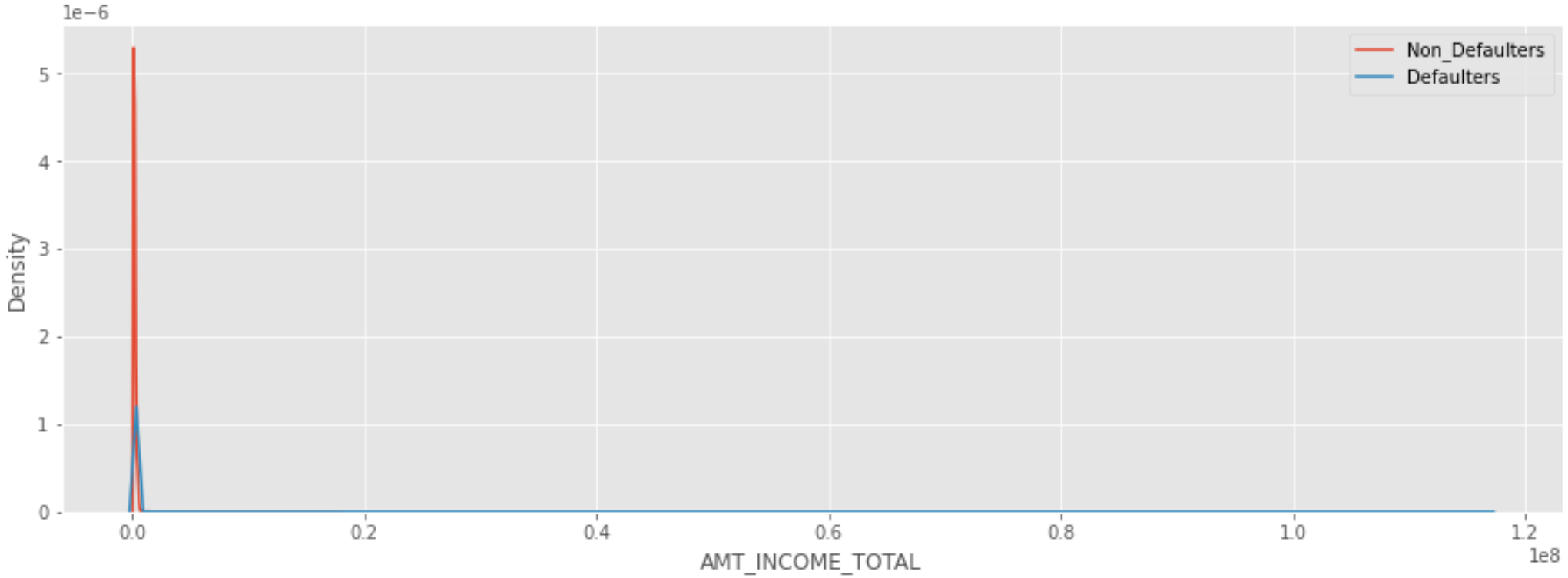
- By this, AMT_GOODS_PRICE has no impact on defaulters and non-defaulters.
- But there are more applicants if the price of the goods for which the loan is given is less.

AMT_ANNUITY



More number of applicants have less loan annuity

AMT_INCOME_TOTAL

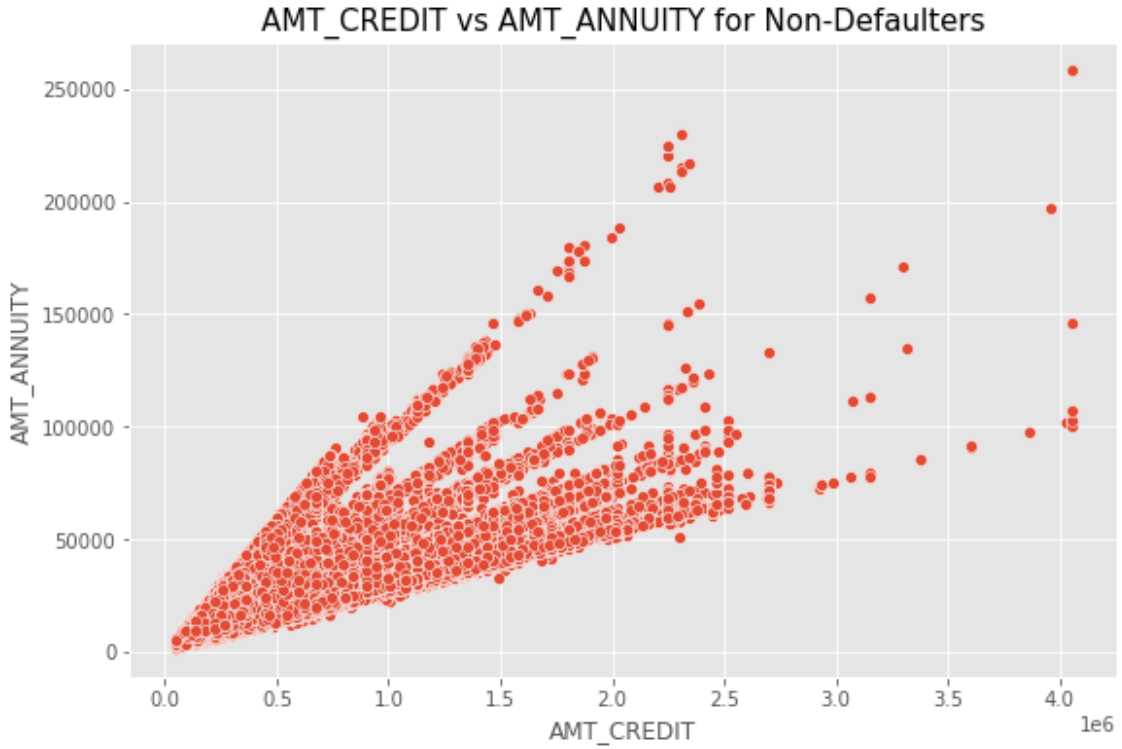


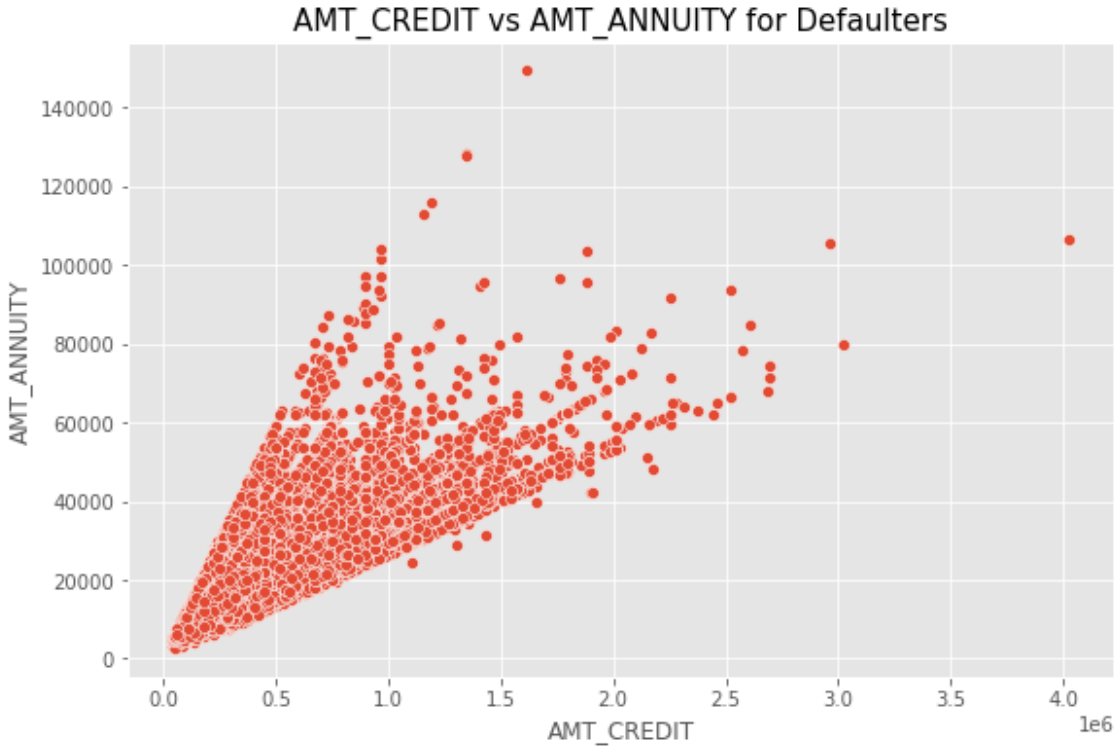
People with lower income are more likely to default.

BIVARIATE ANALYSIS

CONTINUOUS - CONTINUOUS

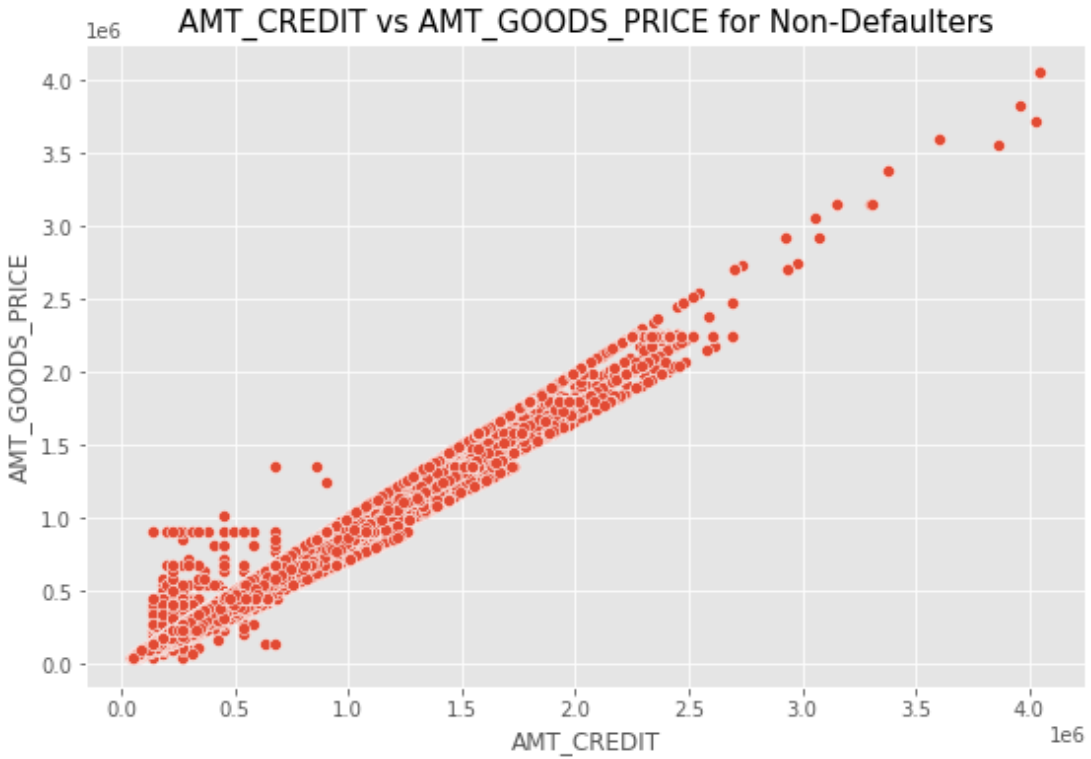
AMT_CREDIT & AMT_ANNUITY

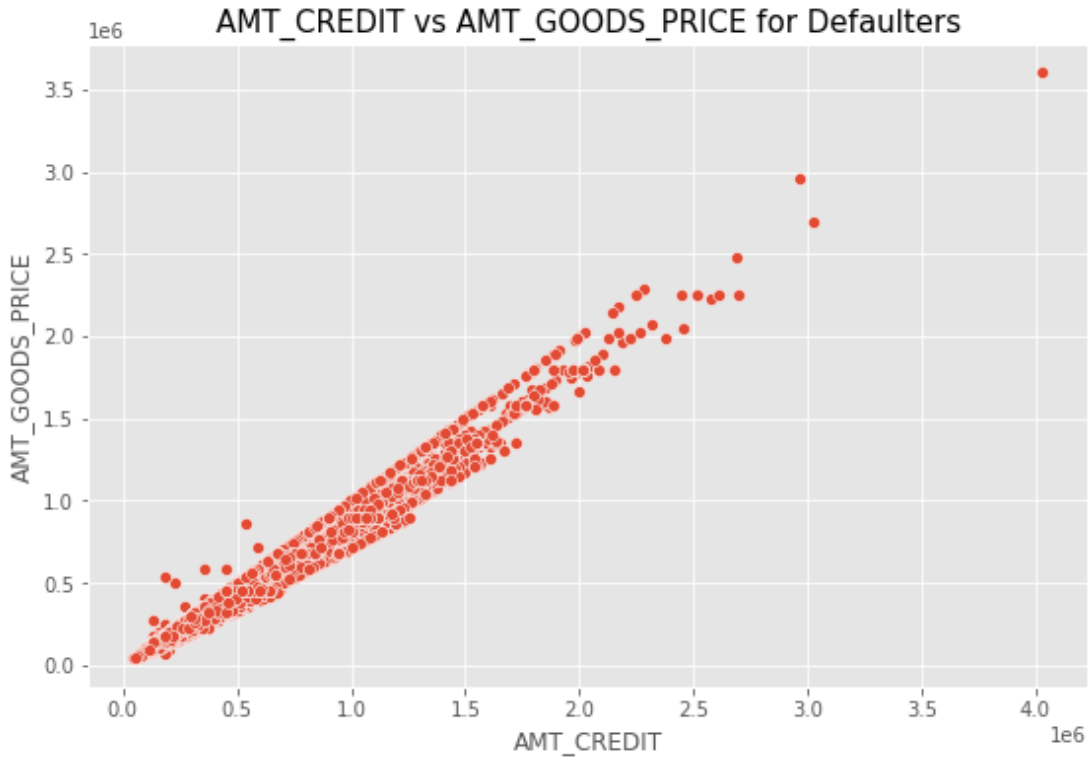




Defaulters take more credit for the annuity that they have.

AMT_CREDIT & AMT_GOODS_PRICE



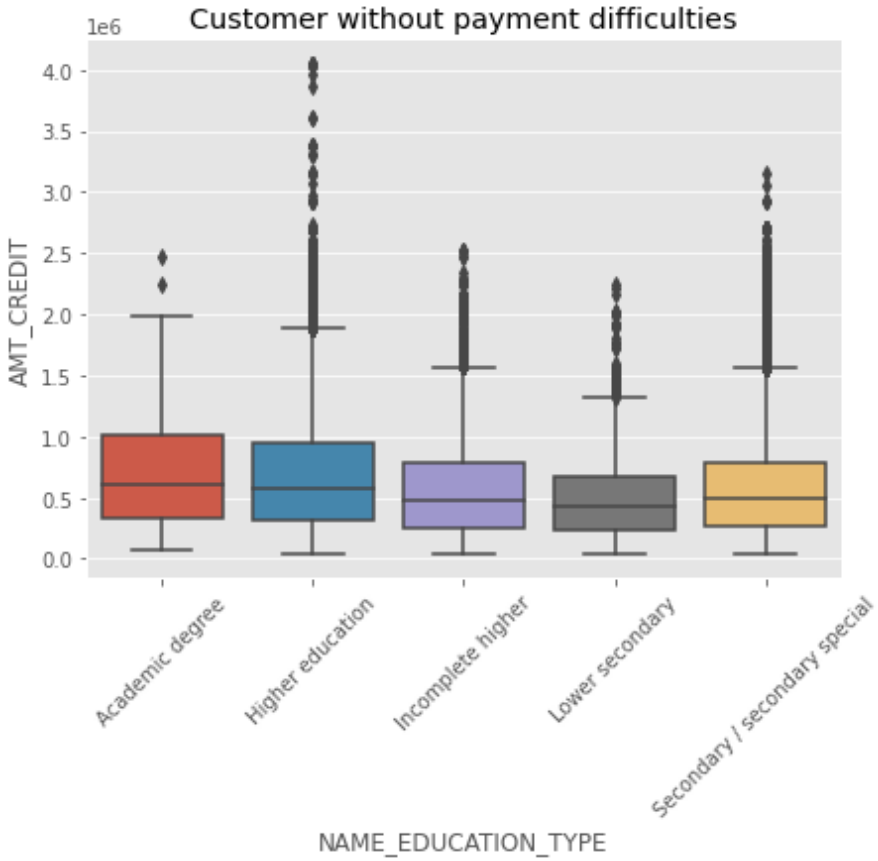


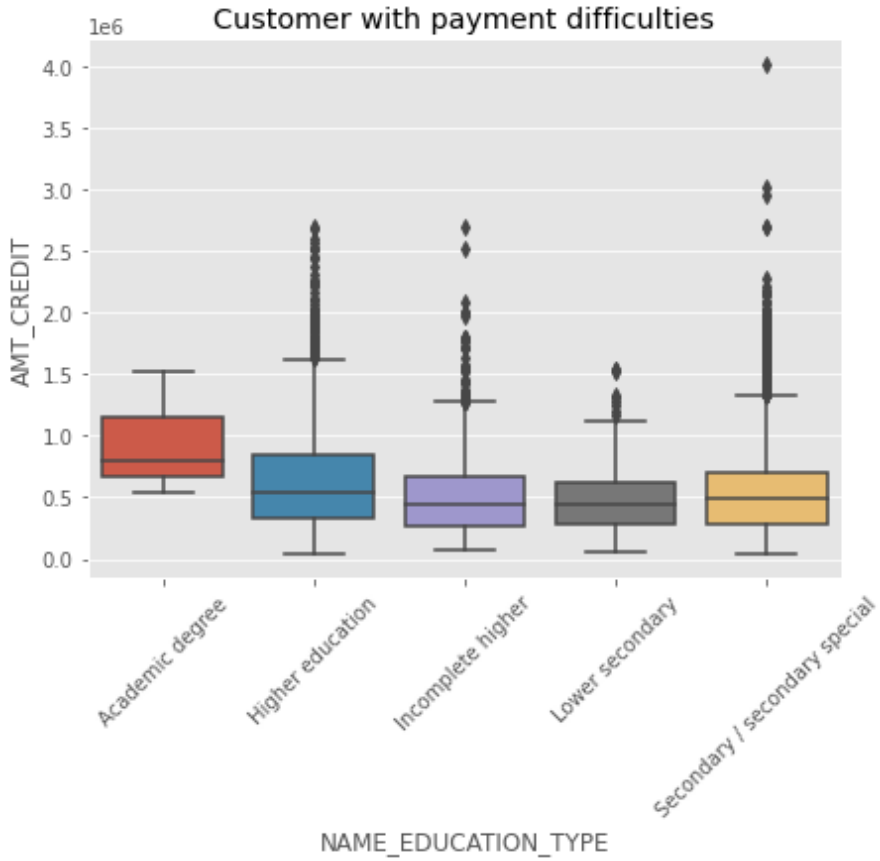
We can see that goods price is positively correlated with credit amount.

BIVARIATE ANALYSIS

CATEGORICAL-NUMERICAL

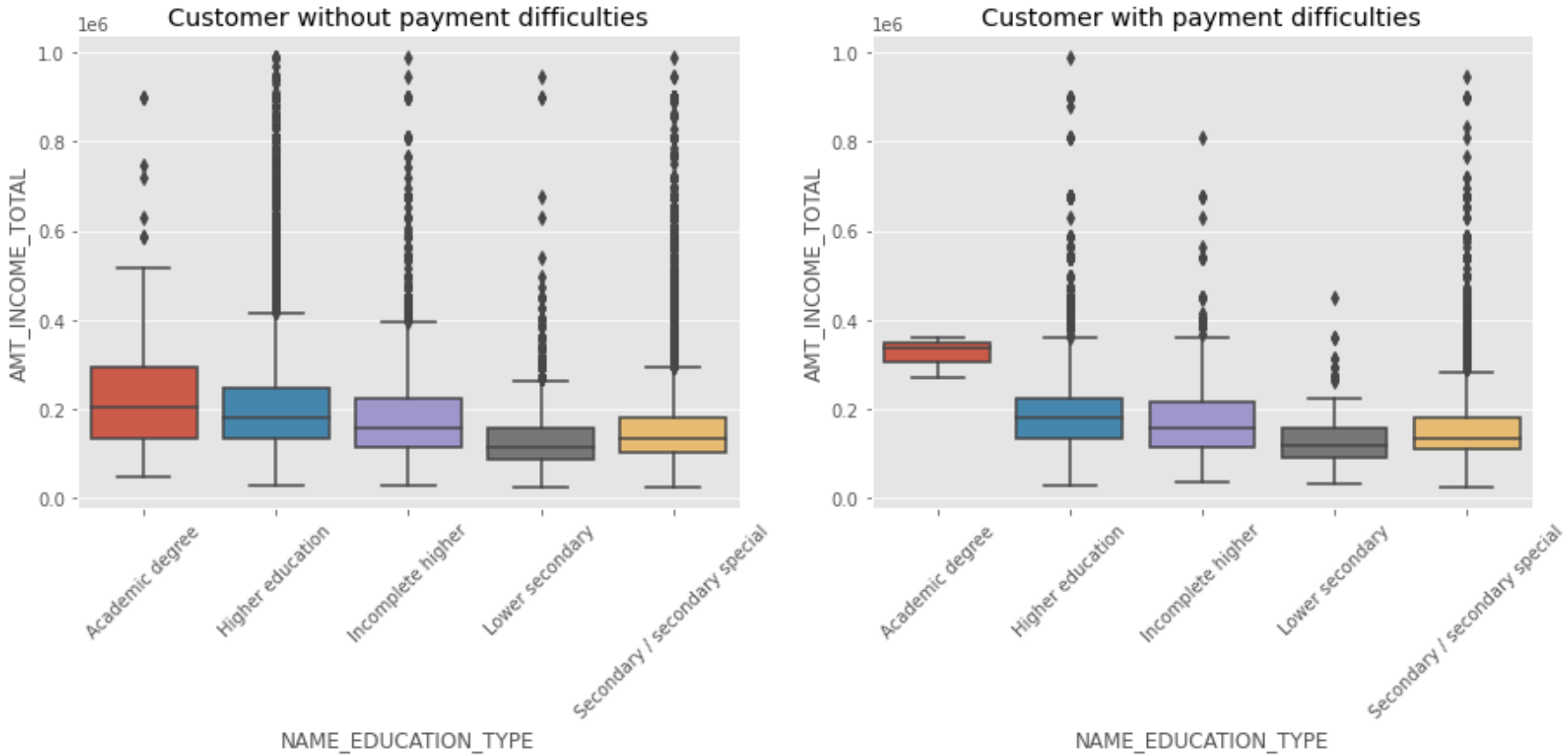
NAME_EDUCATION_TYPE & AMT_CREDIT





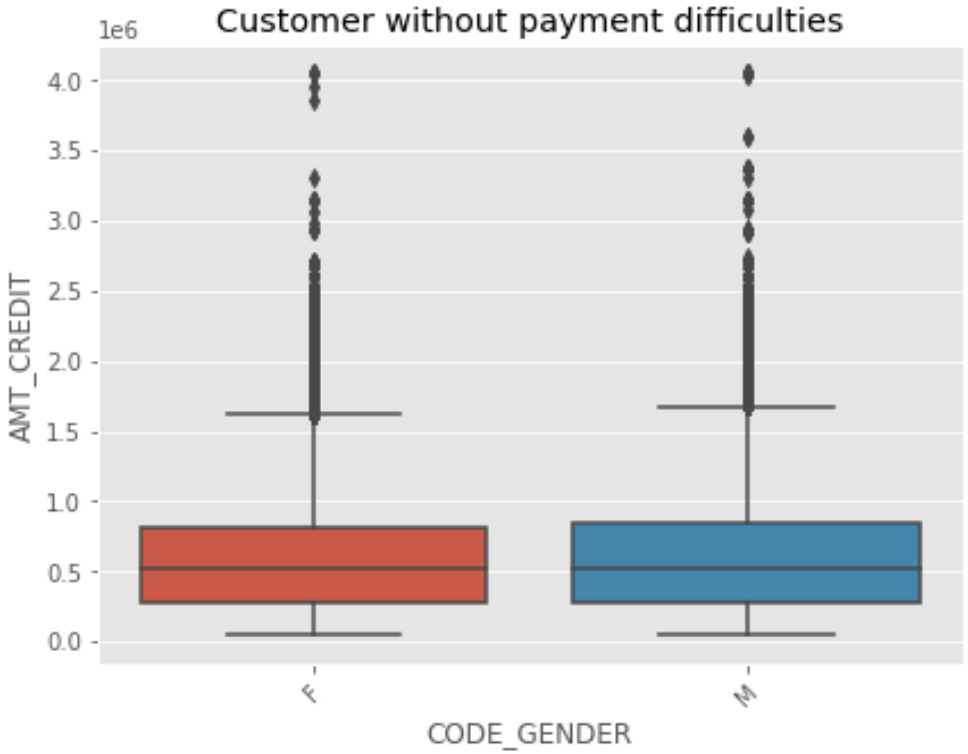
The average credit amount for the defaulters is high when compared to the non-defaulters in case of academic degree.

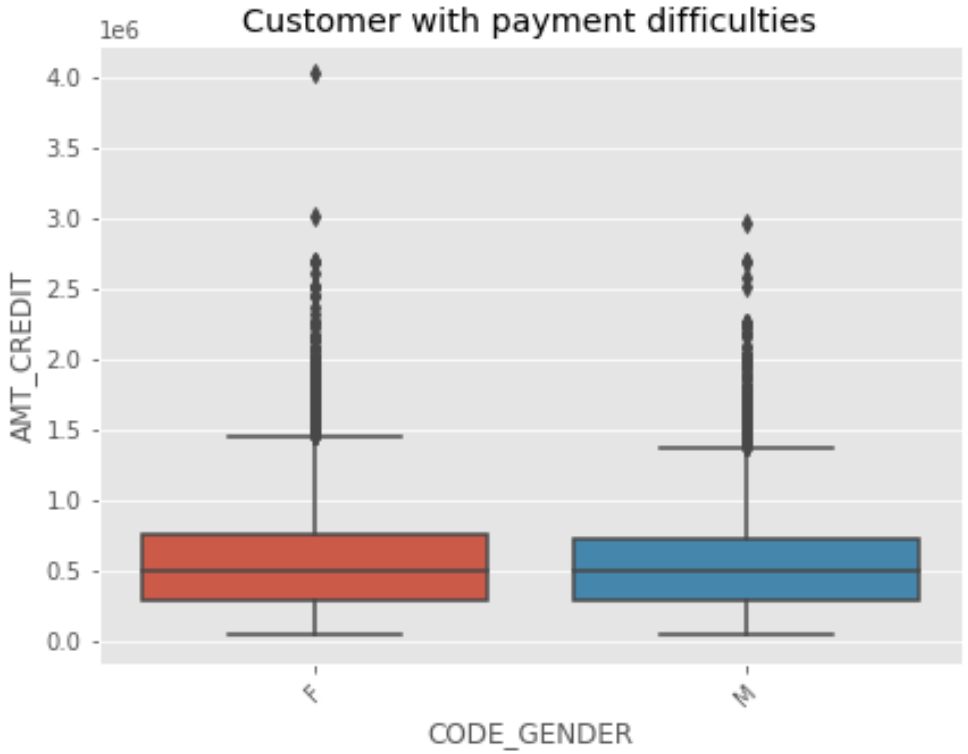
AMT_INCOME_TOTAL & NAME_EDUCATION_TYPE



The average income amount for the defaulters is clearly high when compared to the non-defaulters in case of academic degree.

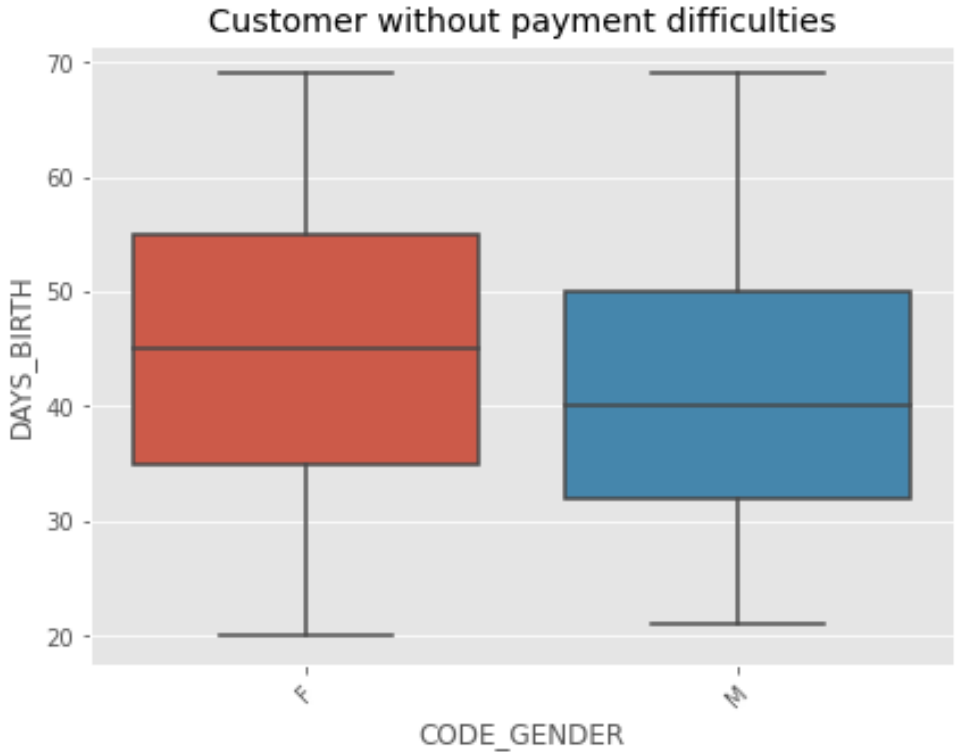
AMT_CREDIT & CODE_GENDER

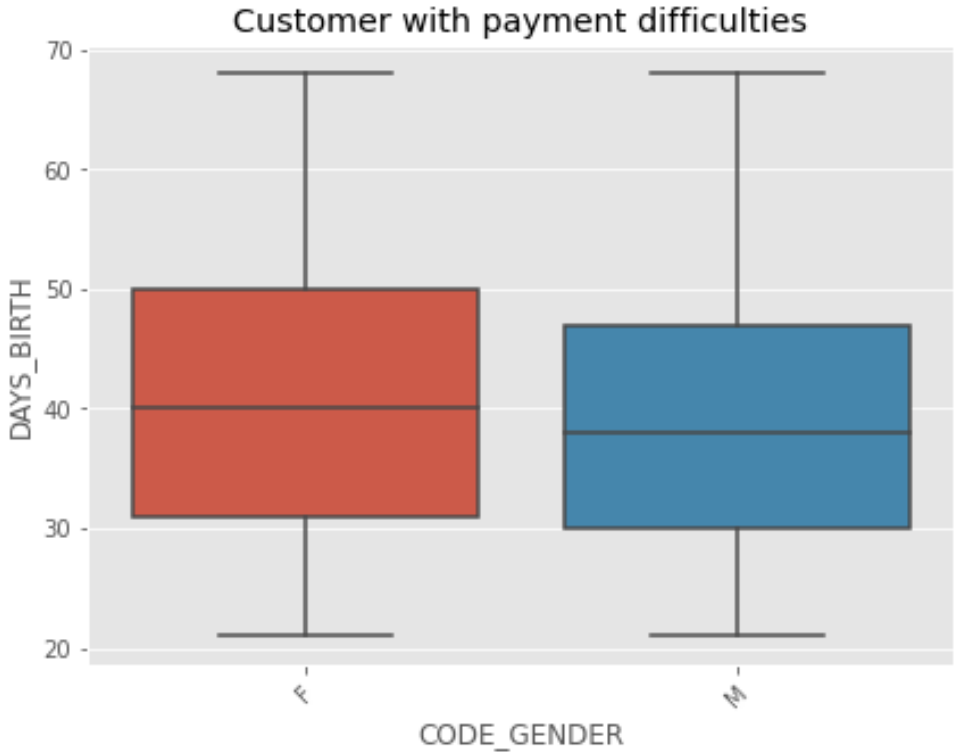




The average credit amount for the defaulters and the non-defaulters is almost same when compared with genders.

DAYS_BIRTH & CODE_GENDER



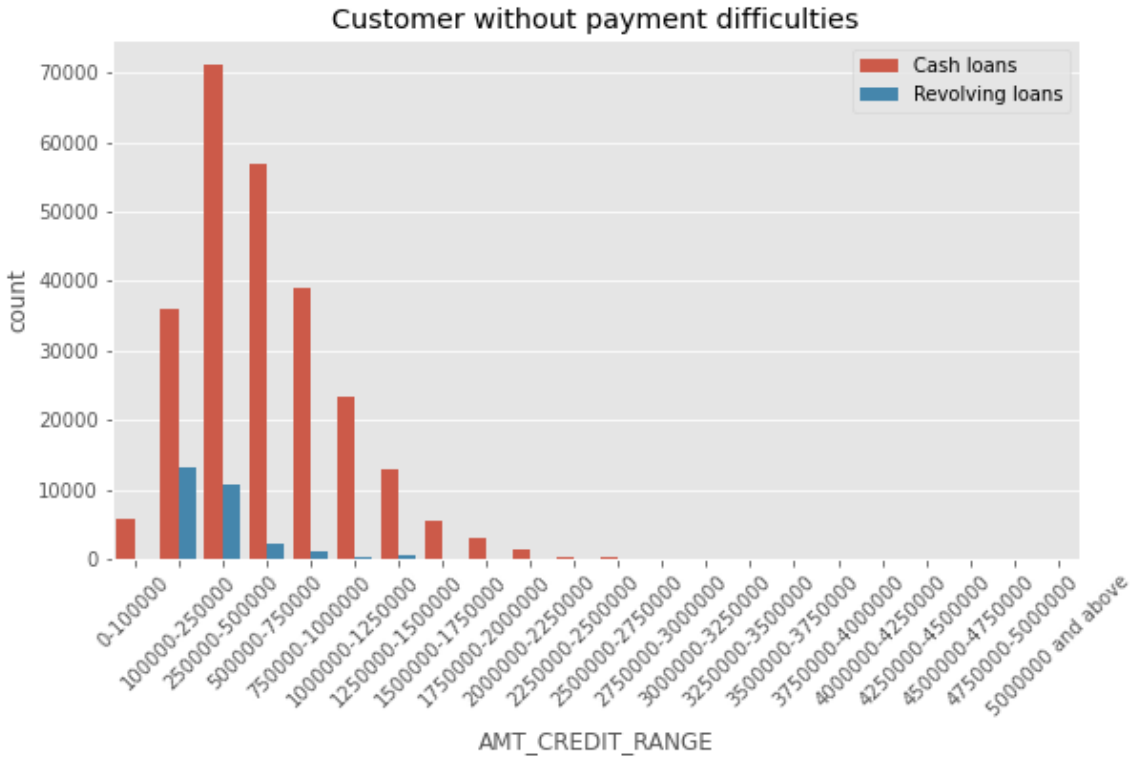


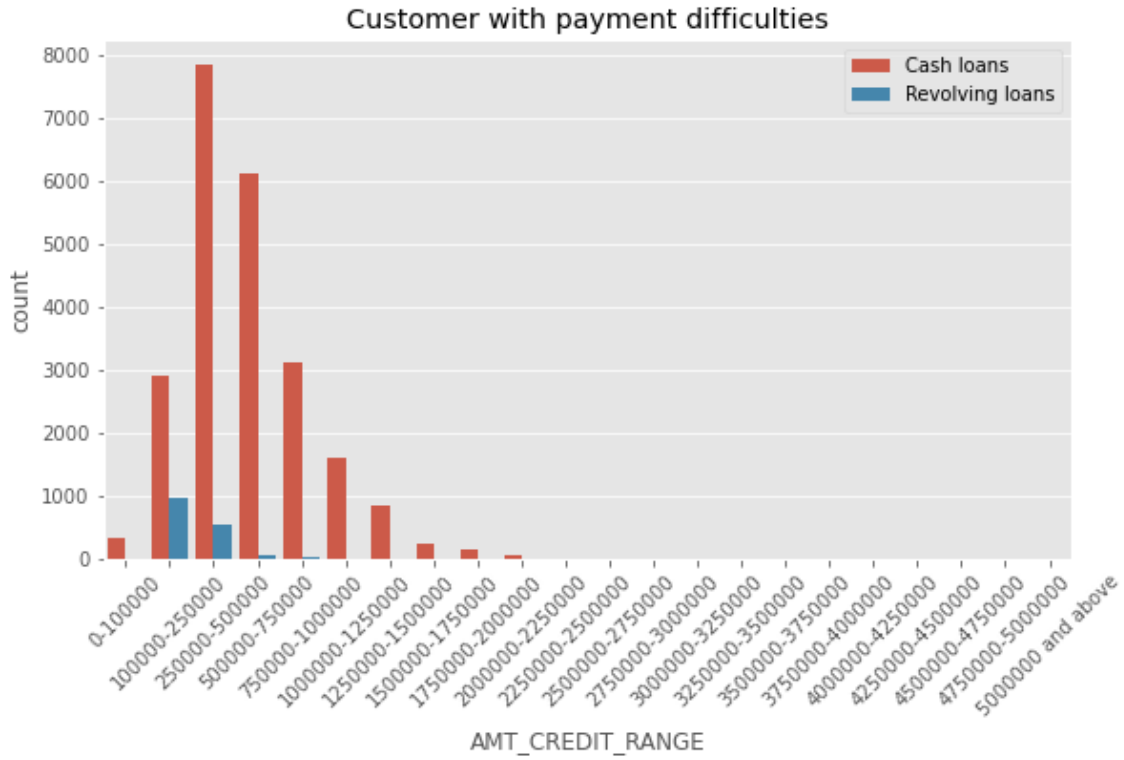
- The average age for the defaulters and the non-defaulters is almost same with Males.
- The average age for the non-defaulters is high when compared to the defaulters with females.

BIVARIATE ANALYSIS

CATEGORICAL - CATEGORICAL

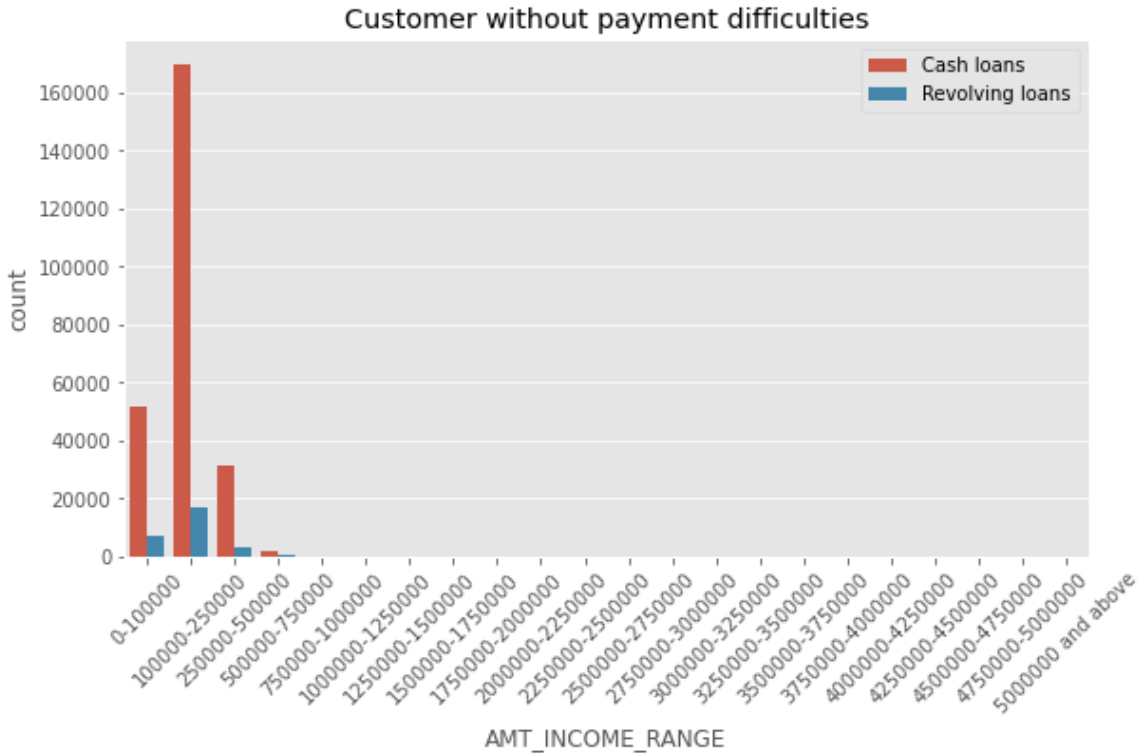
AMT_CREDIT_RANGE & NAME_CONTRACT_TYPE

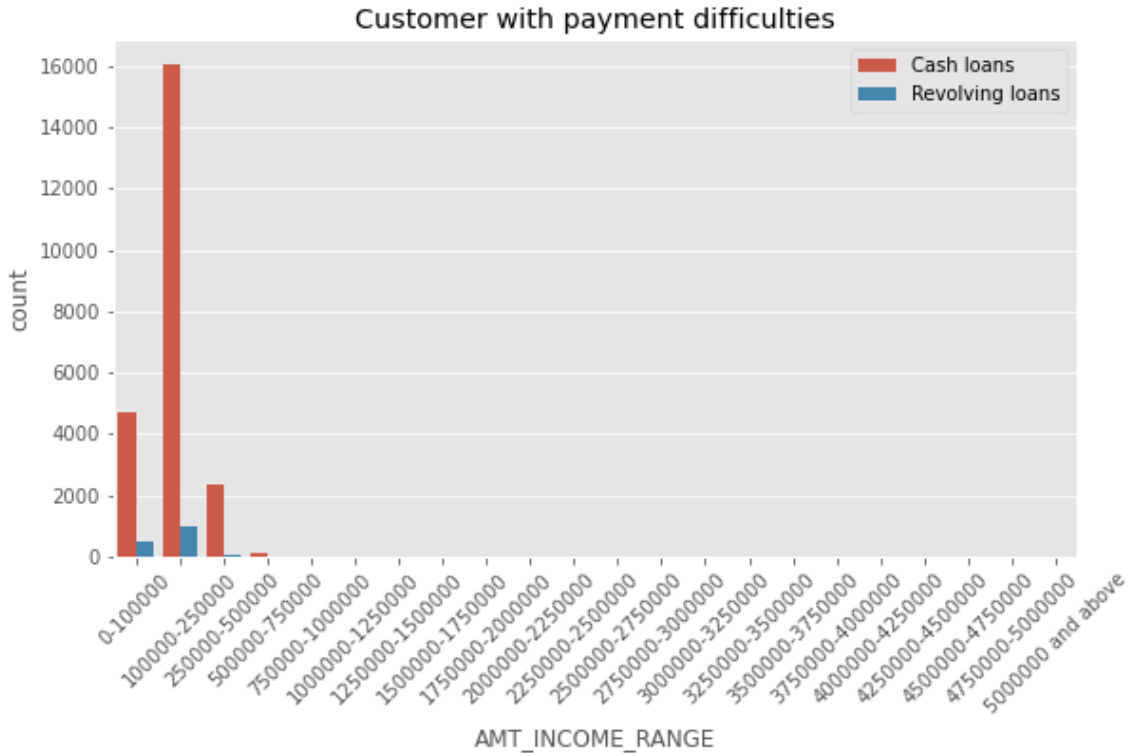




- We can clearly see here that most of the applicant's loan contract type is `cash loans`
- If applicants credit amount is greater than 7,50,000/- they have no difficulty in paying back the loan. (In case of revolving loans).
- Defaulters are more with respect to cash loans...Bank need to consider this type most...

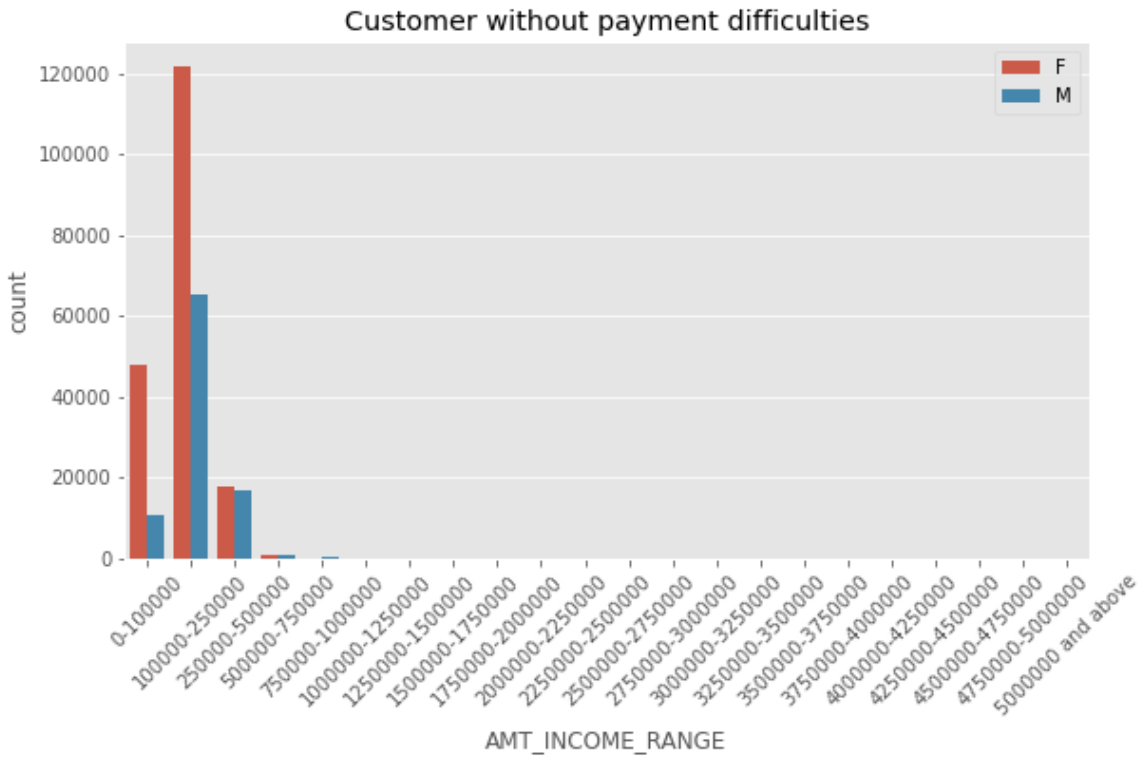
AMT_INCOME_RANGE & NAME_CONTRACT_TYPE

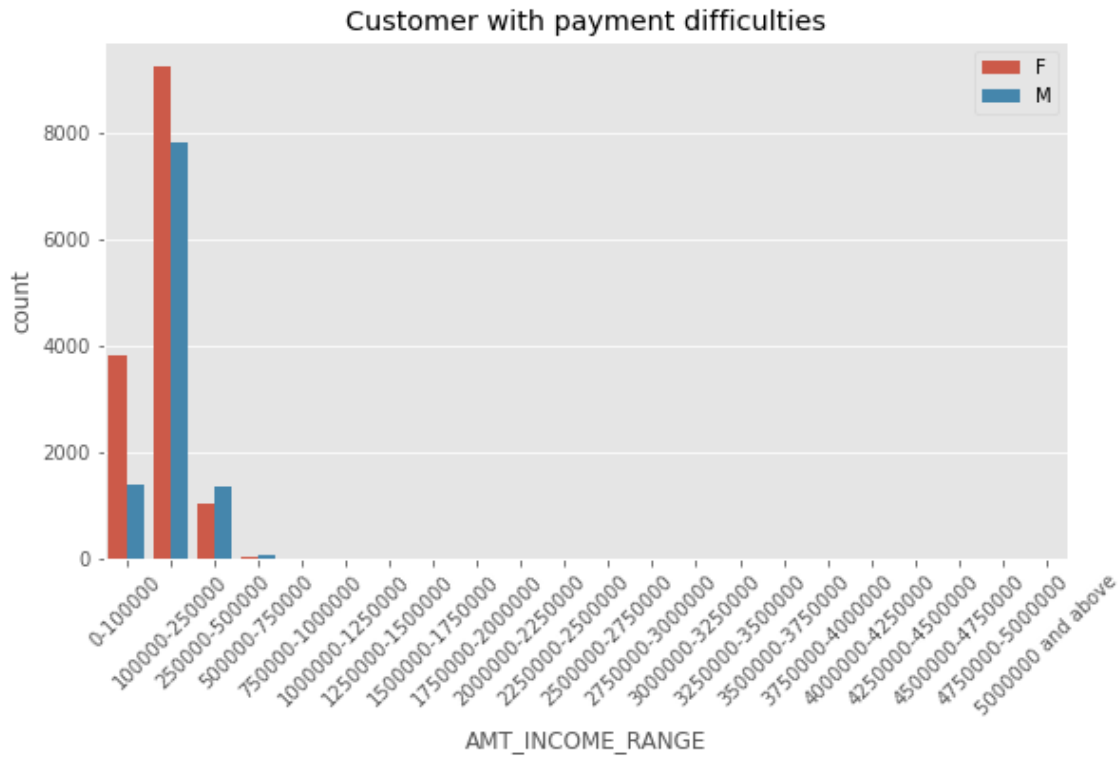




- We can clearly see here that most of the applicant's loan contract type is `cash loans`
- If applicant's income amount is greater than 250000/- they have no difficulty in paying back the loan
- Defaulters are more with respect to cash loans...Bank need to consider this type most..(and income below 2,50,000/-)

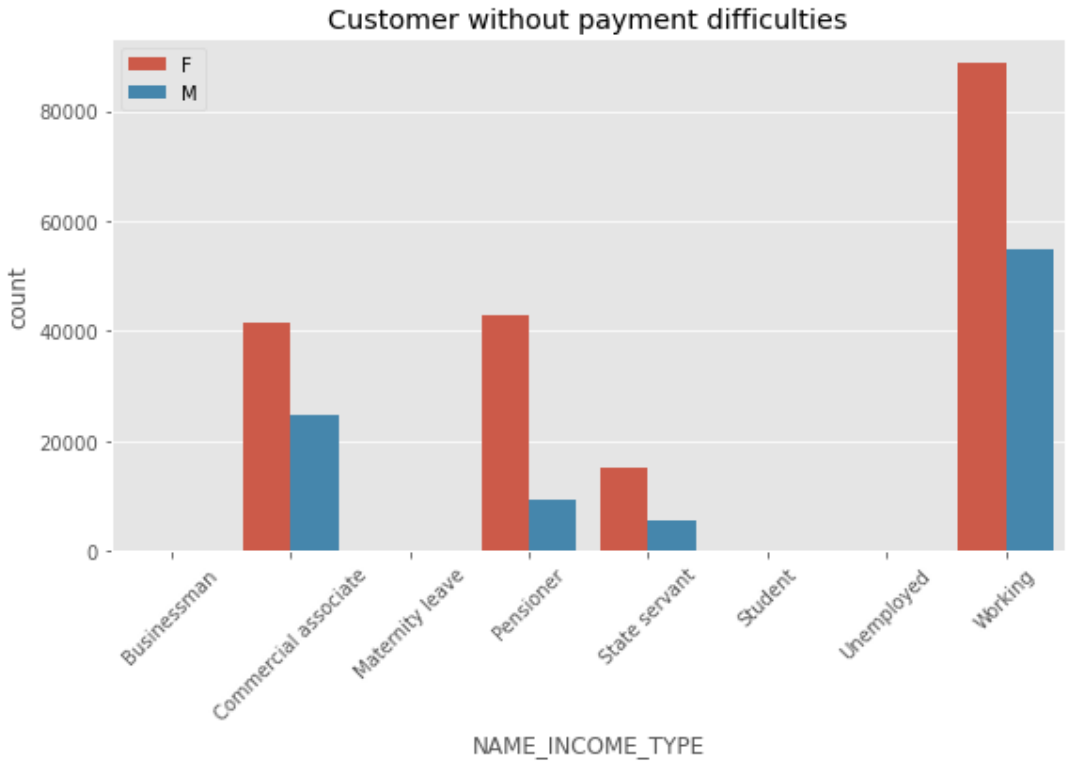
AMT_INCOME_RANGE & CODE_GENDER

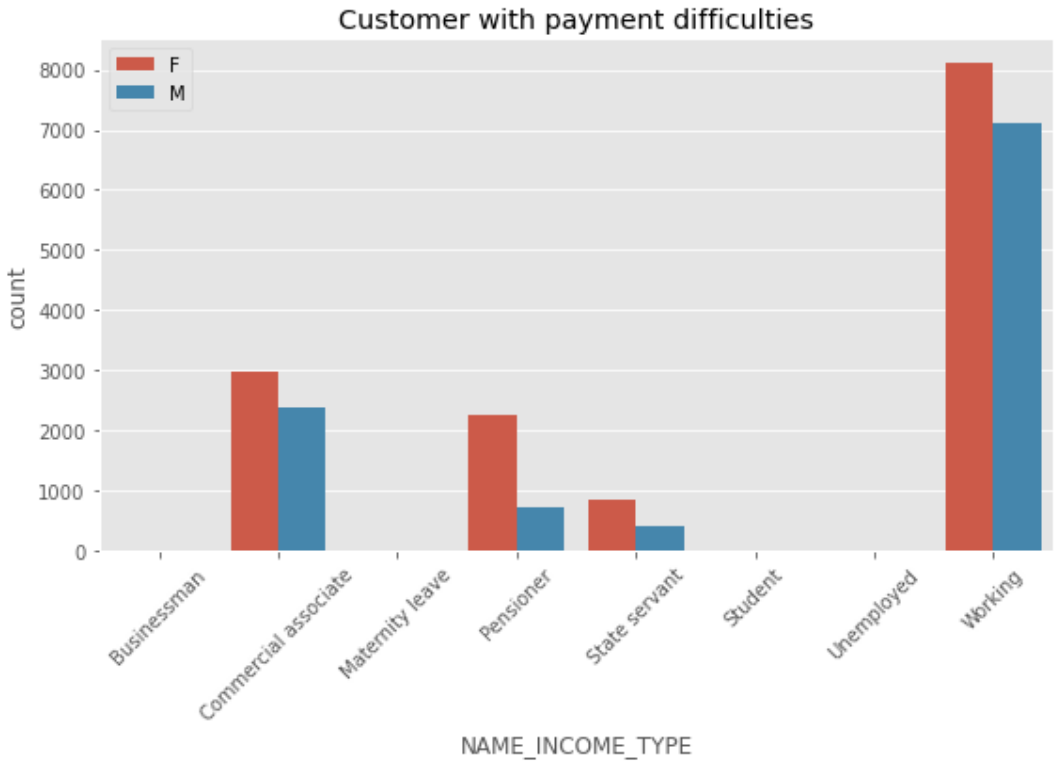




- We can clearly see here that there are a greater number of the female applicants
- Females with income less than 2,50,000/- are most likely to default

NAME_INCOME_TYPE & CODE_GENDER

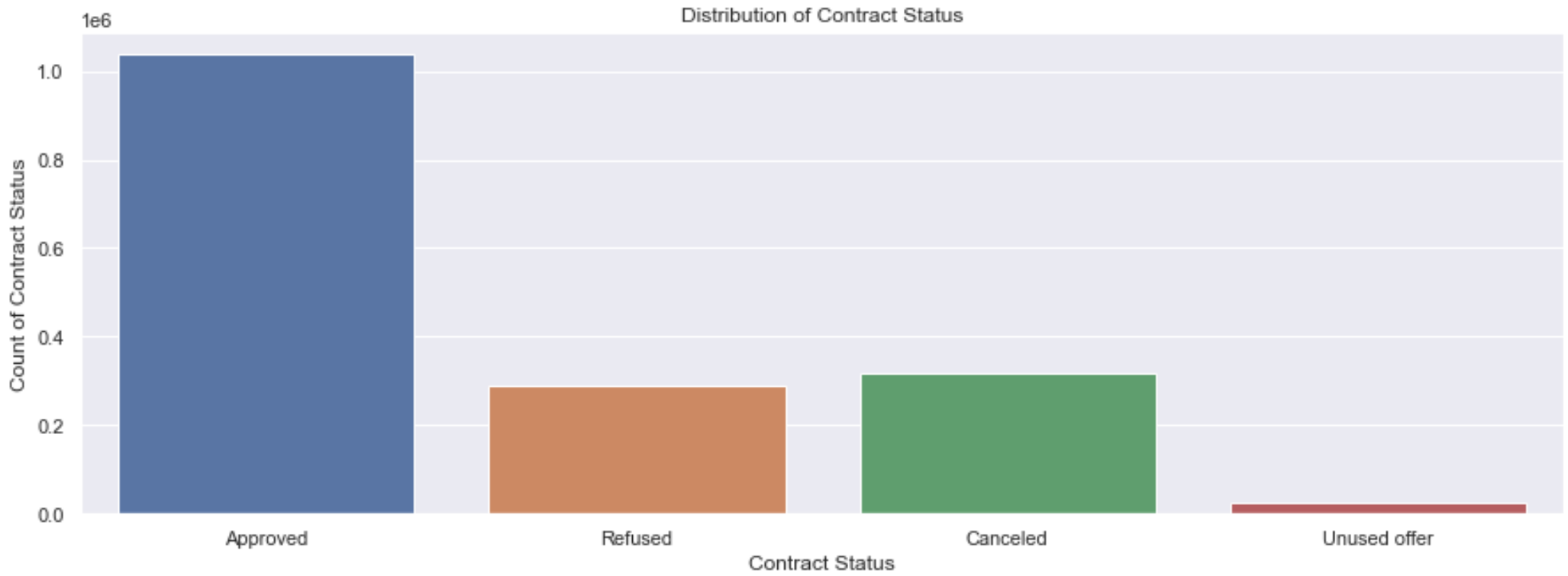




Businessman, students, unemployed applicants are not there in the defaulters (reasons could be...for students, they are still not employed yet. They might not take any loan...same with the unemployed applicants)

PREVIOUS APPLICATION DATA

TARGET VARIABLE (NAME_CONTRACT_STATUS)

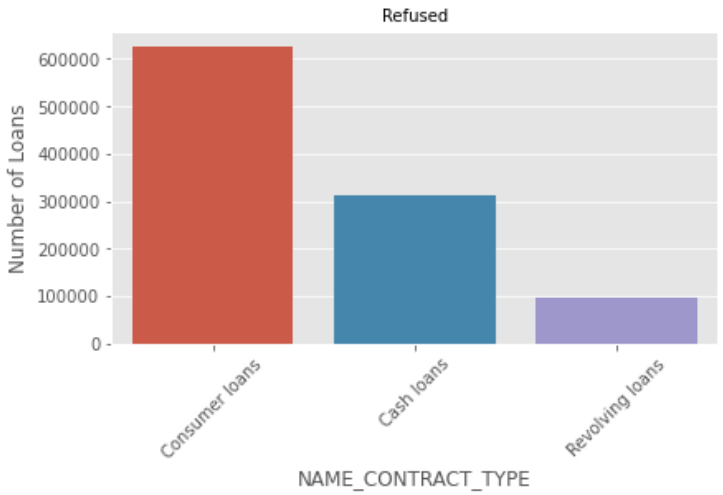


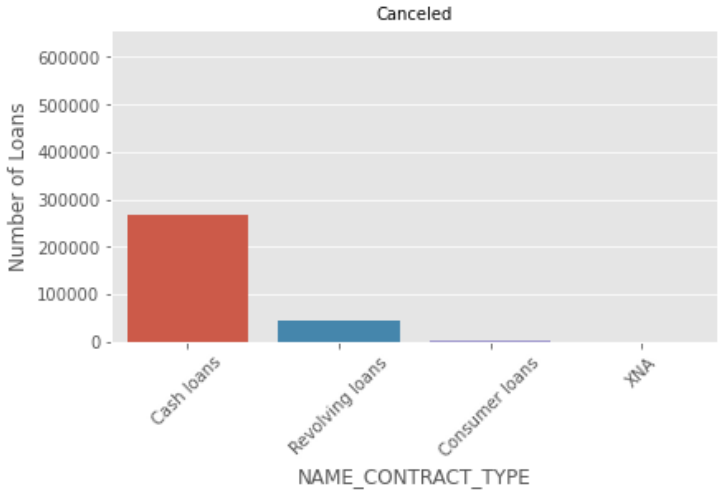
If we observe here, most of the applications were approved, and very less are unused

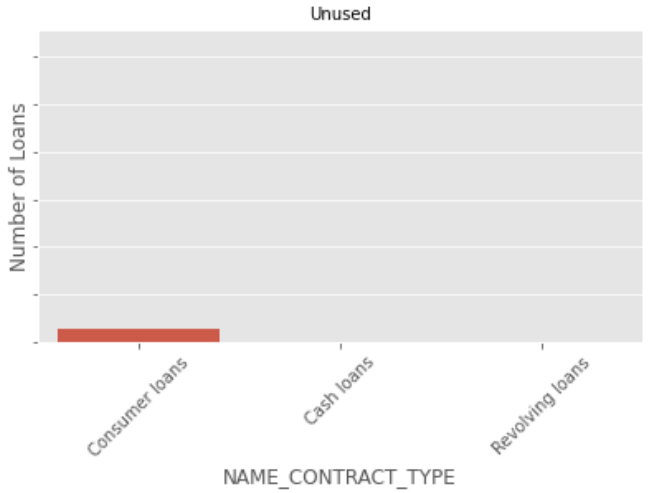
UNIVARIATE ANALYSIS

CATEGORICAL COLUMNS

NAME_CONTRACT_TYPE

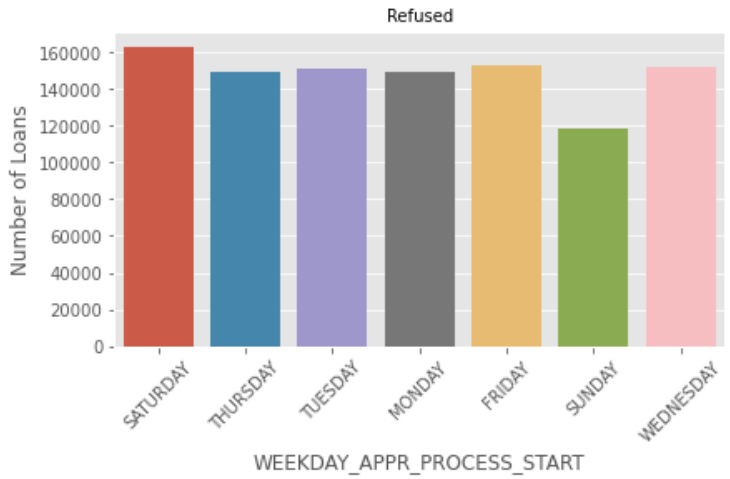


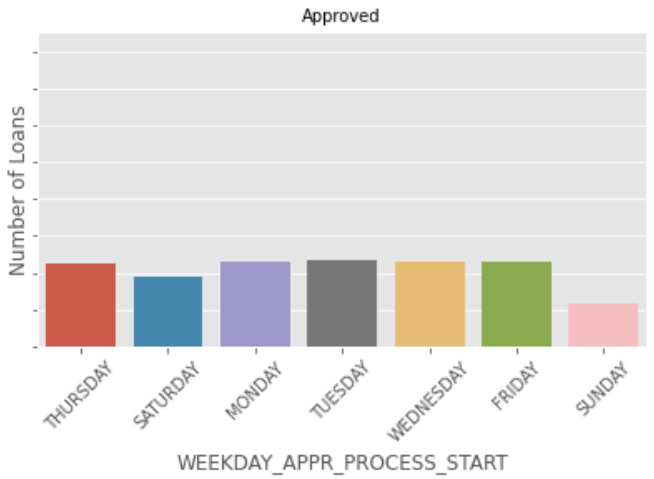


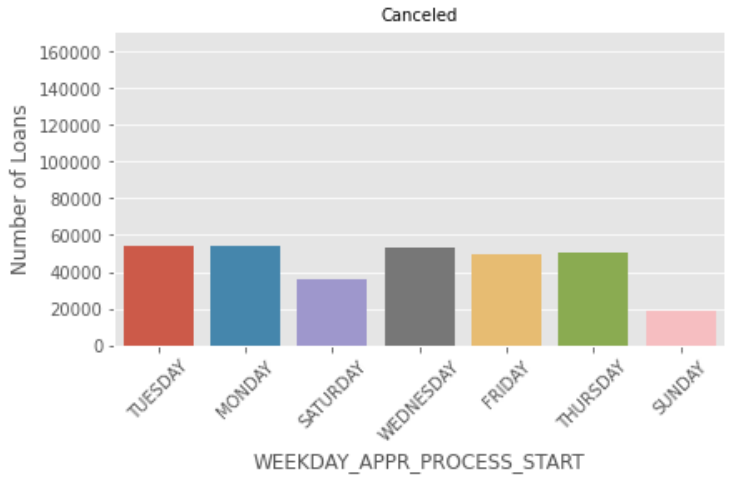


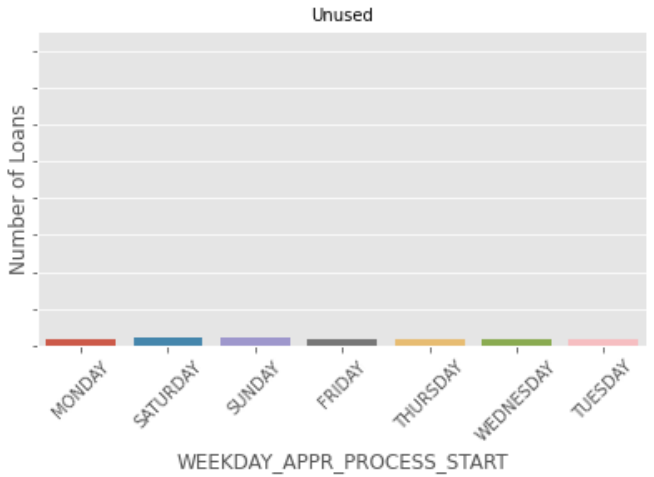
Here we can see that the Revolving loan is much more acceptable as compare to the cash and consumer loans.

WEEKDAY_APPR_PROCESS_START



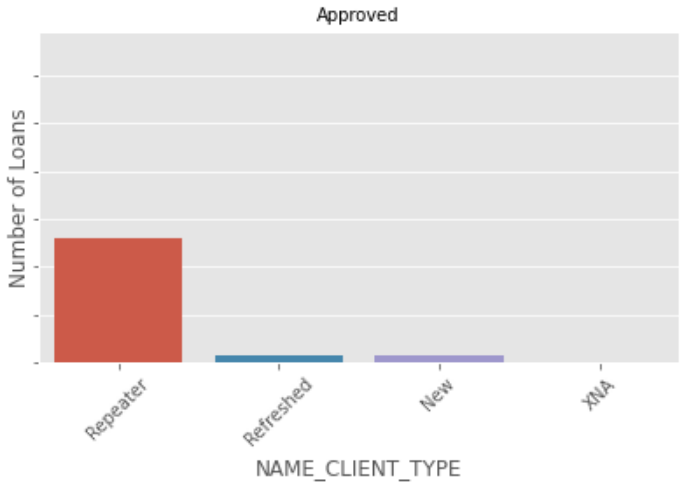


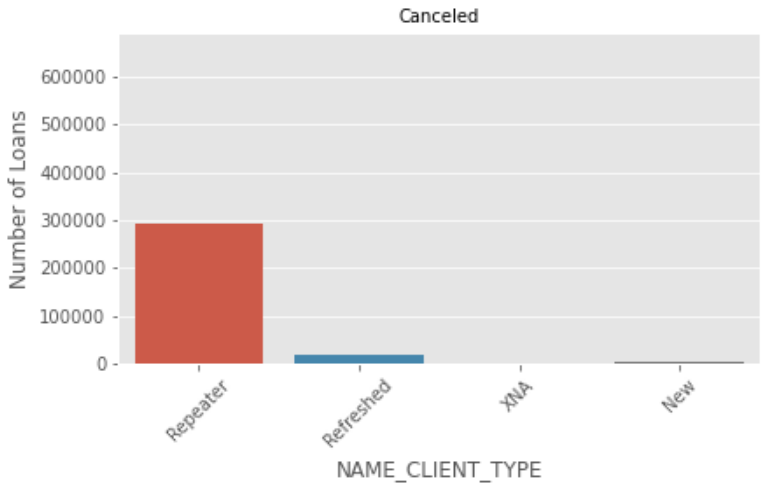


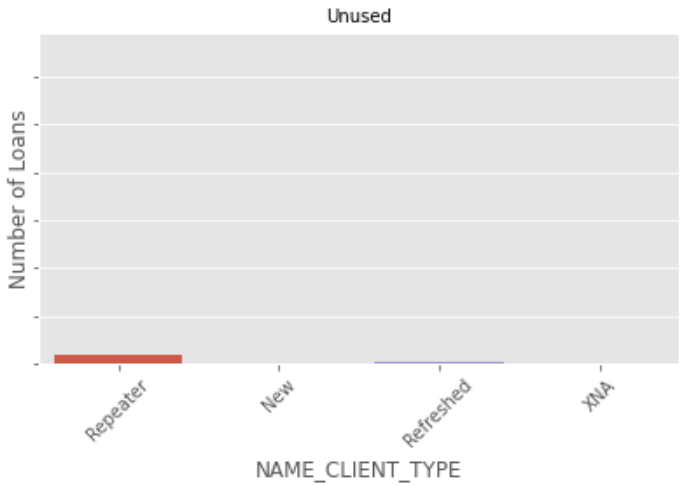


- We observe that there are a smaller number of applicants that come in the weekends.
- Most of the applications were refused on Saturday..

NAME_CLIENT_TYPE

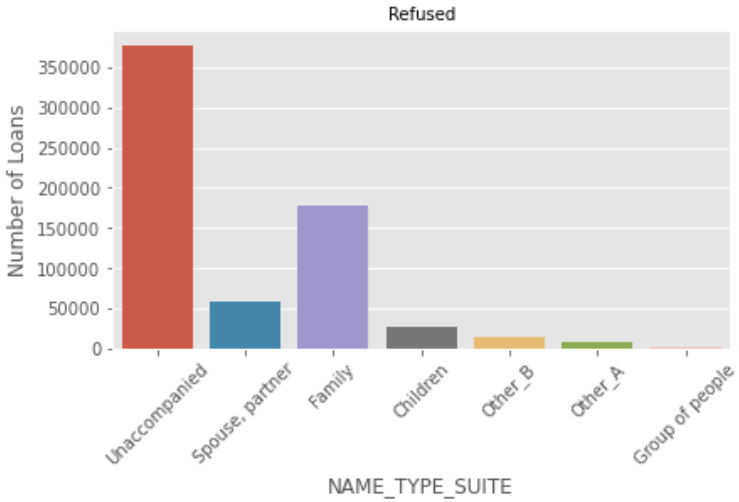


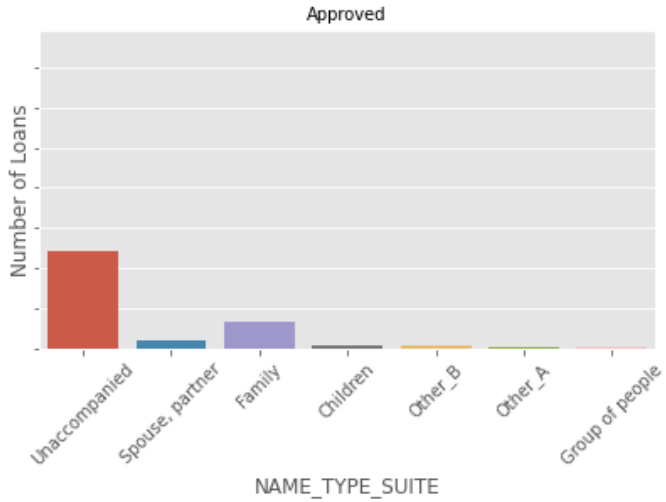


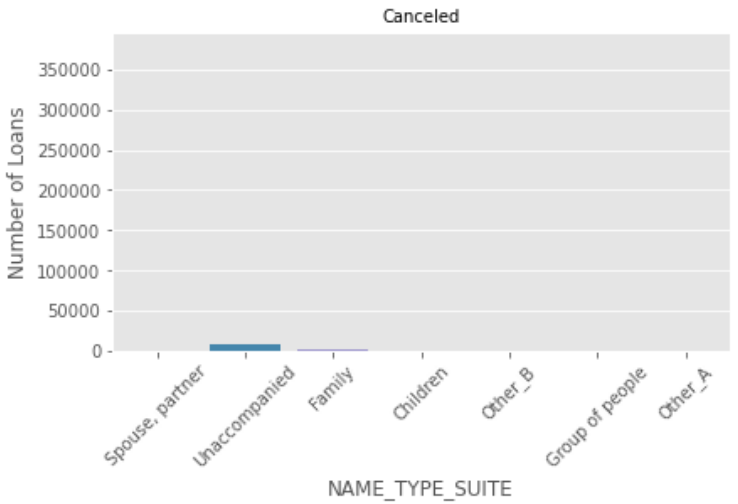


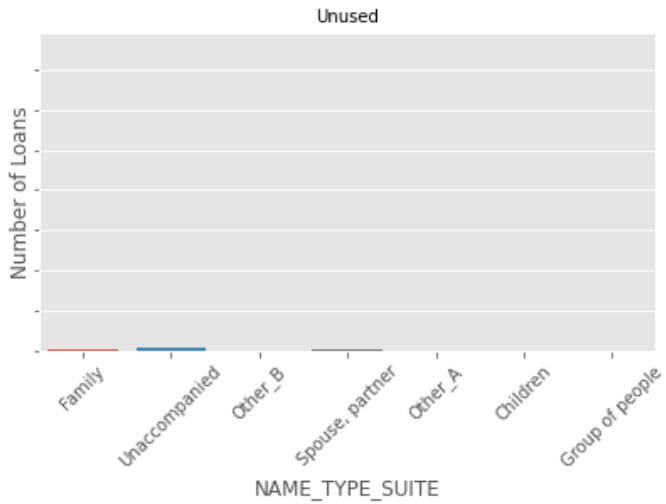
Here we can see that the Repeater is getting more Refused but also, we can see that the it also getting more approved and even that it is getting more canceled and more us used.

NAME_TYPE_SUITE



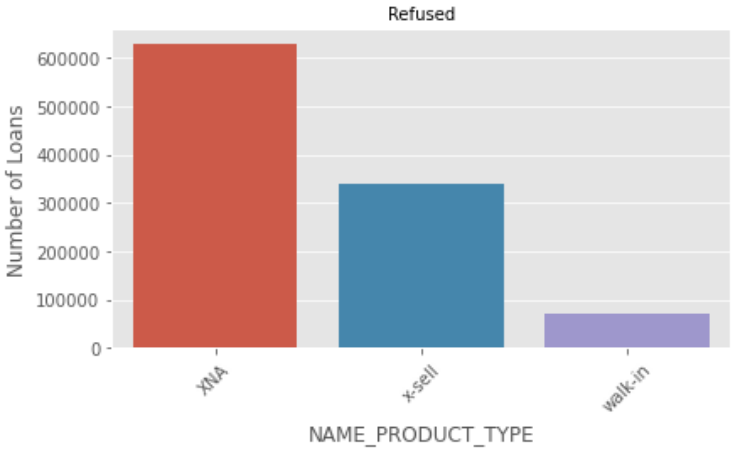


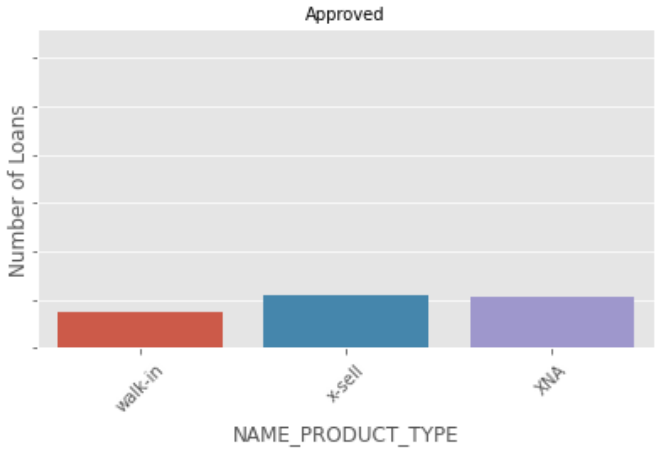


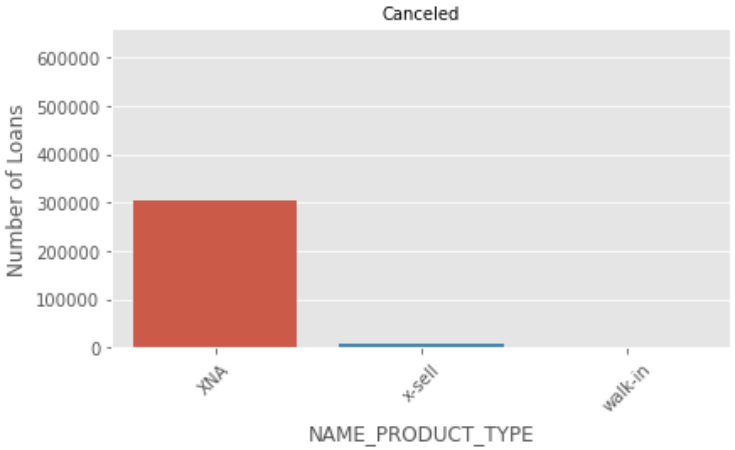


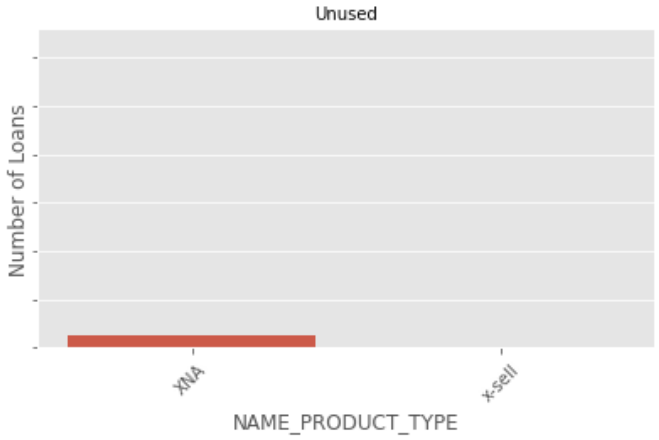
- Most of the loans were refused if the applicant has no company when applying.
- No cancelled loans, when the applicant accompanied with family or spouse.
- There are a greater number of refused loans when the applicants accompanied with family

NAME_PRODUCT_TYPE



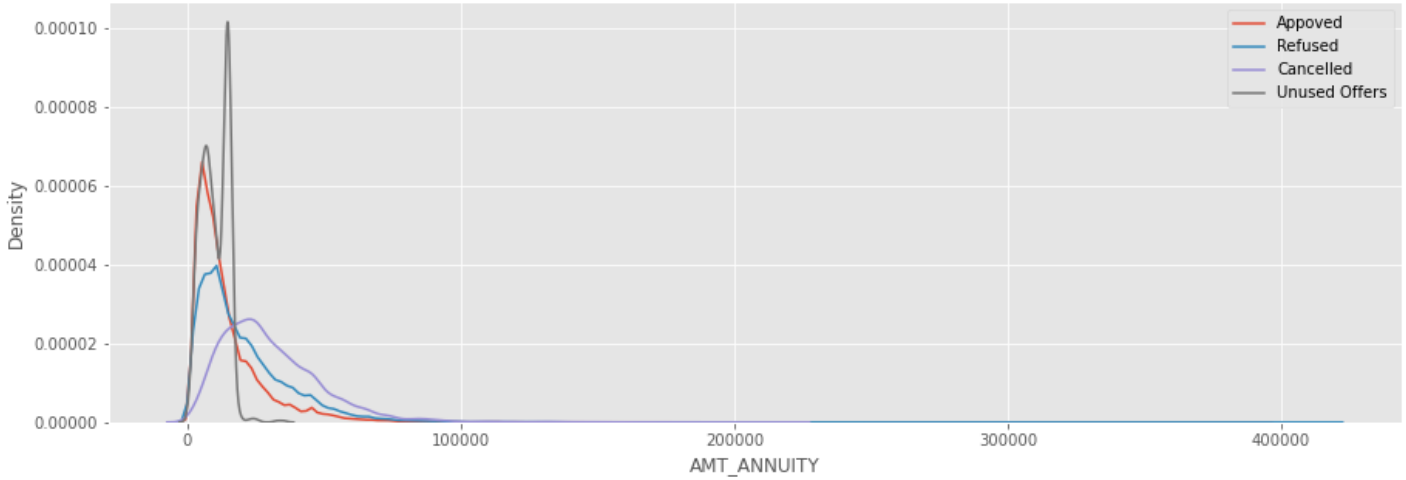


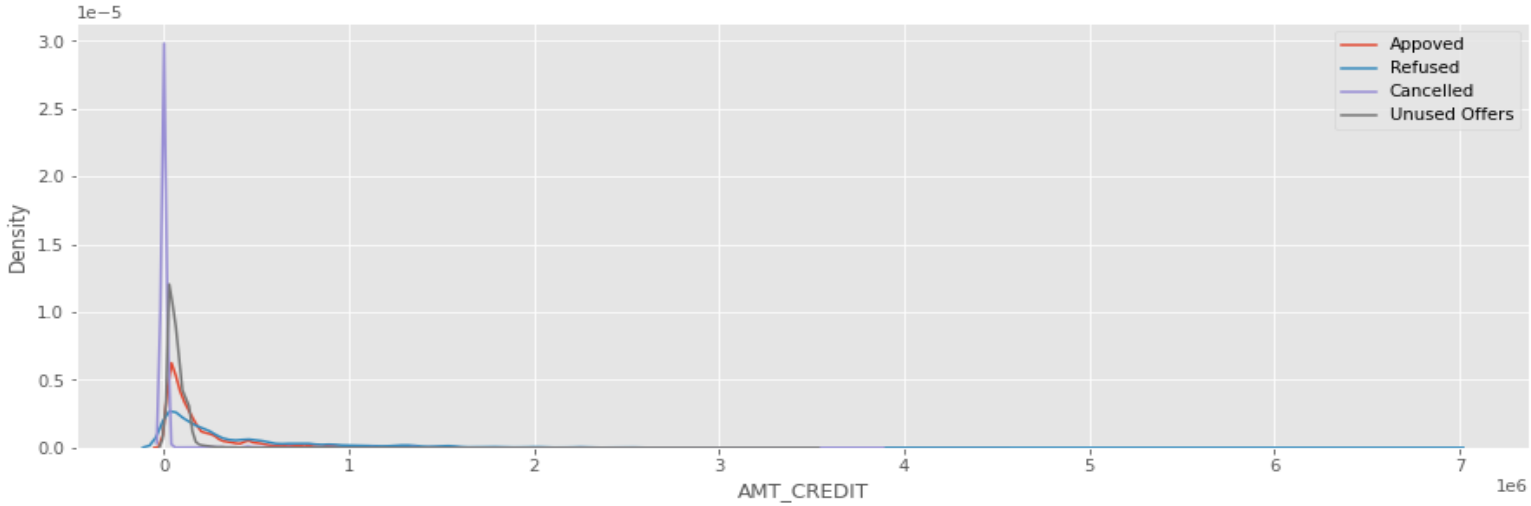




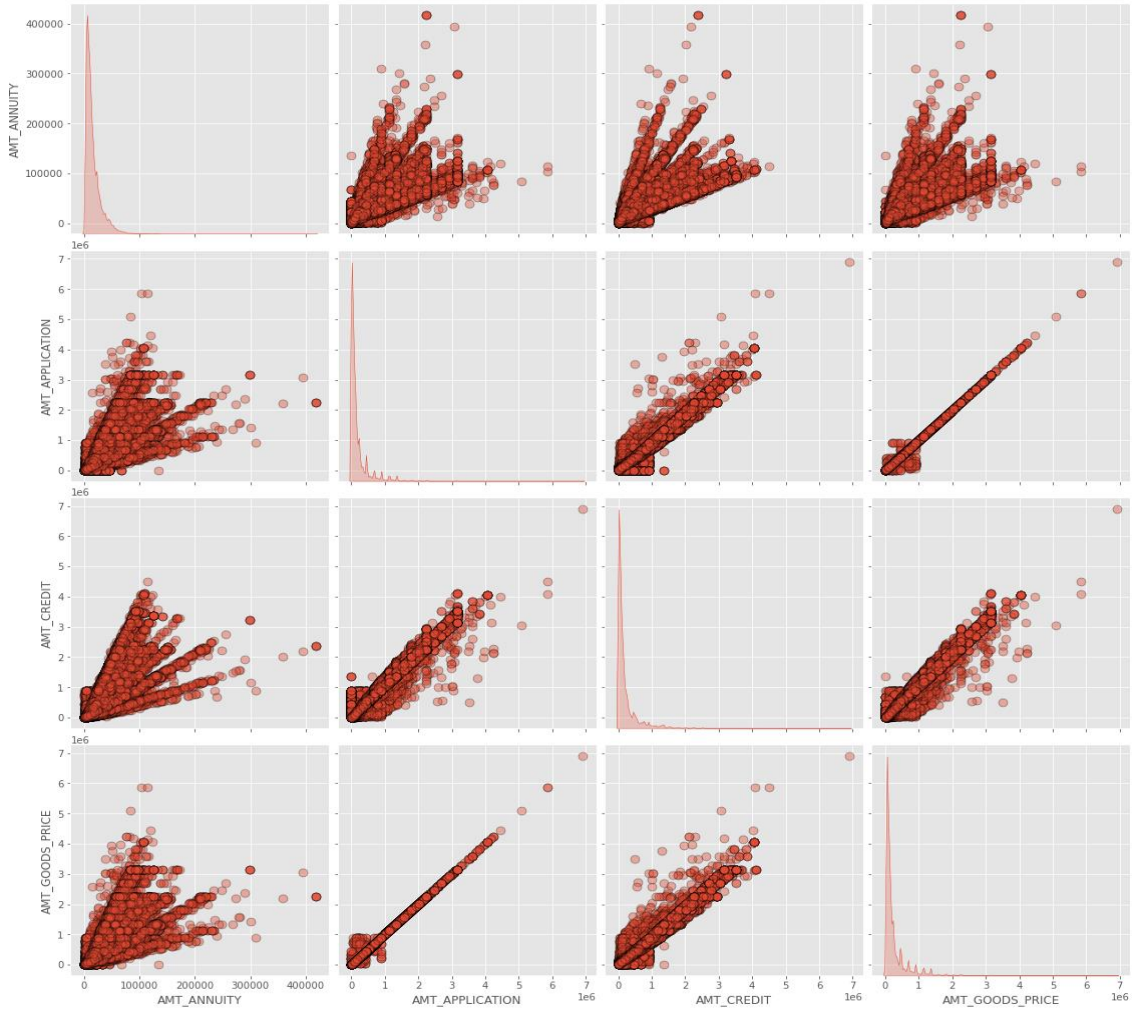
More number of refused loans if the product type is x-sell.

UNIVARIATE NUMERICAL ANALYSIS





BIVARIATE NUMERICAL ANALYSIS



- 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

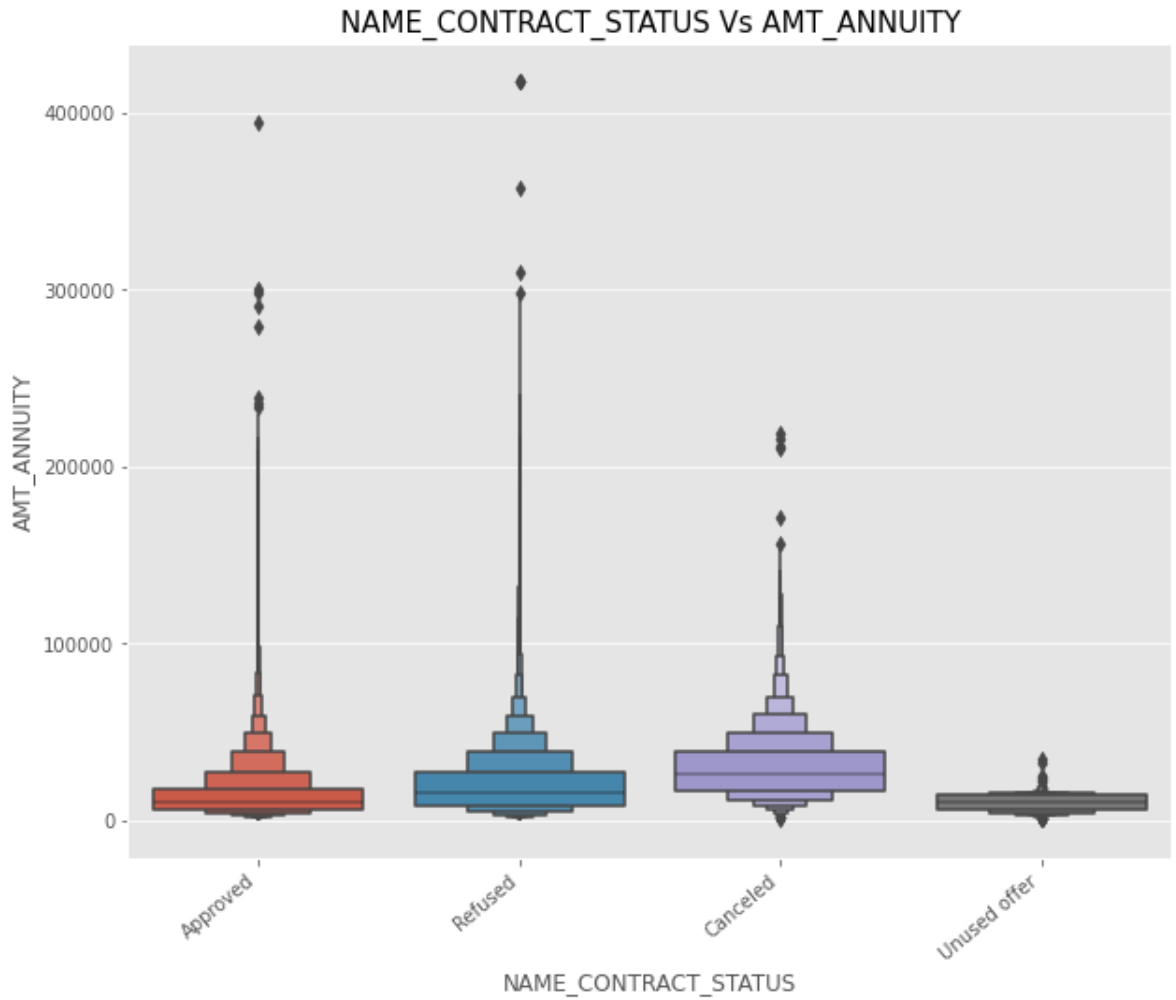
  (2)Final credit amount on the previous application that was approved by the bank

  (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 the goods price of good that client asked for on the previous application.

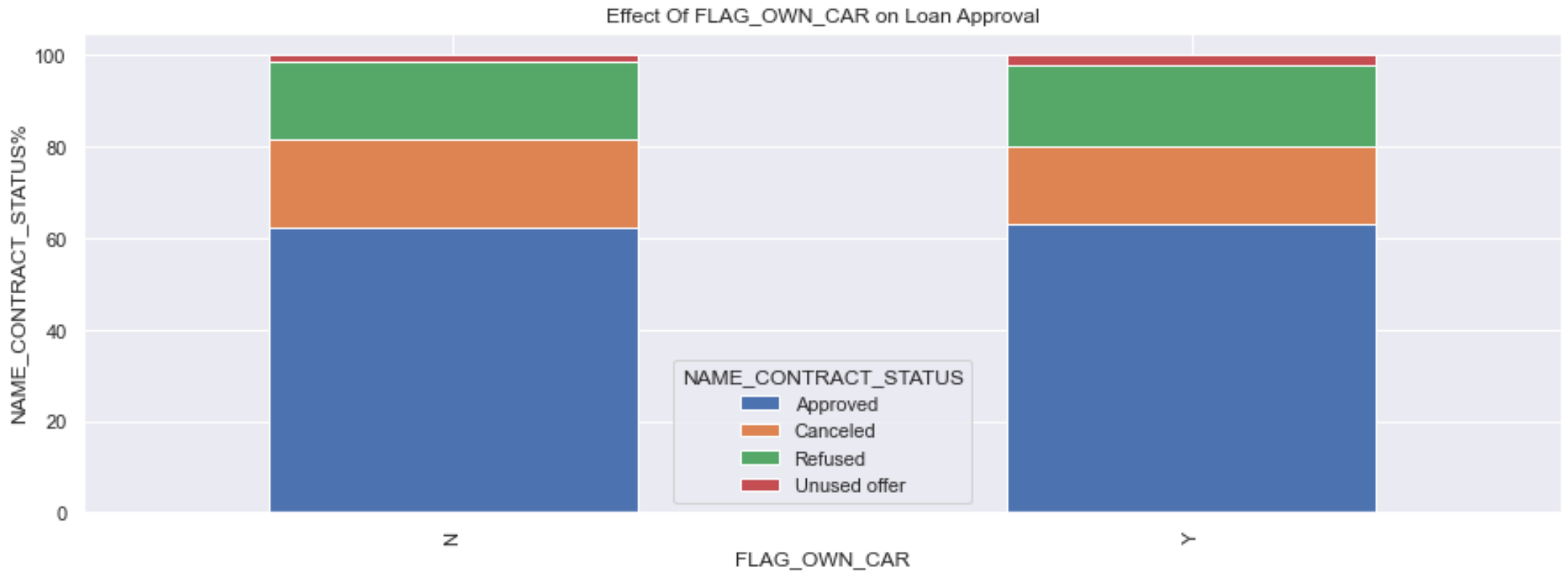
BIVARIATE CATEGORICAL AND NUMERICAL ANALYSIS



- From the 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.

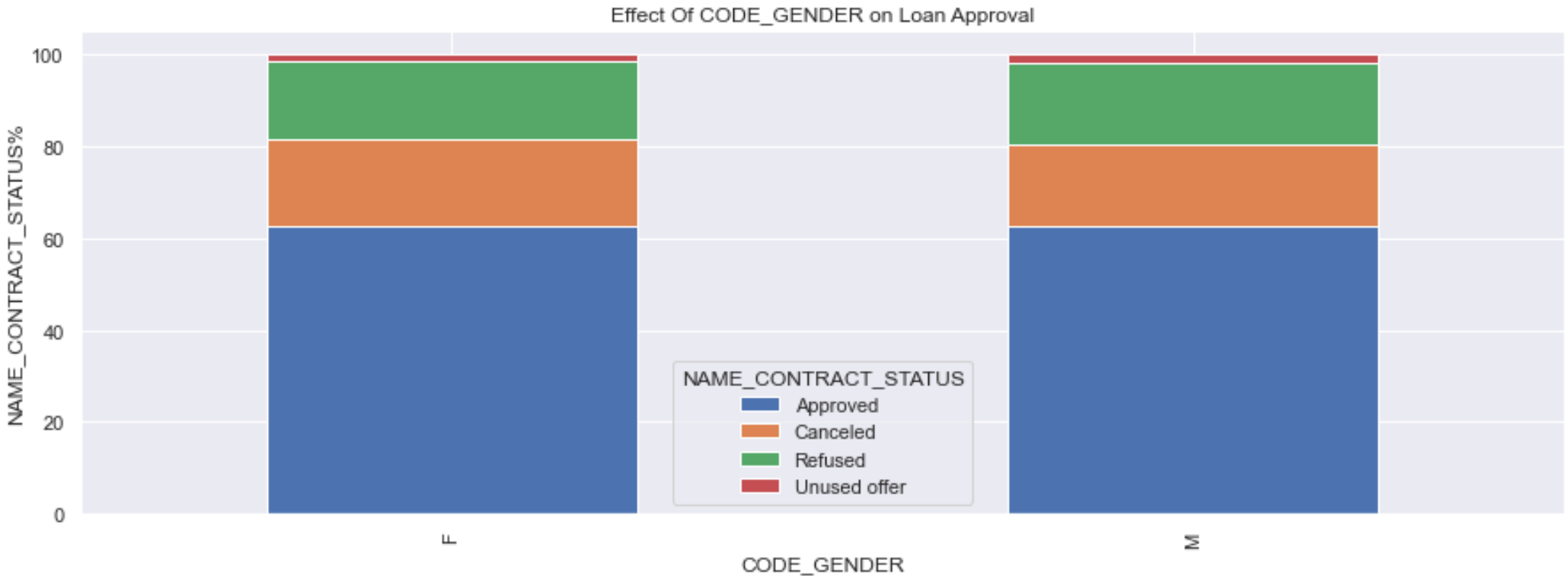
MERGING BOTH APPLICATION DATA AND PREVIOUS APPLICATION DATA

UNIVARIATE ANALYSIS (FLAG_OWN_CAR)



- We see that car ownership doesn't have any effect on application approval or rejection. But we saw earlier that the people who has a car has lesser chances of default.
- The bank can add more weightage to car ownership while approving a loan amount

CODE_GENDER W.R.T TARGET VARIABLE



- We see that 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.

TARGET(APPLICATION DATA) VS NAME_CONTRACT_STATUS (TARGET FOR PREVIOUS APPLICATION DATA)

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

MAIN POINTS THAT SEEMS TO INFLUENCE THE RISK OF DEFAULT AND THE ACTIONS THAT THE BANK SHOULD TAKE:--

- Defaulters are very low if the income of the applicant is more than 5,00,000/-
- Top Correlation in both, defaulters and non-defaulters are similar.
- Increase in the age, the number of the defaulters are getting reduced (Reason could be, as age increases, people were getting employed or their income increases…so that they are able to pay back the loan amount).
- Women take more loans.
- Bank can consider better if the applicant is having car.
- Students, Businessman are not defaulted. So, bank can consider them.

THANK YOU