Introduction

This case study aims to give us an idea of applying EDA in a real business scenario. In this case study, we develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Objectives

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (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 utilize this knowledge for its portfolio and risk assessment.

Two data sets were provided for this case study, namely - Application Data and Previous Application Data.

Application Data is analyzed first.

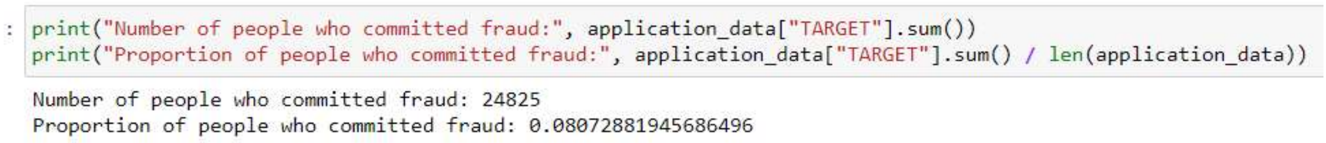
Steps performed for analysis of application data

1. Identification of important variables
    a. Target variable and feature variables
    b. Data types
    c. Categorical or Continuous variable
2. Handling missing values and outliers. Deleted columns which has more than 50% of null values. Identified values to be imputed for missing values and outliers.
   Categorical variable– Impute Mode of the column
   Continuous variable with no outliers- Impute Mean of the column

Continuous variable with outliers- Impute Median of the column

3. Checked the data imbalance and the imbalance ratio.
4. Derived new variables by binning on Continuous variables
5. Performed univariate, bivariate, and multivariate analysis.
6. Visualized data with respect to target variable

Proportion of people who committed fraud



```
: print("Number of people who committed fraud:", application_data["TARGET"].sum())
  print("Proportion of people who committed fraud:", application_data["TARGET"].sum() / len(application_data))

  Number of people who committed fraud: 24825
  Proportion of people who committed fraud: 0.08072881945686496
```

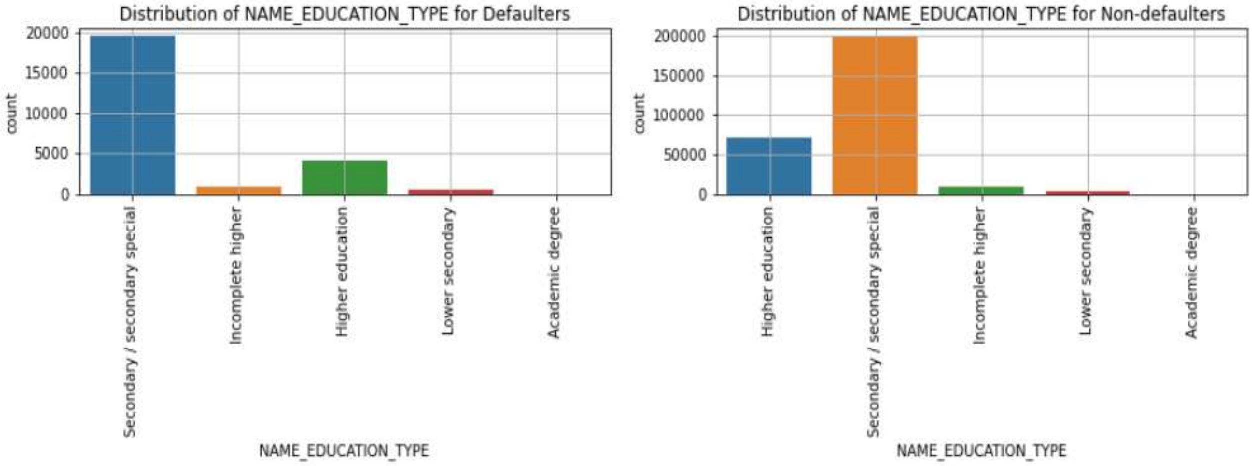
The dataset is centered around variable **TARGET**.

Values that Target variable takes-

1 - when client has payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample,

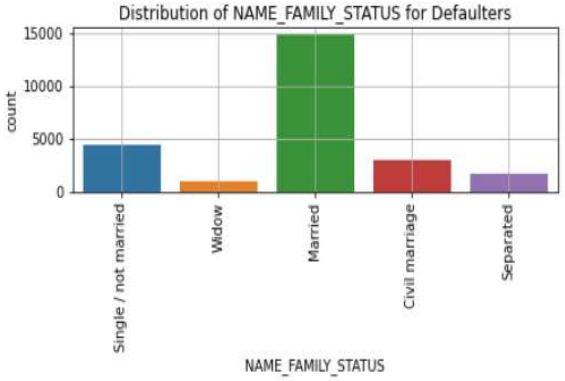
0 - all other cases

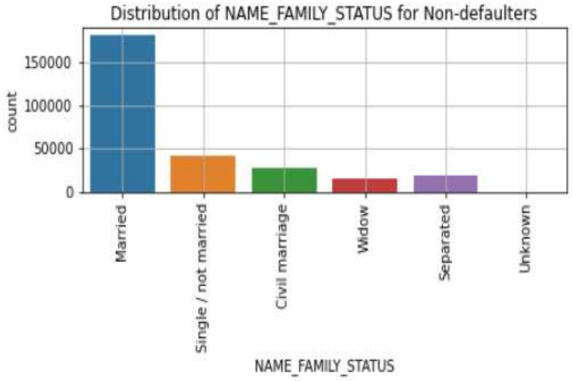
Univariate analysis of Categorical Variables



Observations:

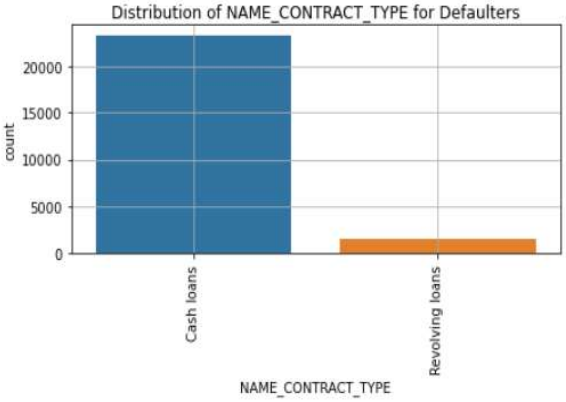
- Among both defaulters and non-defaulters, distribution of secondary/special educated clients are the highest and Academic degree educated clients are the least
- Distribution of Higher education is lesser amongst defaulters. Higher the education level, lower the default rate. Higher is the education level, the more they earn and hence easier to pay off the loan.

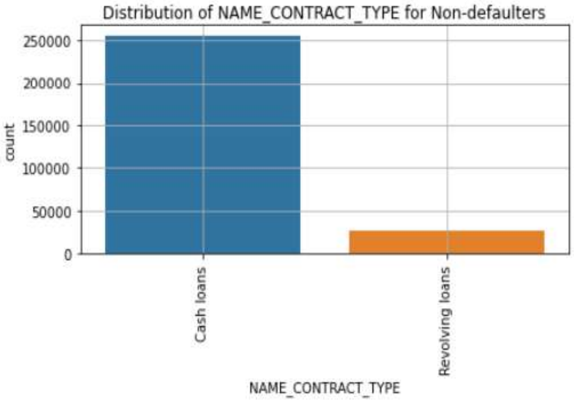




Observations:

- Married applicants take the maximum number of loans. Being married is not a causation for defaulting the loan as seen from the graph
- Single/not married are higher in numbers in defaulter's category.
- There is decrease in the percentage of married and widowed applicants in the defaulter category



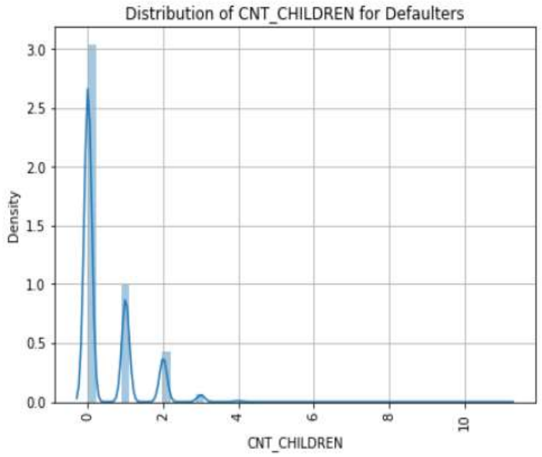


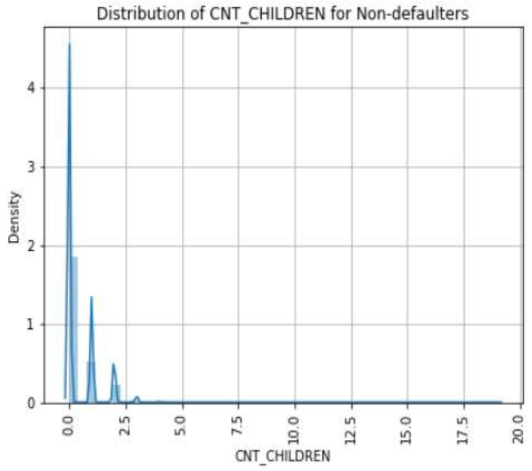
Observations:

- Revolving loans are lesser in the defaulter category. Hence, revolving loans are comparatively safer

- Cash loans are preferred over revolving loans

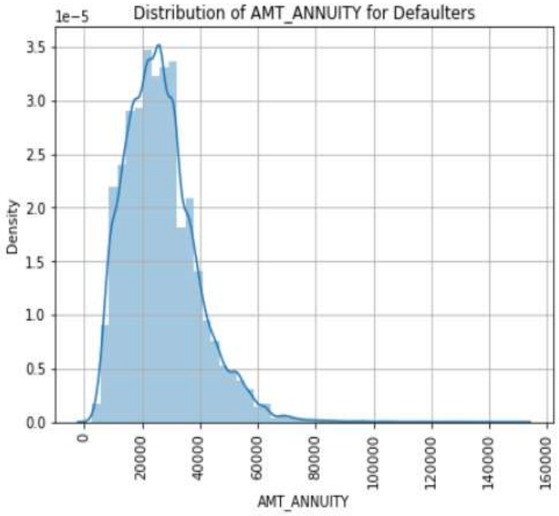
Univariate Analysis Continuous Numerical Variables

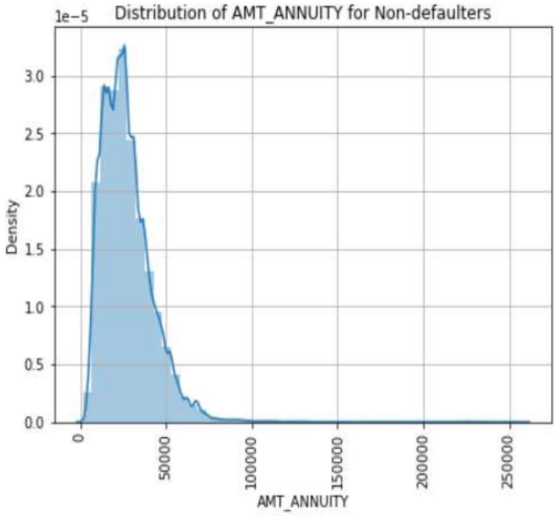

Distribution of CNT_CHILDREN for Defaulters


Distribution of CNT_CHILDREN for Non-defaulters

Observations:

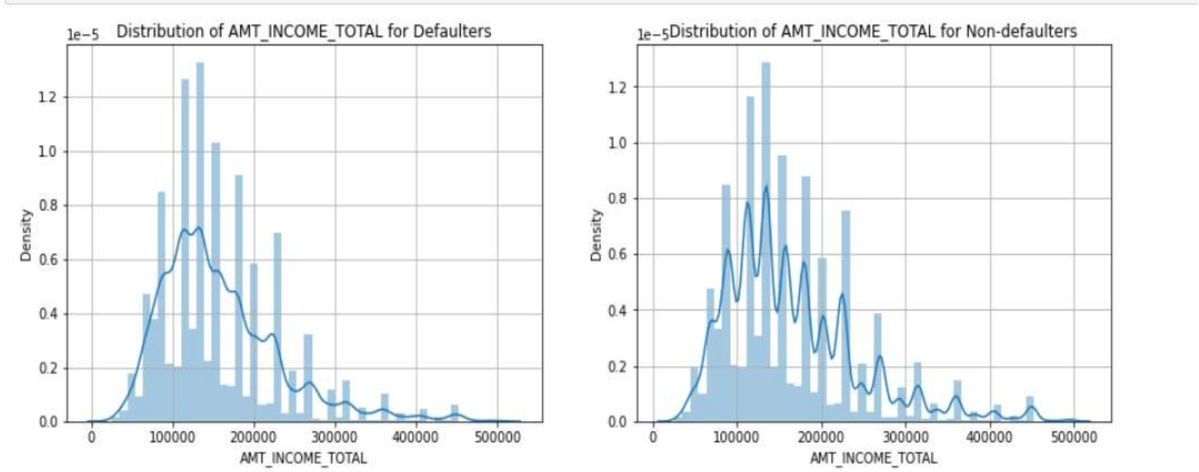
- People with 0 children are much more among non-defaulters which shows that people who are applying for loan have less people dependent financially on them which will lessen the loan payment difficulties.


Distribution of AMT_ANNUITY for Defaulters


Distribution of AMT_ANNUITY for Non-defaulters

Observations:

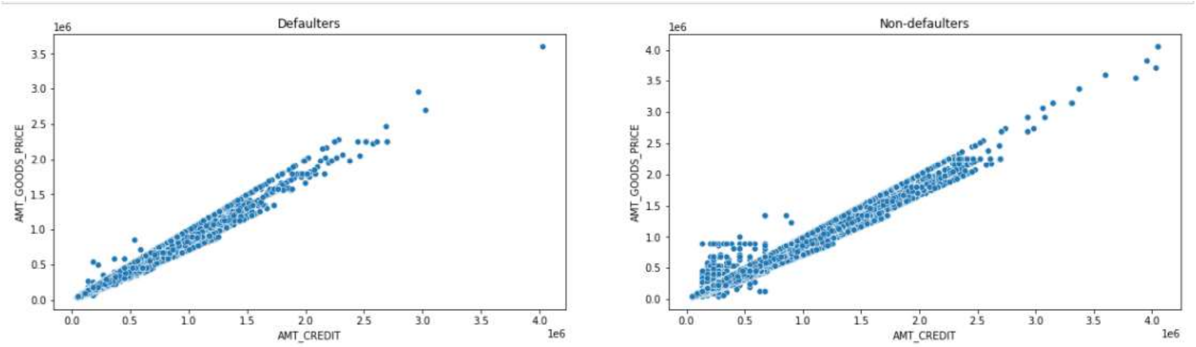
- Lower loan annuity, higher the number of loans
- Default rate is higher for lower annuity
- The KDE for AMT_ANNUITY almost resembles a normal distribution with a right skew



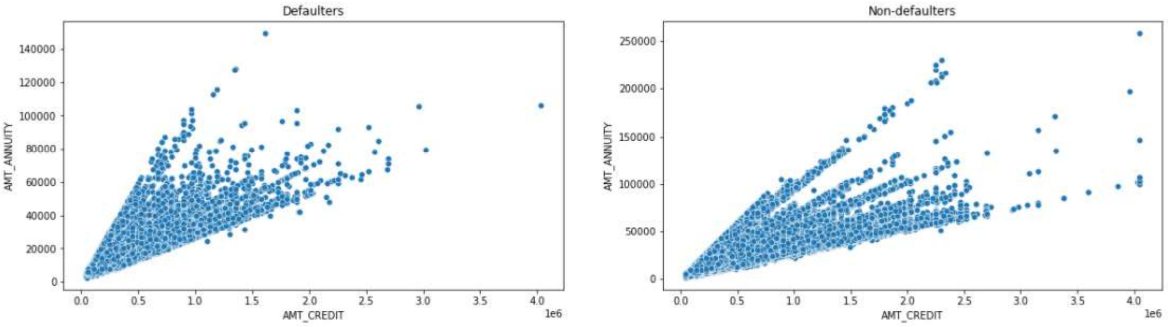
Observations:

- People with lower income have higher default rate
- The density of people earning between 1,00,000-2,00,000 are more likely to apply for the loans and pay them on time. There is a skew to the right in non-defaulter graph which also shows people with higher salaries present as well.

Bivariate Analysis of Numerical-Numerical Variables

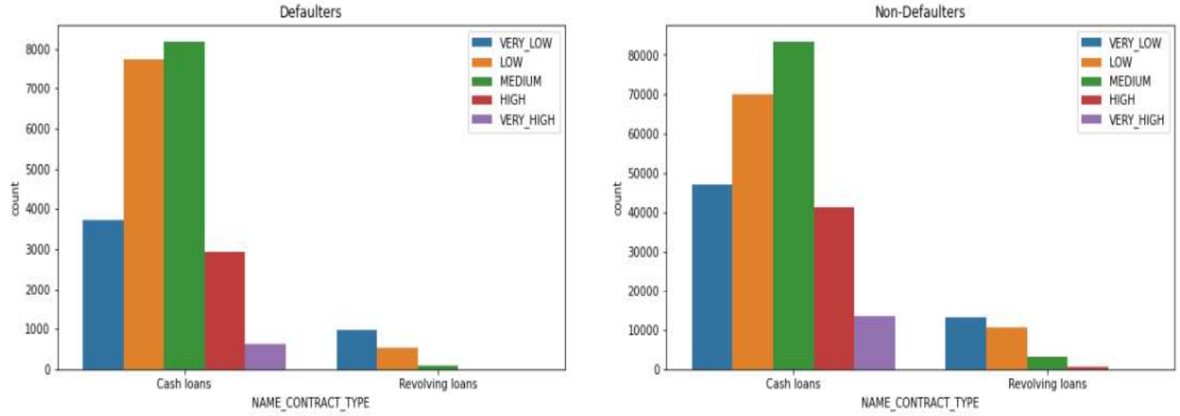


Observation: AMT_GOODS_PRICE and AMT_CREDIT is positively correlated. As goods price increases, the credit also increases.



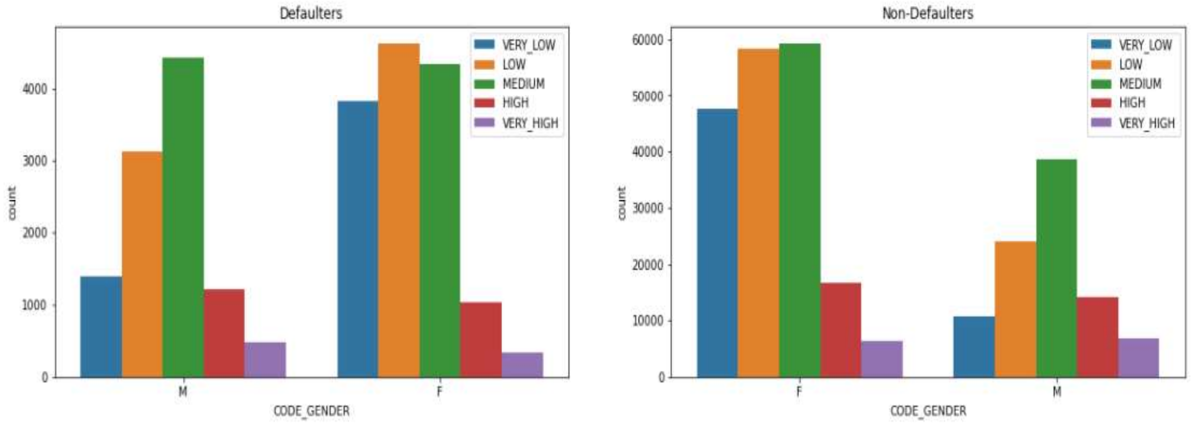
Observation: Non-defaulters take more credit for the same annuity as compared to defaulters

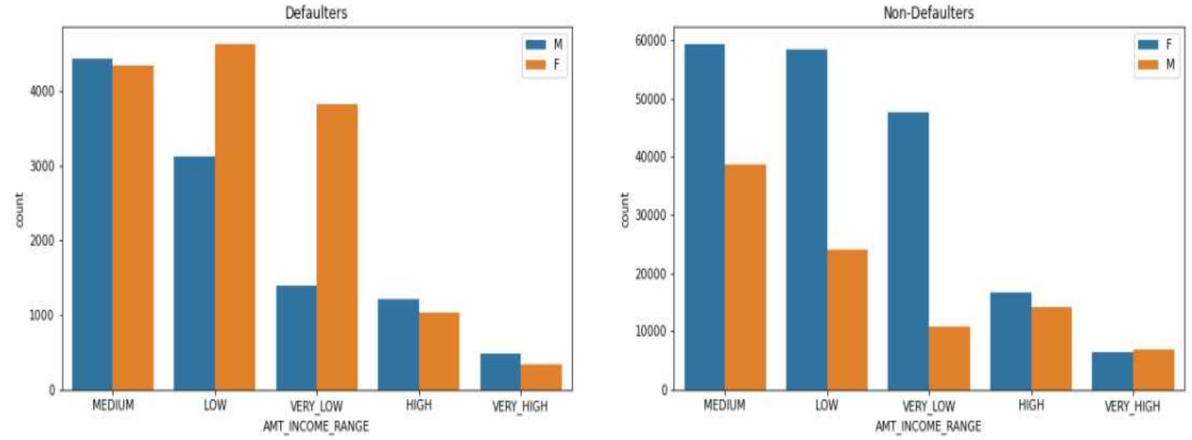
Bivariate Analysis of Categorical-Categorical Variables



Observations:

- Revolving loans are not as popular as cash loans
- Distribution of defaulters is higher for low final credit amount on the previous application.

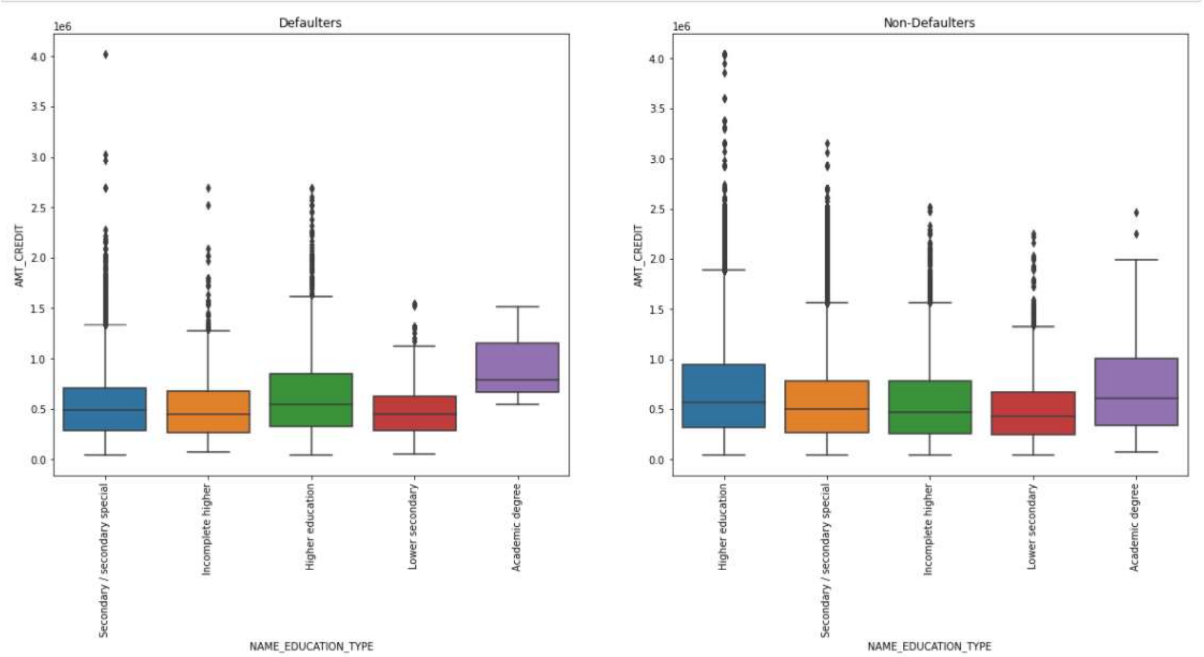




Observations:

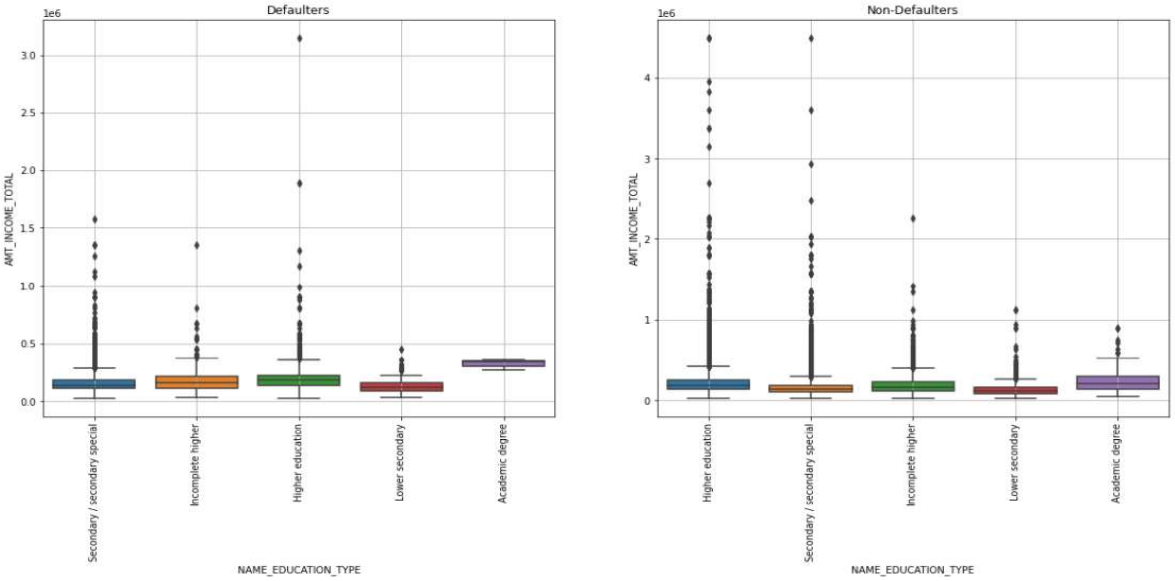
- Female defaulters are higher in number than male defaulters.
- Females with low and very low income default the most
- Females’ non-defaulters are also higher in number than male non-defaulters.
- Female defaulters are high in non-defaulter except in very-high income range

Numerical- Categorical bivariate analysis



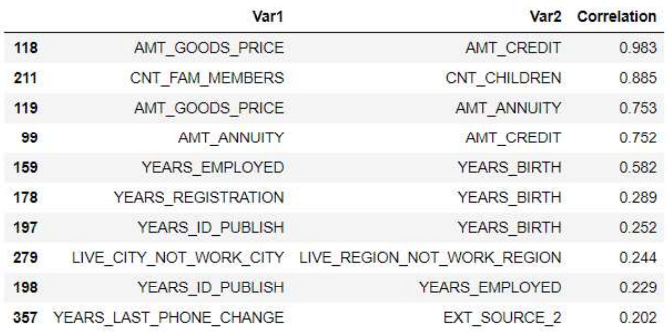
Observations:

- Range of applicants with Secondary/secondary Special and that of Academic degree are different for defaulter and non-defaulters. The distribution is similar for the rest of the education types for defaulter and non-defaulters
- Higher number of outliers in non-defaulter's data frame than defaulters.

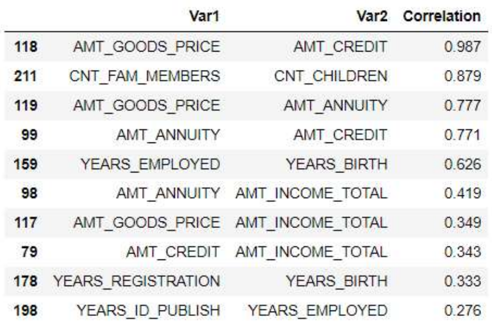


Observation: Outliers are higher in numbers in non-defaulter data frame

Correlation between Defaulters variables

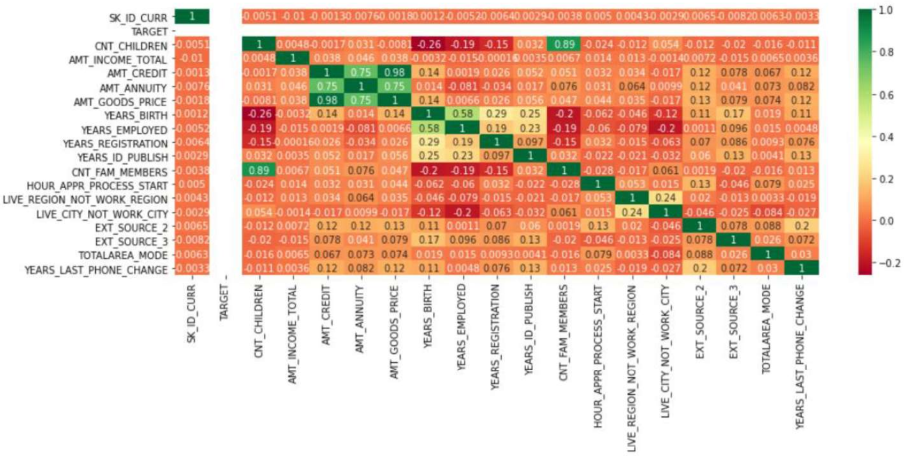


	Var1	Var2	Correlation
118	AMT_GOODS_PRICE	AMT_CREDIT	0.983
211	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885
119	AMT_GOODS_PRICE	AMT_ANNUITY	0.753
99	AMT_ANNUITY	AMT_CREDIT	0.752
159	YEARS_EMPLOYED	YEARS_BIRTH	0.582
178	YEARS_REGISTRATION	YEARS_BIRTH	0.289
197	YEARS_ID_PUBLISH	YEARS_BIRTH	0.252
279	LIVE_CITY_NOT_WORK_CITY	LIVE_REGION_NOT_WORK_REGION	0.244
198	YEARS_ID_PUBLISH	YEARS_EMPLOYED	0.229
357	YEARS_LAST_PHONE_CHANGE	EXT_SOURCE_2	0.202

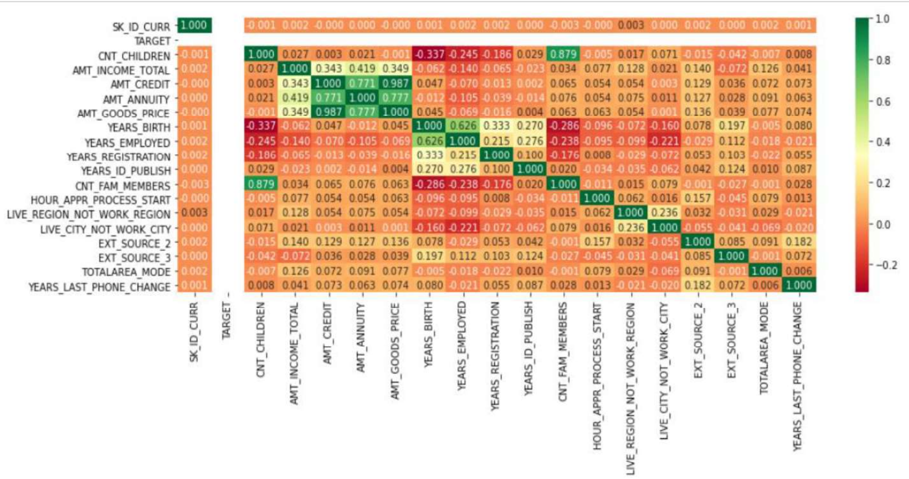


	Var1	Var2	Correlation
118	AMT_GOODS_PRICE	AMT_CREDIT	0.987
211	CNT_FAM_MEMBERS	CNT_CHILDREN	0.879
119	AMT_GOODS_PRICE	AMT_ANNUITY	0.777
99	AMT_ANNUITY	AMT_CREDIT	0.771
159	YEARS_EMPLOYED	YEARS_BIRTH	0.626
98	AMT_ANNUITY	AMT_INCOME_TOTAL	0.419
117	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349
79	AMT_CREDIT	AMT_INCOME_TOTAL	0.343
178	YEARS_REGISTRATION	YEARS_BIRTH	0.333
198	YEARS_ID_PUBLISH	YEARS_EMPLOYED	0.276

Defaulters heat map

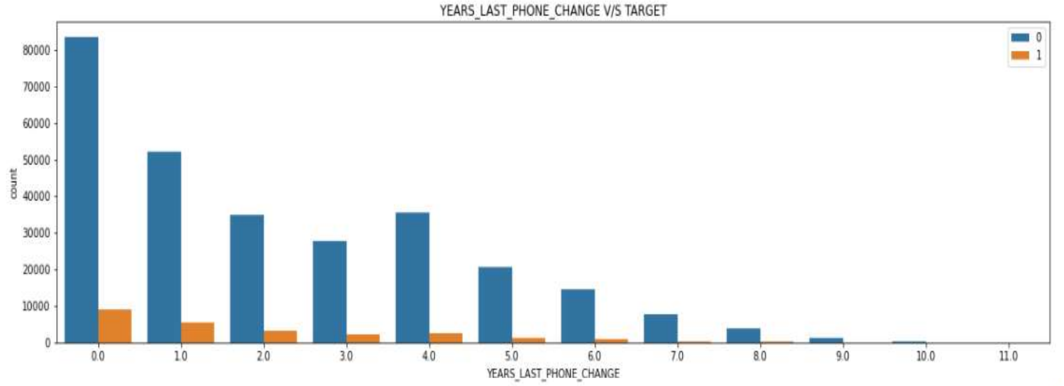


Non-Defaulters heat map

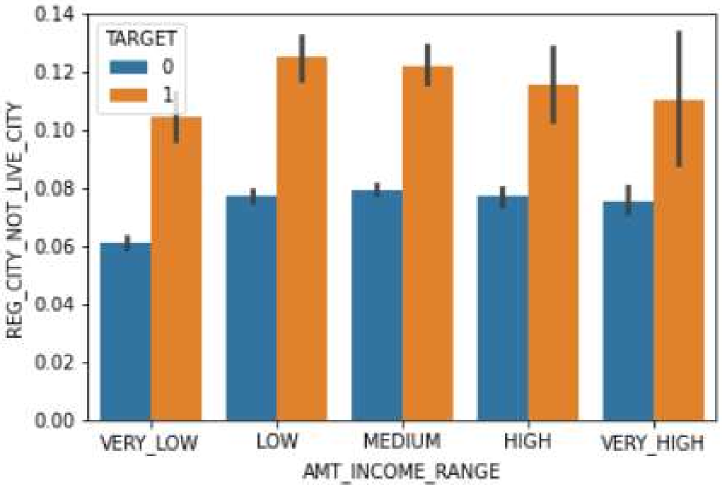


Observation: Top 5 correlations are similar for defaulters and non-defaulters

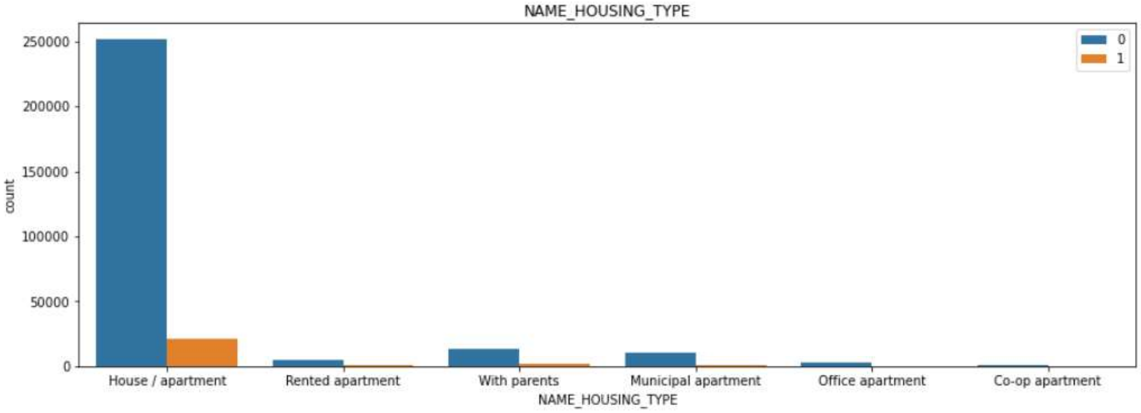
Analysis with respect to target variable



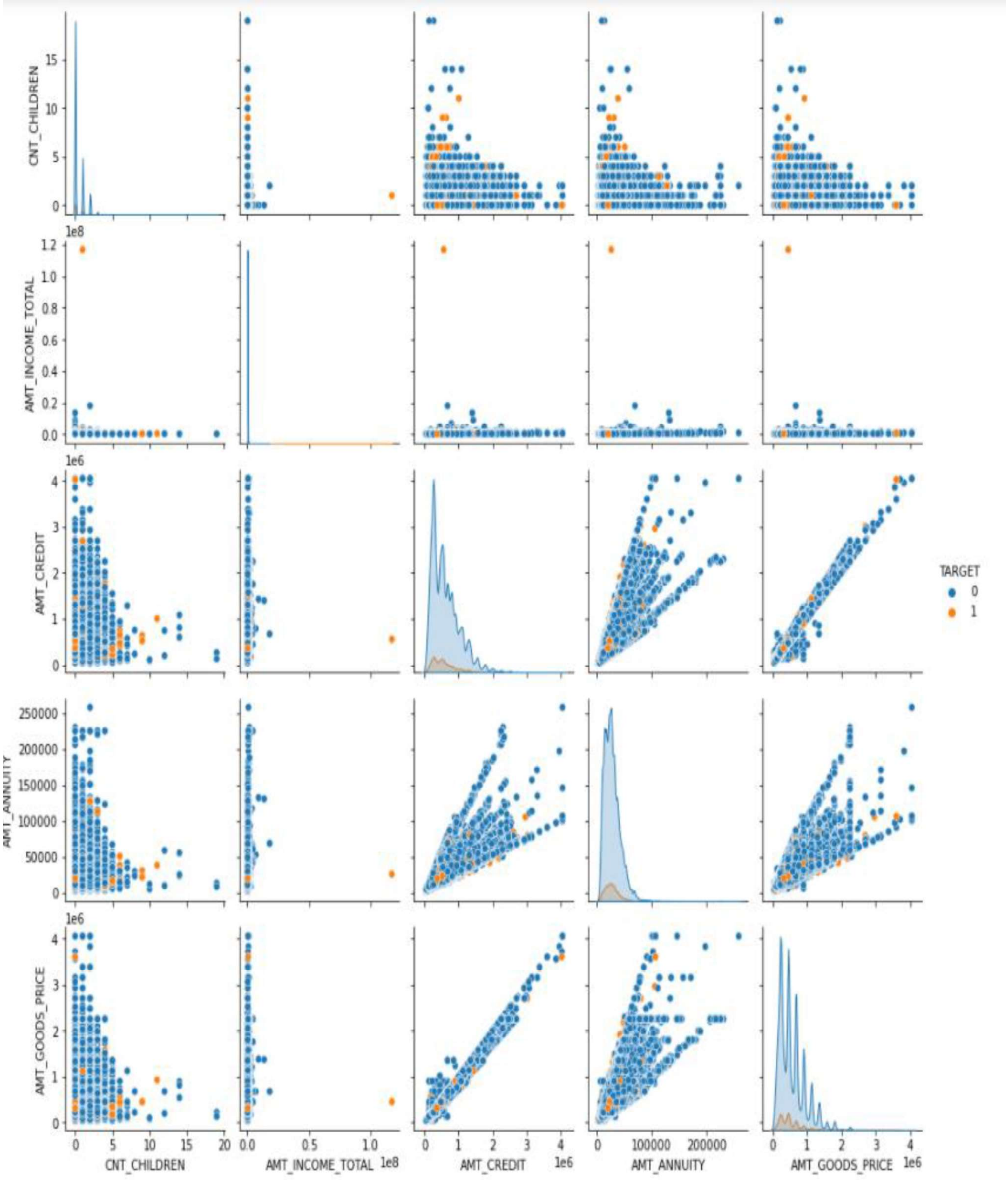
Observation: Non-defaulters retain their phone numbers for a longer time.



Observation: When applicants' permanent address does not match contact address at city level, there is significantly higher default rate when compared to region level.



Observation: Applicants who own a house/apartment have high default rate, but also repay the loans then highest number of times.

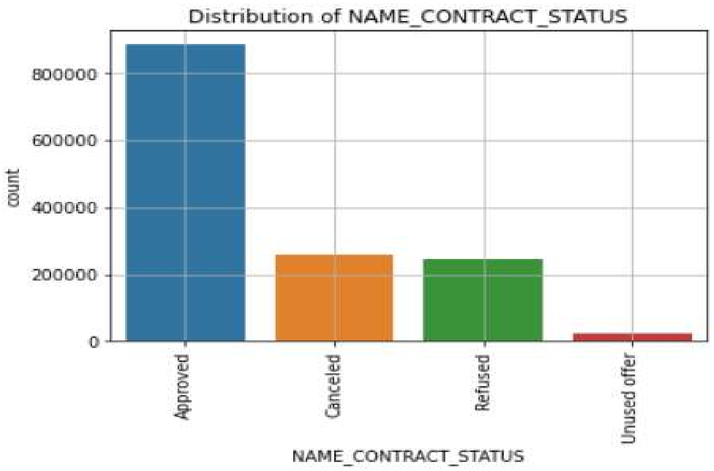


Observations:

- With increasing count of children, the applicant starts facing payment problem irrespective of income, credit amount, annuity and goods price with an exception if the salary of applicant is very high and age is relatively high.
- 'AMT_CREDIT and AMT_ANNUITY' , 'AMT_CREDIT and AMT_GOODS_PRICE' , 'AMT_ANNUITY AND AMT_GOODS_PRICE' have a rising relation if not completely linear relation with few exceptions.
- 'AMT_GOODS_PRICE and AMT_CREDIT' has an increasing relationship.
- CNT_FAM_MEMBERS is not related to the other variables.

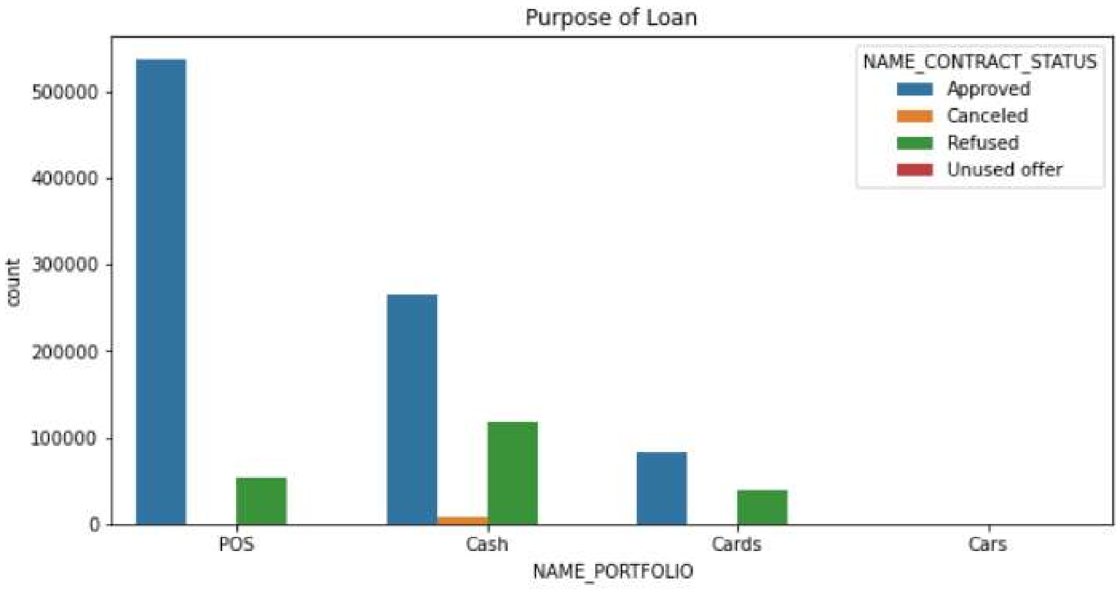
Merging application data and previous application data

Left join of application data and previous application data is performed to retain all the rows in application data. By left join, we get historical application data for each applicant.



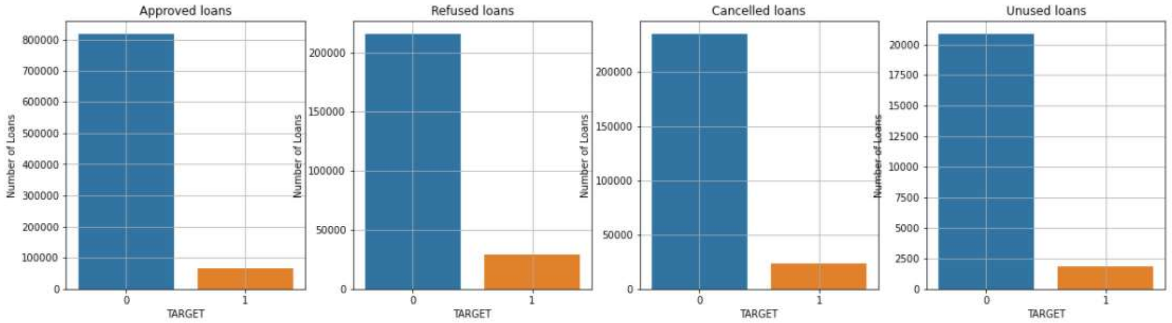
Observation: 61% of the previous loan is approved. 17-18% of the previous loans are either cancelled or refused.

Analysis with respect to previous loan status variable

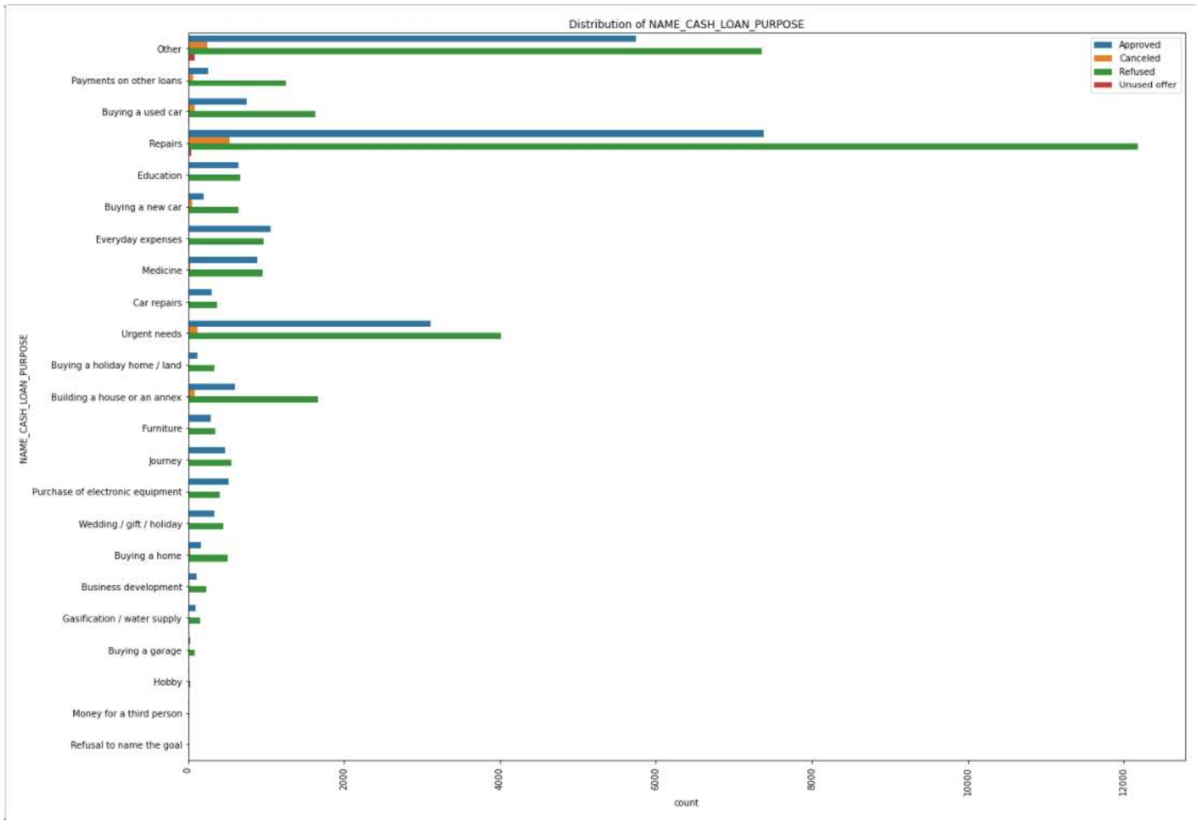


Observations:

- We observe unused offers when loans are taken for POS and highest number of loans are taken for POS
- Cash loans are refused the highest

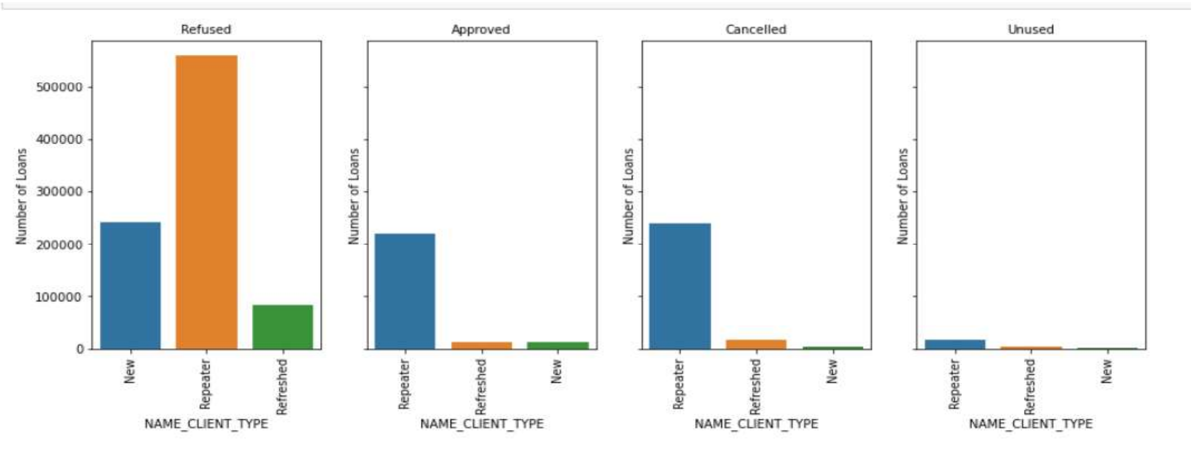


Observation: Applicants whose loans which were previously refused or cancelled have a higher default rate



Observations:

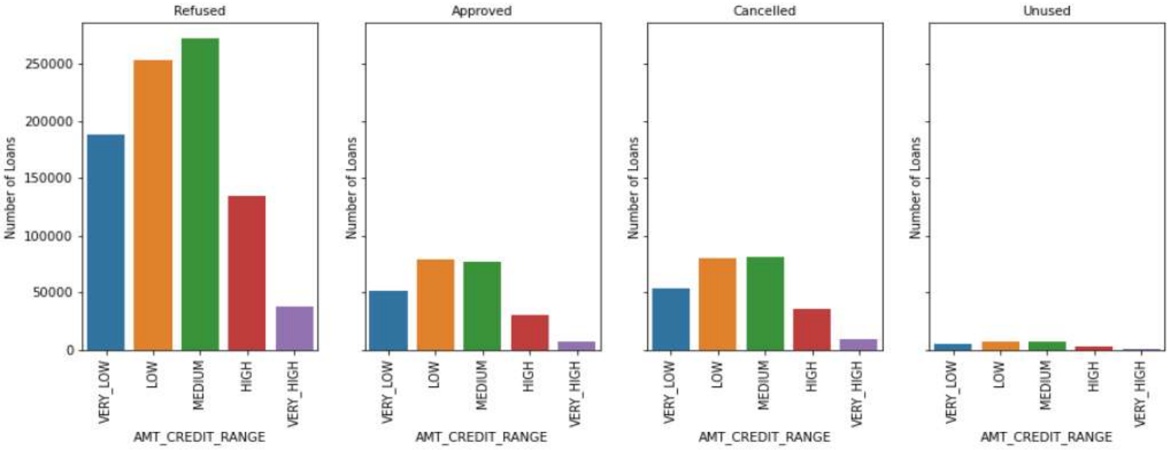
- Loans have been requested maximum for Repairs, Urgent needs, and other categories.
- Highest number of cash loans are taken for repairs. Also, the number of approved and refused cash loans are the highest when applied for repairs



Observations:

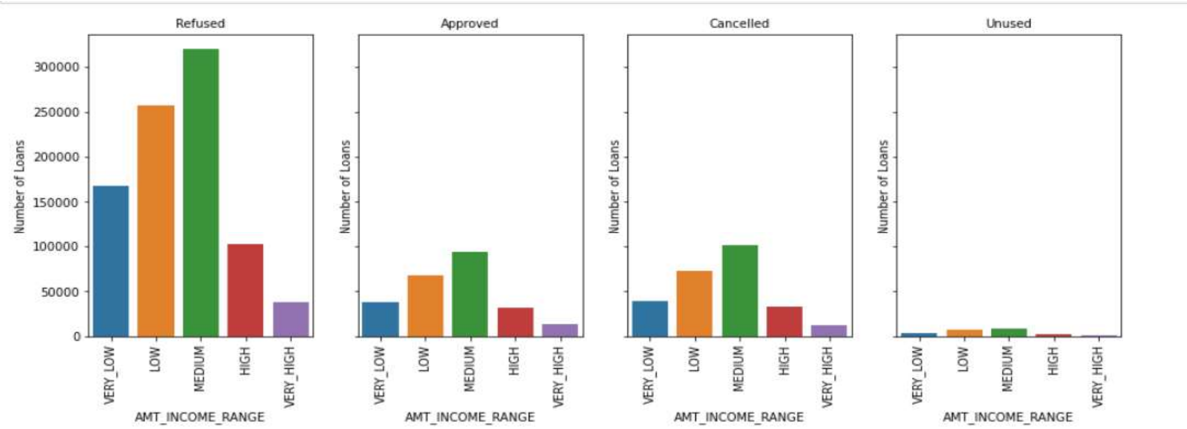
- Repeaters' loans are also highly approved.

- Repeaters' loans are also highly refused, cancelled and unused.
- New clients' loans are highly refused and are rarely approved or cancelled



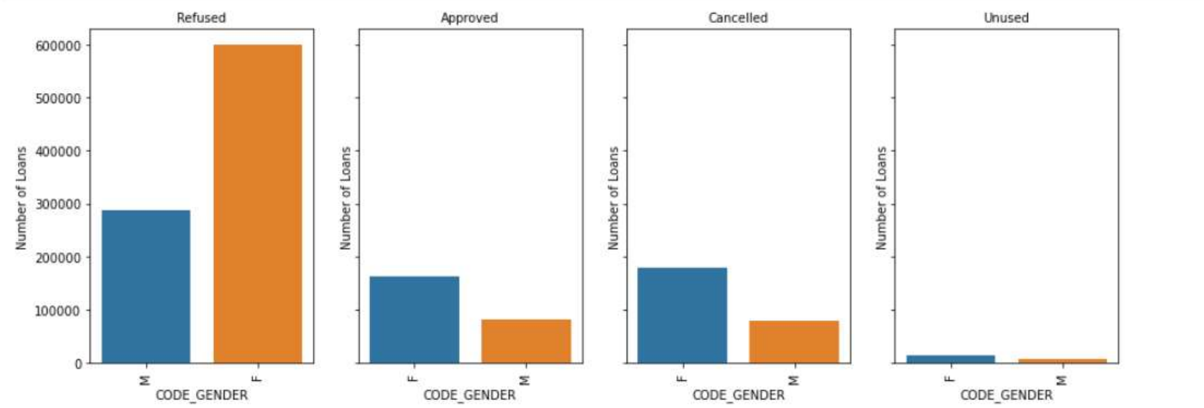
Observations:

- Number of refused loans is the greatest for each category of AMT_CREDIT_RANGE
- Distribution of approved and cancelled loans are similar as seen from the graph
- Medium AMT_CREDIT_RANGE are highly refused followed by low AMT_CREDIT_RANGE followed by very low AMT_CREDIT_RANGE.



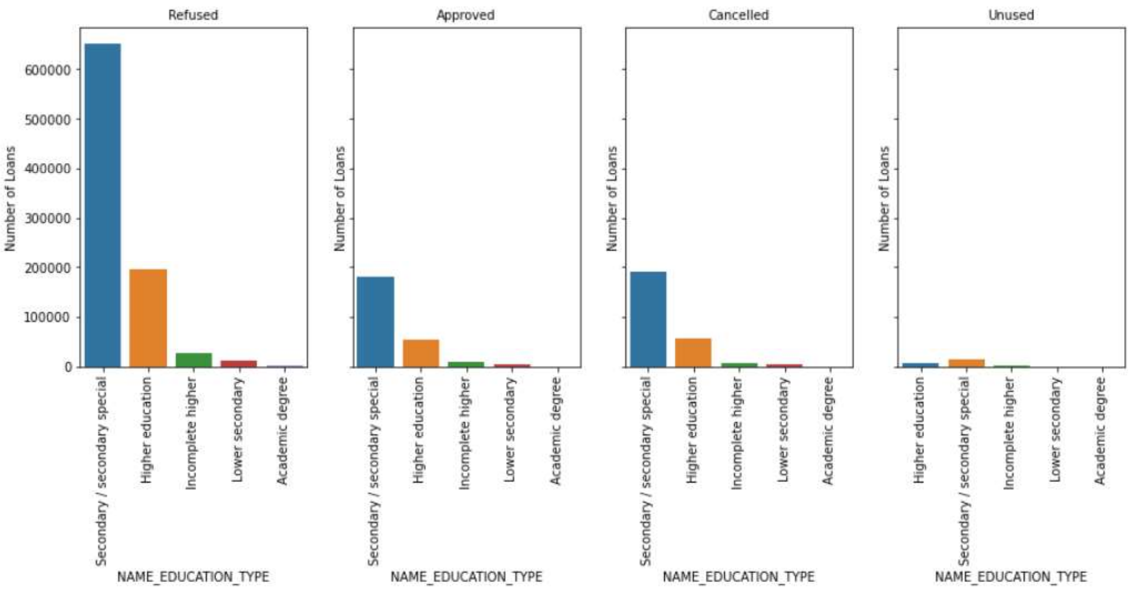
Observations:

- Medium AMT_INCOME_RANGE has the highest approval
- Distribution of approved and cancelled loans are similar as seen from the graph
- The shape of distribution graphs looks similar for all categories of income range



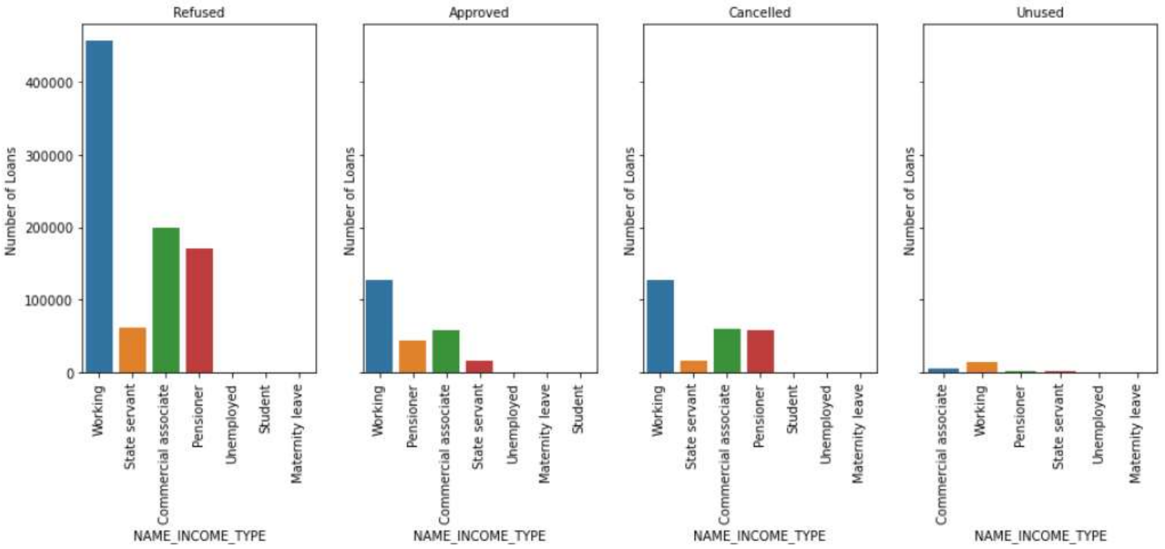
Observations:

- Loans are highly refused, approved, cancelled and unused for females
- Males have an average number of loans refused, approved, cancelled and unused.



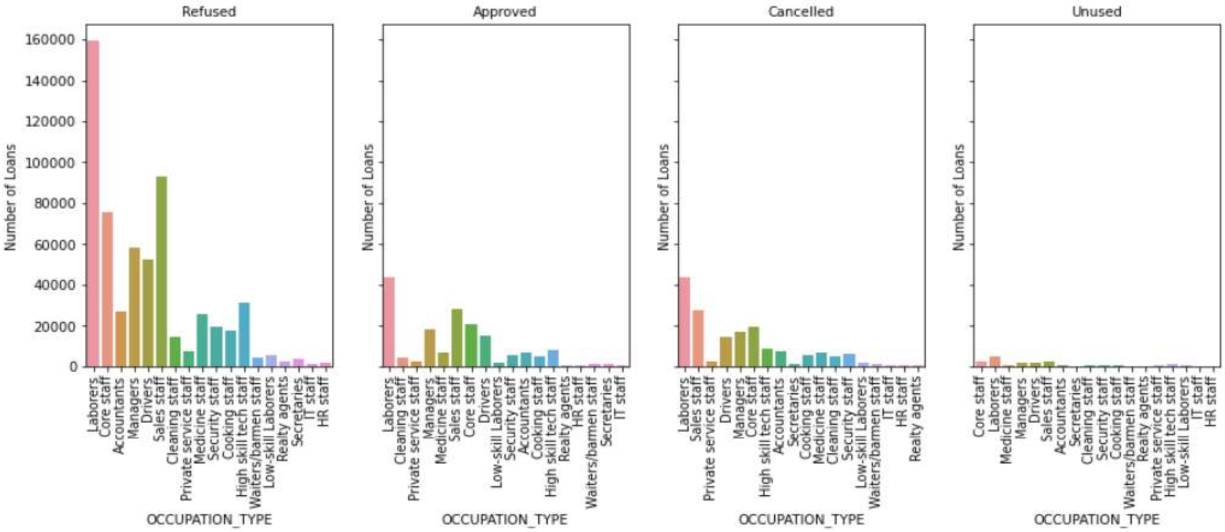
Observations:

- Distribution of loans disbursed to Secondary/Secondary special and Higher education are similar in the first three graphs.
- Secondary/Secondary special has the highest distribution in the first three graphs

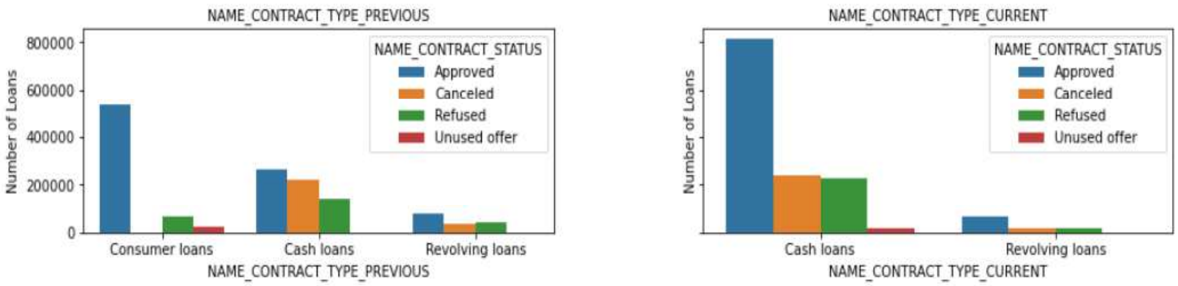


Observations:

- Working applicants have the highest number of approved loans.
- Pensioner applicants' loan are highly approved when compared to refused or cancelled

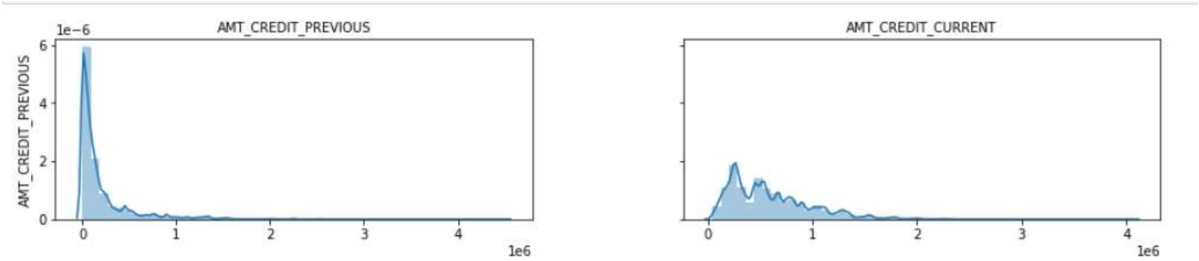


Observation: Laborers loans are highly approved, refused and cancelled followed by Sales staff



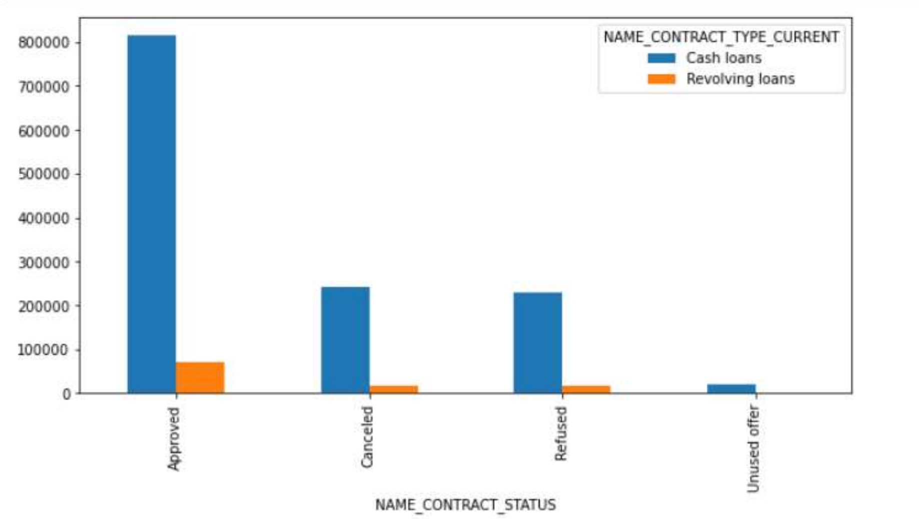
Observations:

- Currently, only two types of loans are offered - namely cash and revolving. Previously, three types of loans were offered - namely consumer, cash and revolving
- Number of cash loans approved are far higher than in previous applications

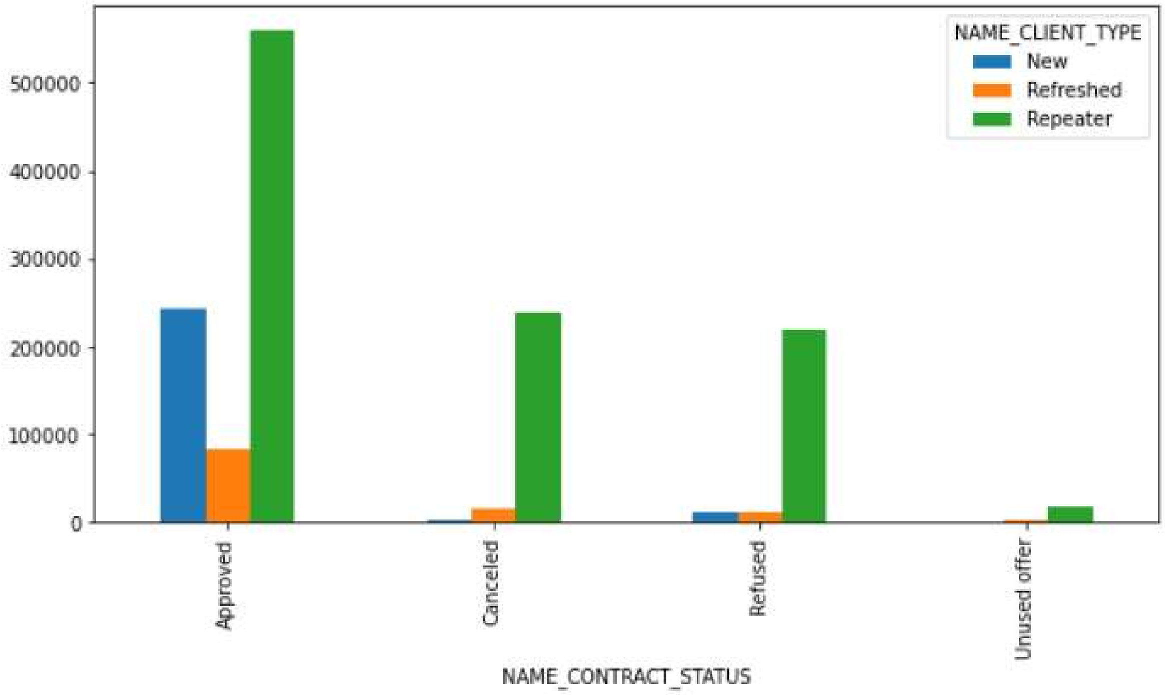


Observation: Number of loans taken were highest for lower credit limit in previous applications. This does not seem to be the case in current application

Bi-variate Categorical Analysis

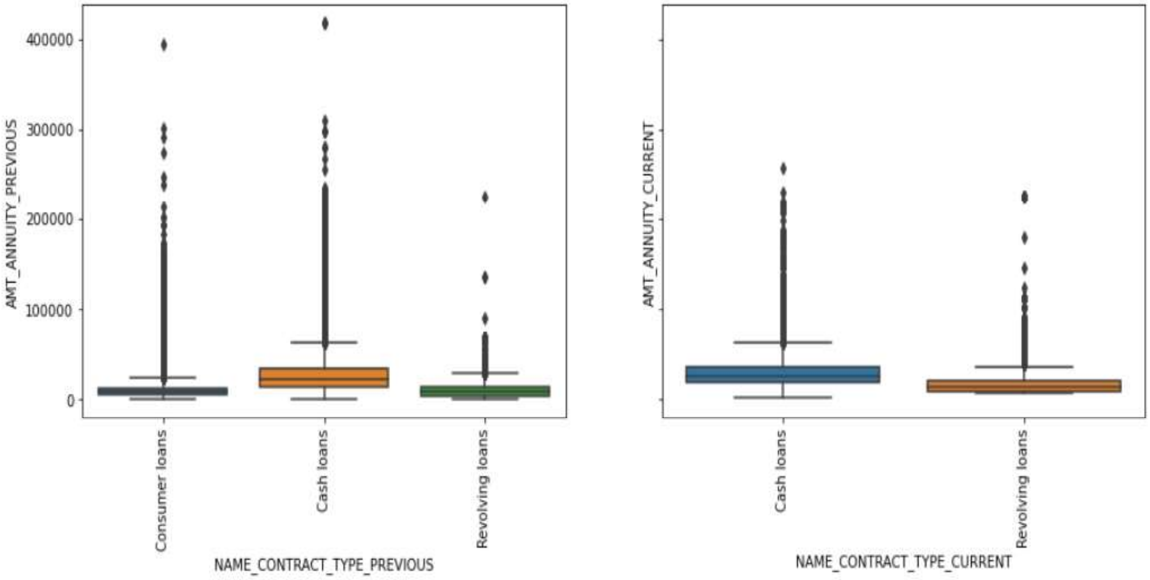


Observation: Cash loans are the highest approved loans.



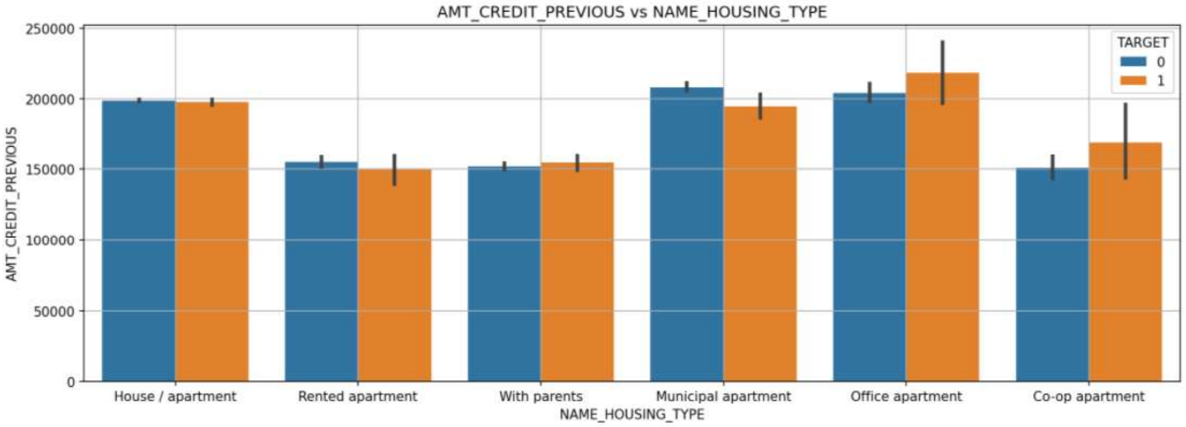
Observation: Repeated applicants have high approval rate

Bi-variate Continuous Analysis



Observations:

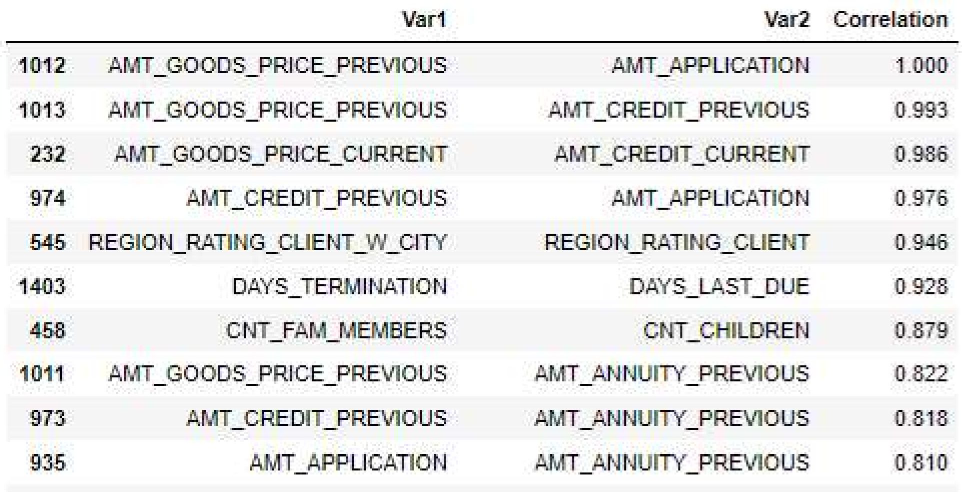
- Cash loans has the highest amount of annuity.
- Revolving loans has the lowest amount of annuity.



Observations:

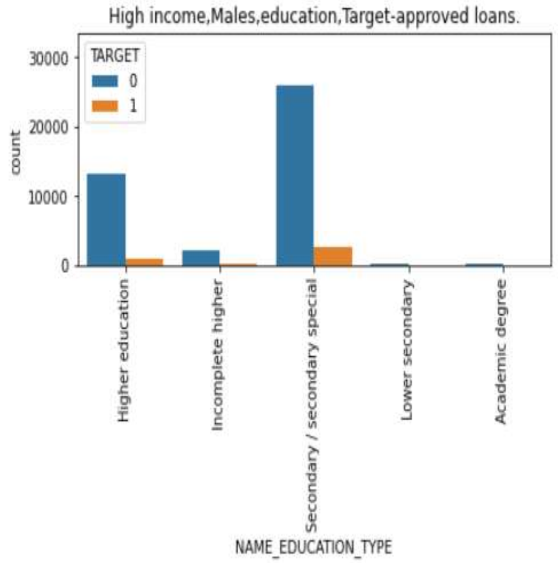
- Credit range for House/apartment, Municipal apartments and Office apartment housing type is the highest.
- Clients who own a co-op apartment and office departments have higher default rate

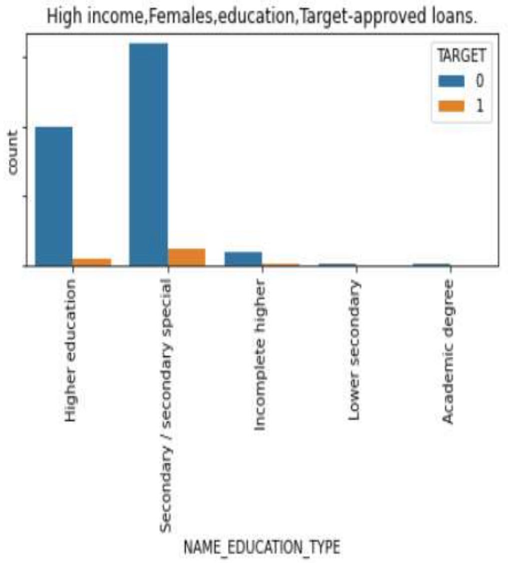
Correlation between previous and application data variables

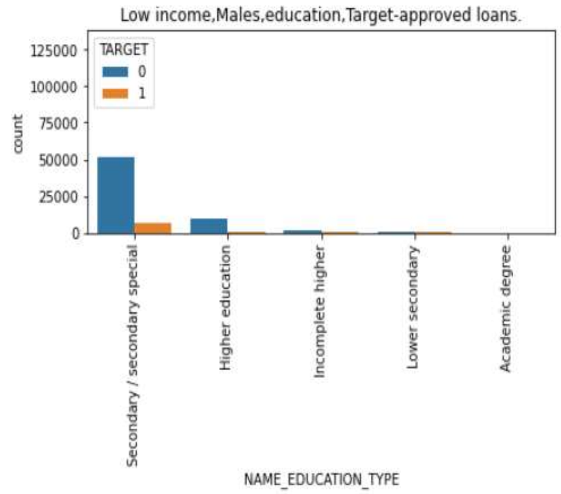


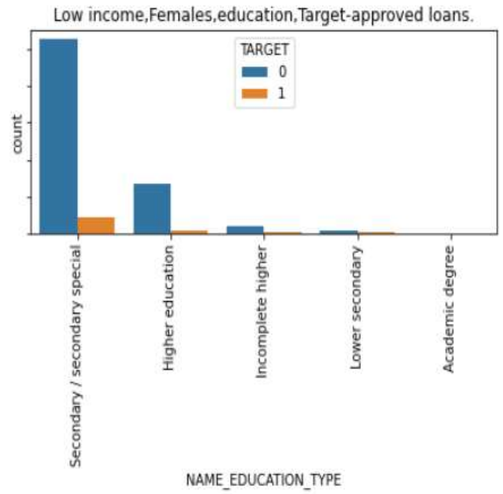
	Var1	Var2	Correlation
1012	AMT_GOODS_PRICE_PREVIOUS	AMT_APPLICATION	1.000
1013	AMT_GOODS_PRICE_PREVIOUS	AMT_CREDIT_PREVIOUS	0.993
232	AMT_GOODS_PRICE_CURRENT	AMT_CREDIT_CURRENT	0.986
974	AMT_CREDIT_PREVIOUS	AMT_APPLICATION	0.976
545	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.946
1403	DAYS_TERMINATION	DAYS_LAST_DUE	0.928
458	CNT_FAM_MEMBERS	CNT_CHILDREN	0.879
1011	AMT_GOODS_PRICE_PREVIOUS	AMT_ANNUITY_PREVIOUS	0.822
973	AMT_CREDIT_PREVIOUS	AMT_ANNUITY_PREVIOUS	0.818
935	AMT_APPLICATION	AMT_ANNUITY_PREVIOUS	0.810

Observation: A very high correlation is observed between Application Amount and the amount of the goods from already processed loans.

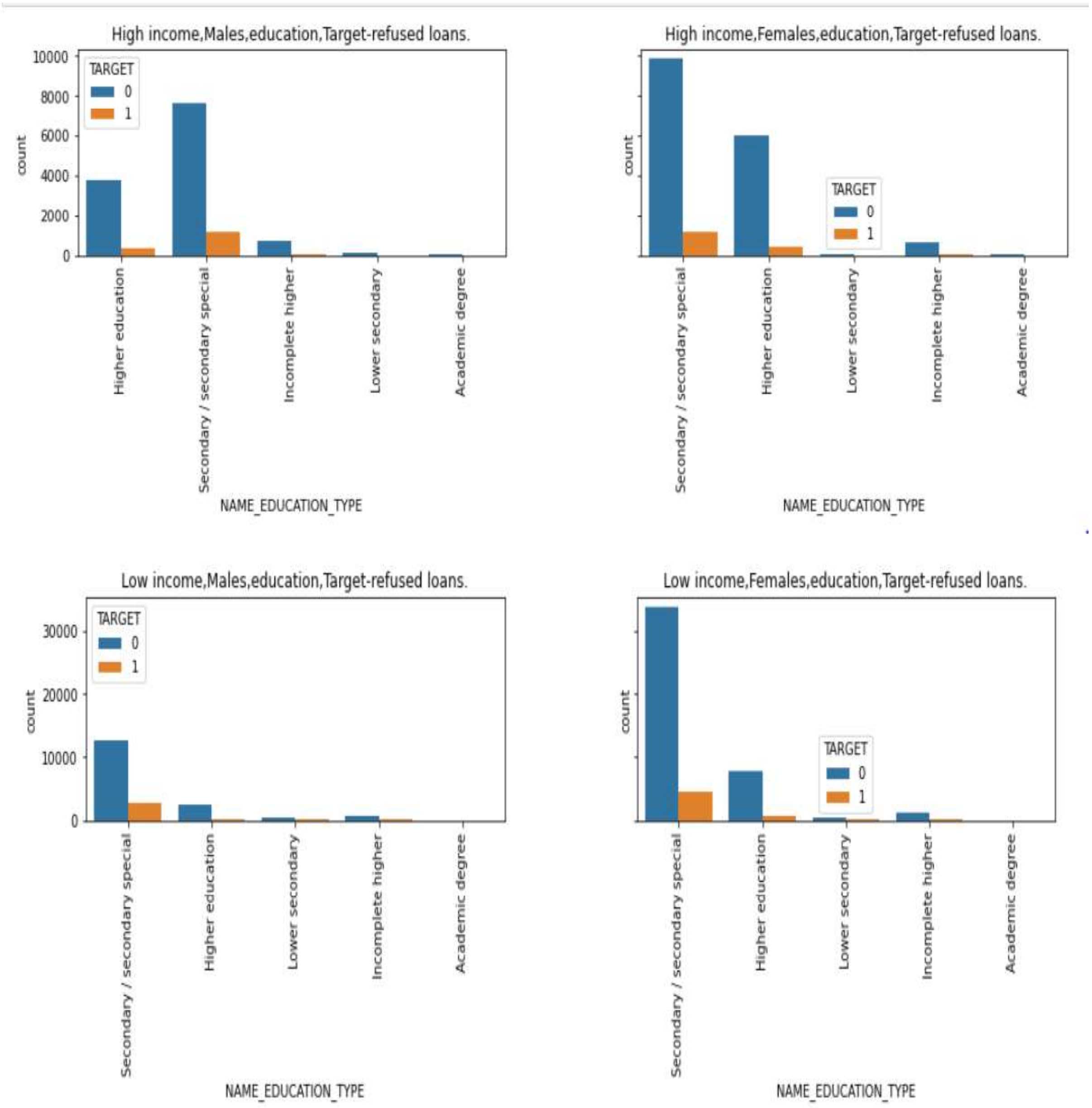








Observation: Out of all the loans the bank has approved for low- and high-income applicants, the number of females that are able to repay the loans are much higher than males



Observations:

- Refused loans, a HUGE proportion of applicants were refused of loans even if they wouldn't have ended defaulting especially the females.
- Applicants with secondary or secondary special education show highest trend in refused as well as approved loans

Conclusion

Driver variables to consider for loan prediction to minimize risk loss:

1. NAME_EDUCATION_TYPE
2. AMT_INCOME_TOTAL
3. AMT_CREDIT
4. AMT_ANNUITY
5. NAME_INCOME_TYPE
6. CODE_GENDER
7. NAME_HOUSING_TYPE

Recommendation to bank

- Maximum number of loans applied by existing customers were previously refused. So now when they have applied again, if total income has increased or credit amount has decreased or the annuity has decreased such that now the payment is possible, then loans can be approved to them. Also, if now the target variable for those applicants shows 0, their loans can be approved.
- Approve more applicants with more high annual income, secondary and highly educated and less credit amount. This category is the least likely to default.
- Males with high income and secondary/secondary special and higher education are more likely to default. Females with the same characteristics have lower default rate.
- Revolving loans are lesser in the defaulter category. Hence, revolving loans are comparatively safer. The bank has lost quite a good number of profits by not approving revolving loans
- Banks should focus on Student, Pensioner and Businessman for successful payments. Pensioner clients, although not earning, repay the loans. Students haven't defaulted a loan although they are not employed.
- Clients who own a co-op apartment and office departments have higher default rate for higher final credit amount on the previous application. Credit range for House/apartment, Municipal apartments and Office apartment housing type is the highest and default rate is low. Banks must focus on clients with House/apartment, Municipal apartments, and Office apartment housing type.
- Clients whose registered city is not the same as the city where they live or work, default rate is higher. This may be due to higher living expenses while staying away from their families. Banks must focus on clients who work in the same city as the registered city.
- People with zero children are much more among non-defaulters which shows that people who are applying for loan have less people dependent financially on them which will lessen the loan payment difficulties. Banks must focus on them.
- Defaulters usually come unaccompanied while applying for a loan. Banks must be informed of this trend.