Risk Analytics in banking and financial services EDA

Group case study by Gayatri Challapalli

Objective

This case study aims to identify patterns present in the data through EDA(Exploratory data analysis) .The data used in this analysis consists the current and previous application details of the customers who has applied for a loan (Cash/Revolving) from a bank/financial company .

Lets identify the patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

Approach

1. Data Cleaning
2. Binning the continuous column
3. Check for Data imbalance
4. Univariate Analysis
5. Bivariate Analysis
6. Correlation Matrix
7. Insights from the analysis

Note :
We followed the same approach for application and previous application data and then merged them to get the final insights.

Data Cleaning - missing values

Application data has 307511 rows and 122 columns.

Identify the columns having missing values (NAN)<40% and exclude them as they might not help in the analysis.

Evaluate the missing value depending on importance of columns for analysis:

1) For categorical columns we can fill missing values with Mode
2) For continuous columns:
    a) Fill with mean if there is no outliers.
    b) Fill with median if there is outliers in data

Note :
We have imputed the missing values as per the above mentioned approach in the python notebook.

Data Cleaning - missing values

4.2.3 NAME_TYPE_SUITE column Imputation

```
AD['NAME_TYPE_SUITE'].value_counts()
```

```
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
```

Inference :

As NAME_TYPE_SUITE has categorical data, hence can impute the missing values with mode.

```
inputVAL = AD['NAME_TYPE_SUITE'].mode()[0]
print(f'The mode of the column: {inputVAL}')
```

```
The mode of the column: Unaccompanied
```

```
# Replacing the null values by mode
AD['NAME_TYPE_SUITE'].fillna(inputVAL,inplace=True)
```

We are imputing the NAN's of NAME_TYPE_SUITE column with its mode as its a categorical variable.

Checking for outliers

Outlier : An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

Approach to identify outliers:

The box plot helps to identify the outliers as it describes the behavior of the data in the middle as well as at the ends of the distributions with the help of median and the lower(25th percentile and Q1) and upper quartiles(75th percentile and Q3)

lower inner fence: Q1 - 1.5*IQ upper inner fence: Q3 + 1.5*IQ

lower outer fence: Q1 - 3*IQ upper outer fence: Q3 + 3*IQ

A point beyond an inner fence on either side is considered a mild outlier. A point beyond an outer fence is considered an extreme outlier.

Treat the outliers :

a) Drop the outlier if there is an error in data

b) Cap the data to 95th or 99th or 5th percentile depending upon data.

Checking for outliers

```
#Checking for outliers in DAYS_EMPLOYED column

plt.figure(figsize=(10,2))
sns.boxplot(AD.DAYS_EMPLOYED)
plt.show()
```

4.5.1 Treating the outliers

```
#Excluding values outside 99%ile in each of the 3 variables
AD=AD[AD.DAYS_EMPLOYED<np.nanpercentile(AD['DAYS_EMPLOYED'], 99)]
```

In the Days_employed columns we are treating the outliers by capping it to 99th percentile

Binning (Numeric to categorical)

Binning also known as discretization is the process of transforming numerical variables into categorical.Binning also helps us to quickly evaluate outliers, invalid or missing values for numerical values.

Example : Columns like AMT_INCOME_TOTAL,AMT_CREDIT,YEARS_BIRTH can be binned to make our analysis easy.

Approach to bin the numeric columns:

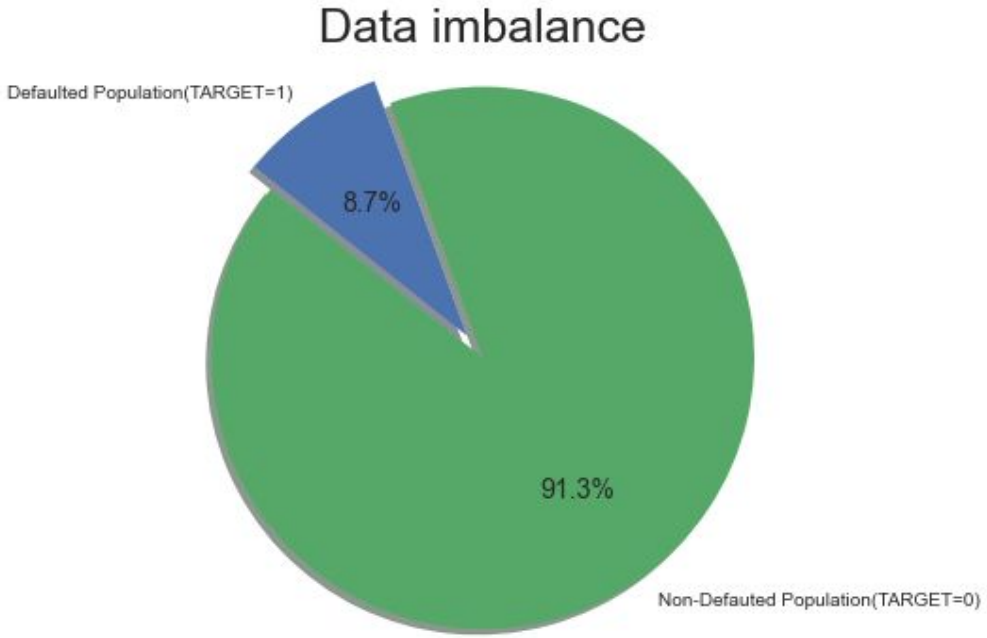
We use the pd.qcut method and specify the labels and quantile cuts

```
: #Creating binned var of AMT_INCOME_TOTAL
range_labels = ['Very Low', 'Low', "Medium", 'High', '
AD["AMT_INCOME_RANGE"] = pd.qcut(AD.AMT_INCOME_TOTAL,
                                 q=[0, .2, .4, .6, .8, 1]
                                 labels=range_labels)
```

```
: AD['AMT_INCOME_RANGE'].value_counts()
```

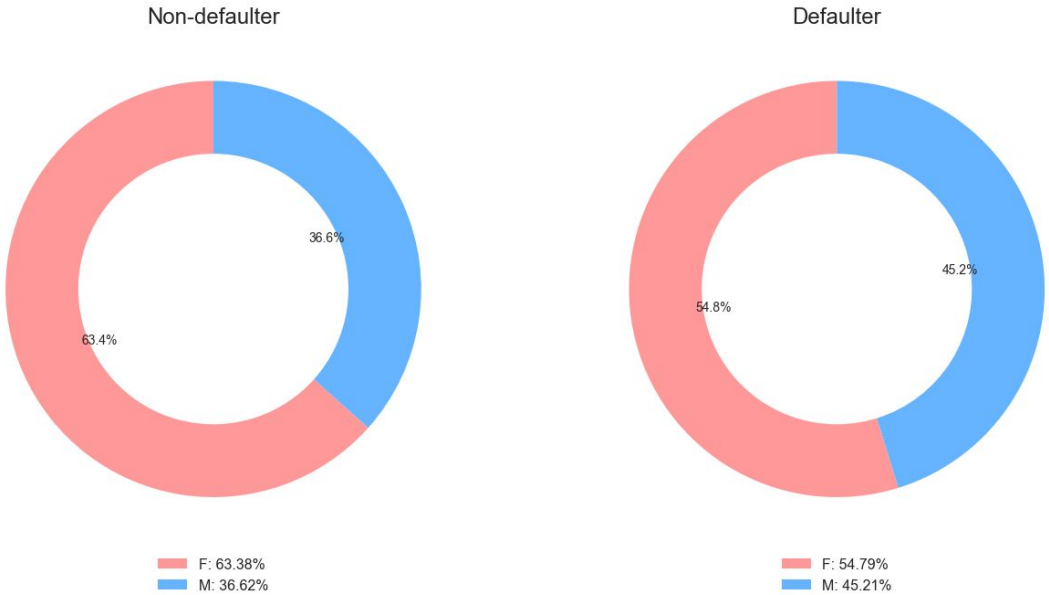
```
: Very Low     73180
  High         60643
  Low          40399
  Very high    38101
  Medium       34537
  Name: AMT_INCOME_RANGE, dtype: int64
```

Data imbalance



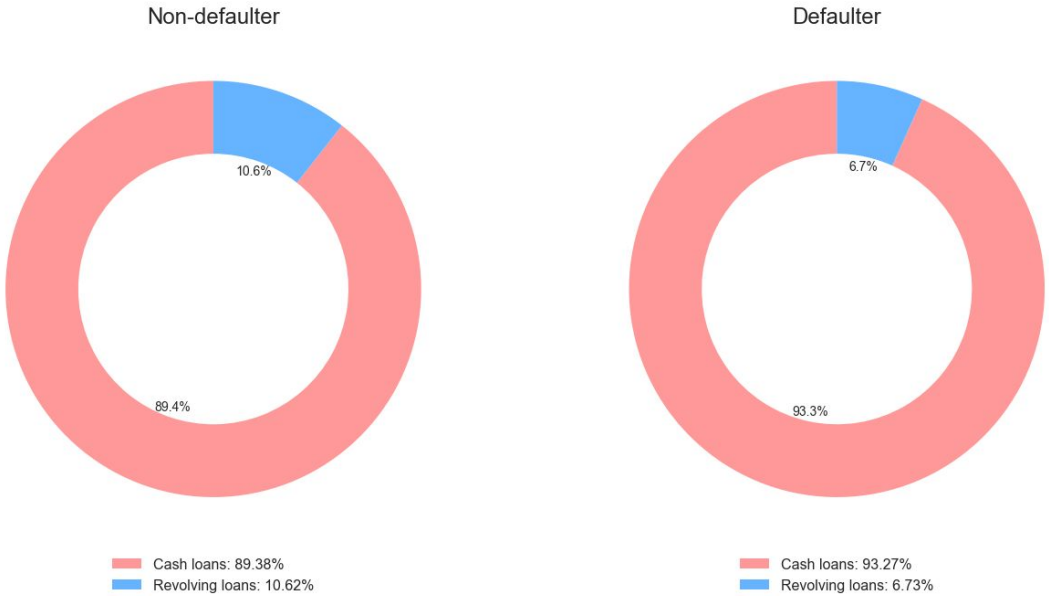
We can see that the Application data has high imbalance with Defaulted population at 8.7% as compared to non-defaulter population at 91.3% Imbalance ratio is 11.4.

Univariate analysis - Gender [CATEGORICAL]



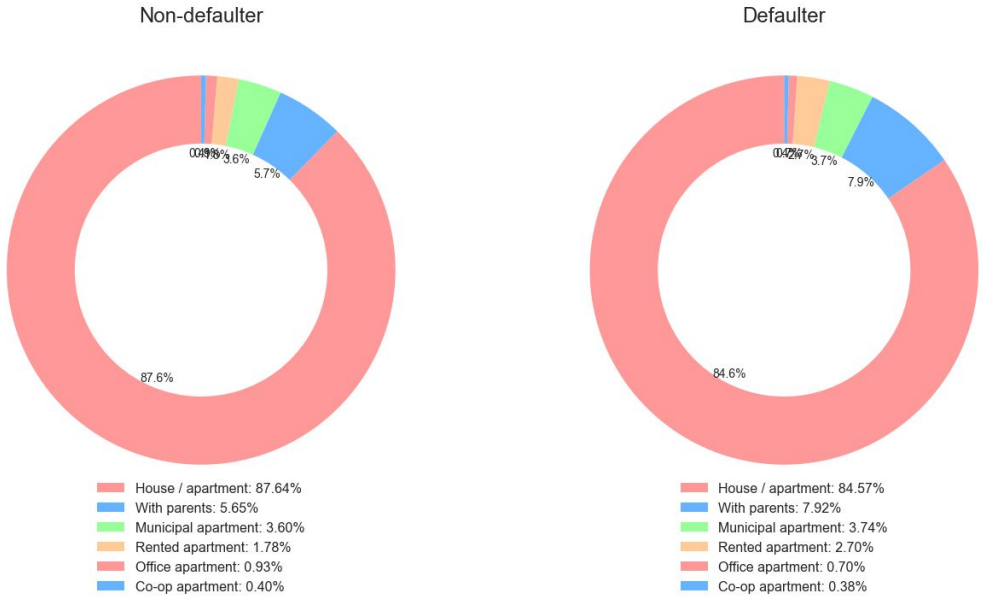
1. Female clients are applying more loans under both defaulters and non-defaulters
2. The percentage of males in the defaulter is more than that in the non-defaulters

Univariate analysis - Type of loan [CATEGORICAL]



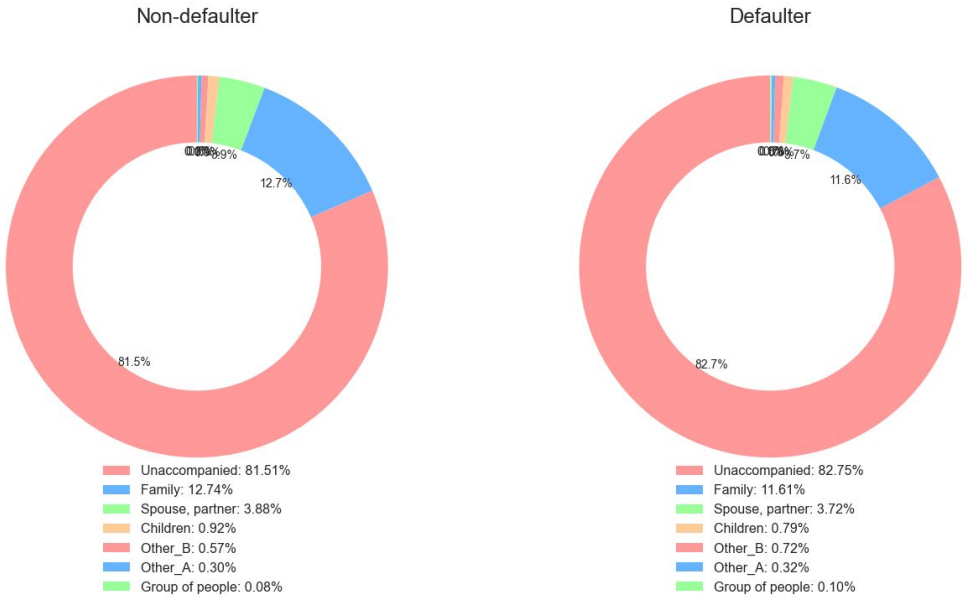
We can notice that revolving loans are lesser in the defaulted population. Hence we can infer that these loans are comparatively safer. As in revolving loans once you repay the amount owed, the credit becomes available to draw on again thus clients may repay due to its repayment and re-borrowing flexibility

Univariate analysis - Housing type [CATEGORICAL]



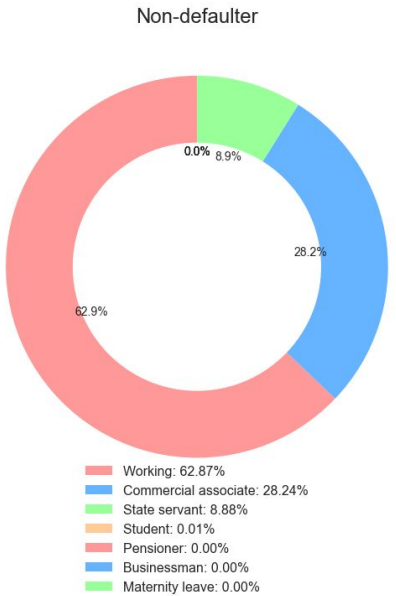
1. Percentage of people living in own house/apartment is more in case of non-defaulters.Hence people with own house have comparatively less chance to default the loan.
2. People living with parents and living in rented apartments are more likely to default the loan may be due to more expenditure.

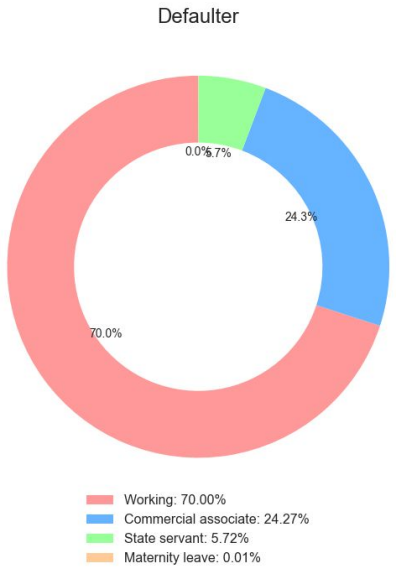
Univariate analysis - Name type suite [CATEGORICAL]



Most of the clients have no one who accompanied them in both defaulter and non-defaulter case.

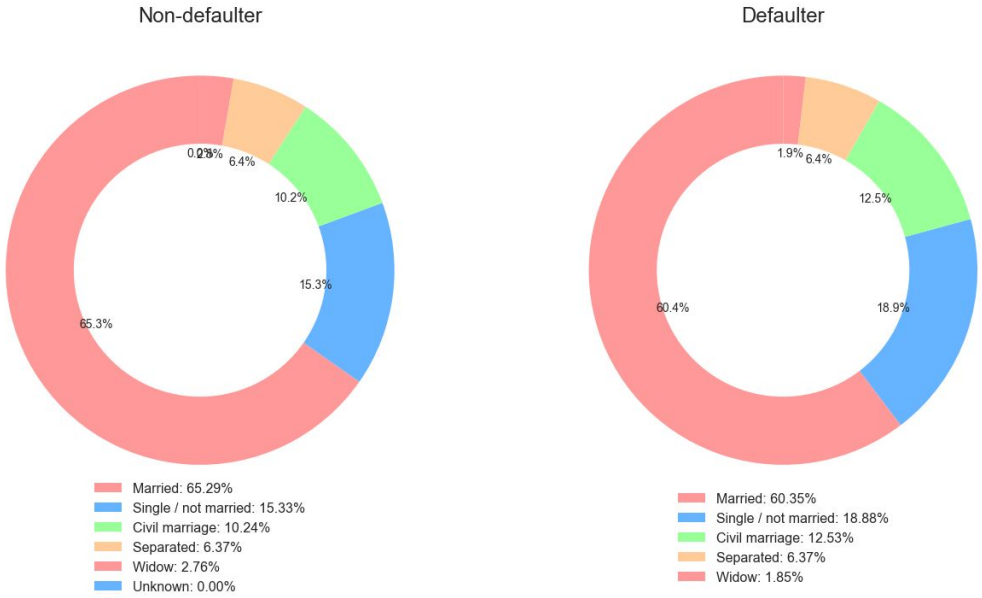
Univariate analysis - Income type [CATEGORICAL]





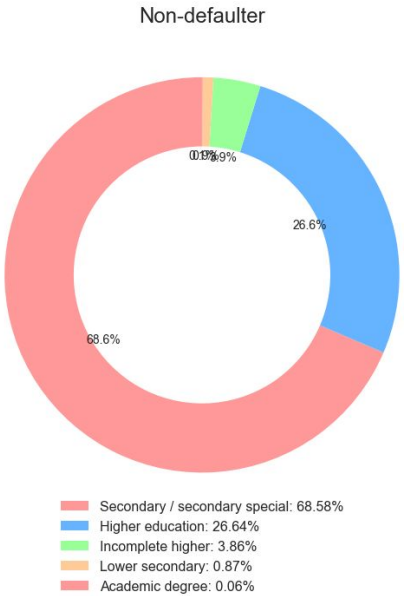
1. Defaulters with have a higher percentage of income from working and commercial associate and less percentage of income from pension than non-defaulters do.
2. There are some non-defaulters having income from their own businesses and their schools. Defaults do not have income from those sources.

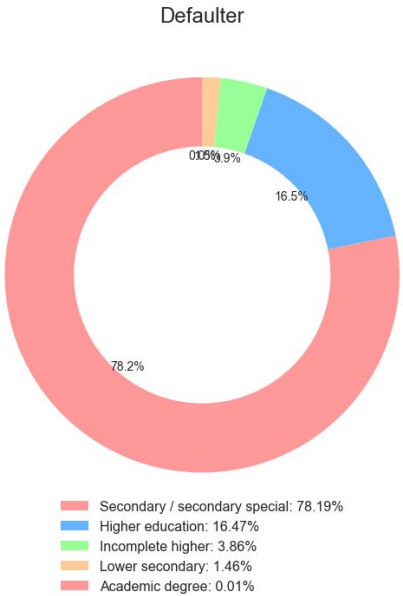
Univariate analysis - Family status [CATEGORICAL]



Married people pay loans on time when compared with single and people under civil marriages.Hence the percentages of non-defaulters who are married is higher than that of defaulters, while defaulters have a higher percentage of single and civil married people.

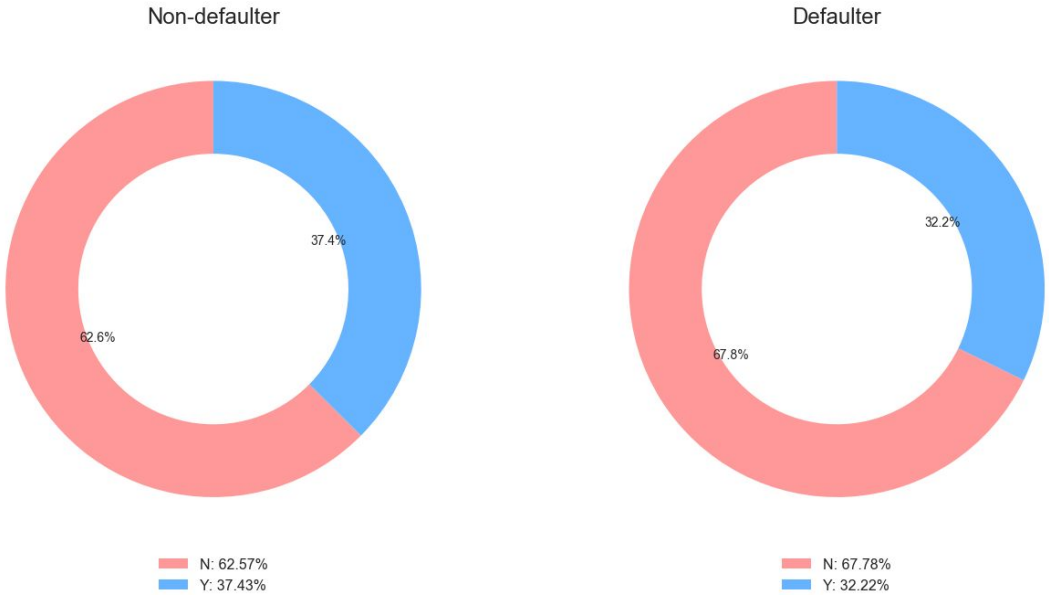
Univariate analysis - Education [CATEGORICAL]





It can be observed that the defaulters has a high percentage of secondary education than the non-defaulters this may be due to education_loans or unemployment.

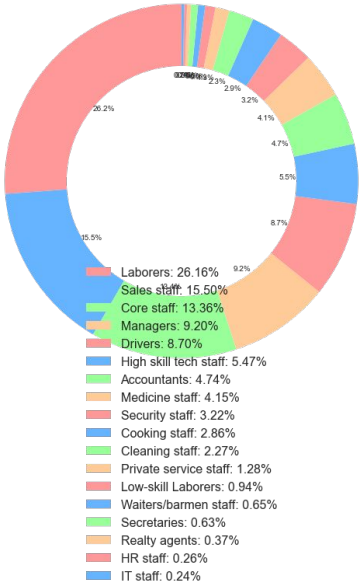
Univariate analysis - Own cars [CATEGORICAL]



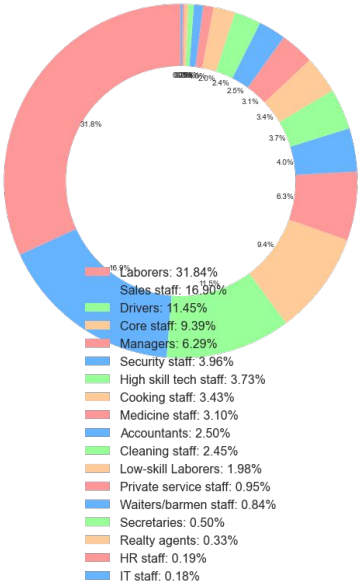
The percentage of people owning a car is higher in defaulters than non_defulters.

Univariate analysis - Occupation Type[CATEGORICAL]

Non-defaulter

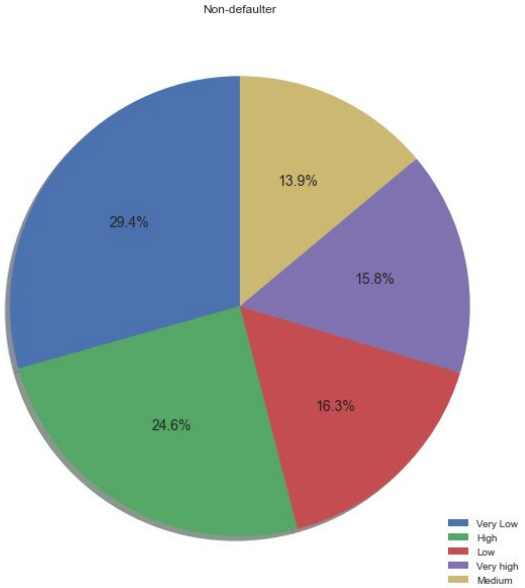


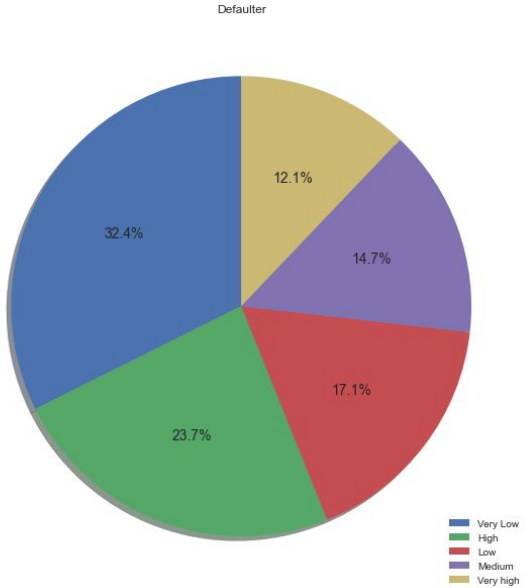
Defaulter



Laborers and other low income people like sales staff , drivers are more in defaulters

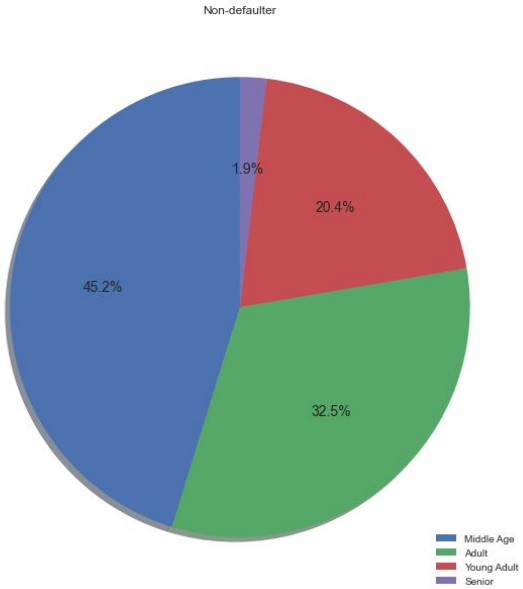
Univariate analysis - Income_Range [CATEGORICAL]

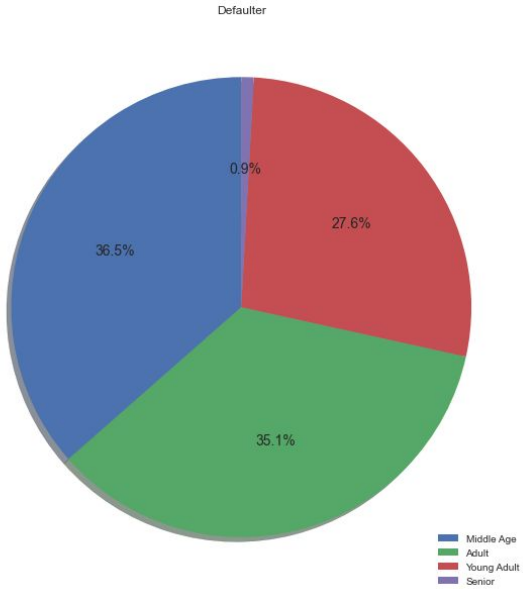




Most of people have a low income under defaulters when compared with non-defaulters.

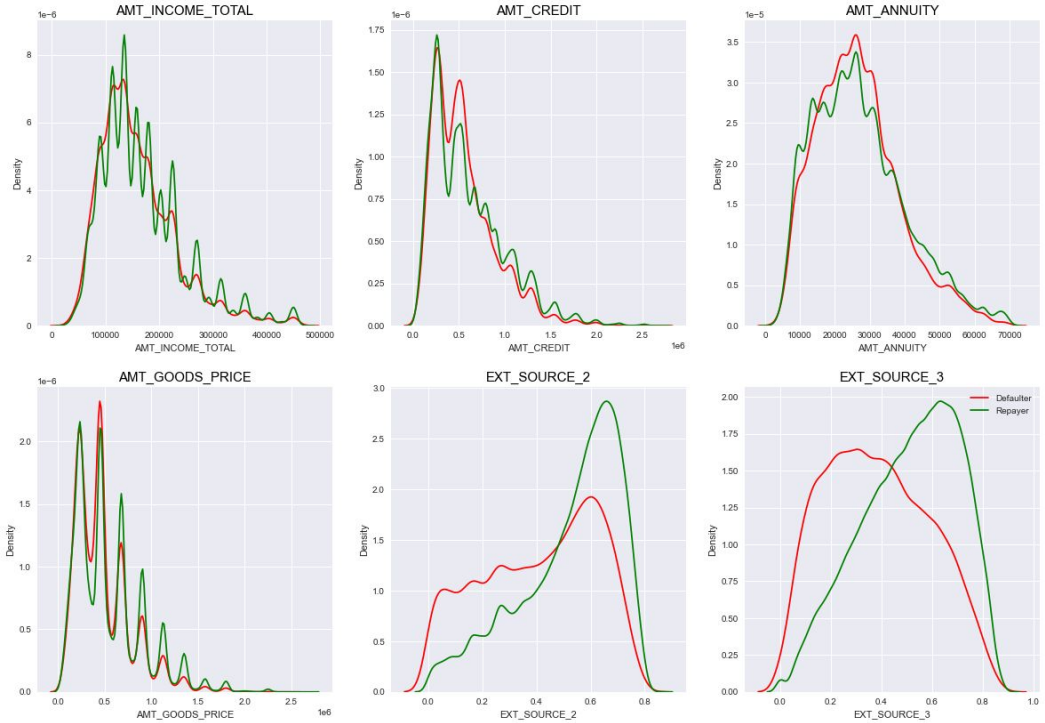
Univariate analysis -Age [CATEGORICAL]





Middle-aged and senior citizens are more reliable than others.:

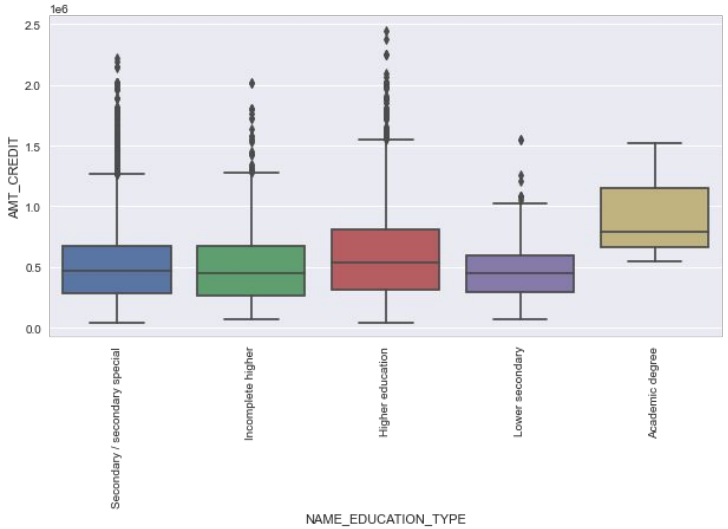
Univariate analysis for continuous variables



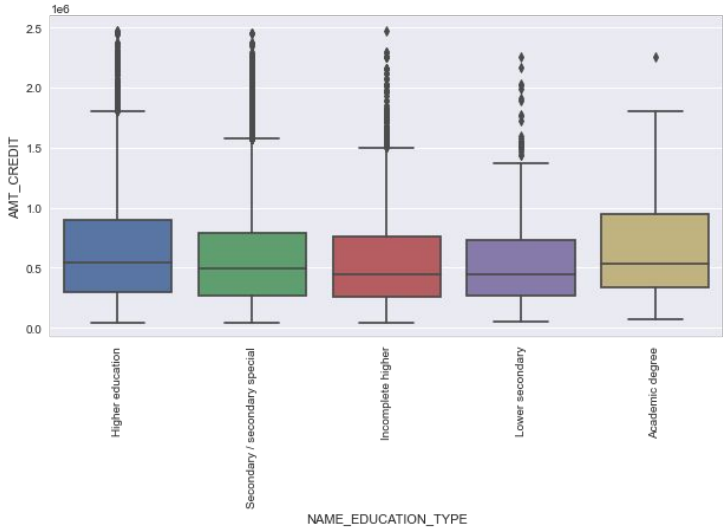
1. Most no of loans are given for goods price below 10 lakhs
2. Most people pay annuity below 50000 for the credit loan
3. Credit amount of the loan is mostly less than 10 lakhs
4. EXT_SOURCE_2 and EXT_SOURCE_3 show a significant difference between defaulters and repayers, both are high for repayers. These two can be marked important for loan approvals

Education type vs Amount credit

Defaulters



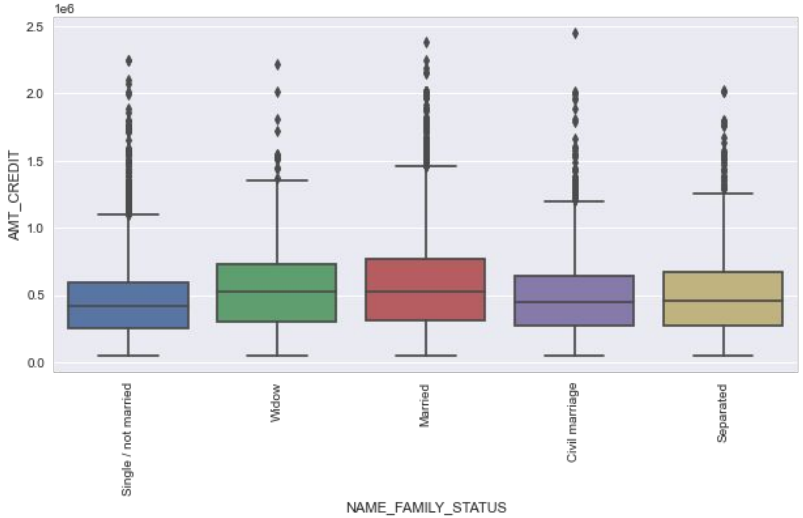
Non-Defaulters



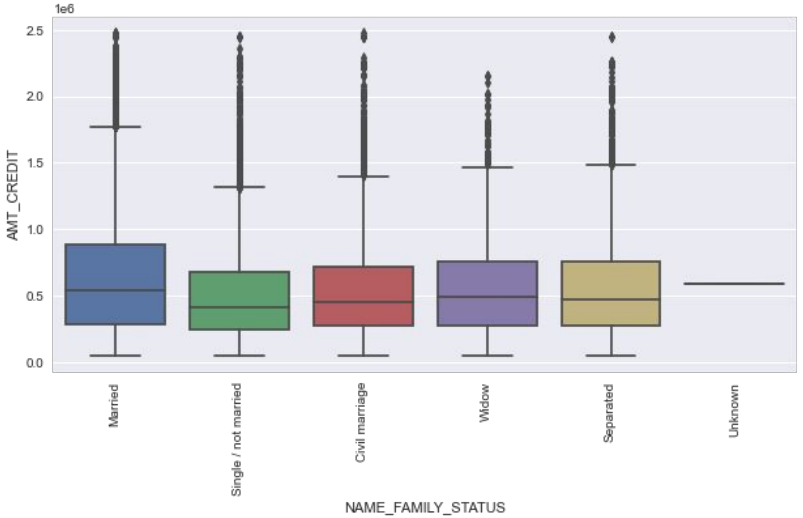
People with academic degree and high education have more credits when compared with others.

Family status vs Amount credit

Defaulters



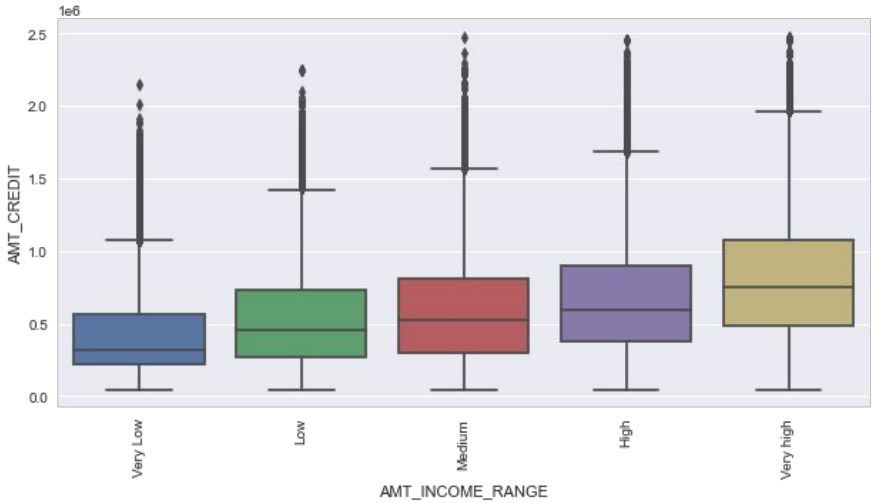
Non-Defaulters



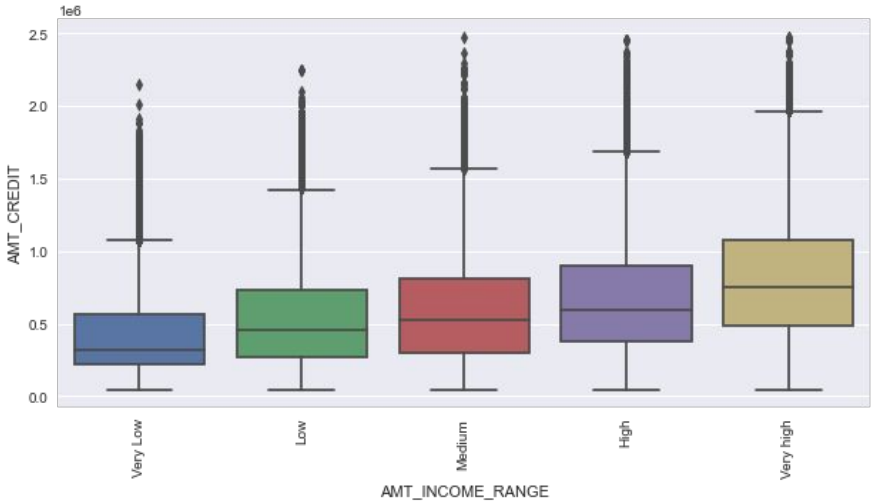
People under married, civil married and separated categories have more credits

Income range vs Amount credit

Defaulters



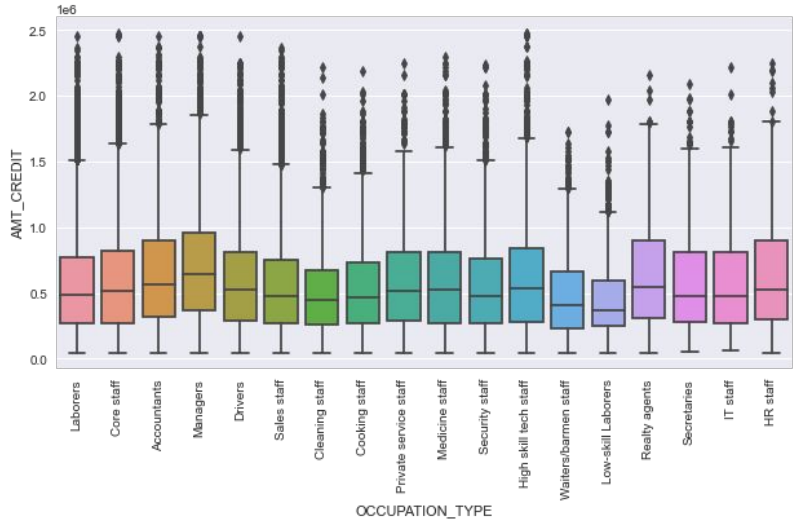
Non-Defaulters



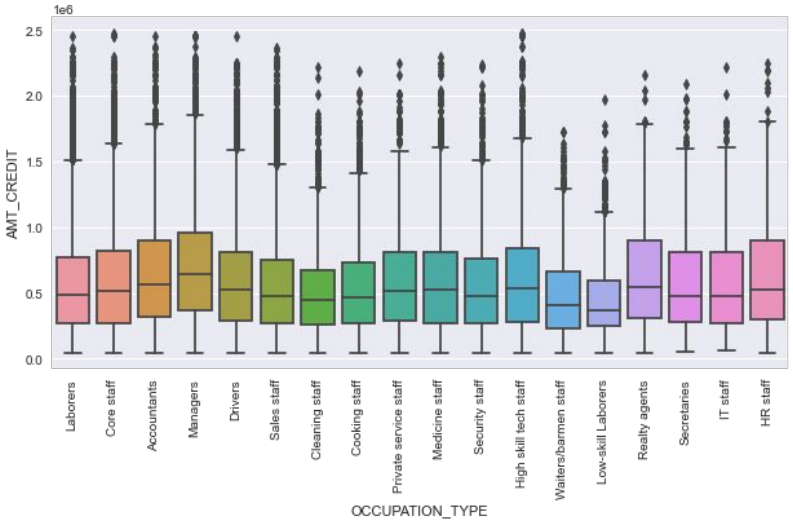
People with higher income seems to have high credit.May be due to there higher income rage there loan limit will be more.

Occupation_Type vs Amount credit

Defaulters

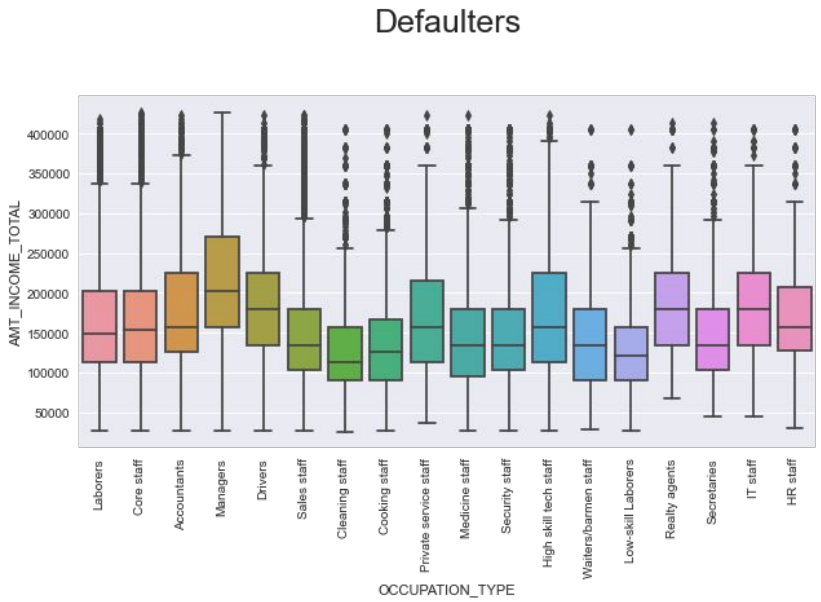


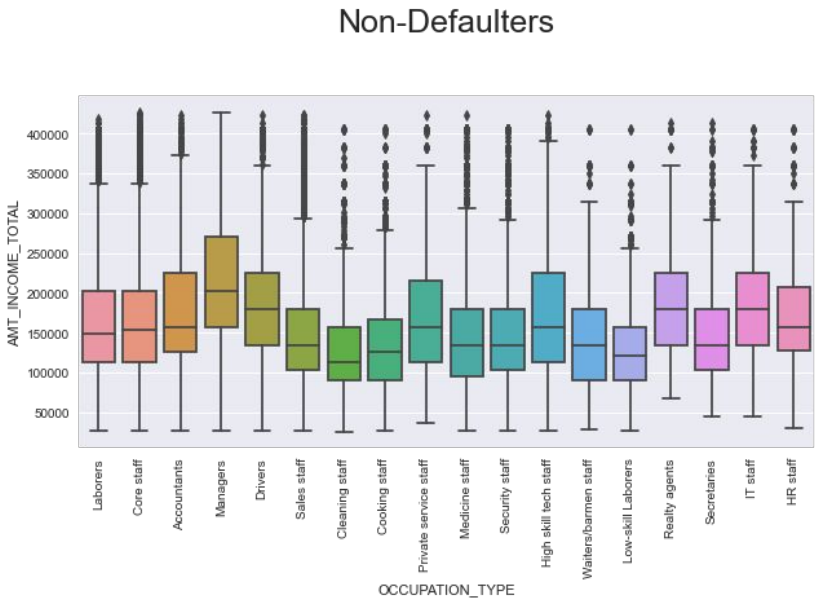
Non-Defaulters



Among defaulters managers and accountants are having more credit amount.Low credit amount are for low skill and waiters.

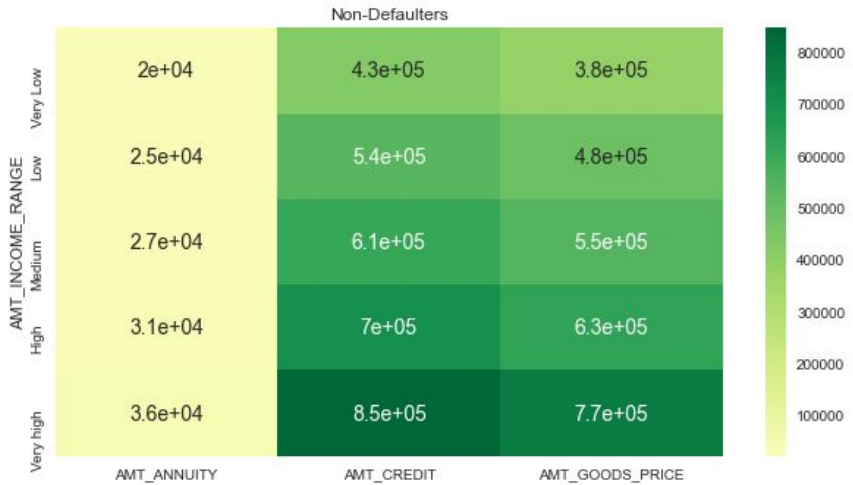
Occupation_Type vs Income range





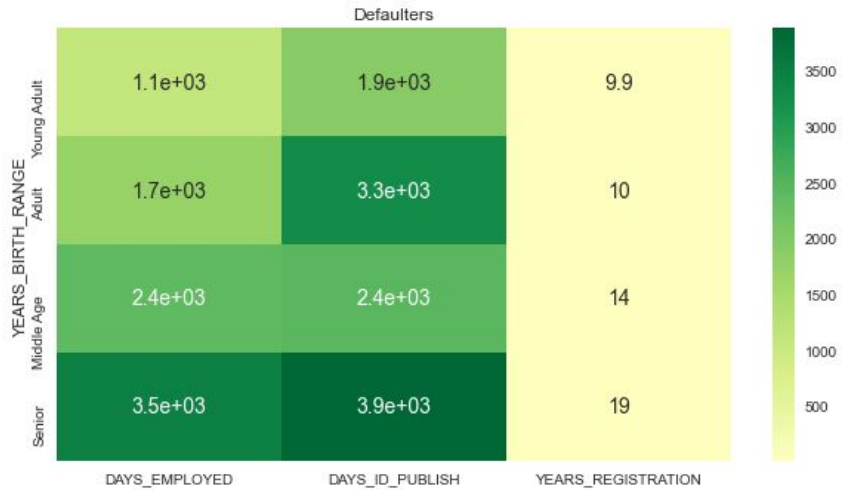
Among defaulters, Managers has more income total and cleaning/cooking/low skill laborers have low income range which is obvious.

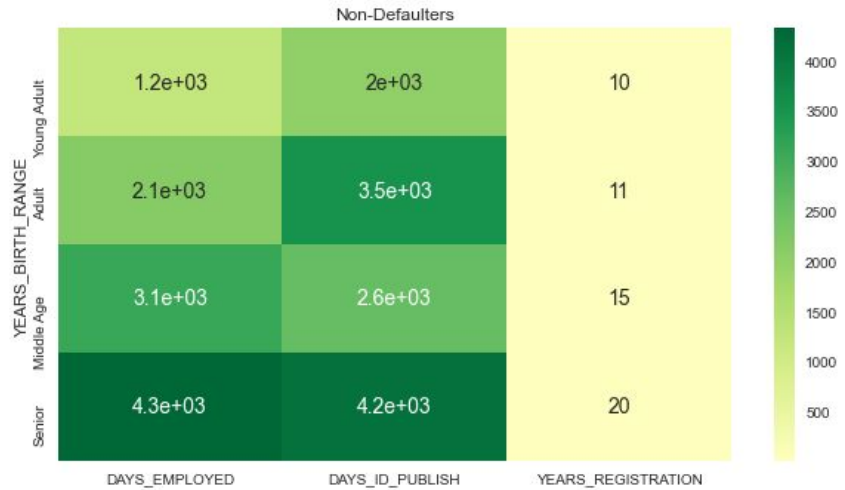
Credit, Annuity, Goods_Price vs Income_Range



1. The average amounts of credit and the average amounts of goods price increase when the income increases.
2. Applications with very low income has high average amounts of credit and high average amounts of annuity with respect to defaulters than those of non-defaulters
3. For the applications with higher income ranges, average amounts of annuity and average amounts of goods price of defaulters are lower than those of non-defaulters.:

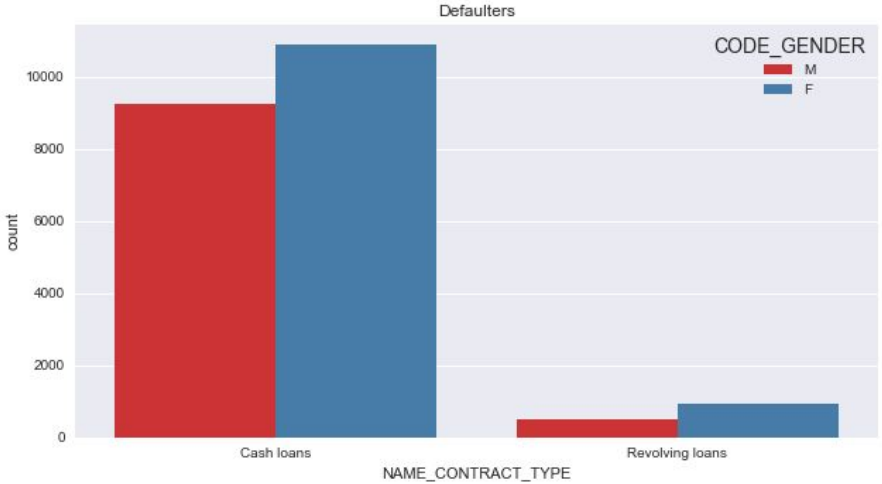
Days_employed,Years_Registration,Days_to_publish vs Years_birth_range

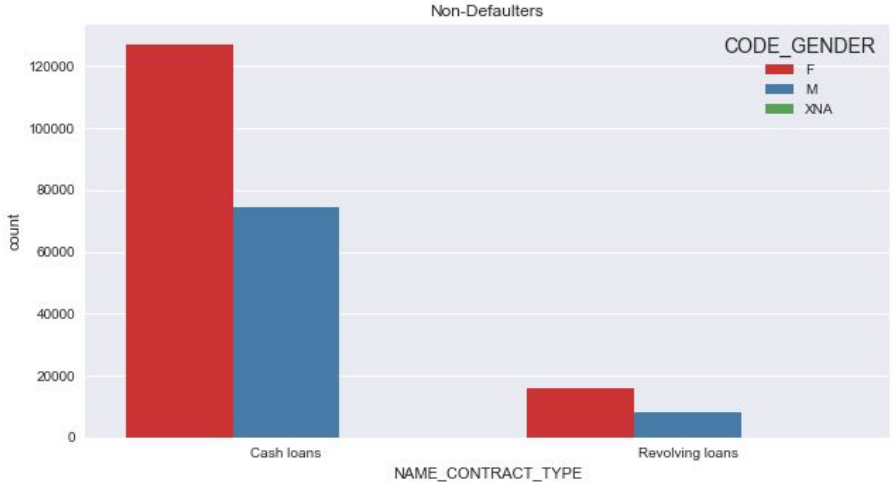




1. DAYS_EMPLOYED is highest for seniors in both Defaulters and non-Defaulters.
2. For all age groups other than seniors, the median values of DAYS_EMPLOYES , DAYS_ID_PUBLISH and DAYS_REGISTRATION of non-defaulters are always higher than those on defaulters.
3. For the group age of seniors, the median values of DAYS_EMPLOYED of the 2 sets of data are equal. The median values of DAYS_ID_PUBLISH and DAYS_REGISTRATION of Non-Defaulters are always higher than those on Defaulters.

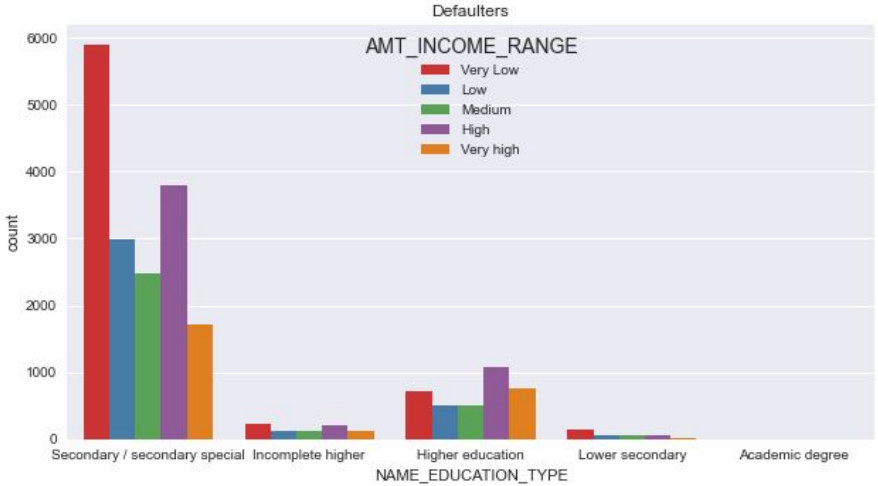
Bivariate analysis -Contract type vs Gender

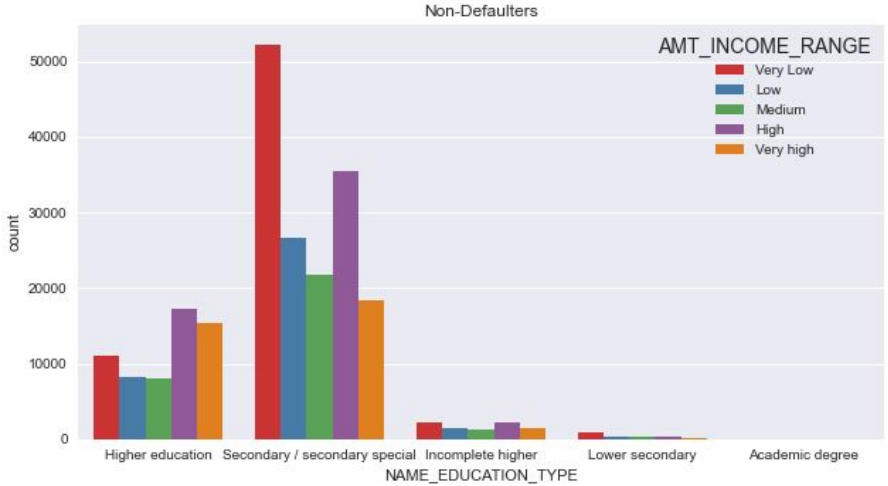




Female clients are applying more loans under both defaulters and non-defaulters

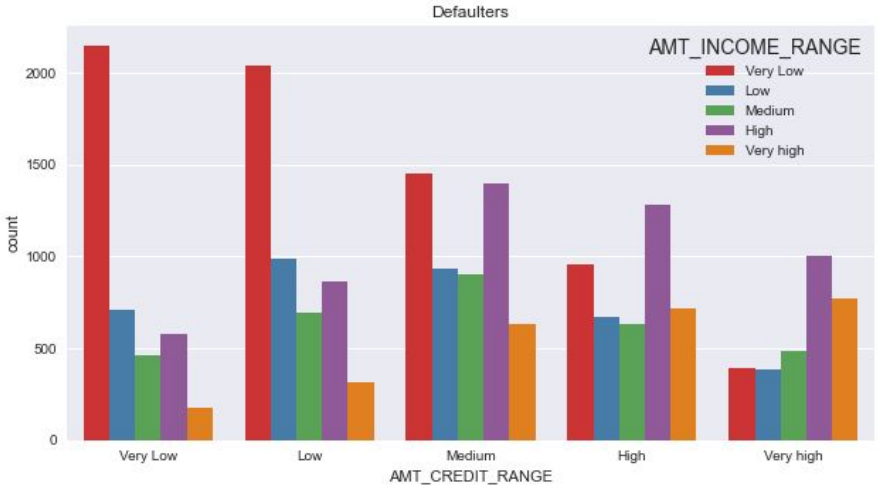
Bivariate analysis -Education Type vs Income-Range

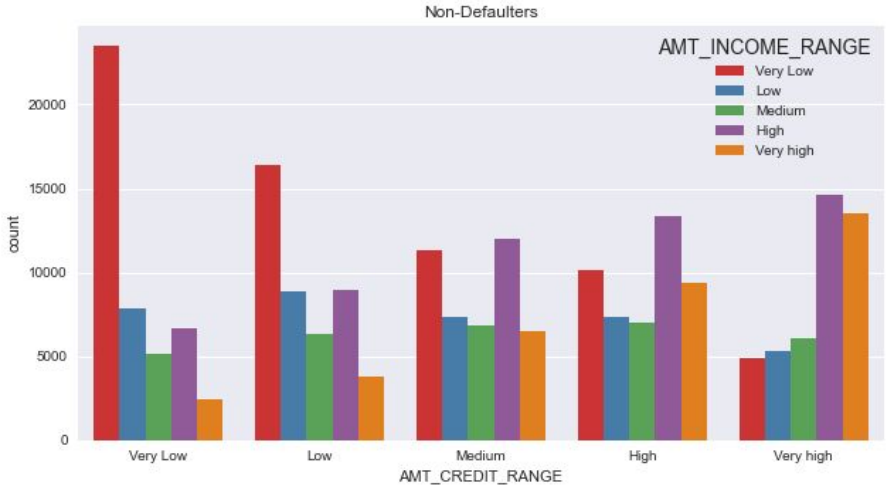




The majority of Defaulters and non-Defaulters just finished their secondary schools and have low income.

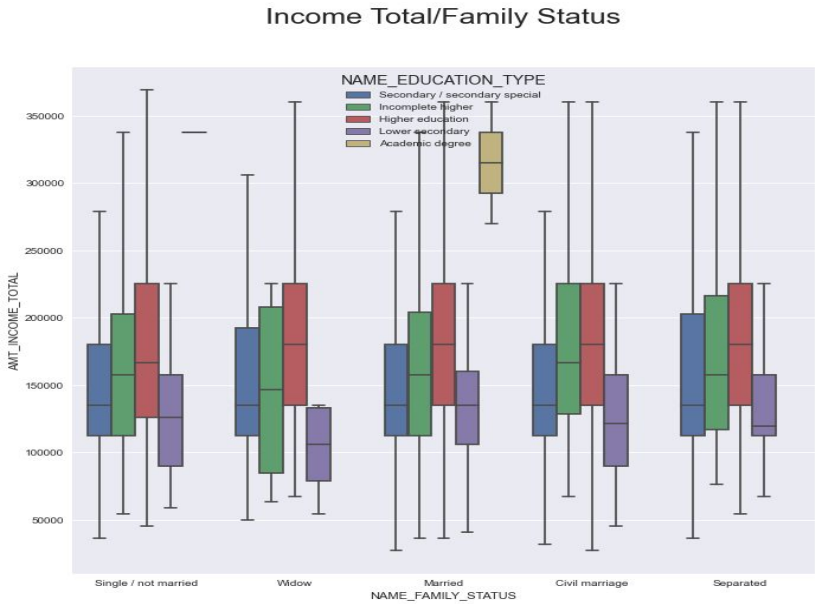
Bivariate analysis -Income range vs Amount credit





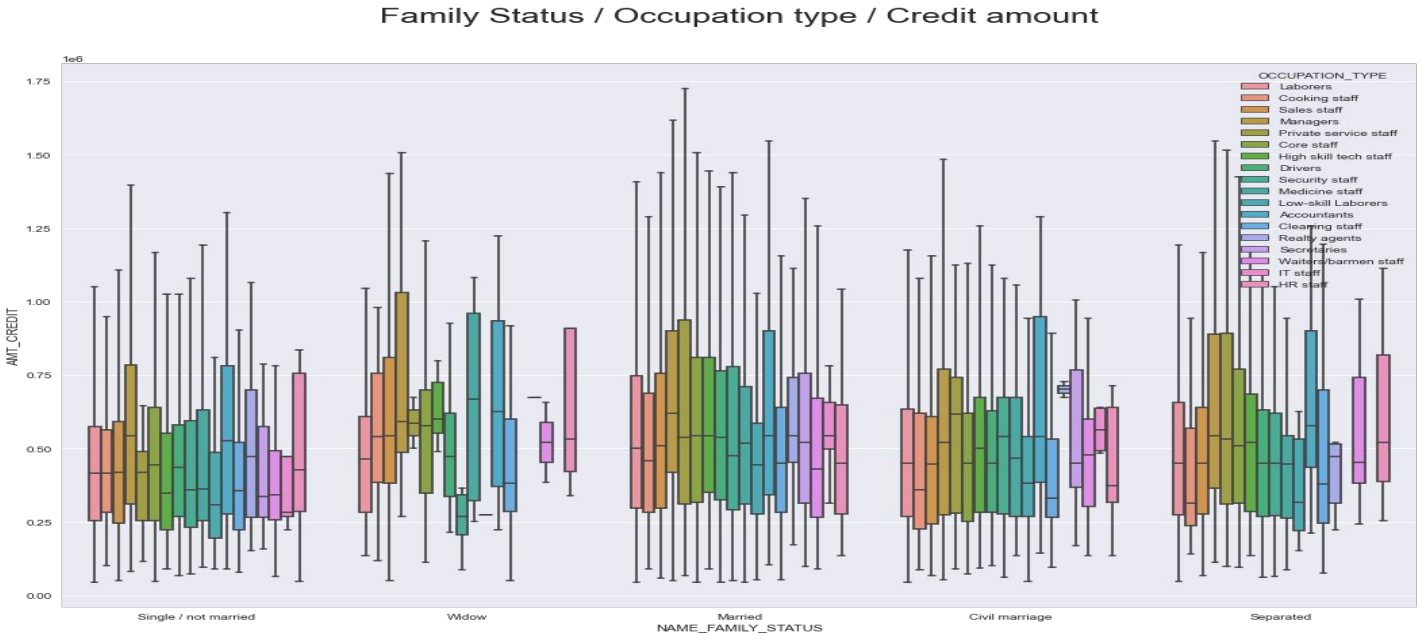
Looks like no much difference between defaulters and non_defaulters here in there distribution.

Segmented bivariate analysis -Income range vs Family status with respect to Education_type



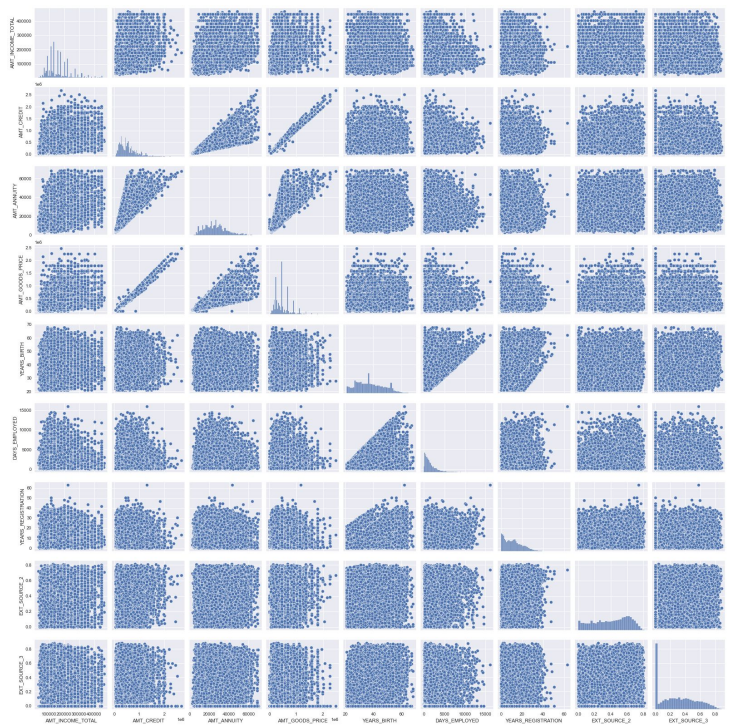
Among defaulters,academic degree holders has highest income total and lower secondary education has lowest income.

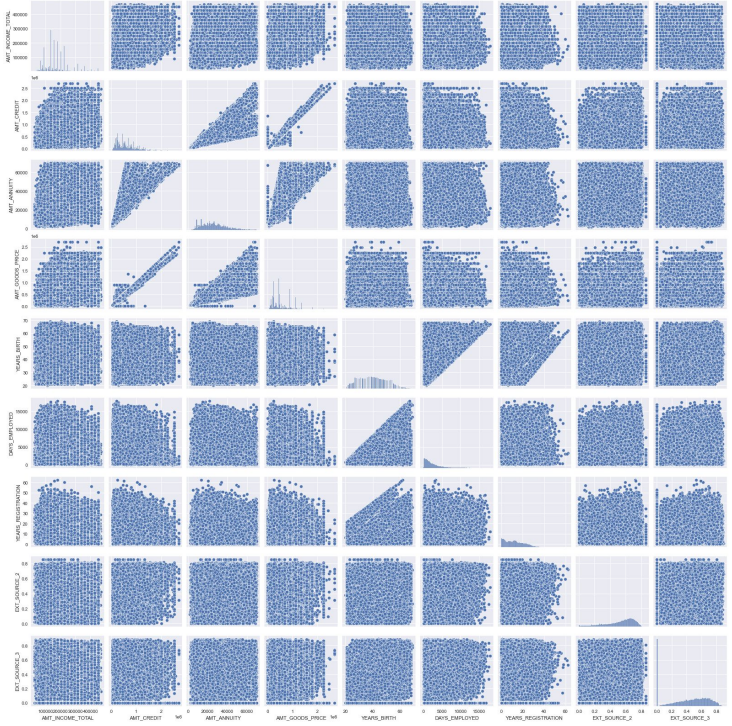
Segmented bivariate analysis -Credit amount vs Family status with respect to Education_type



For defaulters highest and lowest credit amounts are holding by widows. The highest credit amount people are working as accountants or managers.The lowest credit amount people are working as low skilled laborers

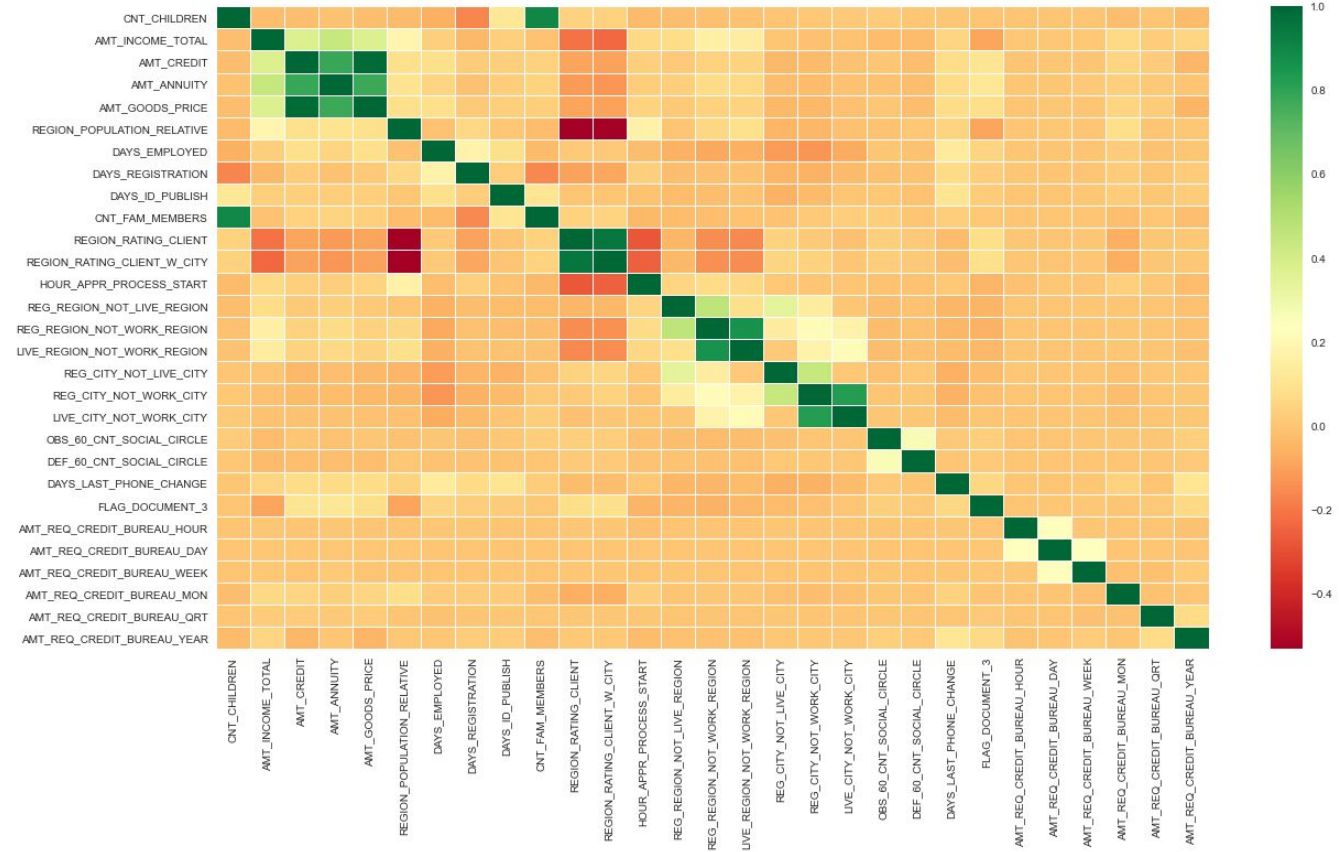
Bivariate analysis of continuous variables - Defaulters vs Repayers





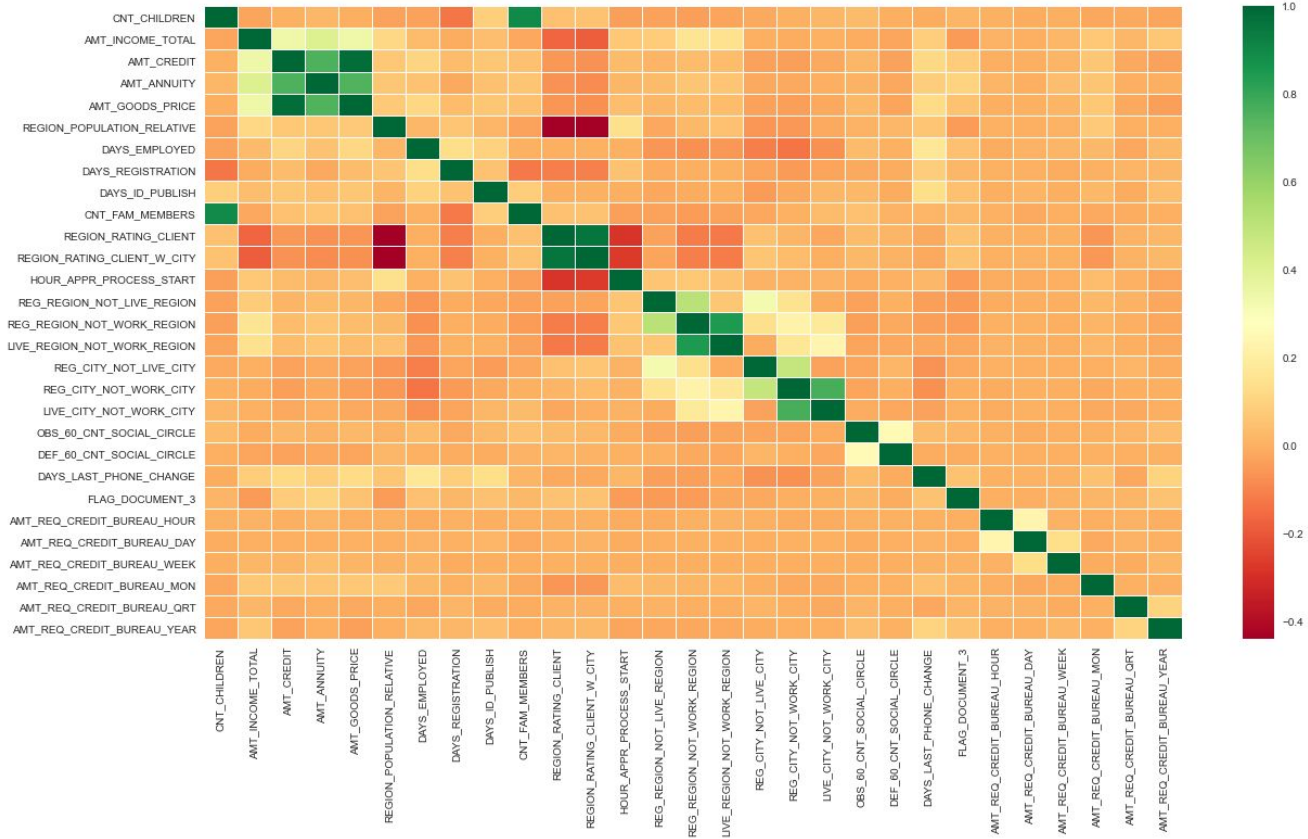
We can see a high correlation between credit amount and goods price which is obvious.

Correlation matrix for repayers



AMT_CREDIT is highly correlated with AMT_GOODS_PRICE,AMT_ANNUITY.We see that repayers have high correlation in number of days employed.

Correlation matrix for defaulters



1. AMT_CREDIT is highly correlated with AMT_GOODS_PRICE which is same for defaulters/repayers.
2. The AMT_ANNUITY correlation with AMT_CREDIT has slightly reduced in defaulters(0.75) when compared to repayers(0.77)

Application_Data insights

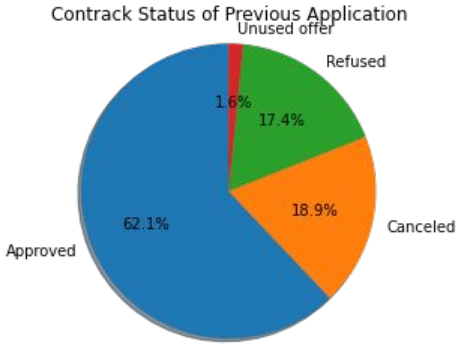
1. The imbalance ratio is too high i.e non defaulters data is 11.4 times more than defaulters
2. Married people , females , people with low income and most of the people living with rented apartments and living with parents ,people with secondary education ,middle aged people are taking more loans.
3. People are preferring more cash loans than revolving,but revolving loans looks safer.
4. people with high income has more credit when compared with low income scale.
5. Academic degree holders are asking more credit amount
6. Widows and people who are on maternity leave tend to be more defaulters under females.
7. The percentage of male defaulters are more when compared with females.
8. Married people are safer as majority of them are non-defaulters while singles default more.
9. Among adults and young adults , default rate is high while middle aged and seniors are more reliable.

ANALYSIS OF PREVIOUS APPLICATION DATA

UNIVARIATE ANALYSIS (CATEGORICAL)

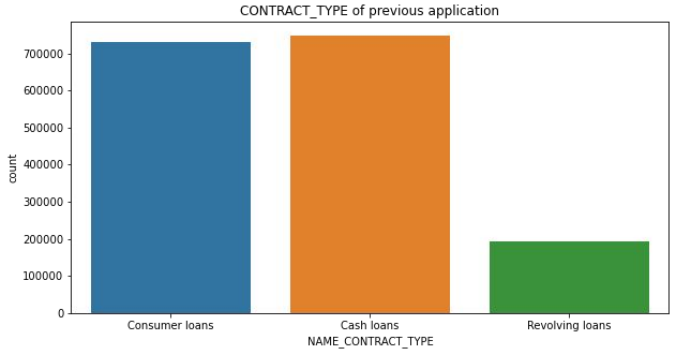
NAME_CONTRACT_STATUS:

Inference: Most applications were approved. Only a very small percentage of applications was unused offer.



NAME_CONTRACT_TYPE:

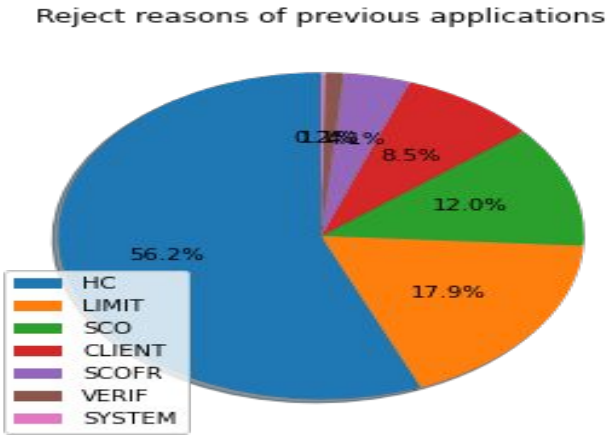
Inference: Number of applications for consumer loans and cash loans are greater than revolving loans.



UNIVARIATE ANALYSIS (CATEGORICAL)

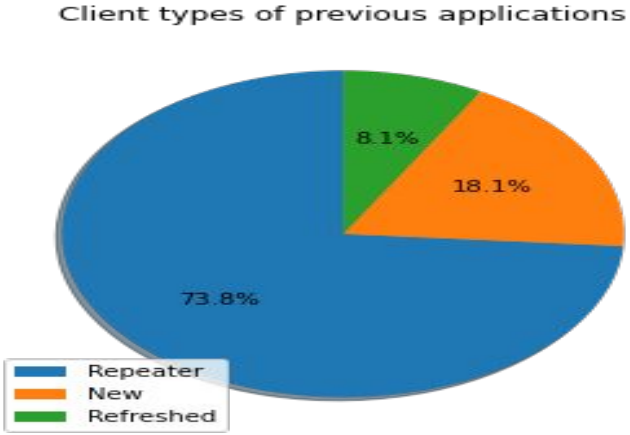
REASONS FOR REJECTING APPLICATIONS:

Inference: Most of application was rejected because of HC.



CLIENT TYPES : Was the client old or new client when applying for the previous application?

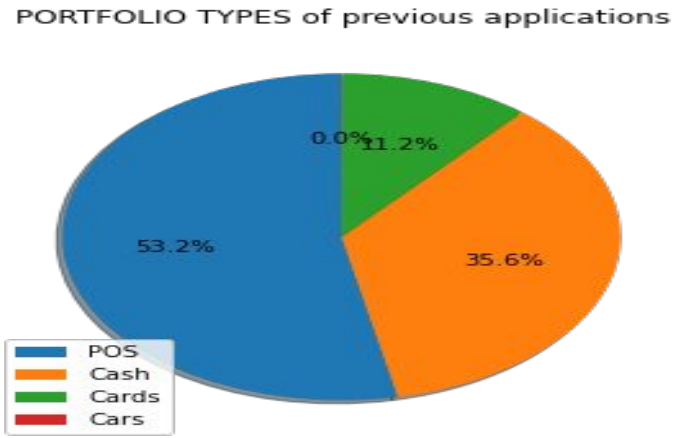
Inference: Most of the client are Repeater.



UNIVARIATE ANALYSIS (CATEGORICAL)

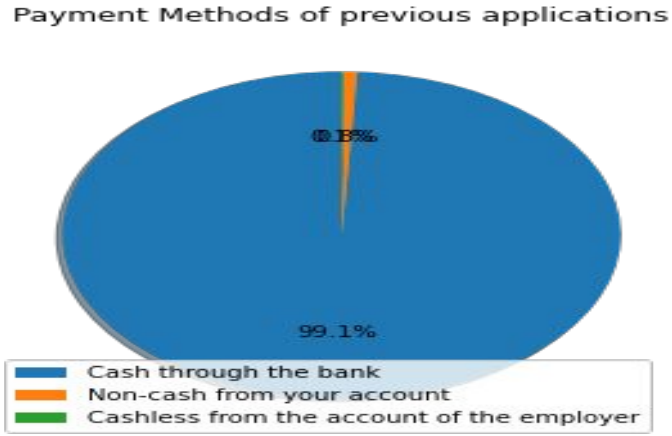
TYPE OF PORTFOLIO:

Inference: Most of the Previous Application was POS.



PAYMENT METHODS:

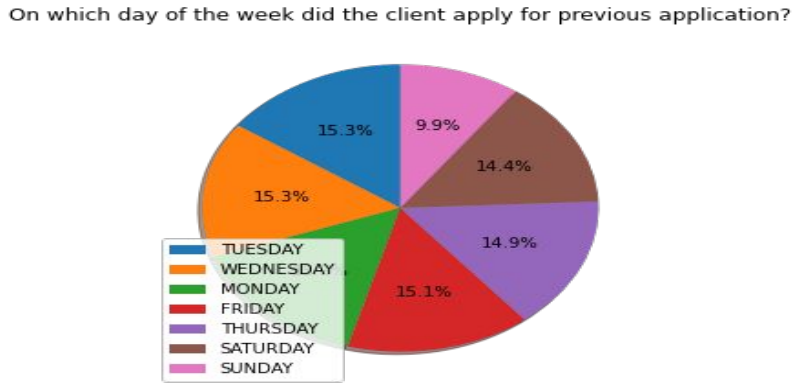
Inference: 99% client paid cash through the bank



UNIVARIATE ANALYSIS (CATEGORICAL)

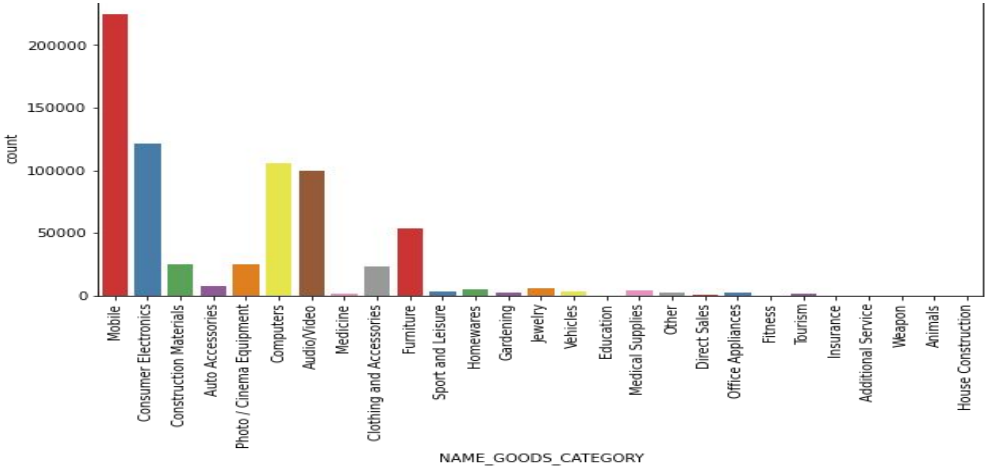
WEEK DAY:

Inference: There were less applications on weekends than on the weekdays.



GOODS CATEGORY:

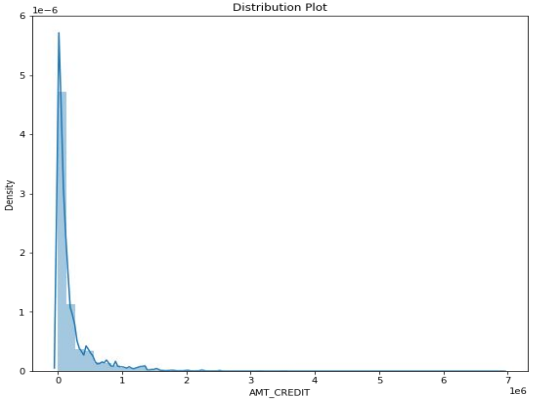
Inference: The majority of clients applied for mobiles, consumer electronics, computers, audio/video and furniture.

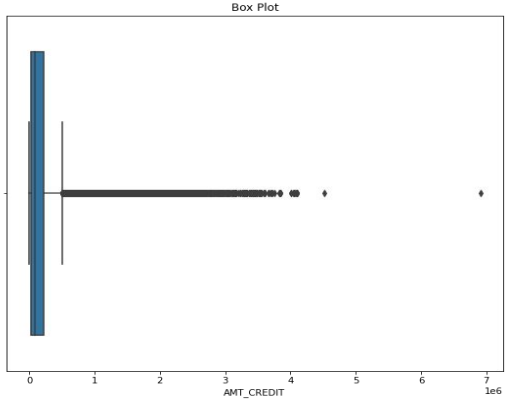


UNIVARIATE ANALYSIS (CONTINUOUS)

Credit Amount:

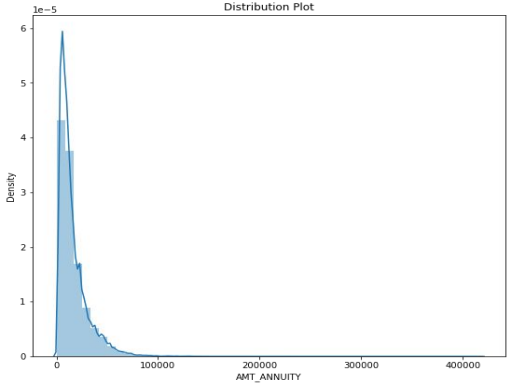
Inference: There are some outliers. Most of the amount of the credit was less than 500000.

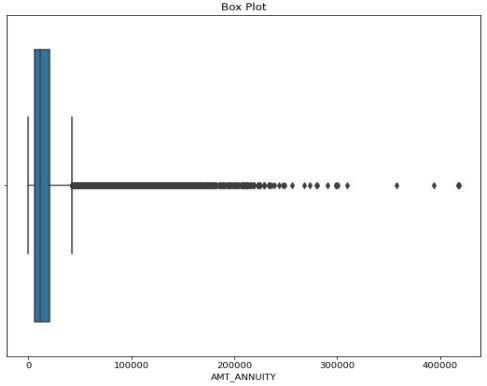




Annuity Amount:

Inference: There are some outliers. Most of the amount of the annuity was less than 500000. Similar to amount of credit amount

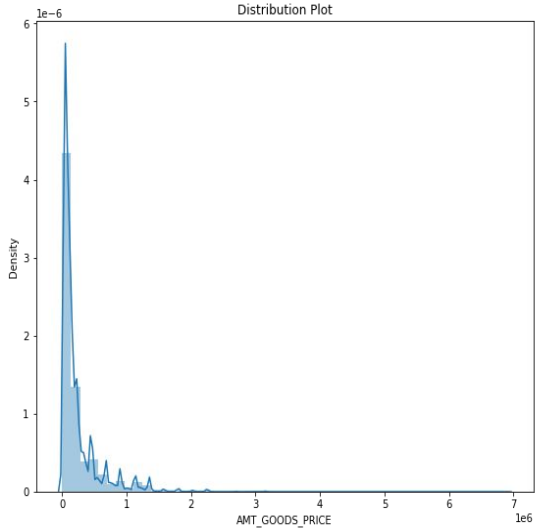


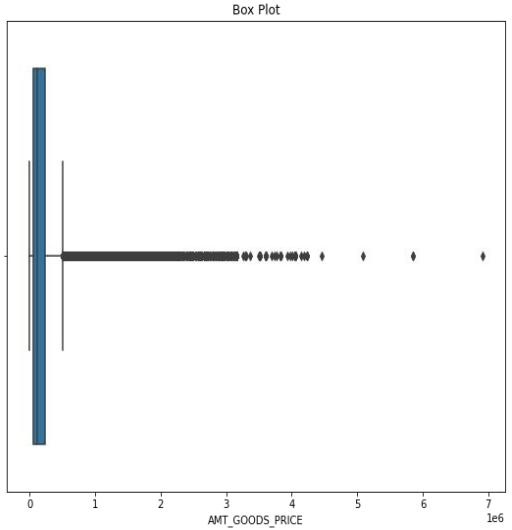


UNIVARIATE ANALYSIS (CONTINUOUS)

Goods Price Amount:

Inference: There are some outliers. Most of the amount of the annuity was less than 500000. Similar to amount of credit amount





BIVARIATE ANALYSIS ON NUMERICAL COLUMN

Annuity of previous application has a very high and positive influence over (Increase of annuity increases below factors):

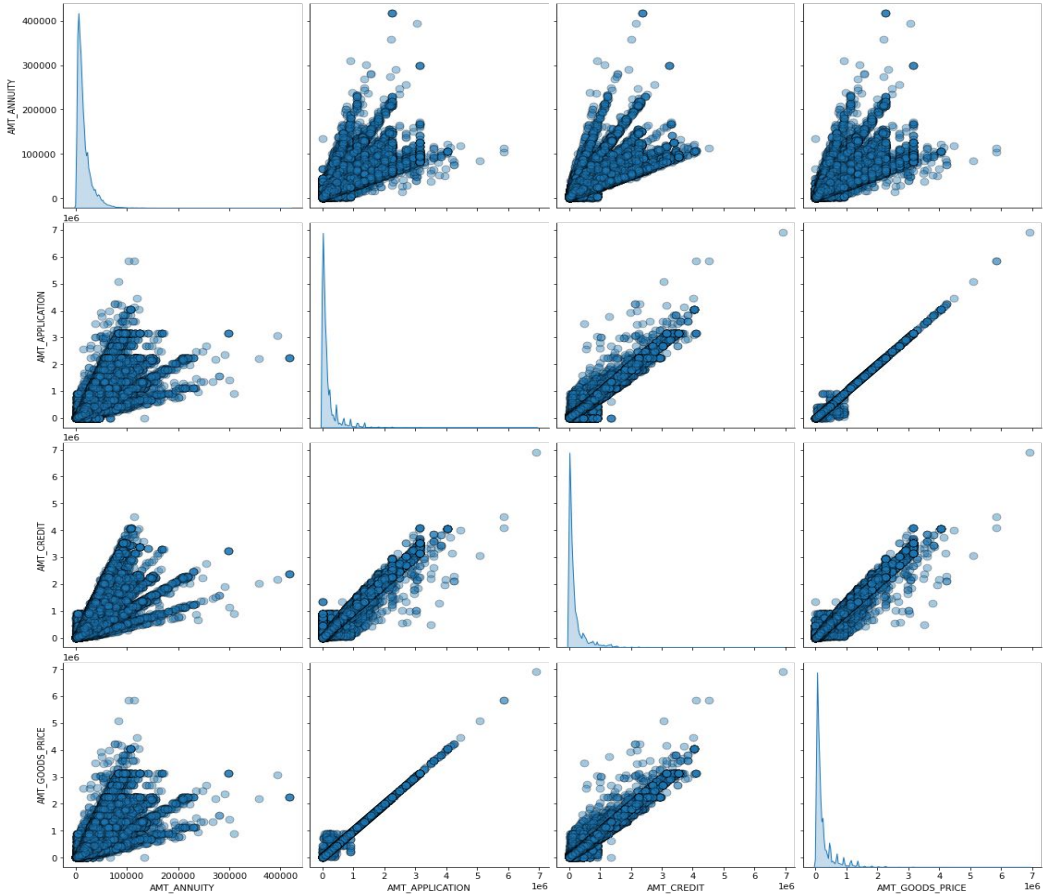
(1) How much credit did client asked on the previous application

(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

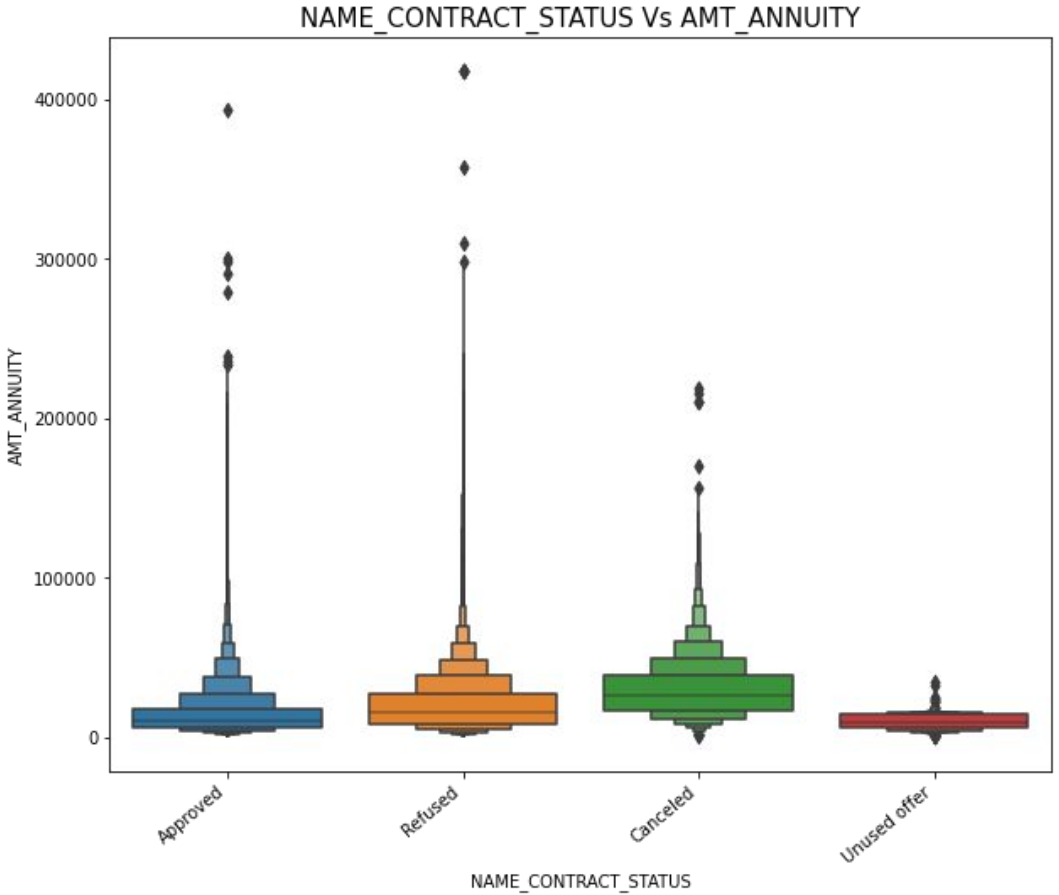
Final credit amount disbursed to the customer previously, after approval is highly influence by the application amount and also the goods price of good that client asked for on the previous application



BIVARIATE ANALYSIS ON CATEGORICAL VS NUMERICAL COLUMNS

INFERENCE:

We can infer that when the AMT_CREDIT is too low, it gets cancelled/unused most of the time.



Previous Application_Data insights

1. Most of the applications in previous data file was approved and a very less percent of applications were unused even after approval.
2. Number of applications for consumer loans and cash loans are greater than revolving loans.
3. Majority of applications was rejected because of HC.
4. Majority of clients which applied for loan was Repeater.
5. 99% client paid cash through the bank (Payment method used by the clients).
6. We can say that number of applications were less on weekend compared to weekdays.
7. Majority of clients applied for ‘ mobiles, consumer electronics, computers, audio/video and furniture’ these category of loans.
8. We can infer that when the AMT_CREDIT is too low, Application of clients gets cancelled/unused most of the time.

ANALYSIS ON APPLICATION AND PREVIOUS APPLICATION DATA AFTER MERGING

UNIVARIATE ANALYSIS (CATEGORICAL)

Previous CONTRACT_STATUS of Defaulters vs Non-Defaulters:

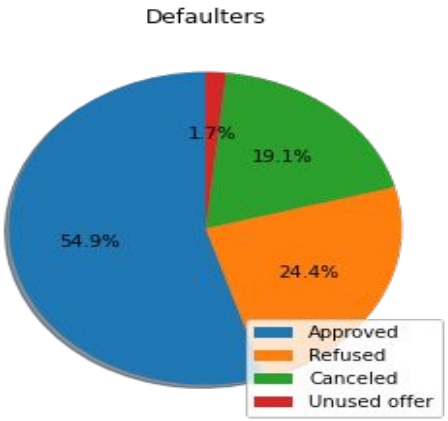
Inference: The percentage of applications from defaulters being refused is higher than that of the non-defaulters.

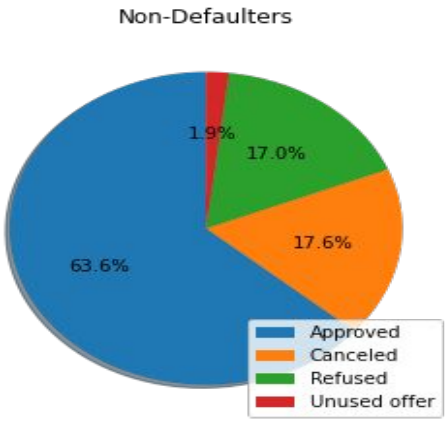
The percentage of applications from non-defaulters being approved is higher than that of the defaulters.

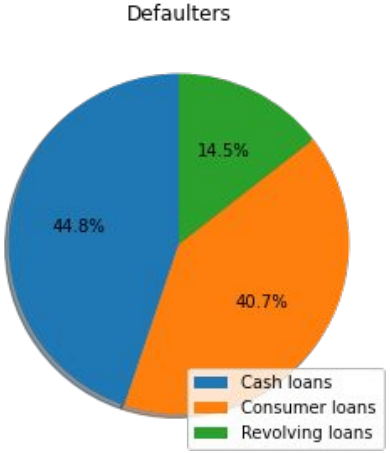
NAME_CONTRACT_TYPE:

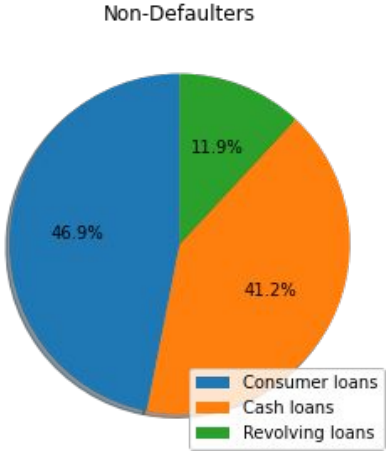
Inference: The percentages of applications from defaulters for cash loans and revolving loans were higher than those of the non-defaulters.

The percentage of applications from non-defaulters for consumer loans was higher than that of the defaulters.







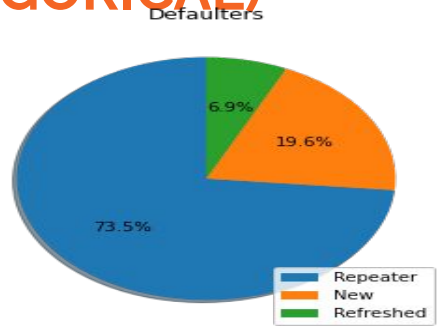


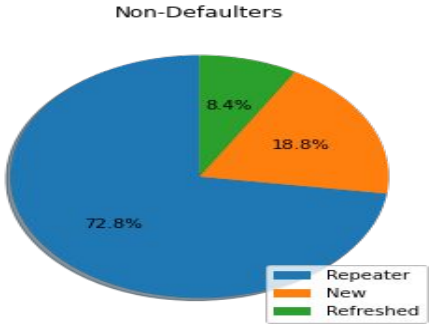
UNIVARIATE ANALYSIS (CATEGORICAL)

CLIENT TYPES : Was the client old or new when applying for the previous application.

Inference: The percentages of defaulters previous applications from new and repeaters clients were higher than those of the non-defaulters.

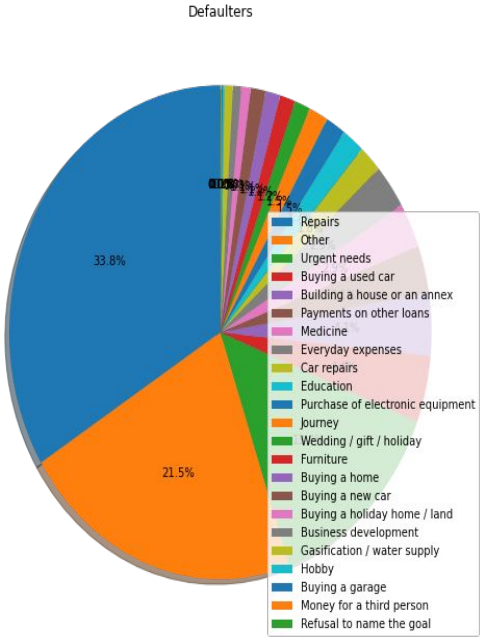
The percentage of non-defaulters previous applications from refreshed clients was higher than those of the defaulters.

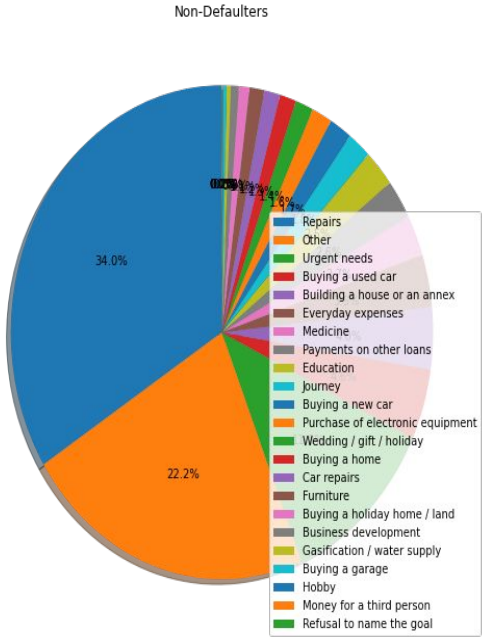




NAME_CASH_LOAN_PURPOSE:

Inference: The percentages of defaulters previous applications refused to name the goal were higher than those of the non-defaulters.





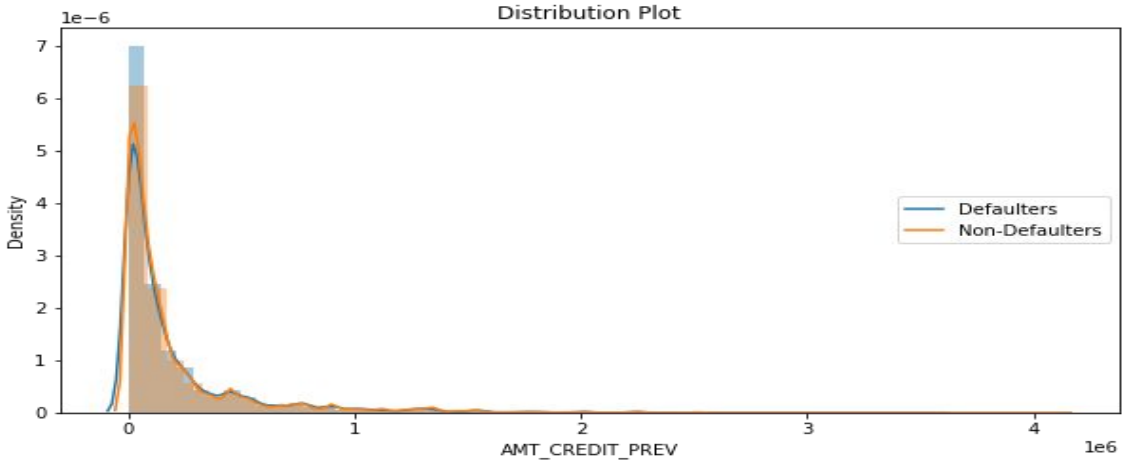
UNIVARIATE ANALYSIS (CONTINUOUS)

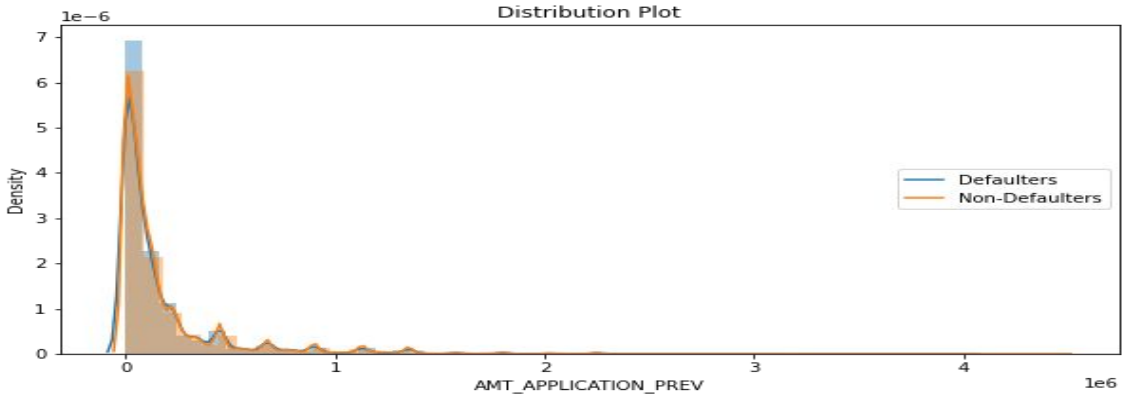
Credit Amount:

Inference: There were some outliers. Most of the amount of the credit was less than 500000. The pattern is the quite similar for defaulters and non-defaulters.

Annuity Amount:For how much credit did client ask on the previous application?

Inference: Similar to the amount of credit, there were some outliers. Most of the amount of the annuity was less than 50000. The pattern is the quite similar for defaulters and non-defaulters.



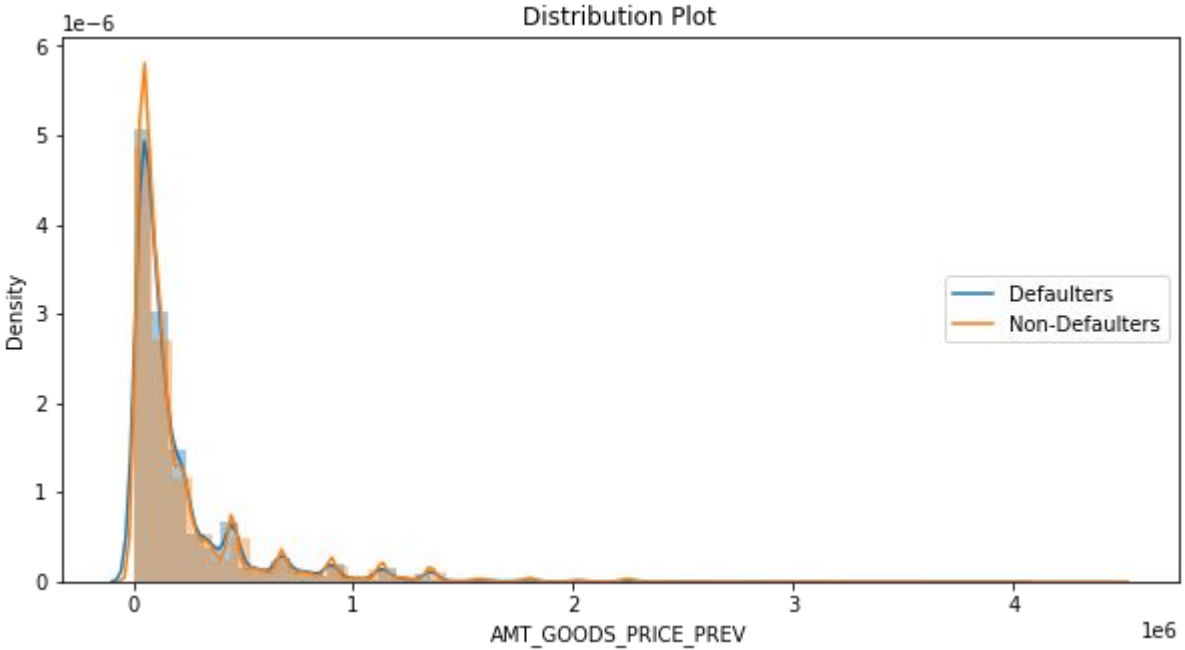


UNIVARIATE ANALYSIS (CONTINUOUS)

Goods Price Amount:

Inference:

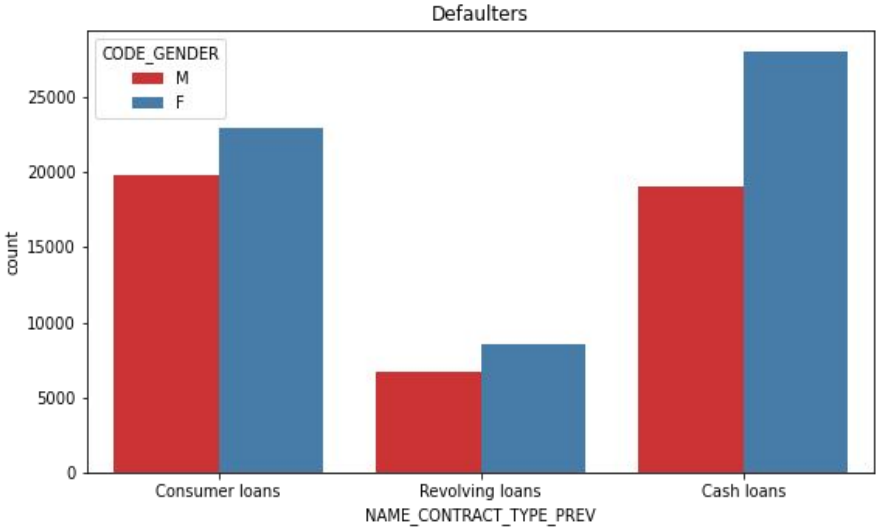
Similar to the credit amount, there were some outliers. Most of the amount of the credit was less than 500000. The pattern is the quite similar for defaulters and non-defaulters.

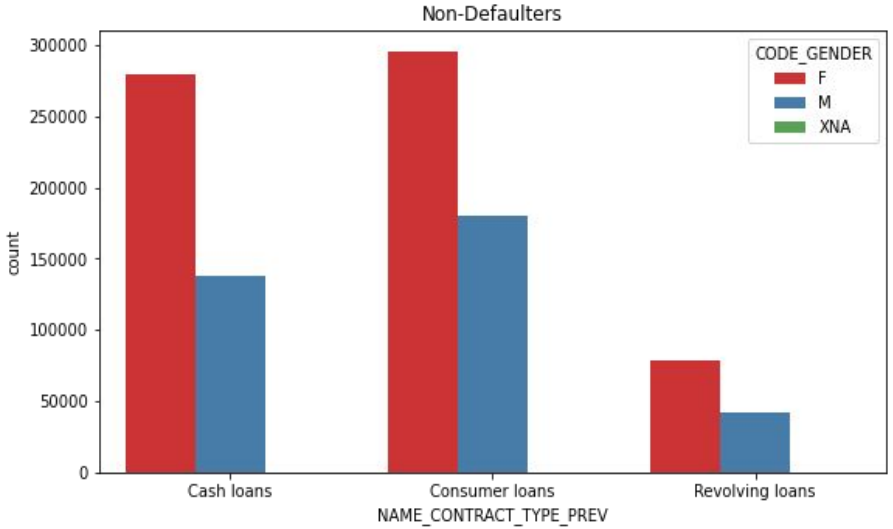


BIVARIATE ANALYSIS (CATEGORICAL VS CATEGORICAL)

Previous Contract & Gender:

Inference: Similar to current loans, for both Defaulters and non-Defaulters, there are more females having all types of previous contracts.



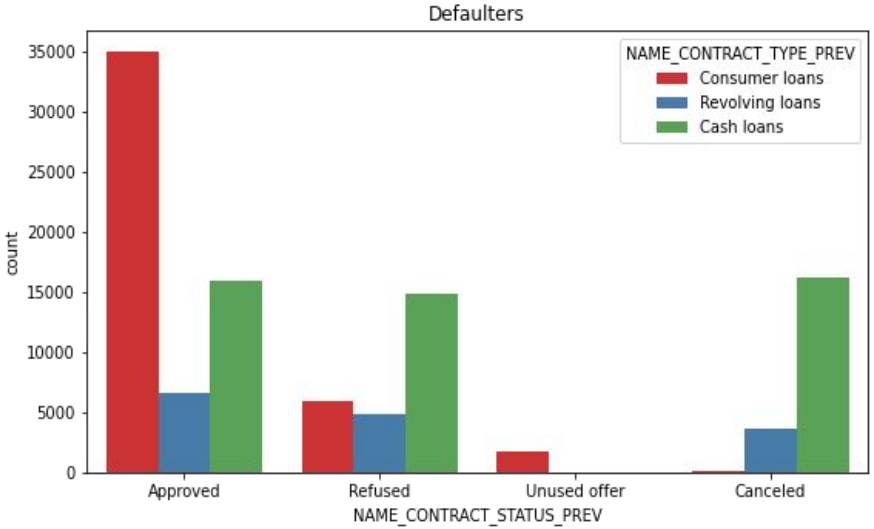


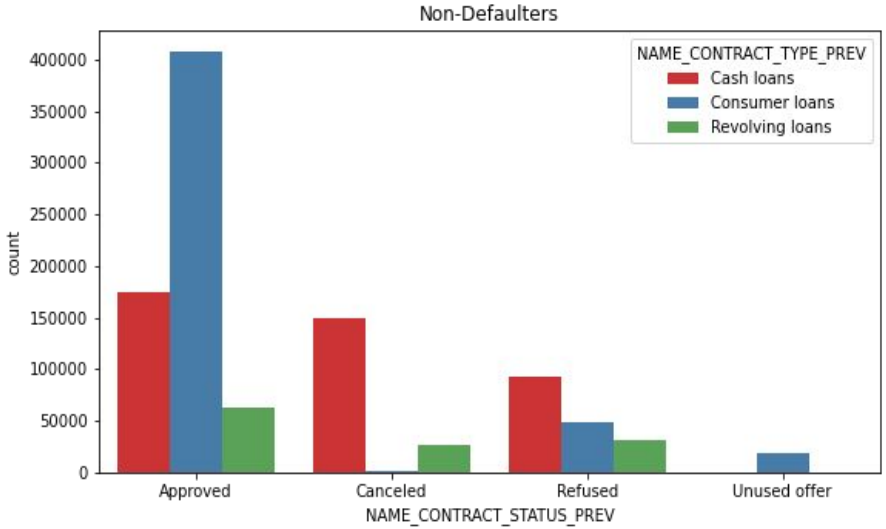
BIVARIATE ANALYSIS (CATEGORICAL VS CATEGORICAL)

Contract types & statuses:

Inference:

There is no significant insight from here.

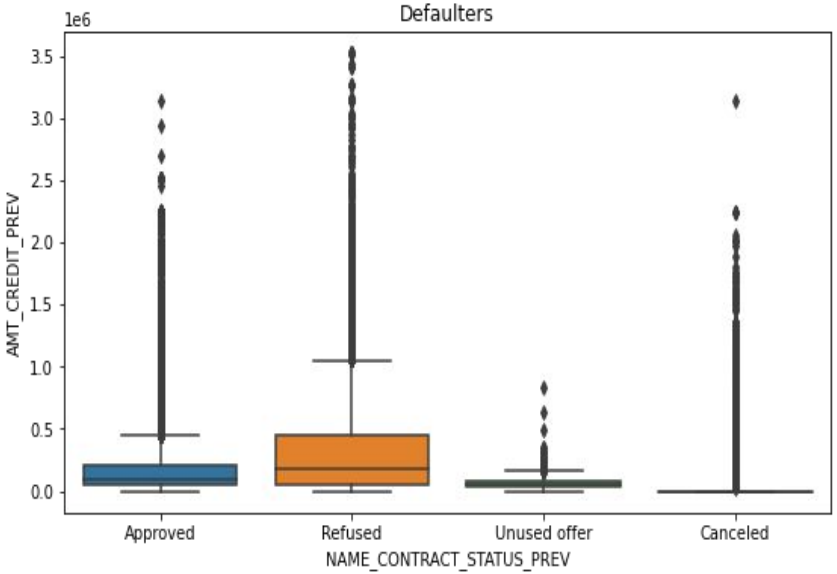


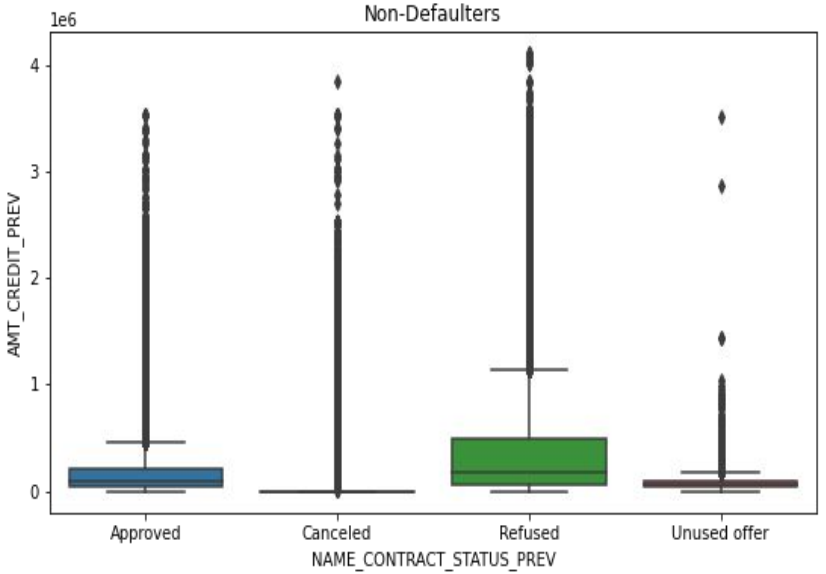


BIVARIATE ANALYSIS (CATEGORICAL VS CONTINUOUS)

Previous Contract Status & Amount of credit:

Inference: Similar for both defaulters and non-defaulters, applications being refused had higher credits.

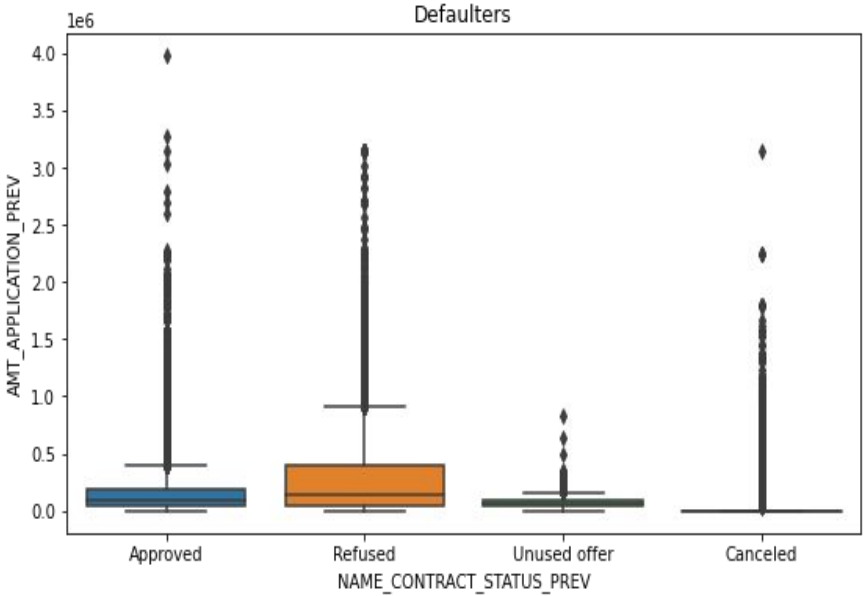


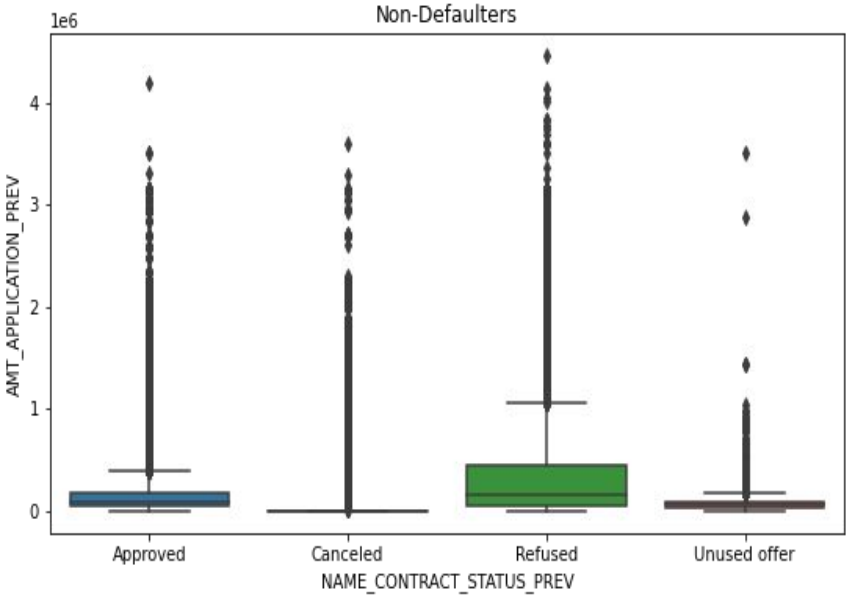


BIVARIATE ANALYSIS (CATEGORICAL VS CONTINUOUS)

Previous Contract Status & Amount of credit clients asked for:

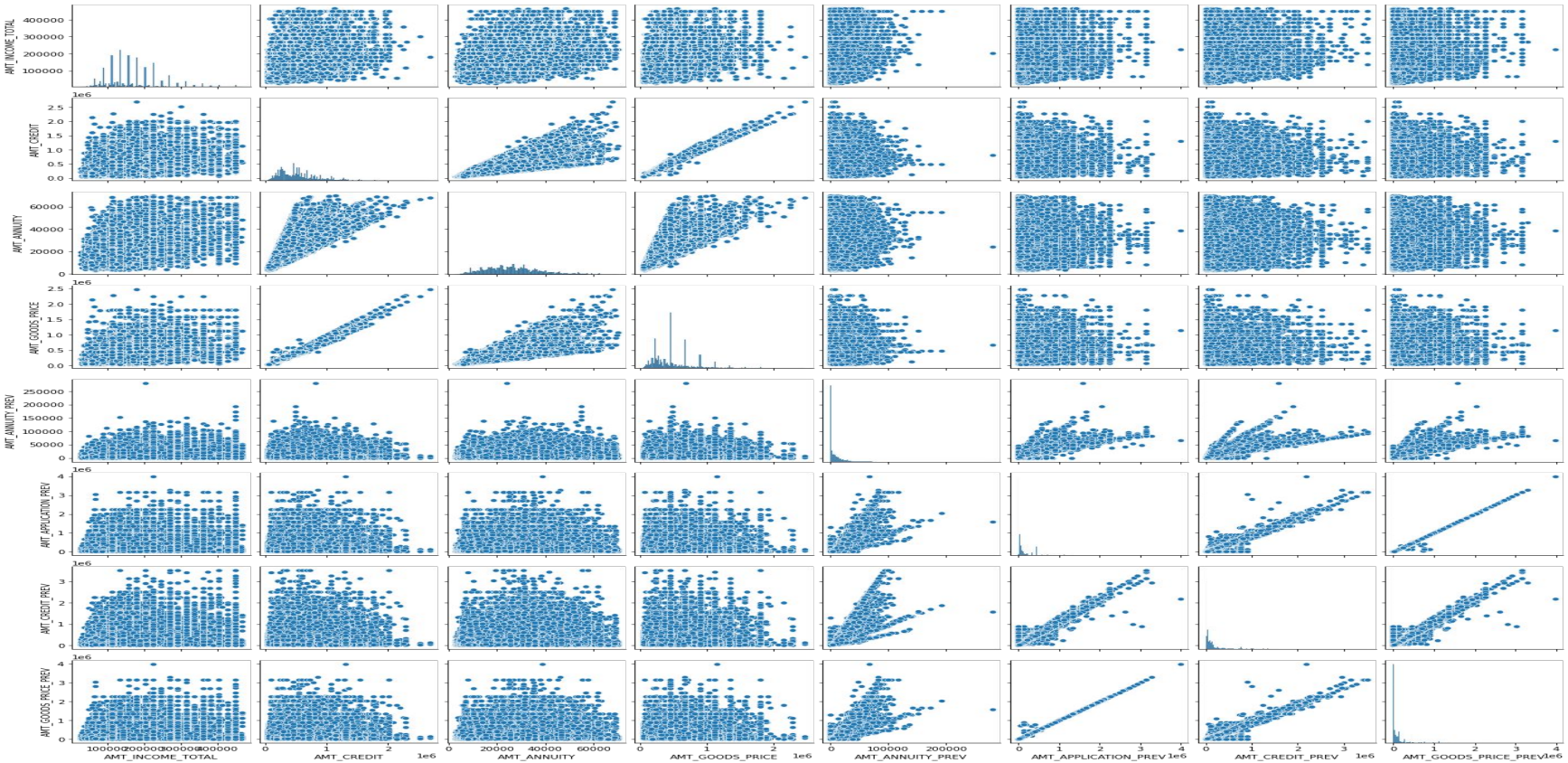
Inference: Similar for both defaulters and non-defaulters, applications being refused had higher credits.





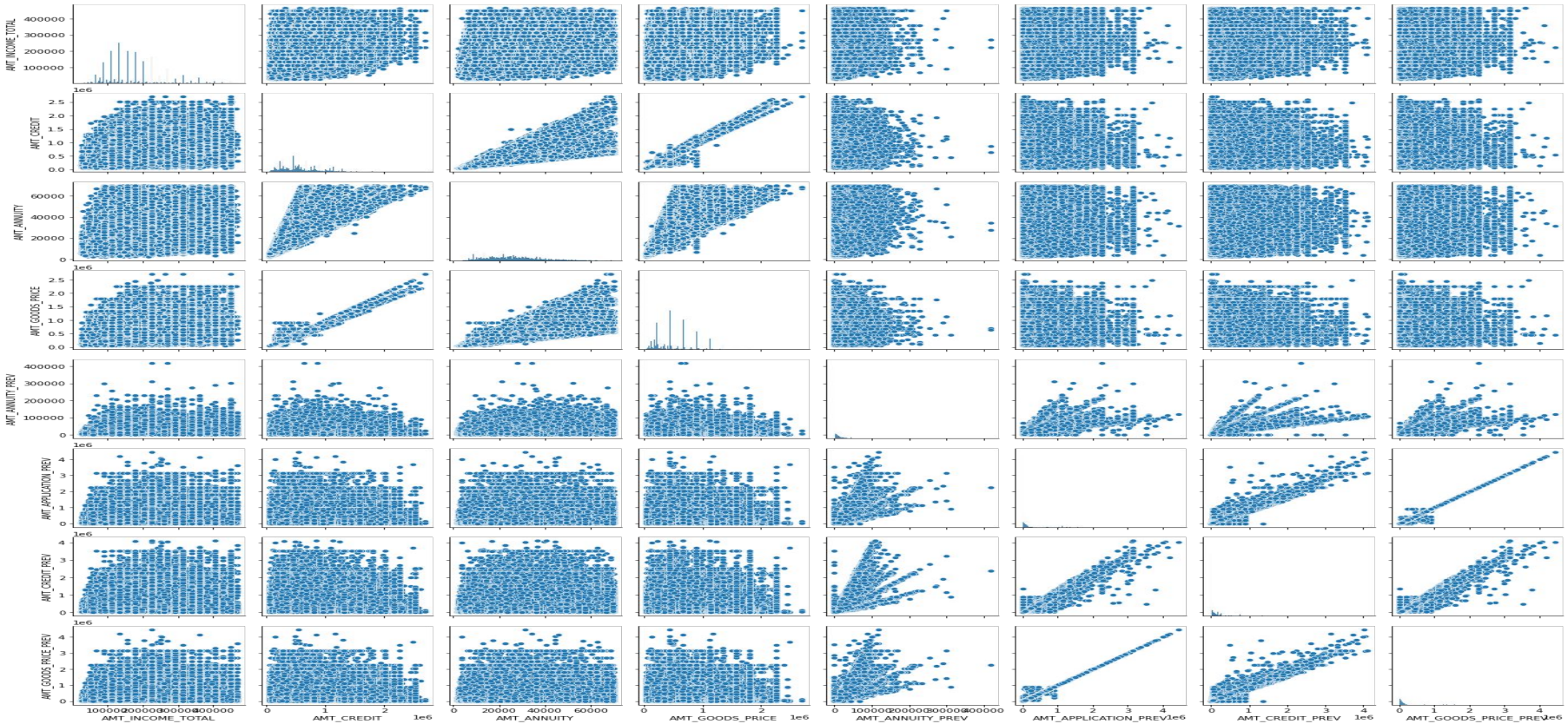
BIVARIATE ANALYSIS (CONTINUOUS VS CONTINUOUS)

Pairplot of multiple Numerical columns for Defaulters



BIVARIATE ANALYSIS (CONTINUOUS VS CONTINUOUS)

Pairplot of multiple Numerical columns for Non- Defaulters



BIVARIATE ANALYSIS (CONTINUOUS VS CONTINUOUS)

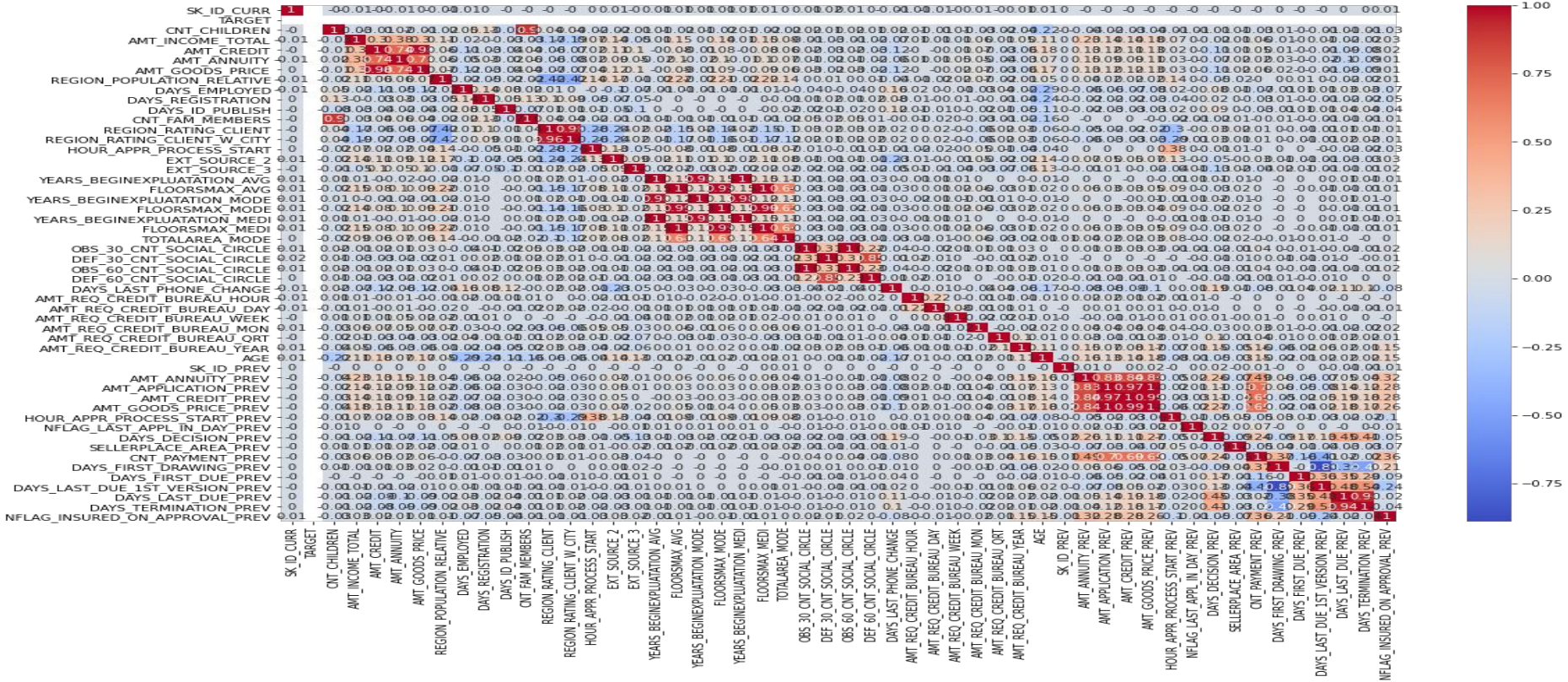
Inferences from Pair Plot of numerical columns of Defaulters and Non-Defaulters:

From the pair plots above,

- It can be seen that there is a high correlation between credit amount and goods price.
- Similarly, there are a high correlation between previous credit amount and previous goods price .
- Also there is high correlation between previous credit applied by clients and previous goods price.

CORRELATION BETWEEN DIFFERENT VARIABLES

Correlation for clients with payment difficulties (Defaulters):



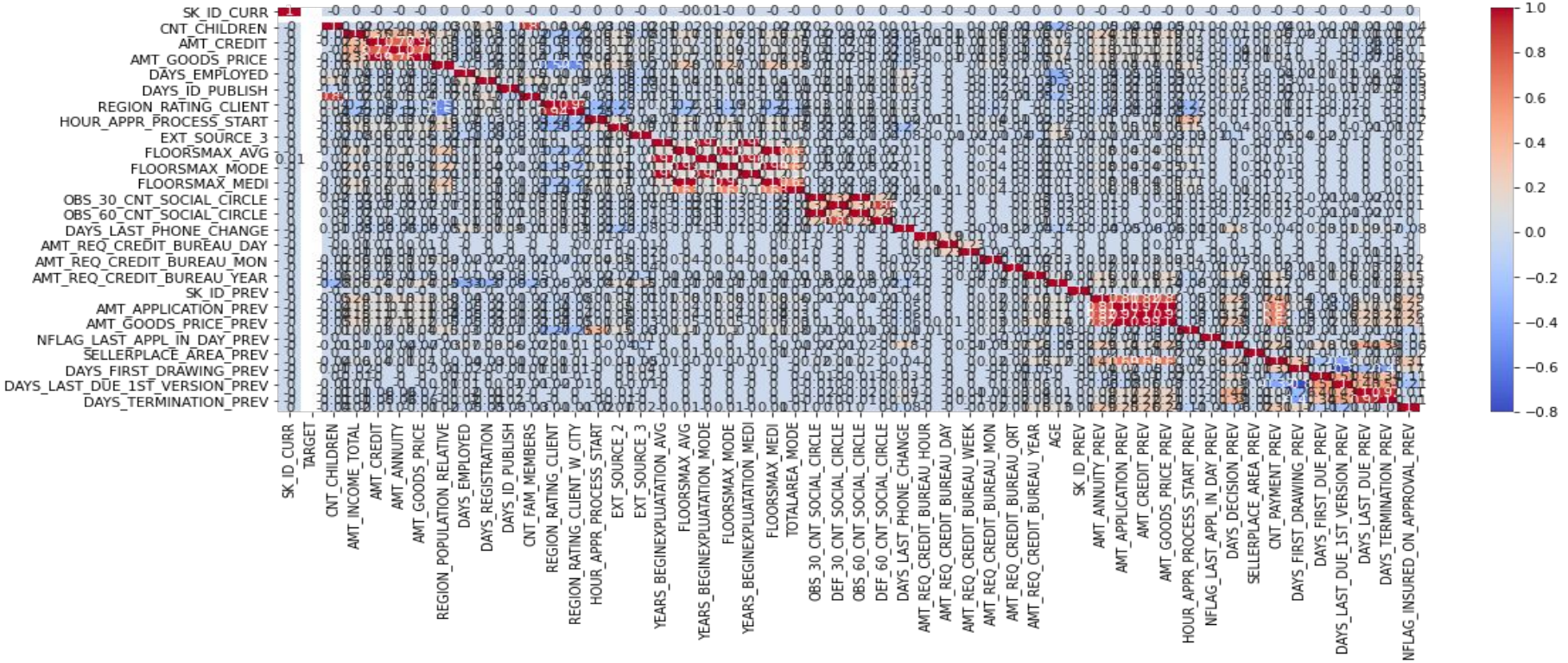
CORRELATION BETWEEN DIFFERENT VARIABLES

Top 10 Correlation for clients with payment difficulties (Defaulters):

AMT_APPLICATION_PREV	AMT_GOODS_PRICE_PREV	1.00
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MEDI	1.00
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	1.00
FLOORSMAX_AVG	FLOORSMAX_MEDI	1.00
AMT_CREDIT_PREV	AMT_GOODS_PRICE_PREV	0.99
FLOORSMAX_AVG	FLOORSMAX_MODE	0.99
FLOORSMAX_MODE	FLOORSMAX_MEDI	0.99
YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_MEDI	0.98
AMT_CREDIT	AMT_GOODS_PRICE	0.98
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MODE	0.98

CORRELATION BETWEEN DIFFERENT VARIABLES

Correlation for clients with payment difficulties (Non-Defaulters):



CORRELATION BETWEEN DIFFERENT VARIABLES

Top 10 Correlation for clients with payment difficulties (Defaulters):

OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	1.00
FLOORSMAX_AVG	FLOORSMAX_MEDI	1.00
AMT_APPLICATION_PREV	AMT_GOODS_PRICE_PREV	1.00
AMT_CREDIT_PREV	AMT_GOODS_PRICE_PREV	0.99
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MEDI	0.99
FLOORSMAX_MODE	FLOORSMAX_MEDI	0.99
AMT_CREDIT	AMT_GOODS_PRICE	0.99
FLOORSMAX_AVG	FLOORSMAX_MODE	0.99
YEARS_BEGINEXPLUATATION_AVG	YEARS_BEGINEXPLUATATION_MODE	0.97
AMT_APPLICATION_PREV	AMT_CREDIT_PREV	0.97

Merged Data Insights

The percentage of applications from defaulters being refused is higher than that of the non-defaulters and the percentage of applications from non-defaulters being approved is higher than that of the defaulters.

The percentages of applications from defaulters for cash loans and revolving loans were higher than those of the non-defaulters. And the percentage of applications from non-defaulters for consumer loans was higher than that of the defaulters.

The percentages of defaulters (previous applications) from new and repeaters clients were higher than those of the non-defaulters.The percentage of non-defaulters (previous applications) from refreshed clients was higher than those of the defaulters.

Similar for both defaulters and non-defaulters, applications being refused had higher credits.

CONCLUSION (Driving factors):

Non-Defaulters:

- More clients with high income
- More middle-age clients and seniors
- More married people
- Higher education
- More ‘approved’ previous applications
- More ‘Consumer Loans' previous applications

Defaulters:

- More clients with low income
- More adults and young adults
- More single people
- Secondary education
- More ‘Refused’ previous applications
- More ‘Revolving Loans’ previous application

Loan Clients in general:

- Laborers occupation
- Secondary education
- Married people
- Middle age
- Low income
- Applied the loans for goods price less than 2,000,000

THANK YOU