EDA CASE STUDY Submitted by: Anush Kumar & Mitthi Jyoti Sharma

CASE SCENARIO

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants are capable of repaying the loan are not rejected.When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

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

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

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

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

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 (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.

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 utilise this knowledge for its portfolio and risk assessment.

Data Analysis

i. Missing Data:

The percentage of missing values in the data set are as follow:

APPLICATION DATA SET	
VARIABLE NAME	PERCENTAGE
LANDAREA_AVG	59.376738
APARTMENTS_MODE	50.749729
HOUSETYPE_MODE	50.176091
OCCUPATION_TYPE	31.345545
AMT_REQ_CREDIT_BUREAU_DAY	13.501631
AMT_REQ_CREDIT_BUREAU_YEAR	13.501631
AMT_REQ_CREDIT_BUREAU_WEEK	13.501631
AMT_REQ_CREDIT_BUREAU_MON	13.501631
AMT_REQ_CREDIT_BUREAU_QRT	13.501631
AMT_GOODS_PRICE	0.090403
AMT_ANNUITY	0.003902
CNT_FAM_MEMBERS	0.000650
DAYS_LAST_PHONE_CHANGE	0.000325

PREVIOUS APPLICATION DATA SET	
VARIABLE NAME	PERCENTAGE
AMT_DOWN_PAYMENT	53.348211
AMT_GOODS_PRICE	22.980235
AMT_ANNUITY	22.221491
CNT_PAYMENT	22.221205

Treatment of missing values: The missing values are replaced with the mode value of the respective column. Rows with ‘XNA’ are dropped from the data set and required changes in the data type is also done.

ii. Outliers

The number of outliers in the application data set are as follow:

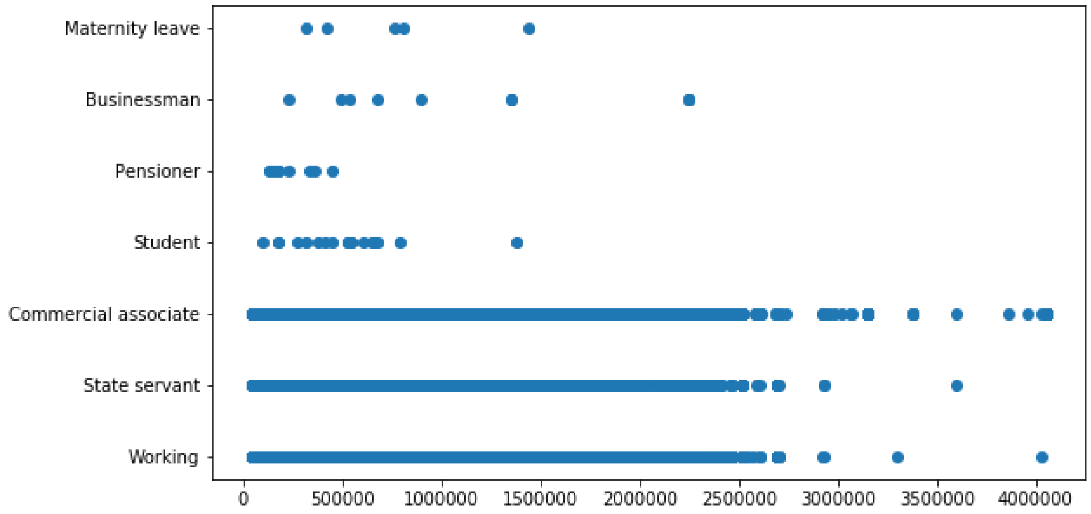
AMT_CREDIT (Final Credit amount): 5426

AMT_INCOME_TOTAL (Income of the client): 8752

CNT_FAM_MEMBERS (No of family members of the client): 3929

Scatter plots for detecting outliers:

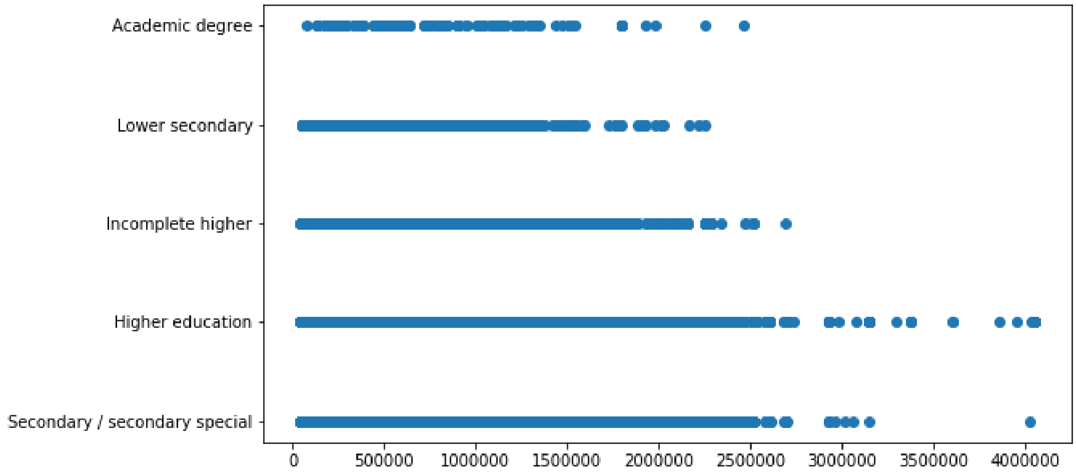
a) Between Amount credit and Income type of the client



(The x-axis represents the amount credit and y-axis represents the income type)

From the graph, one can notice that the amount credited for commercial associate, state servant and working individuals are frequent up to 25 lakhs. Very few loans are credited for amount beyond 25 lakhs for these three categories.

b) Amount credited and education type



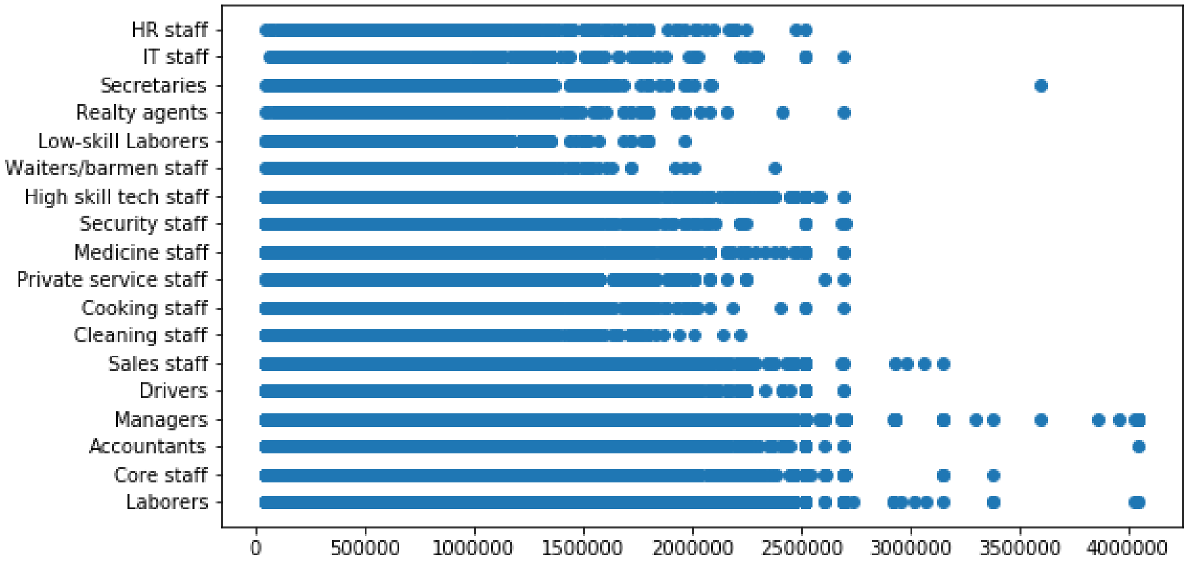
(The x-axis represents the amount credit and y-axis represents the education type)

From this graph, one can notice that the amount credited is frequent for secondary/secondary special, higher education, and is almost up to 25 lakhs, beyond which we notice rare cases of amount credited.

For those with incomplete higher, the maximum amount credited frequently is within 20 lakhs. Very few are given beyond 20 lakhs.

For those with lower secondary, the maximum amount credited frequently is within 15 lakhs and for those with academic degree, amount credited is less frequent compared to others and the maximum credited is 25 lakhs.

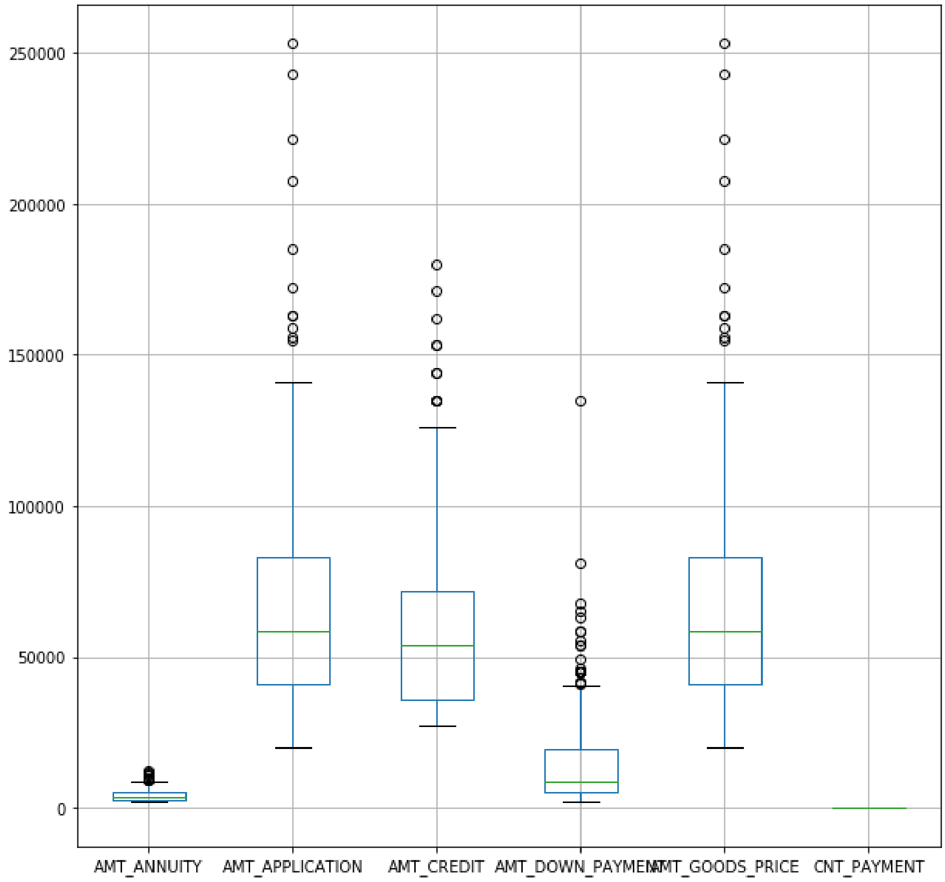
c) Amount credited and occupation type



Laborers, Core staff, accountants, managers, sales staff, medicine staff, high skill tech staff are credit loan frequently and the maximum amount credited to them 25lakhs, beyond which the frequency reduces and few outliers are noticed (rare cases of loan credited).

For others, the maximum amount is within 30 lakhs.

In previous application data set, the outliers are detected using boxplots for particular variables. They are as follows:



In the above box plot, it can be noticed that for amount of application and goods prices, there are outliers beyond 150,000. Whereas for amount credited, the outliers are noticed approximately from 130,000. For Down payment, the outliers are noticed from approximately 40000 and for amount of annuity, the outliers are within 50,000 and very few outliers exist for amount annuity.

iii. Data Imbalance

In the current application data set, the data imbalance is approximately 11:1.

Not-Defaulters-230,298
Defaulters-21,835

In previous application data set, the data imbalance is approximately 14:1.

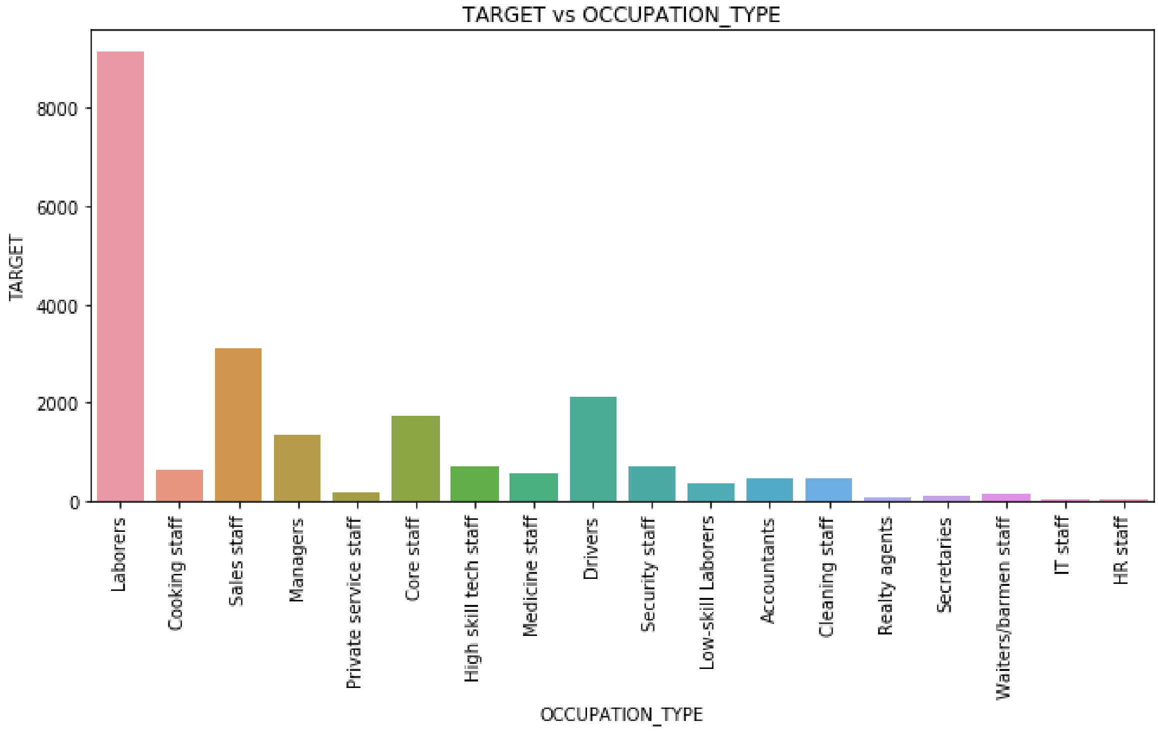
Not Defaulters- 271
Defaulters- 19

iv. Univariate Analysis

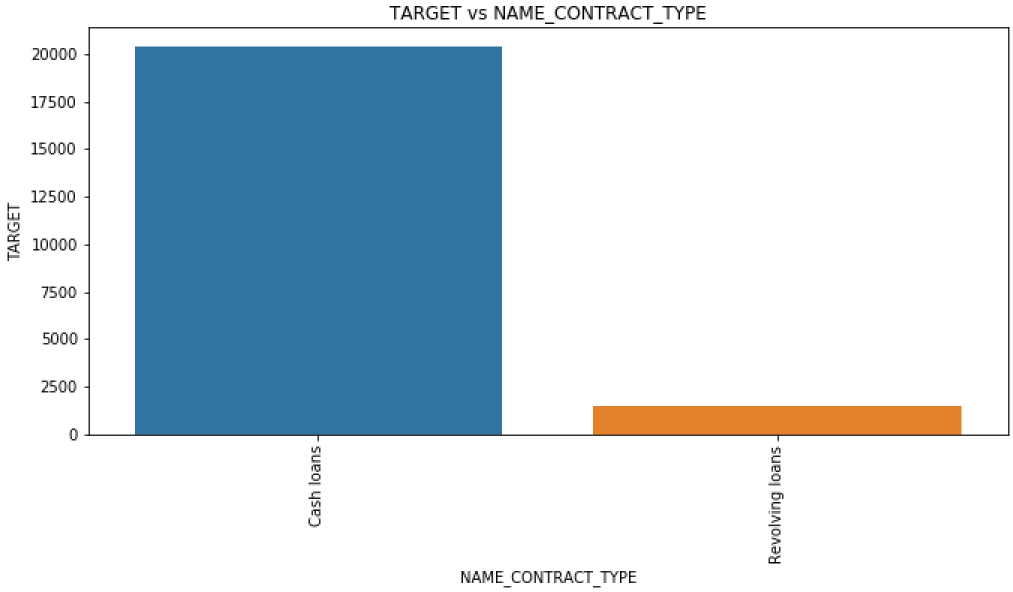
Current Application Data set:

I. **For defaulters:**

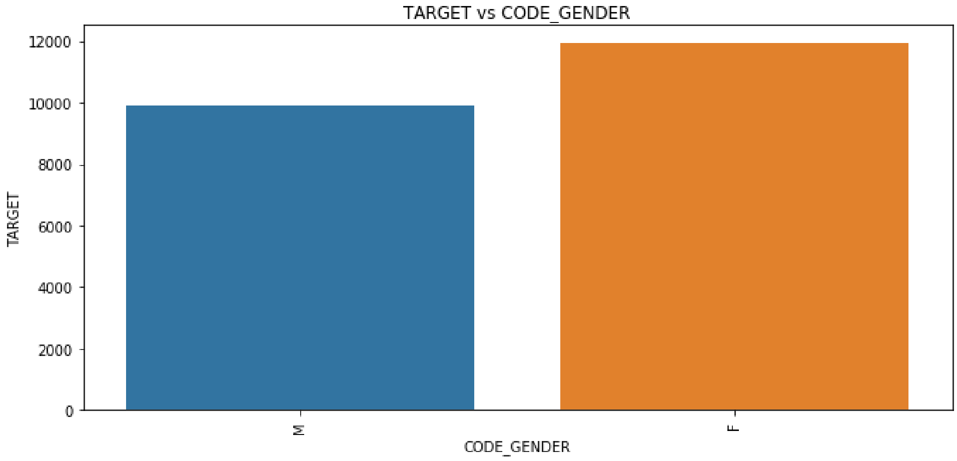
The following are the bar graphs representing defaulters and various variables.



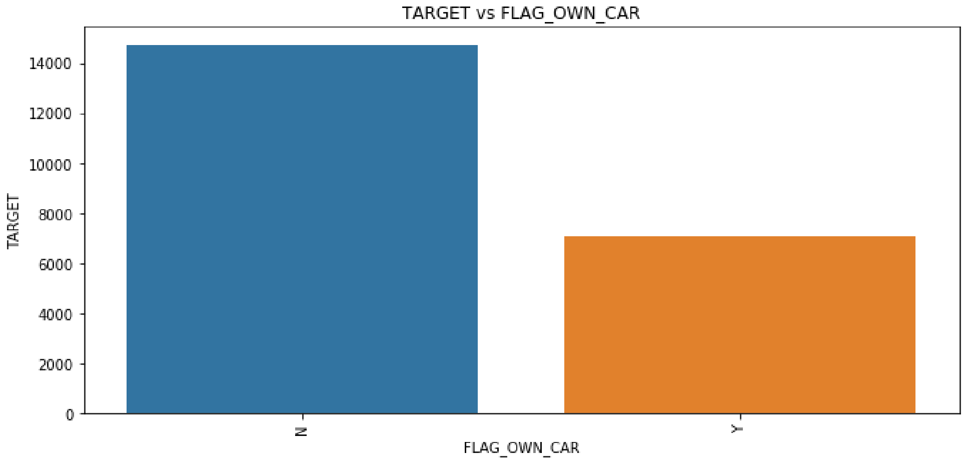
From the graph above, we can notice that the laborers are significantly higher defaulters, followed by sales staff, drivers, core staff and so on.



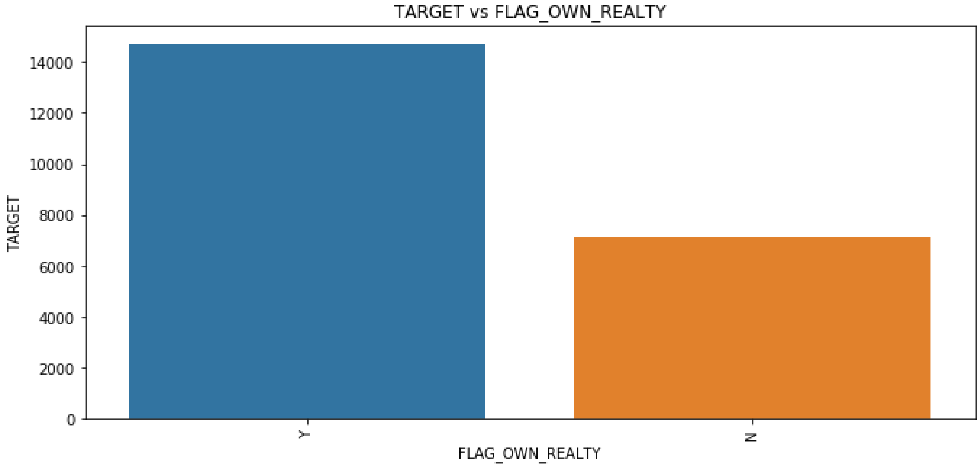
From the above graph we can notice that those take cash loans default significantly higher than those who take revolving loans.



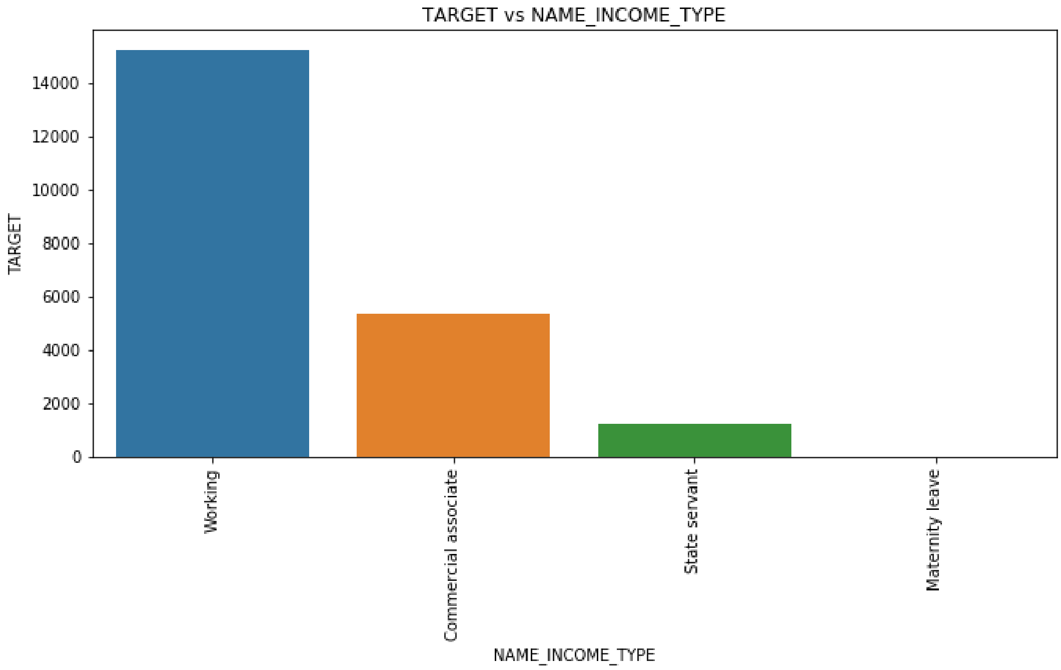
From the above graph, it can be noticed that females are higher defaulters than males.



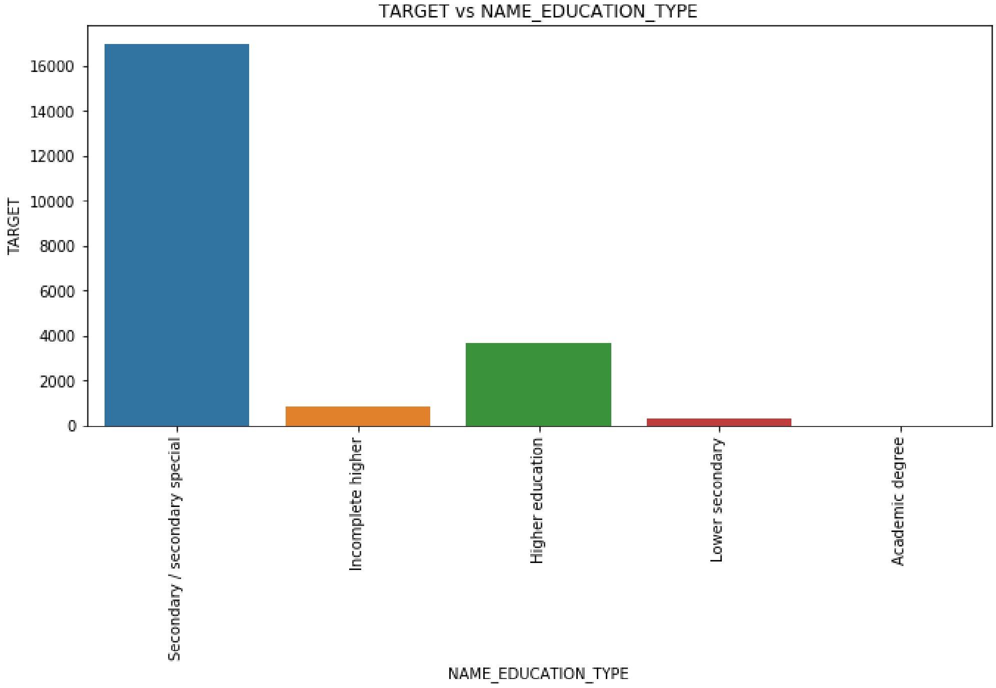
From the above graph, it can be noticed that those who don't own cars default higher than those who own cars.



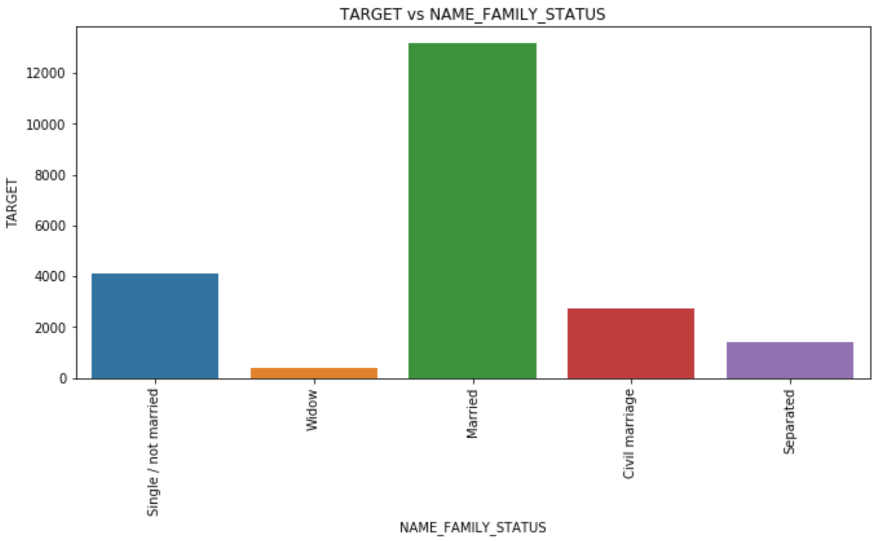
From the above graph, it can be noticed that those who own house or flats default higher than those who do not own any house or flat.



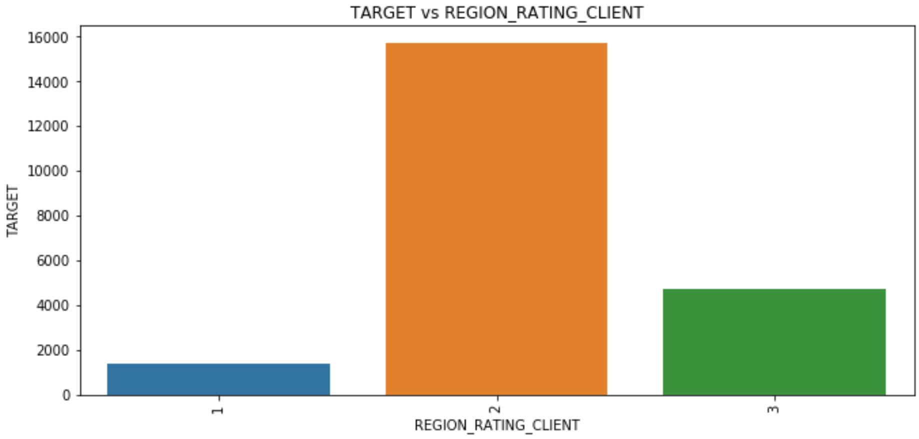
From the above graph, it can be noticed that defaulters among working individuals is very high.



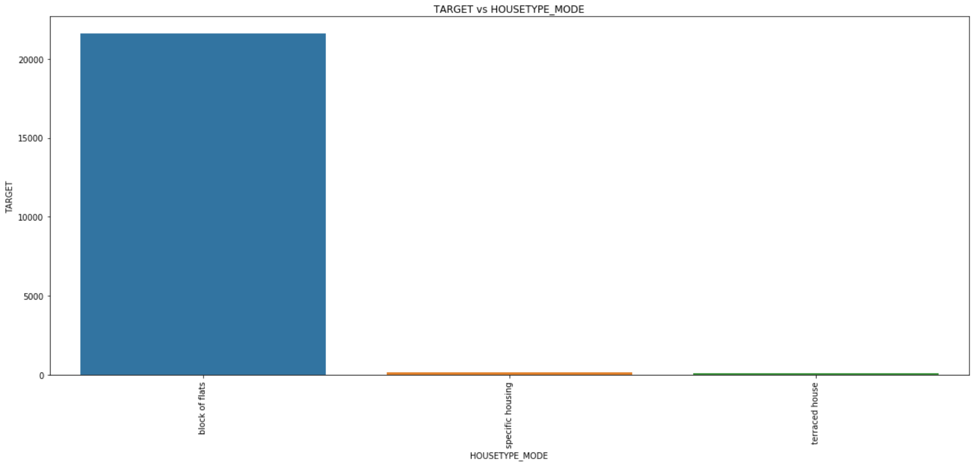
From the above graph, it can be noticed that defaulters among secondary/secondary special educated type are very high.



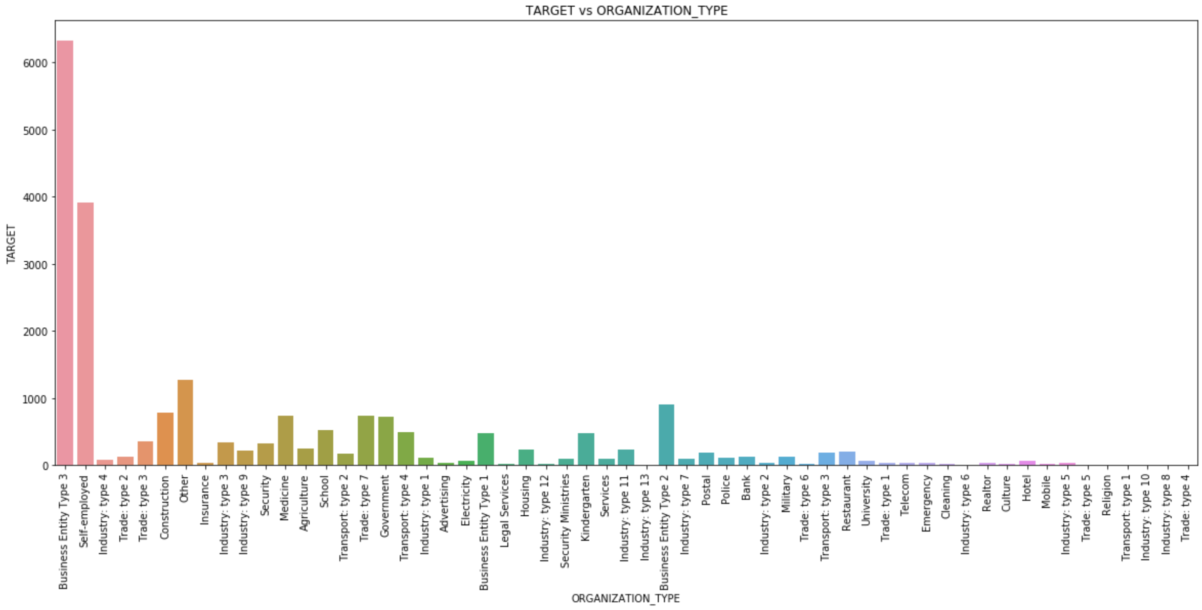
From the above graph, it can be noticed that married people default significantly higher than any other.



From the above graph, it can be noticed that those clients with region rating 2, default significantly higher than other two regions.

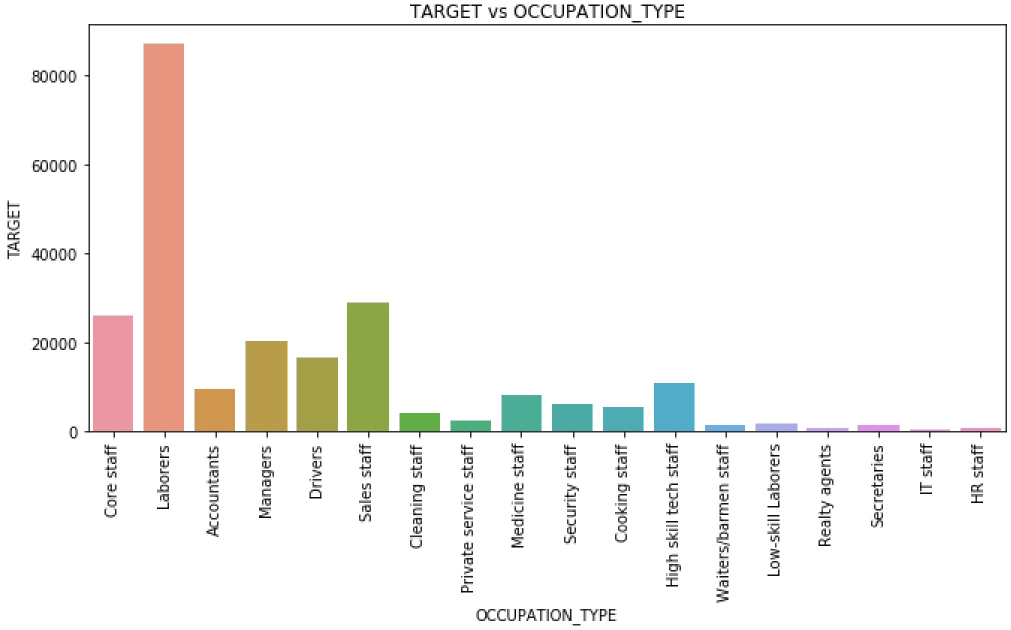


From the above graph, it can be noticed that people who live in blocks of flats default significantly higher than other two household types.

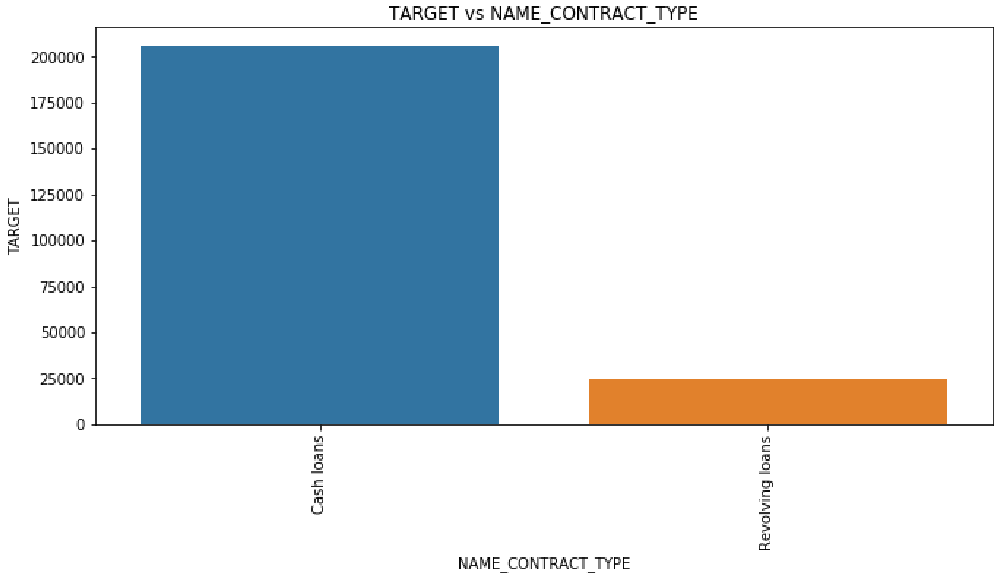


From the above graph, it can be noticed that business entity type-3 default the highest, followed by self-employed, business entity type 2, and so on.

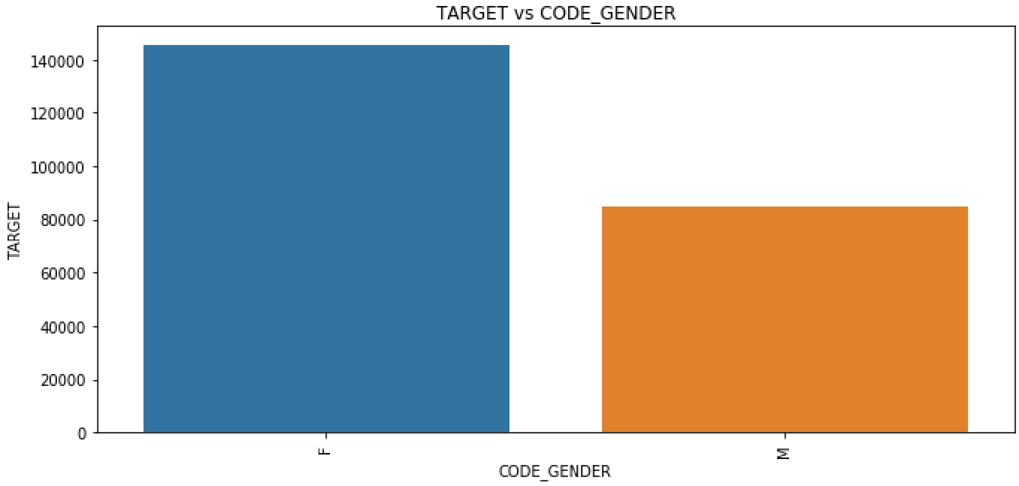
II. For Not defaulters



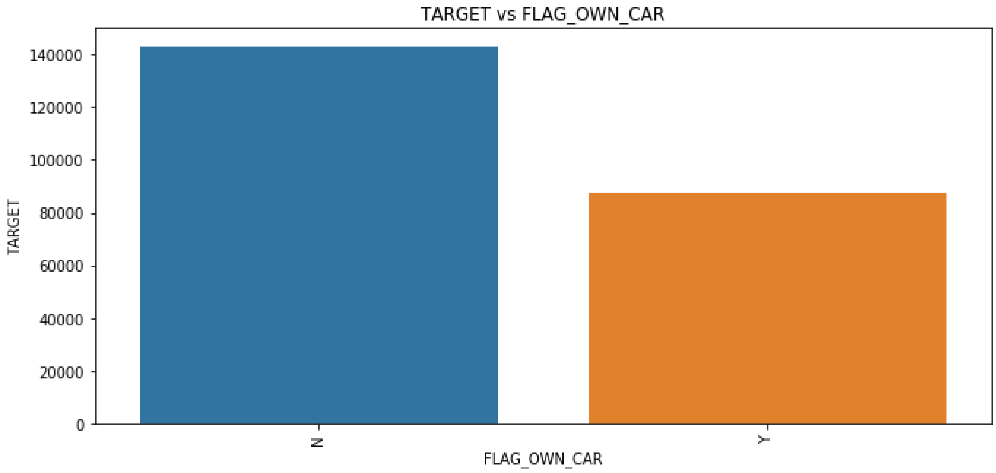
From the above graph, it can be noticed that laborers pay their loan on time and are significantly higher than others.



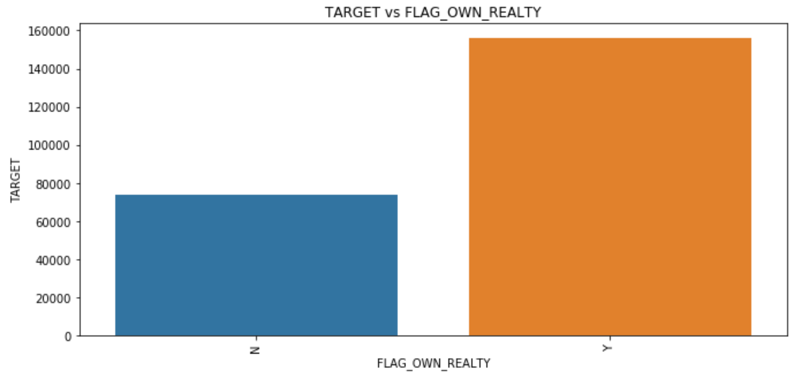
From the above graph, it can be noticed that those who take cash loans pay their loans on time and are significantly higher than those who take revolving loans.



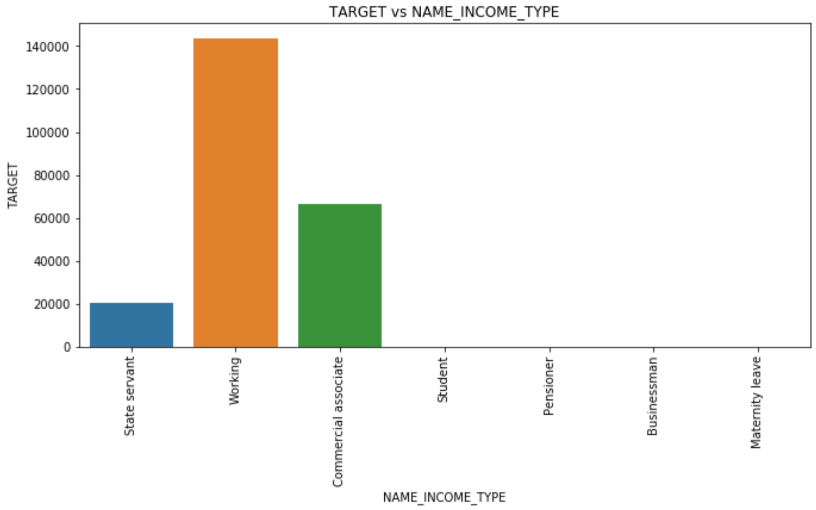
From the above graph, it can be noticed that females pay their loans on time compared to male.



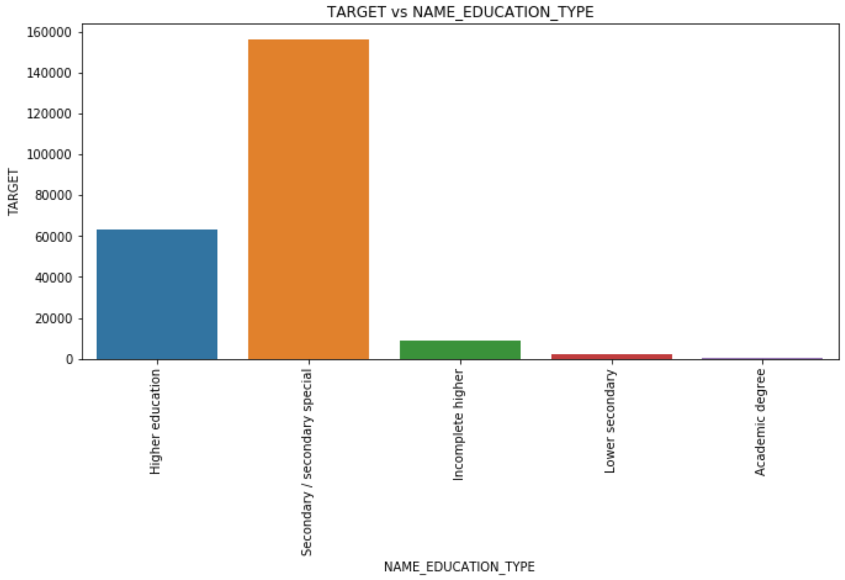
From the above graph, it can be noticed that those who do not own car pay their loans on time.



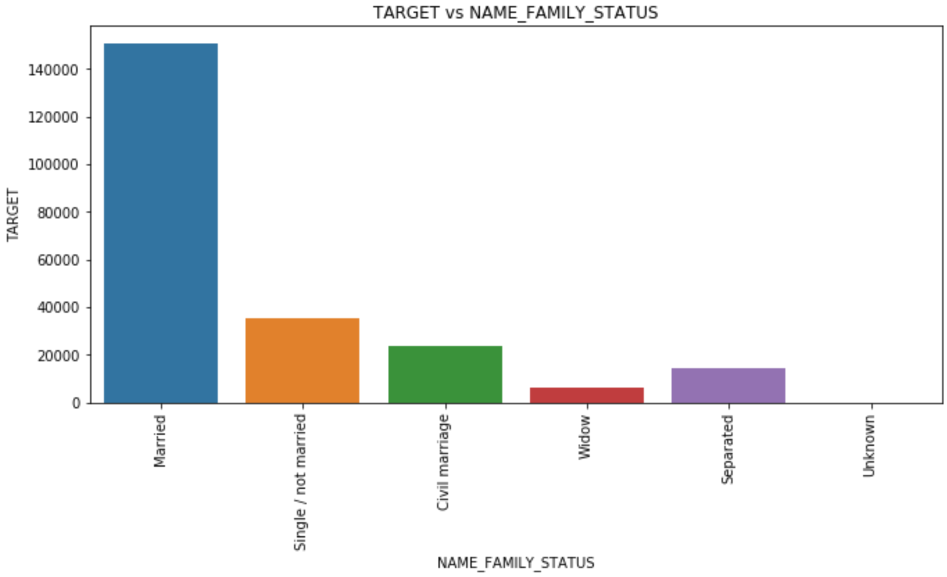
From the above graph, it can be noticed that those who own house or flats pay their loans on time.



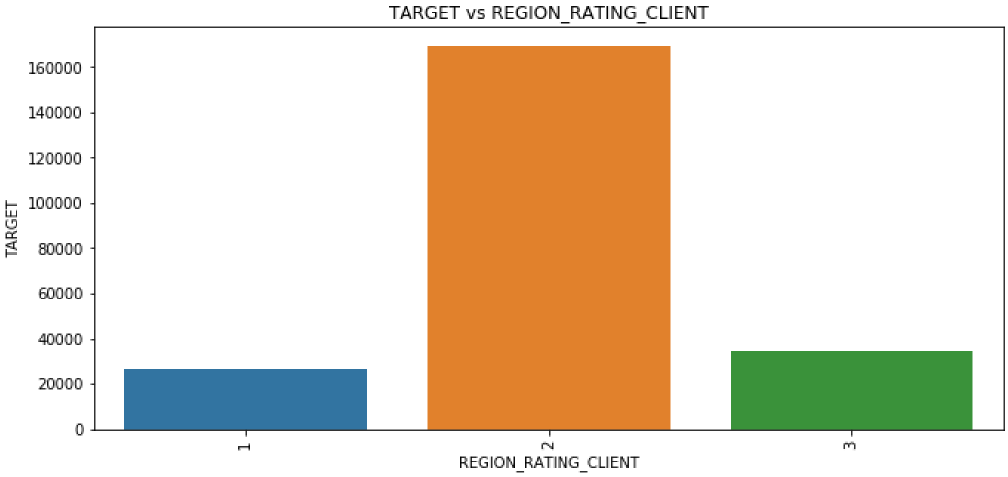
From the above graph, it can be noticed that working individuals pay their loans on time, followed by commercial associate.



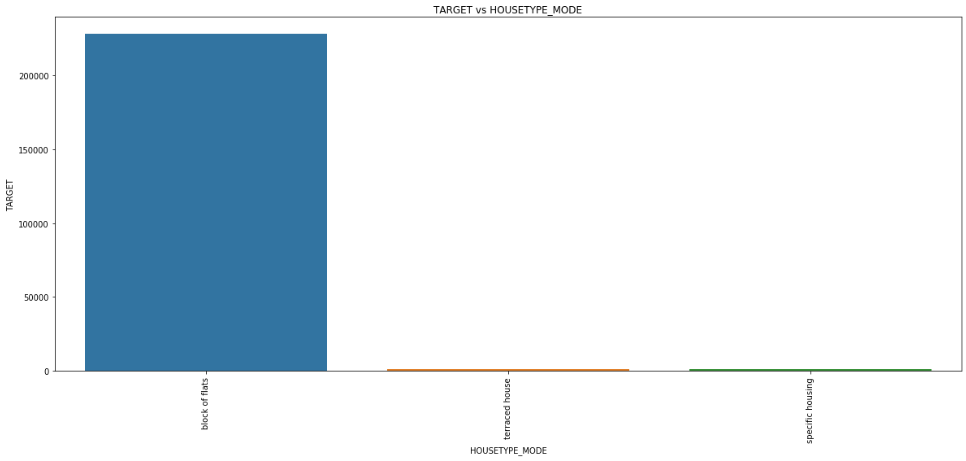
From the above graph, it can be noticed that secondary/ secondary special individuals pay their loans on time.



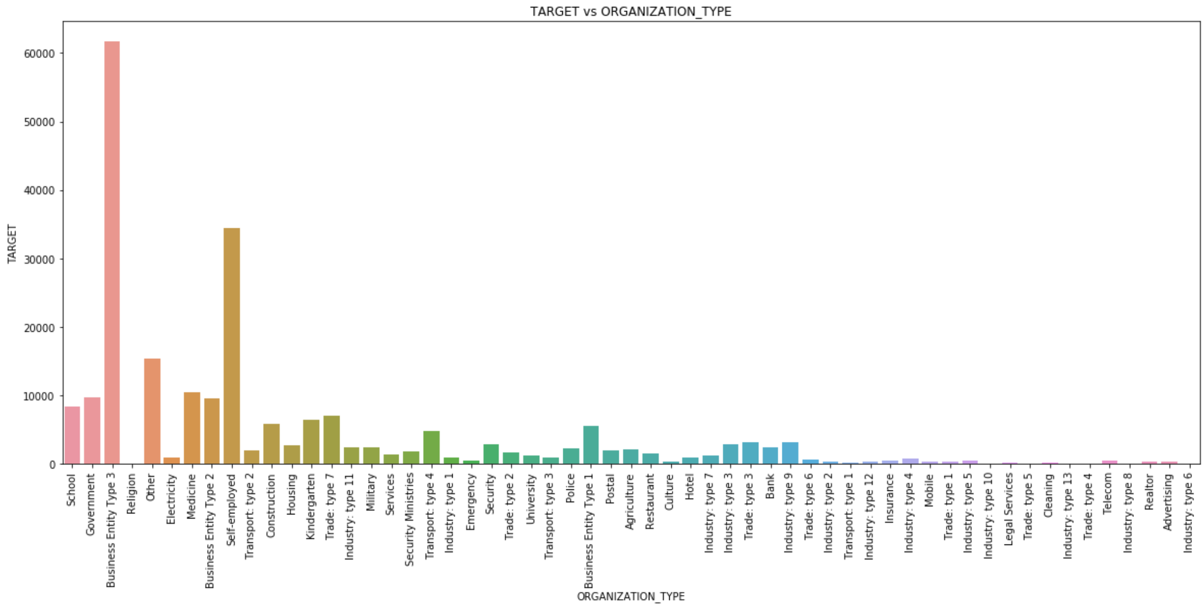
From the above graph, it can be noticed that married people pay their loans on time.



From the above graph, it can be noticed that those who region rating of 2 pay their loans on time.

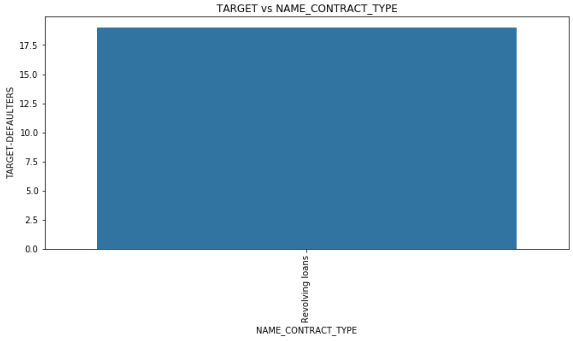


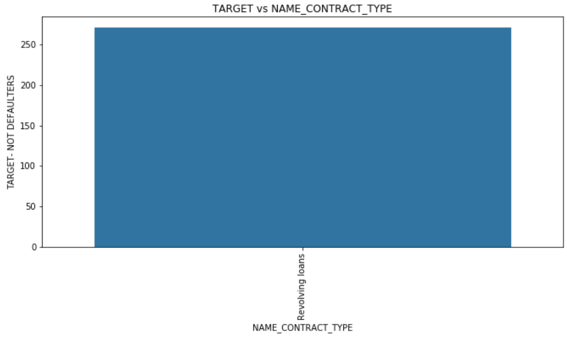
From the above graph, it can be noticed that those who live in blocks of flats pay their loans on time.



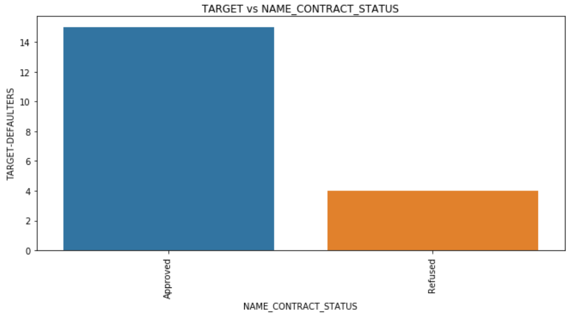
From the above graph, it can be noticed that those who have business type 3 pay their loans on time.

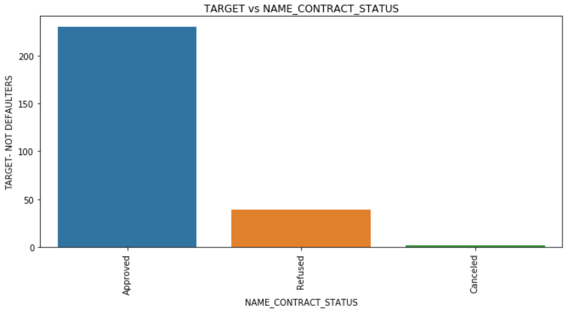
Previous Application Data set



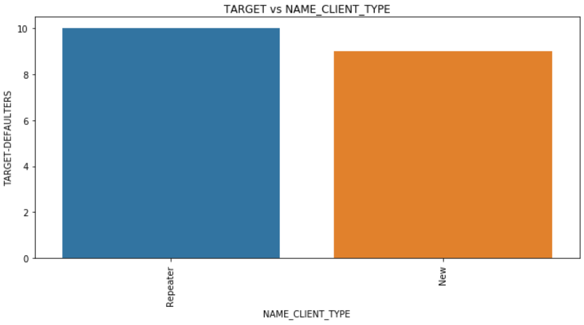


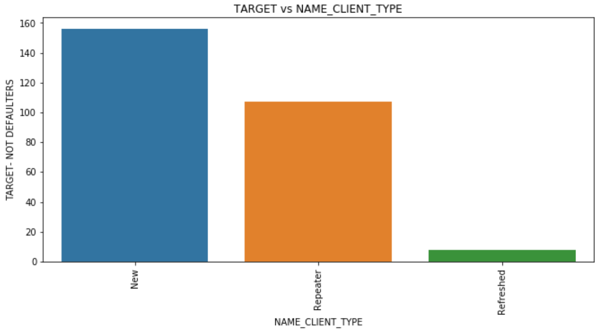
From the above graph, it can be noticed that only revolving loans is the only criteria so no comparison is possible.



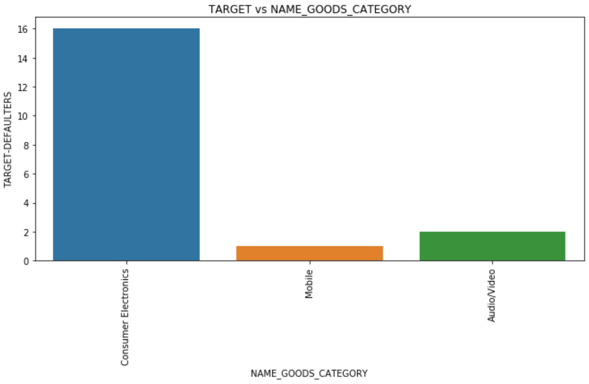


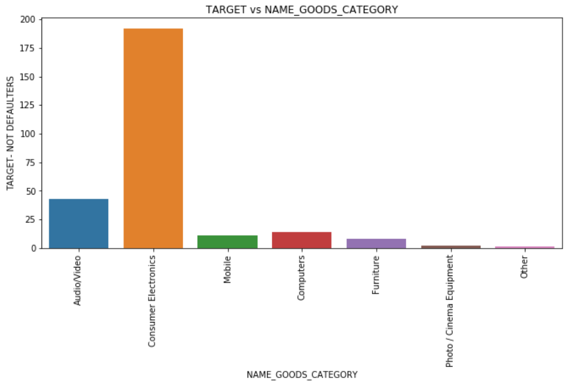
From the above graph, it can be noticed that those whose loans get approved default the highest and pay the loan time as well.



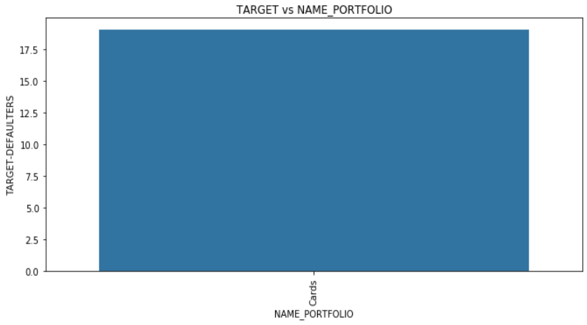


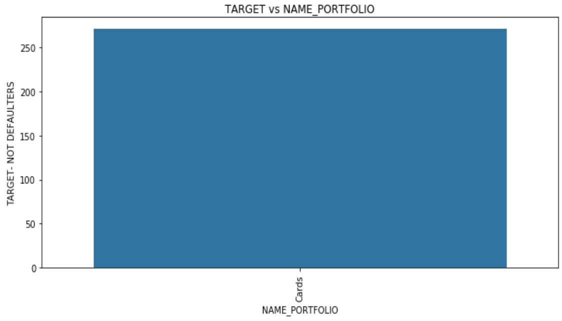
From the above graph, it can be noticed that repeater clients default the highest and are very closely followed by new clients. However, new clients are also the ones who pay their loans on time followed by repeater clients.



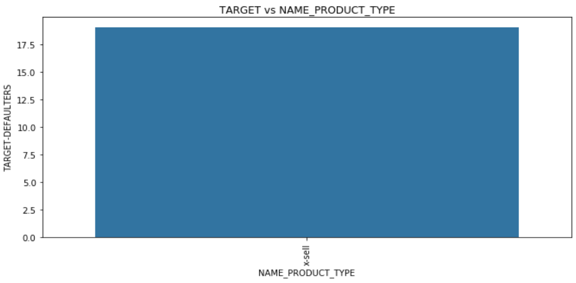


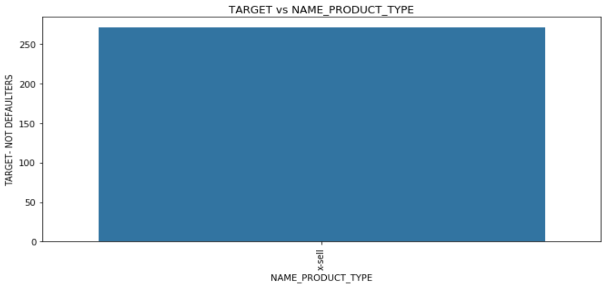
From the above graphs, it can be noticed that those who buy consumer electronic goods default the highest and pay their loans also on time.





From the above graph, it can be noticed that there’s only one criterion. So, no comparison is possible.

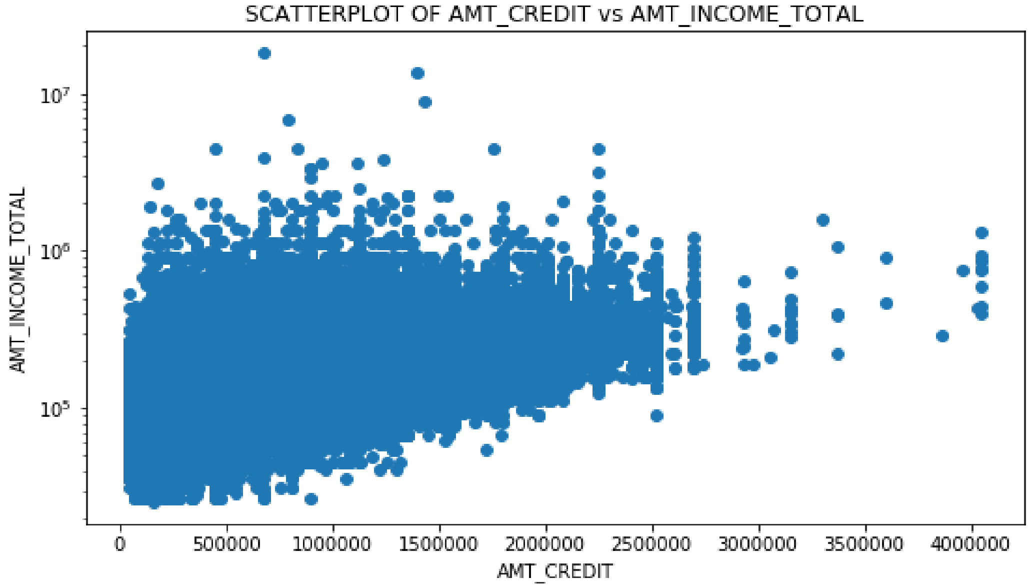




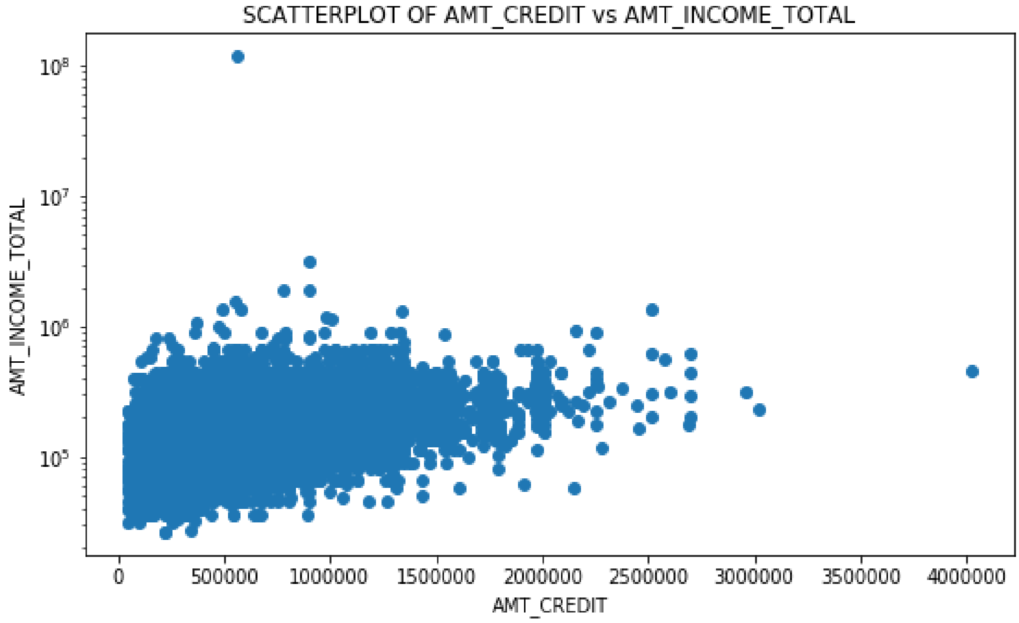
From the above graph, it can be noticed that there’s only one criterion. So, no comparison is possible.

v. Bivariate Analysis

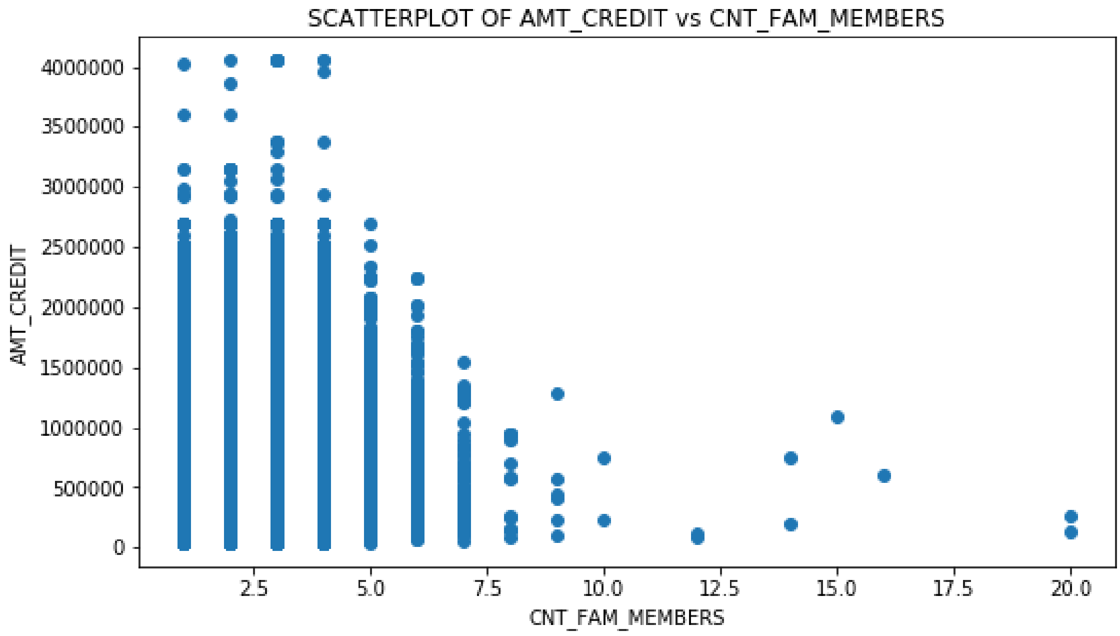
For current application data set:



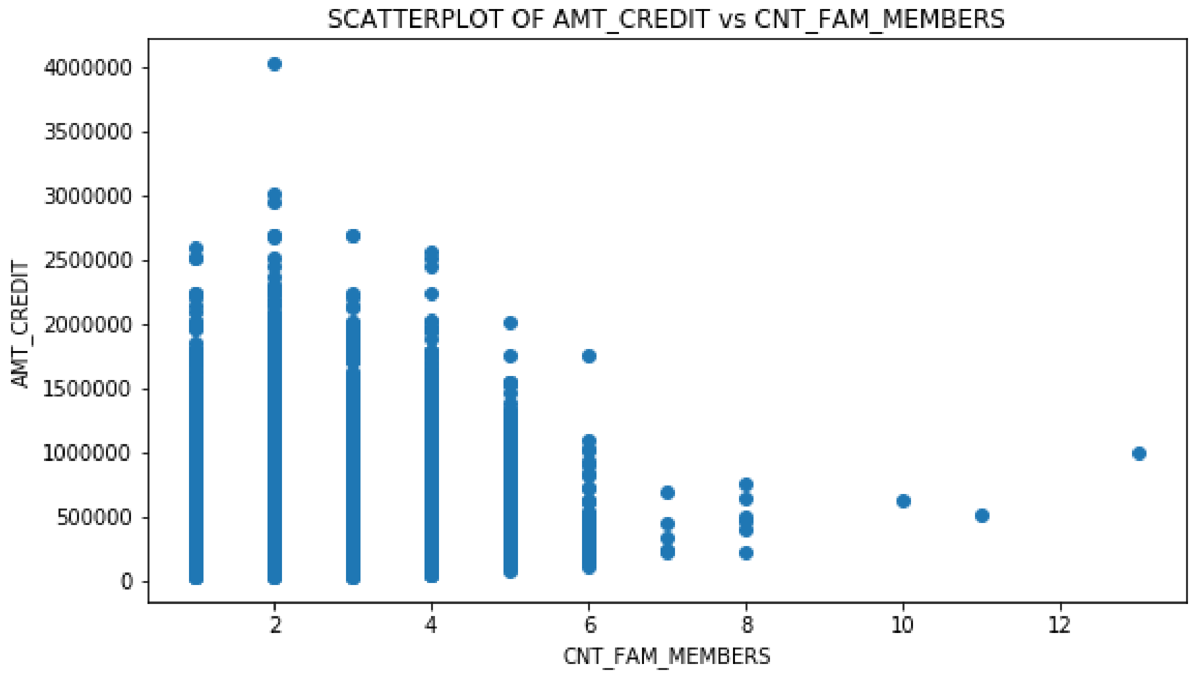
From the above chart we could observe that the density of customers whose total income is from 1 Lakh to 1 Cr are those who repay their loan amounts regularly and it is safe to provide further funds and higher credit amount.



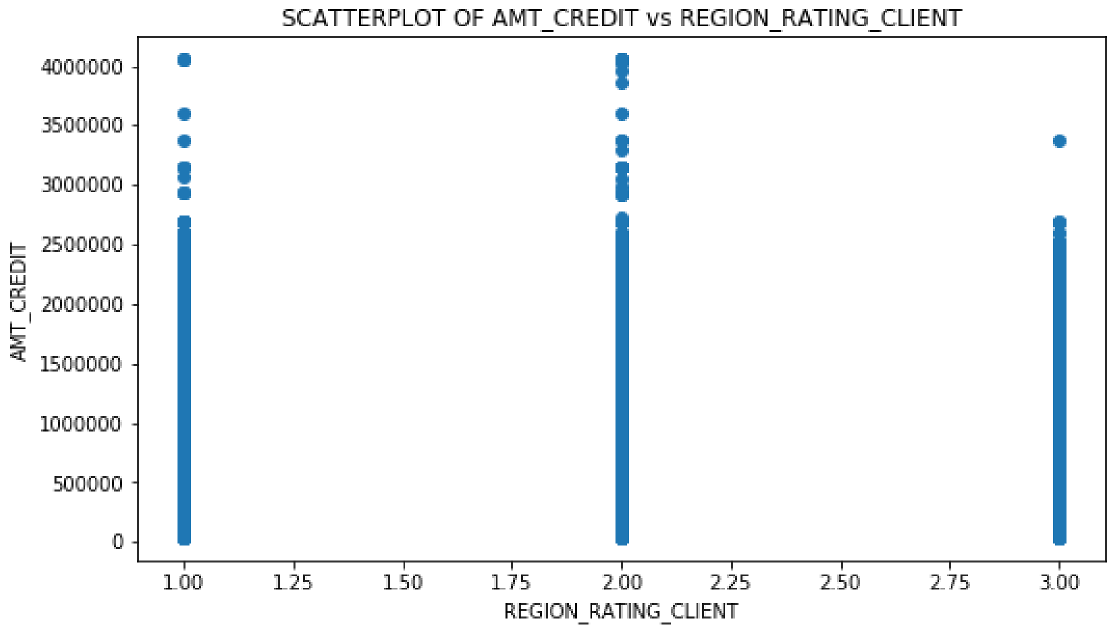
From the scatterplot we can observe that the credit amount is dense for the customers with Income total of 1 Lakh and 1 Cr and they default in repayment.



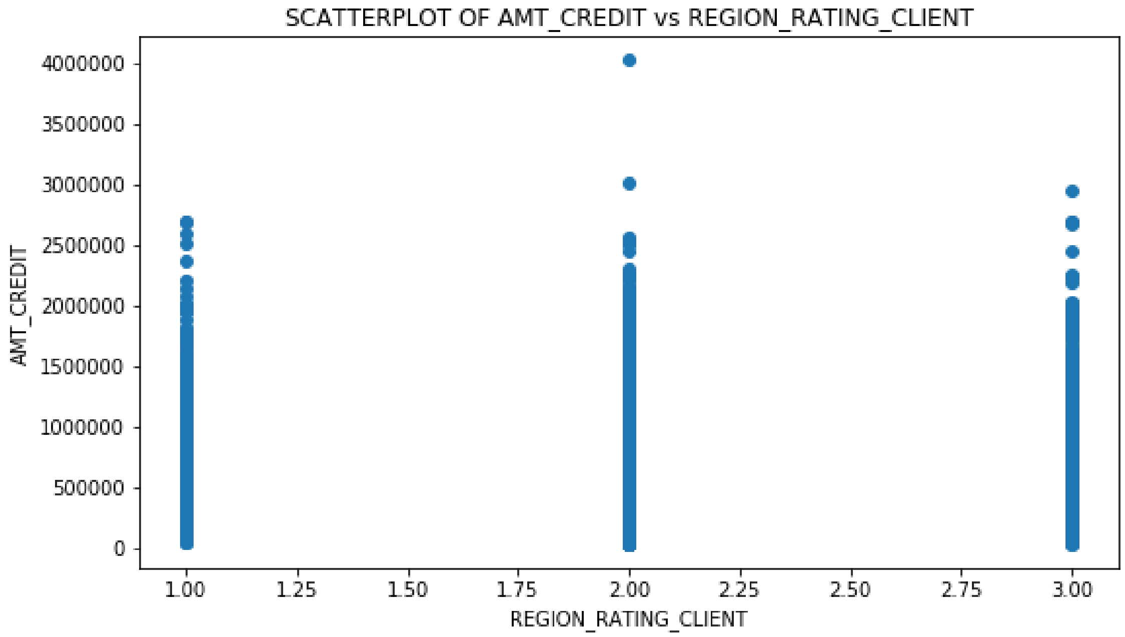
We could observe that lesser the count of the members in the family higher they get the credit amount and banks can have a confidence that the customer will be able to repay his loan.



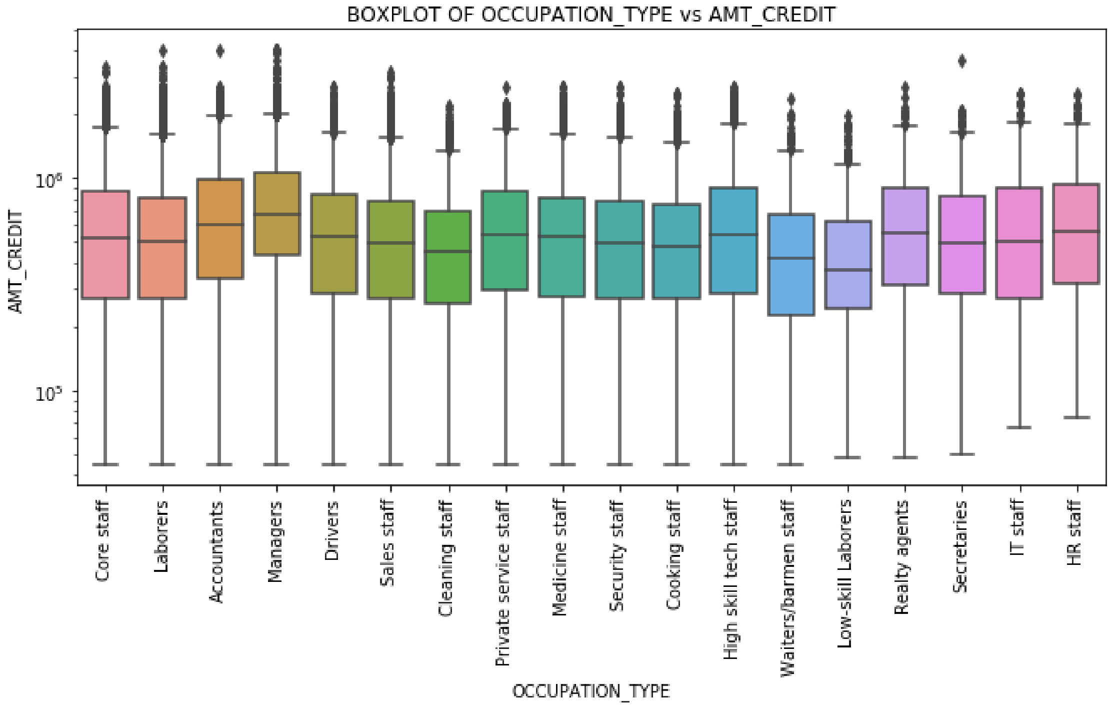
From the above we could see that the customers who face difficulties to repay the loan are having a count of family members are from 1 to 6 and they get a higher credit amount as loan.



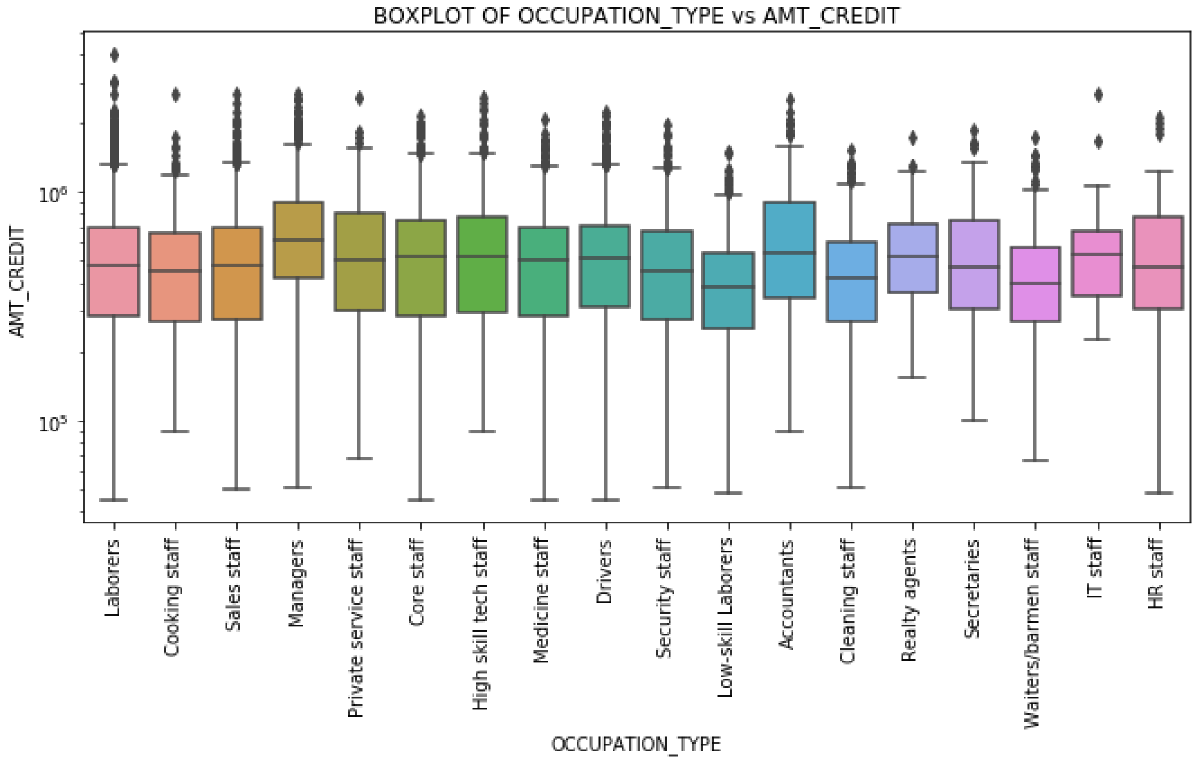
Here the loan amount credit tends to increase if the Region_Rating_Client is 1 or 2. When the customer comes from a region where usually people repay their loans, then it is a confidence of the company that a person from this region will be able to repay the loans.



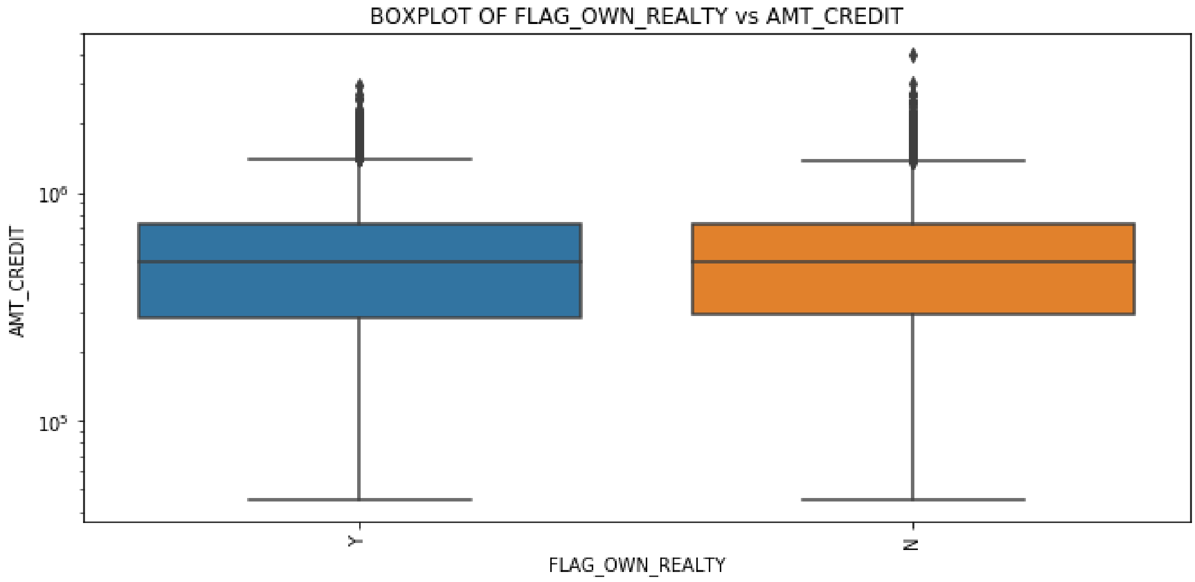
As you can see that the customers who live in the region client rating of 1 and 2 get highest amount credited for loans and have difficulties in repaying the loan.



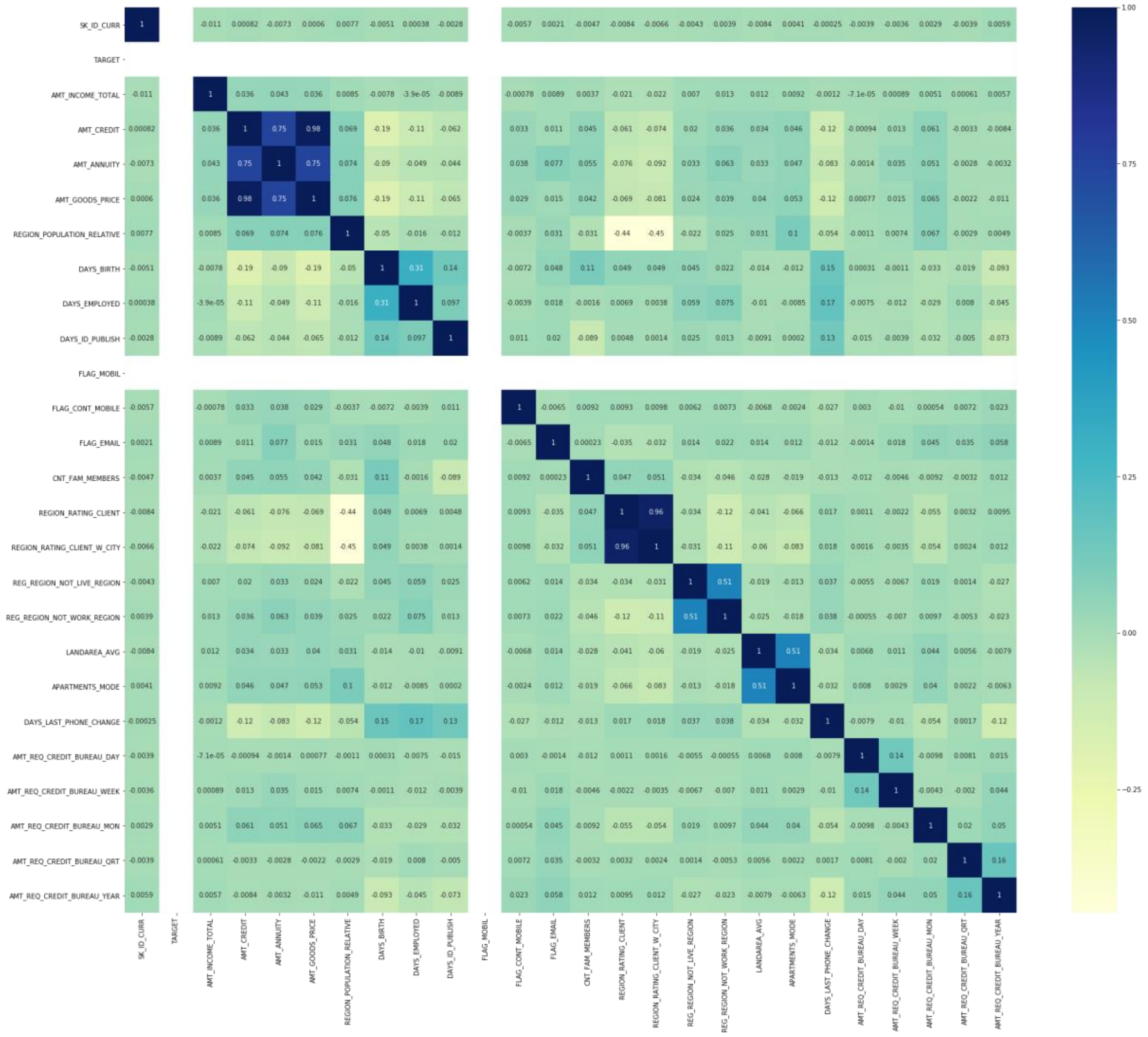
Reality Agents, Managers, High skilled Labourers are the customers who get higher credit amounts and pay their loans on time without defaults.



As per the above box plot, Accountants and Managers get higher credit amount and have difficulties in repaying the loans.



There is no effect on the credit amount of loans even if the customer has or does not have own house. Both get almost the same credit amount for the loans.



The heatmap above shows that the correlation is high between the numerical columns of defaulters such as

1.AMT_CREDIT

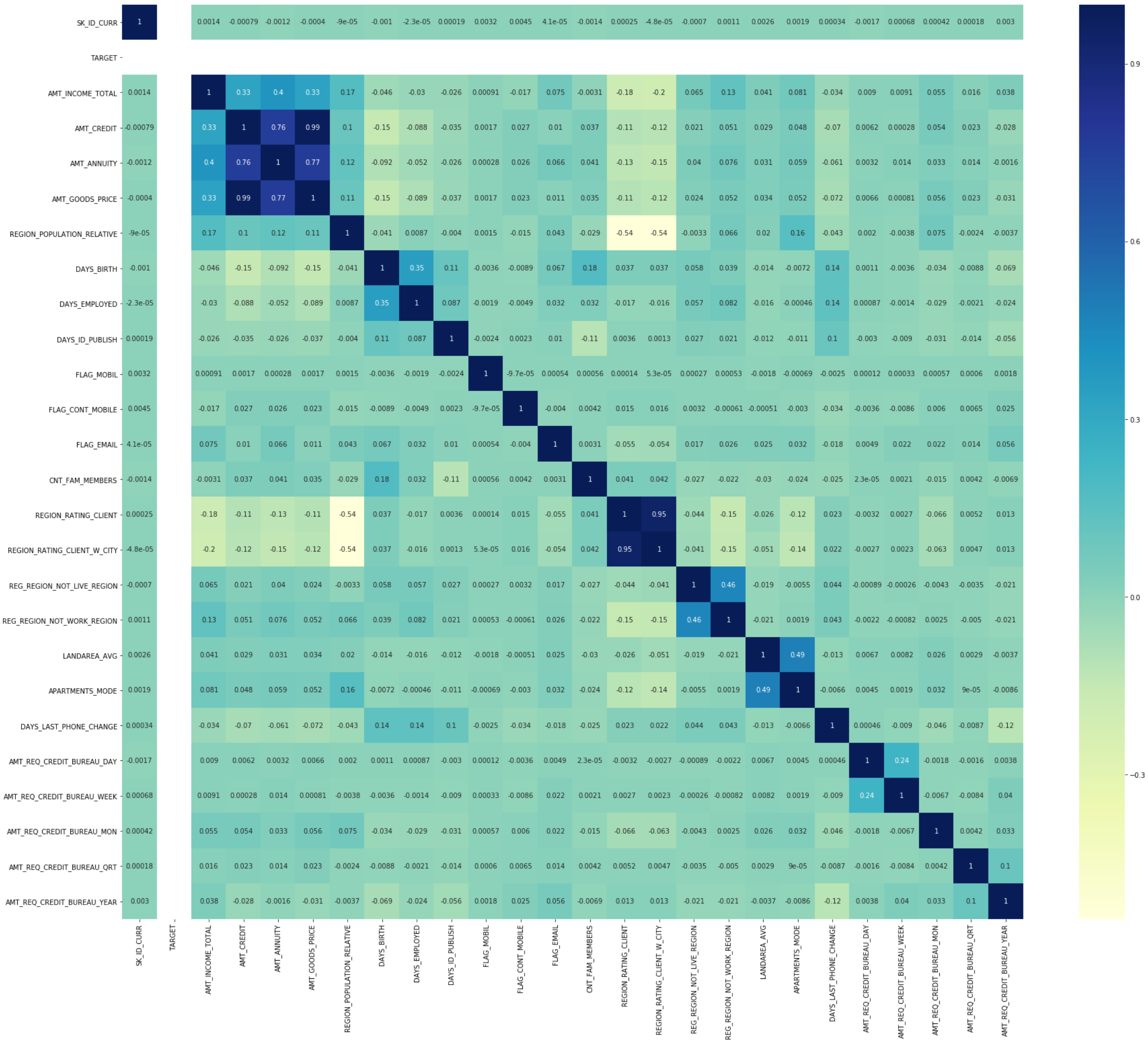
2.AMT_ANNUITY

3.AMT_GOODS_PRICE

4.DAYS_EMPLOYED

5.REGION_RATING_CLIENT

6.REGION_RATING_CLIENT_W_CITY



The heatmap above shows that the correlation is high between the numerical columns of not-defaulters such as:

1.AMT_CREDIT

2.AMT_ANNUITY

3.AMT_GOODS_PRICE

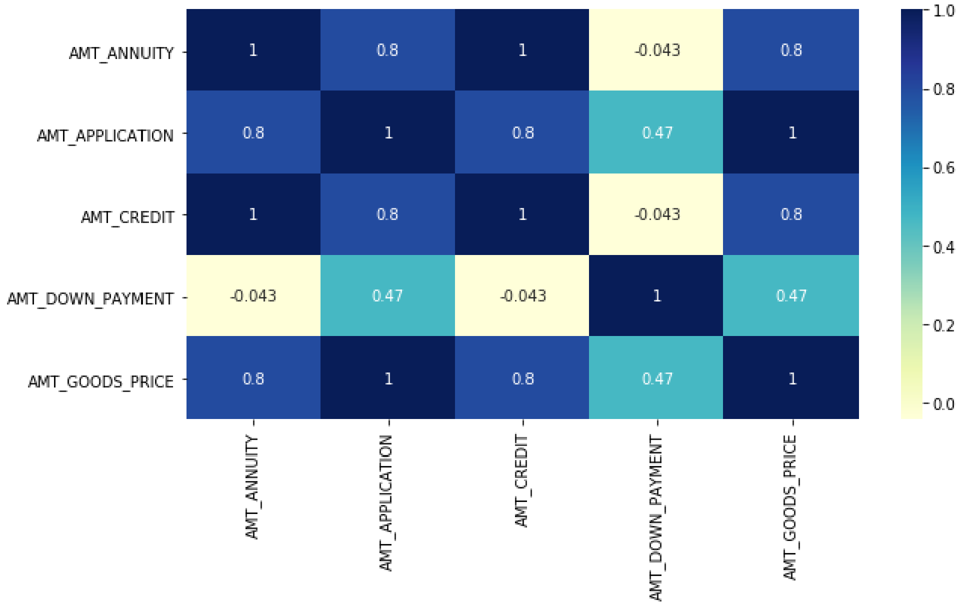
4.DAYS_EMPLOYED

5.REGION_RATING_CLIENT

6.REGION_RATING_CLIENT_W_CITY

For Previous application data set:

Defaulters:



	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE
AMT_ANNUITY	1	0.8	1	-0.043	0.8
AMT_APPLICATION	0.8	1	0.8	0.47	1
AMT_CREDIT	1	0.8	1	-0.043	0.8
AMT_DOWN_PAYMENT	-0.043	0.47	-0.043	1	0.47
AMT_GOODS_PRICE	0.8	1	0.8	0.47	1

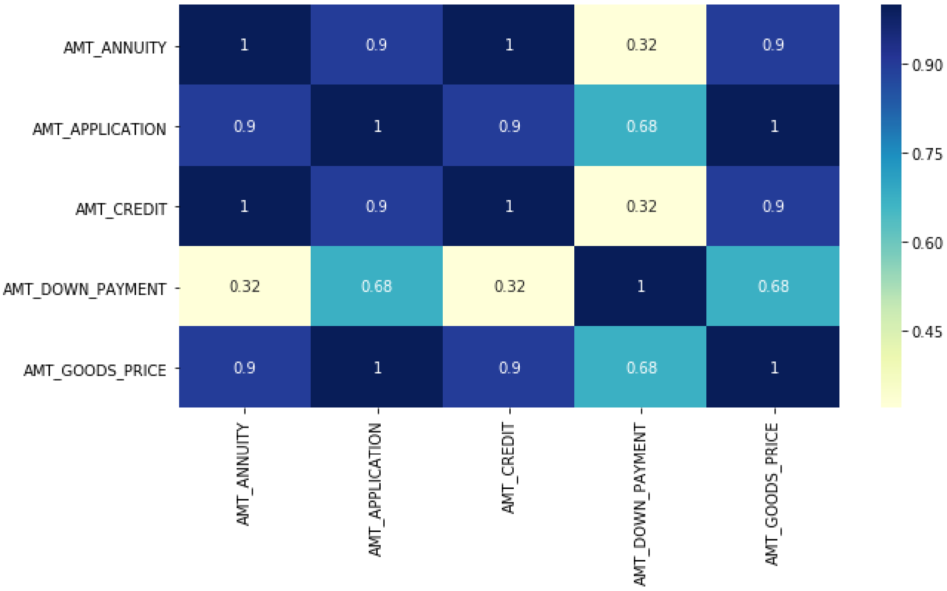
Perfect correlation is seen among certain variables:

a. Amount of credit and amount of annuity
b. Amount of goods price and amount of application

It can be noticed from the above heat map that the highest correlation exists between the following variables:

a. Amount annuity and amount of application
b. Amount of application and amount of credit
c. Amount of goods prices and amount of annuity
d. Amount of goods prices and amount of credit

For Not defaulters:



Perfect correlation is seen among certain variables:

c. Amount of credit and amount of annuity
d. Amount of goods price and amount of application

It can be noticed from the above heat map that the highest correlation exists between the following variables:

e. Amount annuity and amount of application
f. Amount of application and amount of credit
g. Amount of goods prices and amount of annuity
h. Amount of goods prices and amount of credit

Findings and Conclusion

From the univariate analysis and bivariate analysis of current application data set, we notice that the same variables have high number of defaulters and non-defaulters. However, the number of non-defaulters for every category is almost 10 times higher than those for defaulters.

Variable	Category	Defaulters(frequency)	Non-Defaulters(frequency)
Occupation type	Laborers	8000+	80,000+
Type of loan	Cash Loan	20,000	2,00,000+
Gender	Females	12000+	1,40,000+
Car	Doesn't own	14000+	1,40,000+
Realty	Owns	14000+	1,40,000+
Income type	Working	14000+	1,40,000+
Education	Secondary	16000+	1,60,000+
Client Status	Married	12000+	1,40,000+
Region rating	2	16000+	1,60,000+
House type	Blocks of flats	20,000+	2,00,000+
Organization type	Business entity 2	6000+	60,000+

We are already aware that the data imbalance ratio is approximately 1:11(1:10.55). Hence the difference between the number of defaulters and non-defaulters can be understood as a result of data imbalance.

The heatmaps results show that following continuous variables have high correlation amongst each other for both defaulters and non-defaulters: Amount of goods prices, amount of annuity and amount of goods prices and amount of credit.

The analysis of scatter plots between amount credited and other variables like family size, region rating and income total reveals the same trend as that of categorical variables.

From the univariate and bivariate analysis of previous application data set, we notice the following result:

Variable	Category	Defaulters(frequency)	Non-Defaulters(frequency)
Name contract type	Revolving loans	17+	250+
Contract Status	Approved	24+	200+
Client type	Repeater	20	100
Client type	New	Between 8-10	140+
Goods category	Consumer electronics	26	Between 175-200
Name portfolio	Cards	17+	250+

For contract type, new client type and name portfolio, the difference between defaulters and non-defaulters is the same as data imbalance ratio.

However, for repeater client type, for those with approved contract status, and for those who buy consumer electronics goods, the differences are not consistent with data imbalance ratio.

The correlation analysis of both defaulters and non-defaulters reveals that Perfect correlation is seen among certain variables: amount of credit and amount of annuity, Amount of goods price and amount of application.

It can be noticed from the above heat map that the high correlation exists between the following variables:

a. Amount annuity and amount of application

b. Amount of application and amount of credit

c. Amount of goods prices and amount of annuity

d. Amount of goods prices and amount of credit

Hence, we can conclude that the bank needs to be careful about the following client types who can prove to be defaulters:

Laborers, People who take Cash Loan, Females, those who don't own a car, those who own a flat and are living in blocks of flats, are Working Individuals, have Secondary/secondary special education, who are Married, who come from regional rating 2 places, and whose organization type is of business entity 2.

While giving loans, they should look at the past history of the client and notice the following factors:

a. The kind of contract that they want to have because revolving loans are defaulters whereas cash loans comparatively default less.
b. Whether their contract status in past has been Approved or not. If they get their loans easily approved, they might become careless about paying it back on time.
c. Those who are Repeaters, also can default higher than others. The same can be expected if they are new clients.
d. Clients who default buy more of Consumer electronics goods than any other kind. Therefore, such loans should be strictly followed through.
e. Those who have cards tend to default higher than any other portfolio type. So, routine follow up should be done for those who hold various cards of the banks.

Apart from this, both previous and current application data set reveal high correlation of defaulters and non-defaulters for three continuous variables: amount of annuity, amount of credit and amount of goods prices.

The previous application data set also reveals that amount of application is perfectly correlated with amounts of goods price and is highly correlated with amount of credit and annuity.

Hence, the amount of application of loan by every client is an important criterion. Since there's positive correlation with amount of credit and amount of annuity, in case of both defaulters and non-defaulters, one needs to be very careful about it. If amount of application is high, there will be high chances of defaulting as well as defaulting. Similarly, we can understand for all the variables that are highly correlated.

Example case: Suppose a married female applies for a cash loan to the bank for furniture.

We know from our analysis that married and female criteria are defaulters. But Cash loan an furniture purchase are usually non-defaulters. The bank should do a background check and find out details if this person is repeater or a new client. If she has had approved status before or is any other type. What kind of organization she works in and the kind of portfolio she holds.

Suppose the bank figures out that the female is a new client and doesn't have any status from past record and she works as a salaried employee who holds a debit as well as credit card. She doesn't have any overdraft on her credit card either. In a such a situation the bank can take the risk of approving the loan for low value of amount of application.

Note: While drawing final conclusion, data imbalance has been ignored.