CREDIT EDA CASE STUDY

Group Name:
1. Shantam
2. Swati Hota

OBJECTIVE

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 utilize this knowledge for its portfolio and risk assessment.
To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

The given 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):
Approved: The Company has approved loan Application
Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
Refused: The company had rejected the loan (because the client does not meet their requirements etc.).
Unused offer: Loan has been cancelled by the client but on different stages of the process.
In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Three datasets given:
1. ***'application_data.csv'*** contains all the information of the client at the time of application.
The data is about whether a **client has payment difficulties.**

2. ***'previous_application.csv'*** contains information about the client's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

3. ***'columns_description.csv'*** is data dictionary which describes the meaning of the variables.

ANALYSIS

- Missing values ratio acknowledgement
- Data Imbalance
- Outliers
- Removing missing values
- Univariate analysis of the new and old applications
- Bivariate analysis of the new and old applications
- Correlation

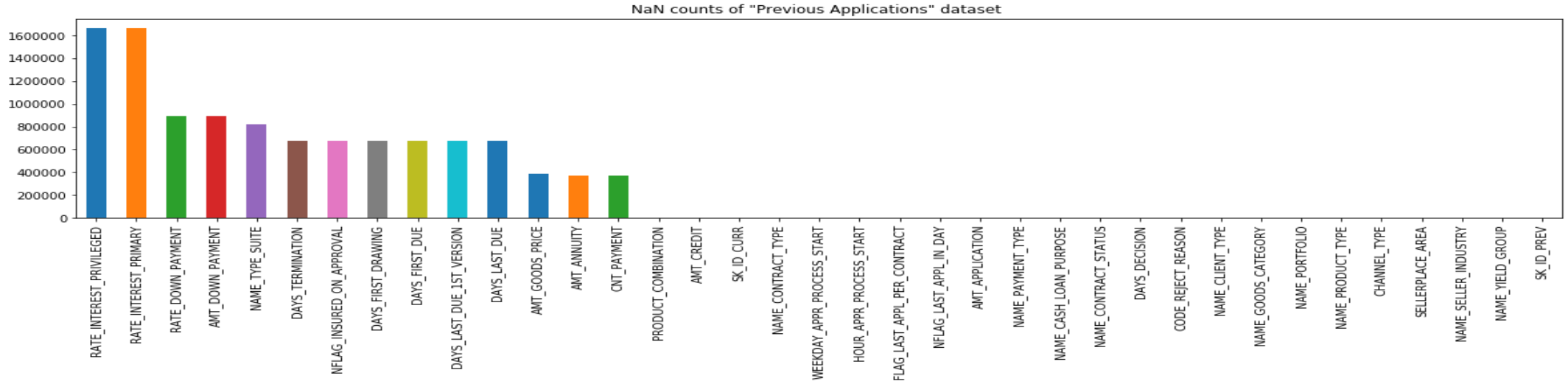
Data Quality checks and handling missing values:

Two data sets were provided as part of this case study

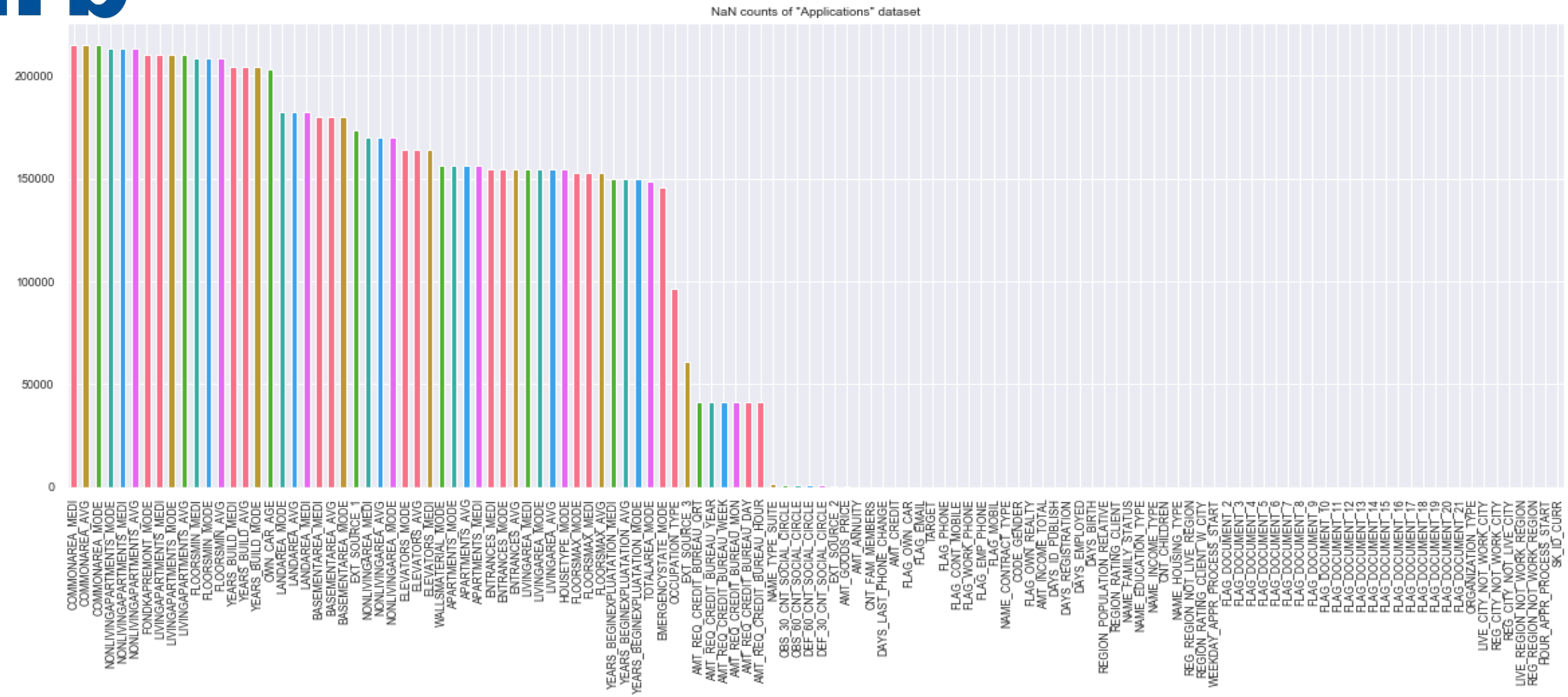
- Application Data
- Previous application data

To find out the missing values and handle it :
We have observed there were many missing values in this data set, so using the indexing and count of missing values in each column, identified the % of missing values for each column.

- And dropped all columns from data frame for which missing values % is more than 50. We have also found few more columns which are having around 47% missing values. Since these are almost around 50% ,we have removed these columns as well.
- To extend the data enhancements and maintain a data frame with accurate values, we have identified columns with NULL values and even applied the same 47% logic and removed those columns from the data frame.



Insight: Near about 37% columns are having fully or partially missing values in **"PREVIOUS -APPLICATION" dataset.**



Insight: Near about 46% columns are having fully or partially missing values in **"APPLICATION" dataset.**

Data imbalance

For "TARGET" Class, we are having a huge data imbalance for Class-0 and Class-1

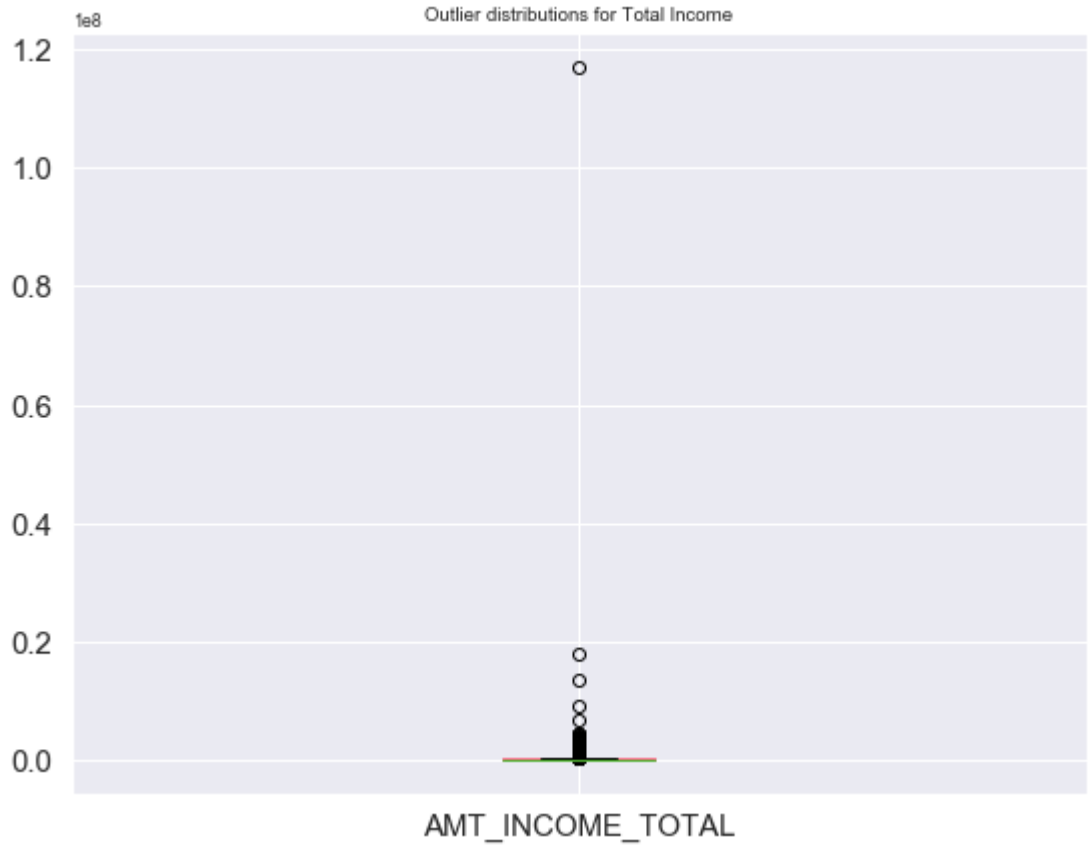
Class-0 : 92% data
Class-1 : 8% data

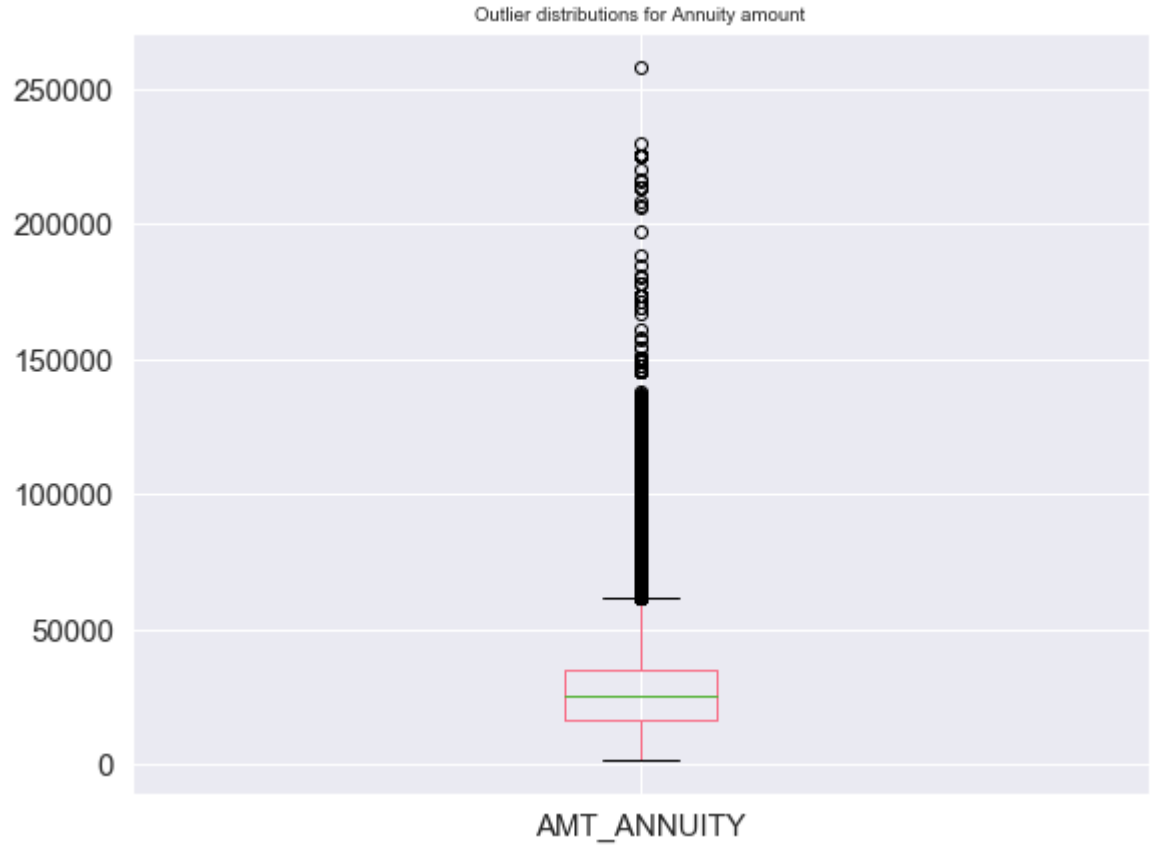
Checking the Data frame for outlier values and analyze them

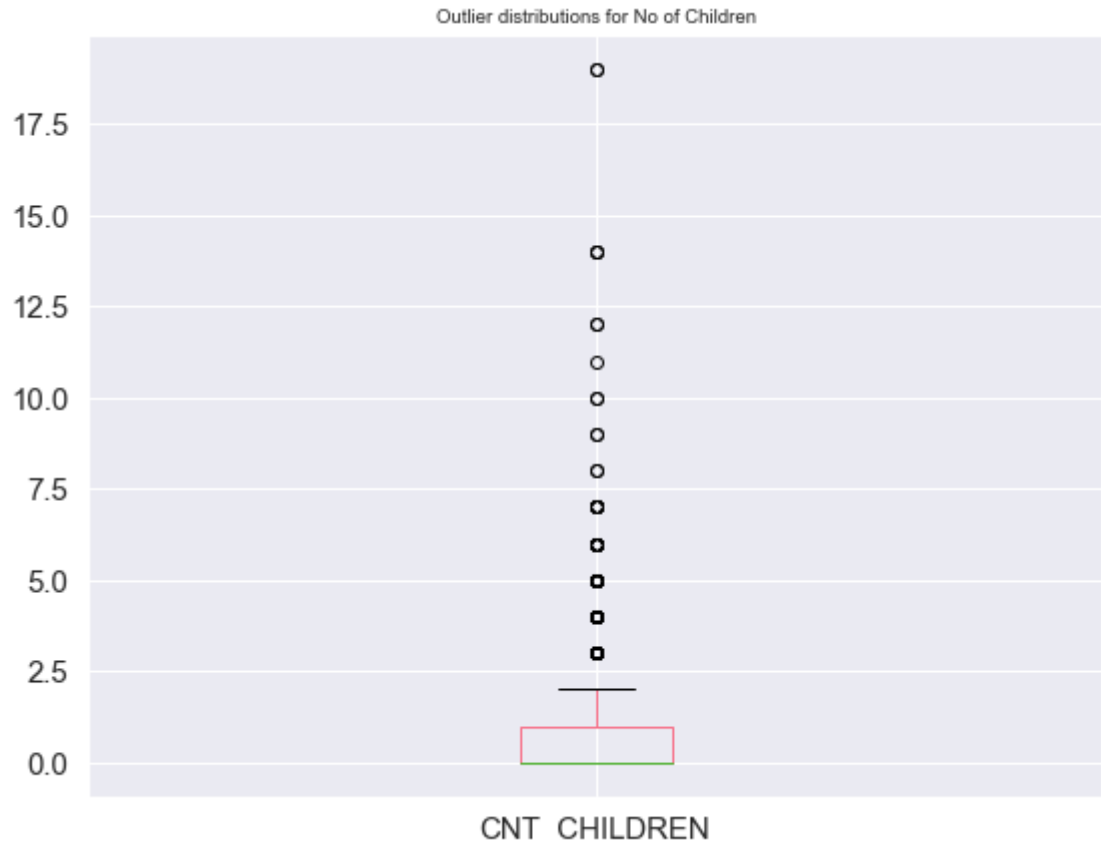
- Checking the outliers for columns and understanding the reason to mention that as an outlier.
- Here in our analysis to find out the outliers, we have considered few numerical columns and analyzed the statistics of them.

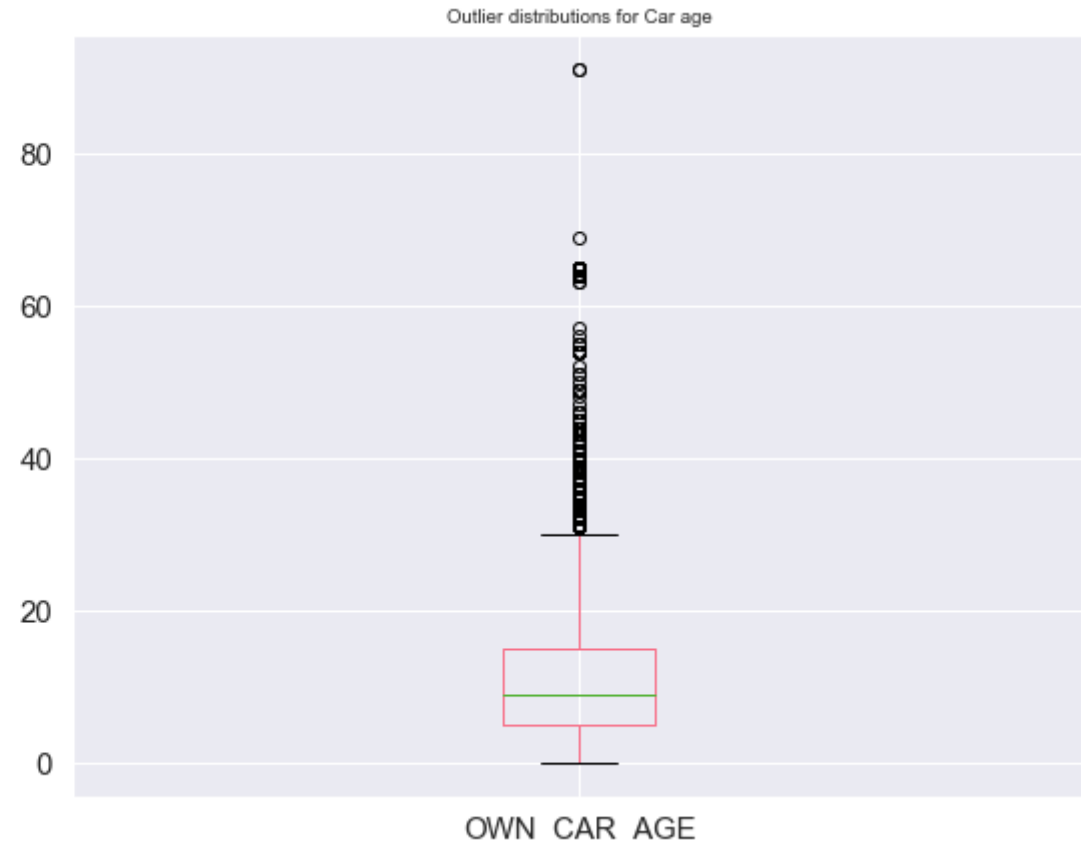
Followings are few variables that have the outlier value(s)

- AMT_INCOME_TOTAL
- AMT_ANNUITY
- CNT_CHILDREN
- OWN_CAR_AGE


Outlier distributions for Total Income
1e8
1.2
1.0
0.8
0.6
0.4
0.2
0.0
AMT_INCOME_TOTAL


Outlier distributions for Annuity amount
250000
200000
150000
100000
50000
0
AMT_ANNUITY


Outlier distributions for No of Children
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
CNT_CHILDREN

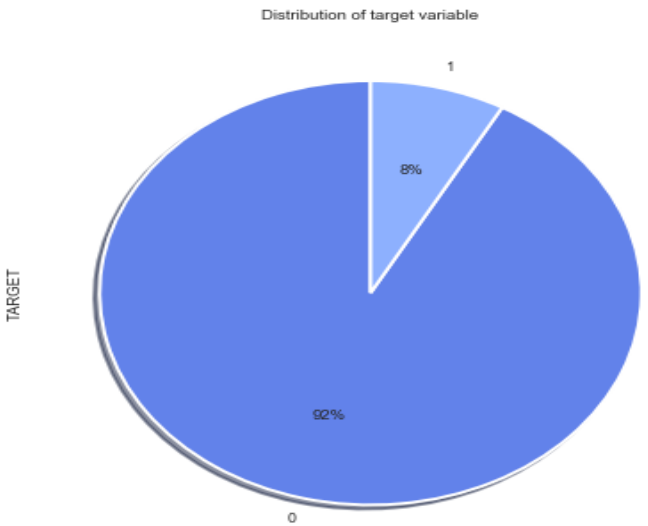

Outlier distributions for Car age
80
60
40
20
0
OWN_CAR_AGE

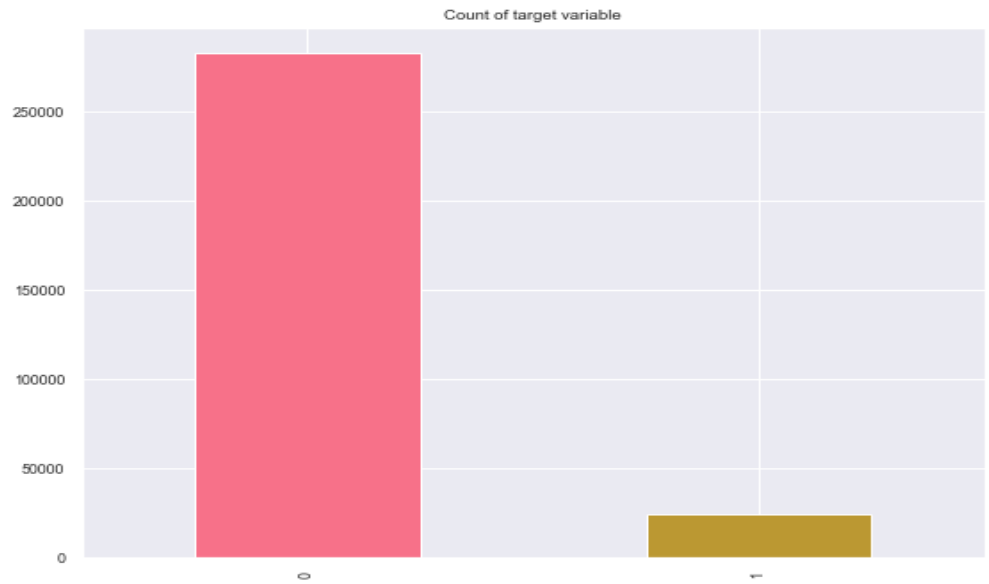
UNI AND BI VARIATE ANALYSIS

UNIVARIATE ANALYSIS:

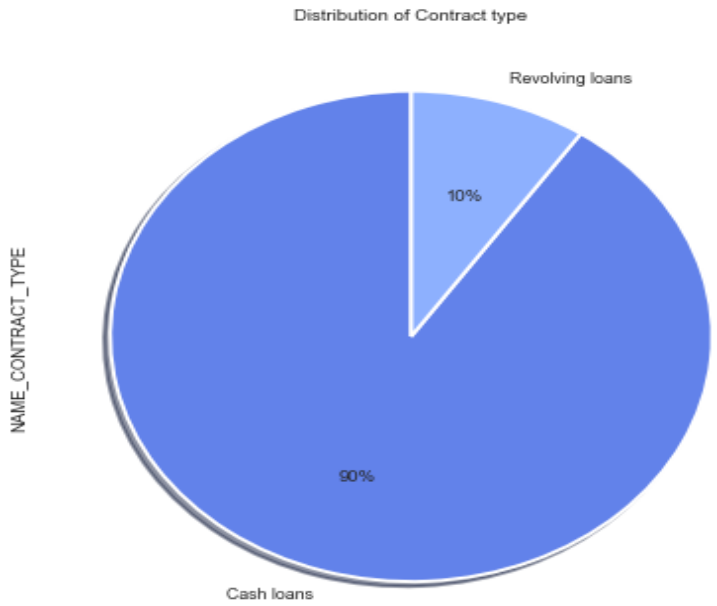
Univariate analysis is perhaps the simplest form of statistical analysis. Like other forms of statistics, it can be inferential or descriptive. The key fact is that only one variable is involved. Univariate analysis can yield misleading results in cases in which multivariate analysis is more appropriate.

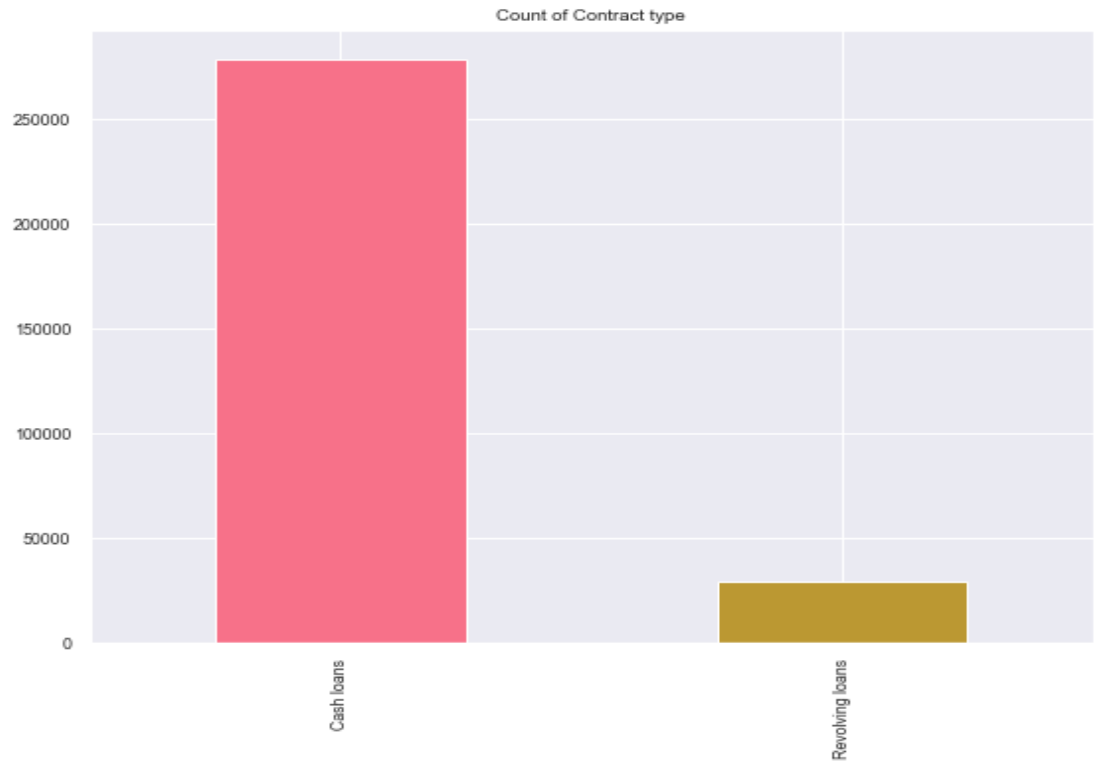
- While working with the data frame based on Target variable, **As shown in the below screenshot, we can clearly see the imbalance between Target type 1 and 0**.
- **Ratio is of 92 : 8**



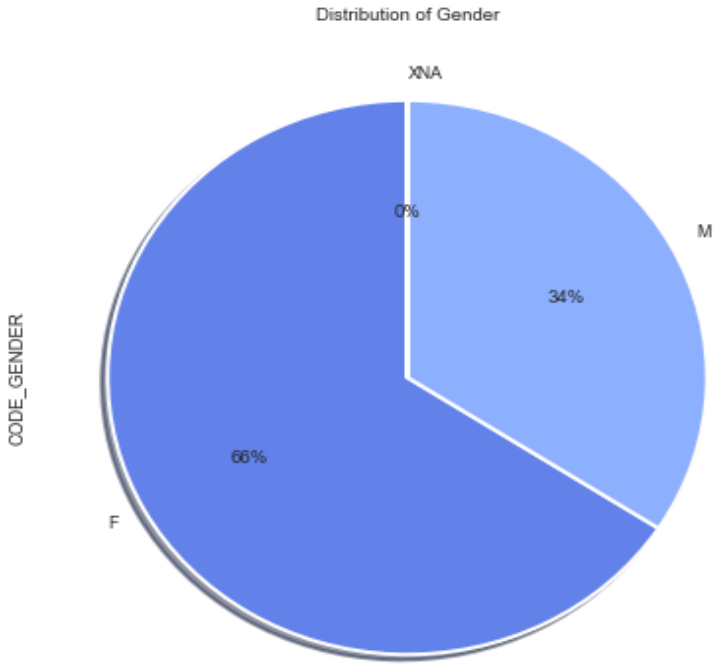


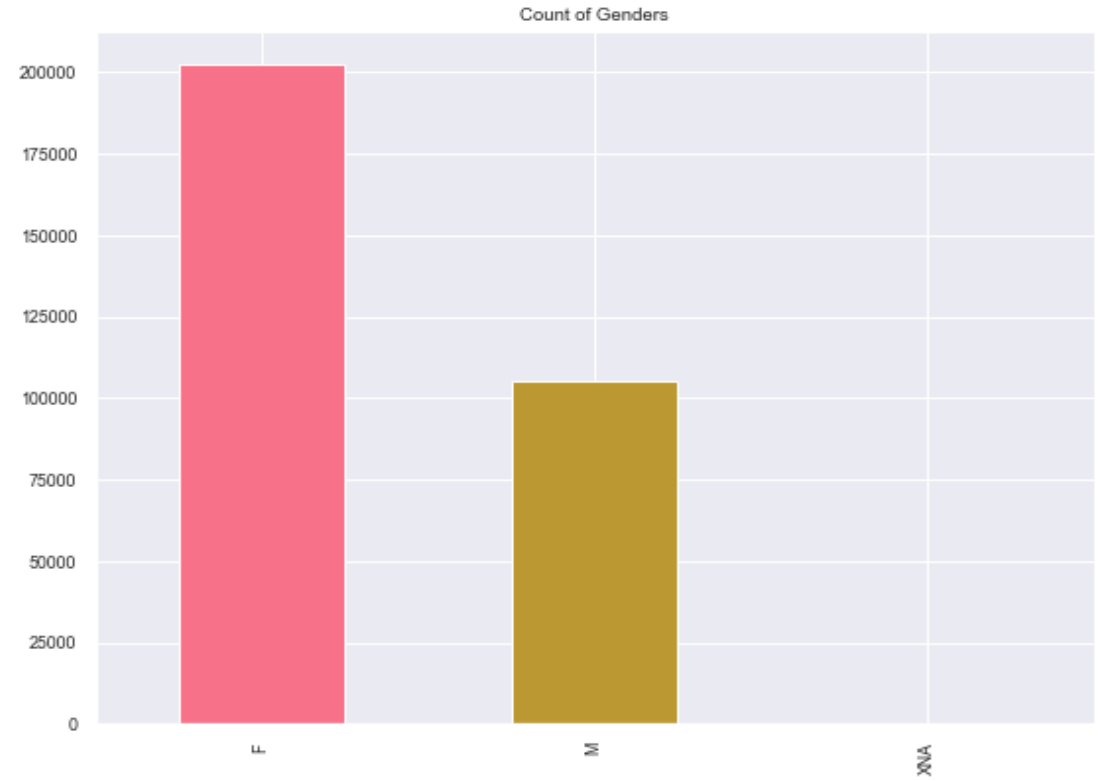
Insight: Out of total new loan applications about 8% are defaulters.



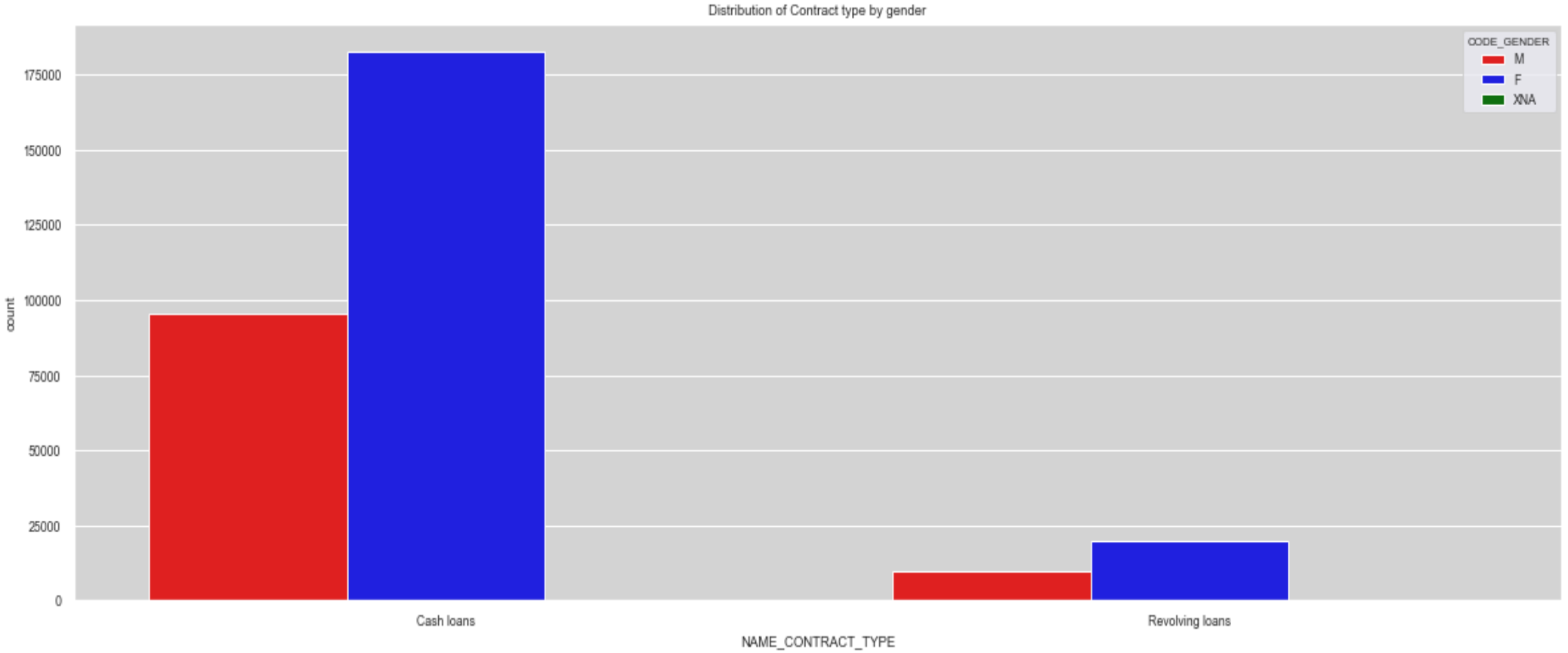


Insight: 90% loans are Cash loan, so most people prefer cash loans.

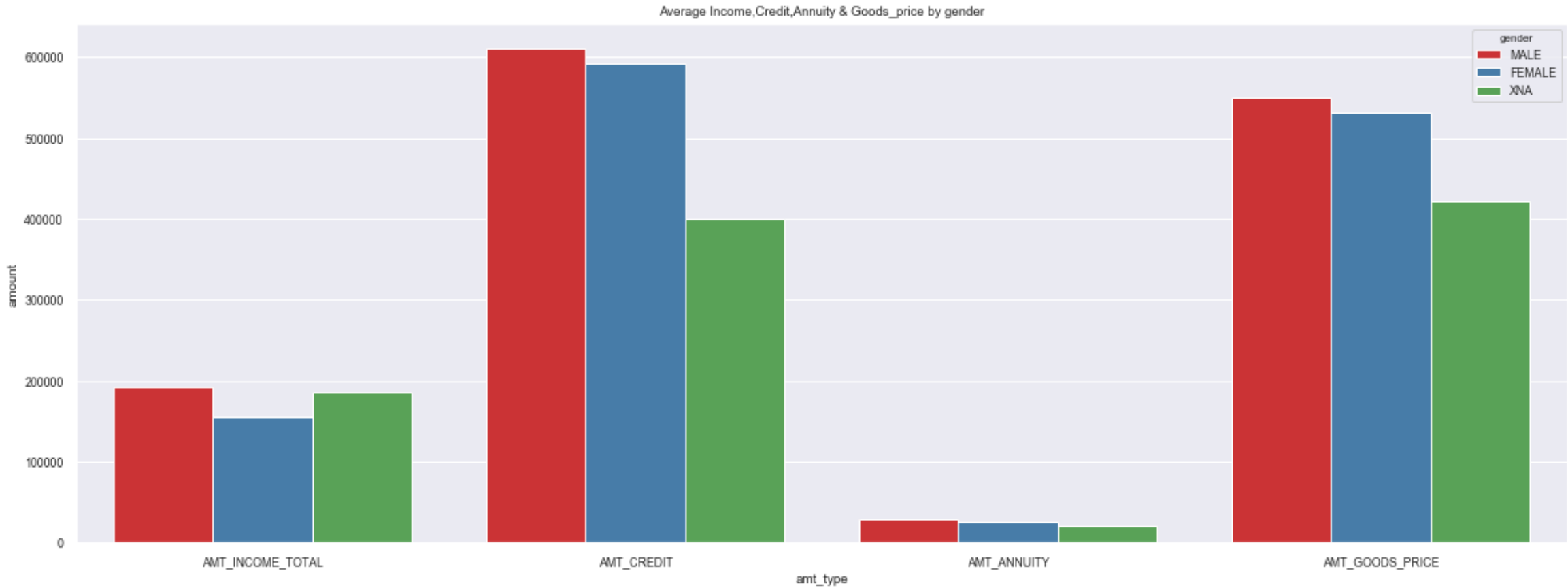




Insight: Out of total new loan applications 66% are female clients, Female clients are taking more loan than Male clients.



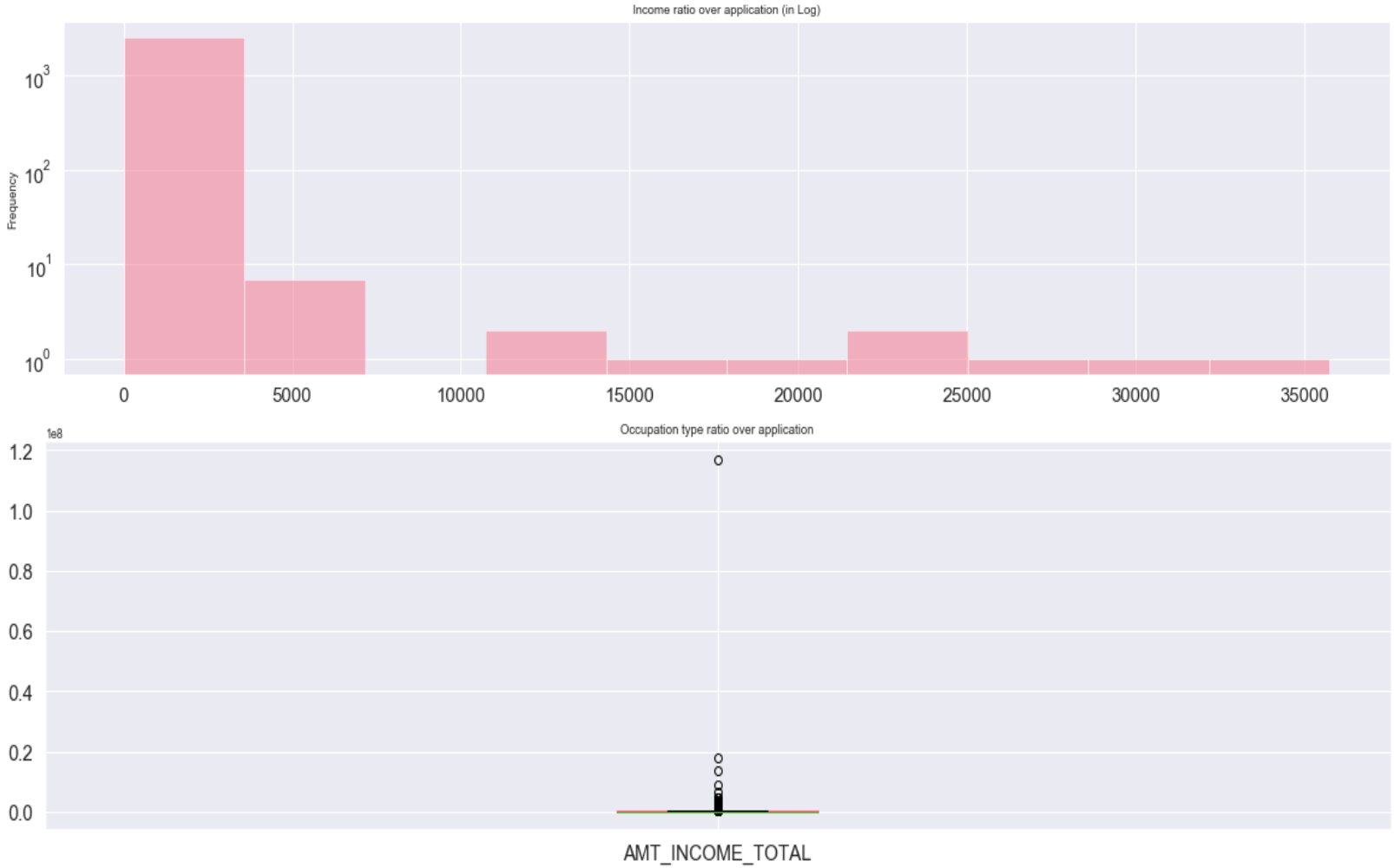
Insight: Here in both the scenario Female clients are taking more loans than Male clients whether its cash loan or revolving loan.



Insight: Here in above case, we observe that Male clients tops in almost every category with respect to the average amount.

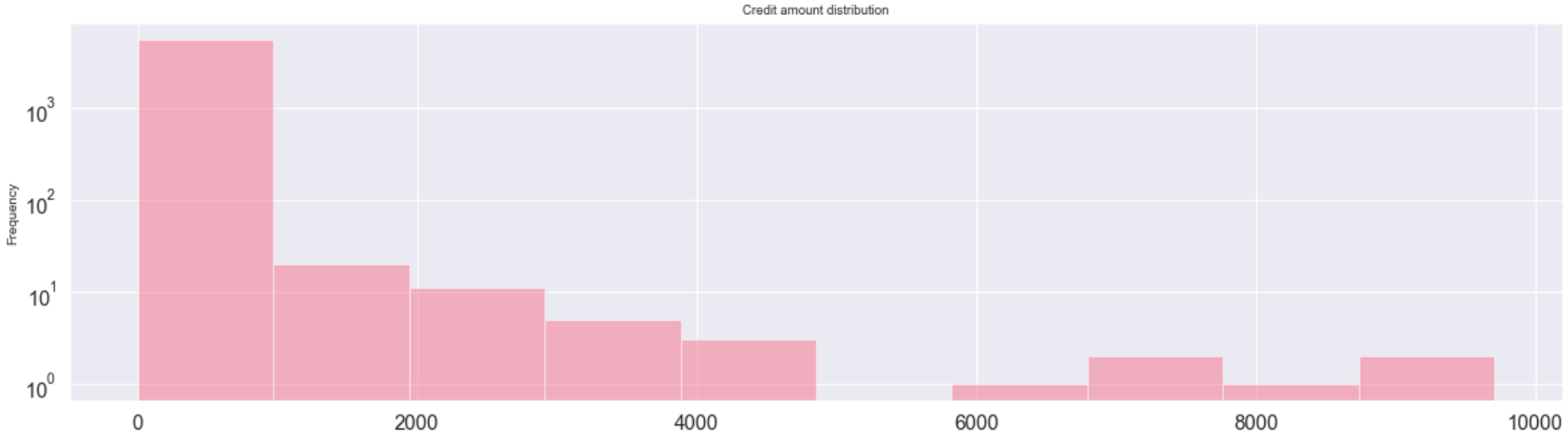
Income Distribution

AMT_INCOME_TOTAL Income Distribution

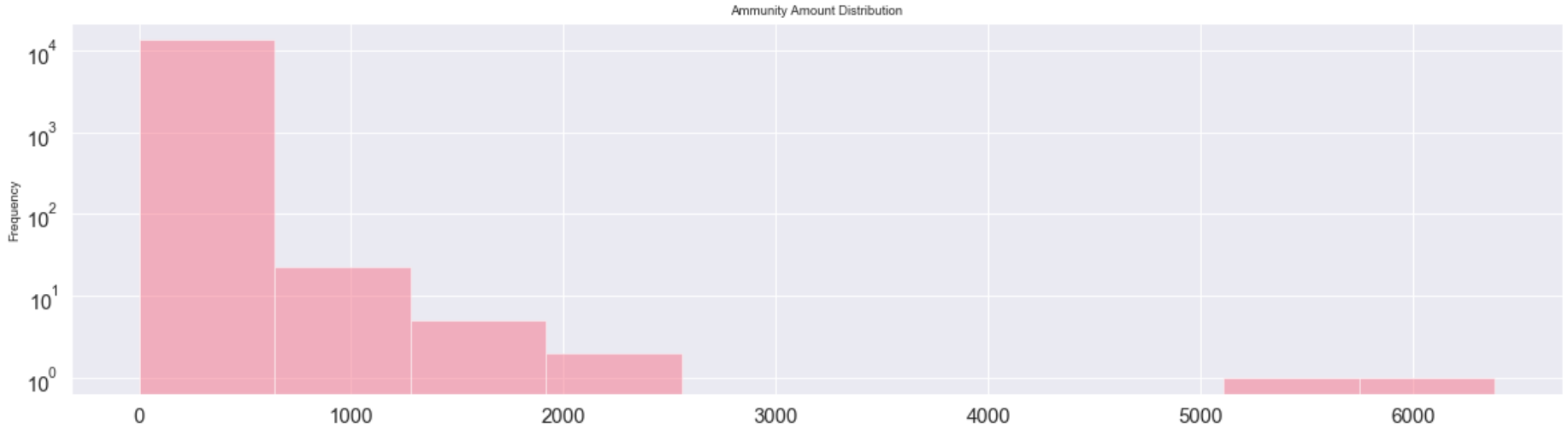


Insight: Mostly data is normally distributed with an exception outlier value in the "AMT_INCOME_TOTAL" column

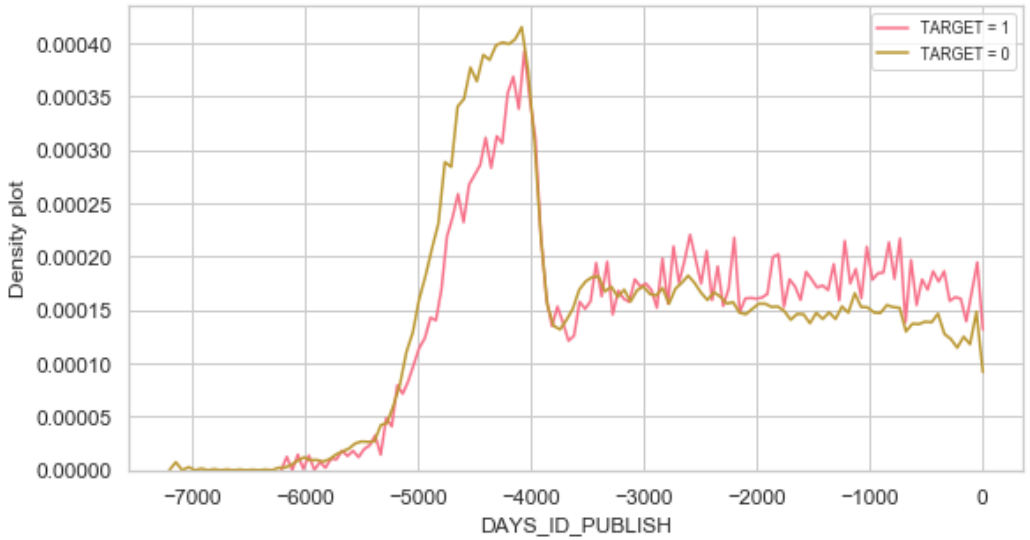
Amount Credit Distribution (In Log values)

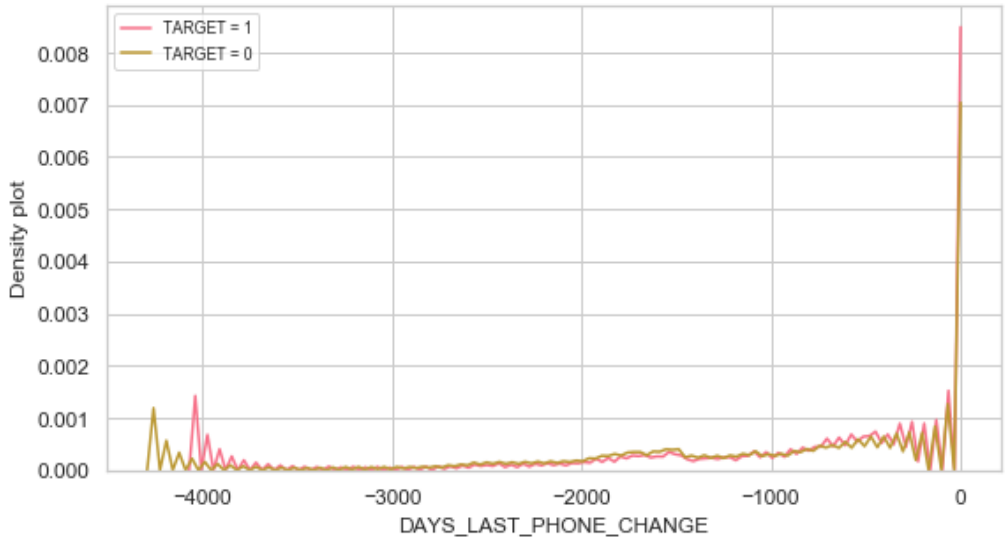


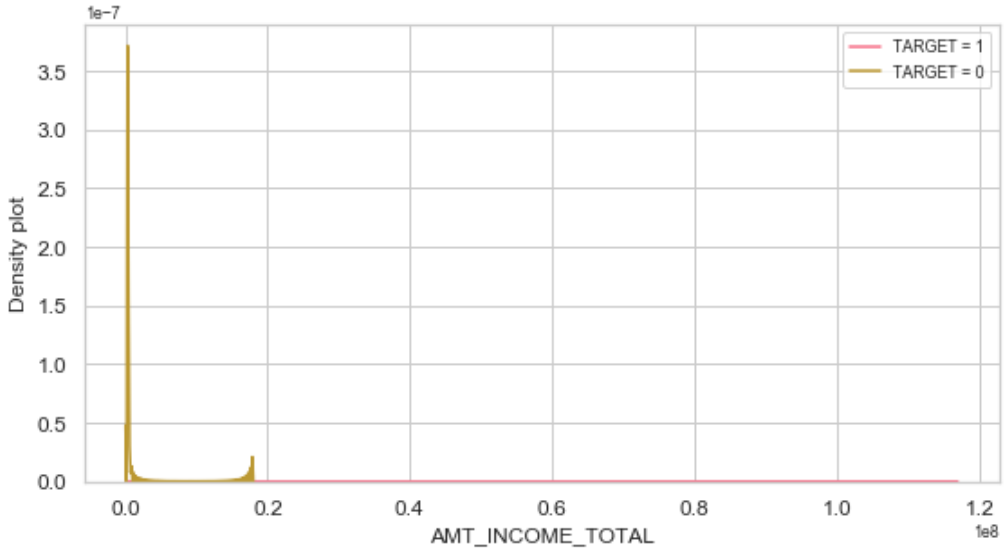
Annuity Distribution (In Log values)

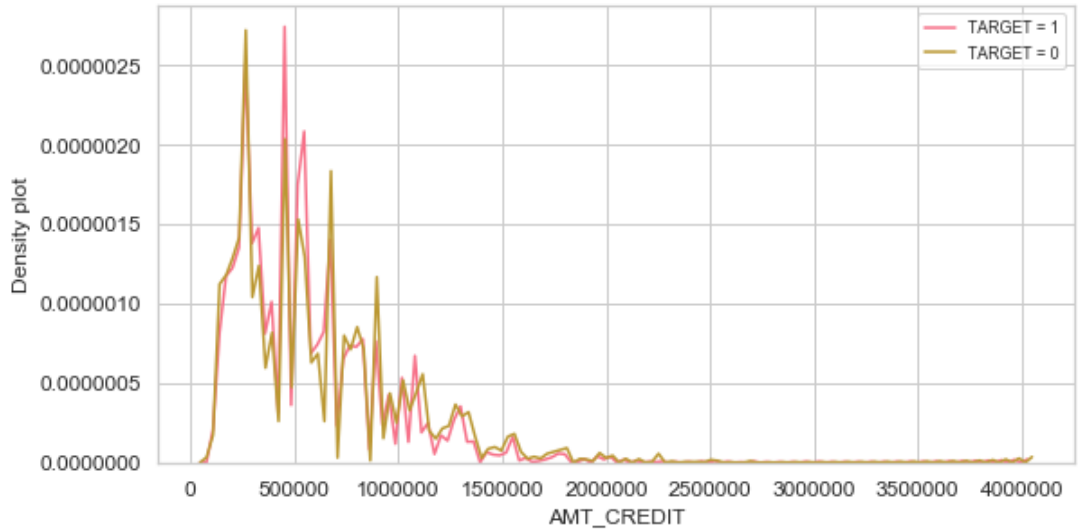


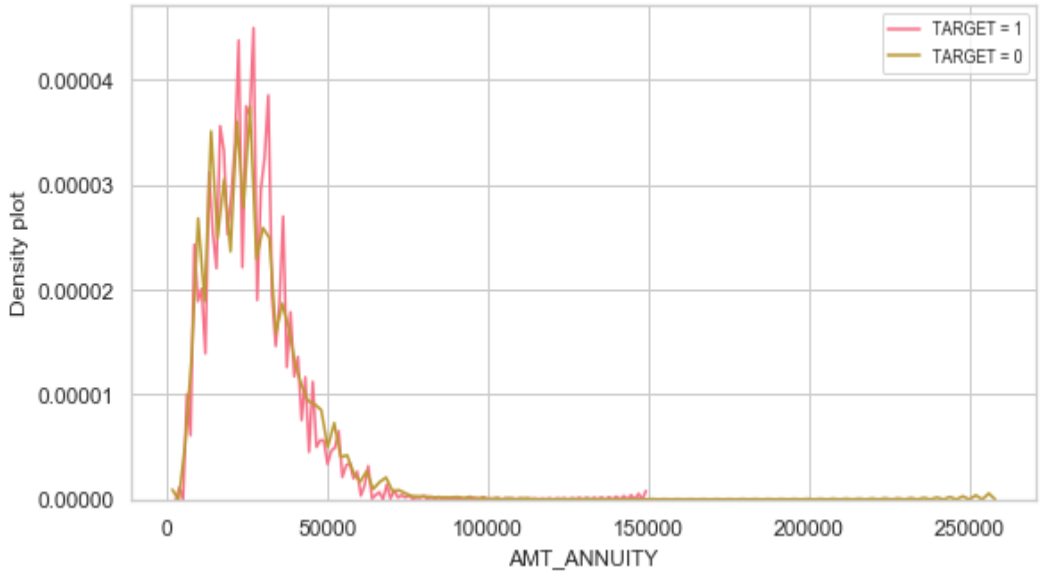
Comparative distribution of various variables for Target=0 and Target=1

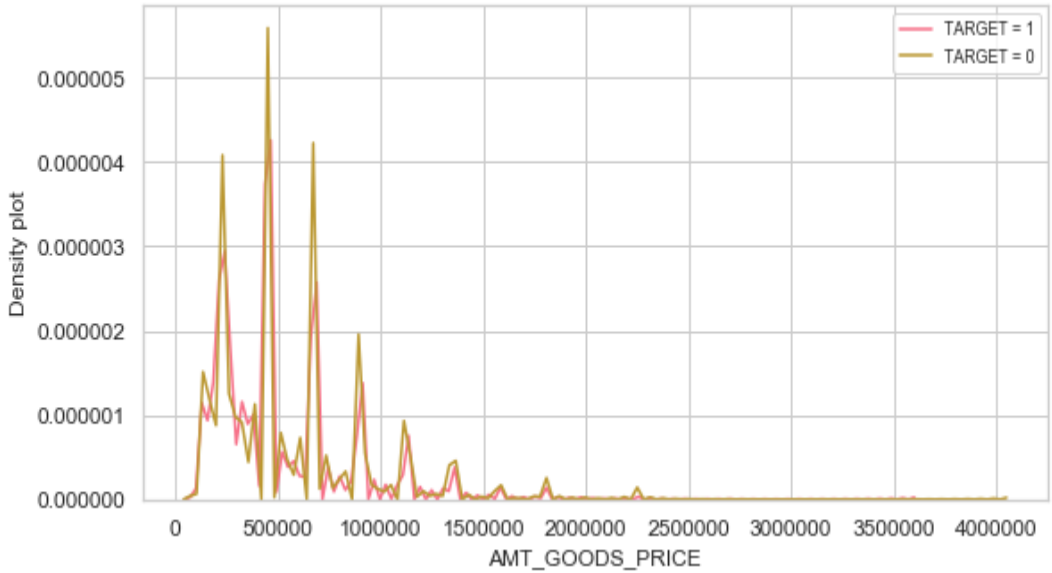


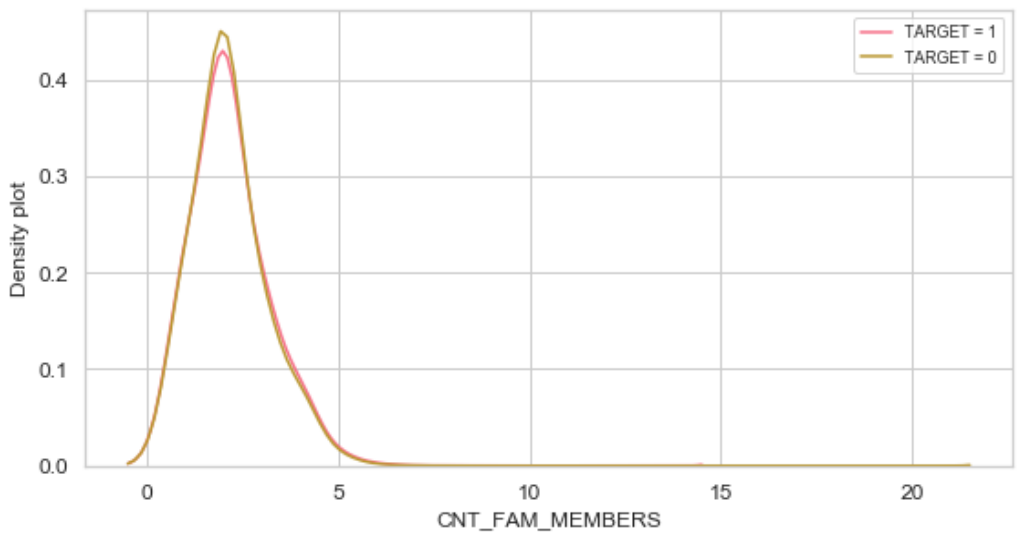


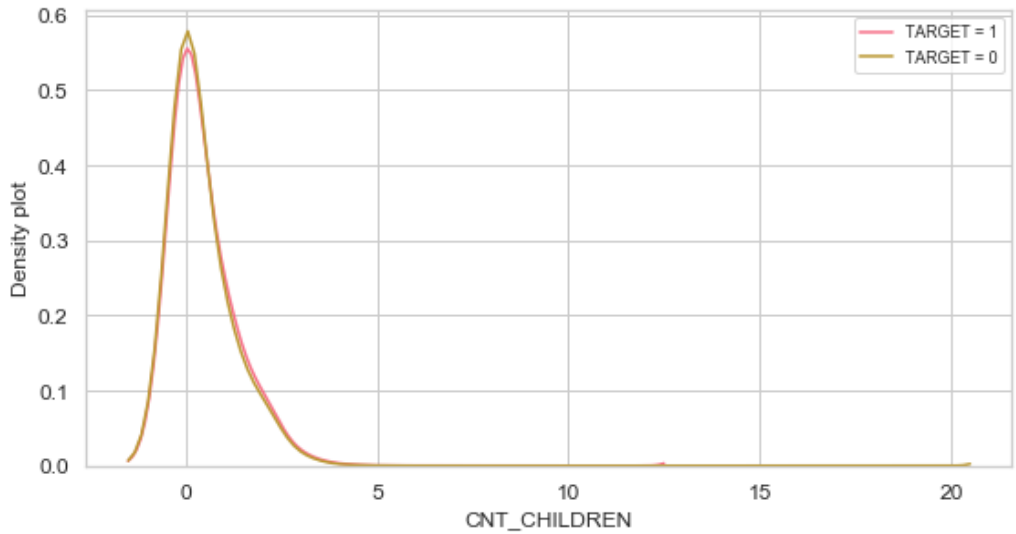


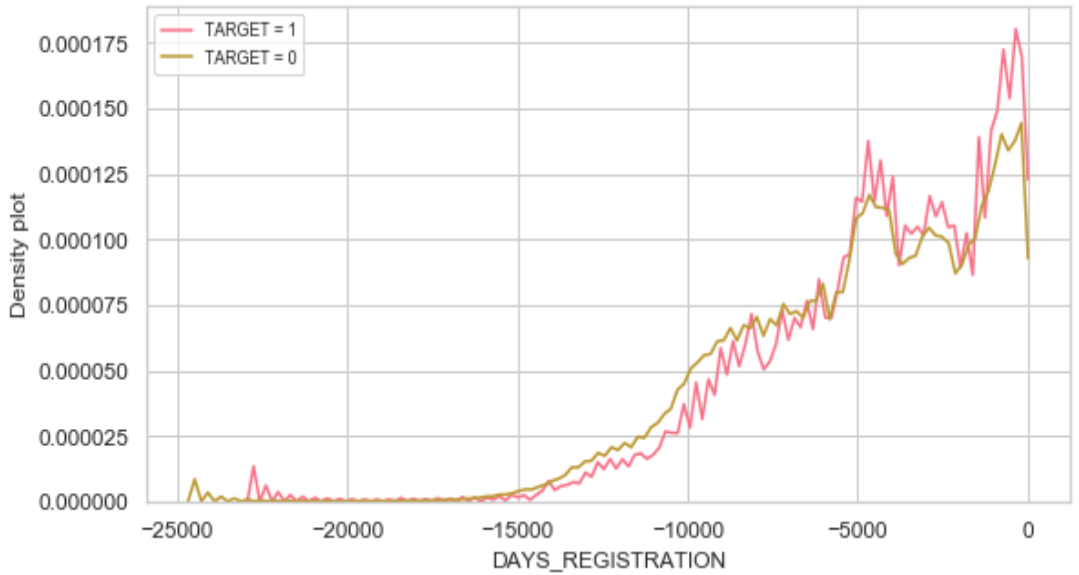


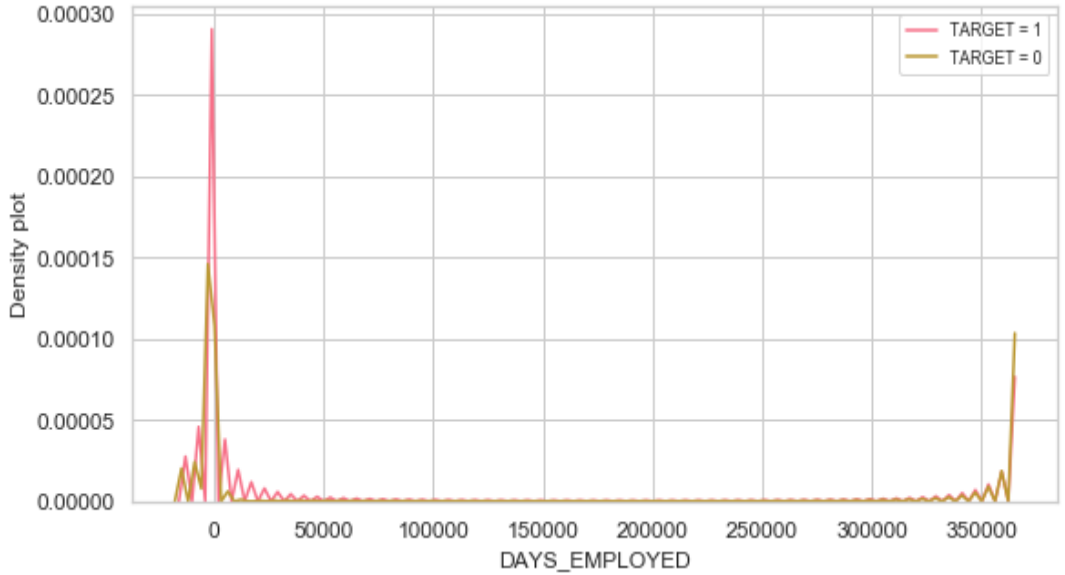

TARGET = 1
TARGET = 0
Density plot
0.00004
0.00003
0.00002
0.00001
0.00000
0
50000
100000
150000
200000
250000
AMT_ANNUITY


TARGET = 1
TARGET = 0
Density plot
0.000005
0.000004
0.000003
0.000002
0.000001
0.000000
0
500000
1000000
1500000
2000000
2500000
3000000
3500000
4000000
AMT_GOODS_PRICE


TARGET = 1
TARGET = 0
Density plot
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
CNT_FAM_MEMBERS


TARGET = 1
TARGET = 0
Density plot
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
CNT_CHILDREN

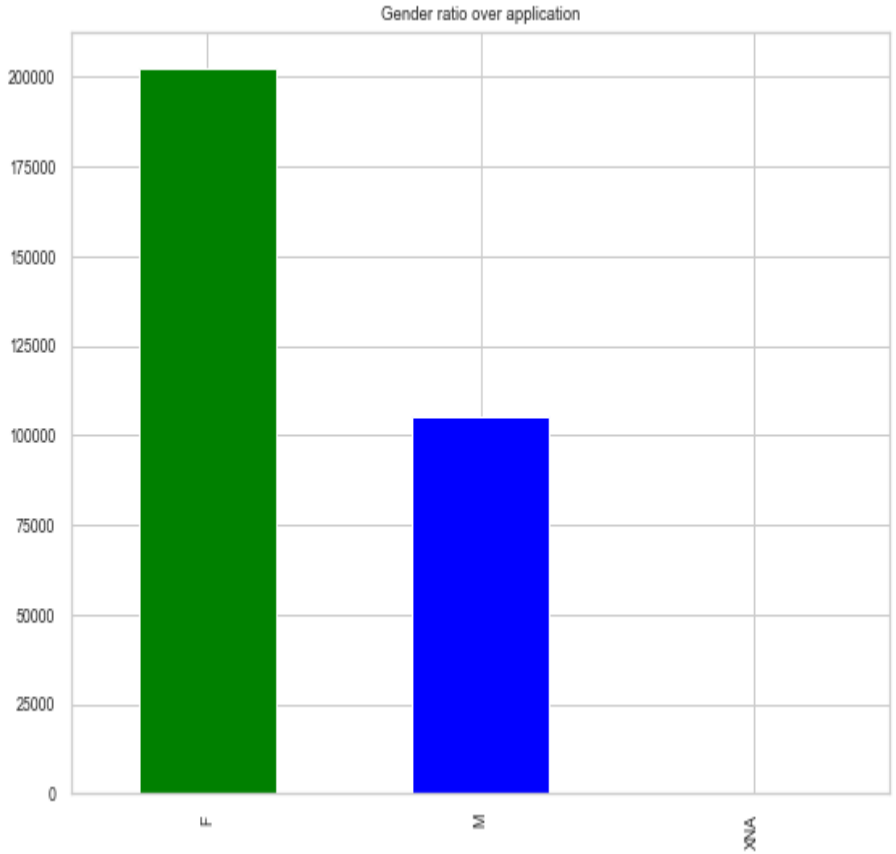


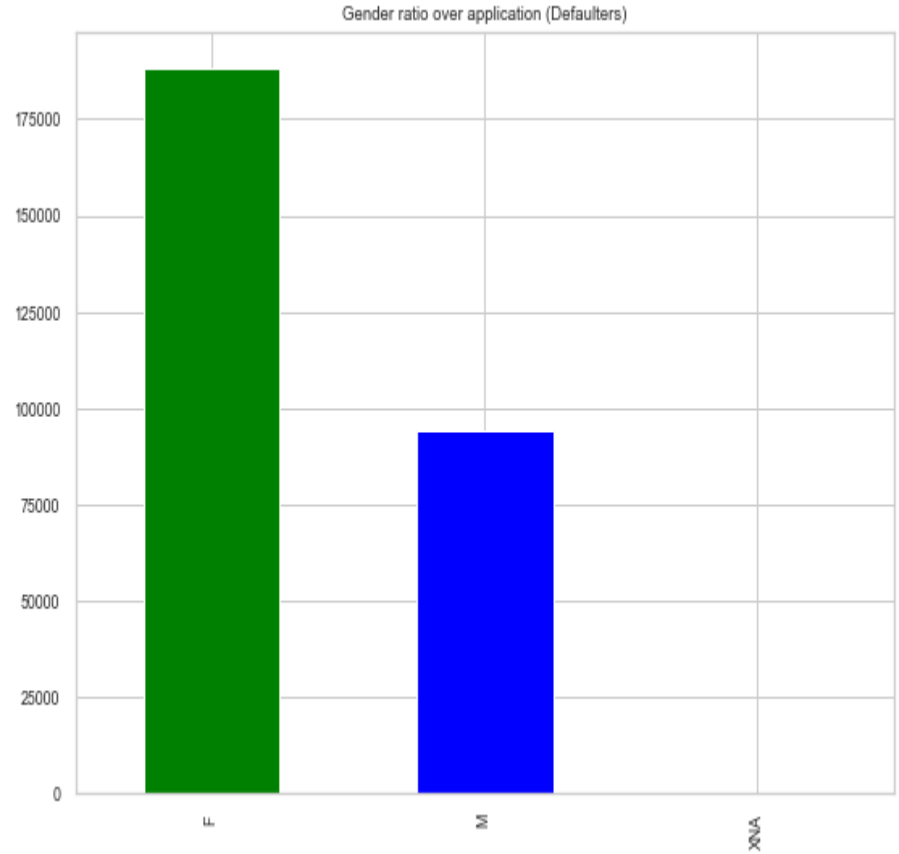


Insight: In every graph the defaulters are fluctuating mostly same as other variable fluctuate.

Ratio on the various categorical dimensions

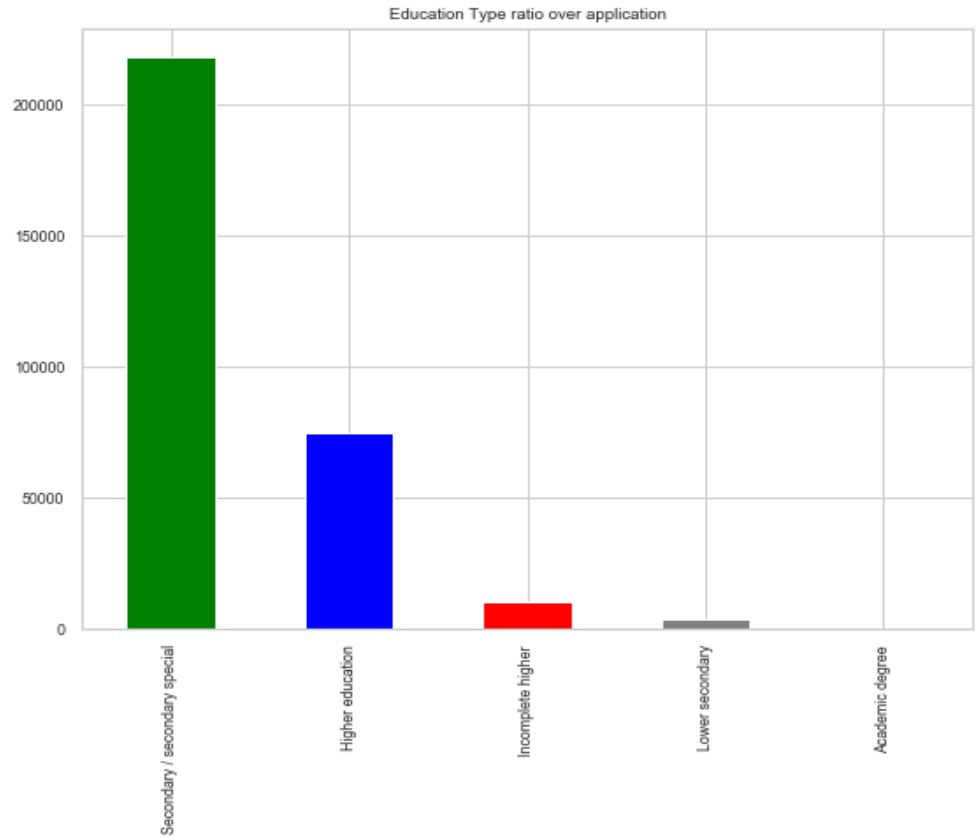
Gender Ratio over application

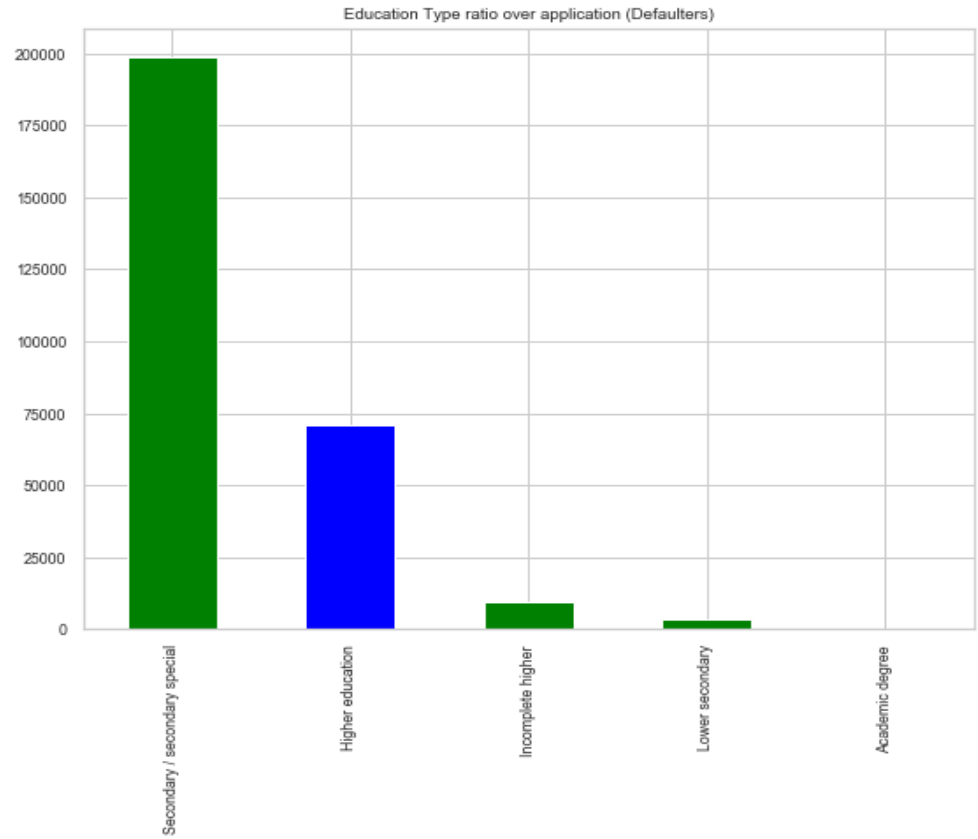




Insight: Female clients are most likely to be defaulters than the male clients

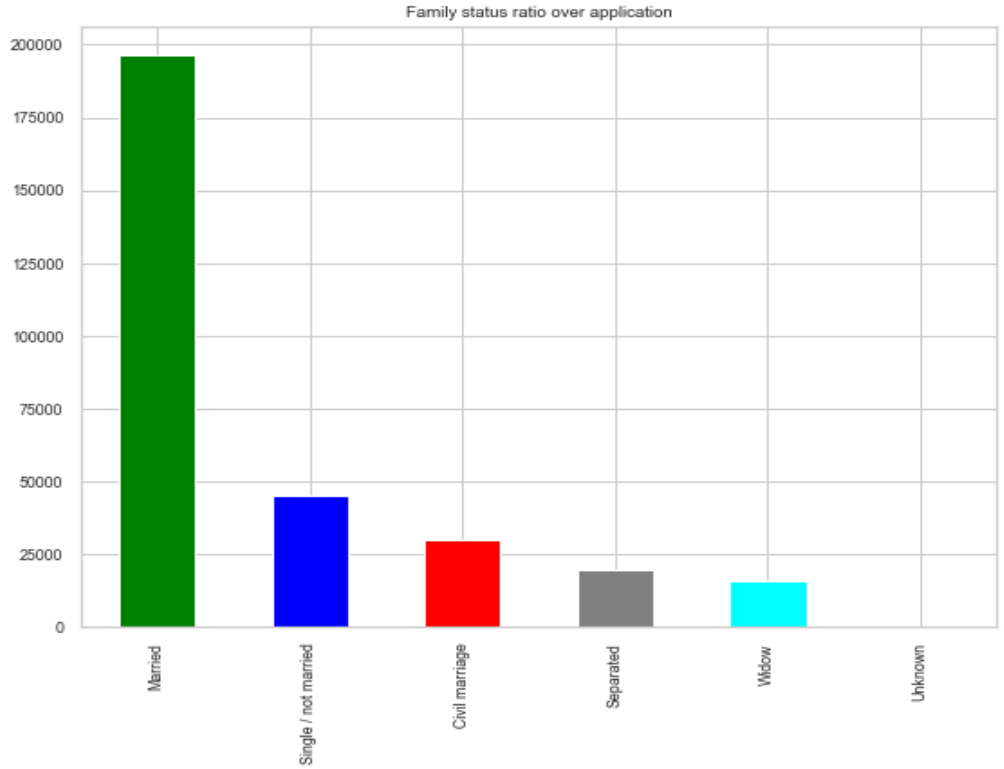
Education Ratio over application

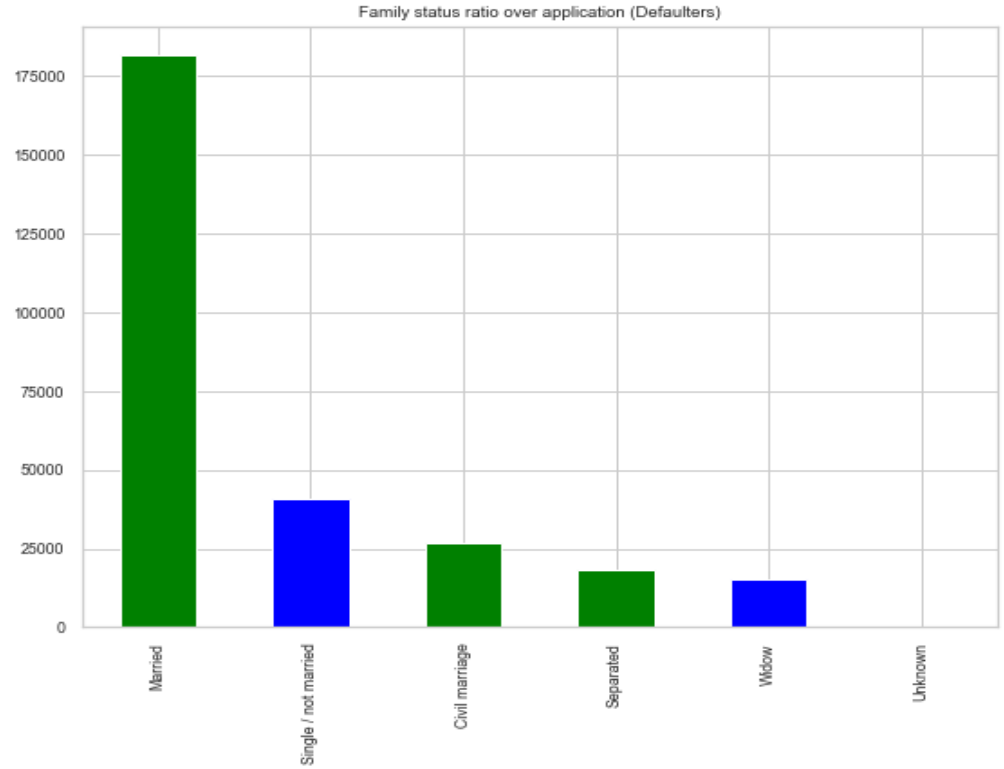




Insight: "Secondary/Secondary Special" Education is more likely to be defaulters.

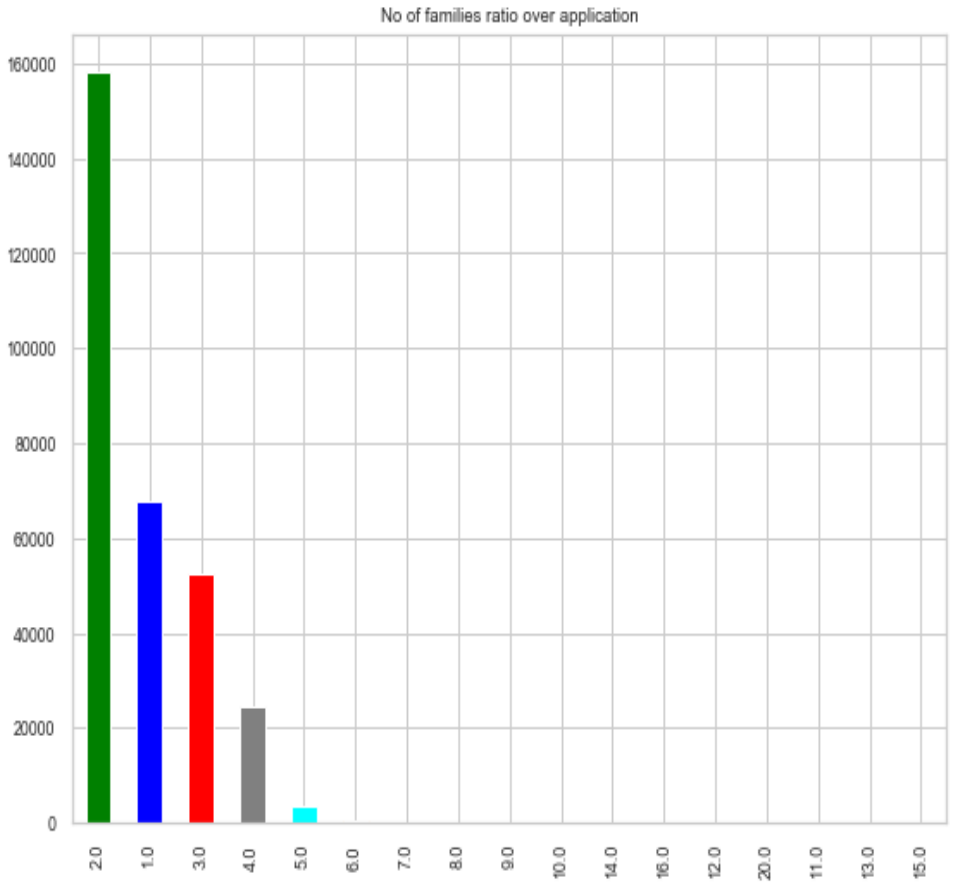
Family Status Ratio over application

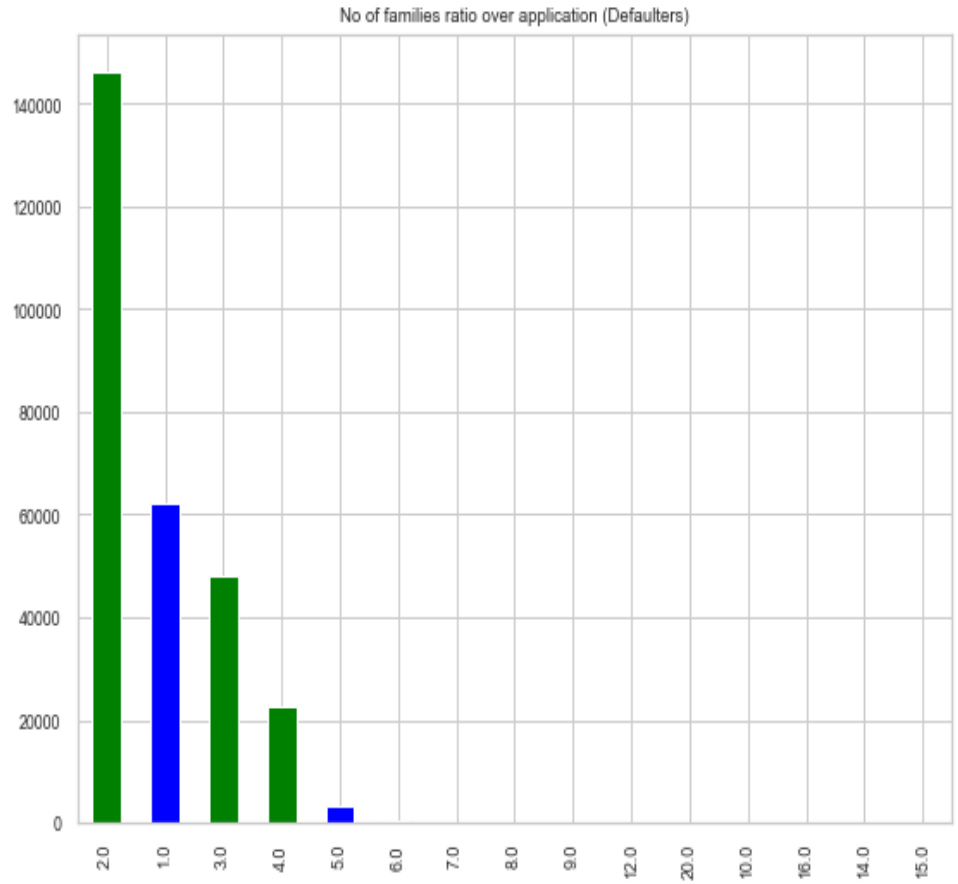




Insight: Married people are most likely to be defaulters

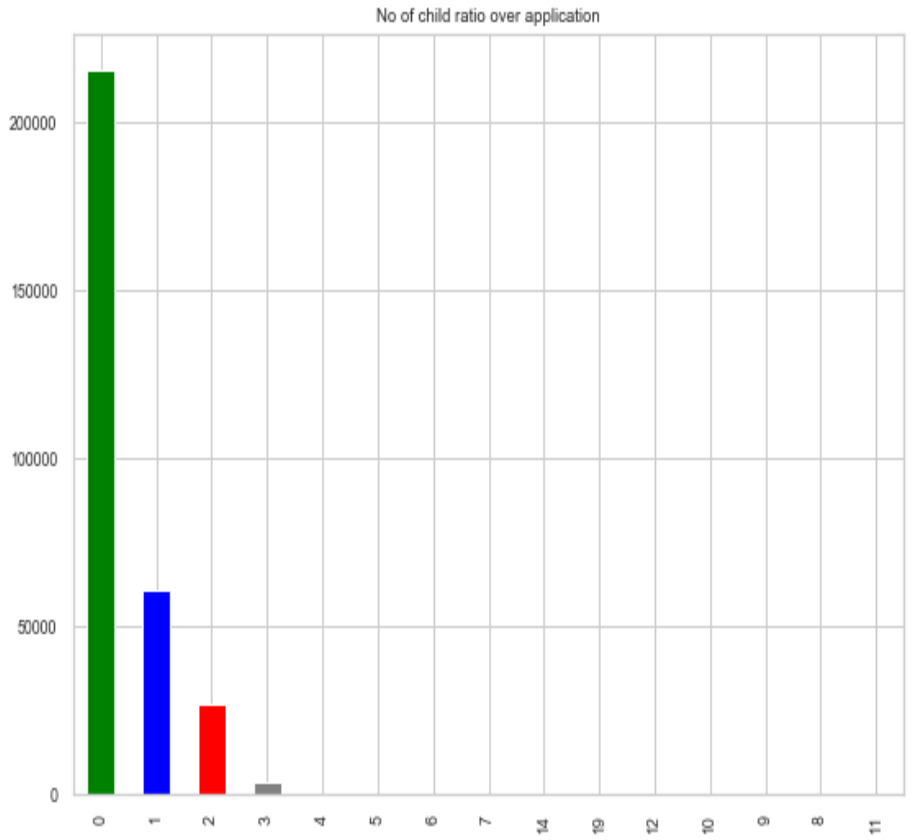
No of families ratio over application : CNT_FAM_MEMBERS

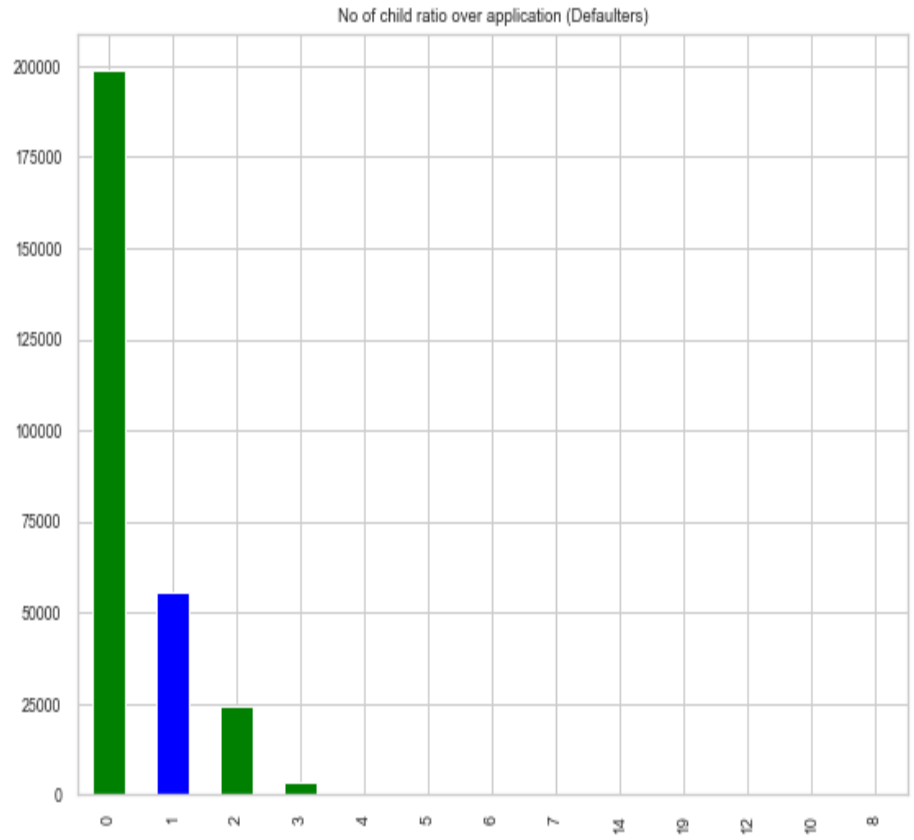




Insight: People who has two family members are most likely to be defaulters.

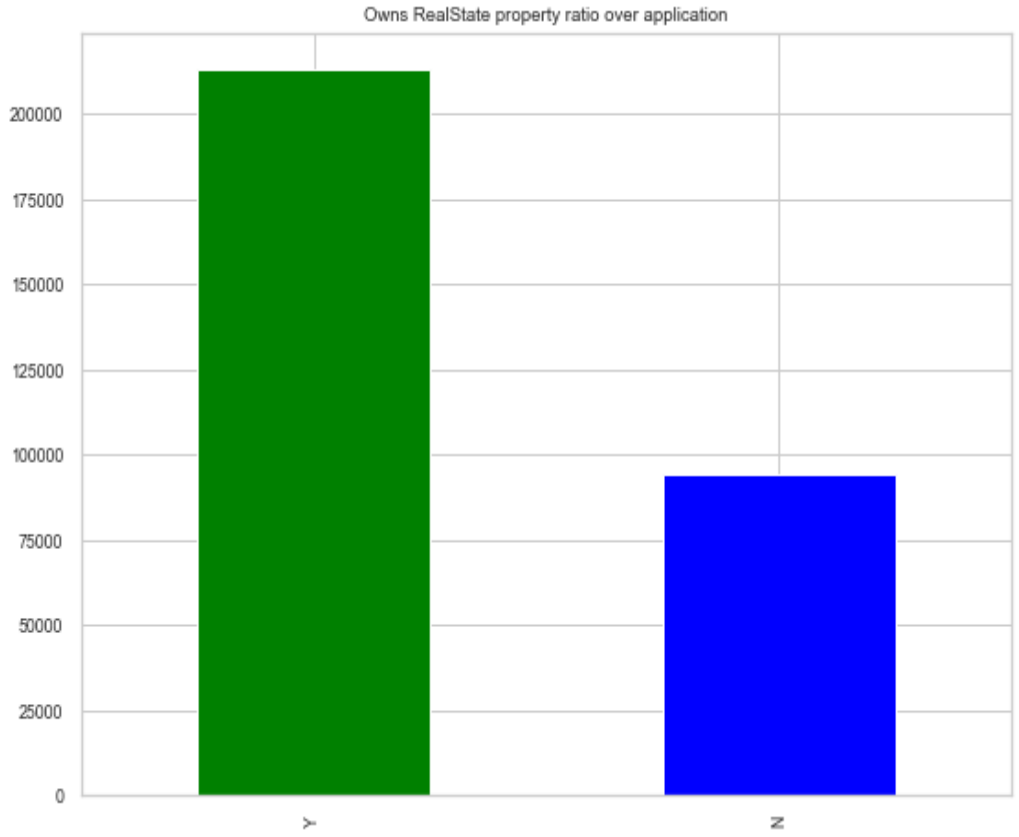
CNT_CHILDREN ratio over application

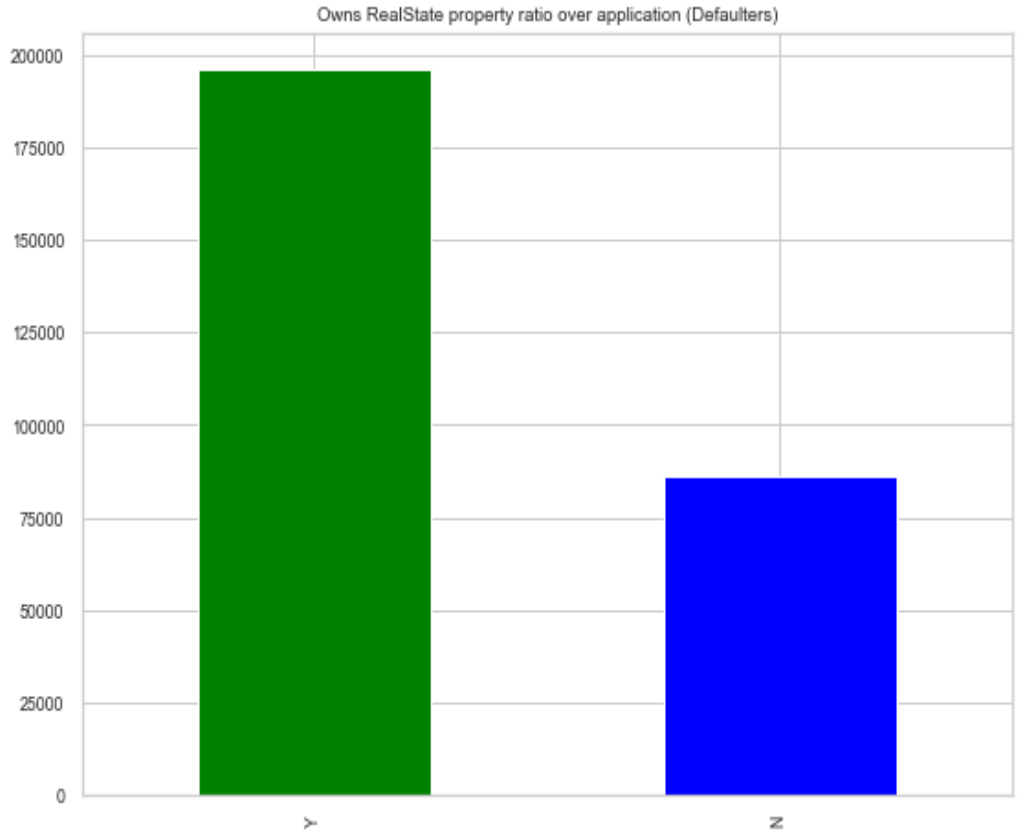




Insight: People who has no children in their family are most likely to be defaulters.

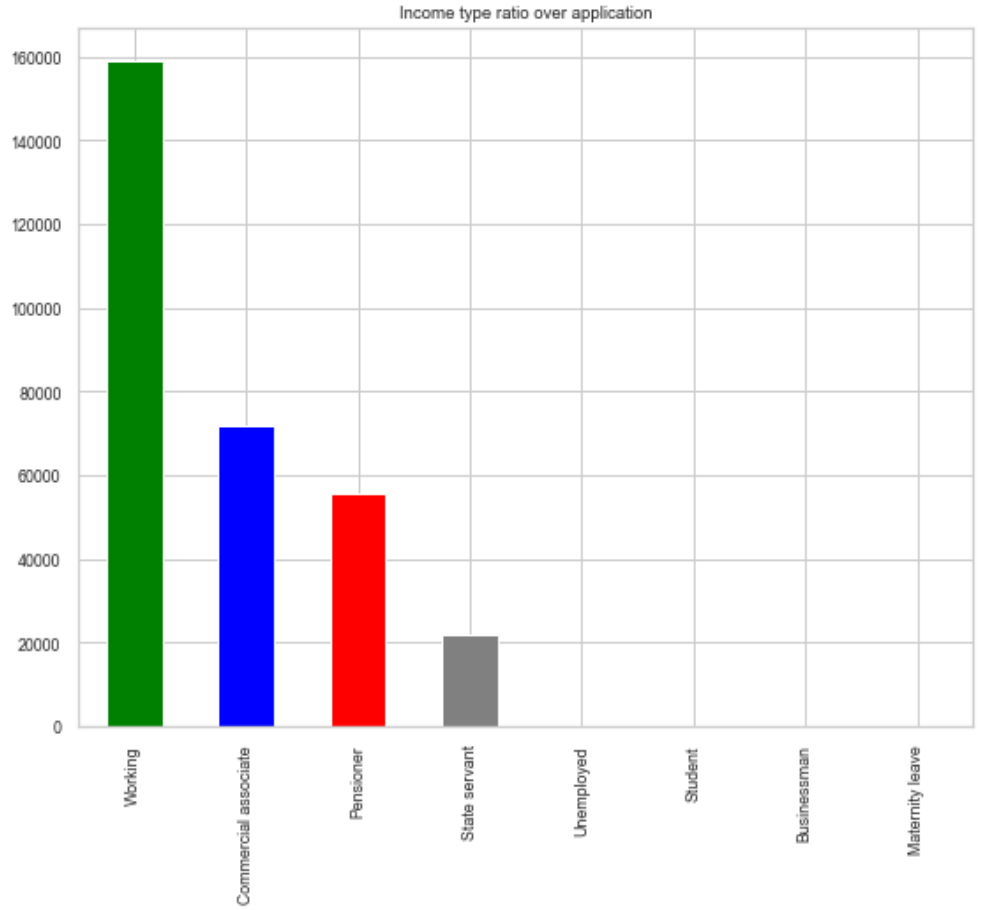
FLAG_OWN_REALTY

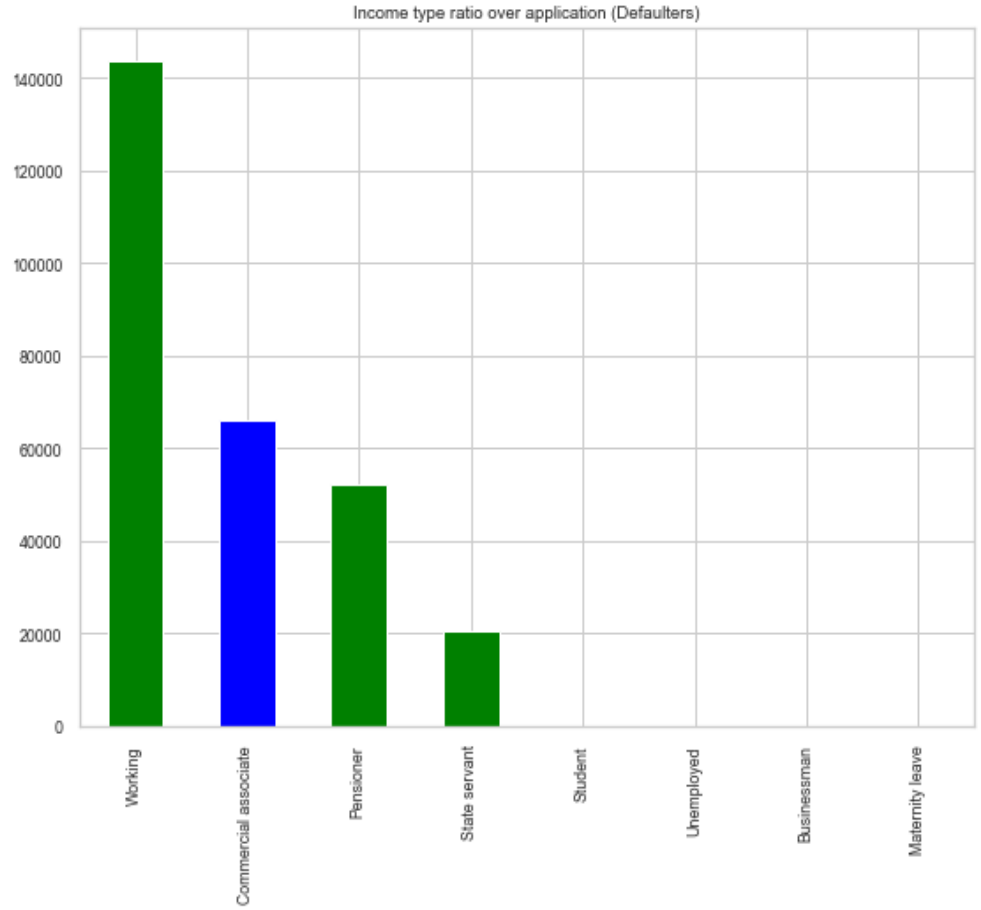




Insight: People who has Real State properties are most likely to be defaulters.

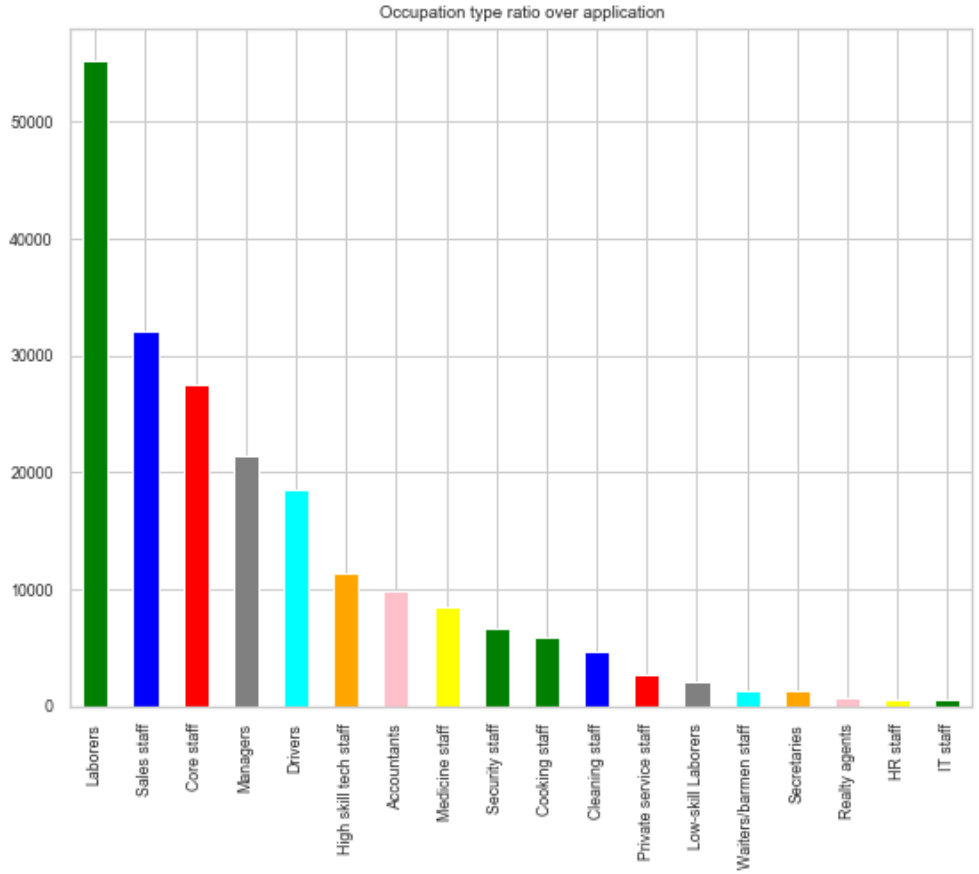
Income type ratio over application

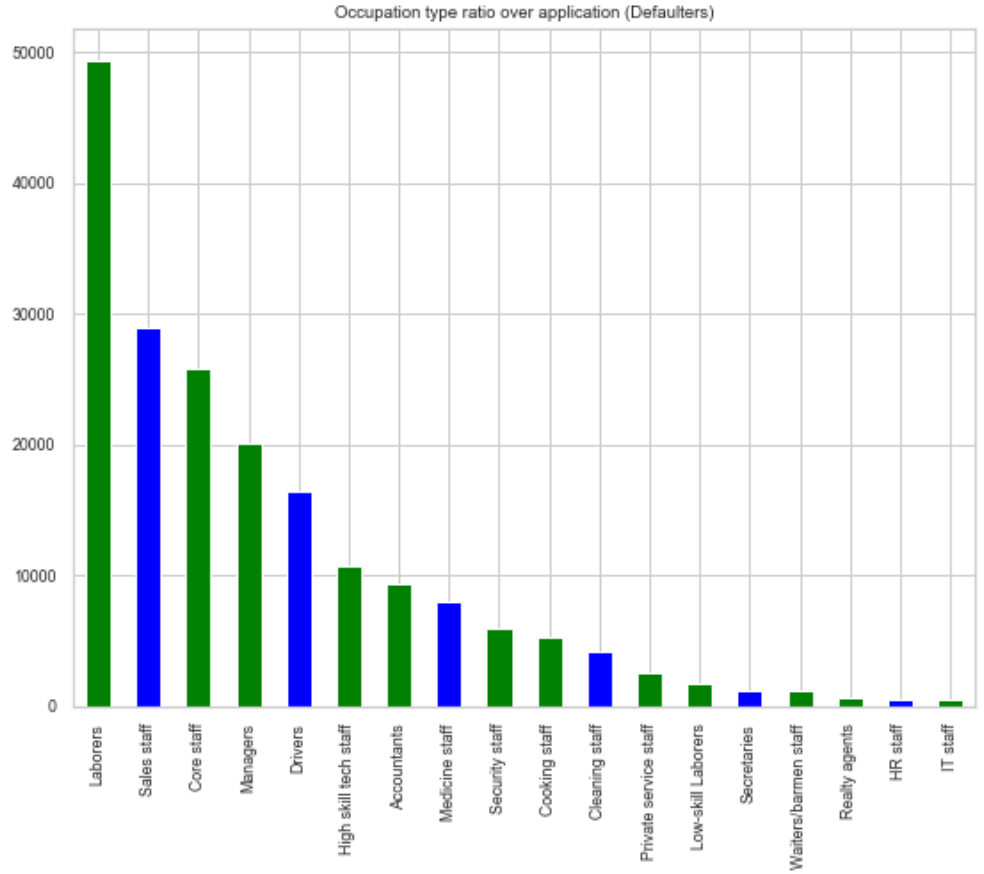




Insight: People who are Working, are most likely to be defaulters.

OCCUPATION_TYPE ratio over application





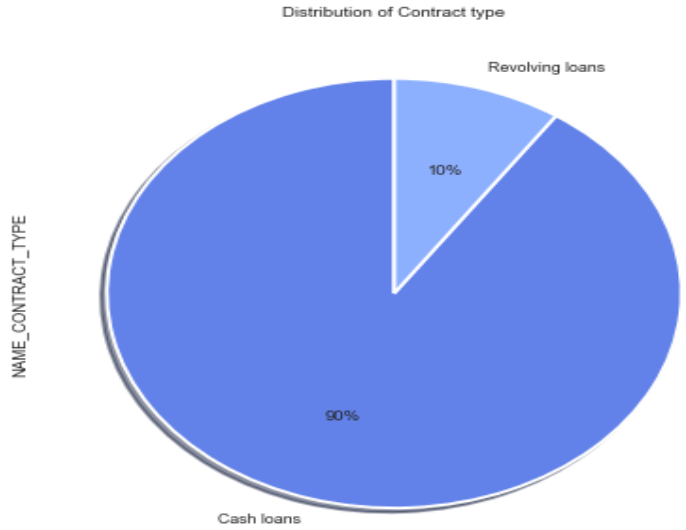
Insight: People who are Laborers, are most likely to be defaulters.

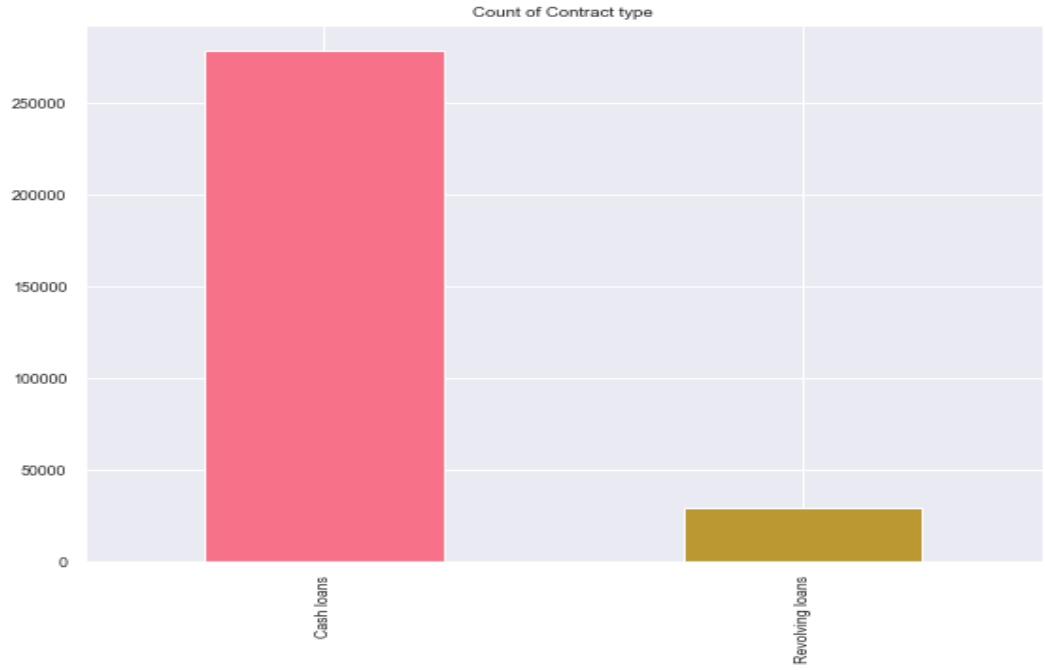
SEGMENTED UNIVARIATE ANALYSIS:

Univariate analysis is perhaps the simplest form of statistical analysis. Like other forms of statistics, it can be inferential or descriptive. The key fact is that only one variable is involved. Univariate analysis can yield misleading results in cases in which multivariate analysis is more appropriate

Insight: 10% out of total client population have difficulties in revolving loans.

Insight: 90% are of cash loans.

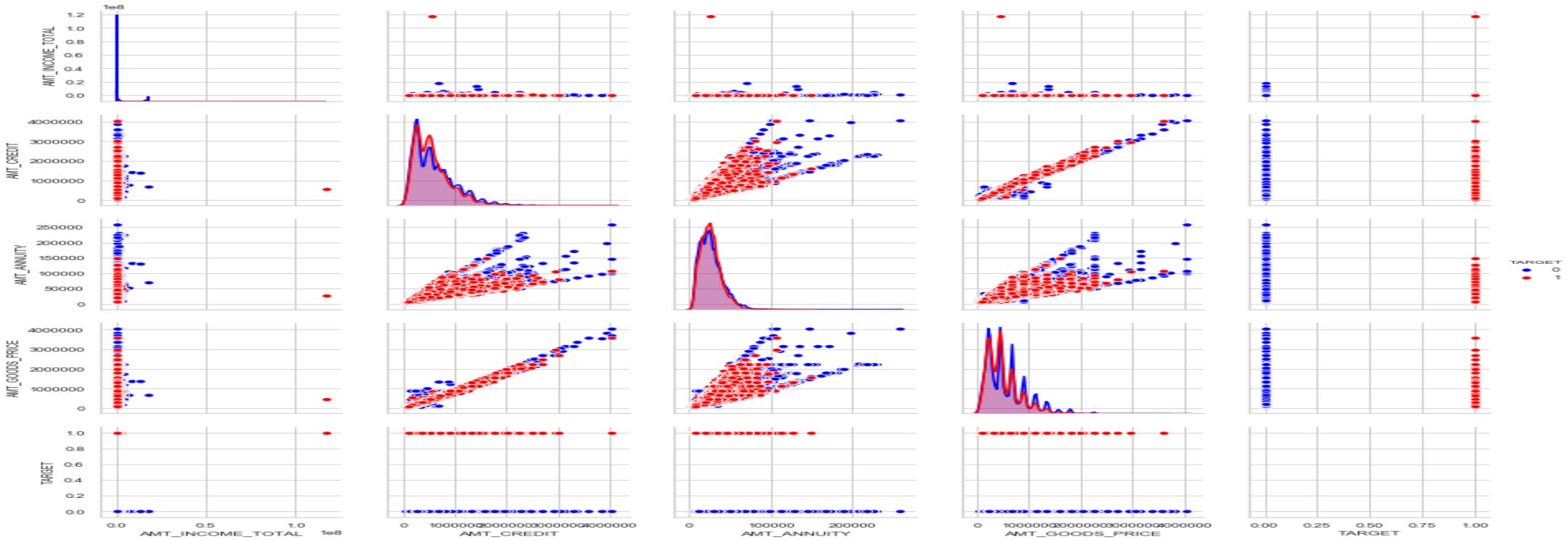




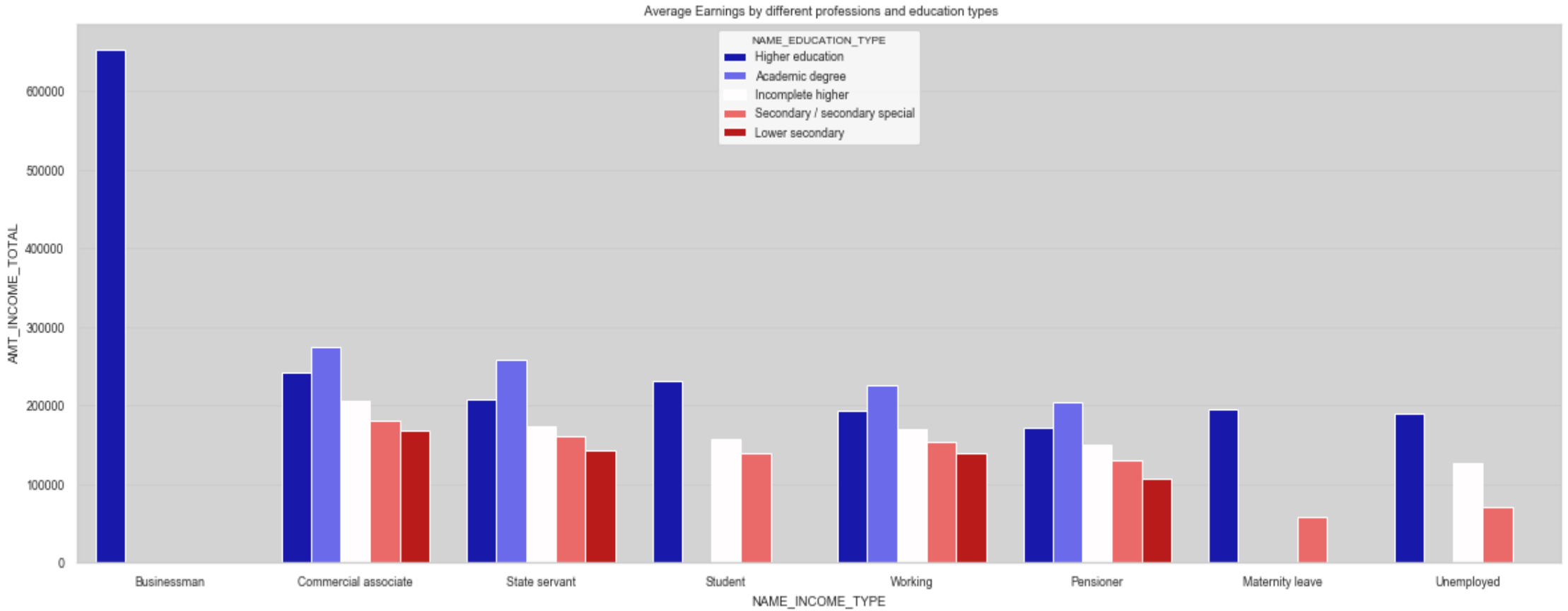
BIVARIATE ANALYSIS:

Bivariate analysis is the simultaneous analysis of two variables (attributes). It explores the concept of relationship between two variables, whether there exists an association and the strength of this association, or whether there are differences between two variables and the significance of these differences.

Bivariate Analysis For Amount

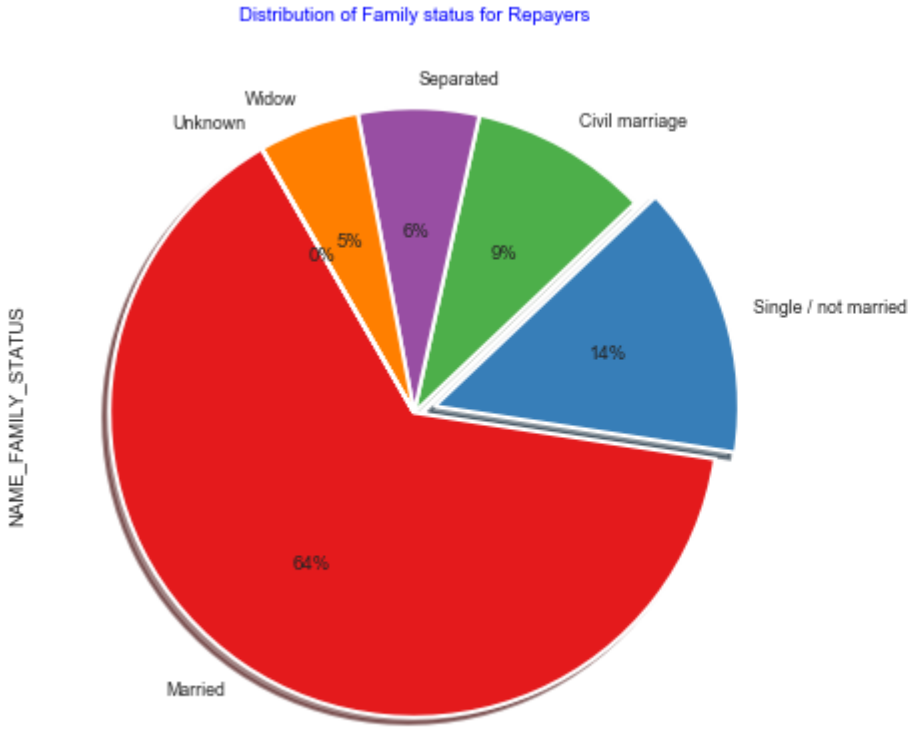


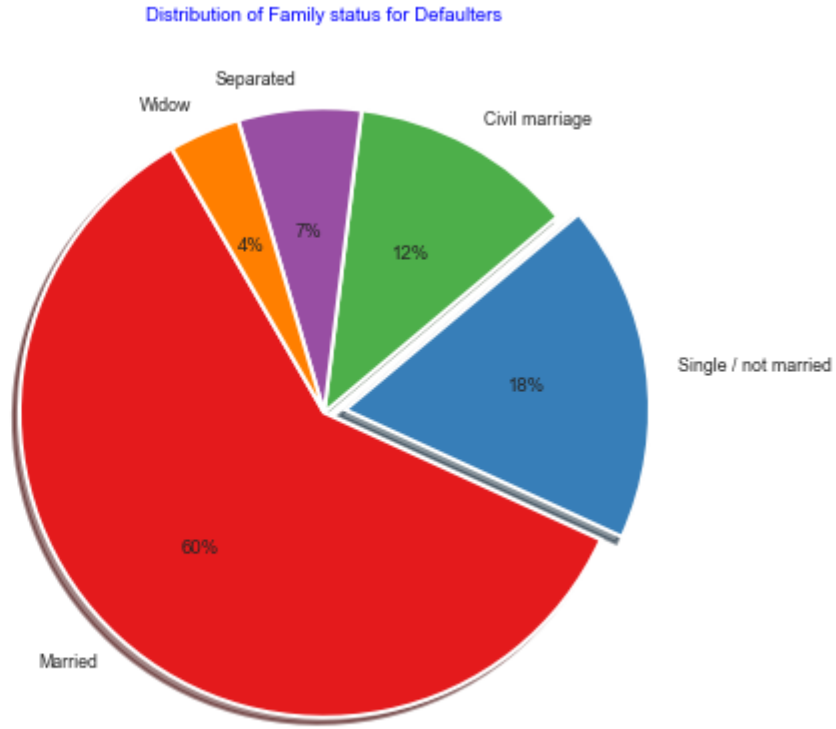
Average Earnings by different professions and education types



Insight: Businessman and Students who are having Higher education, are having more Income. For the rest, who are having Academic degree are having more income.

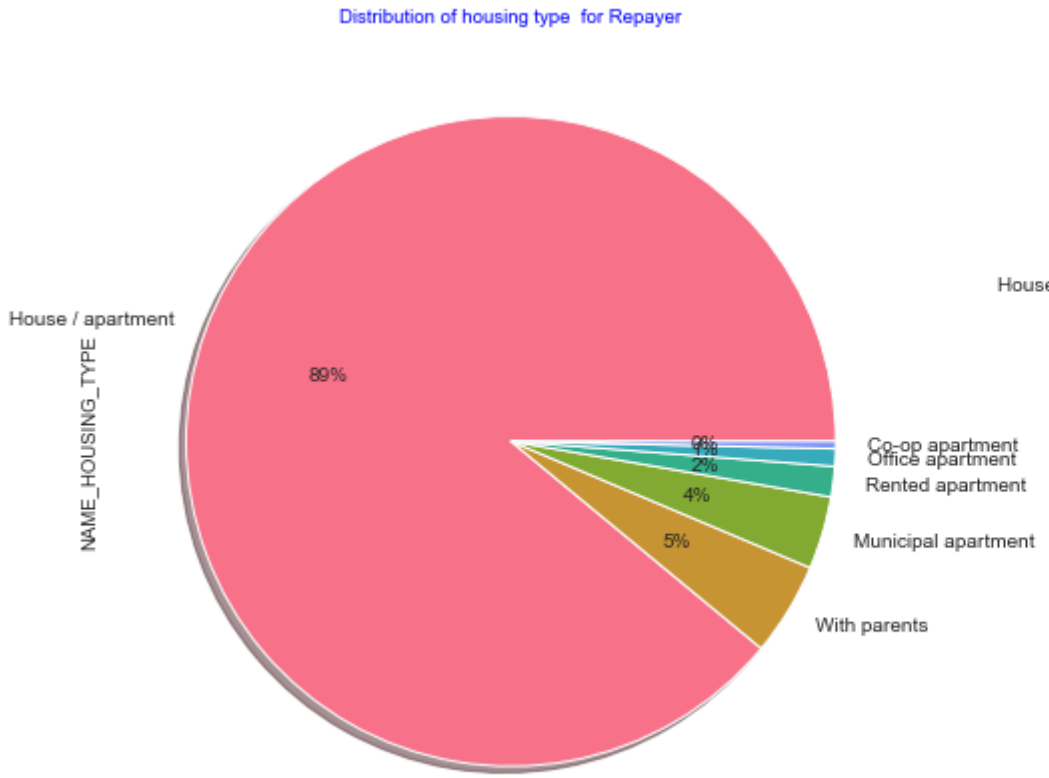
Distribution of Education type by loan repayment status

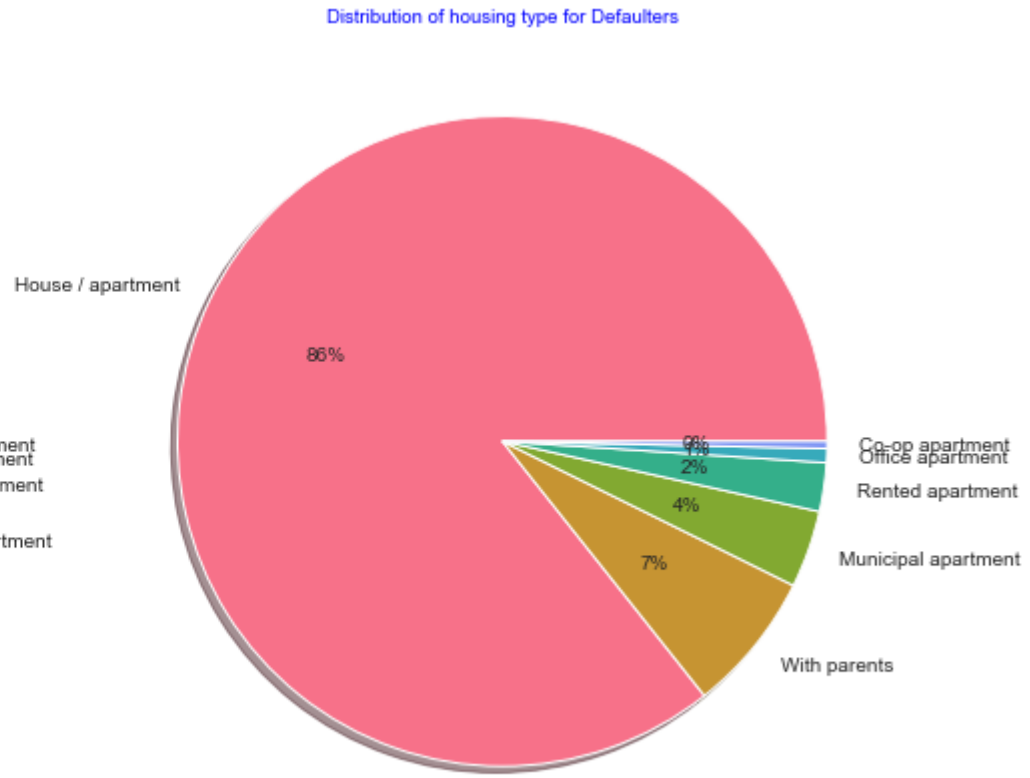




Insight: People with 60% who are married, are most likely to be defaulters.

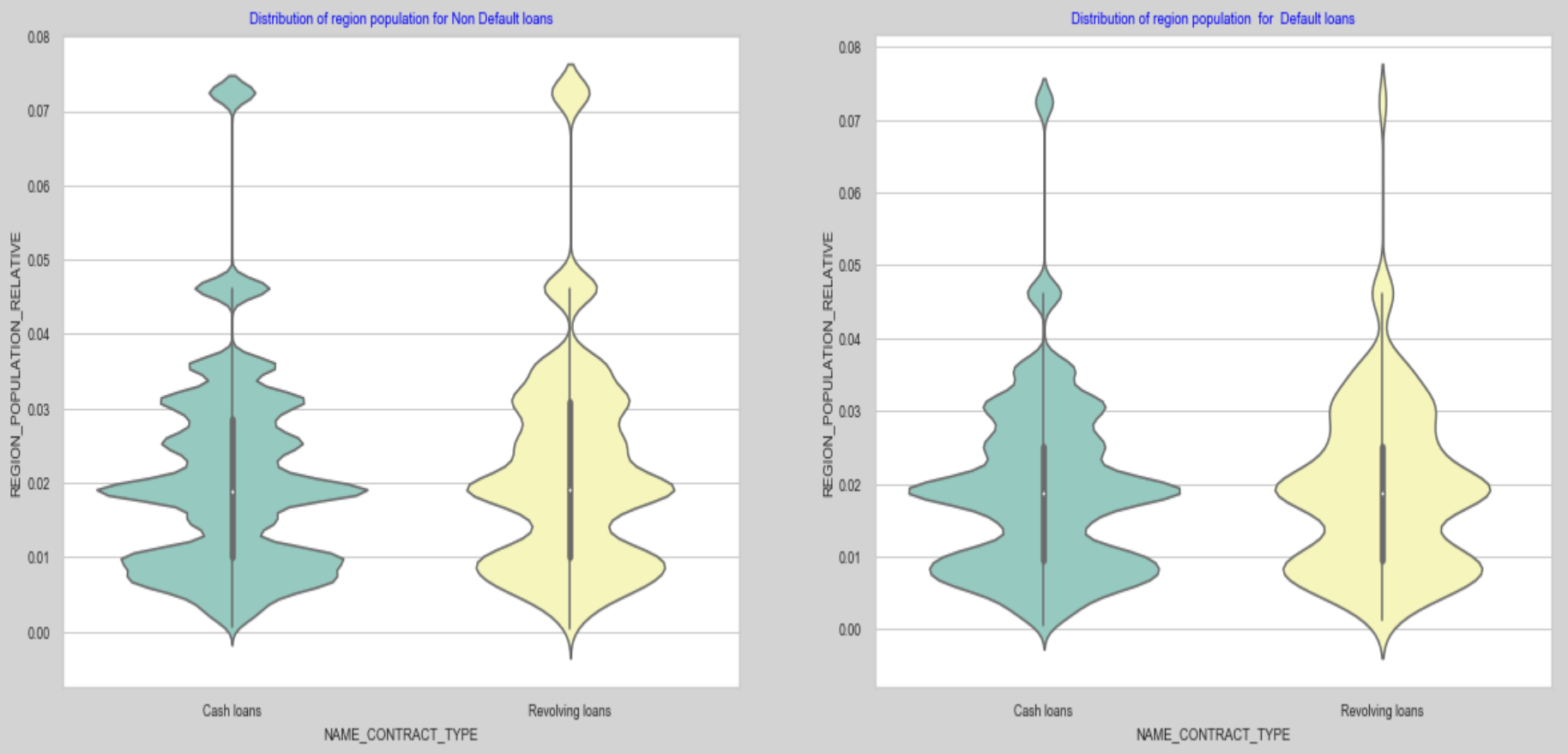
Distribution of housing type for Defaulters





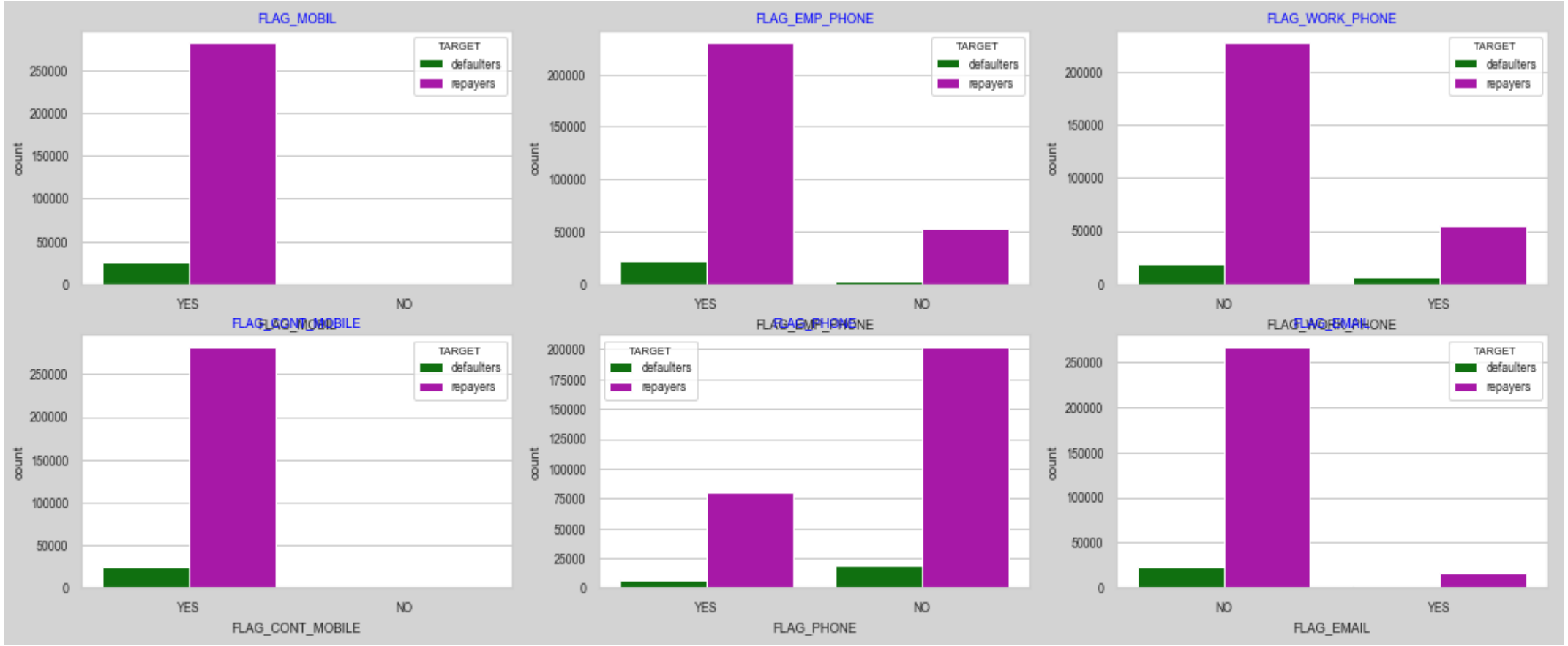
Insight: People(mostly 86%) who lives in House/Apartment, are most likely to be defaulters.

Distribution normalized population of region where client lives by loan repayment status

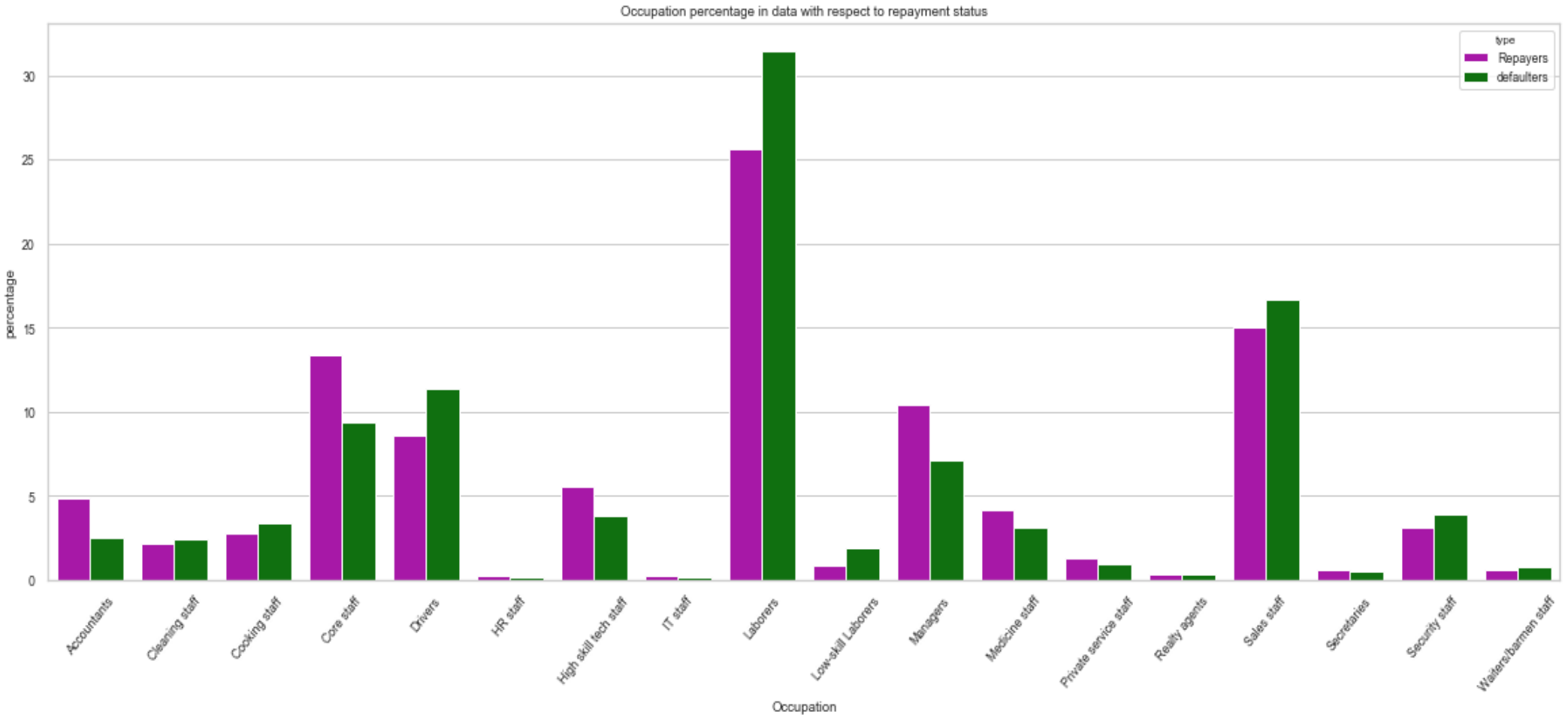


Insight: In High population density regions people are less likely to default on loans.

Distribution of contact information provided by client

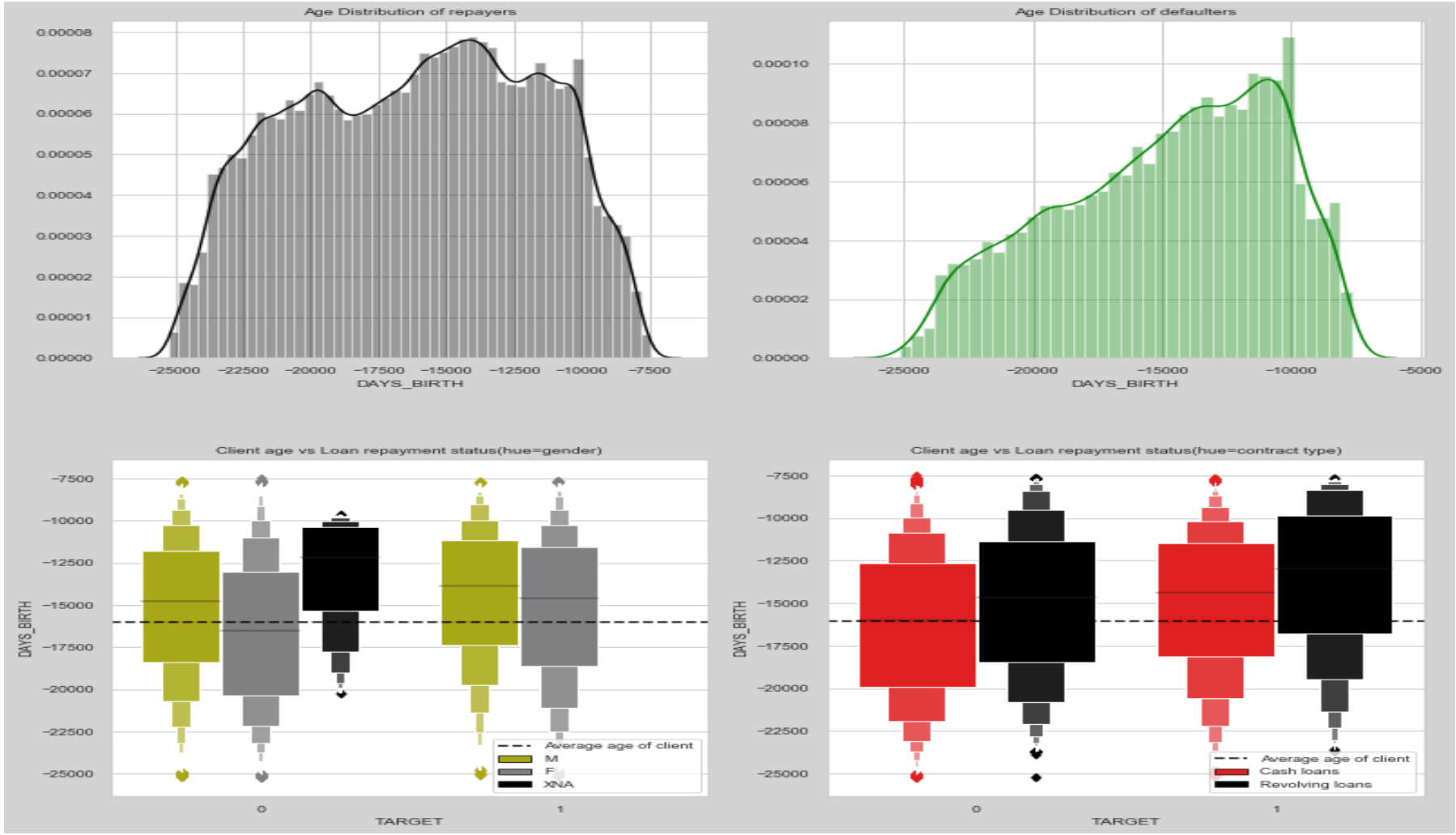


Occupation percentage in data with respect to repayment status



Insight: Occupations like Cleaning staff ,Cooking staff, Drivers ,Laborers , Low-skill Laborers ,Sales staff, Security staff are more likely to default in loans.

Distribution of Client's age



Insight: As per the graph we observe that people with middle and old ages are repayors and mostly old age people are defaulters

CORRELATION

In statistics, dependence or association is any statistical relationship, whether causal or not, between two random variables or bivariate data. In the broadest sense correlation is any statistical association, though it commonly refers to the degree to which a pair of variables are linearly related.

Correlation between variables (Categorized by "TARGET" variable)

Getting Correlation from ***New Loan Application***

	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
SK_ID_CURR	1.000000	-0.002108	-0.001129	-0.001820	-0.000343	-0.000433	-0.000232
TARGET	-0.002108	1.000000	0.019187	-0.003982	-0.030369	-0.012817	-0.039645
CNT_CHILDREN	-0.001129	0.019187	1.000000	0.012882	0.002145	0.021374	-0.001827
AMT_INCOME_TOTAL	-0.001820	-0.003982	0.012882	1.000000	0.156870	0.191657	0.159610
AMT_CREDIT	-0.000343	-0.030369	0.002145	0.156870	1.000000	0.770138	0.986968

rows × 90 columns

Getting Correlation from ***Previous Loan Application***

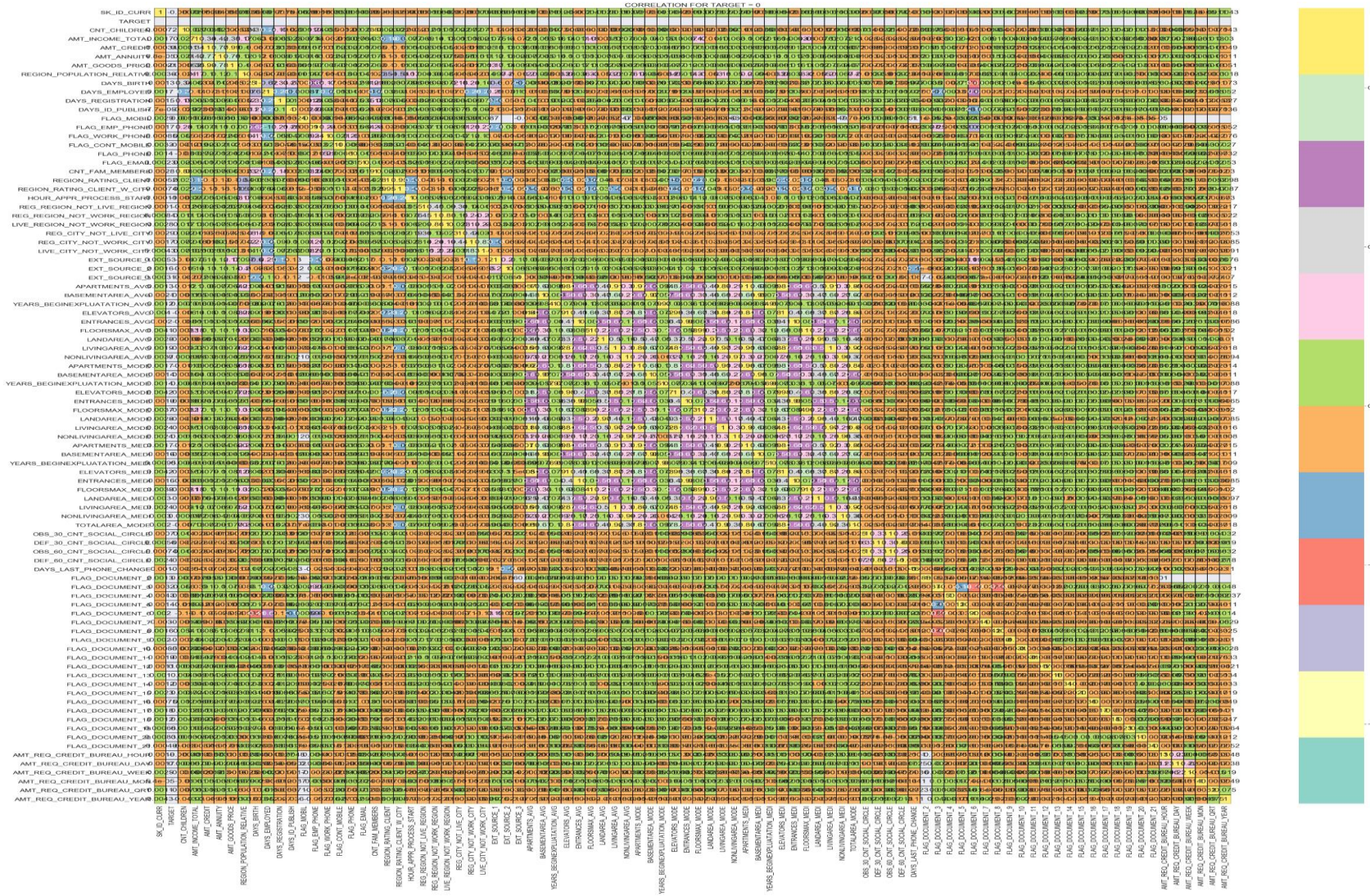
	SK_ID_PREV	SK_ID_CURR	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE
SK_ID_PREV	1.000000	-0.000321	0.011459	0.003302	0.003659	-0.001313	0.015293
SK_ID_CURR	-0.000321	1.000000	0.000577	0.000280	0.000195	-0.000063	0.000369
AMT_ANNUITY	0.011459	0.000577	1.000000	0.808872	0.816429	0.267694	0.820895
AMT_APPLICATION	0.003302	0.000280	0.808872	1.000000	0.975824	0.482776	0.999884
AMT_CREDIT	0.003659	0.000195	0.816429	0.975824	1.000000	0.301284	0.993087

5 rows × 21 columns

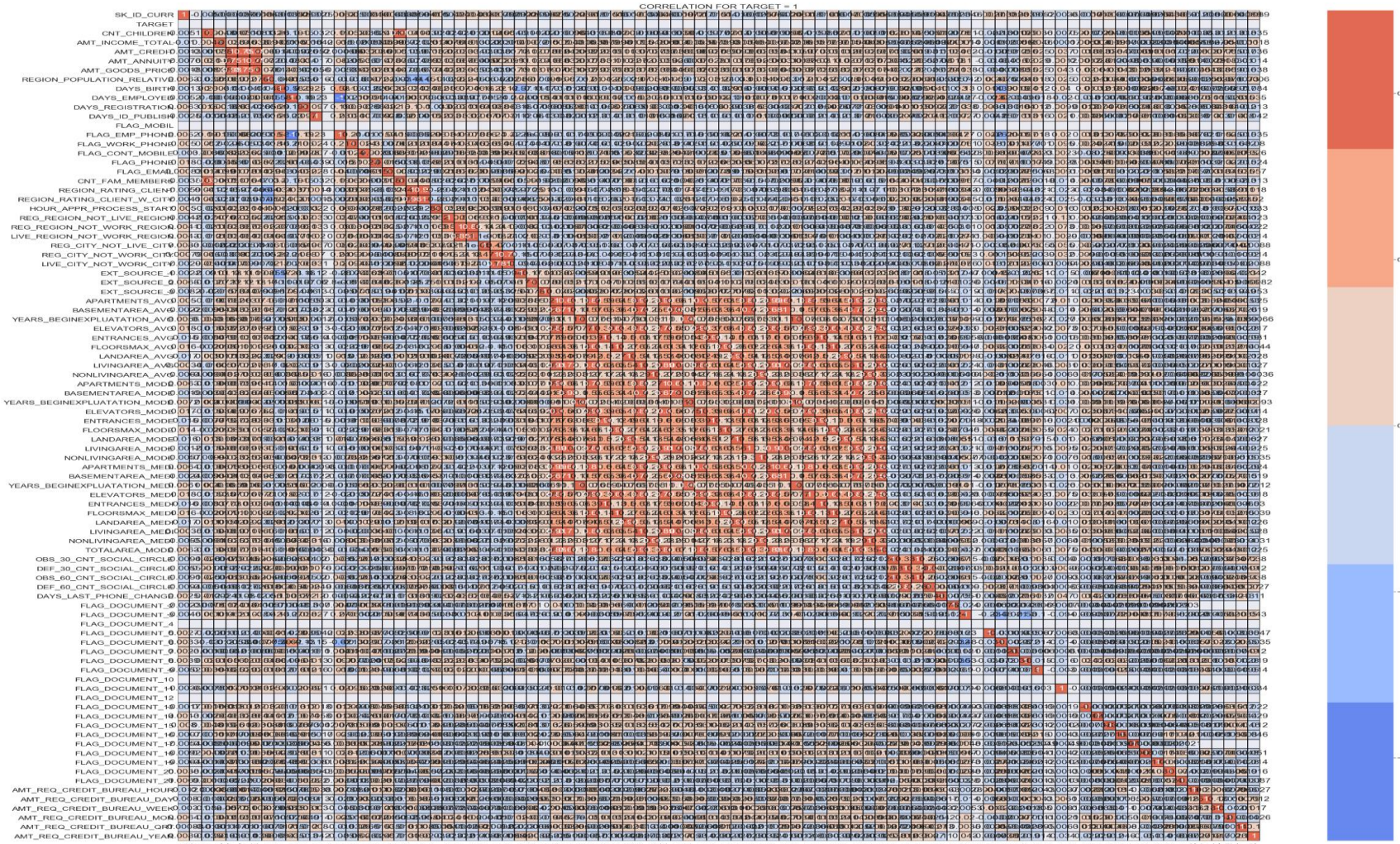
HEAT MAP FOR CORRELATION DATA ON NEW APPLICATION

Categorizing data by "TARGET" variable

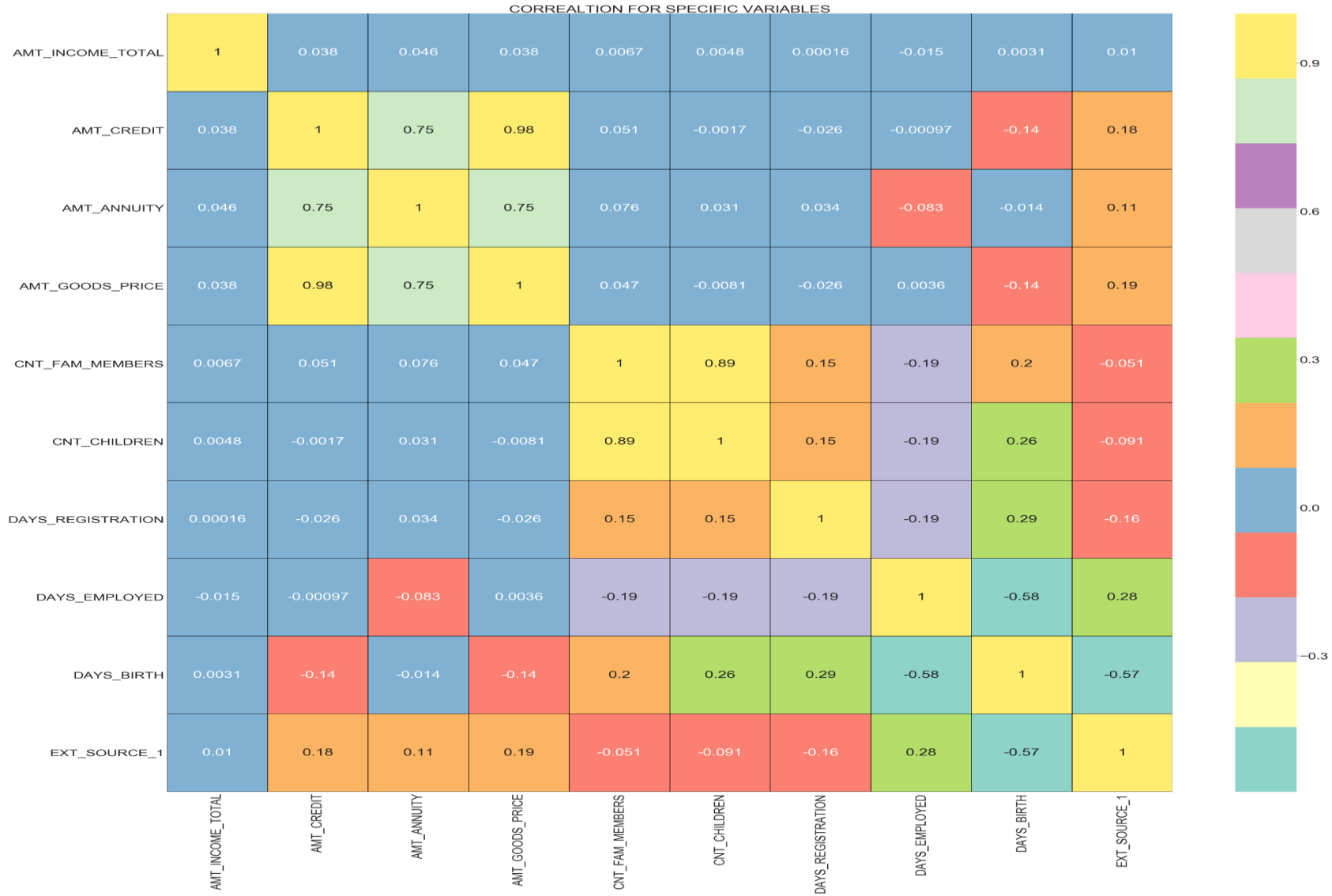
HEAT MAP FOR CORRELATION DATA ON TARGET=0



HEAT MAP FOR CORRELATION DATA ON TARGET=1



CORREALTION FOR SPECIFIC VARIABLES



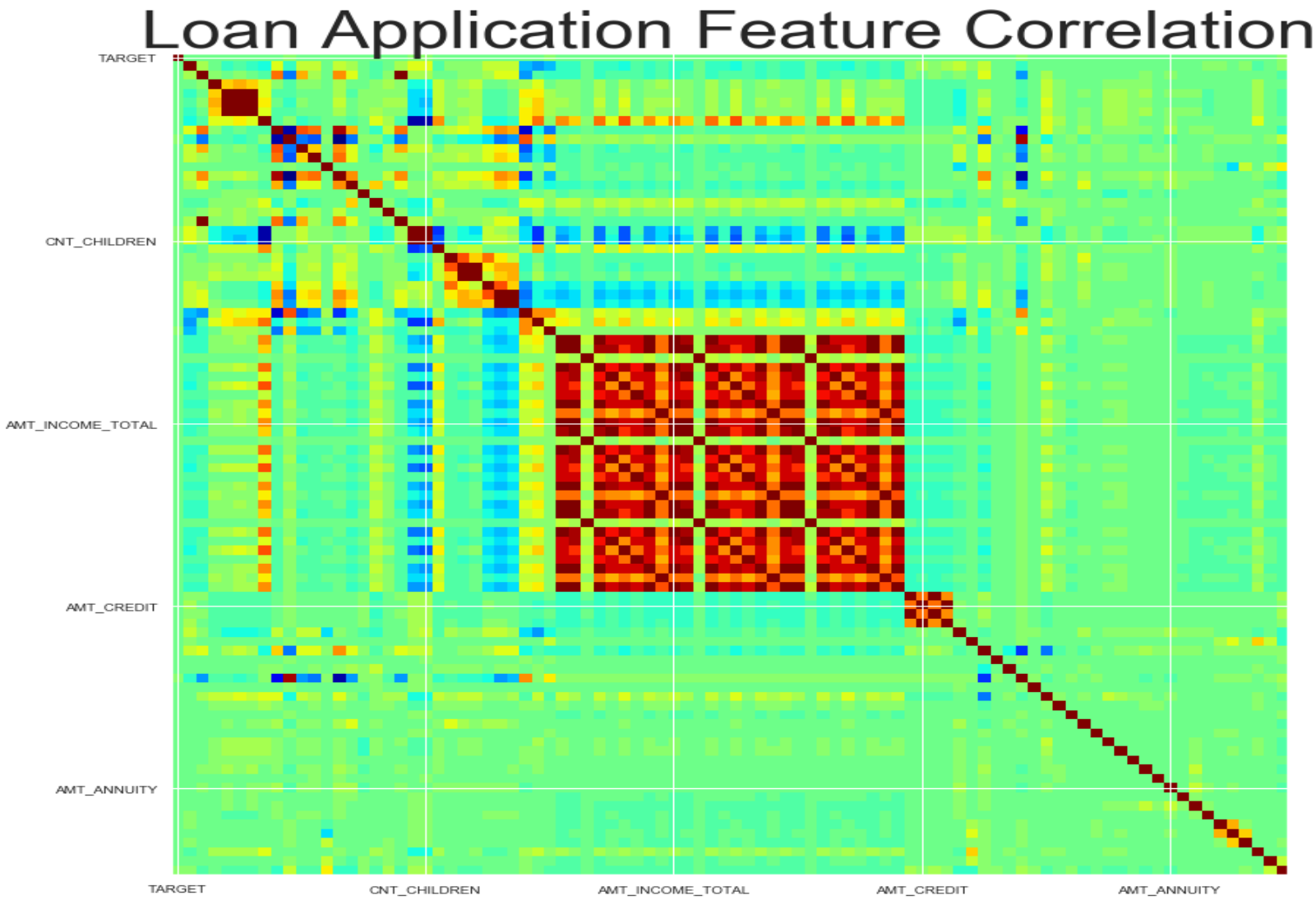
Highly Correlated Variables

TOP 4 CORRELATED VARIABLES	
AMT_CREDIT	AMT_ANNUITY
AMT_ANNUITY	AMT_GOODS_PRICE
AMT_GOODS_PRICE	AMT_CREDIT
CNT_FAM_MEMBERS	CNT_CHILDREN

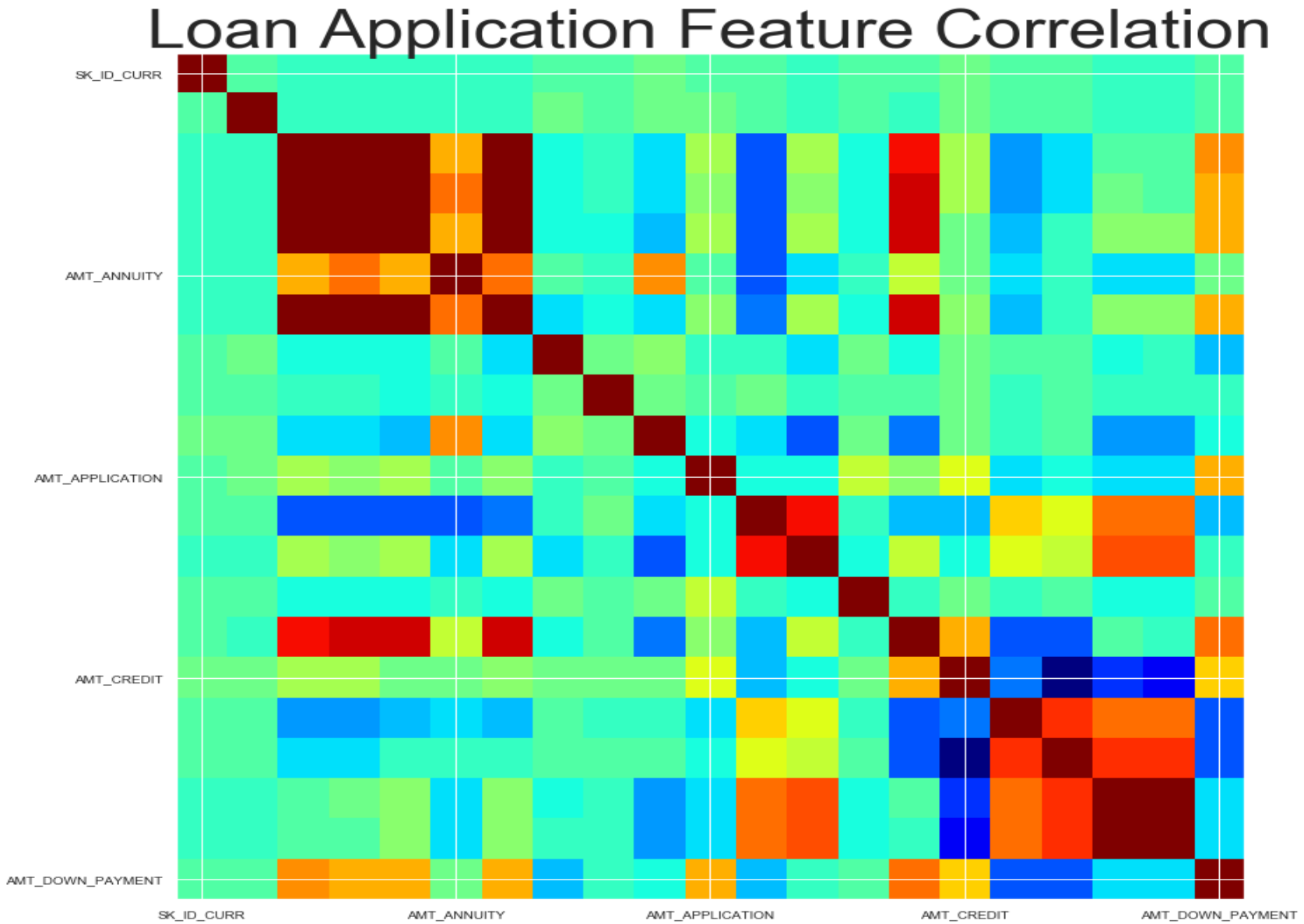
 # TOP 7 Correlated Variables:

1. AMT_CREDIT
2. AMT_ANNUITY
3. AMT_GOODS_PRICE
4. CNT_FAM_MEMBERS
5. DAYS_EMPLOYED
6. DAYS_BIRTH
7. EXT_SOURCE_1

Plot the Heat Map For New Loan Application



Plot the correlation pattern with New Loan Application



Extracting Highly correlated variables from New Loan Applications

	FLAG_MOBIL	DAYS_ID_PUBLISH	DAYS_BIRTH	CNT_CHILDREN	CNT_FAM_MEMBERS	AMT_GOODS_PRICE	AMT_CREDIT	AMT_ANNUITY	A
count	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	14.000000	
mean	0.029782	0.129795	0.186483	0.188221	0.198020	0.178859	0.181524	0.190996	
std	0.287175	0.301759	0.344280	0.376274	0.367062	0.457102	0.455292	0.448615	
min	-0.275207	-0.055193	-0.103767	-0.030668	-0.034020	-0.227535	-0.222726	-0.218532	
25%	-0.034794	-0.033180	-0.019817	-0.021552	-0.014191	-0.098394	-0.094630	-0.102904	
50%	-0.023455	-0.007490	-0.011334	0.012561	0.037756	-0.045742	-0.042903	-0.013116	
75%	-0.016453	0.124909	0.447757	0.100159	0.106021	0.277334	0.277246	0.322332	
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

Plotting the correlation pattern with Specific variables of Previous Application