Credit EDA Case Study

Anik Chakraborty
Hareesh Adhimoolam

Business 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 utilise this knowledge for its portfolio and risk assessment.

Taken Approach

- In Application dataset there are 307511 rows and 122 columns.
- Columns that have missing values more than 50% are dopped as columns with that much of missing values may not be helpful to depict the insight. After dropping such columns, there are 81 columns in the dataset.
- It can be seen for days columns; the values are in negative so we those were converted into positive values.
- Few columns denotes Number of enquires, count of family members, no. of social media connects and no. of days. So, these columns can not be in float, so these columns were also converted in integers.
- There are few columns having integer types but contains categorical data. Converting these columns into data type object is required to perform univariate and bivariate analysis properly. We checked columns having less unique values (most of those are Flag columns) and converted those columns into object type.
- Columns having less than 14% Missing Values were analyzed and suitable imputation methods have been mentioned in next slide.
- Application data has been divided into two datasets one where TARGET=1 (client with payment difficulties) and one where TARGET=0 (Other clients)
- Then outlier detection, univariate analysis, bivariate analysis, correlation analysis are performed
- For Previous Application data same analysis has been performed.
- At the end Conclusion, Important insights are provided.

Application Data

Suggesting Missing Value Treatment

AMT_ANNUITY has continuous value and it's distribution is not symmetrical, we can use Median imputation to treat missing values in this column.

AMT_GOODS_PRICE has continuous value and it's distribution is not symmetrical, we can use Median imputation to treat missing values in this column.

NAME_TYPE_SUITE is a categorical column; we can use mode to impute the missing values.

CNT_FAM_MEMBERS and NAME_FAMILY_STATUS: Both of these Columns have 2 missing values, In NAME_FAMILY_STATUS column the missing value is mentioned as 'Unknow' and in CNT_FAM_MEMBERS columns these are NA values. We can replace these with mode of those two columns.

EXT_SOURCE_2 : it's a continuous variable, we can replace missing values using Median imputation.

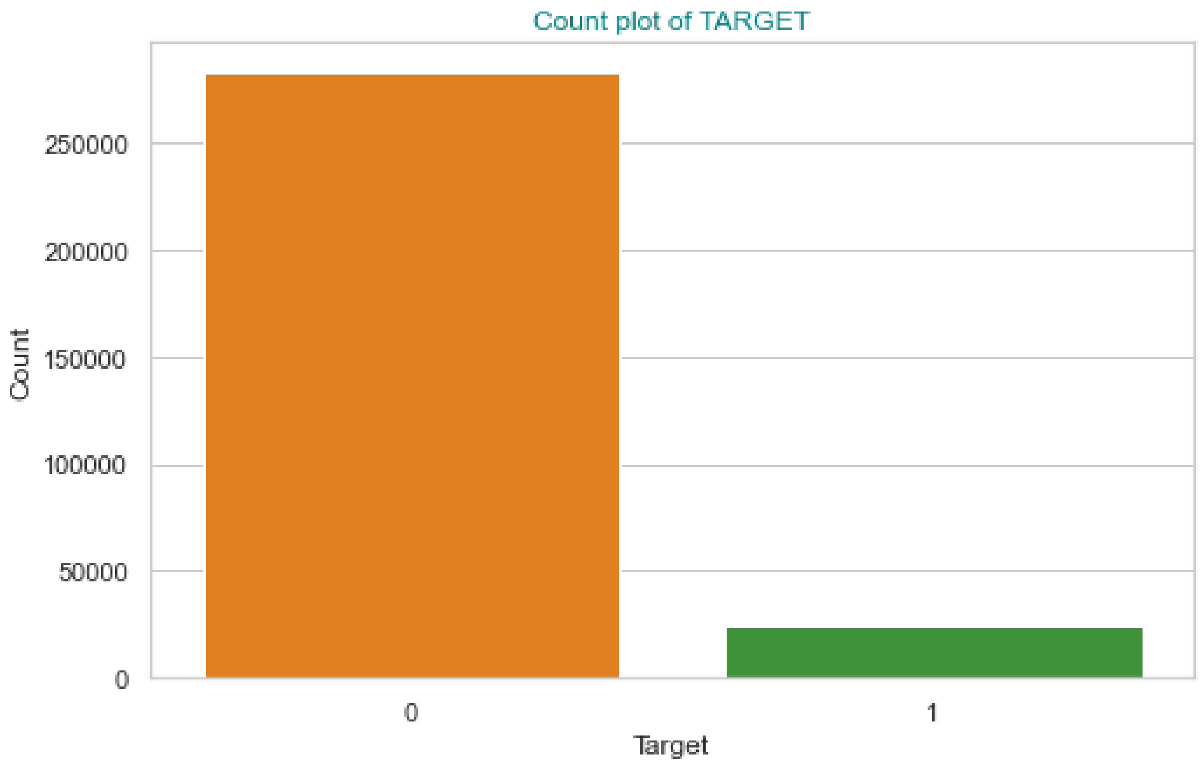
*_CNT_SOCIAL_CIRCLE: Mode imputation can be used for these columns.

DAYS_LAST_PHONE_CHANGE: There is a single missing value, we can use median imputation here.

AMT_REQ_CREDIT_* columns denote Number of enquiries to Credit Bureau about the client # Months/Days/Weeks before application. We can use mode imputation. That will generalize it, so we are basically replacing the missing values with number of enquires generally bank does for most of the customers.

Checking Class Imbalance for TARGET

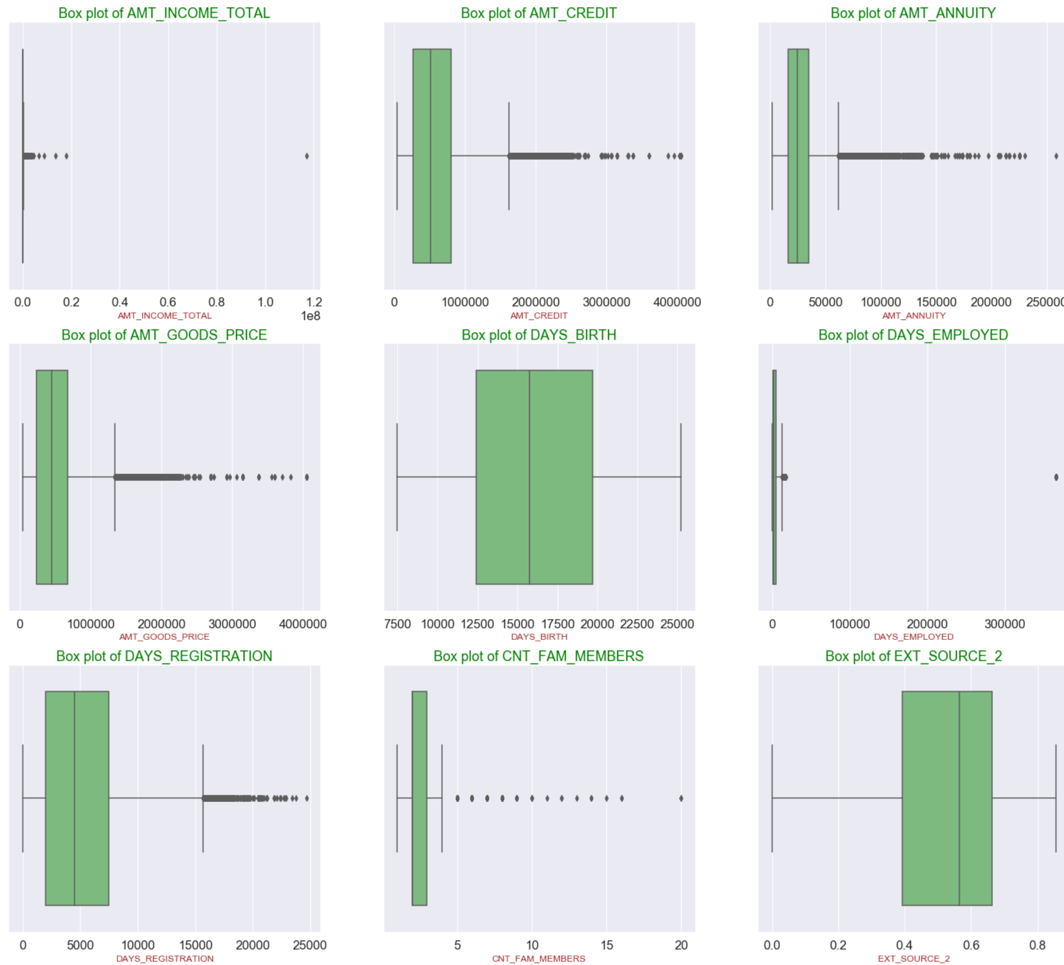
There is a huge class imbalance in the dataset. Almost 91.9% data is of Target 0 and only 8.1% data is of Target 1.




```
0 91.927118
1 8.072882
Name: TARGET, dtype: float64
```

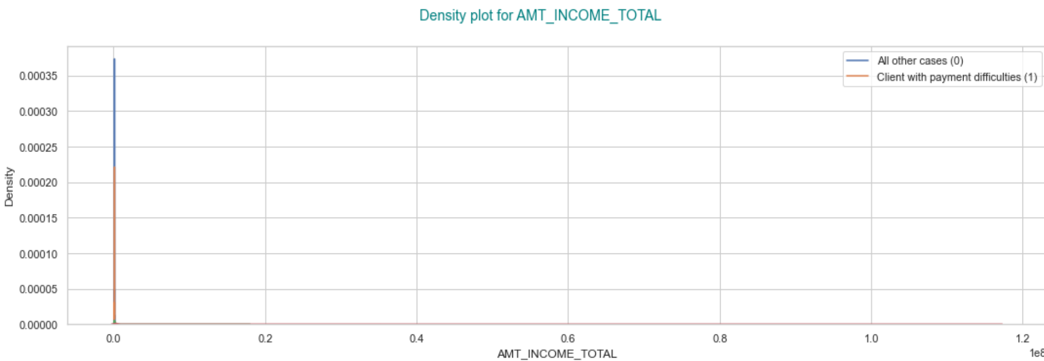
Univariate Analysis of Numeric Variables

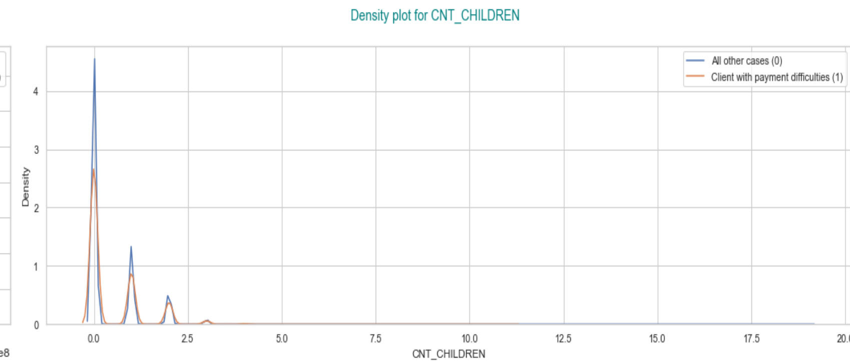
Analyzing and Detecting Outliers

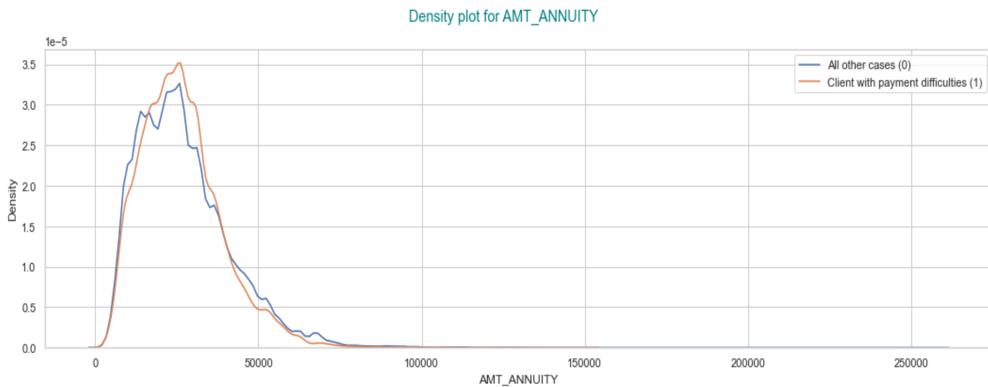


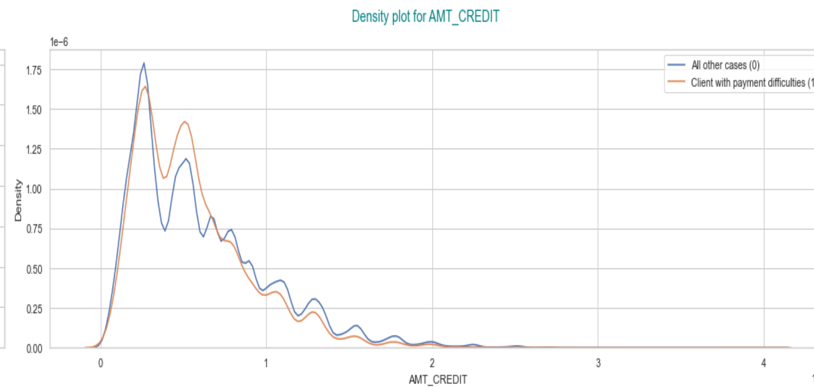
- AMT_INCOME_TOTAL (Income of the client) column we can see income of a client is 117M, that is an outlier.
- For AMT_CREDIT columns: We can see very few Credit amount of the loan are above 3 Million.
- AMT_ANNUITY: For Loan Annuity amount we can see value above 250k, that can be treated as outlier.
- AMT_GOODS_PRICE: For consumer loans it is the price of the goods for which the loan is given. There is an outlier above 4 Million. Infect there are very few observations between 3.5 and 4 Million.
- DAYS_BIRTH: Client's age in days at the time of application. There is no outlier. Most of the clients have age between (12500/365) ~ 34 Years to (20000/365) ~ 55 Years.
- DAYS_EMPLOYED: How many days before the application the person started current employment. We can see there are 55374 observations having DAYS_EMPLOYED= 365243 Days that is 1000 years. It's surely some garbage data and should be treated as missing value. If we ignore this value (365243) and again boxplot for DAYS_EMPLOYED, we can see some outlier above value of 17500 days.
- DAYS_REGISTRATION: How many days before the application did client change his registration. We can see few values between 24k and 25k. These should be treated as outliers.
- CNT_FAM_MEMBERS: How many family members does client have. We can see outliers at value 20. We can also see there are very few observations above 15.
- EXT_SOURCE_2: Normalized score from external data source. We can not see any outliers. Most of the scores are between .4 and .65

Univariate Analysis of Numerical Variables



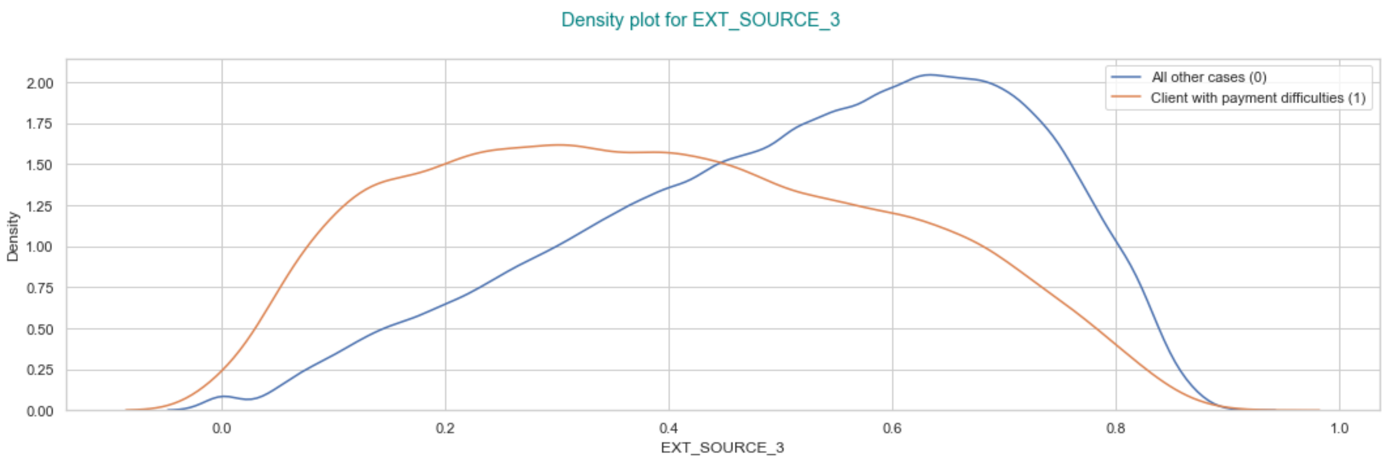


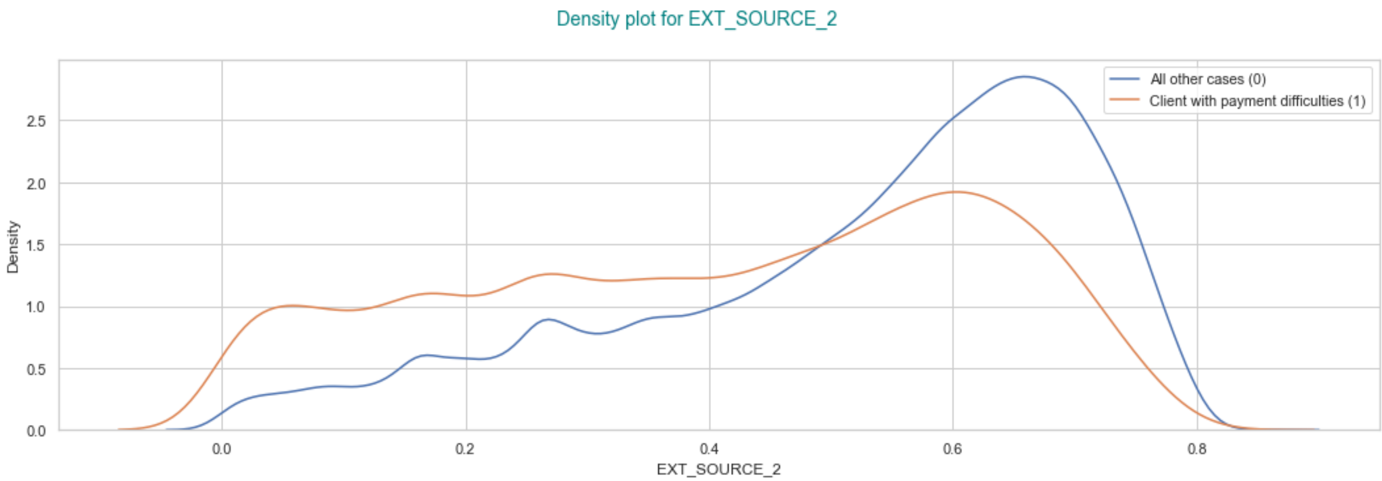




- As AMT_INCOME_TOTAL is very distributed plotting it directly does not give good insight. Previously we have created bins based on income ['Very Low', 'Low', 'Lower-Medium', 'Higher-Medium', 'High', 'Very High']
- Here in the first plot, we are comparing distribution of AMT_INCOME_TOTAL for LOWER-MEDIUM income group.
- There are not much difference in distribution of these numeric variable for TARGET=0 and TARGET=1 observations.

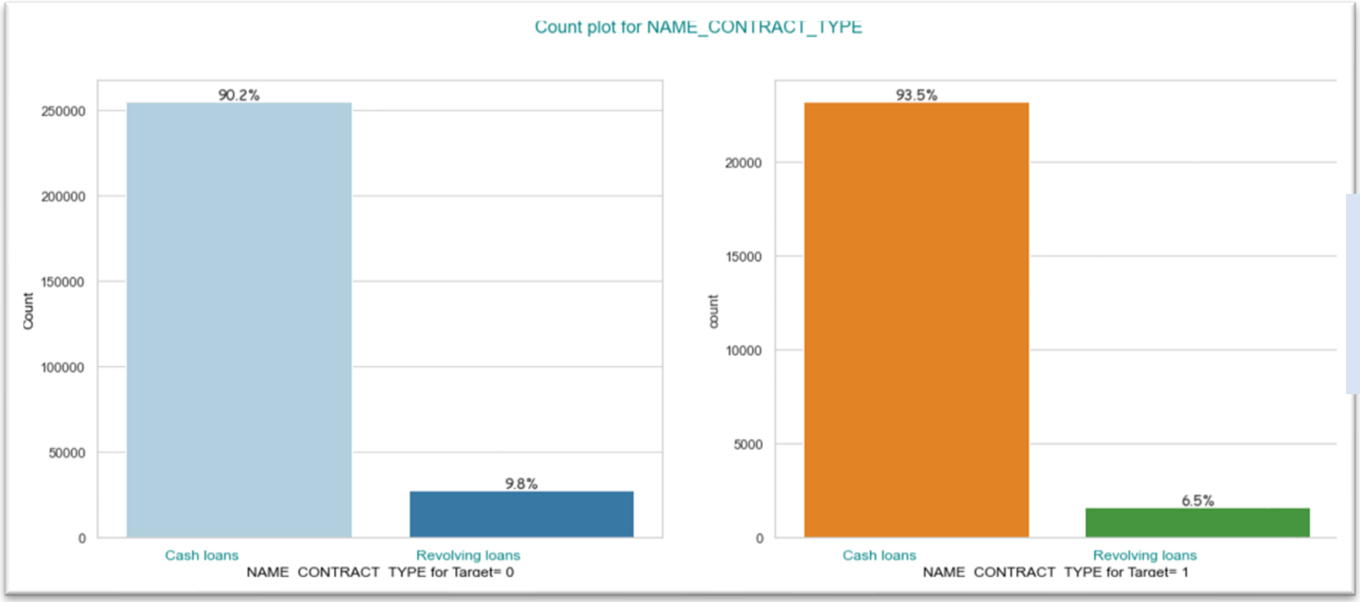
Univariate Analysis of Numerical Variables



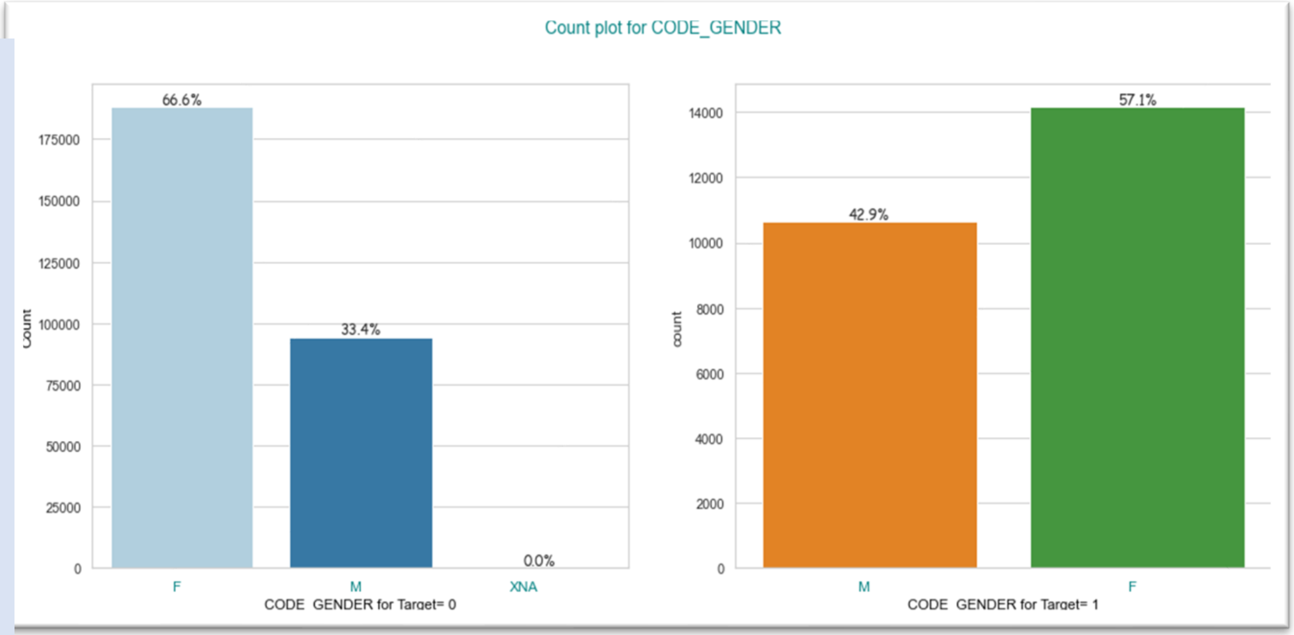


- EXT_SOURCE_2 and EXT_SOURCE_3 denote Normalized score from external data source. In both the cases specially for EXT_SOURCE_3 we can see clients who are facing difficulties in loan repayment have lesser mode value that other group. Bank should give more importance to EXT_SOURCE_2 and EXT_SOURCE_3 scores specially on EXT_SOURCE_3 score before approving loan application, if the scores are available.

Univariate Analysis of Numerical Variables

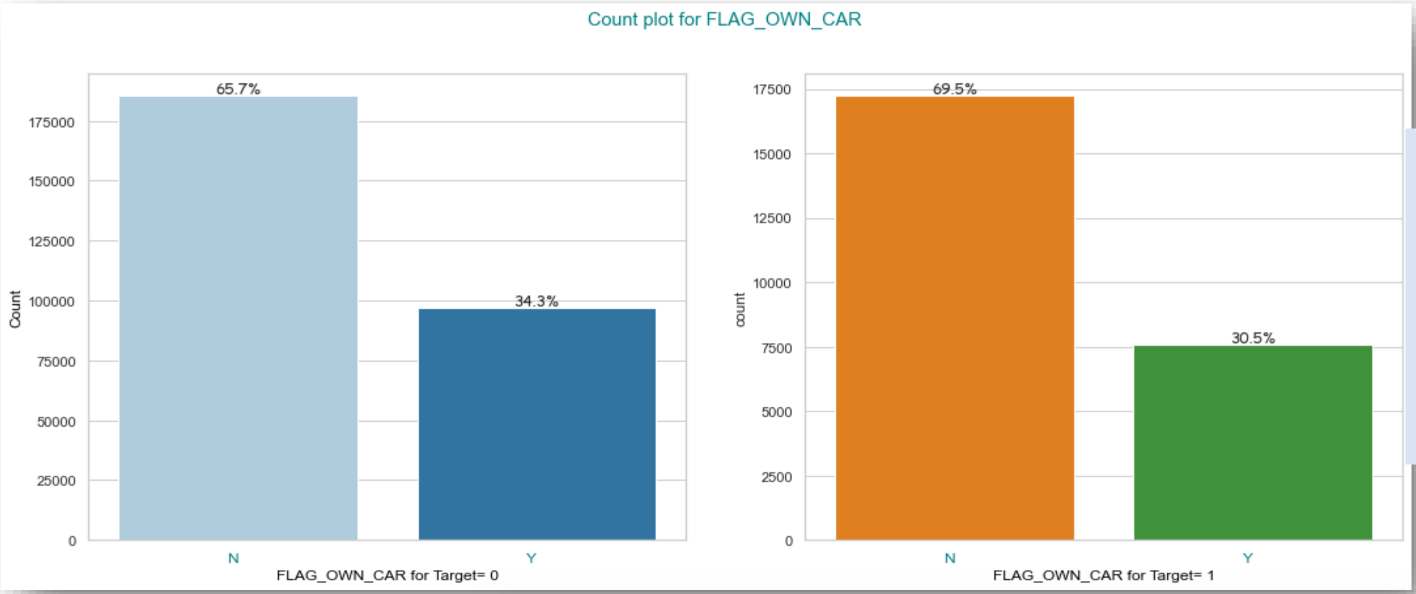


- Number of Cash loan is higher than revolving loan, in both the cases (For Target=0 and 1)

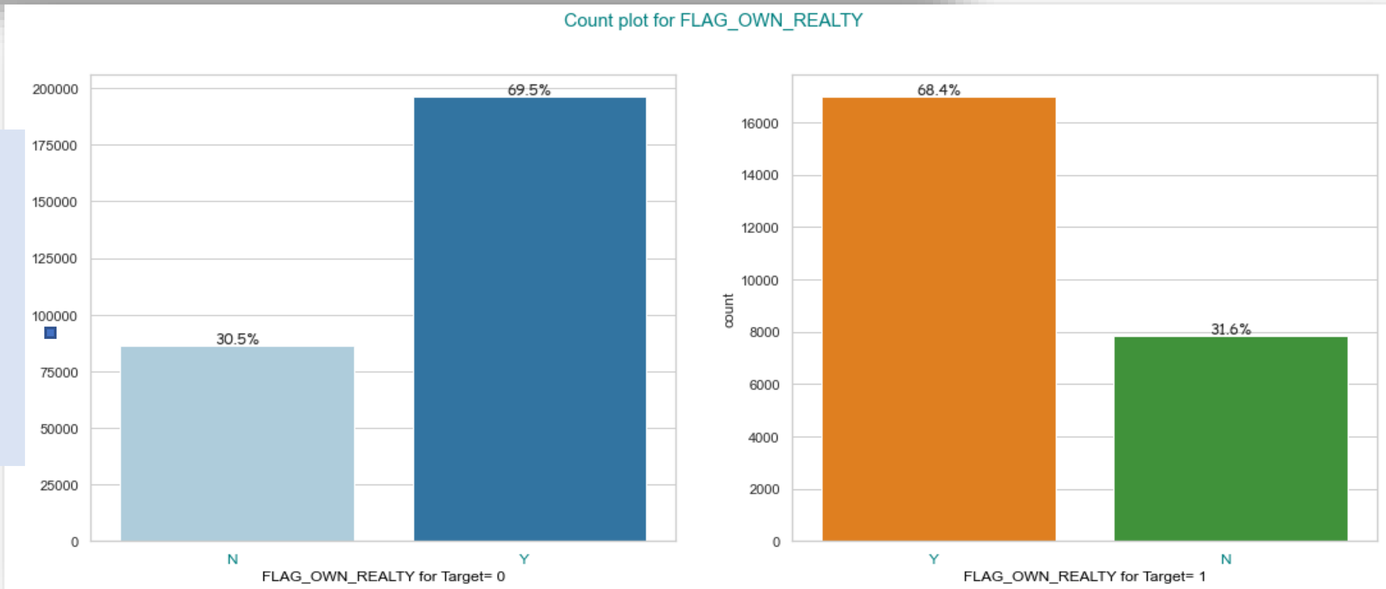


- From the visual we can see number of Female clients are much higher than the number of Male clients. Again, if we compare the proportion, Male to Female proportion is almost 1:2 for Target=0. But for Target=1 Male client to Female client proportion is higher in compare to Target 0. So Female clients are less likely to face payment difficulties than Male clients.

Univariate Analysis of Categorical Variables



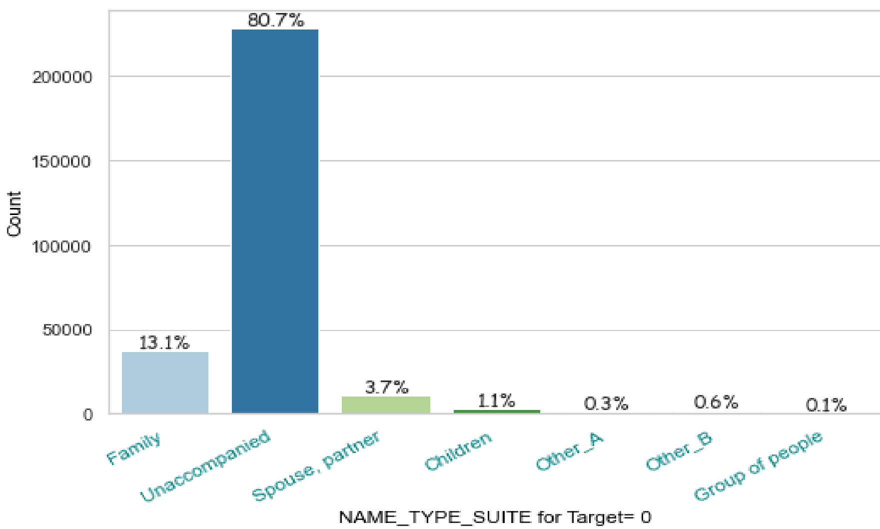
- If we compare the proportions in both the graphs, there is not much of a difference. Clients having car/cars are slightly better in repaying the loans.

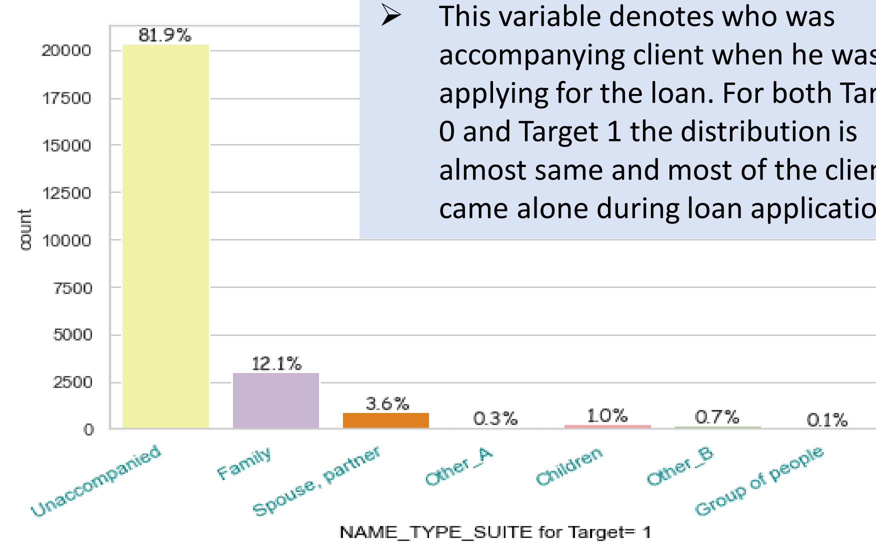


- This flag denotes if the client owns a house or flat. Ratio of 'Y' and 'N' flags are almost same in both the groups. Most of the clients own house or flat.

Univariate Analysis of Categorical Variables

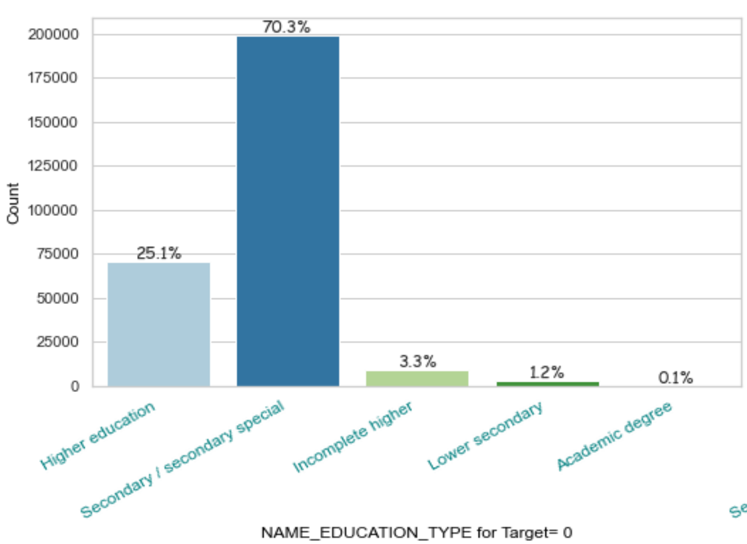
Count plot for NAME_TYPE_SUITE

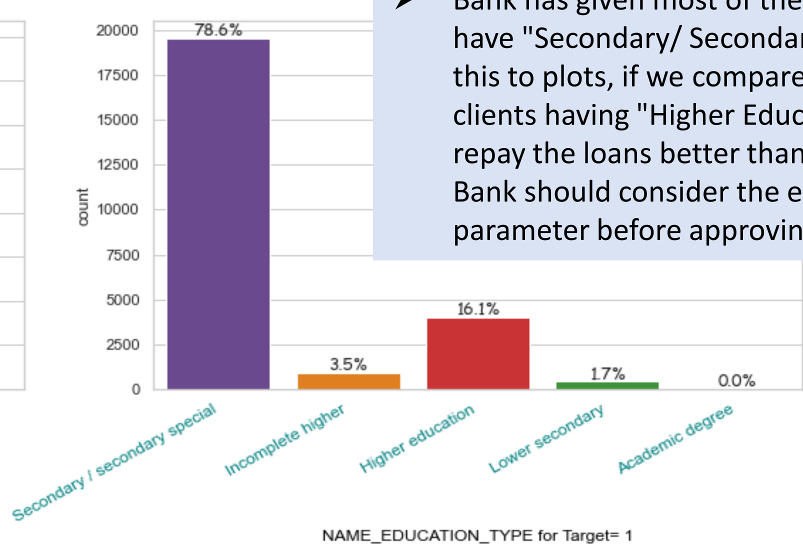




- ➢ This variable denotes who was accompanying client when he was applying for the loan. For both Target 0 and Target 1 the distribution is almost same and most of the client came alone during loan application.

Count plot for NAME_EDUCATION_TYPE

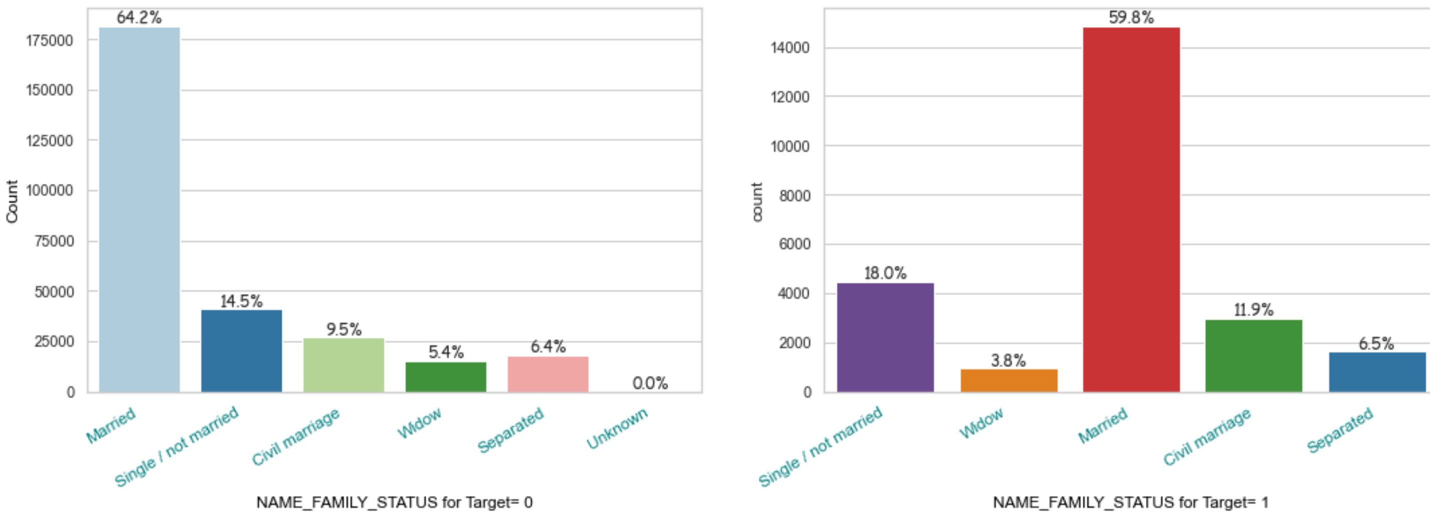




- ➢ Bank has given most of the loans to the people who have "Secondary/ Secondary Special" education. In this to plots, if we compare, we can see only the clients having "Higher Education" are more likely to repay the loans better than other education groups. Bank should consider the education as an important parameter before approving the loan.

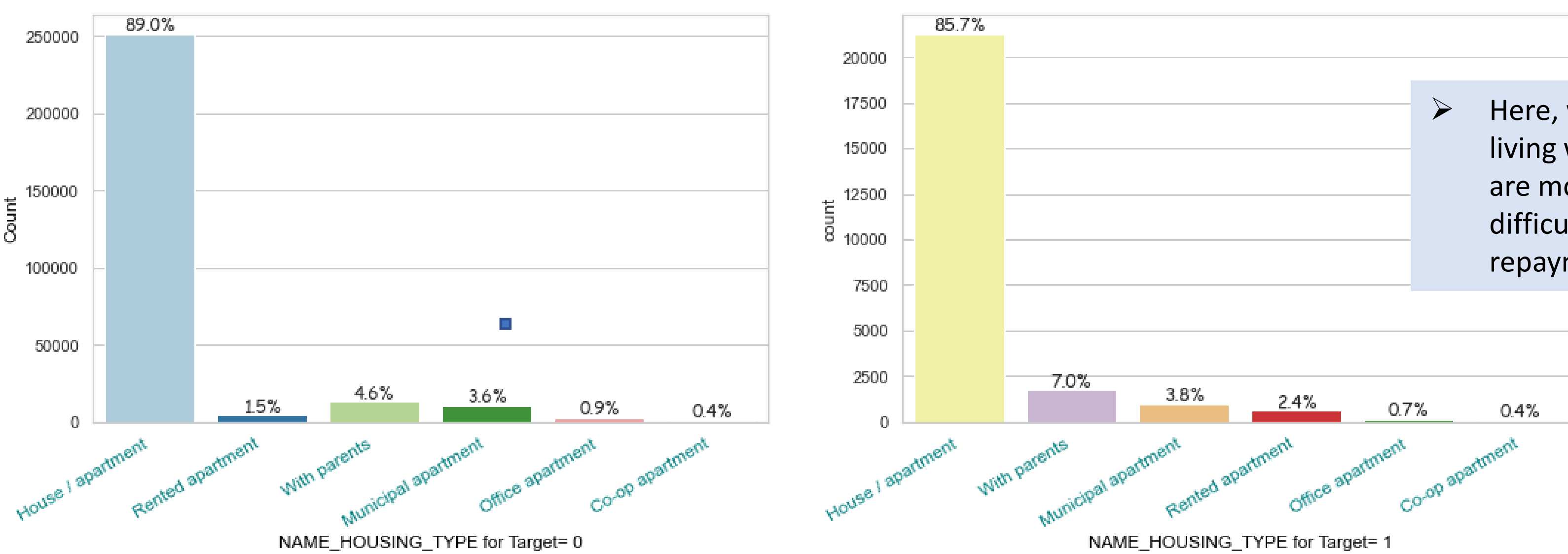
Univariate Analysis of Categorical Variables

Count plot for NAME_FAMILY_STATUS



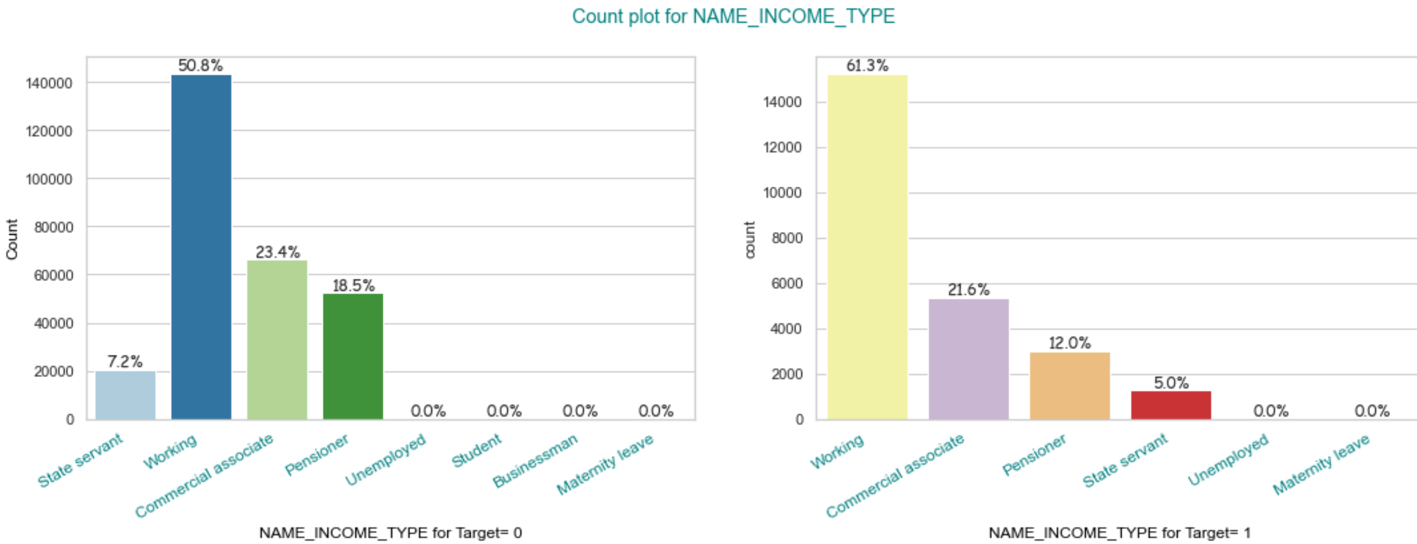
- Most of the clients are "Married" and they are less likely to face difficulties in loan repayment. Where Single, Window and other groups are more likely to default on the loan. The reason could be Married people may do more financial planning in compare to single peoples.

Count plot for NAME_HOUSING_TYPE

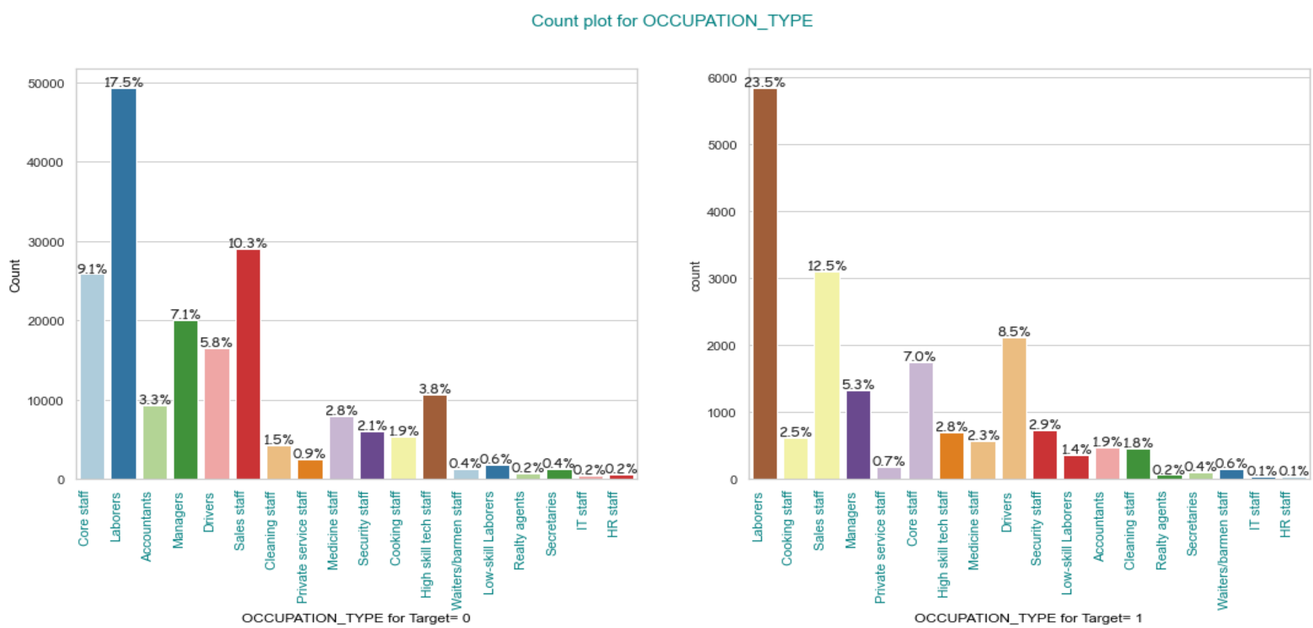


- Here, we can the clients living with their parents are more likely to face difficulties in loan repayment.

Univariate Analysis of Categorical Variables

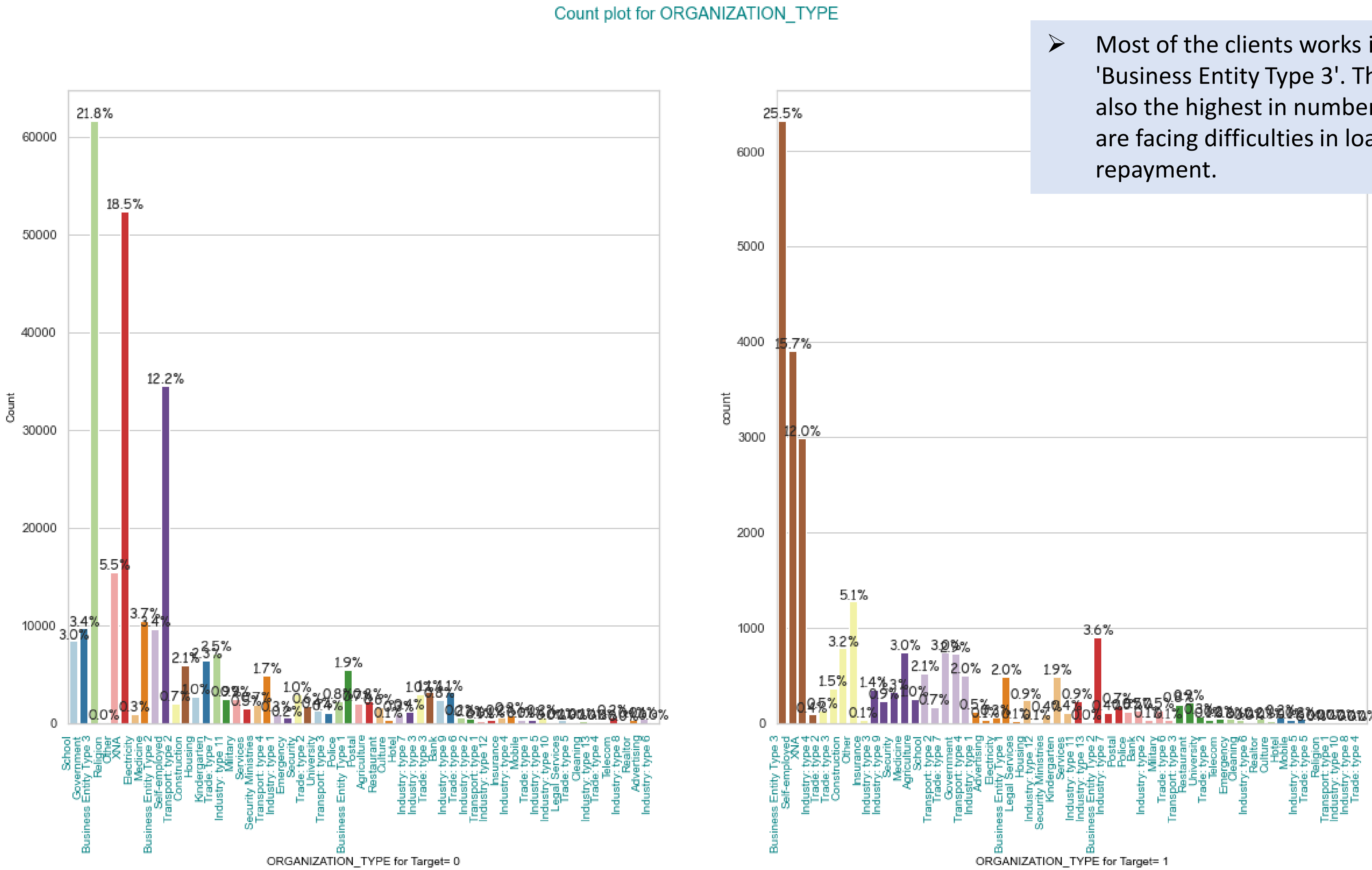


- ➢ Most of the clients taken loan are of working class. Also, they have bit higher chance of failing to repay the loan. As 61.3% of clients with payment difficulties are of working class, where it's bit lesser, 50.8% for Target 0.
- ➢ Pensioners are more likely to repay the loan in compare to other income classes. The reason could be, the pensioners have a fixed stable income, and they may take loan of a calculated amount, so that the monthly installment can be covered using the pension income.



- ➢ Most of the clients are Laborers and they are facing more difficulties in repaying the loan in compare to Managers, Accountants, Core Staff, High Skill tech Staff etc. higher paying jobs.

Univariate Analysis of Categorical Variables



- Most of the clients works in 'Business Entity Type 3'. They are also the highest in number who are facing difficulties in loan repayment.

Correlation Analysis of Numeric Variables

➢ Top 10 correlated variable for client with payment difficulties (1) and other (0) both are same. Pearson correlation coefficient values of those variables are also almost same.

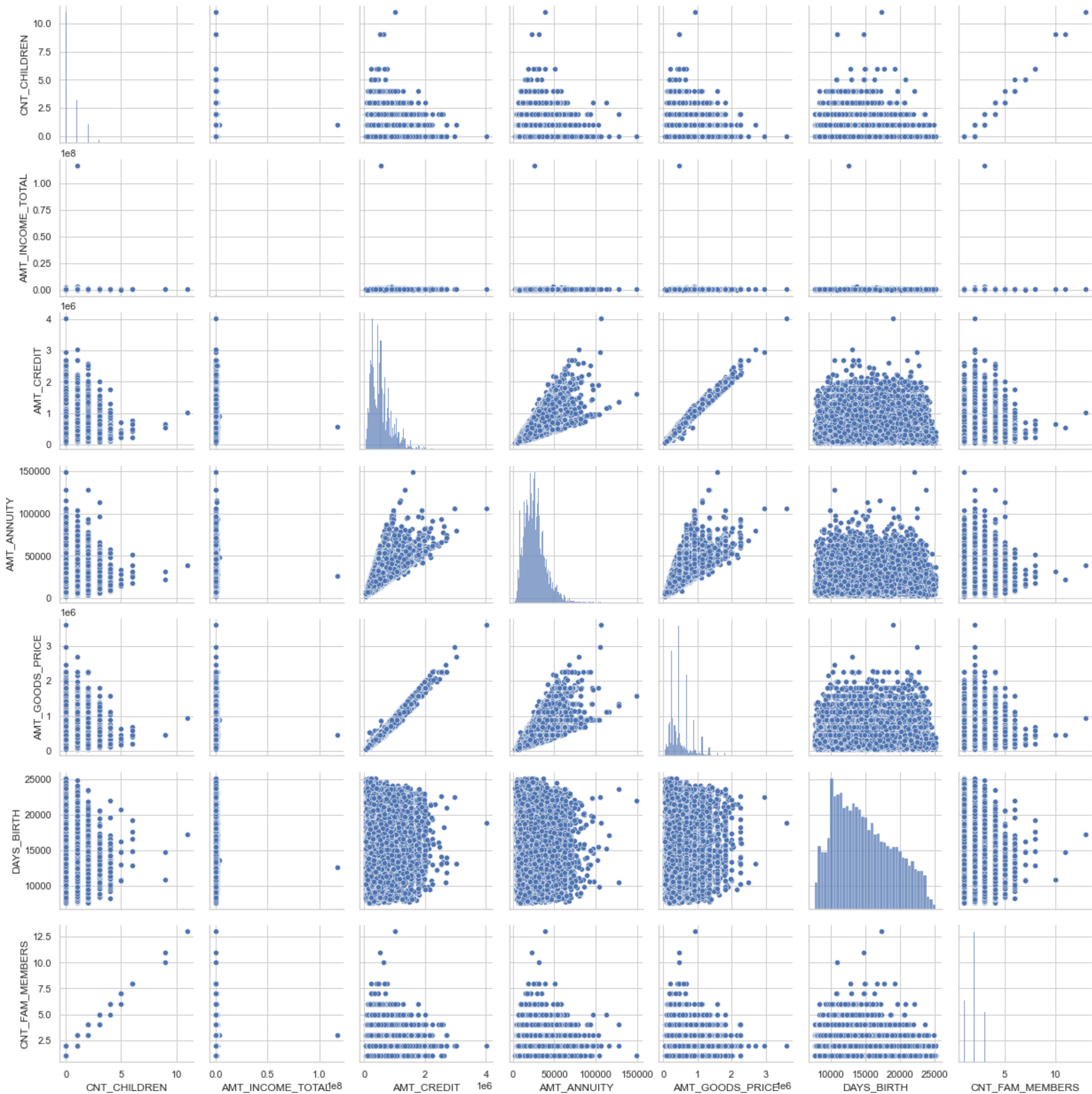
TARGET=0

Var 1	Var 2	correlation	abs_correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508	0.998508
FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997018	0.997018
YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.993582	0.993582
FLOORSMAX_MEDI	FLOORSMAX_MODE	0.988153	0.988153
AMT_GOODS_PRICE	AMT_CREDIT	0.987250	0.987250
FLOORSMAX_MODE	FLOORSMAX_AVG	0.985603	0.985603
YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.971032	0.971032
YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.962064	0.962064
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571	0.878571
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859332	0.859332

TARGET=1

Var 1	Var 2	correlation	abs_correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269	0.998269
FLOORSMAX_MEDI	FLOORSMAX_AVG	0.997187	0.997187
YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_AVG	0.996124	0.996124
FLOORSMAX_MEDI	FLOORSMAX_MODE	0.989195	0.989195
FLOORSMAX_MODE	FLOORSMAX_AVG	0.986594	0.986594
AMT_GOODS_PRICE	AMT_CREDIT	0.983103	0.983103
YEARS_BEGINEXPLUATATION_MODE	YEARS_BEGINEXPLUATATION_AVG	0.980466	0.980466
YEARS_BEGINEXPLUATATION_MEDI	YEARS_BEGINEXPLUATATION_MODE	0.978073	0.978073
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484	0.885484
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994	0.868994

Bivariate Analysis of Numerical - Numerical Variables



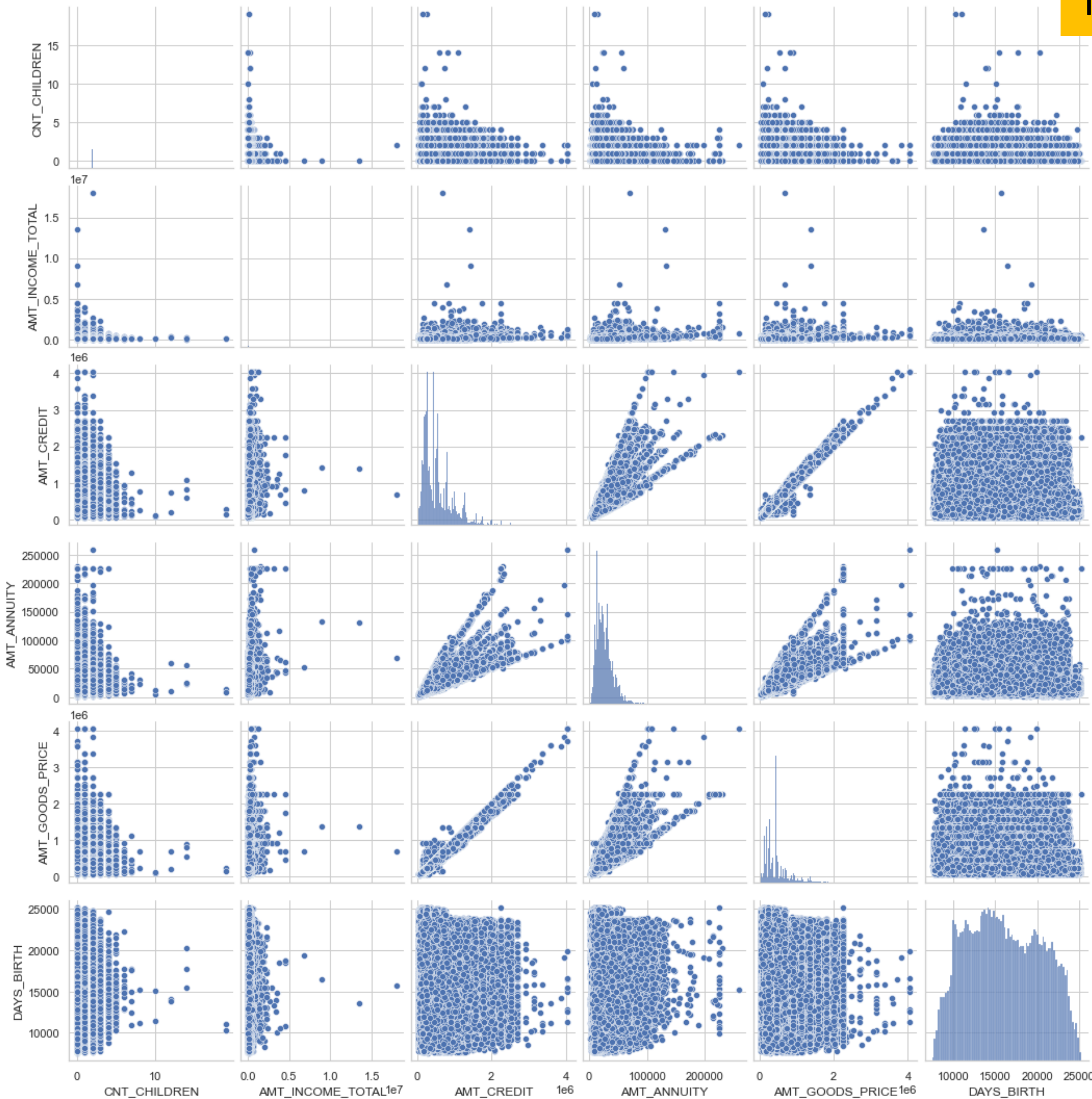
TARGET=1

Bivariate analysis of Numerical variables for clients with payment difficulties (TARGET=1).

Combined observations are mentioned in next slide.

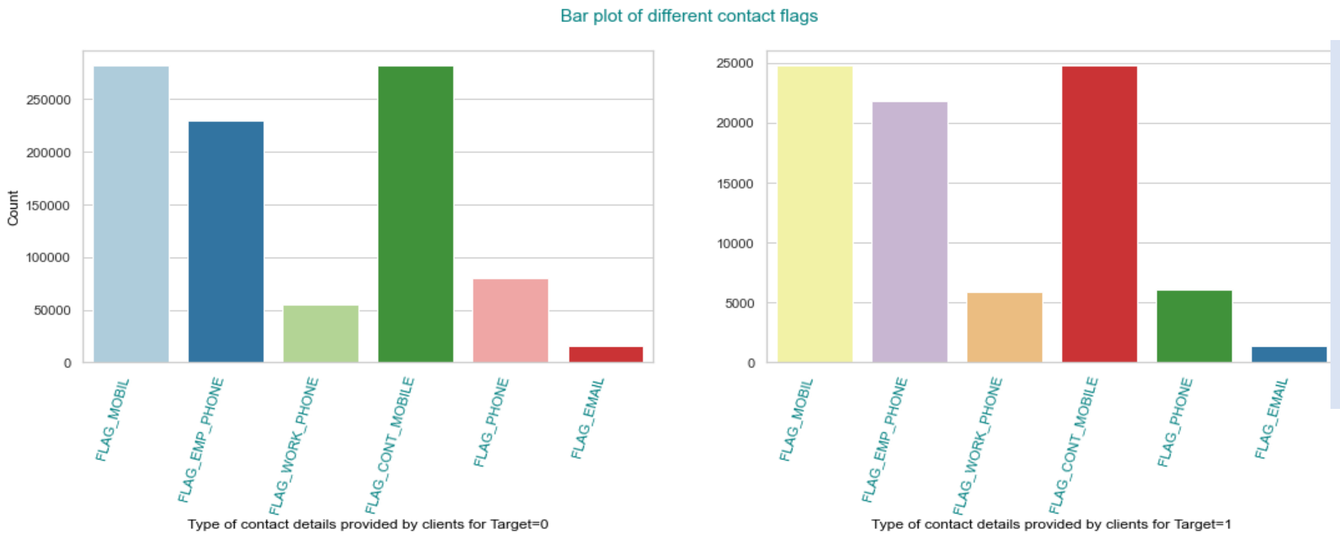
Bivariate Analysis of Numerical - Numerical Variables

TARGET=0

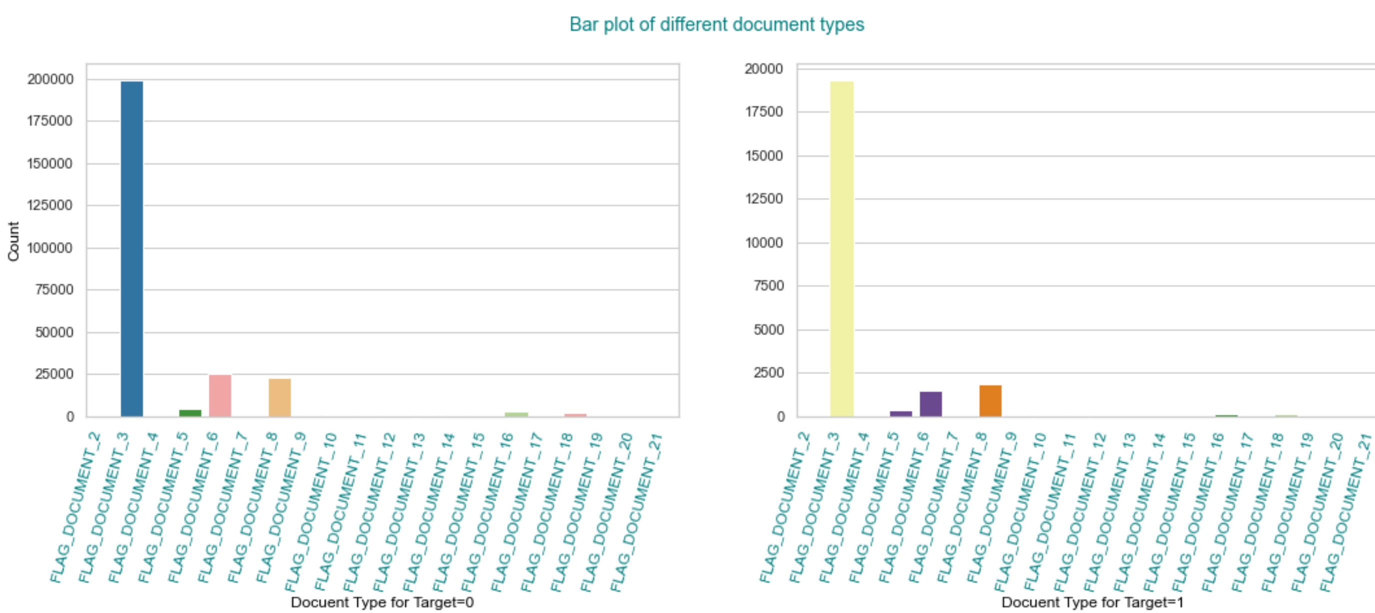


- There is a strong positive linear relation between AMT_GOODS_PRICE and AMT_CREDIT.
- There is a strong positive linear relation between CNT_CHILDREN and CNT_FAM_MEMBER.
- There is a weak positive linear relation between AMT_ANNUITY and AMT_GOOD_PRICE.
- There is a weak positive linear relation between AMT_CREDIT and AMT_ANNUITY.
- Outliers are already identified using Boxplot earlier.

Bivariate Analysis of Categorical - Categorical Variables



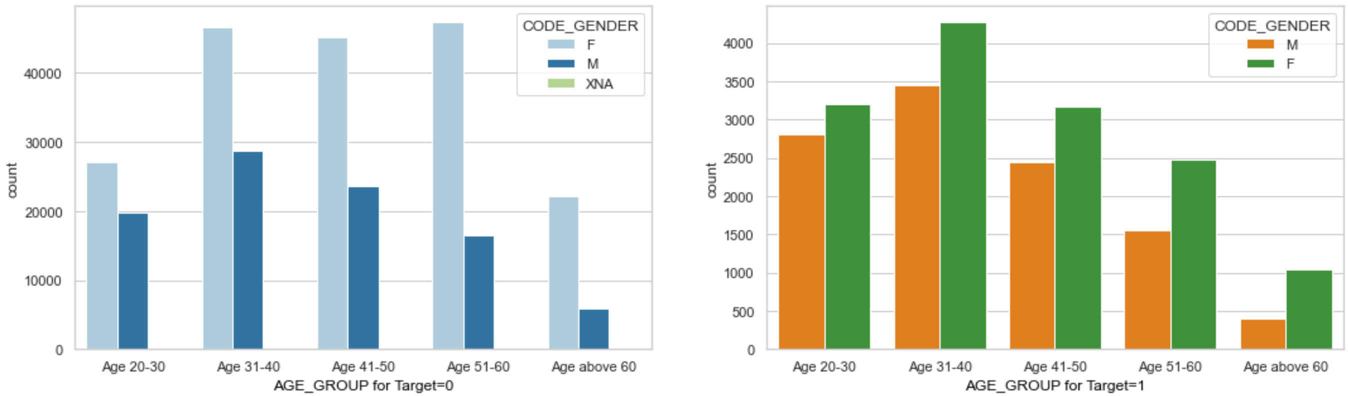
- Most of the clients have provided their mobile number and value of FLAG_CONT_MOBILE and FLAG_MOBIL both are almost same, that means the clients were reachable on provided mobile number. So, there is no significant case of fraud.



- Most of the clients have submitted Document 3 during loan application. probably it's a mandatory document for a loan application. Clients who have also submitted Document 6 during loan application have lesser chance to face difficulties in paying loan installments. Bank needs to check Document 6 and making it a mandatory document may help to reduce defaults on loans.

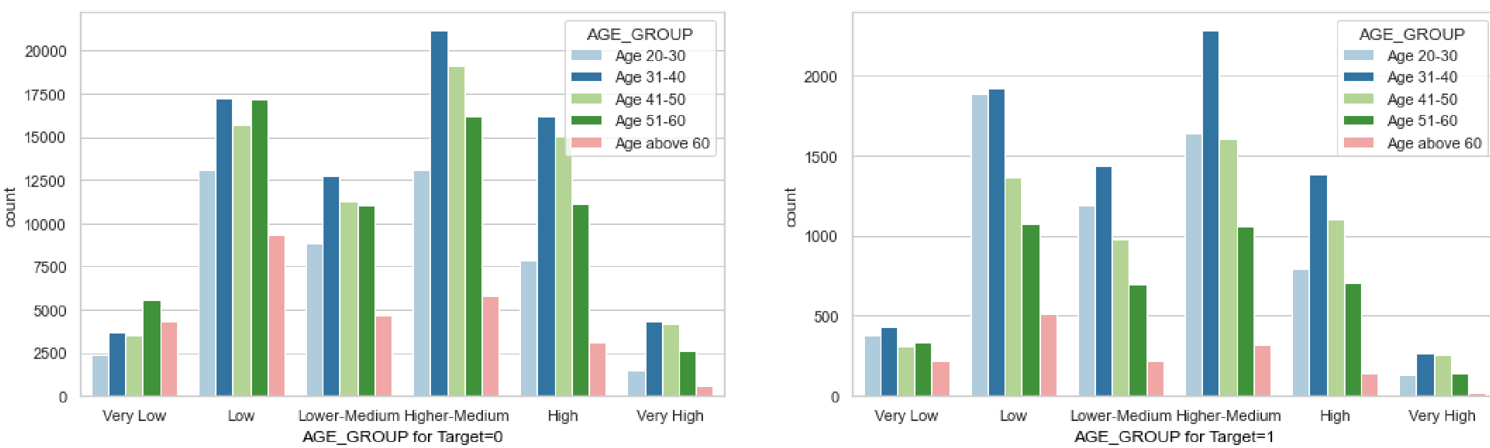
Bivariate Analysis of Categorical - Categorical Variables

AGE_GROUP vs CODE_GENDER



- It can be observed that Age group 20-30 and 31-40 face more difficulties in loan repayments in compare to other age groups. With increased age people more likely to face less difficulty in paying the loan. Here we can see people above age of 60 has the least chance of facing difficulties in paying loan. Previously we saw the same kind of insights for Pensioners. 'Above 60 years age' and 'Pensioners' both points to same group of peoples.

AGE_GROUP vs INCOME_SLAB

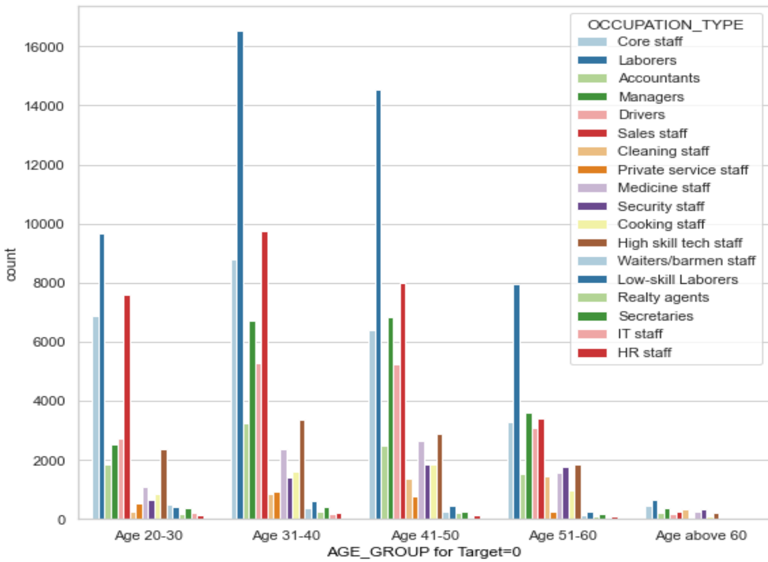


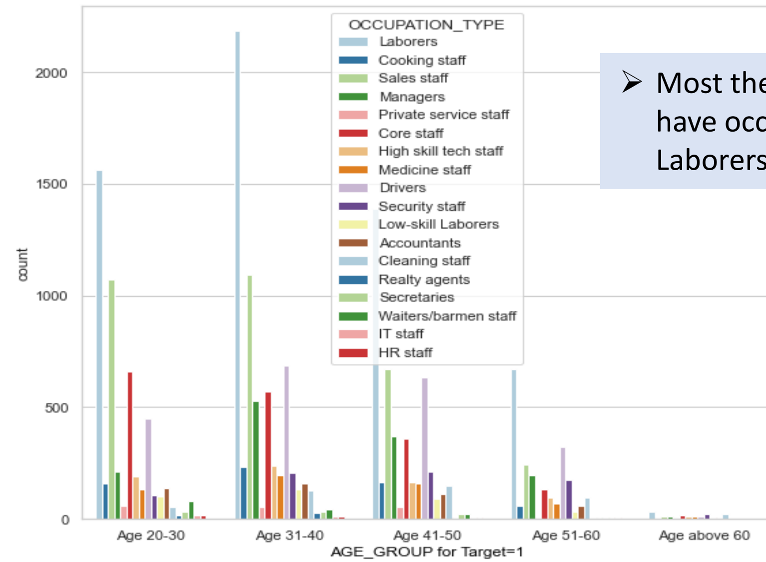
- Low and Higher-Medium Income Slab people have highest proportion of facing difficulties in paying the loan. Again, in these two sections people having Age group 20-30 and 31-40 are the highest in numbers.

So, peoples of age group 20-40 of Low ad Higher-Medium income groups are most likely face payment difficulties.

Bivariate Analysis of Categorical - Categorical Variables

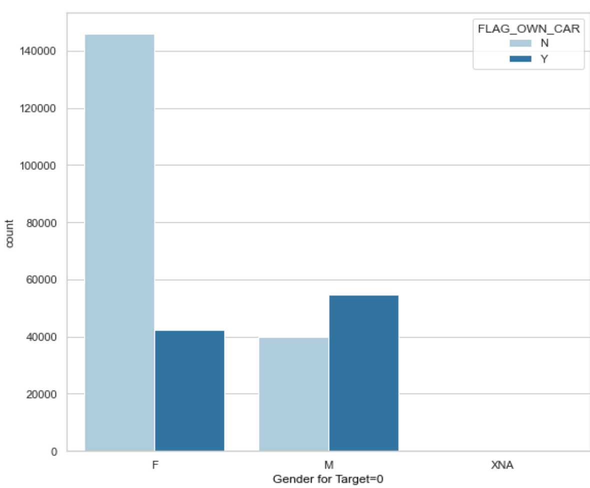
AGE_GROUP vs OCCUPATION_TYPE

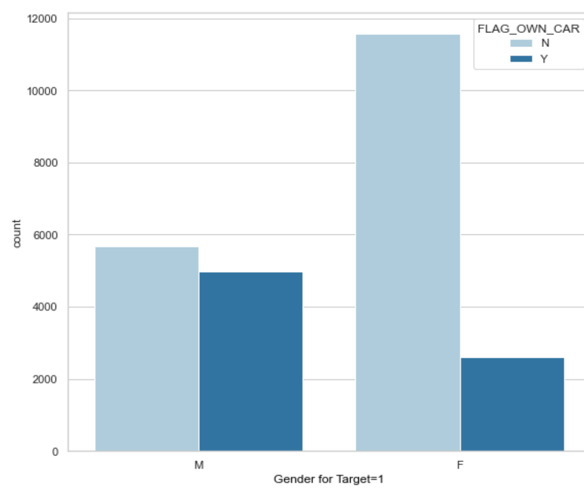




- Most the defaulters have occupation type Laborers.

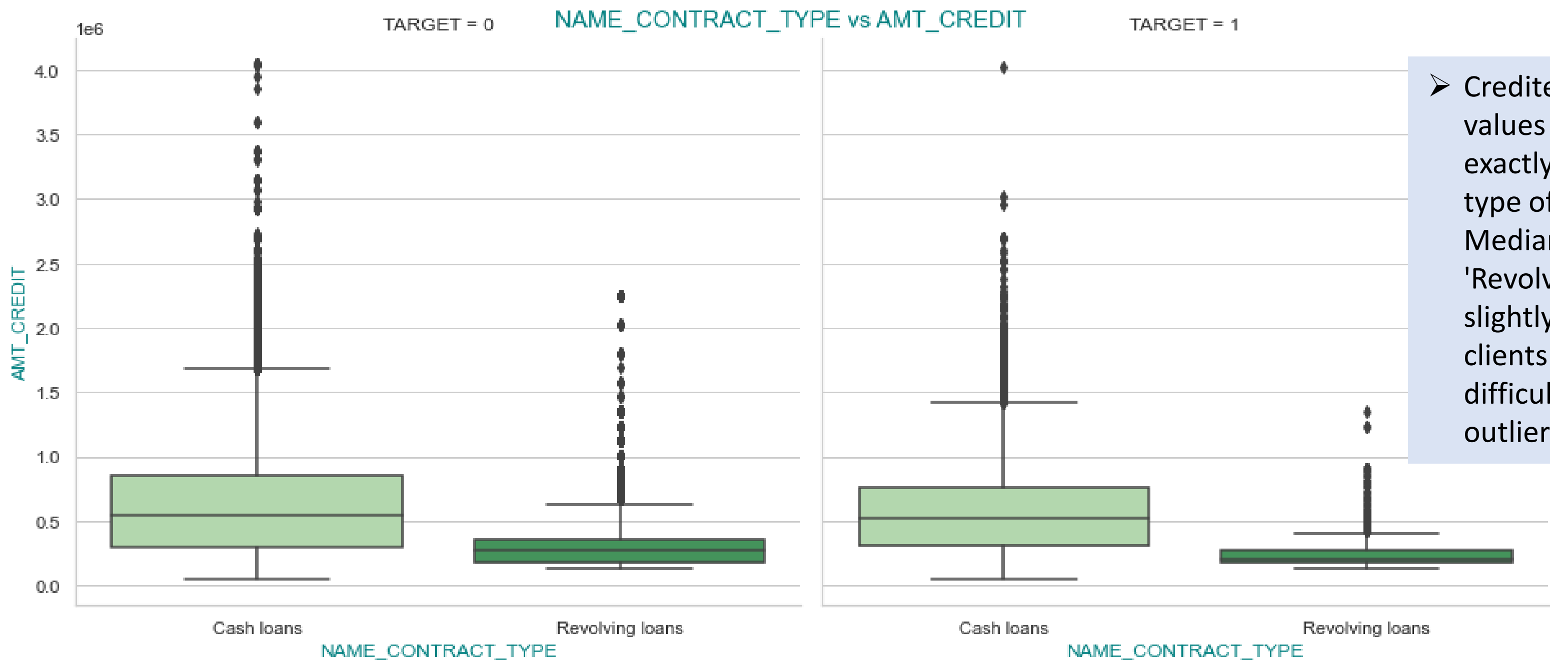
CODE_GENDER vs FLAG_OWN_CAR



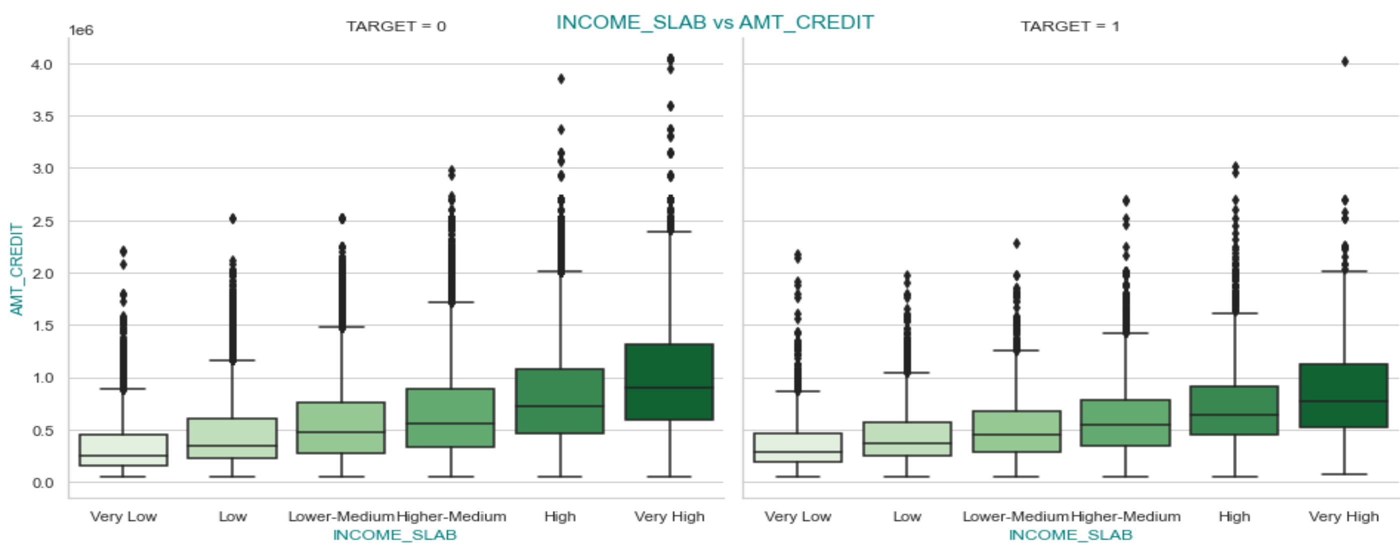


- Most of the customers as well as most of the people who are facing problem in paying back the loan are Female and most of the Female don't own a car. For Male and Female both, if they have car, they are less likely to face difficulties in repaying the loan.

Bivariate Analysis of Categorical - Numerical Variables

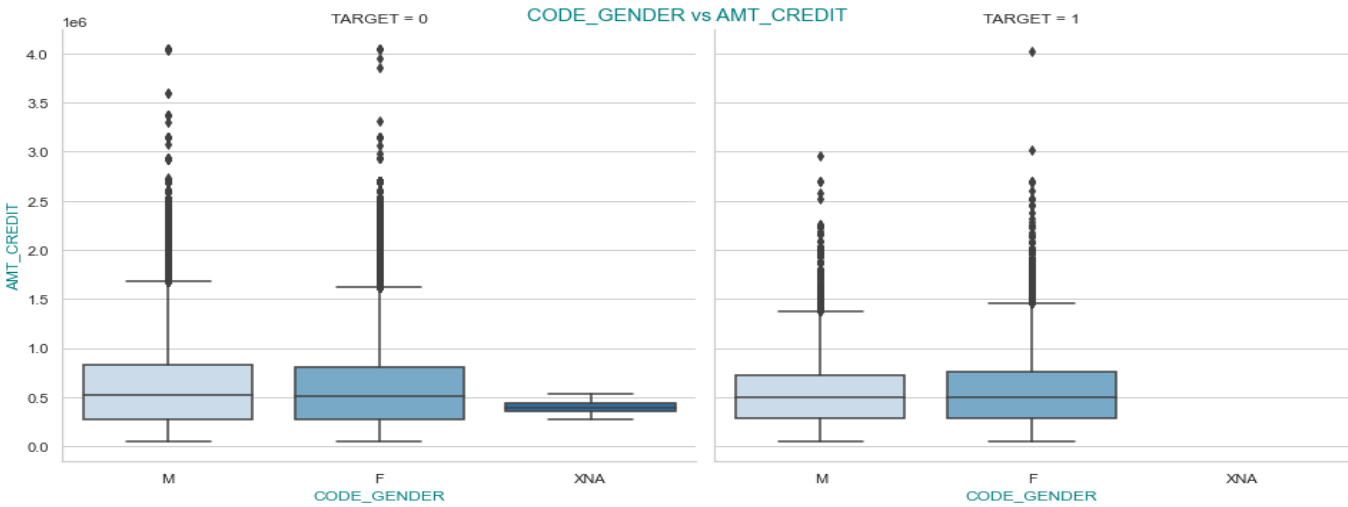


- Credited amount median values of 'Cash loans' is exactly same for both type of clients (0,1). Median value of 'Revolving loans' is slightly lower for the clients with payment difficulties. We can see outliers in both the plots.

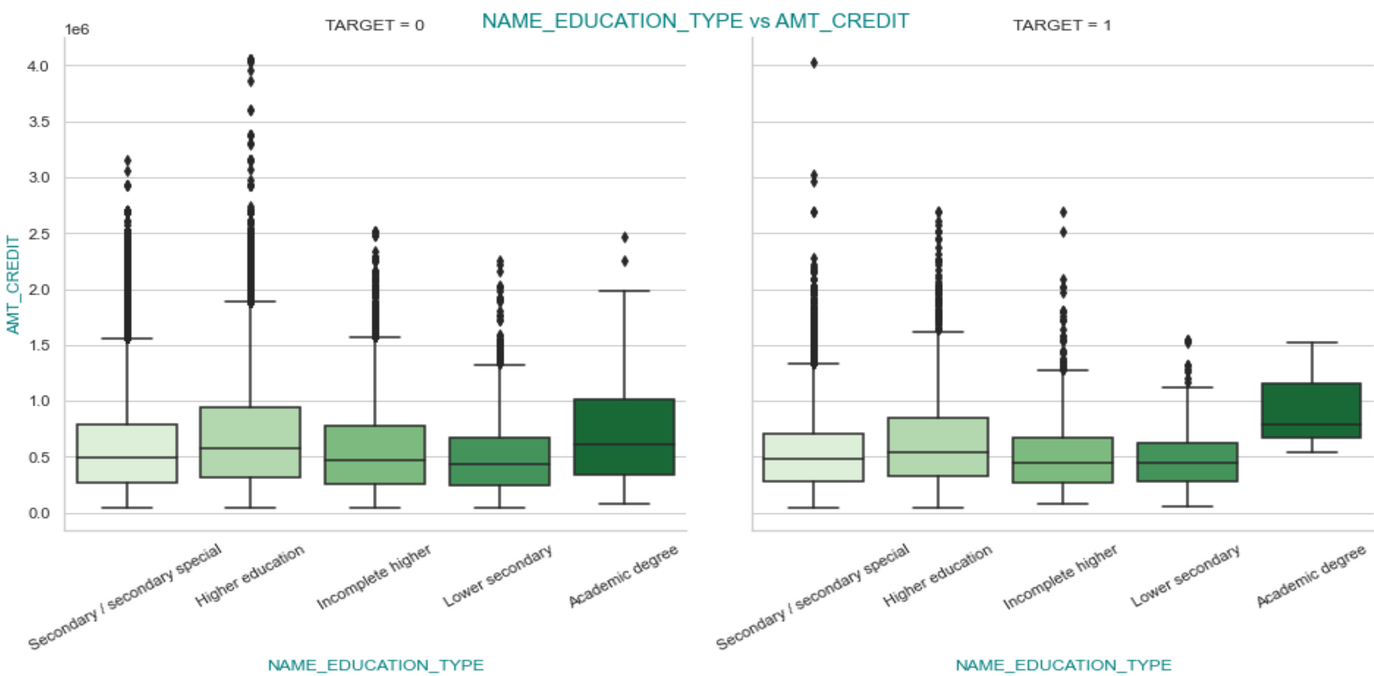


- Outliers can be observed in both the plots. Median values of credited amount for different INCOME SLABS are almost same in both TARGET type. We can also see increase in credited amount median value for higher income slabs.

Bivariate Analysis of Categorical - Numerical Variables

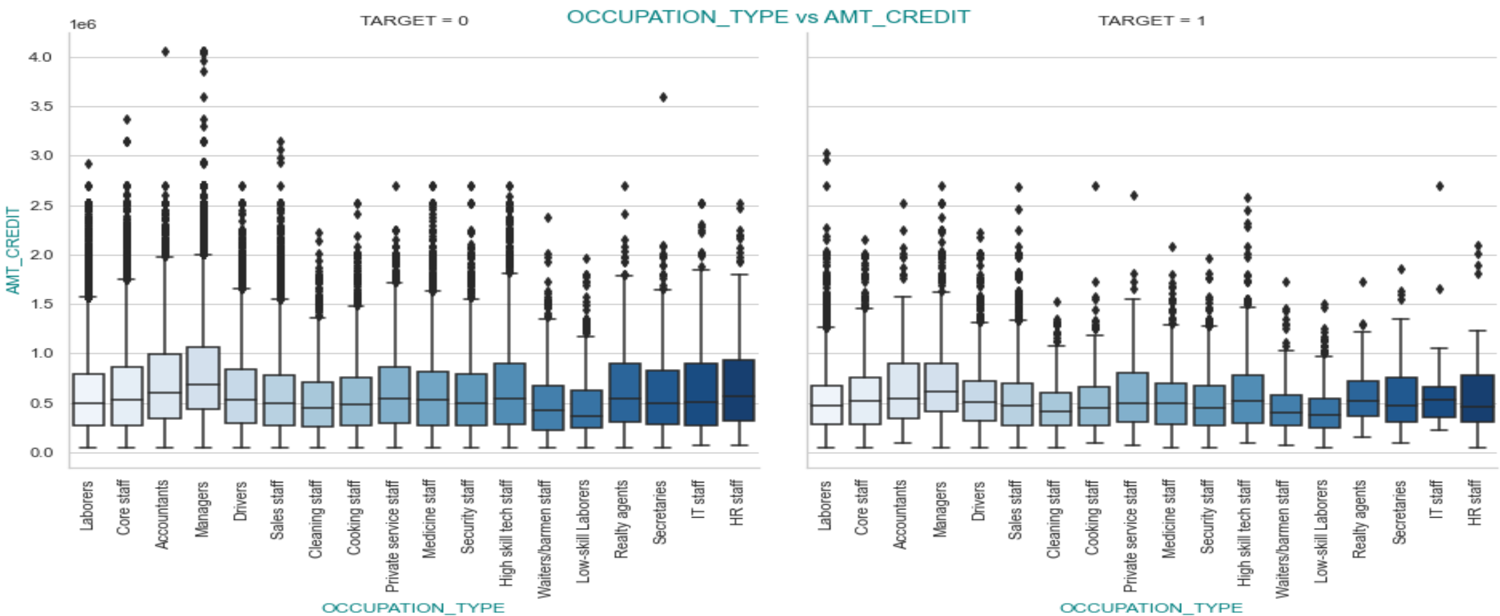


- There is not much difference in the plots for TARGET=0 and 1.

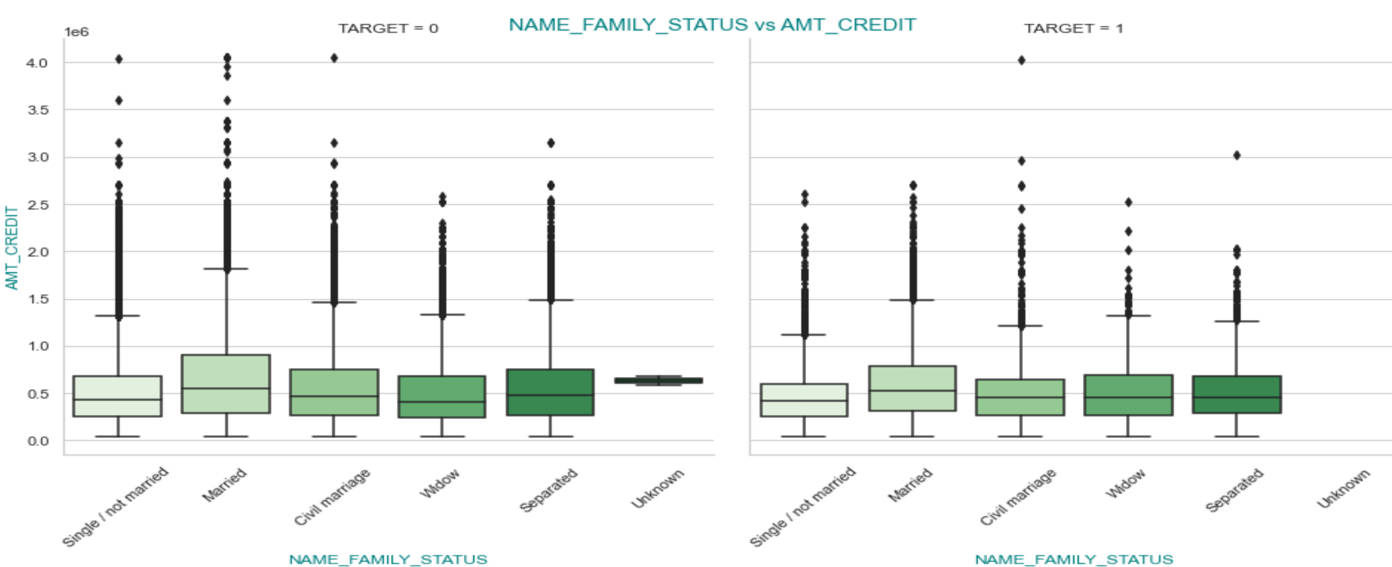


- Median value of Amount credit is higher for clients having 'Academic degree' and for Target=1 this value is even higher and most of the values are between 50 to 75 percentile.

Bivariate Analysis of Categorical - Numerical Variables



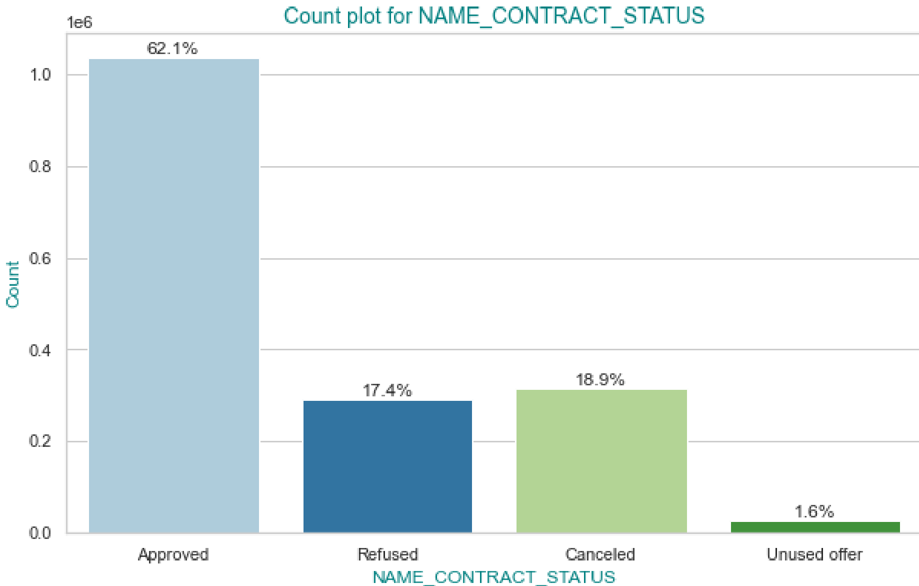
- Median value of Credited amount is the highest for Managers and Lowest for Low-skill Laborers in both the plots.

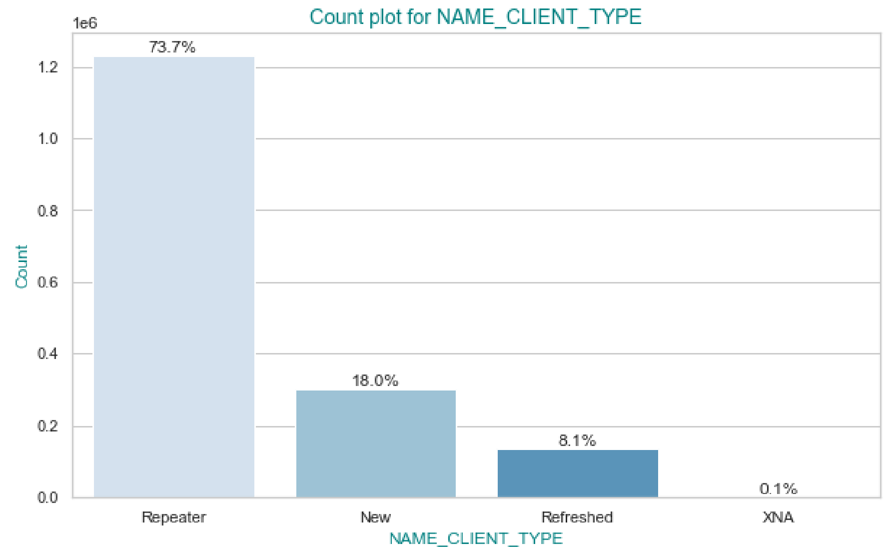


- Median value of Credited amount is the highest for Married clients and lowest for Single/ not married clients in both the cases. Median values of different groups are almost same for TARGET=0 and TARGET=1 clients.

Previous Application Data

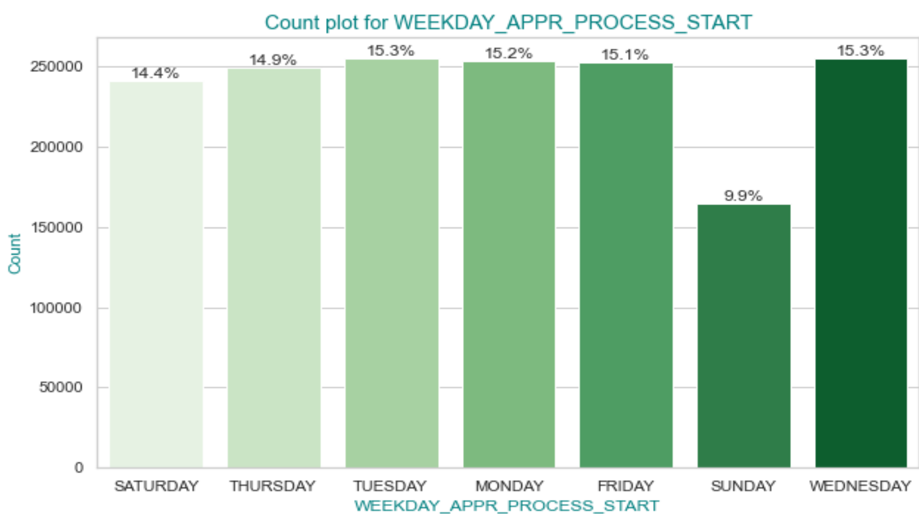
Univariate Analysis of Numeric Variables





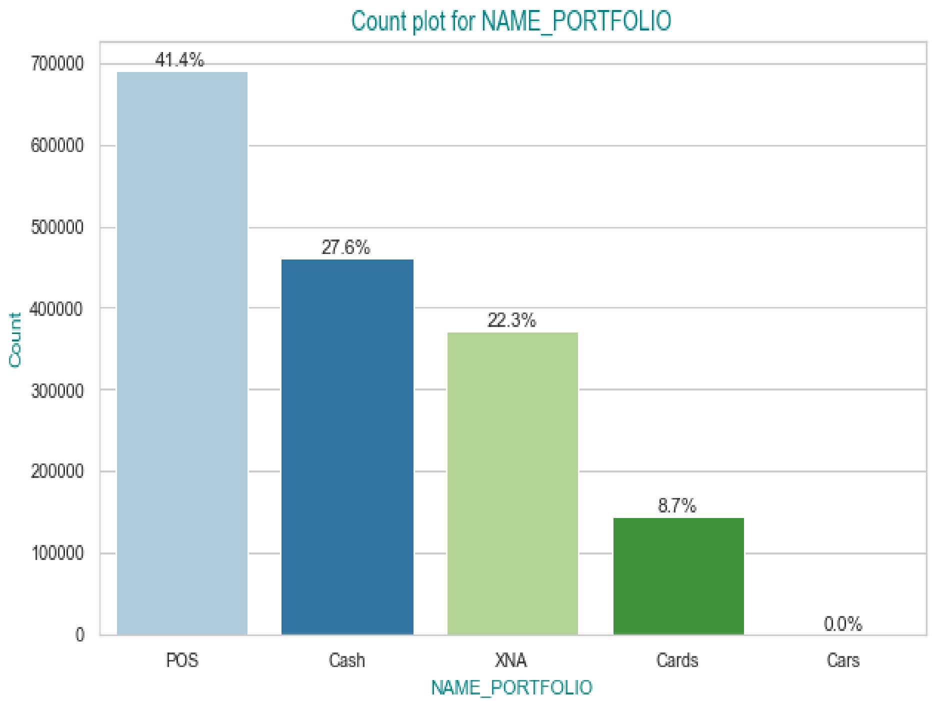
- Most of the loan applications are 'Approved' and only 1.6% are 'Unused offer'.

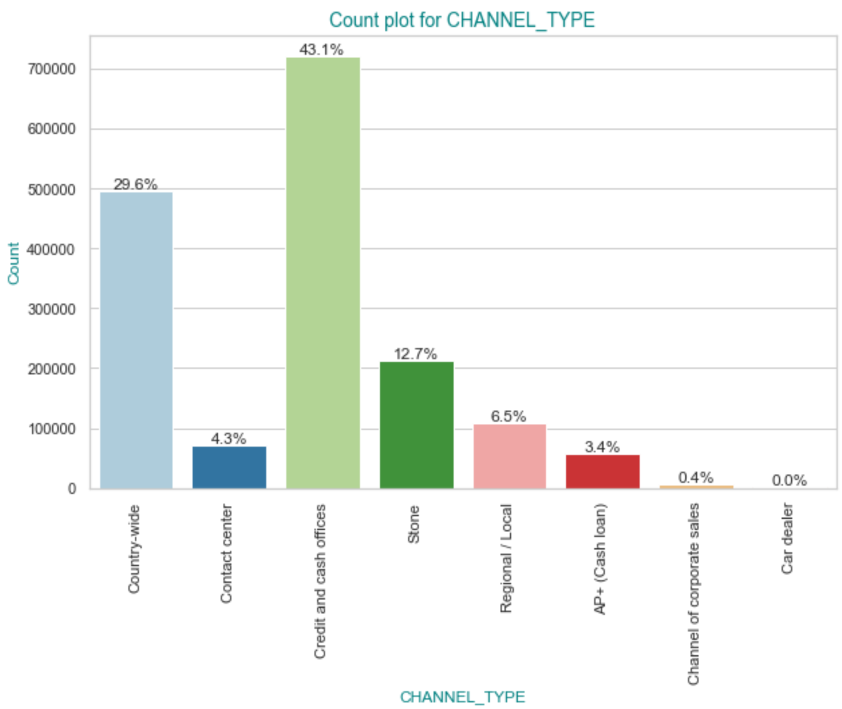
- Most of the clients in previous applications are Repeater clients.



- On Sunday numbers of loan applications are the least and it's highest on Tuesday and Wednesday.

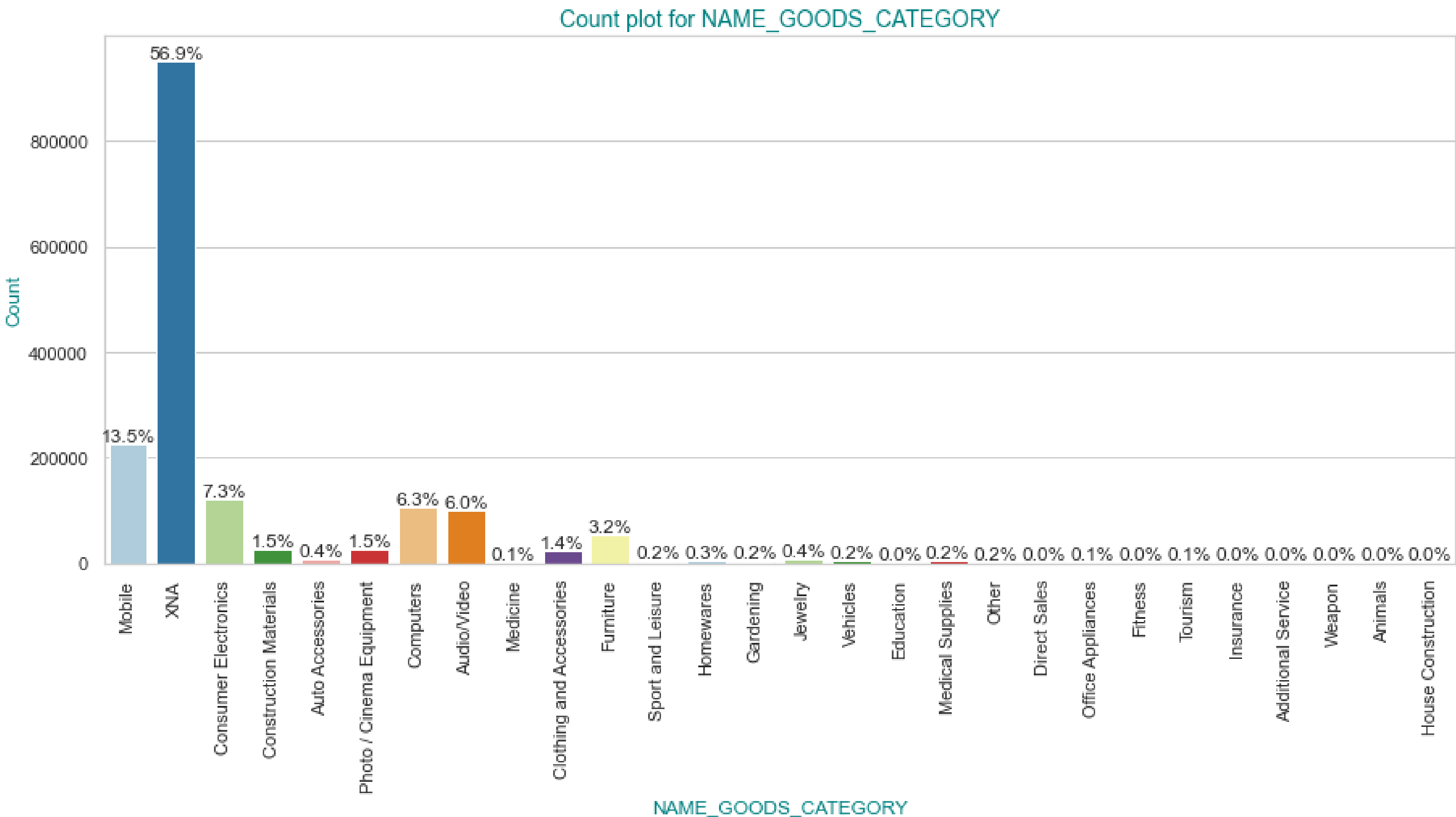
Univariate Analysis of Numeric Variables





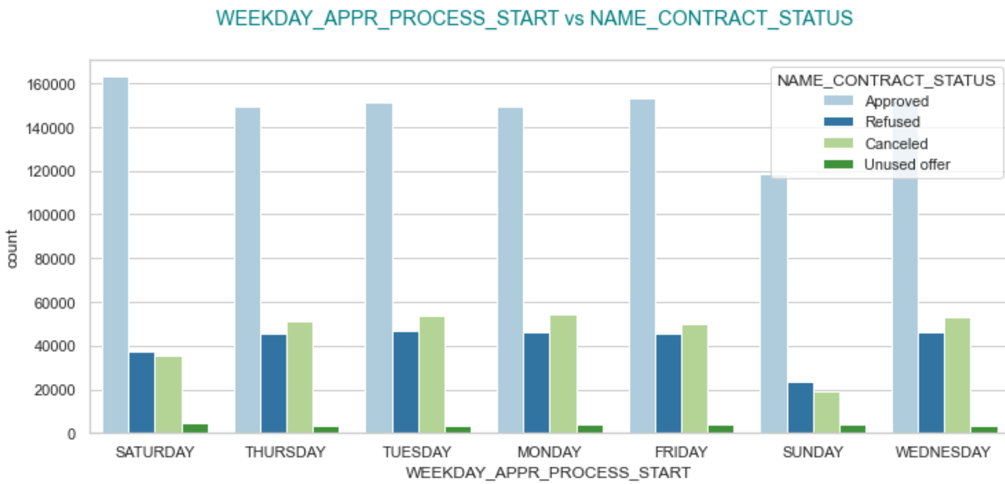
- Most of the previous applications are related to Point of Sale (POS) then followed by Cash.
- Most of the clients on the previous application were acquired through 'Credit and Cash office'.

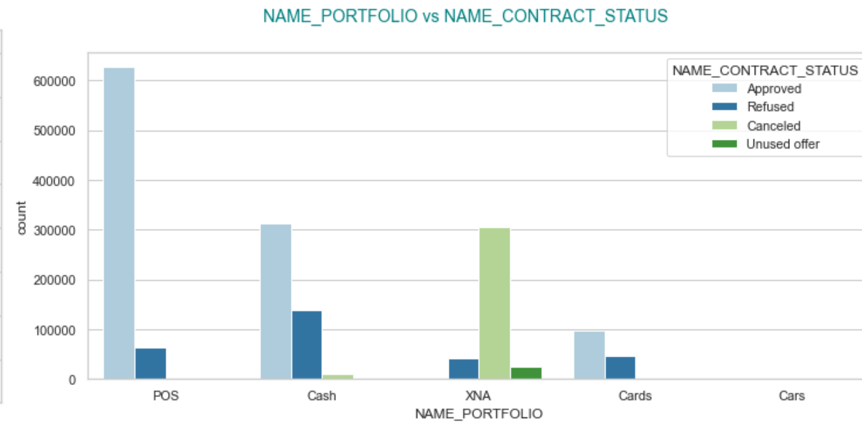
Univariate Analysis of Numeric Variables

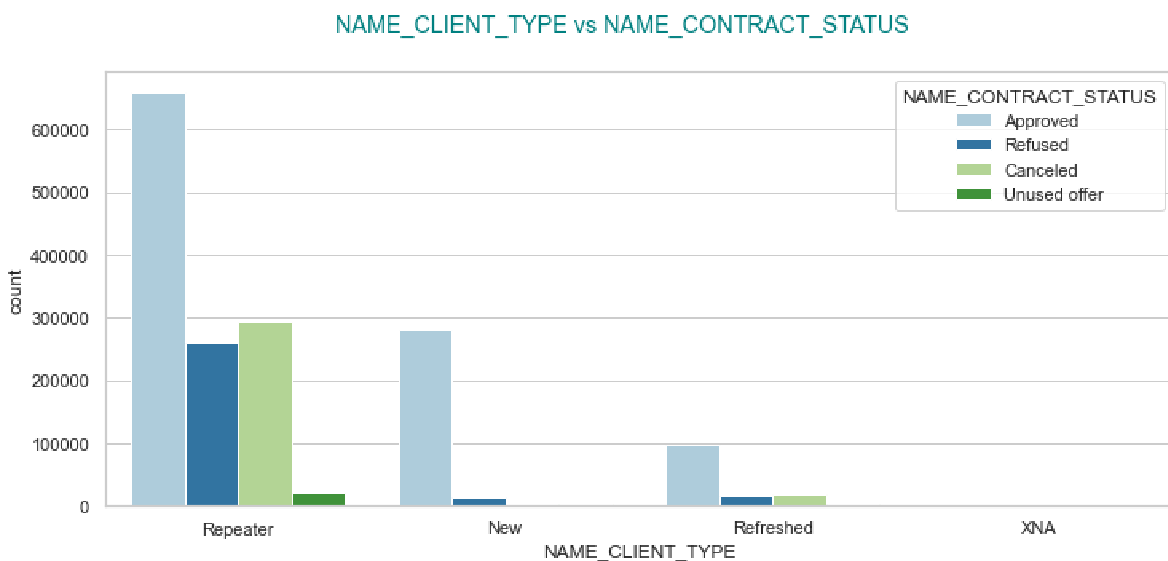


➢ Most of the previous application Loans are taken for Mobile, Consumer Electronics, Computers, Audio/Video.

Bivariate Analysis of Categorical - Categorical Variables

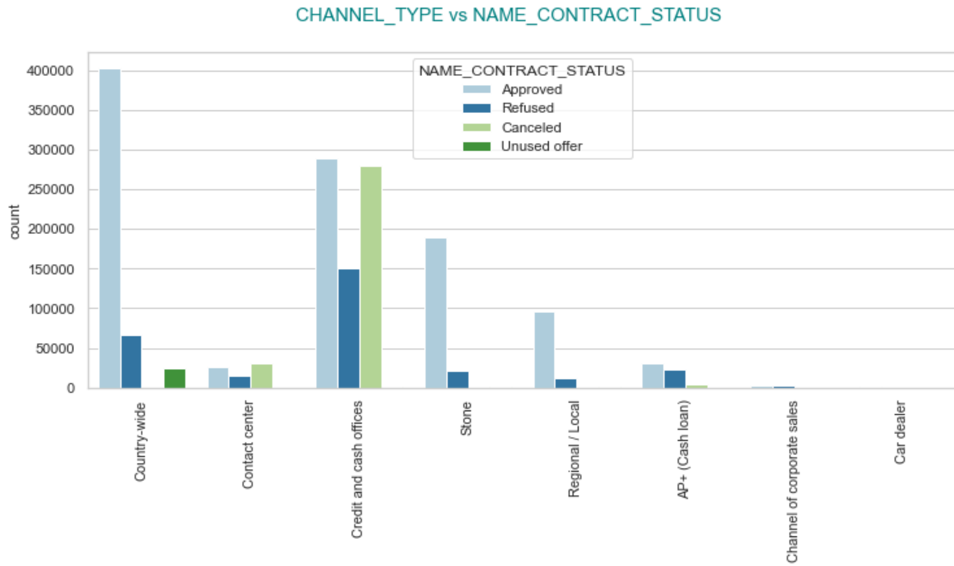




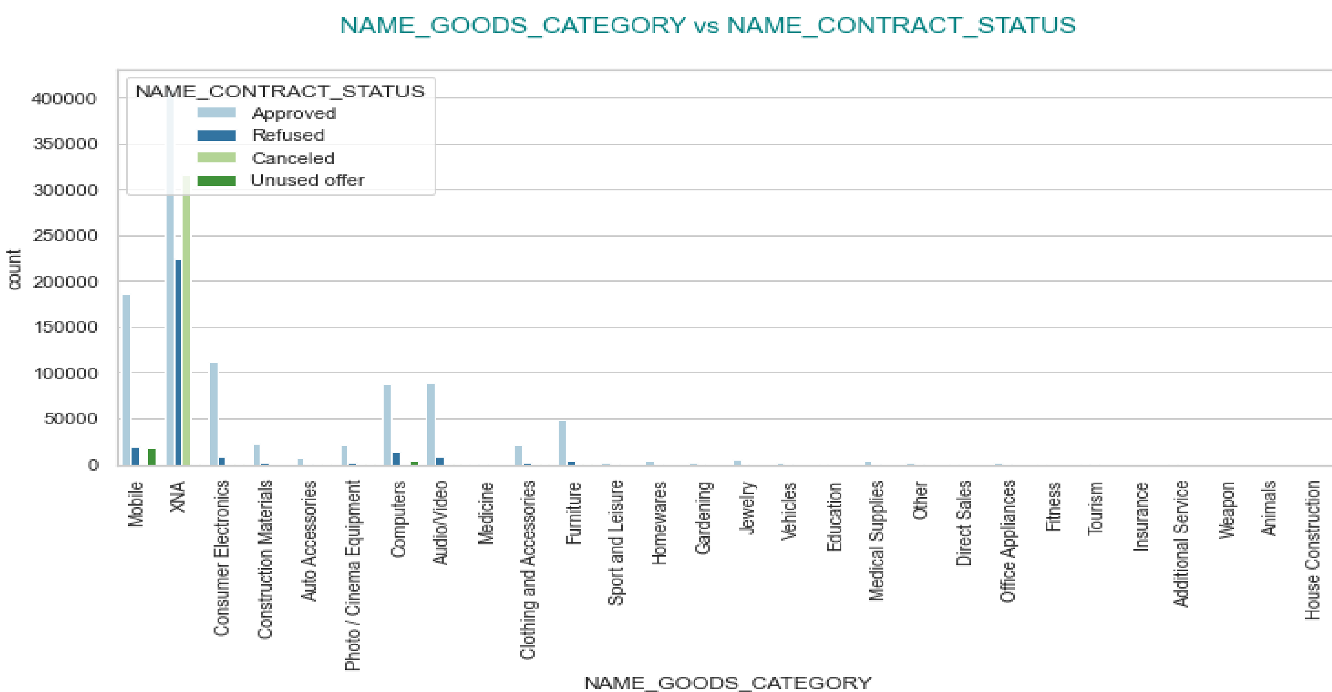


- Loans applied on Saturday has the highest chance of getting approved.
- Most of the previous applications are from Repeater clients. There are no canceled status for new clients. Most of the unused offers are from Repeater customers. Probably it's easier for the Repeater customers to get a loan approved, so few of them might have approved loans though they did not use it later.
- For POS and CARDS there are no unused loans. Chance of a loan application of getting Refused is higher when it's Cash or Card. The loan is very unlikely to get refused if it's POS.

Bivariate Analysis of Categorical - Categorical Variables

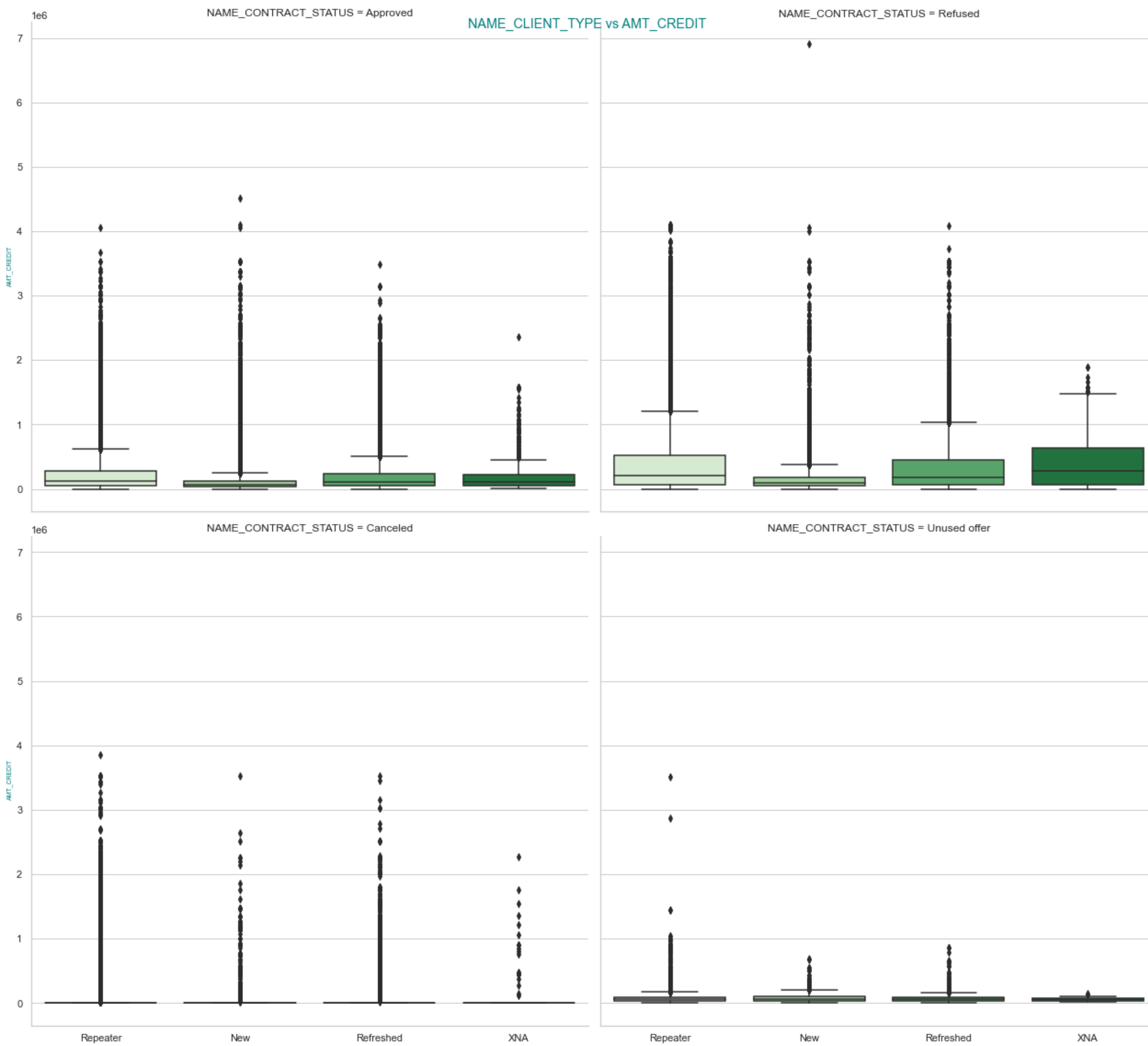


- Clients acquired by bank through 'Credit and Cash offices' have highest chance of getting refused and Cancelled. The loan application most likely will not getting Cancelled if the customer is acquired through 'Country-wide' channel. 'Contact center' channel clients also have very lower chance of loan getting approved.



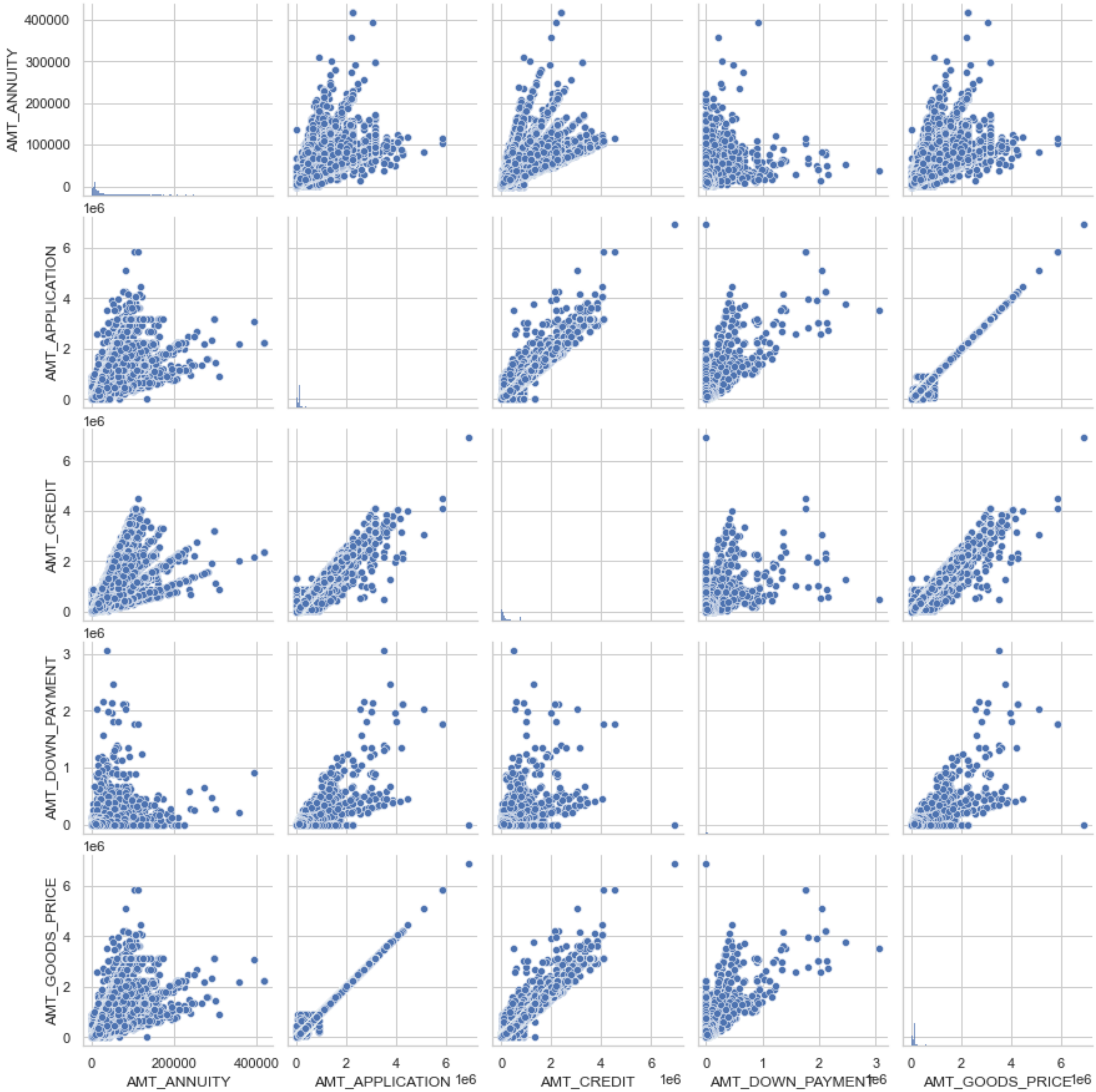
- Most of loan applications belongs to Mobile goods category. There is a very high chance for the application to get approved if the application is for the good's category Mobile, Computers, Audio/Video, Furniture.

Bivariate Analysis of Categorical - Numerical Variables



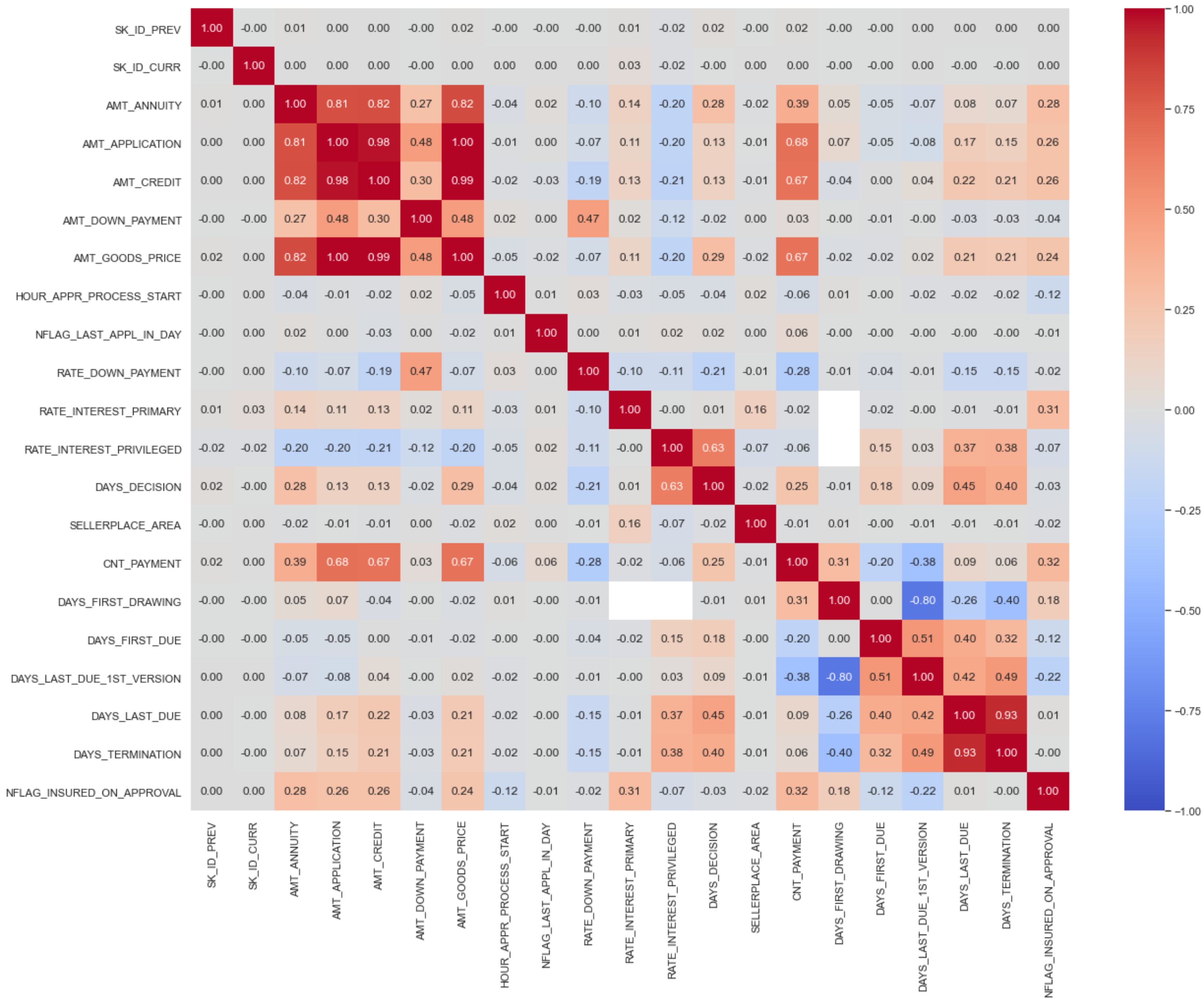
- Median value of Credited Amount of Approved applications are lower in compare to Refused. IQR for all type of applicants for approved applications more compact than Refused application. We may say applications are getting approved if the credited amount is within a particular range.

Bivariate Analysis of Numerical - Numerical Variables



It's obvious that there will be a very strong positive linear relation between AMT_APPLICATION and AMT_GOODS_PRICE. Medium positive linear relation is there between AMT_APPLICATION and AMT_CREDIT.

Correlation Analysis



Conclusion (1) – Application Data

- In sample number of Female clients are much higher than the number of Male clients. Female clients are less likely to face payment difficulties than Male clients. Bank should consider it as a feature to predict loan defaulters.
- Clients who have submitted Document 6 during loan application have lesser chance to have payment difficulties. Bank needs to check Document 6 during loan application. By making it a mandatory document may help to reduce defaults on loans.
- There is a slightly higher chance of clients paying the loan in time if the client owns at least one car. Bank may gather this insight while approving a loan application.
- Clients having "Higher Education" are more likely to repay the loans better than other education groups. Bank should consider the education as an important parameter before approving the loan.
- Most of the clients are "Married" and they are less likely to face difficulties in loan repayment. Where Single, Window are more likely to default on the loan. Clients living with their parents are more likely to default on a loan. Bank should consider this feature along with other features while approving a loan.
- Peoples of age group 20-40 of Low ad Higher-Medium income groups and occupation type Laborers are most likely face payment difficulties. So, applicants matches all these criteria should be validated on other parameters before approving their applications.
- Most of the clients are Laborers and they are facing more difficulties in repaying the loan in compare to Managers, Accountants, Core Staff, High Skill tech Staff etc. higher paying jobs. Bank should consider applicant's current job status before approving the loan application.
- Pensioners are more likely to repay the loan in compare to other income classes. The reason could be, the pensioners have a fixed stable income, and they may take loan of a calculated amount, so that the monthly installment can be covered using the pension income. Bank should encourage for pensioner clients as it’s comparatively profitable for the bank.

Conclusion (2) – Previous Application Data

- Loans applied on Saturday has the highest chance of getting approved.
- Most of the previous applications are from Repeater clients. There are no canceled status for new clients. Most of the unused offers are from Repeater customers. Probably it's easier for the Repeater customers to get a loan approved, so few of them might have approved loans though they did not use it later.
- For POS and CARDS there are no unused loans. Chance of a loan application of getting Refused is higher when it's Cash or Card. The loan is very unlikely to get refused if it's POS.
- Most of loan applications belongs to Mobile goods category. There is a very high chance for the application to get approved if the application is for the good's category Mobile, Computers, Audio/Video, Furniture.
- Clients acquired by bank through 'Credit and Cash offices' have highest chance of getting refused and Cancelled. The loan application most likely will not getting Cancelled if the customer is acquired through 'Country-wide' channel. ‘Contact center' channel clients also have very lower chance of loan getting approved.