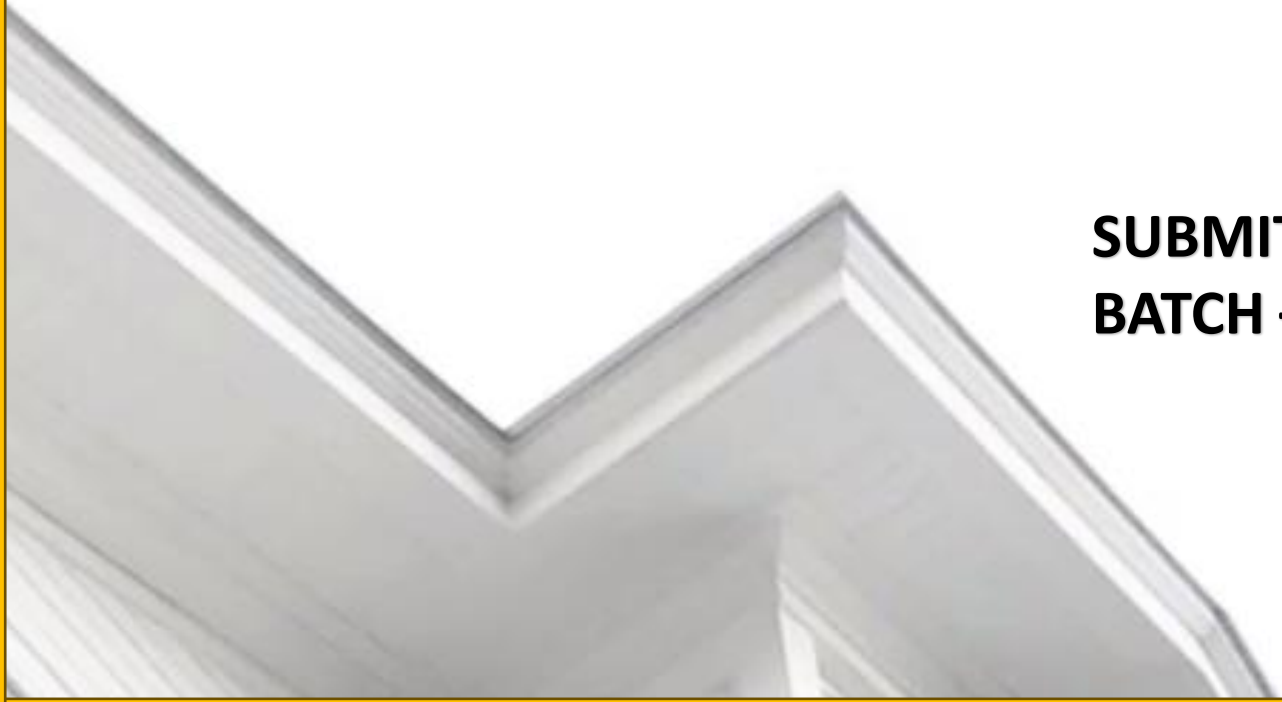
CREDIT EDA ASSIGNMENT



SUBMITTED BY –ANKITA SETHI
BATCH –DSC-56

BUSINESS OBJECTIVE

The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

PROBLEM STATEMENT

- The objective of this case study is to analyse whether aa client can pay their instalments & the data will be used for taking actions like reducing the amount of loan , rejecting the loan & giving the loan at higher interest rates for the risky applicants etc.
  This will make sure that the applicants who can repay the loans are not rejected.

- When an applicant applies for the loan four types of decisions could be taken by client/company:

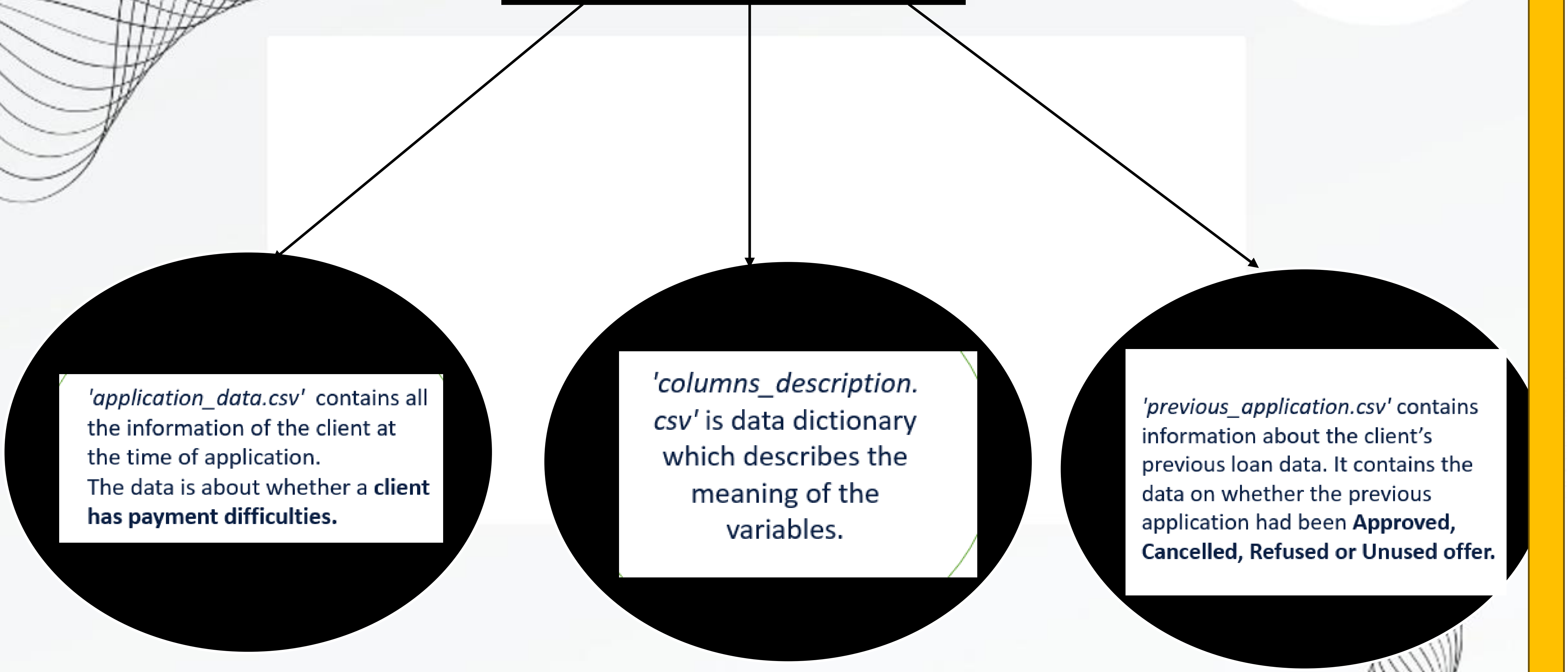
a)Approved
b) Cancelled
c) Refused
d) Unused

Two types of risks are associated with the bank's decision

1.If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.

2.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.

3 Types of data set



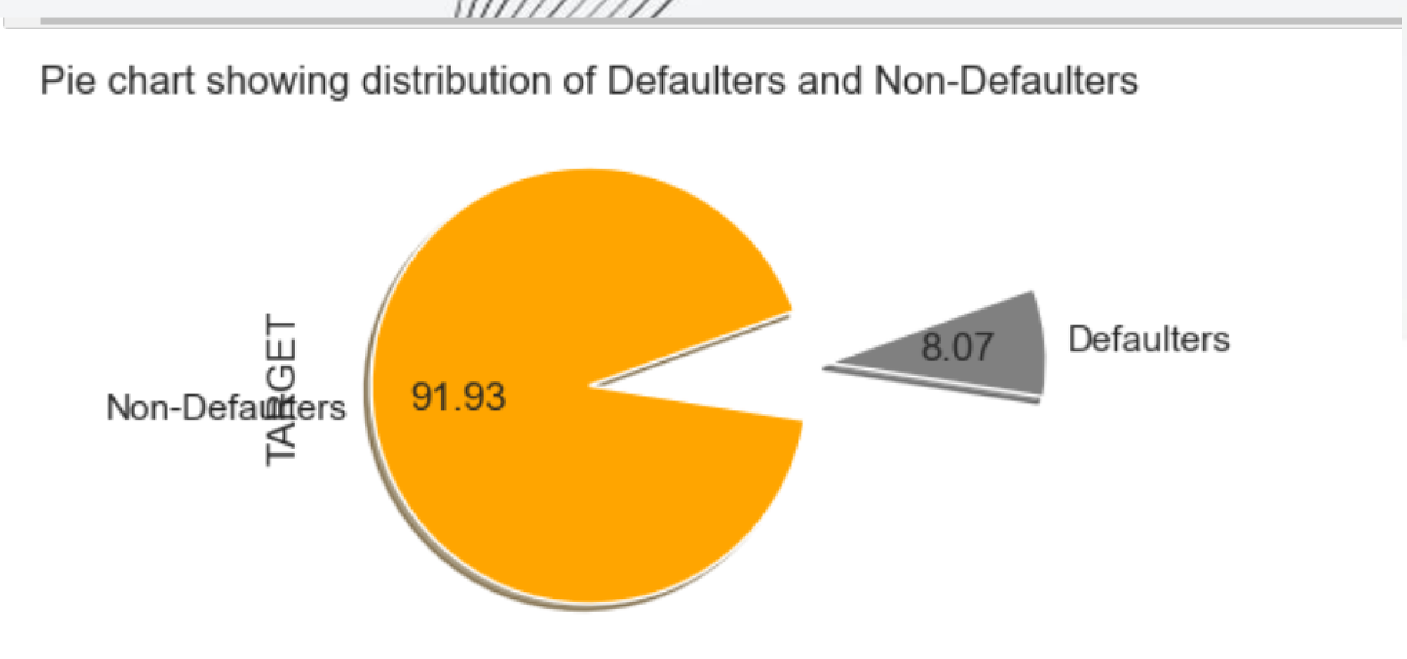
'application_data.csv' contains all the information of the client at the time of application.
The data is about whether a **client has payment difficulties.**

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

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

STEPS INVOLVED IN ANALYSIS

1. Importing libraries in jupyter notebook.
2. Importing csv files.
3. Checking shape, info , dtypes ,describe ,nunique values, missing values.
4. Dropping missing values.
5. Imputing missing values.
6. Checking Imbalance.
7. Detection of outliers.
8. Creating bins.
9. Segregating data into defaulters and non defaulters.
10. Performing analysis (univariate ,bivariate and correlation on numerical and categorical data).
11. Merging 2 data sets.
12. Performing further analysis.
13. Conclusion.



Pie chart showing distribution of Defaulters and Non-Defaulters

IMBALANCE IN DATA

Observation

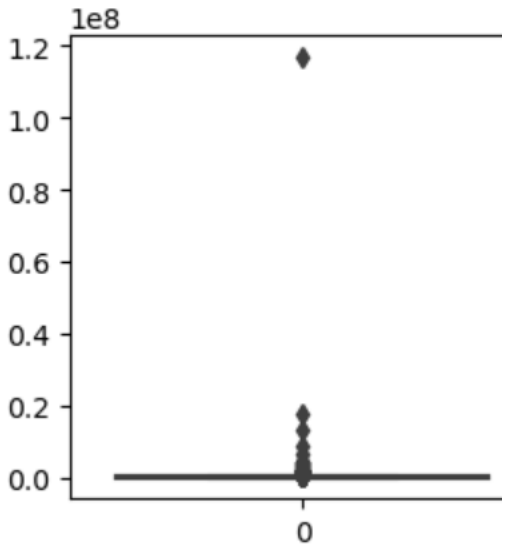
1.We can clearly see from the percentage values that the percentage of defaulter is 8.07 and non-defaulter is 91.92.

2. Pie chart is also shown for a target column where orange portion represents the Non-Defaulters and grey portion shows the defaulters.

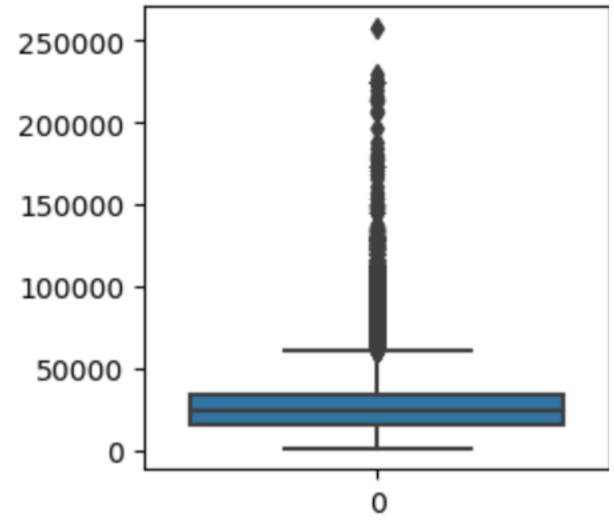
3. From the percentages of target we can see that the data is highly imbalanced.

DETECTION OF OUTLIER

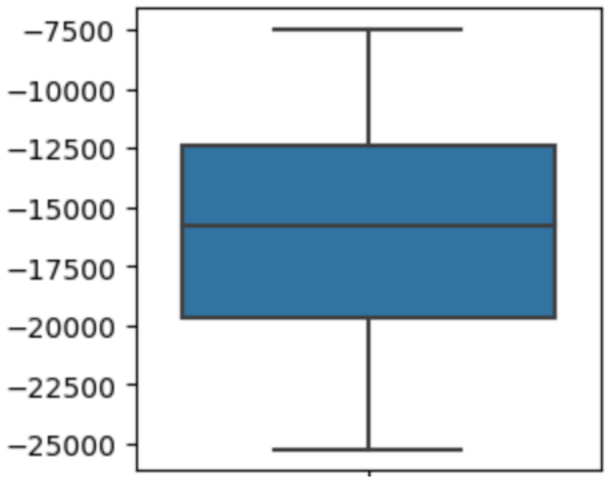
boxplot for AMT_INCOME_TOTAL



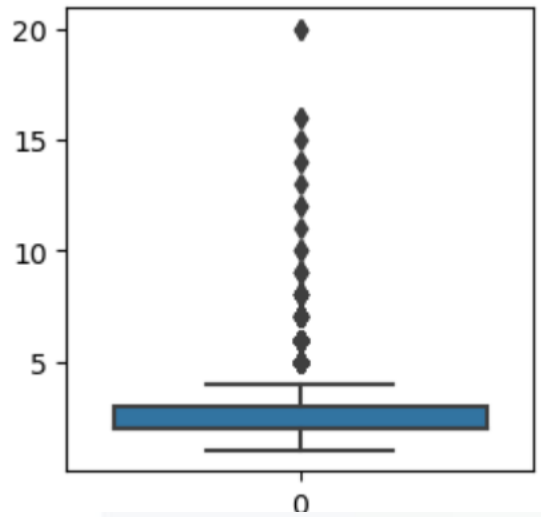
boxplot for AMT_ANNUITY



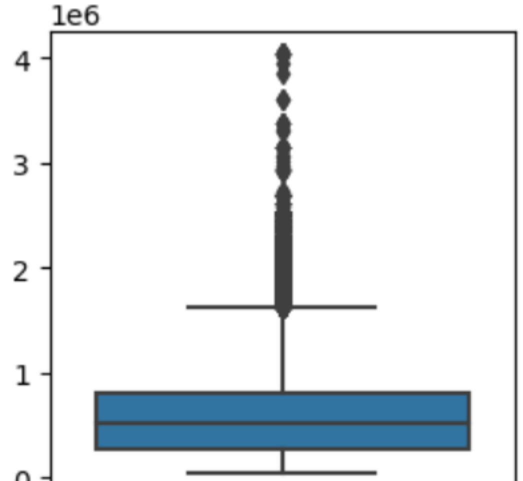
boxplot for DAYS_BIRTH



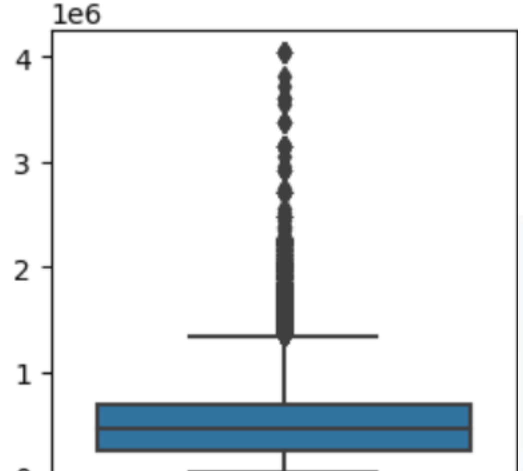
boxplot for CNT_FAM_MEMBERS



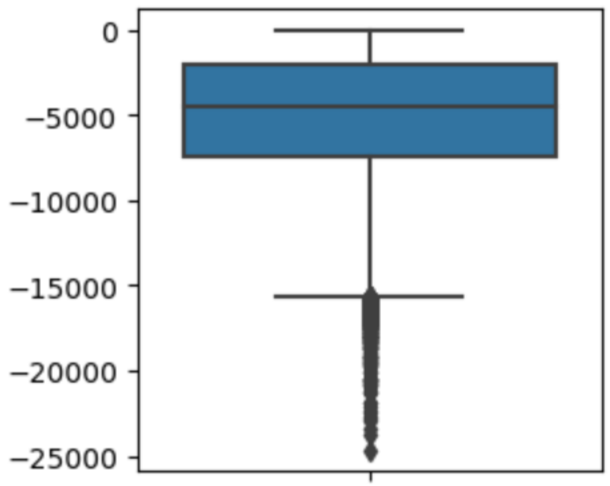
boxplot for AMT_CREDIT



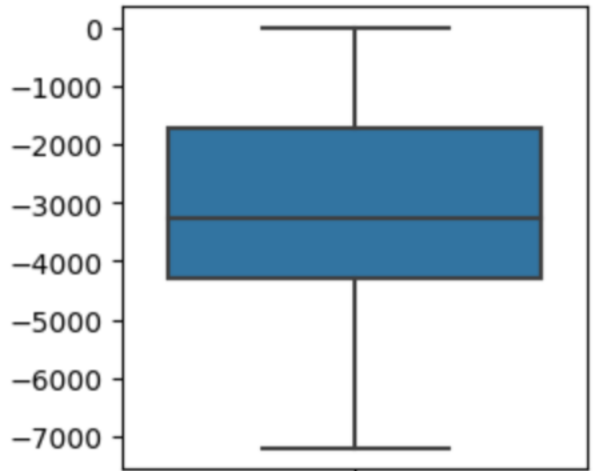
boxplot for AMT_GOODS_PRICE



boxplot for DAYS_REGISTRATION



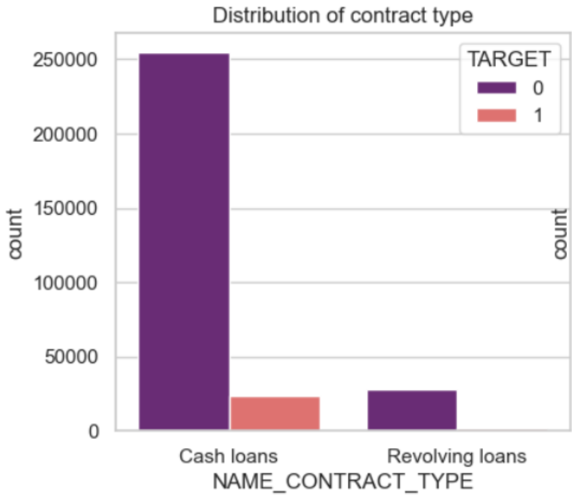
boxplot for DAYS_ID_PUBLISH

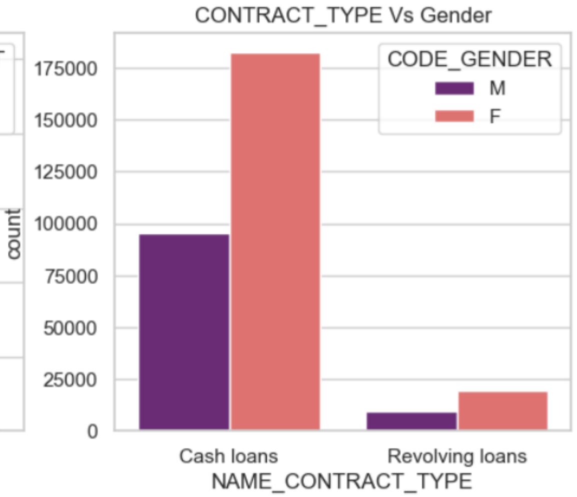


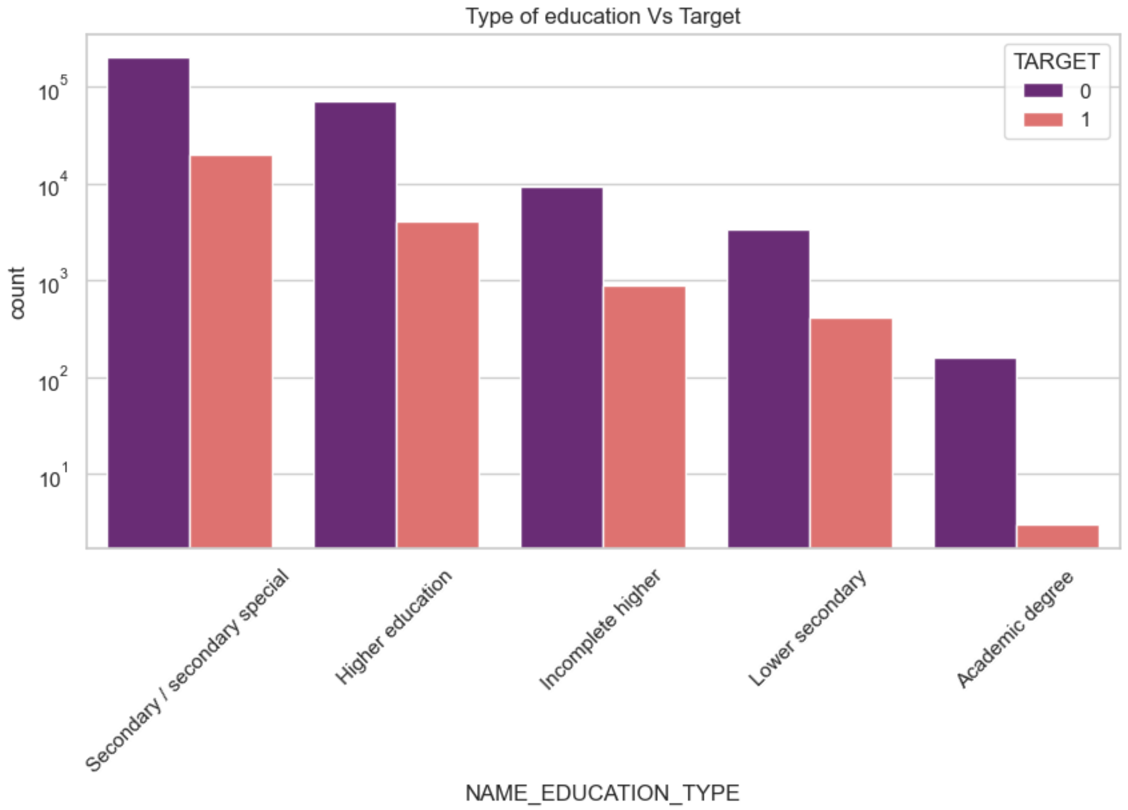
OBSERVATION OF OUTLIERS

1.Outlier- Value beyond the normal value is known as Outlier.
2.To find out outlier in continuous variable we generally use box plot.
3.Points lying outside the box is known as Outlier.
4. In boxplot for AMT_INCOME_TOTAL we can see a potential outlier we cant say its an outlier but yes its a potential outlier as a person can have high income also.
5. In amount credit if we compare 99th percentile value and maximum value there is a huge difference hence we can observe so many outliers in case of AMT_CREDIT.
6. Box plots for all days columns shows the number of days are in negative which is not possible even days of birth is shown in negative.
7. CNT_CHILDREN- Represents Number of children the client has.
As we can clearly see there are so many outliers in case of children of client.
Probability of having 19 children is quite less so it will come under outlier.
we can even observe some values such as 14,12,11,9....so on.

UNIVARIATE ANALYSIS- CATEGORICAL COLUMNS







OBSERVATIONS

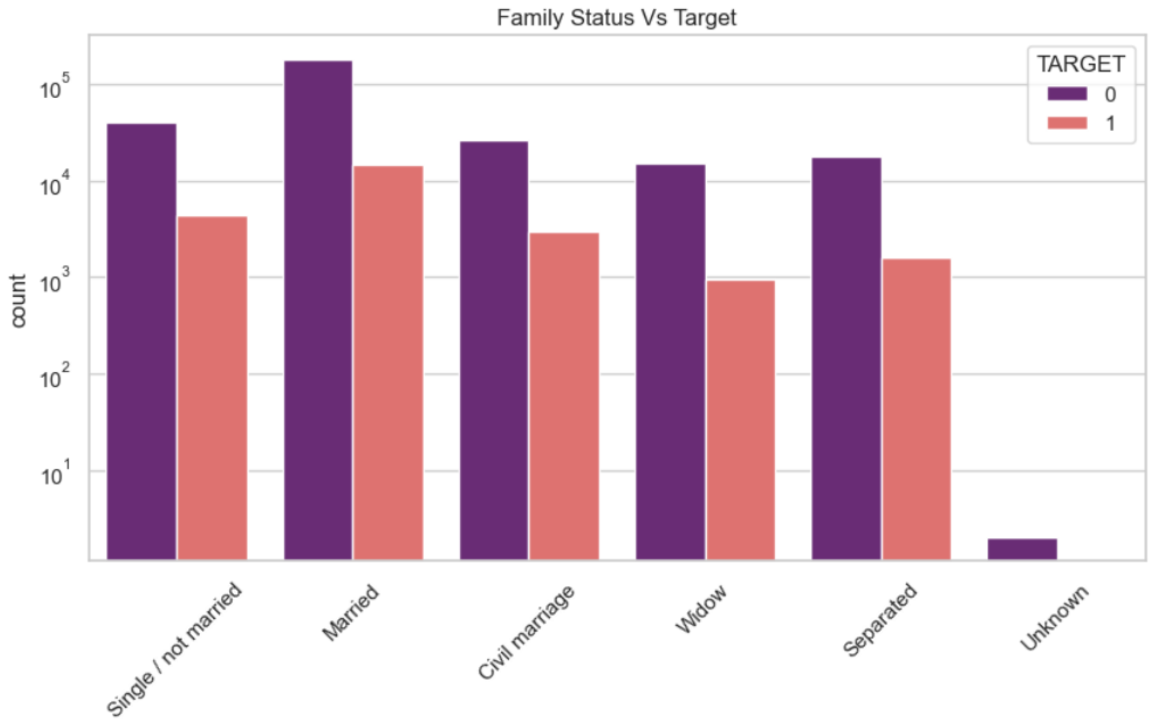
1.From the first plot ie.Distribution of contract type shows that the number of non-defaulter are more for the one who is taking cash loans as compared to the Revolving loans.

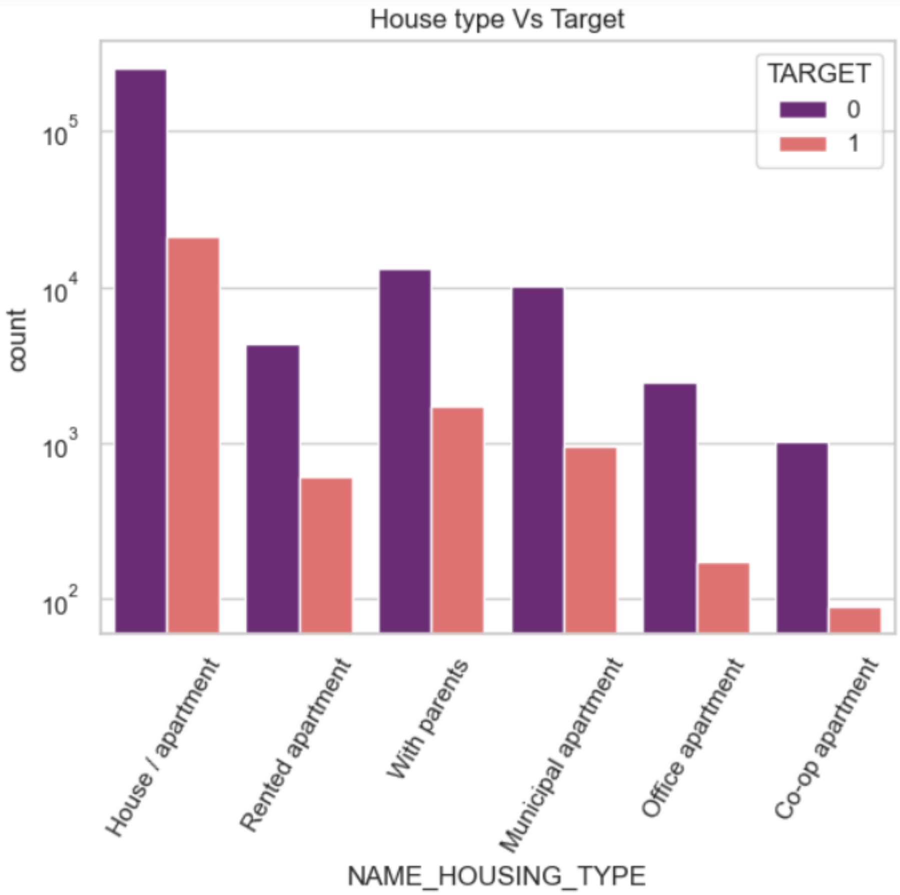
2. In the second plot i.e., CONTRACT_TYPE Vs Gender shows that the Females are taking more loans as compared to males and females are taking more Cash loans as compared to Revolving loans. Even if we talk about males they are taking more cash loans as compared to Revolving loans.

OBSERVATIONS

1.From the above graph it is clear that the number of defaulters are more in Secondary/Secondary special than in Higher education and the number of loan applicants are more in this category only ie. Secondary/Secondary special .

2.There are least number of defaulters in case of Academic Degree.



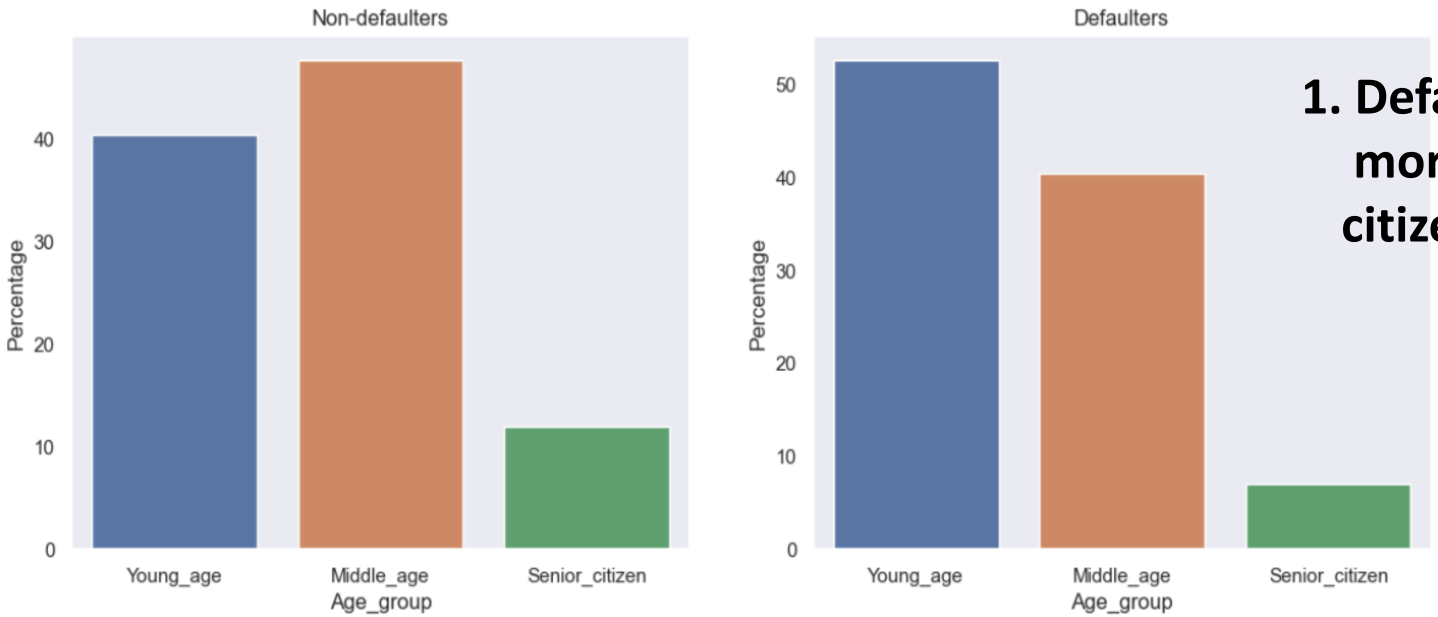


OBSERVATIONS

1. From the above plot we can observe that the number of defaulters are more in Married category and then in Civil marriage.
2. We can observe that married people are more towards applying loan.
3. Less number of defaulters are in widow category.

OBSERVATIONS

1. The one who lives in his /her own house/apartment are more toward applying loan.
2. The number of defaulters are very less in rented apartment, office apartment & co-op apartment.
3. It is very much clear the one who is owner of house/apartment has taken loan in more number.



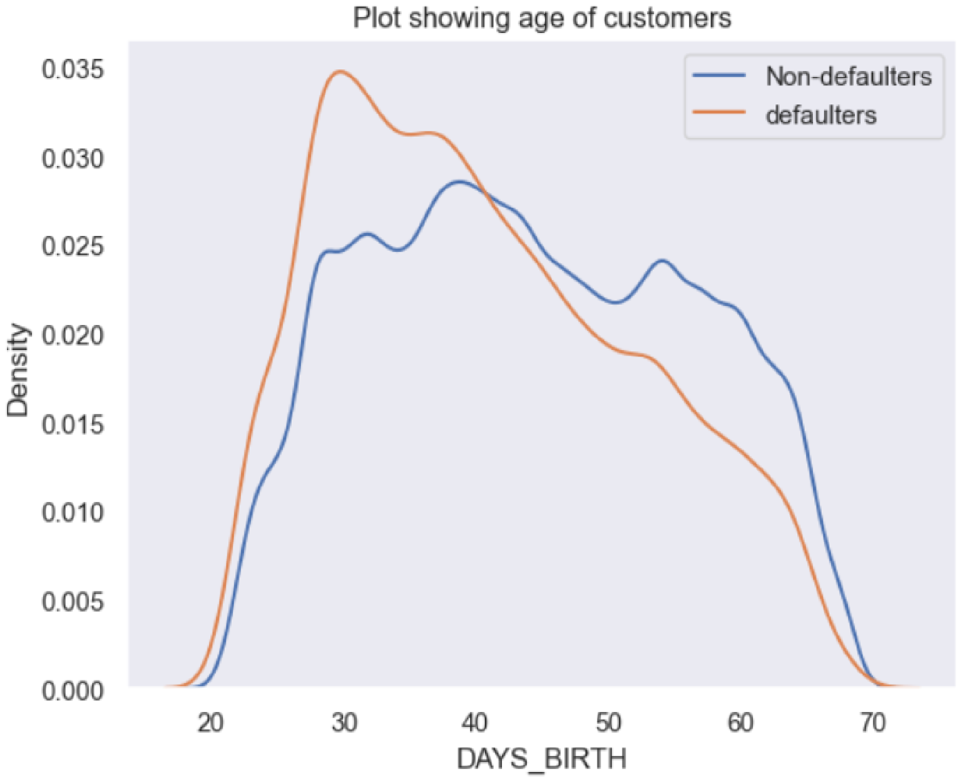
OBSERVATION

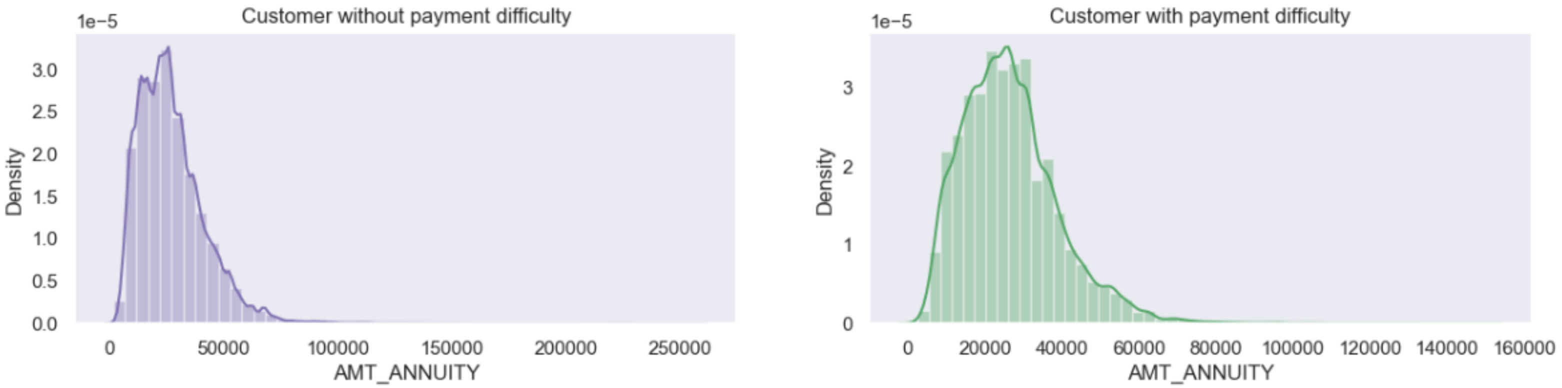
1. Defaulters- We can see that young people are more likely to be a defaulter where as senior citizen are less likely to be a defaulter.

CONTINUOUS COLUMN ------ UNIVARIATE ANALYSIS

OBSERVATIONS

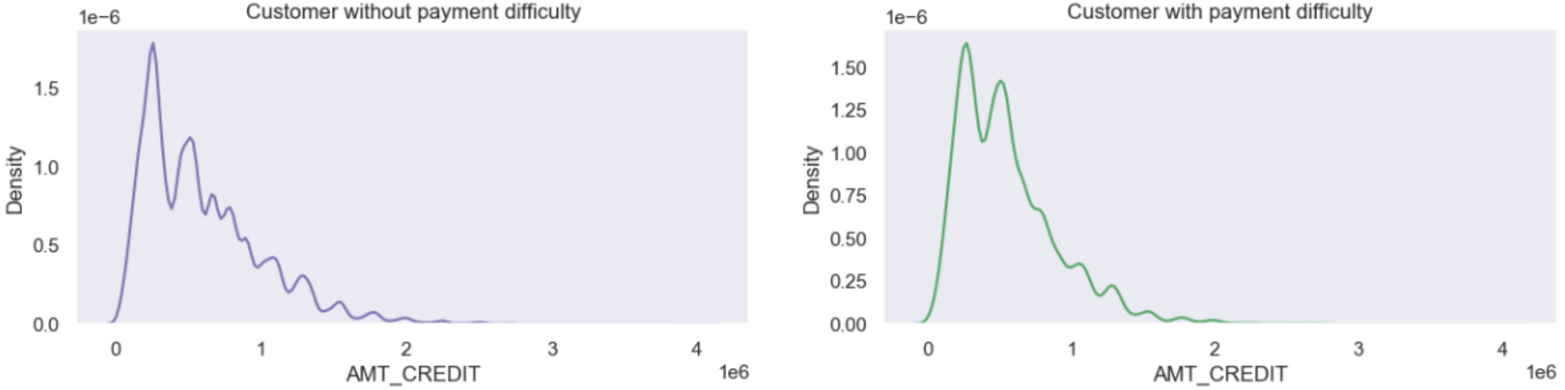
1. If we consider DAYS_BIRTH ie age in years we can see that in case of defaulters median lies at 40 and for non defaulters median lies at 43 approximately.
2. Customer with payment difficulty lies in 31 years to 50 years & the customers without payment difficulty falls in range of 34 years-55years.A peak is seen at 58-60 which means retired people are taking more loans





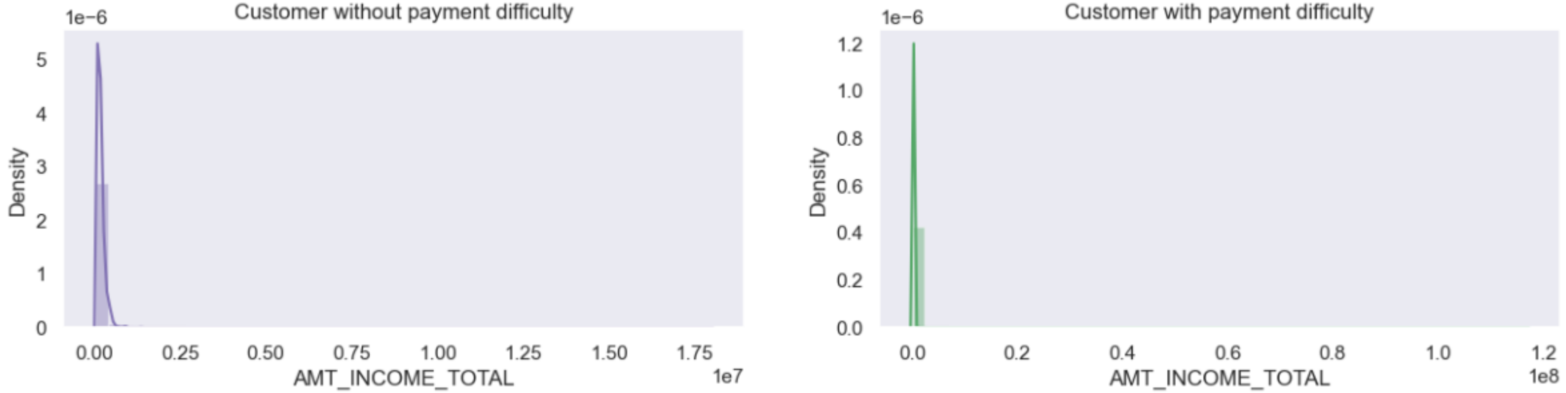
OBSERVATIONS

For AMT_ANNUITY the plots are quite similar. It shows that AMT_ANNUITY for defaulters has maximum count in 25000-30000 for non defaulters it is 35000.



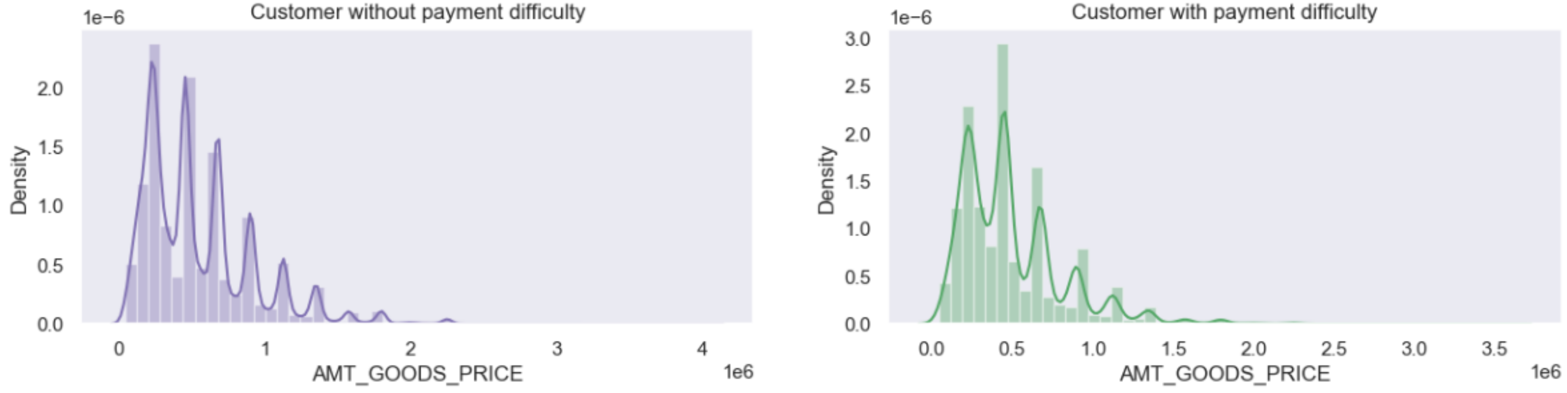
OBSERVATIONS

1. AMT_CREDIT- for defaulters the density lies in region 1.25-1.5.



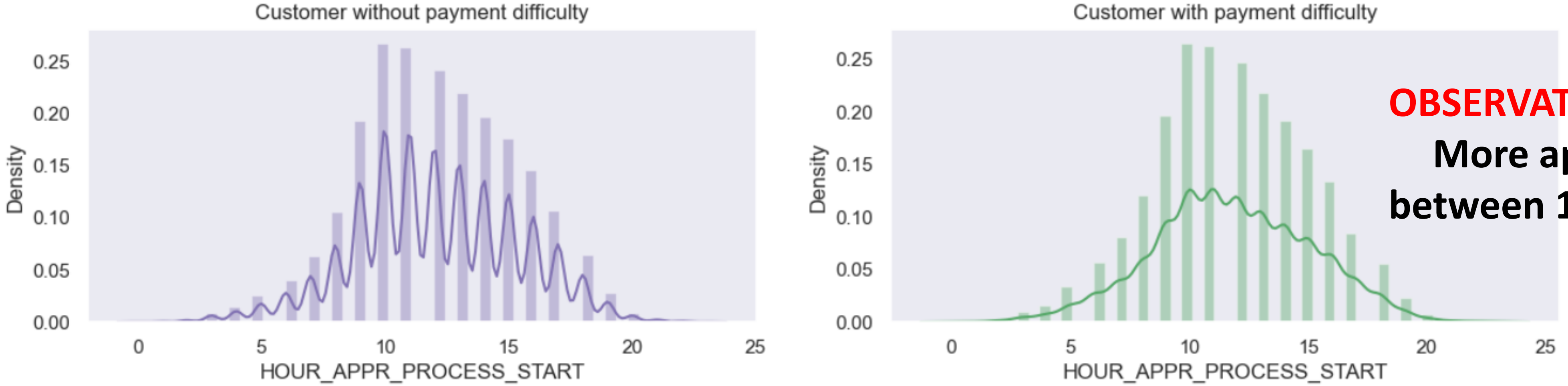
OBSERVATIONS

People with lower total income are likely to be defaulter.



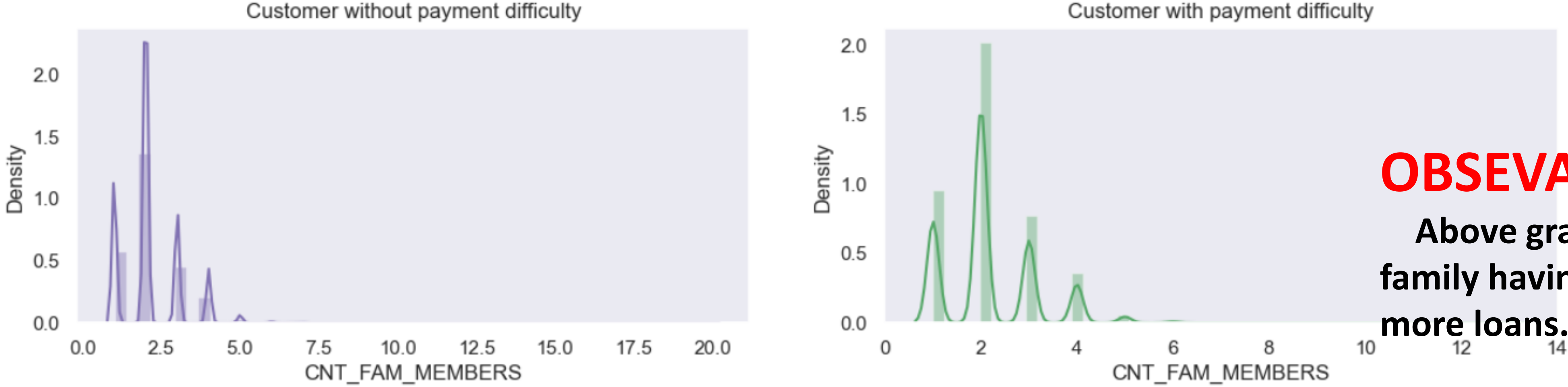
OBSERVATIONS

In first graph maximum density is around 0.5 and for defaulters it is above 0.5



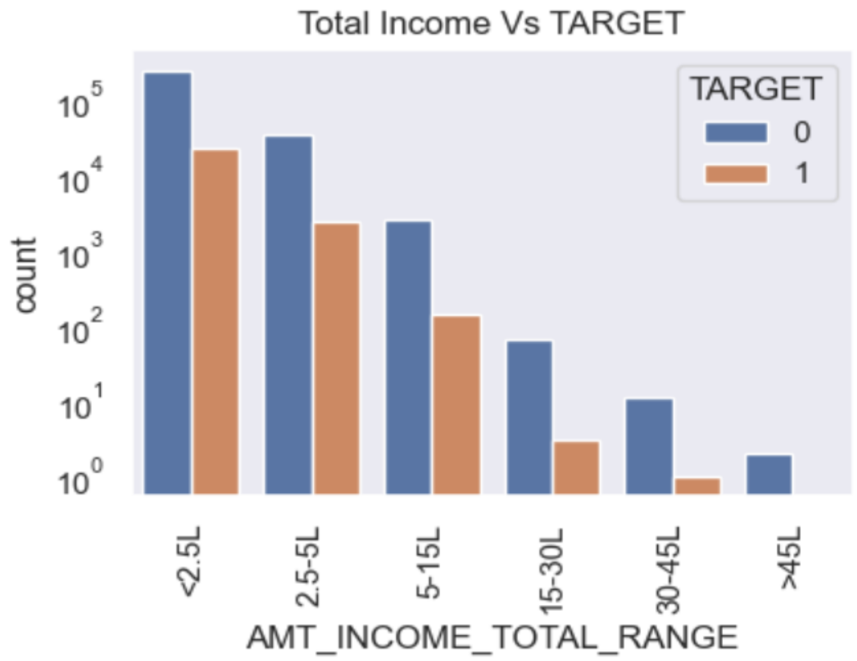
OBSERVATION

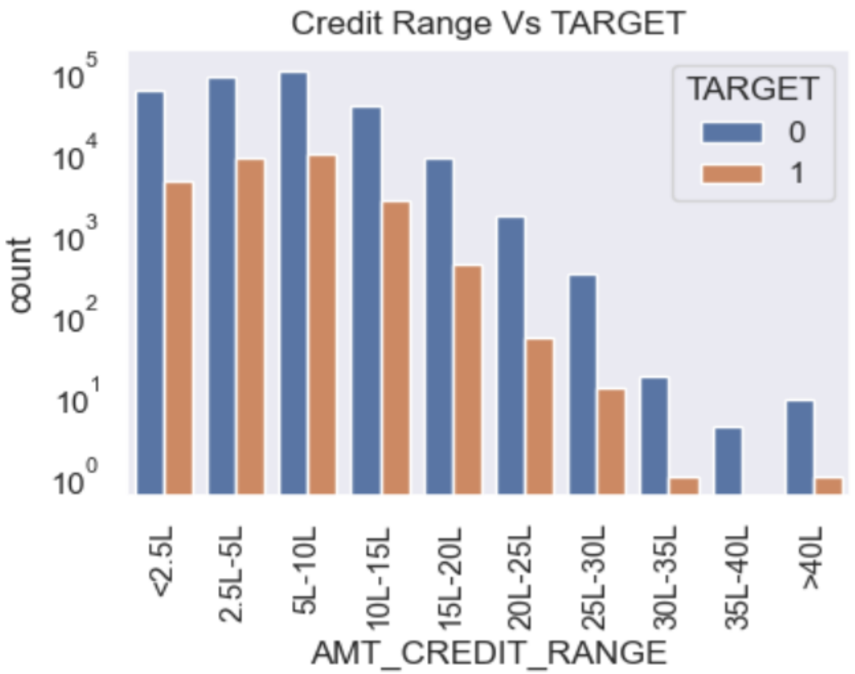
More applicants are filed between 10 am- 2:30pm.



OBSEVATIONS

Above graph shows that the family having 2-3 children are taking more loans.





OBSERVATION

AMT_INCOME_TOTAL_RANGE

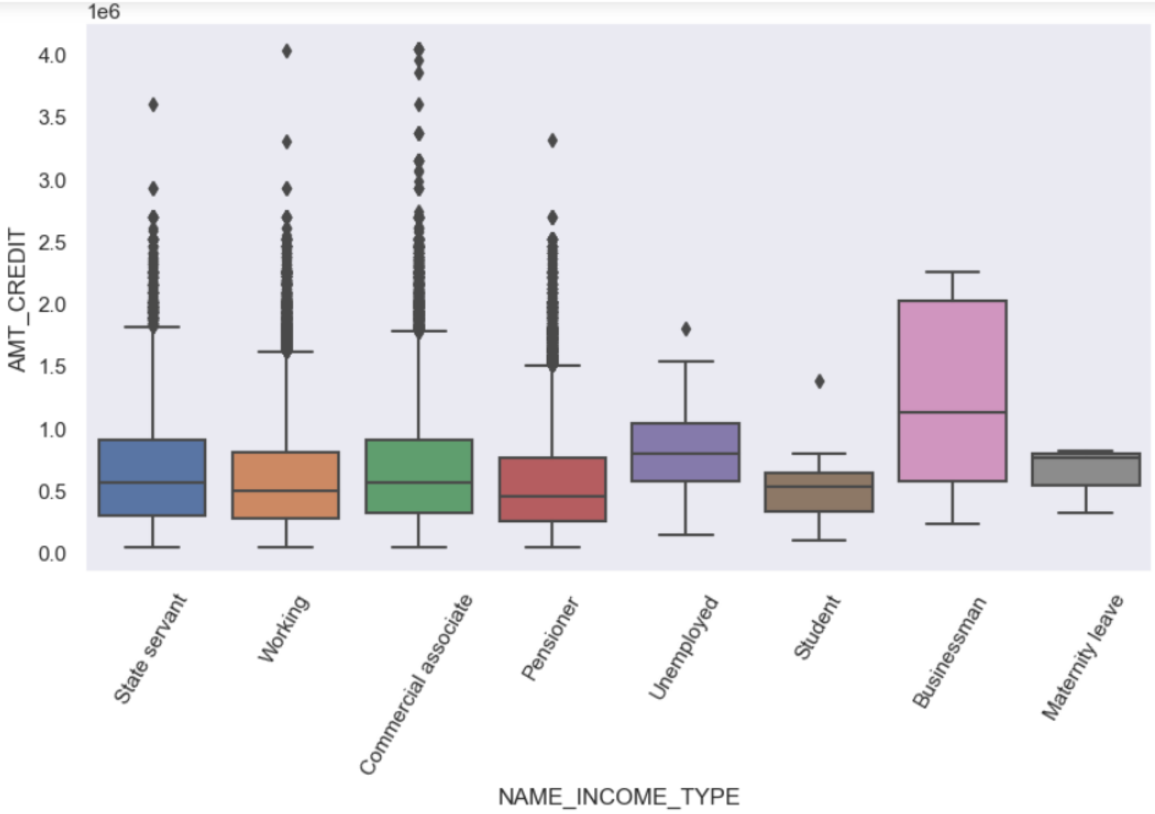
a)Less number of defaulters are present when income is more than 15L and no defaulter whose income is more than 45L.

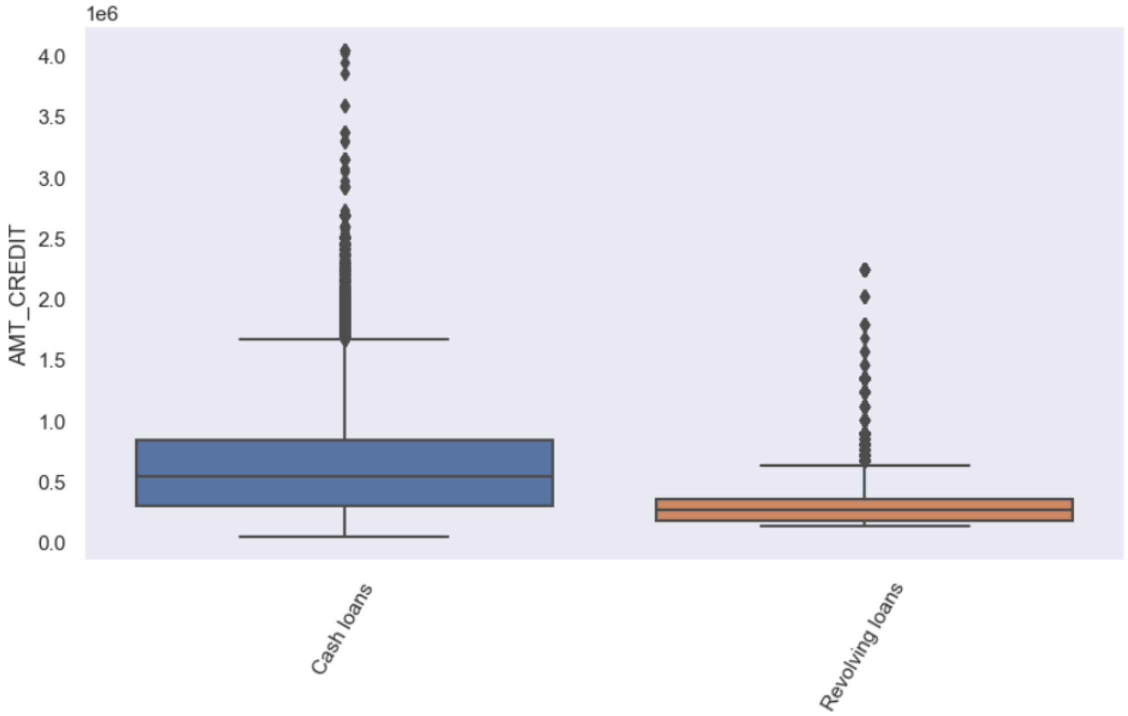
b)More defaulter can be seen at lower side of income.

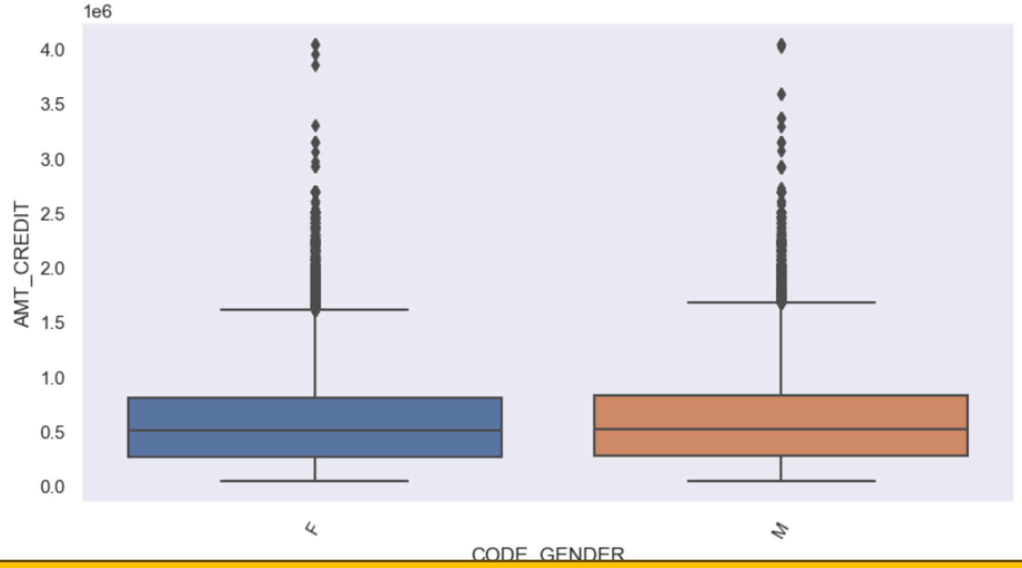
AMT_CREDIT_RANGE

a)More number of defaulters are present whose credit range is lower. as the credit range increases the amount of defaulters decreases.

BIVARIATE - Categorical columns VS Numerical -target 0





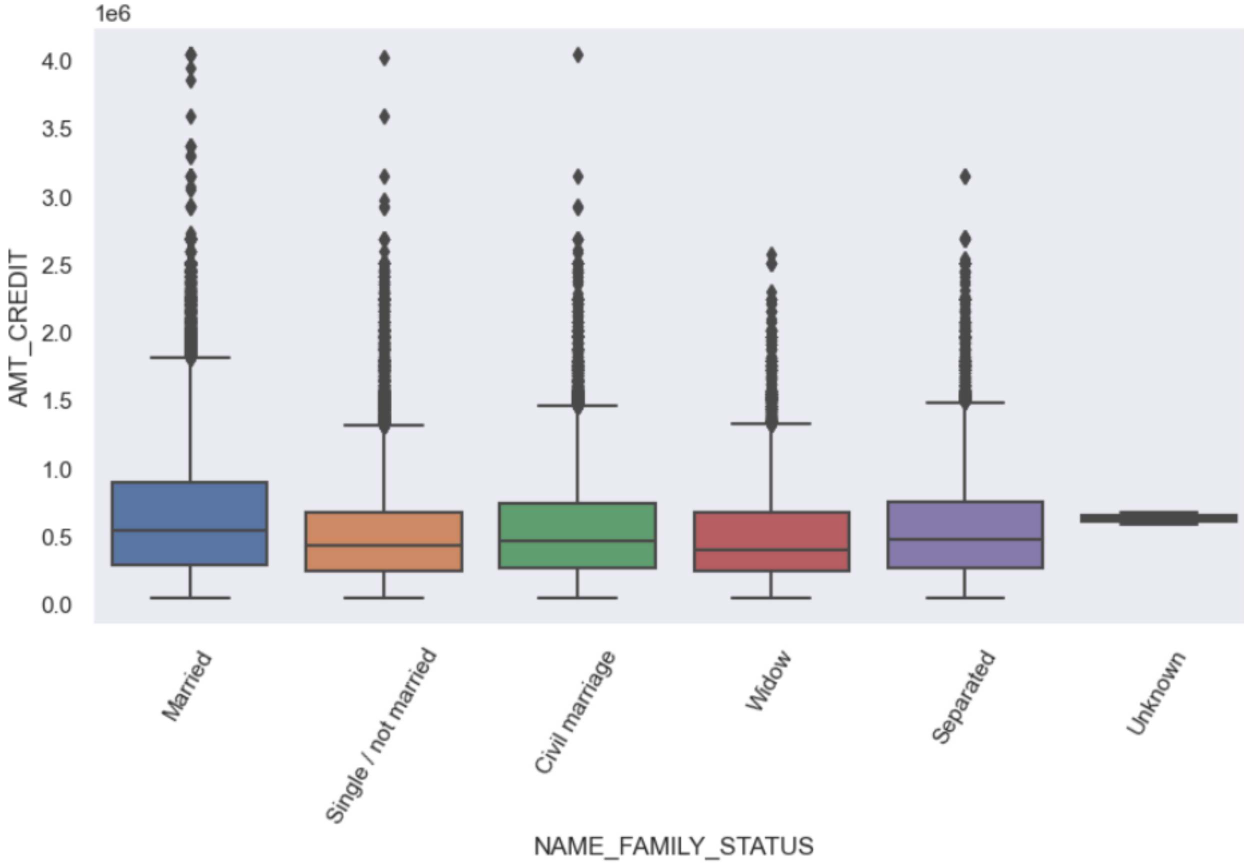


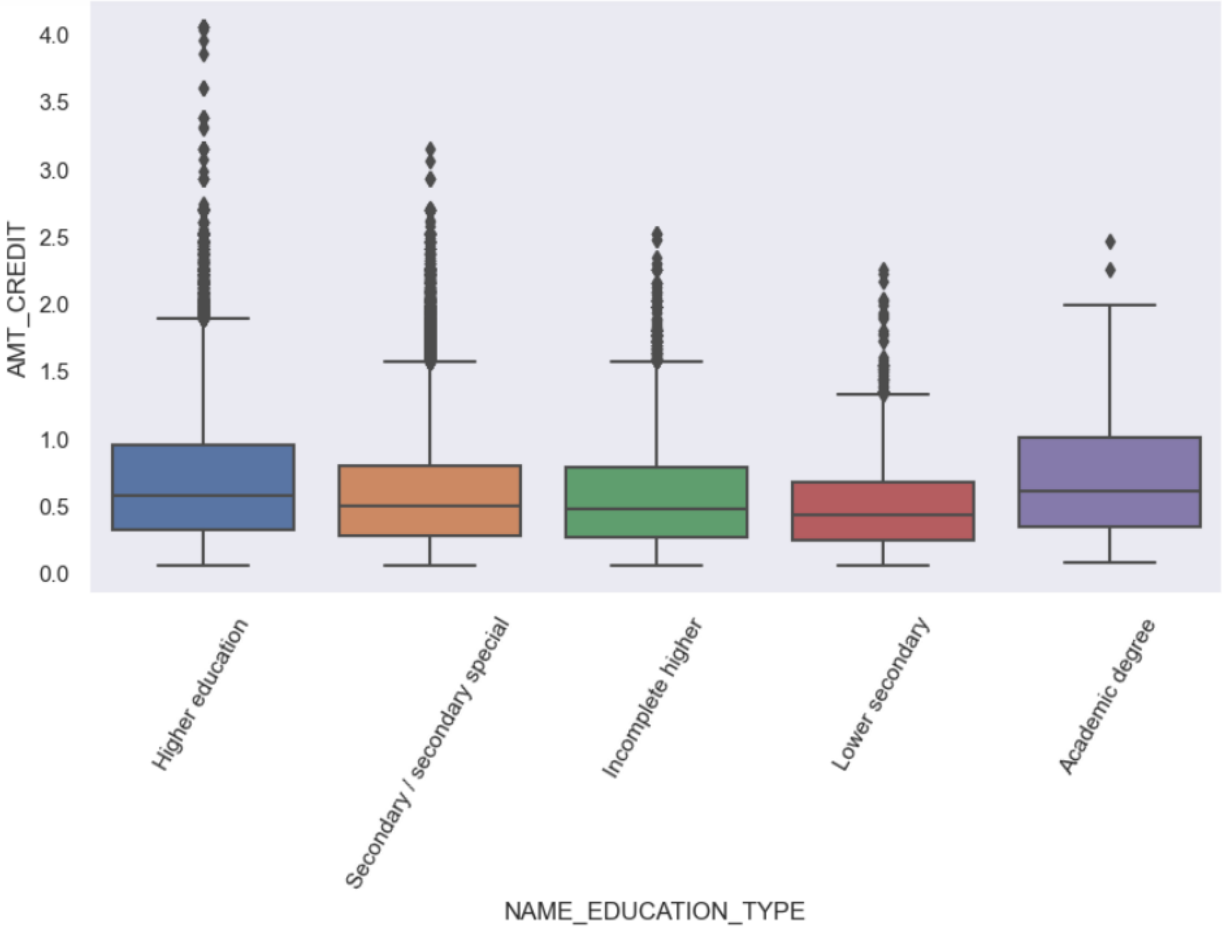
OBSERVATIONS

1.Credit amount of loans are very low for revolving loan.

2.No such difference in code_gender.

3.Businessmen have more amount of credit.





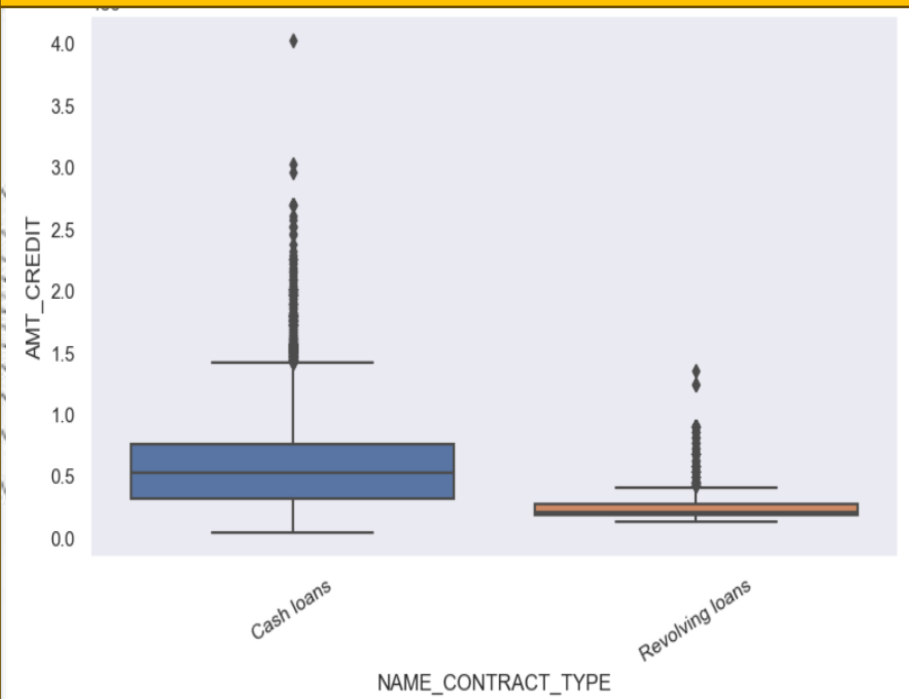
OBSERVATIONS

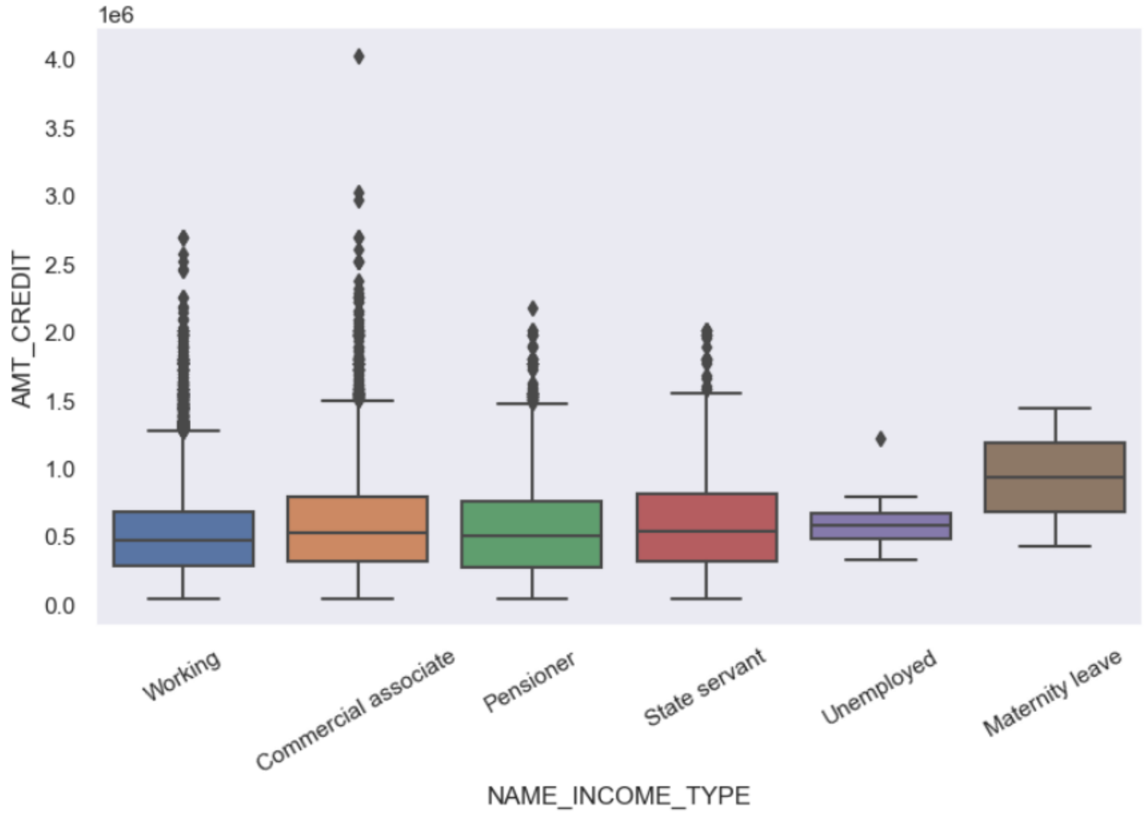
Marriage, civil marriage and separated have higher amount credit.

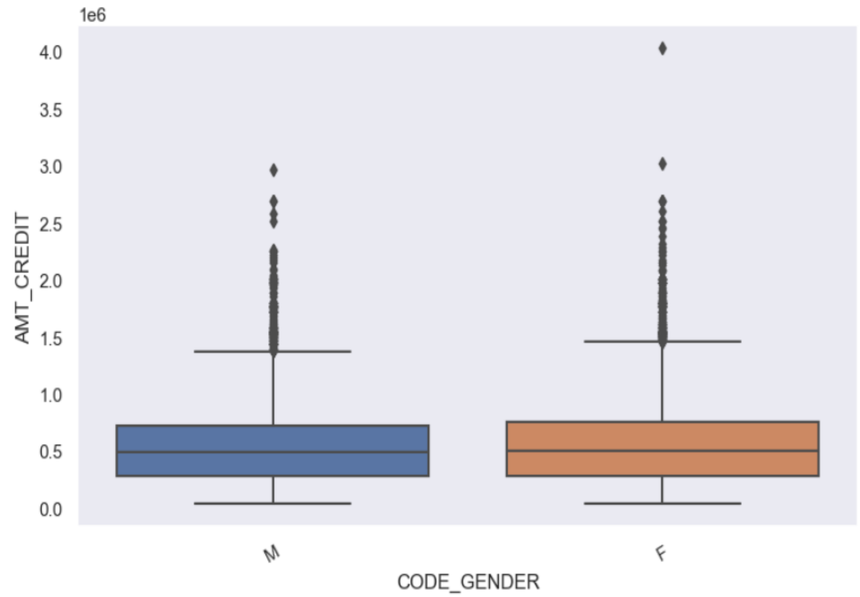
OBSERVATIONS

Academic and higher degree have more amount credit.

BIVARIATE - Categorical columns VS Numerical –target1

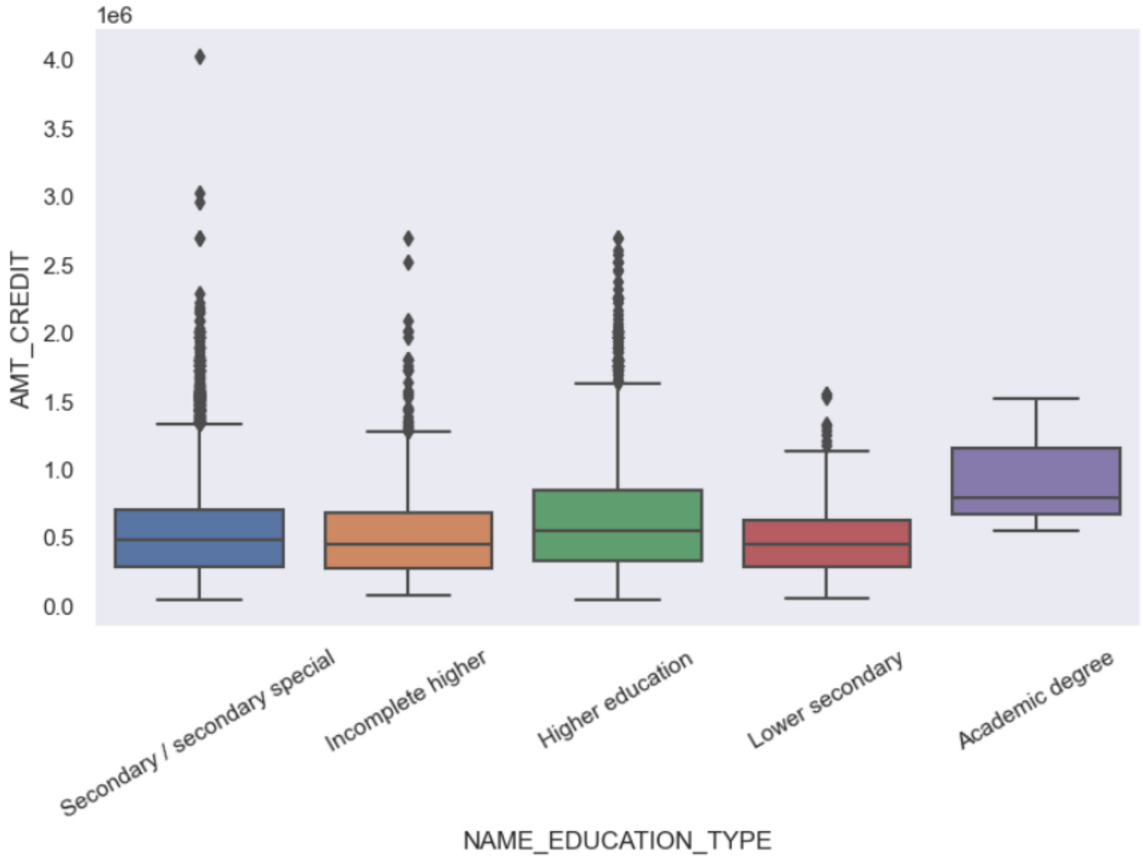






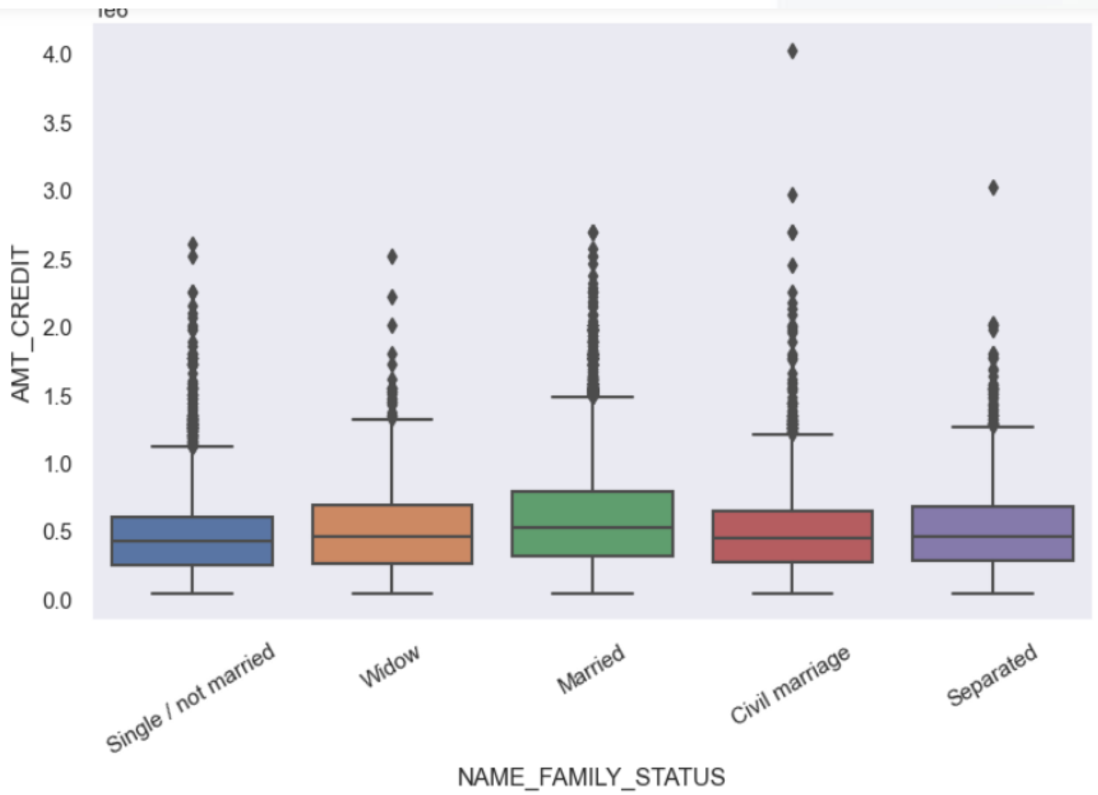
OBSERVATIONS

1. More cash loans are taken by defaulters.
2. More number of outliers in female.
3.No defaulter from business class.



OBSERVATION-

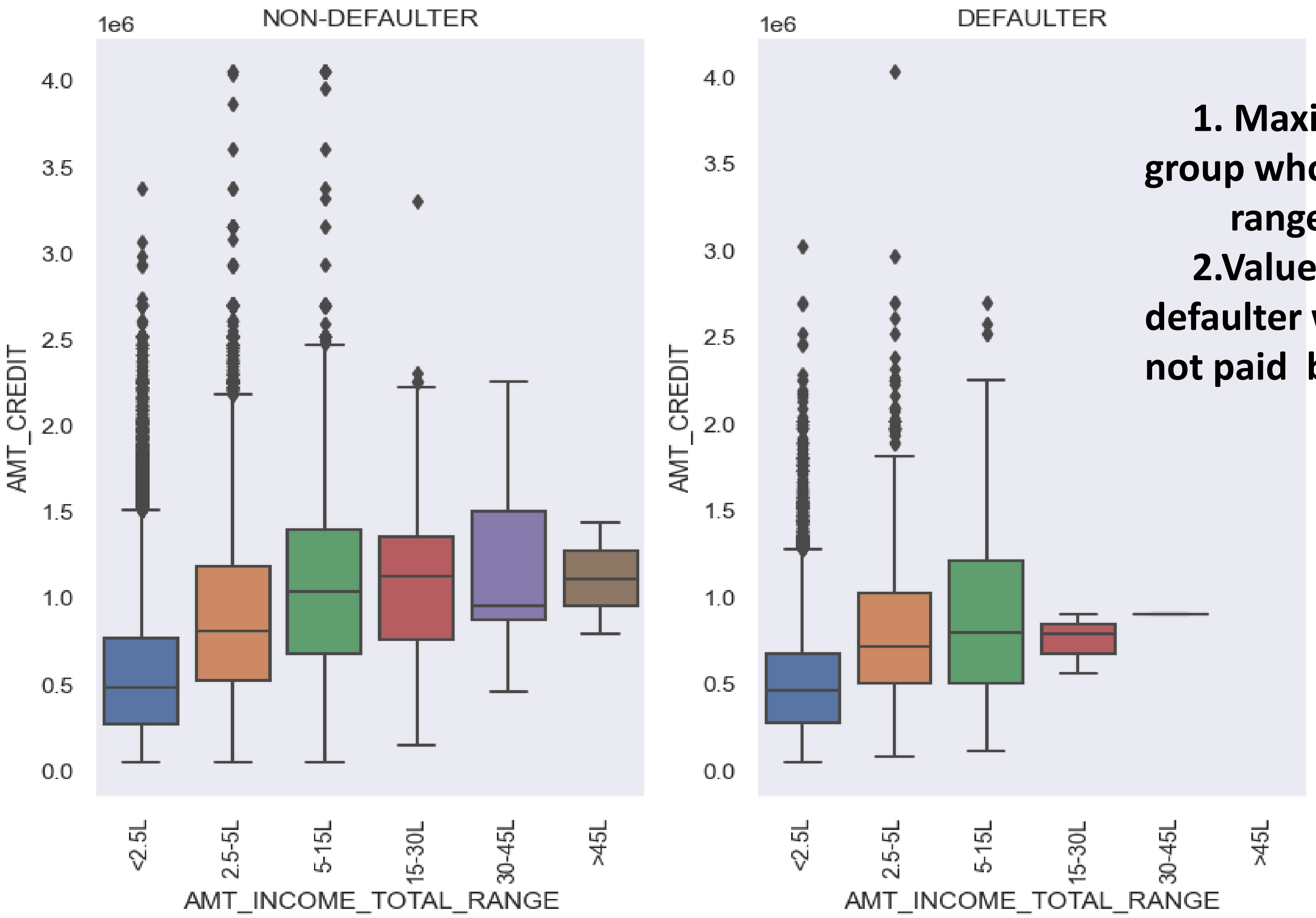
More amount credit for academic degree and then in higher education.



OBSERVATION

Married people tend to have more credit amount.

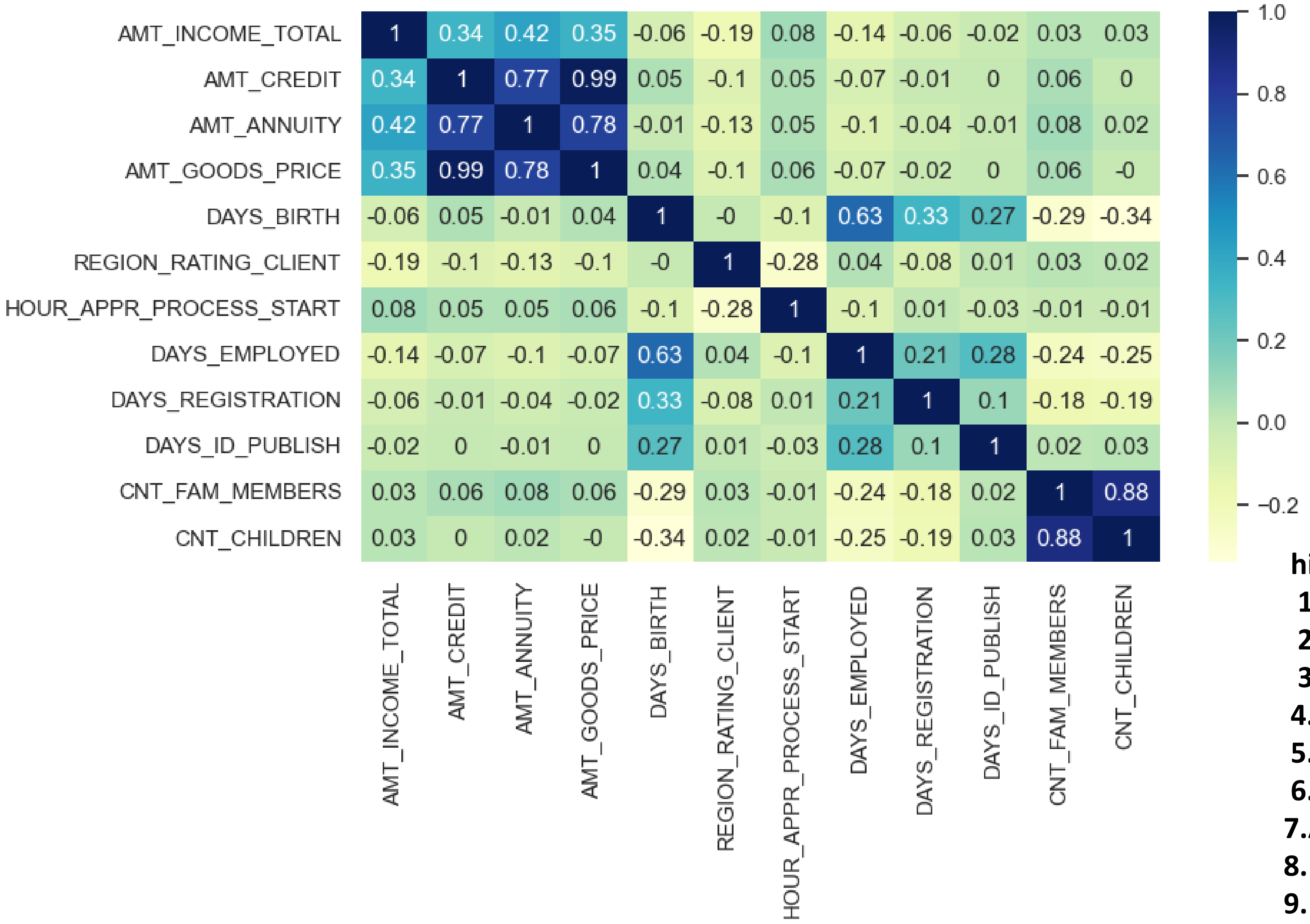
BIVARIATE - Numerical Vs Category



Observations

1. Maximum number of loans are given to the group whose
range is 5-15L.

2.Value of median is for the range 15-30L in case of defaulter which is not good as if the loan amount is not paid back bank can suffer various losses.



	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	REGION_RATING_CLIENT	HOUR_APPR_PROCESS_START	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	CNT_CHILDREN
AMT_INCOME_TOTAL	1	0.34	0.42	0.35	-0.06	-0.19	0.08	-0.14	-0.06	-0.02	0.03	0.03
AMT_CREDIT	0.34	1	0.77	0.99	0.05	-0.1	0.05	-0.07	-0.01	0	0.06	0
AMT_ANNUITY	0.42	0.77	1	0.78	-0.01	-0.13	0.05	-0.1	-0.04	-0.01	0.08	0.02
AMT_GOODS_PRICE	0.35	0.99	0.78	1	0.04	-0.1	0.06	-0.07	-0.02	0	0.06	-0
DAYS_BIRTH	-0.06	0.05	-0.01	0.04	1	-0	-0.1	0.63	0.33	0.27	-0.29	-0.34
REGION_RATING_CLIENT	-0.19	-0.1	-0.13	-0.1	-0	1	-0.28	0.04	-0.08	0.01	0.03	0.02
HOUR_APPR_PROCESS_START	0.08	0.05	0.05	0.06	-0.1	-0.28	1	-0.1	0.01	-0.03	-0.01	-0.01
DAYS_EMPLOYED	-0.14	-0.07	-0.1	-0.07	0.63	0.04	-0.1	1	0.21	0.28	-0.24	-0.25
DAYS_REGISTRATION	-0.06	-0.01	-0.04	-0.02	0.33	-0.08	0.01	0.21	1	0.1	-0.18	-0.19
DAYS_ID_PUBLISH	-0.02	0	-0.01	0	0.27	0.01	-0.03	0.28	0.1	1	0.02	0.03
CNT_FAM_MEMBERS	0.03	0.06	0.08	0.06	-0.29	0.03	-0.01	-0.24	-0.18	0.02	1	0.88
CNT_CHILDREN	0.03	0	0.02	-0	-0.34	0.02	-0.01	-0.25	-0.19	0.03	0.88	1

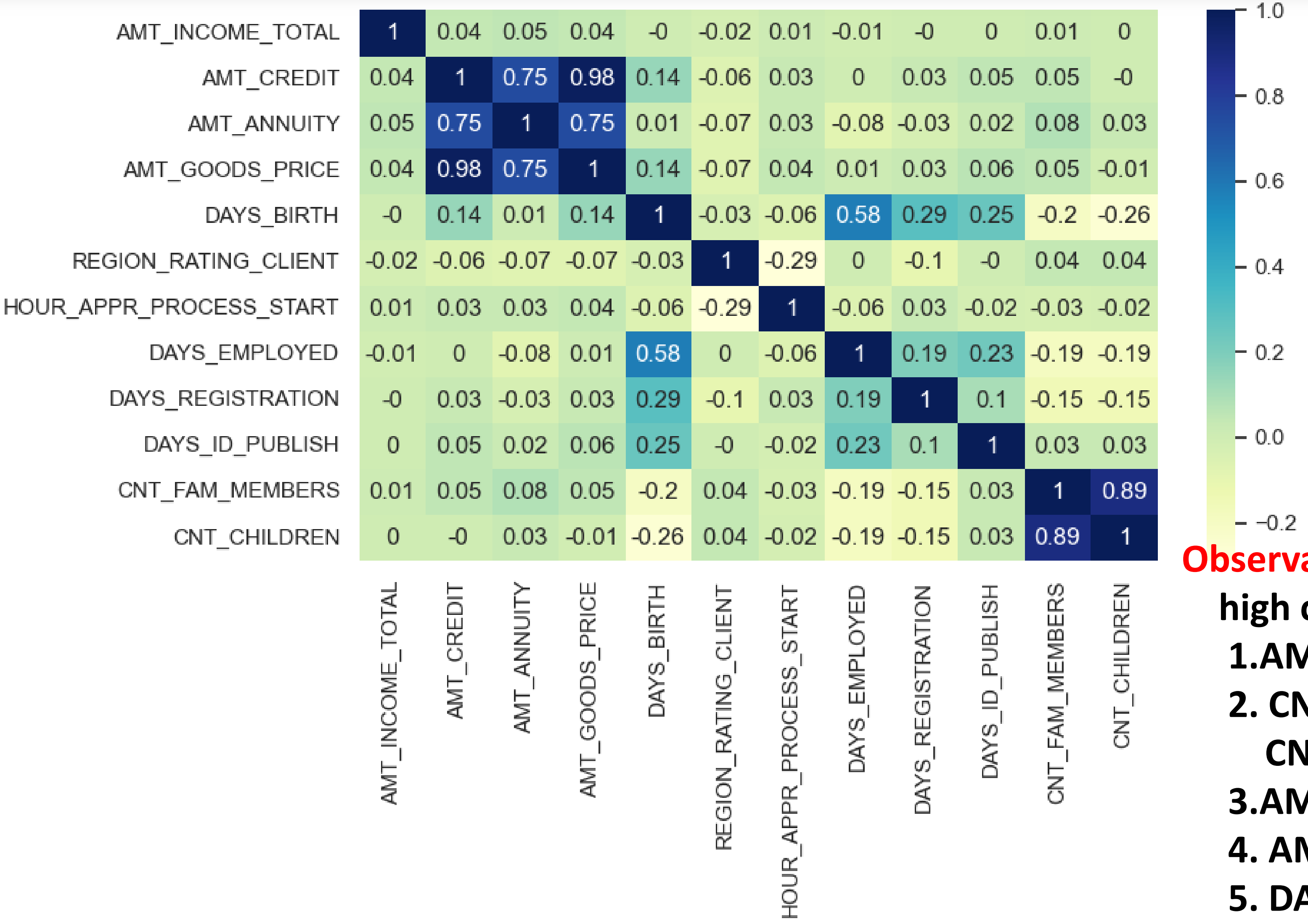
HEATMAP for Non-Defaulter

Observations

high correlation can be seen in

1. AMT_CREDIT and AMT_GOODS_PRICE =0.99
2. CNT_FAM_MEMBERS and CNT_CHILDREN =0.88
3.AMT_ANNUITY and AMT_GOODS_PRICE=0.78
4.AMT_CREDIT and AMT_ANNUITY =0.77
5. DAYS_BIRTH and DAYS_EMPLOYED =0.63
6.AMT_ANNUITY and AMT_INCOME_TOTAL =0.42
7.AMT_GOODS_PRICE and AMT_INCOME_TOTAL =0.35
8. AMTT_TOTAL_INCOME and AMT_CREDIT=0.34
9. DAYS_REGISTRATION and DAYS_BIRTH =0.33
10. DAYS_BIRTH and DAYS_ID_PUBLISH =0.27

HEATMAP for Defaulter



	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	REGION_RATING_CLIENT	HOUR_APPR_PROCESS_START	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	CNT_CHILDREN
AMT_INCOME_TOTAL	1	0.04	0.05	0.04	-0	-0.02	0.01	-0.01	-0	0	0.01	0
AMT_CREDIT	0.04	1	0.75	0.98	0.14	-0.06	0.03	0	0.03	0.05	0.05	-0
AMT_ANNUITY	0.05	0.75	1	0.75	0.01	-0.07	0.03	-0.08	-0.03	0.02	0.08	0.03
AMT_GOODS_PRICE	0.04	0.98	0.75	1	0.14	-0.07	0.04	0.01	0.03	0.06	0.05	-0.01
DAYS_BIRTH	-0	0.14	0.01	0.14	1	-0.03	-0.06	0.58	0.29	0.25	-0.2	-0.26
REGION_RATING_CLIENT	-0.02	-0.06	-0.07	-0.07	-0.03	1	-0.29	0	-0.1	-0	0.04	0.04
HOUR_APPR_PROCESS_START	0.01	0.03	0.03	0.04	-0.06	-0.29	1	-0.06	0.03	-0.02	-0.03	-0.02
DAYS_EMPLOYED	-0.01	0	-0.08	0.01	0.58	0	-0.06	1	0.19	0.23	-0.19	-0.19
DAYS_REGISTRATION	-0	0.03	-0.03	0.03	0.29	-0.1	0.03	0.19	1	0.1	-0.15	-0.15
DAYS_ID_PUBLISH	0	0.05	0.02	0.06	0.25	-0	-0.02	0.23	0.1	1	0.03	0.03
CNT_FAM_MEMBERS	0.01	0.05	0.08	0.05	-0.2	0.04	-0.03	-0.19	-0.15	0.03	1	0.89
CNT_CHILDREN	0	-0	0.03	-0.01	-0.26	0.04	-0.02	-0.19	-0.15	0.03	0.89	1

Observations ----- DEFAULTERS ------- TARGET 1

high correlation can be seen in

1.AMT_CREDIT and AMT_GOODS_PRICE =0.98

2. CNT_FAM_MEMBERS and CNT_CHILDREN=0.89

3.AMT_CREDIT and AMT_ANNUITY =0.75

4. AMT_ANNUITY and AMT_GOODS_PRICE =0.75

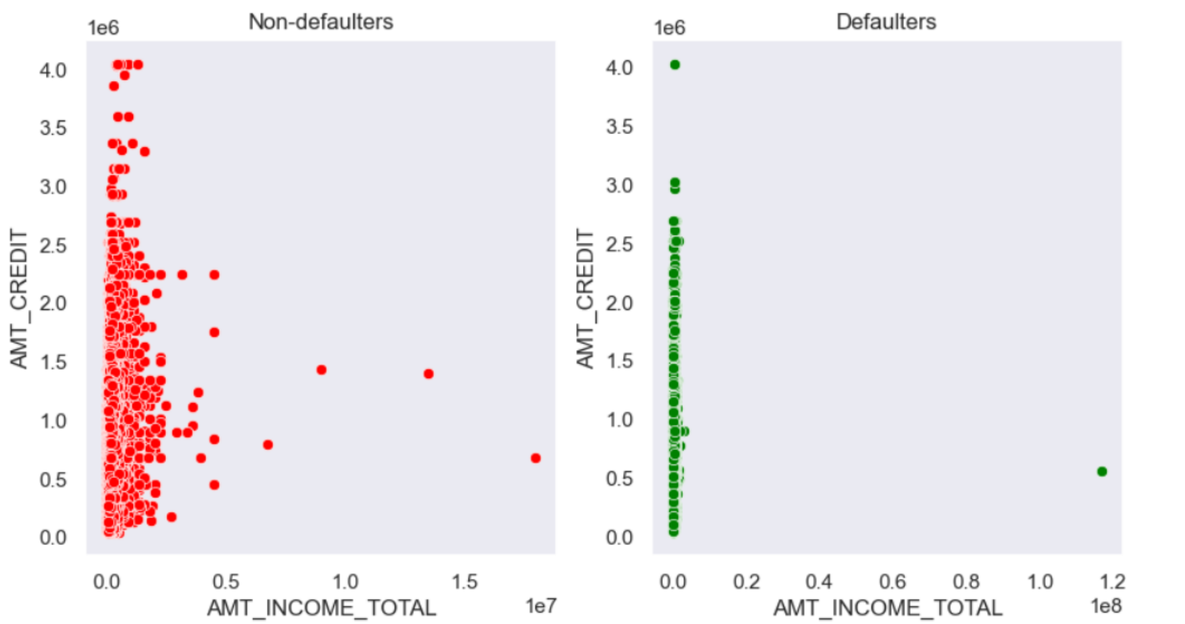
5. DAYS_BIRTH and DAYS_EMPLOYED =0.58

6. DAYS_REGISTRATION and DAYS_BIRTH =0.29

7. DAYS_BIRTH and DAYS_ID_PUBLISH =0.25

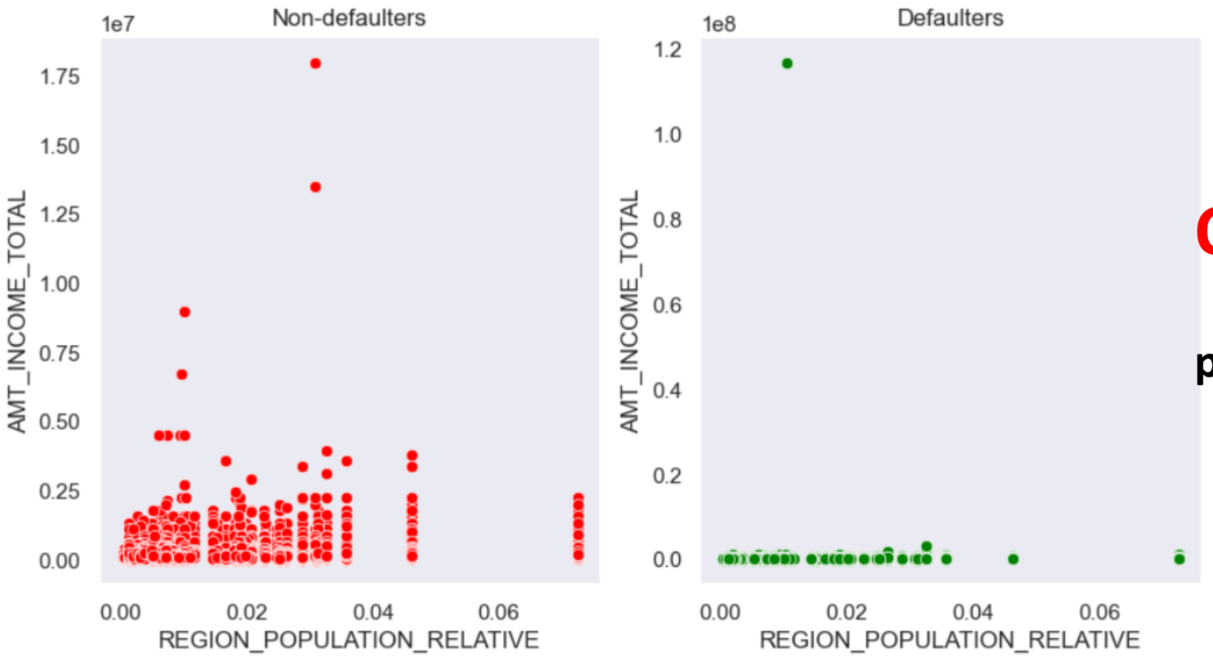
8.DAYS_ID_PUBLISH and DAYS_EMPLOYED =0.23

Bivariate analysis on continuous columns



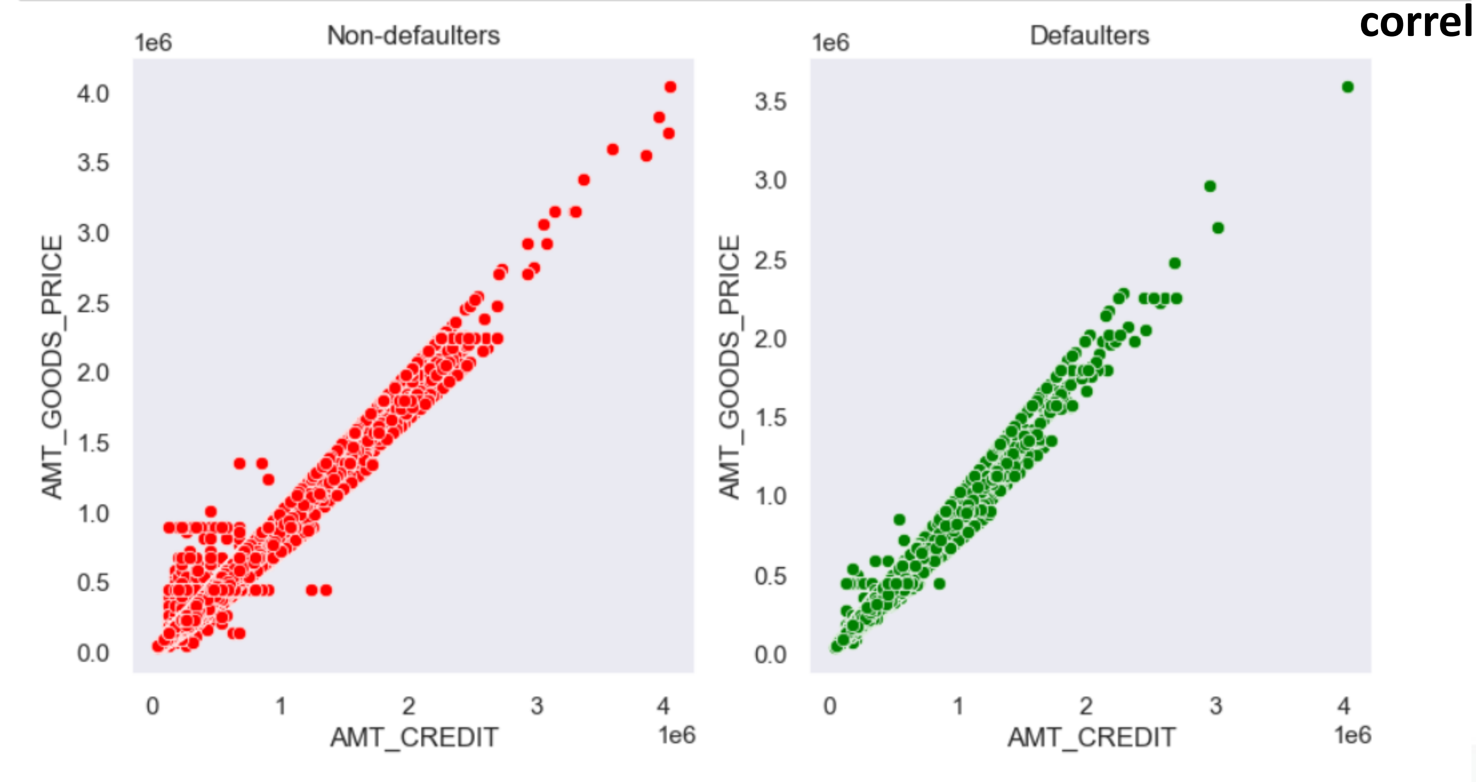
OBSERVATIONS

In case of defaulters we can see the values are more concentrated towards the lower side of income and credit amount.



OBSERVATIONS

Most of the defaulters have very low income where region population is less dense.

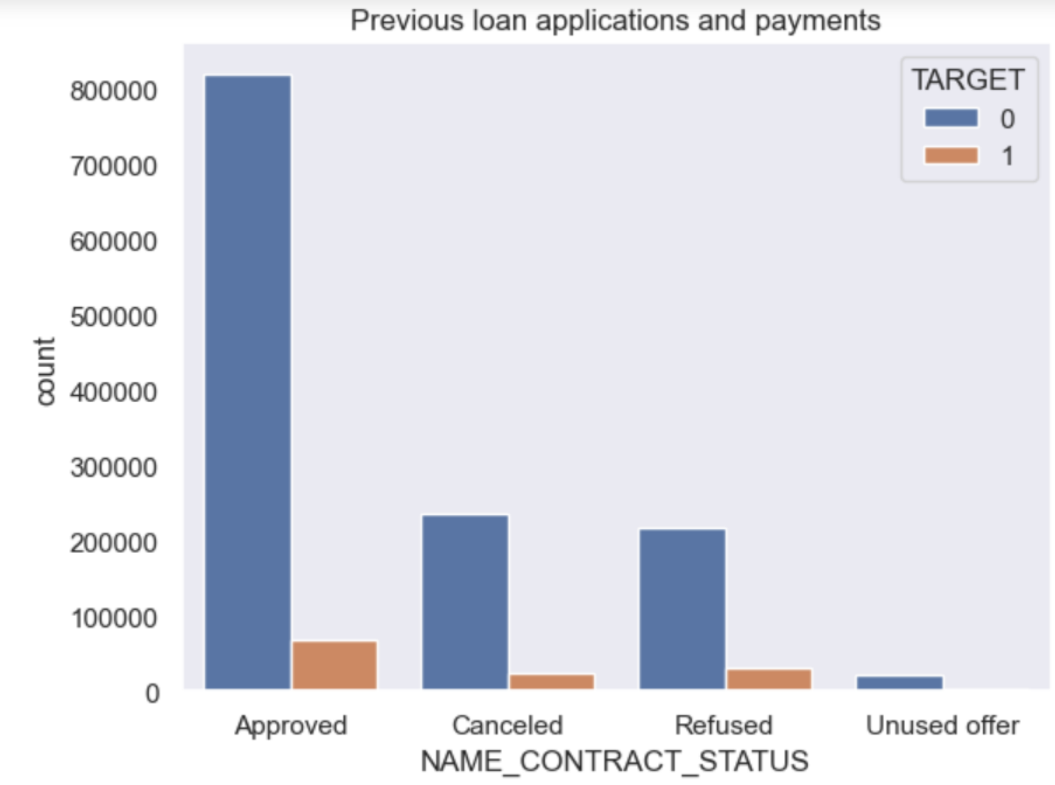


Observation

From the heat map above we can see positive high correlation between AMT_CREDIT,AMT_GOODS_PRICE

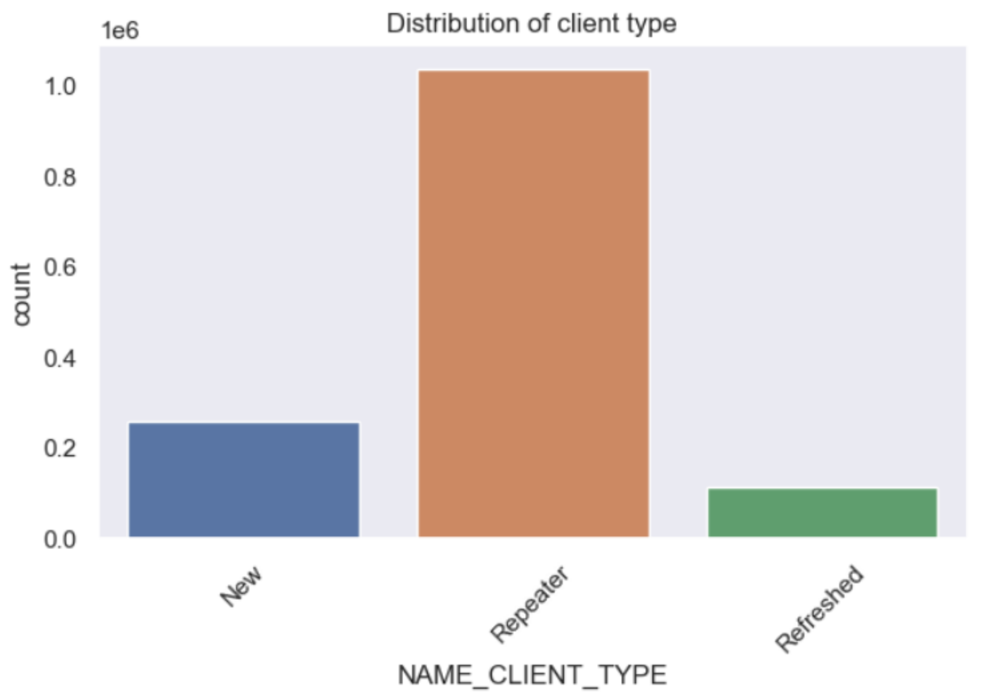
from the scatter plot also we can see high correlation.

MERGING application and previous data



Imbalance on merging data

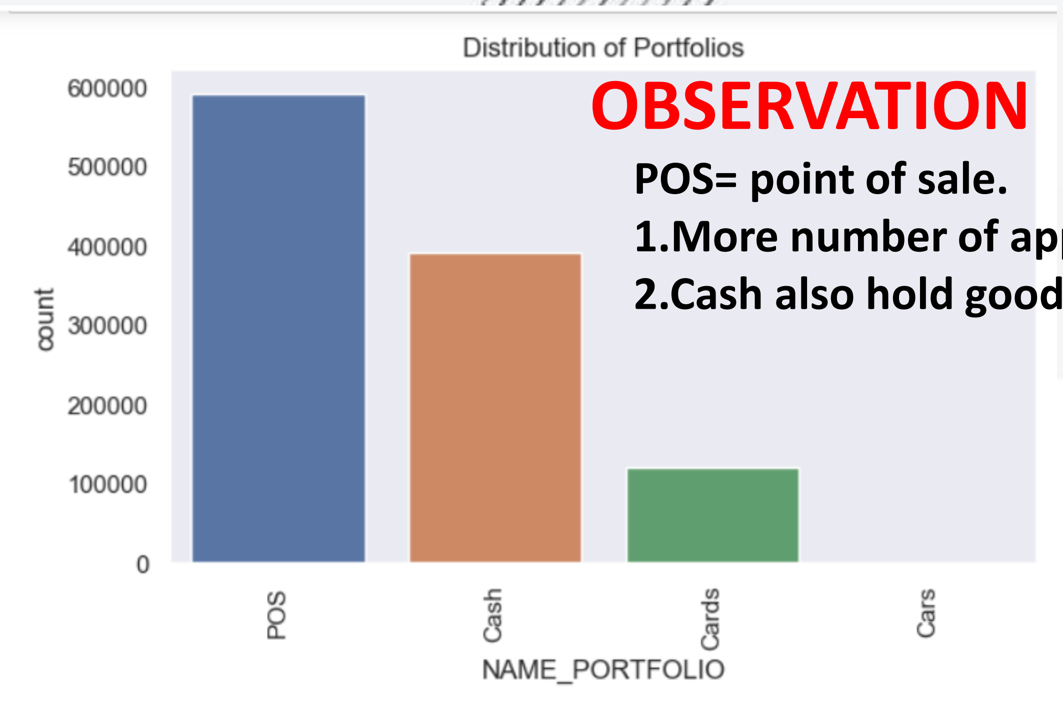
-Here we can see imbalance as we can see approved count for non defaulter is more as compared to defaulter.



OBSERVATIONS

1.Mostly the clients were Repeater.

2.New clients are very less as compared to repeaters.



OBSERVATION

POS= point of sale.

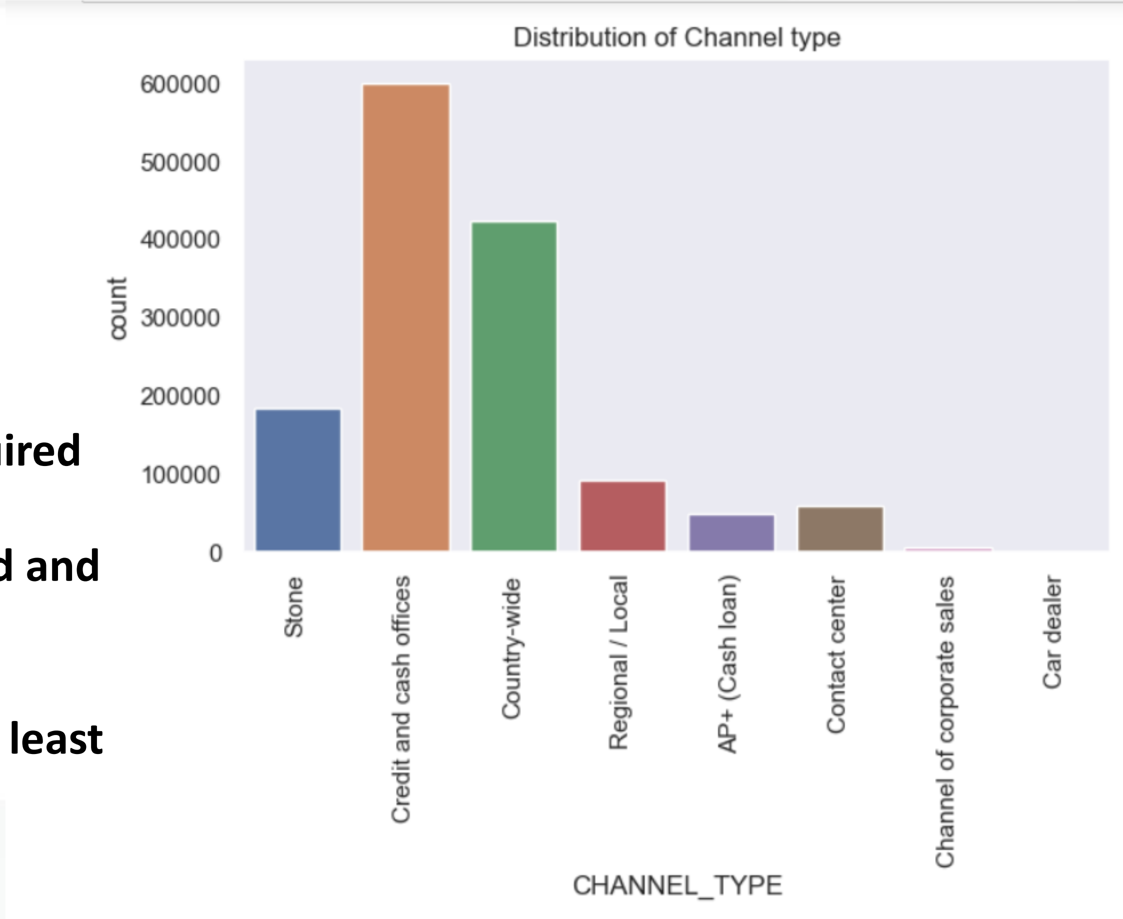
1.More number of application in previous data is for POS.

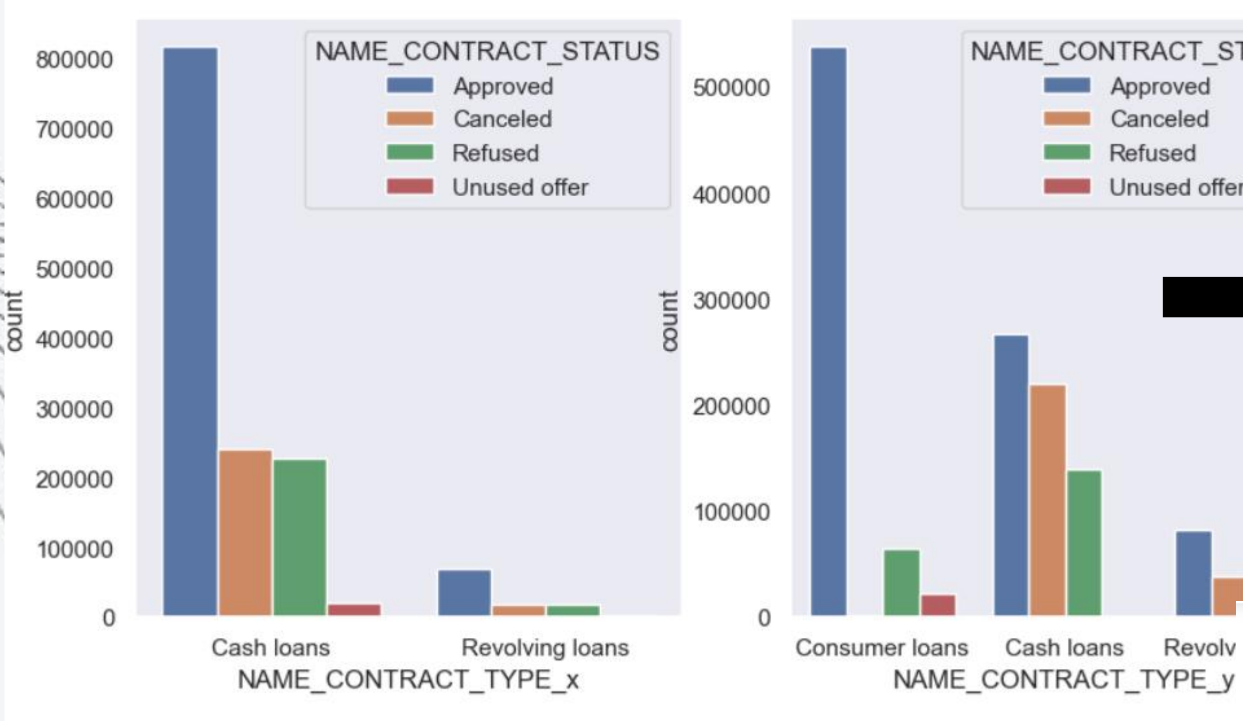
2.Cash also hold good percentage

Observations

Channel_type shows Through which channel we acquired the client on the previous application

1. It is clear from the plot given above that Credit card and cash offices has mostly used
   after this count of countrywide is more .
2. Car dealer ,channel of cooperate sales and AP+ was least used.





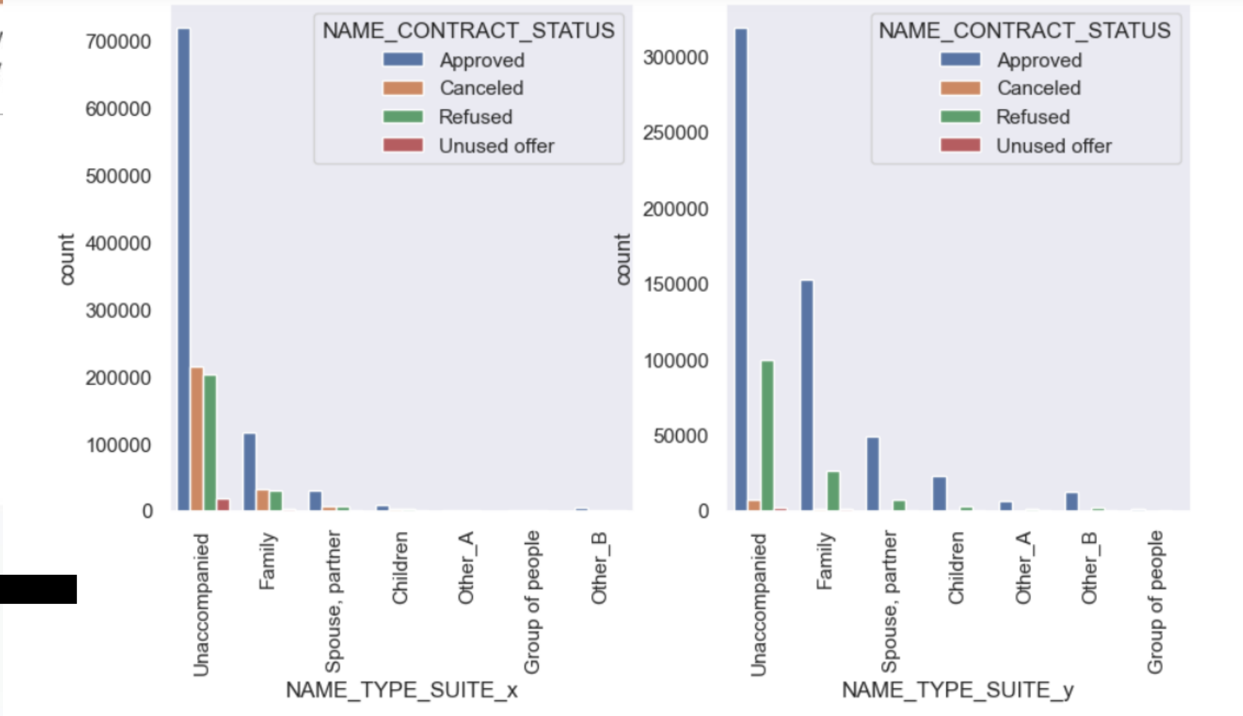
OBSERVATIONS

1.In application data 2 loans were given but previously consumer loan was also given.

2.Number of cash loan approved is highest in application set i.e, the current one and in previous data consumer loans has the maximum value.

NAME_CONTRACT_TYPE_x= data from application data

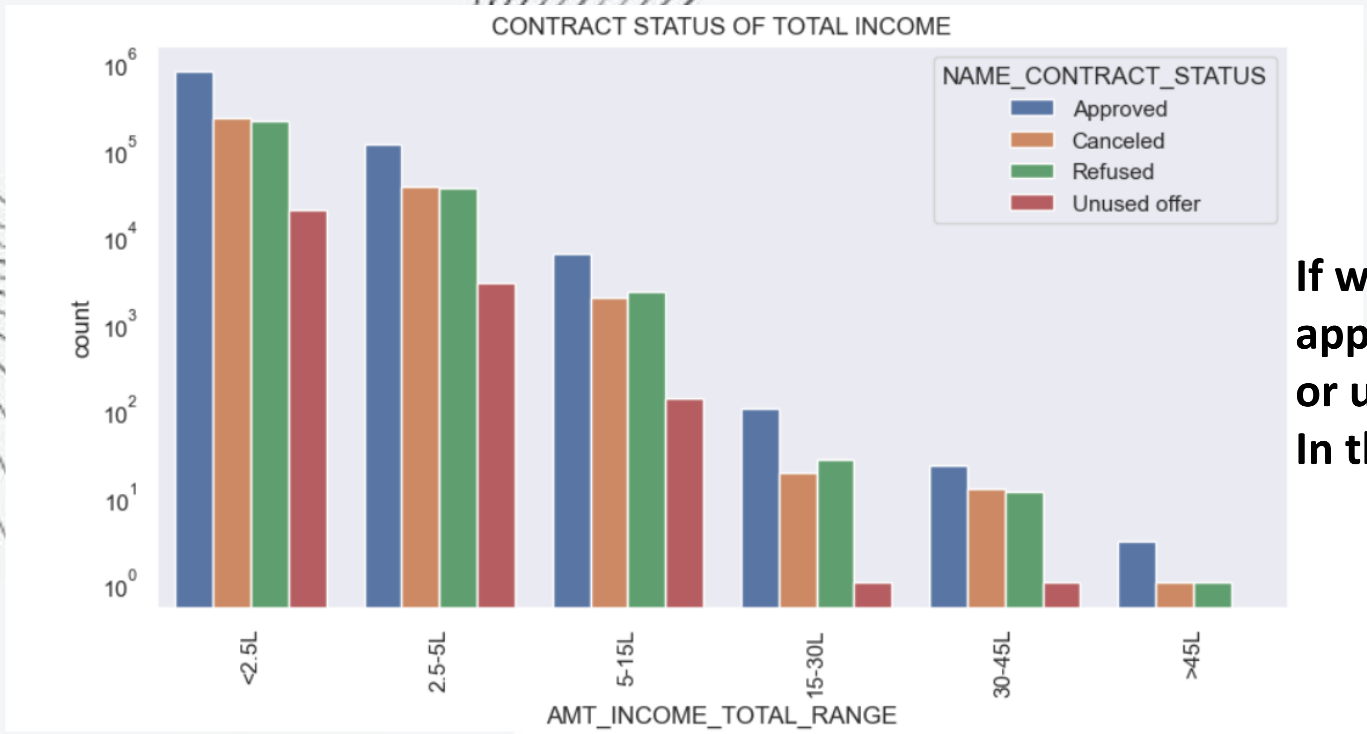
NAME_CONTRACT_TYPE_y= data from previous data



OBSERVATION

In both previous and current data(application_data) unaccompanied one loans were approved. they are maximum in number.

Even who are living with family are applying more for the loan and loan approval ratio is high.

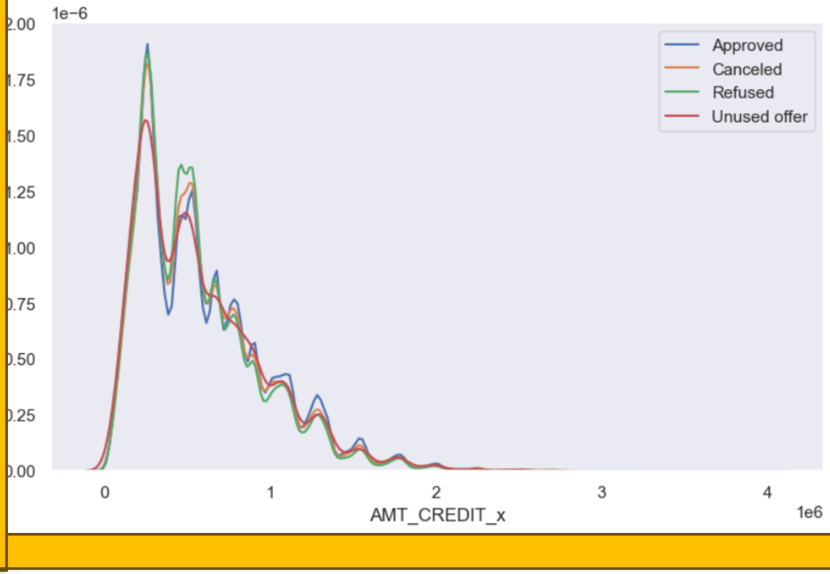


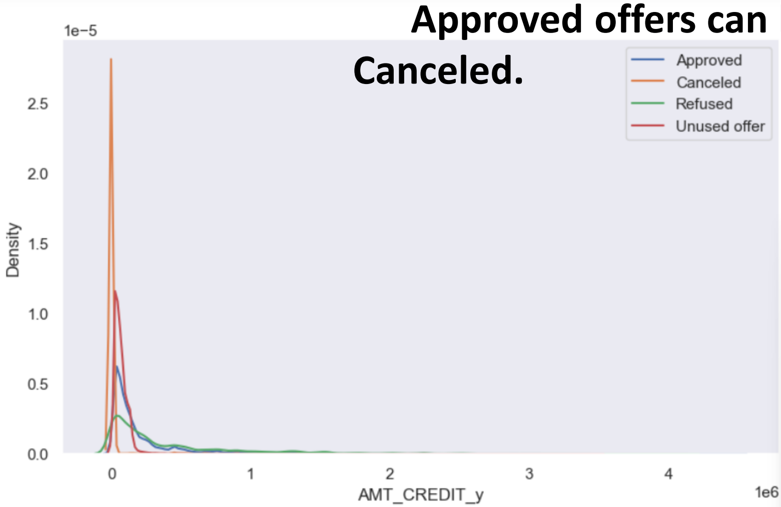
OBSERVATION

If we compare ratios then good amount of loan was approved for 15-30L range as number of canceled, refused or unused is less as compared to approved.

In the case above 45L no unused offer is there.

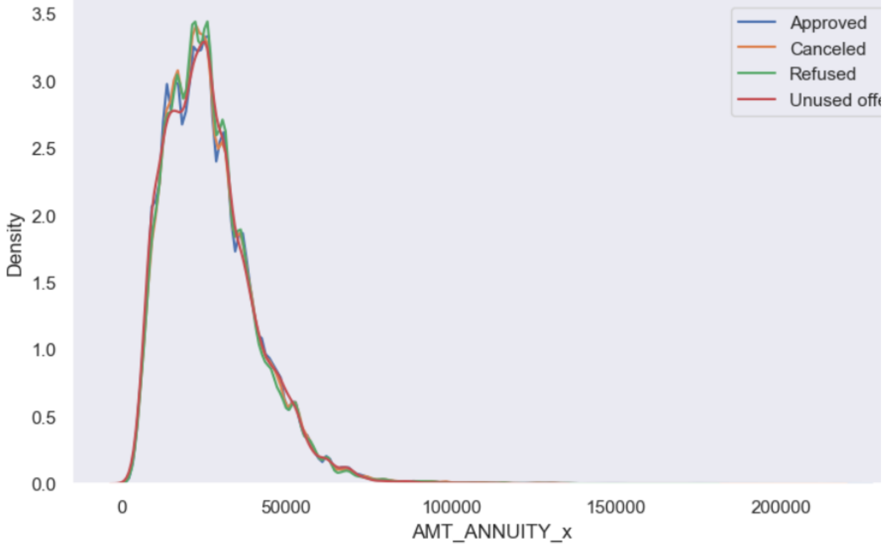
Univariate for Numerical columns

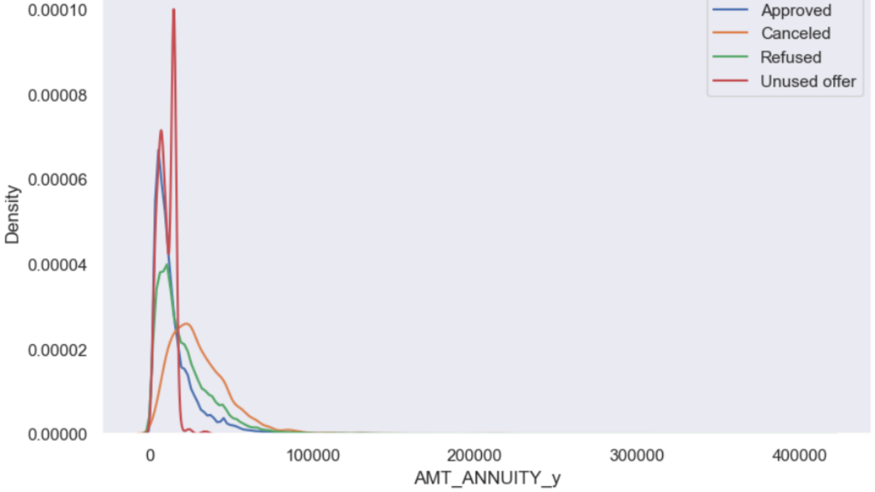


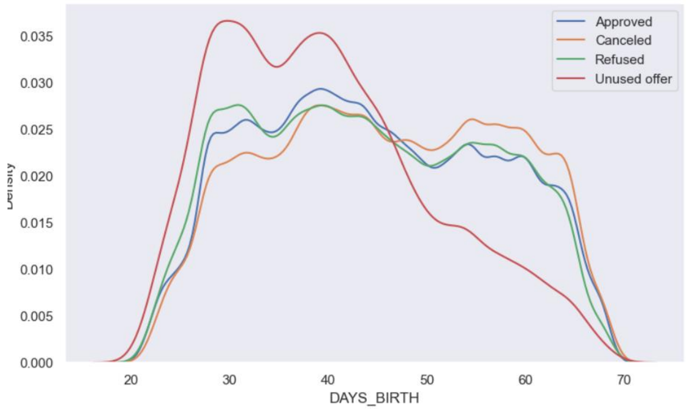


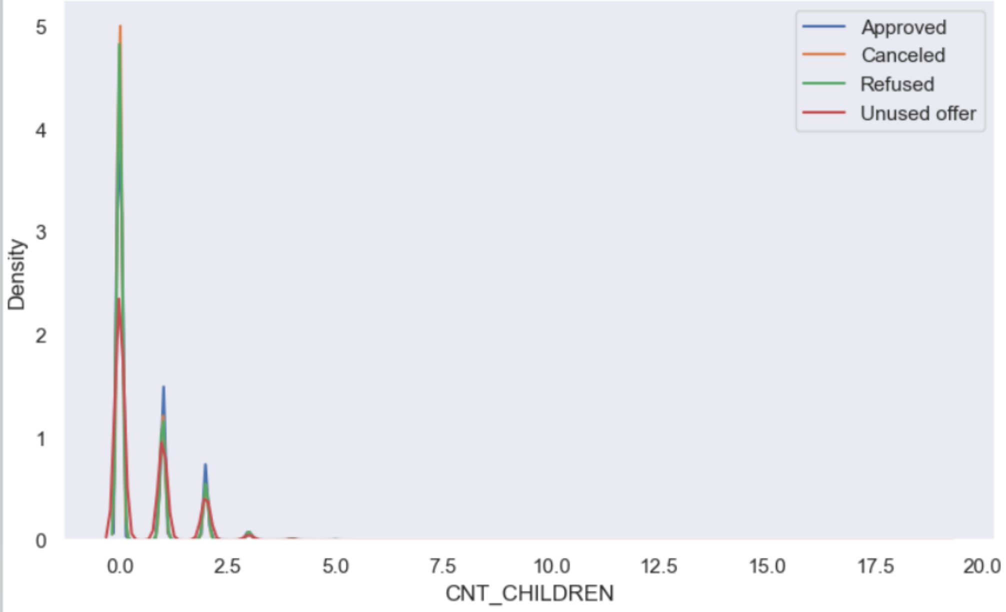
OBSERVATIONS

1. AMT_CREDIT_x represents amount credit of Application data AMT_CREDIT_y represents previous data. Currently high number of Approved offers can be seen whereas in previous data we can see Canceled.









Univariate for Numerical columns

OBSERVATIONS

1.. AMT_ANNUITY_x represents amount annuity of Application data and AMT_ANNUITY_y represents previous data.Previously the bank had high unused offers and in the current data high unused offers for AMT_ANNUITY ie,both are similar.

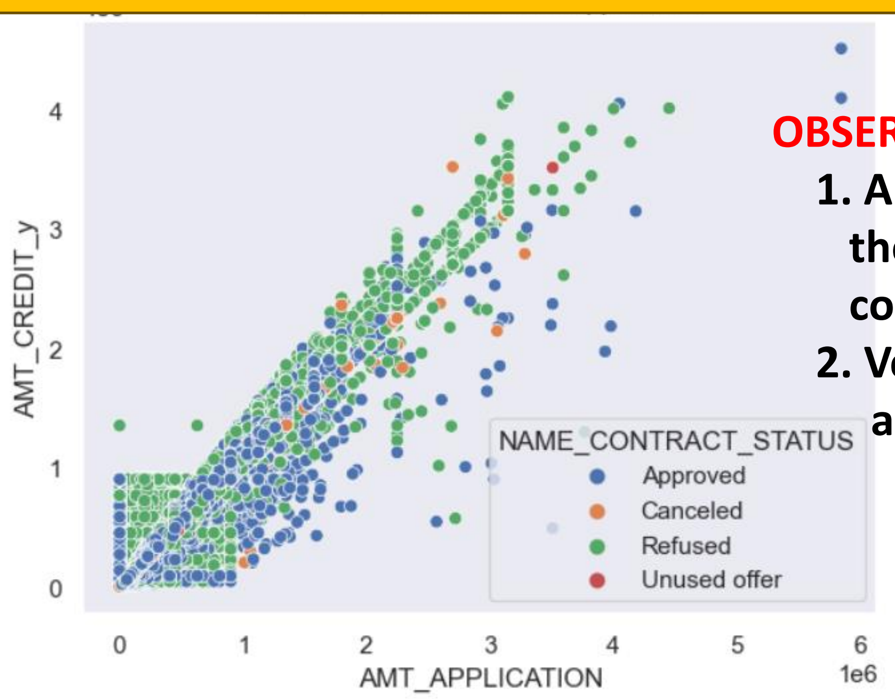
2. DAYS_BIRTH represents age in years high amount of unused offers can be seen
in age group 20-40 years, canceled offers are more towards above 60,approved loans has high density around 30-45,refused offers have high density in range 25-35 years.

4.Nuclear family tends to make more loans.

Bivariate Analysis-numerical columns

OBSERVATIONS

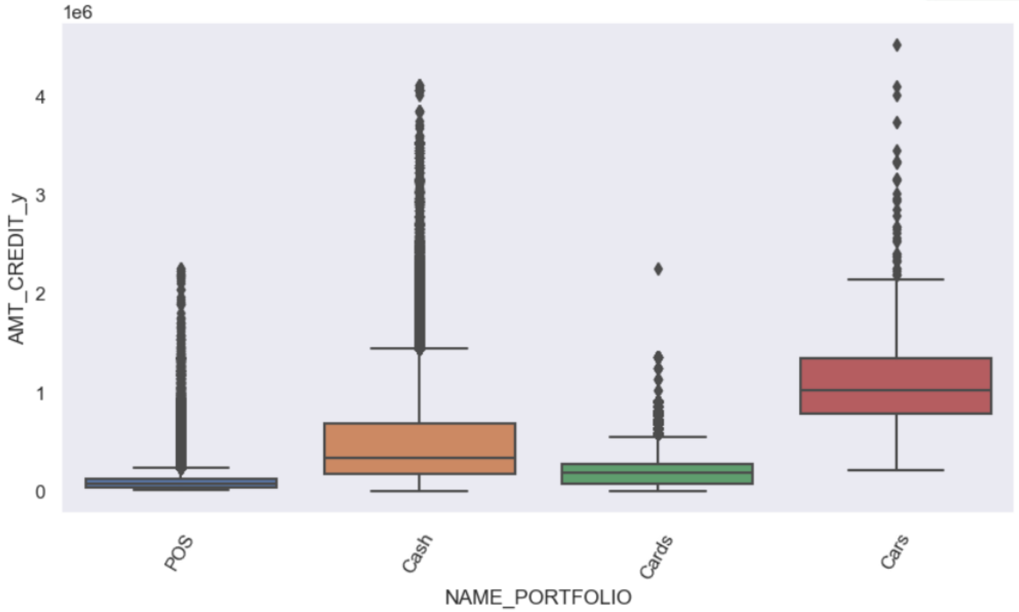
1. AMT_APPLICATION is For how much credit did client ask on the previous application we can see that points are concentrated towards the lower range.
2. Very less applications were approved having high amount application.

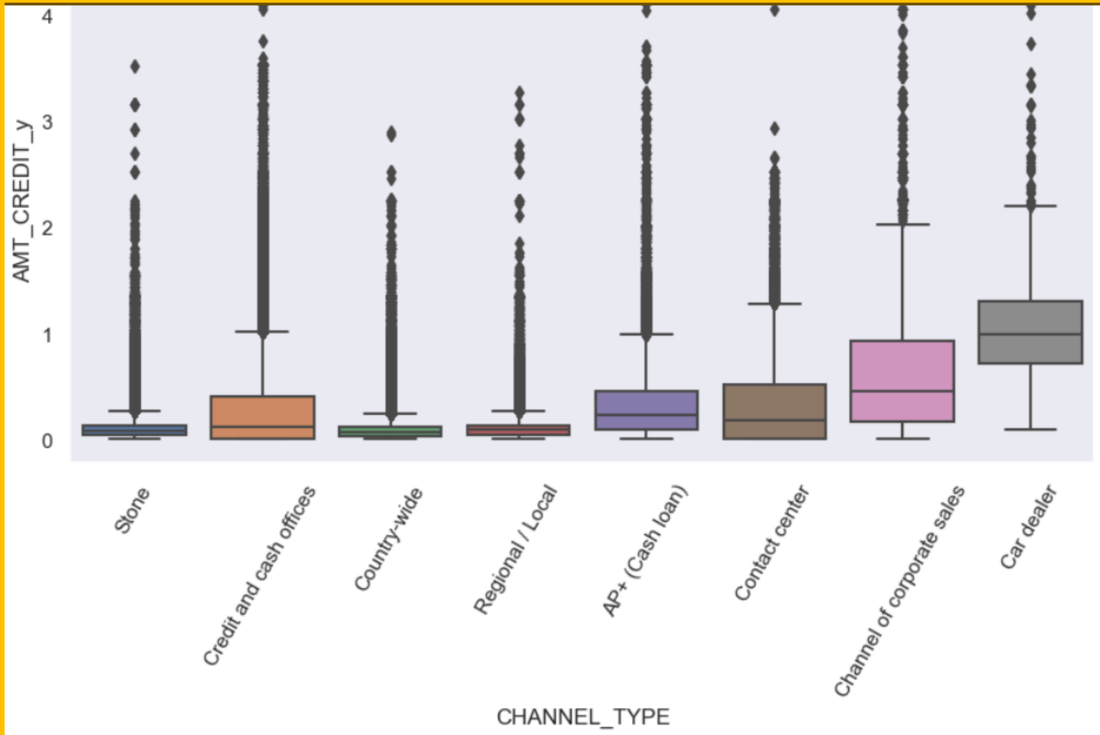


Bivariate analysis for categorical

OBSERVATIONS

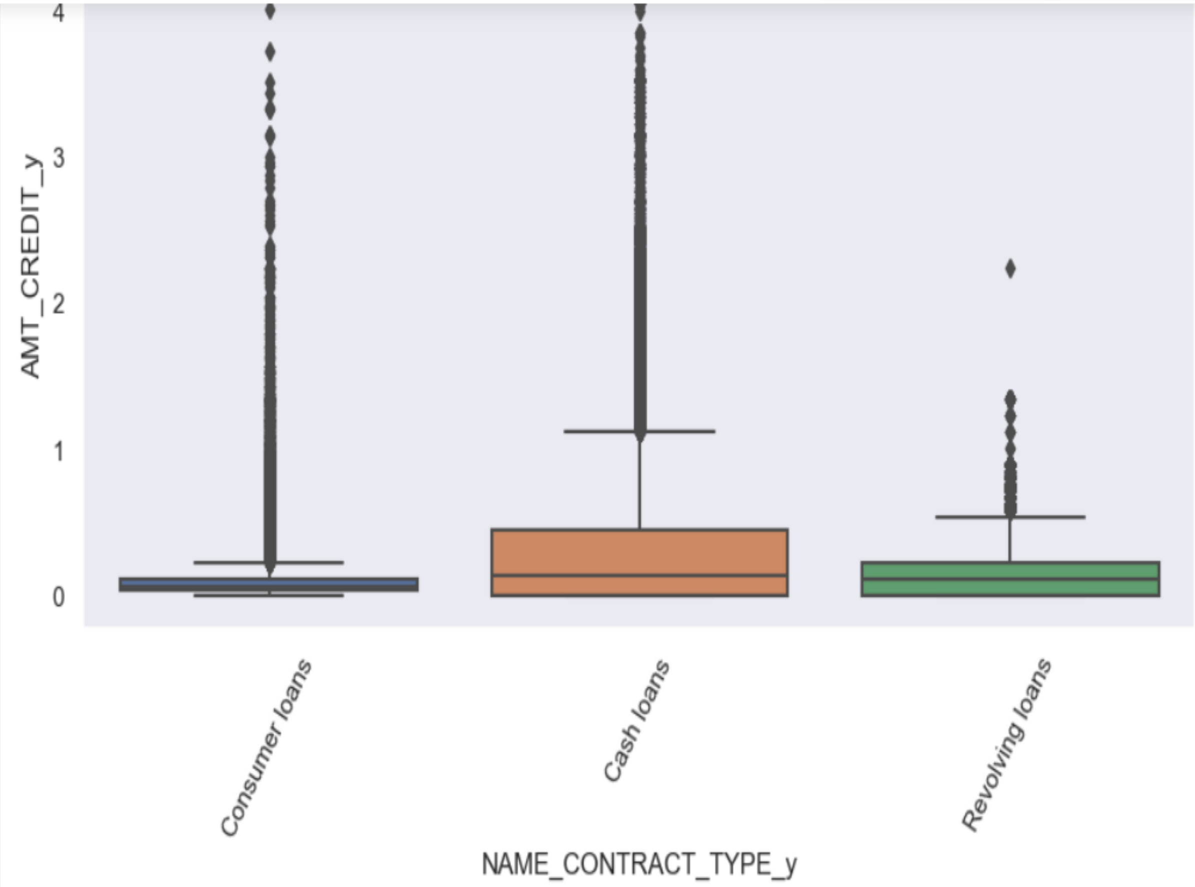
1. The amount credited by cars is the most followed by cash in portfolios.





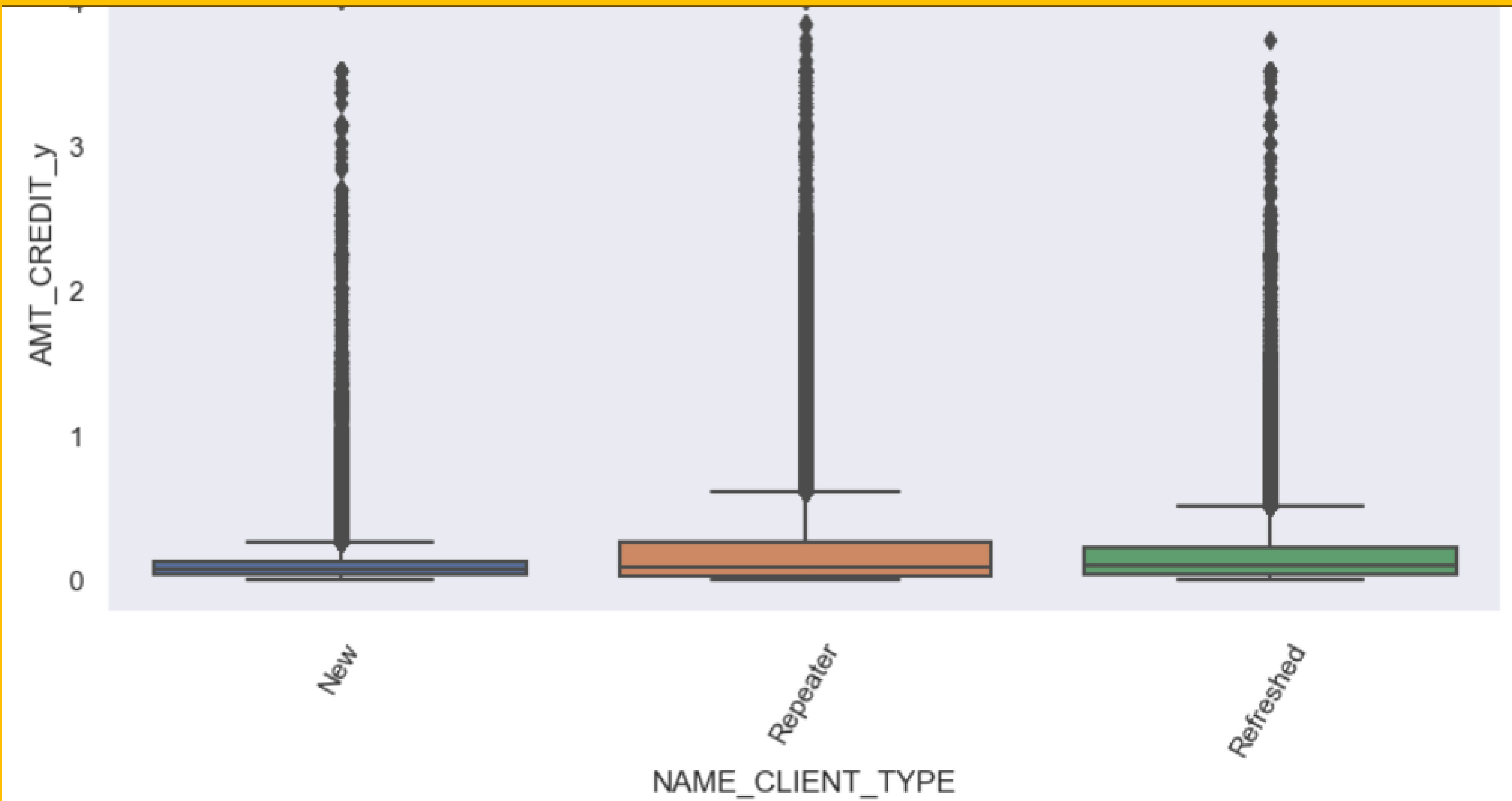
OBSERVTIONS

The amount credited by car dealer is the most followed by channel of corporate sales, contact center, credit and cash offices and then regional /local ,country - wide ,stone has almost same.



OBSERVTIONS

Amount credited by cash loans is maximum.



OBSERVATIONS

Amount credited by Repeaters is more as compared to the other one.

CONCLUSION

The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

1.**TARGET** for application data set = Column Target representing a number whether the client is defaulter or not.

Target 0 = Non - Defaulter

Target 1 = Defaulter

2. Target for previous data set is **NAME_CONTRACT_STATUS** which represent four values of loan status below are the percentages.

Approved	-	62.679378
Canceled	-	18.351900
Refused	-	17.357984
Unused offer	-	1.610737

OVERALL ANALYSIS

1.The main columns on which the bank should focus are:

a) AMT_INCOME
b) AMT_CREDIT
c) AMT_ANNUITY
d) CODE_GENDER
e) NAME_EDUCATION_TYPE
f) DAYS_BIRTH
g) CNT_CHILDREN
h) AMT_GOODS_PRICE
i) NAME_FAMILY_STATUS
j) NAME_HOUSING_TYPE
h) OCCUPATION_TYPE

2. We can clearly observe that the number of female clients are more in the bank so bank should focus on them.
3. They should focus more on Business men , pensioner and students as the chance of becoming defaulter is quite less.

4. Defaulters are more in case of young age group that is below 40 years.
5. Less number of defaulters are in case of academic degree.
6. Single people are likely to pay their loan at time.
7. The one who has rented apartment is likely to be non-defaulter.
8. More is the income less number of defaulters. There are less defaulters when the salary is more than 15L.
9. Credit amount of loans are very low for revolving loan.
10.Marriage,civil marriage and separated have higher amount credit.
11.Number of repeaters are quite high as compared to new one. In order to maintain good business bank should focus on bringing new clients.
12.Credit card and cash offices has mostly used as a channel type. Bank should focus on other channels also.
13.People living with family are applying for loan.
14.The amount credited by cars is the most in case of portfolios.
15. The amount credited by car dealer is the most.
16. Amount credited by cash loans is maximum.
17.Amount credited by Repeaters is more as compared to the other one.So the bank should try to retain its clients.
18. Bank should focus on working professionals as the chance of becoming defaulter is quite less.
19.It is more likely that the applicants who paid the previous loans on time will pay the current loan on time as well as compared to the applicants whose previous loans were rejected.
20.There are only 7.5% of approved applicants who defaulted in the current loan.
21. 88.0% of the applicants who were previously refused were able to pay current loan.

THANK YOU !!