Problem Statement / Use Case

Bank and financial institutions are supposed to provide loans to their customers however they need to take an intelligent decision before disbursing any loan. If the applicant becomes a defaulter then it is a financial loss or if Banks do not offer loan services to their customers then It is a loss of potential business.

The given data contains information about loan applications at the time of applying for a loan, it contains 2 scenarios:

1. Client with Payment difficulties.
2. All other Cases.

Business Objective

The company wants to find the important column or attributes which can help to identify the possibility of defaulters i.e. variables that are strong indicators of defaulters.

Also, this case study will identify patterns that indicate if a customer has difficulty in paying their installment which may help to take corrective actions and decisions higher interest rates, reduced loan amount, or denying the loan.

Data Set

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 on whether the previous application had been **approved, Canceled, Refused, or Unused offer.**

3. 'columns_description.csv' is a data dictionary that describes the meaning of the variables.

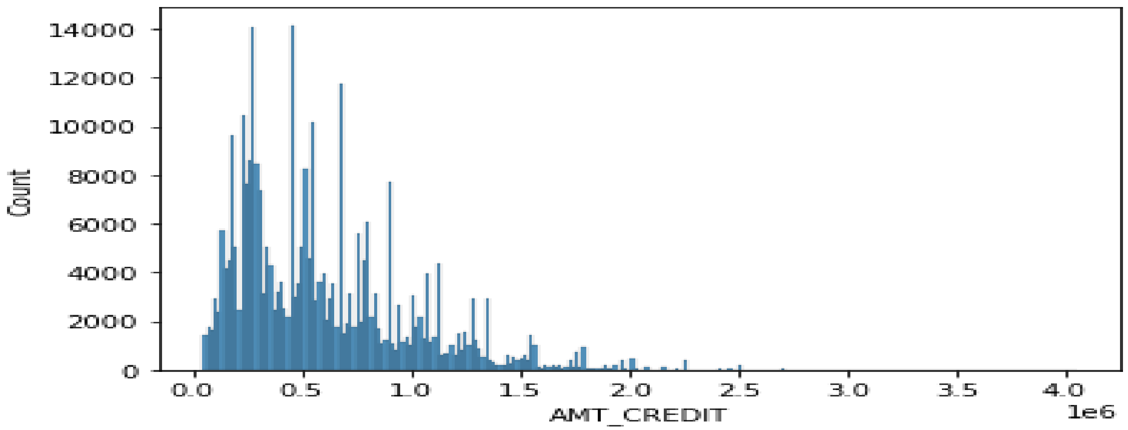
Analysis of Application Data

Approach Followed:-

1. Load the application_data.csv
2. Check the structure and the Metadata
3. Find the Missing value % in each column
4. Drop the columns having higher Missing Value %(>45% in this dataset) as higher missing values will impact analysis.
5. Identification/Imputation of Missing values of the remaining columns.
6. Identification/Division of important columns for further Analysis into Continuous and Categorical Columns
7. Outlier Identification/Analysis through Boxplot.
8. Outlier Treatment.
9. Univariate Analysis for Continuous Data
10. Bivariate Analysis for Continuous Vs Categorical Data
11. Finding the Data Imbalance.
12. Segmented Analysis for TARGET
13. Multivariate Analysis for Continuous Data
14. Multivariate Analysis for Categorical Data
15. Assumption: XNA value in Organization Type is considered as Null

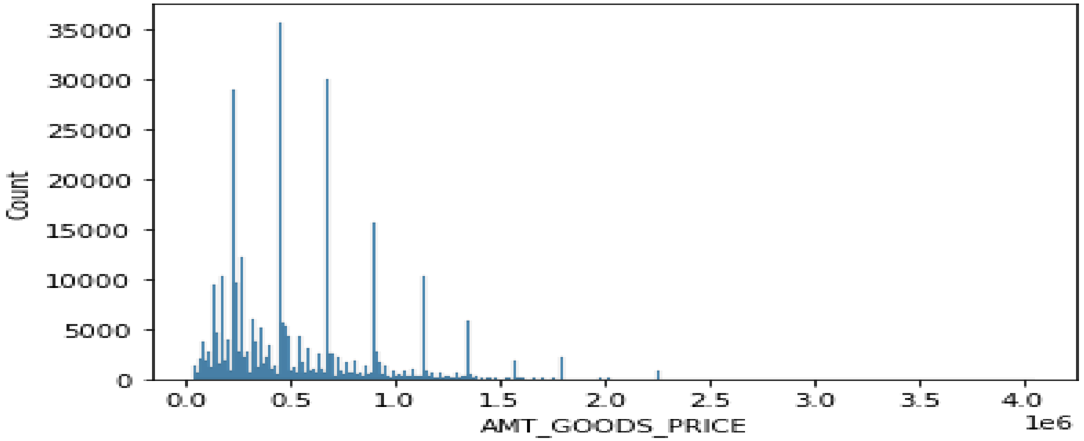
Univariate Analysis for Continuous Columns in Application data

Distribution of AMT_CREDIT

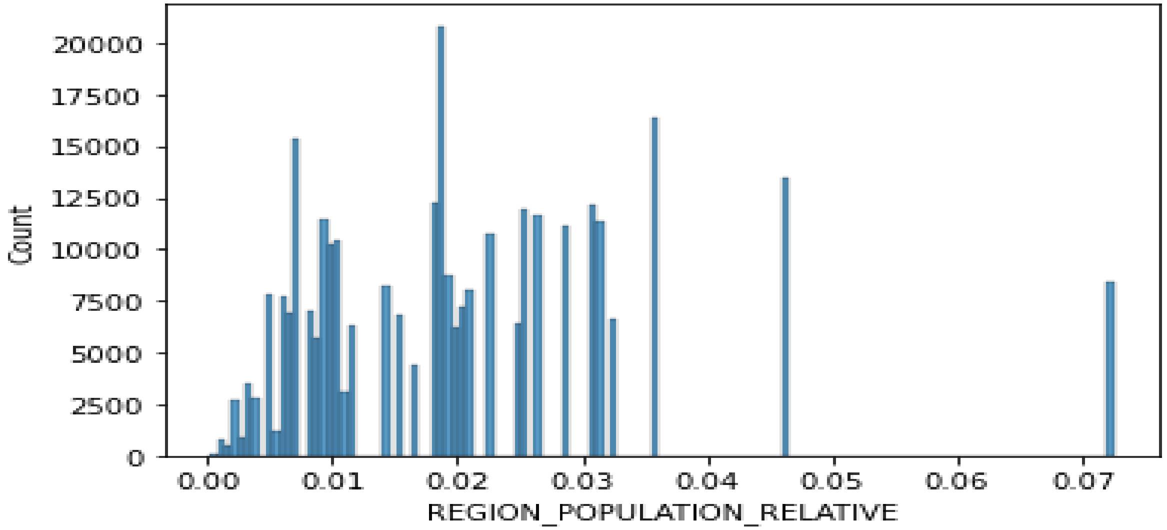


Maximum Clients get amount credited is < 50K

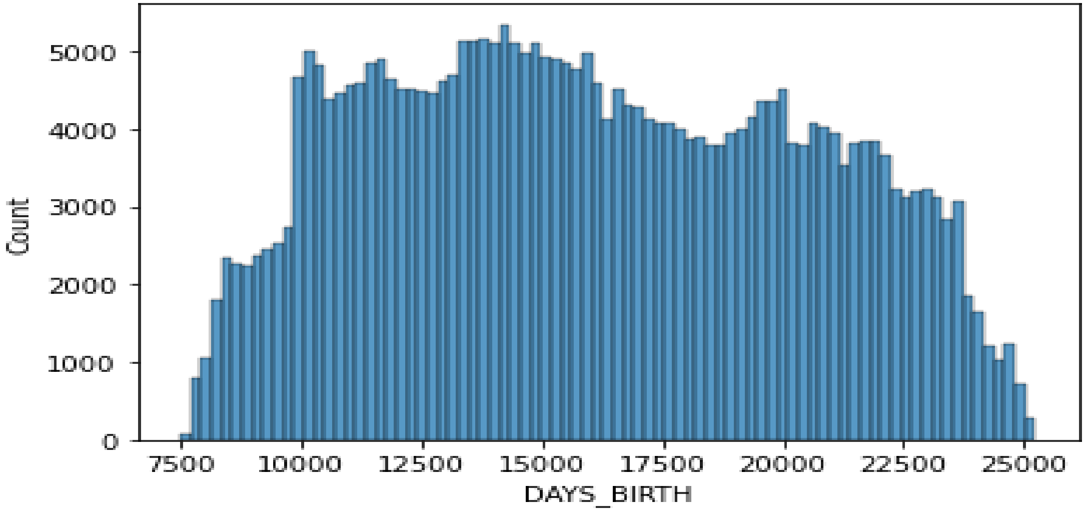
Distribution of AMT_GOODS_PRICE



Largest no. of Clients gets loan against 5L AMT_GOODS_PRICE



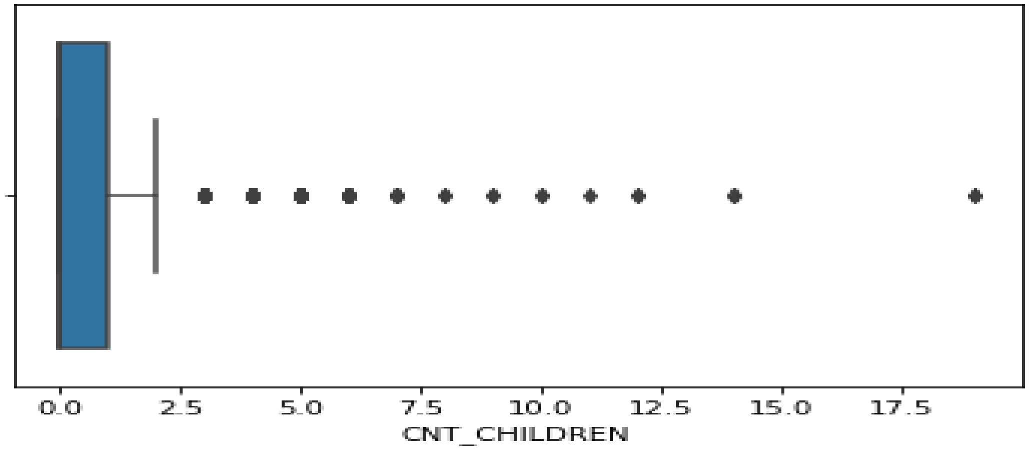
Distribution of REGION_POPULATION_RELATIVE



Distribution of Days of birth

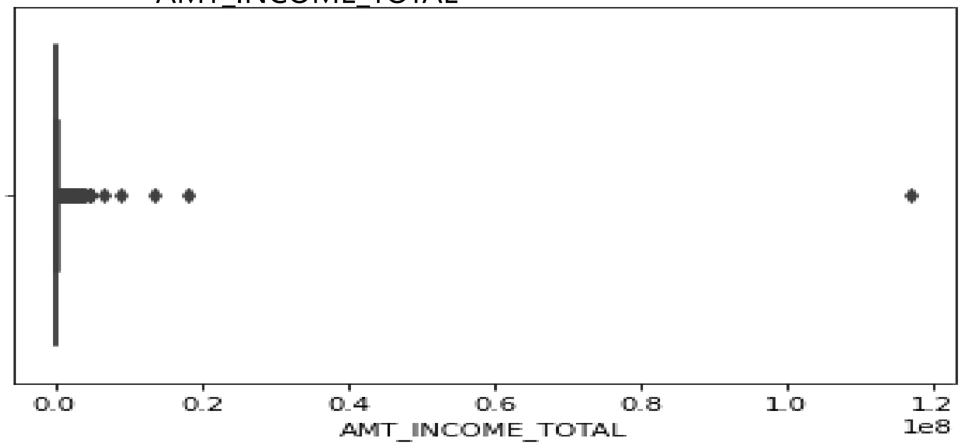
Outlier Analysis (Boxplot) in Application data

CNT_Children

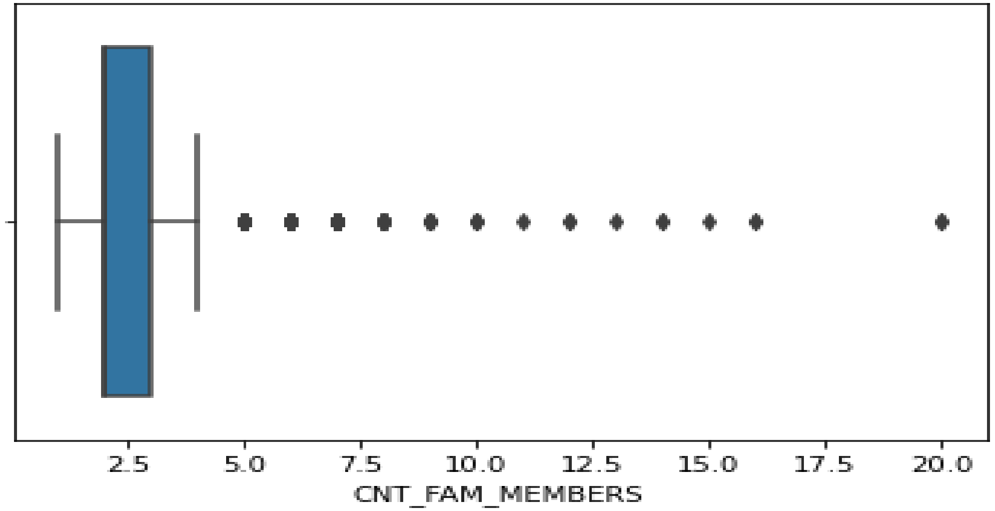


Boxplots clearly show the values above 2.5 are outliers

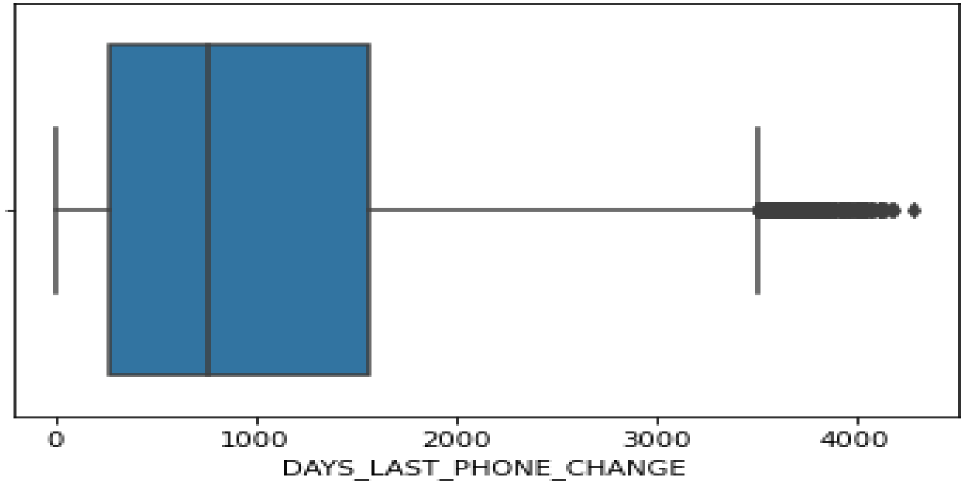
AMT_INCOME_TOTAL



Applicants with Income above 900K are outliers.

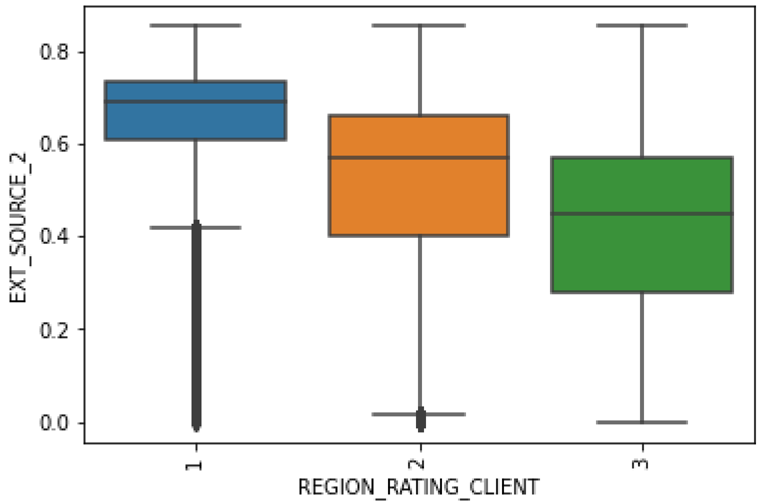


Applicants with 5 or more family members are clearly outliers

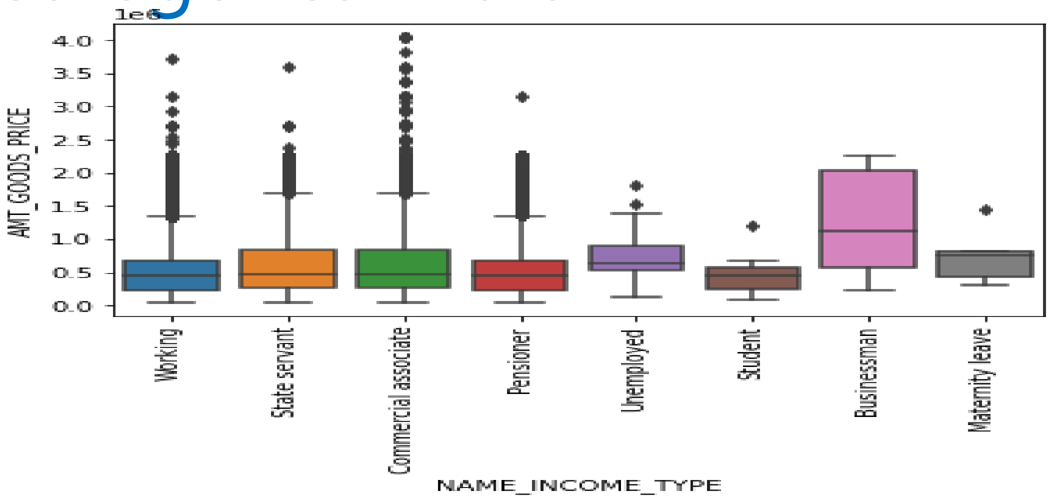


Applicants with `DAYS_LAST_PHONE_CHANGE` above 3514.0 are outliers

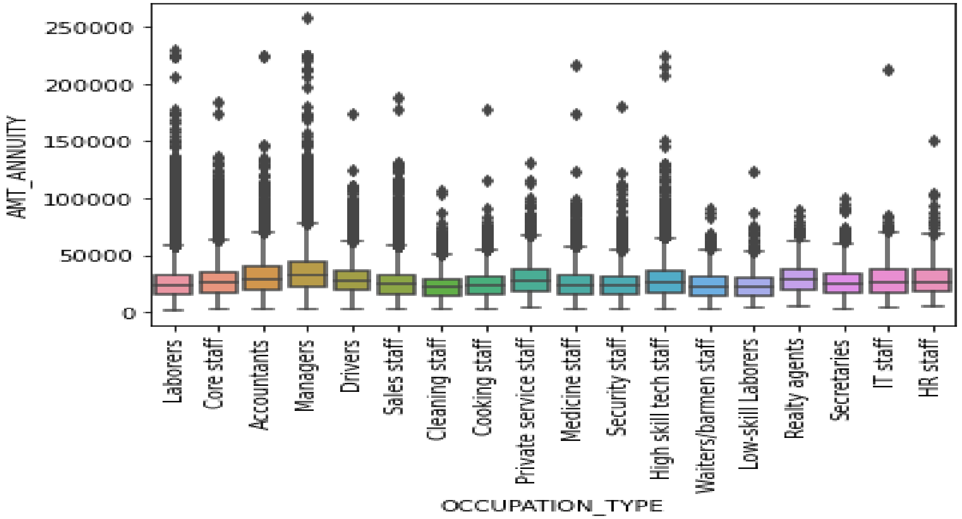
Bivariate Analysis for Continuous Vs Categorical Data



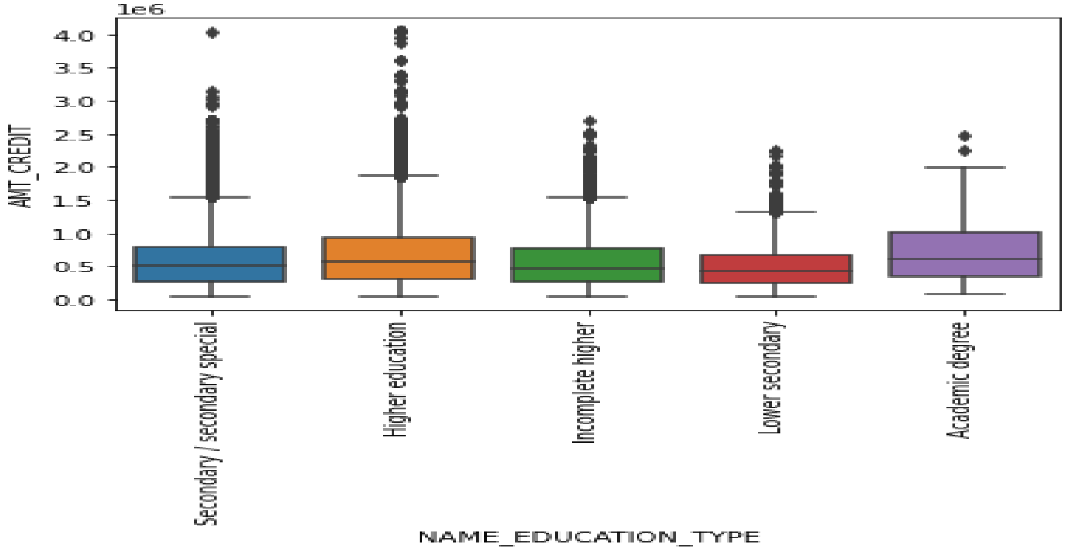
Client staying in Region Ration 1
have higher EXT Source2 data.



Loaned Amount is higher for Businessman



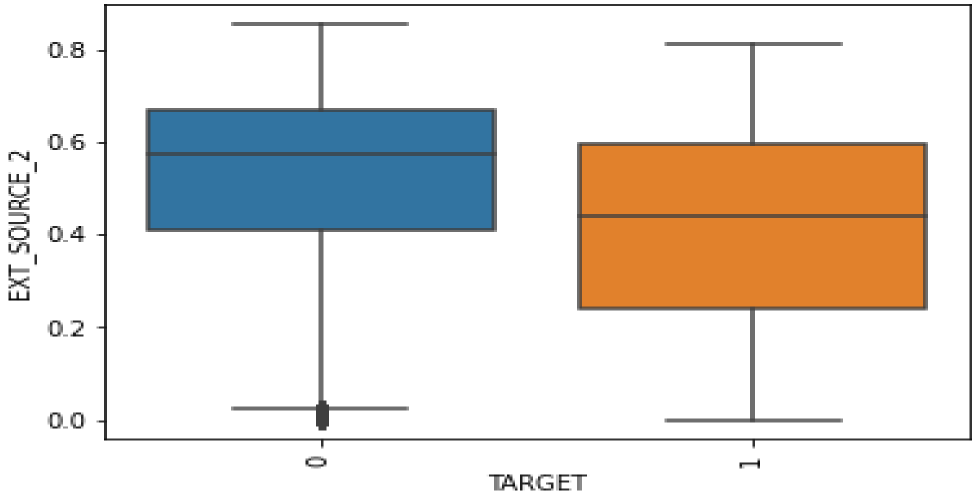
Amount Annuity is higher for highly skilled occupation Type

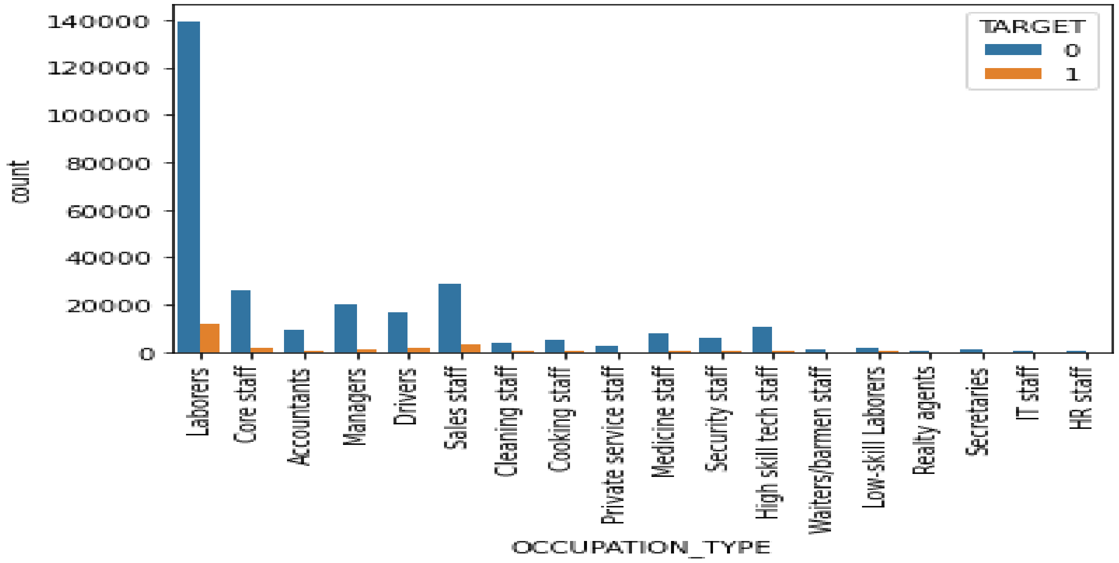


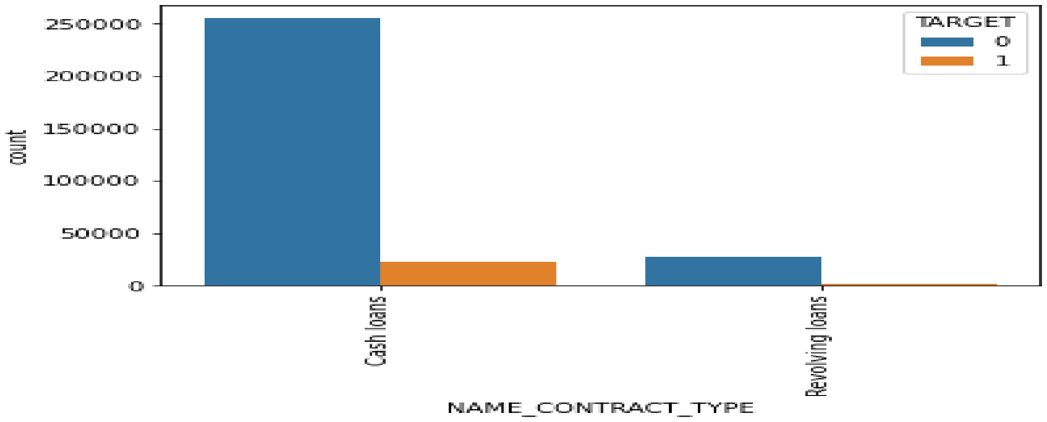
Amount Credited to Academic Degree is higher

Bivariate Analysis for Categorical Vs TARGET

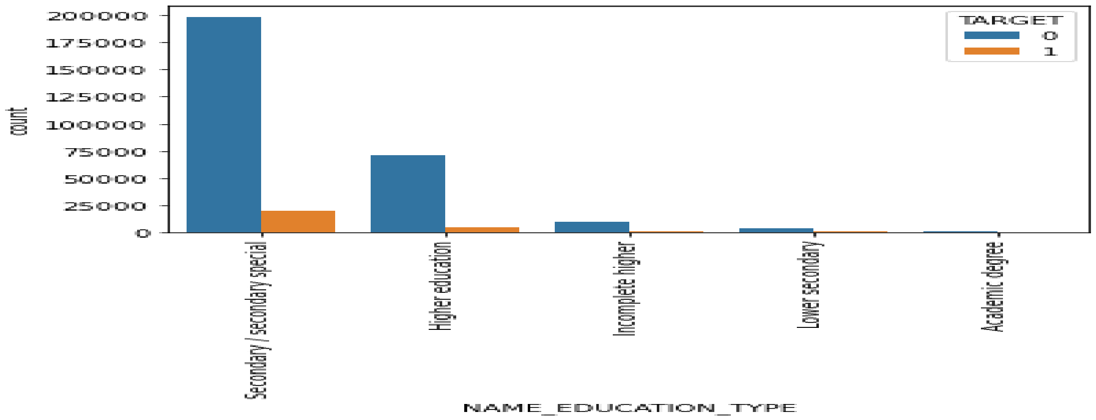
(Defaulters/Having difficulties in payment (TARGET==1),Non Defaulters/All other cases (TARGET ==0)





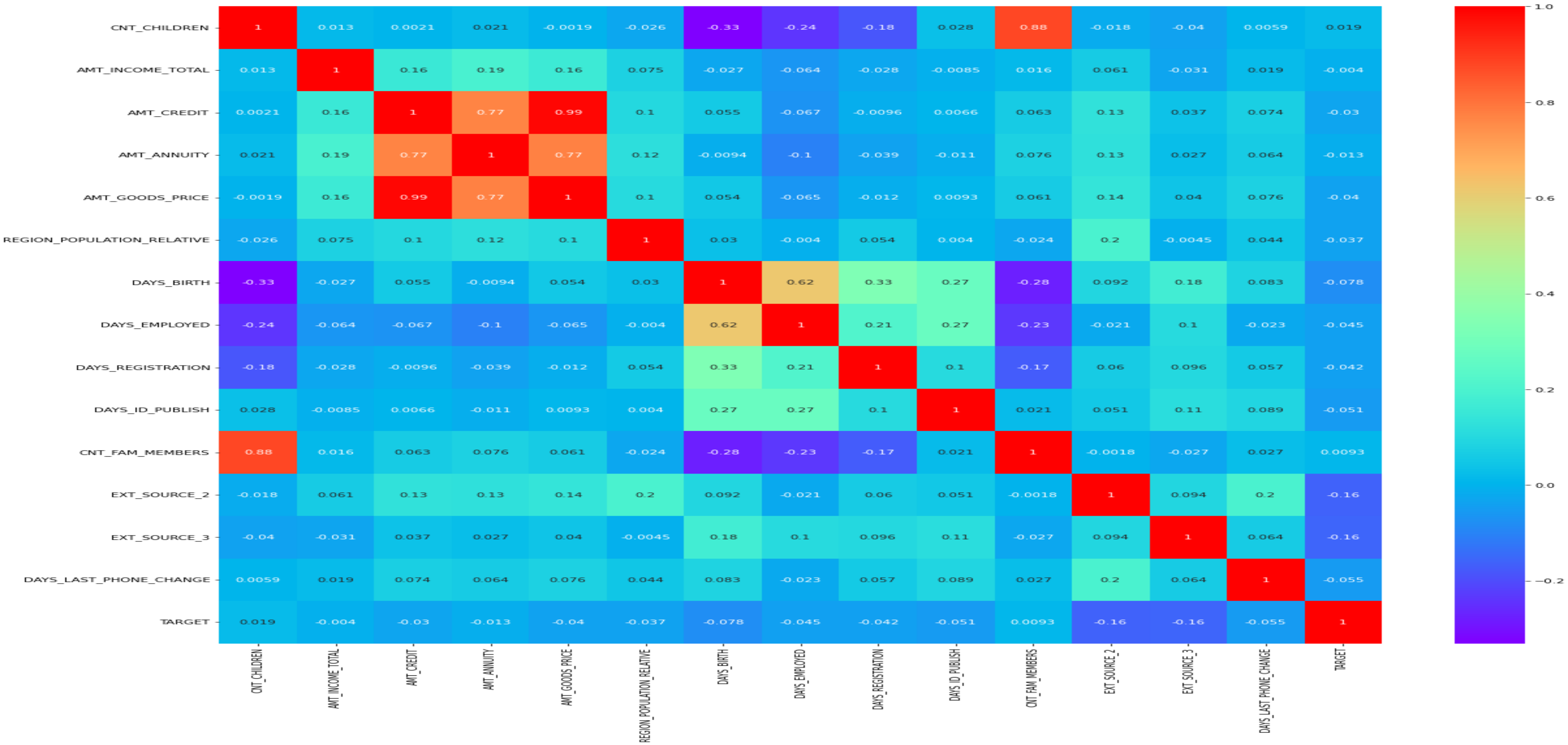


Defaulters are higher in Cash Loans



Defaulters are higher for Secondary/Secondary Special

Multivariate Analysis for Continuous Columns

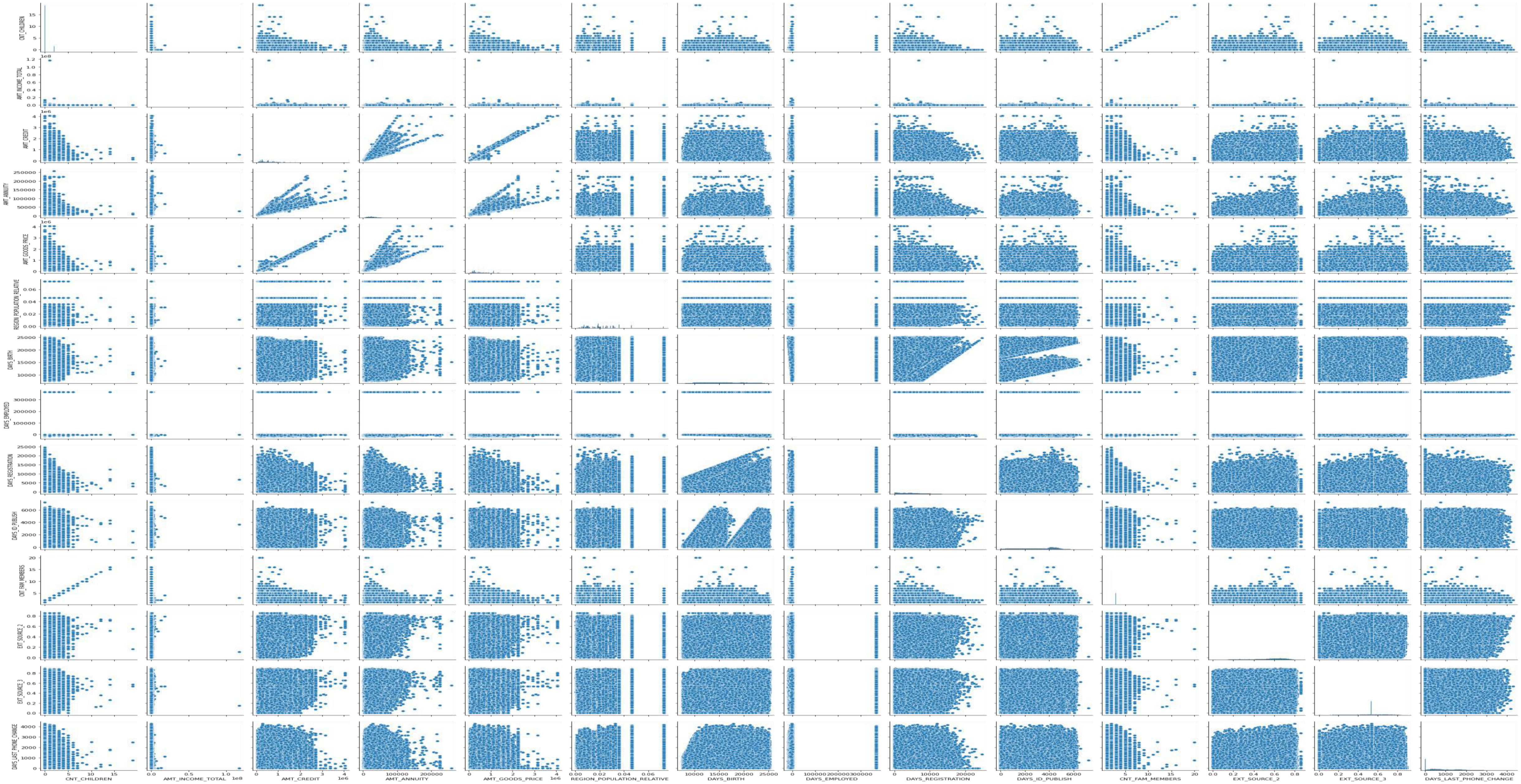


	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	EXT_SOURCE_2	EXT_SOURCE_3	DAYS_LAST_PHONE_CHANGE	TARGET
CNT_CHILDREN	1	0.013	0.0021	0.021	-0.0019	-0.026	-0.33	-0.24	-0.18	0.028	0.88	-0.018	-0.04	0.0059	0.019
AMT_INCOME_TOTAL	0.013	1	0.16	0.19	0.16	0.075	-0.027	-0.064	-0.028	-0.0085	0.016	0.061	-0.031	0.019	-0.004
AMT_CREDIT	0.0021	0.16	1	0.77	0.99	0.1	0.055	-0.067	-0.0096	0.0066	0.063	0.13	0.037	0.074	-0.03
AMT_ANNUITY	0.021	0.19	0.77	1	0.77	0.12	-0.0094	-0.1	-0.039	-0.011	0.076	0.13	0.027	0.064	-0.013
AMT_GOODS_PRICE	-0.0019	0.16	0.99	0.77	1	0.1	0.054	-0.065	-0.012	0.0093	0.061	0.14	0.04	0.076	-0.04
REGION_POPULATION_RELATIVE	-0.026	0.075	0.1	0.12	0.1	1	0.03	-0.004	0.054	0.004	-0.024	0.2	-0.0045	0.044	-0.037
DAYS_BIRTH	-0.33	-0.027	0.055	-0.0094	0.054	0.03	1	0.62	0.33	0.27	-0.28	0.092	0.18	0.083	-0.078
DAYS_EMPLOYED	-0.24	-0.064	-0.067	-0.1	-0.065	-0.004	0.62	1	0.21	0.27	-0.23	-0.021	0.1	-0.023	-0.045
DAYS_REGISTRATION	-0.18	-0.028	-0.0096	-0.039	-0.012	0.054	0.33	0.21	1	0.1	-0.17	0.06	0.096	0.057	-0.042
DAYS_ID_PUBLISH	0.028	-0.0085	0.0066	-0.011	0.0093	0.004	0.27	0.27	0.1	1	0.021	0.051	0.11	0.089	-0.051
CNT_FAM_MEMBERS	0.88	0.016	0.063	0.076	0.061	-0.024	-0.28	-0.23	-0.17	0.021	1	-0.0018	-0.027	0.027	0.0093
EXT_SOURCE_2	-0.018	0.061	0.13	0.13	0.14	0.2	0.092	-0.021	0.06	0.051	-0.0018	1	0.094	0.2	-0.16
EXT_SOURCE_3	-0.04	-0.031	0.037	0.027	0.04	-0.0045	0.18	0.1	0.096	0.11	-0.027	0.094	1	0.064	-0.16
DAYS_LAST_PHONE_CHANGE	0.0059	0.019	0.074	0.064	0.076	0.044	0.083	-0.023	0.057	0.089	0.027	0.2	0.064	1	-0.055
TARGET	0.019	-0.004	-0.03	-0.013	-0.04	-0.037	-0.078	-0.045	-0.042	-0.051	0.0093	-0.16	-0.16	-0.055	1

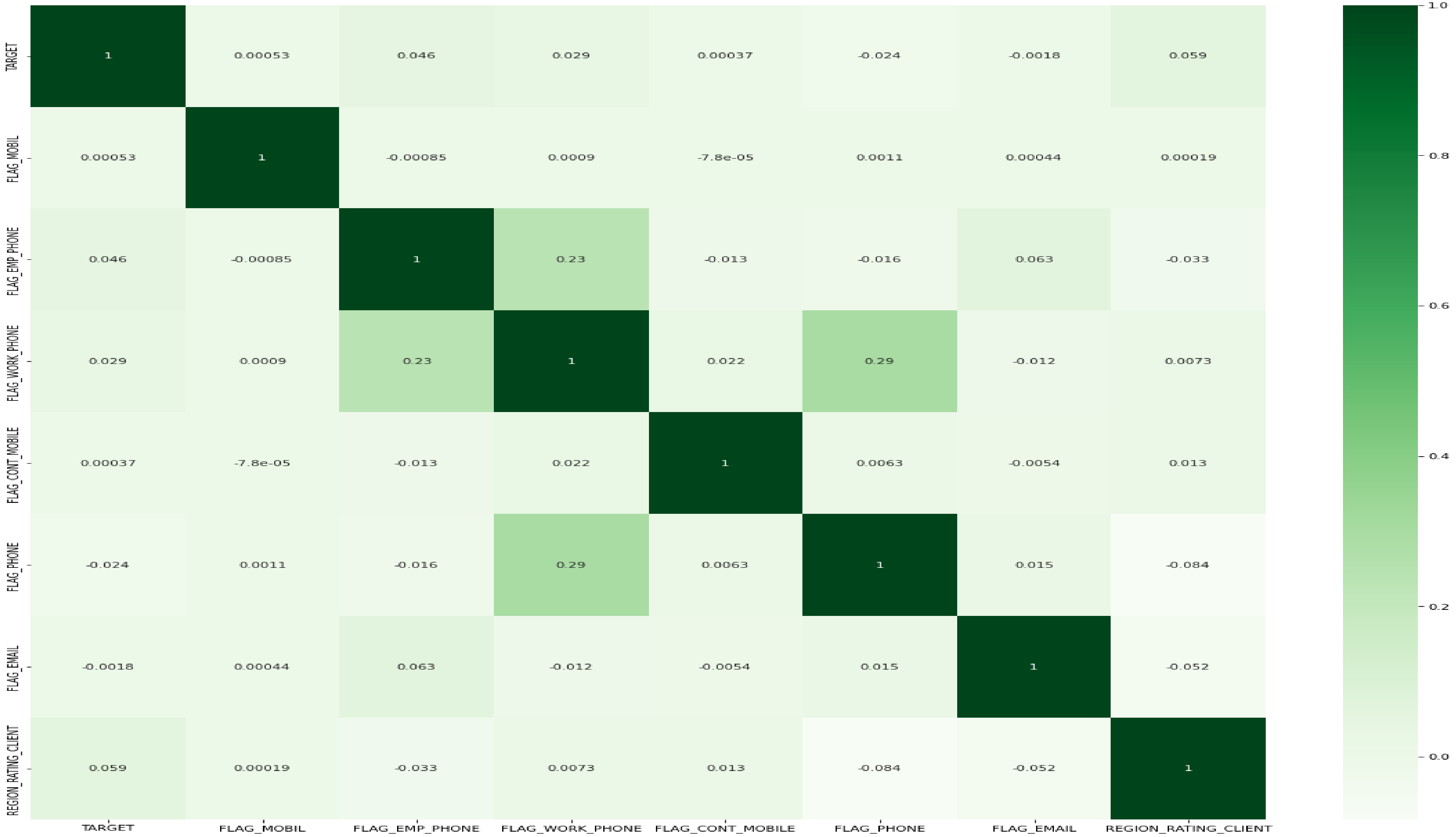
AMT_CREDIT has a high correlation with AMT_GOOD_PRICE
AMT_CREDIT has a high correlation with AMT_ANNUITY

AMT_ANNUITY has a high correlation with AMT_GOOD_PRICE
CNT_CHILDREN has a high correlation with CNT_FAM_MEMBERS

Multivariate Analysis for Continuous Columns –Pair Plot



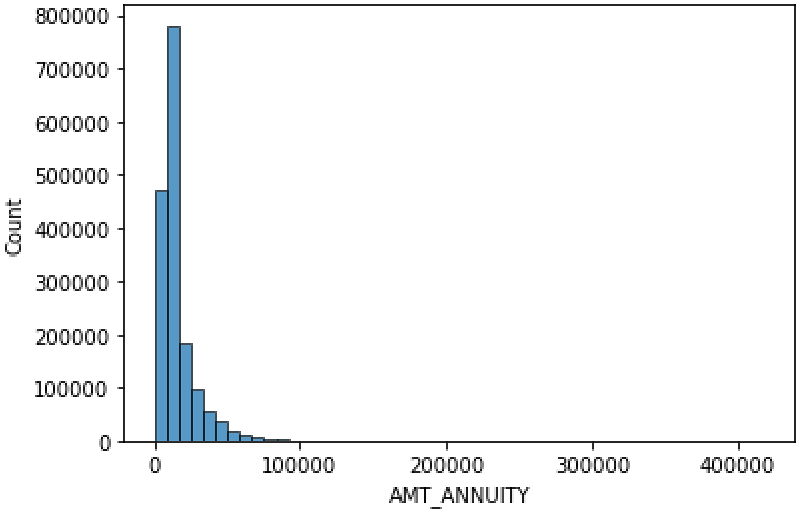
Multivariate Analysis for Categorical Columns

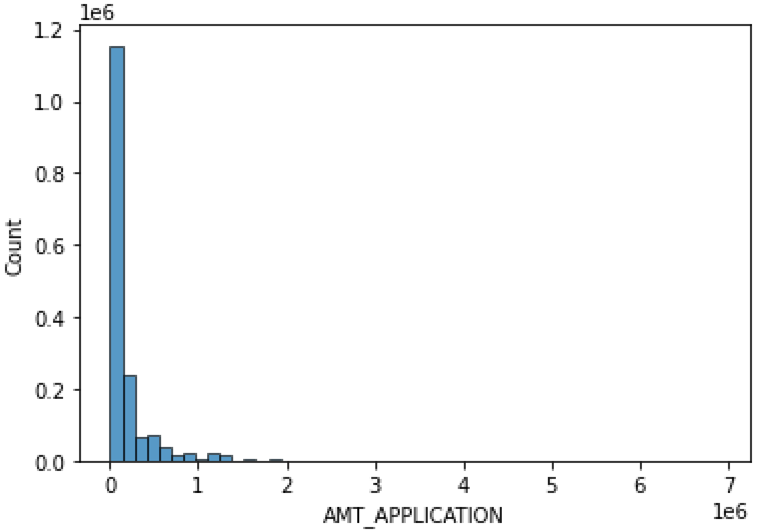


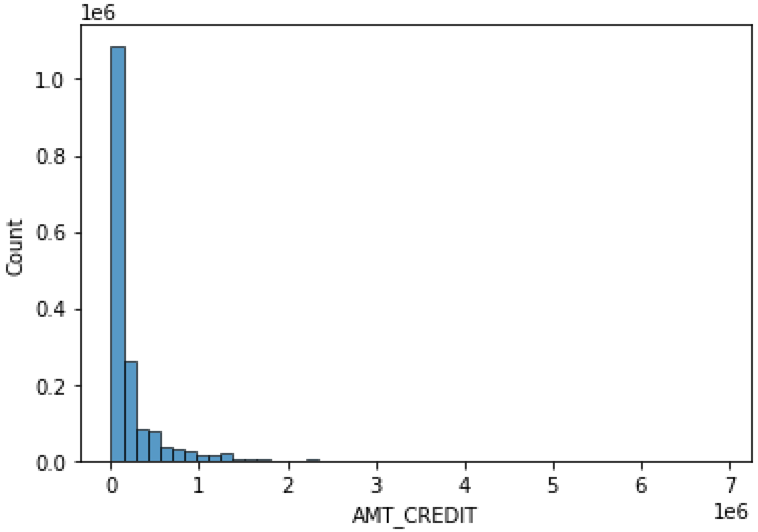
	TARGET	FLAG_MOBIL	FLAG_EMP_PHONE	FLAG_WORK_PHONE	FLAG_CONT_MOBILE	FLAG_PHONE	FLAG_EMAIL	REGION_RATING_CLIENT
TARGET	1	0.00053	0.046	0.029	0.00037	-0.024	-0.0018	0.059
FLAG_MOBIL	0.00053	1	-0.00085	0.0009	-7.8e-05	0.0011	0.00044	0.00019
FLAG_EMP_PHONE	0.046	-0.00085	1	0.23	-0.013	-0.016	0.063	-0.033
FLAG_WORK_PHONE	0.029	0.0009	0.23	1	0.022	0.29	-0.012	0.0073
FLAG_CONT_MOBILE	0.00037	-7.8e-05	-0.013	0.022	1	0.0063	-0.0054	0.013
FLAG_PHONE	-0.024	0.0011	-0.016	0.29	0.0063	1	0.015	-0.084
FLAG_EMAIL	-0.0018	0.00044	0.063	-0.012	-0.0054	0.015	1	-0.052
REGION_RATING_CLIENT	0.059	0.00019	-0.033	0.0073	0.013	-0.084	-0.052	1

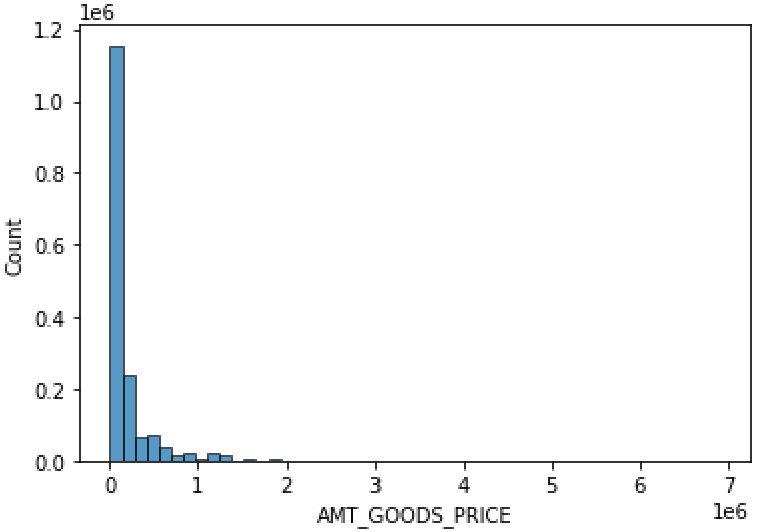
Analysis of Previous Application Data

Univariate Analysis for Continuous Columns in Previous Data

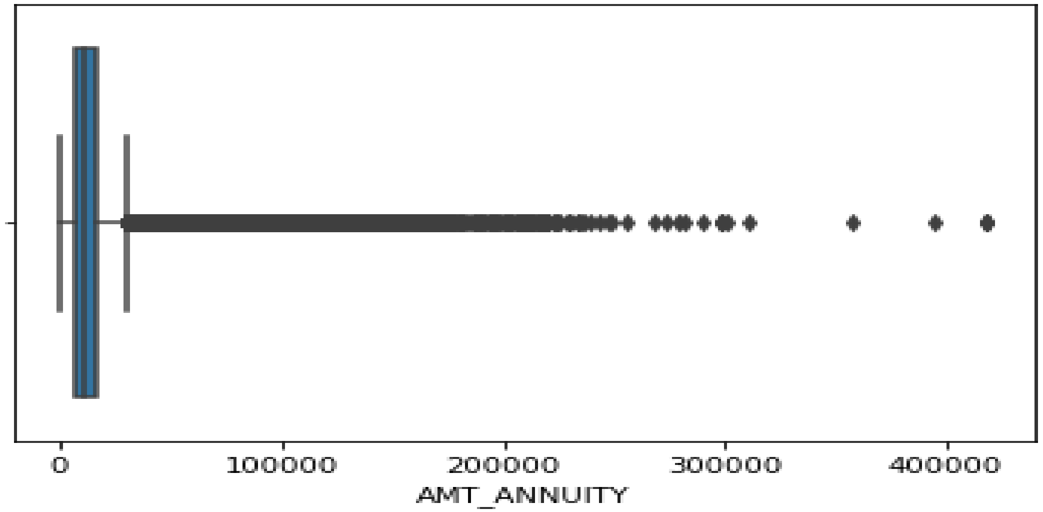




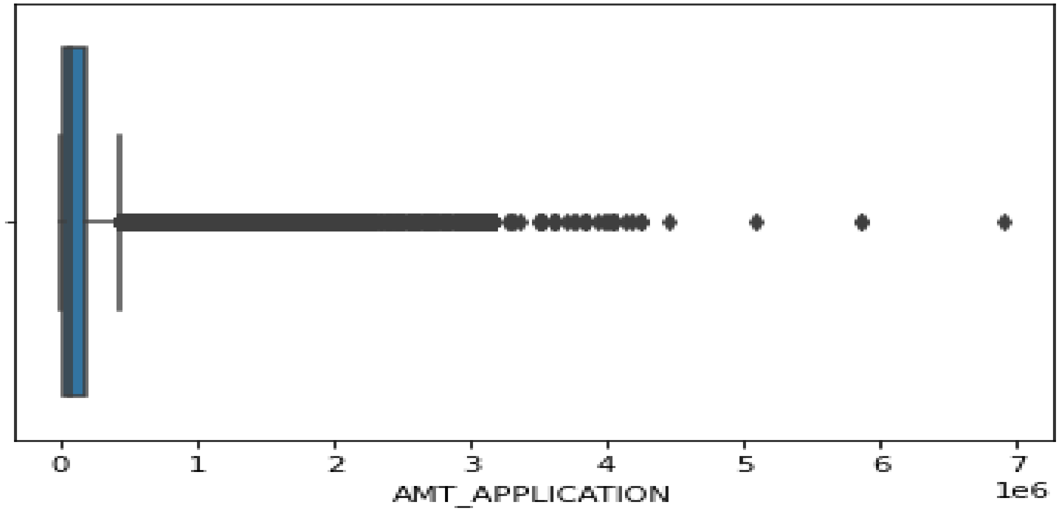




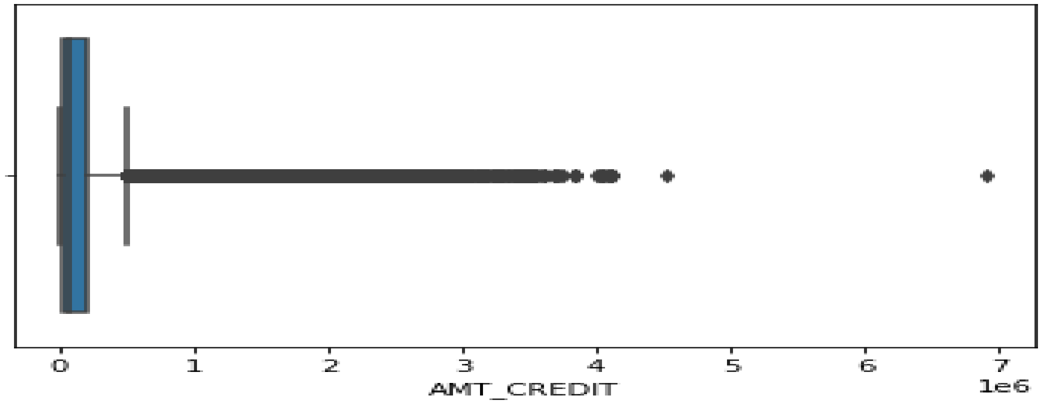
Outlier Analysis (Boxplot) in Previous Data



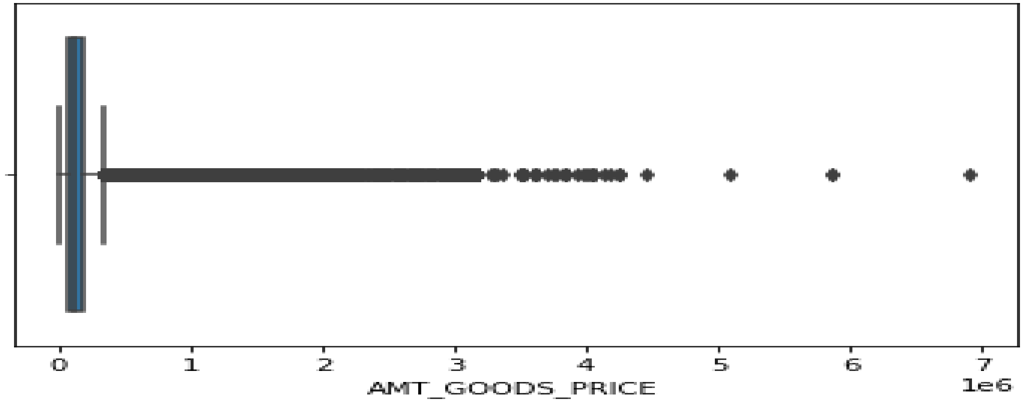
```
Outliers for AMT_ANNUITY
IQR = 9276.930000000002 : Floor = -6368.298750000004
Capping = 30739.421250000007
```



```
Outliers for AMT_APPLICATION
IQR = 161640.0 : Floor = -223740.0 : Capping = 422820.0
```

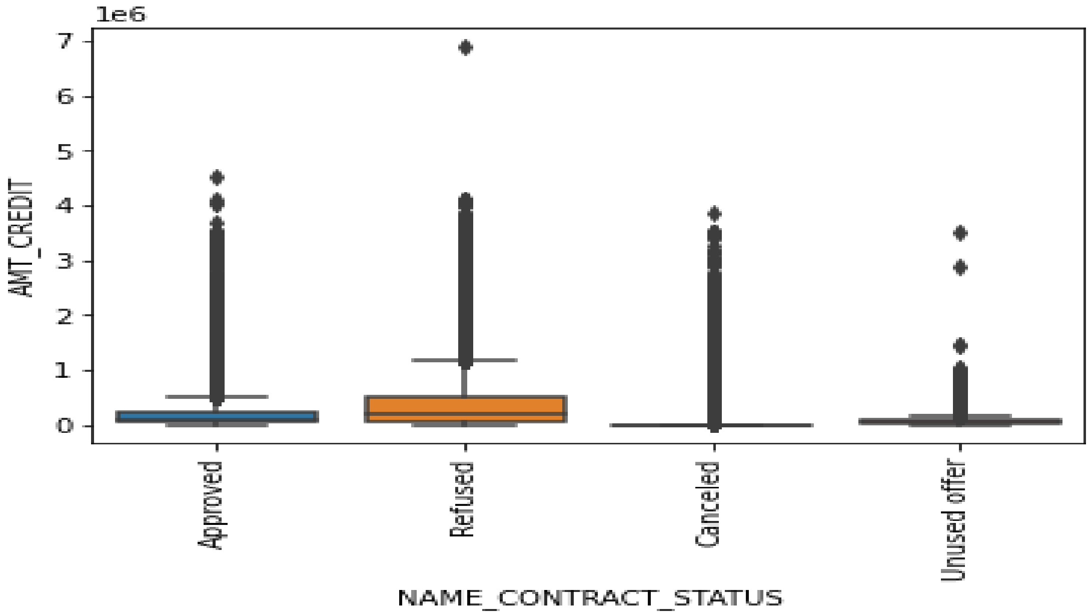


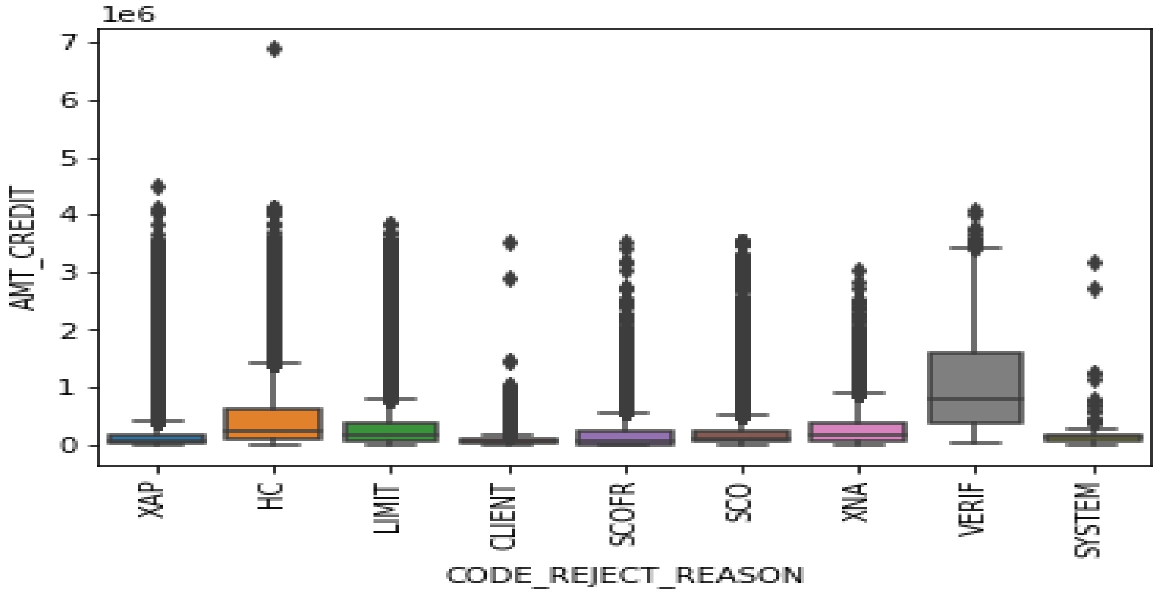
```
Outliers for AMT_CREDIT
IQR = 192258.0 : Floor = -264226.5 : Capping = 504805.5
```

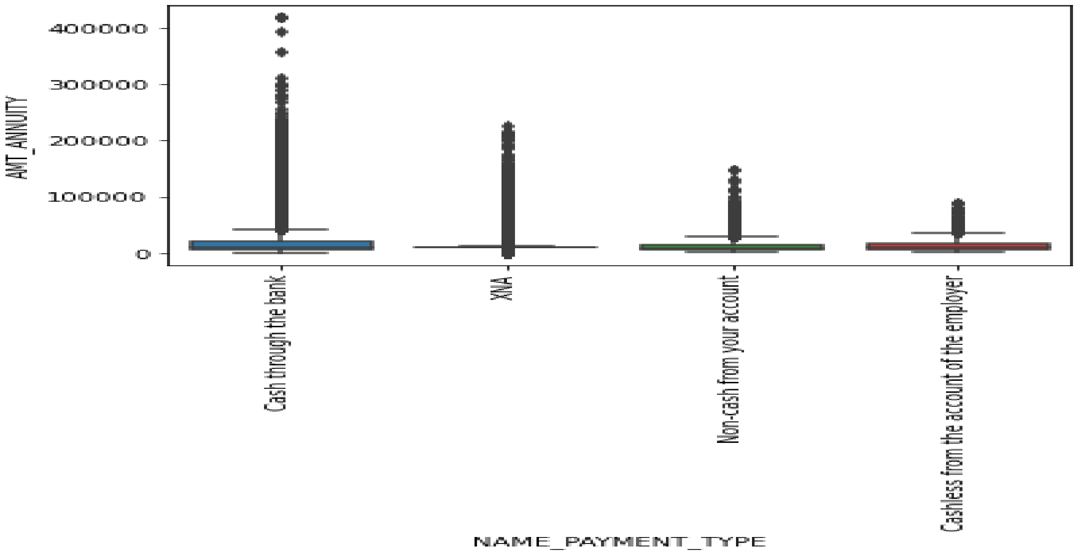


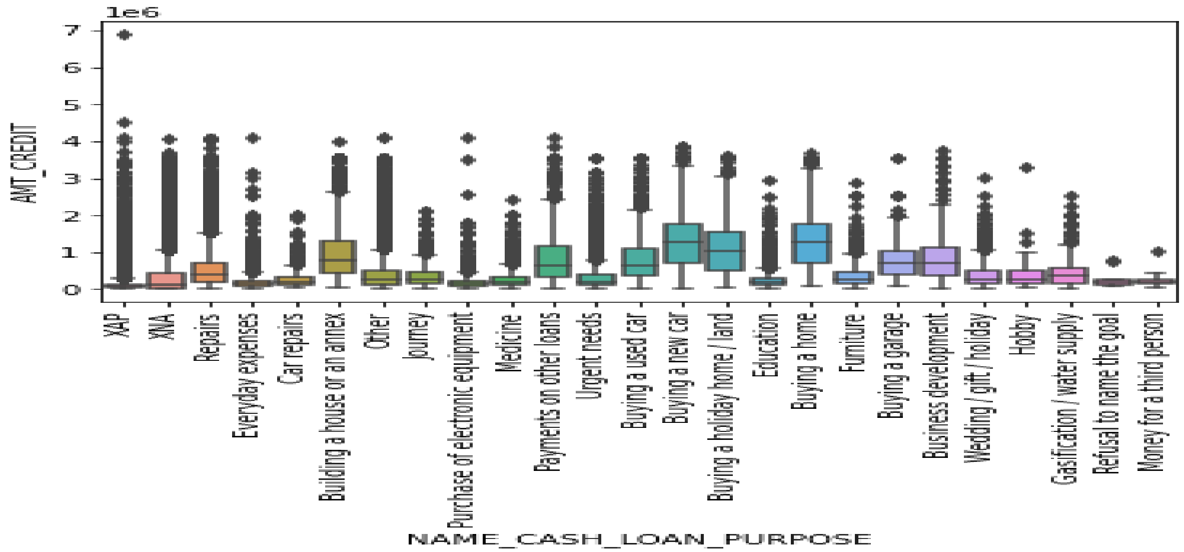
```
Outliers for AMT_GOODS_PRICE
IQR = 112905.0 : Floor = -101857.5 : Capping = 349762.5
```

Bivariate Analysis in Previous Data

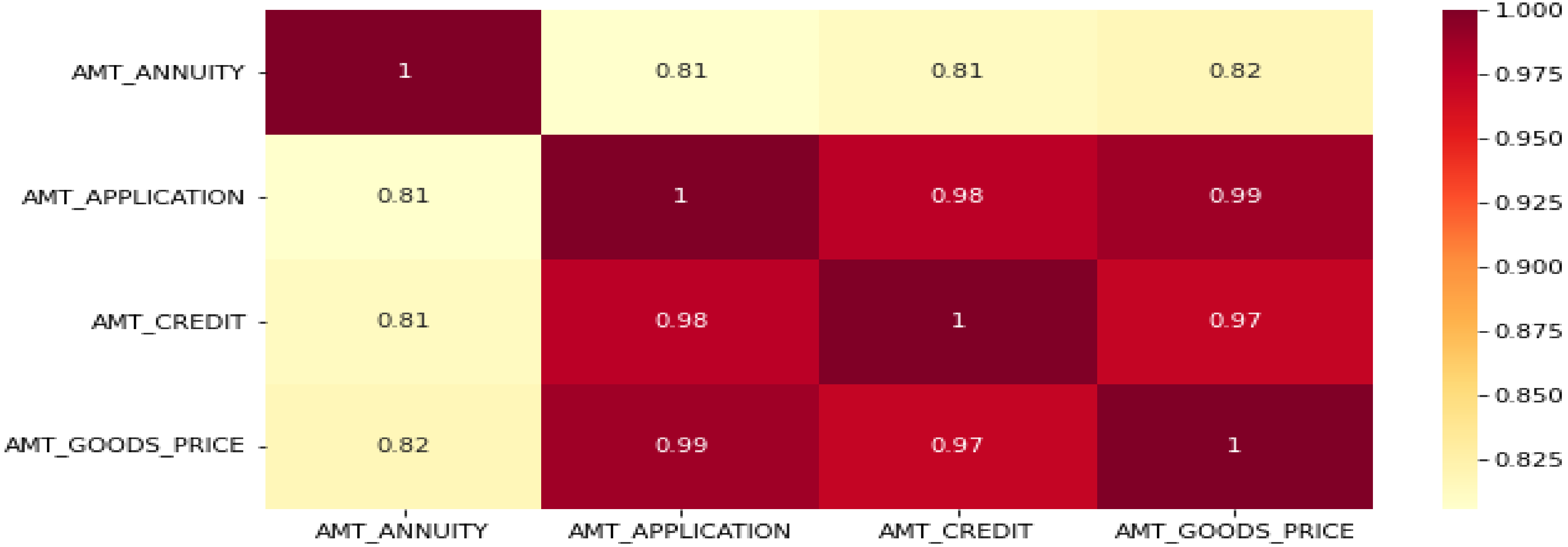








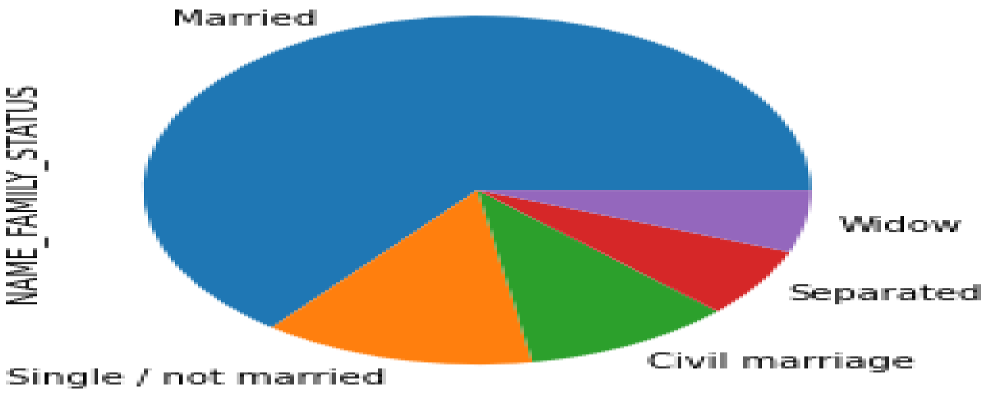
Multivariate Analysis of Previous Application Data

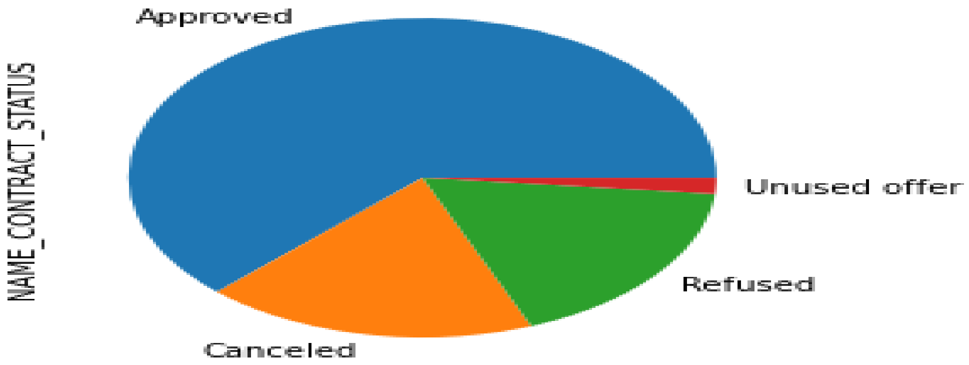


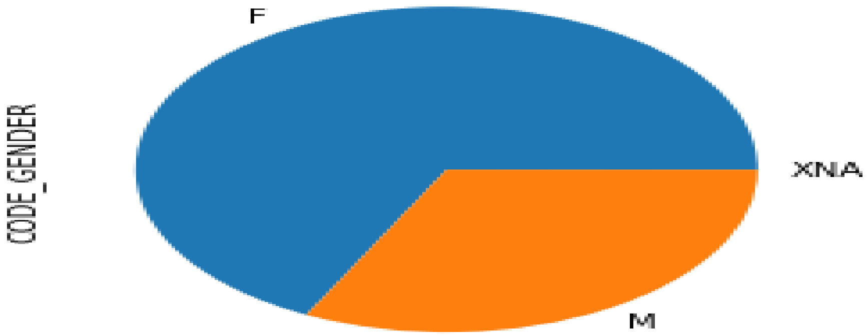
- There is a high correlation between AMT_APPLICATION and AMT_GOODS_PRICE
- There is a high correlation between AMT_APPLICATION and AMT_CREDIT
- There is a high correlation between AMT_GOODS_PRICE and AMT_CREDIT

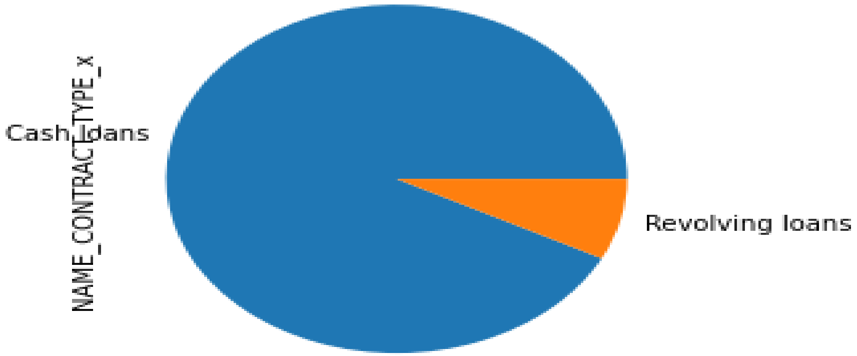
Analysis of Merged Dataset(Application Data + Previous Data)

Univariate Analysis of Merged Dataset

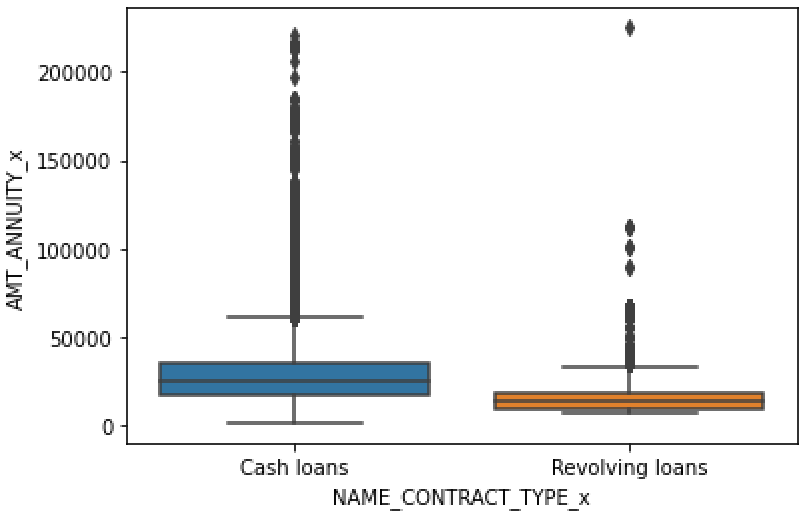


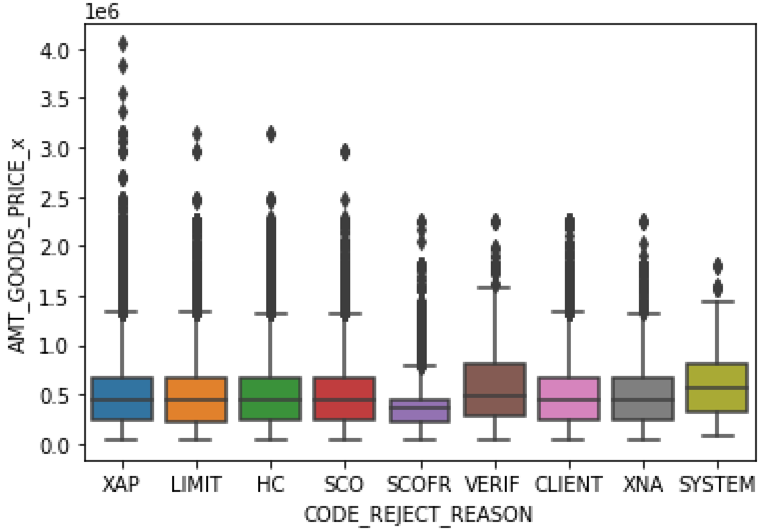


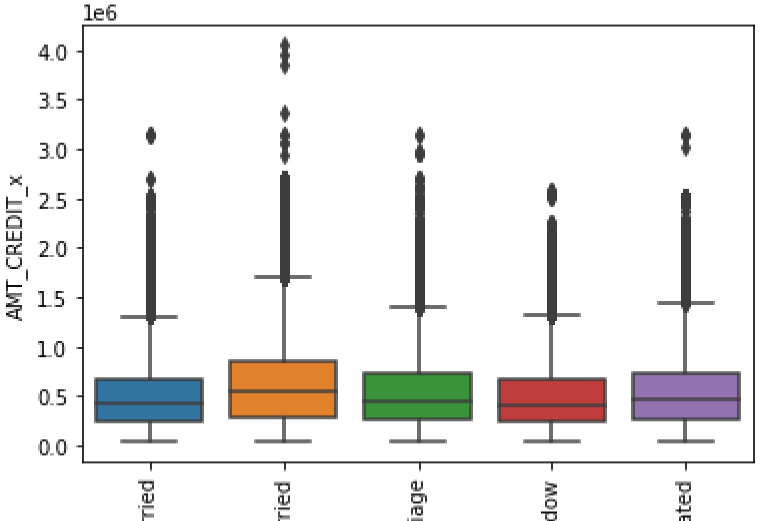


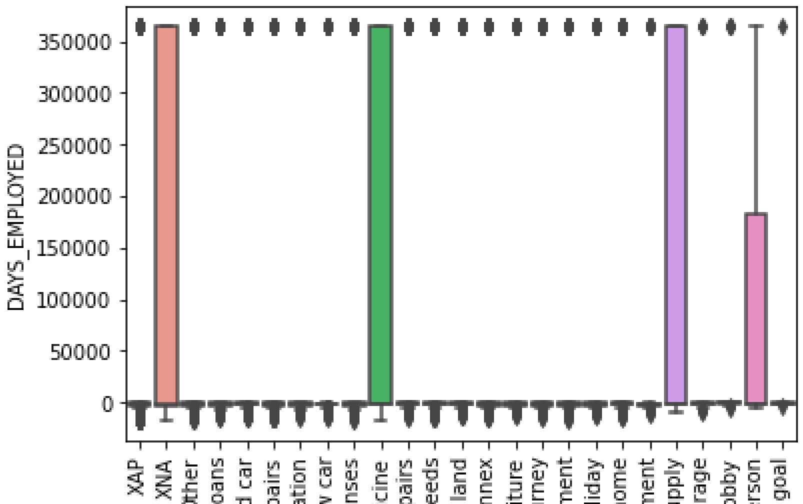


Bivariate Analysis of Merged Dataset

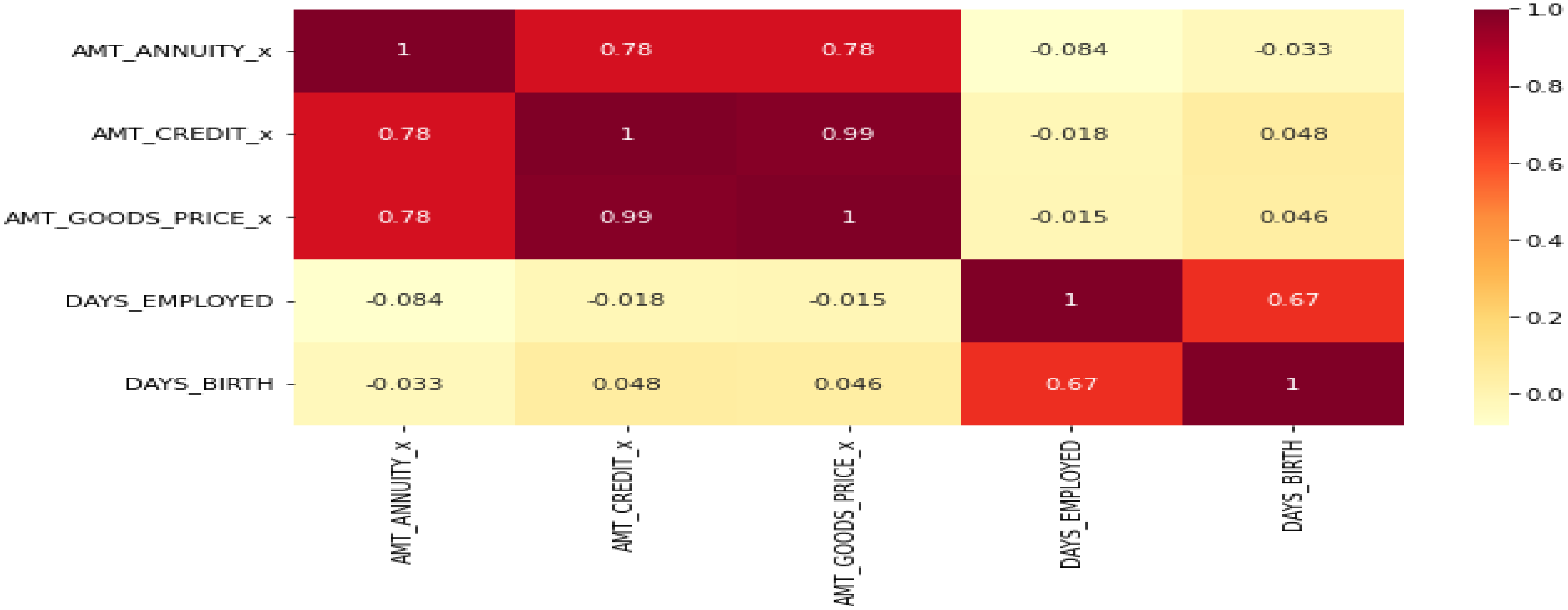








Multivariate Analysis of Merged Dataset

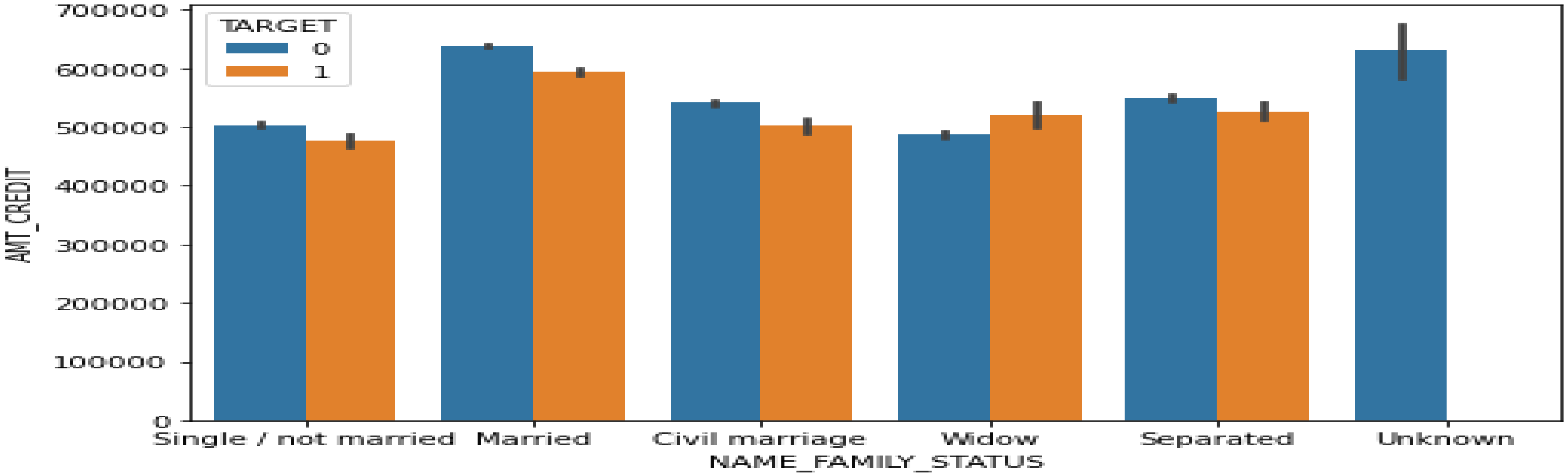


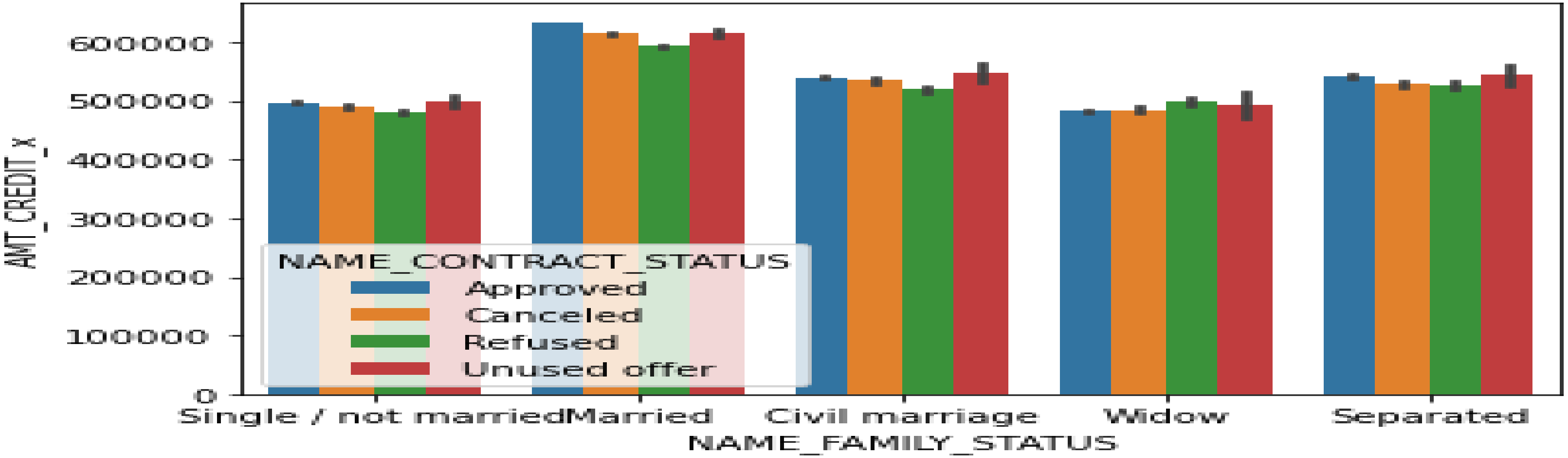
	AMT_ANNUITY_x	AMT_CREDIT_x	AMT_GOODS_PRICE_x	DAYS_EMPLOYED	DAYS_BIRTH
AMT_ANNUITY_x	1	0.78	0.78	-0.084	-0.033
AMT_CREDIT_x	0.78	1	0.99	-0.018	0.048
AMT_GOODS_PRICE_x	0.78	0.99	1	-0.015	0.046
DAYS_EMPLOYED	-0.084	-0.018	-0.015	1	0.67
DAYS_BIRTH	-0.033	0.048	0.046	0.67	1

- There is a high correlation between AMT_CREDIT_x and AMT_GOODS_PRICE_x
- There is a good correlation between AMT_ANNUITY_x and AMT_CREDIT_x
- There is a high correlation between AMT_GOODS_PRICE _x and AMT_ANNUITY_x

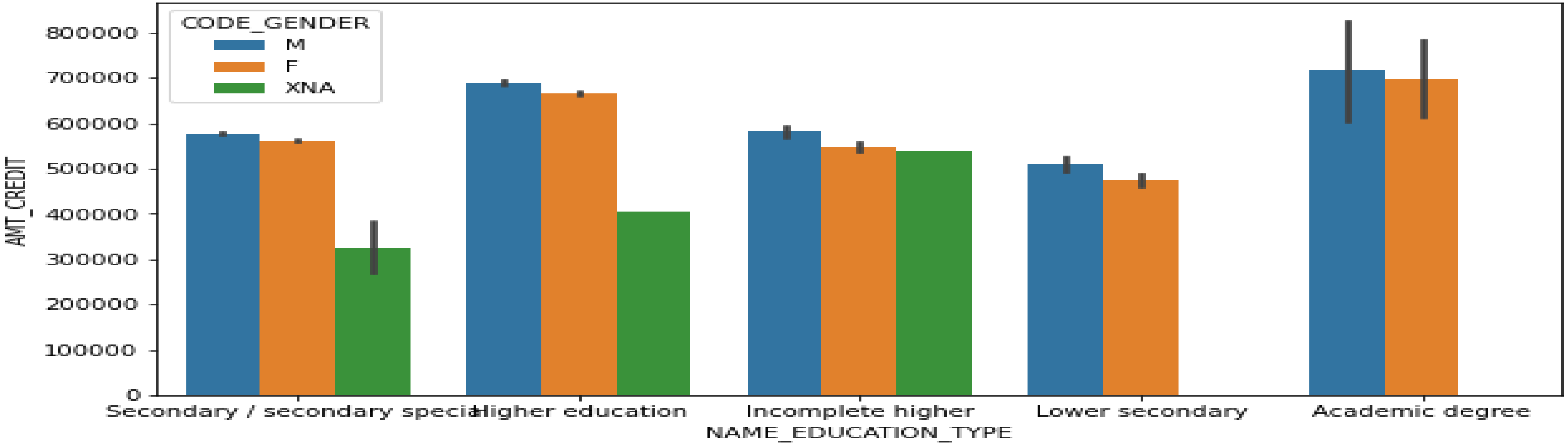
Conclusion

- We can approach the **client category as married people** for granting loans.
- It is clearly indicated from the graph plotted that married people have the highest number of the approval.

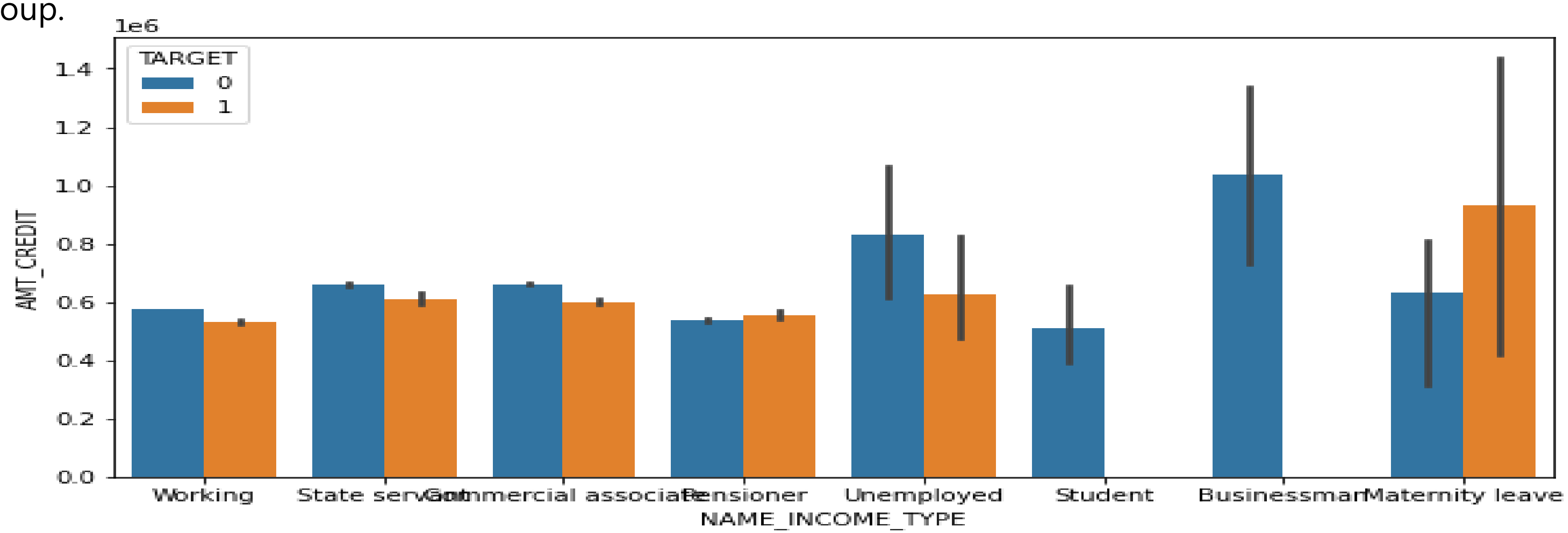




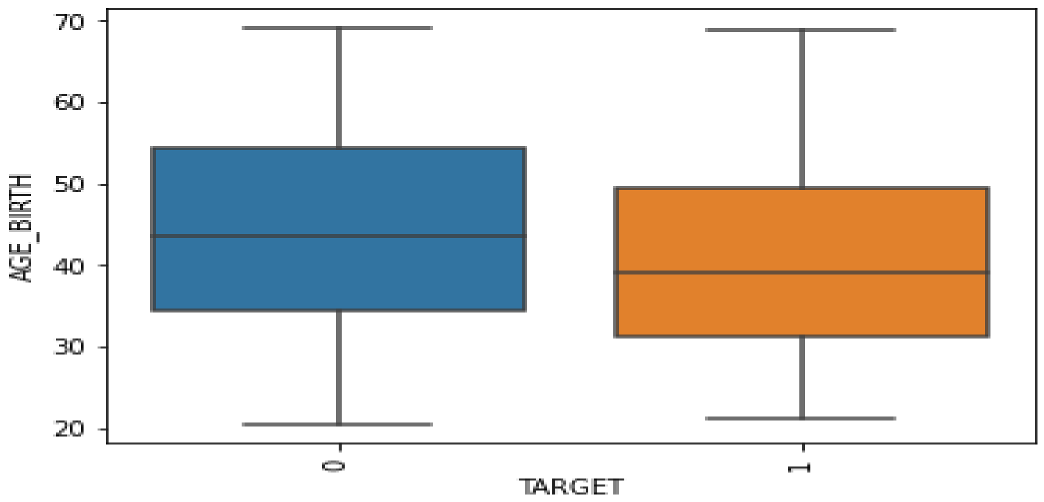
We can approach the **males with academic degrees followed by males with higher education** for granting loans.

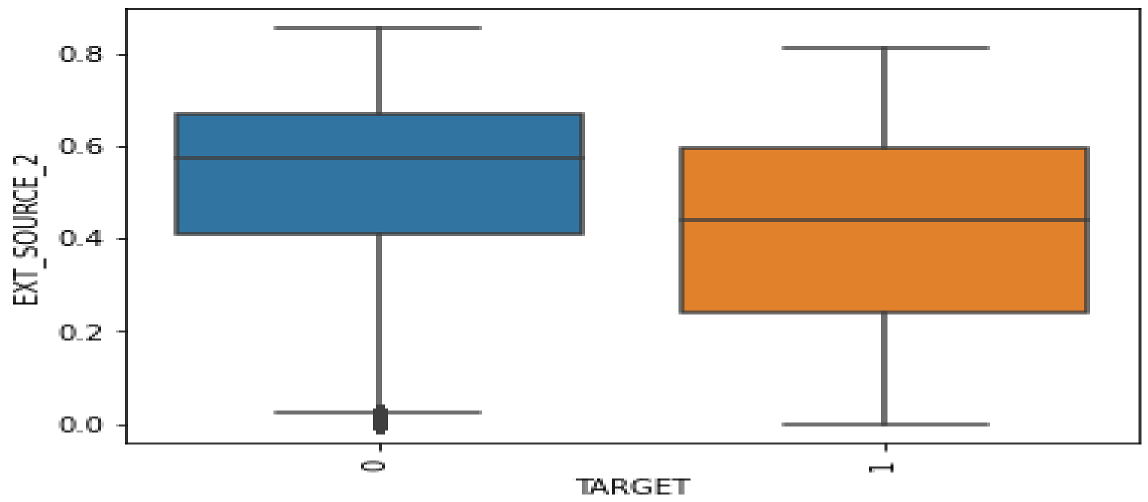


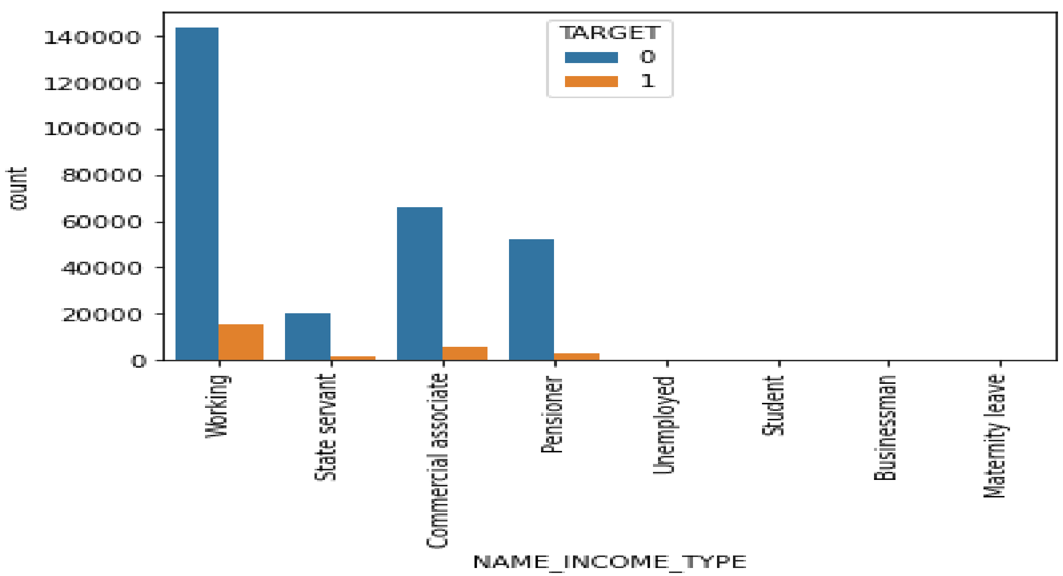
We can approach the **Students & businesses man** for granting loans as we can see that there are no defaulters in this group.

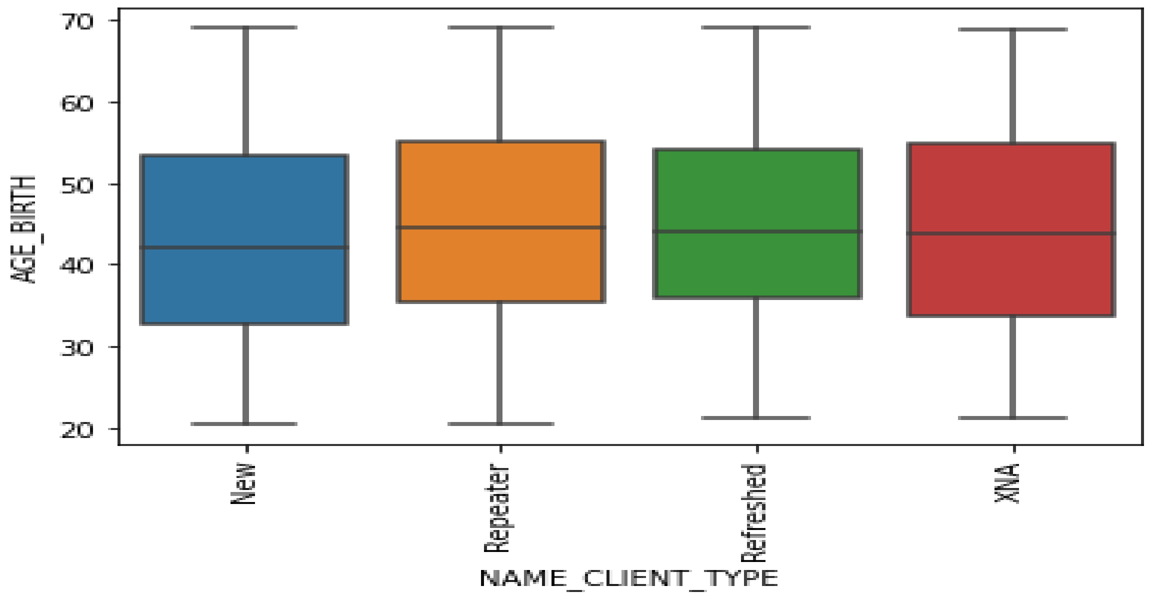


Additional analysis based on other features


AGE_BIRTH
TARGET


EXT_SOURCE_2
TARGET


count
TARGET
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE


AGE_BIRTH
New
Repeater
Refreshed
XNA
NAME_CLIENT_TYPE