EDA ASSIGNMENT

CREDIT RISK ANALYSIS

PROBLEM STATEMENT

Loan Providing company faces challenges for giving loans to people due to there insufficient or nonexistent credit histories.

When evaluating a loan application, the company faces two types of risks:

- If the applicant is likely to repay the loan, rejecting the application results in a loss of business.
- If the applicant is likely to default, approving the loan may lead to a financial loss for the company.

The provided data includes the information about loan application.

- Clients with payment difficulties: Individuals with late payments exceeding a specified number of days on at least one of the initial loan installments.
- Applications where payments are made on time.

For each loan application, there are four potential decisions:

- Approved: The company approves the loan application.
- Cancelled: The client cancels the application during approval, either due to a change of mind or receiving unfavorable pricing.
- Refused: The company rejects the loan application based on client requirements, among other factors.
- Unused offer: The client cancels the loan after approval at different stages of the process.

Result Expected:

- Understanding the data.
- Identify the missing data and handle the missing data.
- Identify the potential outlier.
- Identifying an imbalance in the data.

- Sanity checks.
- Univariate analysis .
- Bivariate analysis.
- Mutivariate analysis.
- Suggestion
- Merging the two datasets.
- Analysis on combined dataset.
- Final Reccomendation.

Understanding Data = understanding every column of the data , identifying the target variable in *application_data.csv and* target variable in *previous_application.csv.*

- *The target variable in application_data.csv is* **TARGET.**
- *The target variable in previous_application.csv is* **NAME_CONTRACT_STATUS.**

Identifying the missing data and handle missing = We can identify the missing data by using syntax.

(data.isnull().value_counts()/data.shape[0])*100

- **data** here represent the name of the dataframe, data.shape give the length of row and then multiplying by 100 give percentage of null value in each column.
- We have the option to remove columns with 40% or more null values, and approximately 0.567% of the data can also be deleted.

Checking for Outlier = We can check the outlier by using boxplot. Identifying the potential outlier.

- We can drop certain percentage of data point with extreme values from the extreme ends
- We can bin the data into group or intervals
- If outlier due to missing value, we can impute or replacing them with appropriate values.

Identifying imbalance in data = We can identify the imbalance of data by examining the distribution of the target variable.

- For calculating imbalance in ratio we can divide the number of instances in the minority class by the majority class

- We can use visualizations such as bar plots, pie charts, or histograms to represent the distribution of classes.

Sanity Checks = We can do some basic sanity checks whether is there any irregularities in the data.

Univariate Analysis = univariate analysis means analysis on single variable.

- We can use Visualization like line chart, boxplot, countplot, histogram.

Bivariate Analysis = Bivariate analysis means analysis on two variable

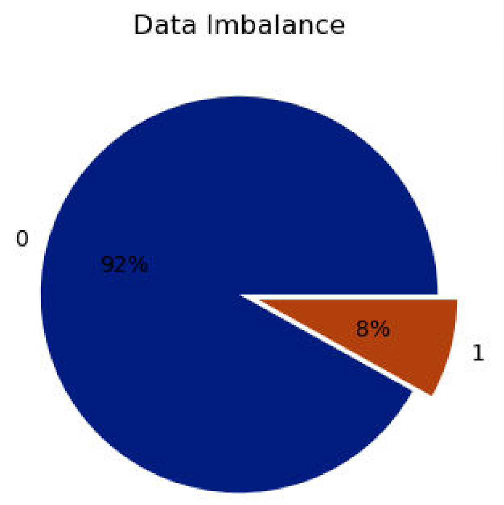
- We can use visualization like Scatter plots, Line charts, Bubble charts , Heatmaps, Stacked Bar Chart, Grouped Bar charts, Pair Plots

Multivariate Analysis = Multivariate Analysis means analysis on more than two variable

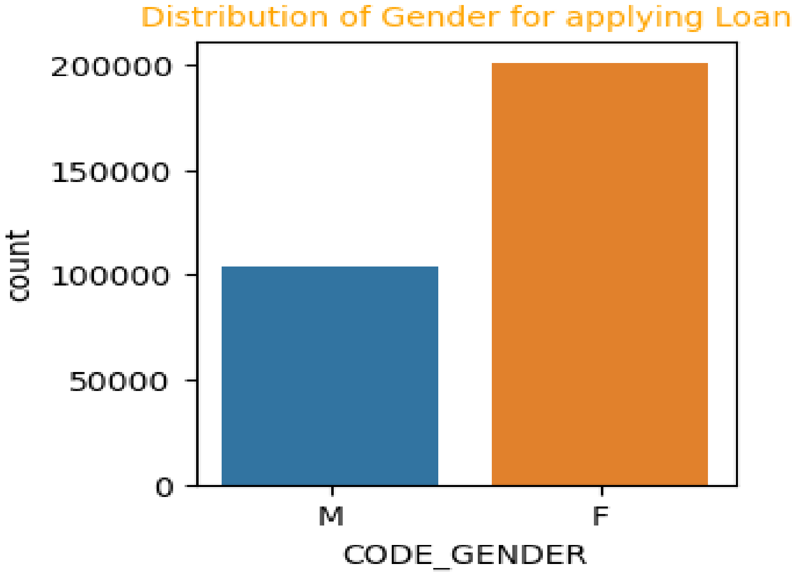
- Visualization we can use Heatmap,pairplot etc.
- Suggestion Any suggestion after analysing the partical data
- Combining Both the dataframe we can join the two dataframe and analyze the Data
- Final recommendation

Graphs and Insight

- Checking the data imbalance we can defaulter is less then the non
- defaulter and the imbalance ratio is 11.6%

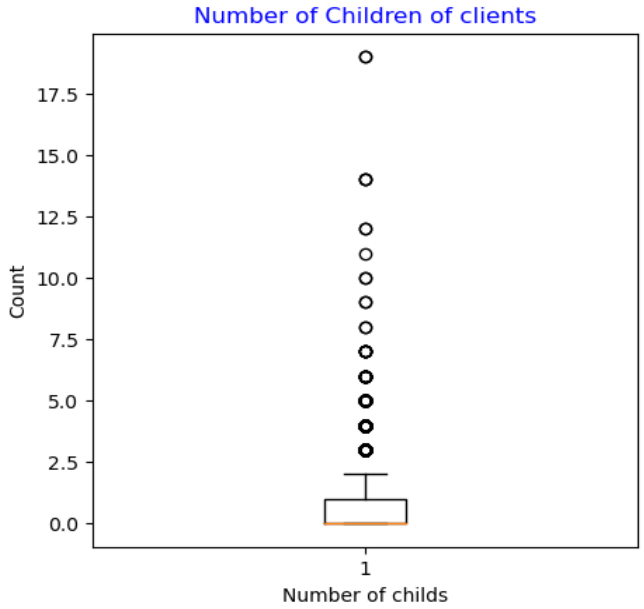


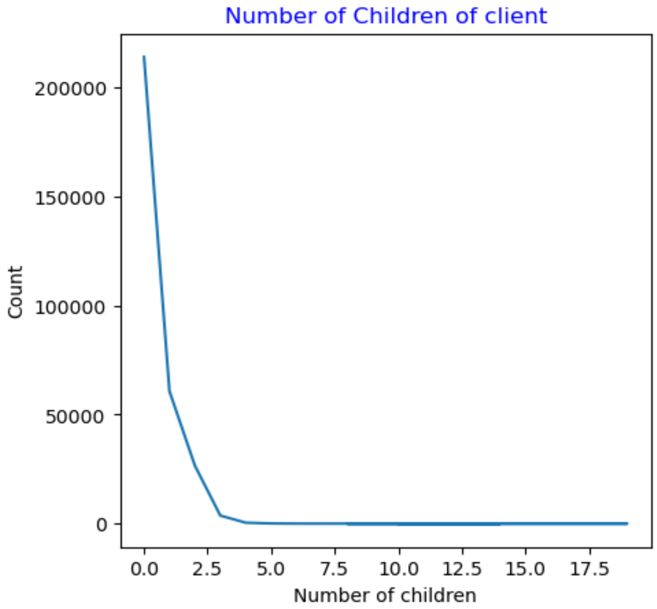
- Females are more prone to applying loan then female



Children Counts

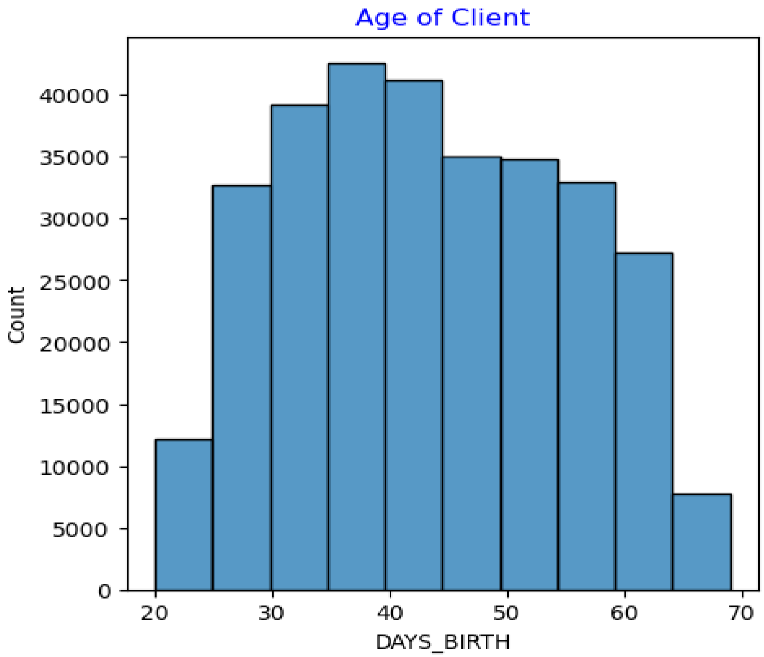
- Most of clients having 0 children

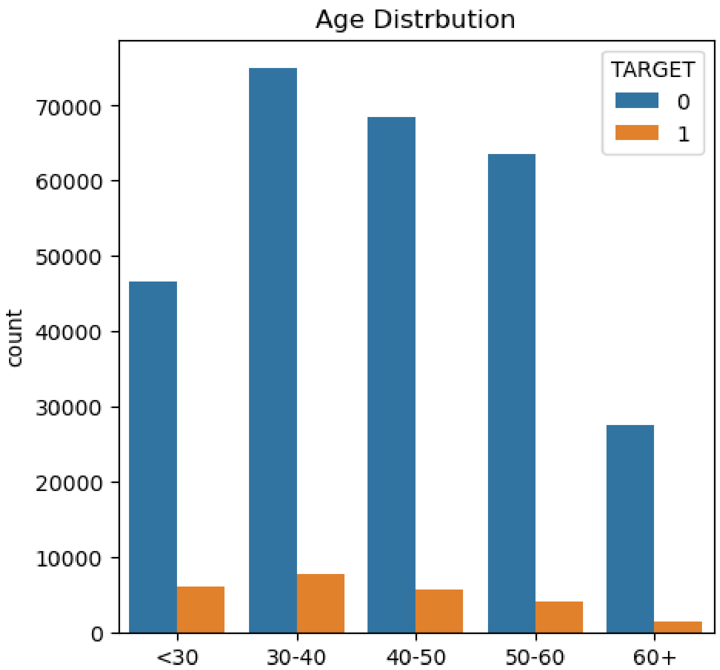


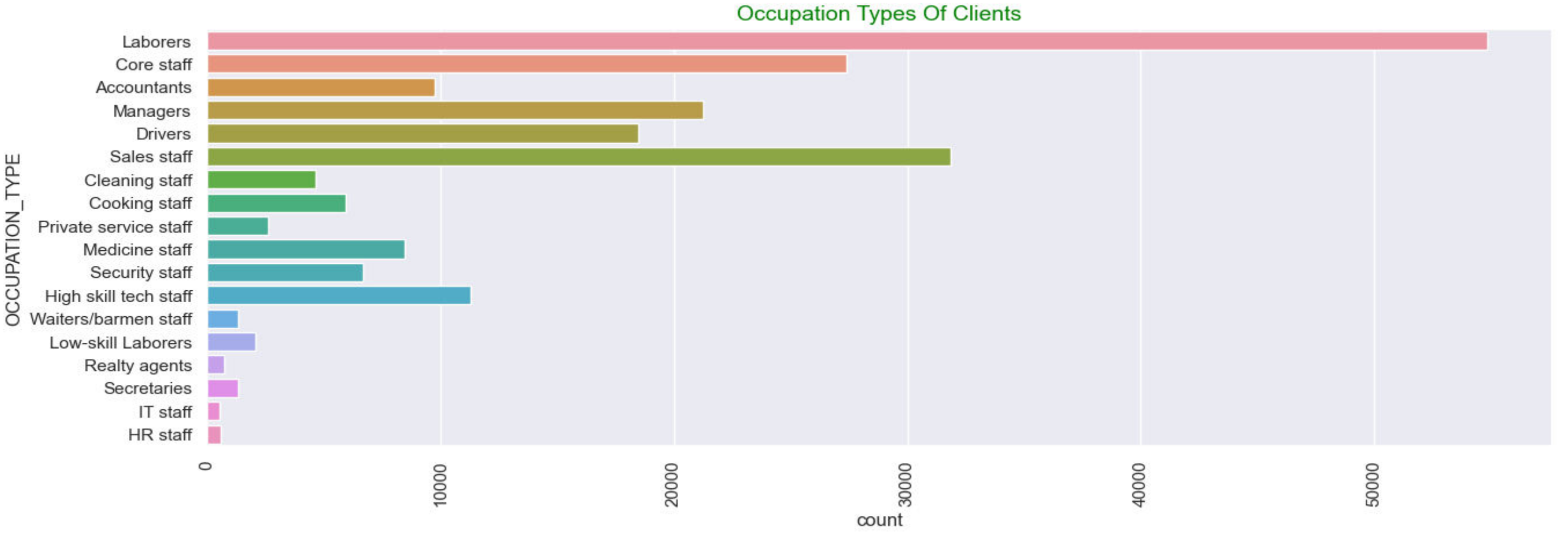


- Most of clients belong to age group between 30 to 40 yrs, this age group is more to default in loan payement.

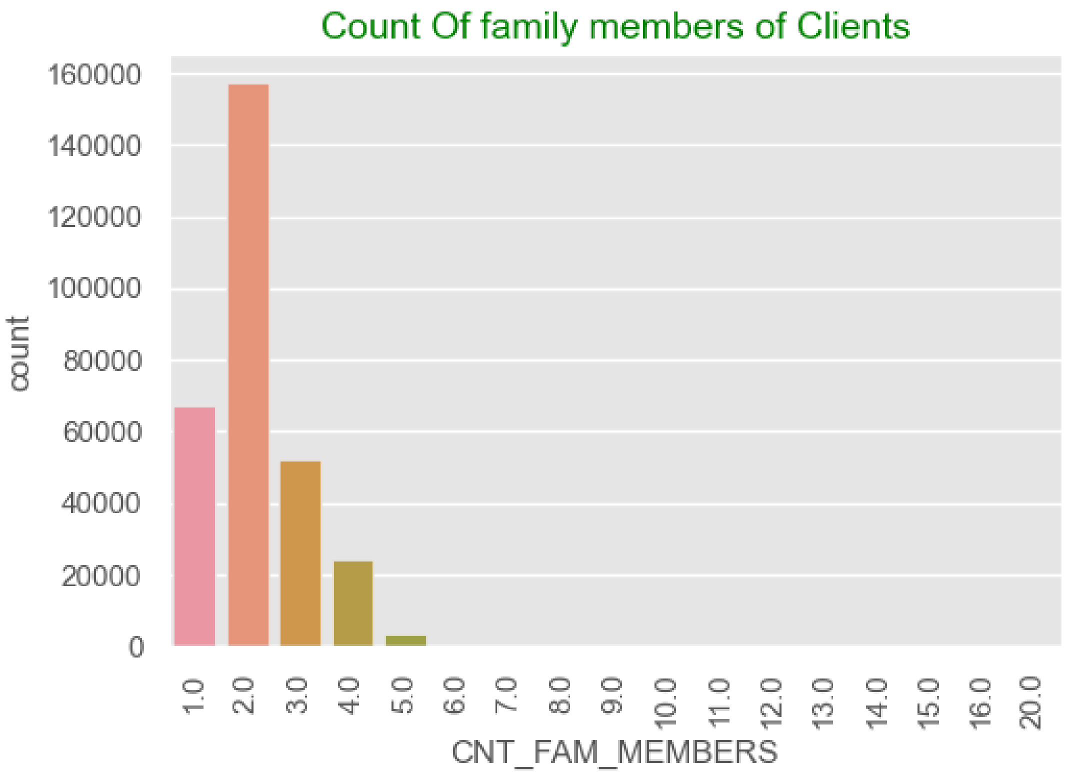
Distribution of data


Age of Client


Age Distrbution

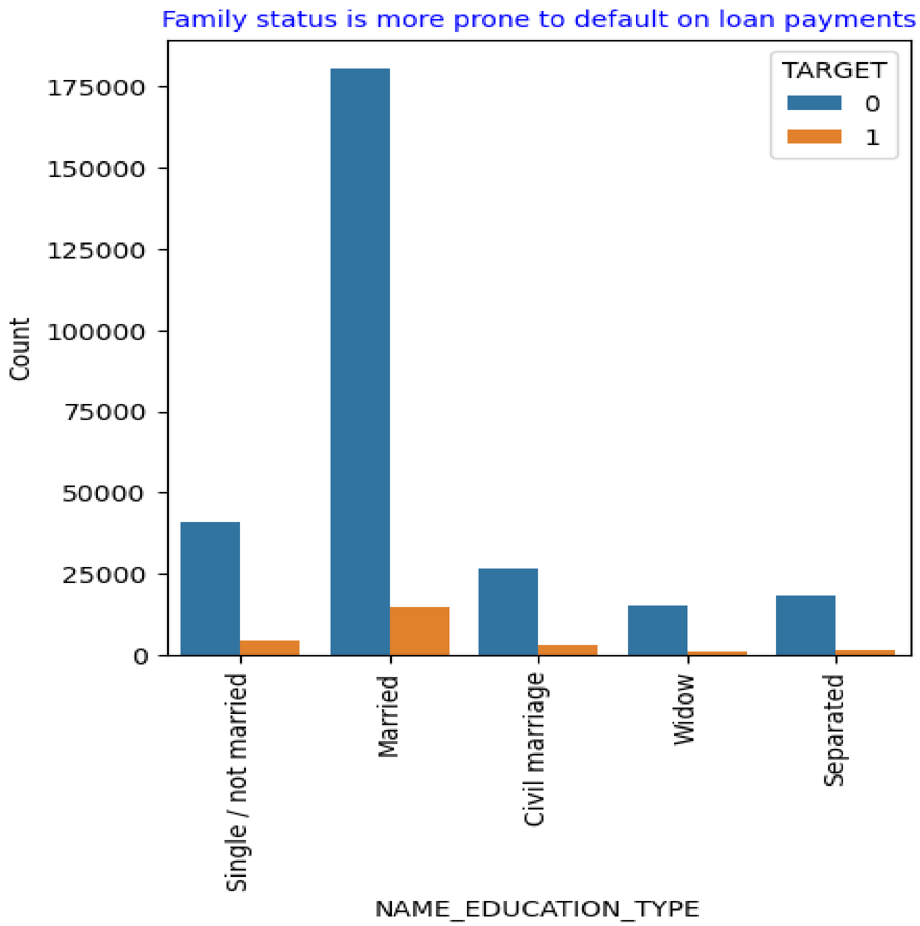

Occupation Types Of Clients

- Most of the client have occupaton type laborer.

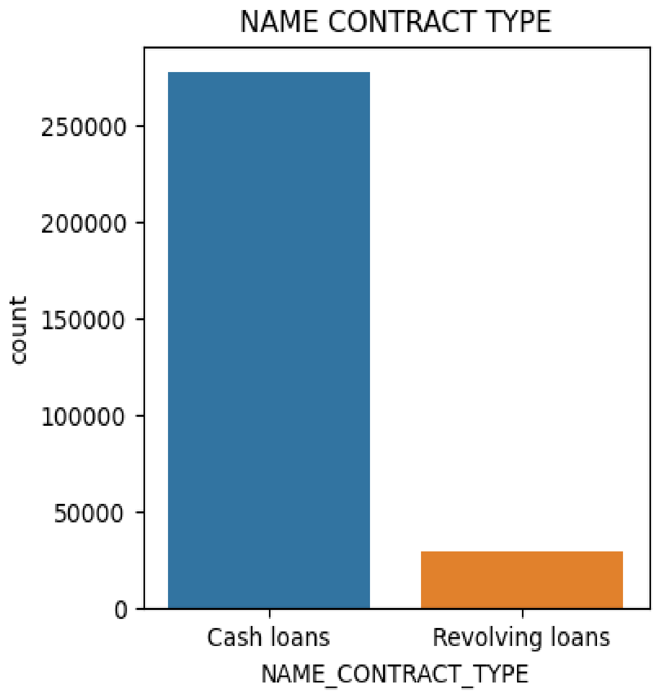


- Count of the family mostly clients are having 2 members

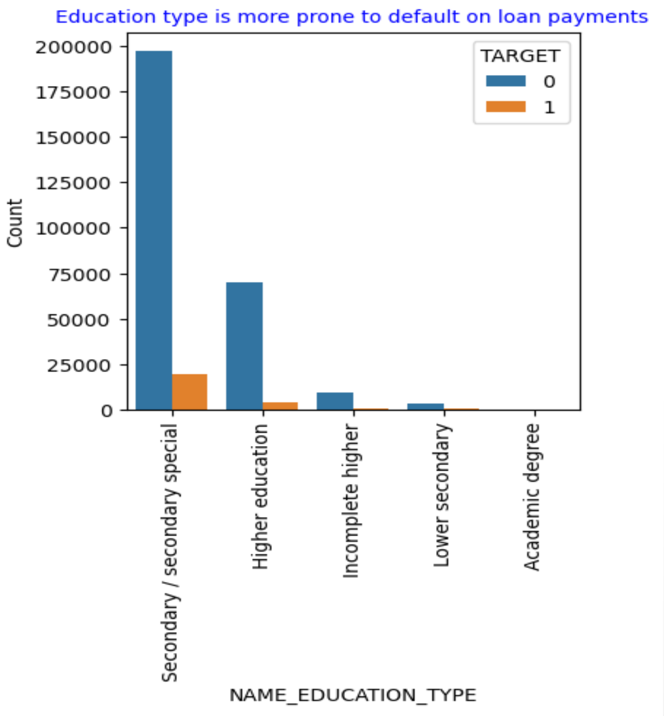
- Most of the client have family status ids married and the married peoples are the one who are more prone to take the loan.



- Most of client have contract type is cash loans.

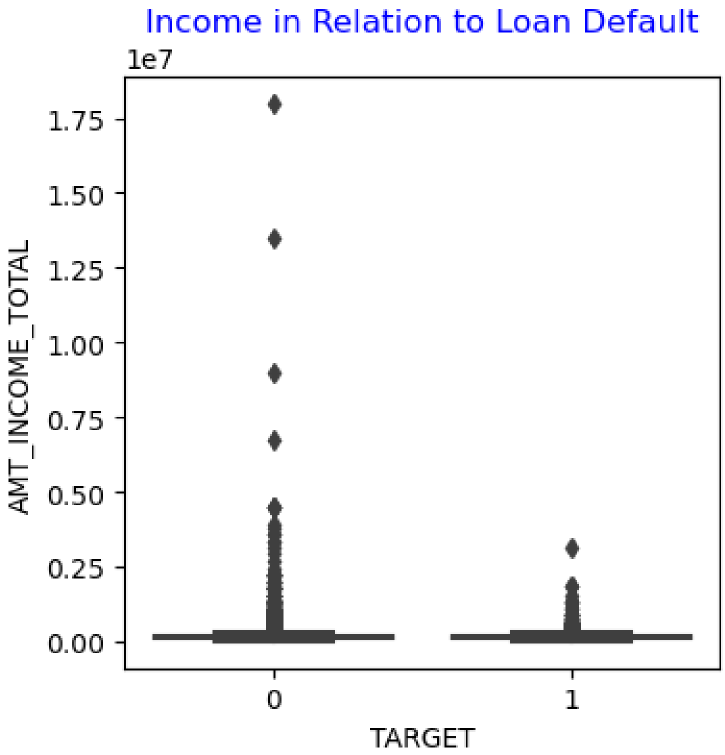


- Most of the client have education type is Secondary



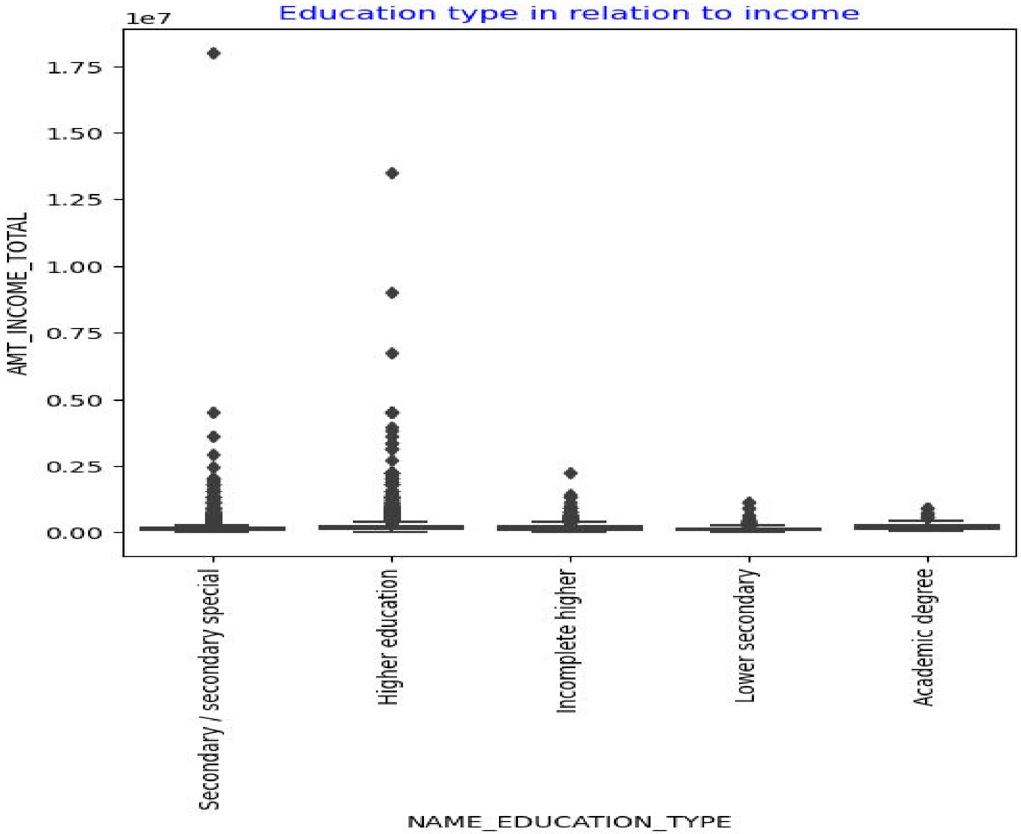
Income and Credit in relation to default

- Low income people are face more to pay there loan.

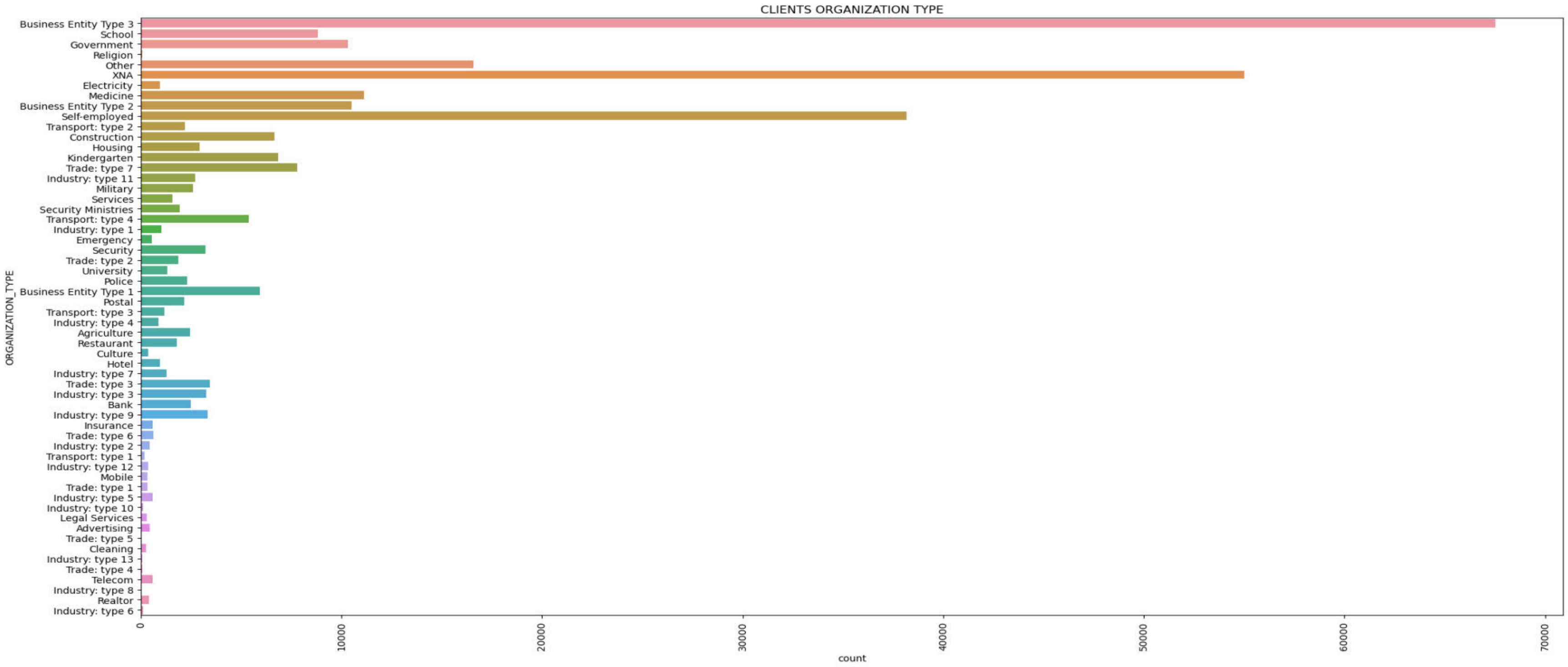


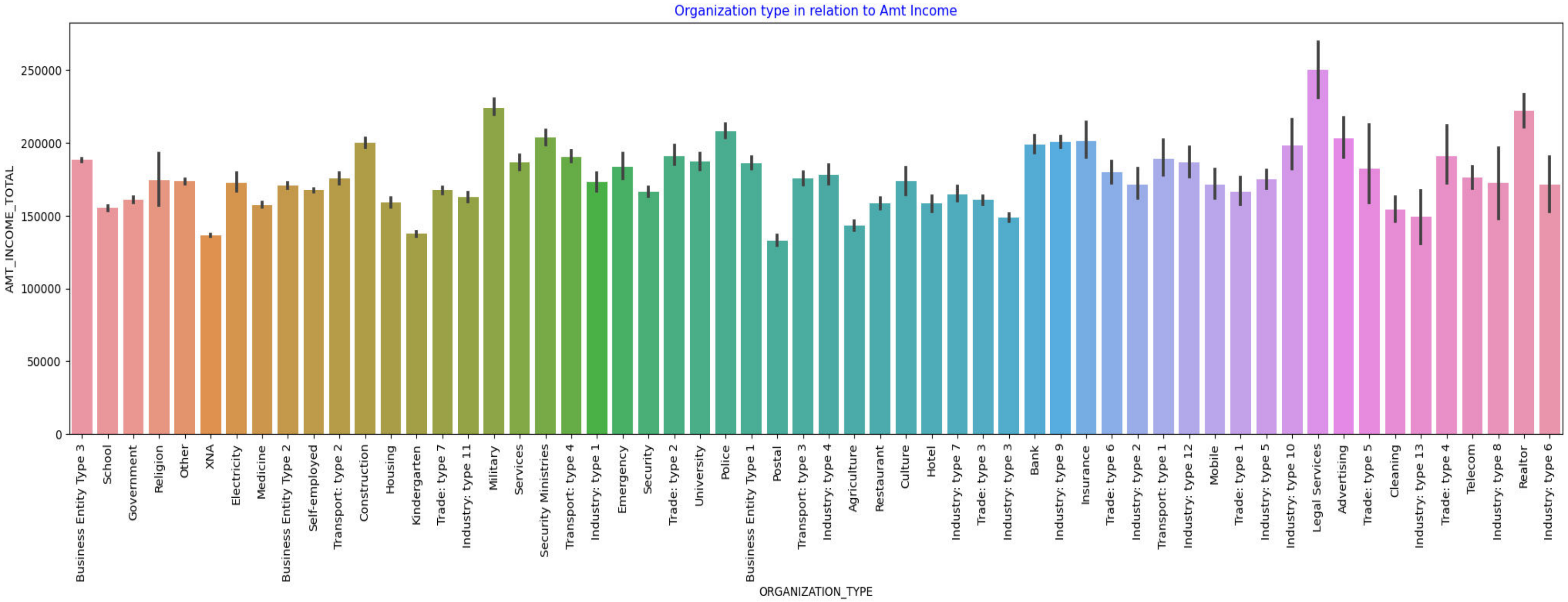
- Credit amount in relation to default of payment

- Relationship between education and income secondary and higher educated clients are earning more.

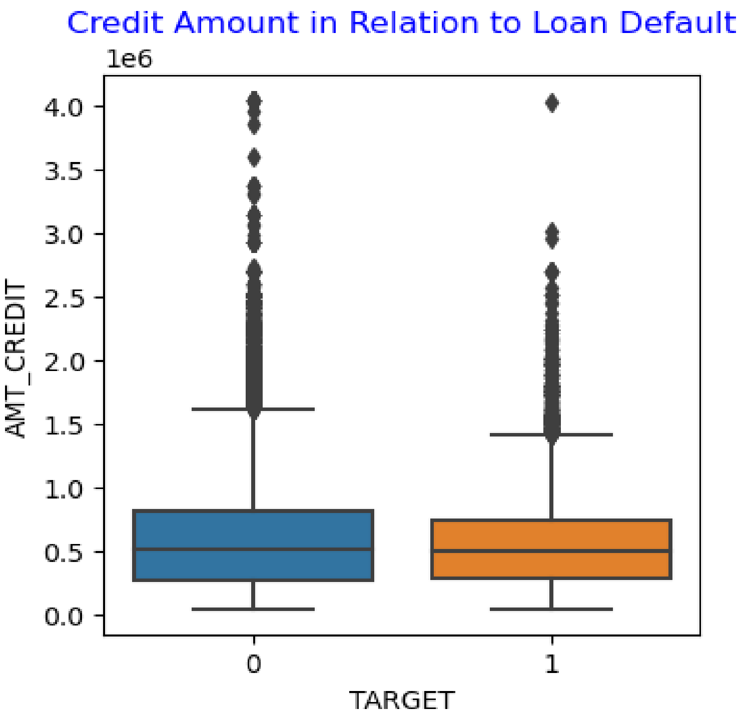


- Relationship between organization type and income and which organization type are more prone to take loan

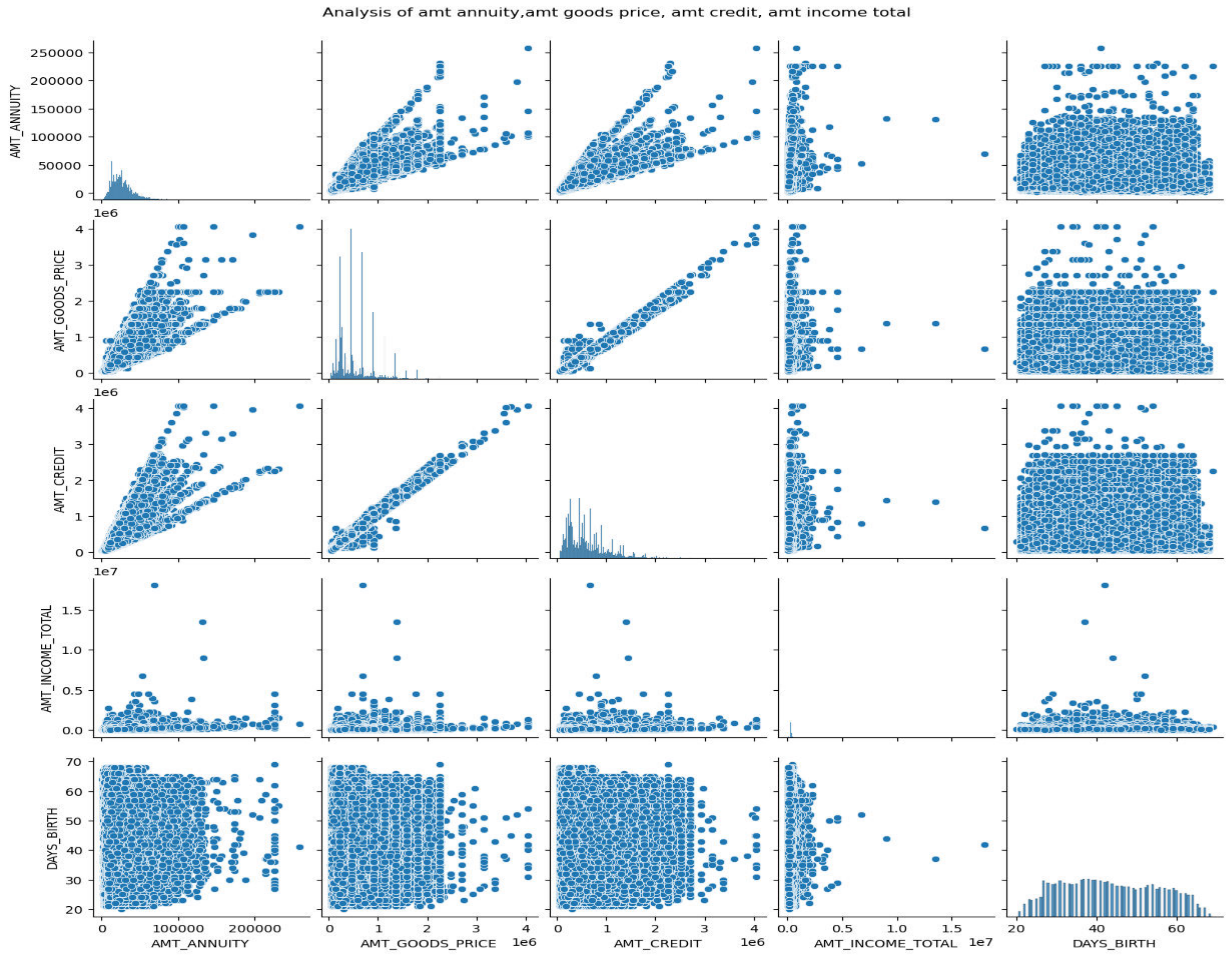

CLIENTS ORGANIZATION TYPE
ORGANIZATION_TYPE
count
Business Entity Type 3
School
Government
Religion
Other
XNA
Electricity
Medicine
Business Entity Type 2
Self-employed
Transport: type 2
Construction
Housing
Kindergarten
Trade: type 7
Industry: type 11
Military
Services
Security Ministries
Transport: type 4
Industry: type 1
Emergency
Security
Trade: type 2
University
Police
Business Entity Type 1
Postal
Transport: type 3
Industry: type 4
Agriculture
Restaurant
Culture
Hotel
Industry: type 7
Trade: type 3
Industry: type 3
Bank
Industry: type 9
Insurance
Trade: type 6
Industry: type 2
Transport: type 1
Industry: type 12
Mobile
Trade: type 1
Industry: type 5
Industry: type 10
Legal Services
Advertising
Trade: type 5
Cleaning
Industry: type 13
Trade: type 4
Telecom
Industry: type 8
Realtor
Industry: type 6
0
10000
20000
30000
40000
50000
60000
70000


Organization type in relation to Amt Income
AMT_INCOME_TOTAL
250000
200000
150000
100000
50000
0
ORGANIZATION_TYPE
Business Entity Type 3
School
Government
Religion
Other
XNA
Electricity
Medicine
Business Entity Type 2
Self-employed
Transport: type 2
Construction
Housing
Kindergarten
Trade: type 7
Industry: type 11
Military
Services
Security Ministries
Transport: type 4
Industry: type 1
Emergency
Security
Trade: type 2
University
Police
Business Entity Type 1
Postal
Transport: type 3
Industry: type 4
Agriculture
Restaurant
Culture
Hotel
Industry: type 7
Trade: type 3
Industry: type 3
Bank
Industry: type 9
Insurance
Trade: type 6
Industry: type 2
Transport: type 1
Industry: type 12
Mobile
Trade: type 1
Industry: type 5
Industry: type 10
Legal Services
Advertising
Trade: type 5
Cleaning
Industry: type 13
Trade: type 4
Telecom
Industry: type 8
Realtor
Industry: type 6

- Relationship between credit and amount to loan default



- Comparision of all the numerical variable of the application data.



Analysis of amt annuity,amt goods price, amt credit, amt income total

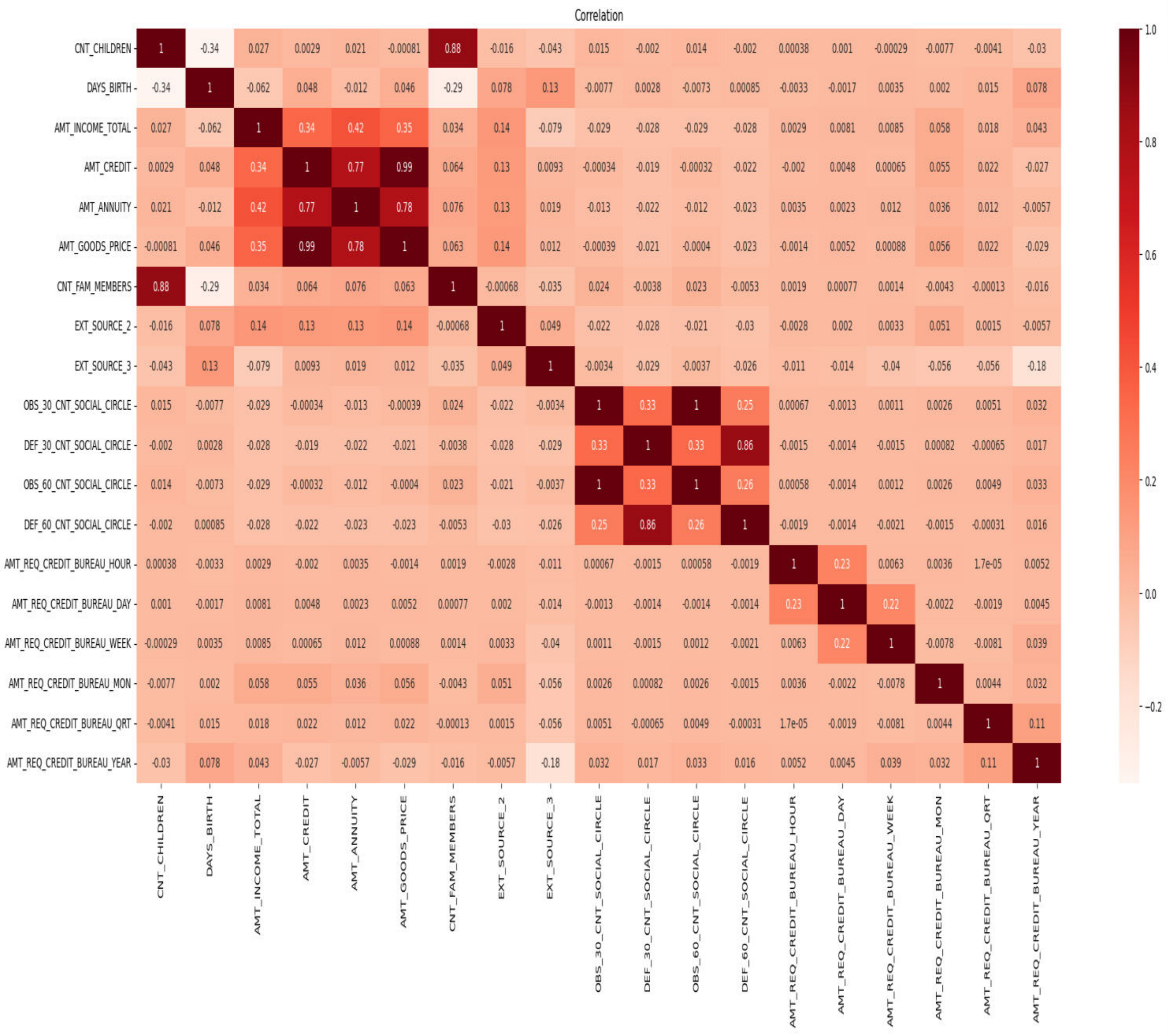
- What we infer from above pair plot

  - The correlation between an increase in the annuity amount and an increase in age is not very apparent in other words it is not evident higher the amount annuity higher the age.

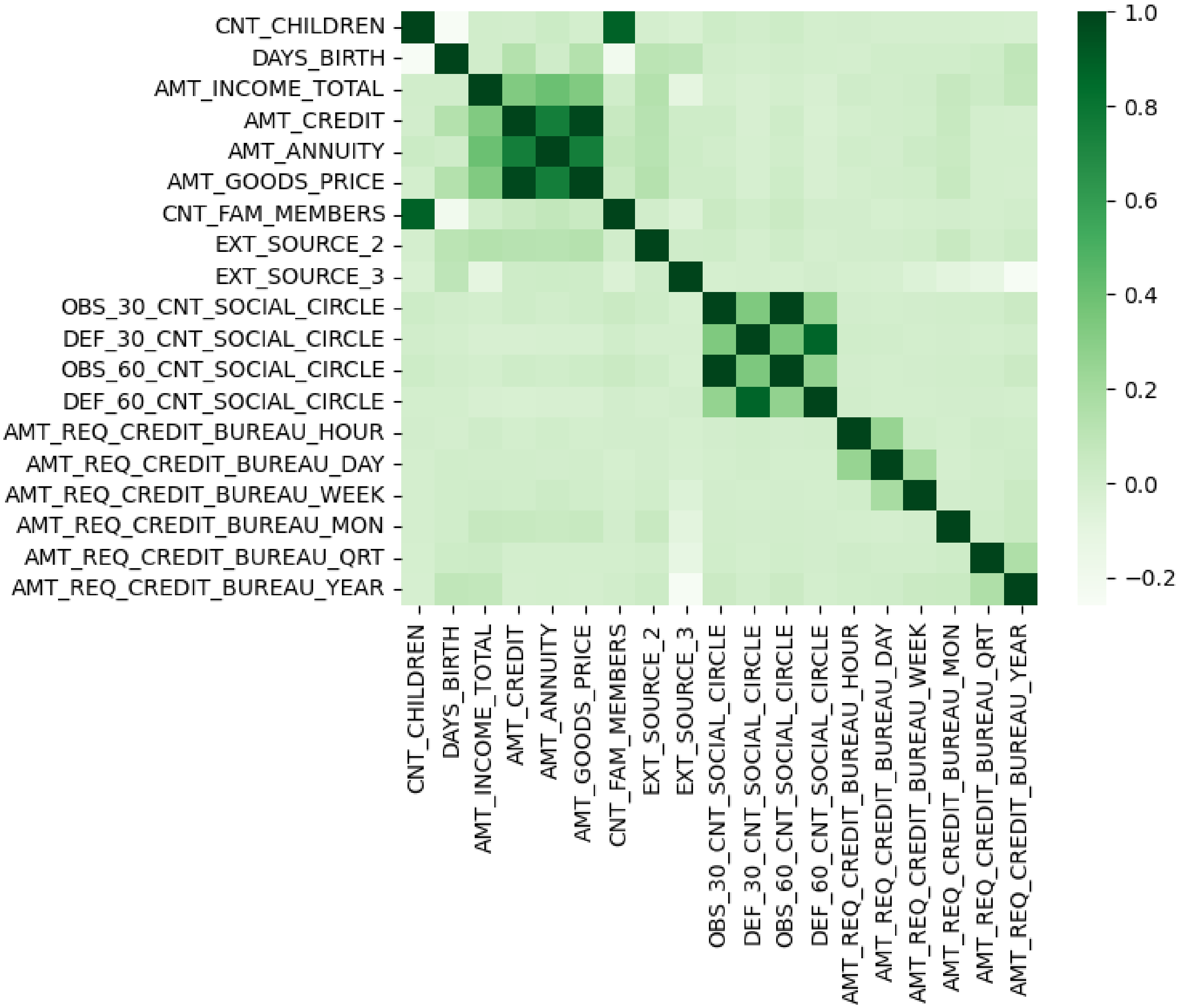
  - AMT_ANNUITY and AMT_INCOME have very weak correlation but AMT_CREDIT(Credit amount of the loan) and AMT_GOODS_PRICE(it is the price of the goods for which the loan is given) have strong linear correlation

  - Increase in good price and increase in credit amount increase in amount of annuity.

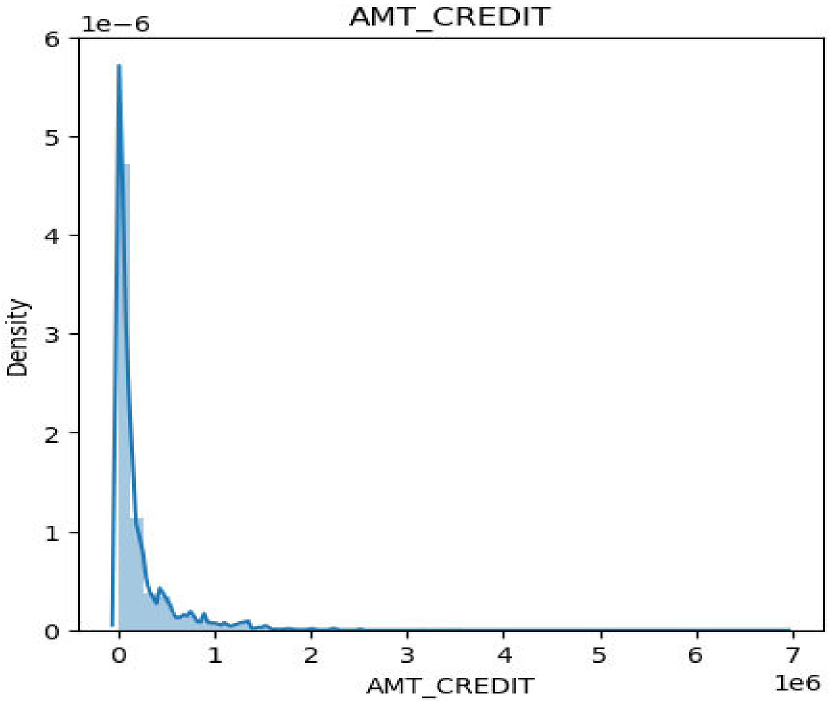
- Finding the correlation for on the basic of target variable
- Divided the targeted variable into two parts Defaulters and no defaulters
- Defaulters represent by 1.
- Non Defaulters Represent by 0.



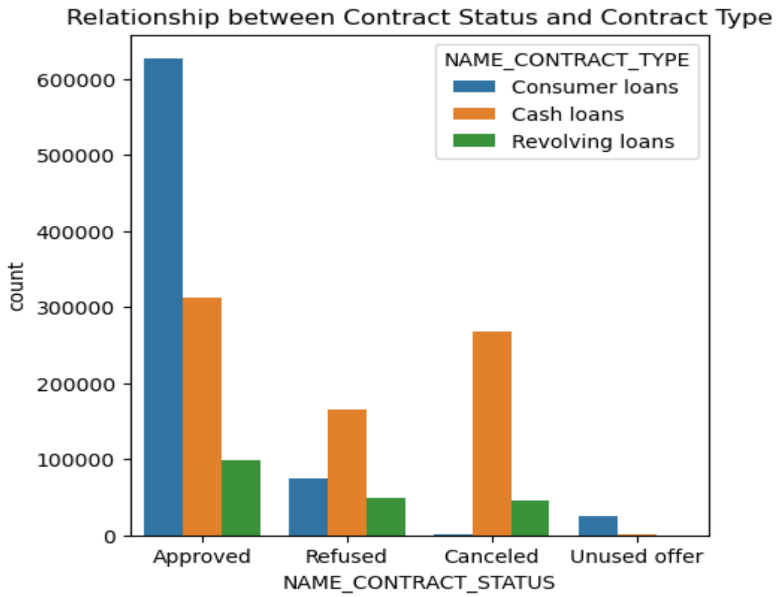
	CNT_CHILDREN	DAYS_BIRTH	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_FAM_MEMBERS	EXT_SOURCE_2	EXT_SOURCE_3	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1	-0.34	0.027	0.0029	0.021	-0.00081	0.88	-0.016	-0.043	0.015	-0.002	0.014	-0.002	0.00038	0.001	-0.00029	-0.0077	-0.0041	-0.03
DAYS_BIRTH	-0.34	1	-0.062	0.048	-0.012	0.046	-0.29	0.078	0.13	-0.0077	0.0028	-0.0073	0.00085	-0.0033	-0.0017	0.0035	0.002	0.015	0.078
AMT_INCOME_TOTAL	0.027	-0.062	1	0.34	0.42	0.35	0.034	0.14	-0.079	-0.029	-0.028	-0.029	-0.028	0.0029	0.0081	0.0085	0.058	0.018	0.043
AMT_CREDIT	0.0029	0.048	0.34	1	0.77	0.99	0.064	0.13	0.0093	-0.00034	-0.019	-0.00032	-0.022	-0.002	0.0048	0.00065	0.055	0.022	-0.027
AMT_ANNUITY	0.021	-0.012	0.42	0.77	1	0.78	0.076	0.13	0.019	-0.013	-0.022	-0.012	-0.023	0.0035	0.0023	0.012	0.036	0.012	-0.0057
AMT_GOODS_PRICE	-0.00081	0.046	0.35	0.99	0.78	1	0.063	0.14	0.012	-0.00039	-0.021	-0.0004	-0.023	-0.0014	0.0052	0.00088	0.056	0.022	-0.029
CNT_FAM_MEMBERS	0.88	-0.29	0.034	0.064	0.076	0.063	1	-0.00068	-0.035	0.024	-0.0038	0.023	-0.0053	0.0019	0.00077	0.0014	-0.0043	-0.00013	-0.016
EXT_SOURCE_2	-0.016	0.078	0.14	0.13	0.13	0.14	-0.00068	1	0.049	-0.022	-0.028	-0.021	-0.03	-0.0028	0.002	0.0033	0.051	0.0015	-0.0057
EXT_SOURCE_3	-0.043	0.13	-0.079	0.0093	0.019	0.012	-0.035	0.049	1	-0.0034	-0.029	-0.0037	-0.026	-0.011	-0.014	-0.04	-0.056	-0.056	-0.18
OBS_30_CNT_SOCIAL_CIRCLE	0.015	-0.0077	-0.029	-0.00034	-0.013	-0.00039	0.024	-0.022	-0.0034	1	0.33	1	0.25	0.00067	-0.0013	0.0011	0.0026	0.0051	0.032
DEF_30_CNT_SOCIAL_CIRCLE	-0.002	0.0028	-0.028	-0.019	-0.022	-0.021	-0.0038	-0.028	-0.029	0.33	1	0.33	0.86	-0.0015	-0.0014	-0.0015	0.00082	-0.00065	0.017
OBS_60_CNT_SOCIAL_CIRCLE	0.014	-0.0073	-0.029	-0.00032	-0.012	-0.0004	0.023	-0.021	-0.0037	1	0.33	1	0.26	0.00058	-0.0014	0.0012	0.0026	0.0049	0.033
DEF_60_CNT_SOCIAL_CIRCLE	-0.002	0.00085	-0.028	-0.022	-0.023	-0.023	-0.0053	-0.03	-0.026	0.25	0.86	0.26	1	-0.0019	-0.0014	-0.0021	-0.0015	-0.00031	0.016
AMT_REQ_CREDIT_BUREAU_HOUR	0.00038	-0.0033	0.0029	-0.002	0.0035	-0.0014	0.0019	-0.0028	-0.011	0.00067	-0.0015	0.00058	-0.0019	1	0.23	0.0063	0.0036	1.7e-05	0.0052
AMT_REQ_CREDIT_BUREAU_DAY	0.001	-0.0017	0.0081	0.0048	0.0023	0.0052	0.00077	0.002	-0.014	-0.0013	-0.0014	-0.0014	-0.0014	0.23	1	0.22	-0.0022	-0.0019	0.0045
AMT_REQ_CREDIT_BUREAU_WEEK	-0.00029	0.0035	0.0085	0.00065	0.012	0.00088	0.0014	0.0033	-0.04	0.0011	-0.0015	0.0012	-0.0021	0.0063	0.22	1	-0.0078	-0.0081	0.039
AMT_REQ_CREDIT_BUREAU_MON	-0.0077	0.002	0.058	0.055	0.036	0.056	-0.0043	0.051	-0.056	0.0026	0.00082	0.0026	-0.0015	0.0036	-0.0022	-0.0078	1	0.0044	0.032
AMT_REQ_CREDIT_BUREAU_QRT	-0.0041	0.015	0.018	0.022	0.012	0.022	-0.00013	0.0015	-0.056	0.0051	-0.00065	0.0049	-0.00031	1.7e-05	-0.0019	-0.0081	0.0044	1	0.11
AMT_REQ_CREDIT_BUREAU_YEAR	-0.03	0.078	0.043	-0.027	-0.0057	-0.029	-0.016	-0.0057	-0.18	0.032	0.017	0.033	0.016	0.0052	0.0045	0.039	0.032	0.11	1


CNT_CHILDREN
DAYS_BIRTH
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
CNT_FAM_MEMBERS
EXT_SOURCE_2
EXT_SOURCE_3
OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
AMT_REQ_CREDIT_BUREAU_HOUR
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN
DAYS_BIRTH
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
CNT_FAM_MEMBERS
EXT_SOURCE_2
EXT_SOURCE_3
OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
AMT_REQ_CREDIT_BUREAU_HOUR
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_YEAR
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

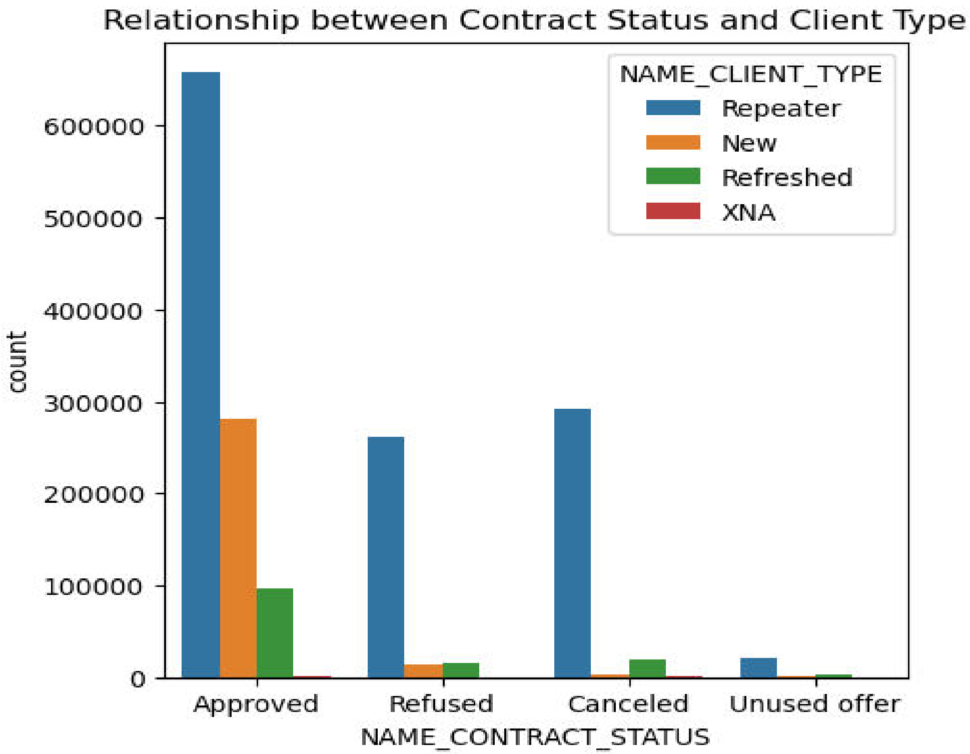
- Most of the client amount credit between 0 to 100000.



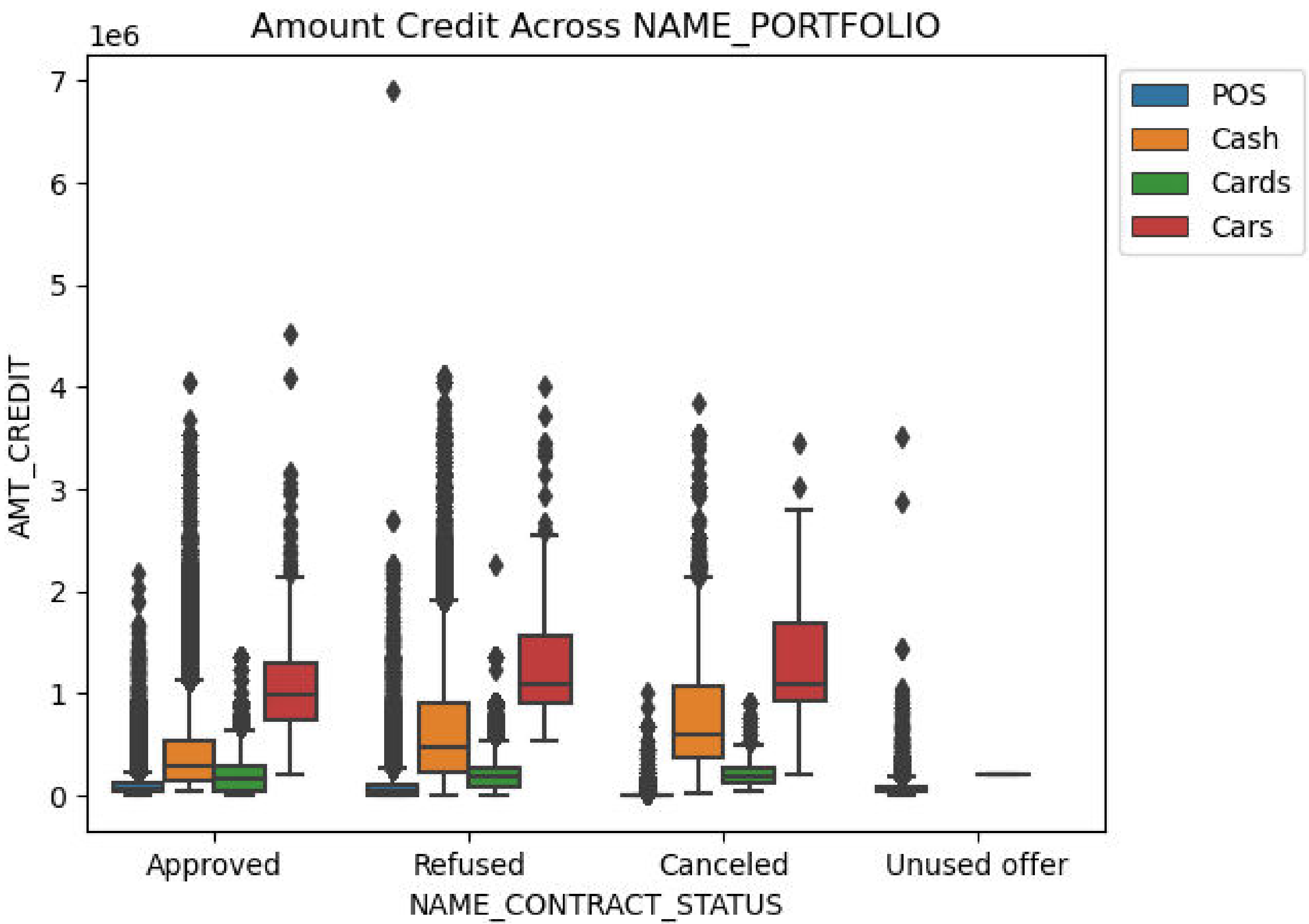
- Highest number of loan approved is consumer loan and highest numer of loan is rejested and canceled is cash loan.



- Most of the client type is Repeater. Repeater loans are most likely to approved.

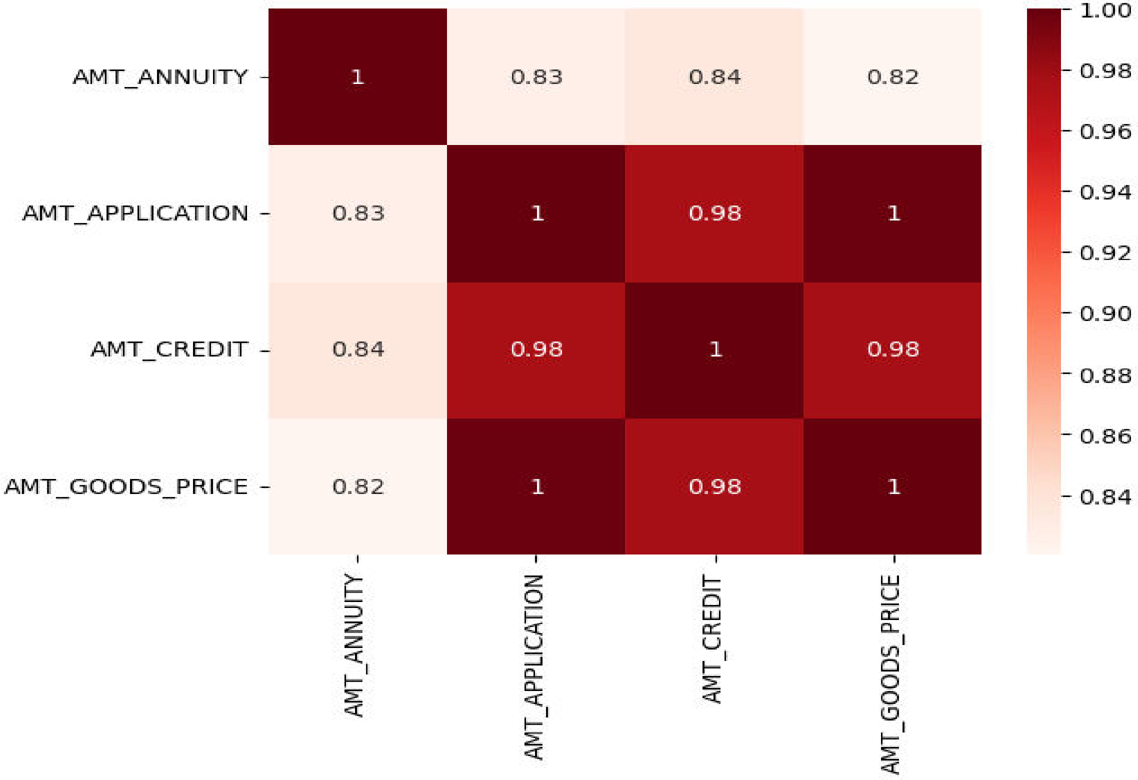


- Considerable number of cash loans face rejection.
- Significant portion of cash loans are canceled.


Amount Credit Across NAME_PORTFOLIO
1e6
7
6
5
4
3
2
1
0
AMT_CREDIT
Approved
Refused
Canceled
Unused offer
NAME_CONTRACT_STATUS
POS
Cash
Cards
Cars

Correlation of numerical variable of previous data

- All the columns AMT_ANNUITY, AMT_GOODS_PRICE, AMT_CREDIT, AMT_APPLICATION] have strong positive linear correlation with each other which is evident in below heatmap.



Final Observation

- 54.99% of loan are approved in case of defaulters
- Education types secondary special have higher rate of loan approved
- Working professional have higher rate of loan approved
- Age group 30 -40 higher rate of loan approved
- Managers tend to request larger credit amounts but have a higher incidence of defaulting on loan payments.

Final Conclusion

- females are applying more for loans in compared to men
- age group between 30 to 40 are applying more for loans.
- married peoples are applying for more loan
- Individuals who apply for multiple loans typically have a family size of two and do not have any children.

From above all the conclusion ***

Probably Newly Married people age group between 30 to 40   are applying more for loan and having no child.

THANK YOU

DEEPANJANA ROY