EDA Assignment

VEERA KUMAR. M

EDA Assignment:

Bank loan data sets has been given to analyse the candidates applied for loan. Key feature is to analyse the candidate details like occupation, education, employed years and age etc., Basis approval has to given to the client / customer or not.

Following key things.

1. Data should be loaded in python notebook and data structure should be completely analysed to find the null values.
2. Once null values have been identified, some data in the data frame has to be removed based on the threshold.
3. Then outliers in the plot needs to be managed by techniques and visualisation.
4. Then visualisation needs to be carried out to show relationship between the data to relate with business requirement of loan approval sanction.

Business Understanding

When the company receives a loan application, the company has to decide for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision:

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

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- **All other cases:** All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

1. **Approved:** The Company has approved loan Application
2. **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.
3. **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
4. **Unused offer:** Loan has been cancelled by the client but at different stages of the process.

THOUGHT PROCESS

Exporting Data:

- Load the given Data (csv) file in the jupyter python notebook

Understanding of Data:

- Analyze the data structure and model based on size, shape, info and describe features in python

Data Cleaning

- Identification of missing values
- Removal of missing values based on Threshold percentage (50%). If a column in a data frame contains more than 50% missing values, those columns has to be removed.
- Once the missing values are imputed with appropriate values , outliers needs to identified.
- Treatment of outliers using flooring and capping

Data Analysis:

- ❑ Data imbalance has to be measured to segregate the data frame based on key column.
- ❑ Plot the data in univariate , bivariate , segmented variate and multivariate.
- ❑ Understanding the relationship between the different numerical and categorical columns.
- ❑ Understand the data correlation.

Conclusion :

1. Using the past data and current data analyse the approval methods and recommend application categories to be approved

Data preparation for analysis:

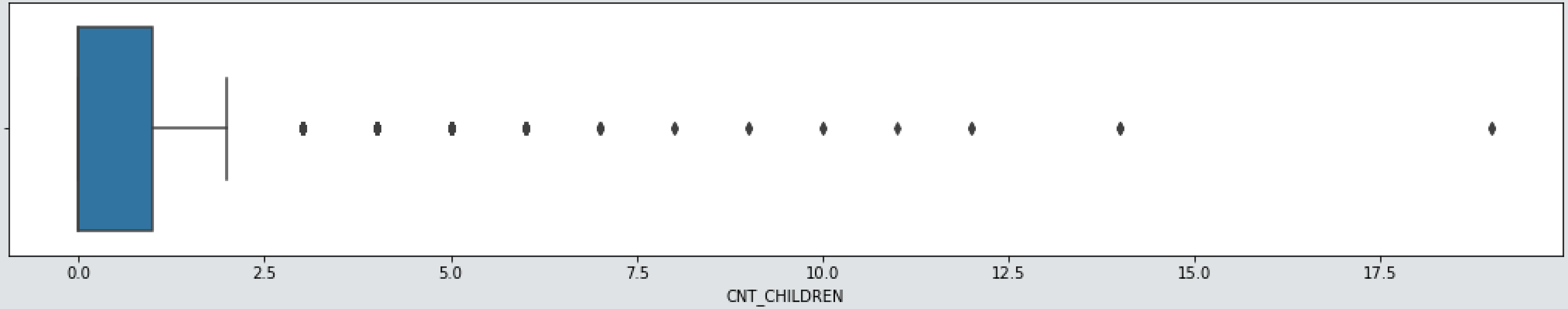
1. Export the data.
2. Identify the missing values and remove the columns which are having missing values greater than 50%
3. Analysing the outliers using box plot and using the flooring and capping methods to treat the outliers
4. Binning of columns for better visualisation.
5. Check the data imbalance based on target variable.
6. Segregate the data frame into Defaulters and others

Analysis of Data :

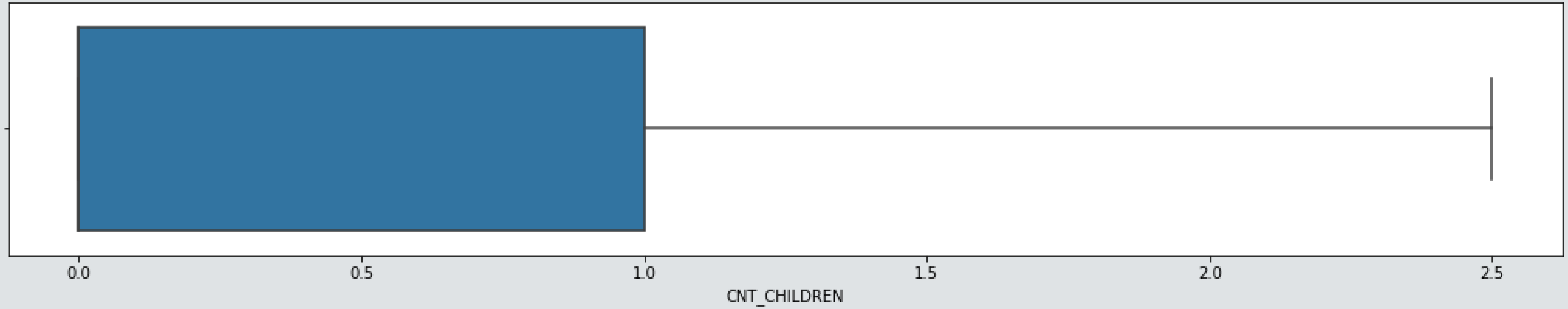
Univariate to understand the density of data in the data frames (application data & Previous data)
Bivariate analysis to understand the data spread density between the variables.
Multivariate analysis to understand the correlation

Taking CNT_CHILDRED column as example :

Identification of outliers : (Outliers before flooring and capping treatment)



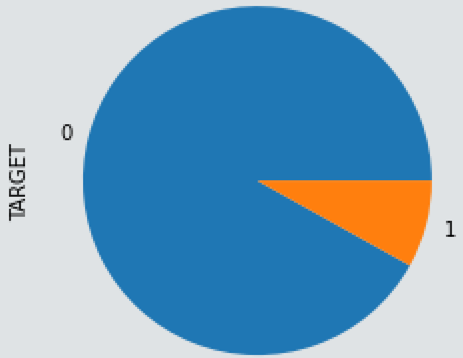
Treatment of outliers : (Outliers After flooring and capping treatment)



DATA IMBALANCE

- Checking the data imbalance in a data frame in respect to key column ‘TARGET’ data.

From the data distribution in terms of column TARGET it seems that 8% of data are in default and remaining are non defaulters.

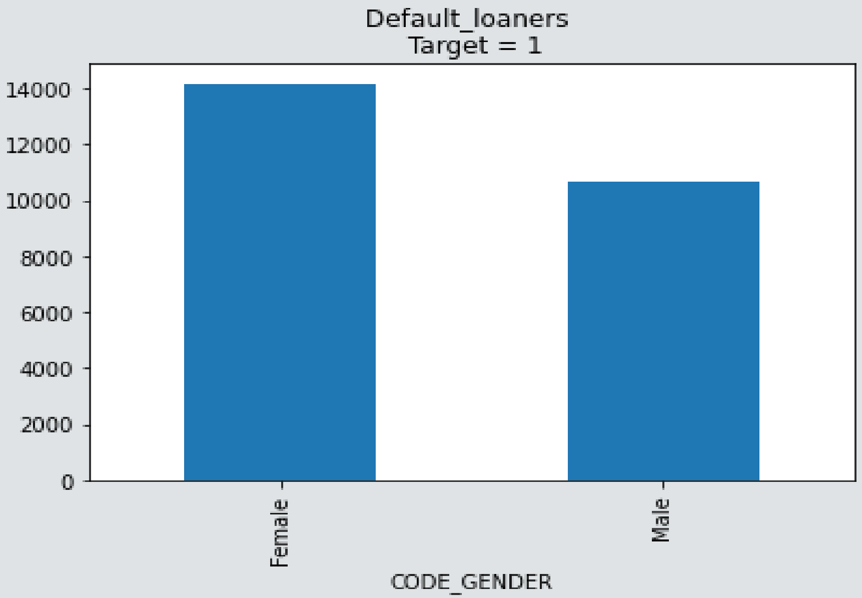


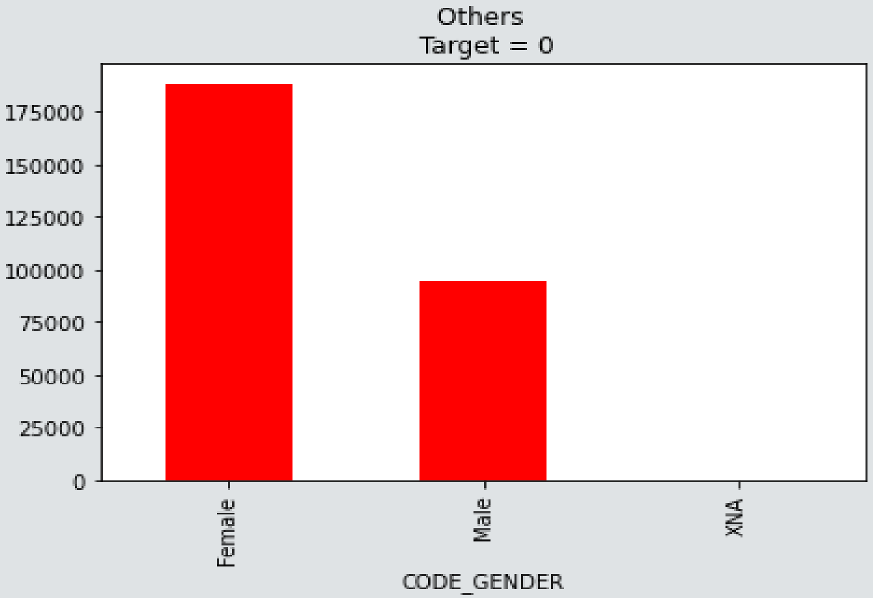
From the split dataset based on TARGET variable, it seems that 11.39 percentage of data imbalance is there.

Uni-Variate analysis

- Analysing the single variable is called Uni-variate analysis.
- To find out the pattern of data distribution in a data frame.

Uni-Variate Analysis : Code gender (Target = 1 vs Target = 0)

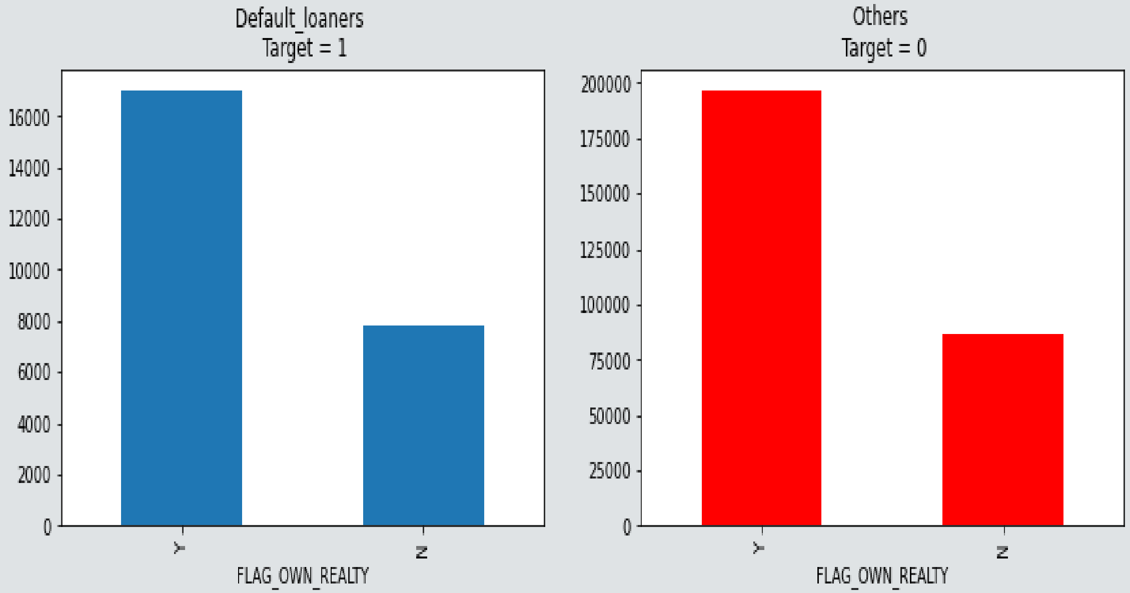




✓ In both the datasets defaulters and others , Females are having high payment difficulty and other more than Male.

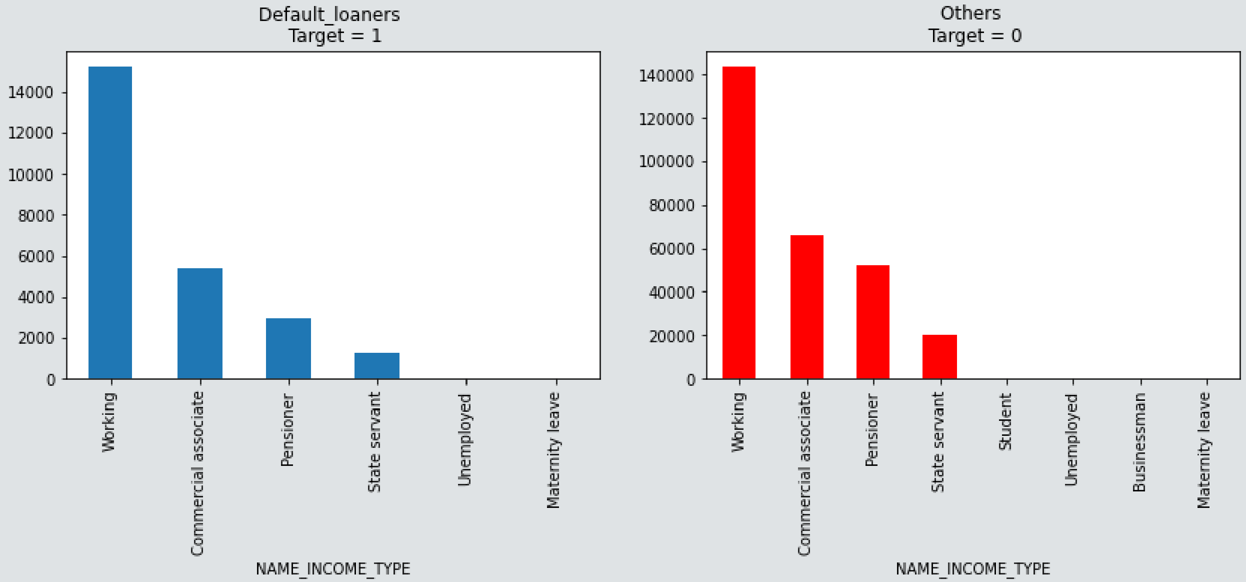
Uni-Variate Analysis in terms of Target (0 and 1)

Analysis of Flag own realty



- Above graph concludes that people with more own realty (assets) are default in making late payment .
- People not owning realty are making payment on time.

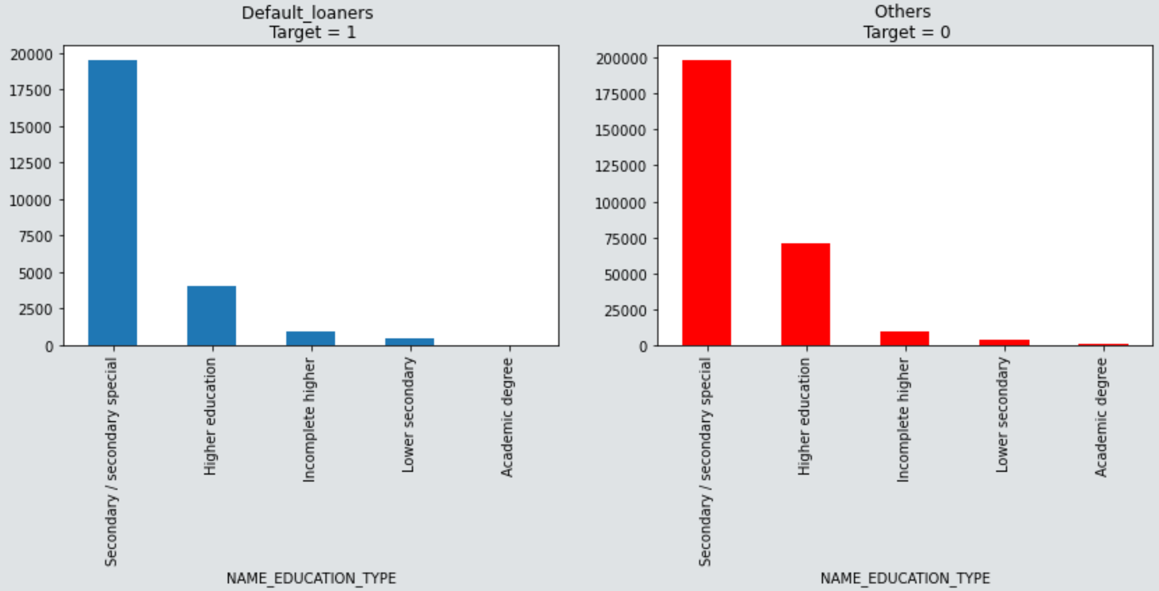
Analysis of Name income type



- Above graph concludes that working professional are challenge in making payment on time.
- State servant and students are making payment on time.

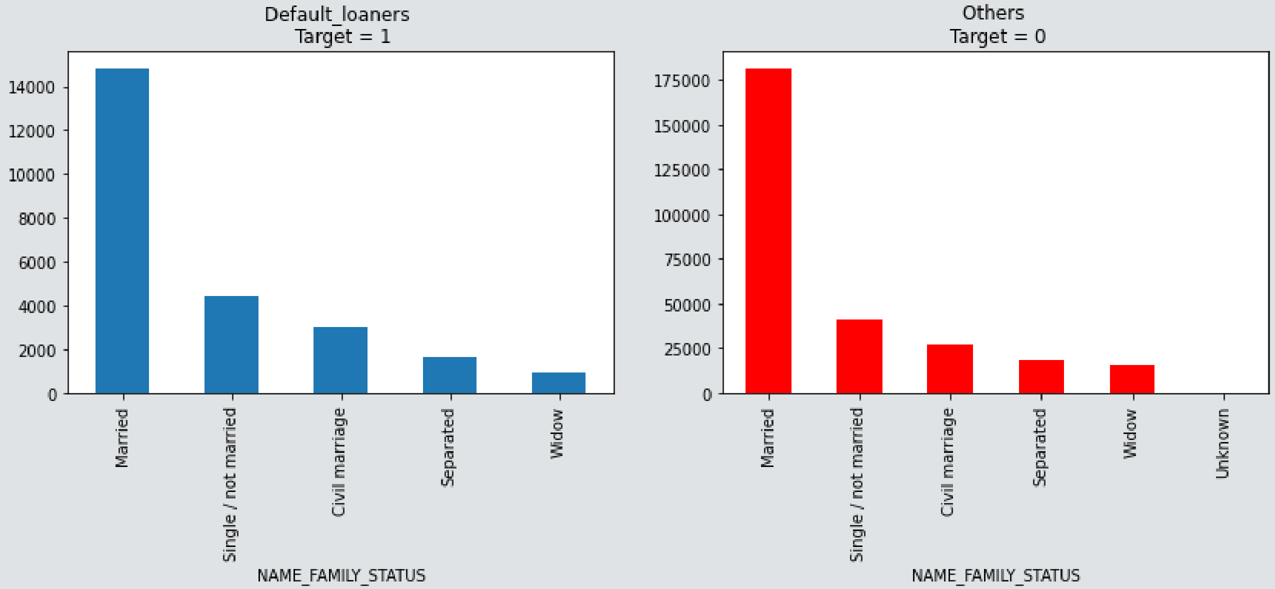
Uni-Variate Analysis in terms of Target (0 and 1)

Analysis of Education type :



- Above graph concludes that people with secondary / secondary special category are default in making late payment .
- People with incomplete higher and lower secondary are making payment on time.

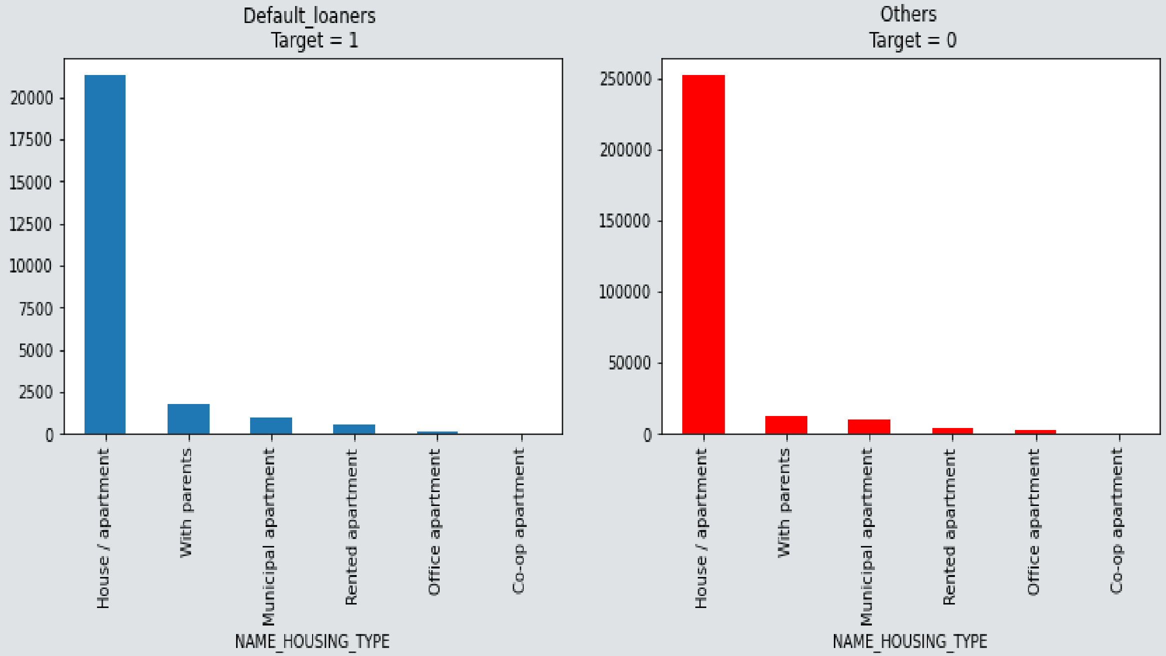
Analysis of Name Family status :



- Above graph concludes that Married people are challenge in making payment on time in both the categories.
- Separated and widows categories are making payment on time.

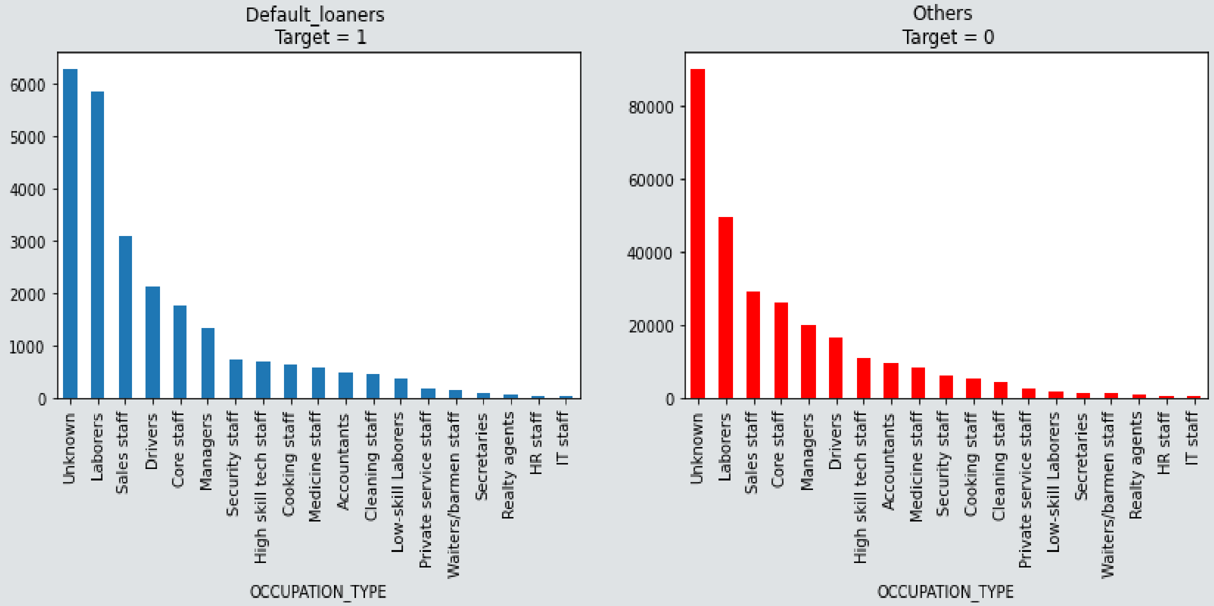
Uni-Variate Analysis in terms of Target (0 and 1)

Analysis of Housing type:



- Above graph concludes that people in House / apartment category are default in making late payment .
- People with office apartment and co-op apartment are making payment on time.

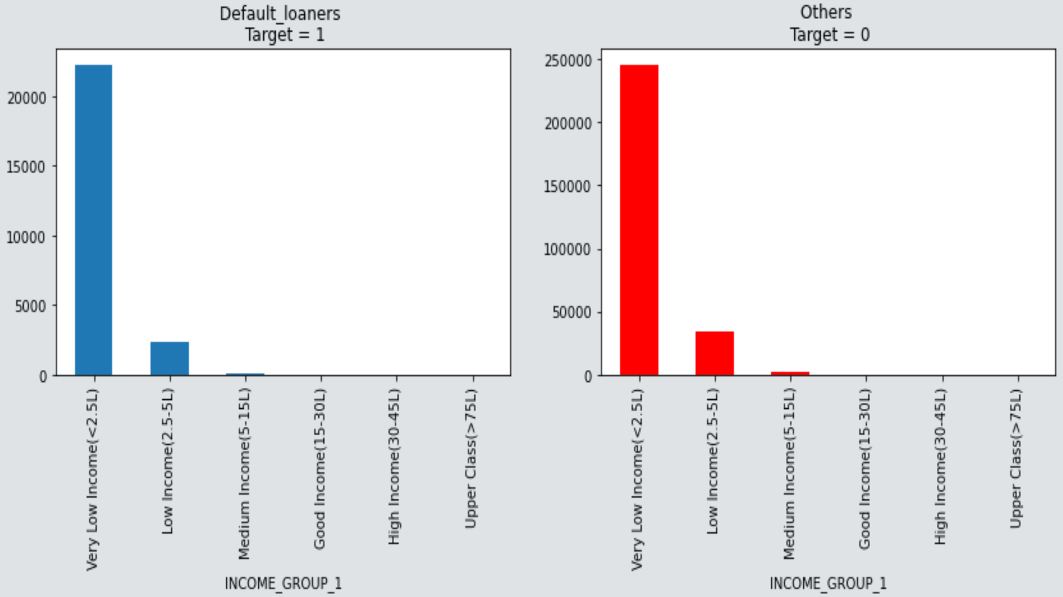
Analysis of Name Occupation type :



- Above graph concludes that unknown category and labour category are challenge in making payment on time .
- People in Private service staff and barmen staff are making payment on time.

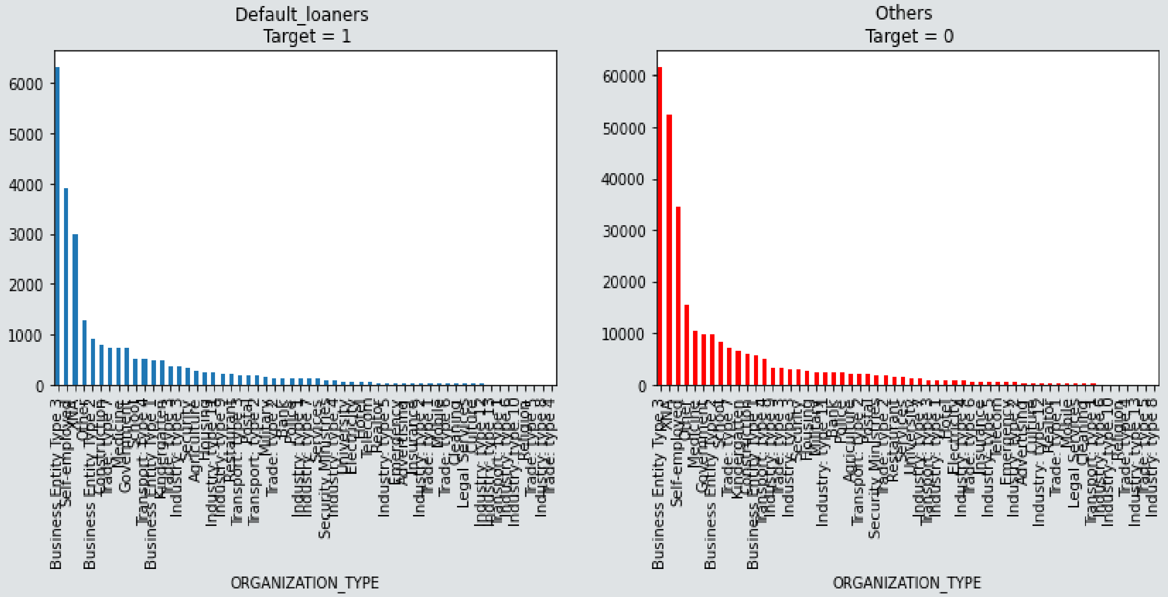
Uni-Variate Analysis in terms of Target (0 and 1)

Analysis of Income group type:



- Above graph concludes that people very low income category are default in making late payment .
- Medium, Good category income people are making payment on time.

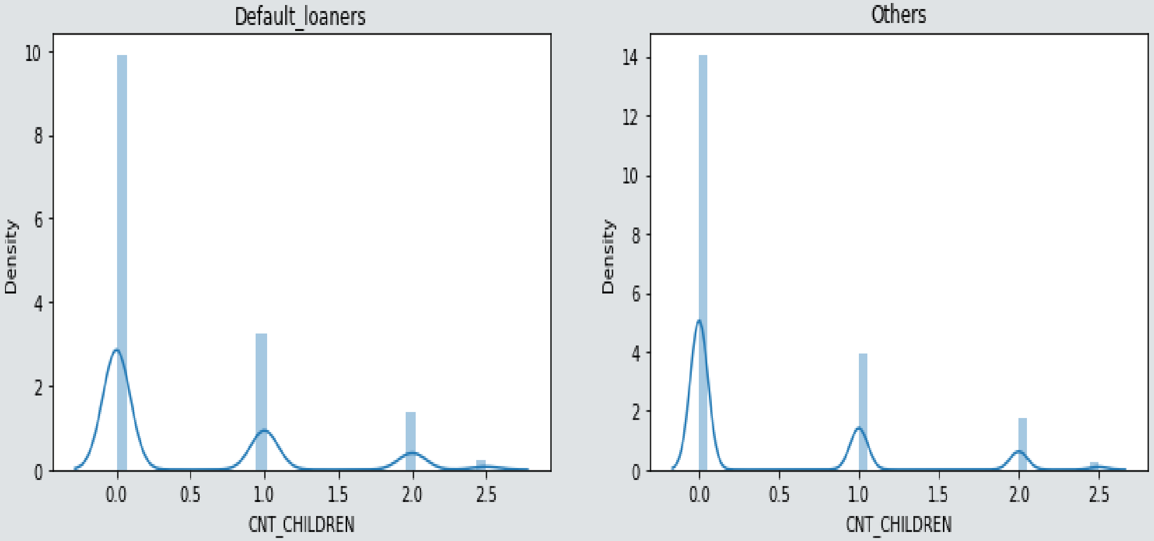
Analysis of Name Occupation type :



- Above graph concludes that Business Entry type and semi employed category are challenge in making payment on time .
- People employed in insurance, legal industry type I and type II are making payment on time.

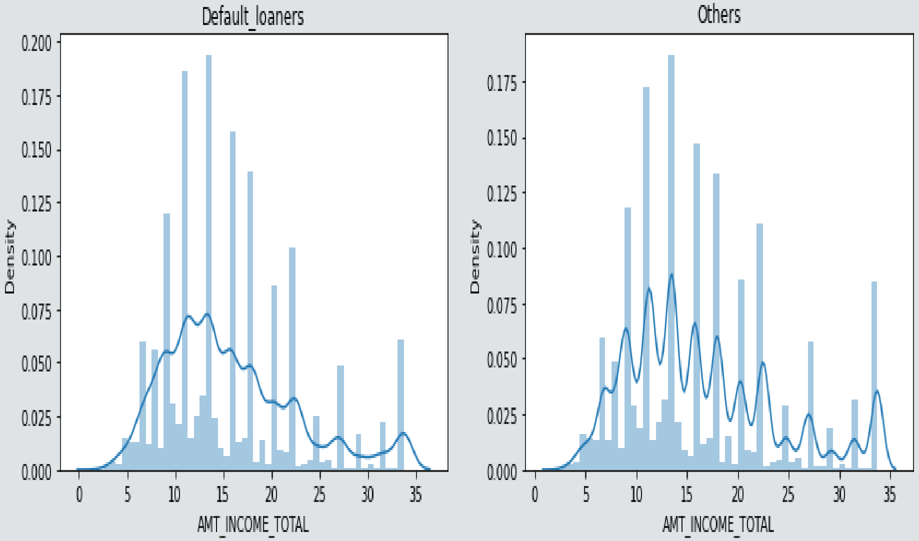
Univariate analysis

Count of Children Columns :



- Above graph showing the count of applicants without children are likely to make payments on time.
- Also client with children are finding difficult in making payment on time.

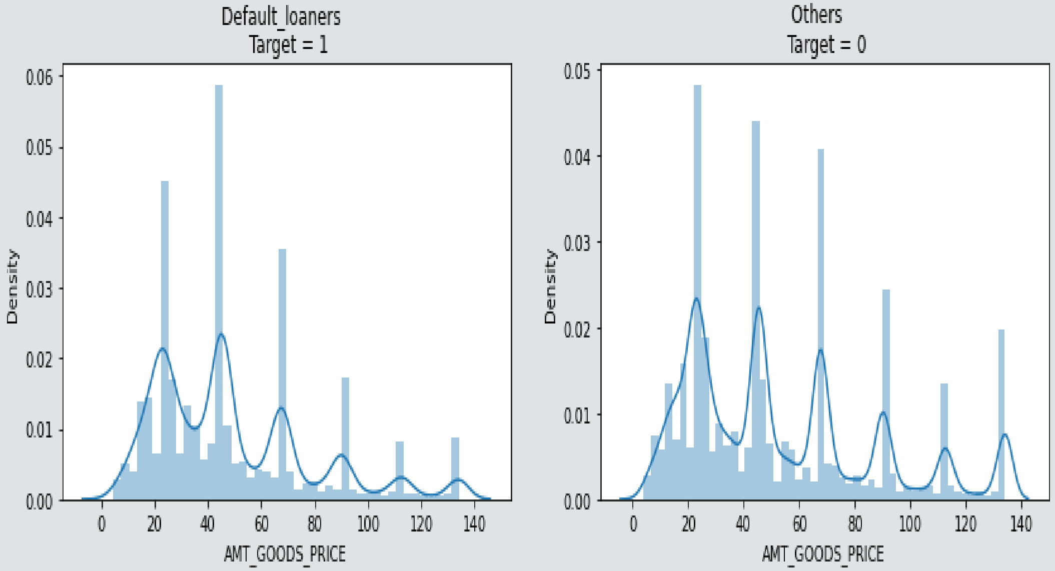
Income Total :



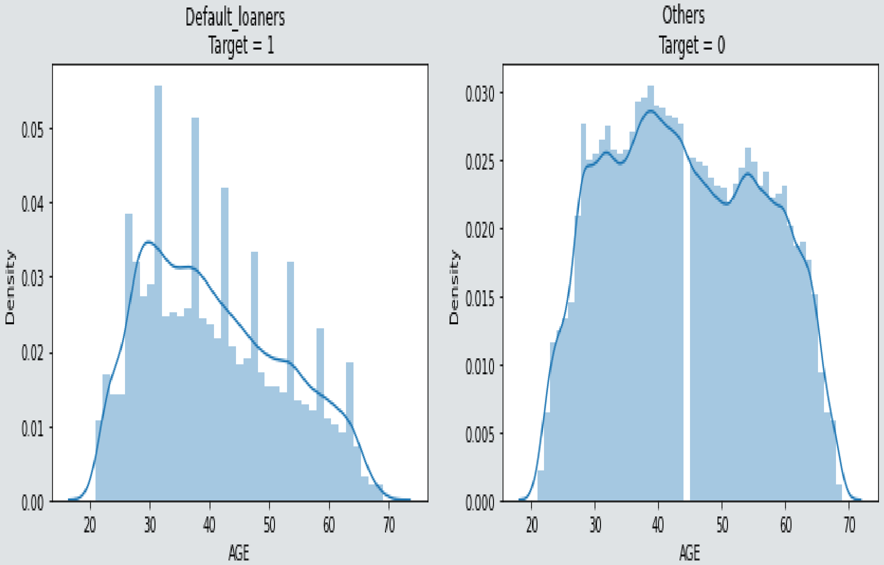
Above graph shows the income of applicants between 10 L and 15L are likely to make payments on time.

Univariate analysis

Amt_Goods_Price :



AGE (in years) :

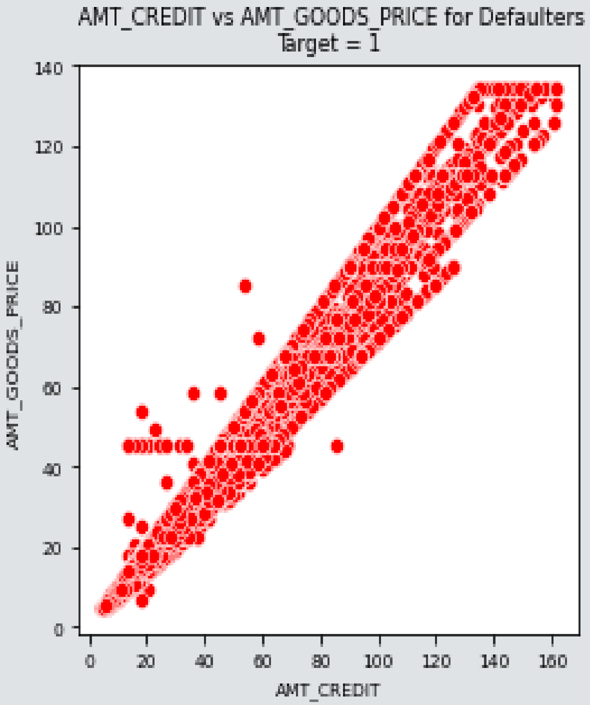


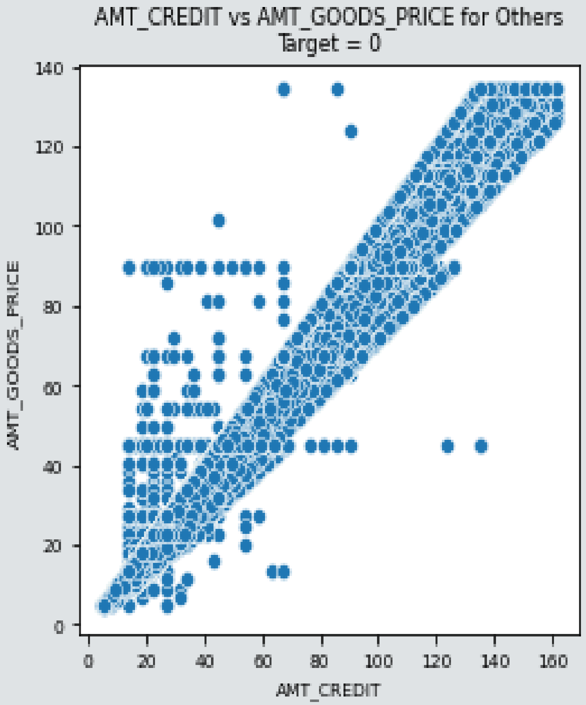
Above graph showing the AMT_GOODS_PRICE in the range of 20 – 70 are having more payment difficulties to make payments on time.

Above graph shows the age of applicants in both client with payment difficulties and others

Bi-Variate Analysis:

AMT_CREDIT vs AMT_GOODS_PRICE

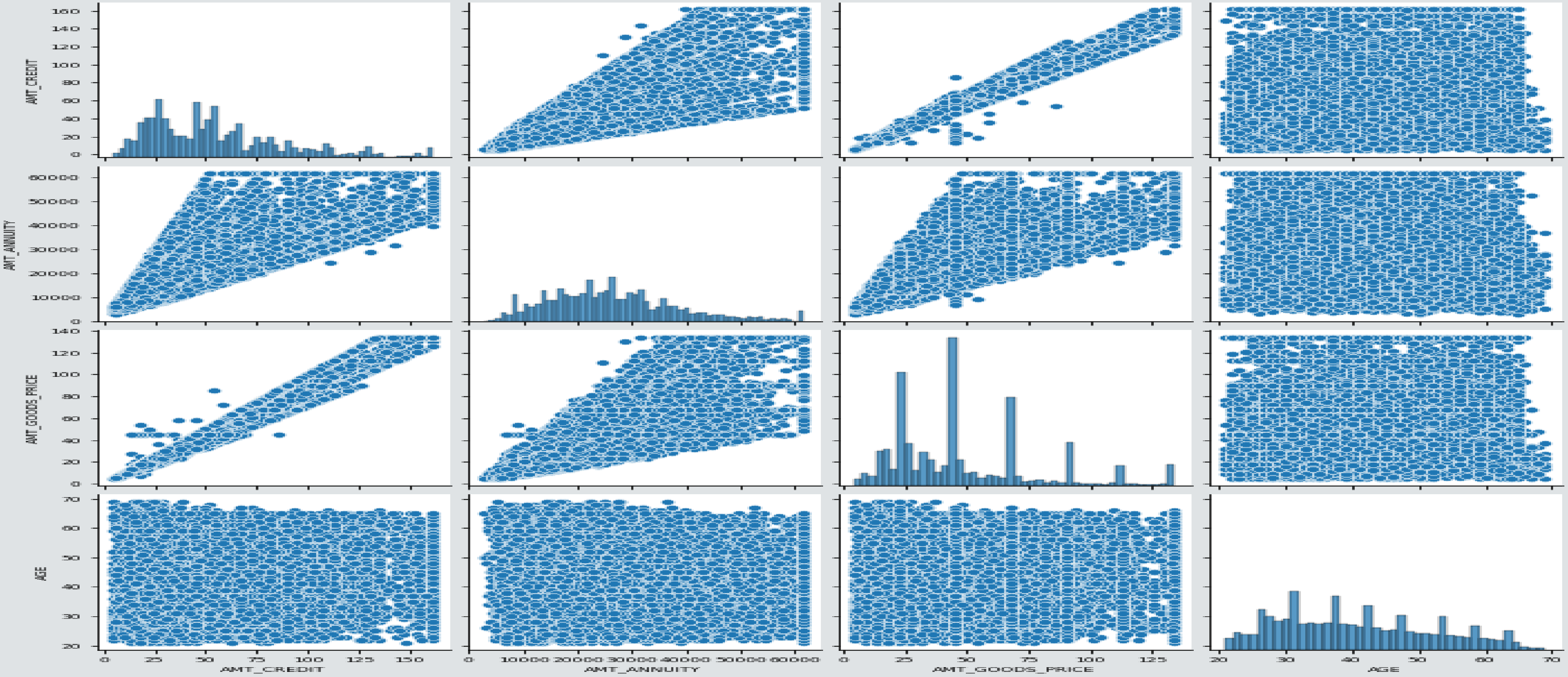




AMT_GOODS_PRICE` and `AMT_CREDIT` have strong positive correlation. This means that as Goods price increases, so does Credit Amount

Bi-Variate Analysis:

Bivariate analysis of categorical variables in terms of 1 (defaulters)



❑ Above graph shows the relation trend between the Amt_credit , Amt_Annuity, Amt_Goods_price and Age.

❑ Amt_credit vs Amt_goods_price , Amt_credit vs Amt_annuity are in positive correlation.

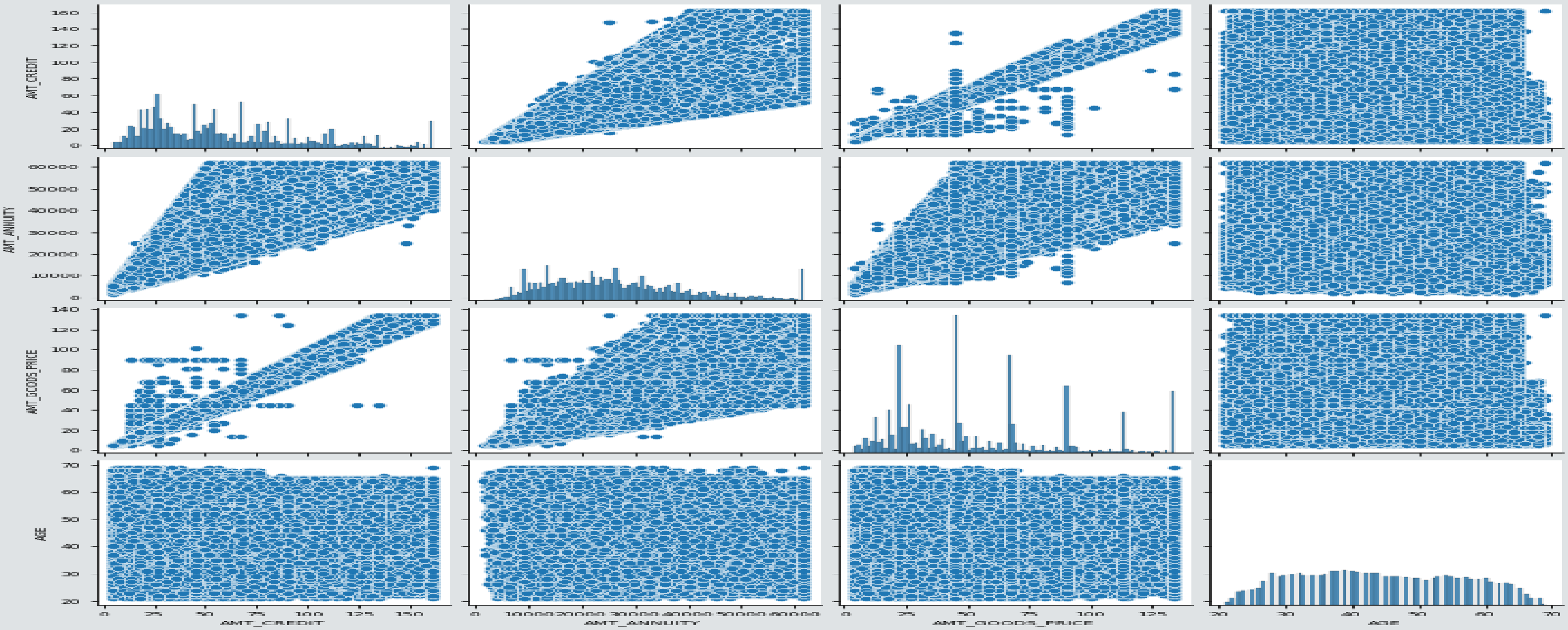
❑ In terms of Amt_annuity , Amt_goods_price and Amt_credit are in positive correlation.

❑ In terms of Amt_Goods_price, Amt_annuity and Amt_credit are in positive correlation.

❑ Interems of Age, Amt_good_price,Amt_annuity and Amt_credit are distributed widely in arandom manner.

Bi-Variate Analysis:

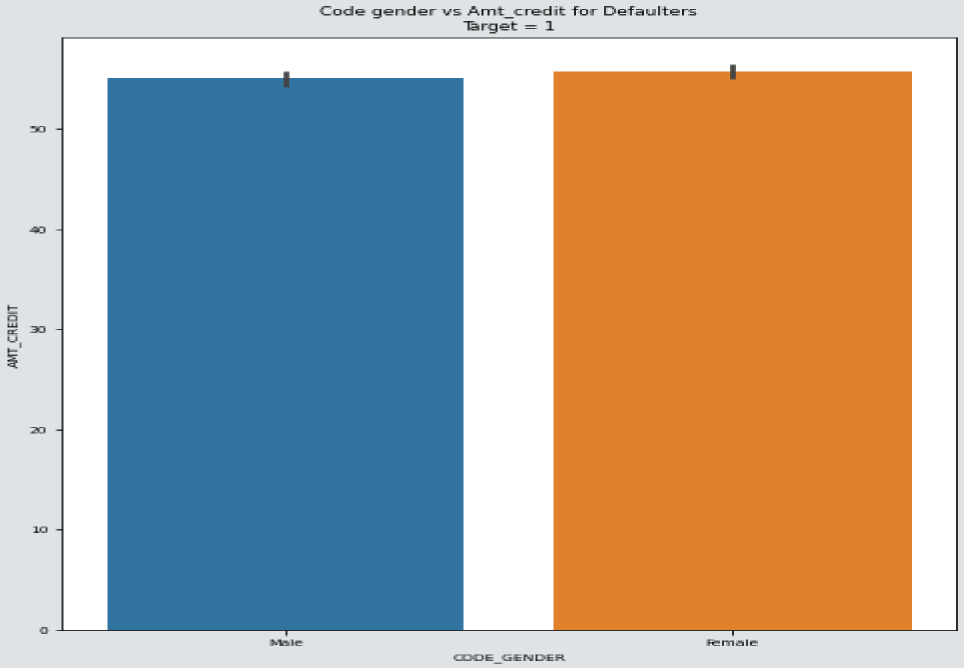
Bivariate analysis of categorical variables in terms of 0 (Others)

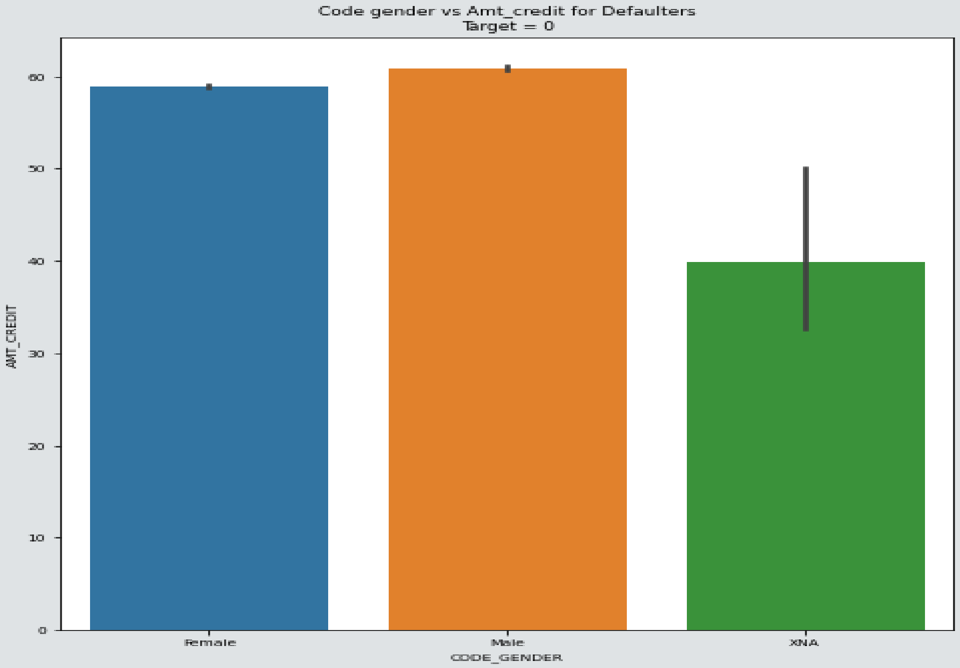


- ❑ Above graph shows the relation trend between the Amt_credit , Amt_Annuity, Amt_Goods_price and Age.
- ❑ Amt_credit vs Amt_goods_price , Amt_credit vs Amt_annuity are in positive correlation.
- ❑ In terms of Amt_annuity , Amt_goods_price and Amt_credit are in positive correlation.
- ❑ In terms of Amt_Goods_price, Amt_annuity and Amt_credit are in positive correlation.

Bi-Variate Analysis:

AMT_CREDIT vs CODE_GENDER

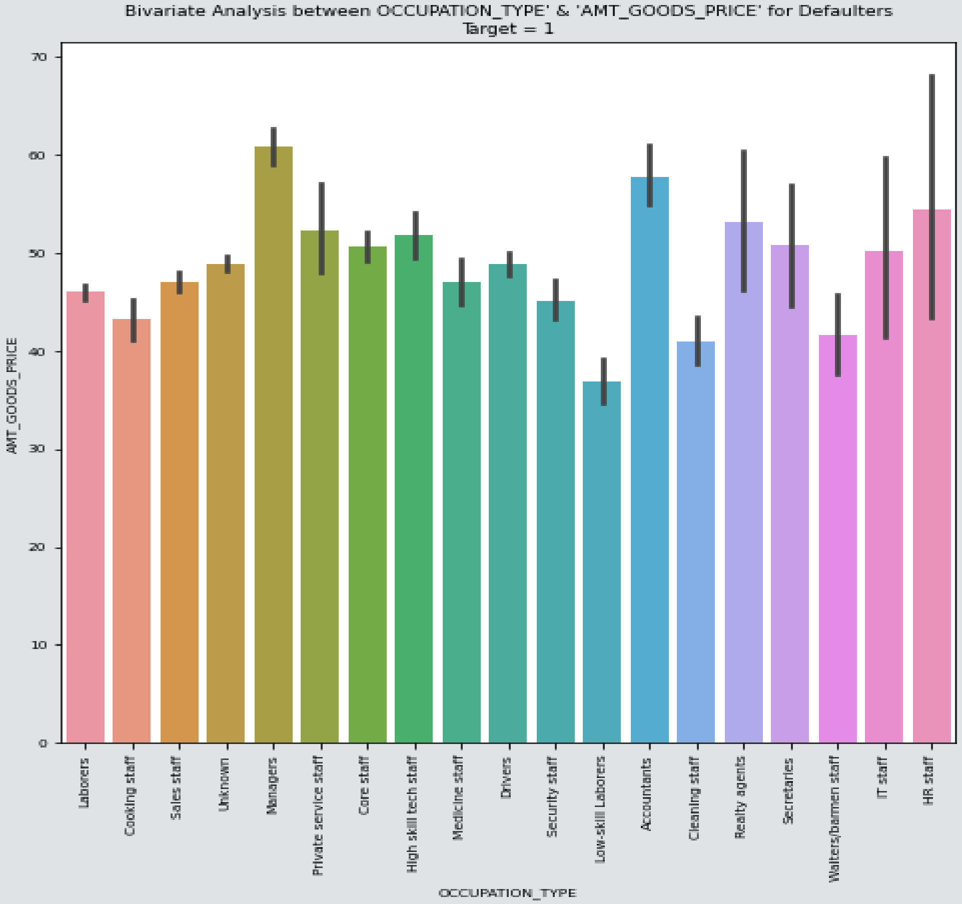


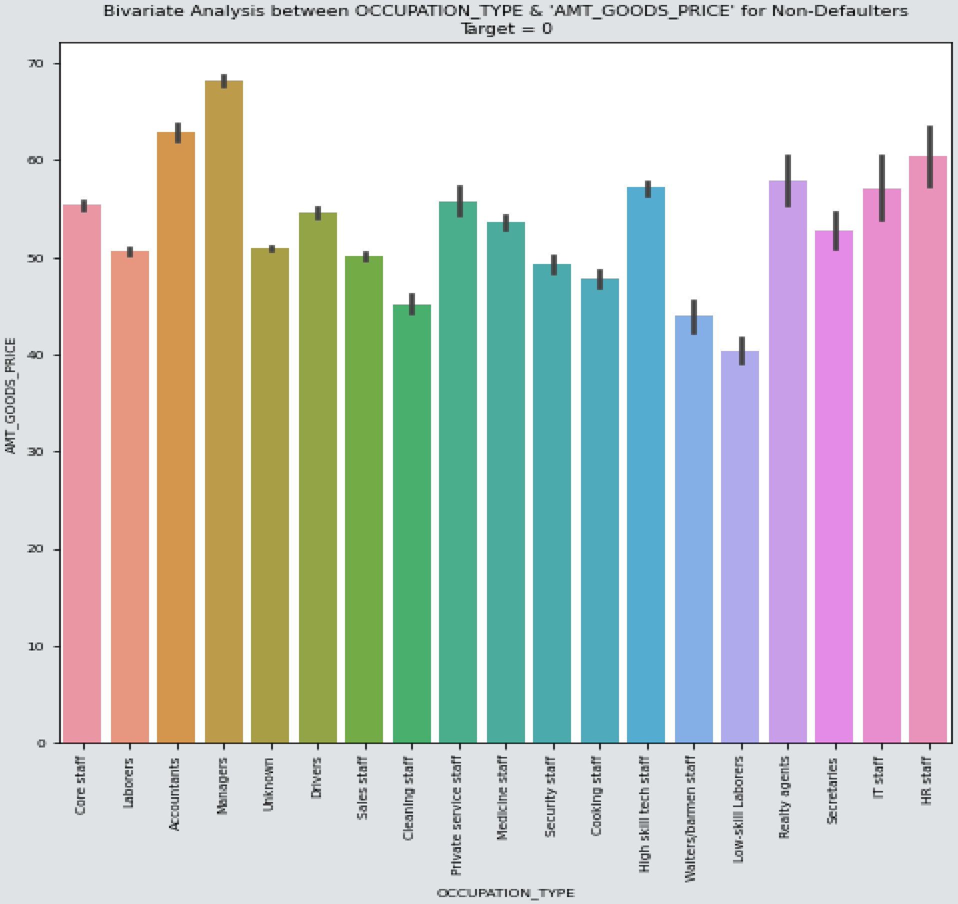


- ❑ AMT_CREDIT Vs CODE_GENDER shows that In Defaulter (Target is 1) Male and Female are having same difficulties.
- ❑ In terms of others in target variable, Male are likely to have higher correlation when compare to female and XNA.

Bi-Variate Analysis:

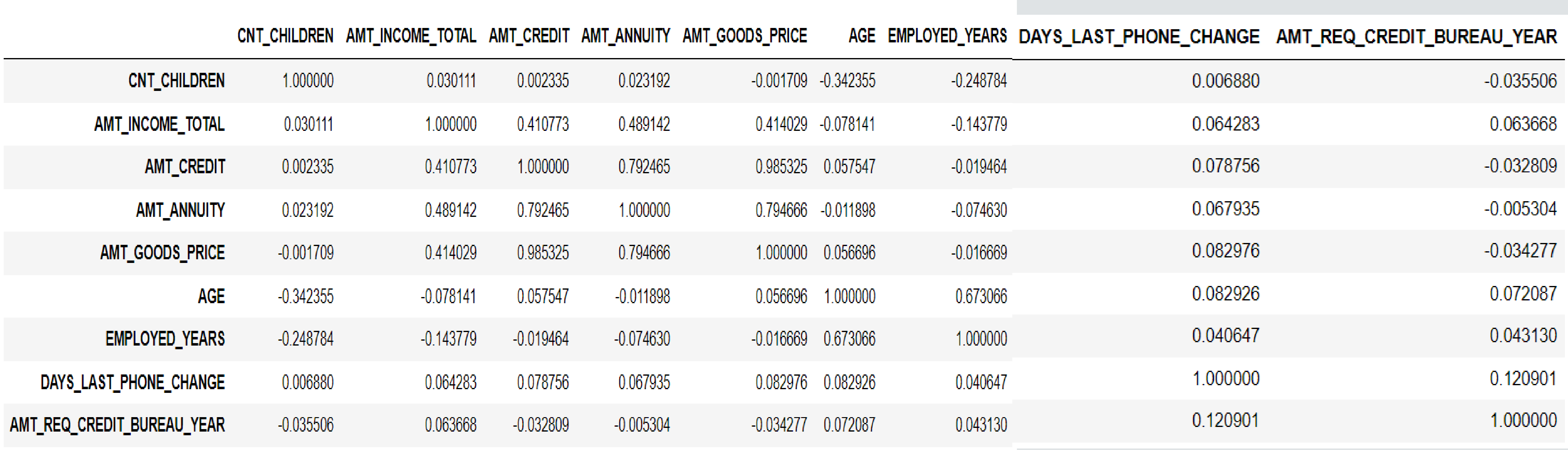
AMT_CREDIT vs AMT_GOODS_PRICE





❑ AMT_GOODS_PRICE` and `OCCUPATION TYPE` are have higher loan defaulters (Target is 1 and 0) of top 3 category of Managers and Accountants and HR staff .

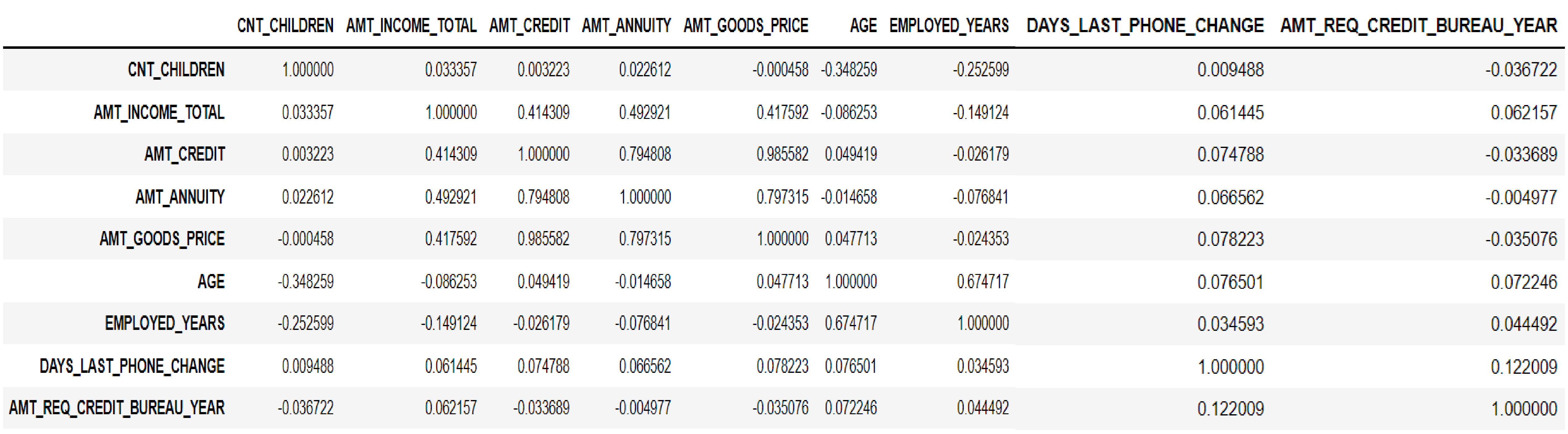
Correlation of numerical columns (Continuous columns terms of TARGET = 1 having payment difficulty)



	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	AGE	EMPLOYED_YEARS	DAYS_LAST_PHONE_CHANGE	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1.000000	0.030111	0.002335	0.023192	-0.001709	-0.342355	-0.248784	0.006880	-0.035506
AMT_INCOME_TOTAL	0.030111	1.000000	0.410773	0.489142	0.414029	-0.078141	-0.143779	0.064283	0.063668
AMT_CREDIT	0.002335	0.410773	1.000000	0.792465	0.985325	0.057547	-0.019464	0.078756	-0.032809
AMT_ANNUITY	0.023192	0.489142	0.792465	1.000000	0.794666	-0.011898	-0.074630	0.067935	-0.005304
AMT_GOODS_PRICE	-0.001709	0.414029	0.985325	0.794666	1.000000	0.056696	-0.016669	0.082976	-0.034277
AGE	-0.342355	-0.078141	0.057547	-0.011898	0.056696	1.000000	0.673066	0.082926	0.072087
EMPLOYED_YEARS	-0.248784	-0.143779	-0.019464	-0.074630	-0.016669	0.673066	1.000000	0.040647	0.043130
DAYS_LAST_PHONE_CHANGE	0.006880	0.064283	0.078756	0.067935	0.082976	0.082926	0.040647	1.000000	0.120901
AMT_REQ_CREDIT_BUREAU_YEAR	-0.035506	0.063668	-0.032809	-0.005304	-0.034277	0.072087	0.043130	0.120901	1.000000

Above table showing the correlation of continuous columns (numerical Data columns)

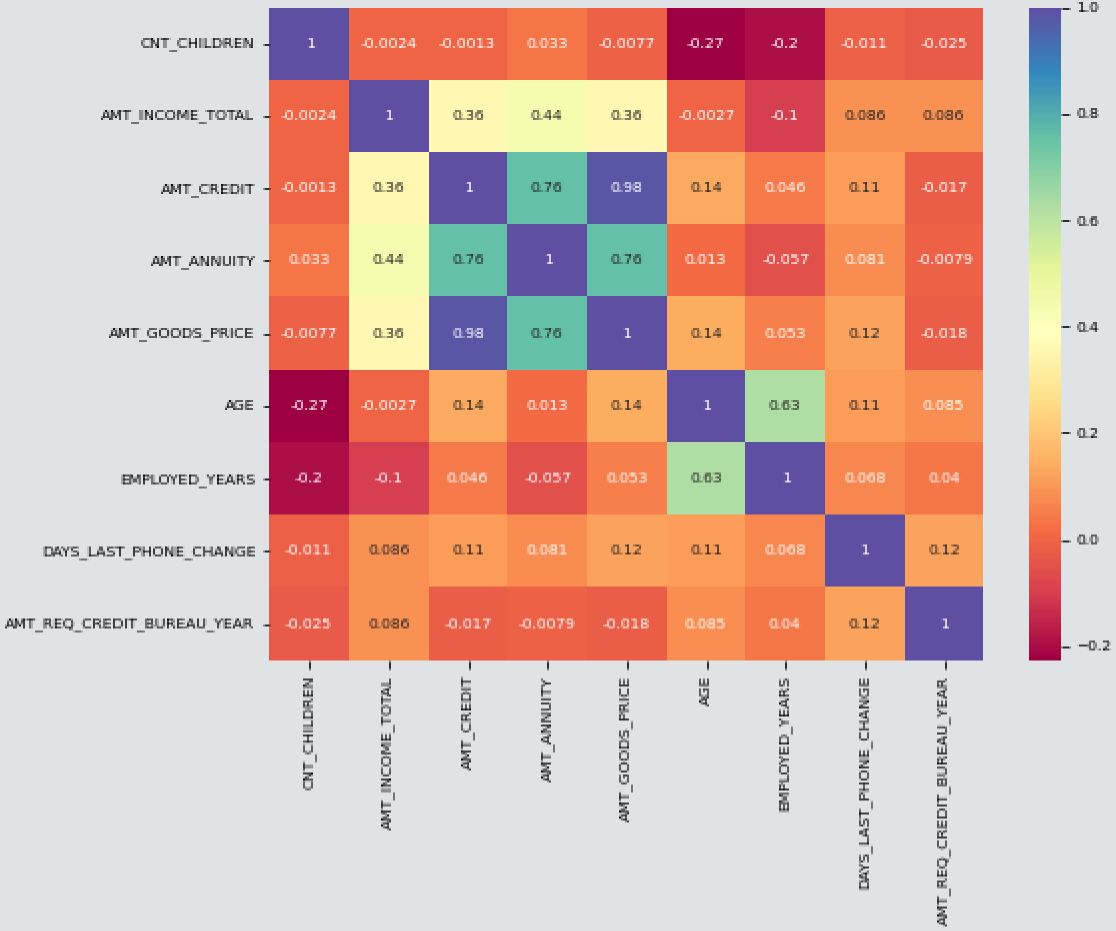
Correlation of numerical columns (Continuous columns terms of TARGET = 0 others)



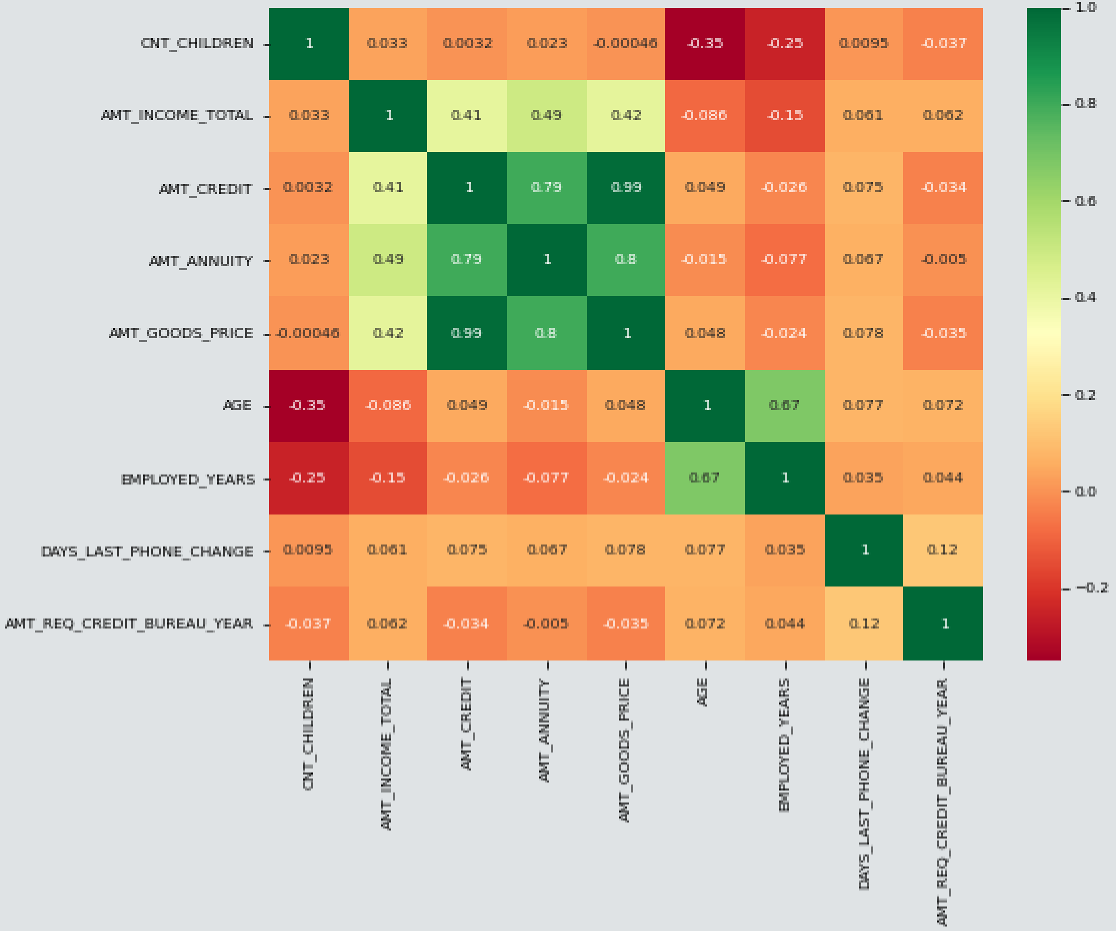
	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	AGE	EMPLOYED_YEARS	DAYS_LAST_PHONE_CHANGE	AMT_REQ_CREDIT_BUREAU_YEAR
CNT_CHILDREN	1.000000	0.033357	0.003223	0.022612	-0.000458	-0.348259	-0.252599	0.009488	-0.036722
AMT_INCOME_TOTAL	0.033357	1.000000	0.414309	0.492921	0.417592	-0.086253	-0.149124	0.061445	0.062157
AMT_CREDIT	0.003223	0.414309	1.000000	0.794808	0.985582	0.049419	-0.026179	0.074788	-0.033689
AMT_ANNUITY	0.022612	0.492921	0.794808	1.000000	0.797315	-0.014658	-0.076841	0.066562	-0.004977
AMT_GOODS_PRICE	-0.000458	0.417592	0.985582	0.797315	1.000000	0.047713	-0.024353	0.078223	-0.035076
AGE	-0.348259	-0.086253	0.049419	-0.014658	0.047713	1.000000	0.674717	0.076501	0.072246
EMPLOYED_YEARS	-0.252599	-0.149124	-0.026179	-0.076841	-0.024353	0.674717	1.000000	0.034593	0.044492
DAYS_LAST_PHONE_CHANGE	0.009488	0.061445	0.074788	0.066562	0.078223	0.076501	0.034593	1.000000	0.122009
AMT_REQ_CREDIT_BUREAU_YEAR	-0.036722	0.062157	-0.033689	-0.004977	-0.035076	0.072246	0.044492	0.122009	1.000000

Above table showing the correlation of continuous columns (numerical Data columns)

Heat Map Analysis of numerical variables for both target (0 and 1) – Correlation Matrix



Correlation heatmap data of Defaulters
Target = 1



Correlation heatmap data of Others_1
Target = 0

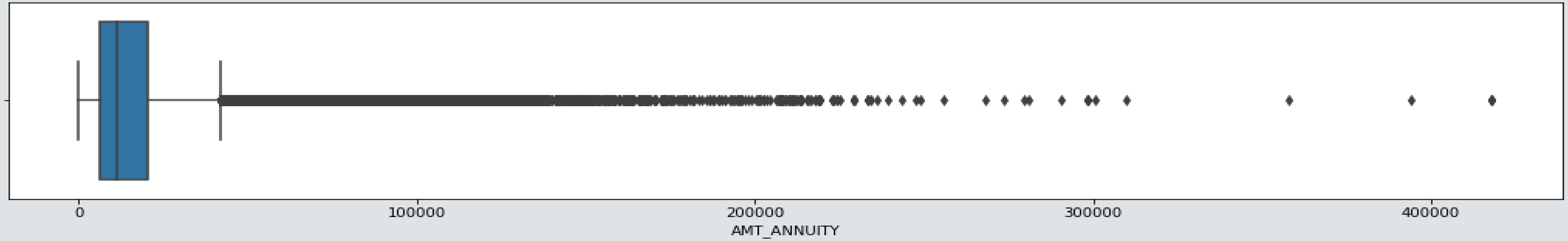
- ✓ **Correlation of Others_1** :Columns with Top correlations are highlighted in light greenish to dark green shade.
- ✓ **Correlation of Defaulters** :Top correlation variables are highlighted in light blue to dark blue
- ✓ Amt_credit , Amt Goods and amt_annuity , Age and Employed yearsv are have strong correlation in payment difficulty data frame.
- ✓ On comparing the correlation matrix, both are having same type of correlation features.

Analysing the Previous application data :

- Export the previous application data in jupyter notebook
- Analyse the data details by using size, shape and describe
- Identification and removing of missing values in each columns based on threshold values
- Identification of outliers based on box.
- Treating the outliers based on flooring and capping techniques.
- Analyse each columns by univariate analysis.
- Merging the application data and Previous application data frames to analyse the contract approval status.
- Using uni variate and multivariate analysis to have a clear picture of approval status.

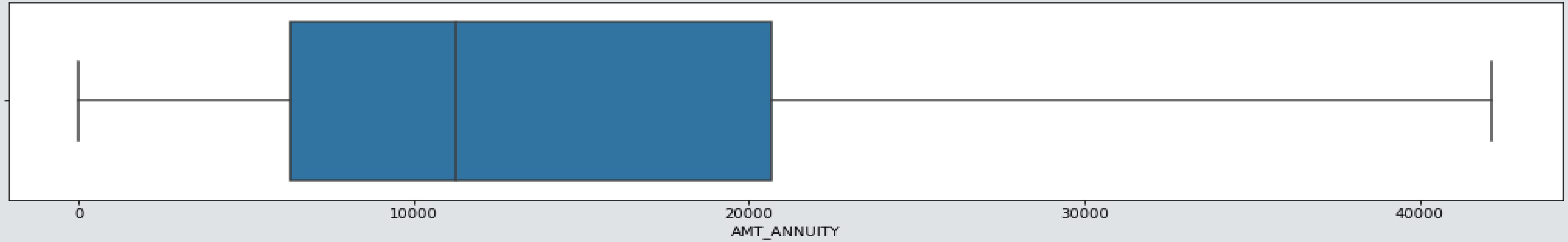
Finding outliers by plotting box plot in previous data

AMT_ANNUITY_COLUMN with Outliers :



From the above plot we can see some values are present after the 95th percentile and they needs to be imputed with median values to get the data analysis accurate.Hence they are treated with flooring and capping methods.

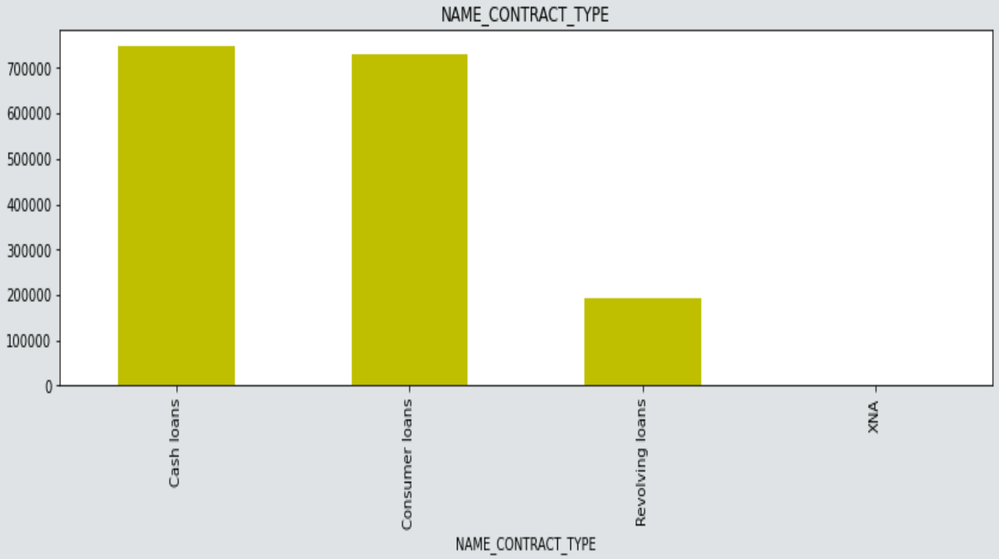
AMT_ANNUITY_COLUMN after treatment of Outliers :



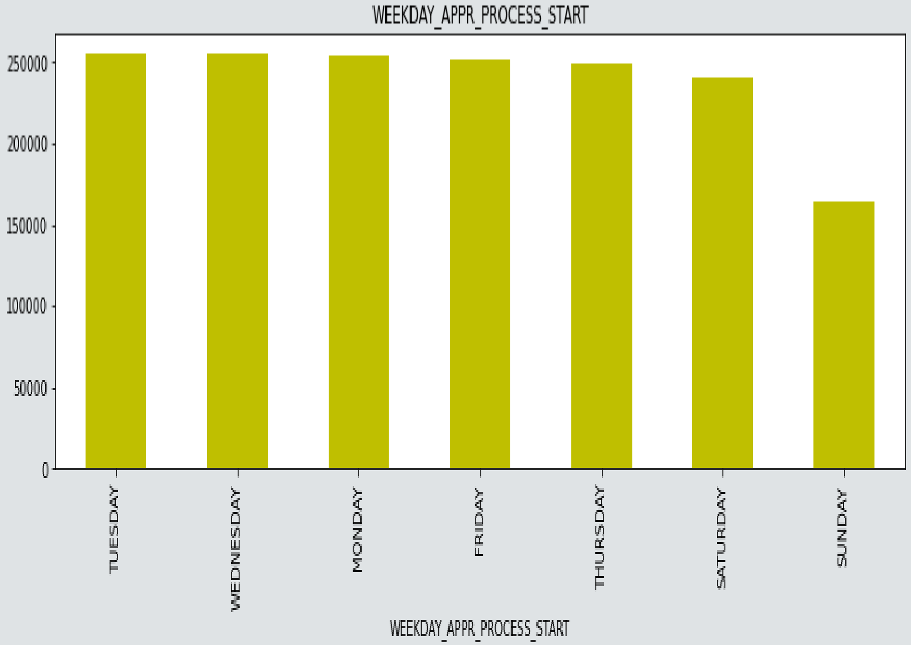
Above plot showing the datas perfectly when compared to data with outliers.

Uni-Variate Analysis:

NAME_CONTRACT_TYPE :

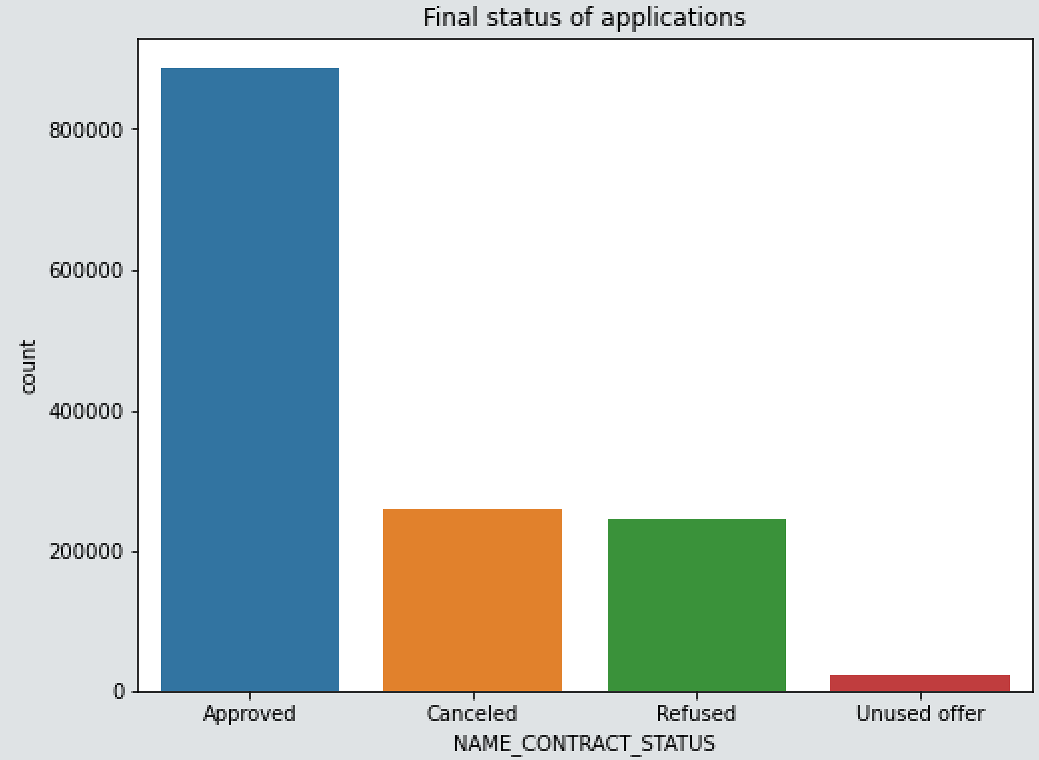


WEEKDAY_APPR_PROCESS_START

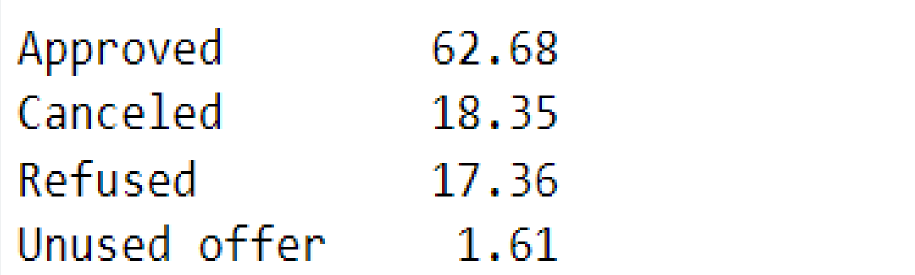


- ❑ From the above bank is providing cash loans mostly and revolving loans in lower volumes.
- ❑ From the WEEKDAY_APPR_PROCESS_START, plot shows Tuesday and Wednesday are high volumes are processed and in Saturday and Sunday less

Univariate analysis - NAME_CONTRACT_STATUS



Percentages distribution :

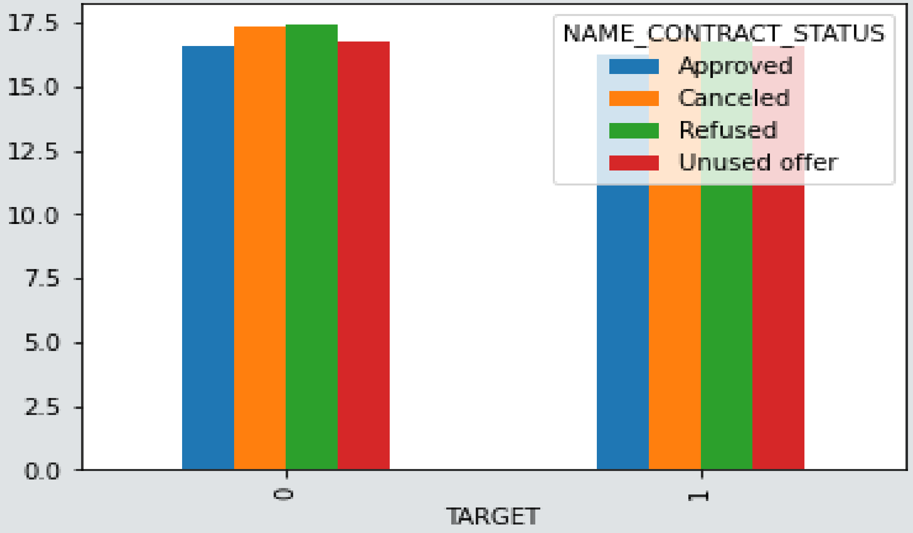


```
Approved        62.68
Canceled        18.35
Refused         17.36
Unused offer     1.61
```

Above plot shows the distribution of status of application. From the above we conclude that more previous applications are approved and their percentages are as above.

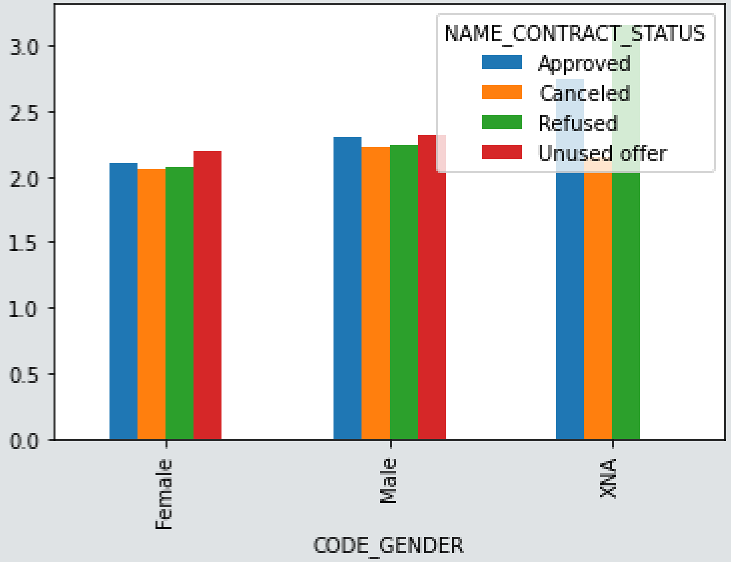
Multi-Variate Analysis

AMT_INCOME_TOTAL vs TARGET vs NAME_CONTRACT_STATUS



- Above graph concludes that approved and cancelled are having the more or less same proportionate distribution

CNT_FAM_MEMBERS vs CODE_GENDER vs NAME_CONTRACT_STATUS



- Above graph concludes that final application status of approval for male and female candidates like to same when compared to XNA.
- XNA Category is having more refusal count than others.

CONCLUSION

- ✓ Bank can give more loans to male candidate considering the payment difficulty.
- ✓ Bank can approve loans for clients who are not owning any realty
- ✓ Bank can provide loans to students and working professional as they are paying the dues on time
- ✓ Incomplete higher and lower secondary are likely to make payment on time and therefore approval for this categories may be increased.
- ✓ Bank can approve Separated and widowed category loan application.
- ✓ Ban can provide loans to office apartment and working rental categories as they are making on time payment.
- ✓ Bank can increase loans to IT staffs and HR staffs.
- ✓ Bank can provide income slab of 10 lakhs to 15 lakhs
- ✓ Bank can provide loans to client age between 30 to 40.