GROUP PROJECT 1

Credit EDA Case Study

Submitted by:-

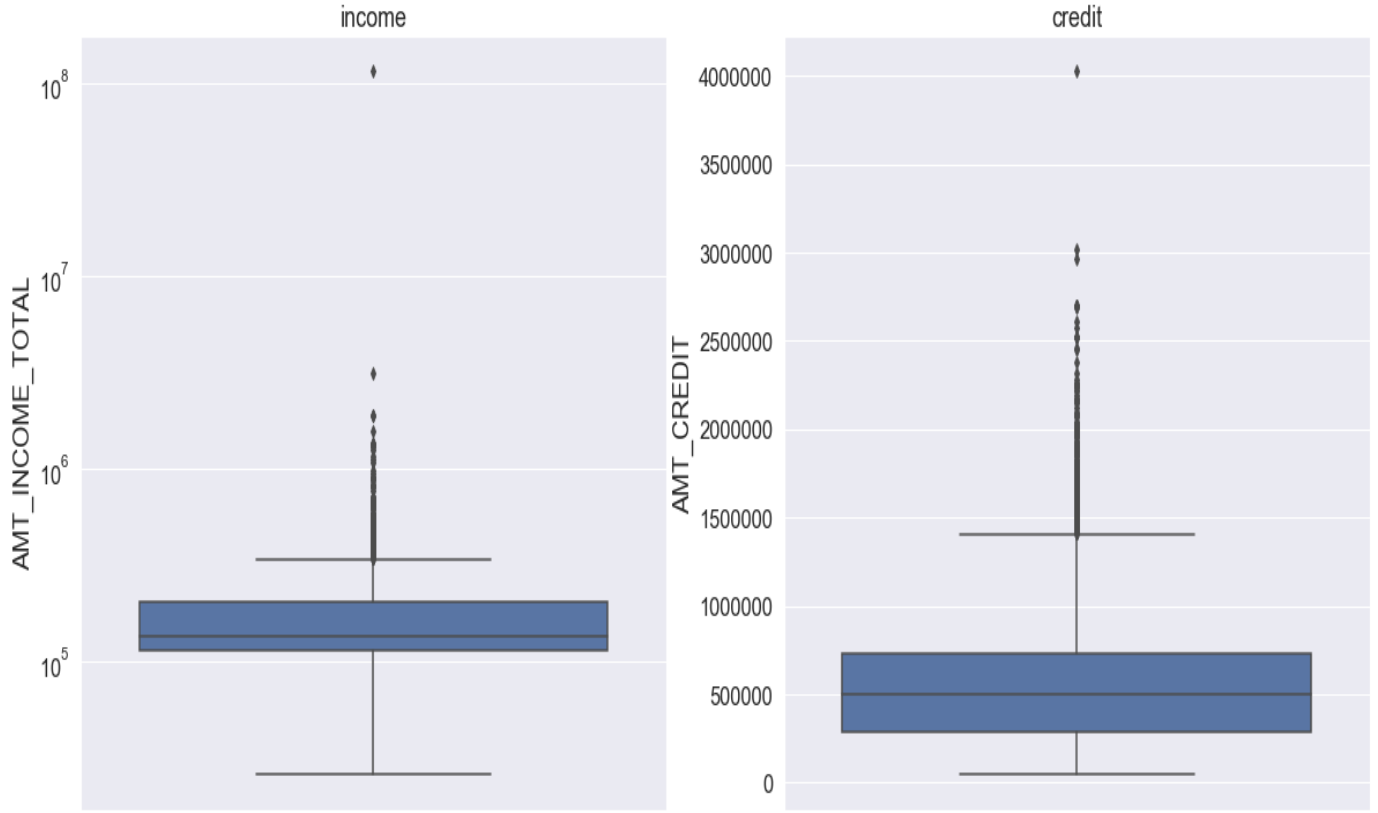
Krutika Shahu
Ankur Agarwal

Outliers:

These **box plots** help us to clearly identify the outliers in **Income, Credit, Years of employment and no. of EMIs.**

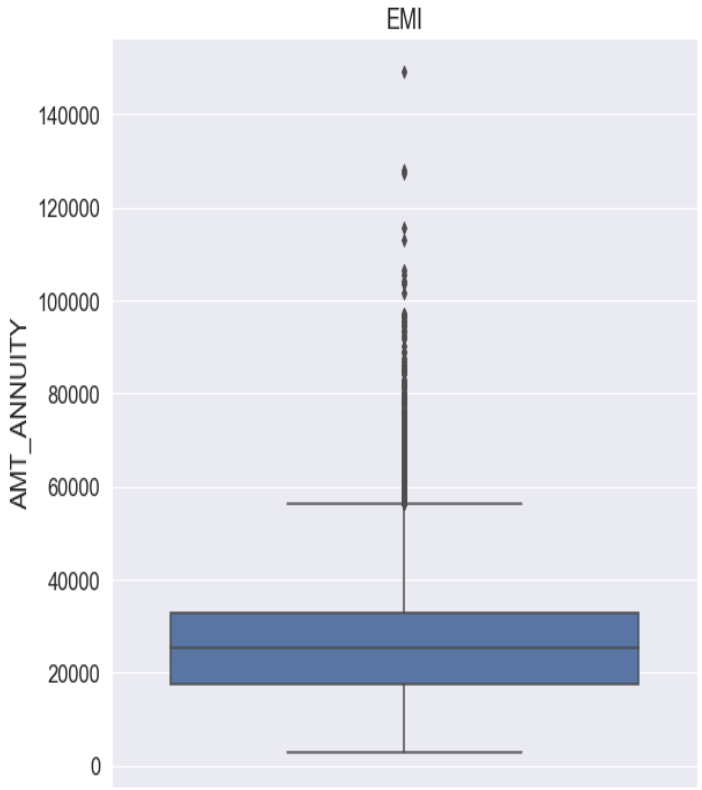
- **Income:** We can clearly identify that the range of the Income is between 10 to power 5 and 10 to power 6. The 25,50,75 and 100 percentiles are within the range of 10 to power 5 and 10 to power 6. And rest are the outliers.

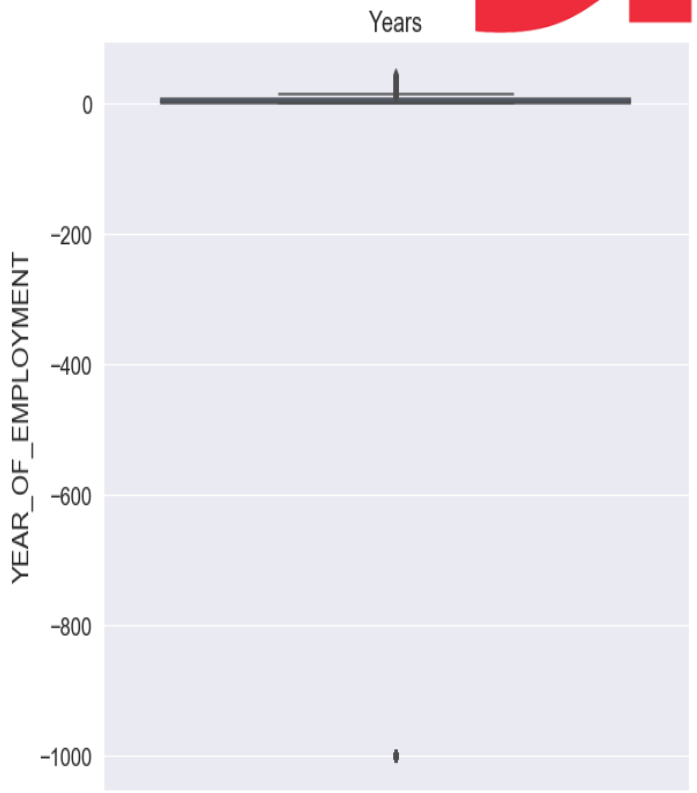
- **Credit:** Similarly, we can see that the 25,50,75 and 100 percentile range for Credit is between is 100000 to 1500000 and rest of the values are outliers



- **AMT_ANNUTY** : The annuity amounts are in a similar fashion distributed between 1000 to 58000

- **Years of employment :** Years of employment has a very unique outlier value which goes beyond -1000 years. It is a clear indicator outlier data which could be a result of human error while entering the data or default random value.



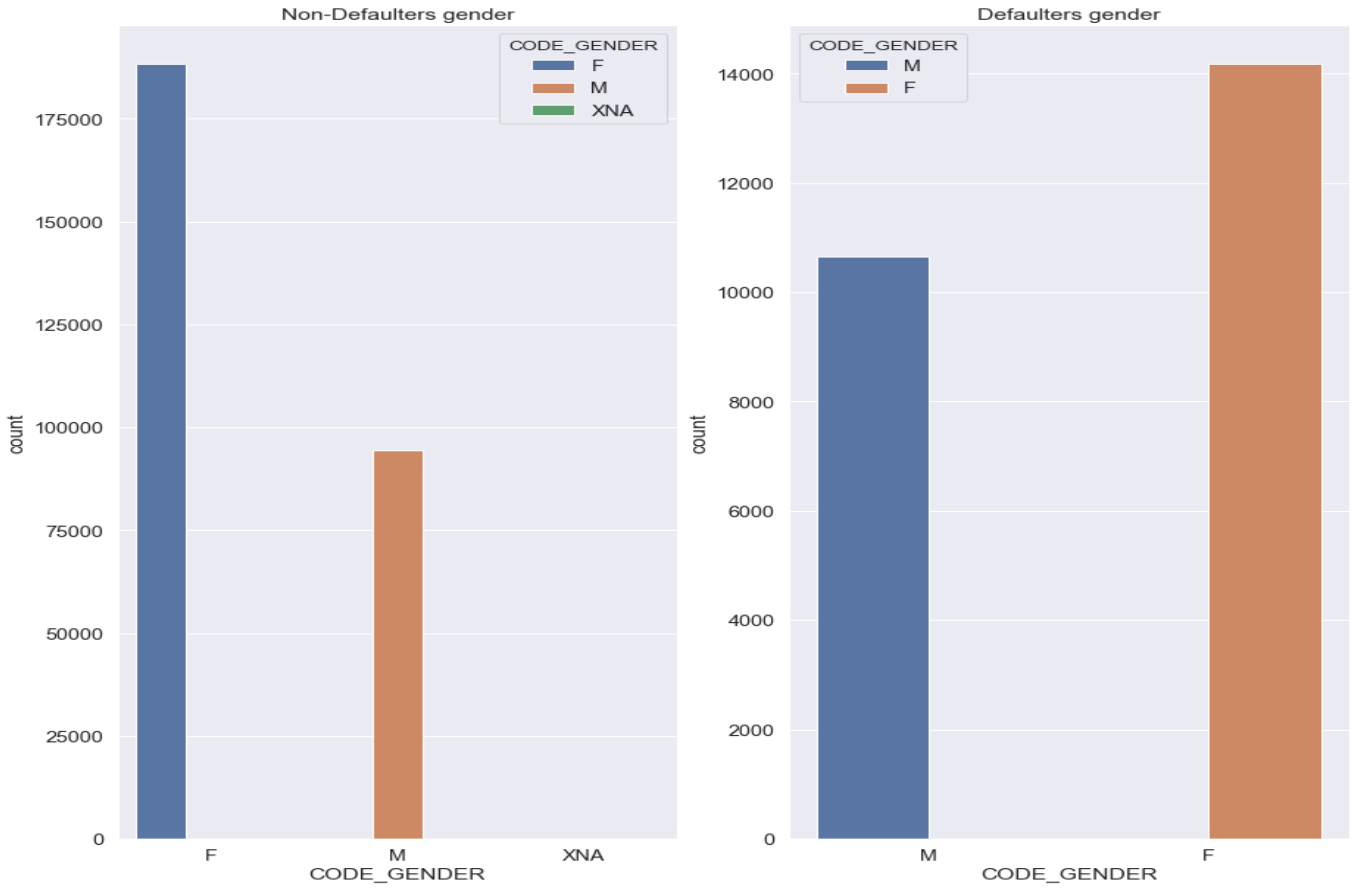


Univariate Analysis

The Graph on this page shows the Univariate Analysis for **Gender.**

The 1st graph is for Non-Defaulters and the 2nd Graph displays Defaulters.

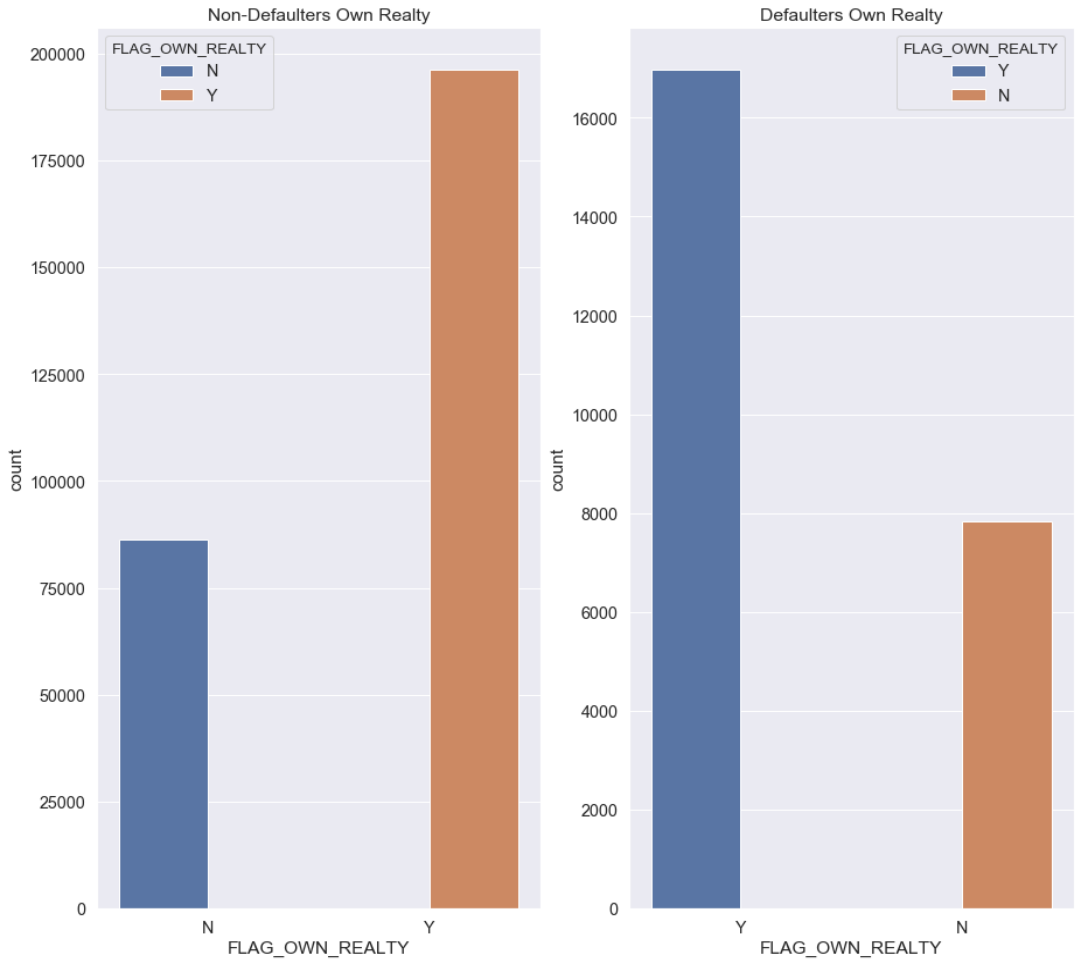
Clearly, in the defaulters target group the **gender-female** is the one which has the highest count and this makes female gender a potential candidate who can default the repayment of loans.



The Graph on this page shows the Univariate Analysis for **People Who Own Realty Property.**

The 1st graph is for Non-Defaulters and the 2nd Graph displays Defaulters.

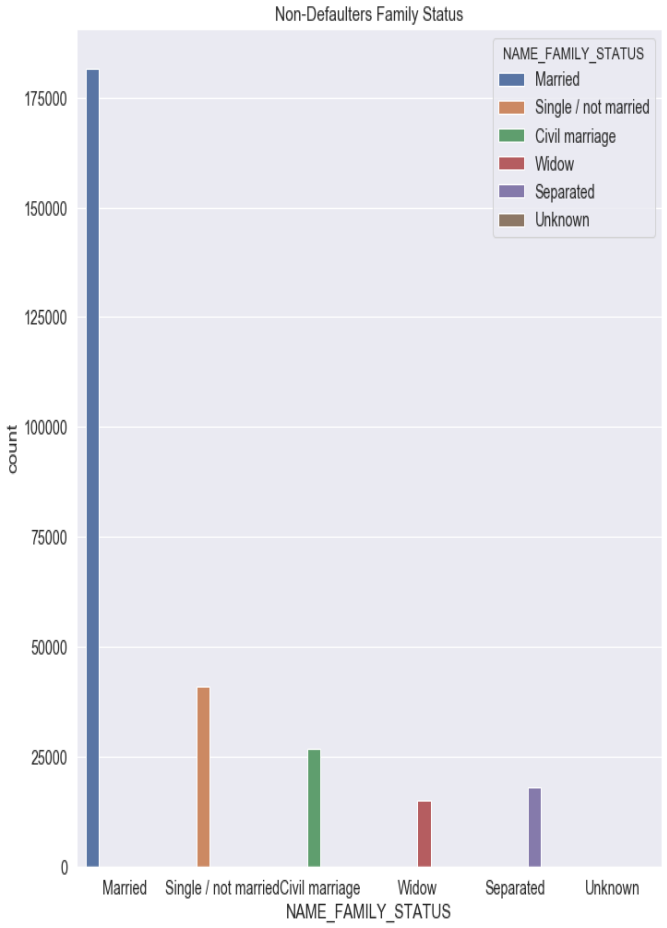
In the defaulters target group(on the right), the people owning a realty property are noticed to have the highest number and we can conclude that the people owning a realty can default payments, potentially because they have an added m responsibility.

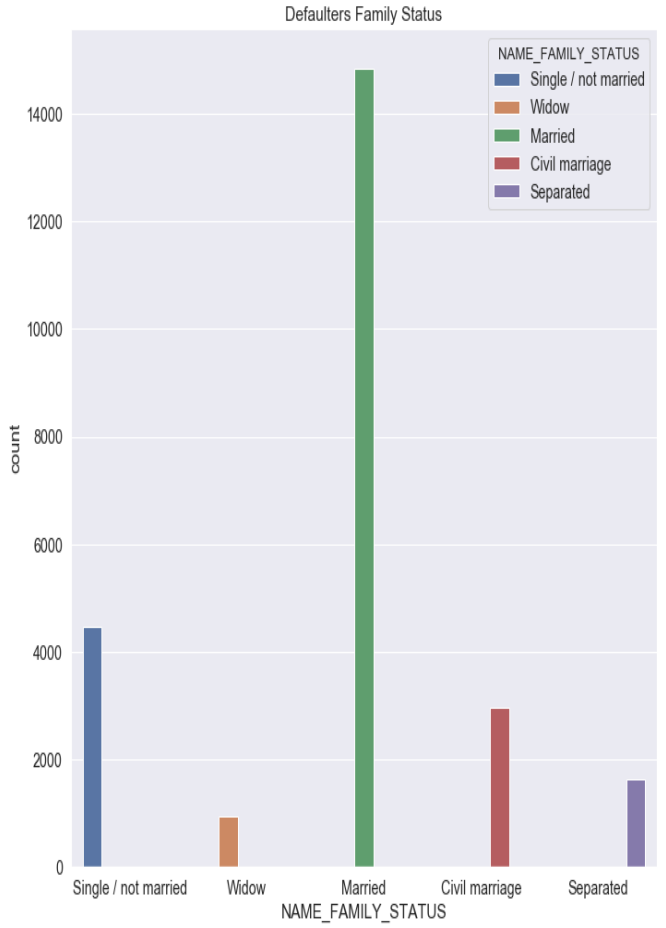


Univariate Segmented Analysis

For Univariate segmented analysis, we have segmented the target=1 group on the basis of their Family status and it is noticed that the people who have married or single status are the ones highest in the defaulting the loans.

Family status, thus becomes an important aspect of analysis to identify the defaulters.

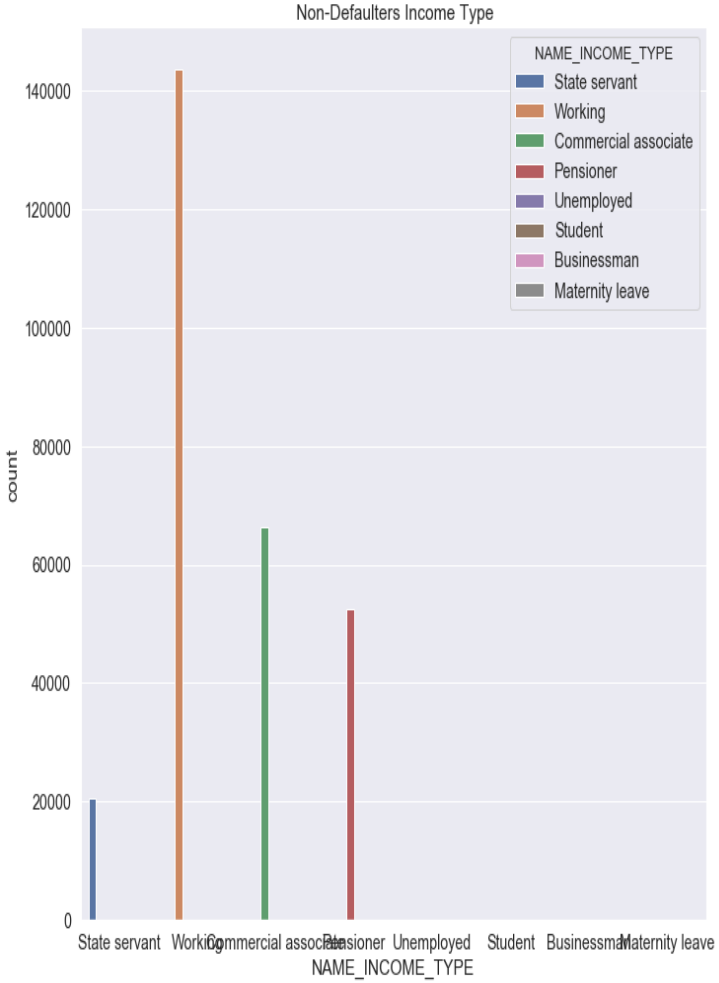


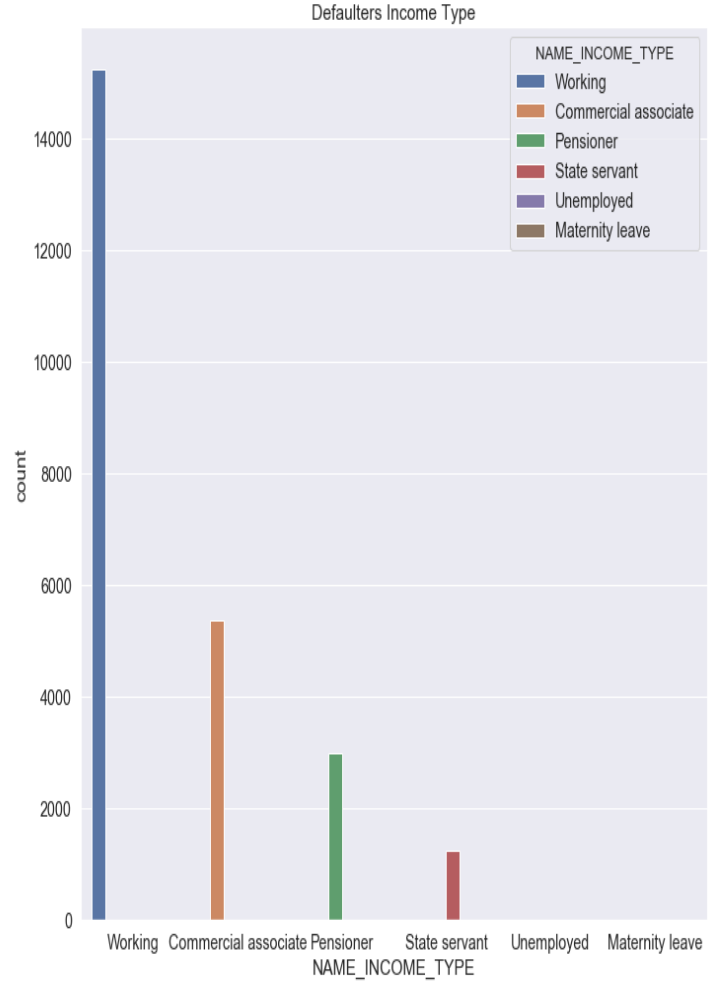


For Univariate segmented analysis, we have segmented the target=1 group on the basis of their Income type and it is noticed that the people who are working have the highest number in the defaulting the loans.

Also, if we closely analyse, then the people in category of business and students have never defaulted the loans(non defaulters group, on the left).

Thus, we can conclude, that while giving credits we have to be mindful of the people or are not business people or students.

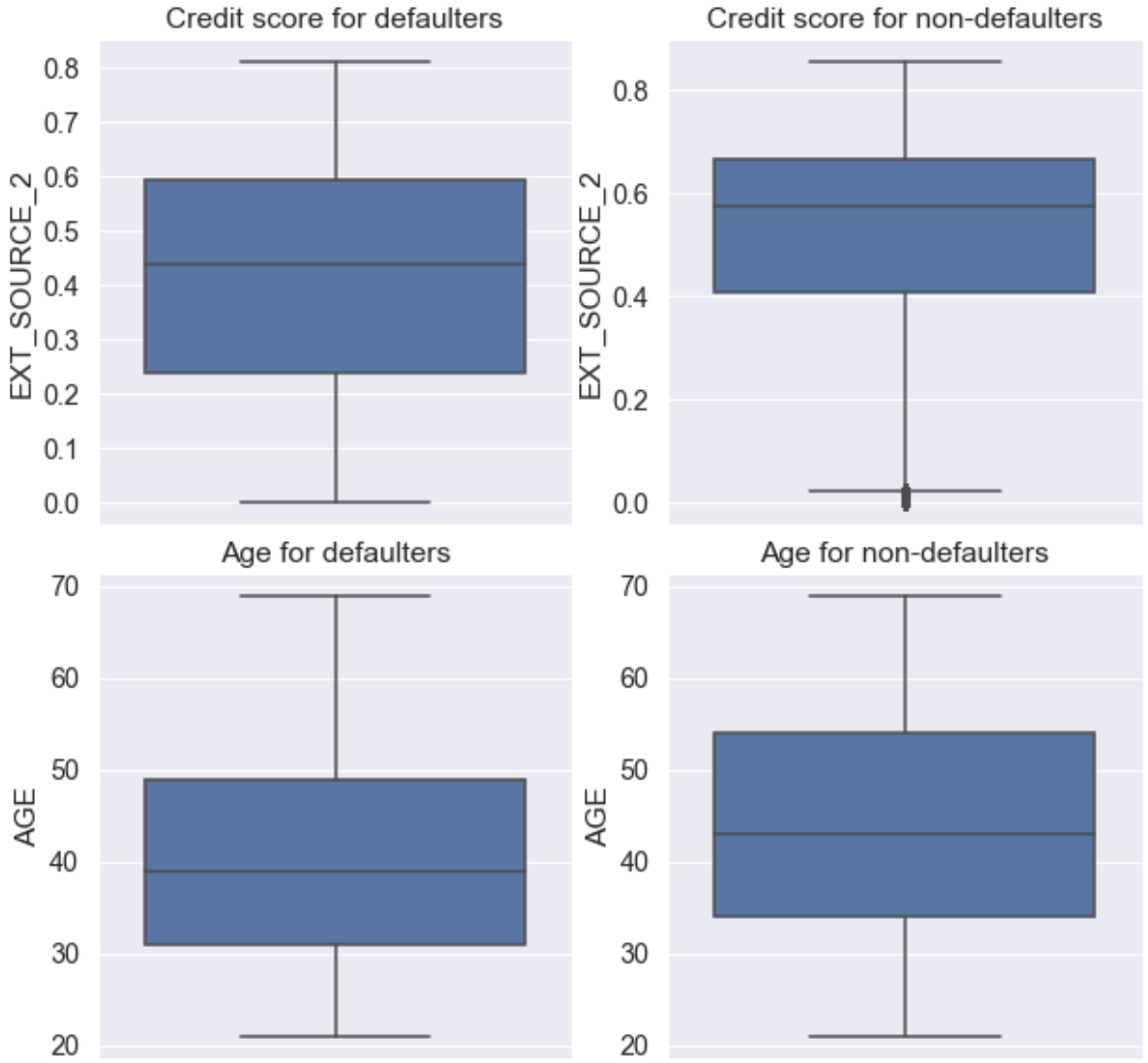




Univariate Quantitative columns analysis

- For univariate analysis of quantitative variables, **the credit score or ext_source_2** is a crucial variable.
- It is evident from the plot that the defaulters have much lesser credit rating as compared to the non-defaulters group. The range from 25th to 75th percentile of defaulters lie in 0.25 to 0.6, while it is as much higher as 0.4 to 0.65 for non defaulters. Hence, while sending out loans, it is imperative to look for clients with higher credit rating.

- **2nd quantitative variable** we have use for univariate non-categorical analysis is **Age** of the customer. It is seen as a trend that the age of the defaulter group is relatively lesser than the that of non-defaulter group.
- The median of the age of non-defaulters is above 40 and median of age of defaulters is below 40; also the 75th percentile of non defaulters is much higher than that of defaulters, which clearly indicates that the defaulters are of much younger age and financial liberty to repay loans at a younger age would be less as compared to the senior age group.

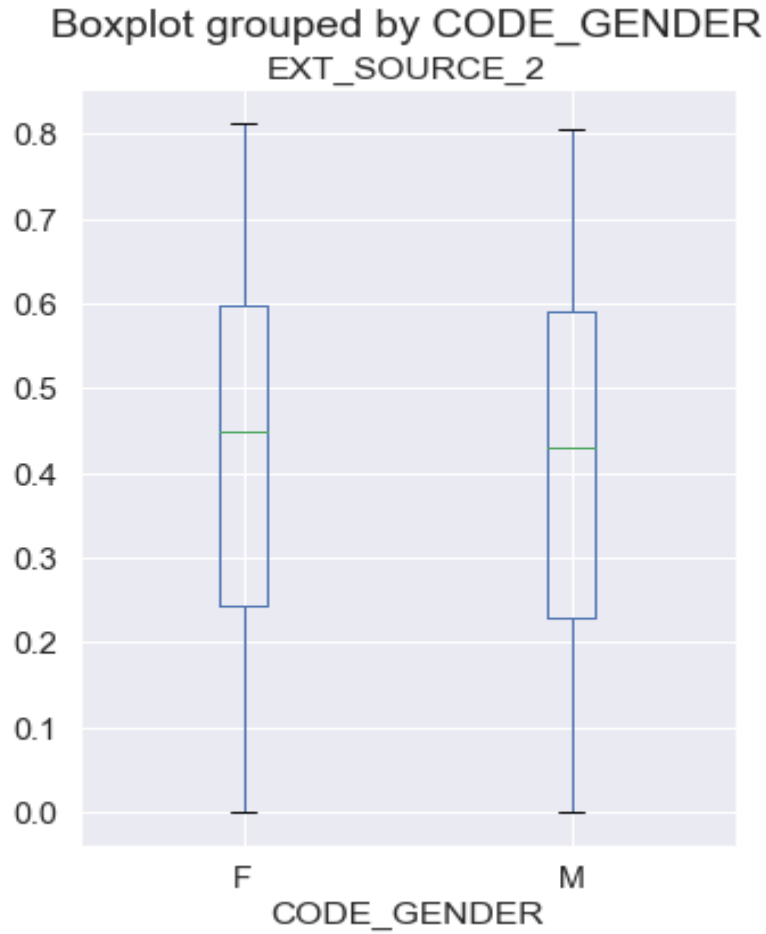


Bivariate Analysis

The Graph shows the bivariate analysis for two variables - **credit score on the basis of gender for the defaulters group.**

- As, we have analyzed these two variables separately in univariate analysis and concluded that these two are potential candidates to be mindful of while distributing the loans, the below chart helps us to drill down further to see, that the credit score of the females is higher than that of males.

- This means that, in the group of defaulters, because the rating of males is low, we shall make the male gender as flag with this credit rating.

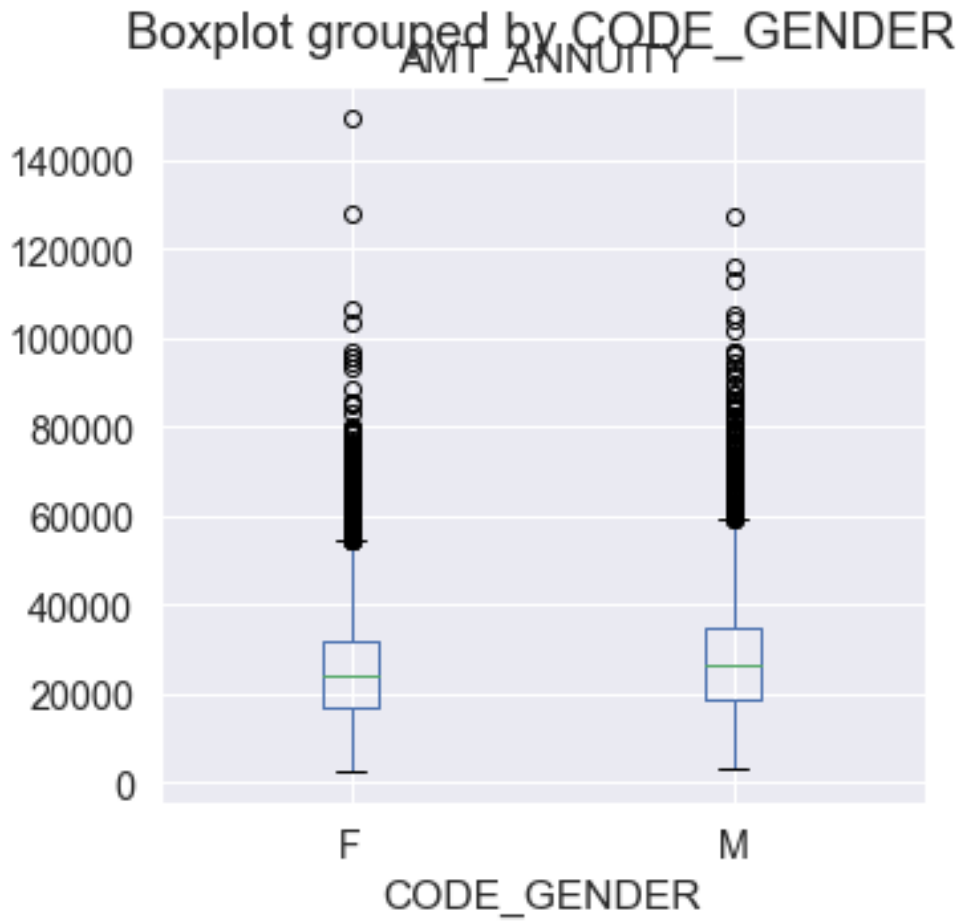


example-2

The Graph shows the difference between the annuity paid by the customer group by gender in target=1 group.

- It turns out that the male population has a higher annuity to be paid as compared to the female population.

- In the above, graph we saw that males have a lesser credit rating. It is a possibility, that because of higher annuity amount, their credit rating is low as repayments could be a problem.



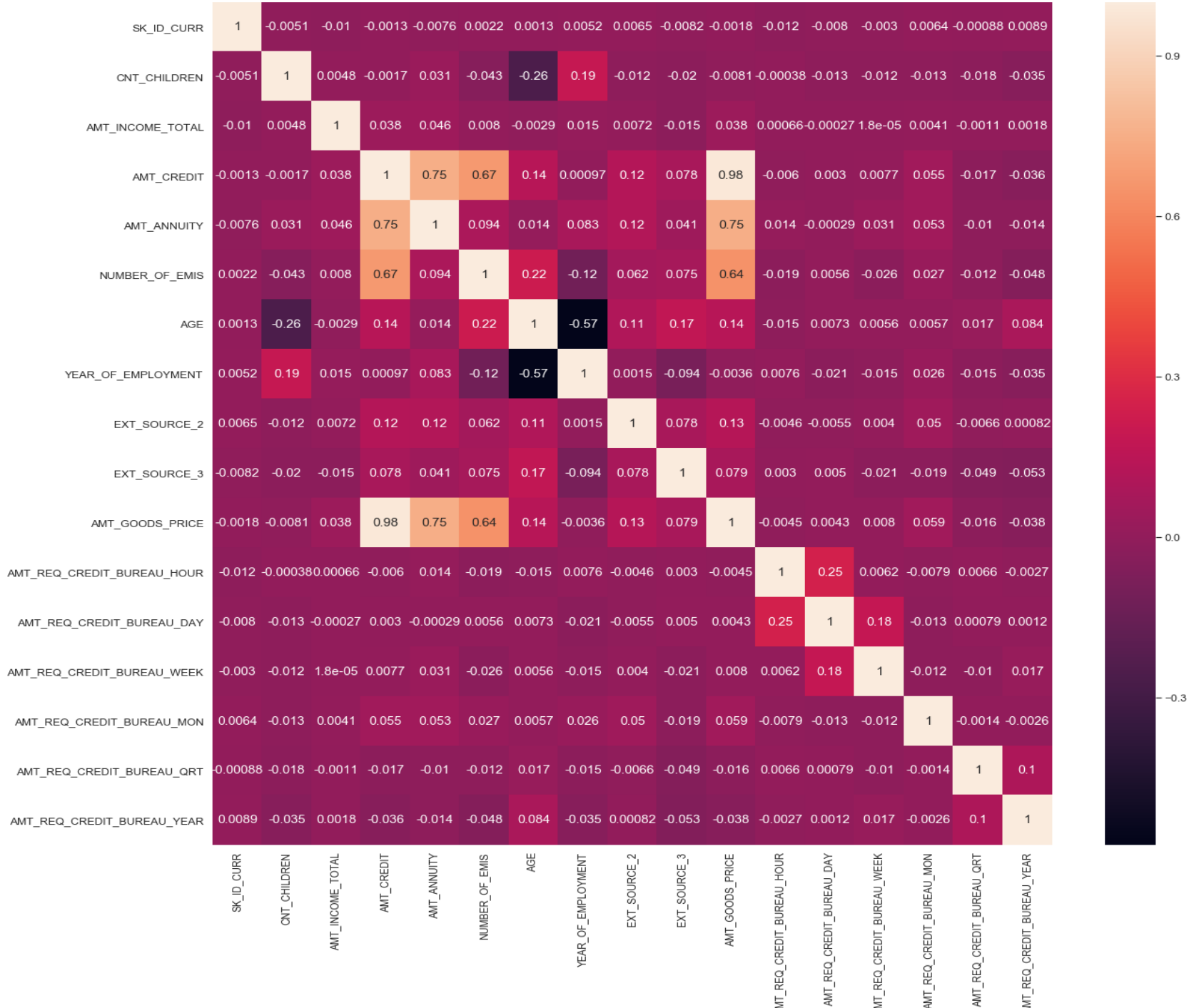
Bivariate analysis with correlation :

The below result gives a clear picture of how annuity, credit and income are positively corelated to each other.
This means that, as the income goes up, the credit and annuity go up with it. This means, that the person, with a higher income can gets a higher credit with higher annuity. This is for target=1, which means that the higher income attracts higher credit and annuity and can be a reason for default payments.

	AMT_ANNUITY	AMT_CREDIT	AMT_INCOME_TOTAL
AMT_ANNUITY	1.000000	0.752195	0.046421
AMT_CREDIT	0.752195	1.000000	0.038131
AMT_INCOME_TOTAL	0.046421	0.038131	1.000000

Heat Map Presentation for Defaulters.

The Grape shows the correlation between various variables for Defaulters.



As show in the above Graph, the top 10 DEFAULTERS having Payment difficulties are :

- **AMT_CREDIT** and **AMT_GOODS_PRICE** having correlation of 0.983103
- **AMT_ANNUITY** and **AMT_GOODS_PRICE** having correlation of 0.752699
- **AMT_CREDIT** and **AMT_ANNUITY** having correlation of 0.752195
- **AMT_CREDIT** and **NUMBER_OF_EMIS** having correlation of 0.672320
- **NUMBER_OF_EMIS** and **AMT_GOODS_PRICE** having correlation of 0.644761
- **AMT_REQ_CREDIT_BUREAU_HOUR** and **AMT_REQ_CREDIT_BUREAU_DAY** having correlation of 0.246741
- **NUMBER_OF_EMIS** and **AGE** having correlation of 0.216718
- **CNT_CHILDREN** and **YEAR_OF_EMPLOYMENT** having correlation of 0.191942
- **AMT_REQ_CREDIT_BUREAU_DAY** and **AMT_REQ_CREDIT_BUREAU_WEEK** having correlation of 0.184098
- **AGE** and **EXT_SOURCE_3** having correlation of 0.171801

Insight

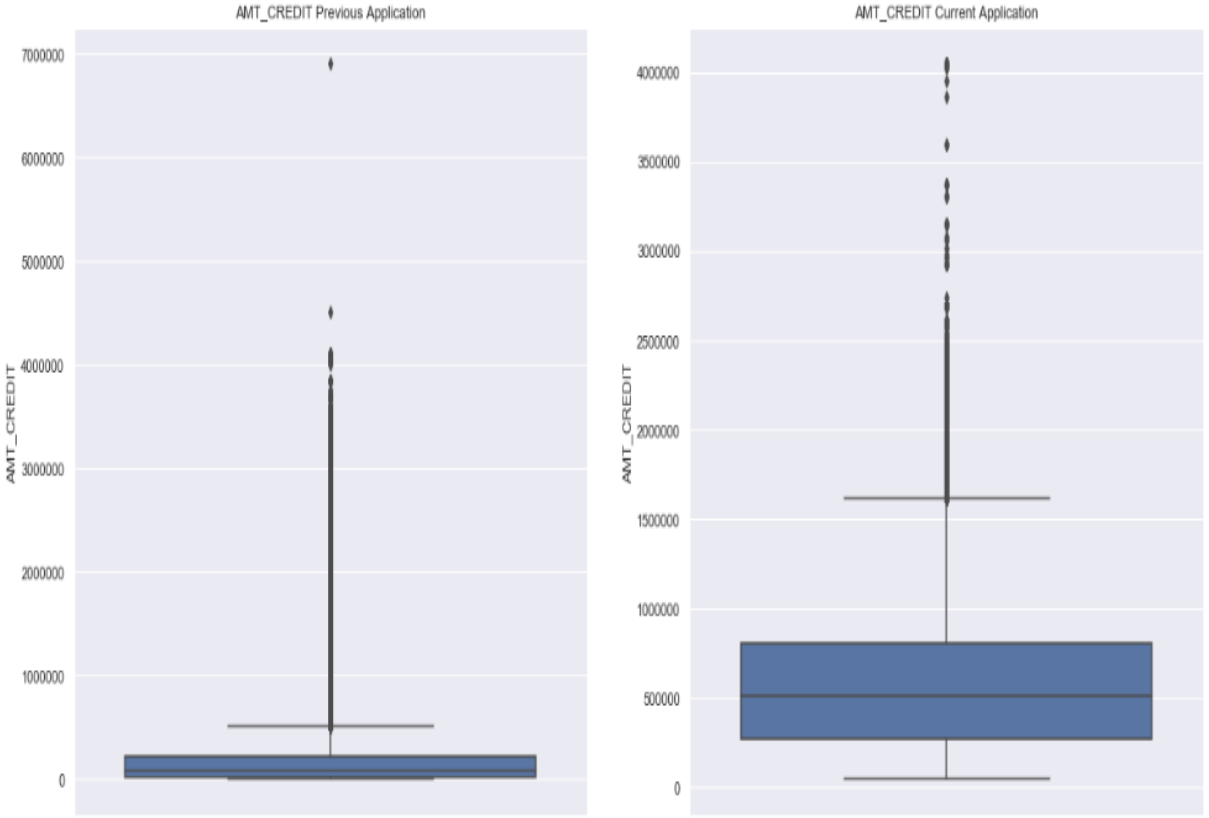
The very obvious inference is that goods price and credit amount are very strongly correlated. This means that the as the price of the goods go up, the credit given goes up to. And this would be because the application of the loan would be of higher amount. Similarly, because the annuity amount and the credit are positively correlated too, as the credit amount goes up, the annuity amount does.

Based on the above graphs, plots and calculations the Data Imbalance ratio here is 91.2 : 8.8

Previous Data-Univariate Analysis

This plot gives the distribution of the Credit given out in the previous application data set. It is good to notice that the full range of the of credit amount is distributed between 0 to 50000, while if we see on the right, which is the plot of credit amount of the current application data set, the median/50th percentile is at 50k.

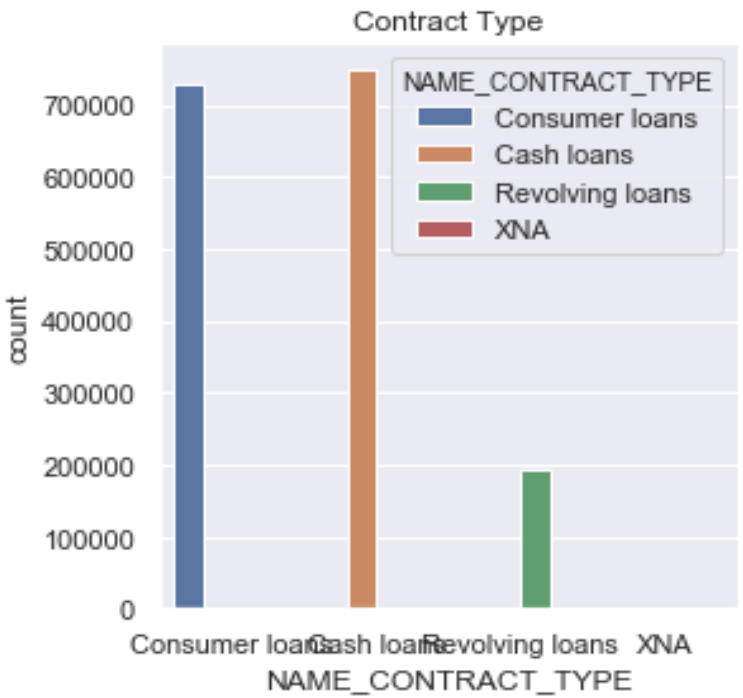
This means that the credit amount being sent out is much higher in the current application as compared to the previous app.



Previous Data-Segmented Univariate Analysis

From the below plot it is evident that the no. of cash loans sent out in the previous application is the highest.

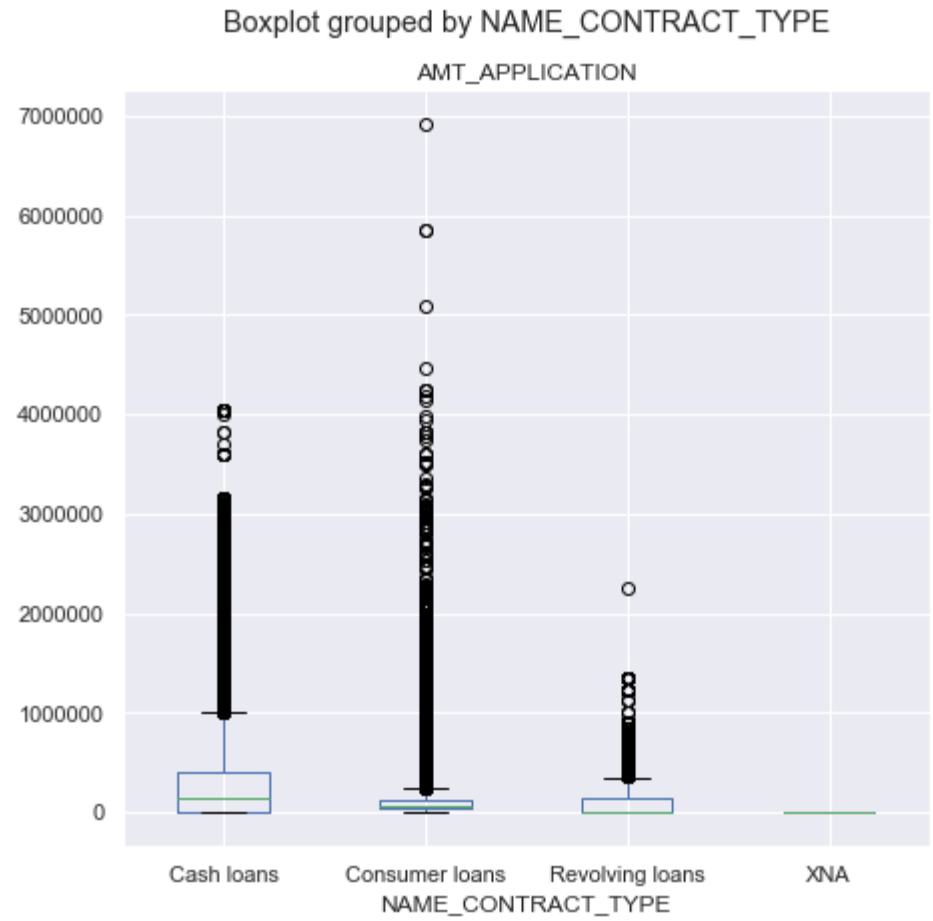
It would be interesting to find out whether the amount of the loan applied is also highest for cash loans or not. We will check this in the next segment of Bivariate analysis.



Previous Data-Bivariate Analysis

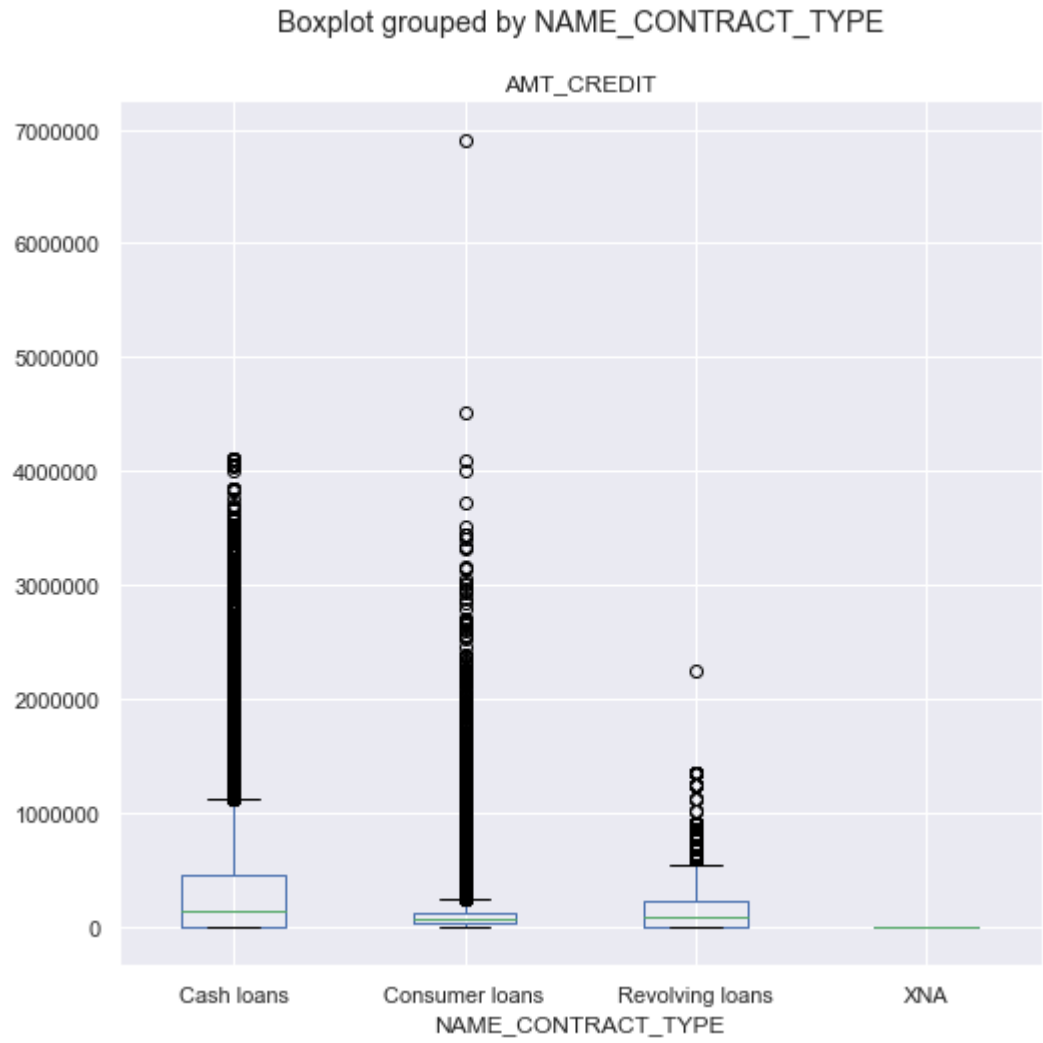
In the previous plot, we saw that the no. of cash loans was the highest in the previous app data. The below plot, drills down further to give us the insight that the application amount of cash loans sent were more than that of any other contract type.

Furthermore, it would be interesting to know that the credits given lie in the same range as of applied amount or not.



Previous Data-Bivariate Analysis

In the previous plot we saw the distribution of the applied amount wrt to the contract type and it cash loans had the highest amount. Now, we are going to see the distribution of the credit amount wrt to the contract type. And we can observe that the amount of credit given is highest for cash loans only. But if we observe keenly, we can see that the distribution of applied and credit amounts are different for cash loans. Amount given as credit is more than amount applied.
The 25th and 50th percentile seems similar, but definitely the 75th and 100th percentile of the credit amount is more than that of applied amount.



Previous Data-Bivariate analysis with corelation

The below correlation matrix, helps us to understand that the three amount variables move positively in the previous application.

Most strongly correlated are the credit amount and the goods price, at .99. this means that the credit given is near about the price of the good for which the loan is seeked. Similarly, the amount of good's price and amount of loan applied are strongly correlated too. And then the Credit amount and applied amount, which are strongly correlated at 0.97.
The inference we can draw is that- as the price of the good goes up, the person seeking the loan, applies for a greater amount.
And hence, when the applied amount goes up, the credit amount is given goes up too.

	AMT_APPLICATION	AMT_GOODS_PRICE	AMT_CREDIT
AMT_APPLICATION	1.000000	0.999884	0.975824
AMT_GOODS_PRICE	0.999884	1.000000	0.993087
AMT_CREDIT	0.975824	0.993087	1.000000

Analysis on variables to identify loan defaulters

Gender- It is a critical variable to consider while sending out the loans; though the no. of defaulters are females is higher than that of males but the data showed us that the credit rating of male defaulters is low. Furthermore the annuity of males is higher which may tend to make them default the loans. Hence Gender is an important aspect while finding defaulters.

Income Status- It was clearly observed in the analysis that the Business and Student group of people were never in the defaulters group, but rest others were in the defaulter group. Hence we should be mindful of the income type.

Goods Price- In both the data sets we have seen that the goods price is positively correlated to the credit and annuity, which means when the loan is applied for an expensive good, the credit and annuity go up and hence the risk of defaulting the loan.

Age- It was closely observed that the people of higher age group were able to repay loans better as compared to the people of younger age. Hence, Age becomes an important factor to consider.

Type of Contract- It has been observed in both the data sets, that the max. applications came for cash loans and the amount given as credit was highest for cash loans. Hence, cash loans become more prone to get defaulted and thus we shall check the type of loan to find the defaulters.