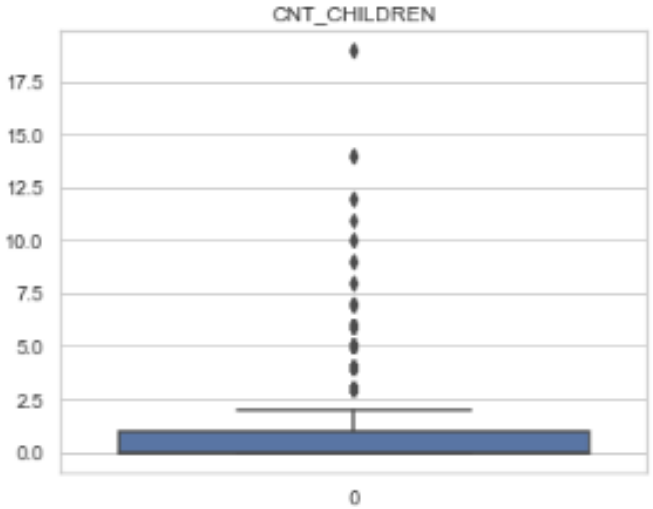
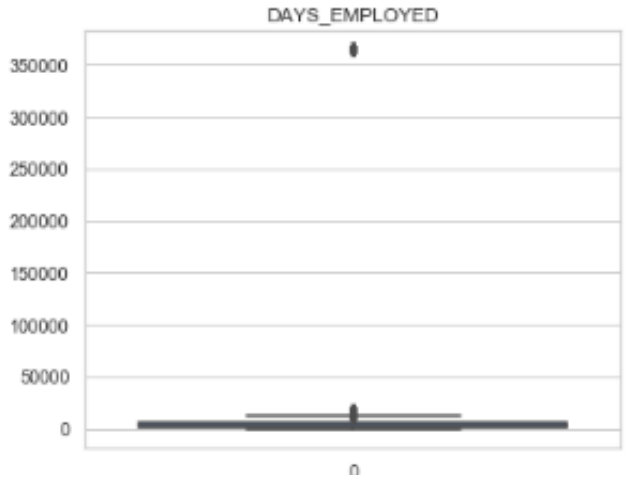
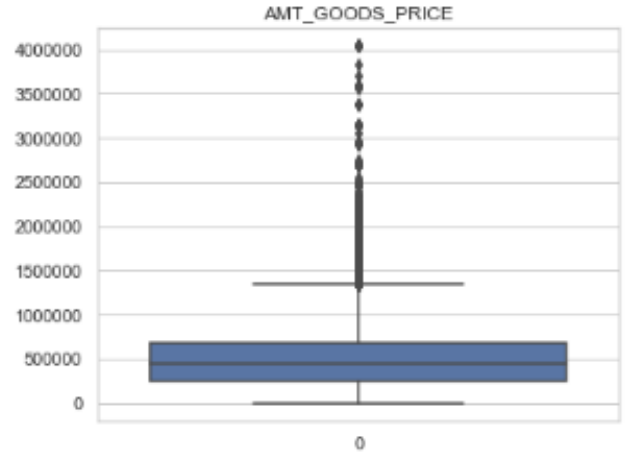
EDA_Group_Case_Study

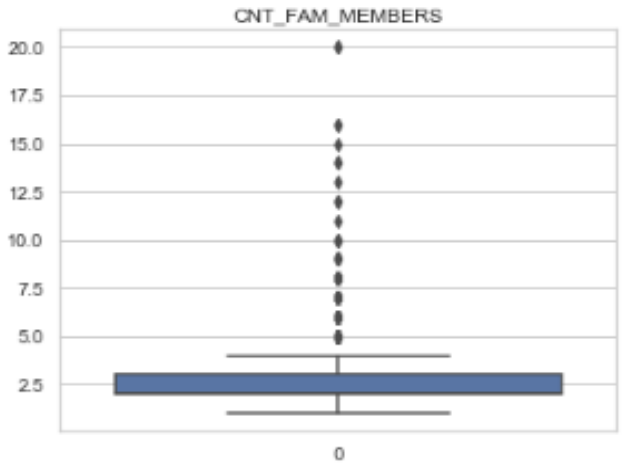
Ankur Sugandhi

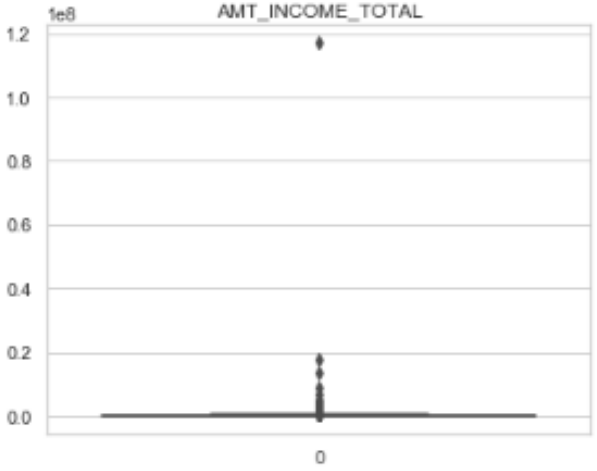
Outliers

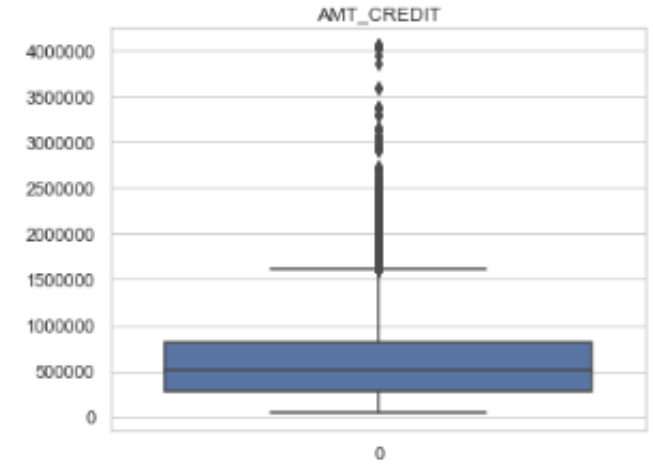






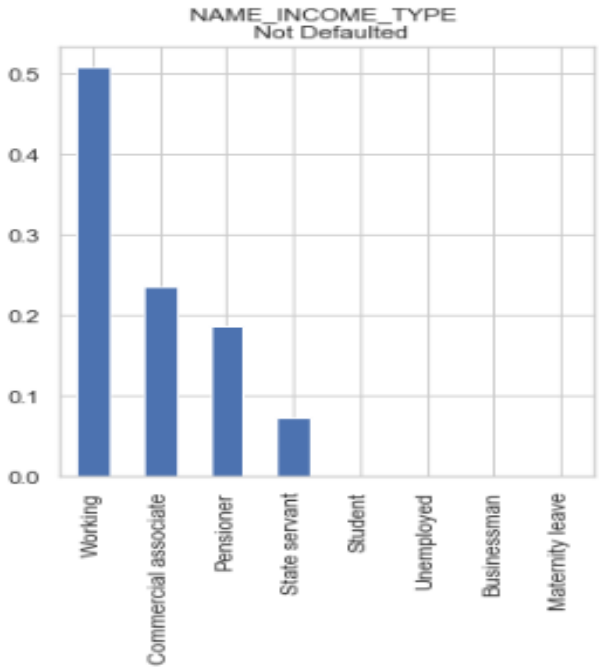


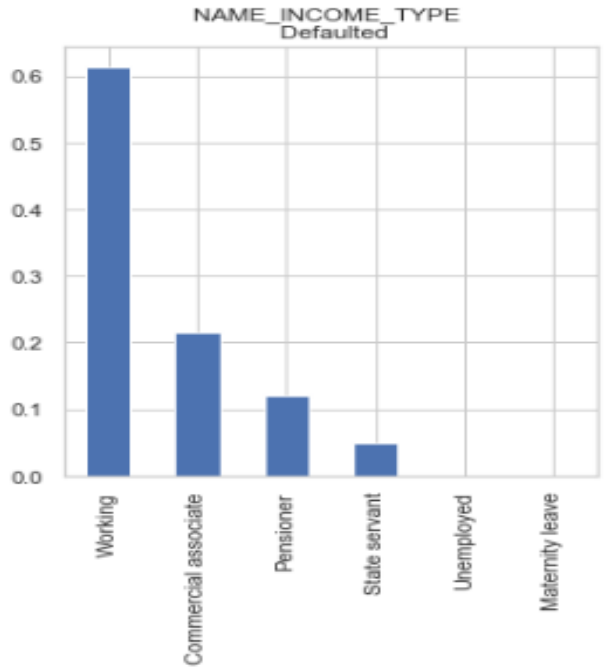


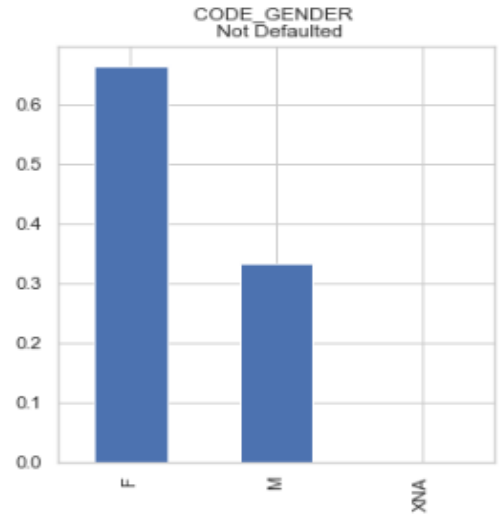


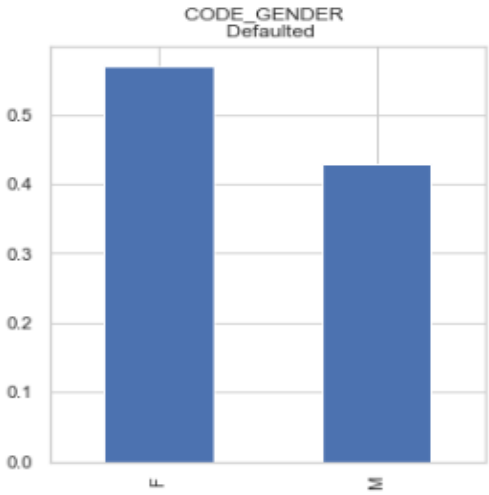
The attributes selected CNT_CHILDREN, DAYS_EMPLOYED, AMT_GOODS_PRICE, CNT_FAM_MEMBERS, AMT_INCOME_TOTAL, AMT_CREDIT, above graphs display the outliers for the same.
CNT_CHILDREN, CNT_FAM_MEMBERS, AMT_CREDIT and AMT_GOODS_PRICE has higher outliers compared to DAYS_EMPLOYED and AMT_INCOME_TOTAL

univariate analysis for categorical variables in both the datasets





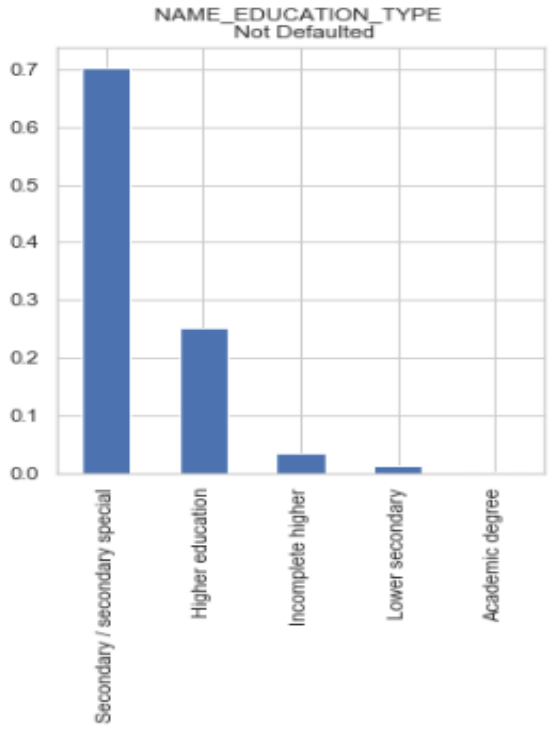


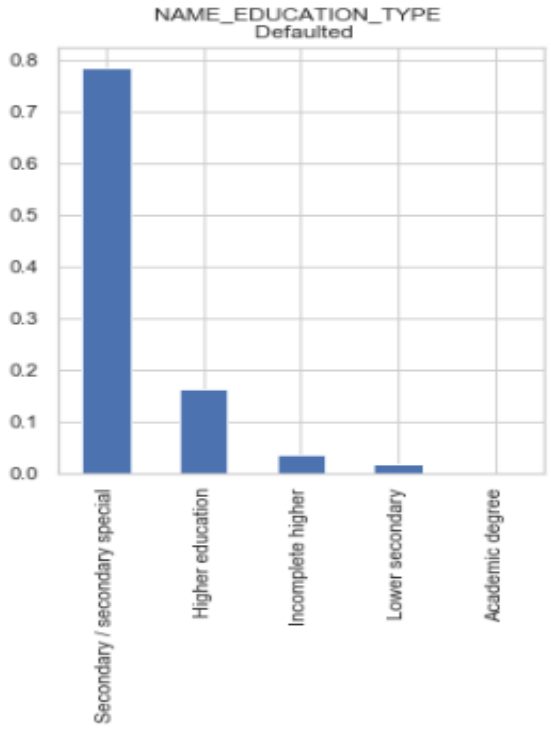


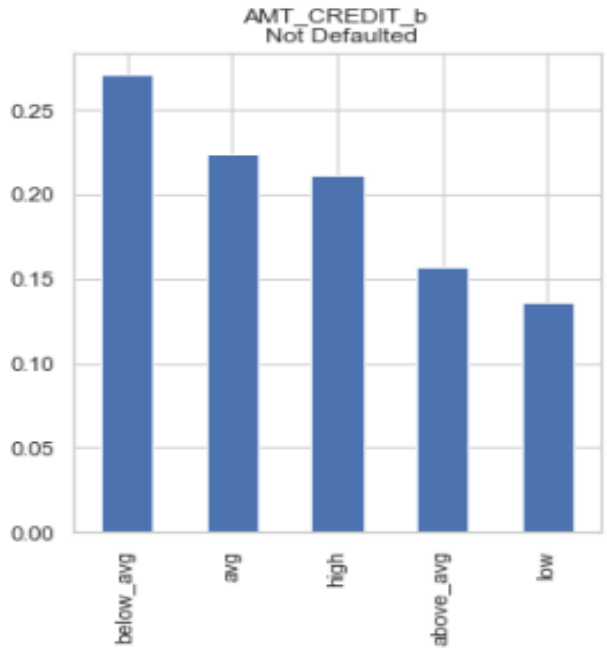
For the characteristic NAME_INCOME_TYPE, no specific correlation found w.r.t working, commercial associate, pensioner, state servant, student, unemployed and businessmen. But if we check the absolute numbers, defaulted rate is more in case of working people

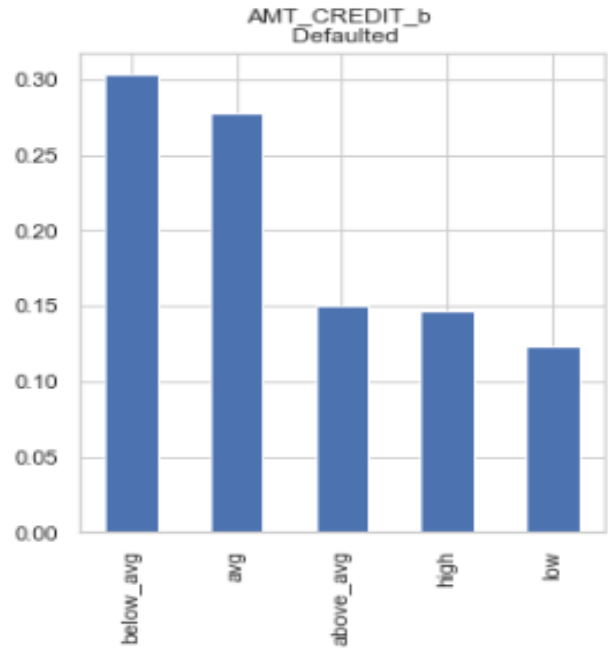
For the characteristic CODE_GENDER, attributes Female is higher for both in Defaulted and Not Defaulted hence no correlations observed. But in the case of Males Not defaulted is less compared to Defaulted hence the correlation that the defaulted males are more than non defaulted.

Univariate analysis for categorical variables in both the datasets









For the characteristic NAME_EDUCATION_TYPE, no specific correlation found w.r.t secondary special, higher education, incomplete higher education, lower secondary education and academic degree. But if we follow the absolute numbers, defaulted rate is more in case of secondary special.

For the characteristic AMT_CREDIT_b, defaulted is more for below average and average, whereas for high, above average and low amount credit the defaulted are less. It shows that the high and above high are would be managing because of the discipline.

Lets find correlation between the variables for both df1 & df2 datasets

df1

	Var1	Var2	Correlation	Correlation_abs
46	AMT_GOODS_PRICE	AMT_CREDIT	0.986966	0.986966
99	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571	0.878571
83	DAYS_EMPLOYED	DAYS_BIRTH	0.626114	0.626114
45	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349473	0.349473
23	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799	0.342799
66	DAYS_BIRTH	CNT_CHILDREN	-0.336966	0.336966
105	CNT_FAM_MEMBERS	DAYS_BIRTH	-0.285823	0.285823
95	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.276663	0.276663
94	DAYS_ID_PUBLISH	DAYS_BIRTH	0.271314	0.271314
77	DAYS_EMPLOYED	CNT_CHILDREN	-0.245174	0.245174

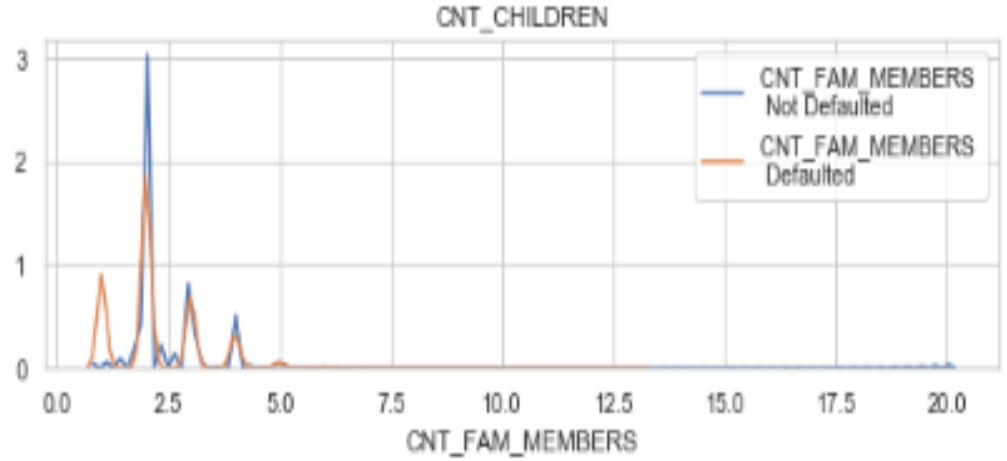
df2

	Var1	Var2	Correlation	Correlation_abs
46	AMT_GOODS_PRICE	AMT_CREDIT	0.982854	0.982854
99	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484	0.885484
83	DAYS_EMPLOYED	DAYS_BIRTH	0.582185	0.582185
66	DAYS_BIRTH	CNT_CHILDREN	-0.259109	0.259109
94	DAYS_ID_PUBLISH	DAYS_BIRTH	0.252863	0.252863
95	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.229090	0.229090
105	CNT_FAM_MEMBERS	DAYS_BIRTH	-0.203267	0.203267
77	DAYS_EMPLOYED	CNT_CHILDREN	-0.192864	0.192864
106	CNT_FAM_MEMBERS	DAYS_EMPLOYED	-0.186515	0.186515
70	DAYS_BIRTH	AMT_GOODS_PRICE	0.135516	0.135516

For df1 number 46, 99 and 83 has the highest correlation and 105, 95, 94 and 77 has the lowest correlation
For df2 number 46, 99 and 83 has the highest correlation and 70, 106 and 77 has the lowest correlation

In both the data base the highest correlating tables are same

Performing univariate for numerical variables in both the datasets and compare it



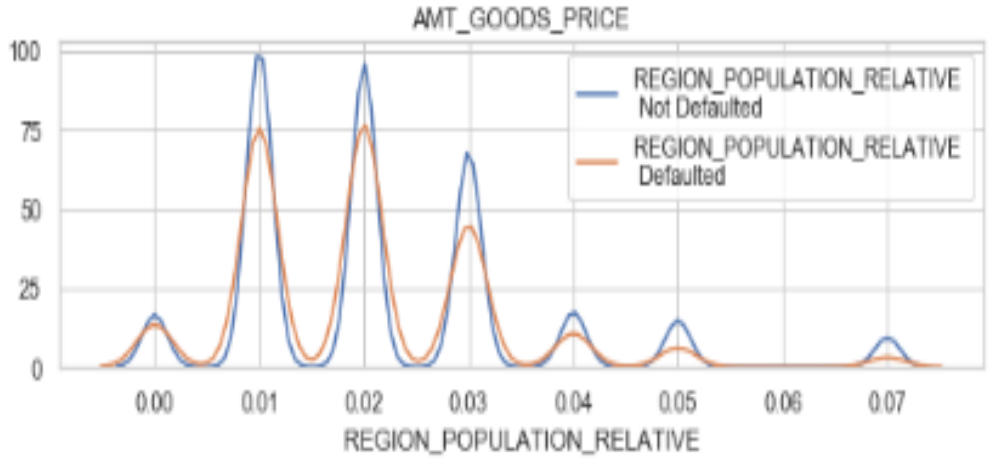
For the characteristic Count Children the trend for defaulted and not defaulted follows the same trend

The rate of Not defaulted is more for the Count of Family members between 1-2.5

For the characteristic Days Employed the trend for defaulted and not defaulted follows the same trend

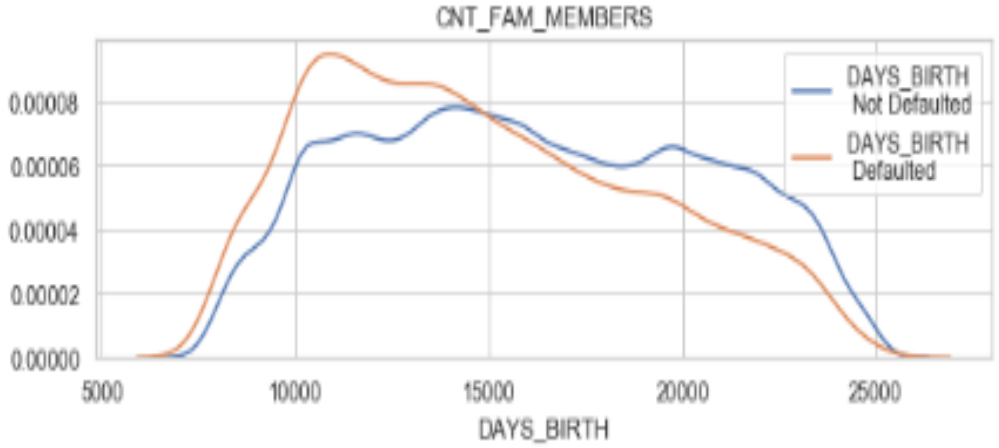
We cannot interpret any direct understanding from the above graph

Performing univariate for numerical variables in both the datasets and compare it



For the characteristic Amt Goods Price, the trend for defaulted and not defaulted follows the same trend

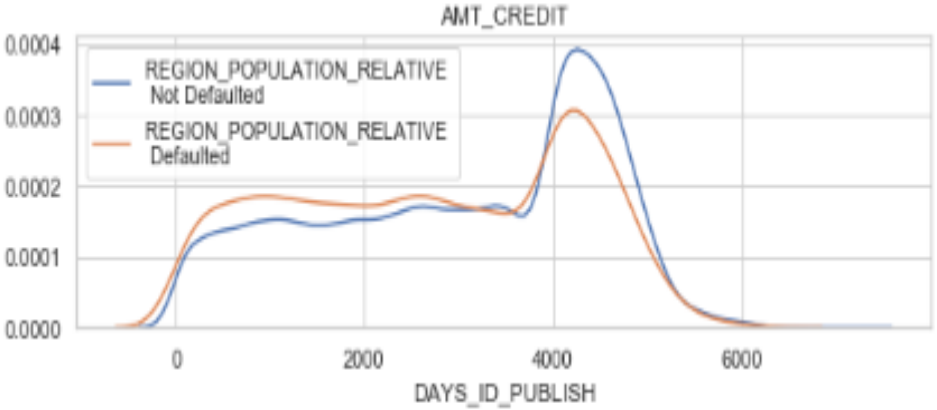
We cannot interpret any direct correlation from the above graph, but one thing we can understand the Not Defaulted is higher for all the cases of Region Population Relative



For the characteristic Count Family Members in the range of 10000 to 15000 Days of Birth and if family members are high then Defaulted are more, similarly as the Days of Birth increases the rate of Defaulted people decreases in other words the Not Defaulted increased as the days of birth increases

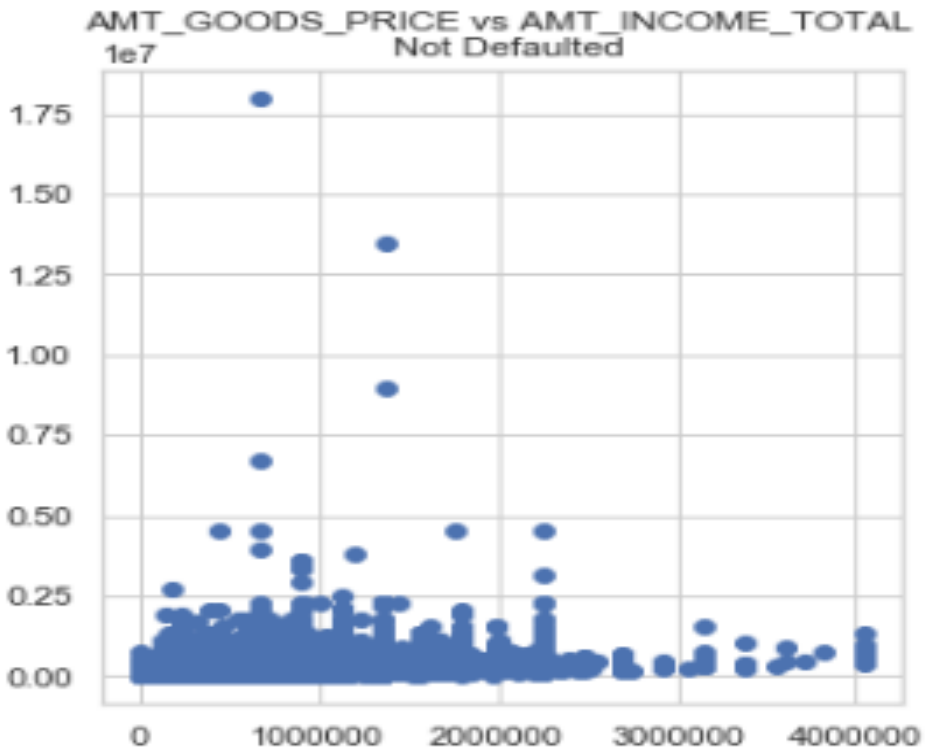
Performing univariate for numerical variables in both the datasets and compare it

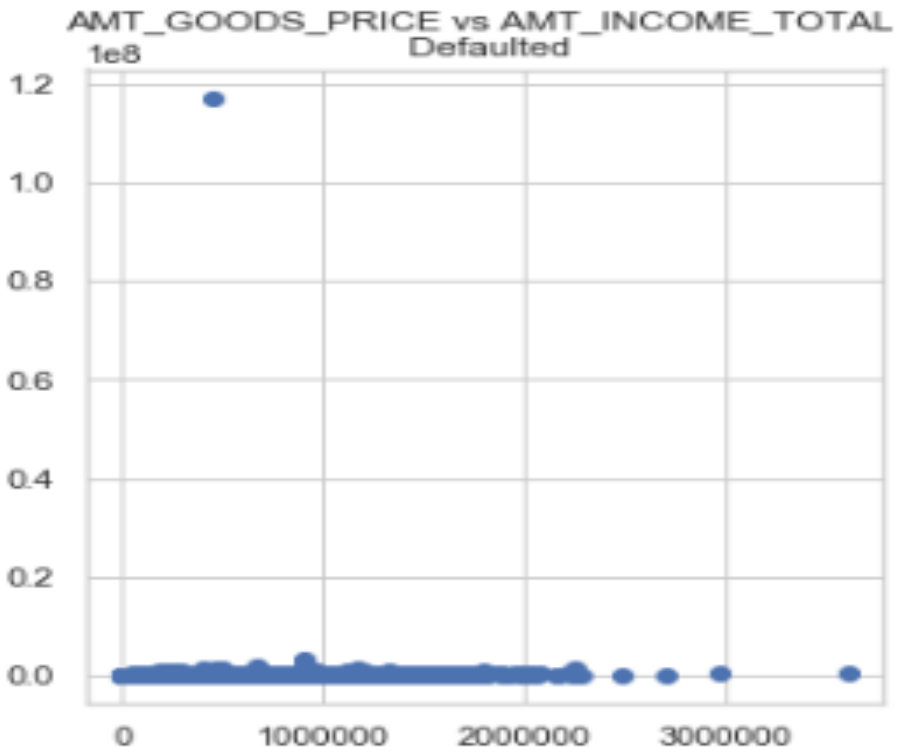
For the characteristic AMT_INCOME_TOTAL with respect to DAYS_EMPLOYED has no correlation, and we cannot interpret any data from the above graph.



In the Characteristic AMT_CREDIT we can understand from the above graphical representation that as the Days of ID Publish increases the Not Defaulted also increases

Performing Bivariate analysis for numerical variables

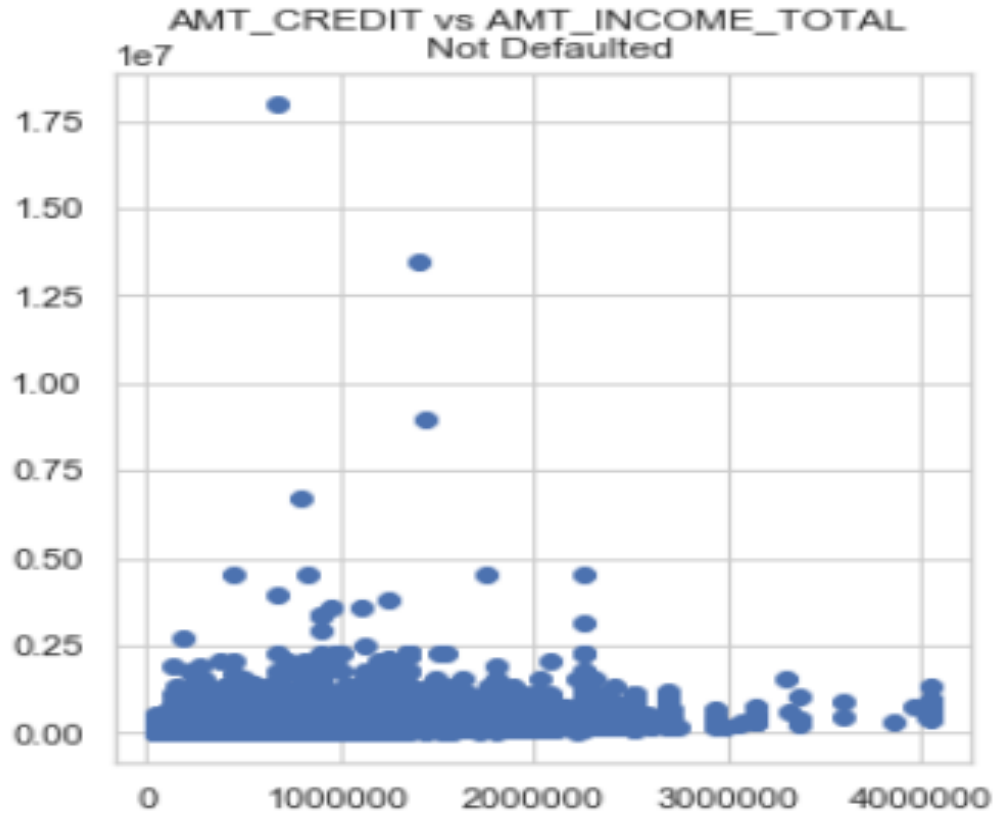


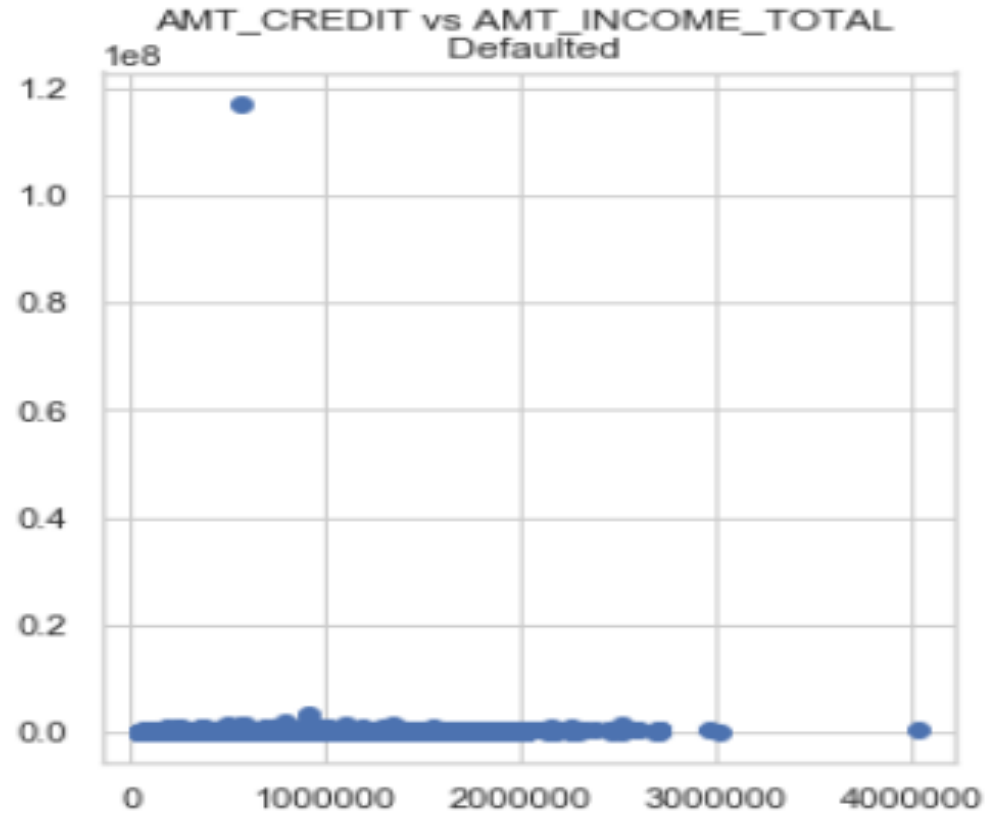


Amt Good Price Vs Amt Income Total,
In the case of Not Defaulted if we avoid the outliers we can understand that for Not Defaulted , lower the Amt Income Total higher the Amt Good Price

In the case of Defaulted the graphical representation shows no correlation

Performing Bivariate analysis for numerical variables

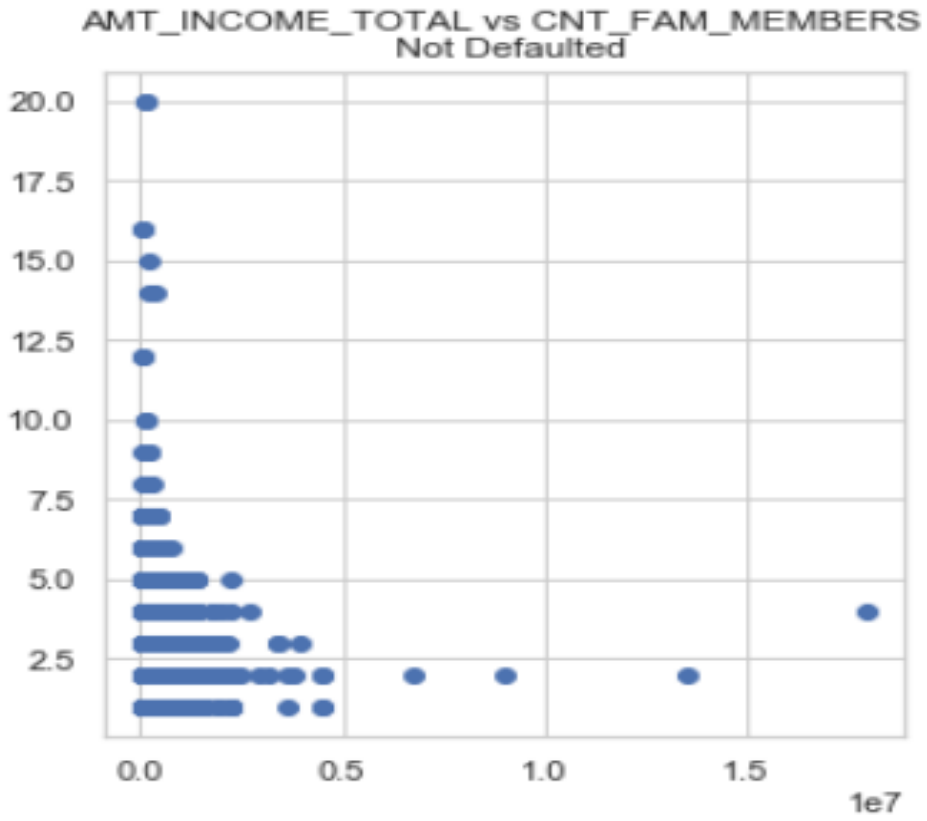


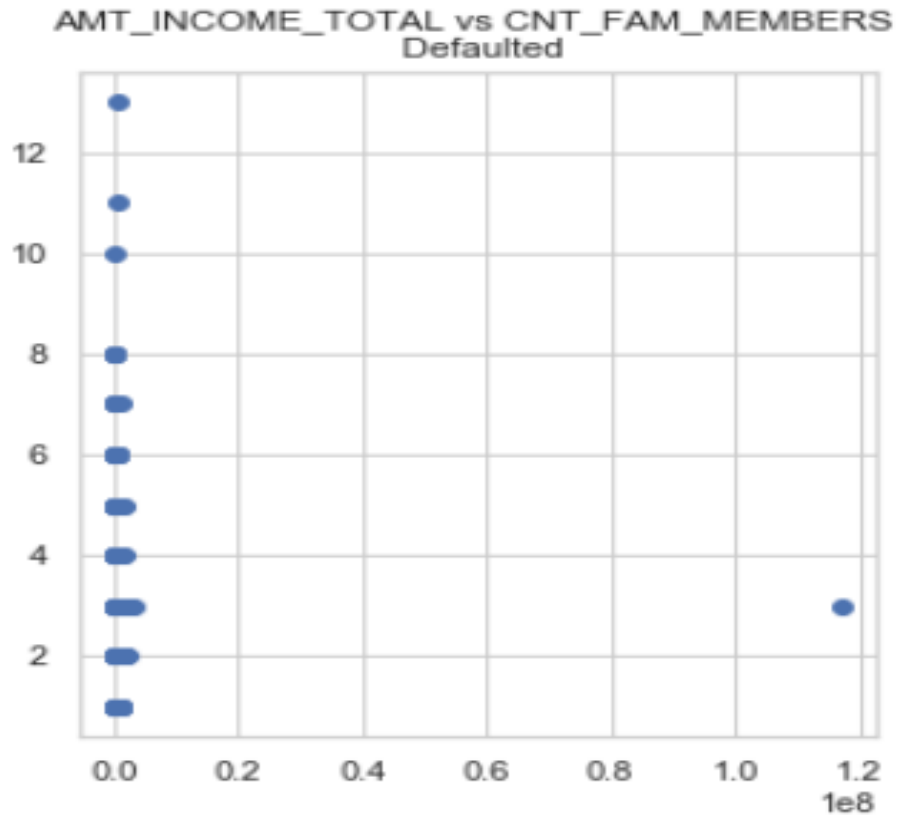


Amt Credit Vs Amt Income Total,
In the case of Not Defaulted if we avoid the outliers we can understand that for Not Defaulted , lower the Amt Income Total higher the Amt Credit

In the case of Defaulted the graphical representation shows no correlation

Performing Bivariate analysis for numerical variables

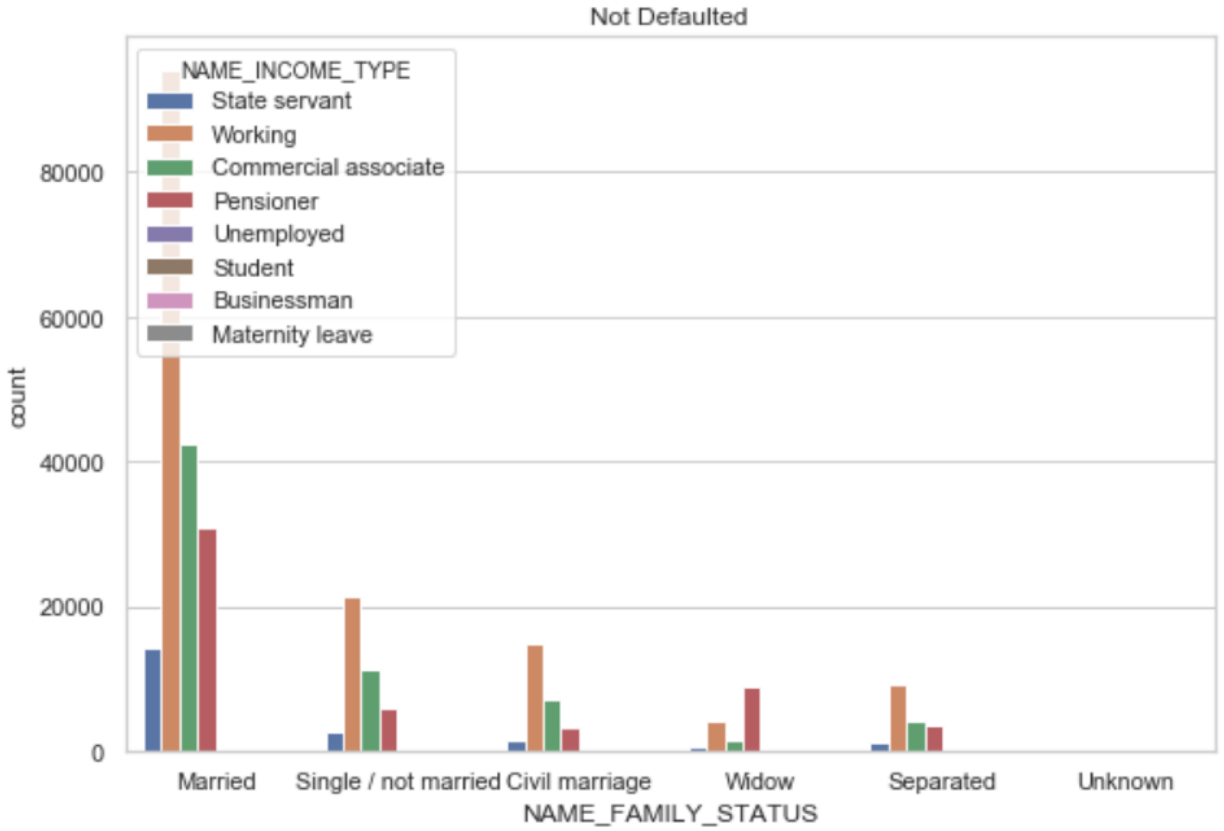


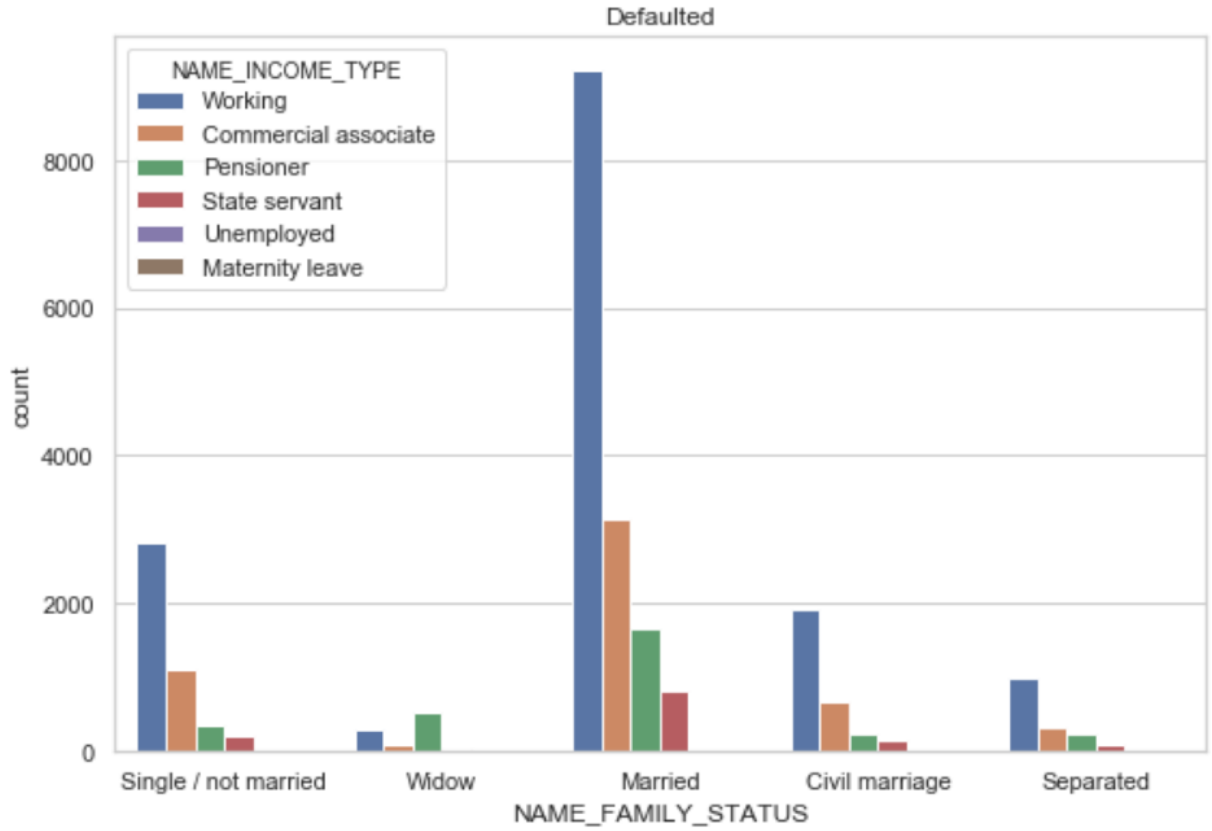


Amt Income Total Vs Cnt Fam Members,
In the case of Not Defaulted if we avoid the outliers from the data we can understand that , in the lower Amt Income as the maximum family member is 1 hence they are Not Defaulted.

In the case of Defaulted the graphical representation shows no correlation

bivariate analysis for categorical variables

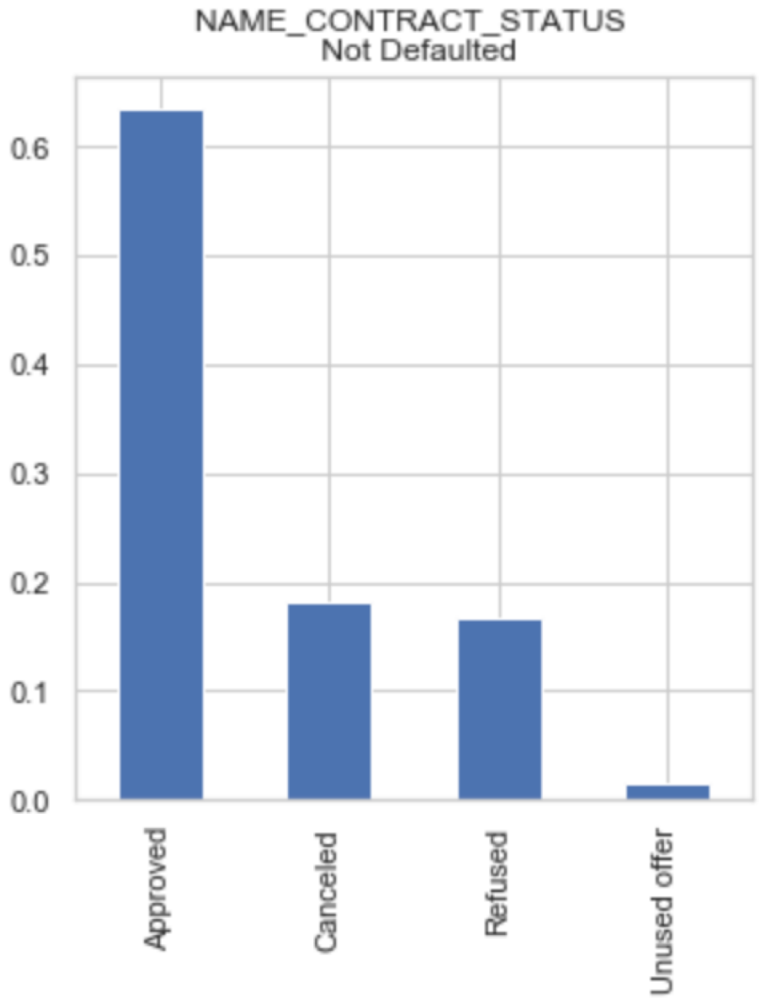


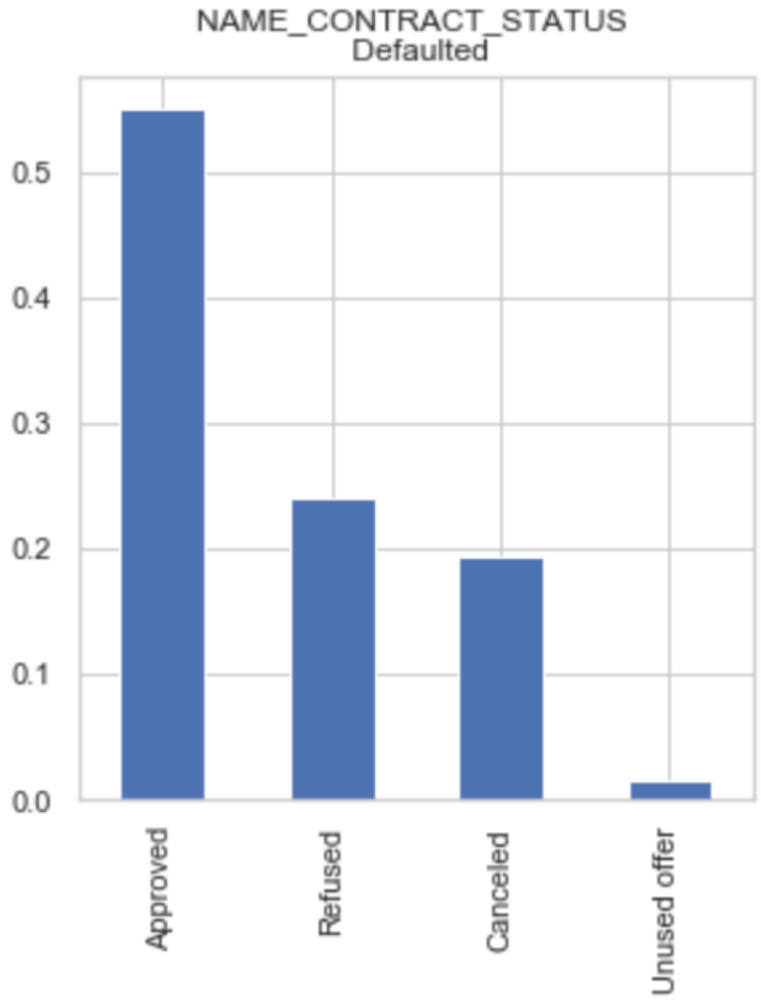


NAME_FAMILY_STATUS Vs NAME_INCOME_TYPE,
The above observation indicates that single/non-married and are slightly more indicative to default

Merging the Application and previous application dataset to understand the pattern

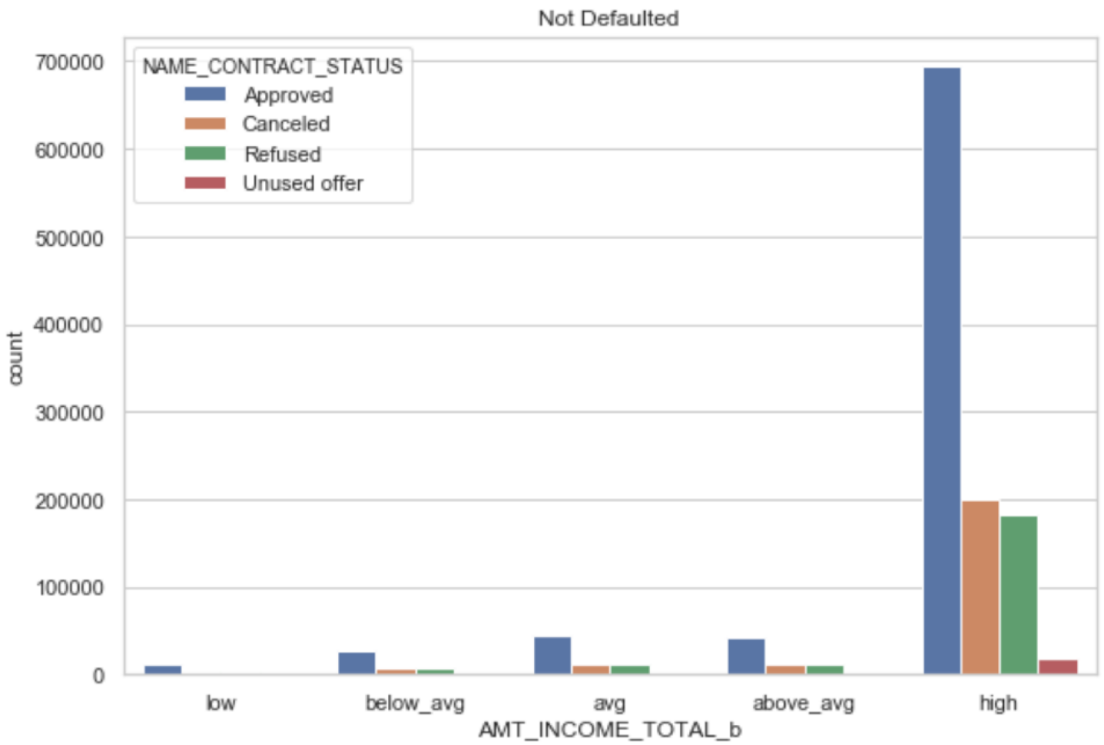
Univariate analysis for NAME_CONTRACT_STATUS

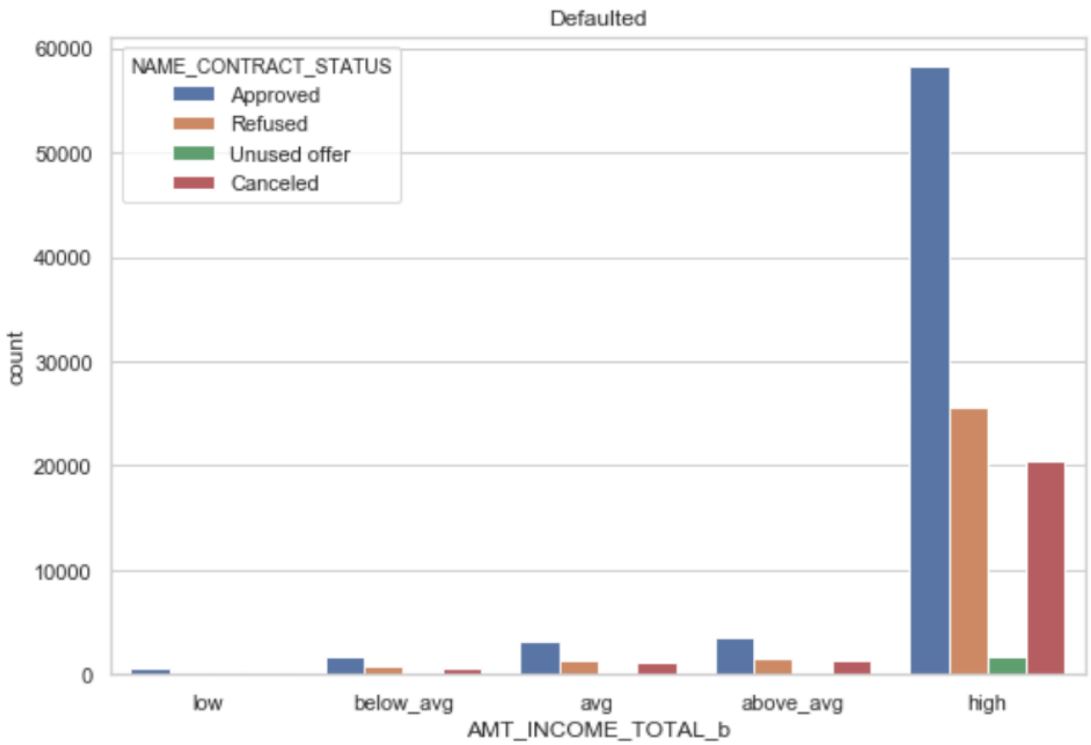




Name_Contract_status,
The above observation indicates 10% more refused cases in defaulted status

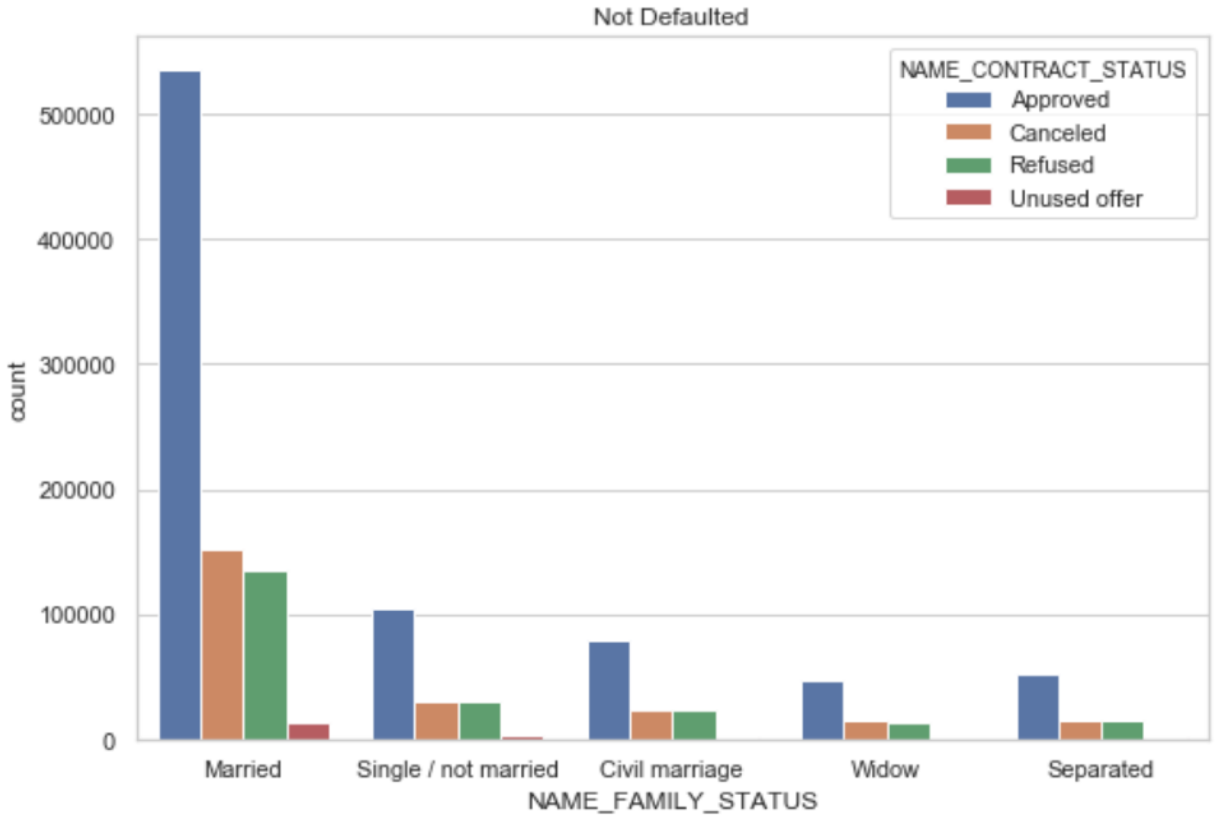
Bivariate analysis for NAME_CONTRACT_STATUS

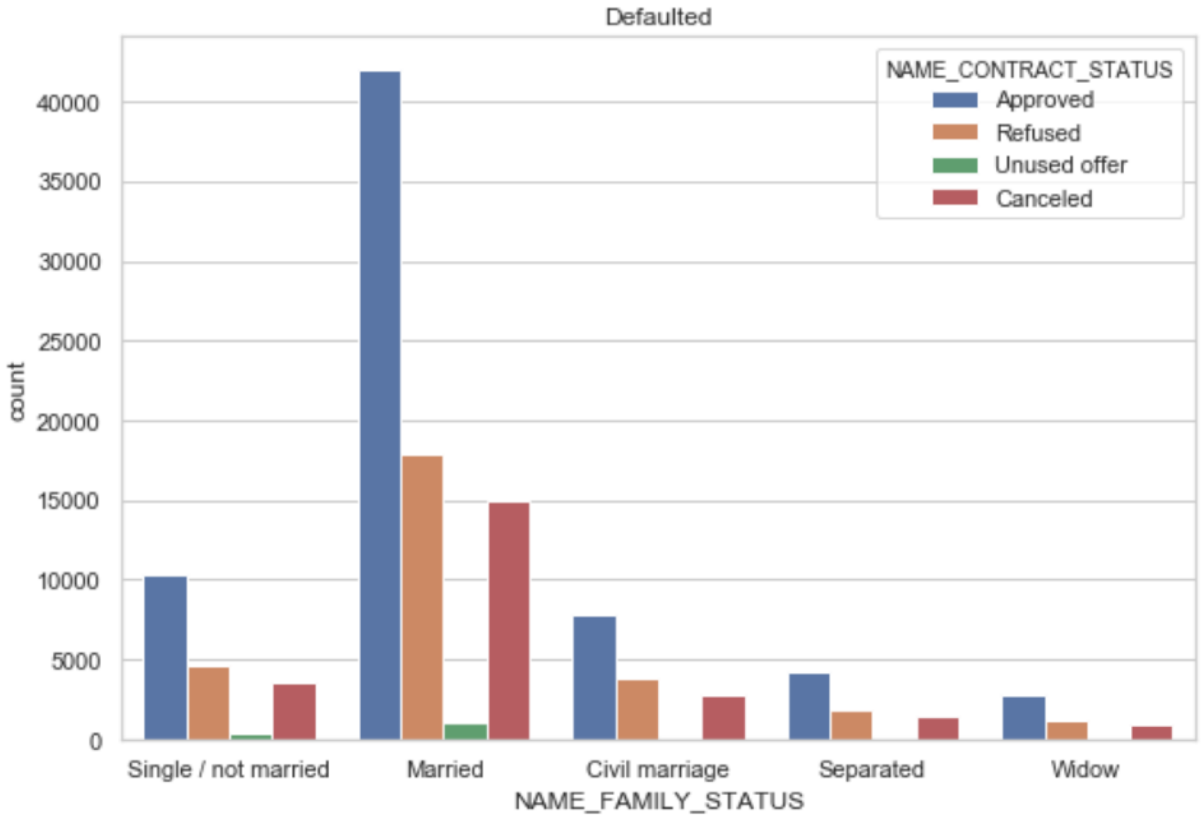




Name_Contract_status Vs AMT_INCOME_TOTAL_B
You can see that the refused status is more in defaulted graph for different income categories

Bivariate analysis for NAME_CONTRACT_STATUS





Name_Contract_status Vs NAME_FAMILY_STATUS
You can see that the refused status is more in defaulted graph for different Family_Status categories