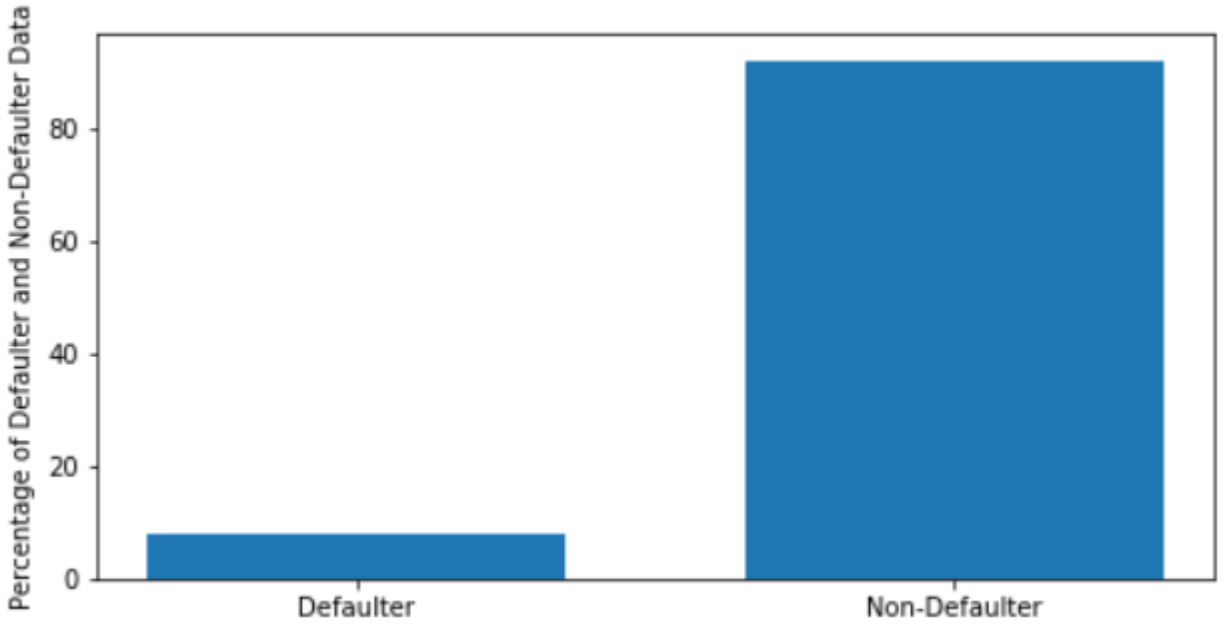
Bank loan (EDA) Case Study

Univariate Analysis for Categorical Variables from application_data.csv

Data Imbalance Ratio between Target 0 and Target 1

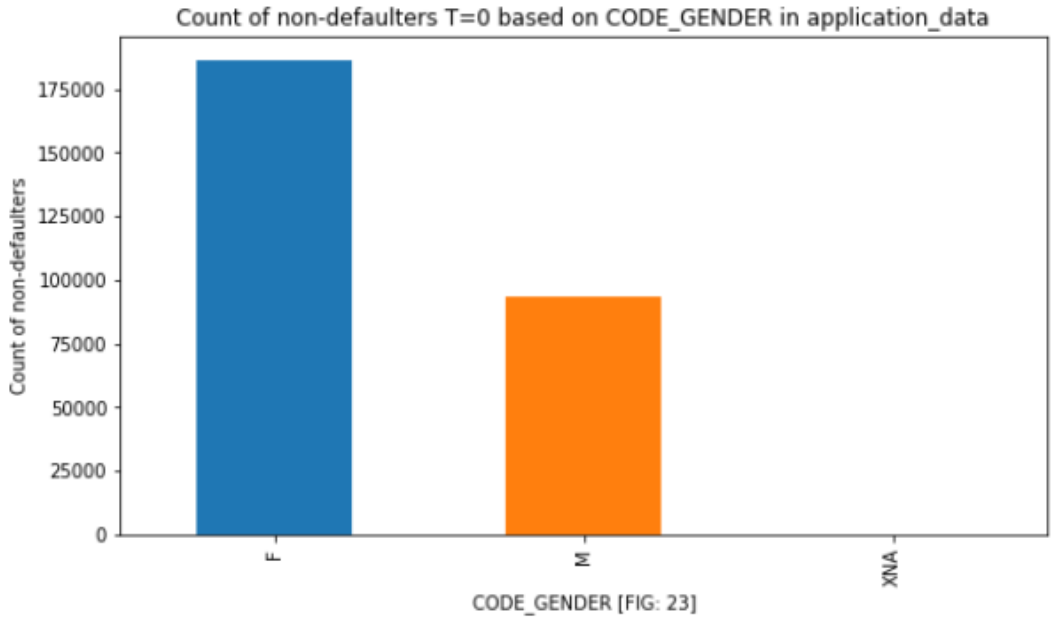


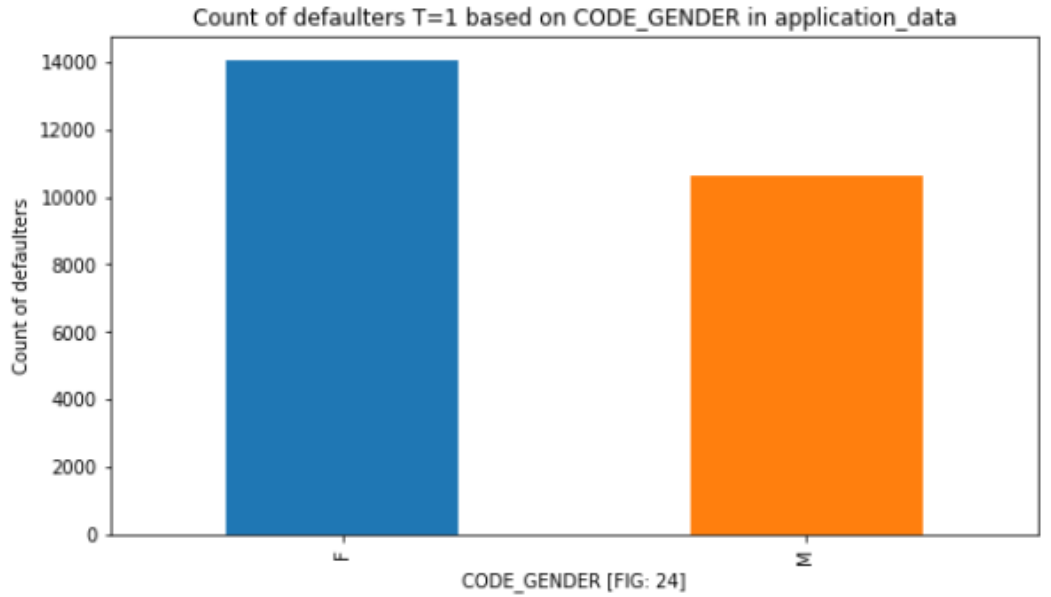
From the above graph we can very clearly see that there is an imbalance between the defaulters data and the non-defaulters. As mentioned earlier, imbalance of data will tend to create a biased model.

Straight forward inference from the above plot is that most of the loan applicants (around 91.93%) are likely to pay their loan/EMI on time (Target 0) and only around 8.07% of the applicants are expected to default on their loans.

The imbalance ratio is hence: 91.93:8.07

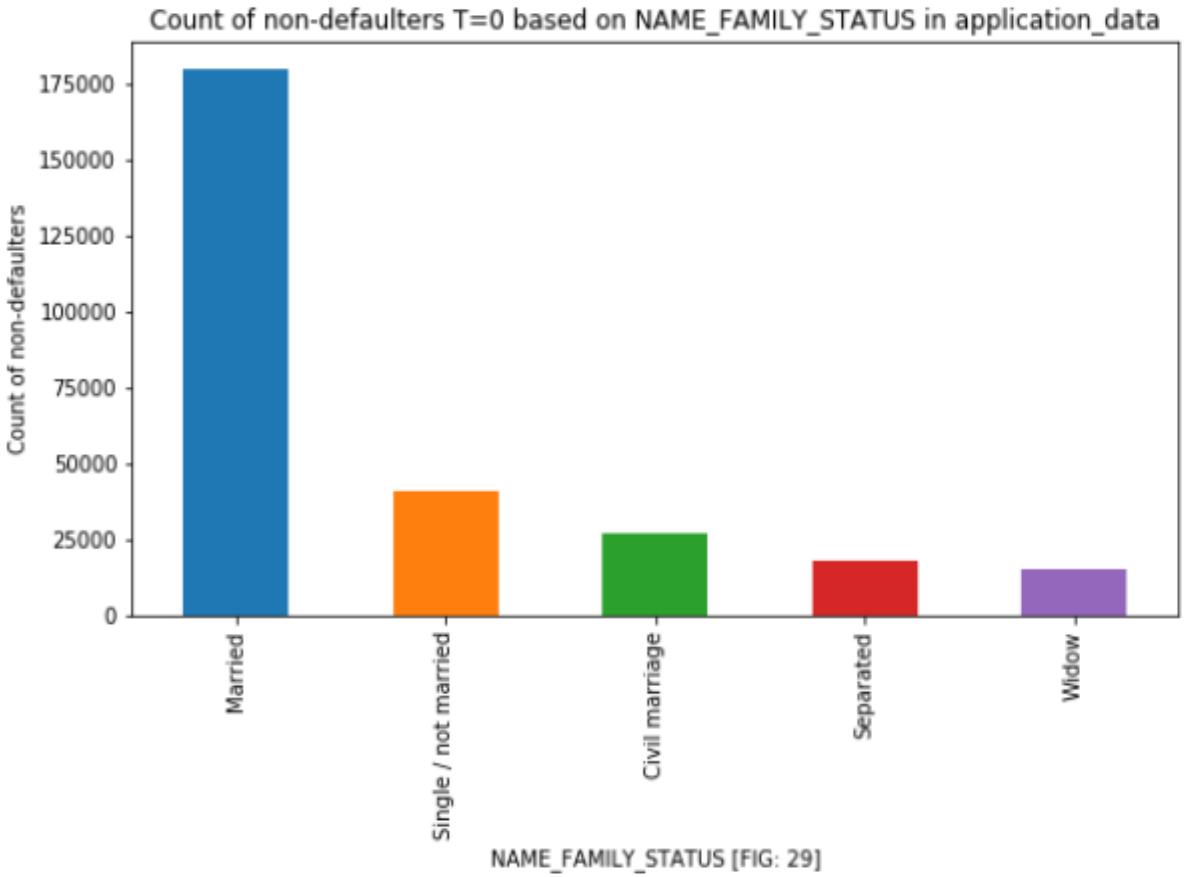
Target relationship with Applicant Gender

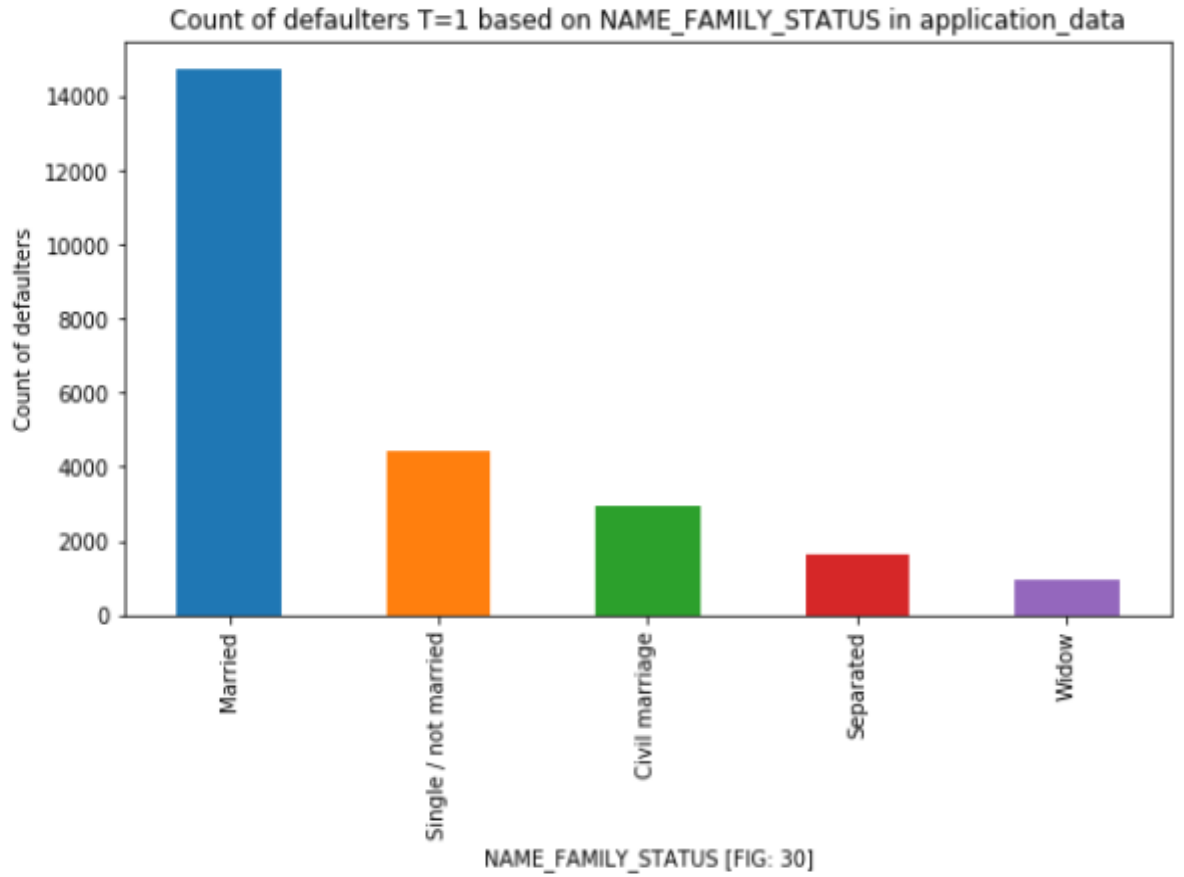




One of the five categorical variables chosen for univariate analysis, which is a fairly important factor, is gender of applicant. The above plot clearly indicates that in either case (i.e. Target 0 and Target 1), the number of Female applicants is higher than males.

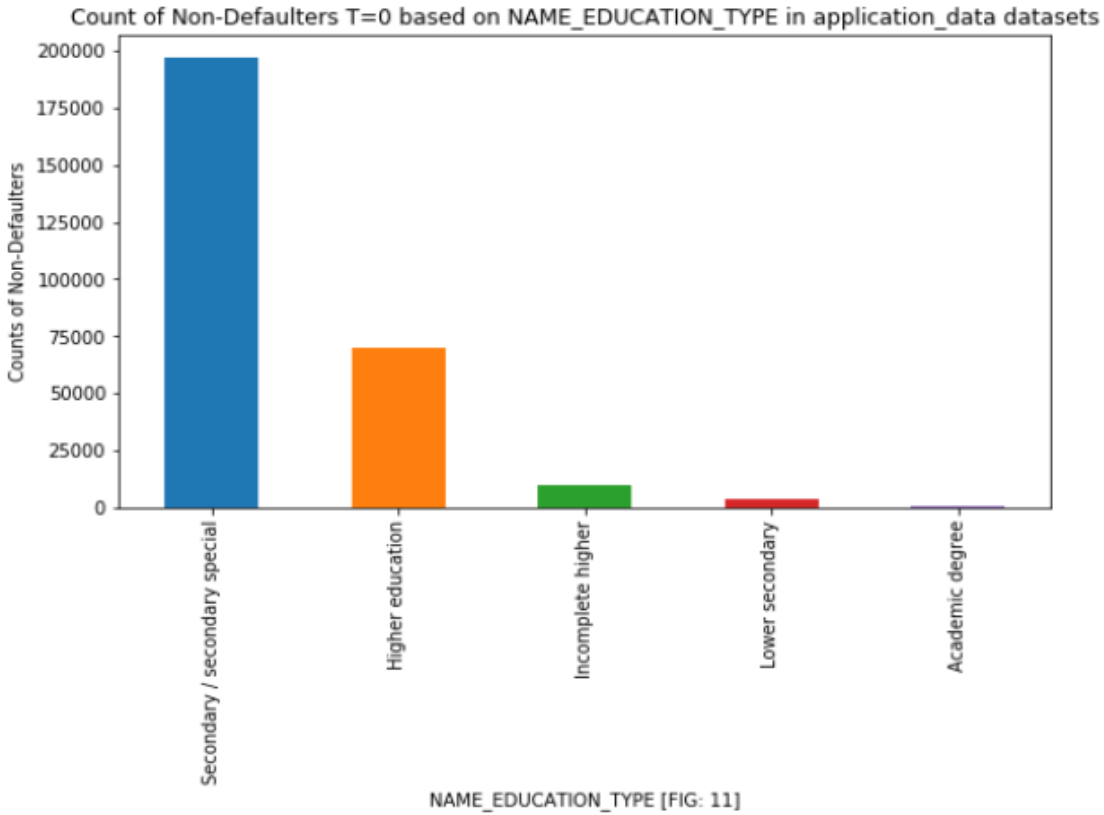
Target relationship with Family Status

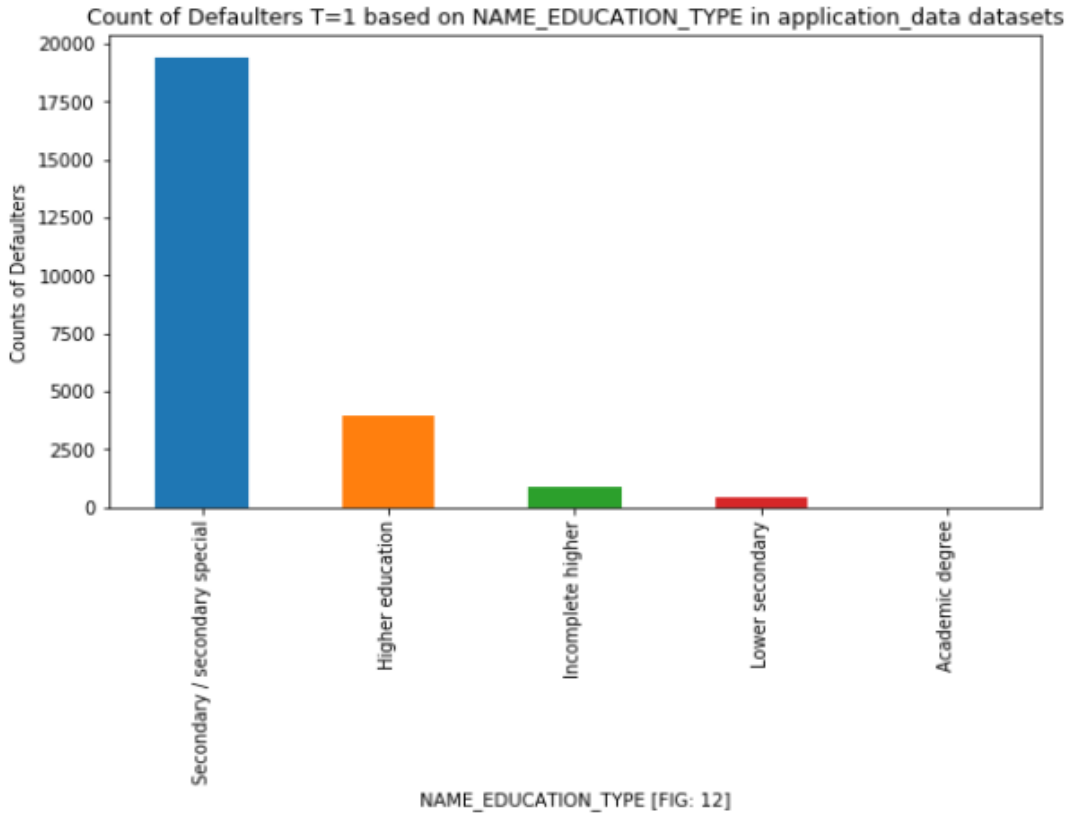




Target 0/1 plot against Family status shows that married people make up the highest number of applicants. This makes sense as married people have a lot of financial responsibilities and hence need monetary help.

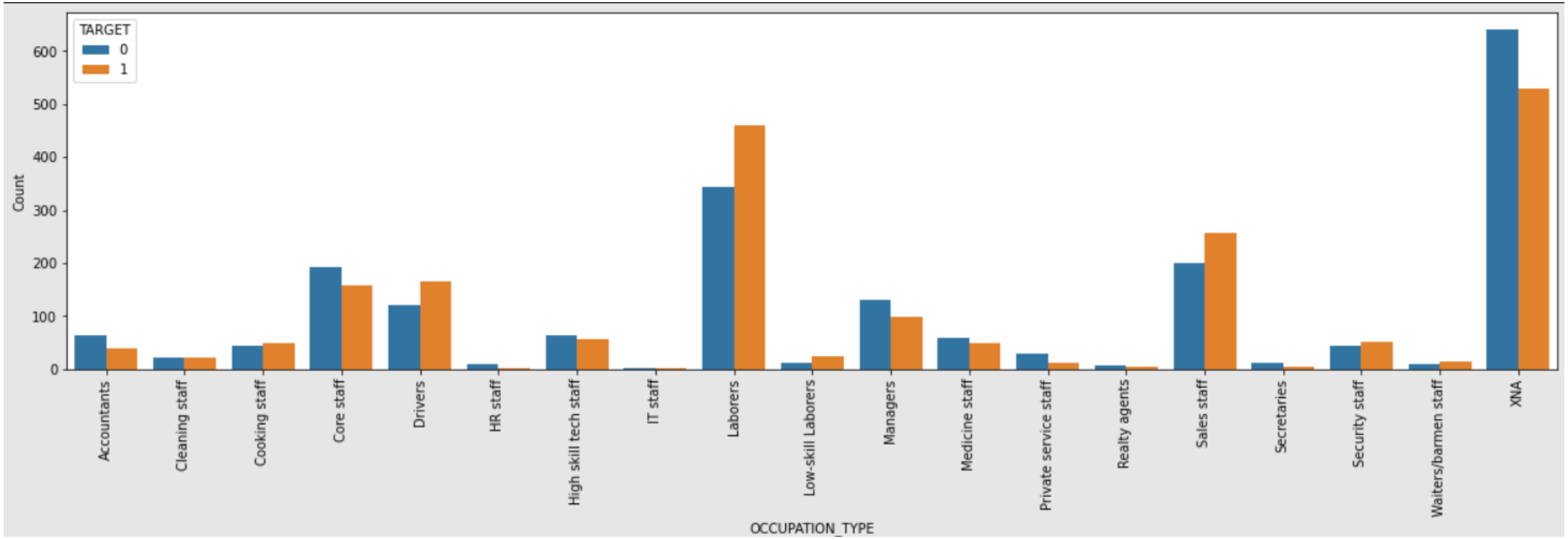
Target relationship with Education Type





The above plot clearly shows that clients with Secondary/Secondary special education submit the most loan applications.

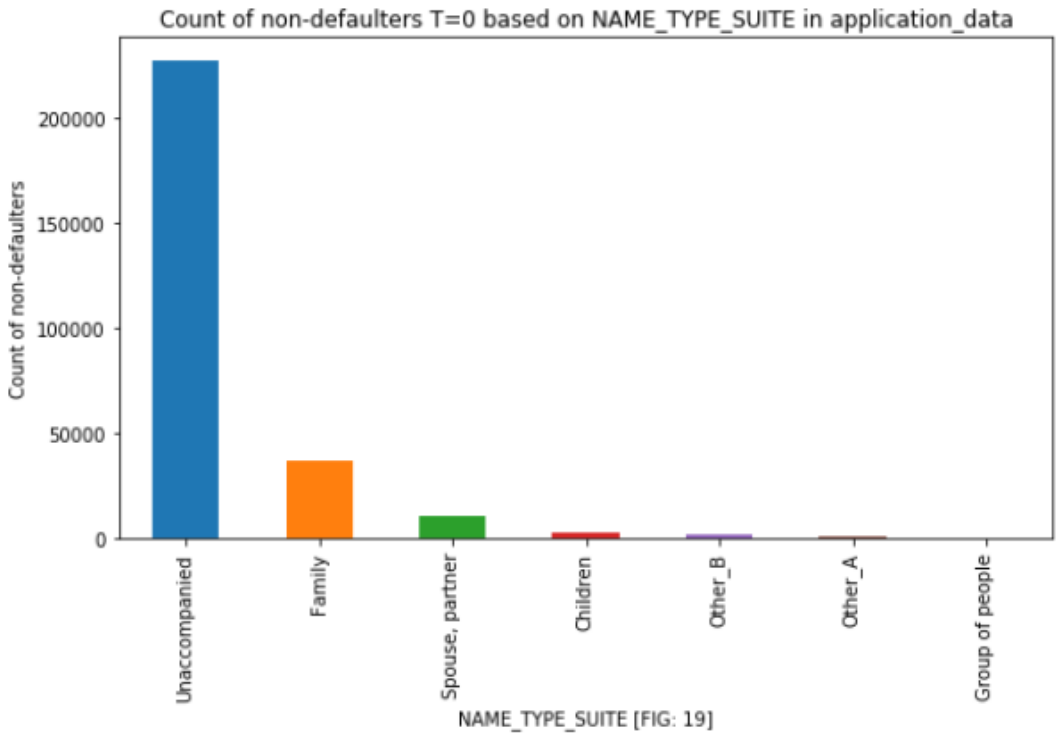
Target relationship with Occupation Type

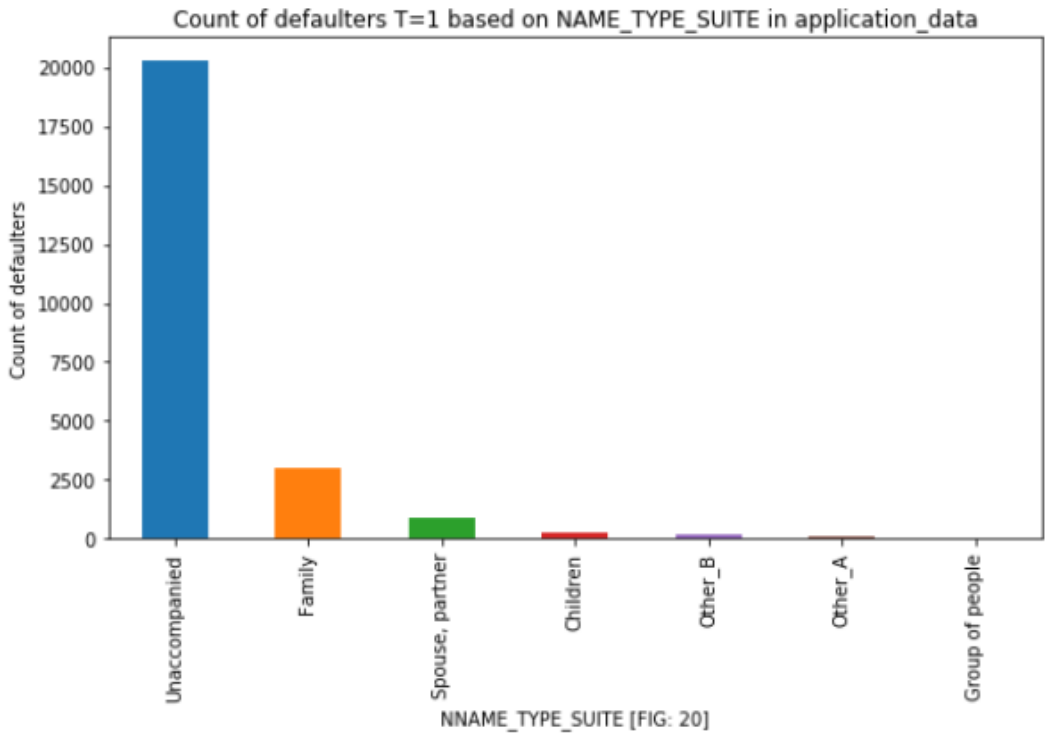


Among all applicant occupation types, most of the applications are received from the laborer class; Laborers do tend to need financial assistance to set themselves up, make their ends meet and, to look after their families and also be able to send some moneyback home.

Note- As Occupation Type is important aspect in analysis, so keep in mind we have imputed the 31 percent missing valus with XNA.

Target relationship with Suite Type

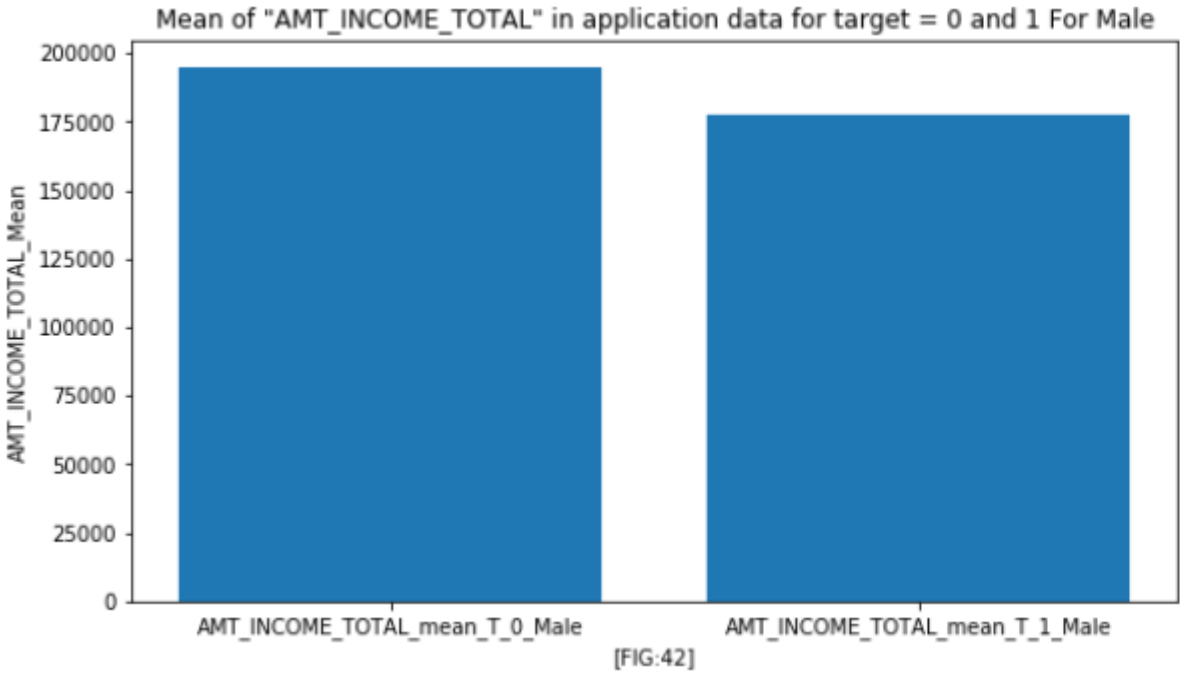




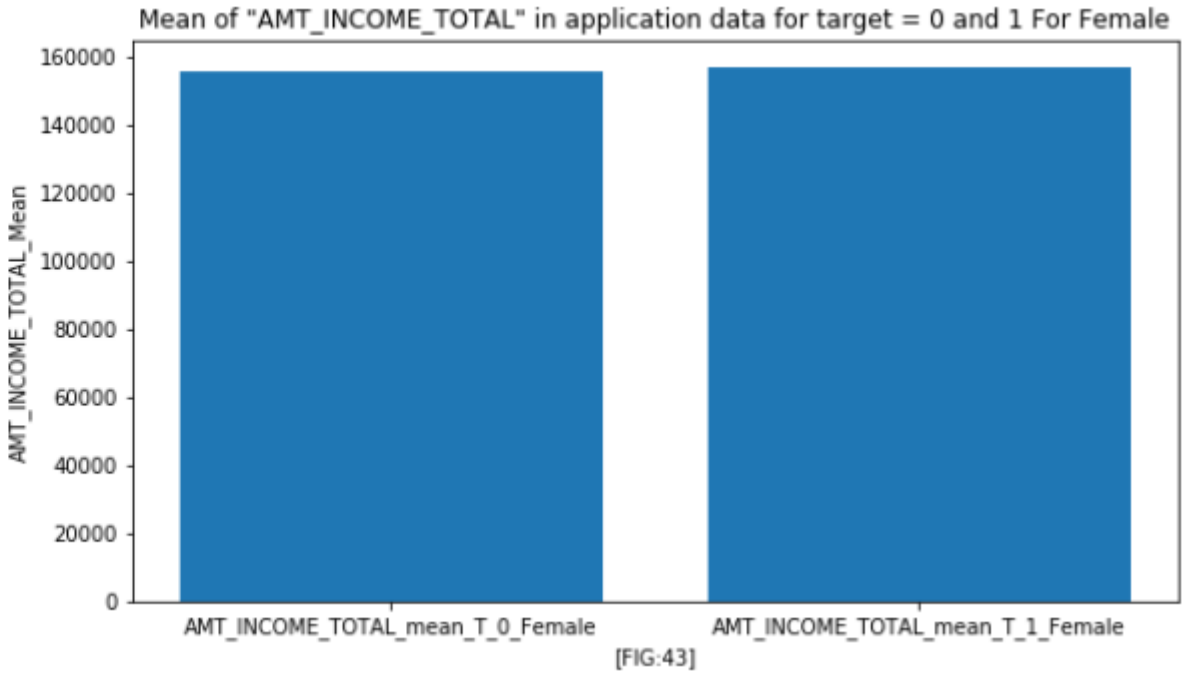
Suite type column indicates who accompanied the applicant when they approached the bank for loan. What the graph shows is exactly what happens; most people go on their own to apply/submit for loan.

Univariate Analysis for Numerical Variables from application_data.csv

Segmented Univariate Analysis on Numerical variable(AMT_INCOME_TOTAL) for Target 1 and 0 on the basis of Gender



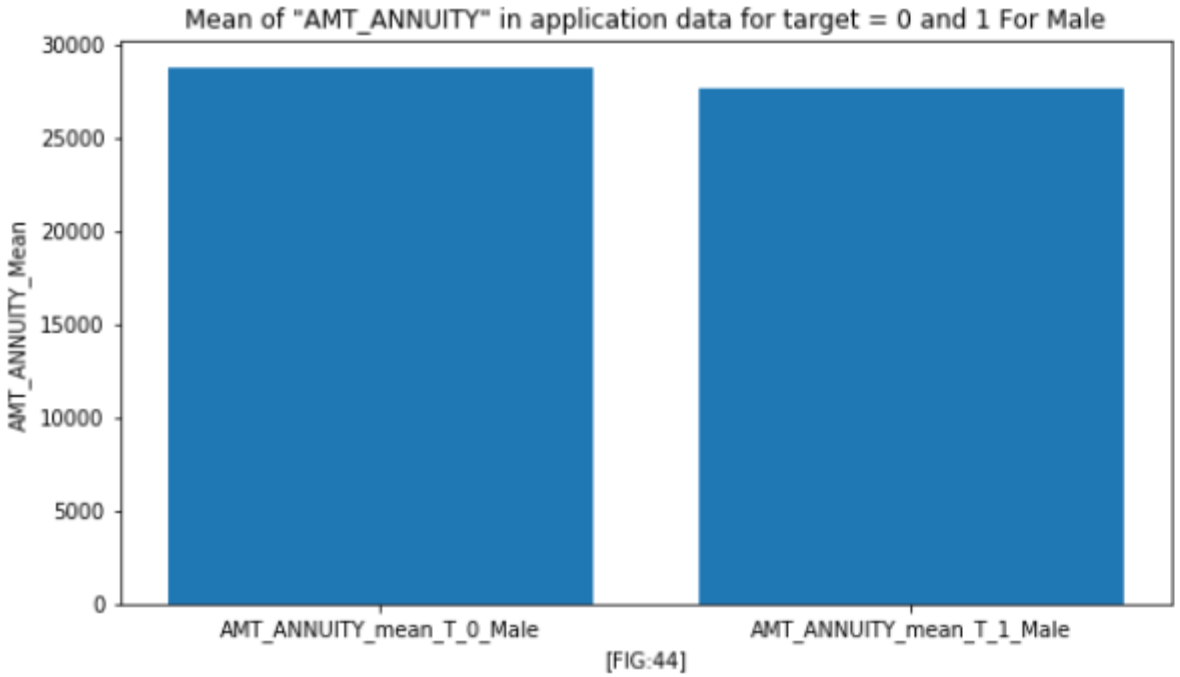
[FIG:42]



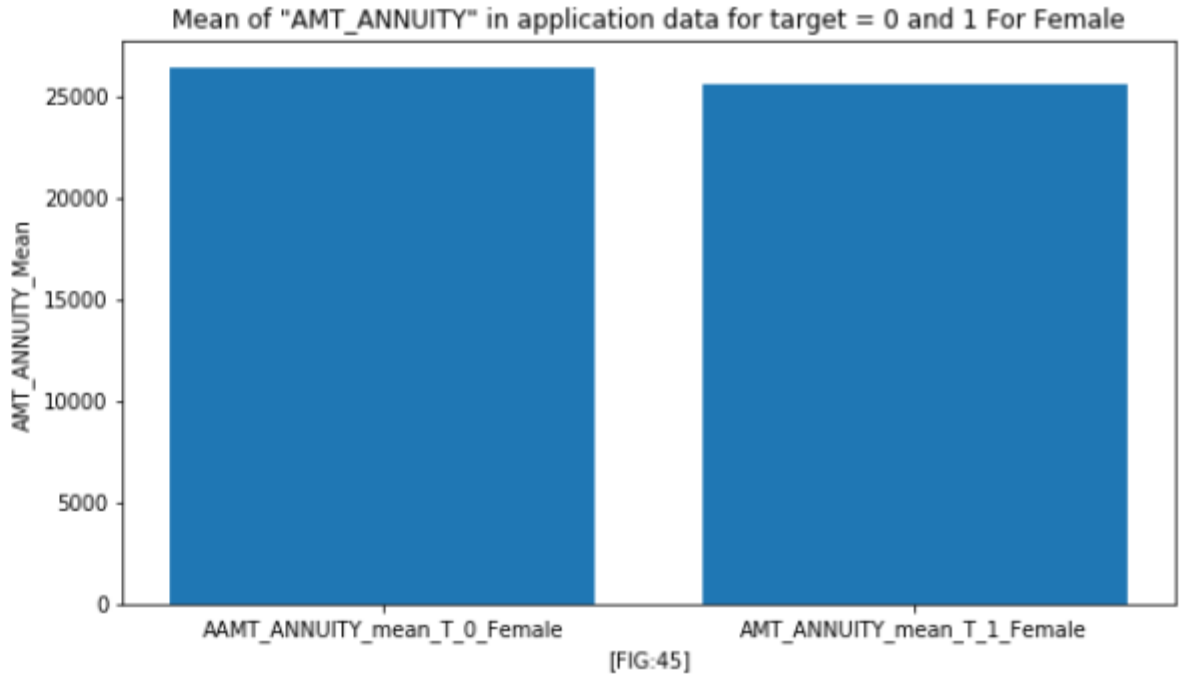
[FIG:43]

We have to extra careful in providing loan for applications holding less income since they may most likely to become defaulters.

Segmented Univariate Analysis on Numerical variable(AMT_ANNUITY) for Target 1 and 0 on the basis of Gender



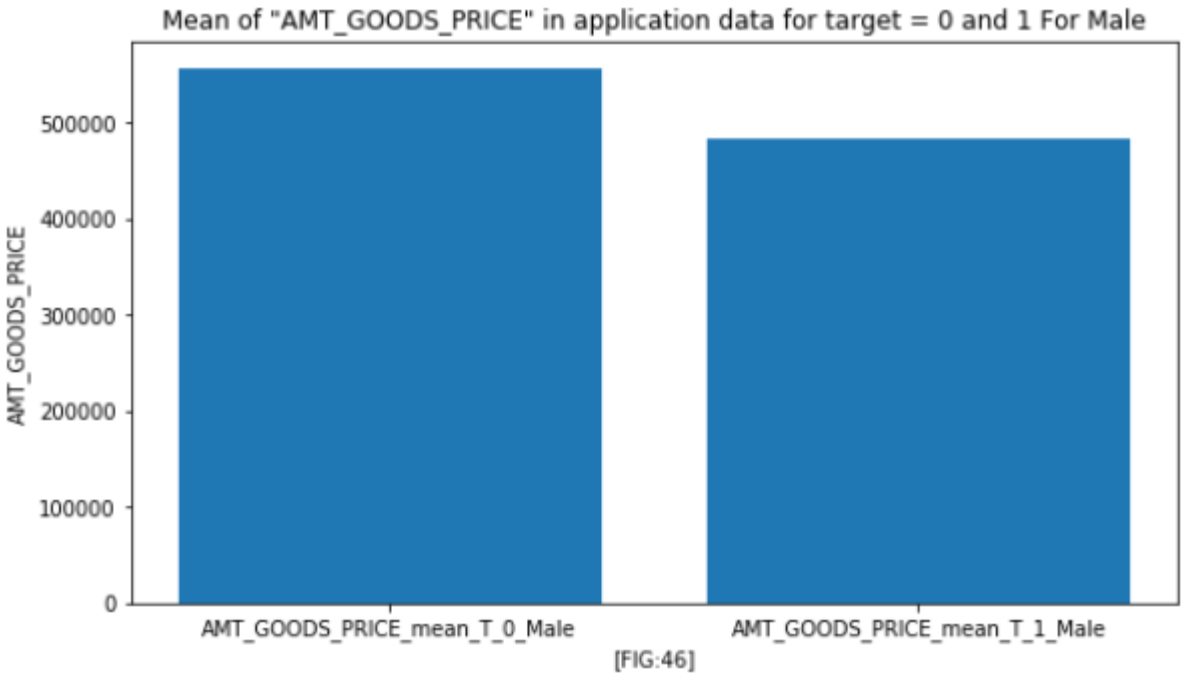
[FIG:44]



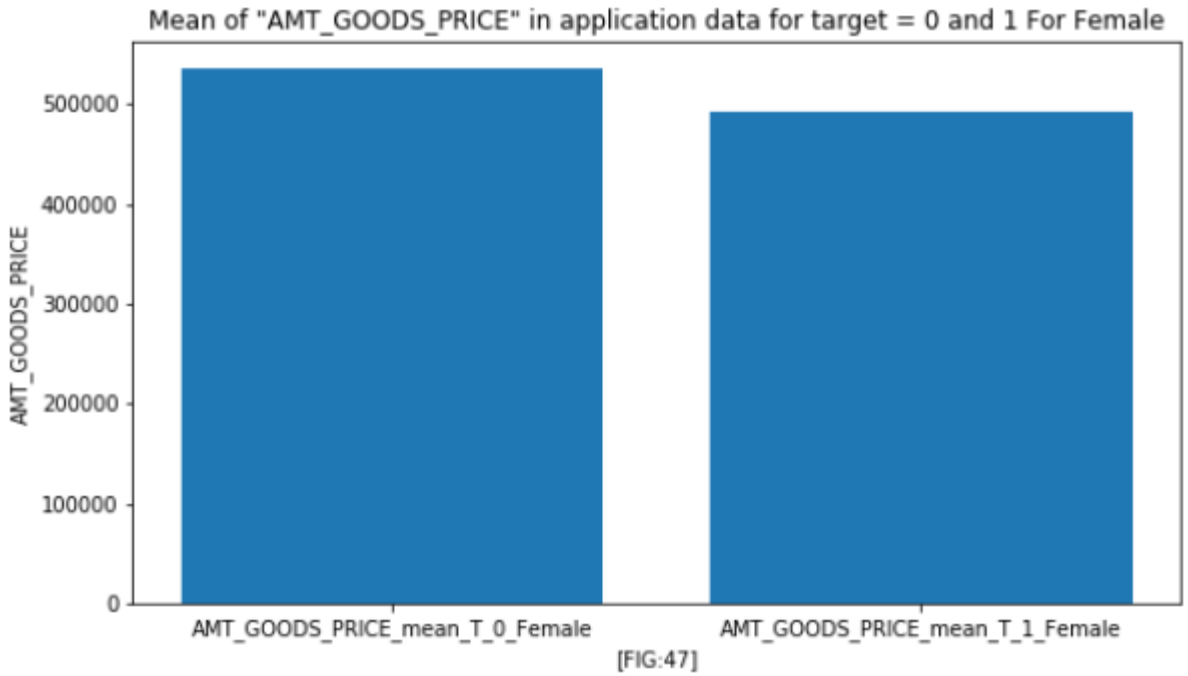
[FIG:45]

We have to extra careful in providing loan for applications holding less AMT_ANNUITY since they may most likely to become defaulters

Segmented Univariate Analysis on Numerical variable(AMT_GOODS_PRICE) for Target 1 and 0 on the basis of Gender



[FIG:46]

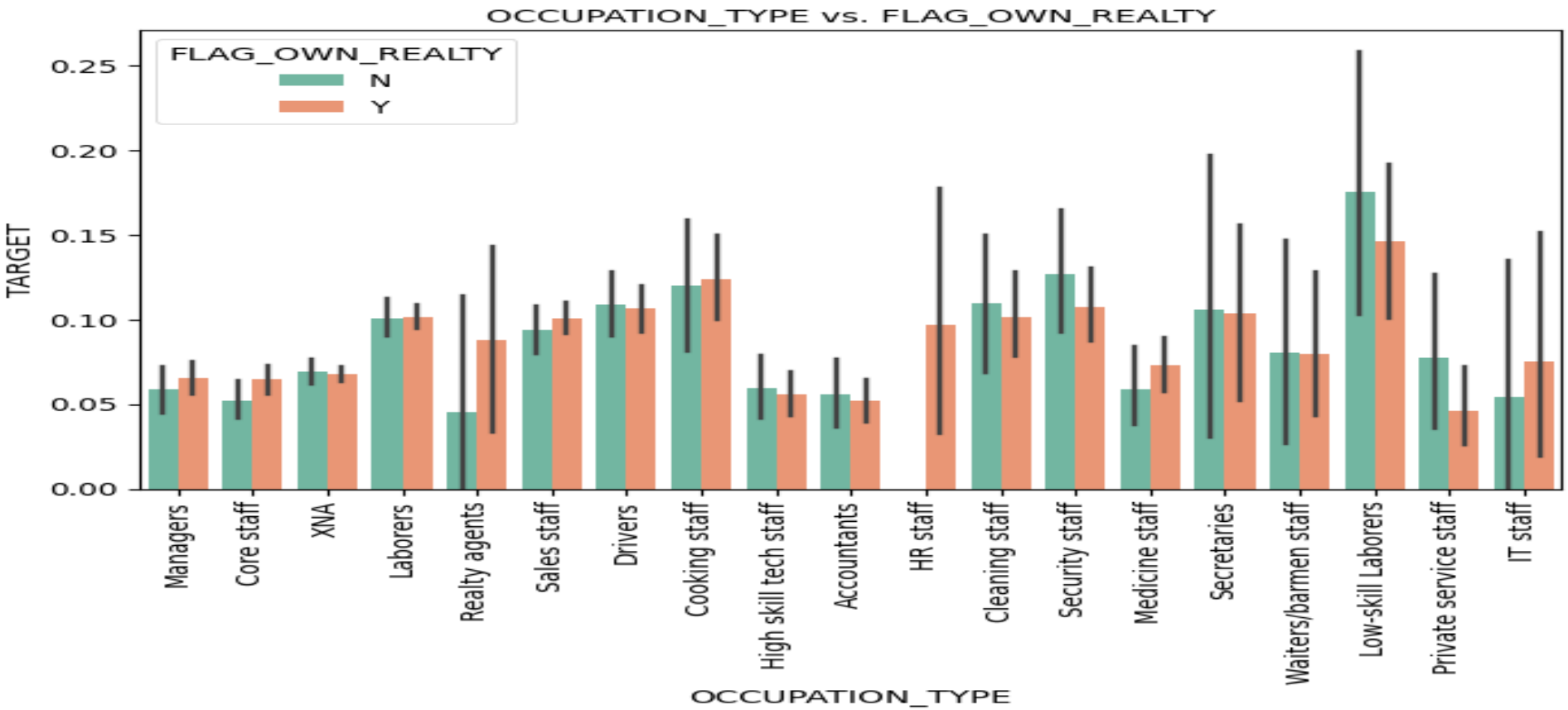


[FIG:47]

We have to extra careful in providing loan for applications holding less AMT_GOODS_PRICE since they may most likely to become defaulters

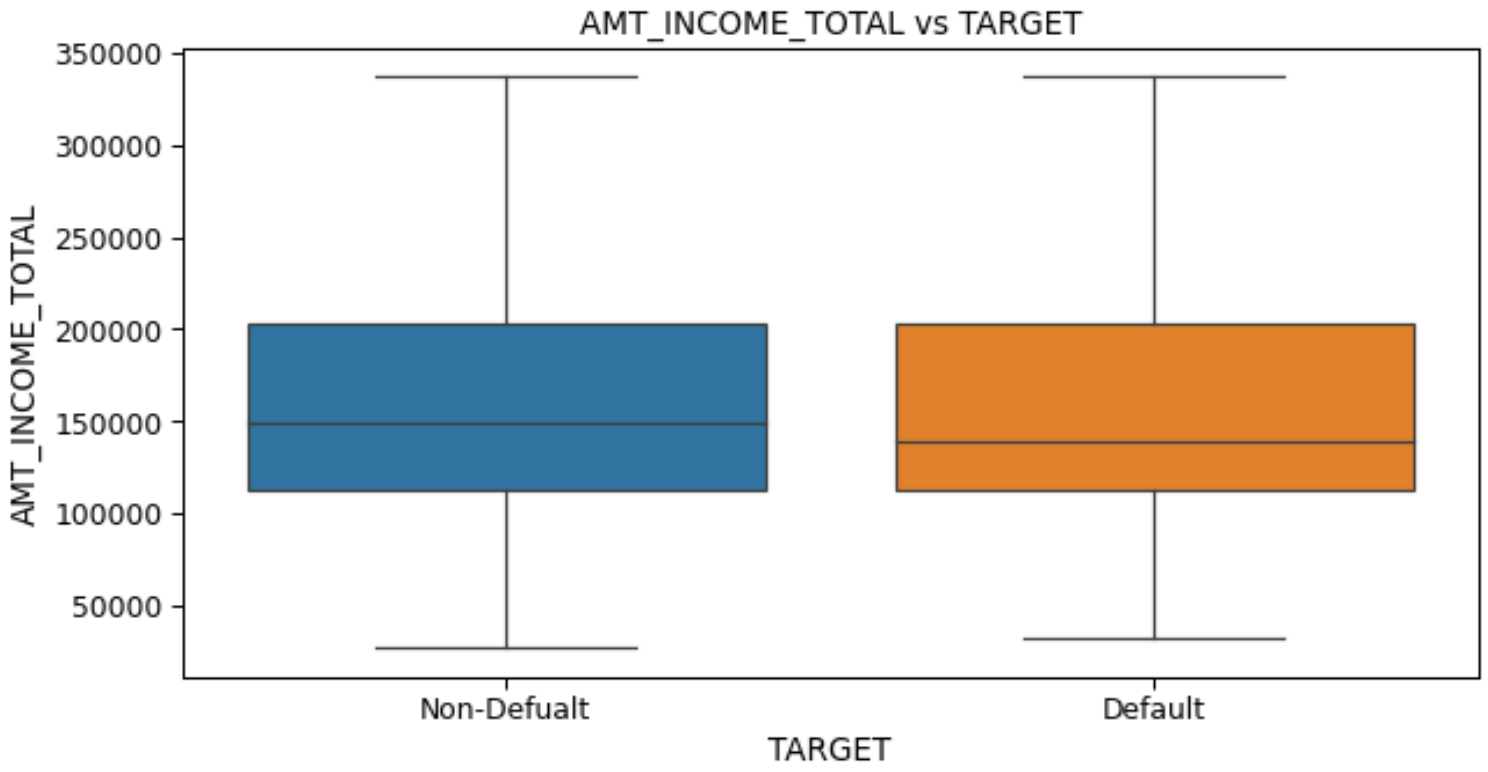
Bivariate Analysis for Numerical Variable

Target versus Occupation Type w.r.t. Own Property



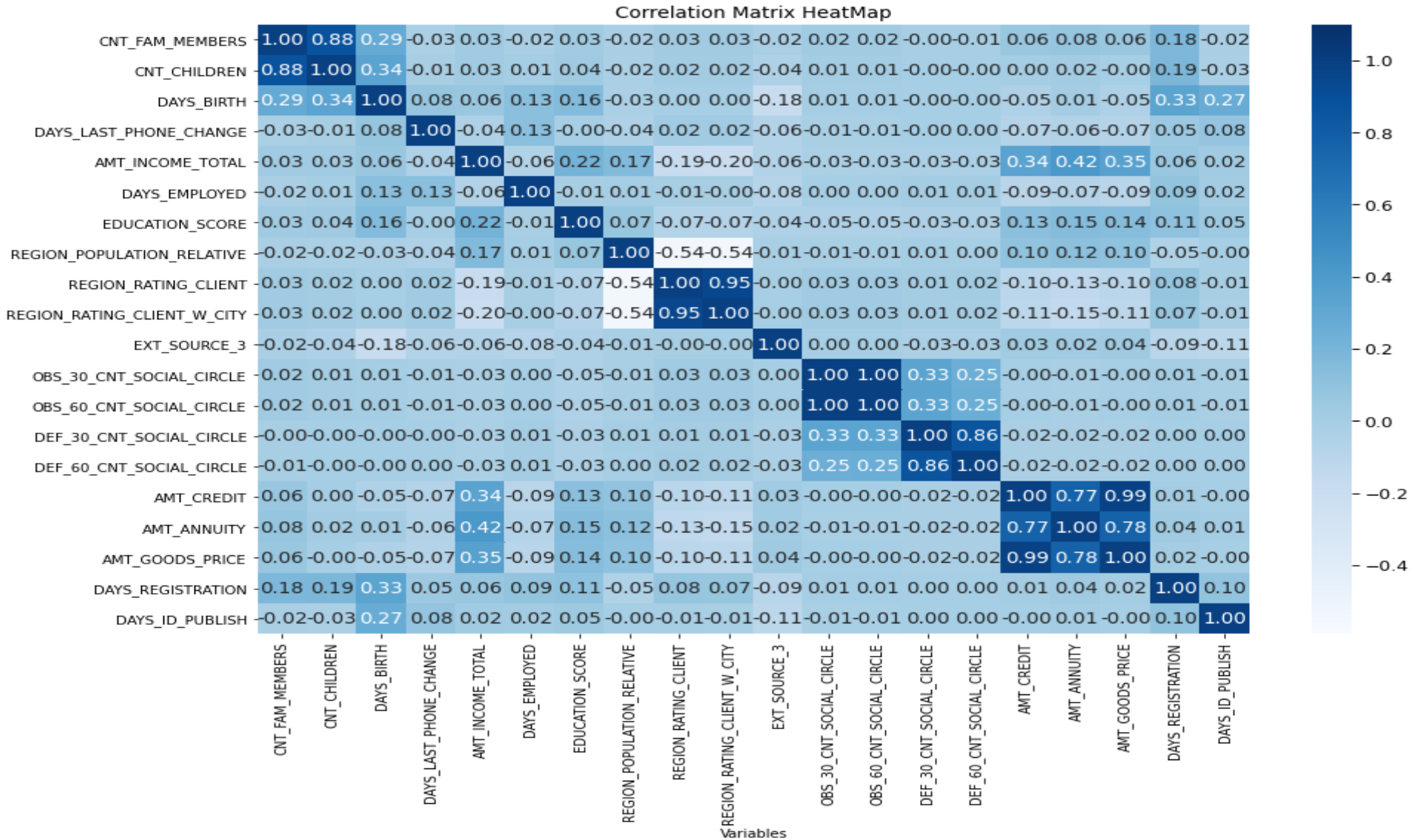
Bivariate plot of target against type of client occupation tells us that Low-skilled labourers who do not have own property tend to skip paying their loan on time

Target versus Amount Income Total



Plot shows a near identical distribution of income for both defaulters and non-defaulters

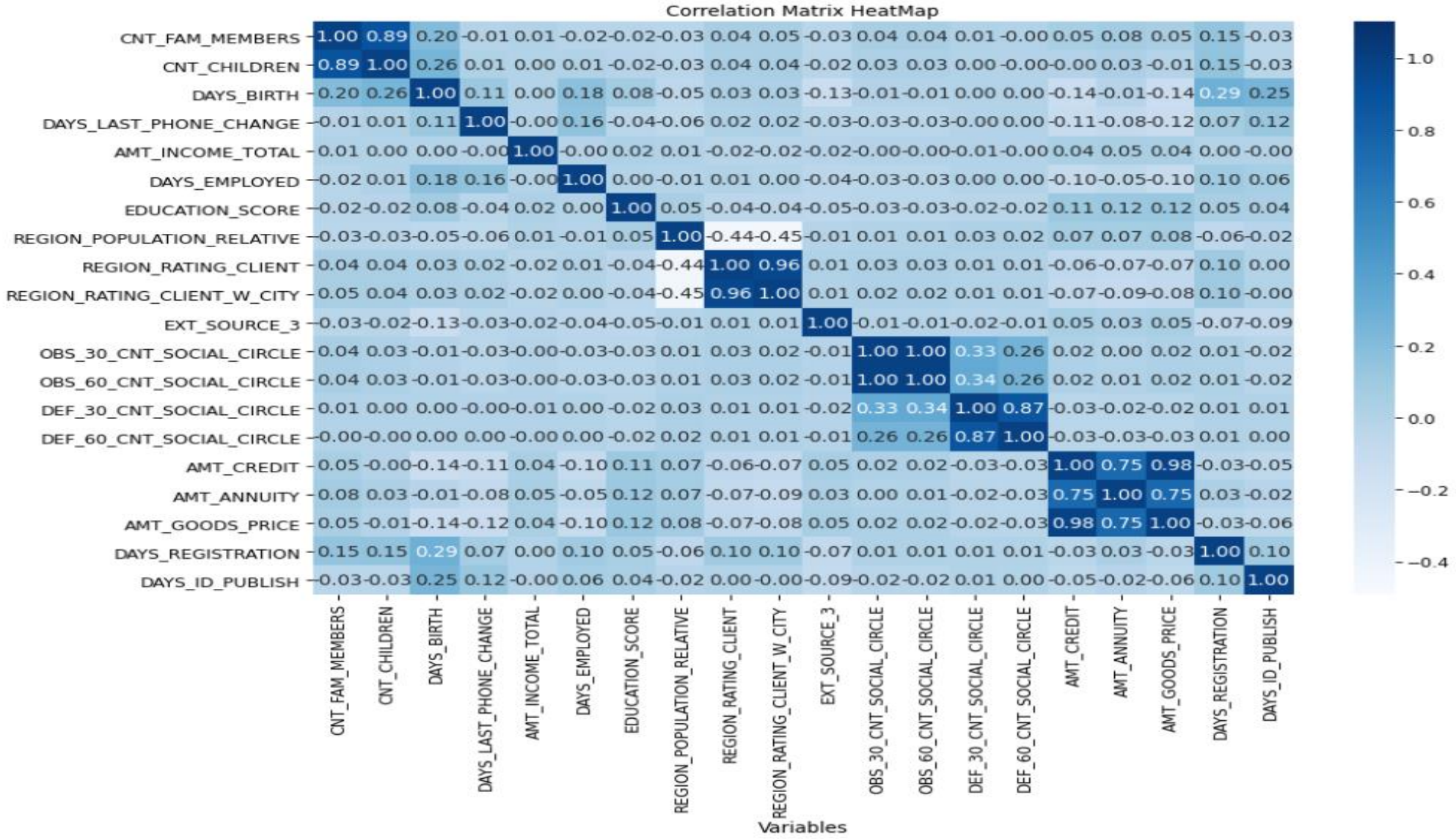
correlation between the numerical variables For Target=0



Top 10 correlation between the numerical variables For Target=0

V1	V2	Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998513
AMT_GOODS_PRICE	AMT_CREDIT	0.987260
EGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.949905
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878681
IVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861303
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859458
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.830488
AMT_GOODS_PRICE	AMT_ANNUITY	0.775838
AMT_ANNUITY	AMT_CREDIT	0.770379
FLAG_EMP_PHONE	DAYS_BIRTH	0.622090

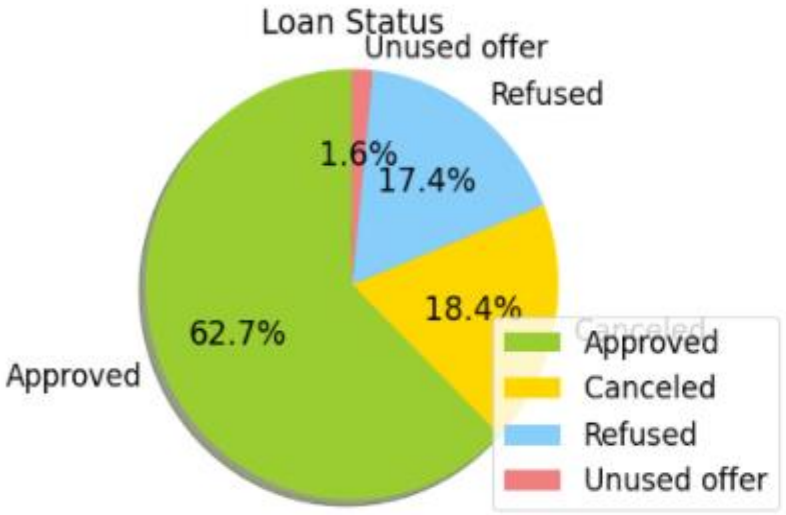
correlation between the numerical variables For Target=1

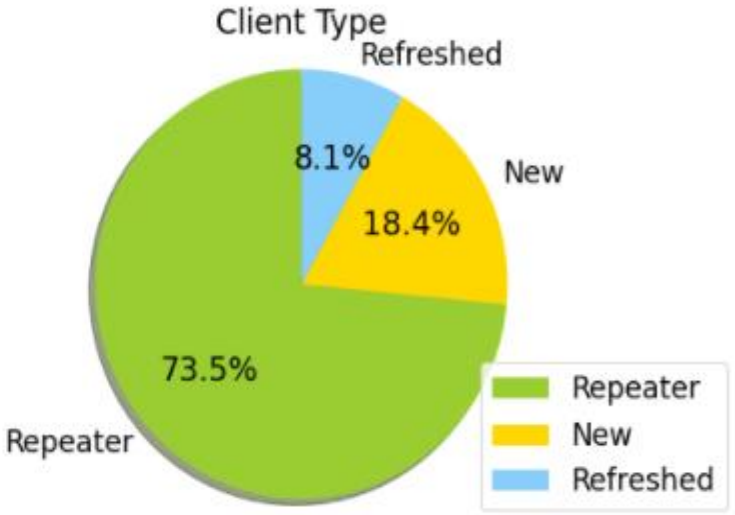


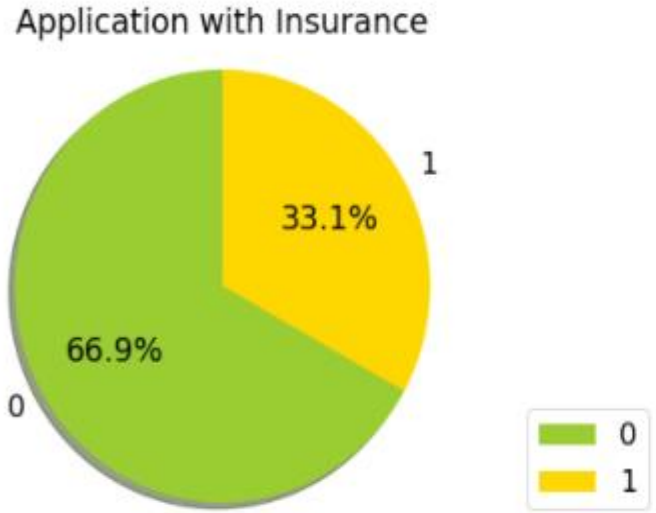
Top 10 correlation between the numerical variables For Target=1

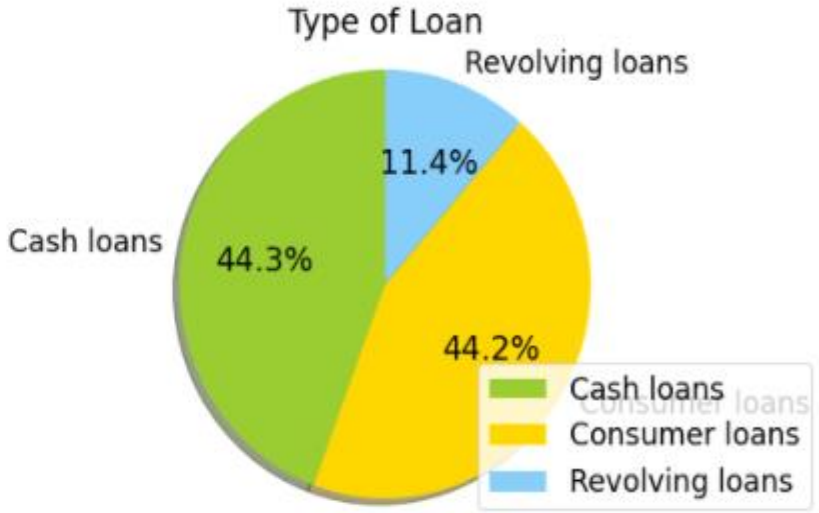
V1	V2	Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998286
AMT_GOODS_PRICE	AMT_CREDIT	0.983065
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956477
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885556
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.869761
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847260
LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778110
AMT_GOODS_PRICE	AMT_ANNUITY	0.752206
AMT_ANNUITY	AMT_CREDIT	0.751400
FLAG_DOCUMENT_6	FLAG_EMP_PHONE	0.617071

Analysis of data from previous application records









Below are the Statistics and analyses of records from previous applications

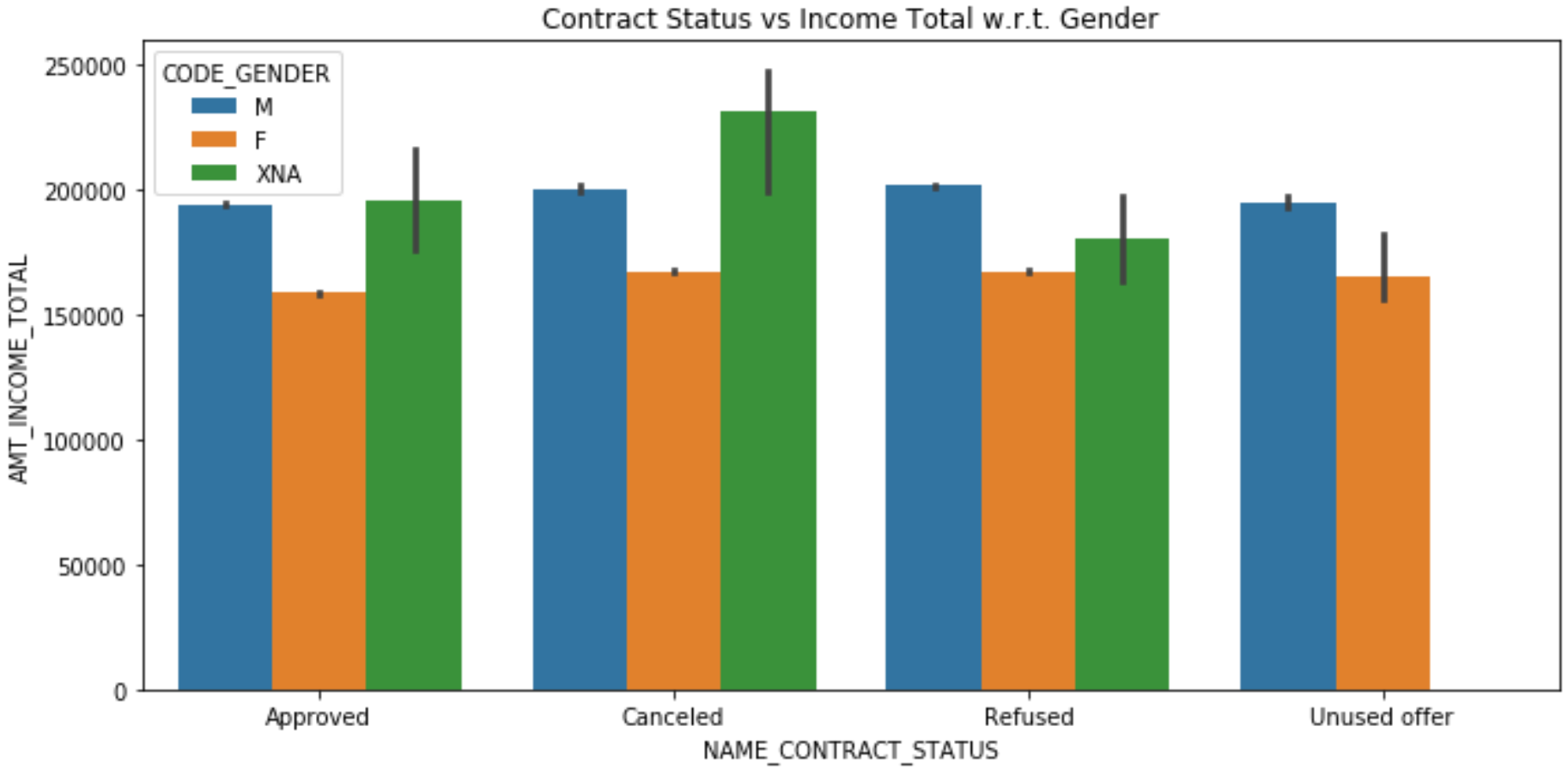
Comparison:

- Approved: 62.7 %
- Cancelled: 18.9 %
- Refused: 17.4 %
- Unused offer: 1.58 %
- 73.5 % were repeater clients who applied for loan previously.
- Approx. 67% clints requested insurance during the previous application
- Cash loans - 44.76 %
- Consumer loans - 43.66 %
- Revolving loan - 11.57 %

Bivariate Analysis on merged data set (application data and previous data)

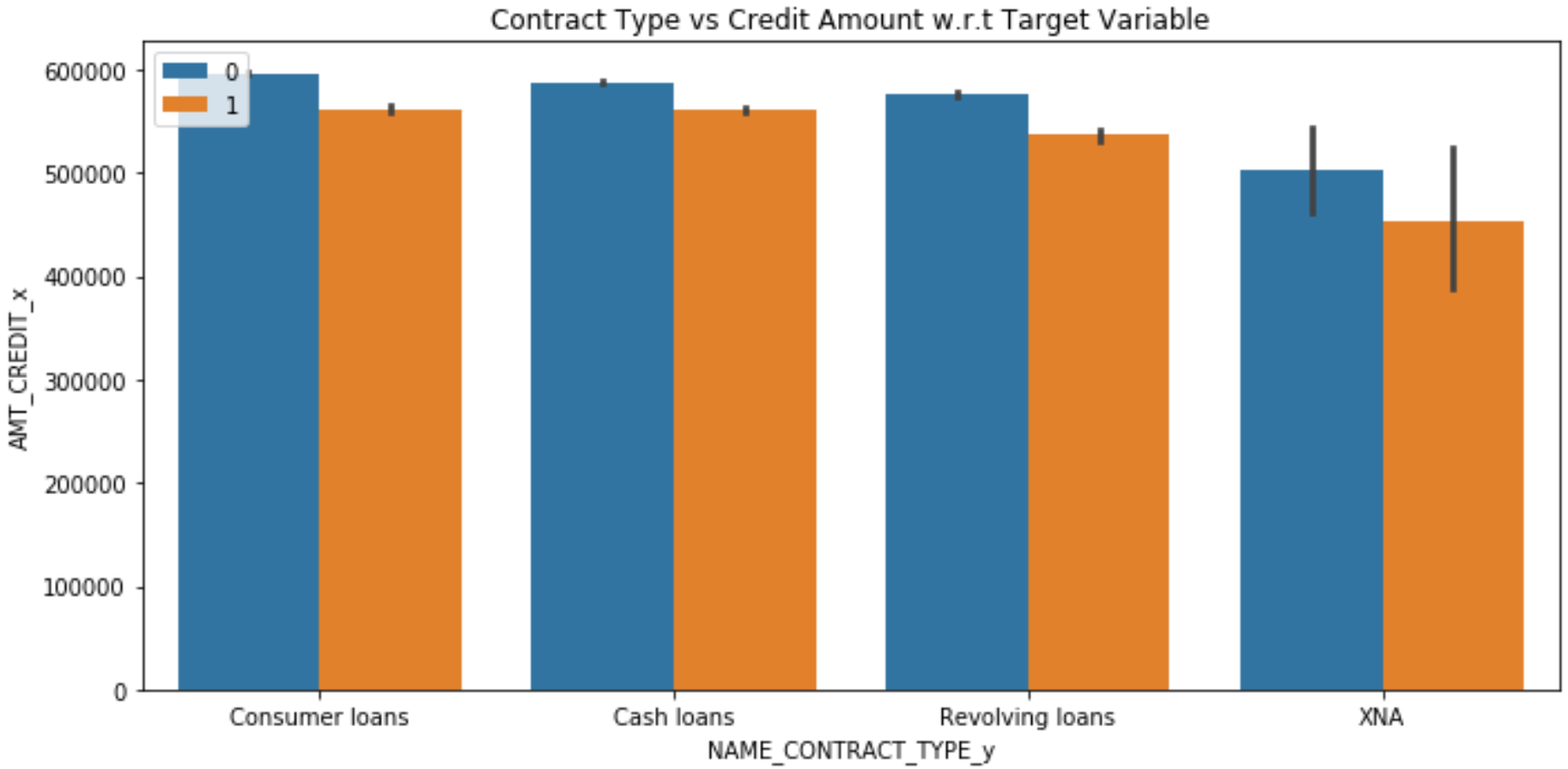
- *For every SK_ID_CURR in application data, there is a corresponding SK_ID_CURR in previous application file. Each SK_ID_CURR is associated with 1 or more SK_ID_PREV which represent a history of loan applications by the customer.*
- *Each SK_ID_PREV (previous loan history) has a TARGET variable (renamed to TARGET_x in merged file) that represent the banks decision to approve/refuse the loan.*
- *Some columns (with -ve data values) have several +ve "365243" (1000 years) entries where data is not available. If these columns are needed for analysis, we will impute them with "median" of those columns (not the mean, since mean is highly skewed).*
- *We observed that in pa (previous application data) there is consumer loan segment which is not in application data.*

Contract Status versus Income Total w.r.t. Gender



Visualization shows that all four contract statuses for male applicants tend to round off at the 200000 income mark

Contract Type versus credit amount w.r.t. Target Variable



People taking consumer loans of close to 600000 are likely to both pay on time and default compared to other types

Conclusion

- In 6.1.3 It is an evident that if the categorical variables [PRODUCT_COMBINATION, NAME_YIELD_GROUP, CHANNEL_TYPE, NAME_PORTFOLIO] contains the categories [cash X-sell: low, XNA, Credit and cash offices, Cash] then the applicant is more likely to become a defaulter and the loan providing company should be more careful in providing loan for these applicants.
- On the other hand, if the above mentioned categorical variables contains the categories [POS household with interest, middle, country-wide, POS] then the is not likely to become a defaulter and the loan providing company should not be cancelling the application in providing the loan for these applicants.
- For male applications income (AMT_INCOME_TOTAL) plays a role in becoming a default or not. Higher the income it is likely to become defaulter. So the loan lending company should be careful in providing loans for male applicants who has less income.
- In considering the numerical variable AMT_ANNUITY, in both male and female, if the value of AMT_ANNUITY is less then the applicants are most likely to become defaulters, so the lending company should be more careful in lending loans to these applicants.
- In considering the numerical variable AMT_GOODS_PRICE, in both male and female, if the value of AMT_GOODS_PRICE is less then the applicants are most likely to become defaulters, so the lending company should be more careful in lending loans to these applicants.
- In considering the numerical variable AMT_APPLICATION, in general, if the value of AMT_APPLICATION is high then the applicants are most likely to become defaulters, so the lending company should be more careful in lending loans to these applicants.