CREDIT EDA Assignment

NIHARIKA DUSA

INTRODUCTION

- 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 for 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.

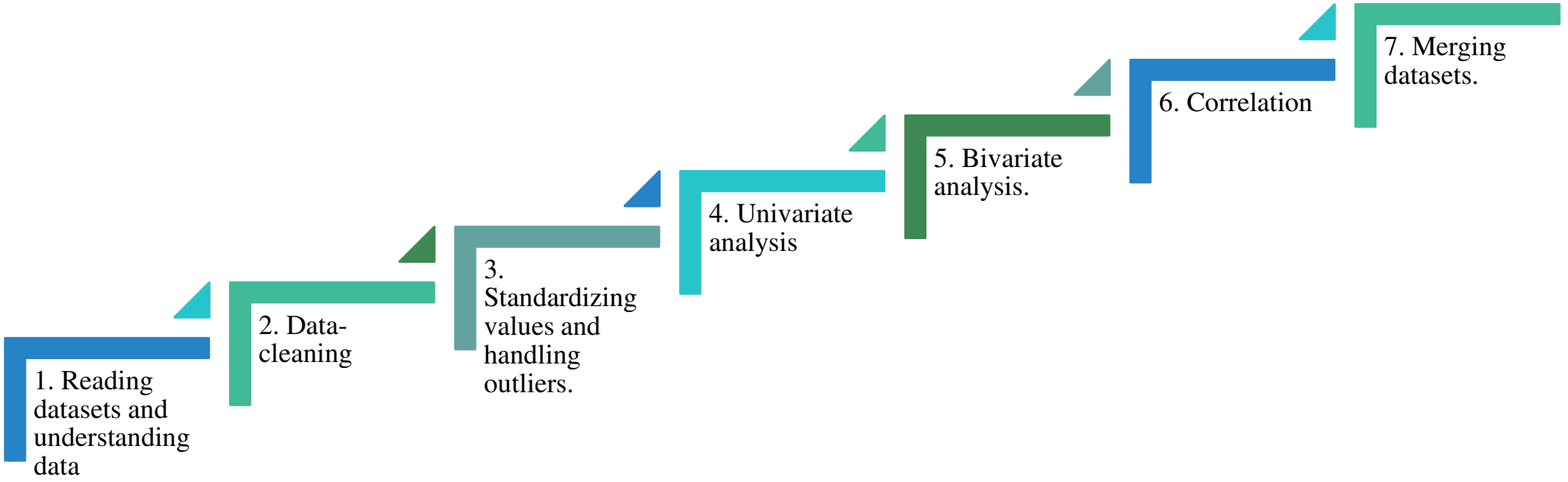
Problem statement

- The loan-providing companies find it hard to give loans to people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming defaulters.
- Suppose you work for a consumer finance company that specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

Datasets used:

- application.csv: This is current dataset
- Previous_application.csv: This is previous data about loans.

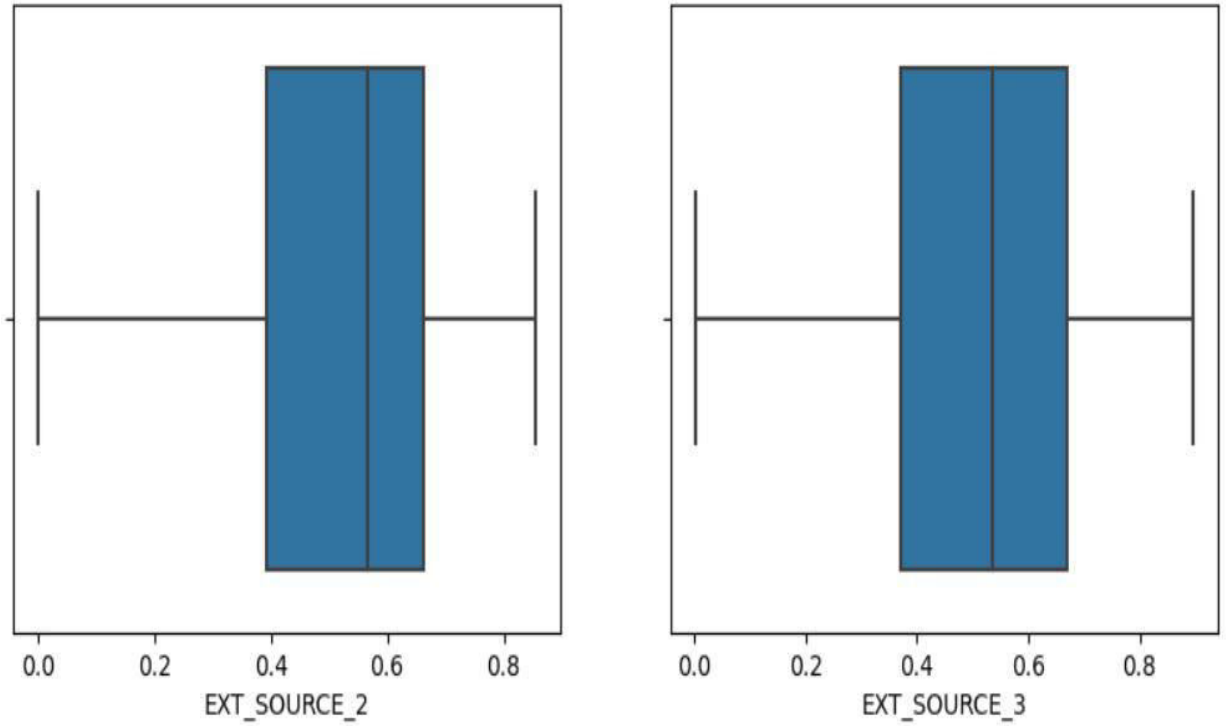
Steps involved



Data-Cleaning

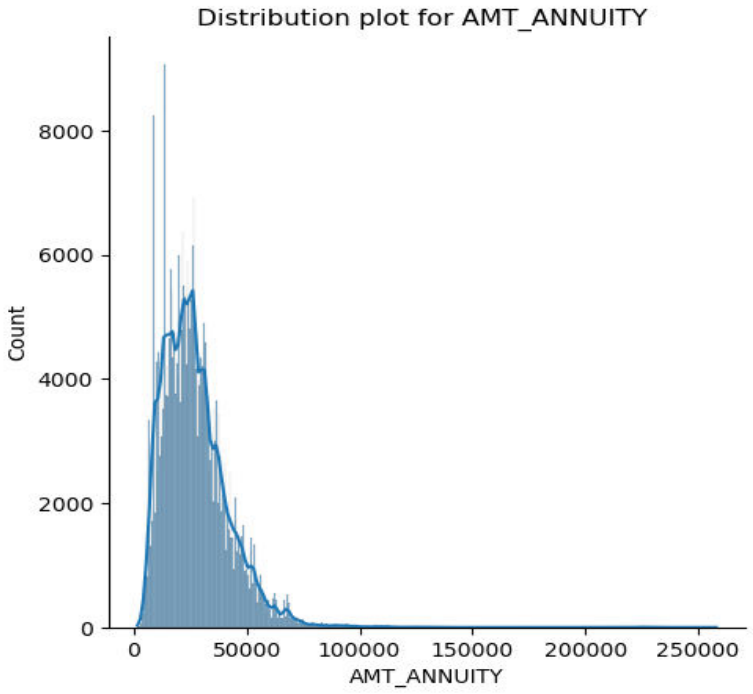
- I have used box plots and distribution plots to check for outliers and skewness and then imputed missing values with mean, median, or mode accordingly.

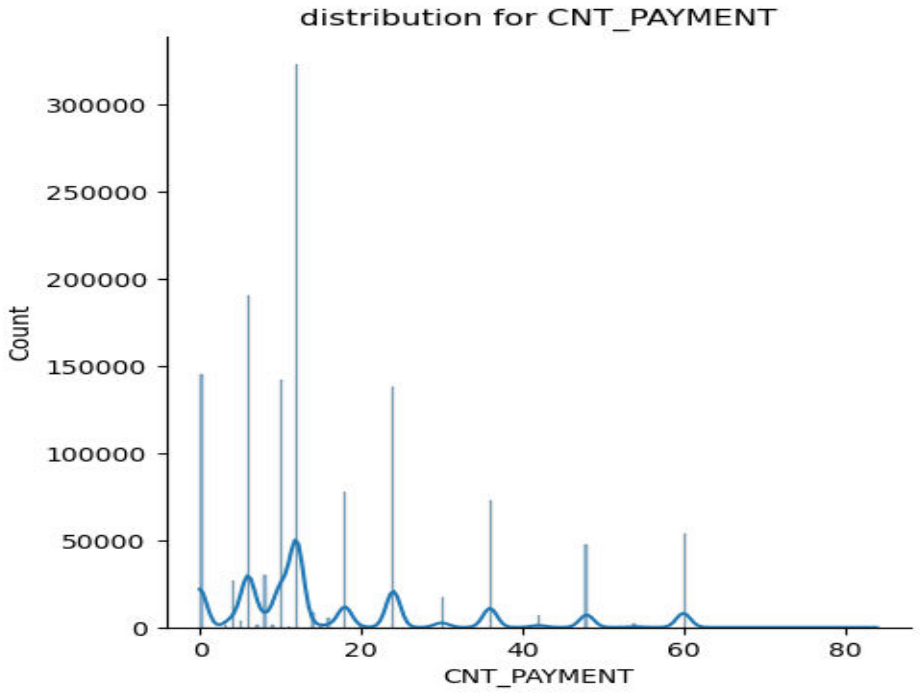
Box-plots for EXT_SOURCE_2 and EXT_SOURCE_3



Data-cleaning

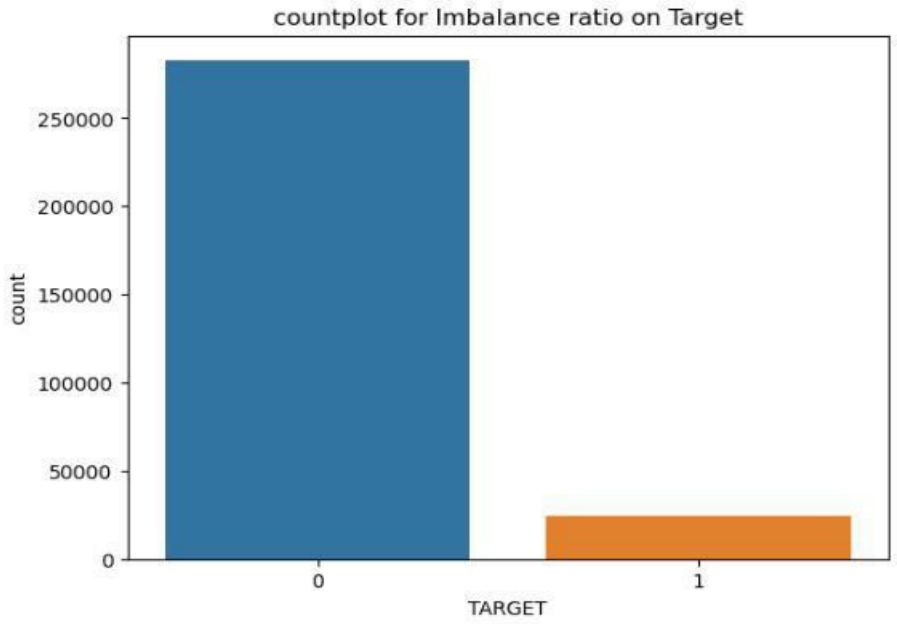
- For the AMT_ANNUTIY, CNT_PAYMENT column I have used a distribution plot to calculate the skew and imputed missing values with the median. Similarly done for all columns.

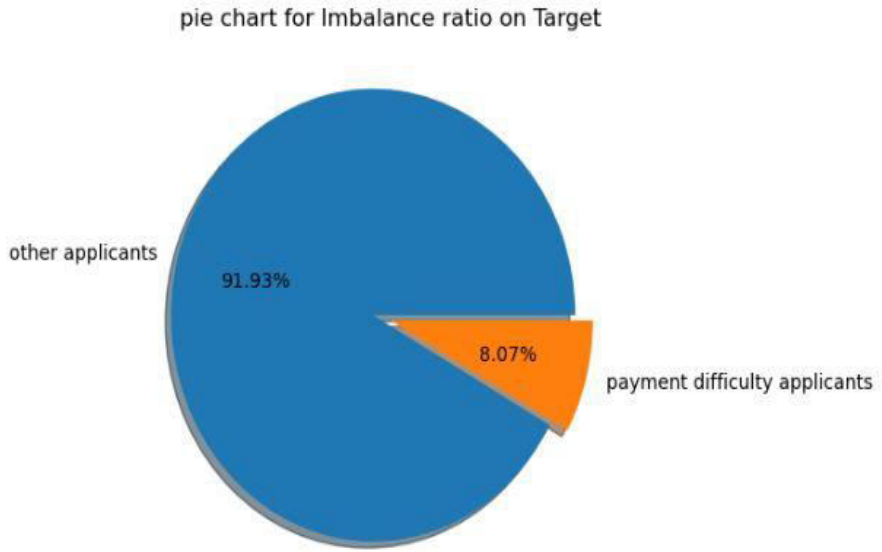




Data Imbalance

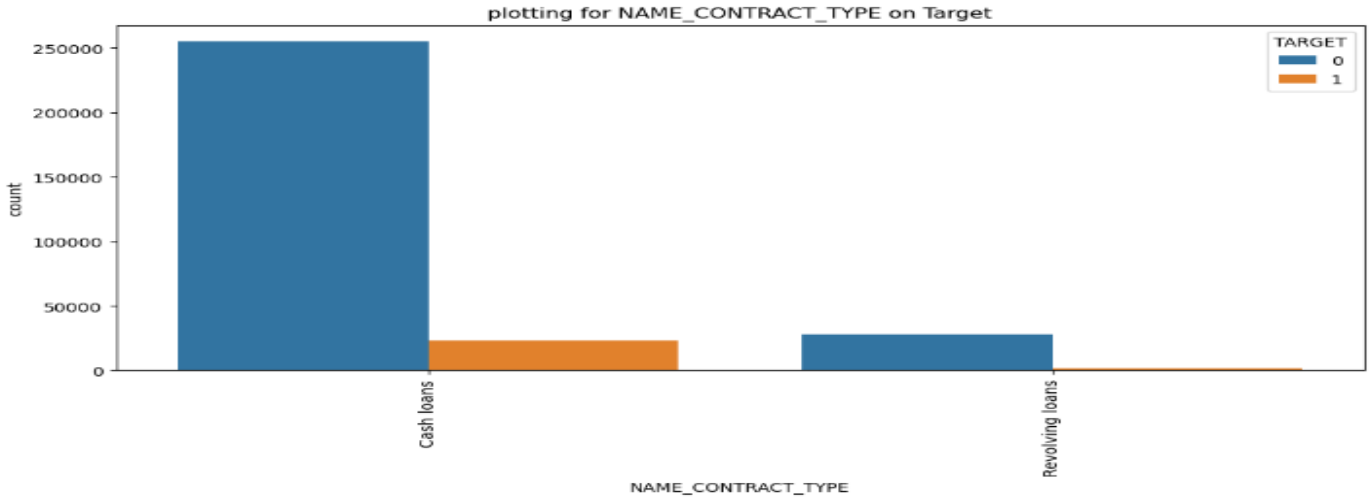
- There is huge data imbalance on Target for on-time payment applicants and payment difficulty applicants .
- The imbalance ratio is 11.39.

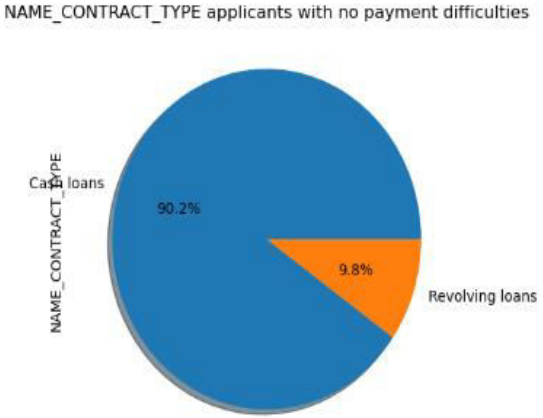


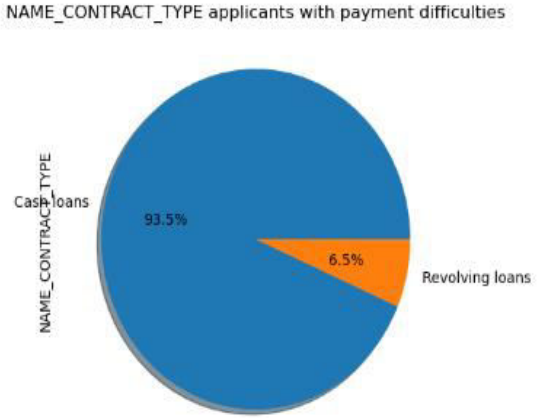


Univariate analysis for Categorical columns

- Seems like most of the clients are opting for Cash loans than revolving loans.

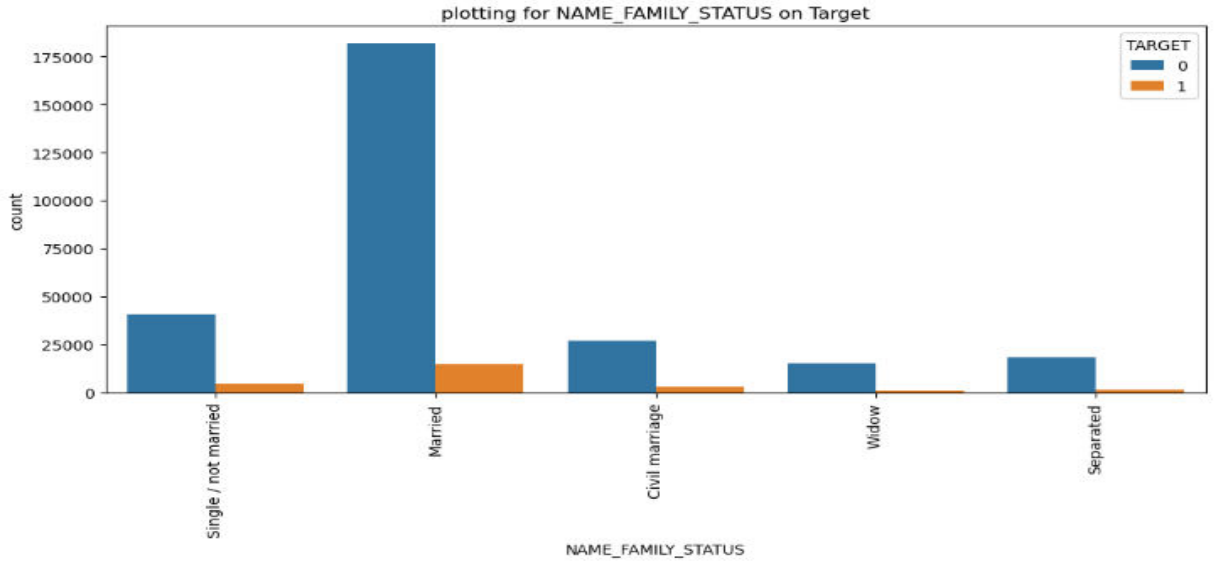


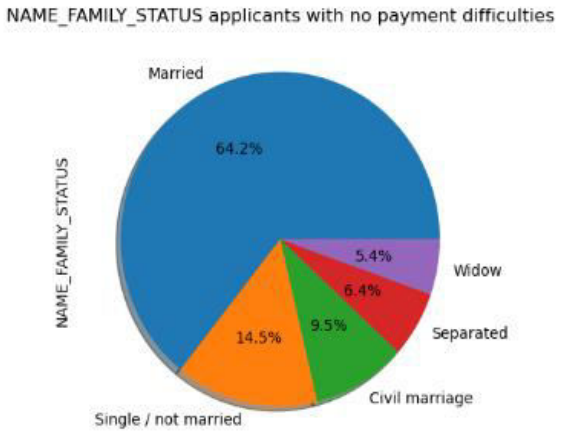


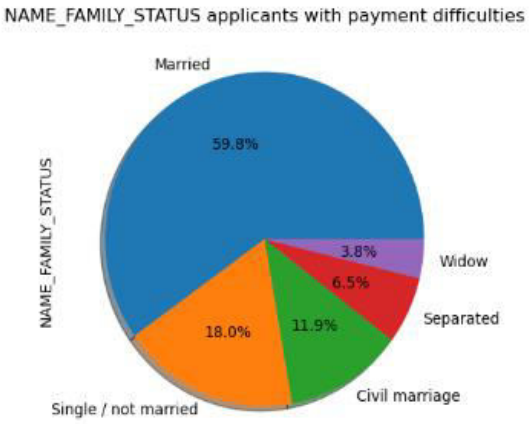


Univariate analysis for Categorical columns

- Applicants who are married are more in both as defaulters and non defaulters.

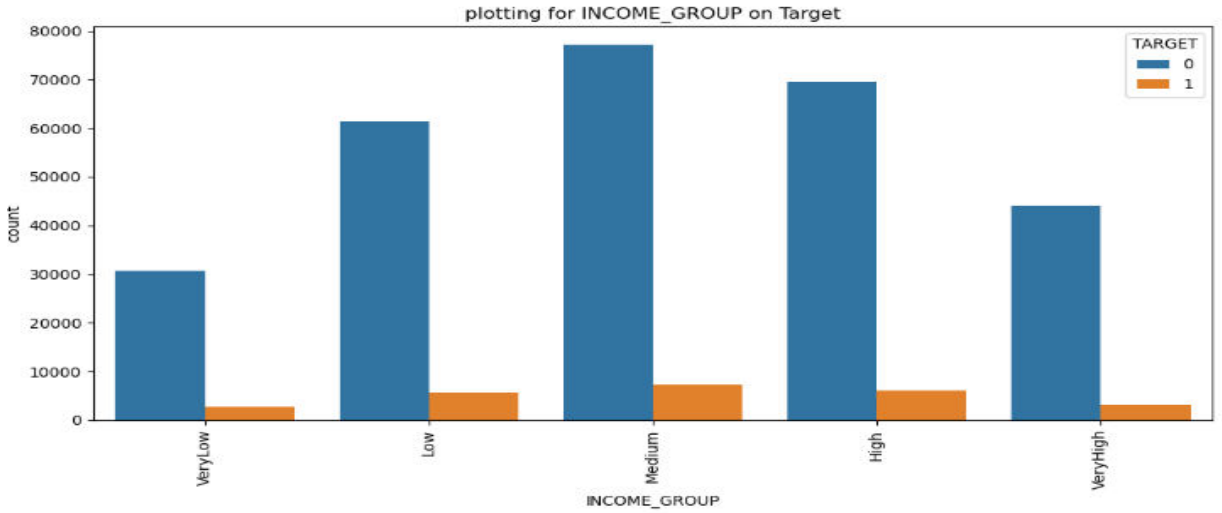


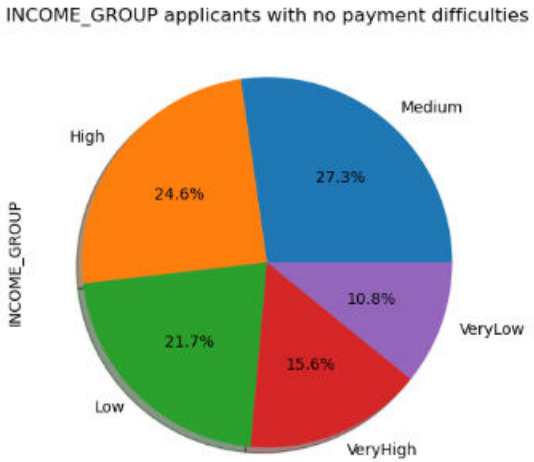


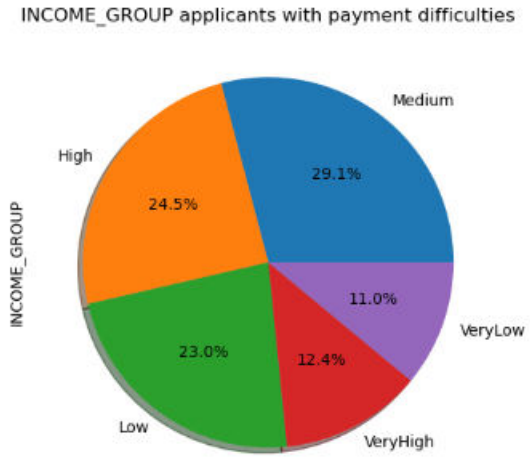


Univariate analysis for Categorical columns

- Applicants with medium income are both defaulters and nondefaulters.

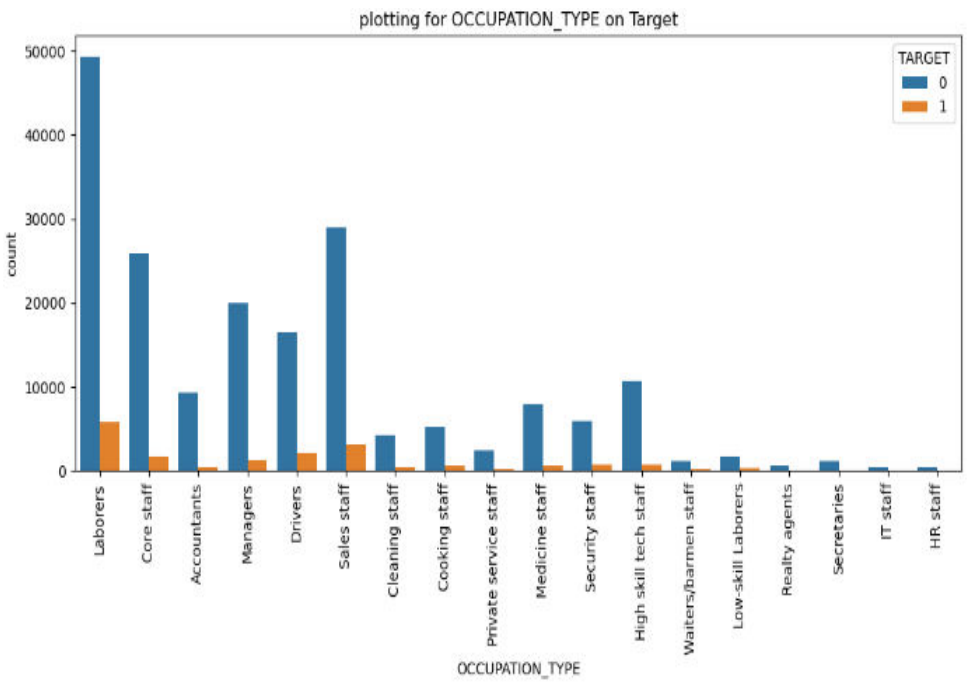


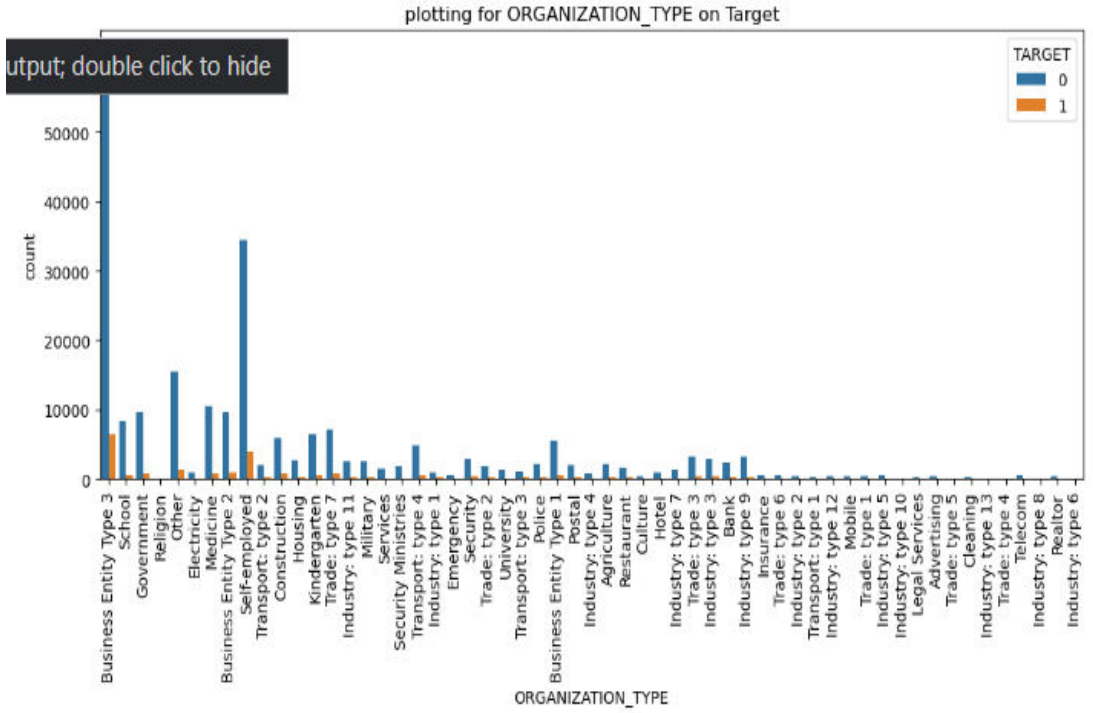




Univariate analysis for Categorical columns

- Analysis for OCCUPATIONTYPE and ORAGANIZATION_TYPE

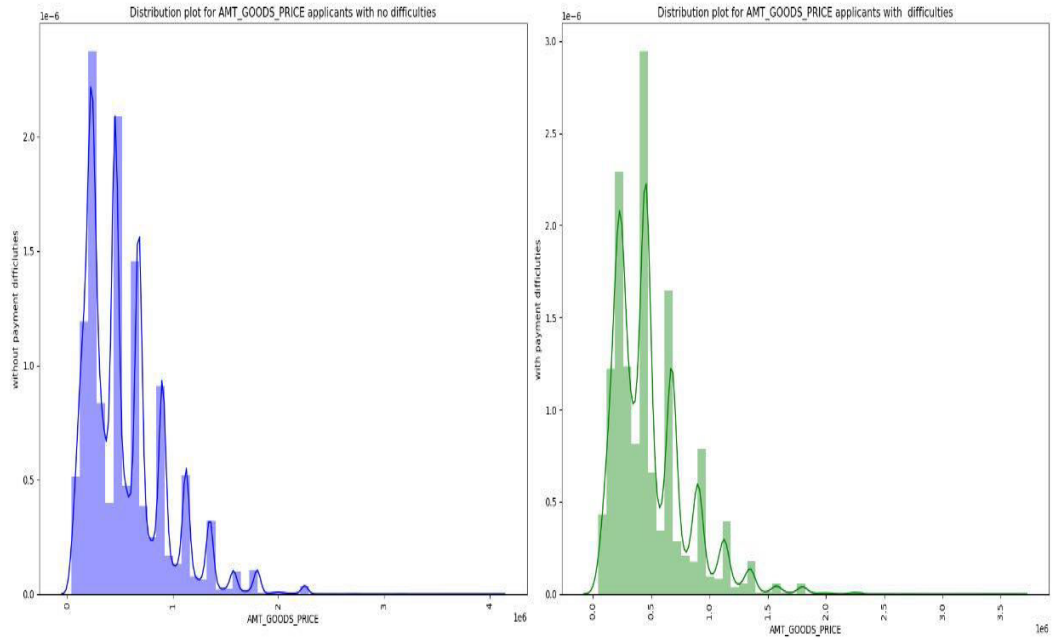




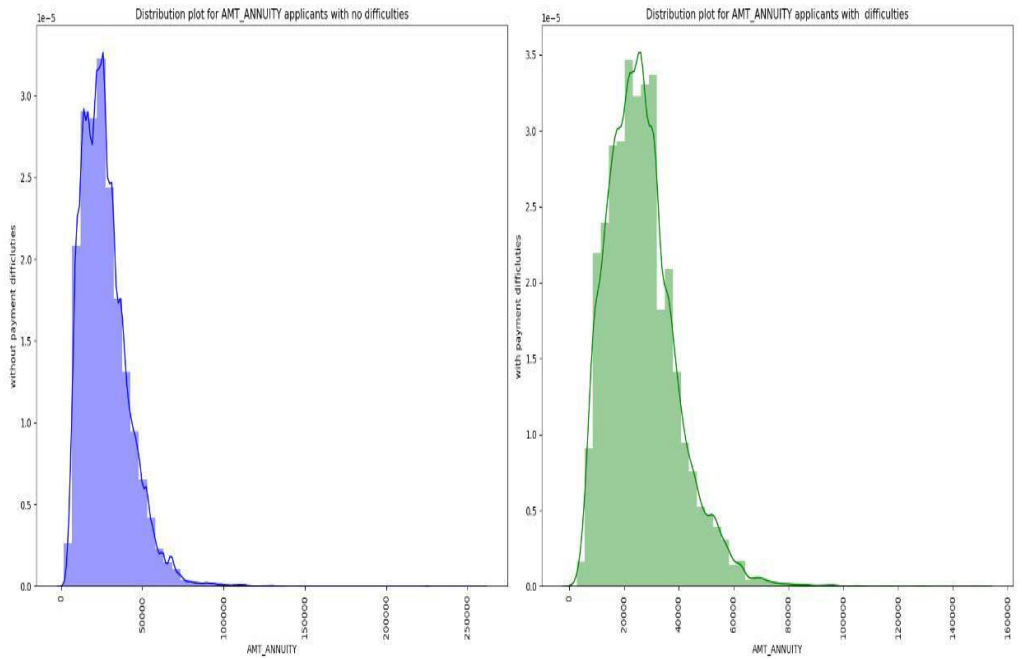
Univariate analysis for numerical columns

- We cant find much difference between defaulters and non defaulters.

Graph for AMT_GOODS_PRICE



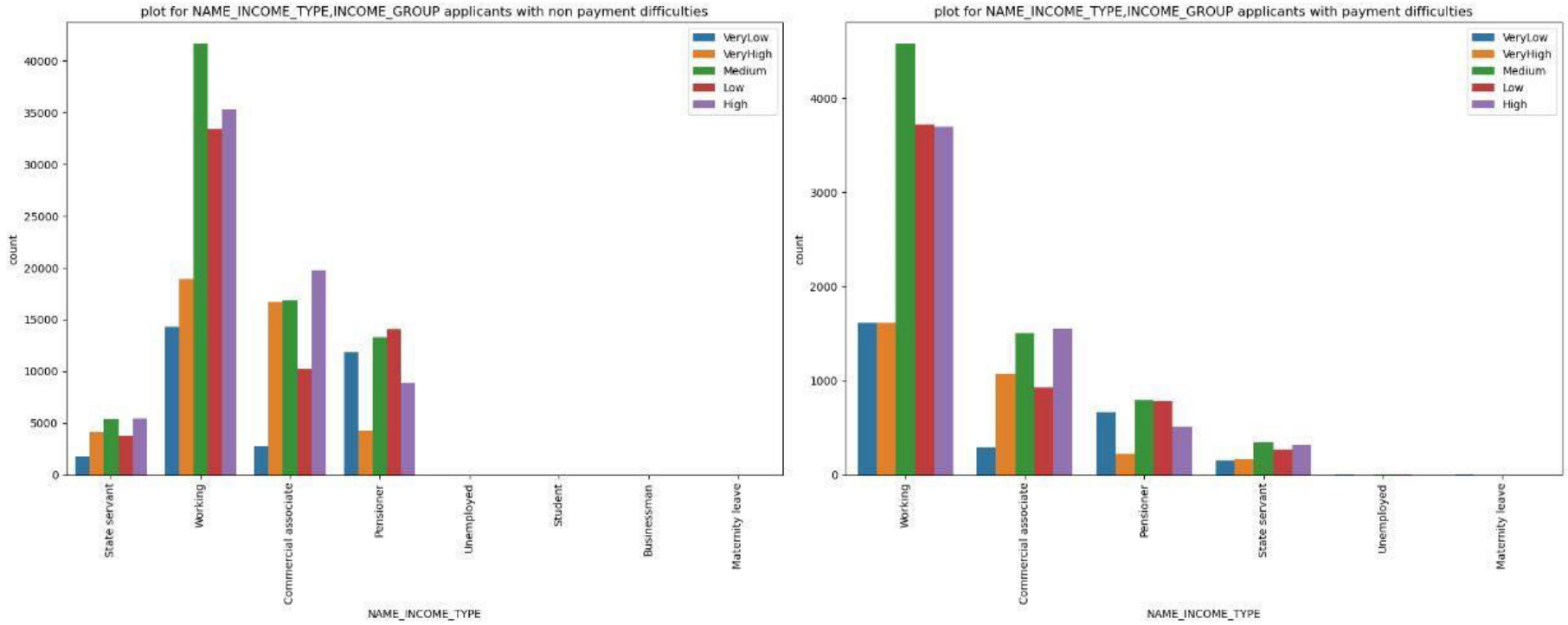
Graph for AMT_ANNUITY



Bivariate Analysis

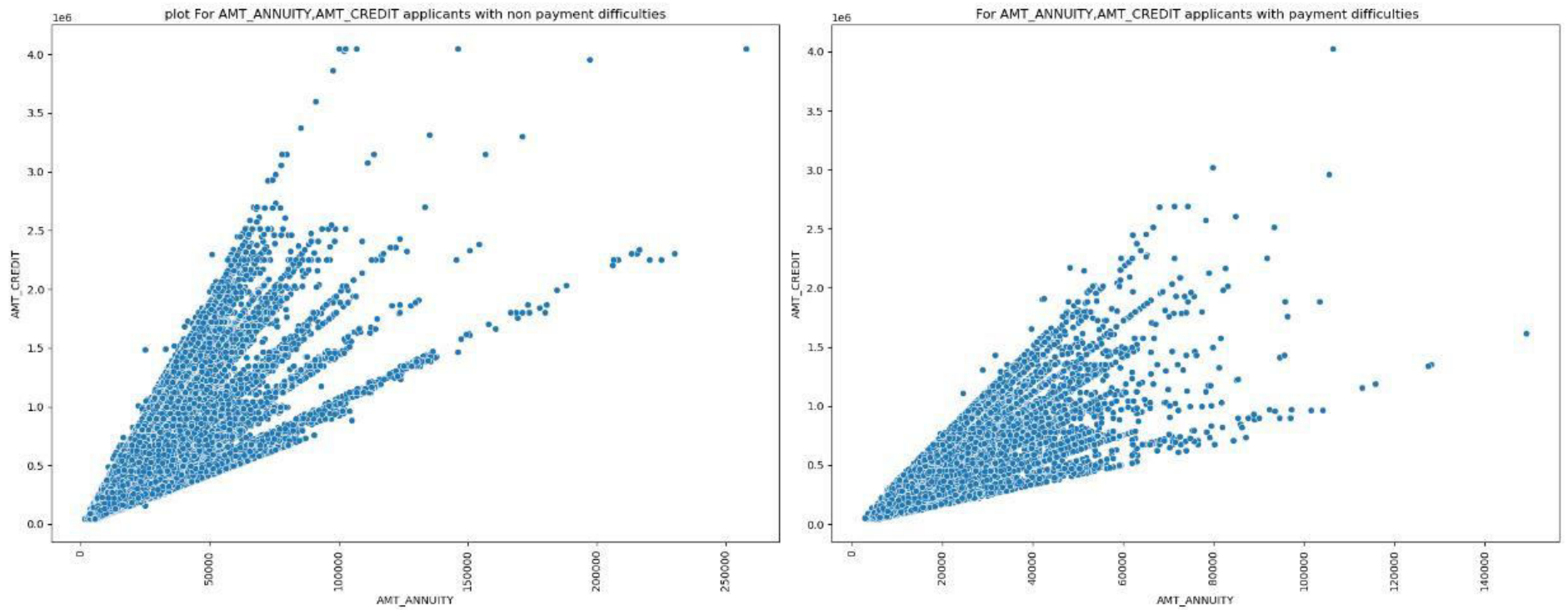
- We can see applicants who are working and has medium income are more with payment difficulties

Graph for NAME_INCOME_TYPE,INCOME_GROUP



Bivariate Analysis

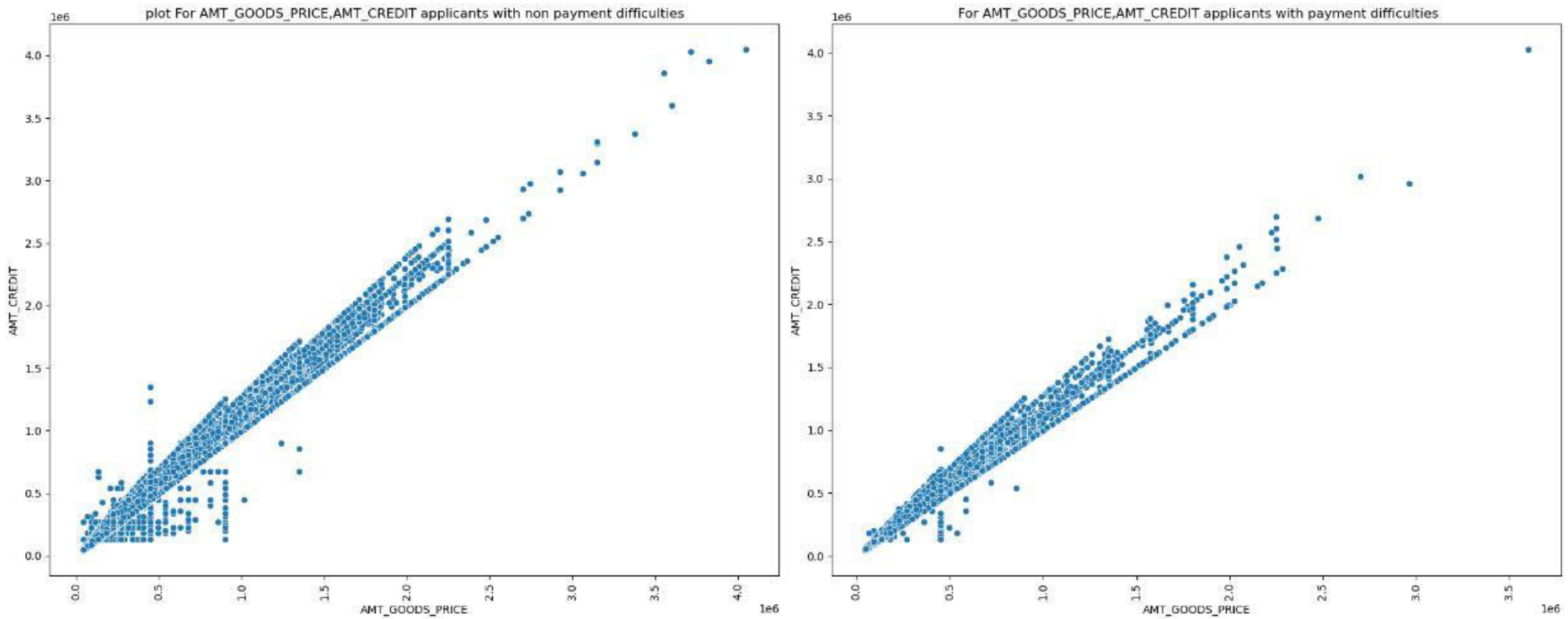
Graph for AMT_ANNUITY,AMT_CREDIT



AMT_ANNUITY AND AMT_CREDIT ARE HIGHLY CORRELATED

Bivariate Analysis

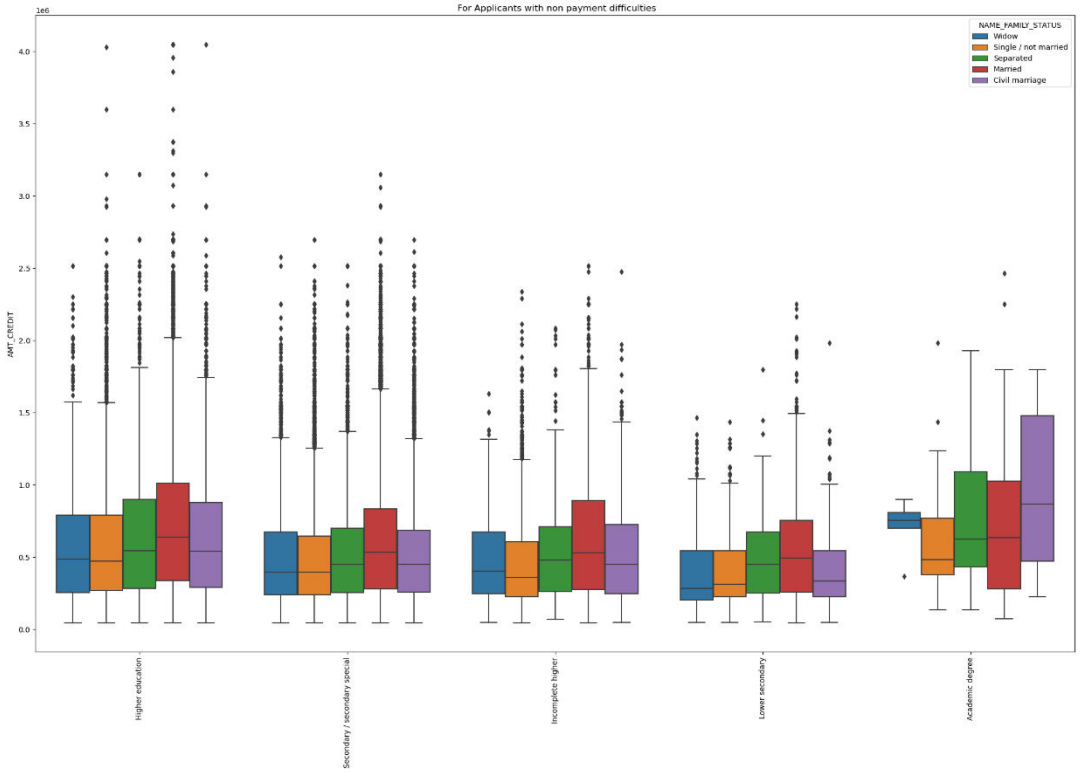
Graph for AMT_GOODS_PRICE,AMT_CREDIT

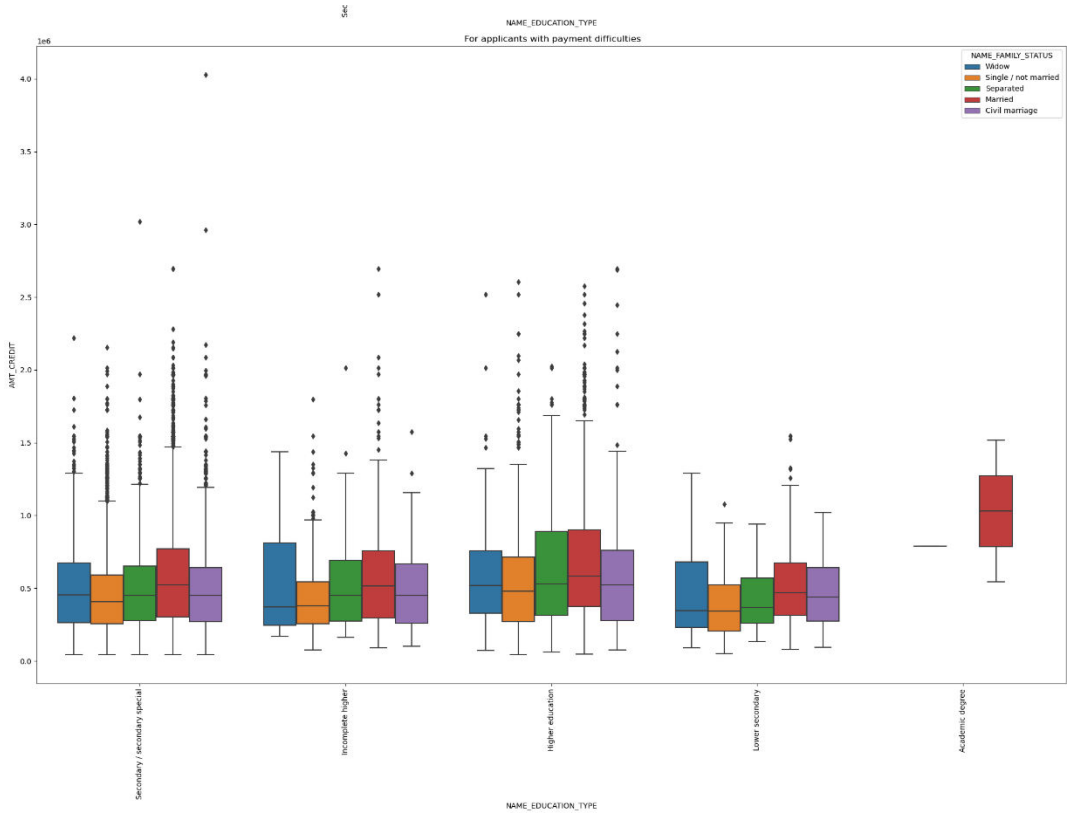


AMT_GOODS_PRICE and AMT_CREDIT ARE HIGHLY CORRELATED

Bivariate Analysis

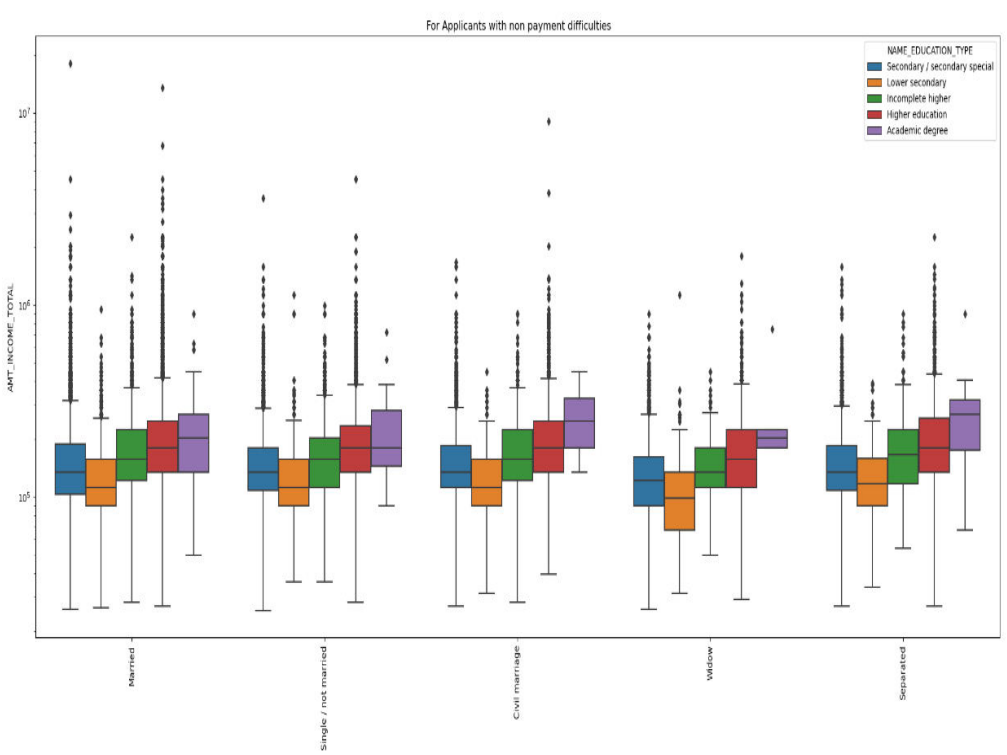
- Applicants with all Education types other than Academic degree have more outliers for non payment difficulties.

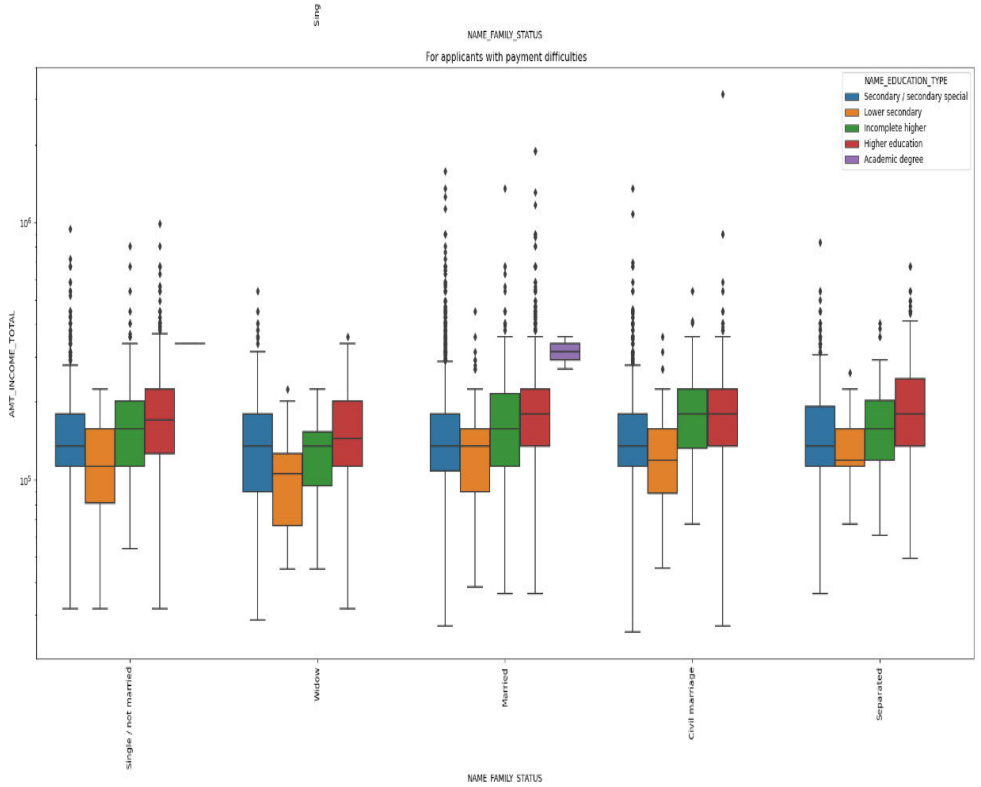




Bivariate Analysis

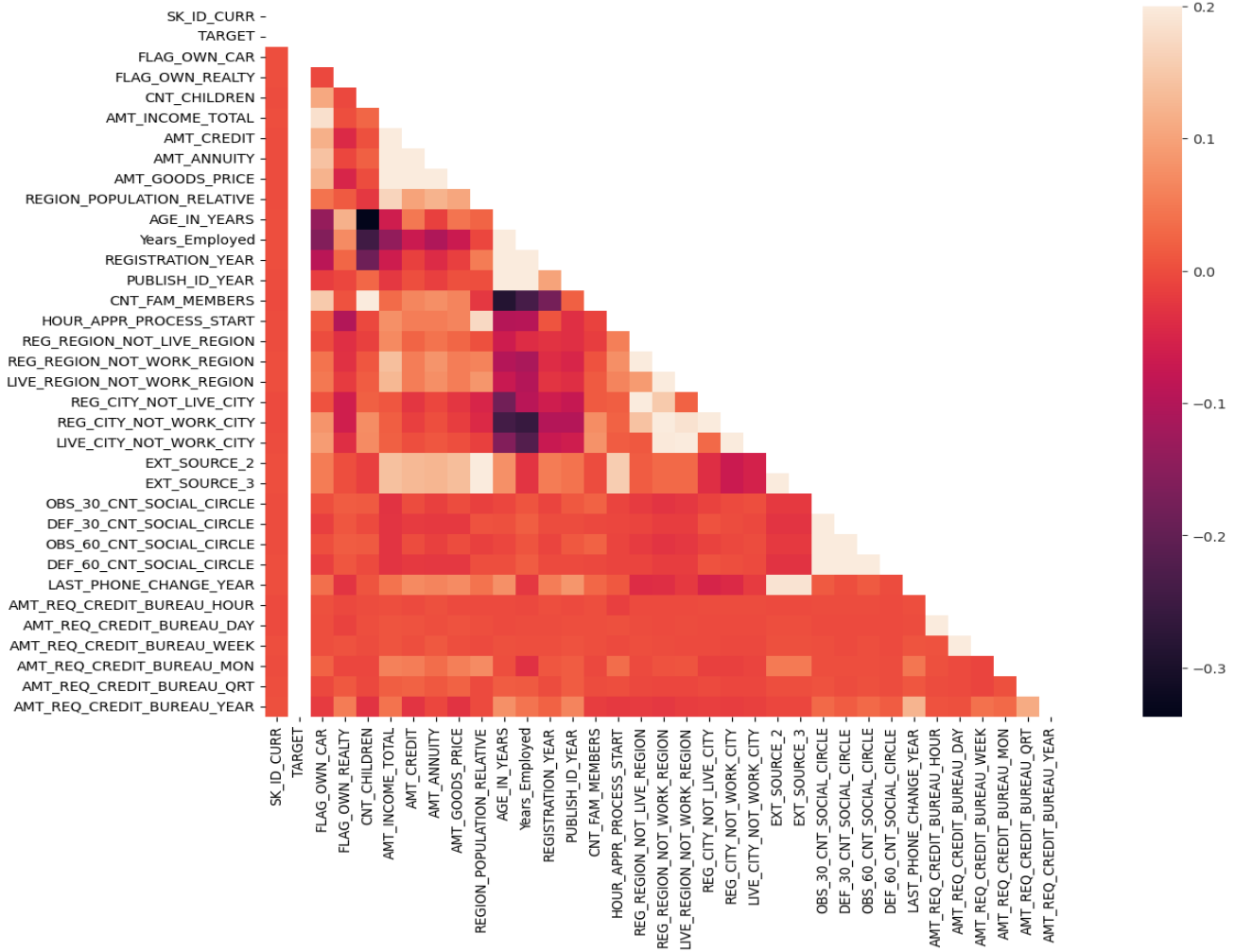
- Applicants who have payment difficulty has very less income when compared to nonpayment difficulty





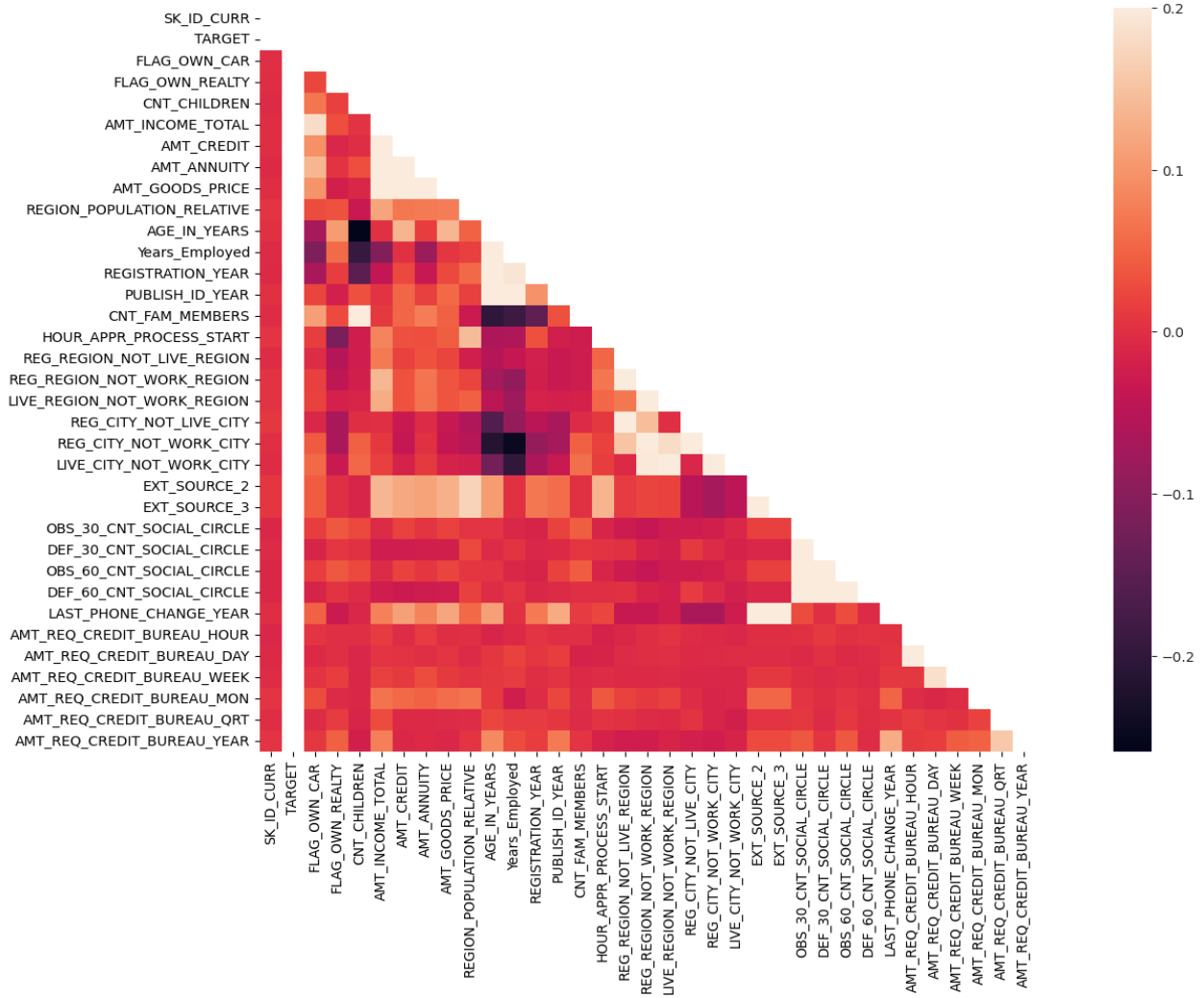
Correlations

	First_column	Second_column	Correlation%
934	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	99.85
286	AMT_GOODS_PRICE	AMT_CREDIT	98.70
494	CNT_FAM_MEMBERS	CNT_CHILDREN	87.86
647	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	86.19
970	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	85.94
755	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	83.04
287	AMT_GOODS_PRICE	AMT_ANNUITY	77.64
251	AMT_ANNUITY	AMT_CREDIT	77.13
395	Years_Employed	AGE_IN_YEARS	62.61
611	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	44.61



Correlations

	First_column	Second_column	Correlation%
934	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	99.83
286	AMT_GOODS_PRICE	AMT_CREDIT	98.28
494	CNT_FAM_MEMBERS	CNT_CHILDREN	88.55
970	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	86.90
647	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	84.79
755	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	77.85
287	AMT_GOODS_PRICE	AMT_ANNUITY	75.23
251	AMT_ANNUITY	AMT_CREDIT	75.22
395	Years_Employed	AGE_IN_YEARS	58.22
611	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	49.79



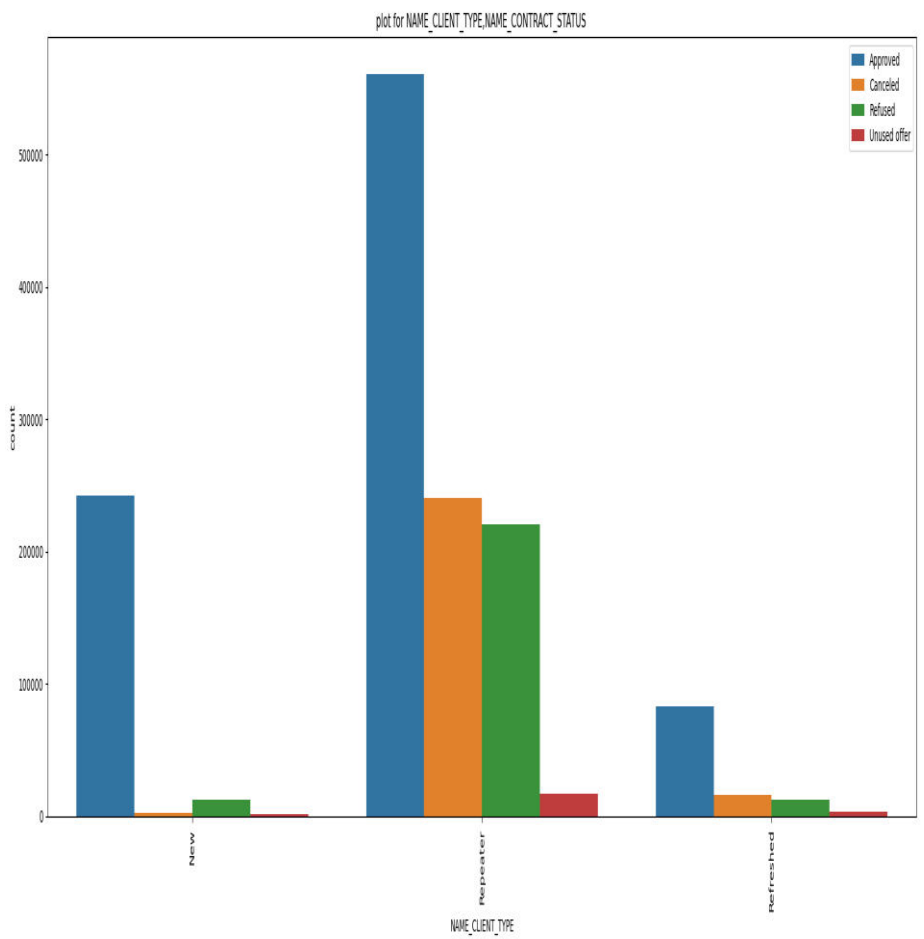
Merging Datasets

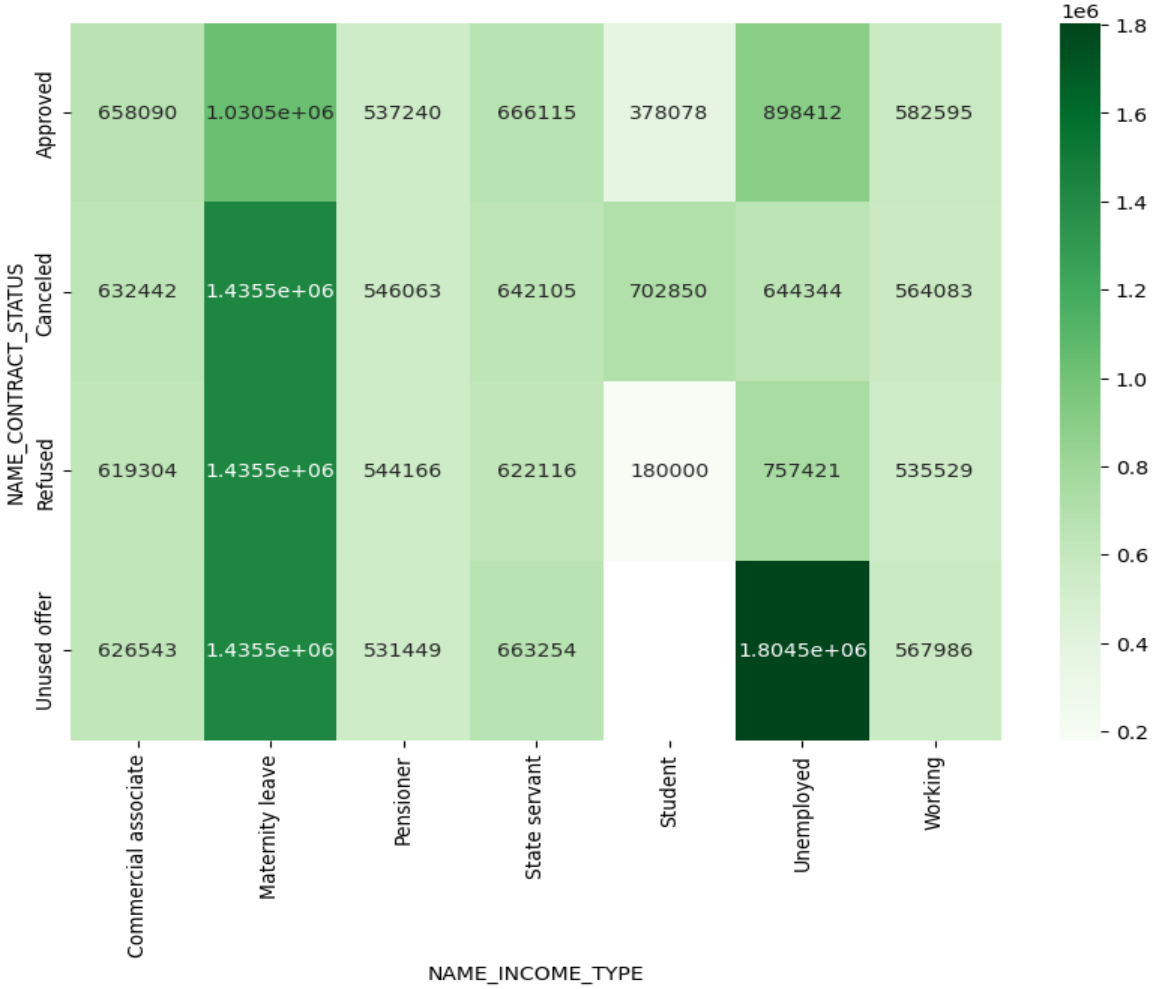
- After performing data cleaning and handling all the outliers from two datasets, we next merge the datasets.
- I have merged the datasets using left join, o column SD_ID_CURR
- I have performed analysis on merged datasets

Observation:

- The unemployed category is offered high credit.
- And for maternity leave applicants also.
- compared to approved other categories like unused refused and canceled have fewer credits

Analysis on Merge dataset





Conclusion

We can target the following applicants

- Applicants in the age group 30-50.
- Applicants with an academic degree.
- Applicants who are repeaters
- married applicants.
- Applicants who have medium income.
- Male applicants can also come under defaulters.

THANK YOU