Loan Application Data Analytics

Defaulters VS Non-Defaulters

EDA performed by:
Arpit Vijay (arpitvj1993@gmail.com)
Dhananjay Punekar (dhananjay.punekar36@gmail.com)

Objective of the EDA:

- We are given this case study to identify the patterns and factors by which company can decide about giving the loan on the basis of gain information and understanding.
- Identify patterns which indicate if a client has difficulty paying their installments.
- Understanding the driving factors (or driver variables) behind loan default.
- Finding the top 10 correlation for the Client with payment difficulties and all other cases.

Our Approach

- Data cleaning
  - Dropping columns with more than 40% null values.
  - Converting the columns to proper data types.
  - Convert negative values in some date columns to positive.
- Finding outliers

(We have only found and visualised the outliers but have not performed any outlier treatment as it was optional !)

- Determine the correlation for Target 0 and Target 1
- Univariate analysis
- Bivariate analysis
- Analysing the merged dataset
- Summary and conclusion

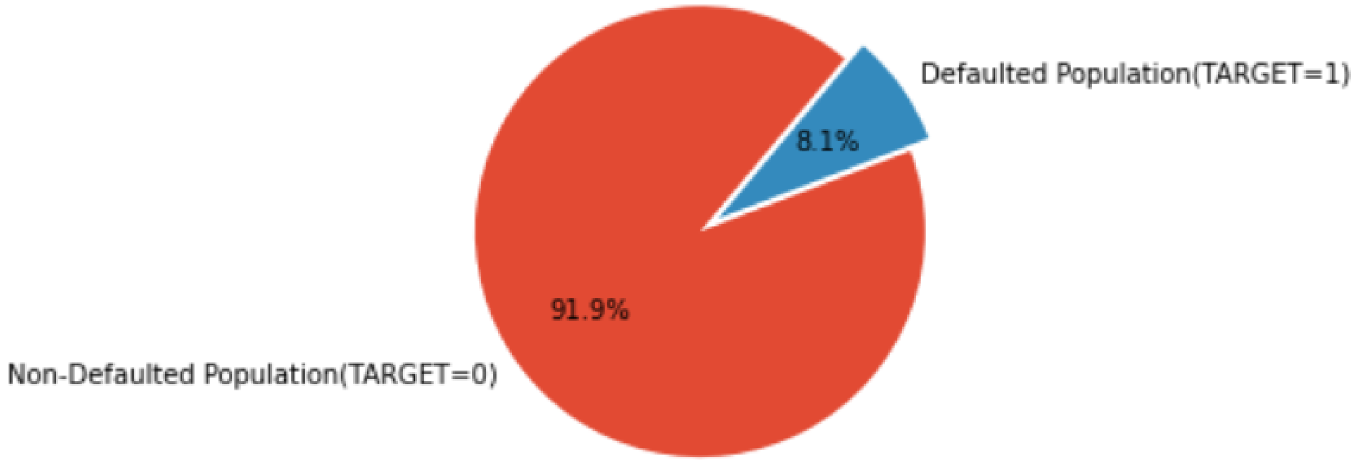
Application Data

Analysis

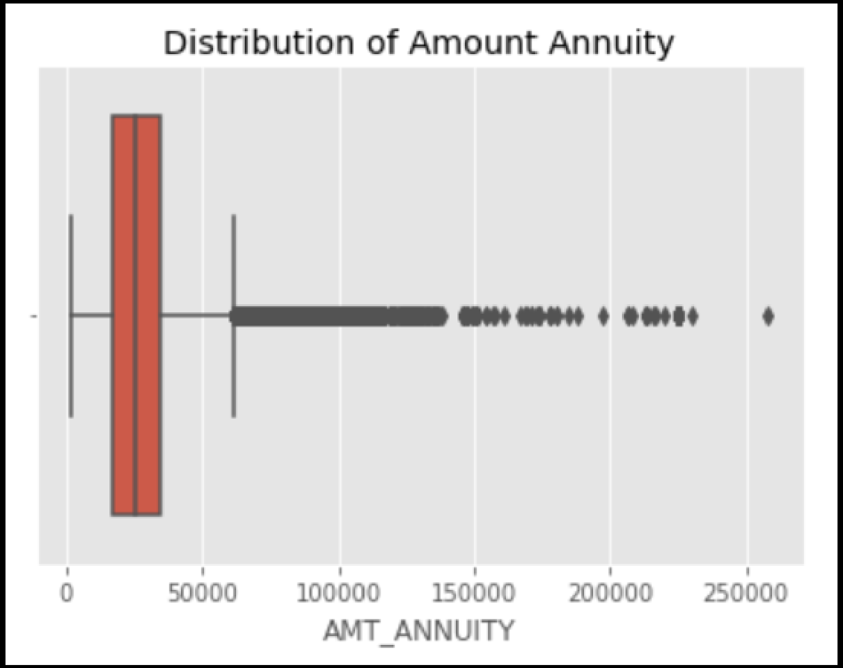
Understanding data at a high level:

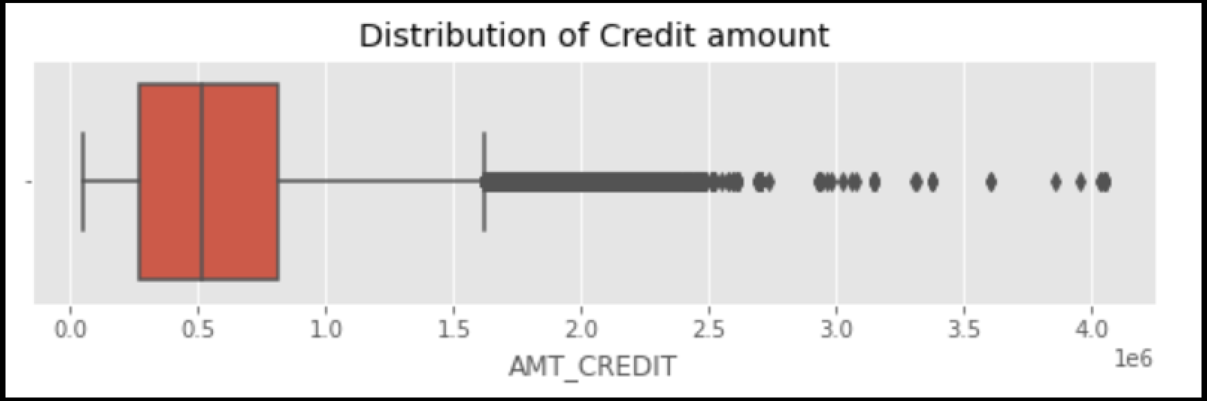
We are mainly targeting to see if an applicant is a defaulter or a non defaulter.

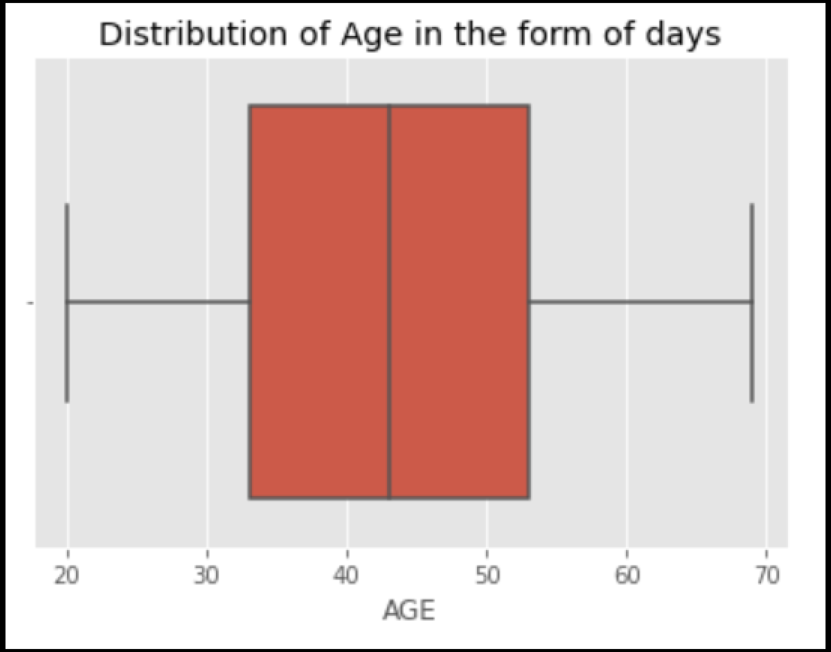
The given data, after a few treatments, has about 91.9% Non-Defaulter and 8.1% defaulters

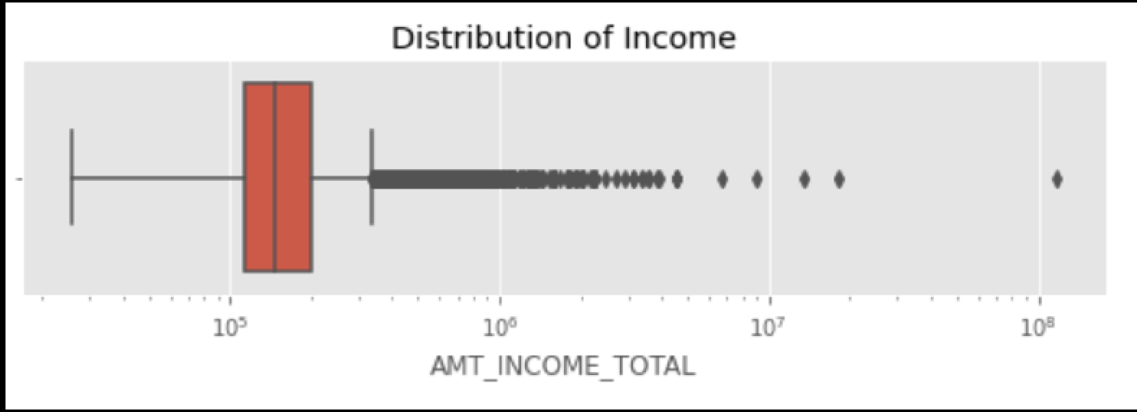


Outlier search





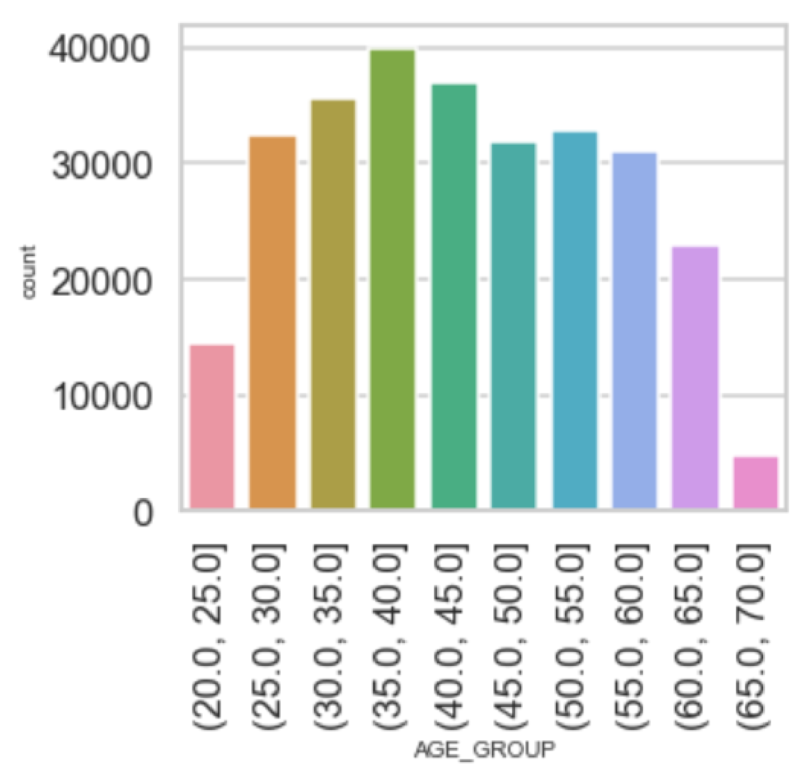




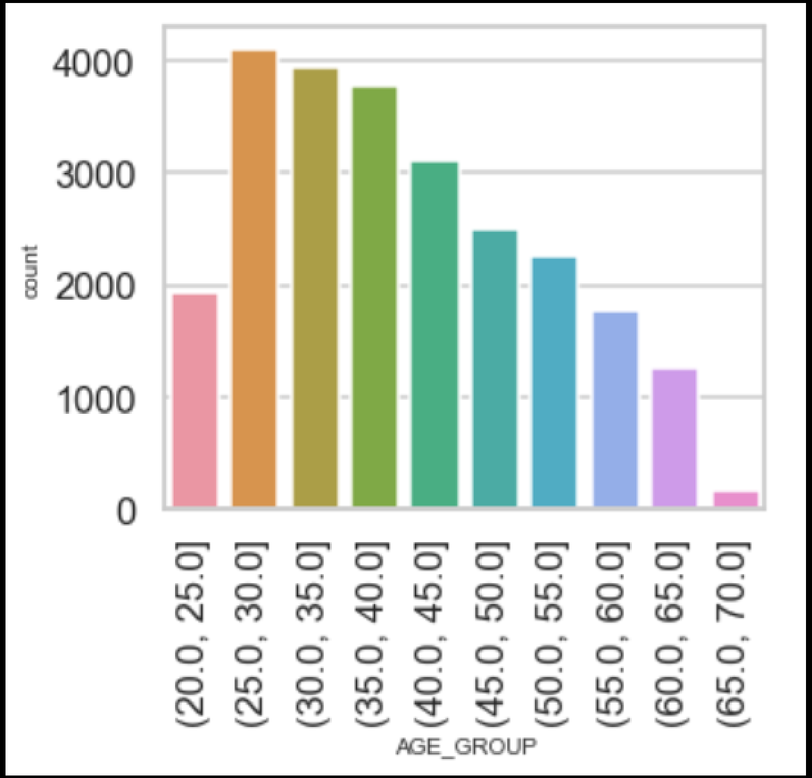
Which income range are the major applicants?

People in the age group of 25 to 45 apply highest for the loans

Target 0 (Non-Defaulters)

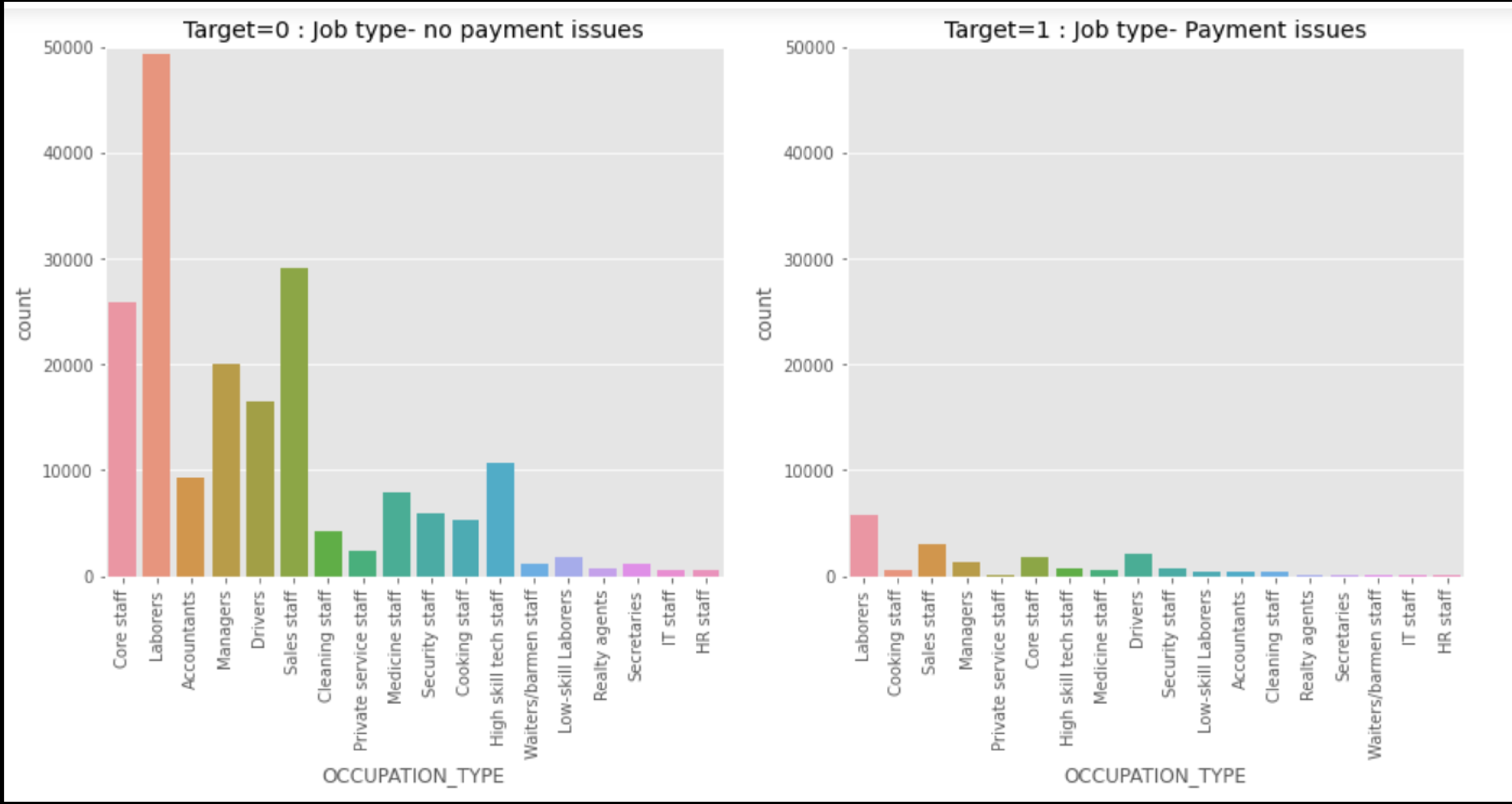


Target 1 (Defaulters)



Count of applicants by OCCUPATION_TYPE

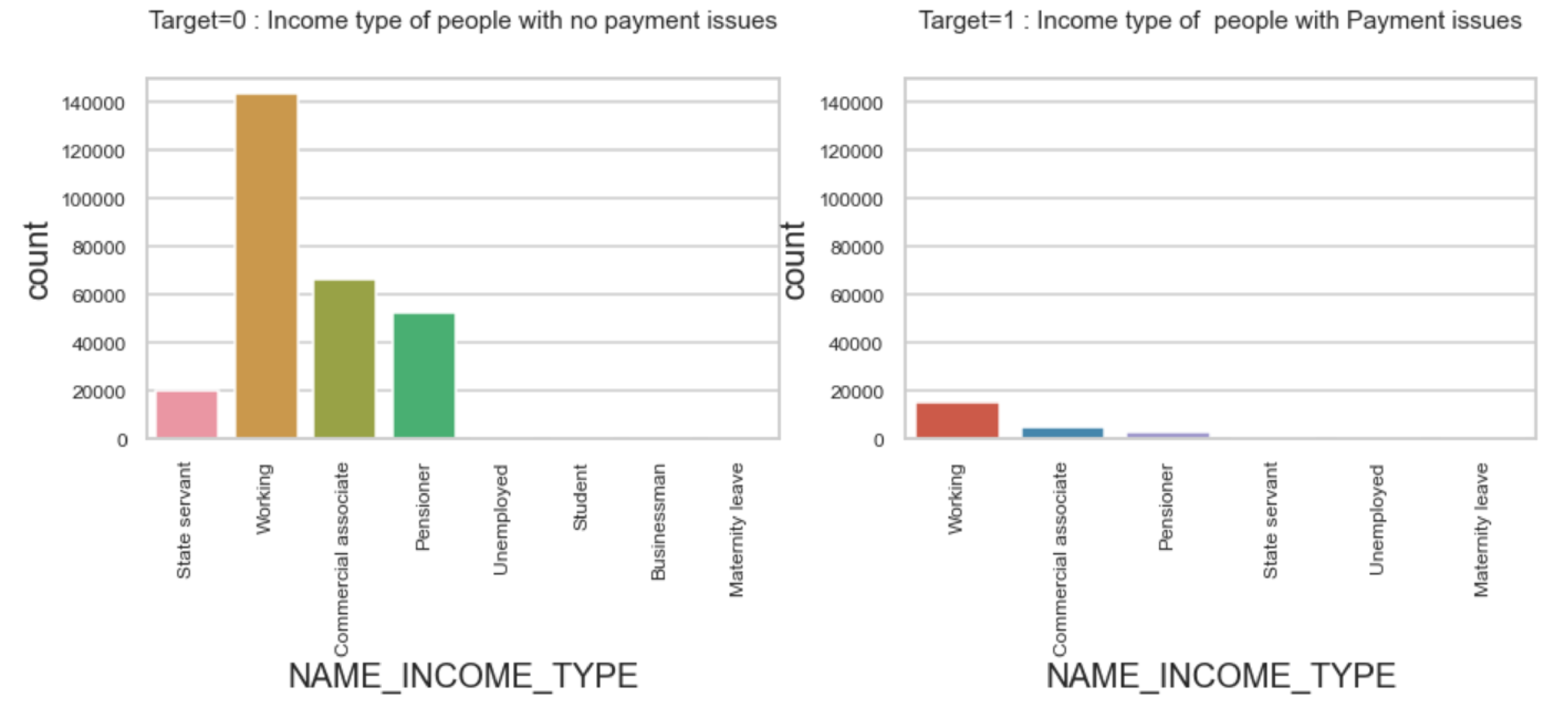
Univariate Analysis



Labourers are the most frequent applicants

Count of applicants by INCOME_TYPE

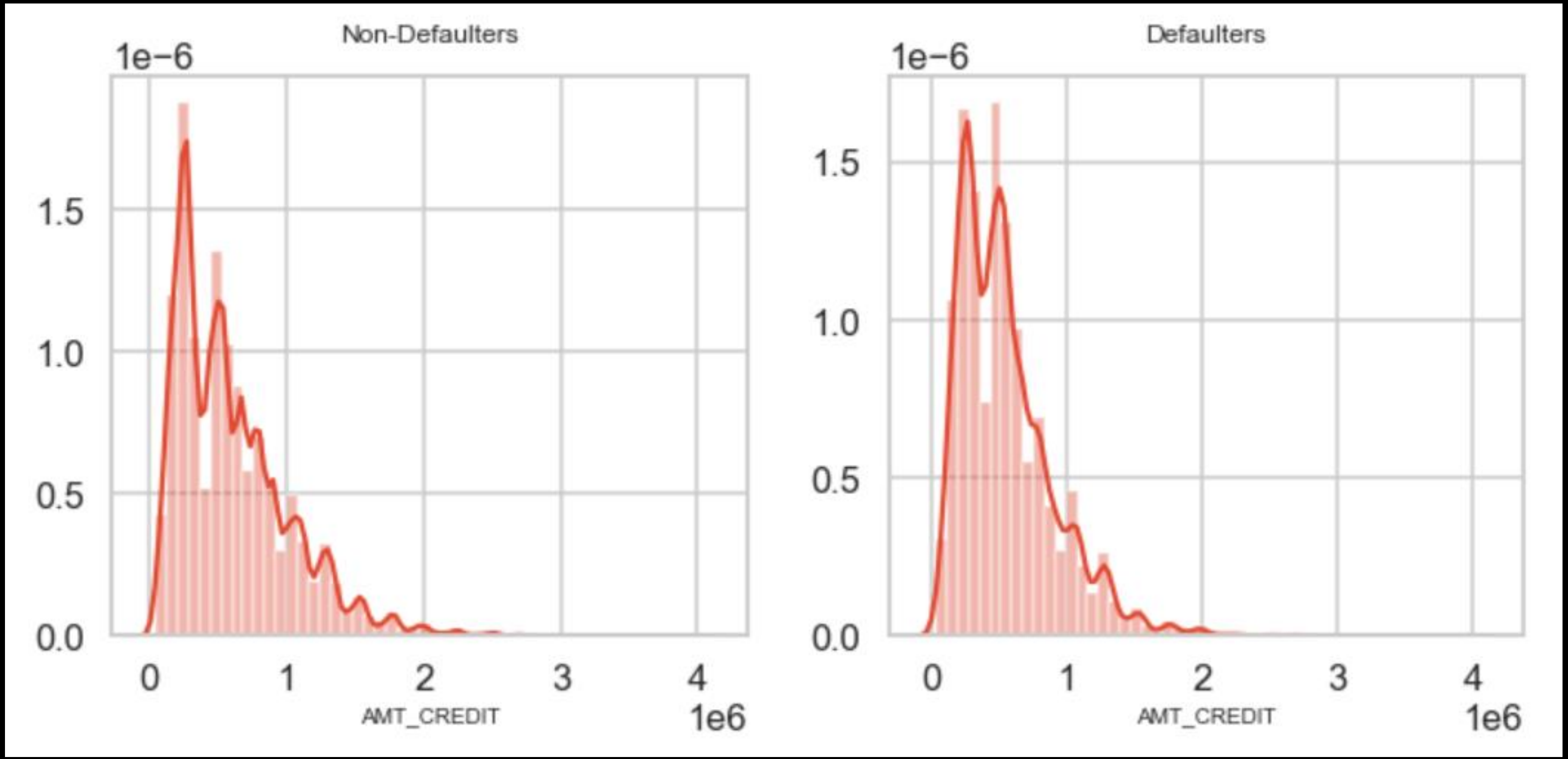
Univariate Analysis



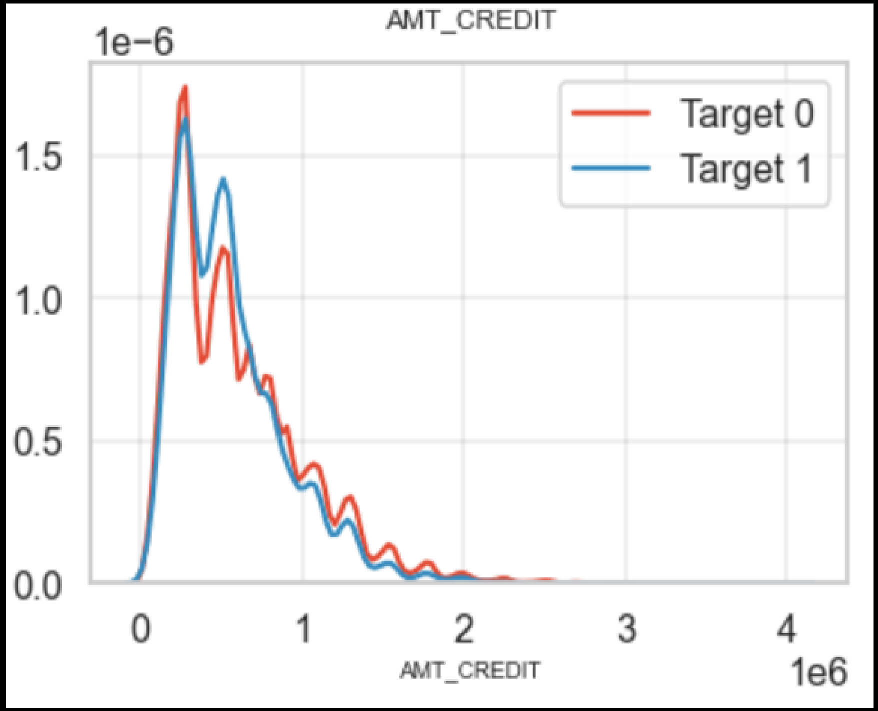
Working class people are the most frequent applicants

Distribution of Amount Credit

Numerical Univariate Analysis



The above 2 graphs can be plotted overlapping to get better insights

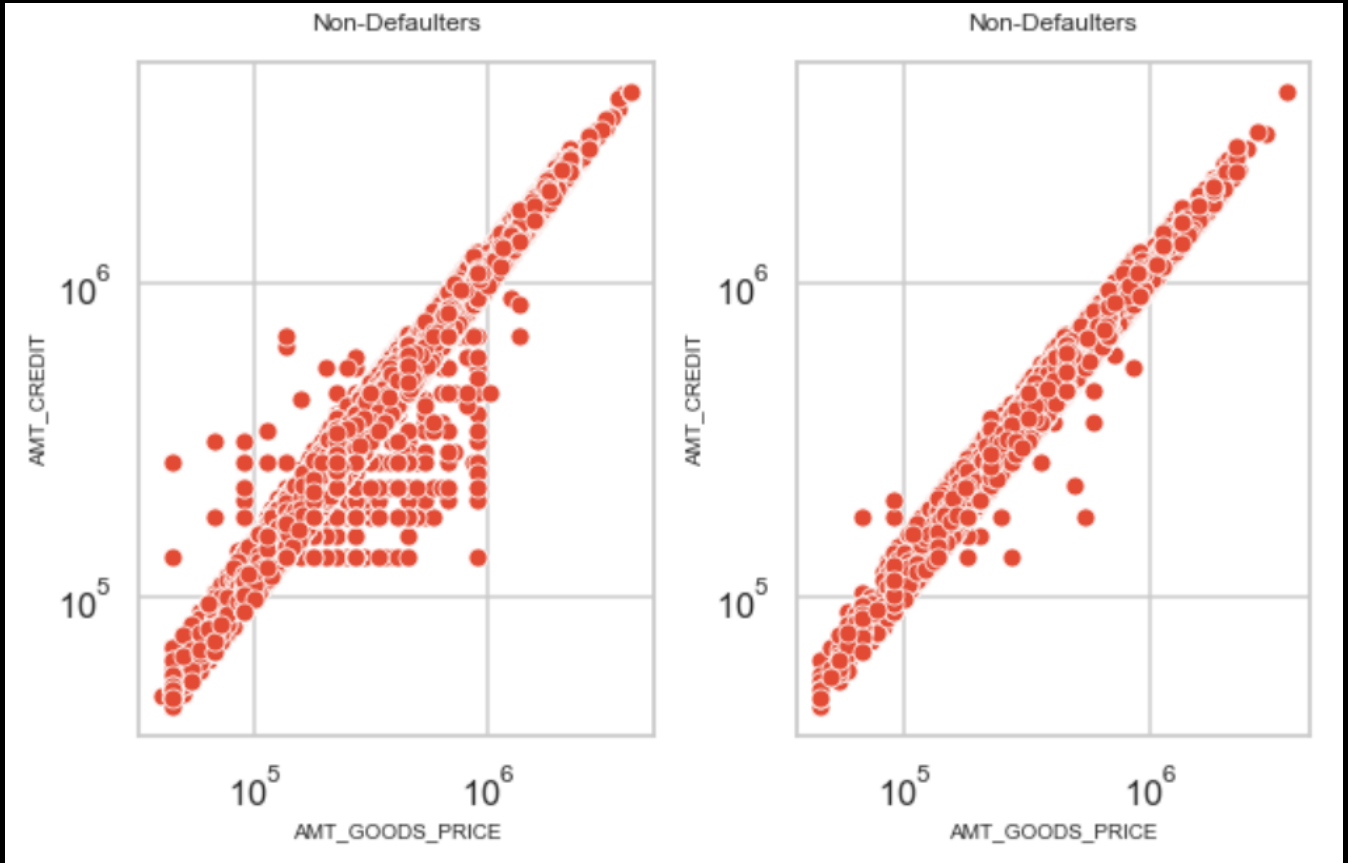


Overlapping plot

Credit amount is left skewed

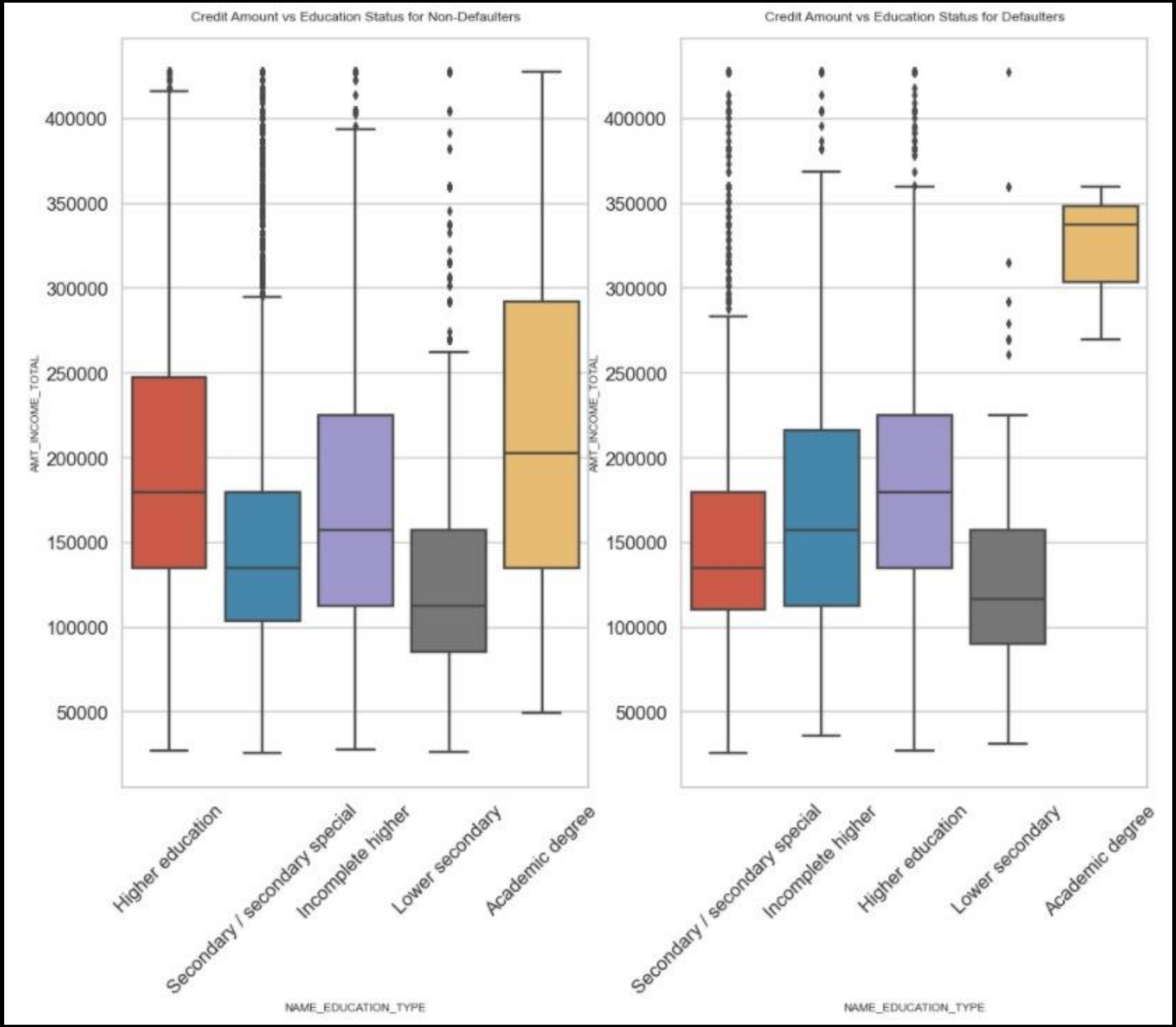
Variation of Amount Credit with Amount Good Price

Numerical Bivariate Analysis



Amount Credit and Amount goods price show a strong positive correlation

Variation of Amount Total Income across Education Type



Applicants with a Academic degree have a higher total income

Correlation Matric for Target 0 and Target 1

Target 1
Defaulter

Var1	Var2	Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998269
AMT_GOODS_PRICE	AMT_CREDIT	0.983103
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.868994
AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
AMT_ANNUITY	AMT_CREDIT	0.752195
DAYS_EMPLOYED	DAYS_BIRTH	0.582185
AMT_ANNUITY	AMT_INCOME_TOTAL	0.430682
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.354769
AMT_CREDIT	AMT_INCOME_TOTAL	0.353619
OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.337181

Target 0
Non-Defaulter

Var1	Var2	Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998508
AMT_GOODS_PRICE	AMT_CREDIT	0.987253
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.859289
AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
AMT_ANNUITY	AMT_CREDIT	0.771308
DAYS_EMPLOYED	DAYS_BIRTH	0.626116
AMT_ANNUITY	AMT_INCOME_TOTAL	0.488599
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.419921
AMT_CREDIT	AMT_INCOME_TOTAL	0.414447
DAYS_BIRTH	CNT_CHILDREN	-0.336966

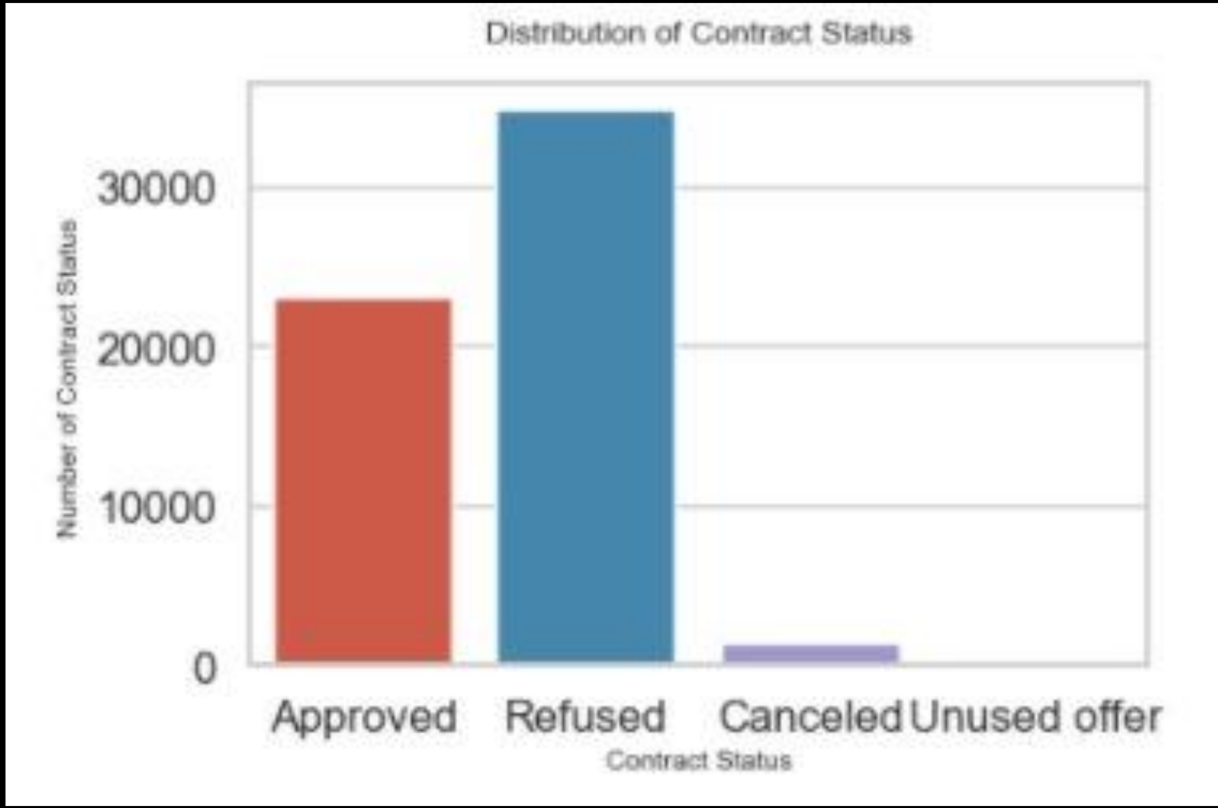
Inference: For Default and Non-default population, the Top 10 correlations are same

Merged Data Set

Application Data and Previous Application Data

Analysis

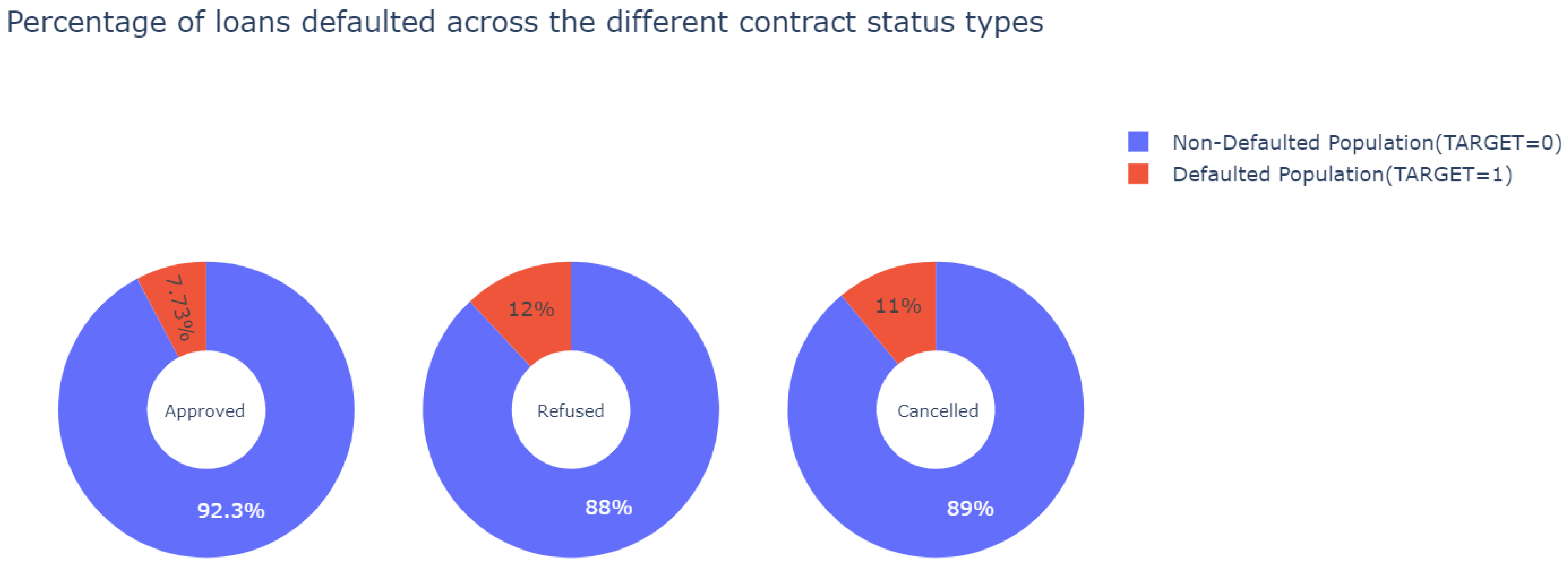
Application status for common applicants



Most of the applicants between previous and current applicants have been refused credit previously

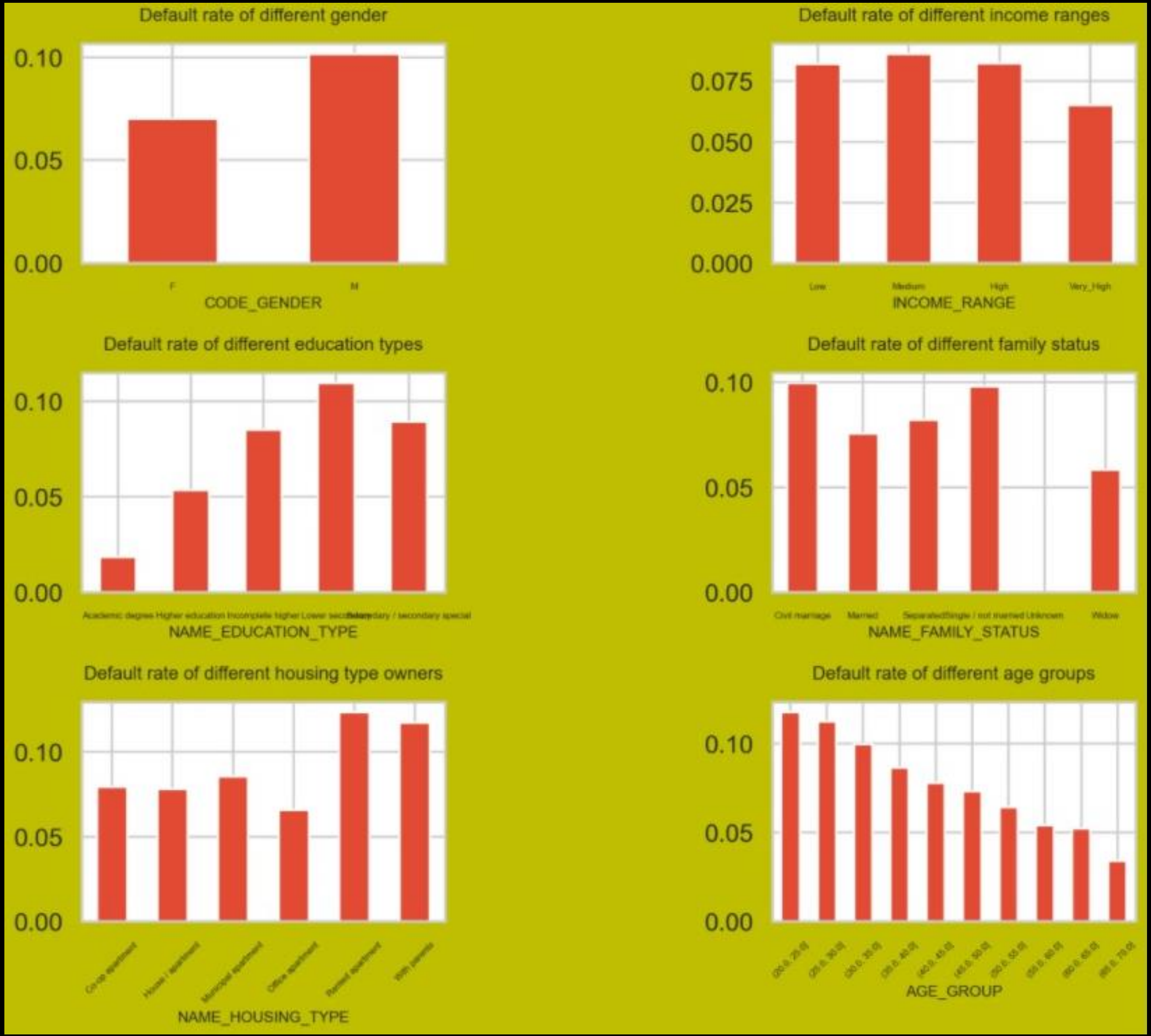
Analysis on merged data set

Loans which have been refused and cancelled before have more chances to default as compared to the approved ones.



Default Rate Analysis

Default rate is assumed as the ratio of number of defaulted on total number of loans in the segment we are observing



Final Conclusions

- People with 'Low' income range have higher chances of defaulting, therefore we should focus on other income ranges over this.
- People with 'Lower secondary' education and 'Single' status have the highest default rate. Therefore, we should be very careful while providing them loans. An authenticated guarantor's presence should be considered mandatory.
- Among both genders, even though there are more females applicants, it is still observed that females are lesser defaulters than males. Therefore, providing loans to females over males can be a plus point.
- People with 'Rented apartments' as their housing type are the highest defaulters. Therefore, we should check the security assets as well as the income of the applicant thoroughly.
- Age group (20-25) are the highest defaulters. Whereas, income stability is better in the age groups from 25 to 45 and they are less likely to default. Therefore, we should offer more loans to (25-45) age groups.
- People with housing types - 'office apartments' and 'with parents' are least likely to default as compared to other categories. Therefore, we should focus more on providing loans to these applicants.
- Banks should focus less on income type ‘Working’ as they have most number of unsuccessful payments.
- Banks should focus more on contract type ’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments with loan purpose ‘Repair’ is having higher number of unsuccessful payments on time

Thank You