Credit EDA Case Study Presentation

By

Mr. S. Krishna Kumar

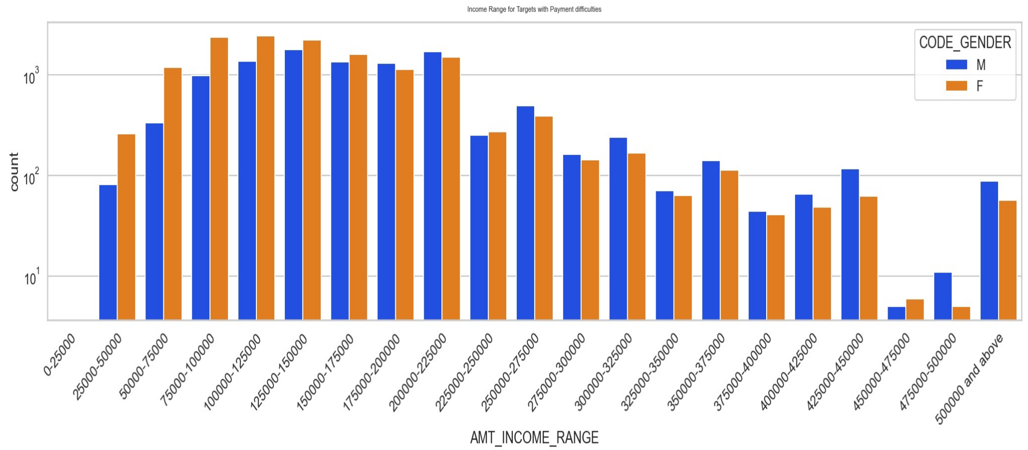
Purpose of the Exercise

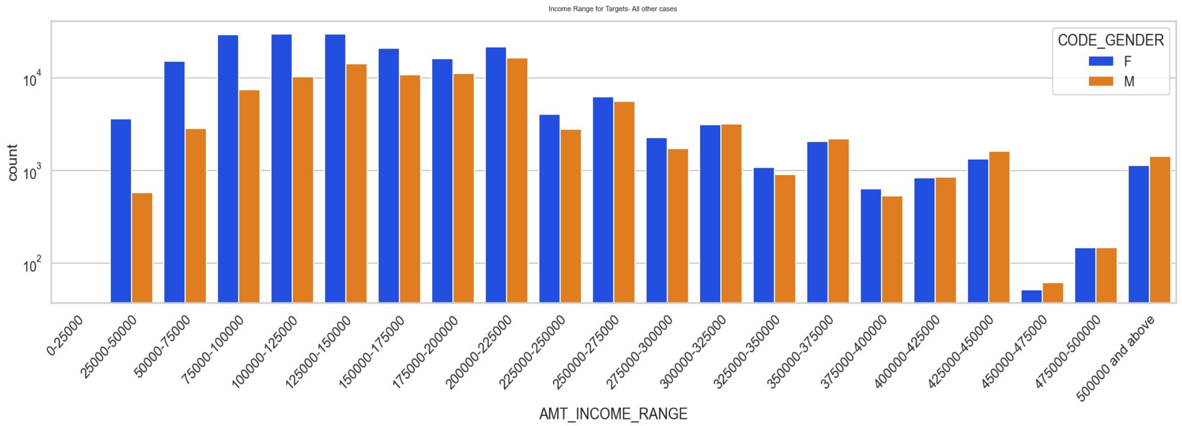
- Perform EDA on current application data & previous application data to predict the credit worthiness of an applicant's profile. This will help the company to avoid lending loans to defaulters.
- Finding potential profile and provide loans to them.

Steps to EDA

1. Understanding the datasets
2. Data-Cleaning, Imputing and Bucketing columns wherever necessary
3. Perform univariate and bivariate analysis and find correlation between column variables.
4. Merge previous application data and current application data
5. Perform univariate and bivariate analysis
6. Conclude risks and recommendations

Distribution of Income and Code-gender





Observation:

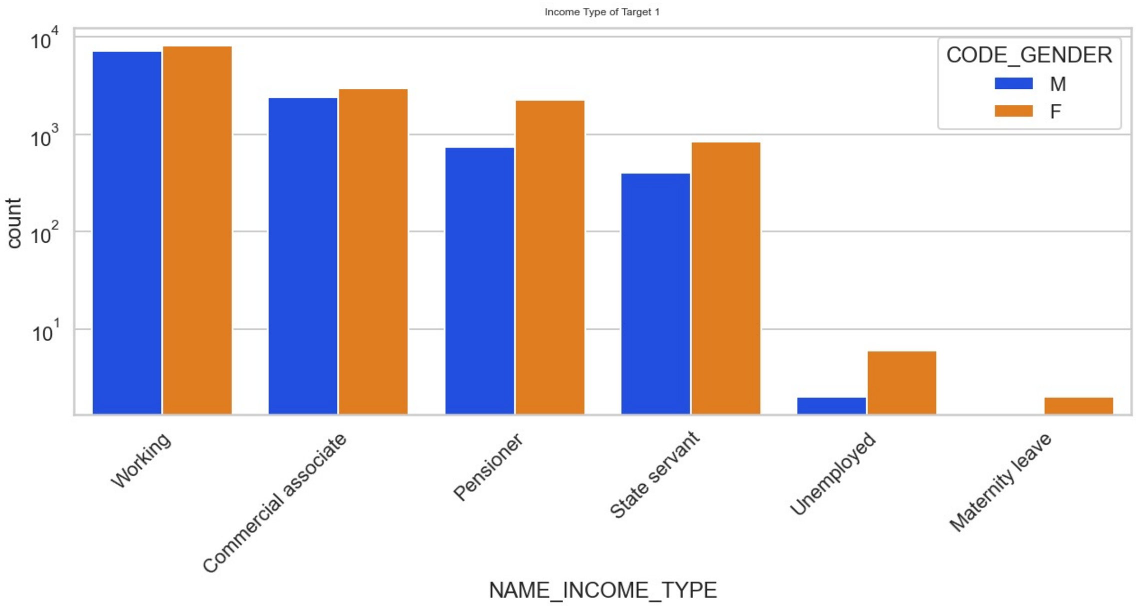
From the above graph for Target=1 (Defaulters) the following points can be concluded.

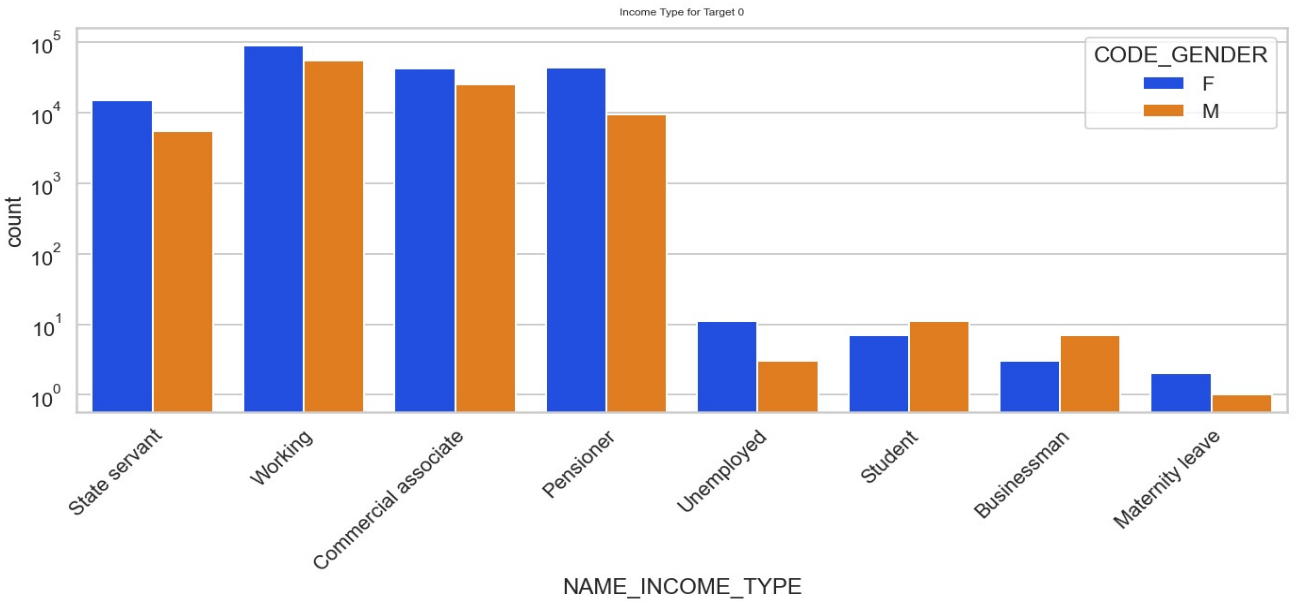
1. The income range of males were more compared to females at ranges 500000 and above .
2. This graph show that males are more than female in having income ranges between 175000 and 225000.

From the above graph for Target=0 (Non-Defaulters) the following points can be concluded.

1. The income range of female counts are higher than male at ranges 25000 and 200000
2. This graph show that females are more than male in having income ranges for that range.

Distribution of Profession and Code Gender





Observation:

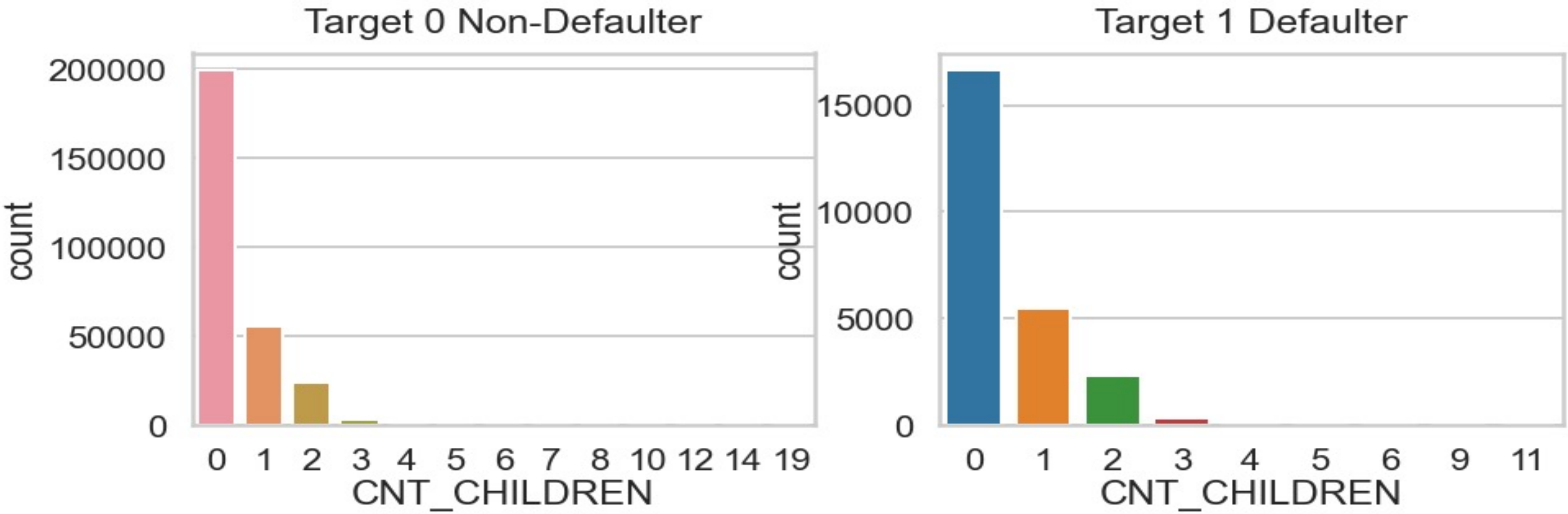
From the above graph for Target=1 (Defaulters) the following points can be concluded.

1. Income type working, commercial associate, pensioner and state servant defaulted more compared to unen
2. Unemployed has lesser credit compared to other categories.

From the above graph for Target=0 (Non-Defaulters) the following points can be concluded.

1. Females defaulted less compared to males.
2. Students and Businessman never defaulted based on the above data.

Distribution of Number of children

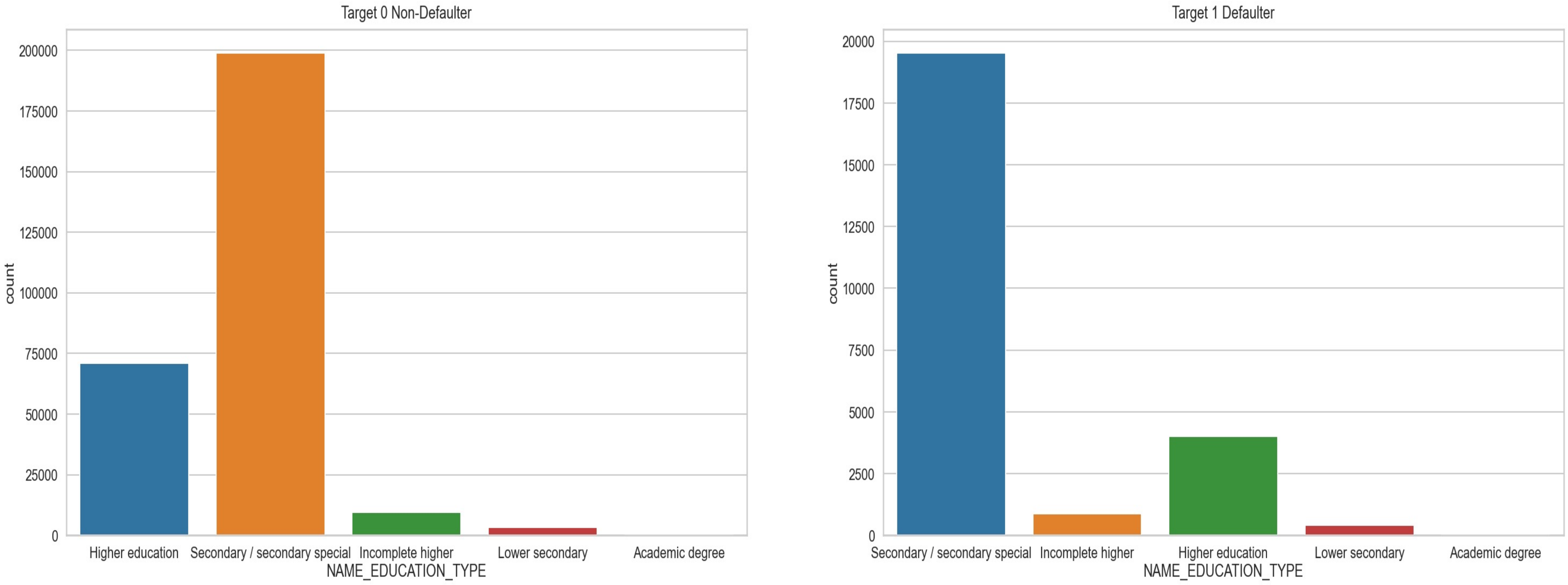


Observation:

From the above graph the following points can be concluded.

- Both the graph pretty much looks the same. So, children count doesn't determine the chances of being a defa
or a non-defaulter.

Distribution of defaulters with Education type

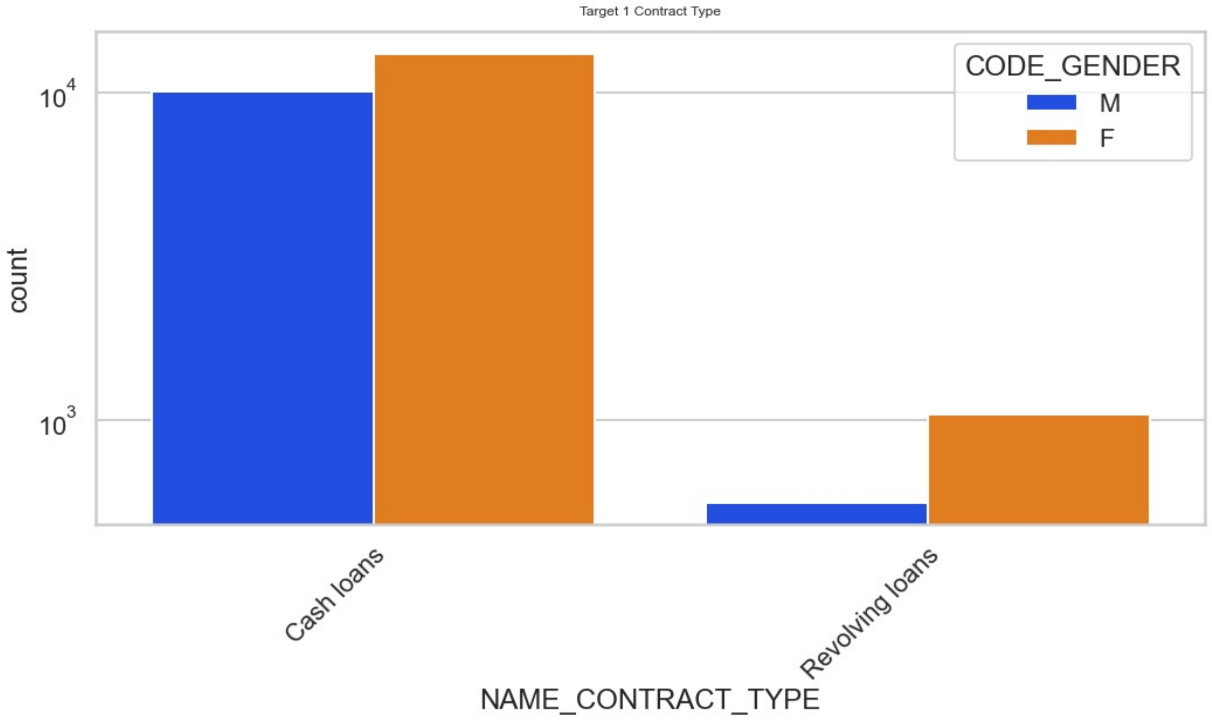


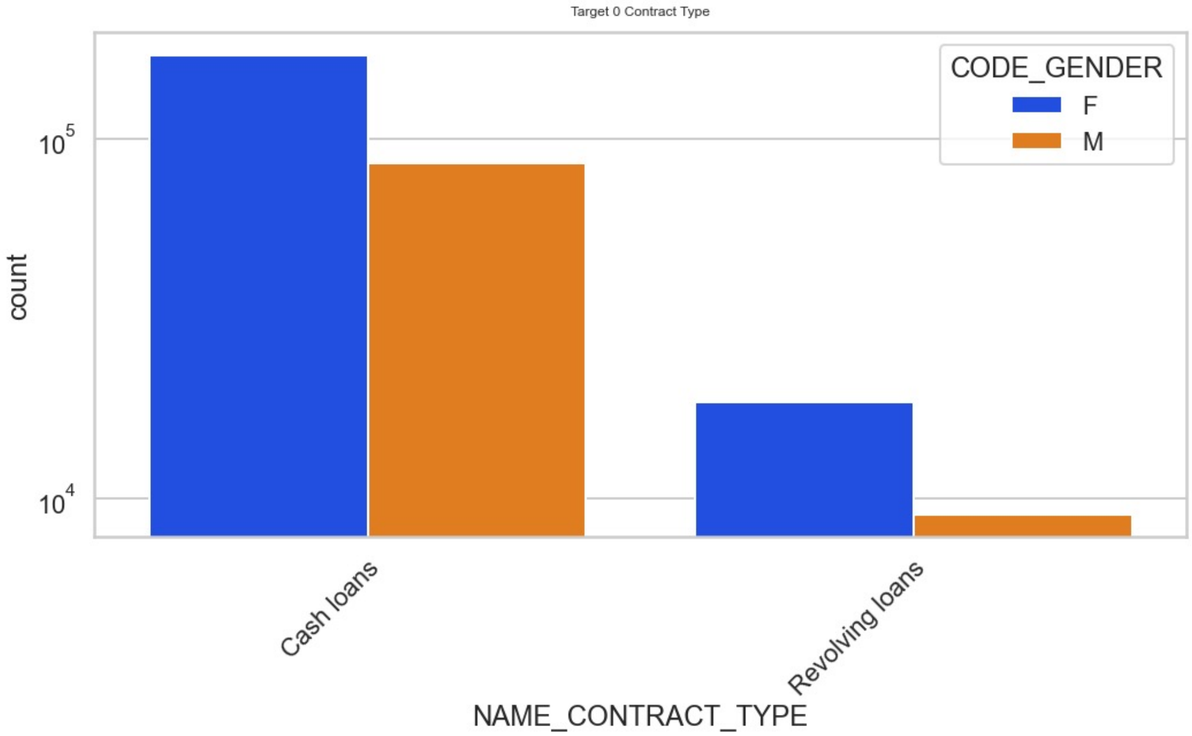
Observation:

From the above graph the following points can be concluded.

- People with Secondary education defaulted the most.

Distribution of defaulters with Contract type





<u>Observation</u>:

From the above graph for Target=1 (Defaulters) the following points can be concluded.

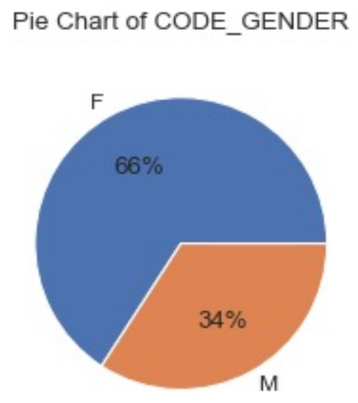
1. Individuals getting Cash loans defaulted the most than Revolving loans.
2. Female revolving loans were more compared to males.

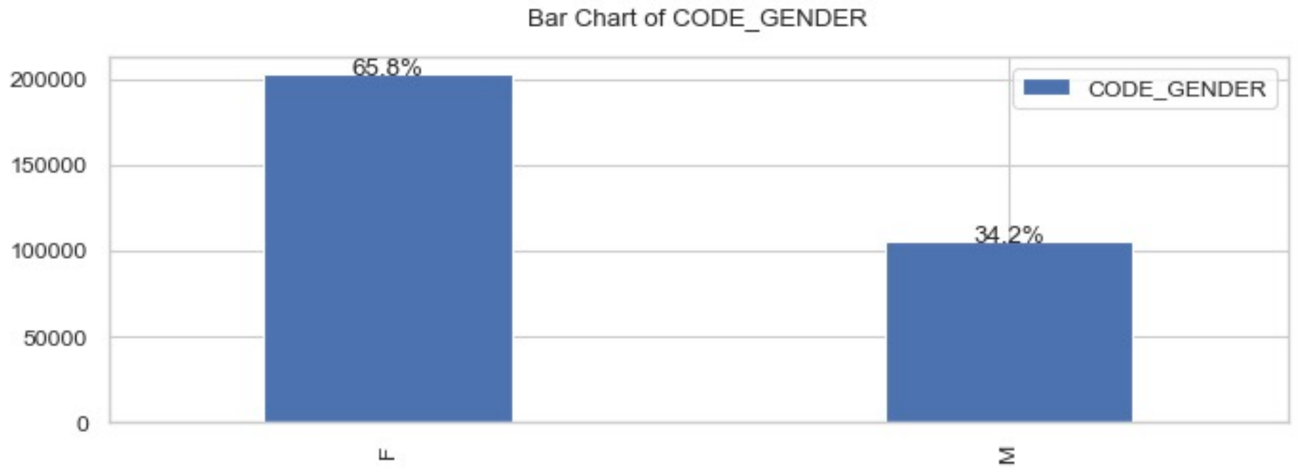
From the above graph for Target=0 (Non-Defaulters) the following points can be concluded.

1. Females defaulted less both on Cash loans and Revolving loans
2. Revolving loans were availed the least on both the scenarios

Continous Univariate Analysis:

Proportion of applicants wrt Gender

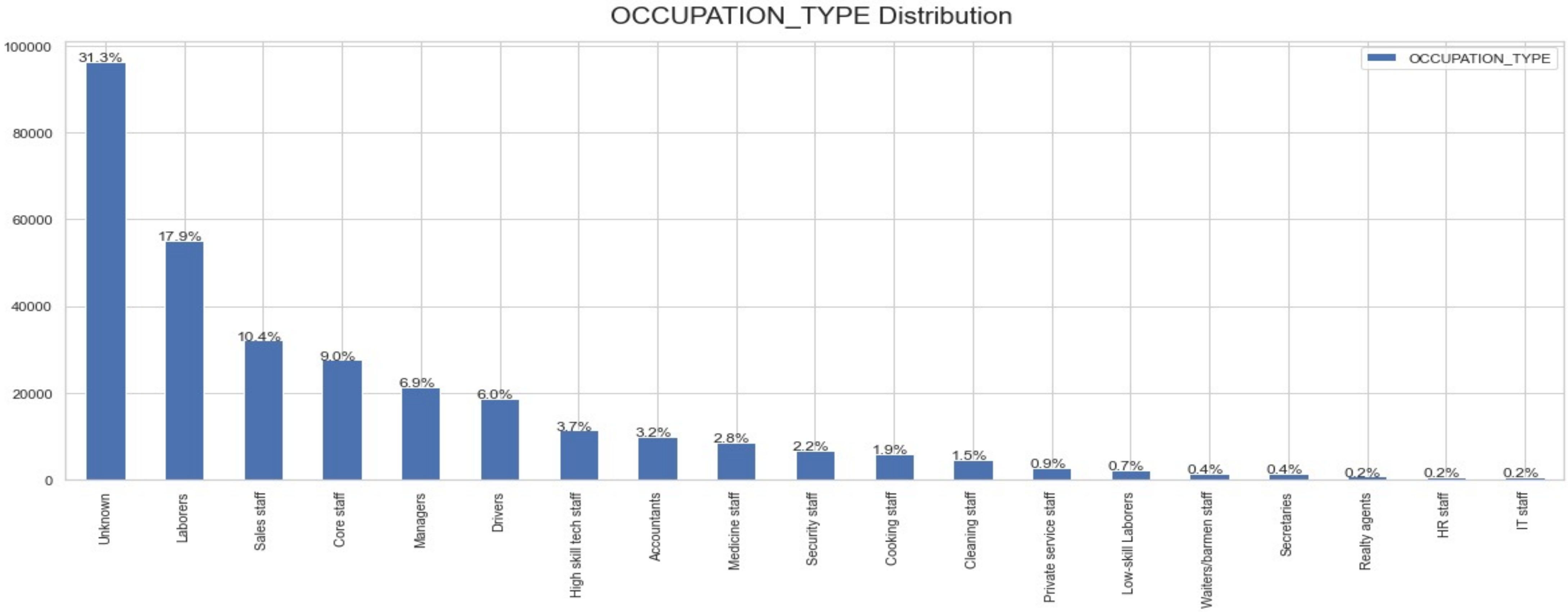




Overview:

- Female applicants are more (66%) compared to the male applicants(34.2%).

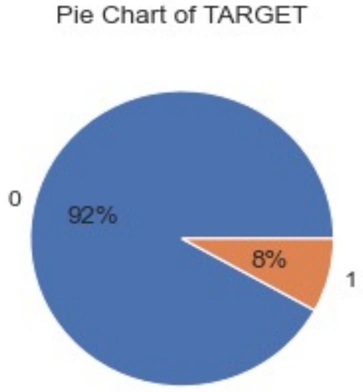
Proportion of applicants wrt Occupation

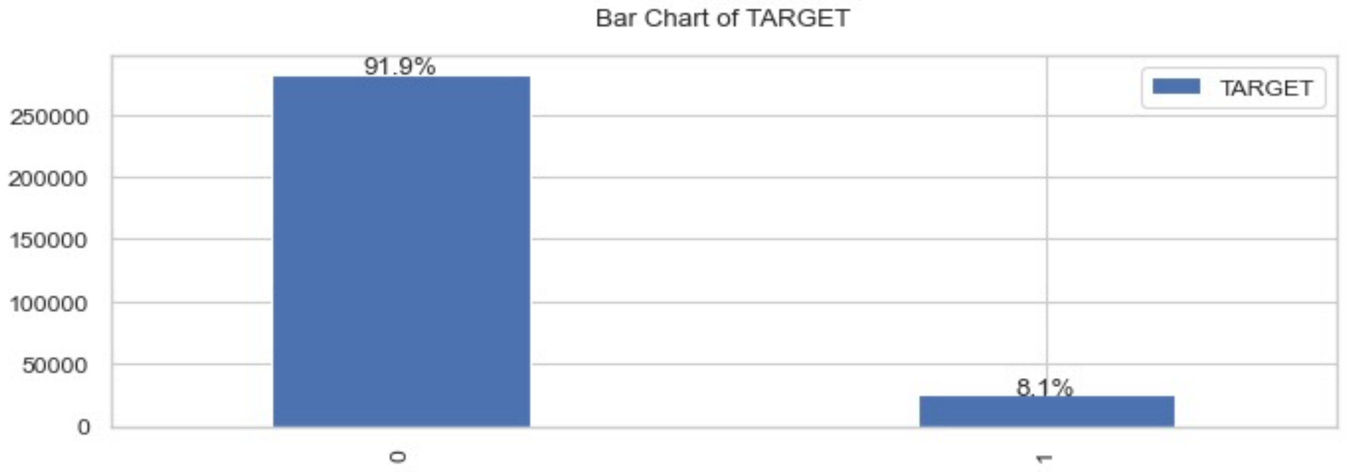


Overview:

- Most of the values in Occupation type were missing and imputed with unknown.
- Labourers applied the most and after that Sales staff, core staff and managers preceded after them.

Proportion of applicants wrt defaulters

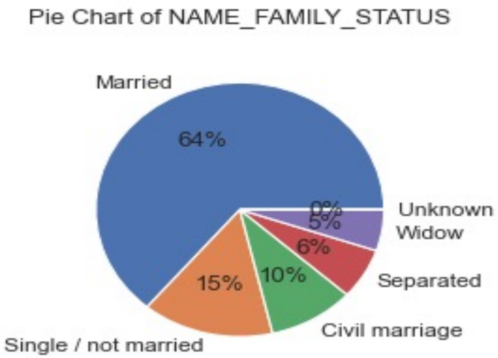


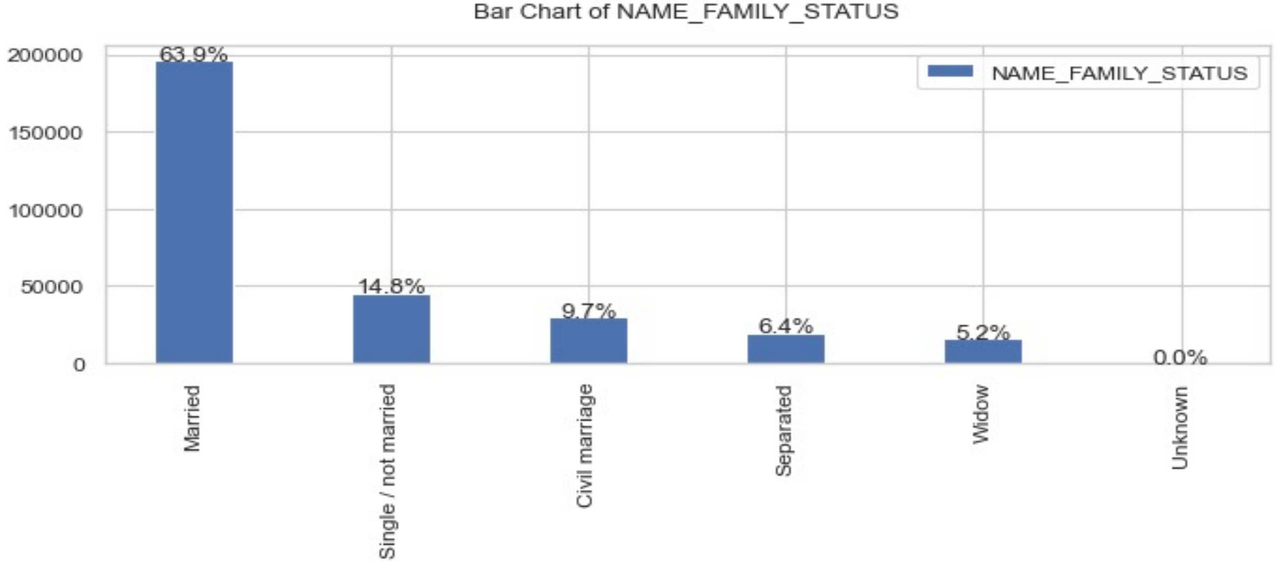


Overview:

- Proportion of people with payment difficulties were quite less.

Proportion of applicants wrt Family status

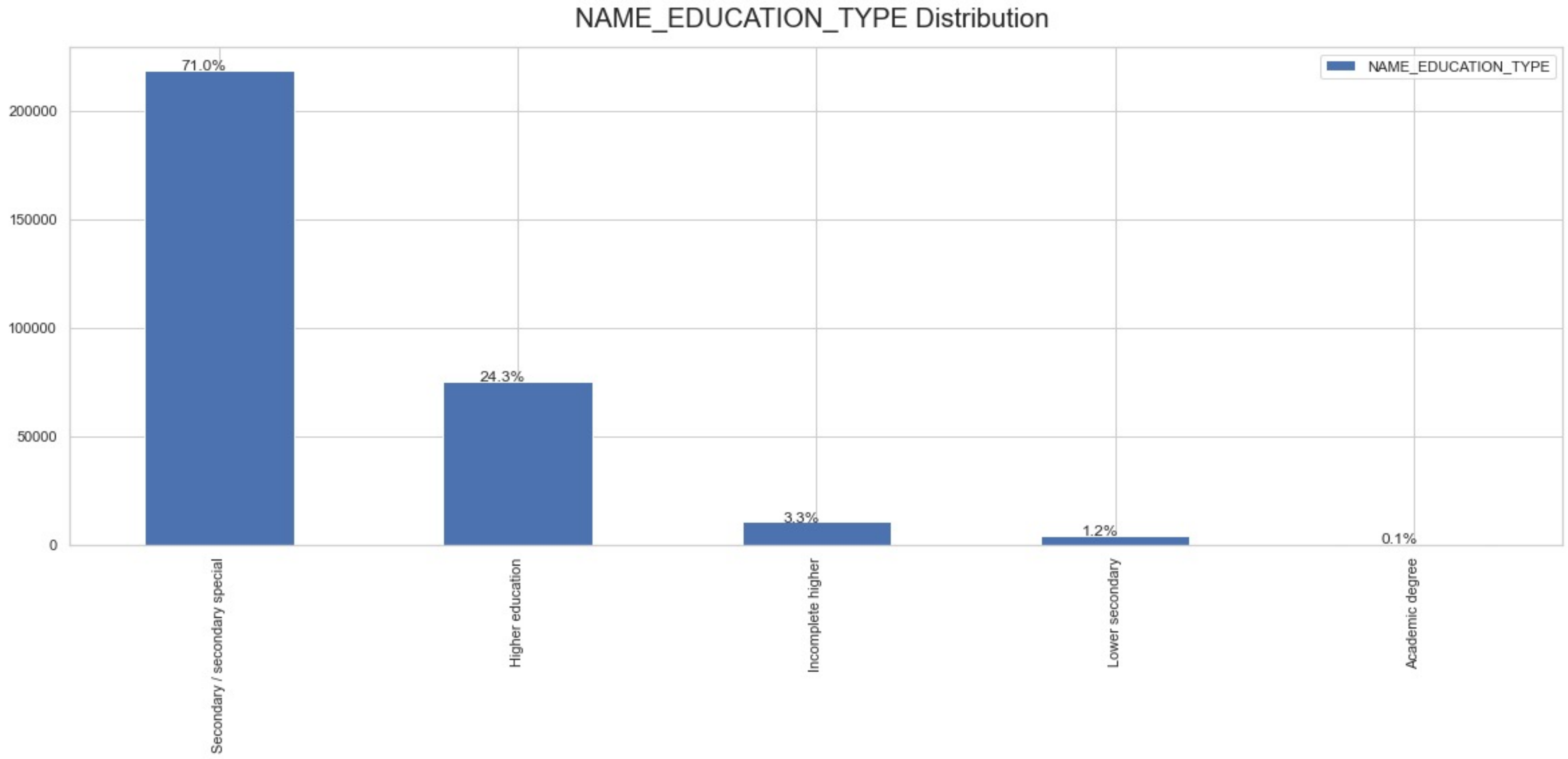




Overview:

- Married applicants (63.9%) were most followed by Single/not married(14.8%)

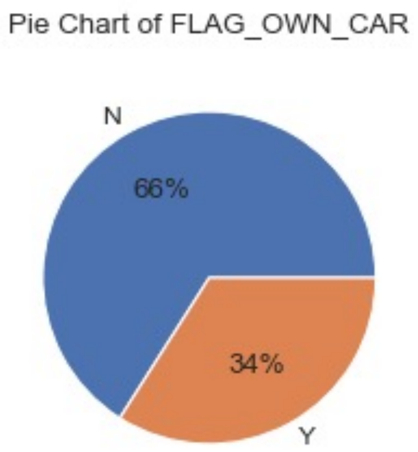
Proportion of applicants wrt Education

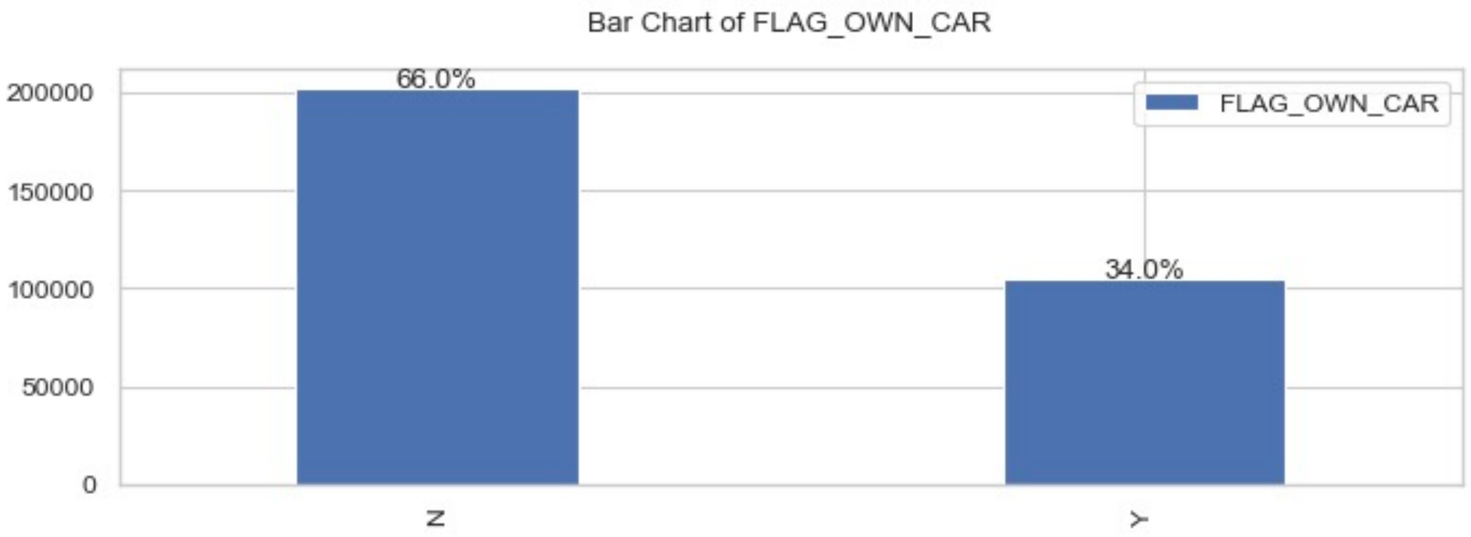


Overview:

- Most of the loan applications was for Secondary Education(71%) followed by Higher education(24.3%).

Proportion of applicants wrt Car Ownership

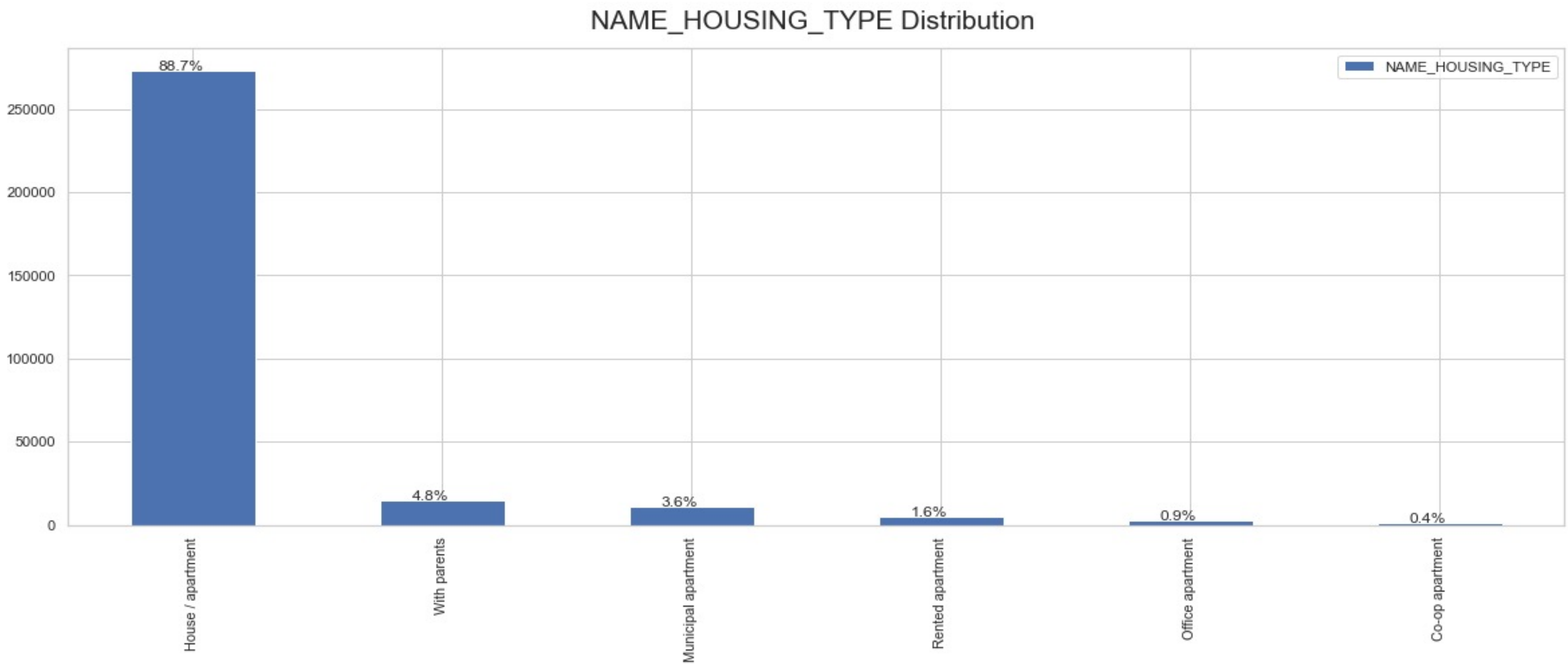




Overview:

- Applicants who doesn't own a car applied the most(66%)

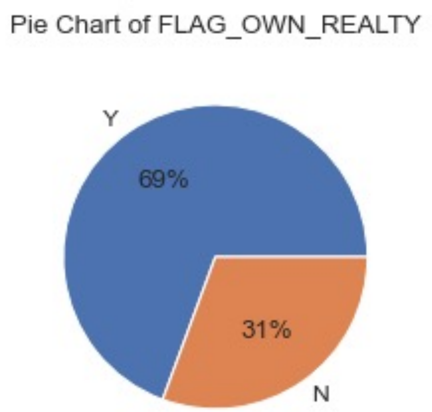
Proportion of applicants wrt Housing

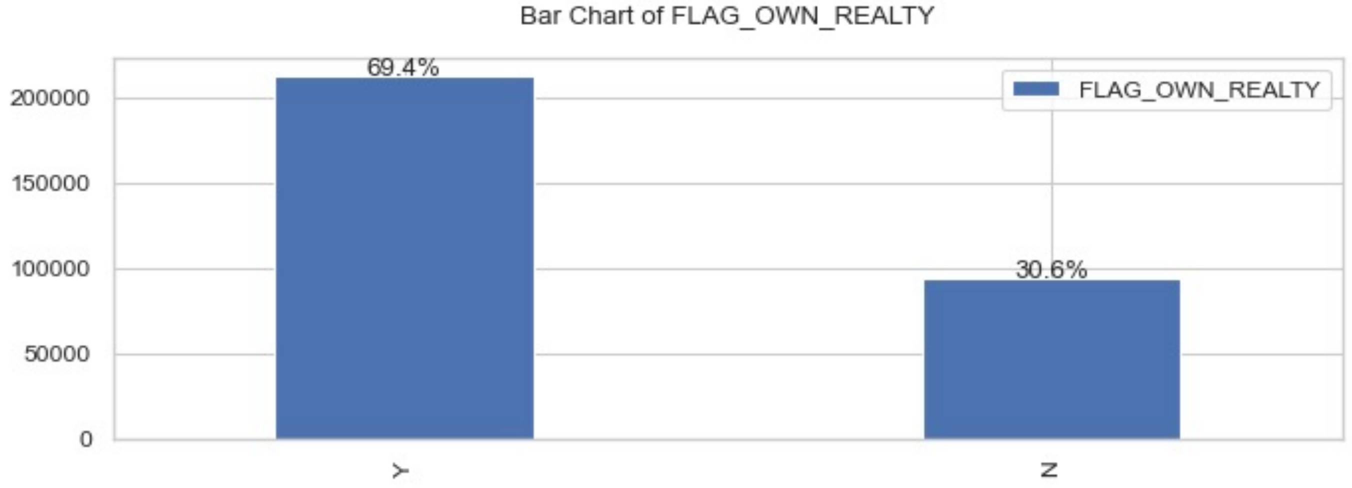


Overview:

- Individuals who owned a House/apartment applied for loan the most.

Proportion of applicants wrt Realty Ownership



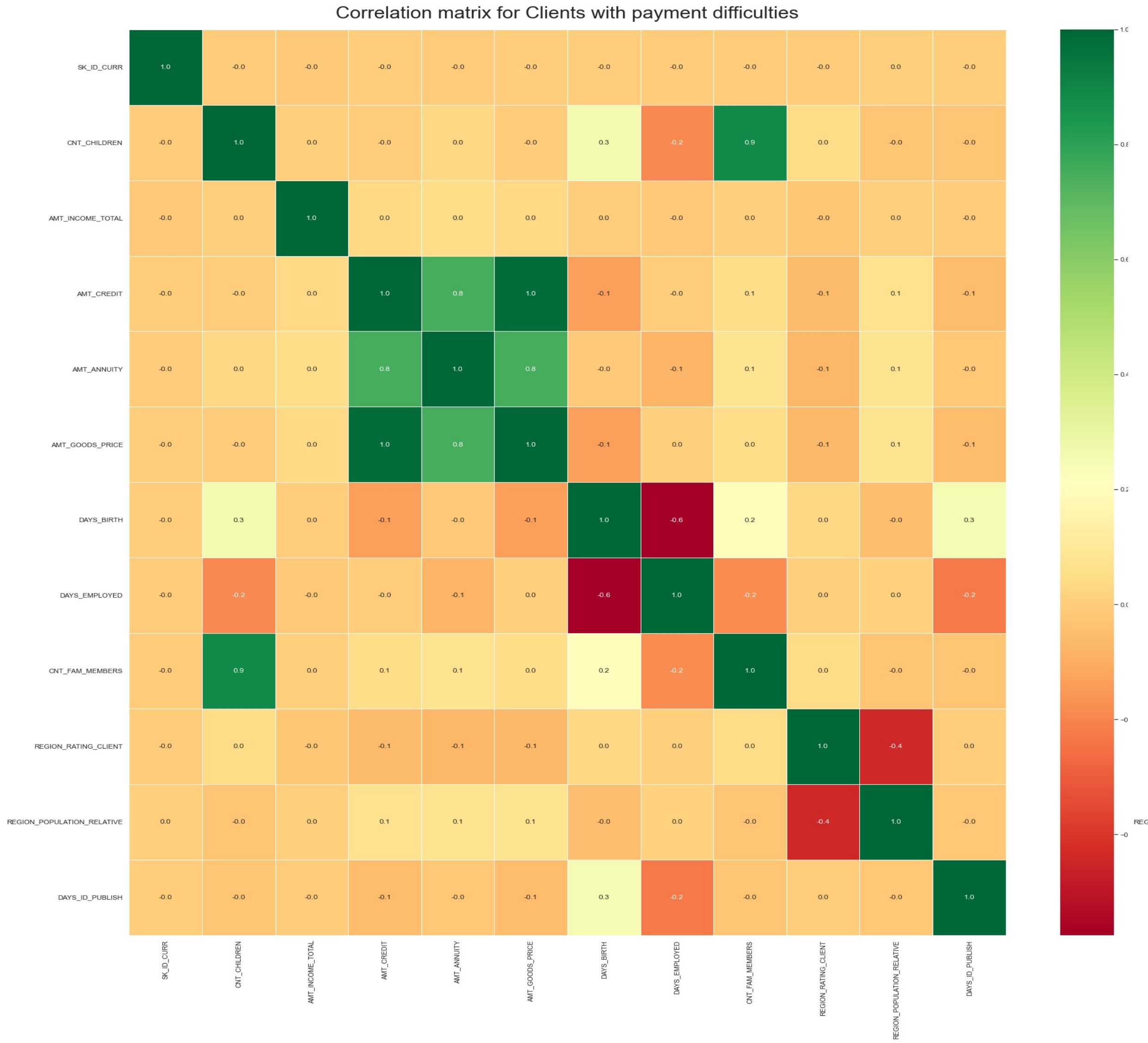


Overview:

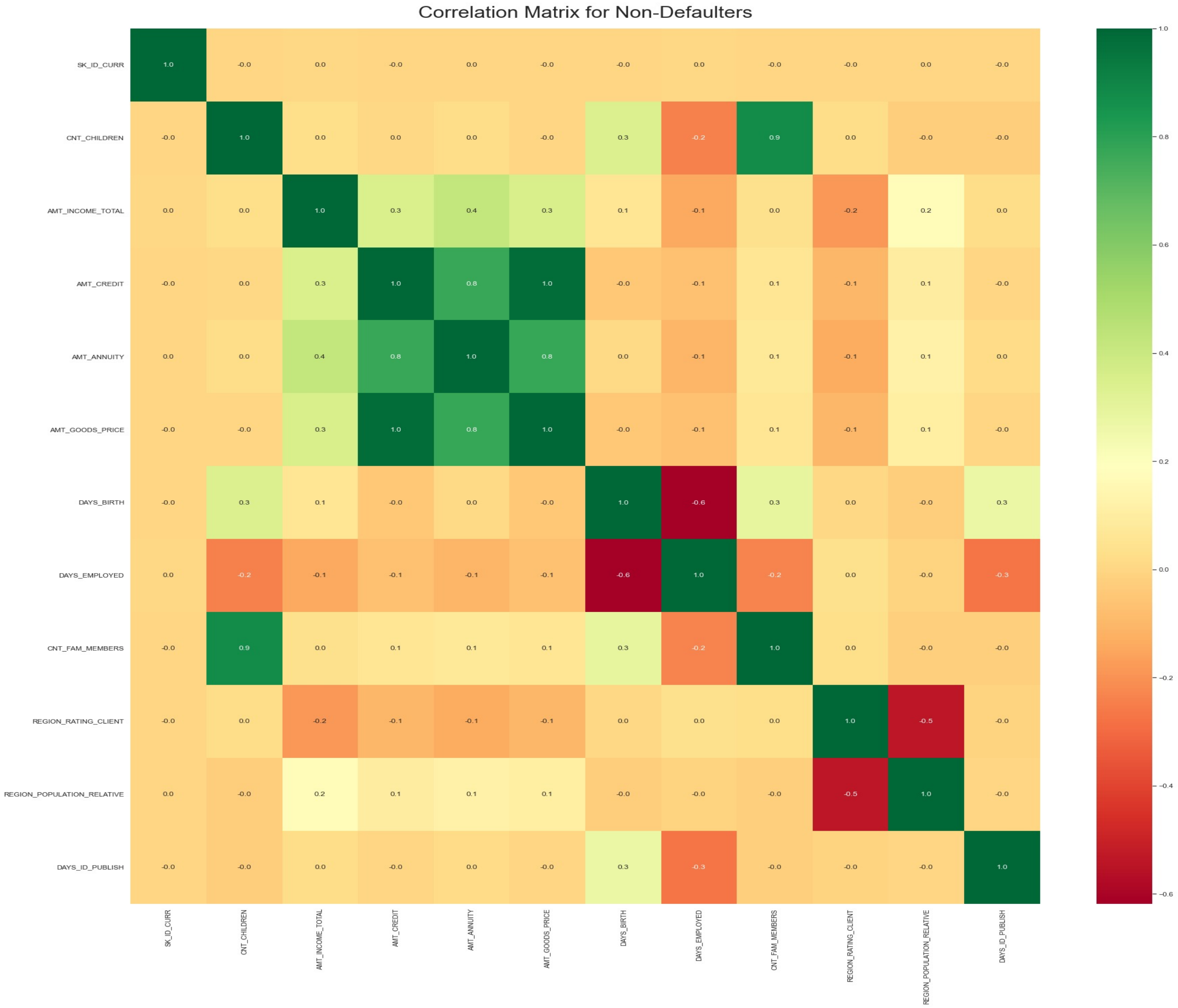
- Individuals who owned a House or flat applied for loan the most.

BIVARIATE ANALYSIS

Analysing correlation for numerical columns for both Target 0 and Target1



For Target 1



For Target 0

Observation- For defaulters and non-defaulters

Top 10 correlations for relevant columns - Payment Difficulties ¶

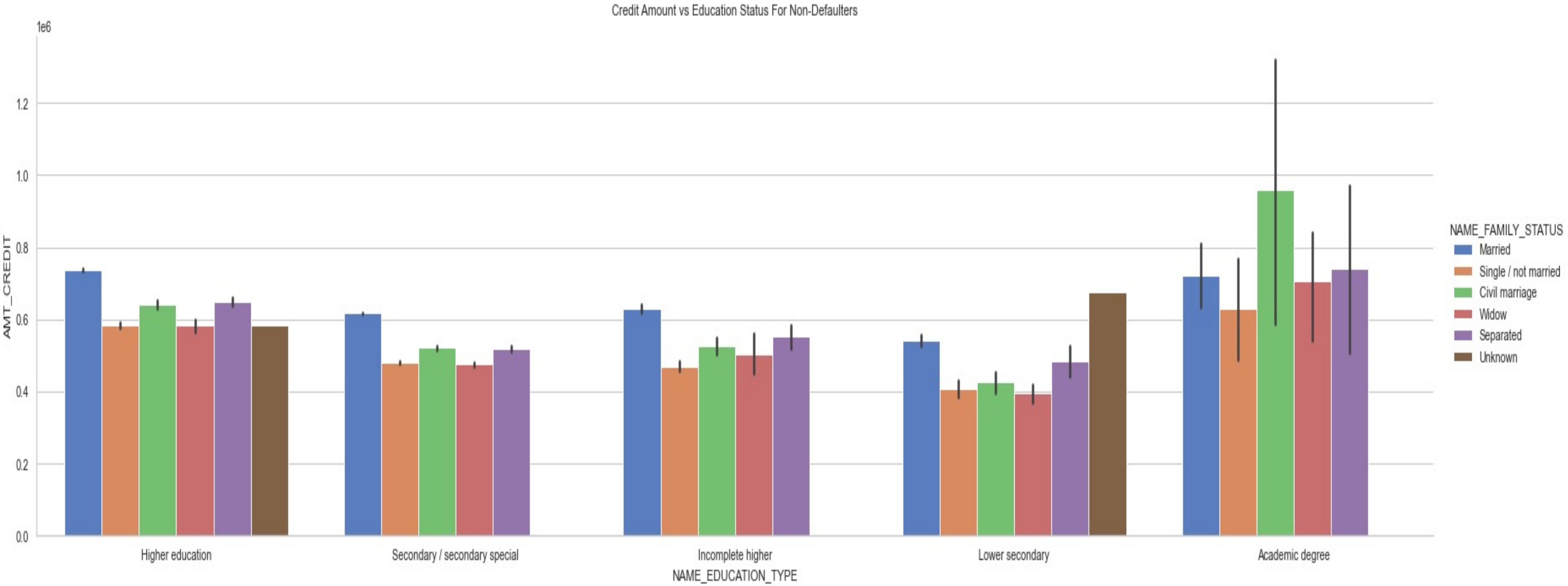
	Column 1	Column 2	Correlation
	AMT_GOODS_PRICE	AMT_CREDIT	0.982783
	CNT_CHILDREN	CNT_FAM_MEMBERS	0.885484
	AMT_GOODS_PRICE	AMT_ANNUITY	0.752295
	AMT_CREDIT	AMT_ANNUITY	0.752195
	DAYS_BIRTH	DAYS_EMPLOYED	0.575097
	REGION_POPULATION_RELATIVE	REGION_RATING_CLIENT_W_CITY	0.446977
	REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	0.443236
	CNT_CHILDREN	DAYS_BIRTH	0.259109
	EXT_SOURCE_2	REGION_RATING_CLIENT	0.250335
	EXT_SOURCE_2	REGION_RATING_CLIENT_W_CITY	0.248619

Top 10 Correlation for relevent columns - Non-defaulters

	Column 1	Column 2	Correlation
	AMT_GOODS_PRICE	AMT_CREDIT	0.987022
	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149
	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
	AMT_ANNUITY	AMT_GOODS_PRICE	0.776421
	AMT_CREDIT	AMT_ANNUITY	0.771297
	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349426
	AMT_INCOME_TOTAL	AMT_CREDIT	0.342799
	CNT_CHILDREN	DAYS_BIRTH	0.336966
	EXT_SOURCE_2	REGION_RATING_CLIENT	0.291350
	DAYS_EMPLOYED	CNT_CHILDREN	0.243356

- High correlation between 2 variants for customers with difficulty paying and the other for the price of goods and the amount of credit.
- The credit score (from ext source 2) of the applicant has a strong relationship with the region of the client. It has a slightly stronger connection to those who do not have difficulty paying (0.29) compared to those had payment problems (0.25).
- The price of the goods and the amount of revenue are closely related in all other cases compared to the payment difficulty. This could mean that those who do not have a hard time paying should check their income and the goods they want to buy in a better way.
- The region_rating_client and region_rating_client_w_city are directly proportional to each other and have higher positive correlation for customers without payment difficulties as opposed to customers with payment difficulties

Credit Amount vs Education Status For Non-Defaulters

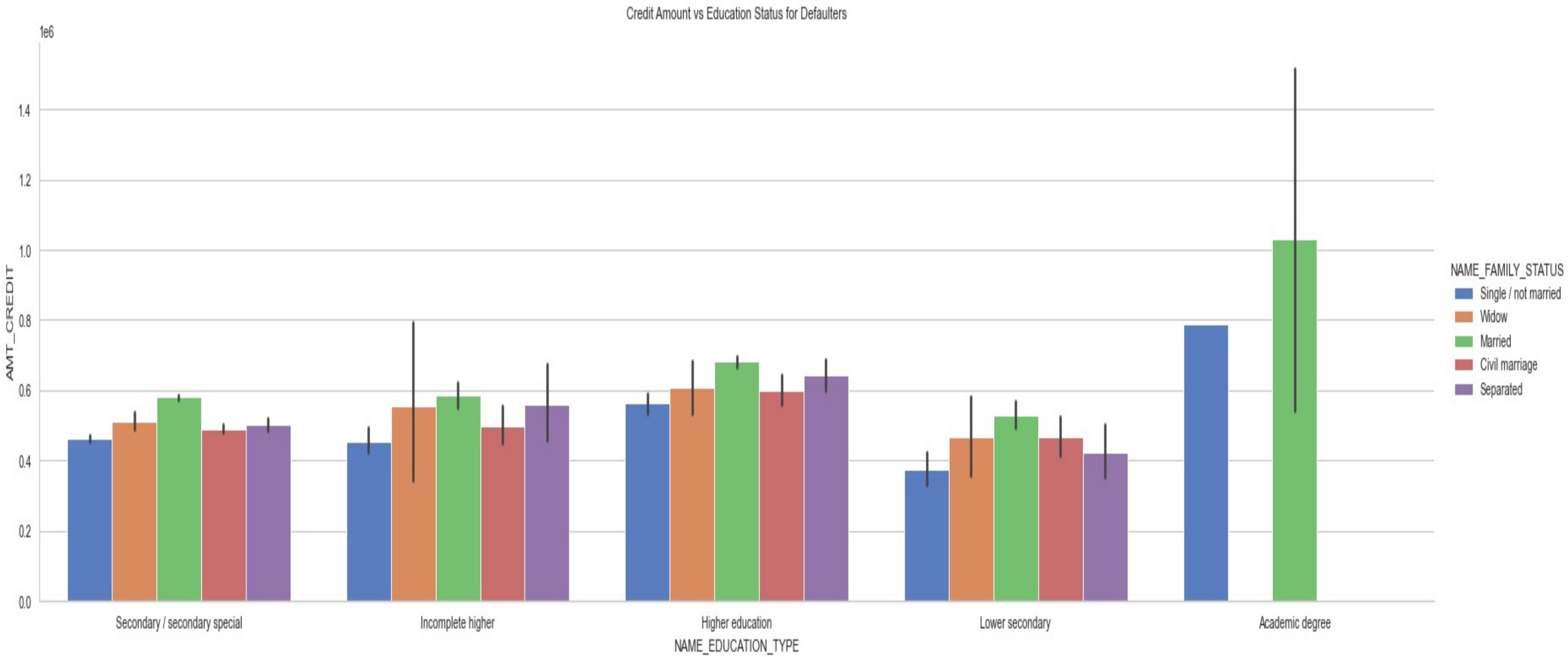


Observation

From the above graph the following can be concluded:

- Most of the loan was received for Acadmeic degree and especially Civil marriage applicants were most.
- Lower secondary individuals borrowed the least amount and widows borrowed the least among them.

Credit Amount vs Education Status For Defaulters



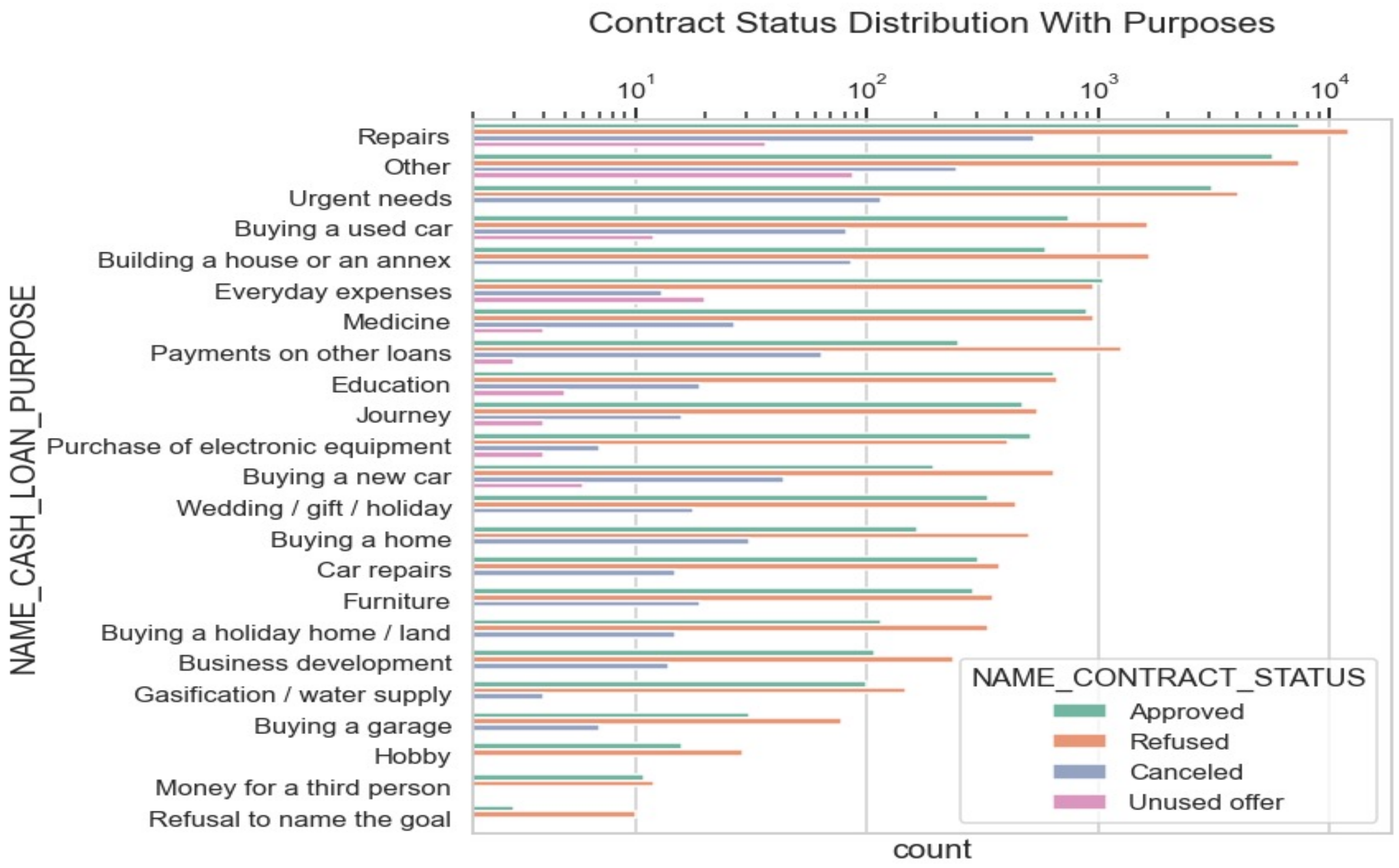
Observation

From the above graph the following can be concluded:

- Applicants with lower secondary education borrowed the least.
- In Lower secondary individuals widows borrowed the least and married defaulted the most.

Previous Application Analysis after merging data

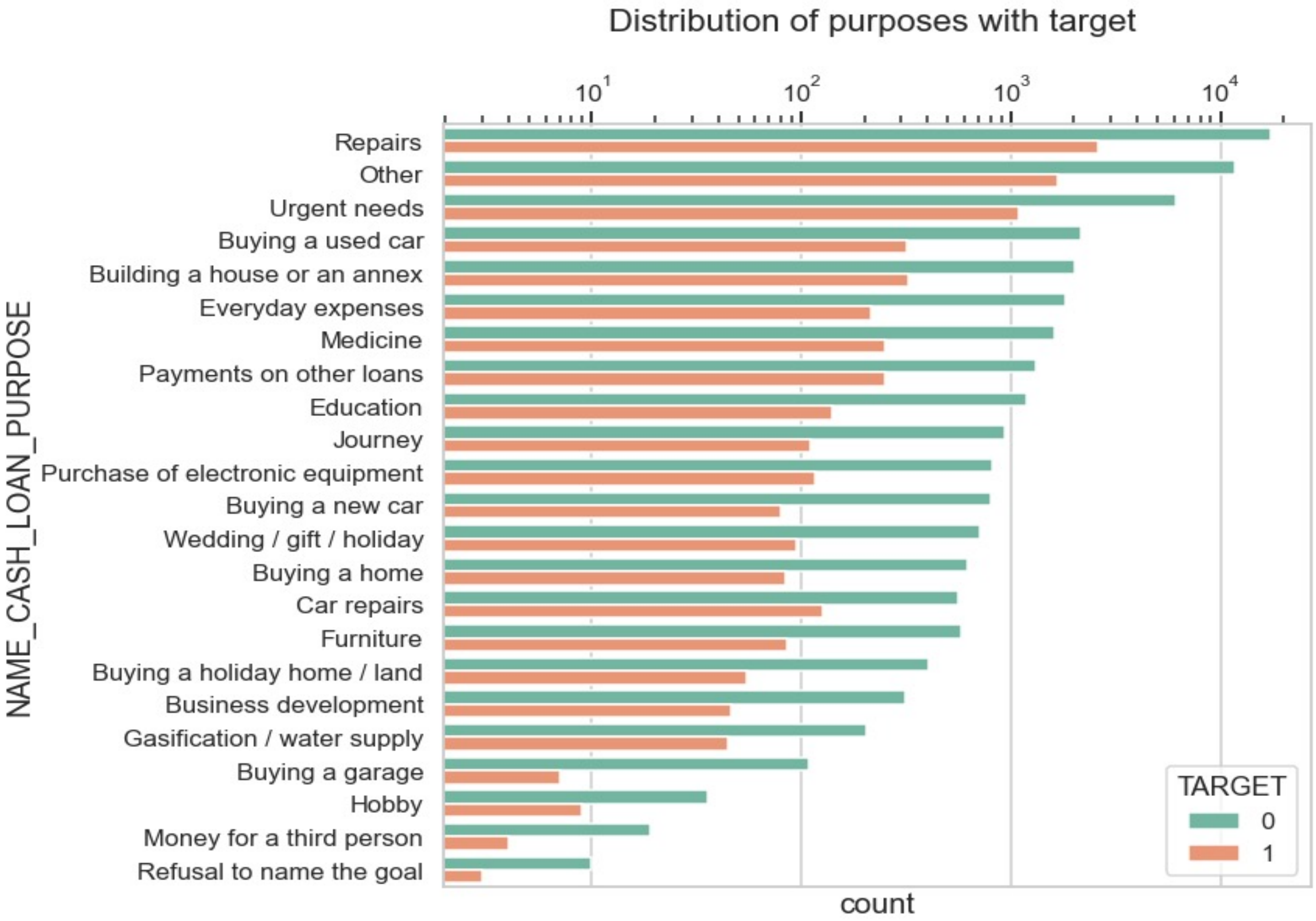
Distribution of Contract Status With Loan Purposes



Observation

- Loan refusals are more for Repairs purposes.
- Equal number of approvals and refusals for Education, Medicine and Everyday expenses

Distribution of Loan Purpose With Defaulters and Non-defaulters



Observation

- Loans on Repairs are defaulted the most

Conclude Recommendations and Risks

Conclusion after EDA on the dataset

- The bank approved loans more to females
- Proportion of defaulters is 8.1%
- Cash loans defaulted the most. Banks should consider giving more Revolving loans.
- Individuals with higher income default less.
- Banks should reconsider Repair loans as it is defaulted most.
- Applicants previously Refused / Canceled – higher default rate
- Female individuals with higher education default the least.
- Single clients default more: giving loans to married clients is safer