CREDIT EDA ASSIGNMENT

BY

RACHIN SALIM.

CREDIT EDA ASSIGNMENT

Introduction

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Understanding

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 to 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.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

Approved: The Company has approved loan Application

Cancelled: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.

Refused: The company had rejected the loan (because the client does not meet their requirements etc.).

Unused offer: Loan has been cancelled by the client but at different stages of the process.

This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

Assumptions

Since the data provided by client is huge it is natural to assume that the given data set is not clean.

We need to check and handle if present for the following before starting the analysis.

- Missing/Null Values
- Outliers
- Are the values Standardised
- Data Types
- Unwanted Rows/Columns

Methodology

Import/Load Data

We have to import/load the data set into the Python IDE. (Jupyter notebook was used for the entire EDA for this assignment)

In this EDA we have 2 Dataset:

- **application_data.csv**
- **previous_application.csv**

Data Understanding

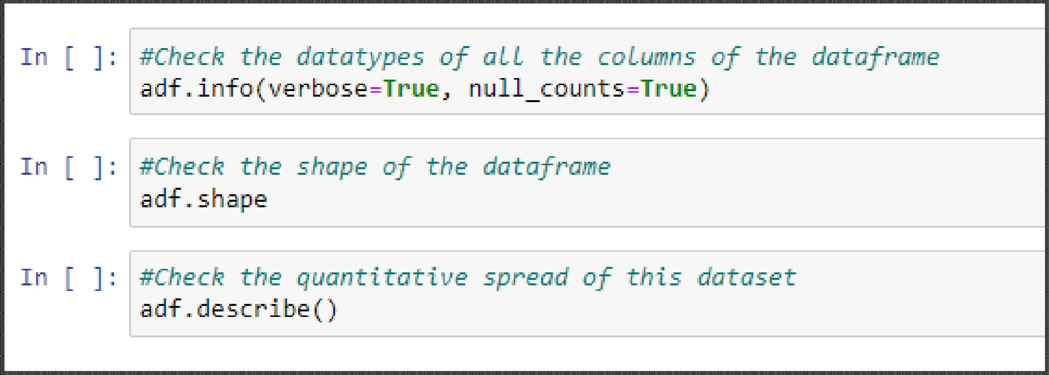
We need to understand the data prior to start of EDA. It is requires that we have basic idea about the data we are going to handle.

For that we will first take the first data:

- **application_data.csv –** This dataset contains all information of the current applications.

First step we need to do after importing the values is to find the following:

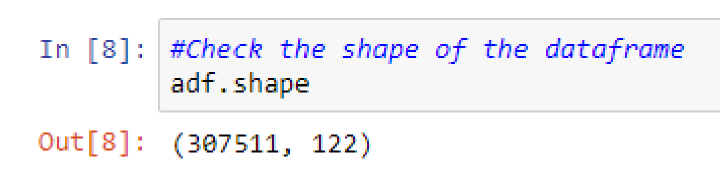
- ❖ df.info()
- ❖ df.shape
- ❖ df.describe()



```
In [ ]: #Check the datatypes of all the columns of the dataframe
        adf.info(verbose=True, null_counts=True)

In [ ]: #Check the shape of the dataframe
        adf.shape

In [ ]: #Check the quantitative spread of this dataset
        adf.describe()
```

This will give us a basic understanding of the data.

Fore example df.shape will give the number of rows and columns of the data frame.



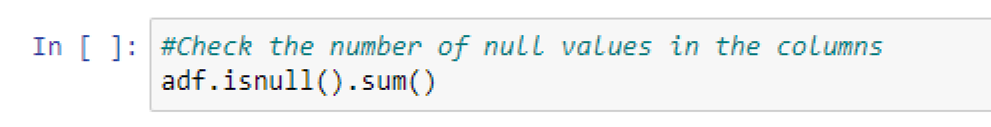
```
In [8]: #Check the shape of the dataframe
        adf.shape

Out[8]: (307511, 122)
```

Once data understanding is completed we can move to Data Cleaning.

Data Cleaning.

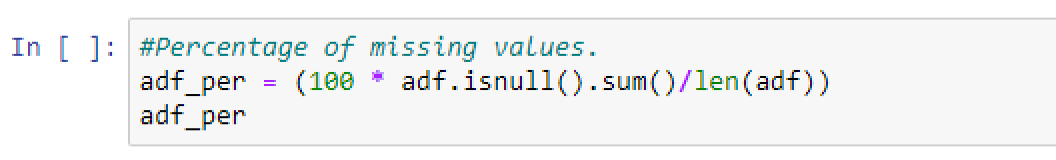
Data cleaning is done to prepare the dataset for analysis.

We need to check for the null value

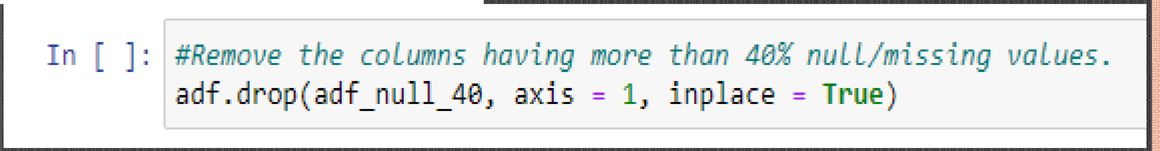


```
In [ ]: #Check the number of null values in the columns
        adf.isnull().sum()
```

Once we found out the null values we can find out the percentage of null values present in each column and drop those columns preferably more than 30-40%.

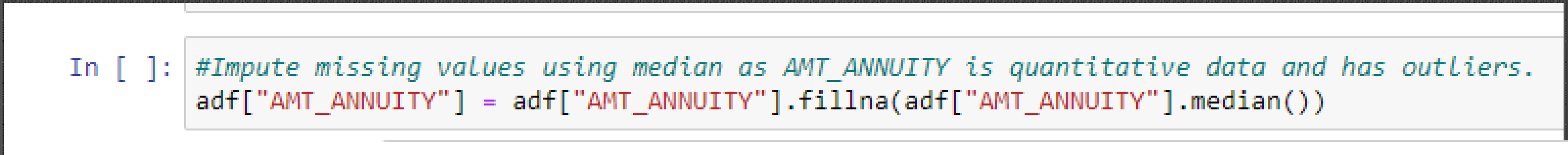


```
In [ ]: #Percentage of missing values.
        adf_per = (100 * adf.isnull().sum()/len(adf))
        adf_per
```

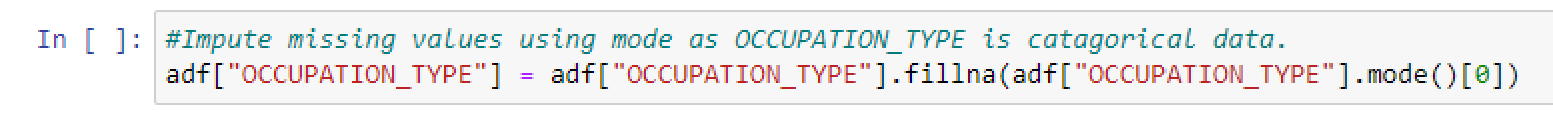


```
In [ ]: #Remove the columns having more than 40% null/missing values.
        adf.drop(adf_null_40, axis = 1, inplace = True)
```

Once the columns above 40% of null values are deleted we need to focus on the rest of the null value columns. We will impute them with mode for categorical columns and median for numerical columns.

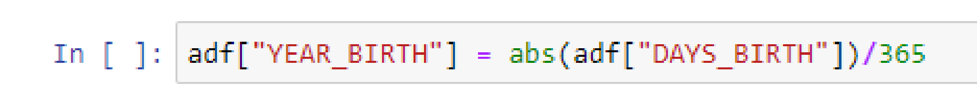


```
In [ ]: #Impute missing values using median as AMT_ANNUITY is quantitative data and has outliers.
        adf["AMT_ANNUITY"] = adf["AMT_ANNUITY"].fillna(adf["AMT_ANNUITY"].median())
```



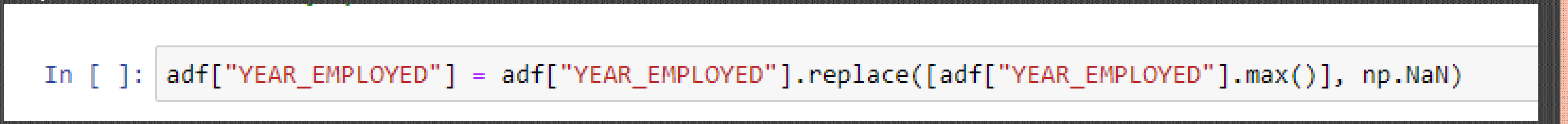
```
In [ ]: #Impute missing values using mode as OCCUPATION_TYPE is catagorical data.
        adf["OCCUPATION_TYPE"] = adf["OCCUPATION_TYPE"].fillna(adf["OCCUPATION_TYPE"].mode()[0])
```

Once we have imputed the null values we can drop the unwanted columns.
Next step is to do standardization of the values. For example here i have found out the absolute value and also converted Age in Days to Years



```
In [ ]: adf["YEAR_BIRTH"] = abs(adf["DAYS_BIRTH"])/365
```

After Standardization is complete we can check for outliers. We can either drop the outlier, impute the values using median or we can segregate the values into baskets/bins.



```
In [ ]: adf["YEAR_EMPLOYED"] = adf["YEAR_EMPLOYED"].replace([adf["YEAR_EMPLOYED"].max()], np.NaN)
```

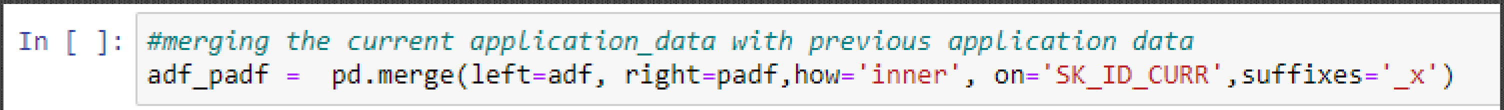
Here in this example I have replaced the Outlier value with np.NaN so it will not create error while doing any operations on the column.

We need to check for the Data types to ensure that all the values are converted to correct data types for starting the analysis.

Once we have finished this we will move on to next data set previous_application.csv. This dataset will give us all the information about previous applications.

Now we have to repeat all the steps done for application_data for previous_application.csv.

After cleaning the second dataset, I merged both the datasets.



```
In [ ]: #merging the current application_data with previous application data
        adf_padf =  pd.merge(left=adf, right=padf,how='inner', on='SK_ID_CURR',suffixes='_x')
```

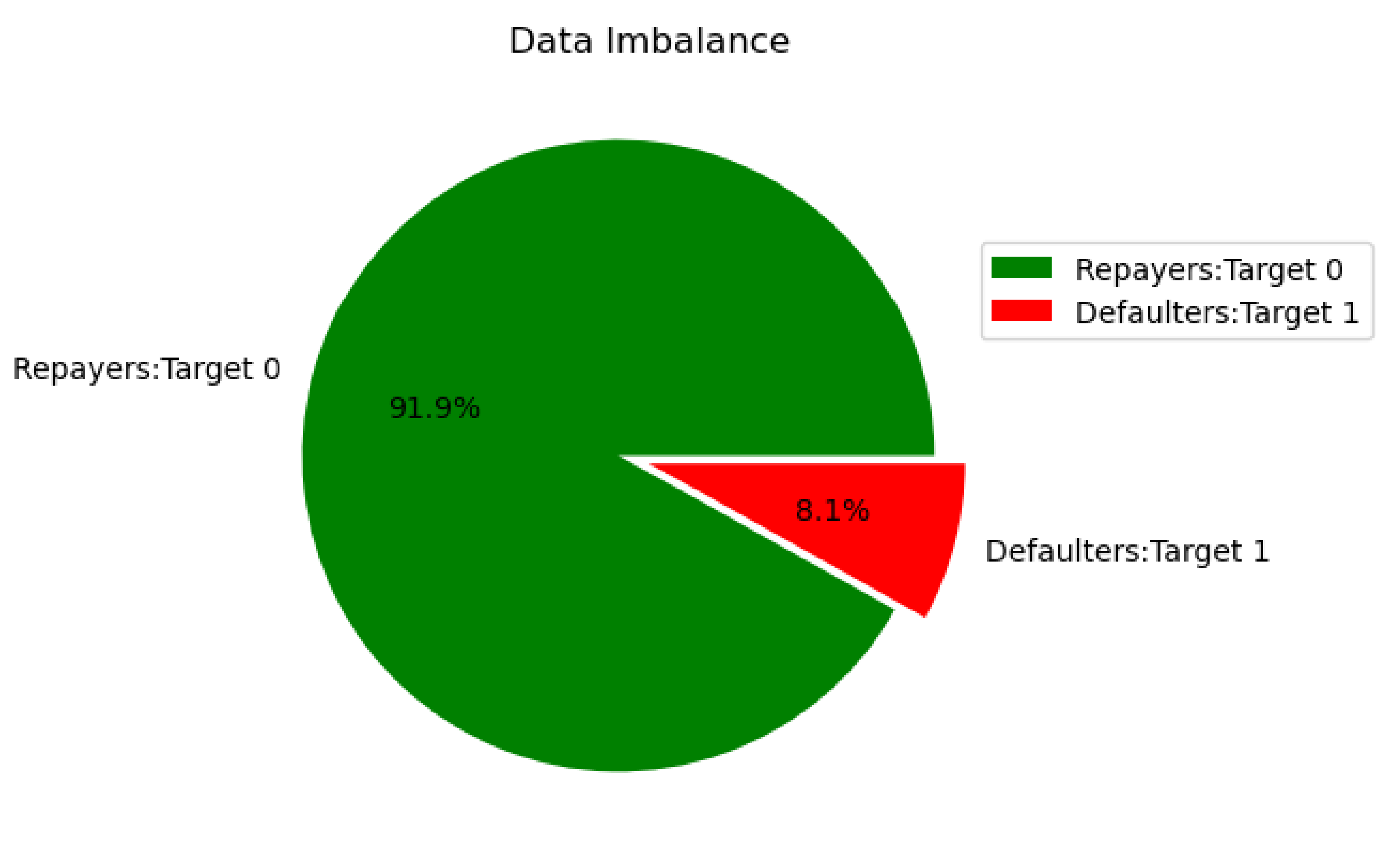
Once the new dataset is created I once more carried out all checks in data cleaning on the new dataset to confirm its readiness for analysis.

Once the data set is clean we move on to analysis.

Current and previous Application Data Analysis

Data Imbalance

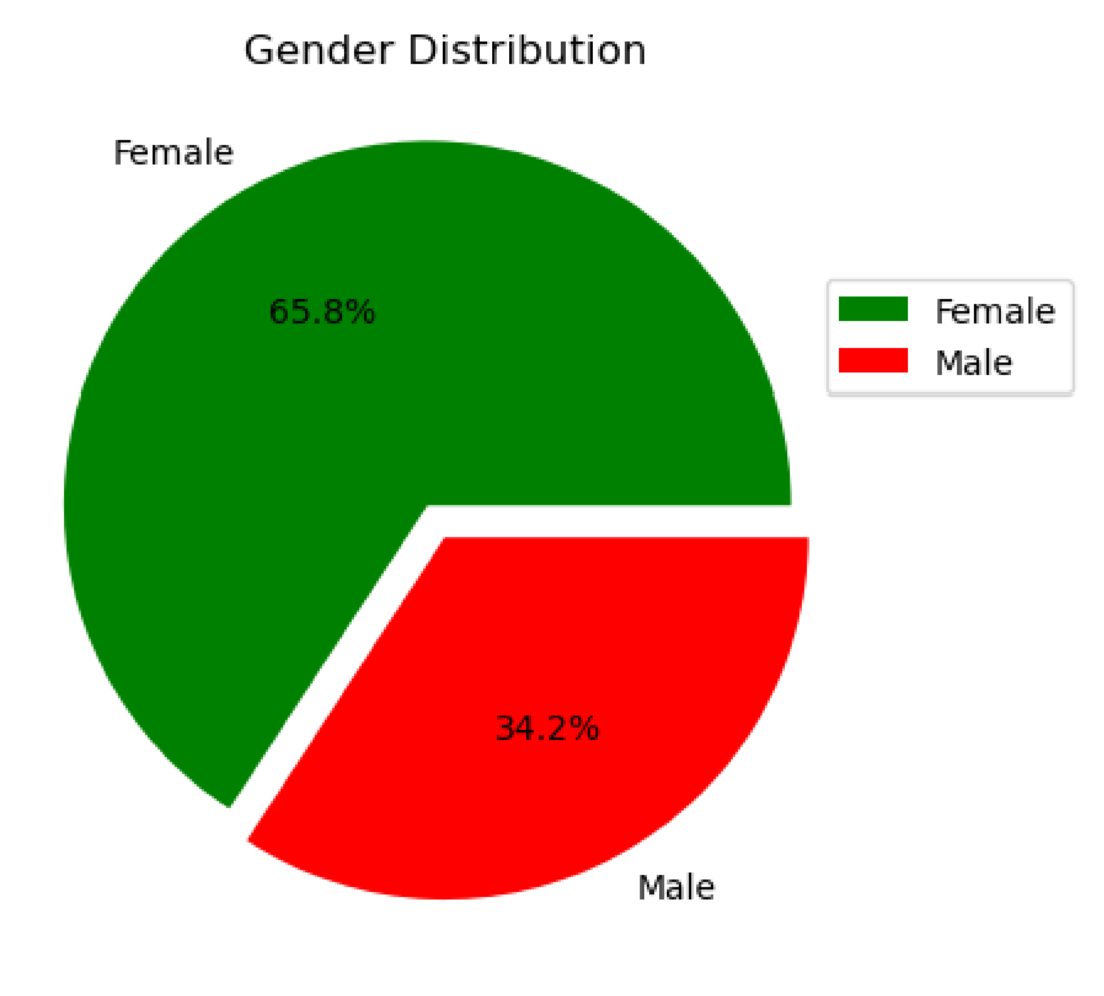
- Number of Clients with no payment difficulties are very high compared to Defaulters
- There are 8.07% clients that have payment difficulties (Defaulters) and 91.9% are having no difficulties (Repayers).



Gender Count

- Number of Female Clients are almost double compared to Male Clients.

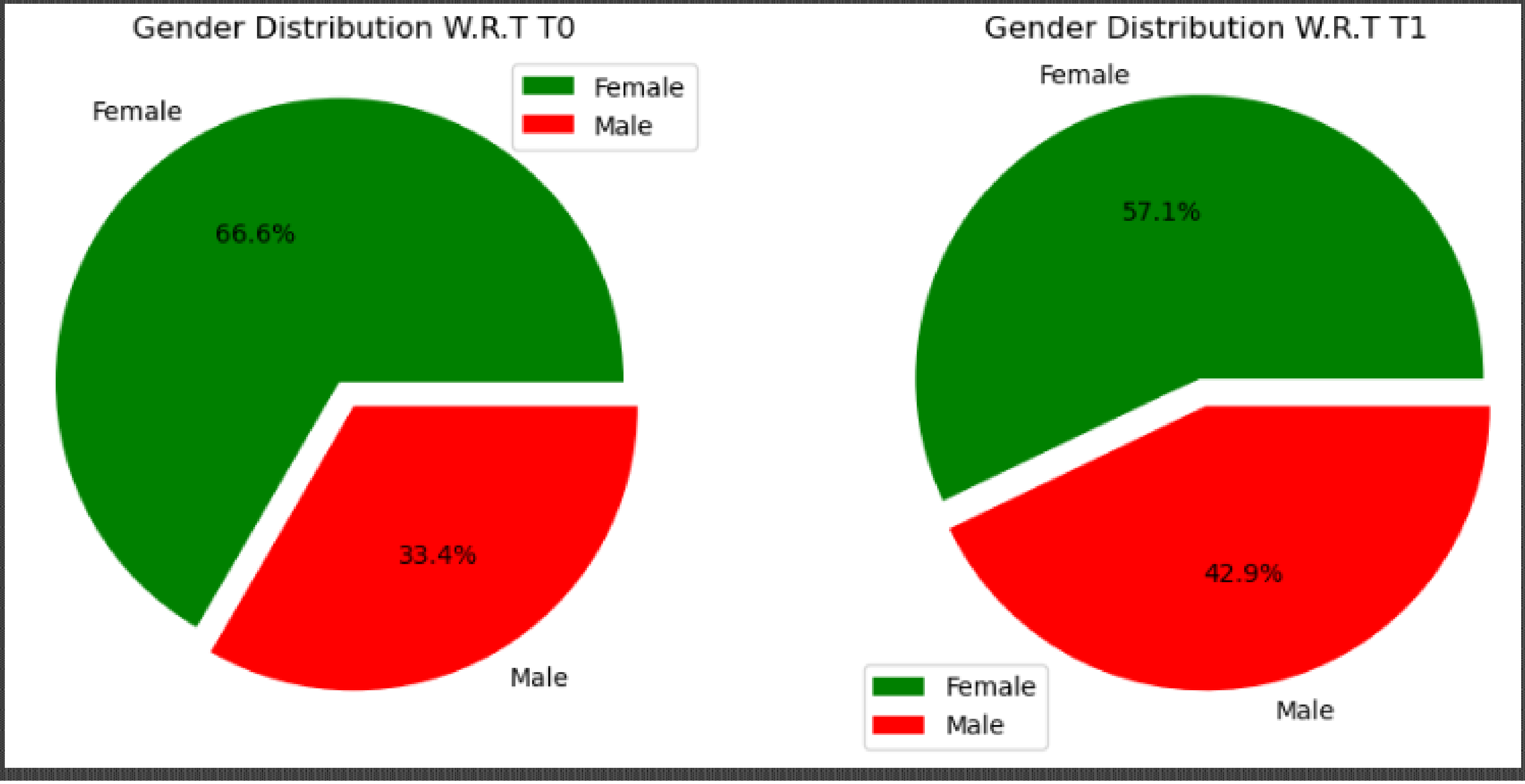
- There are 65.8% Female Clients compare to 34.2% Male Clients.



Gender Count With Respect To Target Variables.

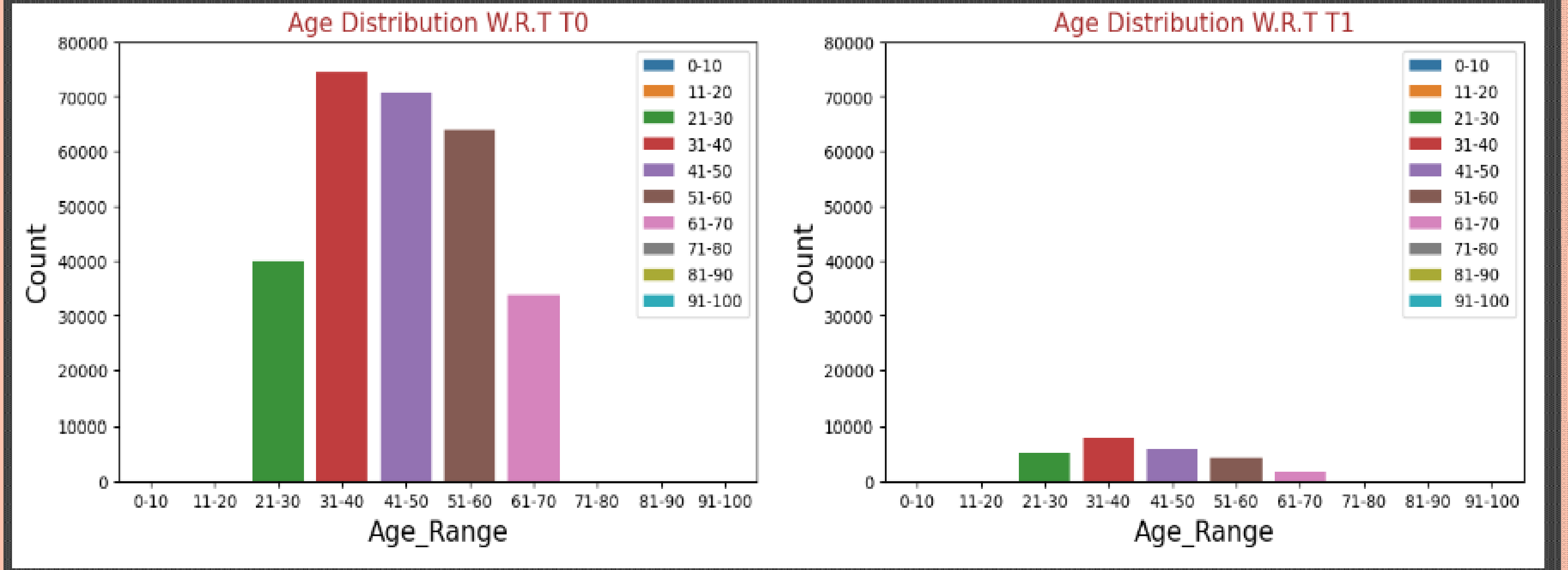
➢66.6% Female clients are repayers whereas 33.4% male clients repayers.

➢57.1% Female clients are defaulters while 42.9% male clients are defaulters.



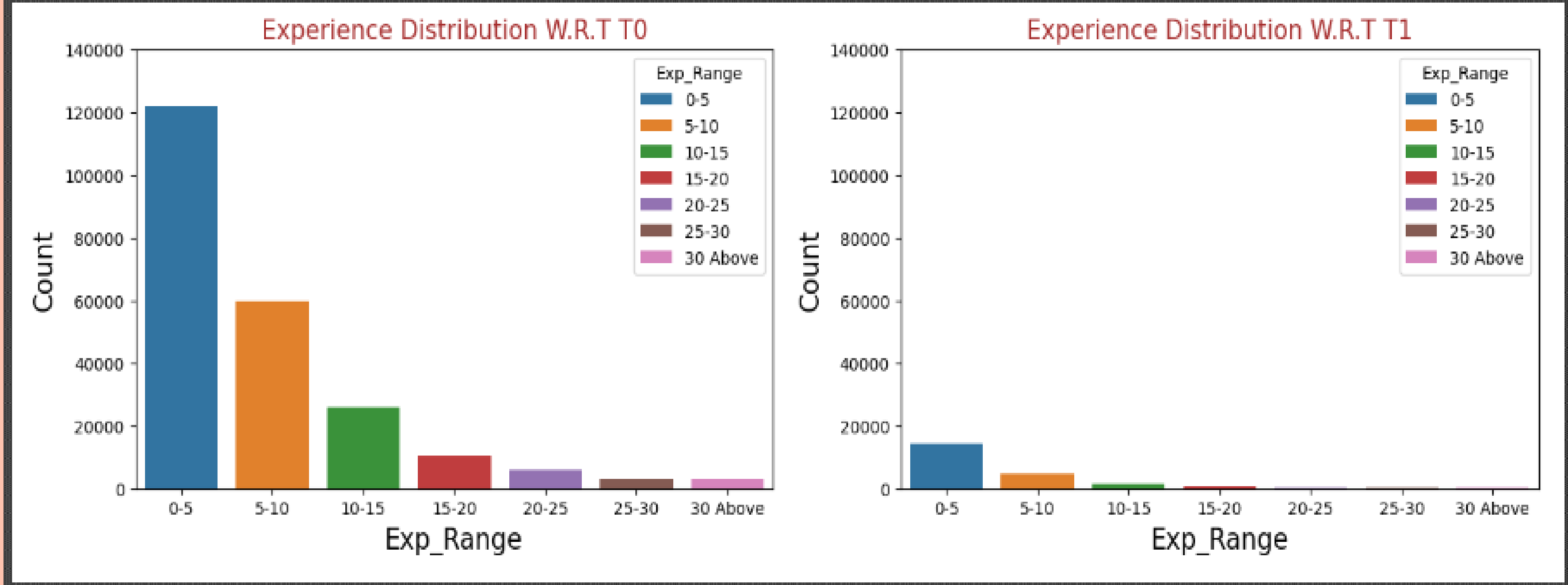
UNIVARIATE ANALYSIS

Gender Count With Respect To Target Variables.



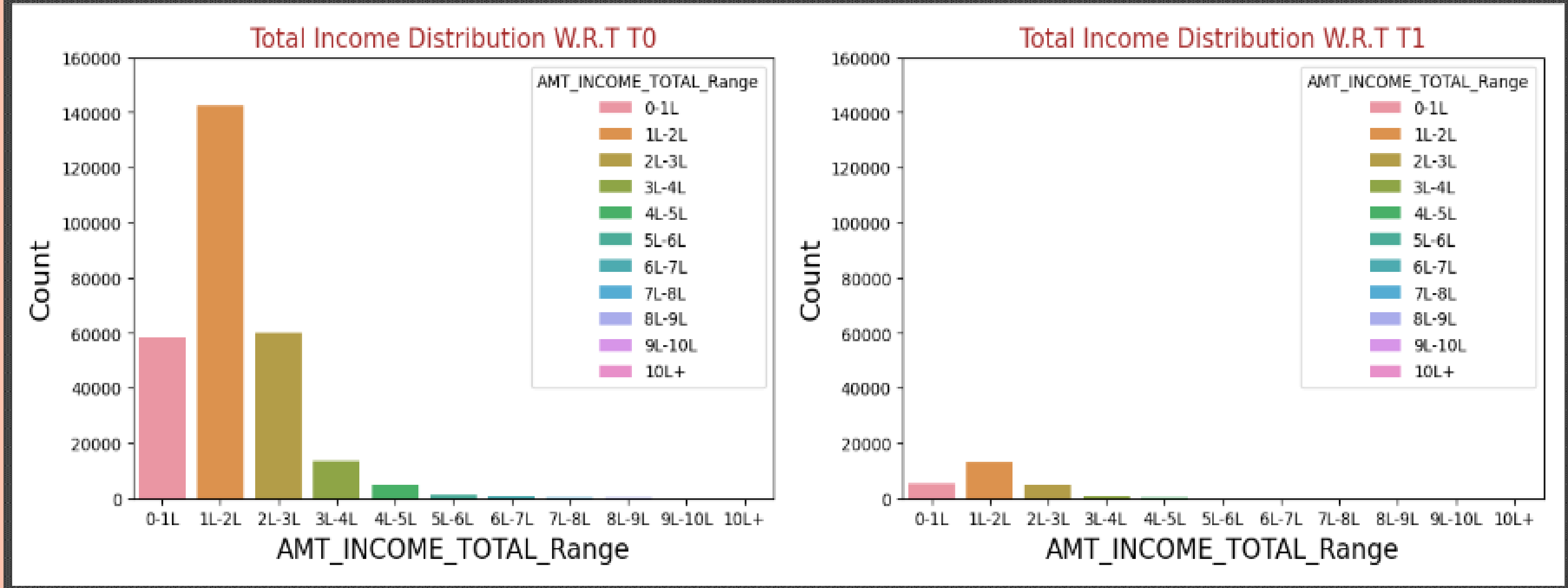
- ➢Customers without payment difficulties are in the age range of 30-40 years followed by 40-50 years
- ➢Customers below 30 years are more likely to default.

Experience Distribution With Respect To Target Variables.



- Customers having less than 5 years of experience are mostly having no payment difficulties and customers are less likely to take loans as their experience increases.

Total Income With Respect To Target Variables.

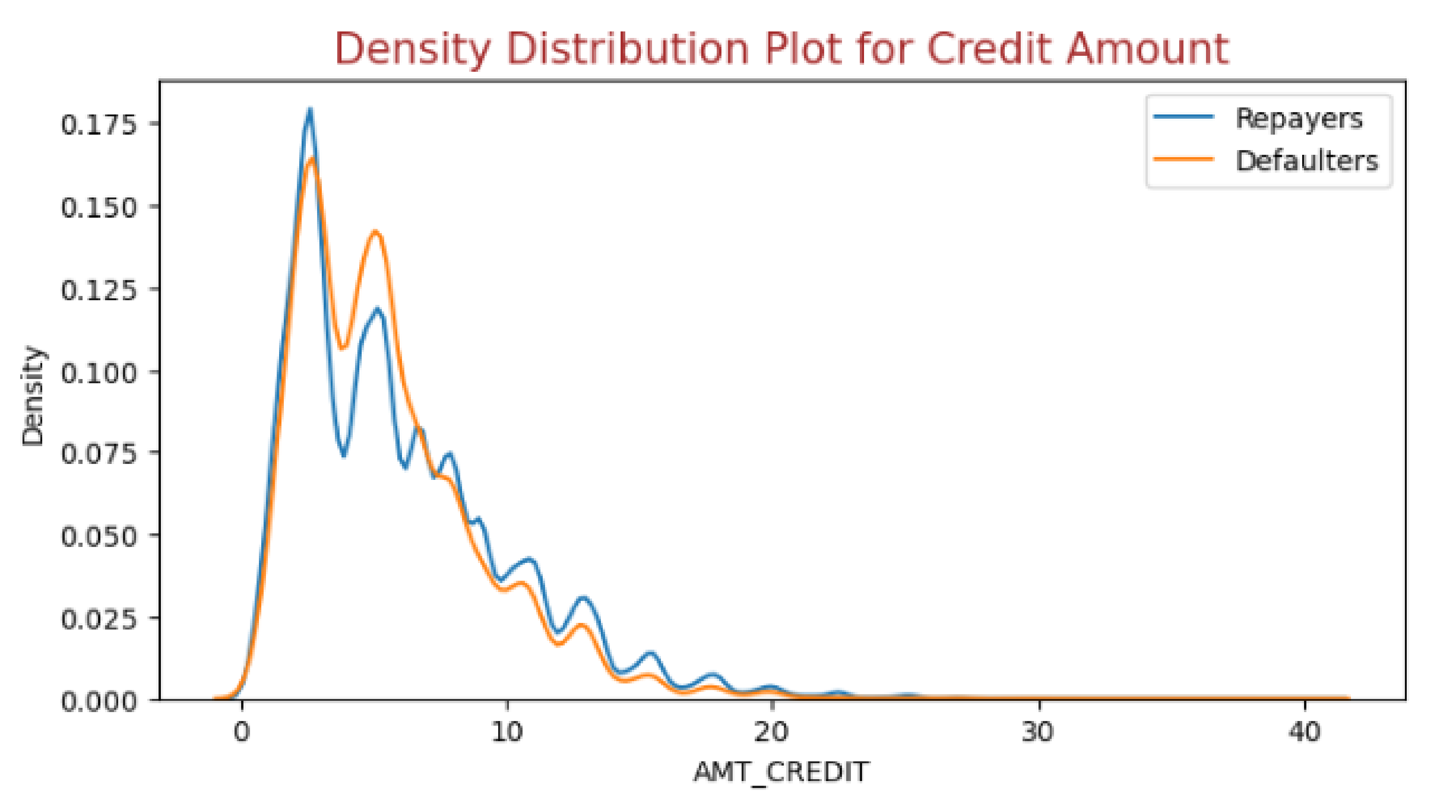


➢Customers without payment difficulties are having a salary range of 1L - 2L.

➢Customers below 30 years are more likely to default.

Density Distribution Plot for Credit Amount.

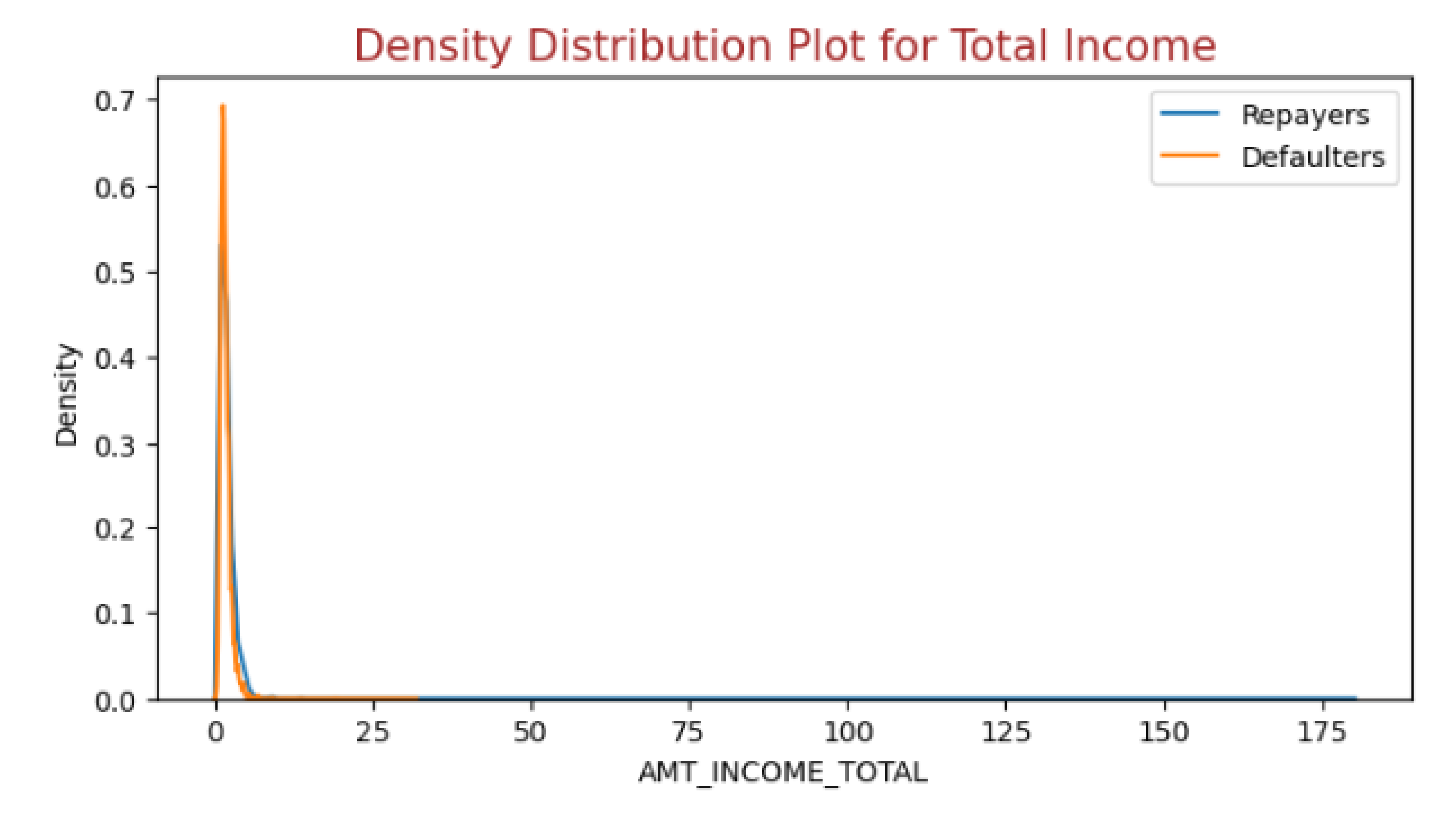
- Credit amount for maximum number of loans distributed is under 10L.
- As the credit amount increases the density increases and after reaching the peak value it starts to decrease.
- Both Repayers and Defaulters shows almost similar pattern in density distribution.



Density Distribution Plot for Total Income.

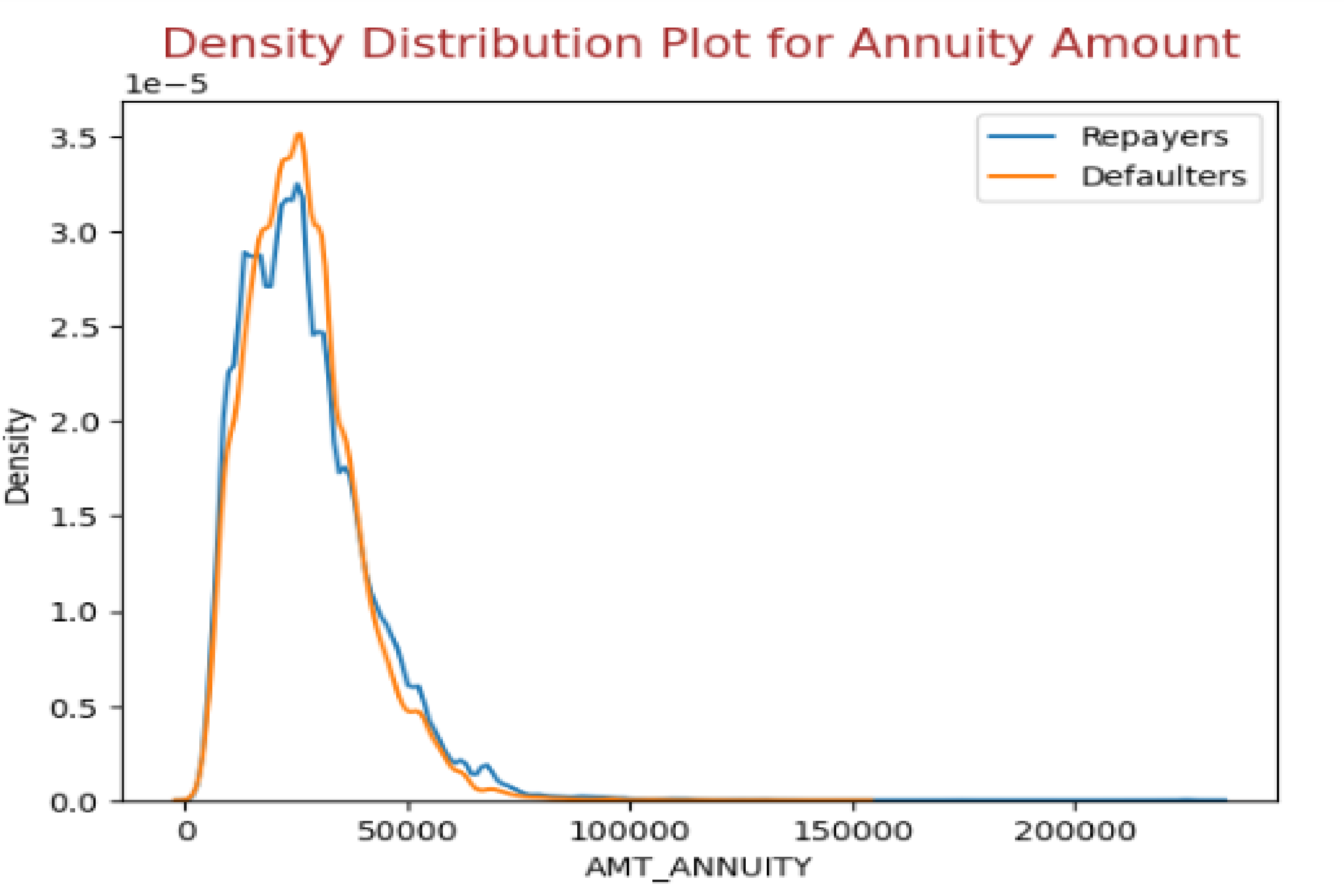
- Maximum Number of customers are having a total income of under 4 Lakhs

- Both Repayers and Defaulters shoe a similar pattern in Density Distribution.

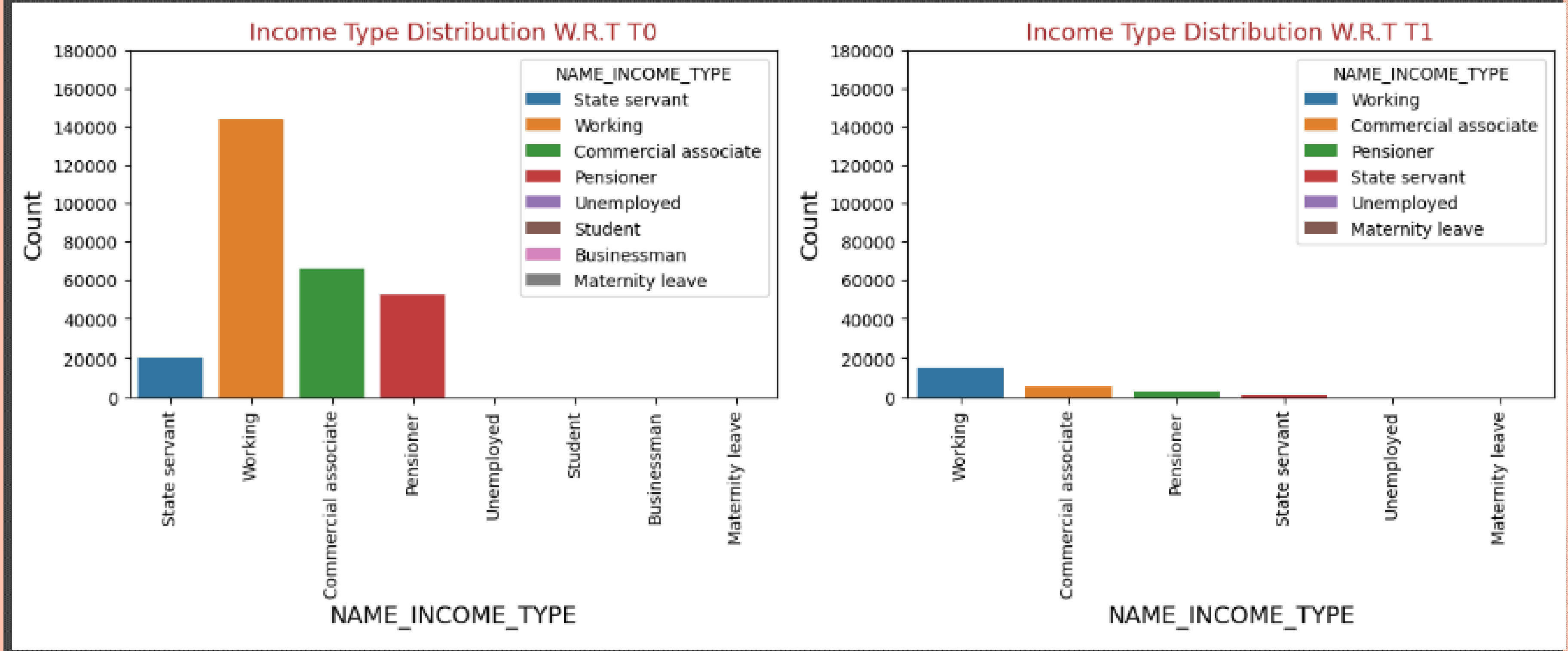


Density Distribution Plot for Annuity Amount.

- Maximum Number of customers are having a Annuity amount of less than 50000.
- Both Repayers and Defaulters shoe a similar pattern in Density Distribution.

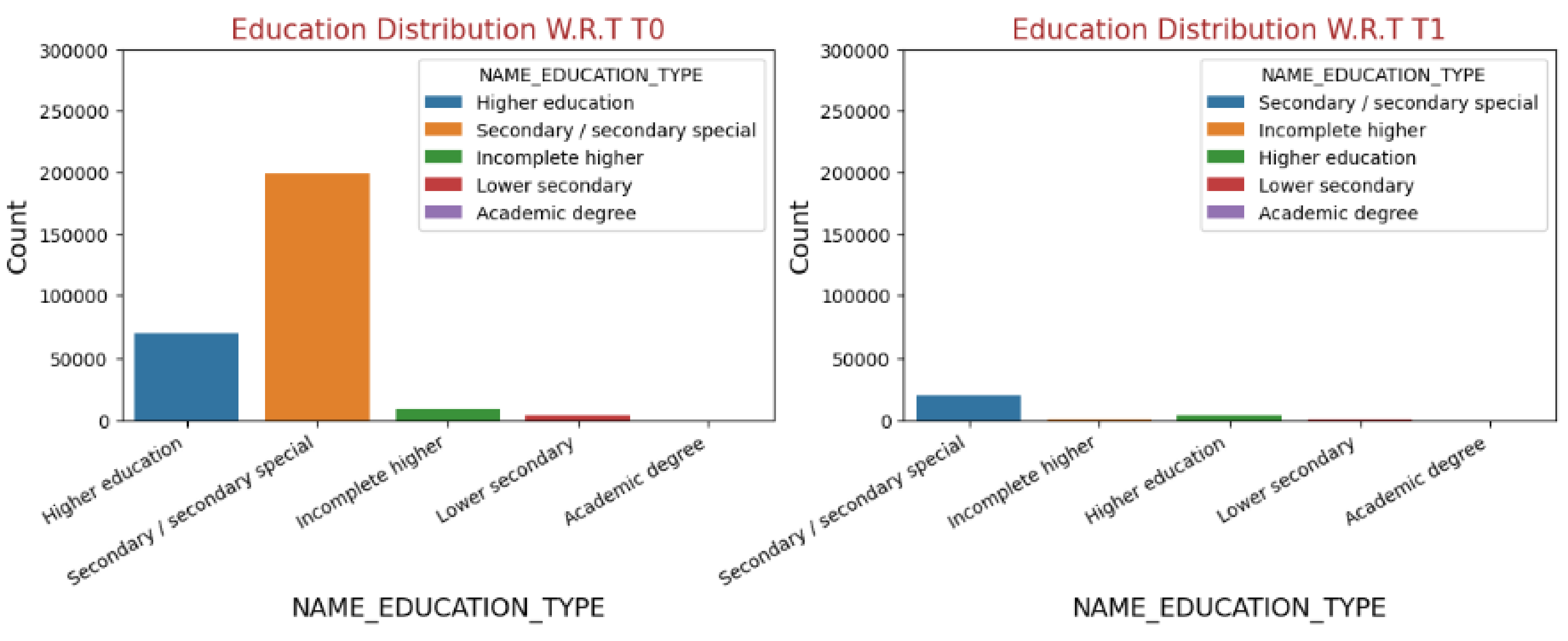


Income Type Distribution With Respect To Target Variables.



- ➢Maximum customers under Working category have less payment difficulties.
- ➢Unemployed, Pensioners and Maternity leave category are more likely to default.

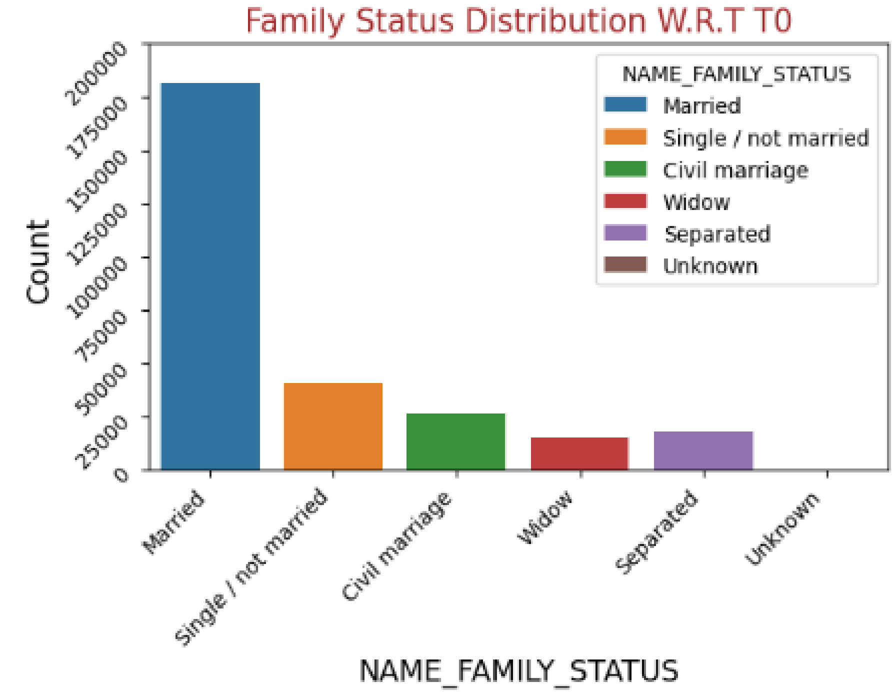
Education Type Distribution With Respect To Target Variables.

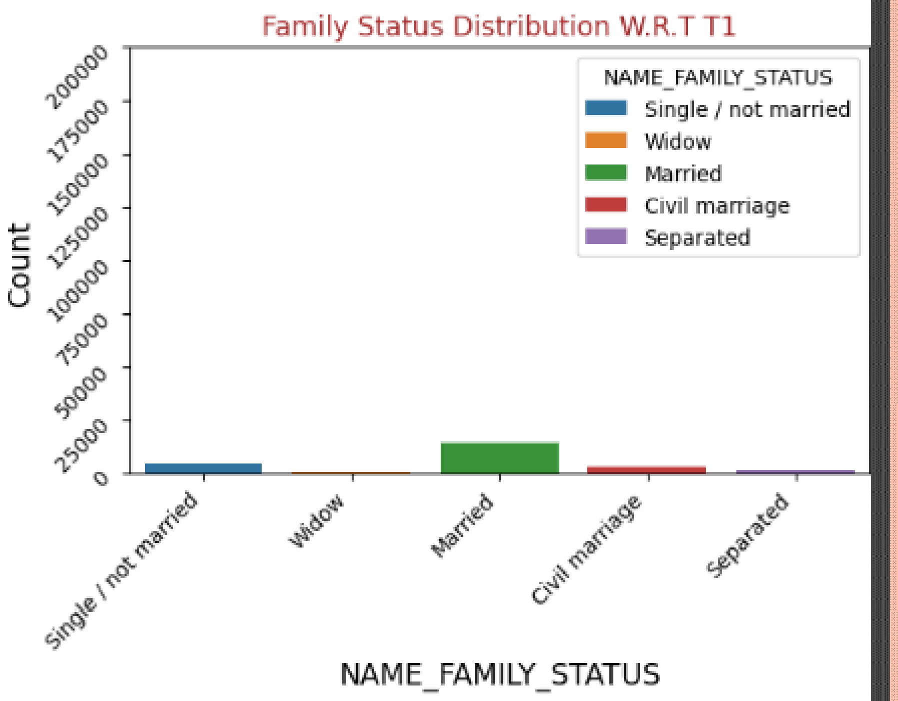


- Maximum Repayers are under the category Secondary/secondary special Education followed by higher education.

- Incomplete higher and Lower secondary are more likely to default.

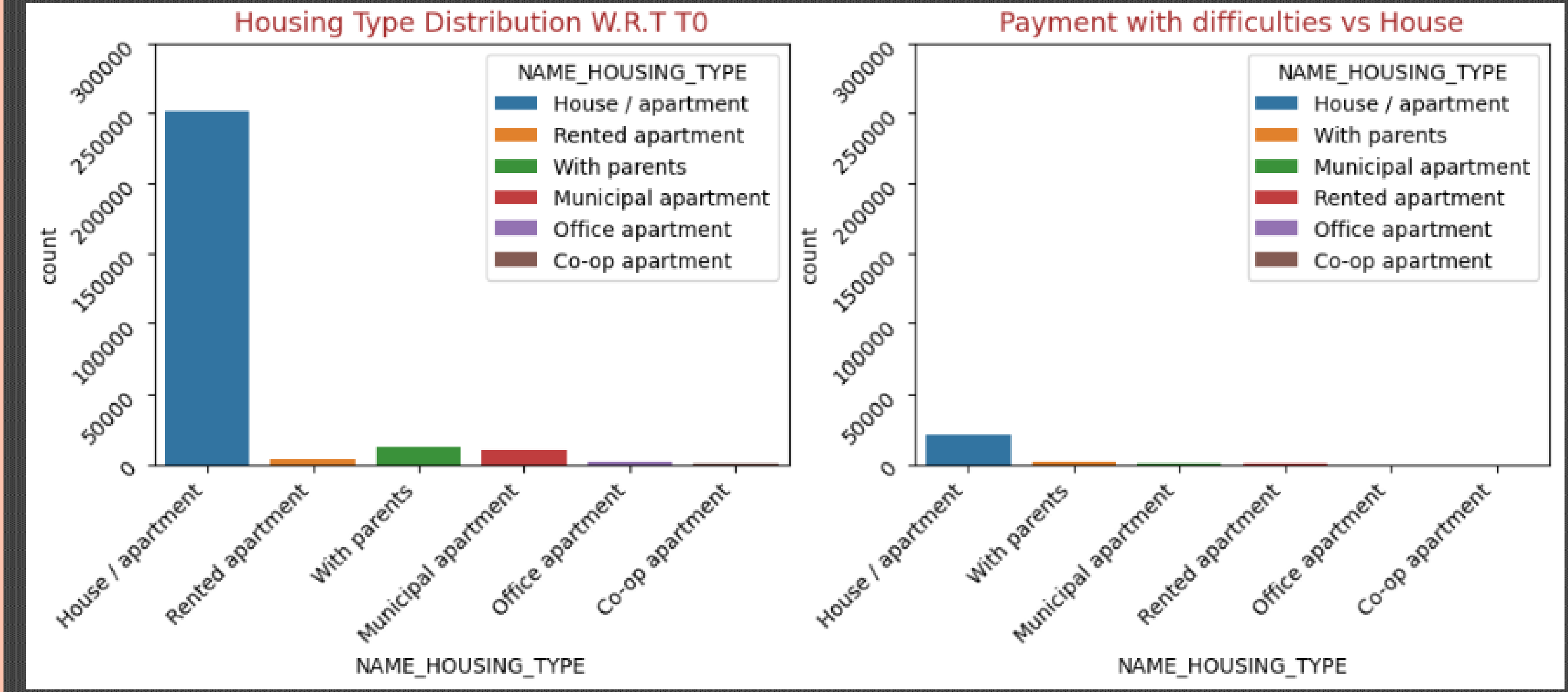
Family Status Distribution With Respect To Target Variables.





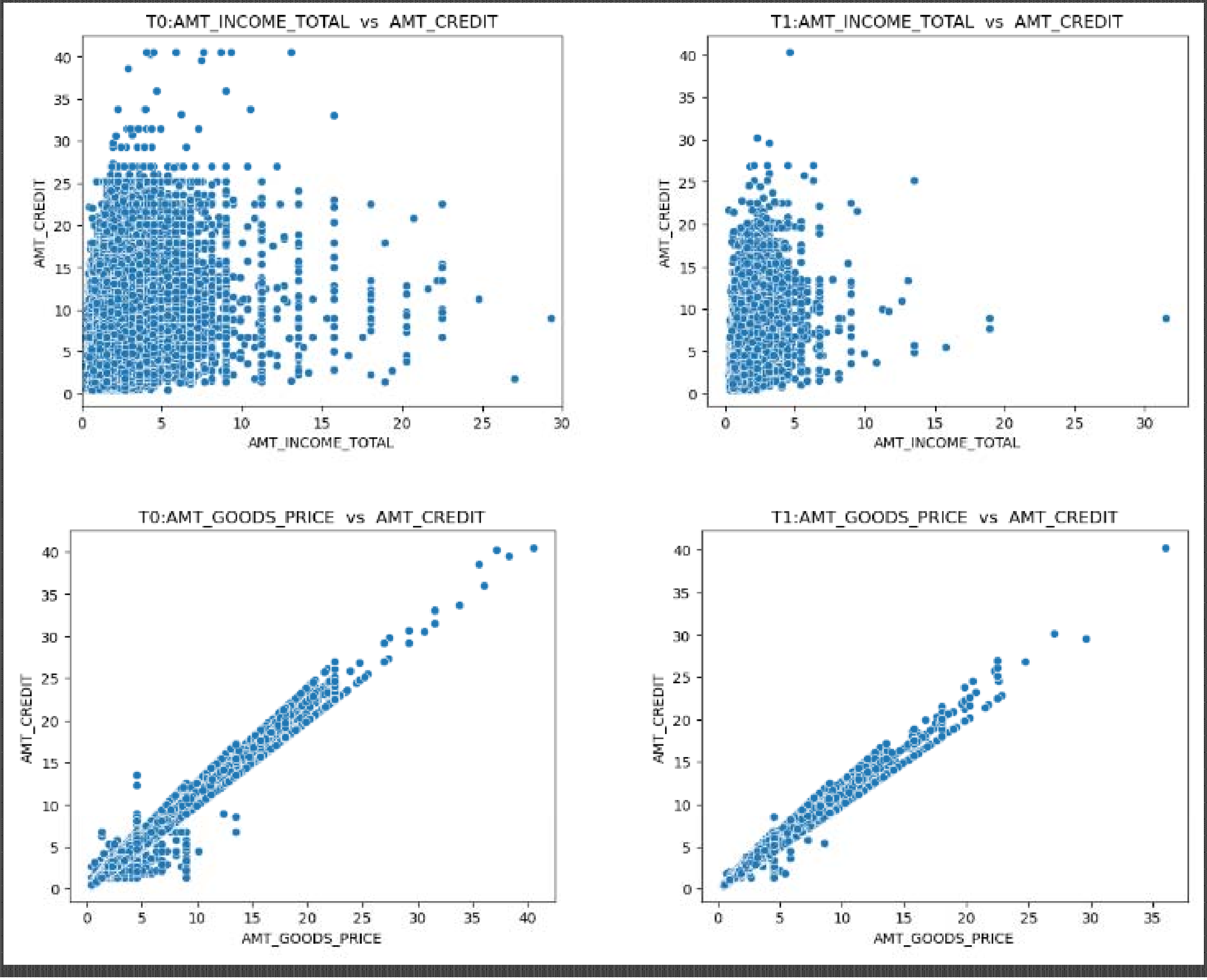
- ➢Married customers have less payment difficulties.
- ➢Separated, Civil Marriage, Single and Widows are more likely to default.

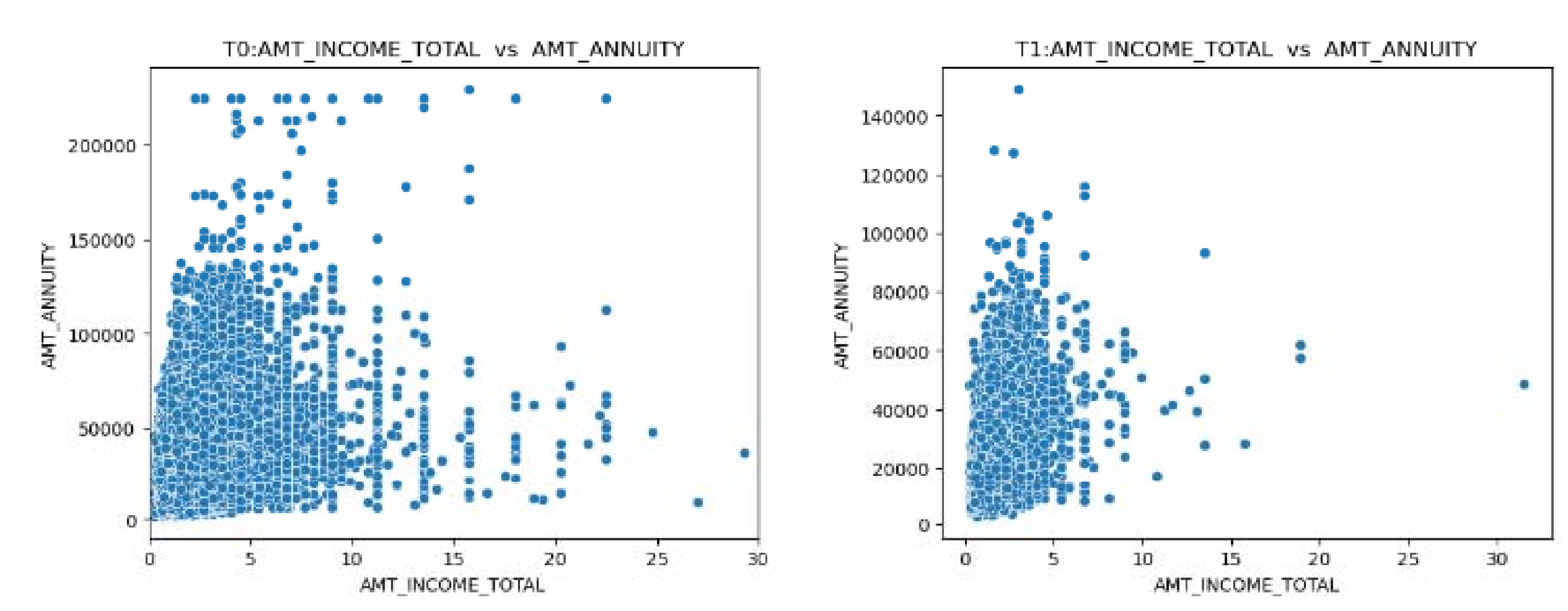
Housing Type Status Distribution With Respect To Target Variables.



- Customers owning House/Apartment have less payment difficulties.
- Customers living in Office apartment and Co-op apartment are more likely to default.

Bivariate Analysis





➢Credit amount for those who dont have payment difficulties is higher than those with payment difficulties.

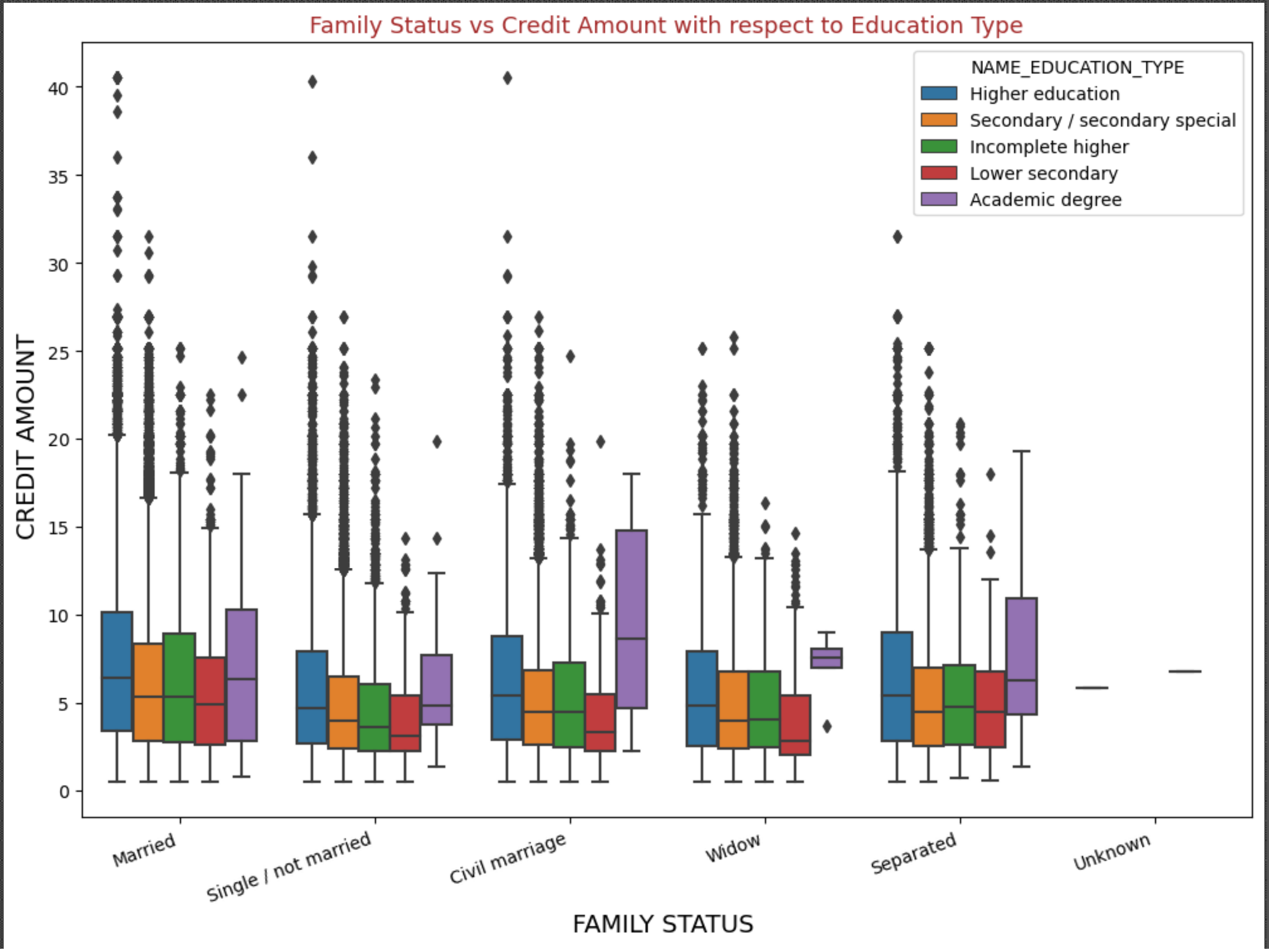
➢AMT_CREDIT and AMT_GOODS_PRICE are highly correlated. Customers with higher goods price and dont have payment difficulties have higher credit amount than those with higher goods price but are having payment difficulties.

➢Most of the defaulters are having Less than 5L total income

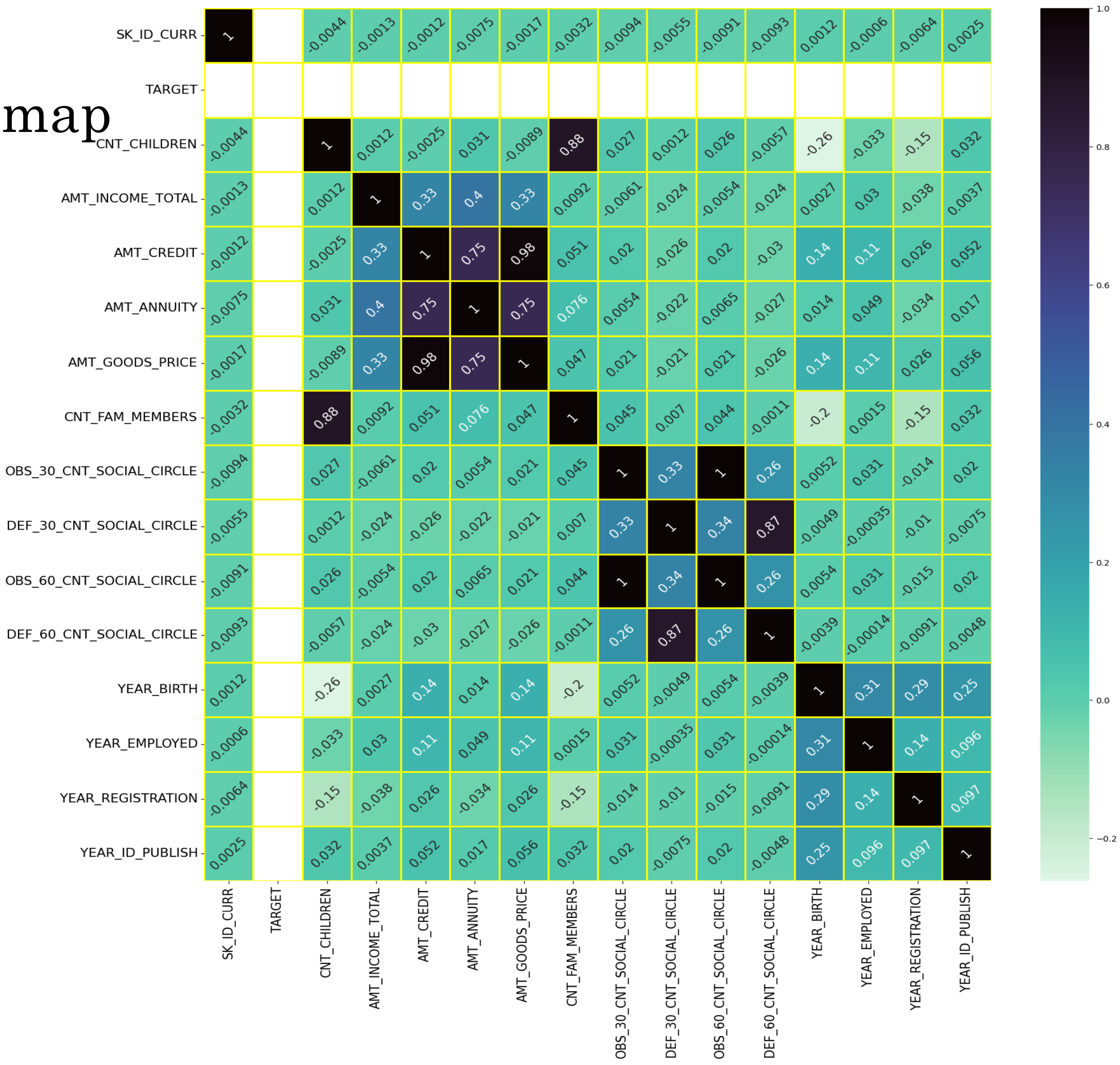
Family Status vs Credit Amount with respect to Education Type

➢Customers who are Married with Higher education followed by single with higher education has the higher credit than the rest.

➢Education type of Lower Secondary in all other family status has lowest credit.



Heat map



	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_FAM_MEMBERS	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	YEAR_BIRTH	YEAR_EMPLOYED	YEAR_REGISTRATION	YEAR_ID_PUBLISH
SK_ID_CURR	1		-0.0044	-0.0013	-0.0012	-0.0075	-0.0017	-0.0032	-0.0094	-0.0055	-0.0091	-0.0093	0.0012	-0.0006	-0.0064	0.0025
TARGET																
CNT_CHILDREN	-0.0044		1	0.0012	-0.0025	0.031	-0.0089	0.88	0.027	0.0012	0.026	-0.0057	-0.26	-0.033	-0.15	0.032
AMT_INCOME_TOTAL	-0.0013		0.0012	1	0.33	0.4	0.33	0.0092	-0.0061	-0.024	-0.0054	-0.024	0.0027	0.03	-0.038	0.0037
AMT_CREDIT	-0.0012		-0.0025	0.33	1	0.75	0.98	0.051	0.02	-0.026	0.02	-0.03	0.14	0.11	0.026	0.052
AMT_ANNUITY	-0.0075		0.031	0.4	0.75	1	0.75	0.076	0.0054	-0.022	0.0065	-0.027	0.014	0.049	-0.034	0.017
AMT_GOODS_PRICE	-0.0017		-0.0089	0.33	0.98	0.75	1	0.047	0.021	-0.021	0.021	-0.026	0.14	0.11	0.026	0.056
CNT_FAM_MEMBERS	-0.0032		0.88	0.0092	0.051	0.076	0.047	1	0.045	0.007	0.044	-0.0011	-0.2	0.0015	-0.15	0.032
OBS_30_CNT_SOCIAL_CIRCLE	-0.0094		0.027	-0.0061	0.02	0.0054	0.021	0.045	1	0.33	1	0.26	0.0052	0.031	-0.014	0.02
DEF_30_CNT_SOCIAL_CIRCLE	-0.0055		0.0012	-0.024	-0.026	-0.022	-0.021	0.007	0.33	1	0.34	0.87	-0.0049	-0.00035	-0.01	-0.0075
OBS_60_CNT_SOCIAL_CIRCLE	-0.0091		0.026	-0.0054	0.02	0.0065	0.021	0.044	1	0.34	1	0.26	0.0054	0.031	-0.015	0.02
DEF_60_CNT_SOCIAL_CIRCLE	-0.0093		-0.0057	-0.024	-0.03	-0.027	-0.026	-0.0011	0.26	0.87	0.26	1	-0.0039	-0.00014	-0.0091	-0.0048
YEAR_BIRTH	0.0012		-0.26	0.0027	0.14	0.014	0.14	-0.2	0.0052	-0.0049	0.0054	-0.0039	1	0.31	0.29	0.25
YEAR_EMPLOYED	-0.0006		-0.033	0.03	0.11	0.049	0.11	0.0015	0.031	-0.00035	0.031	-0.00014	0.31	1	0.14	0.096
YEAR_REGISTRATION	-0.0064		-0.15	-0.038	0.026	-0.034	0.026	-0.15	-0.014	-0.01	-0.015	-0.0091	0.29	0.14	1	0.097
YEAR_ID_PUBLISH	0.0025		0.032	0.0037	0.052	0.017	0.056	0.032	0.02	-0.0075	0.02	-0.0048	0.25	0.096	0.097	1

Deduction

Highly Positively Correlated Variables.

❖- AMT_CREDIT vs AMT_GOODS_PRICE

❖- AMT_CREDIT vs AMT_ANNUITY

❖- AMT_ANNUITY vs AMT_GOODS_PRICE

❖- CNT_FAM_MEMBERS vs CNT_CHILDREN

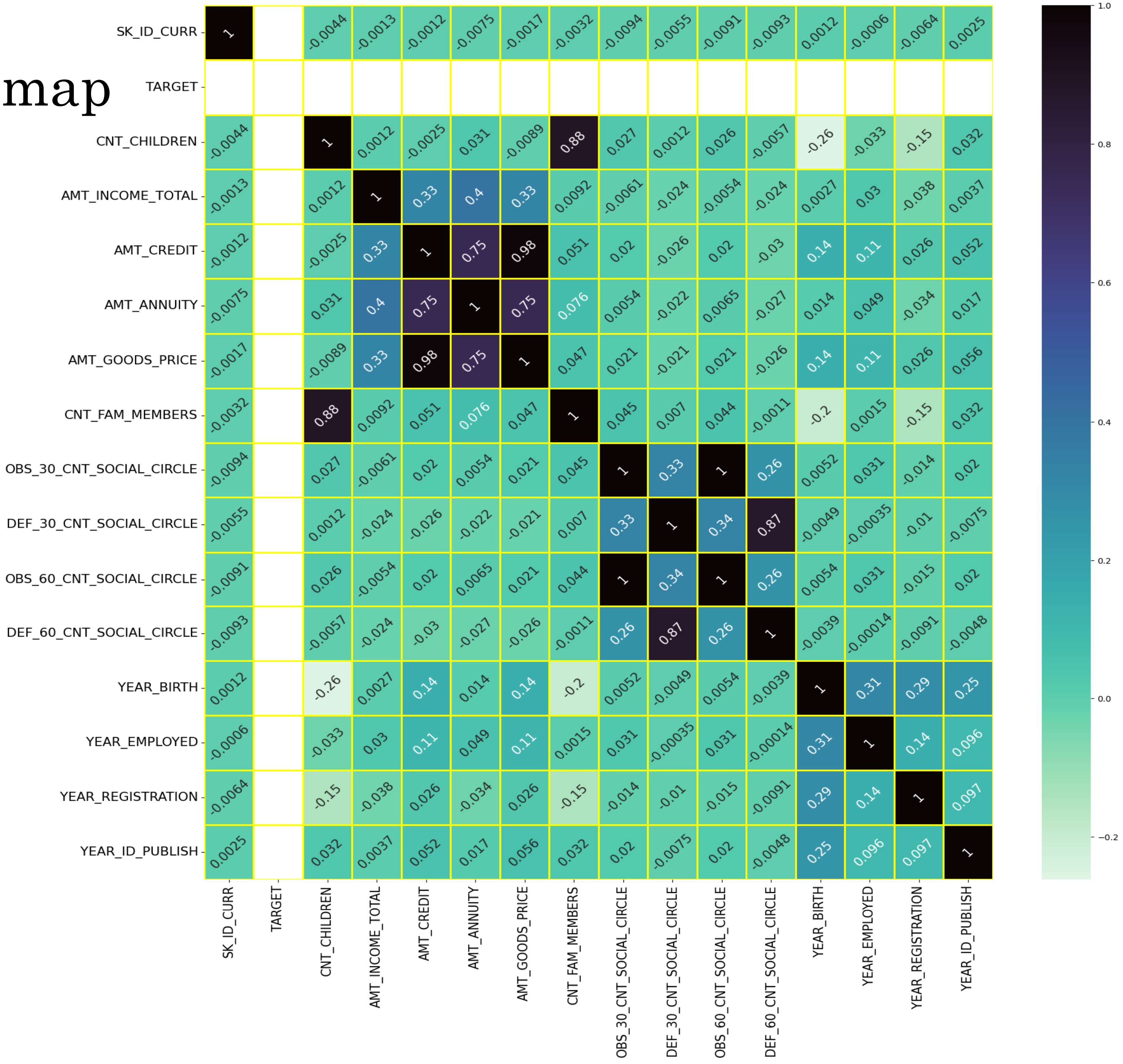
❖-DEF_30_CNT_SOCIAL_CIRCLE vs DEF_60_CNT_SOCIAL_CIRCLE

❖-OBS_30_CNT_SOCIAL_CIRCLE vs OBS_60_CNT_SOCIAL_CIRCLE

Highly Negatively Correlated Variables.

❖- There are no Highly Negatively Correlated Variables**.**

Heat map



	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	CNT_FAM_MEMBERS	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	YEAR_BIRTH	YEAR_EMPLOYED	YEAR_REGISTRATION	YEAR_ID_PUBLISH
SK_ID_CURR	1		-0.0044	-0.0013	-0.0012	-0.0075	-0.0017	-0.0032	-0.0094	-0.0055	-0.0091	-0.0093	0.0012	-0.0006	-0.0064	0.0025
TARGET																
CNT_CHILDREN	-0.0044		1	0.0012	-0.0025	0.031	-0.0089	0.88	0.027	0.0012	0.026	-0.0057	-0.26	-0.033	-0.15	0.032
AMT_INCOME_TOTAL	-0.0013		0.0012	1	0.33	0.4	0.33	0.0092	-0.0061	-0.024	-0.0054	-0.024	0.0027	0.03	-0.038	0.0037
AMT_CREDIT	-0.0012		-0.0025	0.33	1	0.75	0.98	0.051	0.02	-0.026	0.02	-0.03	0.14	0.11	0.026	0.052
AMT_ANNUITY	-0.0075		0.031	0.4	0.75	1	0.75	0.076	0.0054	-0.022	0.0065	-0.027	0.014	0.049	-0.034	0.017
AMT_GOODS_PRICE	-0.0017		-0.0089	0.33	0.98	0.75	1	0.047	0.021	-0.021	0.021	-0.026	0.14	0.11	0.026	0.056
CNT_FAM_MEMBERS	-0.0032		0.88	0.0092	0.051	0.076	0.047	1	0.045	0.007	0.044	-0.0011	-0.2	0.0015	-0.15	0.032
OBS_30_CNT_SOCIAL_CIRCLE	-0.0094		0.027	-0.0061	0.02	0.0054	0.021	0.045	1	0.33	1	0.26	0.0052	0.031	-0.014	0.02
DEF_30_CNT_SOCIAL_CIRCLE	-0.0055		0.0012	-0.024	-0.026	-0.022	-0.021	0.007	0.33	1	0.34	0.87	-0.0049	-0.00035	-0.01	-0.0075
OBS_60_CNT_SOCIAL_CIRCLE	-0.0091		0.026	-0.0054	0.02	0.0065	0.021	0.044	1	0.34	1	0.26	0.0054	0.031	-0.015	0.02
DEF_60_CNT_SOCIAL_CIRCLE	-0.0093		-0.0057	-0.024	-0.03	-0.027	-0.026	-0.0011	0.26	0.87	0.26	1	-0.0039	-0.00014	-0.0091	-0.0048
YEAR_BIRTH	0.0012		-0.26	0.0027	0.14	0.014	0.14	-0.2	0.0052	-0.0049	0.0054	-0.0039	1	0.31	0.29	0.25
YEAR_EMPLOYED	-0.0006		-0.033	0.03	0.11	0.049	0.11	0.0015	0.031	-0.00035	0.031	-0.00014	0.31	1	0.14	0.096
YEAR_REGISTRATION	-0.0064		-0.15	-0.038	0.026	-0.034	0.026	-0.15	-0.014	-0.01	-0.015	-0.0091	0.29	0.14	1	0.097
YEAR_ID_PUBLISH	0.0025		0.032	0.0037	0.052	0.017	0.056	0.032	0.02	-0.0075	0.02	-0.0048	0.25	0.096	0.097	1

Deduction

Highly Positively Correlated Variables.

❖- AMT_CREDIT vs AMT_GOODS_PRICE

❖- AMT_CREDIT vs AMT_ANNUITY

❖- AMT_ANNUITY vs AMT_GOODS_PRICE

❖- CNT_FAM_MEMBERS vs CNT_CHILDREN

❖-DEF_30_CNT_SOCIAL_CIRCLE vs DEF_60_CNT_SOCIAL_CIRCLE

❖-OBS_30_CNT_SOCIAL_CIRCLE vs OBS_60_CNT_SOCIAL_CIRCLE

Highly Negatively Correlated Variables.

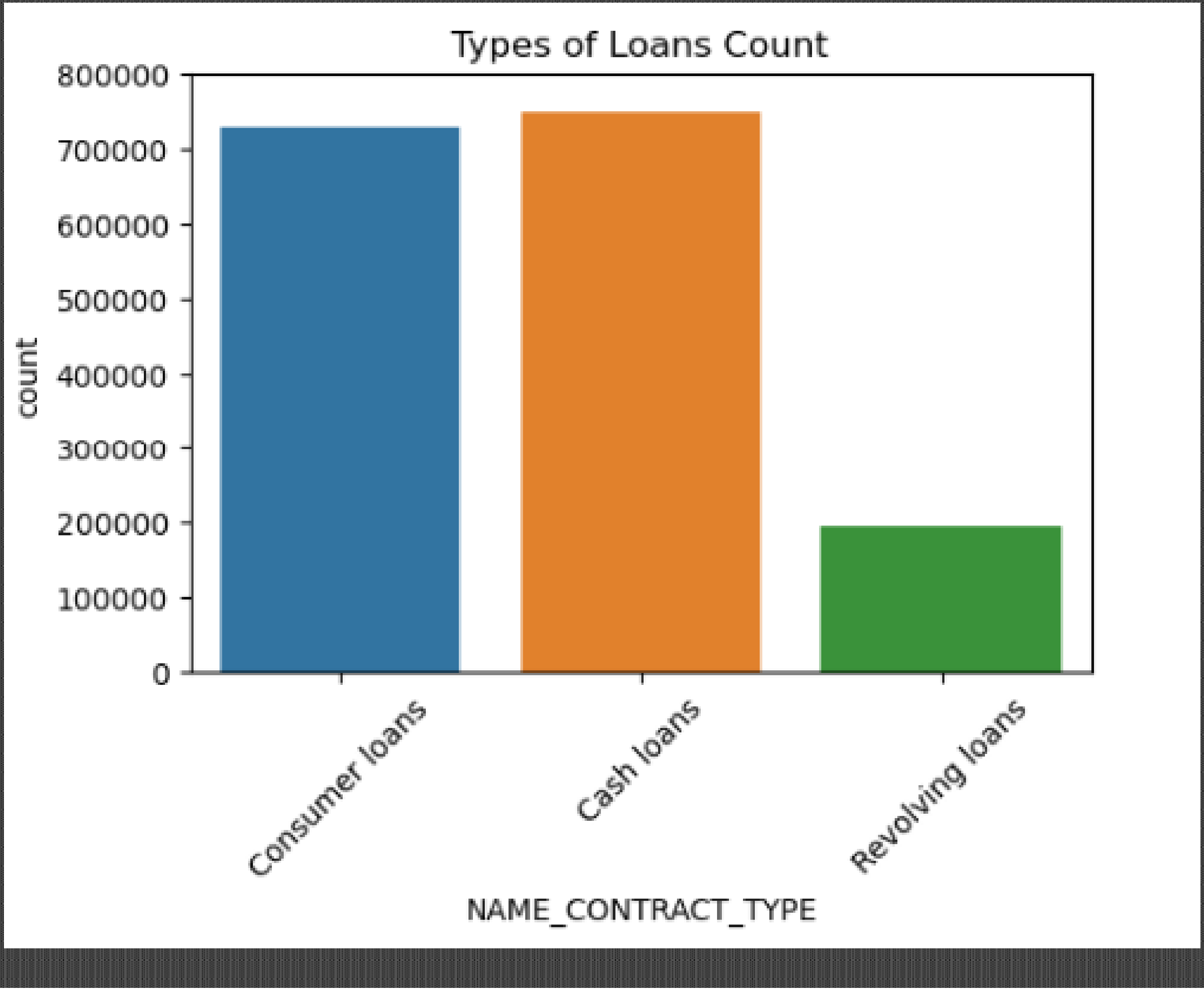
❖- There are no Highly Negatively Correlated Variables.

So from the above heat maps we can conclude that the variables correlated in T0 and T1 are same with small variation in the correlation values.

-

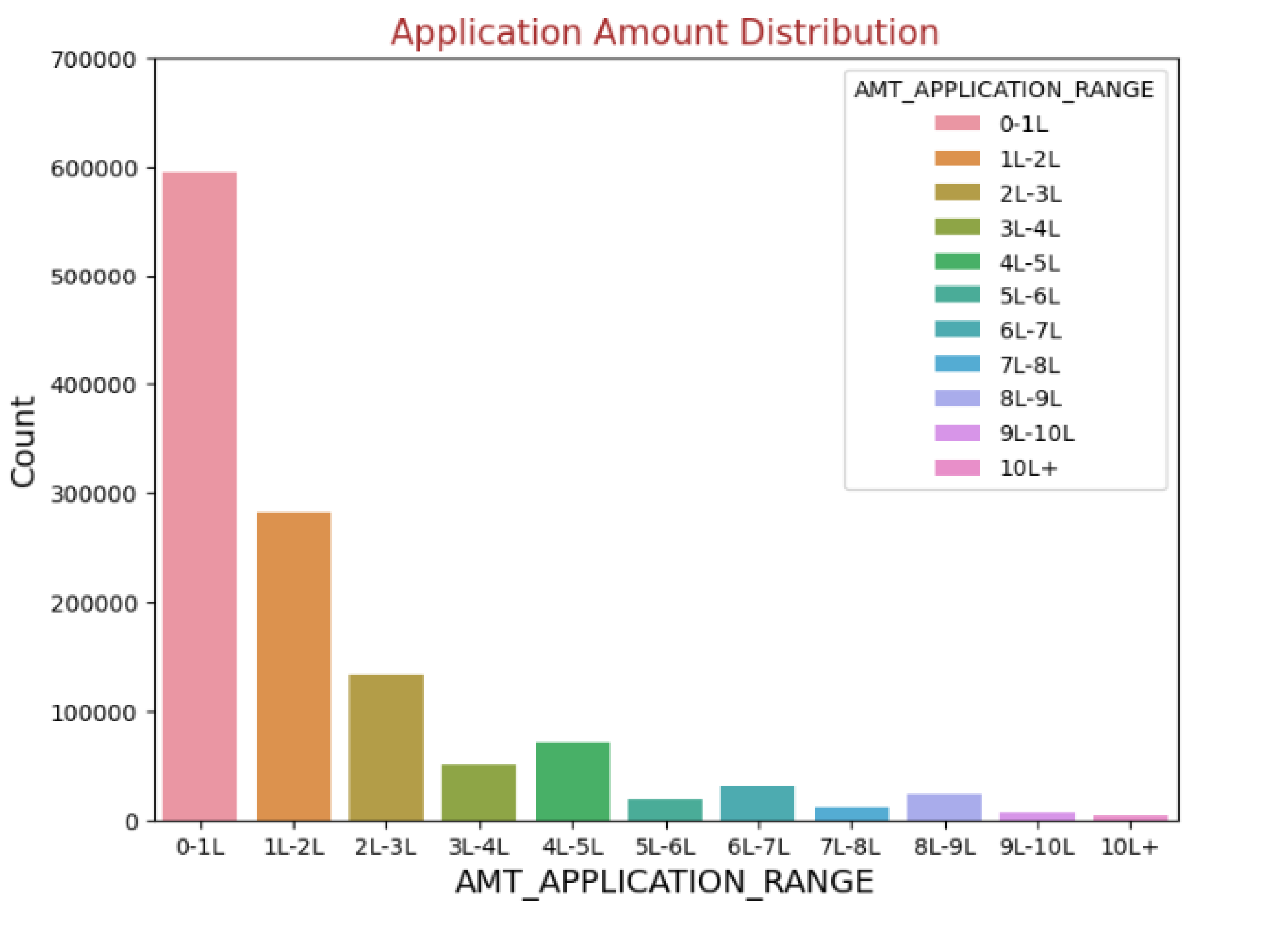
Distribution of Types of Loan

- Maximum customers prefer Cash Loans followed by Consumer loans..



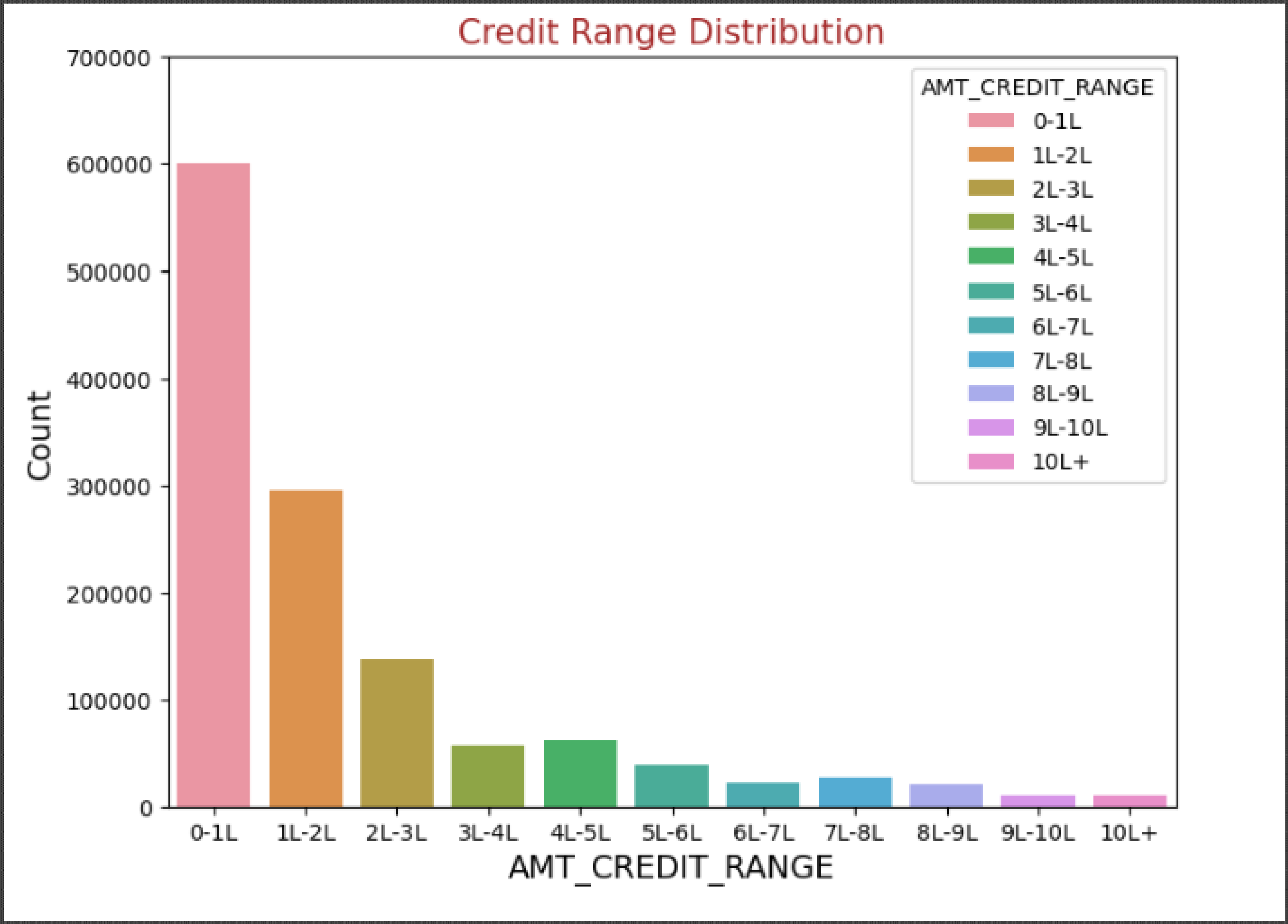
Application Amount Distribution

- Maximum customers applied for a loan of 0-1 Lakhs followed by 1-2 Lakhs.

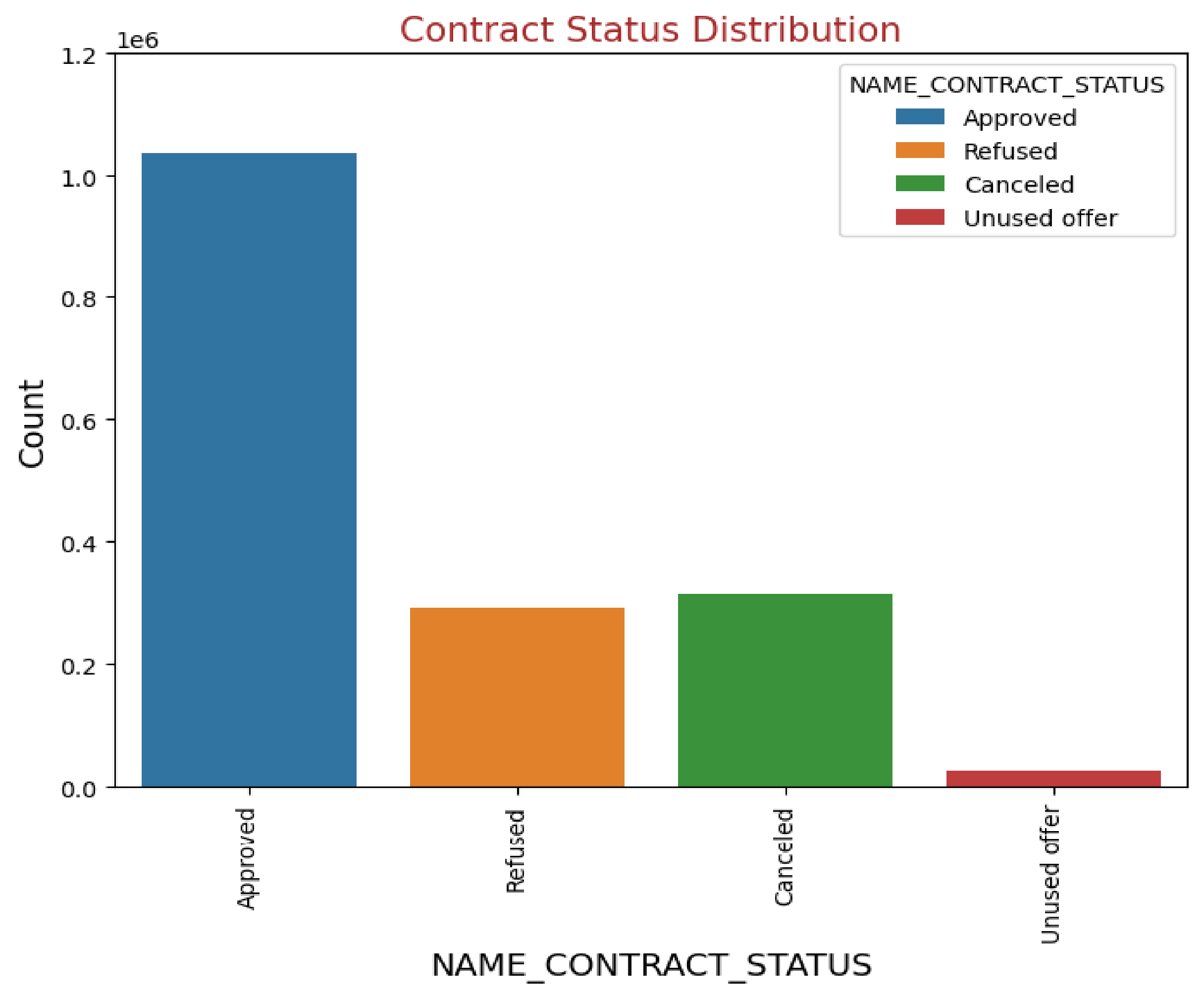


Credit Range Distribution

- Maximum customers received loans of 0-1 Lakhs followed by 1-2 Lakhs.



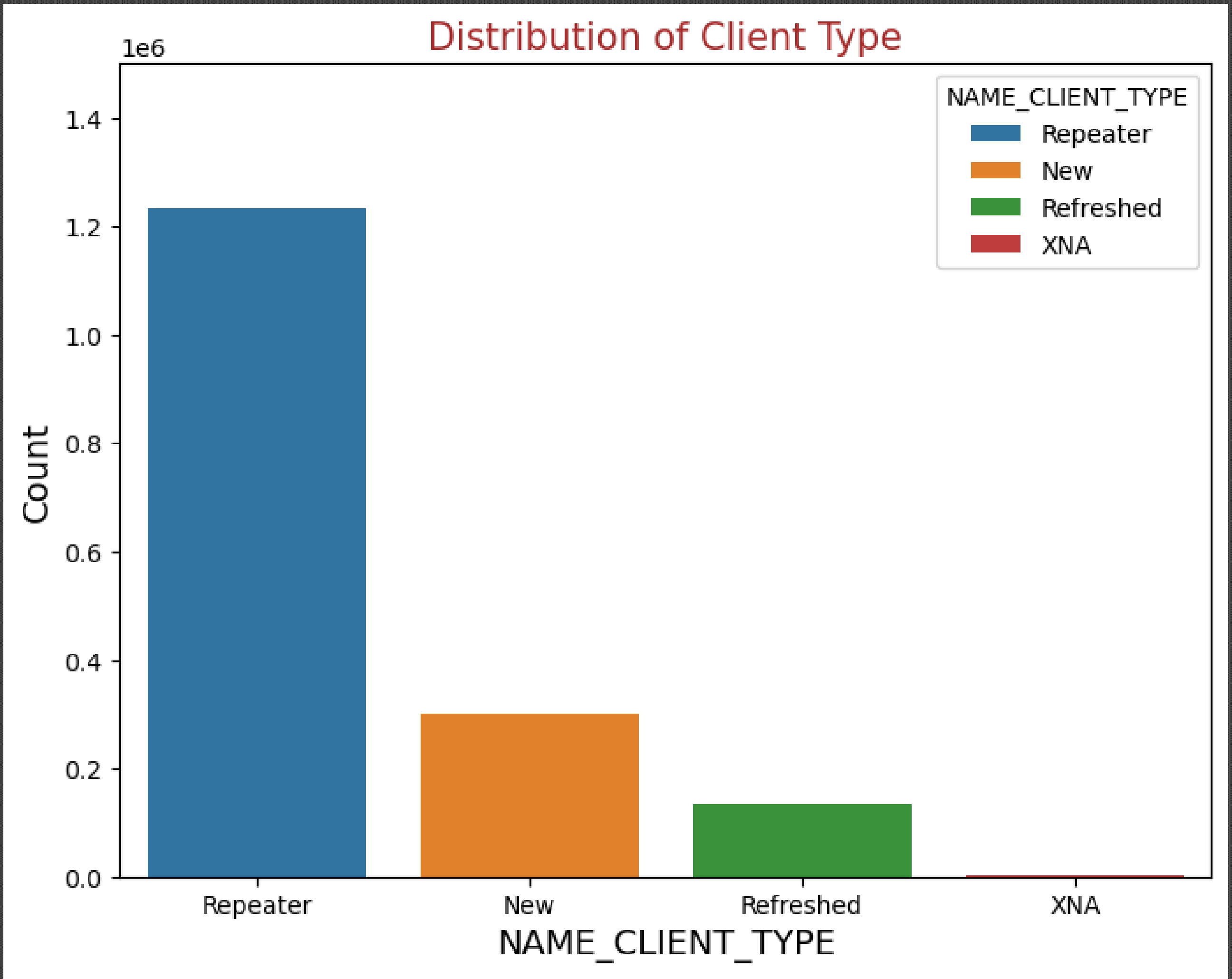
Contract Status Distribution



- Bank has approved more applications than other 3 contract staus.
- Refused and Cancelled applications are having almost similar counts.

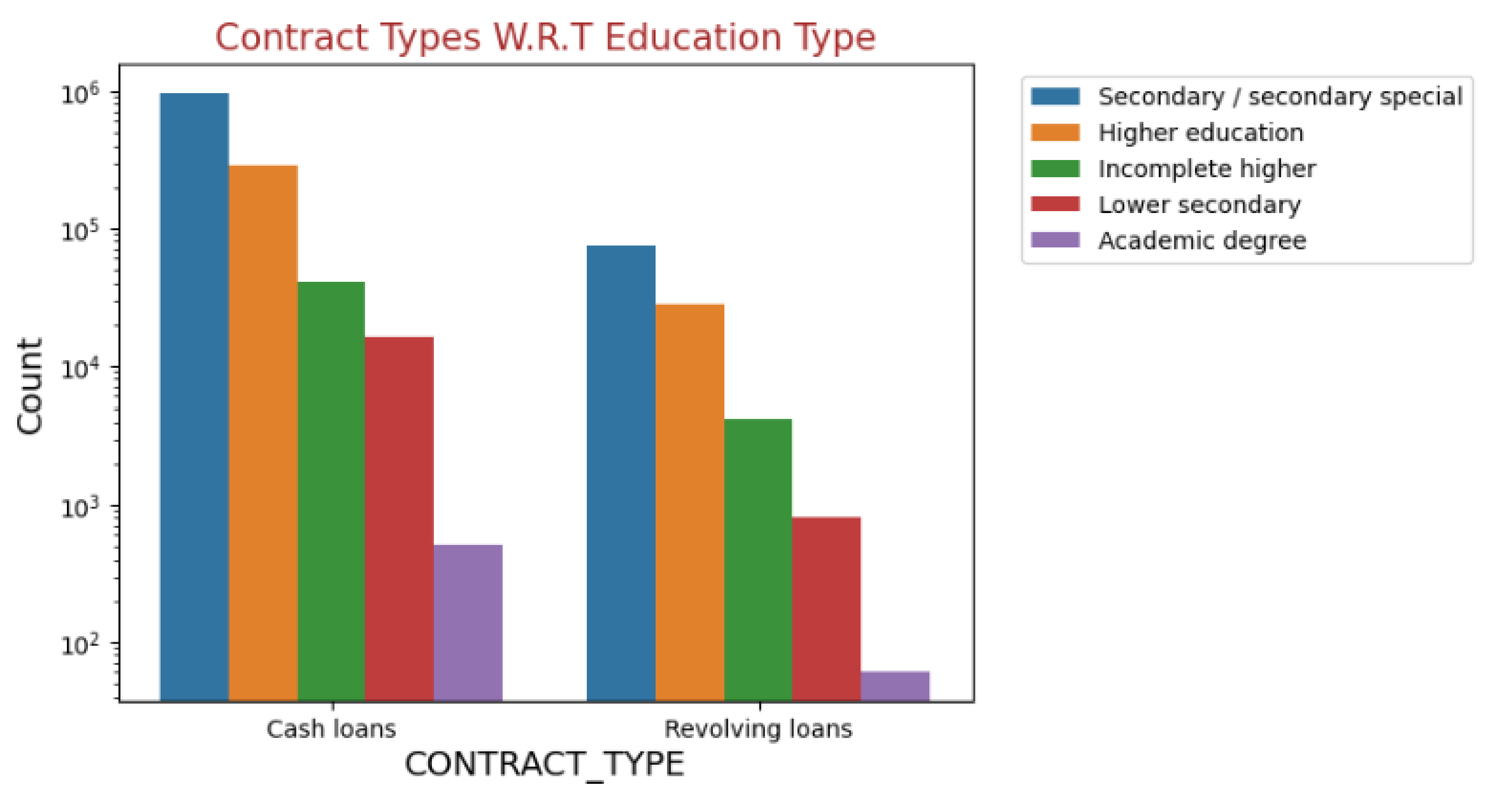
Client Type Distribution

- Maximum number of customers are Repeaters, i'e they have done business with bank before..



MERGED CURRENT AND PREVIOUS APPLICATION DATA ANALYSIS

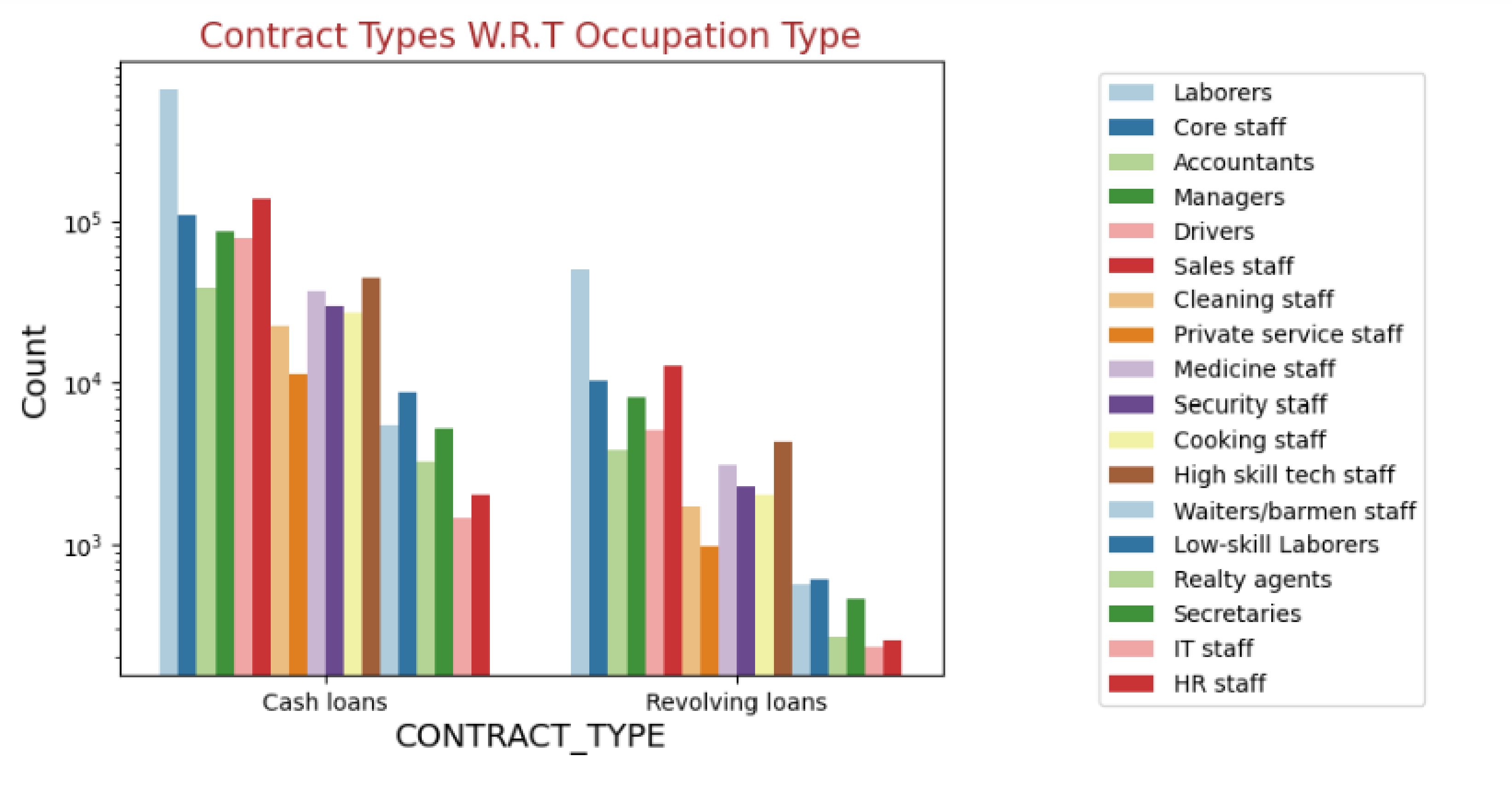
Distribution of Contract Types with respect to Education Type



- Most clients with all kinds of education types prefer cash loans over revolving loans.

- Client with Secondary/Secondary special education types takes more number of loans compared to others.

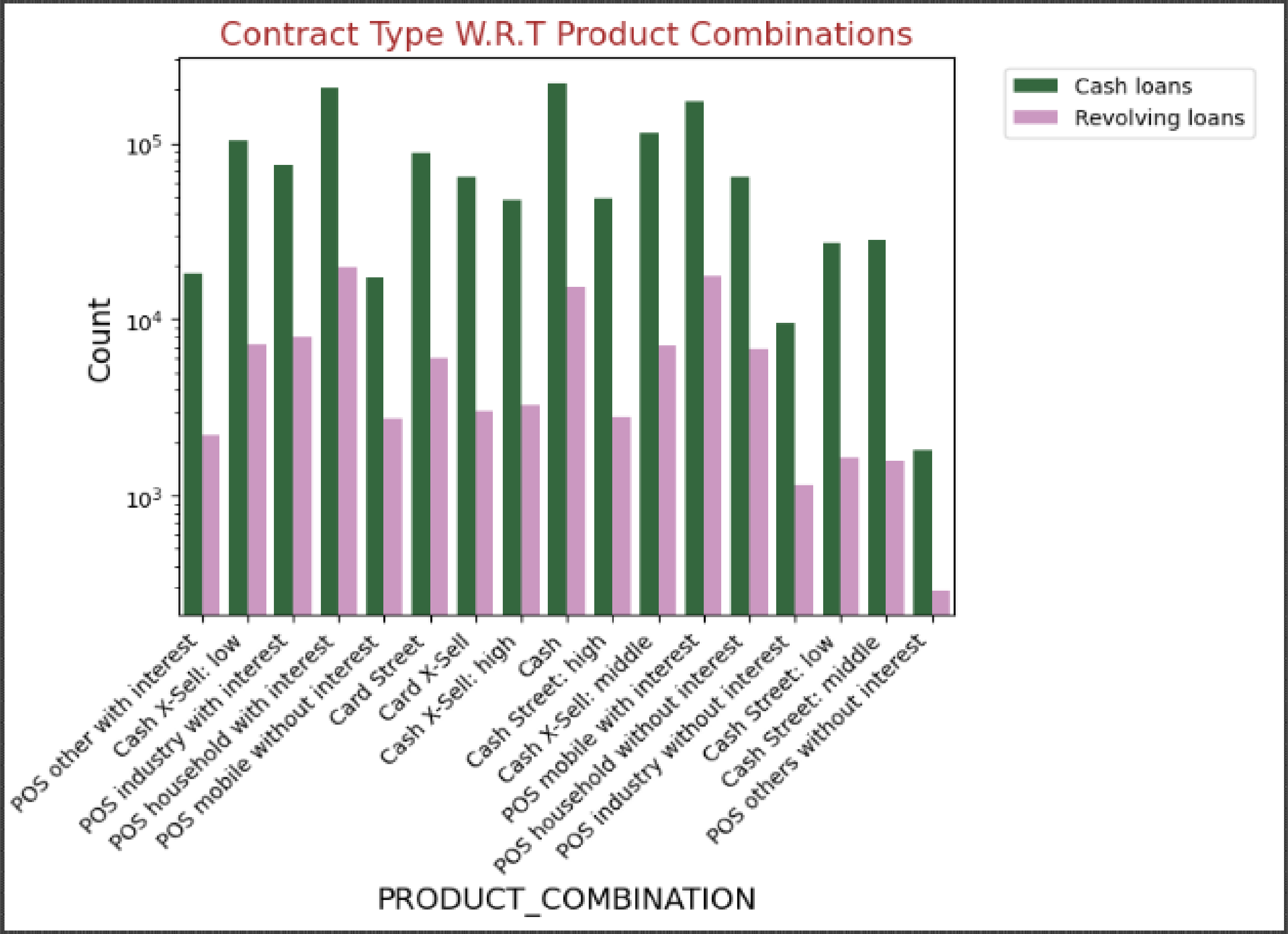
Distribution of Contract Types with respect to Occupation Type



➢Max number of clients preferring cash loans and revolving comes under Laborer category.

➢Least number of clients preferring cash loans and revolving comes under IT Staff category.

Distribution of Contract Types with respect to Product Combinations

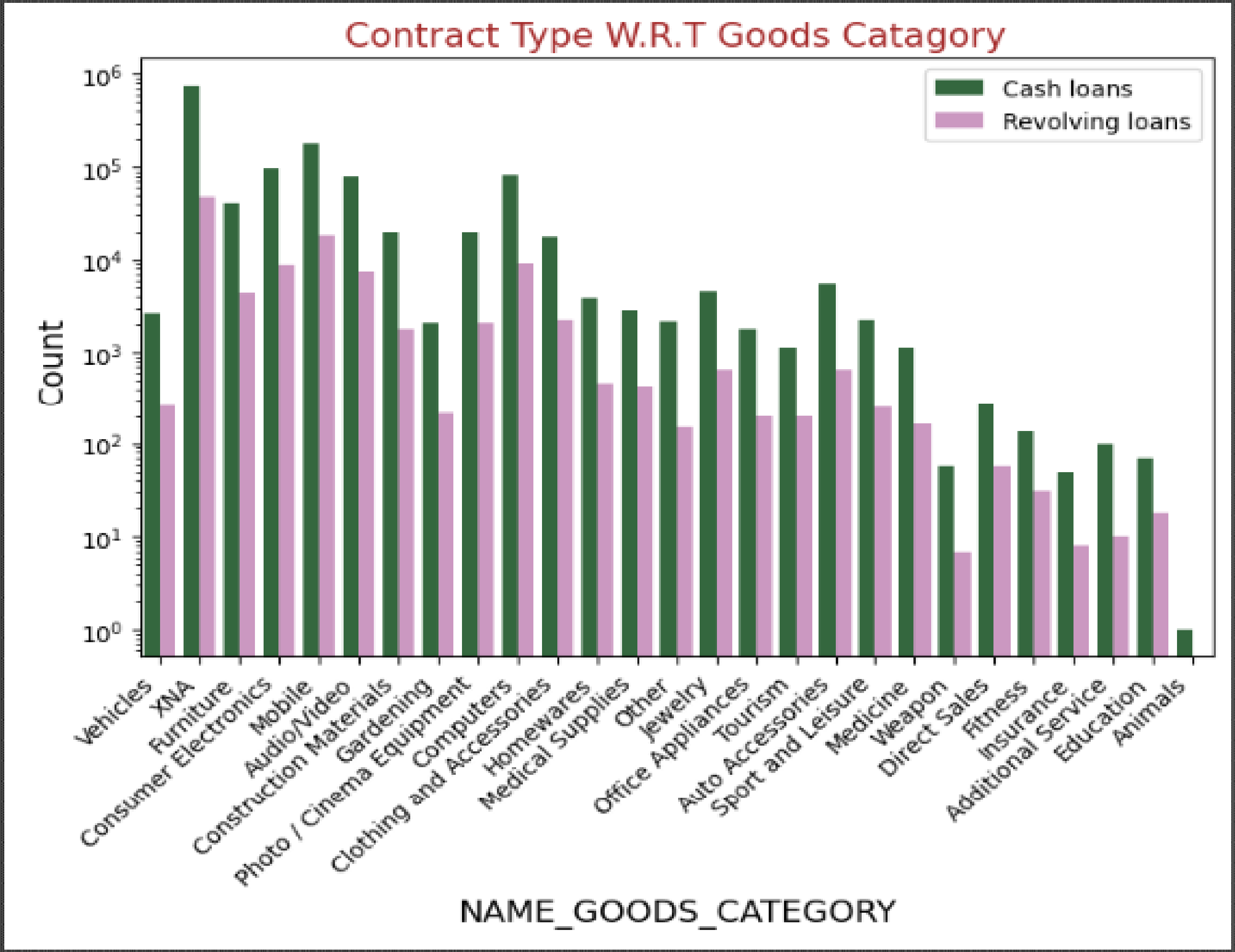


- Maximum number of cash loans are for Product combination of POS household with interest followed by Cash.
- Least number of loans are for product combination of POS others without interest

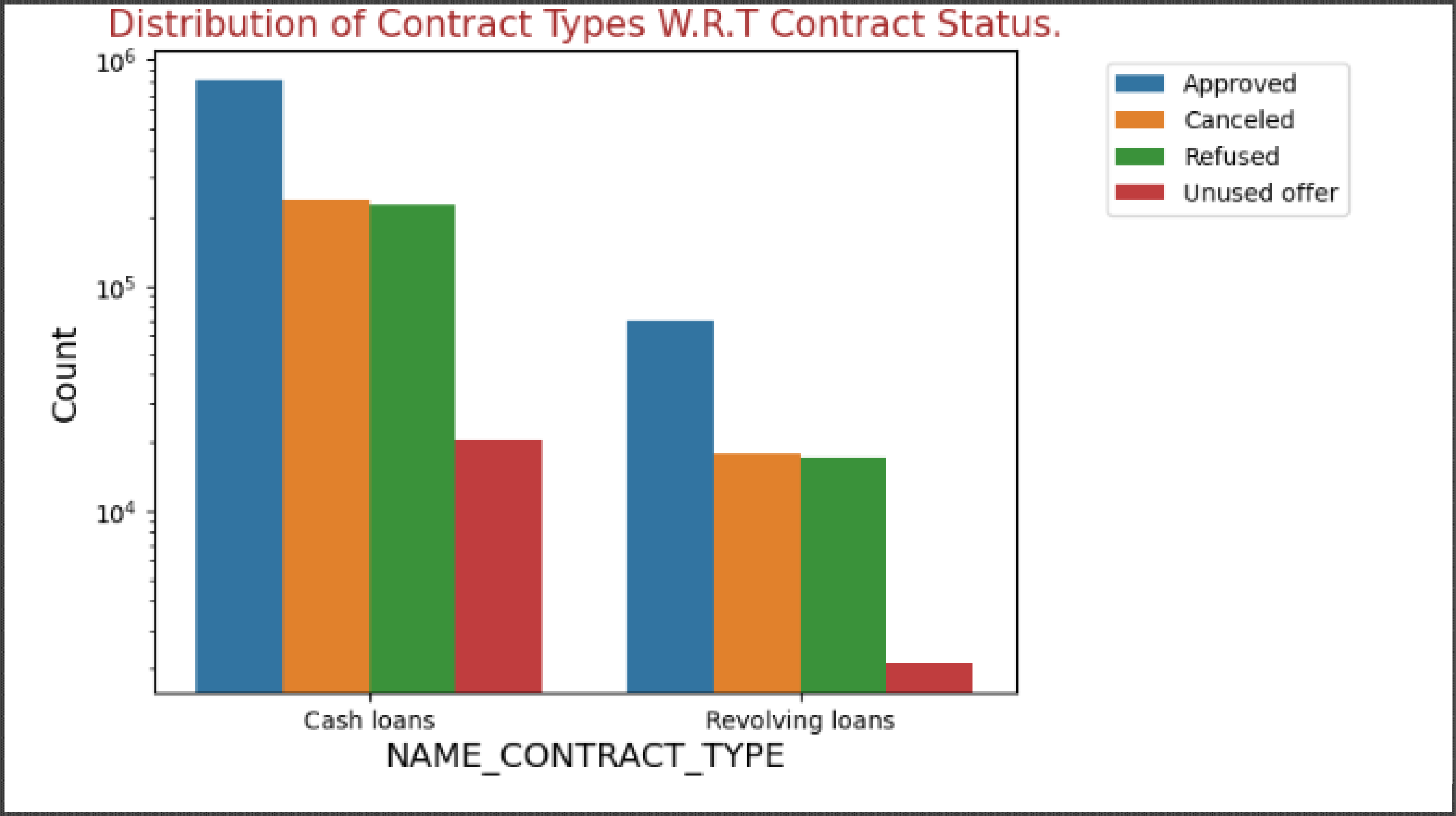
Distribution of Contract Types with respect to Goods Category

- ➢Maximum number of loans are for XNA followed by Mobile and Electronics.

- ➢Least number of loans are for Animals.



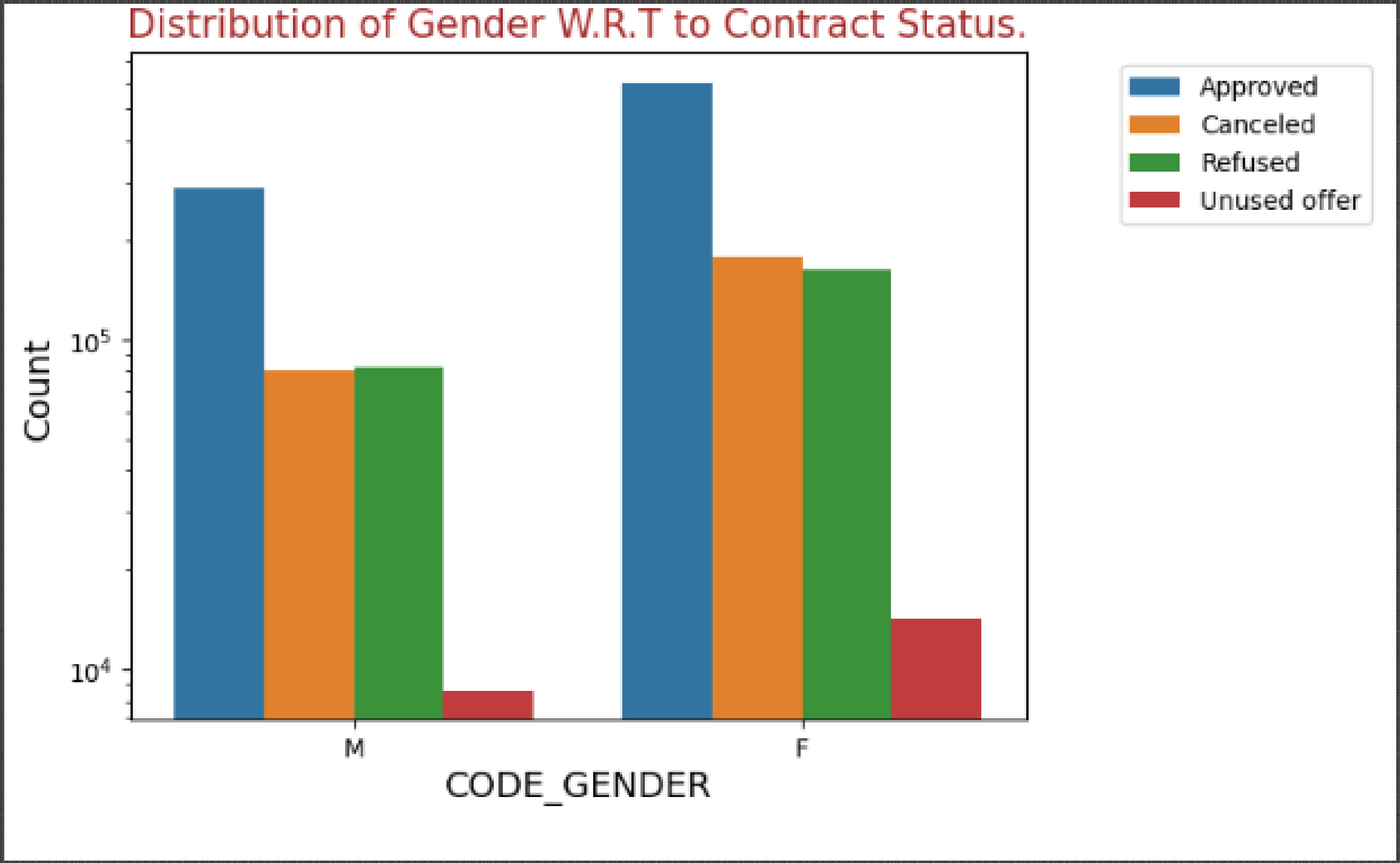
Distribution of Contract Types with respect to Contract Status

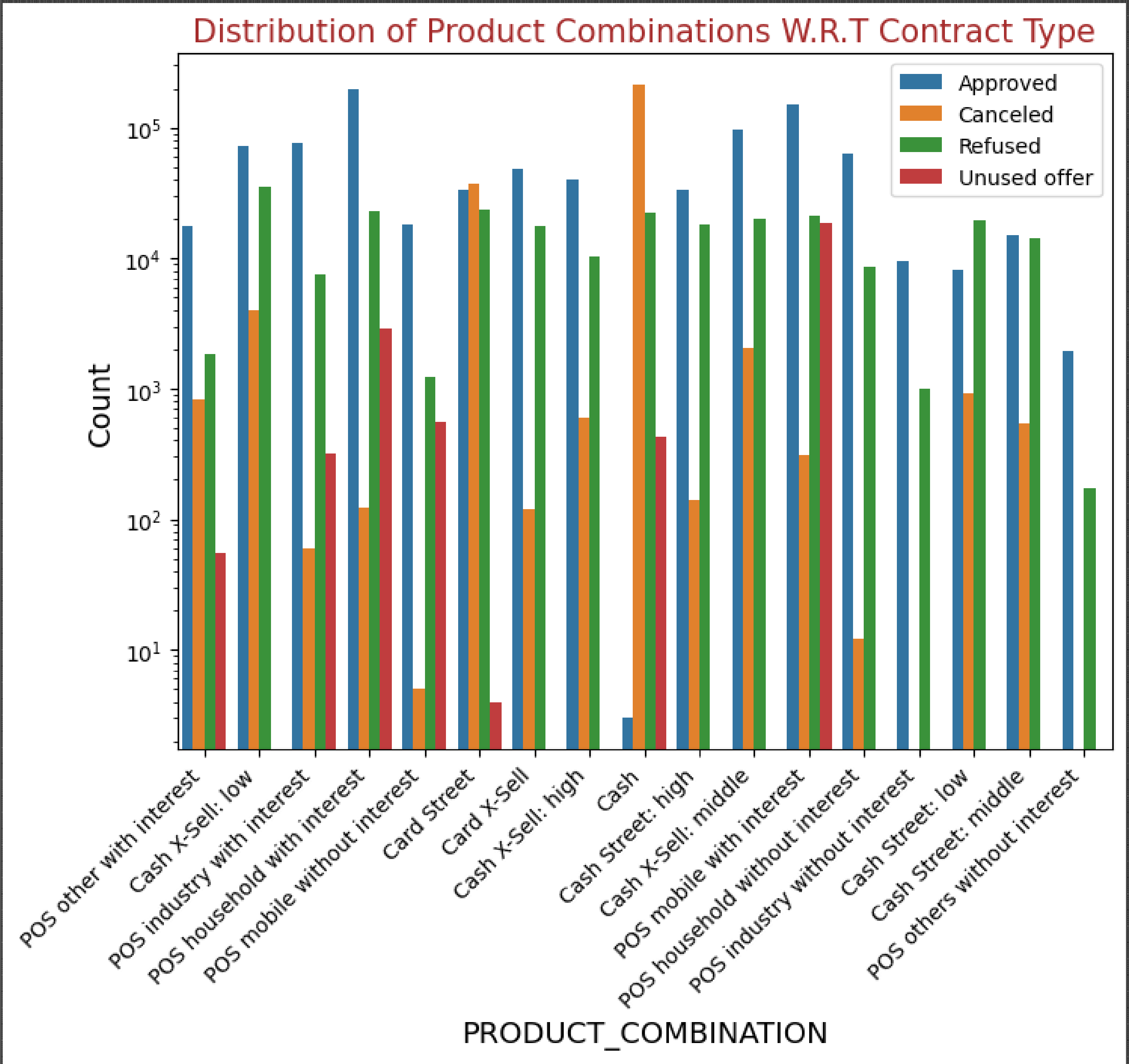


- ➢Maximum number of approved loans are from Cash Loan contract type.
- ➢Contract status count is higher for Cash loans compared to Revolving loans
- ➢Customers tend to accept more offers given for Revolving Loans.

Distribution of Gender with respect to Contract Status

- Female Customers have Maximum approved loans.
- Male clients use most of the offers of loans as unused offers are very less for male clients than that of female client




Distribution of Product Combinations W.R.T Contract Type
Approved
Canceled
Refused
Unused offer
Count
10^5
10^4
10^3
10^2
10^1
POS other with interest
Cash X-Sell: low
POS industry with interest
POS household with interest
POS mobile without interest
Card Street
Card X-Sell
Cash X-Sell: high
Cash
Cash Street: high
Cash X-Sell: middle
POS mobile with interest
POS household without interest
POS industry without interest
Cash Street: low
Cash Street: middle
POS others without interest
PRODUCT_COMBINATION

Distribution of Product Combinations W.R.T Contract Type

MAXIMUM COUNTS

a) Maximum number of approved loans are for POS household with interest.

b) Maximum number of cancelled loans are for Cash.

c) Maximum number of refused loans are for Cash X-Sell: low.

d) Maximum number of unused offers are for POS mobile with interest

MINIMUM COUNTS

a) Minimum number of approved loans are for Cash.

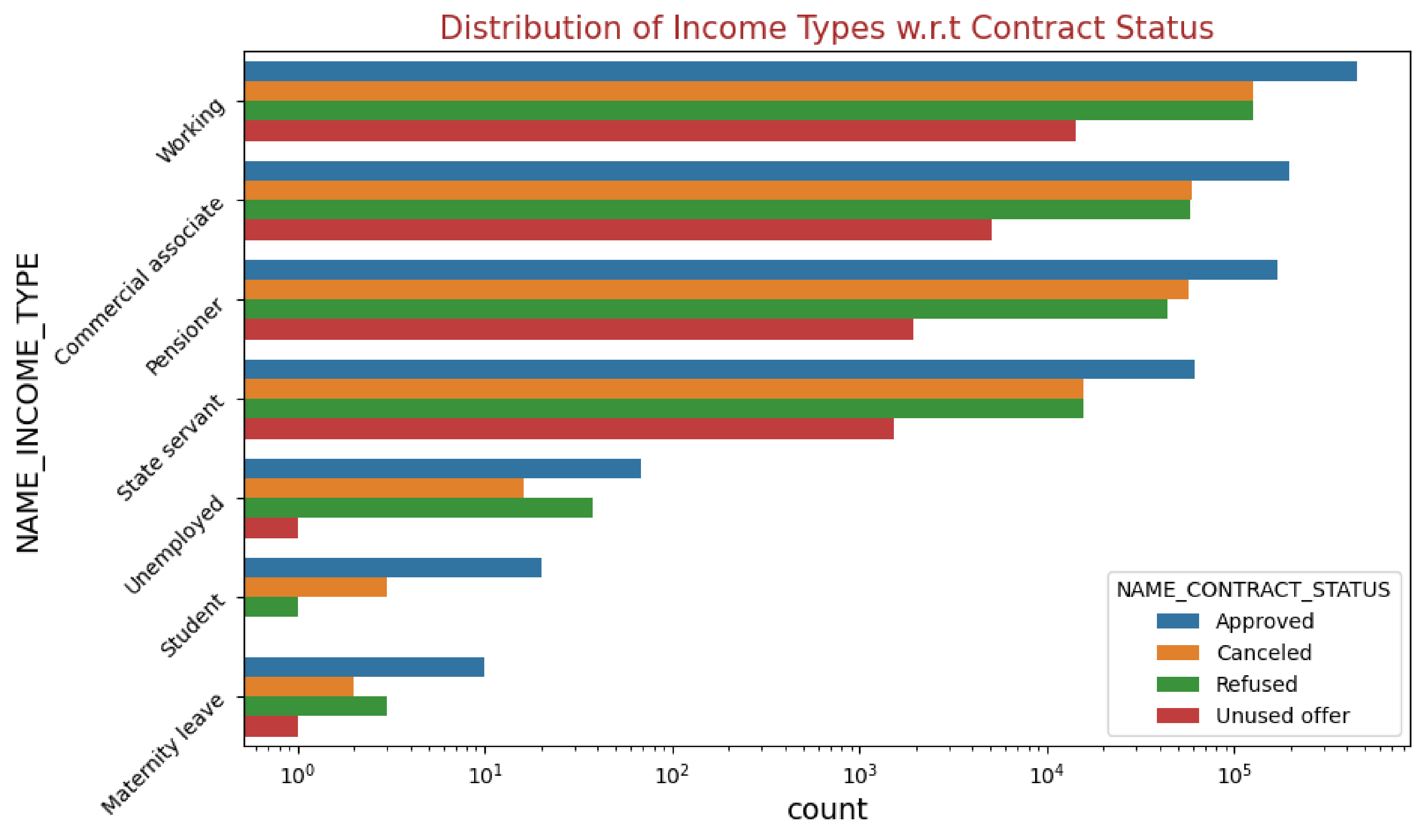
b) Minimum number of cancelled loans are for POS industry without interest, POS others without interest.

c) Minimum number of refused loans are for POS others without interest.

d) Minimum number of unused offers are for:

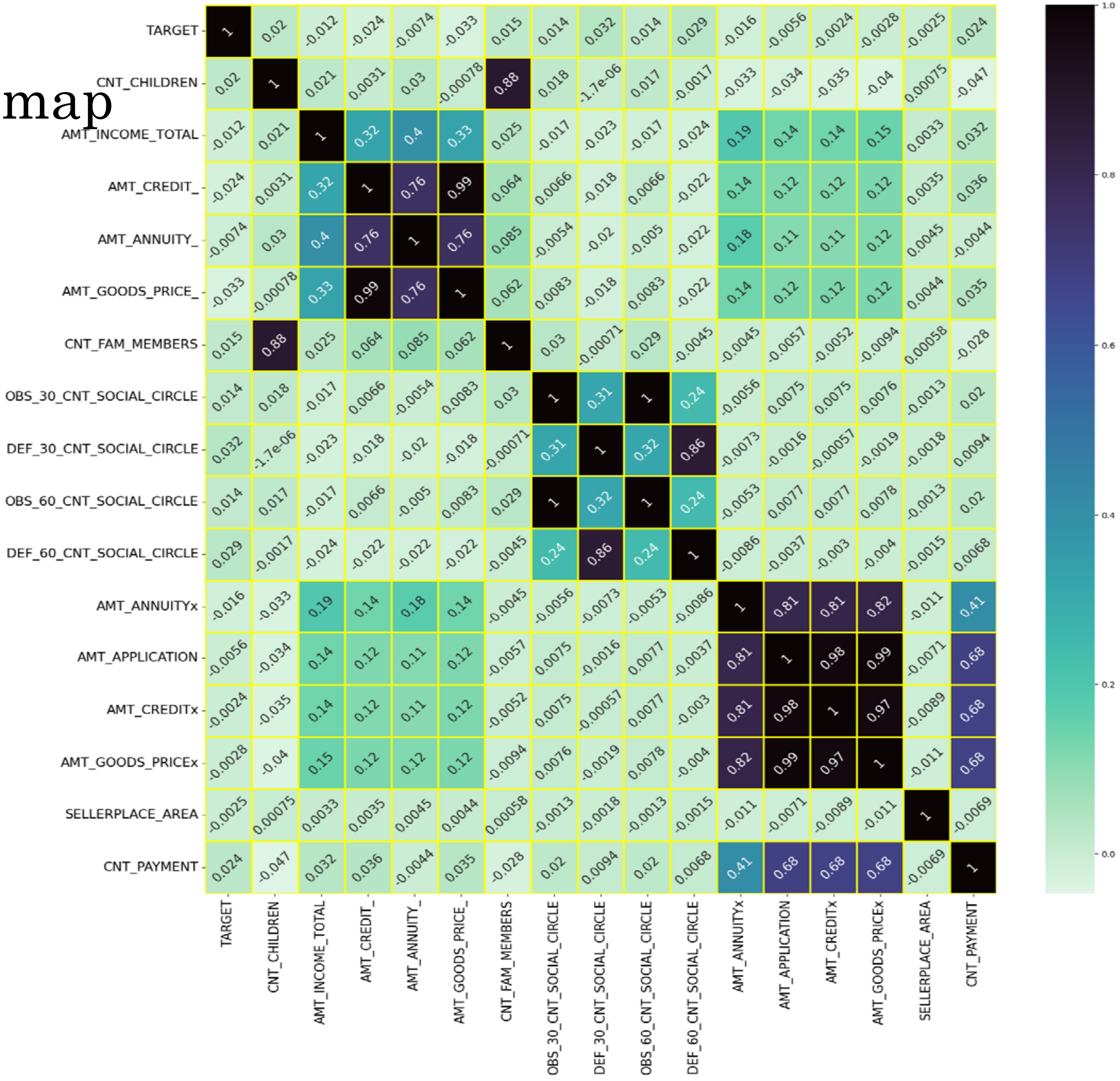
1. Cash X-Sell: low
2. Card X-Sell
3. Card X-Sell: high¶
4. Cash Street: high
5. Cash X-Sell: middle
6. POS household without interest
7. POS industry without interest
8. Cash X-Sell: low
9. Cash Street: middle
10. Cash POS others without interest

Distribution of Income Types w.r.t Contract Status



- Working Category has the maximum counts for all Contract status.
- Student Category has no unused offers

Heat map



	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT_	AMT_ANNUITY_	AMT_GOODS_PRICE_	CNT_FAM_MEMBERS	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	AMT_ANNUITYx	AMT_APPLICATION	AMT_CREDITx	AMT_GOODS_PRICEx	SELLERPLACE_AREA	CNT_PAYMENT
TARGET	1	0.02	-0.012	-0.024	-0.0074	-0.033	0.015	0.014	0.032	0.014	0.029	-0.016	-0.0056	-0.0024	-0.0028	-0.0025	0.024
CNT_CHILDREN	0.02	1	0.021	0.0031	0.03	-0.00078	0.88	0.018	-1.7e-06	0.017	-0.0017	-0.033	-0.034	-0.035	-0.04	0.00075	-0.047
AMT_INCOME_TOTAL	-0.012	0.021	1	0.32	0.4	0.33	0.025	-0.017	-0.023	-0.017	-0.024	0.19	0.14	0.14	0.15	0.0033	0.032
AMT_CREDIT_	-0.024	0.0031	0.32	1	0.76	0.99	0.064	0.0066	-0.018	0.0066	-0.022	0.14	0.12	0.12	0.12	0.0035	0.036
AMT_ANNUITY_	-0.0074	0.03	0.4	0.76	1	0.76	0.085	-0.0054	-0.02	-0.005	-0.022	0.18	0.11	0.11	0.12	0.0045	-0.0044
AMT_GOODS_PRICE_	-0.033	-0.00078	0.33	0.99	0.76	1	0.062	0.0083	-0.018	0.0083	-0.022	0.14	0.12	0.12	0.12	0.0044	0.035
CNT_FAM_MEMBERS	0.015	0.88	0.025	0.064	0.085	0.062	1	0.03	-0.00071	0.029	-0.0045	-0.0045	-0.0057	-0.0052	-0.0094	0.00058	-0.028
OBS_30_CNT_SOCIAL_CIRCLE	0.014	0.018	-0.017	0.0066	-0.0054	0.0083	0.03	1	0.31	1	0.24	-0.0056	0.0075	0.0075	0.0076	-0.0013	0.02
DEF_30_CNT_SOCIAL_CIRCLE	0.032	-1.7e-06	-0.023	-0.018	-0.02	-0.018	-0.00071	0.31	1	0.32	0.86	-0.0073	-0.0016	-0.00057	-0.0019	-0.0018	0.0094
OBS_60_CNT_SOCIAL_CIRCLE	0.014	0.017	-0.017	0.0066	-0.005	0.0083	0.029	1	0.32	1	0.24	-0.0053	0.0077	0.0077	0.0078	-0.0013	0.02
DEF_60_CNT_SOCIAL_CIRCLE	0.029	-0.0017	-0.024	-0.022	-0.022	-0.022	-0.0045	0.24	0.86	0.24	1	-0.0086	-0.0037	-0.003	-0.004	-0.0015	0.0068
AMT_ANNUITYx	-0.016	-0.033	0.19	0.14	0.18	0.14	-0.0045	-0.0056	-0.0073	-0.0053	-0.0086	1	0.81	0.81	0.82	-0.011	0.41
AMT_APPLICATION	-0.0056	-0.034	0.14	0.12	0.11	0.12	-0.0057	0.0075	-0.0016	0.0077	-0.0037	0.81	1	0.98	0.99	-0.0071	0.68
AMT_CREDITx	-0.0024	-0.035	0.14	0.12	0.11	0.12	-0.0052	0.0075	-0.00057	0.0077	-0.003	0.81	0.98	1	0.97	-0.0089	0.68
AMT_GOODS_PRICEx	-0.0028	-0.04	0.15	0.12	0.12	0.12	-0.0094	0.0076	-0.0019	0.0078	-0.004	0.82	0.99	0.97	1	-0.011	0.68
SELLERPLACE_AREA	-0.0025	0.00075	0.0033	0.0035	0.0045	0.0044	0.00058	-0.0013	-0.0018	-0.0013	-0.0015	-0.011	-0.0071	-0.0089	-0.011	1	-0.0069
CNT_PAYMENT	0.024	-0.047	0.032	0.036	-0.0044	0.035	-0.028	0.02	0.0094	0.02	0.0068	0.41	0.68	0.68	0.68	-0.0069	1

Deduction

❑**Highly Positively Correlated Variables.**

❖- AMT_CREDIT vs AMT_GOODS_PRICE

❖- AMT_CREDIT vs AMT_ANNUITY

❖- AMT_ANNUITY vs AMT_GOODS_PRICE

❖- CNT_FAM_MEMBERS vs CNT_CHILDREN

❖-DEF_30_CNT_SOCIAL_CIRCLE vs DEF_60_CNT_SOCIAL_CIRCLE

❖-OBS_30_CNT_SOCIAL_CIRCLE vs OBS_60_CNT_SOCIAL_CIRCLE

❖-AMT_ANNUITYx vs AMT_APPLICATION

❖-AMT_ANNUITYx vs AMT_CREDITx

❖-AMT_ANNUITYx vs AMT_GOODS_PRICEx

❖-AMT_APPLICATION vs AMT_CREDITx

❖-AMT_APPLICATION vs AMT_GOODS_PRICEx

❖-AMT_GOODS_PRICEx vs AMT_CREDITx

❖-CNT_PAYMENT vs AMT_APPLICATION

❖-CNT_PAYMENT vs AMT_CREDITx

❖-CNT_PAYMENT vs AMT_GOODS_PRICEx

❑**Highly Negatively Correlated Variables.**

❖- There are no Highly Negatively Correlated Variables.

Conclusion

After finishing analysis on the Bank dataset, it can be concluded that there are factors related to the customer with which the bank would be able to predict both Repayers and Defaulters. Those contributing factors can be summarized as below.

A. Repayers:

- ➢Gender: Females are less likely to default.
- ➢Education Type: Secondary/secondary special Education followed bt Higher Education has more Repayers
- ➢Income Type: Maximum repaying customers are under working catagory.
- ➢Housing Type: Maximum repaying customers owning House/Apartment followed by customers living with parents
- ➢Age: Customers in Age range of 30-40 are very less likely to default.
- ➢Total Income: Maximum customers without any repaying difficulties are having a salary range of 1L - 2L.
- ➢Family Status: Maximum customers without any repaying difficulties are married.

A. Defaulter:

1) Gender: Men default at a relatively higher rate.
2) Education Type: Customers with Lower Secondary education have a slightly higher defaulting rate compared to same category repayers.
3) Income Type: Customers who are either at Maternity leave OR Unemployed tends to default more.
4) Age: Customers in Age range below 30 are very likely to default.
5) Family Status: Customers other than married are likely to default.

B) Other Factors Noticed during Analysis.

1) Credit amount for those who don't have payment difficulties is higher than those with payment difficulties.
2) AMT_CREDIT and AMT_GOODS_PRICE are highly correlated. Customers with higher goods price and don't have payment difficulties have higher credit amount than those with higher goods price but are having payment difficulties.
3) Most of the defaulters are having Less than 5L total income.

THANK YOU