EDA CASE STUDY

Vikramjit Bora and Anjali Agarwal

Upgrad C28 January batch

5th May 2021

Problem Statement:

This is a Credit Risk analysis problem from Banking and Financial Services domain.

Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. Hence, you need to understand how data is used to minimize the risk of losing money while lending to customers. You also need to analyse how customer attributes and loan attributes influence the tendency of default. The company can utilize this knowledge for its portfolio and risk assessment.

Analysis Approach:

Below are the steps followed for analysis of this case study:

First performed the above steps on applications data, then performed the similar steps on previous data and then finally for the merged dataset.

1.Read the application dataset

Performed a non-graphical EDA

2. Data Cleaning

- Missing Value Analysis – removed columns with more than 45% data missing
- Removed rows where columns had less than 1% null values
- Imputed missing values using Central Tendency
- Removed rows which have junk data in few columns
- In order to better understand and analyze the data, added a few derived columns such as converted 'days since birth' to 'years since birth'(DAYS_BIRTH to YEARS_BIRTH)
- Outlier Analysis

3. Checked and corrected the data types

4. Binned continuous variables

5. Checked the Target variable

6. Performed Univariate Analysis(one variable at a time) on Categorical and Numerical variables

7. Performed Bivariate Analysis(2 variables at a time)

- Categorical vs Numerical variables
- Categorical vs Categorical Variables
- Numerical vs Numerical Variables

8. Checked Correlations

Top 10 correlation pairs

9. Performed Multi-variate analysis(more than 2 variables at a time)

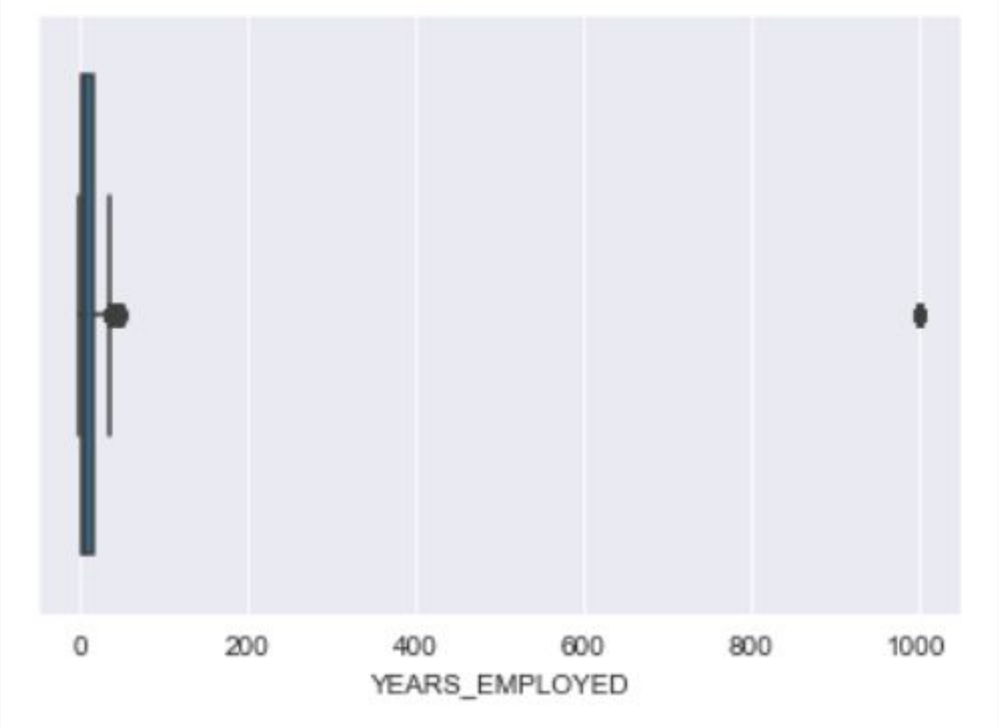
10. Derived final conclusions

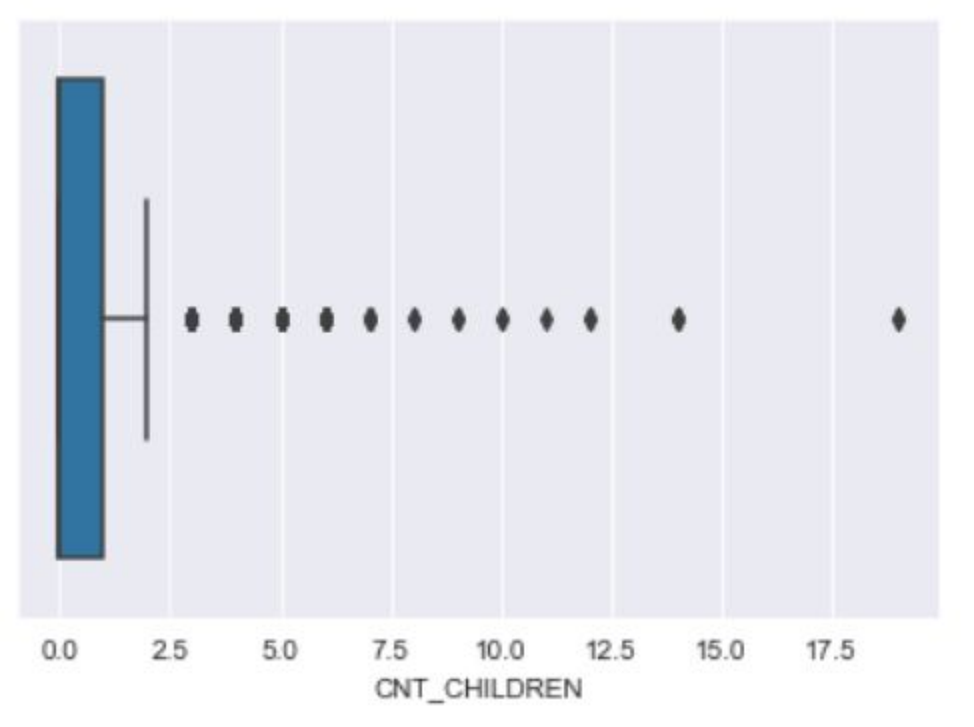
Analysis on Applications Data

Outlier treatment

Insights:

- The value of 1000 is an outlier for the YEARS_EMPLOYED column. Since this is not possible, we may impute the value with the median.
- 99.9 percentile of Number of children of clients is at 4. Any number above 4 is an unlikely event and can be considered as an outlier.

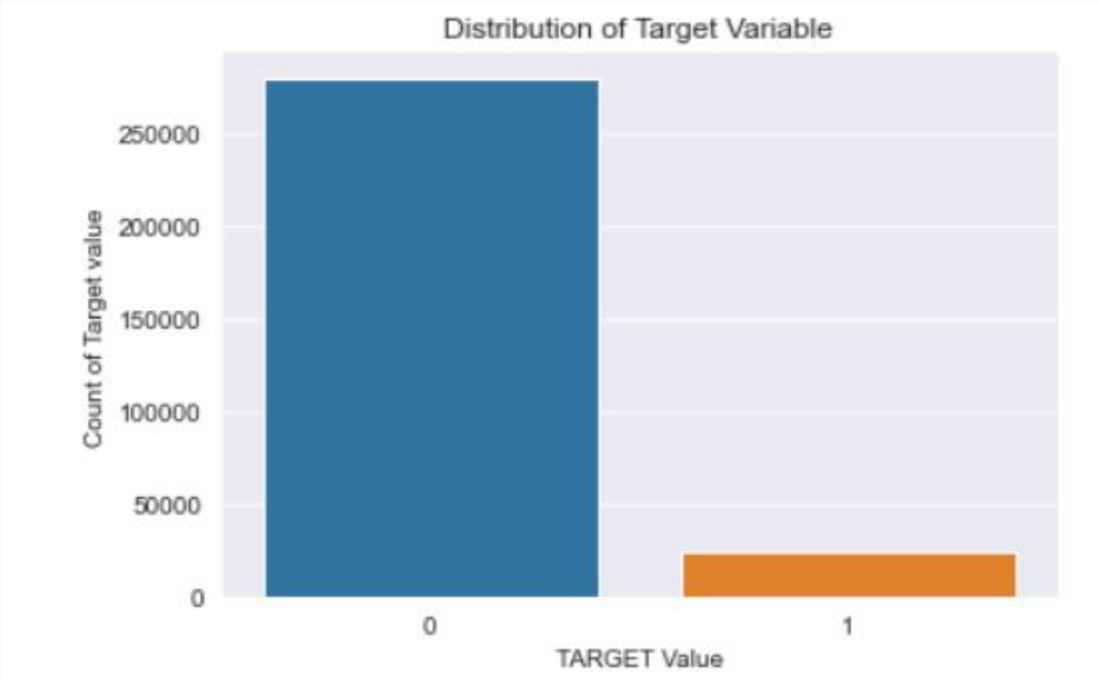




Checking the Target Variable

Insights:

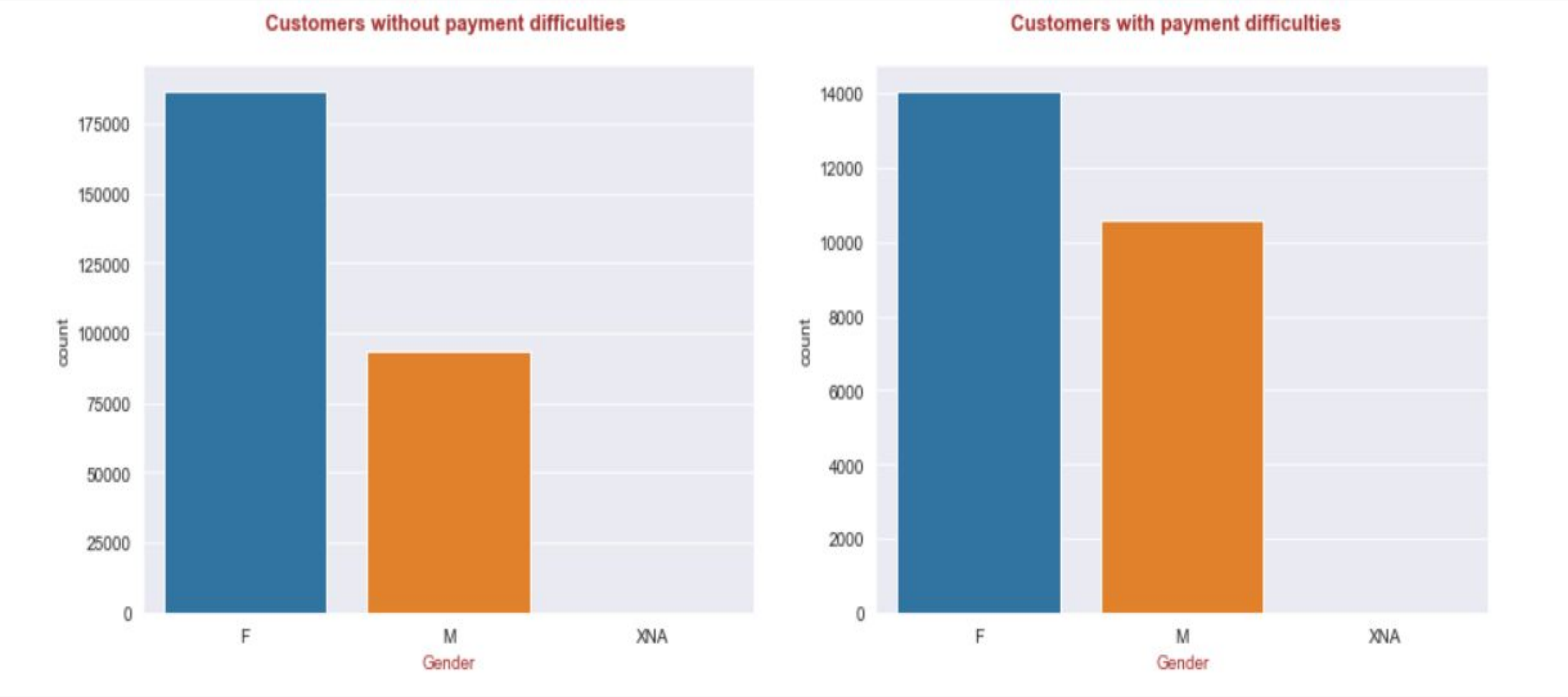
- The ratio of customers not facing loan repayment difficulties vs those facing loan repayment difficulties is 11.33:1
- The data is highly imbalanced with approximately only 8% of the data being for clients with payment difficulties. So in order to analyze the data better, the application dataset has been into two different data frames based on target variable's value.



Comparing Payment Difficulties vs Non Payment Difficulties on the basis of Gender

Insights:

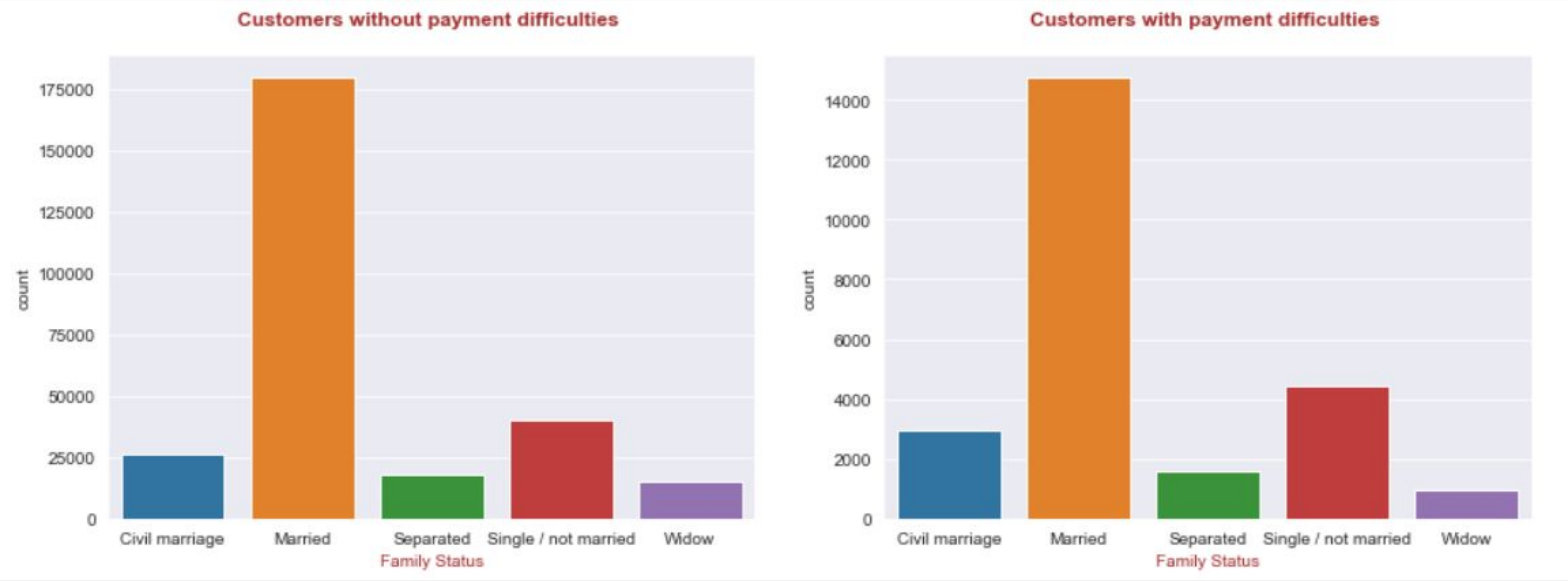
Females are the majority in both the cases although there is an increase in the percentage in Male Payment Difficulties compared to Non-Payment Difficulties



Comparing the Payment Difficulties and Non Payment Difficulties on the basis of Family status

Insights:

Married couples are the majority in both the cases.



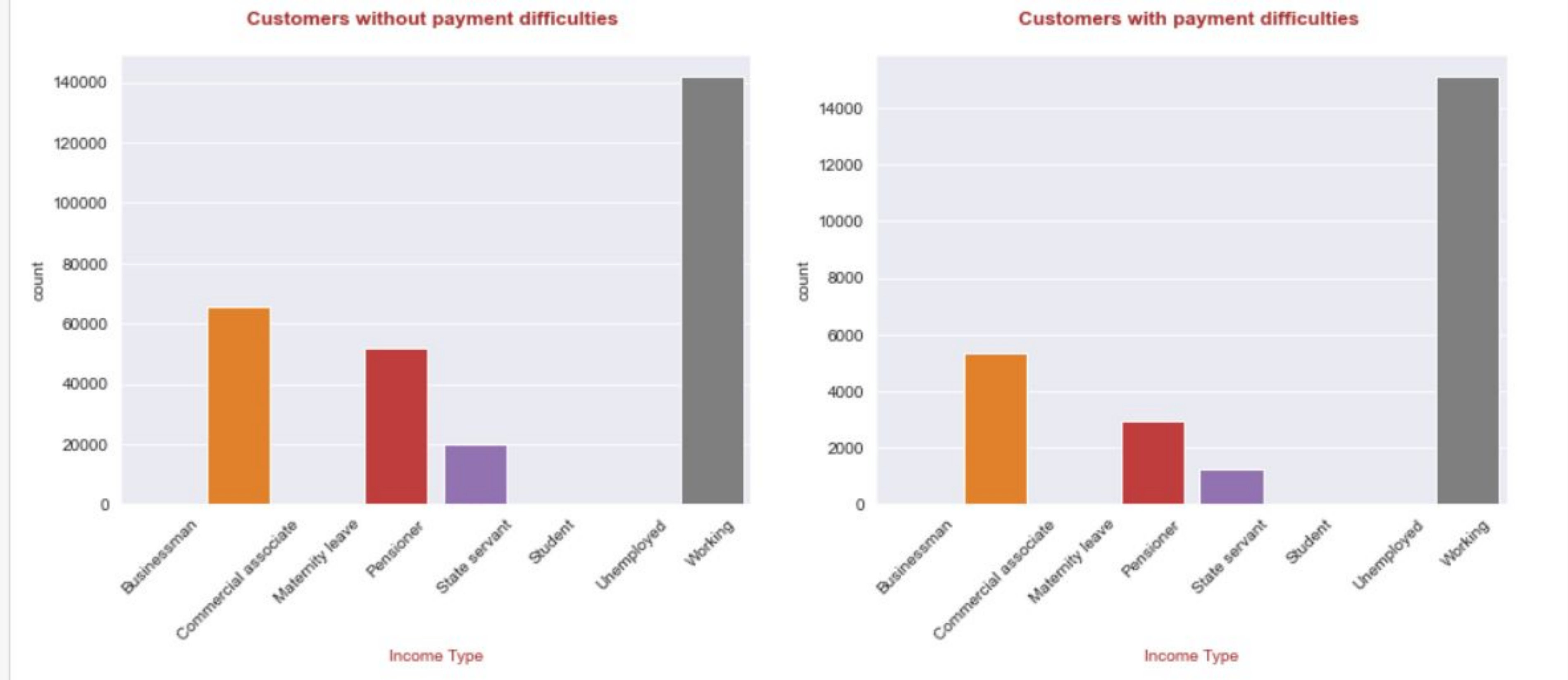
Customers without payment difficulties

Customers with payment difficulties

Comparing the Payment Difficulties and Non Payment Difficulties on the basis of Income Type

Insights:

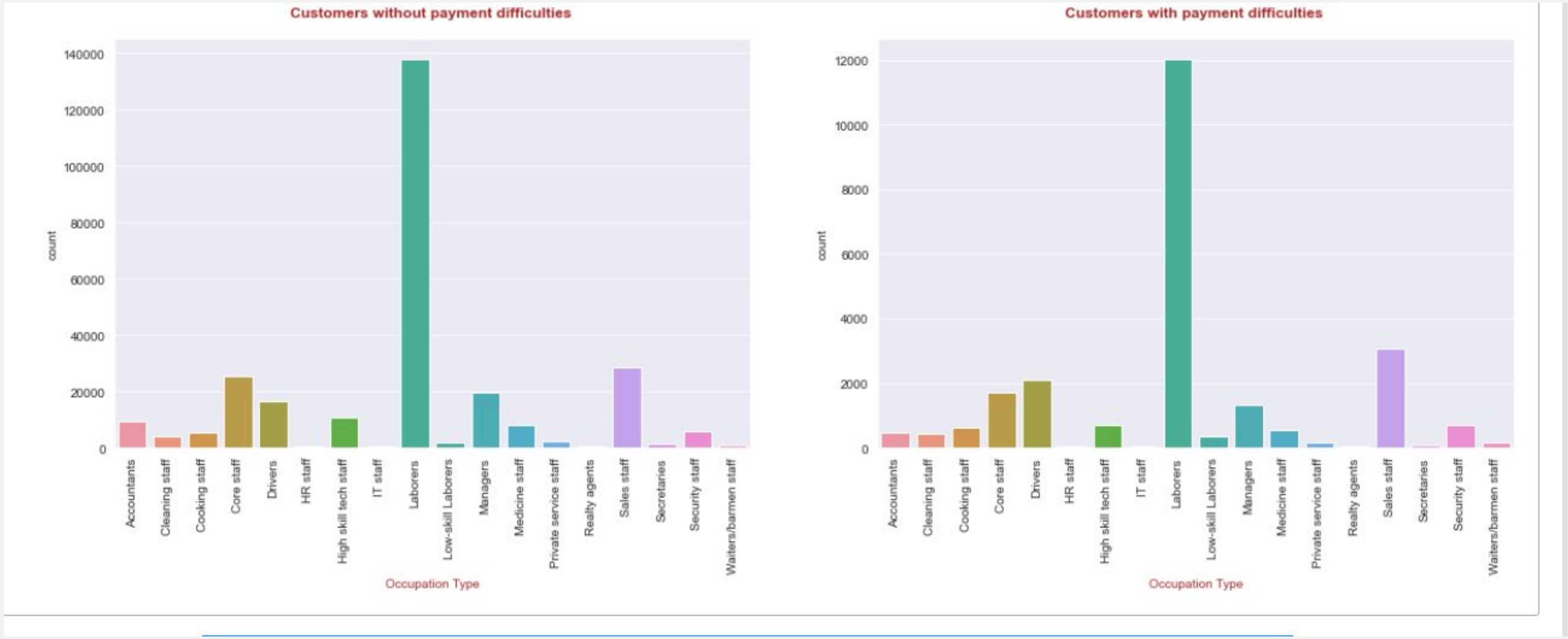
Working people are the majority in both the cases.



Comparing the Payment Difficulties and Non Payment Difficulties on the basis of Occupation Type

Insights:

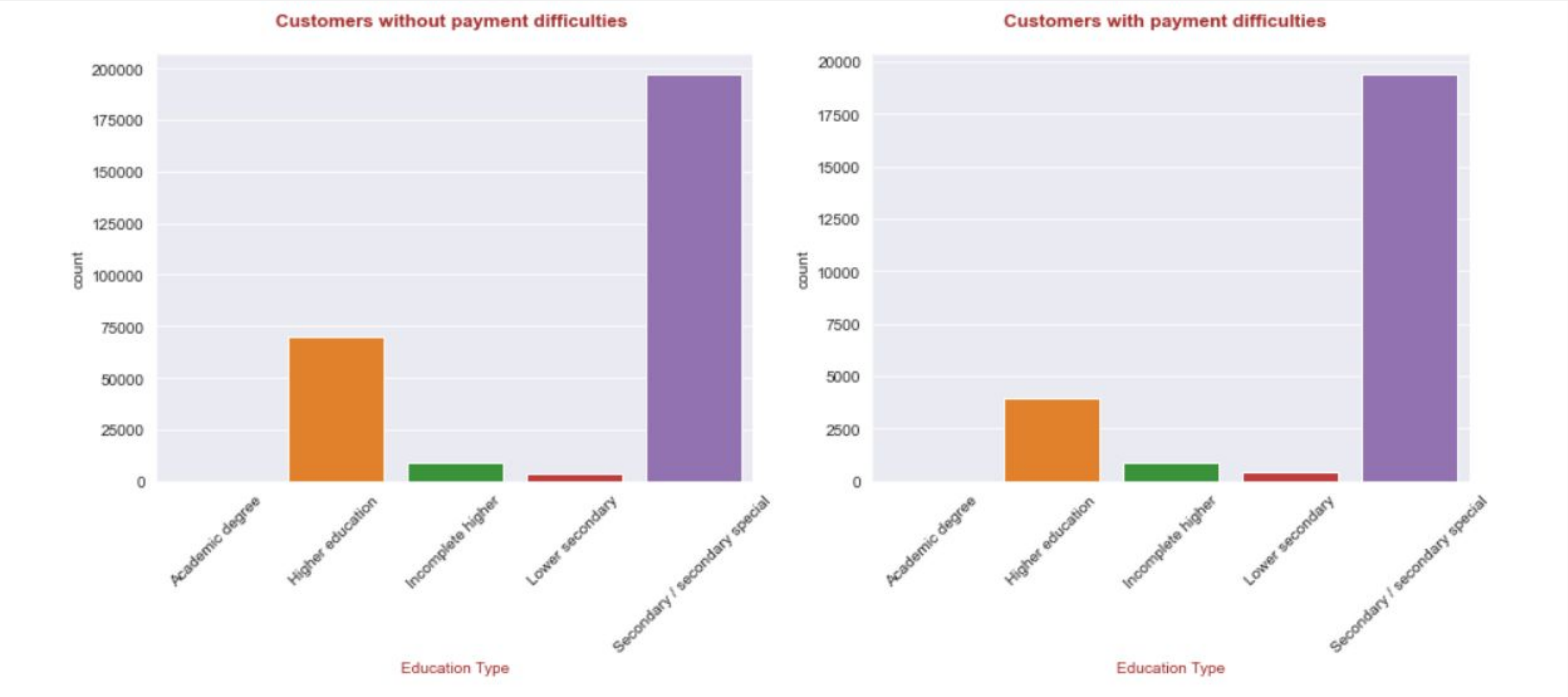
Laborers are having more difficulties in repaying the loan. Sales staff, Core staff, Drivers also have payment difficulties.



Comparing the Payment Difficulties and Non Payment Difficulties on the basis of Education Type

Insights:

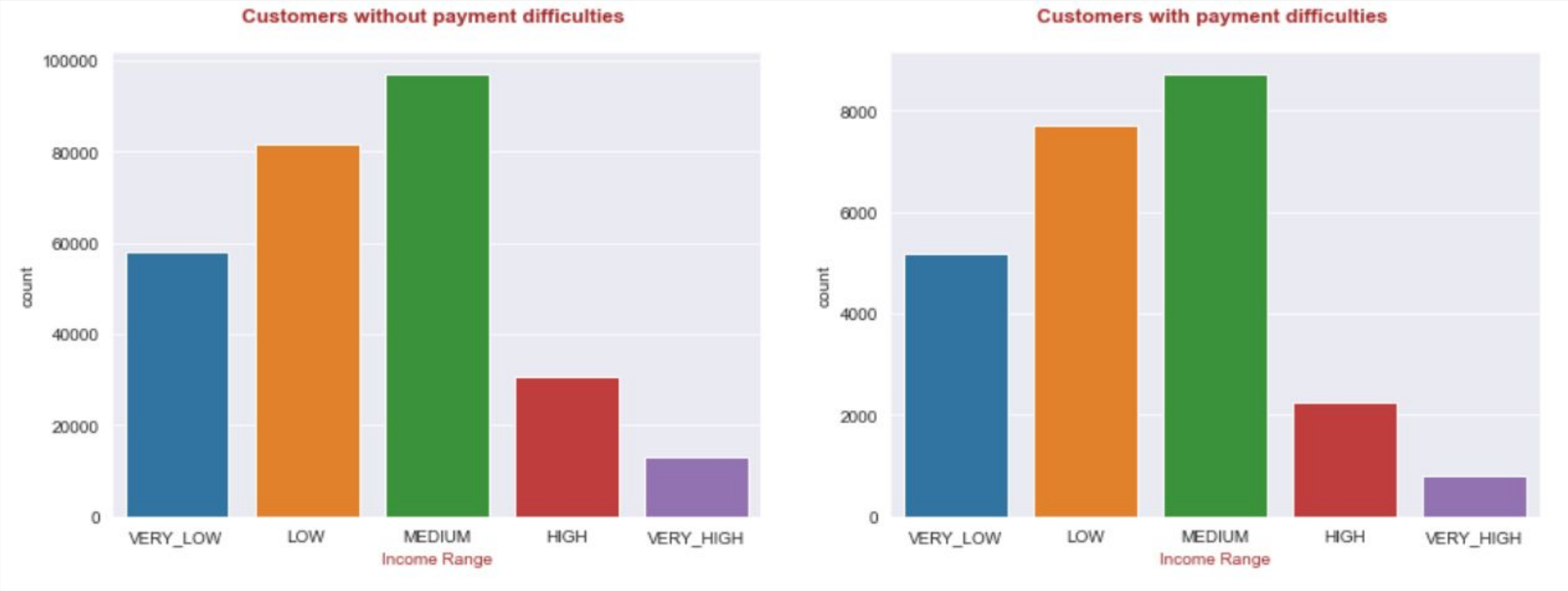
Customers who have completed higher education comparatively have less difficulty in loan repayment.



Comparing the Payment Difficulties and Non Payment Difficulties on the basis of Income Range

Insights:

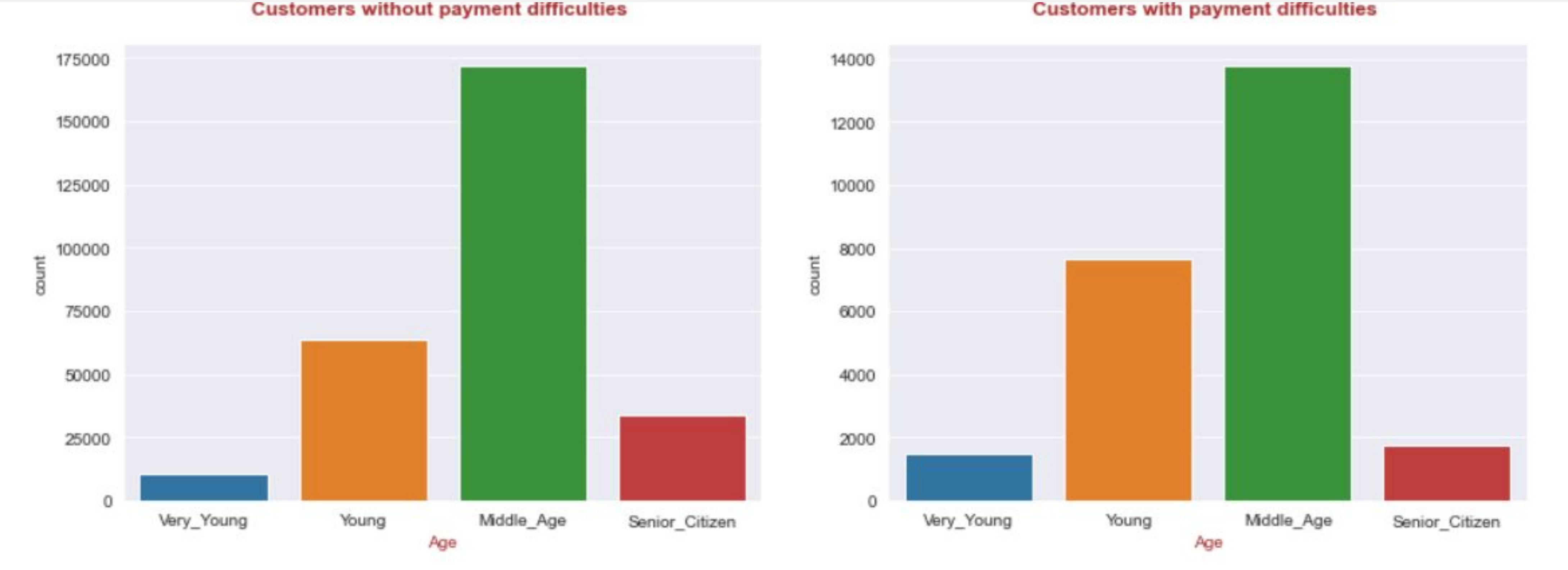
There is an increase in the proportion of Loan Payment Difficulties for whose income is low or very low as compared to the corresponding proportion for Non Payment Difficulties



Comparing the Payment Difficulties and Non Payment Difficulties on the basis of Age categories

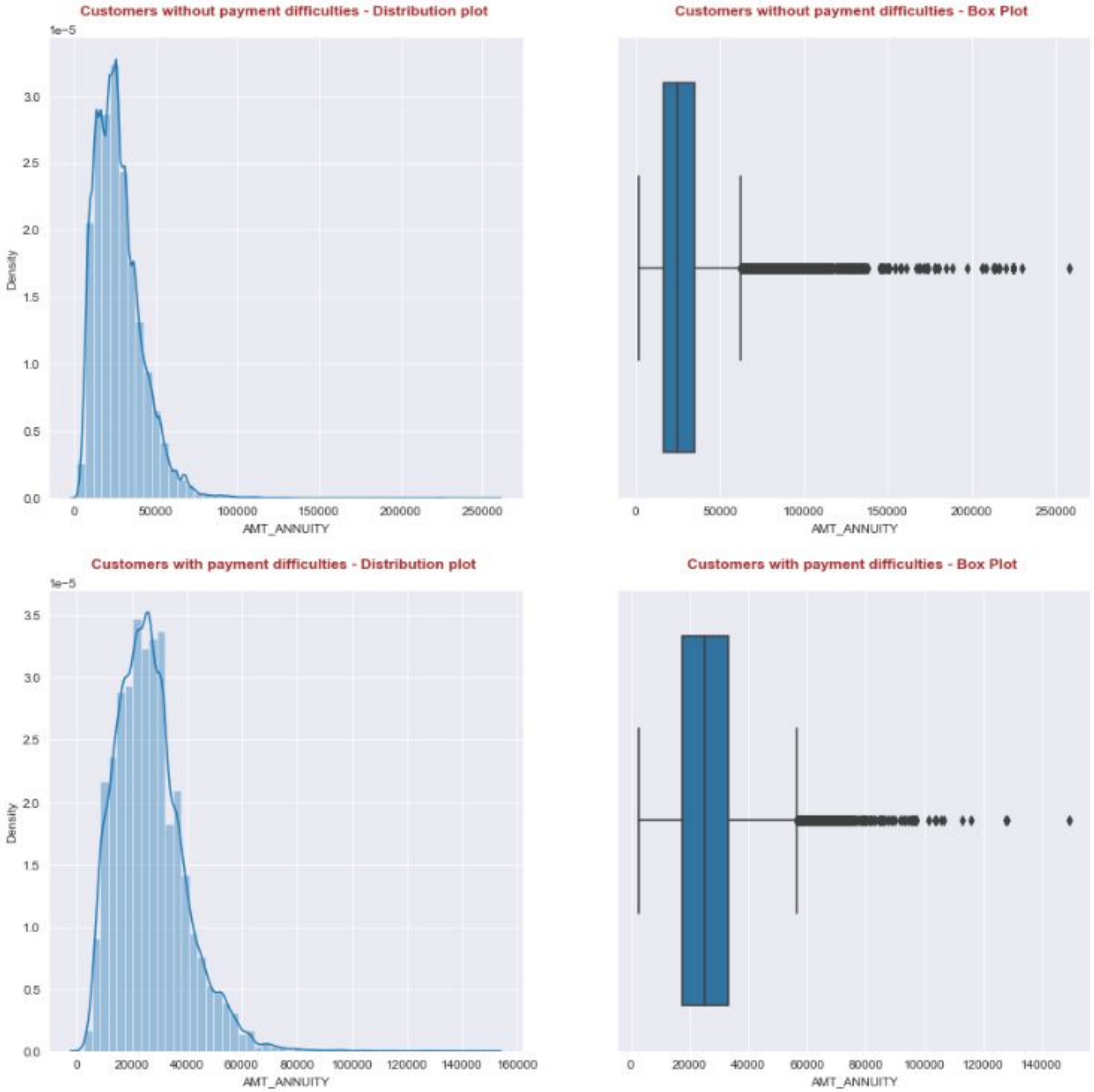
Insights:

Middle Aged customers have a considerably higher proportion of Loan Payment Difficulties



Univariate Analysis:

Comparing the Payment Difficulties and Non Payment Difficulties on the basis of Annuity Amount

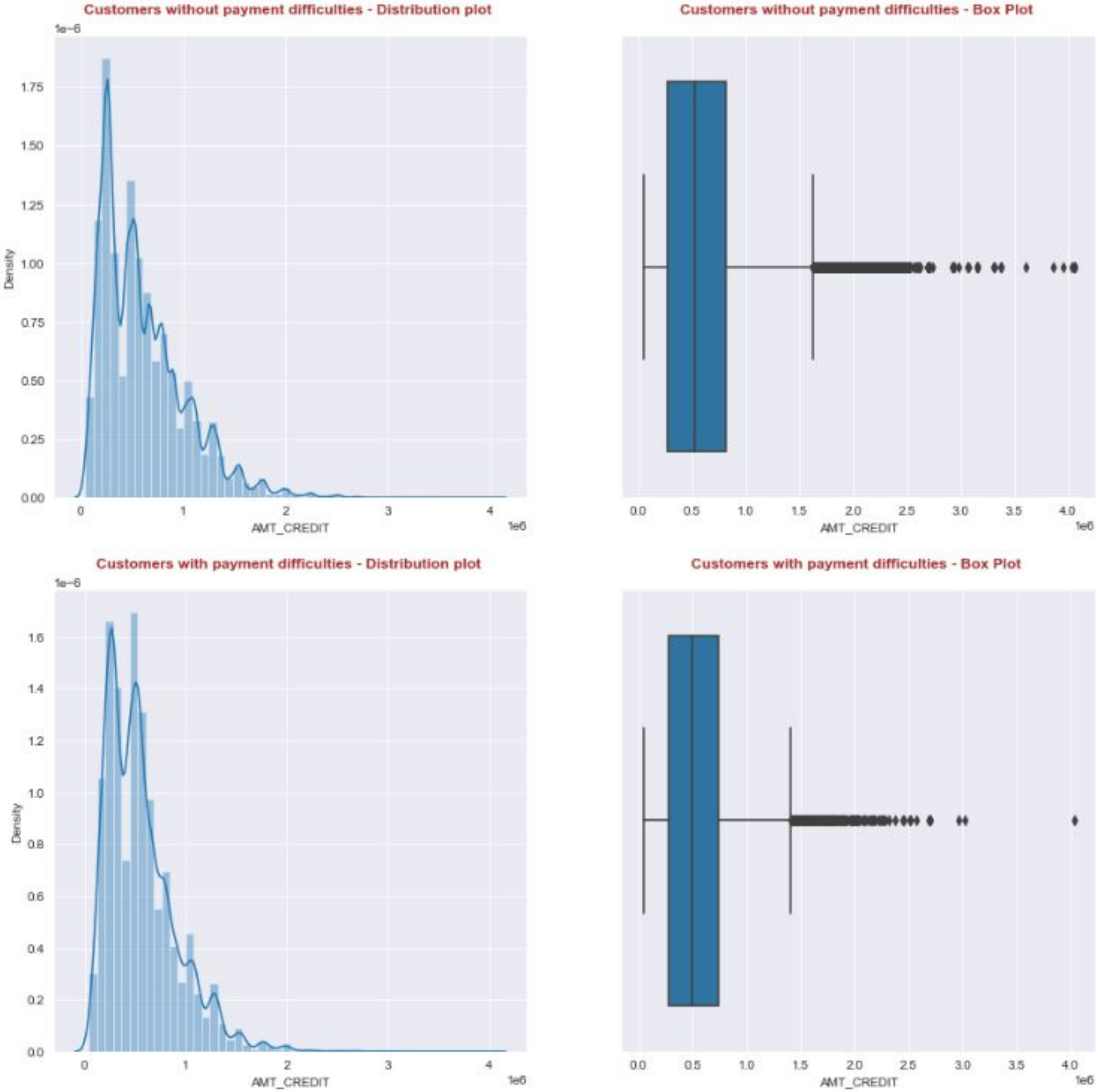


Insights:

The distribution of AMT_ANNUITY for both target 0 and 1 is right-skewed and have some outliers.

Univariate Analysis:

Comparing the Payment Difficulties and Non Payment Difficulties on the basis of credit amount

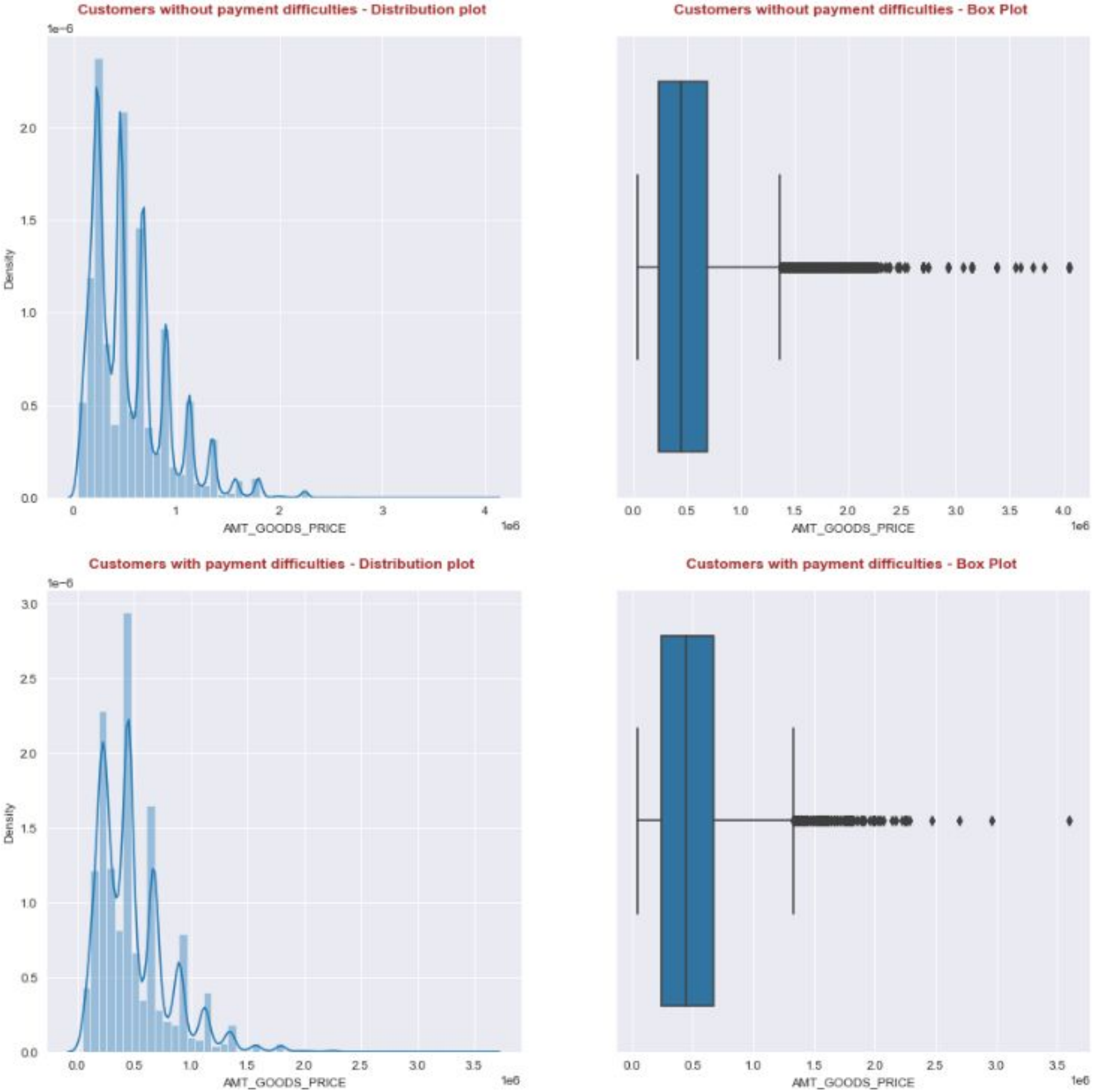


Insights:

The distribution of AMT_CREDIT for both target 0 ar 1 is right-skewed and have some outliers.

Univariate Analysis:

Comparing the Payment Difficulties and Non Payment Difficulties on the basis of goods prices



Insights:

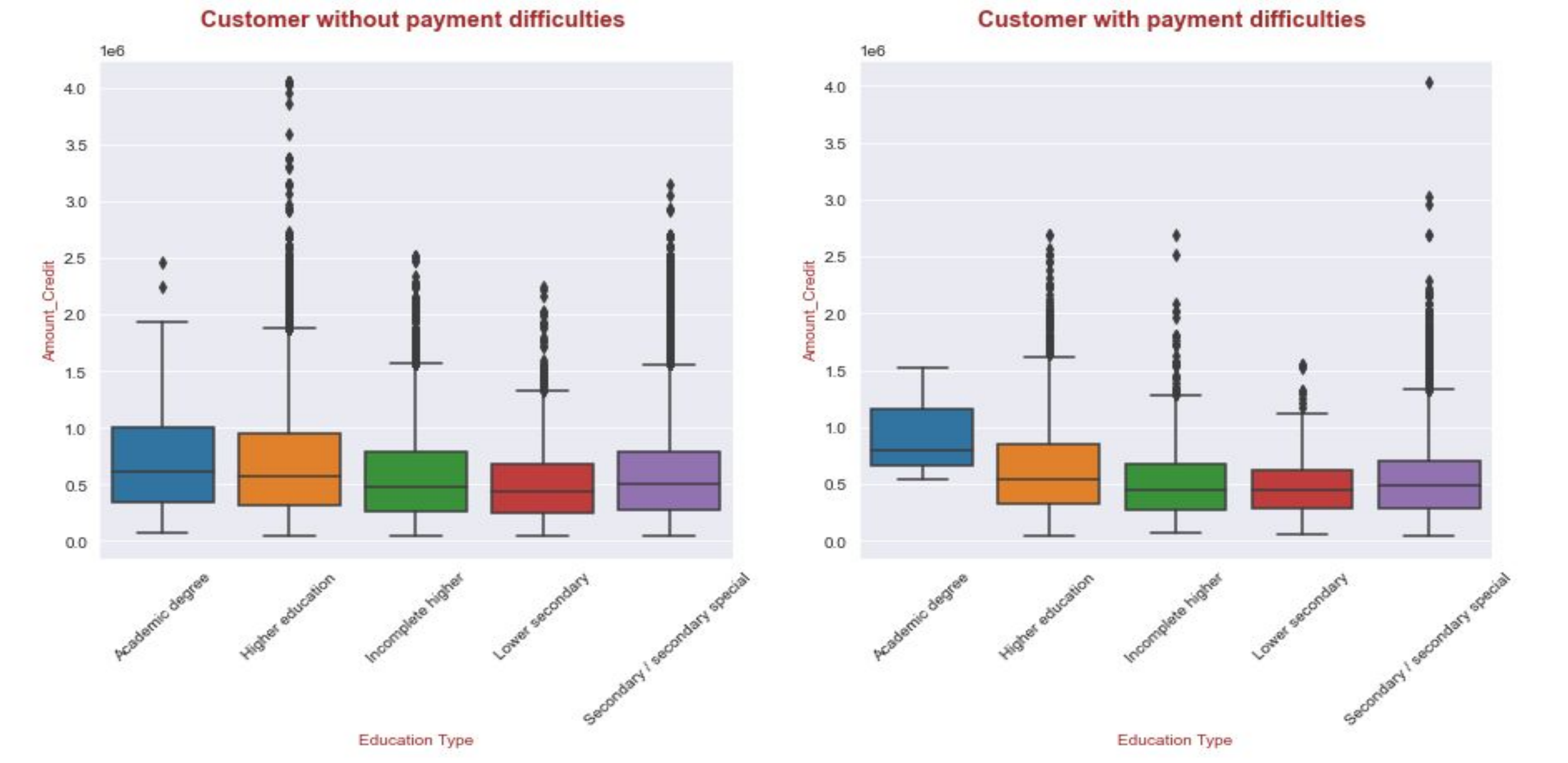
The distribution of AMT_GOODS_PRICE for both target 0 and 1 is right-skewed and have some outliers.

Bivariate Analysis:

Comparing the Payment Difficulties and Non Payment Difficulties on the basis of Education Type

Insights:

- The median for credit amount increases from lower secondary to Academic degree education types for customers without payment difficulties.
- The median credit amount for customers with academic degree is higher than other education types where customers are facing difficulty in loan payment.

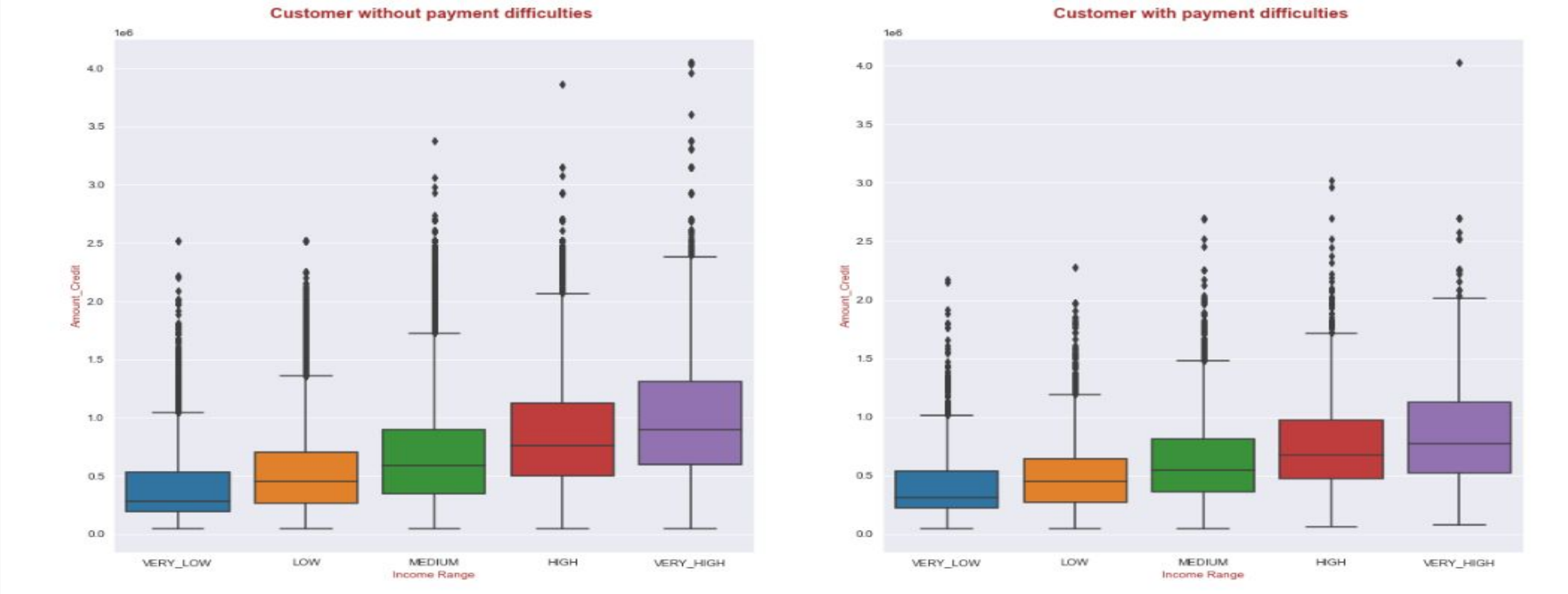


Bivariate Analysis:

Comparing the Payment Difficulties and Non Payment Difficulties on the basis of Income range

Insights:

- As income increases, loan credit amount also increases for both customers with and without payment difficulties.

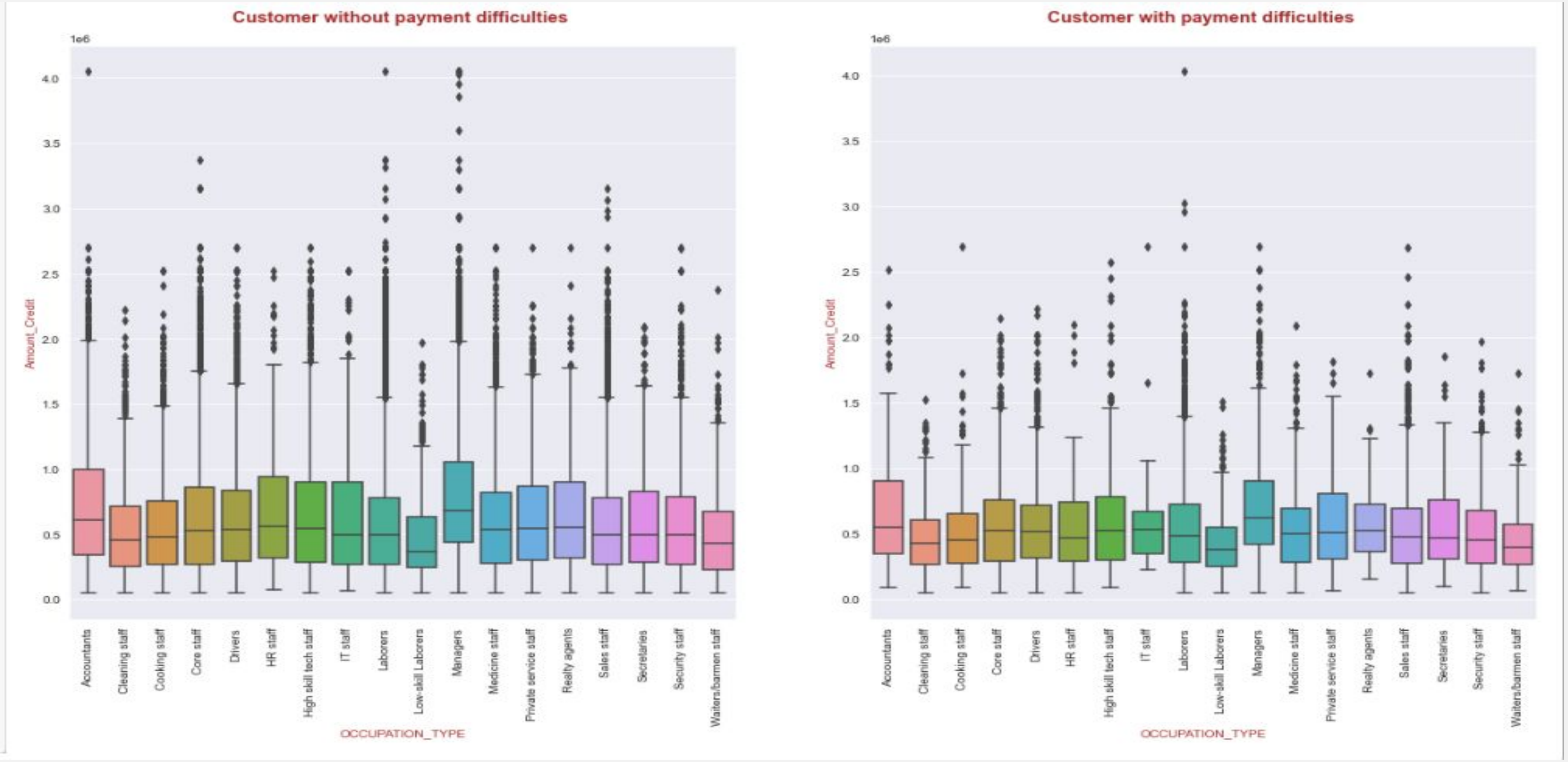


Bivariate Analysis:

Comparing the Payment Difficulties and Non Payment Difficulties on the basis of Occupation type

Insights:

The range of the customers without payment difficulties is more compared to the customers with payment difficulties.



Customer without payment difficulties

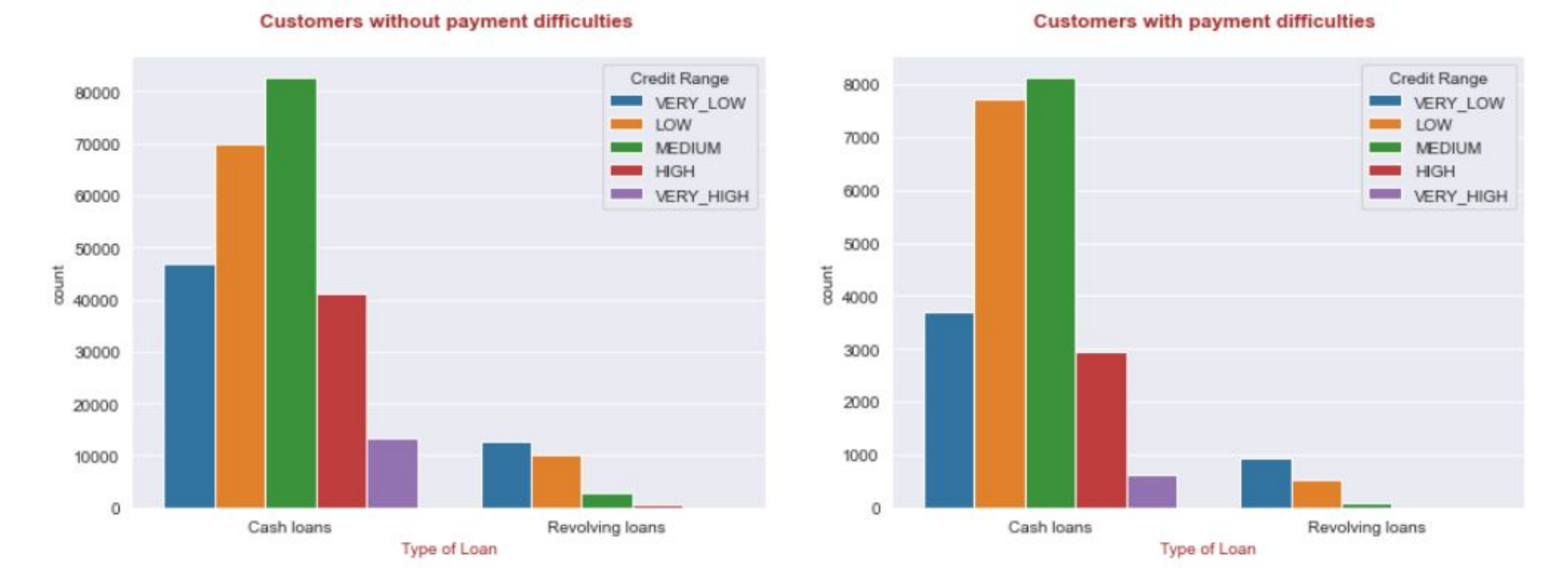
Customer with payment difficulties

Categorical vs Categorical Analysis:

NAME_CONTRACT_TYPE
vs
AMT_CREDIT_RANGE

Insights:

Most number of cash loans is in Medium Credit range for both customers with and without payment difficulties.
However, Low credit range for cash loans is very close to number of medium credit range loans for customers with payment difficulties.



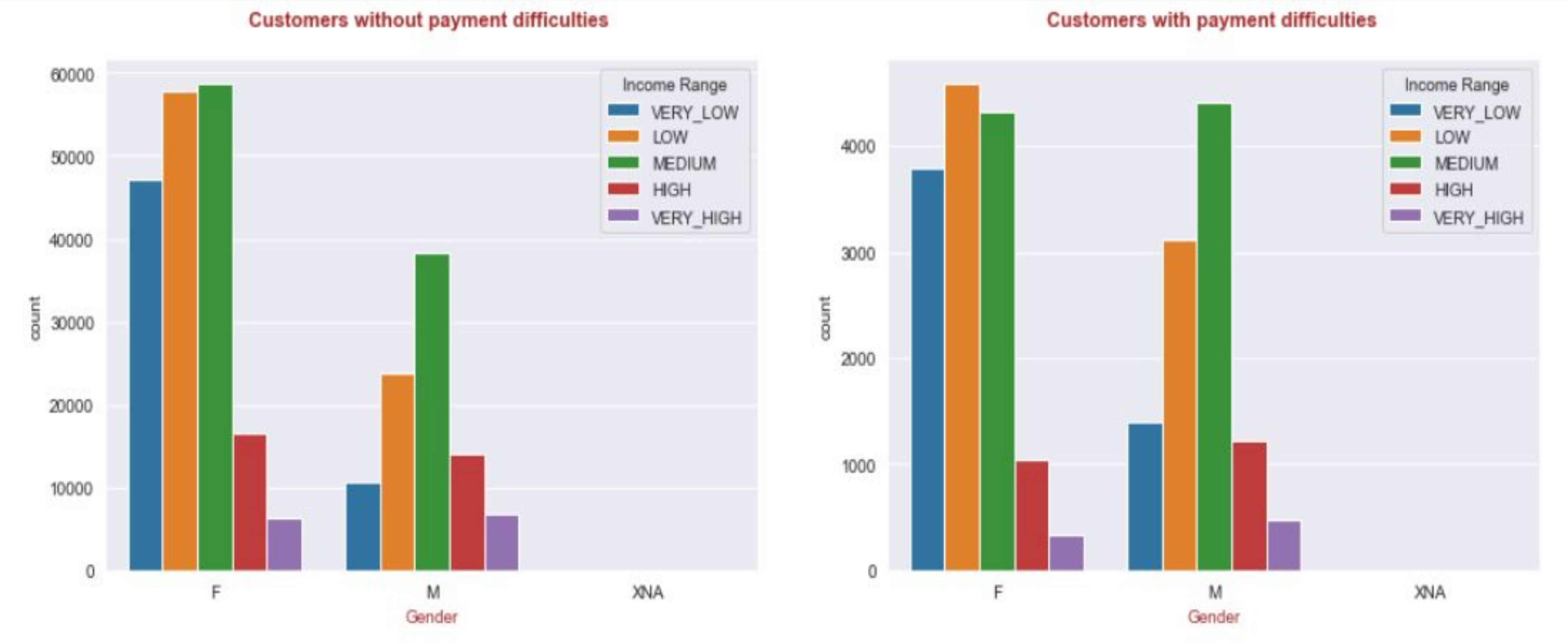
Categorical vs Categorical Analysis:

CODE_GENDER
vs
AMT_INCOME_RANGE

Insights:

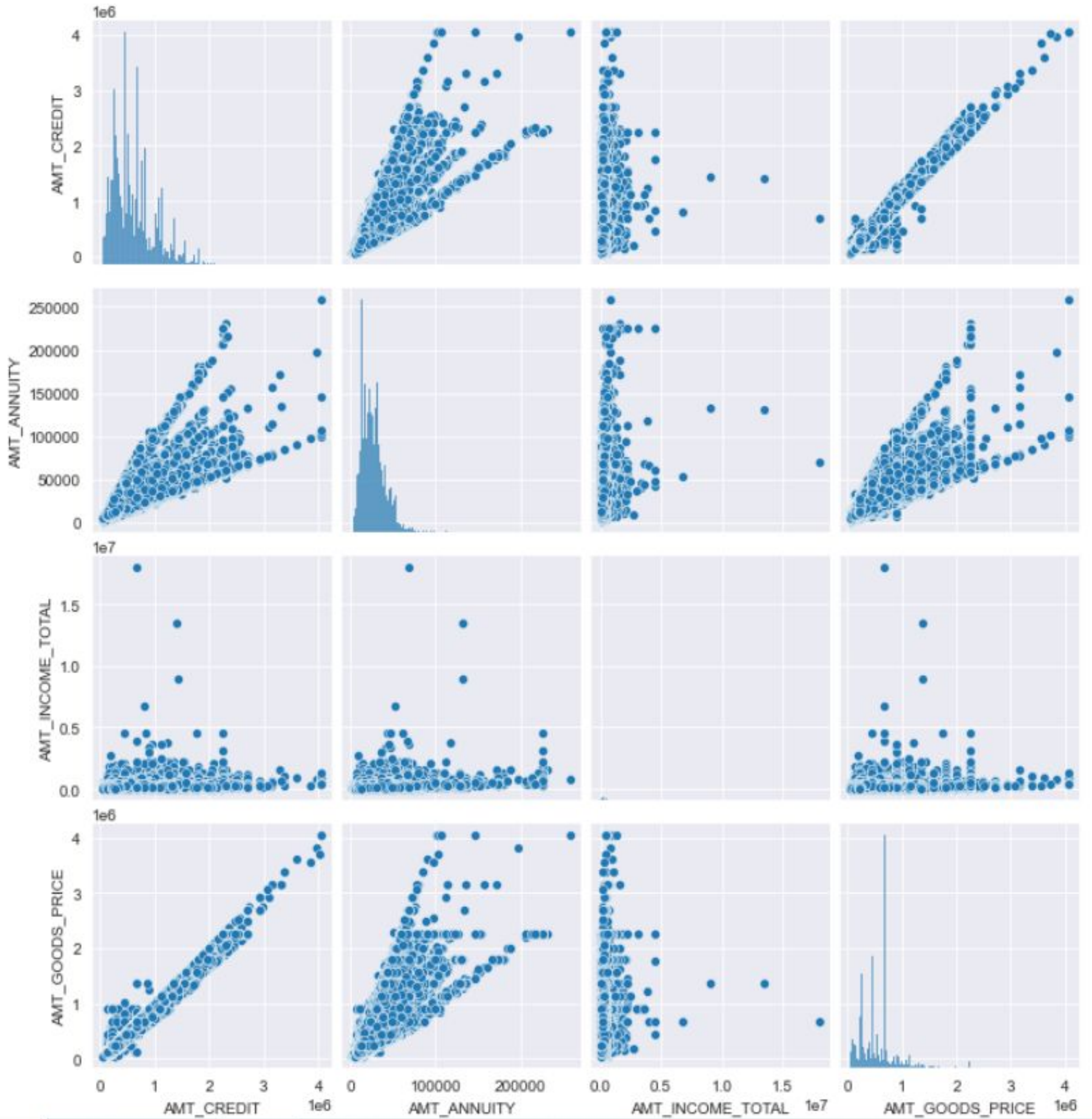
Similar number of males with medium income range and females with medium or low income range have difficulty in loan payment.

However, more females in the medium, low and very low income range have no payment difficulty compared to males in same income range.



Numerical vs Numerical variables Analysis:

Comparing 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_INCOME_TOTAL', 'AMT_GOODS_PRICE' variables for the customers without Payment Difficulties

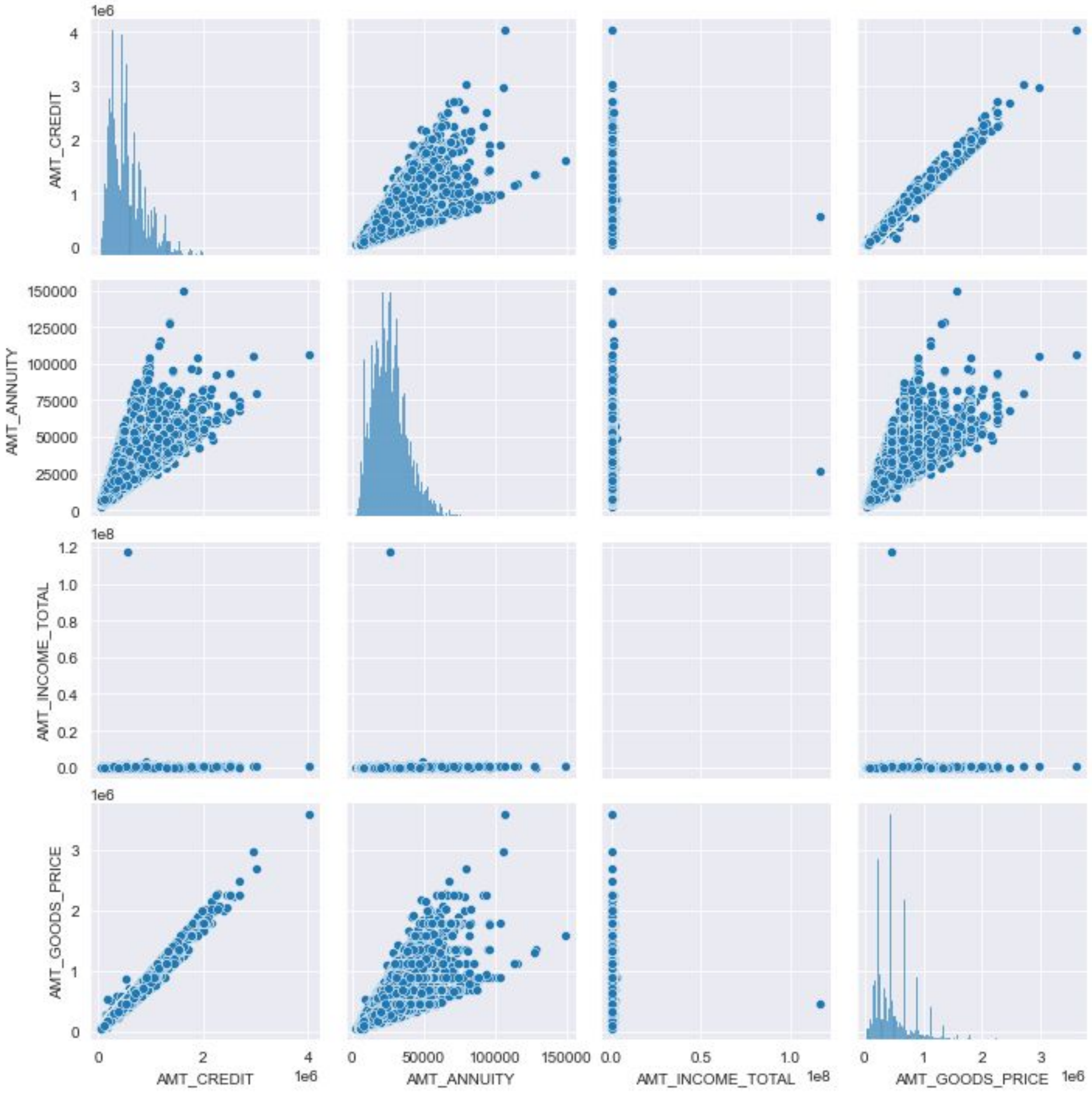


Insights:

- Annuity amount has a positive correlation with credit amount.
- Goods price has a strong positive linear relationship with credit amount.
- Goods price has a positive correlation with annuity amount.

Numerical vs Numerical variables Analysis:

Comparing 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_INCOME_TOTAL', 'AMT_GOODS_PRICE’ variables for the customers with Payment Difficulties



Insights:

Similar insights as customers without payment difficulty:

- Annuity amount has a positive correlation with credit amount.
- Goods price has a strong positive linear relationship with credit amount.
- Goods price has a positive correlation with annuity amount.

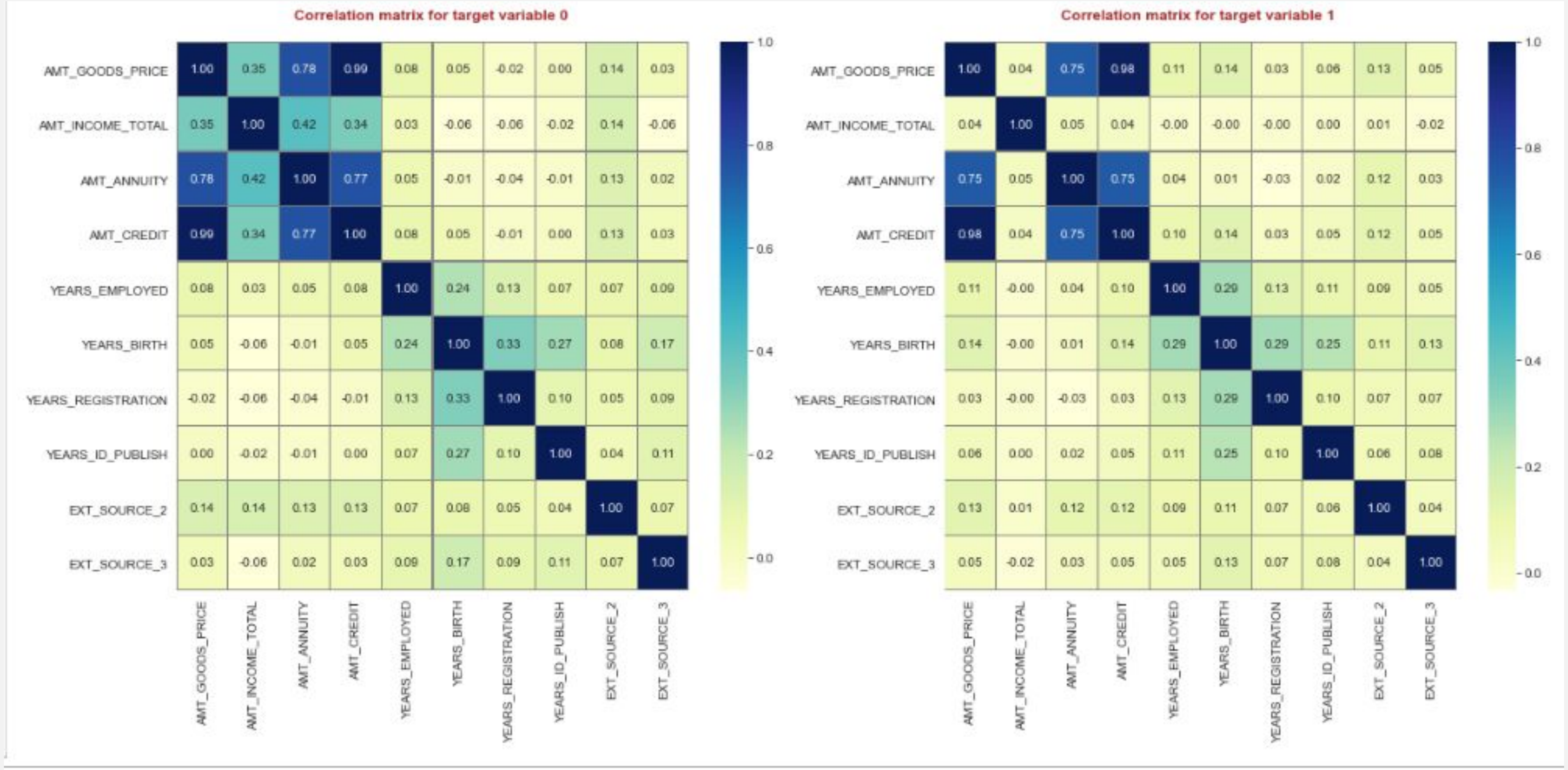
Correlation:

Correlating the numerical variables for customers with Payment Difficulties vs those with No Payment Difficulties

Insights:

The below 3 pairs are highly correlated:

- AMT_GOODS_PRICE and AMT_CREDIT
- AMT_GOODS_PRICE and AMT_ANNUITY
- AMT_ANNUITY and AMT_CREDIT



Correlation matrix for target variable 0

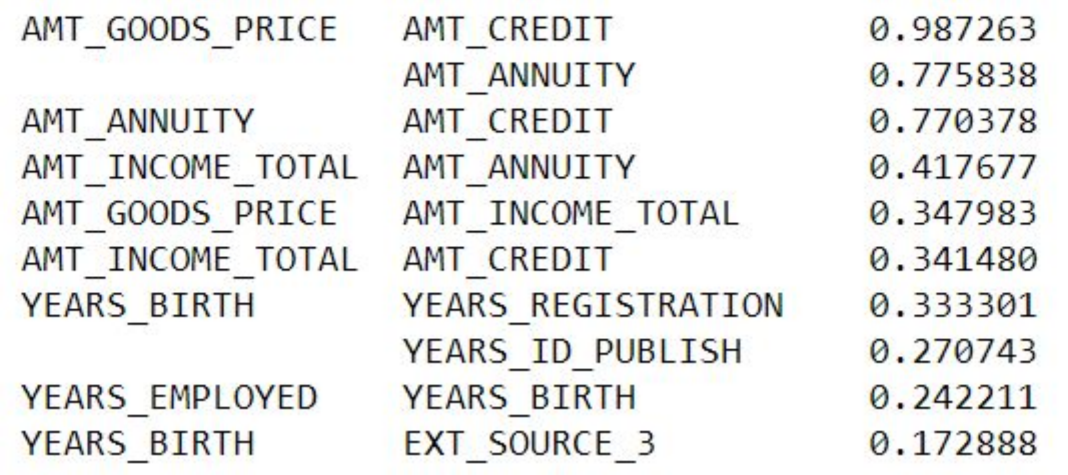
	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	AMT_ANNUITY	AMT_CREDIT	YEARS_EMPLOYED	YEARS_BIRTH	YEARS_REGISTRATION	YEARS_ID_PUBLISH	EXT_SOURCE_2	EXT_SOURCE_3
AMT_GOODS_PRICE	1.00	0.35	0.78	0.99	0.08	0.05	-0.02	0.00	0.14	0.03
AMT_INCOME_TOTAL	0.35	1.00	0.42	0.34	0.03	-0.06	-0.06	-0.02	0.14	-0.06
AMT_ANNUITY	0.78	0.42	1.00	0.77	0.05	-0.01	-0.04	-0.01	0.13	0.02
AMT_CREDIT	0.99	0.34	0.77	1.00	0.08	0.05	-0.01	0.00	0.13	0.03
YEARS_EMPLOYED	0.08	0.03	0.05	0.08	1.00	0.24	0.13	0.07	0.07	0.09
YEARS_BIRTH	0.05	-0.06	-0.01	0.05	0.24	1.00	0.33	0.27	0.08	0.17
YEARS_REGISTRATION	-0.02	-0.06	-0.04	-0.01	0.13	0.33	1.00	0.10	0.05	0.09
YEARS_ID_PUBLISH	0.00	-0.02	-0.01	0.00	0.07	0.27	0.10	1.00	0.04	0.11
EXT_SOURCE_2	0.14	0.14	0.13	0.13	0.07	0.08	0.05	0.04	1.00	0.07
EXT_SOURCE_3	0.03	-0.06	0.02	0.03	0.09	0.17	0.09	0.11	0.07	1.00

Correlation matrix for target variable 1

	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	AMT_ANNUITY	AMT_CREDIT	YEARS_EMPLOYED	YEARS_BIRTH	YEARS_REGISTRATION	YEARS_ID_PUBLISH	EXT_SOURCE_2	EXT_SOURCE_3
AMT_GOODS_PRICE	1.00	0.04	0.75	0.98	0.11	0.14	0.03	0.06	0.13	0.05
AMT_INCOME_TOTAL	0.04	1.00	0.05	0.04	-0.00	-0.00	-0.00	0.00	0.01	-0.02
AMT_ANNUITY	0.75	0.05	1.00	0.75	0.04	0.01	-0.03	0.02	0.12	0.03
AMT_CREDIT	0.98	0.04	0.75	1.00	0.10	0.14	0.03	0.05	0.12	0.05
YEARS_EMPLOYED	0.11	-0.00	0.04	0.10	1.00	0.29	0.13	0.11	0.09	0.05
YEARS_BIRTH	0.14	-0.00	0.01	0.14	0.29	1.00	0.29	0.25	0.11	0.13
YEARS_REGISTRATION	0.03	-0.00	-0.03	0.03	0.13	0.29	1.00	0.10	0.07	0.07
YEARS_ID_PUBLISH	0.06	0.00	0.02	0.05	0.11	0.25	0.10	1.00	0.06	0.08
EXT_SOURCE_2	0.13	0.01	0.12	0.12	0.09	0.11	0.07	0.06	1.00	0.04
EXT_SOURCE_3	0.05	-0.02	0.03	0.05	0.05	0.13	0.07	0.08	0.04	1.00

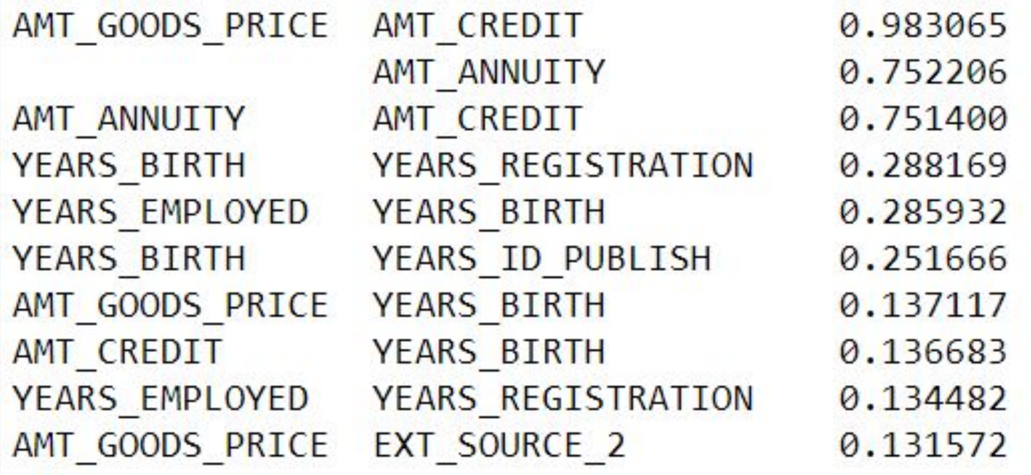
Top 10 Correlations:

For customers without payment difficulties:



```
AMT_GOODS_PRICE    AMT_CREDIT             0.987263
                   AMT_ANNUITY            0.775838
AMT_ANNUITY        AMT_CREDIT             0.770378
AMT_INCOME_TOTAL   AMT_ANNUITY            0.417677
AMT_GOODS_PRICE    AMT_INCOME_TOTAL       0.347983
AMT_INCOME_TOTAL   AMT_CREDIT             0.341480
YEARS_BIRTH        YEARS_REGISTRATION     0.333301
                   YEARS_ID_PUBLISH       0.270743
YEARS_EMPLOYED     YEARS_BIRTH            0.242211
YEARS_BIRTH        EXT_SOURCE_3           0.172888
```

For customers with payment difficulties:



```
AMT_GOODS_PRICE   AMT_CREDIT            0.983065
                  AMT_ANNUITY           0.752206
AMT_ANNUITY       AMT_CREDIT            0.751400
YEARS_BIRTH       YEARS_REGISTRATION    0.288169
YEARS_EMPLOYED    YEARS_BIRTH           0.285932
YEARS_BIRTH       YEARS_ID_PUBLISH      0.251666
AMT_GOODS_PRICE   YEARS_BIRTH           0.137117
AMT_CREDIT        YEARS_BIRTH           0.136683
YEARS_EMPLOYED    YEARS_REGISTRATION    0.134482
AMT_GOODS_PRICE   EXT_SOURCE_2          0.131572
```

Insights:

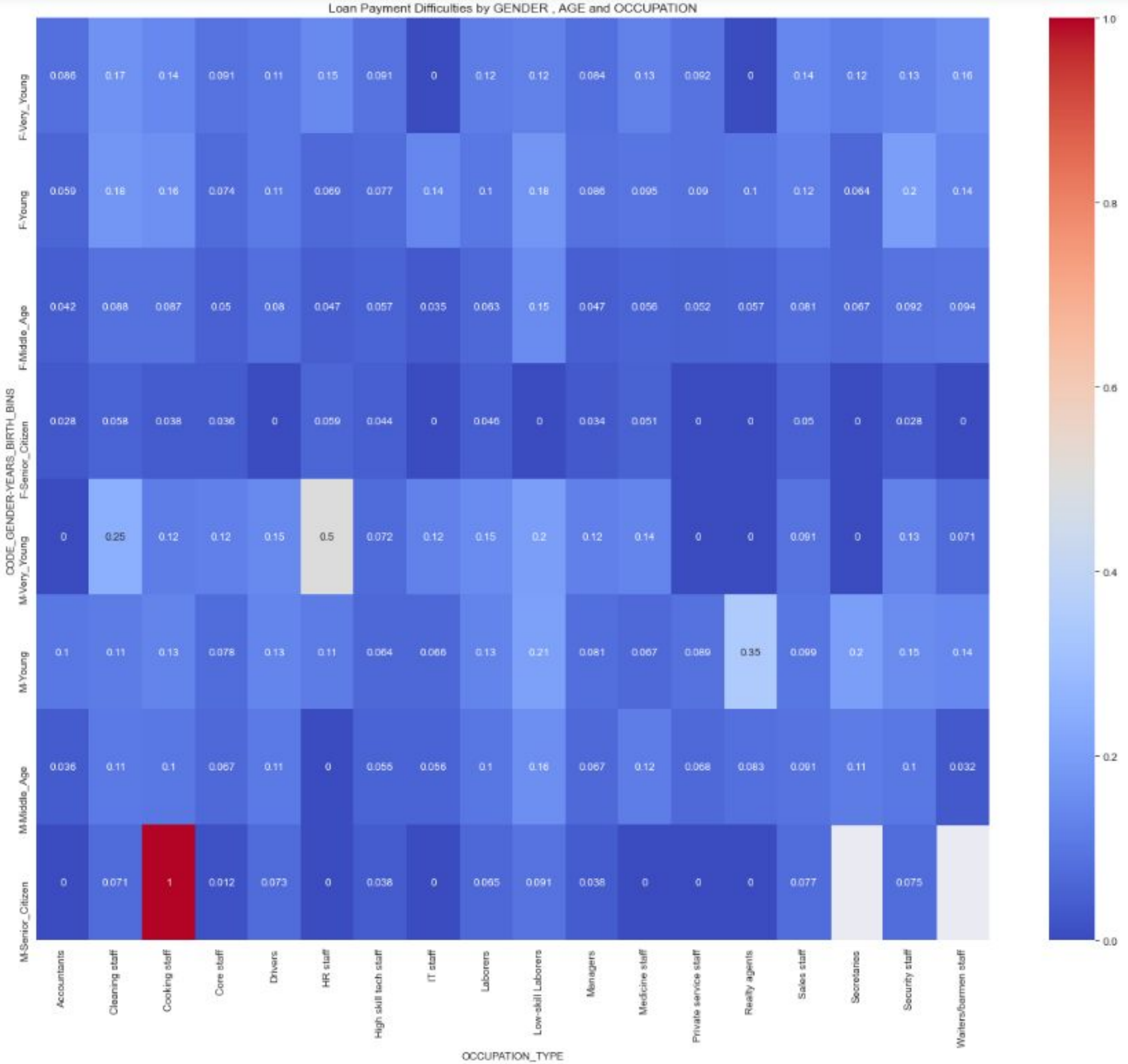
We observe that among the top 10 correlations from both the data frames, the top 3 correlations are almost similar.

- AMT_GOODS_PRICE and AMT_CREDIT
- AMT_GOODS_PRICE and AMT_ANNUITY
- AMT_ANNUITY and AMT_CREDIT

These variables have a high positive correlation for both customers with payment difficulty as well as for customers without payment difficulty.

Multivariate Analysis:

Analysing demographic attributes with target variable

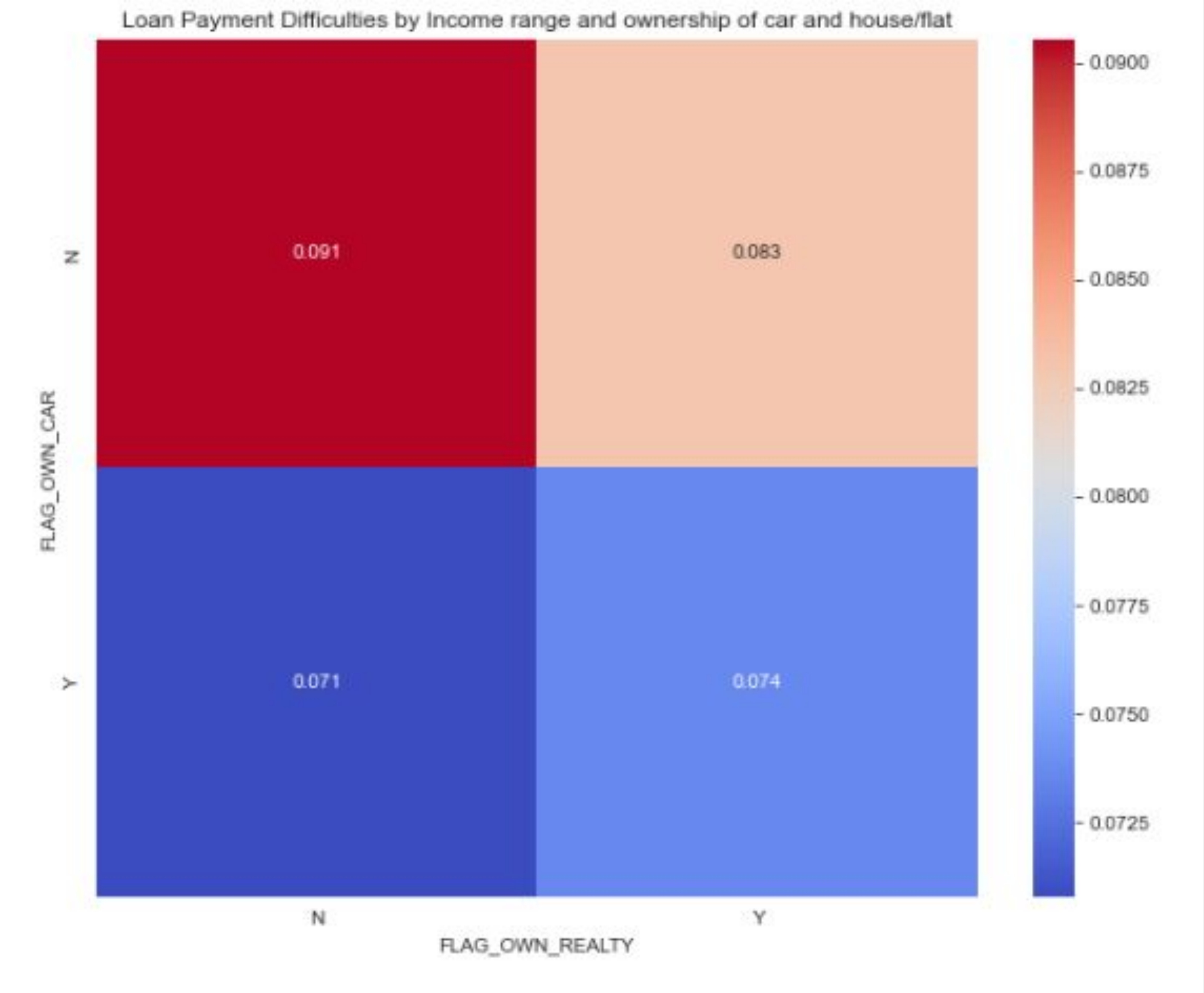


Insights:

Male senior cooking staffs and male very young HR staffs are having maximum difficulty in making loan repayment.

Multivariate Analysis:

Analysing ownership of car and house/flat variables attributes with target variable

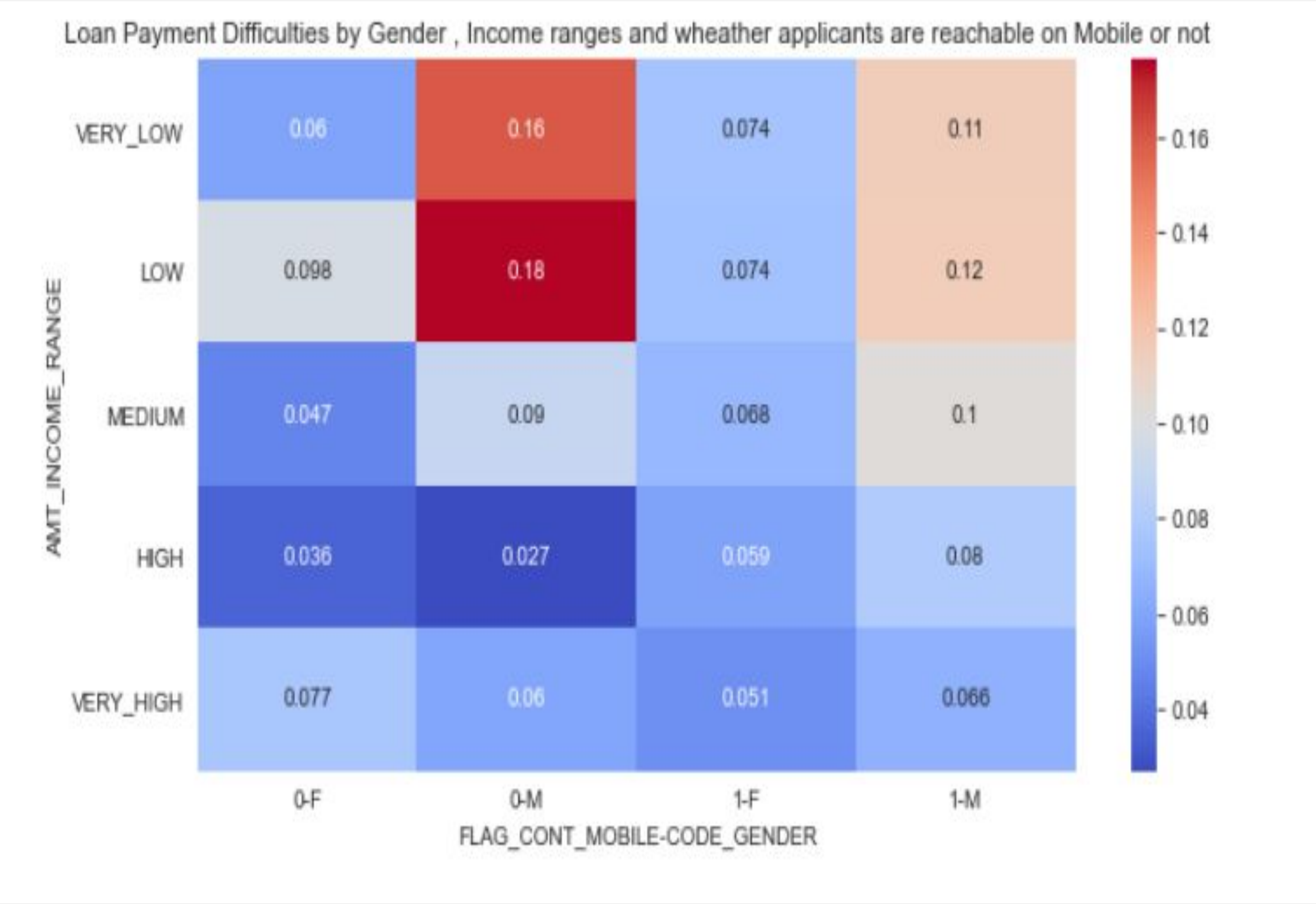


Insights:

Customers not having cars along with house/flat are having difficulty in making loan repayments.

Multivariate Analysis:

Analysing Gender, Income range and whether applicants are reachable on mobile or not attributes with target variable

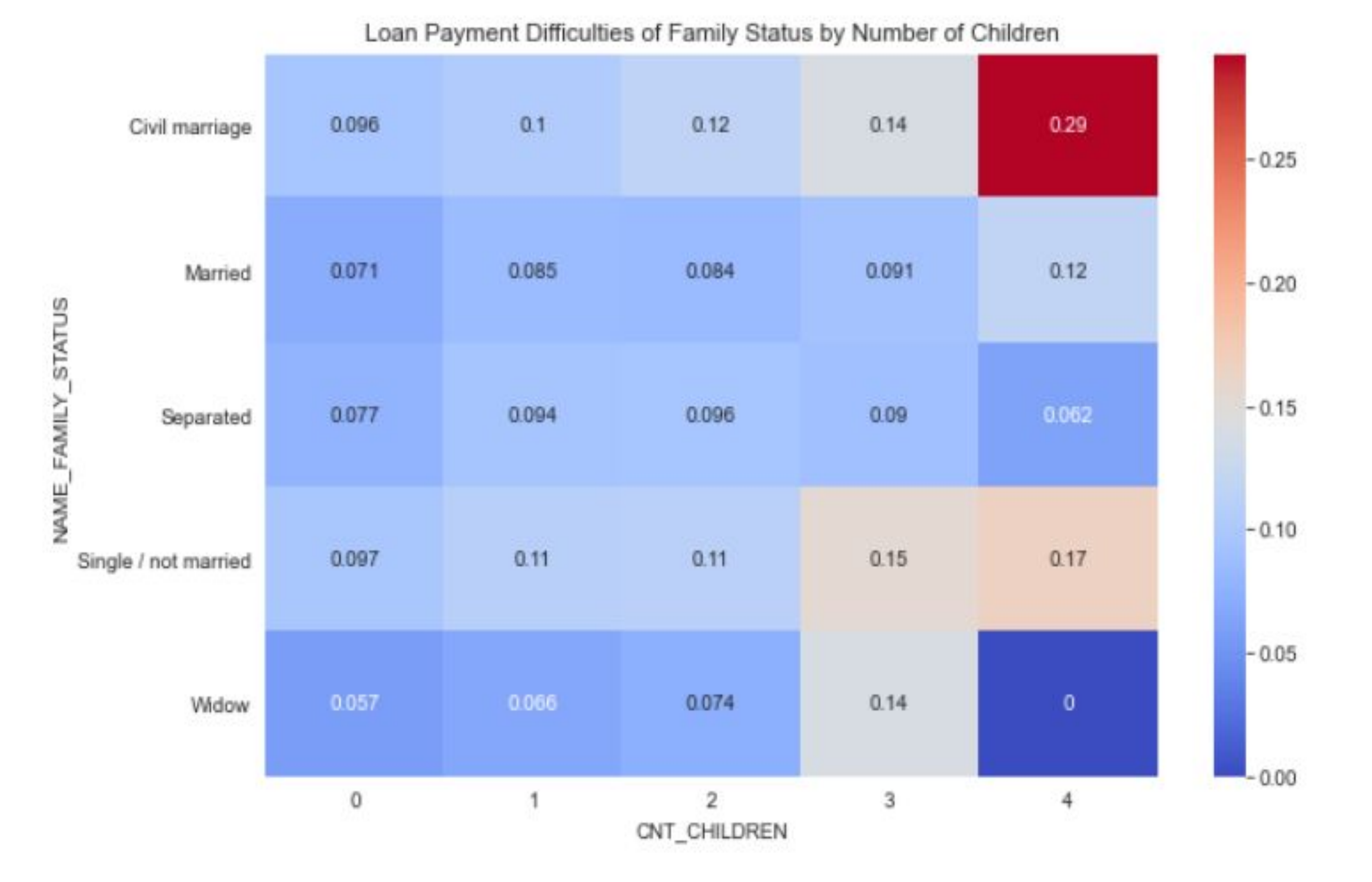


Insights:

Male customers who are not reachable on phone and have low incomes are having difficulty in making loan repayments.

Multivariate Analysis:

Analysing impact of number of children and family status on probability of defaulting on loan

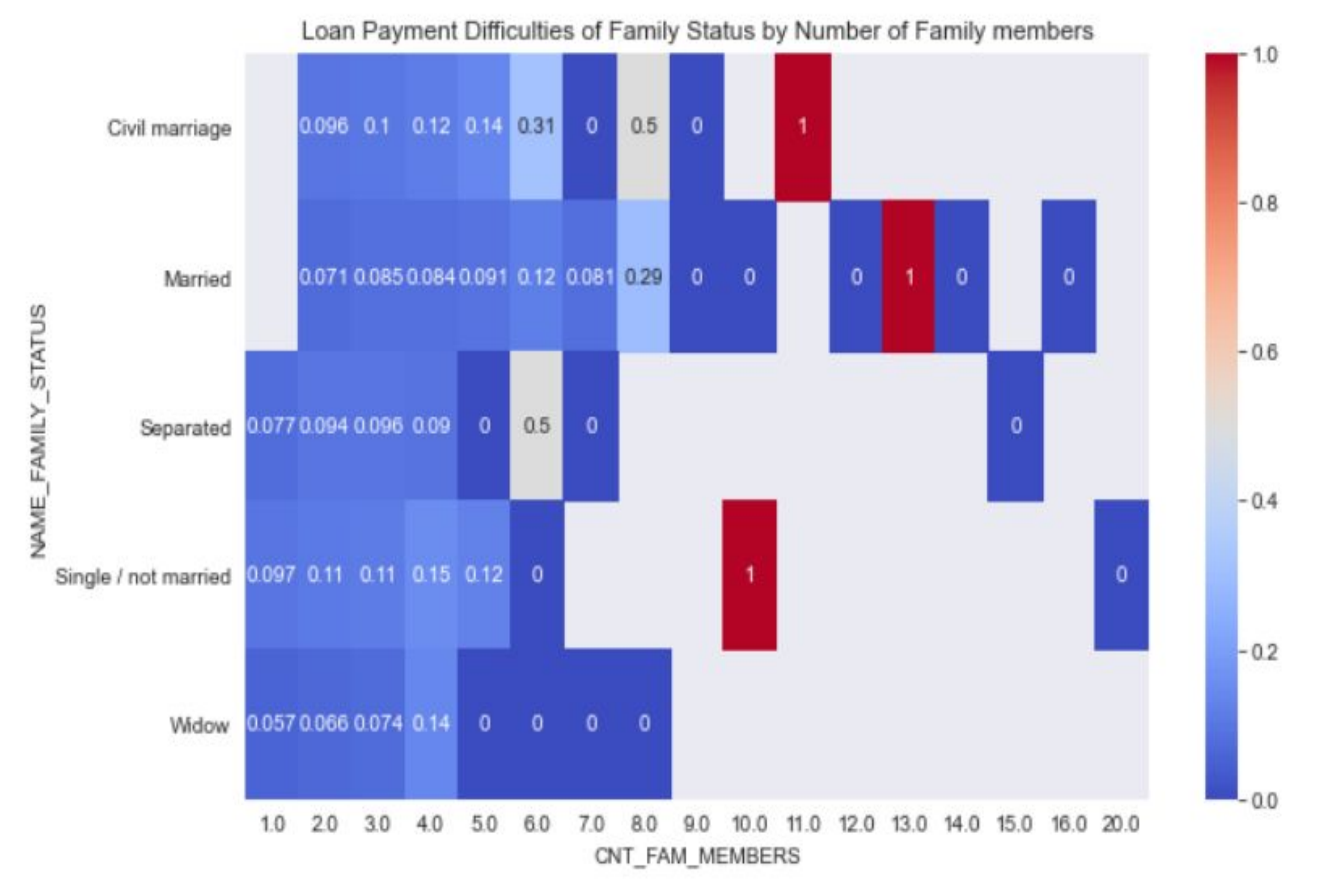


Insights:

Loan repayment difficulties are seen more in families with more children as compared families with less children.

Multivariate Analysis:

Analysing impact of number of family members and family status on probability of defaulting on loan

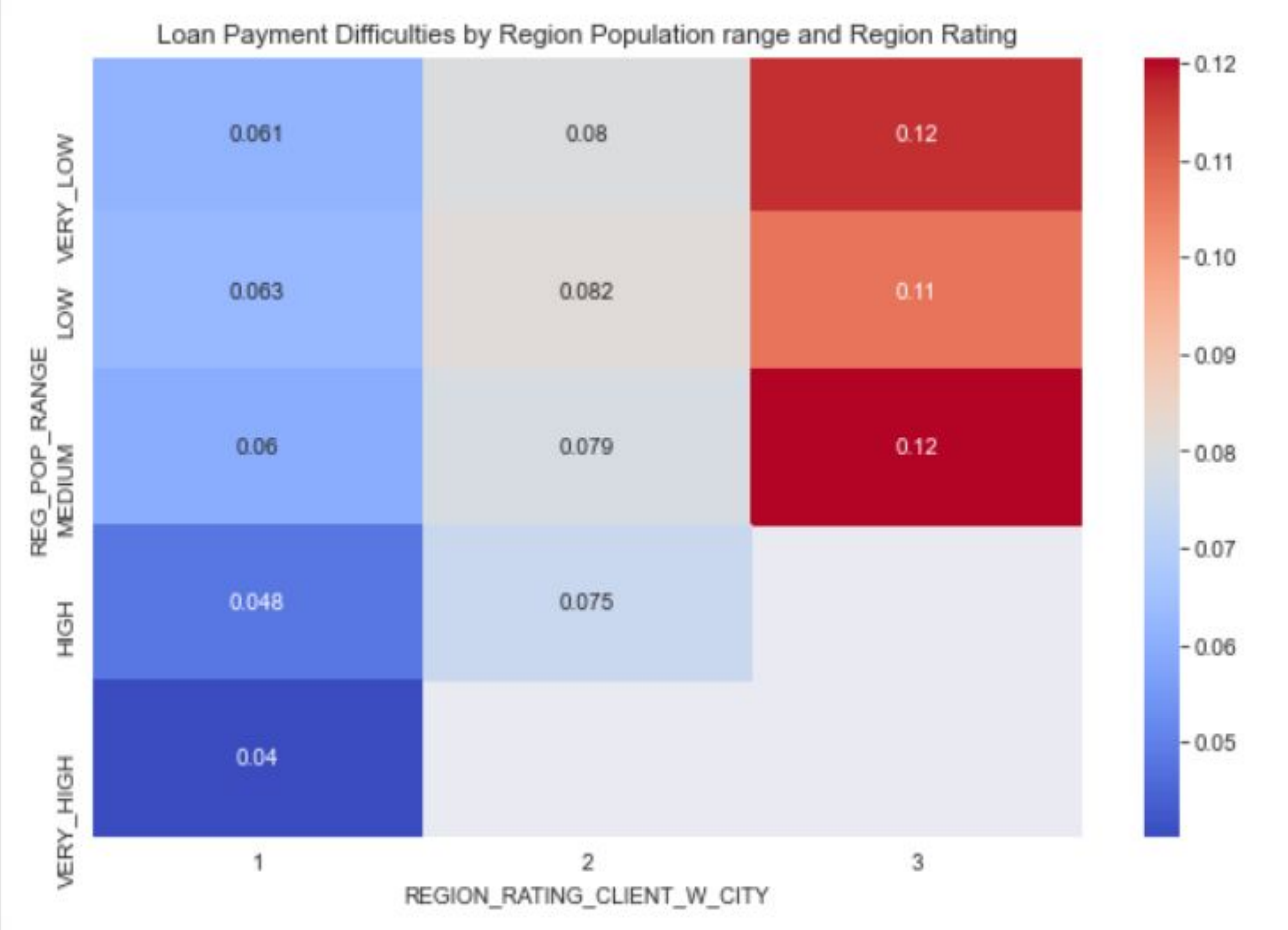


Insights:

Loan repayment difficulties are seen more in more-membered families as compared to less-membered families.

Multivariate Analysis:

Analysing Loan Payment Difficulties by Region Population range and Region Rating

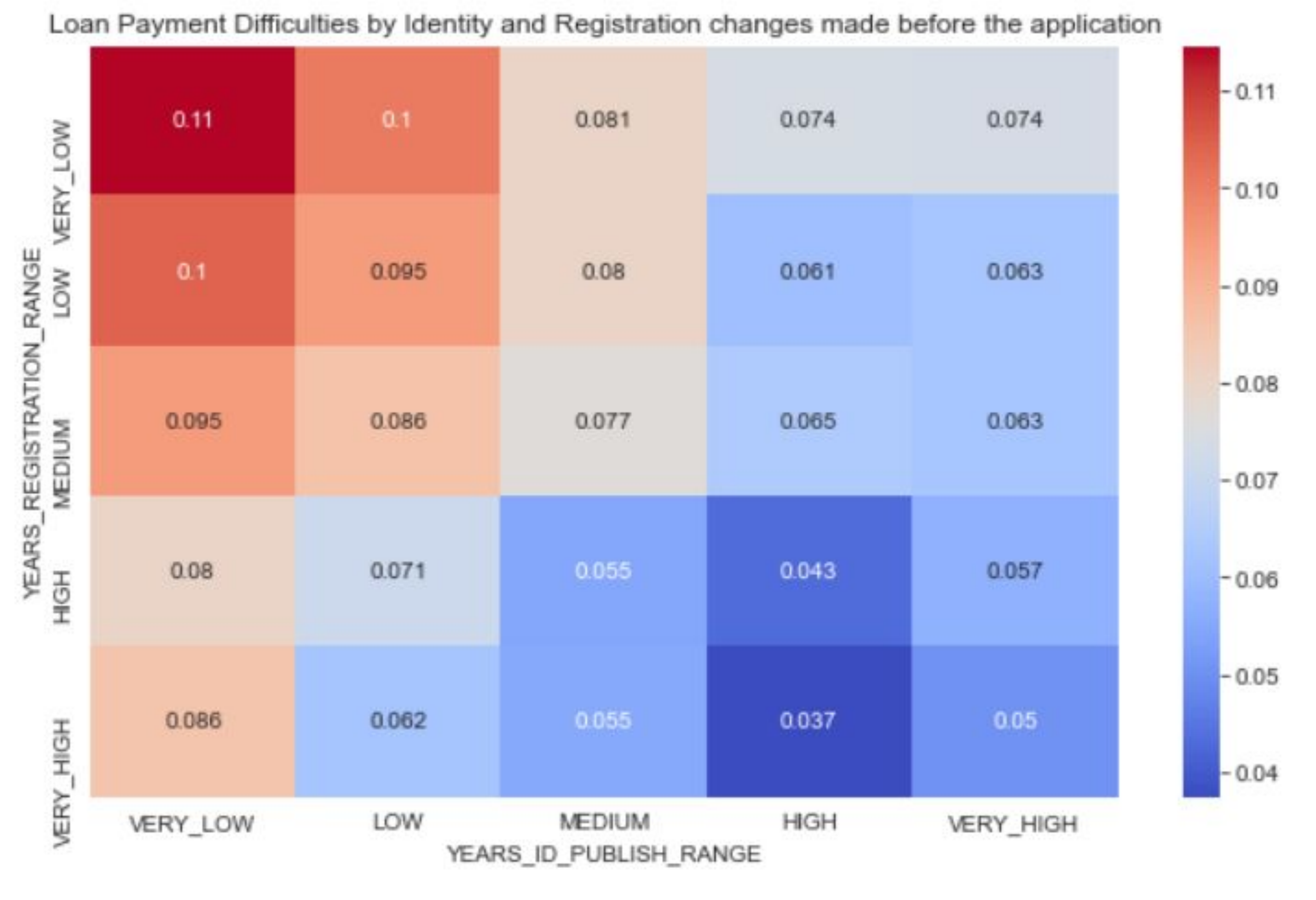


Insights:

Customers with REGION_POPULATION_RELATIVE < 0.03 are having more difficulties in repaying for loans. Also, higher the customer's Region Rating, higher the chances of his/her defaulting.

Multivariate Analysis:

Analysing Loan Payment Difficulties by Identity and Registration changes made before the application

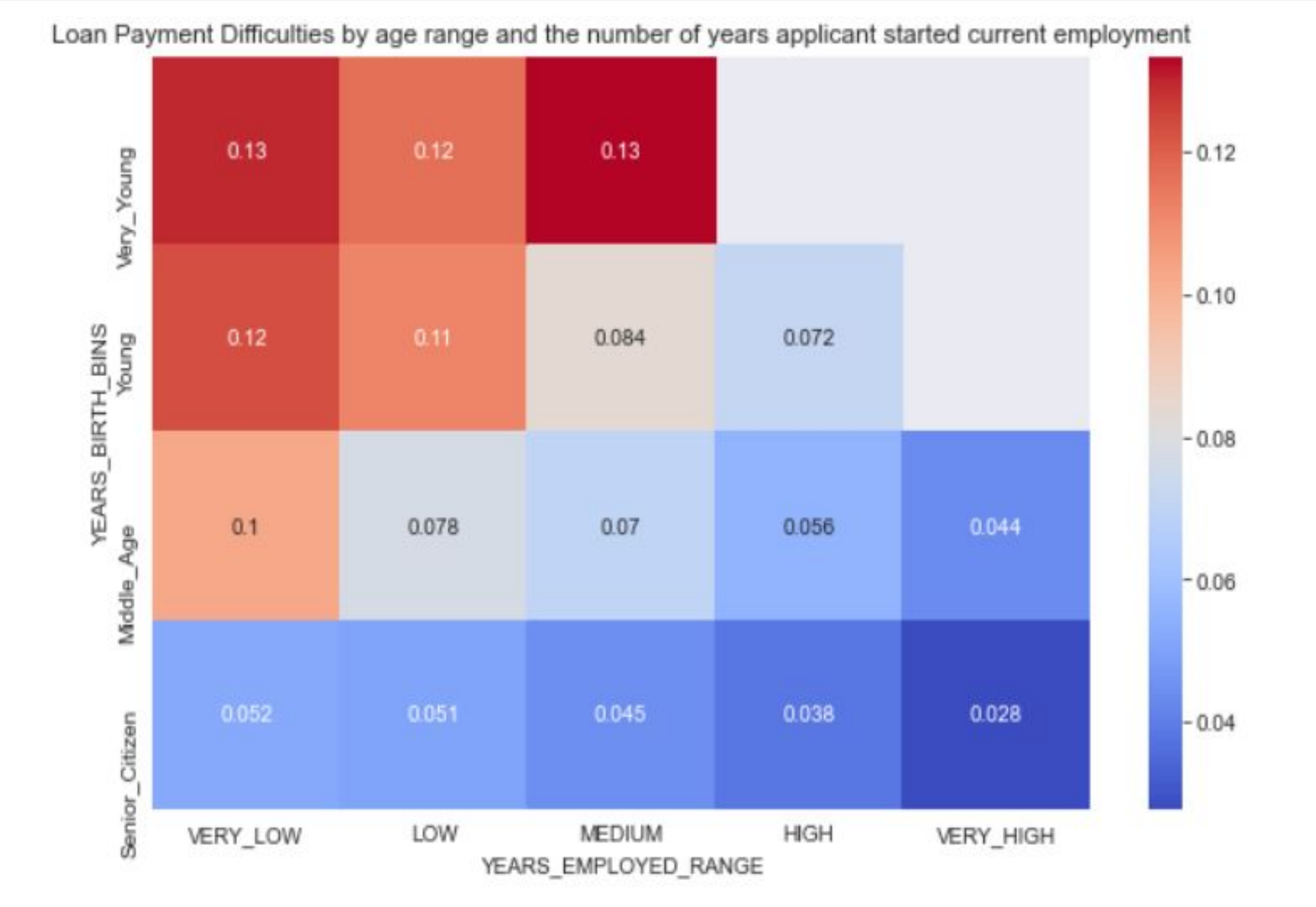


Insights:

Customers who changed their registration in last 12 years and ID in 9 years before the application, are having more difficulties in repaying for loans as compared to those who have changed before 12 and 9 years timeframe respectively.

Multivariate Analysis:

Analysing Loan Payment Difficulties by age range and the number of years applicant started current employment

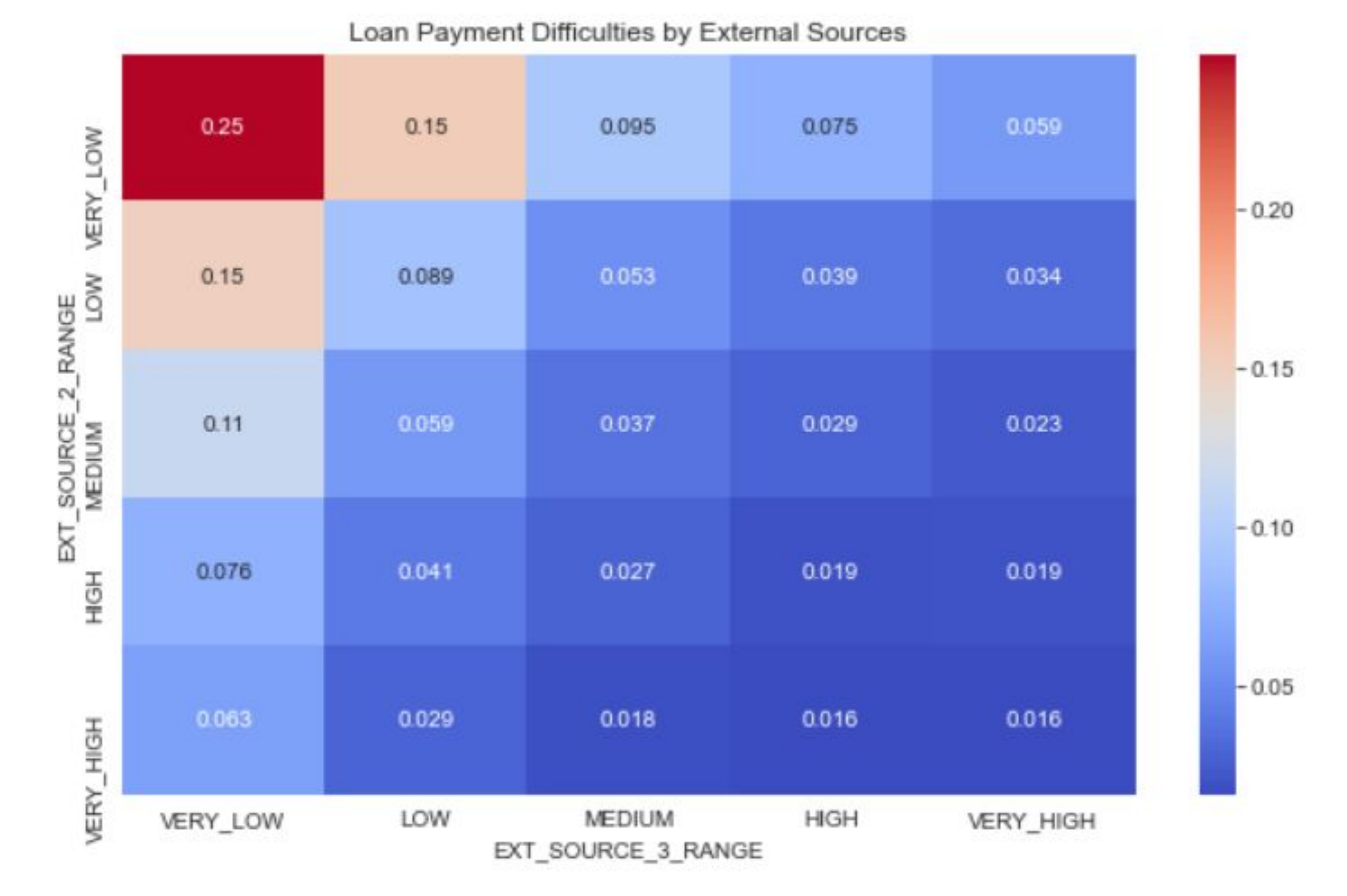


Insights:

Non-senior citizens(less than 56 years age) who have spent less than 8.8 years in their current employment are having more difficulties in repaying for loans.

Multivariate Analysis:

Analysing Loan Payment Difficulties by External Sources

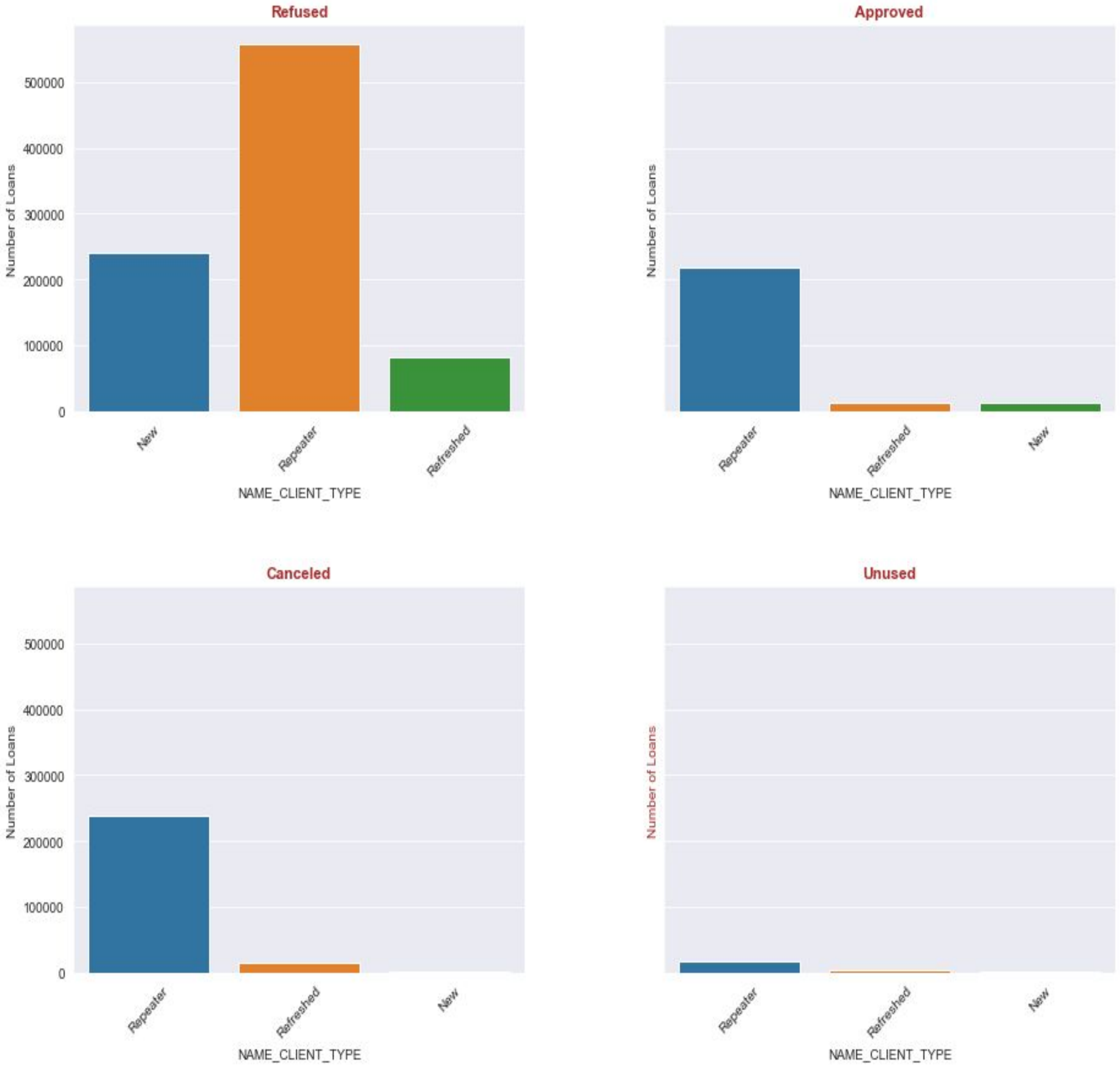


Insights:

Customers who have received ratings lower than 0.57 and 0.53 from external sources 2 and 3 are having more difficulties in repaying for loans.

Analysis on Merged Data (Applications Data + Previous Applications Data

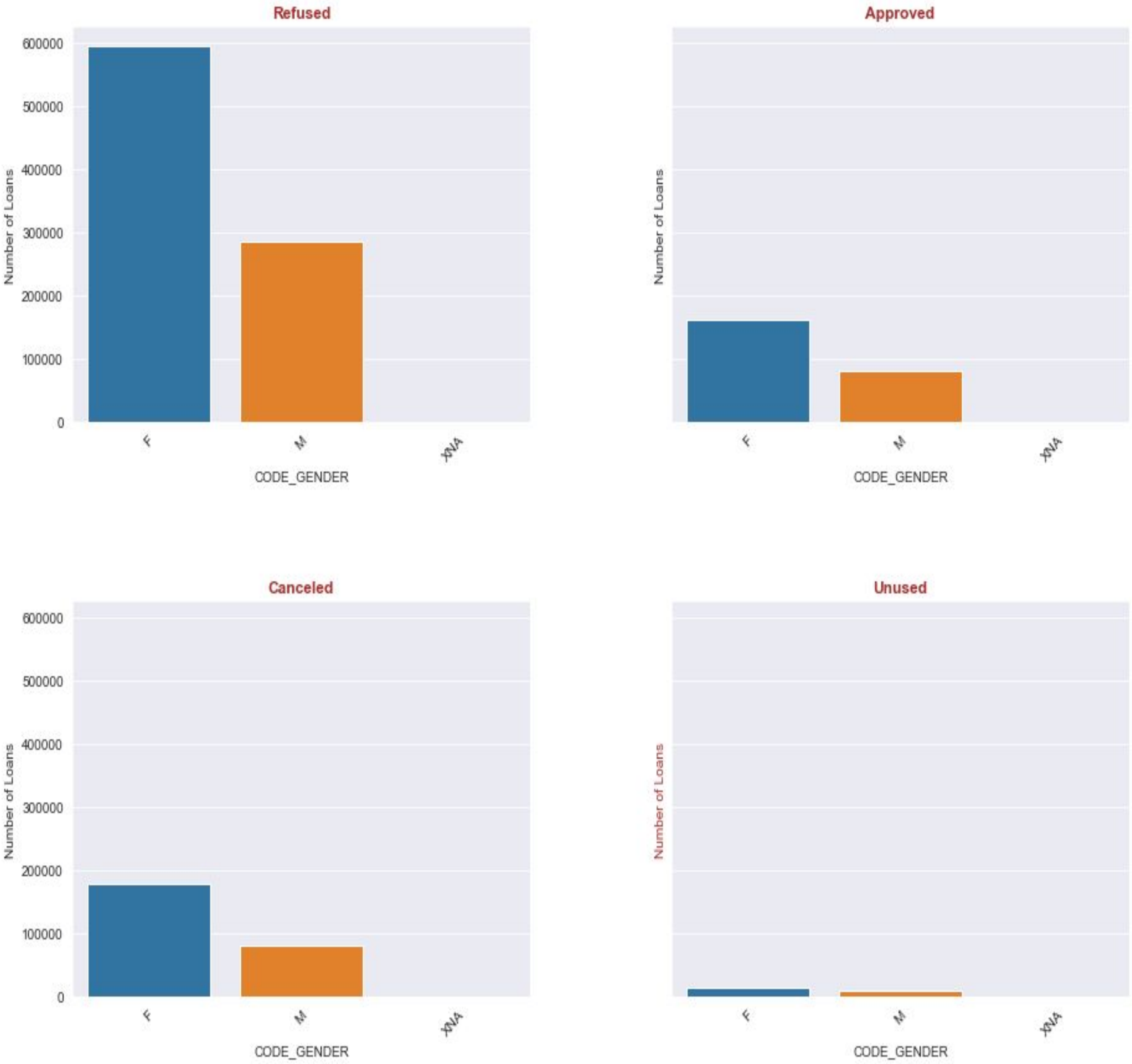
Creating subplots to analyse NAME_CLIENT_TYPE and number of loans for different NAME_CONTRACT_STATUS categories



Insights:

Repeater category of customers are most likely to be Refused an application.
However, it is also true that for approved loans and cancelled loans, the most number of customers are from the repeater category itself though the proportion is smaller

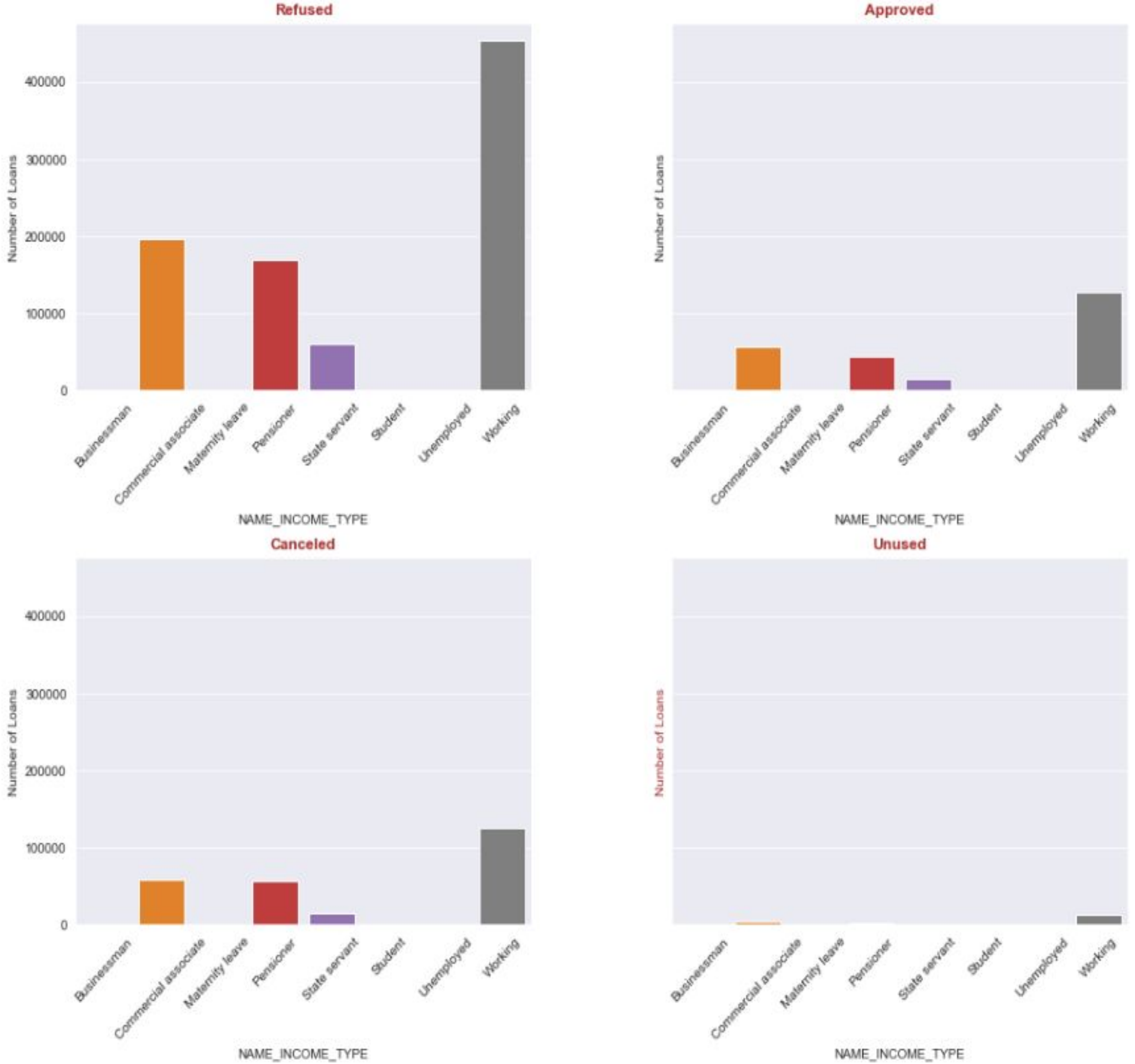
Creating subplots to analyse GENDER and number of loans for different NAME_CONTRACT_STATUS categories



Insights:

The number of females in getting their loan application approved or refused or cancelled is more than males.

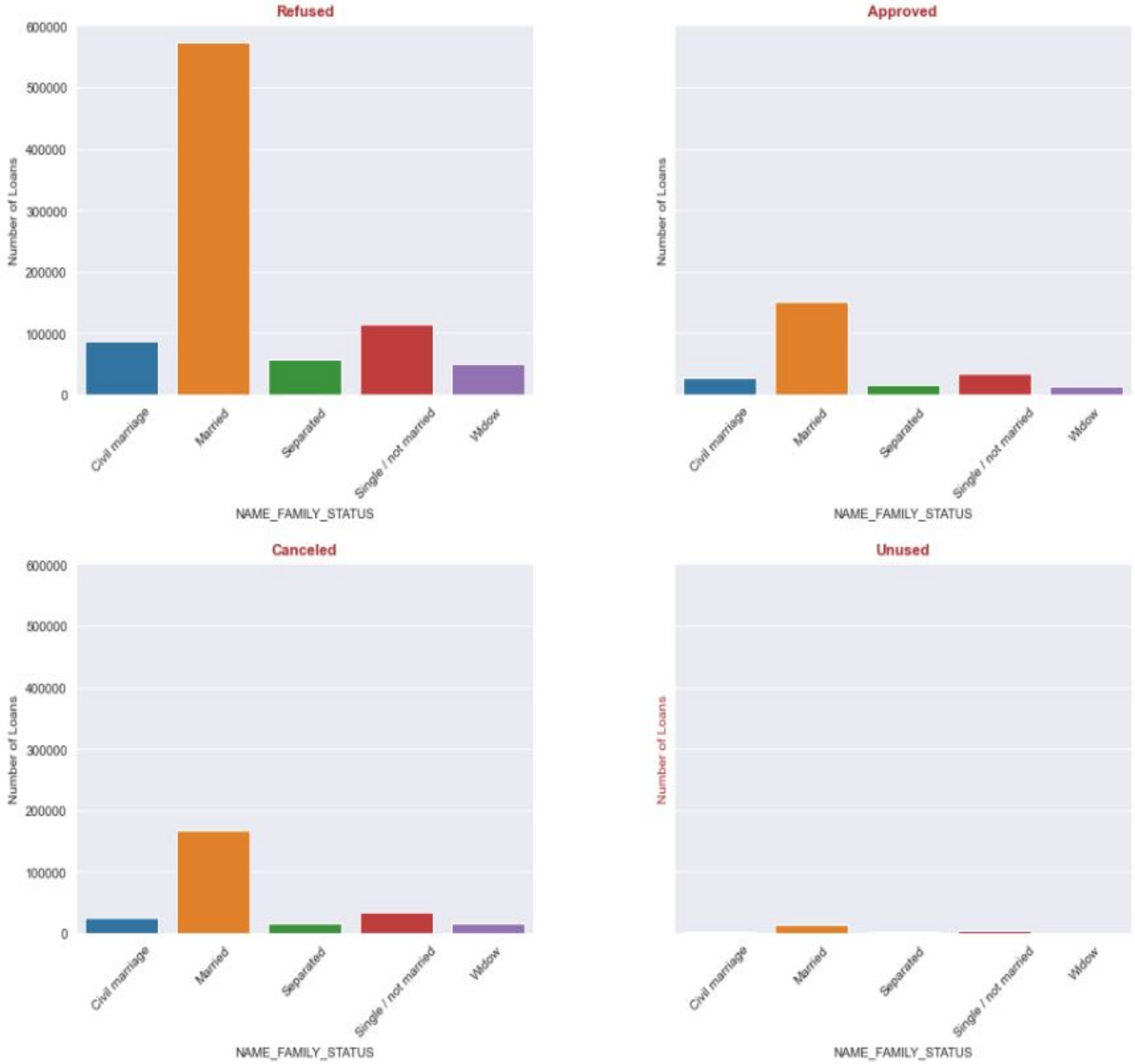
Creating subplots to analyse INCOME TYPE and number of loans for different NAME_CONTRACT_STATUS categories



Insights:

Working customers have the highest number of refused, approved, canceled and unused offer among all customers. This is expected as number of working customers applying for loans will also be more compared to other categories. Moreover, for Pensioner and Commercial associate customers, there is a considerably higher number of applications which have been refused.

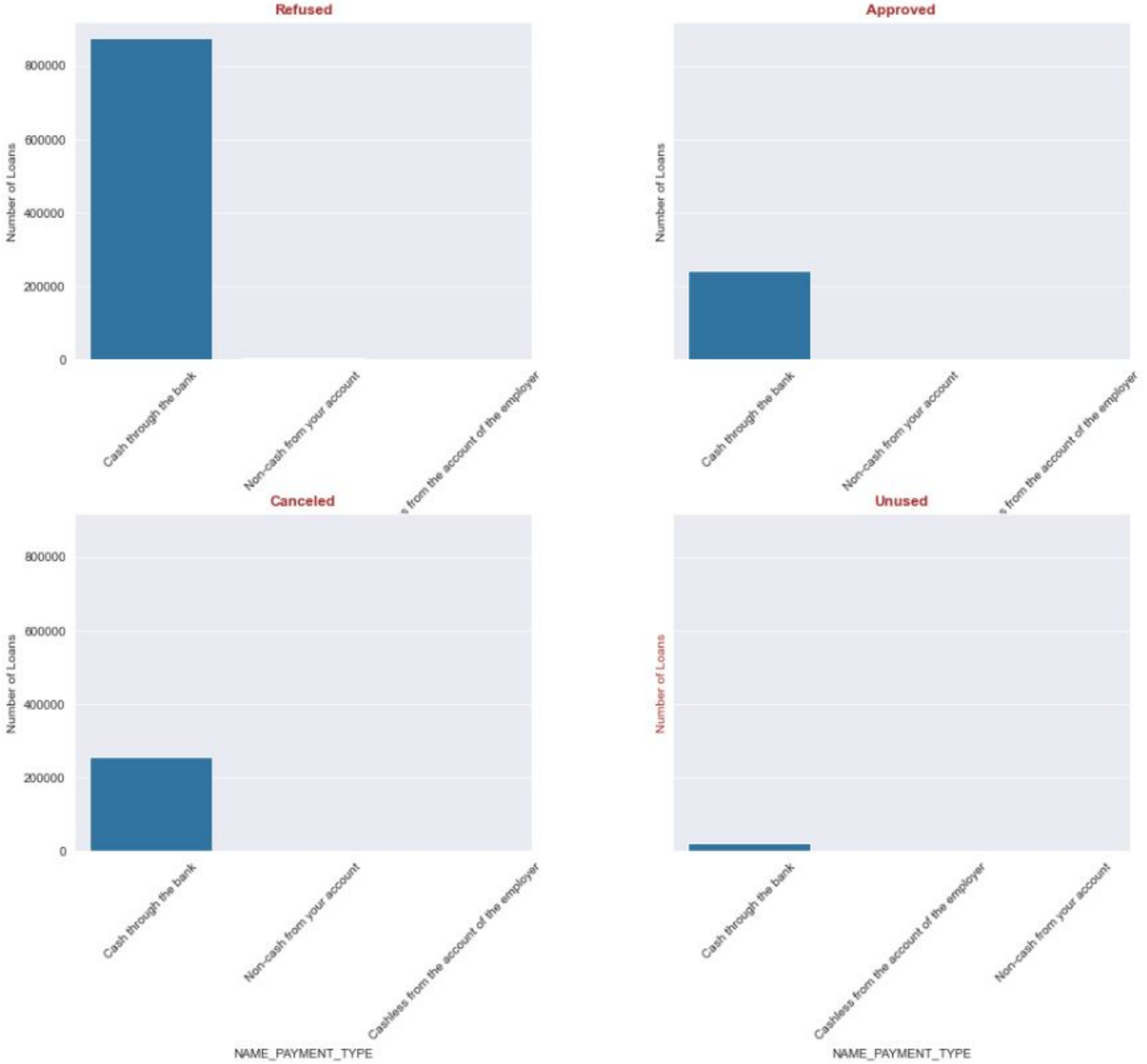
Creating subplots to analyse FAMILY STATUS and number of loans for different NAME_CONTRACT_STATUS categories



Insights:

Married customers are applying and taking more loans. However, the proportion of applications which are refused are much higher for married customers compared to other categories.

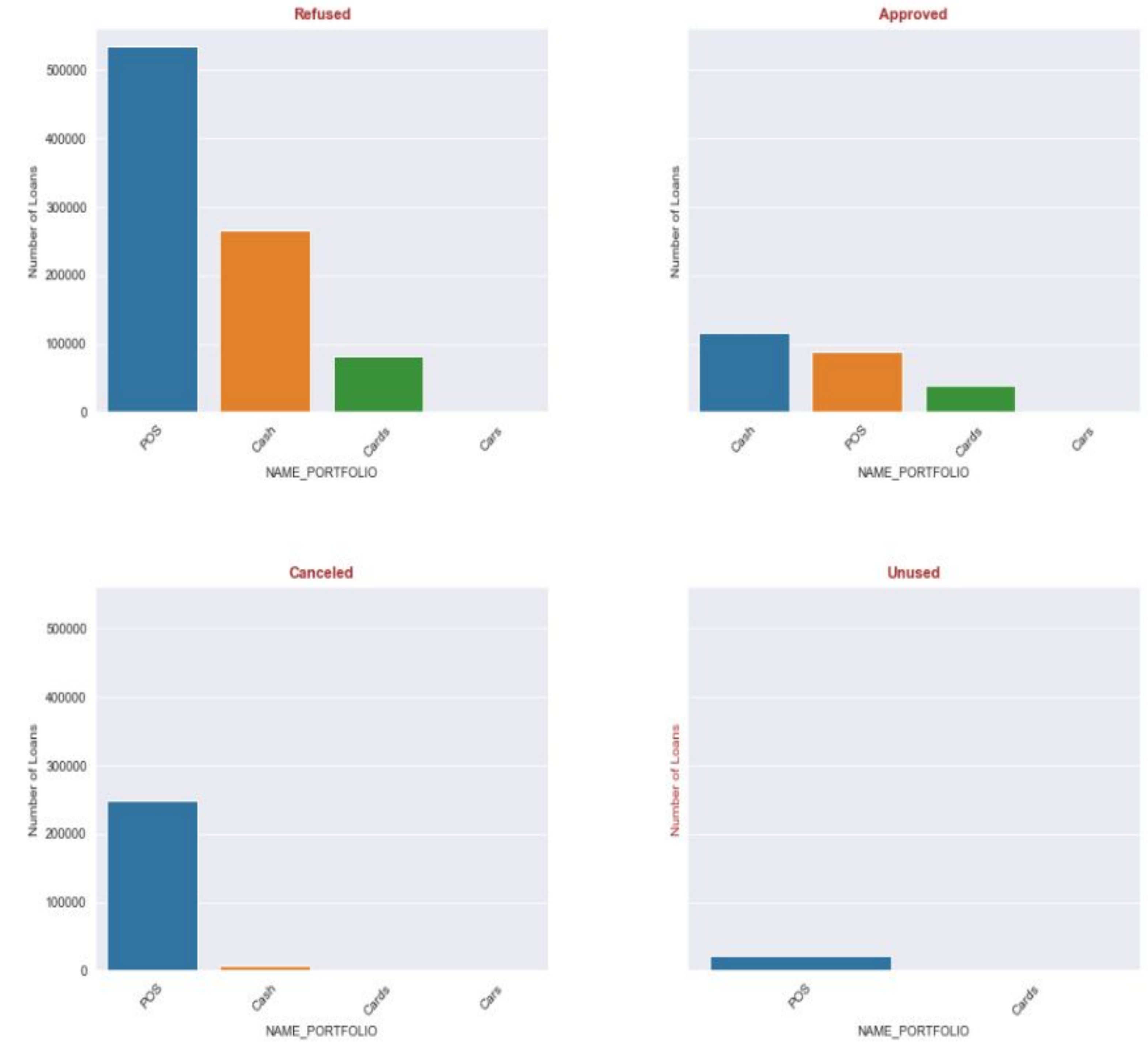
Creating subplots to analyse PAYMENT TYPE and number of loans for different NAME_CONTRACT_STATUS categories



Insights:

Most payment types for loan applications is Cash through bank. The number of approved and cancelled applications for Cash through bank payment type is almost same. However, the number of refused applications for the same payment type is very high

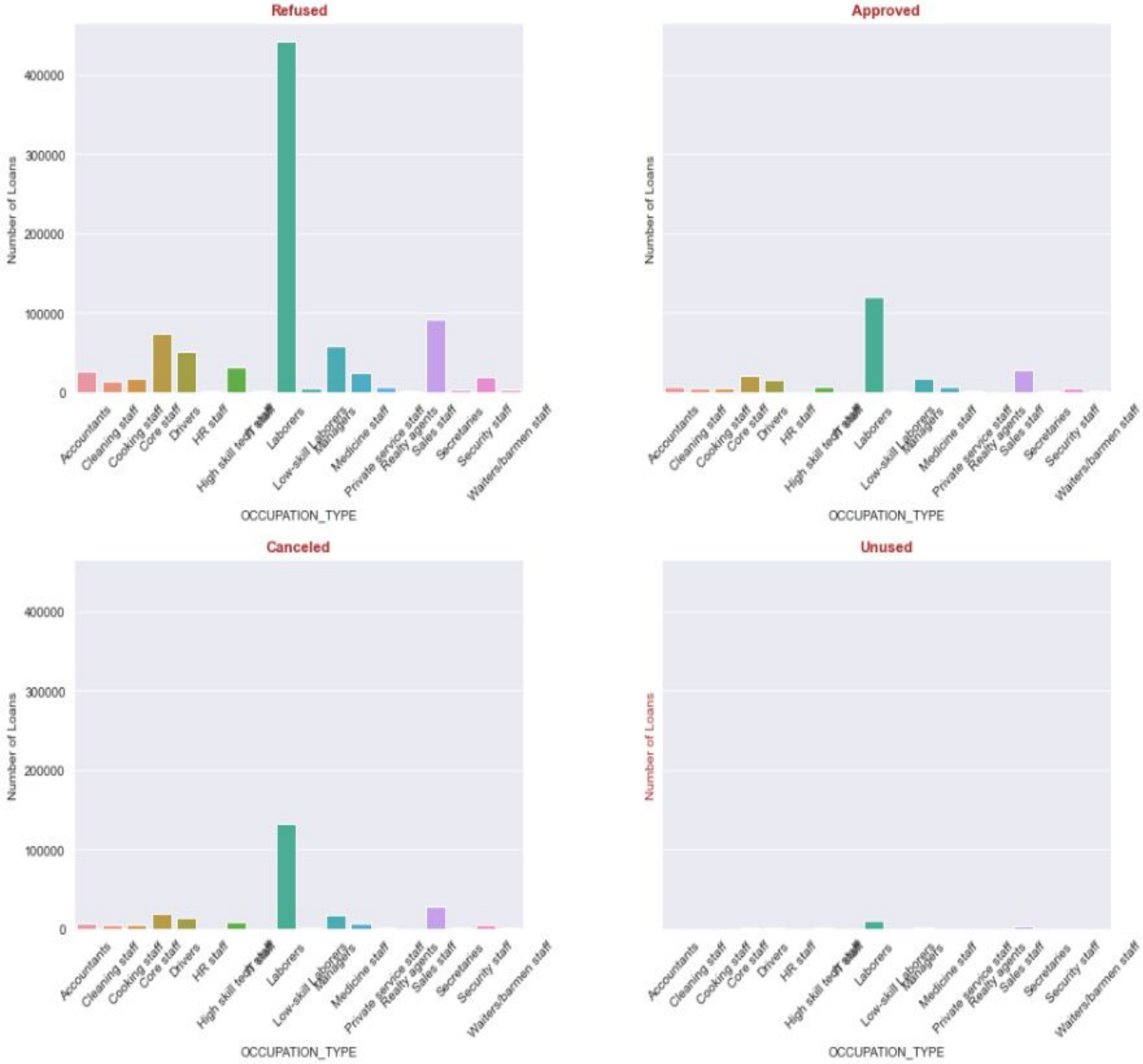
Creating subplots to analyse PORTFOLIO and number of loans for different NAME_CONTRACT_STATUS categories



Insights:

Highest number of approved loans is for cash loans. Highest number of applications which have been refused is for POS loans.

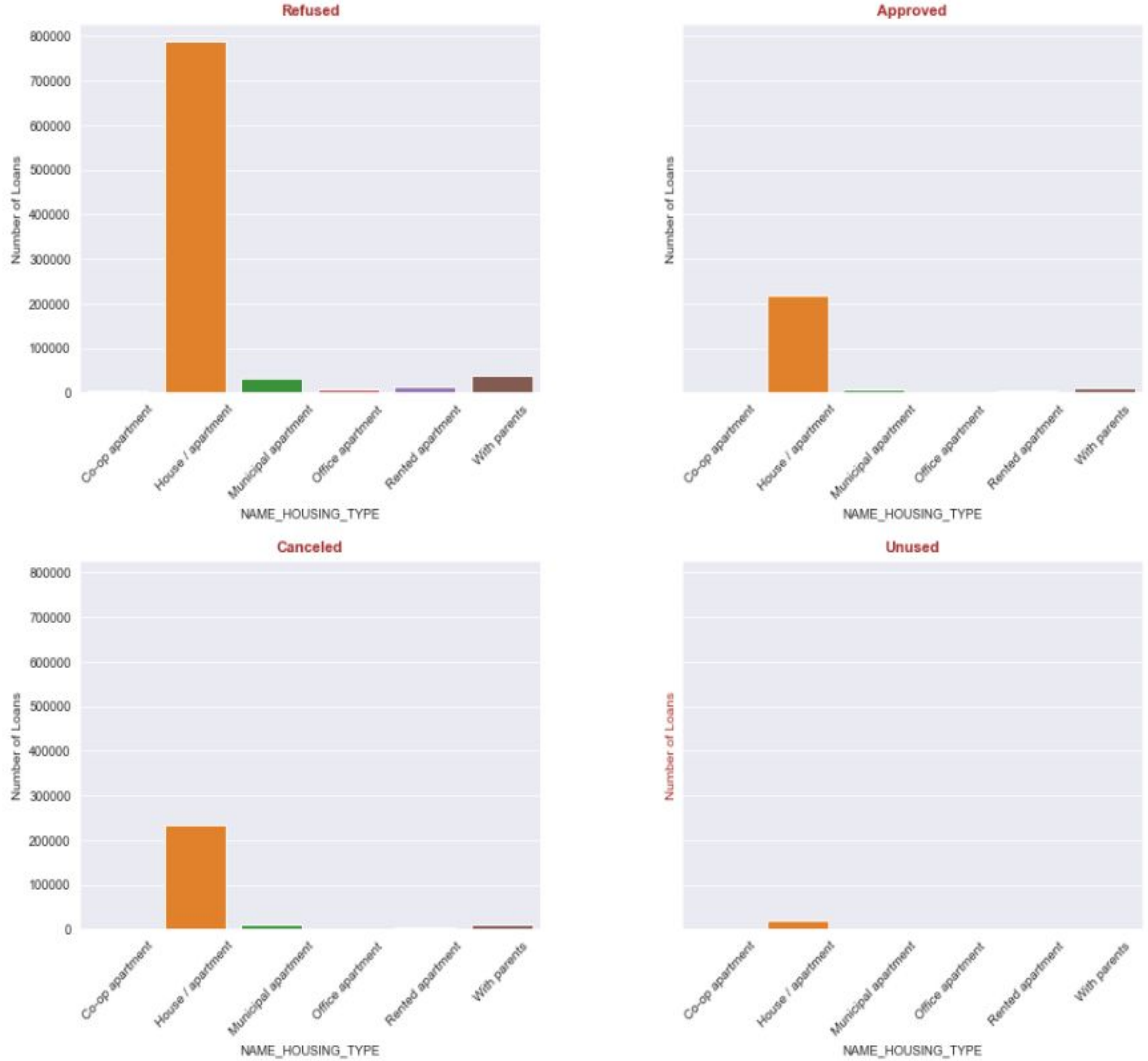
Creating subplots to analyse OCCUPATION and number of loans for different NAME_CONTRACT_STATUS categories



Insights:

Laborers have the most number of approved as well as refused applications. Also, they have a pretty high number of canceled applications as well.

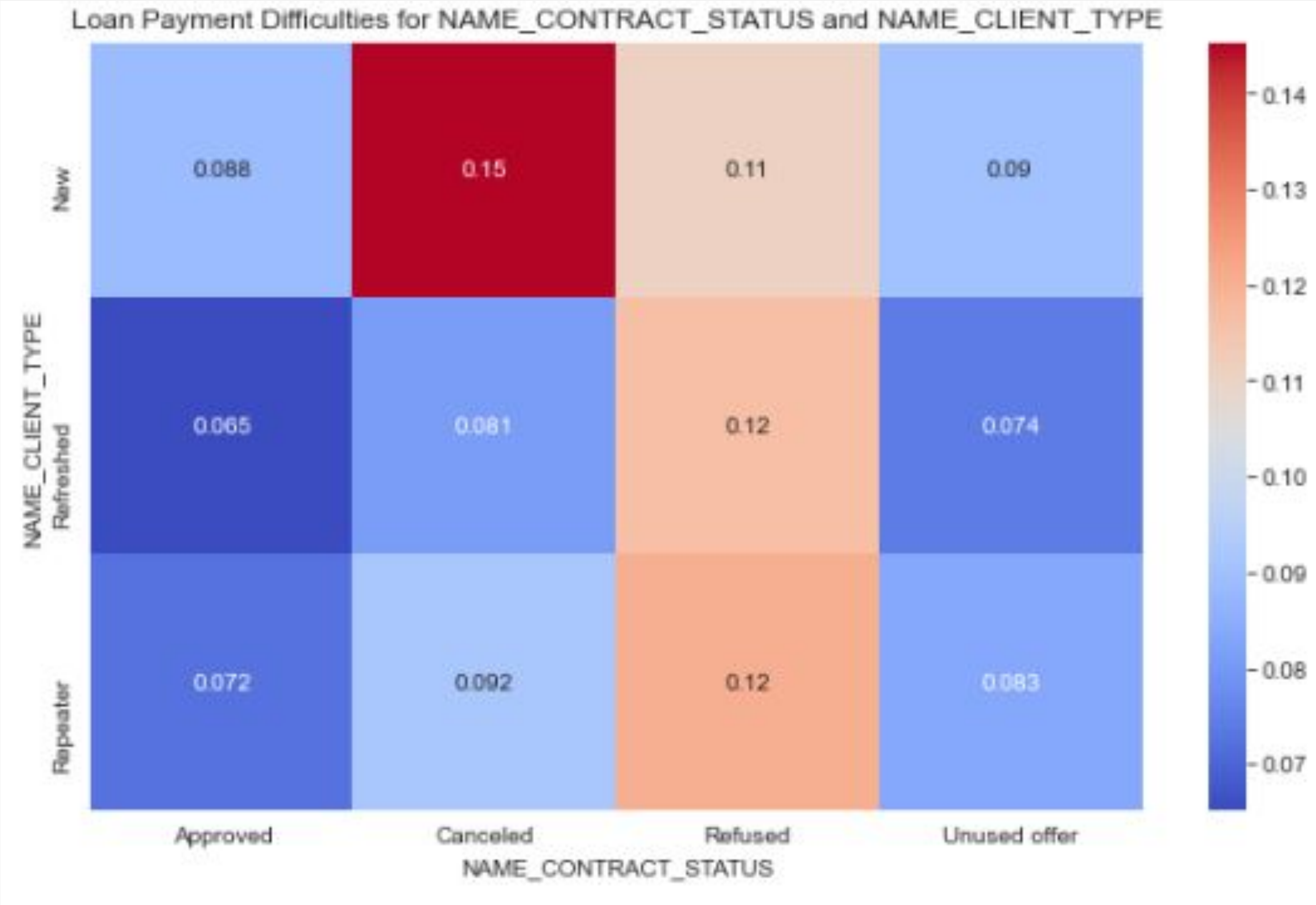
Creating subplots to analyse HOUSING_TYPE and number of loans for different NAME_CONTRACT_STATUS categories



Insights:

Most number of applications are from customers in house/apartment. This is true for approved, refused, canceled or unused applications.

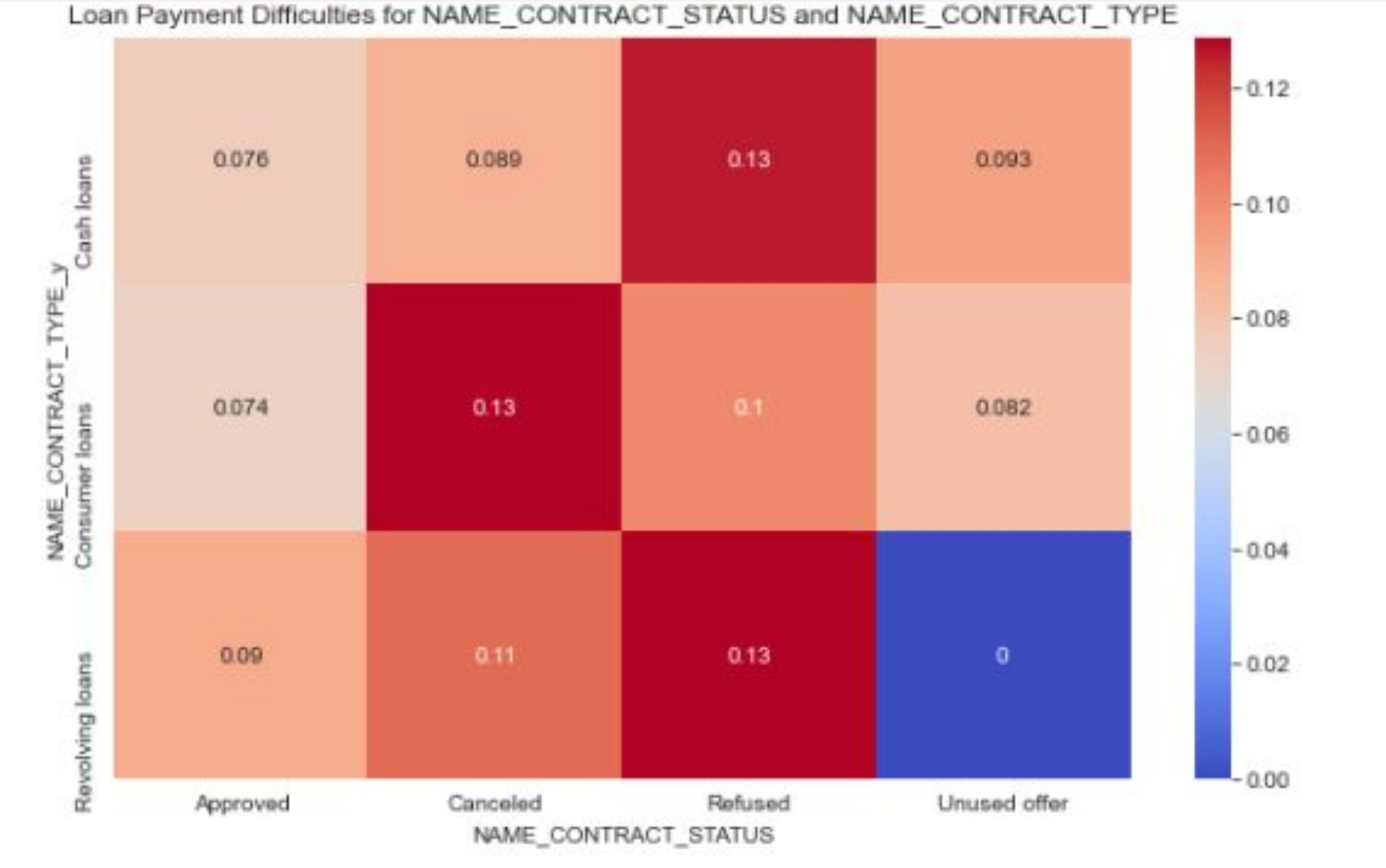
Loan Payment Difficulties for NAME_CONTRACT_STATUS and NAME_CLIENT_TYPE



Insights:

Customers who were 'New' and had 'Cancelled' previous application tend to have more difficulty in loan payment in current application.

Loan Payment Difficulties for NAME_CONTRACT_STATUS and NAME_CONTRACT_TYPE



Insights:

Customers with 'Revolving loans' and with 'Refused' previous application or 'Consumer loans' with 'Canceled' previous application or 'Cash loans' with 'Refused' previous application tend to have more difficulty in loan payment in current application.

Overall Insights:

Applications data

- The count of ‘Maternity Leave’ in ‘NAME_INCOME_TYPE’ is very less and it also has high % of payment difficulties. Hence, client with income type as ‘Maternity leave’ are the driving factors for Loan Defaulters.

- The count of ‘Low skilled Laborers’ in ‘OCCUPATION_TYPE’ is comparatively less and it also has high % of payment difficulties. Hence, client with occupation type as ‘Low skilled Laborers’ are the driving factors for Loan Defaulters.

- The count of ‘Lower Secondary’ in ‘NAME_EDUCATION_TYPE’ is comparatively less and it also has high % of payment difficulties. Hence, client with education type as ‘Lower Secondary’ are the driving factors for Loan Defaulters.

- Male senior cooking staffs and male very young HR staffs are having maximum difficuly in loan payment.

- Customers not having cars along with house/flat are having difficulty in making loan payments.

- Male customers who are not reachable on phone and have low incomes are having difficulty in making loan payments.

- Loan payment difficulties are seen more in more-membered families as compared to less-membered families.

- Customers with 'REGION_POPULATION_RELATIVE' < 0.03 are having more difficulties in paying for loans. Also, higher the customer's Region Rating, higher the chances of his/her defaulting.

- Customers who changed their registration in last 12 years and ID in 9 years before the application, are having more difficulties in paying for loans as compared to those who have changed before 12 and 9 years timeframe respectively.

- Non-senior citizens(less than 56 years age) who have spent less than 8.8 years in their current employment are having more difficulties in paying for loans.

- Customers who have received ratings lower than 0.57 and 0.53 from external sources 2 and 3 are having more difficulties in paying for loans

Overall Insights:

Previous Applications data

- The count of ‘Refused’ in ‘NAME_CONTRACT_STATUS’ is comparatively less and it also has high % of payment difficulties. Hence, client with contract status as ‘Refused’ in previous application are the driving factors for Loan Defaulters.

- The count of ‘Revolving Loans’ in ‘NAME_CONTRACT_TYPE’ is comparatively less and it also has high % of payment difficulties. Hence, client with contract type as ‘Revolving loans’ in previous application are the driving factors for Loan Defaulters.

- It can be observed from the graph that Clients with 'Revolving loans' and with 'Refused' previous application tend to have more % of payment difficulties in current application. Since the count of both 'Revolving loans' and ‘Refused’ is comparatively less, clients with ‘Revolving Loans’ and ‘Refused’ previous application are driving factors for Loan Defaulters

Thank You!