CREDIT EDA CASE STUDY

SHREYAS R

DEEPIKA BHATT

INTRODUCTION

- **Available Data:**

- The Data Frame “data” contains the data of client who has recently applied for loan and data frame “Previous_data” contains the information about the previous loans.
- **“SK_ID_CURR”** column name is common in both the data frames. Thus we can merge the data of both the data frame on the basis of the mentioned column.

- **Merged Data:**

- The data frame “merged_data” contains the merged data of “data” and “Previous_data”

- **Problem Statement:**

- 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:

1) If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
2) 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.

APPROACH & METHODOLOGY

DESCRIPTION OF DATA

- Shape
- Info
- Data Type

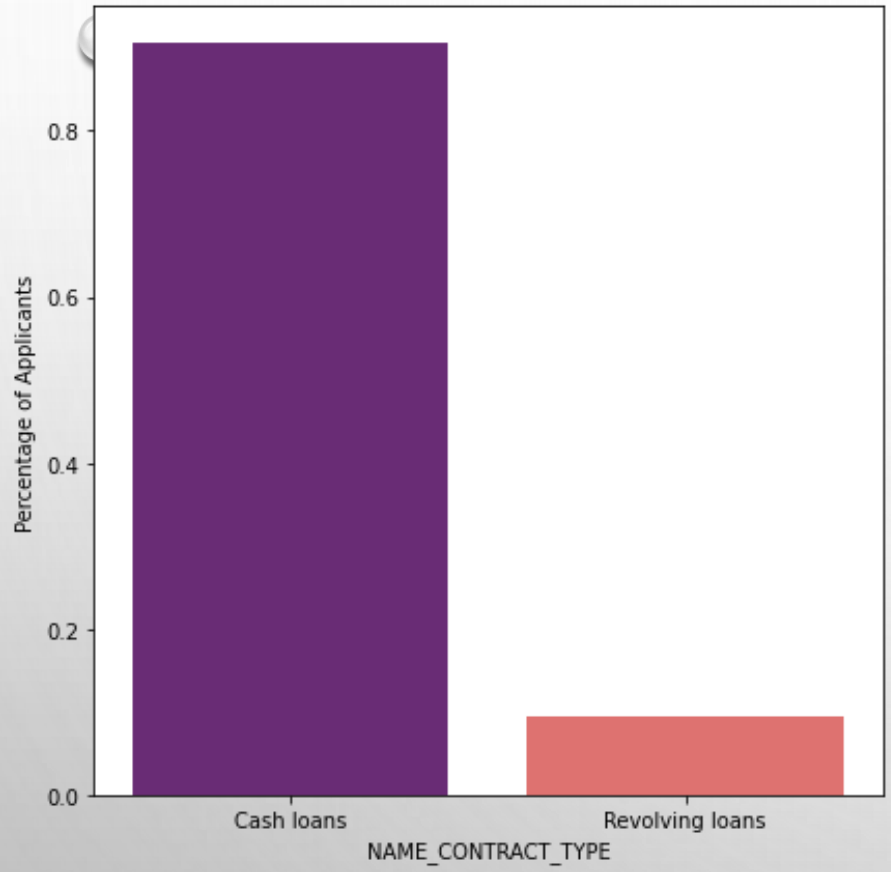
DATA QUALITY CHECK

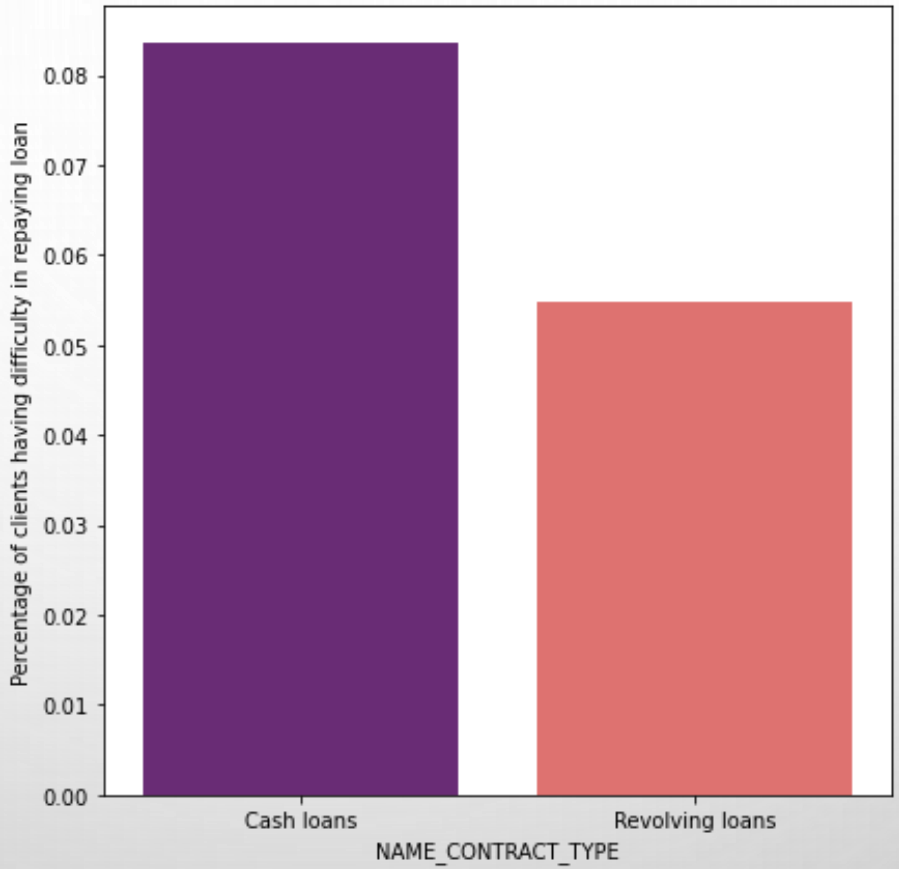
- Handling the Missing/Null values
- Handling the Outliers
- Standardize the columns

DATA ANALYSIS

- Checking the Imbalance percentage of the Target variable
- Univariate/Segmented Univariate Analysis of Categorical and Numerical variables
- Bivariate Analysis of Numerical-Numerical, Categorical-Numerical, and Categorical-Categorical variables

UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS ON CONTRACT TYPE

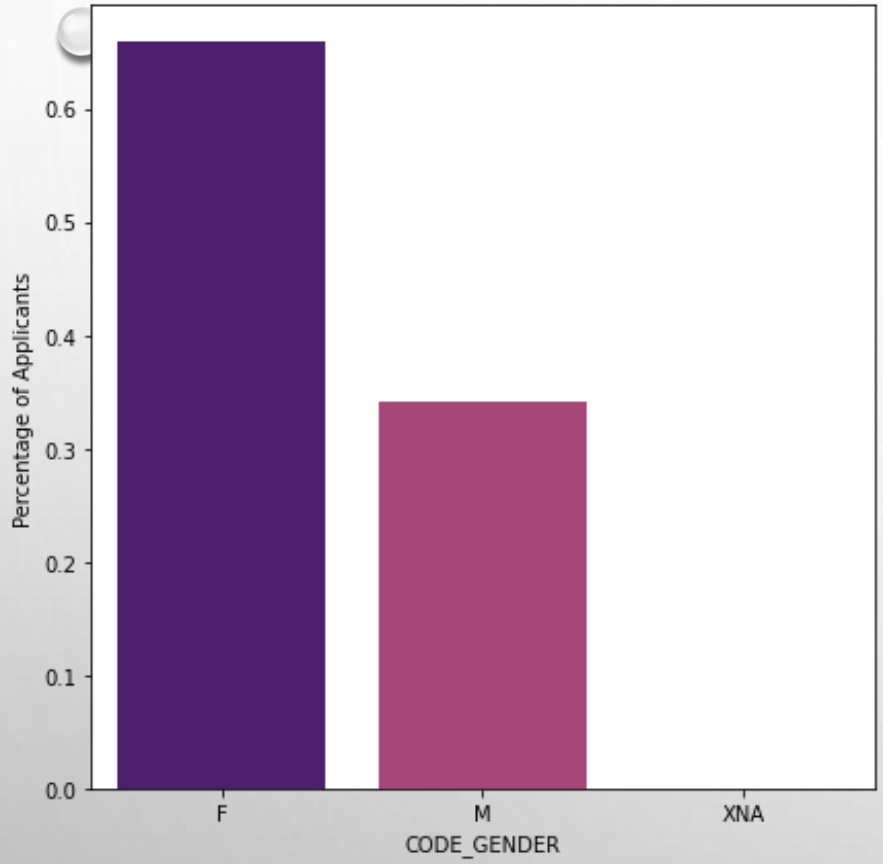


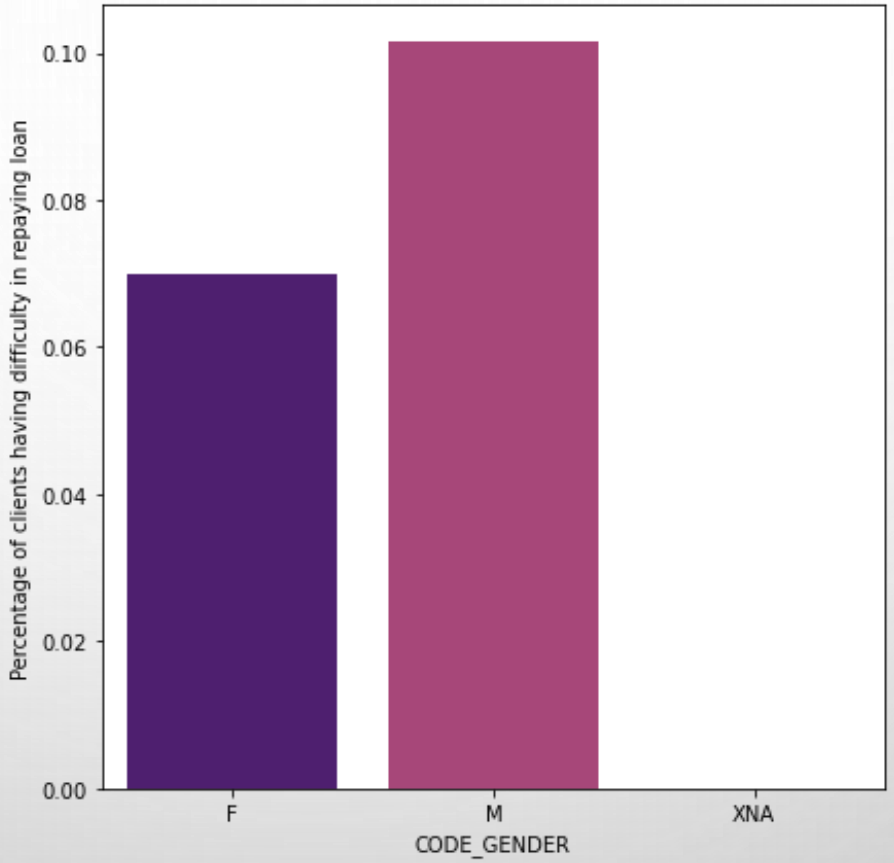


1.Major portion of the clients have opted for Cash loans(Approximately 85-90%), whereas 10-15% have opted for Revolving loans.

2.The mean for Cash loans is approximately 8.5%, whereas for revolving loans it is approximately 5.5%. This shows that the clients who have opted for cash loans are more likely to default.

UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS ON GENDER

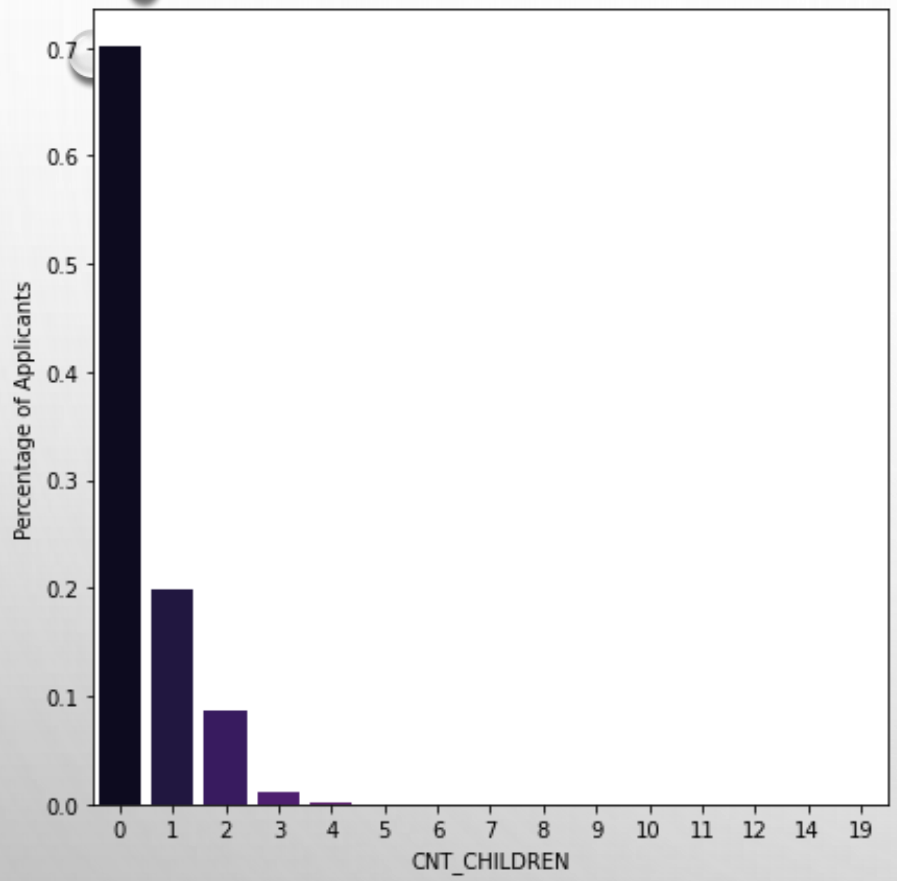


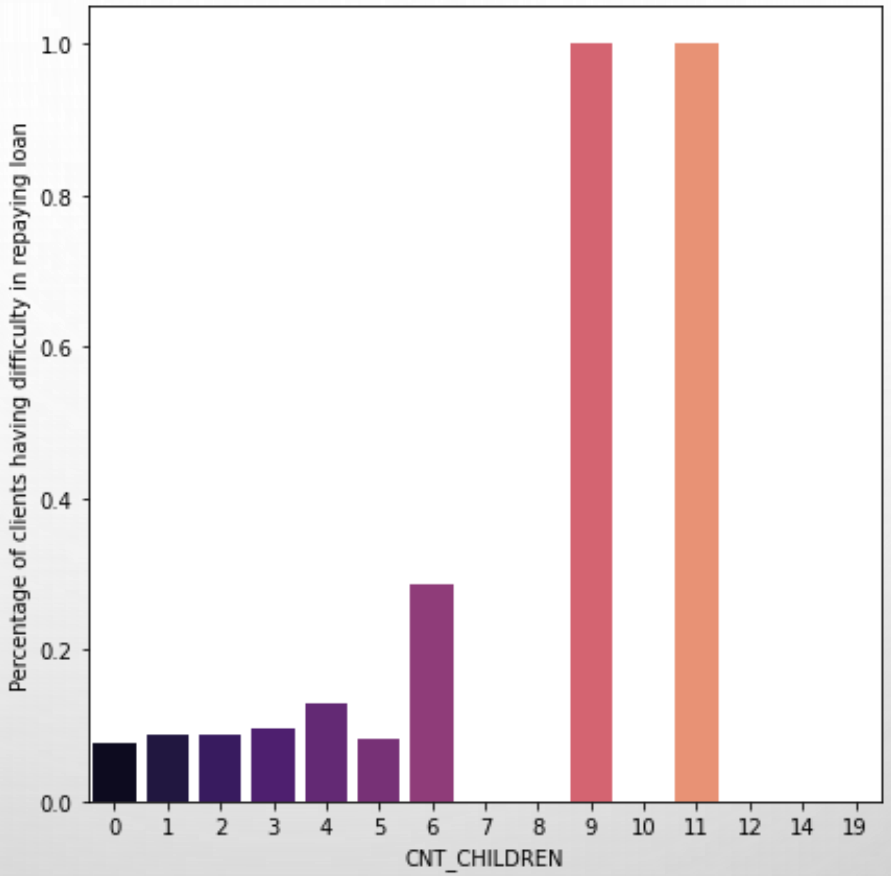


1.There are considerably more female clients (approximately 65%) than male clients (approximately 35%)

2.Male clients are more likely to default (approximately 10%) when compared to female clients (approximately 7%)

UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS ON CHILDEREN COUNT

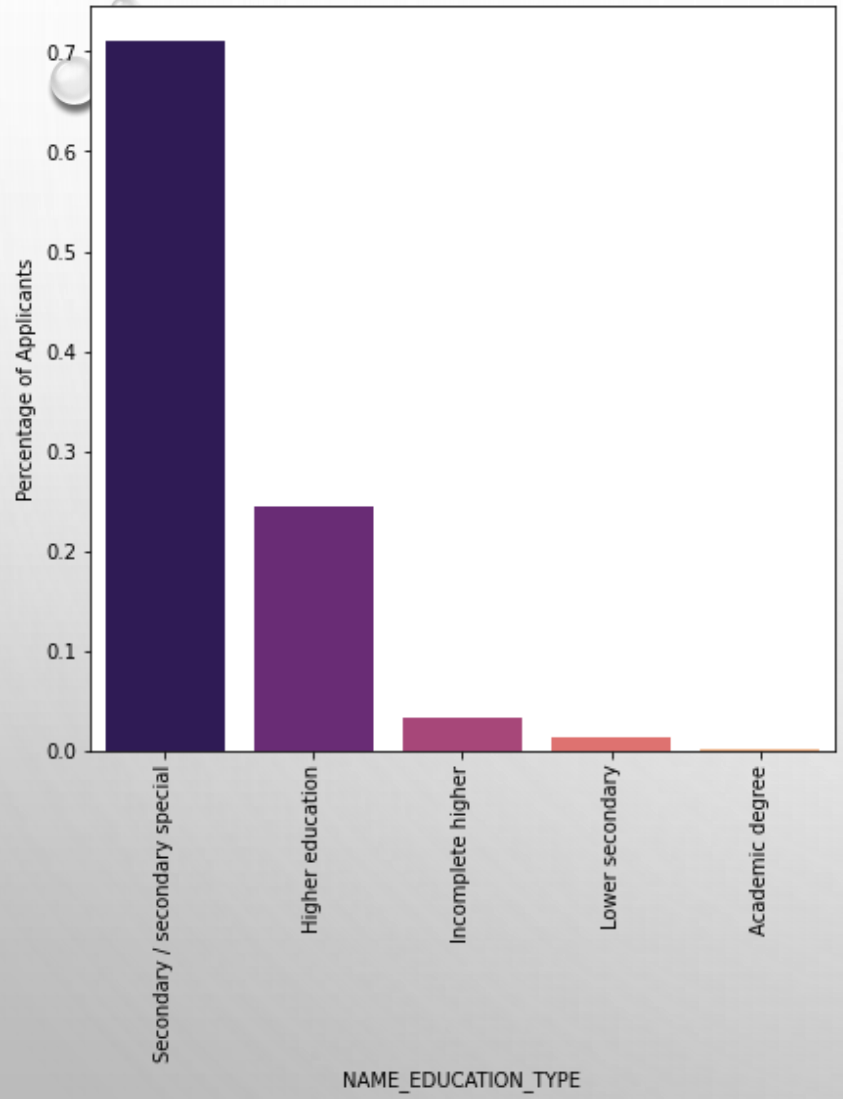


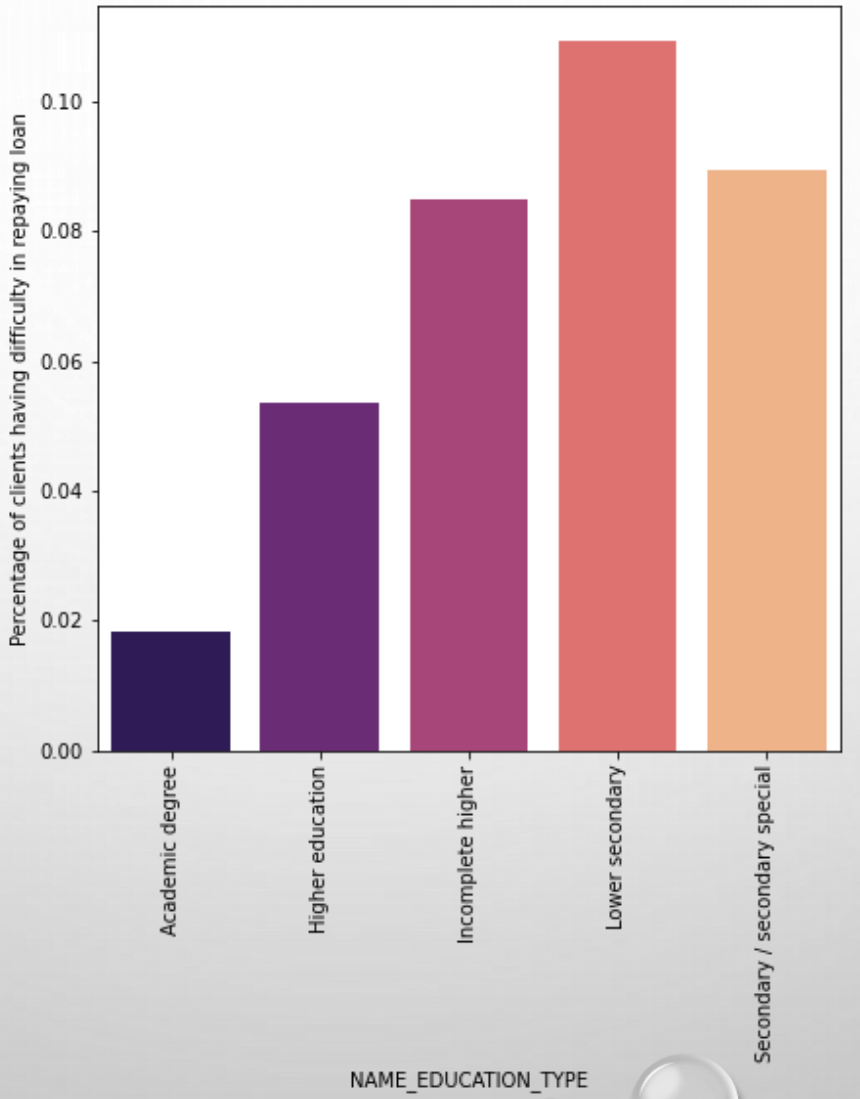


1.Majority of the clients do not have children (approximately 70%), and about 20% of the applicants have 1 child and less than 10% of the clients have 2 children.

2.Looks like the chances of not repaying the loan increases with the number of children the clients has. And the clients who have 9 or 11 children have approximately 100% chance of defaulting.

UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS ON EDUCATION TYPE

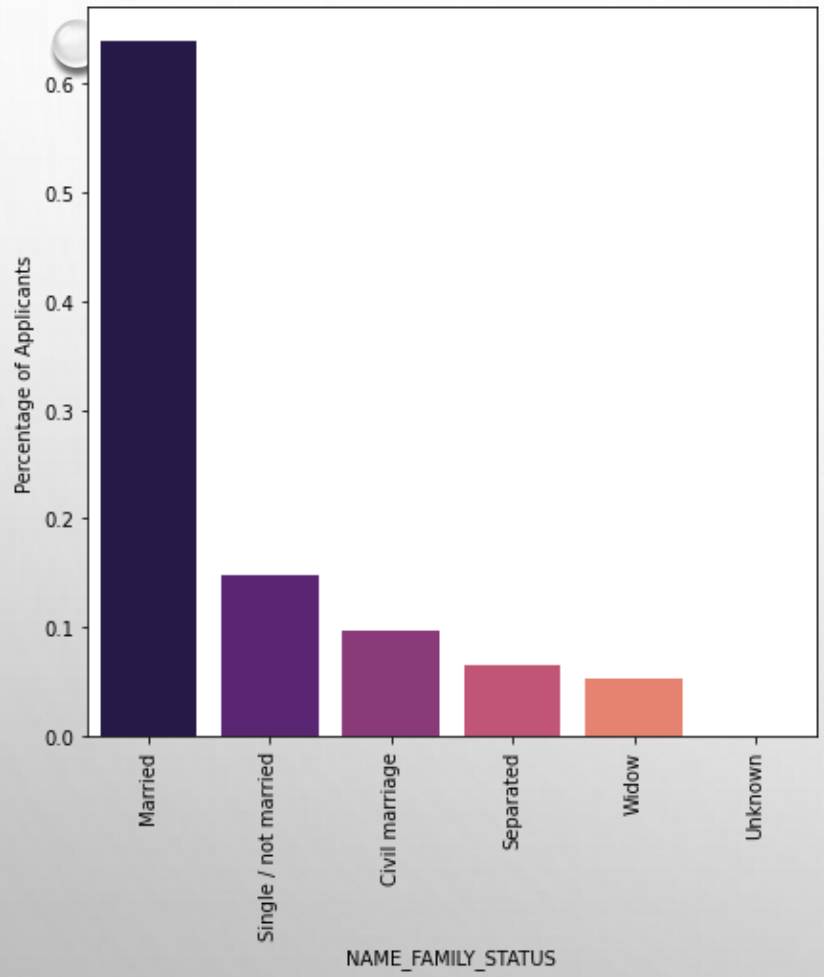


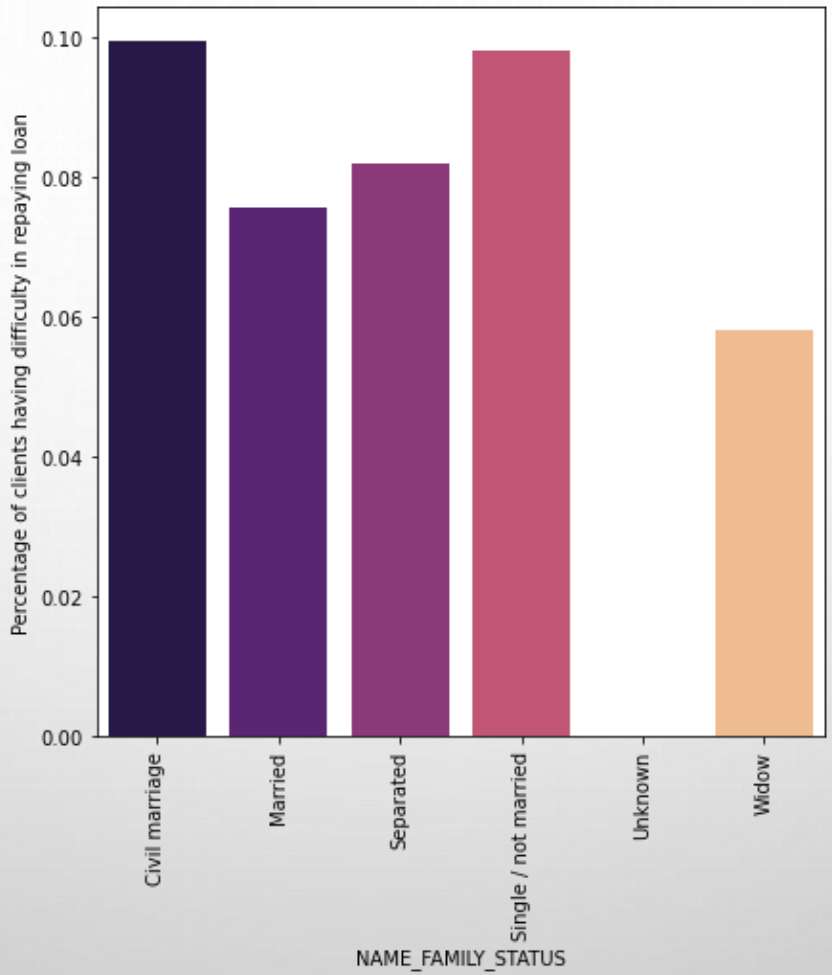


1.Majority of the clients have completed Secondary education (~70%), followed by clients who have completed higher education (about 25%), followed by incomplete higher education and lower secondary education.

2.Clients who have have completed lower secondary education are most likely to default (around 10%), followed by incomplete higher (around 8%) and secondary education (around 8%), and the clients who have an academic degree are least likely to default (nearly 2%)

UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS ON FAMILY STATUS

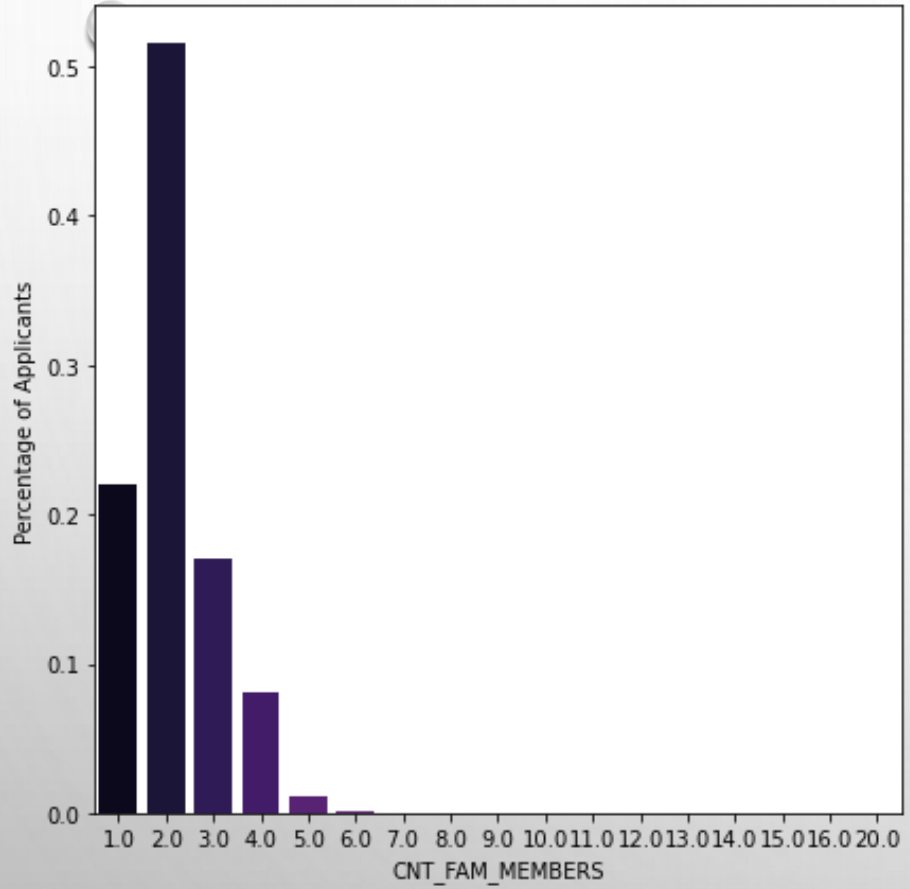


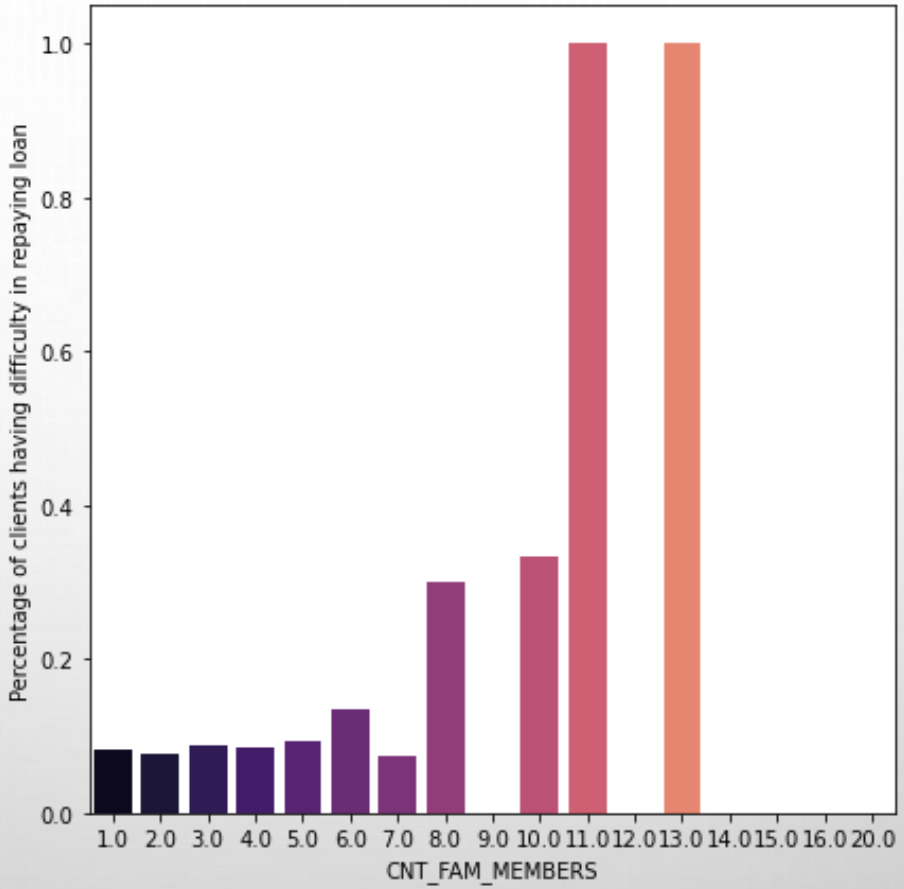


1.More than half of the clients are married (~60%), about 15% of the clients are not married, and about 10% of the clients have undergone civil marriage followed by separated and widows.

2.Single clients and clients who have undergone civil marriage are mosy likely to defualt (nearly 10%), followed by separated and married clients (approximately 8%) and widows have the least default rate of around 6%.

UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS ON FAMILY MEMBER COUNT

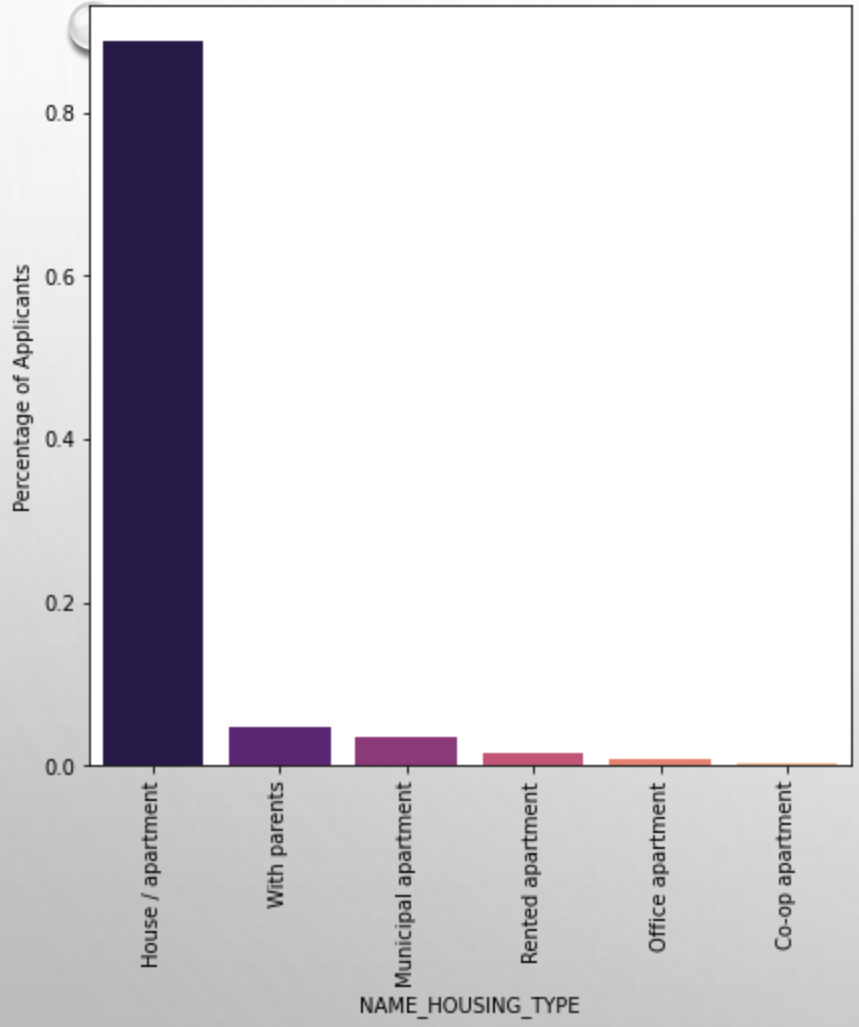


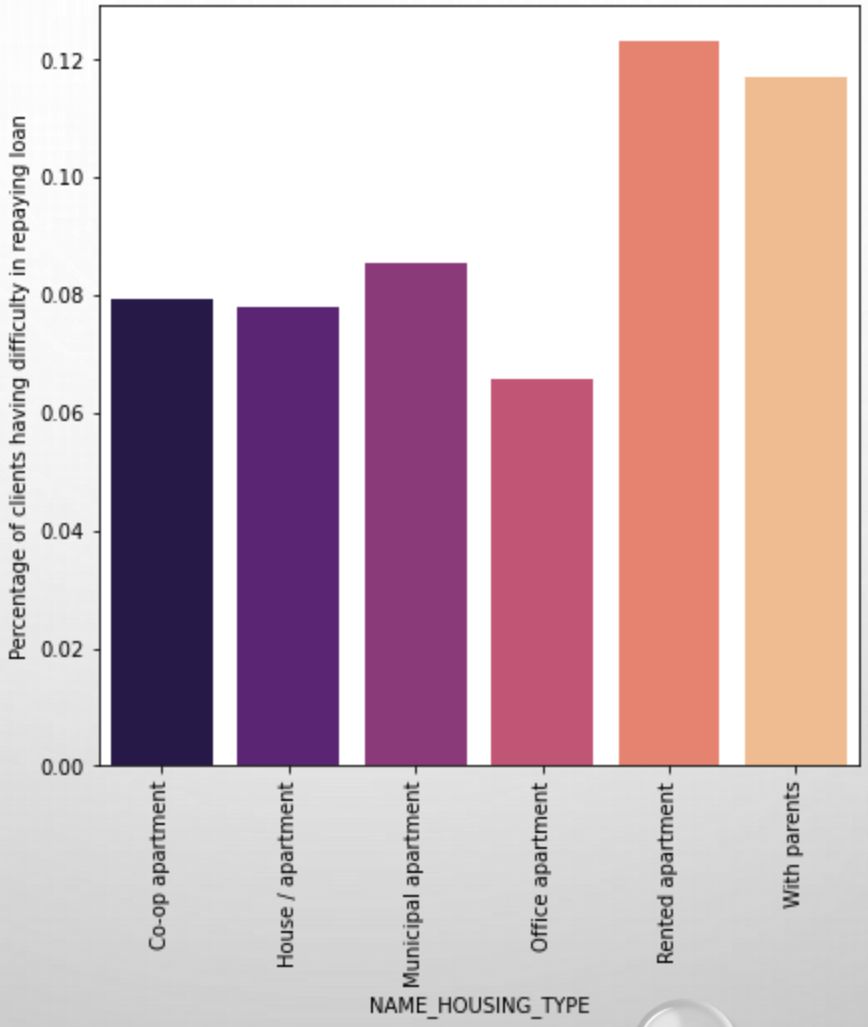


1.Clients with family members of 2 are the top category (approximately 50%), followed by 1 (single people), 3 (probably families with 1 child) and 4 family members.

2.Looks like the default rate increases with the increase in family members, also clients with 11 and 13 family members have 100% default rate.

UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS ON HOUSING TYPE

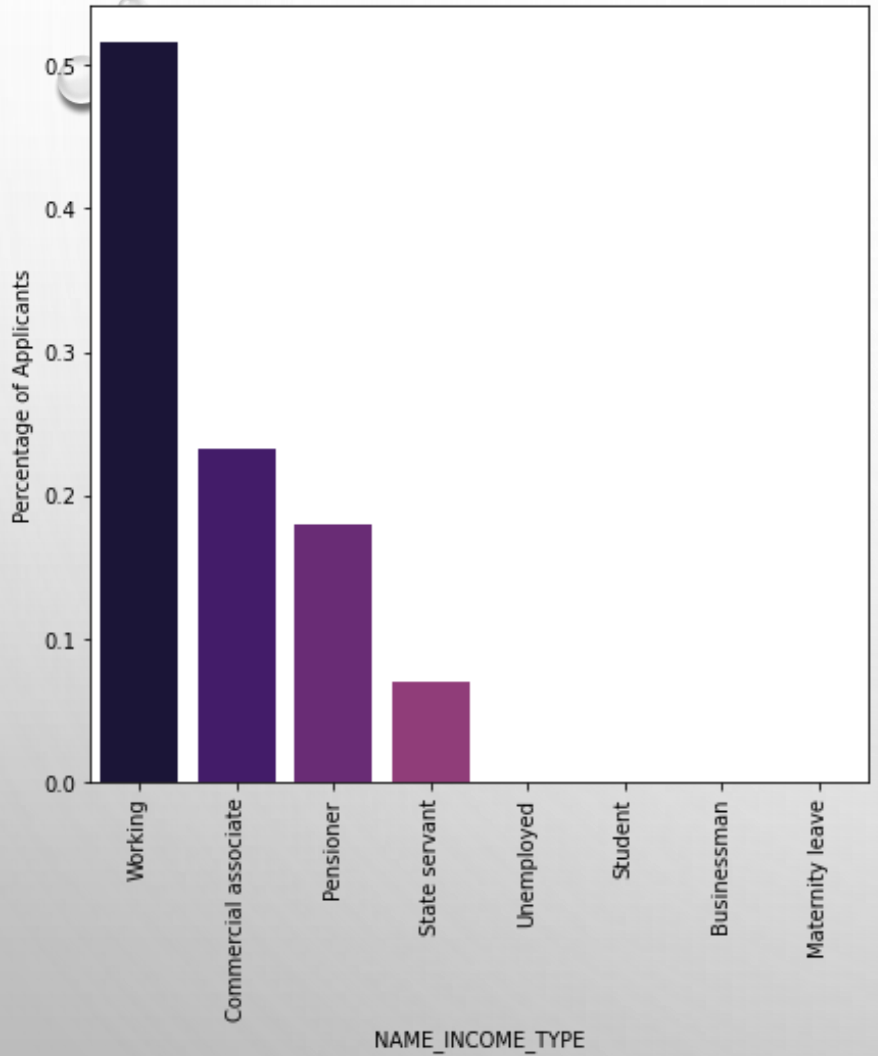


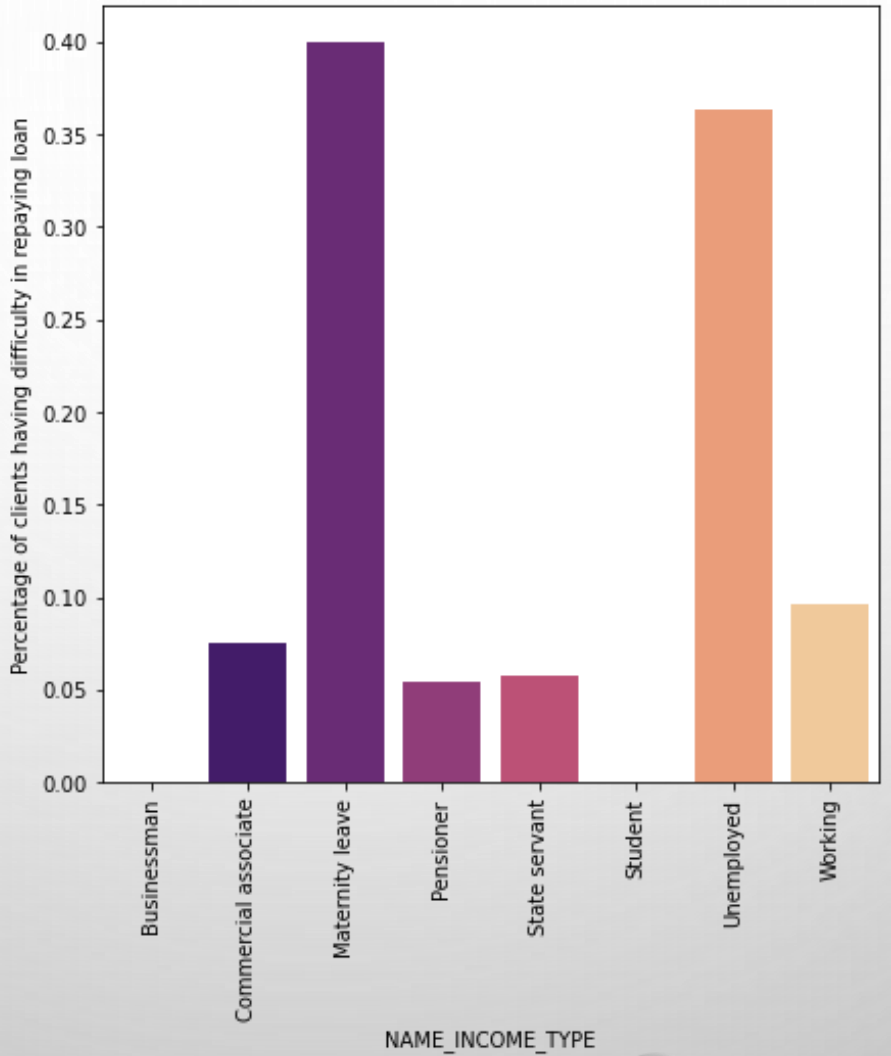


1.Almost 90% of the clients live in a House/apartment, followed by living with parents, municipal apartment etc.

2.Clients who live in a rented apartment have the highest default rate of around 12%, followed by clients living with parents (almost 12%), followed by municipal apartments etc.

UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS ON INCOME TYPE

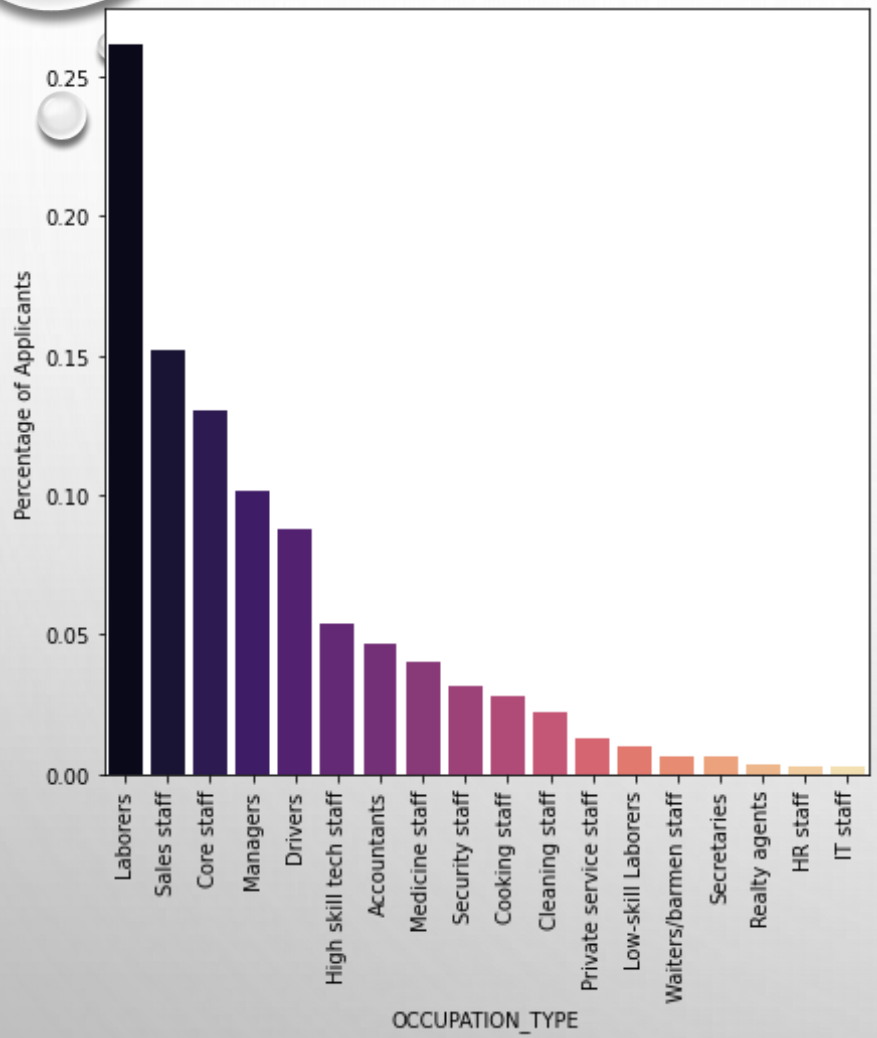


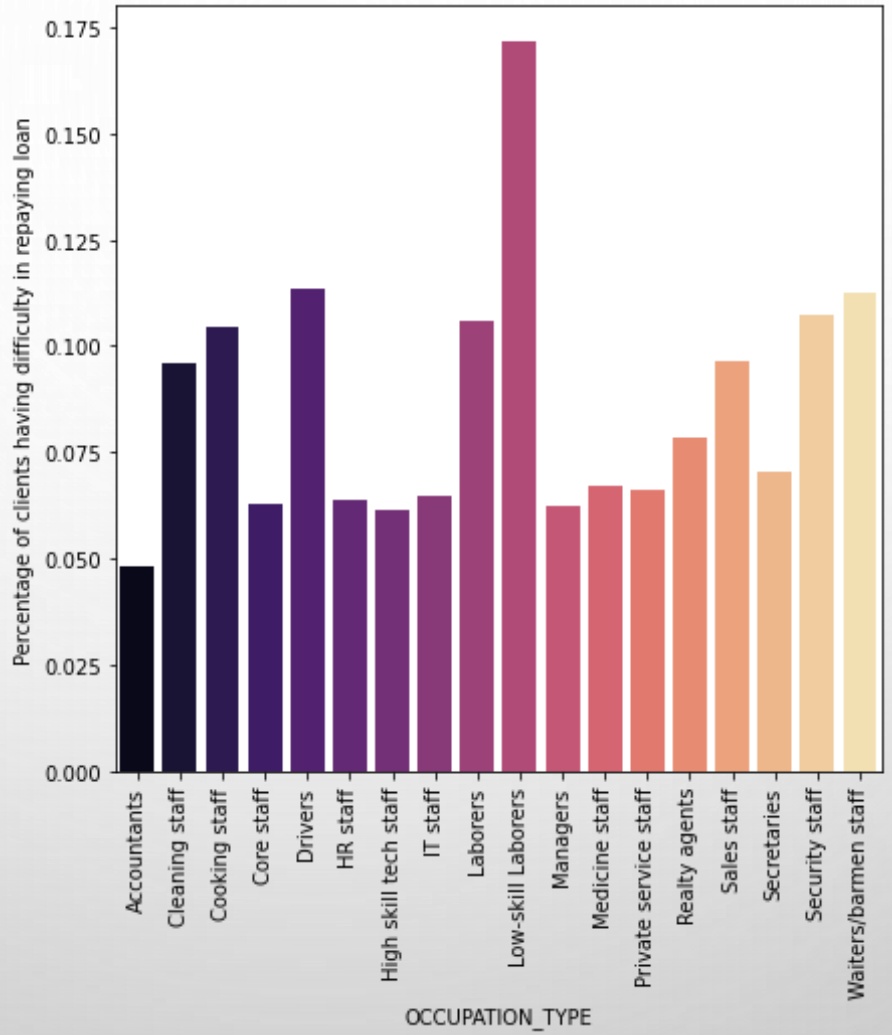


1.More than half of the clients are working, about 20-25% are commercial associates, nearly 20% are pensioners and nearly 10% are state servants.

2.Clients who are on maternity leave and unemployed clients are most likely to default (approximately 35-40%), the remaining income types have around 10% chance of defaulting.

UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS ON OCUPATION TYPE

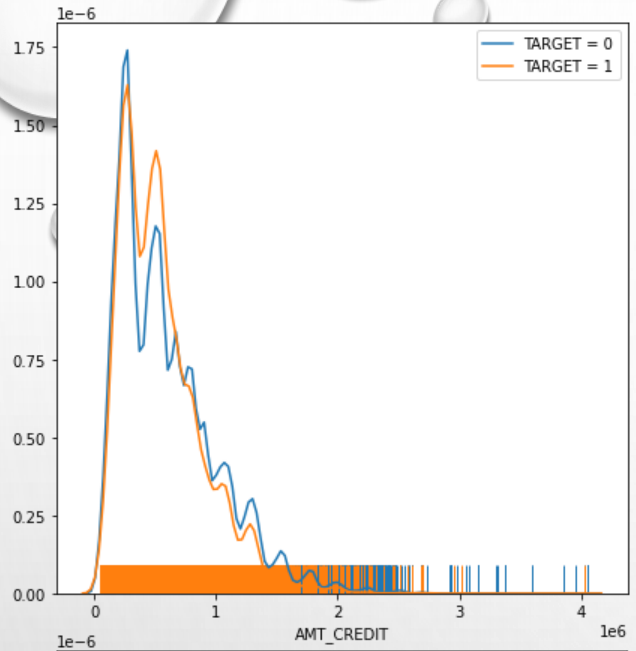


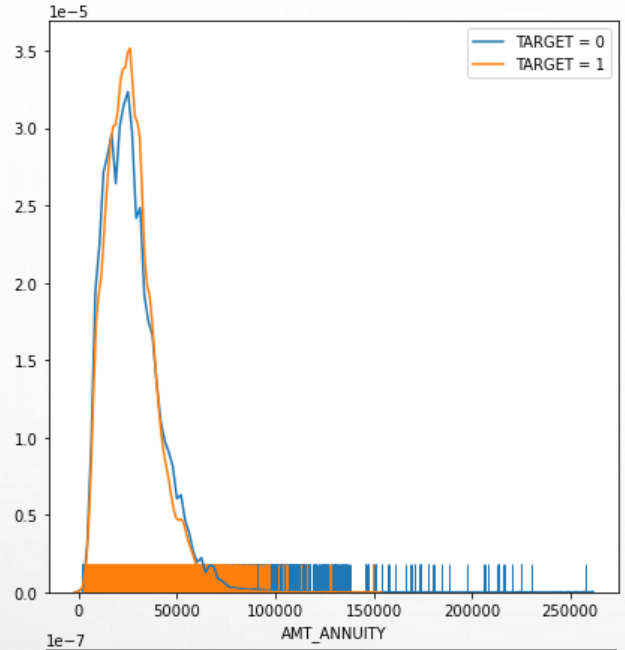


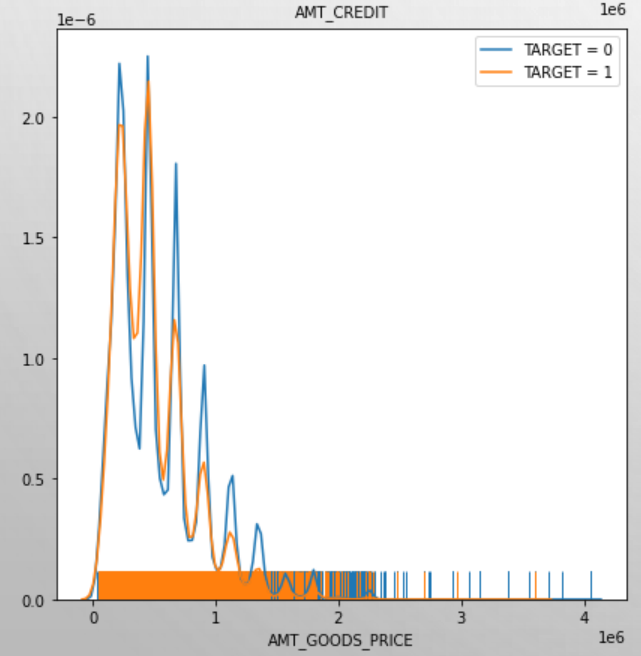
1.Most of the clients are Laborers(approximatey 25%), followed by Sales staff, Core staff, Managers etc. IT staff take the least number of loans.

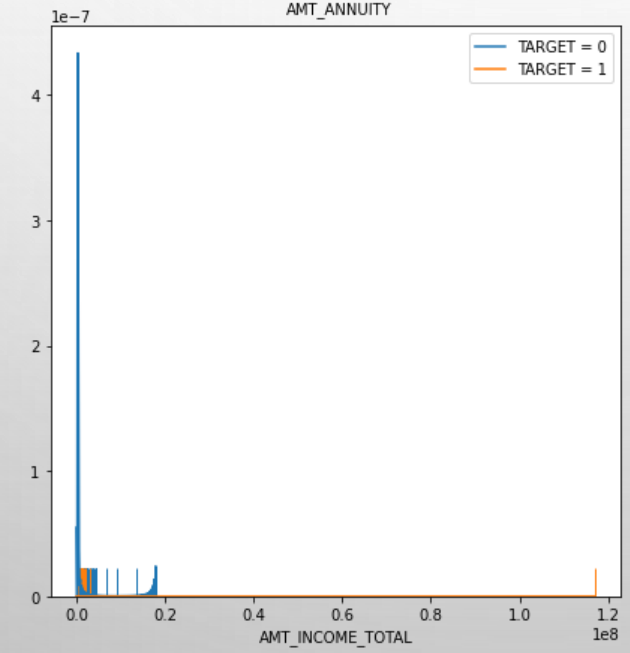
2.Low-skill labourers have the highest default rate, followed by Drivers, Waiters/barmen staff, Security staff, Laborers, Cooking staff, sales staff, cleaning staff etc.

DENSITY PLOT OF CONTINIOUS VARIABLES









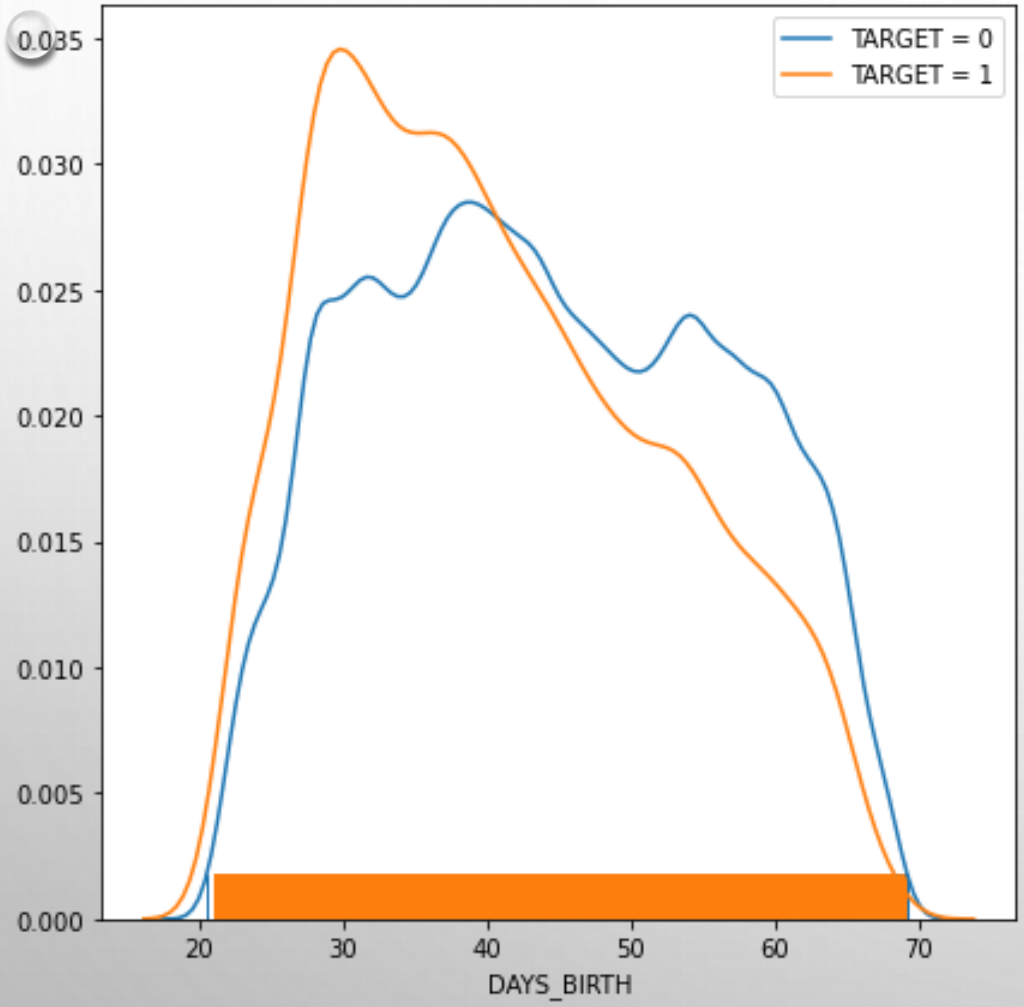
1.The credit amount of the loan lies mostly around 200000-1000000 for both, client with payment difficulties(TARGET=1) and all other cases(TARGET=0)

2.The annuity amount seems to be similar for both target types, they are concentrated at around 50000.

3.The goods for which the clients have received loan are mostly concentrated between 200000-1000000

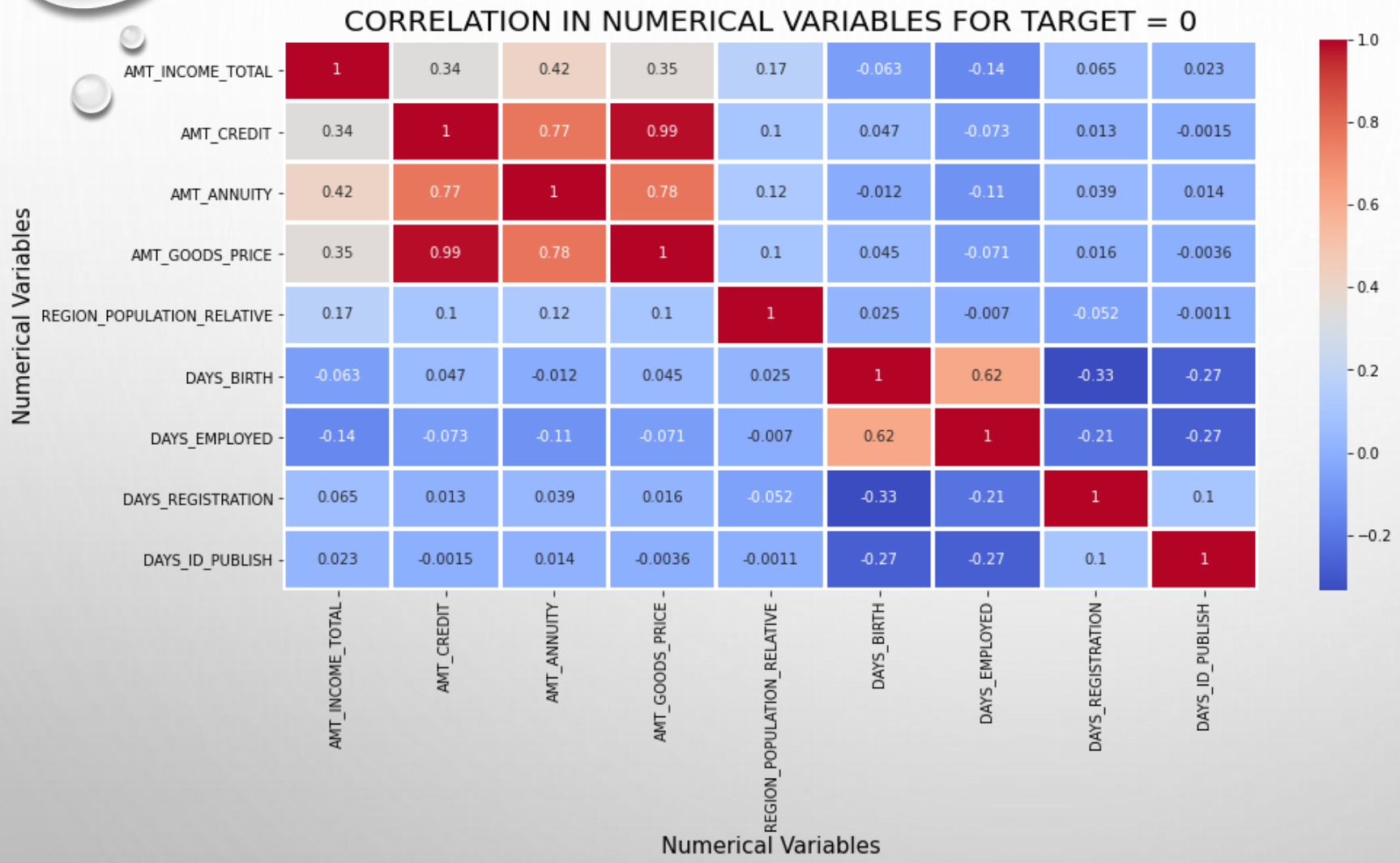
4.The income of clients of both target types mostly lie at around 10000000. We can see some outliers as well.

DENSITY PLOT OF CONTINIOUS VARIABLE



- This plot has been converted from days to years
- When we compare the two target types, we see that the clients who have difficulty in payment are relatively younger and most of them lie at around 27 years old. Therefore the younger population have the highest default rate.

BIVARIATE NUMERICAL VARIABLE ANALYSIS



1.Credit amount and Goods price amount have the highest correlation of 0.99, which is fairly obvious since the client is opting for a loan equal to the price of his goods.

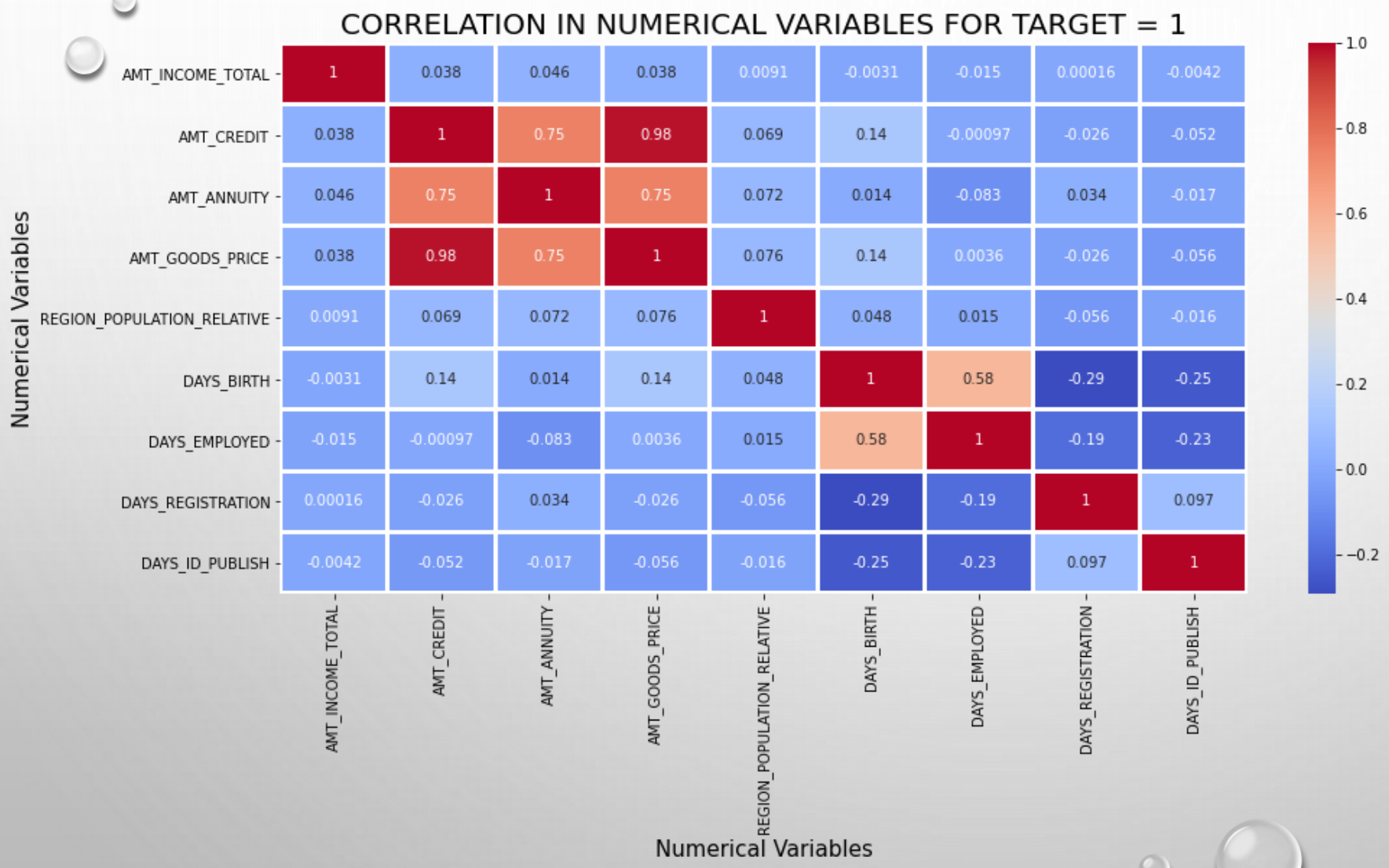
2.Annuity amount and goods price amount have a correlation of 0.78. Since the annuity is decided by the price of the goods (credit amount).

3.Credit amount and annuity amount have a correlation of 0.77. As the annuity amount depends on the credit amount.

4.Income amount and annuity amount has a correlation of 0.42. Annuity might be decided based on the income of the client.

5.Correlation between the income amount and credit amount is pretty good (0.34) and also between annuity amount and income amount(0.42), might be the reasons why these clients are able to repay their loans.

BIVARIATE NUMERICAL VARIABLE ANALYSIS



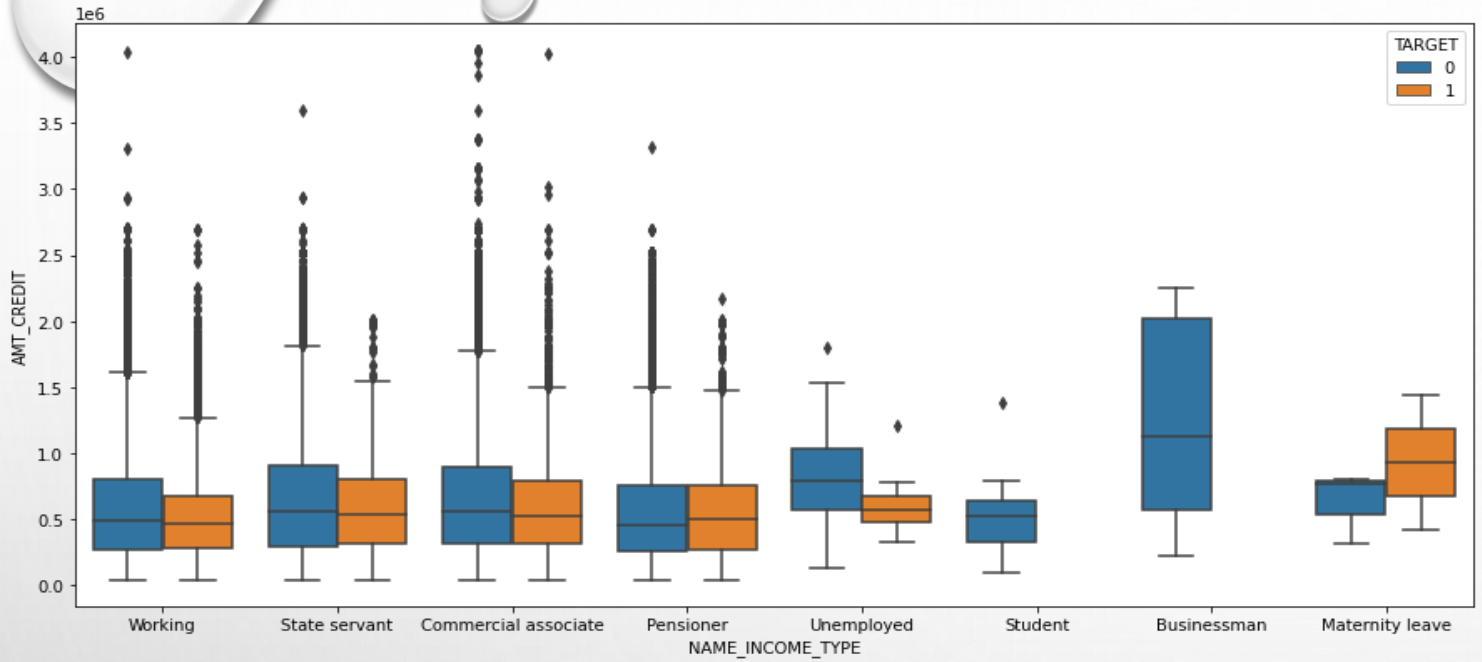
1.Annuity amount and goods price amount have a correlation of 0.75. Since the annuity is decided by the price of the goods (credit amount).

2.Credit amount and Goods price amount have the highest correlation of 0.98, which is fairly obvious since the client is opting for a loan equal to the price of his goods.

3.Credit amount and annuity amount have a correlation of 0.75. As the annuity amount depends on the credit amount.

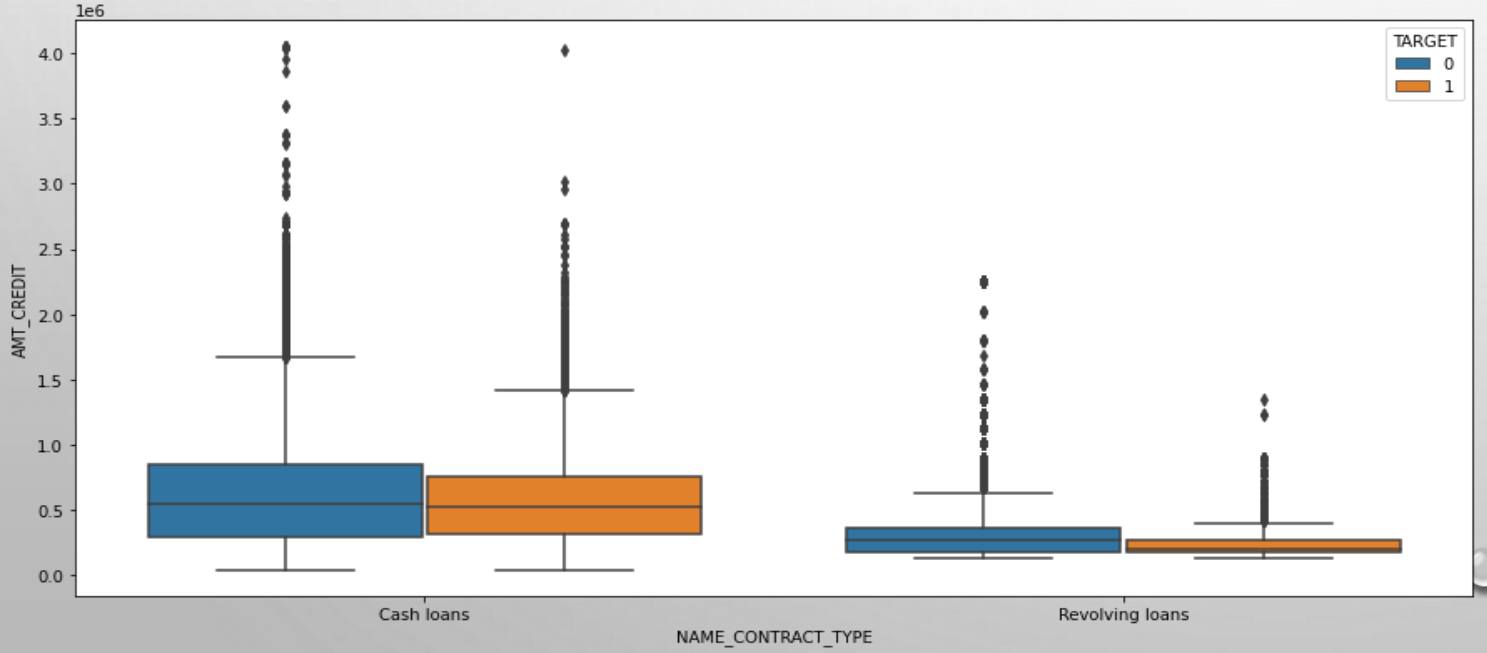
4.Correlation between the income amount and credit amount is very less (0.038) and also between annuity amount and income amount(0.046), might be the reasons for difficulty in repaying

BIVARIATE ANALYSIS ON CONTINIOUS AND CATEGORICAL VARIABLE



1.The clients who are businessmen and students do not have a boxplot for TARGET=1 which shows there are no defaulters

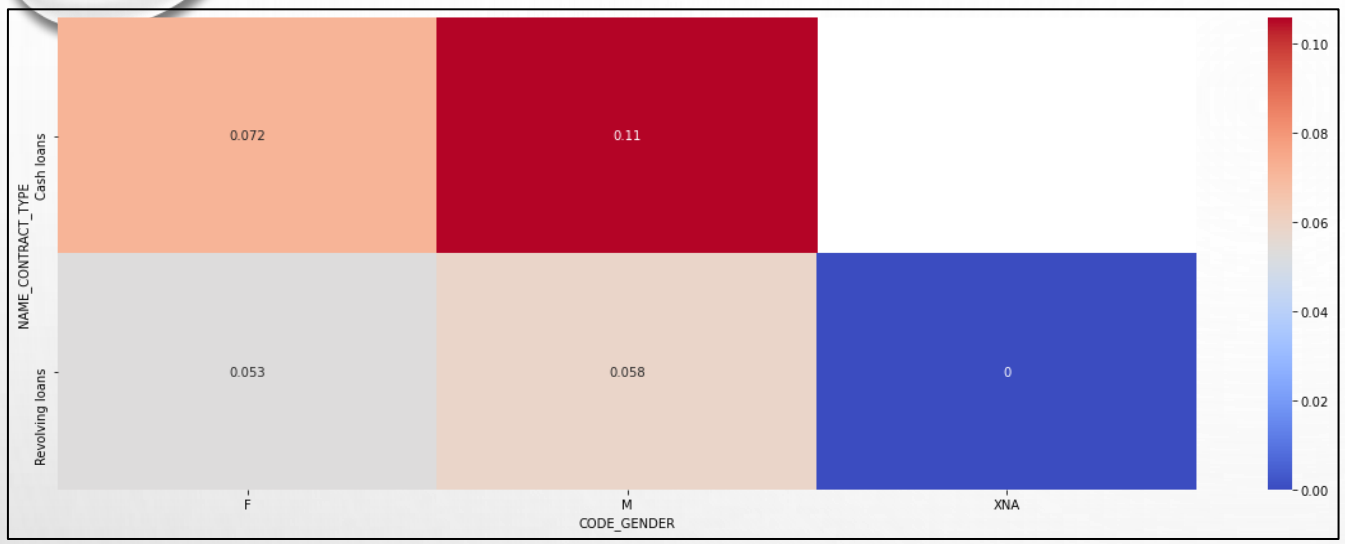
2.The businessmen have a big IQR which suggests that the credit amount has a huge range.

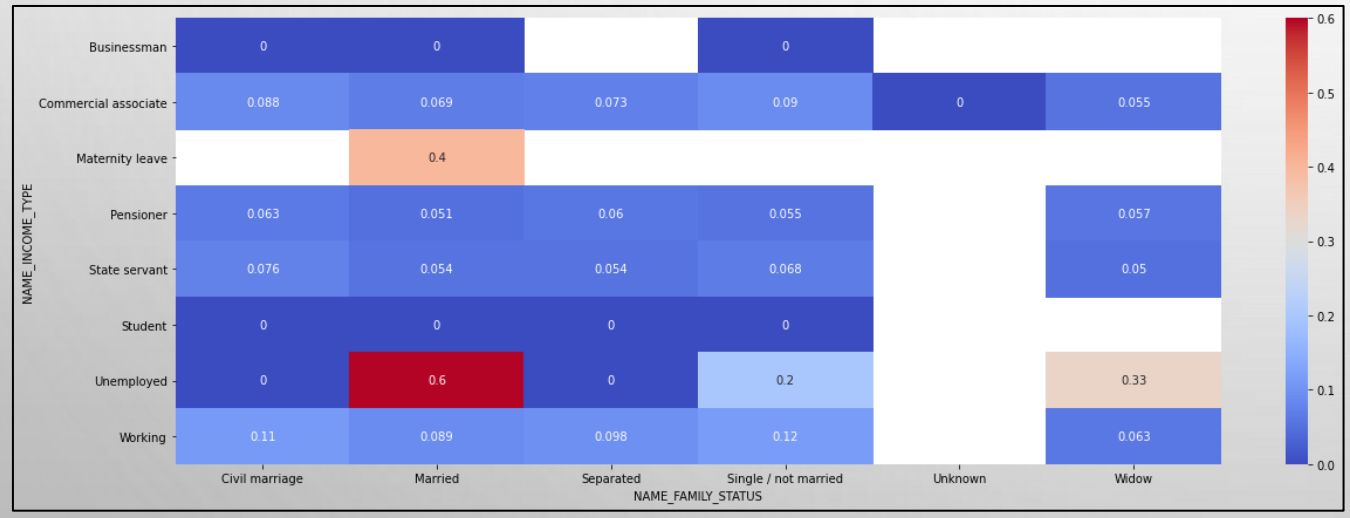


1.It is clearly evident that the credit amounts are higher for cash loans when compared to revolving loans

2.The credit amount is concentrated at the lower end of the IQR for TARGET=1, which suggests that most of the clients who have difficulty in repaying the loans have lesser credit amount.

BIVARIATE CATEGORICAL VARIABLE ANALYSIS BY CONSIDERING THE TARGET VARIABLE VALUE





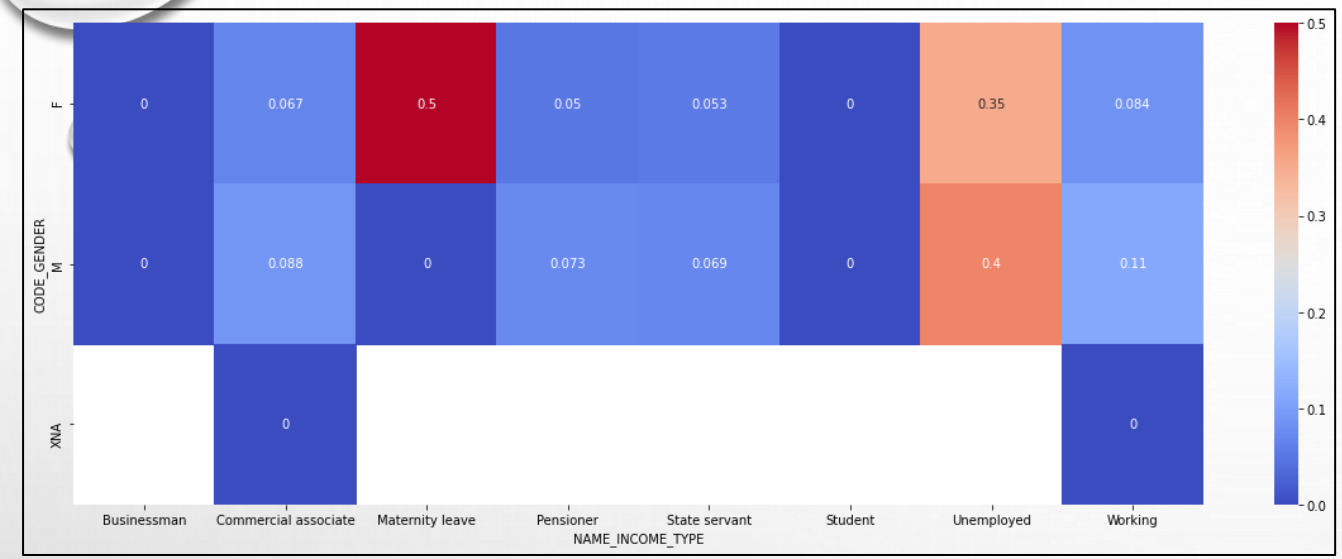
1.Male clients who have opted for cash loans are most likely to default.

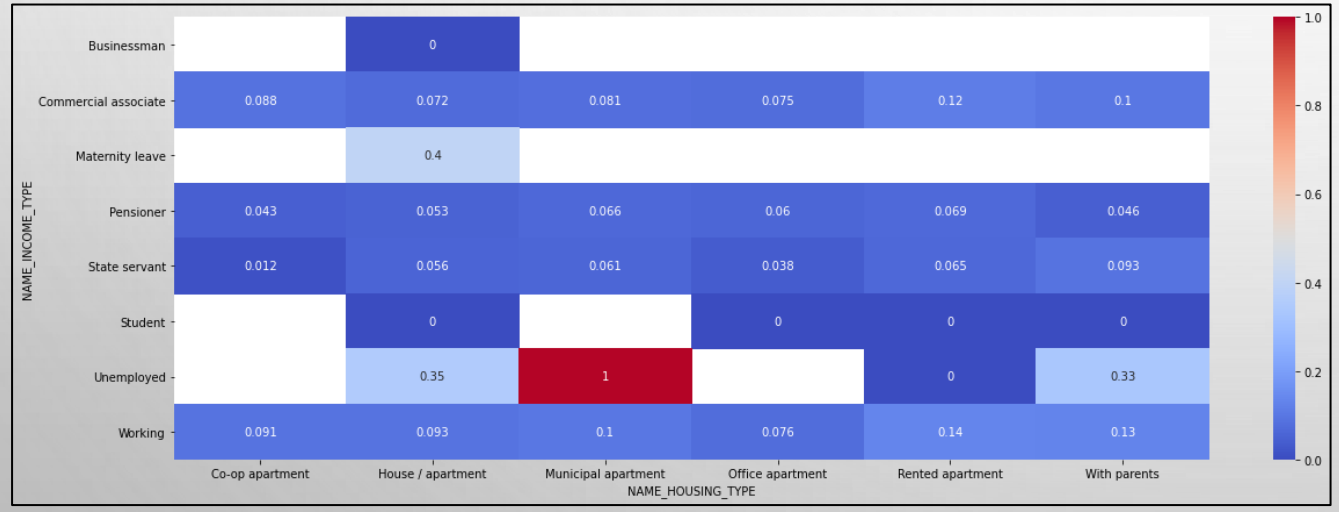
2.Clients who are unemployed and married are most likely to default.

3.Clients on maternity leave and married have a high default rate.

4.Unemployed clients who are also widows have a high default rate as well.

BIVARIATE CATEGORICAL VARIABLE ANALYSIS BY CONSIDERING THE TARGET VARIABLE VALUE





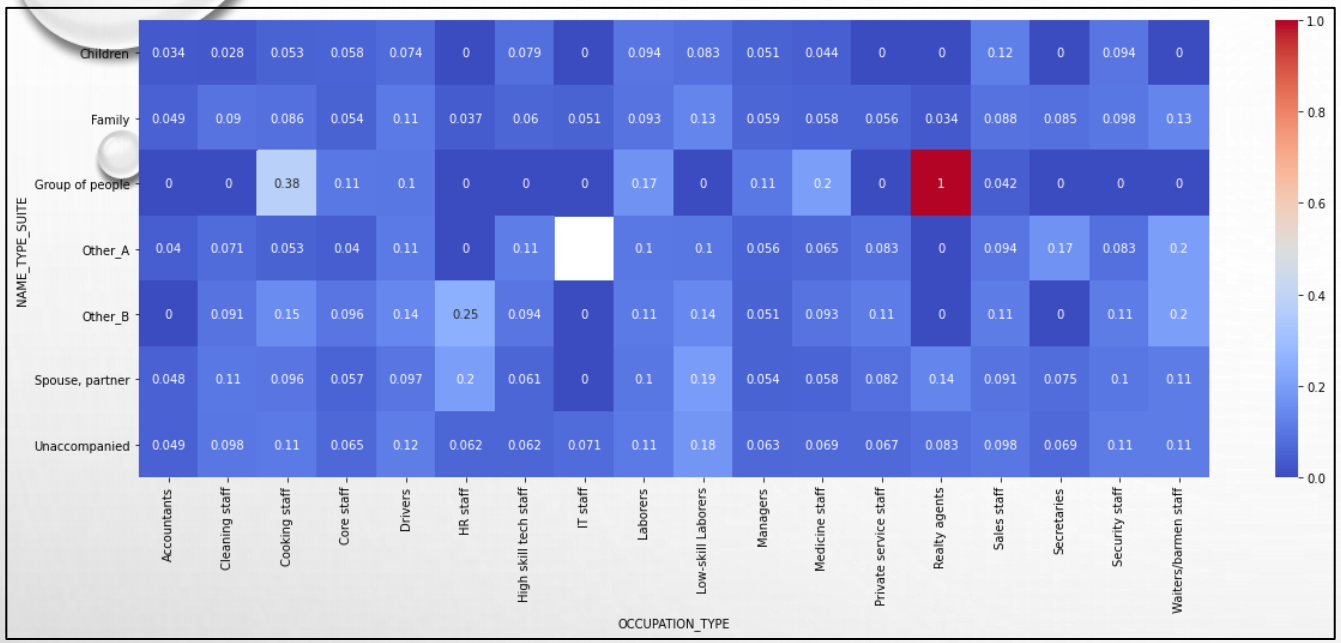
1.Women on maternity leave are most likely to default.

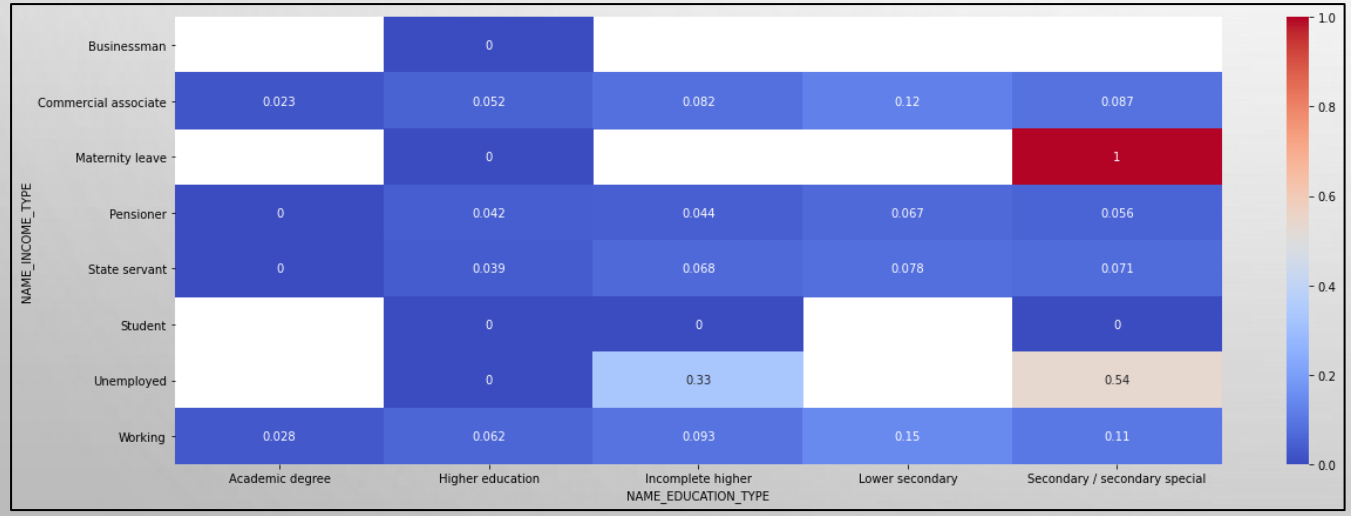
2.Also unemployed clients have a high default rate irrespective of gender.

3.Clients who are unemployed in general have high default rate, specifically the ones living in a municipal apartment.

4.Clients on maternity leave and living in a house/apartment also have high default rate.

BIVARIATE CATEGORICAL VARIABLE ANALYSIS BY CONSIDERING THE TARGET VARIABLE VALUE





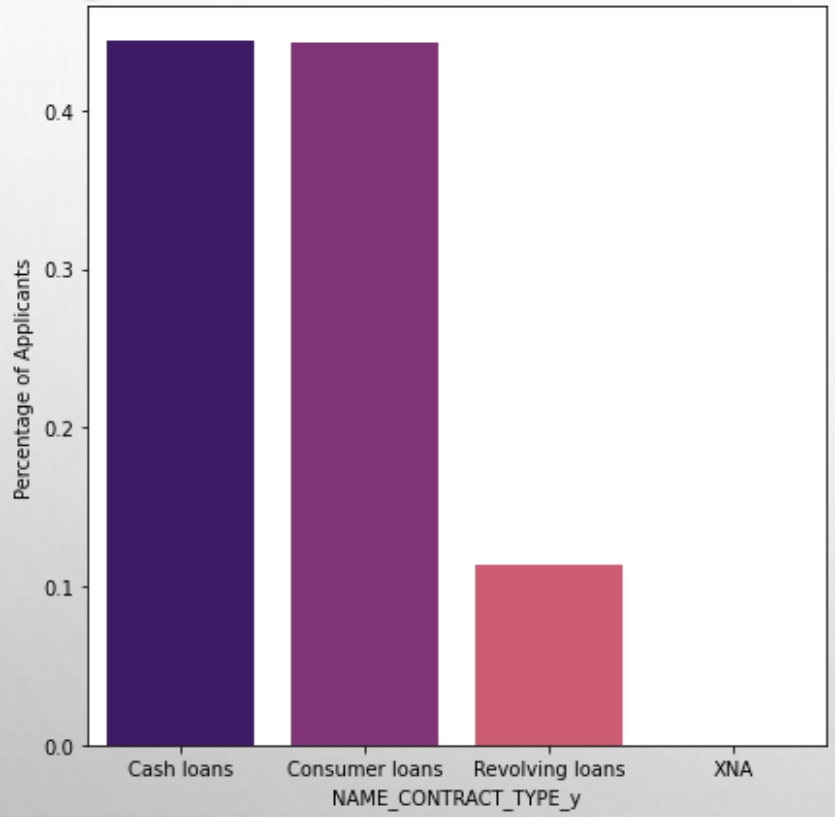
1.Clients who are realty agents and were accompanied by a group of people while applying for loan are most likely to default.

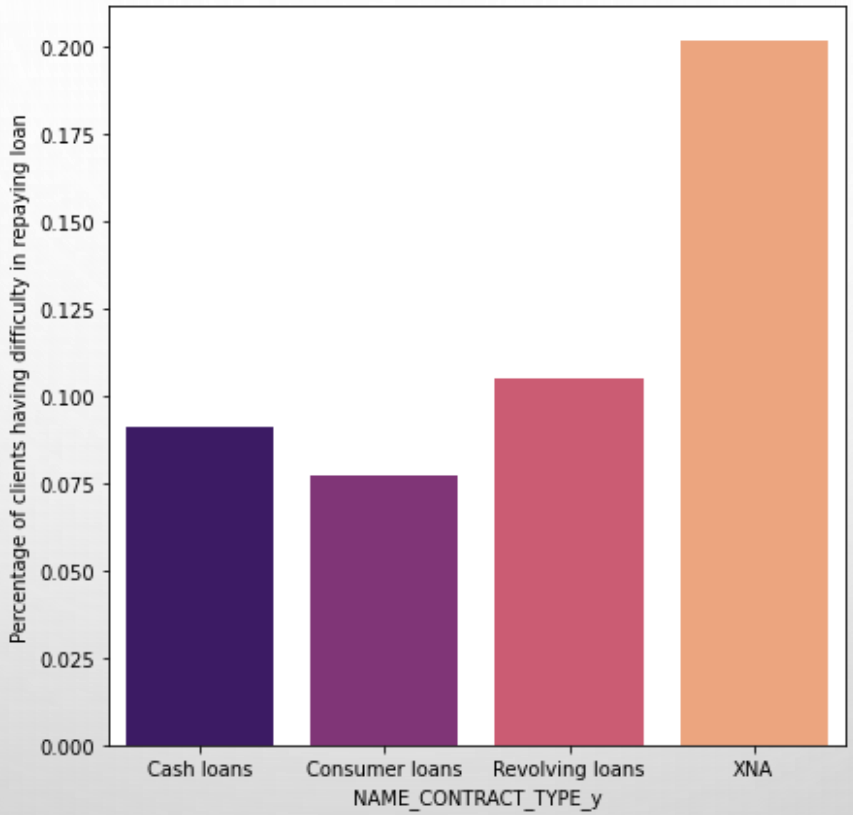
2.Cooking staff accompanied by a group of people while applying for loan have a high default rate.

3.Clients who are on maternity leave and have completed secondary education are most likely to default.

4.Unemployed clients who have completed secondary education have a high default rate.

UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS ON CONTRACT TYPE (Merged Data)

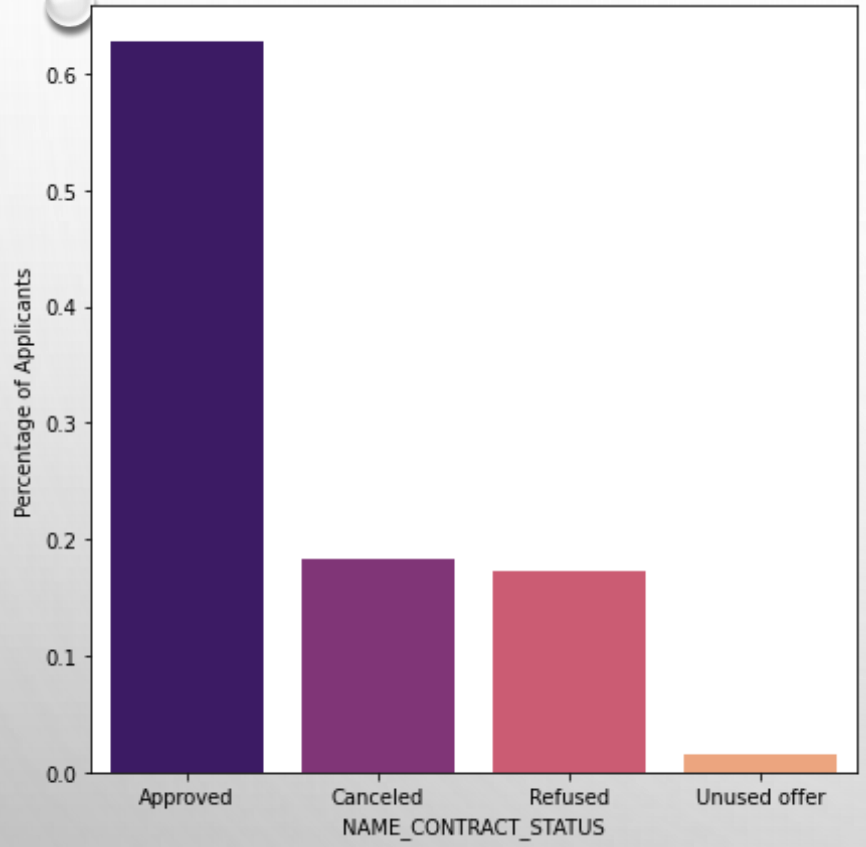


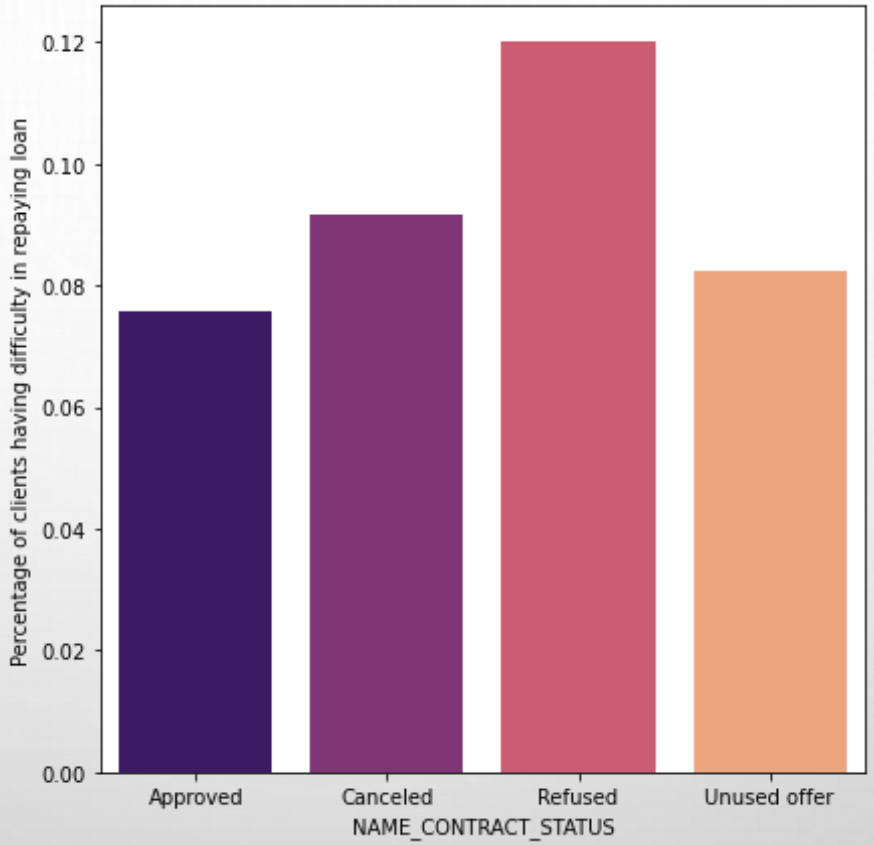


1.From the previous data most of the clients have opted for cash and consumer loans (approximately 45% each) and about 1% have opted for revolving loans.

2.Clients who have opted for revolving loans have the highest default rate, followed by cash loans and consumer loans. Also to be noted that revolving loans category was not present in application data.

UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS ON CONTRACT STATUS (Merged Data)

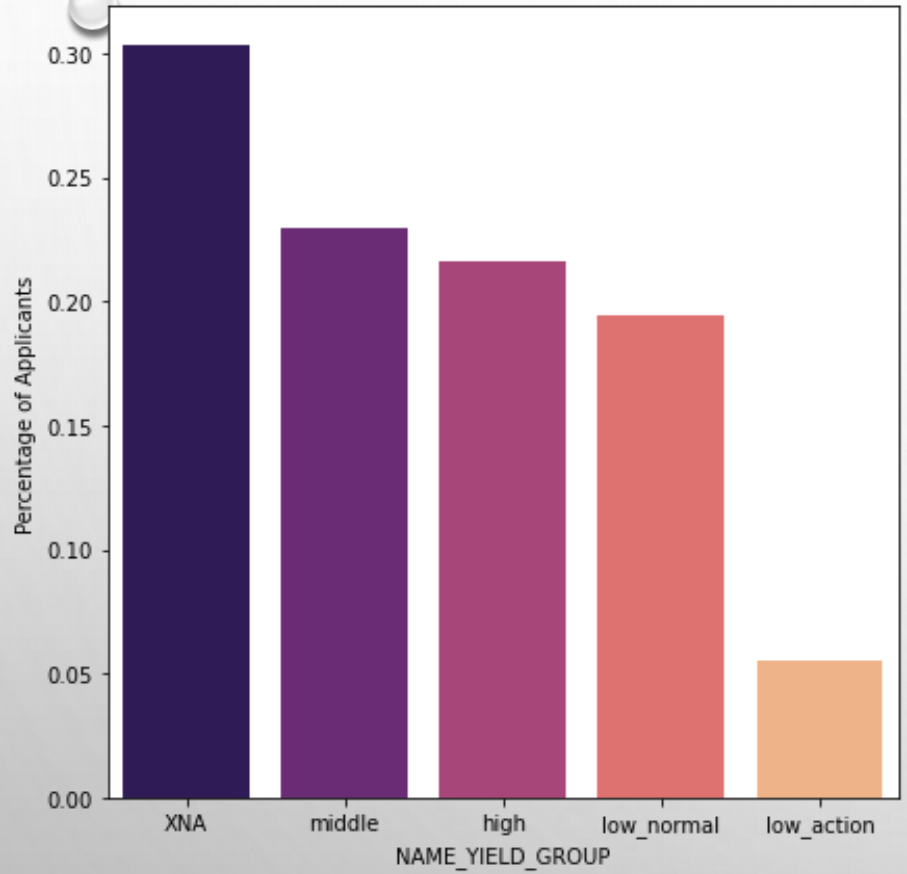


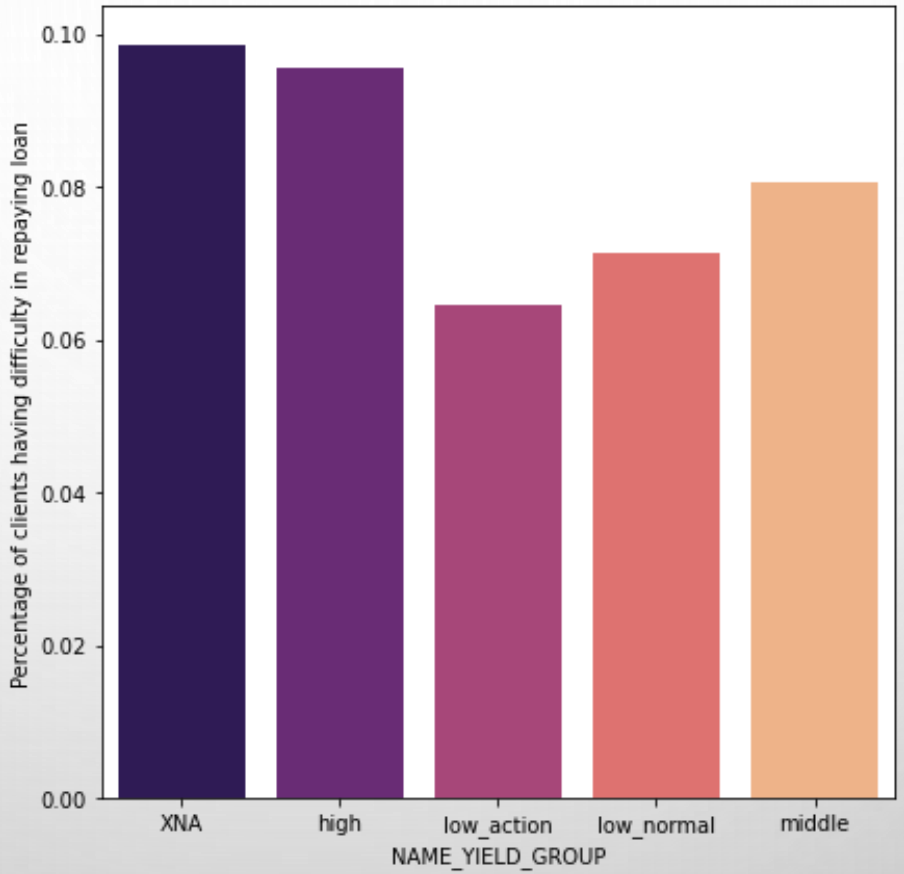


1.Majority of the clients had their loans approved (about 60%), followed by clients who have cancelled their application (nearly 20%) followed by clients who had their applications rejected (nearly 20%).

2.The clients who got their applications rejected are most likely to default (about 12%), followed by clients who cancelled (around 9%), followed by clients who cancelled at different stages of the application process. As expected, the clients who had their applications approved are least likely to default.

UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS ON YIELD GROUP (Merged Data)

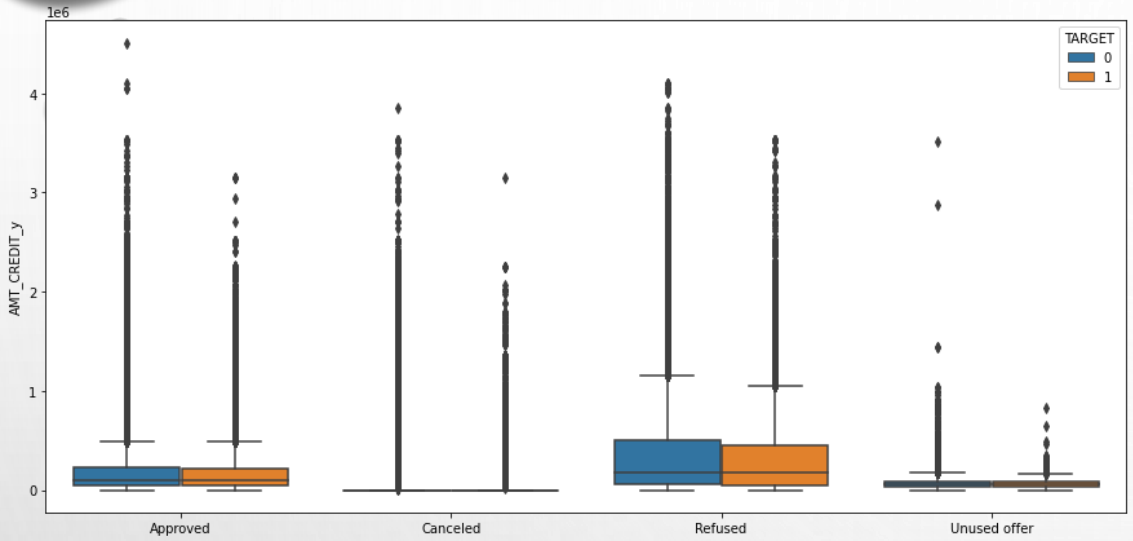




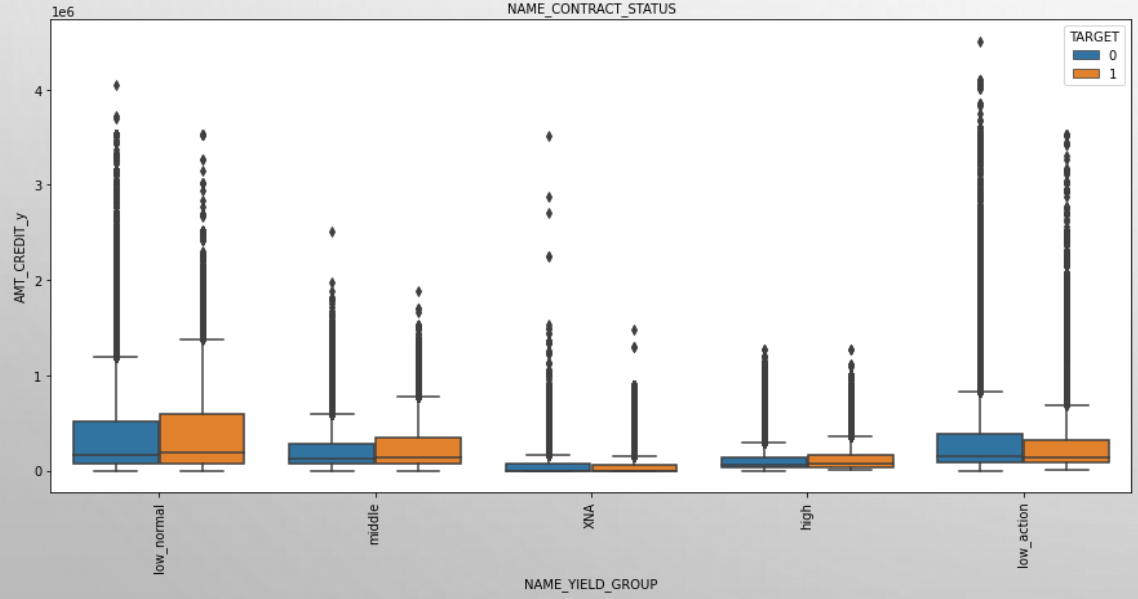
1.Around 20% of the clients have middle, followed by high and low_normal interest rates.

2.Clients with high interest rate are most likely to default (nearly 10%), followed by middle , low_normal.

BIVARIATE ANALYSIS
ON CREDIT AMOUNT & CONTRACT STATUS (Merged Data)



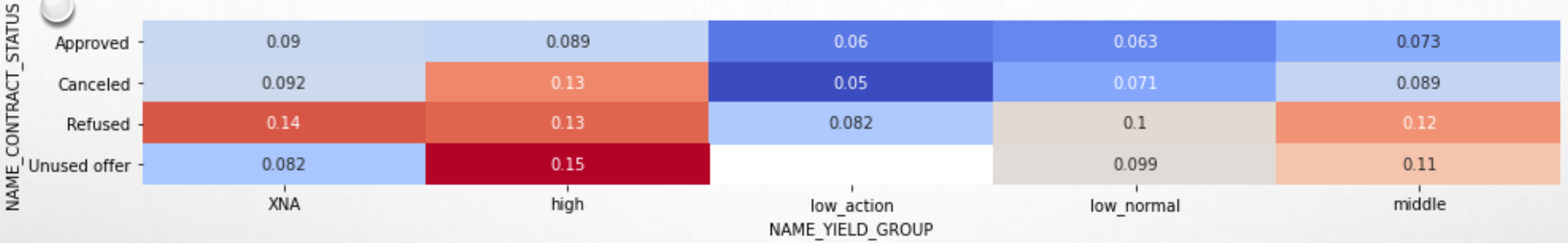
1.The loans have been refused most likely due to clients demanding high credit amounts, which is evident by the IQR being in the higher range.



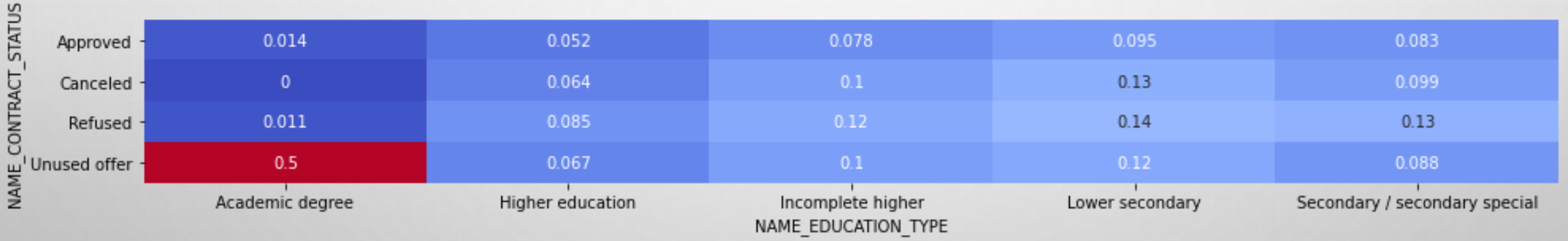
1.It is evident that when the clients are offered higher credit amounts their interest rate is relatively lesser when compared to lower credit amounts.

2.Therefore higher the credit amount lesser is the interest rate.

BIVARIATE CATEGORICAL VARIABLE ANALYSIS OF CONTRACT STATUS AND OTHER VARIABLES BY CONSIDERING THE TARGET VALUE (Merged data)

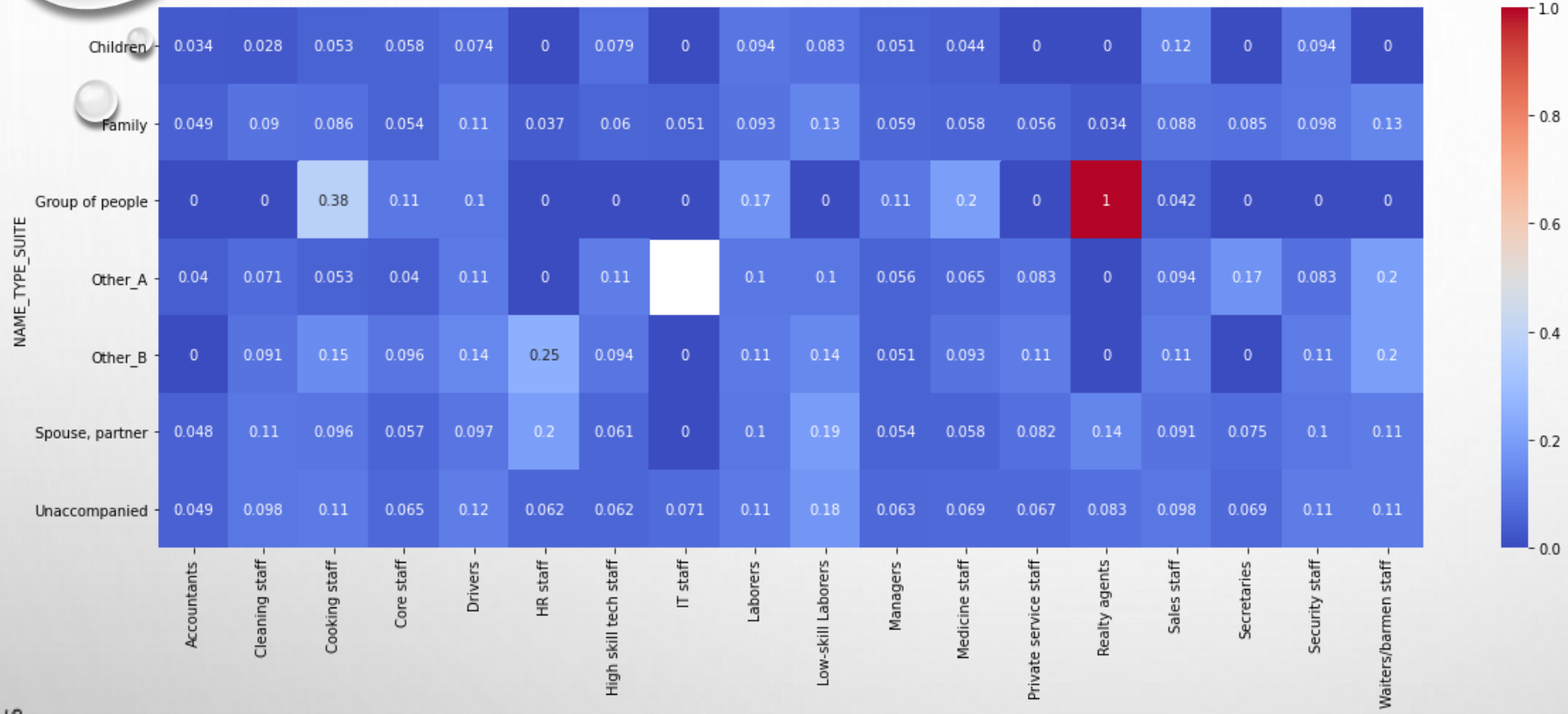


Clients who have high interest rate for their loans and having contract status 'Unused offer' are more likely to default.



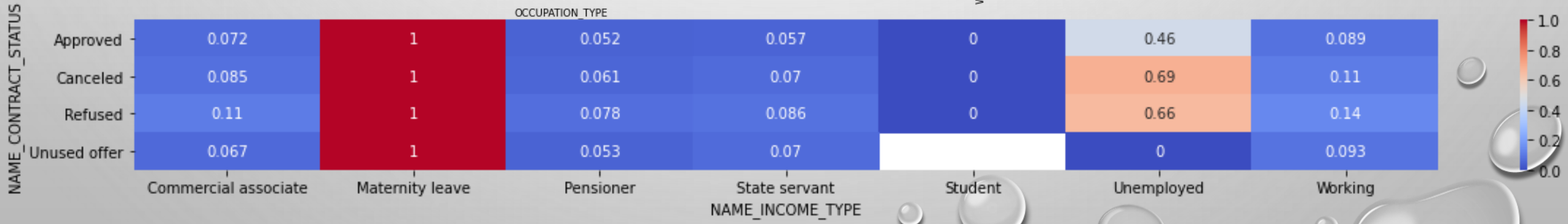
Clients who have an academic degree whose contract status is unused offer are more likely to default.

BIVARIATE CATEGORICAL VARIABLE ANALYSIS OF CONTRACT STATUS & OTHER VARIABLES BY CONSIDERING THE TARGET VALUE (Merged data)



1.Clients who are Realty agents and accompanied by a group of people while applying for the loan are most likely to default.

2.Clients who are on maternity leave are more likely to default irrespective of Contract status



CONCLUSION

Clients to avoid:

i. In general clients who are on maternity leave are more likely to default.
ii. In general clients who are unemployed are more likely to default.
iii. In general clients who have relatively more number of children are more likely to default.
iv. Unemployed Male clients who have opted for Cash loans/ having relatively more number of children are more likely to default.
v. Clients who have lower-secondary education/ living in a rented apartment/ with parents are more likely to default.
vi. Male Clients who are relatively younger / low-skilled laborer's / having more number of family members are more likely to default.

Clients to look for:

i. Clients who have opted for Revolving loans are less likely to default.
ii. Clients with contract status 'Approved' are less likely to default irrespective of other variables.
iii. Clients who are businessmen or students are less likely to default.
iv. Clients with an academic degree are less likely to default.
v. Clients living in an office apartment are less likely to default.