# CREDIT EDA CASE STUDY

*By – Devanshi Sinha*

# INTRODUCTION

## 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.

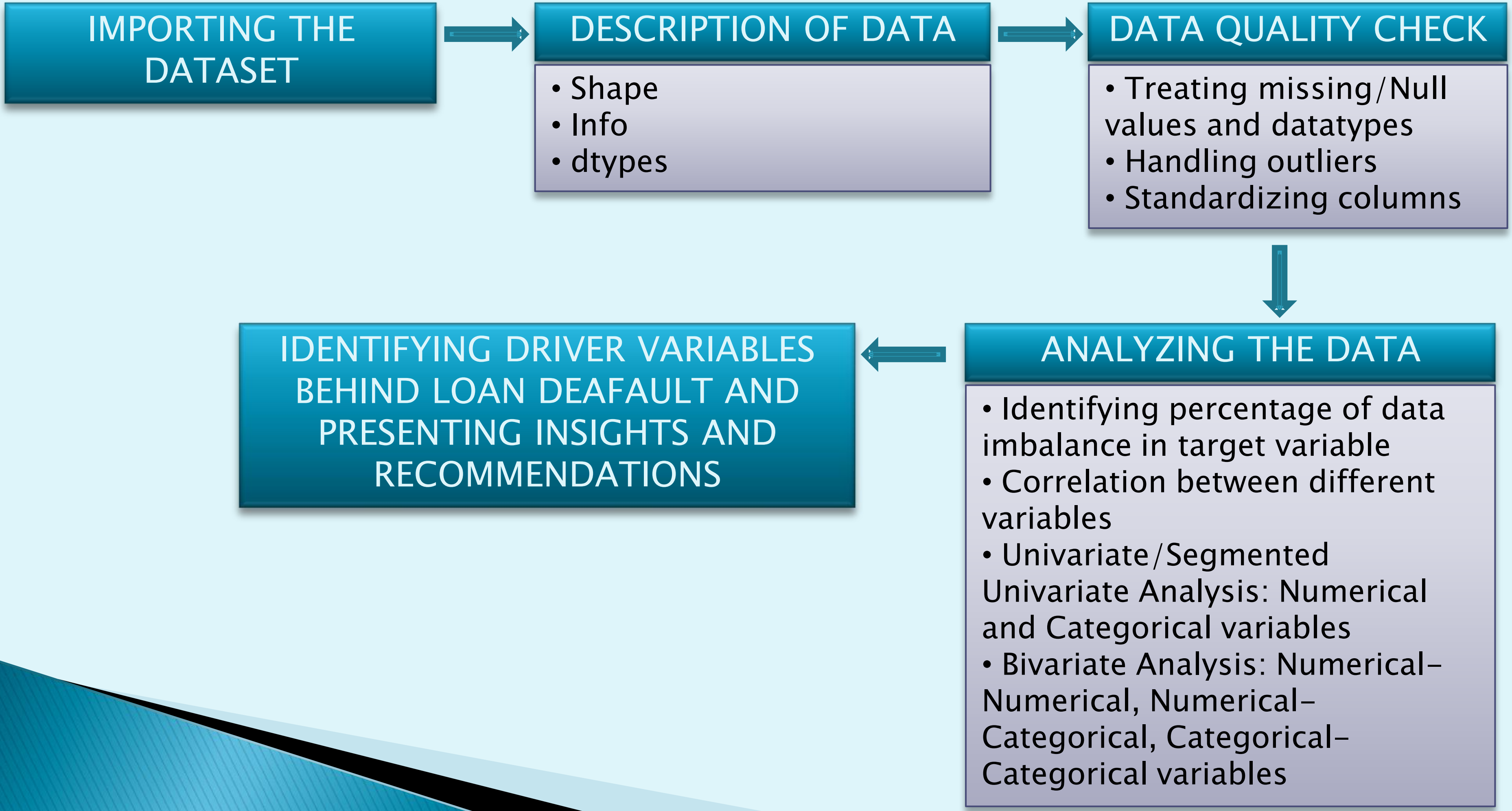
## Available Data:

- Read into the *app_data* dataframe*: 'application_data.csv'* contains all the information of the client at the time of application.
- Read into the *prev_app* dataframe*: 'previous_application.csv'* contains information about the client's previous loan data.
- *'columns_description.csv'* is data dictionary which describes the meaning of the variables.
- 'SK_ID_CURR' is a column name common in both the data frames, so we merge the data frame based on this client id.
- *merged_data* is the new dataframe created after merging the *prev_app* dataframe with *app_data*.

# OBJECTIVE

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.

# ANALYSIS APPROACH



IMPORTING THE DATASET

DESCRIPTION OF DATA

- Shape
- Info
- dtypes

DATA QUALITY CHECK

- Treating missing/Null values and datatypes
- Handling outliers
- Standardizing columns

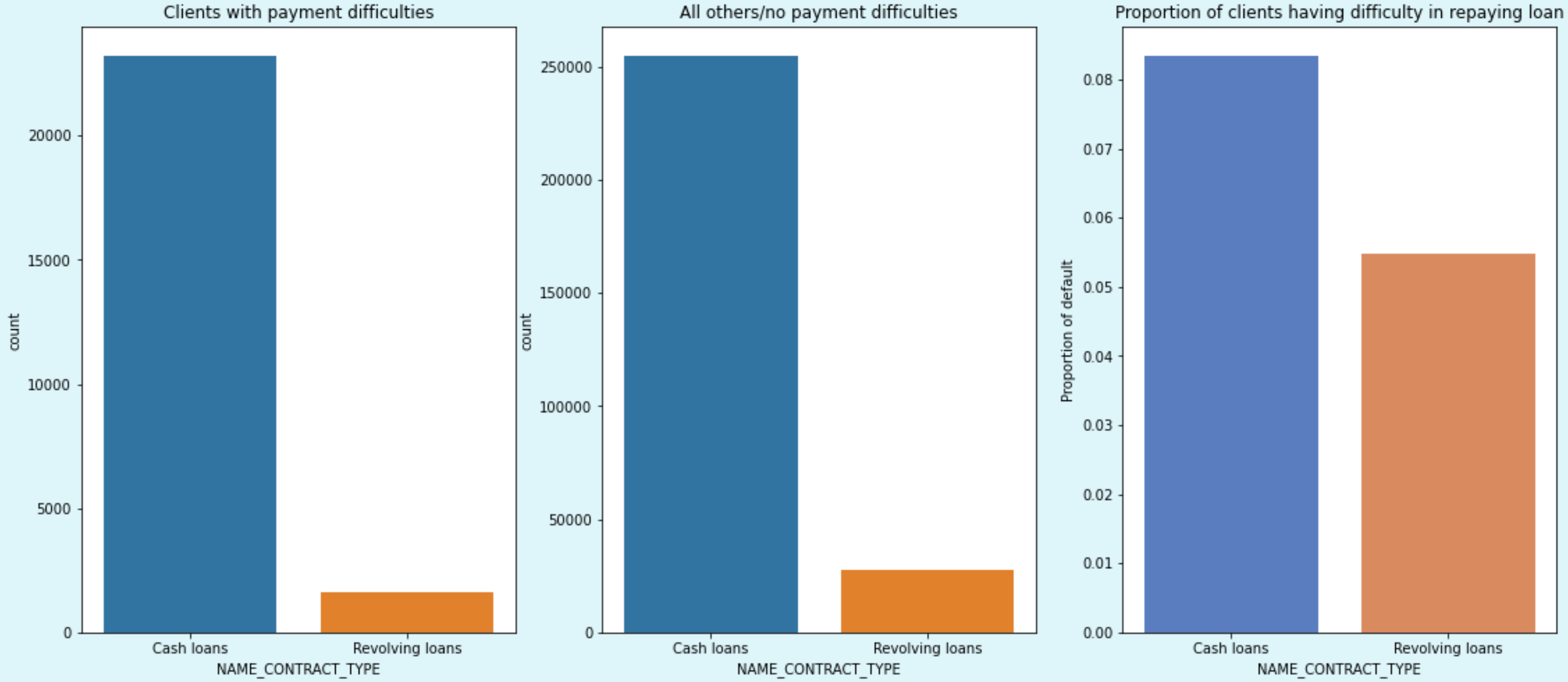
ANALYZING THE DATA

- Identifying percentage of data imbalance in target variable
- Correlation between different variables
- Univariate/Segmented Univariate Analysis: Numerical and Categorical variables
- Bivariate Analysis: Numerical-Numerical, Numerical-Categorical, Categorical-Categorical variables

IDENTIFYING DRIVER VARIABLES BEHIND LOAN DEAFAULT AND PRESENTING INSIGHTS AND RECOMMENDATIONS

# UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS

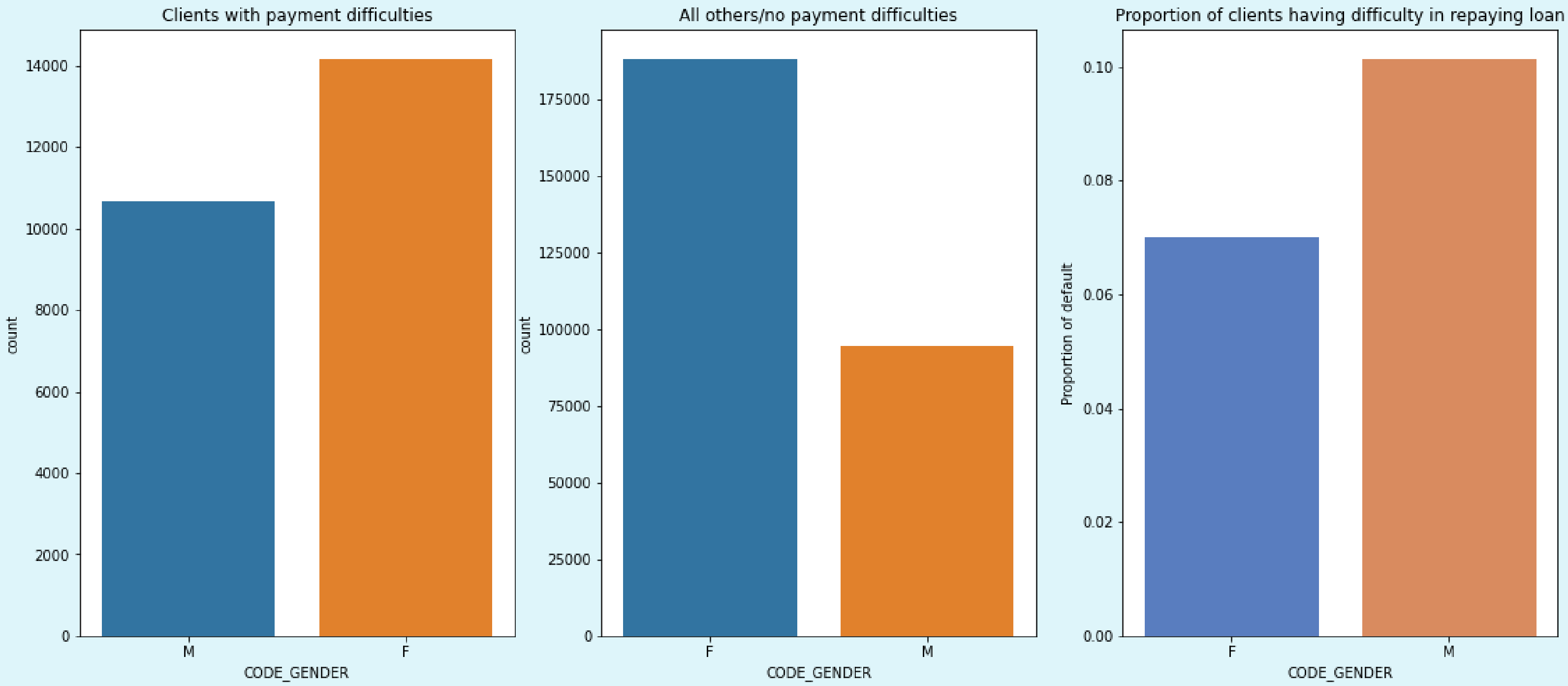
## CONTRACT TYPE



1. We can very clearly observe that the proportion of cash loans taken (about 91-95%) is higher than the proportion of revolving loans taken (about 5-8%) by both clients with payment difficulties and with no payment difficulties.
2. It can also be concluded from the above information that proportion of default of clients in cash loans (~ 8%) is higher than default in resolving loan (~5%).

# UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS

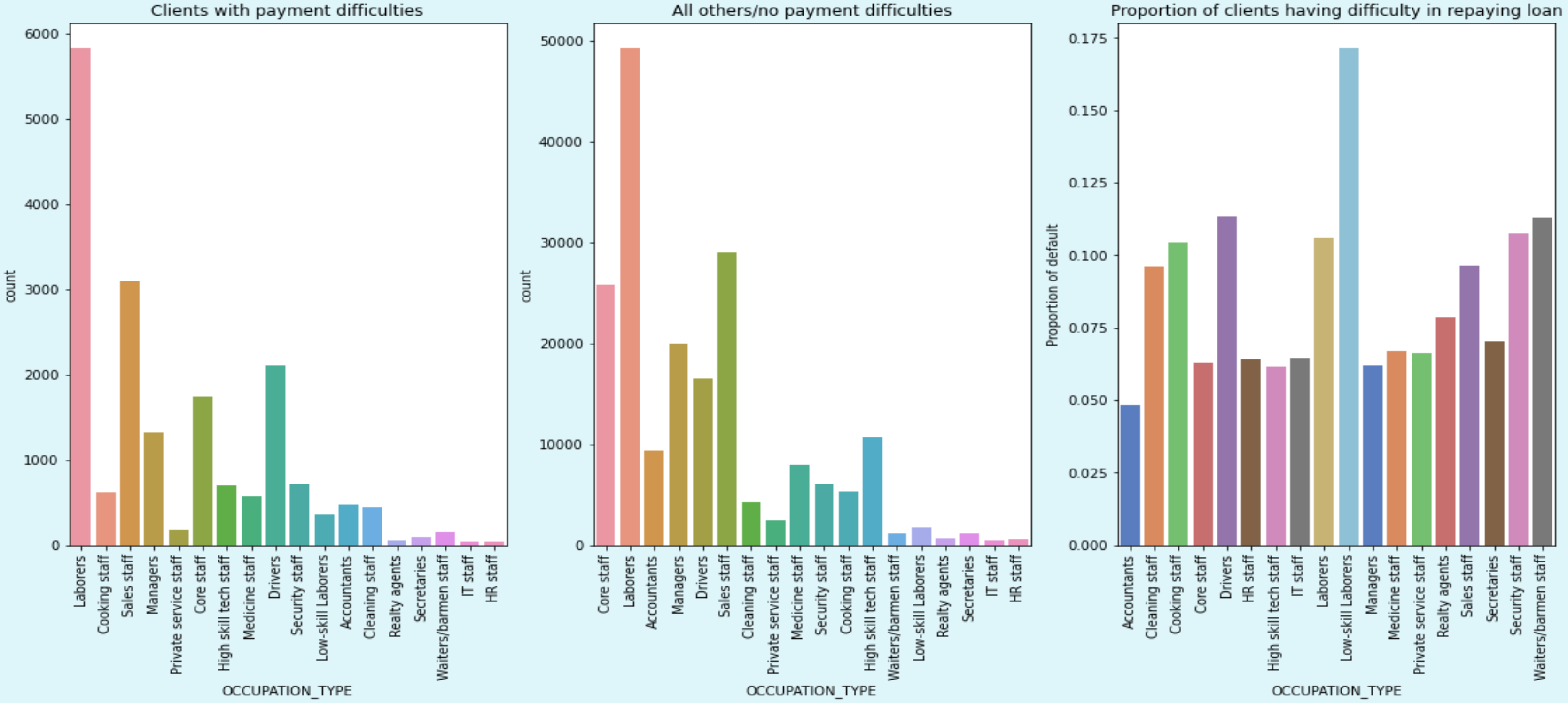
## GENDER



1. We can very clearly observe that the proportion of females clients (about 57-67%) is more than male clients (about 33-43%)in both clients with payment difficulties and with no payment difficulties.
2. We can also conclude from the above information that the proportion of male clients defaulting (~ 10%) is greater as compared to female clients defaulting (~ 7%).

# UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS

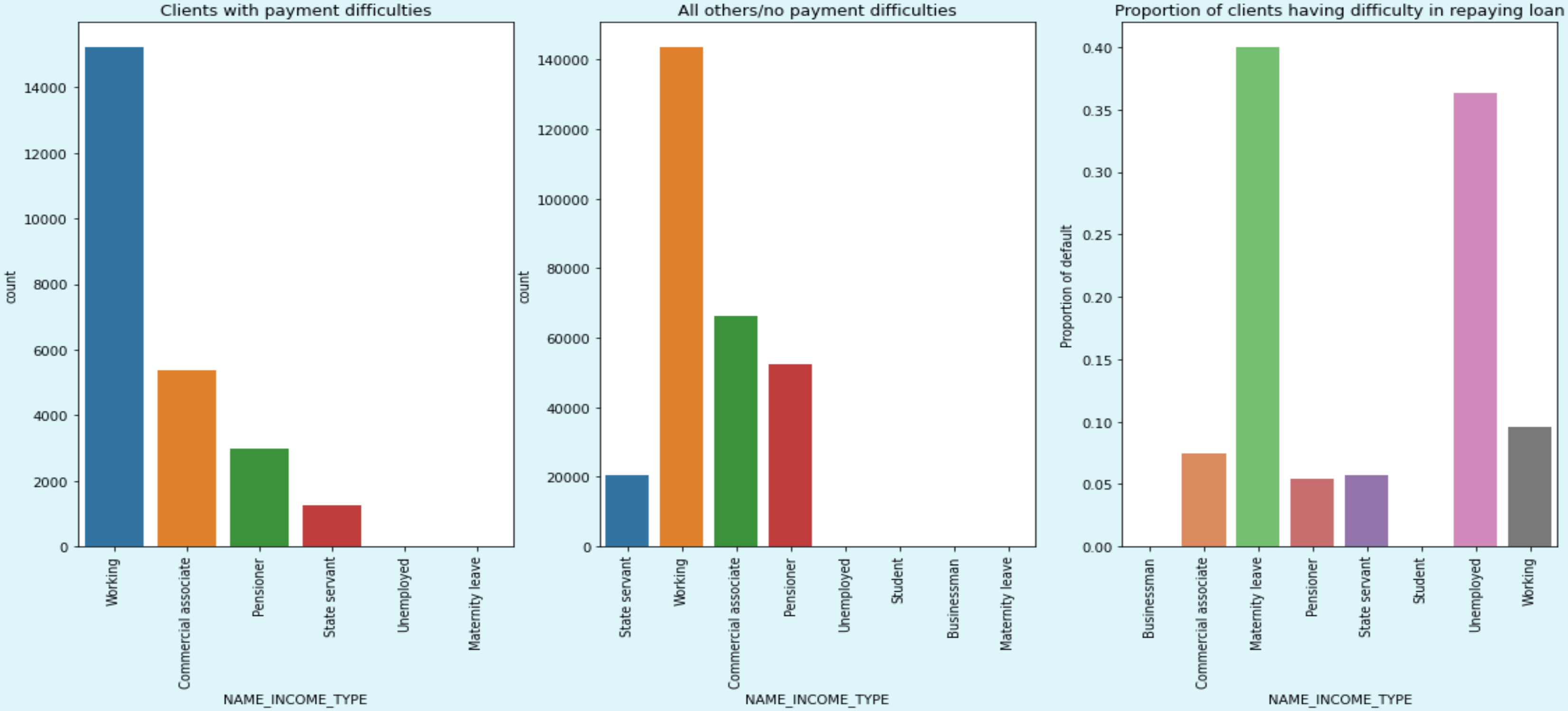
## OCCUPATION TYPE



1. We can very clearly observe that most clients belong to laborers category (about 25-30%) followed by Sales staff (about 15-17%) and least to the HR category and so on.
2. We can also conclude from the above information that Low-skill laborers have the highest default rate (~ 17%) while accountants have the lowest default rate (~ 5%).

# UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS

## INCOME TYPE


Clients with payment difficulties

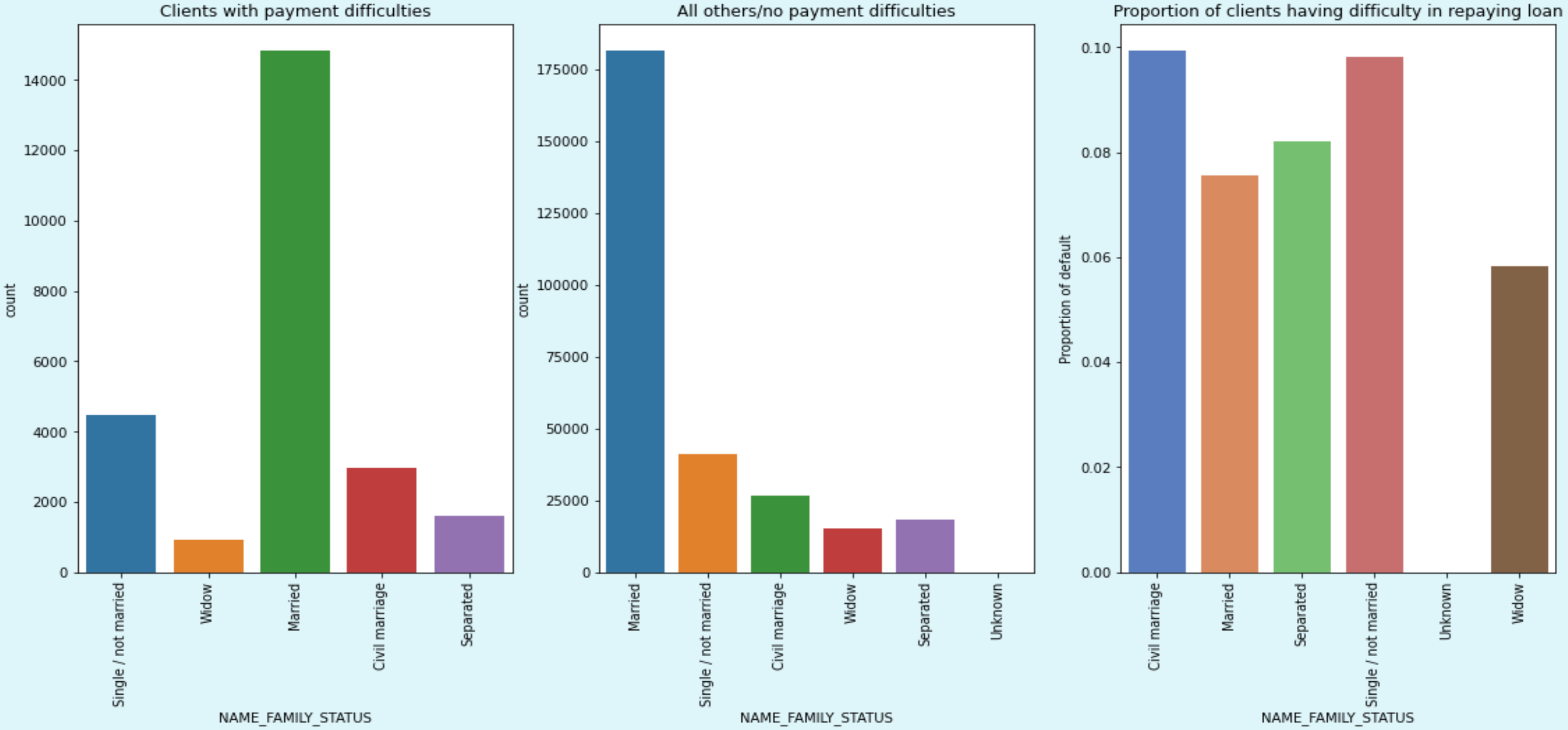
All others/no payment difficulties

Proportion of clients having difficulty in repaying loan

1. We can clearly observe that most clients are working, about 20-25% are commercial associates and least number of clients are on maternity leave and so on.
2. We can also conclude from the above information that clients who on maternity leave and are unemployed have the highest default rate while the rest of the income types have less than 10% chance of defaulting.

# UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS

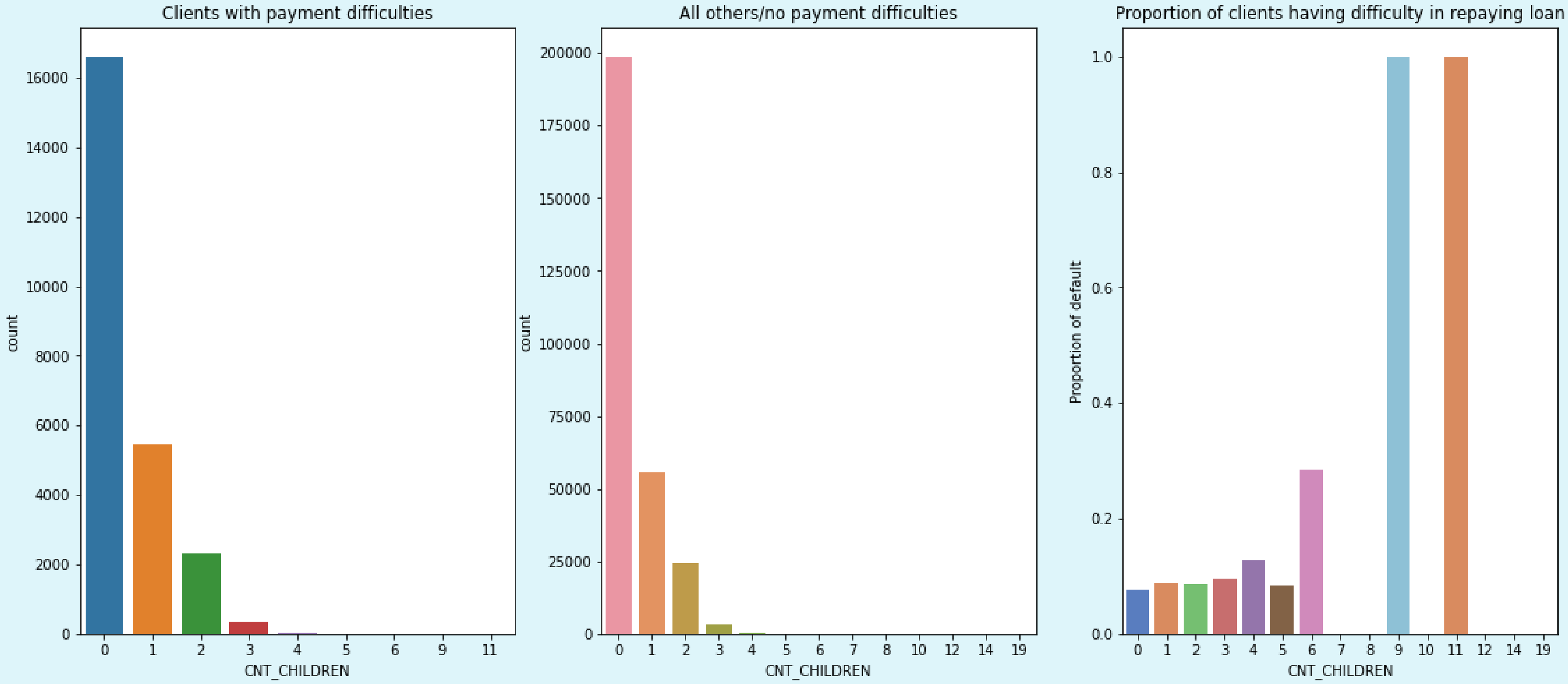
## FAMILY STATUS



1. We can very clearly observe that most clients are married (~60%), about 15% of the clients are not married, with the least number of clients being widows
2. We can also conclude from the above information that clients who have undergone civil marriage or clients who single are most likely to default (~ 10%) with widows having the least default rate (~ 6%).

# UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS

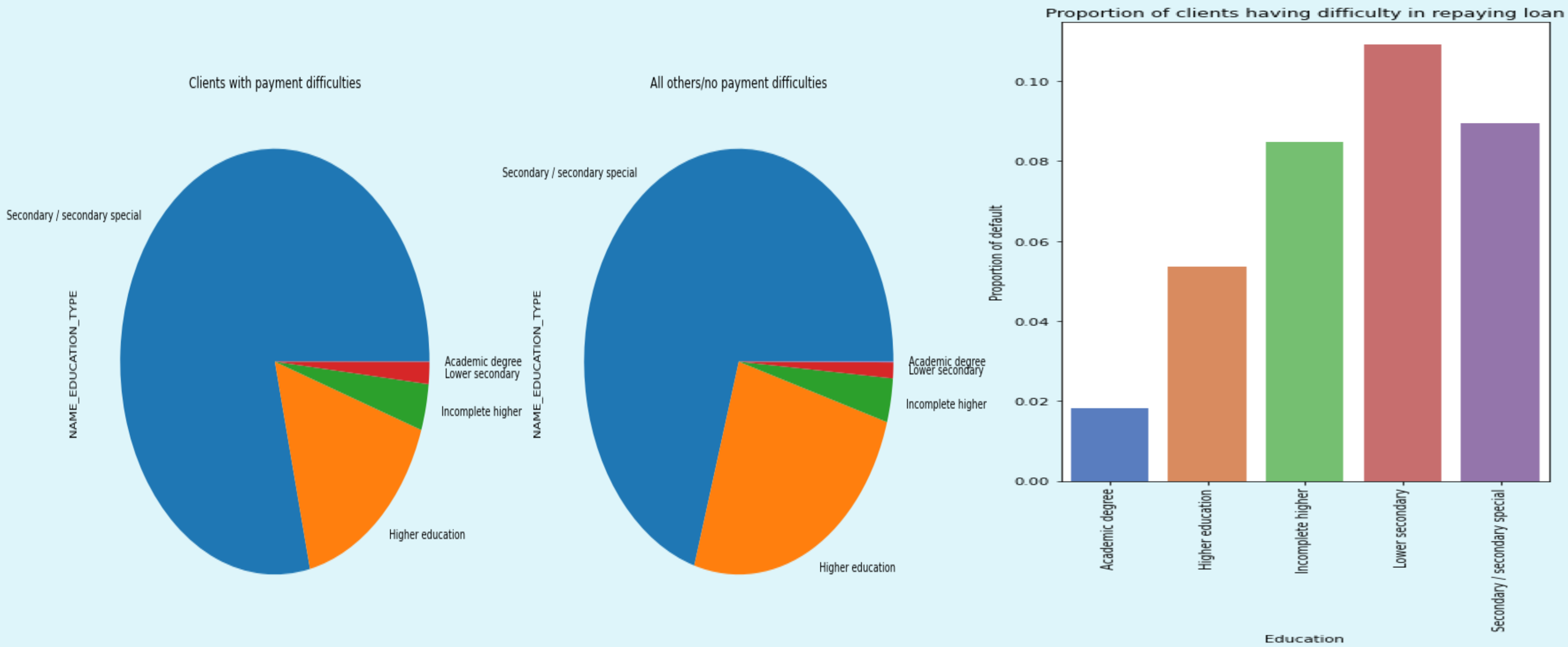
## CHILDREN COUNT



1. We can very clearly observe that most clients do not have children (~ 60%), followed by clients who has 1 child (~ 20%) and less than 10 % clients having 2 or more than children.
2. We can also conclude from the above information that clients who have 9 or 11 children have approximately 100% chance of defaulting. In general it seems like chances of defaulting increases with the number of children.

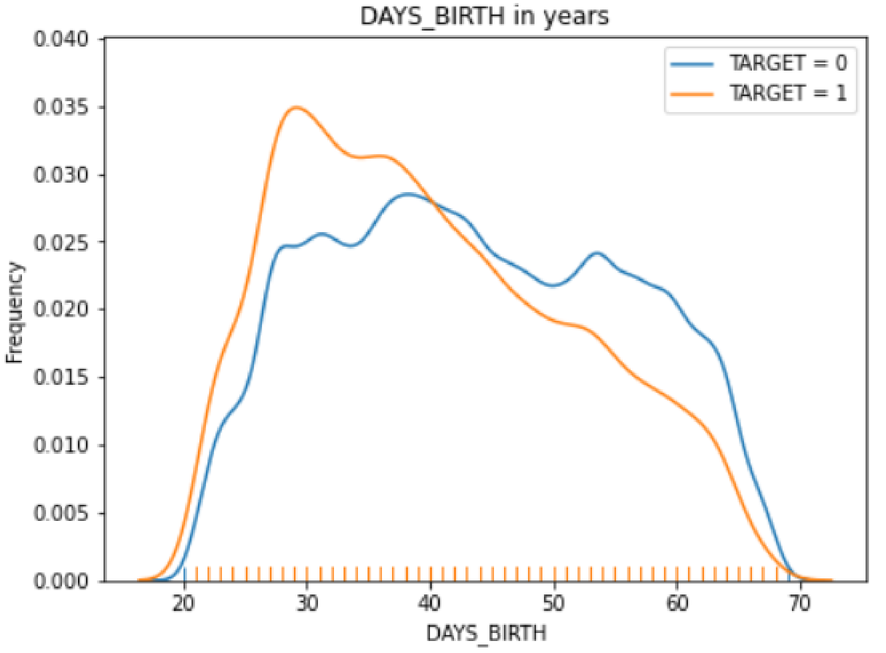
# UNIVARIATE/ SEGMENTED UNIVARIATE ANALYSIS

## EDUCATION TYPE



1. We can very clearly observe that most clients have completed secondary education (approx. 70-80%), followed by clients who have completed their higher education and so on.
2. We can also conclude from the above information that clients having lower secondary education have the highest default rate (~ 11%) followed by clients having secondary education while the clients having academic degrees are least likely to default (~2%).

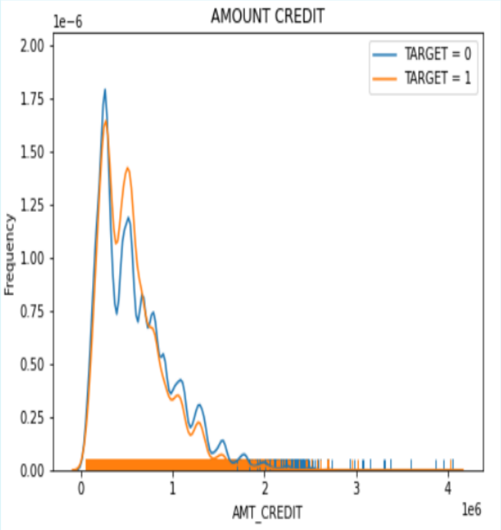
# ANALYSIS OF CONTINUOUS VARIABLES

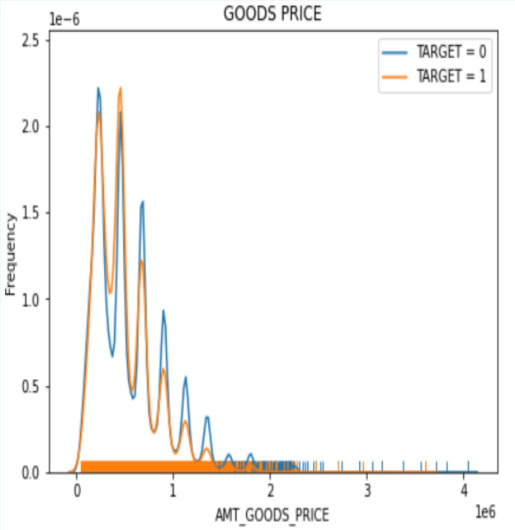


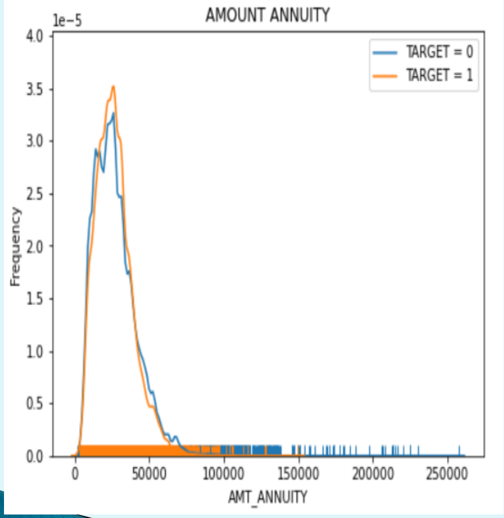
1. From the above graph comparing both the target variables (0 & 1) we observe clients of all ages, with the most number of clients being around 40 years of age.
2. We can also conclude from above that relatively younger clients have higher default rate with most of them lying around 30 years of age.

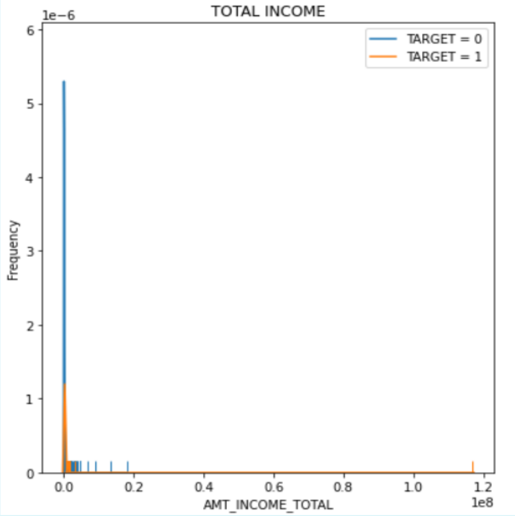
**Note-** The above plot for DAYS_BIRTH has been converted from days to years for better analysis.

# ANALYSIS OF CONTINUOUS VARIABLES





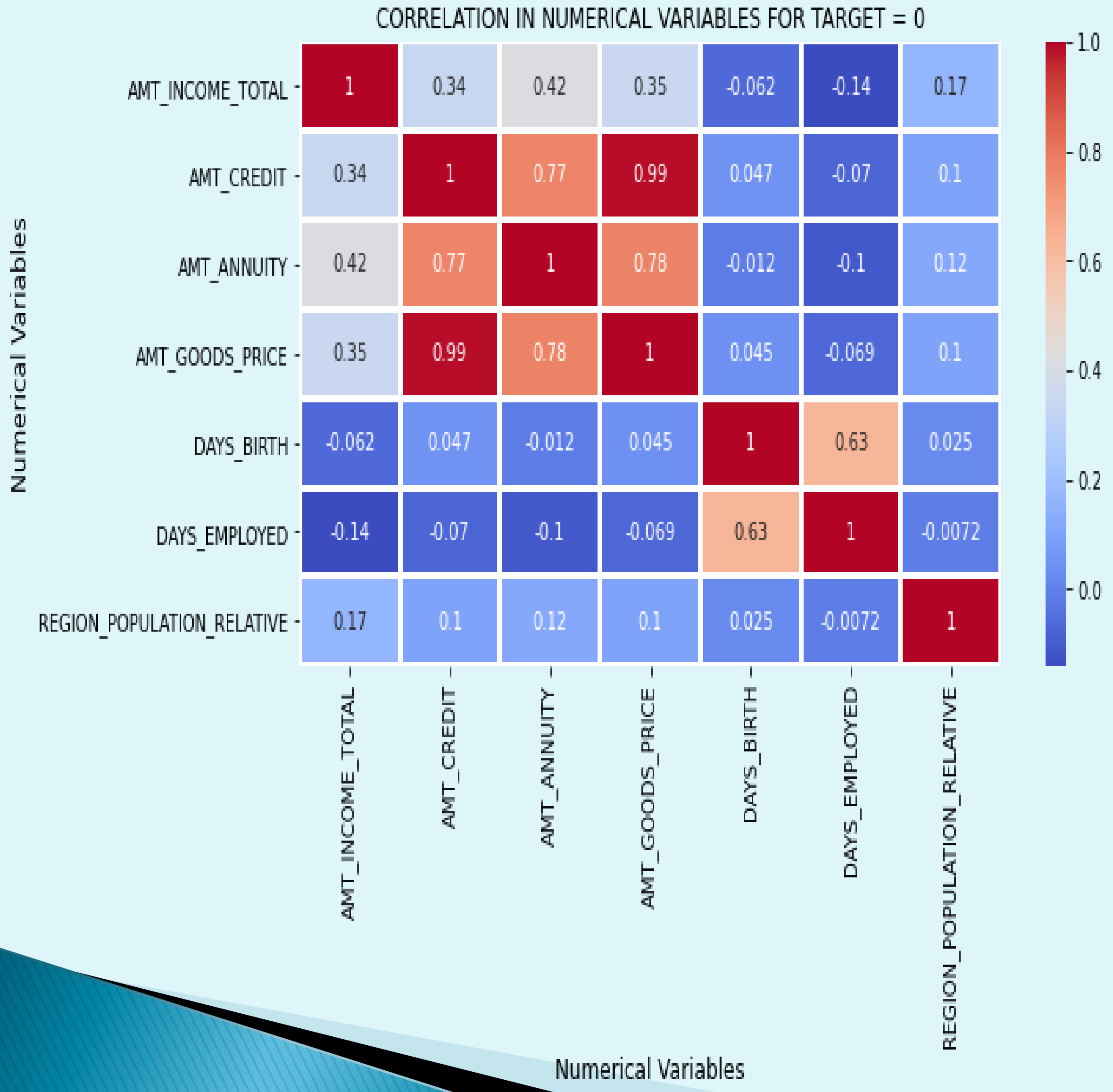




1. We can observe that credit amount of the loan lies mostly around 200000-1000000 for both target variables (0 & 1).
2. We can also conclude that the lesser loan credit amount, the higher the default rate.
3. We can observe that goods for which the clients have received loan are mostly concentrated in the range 200000-700000 for both target types.
4. Also, at this range we can see clients are more defaulted than at the higher range of the goods price.
5. We can clearly observe the annuity amount to be similar for both target types mostly concentrated around 40000.
6. We can observe that income of clients of both target types (0 & 1) lie mostly around 10000000 with the presence of some outliers as well.

# BIVARIATE ANALYSIS

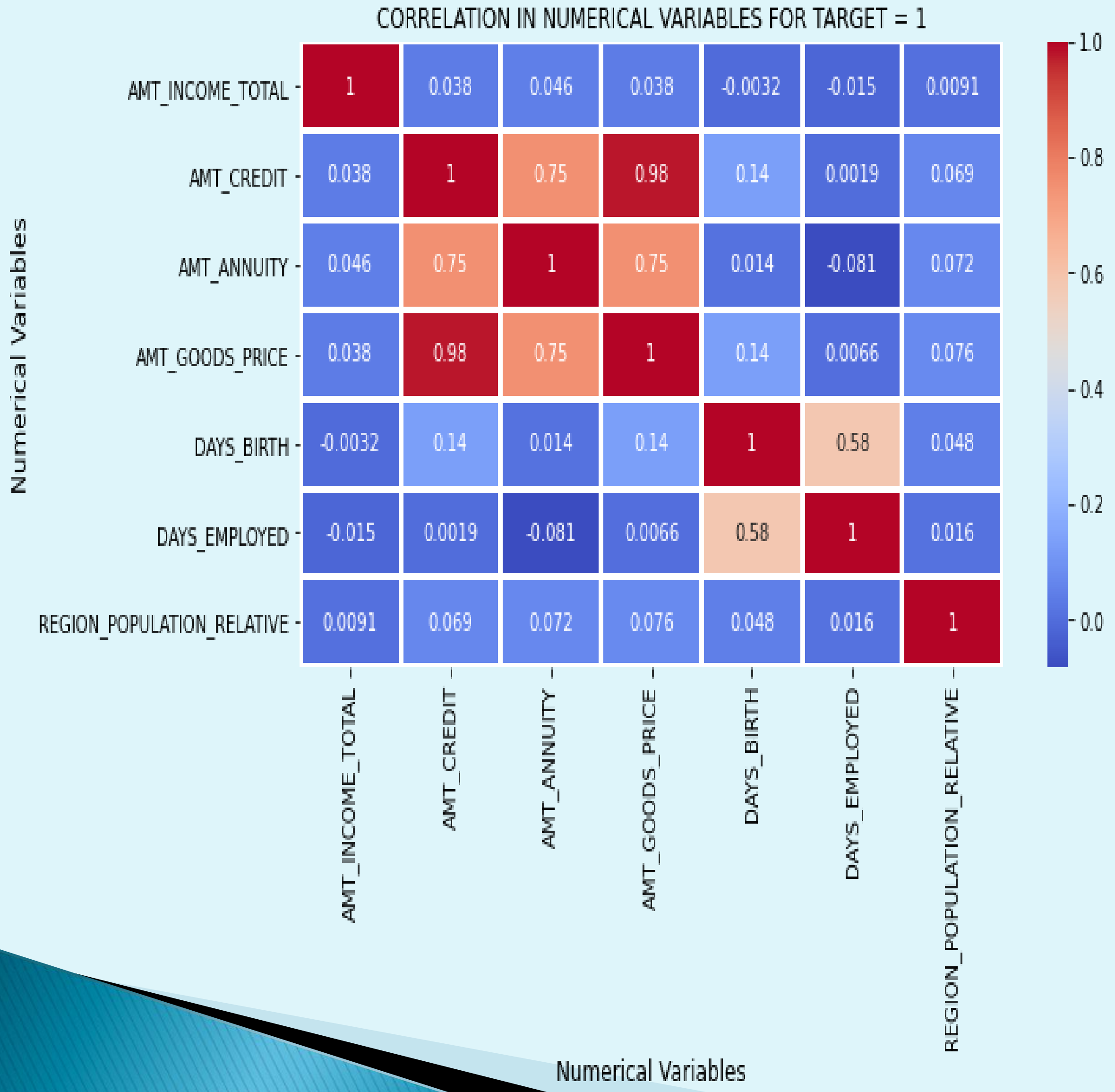
## Numerical- Numerical Analysis



1. Credit amount and Goods price amount has the highest correlation of 0.99, which is obvious since the client is opting for a loan which is equal to the price of those goods.
2. Annuity amount and Goods price amount also have a high correlation of about 0.78 which is pretty obvious since the annuity amount is decided by the price of the goods (also credit amount).
3. Therefore, as suggested in the above two points Credit amount and annuity amount also have a correlation of 0.77 as the annuity amount is decided based upon on the credit amount.
4. Income amount has a correlation of 0.42 with annuity amount suggesting that the annuity amount might be decided based upon the client's total income.
5. Income amount and credit amount have a correlation of 0.34 and goods price and income amount have a correlation of 0.35, suggesting why the clients are able to repay their loans.

# BIVARIATE ANALYSIS

## Numerical- Numerical Analysis

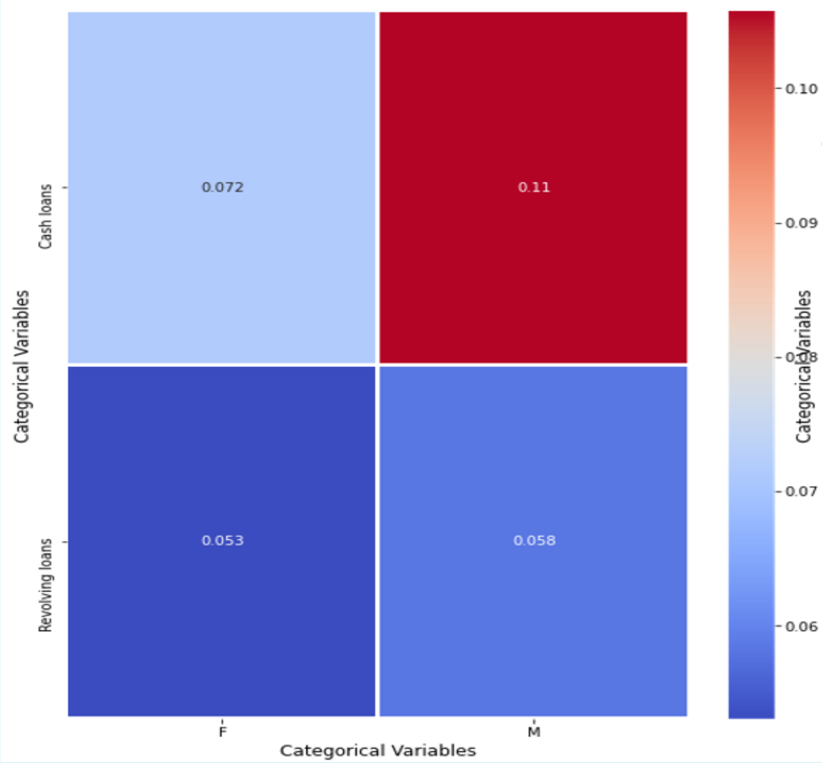


| | AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | DAYS_BIRTH | DAYS_EMPLOYED | REGION_POPULATION_RELATIVE |
|---|---|---|---|---|---|---|---|
| AMT_INCOME_TOTAL | 1 | 0.038 | 0.046 | 0.038 | -0.0032 | -0.015 | 0.0091 |
| AMT_CREDIT | 0.038 | 1 | 0.75 | 0.98 | 0.14 | 0.0019 | 0.069 |
| AMT_ANNUITY | 0.046 | 0.75 | 1 | 0.75 | 0.014 | -0.081 | 0.072 |
| AMT_GOODS_PRICE | 0.038 | 0.98 | 0.75 | 1 | 0.14 | 0.0066 | 0.076 |
| DAYS_BIRTH | -0.0032 | 0.14 | 0.014 | 0.14 | 1 | 0.58 | 0.048 |
| DAYS_EMPLOYED | -0.015 | 0.0019 | -0.081 | 0.0066 | 0.58 | 1 | 0.016 |
| REGION_POPULATION_RELATIVE | 0.0091 | 0.069 | 0.072 | 0.076 | 0.048 | 0.016 | 1 |

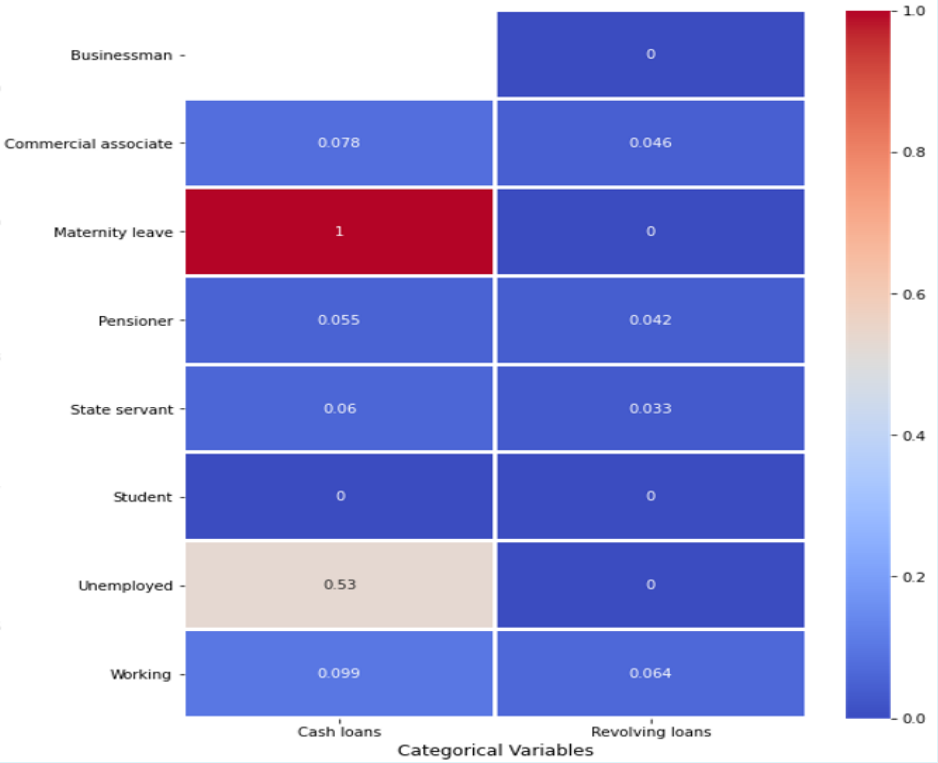
1. Credit amount and Goods price amount has the highest correlation of 0.98, which is obvious since the client is opting for a loan which is equal to the price of those goods.
2. Annuity amount and Goods price amount also have a high correlation of about 0.75 which is pretty obvious since the annuity amount is decided by the price of the goods (also credit amount).
3. Therefore, as suggested in the above two points Credit amount and annuity amount also have a correlation of 0.75 as the annuity amount is decided based upon on the credit amount.
4. Income amount and credit amount have a small correlation of 0.038 and annuity amount and income amount have a correlation of 0.046, suggesting why the clients are facing difficulty while repaying their loans.

# BIVARIATE ANALYSIS

## Categorical- Categorical Analysis



*CONTRACT TYPE VS GENDER*

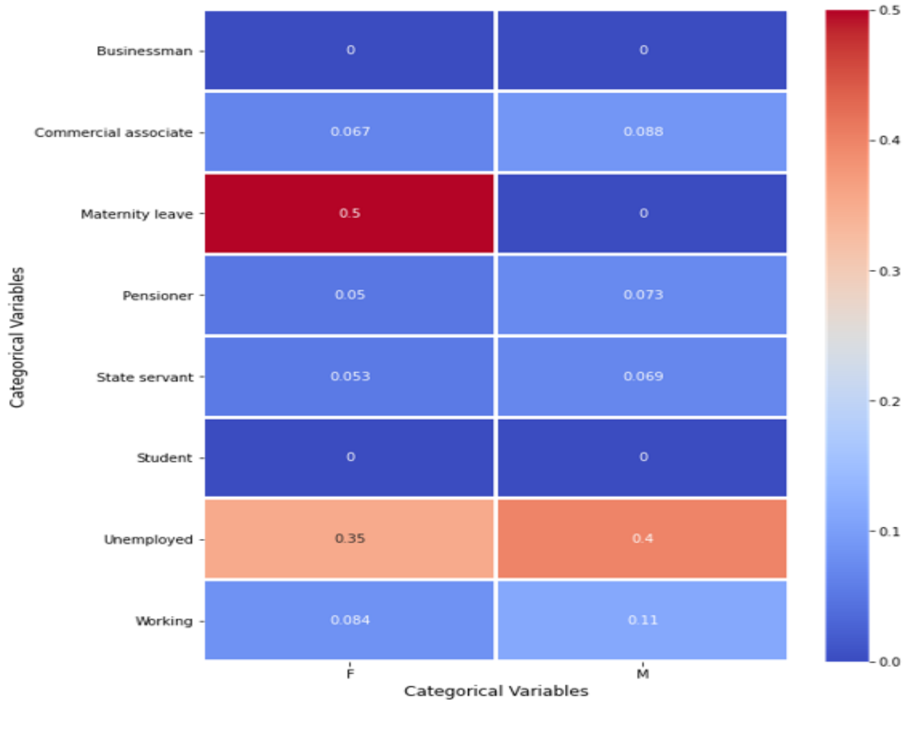


*INCOME TYPE VS CONTRACT TYPE*

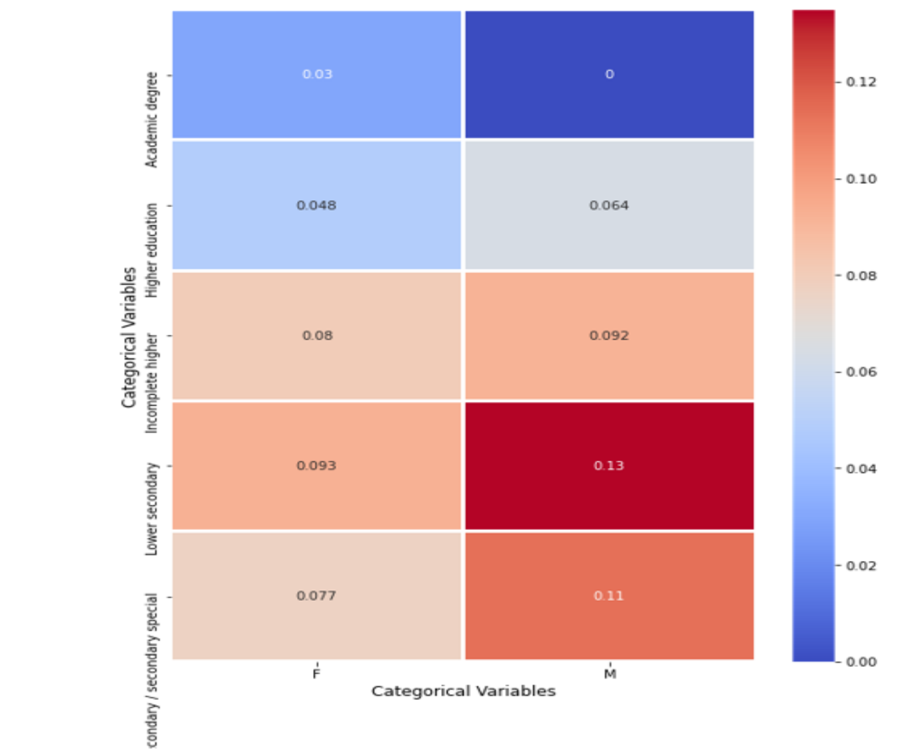
1. It is clearly evident that male clients who have opted for cash loans are most likely to default.
2. Clients on maternity leave are most likely to default.
3. Unemployed clients who have opted for cash loans are more likely to default.

# BIVARIATE ANALYSIS

## Categorical- Categorical Analysis



*INCOME TYPE VS GENDER*

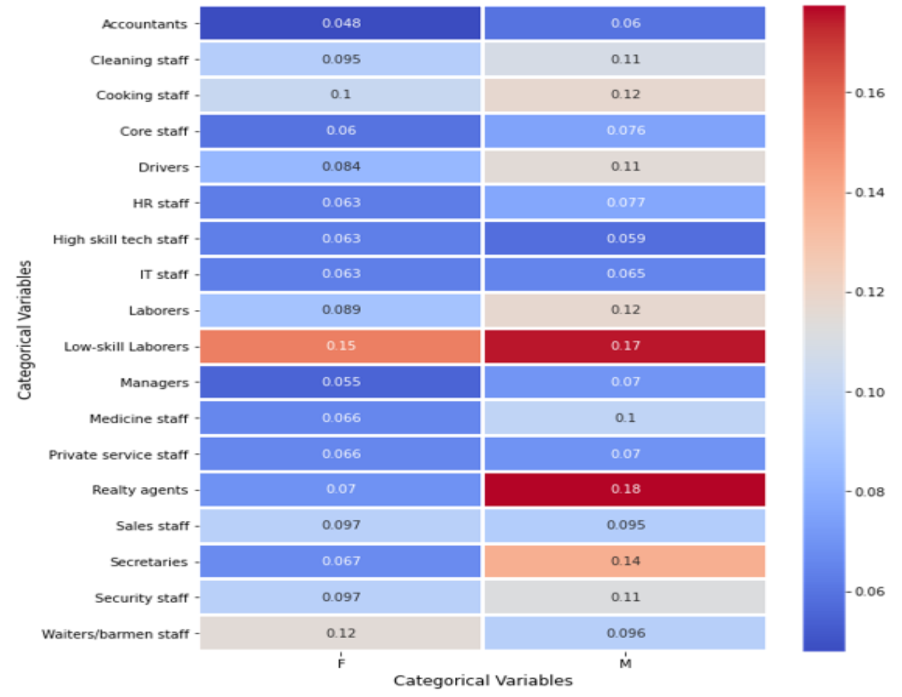


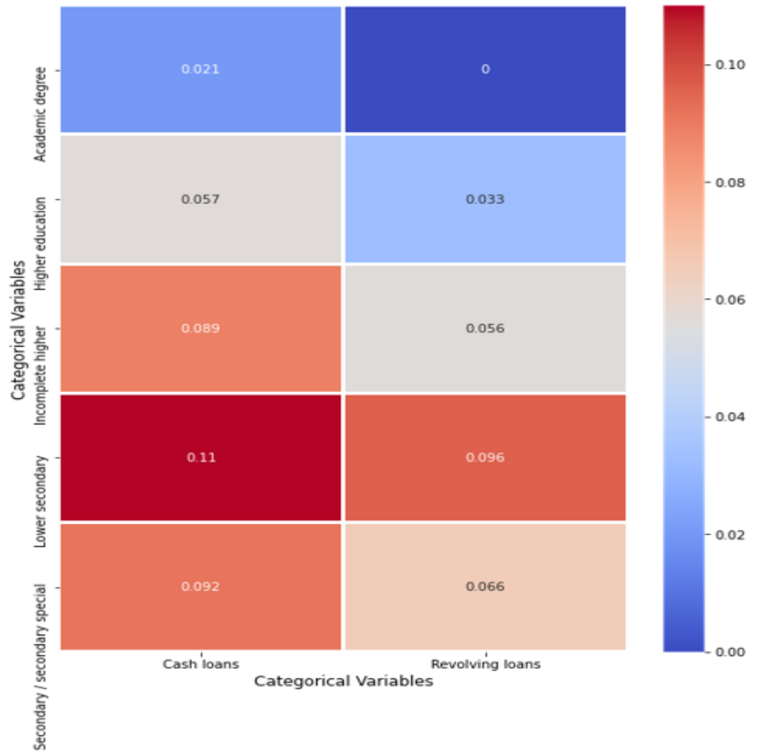
*EDUCATION TYPE VS GENDER*

1. Female clients on maternity leave are most likely to default.
2. Unemployed clients have a high default rate irrespective of gender.
3. Male clients are more likely to default irrespective of the education type, with male client having lower secondary education having highest default rate.

# BIVARIATE ANALYSIS

## Categorical- Categorical Analysis





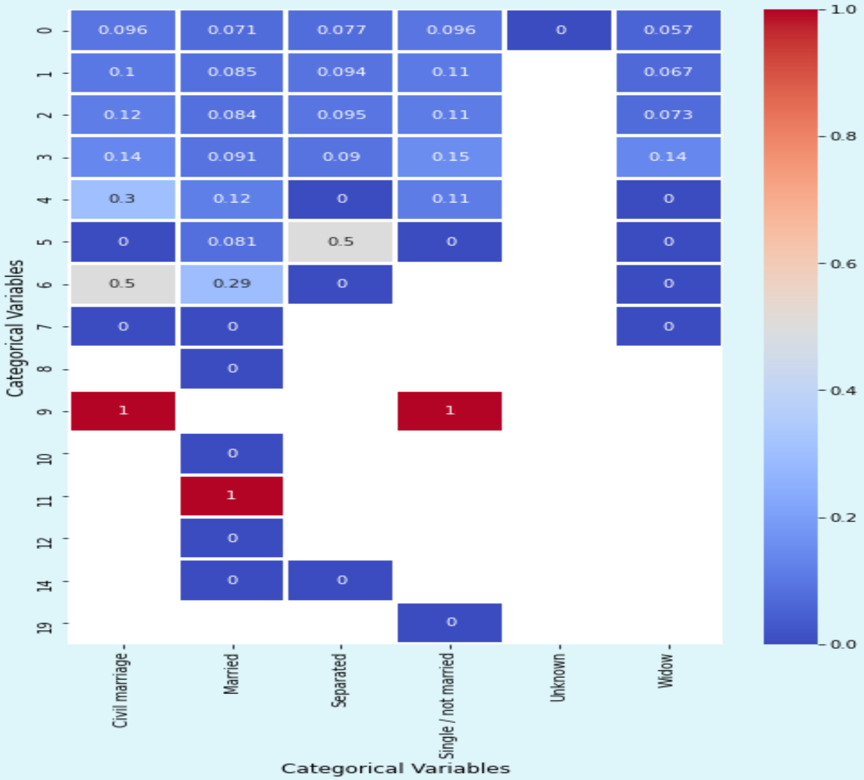
*OCCUPATION TYPE VS GENDER*

*EDUCATIONT YPE VS CONTRACT TYPE*

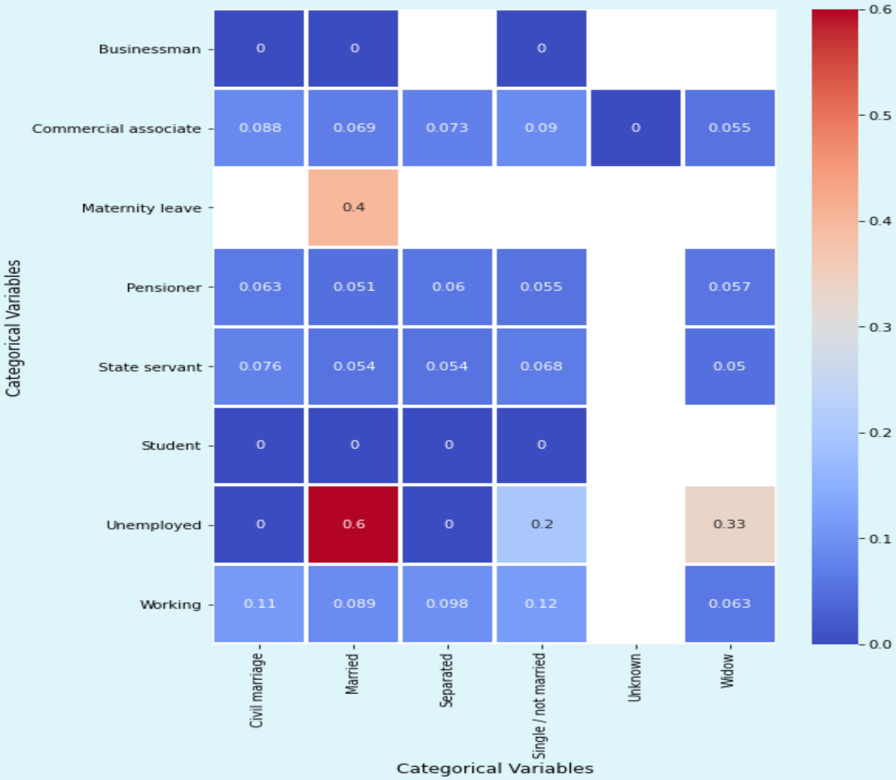
1. Male clients who are realty agents are more likely to default as compared to clients in other job types.
2. Low skill laborers have comparably higher defaulting chance and accountants have comparably lower defaulting chance  irrespective of gender.
3. Clients with lower secondary education are more likely to default irrespective of loan type.
4. Clients who opted for cash loans are most likely to default irrespective of their education level.

# BIVARIATE ANALYSIS

## Categorical- Categorical Analysis



*CHILDREN COUNT VS FAMILY STATUS*

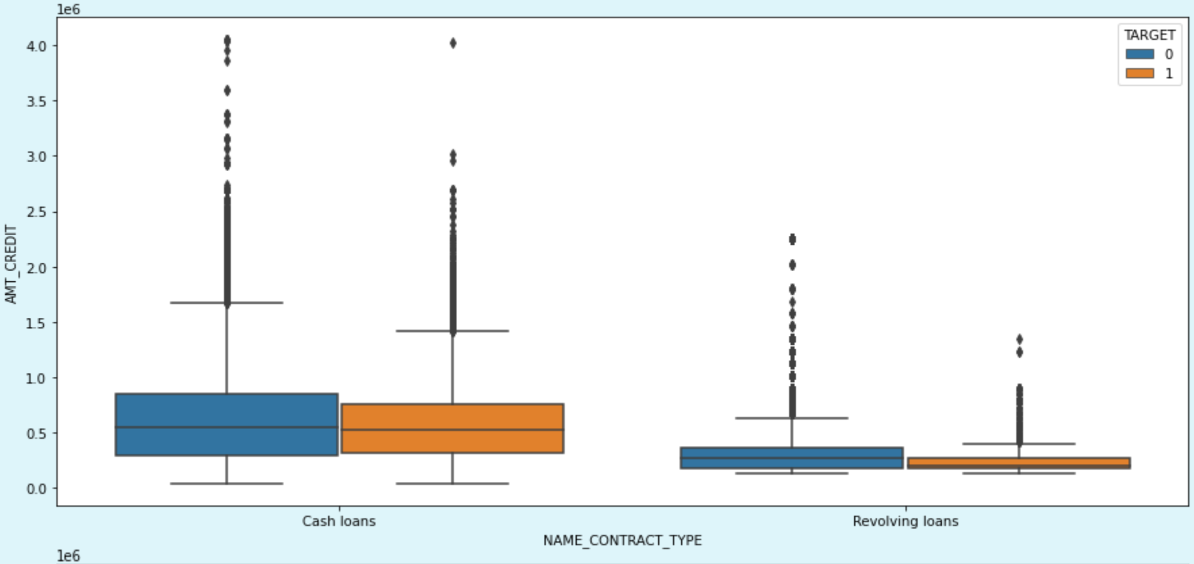


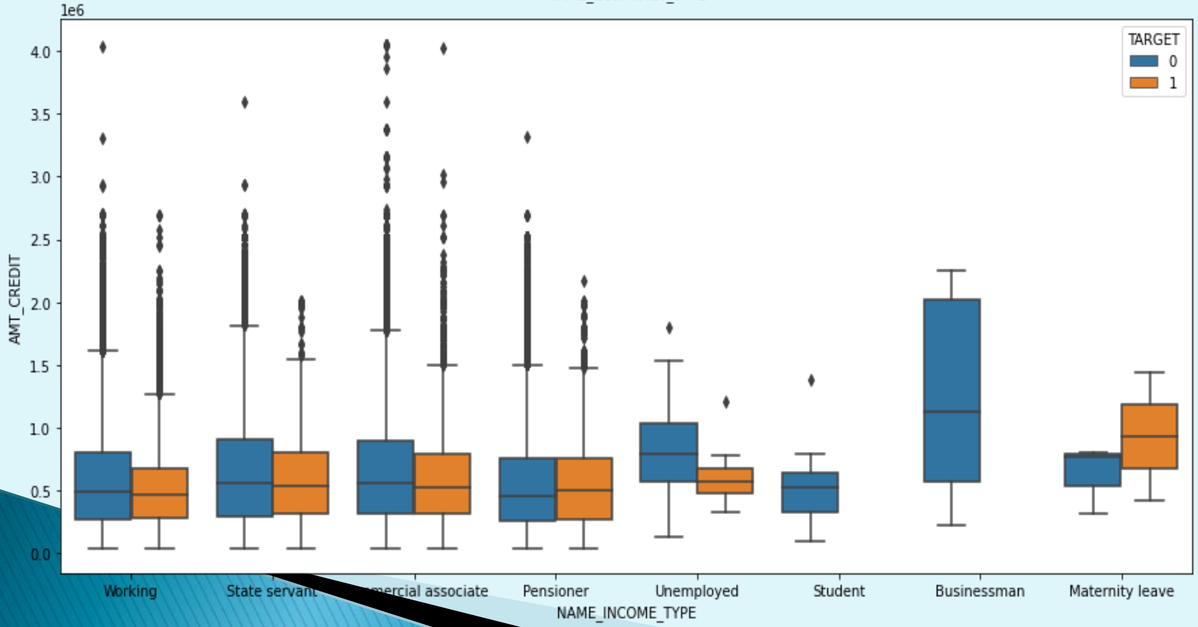
*INCOME TYPE VS FAMILY STATUS*

1. Clients having 9 or 11 children are most likely to default.
2. Clients who have 5 children and are separated have a high default rate.
3. Clients who have undergone civil marriage and have 6 children are also most likely to default.
4. Clients who are unemployed and married are most likely to default.
5. Clients who are unemployed and are widow also have a high default rate.
6. Married clients who are on maternity leave also a high default rate.

# BIVARIATE ANALYSIS

## Numerical- Categorical Analysis

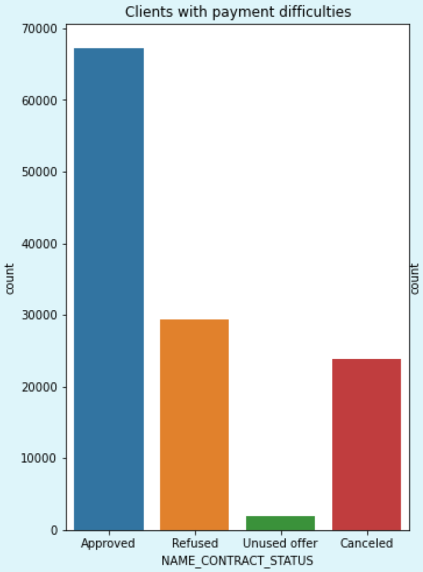


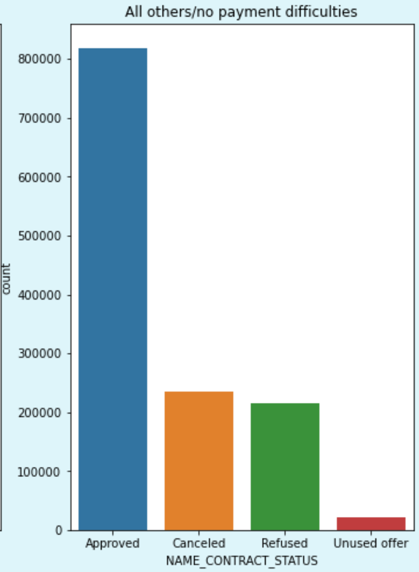


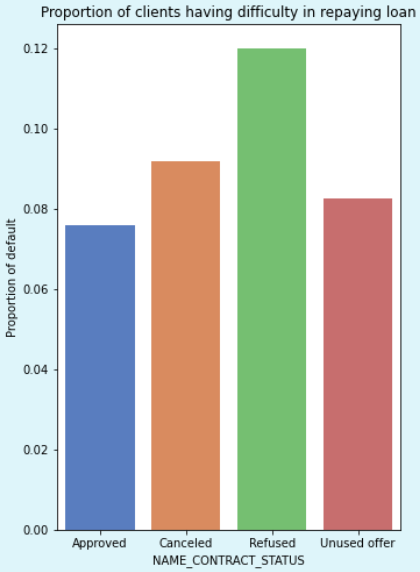
1. From the boxplot we can see that the credit amount for cash loans is higher than for revolving loans.
2. We observe that the credit amount for Target =1 clients is concentrated in the lower end of the box plot, which indicates the difficulty in loan payment for lesser credit amounts.
3. We observe that there are no boxplots for Businessman and Student category in target =1 which indicates they do not have difficulties in repaying loan (no defaulters).
4. The box plot for Businessman has a big IQR indicating a huge range of credit amount for the client.

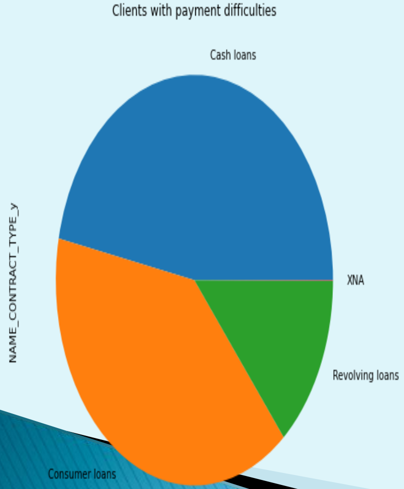
# UNIVARIATE ANALYSIS/ SEGMENTED UNIVARIATE ANALYSIS

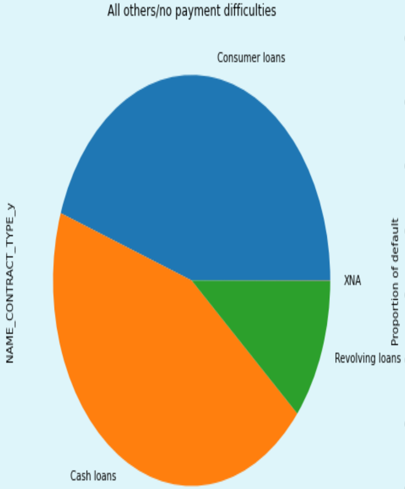
(for merged data)

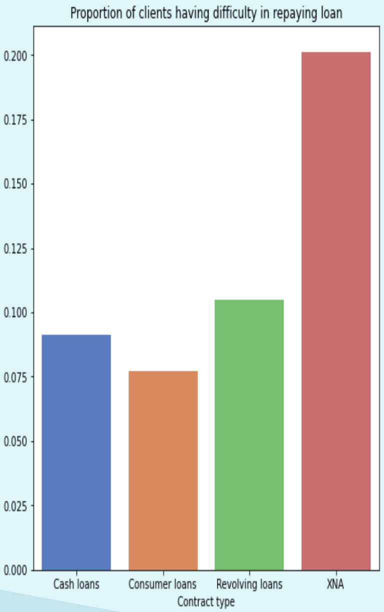








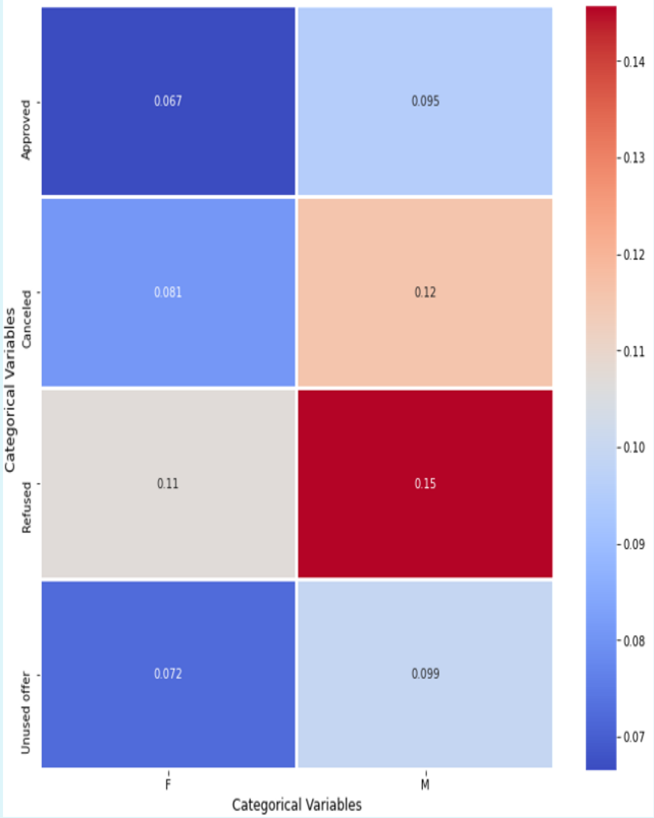




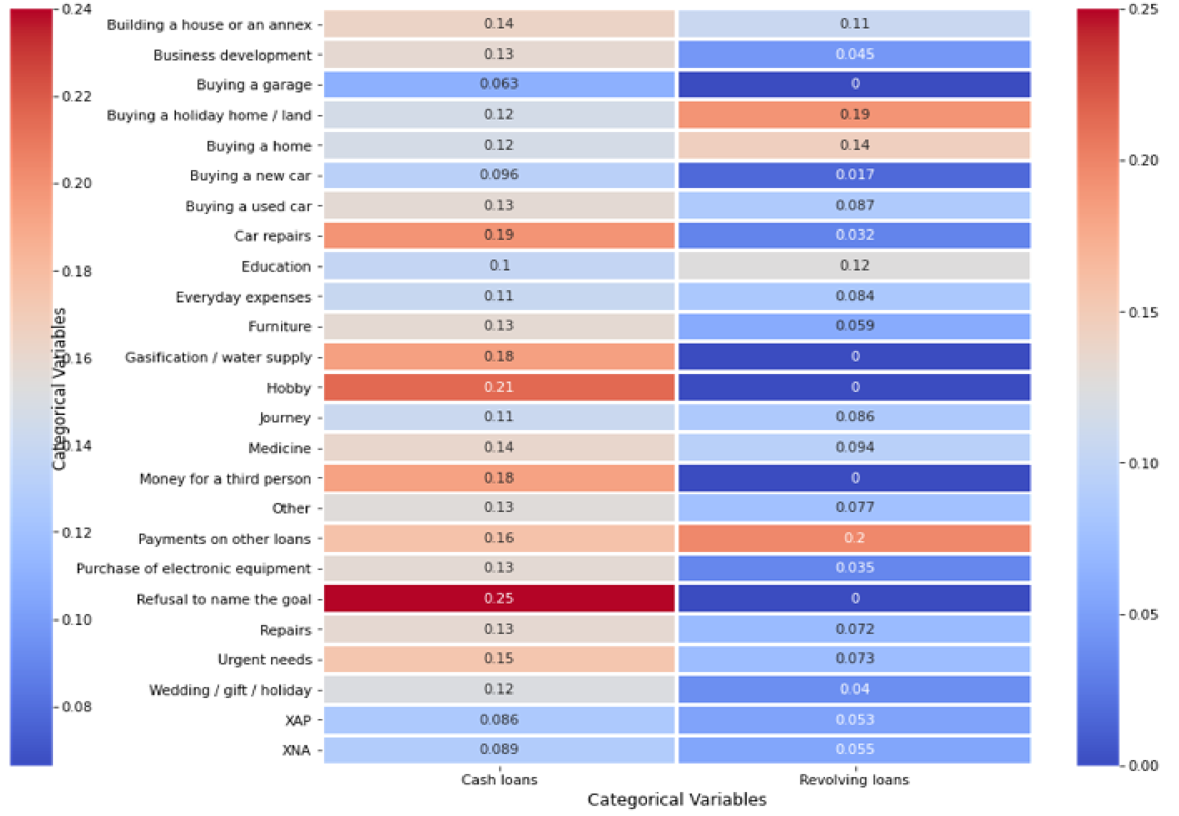
1. We can observe that majority of the clients' loans got approved (about 55-65%), while there were least number of clients with unused loans.
2. Clients who got their applications rejected are most likely to default, followed by clients who cancelled their loans while the clients whose applications got approved are least likely to default.
3. We can observe from the above previous application dataset that most of the clients have opted for cash and consumer loans (about 40%-45%) while about 10% have opted for revolving loans.
4. Clients who have opted for revolving loans have the highest default rate as compared to clients who opted for cash loans and consumer loans.

# BIVARIATE ANALYSIS

(for merged data)



CONTRACT STATUS VS GENDER

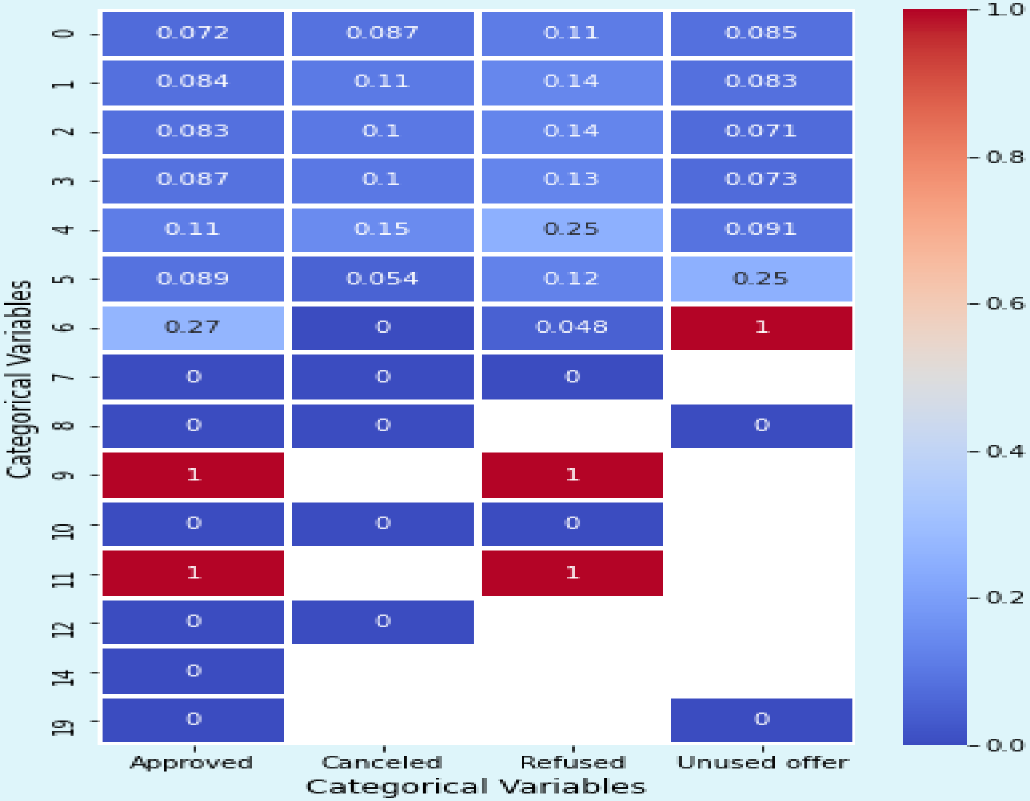


CASH LOAN PURPOSE VS CONTRACT TYPE

1. Male clients whose of contract status is refused are more likely to default.
2. Clients who opted for cash loans and refused to share their goal for loans have highest default rate, and clients who opted for cash loans for car repairs, gasification/ water supply, money for a third person are more likely to default.
3. Clients who opted for payments on other loans and buying a holiday home/land are more likely to default on revolving loans.

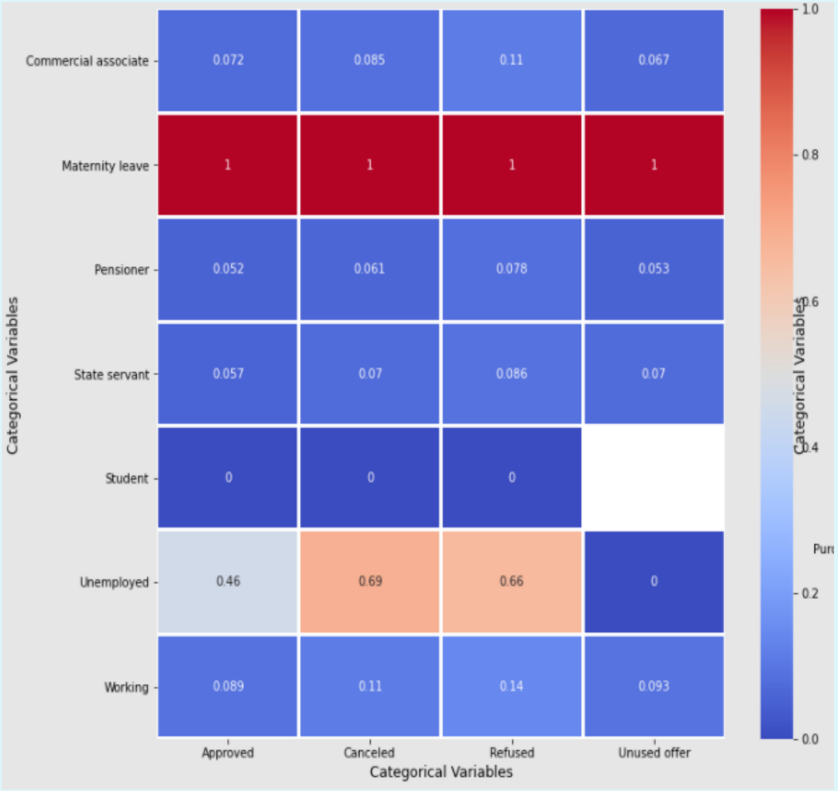
# BIVARIATE ANALYSIS

(for merged data)



| | Approved | Canceled | Refused | Unused offer |
|---|---|---|---|---|
| 0 | 0.072 | 0.087 | 0.11 | 0.085 |
| 1 | 0.084 | 0.11 | 0.14 | 0.083 |
| 2 | 0.083 | 0.1 | 0.14 | 0.071 |
| 3 | 0.087 | 0.1 | 0.13 | 0.073 |
| 4 | 0.11 | 0.15 | 0.25 | 0.091 |
| 5 | 0.089 | 0.054 | 0.12 | 0.25 |
| 6 | 0.27 | 0 | 0.048 | 1 |
| 7 | 0 | 0 | 0 | |
| 8 | 0 | 0 | | 0 |
| 9 | 1 | | 1 | |
| 10 | 0 | 0 | 0 | |
| 11 | 1 | | 1 | |
| 12 | 0 | 0 | | |
| 14 | 0 | | | |
| 19 | 0 | | | 0 |

CHILDREN COUNT VS CONTRACT STATUS



| | Approved | Canceled | Refused | Unused offer |
|---|---|---|---|---|
| Commercial associate | 0.072 | 0.085 | 0.11 | 0.067 |
| Maternity leave | 1 | 1 | 1 | 1 |
| Pensioner | 0.052 | 0.061 | 0.078 | 0.053 |
| State servant | 0.057 | 0.07 | 0.086 | 0.07 |
| Student | 0 | 0 | 0 | |
| Unemployed | 0.46 | 0.69 | 0.66 | 0 |
| Working | 0.089 | 0.11 | 0.14 | 0.093 |

INCOME TYPE VS CONTRACT STATUS

1. Clients having 6 children and unused offer are most likely to default.
2. Clients with 9 or 11 children with either approved or refused loan offer have the highest default rate.
3. In general, clients having more number of children are more likely to default irrespective of contract status.
4. Clients who are on maternity leave are more likely to default irrespective of Contract status

# CONCLUSION

**Insights and recommendations:**

➢ ***Clients to avoid (most likely to default):***

1. In general clients who are on maternity leave are most likely to default.
2. In general clients who are unemployed are most likely to default.
3. Male clients who are unemployed/ have opted for Cash loans/ have contract status as refused are more likely to default.
4. Clients who have lower-secondary education/ have undergone civil marriage/ are single are more likely to default.
5. In general clients who have relatively more number of children are more likely to default.
6. Male Clients who are realty agents/ low-skilled laborers are more likely to default.

➢ ***Clients to seek (not likely to default):***

1. Clients who have opted for Revolving loans are less likely to default.
2. In general accountants are less likely to default.
3. Clients who are businessmen or students are least likely to default.
4. In general clients with contract status as 'Approved' are less likely to default.
5. In general clients with an academic degree are less likely to default.
6. In general clients who are widiws are less likely to default.