

# EDA ASSIGNMENT-BANK LOAN

Submitted To:- IIIT Bangalore(Upgrad)

Submitted By:- Ritika Sen

Dated:- 01/11/2022

# CONTENTS

- 1.Problem statement
- 2.Assumptions
- 3.Data Understanding
- 4.Data manipulation
- 5. Graphical Analysis
- 6.Conclusion

# PROBLEM STATEMENT

- **The main problem:-**
  - 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.
- **The action required:-**
  - This case study aims to identify patterns which indicate if a client has difficulty paying their instalments 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

- 1. Defaulters will be the customers which are unemployed or earn less or having more children.
- 2. Young people may be there in a good number in the list of defaulters due to inappropriate or less knowledge on how to manage finances.
- 3. There are more chances that the employed people or higher earning people will pay the loan on time.
- 4. Gender might not affect the defaulters or repayers statistics.

# DATA UNDERSTANDING

- **1.Application\_dataset:-**

- It contains all the information of the client at the time of application.
- The data is all about whether a client has payment difficulties or not.
- There are a total of 3,07,511 number of rows and 122 number of columns.
- There are three types of datatypes available:- float64 : 65
- int64 : 41
- object/string : 16
- There are negative values in some of columns.
- There are days and hour which are not in format. In the dataset.
- There are 67 columns comprising missing values majority of them have more than 40% missing values.

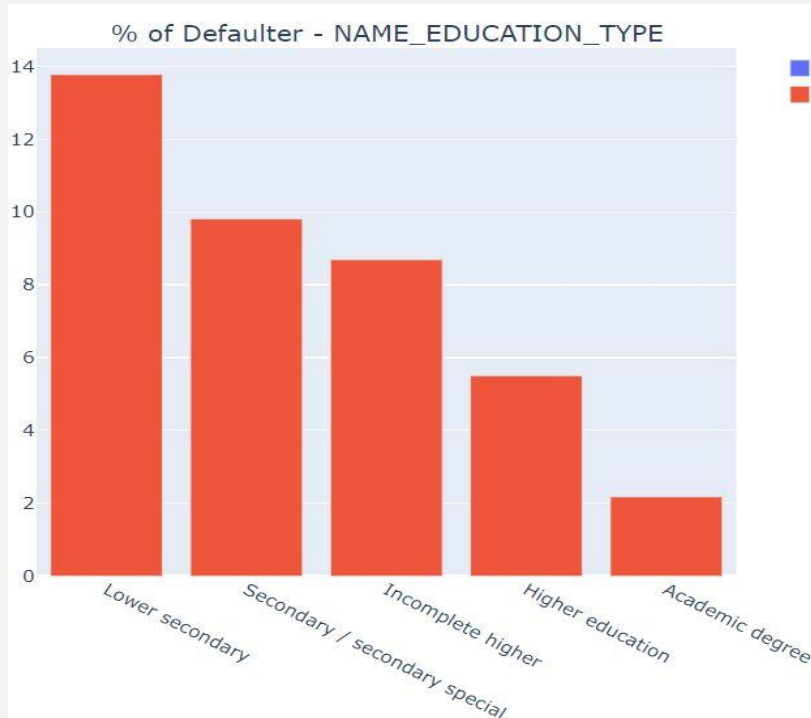
- **2.Previous\_Application\_Dataset:-**

- It contains information about the client's previous loan data.
- It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.
- There are 16,70,214 number of rows and 37 number of columns.
- Total number of datatypes available:float64 -15
- int64 -6
- object/string - 16
- There are negative values in some of the columns.
- There are days and hour which are not in the correct format.
- There are 16 columns which have missing values and majority of contain more than 40% missing values.

# DATA MANIPULATION

- Dropped columns with 40% null values except some important features.
- Analyzed FLAG data and dropped unwanted FLAG features. FLAG\_OWN\_REALTY has two types of values Y and N.
- We will change this to 1 and 0 In OCCUPATION\_TYPE there are a lot of missing values we will fill these values with Unknown.
- Fill median values in AMT\_ANNUITY column Worked with XNA values of CODE\_GENDER column Worked with XNA values of ORGANISATION\_TYPE column Convert columns DAYS\_BIRTH, DAYS\_EMPLOYED, DAYS\_REGISTRATION, DAYS\_ID\_PUBLISH, DAYS\_LAST\_PHONE\_CHANGE from negative to positive and convert the values to columns AMT\_INCOME\_TOTAL, AMT\_CREDIT, AMT\_ANNUITY,
- AMT\_GOODS\_PRICE YEARS\_EMPLOYED used to have outlier value around 1000 years which is impossible and hence it has been removed.

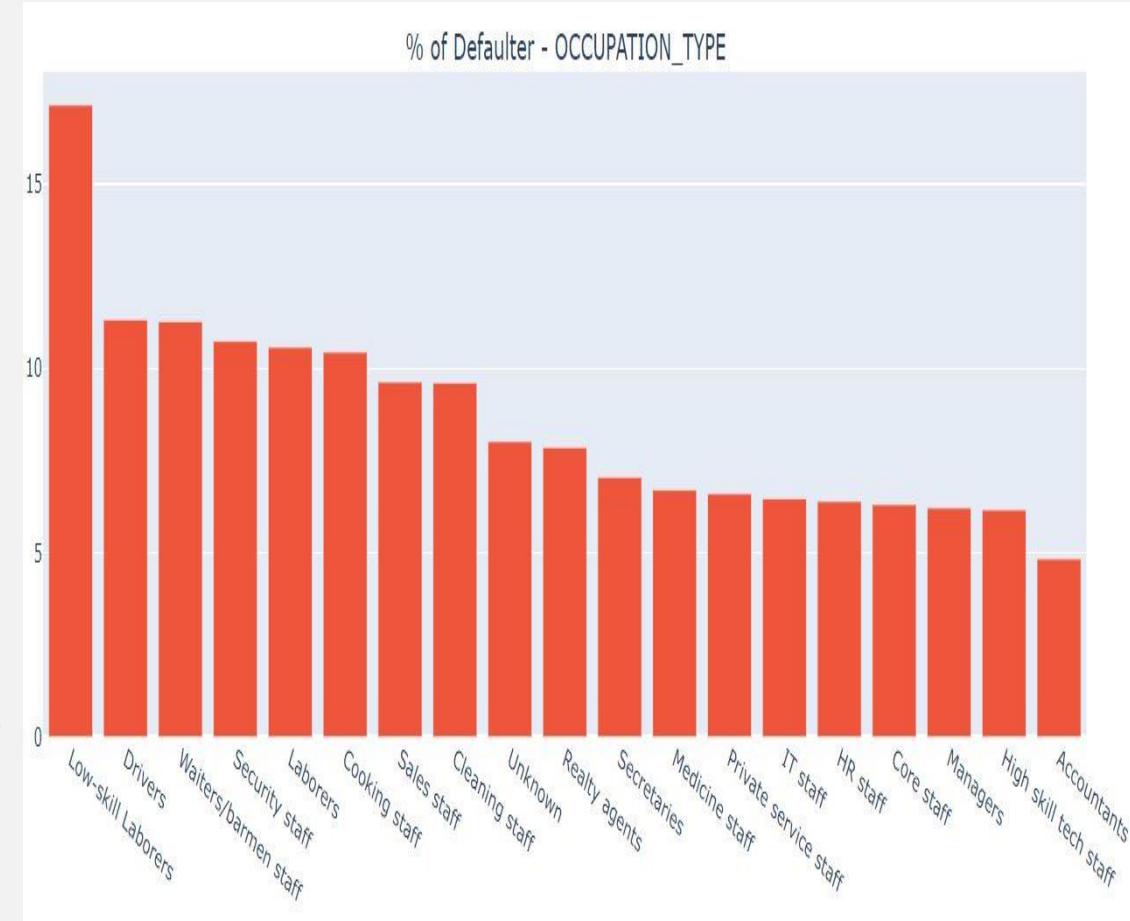
# GRAPHICAL ANALYSIS



**For Graph 1**-Clients with Secondary and Higher Education took more loan.

Clients with lower secondary education has higher percentage of Defaulters even though they are minority compared to other Educational level loan takers. Hence, Lower secondary education clients are risky for the company.

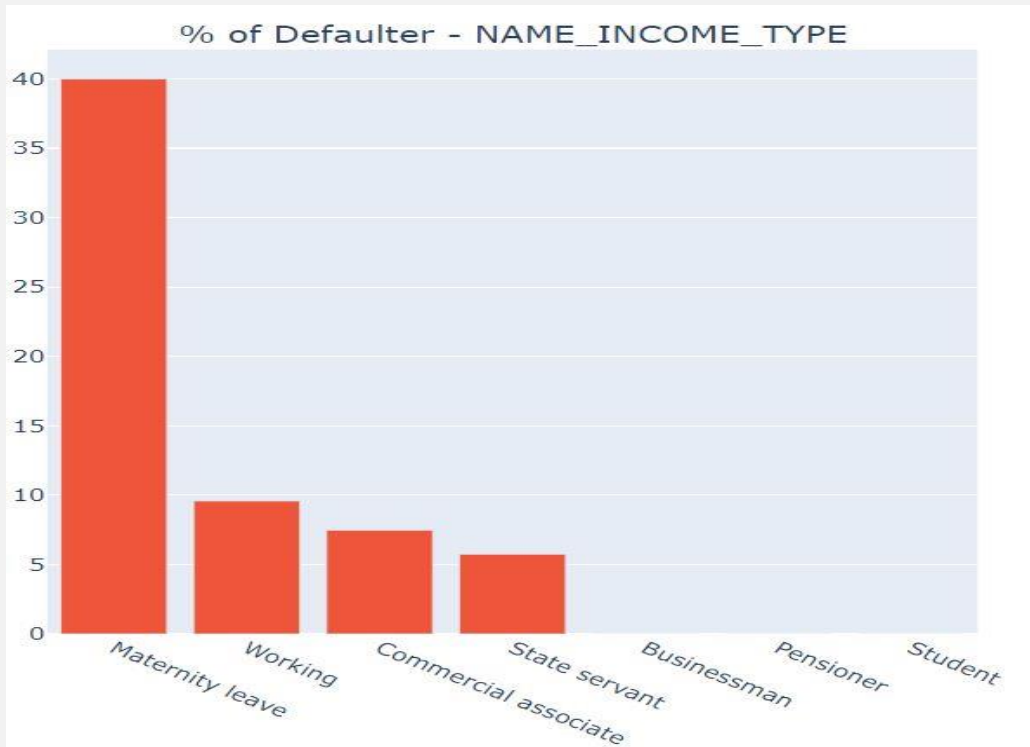
**For Graph 2**-Low-skill Laborers (17%), Drivers(11%), Waiters/barmen staff(11.27%), Security staff(10.74%) etc. are the defaulters. Hence, Clients with less income range are more likely to be a loan defaulters than high income range clients. So, Occupation\_type and Defaulters are inversely proportional to each other.



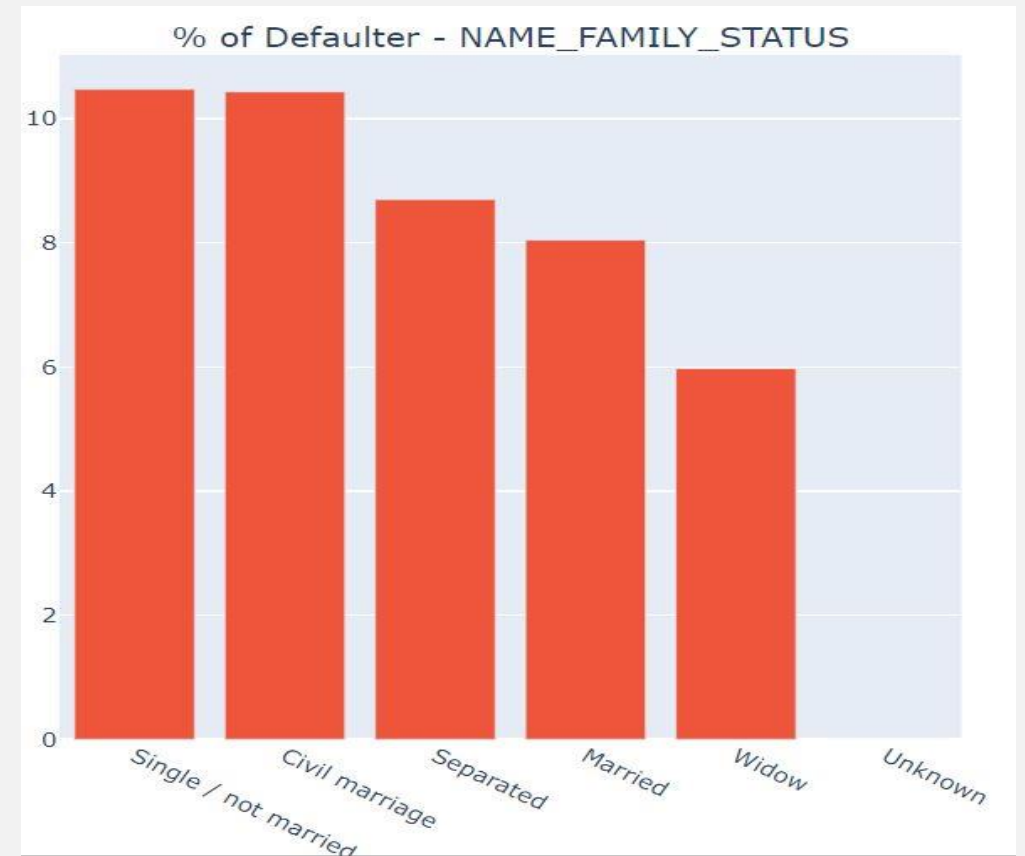


# GRAPHICAL ANALYSIS

- For Graph 3-Working class people are the loan takers.
- Unemployed or Maternity leave clients have very high defaulters even though they are minority compared to other income types.

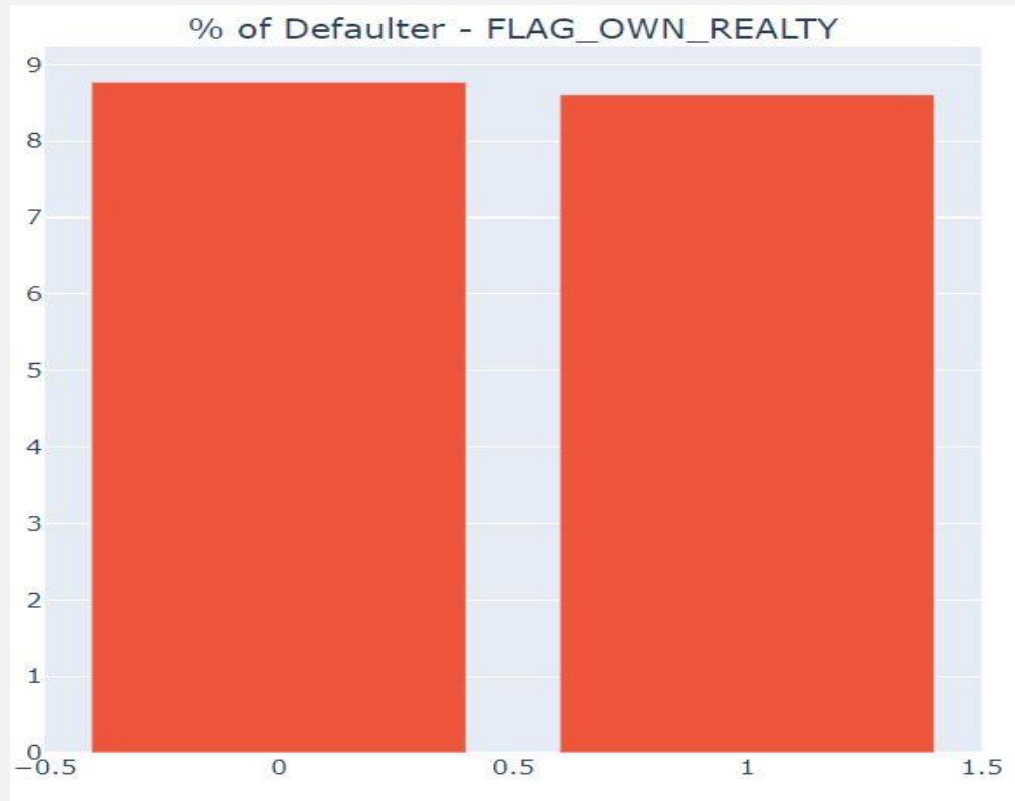


- For Graph 4-Civil marriages (10.43%) and single people (10.47%) have higher chance of defaulter than others.

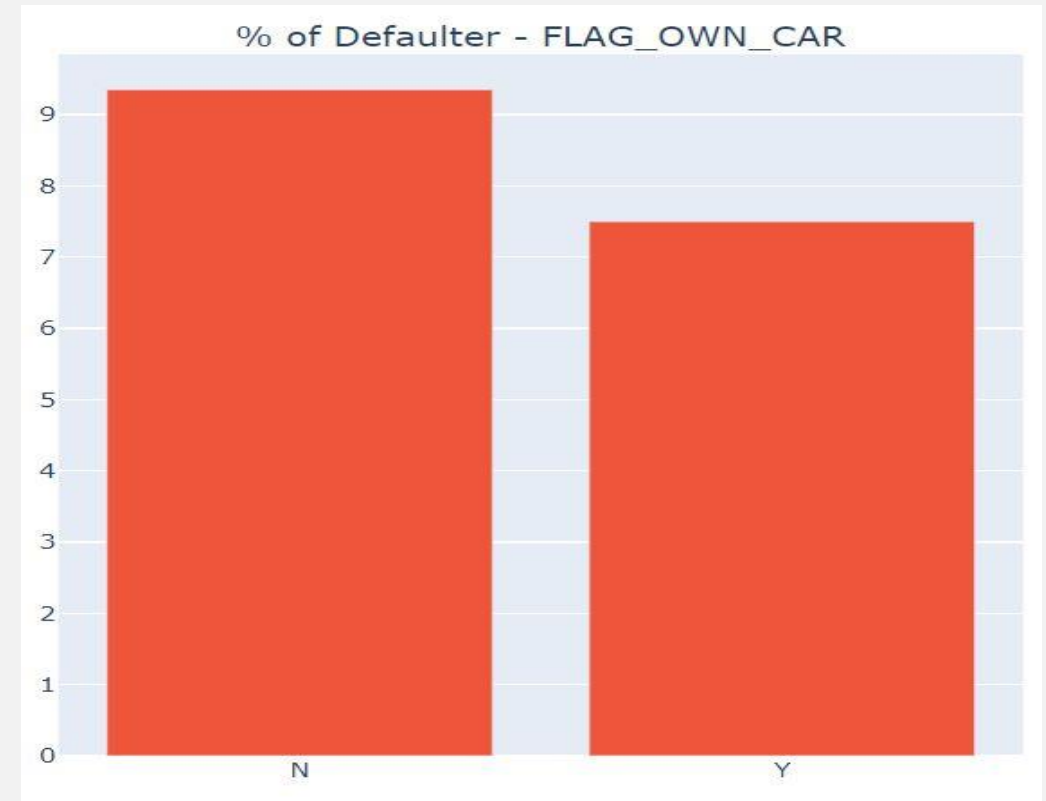


# GRAPHICAL ANALYSIS

- **For Graph 5-** The clients who own real estate are more than double of the ones that don't own. The defaulting rate of both categories are around the same (~8%). Thus we can infer that there is no correlation between owning a realty and defaulting the loan.



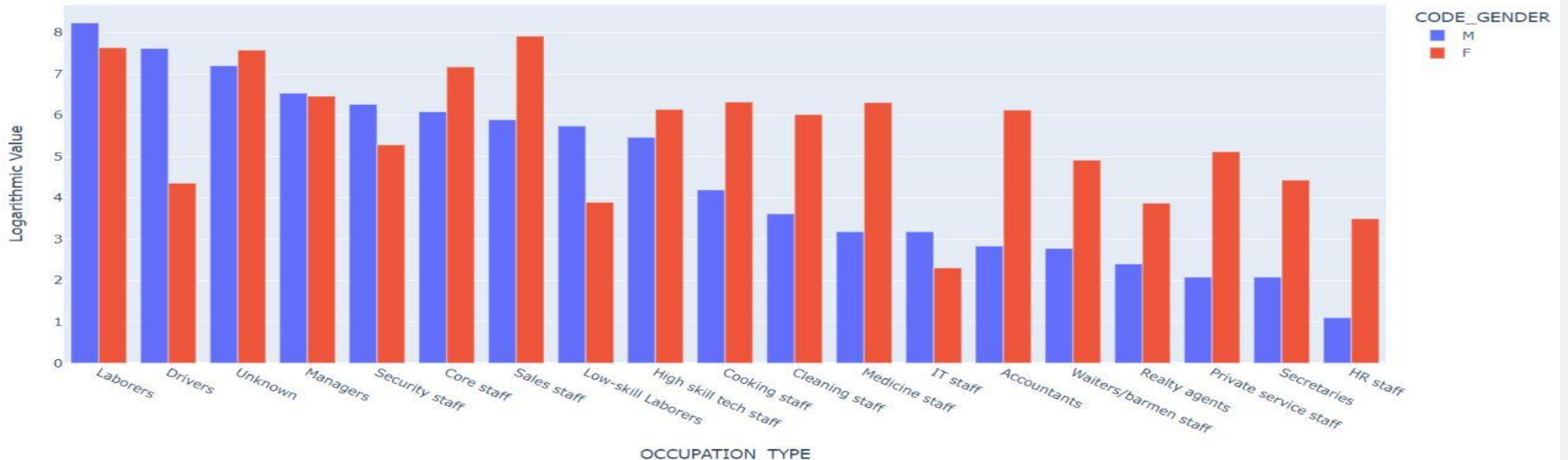
- **For Graph 6-** Clients with car are trust able on loan payments because we can see clients who owns car are more likely to repay the loan. We can say that clients without car are more likely to not repay the loan.



# GRAPHICAL ANALYSIS

- For Graph 7-Females with occupation type: - Accountants, Managers, core staff, HR staff etc. professions are the most defaulters as compared to their counterpart males. Males with occupation type: Realty agent, Low skilled labourers etc. professions are the most defaulters as compared to their counterpart female.

Distribution of OCCUPATION\_TYPE - Target 1

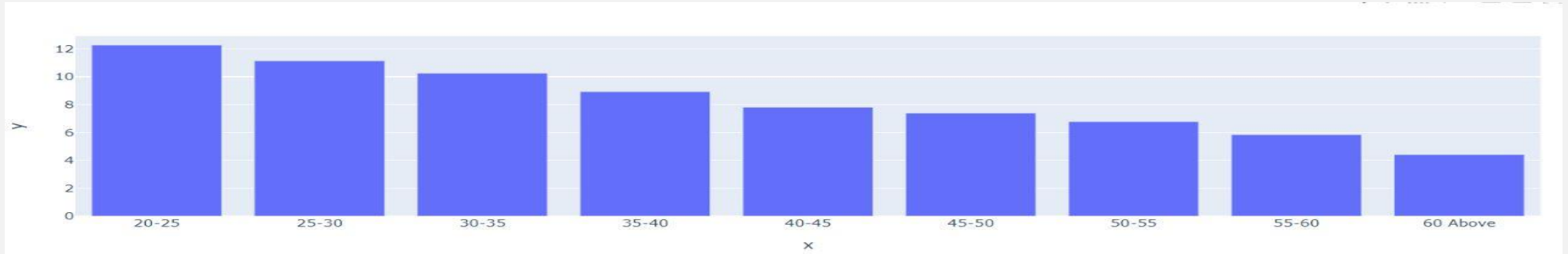


# GRAPHICAL ANALYSIS

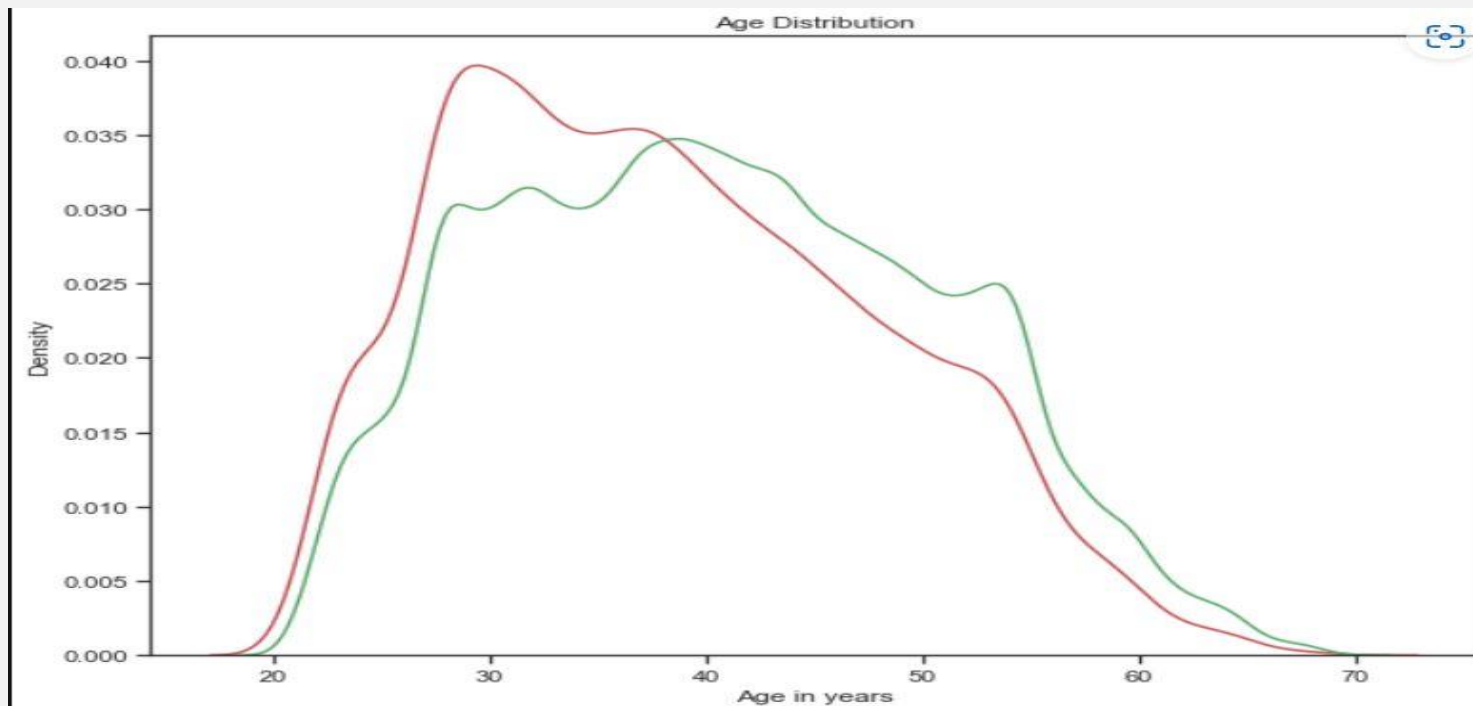
- **For Graph 8-** Male clients with less income range are the most defaulters. Default rate keeps decreasing as the income range increases i.e. inversely proportional.



# GRAPHICAL ANALYSIS

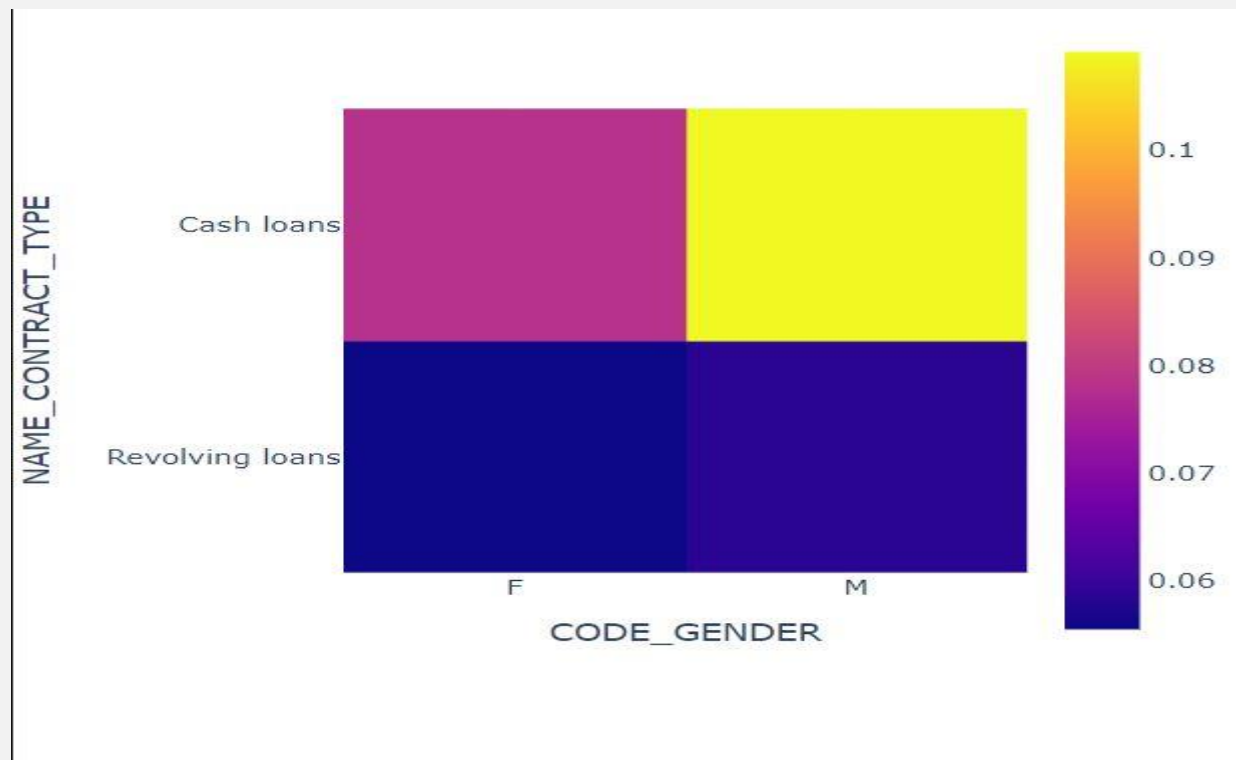


- For Graph 9-We can see that young people have 12.2% default rate while older people have only 4%.
- We need to make a focus on young people to make them learn about financial management to reduce default rate.



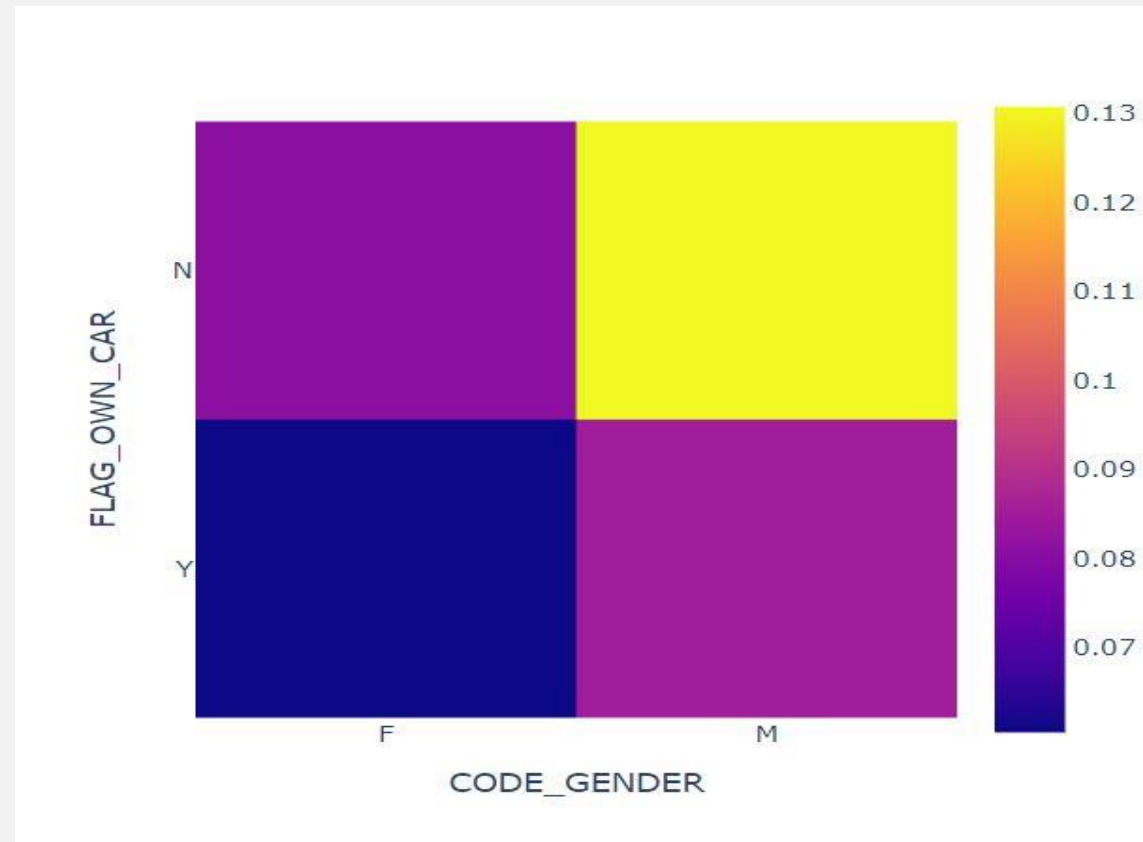
# GRAPHICAL ANALYSIS

- For Graph 10-As per the heatmap, Male candidates with cash loans are the one's with highest default rate as compared to other three sub categories,.
- But analyzing along with the bar plot, it's clear that both males/females having cash loans have almost equal ratio in defaulting.
- But looking at the revolving loans, even though the numbers are less, but the females who have taken revolving loans are the highest defaulters as compared to their male counter parts.



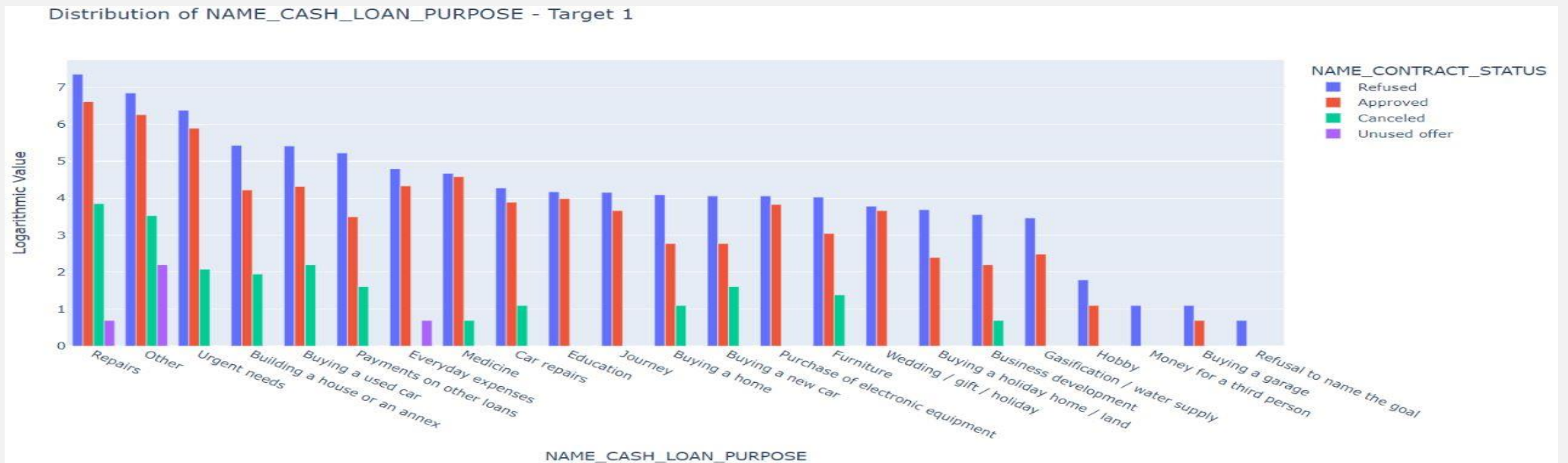
# GRAPHICAL ANALYSIS

- For Graph I I - According to heatmap, Male clients without car are the ones with highest default rate compared to the three sub categories.
- But according to bar plot, we have female clients more than male clients with highest default rate .



# GRAPHICAL ANALYSIS

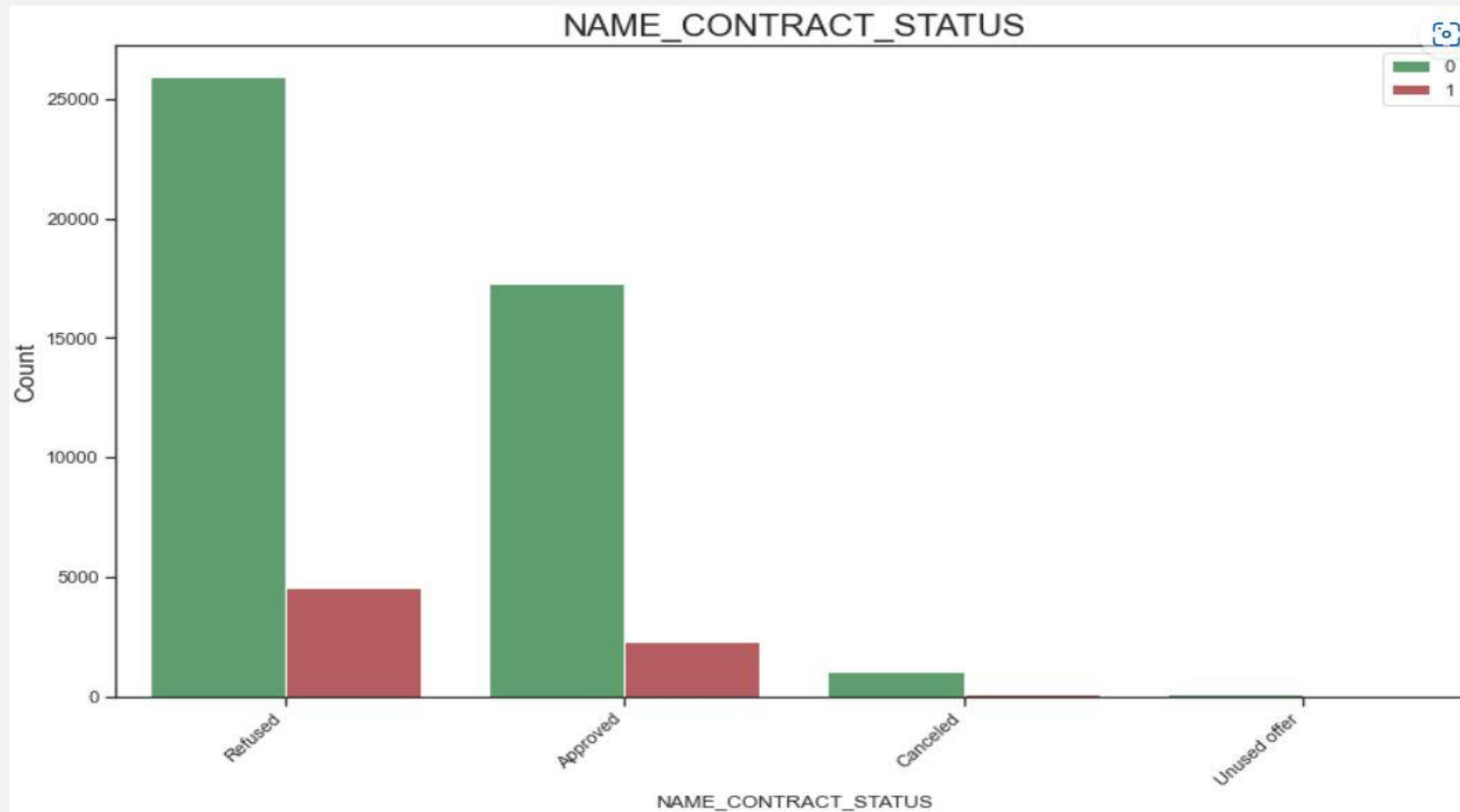
- For Graph 12 -Loan taken for the purpose of Repairs looks to have highest default rate Huge number application have been rejected by bank or refused by client which are applied for Repair or Other.
- From this, we can infer that repair is considered high risk by bank. Also, either they are rejected or bank offers loan on high interest rate which is not feasible by the clients and they refuse the loan.
- There are few places where loan payment is significant higher than facing difficulties. They are 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car' and 'Education' Hence we can focus on these purposes for which the client is having for minimal payment difficulties.





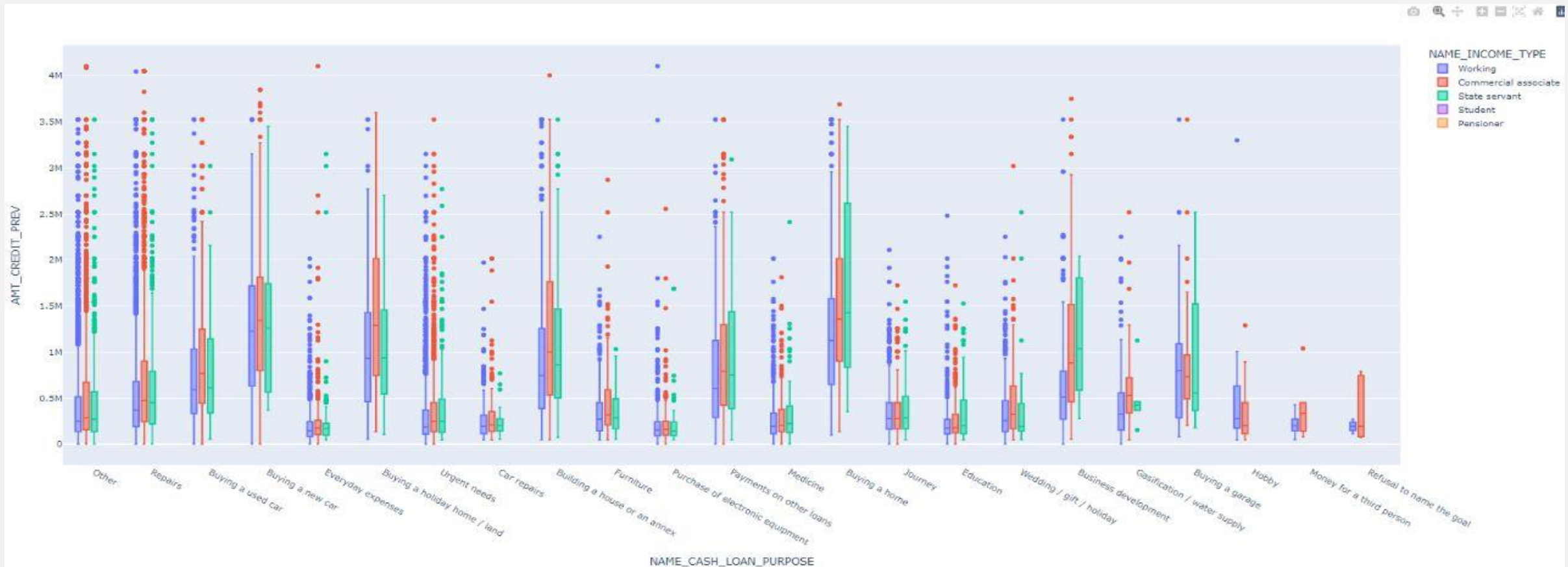
# GRAPHICAL ANALYSIS

- For Graph 13-90% of the previously cancelled client have actually repaid the loan. Revising the interest rates would increase business opportunity for these clients.
- 88% of the clients who have been previously refused a loan have paid back the loan in current case. Refusal reason should be recorded for further analysis as these clients could turn into potential repaying customer.



# GRAPHICAL ANALYSIS

- For Graph 14-Credit amount loan is higher in general for purposes like Buying a new car, Buying a holiday home/land, Buying a house or an annex buying a home.



# CONCLUSION

- Female applicants without cars are the most defaulters.
- Clients with low income range are the most in defaulters.
- Labourers, Sale staff, Drivers are the most defaulters in the dataset.
- Females with occupation HR\_STAFF, waitress, secretary are the most defaulters.
- Clients of young age less than 28 are most defaulters in dataset.
- After merging both the data.
- Banks should focus more on contract type 'Student' , 'Pensioner' and 'Businessman' with housing 'type other than 'Co-op apartment' for successful payments.
- After merging both the data, we came into the conclusion that Working, commercial associates & state servants are mostly defaulters.