# CREDIT Risk Analysis-EDA

ASHISH ARORA – DS45

# Problem Statement:

**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 default, then approving the loan may lead to a financial loss for the company.**
- **If the applicant is likely to repay the loan, then not approving the loan results in a opportunity cost of losing a business to the company.**

client with payment difficulties

# Data Provided:

**1. Previous Application Loan Data.**

- Approved: The Company has approved loan Application
- Cancelled: The client cancelled the application sometime during approval.
- Refused: The company had rejected the loan.
- Unused offer: Loan has been cancelled by the client but at different stages of the process.

**2. Current Application Loan Data.**

- Defaulter (1): Client with payment difficulties.
- Re-payer (0): All other cases.

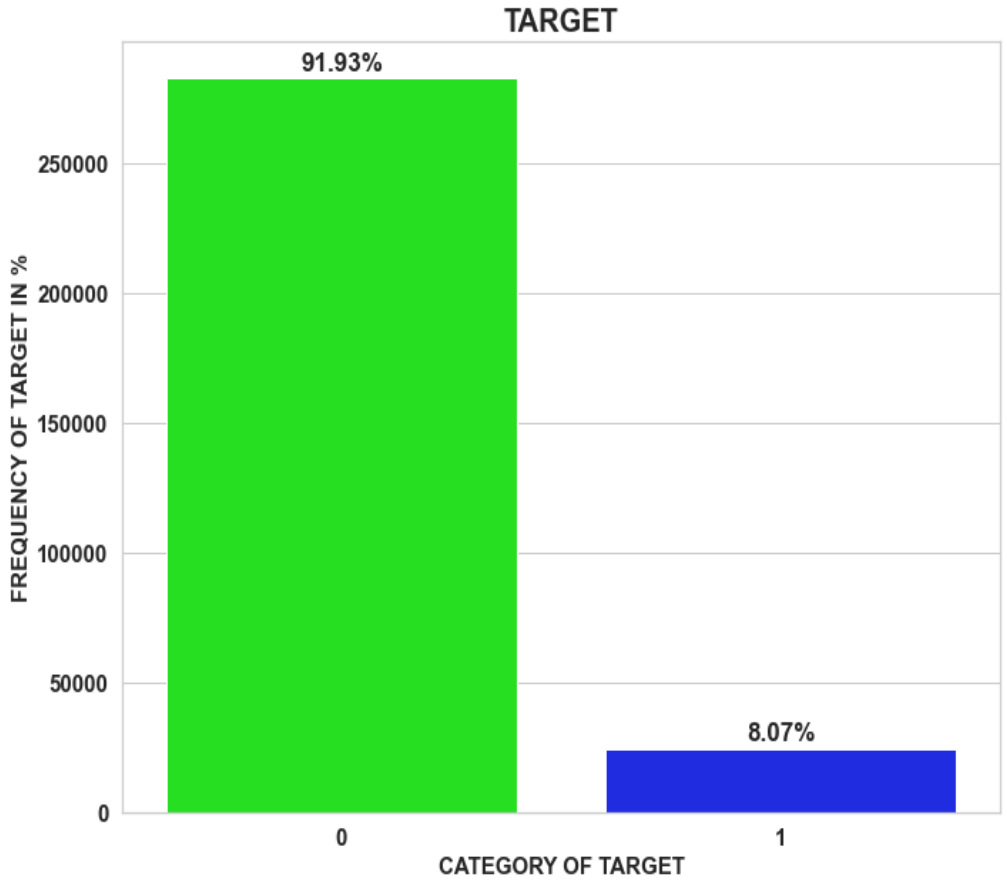
# Business Objective:

- Identification of such applicants  who are capable of repaying the loan and are not rejected.
- How consumer attributes and loan attributes influence the tendency of default.
- Finding the driving factors behind loan default.

# Steps Involved

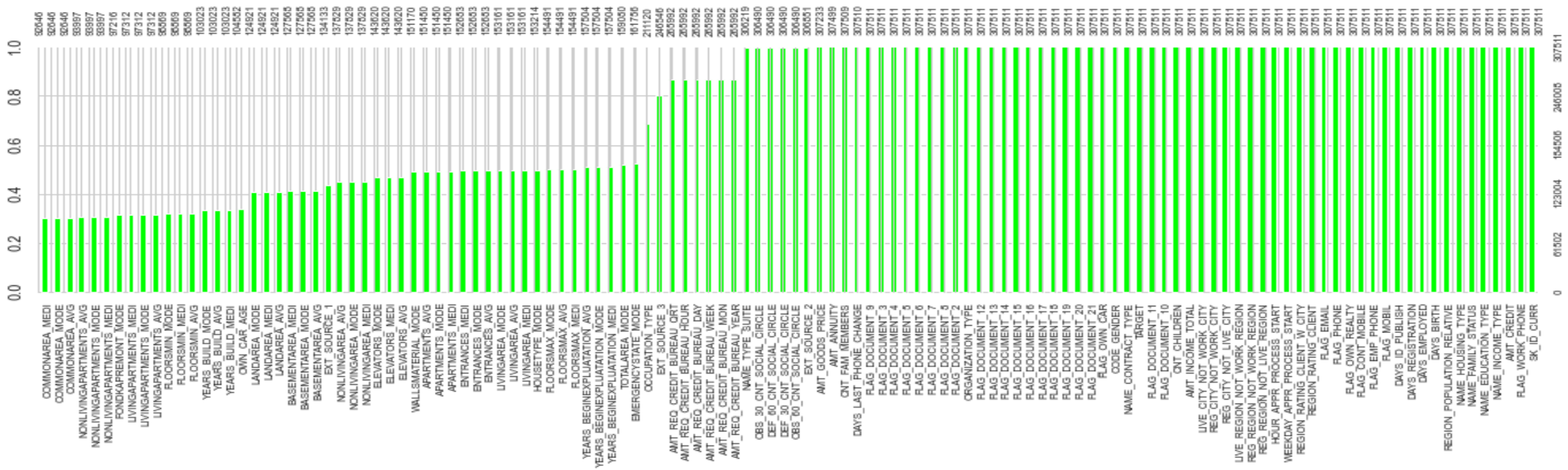
- **Data Cleaning**
  - Null Value
  - Standardization
  - Deleting Unnecessary columns
- **Data Manipulation**
  - Null value imputation
  - Outliers Treatment
  - Data Type Conversion
- **Imbalanced Data analysis**
  - Finding Ratios
- **Categorical Variable Analysis**
  - Univariate
  - Univariate and Bivariate
  - Bivariate and Multivariate
- **Numerical Variable Analysis**
  - Univariate
  - Bivariate
  - Multivariate
- **Conclusions**
  - Decisive factors

# Defaulter Count Ratio




- With the given information it can be interpret that our dataset is imbalanced. Most clients doesn't have difficulties in payment.
- However our priority is to identify all those prospects which may lead a client to default. As even a single default will impact company profits negatively with huge loss.
- So every bit of information which will be robust to default will be primary for us.

# Null Value Estimation




***Insights*: Bar plot depicts there are around 50-60 columns which have missing values**

# Dropping columns having null values above 40%

In [35]:
```
def null_count_comparison(df, target_col):
    my_dict = {}
    for i in df[target_col].unique():
        my_dict[i]  = df[df[target_col] == i].isnull().sum()

    dd = pd.DataFrame(my_dict)
    dd = dd.loc[~(dd==0).all(axis=1)]
    dd["Total"] = dd.sum( axis =1)
    return round(dd/df.shape[0]*100, 4).sort_values(by="Total", ascending =False)
```

In [36]:
```
null_count= null_count_comparison(curr_app,"TARGET")
null_count
```

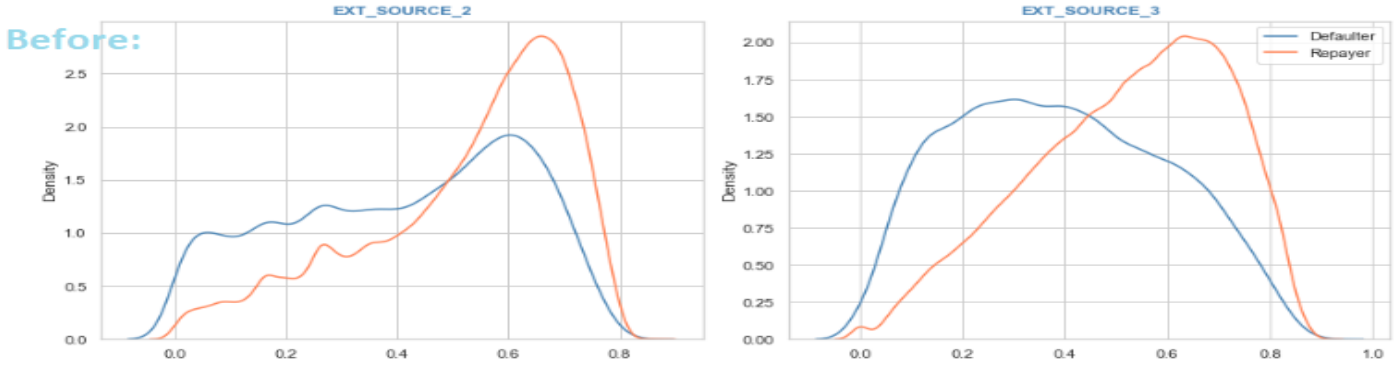
Out[36]:

| | 1 | 0 | Total |
|---|---|---|---|
| COMMONAREA_MEDI | 5.991000 | 63.881300 | 69.872300 |
| COMMONAREA_AVG | 5.991000 | 63.881300 | 69.872300 |
| COMMONAREA_MODE | 5.991000 | 63.881300 | 69.872300 |
| NONLIVINGAPARTMENTS_MEDI | 5.959800 | 63.473200 | 69.433000 |
| NONLIVINGAPARTMENTS_MODE | 5.959800 | 63.473200 | 69.433000 |
| NONLIVINGAPARTMENTS_AVG | 5.959800 | 63.473200 | 69.433000 |
| FONDKAPREMONT_MODE | 5.894100 | 62.492100 | 68.386200 |
| LIVINGAPARTMENTS_MODE | 5.893100 | 62.461800 | 68.355000 |
| LIVINGAPARTMENTS_MEDI | 5.893100 | 62.461800 | 68.355000 |
| LIVINGAPARTMENTS_AVG | 5.893100 | 62.461800 | 68.355000 |
| FLOORSMIN_MODE | 5.856100 | 61.992600 | 67.848600 |

- With the given code of block, we were able to create a dataframe containing name of all those columns having missing values and proportion of missing value with respect to target variable.
- Total 67 columns have null values, out of which 49 columns have more than 40% values as null.
- All the null columns which are above 40 are normalized data and related to the client properties attributes. No values can be found out of this. So we dropped them.

# Data Manipulation : Impute Outliers

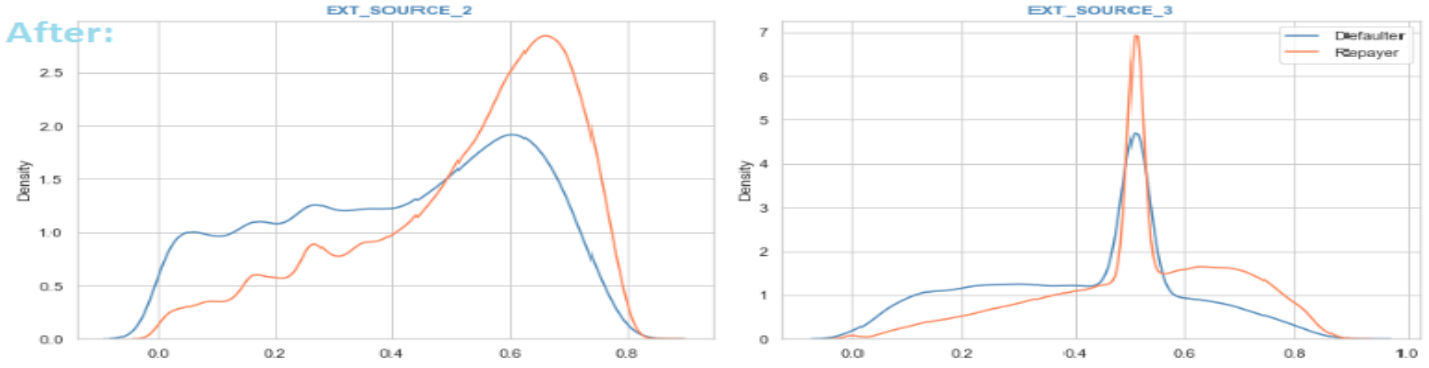
Before:




**Insights**

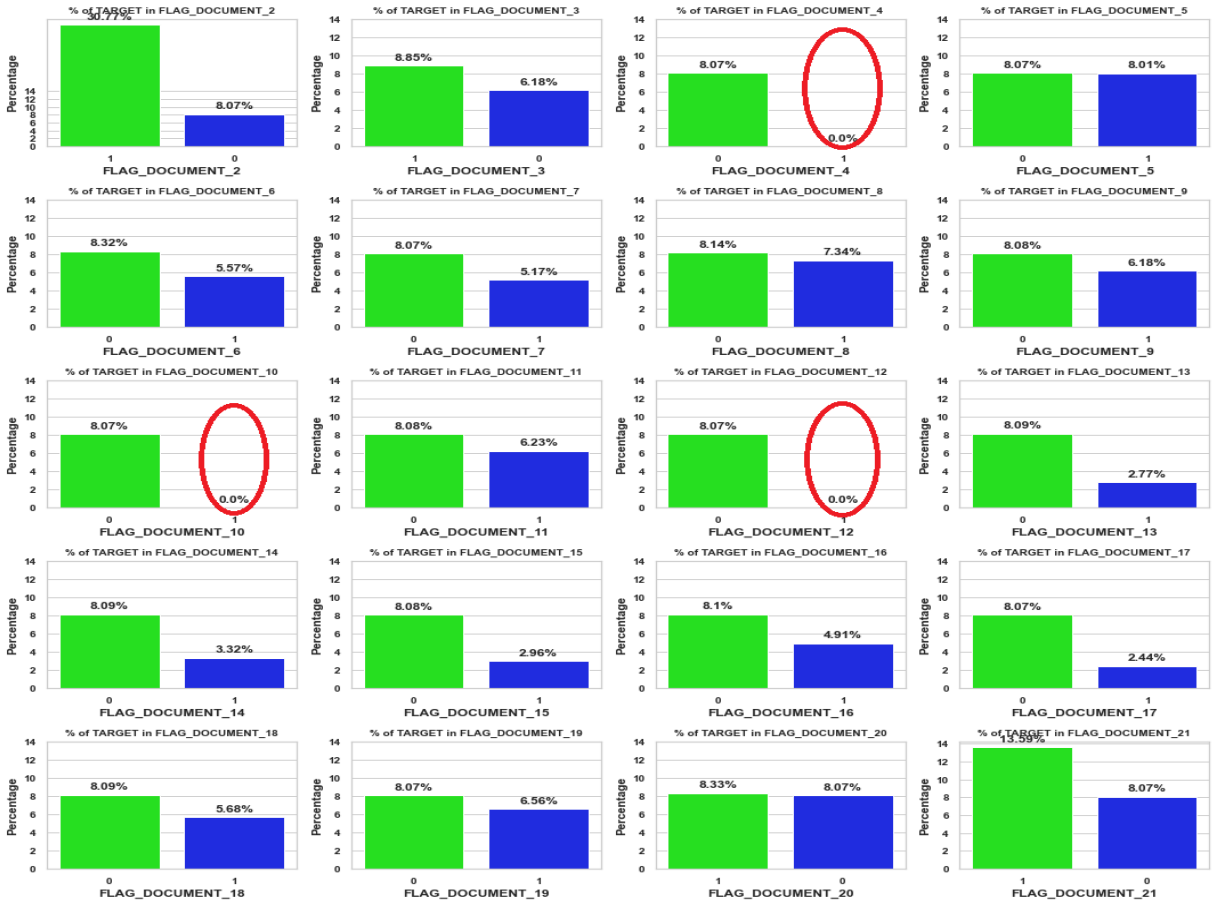
- **For Ext_source_2, we can said that for repayer and defaulter distribution are left skewed.**
- **Same as the case for Exit_source_3.**
- **Better fill the na values for both the values with MEAN as mean lies before median in left skewed distribution.**

After:




- All the information about the enquiries to Credit Bureau about the client in different periods has outliers and missing values.
- Almost all variables about the enquires to credit bureau have high peak over value 0. Even IQR is 0.
- As data is normalized, we can't even fix the outliers. All these variable share equal proportion of missing values so i decide to keep yearly enquiry column rest drop and fill its value with median.

# Data Manipulation: Deleting Unnecessary columns



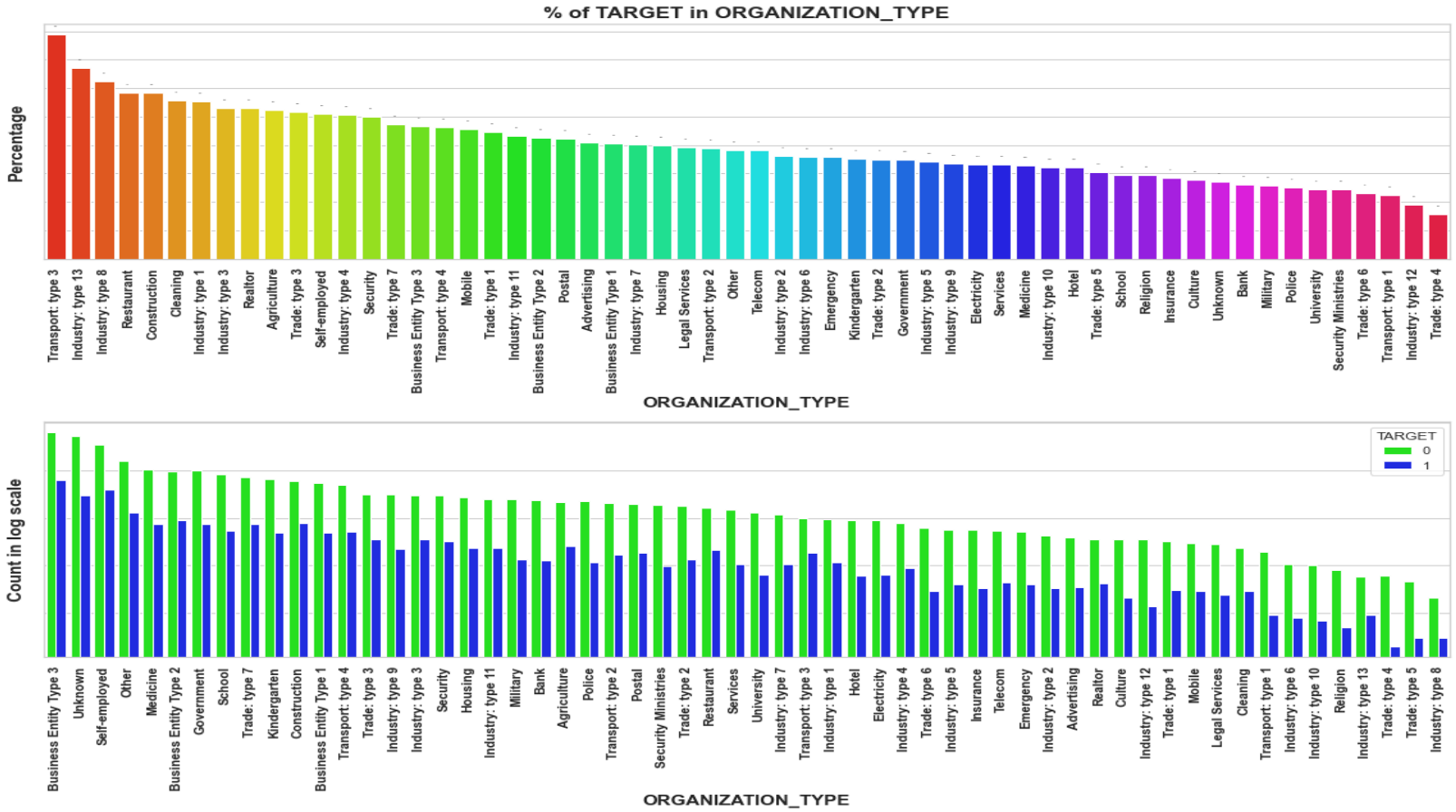

- 'FLAG_DOCUMENT_4', 'FLAG_DOCUMENT_10' and 'FLAG_DOCUMENT_12' have no associativity with the default rate.
- Though all other flag documents has some associativity of being default, For us even a small percentage which may lead to default is crucial.
- Hence we would keep all the rest flag documents columns and will drop these 3 columns

# Data Manipulation : Filling Missing Values

| | | AMT_ANNUITY | |
|---|---|---|---|
| | CODE_GENDER | F | M |
| ORGANIZATION_TYPE | OCCUPATION_TYPE | | |
| Other | Accountants | 723.000000 | 15.000000 |
| | Cleaning staff | 442.000000 | 36.000000 |
| | Cooking staff | 361.000000 | 20.000000 |
| | Core staff | 607.000000 | 122.000000 |
| | Drivers | 72.000000 | 874.000000 |
| | HR staff | 40.000000 | 3.000000 |
| | High skill tech staff | 517.000000 | 266.000000 |
| | IT staff | 17.000000 | 31.000000 |
| | Laborers | 1116.000000 | 1899.000000 |
| | Low-skill Laborers | 25.000000 | 90.000000 |
| | Managers | 821.000000 | 508.000000 |
| | Medicine staff | 982.000000 | 52.000000 |
| | Private service staff | 121.000000 | 17.000000 |
| | Realty agents | 41.000000 | 6.000000 |
| | Sales staff | 814.000000 | 109.000000 |
| | Secretaries | 131.000000 | 5.000000 |
| | Security staff | 202.000000 | 323.000000 |
| | Unknown | 3691.000000 | 1501.000000 |
| | Waiters/barmen staff | 71.000000 | 10.000000 |
| XNA | Cleaning staff | 2.000000 | NaN |
| | Unknown | 45269.000000 | 10103.000000 |

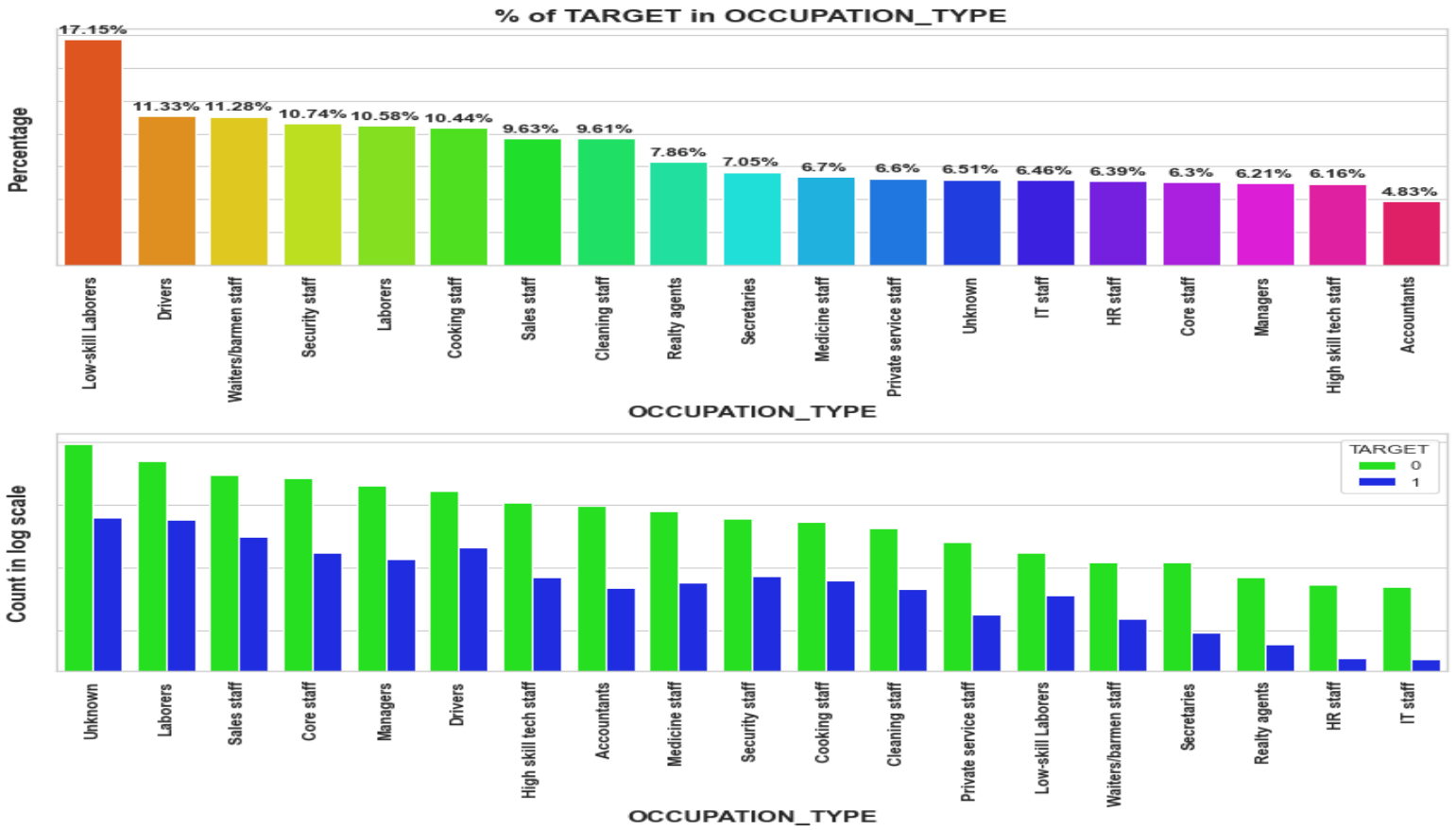
**Inferences:**
-- It is evident that xna organisation type is aslo not known for occupation type.
-- Hence it is not possible to impute with any others.
-- Better to keep it as separate category and name it "Unknown".**

- In organization type, there was a category called others as well as XNA category, so in order to check can we make XNA fill with others we created a pivot table and find the associativity of others with occupation type and gender.
- It was evident as XNA organization type was also not known for occupation type.
- So we impute the XNA values with unknown.




- Transport type 3 and industry type 13 are most prone to default.
- Whereas Industry type 4 and Trade Type 4 are least prone to default.
- However Business entity Type 3 has applied for most loans and industry type 8 has applied for least loans.

# Occupation Type




- After filling null values, people who are not known followed by laborers and sales staff have taken the most number of loans and then the sales staff. High skilled jobs like IT, accountant and secretariat has taken low loans and also has low default rate.

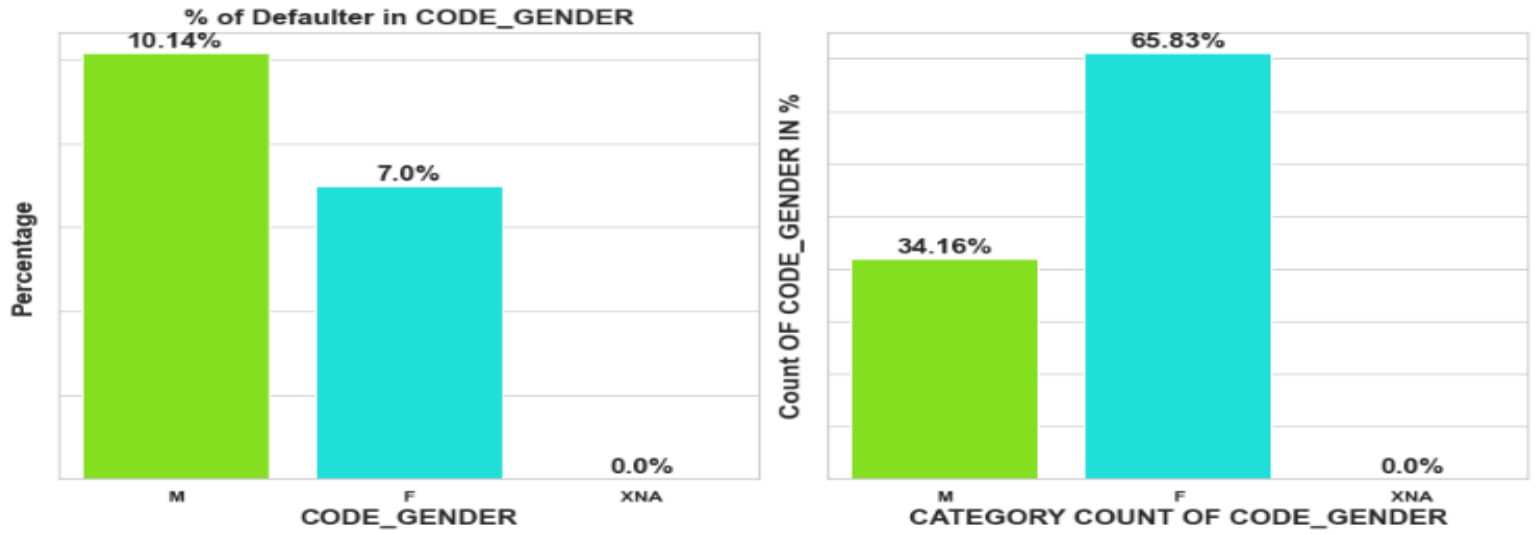
- However, Low skill laborers, drivers and waiters have more chance of doing default.

# Gender Imbalance: 65% Female

**Replace it with the "F" as mostly are female**

```
curr_app["CODE_GENDER"].replace("XNA", 'F', inplace =True)
```

```
comparison_categorical_count_plot(df = curr_app, target_var = "TARGET", attribute = "CODE_GENDER", scale_log = False,
                        y_Ticks_to_rotate = 360, x_Ticks_to_rotate = 360, layout_vert=False, annot_size = 17, fig_size=(13, 6) )
```



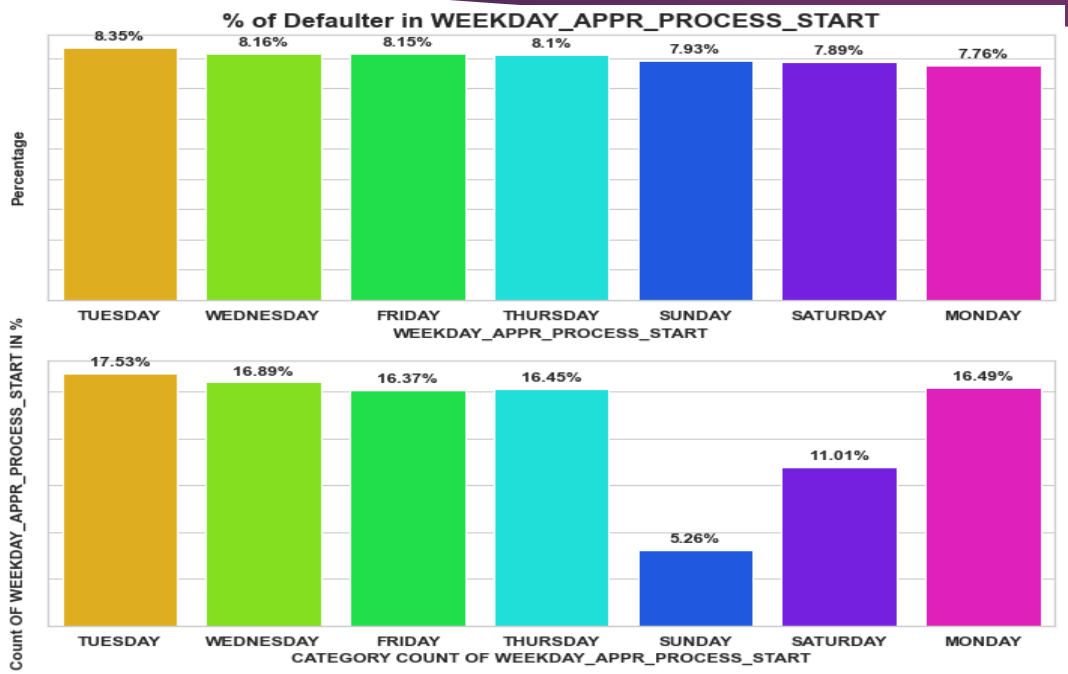

**Insights**

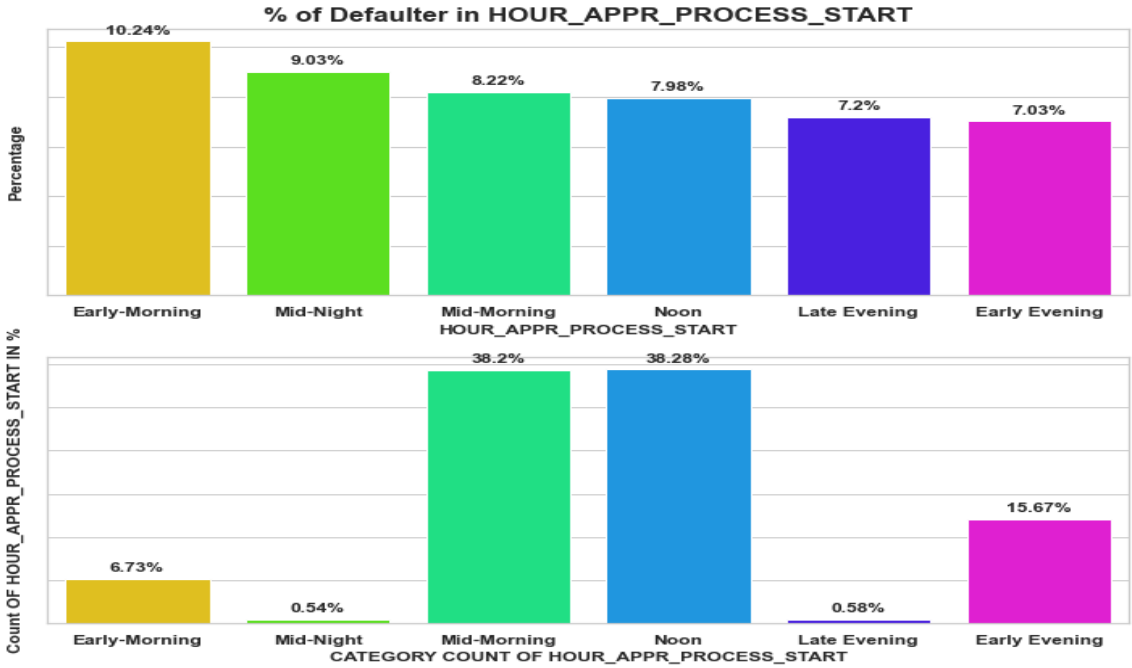
- **The given information tells that 66% loans are taken from female and 34% loans are taken male.**
- **However male client's are more prone to default.**

Activa
Go to S

- There were 4 values which were not known so we replaced it with females. The given information tells that 66% loans are taken from female and 34% loans are taken male.
- However male client's are more prone to default.

# Role of week and hours in Loan Application

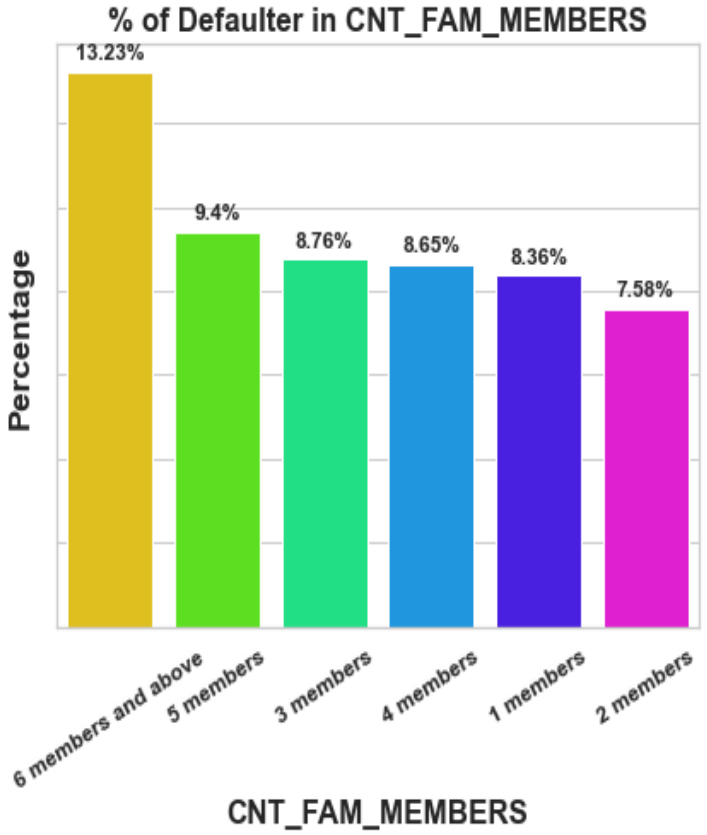


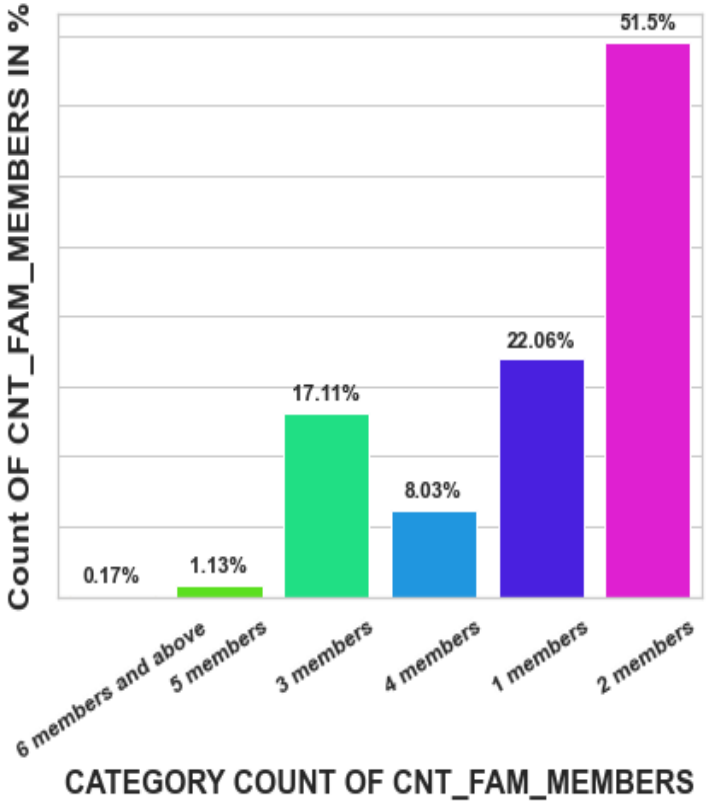




- On Tuesday maximum number of loan applications are filled.
- On Saturday and Sunday least number of loan applications are filled as these are weekends as well. Almost all days share same probability to default.
- However Tuesday are more prone to default cases whereas Monday are less prone to default case

- With the given graph we can say that Most of the clients apply loans at first half of the day.
- And to be precise in collectively 76% people apply loan at Mid Morning and at Noon hours.
- Clients who apply loans at Earning Morning are more prone to default.
- Whereas clients who apply at late early evening are less prone to default.

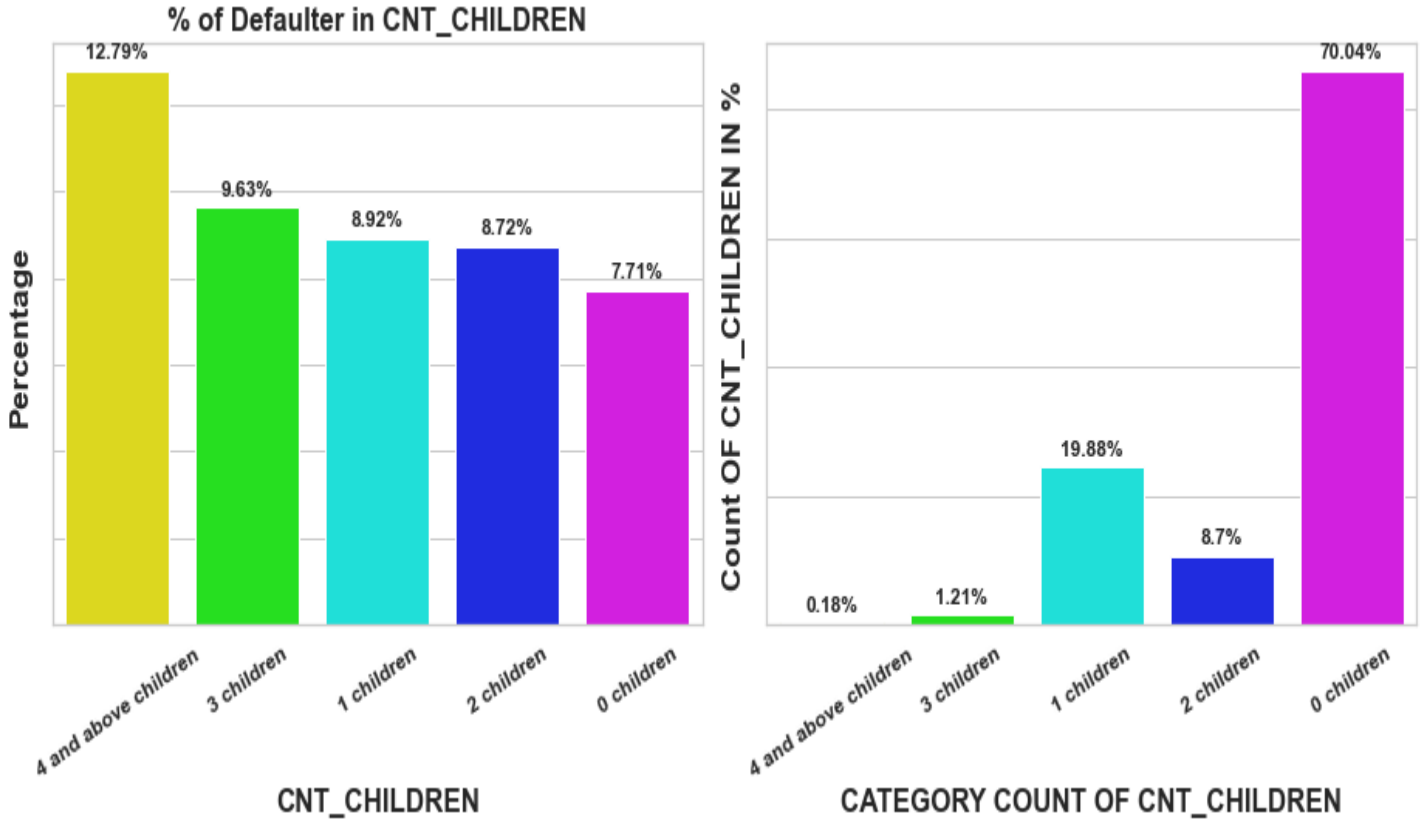
# Default Rate: Based on Family Members




- First we bin the family members count, as there are many outliers.
- Clients with 2 family members has applied for the maximum loans and they also have less default rate.
- Clients having family members above 6 have worst default rate comparing to other family member groups.

# Default Rate: Based on Children count

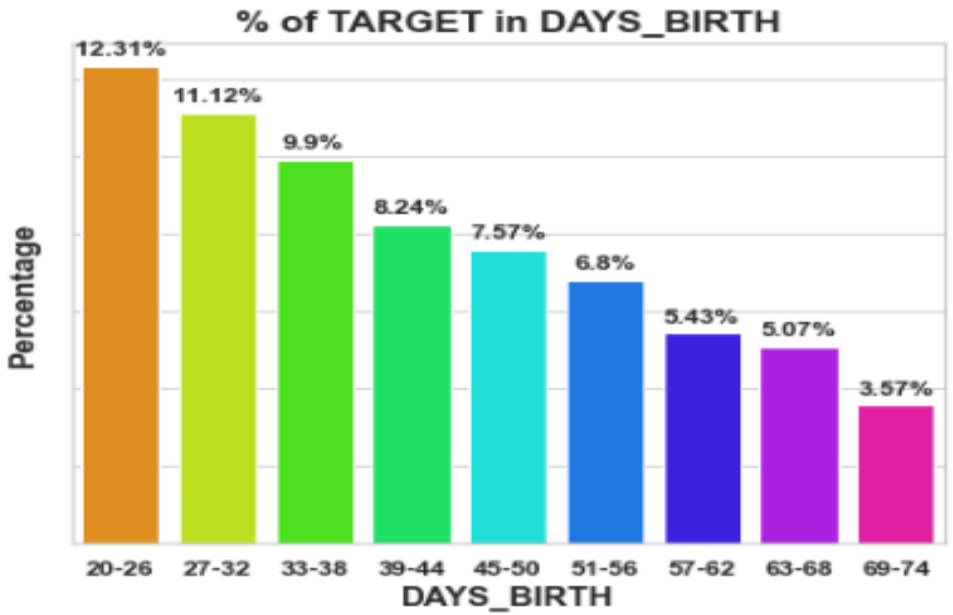


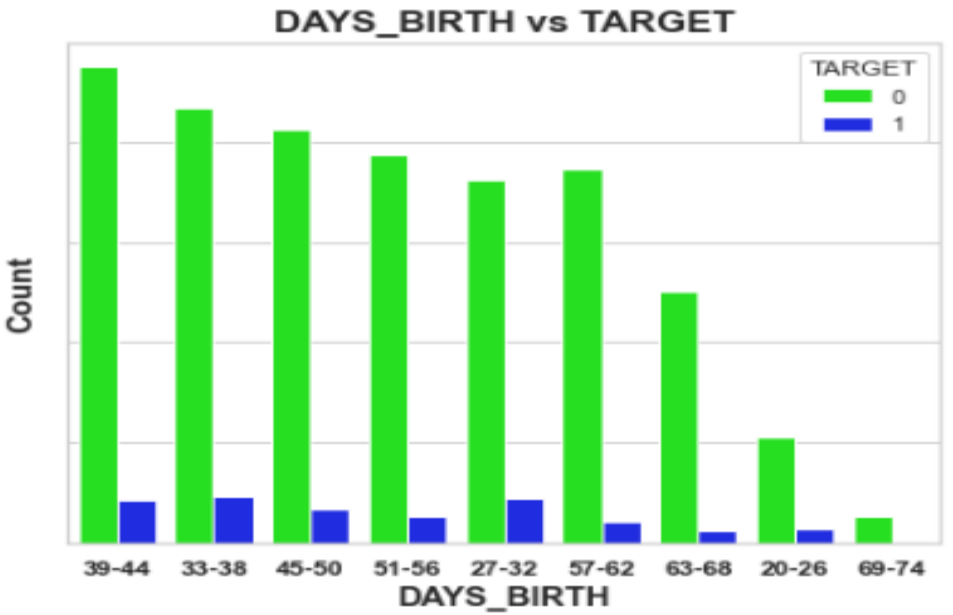




- Firstly, we bin the children count, as there are many outliers.
- Client's which have zero children has applied for the most loans and on top of it they have less default rate.
- 3 and above children have applied for least loans and to the given information they have high default case.

# Default Rate: Age of client



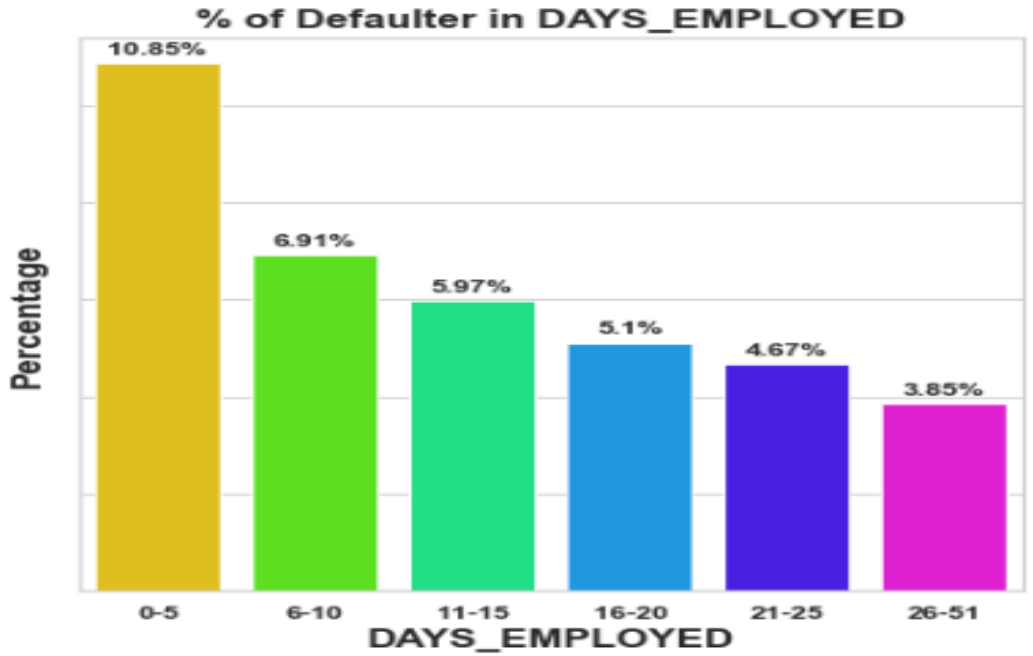



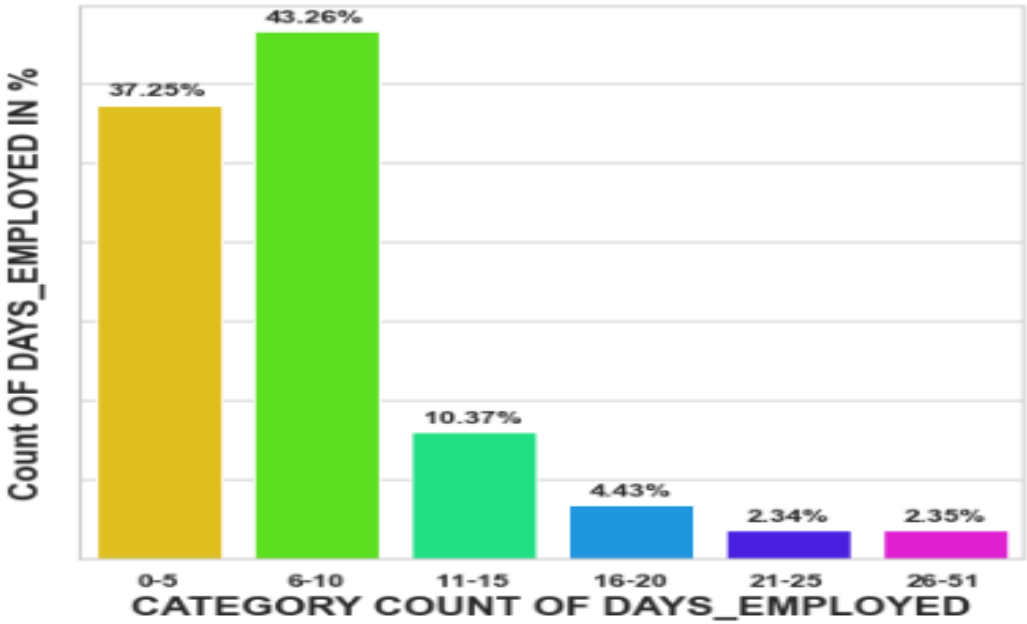

**Insights**

- Client's at age between 39-44 has applied the most numbe of LOANS
- Moreover their probabilty of getting default is 7.5%, which is not that good but also not bad.
- Clients at age between 20-26 have applied for second least loans, though their prob of doing default is high.
- Clients whose age is 69-44 has applied most least loans, Though they are least prone to default.

Activa

# Default Rate: Client's Employment



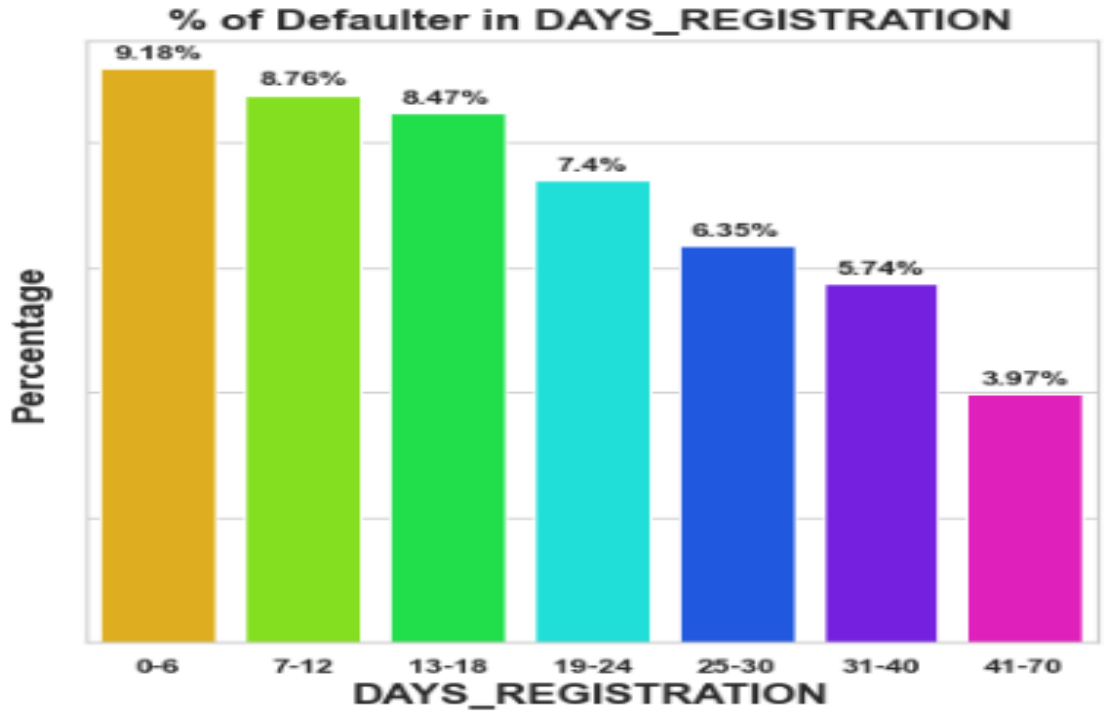



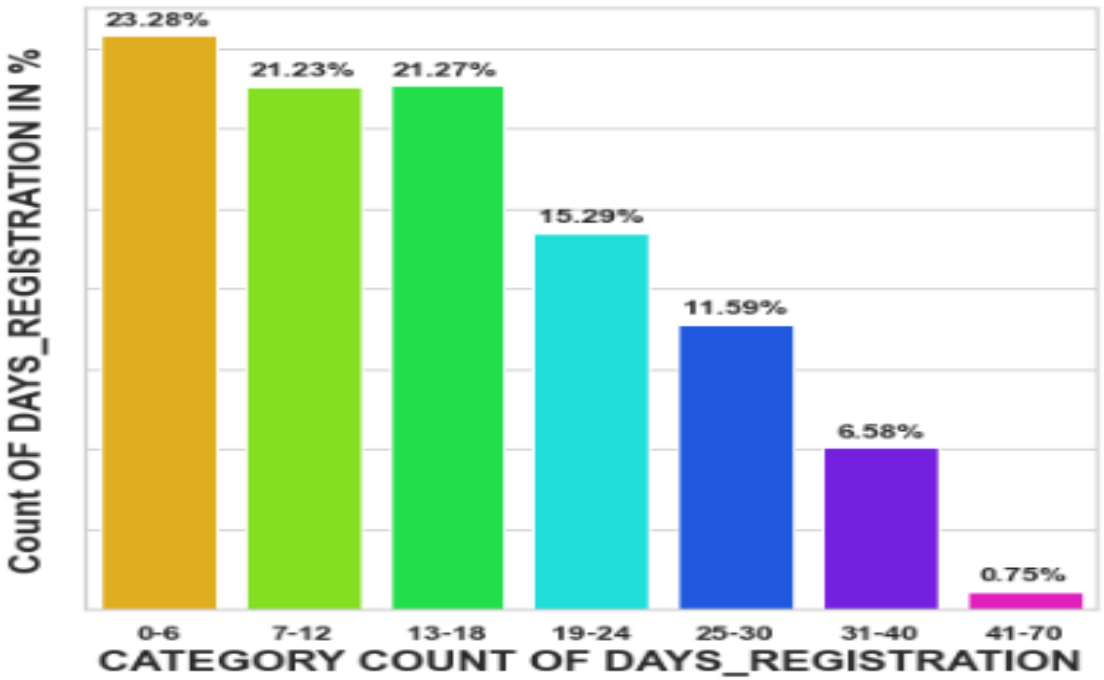

**Insights**

- Client's WHO ARE EMPLOYED FOR 6-10 YEARS AND FOLLOWED BY 0-5 YEARS HAS COLLECTIVELY AAPLIED FOR 80 OF LOANS%
- Clients whose job employed years is in 0-5 years are mostly prone to default
- As the years of work experience increases loans are required less for those applicants and even the default ratio also increases

Acti
Go to

# Default Rate: Client's Registration change

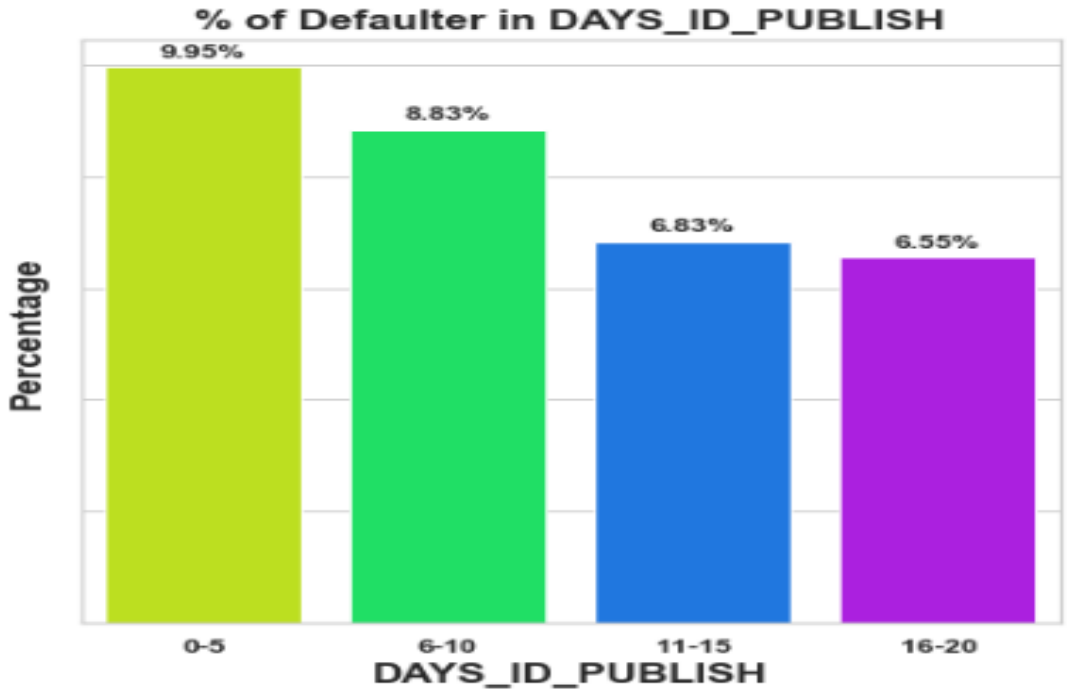


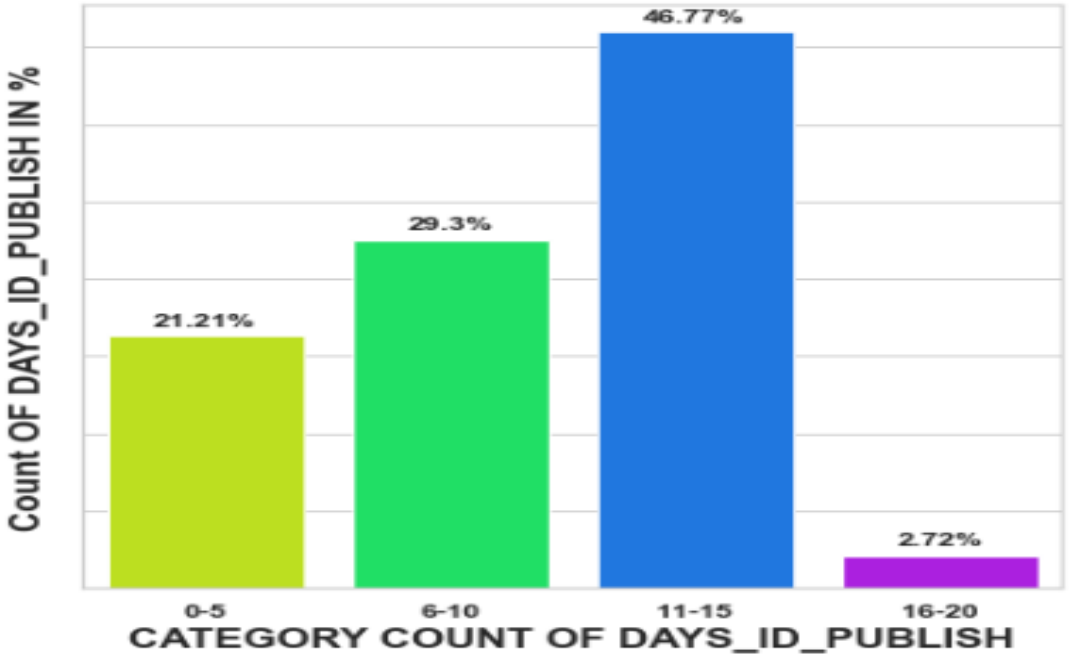




**Insights**

- Out of total applications, 23% of clients have changed the registation 0-6 years followed by 7-12 years.
- As the years of registartion increases default ratio also decreases.

# Default Rate: Client's Document change




**Insights**

- Out of all the client's clients who changes his document 11-15 times are highest and lowest id for 16-20 times.
- However such clients are not prone to defualt when compare to clients who change their identity documents 0-5 times.
- Their probability of defaulting is 10% whereas for 11-15 times category it is 7%.

# Correlation Coefficient

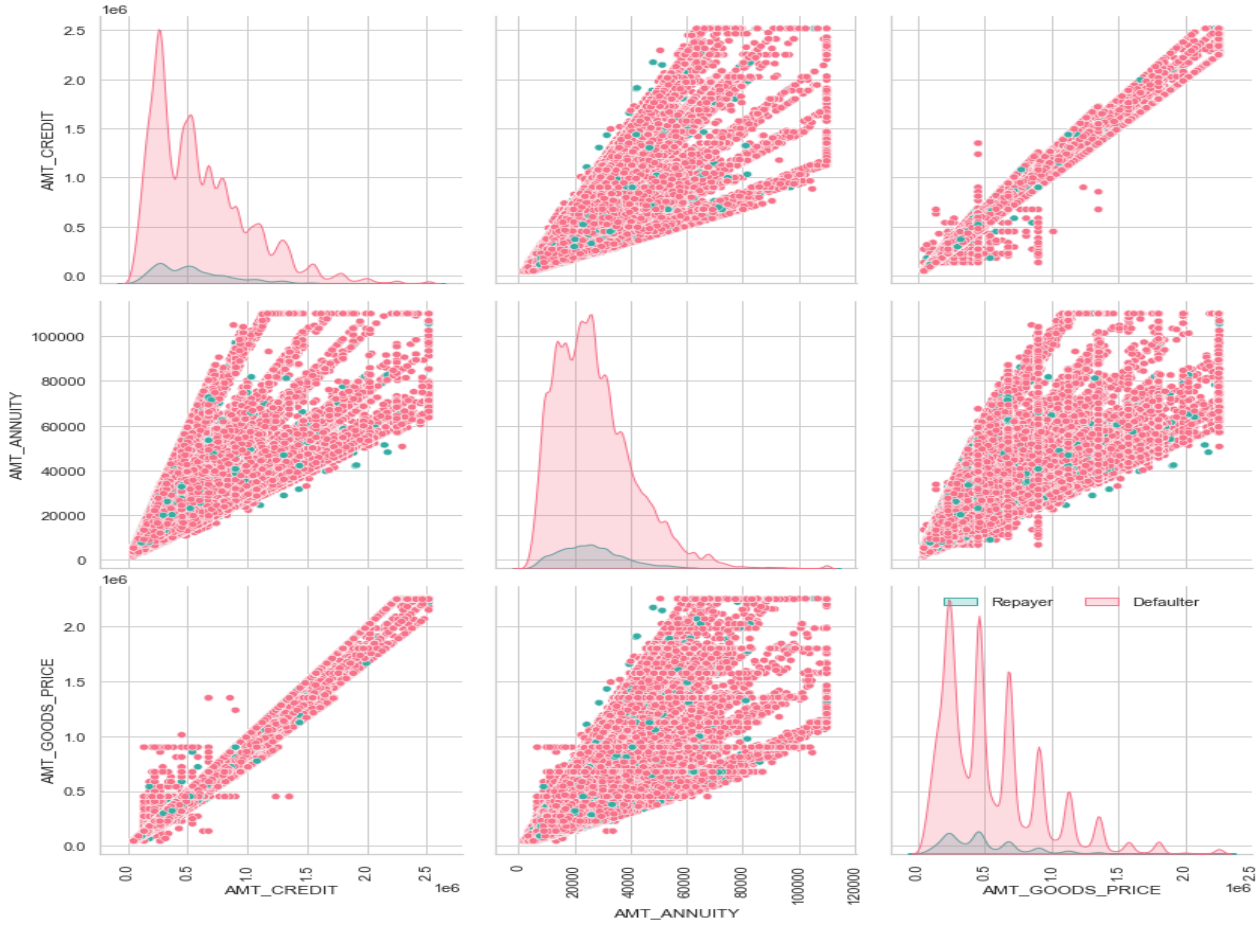

## Re-payer: Correlation Coefficient

- **Amount Credit and Amount Annuity = 0.78**
- **Amount Goods Price and Amount Annuity = 0.78**
- **Amount Goods Price and Amount Credit = 0.99**

## Defaulter: Correlation Coefficient

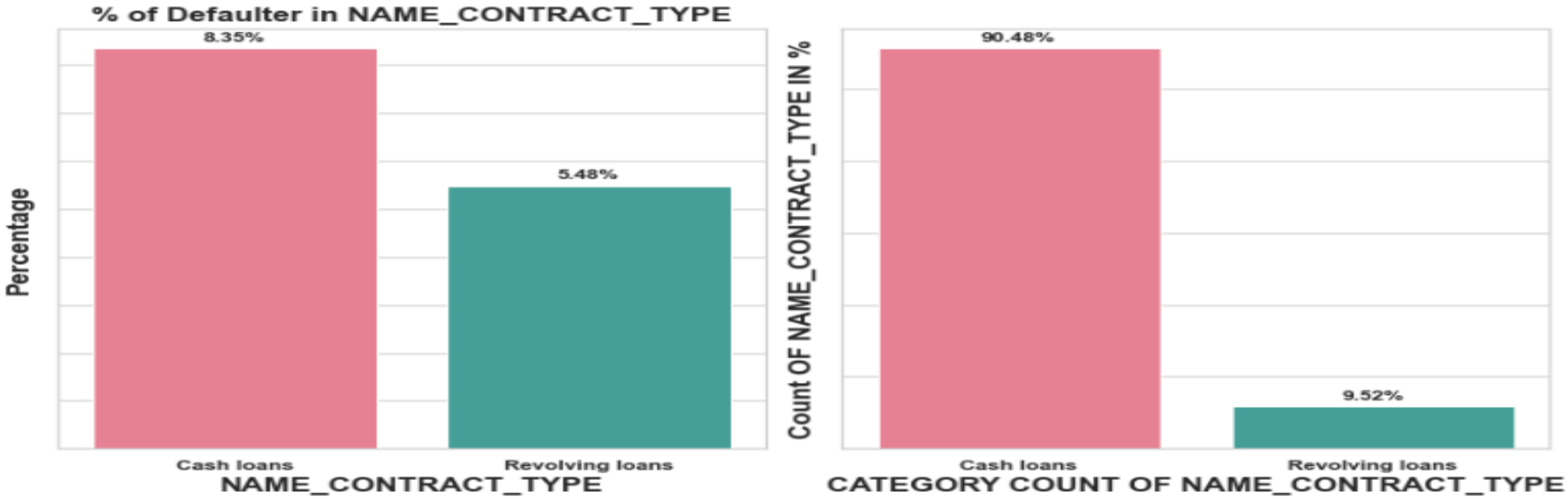
- **Amount Credit and Amount Annuity = 0.75**
- **Amount Goods Price and Amount Annuity = 0.75**
- **Amount Goods Price and Amount Credit = 0.98**

# Income, Amount Credit and Amount Annuity, Amount Goods Price




- **Other than Income, Amount Credit and Amount Annuity, Amount Goods Price has good correlation.**
- **Amount Annuity, and Amount Goods Price has strong positive relationship.**
- **The given scenario is same for re-payer and defaulter.**

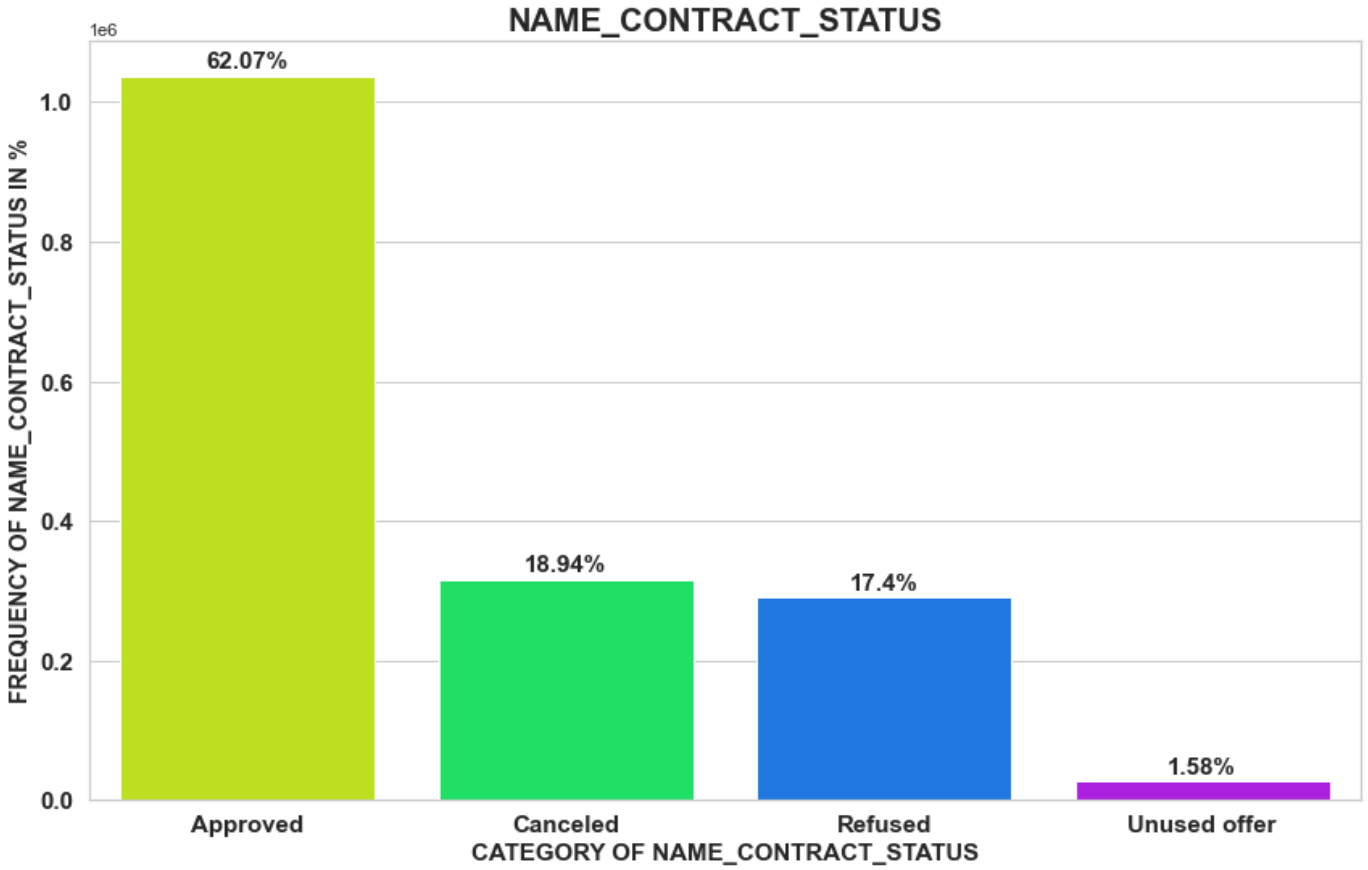
# Default Rate: Contract Type




**Insights**

- 90% loans are cash loans and only 10% laons are revoilving loans.
- As cash loans are high in number, thus probabilty of default is more with it as comapre to revolving loans.
- Cash loan default rate is 8.35% and revolving loans has default rate is 5.5%.

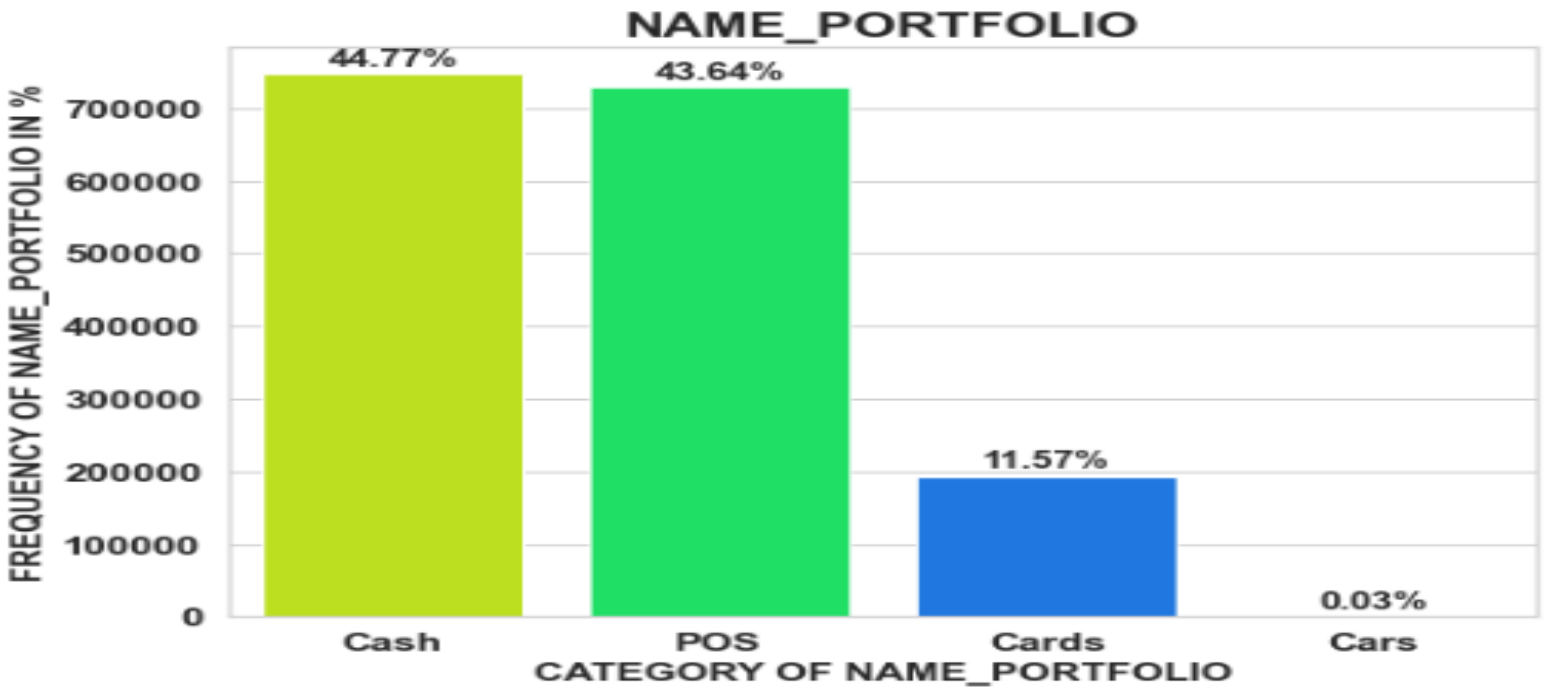
# Previous Applications: Loan Status




- About 62% loans were accepted and 17% applicants were refused to get loan.

- Around 19% loans were canceled by the applicants during the application process due to reason like interest rate, or they might have good option or they changed their mind, and 1% of applicants opt out to take loan

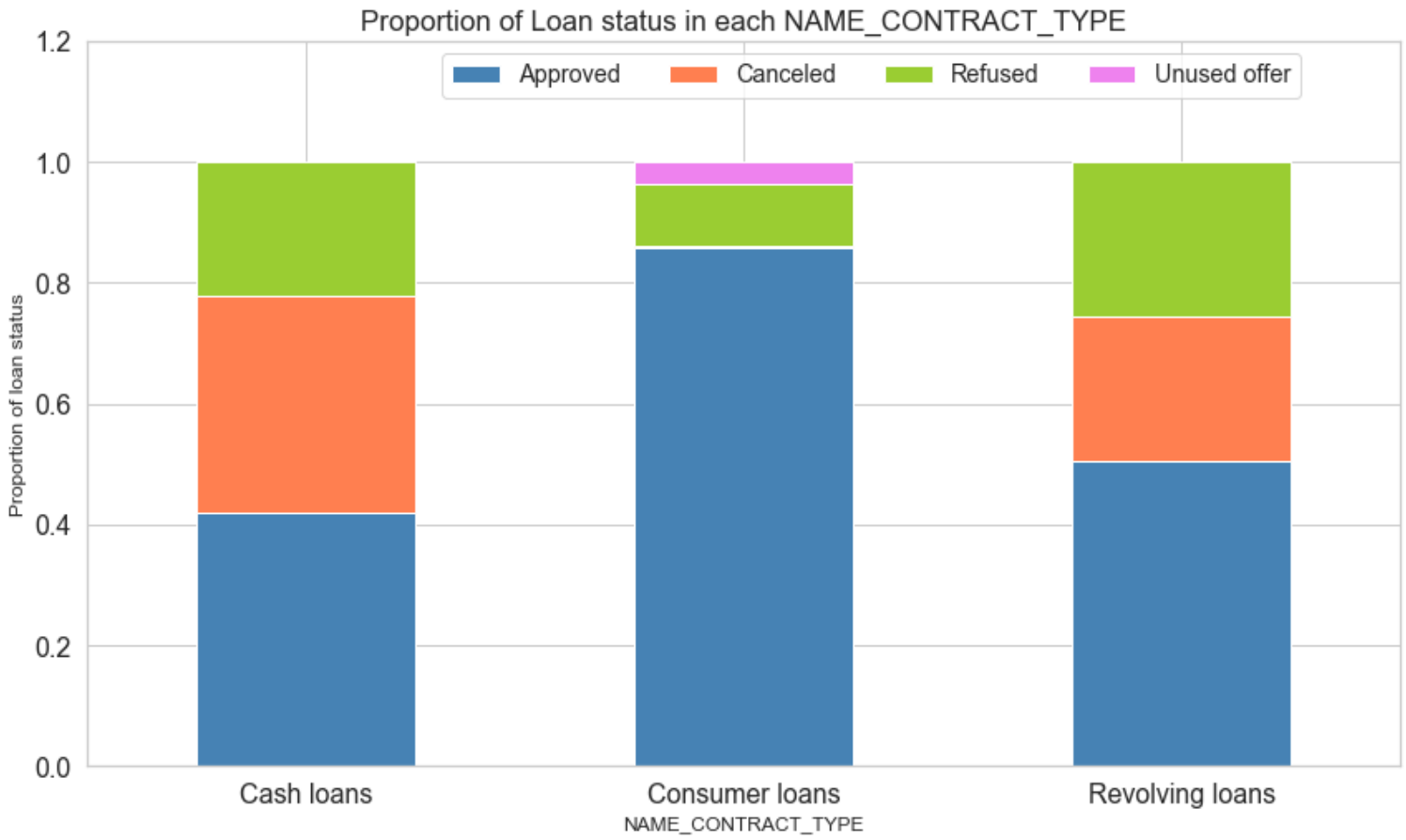
# Loan: Portfolio Type




**Insight**

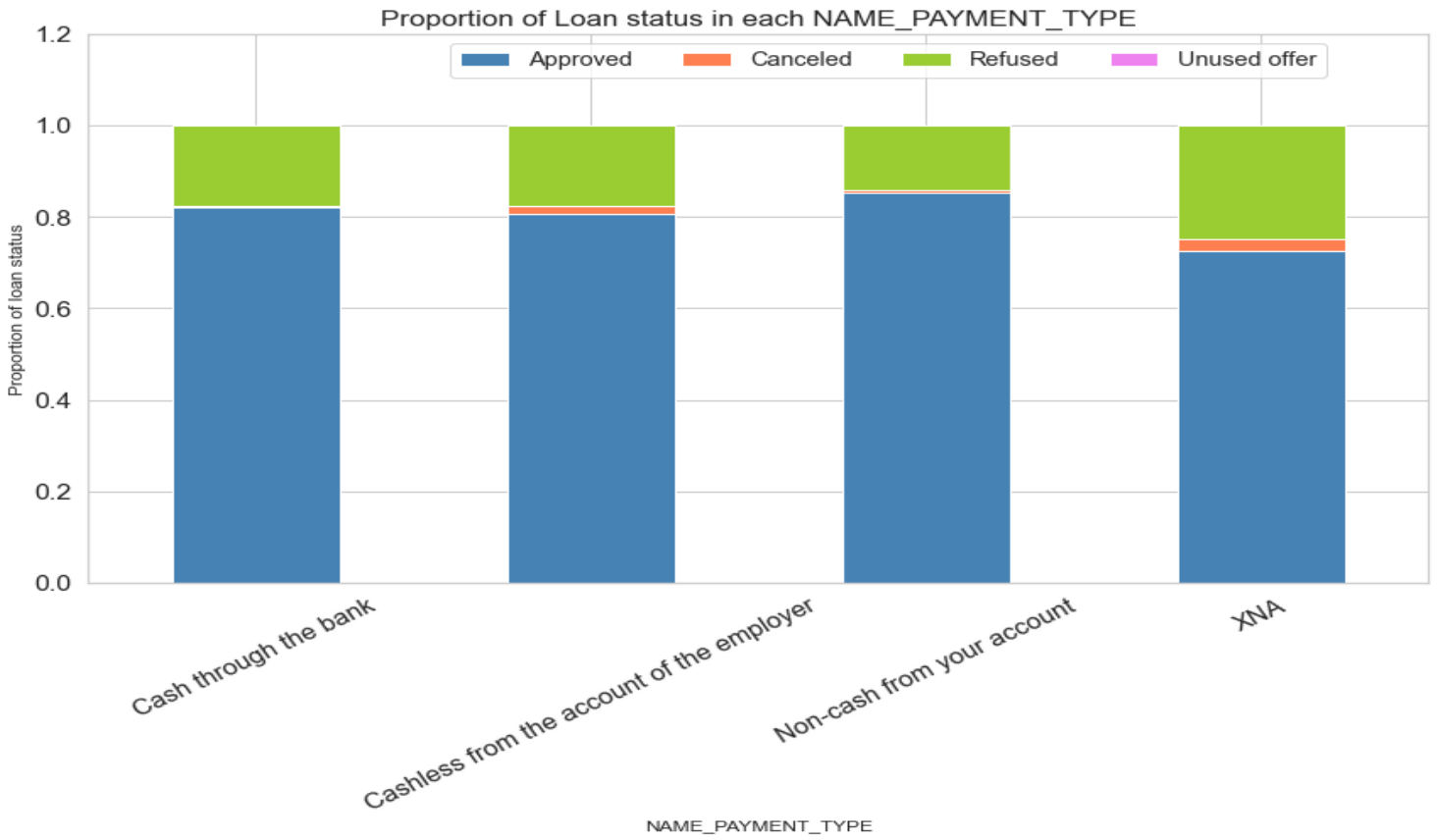
- Loan is mostly for cash and point of sale financing.
- About 12% for cards payment.
- Least for cars only 0.03%.

# Loan Status: Contract Type



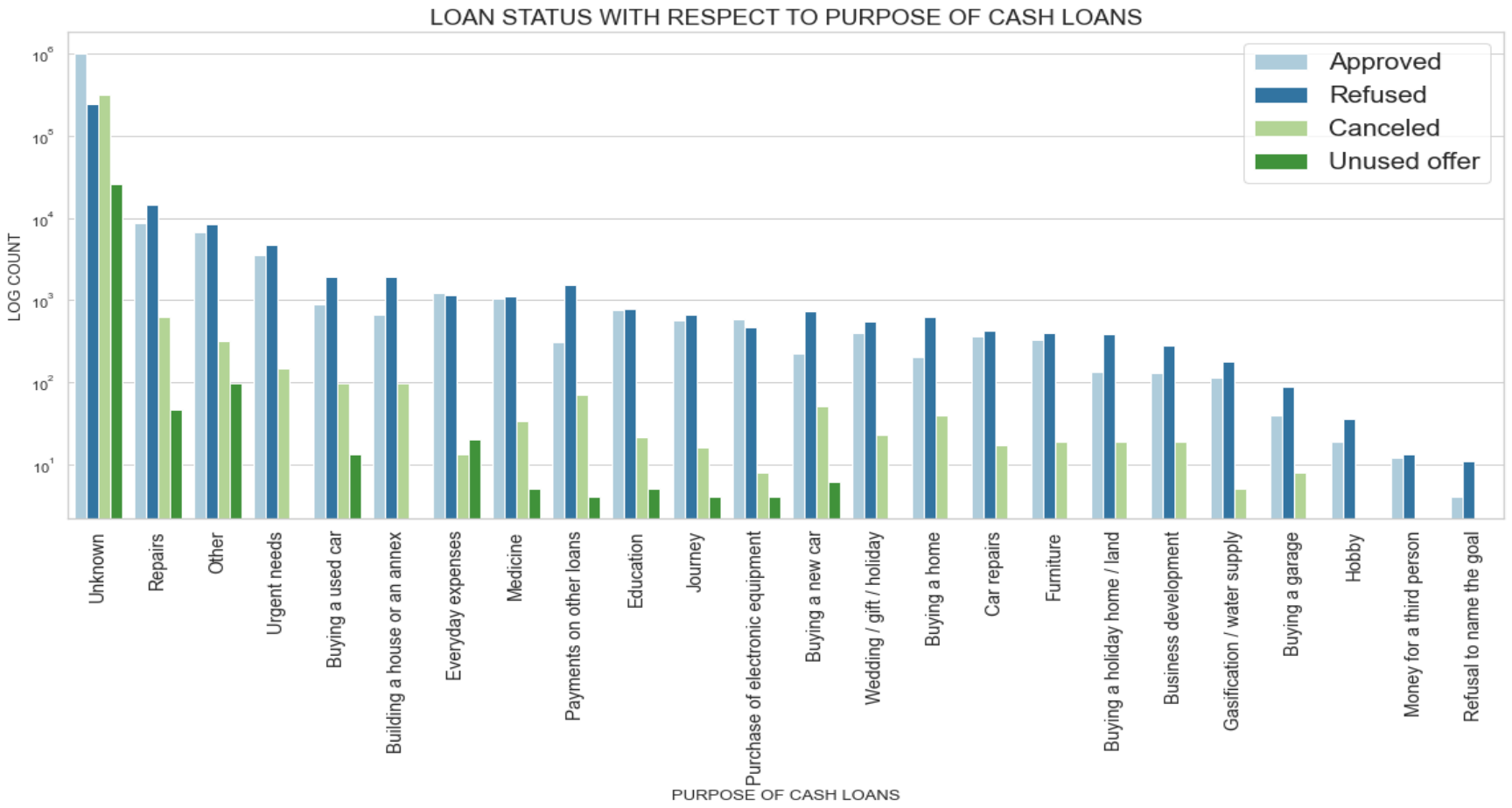

- Revolving Loans has the high refused rate of all contract types.
- 85% of Consumer loans are approved.
- 35% cash loans and 23% of revolving loans are cancelled by client.

# Loan Status: Payment Type



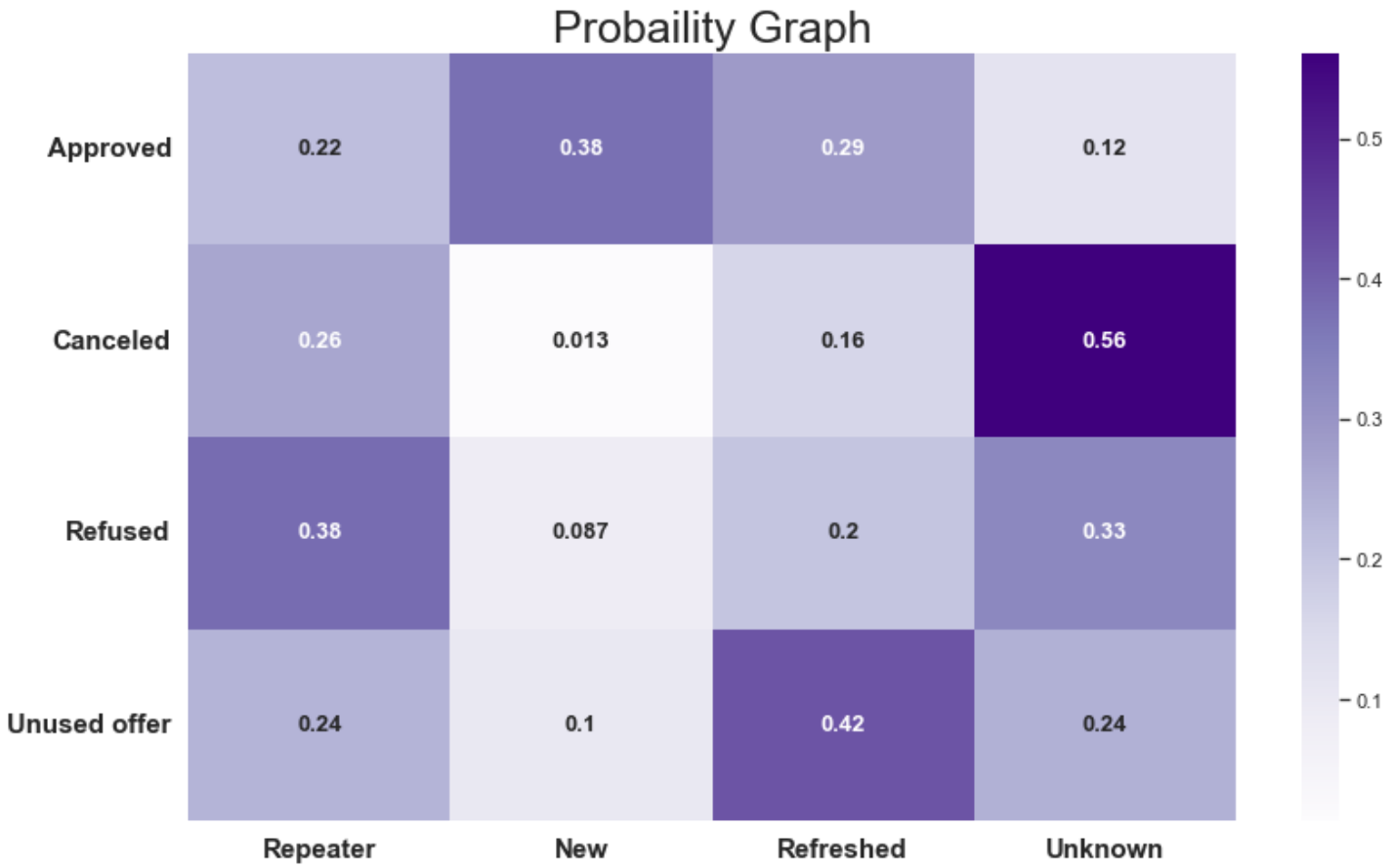

- It is clearly evident that for unknown payment are the ones where most loans are cancelled.
- Cashless payment from account has respectively more chance of loan getting default.
- Though the cash via bank and cashless from employees has the maximum refusal rate comparing to others two payment type.

# Loan Status: Purpose of cash Loans



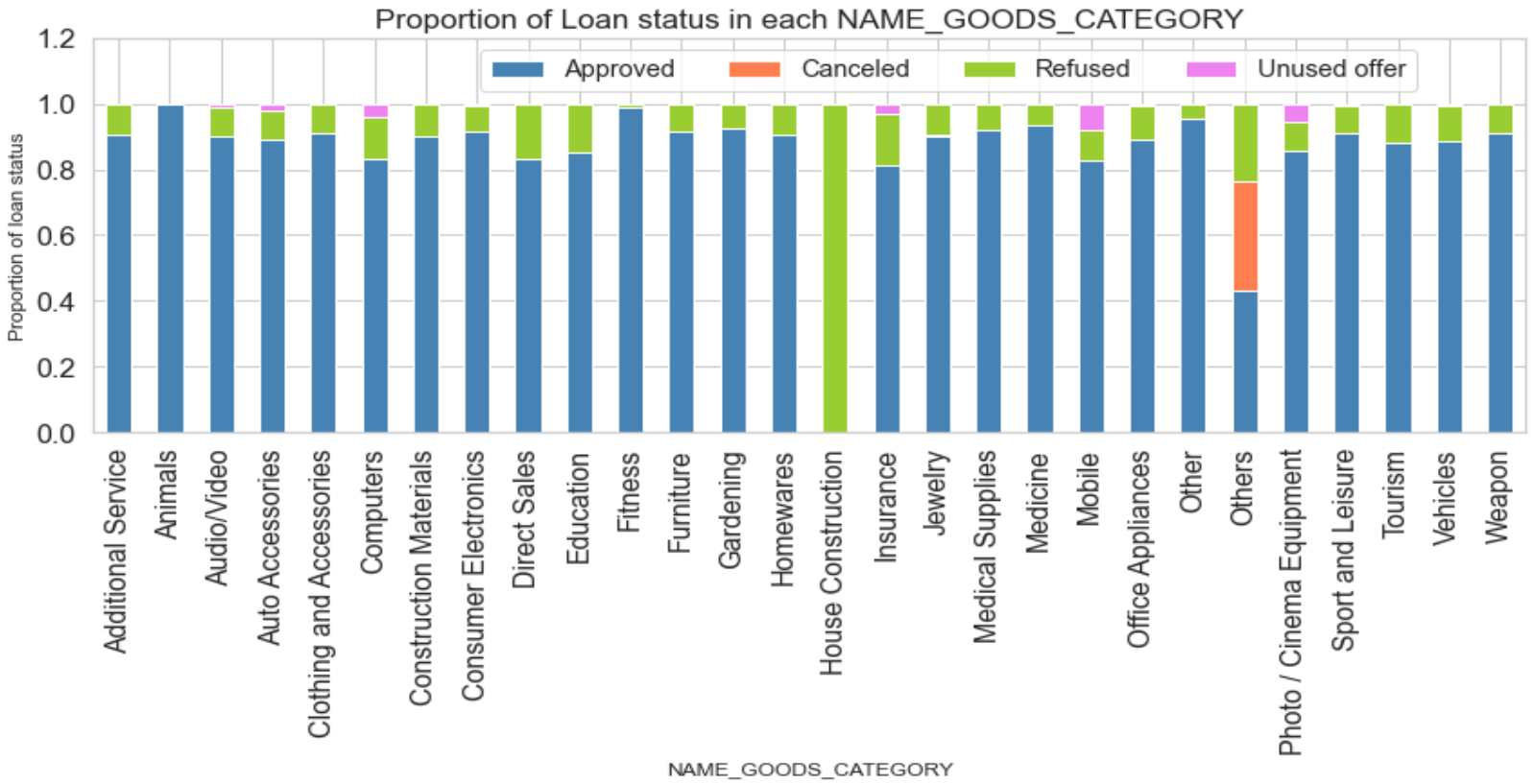

- Unknown purpose of loans are approved and cancel the most.
- Payments for other loans are approved the least and refused the man.
- Loans for Everyday expenses are cancelled the least.

# Loan status: CLIENT TYPE



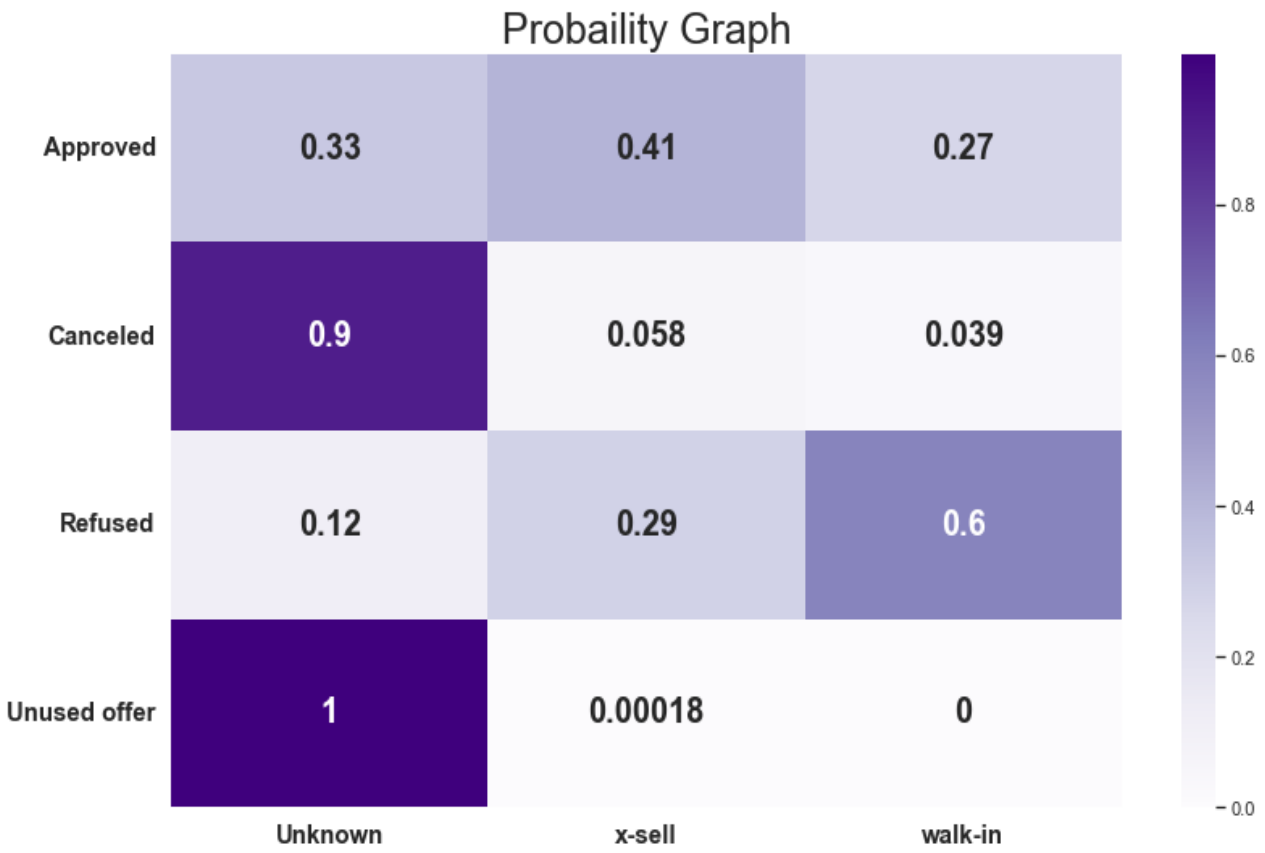

- Clients which are unknown has 56% chance of cancelling their loans and moreover their loan will get also refused more.
- Thus low approval rate.
- New Clients has the best probality in getting approval of loan.
- Clients which are repeater, their loans are more prone to get refused than any other customer.

# Loan Status: Goods Category



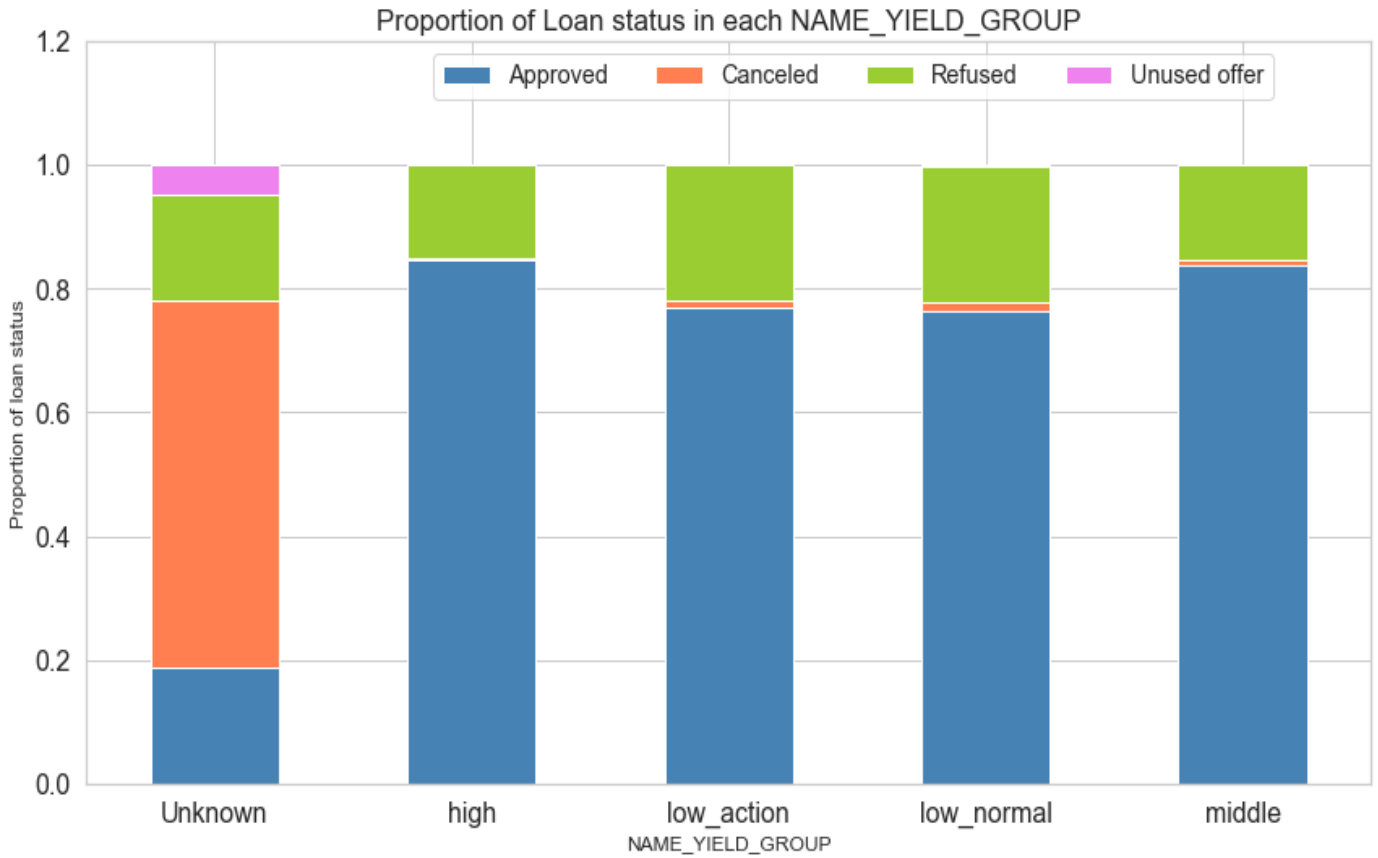

- Loans which are for house construction will not get approved.
- Loans which are taken for animals will get approved easily.
- Unused offers most in mobiles, cinema equipments and computers.
- All canceled loans are part of those clients whose Goods category are not known.

# Loan Status: Product Type



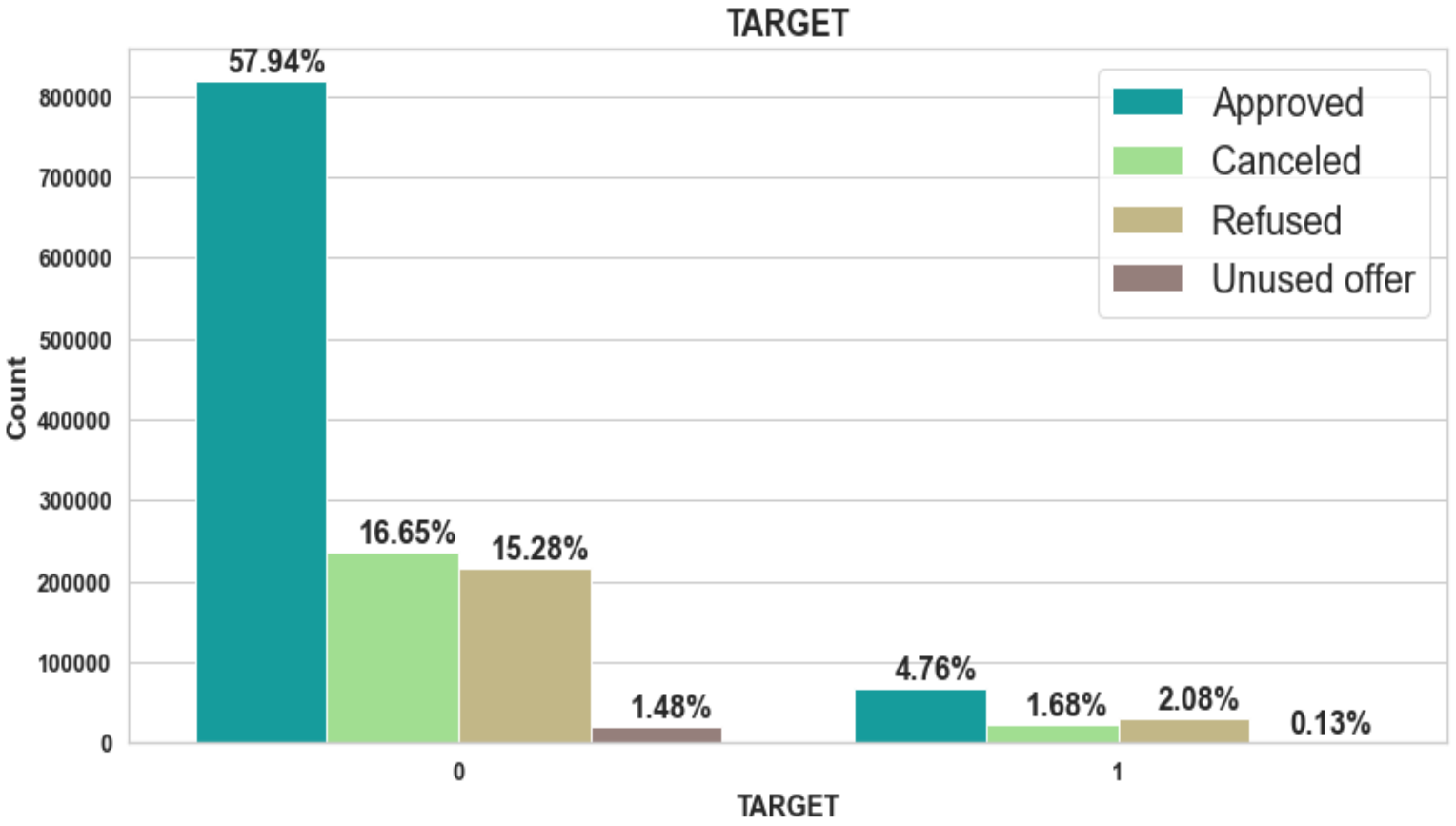

- Loans which are x_sell has better approval rate than walk-in.
- Loans are majorly refused if it is walk-in.
- Loans which are not known often gets canceled or approved and less refusal rate.

# Loan Status: Yield Group



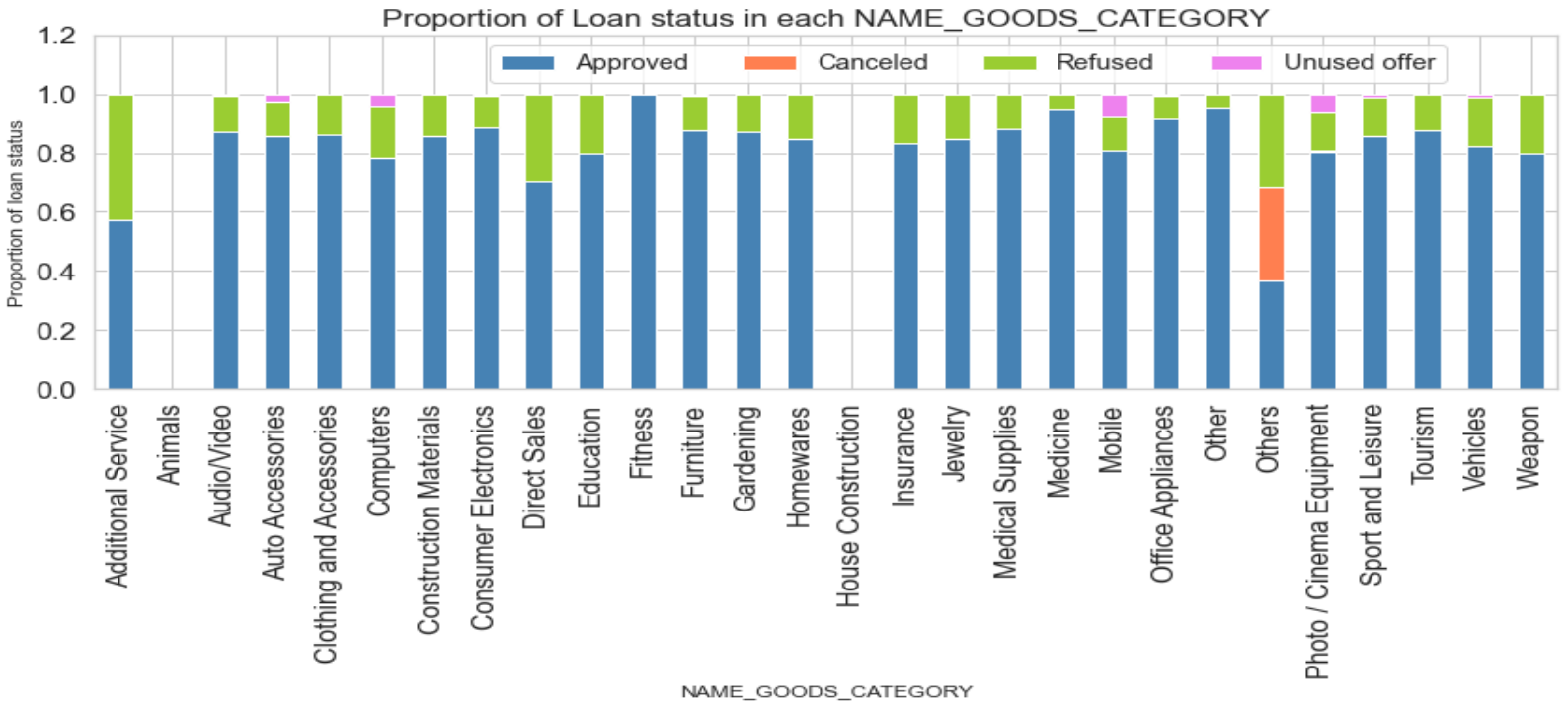

- Clients which are in middle and high yoeld group has better approval rate.
- Clients whose yield are not are mostly get canceled.
- Clients which are in low action and low normal yield group are more to refusal.

# Loan status vs Defaulter & Non Defaulters



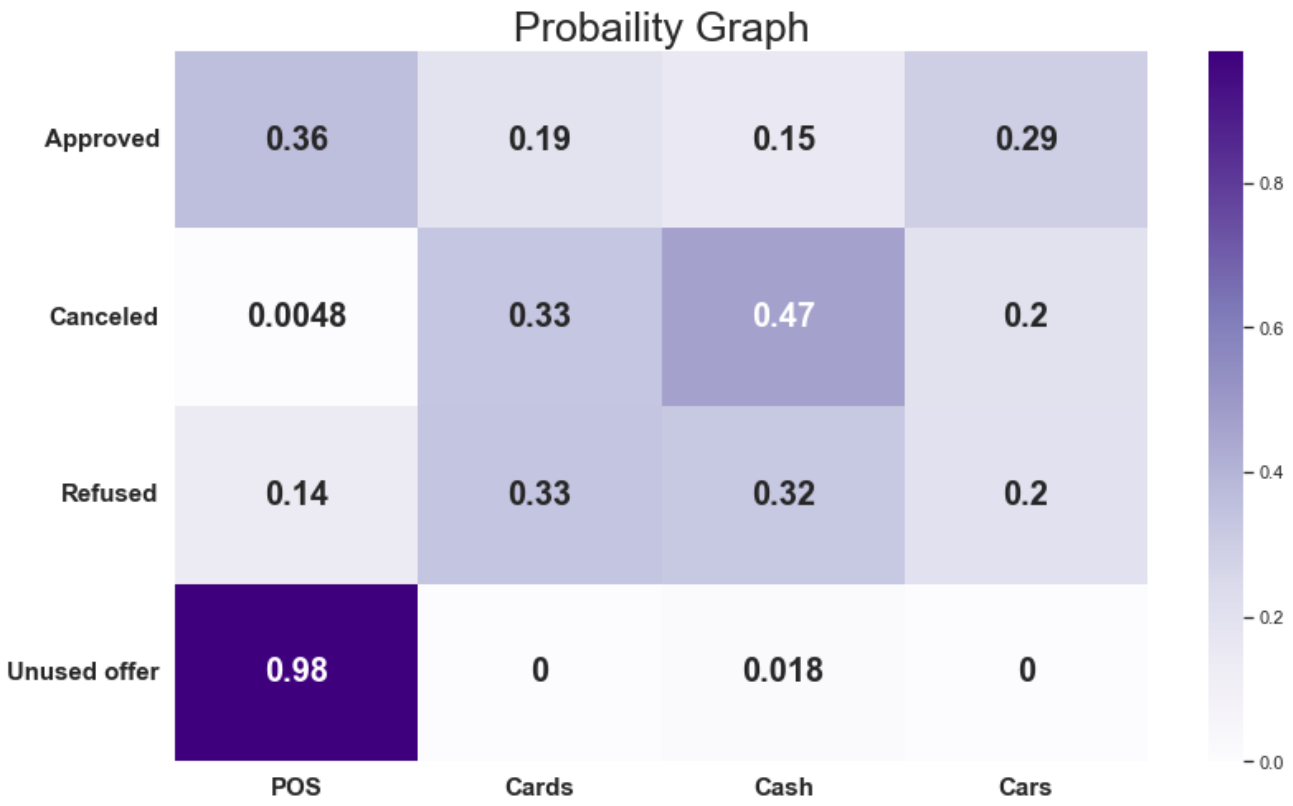

- Out of all the defaulters,
- 4.76% loans were approved.
- 1.68% loans were cancelled.
- 2.08% loans were refused.
- 0.13% loans were unused offer.

# Defaulter: Goods Category




- **Loans for fitness are all approved.**
- **Loans for medicines are less refused.**
- **Loans for addition services and direct sales are mostly refused.**

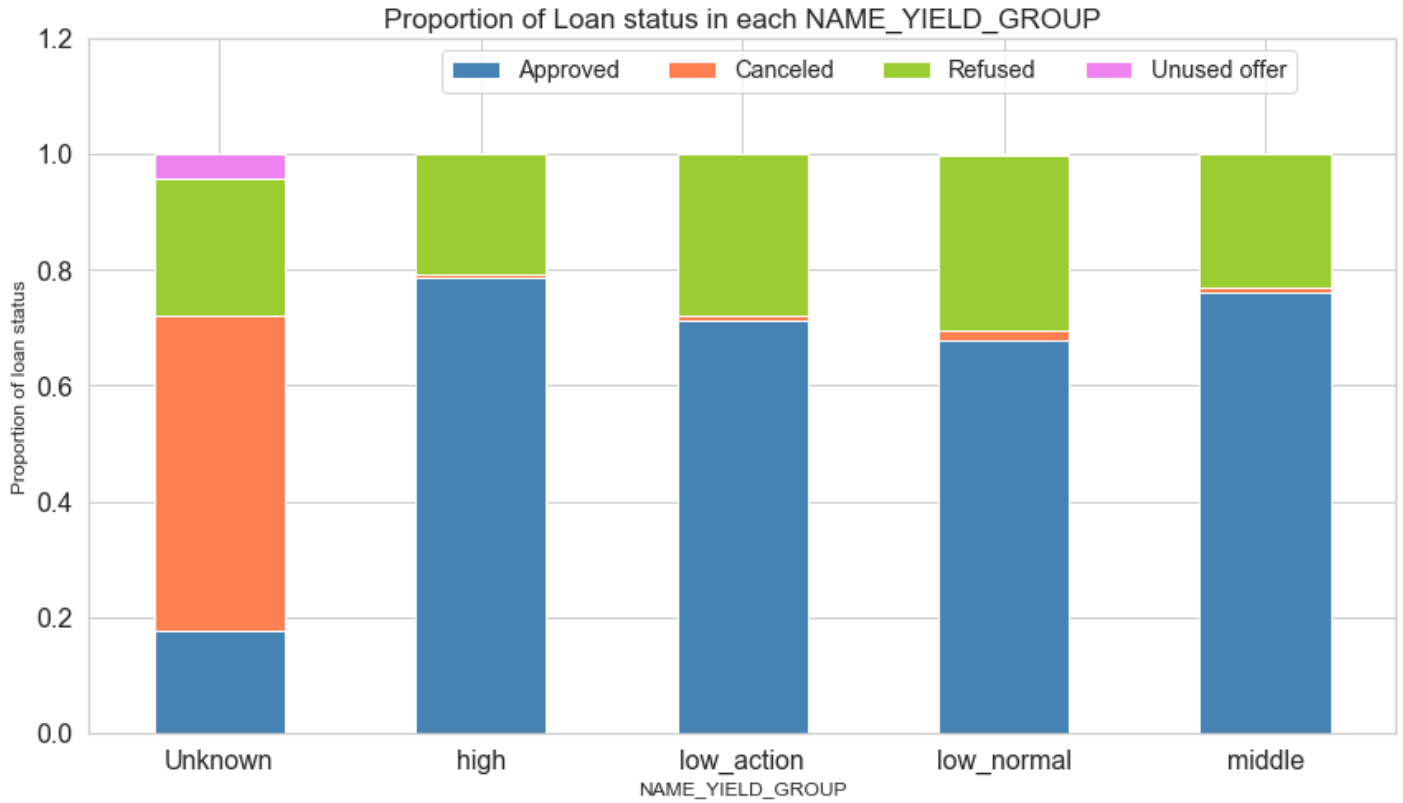
# DEFAULTER: PORTFOLIO TYPE




**Insights**

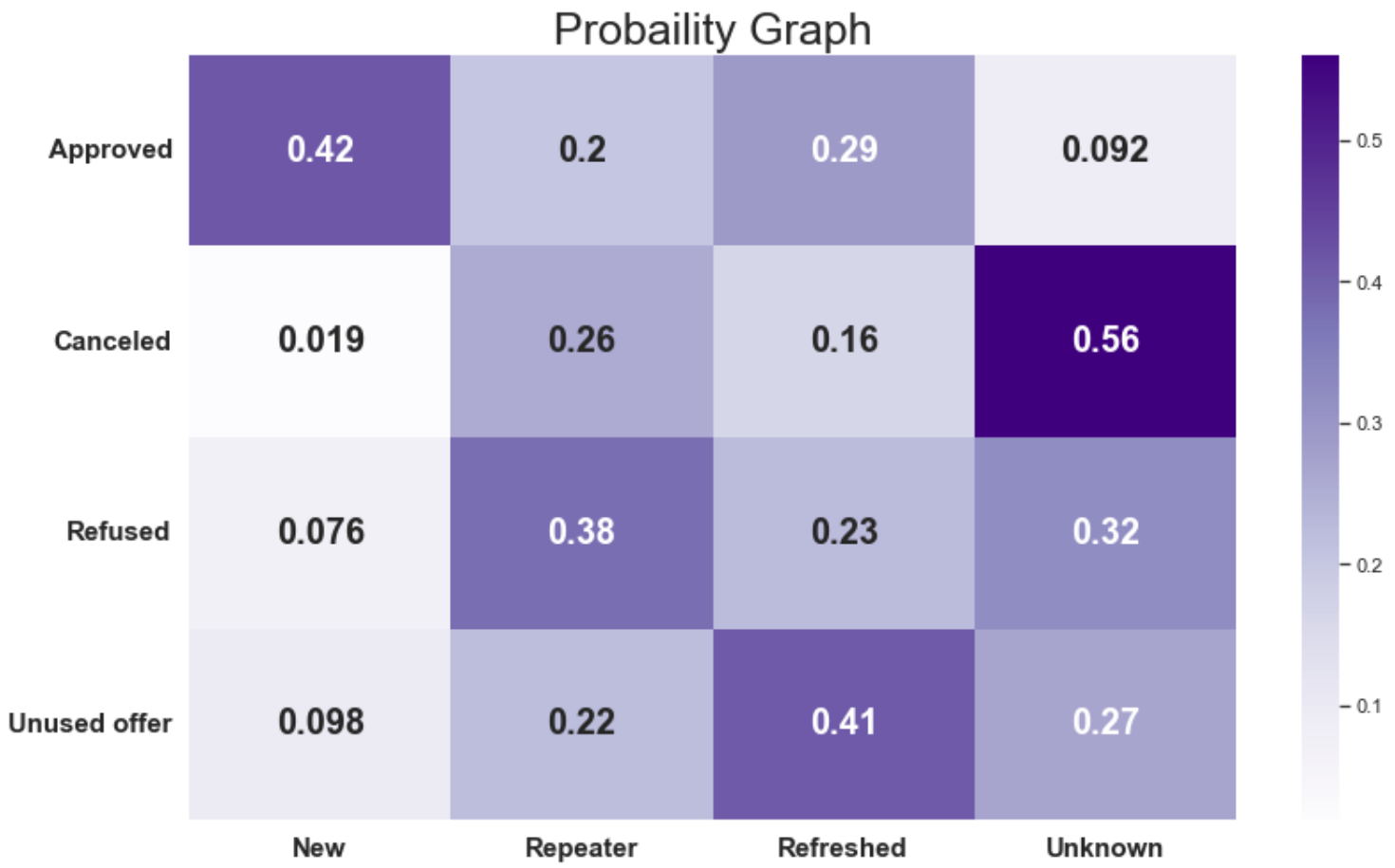
- **For defaulter if it is a cash loan then it has the least chances of getting approved.**
- **Pos still has the best chance from all other categories.**
- **Though loan for pos is refused the most.**

# Defaulter: Yield Group



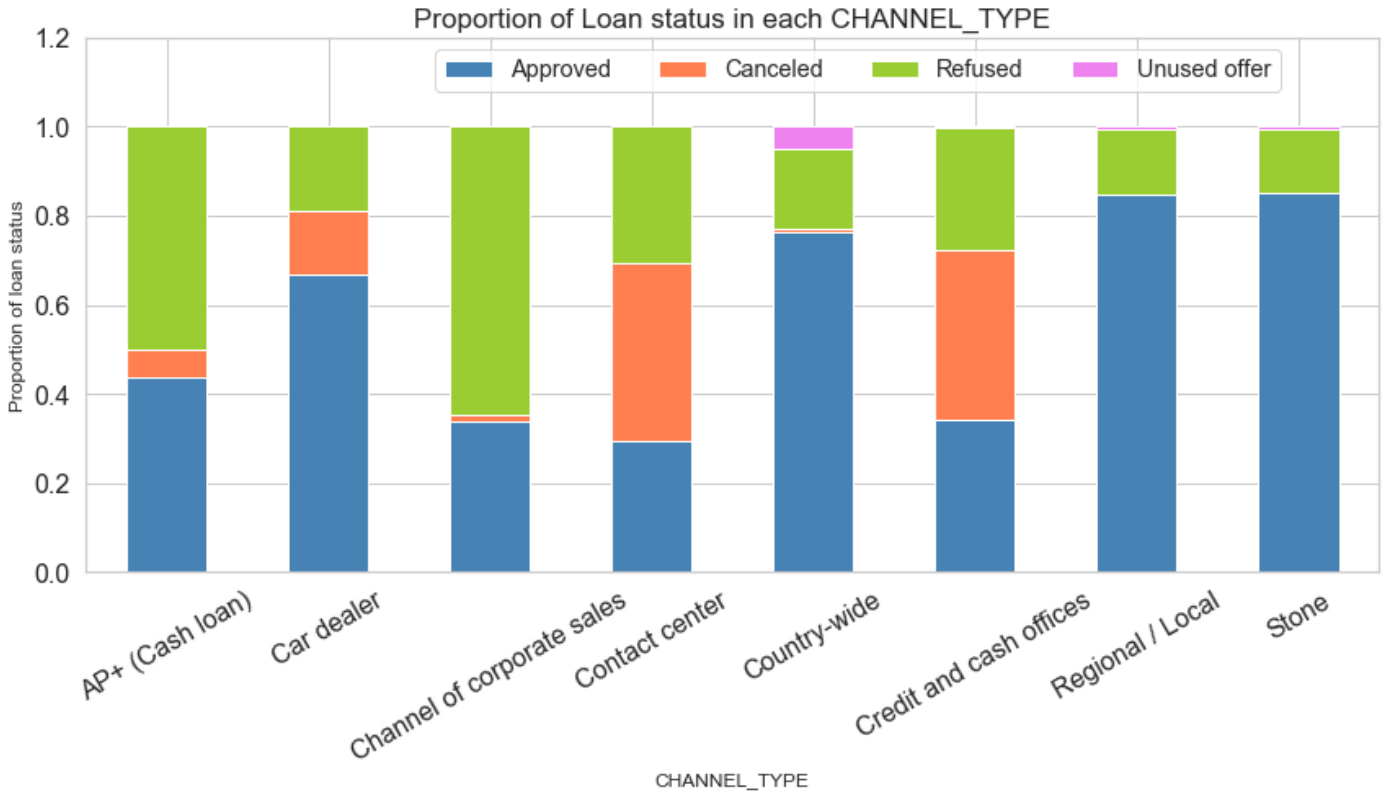

- **For defaulters chance of loans getting approved is good for high and middle yield groups.**

# Defaulter: Client Type



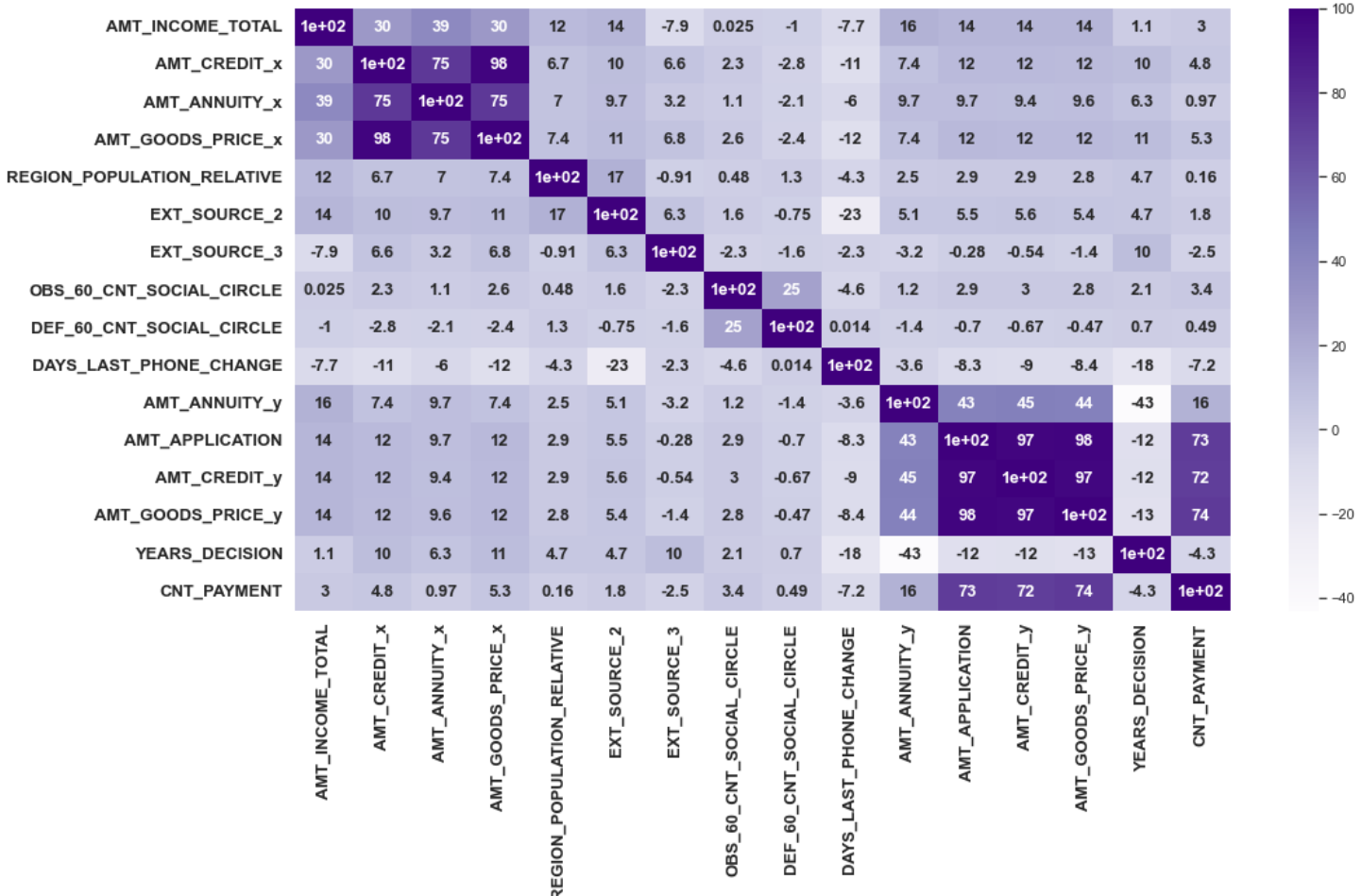

- **Repeater loans are refused the most.**
- **New client which categorized as default has good chance of getting approved from others.**

# Defaulter: Channel Type



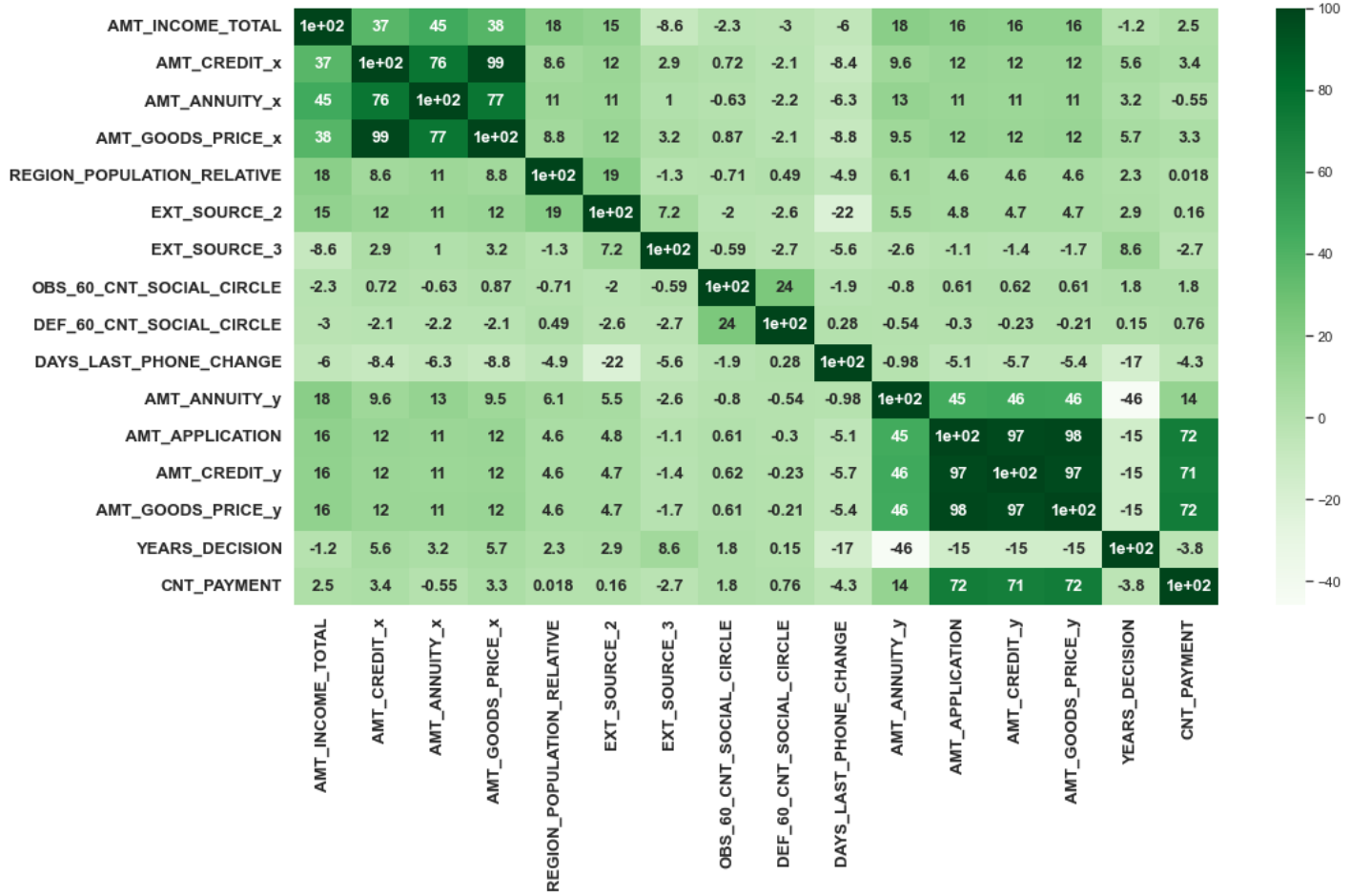

- **Client which is categorizes as defaulter, and if their loans get approved chances would be that client was acquired store channel or country wide channel.**
- **If channel was credit and cash offices than the chances are least.**

# Defaulter: Top 10 Correlated variables




| | level_0 | level_1 | 0 |
|---|---|---|---|
| 0 | AMT_CREDIT_x | AMT_GOODS_PRICE_x | 0.982538 |
| 1 | AMT_APPLICATION | AMT_GOODS_PRICE_y | 0.978394 |
| 2 | AMT_APPLICATION | AMT_CREDIT_y | 0.974891 |
| 3 | AMT_GOODS_PRICE_y | AMT_CREDIT_y | 0.966622 |
| 4 | AMT_ANNUITY_x | AMT_CREDIT_x | 0.746966 |
| 5 | AMT_ANNUITY_x | AMT_GOODS_PRICE_x | 0.746087 |
| 6 | AMT_GOODS_PRICE_y | CNT_PAYMENT | 0.740928 |
| 7 | CNT_PAYMENT | AMT_APPLICATION | 0.727391 |
| 8 | AMT_CREDIT_y | CNT_PAYMENT | 0.723440 |
| 9 | AMT_CREDIT_y | AMT_ANNUITY_y | 0.446033 |

# Non-Defaulter: Top 10 Correlated Variable




| | | | |
|---|---|---|---|
| 0 | AMT_GOODS_PRICE_x | AMT_CREDIT_x | 0.986412 |
| 1 | AMT_APPLICATION | AMT_GOODS_PRICE_y | 0.984025 |
| 2 | AMT_APPLICATION | AMT_CREDIT_y | 0.974952 |
| 3 | AMT_CREDIT_y | AMT_GOODS_PRICE_y | 0.970911 |
| 4 | AMT_GOODS_PRICE_x | AMT_ANNUITY_x | 0.765996 |
| 5 | AMT_ANNUITY_x | AMT_CREDIT_x | 0.762273 |
| 6 | CNT_PAYMENT | AMT_GOODS_PRICE_y | 0.723636 |
| 7 | CNT_PAYMENT | AMT_APPLICATION | 0.715561 |
| 8 | CNT_PAYMENT | AMT_CREDIT_y | 0.708940 |
| 9 | AMT_ANNUITY_y | AMT_CREDIT_y | 0.462282 |

# Decisive Factors: Defaulter

- Men are at relatively higher default rate
- Clients with Lower Secondary & Secondary education
- Clients who are either at Maternity leave OR Unemployed default a lot.
- Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as the default rate is huge.
- Avoid young people who are in age group of 20-40 as they have higher probability of defaulting
- Client who have less than 5 years of employment have high default rate.
- Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.

# Decisive Factors: Re-Payer

- Academic degree has less defaults.
- Student and Businessmen have no defaults.
- Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%
- People above age of 50 have low probability of defaulting
- Clients with 40+ year experience having less than 1% default rate
- Loans bought for Hobby, Buying garage are being repayed mostly.
- People with zero to two children tend to repay the loans.

# Thank You!