# CASE STUDY – EXPLORATORY DATA ANALYSIS ON CREDIT ASSESSMENT

Submitted By:
Name: Pavithra Sri S
Date : 27.10.2024

# PROBLEM STATEMENT AND ITS OBJECTIVE

When the consumer finance 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. Risk of Non-Approval for Creditworthy Applicants
2. Risk of Approving Loans to Potential Defaulters.

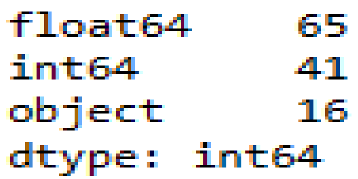
## Objective:

Aim is to identify the factors and patterns within the data that differentiates those who likely to repay the loan and those who become defaults by **Exploratory data analysis (EDA).** This will bring out the insights hiding behind the data with which lending companies can take meaningful decisions in future.

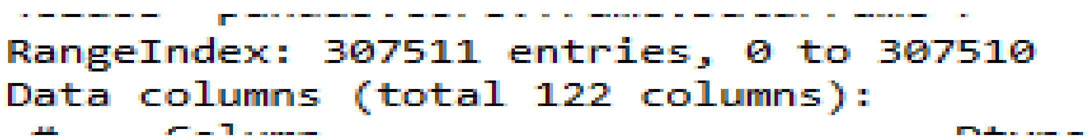
# DATASET UNDERSTANDING

## Application Dataset

- It has 307511 rows and 122 rows
- It has 3 types of data - float, int and object.
- The columns has null values, days are not in correct format.



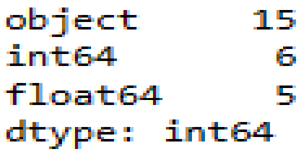
```
float64     65
int64       41
object      16
dtype: int64
```



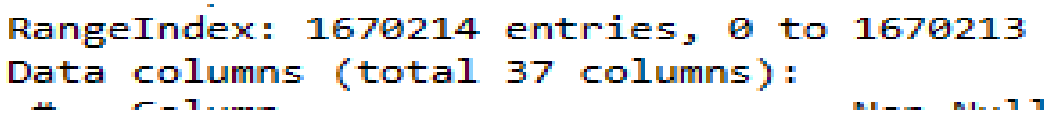
```
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 122 columns):
```

## Previous_application dataset

- There are 1670214 rows and 37 columns.
- 16 Categorical and 21 numerical columns are there.
- There are 3 data types : object, int and float.
- The columns has null values, days are not in correct format.



```
object      15
int64        6
float64      5
dtype: int64
```

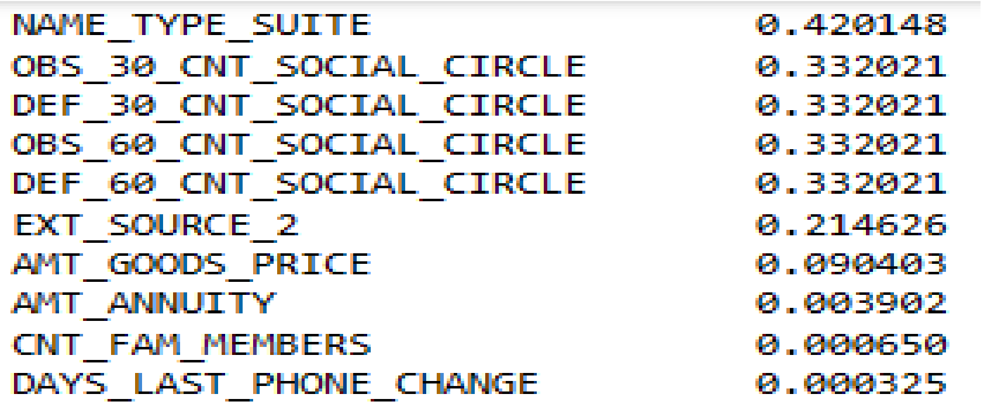


```
RangeIndex: 1670214 entries, 0 to 1670213
Data columns (total 37 columns):
```

# Data Cleaning and Manipulation

## Null Value Handling:

- Checking and identifying null values in the dataset.
- Removed the columns with a null percentage greater than 50%  (41 columns) and suggested imputation methods for those columns with null percentage less than 13%.



```
NAME_TYPE_SUITE                 0.420148
OBS_30_CNT_SOCIAL_CIRCLE        0.332021
DEF_30_CNT_SOCIAL_CIRCLE        0.332021
OBS_60_CNT_SOCIAL_CIRCLE        0.332021
DEF_60_CNT_SOCIAL_CIRCLE        0.332021
EXT_SOURCE_2                    0.214626
AMT_GOODS_PRICE                 0.090403
AMT_ANNUITY                     0.003902
CNT_FAM_MEMBERS                 0.000650
DAYS_LAST_PHONE_CHANGE          0.000325
```

## Data Type Anomalies:

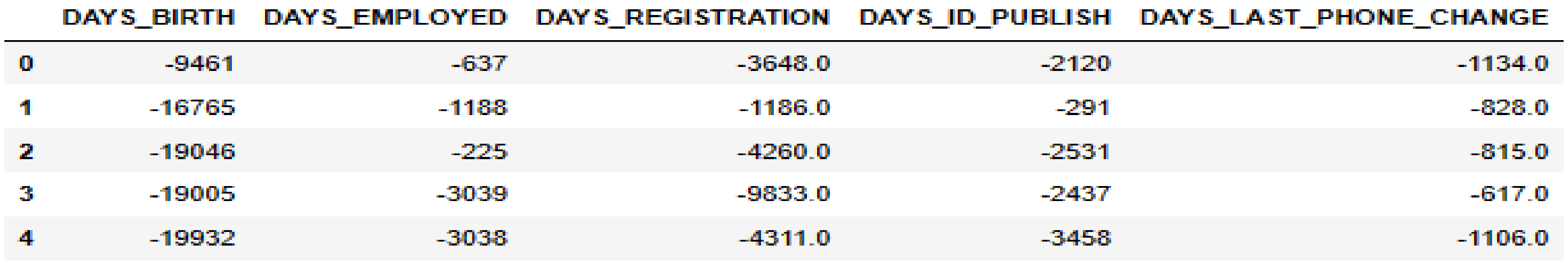
- Examining the data types of each column for inconsistencies and correcting if any for accurate analysis.

## Outlier Detection and Treatment:

- Identifying outliers in the dataset.
- Utilized statistical methods like IQR (Interquartile Range), box plot to detect outliers.
- Suggested methods to treat outliers either by removing them or by methods like imputation, depending on the nature of the data.

## Standardizing the data:

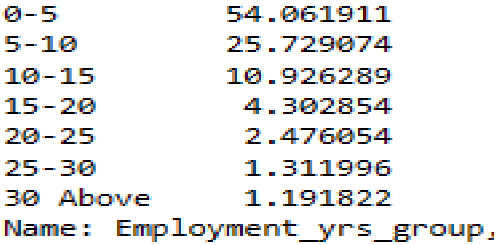
- Standardizing the negative values with abs() function for the below columns.



| | DAYS_BIRTH | DAYS_EMPLOYED | DAYS_REGISTRATION | DAYS_ID_PUBLISH | DAYS_LAST_PHONE_CHANGE |
|---|---|---|---|---|---|
| 0 | -9461 | -637 | -3648.0 | -2120 | -1134.0 |
| 1 | -16765 | -1188 | -1186.0 | -291 | -828.0 |
| 2 | -19046 | -225 | -4260.0 | -2531 | -815.0 |
| 3 | -19005 | -3039 | -9833.0 | -2437 | -617.0 |
| 4 | -19932 | -3038 | -4311.0 | -3458 | -1106.0 |

- Binning has been done for the below columns

DAYS_BIRTH, DAYS_EMPLOYED, AMT_INCOME_TOTAL.

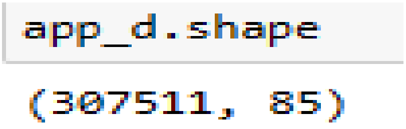


```
0-5          54.061911
5-10         25.729074
10-15        10.926289
15-20         4.302854
20-25         2.476054
25-30         1.311996
30 Above      1.191822
Name: Employment_yrs_group,
```

# Readiness of the dataset after cleaning and Manipulation

### Application Dataset

- It has 307511 rows and 85 rows
- It has 3 types of data - float, int and object.
- The columns with null percentage >50% has been removed and binning has been done for 2 columns for better analysis .
- The columns with negative values has been converted into positive values.
- Proper imputation methods for treating null values and outliers has been suggested.

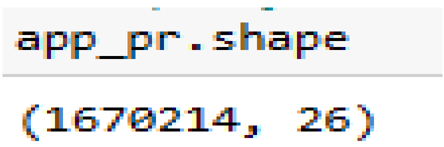


```
app_d.shape

(307511, 85)
```

### Previous_application dataset

- There are 1670214 rows and 26 columns.
- 15 Categorical and  11 numerical columns are there.
- There are 3 data types : object, int and float.
- The columns with negative values has been converted into positive values.
- Proper imputation methods for treating null values and outliers has been suggested.



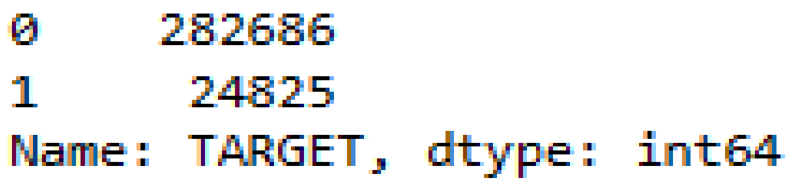
```
app_pr.shape

(1670214, 26)
```

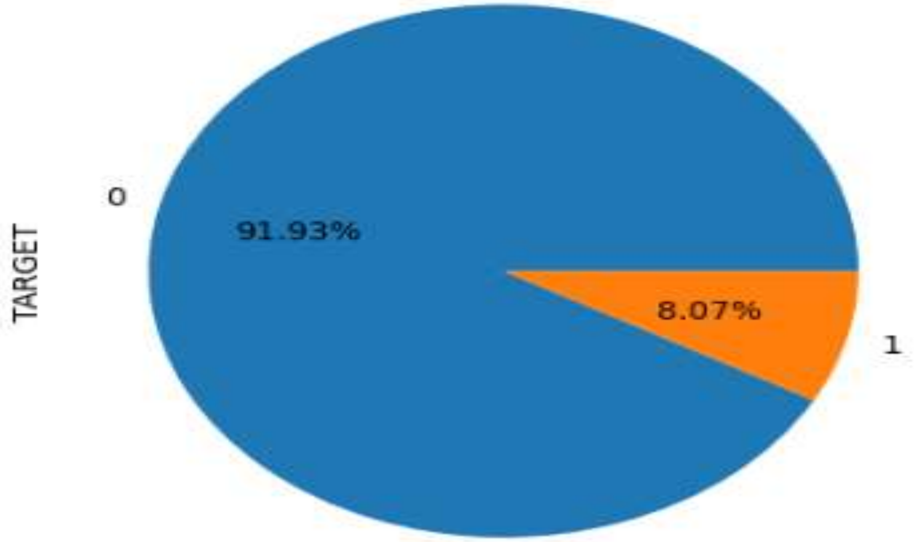
# Imbalance of the target variable

**Target variable:**

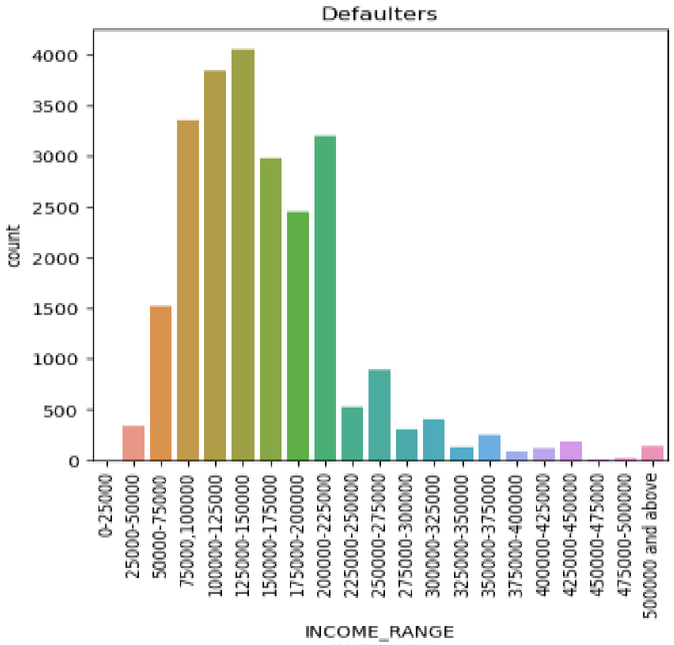
- 1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample
- 0 - all other cases

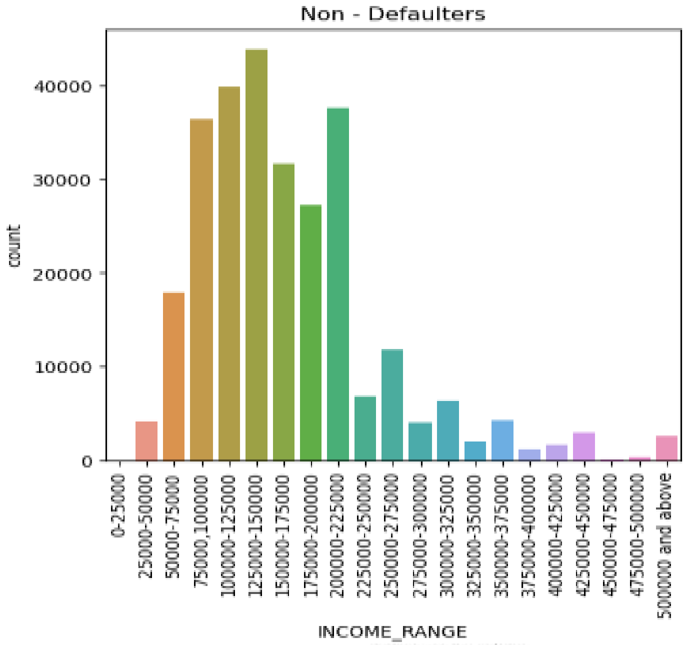


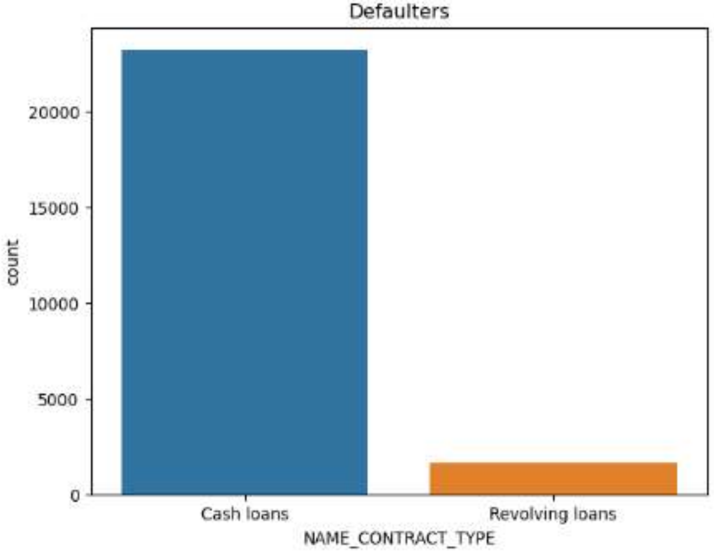
```
0    282686
1     24825
Name: TARGET, dtype: int64
```

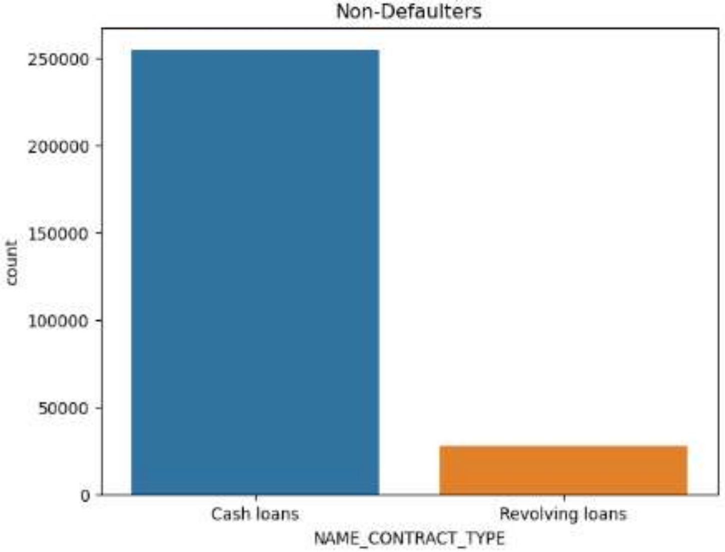


# Univariate analysis on categorical columns



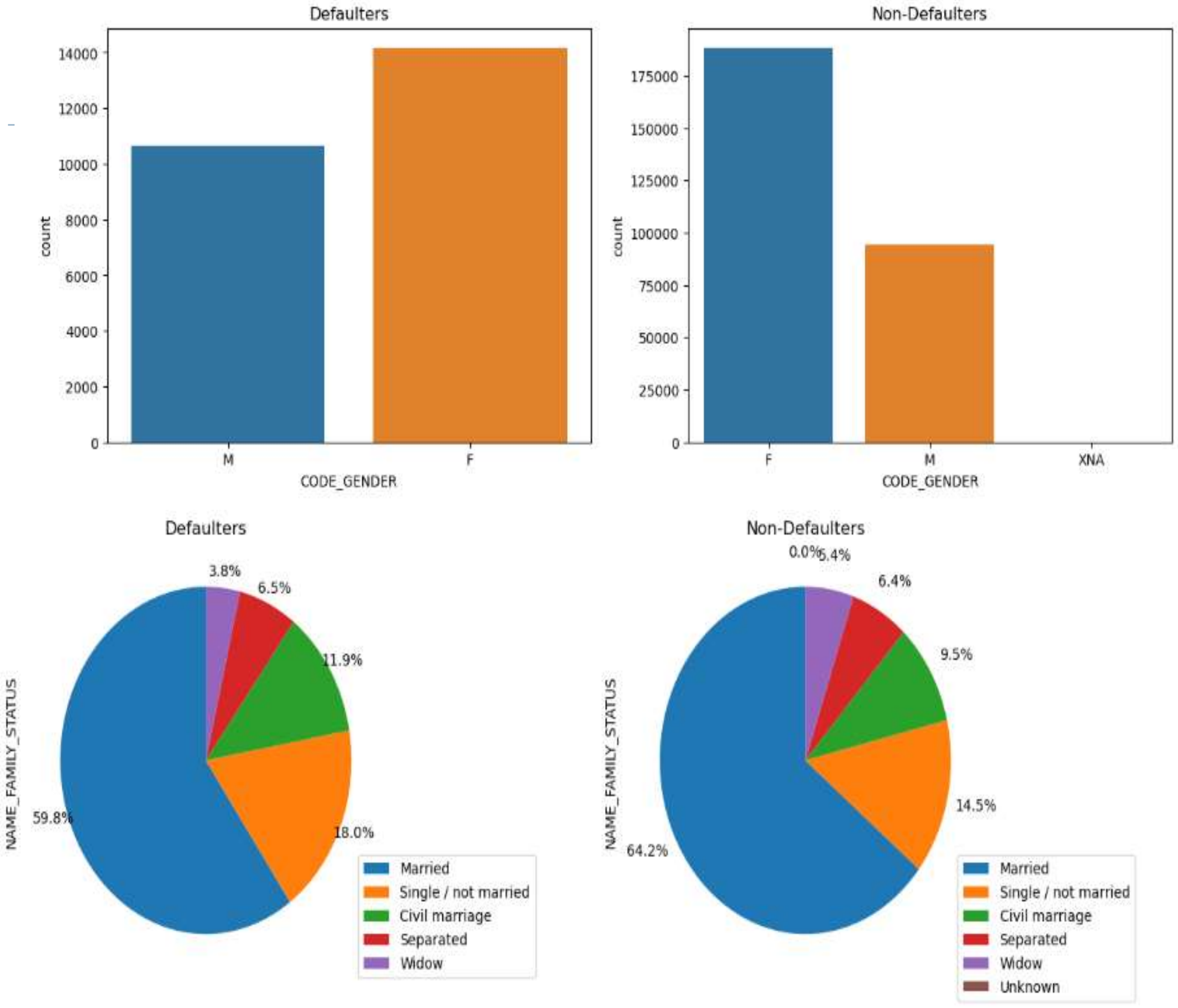






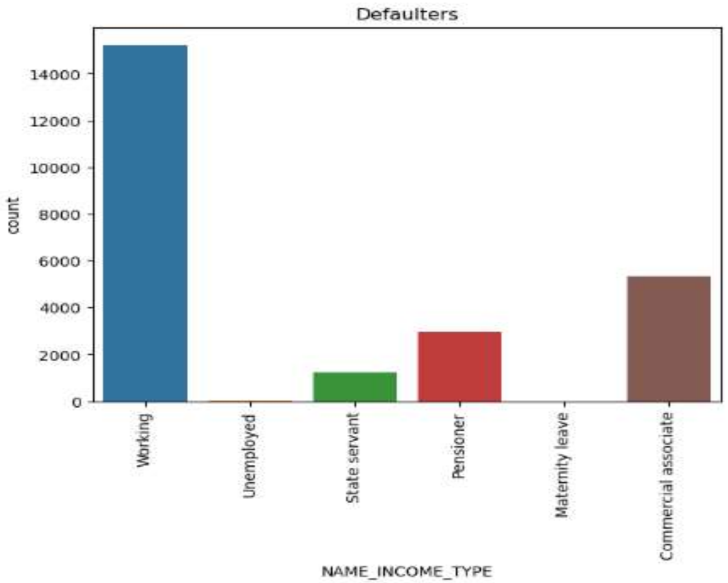
- The analysis suggests that individuals with an income range of 100,000 to 150,000 constitute a substantial portion in both defaulter and non-defaulter groups

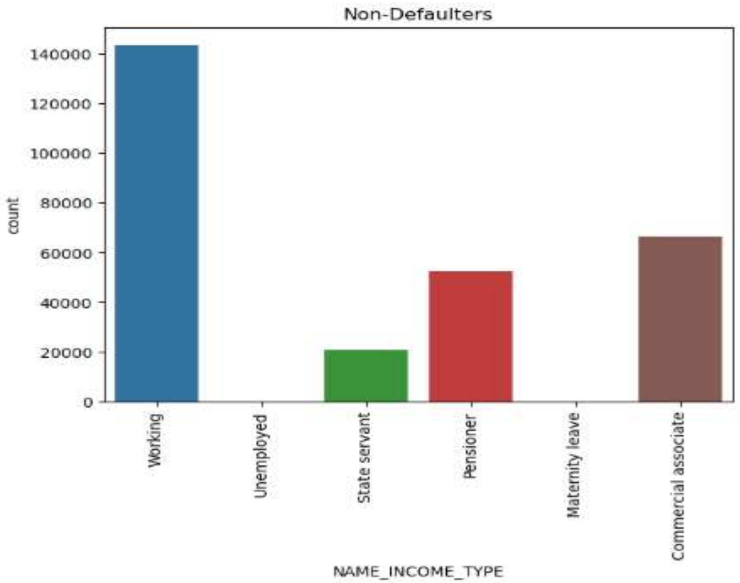
- Notably, in the cash loans category, the proportion of defaulters is significantly lower than non-defaulters, indicating a relatively lower risk associated with this loan type. Conversely, while the number of defaulters in revolving loans is lower, the proportion of defaulters to non-defaulters is comparatively higher.

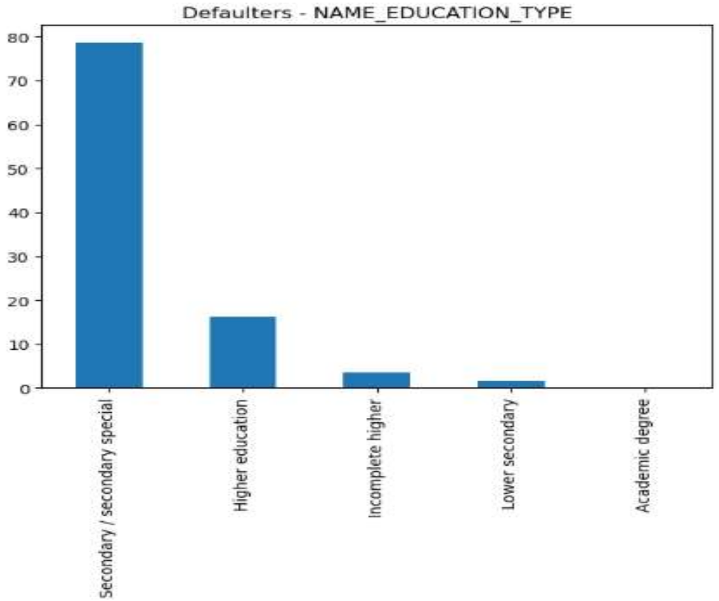


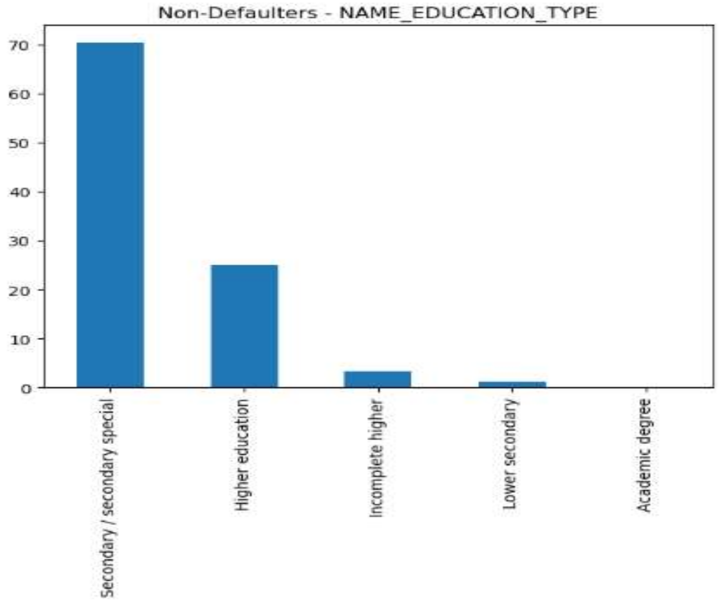
- From the count plot, we can see that MALE gender is contributing to more number of defaults than FEMALE.

- People who is married or widow are more likely to repay the loan comparatively. But even though there is no such a strong correlation. 2. On the other hand, Giving loans to people who is not married or done civil marriage or seperated is little riskier.



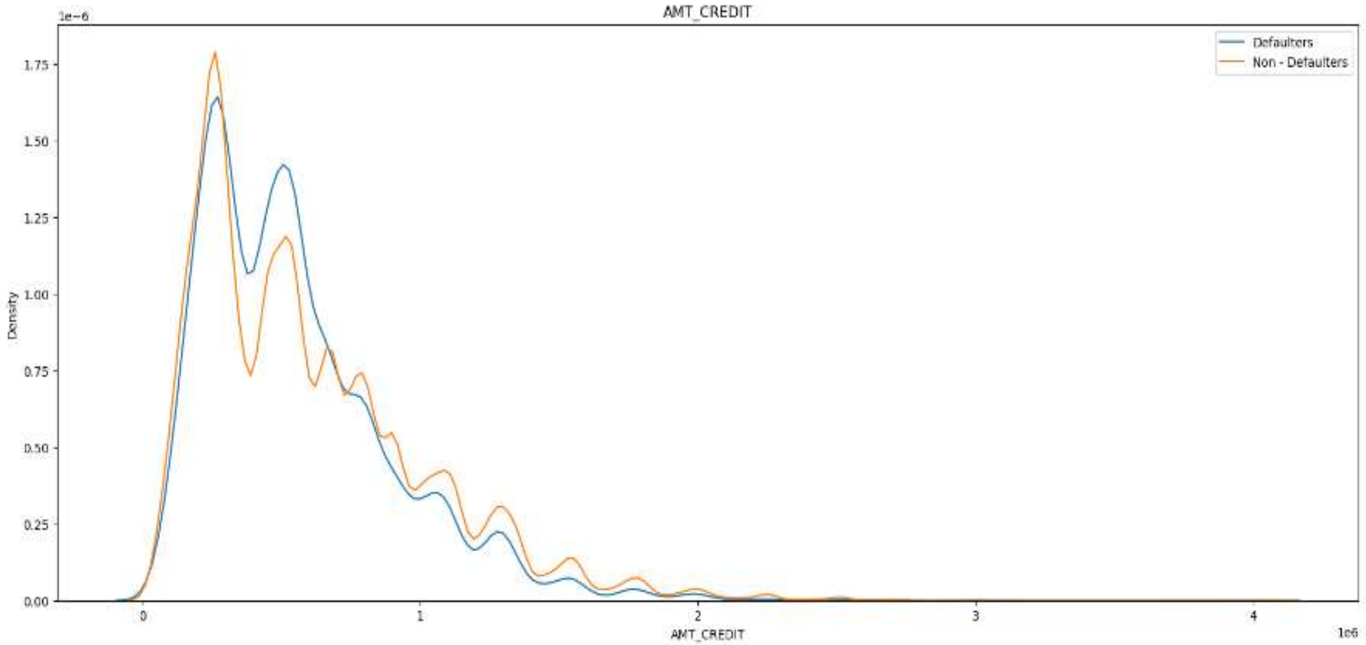


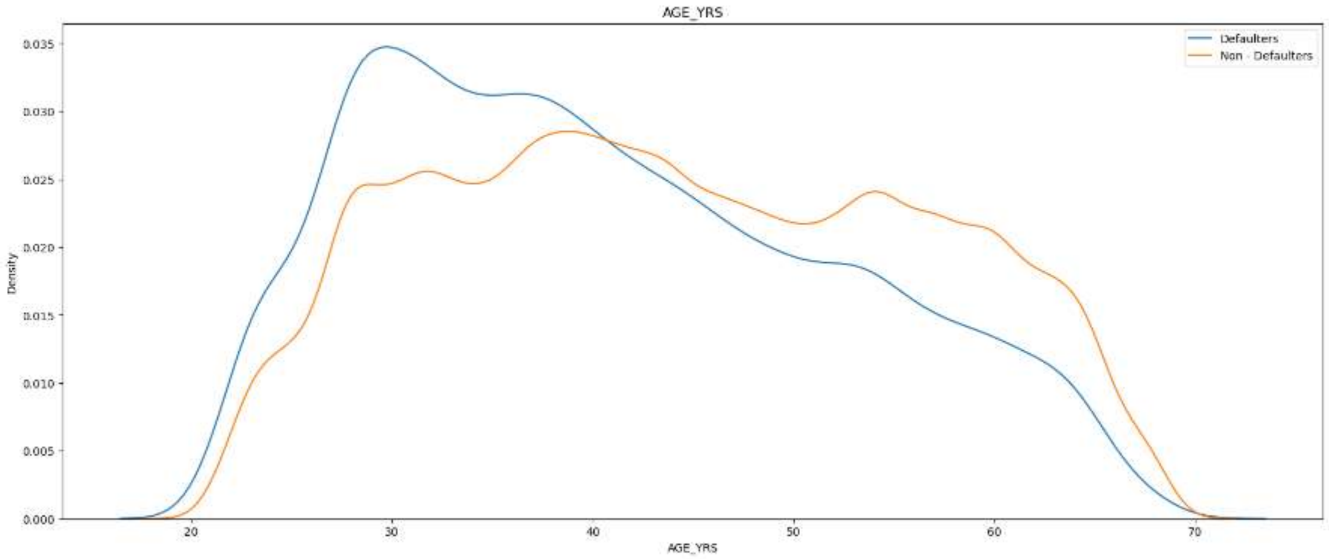




- Students and businessman have no difficulties in payment which is evident from bar plot. Pensioner have higher percentage in non defaulters which indicates that it is non riskier to give loans to pensioners.

- From the above analysis, we can infer that, there is higher percentage of people with higher education who has no payment difficulties which indicates that there is no much risk in providing loans.

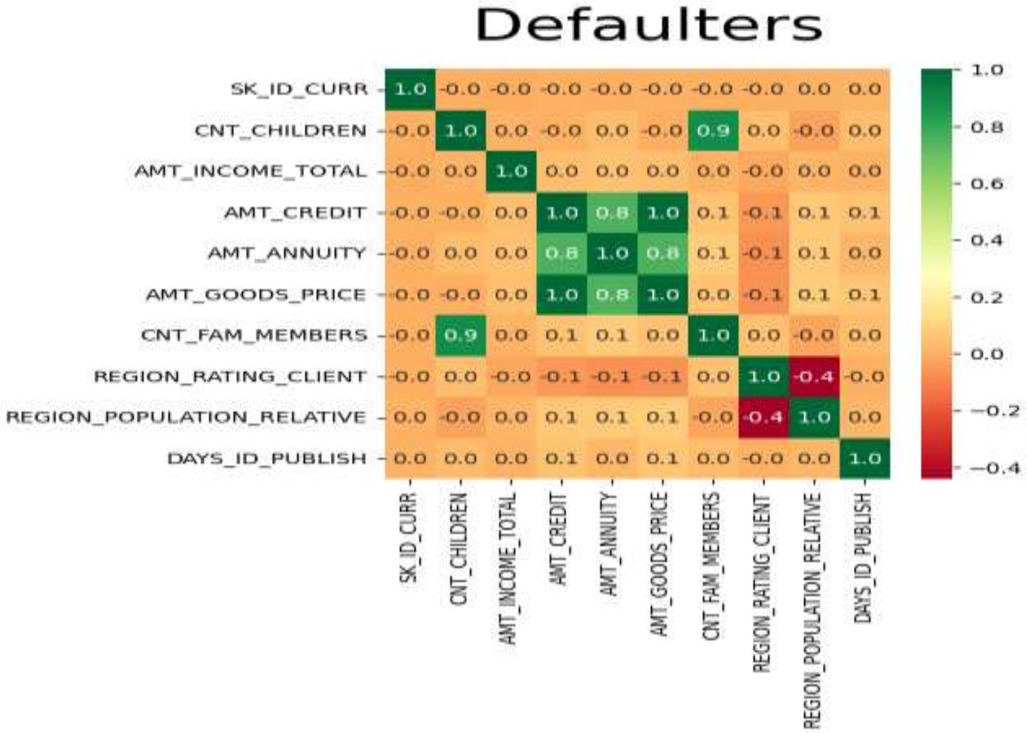




•As the credit amount increases, there is a gradual decrease in the number of defaulters, suggesting a potential correlation between higher credit amounts and improved repayment outcomes.

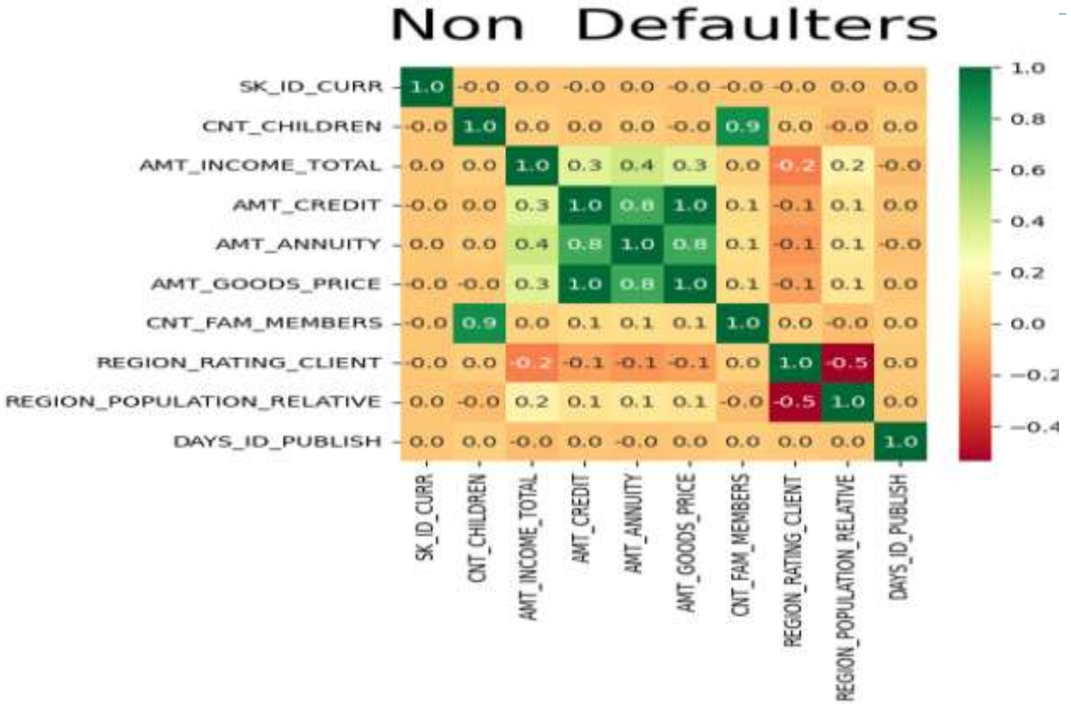
•Individuals aged 20 to 40 seem to face difficulties in repayment, while those aged between 40 and 70 exhibit a higher likelihood of repaying the loan on time.

# BIVARIATE/MULTIVARIATE ANALYSIS



Defaulters

| | SK_ID_CURR | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | CNT_FAM_MEMBERS | REGION_RATING_CLIENT | REGION_POPULATION_RELATIVE | DAYS_ID_PUBLISH |
|---|---|---|---|---|---|---|---|---|---|---|
| SK_ID_CURR | 1.0 | -0.0 | -0.0 | -0.0 | -0.0 | -0.0 | -0.0 | -0.0 | 0.0 | 0.0 |
| CNT_CHILDREN | -0.0 | 1.0 | 0.0 | -0.0 | 0.0 | -0.0 | 0.9 | 0.0 | -0.0 | 0.0 |
| AMT_INCOME_TOTAL | -0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | -0.0 | 0.0 | 0.0 |
| AMT_CREDIT | -0.0 | -0.0 | 0.0 | 1.0 | 0.8 | 1.0 | 0.1 | -0.1 | 0.1 | 0.1 |
| AMT_ANNUITY | -0.0 | 0.0 | 0.0 | 0.8 | 1.0 | 0.8 | 0.1 | -0.1 | 0.1 | 0.0 |
| AMT_GOODS_PRICE | -0.0 | -0.0 | 0.0 | 1.0 | 0.8 | 1.0 | 0.0 | -0.1 | 0.1 | 0.1 |
| CNT_FAM_MEMBERS | -0.0 | 0.9 | 0.0 | 0.1 | 0.1 | 0.0 | 1.0 | 0.0 | -0.0 | 0.0 |
| REGION_RATING_CLIENT | -0.0 | 0.0 | -0.0 | -0.1 | -0.1 | -0.1 | 0.0 | 1.0 | -0.4 | -0.0 |
| REGION_POPULATION_RELATIVE | 0.0 | -0.0 | 0.0 | 0.1 | 0.1 | 0.1 | -0.0 | -0.4 | 1.0 | 0.0 |
| DAYS_ID_PUBLISH | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.1 | 0.0 | -0.0 | 0.0 | 1.0 |

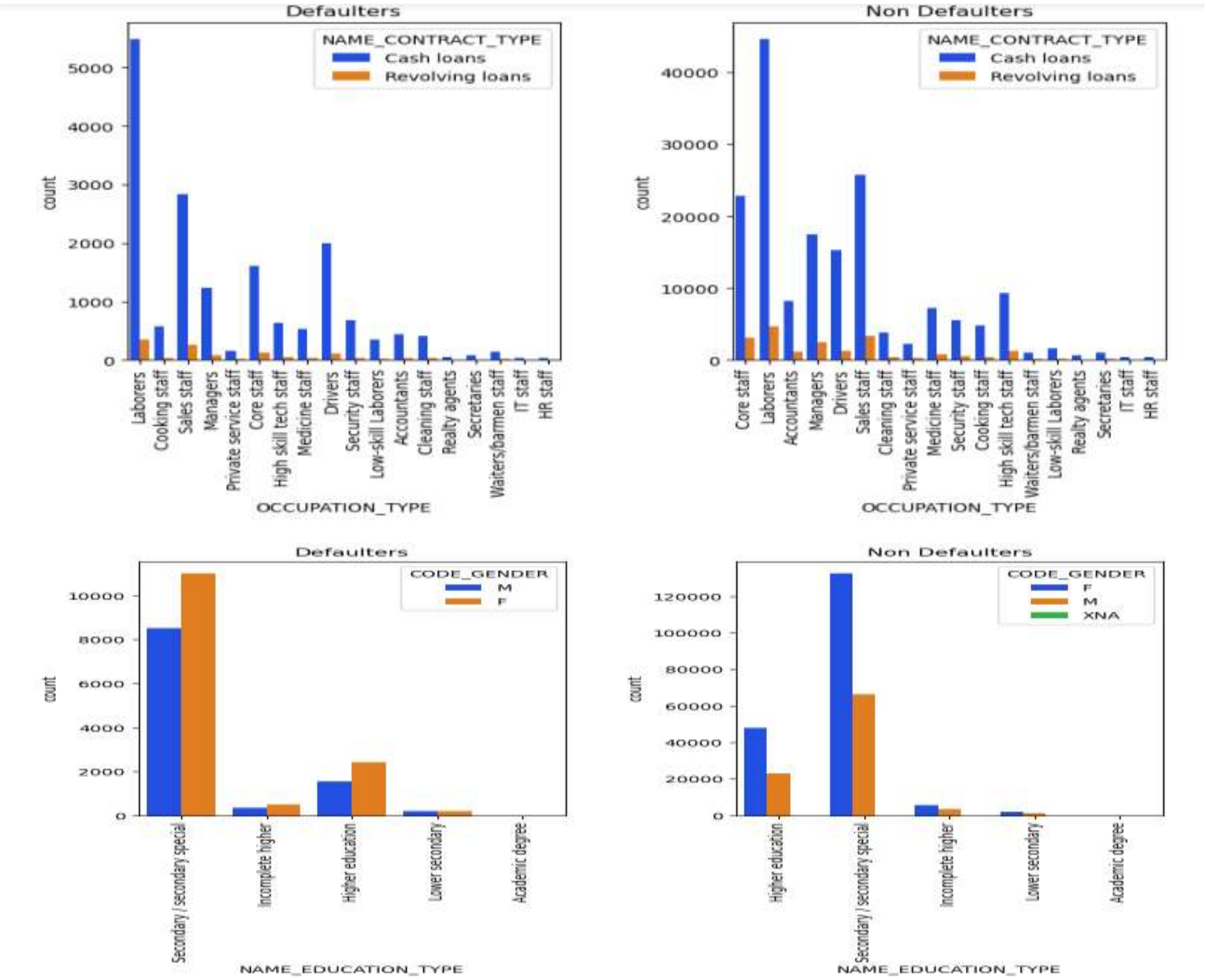


Non Defaulters

| | SK_ID_CURR | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | AMT_ANNUITY | AMT_GOODS_PRICE | CNT_FAM_MEMBERS | REGION_RATING_CLIENT | REGION_POPULATION_RELATIVE | DAYS_ID_PUBLISH |
|---|---|---|---|---|---|---|---|---|---|---|
| SK_ID_CURR | 1.0 | -0.0 | 0.0 | -0.0 | 0.0 | -0.0 | -0.0 | -0.0 | 0.0 | 0.0 |
| CNT_CHILDREN | -0.0 | 1.0 | 0.0 | 0.0 | 0.0 | -0.0 | 0.9 | 0.0 | -0.0 | 0.0 |
| AMT_INCOME_TOTAL | 0.0 | 0.0 | 1.0 | 0.3 | 0.4 | 0.3 | 0.0 | -0.2 | 0.2 | -0.0 |
| AMT_CREDIT | -0.0 | 0.0 | 0.3 | 1.0 | 0.8 | 1.0 | 0.1 | -0.1 | 0.1 | 0.0 |
| AMT_ANNUITY | 0.0 | 0.0 | 0.4 | 0.8 | 1.0 | 0.8 | 0.1 | -0.1 | 0.1 | -0.0 |
| AMT_GOODS_PRICE | -0.0 | -0.0 | 0.3 | 1.0 | 0.8 | 1.0 | 0.1 | -0.1 | 0.1 | 0.0 |
| CNT_FAM_MEMBERS | -0.0 | 0.9 | 0.0 | 0.1 | 0.1 | 0.1 | 1.0 | 0.0 | -0.0 | 0.0 |
| REGION_RATING_CLIENT | -0.0 | 0.0 | -0.2 | -0.1 | -0.1 | -0.1 | 0.0 | 1.0 | -0.5 | 0.0 |
| REGION_POPULATION_RELATIVE | 0.0 | -0.0 | 0.2 | 0.1 | 0.1 | 0.1 | -0.0 | -0.5 | 1.0 | 0.0 |
| DAYS_ID_PUBLISH | 0.0 | 0.0 | -0.0 | 0.0 | -0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |

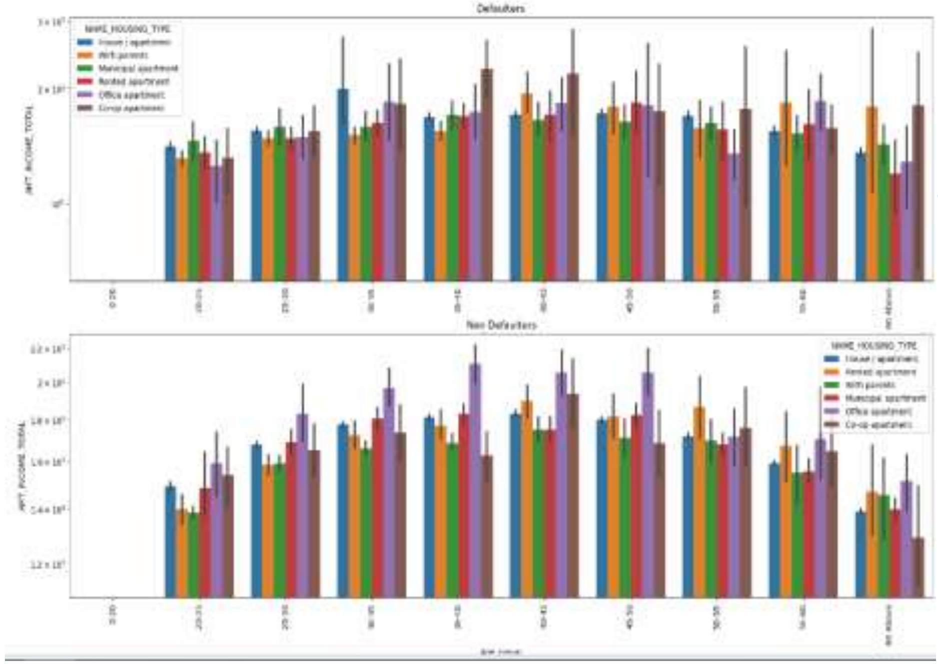
These columns are highly positively correlated with each other:

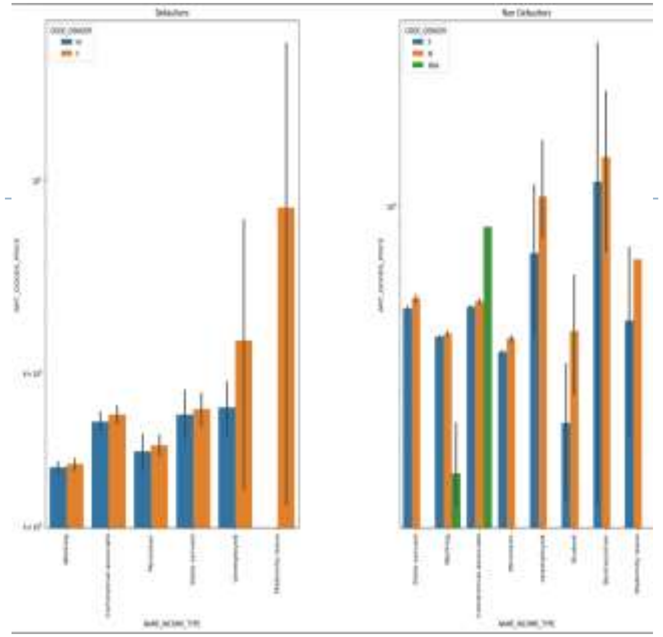
AMT_CREDIT - AMT_GOODS_PRICE
CNT_FAM_MEMBERS - CNT_CHILDREN
AMT_GOODS_PRICE - AMT_ANNUITY
AMT_CREDIT - AMT_ANNUITY

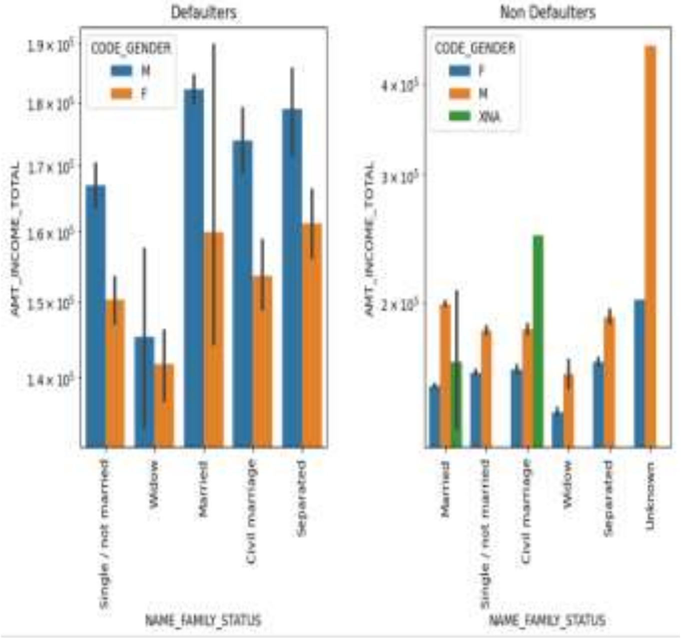


- From the above plot we could see that in secondary special, male have difficulties in on payment and females are more likely to repay the amount on time. Females with higher education are more likely to repay the amount.

- From the above plot we could see that in core staff, drivers with cash loans are more likely to be defaulters.



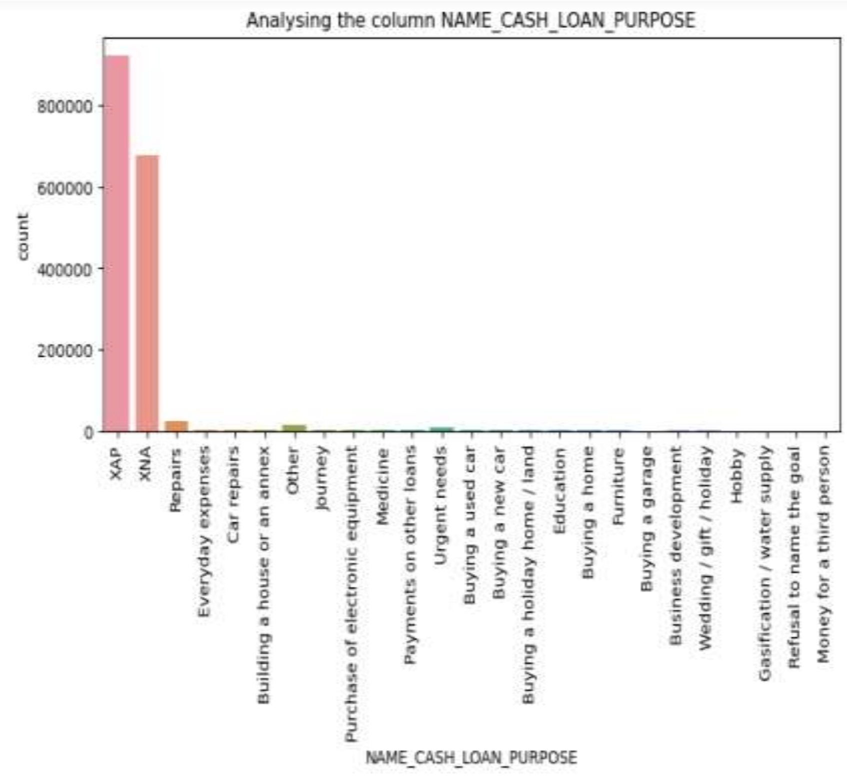


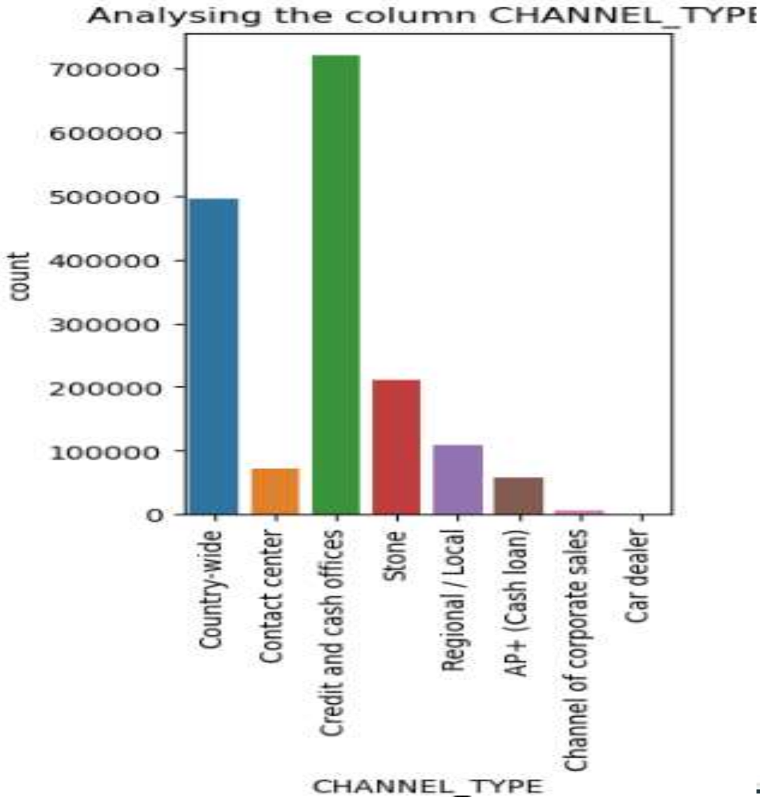


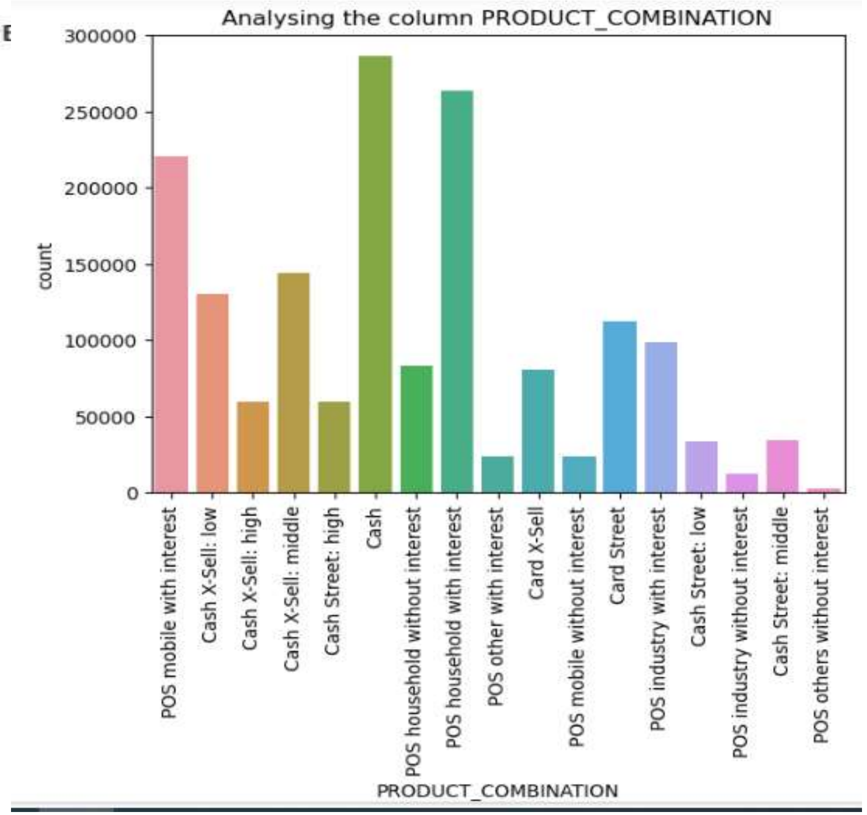
- People aged above 60 and living in Co-op apartment had shown a significant increase in payment difficulties.

- Students and businessman of both the genders are showing good result in making their payments on time.

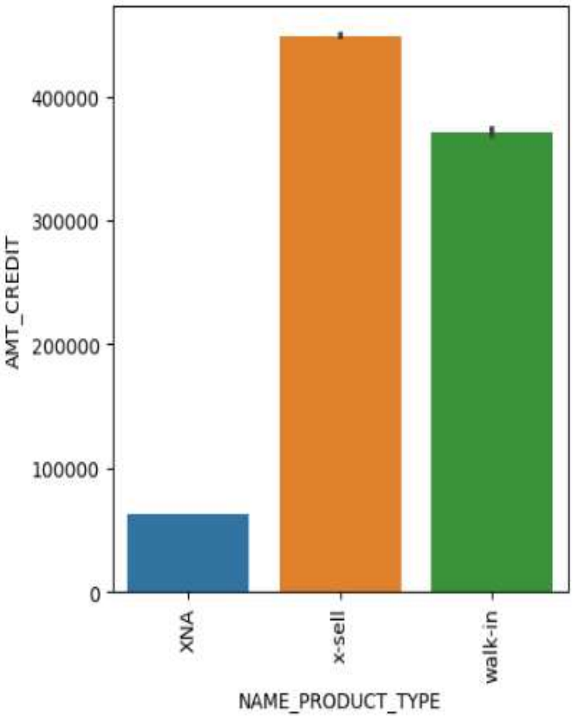
- Widowed male has less difficulty in doing payment on time.

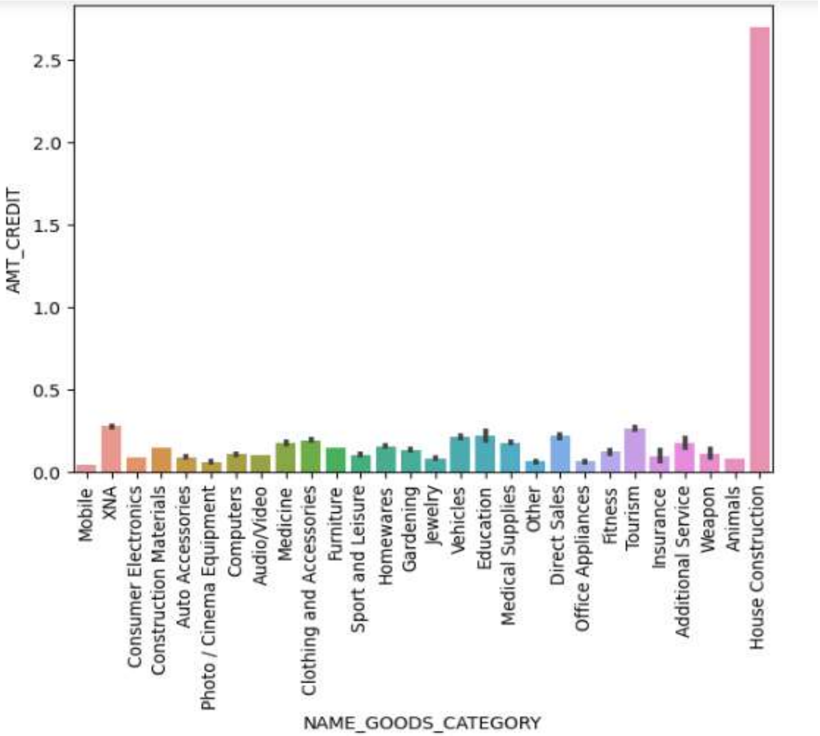


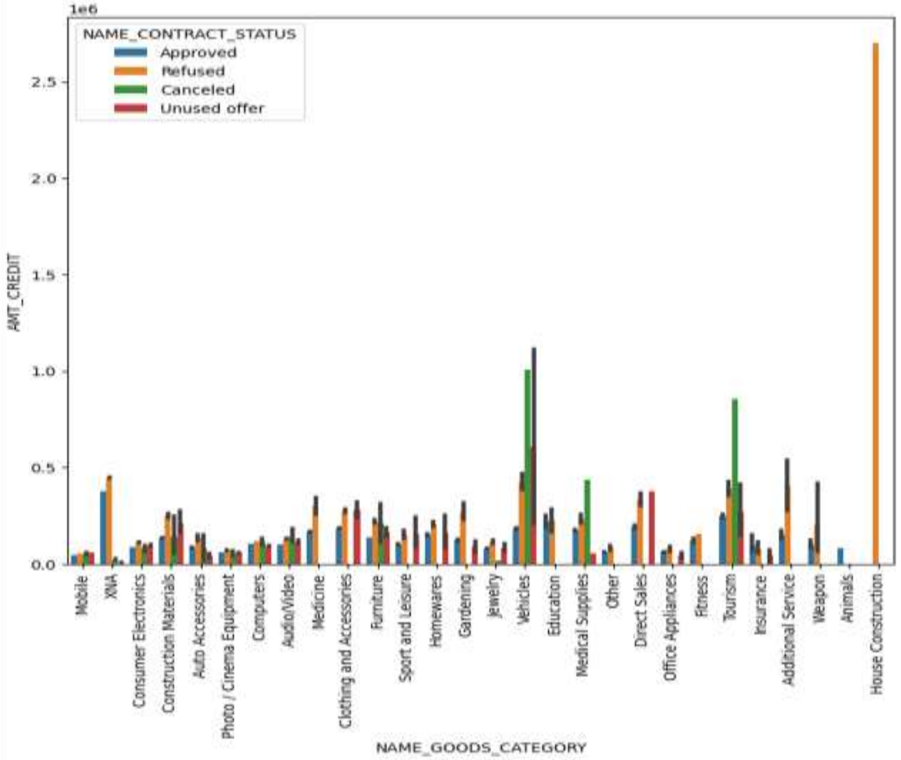




- Significant portion of loan applicants is acquired through credit and cash.The channels of corporate sales and car dealers seem to have lower observation levels, suggesting that fewer applicants are acquired through these channels.
- In the analysis of the graph, it is evident that the product combination labeled 'cash' has received the highest number of loan sanctions, indicating a significant demand or preference for this particular product. Conversely, the combination 'POS others without interest' shows the least number of loan sanctions.
- It has lot of missing values to be actioned which will impact the analysis. Apart from incorrect data, major reason for opting loans is for repairs and the least is people refused to mention their reason for opting loans.



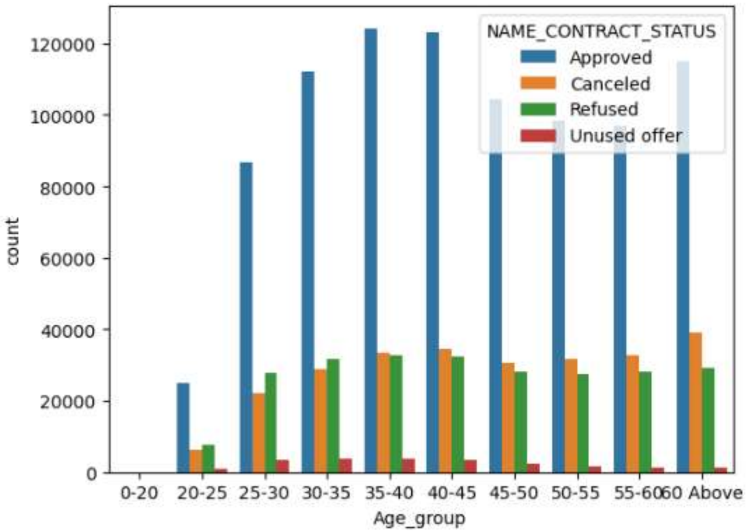


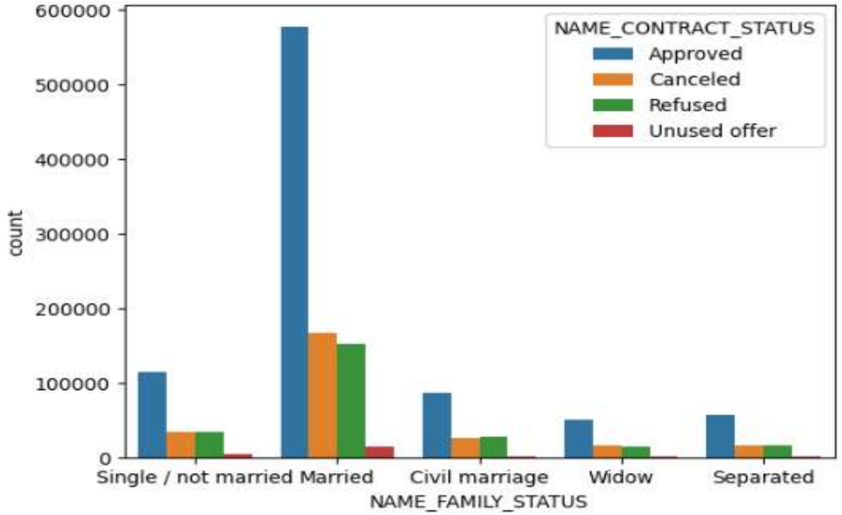


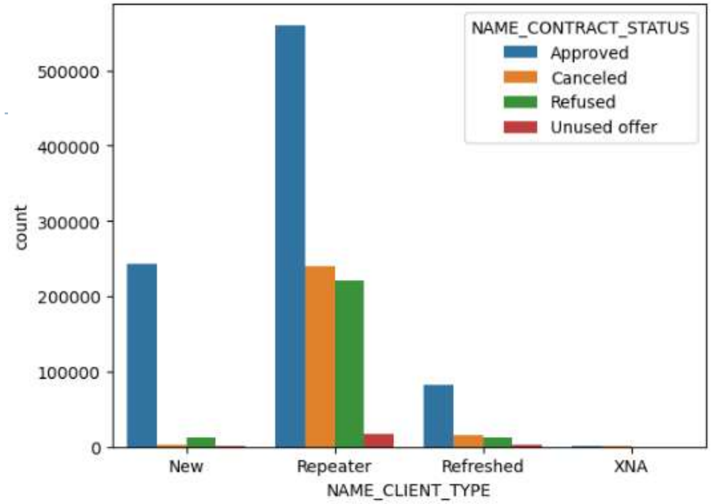
•The customers classified as "x-sell" tend to have a higher number of credit sanctions.The customers labeled as "walk-in" have fewer credit sanctions. This may indicate that the "x-sell" category represents customers who are more actively engaged with the business, leading to higher credit requirements or a higher likelihood of taking up credit offers.

•From the above we could see that people are mostly availing loans for their house construction and next to it is tourism.

•From the above we could see that Loan has been approved largely for tourism and next is education. For construction we could see that ,mostly it is refused.





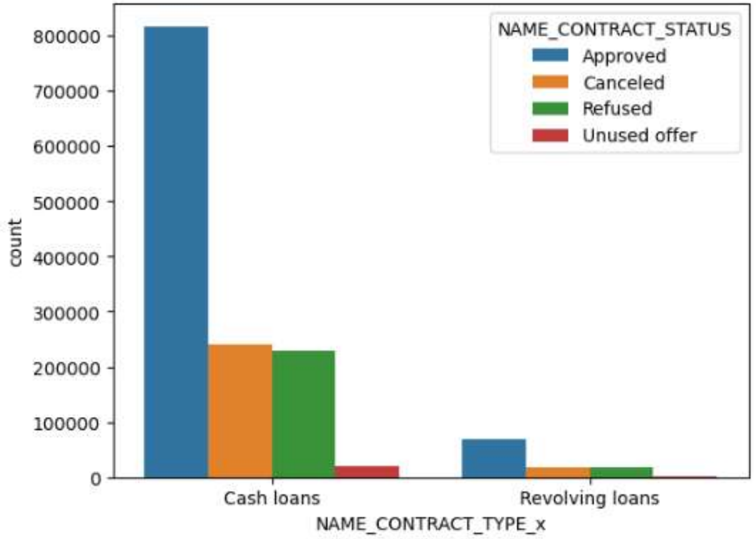


- Applicants under the age 30 to 35 and 35 to 40 has got their loans approved in majority and people between age 20 to 25 and 60 above has got least refusal.

- Married people has got majority of loans approved.

- Repeater has got majority of loans approved.

- Cash loans has got approved large number of times.



# Conclusion:

- **Income Influence:**

  Individuals with an income range of 100,000 to 150,000 constitute a significant portion in both defaulter and non-defaulter groups, highlighting their prevalence in loan applications.

- **Loan Type and Risk:**

  Cash loans exhibit a lower proportion of defaulters compared to non-defaulters, indicating a relatively lower risk associated with this loan type. In contrast, while the number of defaulters in revolving loans is lower, the proportion of defaulters to non-defaulters is comparatively higher, suggesting a potentially higher risk.

- **Gender and Marital Status Impact :**

  Males contribute more to defaults than females, as revealed by count. This underscores the importance of gender in assessing default risk. Married or widowed individuals are more likely to repay loans, indicating a positive correlation. Conversely, loans to those not married, in civil partnerships, or separated pose a slightly higher risk.

- **Education Level and Gender Impact :**

  Higher education correlates with a lower likelihood of payment difficulties, suggesting that individuals with advanced education are less risky borrowers. In the secondary special education category, males face difficulties in payment, while females with higher education are more likely to repay on time.

- **Credit Amount and Defaults:**
  - There is a gradual decrease in the number of defaulters as the credit amount increases, suggesting a potential correlation between higher credit amounts and improved repayment outcomes.
- **Age and Housing Type Impact:**
  - Individuals aged 20 to 40 face more difficulties in repayment, while those between 40 and 70 exhibit a higher likelihood of timely repayments. Individuals above 60 living in Co-op apartments show a significant increase in payment difficulties
- **Correlated Columns:**
  - Several pairs of columns show high positive correlation, emphasizing the need to consider these relationships in analysis and modeling.
- **Occupation and Default Risk:**
  - Core staff and drivers with cash loans show distinct repayment behaviors, with drivers more likely to be defaulters in this category.
- **Profession Impact:**
  - Students and businessmen of both genders consistently exhibit good results in making payments on time.
- **Credit Acquisition Channels:**
  - "X-sell" customers tend to have a higher number of credit sanctions, indicating more active engagement or higher credit requirements.

- **Profession Impact:**
  - Students and businessmen of both genders consistently exhibit good results in making payments on time.
- **Credit Acquisition Channels:**
  - "X-sell" customers tend to have a higher number of credit sanctions, indicating more active engagement or higher credit requirements.
- **Loan Purpose Influence:**
  - Loans are frequently sought for house construction and tourism, with house construction being the most common reason.
- **Product Combination Preferences and Prevalent Reasons for Opting Loans:**
  - The product combination labeled 'cash' receives the highest number of loan sanctions, while 'POS others without interest' has the least, indicating varying demand for different products. Major reasons for opting loans include repairs, with a noteworthy number of loans sanctioned for education and tourism.

The analysis provides a holistic view of factors influencing loan approval and default patterns, encompassing income, loan type, demographics, occupation, education, and more. These insights can guide targeted strategies for risk assessment, customer engagement, and product development to enhance the efficiency and effectiveness of the lending process.