Credit EDA Assignment

DC C62

Presented by
Tirumalesh Kommavarapu

INDEX

Objective

- 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:
- 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 loan providing companies needs to understand the nature of defaulter who are not repaying the loans
- Company wants to understand the driving factor for defaults
- Identifying the defaulter using EDA is the aim of the assignment

Reading the datasets

Dataset used

- application_data.csv
- previous_application.csv

In [1]:
```
# importing necessary libraries
import warnings
import numpy as np, pandas as pd, matplotlib.pyplot as plt, seaborn as sns
warnings.filterwarnings("ignore")
%matplotlib inline
pd.set_option('display.max_columns',None)#To view all the columns in dataset
```

Reading the dataset ¶

In [2]:
```
current= pd.read_csv('C:/Users/tirum/Downloads/Credit EDA/application_data.csv')
previous=pd.read_csv('C:/Users/tirum/Downloads/Credit EDA/previous_application.csv')
```

- Importing the necessary libraries using import method
- Read the both dataset using the pandas library
- Naming the dataframes as current and previous for application_data.csv and previous_application.csv respectively

Exploring Dataset

- current dataframe as shape of (307511, 122) and previous dataframe as shape of (1670214, 37).
- There are total 49 columns having null values more than 35 % in current dataframe.
- There are total 17 columns having null values more than 35 % in previous dataframe

Handling the missing values in the current dataframe

- Dropping the irrelevant columns and columns which has null values more than 35 %
- Dropping the row record that have very few missing values in their columns
- Filling the null values with mode value for category columns
- Mean and median is filled for numerical columns
- Mean is filled for which data is symmetrical distributed
- If the data is skewed then median is filled in missing values

Outlier Analysis

1. CNT_CHILDREN:

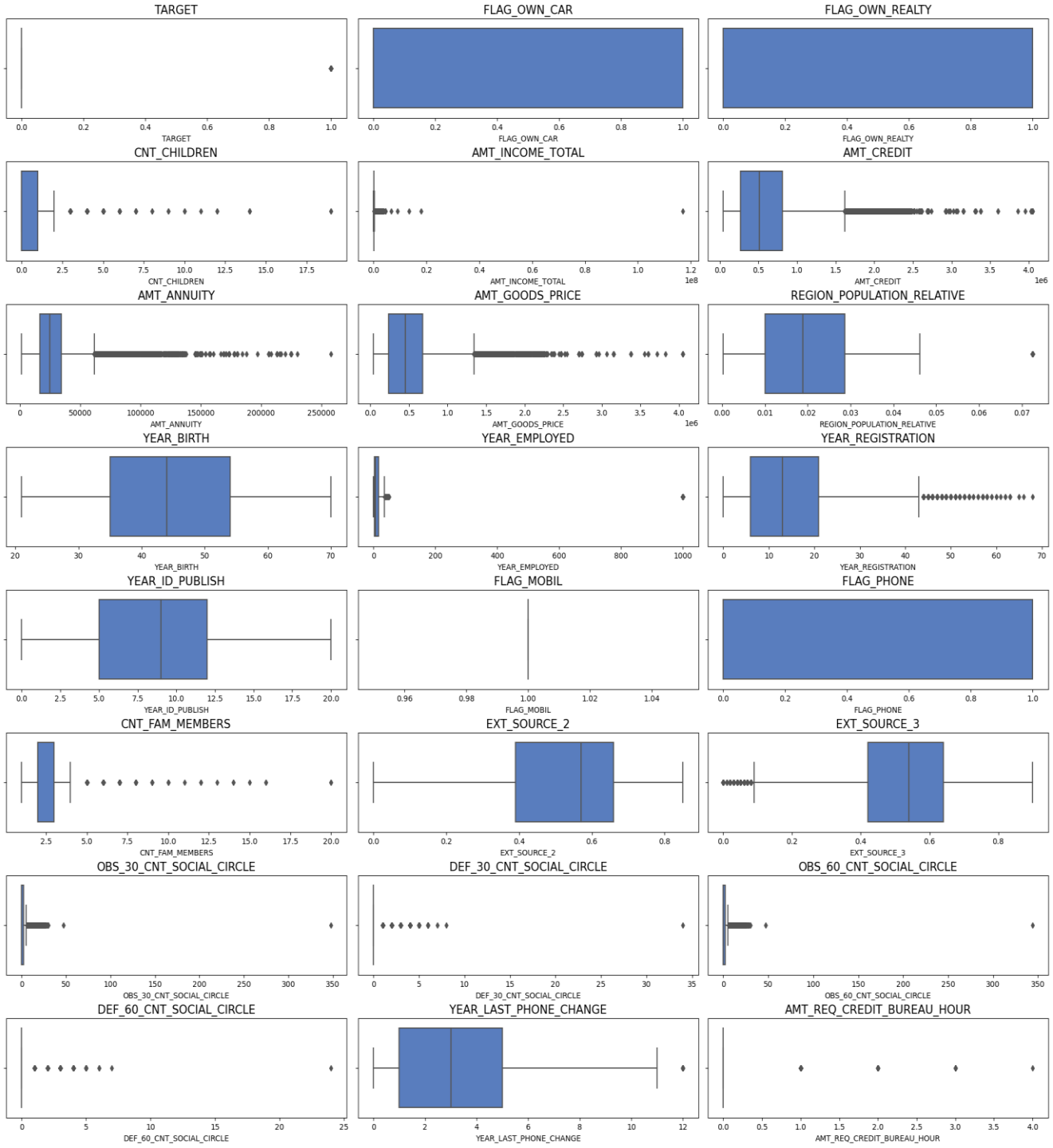
- Outliers are observed in the range from 3 to 19.
- There is a significant difference between the 90th, 95th, and 100th quantiles.

2. AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, and AMT_GOODS_PRICE:

- Outliers are present, but considering the nature of these variables (income, credit, annuity, goods price), high values are reasonable.
- High values in these variables may indicate individuals with high income, large credit amounts, high annuities, or expensive goods.

3. REGION_POPULATION_RELATIVE:

- There are a few values considered as outliers.
- All outliers have a relative population density greater than 0.07.



Outlier Analysis

4. YEAR_EMPLOYED:

- Most values are concentrated between 0 and 20
- It have same value at 90th, 95th and 100th quantiles

5. EXT_SOURCE_2 and EXT_SOURCE_3:

- No outliers are observed in both variables.

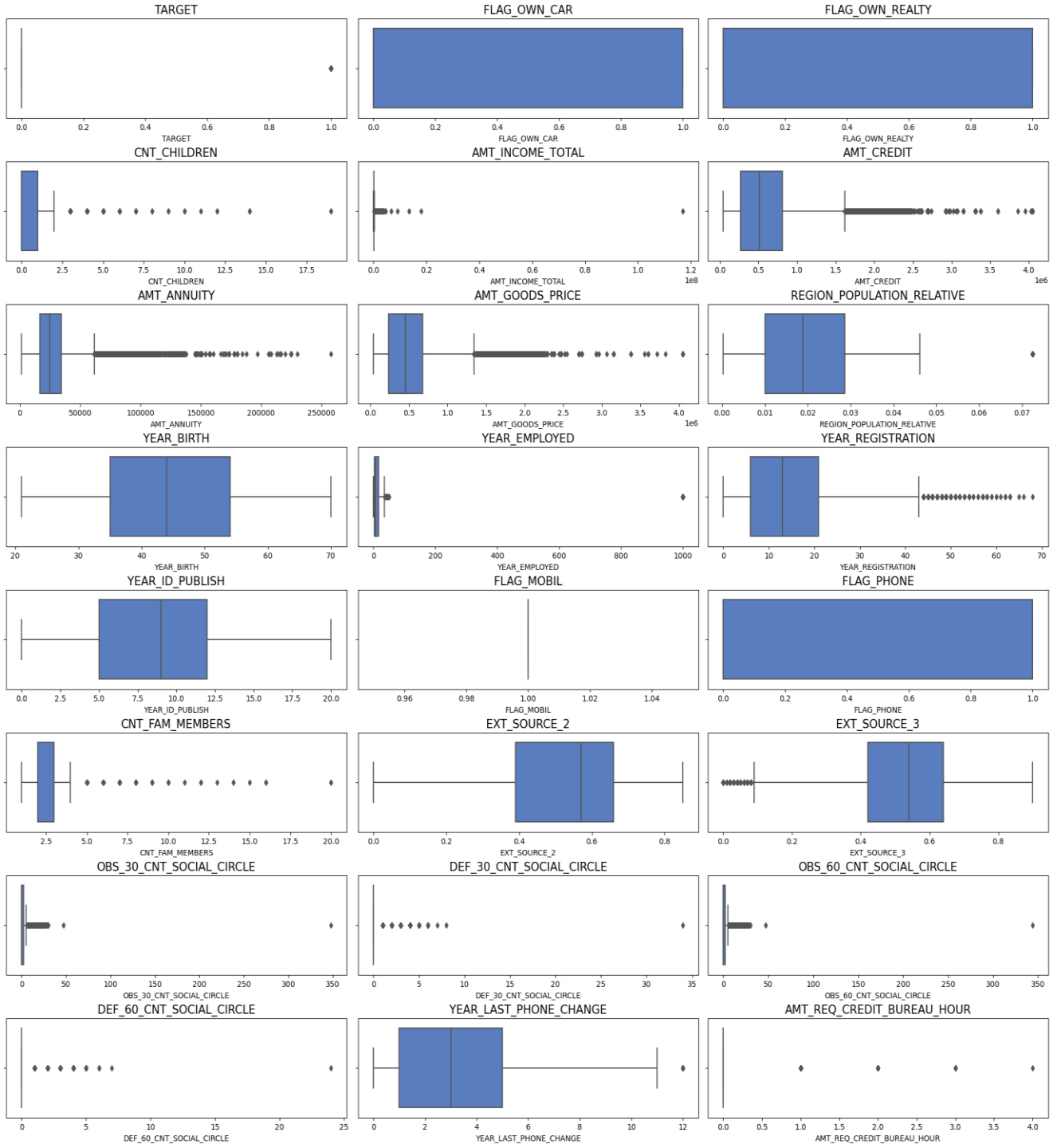
6. OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE:

- Most values are concentrated between 0 and 50.
- A small number of outliers exist, particularly values greater than 20.

7. DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE:

- Most values are between 0 and 50.
- A few outliers are present, especially values exceeding 20.

In summary, while there are outliers in some variables, the context of each variable should be considered. For variables related to financial aspects (income, credit, etc.), high values may be realistic.



Univariate Analysis

1. Target:

- The current dataset, focused on application data, exhibits a notable imbalance in the distribution of default and non-default instances. Specifically, the defaulted population constitutes 8.1% of the dataset, while the non-defaulted population dominates with a share of 91.9%. This results in an imbalanced ratio of 11.3,

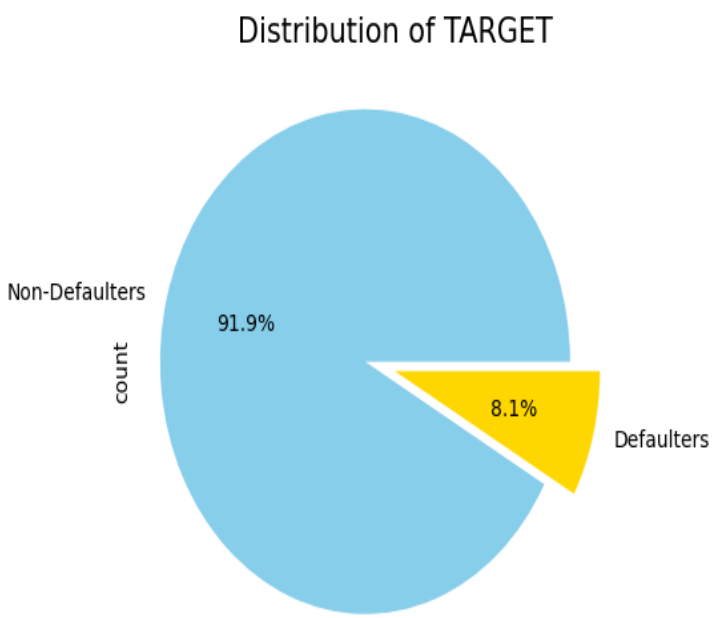
  indicating a substantial disproportion between the two classes.

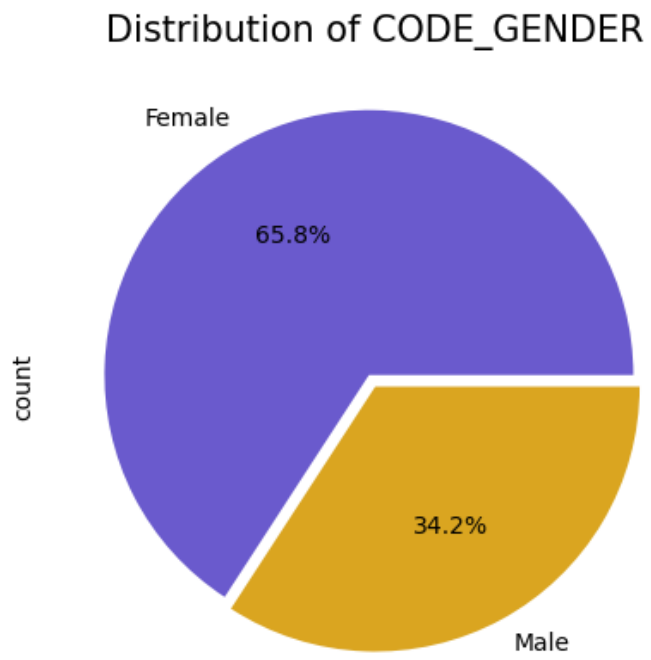
2. NAME_CONTRACT_TYPE

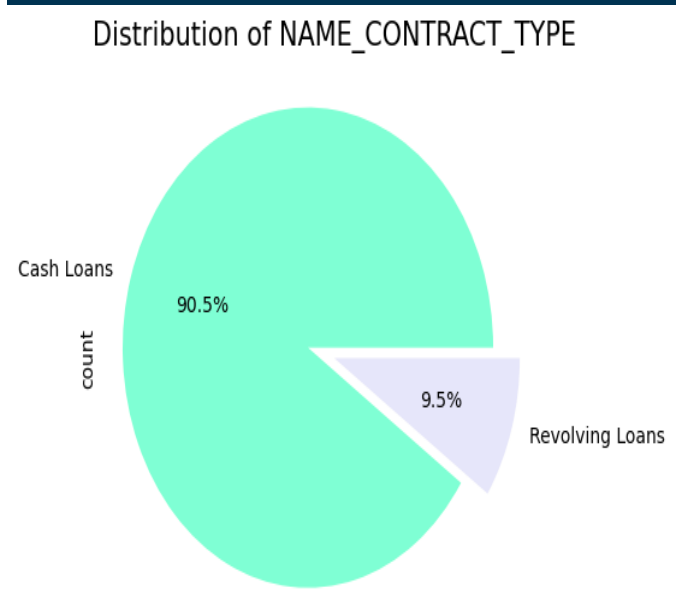
- **Cash Loans**: The majority of contracts fall under the category of "Cash Loans," constituting approximately 90.5% of the total contracts. This indicates that the majority of loan agreements in your dataset are traditional cash loans.
- **Revolving Loans**: The remaining 9.5% of contracts belong to the "Revolving Loans" category. Revolving loans are a type of credit that allows a borrower to repeatedly borrow money up to a certain limit and repay it in installments.
- This suggests a significant imbalance in the distribution of these contract types. The majority of contracts are of the "Cash Loans" type (90.5%), while the "Revolving Loans" type constitutes only a small portion (9.5%).

3. CODE_GENDER

- Majority of loans are taken by Females







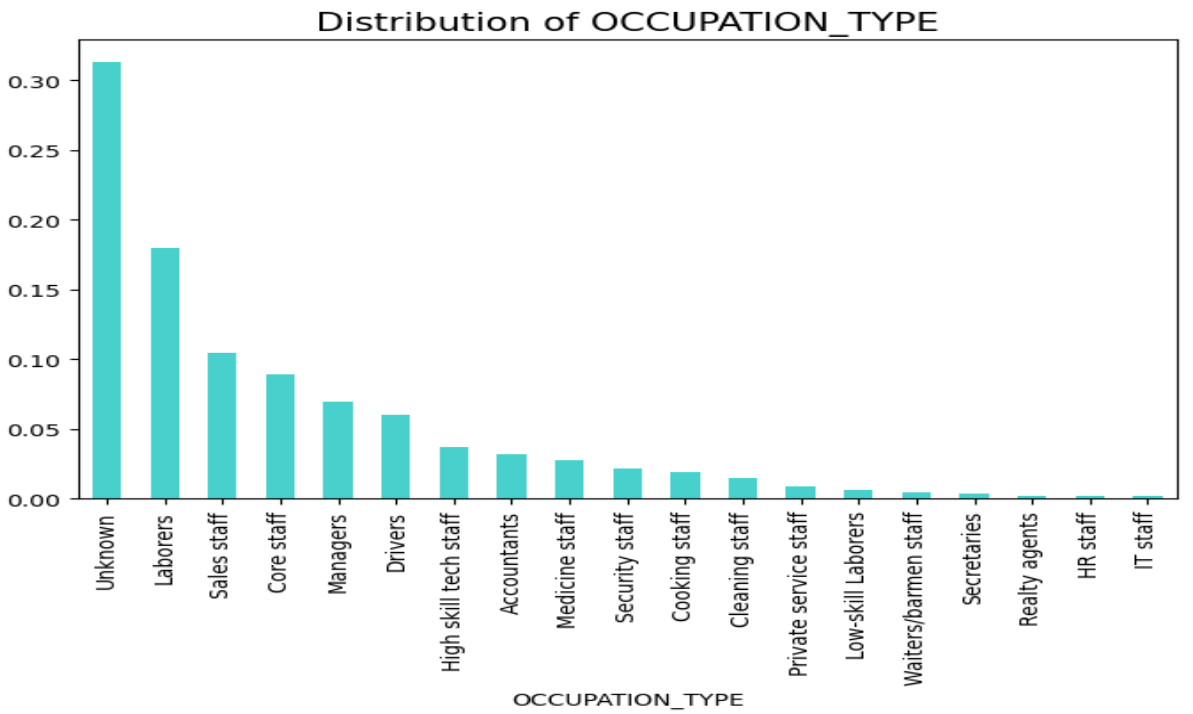
Univariate Analysis

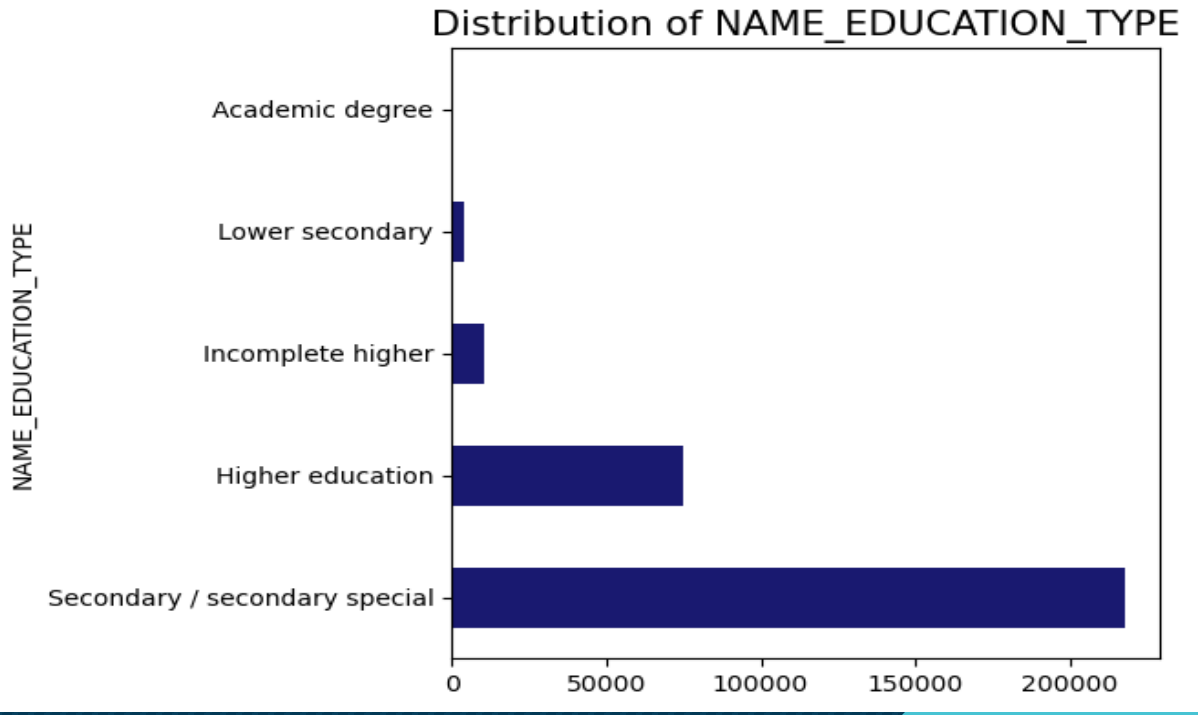
1. **OCCUPATION_TYPE:**

- Majority of Occupation type is Unknown apart from Unknown Laborers & Sales staff are taken more Loans

2. NAME_EDUCATION_TYPE:

- Majority of clients have studied Secondary / secondary special
- Least is the Academic degree





Univariate Analysis

1. Age Group:

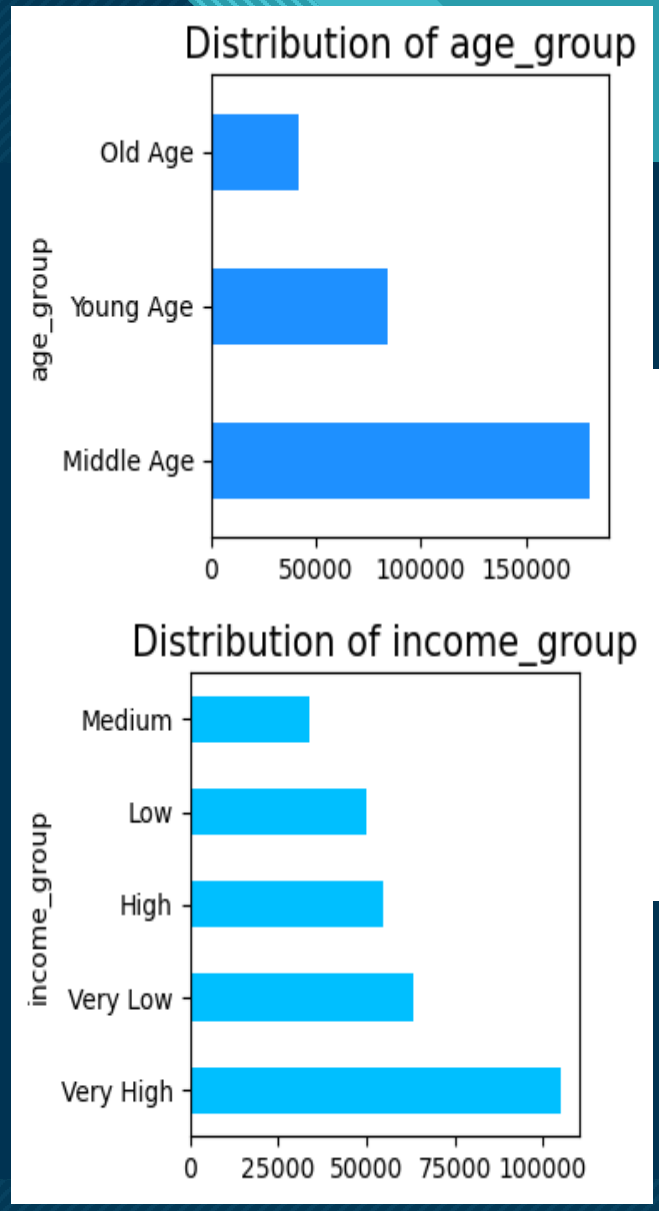
- Most of the clients are under middle age group who are grater than 35 years and less than 60 years

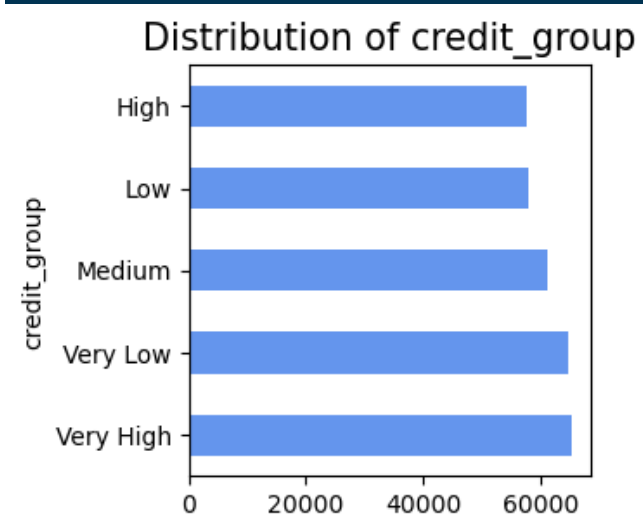
2. Income Group:

- Majority of Loans are taken by Very High Income group who as income more than 225000

3. Credit Group:

- Significantly same Loans are taken by all Credit groups however with small margin Very High group takes more Loans





Bivariate Analysis

Defaulters

1. CODE_GENDER:

- Most of the Defaulter are Females

2. FLAG_OWN_CAR:

- Most of the Defaulter do not have Cars

3. FLAG_OWN_REALTY:

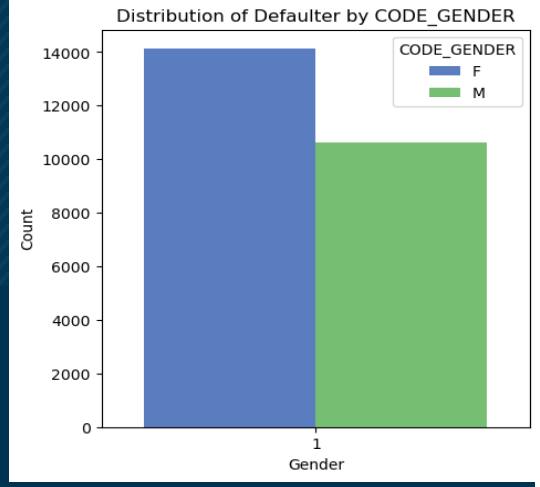
- Most of the Defaulter do have their Own Houses or Land

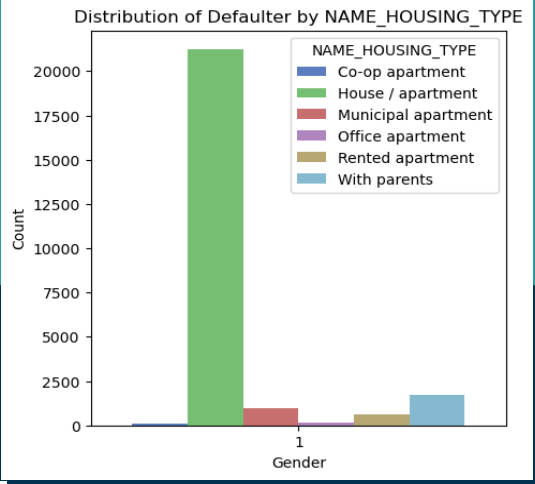
4. NAME_CONTRACT_TYPE

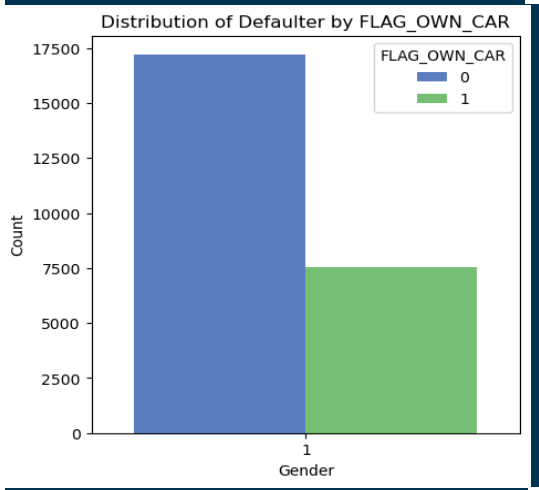
- Most of the Defaulter have Cash Loans

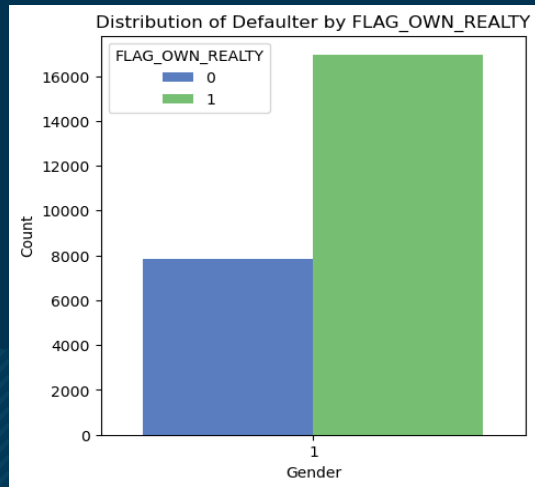
5. NAME_HOUSING_TYPE:

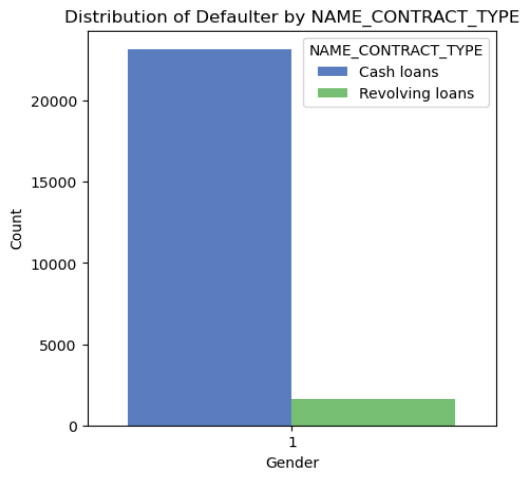
- Most of the Defaulter lives in House/apartment











Bivariate Analysis

Defaulters

6. **NAME_TYPE_SUITE:**

- Most of the defaulters were unaccompanied when applying for the loan.

7. **NAME_FAMILY_STATUS:**

- Most of the Defaulters are married

8. **Age Group**:

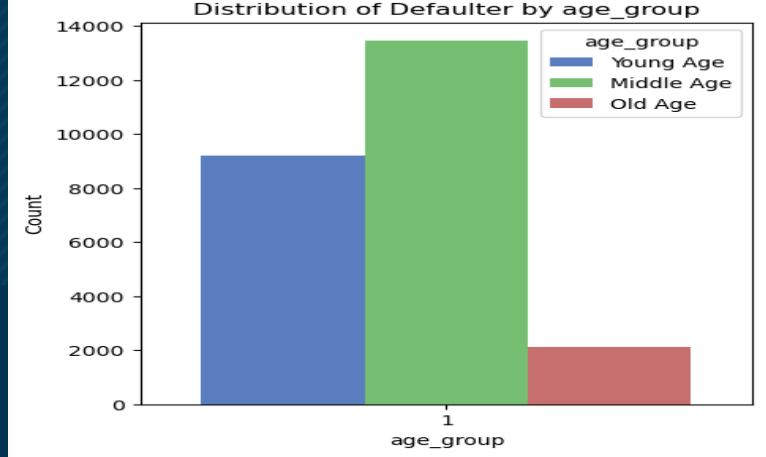
- Most of the Defaulters are Middle Age group

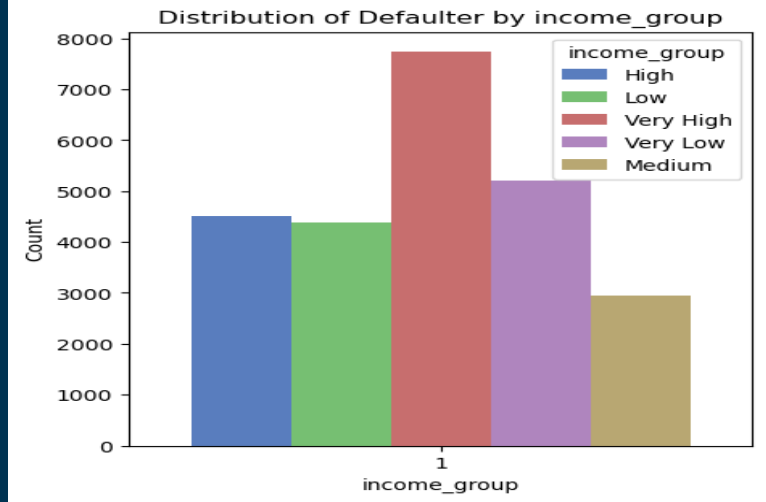
9. **Income Group:**

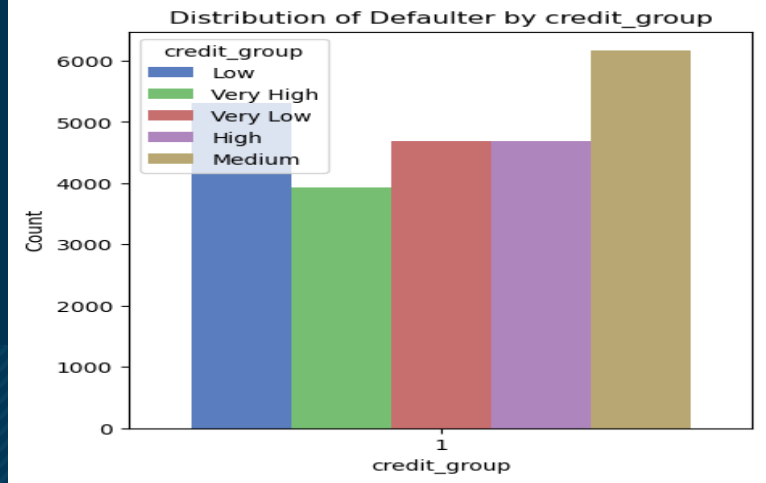
- Most of the Defaulters are related to high income group who are having income greater than 225000

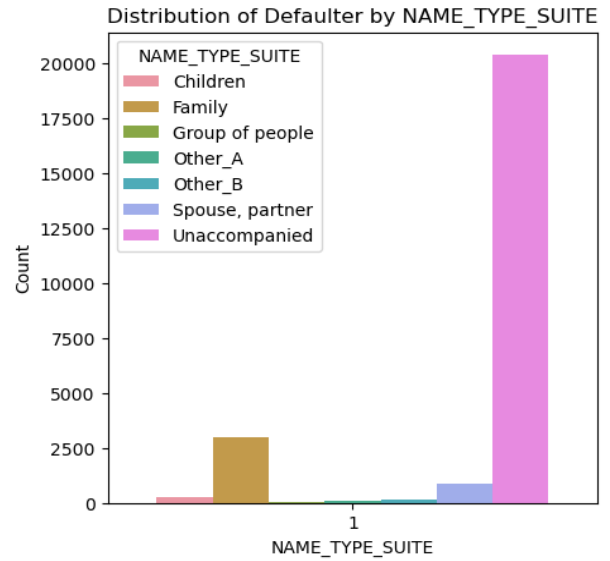
10. **Credit Group**:

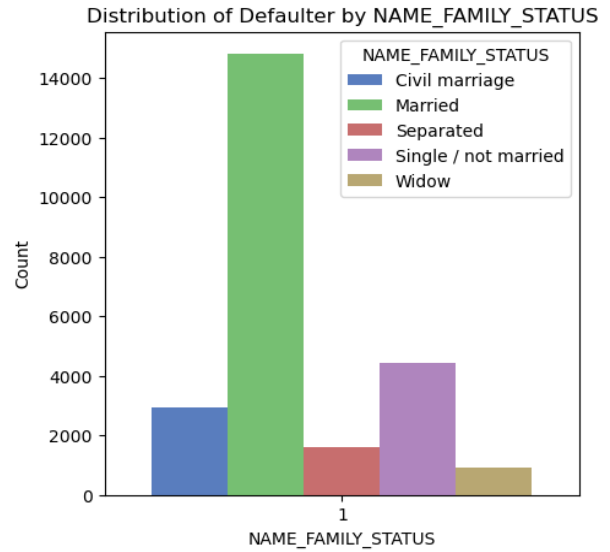
- Most of the defaulters are related to the medium credit group, specifically those with a credit greater than 604,413 and less than 900,000.



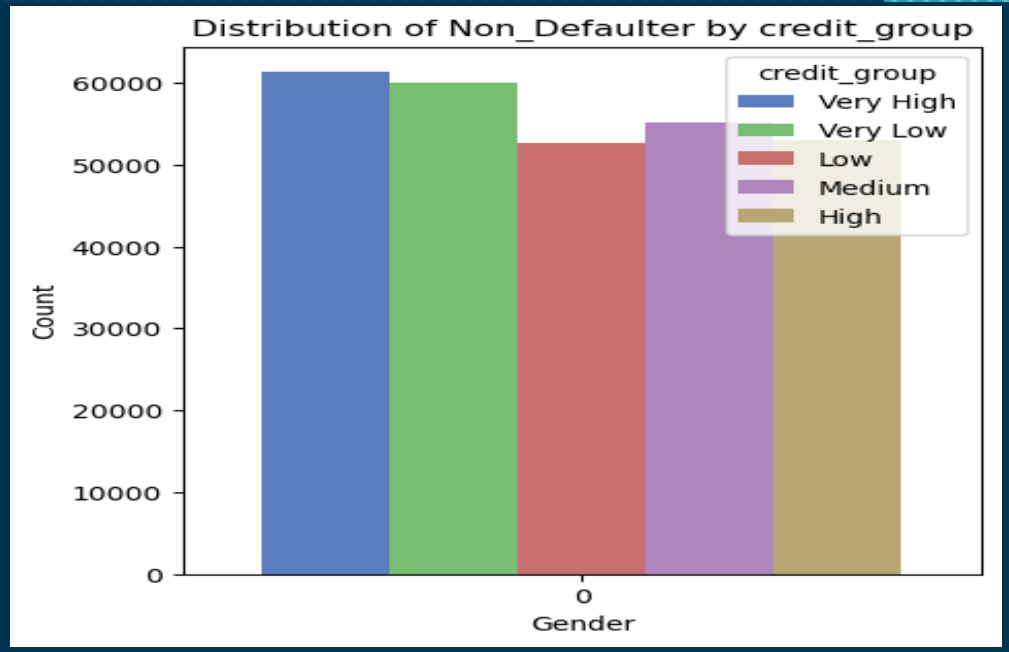








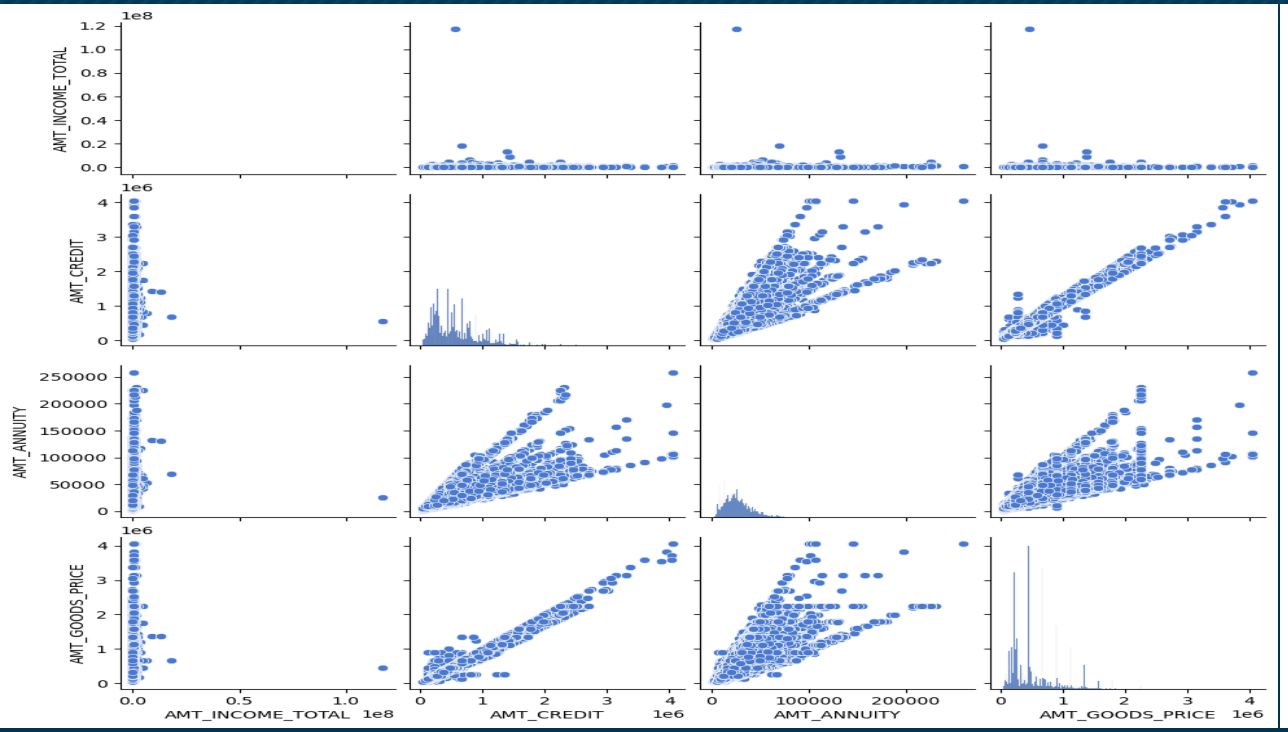
Bivariate Analysis

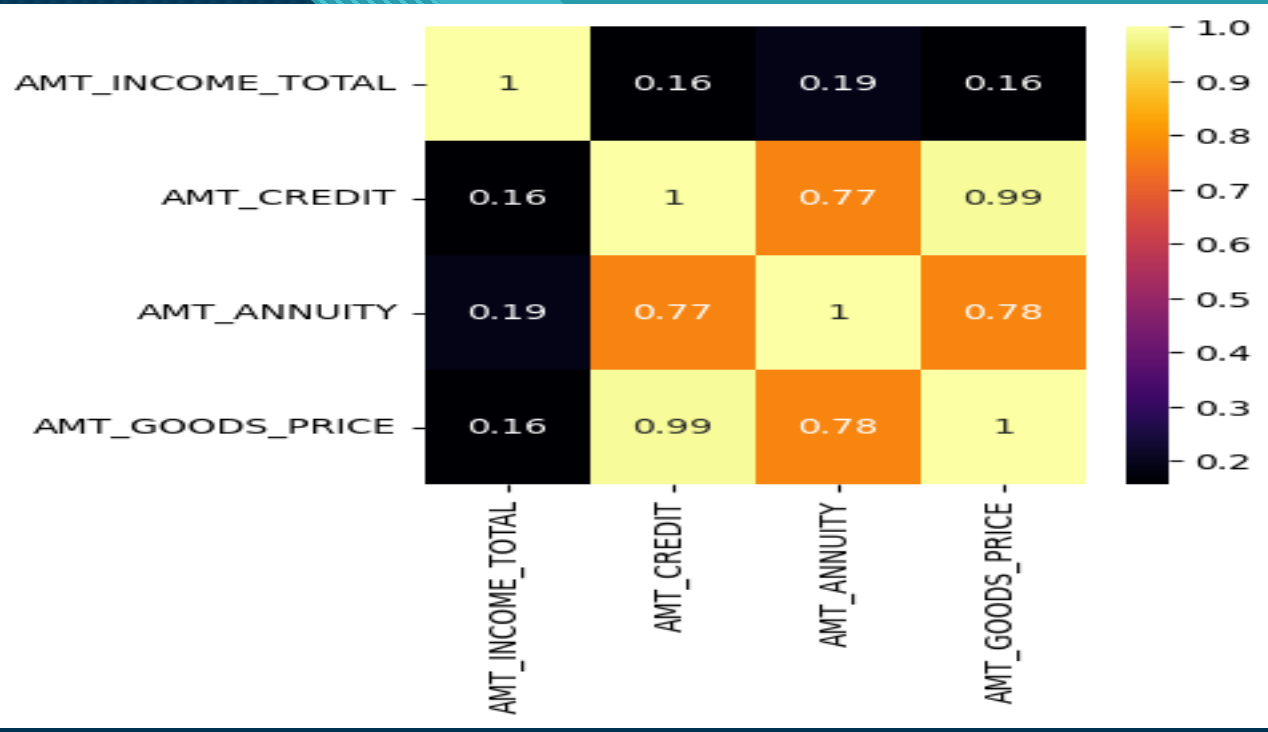


Non Defaulters

- All of the Non-Defaulters graphs are significantly same as Defaulters except credit_group.
- Most of the Non Defaulters are related to the very high credit group, specifically those with a credit greater than 900,000.

Multivariate Analysis

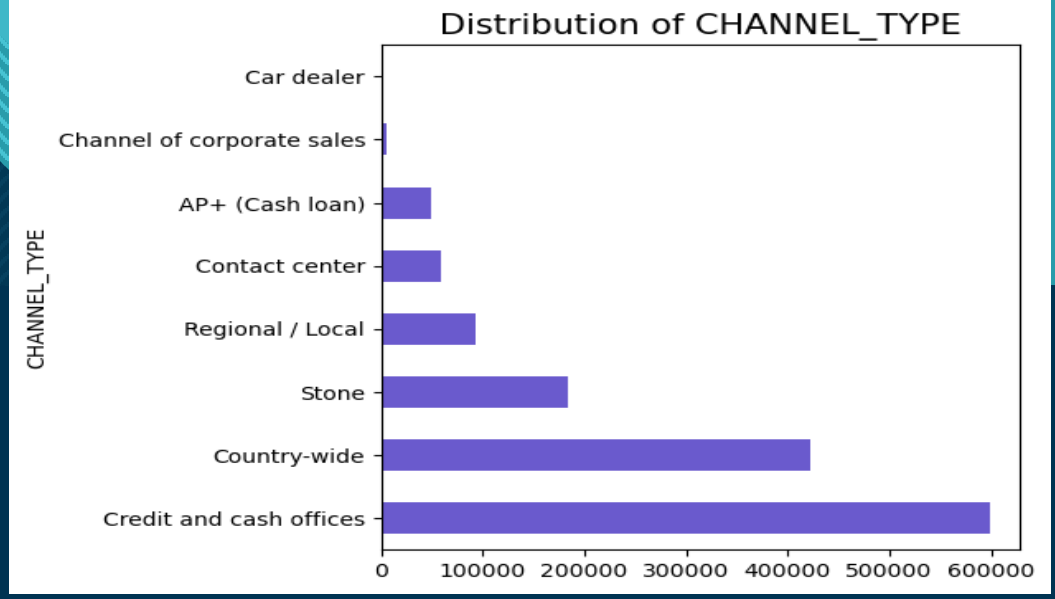




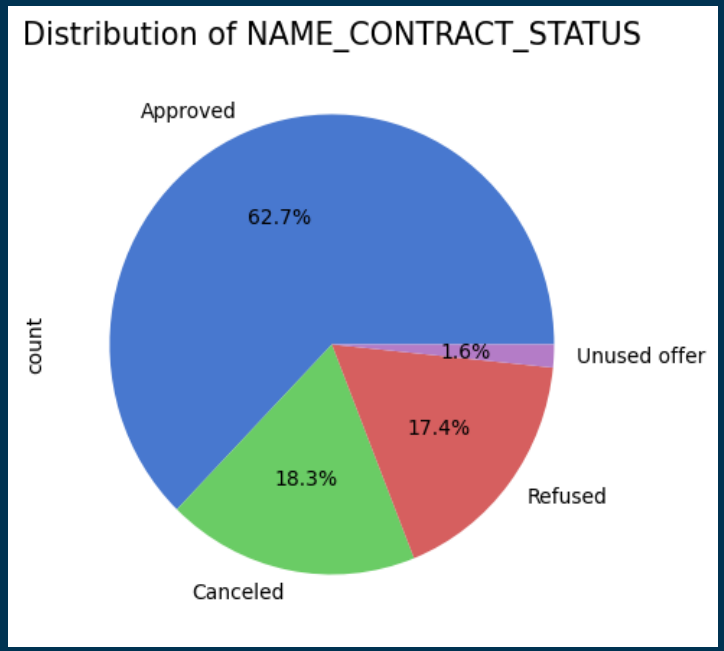
- A strong association between AMT_CREDIT, AMT_ANNUITY and AMT_GOODS_PRICE

Previous Dataframe Analysis
Univariate Analysis

1. **Channel Type:**
- Most of the clients are acquired from Credit and cash offices on the previous loan application

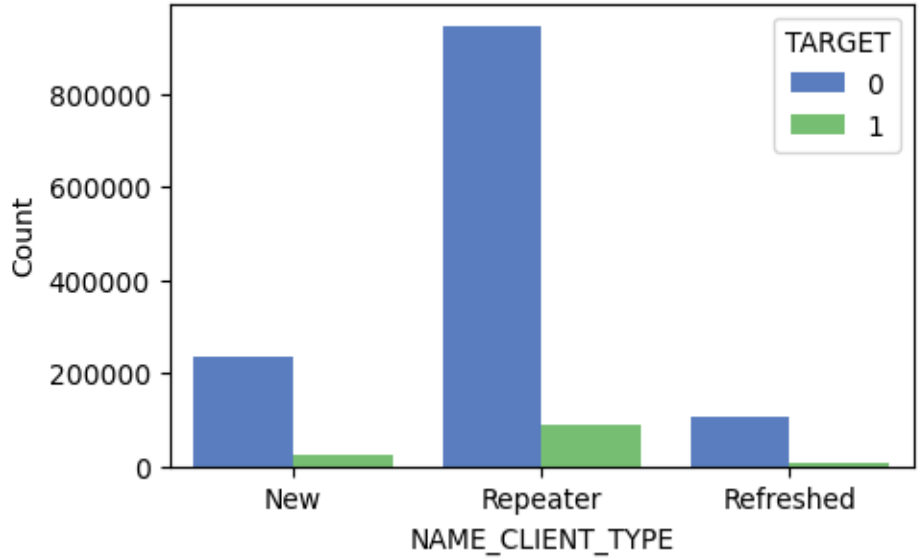


2. NAME_Contract_STATUS:
- 63% percent of the previous loan applications are approved
- 17.4% percent are rejected the previous loan applications

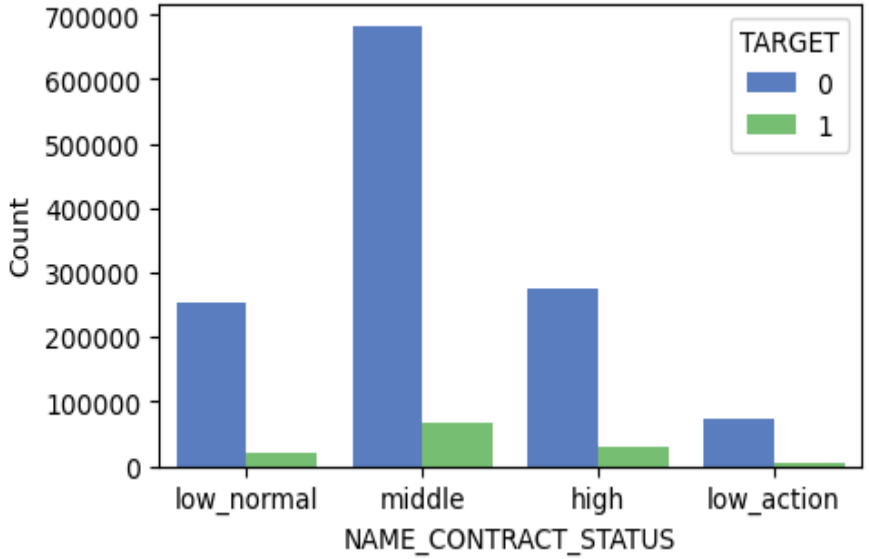


Bivariate Analysis

Distribution of Defaulters and Non-Defaulters NAME_CLIENT_TYPE by TARGET



Distribution of Defaulters and Non-Defaultersby NAME_YIELD_GROUP and TARGET



1. TARGET vs NAME_CLIENT_TYPE:

- As we see that most of the previous loans application from Repeater(Old customers) and Majority of Defaulters are Repeaters

2. TARGET vs NAME_YIELD_GROUP:

- Most of the Defaulters are in medium interest group

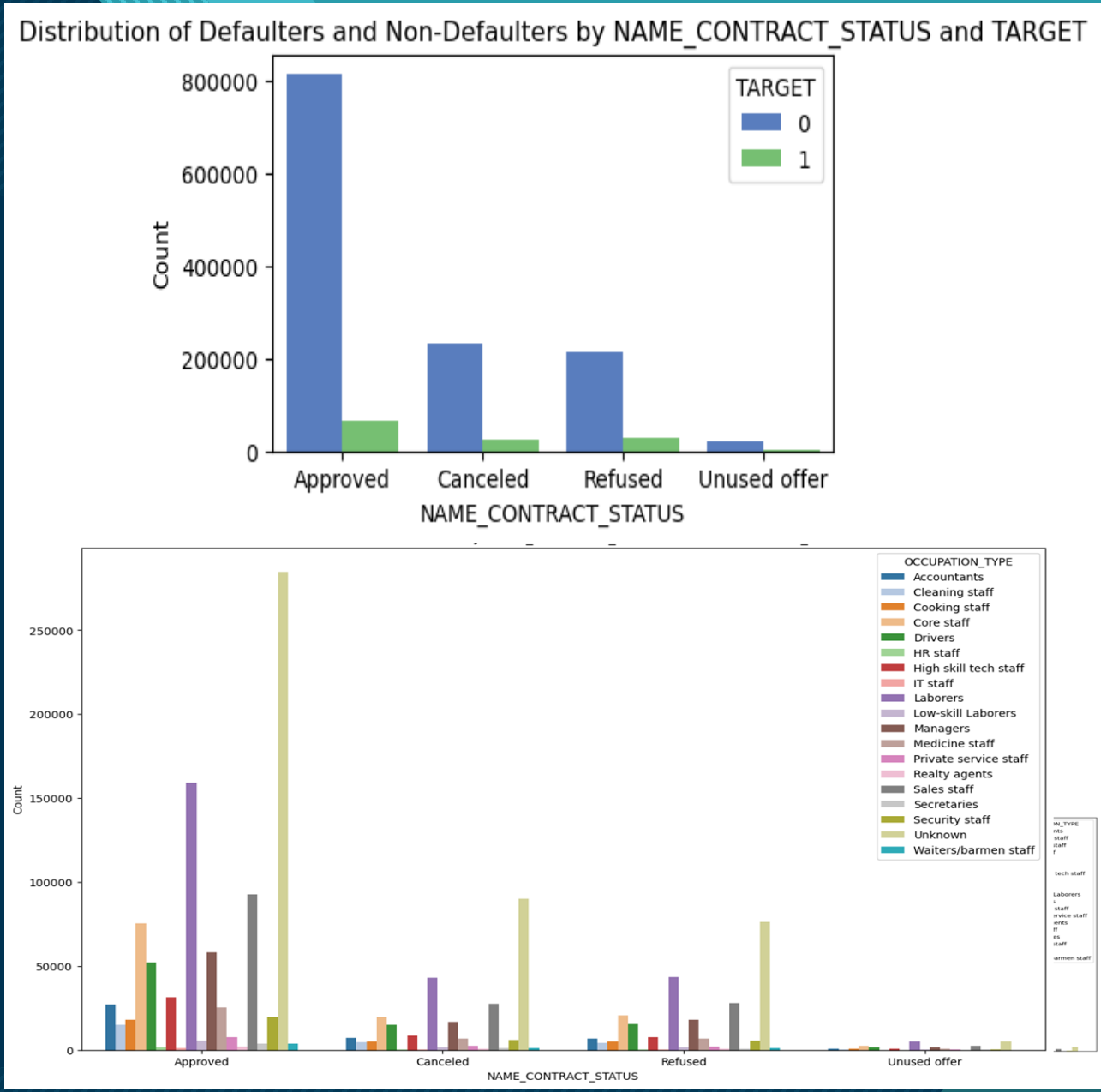
Bivariate Analysis

1. CONTRACT STATUS vs TARGET

- Most of the Loans are Approved
- Most of them are non defaulter and rate of approving for defaulter is more when compare with cancelled, refused and unused offer

2. CONTRACT STATUS vs OCCUPATION

- Most of the loans are approved for applicants with an unknown occupation type. Apart from the unknown category, a higher number of approvals are observed for laborers, sales staff, and core staff. And Same for rest of categories



Thank You