Credit EDA Assignment

–Swati Ingole Nichat

Table of contents

Problem statement

This EDA case study aims to identify the driving factors or the variables which strongly indicates the defaulters (that is the clients who have difficulty in paying the loan). Upon finding factors the company then can utilise the analysed risk factors associated for taking actions against defaulters such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

So the **AIM** is to analyse the patterns in the dataset to **find out the factors** which ensure that :

1. The loan application is not rejected for those clients who are capable of repaying it.
2. The applicant belongs defaulter category.

Approach /Methodology

The approach/ methodology for the problem statement followed would be:

- 1. Understanding the domain/variables (columns) of the dataframe with the data dictionary given (column description) and Google search.
- Import the libraries and the warnings
- Load/read the data
- Check the structure of the data (Normal routine check)
- Application data: Data quality (sanity) check: Missing values and outlier check.
- Check the datatype/anomalies of/in all columns and change/fix the datatype/values accordingly and Data standardization
- Binning
- Data imbalance detection for application data.
- Univariate Analysis
- Bivariate and Multivariate Analysis
- Top correlations
- Similar analysis is followed on previous_data
- Merged data analysis.
- Inference and conclusion

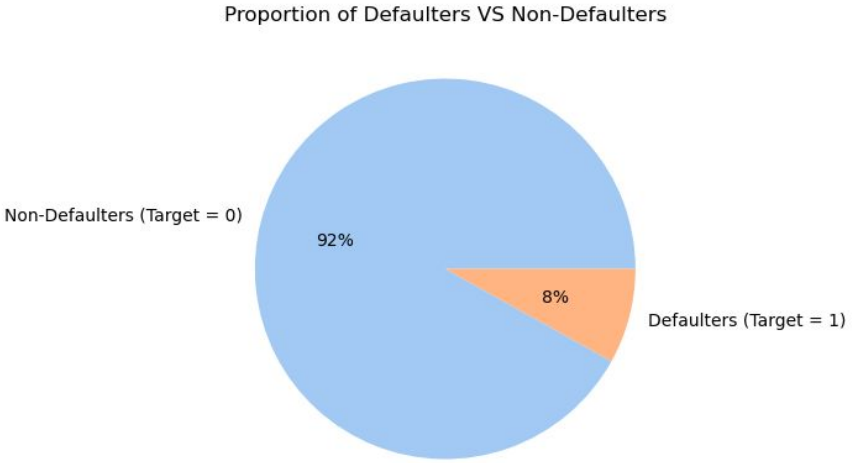
Assumptions:

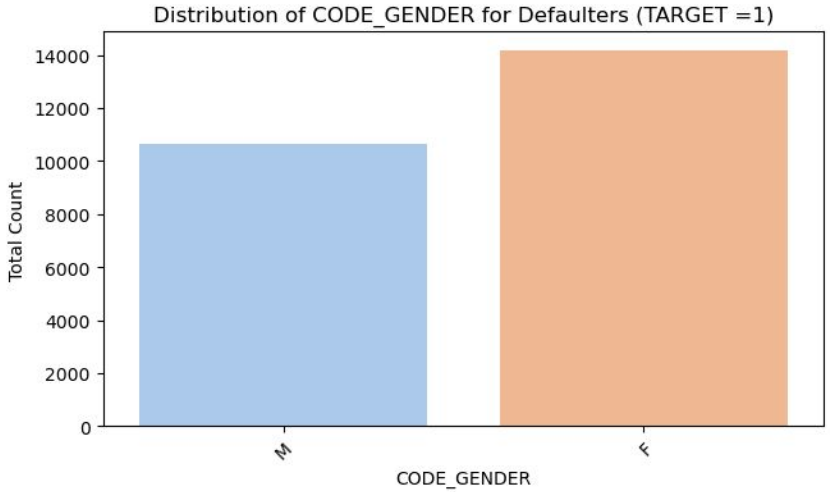
1. The applicants with difficulty in payment are called as defaulters **(TARGET =1)**
2. The applicants without difficulty in payment are called as non defaulters **(TARGET =0)**
3. XNA in the dataset is assumed as unknown as there is no information about it and hence can not be treated as missing.

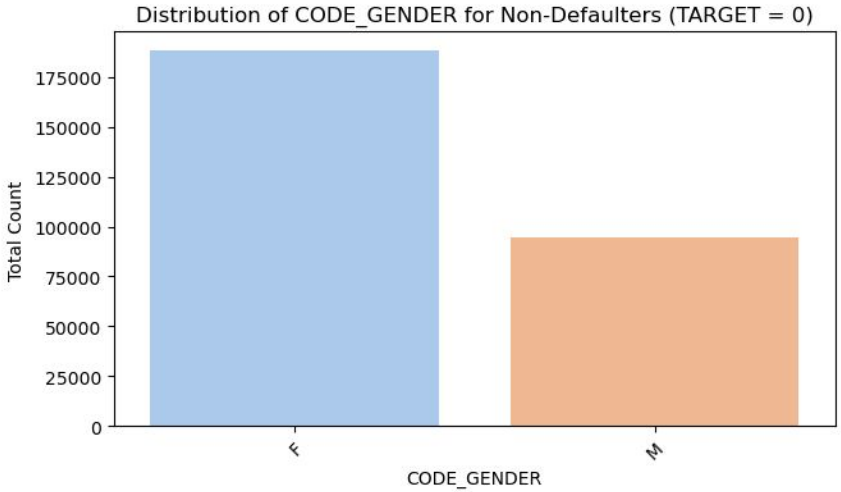
Graphs and Insights:

Proportion of defaulters:

- 8% of the total data are defaulters .
- This implies that in every 11 applicants we have 1 defaulter.
- Male defaulters are more.

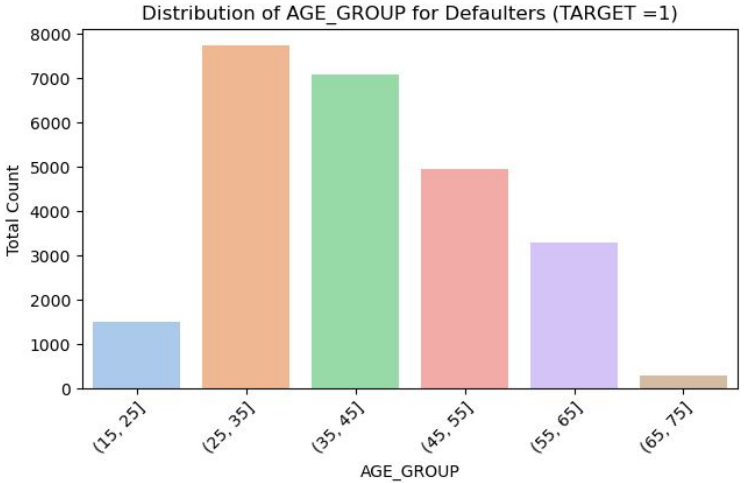


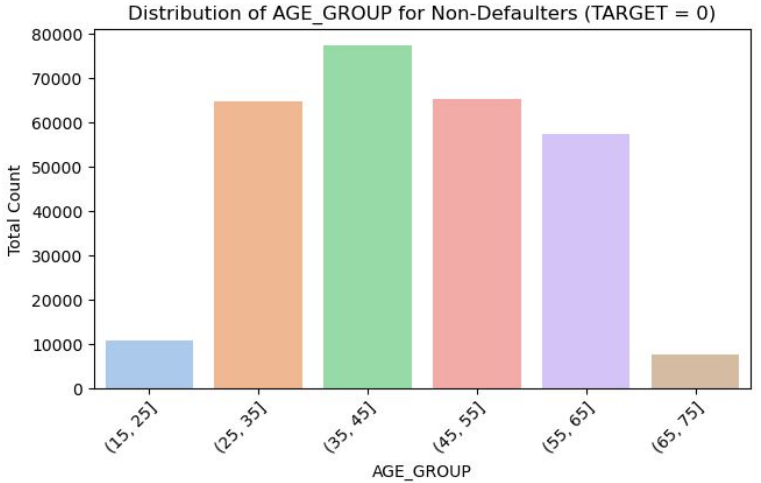


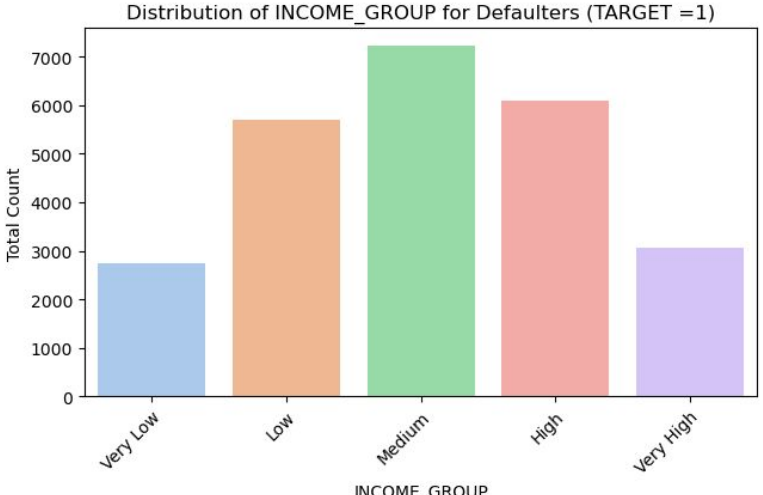


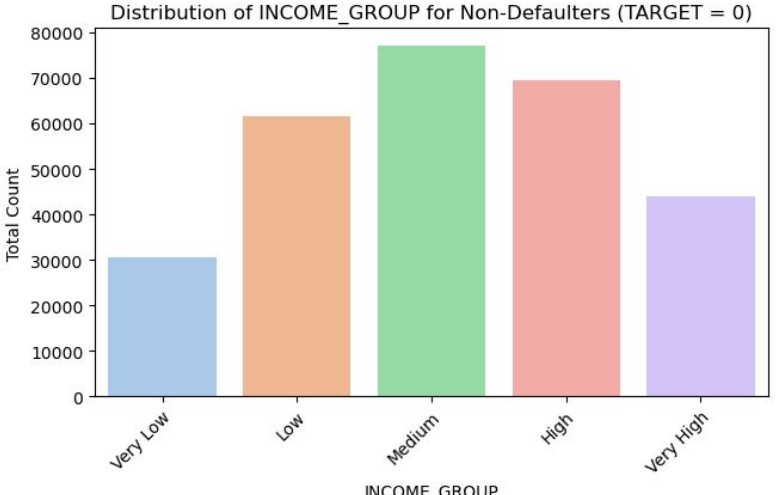
Age and Income

- Medium income range group are more defaulters as their count for loan application is also more.
- Age-group 25-35 is the most risky group to lend loan.
- 35 to 55 age group has highest loan requirement.
- Less defaults from higher age group.



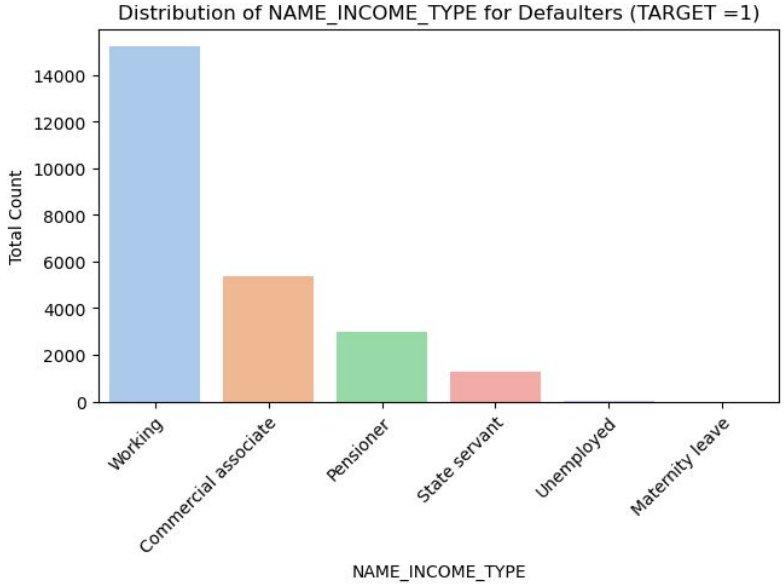


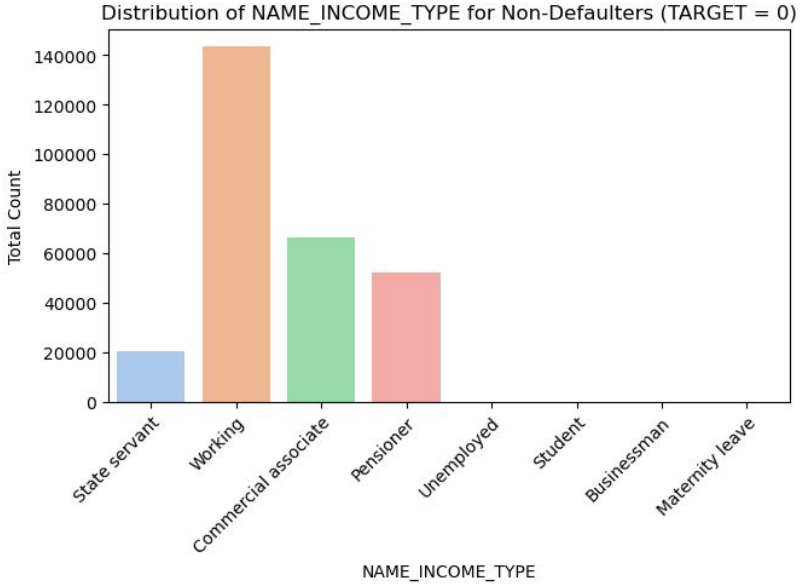




Income type:

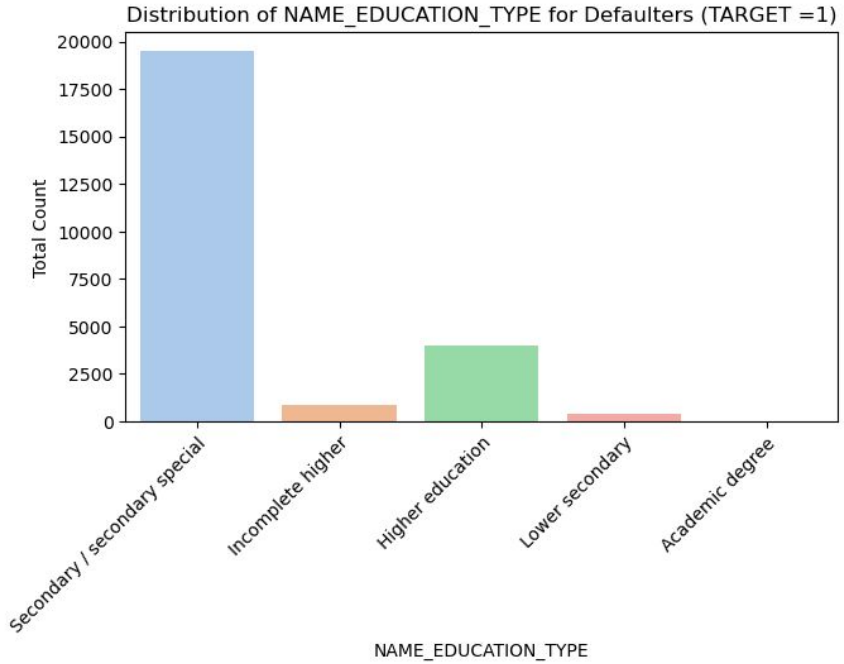
- Working category has high defaulters compared to other groups- 61.32
- Unemployed, student, maternity leave are the groups to be least bothered for considering loan.
- Businessman group applies more for loan but needs further analysis.
- State servants defaults less. (5% of total defaulters)

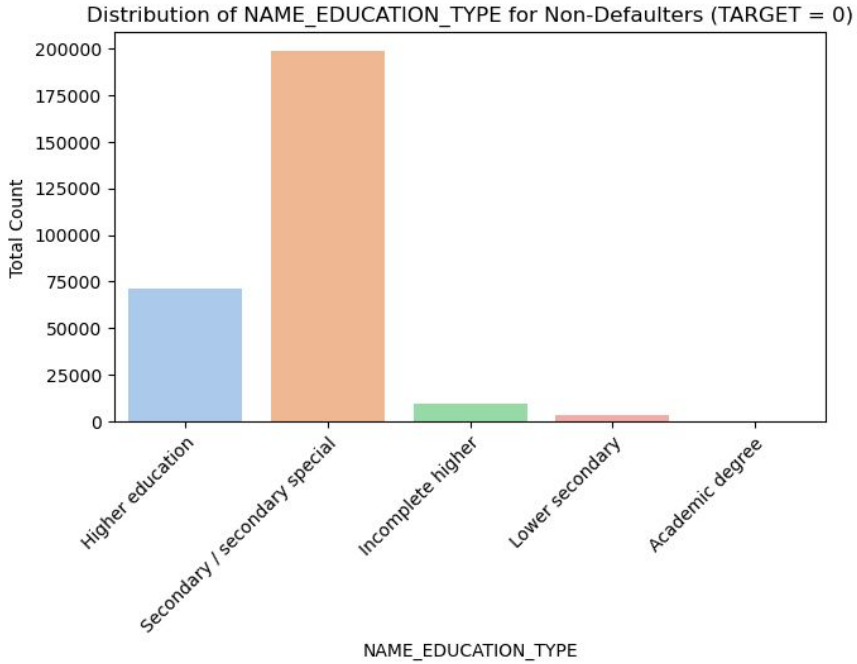




Education:

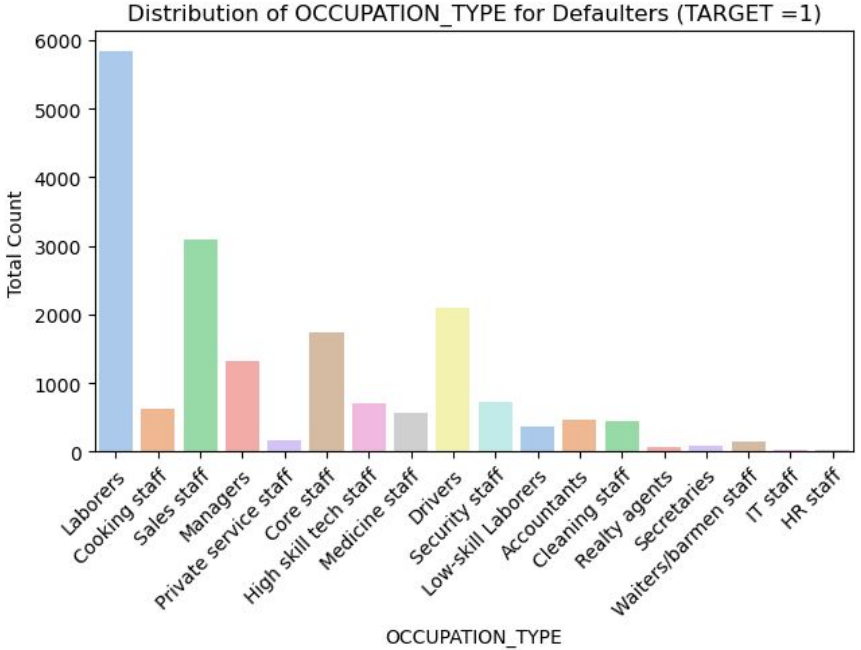
- Higher education group have least tendency to default compared to any other group
- Those who have Lower secondary, Academic degree and Incomplete higher education tends to default more.

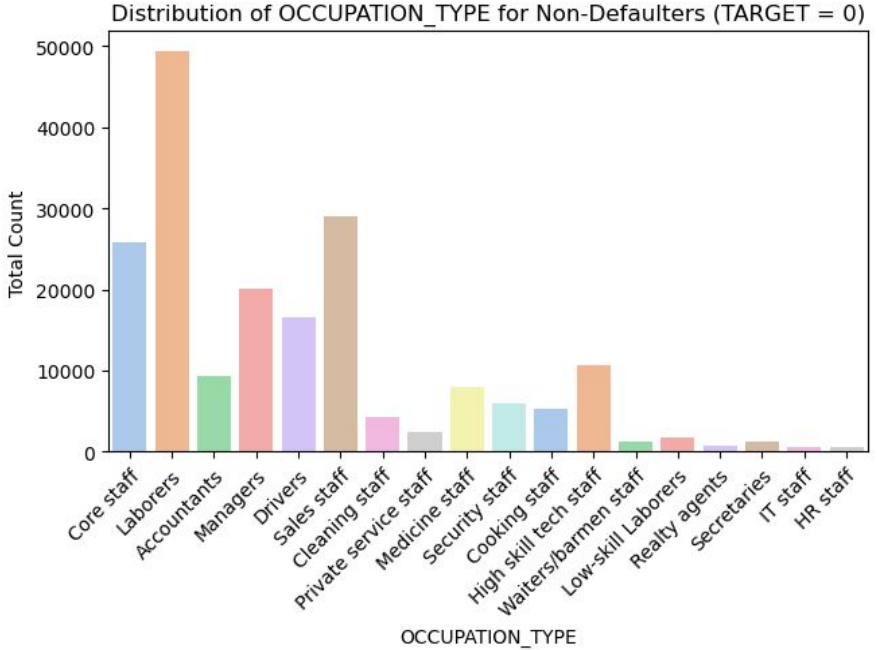




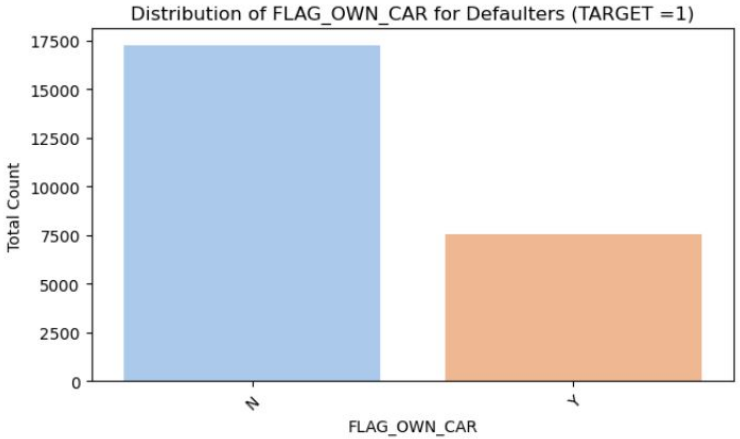
Occupation:

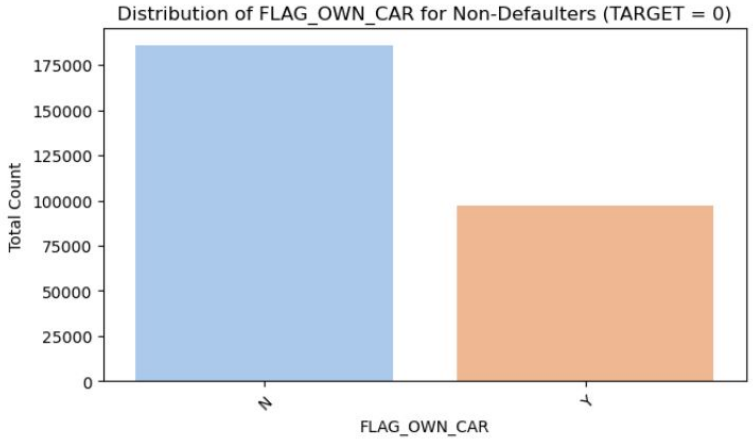
- 37.48% among all the defaulters belongs Laborers group
- Sales staff also have more defaulters- 16.68%, Drivers have 11.36% default.
- Realty agents, Secretaries, IT and HR staff are least defaulters, possible reason being high income group.

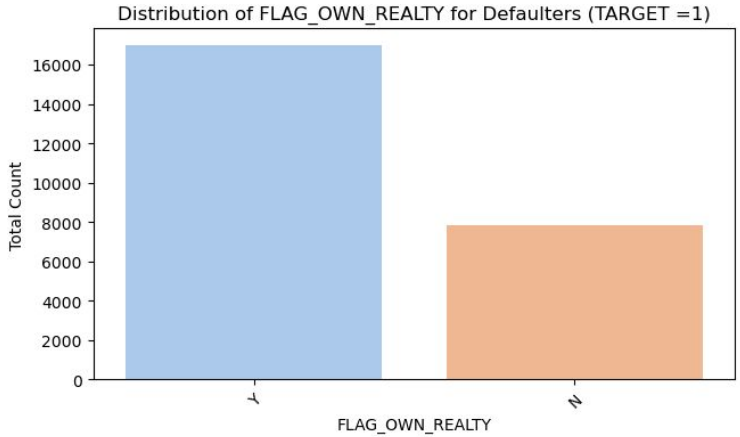




Car and Realty owners:





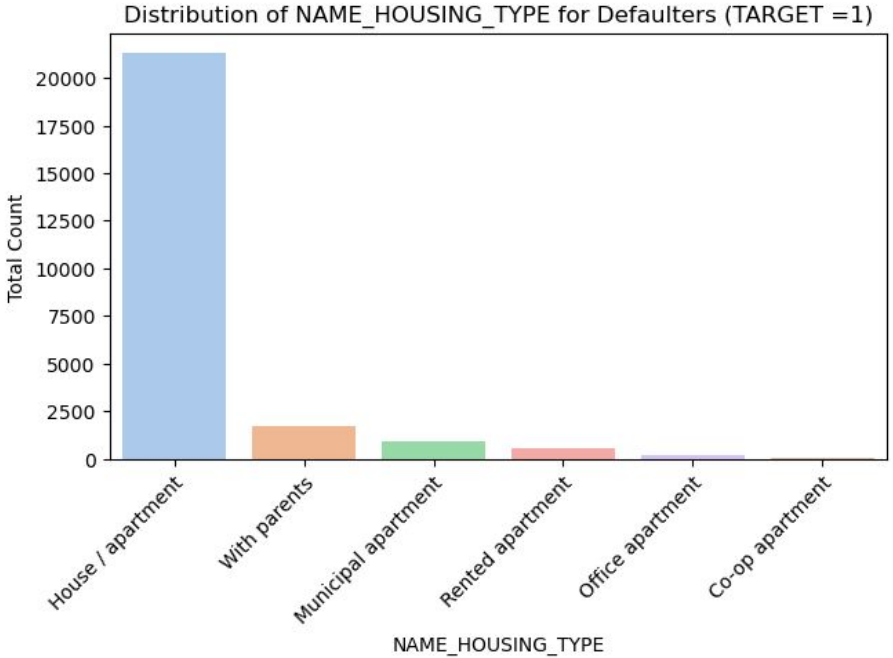


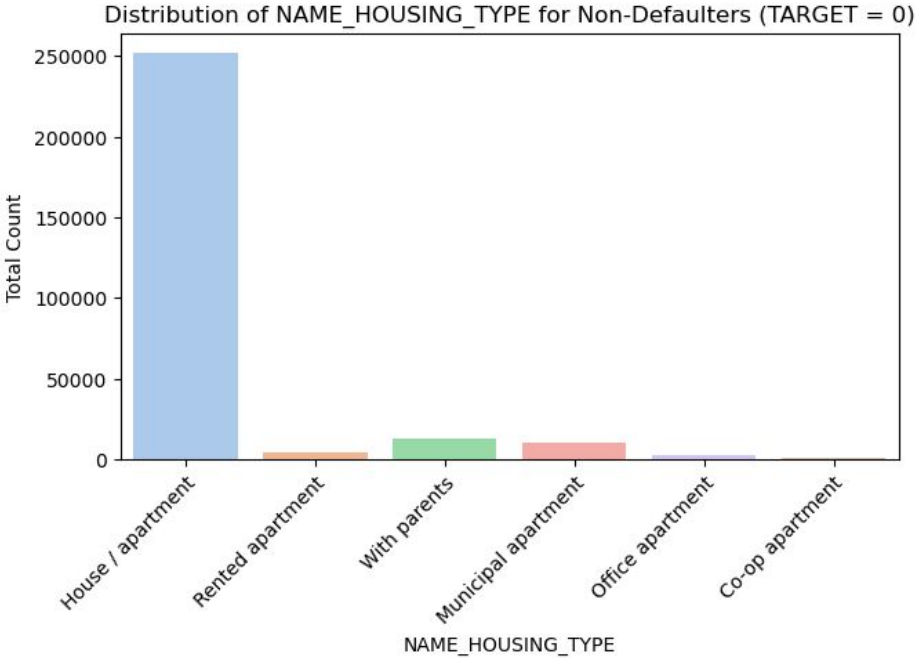
- Car owners are less defaulters. Possible reason: car comes under luxury good and someone always think of luxury items after settling down on basic needs.

- Those applicants who do not own realty are more likely to default than those who own realty.

Housing type

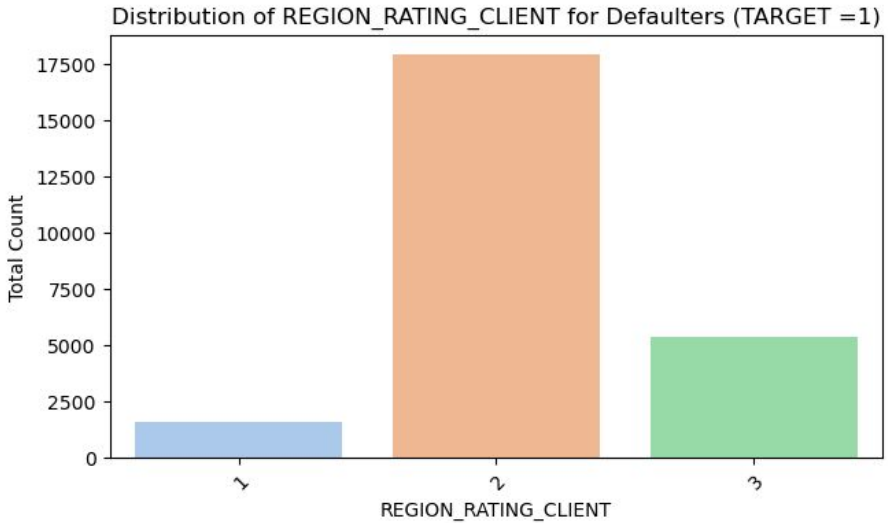
- Applicants who own House/Apartment apply for more loans.
- People living with parents tend to default more often may be because more people living in house is proportional to more expenses.

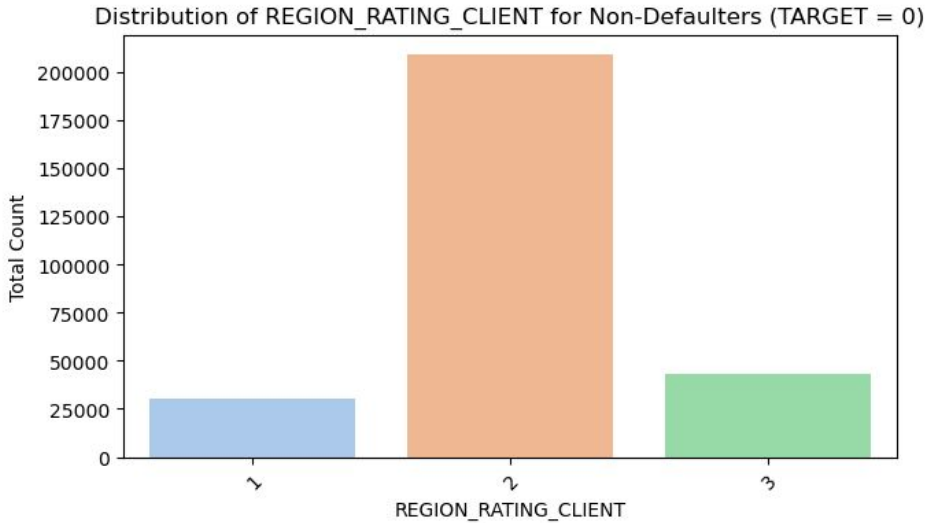




Region wise Client rating

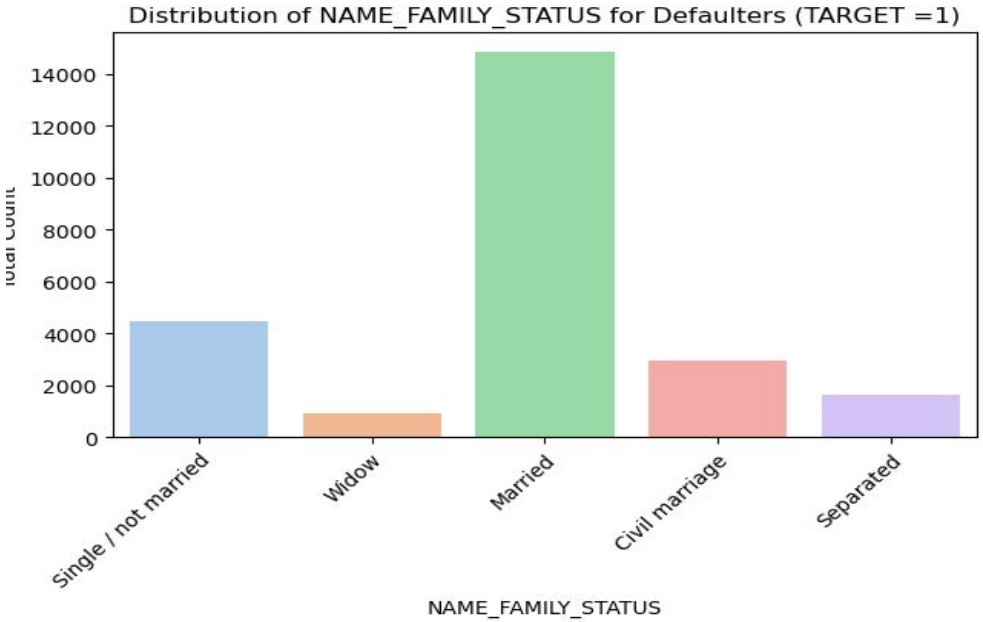
- Clients belonging to region with Rating of 2 are more likely to apply for loan and are low defaulters.
- Clients with region rating 3 are more defaulters.

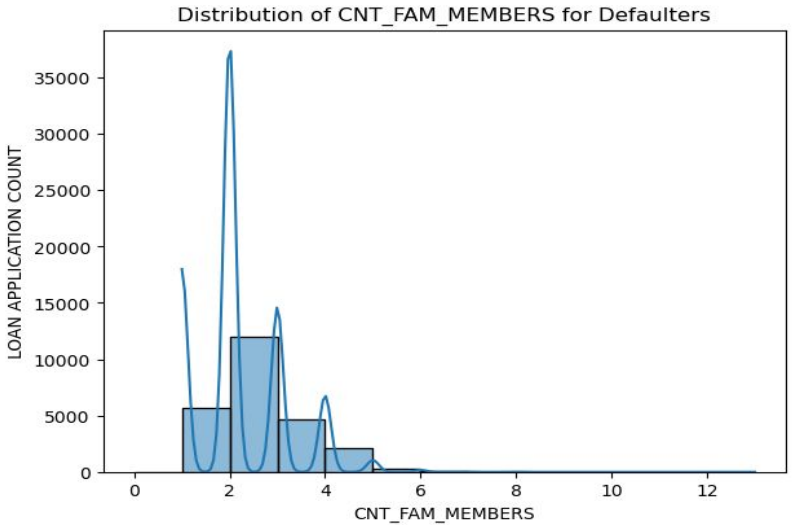


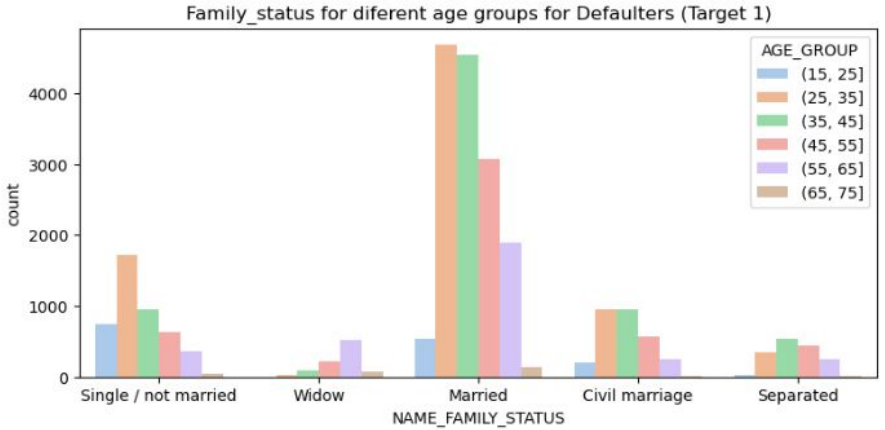


Family status and number of family members:

- Married clients belonging from 25 to 45 age group have more payment difficulties and tend more towards defaulters.
- Single/not married and widow are risky group
- Family with 2 to 3 members are applying more for loan. Also they tend to default more.
- As the number of family members increases their default rate increases.







Top Correlations:

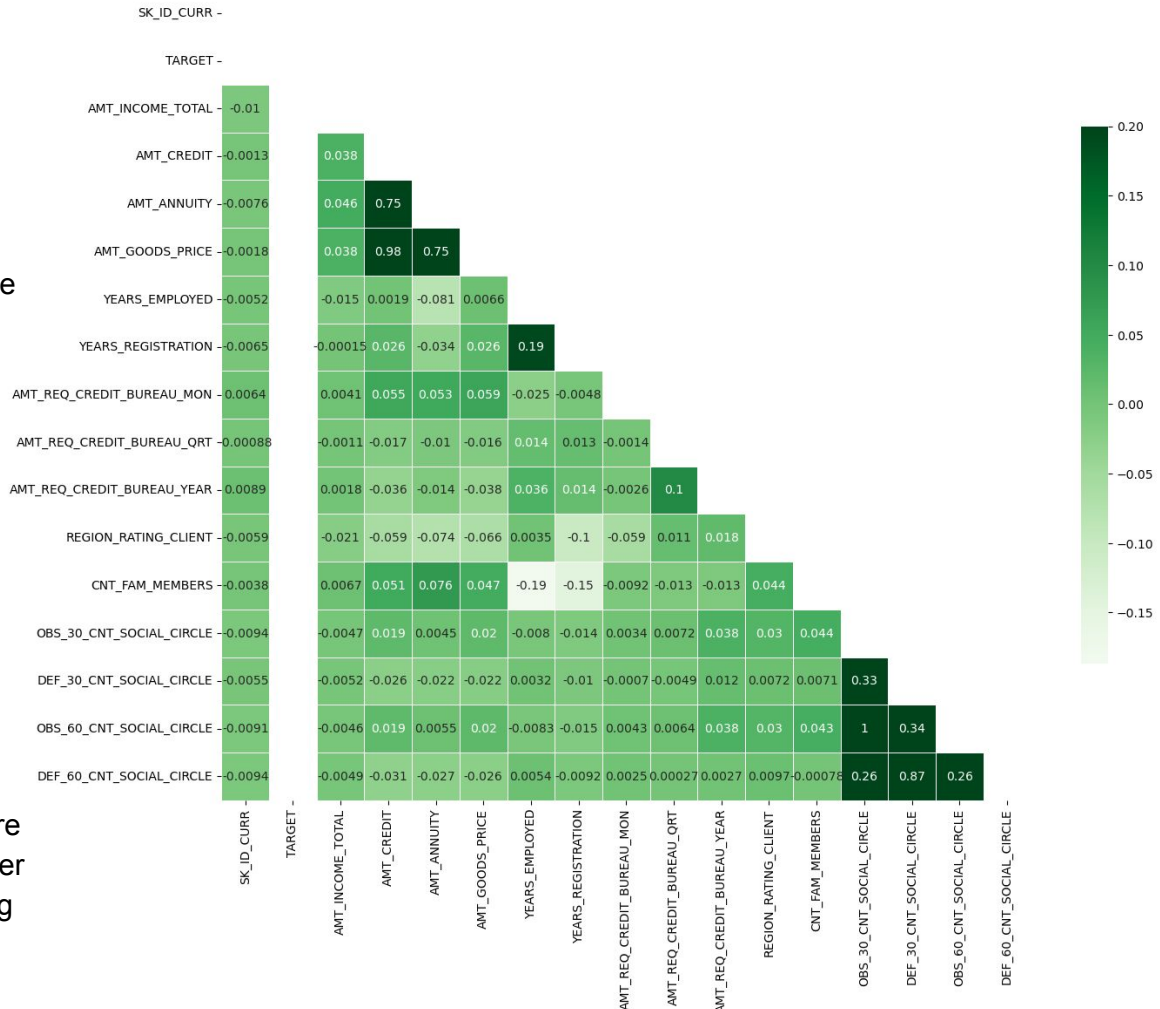
- OBS_60_CNT_SOCIAL_CIRCLE and OBS_30_CNT_SOCIAL_CIRCLE have highest correlation i,e 0.99

- AMT_GOODS_PRICE and AMT_CREDIT also have high correlation in them i,e 0.98

- 3. AMT_ANNUITY and AMT_CREDIT show good correlation

- DEF_60_CNT_SOCIAL_CIRCLE and DEF_30_CNT_SOCIAL_CIRCLE

Note: The high correlation means these variables are strongly associated or related to each other. However their causation needs to be verified before predicting anything.



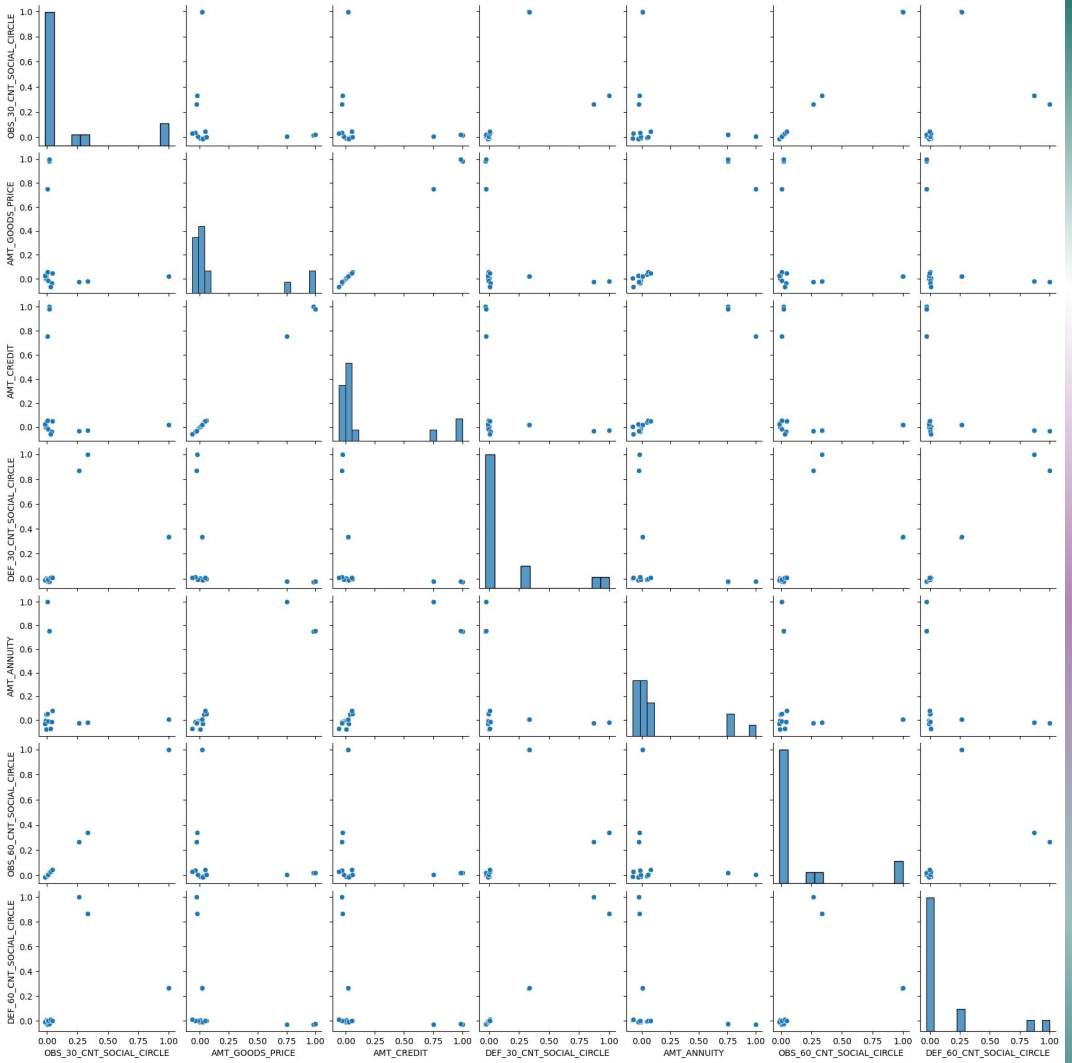
Pattern in correlation

- AMT_GOODS_PRICE and AMT_CREDIT and AMT_ANNUITY shows positive trend. So would be okay to assume that high loan is required for high goods value. AMT_GOODS_PRICE and AMT_CREDIT are proportional to each other.

- OBS_30_CNT_SOCIAL_CIRCLE and OBS_60_CNT_SOCIAL_CIRCLE are strongly related. This seems to be one of the parameters to be considered. Those who are likely more to be more defaulters in 30 DPD are not necessarily but likely to be defaulters for 60 DPD.

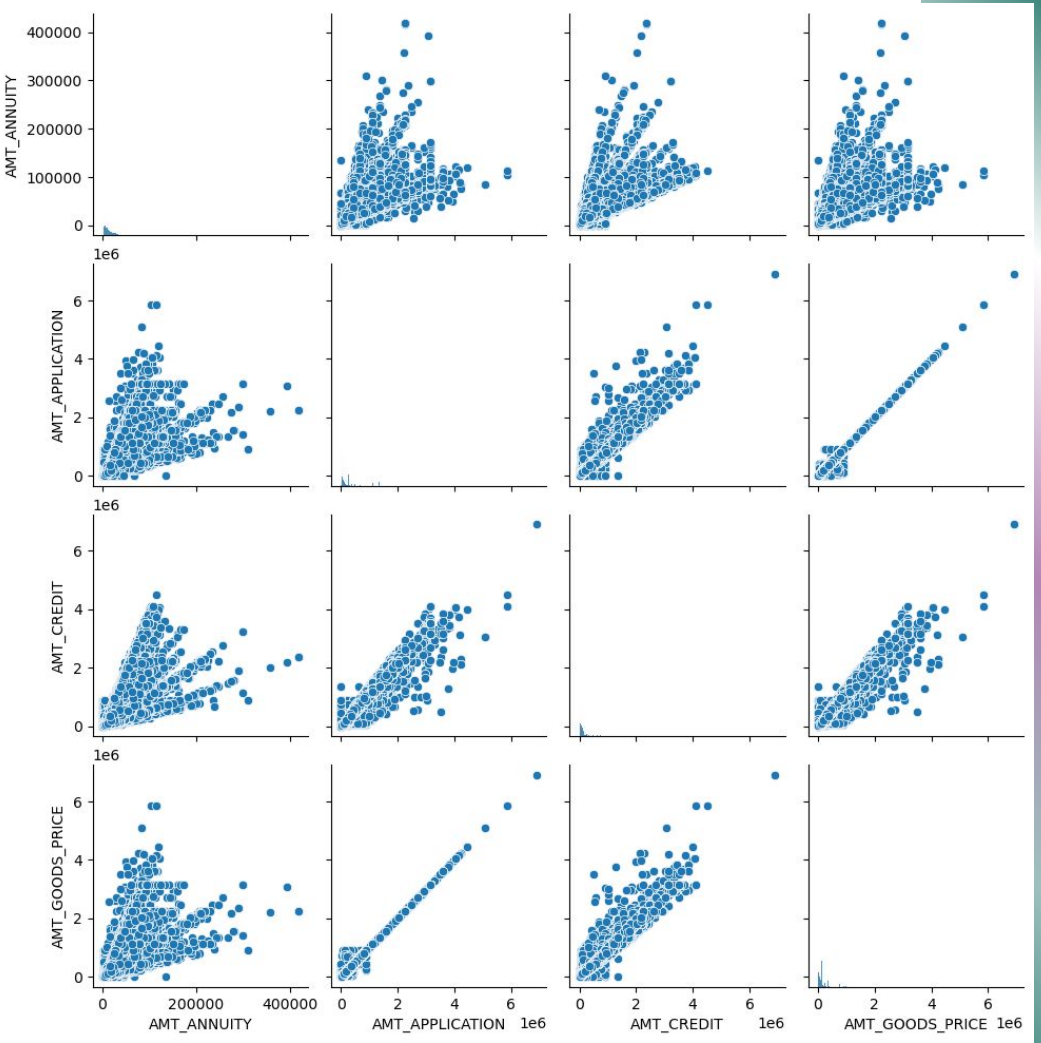
- DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE also shows positive correlation. However they doesn't seem to have any relation with AMT_GOODS_PRICE and AMT_CREDIT.

Majorly banks need the surrounding information for examining if more loans required in same locality or if any other loan requirement is there. To keep track of repayment of loan as well



Bivariate analysis on continuous columns: 'AMT_ANNUITY','AMT_APPLICATION', 'AMT_CREDIT', 'AMT_GOODS_PRICE'

- 'AMT_ANNUITY','AMT_APPLICATION', 'AMT_CREDIT', 'AMT_GOODS_PRICE' have good correlation in them.
- All the parameters follow positive trend.
- Majority of the applicants belong to amount less than 20 Lakhs.
- As the AMT_CREDIT increase beyond 2 Lakhs the defaulter rate increases.



Inference:

Driving variables for Defaulters	Reasoning
❖ CODE_GENDER:	Males are relatively more Defaulters.
❖ INCOME_GROUP	Males with medium and high income group defaults more.
❖ DAYS_BIRTH	Age group of 25 to 35 is the most risky group to grant loan.
❖ NAME_INCOME_TYPE	➔ Working category (61.32% of all defaults) has high defaulters compared to other groups. ➔ Unemployed, student, maternity leave at higher risk of default.
❖ NAME_EDUCATION_TYPE	Lower secondary, Academic degree and Incomplete higher education tends to default more.
❖ OCCUPATION_TYPE	➔ 37.48% among all the defaulters belongs Laborers group ➔ Sales staff also have more defaulters- 16.68%, Drivers have 11.36% high default rate.
❖ REGION_RATING_CLIENT	Rating 3 has highest default.
❖ NAME_FAMILY_STATUS	➔ Married clients belonging from 25 to 45 age group have more payment difficulties and tend more towards defaulters. ➔ Single/not married and widow are risky group

Cont...

Driving variables for Defaulters	Reasoning
❖ CNT_FAMILY_MEMBERS	Family with 2 to 3 members have more defaults. After 6 family members the rate of default increases to 100%
❖ AMT_CREDIT	As the AMT_CREDIT increase beyond 2 Lakhs the defaulter rate increases.
❖ AMT_GOODS_PRICE	The defaulter rate increases if AMT_CREDIT goes beyond 5 Lakhs.

Some more supportive parameters which can be taken into consideration are:

Supporting variables	Reasoning
❖ FLAG_OWN_CAR	Car owners are less defaulters.
❖ NAME_HOUSING_TYPE	As the AMT_CREDIT increase beyond 2 Lakhs the defaulter rate increases.
❖ FLAG_OWN_REALTY	Realty owners are less defaulters.
❖ OBS_30_CNT_SOCIAL_CIR	If the observations for social surroundings are more then client may also needs to be verified

Conclusion

- Females with higher income and education to be given more weightage when granting loans.
- Males with higher income should be given weightage
- State servants very less defaults and can be considered as safe to grant loan.
- Unused loan amount have lower applications which needs verification to identify underlying reason.
- Introducing more revised offers to higher-income group will be beneficial as they tend to default less.
- Refreshed and widow who have unused status earlier needs verification.

Thank you.

Credits:

- Kaggle: Your Machine Learning and Data Science Community
- Matplotlib — Visualization with Python
- Learn Python Programming - Python Tutorial (pythonbasics.org)
- Free Google Slides themes and Powerpoint templates | Slidesgo

Note: The coding platform used is Jupyter Notebook