CREDIT ANALYTICS IN BANKING AND FINANCIAL SERVICES

Objective

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.

In other words, the company wants to understand the **driving factors** (or driver variables) behind **loan default**, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its **portfolio and risk assessment**.

-By Rishabh
Zanwar

EXPLORATORY DATA ANALYSIS (EDA):

Steps and Best Practices:

- Step 1: **Define the Problem**: The company wants to understand the driving factors behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.
- Step 2: **Collect the Data:** Gather the relevant dataset for analysis
  - ➢application_data.csv- It contains all the information of the client at the time of application.
  - ➢previous_application- It contains information about the client's previous loan data.
- Step 3: **Load the Data:** Loading the dataset into Jupyter Notebook for analysis and examine its structure.
- Step 4: **Understand the Data:** Check data types, missing values, sanity checks, and summarize dataset using descriptive statistics.
- Step 5: **Identify Outliers:** Use visualizations and statistical methods to identify and handle outliers.
- Step 6: **Explore Relationships Between Variables:** Use visualizations to explore relationships between variables, both numerical and categorical using univariate, segmented univariate, bivariate analysis.
- Step 7: **Summarize Findings & Make Recommendations:** Summarize key insights and findings from the EDA process and make recommendations or decisions that can help address the problem or achieve the objectives of the analysis.

UNDERSTAND THE DATA

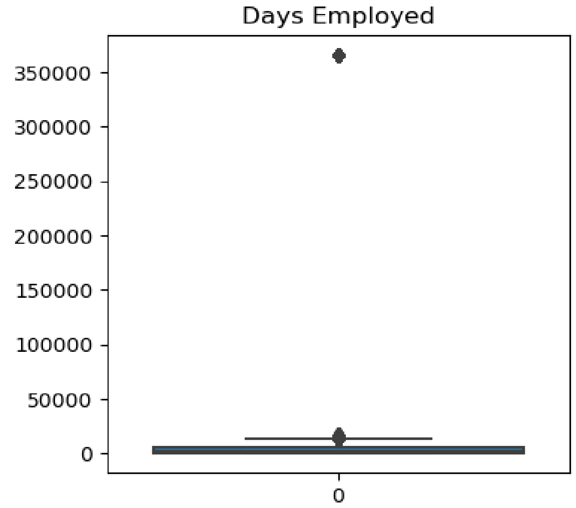
- application_data.csv- It contains 122 columns and 307511 rows with data types float, int, object.
  - ➢ Data Handling and Cleaning
    - ▪ Handling Missing Values:
      - ✓ After checking the data, it was found that the missing values range from 0.00 to 69.87% (in percentage). Lets drop the column having missing value more than 40%.
      - ✓ Remaining column , missing values will be imputed using mean, median, or mode as appropriate.
      - ✓ Certain columns such as 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', and 'DAYS_LAST_PHONE_CHANGE' contain negative values. The absolute function was applied to convert them to positive values.
      - ✓ Additionally, some columns contain extremely high values, such as 'AMT_INCOME_TOTAL’, 'AMT_CREDIT_RANGE’. To facilitate analysis, these columns were binned using appropriate methods.
- previous_application- It contains 37 columns and 1670214 rows with data types float, int, object.
  - ➢ Data Handling and Cleaning
    - ▪ Handling Missing Values:
      - ✓ After checking the data, it was found that the missing values range from 0.00 to 69.87% (in percentage). Lets drop the column having missing value more than 40%.
      - ✓ Remaining column , missing values will be imputed using mean, median, or mode as appropriate.

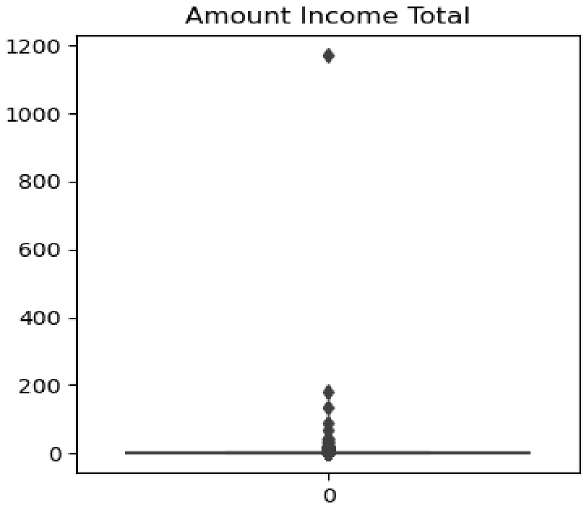
OUTLIERS ANALYSIS USING BOXPLOT

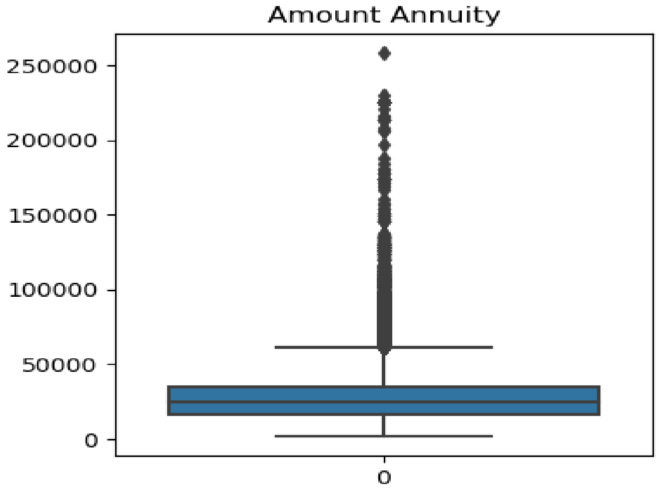
From the dataset, we observe that there is a significant difference between the maximum values and the 75th percentile for columns.

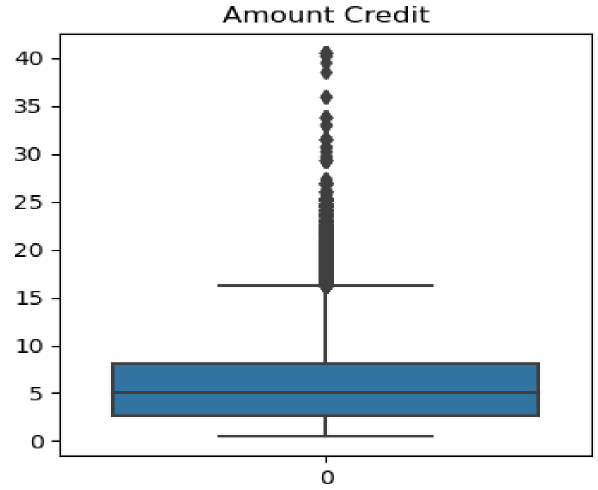
Insight-

- In the **'DAYS_EMPLOYED'** column, an outlier value of 365243 was identified, which corresponds to approximately 1000 years. Such a value is clearly unrealistic and indicates a data entry error or anomaly.
- The **'AMT_INCOME_TOTAL'** column also exhibits a substantial number of outliers. This suggests that some loan applicants have significantly higher incomes compared to the rest of the dataset. While doing risk assessment we have to check on this.
- **AMT_ANNUITY, AMT_CREDIT** also contain Outliers.
- There is a possibility of outliers in the **'CNT_CHILDREN'** column we need to further analyze.



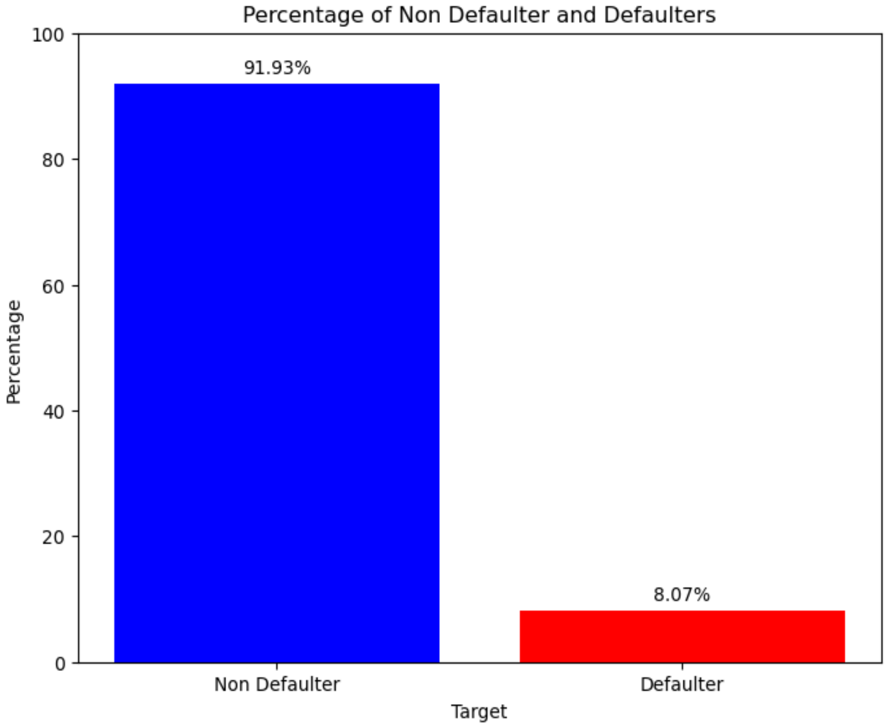






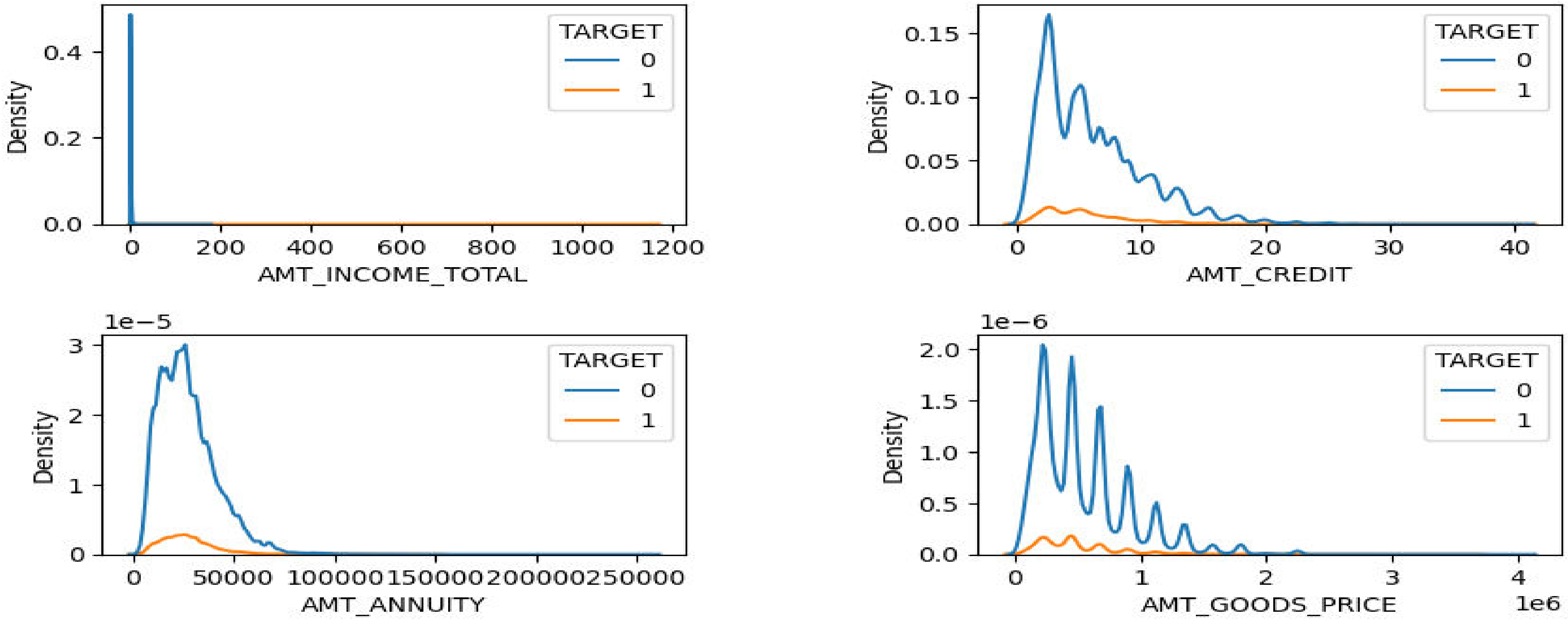
DATA IMBALANCE

- The majority of loan applicants in the dataset are Non Defaulter, constituting 91.93 % of the total.
- Only a small proportion of applicants are Defaulters, accounting for 8.07% of the dataset.
- The significant imbalance between Non Defaulter and Defaulters (11.39:1) indicates an uneven distribution, with Non Defaulter far outnumbering defaulters.



NUMERICAL UNIVARIATE ANALYSIS

- A majority of loans are granted for goods priced below 10 lakhs.
- The majority of individuals pay annuities below 50,000 for their credit loans.
- The credit amount of loans is predominantly less than 10 lakhs.

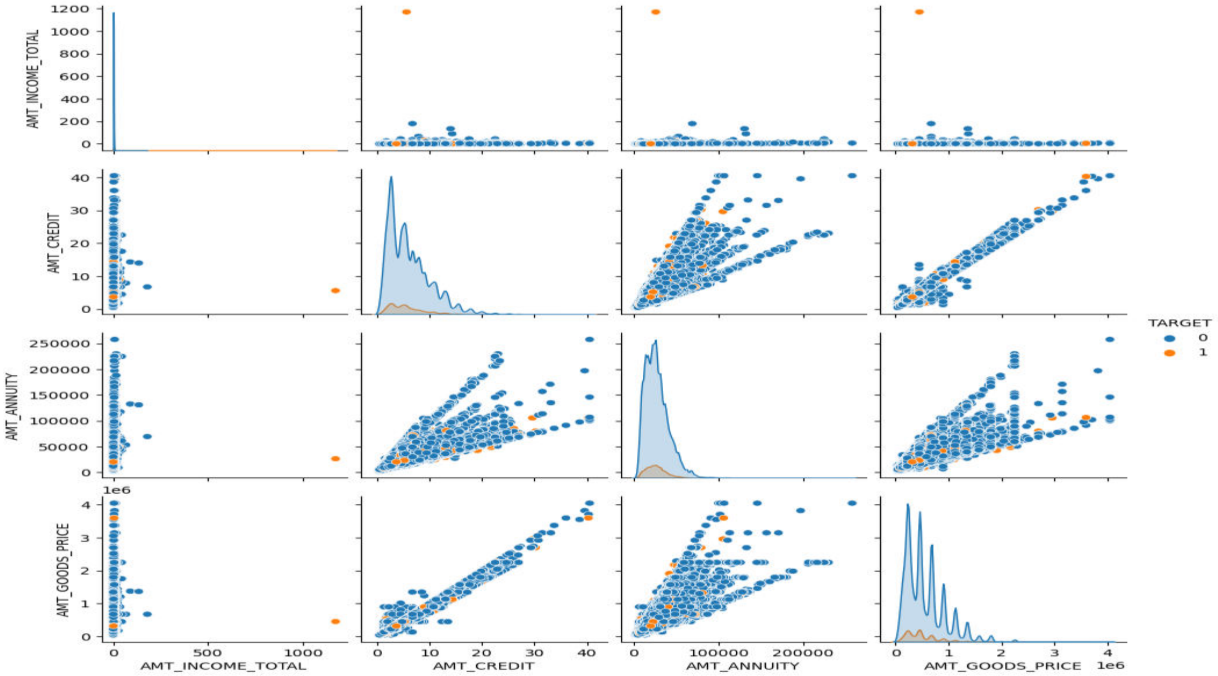


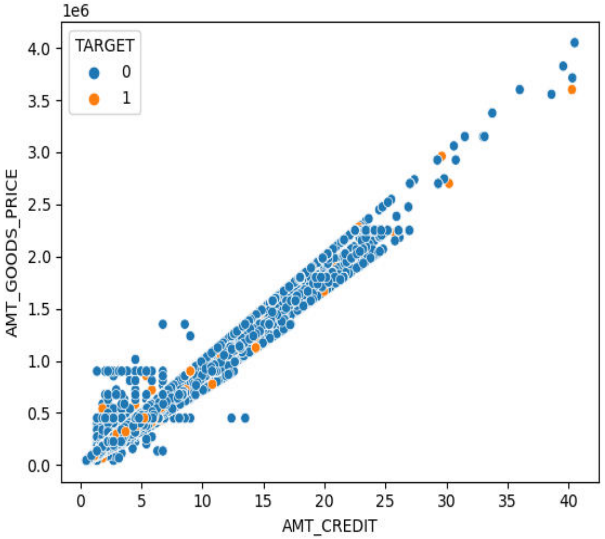
NUMERICAL BIVARIATE ANALYSIS

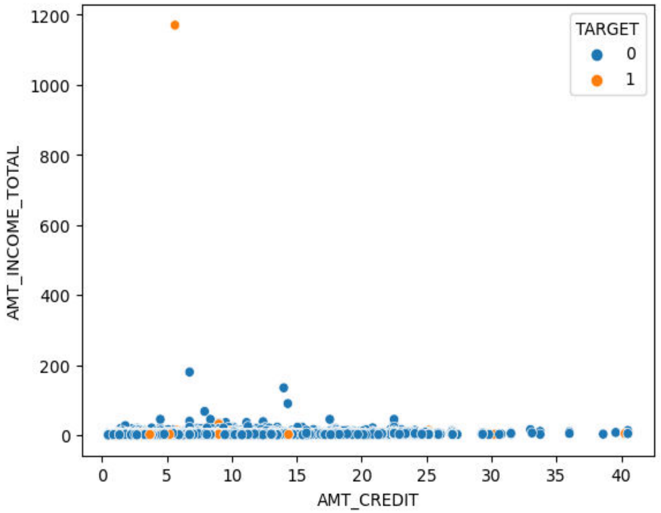
- AMT_CREDIT and AMT_GOODS_PRICE are linearly corelated, furthermore as AMT_CREDIT increases the defaulters are decreasing.

- AMT_CREDIT and AMT_INCOME_TOTAL Individuals with higher incomes, particularly those earning more than 1 million, are less likely to take out loans. Additionally, when the 'AMT_CREDIT' exceeds 1.5, the number of defaulters decreases.

- Note- Pairplot picture is not clear to visible Pls ref Jupyter file

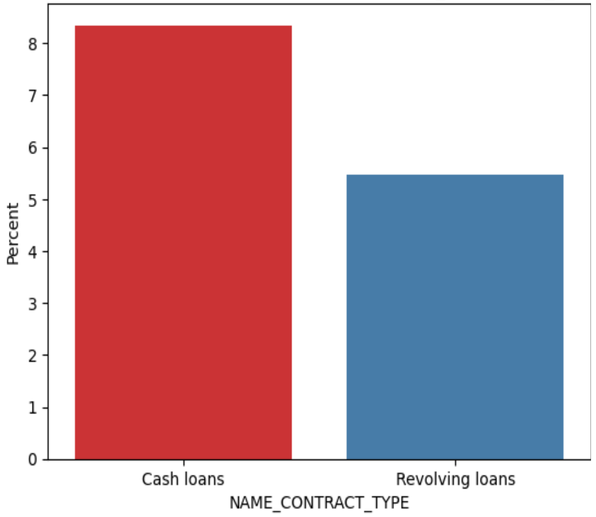


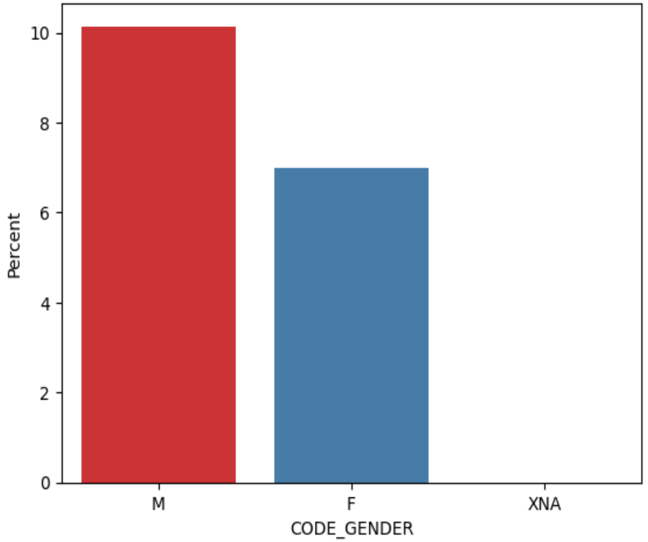


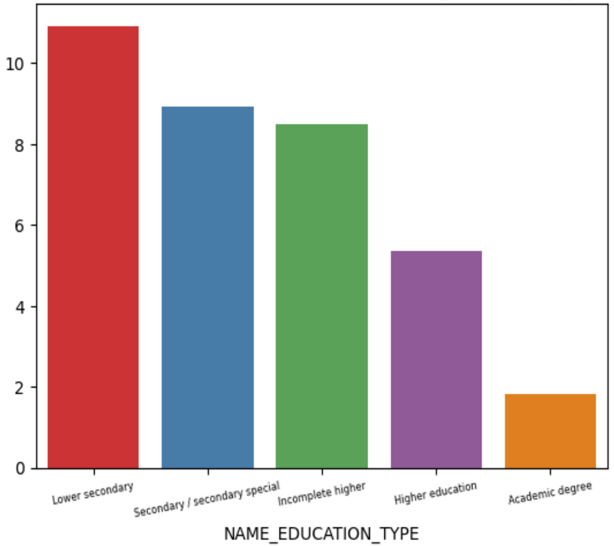


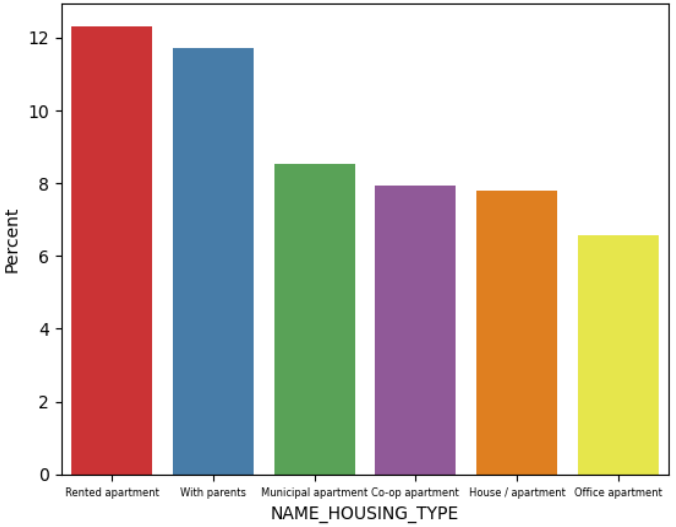
SEGMENTED UNIVARIATE ANALYSIS

- Approximately 8-9% of cash loan applicants and 5-6% of revolving loan applicants are defaulters.
- Males have a higher default rate of approximately 10% compared to females, who have a default rate of about 7%.
- Individuals in the lower secondary education category exhibit the highest default rate, at around 11%, while those with academic degrees are the least likely to default on their loans.
- Individuals living with their parents have a default rate of approximately 11.5%, while those living in rented apartments have a default rate exceeding 12%.


Defaulter Percentage with respect to NAME_CONTRACT_TYPI


Defaulter Percentage with respect to CODE_GENDER


Defaulter Percentage with respect to NAME_EDUCATION_TYPI

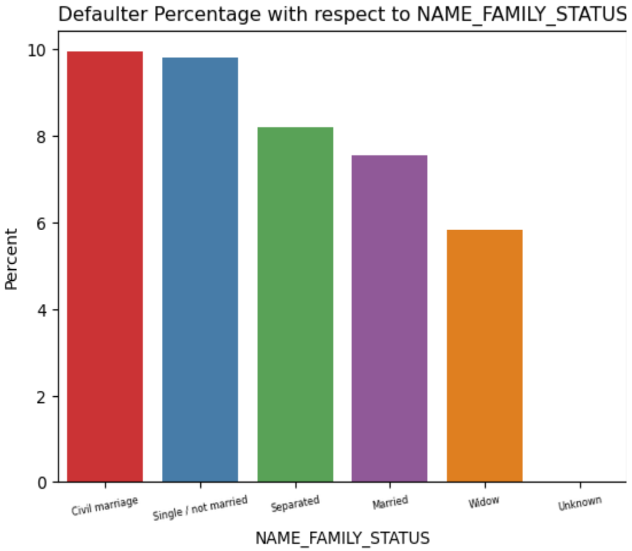

Defaulter Percentage with respect to NAME_HOUSING_TYPE

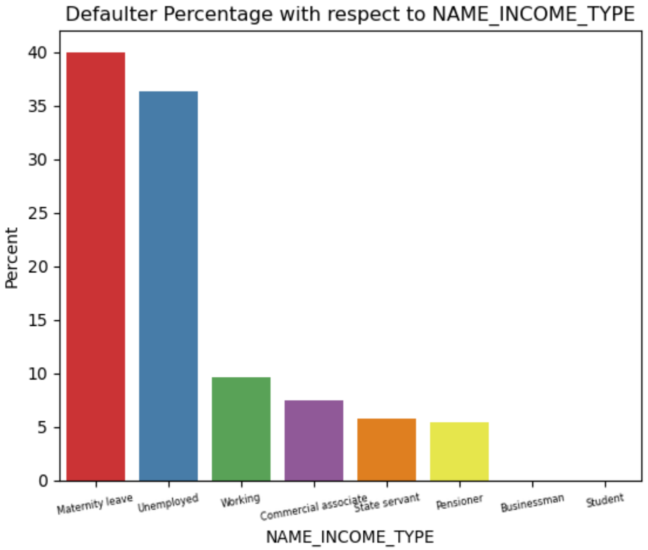
SEGMENTED UNIVARIATE ANALYSIS

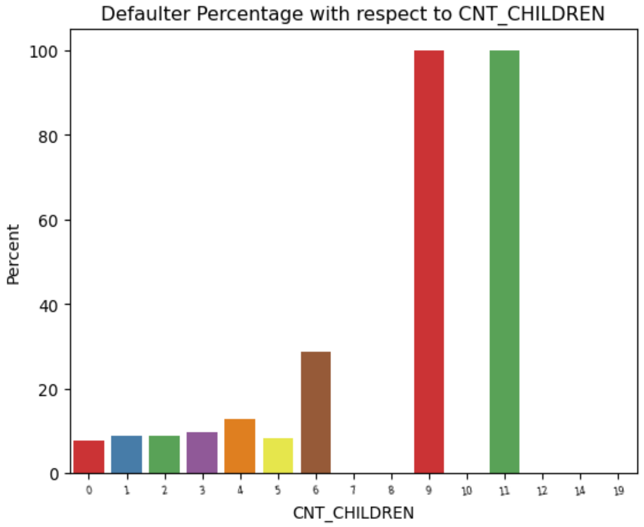
- Civil marriage is associated with the highest percentage of defaulters, approximately 10%, while widows have the lowest percentage, around 6%, with the exception being the "Unknown" category

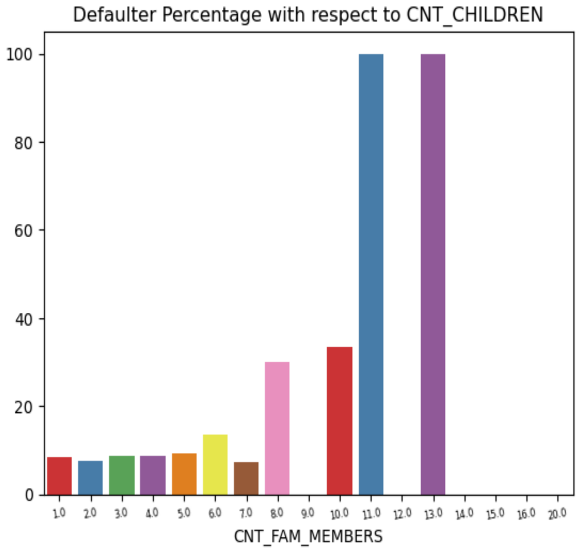
- Applicants on maternity leave and those who are unemployed have the highest defaulting percentages, at 40% and 37% respectively. Other categories fall below the average default rate of around 10%.

- Family member having more than 8 members and children having more than 6 increases the risk of defaulting.


Defaulter Percentage with respect to NAME_FAMILY_STATUS

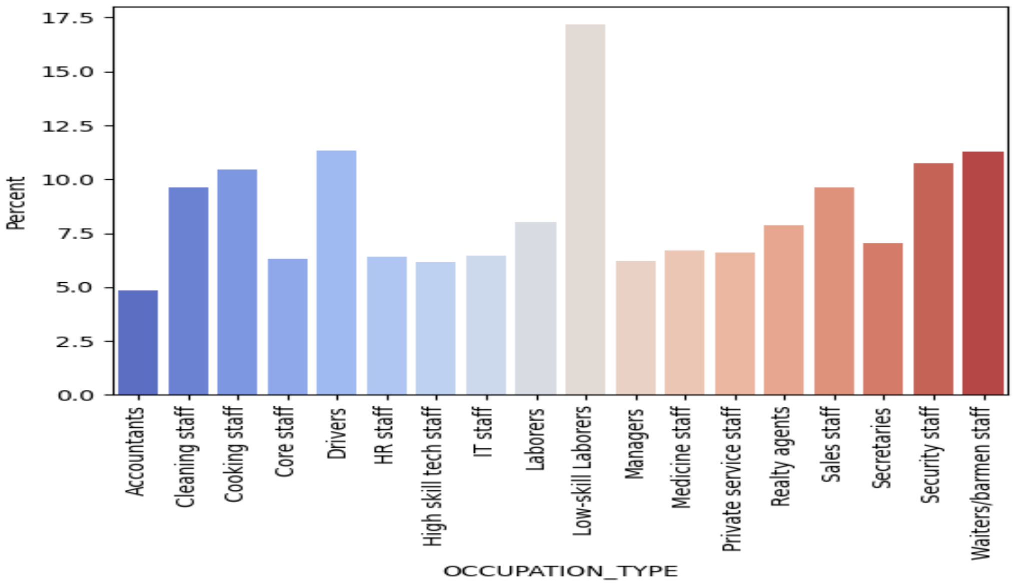

Defaulter Percentage with respect to NAME_INCOME_TYPE

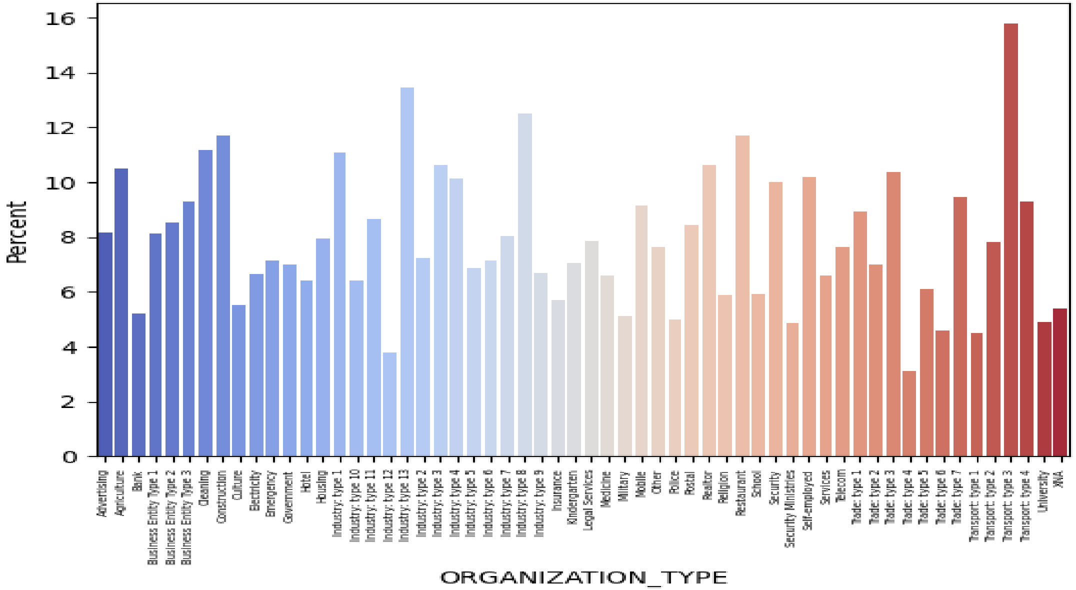

Defaulter Percentage with respect to CNT_CHILDREN


Defaulter Percentage with respect to CNT_CHILDREN

SEGMENTED UNIVARIATE ANALYSIS

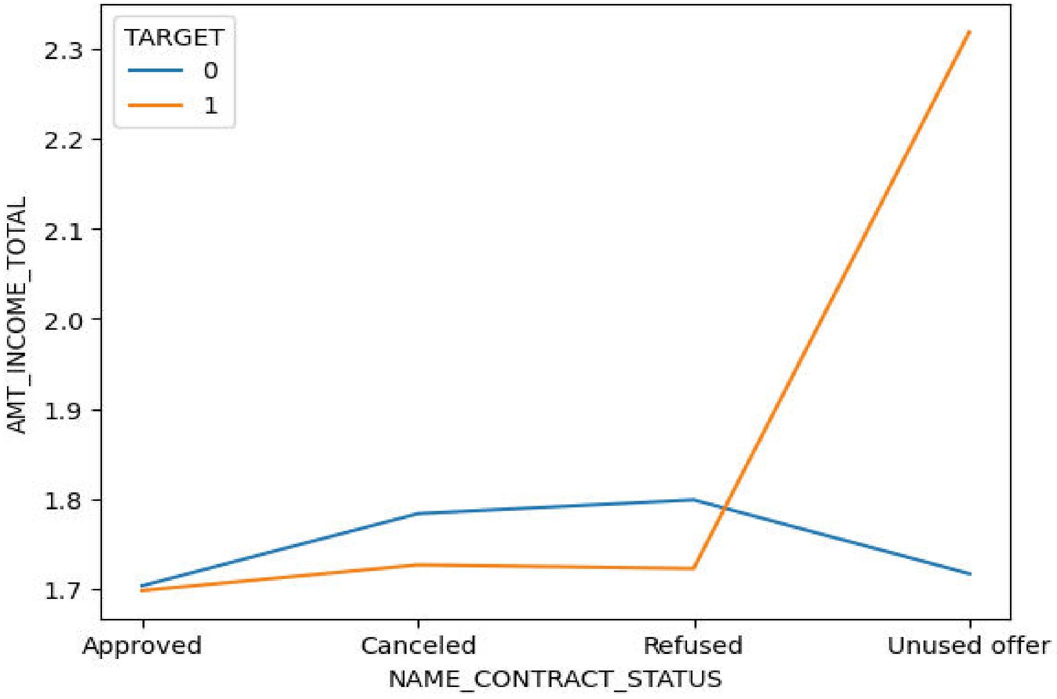
- The category with the highest percentage of defaulters is low-skill laborers, at over 17%. This is followed by drivers, waiters/barmen staff, security staff , and cooking staff.
- The organizations with the highest percentage of defaulters are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%), and restaurants (less than 12%).

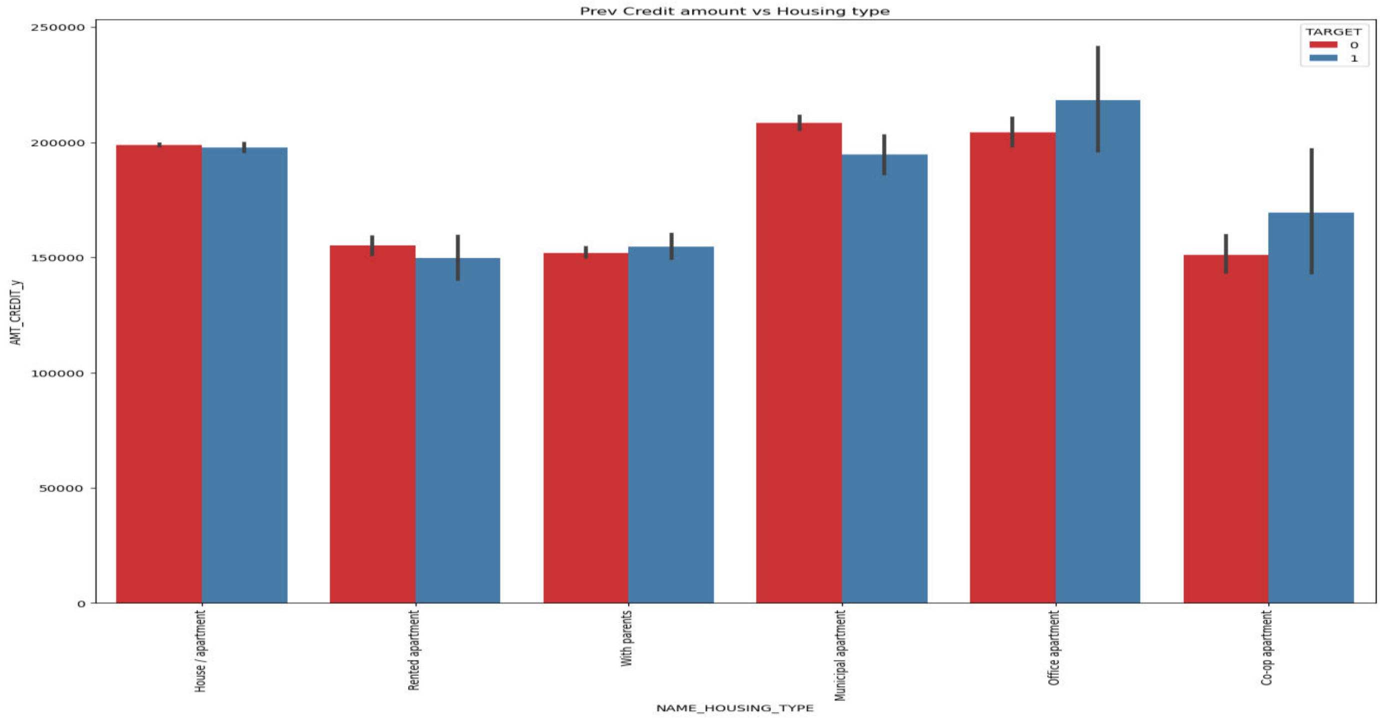




BIVARIATE ANALYSIS

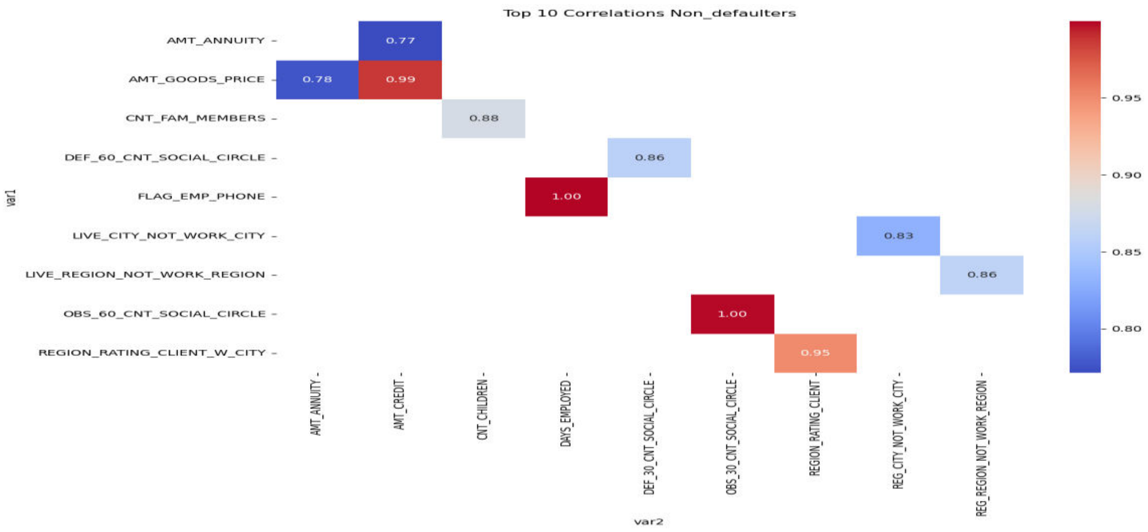
- The point plot show that the people who have not used offer earlier have defaulted even when there average income is higher than others
- Bank should avoid giving loans to the housing type of office apartment as they are having difficulties in payment. Bank can focus mostly on housing type with parents or House\ apartment or rented apartment for successful payments.

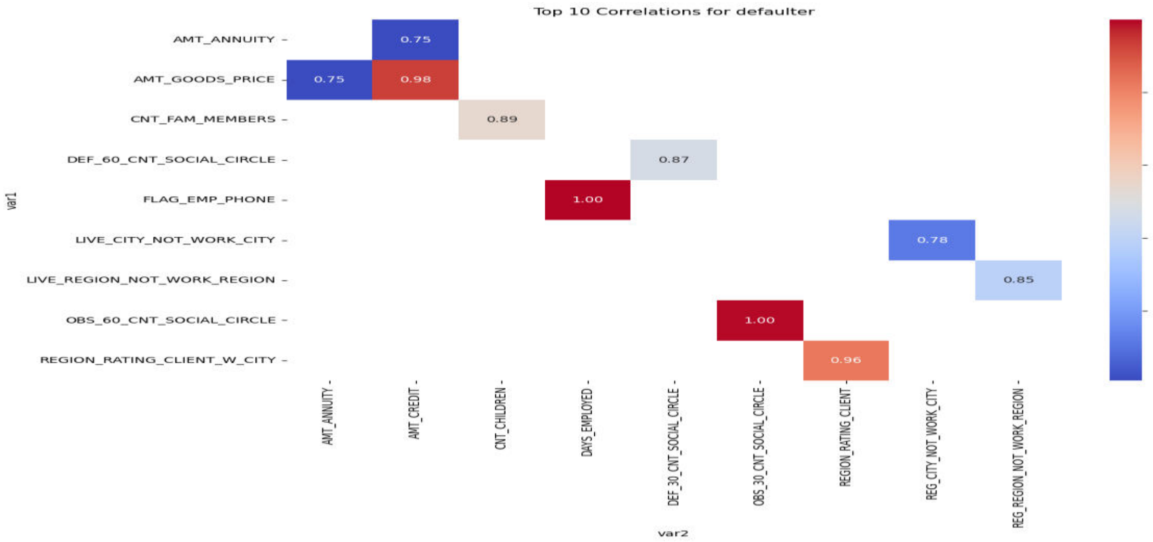



Prev Credit amount vs Housing type

TOP 10 CORRELATION FOR THE DEFAULTER AND NON DEFAULTER

- Credit amount is highly correlated with the Amount goods price, a pattern consistent among both Non defaulter and defaulters.
- The correlation between loan annuity and credit amount is slightly lower in defaulters (0.75) compared to Non defaulter (0.77).
- Among defaulters, there is a significant drop in the correlation between total client income and credit amount (0.038), whereas it is (0.342) among Non defaulter.
- The correlation between the client's age and the number of children is reduced in defaulters (0.259) compared to non defaulter (0.337).





CONCLUSIONS

- After analyzing the datasets, certain client attributes can help to identify whether they are likely to repay the loan or not. The analysis is summarized below, highlighting the contributing factors:
  - ✓The data exhibits a high level of imbalance, with 91.93% of clients experiencing non-payment difficulties and only 8.07% facing payment difficulties. The imbalance ratio is 11.39%.
  - ✓The majority of loans have been taken by females, who have a default rate of just around 7%, indicating that they are a safer bet compared to males with default rate 10%.
  - ✓Clients with secondary education are more likely to apply for loans. However, Academic degree is the safest segment for loan approval, with a default rate of less than 2%
  - ✓. The majority of loans are applied for by individuals in the married category. However, people who are single or have a civil marriage tend to default more with 10 %.
  - ✓The category with the highest percentage of defaulters is low-skill laborers, at over 17%. This is followed by drivers, waiters/barmen staff, security staff , and cooking staff.
  - ✓The organizations with the highest percentage of defaulters are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%), and restaurants (less than 12%).
  - ✓The people who have not used offer earlier have defaulted even when there average income is higher than others.
  - ✓People who get loan for 3-6 Lakhs tend to default more than others and hence having higher interest specifically for this credit range would be ideal.
  - ✓Given that 90% of the applications have a total income of less than 3 lakhs and are at a higher risk of defaulting, offering them loans with higher interest rates compared to other income categories could be considered.

RECOMMENDATION

- Based on the analysis, clients with the following characteristics are recommended for loan approval:
  - ✓Highly educated, preferably female.
  - ✓Have a house or apartment, are married, and have no more than 5 children
  - ✓Low income below 1 million
  - ✓Working in organizations such as Others, Business Entity Type 3, or self-employed
  - ✓Employed as accountants, IT staff, managers.
- Precautions to consider include avoiding organizations in Transport type 3, Industry type 13 as well as low-skill laborers. It is also advisable to avoid offers to previously unused or high-income customers.
- Excise caution when giving loan to 30-40 age group people and carefully examine their loan application as they have large number of defaulters.
- Additionally, it was observed that most of the customers who were previously canceled or refused are now Non defaulters. The bank can further analyze this segment and consider offering them loans to increase business opportunities.
- After granting a loan, if a client wishes to apply for another loan, the bank can conduct a risk analysis based on the following factors:
  - ✓Payment history, Current outstanding balances and debt, Length of time the accounts have been open
  - ✓Amount of available credit being used (credit utilization ratio)
  - ✓Presence of derogatory marks, such as debts sent to collection, foreclosures, or bankruptcies.