Credit-EDA Case Study

Prepared by
Amit

Problem Statement

Business Understanding

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

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:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
2. 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.

Aim of Case Study

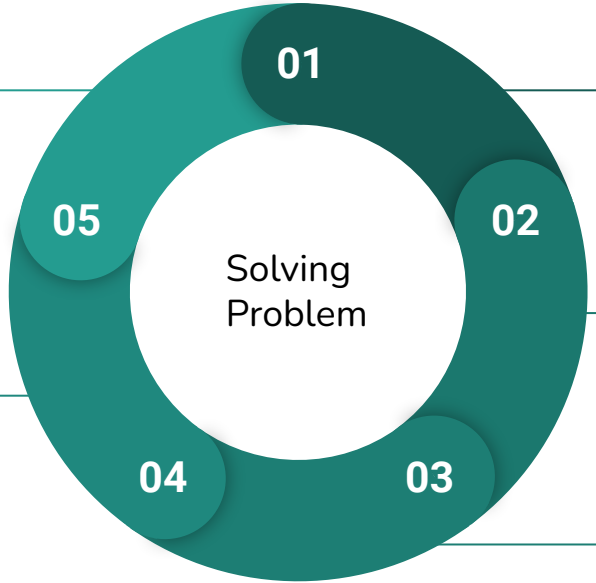
To identify patterns which indicate if a client has difficulty paying their installments, which may be used for taking actions such as

1. Denying the loan, reducing the amount of loan,
2. 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.

The company wants to understand the driving factors (or driver variables) behind loan default, The company can utilize this knowledge for its portfolio and risk assessment.

Data Analysis Approach



Understanding Business Problem

Data Understanding

Data Cleaning

Handling data irregularities such as

Missing Value, Anomalies, Standardizing the values, Filtering data

Missing Values, Outliers Analysis

Data Analysis and Getting Insights

Strategy For Data Analysis

1) Understanding the problem statement and our main aim to solve the problem
2) **Data Sourcing** from platform, then reading it in system for data analysis
3) **Understanding data** – application data, previous data, its data types and meaning from data dictionary
4) **Data Cleaning**

   a) Calculation of Missing values, and dropping the columns with threshold as 40% **assumption,** and removing unnecessary columns which are not important

   b) Segmenting dataset columns in categorical and continuous columns

   c) **Data imputation** with Mean(), Median (), Mode() in place of missing values

   d) Standardizing the values & Outliers analysis

5) **Data Analysis** – Univariate, Bivariate/Multivariate Analysis using different plots and getting insights.

Methods for Data Imputation in Missing Values

1. **Filling with Mode()** in categorical columns, when the mode() value percentage is higher than other values
2. **Filling with other values (“Missing”)** in categorical columns, when null % is high, and the value frequencies is somewhat same
3. **Filling with Mean() and Median() in continuous columns**

   Fill with Mean() when the data is normally distributed, meaning when mean and median are equally the same.

   Fill with Median() when the data distribution is skewed on histogram.

Standardizing the Values

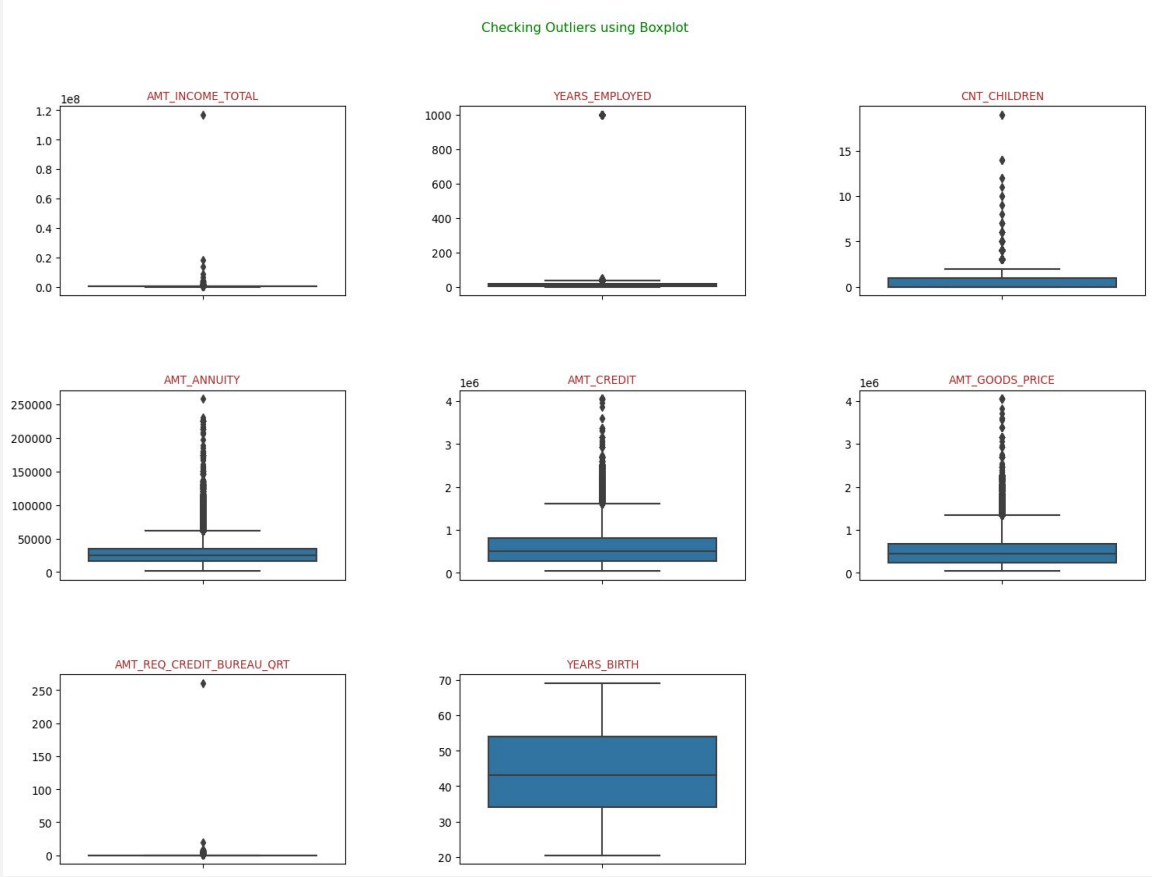
1. Adding new Features (columns) in dataset to understand data better
2. Adding categorical features for continuous columns called **binning.**
3. When numerical data has high range (breaking them in **small buckets**) to understand segments
4. Changing the format of data, like converting Days – ve values to +ve values.
5. Changing Column names to better ones.

Outliers Analysis using Boxplot

There are outliers which hinders business decisions.

Some outliers from the plot as

1. children count value as 19
2. Employment years **1000 years.**
3. **Total income as 117M,** an outlier
   for the given dataset
4. There are outliers, but they are actual data points, only their frequency is low



Data Imbalance

- Reflects an unequal distribution of classes in dataset
- Data imbalance is in the context of Target variable only

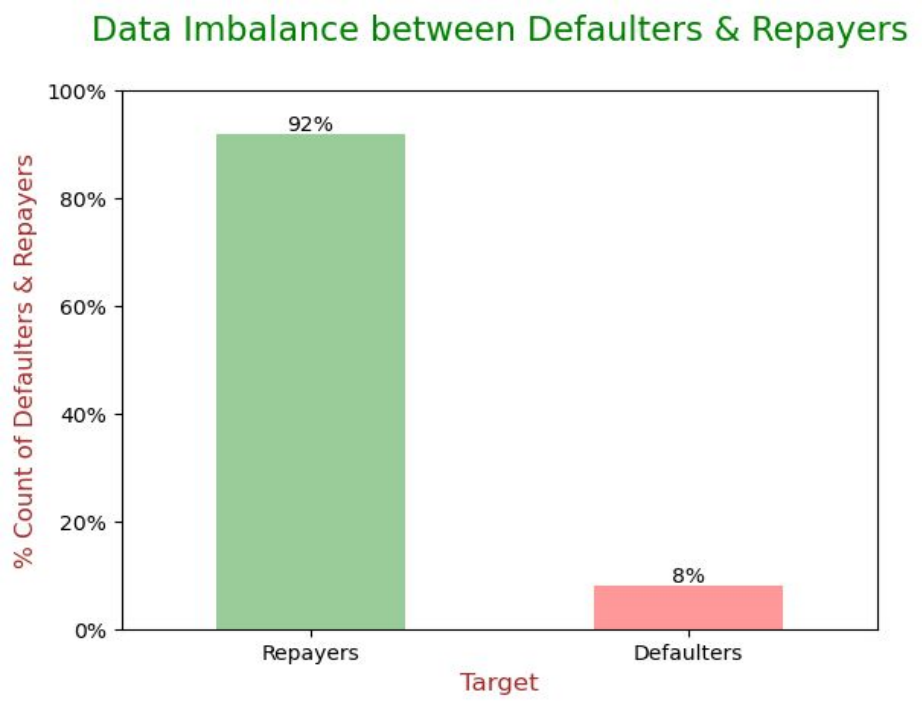
1. **Unequal distribution as**

   92% Re-payers
   8% Defaulters
   Ratio = 11.4 : 1

2. Data Imbalance Ratio, meaning for every 11 repayer there is 1 defaulter

Defaulter are clients who with **payments difficulties**

Re-payers are who successfully paid their loan



Correlation between Numerical columns

- Segmenting the data frame w.r.t. TARGET variable.
  Defaulters (with payment difficulties)
  Re-payers (all other cases)

Defaulters

The top 3 correlations are highly correlated

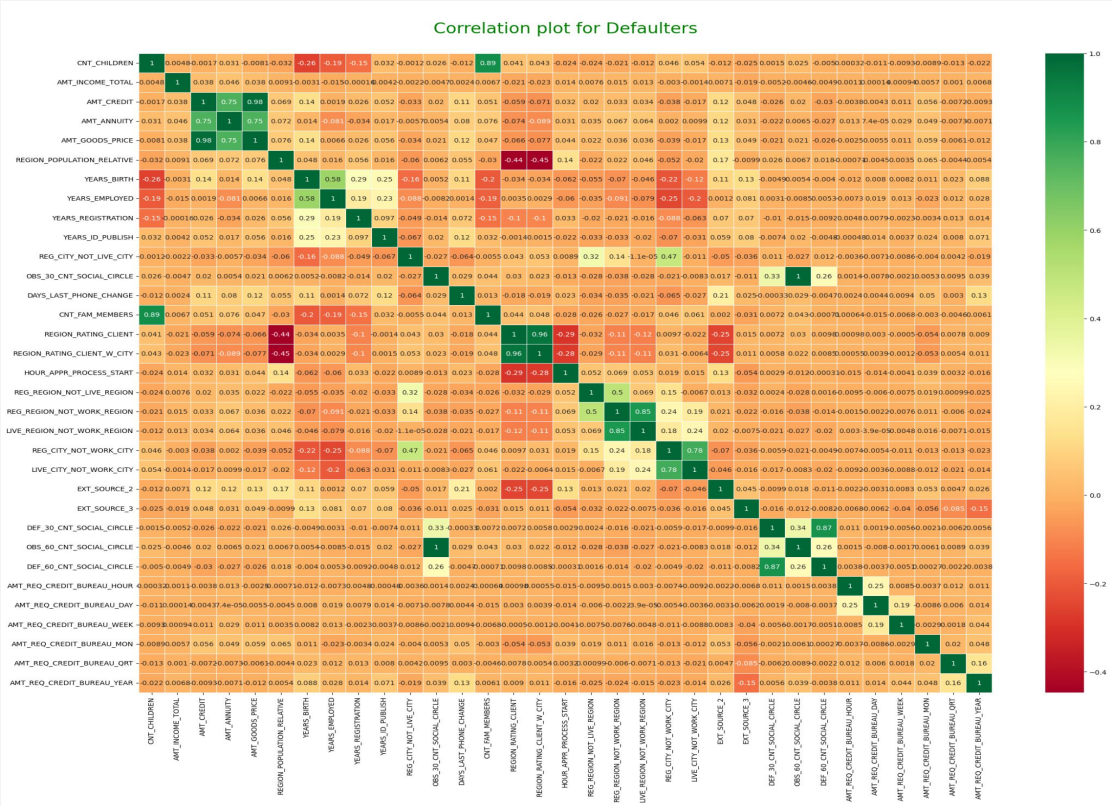
- If one increases, other will also increase
- If the AMT_CREDIT is high, the AMT_GOODS_PRICE is also high as both are correlated with coefficient (0.983)

Top 10 Correlation Combination for Defaulters

	VAR1	VAR2	CORR_COEFFICIENT
0	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.9983
1	AMT_CREDIT	AMT_GOODS_PRICE	0.9828
2	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.9566
3	CNT_CHILDREN	CNT_FAM_MEMBERS	0.8855
4	DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.8690
5	REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.8479
6	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.7785
7	AMT_ANNUITY	AMT_GOODS_PRICE	0.7523
8	AMT_ANNUITY	AMT_CREDIT	0.7522
9	YEARS_EMPLOYED	YEARS_BIRTH	0.5822

Correlation plot for Numerical columns

Defaulters Correlation plot for Numerical columns



Correlation between Numerical columns

- Segmenting the data frame w.r.t. TARGET variable.
  Defaulters (with payment difficulties)
  Re-payers (all other cases)

Re-payers

The top 3 correlations are highly correlated

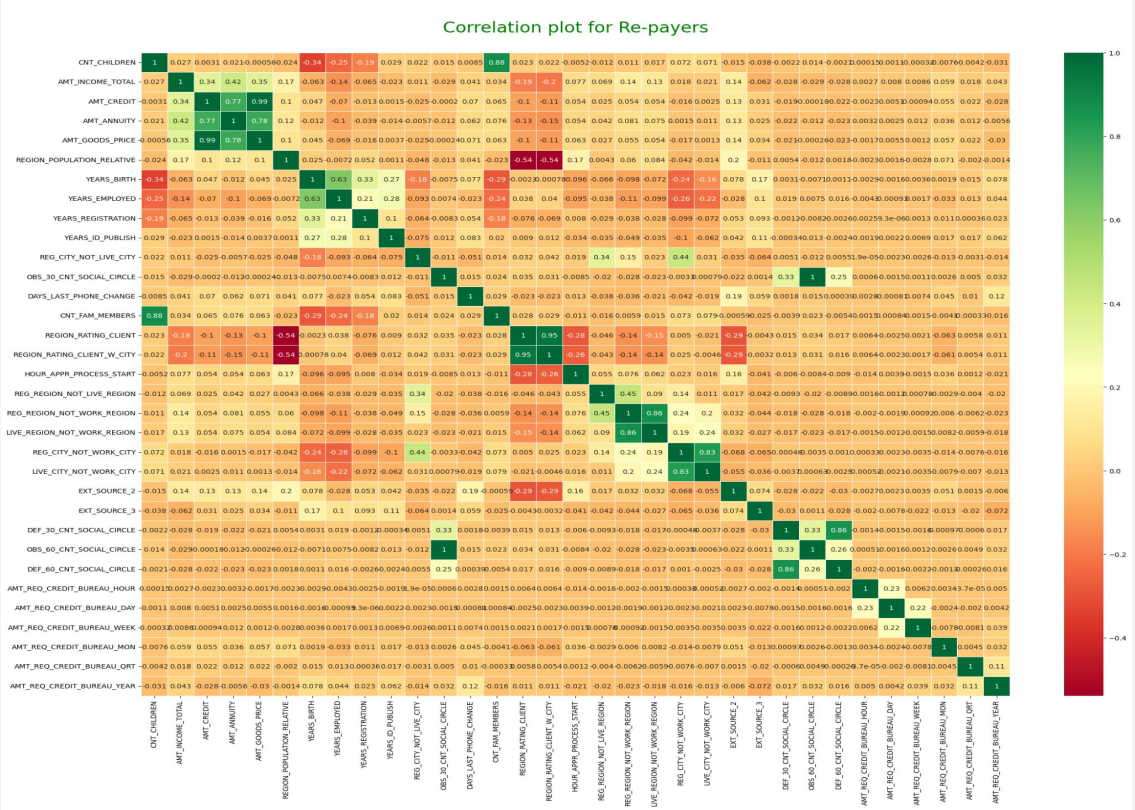
- If one increases, other will also increase
- The overall top 10 correlations has some increment in case of repayers as compare to defaulter
- Experienced and elder clients didn't face payments difficulties as their coefficient is increased (from 0.582 to 0.626) as seen in the both figures).

Top 10 Correlation Combination for Re-payers

	VAR1	VAR2	CORR_COEFFICIENT
0	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.9985
1	AMT_GOODS_PRICE	AMT_CREDIT	0.9870
2	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.9501
3	CNT_FAM_MEMBERS	CNT_CHILDREN	0.8786
4	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.8619
5	DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.8594
6	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.8304
7	AMT_ANNUITY	AMT_GOODS_PRICE	0.7764
8	AMT_CREDIT	AMT_ANNUITY	0.7713
9	YEARS_BIRTH	YEARS_EMPLOYED	0.6261

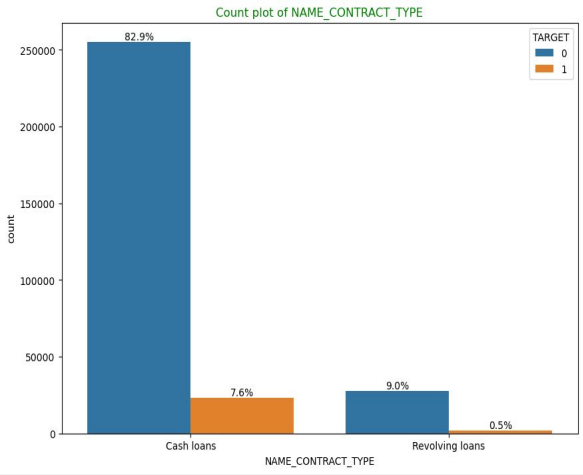
Correlation plot for Numerical columns

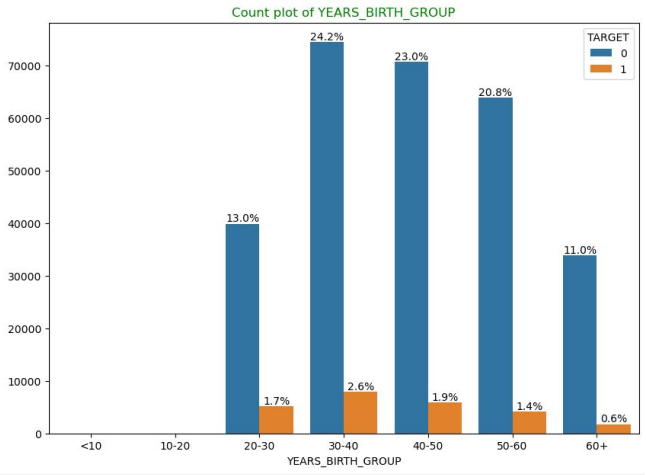
Re-payers Correlation plot for Numerical columns

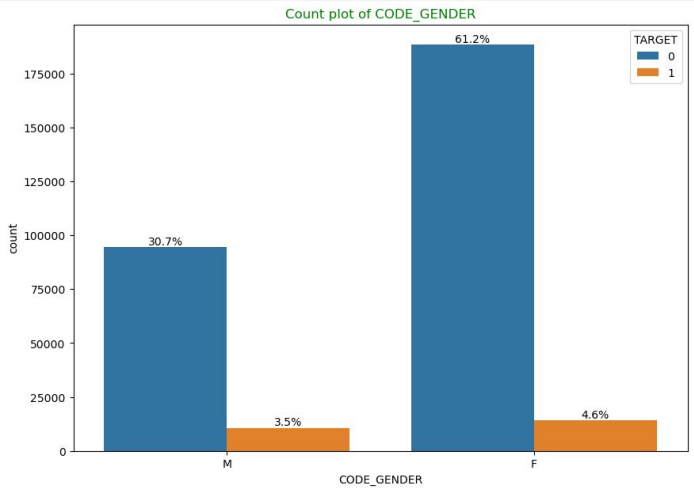


Univariate Analysis for application-data

- **Categorical Analysis**







Insights

Plots tells value counts % in each variable w.r.t to repayers(0) and defaulters (1)

- Clients are taking 83% of cash loans, 9% of Revolving loans
- Majority of clients falls in 30-50 age groups
- Number of Female (66%) clients are more than Males (34%)

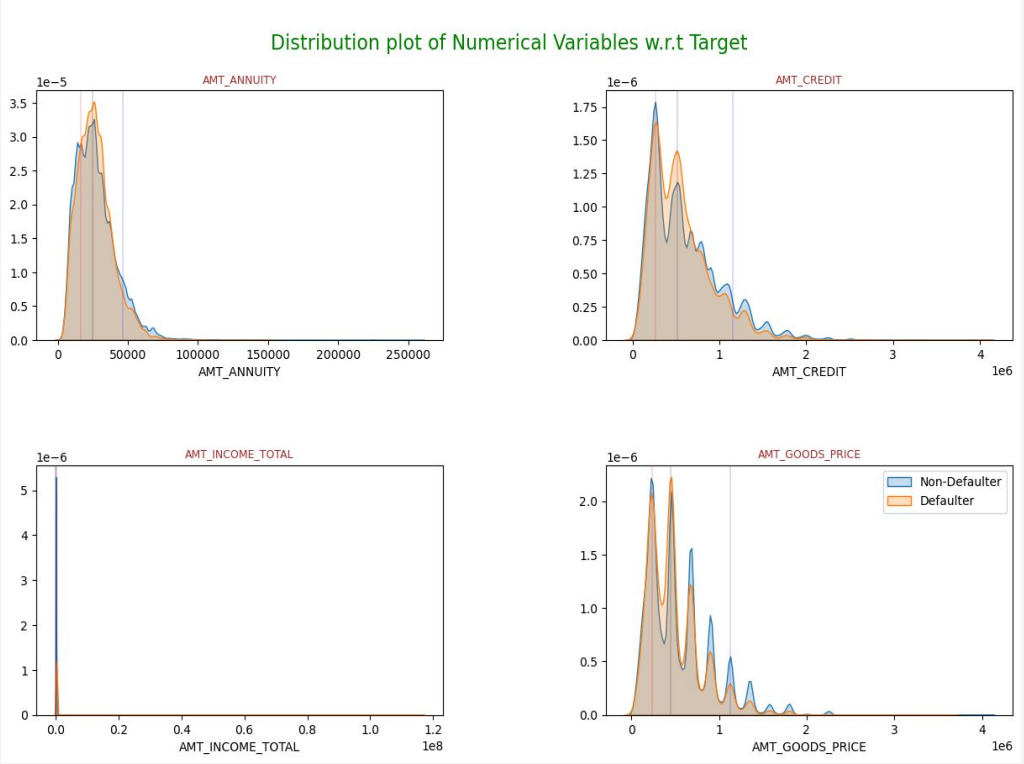
We get similar insights with other plots as well

Univariate Analysis for application-data

- **Continuous Analysis (AMT_XX)**

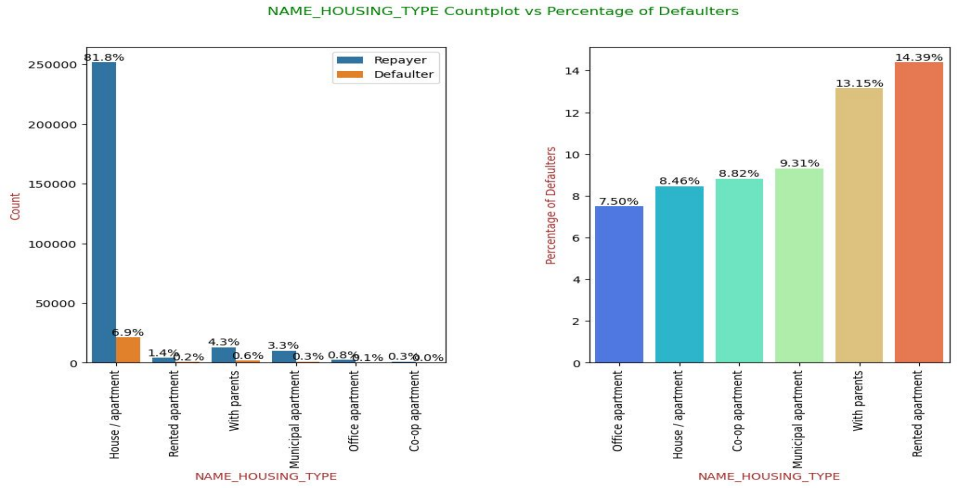
Insights

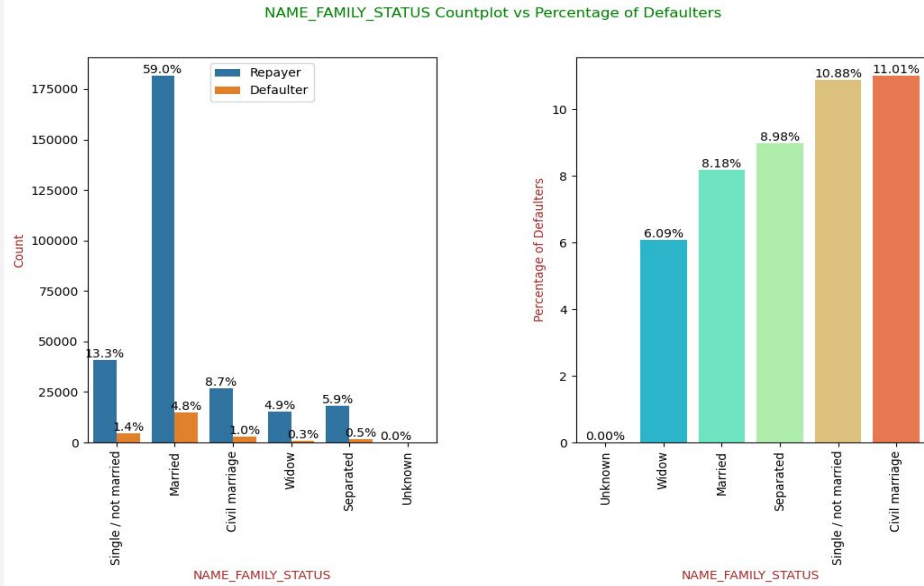
- All plots are skewed towards left, meaning the high frequencies of AMT_XX are at lower amounts
- Most applicants took AMT_ANNUITY of approx 25k
- Less number of applicants are buying expensive goods
- Most Clients buying goods of less amount at 240k.
- 90% of Credit amount is less than 10 lakhs

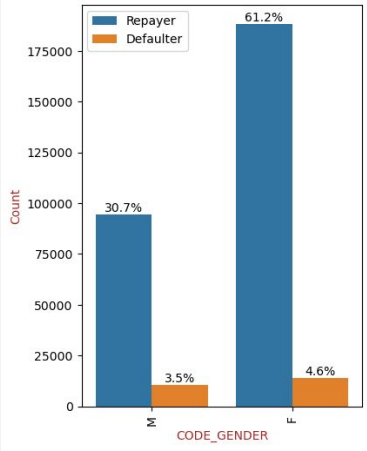


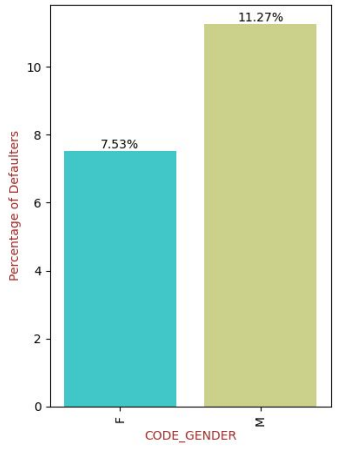
Bivariate Analysis

- **Categorical variables (Variable/s vs TARGET)**







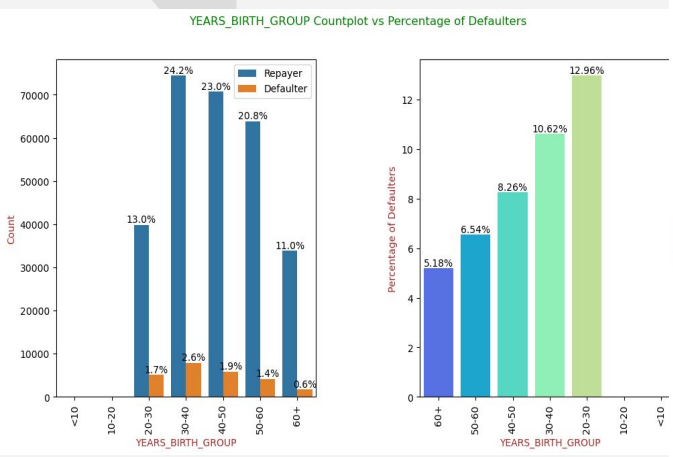


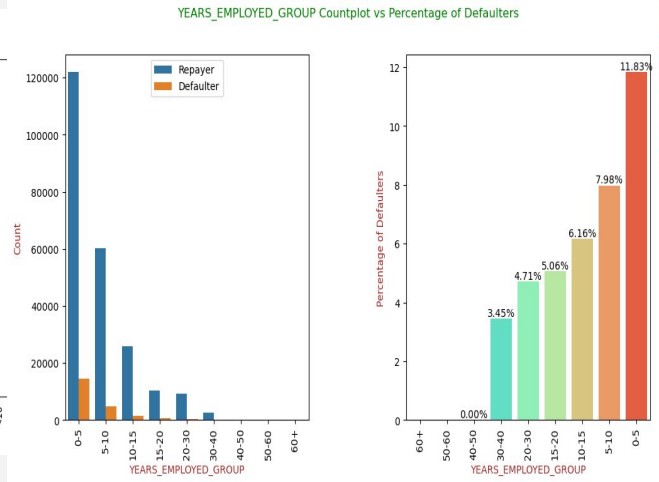
Insights

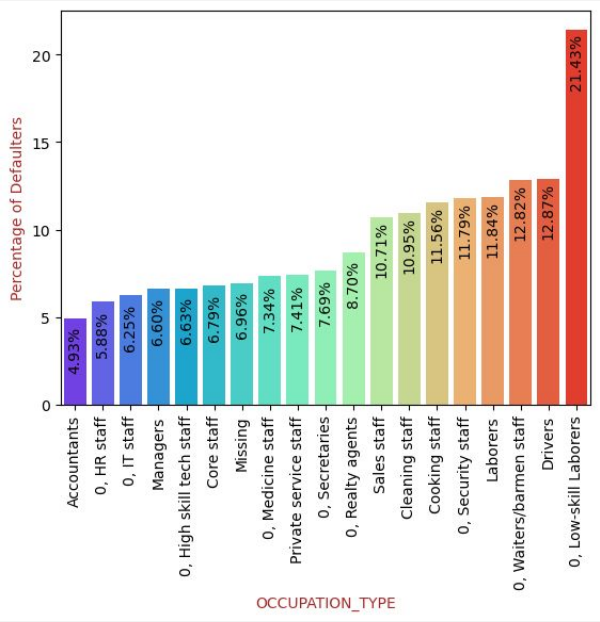
- Applicants living in Rented apartments and living with parents have high defaulter rate (13-15%)
- Civil marriage and Single / not married High defaulter rate approx(10-11 %)
- Females have 92.5% of re-pay rate , 11% of males are defaulters whereas only 7.5 % females are defaulter.

Bivariate Analysis

- **Categorical variables (Variable/s vs TARGET)**





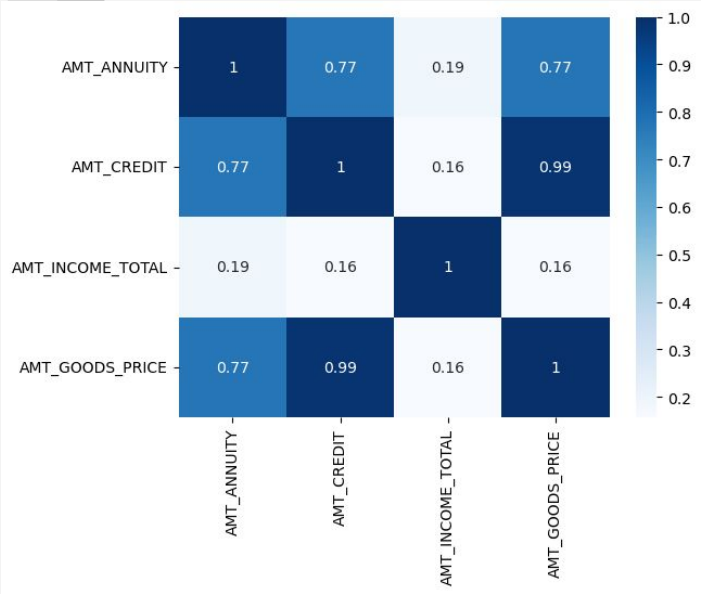


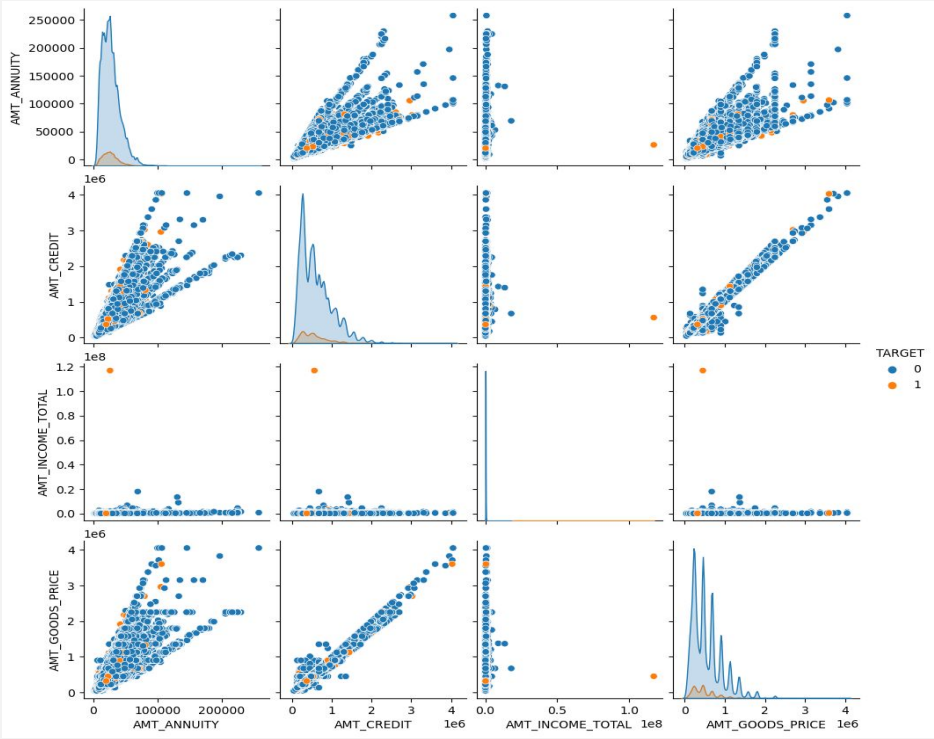
Insights

- Adult applicants in age group 20-30 years has high default rate 13%,clients with 60+ age has 95% repay rate
- 0-5 years of work exp. shows 12 % default rate, clients with 15 years of work exp has 95% repay rate
- Low-skill laborers has less application counts but 20% of them are likely to default, these are **risky applicants**
- Drivers, Waiters/barmen staff and labors (working applicants) shows (10-12%) default rate
- Where Accountants, HR staff, Managers are more likely to pay the loan as they have approx 95% repay rate

Bivariate Analysis

- **Continuous variables (Variable/s vs TARGET)**

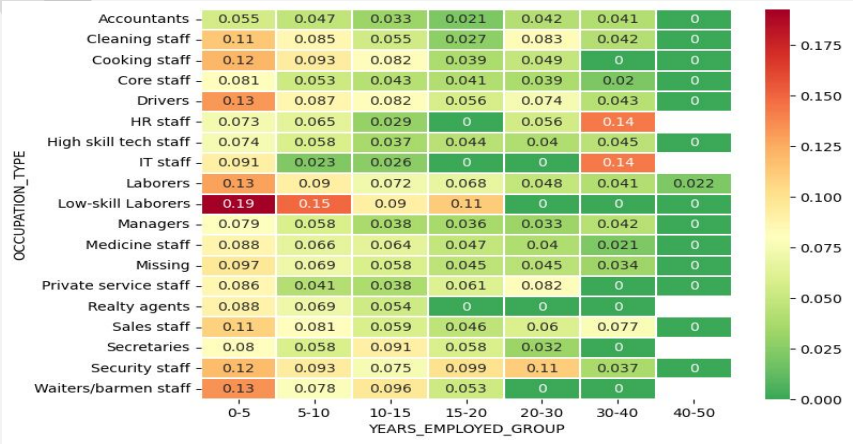


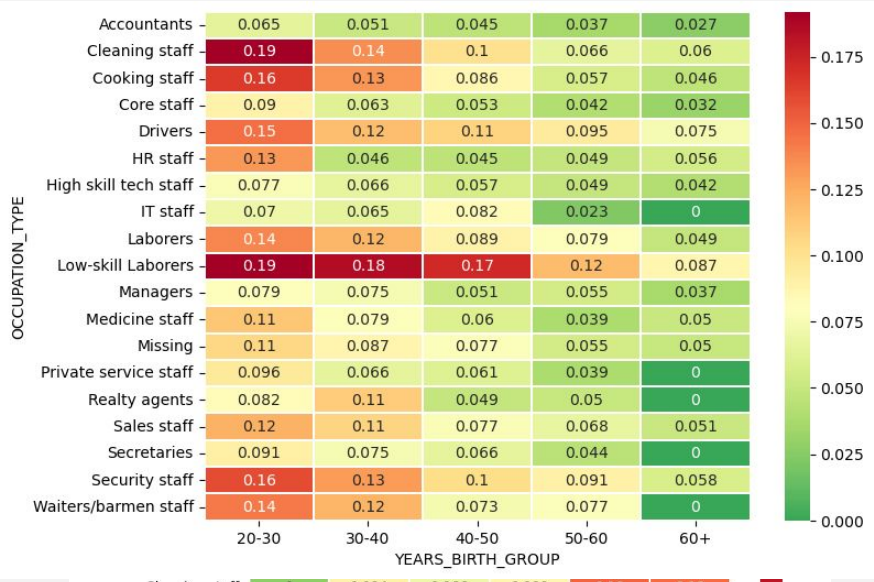


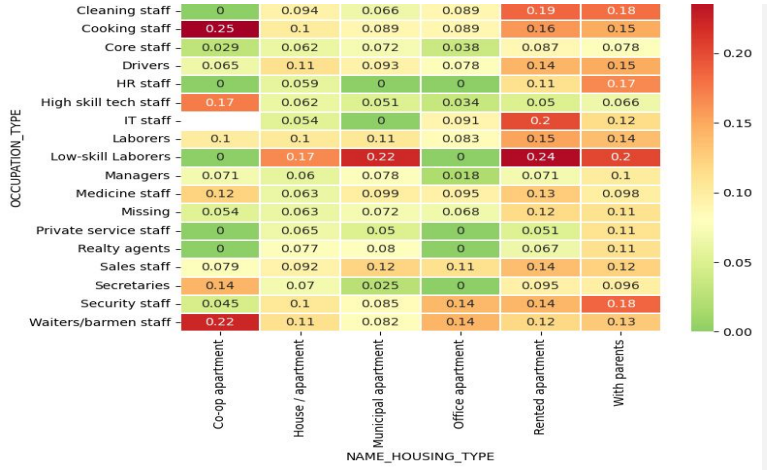
- Applicants with 100k annuity will more likely to repay their loans, but their frequency is low
- In credit amount range > 3M there are approx 20% defaulters out of total applicants who has AMT_CREDIT > 3M, The Defaulters rate is more when the AMT_CREDIT > 3M.
- High correlation between AMT_CREDIT and AMT_GOODS_PRICE with correlation coefficient of 0.99

Multivariate Analysis

(more than 2 variables vs TARGET variable)



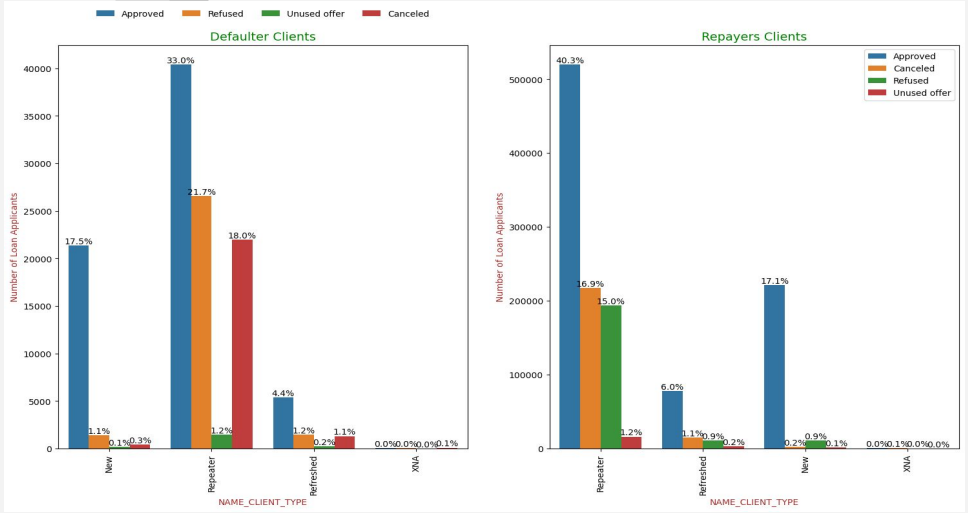


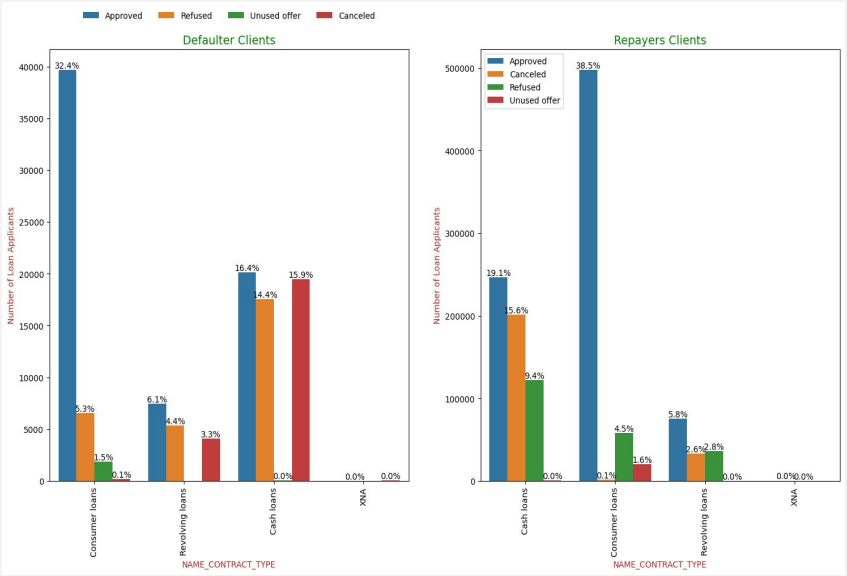


- In 0-5 years work exp., Low-skill laborers shows highest default rate, Loan providers can target Accountants in this segment to increase profit
- Clients living in Co-op apartments and working as cooking staff, Waiters/barmen staff are risky clients.
- Clients in 20-30 age group shows highest default rate

Merged Dataset

- **Bivariate/Multivariate analysis**

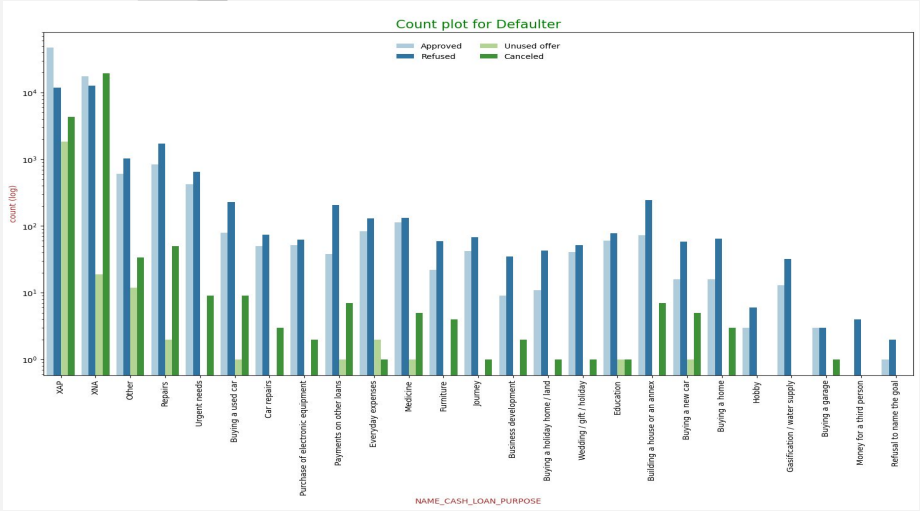


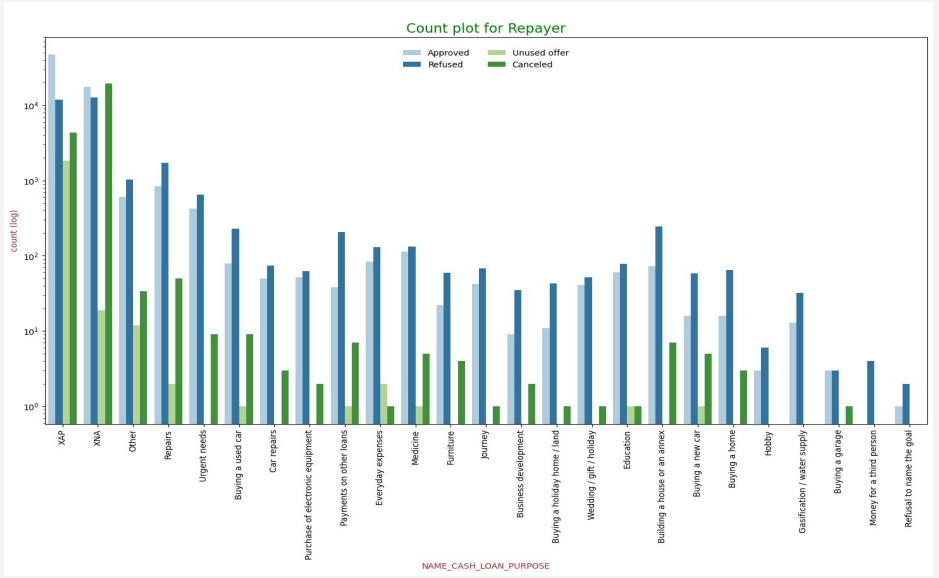


- 33% of loans which gets approved are defaulted by Repeaters, followed by 17% newcomers.
- Defaulters take (32%) of loans as Consumer loans, which gets approved.
- Also There is 38% of approved loan which the repayers pay when they take Consumer loan,19 % of approved loans are repaid from cash loans

Merged Dataset

- **Bivariate/Multivariate analysis**





Insights

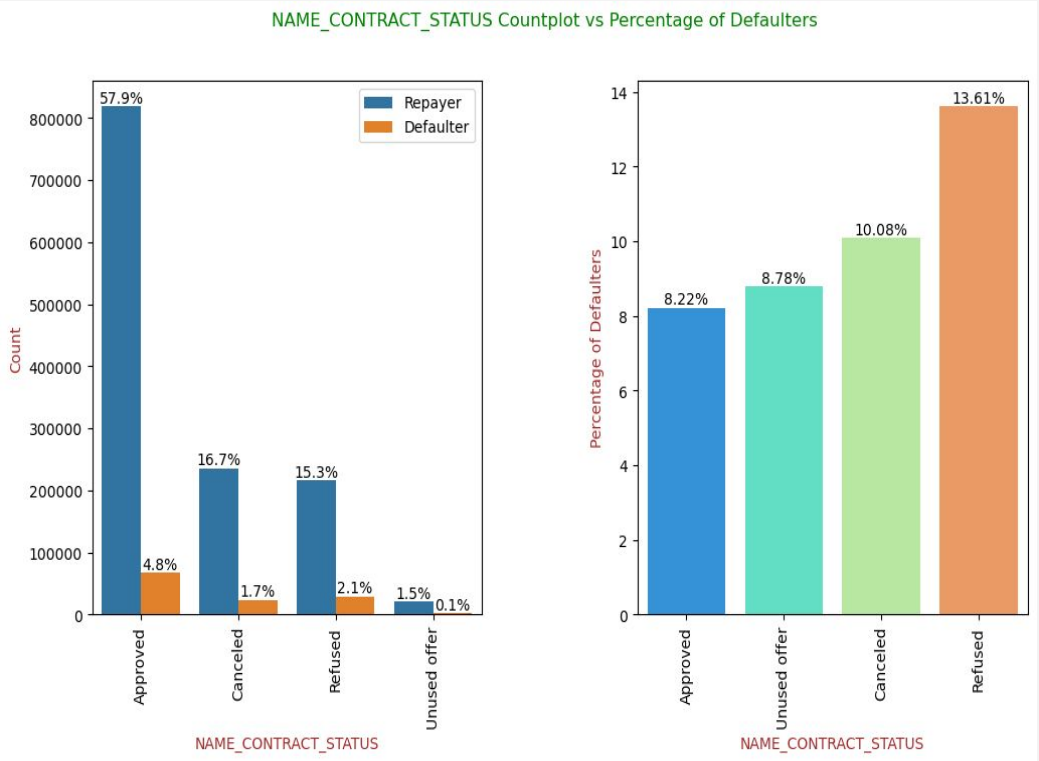
- XAP, XNA are very important information, but they are unknown, team should collect data with more precision, or proper data transformation should be done.
- From Known, Repairs got most refusals in both cases (defaulters, repayers)
- One reason for the canceled decision by clients can be high rate of interest given to loan applicants

Merged Dataset

- **Bivariate/Multivariate analysis**

Insights

- Approved loans has 92% of repayers rate, there is 8% of default rate,
- Refused loans shows high defaulters rate (13.61%) in case of Repeaters (customers repeating loans)



Observation From Analysis

📝 **Decisive Factors for the client to be a Defaulter, Risky Borrowers**

- 🎯 **Meaning clients applications can be Rejected**

🎯 **Driving factors behind difficulties in payments :**

1. **CODE_GENDER** – The Default rate of Males (11.27%) is higher than Females(7.53%)
2. **NAME_CONTRACT_TYPE** – Clients shows more default rate in Cash loans (9.1%) than Revolving loans (5.78%)
3. **NAME_FAMILY_STATUS** – Clients in Civil marriage or Single / not married shows more default rate
4. **NAME_HOUSING_TYPE** – Clients living on Rent or with parents has high default rate of 13-14%.
5. **YEARS_BIRTH_GROUP** – Young applicants in age group 20-30 has high default rate 13%
6. **YEARS_EMPLOYED_GROUP** – People who has 0-5 years of work experience have approx 12 % default rate
7. **REGION_RATING_CLIENT** – Clients who lives in region of rating 3 are more likely to default than who lives in region with 1, 2 ratings
8. **OCCUPATION_TYPE** – Low-skill labourers are the risky clients, has less application counts but 20% of them are likely to default, Drivers, Waiters/barmen staff and labours (working applicants) shows (10-12%) default rate
9. **AMT_CREDIT** – The Defaulters rate is more when the AMT_CREDIT > 3M,There are approx 20% defaulters out of total applicants who have AMT_CREDIT > 3M.
10. **AMT_GOODS_PRICE** – Maximum frequency of people are buying goods worth 240k and 450k (people buying 450k has more defaulters than people buying goods worth 240k)

Summary for Defaulters

- **Decisive Factors for an applicant to be a Defaulter**

Male clients have relative higher default rate than female clients

Low-skill labourers, Drivers, Waiters/barmen staff and labours shows high default rate

Clients living on Rent or with parents has high default rate of 13-14%.

Clients in Civil marriage or Single / not married shows more default rate

Careful with young clients in age group 20-30, has high default rate 13%

When the credit amount goes beyond 3M, there is an increase in defaulters

Observation From Analysis

📝 **Decisive Factors for the client to be a Re-payer, Safe Borrowers**

- 🎯 **Meaning clients applications can be Approved**

🎯 **Driving factors behind successful loan payments :**

1. **CODE_GENDER** – Females are more number of clients and their repaying rate is approx 92%, which is very good.
2. **NAME_CONTRACT_TYPE** – Clients shows more repay rate in Revolving loans than cash loans
3. **NAME_HOUSING_TYPE** – Clients living in their own house / apartment or office apartments has high repay rate of approx 92%.
4. **YEARS_BIRTH_GROUP** – Clients with 50+ age have highest repay rate
5. **YEARS_EMPLOYED_GROUP** – People who has more years of work experience (15+) have approx 95 % repay rate
6. **NAME_INCOME_TYPE** – Clients who are Businessman or Students are good to go clients they have no defaulters
7. **OCCUPATION_TYPE** – Accountants, HR staff, Managers are more likely to pay the loan as they have approx 95% repay rate
8. **AMT_ANNUITY** – Applicants with 1 lakh (100k) Annuity will more likely to repay their loans, but their frequency is low.
9. **NAME_EDUCATION_TYPE** – Academic degree clients shows very high repay rate, they are valuable for banks
10. **REGION_RATING_CLIENT** – Clients who lives in region of rating 1 shows high repay rate

Summary for Re-Payers

- **Decisive Factors for an applicant to be a safer client**

Clients with 50+ age have highest repay rate

Accountants, HR staff, Managers have approx 95% repay rate.

Clients living in their own house / flat or office flats has high repay rate

Academic degree clients shows very high repay rate, they are valuable for banks

Businessman or Students are good to go clients, they have no defaulters

Applicants with 100k Annuity have high repay rate.

✅ Conclusions

Suggestions to prevent the Business loss, as Decreasing Defaulter rate and Increasing Re-Pay rate is our aim.

- **CODE_GENDER**

Loss -Males as Low-Skill Labourers or Realty agents are risky clients, and Loan providers should be careful while giving cash loans. (This insight is from Multivariate analysis)

Profit – Bank should make good customer relations with females to increase profit

- **NAME_HOUSING_TYPE**

Loss – Clients living in Co-op apartments and working as cooking staff are risky clients, if they default loan providers will suffer loss.
Profit – They can be offered with less credit amount.

✅ Conclusions

Suggestions to prevent the Business loss, as Decreasing Defaulter rate and Increasing Re-Pay rate is our aim.

- **NAME_CONTRACT_TYPE**

  Clients who are in IT Staff, Accountants, Managers with high repay rate when the loan is Revolving type , than cash loan.

- **AMT_INCOME_GROUP**

  90 percentile of clients is under 300k income, they show high default rate, so they can be given loan at high interest rate.

- **NAME_EDUCATION_TYPE**

  More loan products can be suggested to people with Academic degrees to increase the profit.

THANK YOU