EDA Case Study: Credit Risk Analytics

By:-
1. Saqib Mohammed
2. Shenaz Rahaman

Problem Statement

The company has to decide for loan approval based on an applicant's profile. There are two types of risks 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.**

Thus the company wants to identify patterns in the dataset to ensure that the applicants capable of repaying the loan are not rejected and understand the influence of consumer and loan attributes on the tendency of default, i.e., the driving factors or strong indicators behind loan default.

Approach

We perform Exploratory Data Analysis (EDA) on the given datasets with the help of following information:

- **Loan payment status as per the 'application_data.csv' file -**
  - **Client with payment difficulties (‘0’) - Defaulter**
  - **All other cases (‘1’) - Repayer**

- **Decision on loan application as per the 'previous_application.csv' file -**
  - **Approved - by the Company**
  - **Cancelled - by the Client**
  - **Refused - by the Company**
  - **Unused offer - by the Client**

Steps of this EDA

- **Data Sourcing -**
  - **Reading the data**
  - **Structure of DataFrames**

- **Data Cleaning -**
  - **Handling Null Values**
  - **Standardize Values**
  - **Null Value Data Imputation**
  - **Identifying Outliers**

- **Data Analysis -**
  - **Imbalance Data**
  - **Univariate Analysis**
    Bivariate Analysis
  - **Merged Dataframe Analysis**

Data Sourcing

- **Reading the data -**
  - The **application_data.csv** and **previous_application.csv** files are read into their respective dataframes **application_df** and **previous_df** using **pd.read_csv().**

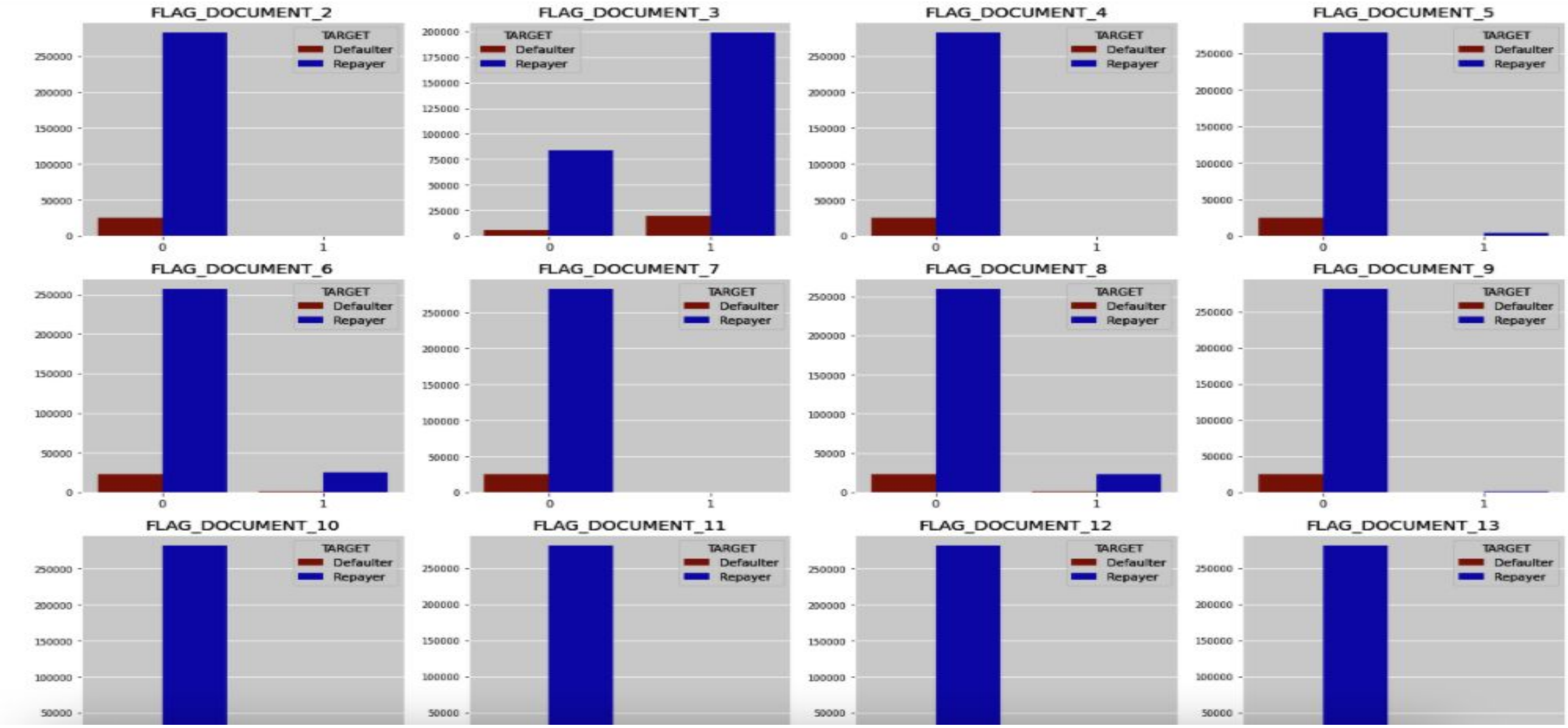
- **Structure of DataFrames -**
  - Basic inspection is performed on the dataframes to determine:
    - Column wise information using **.info()**
    - Dimension using **.shape**
    - Summary statistics of numeric columns using **.describe()**

Data Cleaning

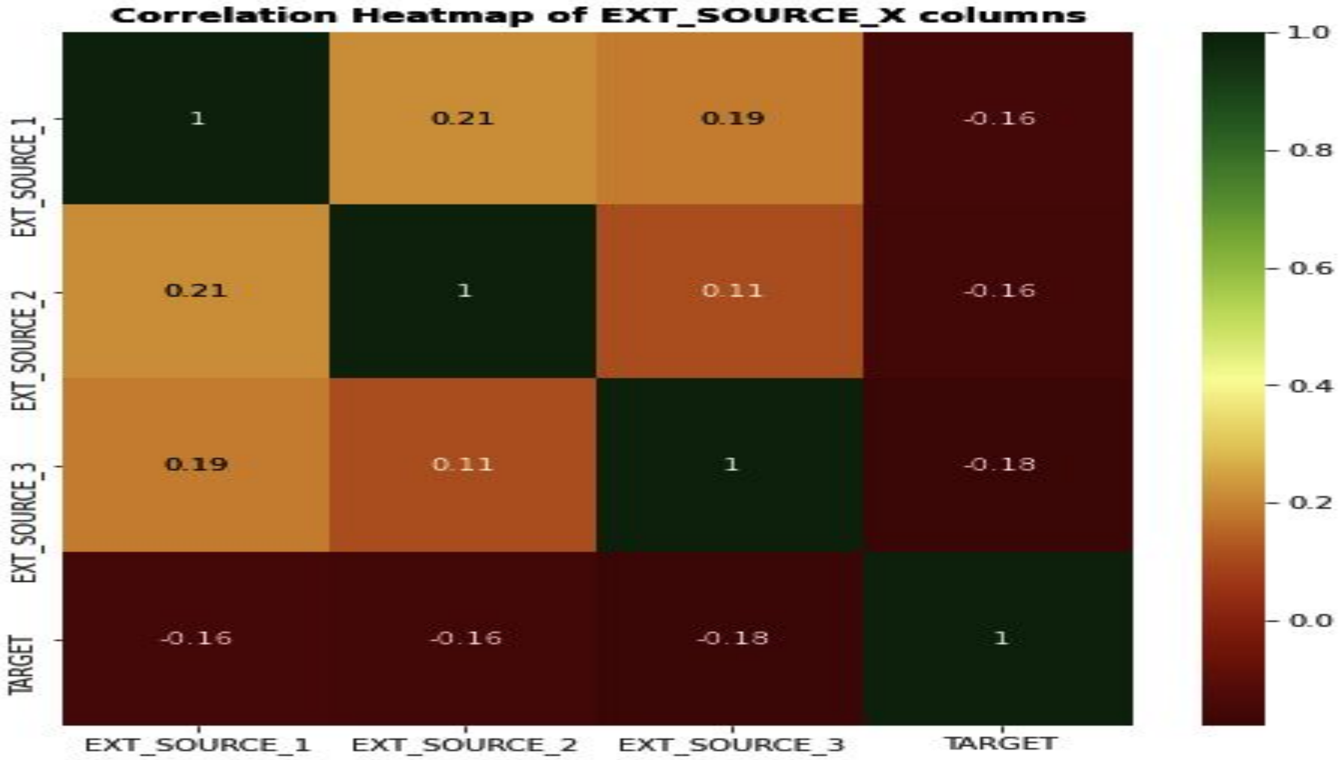
- **Handling Null Values - (application_df)**
    - The **% of missing values** in each column of **'application_df'** is found.
    - It is observed that there are **many columns** with **high missing values** (more than 40%). Such columns are handled by either **dropping** them or **imputing** values in them based on their relevance.
    - There are **49 such columns** found and most of them are **related to different area sizes or apartment owned/rented** by the loan applicant. As such information is **not of much significance for this analysis** we store these columns in a list (**Unwanted_application**) to be dropped from the dataframe.
    - The **significance of other columns (Flag document, EXT_Source, Contact Flag) on Target** are also analysed and non relevant columns are included in the list **Unwanted_application** for them to be dropped for further analysis.
    - Thus there are **only 46 columns remaining after dropping the non relevant columns** from the application_df for this analysis.

○ **Flag Document -**

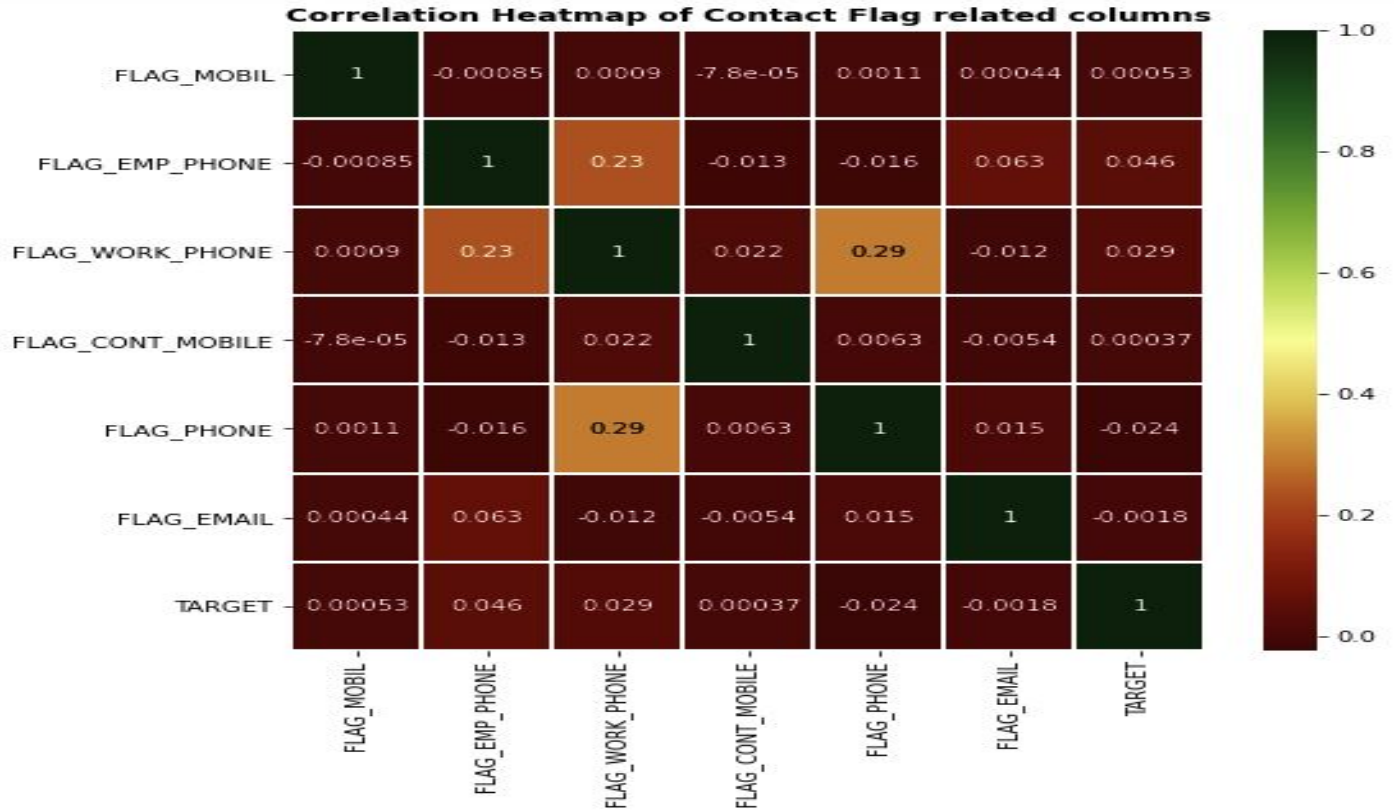
■ It is observed from the countplots that most applicants have **not submitted FLAG_DOCUMENT_X except FLAG_DOCUMENT_3.** Also, such applicant are have a lesser chance of defaulting the loan. Hence, among them only **FLAG_DOCUMENT_3 is of relevance** and the rest can be dropped.



- **EXT_Source -**
  - It is observed from the heatmap that there is almost **no correlation between EXT_SOURCE_X columns and Target column**. So these columns can be dropped as well.



- **Contact Flag -**
  - It is observed from the heatmap that there is **no correlation between Contact flag columns and Target column**. Similarly, these columns can also be dropped.



- **Handling Null Values - (previous_df)**

    - Similarly the **% of missing values** in each column of **'previous_df'** is found.

    - It is observed that there are **many columns** with **high missing values** (more than 40%). Such columns are handled by either **dropping** them or **imputing** values in them based on their relevance.

    - There are **11 such columns** found and most of them are **related to interest rates and days of due payment for the loan.** These columns are stored in a list (**Unwanted_previous**) to be dropped from the dataframe.

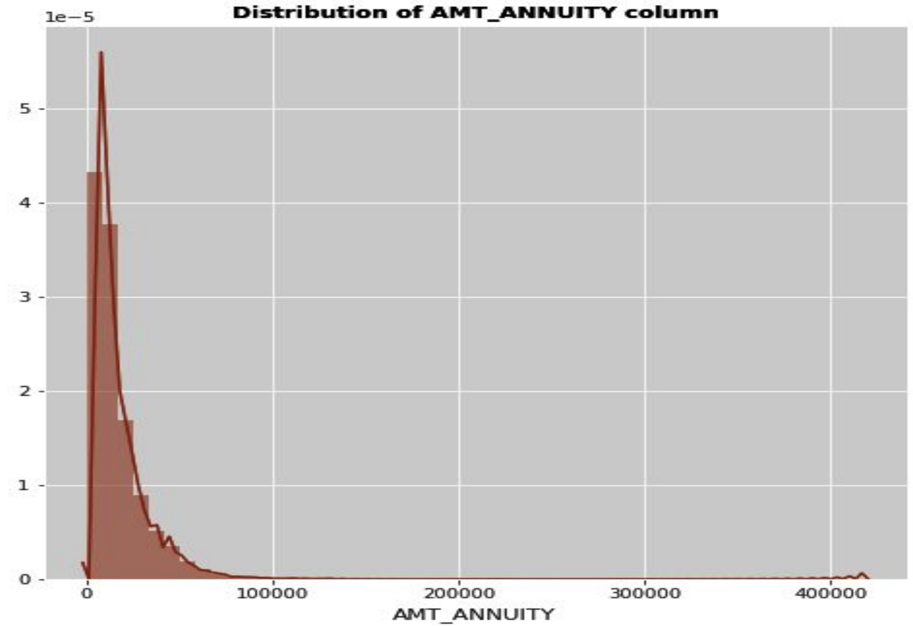
    - Other columns that are not necessary for this analysis are also added to the list **Unwanted_previous** for them to be dropped.

    - Thus there are **only 22 columns remaining after dropping the non relevant columns** from the previous_df for this analysis.

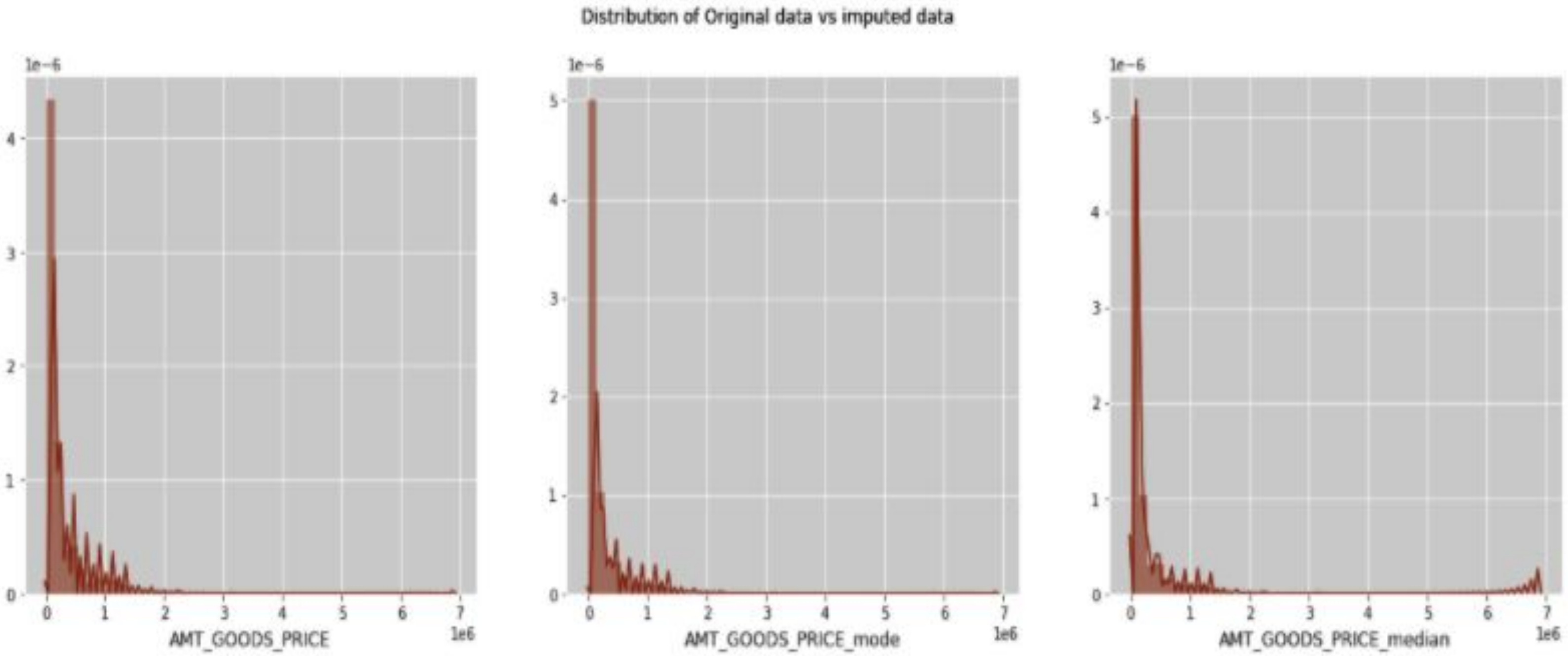
- **Standardize Values -**
    - The columns related to **count of days are converted to their absolute values** as days cannot be negative.
    - The significant **numerical columns are converted to categorical columns** by grouping them into bins.
    - **Datatypes are changed** for certain categorical variables**.**

- **Null Value Data Imputation -**
    - Checking **null value %** of the remaining columns and **imputing/ignoring** them accordingly.
    - **For application_df -**
        - The categorical variable **'NAME_TYPE_SUITE'** with **lower null percentage(0.42%)** is imputed with the most frequent category or **Mode**.
        - The categorical variable **'OCCUPATION_TYPE'** with **higher null percentage(31.35%)** is imputed with a **new category (Unknown)** as assigning any existing category might influence the analysis.
        - Columns representing the **number of enquiries made** are imputed with their r**espective median values** as there are **no outliers** and their respective **means are in decimal,** they cannot be used to impute count of enquiries.
        - **Records in columns** with **very low % of missing values are ignored** for this analysis.

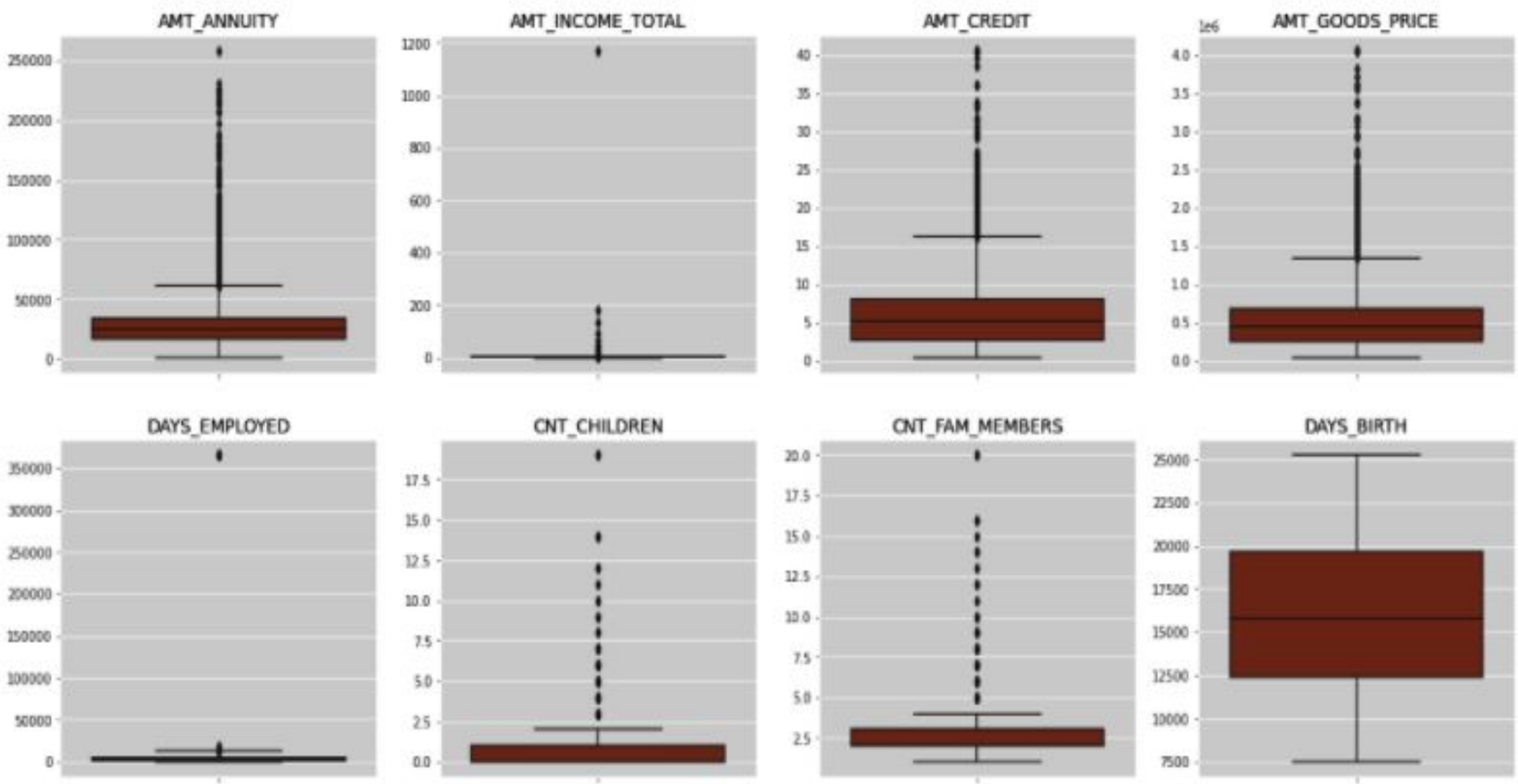
- **For previous_df -**
  - Column **‘PRODUCT_COMBINATION’** with **very low % (<1%) of missing values** and thus **such records are ignored** for this analysis.
  - **Missing values for ‘CNT_PAYMENT’** are **imputed with 0** as the **NAME_CONTRACT_STATUS** for these **indicate that most of these loans were not started**.
  - A single peak at the left side for the **AMT_ANNUITY distribution** indicate skewness, i.e., presence of outliers and are thus **imputed with median values** to avoid exaggeration of data.



- It is observed the **distribution of AMT_GOODS_PRICE data is closer to its distribution when imputed with Mode** and hence it is imputed accordingly.



- **Identifying Outliers - (application_df)**



- **Boxplots** are used to **identify outliers for relevant columns** and the following **observations** are made:

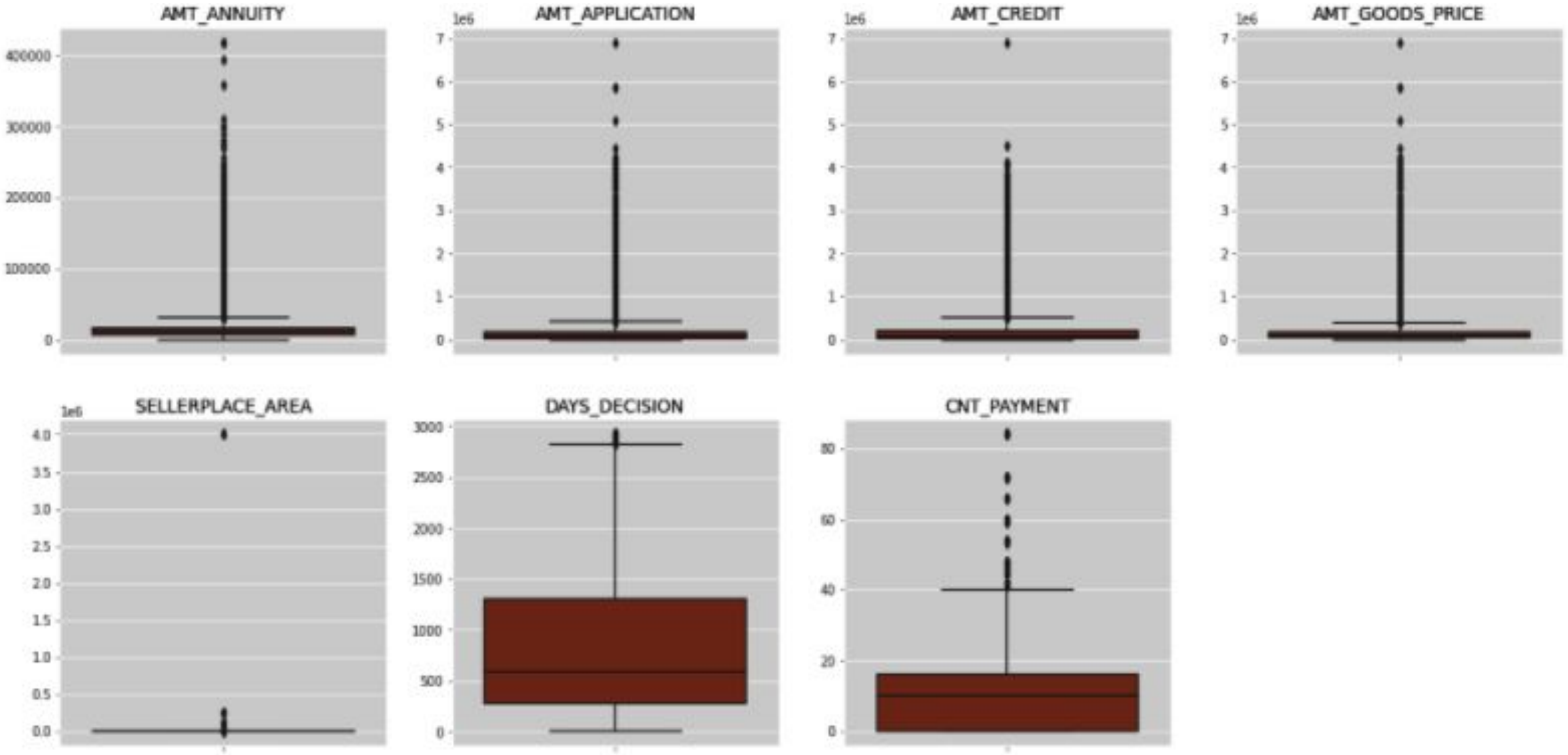
  - **AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE, CNT_CHILDREN, CNT_FAM_MEMBERS** have **some outliers**.

  - **AMT_INCOME_TOTAL** has a **large number of outliers** which indicates that **few of the loan applicants have higher income** as compared to others.

  - **DAYS_BIRTH** has **no outliers** which means the **available data is reliable.**

  - **DAYS_EMPLOYED** has **extreme outlier values at around 350000 days (i.e. about 958 years),** which is **not possible** and hence such values have to be considered as **incorrect entry**.

- **Identifying Outliers - (previous_df)**



- Similarly, **Boxplots** are used to **identify outliers for relevant columns.** The following **observations** are made:

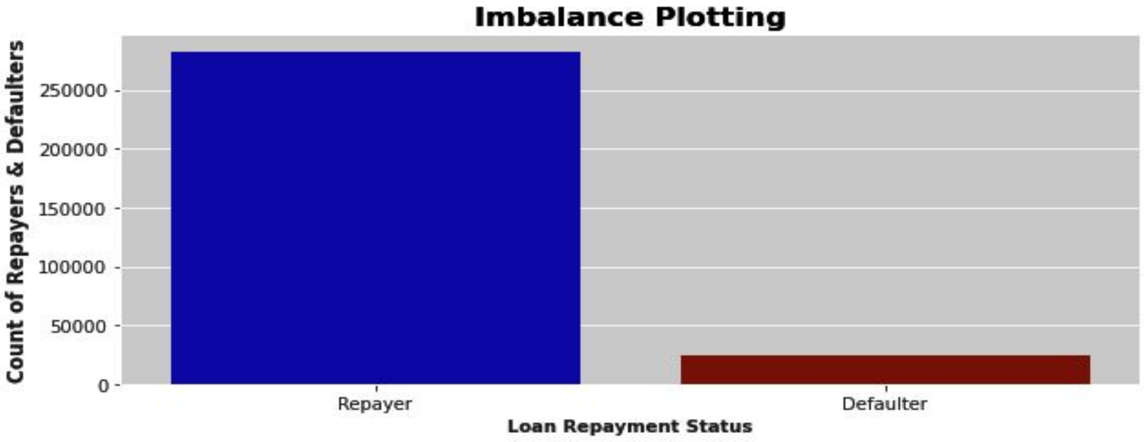
    - **AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA** have a **large number of outliers.**

    - **CNT_PAYMENT** has **fewer outlier values.**

    - **DAYS_DECISION** has **very few outliers** indicating that **decisions for these previous applications were taken long back**.

Data Analysis

- **Imbalance Data -**
  - **Count of the target variable is plotted** to determine the **ratio of Repayers to Defaulters**, which is found to be **91.93 : 8.07**.




```
# % count of the Target variable
application_df['TARGET'].value_counts(normalize=True)*100
```

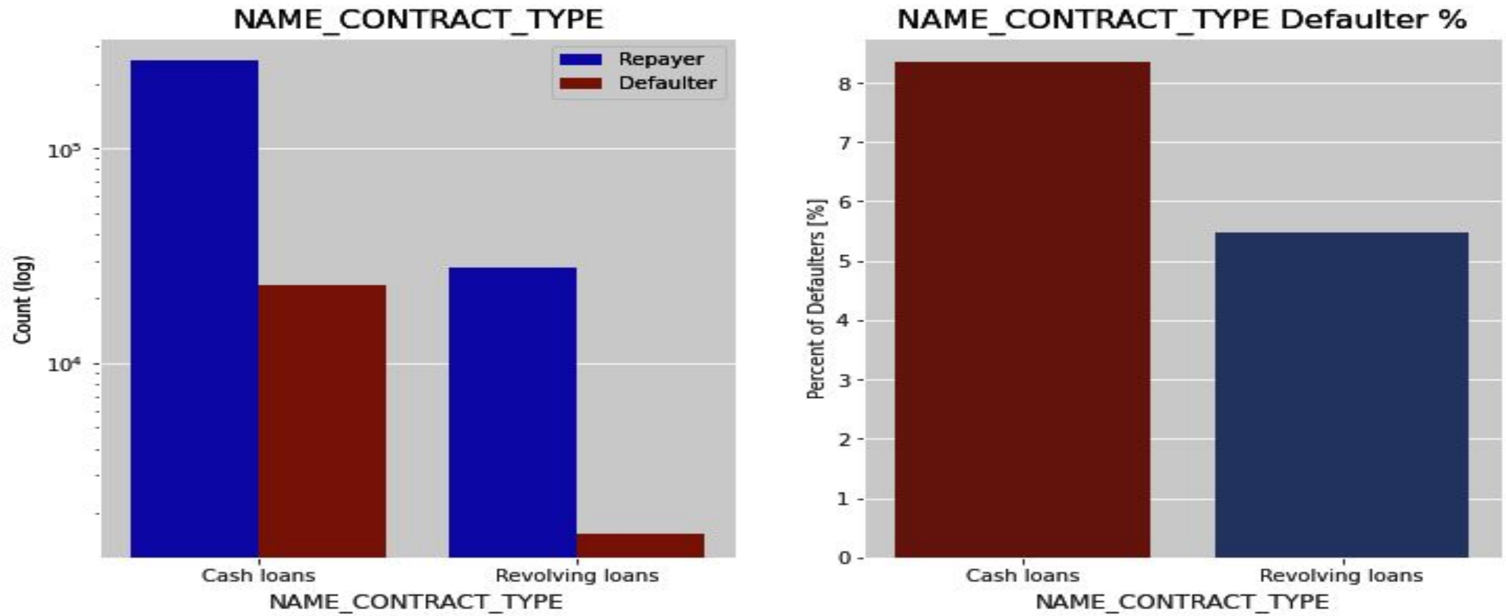
```
0    91.927118
1     8.072882
Name: TARGET, dtype: float64
```

Ratio of imbalance in percentage with respect to Repayer and Defaulter data is 91.93 : 8.07.

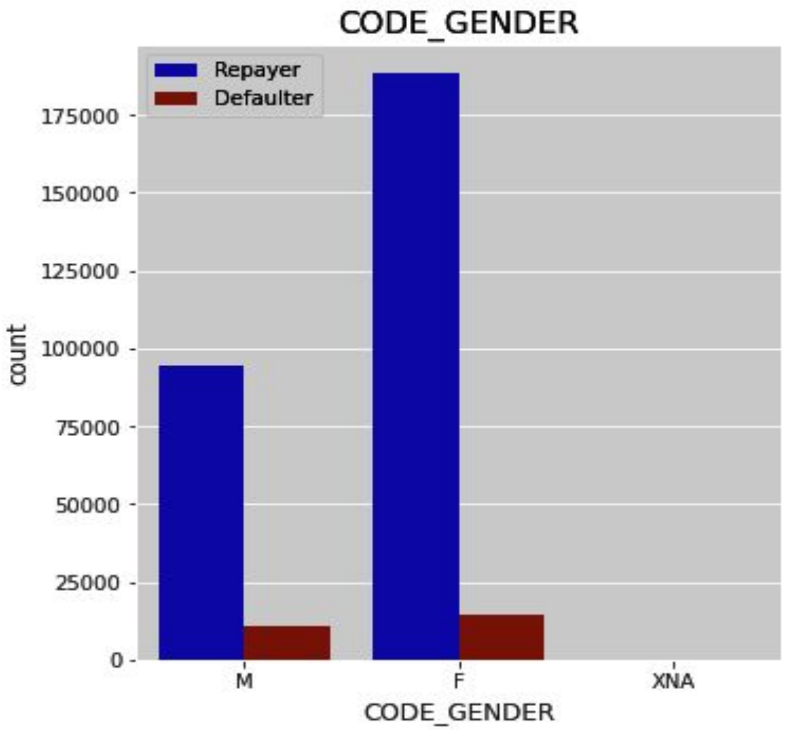
- ## Univariate Analysis -

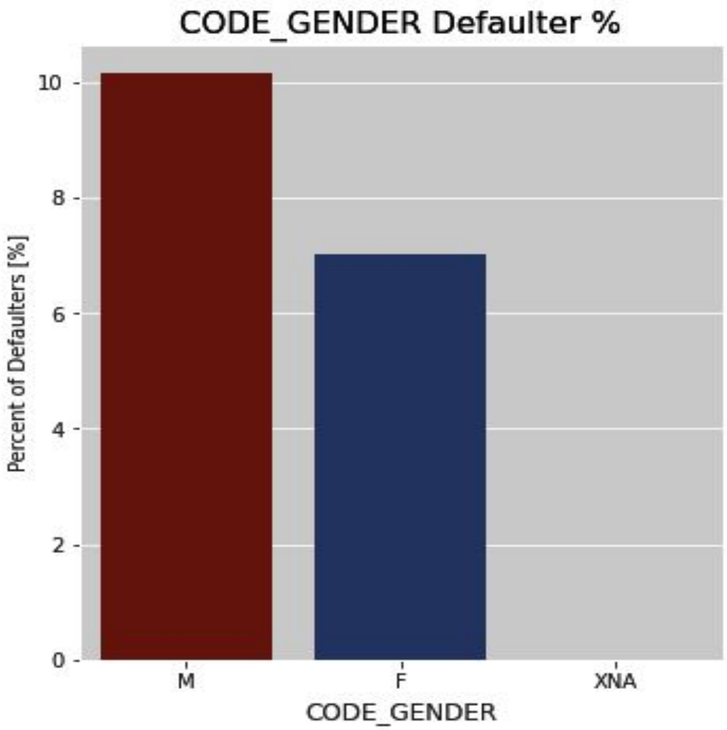
  - It is performed **based on repayment loan status (TARGET)** with the help of **countplot** and **barplot for percentage of defaulters of such columns.** The observation made are as below.

  - **NAME_CONTRACT_TYPE -**
    - The **Contract type - 'Revolving loans'** are just a **small fraction of the total number of loans**. Also, a **larger amount of Revolving loans** when compared to their frequency, are **not being repaid**.

  - 

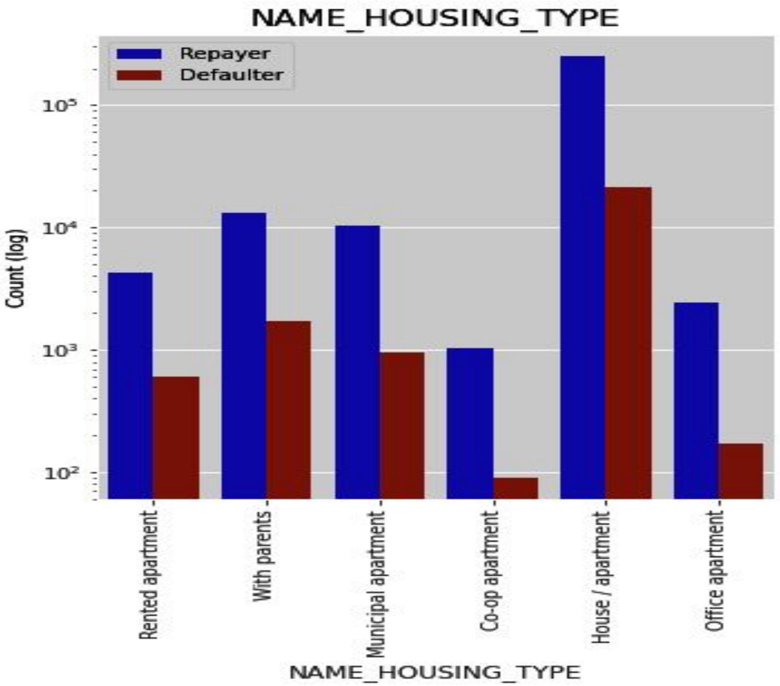
- CODE_GENDER -
    - The **number of female clients is almost twice the number of male clients**. Based on the **% of defaulted credits**, **males** are **more likely to Default** as compared to females.

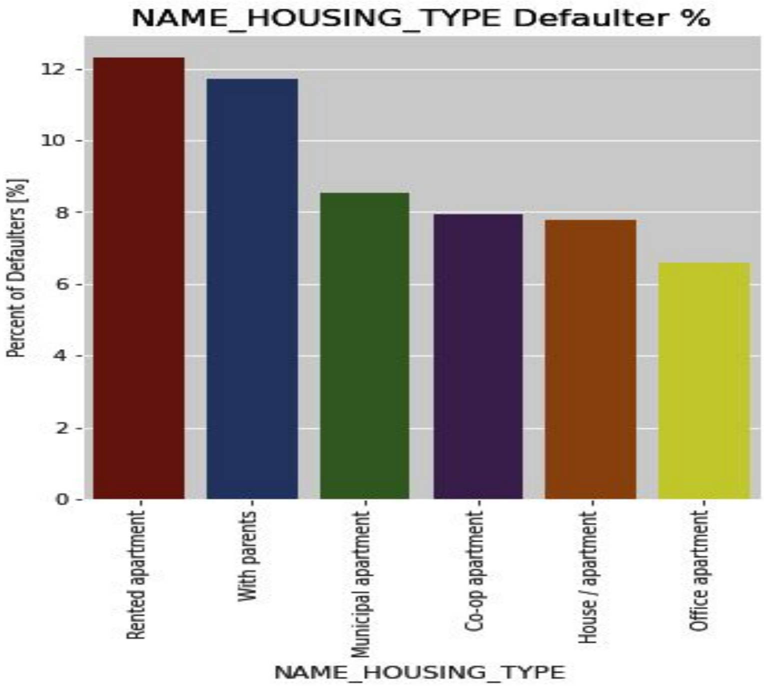
- 



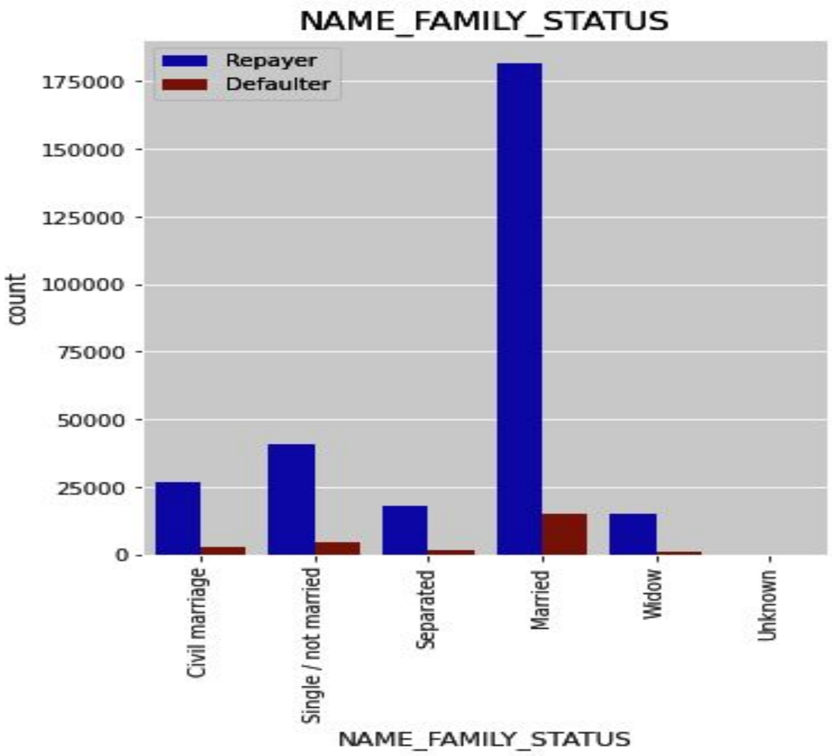
○ **NAME_HOUSING_TYPE -**

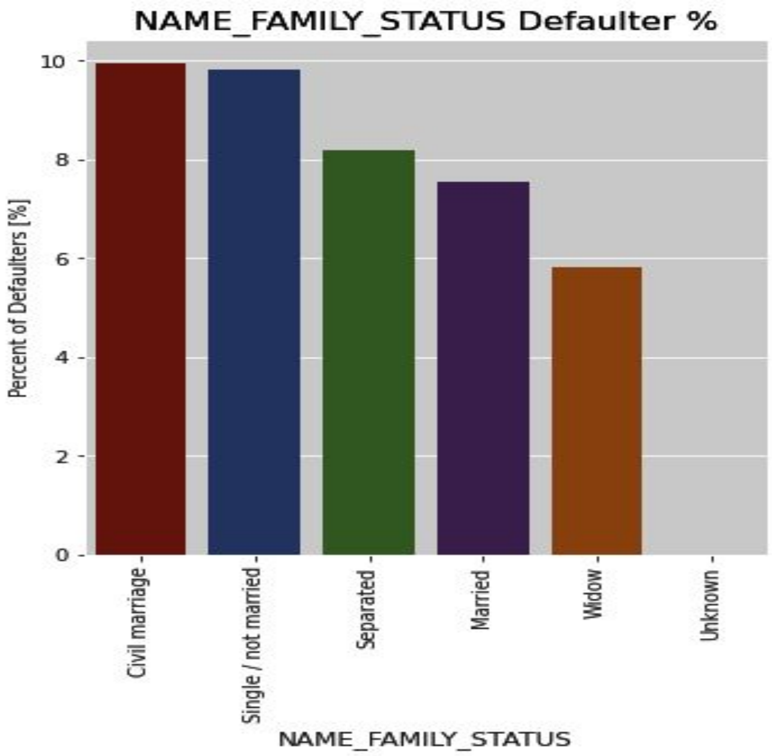
- **Most clients** live in **House/Apartment.**
- **Clients living in Rented Apartments** are **most likely to be Defaulters(>12%)**.
- **Clients living With parents** also have a **higher % (almost 12%) to be Defaulters**.
- **Clients living in the Office apartment** are **least likely to be Defaulters.**



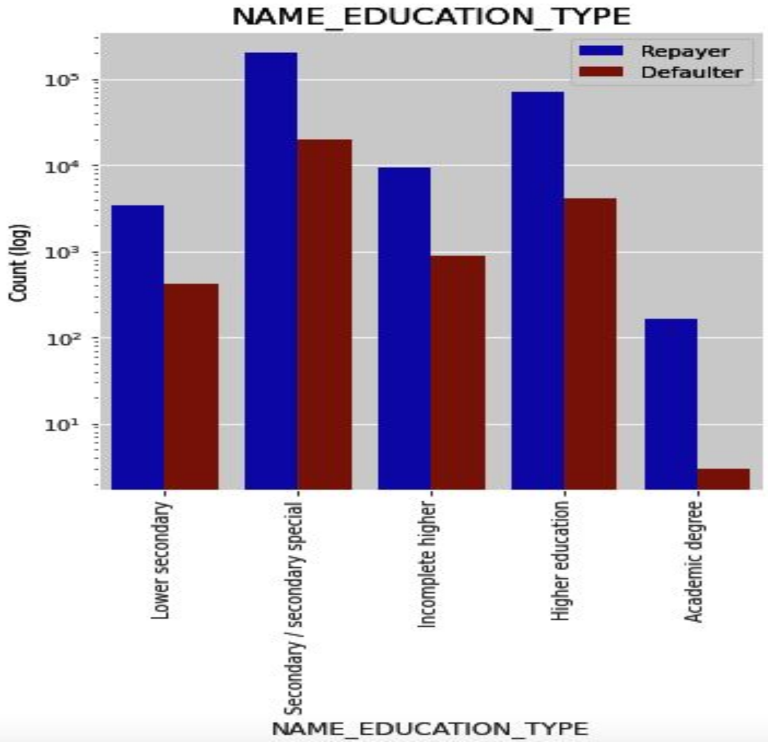


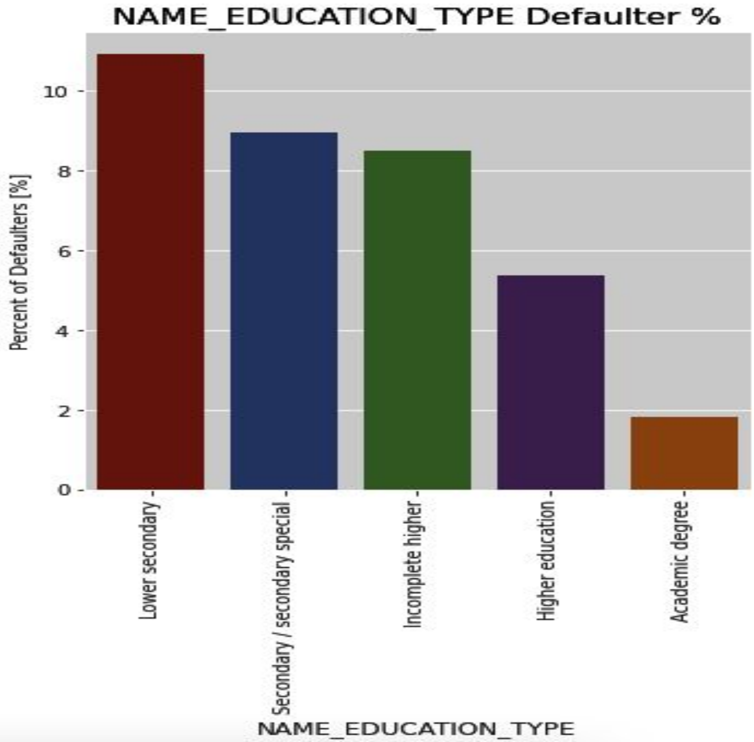
- **NAME_FAMILY_STATUS -**
  - **Most clients** are **Married**.
  - **Clients with Civil marriage** are having **high Defaulter% (10%)** and then **Single/not married (almost 10%)**.
  - **Widow clients** are the **least Defaulters**.





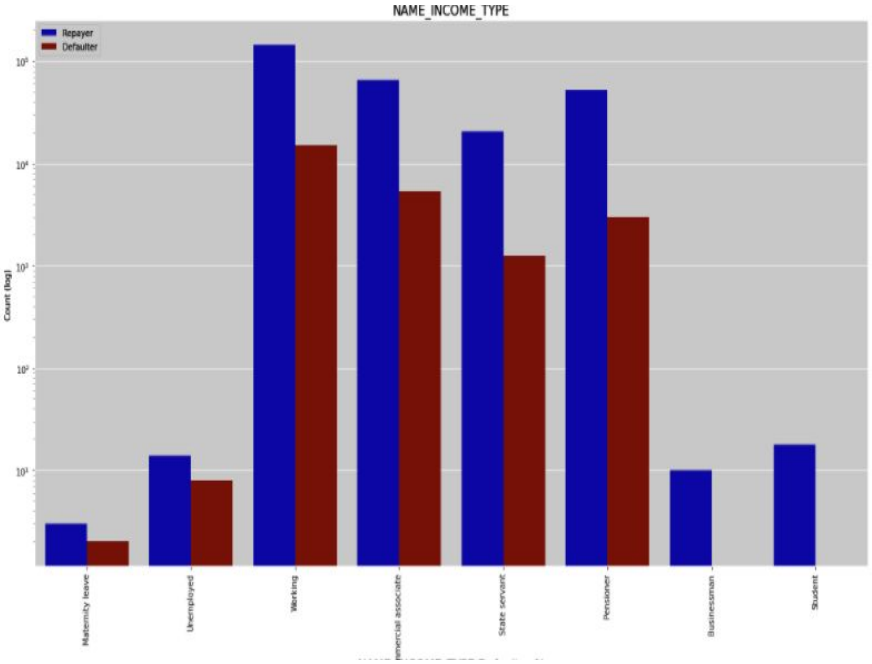
- **NAME_EDUCATION_TYPE -**
    - **Most clients** are having **Secondary/secondary special education**.
    - **Clients with lower secondary** (even though they are of very low numbers) have **high Defaulter% (>10%)** and those with **Secondary/secondary special education (about 9%)**.
    - **Academic degree clients** are the **least Defaulters**.

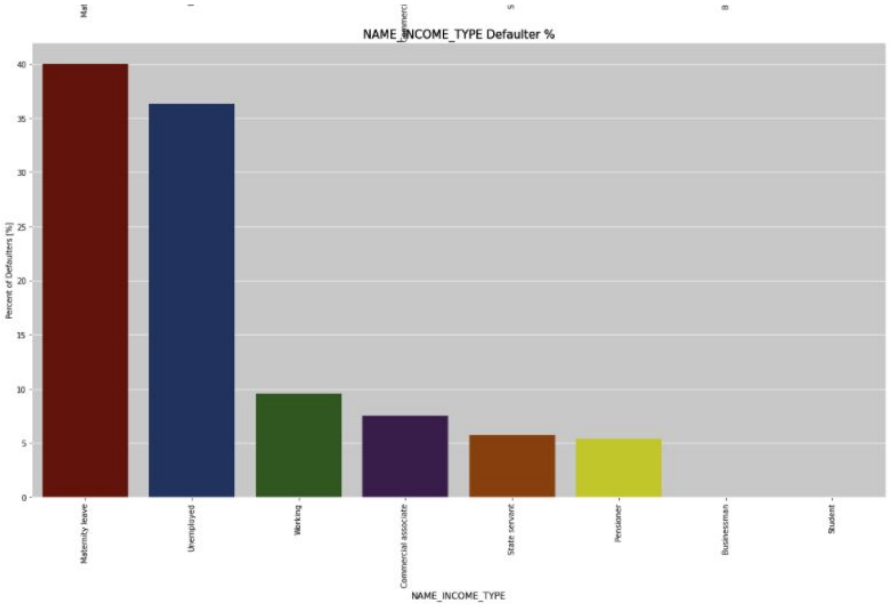




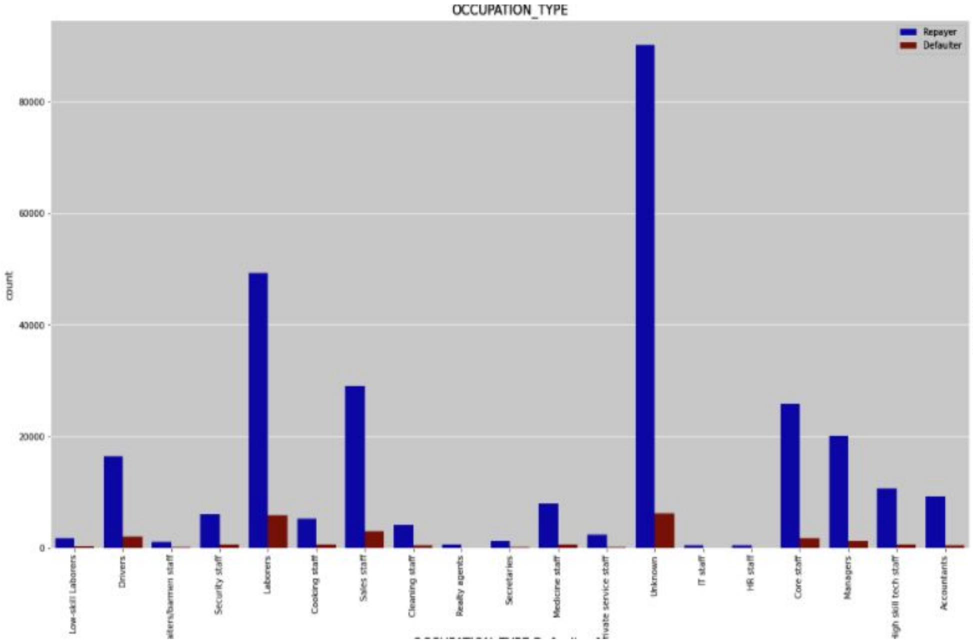
- **NAME_INCOME_TYPE -**

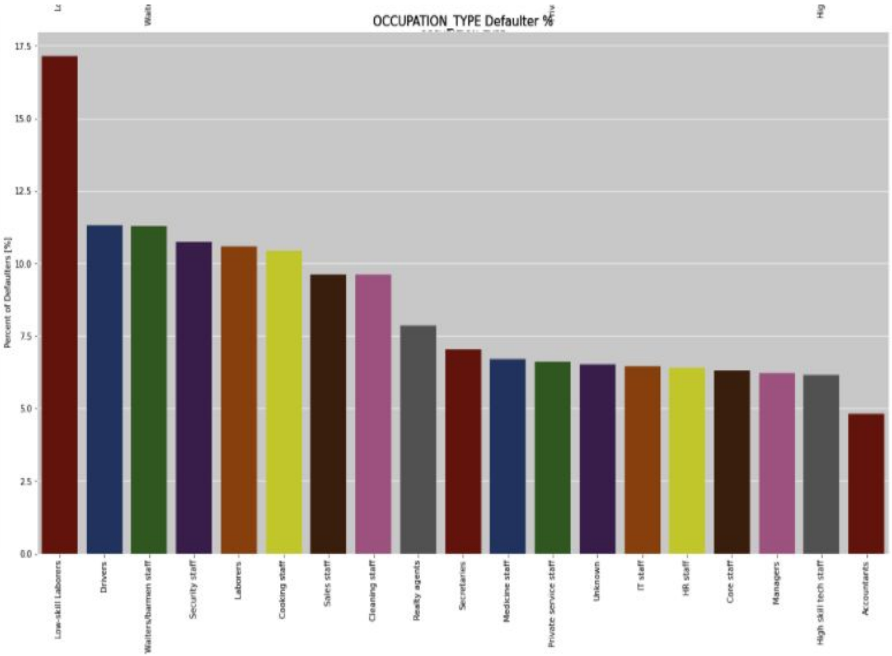
    - **Most clients** have INCOME Category as **Working, Commercial associate, Pensioner and State servant**.
    - **Clients on Maternity leave** (even though they are of very low numbers) have **high Defaulter% (almost 40%)** and those who are **Unemployed (>35%)**.
    - **Pensioner clients** are the **least Defaulters**.
    - **Students and Businessmen**, though less in numbers have **no Default record.** Thus these two categories are **safest for providing loan**.





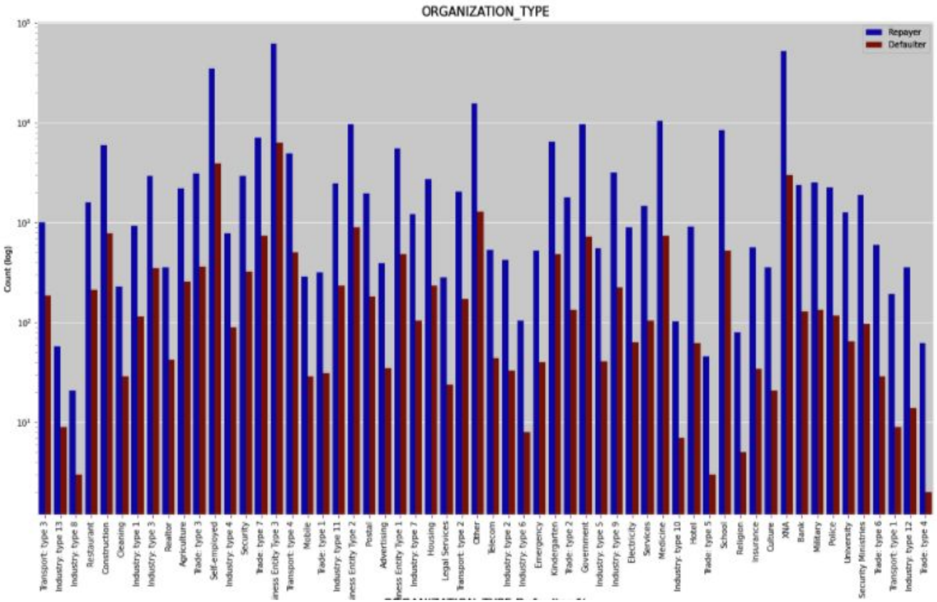
- **OCCUPATION_TYPE -**
    - **Most Clients haven't mentioned** their **Occupation Type**.
    - **Low Skill Laborers** are the **highest Defaulters** (even though they are rare clients) with **>17%, Drivers and Waiters/barmen**, **Security staff, Laborers,** each with **>10%.**
    - **Accountants** are the **least Defaulter Clients with <5%**.

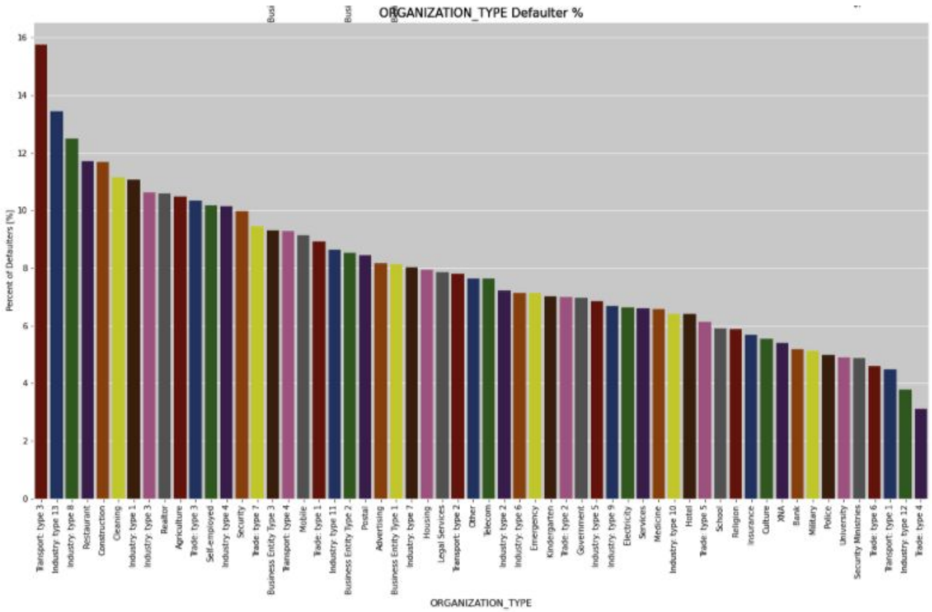




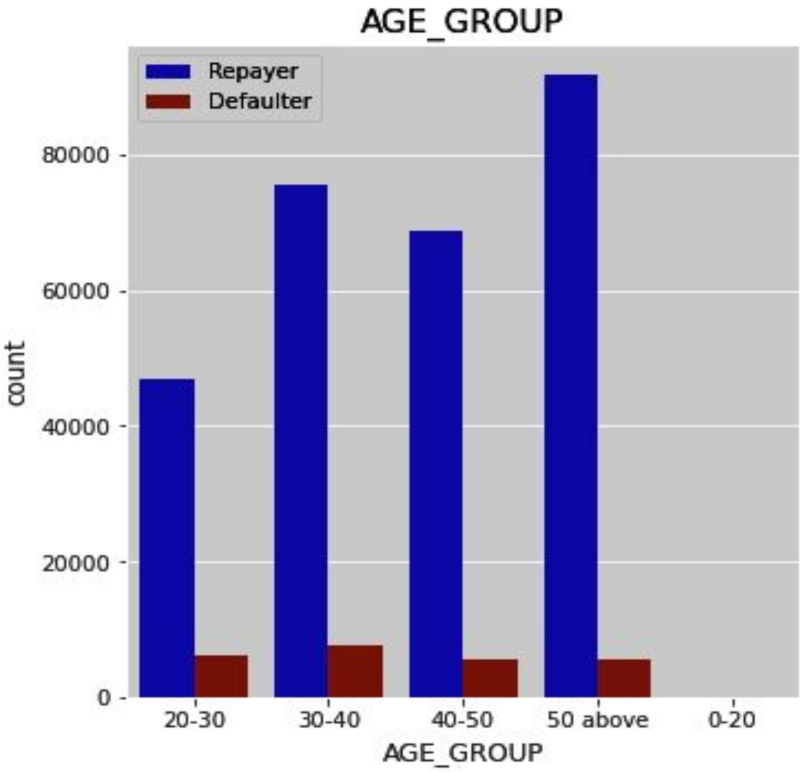
- **ORGANIZATION_TYPE -**

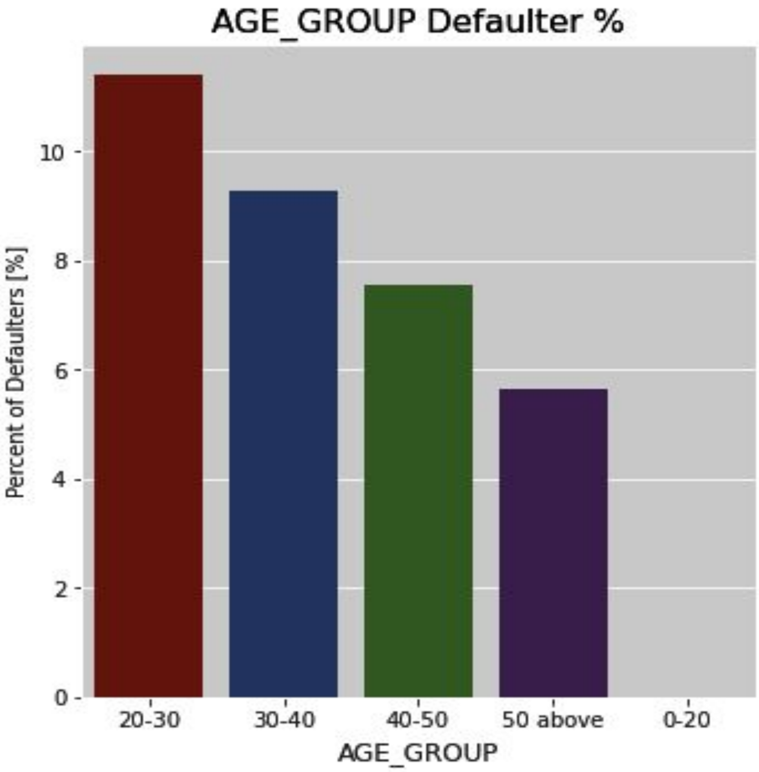
  - Organizations with **highest percent of loans not repaid** are **Transport: type 3 (almost 16%), Industry: type 13 (about 13.5%), Industry: type 8 (about 12.5%), Restaurant and Construction (almost 12% each)**. **Self employed people** have relative **high Default%**.
  - **Most clients** are from **Business Entity Type 3**.
  - **Information is unavailable** for a **large number of clients**.
  - **Trade: type 4 and Industry: type 12** are the **least Defaulters**.



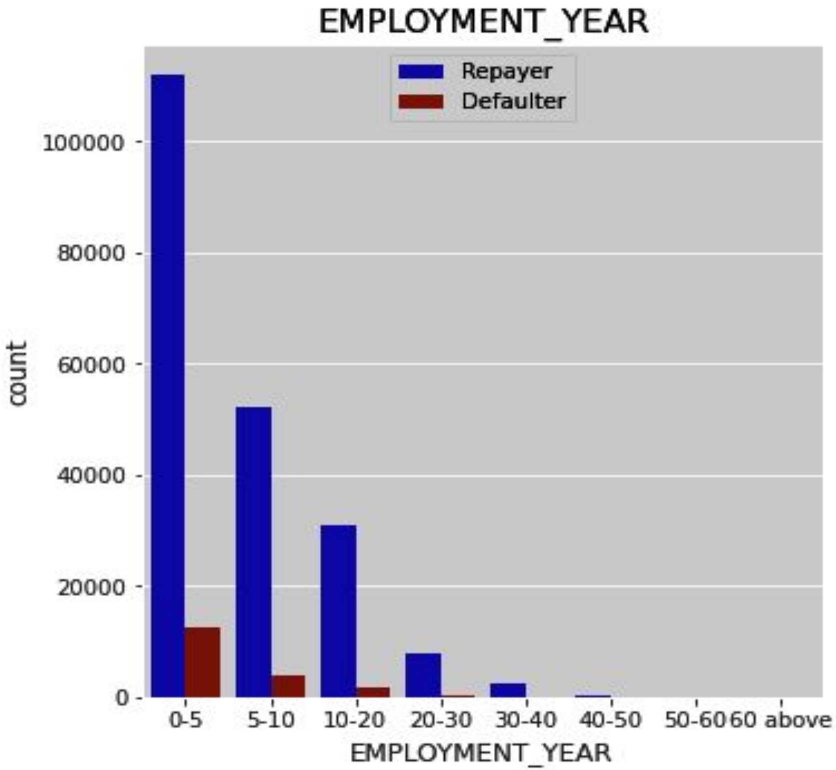


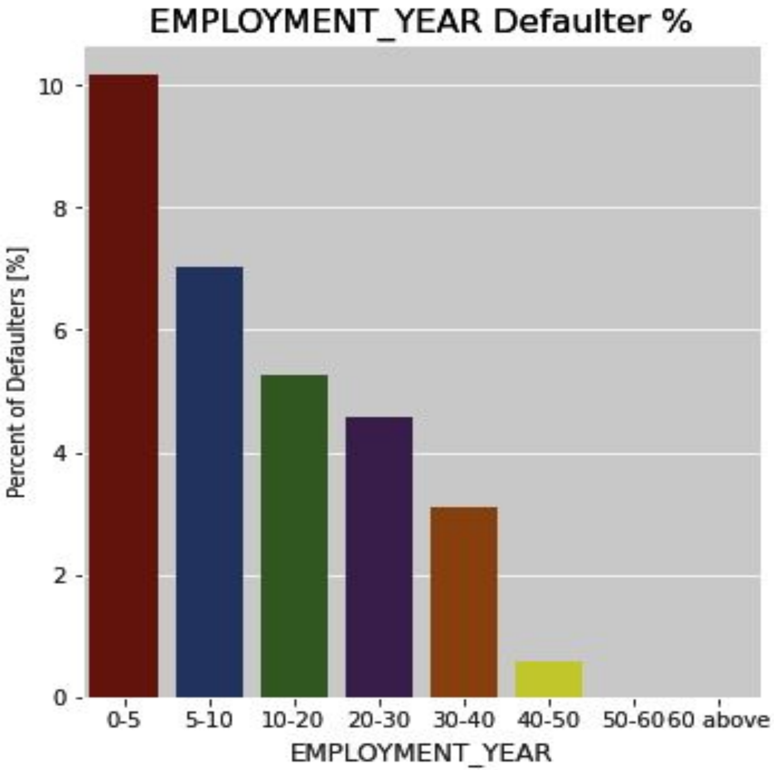
- **AGE_GROUP -**
  - **Most clients** are **above 50 years** of age and are the **Least Defaulters.**
  - **Clients in age group 20-30** are **high Defaulters (>10%)**.



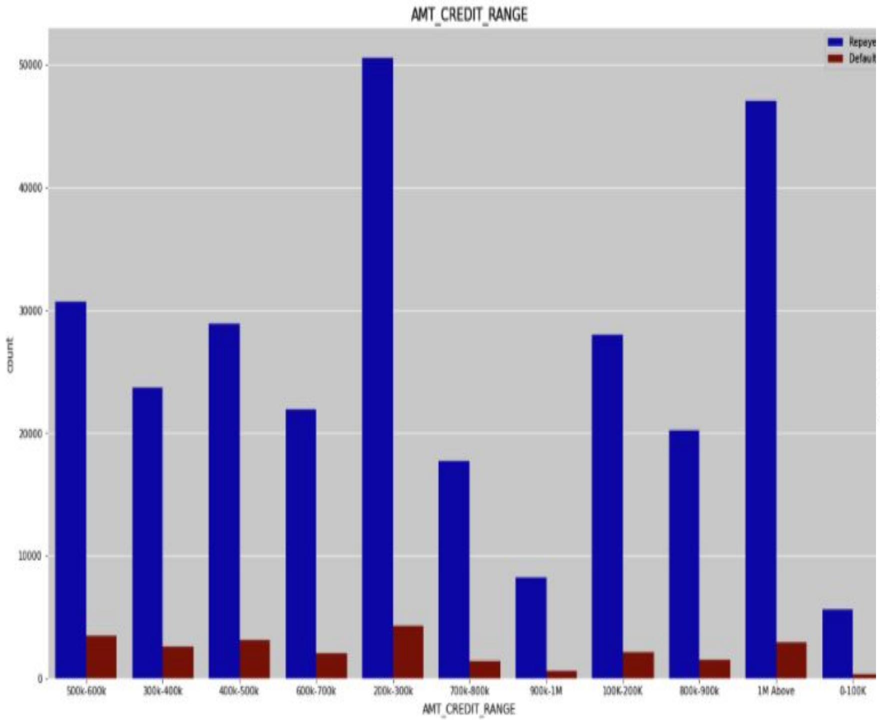


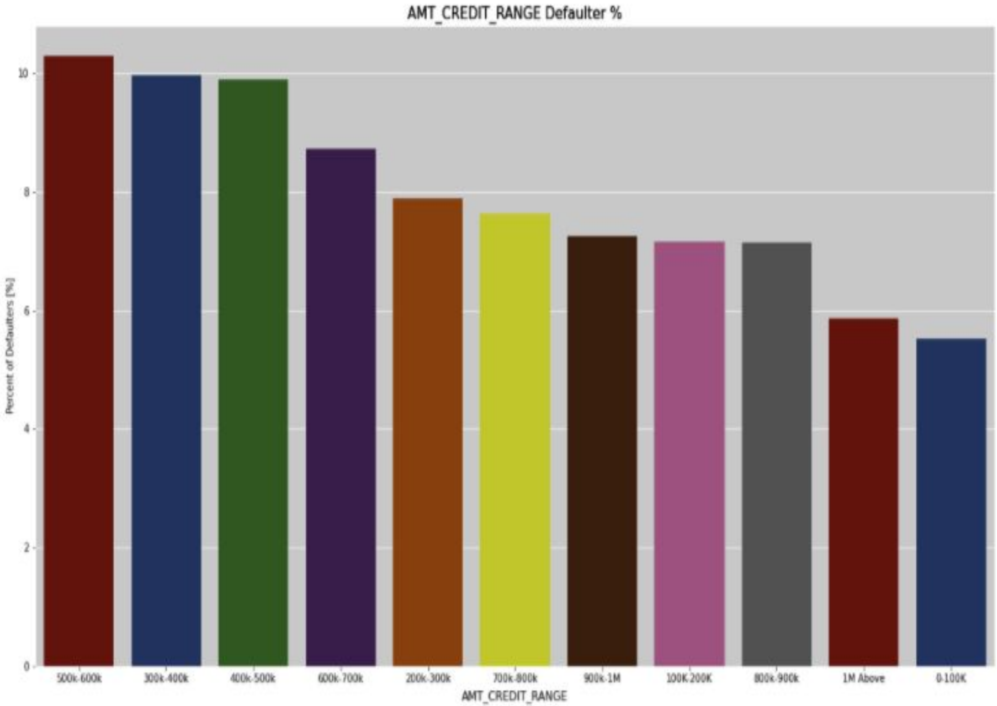
- **EMPLOYMENT_YEAR -**
  - **Most clients** have been **employed for 0-5 years**.
  - With the **increase in years of employment, defaulting % decreases**.



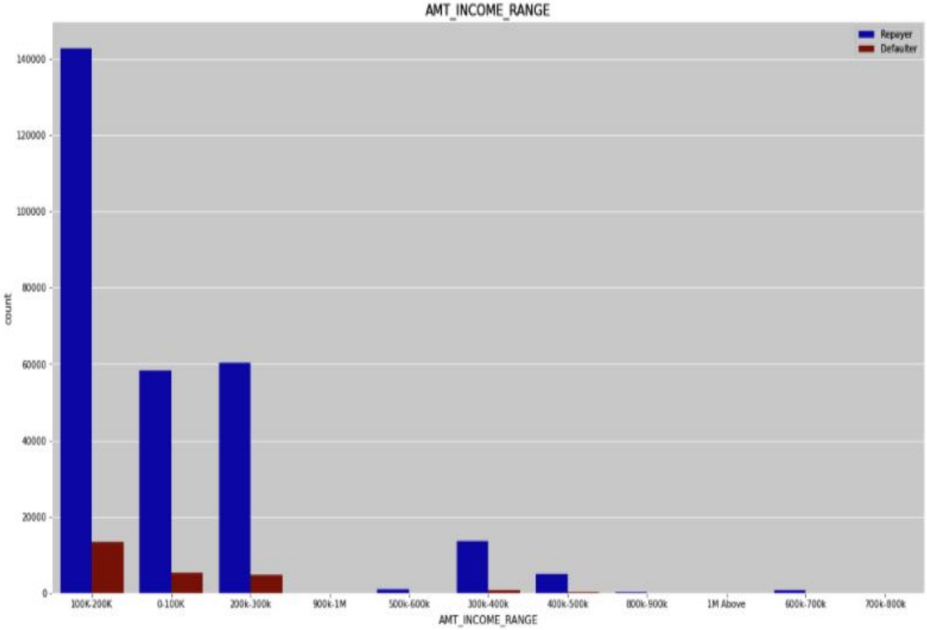


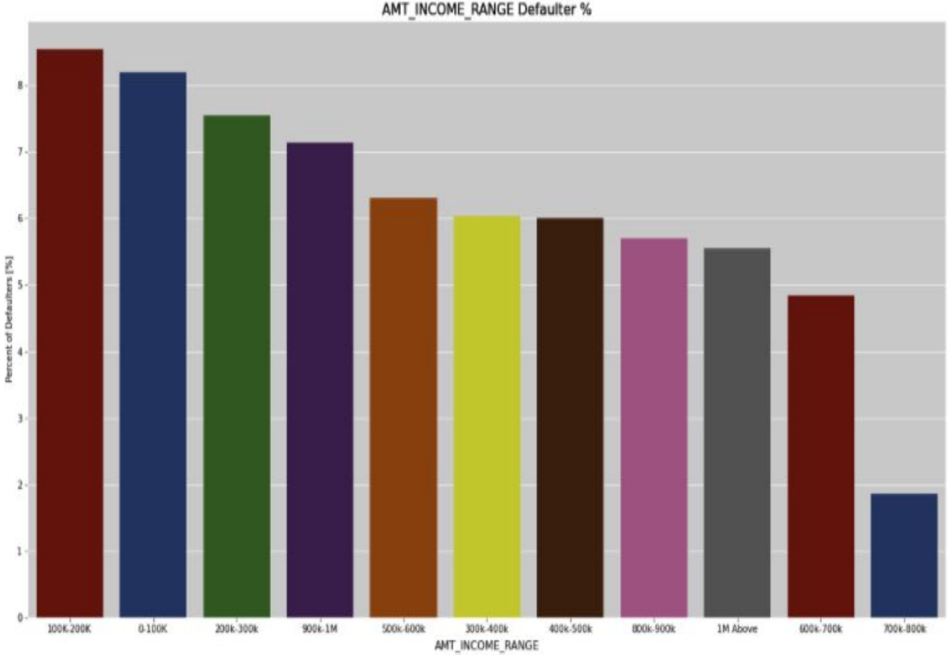
- **AMT_CREDIT_RANGE -**
    - **More than 80% of the loans provided** are for amount **less than 900k**.
    - **Clients with loans for 300-600k** tend to **default more than others**.





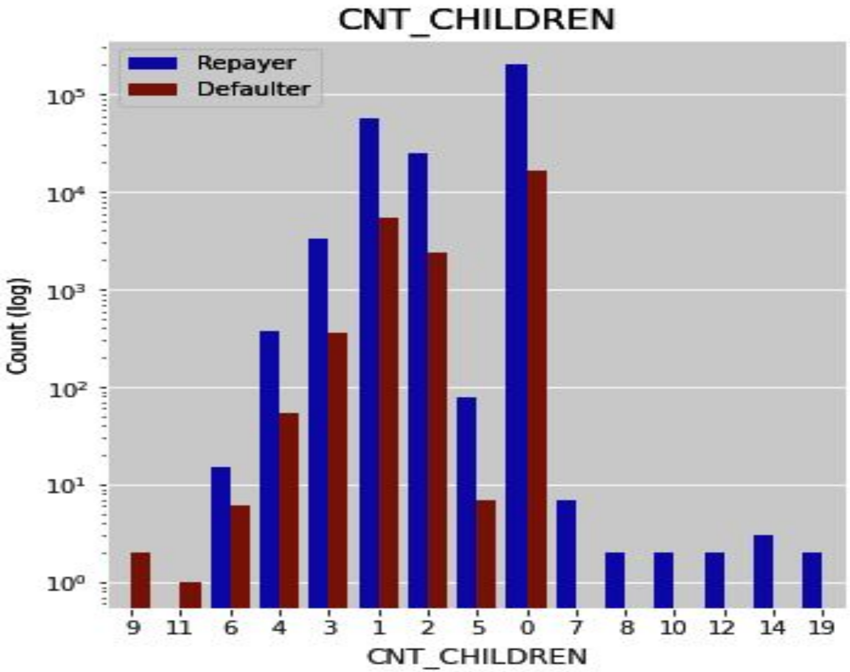
- **AMT_INCOME_RANGE -**
  - **90% of the clients** have **total Income < 300k**.
  - Clients with **Income < 300k** has **higher Defaulting rates**.
  - Clients with **Income > 700k** are **less likely to Default**.

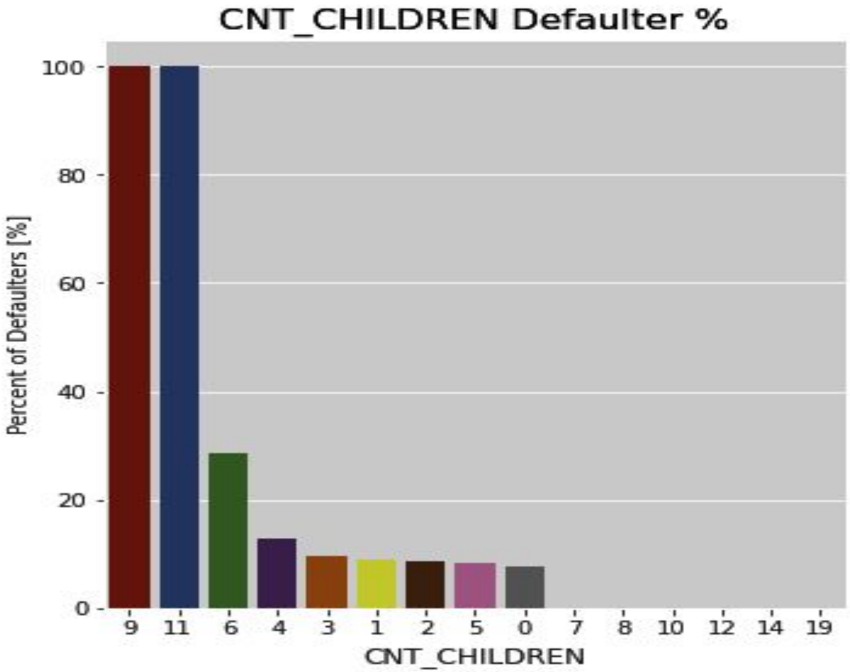




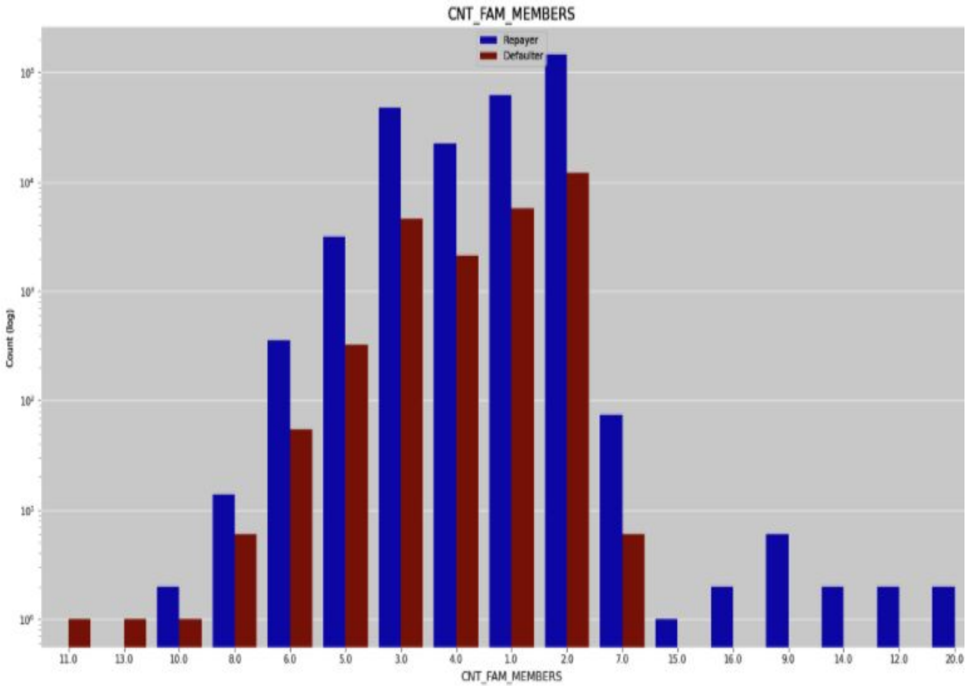
- **CNT_CHILDREN -**

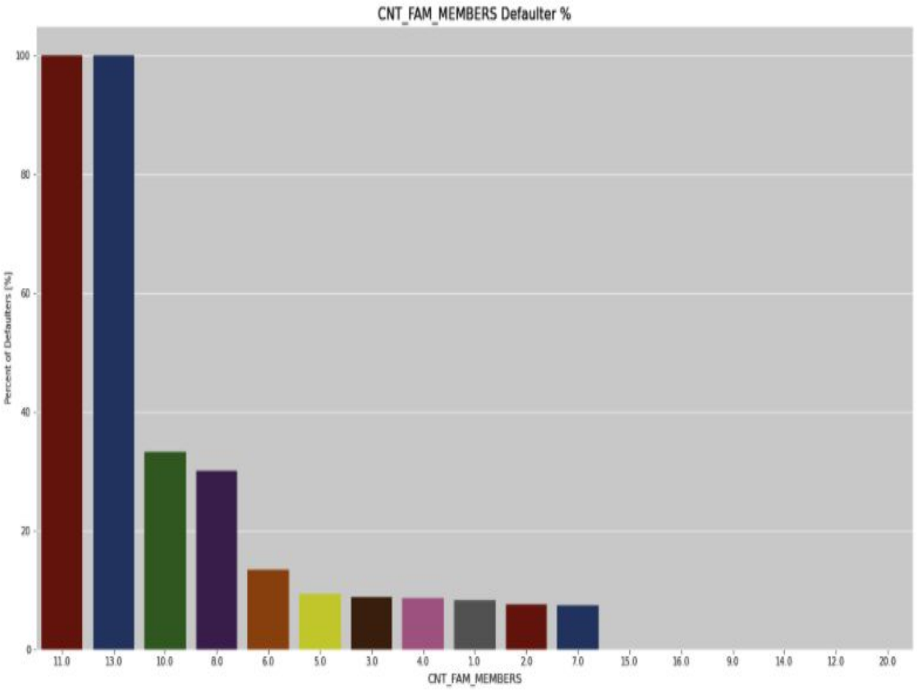
    - **Most** Clients **do not have children.**
    - **Client with > 4 children** have a **very high Default rate**. **100% Default rate** is observed for **clients with 9 or 11 children** (though they are rare applicants).





- **CNT_FAM_MEMBERS -**
  - Family member count follows the same trend as children count where **having more family members increases the risk of Defaulting**.
  - **Clients with 11 or 13 family members** have **100% Default rate**.





- ## Bivariate Analysis -

  - **Top 10 correlations** are found for relevant columns and the **correlating factors are identified** for the **Repayers** and **Defaulters** with the help of **heatmap and pairplot**.

  - For **Repayers**, it is observed that:

    - **Credit amount** is **highly correlated** with **amount of goods price, loan annuity & total income**
    - **Repayers** have **high correlation** in **number of days employed**.

  - For **Defaulters**, it is observed that:

    - **Credit amount** is **highly correlated** with **amount of goods price** which is similar as for Repayers.
    - The l**oan annuity correlation with credit amount** and **correlation for Employment days** have **slightly reduced for Defaulters**.

- **Top 10 Correlation -**

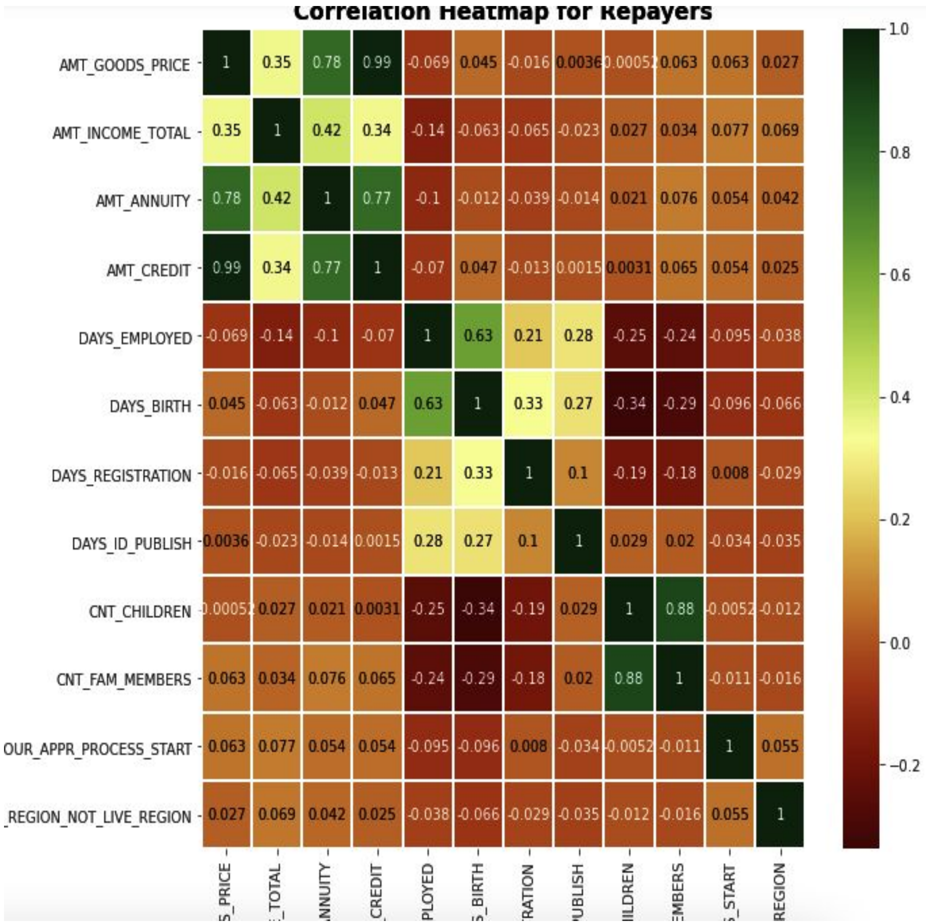
Top 10 correlation for Repayers:

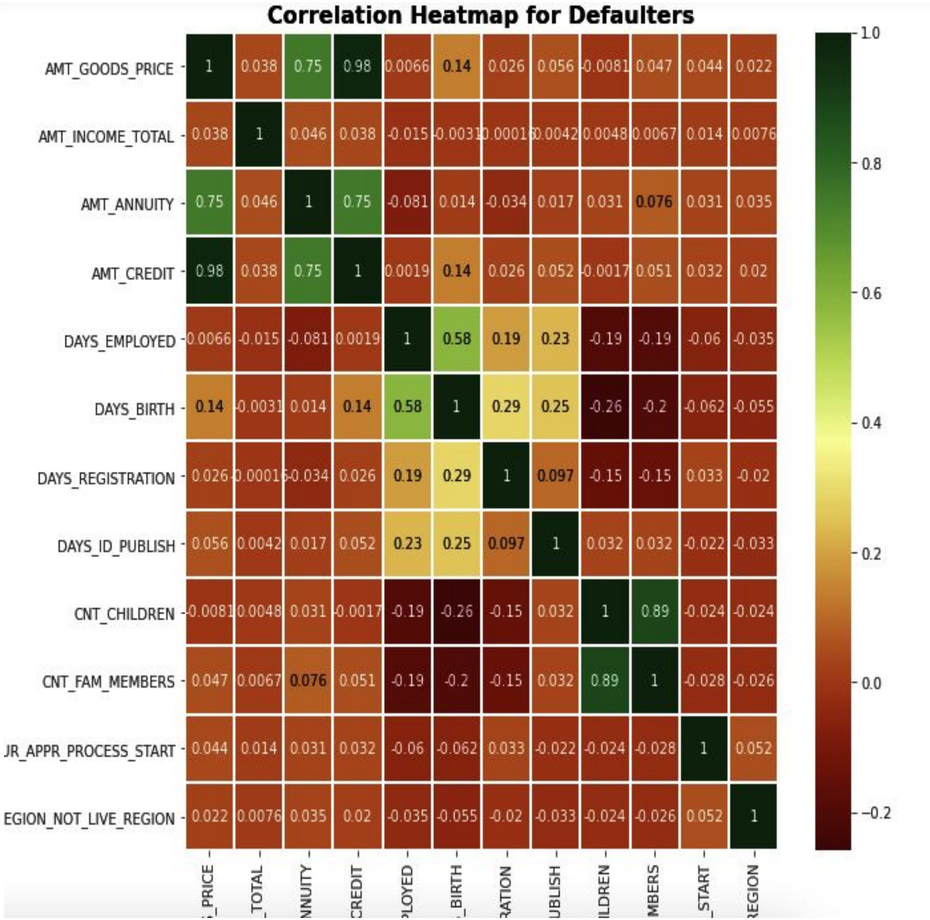
	VAR1	VAR2	Correlation
36	AMT_CREDIT	AMT_GOODS_PRICE	0.987250
116	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
24	AMT_ANNUITY	AMT_GOODS_PRICE	0.776686
38	AMT_CREDIT	AMT_ANNUITY	0.771309
64	DAYS_BIRTH	DAYS_EMPLOYED	0.626114
25	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953
12	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.349462
37	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799
101	CNT_CHILDREN	DAYS_BIRTH	0.336966
77	DAYS_REGISTRATION	DAYS_BIRTH	0.333151

Top 10 correlation for Defaulters:

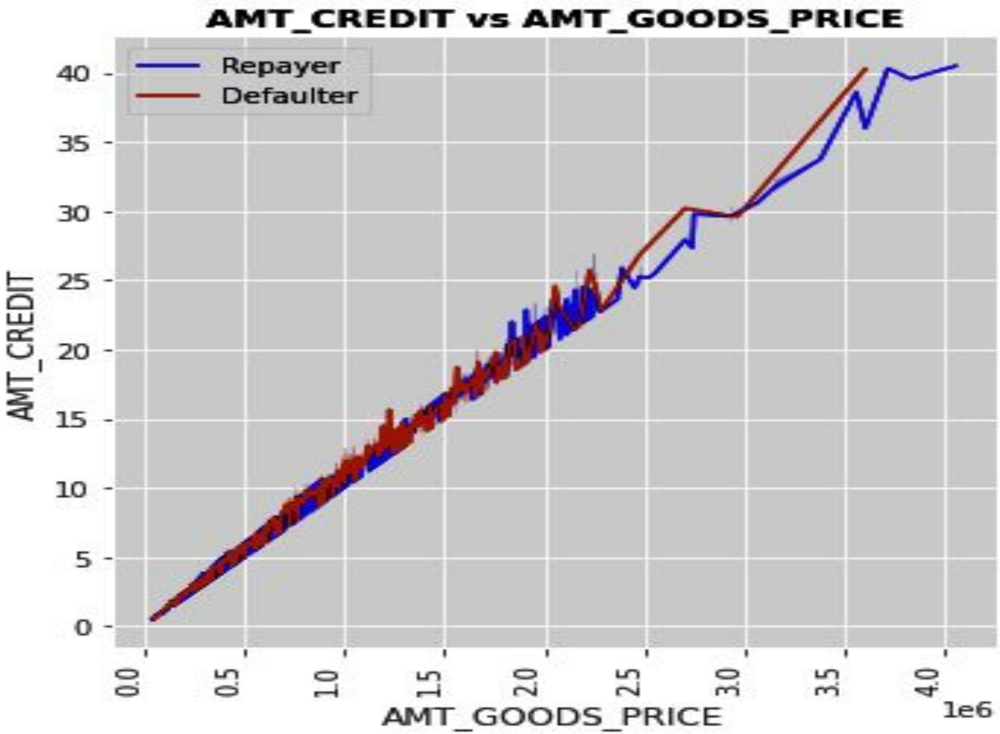
	VAR1	VAR2	Correlation
36	AMT_CREDIT	AMT_GOODS_PRICE	0.983103
116	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
24	AMT_ANNUITY	AMT_GOODS_PRICE	0.752699
38	AMT_CREDIT	AMT_ANNUITY	0.752195
64	DAYS_BIRTH	DAYS_EMPLOYED	0.582185
77	DAYS_REGISTRATION	DAYS_BIRTH	0.289114
101	CNT_CHILDREN	DAYS_BIRTH	0.259109
89	DAYS_ID_PUBLISH	DAYS_BIRTH	0.252863
88	DAYS_ID_PUBLISH	DAYS_EMPLOYED	0.229090
113	CNT_FAM_MEMBERS	DAYS_BIRTH	0.203267

○ Correlation Heatmaps -

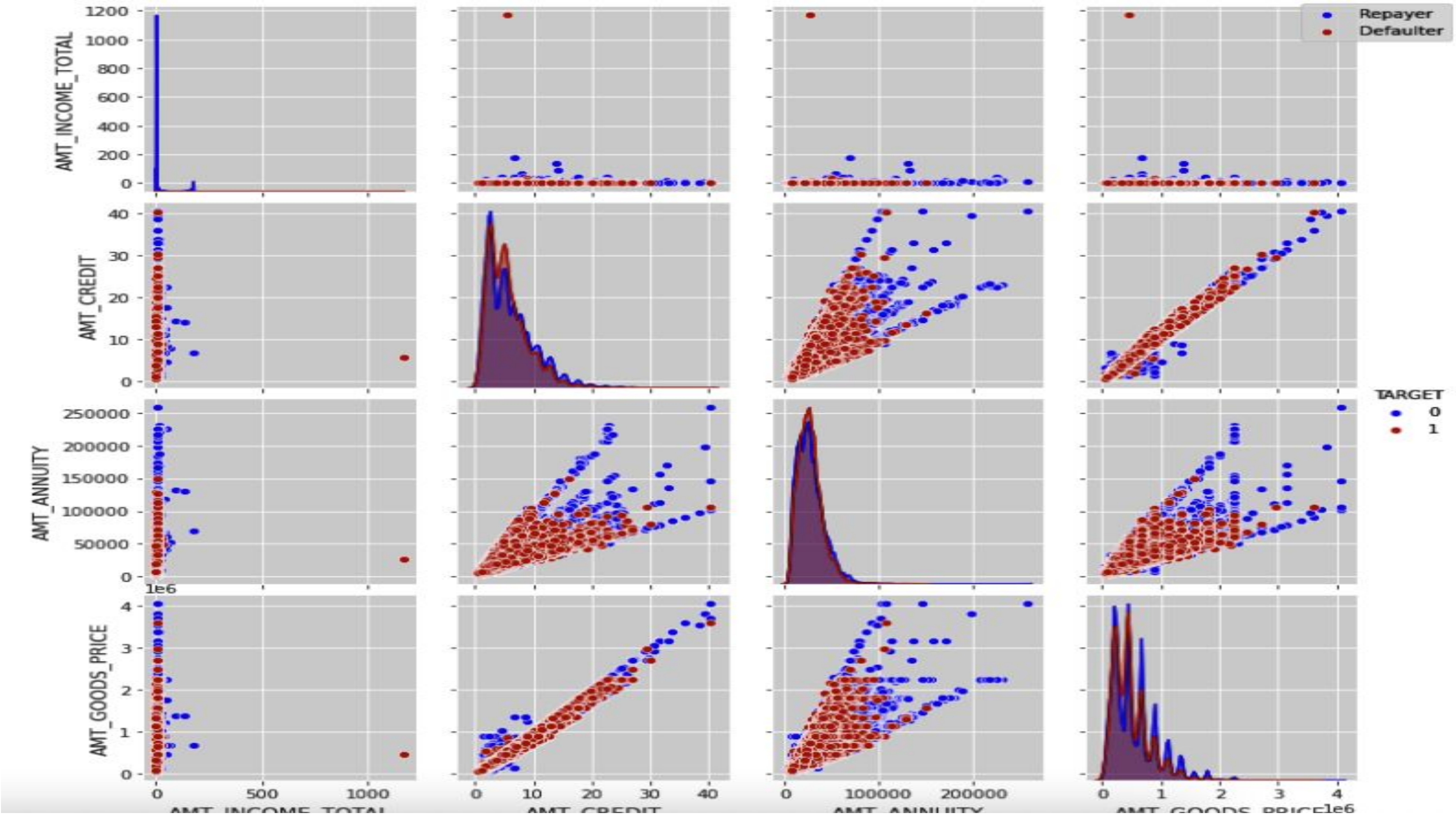




- **Relational plot for Credit Amount and Good Price -**
    - It is observed that when the **credit amount exceeds 3 millions for amount goods price**, there is an **increase in Defaulters.**



- ○ **Pairplot between Amount variables -**
  - It is observed that **AMT_CREDIT and AMT_GOODS_PRICE** are **highly correlated**.



- ## Merged Dataframe Analysis -

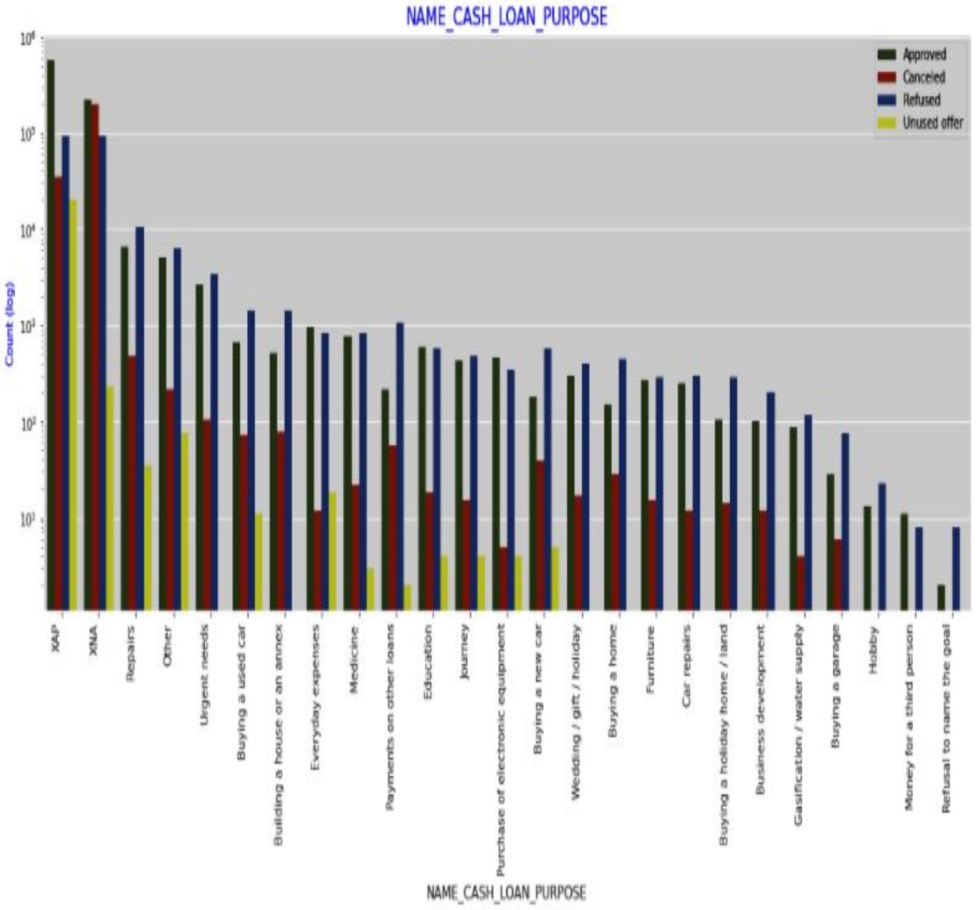
    - Both **‘application_df’** and **‘previous_df’** are merged based on the **current application id** and **common records** are analysed.

    - **Repayer and Defaulter records** are **segregated into two separate dataframes** and **impact of the Decisions taken** for previous loans are **analysed** for relevant records.

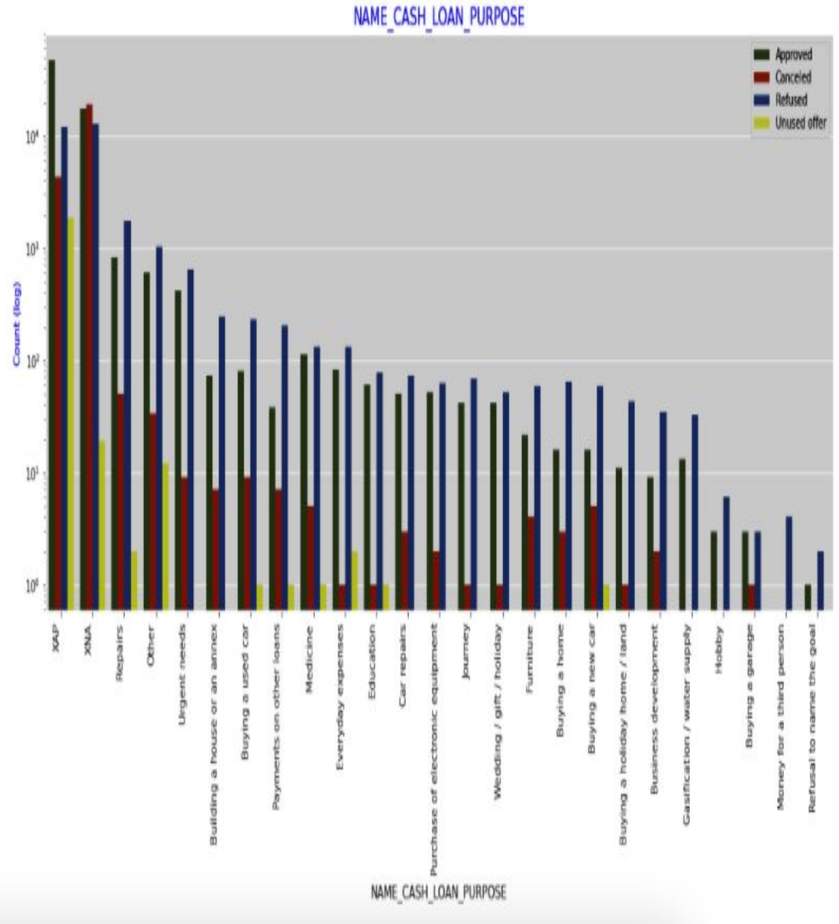
    - Observations for **decisions taken based on loan purpose:**

        - Loan purpose has **high number of unknown values (XAP, XNA)**.
        - Loans taken for the **purpose of Repairs** seems to have **highest Default rate**.
        - A **very high number of applications for the purpose of 'Repairs' and 'Others'** has been **Refused by Company or Canceled by Client.**

For Repayers:



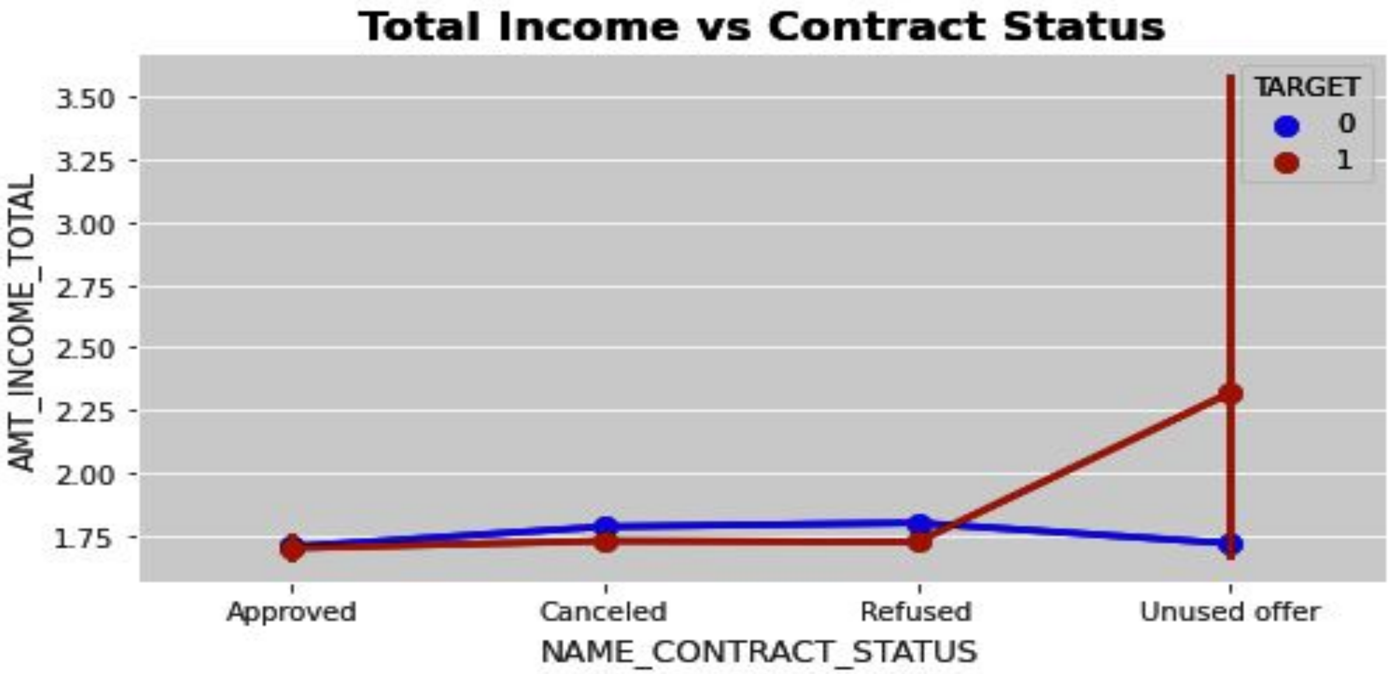
For Defaulters:



- Observations for **decisions taken to identify business or financial loss** -
  - **About 90% of the loans have been Repayed** for cases where the **Client Canceled their application**.
  - **88% of the Clients** who have been previously **Refused a loan by the Company**, have **Repayed the Current loan.**

NAME_CONTRACT_STATUS	TARGET	Counts	Percentage
Approved	0	818856	92.41%
	1	67243	7.59%
Canceled	0	235641	90.83%
	1	23800	9.17%
Refused	0	215952	88.0%
	1	29438	12.0%
Unused offer	0	20892	91.75%
	1	1879	8.25%

- Relationship between **total income and contract status** -
  - It is observed that the **Clients with Unused offer** earlier have **Defaulted** even when their **average income is higher than others**.



Conclusion

- Based on our analysis, the indicators of an applicant to be a **Repayer** or a **Defaulter** can be summarized as below:

Column	Repayer	Defaulter
Level of Education	Academic degree	Lower Secondary & Secondary/secondary special
Income type	Students and Businessmen	Clients on Maternity leave or Unemployed
Organisation type	Trade: type 4 and Industry: type 12	Transport: type 3 (almost 16%), Industry: type 13 (about 13.5%), Industry: type 8 (about 12.5%), Restaurant and Construction (almost 12% each); Self employed people
Age group	Above the age of 50 years	Age group of 20-40 years
Employment years	40+ years of employment	less than 5 years of employment
Income range	Income more than 700,000	less than 300,000
No. of children	zero to two children	more than 8 children
Family Status	Widow clients	Civil marriage and Single/not married
Credit amount	Below 1 million	Beyond 3 million
Loan purpose	Hobby, Buying garage	Repairs, Others

- In order to **mitigate the risks of business loss or financial loss**, the following suggestions can be implemented:
    - **About 90% of the loans have been repayed** for cases where the **Client Canceled their application** previously. Thus, **recording the reason for cancelation** can **help the Company to determine and negotiate terms** with these repaying Customers in future for **increasing their business opportunity.**
    - **88% of the Clients** who have been **previously Refused a loan by the Company**, have **now turned into a repaying Client.** Hence, **documenting the reason for rejection** can **mitigate the business loss** and these clients could be contacted for future loans.