Credit EDA Assignment

Presentation by-

Anugrah Stanley

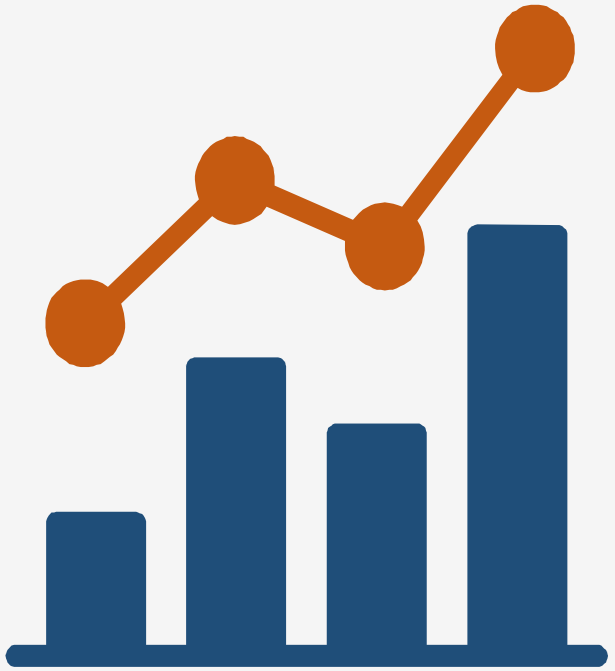
Objective

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 utilize this knowledge for its portfolio and risk assessment.

Work Flow

1. Importing Libraries
2. Data Loading and Sanity Checks
3. Data Cleaning
4. Data Analysis
    - Data Imbalance Analysis
    - Univariate and Segmented Univariate Analysis
    - Bivariate and Multivariate Analysis



1 Importing Libraries

For Mathematical Functions

For Data Manipulation and Analysis

For Visualizations

For Visualizations

2 Data Loading and Sanity Checks

❖ For application_data dataset

- ✓ Loaded the data set using pandas
- ✓ The data frame contains 121 features, 1 target variable and 307511 rows.
- ✓ The data types of features are as follows:
  a. 65 features are float64
  b. 41 features are integer
  c. 16 features are object
- ✓ There are a lot of features with null values but datatypes of these features looks fine.

❖ For previous_application dataset

- ✓ Loaded the data set using pandas
- ✓ The data frame contains 37 features and 1670214 rows.
- ✓ The data types of features are as follows:
  a. 15 features are float64
  b. 06 features are integer
  c. 16 features are object
- ✓ There are a lot of features with null values but datatypes of these features looks fine.

3 Data Cleaning: Missing Value Treatment

❖ For application_data dataset

- ✓ Dropped 49 features which had Null values more than 40%.
- ✓ Features EXT_SOURCE_1 and EXT_SOURCE_2 had Null values but these features showed no correlation with the TARGET variable, hence dropped them.
- ✓ Imputed the existing 31% Null values of feature "OCCUPATION_TYPE" with "Unknown" because imputing it with any existing category might influence the analysis.
- ✓ Imputed the existing 13% Null Values in AMT_REQ_X features with the "Median" because they represent number of enquiries made which cannot be decimal.
- ✓ Imputed the Null Values of feature NAME_TYPE_SUITE with "Unknown“.
- ✓ Imputed marginal existent Null Values of SOCIAL_CIRCLE_X features and DAYS_LAST_PHONE_CHANGE with "Mode".
- ✓ Imputed marginal existent Null Values of the features "AMT_GOODS_PRICE","AMT_ANNUITY","CNT_FAM_MEMBERS" with "Median" due to the presence of outliers.
- ✓ The final shape of the "application_data" dataset stands at (307511, 71) after dropping/imputing all the Null Values.

❖ For previous_application dataset

- ✓ Dropped 11 features which had Null values more than 40%.
- ✓ Imputed the existing Null values of features "AMT_ANNUITY","AMT_GOODS_PRICE" and "AMT_CREDIT" with median, mode and mode respectively because of the presence of outliers and to preserve the original distribution.
- ✓ Imputed Null Values of CNT_PAYMENT with 0 as the NAME_CONTRACT_STATUS for these indicate that most of these loans were not started.
- ✓ Imputing the Null values in PRODUCT_COMBINATION categorical feature with "Mode" as Null values are insignificant in numbers.
- ✓ The final shape of the "previous_application" dataset stands at (1670214, 26) after dropping/imputing all the Null Values.

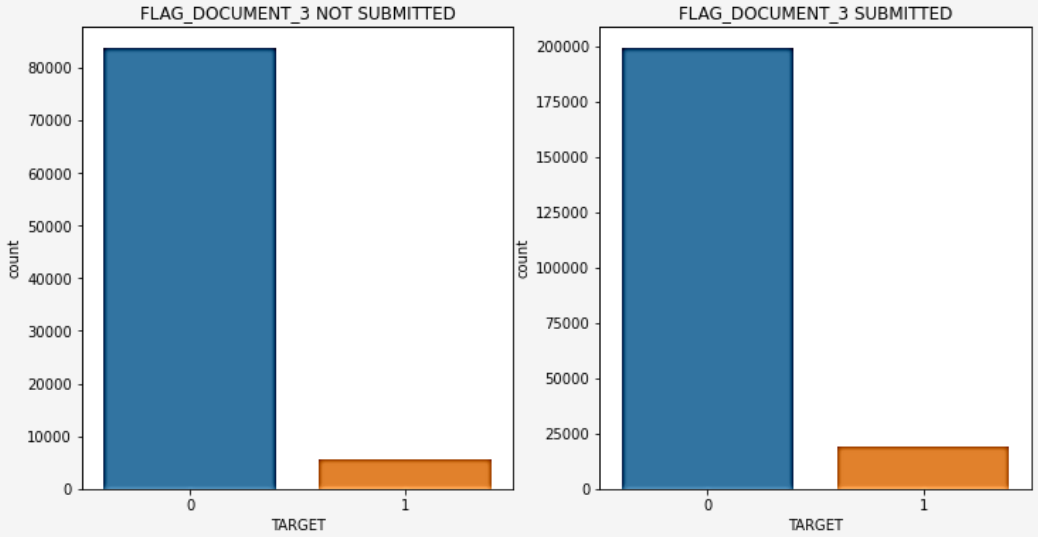
3 Data Cleaning: Fixing Rows and Columns

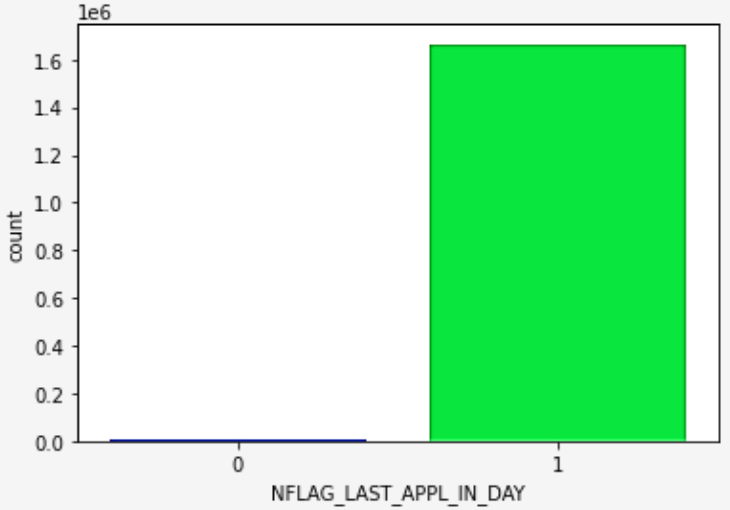
❖ For application_data dataset

- ✓ Upon analyzing the FLAG_DOCUMENT_X features, it was found that most of the applicants have not submitted FLAG_DOCUMENTS_X (except DOC_3) hence dropped those features.
- ✓ Upon further analysis of FLAG_DOCUMENT_3 feature, it was found that it does not bear an impact on the TARGET variable, hence it was dropped.
- ✓ No great correlation was found between the features 'FLAG_MOBIL', 'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE','FLAG_EMAIL' and the 'TARGET' variable, hence these features were dropped.
- ✓ The final shape of the "application_data" dataset after dropping unnecessary features stands at (307511, 45).

❖For previous_application dataset

- ✓ Dropped the feature FLAG_LAST_APPL_PER_CONTRACT and NFLAG_LAST_APPL_IN_DAY as the data is greatly imbalanced.
- ✓ Dropped HOUR_APPR_PROCESS_START as this feature didn't look relevant.
- ✓ The final shape of the "previous_application" dataset after dropping unnecessary features stands at (1670214, 23).





3 Data Cleaning: Fixing Incorrect/Invalid/Unknown Values

❖ For application_data dataset

- ✓ Converted values in features 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'DAYS_LAST_PHONE_CHANGE' from negative to positive as days cannot be negative.
- ✓ Replaced the marginal number of value "XNA" in CODE_GENDER feature with most frequent value "F".
- ✓ Replaced the marginal number of value "Unknown" in NAME_FAMILY_STATUS feature with most frequent value "Married".
- ✓ Replaced the significant number of value "XNA" in ORGANIZATION_TYPE" feature with "Unknown" as replacing it with any other category might influence the analysis.
- ✓ Derived AGE from DAYS_BIRTH, YEARS_EMPLOYED from DAYS_EMPLOYED, YEARS_REGISTRATION from DAYS_REGISTRATION, YEARS_ID_PUBLISH from DAYS_ID_PUBLISH, YEARS_LAST_PHONE_CHANGE from DAYS_LAST_PHONE_CHANGE feature for the ease of analysis.

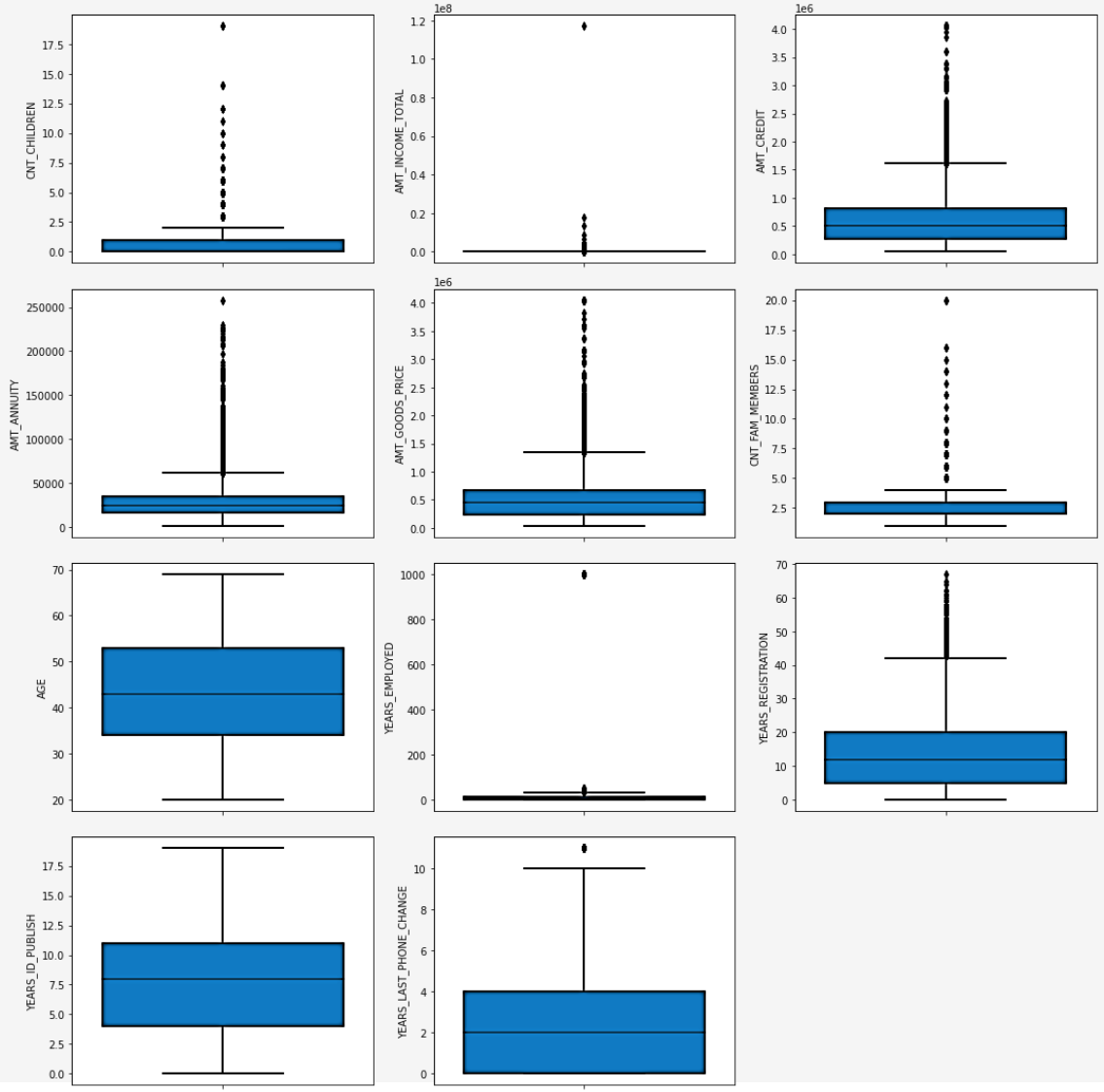
❖ For previous_application dataset

- ✓ Converted values in features 'DAYS_DECISION' from negative to positive as days cannot be negative.
- ✓ Some features have values XAP, XNA in huge numbers. We will continue our analysis without changing these values.
- ✓ Derived YEARS_DECISION from DAYS_DECISION feature from the existing feature for ease of analysis.

3 Data Cleaning: Identifying and Handling Outliers

❖ For application_data dataset

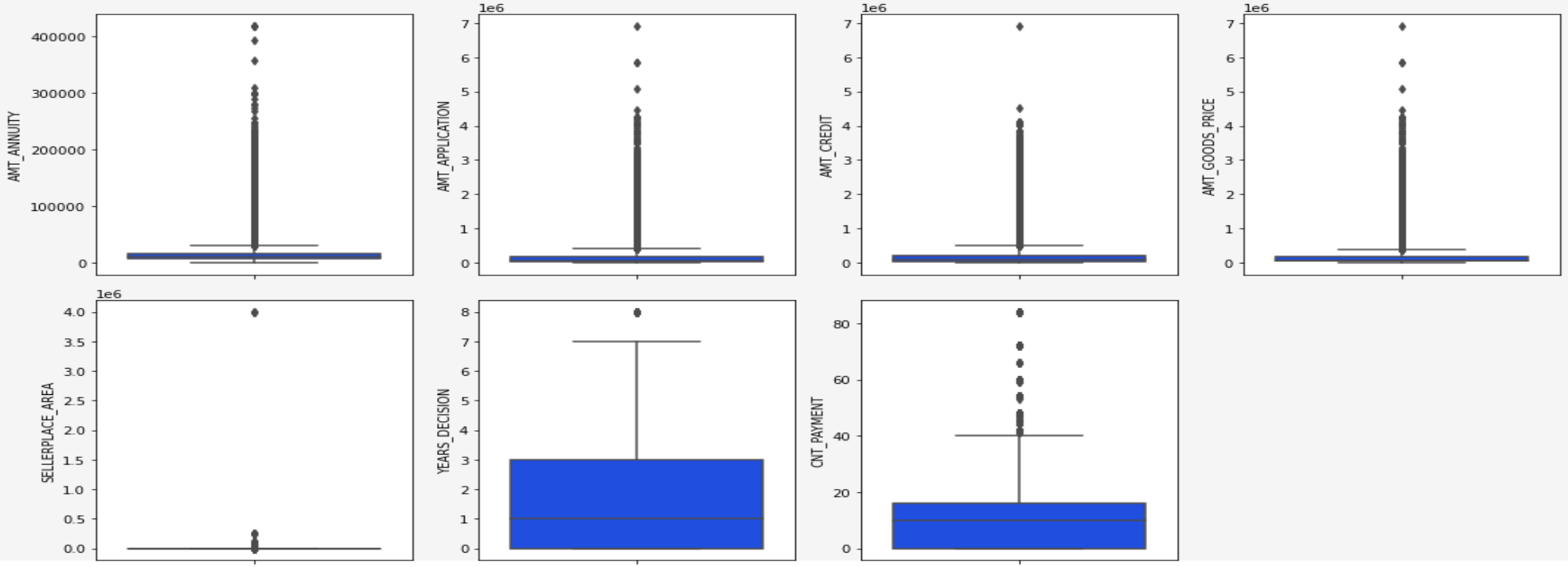
- ✓ CNT_CHILDREN, CNT_FAMILY_MEMBERS features have some number of outliers but the values can't be considered wrong as people can have any number of children or family members.
- ✓ AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE features have outliers but the values are valid.
- ✓ AGE feature has no outliers.
- ✓ YEARS_EMPLOYED has outliers with value around 1000 years, which is practically impossible. However, on further analysis, it was found that they belong to INCOME_TYPE "Pensioner" or "Unemployed". Hence, this insight would be considered during analysis of this feature.
- ✓ YEARS_REGISTRATION feature has outliers but the values are valid.
- ✓ YEARS_ID_PUBLISH feature has no outliers.
- ✓ YEARS_LAST_PHONE_CHANGE has outliers 11, but the values are valid.
- ✓ Created bins for these features for further analysis.



3 Data Cleaning: Identifying and Handling Outliers

❖ For previous_application dataset

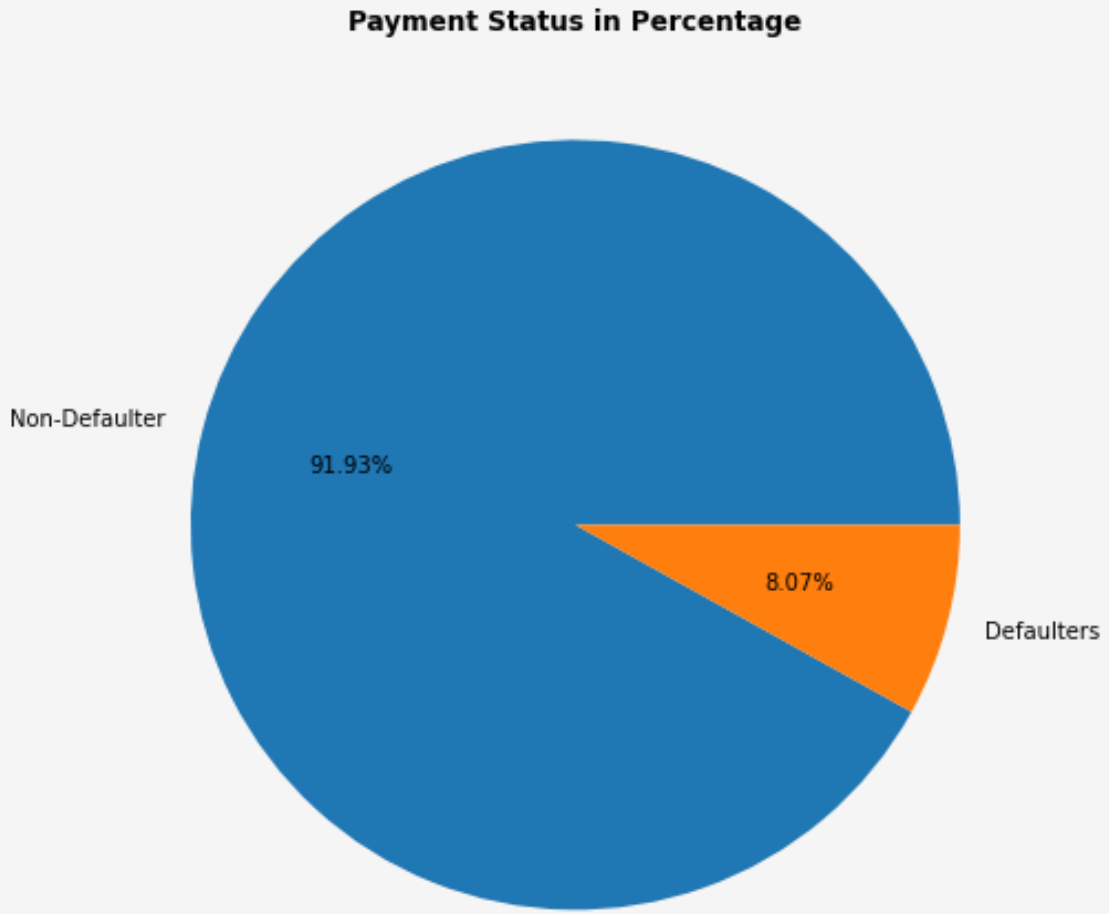
- ✓ AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers but the values are valid.
- ✓ CNT_PAYMENT has few outlier values but valid values.
- ✓ YEARS_DECISION has few outliers indicating that these previous applications decisions were taken long back.



4 Data Analysis: Data Imbalance Analysis

❖ Insights

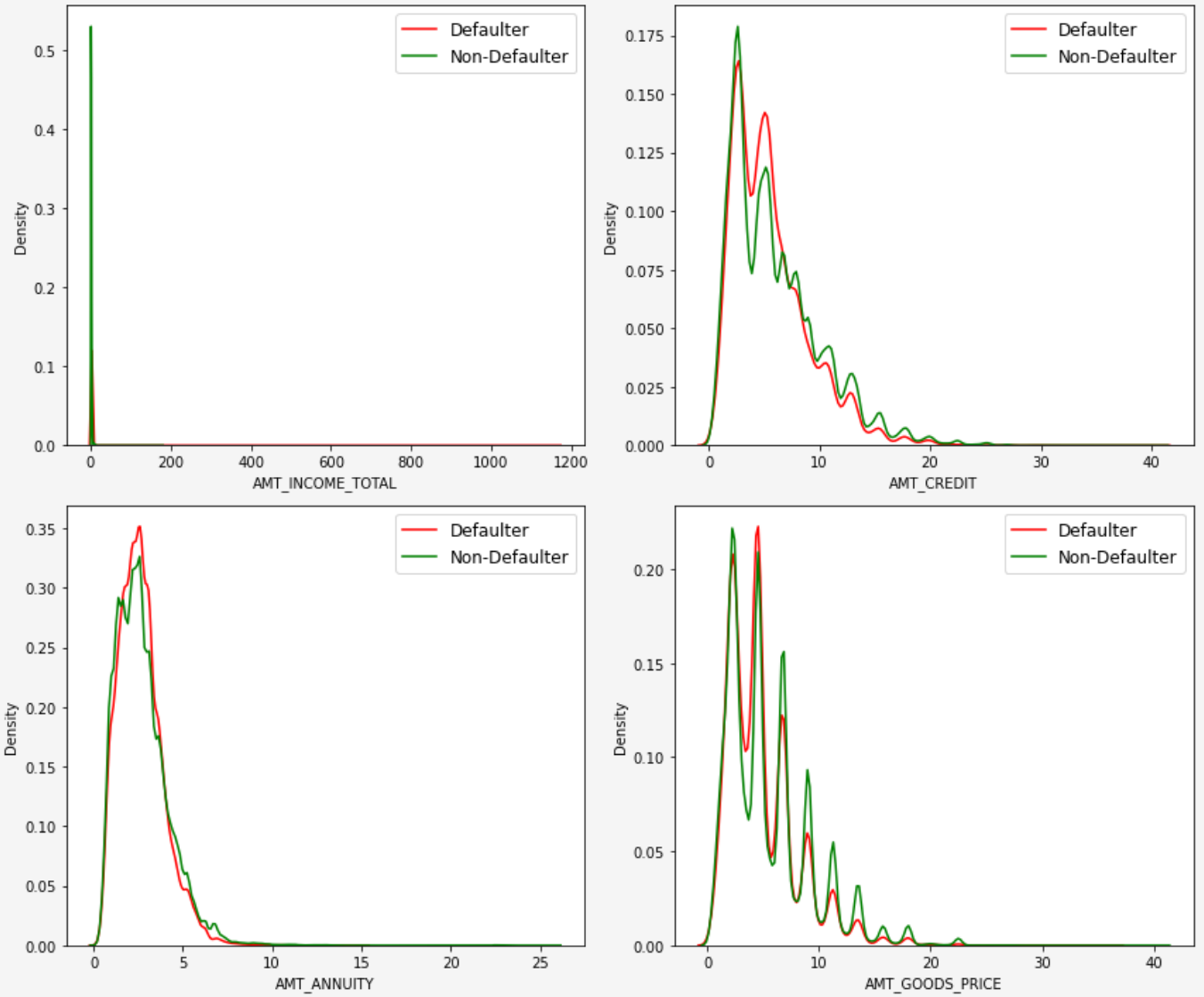
- ✓ The data is highly imbalanced as number of Clients with Payment Difficulties are very less in total population.
- ✓ There are 91.93% On-Time Payment Clients and 8.07% Clients with Payment Difficulties.
- ✓ There are 24825 Clients with Payment Difficulties and 282686 On-TIme Payment Clients.
- ✓ Ratio of Client with Payment Difficulties & On-Time Payment Clients is 11.39%.



4 Data Analysis: Univariate and Segmented Univariate Analysis

❖ Insights

- ✓ Most number of loans are given for goods price below 10 lakhs.
- ✓ Most people pay annuity below 50000 for the credit loan.
- ✓ Credit amount of the loan is mostly less then 10 lakhs.
- ✓ The non-defaulters and defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision.



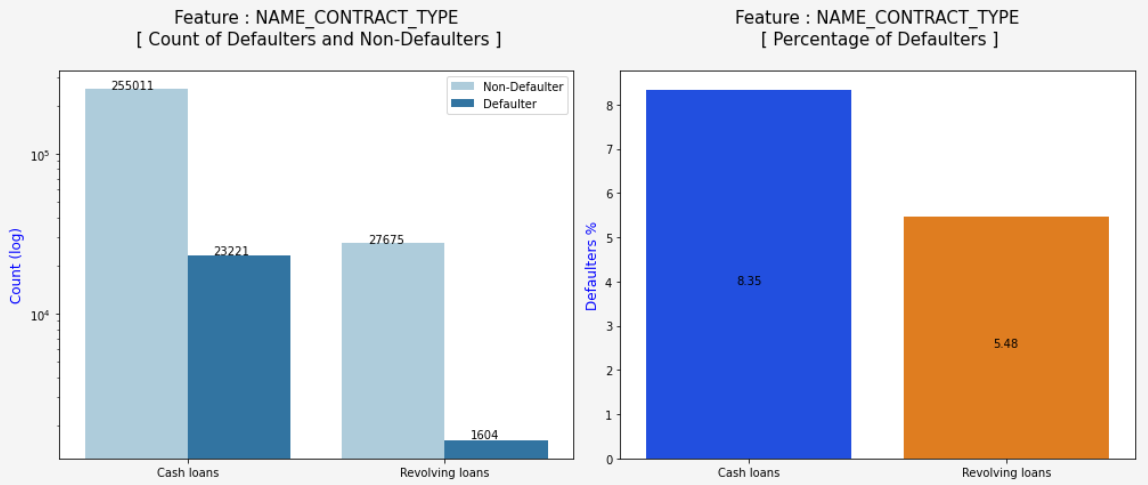
4 Data Analysis: Univariate and Segmented Univariate Analysis

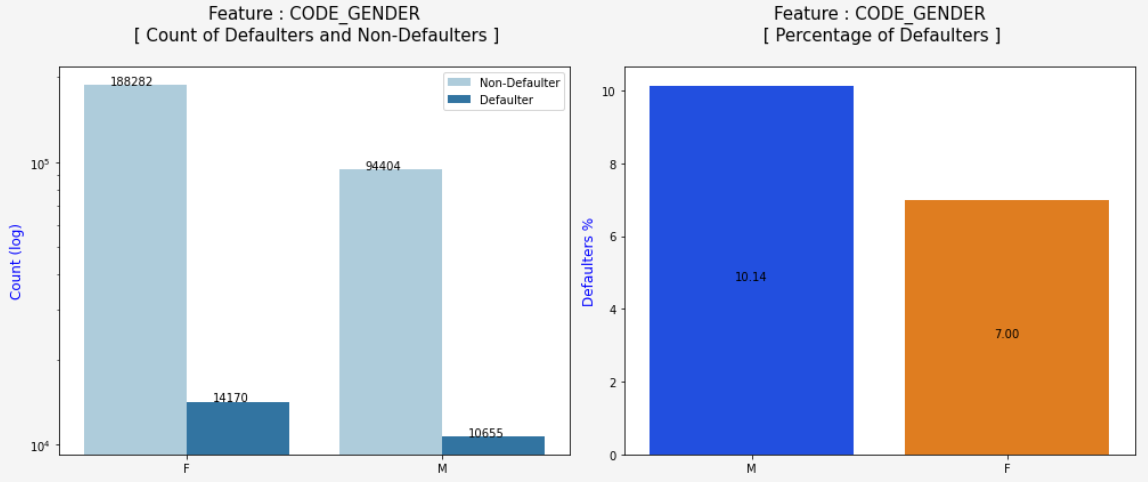
❖ Insights from Contract Type

- ✓ There are more clients for Cash loans than Revolving loans.
- ✓ Defaulter percentage is higher in Cash loans.

❖ Insights from Code Gender

- ✓ There are more Females clients than Males.
- ✓ Defaulter percentage is higher in Males.





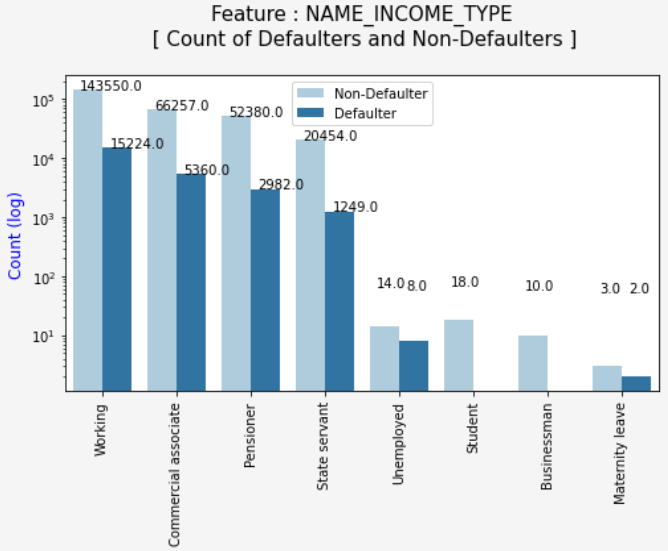
4 Data Analysis: Univariate and Segmented Univariate Analysis

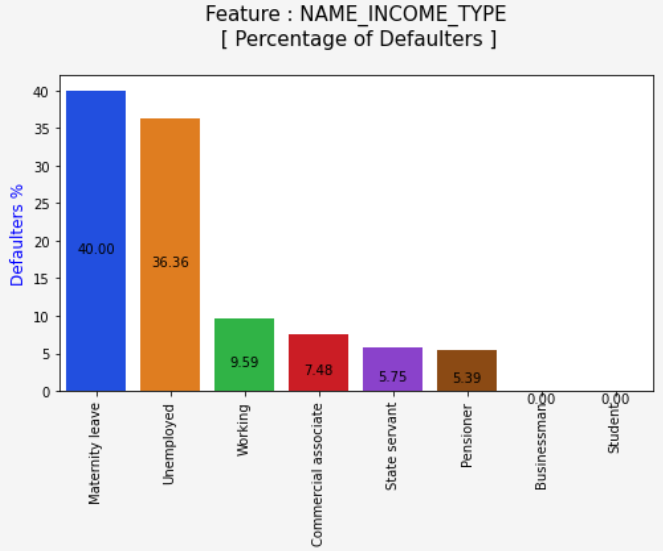
❖ Insights from Income Type

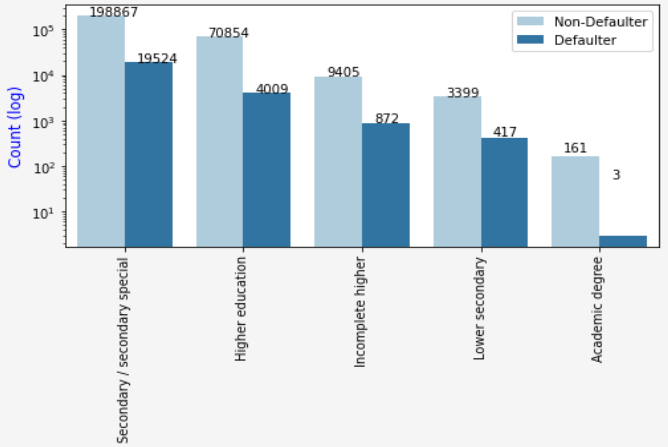
- ✓ Most of the clients have income type as Working, followed by Commercial associate, Pensioner and State servant.
- ✓ Clients who are Unemployed and maternity leave have high default rate.
- ✓ Student and Businessmen, though less in numbers do not have any default record.

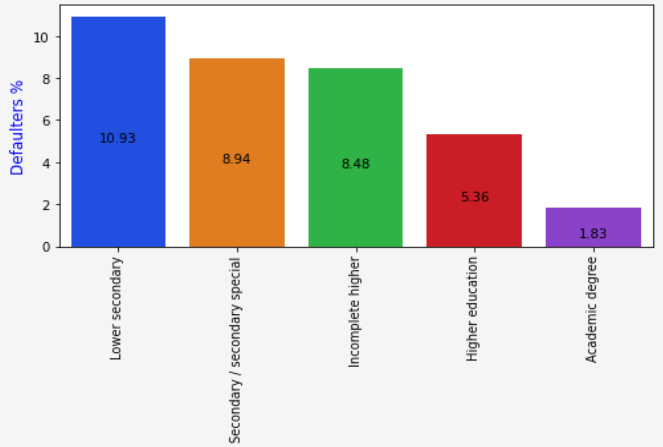
❖ Insights from Education Type

- ✓ Most of the clients have Sec./Sec. special education, followed by Higher education.
- ✓ Clients with education type as Lower secondary category have the highest default rate.
- ✓ Clients with Academic degree have least defaulting rate.









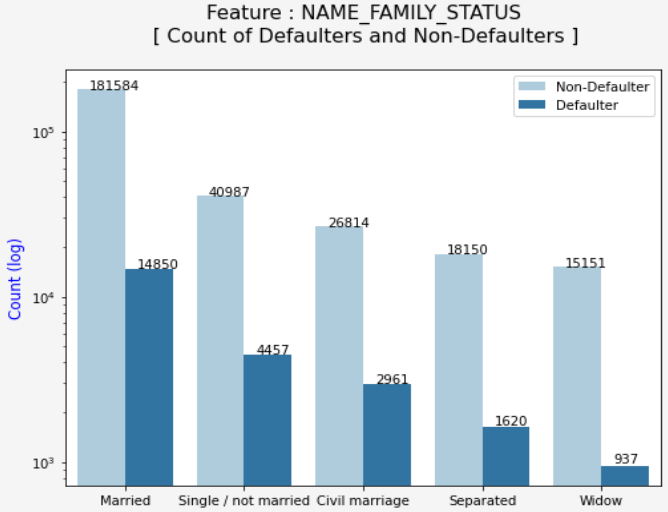
4 Data Analysis: Univariate and Segmented Univariate Analysis

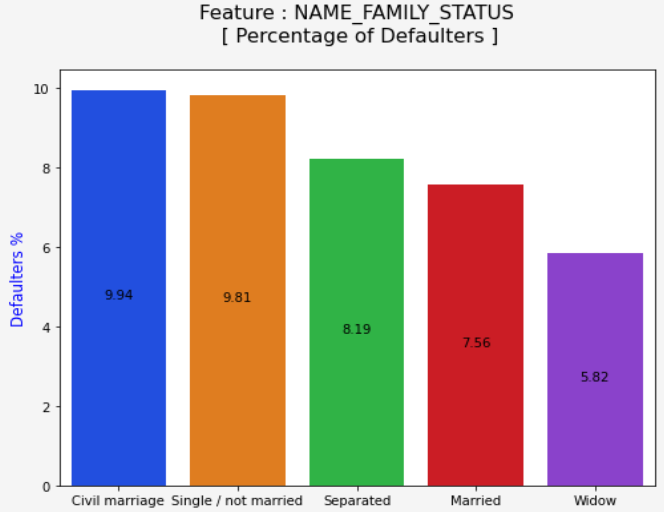
❖ Insights from Family Status

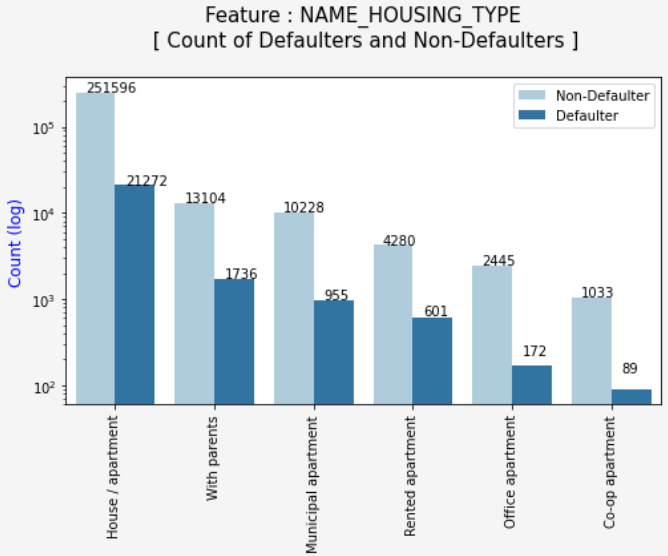
- ✓ Most of the clients are married, followed by Single/not married and civil marriage.
- ✓ Single & civil marriage clients have high default rate.

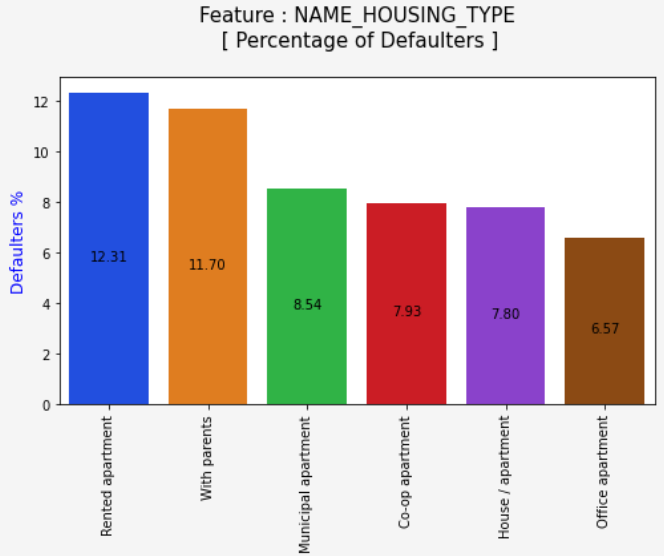
❖ Insights from Housing Type

- ✓ Most of the clients live in House/ Apartment.
- ✓ Clients living with parent & living in rented apartments have high default rate.
- ✓ Clients living in office apartments have lowest default rate.





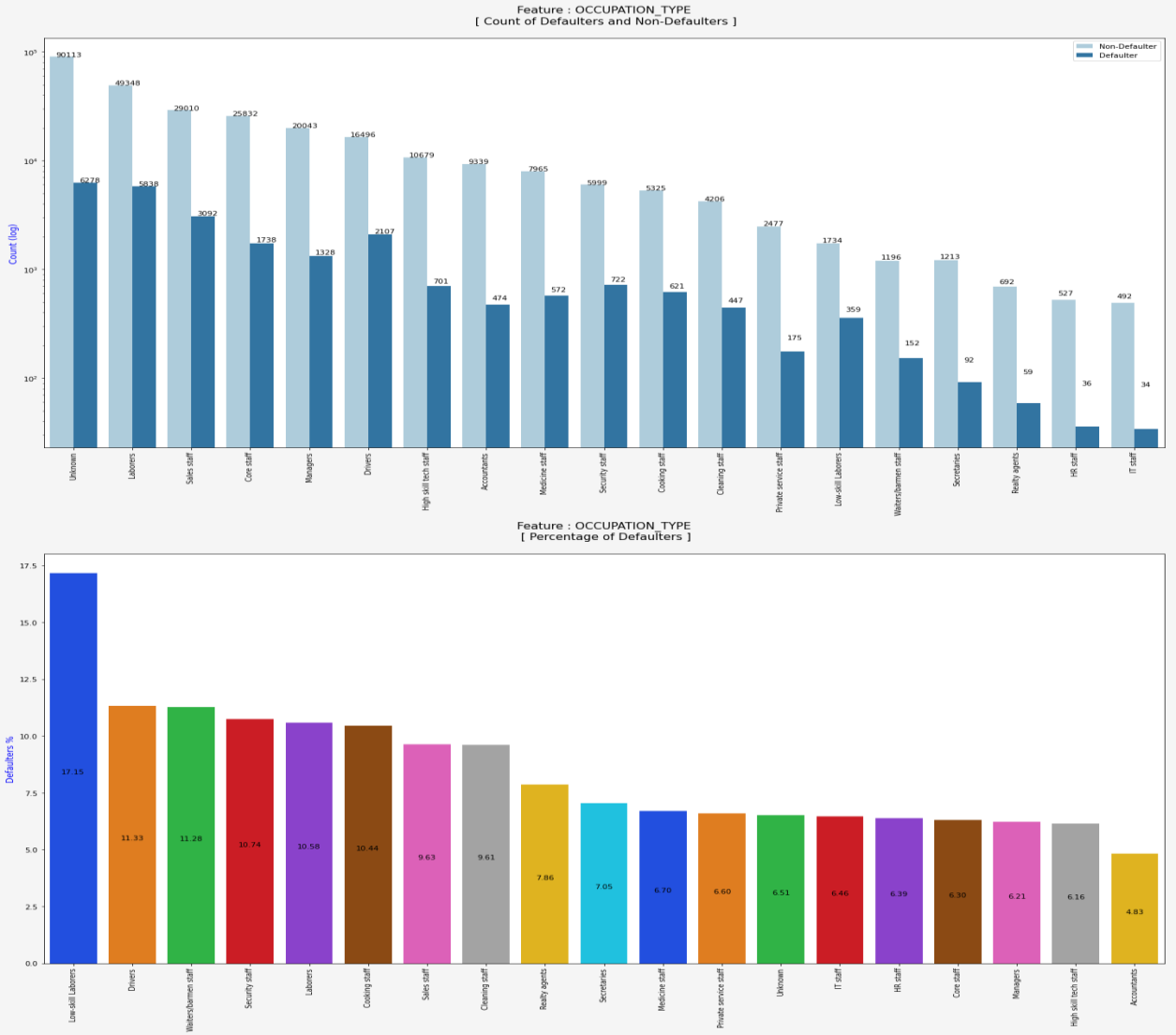




4 Data Analysis: Univariate and Segmented Univariate Analysis

❖ Insights from Occupation Type

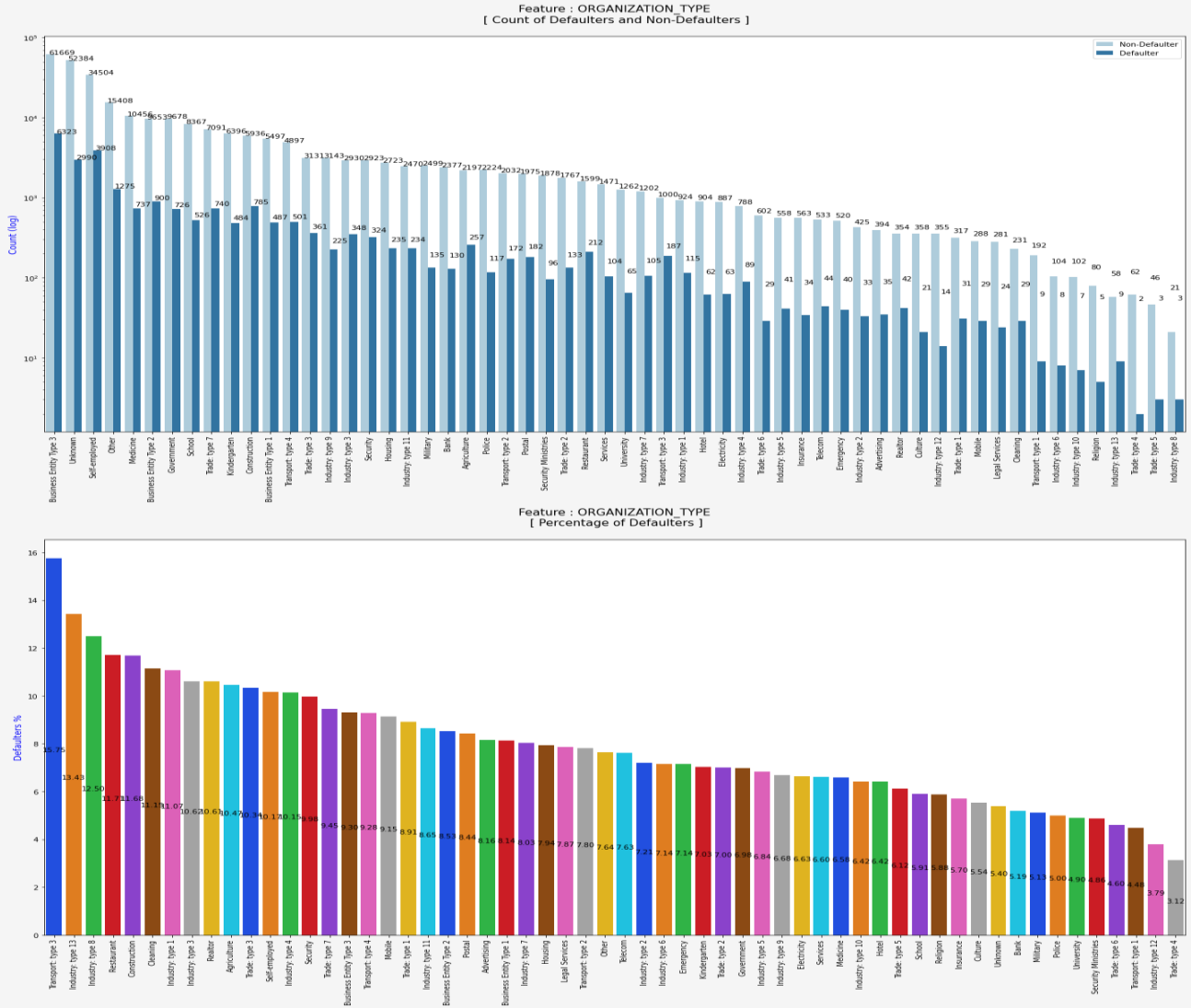
- ✓ Most of the clients are Labour class followed by Sales staff
- ✓ Very few clients have occupation type IT staff.
- ✓ Low-skill labourers have the highest default rate.



4 Data Analysis: Univariate and Segmented Univariate Analysis

❖ Insights from Organization Type

- ✓ Most of the clients are from Business Entity Type 3.
- ✓ For a very high number of applications, Organization type information is unknown(XNA).
- ✓ Transport: type 3, Industry: type 13 , Industry: type 8 , and Restaurant: type 3 are the organizations with the highest percentage of defaulters.
- ✓ Trade Type 4 and 5, Industry type 8 have the least default rate.



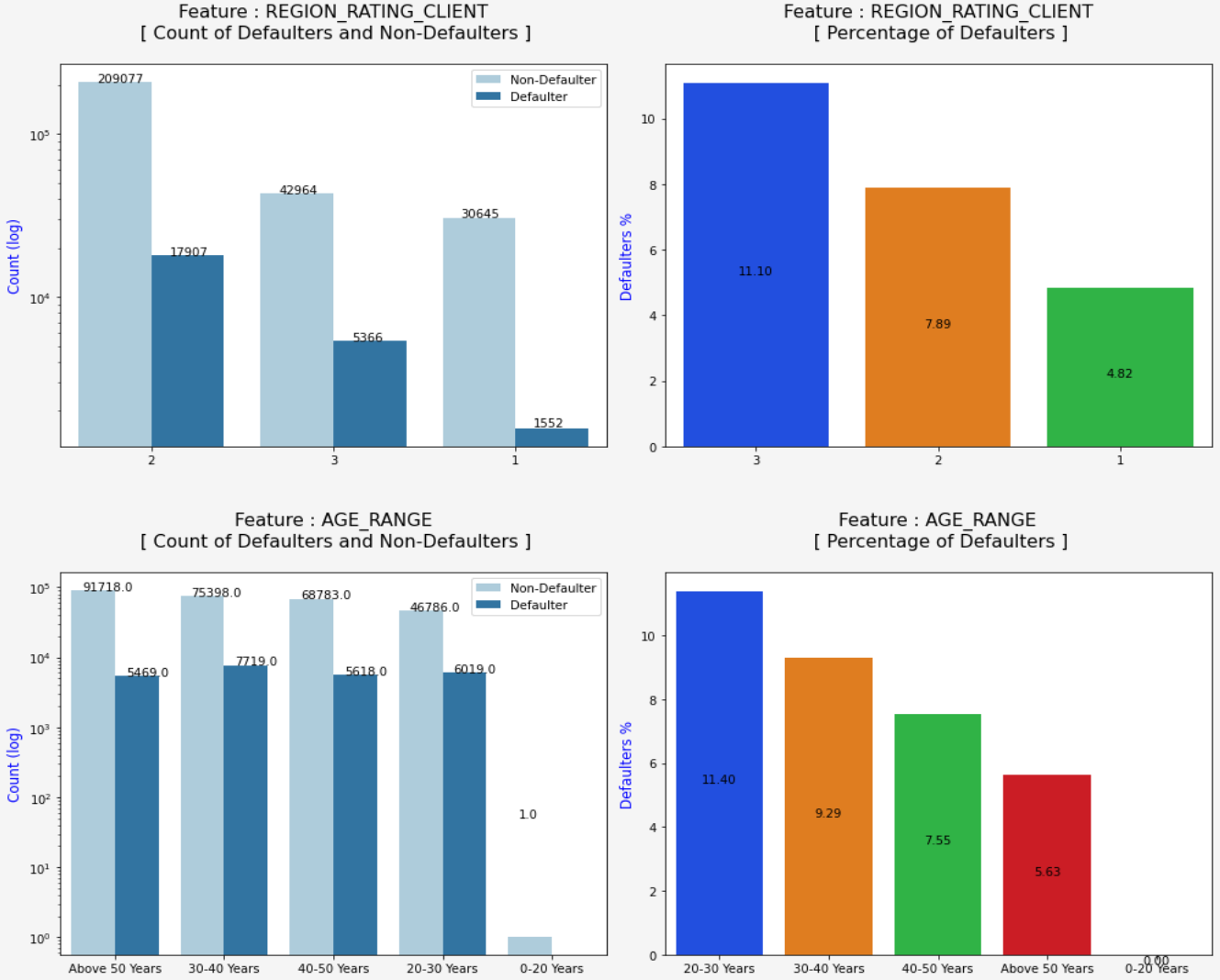
4 Data Analysis: Univariate and Segmented Univariate Analysis

❖ Insights from Region Rating Client

- ✓ Most of the clients are living in the region with Region Rating 2.
- ✓ Clients living in the region with Region Rating 3 have the highest default rate
- ✓ Clients living in the region with Region Rating 1 have the lowest default rate.

❖ Insights from Days Birth

- ✓ Clients in the age group range 20-40 have higher probability of defaulting.
- ✓ Clients above age of 50 have low probability of defaulting.



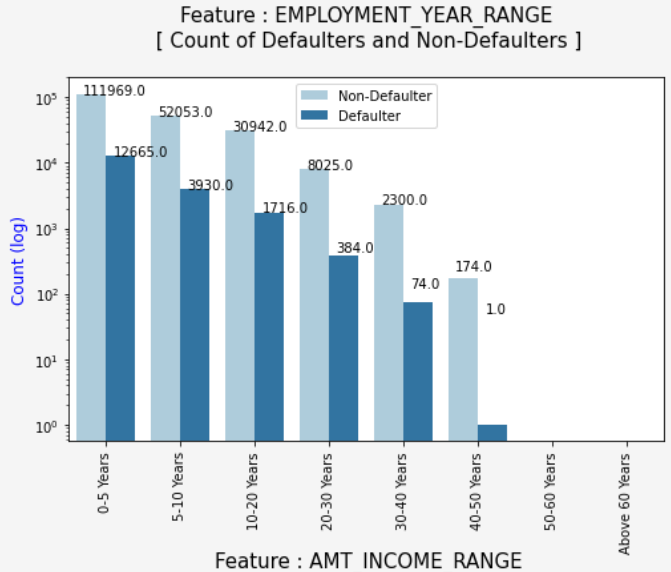
4 Data Analysis: Univariate and Segmented Univariate Analysis

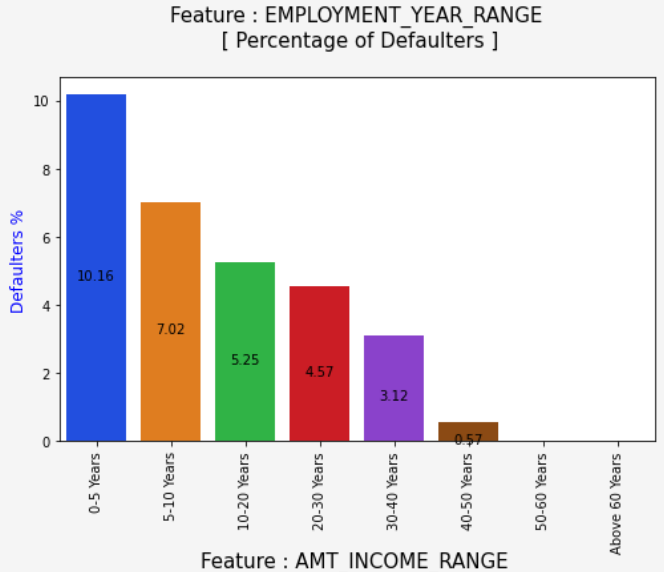
❖ Insights from Days Employment

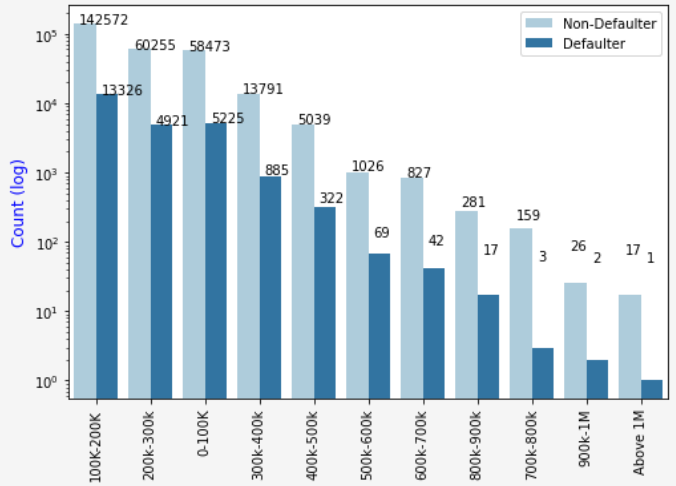
- ✓ Majority of the clients have been employed in between 0-5 years. The defaulting rate of this group is also the highest.
- ✓ With increase of employment year, defaulting rate is gradually decreasing with clients having 40+ year experience having less than 1% default rate.

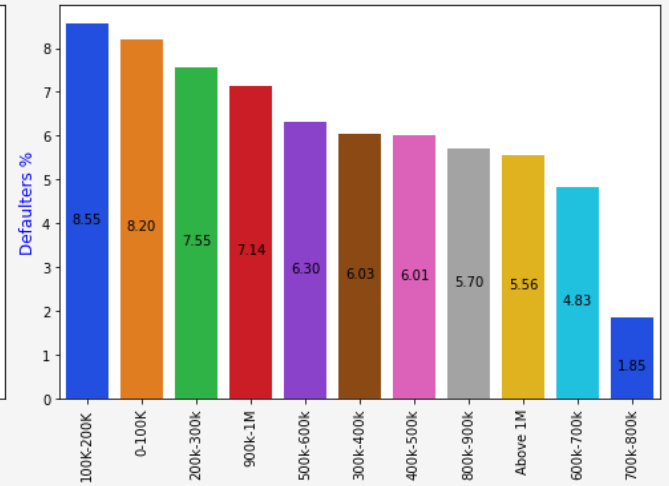
❖ Insights from Amount Income

- ✓ 90% of the applications have Income total less than 3,00,000.
- ✓ Clients with Income less than 3,00,000 have high probability of defaulting.
- ✓ Clients with Income more than 7,00,000 are less likely to default.









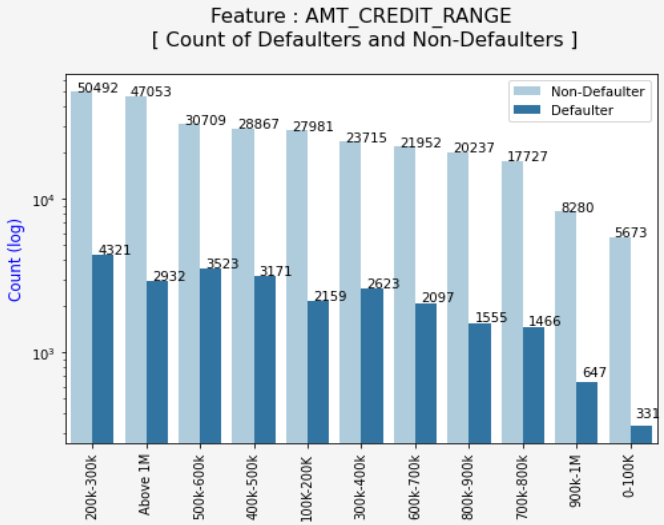
4 Data Analysis: Univariate and Segmented Univariate Analysis

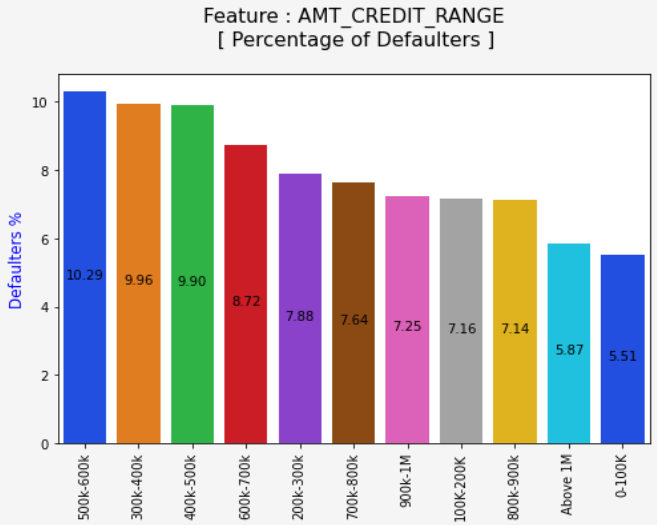
❖ Insights from Amount Credit

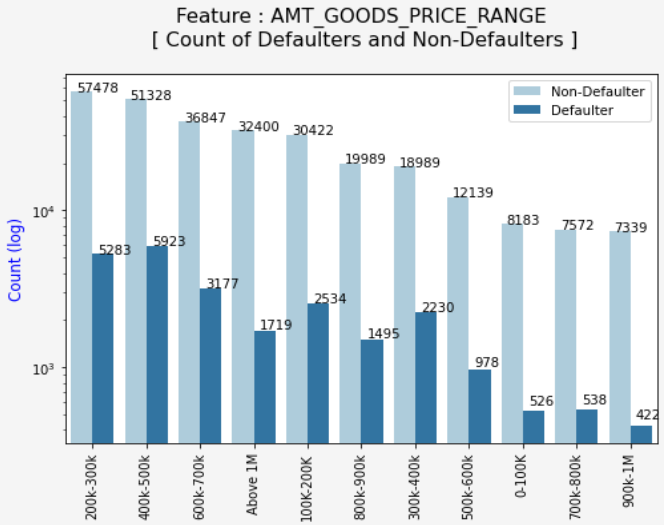
- ✓ More than 80% of the loan provided are for amount less than 9,00,000.
- ✓ People who get loan for 300-600k tend to default more than others.

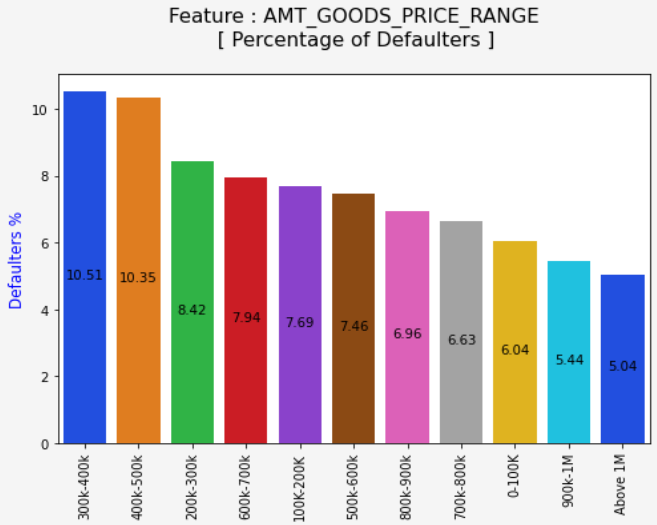
❖ Insights from Amount Goods Price

- ✓ Around 20% of the clients have taken loan for goods in price range 200K-300K.
- ✓ Defaulter percentage of clients taking loan for goods in price range 300k-500k is highest.









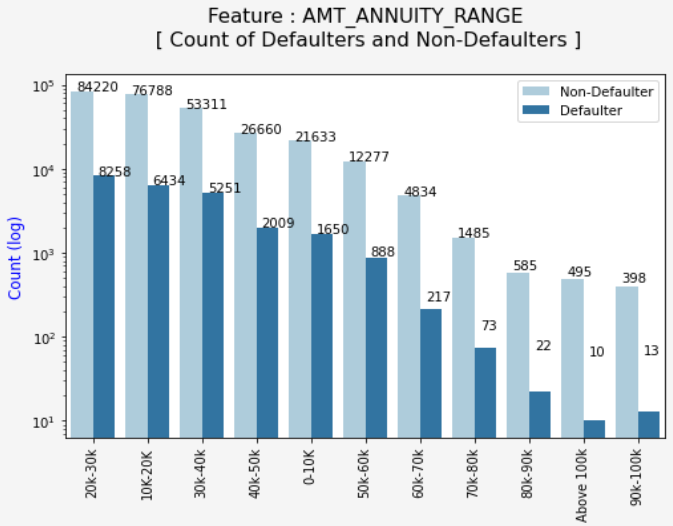
4 Data Analysis: Univariate and Segmented Univariate Analysis

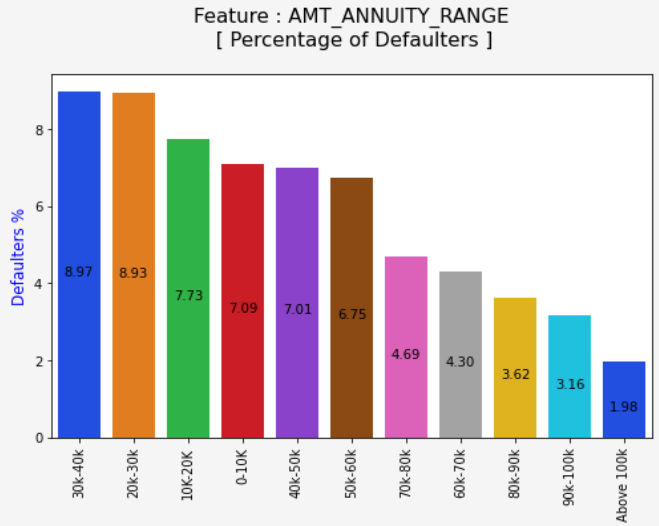
❖ Insights from Amount Annuity

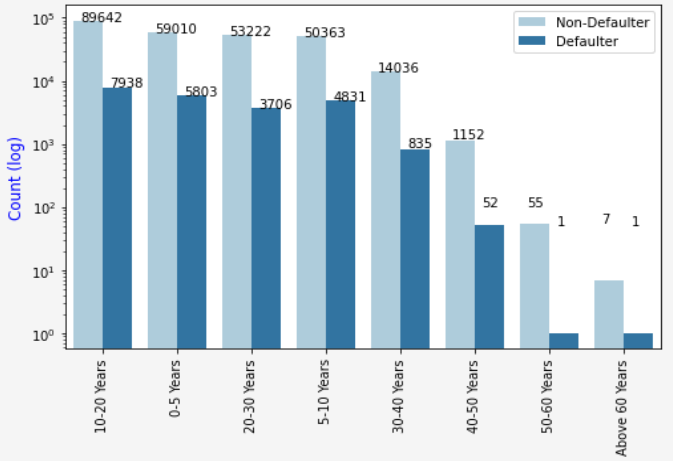
- ✓ More than 70% of the loan provided have loan annuity in the range 10-40k.
- ✓ Defaulter percentage is higher for the clients with loan annuity in the same range.

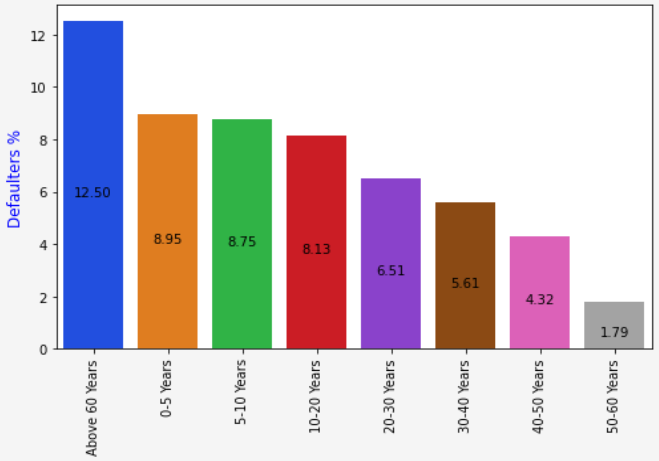
❖ Insights from Days Registration

- ✓ Around 33% of the clients have changed their registration 10-20 years before the application.
- ✓ Defaulter percentage is higher for clients who have changed their registration more than 50 years before the application.









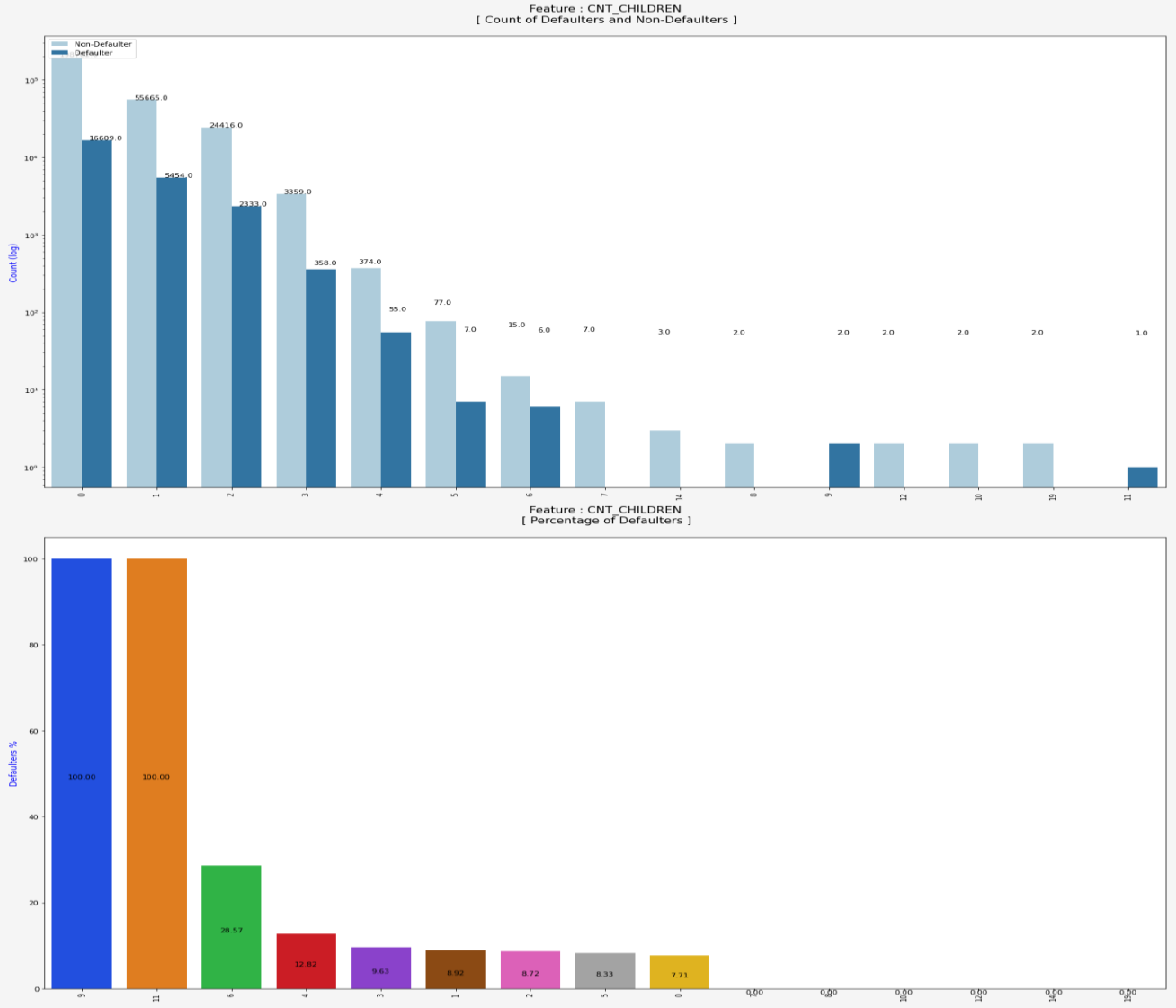
4 Data Analysis: Univariate and Segmented Univariate Analysis

❖ Insights from Count Children

- ✓ Most of the clients do not have children.
- ✓ Client who have more than 4 children has a very high default rate with child count 9 and 11 showing 100% default rate.

❖ Insights from Count Family Members

- ✓ Most of the clients have family members in the range of 1 to 4.
- ✓ Defaulter percentage is positively correlated to the count of family members with 100 % default rate of clients having family members 11 and 13.



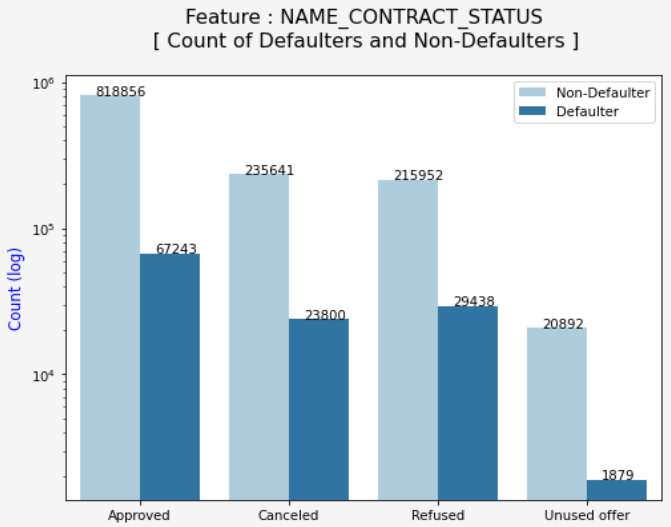
4 Data Analysis: Univariate and Segmented Univariate Analysis

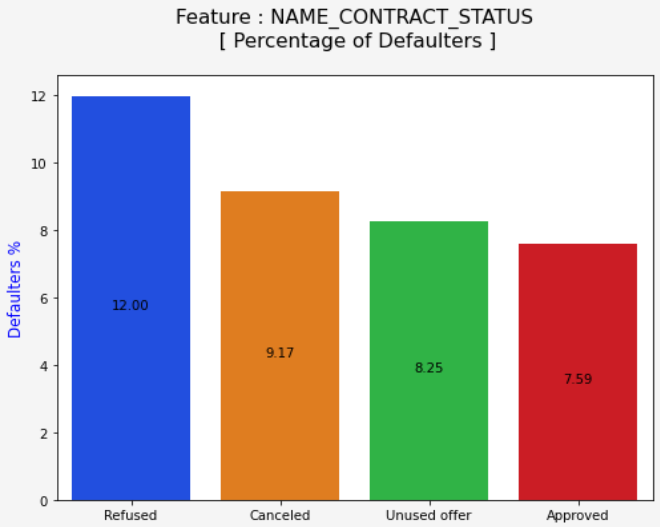
❖ Insights from Name Contract Status

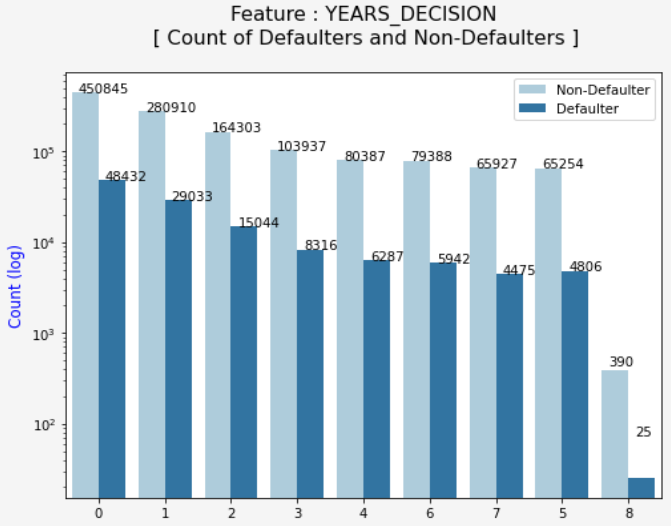
- ✓ 90% of the previously cancelled client have repaid the loan.
- ✓ 88% of the clients who have been previously refused a loan have paid back the loan in current case.
- ✓ 7% of the previously approved loan applicants that defaulted in current loan.

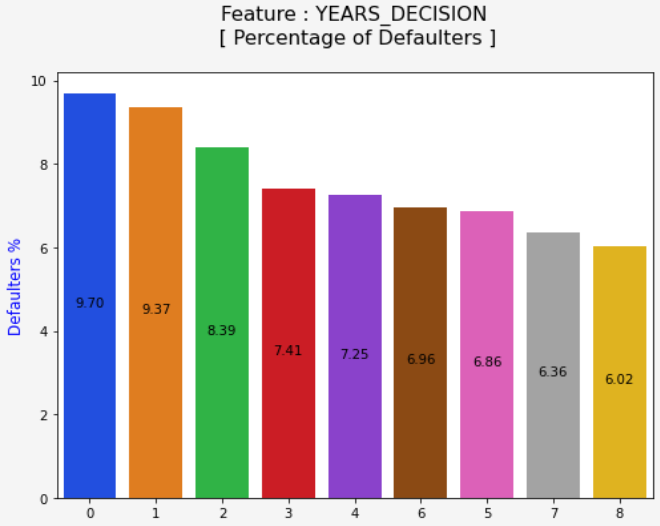
❖ Insights from Days Decision

- ✓ Most of the applicants have applied for a new loan within 1 year of previous loan decision.
- ✓ They also tend to have highest default rate.









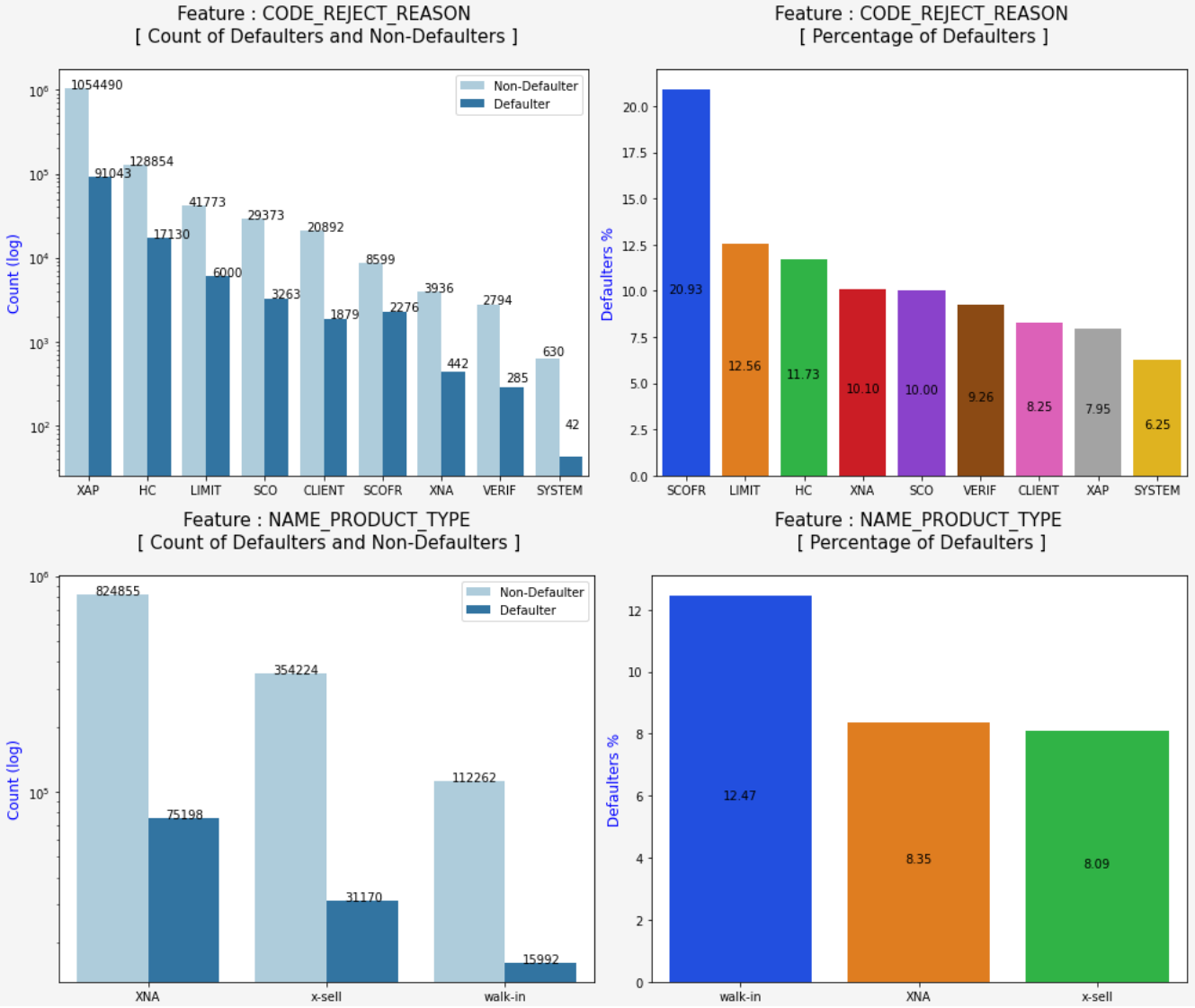
4 Data Analysis: Univariate and Segmented Univariate Analysis

❖ Insights from Code Reject Reason

- ✓ 'SCO', 'LIMIT' and 'HC' are the most common reason of rejection after Unknown Value(XAP).
- ✓ Defaulter percentage is highest for the code 'SCOFR'.

❖Insights from Name Product Type

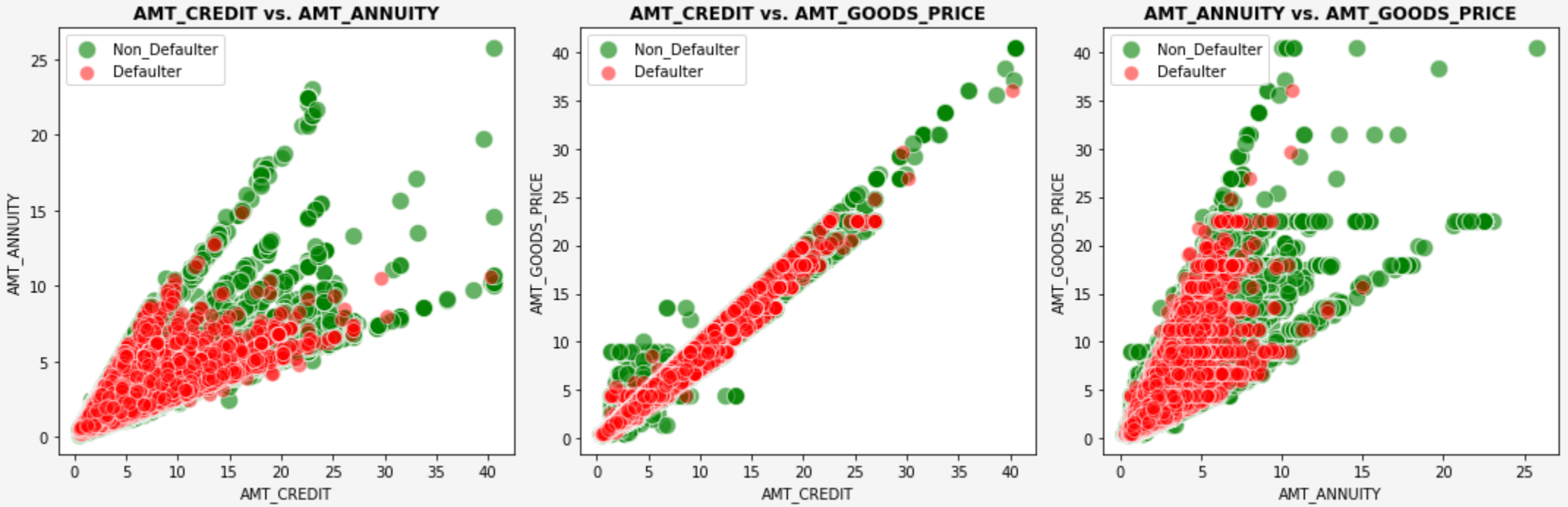
- ✓ It has high number of Unknown Values(XNA).
- ✓ Most of the previous applications were x-sell.
- ✓ Default rate is high where the previous application was walk-in.



4 Data Analysis: Bivariate/ Multivariate Analysis

❖ Insights

- ✓ There are very less defaulters for AMT_CREDIT >3M
- ✓ AMT_CREDIT and AMT_GOODS_PRICE are highly correlated.
- ✓ When AMT_ANNUITY >15000 and AMT_GOODS_PRICE > 3M, there is a lesser chance of defaulters



4 Data Analysis: Bivariate/ Multivariate Analysis

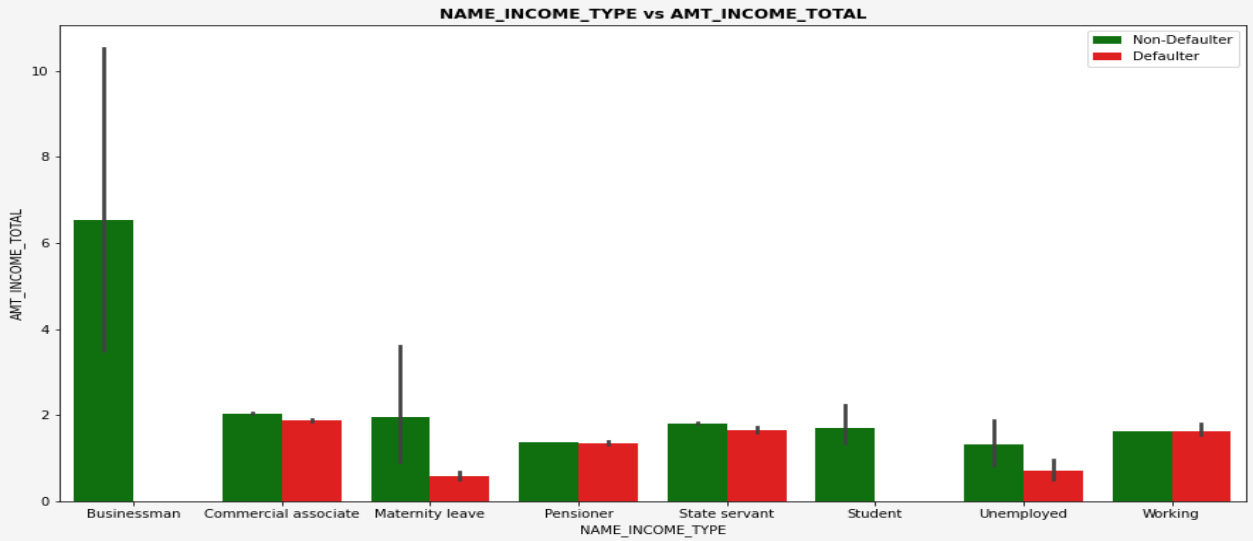
❖ Insights from:

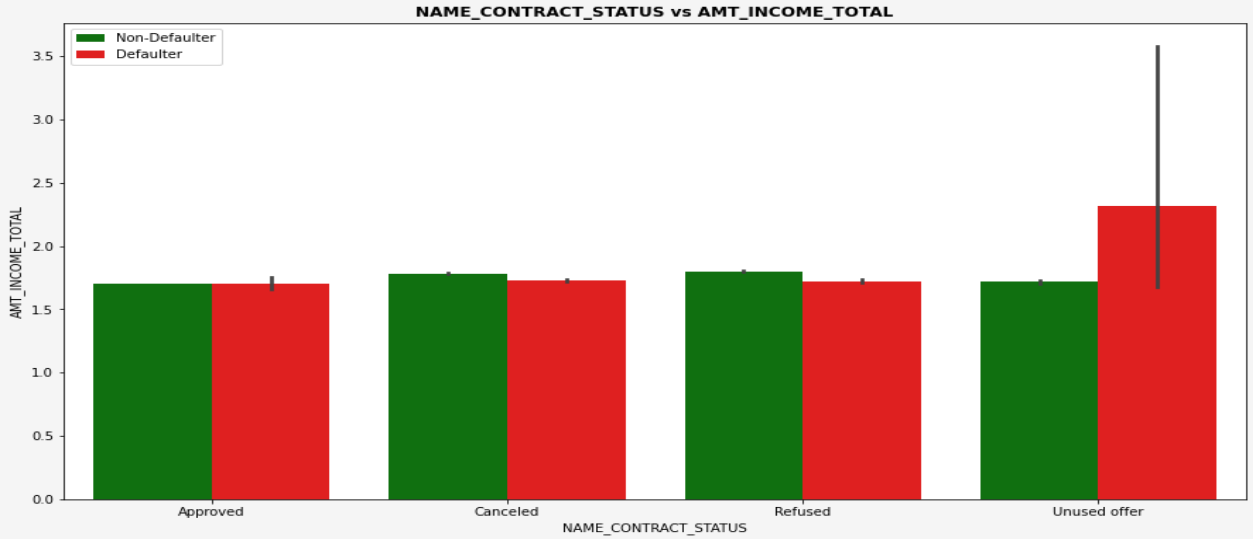
Income Type Name Vs. Amount Income Total

- ✓ Clients with high income, such as businessmen, have a 0% default rate.
- ✓ Clients with low income, less than or equal to 2 lakh, can be either defaulters or non-defaulters.

Name Contract Status Vs. Amount Income Total

- ✓ Clients who have not used offer earlier have defaulted even when there average income is higher than others.



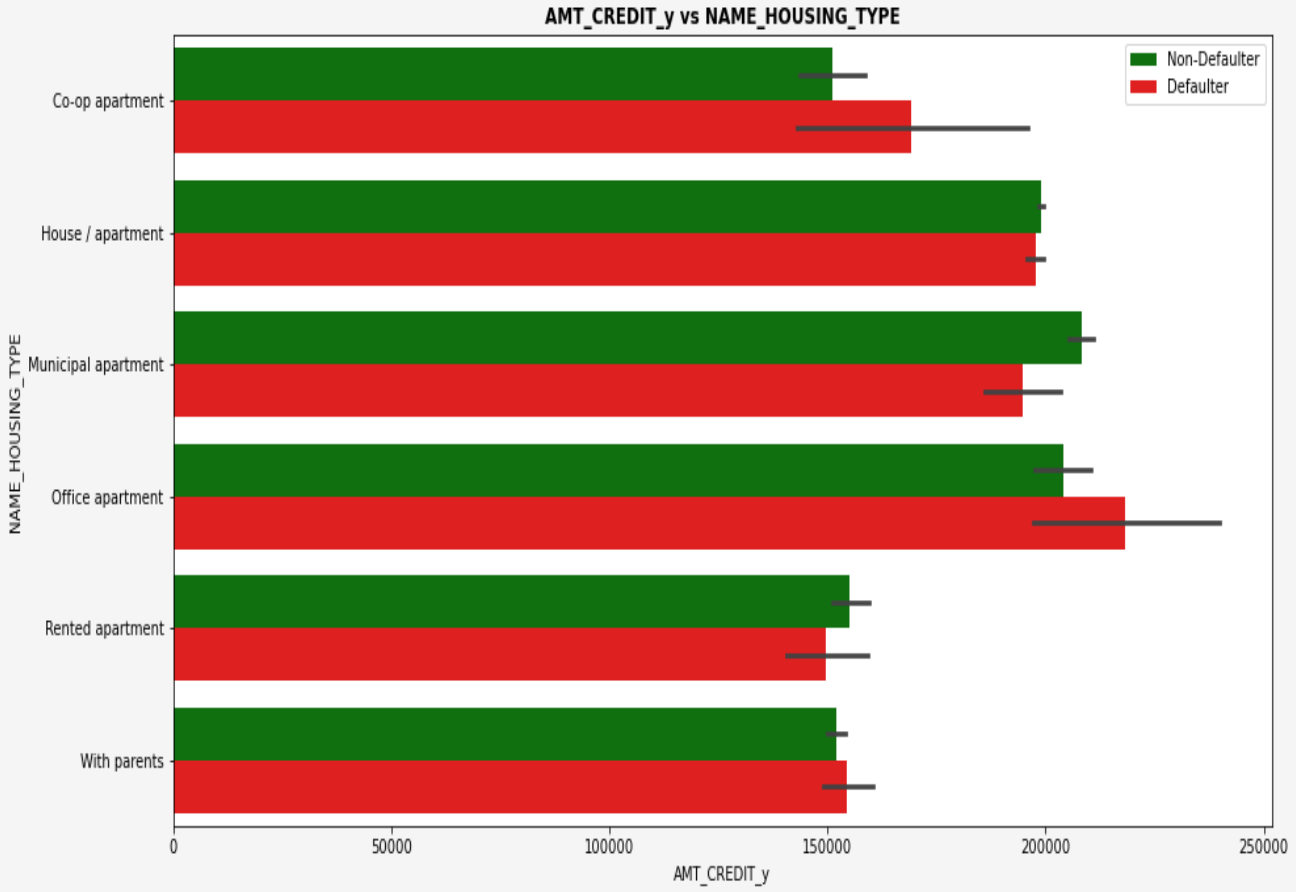


4 Data Analysis: Bivariate/ Multivariate Analysis

❖ Insights from:

Amount Credit_y Vs. Name Housing Type

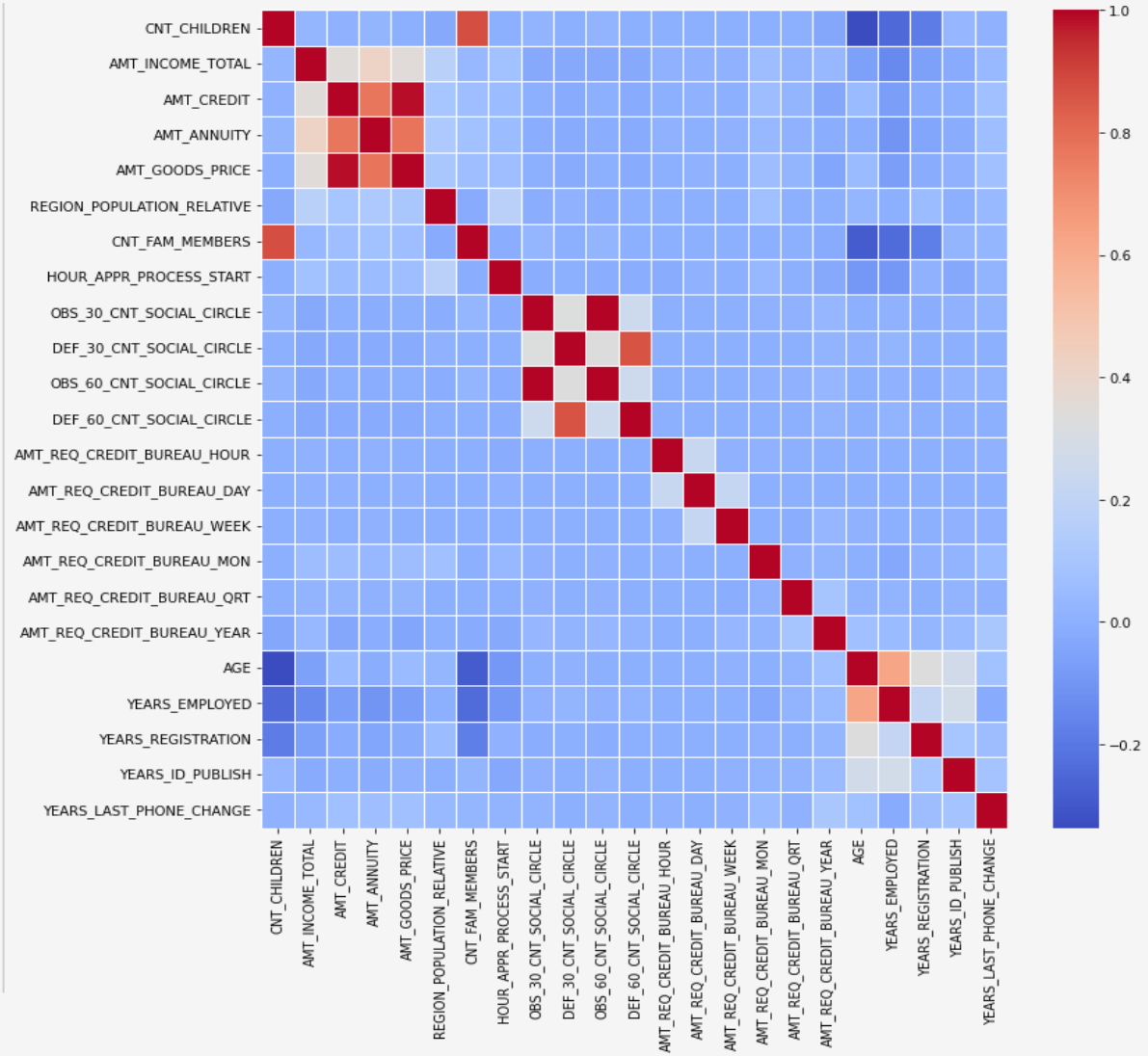
- ✓ "Office apartment" category has the higher credit as compared to others.
- ✓ Clients having housing type Co-op apartment have a very high default rate.
- ✓ Clients living in rented apartment default the least.



4 Data Analysis: Bivariate/ Multivariate Analysis

❖ Top 10 correlation for the Non- Defaulters

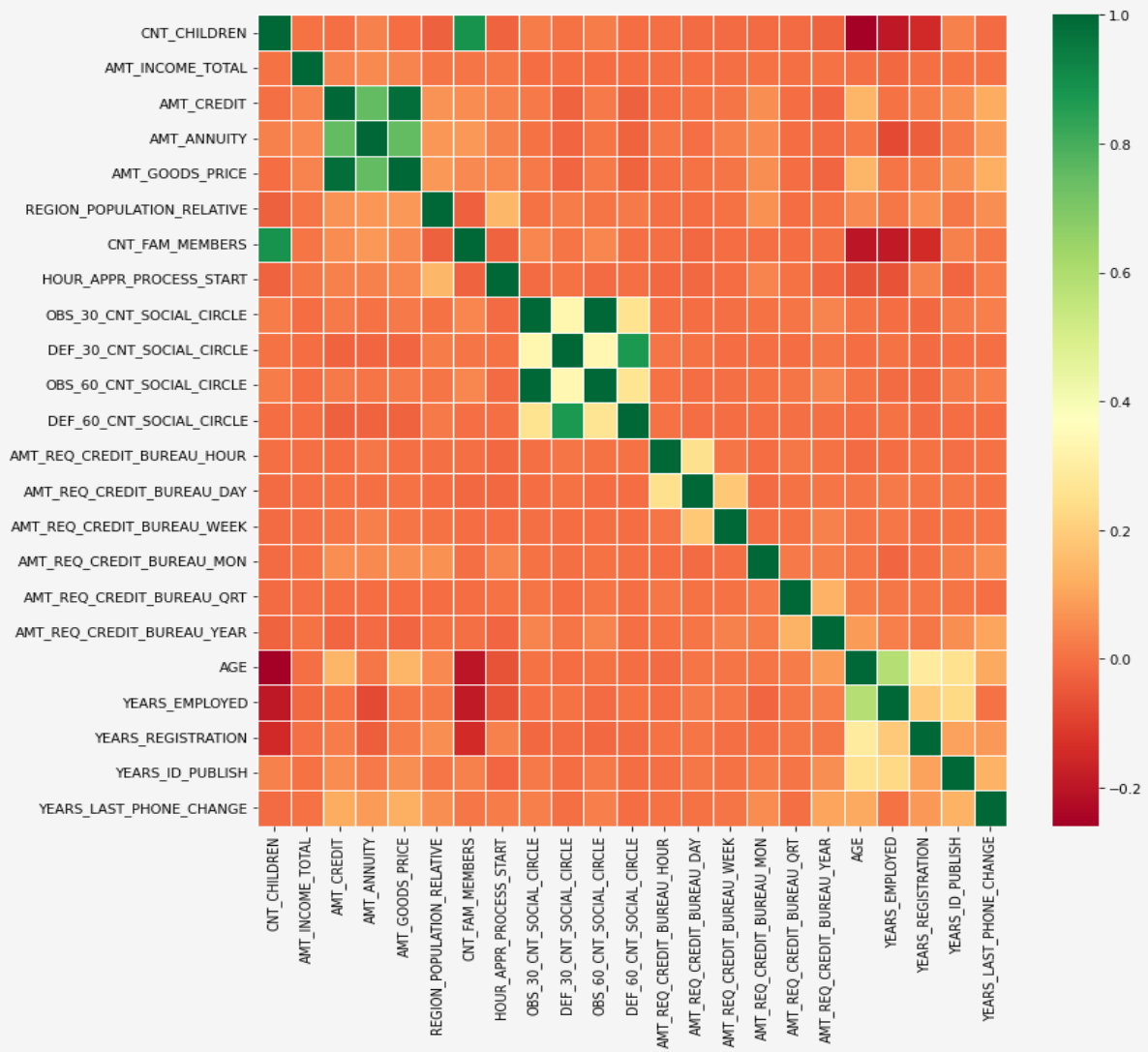
VAR1	VAR2	Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
AMT_GOODS_PRICE	AMT_CREDIT	0.99
CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
AMT_GOODS_PRICE	AMT_ANNUITY	0.78
AMT_ANNUITY	AMT_CREDIT	0.77
YEARS_EMPLOYED	AGE	0.63
AMT_ANNUITY	AMT_INCOME_TOTAL	0.42
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.35
AMT_CREDIT	AMT_INCOME_TOTAL	0.34



4 Data Analysis: Bivariate/ Multivariate Analysis

❖ Top 10 correlation for the Defaulters

VAR1	VAR2	Correlation
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
AMT_GOODS_PRICE	AMT_CREDIT	0.98
CNT_FAM_MEMBERS	CNT_CHILDREN	0.89
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87
AMT_GOODS_PRICE	AMT_ANNUITY	0.75
AMT_ANNUITY	AMT_CREDIT	0.75
YEARS_EMPLOYED	AGE	0.58
OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.34
DEF_30_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.33
YEARS_REGISTRATION	AGE	0.29



Key Insights

ATTRIBUTES	DRIVING FACTORS FOR NON-DEFAULTERS	DRIVING FACTORS FOR DEFAULTERS
NAME_CONTRACT_TYPE:	Client applying for Revolving Loans.	Client applying for Cash Loans.
CODE_GENDER:	Female clients	Male clients
NAME_INCOME_TYPE:	Clients who are student or businessmen	Clients on maternity leave or unemployed
NAME_EDUCATION_TYPE:	Clients having academic degree.	Clients having lower secondary education
REGION_FAMILY_STATUS:	Clients who are either married or widow	Single or Civil Marriage clients
NAME_HOUSING_TYPE:	Clients living in office apartments	Clients living in rented apartments or with parents
OCCUPATION_TYPE:	Clients who are accountants	Clients who are Low-skill labourers
REGION_RATING_CLIENT:	Clients with Rating 1	Clients with Rating 3
ORGANIZATION TYPE:	Trade type-4 and Industry type-12	Transport type-3 and Industry type-13
DAYS_BIRTH:	Clients above 50 years of age	Clients between 20-30 years of age
DAYS_EMPLOYED:	Clients with work-ex of 40-50 years	Clients with work-ex of 0-5 years
AMT_INCOME_TOTAL:	Clients with income range in 700k-800k	Clients with income range in 0-300k
AMT_GOODS_PRICE:	Goods price range above 1M	Goods price range between 300k-500k
AMT_ANNUITY RANGE:	Clients having annuity above 100k	Clients having annuity between 20k-40k
CNT_CHILDREN:	Clients having no children	Clients having more children
CASH_LOAN_PURPOSE:	For buying a garage	For hobby and car repairs
NAME_CLIENT_TYPE:	Refreshed clients	New and repeater clients

Suggestions

- 90% of the clients whose loans were cancelled previously have repaid the loan, hence the company must re-evaluate the reason for cancellation of their loans earlier so that there is an increase in business opportunity.
- 88% of the clients who were refused loan by the company earlier are now a repaying clients, hence company must diligently find the cause of refusal to tackle the business loss.
- The company should provide loans to the high risk clients on a higher interest rate but should be careful about the correlating factors while approving the loans.

Thank You