



# Credit Risk Analysis

**Presentation by-**  
**Amruta Gurav**

# Objective



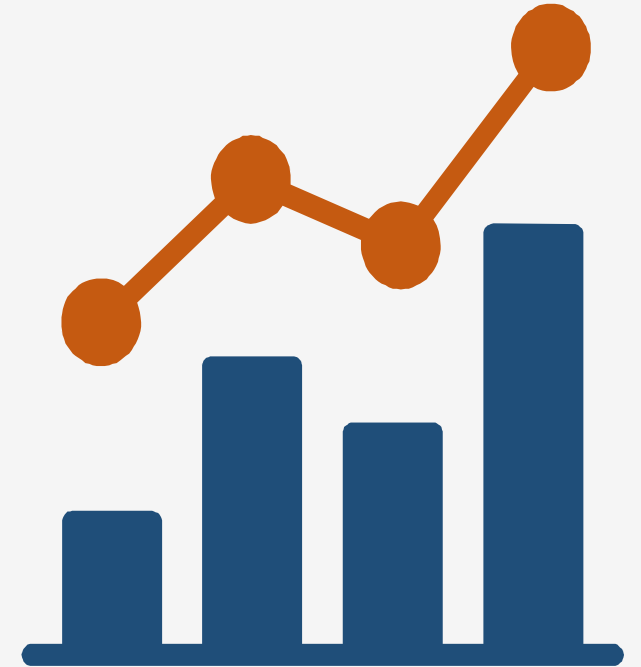
This case study aims to identify patterns which indicate if a client has difficulty paying their instalments, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable 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.

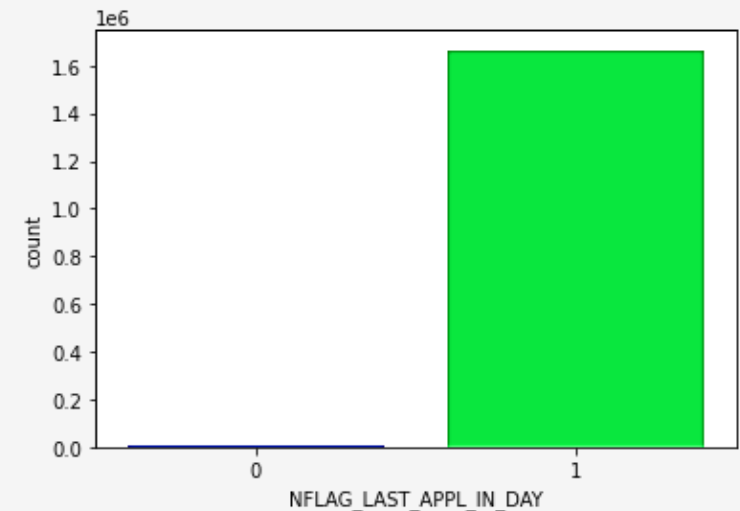
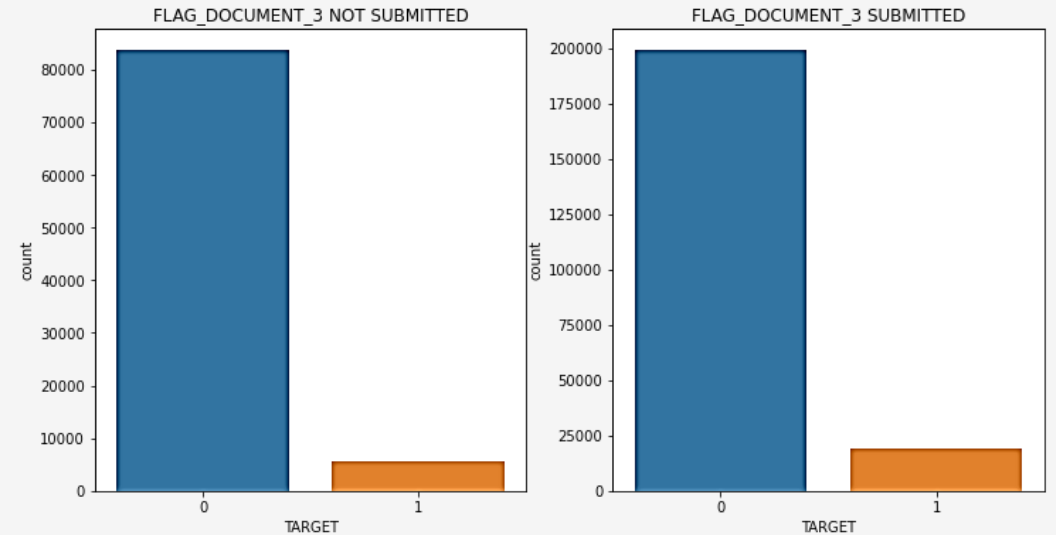
### 3 Data Cleaning: Fixing Rows and Columns

#### ❖ 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

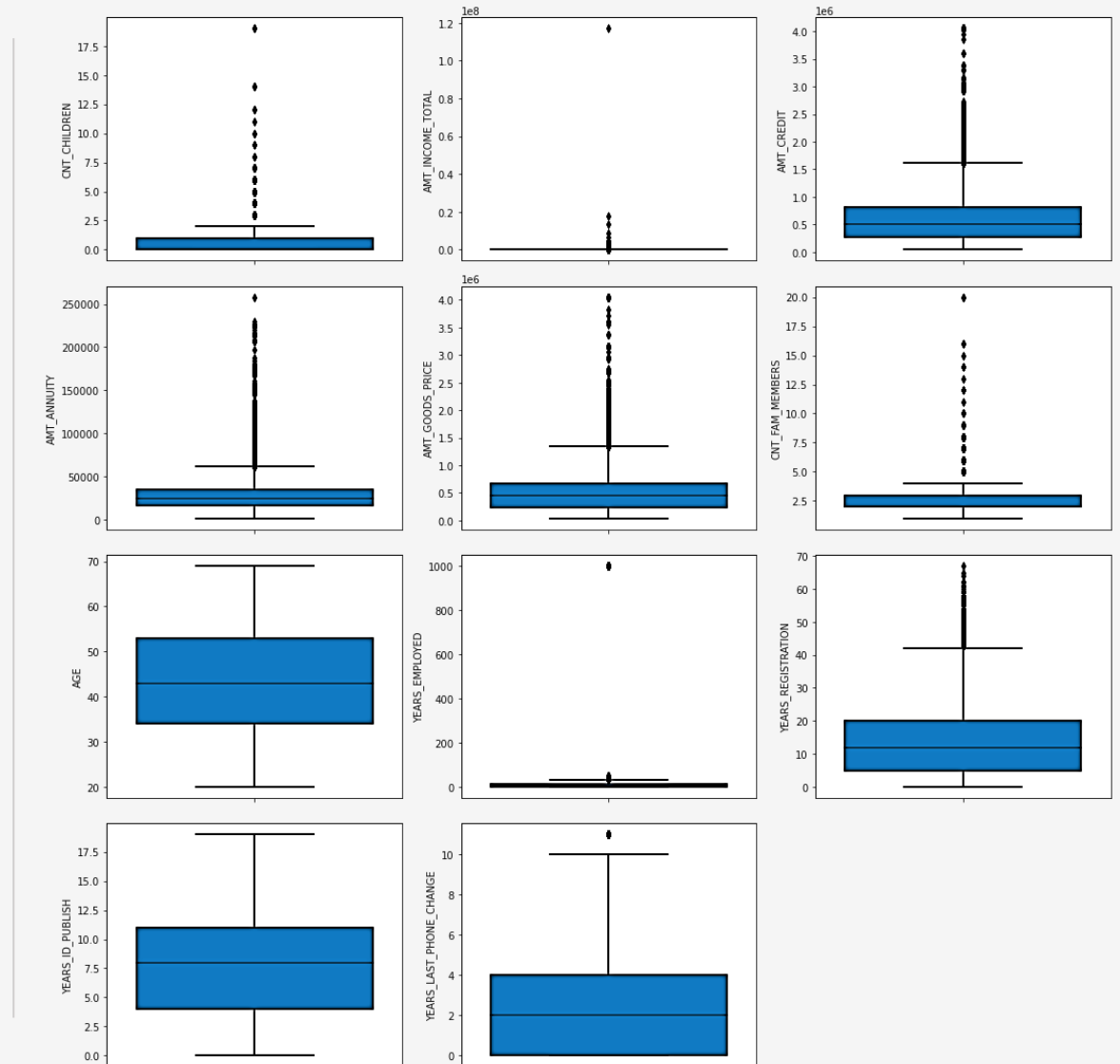
- ✓ Converted values in features 'DAYS\_DECISION' from negative to positive as days cannot be negative.
- ✓ Some features have values XAP, XNA in huge numbers. We will continue our analysis without changing these values.
- ✓ Derived YEARS\_DECISION from DAYS\_DECISION feature from the existing feature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



### 3 Data Cleaning: Identifying and Handling Outliers

#### ❖ For application\_data dataset

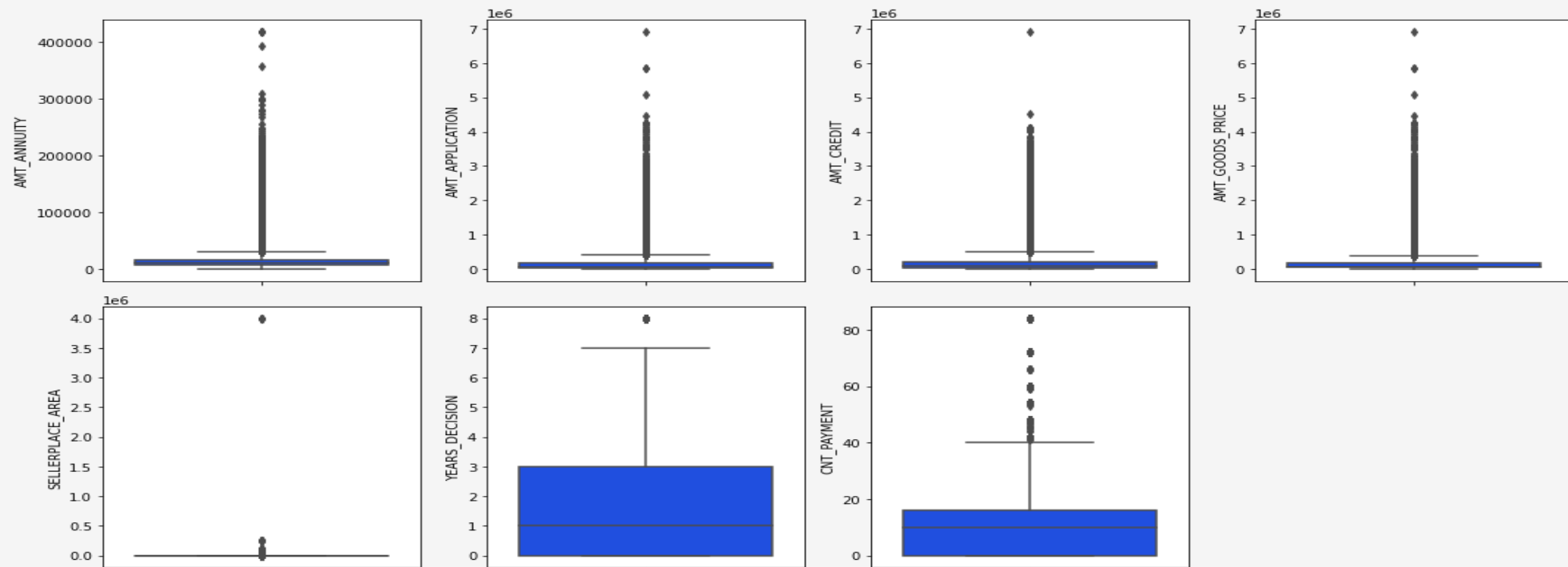
- ✓ CNT\_CHILDREN, CNT\_FAMILY\_MEMBERS features have some number of outliers but the values can't be considered wrong as people can have any number of children or family members.
- ✓ AMT\_INCOME\_TOTAL, AMT\_CREDIT, AMT\_ANNUITY, AMT\_GOODS\_PRICE features have outliers but the values are valid.
- ✓ 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

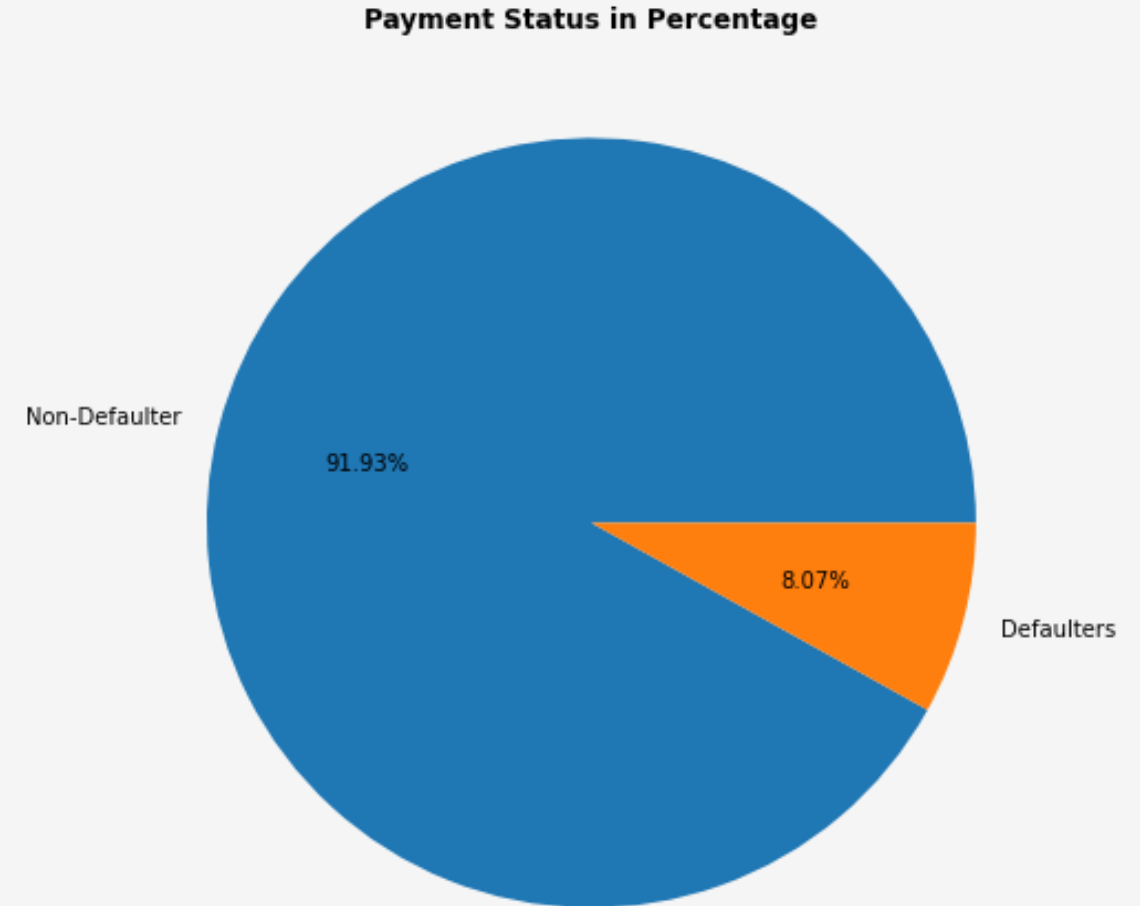
- ✓ AMT\_ANNUIY, AMT\_APPLICATION, AMT\_CREDIT, AMT\_GOODS\_PRICE, SELLERPLACE\_AREA have huge number of outliers but the values are valid.
- ✓ CNT\_PAYMENT has few outlier values but valid values.
- ✓ YEARS\_DECISION has few outliers indicating that these previous applications decisions were taken long 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%.

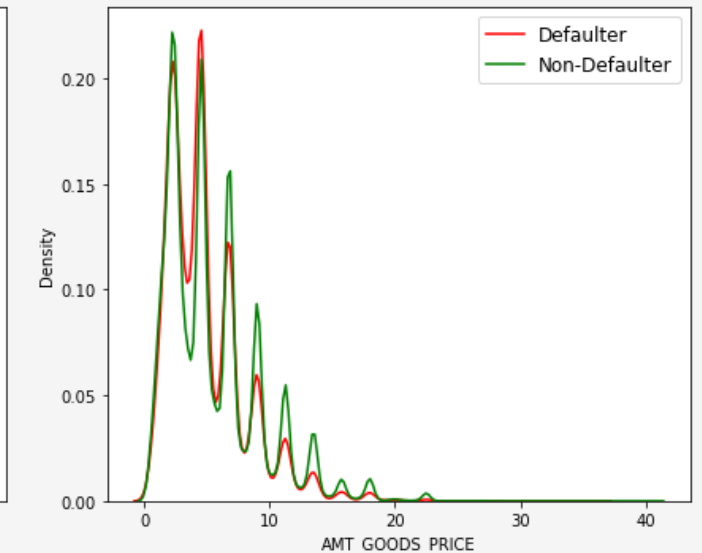
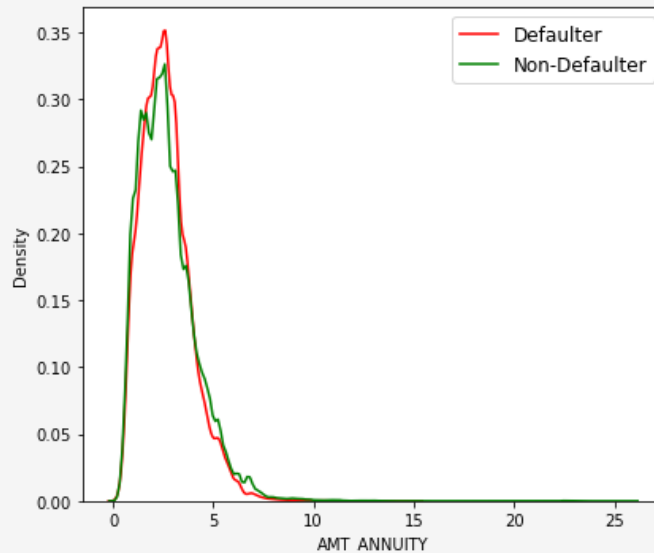
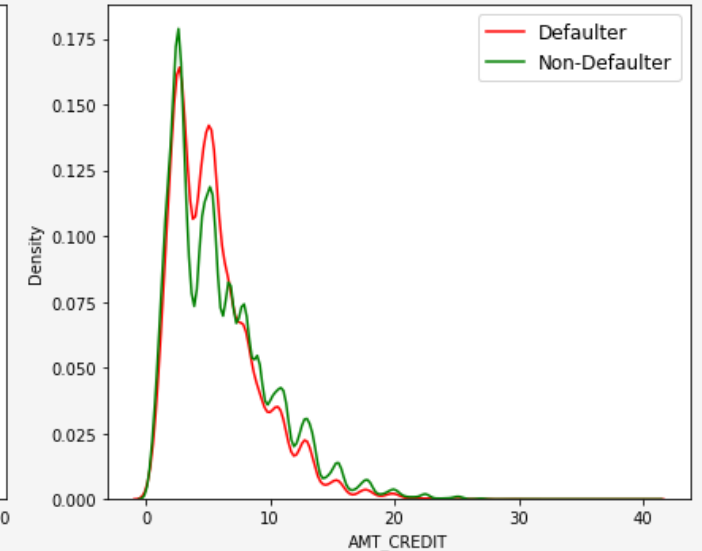
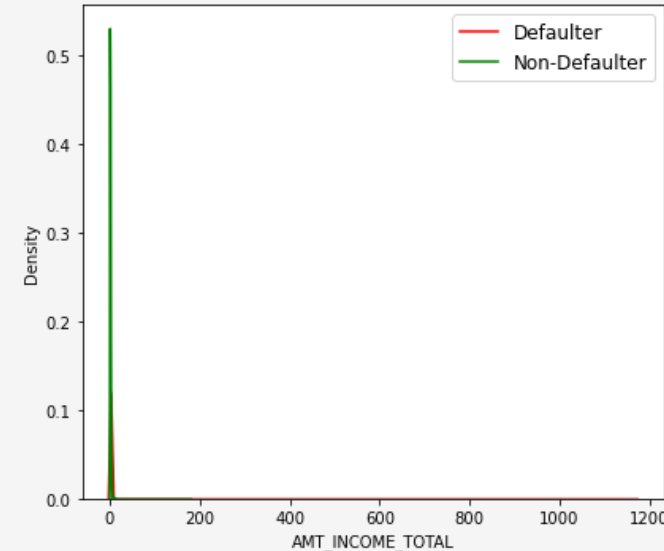


## 4

## 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 than 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.



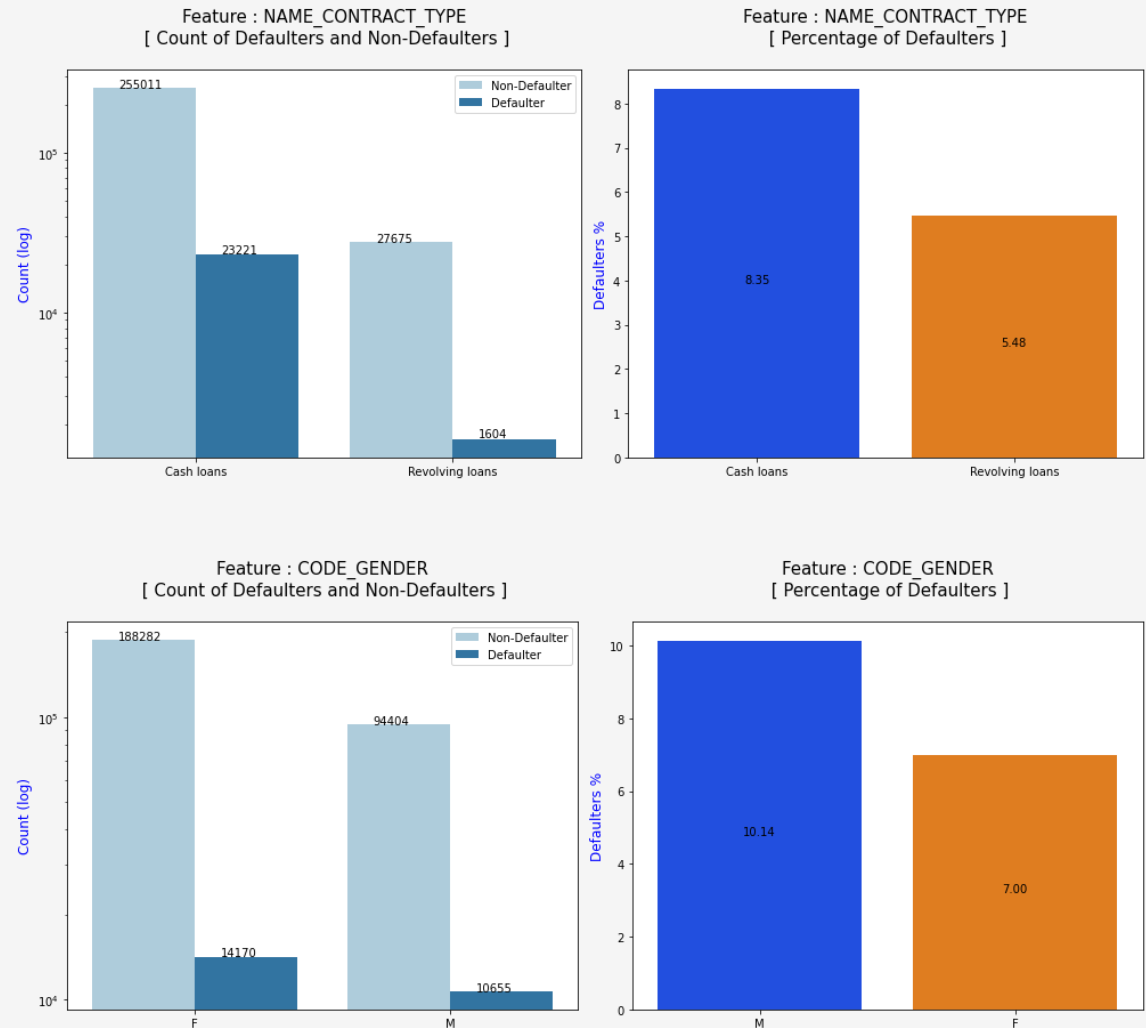
## 4 Data Analysis: Univariate and Segmented Univariate Analysis

### ❖ 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

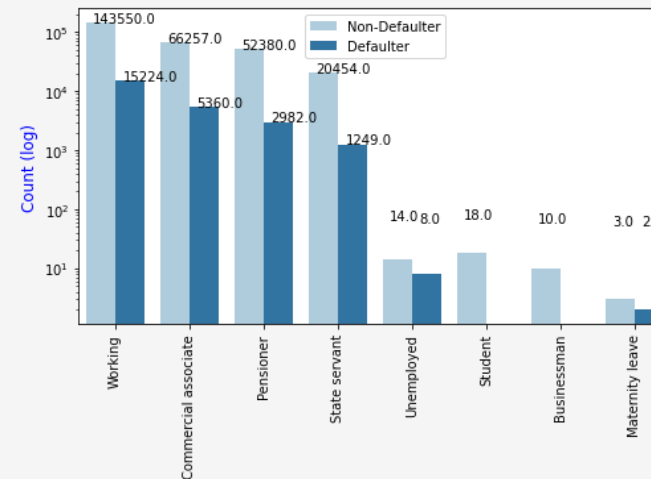
### ❖ Insights from Income Type

- ✓ 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.

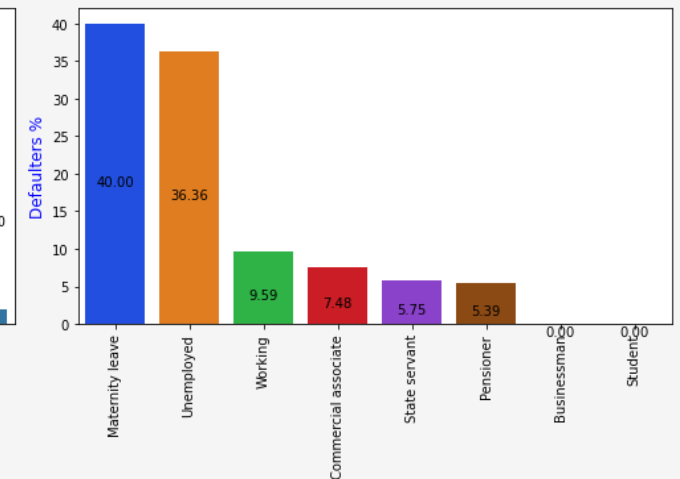
### ❖ Insights from Education Type

- ✓ 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.

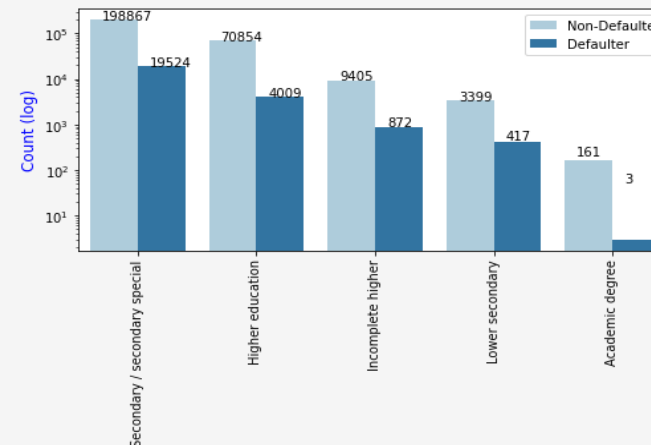
Feature : NAME\_INCOME\_TYPE  
[ Count of Defaulters and Non-Defaulters ]



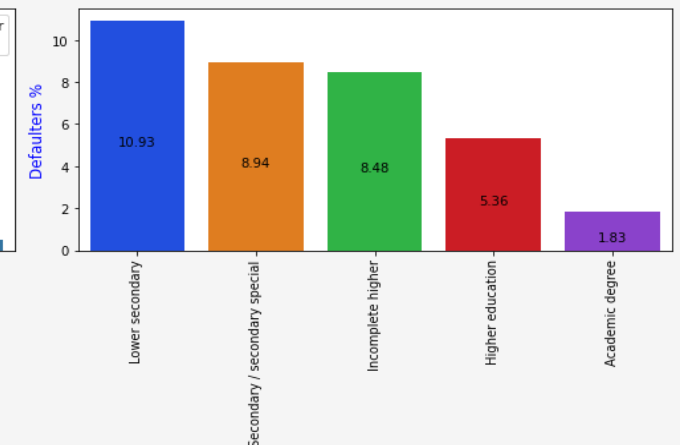
Feature : NAME\_INCOME\_TYPE  
[ Percentage of Defaulters ]



Feature : NAME\_EDUCATION\_TYPE  
[ Count of Defaulters and Non-Defaulters ]



Feature : NAME\_EDUCATION\_TYPE  
[ Percentage of Defaulters ]



## 4 Data Analysis: Univariate and Segmented Univariate Analysis

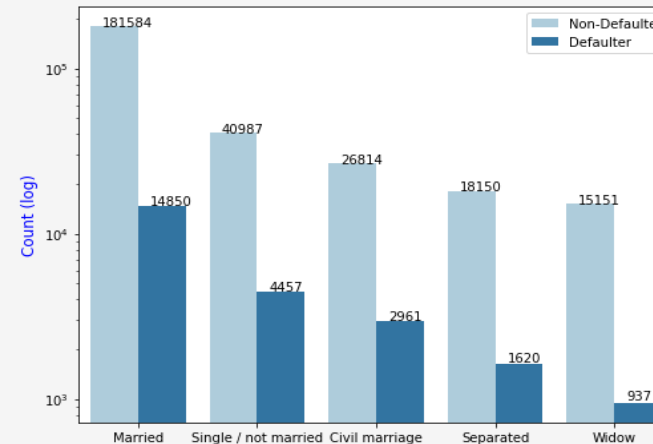
### ❖ Insights from Family Status

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

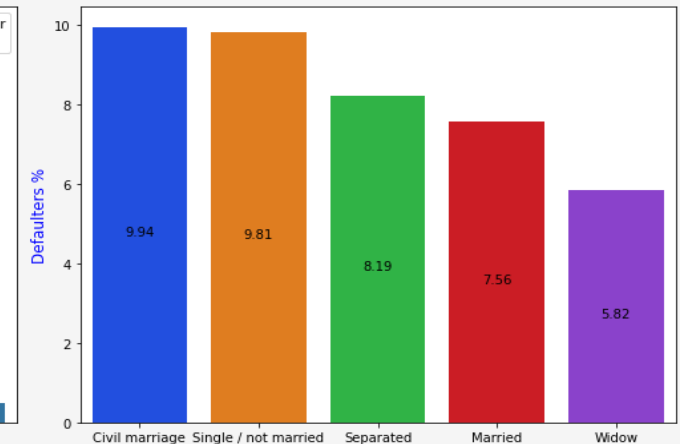
### ❖ Insights from Housing Type

- ✓ 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.

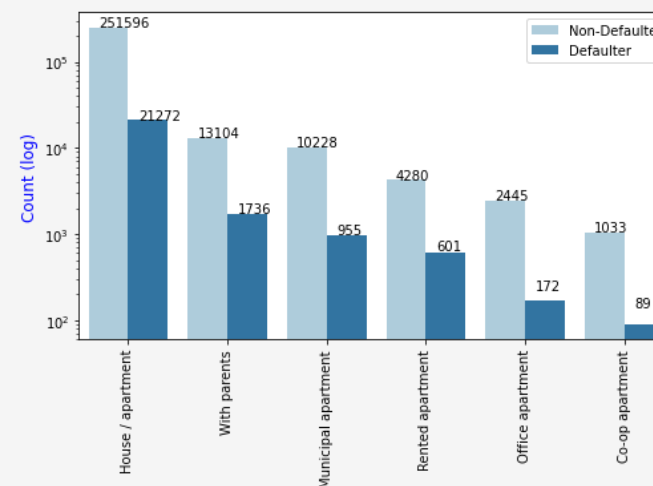
Feature : NAME\_FAMILY\_STATUS  
[ Count of Defaulters and Non-Defaulters ]



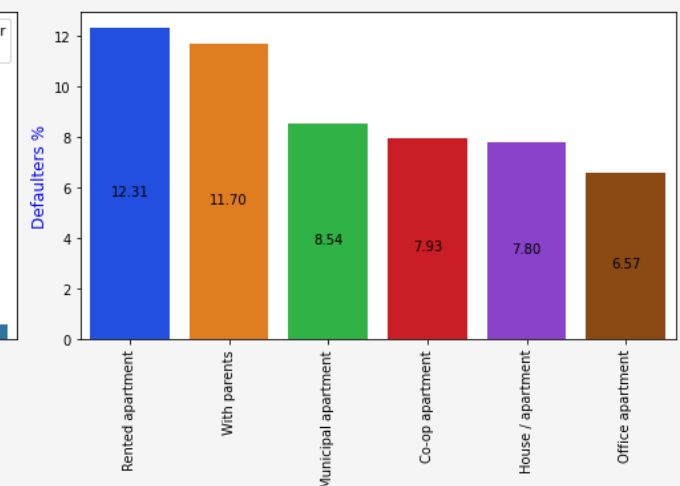
Feature : NAME\_FAMILY\_STATUS  
[ Percentage of Defaulters ]



Feature : NAME\_HOUSING\_TYPE  
[ Count of Defaulters and Non-Defaulters ]



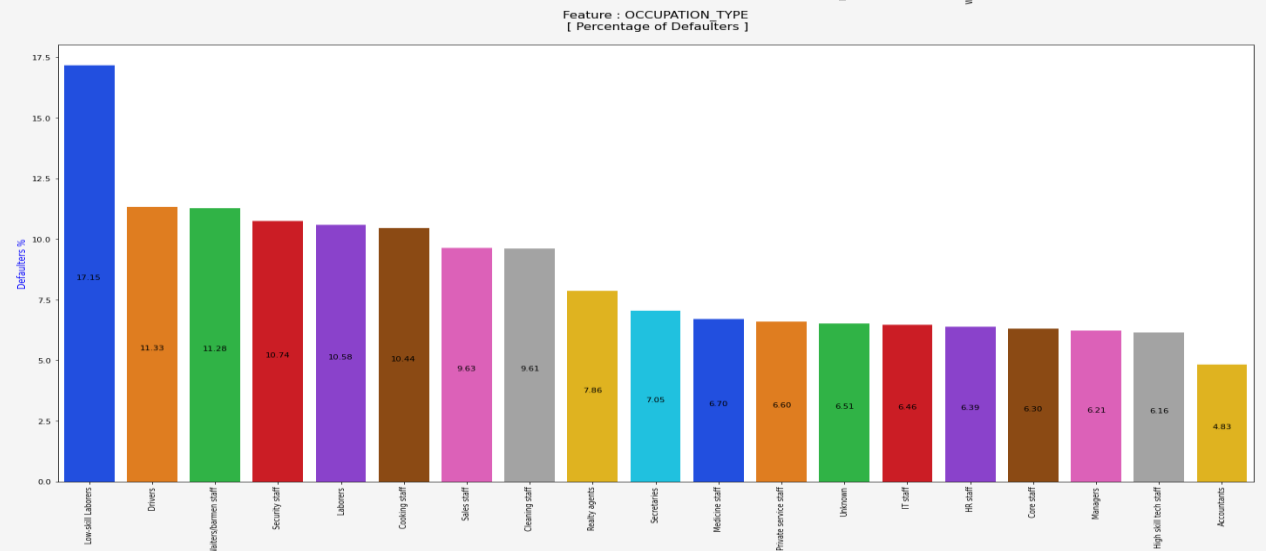
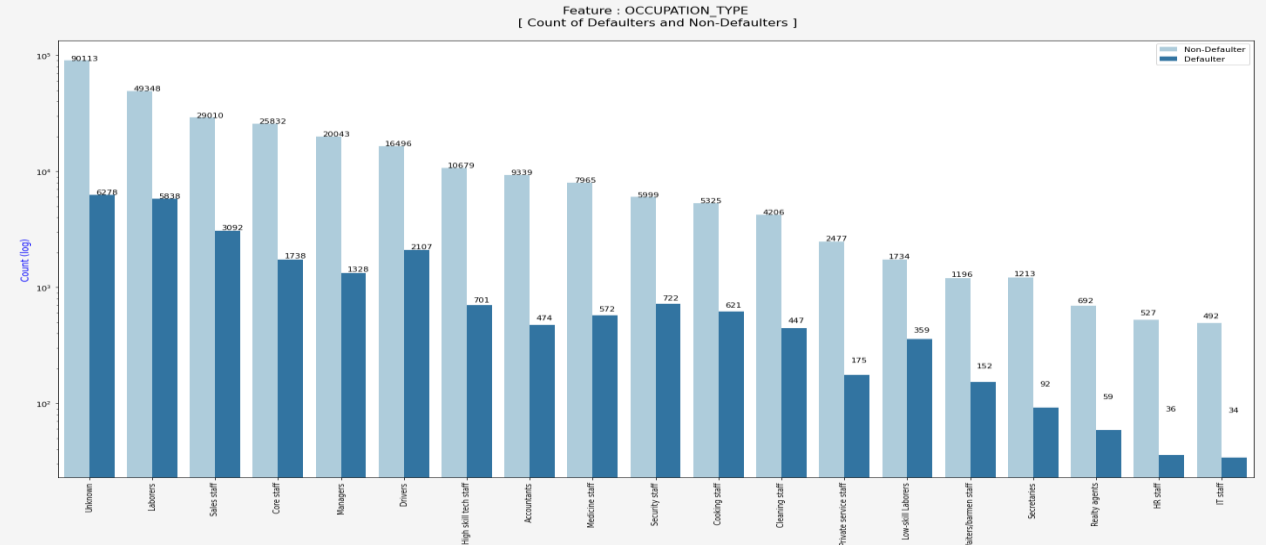
Feature : NAME\_HOUSING\_TYPE  
[ Percentage of Defaulters ]



## 4 Data Analysis: Univariate and Segmented Univariate Analysis

### ❖ Insights from Occupation Type

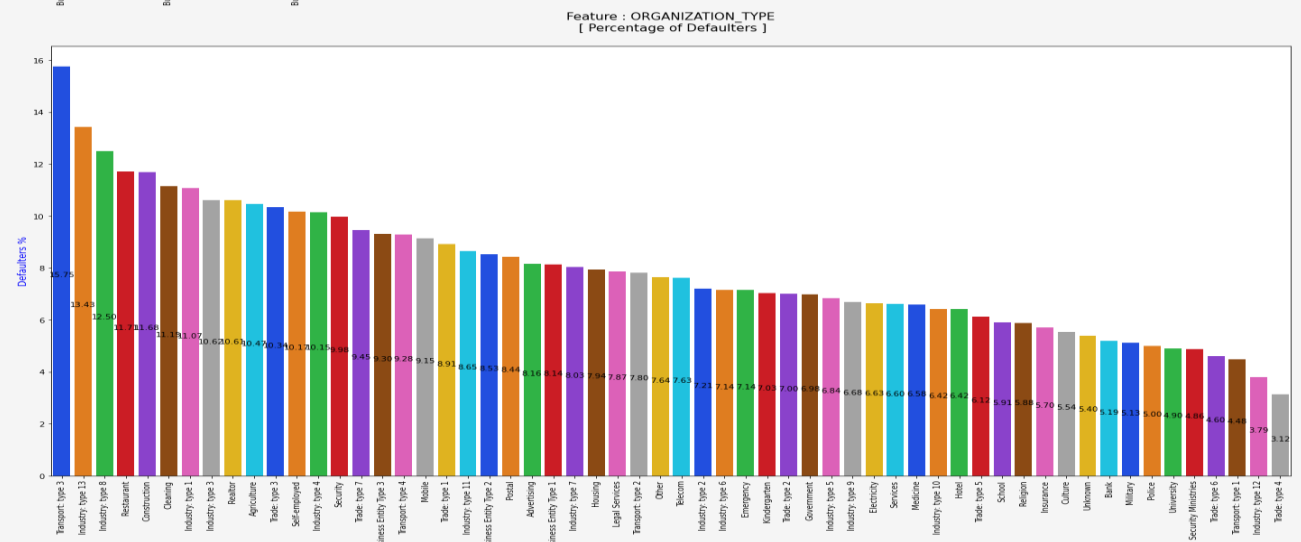
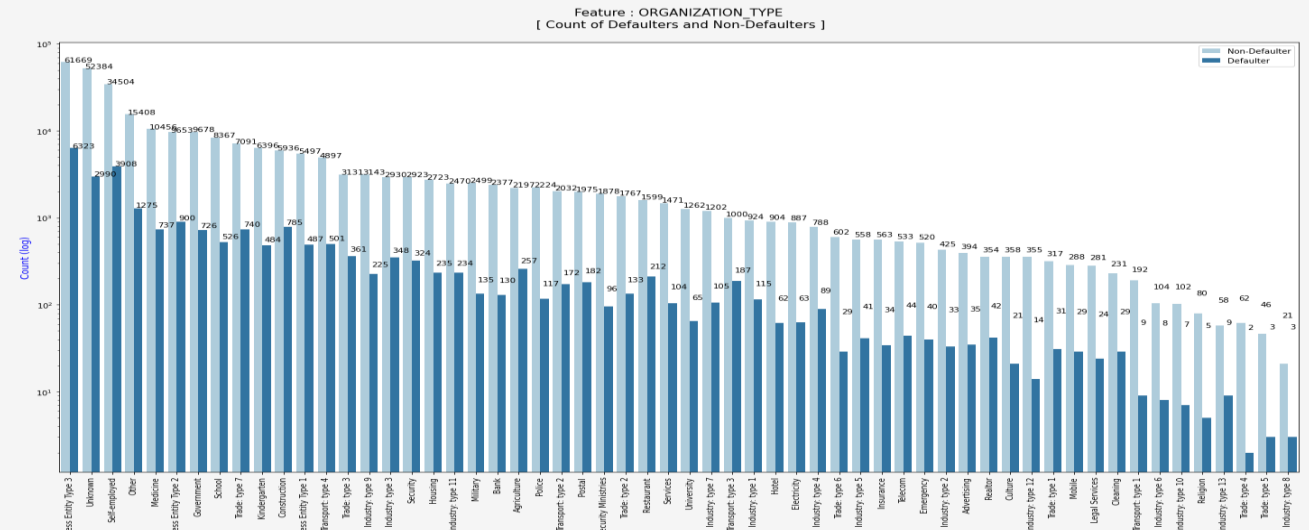
- ✓ Most of the clients are Labour class followed by Sales staff
- ✓ Very few clients have occupation type IT staff.
- ✓ Low-skill labourers have the highest default rate.





## ❖ 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

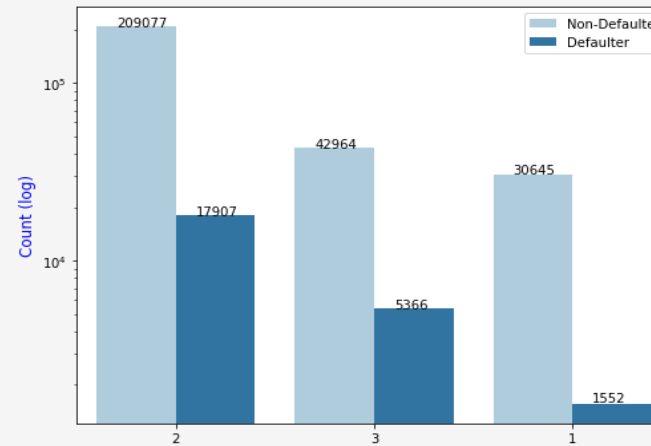
### ❖ Insights from Region Rating Client

- ✓ 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.

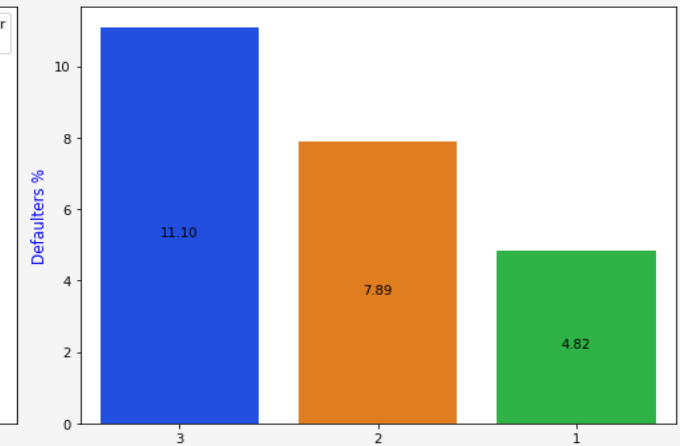
### ❖ Insights from Days Birth

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

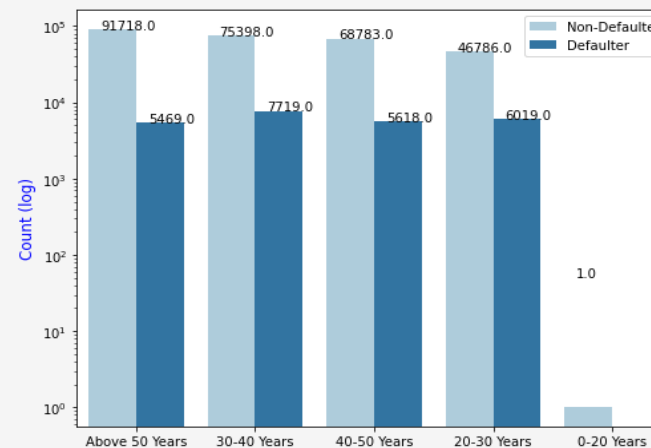
Feature : REGION\_RATING\_CLIENT  
[ Count of Defaulters and Non-Defaulters ]



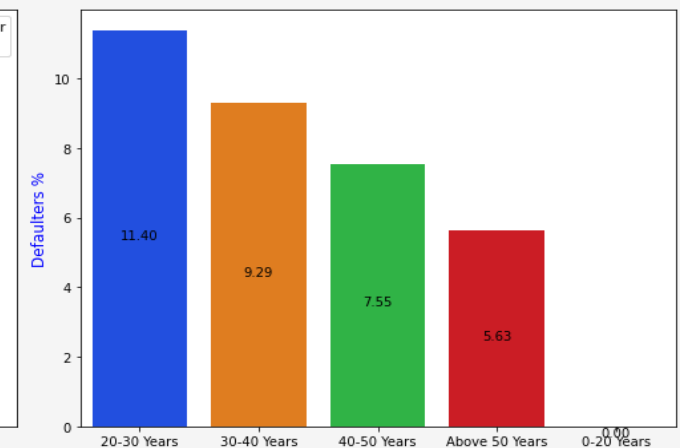
Feature : REGION\_RATING\_CLIENT  
[ Percentage of Defaulters ]



Feature : AGE\_RANGE  
[ Count of Defaulters and Non-Defaulters ]



Feature : AGE\_RANGE  
[ Percentage of Defaulters ]



## 4 Data Analysis: Univariate and Segmented Univariate Analysis

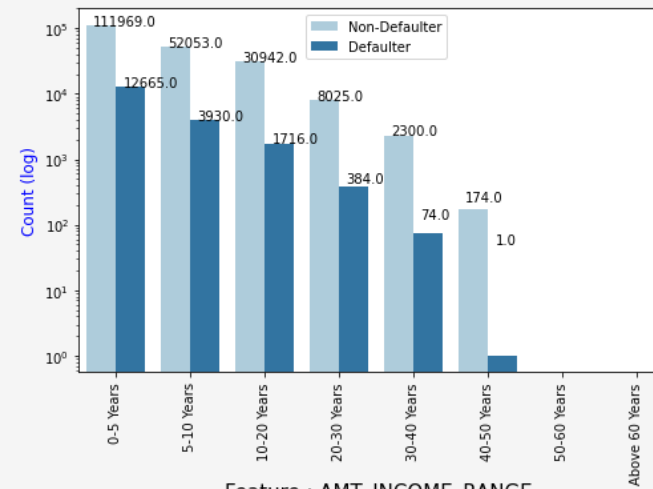
### ❖ Insights from Days Employment

- ✓ 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.

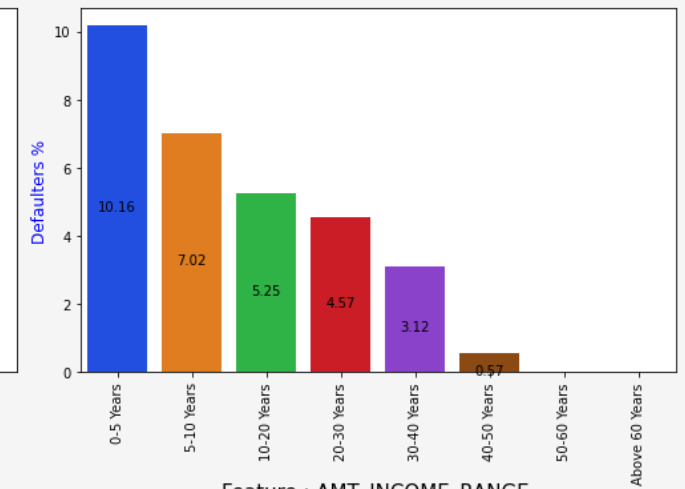
### ❖ Insights from Amount Income

- ✓ 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.

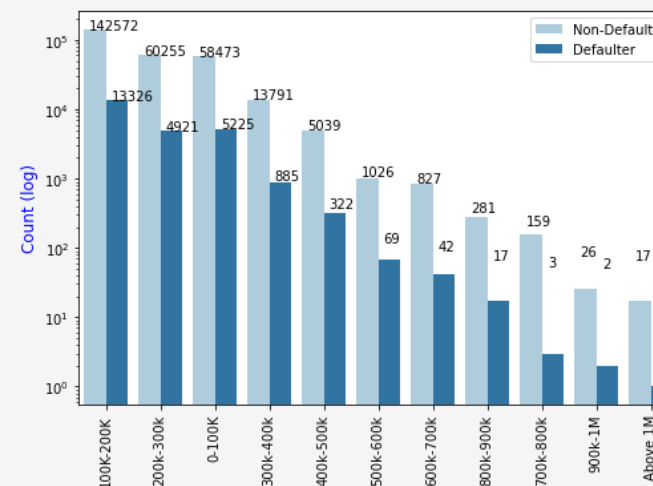
Feature : EMPLOYMENT\_YEAR\_RANGE  
[ Count of Defaulters and Non-Defaulters ]



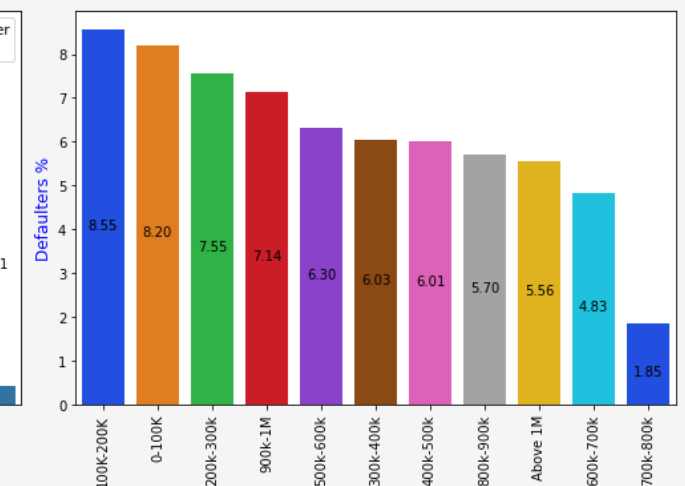
Feature : EMPLOYMENT\_YEAR\_RANGE  
[ Percentage of Defaulters ]



Feature : AMT\_INCOME\_RANGE  
[ Count of Defaulters and Non-Defaulters ]



Feature : AMT\_INCOME\_RANGE  
[ Percentage of Defaulters ]



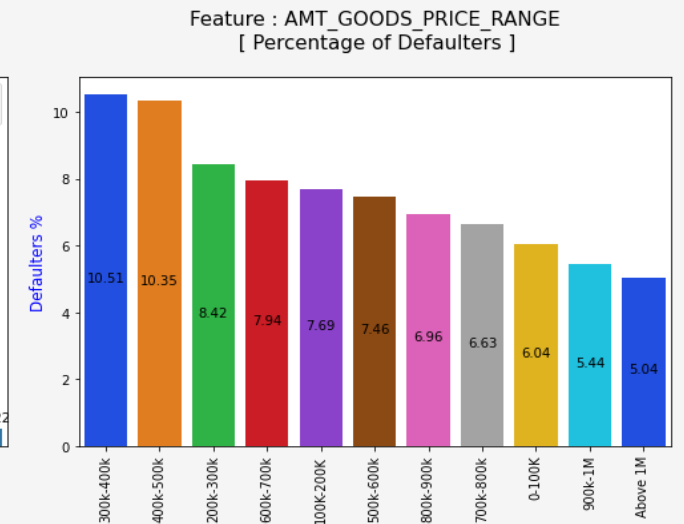
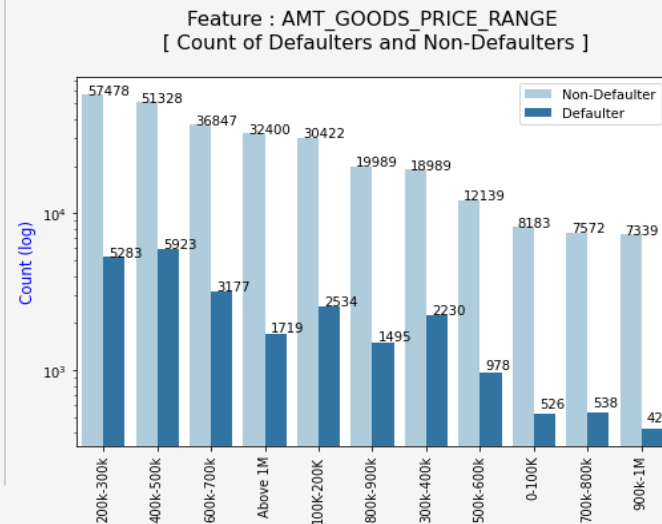
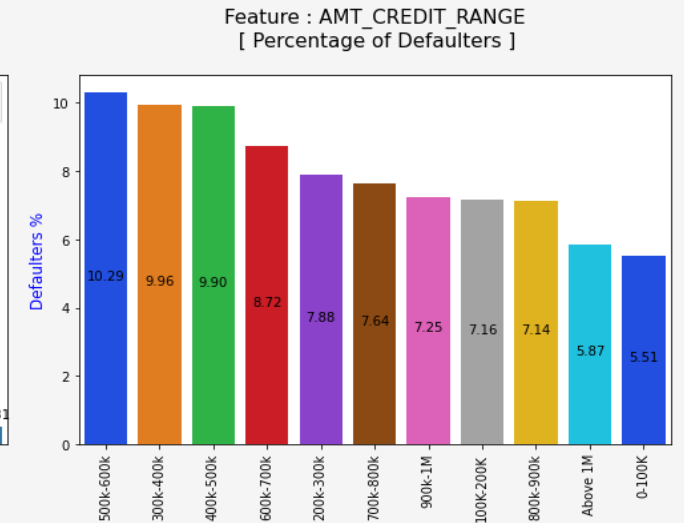
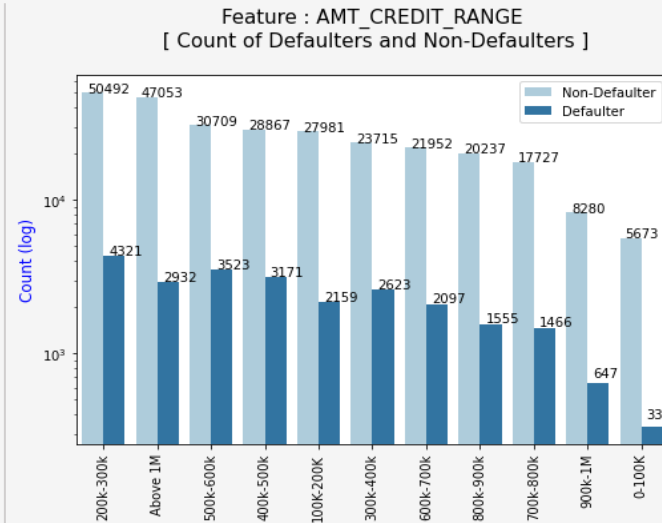
## 4 Data Analysis: Univariate and Segmented Univariate Analysis

### ❖ 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.



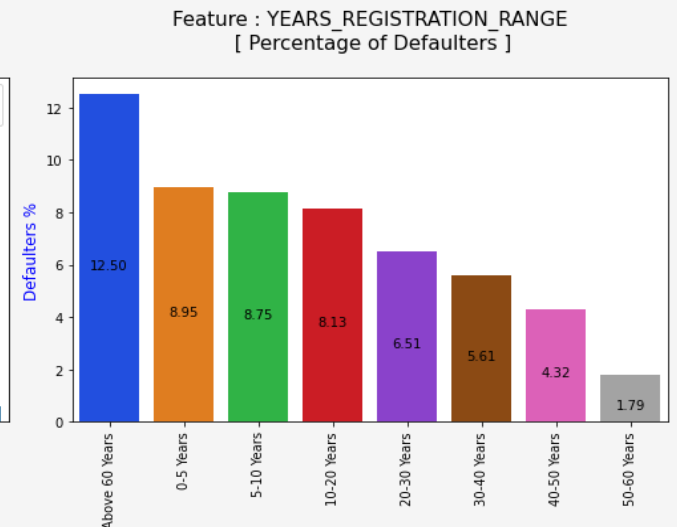
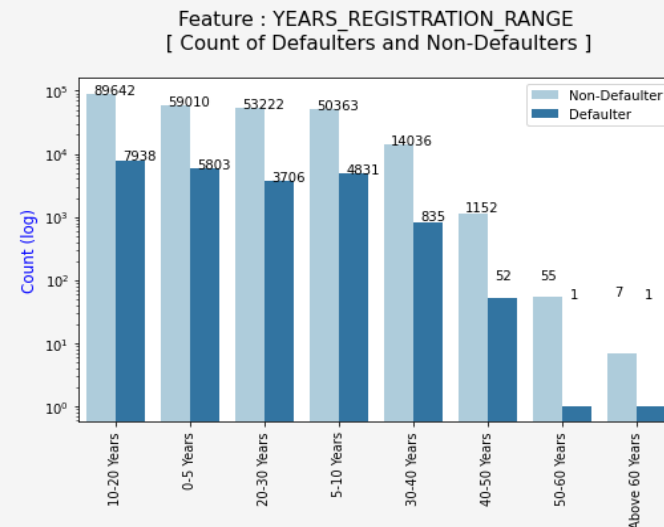
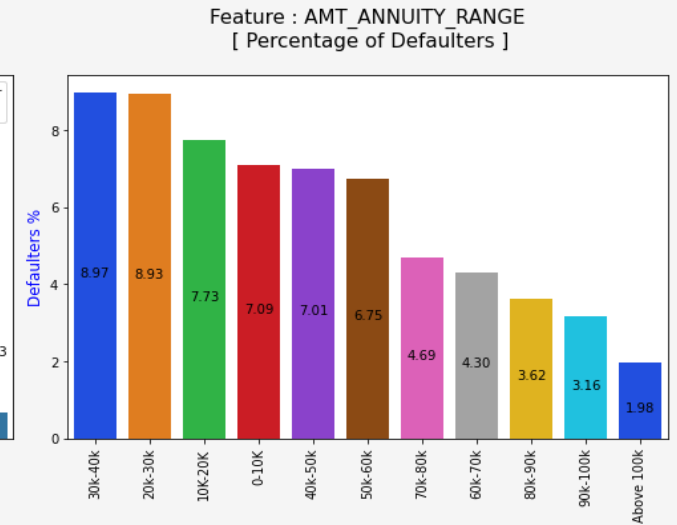
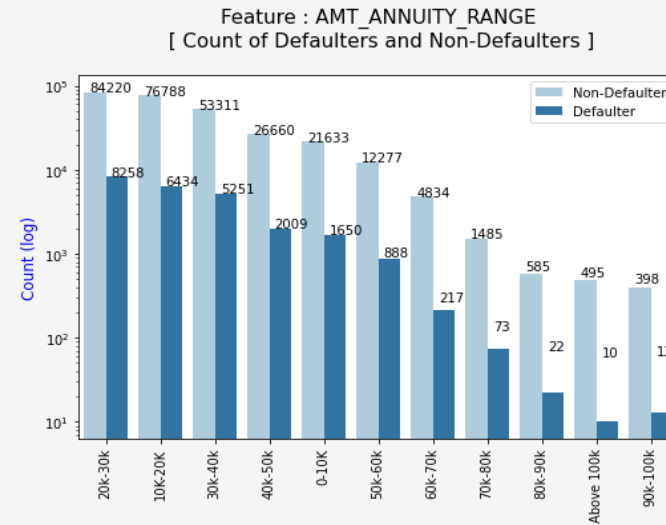
## 4 Data Analysis: Univariate and Segmented Univariate Analysis

### ❖ 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.



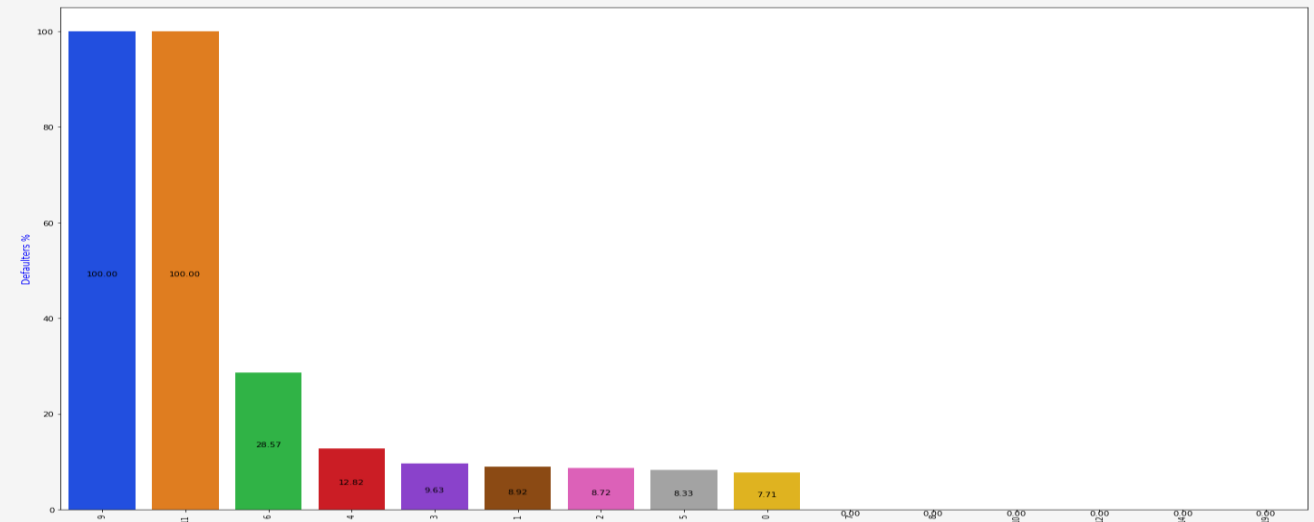
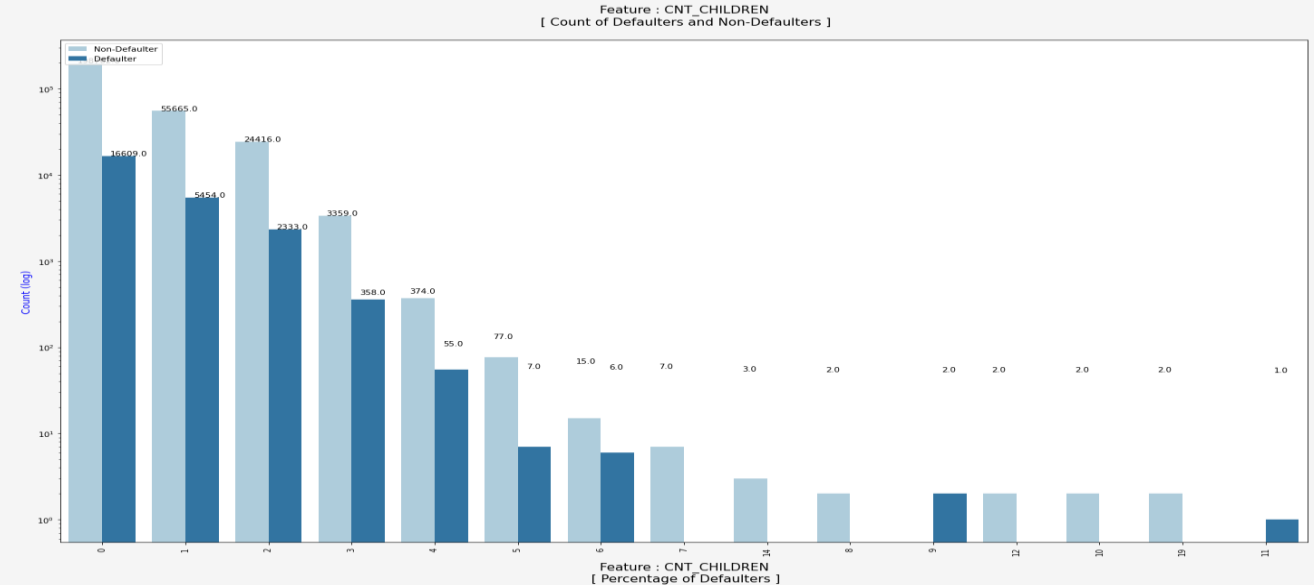
## 4 Data Analysis: Univariate and Segmented Univariate Analysis

### ❖ 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.



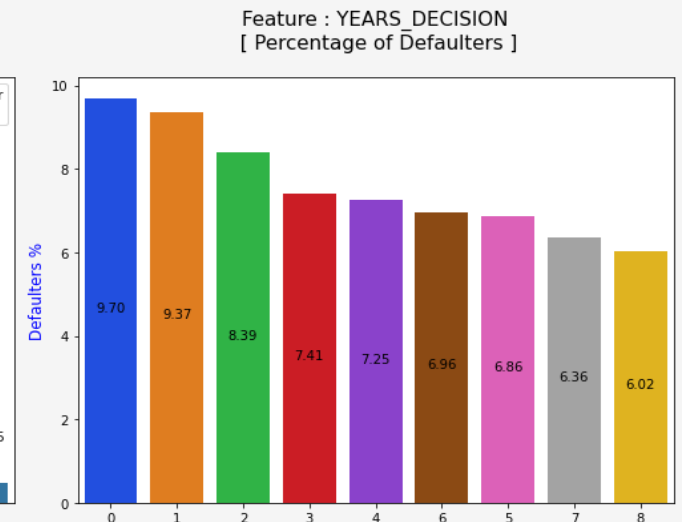
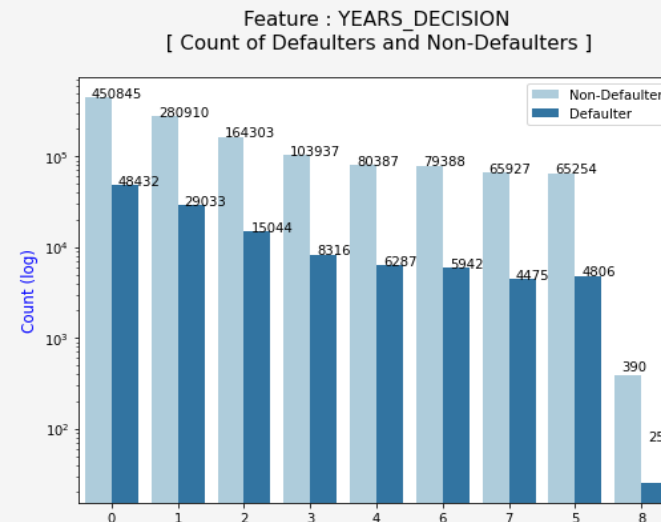
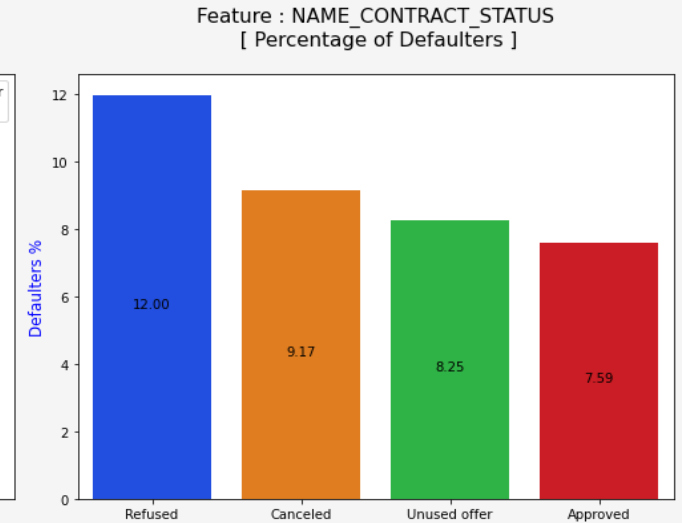
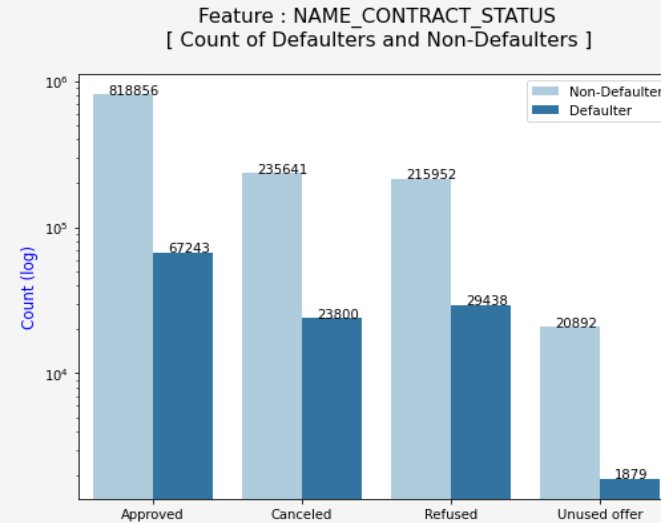
## 4 Data Analysis: Univariate and Segmented Univariate Analysis

### ❖ 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.



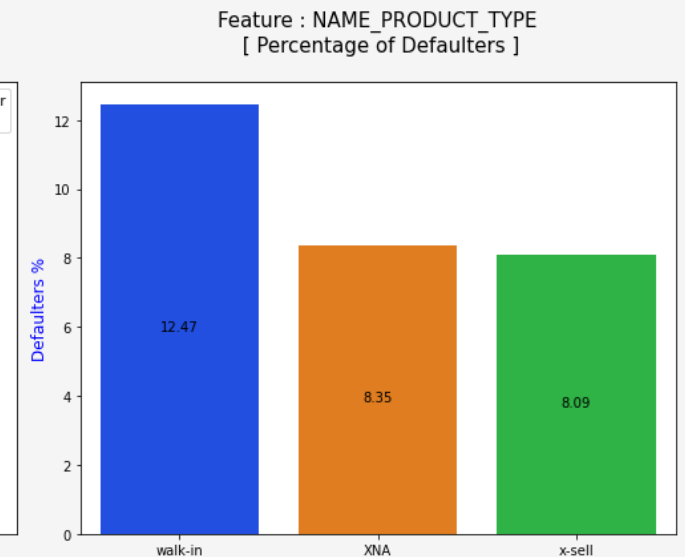
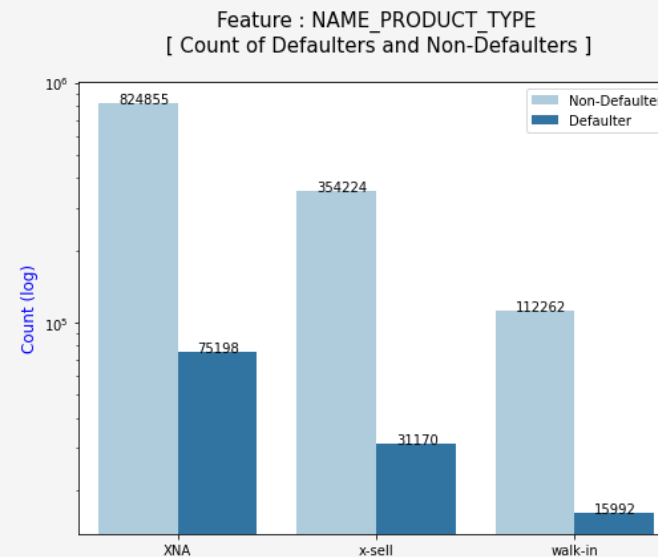
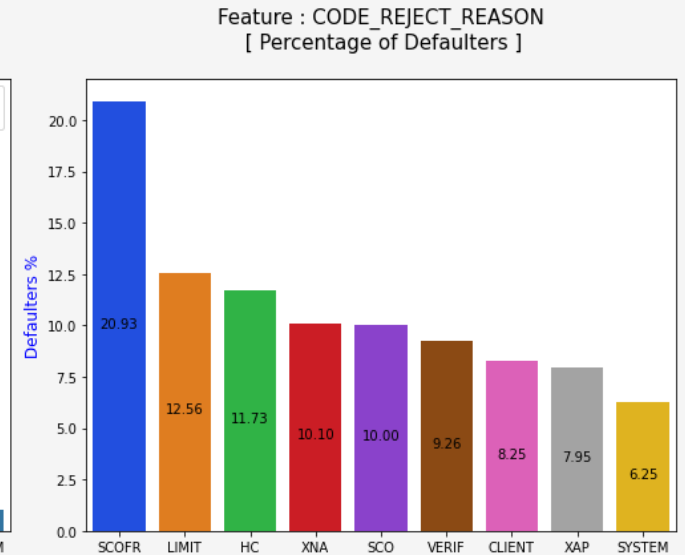
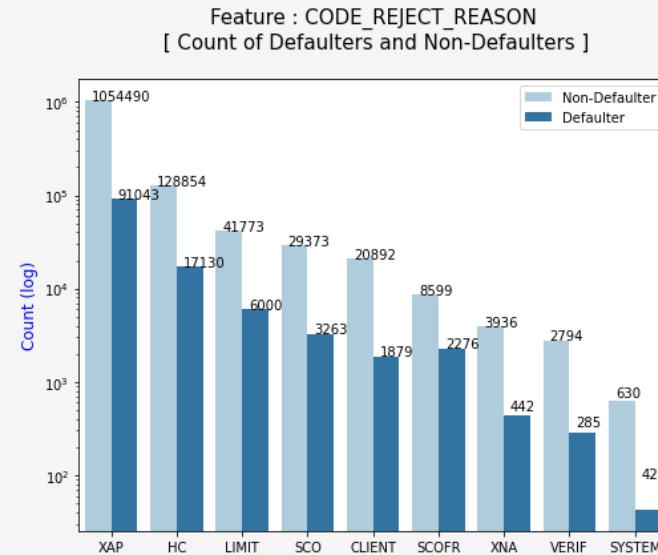
## 4 Data Analysis: Univariate and Segmented Univariate Analysis

### ❖ 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.



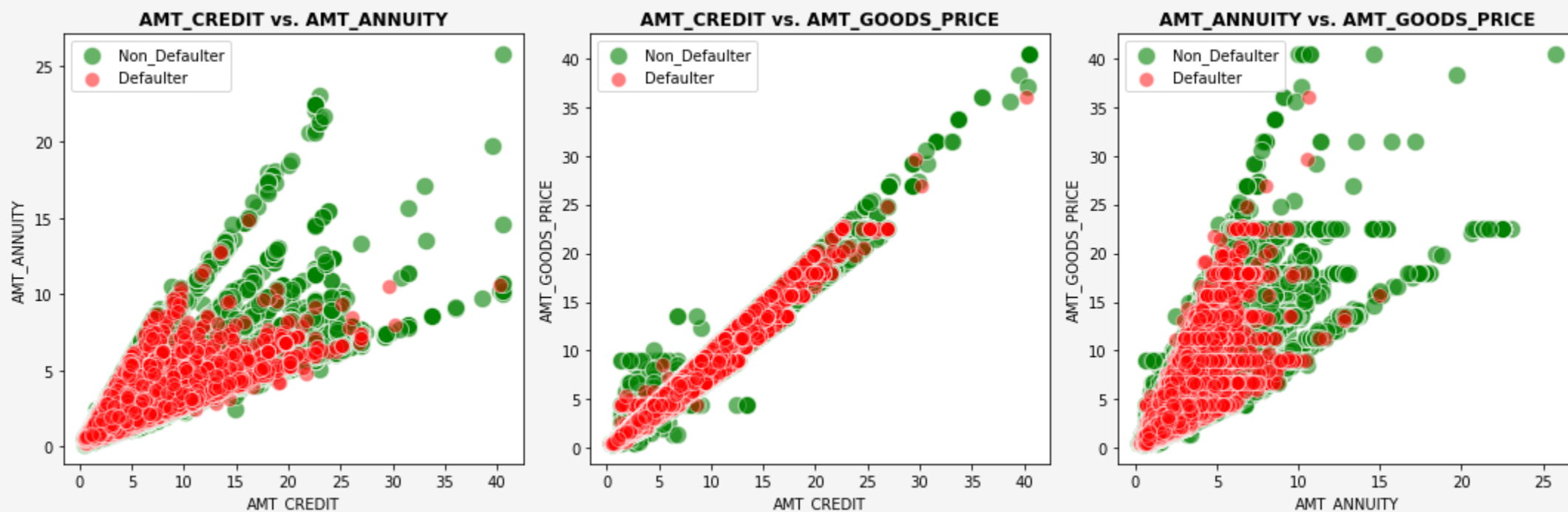


## 4

## Data Analysis: Bivariate/ Multivariate Analysis

### ❖ Insights

- ✓ There are very less defaulters for  $\text{AMT\_CREDIT} > 3\text{M}$
- ✓  $\text{AMT\_CREDIT}$  and  $\text{AMT\_GOODS\_PRICE}$  are highly correlated.
- ✓ When  $\text{AMT\_ANNUITY} > 15000$  and  $\text{AMT\_GOODS\_PRICE} > 3\text{M}$ , there is a lesser chance of defaulters



## 4 Data Analysis: Bivariate/ Multivariate Analysis

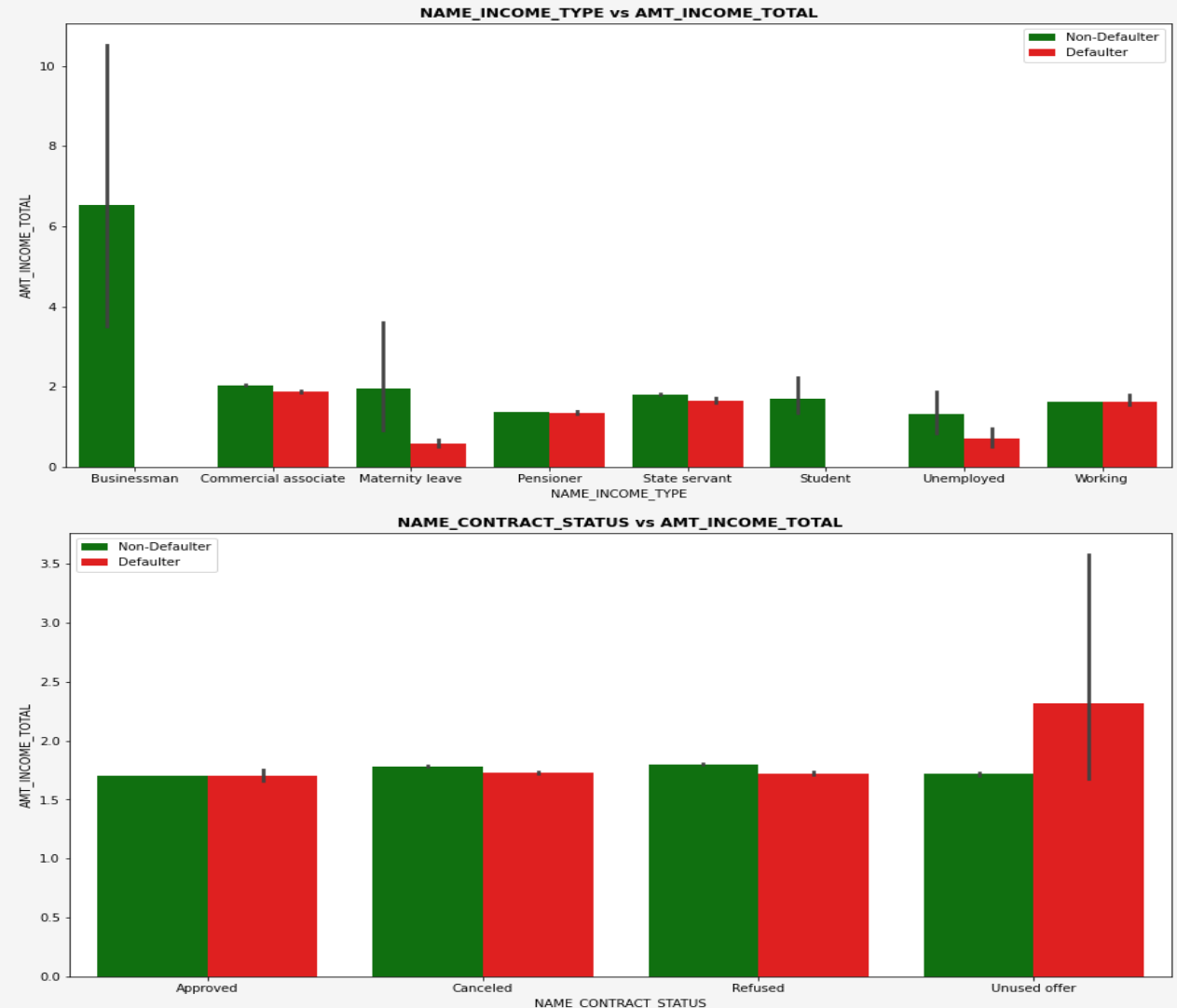
### ❖ Insights from:

#### 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 their average income is higher than others.



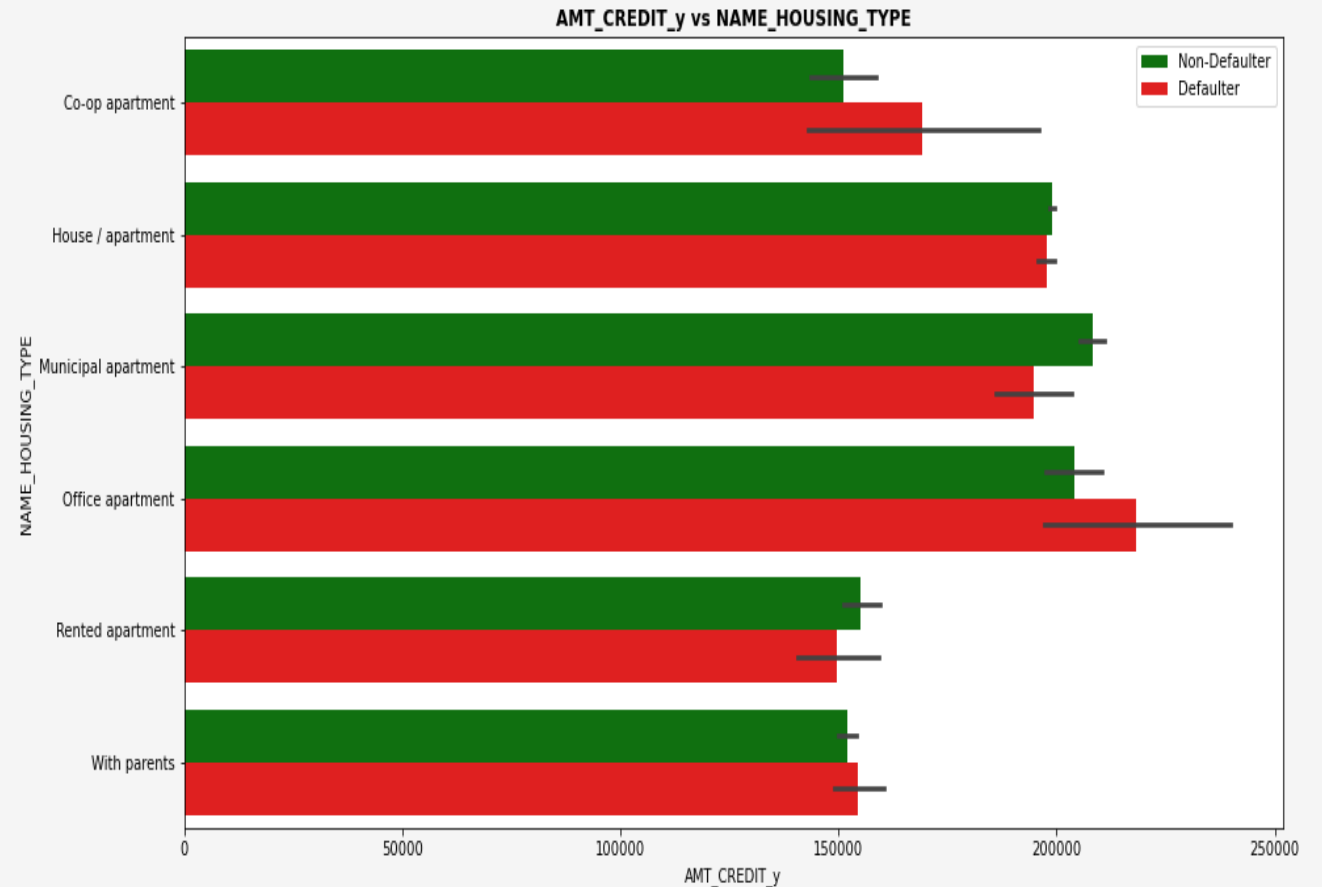
## 4

## Data Analysis: Bivariate/ Multivariate Analysis

### ❖ Insights from:

#### Amount Credit\_y Vs. Name Housing Type

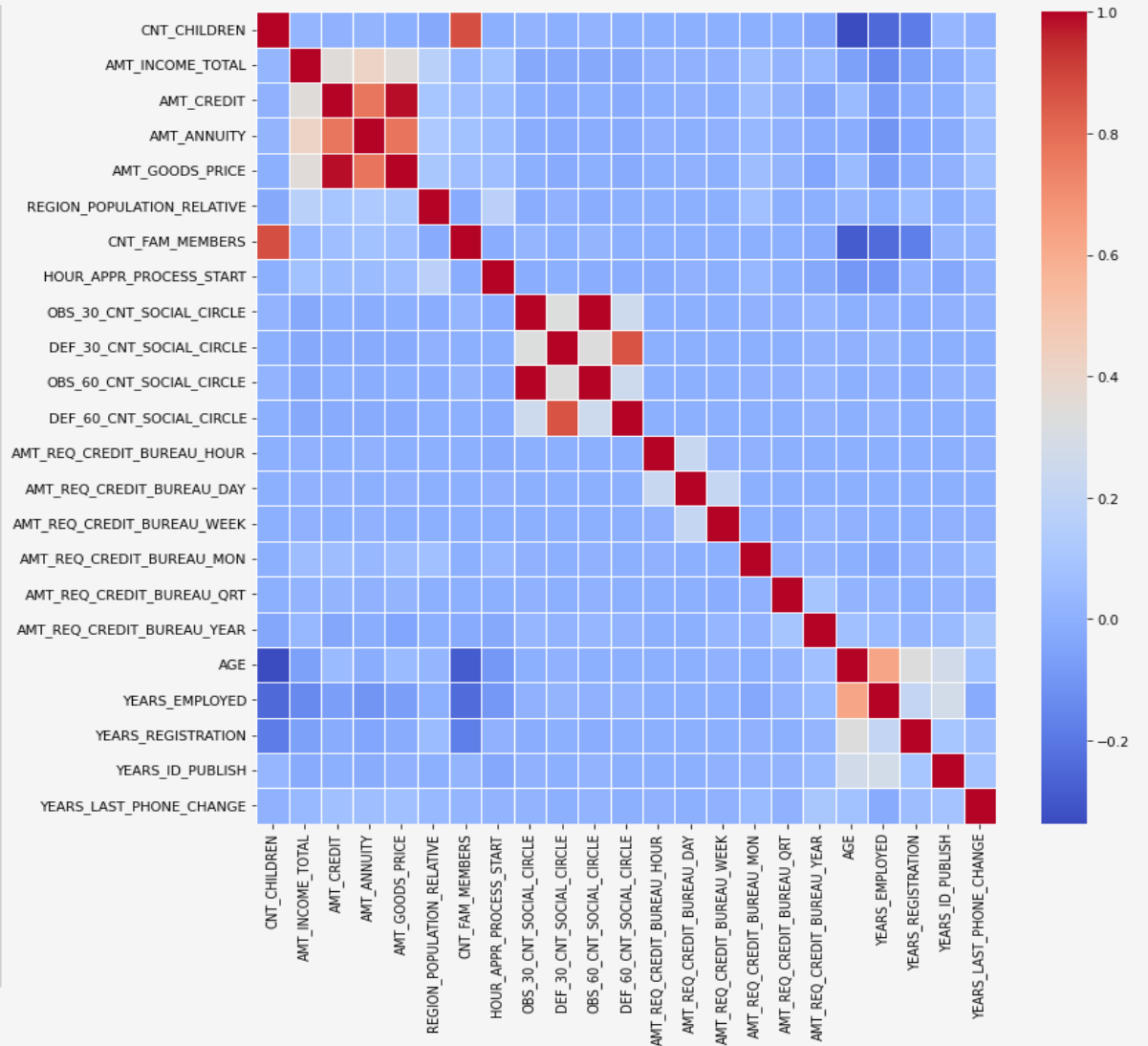
- ✓ "Office apartment" category has the higher credit as compared to others.
- ✓ Clients having housing type Co-op apartment have a very high default rate.
- ✓ Clients living in rented apartment default the least.



## 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



## ❖ Top 10 correlation for the Defaulter

Heatmap visualization showing the correlation matrix for 25 variables. The color scale ranges from -0.2 (dark red) to 1.0 (dark green). The diagonal is dark green (1.0). The matrix is symmetric. Key correlations include: AGE and YEARS\_EMPLOYED (approx. 0.8), AGE and YEARS\_REGISTRATION (approx. 0.6), AGE and YEARS\_ID\_PUBLISH (approx. 0.4), AGE and YEARS\_LAST\_PHONE\_CHANGE (approx. 0.2), and AGE and AMT\_REQ\_CREDIT\_BUREAU\_YEAR (approx. 0.2). The matrix shows a clear pattern of decreasing correlation as the distance between variables increases.

# Key Insights



## DRIVING FACTORS FOR NON-DEFAULTERS

## DRIVING FACTORS FOR DEFAULTERS

### ATTRIBUTES

**NAME\_CONTRACT\_TYPE:** Client applying for Revolving Loans.  
**CODE\_GENDER:** Female clients  
**NAME\_INCOME\_TYPE:** Clients who are student or businessmen  
**NAME\_EDUCATION\_TYPE:** Clients having academic degree.  
**REGION\_FAMILY\_STATUS:** Clients who are either married or widow  
**NAME\_HOUSING\_TYPE:** Clients living in office apartments  
**OCCUPATION\_TYPE:** Clients who are accountants  
**REGION\_RATING\_CLIENT:** Clients with Rating 1  
**ORGANIZATION TYPE:** Trade type-4 and Industry type-12  
**DAYS\_BIRTH:** Clients above 50 years of age  
**DAYS\_EMPLOYED:** Clients with work-ex of 40-50 years  
**AMT\_INCOME\_TOTAL:** Clients with income range in 700k-800k  
**AMT\_GOODS\_PRICE:** Goods price range above 1M  
**AMT\_ANNUITY RANGE:** Clients having annuity above 100k  
**CNT\_CHILDREN:** Clients having no children  
**CASH\_LOAN\_PURPOSE:** For buying a garage  
**NAME\_CLIENT\_TYPE:** Refreshed clients

**NAME\_CONTRACT\_TYPE:** Client applying for Cash Loans.  
**CODE\_GENDER:** Male clients  
**NAME\_INCOME\_TYPE:** Clients on maternity leave or unemployed  
**NAME\_EDUCATION\_TYPE:** Clients having lower secondary education  
**NAME\_FAMILY\_STATUS:** Single or Civil Marriage clients  
**NAME\_HOUSING\_TYPE:** Clients living in rented apartments or with parents  
**OCCUPATION\_TYPE:** Clients who are Low-skill labourers  
**REGION\_RATING\_CLIENT:** Clients with Rating 3  
**ORGANIZATION TYPE:** Transport type-3 and Industry type-13  
**DAYS\_BIRTH:** Clients between 20-30 years of age  
**DAYS\_EMPLOYED:** Clients with work-ex of 0-5 years  
**AMT\_INCOME\_TOTAL:** Clients with income range in 0-300k  
**AMT\_GOODS\_PRICE:** Goods price range between 300k-500k  
**AMT\_ANNUITY RANGE:** Clients having annuity between 20k-40k  
**CNT\_CHILDREN:** Clients having more children  
**CASH\_LOAN\_PURPOSE:** For hobby and car repairs  
**NAME\_CLIENT\_TYPE:** 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**