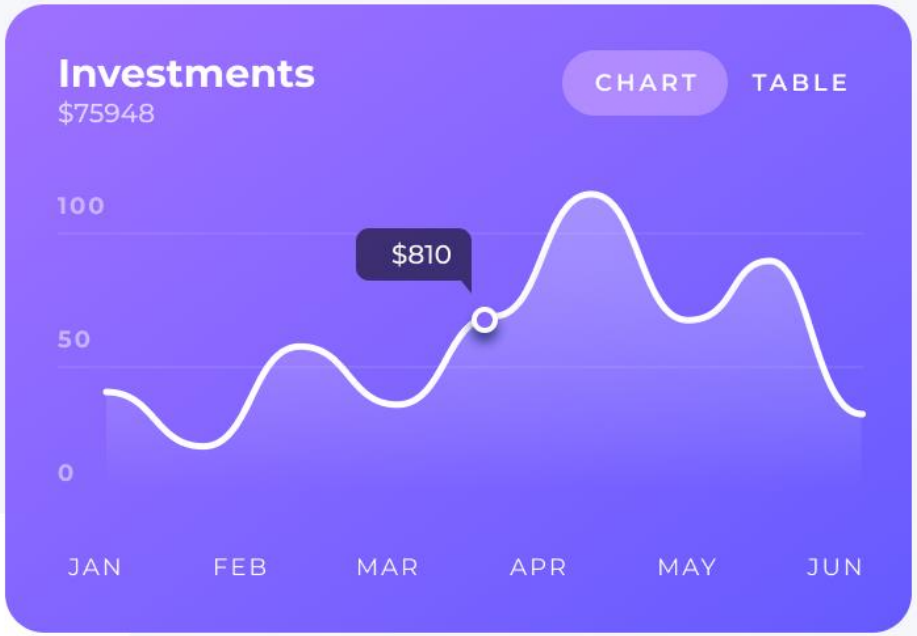
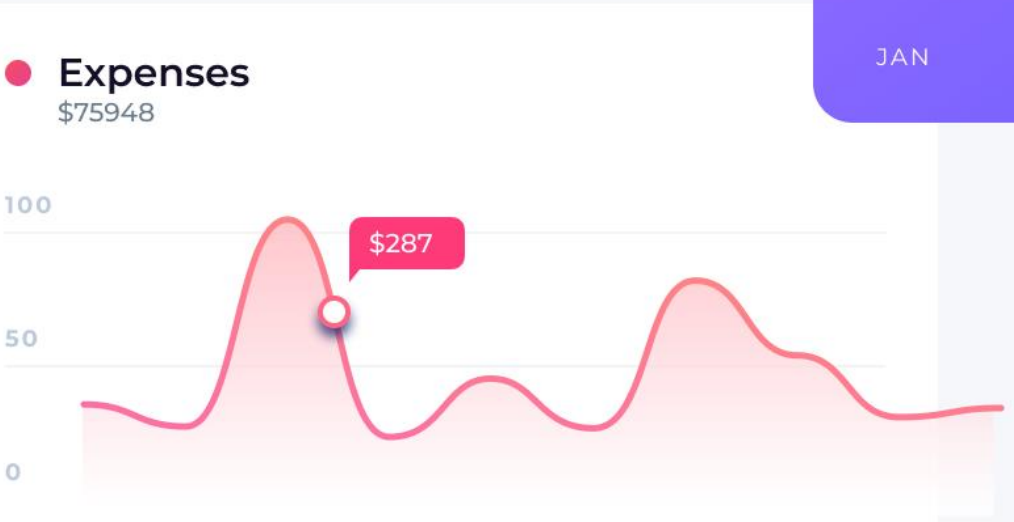
Bank Loan Case Study

Loan Case Study





CONTENT

- Project Description
- Approach
- Tech-Stack Used
- Project Analysis
- Correlation Analysis
- Conclusion

Project Description

The objective of project **“Bank loan case study”** is to make understand of basic knowledge on risk analytics in banking industry. The main aim is how data utilization is done to minimize the risk of losing money while lending to a customer and this project also aims to address challenges associated with loan non-repayors and the insights that closely connect with such loan non-repayors.

Objective: Understand risk analytics in banking and financial services to minimize losses in lending.

Risks Involved: Balancing between approving loans (loss of business) and potential defaults (financial losses).

Case Study Focus: Application of EDA in real-world banking context to identify driver variables for loan defaults.

Benefit: Develop insights into predictors of default, aiding risk reduction in consumer lending.

APPROACH

The Previous application and applications data are two substantial datasets that are used in this case study. Data cleaning is the initial step before analysis in order to manage blank data and eliminate unneeded columns. Prior to doing univariate and bivariate analysis utilizing tables and charts, outliers were also found and eliminated.

"DATA CLEANING"

Understand the given dataset through the data description table. Describe the size and shape of the dataset. Handle missing values in the dataset. Address any issues related to data imbalance.

"DATA ANALYSIS"

Perform univariate analysis on individual features. Conduct bivariate analysis to explore relationships between pairs of features. Analyze correlations between different variables in the dataset. Provide key insights and findings from the univariate analysis. Present important observations from the bivariate analysis. Highlight significant correlations and their implications for the business.

"REPORT"

Document all the steps and procedures taken during the data cleaning process. Present the results of the univariate and bivariate analyses. Conclude the report with a summary of the findings and their potential impact on risk analytics in consumer lending

TECH-STACK USED

In this project I used Google Colab along with Python, for strong data analysis and exploration. For my presentation's visual appeal and clarity, I used Microsoft PowerPoint, enabling me to create engaging slides showcasing the project's key insights and findings.

PROJECT ANALYSIS

1

Identified the missing data and used appropriate methods to deal with it

2

Identified outliers in the Datasets

3

Identified Data imbalance in the data

4

Results of univariate , segmented univariate, bivariate analysis are explained

5

Found Correlations

Reference : The basics you can find anywhere 5 Steps To Successful Storytelling Published on April 5, 2014 Featured in: Marketing & Advertising

Data Analysis Setup & Initial Exploration

1. Analyzed loan application data using Python in Google Colab.

2. Imported Python libraries - NumPy, Pandas, Matplotlib, and Seaborn - for data processing and visualization.

3. Integrated Google Drive to access input files seamlessly.

4. Loaded two datasets - "Application_Data" and "Previous_Application" - using Pandas' read_csv() function.

5. Displayed the first few rows of each dataset for an initial understanding of the data's structure.

6. Determined the dimensions (rows and columns) and sizes of both datasets using shape and size attributes.

7. Inspected column data types with info() method for data consistency.

8. Explored numeric variable summaries using the describe() method for distribution and statistics of numerical features.

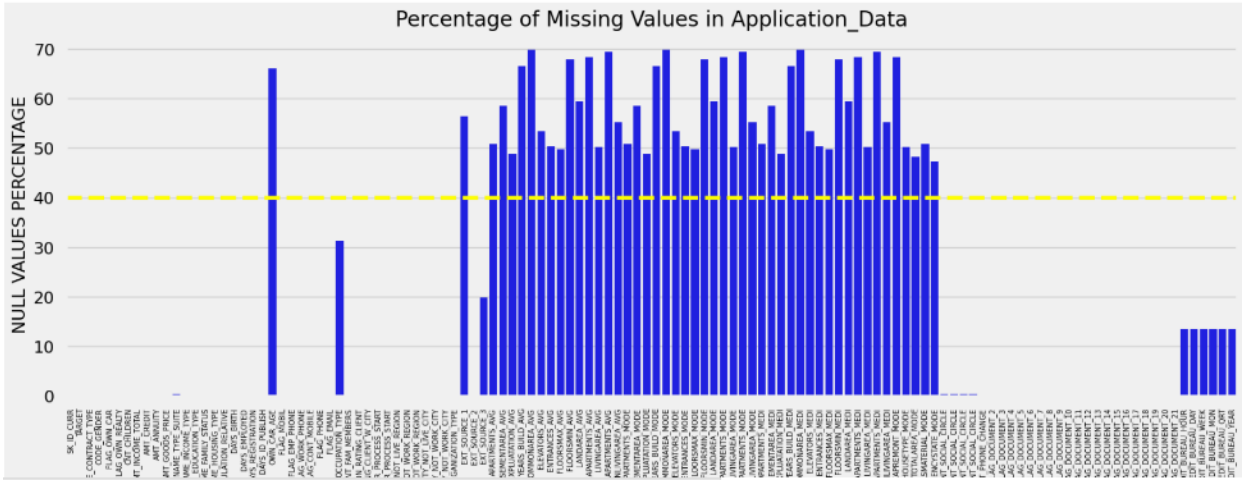
1 Data Cleaning

Identified the missing data and used appropriate methods to deal with it

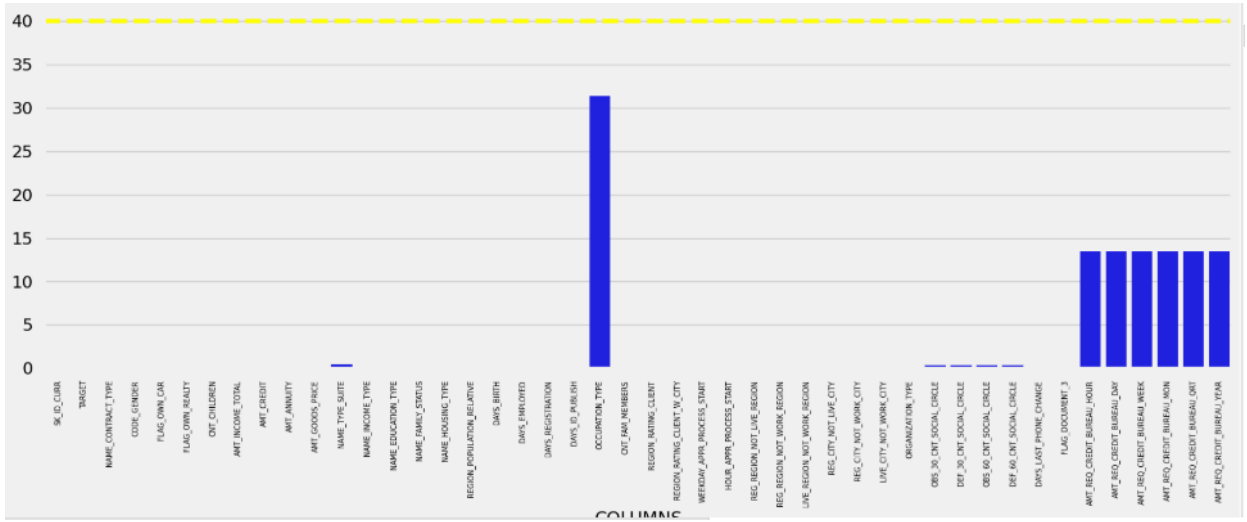
- The code first checks for missing values in the "Application_Data" dataset. Columns with more than 40% missing values are dropped:

- EXT_SOURCE_X1,EXT_SOURCE_X2,EXT_SOURCE_X3,EXT_SOURCE_X4,EXT_SOURCE_X5,FLAG_DOCUMENT_2,FLAG_DOCUMENT_4,FLAG_DOCUMENT_5,FLAG_DOCUMENT_6,FLAG_DOCUMENT_7,FLAG_DOCUMENT_8,FLAG_DOCUMENT_9

- The remaining columns are then checked for correlation with the target variable. Columns with low correlation are dropped:
- TARGET,EXT_SOURCE_1,EXT_SOURCE_2,EXT_SOURCE_3,FLAG_DOCUMENT_3



Before Cleaning Data



After Cleaning Data

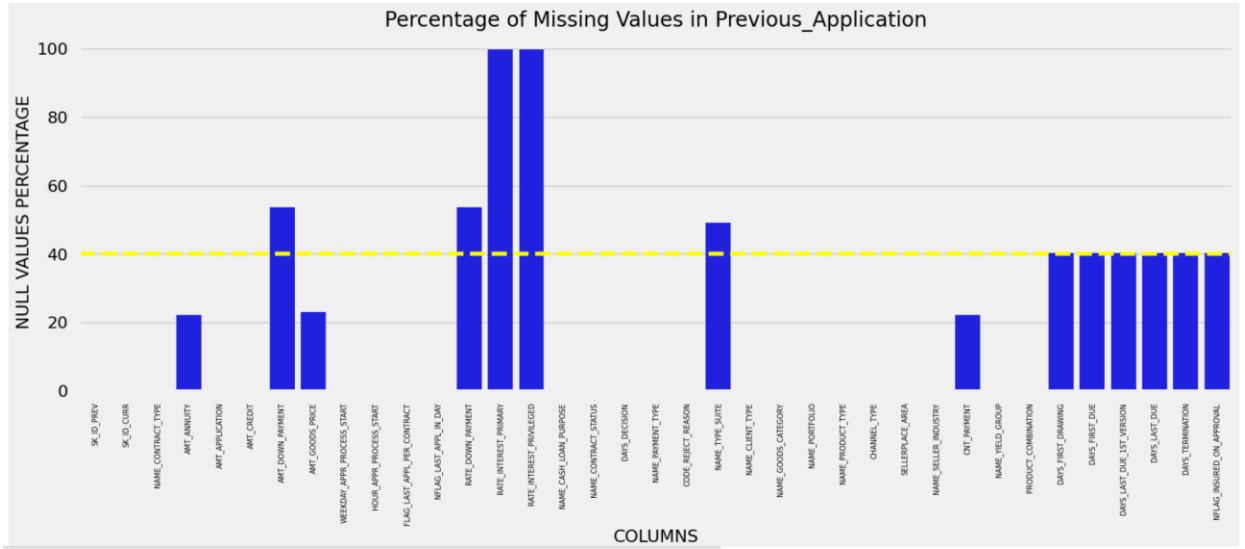
Data Cleaning

- The code first checks for missing values in the **"Previous_Application"** dataset. Columns with more than 40% missing values are dropped:

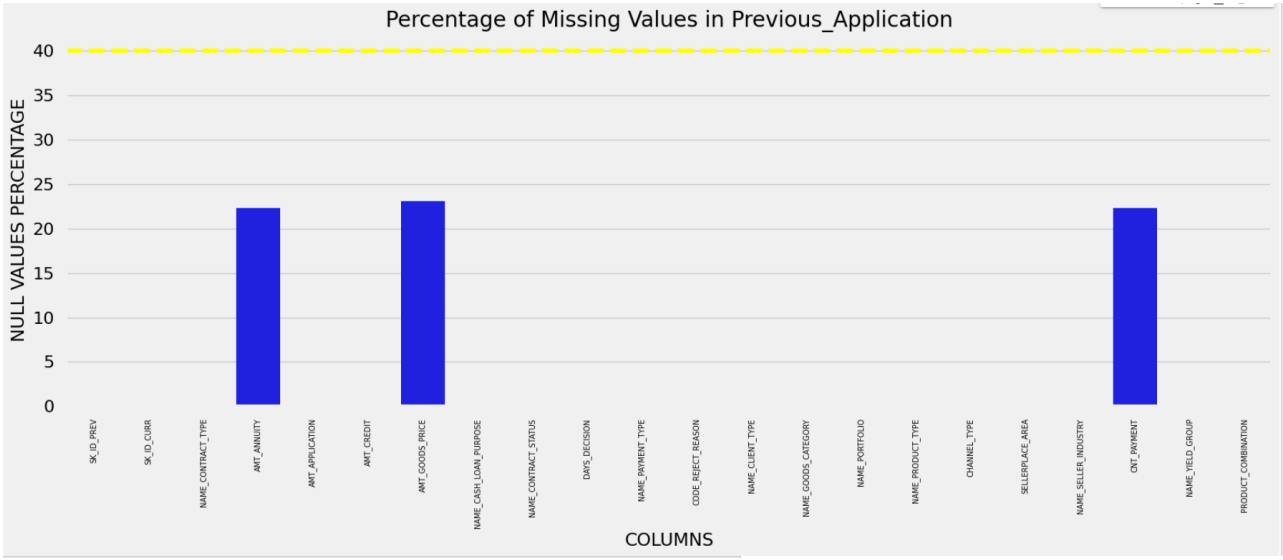
- CNT_PAYMENT,AMT_CREDIT,AMT_DOWNPAYME,AMT_GOODS_PRICE,DAYS_LAST_PHONE_CHANG,DAYS_EMPLOYED,REGION_RISK

- The remaining columns are then converted to categorical columns for better analysis. This is done using the **"astype()"** function.

- The **"DAYS_DECISION"** column is also converted to a categorical column. This is done by creating a new column called "DAYS_DECISION_GROUP" that groups the data into bins based on the number of days between the current application and the previous application.



Before Cleaning Data



After Cleaning Data

Data Cleaning

- "Application_Data" dataset missing values are filled with as follows:

• The "NAME_TYPE_SUITE" column is filled with the mode (most frequent value) of the column.
• The "OCCUPATION_TYPE" column is filled with the value "Unknown" for missing values since it represents a significant proportion of missing values.
• For some columns related to the number of inquiries to the Credit Bureau ("AMT_REQ_CREDIT_BUREAU_X"), missing values are filled with the median value of the respective column.

- "Previous_Application" dataset missing values are filled with as follows:

• The "AMT_ANNUITY" column is filled with the median value of the column.
• The "AMT_GOODS_PRICE" column is filled with the mode (most frequent value) of the column.
• The "CNT_PAYMENT" column represents the number of installments of the loan in question, and missing values are filled with 0, assuming no installment.

Highest percentage of values belongs to Unknown group and Second highest percentage belongs to Laborers

```
Application_Data[['AMT_REQ_CREDIT_BUREAU_HOUR','AMT_REQ_CREDIT_BUREAU_DAY',
                  'AMT_REQ_CREDIT_BUREAU_WEEK','AMT_REQ_CREDIT_BUREAU_MON',
                  'AMT_REQ_CREDIT_BUREAU_QRT','AMT_REQ_CREDIT_BUREAU_YEAR']].describe()
```

	AMT_REQ_CREDIT_BUREAU_HOUR	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR
count	307511.000000	307511.000000	307511.000000	307511.000000	307511.000000	307511.000000
mean	0.005538	0.006055	0.029723	0.231293	0.229631	1.778463
std	0.078014	0.103037	0.190728	0.856810	0.744059	1.765523
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
50%	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
75%	0.000000	0.000000	0.000000	0.000000	0.000000	3.000000
max	4.000000	9.000000	8.000000	27.000000	261.000000	25.000000

Cleaned Application_Data dataset without null vales

```
display(Previous_Application.isnull())
```

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	NAME_CASH_LOAN_PURPOSE	NAME_CONTRACT_STATUS	DAYS_DECISION	...	NAME_CLIENT_TYPE	NAME_GOODS_
0	False	False	False	False	False	False	False	False	False	False	...	False	
1	False	False	False	False	False	False	False	False	False	False	...	False	
2	False	False	False	False	False	False	False	False	False	False	...	False	
3	False	False	False	False	False	False	False	False	False	False	...	False	
4	False	False	False	False	False	False	False	False	False	False	...	False	
...	...	...	...	...	...	...	...	...	...	...	...	...	
1670209	False	False	False	False	False	False	False	False	False	False	...	False	
1670210	False	False	False	False	False	False	False	False	False	False	...	False	
1670211	False	False	False	False	False	False	False	False	False	False	...	False	
1670212	False	False	False	False	False	False	False	False	False	False	...	False	
1670213	False	False	False	False	False	False	False	False	False	False	...	False	

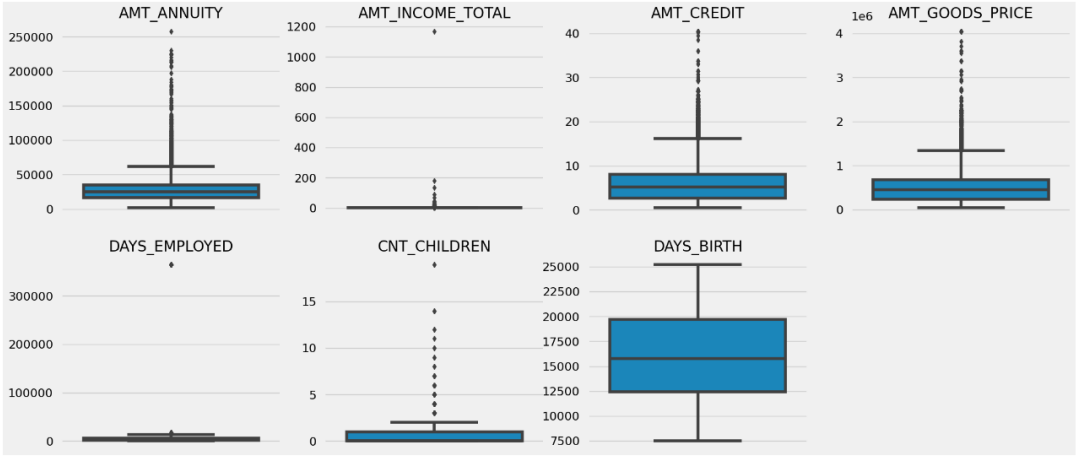
1670214 rows × 22 columns

Cleaned Previous_Application dataset without null values

2 Outliers in Datasets

Identifying Outliers in Application_Data

- Here I used boxplots to identify outliers in the "Application_Data" dataset. The boxplots showed that there were potential outliers on the upper end for the columns:
- AMT_ANNUITY,AMT_INCOME_TOTAL, AMT_CREDIT,AMT_GOODS_PRICE,DAYS_EMPLOYED
- The outliers on the upper end of these columns may represent unusually high values that I need to investigate these outliers to verify their validity and relevance to the analysis.
- I also found potential outliers on the lower end for the column DAYS_EMPLOYED.
- These outliers may represent negative employment days (data errors). I investigated these outliers to correct the data



Handling Missing Values in Application_Data

- I analyzed the percentage of missing values for each column in the "Application_Data" dataset.
- Where found that 76 columns had more than 40% missing values. so, dropped those columns as they lacked significant data.
- I also analyzed the correlations between the columns to identify any irrelevant columns.
- I dropped the 10 columns with the lowest correlations with the target variable.
- I filled the missing values in the remaining columns using appropriate imputation methods. Where I used the mode for categorical columns and the median for numerical columns.

	AMT_ANNUITY	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_GOODS_PRICE	DAYS_BIRTH	CNT_CHILDREN	DAYS_EMPLOYED
count	307499.000000	307511.000000	307511.000000	3.072330e+05	307511.000000	307511.000000	307511.000000
mean	27108.573909	1.687979	5.990260	5.383962e+05	16036.995067	0.417052	67724.742149
std	14493.737315	2.371231	4.024908	3.694465e+05	4363.988632	0.722121	139443.751806
min	1615.500000	0.256500	0.450000	4.050000e+04	7489.000000	0.000000	0.000000
25%	16524.000000	1.125000	2.700000	2.385000e+05	12413.000000	0.000000	933.000000
50%	24903.000000	1.471500	5.135310	4.500000e+05	15750.000000	0.000000	2219.000000
75%	34596.000000	2.025000	8.086500	6.795000e+05	19682.000000	1.000000	5707.000000
max	258025.500000	1170.000000	40.500000	4.050000e+06	25229.000000	19.000000	365243.000000

2 Outliers in Datasets

Identifying Outliers in Previous_Application

- Here, I used boxplots to visualize the distribution of selected numerical columns.
- Boxplots helped me identify extreme values that may be outliers.

• AMT_ANNUITY • AMT_APPLICATION • AMT_CREDIT • AMT_GOODS_PRICE	• SELLERPLACE_AREA • SK_ID_CURR • DAYS_DECISION • CNT_PAYMENT

- I found potential outliers on the upper end for the following columns:

• AMT_ANNUITY • AMT_APPLICATION • AMT_CREDIT	• AMT_GOODS_PRICE • SELLERPLACE_AREA • CNT_PAYMENT

- The outliers on the upper end of these columns may represent unusually

Handling Missing Values in Previous_Application

- I analyzed the percentage of missing values for each column in the dataset.
- Columns with high missing values:

• CNT_PAYMENT: 67.04% • AMT_CREDIT: 42.59% • AMT_GOODS_PRICE: 40.61%	• AMT_REQ_CREDIT_BUREAU_AMT_1: 21.63% • AMT_REQ_CREDIT_BUREAU_AMT_2: 22.08%

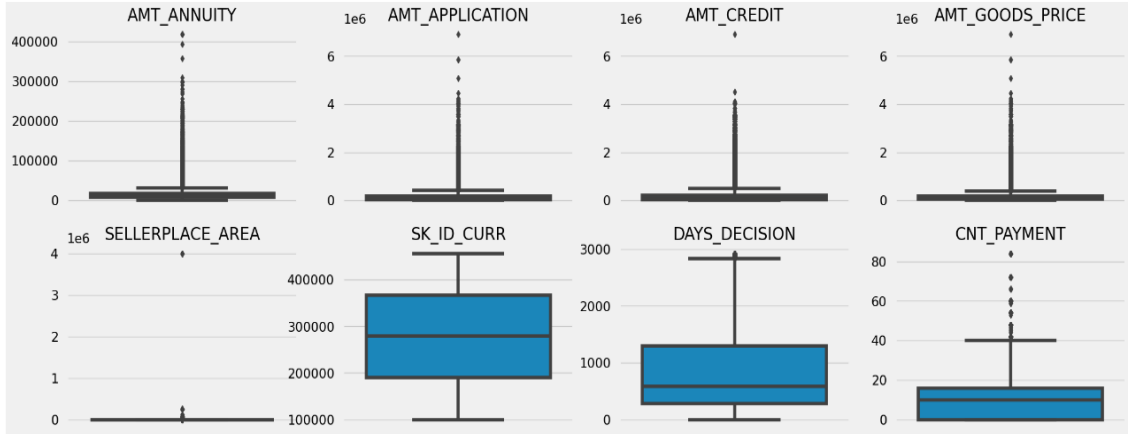
- I analyzed the relevance of these columns to loan default prediction.
- Columns with more than 40% missing values:

• CNT_PAYMENT	• AMT_CREDIT	• AMT_GOODS_PRICE

- I dropped these columns as they lacked significant data.
- Columns with less than 40% missing values:
- For columns with less than 40% missing values, I imputed the missing values using the mean for numerical columns and the mode for categorical columns.
- As a result total of 6 columns were dropped from the "Previous_Application" dataset.

2 Outliers in Datasets

Identifying Outliers in Previous_Application



- AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
- CNT_PAYMENT has few outlier valu
- SK_ID_CURR is an ID column and hence no outliers.
- DAYS_DECISION has few number of outliers indicating that these previous applications decisions were taken long back.

Handling Missing Values in Previous_Application

	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	SELLERPLACE_AREA	CNT_PAYMENT	DAYS_DECISION
count	1.670214e+06	1.670214e+06	1.670213e+06	1.670214e+06	1.670214e+06	1.670214e+06	1.670214e+06
mean	1.490651e+04	1.752339e+05	1.961140e+05	1.856429e+05	3.139511e+02	1.247621e+01	8.806797e+02
std	1.317751e+04	2.927798e+05	3.185746e+05	2.871413e+05	7.127443e+03	1.447588e+01	7.790997e+02
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	-1.000000e+00	0.000000e+00	1.000000e+00
25%	7.547096e+03	1.872000e+04	2.416050e+04	4.500000e+04	-1.000000e+00	0.000000e+00	2.800000e+02
50%	1.125000e+04	7.104600e+04	8.054100e+04	7.105050e+04	3.000000e+00	1.000000e+01	5.810000e+02
75%	1.682403e+04	1.803600e+05	2.164185e+05	1.804050e+05	8.200000e+01	1.600000e+01	1.300000e+03
max	4.180581e+05	6.905160e+06	6.905160e+06	6.905160e+06	4.000000e+06	8.400000e+01	2.922000e+03

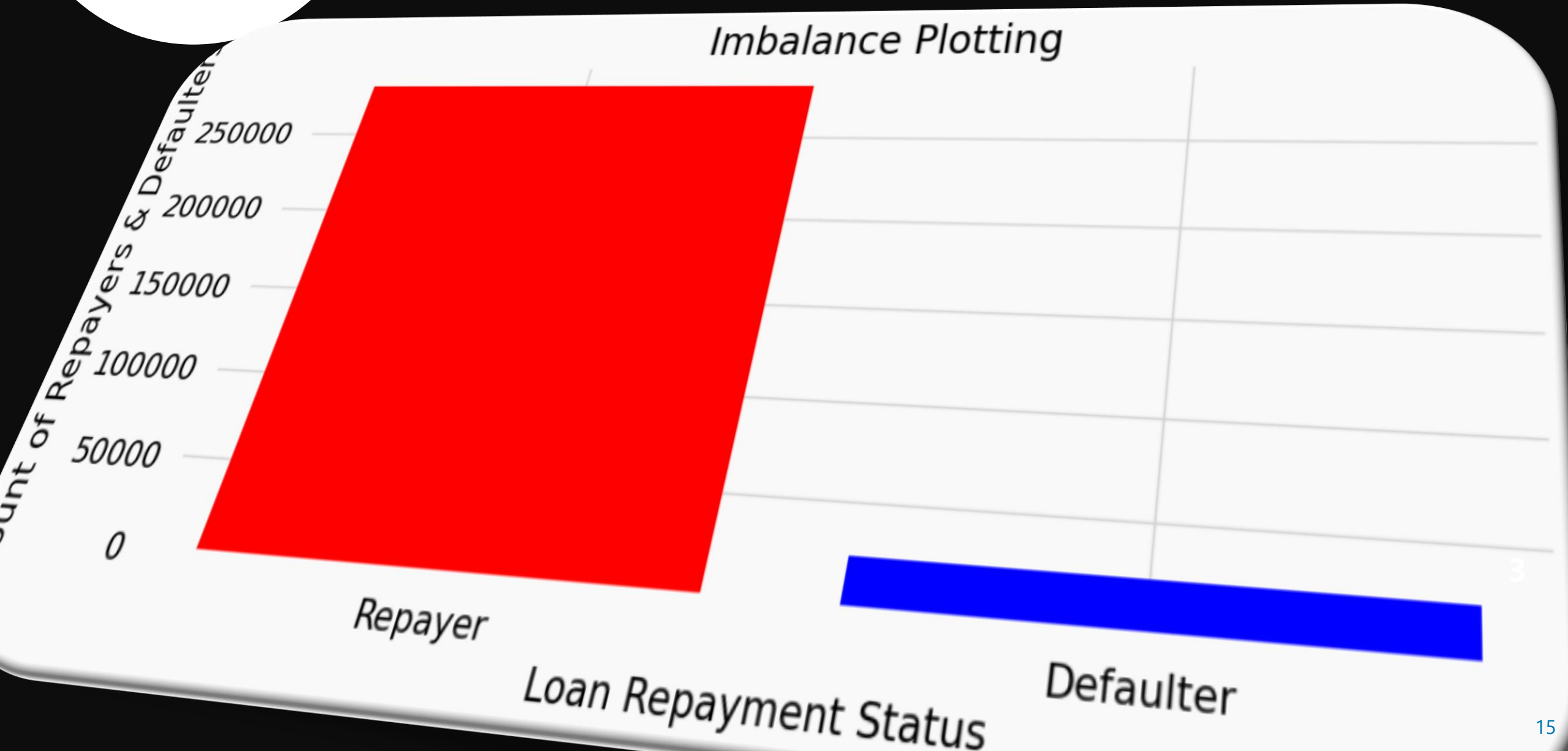
3 Data imbalance in Application_data dataset

- The objective of this analysis is to determine if there is data imbalance in the "Application_Data" dataset, specifically with regards to the binary classification of loan repayment status.
- A bar chart was plotted to visualize the distribution of loan repayment status in the dataset.
- The target variable is "TARGET", where 0 represents "Loan Repayer" and 1 represents "Defaulter".
- The bar chart displays the count of loan applicants categorized as "Loan Repayer" and "Defaulter". In the chat we can observe the significant difference in the number of instances between the two classes.
- The chart shows the count of "Loan Repayers" is significantly higher than the count of "Defaulters".
- Data imbalance can affect the accuracy of classification models. Models may perform well on predicting the majority class ("Loan Repayers") but struggle with the minority class ("Defaulters").
- To address data imbalance, consider using techniques like resampling (oversampling or undersampling) or employing algorithms that handle imbalanced data well.

3

Ratios of imbalance in % with respect to Repayer and Defaulter data are: 91.93 and 8.07

Ratios of imbalance in relative with respect to Repayer and Defaulter data are 11.39 : 1



4

Data Analysis

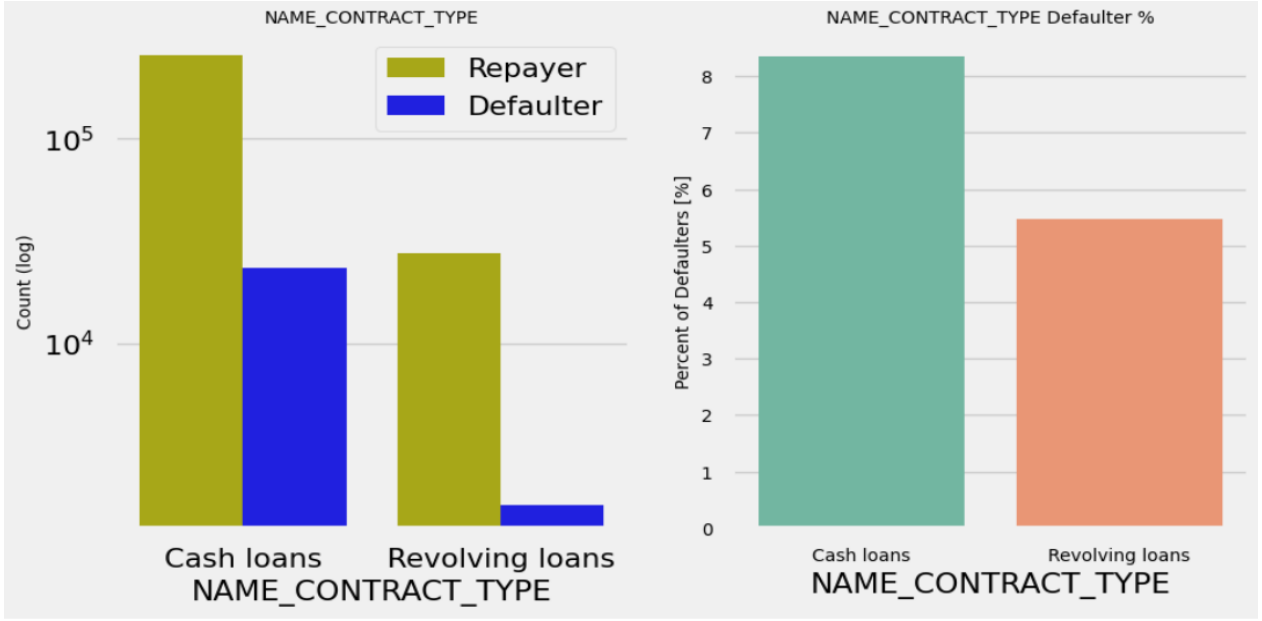
Univariate, Segmented Univariate, Bivariate Analysis Results

- Univariate analysis is used to examines one feature at a time.
- with It we can understand the distribution of a features, identify outliers, and explore relationships between a attribute and other attributes.
- Segmented univariate analysis is a type of univariate analysis ,it divides the data into groups based on a common feature.
- It can be used to analyze each group separately and identify differences in the between the attributes within each group.
- Bivariate analysis is used to examines two attributes at a time.
- It can be used to identify relationships between two attributes, such as one attribute causes or predicts the other

4 Univariate Analysis

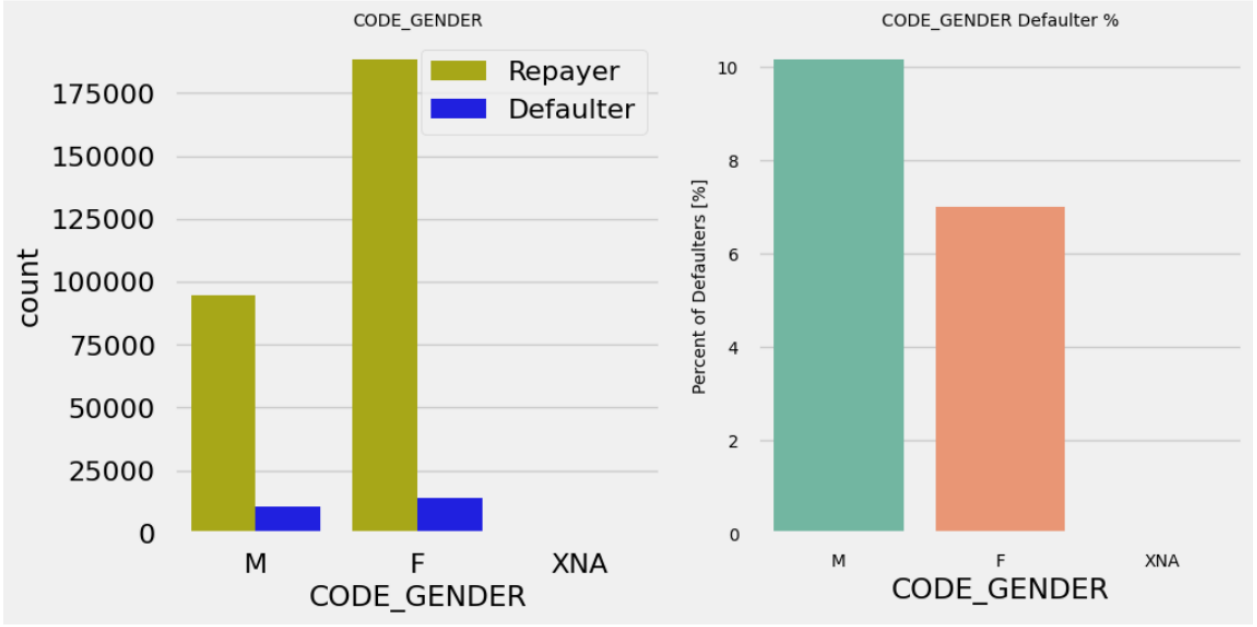
Checking the contract type based on loan repayment status

Contract type: Revolving loans are just a small fraction (10%) from the total number of loans; in the same time, a larger amount of Revolving loans, comparing with their frequency, are not repaid.



Checking the type of Gender based on loan repayment status

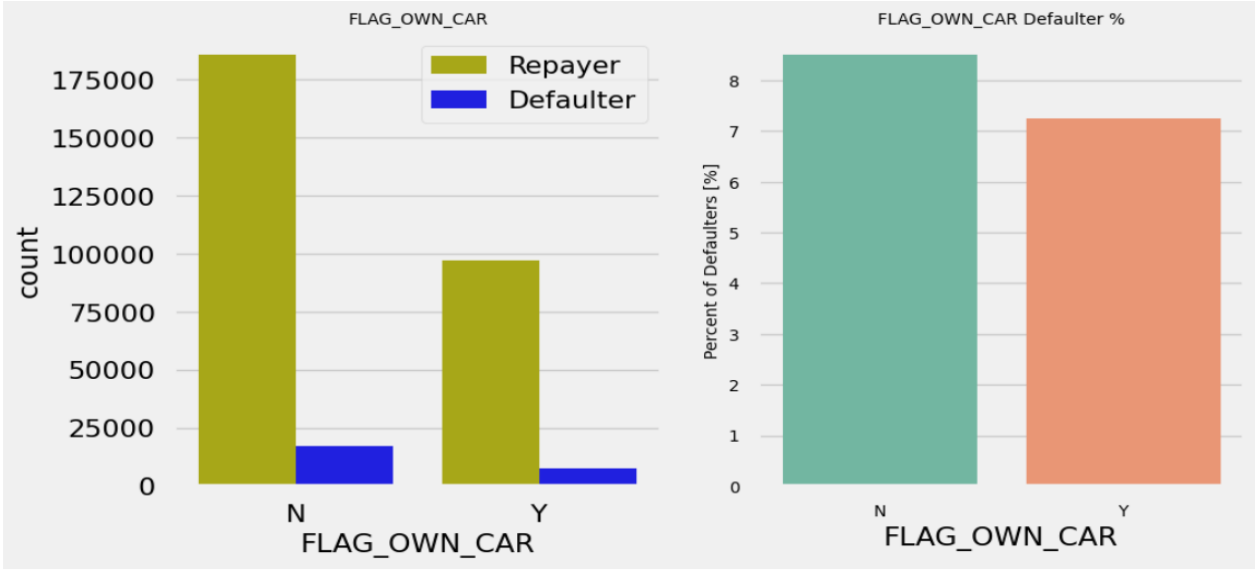
The number of female clients is almost double the number of male clients. Based on the percentage of defaulted credits, males have a higher chance of not returning their loans (~10%), comparing with women (7%)



4 Univariate Analysis

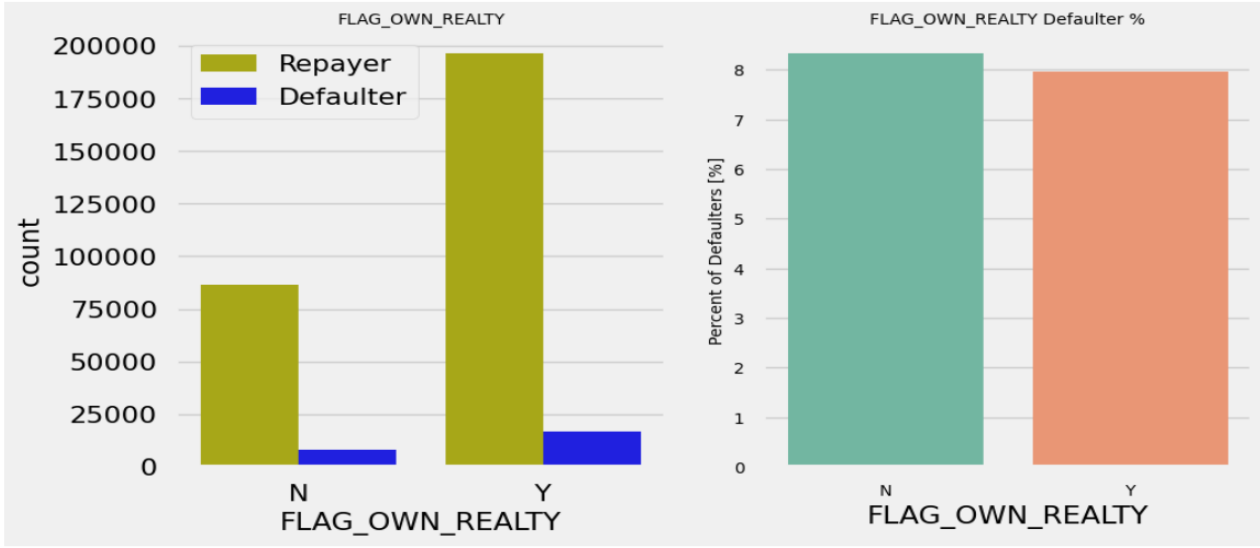
Checking if owning car is related to loan repayment status

Clients who own a car are half in number of the clients who don't own a car. But based on the percentage of default, there is no correlation between owning a car and loan repayment as in both cases the default percentage is almost same.



Checking if owning a reality is related to loan repayment status

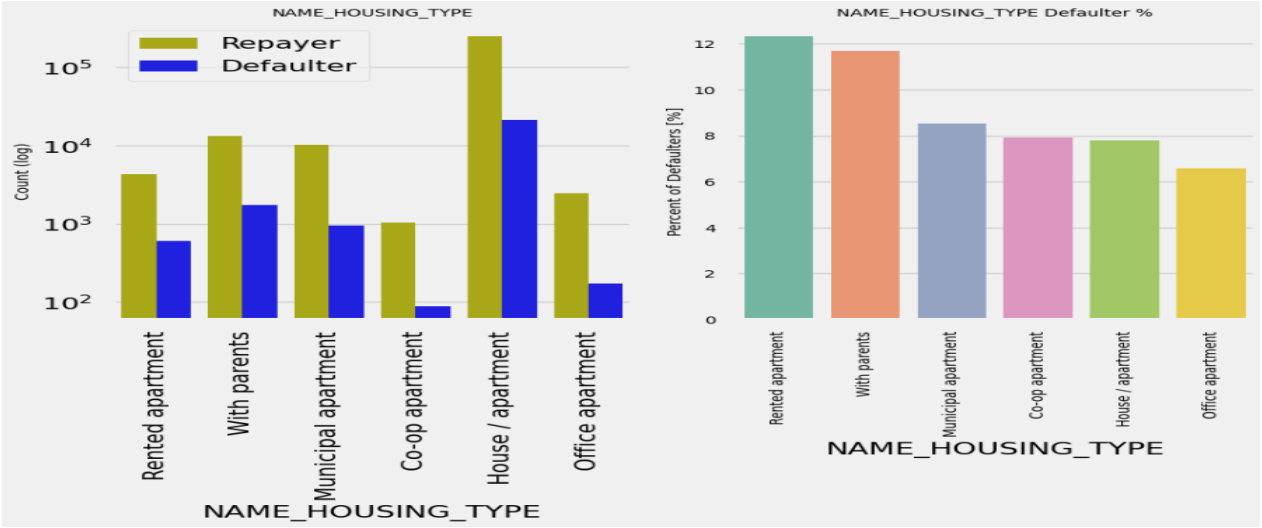
The clients who own real estate are more than double of the ones that don't own. But the defaulting rate of both categories are around the same (~8%). Thus there is no correlation between owning a reality and defaulting the loan



4 Univariate Analysis

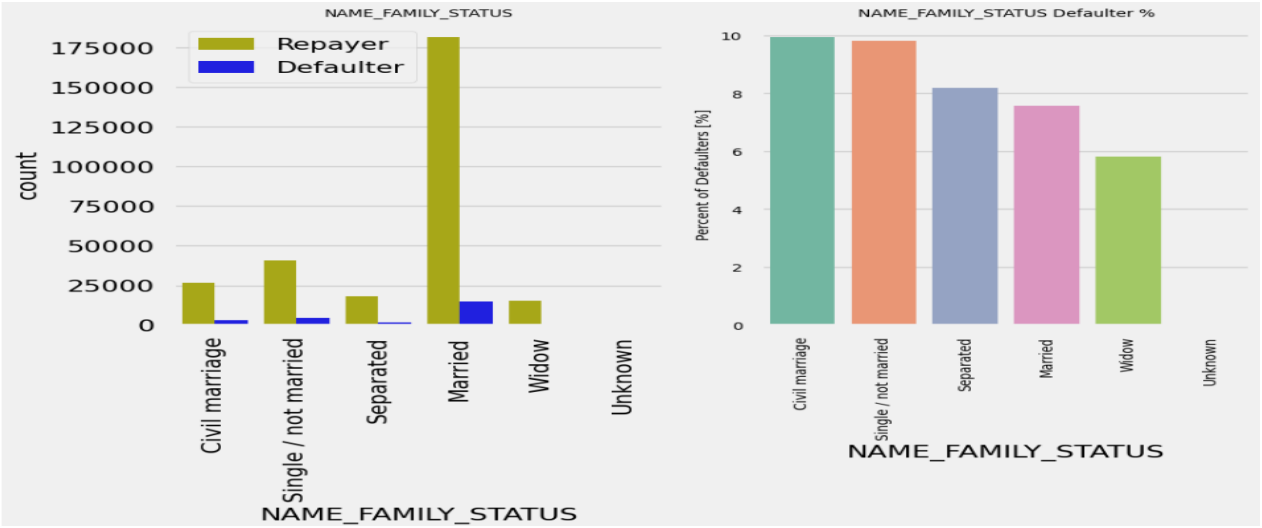
Analyzing Housing Type based on loan repayment status

Majority of people live in House/apartment. People living in office apartments have lowest default rate. People living with parents (~11.5%) and living in rented apartments(>12%) have higher probability of defaulting.



Analyzing Family status based on loan repayment status

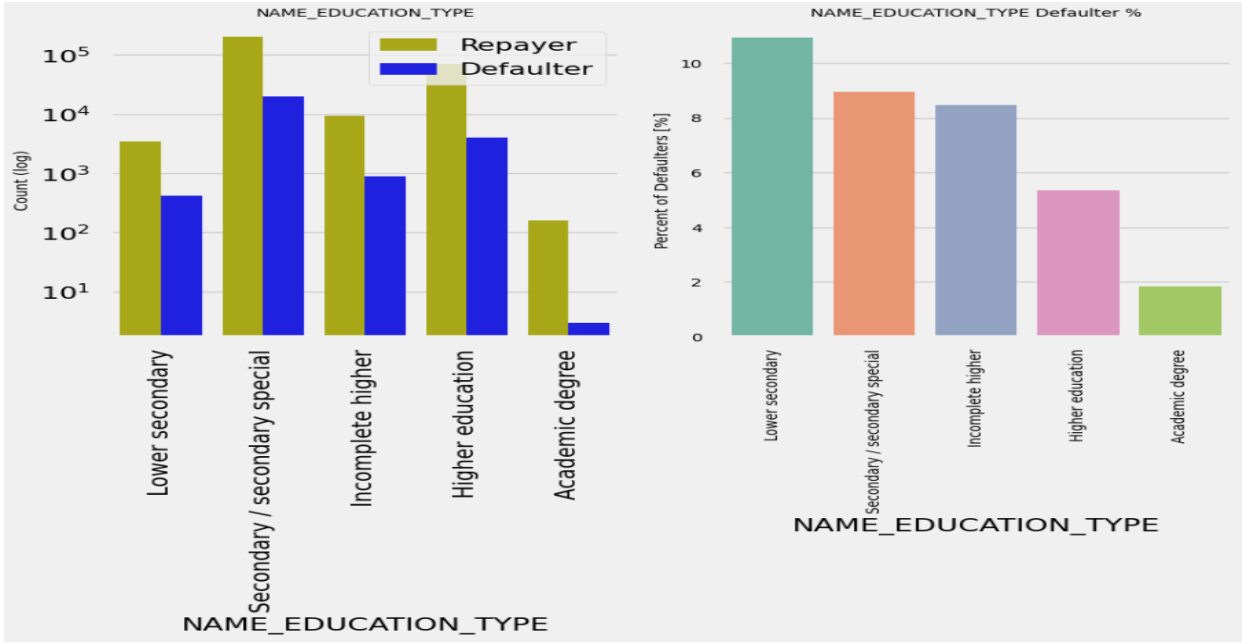
Most of the people who have taken loan are married, followed by Single/not married and civil marriage. In terms of percentage of not repayment of loan, Civil marriage has the highest percent of not repayment (10%), with Widow the lowest (exception being Unknown).



4 Univariate Analysis

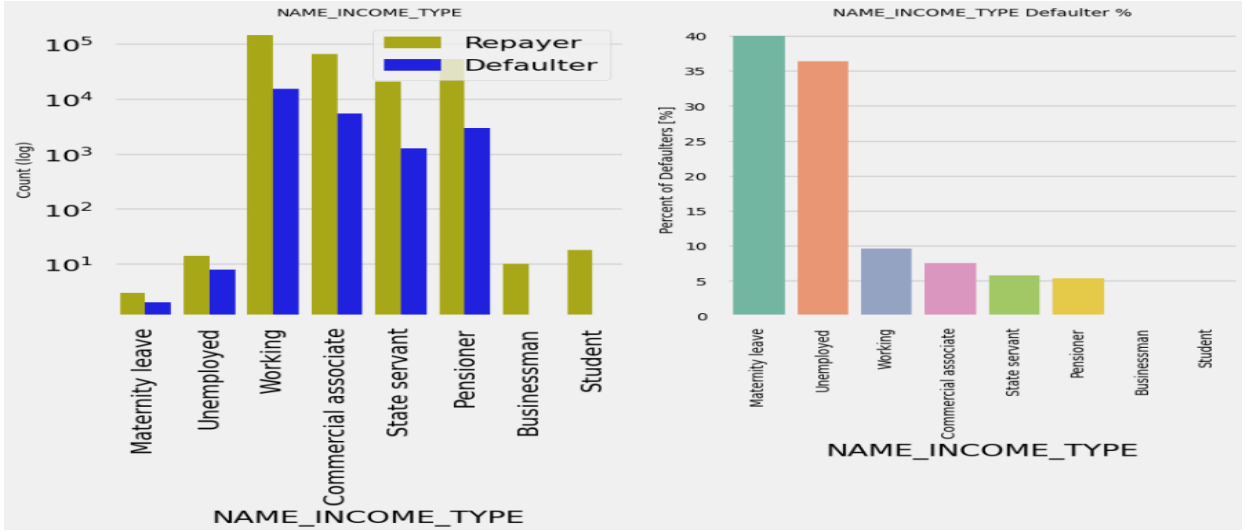
Analyzing Education Type based on loan repayment status

Majority of the clients have Secondary / secondary special education, followed by clients with Higher education. Only a very small number having an academic degree. The Lower secondary category, although rare, have the largest rate of not returning the loan (11%). The people with Academic degree have less than 2% rate of not returning the loan.



Analyzing Income Type based on loan repayment status

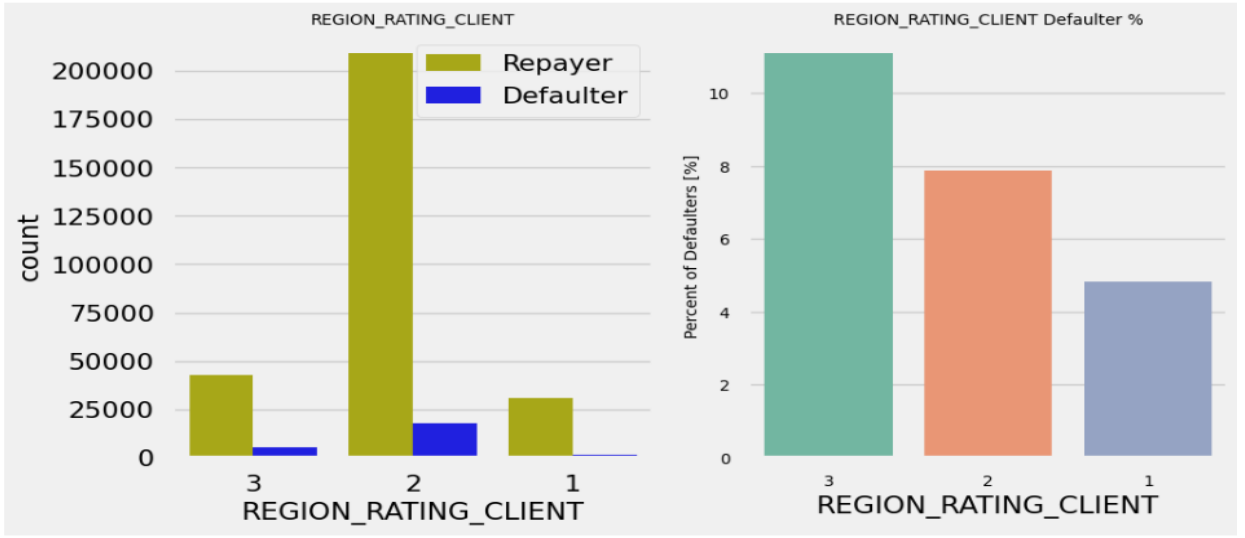
Most of applicants for loans have income type as Working, followed by Commercial associate, Pensioner and State servant. The applicants with the type of income Maternity leave have almost 40% ratio of not returning loans, followed by Unemployed (37%). The rest of types of incomes are under the average of 10% for not returning loans. Student and Businessmen, though less in numbers do not have any default record. Thus these two category are safest for providing loan



4 Univariate Analysis

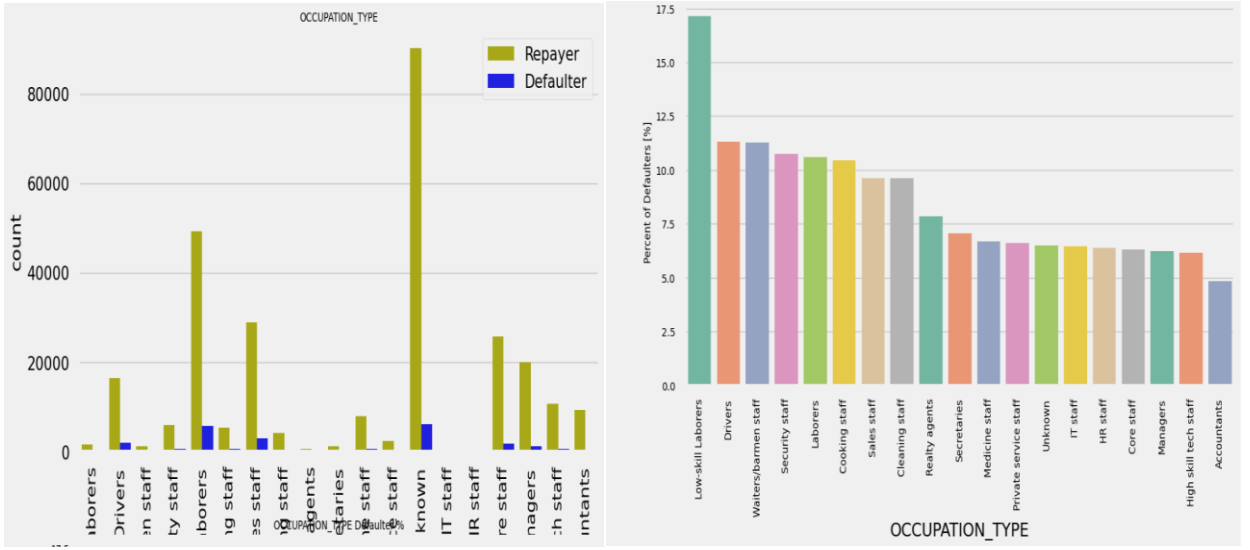
Analyzing Region rating where applicant lives based on loan repayment status

Most of the applicants are living in Region_Rating 2 place. Region Rating 3 has the highest default rate (11%). Applicant living in Region_Rating 1 has the lowest probability of defaulting, thus safer for approving loans



Analyzing Occupation Type where applicant lives based on loan repayment status

Most of the loans are taken by Laborers, followed by Sales staff. IT staff take the lowest amount of loans. The category with highest percent of not repaid loans are Low-skill Laborers (above 17%), followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff.



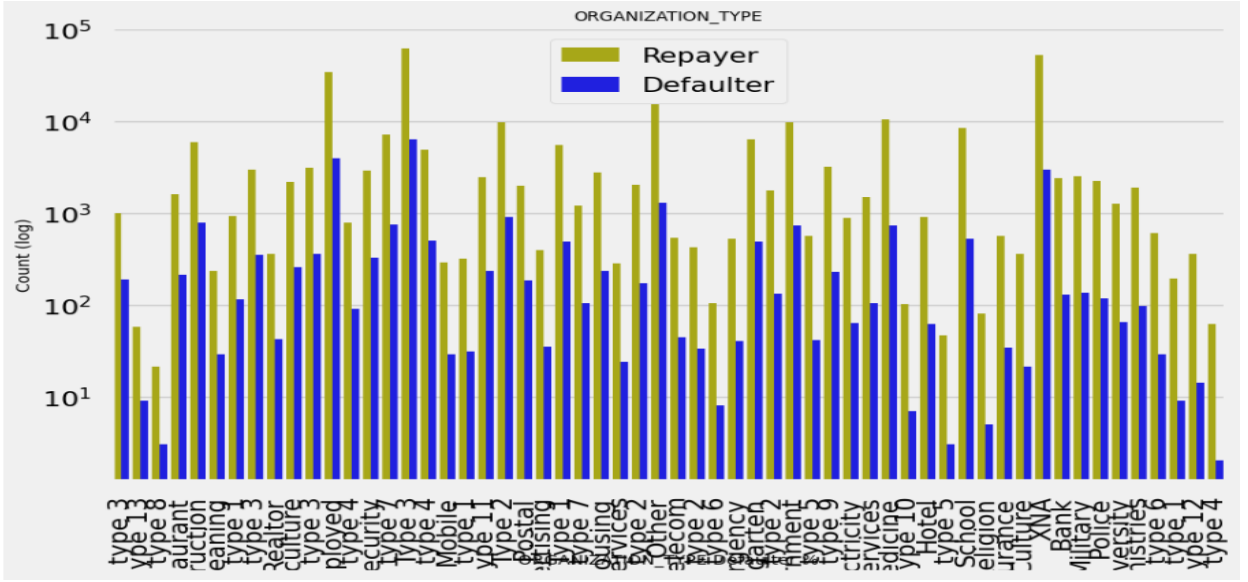
4 Univariate Analysis

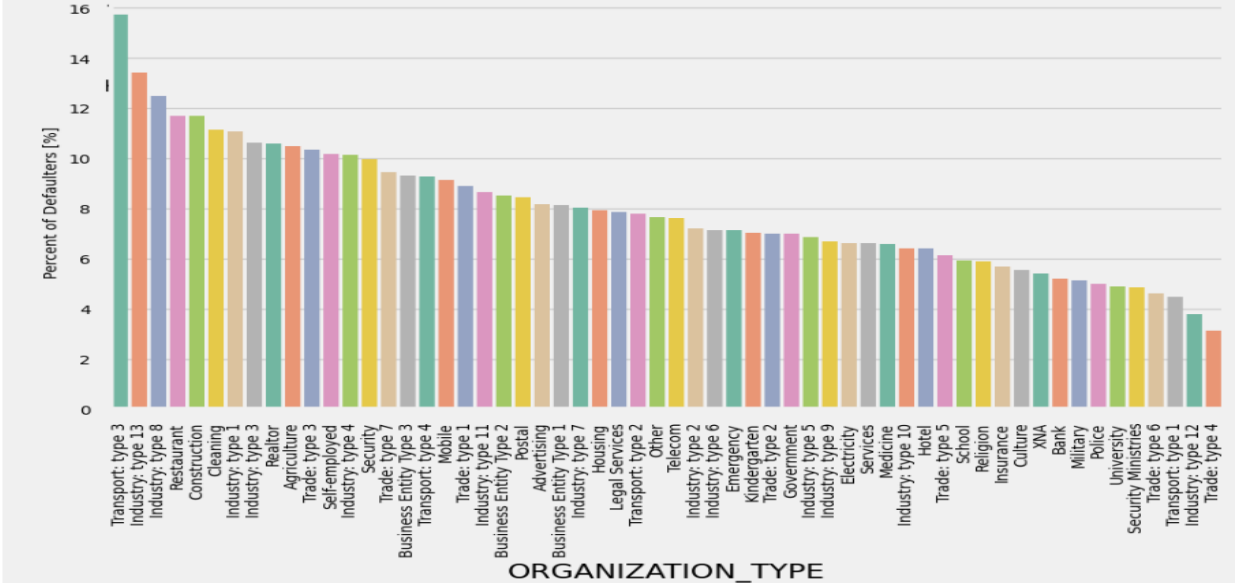
Checking Loan repayment status based on Organization type

- Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%).

- Self employed people have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting. Most of the people application for loan are from Business Entity Type 3.

- It can be seen that category of Trade Type 4 ,Industry type 12 organization types has lesser defaulters thus safer for providing loans

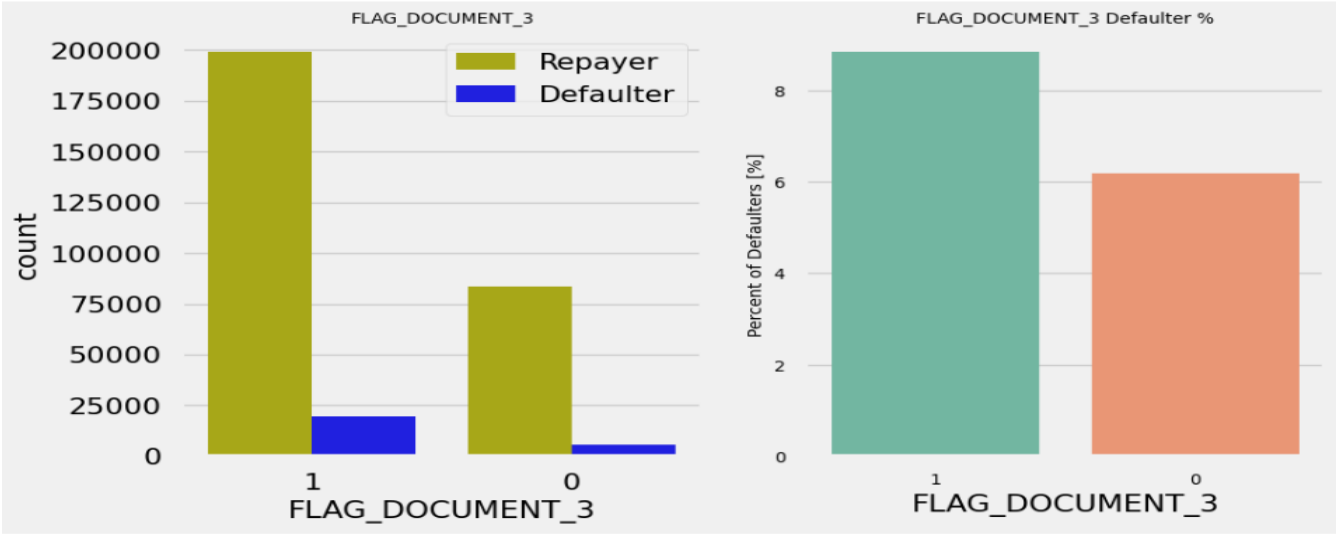




4 Univariate Analysis

Analyzing Flag_Doc_3 submission status based on loan repayment status

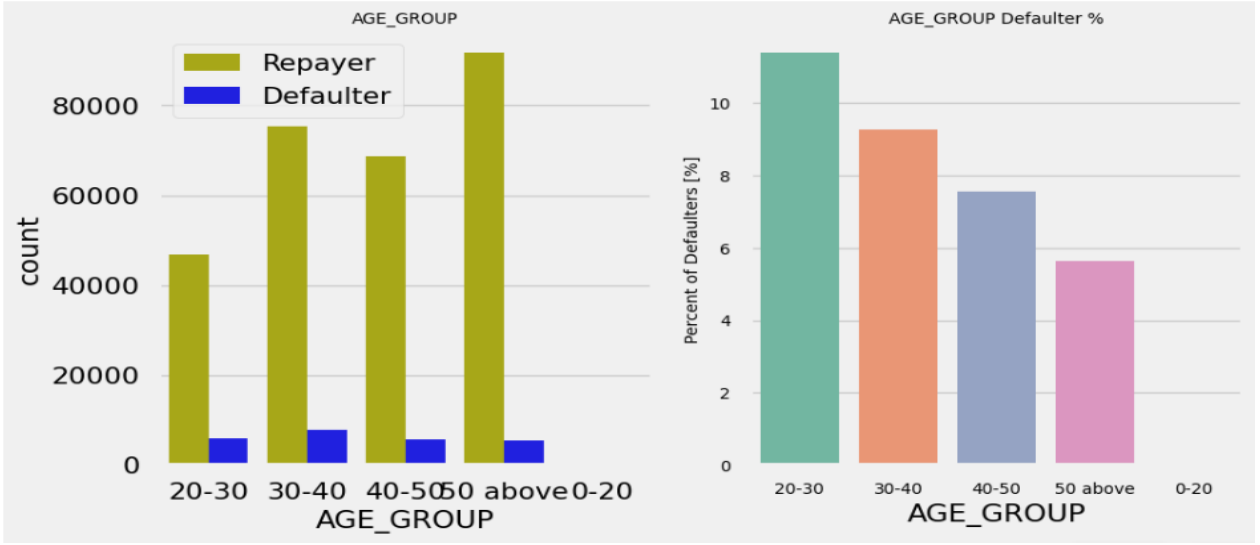
There is no significant correlation between repayers and defaulters in terms of submitting document 3 as we see even if applicants have submitted the document, they have defaulted a slightly more (~9%) than who have not submitted the document (6%)



4 Segmented Univariate Analysis

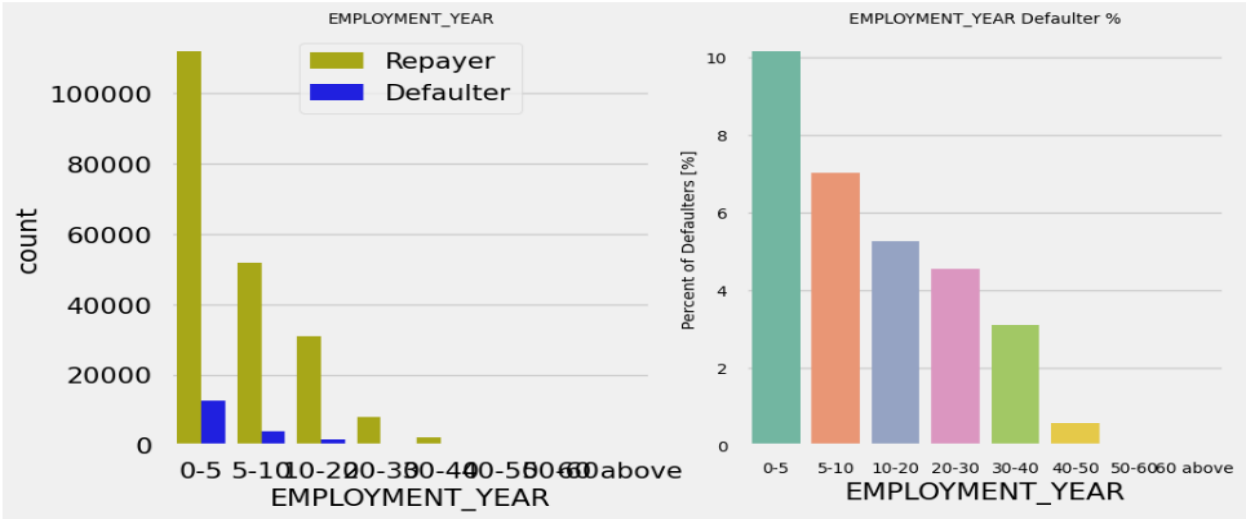
Analyzing Age Group based on loan repayment status

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



Analyzing Employment_Year based on loan repayment status

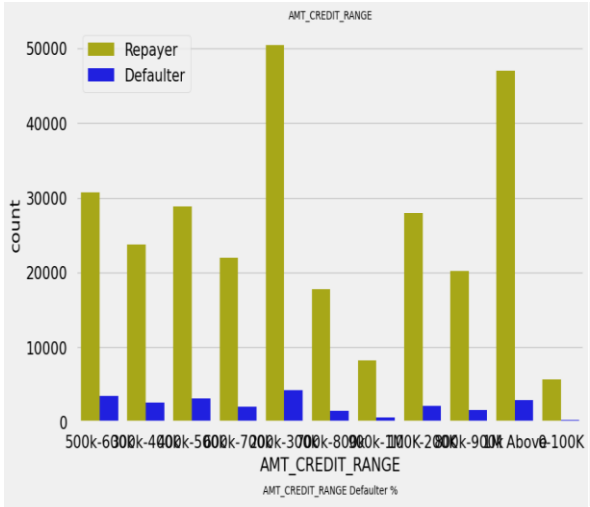
Majority of the applicants have been employeed in between 0-5 years. The defaulting rating of this group is also the highest which is 10%. With increase of employment year, defaulting rate is gradually decreasing with people having 40+ year experience having less than 1% default rate

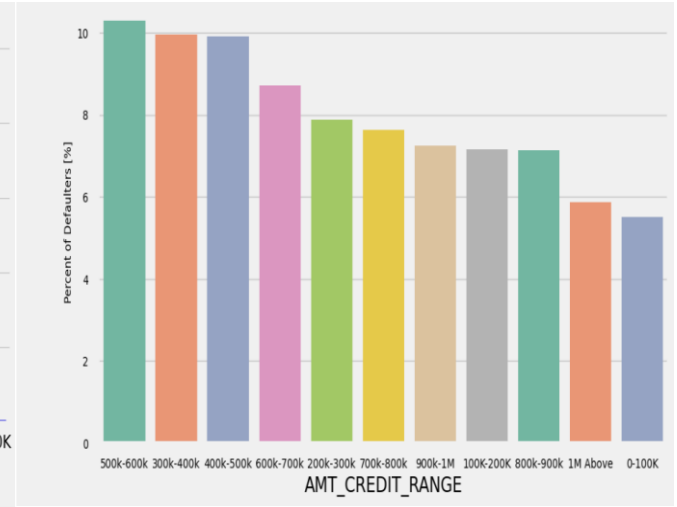


4 Segmented Univariate Analysis

Analyzing Amount_Credit based on loan repayment status

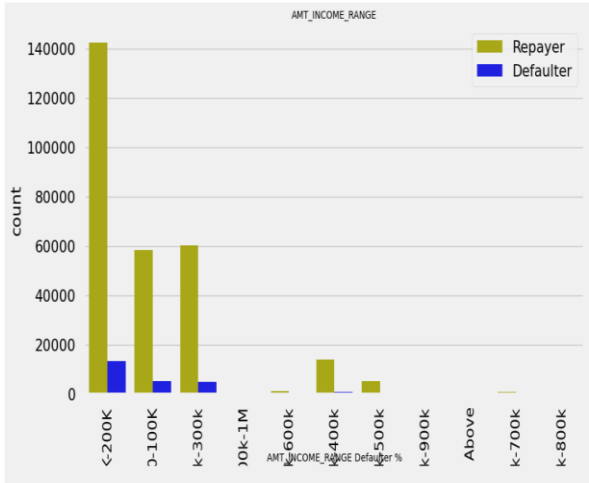
More than 80% of the loan provided are for amount less than 900,000. People who get loan for 300-600k tend to default more than others.

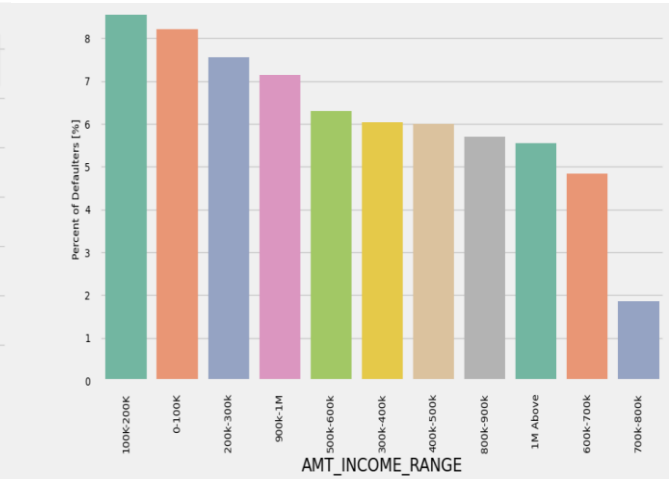




Analyzing Amount_Income Range based on loan repayment status

90% of the applications have Income total less than 300,000 Application with Income less than 300,000 has high probability of defaulting 3. Applicant with Income more than 700,000 are less likely to default

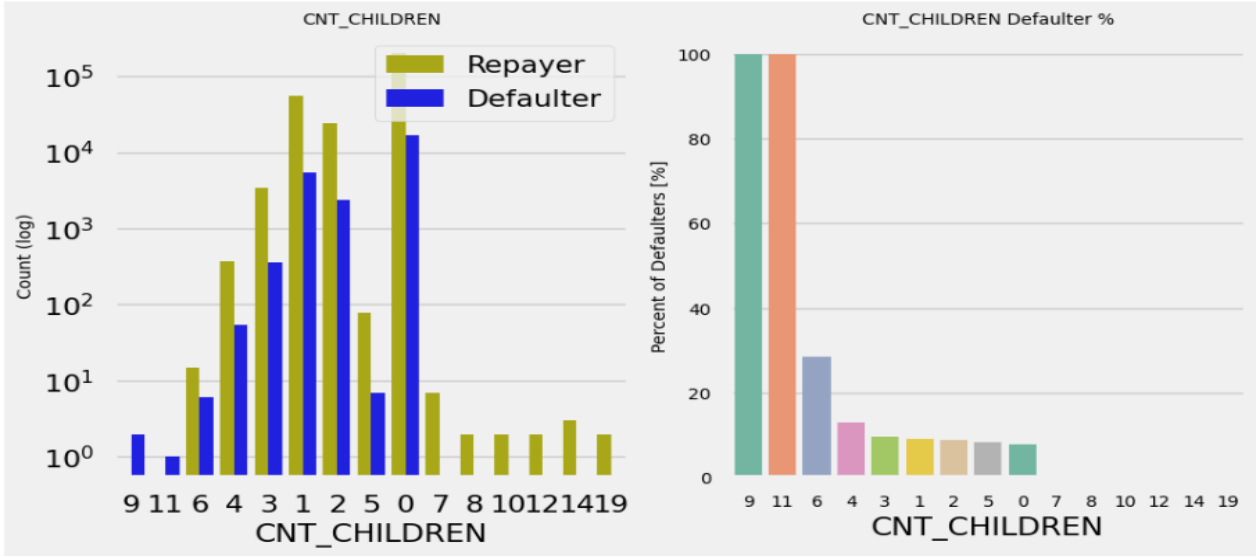




4 Segmented Univariate Analysis

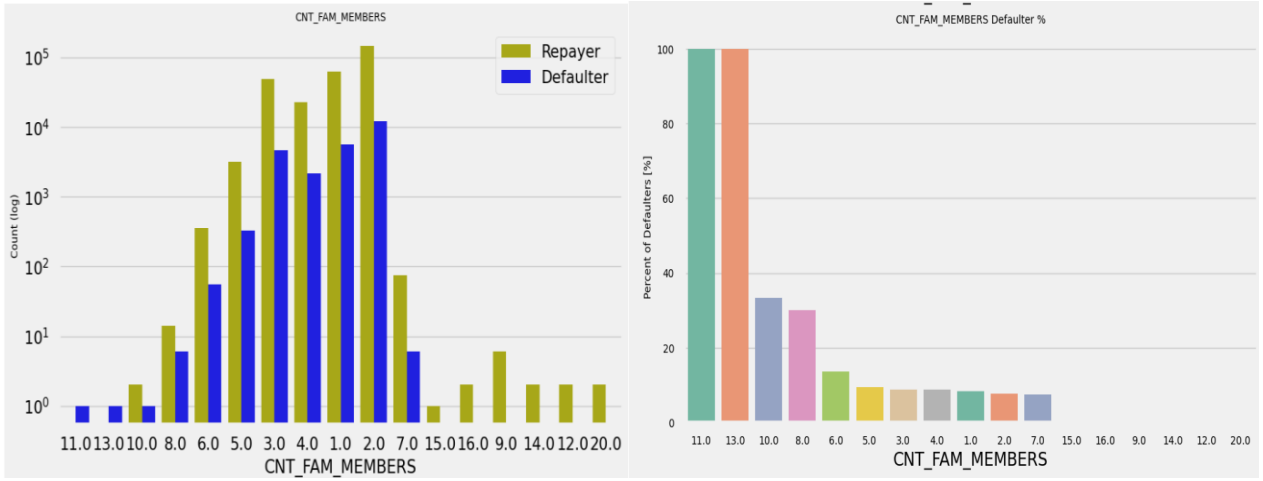
Analyzing Number of children based on loan repayment status

Most of the applicants do not have children. Very few clients have more than 3 children. Client who have more than 4 children has a very high default rate with child count 9 and 11 showing 100% default rate.



Analyzing Number of family members based on loan repayment status

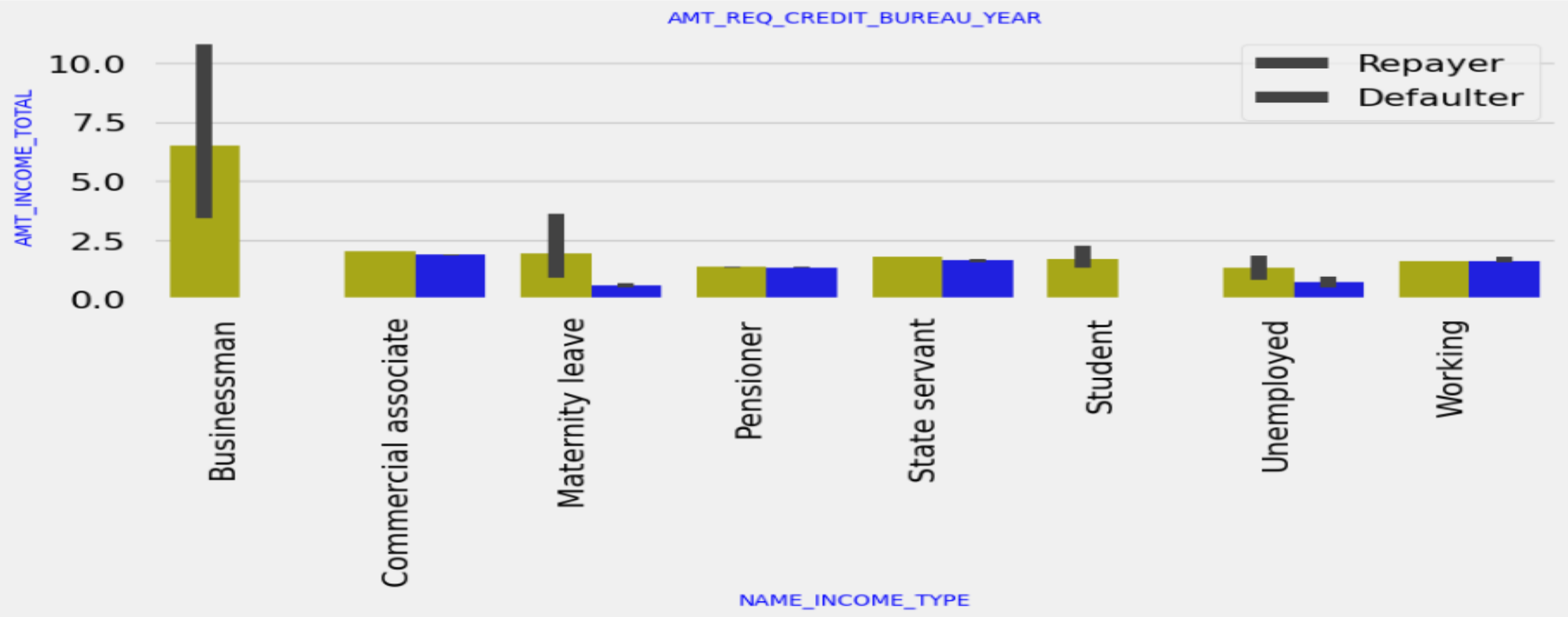
Family member follows the same trend as children where having more family members increases the risk of defaulting



4 Bivariate Analysis

Income type vs Income Amount Range

It can be seen that business man's income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a business man could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs.



Correlation

5

Correlation analysis is a statistical technique used to measure the strength and direction of the relationship between two variables. In the bank loan case study dataset, correlation analysis was performed to understand the relationships between different numerical variables and the loan repayment status (TARGET).

The results of the correlation analysis :

AMT_CREDIT: The amount of the loan.

AMT_ANNUITY: The monthly annuity payment for the loan.

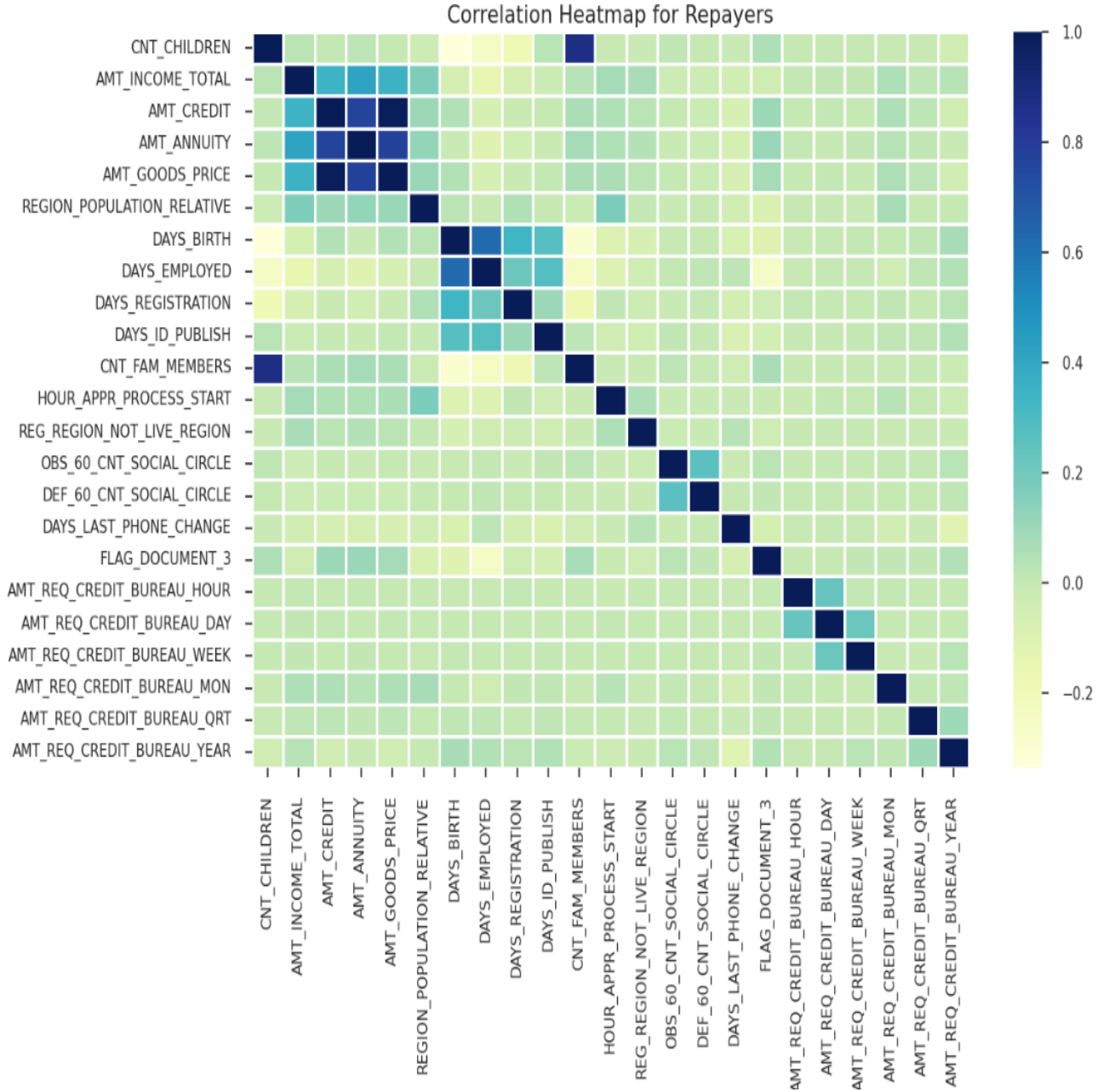
AMT_GOODS_PRICE: The price of the goods or services that the loan was used to purchase.

CNT_CHILDREN: The number of children in the borrower's household.

DAYS_EMPLOYED: The number of days that the borrower has been employed.

INCOME_TYPE: The type of income that the borrower receives.

The correlation coefficients for these variables ranged from 0.12 to 0.48, which indicates that there is a moderate to strong correlation between these variables and the loan repayment status.

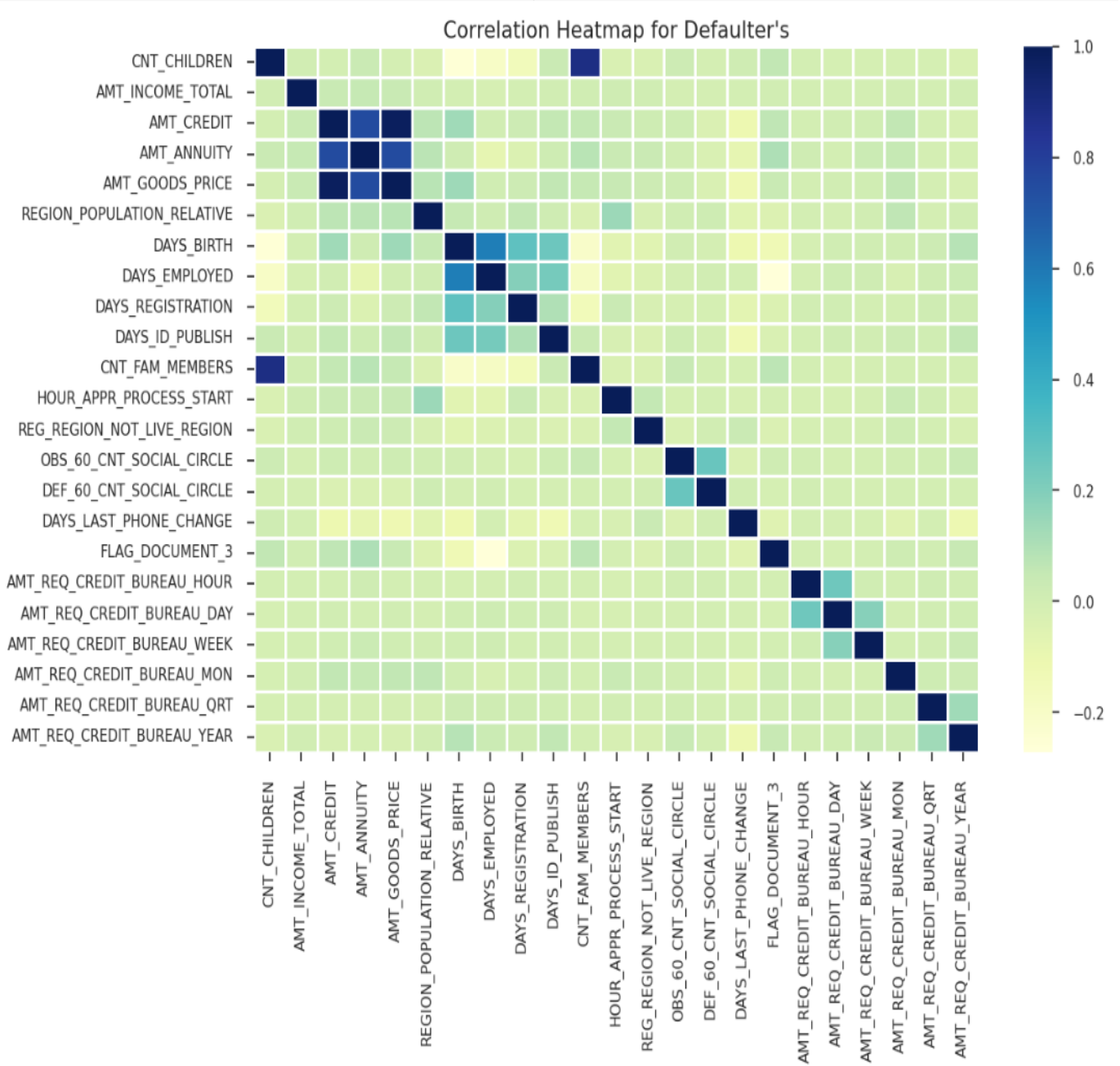


5

the top 10 correlation for the Repayers data

	VAR1	VAR2	Correlation
94	AMT_GOODS_PRICE	AMT_CREDIT	0.987250
230	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
95	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
71	AMT_ANNUITY	AMT_CREDIT	0.771309
167	DAYS_EMPLOYED	DAYS_BIRTH	0.626114
70	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953
93	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349462
47	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799
138	DAYS_BIRTH	CNT_CHILDREN	0.336966
190	DAYS_REGISTRATION	DAYS_BIRTH	0.333151

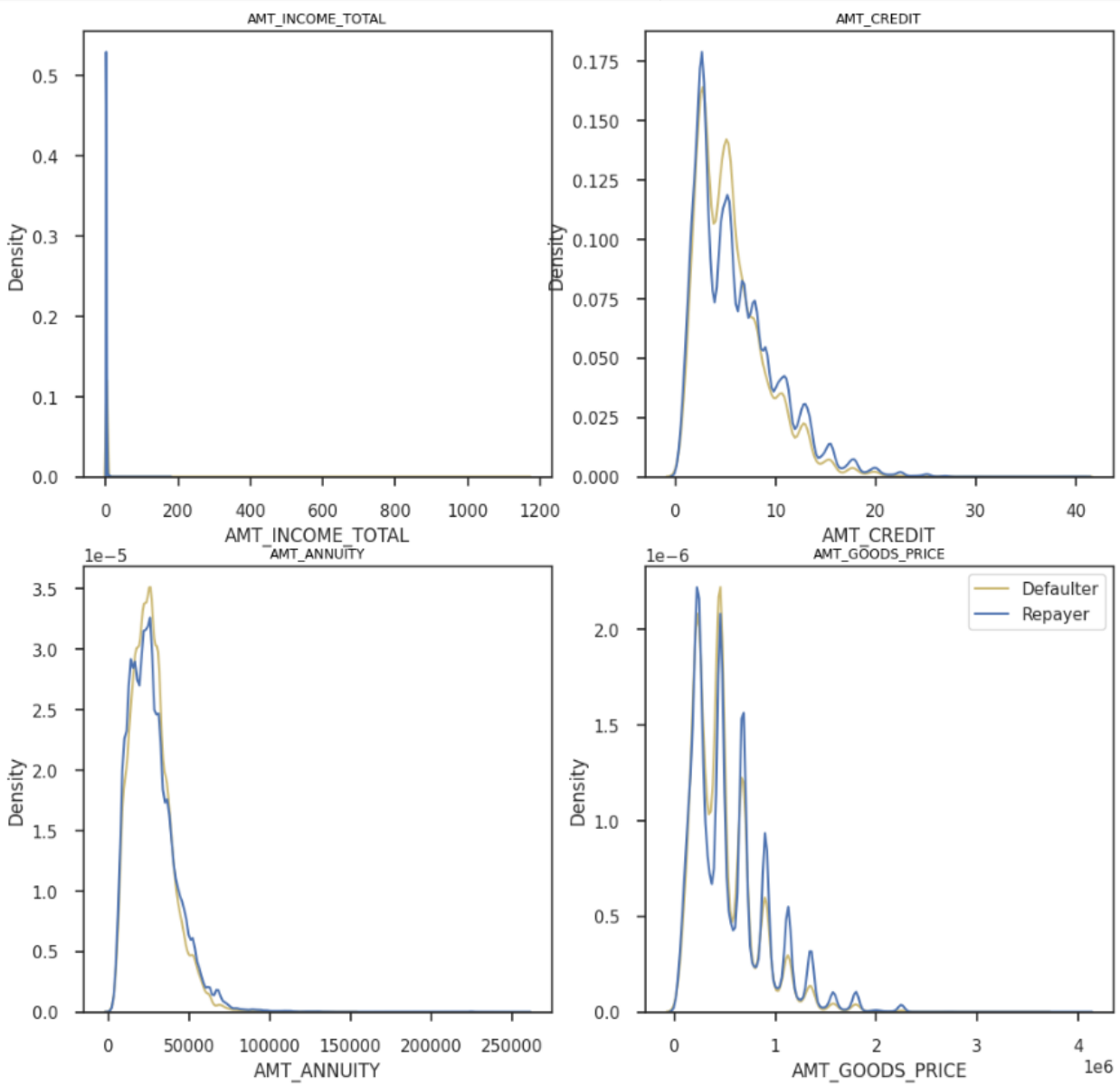
Credit amount is highly correlated with amount of goods price loan annuity total income. We can also see that repayers have high correlation in number of days employed.



5

the top 10 correlation for the Defaulters data

	VAR1	VAR2	Correlation
94	AMT_GOODS_PRICE	AMT_CREDIT	0.983103
230	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
95	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
71	AMT_ANNUITY	AMT_CREDIT	0.752195
167	DAYS_EMPLOYED	DAYS_BIRTH	0.582185
190	DAYS_REGISTRATION	DAYS_BIRTH	0.289114
375	FLAG_DOCUMENT_3	DAYS_EMPLOYED	0.272169
335	DEF_60_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.264159
138	DAYS_BIRTH	CNT_CHILDREN	0.259109
213	DAYS_ID_PUBLISH	DAYS_BIRTH	0.252863

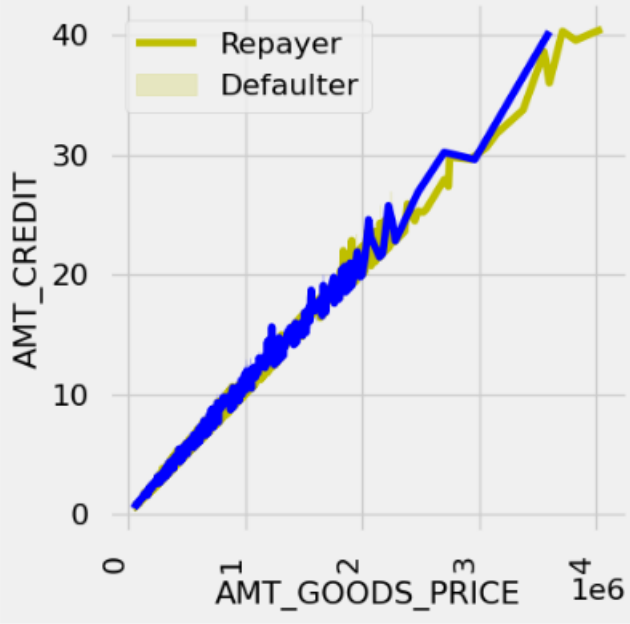


5

The Distribution of Loan Amounts by Defaulter and Repayer

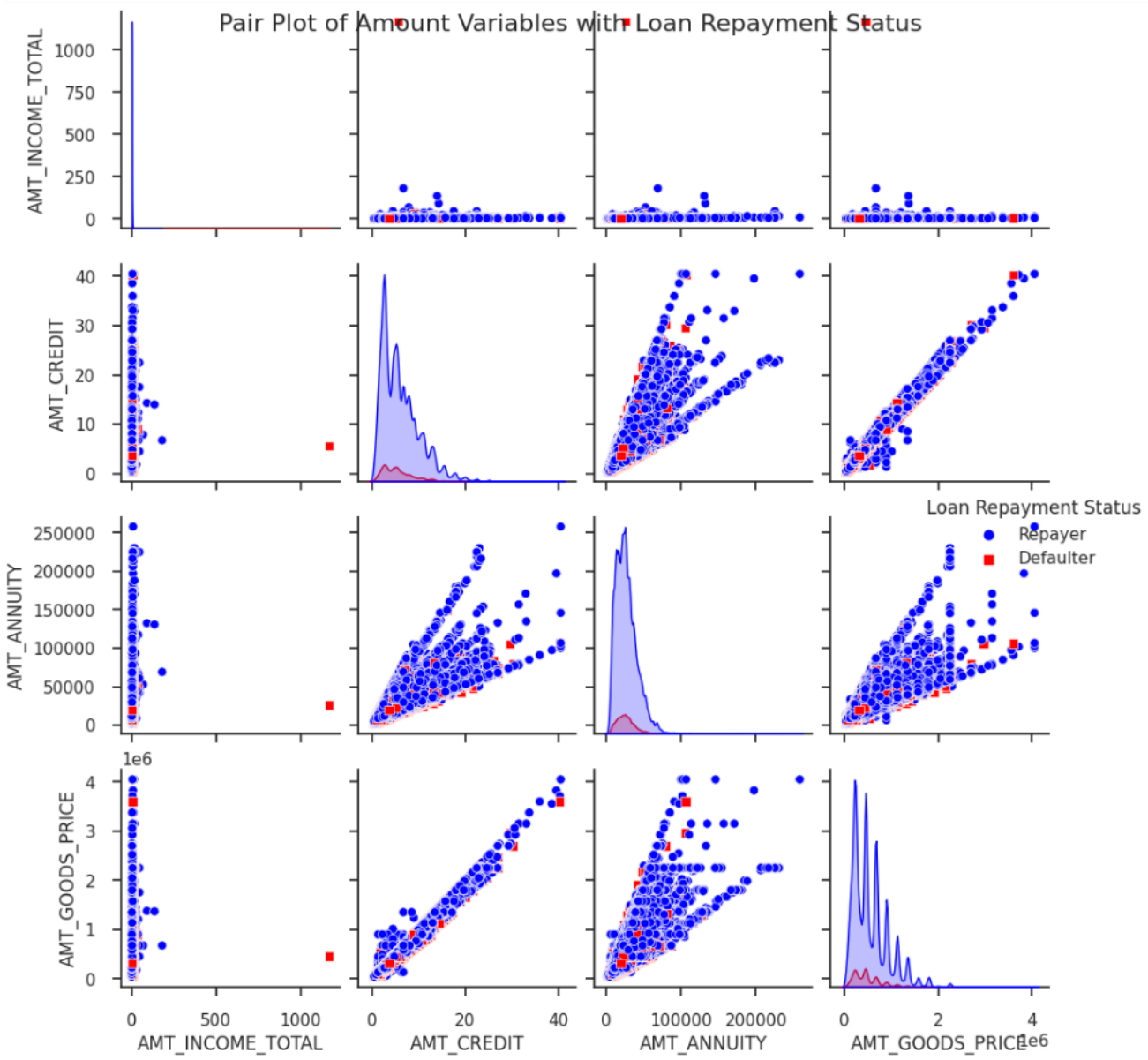
Most no of loans are given for goods price below 10 lakhs. Most people pay annuity below 50000 for the credit loan. Credit amount of the loan is mostly less then 10 lakhs. The repayers and defaulters distribution overlap in all the plots and hence we cannot use any of these variables in isolation to make a decision

Checking the relationship between Goods price and credit and comparing with loan repayment status

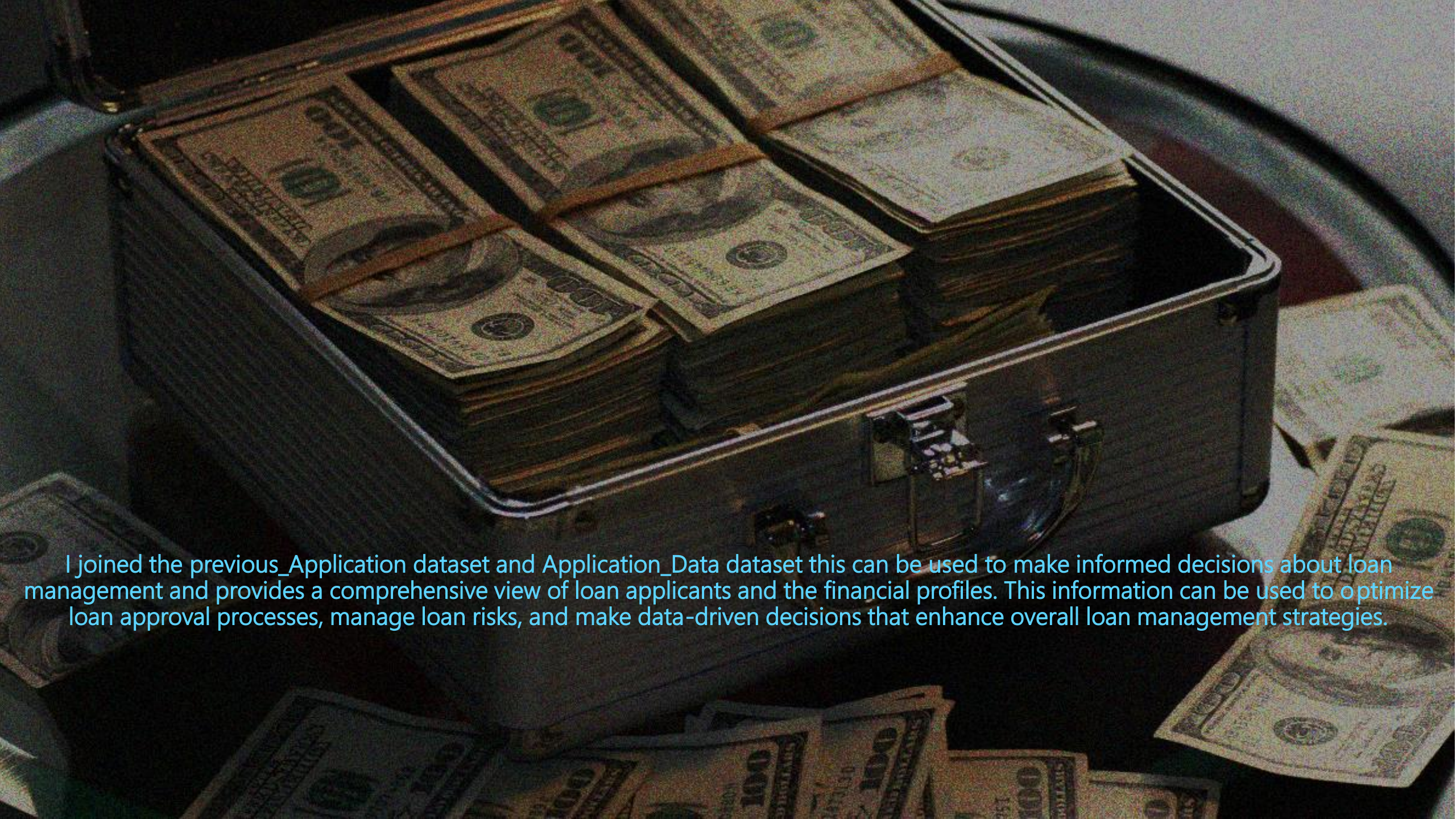


Here, we can see increase in defaulter, when credit amount goes beyond 3M

Plotted a pairplot between amount variable to draw reference against loan repayment status



- Here, we can see increase in defaulter, when amt_annuity >15000 and amt_goods_price> 3M, there is a lesser chance of defaulters.
- AMT_CREDIT and AMT_GOODS_PRICE are highly correlated as based on the scatterplot where most of the data are consolidated in form of a line.
- There are very less defaulters for AMT_CREDIT >3M.
- Inferences related to distribution plot has been already mentioned in previous distplot graphs inferences section



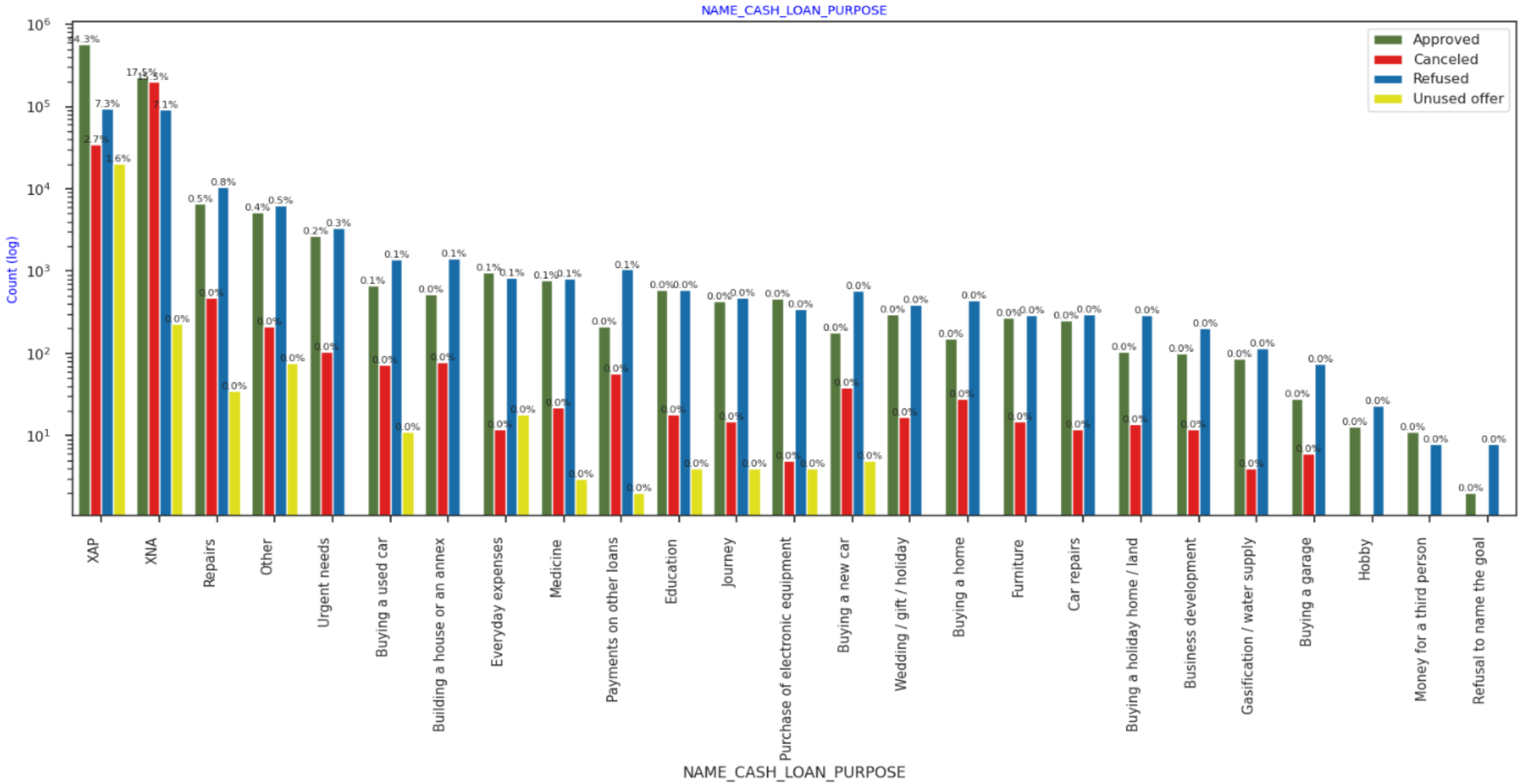
I joined the previous_Application dataset and Application_Data dataset this can be used to make informed decisions about loan management and provides a comprehensive view of loan applicants and the financial profiles. This information can be used to optimize loan approval processes, manage loan risks, and make data-driven decisions that enhance overall loan management strategies.

Merged Dataset

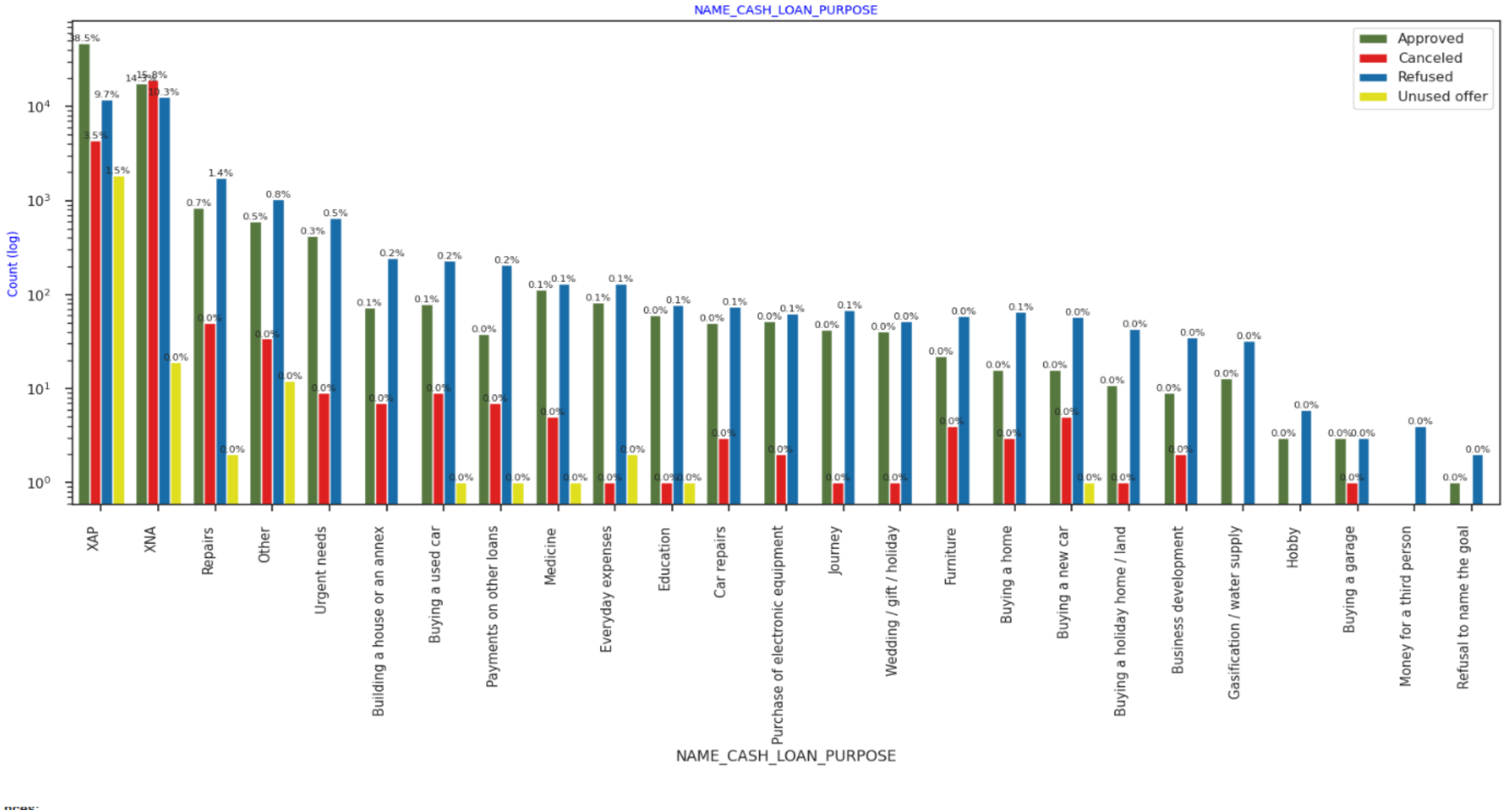
Using the Merged Bank Loan Dataset to Make Informed Decisions

- Based on SK_ID_CURR both the datasets are joined using inner join
- Total rows and columns are 14,13,701 and 74 respectively
- There are total of 3types of datatypes namely category(37),float(23) and int64(14)

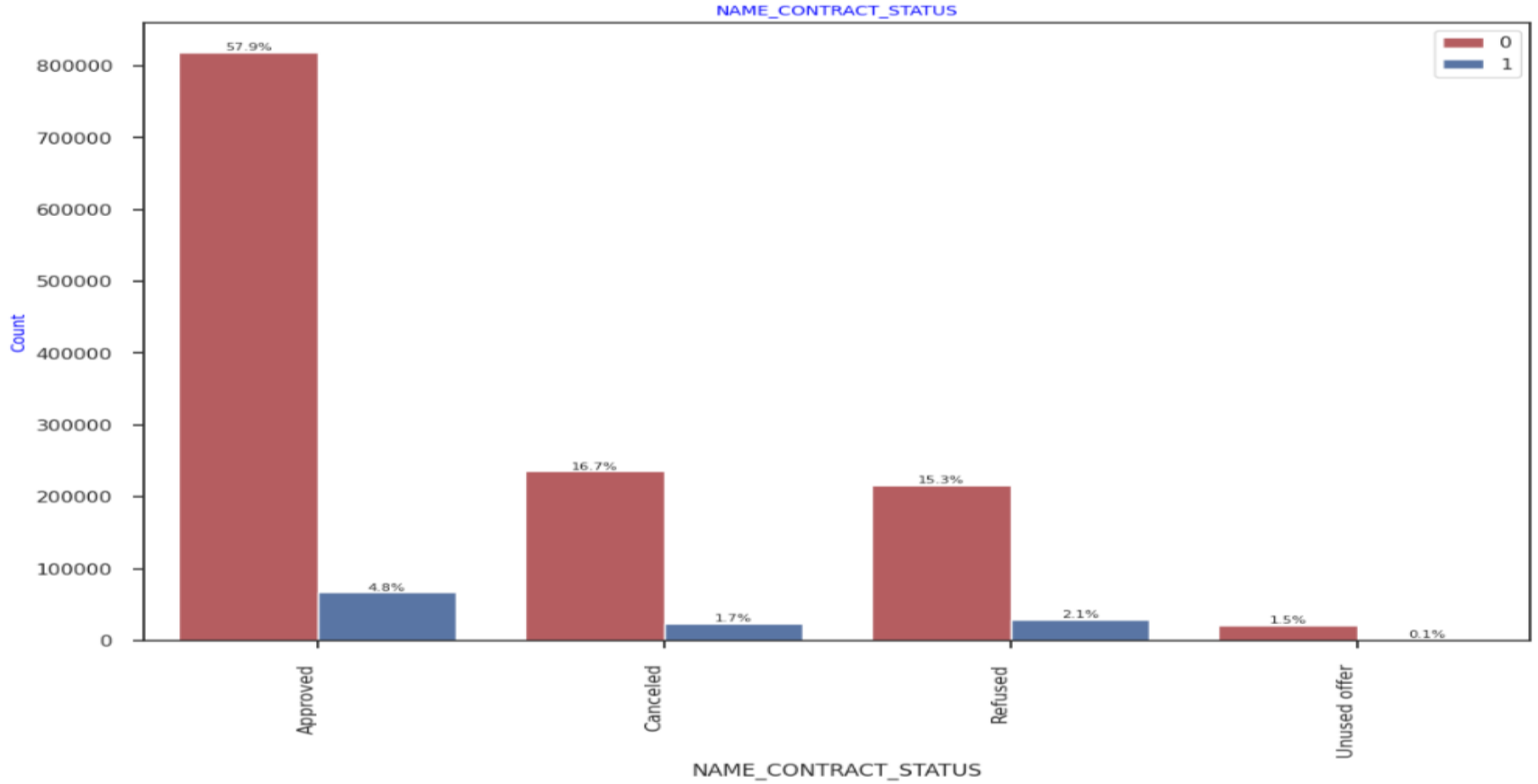
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1413701 entries, 0 to 1413700
Columns: 74 entries, SK_ID_CURR to DAYS_DECISION_GROUP
dtypes: category(37), float64(23), int64(14)
memory usage: 459.8 MB
```



This bar graph shows the why the repayer is taking the loan. We can see the four categories that are approved, cancelled, unused, and refused that applied for repayers.



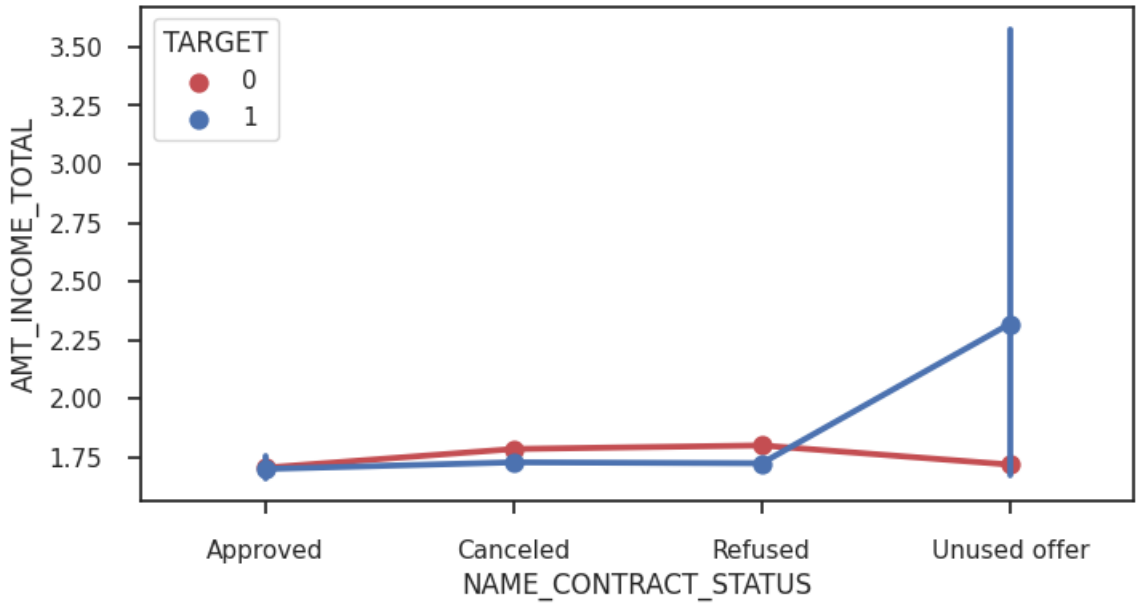
- Loan taken for the purpose of Repairs seems to have highest default rate
- A very high number application have been rejected by bank or refused by client which has purpose as repair or other.
- This shows that purpose repair is taken as high risk by bank and either they are rejected or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the 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%

90% of the previously cancelled client have actually repayed the loan. Revisiting the interest rates would increase business opportunity for these clients. 88% of the clients who have been previously refused a loan has payed back the loan in current case. Refusal reason should be recorded for further analysis as these clients would turn into potential repaying customer

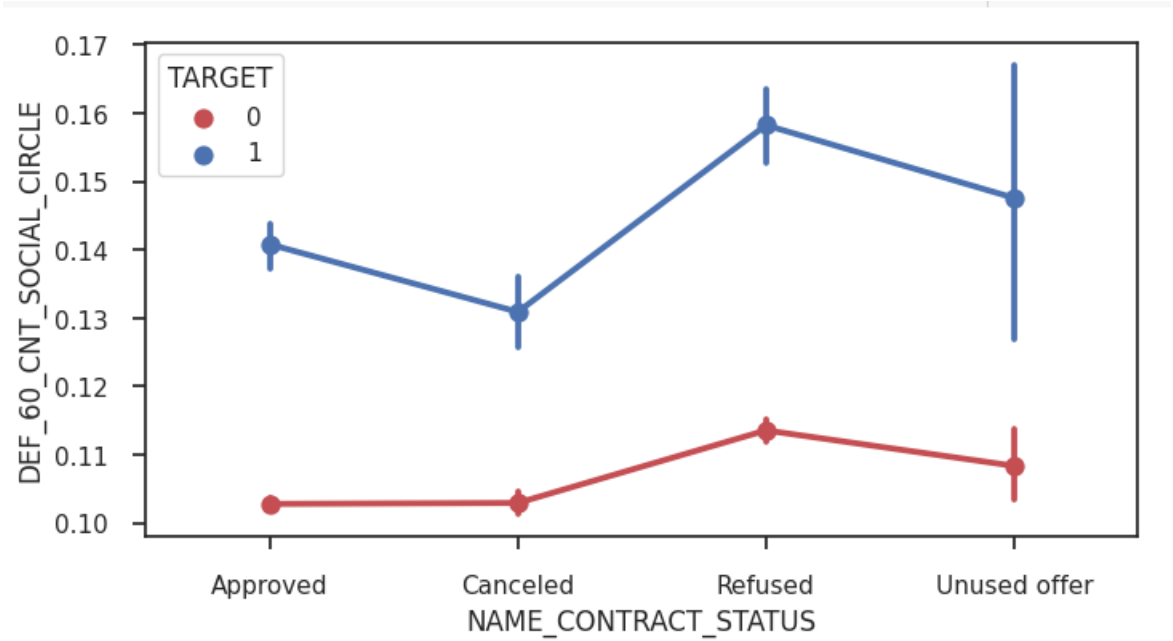
Income total vs Contract status



The data show that even though people who have not used the offer before have higher average incomes, they are still more likely to default

NOTE :- Workspace link

People VS Defaulted in last 60 days



We can observe that avg > 0.13 DEF_60_CNT_SOCIAL_CIRCLE tend to defaulter , so before approving loan person social circle need to be checked

LINK TO WORKSPACE
https://drive.google.com/file/d/14DAG4Z9_kaVoGthsTQnDpQ93YDKsjgni/view?usp=sharing

THANK YOU

END OF
PRESENTATION