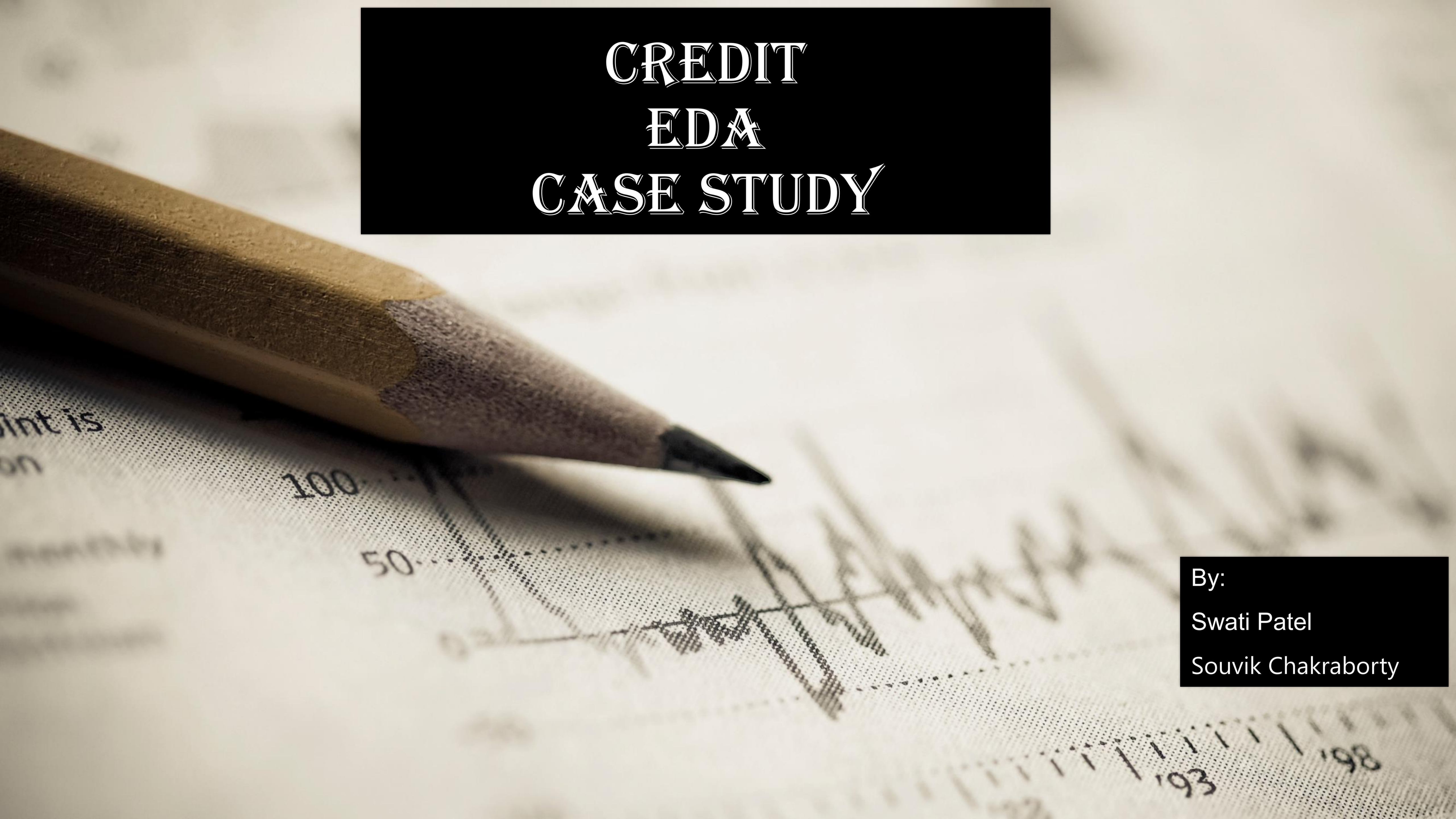


CREDIT
EDA
CASE STUDY

By:

Swati Patel

Souvik Chakraborty

Problem Statement

When a loan providing company receives loan application, there are risks associated with the bank’s decision where the company has to decide for loan approval based on the applicant’s profile. Identifying such applicants who might be having loan repayment difficulties or are actually suitable for loan approval is a big task.

This case study aims to identify patterns which indicate if a client has difficulty paying their installments.
In other words, Identification of such driving factors (or driver variables) behind loan default (the variables which are strong indicators of default) using EDA is the aim of this case study.

Analysis Approach

Since the aim is to identify pattern for the loan defaulter using Exploratory Data Analysis,
we are using various graphs like boxplot, bar chart and pie chart to understand the distribution of different features (columns) provided in the two dataset.

The approach for both the dataset will be as follows.

1. Data Extraction
2. Data Cleaning : Verifying if Datatypes conversion required, Missing Value treatment, Outlier detection and treatment, Data conversion in case required.
3. Data Imbalance
4. Perform Univariate, bivariate analysis on categorical and numerical features.
5. Merging both the datasets to perform EDA on merged dataset.
6. Find the top 10 correlated features.
7. Recommendation.

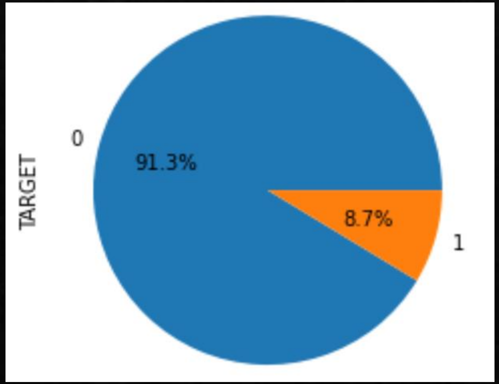
1. Dataset provided:
   - *'application_data.csv'* :Client's information at the time of application. The data is about whether a **client has payment difficulties.**
     (Data: 307511, Features: 122)

   - *'previous_application.csv'* :Client's information about the previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**
     (Data : 1670214, Features: 37)

2. For missing value treatment:
   - Columns having more than 40% of missing values were removed.
   - For categorical features were replaced with mode.
   - For numerical features were replaced with median.

3. For Outlier Detection:
   True numerical features were identified and using the boxplot, outliers were detected. Although outlier treatment was not applied, but it could have been treated using IQR .By replacing the max values with 70% IQR and min values with 25% .

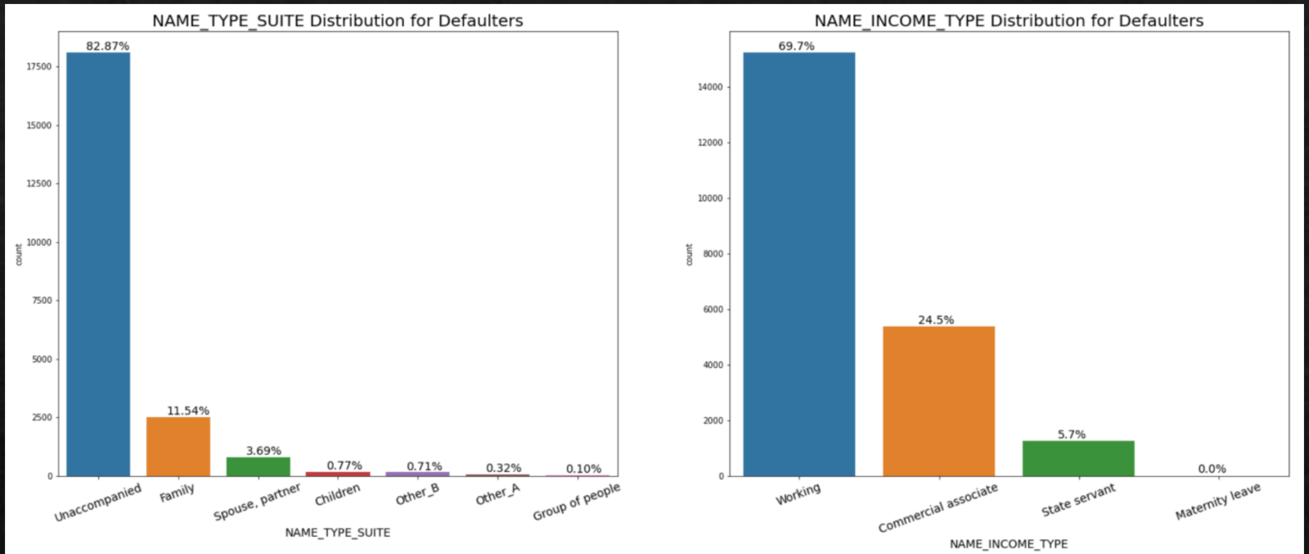
4. Data Imbalance:
   Pie chart plotted on the basis on 'Target' feature. The plot showed highly imbalanced data. Hence to analyze for defaulters we created new dataset having 'Target' as 1. IQR



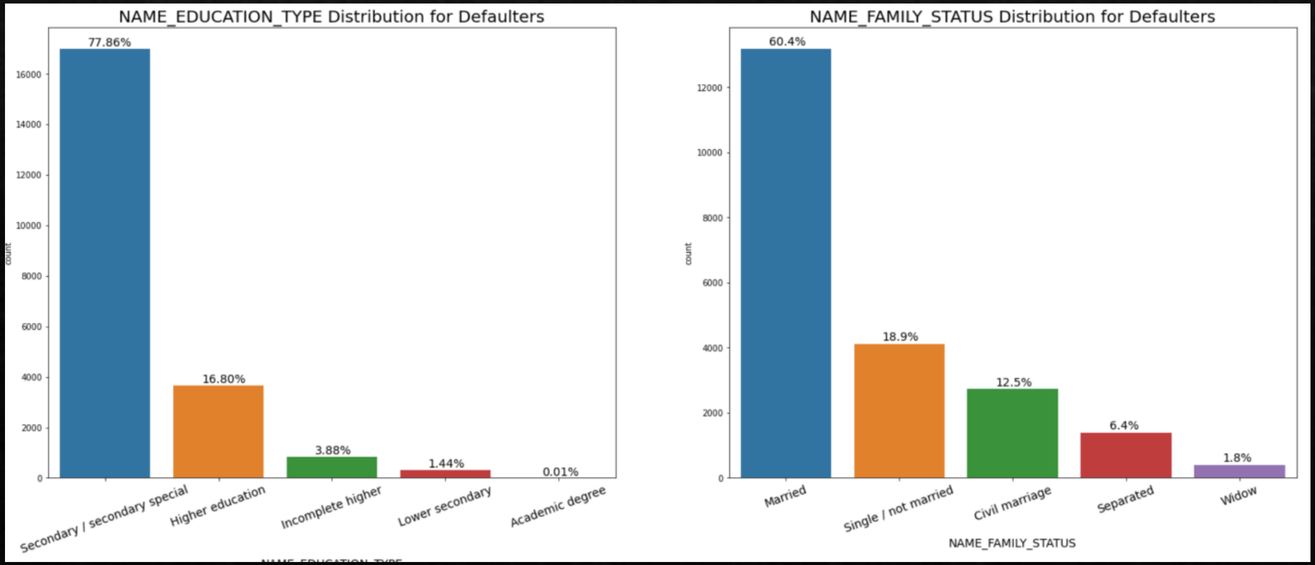
(Detailed work can be referred using the python notebook)

Univariate Analysis

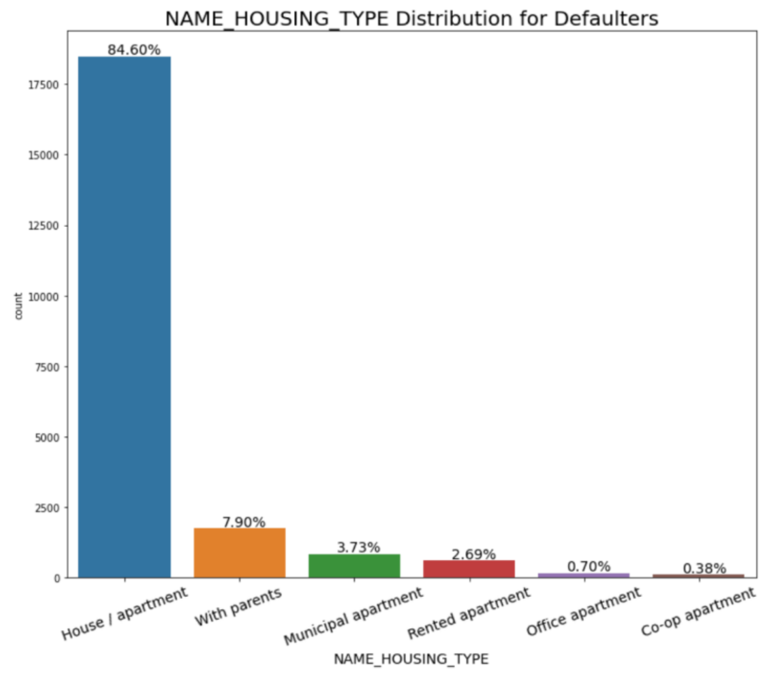
For Application_Data:

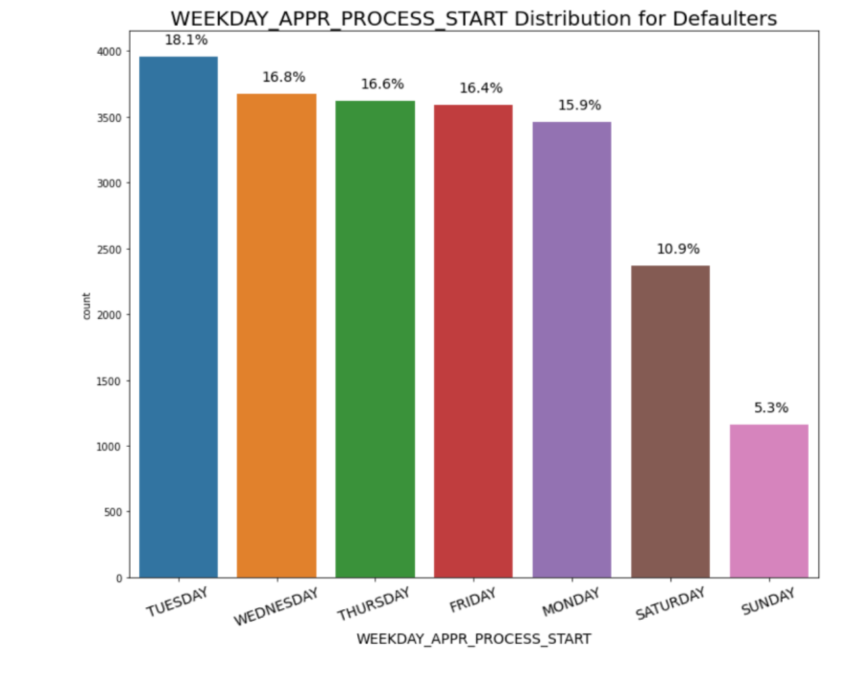


For Suite Type as 'Unaccompanied' and for Income Type as 'Working' percentage of loan difficulties is higher.

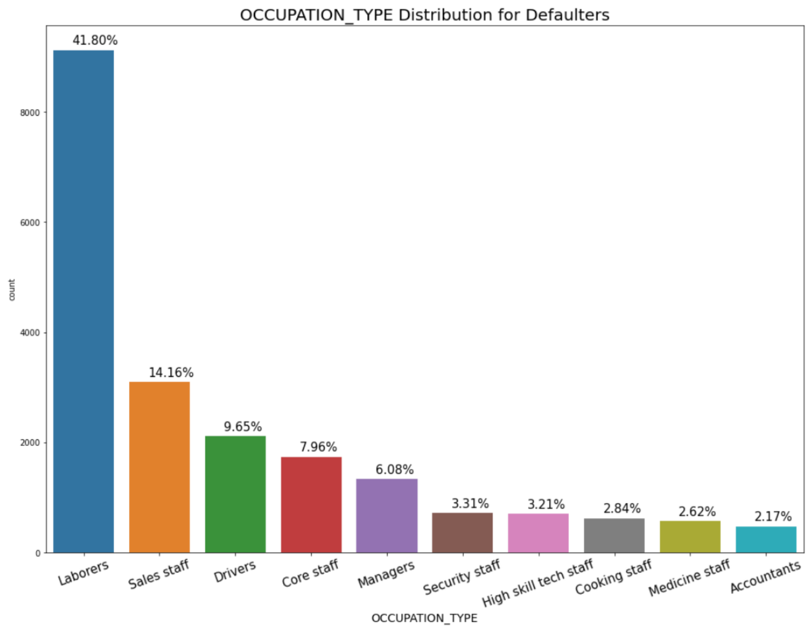


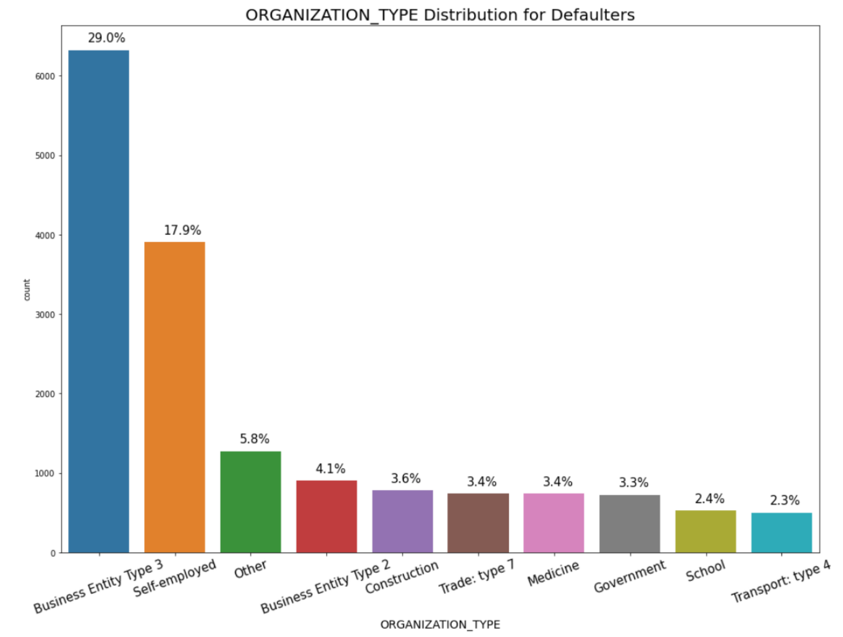
For Education Type as 'Secondary' and for Family_Status Type as 'Married' , percentage of loan difficulties is higher.



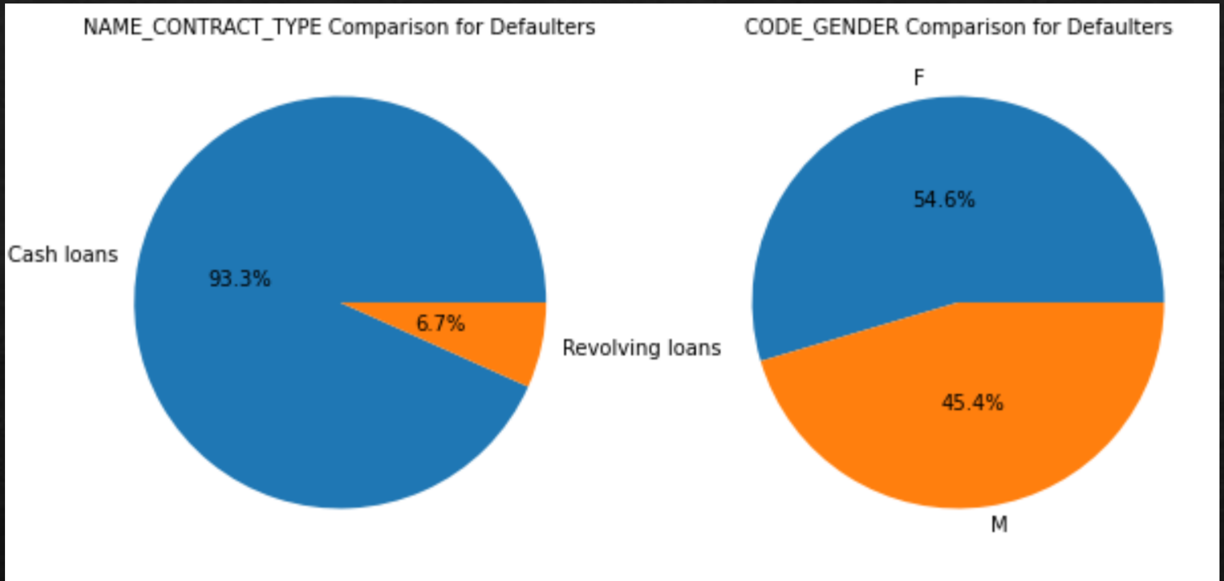


For Housing Type as 'House/Apartment' and for application applied on 'Tuesday', percentage of loan difficulties is higher.

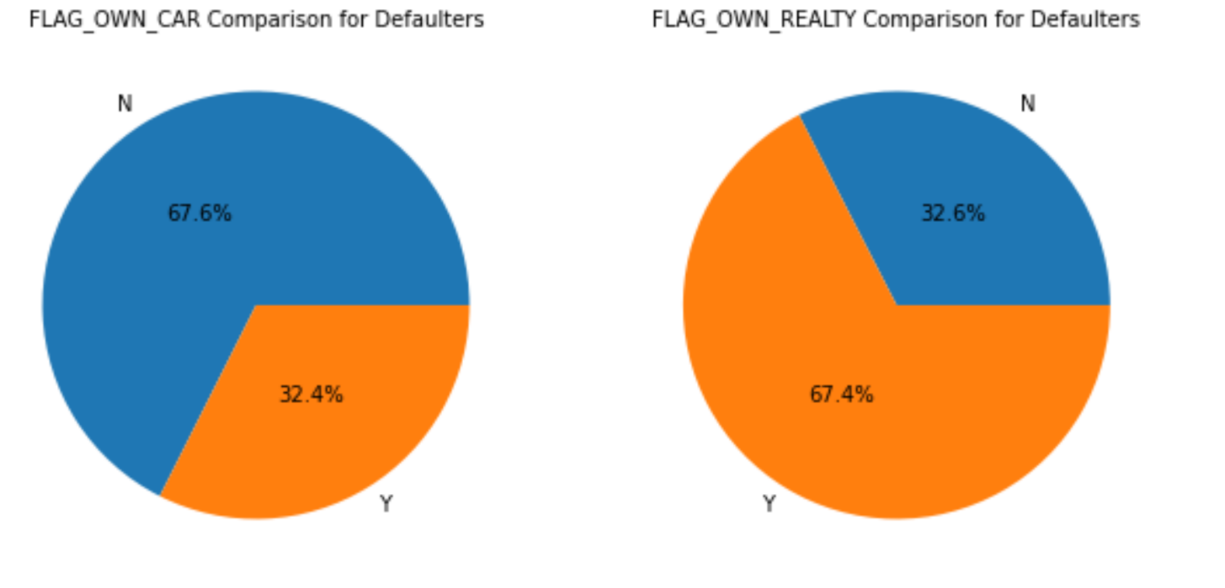




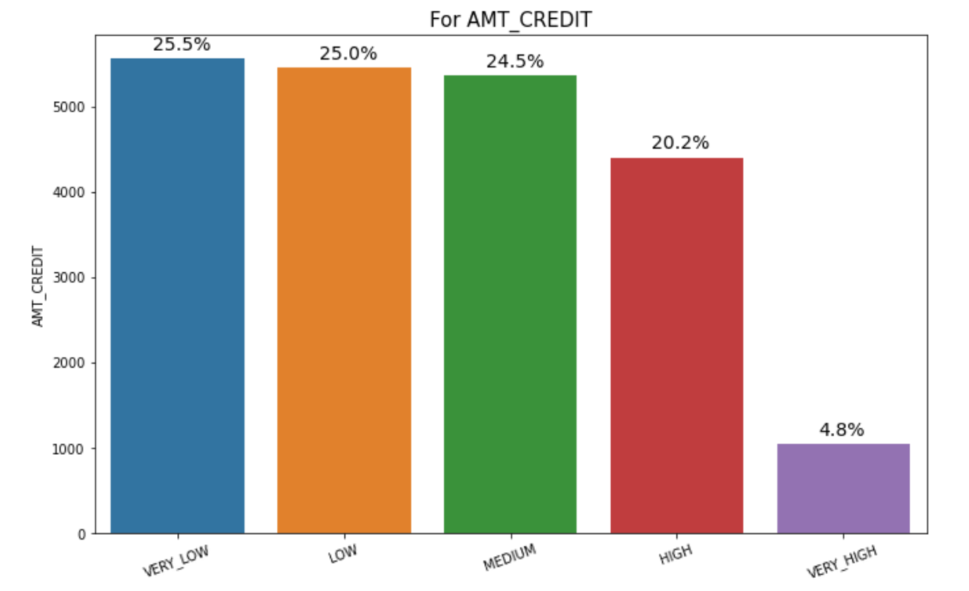
For Occupation types as 'Labourers and for organization type as 'Business Ent-3',percentage of Loan difficulties is higher

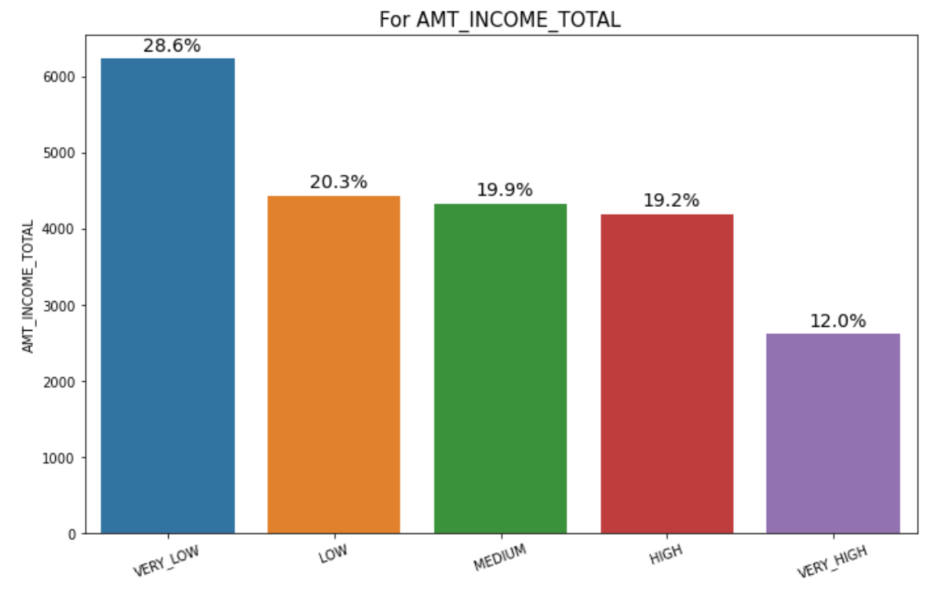


For Contract type as 'Cash Loan' and for Female, % of loan difficulties is higher.

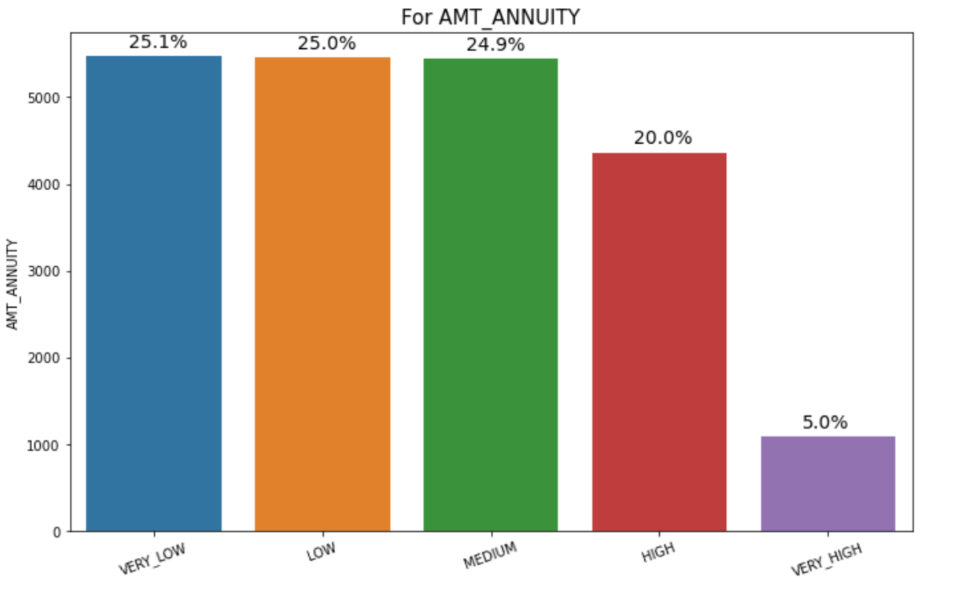


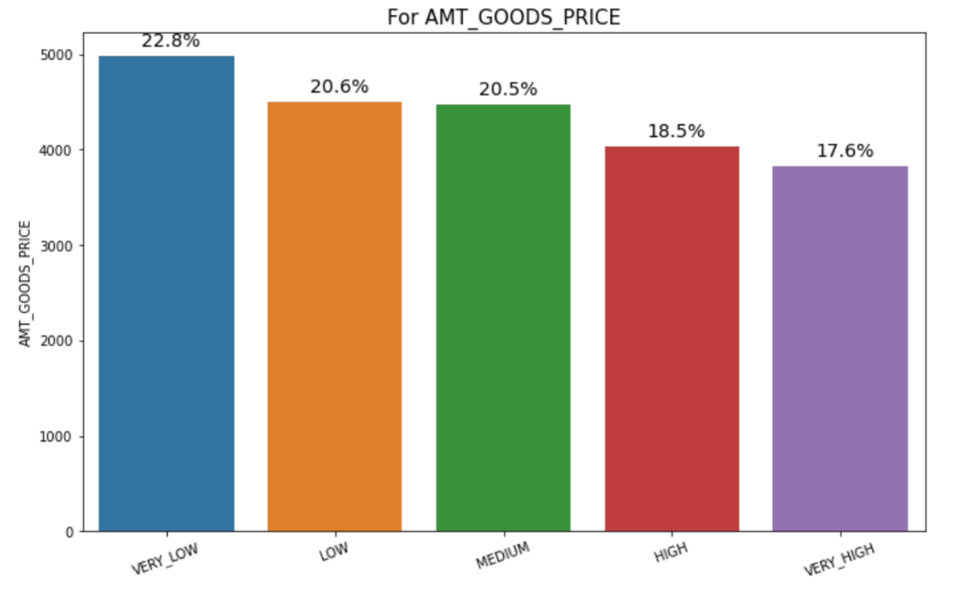
For a customer NOT having a car, or for a customer having a flat, % of loan difficulties is higher.


For AMT_CREDIT


For AMT_INCOME_TOTAL

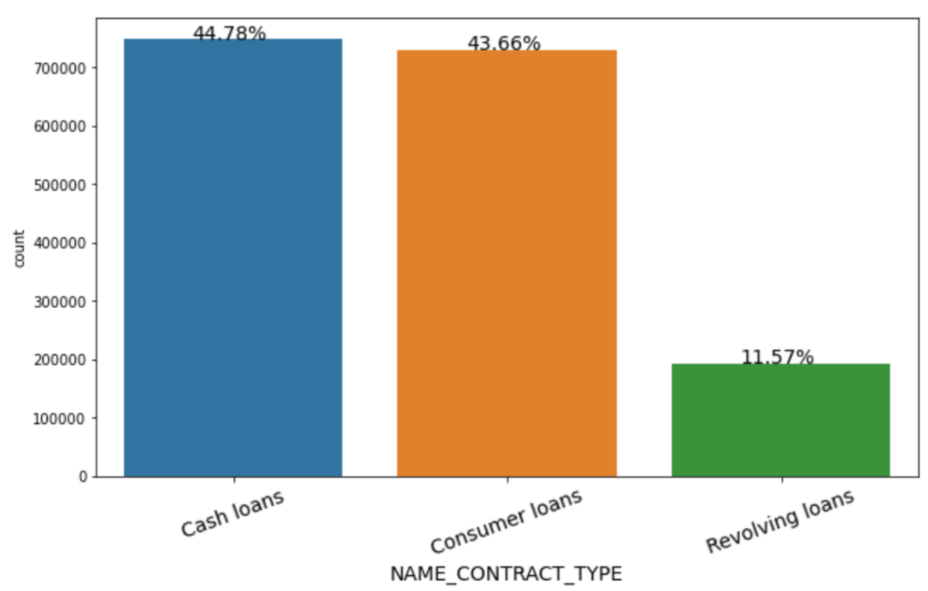
Lower the Credit amount or lower the Income, more % of Loan Difficulties.

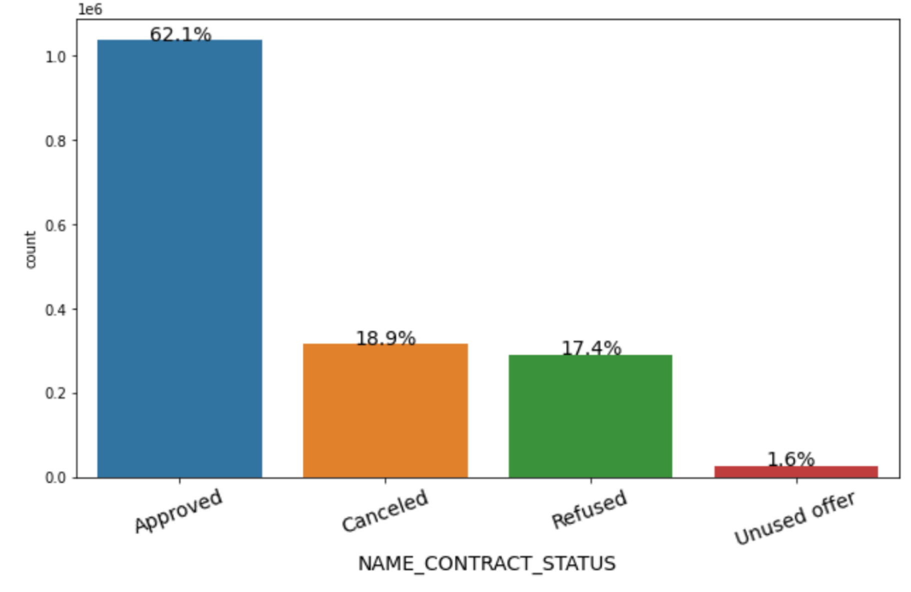

For AMT_ANNUITY


For AMT_GOODS_PRICE

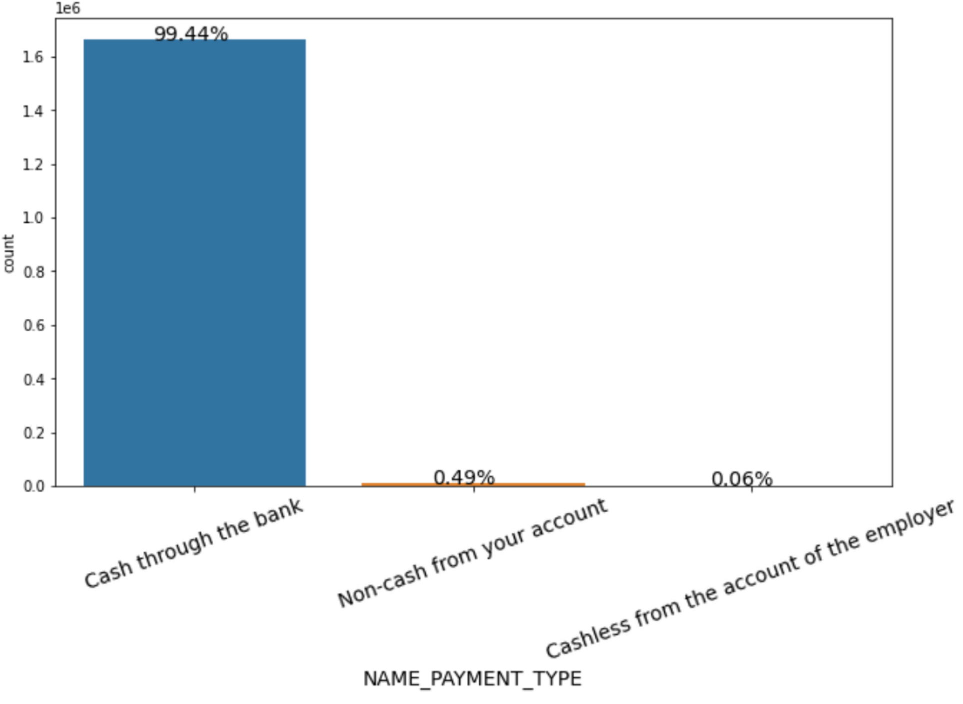
Even Lower the Annuity or lower the previous credit amount, more is the loan difficulty %.

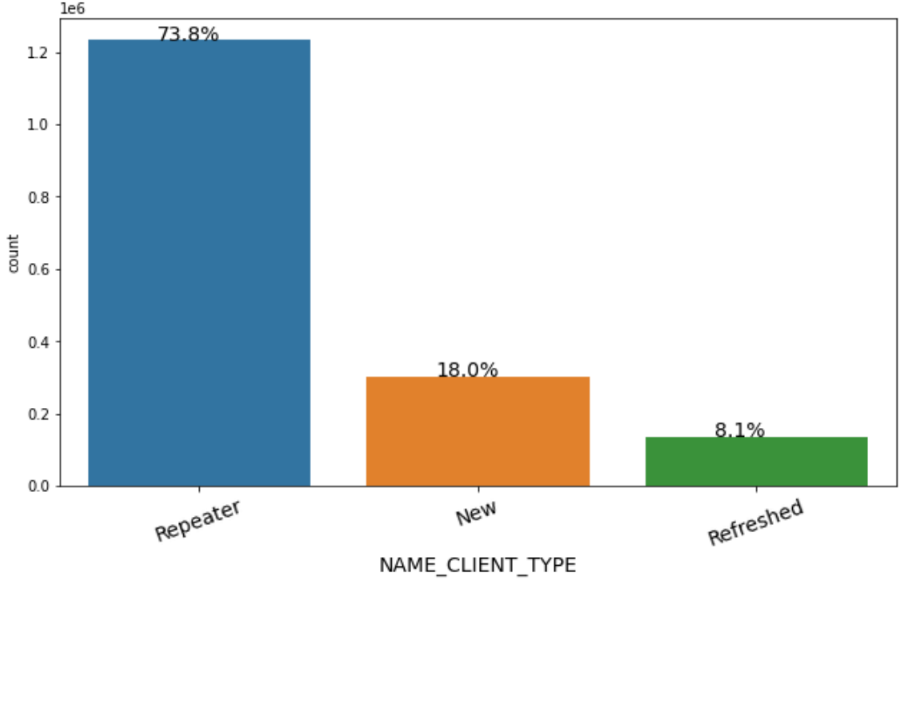
For Previous_Application:



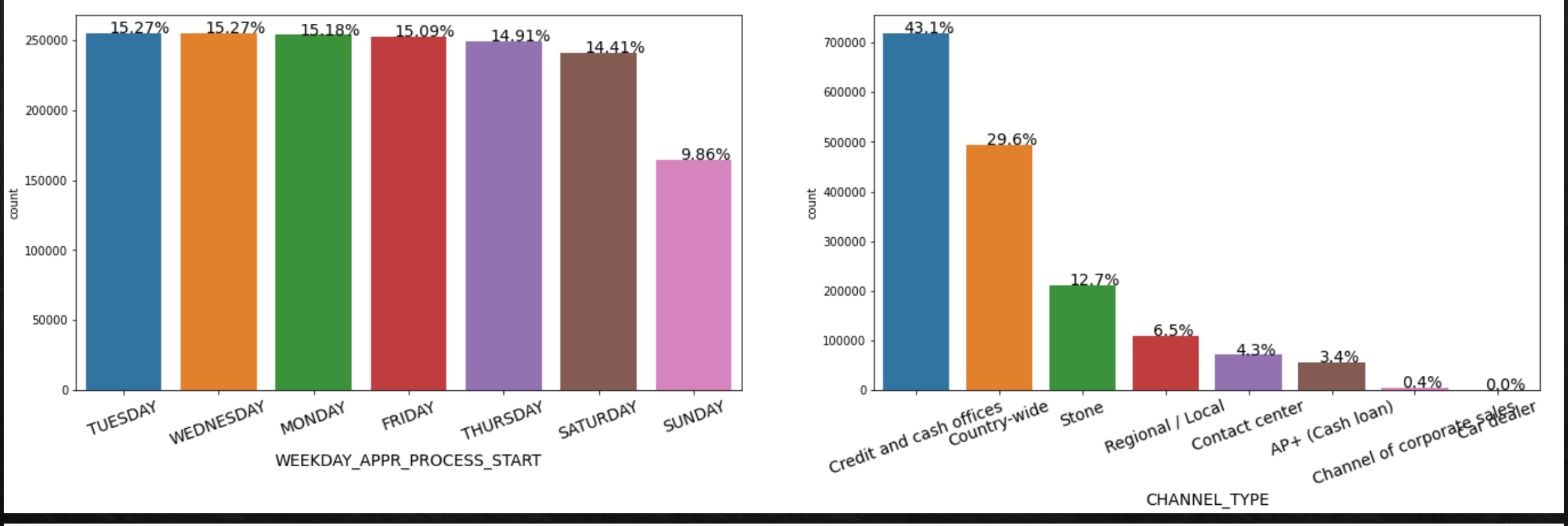


Majority applying for Cash Loans and Majority of loans are approved and very less percentage of loans are 'Unused offers'.

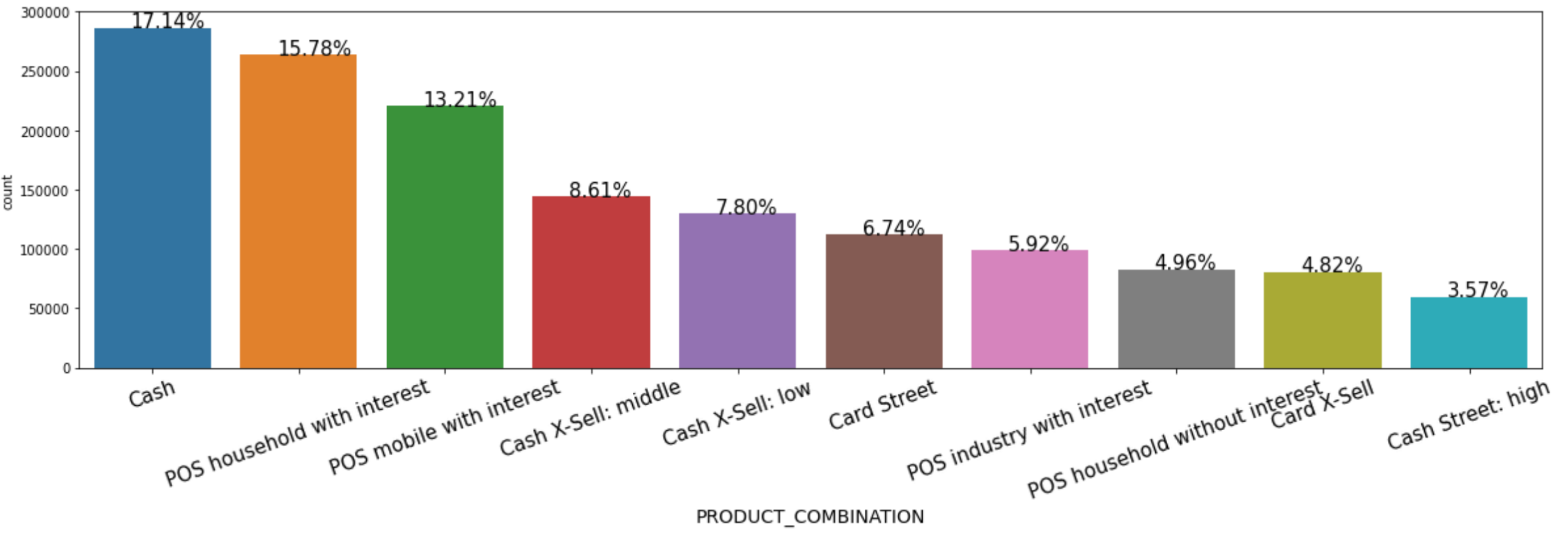




99% choose to pay through banks and majority of the client were repeater.



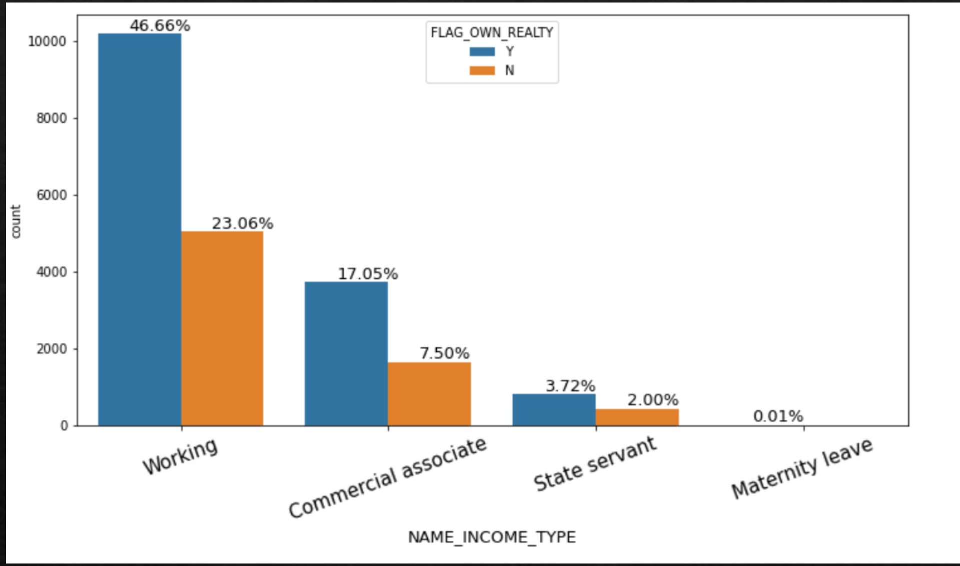
Applicants applying on 'Sunday' are the least. Majority of applicants coming from 'Credit and Cash Offices'.

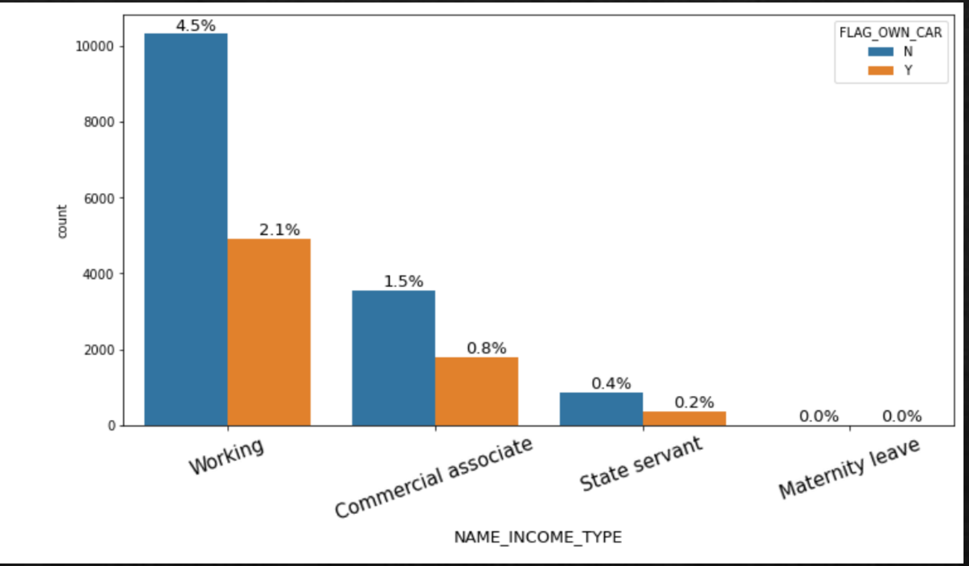


Majority of loans applied for Cash and good amount for 'POS household with interest'.

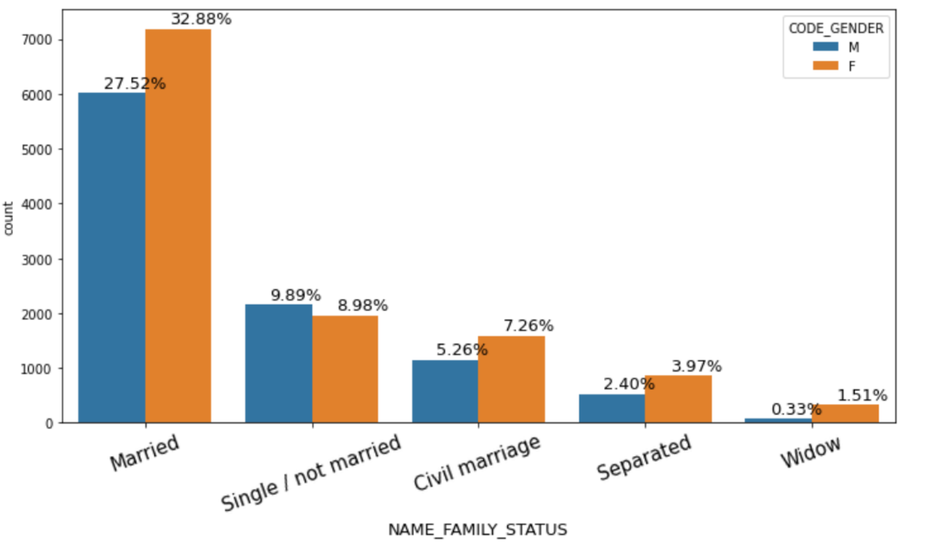
Bivariate Analysis

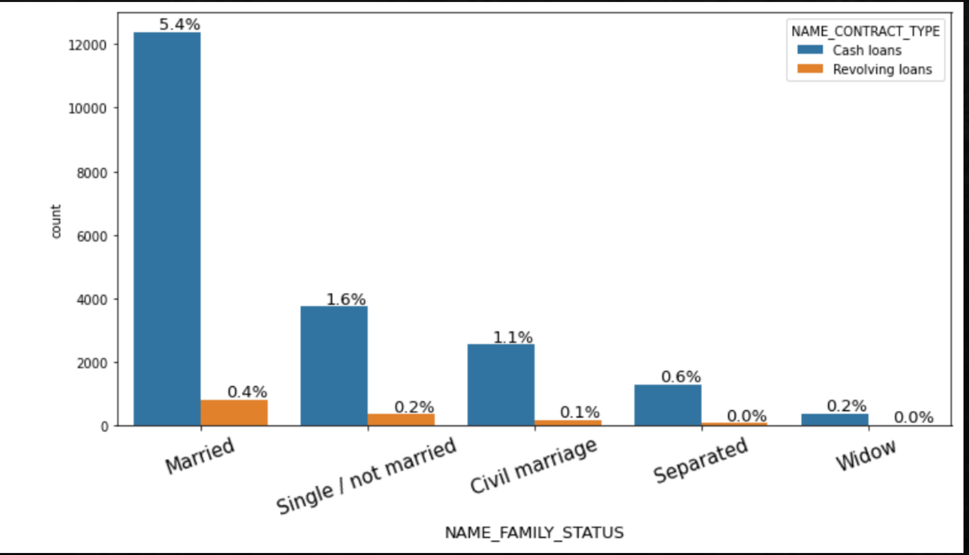
For Customer_application:



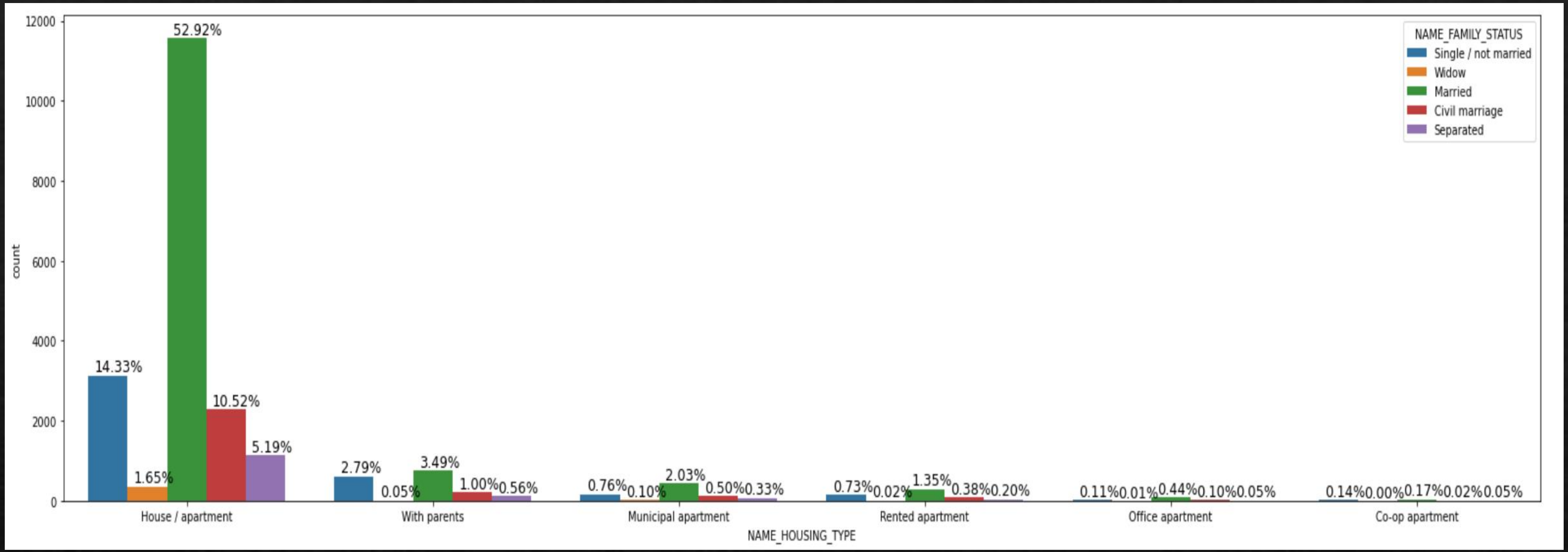


'Working' Income Type having flat or 'Working' Type not owing a car have more loan difficulties.

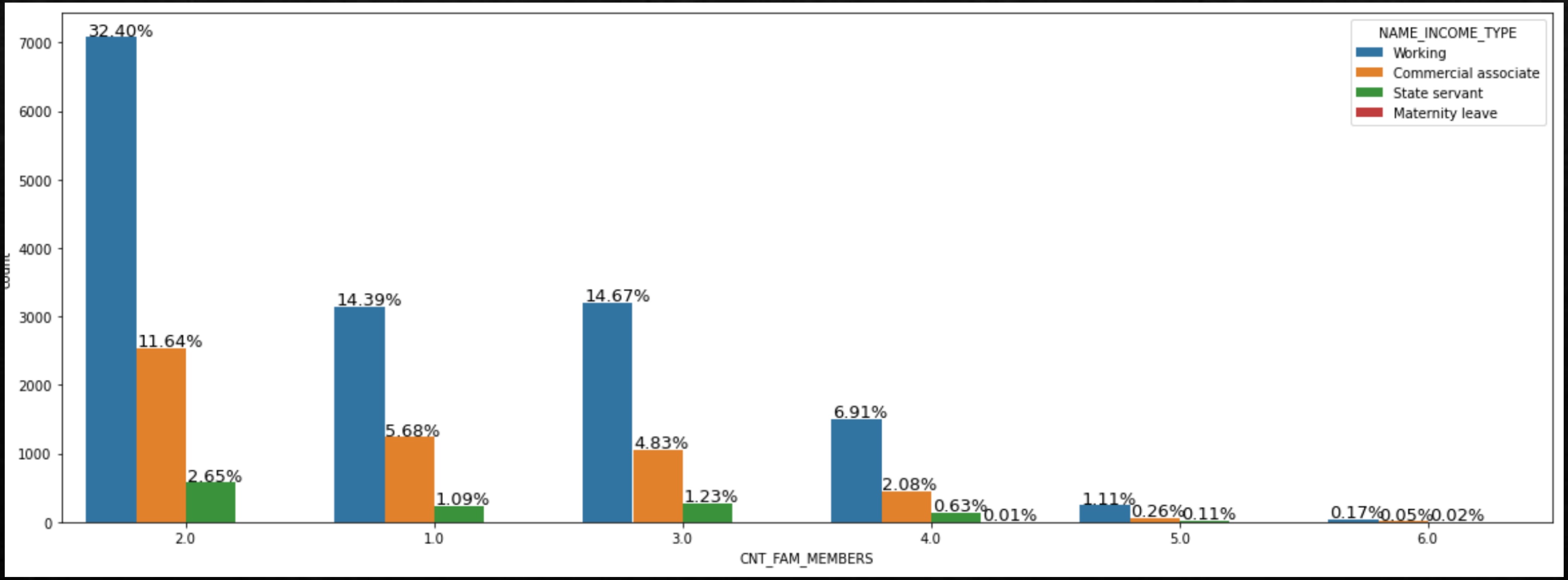




Female Married applicant or Married applicant with contract type as 'Cash Loan' have more % of loan difficulties.

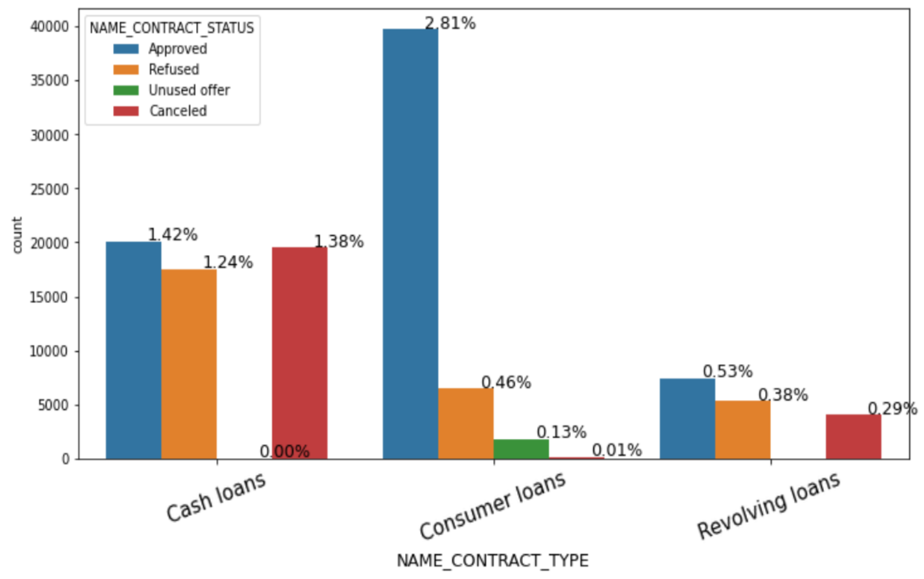


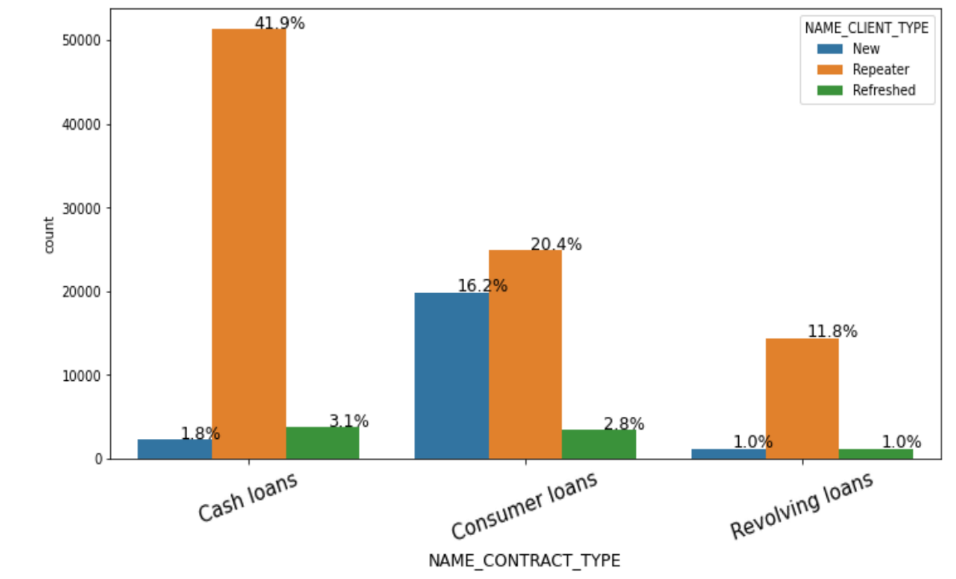
Married Applicant with House/Apartment have more loan difficulties.



Family members with a count of 2 and is working face more loan difficulties.

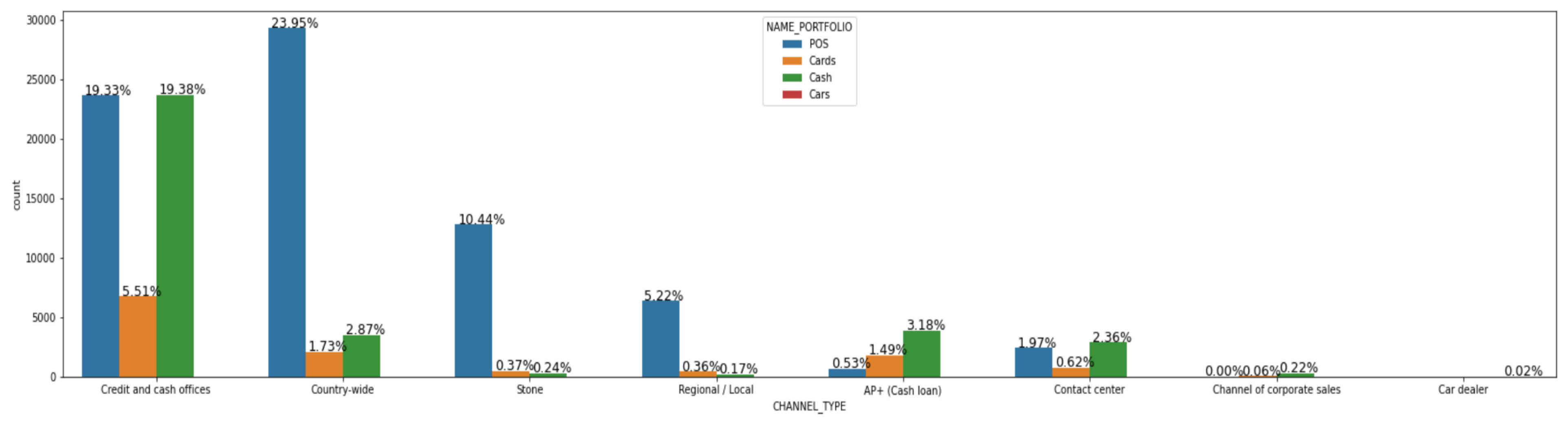
For merged Dataset:



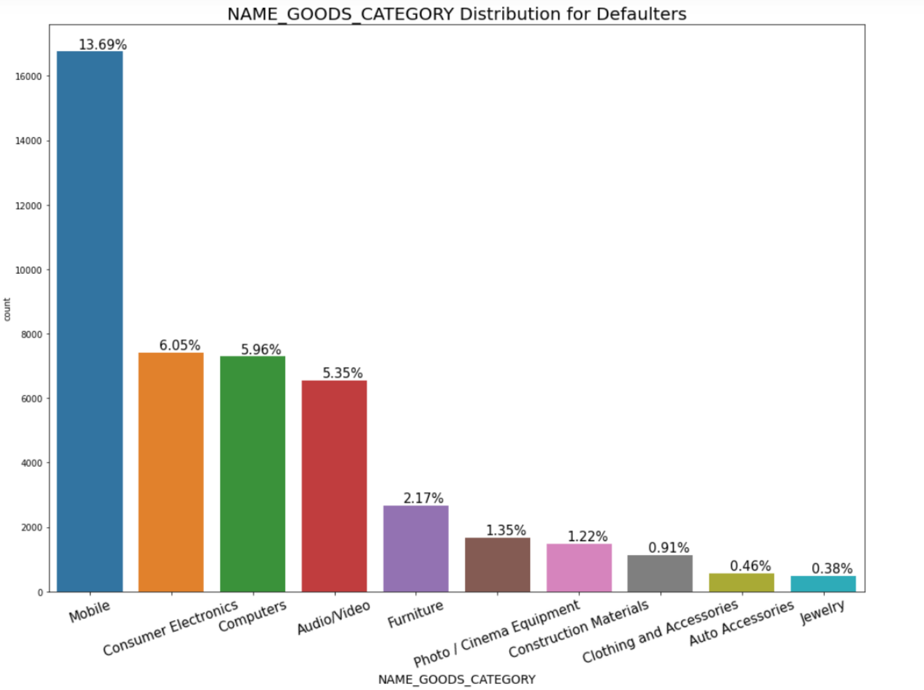


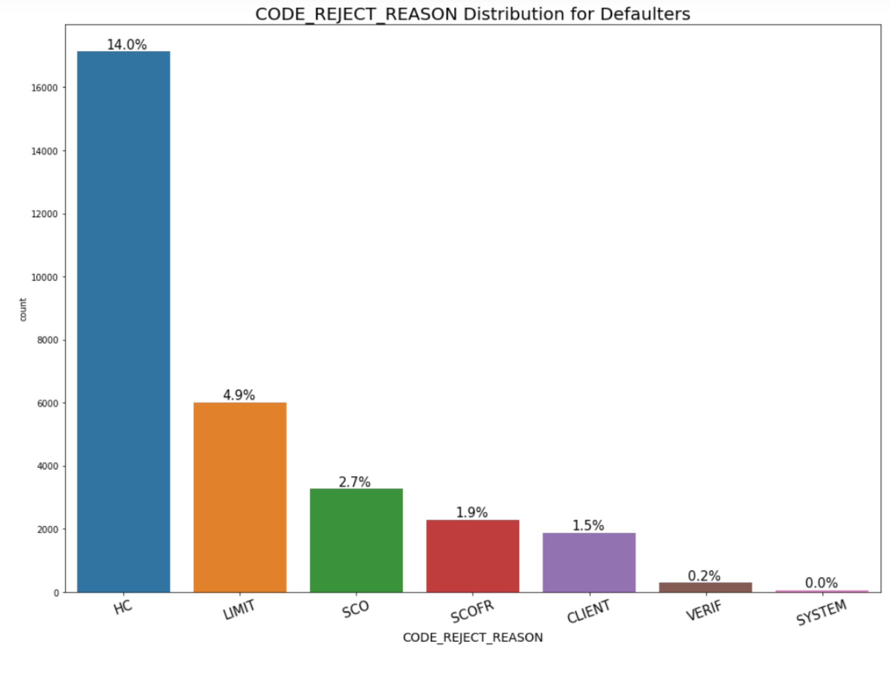
Although for 'Cash Loan ' type % of difficulties is more, but 'Consumer Loan' type with 'Approved' Contract status have maximum majority of Loan Difficulties.

Also the 'Cash Loan' type with 'Repeated' Customer have more Loan Difficulties.

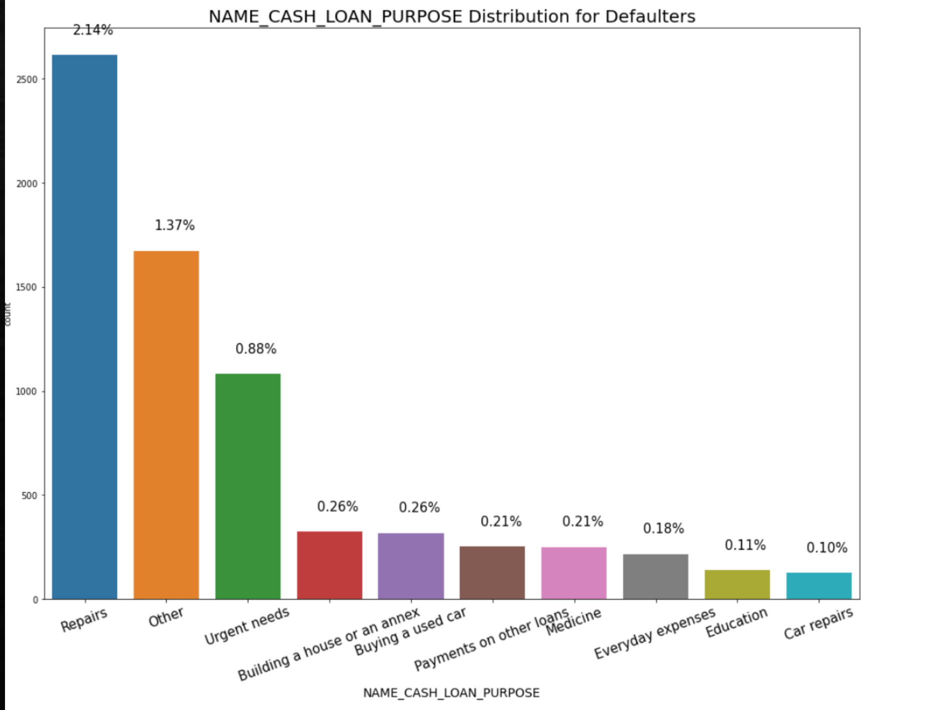


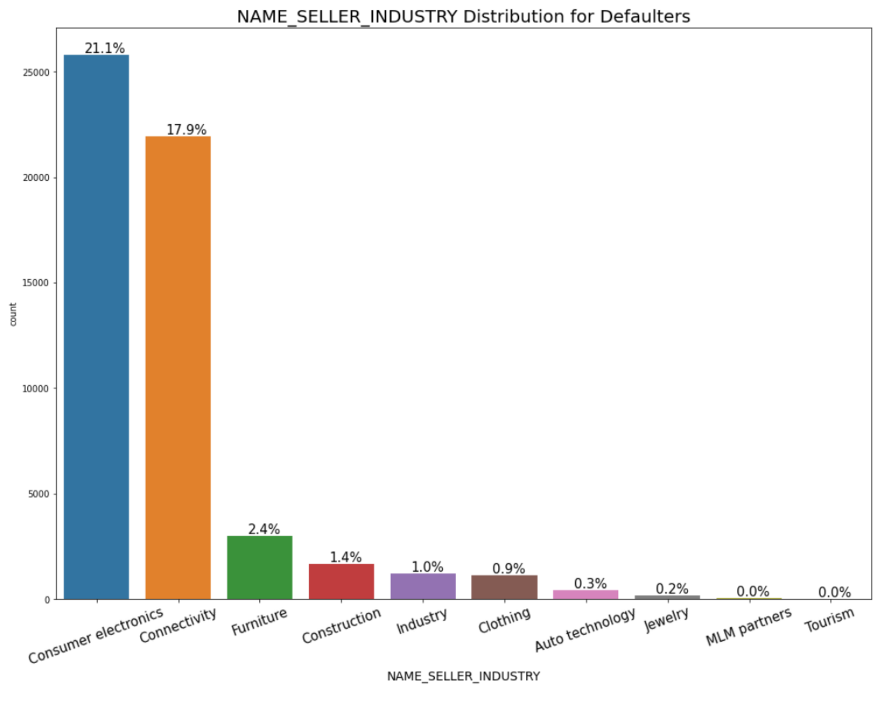
% of Loan Difficulty more for 'Country-wide’ Channel type for 'POS' portfolio.





Majority of loans for mobile have more % of Loan difficulty. HC is the reason majority of applications got rejected.

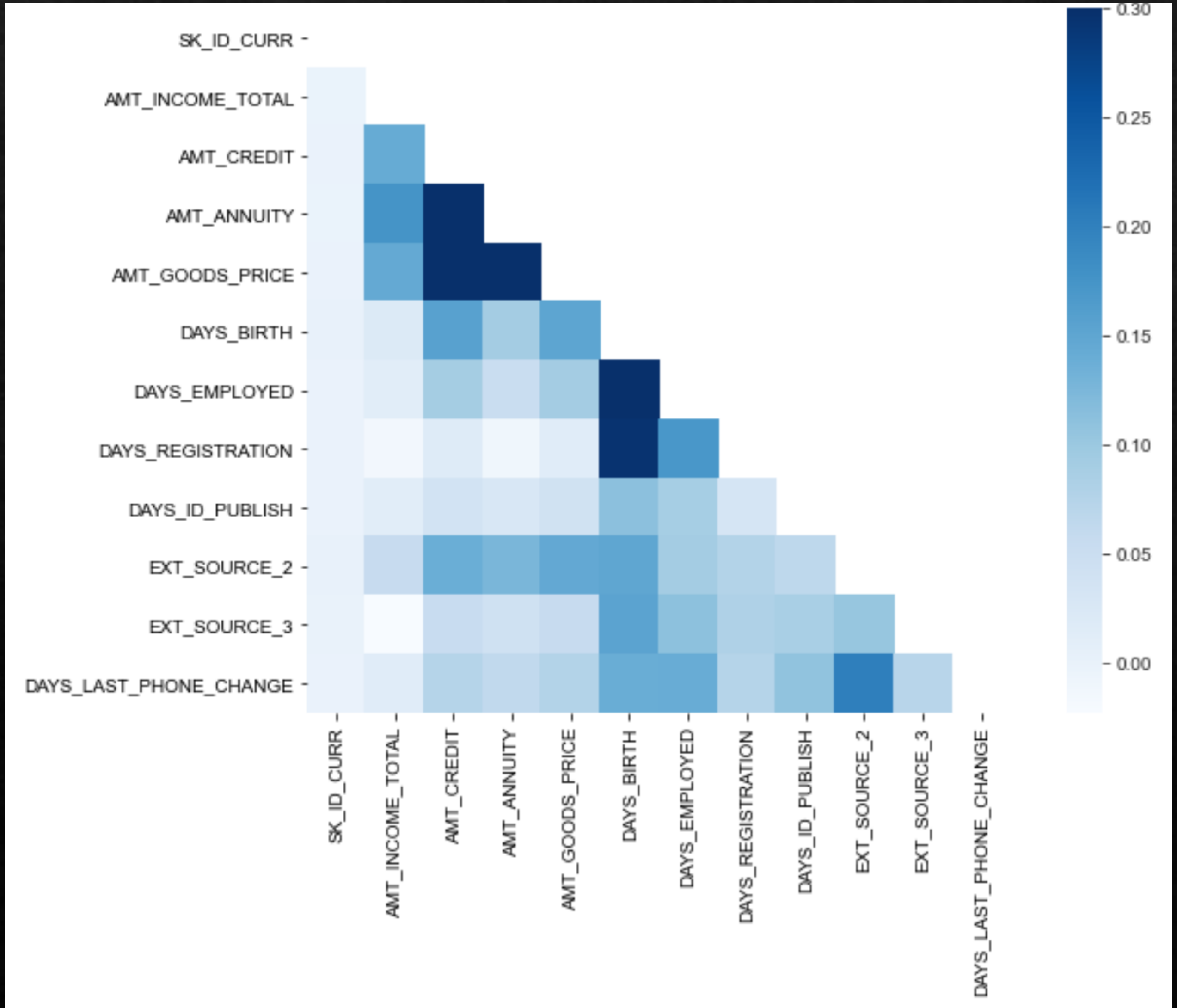




Purpose of cash loan for 'Repairs' have more % of Loan difficulties. Also 'Consumer Industries' have more Loan difficulties.

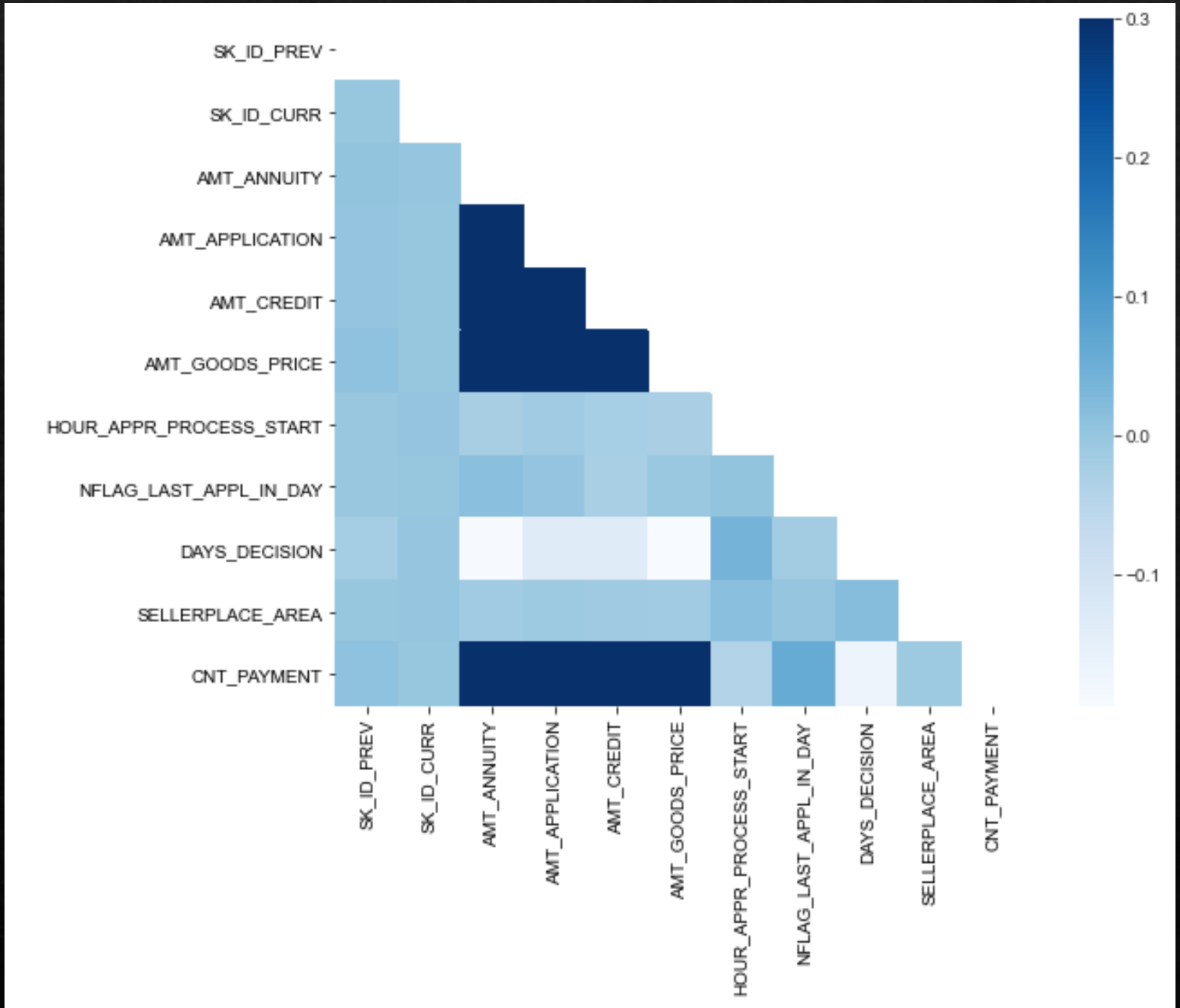
Top 10 Correlation

For Customer_application:



Feature-1	Feature-2	Correlation-Coefficient
AMT_CREDIT	AMT_GOODS_PRICE	0.986191
AMT_GOODS_PRICE	AMT_ANNUITY	0.765439
AMT_CREDIT	AMT_ANNUITY	0.761294
DAYS_EMPLOYED	DAYS_BIRTH	0.352321
DAYS_REGISTRATION	DAYS_BIRTH	0.29593
EXT_SOURCE_2	DAYS_LAST_PHONE_CHANGE	0.201864
AMT_INCOME_TOTAL	AMT_ANNUITY	0.175383
DAYS_REGISTRATION	DAYS_EMPLOYED	0.171754
AMT_CREDIT	DAYS_BIRTH	0.157405
DAYS_BIRTH	EXT_SOURCE_3	0.154222

For Previous_application:



Feature-1	Feature-2	Correlation-Coefficient
AMT_GOODS_PRICE	AMT_APPLICATION	0.987143
AMT_CREDIT	AMT_APPLICATION	0.975824
AMT_GOODS_PRICE	AMT_CREDIT	0.971117
AMT_ANNUITY	AMT_GOODS_PRICE	0.816293
AMT_CREDIT	AMT_ANNUITY	0.81167
AMT_ANNUITY	AMT_APPLICATION	0.805558
CNT_PAYMENT	AMT_APPLICATION	0.681114
AMT_GOODS_PRICE	CNT_PAYMENT	0.676007
CNT_PAYMENT	AMT_CREDIT	0.674387
CNT_PAYMENT	AMT_ANNUITY	0.406187

Conclusion

As per the finding in the previous slide, below are the driving factor for Loan Defaulters since they face majority of Loan difficulties

- Suite Type as *‘’Unaccompanied’*
- Income Type as *‘Working’*
- Education Type as *‘Secondary’*
- Family_Status Type as *‘Married’* with Housing type as ‘*House/Apartment’*
- Customer not having a Car
- Customer having a Flat
- For Occupation types as *‘Labourers’*
- for organization type as *‘Business Ent-3’*
- Family members with a count of ‘*2’* and ‘*Working’*
- *Lower* the Credit amount or *lower* the Income.
- Contract type as *‘Consumer Loan’* with Contract status as *‘Approved’*
- Contract type as *‘Cash Loan’* with *‘Repeated’* Customer
- For Loan purpose as ‘*Repairs’*
- For seller Industries as *‘Consumer electronics’*

Thank You