EDA ASSIGNMENT

Introduction :

This case study aims to give you an idea of applying EDA in a real business scenario. In this case study, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

Problem Understanding :

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.
- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.
- The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:
- The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- All other cases: All other cases when the payment is paid on time.
- When a client applies for a loan, there are four types of decisions that could be taken by the client/company):
- **Approved:** The Company has approved loan Application
- **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
- **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
- **Unused offer:** Loan has been cancelled by the client but on different stages of the process.
- In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

PROBLEM DESCRIPTION :

- This case study aims to identify patterns which indicate if a client has difficulty paying their installments 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 utilise this knowledge for its portfolio and risk assessment.
- To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

Data Understanding :

- This dataset has 3 files as explained below:

1. 'application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.
3. 'columns_description.csv' is data dictionary which describes the meaning of the variables.

Expected Results :

- **Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.**
- **Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)**
- **Hint:** *Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach.*
- **Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.**
- **Identify if there is data imbalance in the data. Find the ratio of data imbalance.**
- **Hint:** *How will you analyse the data in case of data imbalance? You can plot more than one type of plot to analyse the different aspects due to data imbalance. For example, you can choose your own scale for the graphs, i.e. one can plot in terms of percentage or absolute value. Do this analysis for the 'Target variable' in the dataset (clients with payment difficulties and all other cases). Use a mix of univariate and bivariate analysis etc.*
- **Hint:** *Since there are a lot of columns, you can run your analysis in loops for the appropriate columns and find the insights.*
- **Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.**
- **Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.**
- **Include visualisations and summarise the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases.**
- You need to submit one/two Ipython notebook which clearly explains the thought process behind your analysis (either in comments of markdown text), code and relevant plots. The presentation file needs to be in PDF format and should contain the points discussed above with the necessary visualisations. Also, all the visualisations and plots must be done in Python(should be present in the Ipython notebook), though they may be recreated in Tableau for better aesthetics in the PPT file.

Reading and understanding the data :

IMPORTING ALL THE NECCESSARY MODULES

- Importing the dataset (application.csv)
- Understanding the dataset (shape)
- There are 122 columns having various data types like object, int, float and 305711 rows**.**

INSIGHT:

- there are 122 columns and 307511 rows.
- there columns having negative, postive values which includes days. fixing is required
- there are columns with very hight values, columns related to Amount(Price). standardising is required, will perform these task later in the notebook

Notebook setting to display all rows and coloumns to have better clearity on the data

```
notebook setting to display all the rowns and
columns to have better clearity on the data
pd.set_option('display.max_rows', 500)
pd.set_option('display.max_columns', 500)
pd.set_option('display.width', 1000)
pd.set_option('display.expand_frame_repr', False)
```

Data Cleaning & Manipulation :

Null Values

-Dealing with Null values more than 50% :

INSIGHT

- There are 41 columns having null values more than 50% which are related to different area sizes on apartment owned/rented by the loan applicant

-Dealing with null values more than 15% :

from the columns dictionary we can conclude that only 'OCCUPATION_TYPE', 'EXT_SOURCE_3 looks relevant to TARGET column.thus dropping all other columns except 'OCCUPATION_TYPE','EXT_SOURCE_3

- After after dropping 8 columns we are left with 73 columns
- There are 2 more Columns with missing values more than 15%

Analyse & Removing Unneccsary Columns

Starting with EXT_SOURCE_3 , EXT_SOURCE_2. As they have normalised values, now we will understand the relation between these columns with TARGET column using a heatmap

- There seems to be no linear correlation and also from columns description we decided to remove these columns.
- Also we are aware correation doesn't cause causation

Now we will check columns with FLAGS and their relation with TARGET columns to remove irrelevant ones

- For this we will create a dataframe containig all FLAG columns and then plot bar graphs for each column with respect to TARGET column for which "0" will represent as Repayer and "1" will represent as Defaulter

Imputing values :

-Now that we have removed all the unneccesarry columns, we will proced with imputing values for relevent missing columns whereever required

Insight

- Now we have only 7 columns which have missing values more than 1%. Thus, we will only impute them for further analysis and such columns are: OCCUPATION_TYPE, AMT_REQ_CREDIT_BUREAU_YEAR, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_HOUR

Imputing for "OCCUPATION_TYPE" column

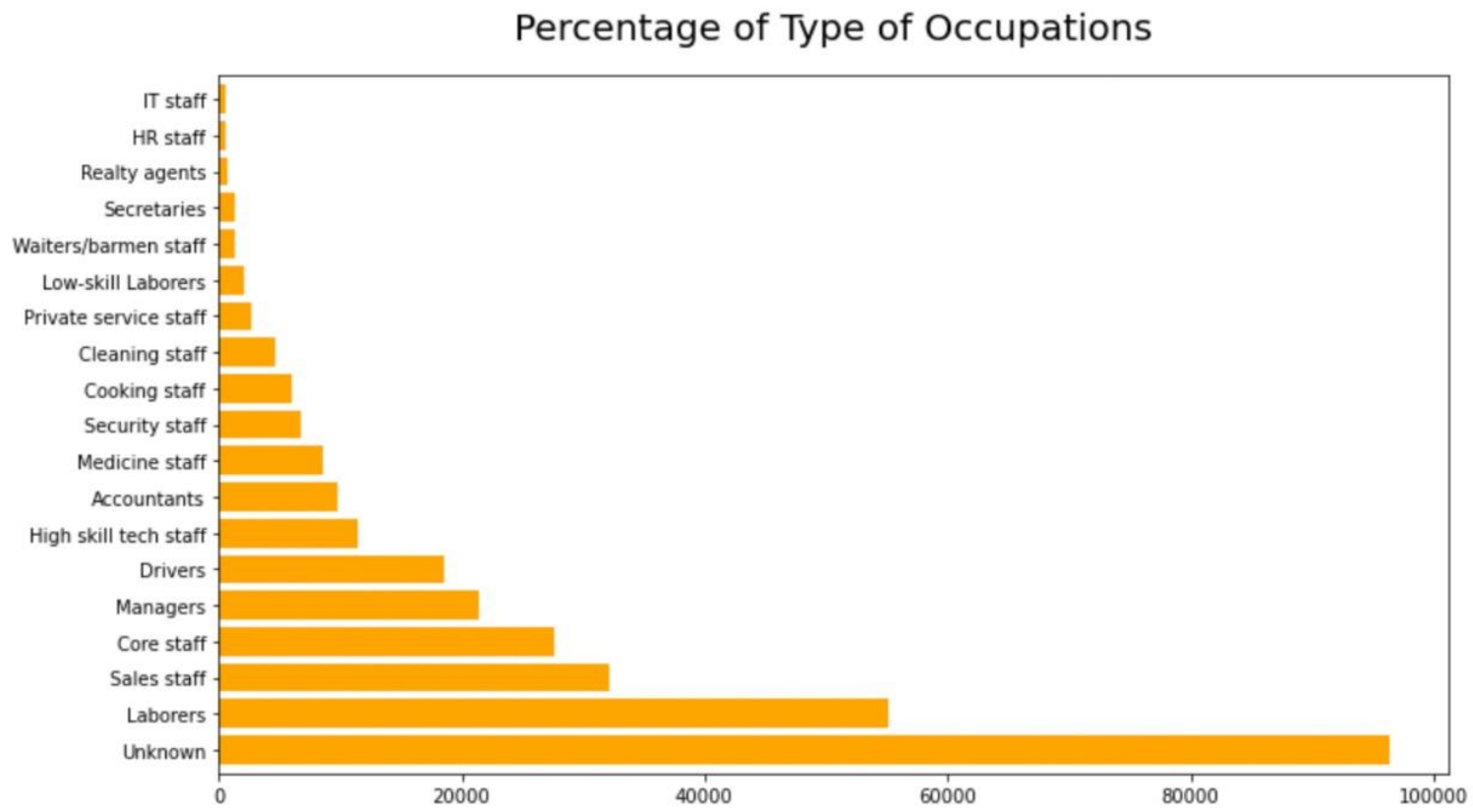
Insight:

- from above it looks like this columnn is categorical one and have missing values of 31.35%. to fix this we will impute another category as "Unknown" for the missing values.

GRAPHICAL REPRESENTATION :

- **Imputing null values with unknown**
- **Plotting a percentage graph having a each category of OCCUPATION_TYPE**

Percentage of type of Occupation



- **Highest percentage of values belongs to Unknown group and Secons belongs to Laborers**

Standardising values :

Insights:

- from above describe result we can see that
- columns AMT_INCOME_TOTAL, AMT_CREDIT, AMT_GOODS_PRICE have very high values, thus will make these numerical columns in categorical columns for better understanding.
- columns DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE which counts days have negative values. thus will correct those values
- convert DAYS_BIRTH to AGE in years , DAYS_EMPLOYED to YEARS EMPLOYED

Taking care of columns:

AMT_INCOME_TOTAL, AMT_CREDIT, AMT_GOODS_PRICE

Dealing with columns :

DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE

-from describe we get that days are in negative that is not usual, so to correct it we use absolute function as below.

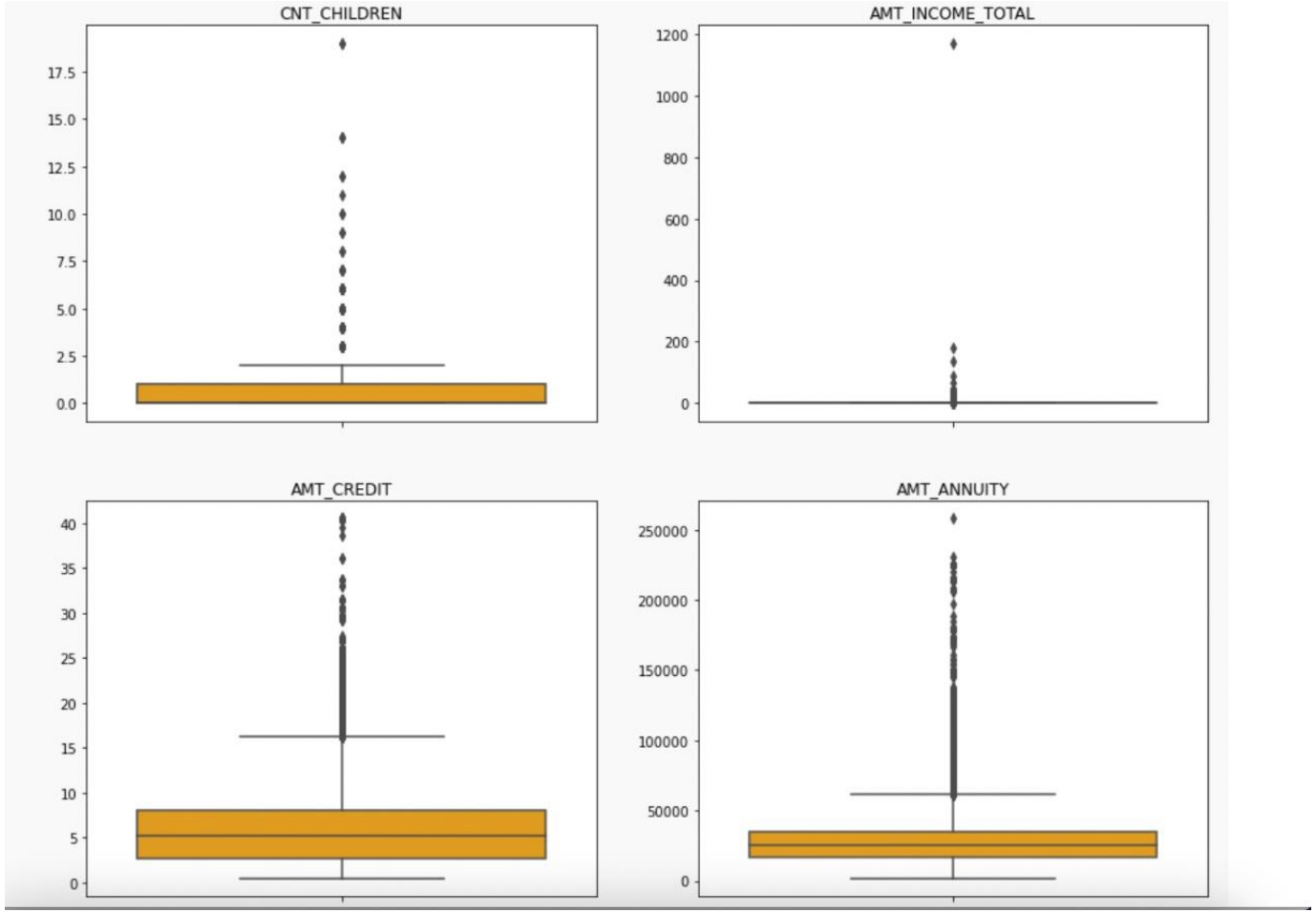
Convert

- DAYS_BIRTH, DAYS_EMPLOYED columns in terms of Years and binning years for better understanding, that is adding two more categorical column

Insight:

- It can be seen that in current application data
- AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
- AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.
- DAYS_BIRTH has no outliers which means the data available is reliable.
- DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.

Identifying Outliers :



Dataset 2 :

"previous_application.csv"

Note: Have followed similar steps done for application_data.csv

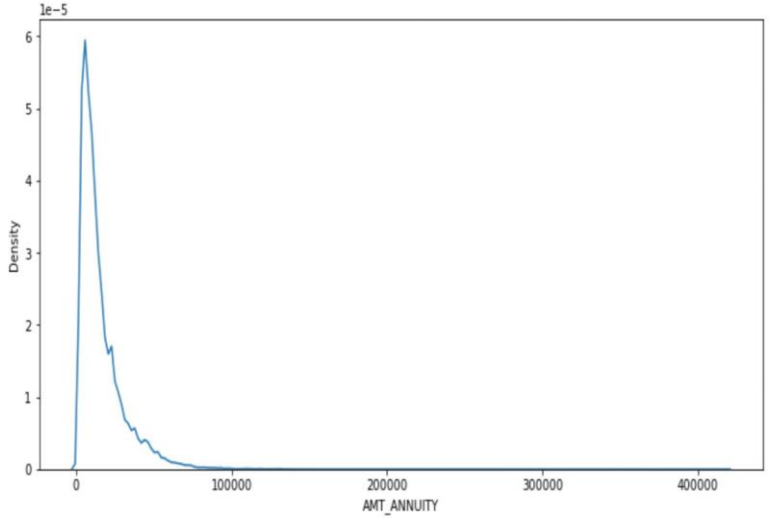
Insight

- there are 37 columns and 1679214 rows.
- there columns having negative, postive values which includes days. fixing is required
- There are missing values in columns 'DAYS_FIRST_DUE', 'DAYS_TERMINATION', 'DAYS_FIRST_DRAWING','DAYS_LAST_DUE_1ST_VERSION', 'DAYS_LAST_DUE' and these columns count days thus will keeping null values as they are
- Almost 35% loan applicatants have applied for a new loan within 1 year of previous loan decision

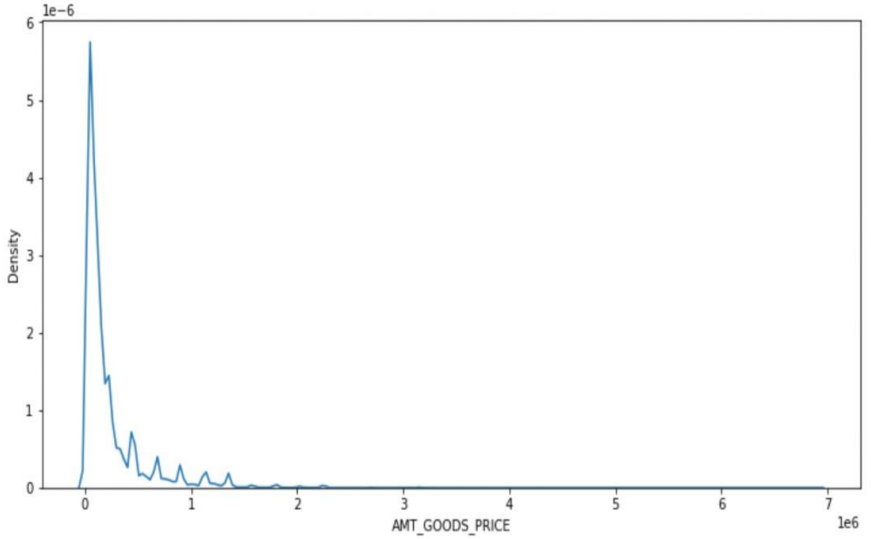
Now dealing with continuos variables "AMT_ANNUITY", "AMT_GOODS_PRICE"

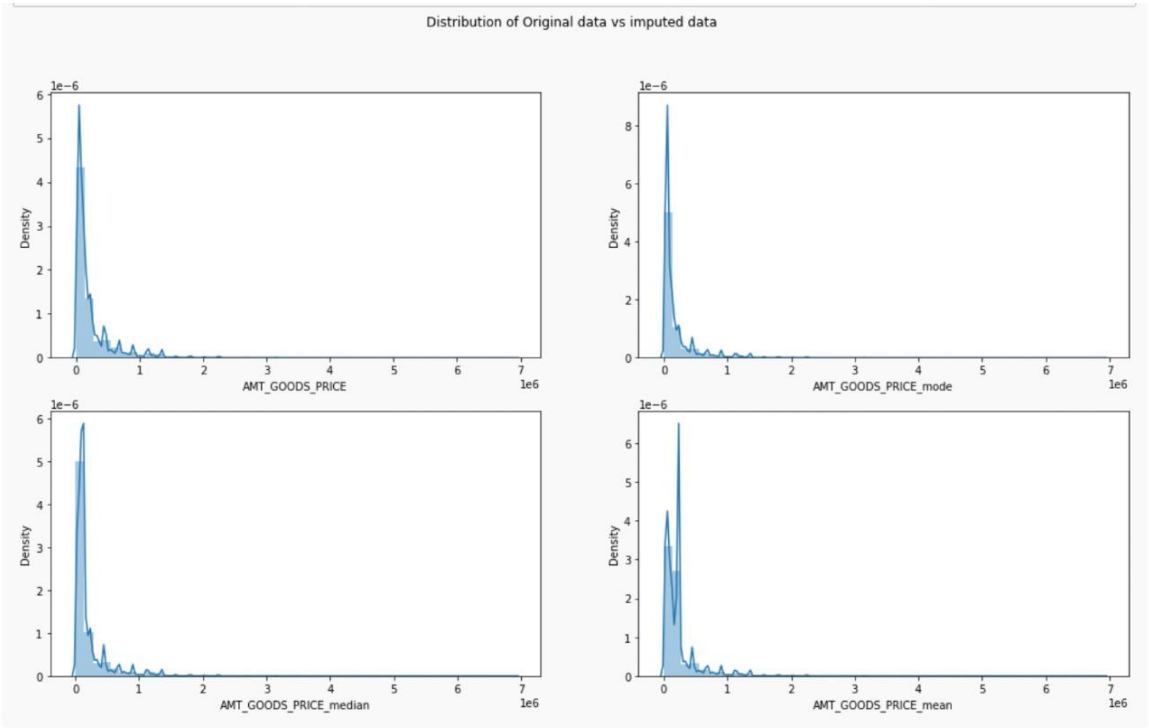
- **To impute null values in continuous variables, we plotted the distribution of the columns and used**
- median if the distribution is skewed
- mode if the distribution pattern is preserved.
- There is a single peak at the left side of the distribution and it indicates the presence of outliers and hence imputing with mean would not be the right approach and hence imputing with median.
- There are several peaks along the distribution. Let's impute using the mode, mean and median and see if the distribution is still about the same.

*



*

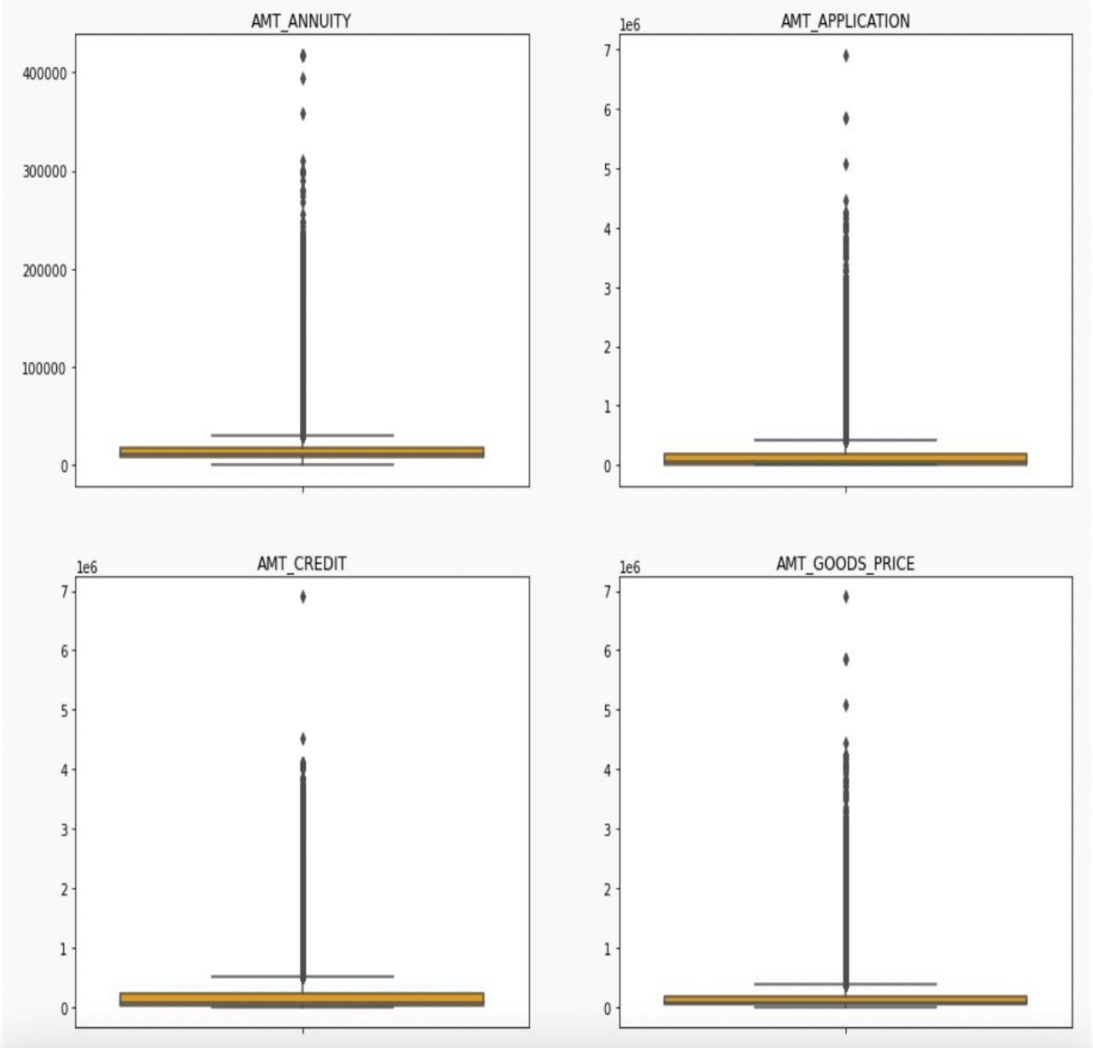


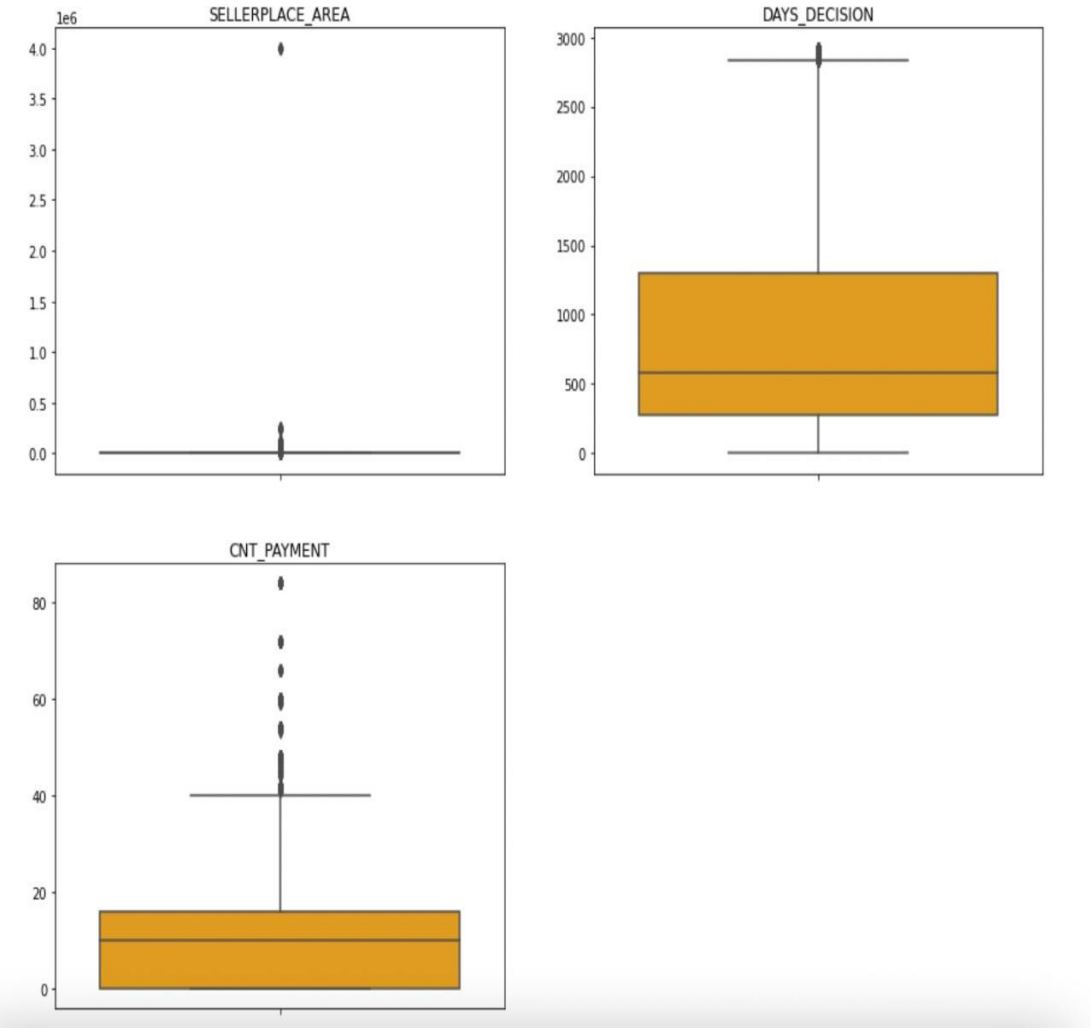


- **The original distribution is closer with the distribution of data imputed with mode in this case, thus will impute mode for missing values**

Finding outliers :

from describe we could find all the columns those wo have high difference between max and 75 percentile and the ones which makes no sense having max value to be so high are captured below





Imbalance Data & Plotting Functions:

Important Function for Univariate analysis

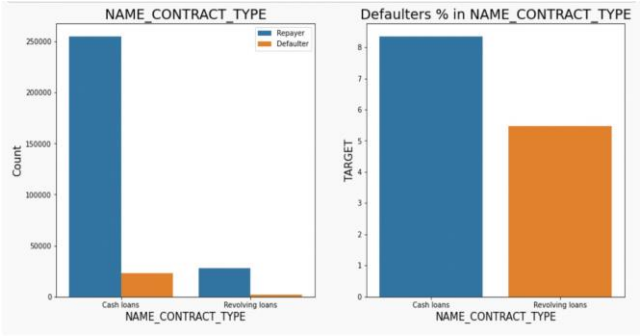
- Creating a function for plotting Variables to do univariate analysis. This function will create two plots

1. Count plot of given column w.r.t TARGET column
2. Percentage of defaulters within that column

- The function is taking 6 arguments

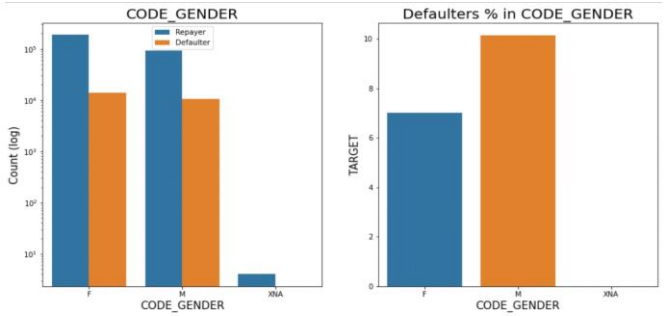
1. dataset : to put the dataset we want to use
2. col : column name for which we need to the analysis
3. target_col : column name for with which we will be comparing
4. ylog : to have y-axis in log10 terms, in case the plot is not readable
5. x_label_angle : to maintain the orientation of x-axis labels
6. h_layout : to give horizontal layout of the subplots

Segmented Univariate Analysis



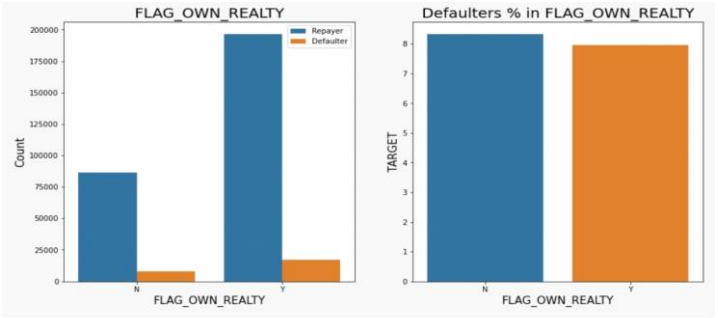
Inferences: Contract type

- Revolving loans are just a small fraction (10%) from the total number of loans
- Around 8-9% Cash loan applicants and 5-6% Revolving loan applicant are in defaulters



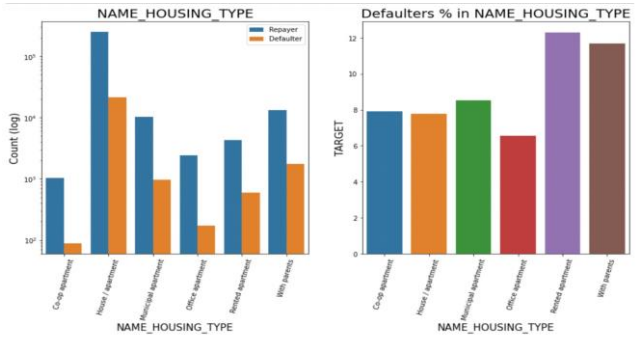
Inferences: Gender Type

- 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 about 10%, comparing with women about 7%



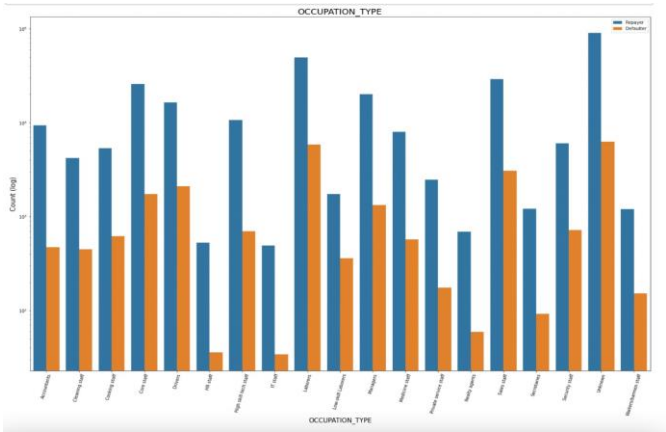
Inferences:

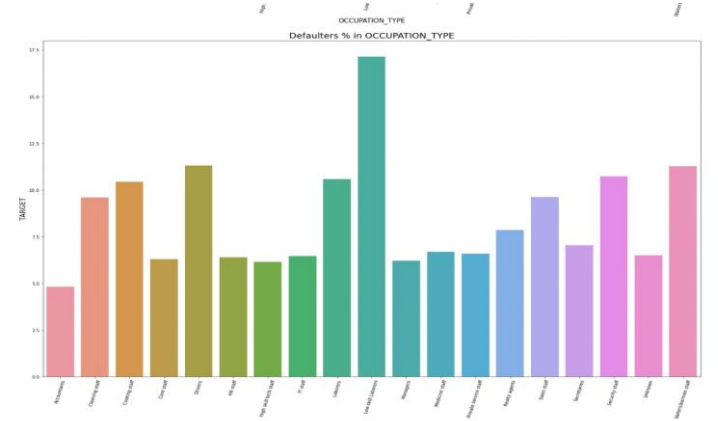
- The clients who own real estate are more than double of the ones that don't own.
- The defaulting rate of both categories are around the same (~8%). Thus we can infer that there is no correlation between owning a reality and defaulting the loan.



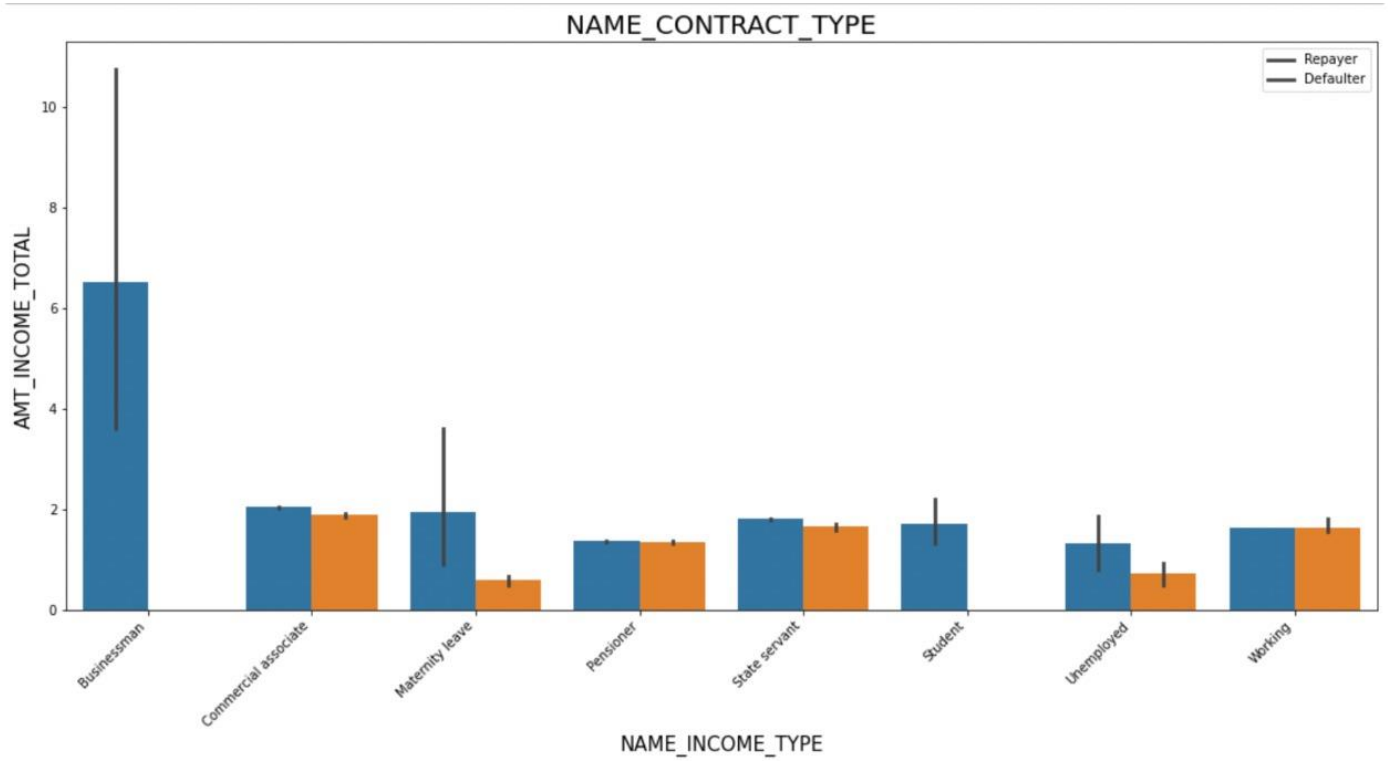
Inferences: Applicant House type

- 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





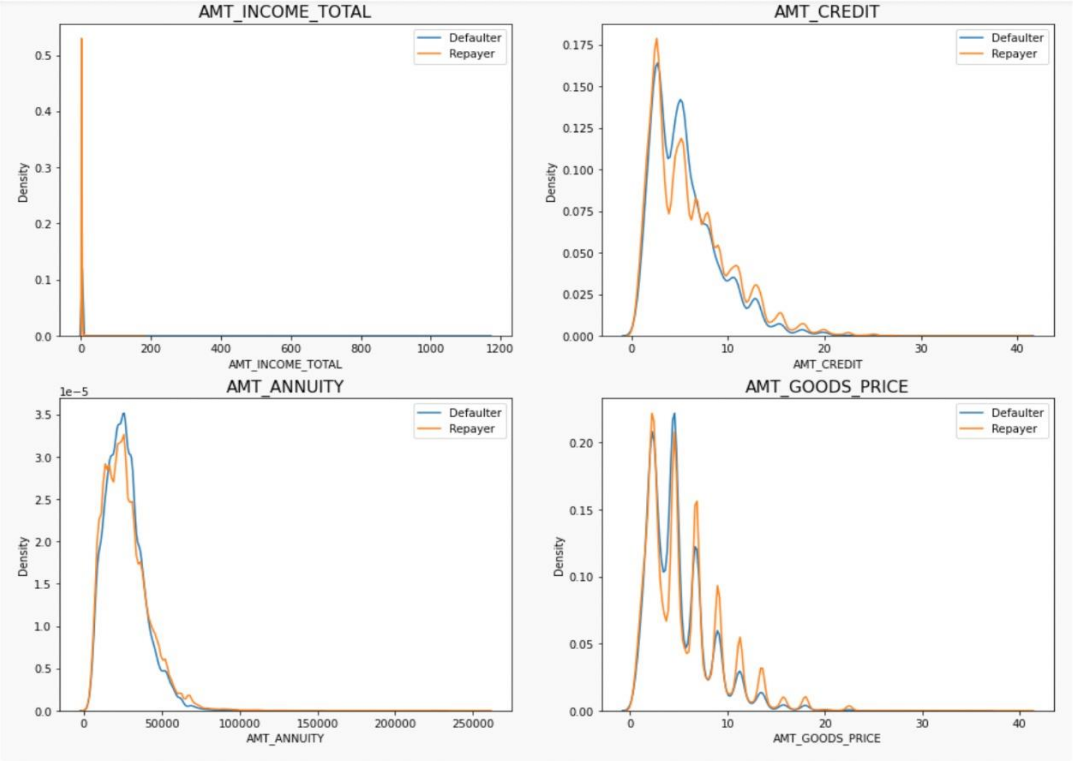
Bivariate or Multivariate Analysis



Inferences:

- **It can be seen that Businessman income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a Businessman could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs**

Univariate Analysis

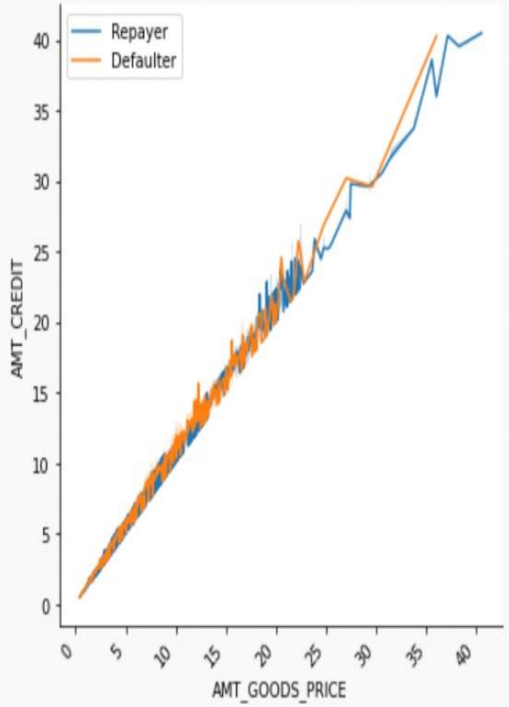


Inferences:

- **Most no of loans are given for goods price below 10 lakhs**
- **Most people pay annuity below 50K for the credit loan**
- **Credit amount of the loan is mostly less then 10 lakhs**

Bivariate Analysis

<Figure size 1080x1080 with 0 Axes>



Inferences:

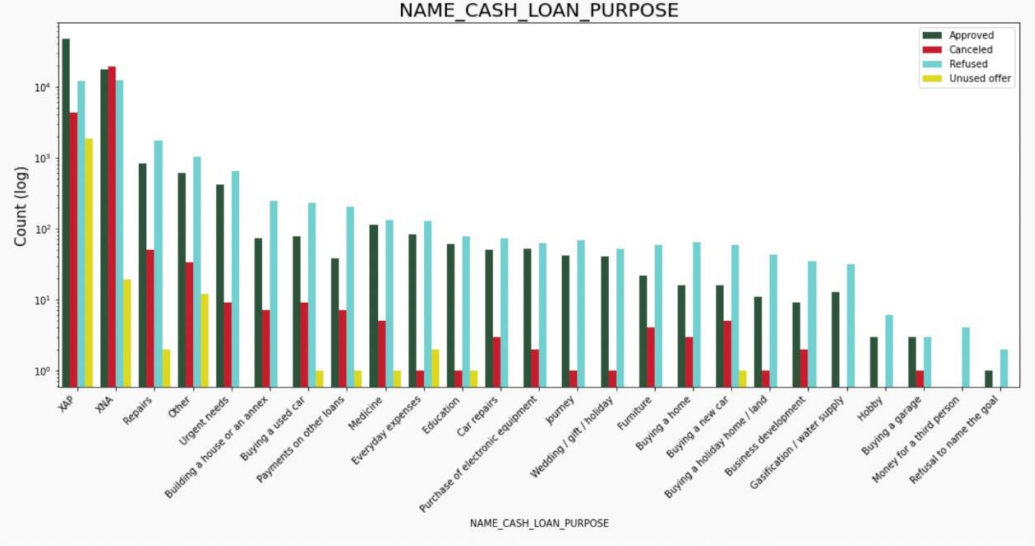
- **When the credit amount goes beyond 30 Lakhs, there is an increase in defaulters.**

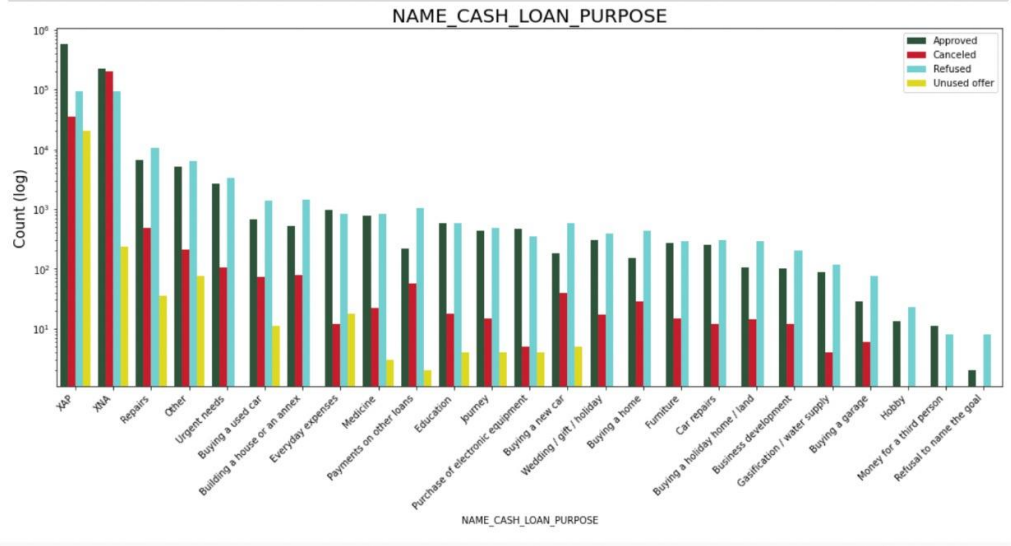
Merged Dataframes Analysis :

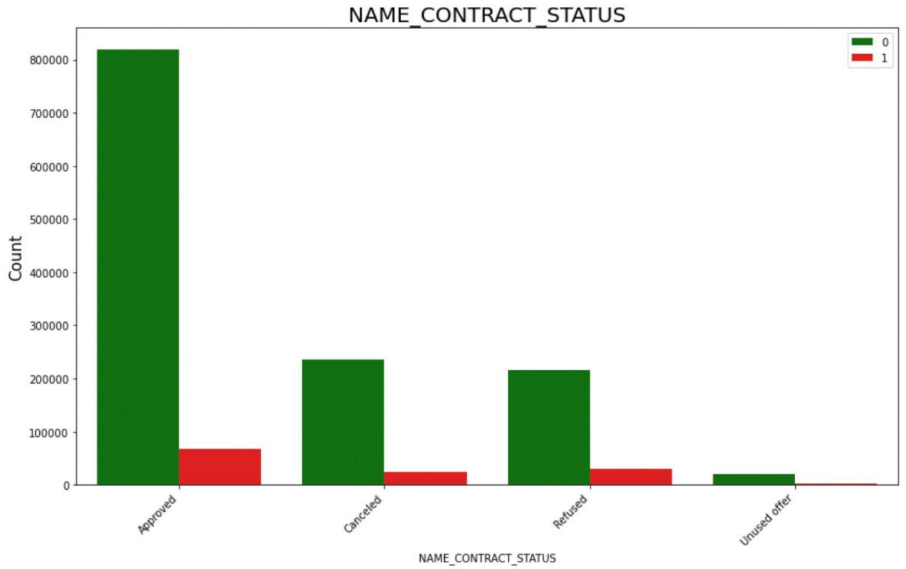
Merge both Dataframe on SK_ID_CURR with inner joins

Inferences:

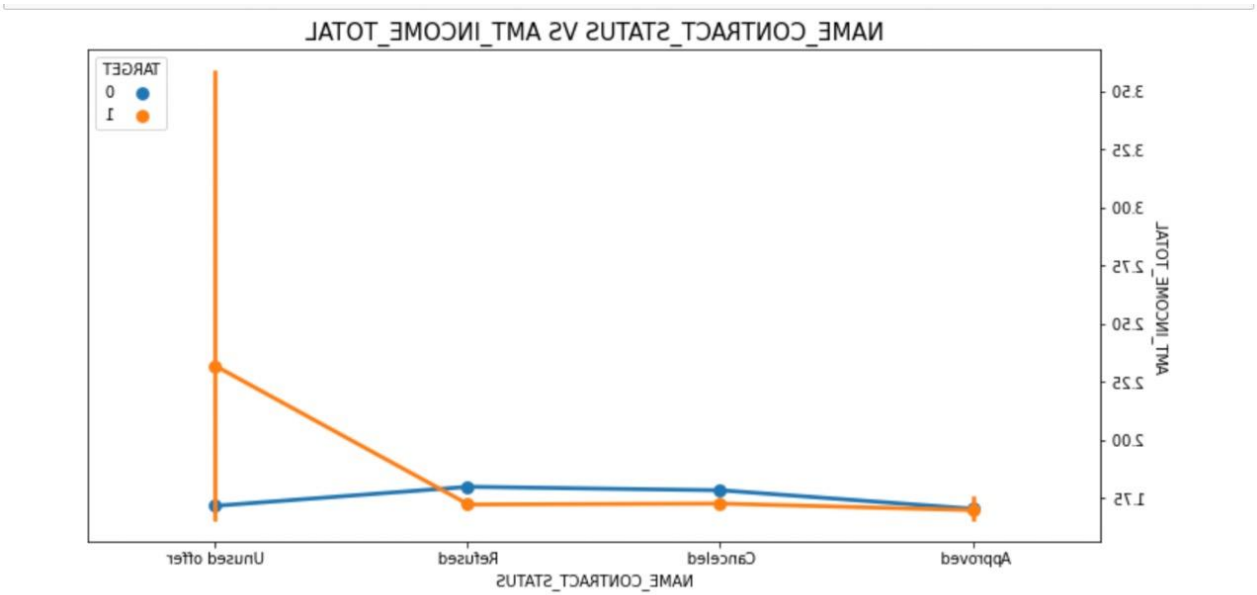
- Loan purpose has high number of unknown values (XAP, XNA)
- Loan taken for the purpose of Repairs looks to have highest default rate
- Huge number application have been rejected by bank or refused by client which are applied for Repair or Other. from this we can infer that repair is considered high risk by bank. Also, either they are rejected or bank offers loan on high interest rate which is not feasible by the clients and they refuse the loan.







		Counts	Percentage
NAME_CONTRACT_STATUS	TARGET		
Approved	0	818856	92.41%



Inferences:

90% of the previously cancelled client have actually repayed the loan. Revising 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 could turn into potential repaying customer.

The point plot show that the people who have not used offer earlier have defaulted even when there average income is higher than others

Clients who have average of 0.13 or higher their DEF_60_CNT_SOCIAL_CIRCLE score tend to default more and thus analysing client's social circle could help in disbursment of the loan.

Conclusions :

After analysing the datasets, there are few attributes of a client with which the bank would be able to identify if they will repay the loan or not. The analysis is consised as below with the contributing factors and categorization:

A. Decisive Factor whether an applicant will be Repayer:

1. NAME_EDUCATION_TYPE: Academic degree has less defaults.
2. NAME_INCOME_TYPE: Student and Businessmen have no defaults.
3. REGION_RATING_CLIENT: RATING 1 is safer.
4. ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%
5. DAYS_BIRTH: People above age of 50 have low probability of defaulting
6. DAYS_EMPLOYED: Clients with 40+ year experience having less than 1% default rate
7. AMT_INCOME_TOTAL:Applicant with Income more than 700,000 are less likely to default
8. NAME_CASH_LOAN_PURPOSE: Loans bought for Hobby, Buying garage are being repayed mostly.
9. CNT_CHILDREN: People with zero to two children tend to repay the loans.

B. Decisive Factor whether an applicant will be Defaulter:

1. CODE_GENDER: Men are at relatively higher default rate
2. NAME_FAMILY_STATUS : People who have civil marriage or who are single default a lot.
3. NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education
4. NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot.
5. REGION_RATING_CLIENT: People who live in Rating 3 has highest defaults.
6. OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as their default rate is huge.
7. 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.
8. DAYS_BIRTH: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting
9. DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.
10. CNT_CHILDREN & CNT_FAM_MEMBERS: Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.
11. AMT_GOODS_PRICE: When the credit amount goes beyond 3lakhs, there is an increase in defaulters.

C. Factors that Loan can be given on Condition of High Interest rate to mitigate any default risk leading to business loss:

1. NAME_HOUSING_TYPE: High number of loan applications are from the category of people who live in Rented apartments & living with parents and hence offering the loan would mitigate the loss if any of those default.
2. AMT_CREDIT: People who get loan for 3-6 Lakhs tend to default more than others and hence having higher interest specifically for this credit range would be ideal.
3. AMT_INCOME: Since 90% of the applications have Income total less than 3Lakhs and they have high probability of defaulting, they could be offered loan with higher interest compared to other income category.
4. CNT_CHILDREN & CNT_FAM_MEMBERS: Clients who have 4 to 8 children has a very high default rate and hence higher interest should be imposed on their loans.
5. NAME_CASH_LOAN_PURPOSE: Loan taken for the purpose of Repairs seems to have highest default rate. A very high number applications have been rejected by bank or refused by client in previous applications as well 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. The same approach could be followed in future as well.

D. NOTE:

- 90% of the previously cancelled client have actually repayed the loan. Record the reason for cancellation which might help the bank to determine and negotiate terms with these repaying customers in future for increase business opportunity.
- 88% of the clients who were refused by bank for loan earlier have now turned into a repaying client. Hence documenting the reason for rejection could mitigate the business loss and these clients could be contacted for further loans.