CREDIT EDA CASE STUDY

BY: SAHIL KUMAR

Email: sahilkumar041095@gmail.com

INTRODUCTION

For the purpose of this case study, two data sets were provided, namely:

• Application Data

• Previous Application Data

First, we took the Application Data dataset for analysis. Data cleaning was done before analysis. Following were the steps followed:

1. Found out the % of missing values in each column so as to determine which value to delete.

```
# Identifying missing values in each column
#Function to calculate null_values to identify % of data is missing in each column
def null_values(df):
    total = df.isnull().sum()
    datatypes = df.dtypes
    unique = df.nunique()
    percent = round((df.isnull().sum()*100/len(df)).sort_values(ascending = False),2)
    # percent = (df.isnull().sum()/df.isnull().count()*100)
    return pd.concat([total, datatypes, unique, percent], axis=1, keys=['Total', 'Data_Type', 'Unique', 'Percent']).sort_values(by="Percent", ascending=False)
```

2. Removed columns with more than 50% NaN values

```
# Dropping columns with High Missing Values more than 50% missing values

cols_to_keep=list(null_values_data[(null_values_data.Percent<50)].index)

cols_to_keep

application_data=app_data[cols_to_keep]
```

Analyse and Removing Unnessary columns

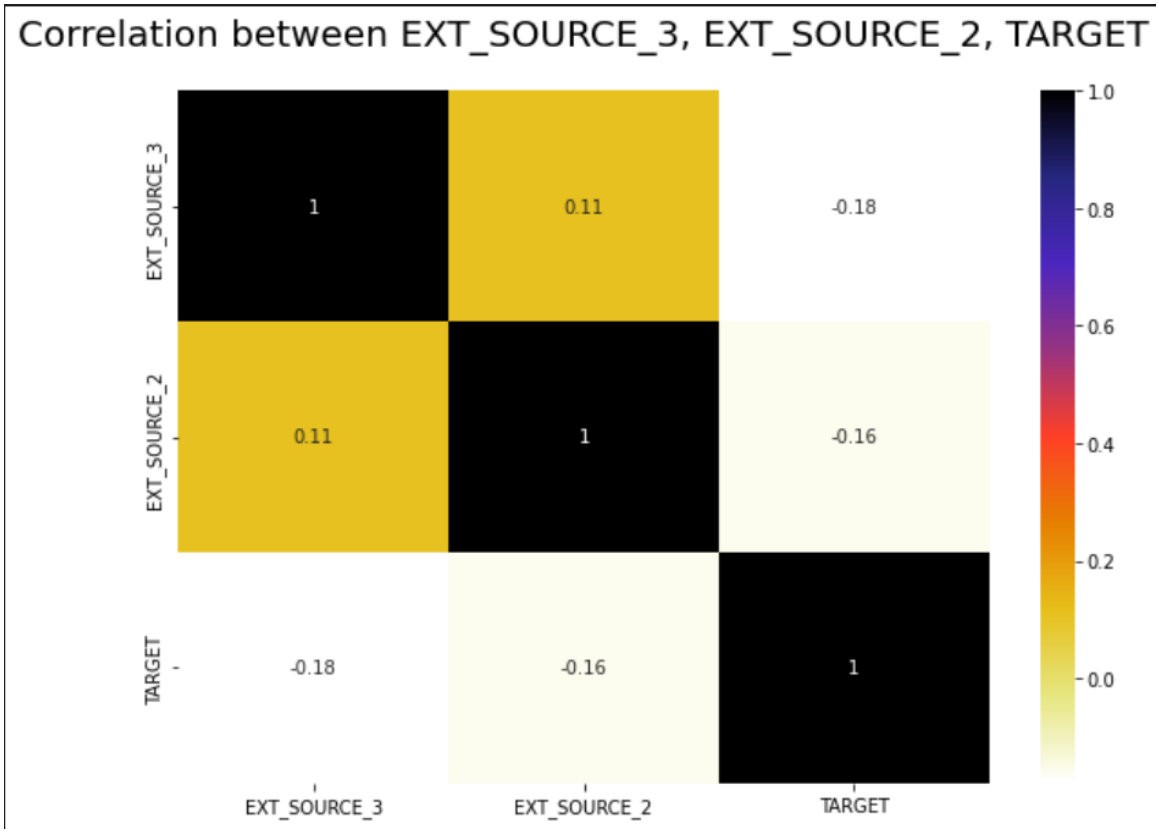
```
Plotting all the graph to find the relation and evaluting for dropping such columns
```

Observation

Columns (FLAG_OWN_REALTY, FLAG_MOBIL ,FLAG_EMP_PHONE, FLAG_CONT_MOBILE, FLAG_DOCUMENT_3) have more repayers than defaulter and from these keeping FLAG_DOCUMENT_3,FLAG_OWN_REALTY,

FLAG_MOBIL more sense thus we can include these columns and remove all other FLAG columns for furhter analysis.

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



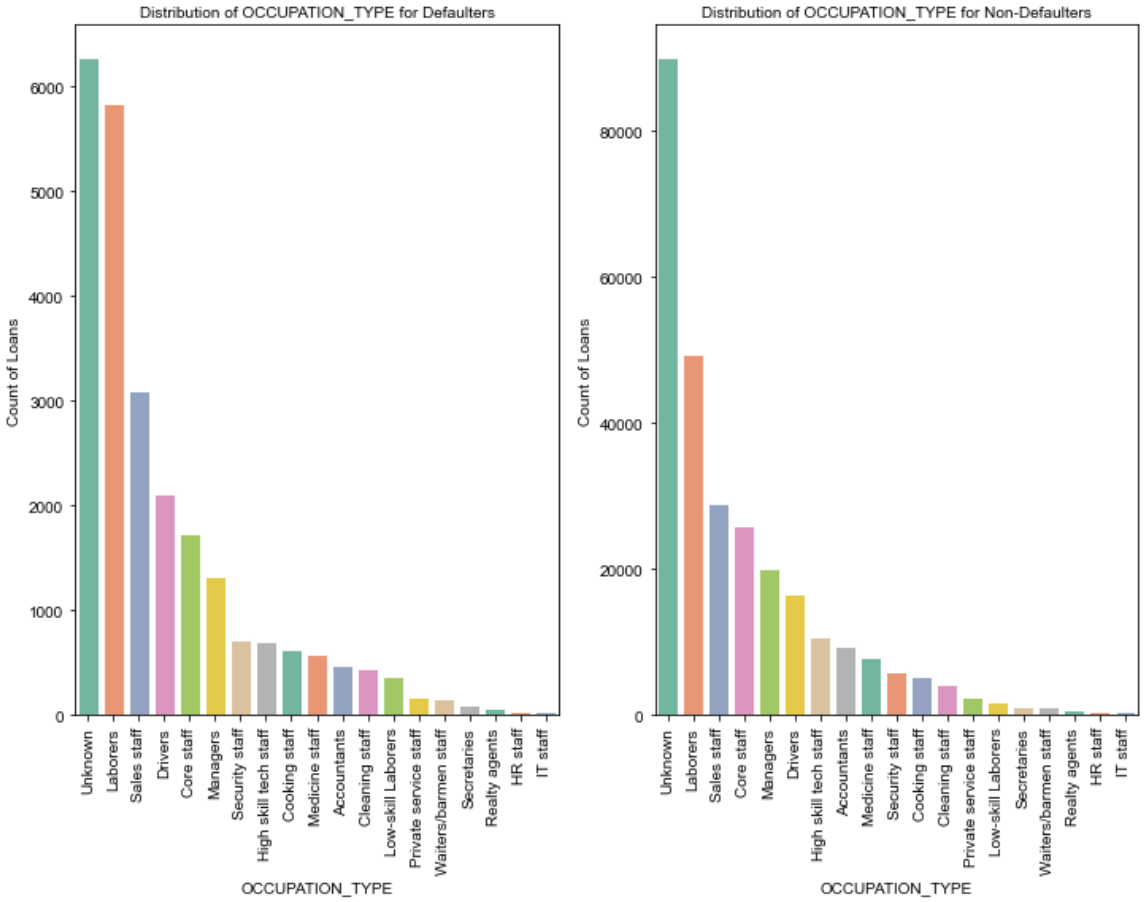
Observation:

*** **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.****

Imputing values

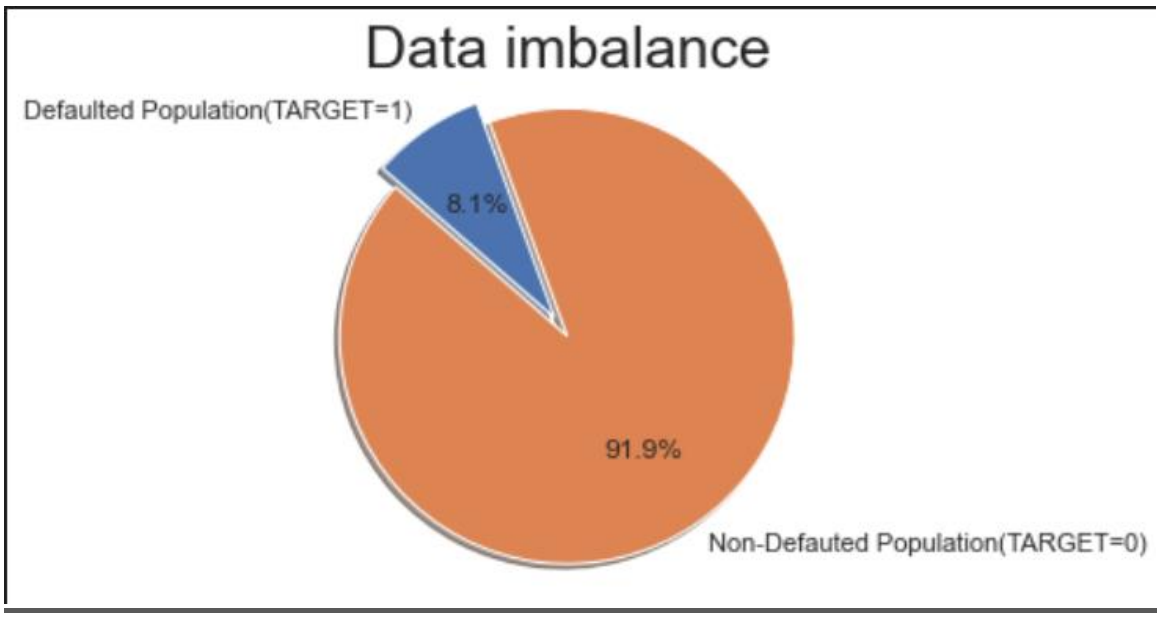
Imputing for "OCCUPATION_TYPE" column



Observation:

We can see that Distribution of OCCUPATION_TYPE for Defaulters and for Non-Defaulters.

Imbalanced Data



Observation

- application_data dataframe that is application data is highly imbalanced. Defaulted population is 8.1 % and non- defaulted population is 91.9% .Ratio is 11.3
- We will separately analyse the data based on the target variable for a better understanding.

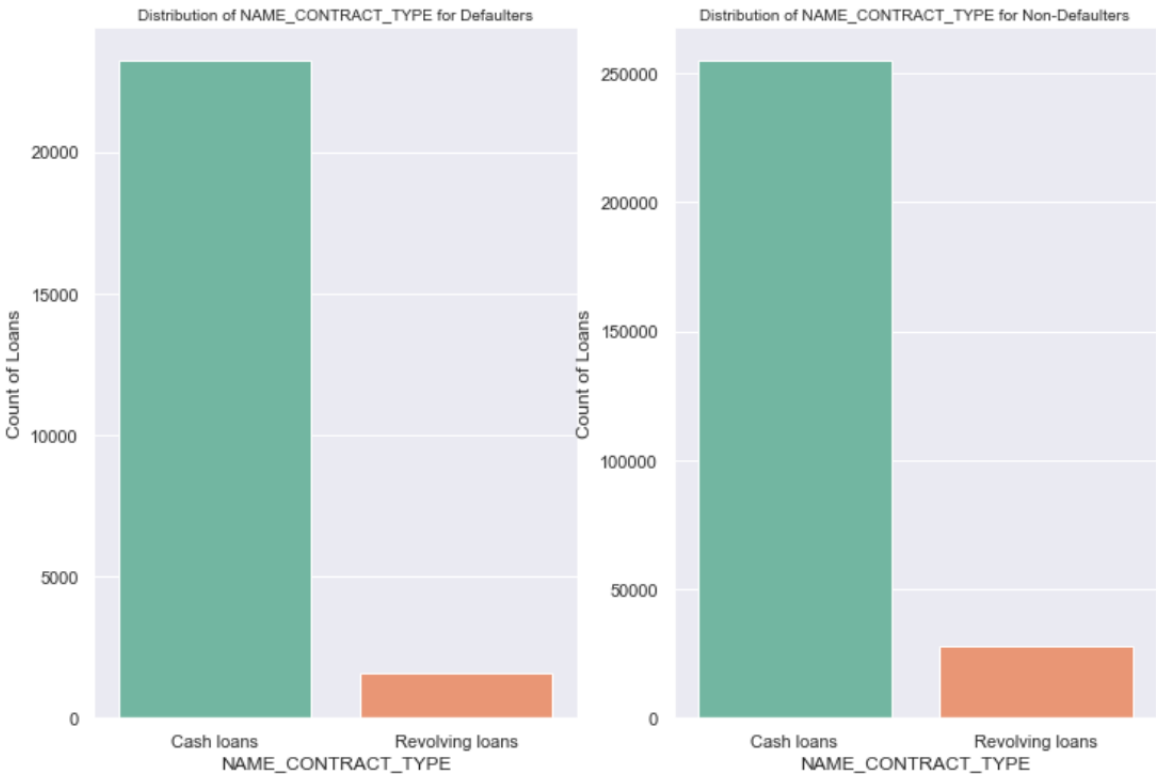
Bi-Variate Analysis of Variables

Gender Distribution in Target

Observation

- It seems like Female clients applied higher than male clients for loan
- 66.6% Female clients are non-defaulters while 33.4% male clients are non-defaulters.
- 57% Female clients are defaulters while 42% male clients are defaulters.

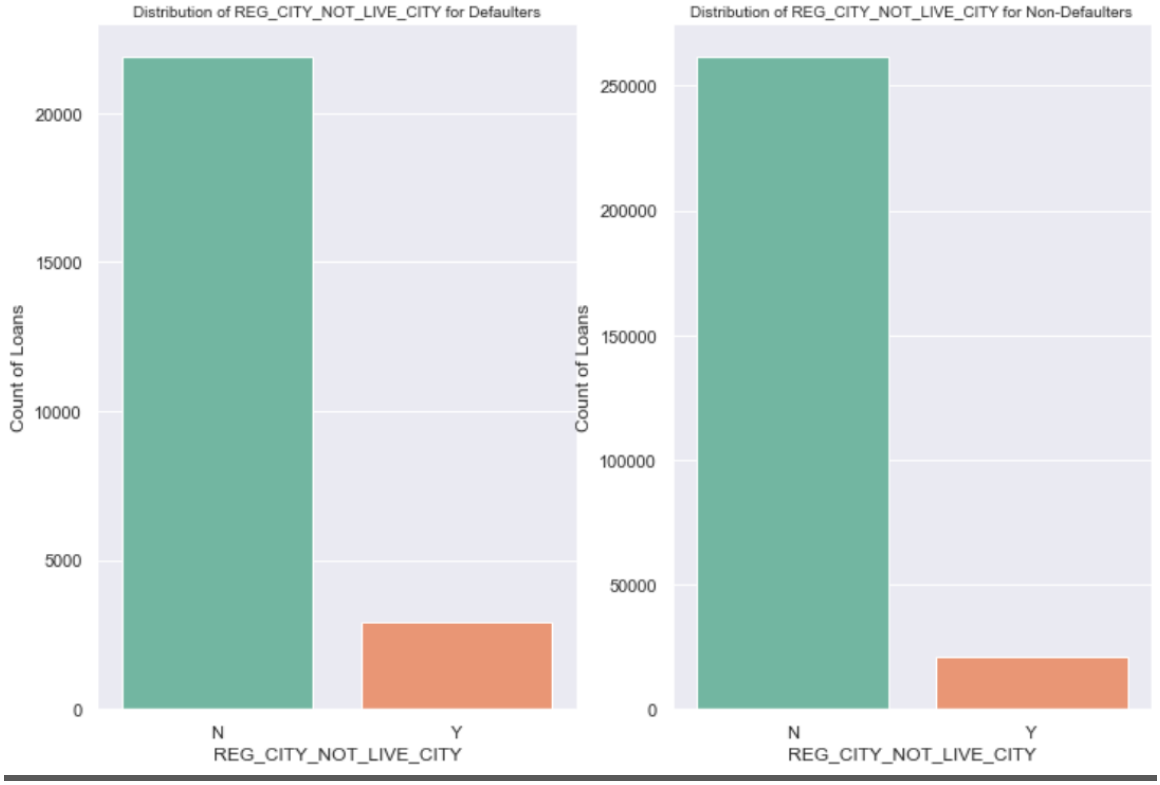
Distribution of NAME_CONTRACT_TYPE for Defaulters and Non-Defaulters

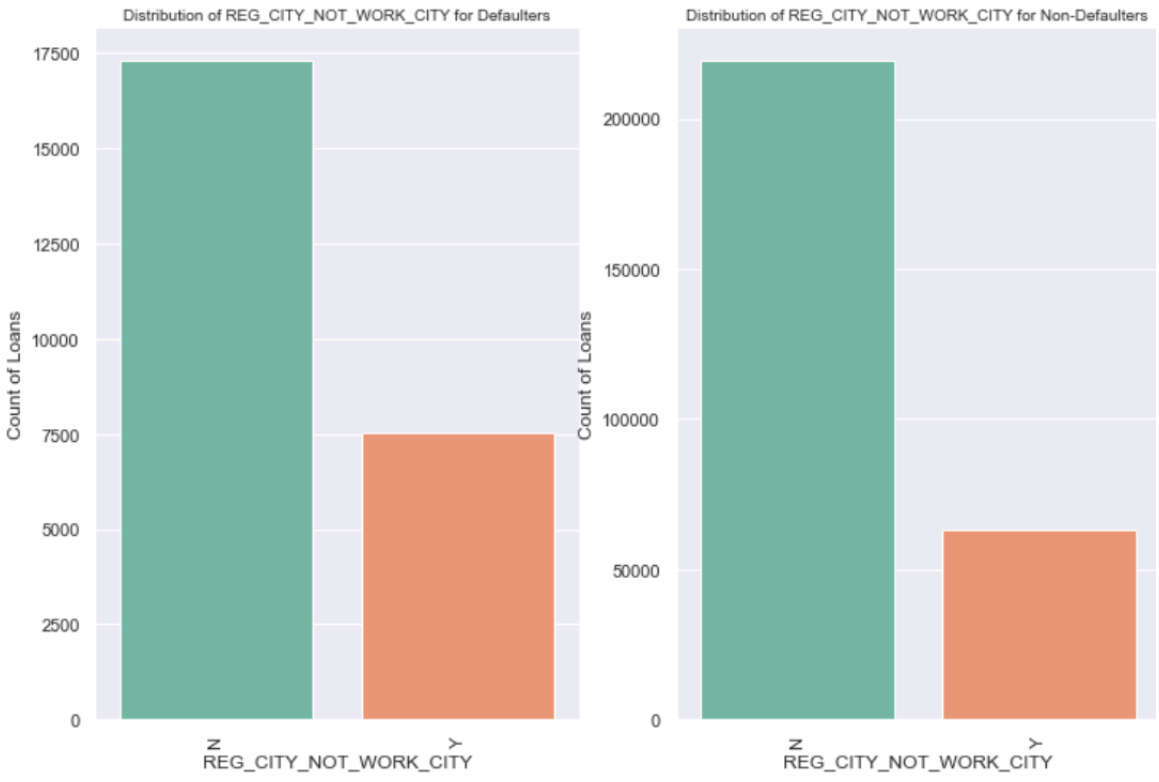


Observation:

We observe that the number of Cash loans is much higher than the number of Revolving loans for both Target = 0 and Target = 1

Distribution of REG_CITY_NOT_LIVE_CITY for Defaulters and Non-Defaulters



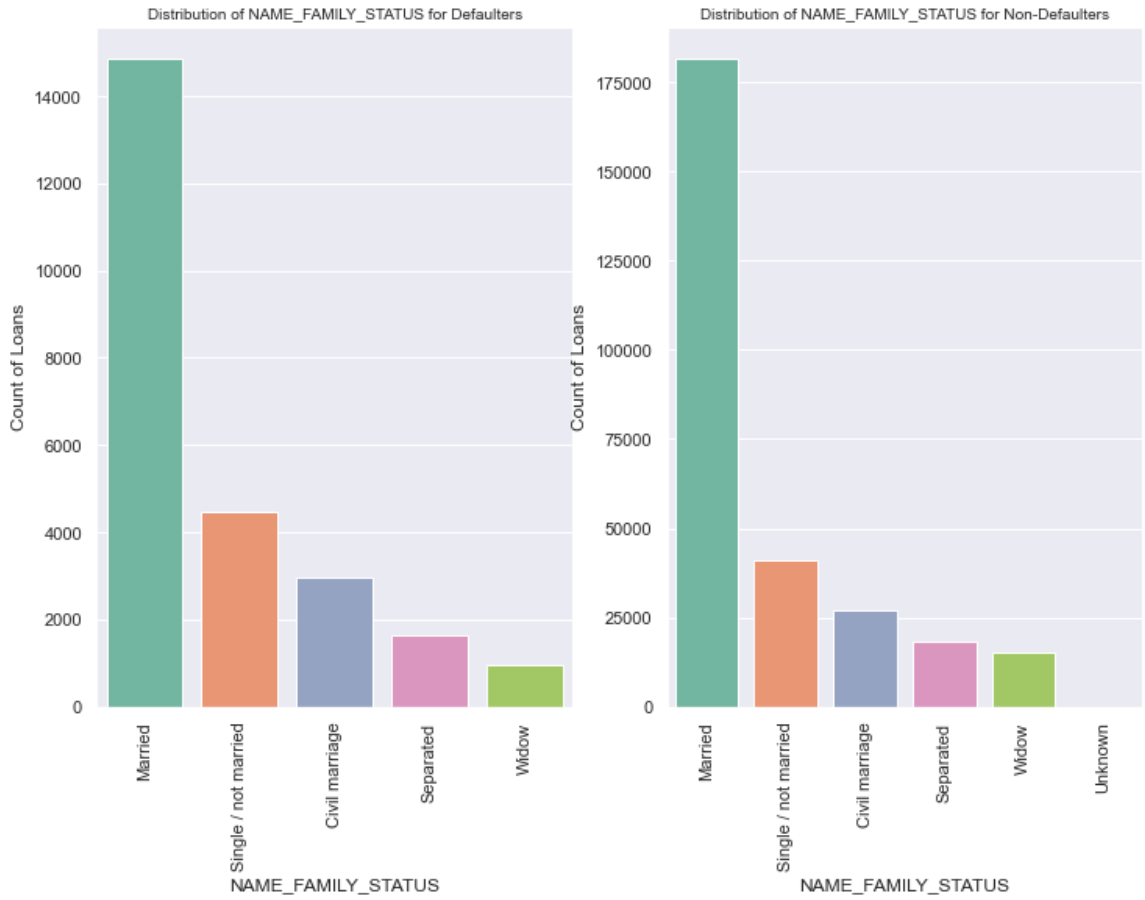


Observation:

We observe that the Ratio of people whose Registration City is not the same as live city or work city is higher in case of defaulters are compared to defaulters.

It tells us that people who live or work in a city different than the registration city are more likely to have payment difficulties.

Distribution of NAME_FAMILY_STATUS for Defaulters and Non-Defaulters

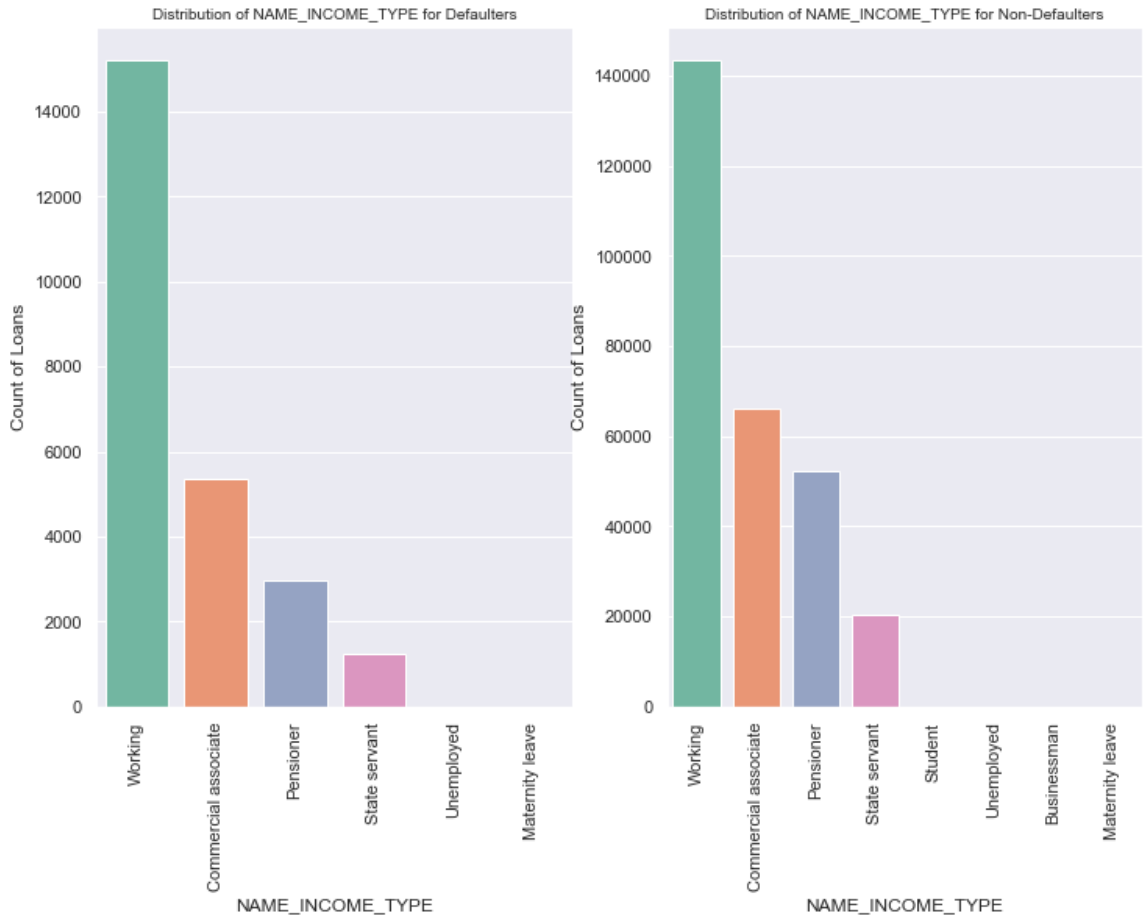


Observation:

Ratio of Single/Unmarried people is more in the left graph.

Single/Unmarried people are more likely to have payment difficulties

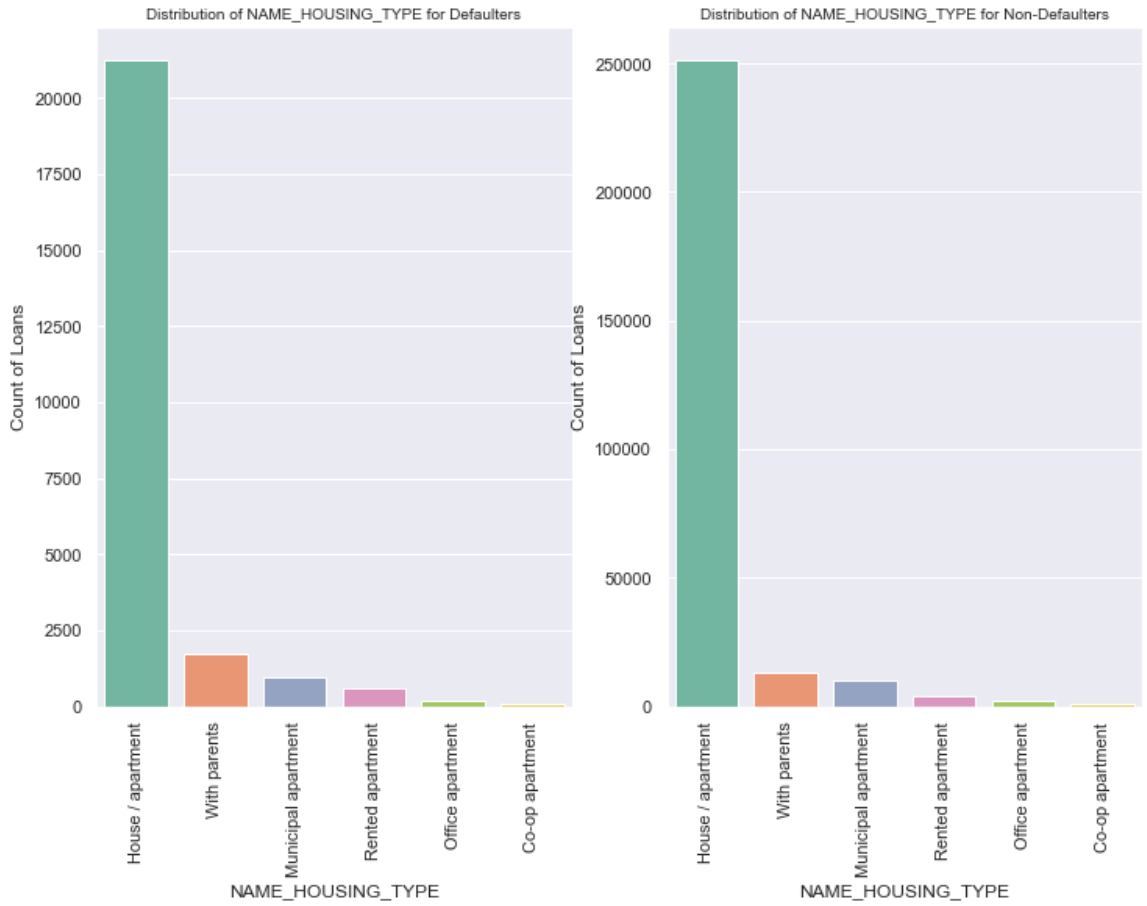
Distribution of NAME_INCOME_TYPE for Defaulters and Non-Defaulters



Observation:

Commercial associates, Pensioner, State Servants have a higher ratio to total in non-defaulters.

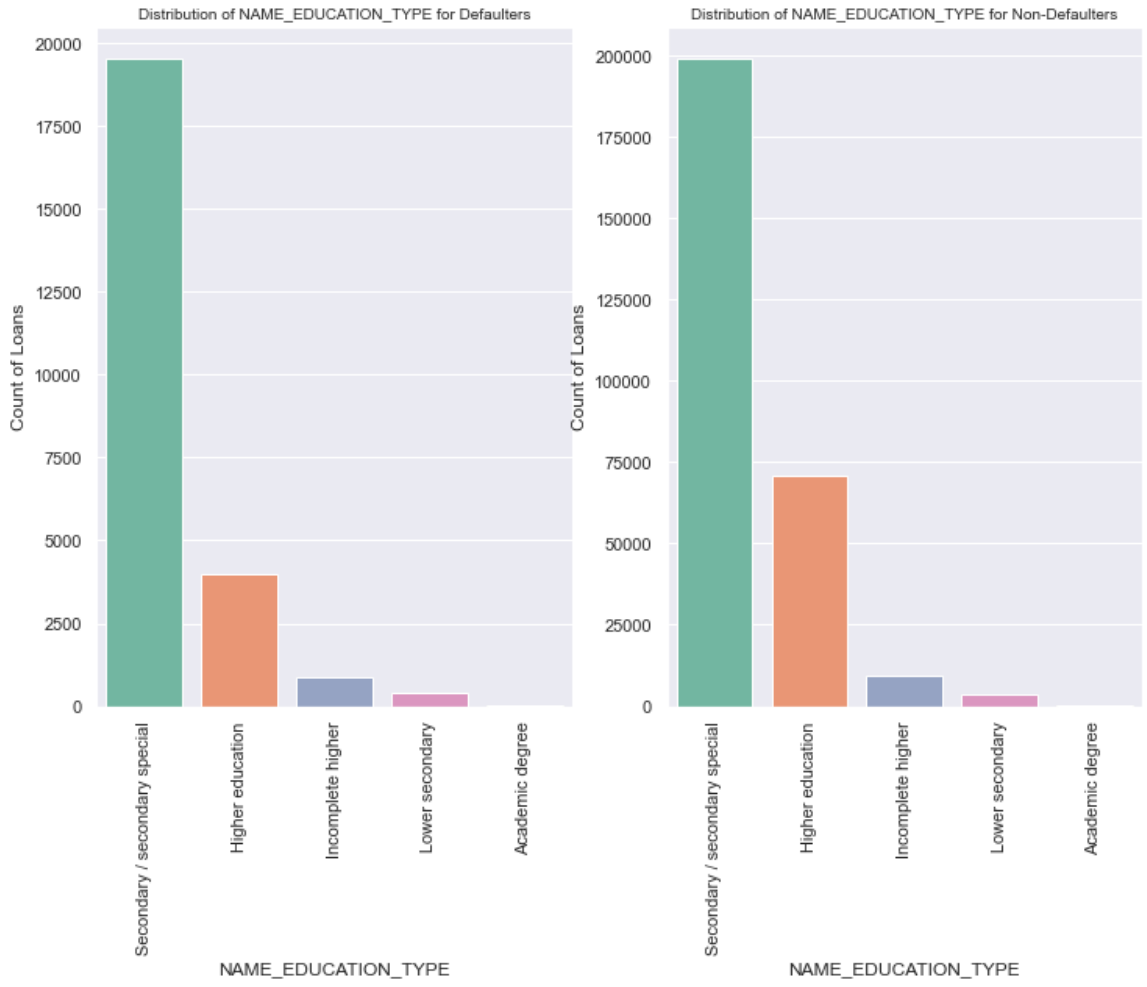
Distribution of NAME_HOUSING_TYPE for Defaulters and Non-Defaulters



Observation:

Most people live in a House/Apartment Ratio of People who live With Parents is more for defaulter than non-defaulters. It tells us that applicant who live with parents have a higher chance of having payment difficulties.

Distribution of NAME_EDUCATION_TYPE for Defaulters and Non-Defaulters

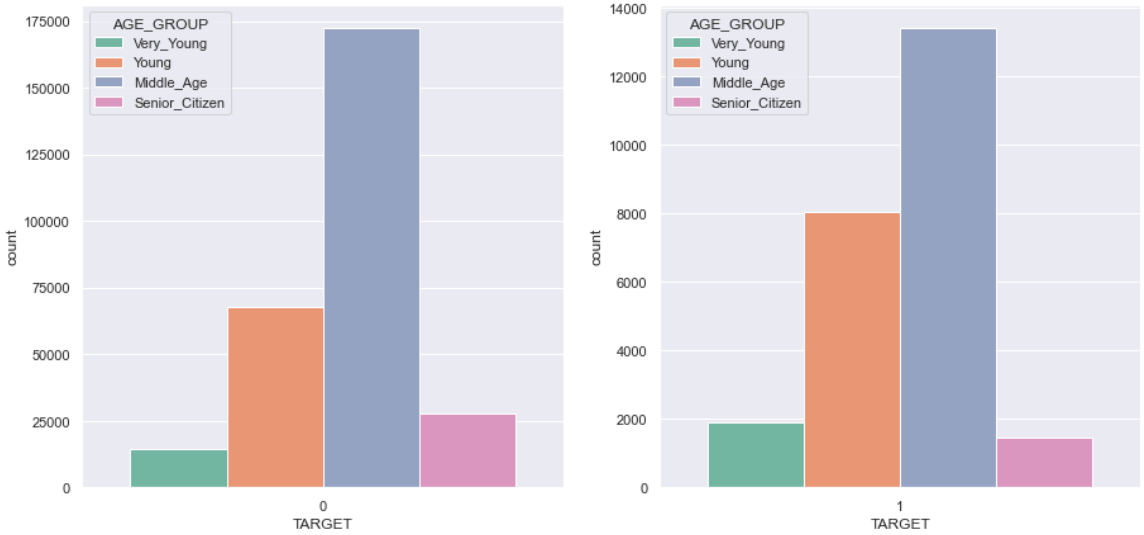


Observation:

While the category with highest count remains same.

- This chart tells us that people with Academic Degree rarely take loans and are rarely defaulters. So they are potentially good customers.
- People with higher education are less likely to have payment difficulties. The Ratio is higher for non-defaulters than defaulters.

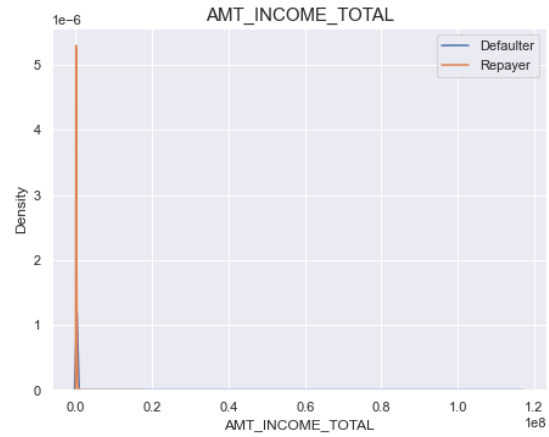
Age Distribution based on Target 0 and Target 1

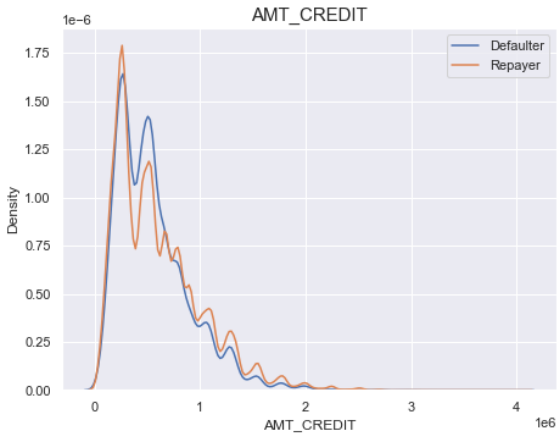


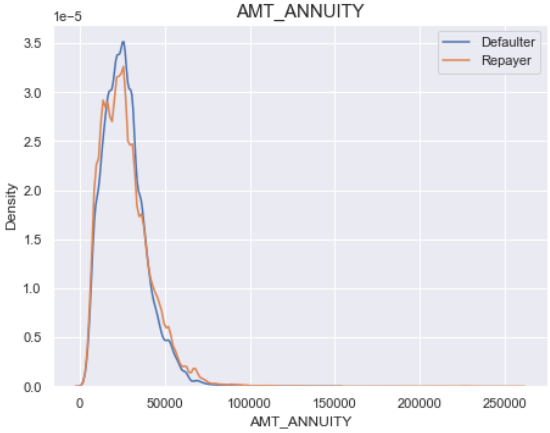
Observation:

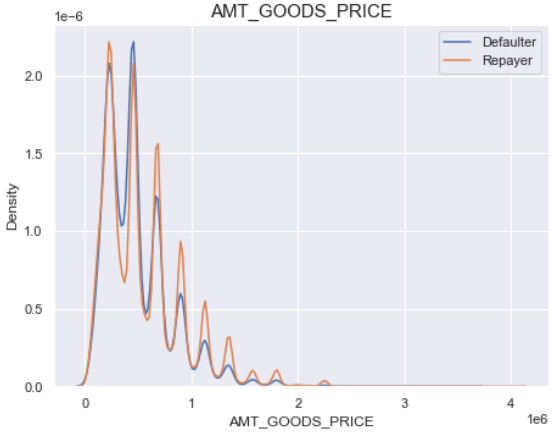
- Middle Age(35-60) the group seems to applied higher than any other age group for loans in the case of Defaulters as well as Non-defaulters.
- Also, Middle Age group facing paying difficulties the most.
- While Senior Citizens(60-100) and Very young(19-25) age group facing paying difficulties less as compared to other age groups.

Numerical Univariate Analysis







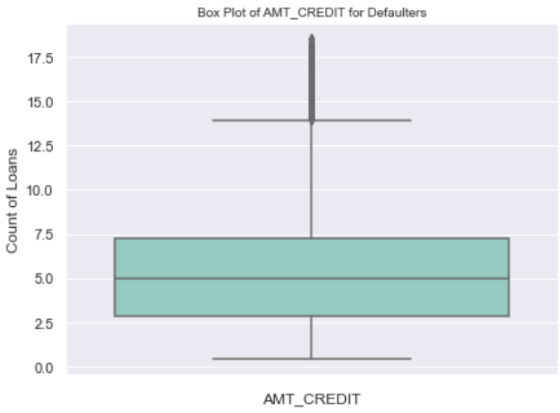


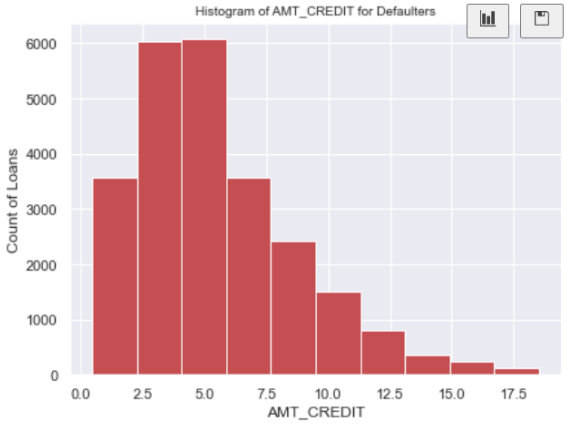
Observation:

- 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
- 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

Identifying Outliers

Based on AMT_CREDIT

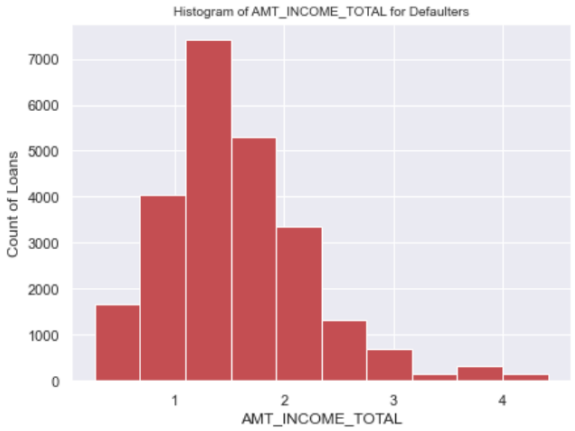




Observation:

we observe that the credit amount lies between 2.5 to around 5.0 for defaulters.

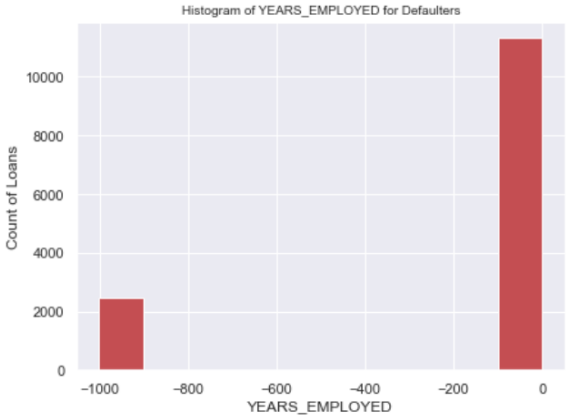
Based on AMT_INCOME_TOTAL



Observation:

This tell us that most people with payment have incomes in the lower range between 1 to 2 which some on the higher end some on the lower

Based on YEARS_EMPLOYED

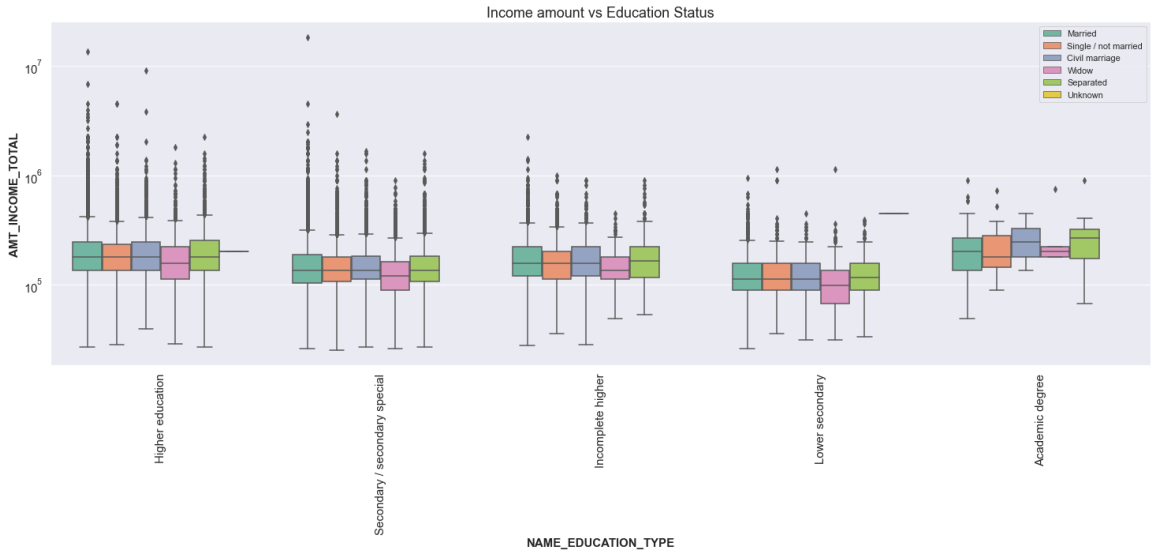


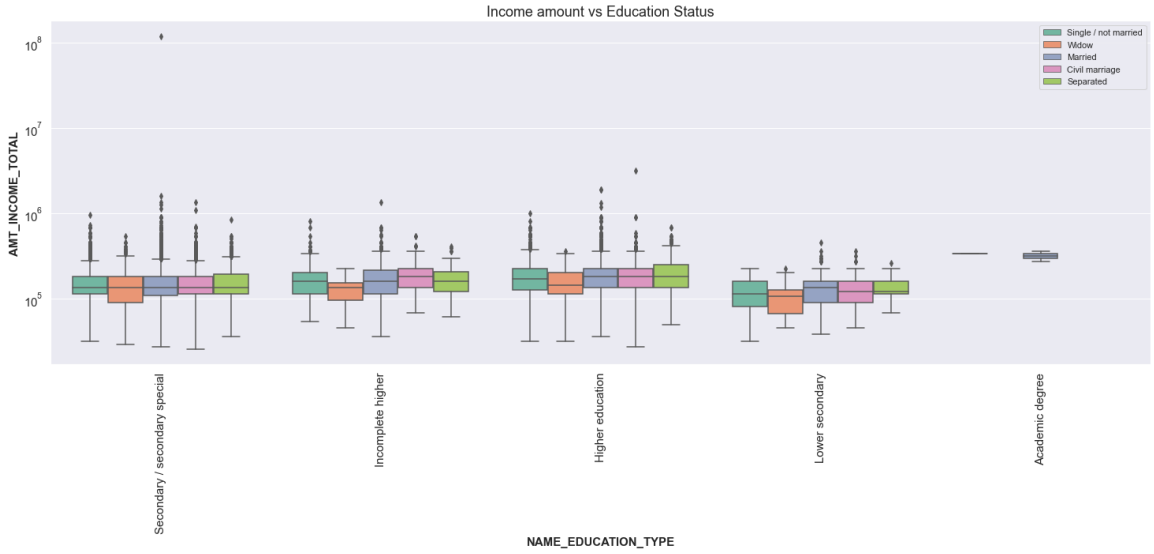
Observation

A large number of entries have 0 which means, a lot of people don't work.

Multi-Variate Analysis of Variables

AMT_INCOME_TOTAL Vs NAME_EDUCATION_TYPE Vs NAME_FAMILY_STATUS for no payment and payment difficulties client (Target 0 and Target 1)





Observation:

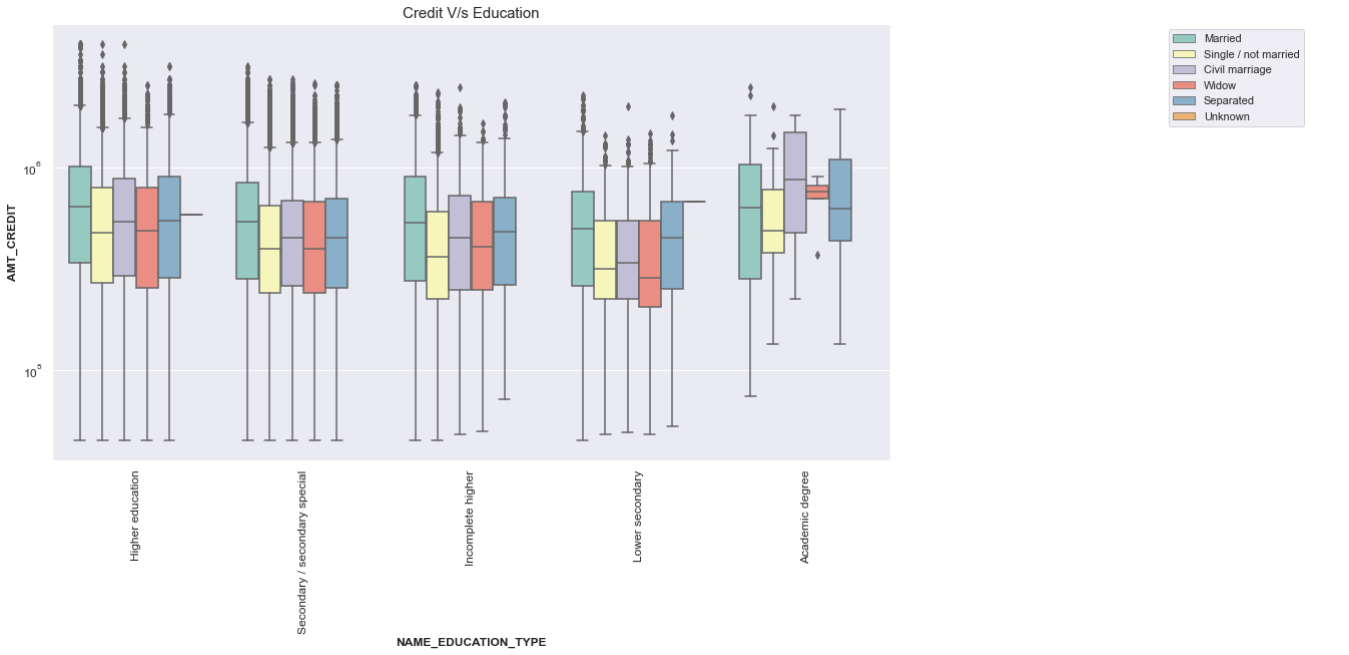
for TARGET 0

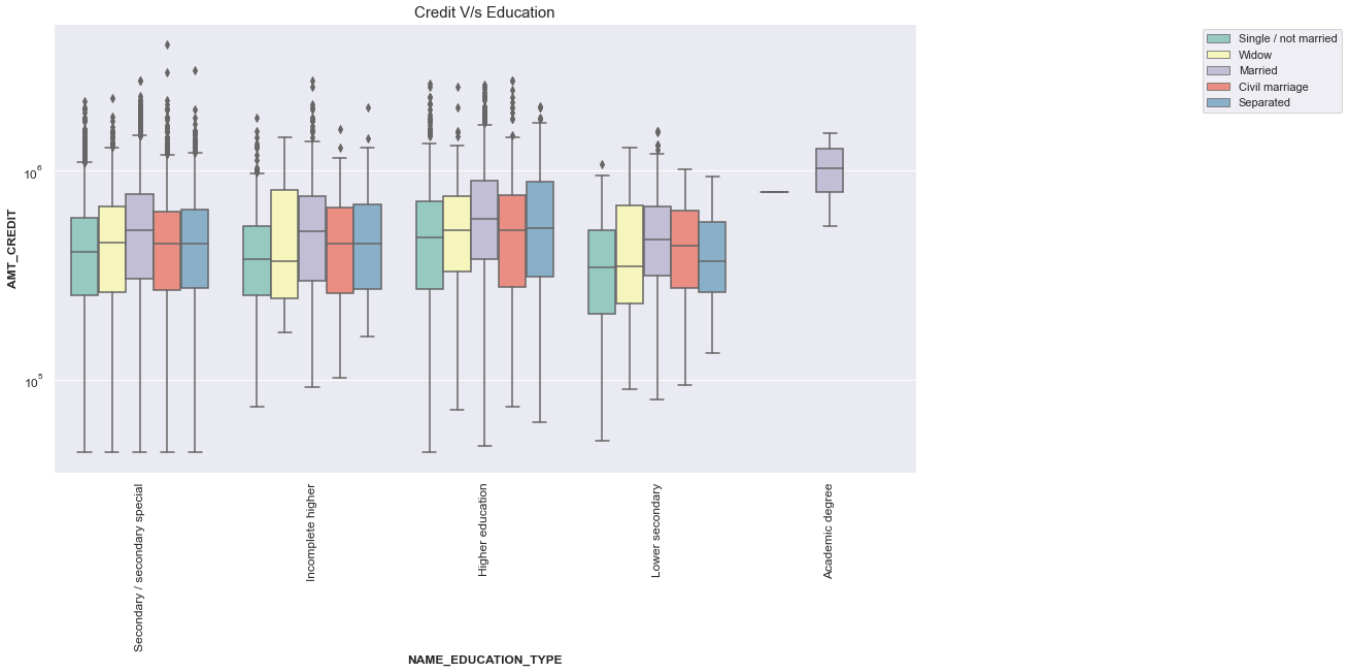
- Widow Client with Academic degree have very few outliers and doesn't have First and Third quartile. Also, Clients with all types of family statuses having academic degrees have very less outliers as compared to other types of education.
- Income of the clients with all types of family status having rest of the education type lie Below the First quartile i.e. 25%
- Clients having Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special have a higher number of outliers.
- From the above figure, we can say that some of the clients having Higher Education tend to have the highest income compared to others.
- Though some of the clients who haven't completed their Higher Education tend to have higher incomes.
- Some of the clients having Secondary/Secondary Special Education tend to have higher incomes.

for TARGET 1

- The income amount for Married clients with an academic degree is much lesser as compared to others.
- (Defaulter) Clients have relatively less income as compared to Non-defaulters.

NAME_EDUCATION_TYPE Vs AMT_CREDIT Vs NAME_FAMILY_STATUS for no payment and payment difficulties client (Target 0 and Target 1)





Observation

for TARGET 0

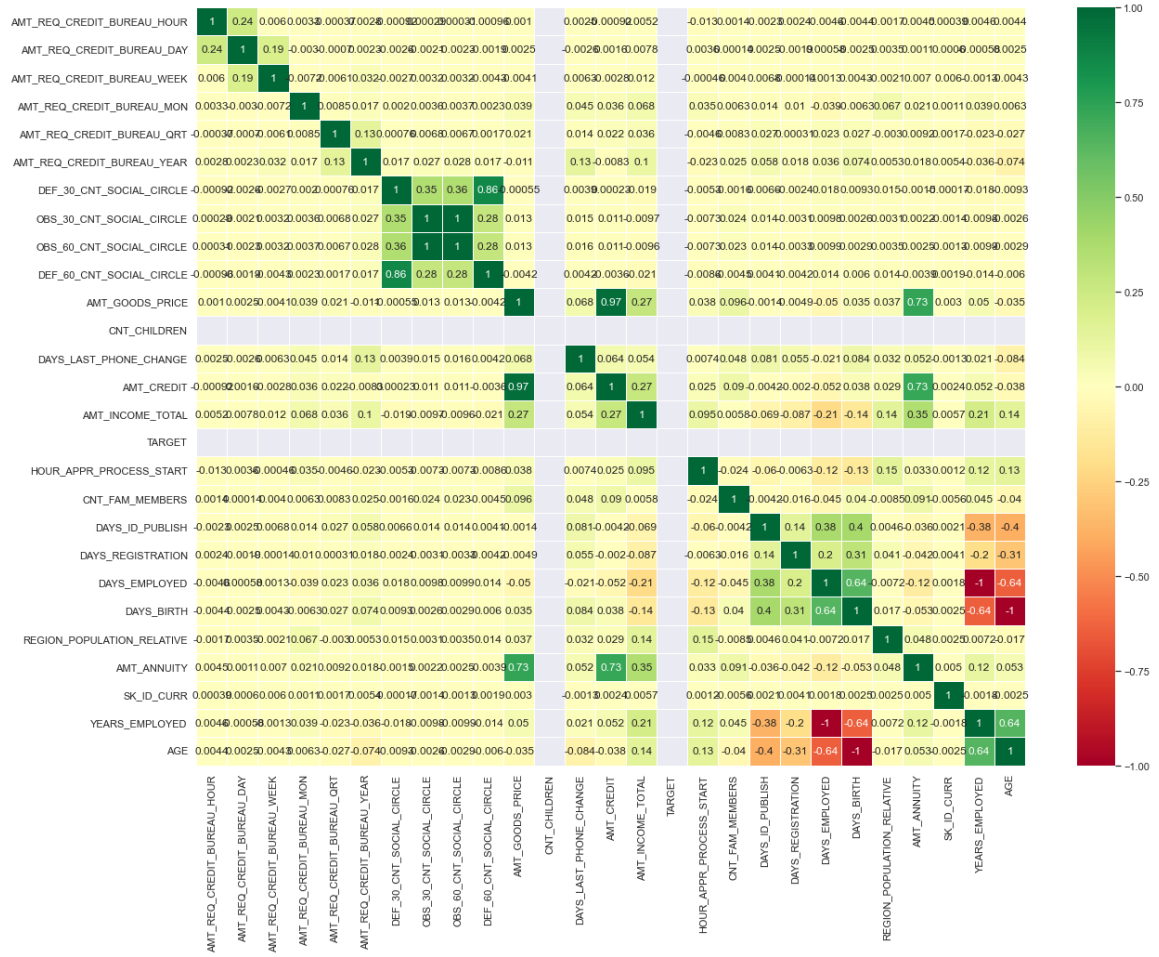
- Clients with different Education types except Academic degrees have a large number of outliers**
- Most of the population i.e. clients' credit amounts lie below 25%.
- Clients with an Academic degree and who is a widow tend to take higher credit loan.**
- Some of the clients with Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special Education are more likely to take a high amount of credit loans.

for TARGET 1

- Married client with academic applied for a higher credit loan. And doesn't have outliers. Single clients with academic degrees have a very slim boxplot with no outliers.
- Some of the clients with Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special Education are more likely to take a high amount of credit loans.

Correlation between numeric variable

Correlation with nondefaulters or Repayers



Observation Correlating factors amongst repayers

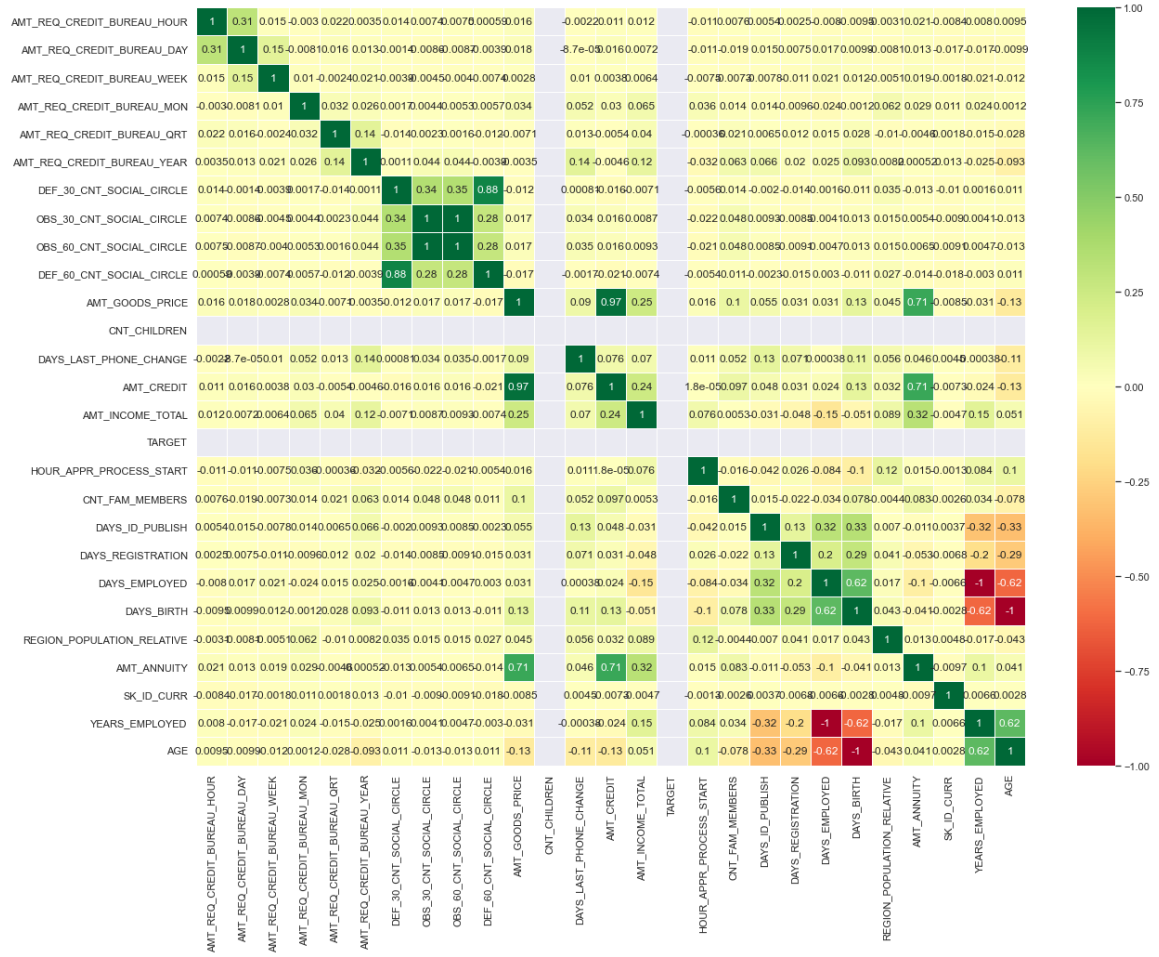
1. AMT_ANNUITY (Loan Annuity) amount is highly correlated with:

- **Goods Price Amount**
- **Amount Credit**
- **Total Income upto some extent**

2. We can also see that repayers have high correlation in number of days employed.

8.2 correlation with defaulters

Correlation with defaulters



Observation: Correlating factors amongst repayers

- Credit amount is highly correlated with good price amount which is same as repayers.
- Loan annuity correlation with credit amount has slightly reduced in defaulters(0.71) when compared to repayers(0.72)
- We can also see that repayers have high correlation in number of days employed(0.64) when compared to defaulters(0.61).
- Days_birth and number of children correlation has reduced to 0.25 in defaulters when compared to 0.33 in repayers.
- There is a slight increase in defaulted to observed count in social circle among defaulters(0.26) when compared to repayers(0.25)

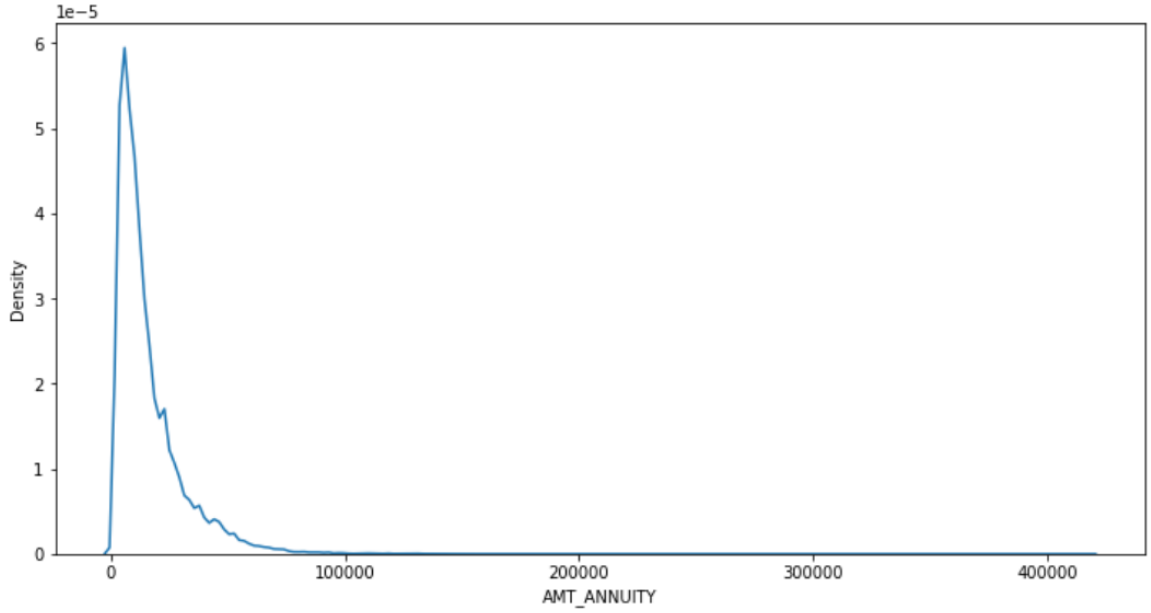
Conclusion

- Now that I have gained insight into the dataset ie performed an Exploratory Data Analysis, So summarizing what I have learnt and did in this case study,
- I have extensively covered pre-processing steps required to analyze data
- I have covered Null value imputation methods
- I have also covered step by step analyzing techniques such as Univariate analysis, Bivariate analysis, Multivariate analysis, etc
- Now we are left with 0 missing values and clead the data based on the scenario what I have explained and did analysis in the above.
- Now We are ready for model training for further analysis to train.

PREVIOUS APPLICATION ANALYSIS

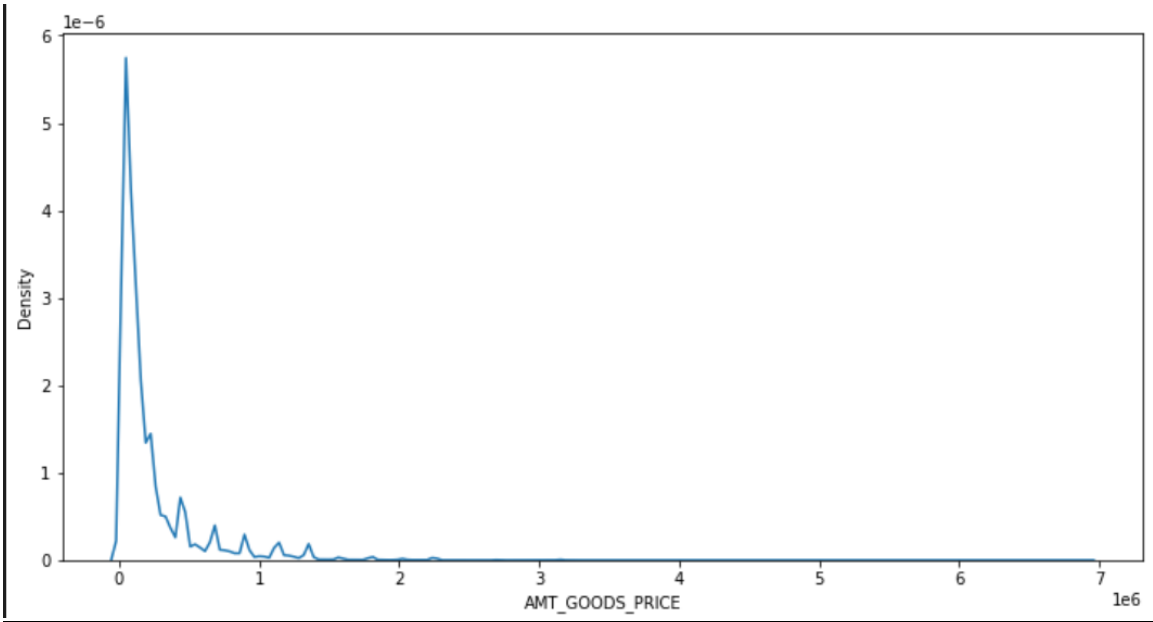
Then we moved on to analysis of the second data set. We performed few data cleaning steps and then moved on to analyzing the data.

Dealing with continuos variables "AMT_ANNUITY", "AMT_GOODS_PRICE"



Observation:

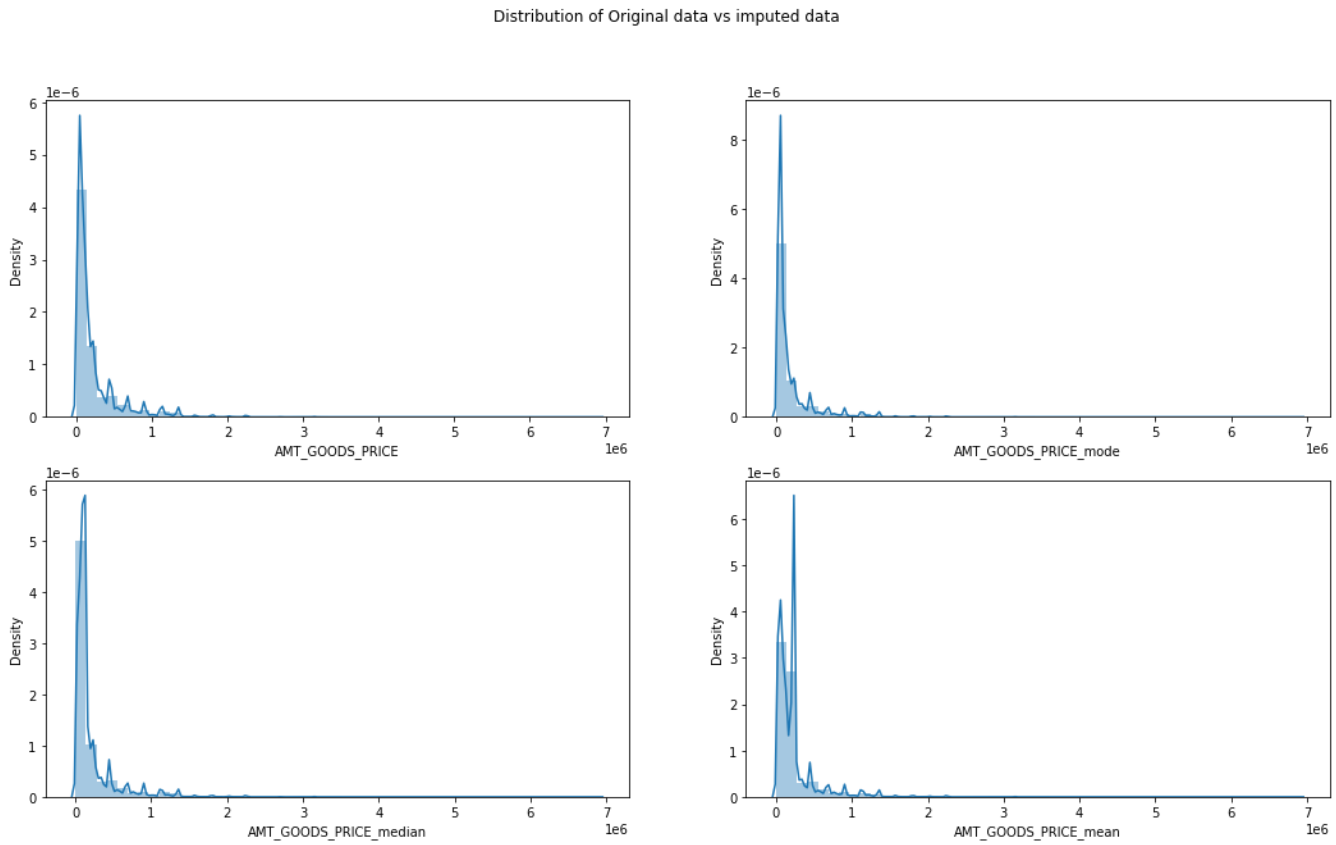
- **There is a single peak at the left side of the distribution and it indicates the presence of outliers and hence imputing with mean would not be the right approach and hence imputing with median.**



Observation

- **There are several peaks along the distribution which is right skewed.
- **Now imputing using the mode, mean and median and see if the distribution is still about the same.**

Creating new dataframe for "AMT_GOODS_PRICE" with columns imputed with mode, median and mean

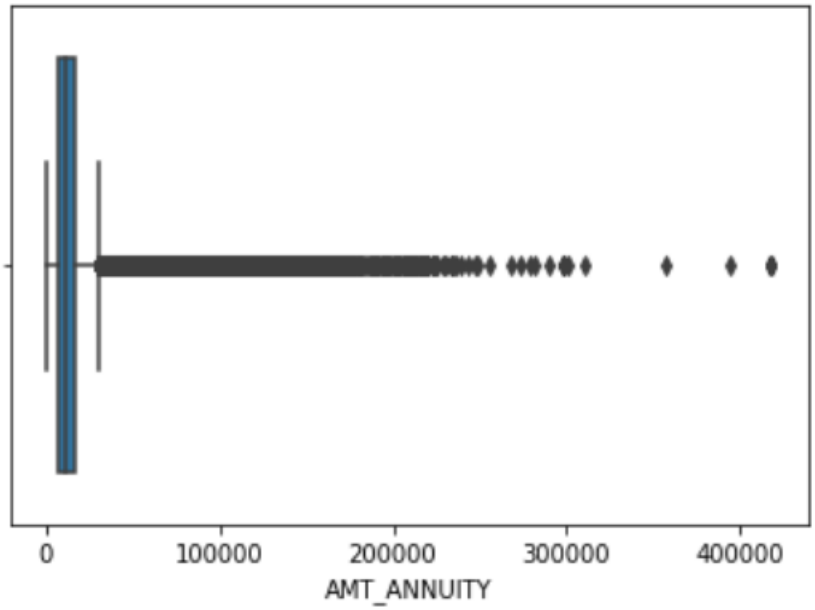


Observation

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

Identifying Outliers

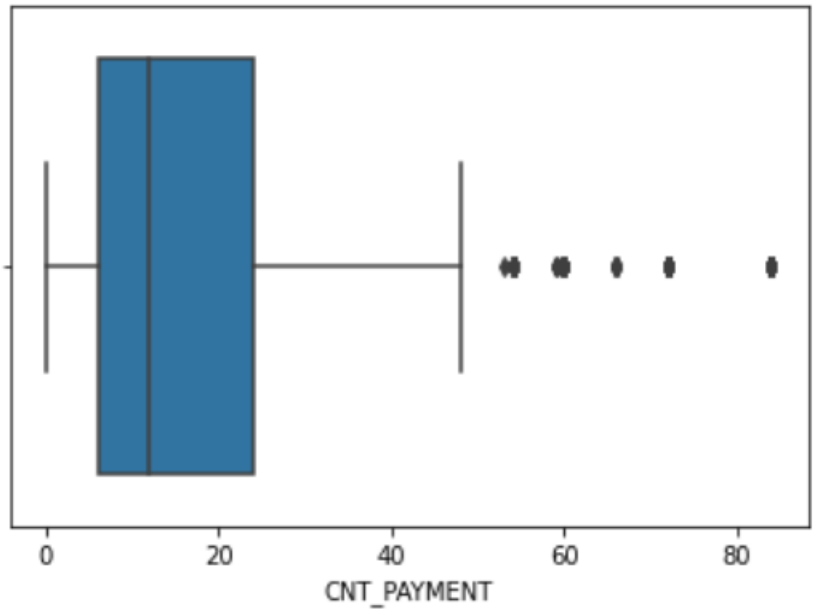
Based on AMT_ANNUITY



Observation

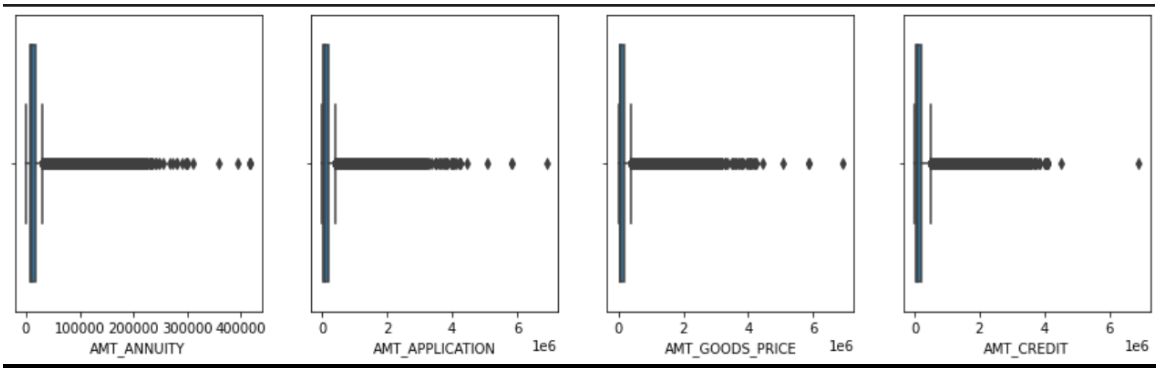
Since the column has many outliers hence imputing with median is the best method.

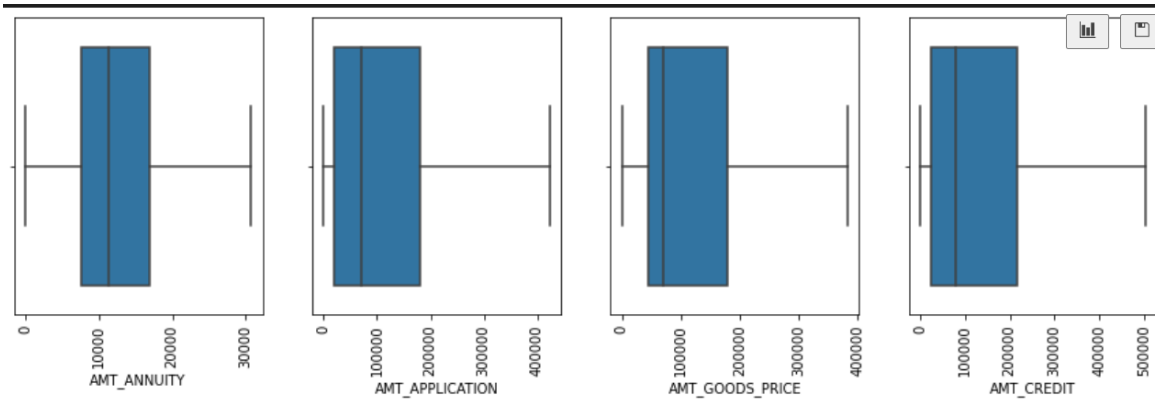
Based on CNT_PAYMENT



Observation

There are some outliers, however it lies between the 5-22 in range of quantile of 0.25-0.75 since the column has outliers hence imputing with median is the best method.

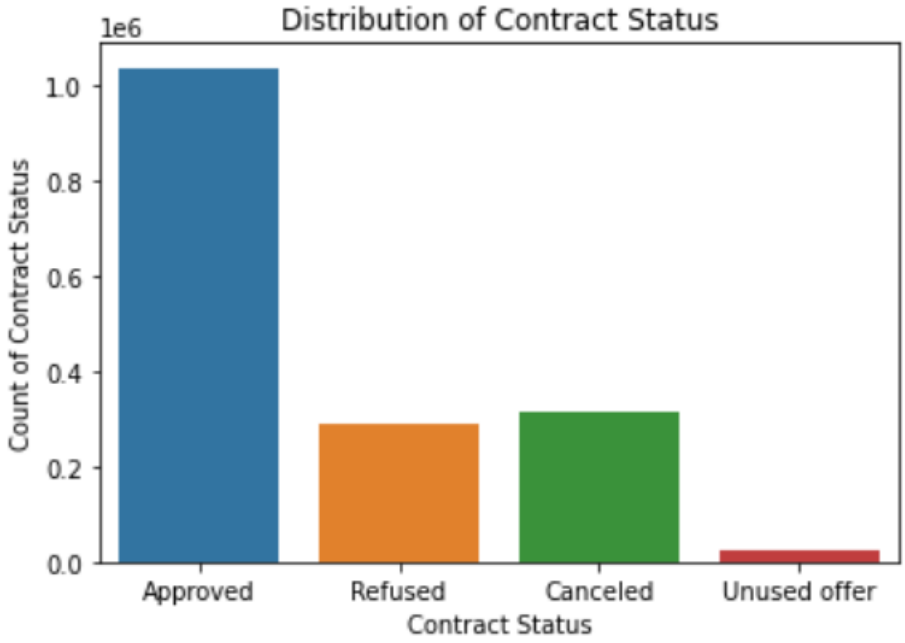




Observation

- Now we have deal with the outliers and removed it based on analysing the above situation and left with no outliers for these data AMT_ANNUITY, AMT_APPLICATION, AMT_GOODS_PRICE, AMT_CREDIT

Distribution of Contract Status

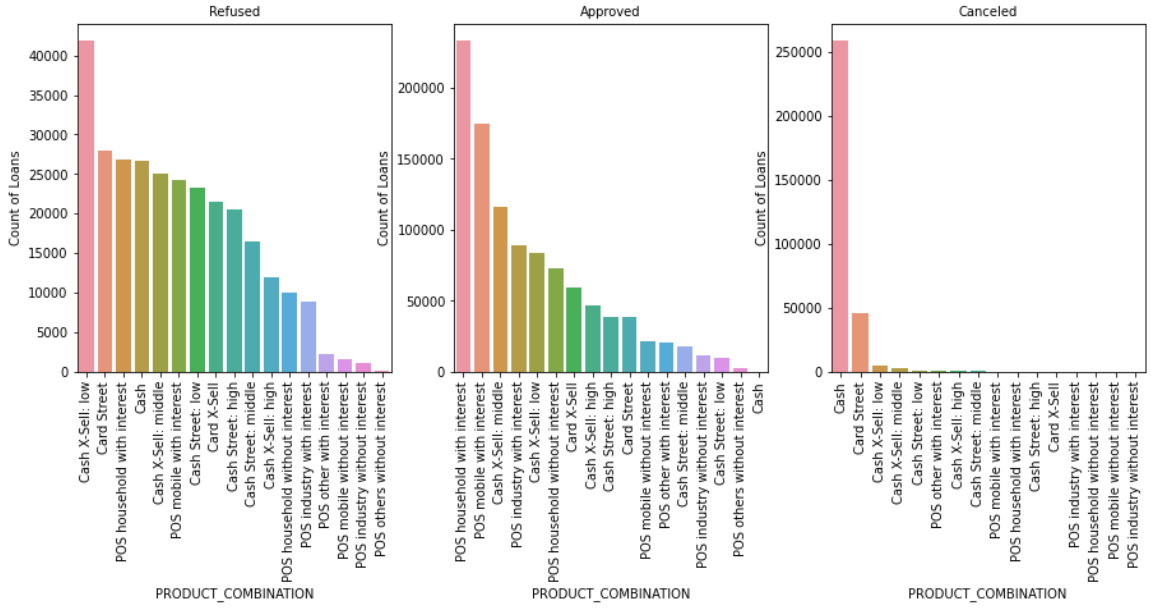


Observation

- High distribution of contract status is Approved whereas lower one is Unused offer

Bi-Variate Analysis of Variables

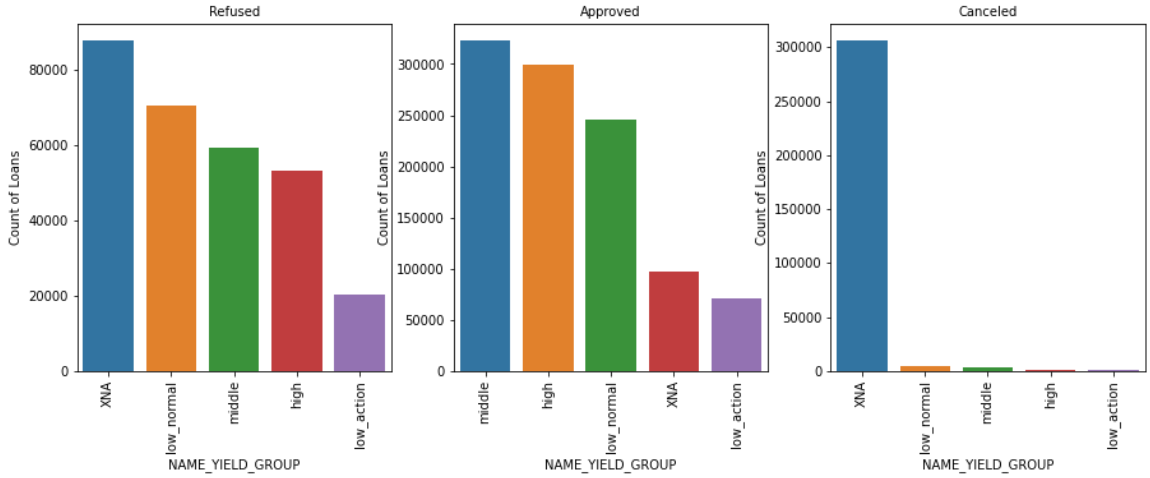
Based on PRODUCT_COMBINATION



Observation

- We observe most number of loans were approved for POS household with interest.
- Most number of refused loans were of Cash X-Sell: Low Product combination
- Most Canceled loans were Cash loans

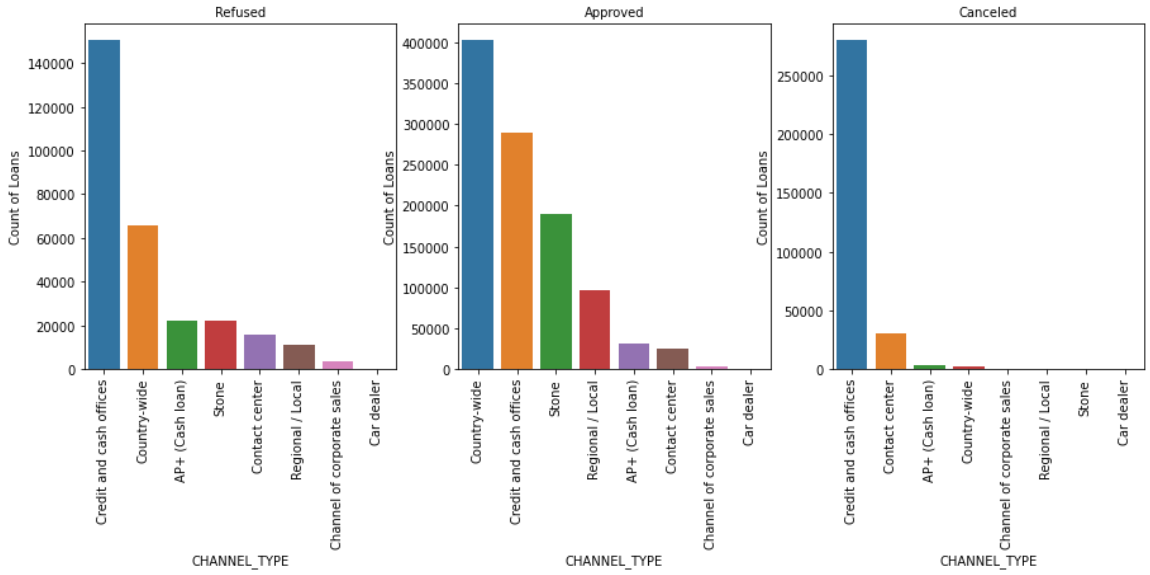
Based on NAME_YIELD_GROUP



Observation

- Most approved loans were from Middle Yield Goup
- Most refused loans were from Yield Goups Not specified

Based on CHANNEL_TYPE

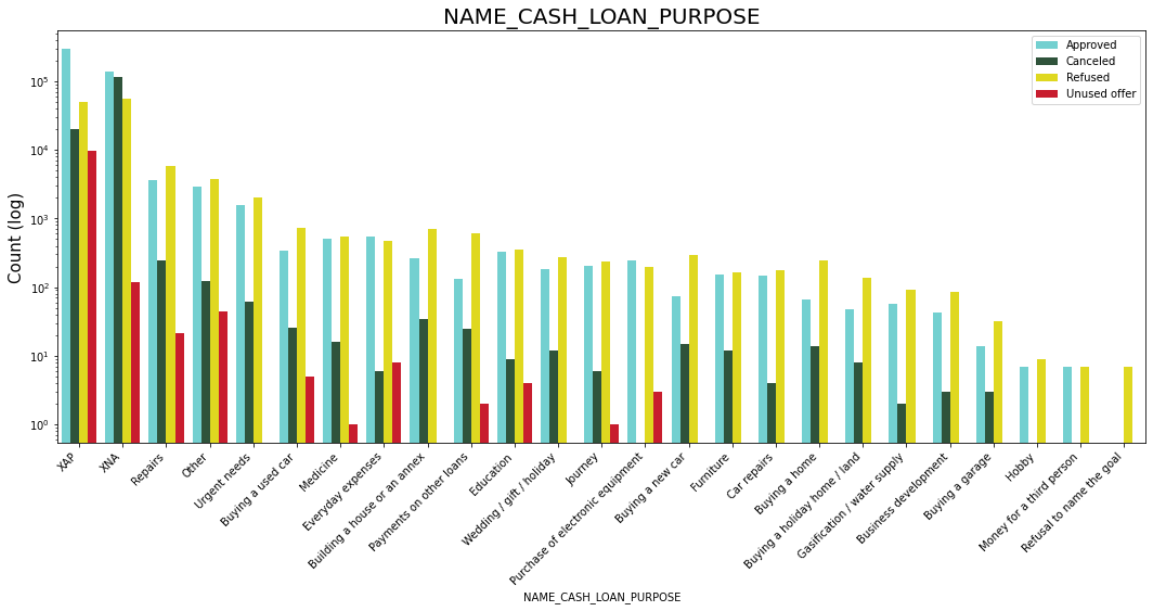


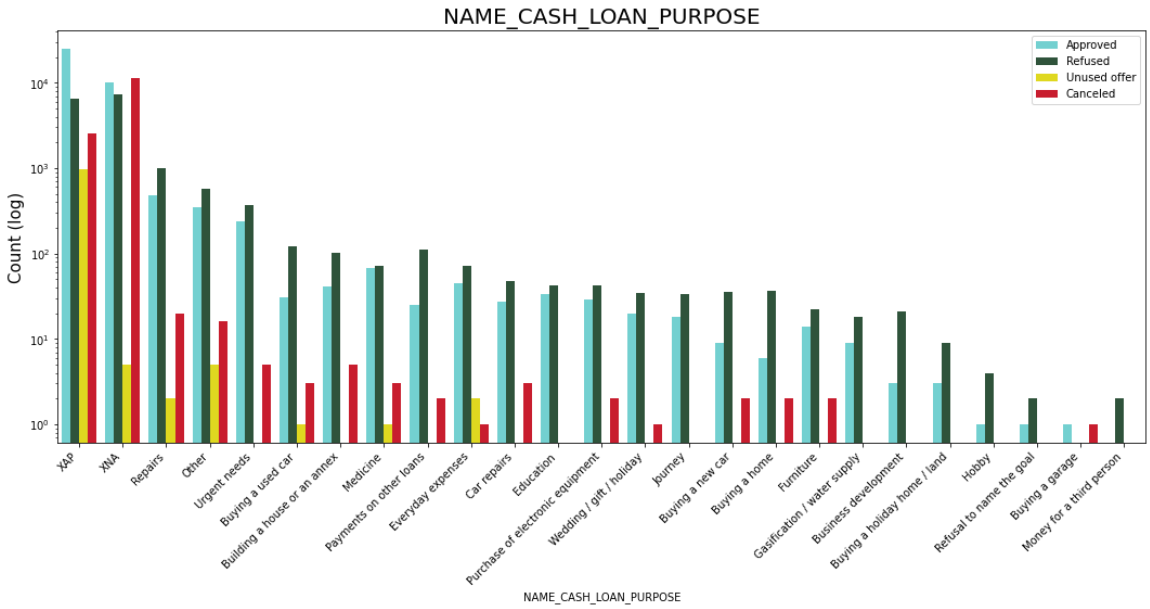
Observation

- Most approved loans were from Country-wide Channel
- Most refused loans were from Credit and Cash Offices Channel

Analysing now with merging these two dataset.

Plotting Contract Status vs purpose of the loan

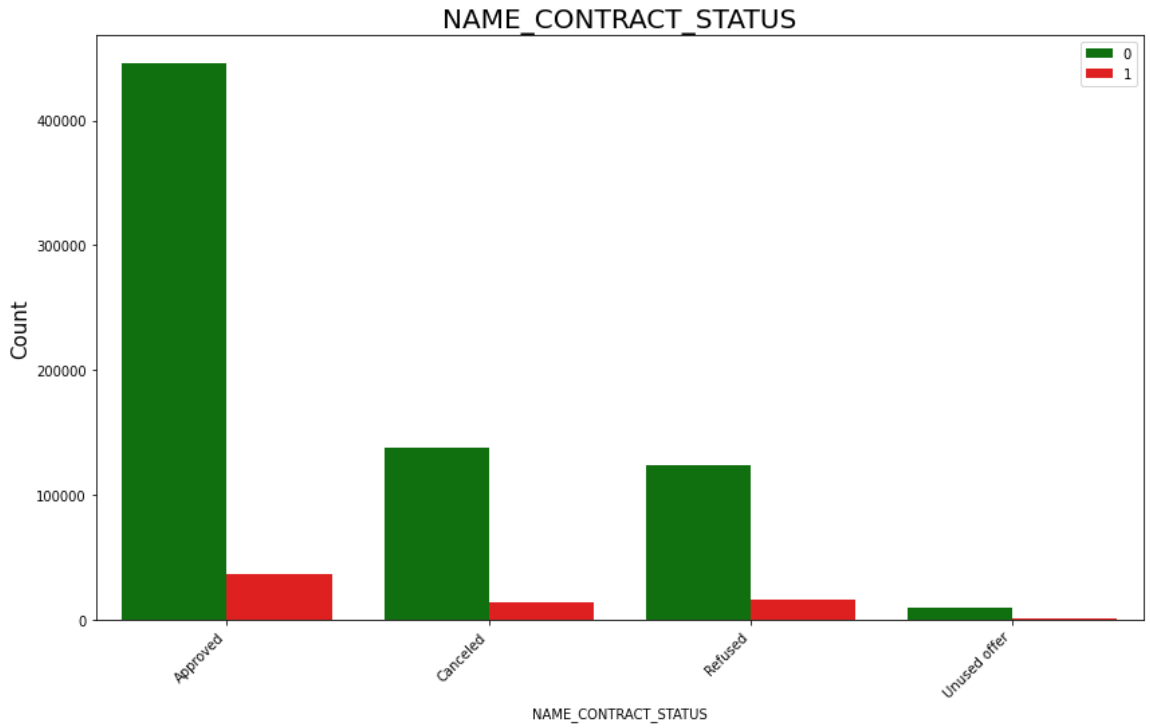




Observation

- **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.**

Checking Contract Status based on loan repayment status whether there is any business loss or financial loss



Observation

- **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.**

Mostly the variables are more or less familiar, as we seen in our Application data, that has been contributing more to the DEFAULTERS prediction

Final Conclusion

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 Defaulter:

- CODE_GENDER: Men are at relatively higher default rate
- NAME_FAMILY_STATUS : People who have civil marriage or who are single default a lot.
- NAME_EDUCATION_TYPE: People with Lower Secondary & Secondary education
- NAME_INCOME_TYPE: Clients who are either at Maternity leave OR Unemployed default a lot.
- OCCUPATION_TYPE: Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as their default rate is huge.
- 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.
- DAYS_BIRTH: Avoid young people who are in age group of 20-40 as they have higher probability of defaulting
- DAYS_EMPLOYED: People who have less than 5 years of employment have high default rate.
- 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.

- AMT_GOODS_PRICE: When the credit amount goes beyond 3lakhs, there is an increase in defaulters.

B. Decisive Factor whether an applicant will be Repayer:

- NAME_EDUCATION_TYPE: Academic degree has less defaults.
- NAME_INCOME_TYPE: Student and Businessmen have no defaults..
- ORGANIZATION_TYPE: Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 1%
- DAYS_BIRTH: People above age of 50 have low probability of defaulting
- DAYS_EMPLOYED: Clients with 40+ year experience having less than 1% default rate
- AMT_INCOME_TOTAL:Applicant with Income more than 650,000 are less likely to default
- CNT_CHILDREN: People with zero to two children tend to repay the loans.

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

- 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.
- 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.
- 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.
- 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.
- 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. Suggestions:

- 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.