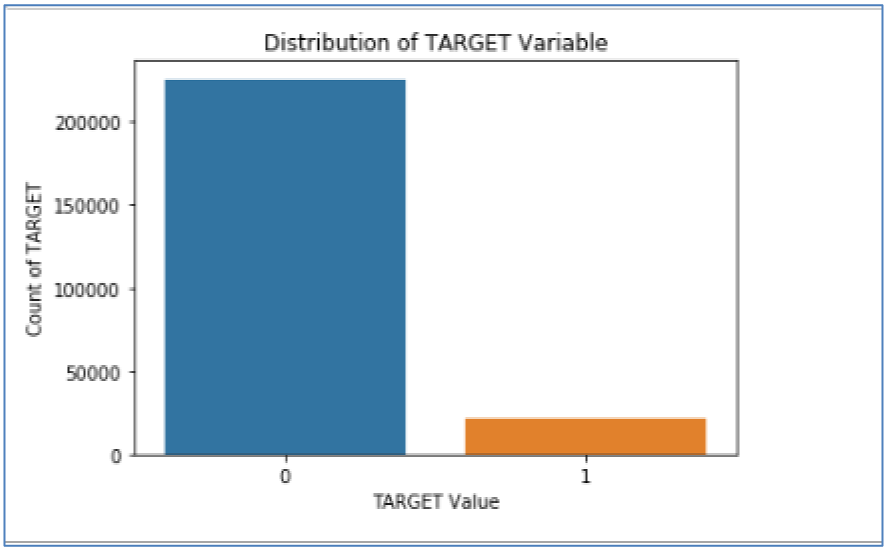
EDA CREDIT CASE STUDY

# Imbalance Percentage for Defaulter (Target 1) and Non-defaulter (Target 0)

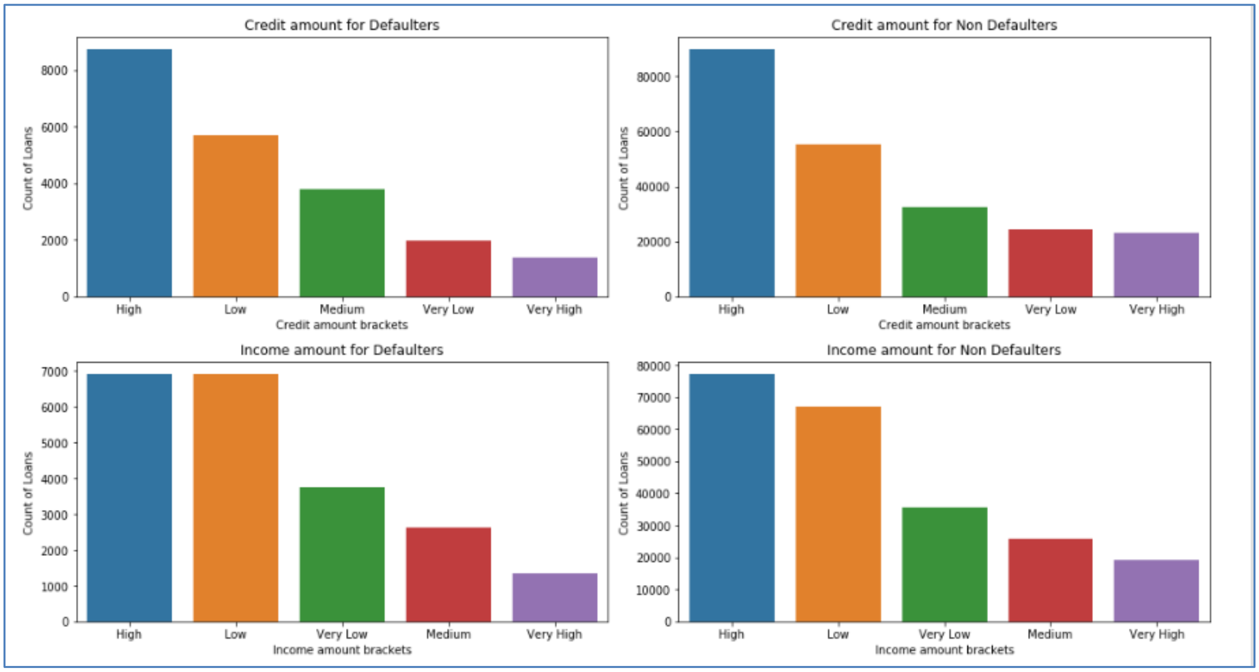
Defaulter percentage is approximate 9% and non-defaulter % is 91% hence data is highly imbalance

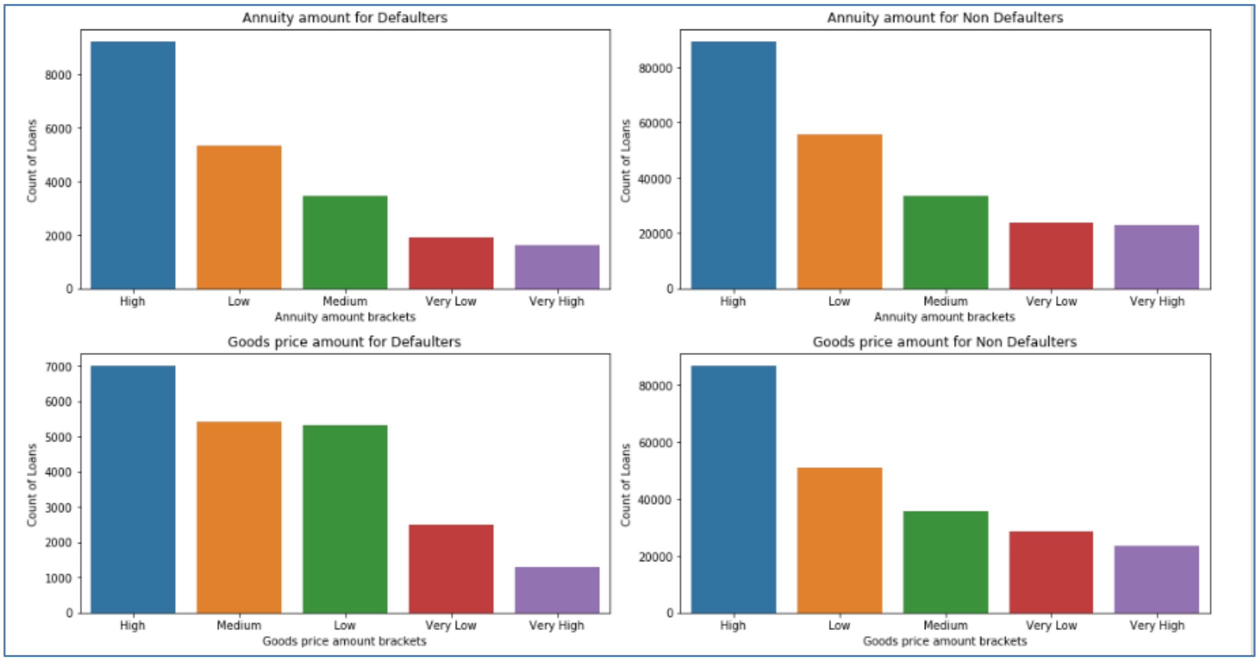


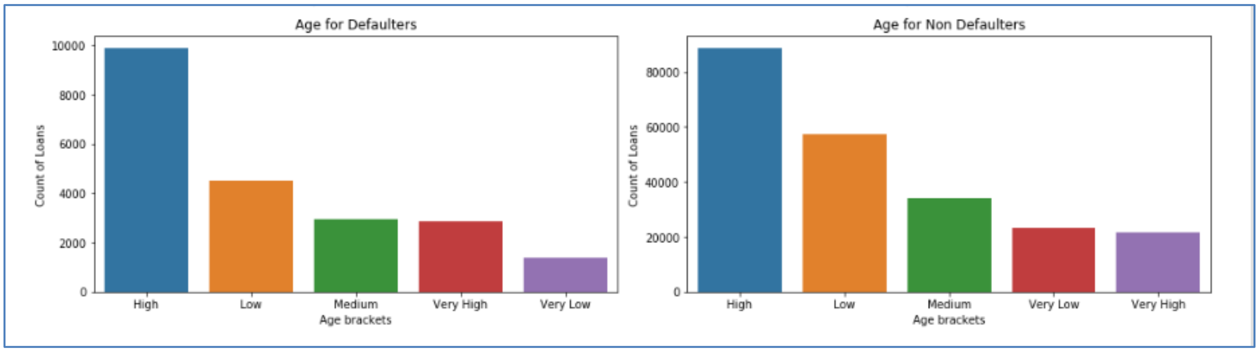
# Univariate Analysis for Continuous Variable showing the Defaulter and Non-defaulter side by side

Lets create categories as 'Very Low' , 'Low' , 'Medium' , 'High' , 'Very High' by binning method for below variables

AMT_CREDIT, AMT_INCOME_TOTAL, AMT_ANNUITY, AMT_GOODS_PRICE, DAYS_BIRTH





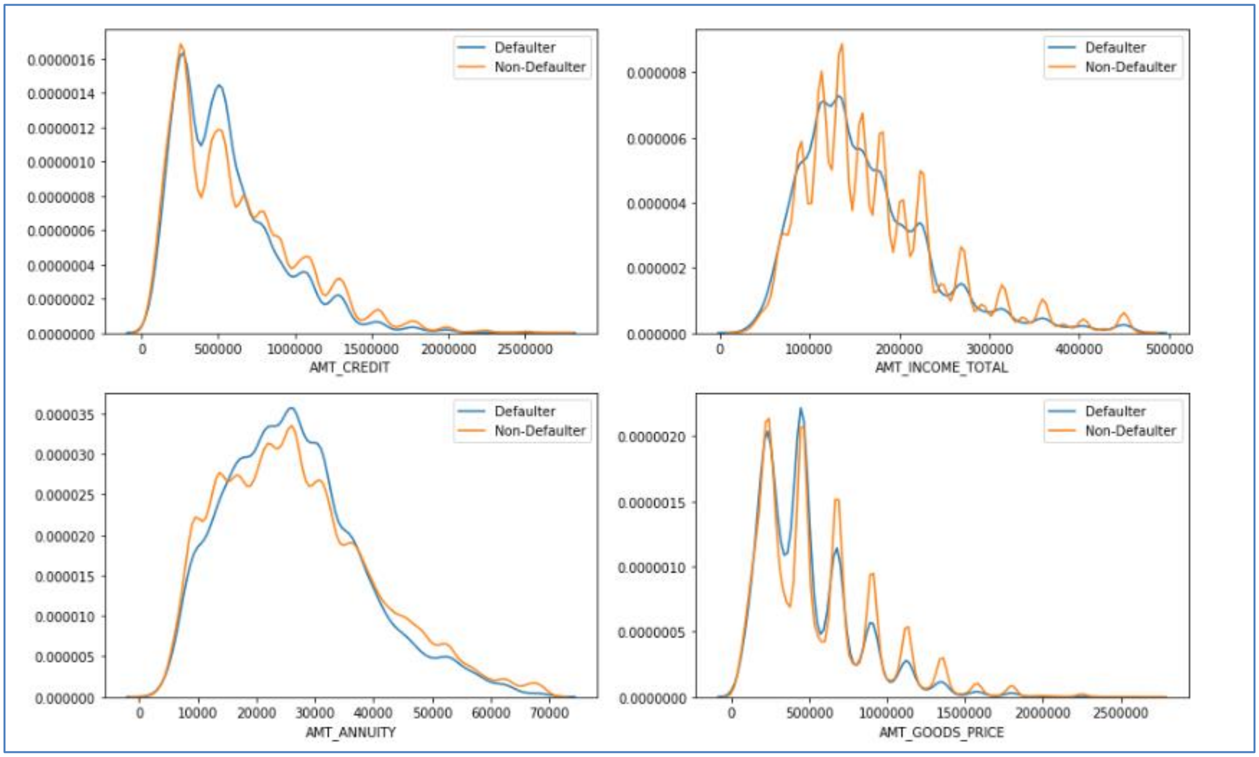


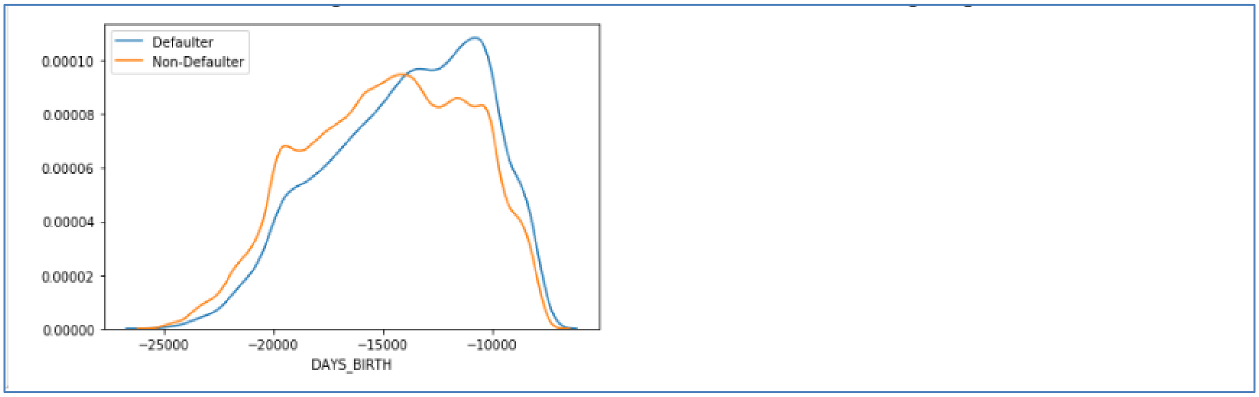
Observations and Inferential

1) From the countplots it has been observed that there is no specific differentiated pattern for above variables between Defaulter and Non-defaulter.

2) There is High demand of High credit amount from defaulters as well as Non-defaulter

Distribution plot for the 5 continuous variables



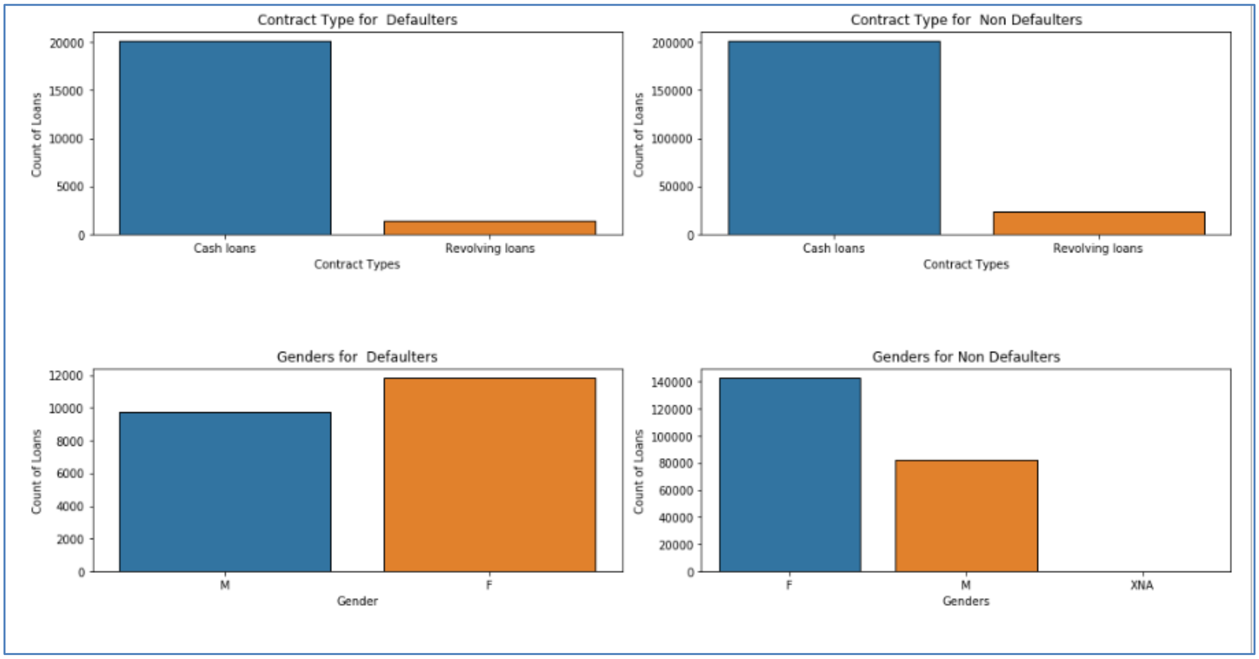


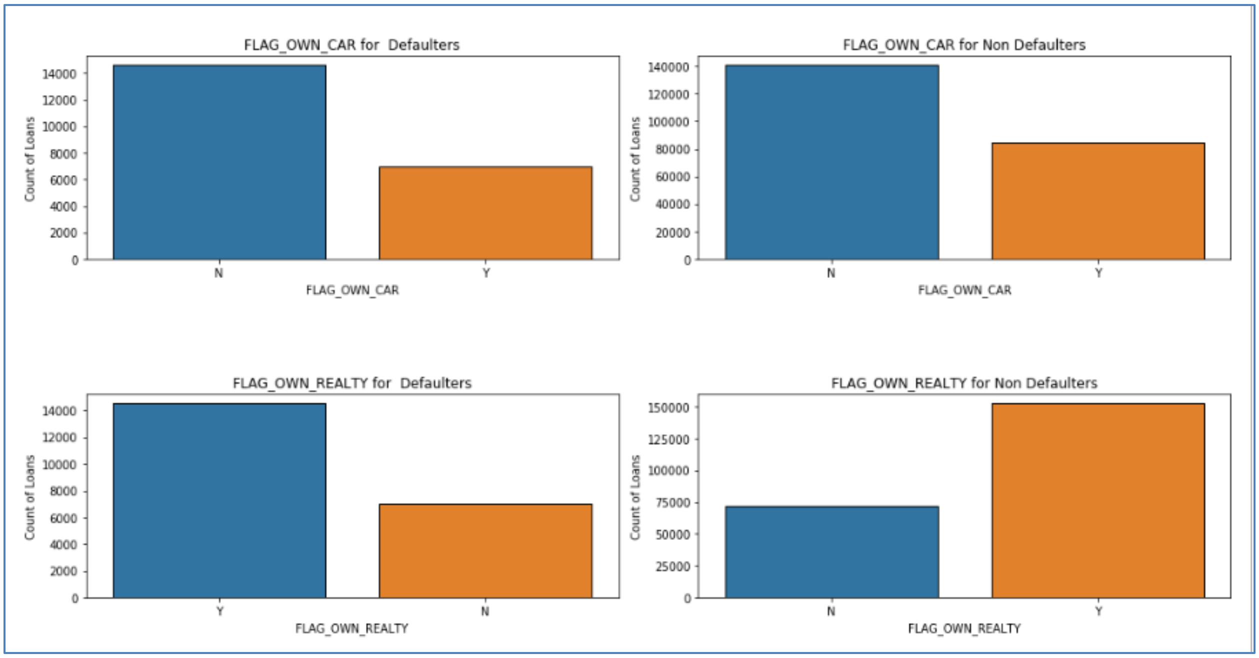
<u>Observations and Inferential</u>

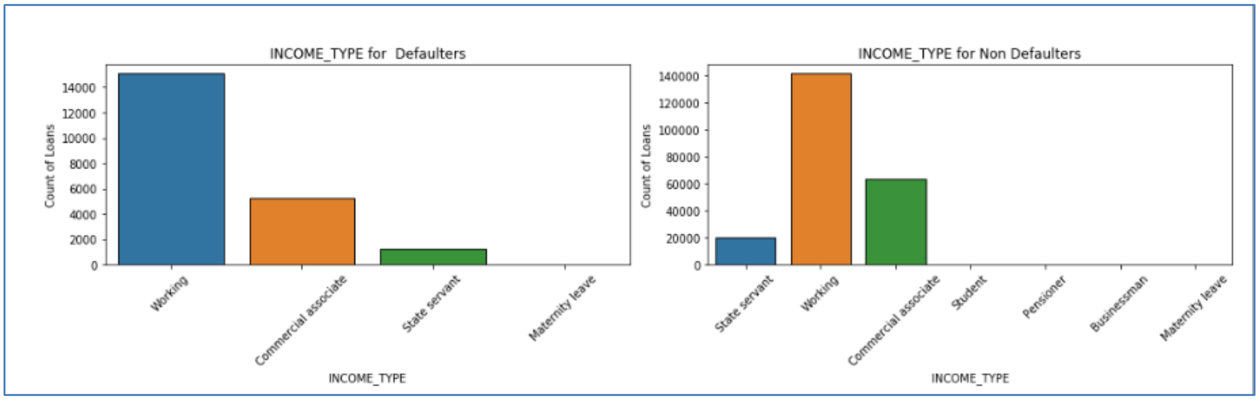
1) For Amount_Credit, Amount_Annuity and Amount_Goods price, distribution is similar
2) For Amount_Income and Days_Birth, there is a difference in pattern
3) Amount_Income Total = There is a large spike at regular interval of Income Amount for Non defaulter, so income amount is steady compare to non-defaulters
4) Days_Birth = Probability of being defaulters at lower Age is much higher than higher Age

Univariate analysis of Categorical variables for Defaulter and Non-defaulters to identify the pattern

Let`s consider Categorical variables i.e. NAME_CONTRACT_TYPE, CODE_GENDER, FLAG_OWN_CAR, FLAG_OWN_REALTY, NAME_TYPE_SUITE, NAME_INCOME_TYPE







Observations and Inferential

1) From the graphs of Contract type we can see that the Revolving loans are small amount compared to Cash loans

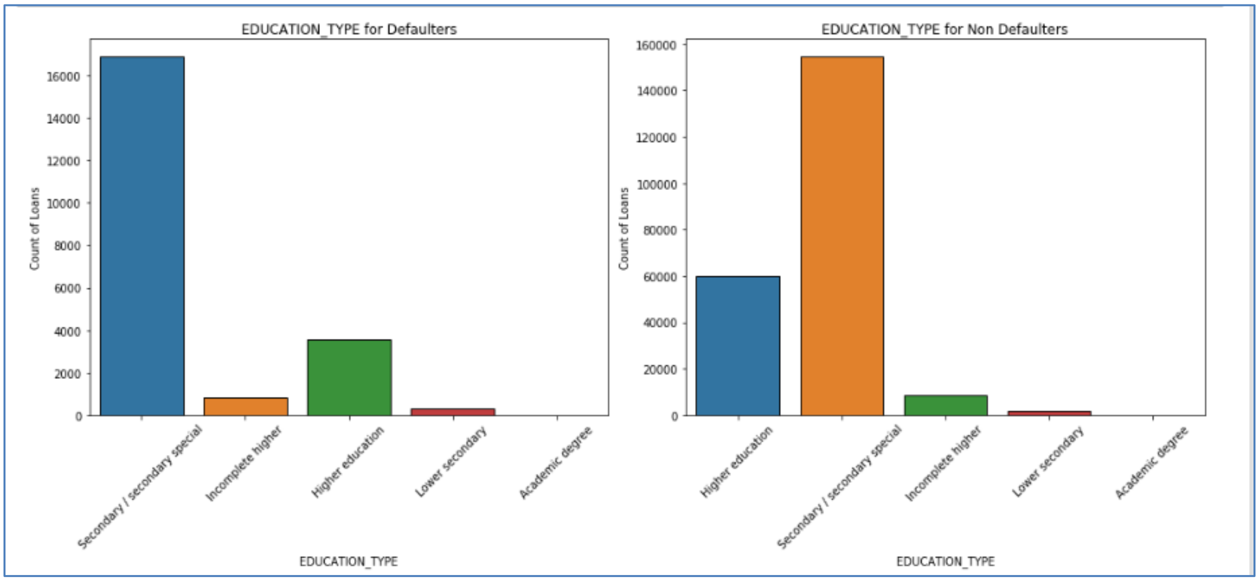
2) The women defaulters are more than men

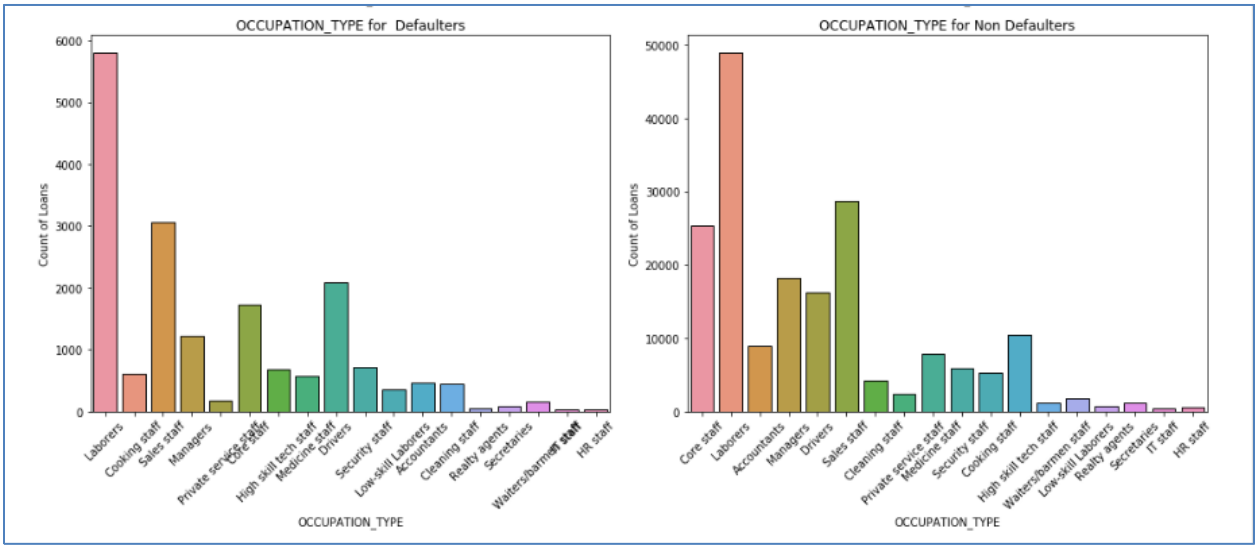
3) Number of defaulters having own car and realty are less compare to defaulters who does not own anything hence these are not directly impacting factors

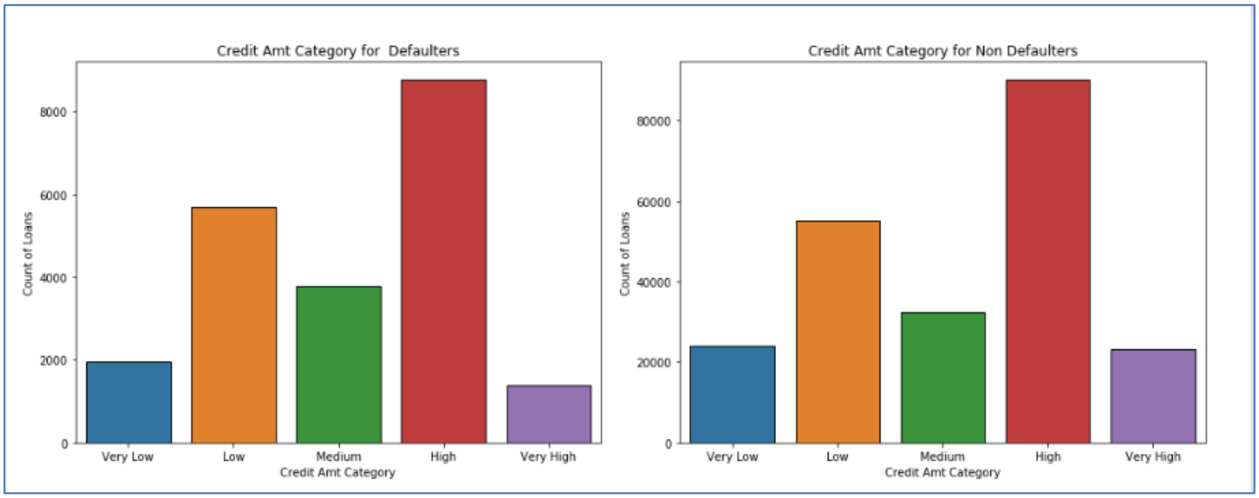
4) By observing income type; working personnel are more defaulter compare to Commercial associates and state governments, more than 60% consumers are working personnel hence we cannot say there are more chances that working personnel are defaulters

# Univariate analysis of Categorical variables for Defaulter and Non-defaulters to identify the pattern

Let`s consider Categorical variables i.e. NAME_EDUCATION_TYPE, NAME_FAMILY_STATUS, NAME_HOUSING_TYPE, OCCUPATION_TYPE, ORGANIZATION_TYPE, Amount Credit Category



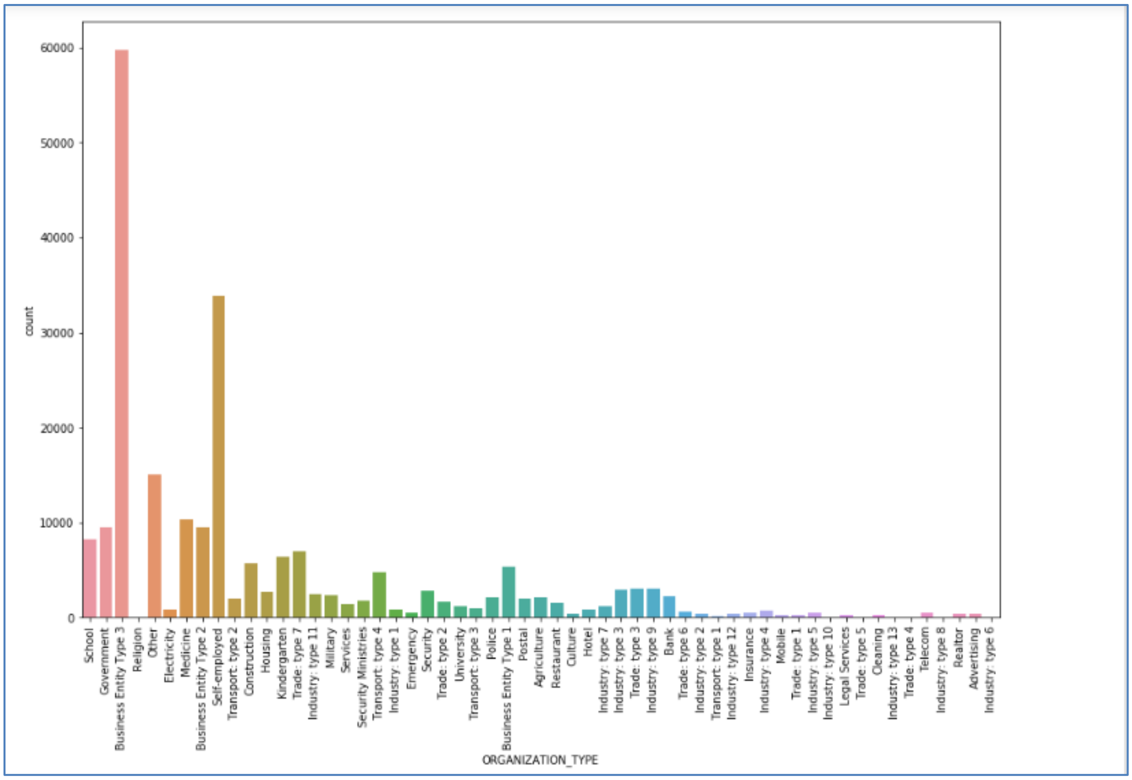


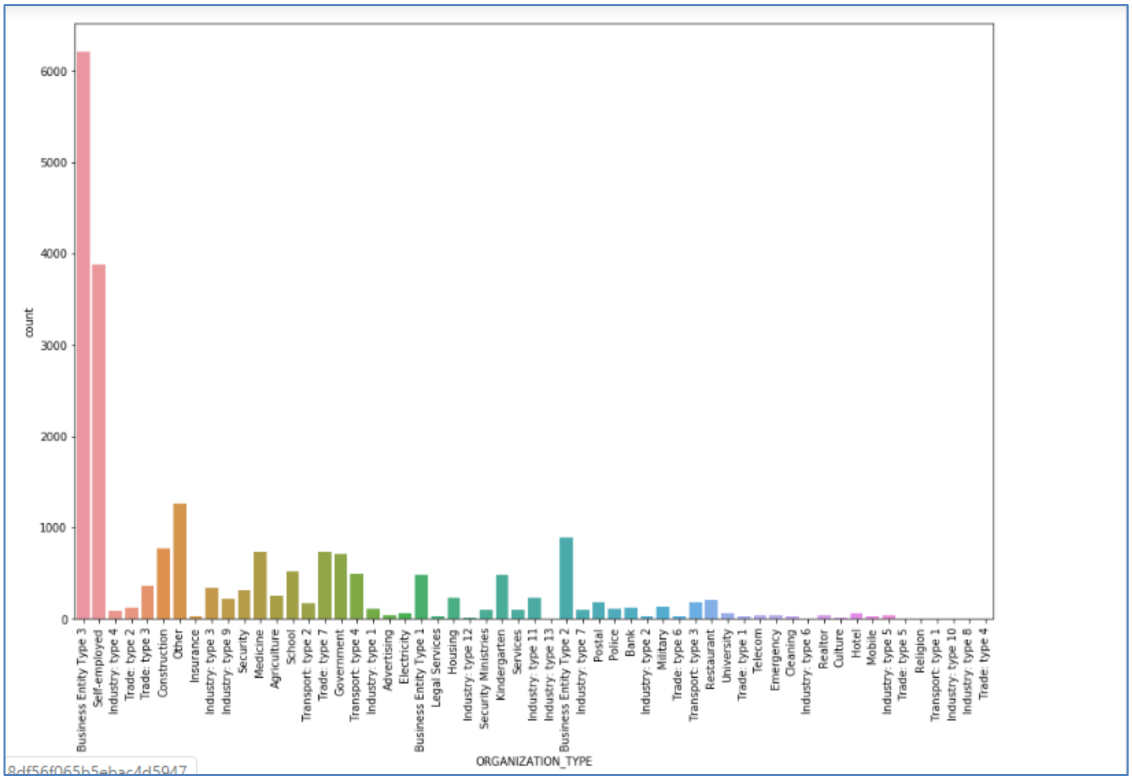


<u>Observations and Inferential</u>

1) By observing education type; consumer educated till secondary/secondary special are more defaulter compare to other qualification levels but 70% consumers are studied till secondary/secondary special hence we cannot say there are more chances that these category would be having more defaulters

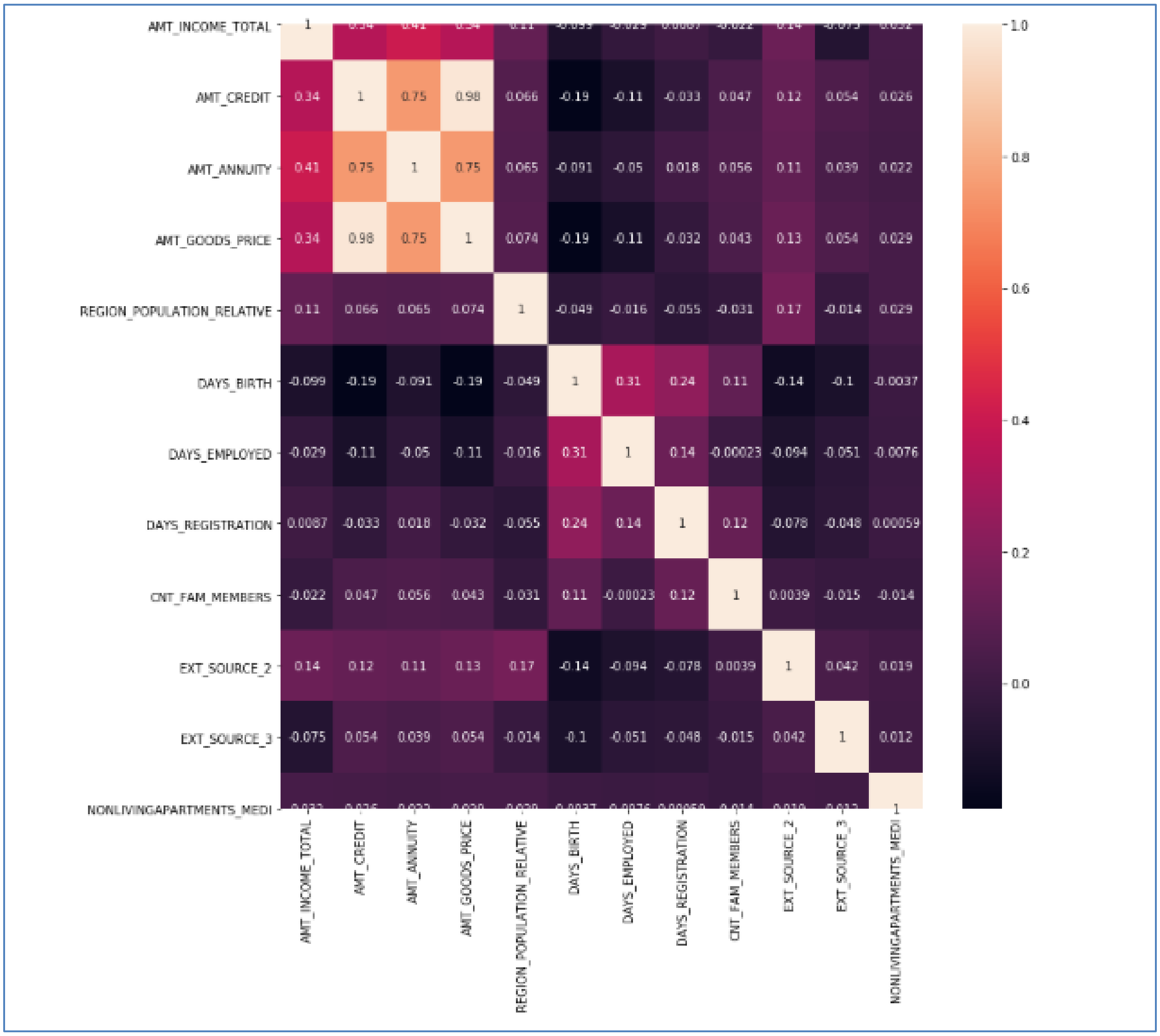
2) Similar patterns has been observed for Occupation type





From above graphs of Org type of defaulter and Non-defaulter, it can be seen that most of the application are received from "Business Entity type 3"

Correlation of Continuous Variables for Defaulters



Top 5 Correlated Variables for Defaulters

1) Amount Goods price and Amount Credits - Very strong correlation

2) Amount Annuity and Amount Credits - Strong correlation

3) Amount Goods price and Amount Annuity - Strong correlation

4) Amount Annuity and Amount total income - Moderate correlation

5) Amount Goods price and Amount Credits - Moderate correlation

Bottom 5 Correlated Variables for Defaulters

1) NONLIVINGAPARTMENTS_MEDI and Days Registration

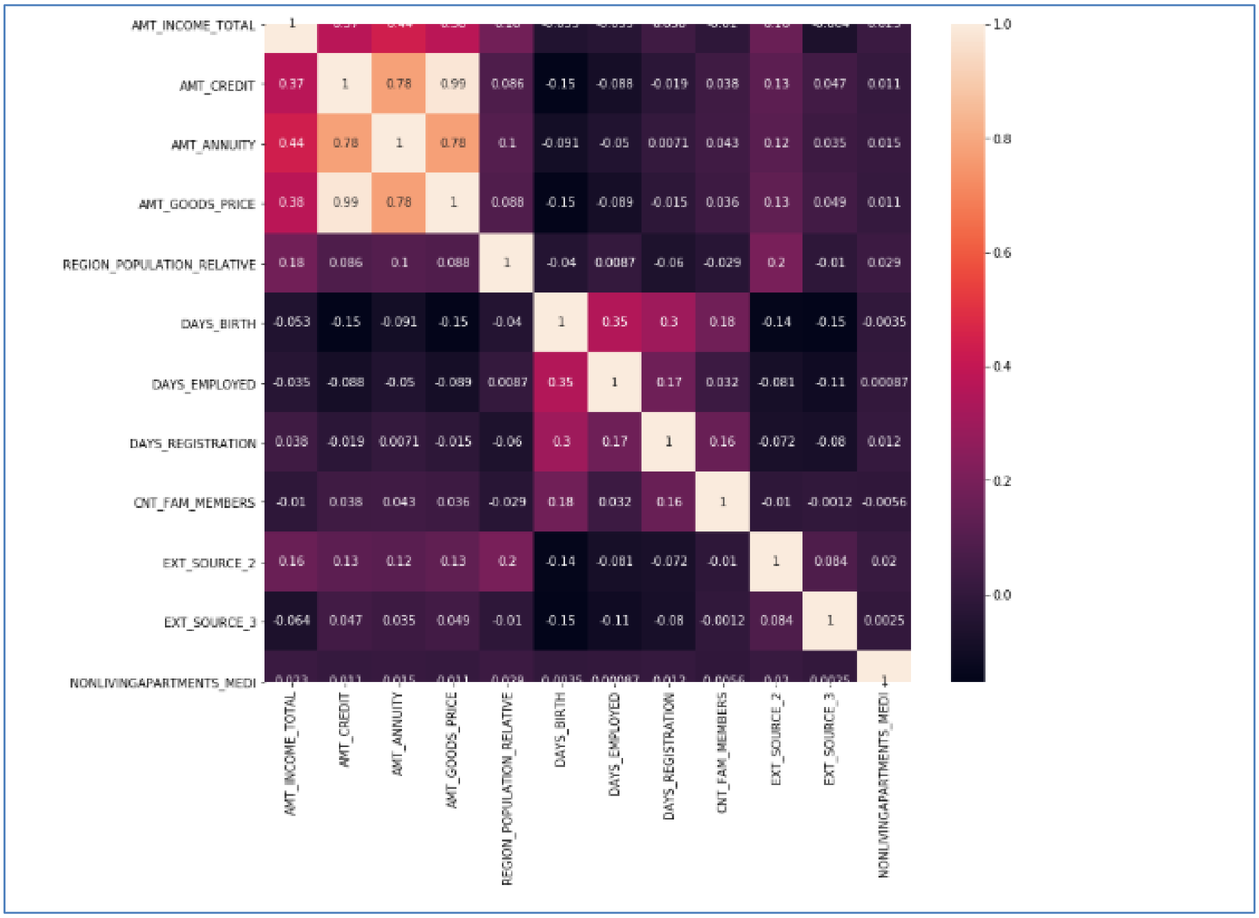
2) NONLIVINGAPARTMENTS_MEDI and Days Birth

3) Ext source 2 and Count of Family Members

4) NONLIVINGAPARTMENTS_MEDI and Days Employed

5) Ext source 3 and Region population relative

# Correlation of Continuous Variables for Non-Defaulters



Top 5 Correlated Variables for Non-Defaulters

1) Amount Goods price and Amount Credits

2) Amount Annuity and Amount Credits

3) Amount Goods price and Amount Annuity

4) Amount Annuity and Amount total income

5) Amount Goods price and Amount Credits

Bottom 5 Correlated Variables for Non - Defaulters

1) NONLIVINGAPARTMENTS_MEDI and Ext source 3

2) NONLIVINGAPARTMENTS_MEDI and Days Employed

3) Ext source 2 and Count of Family Members

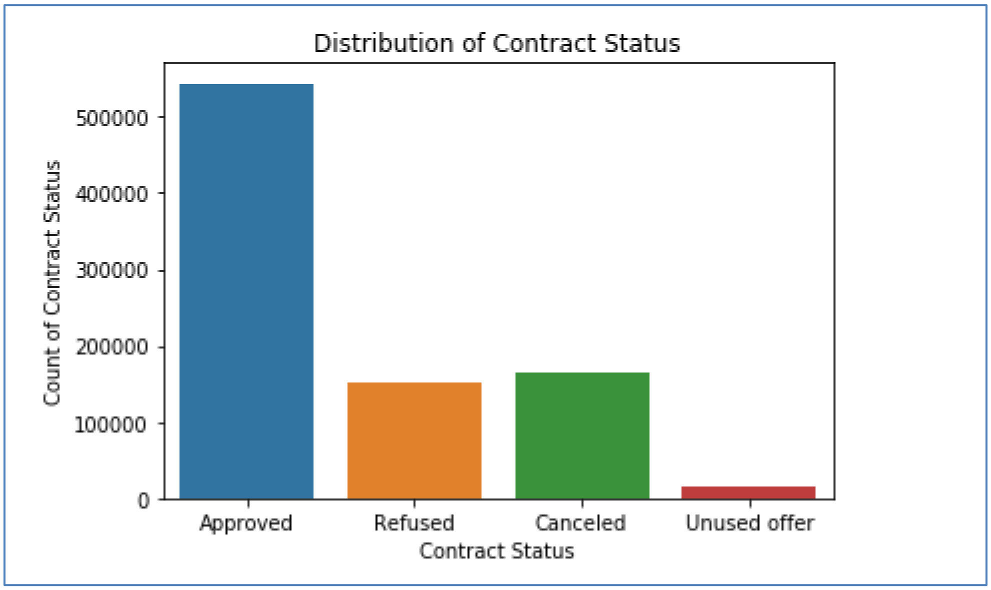
4) NONLIVINGAPARTMENTS_MEDI and Days Birth

5) Count of Family Members and Total Income amount

For Both Defaulter and Non defaulter, top 5 correlated variables are same but correlation factors non-defaulters are higher than the defaulters

Stage 2: Upload previous data and perform analysis

# Distribution of previous loan application status



# Univariate analysis of segmented continuous variables against approved, refused and cancelled loans

Percent of loans approved: 62.21

Percent of loans refused: 17.38

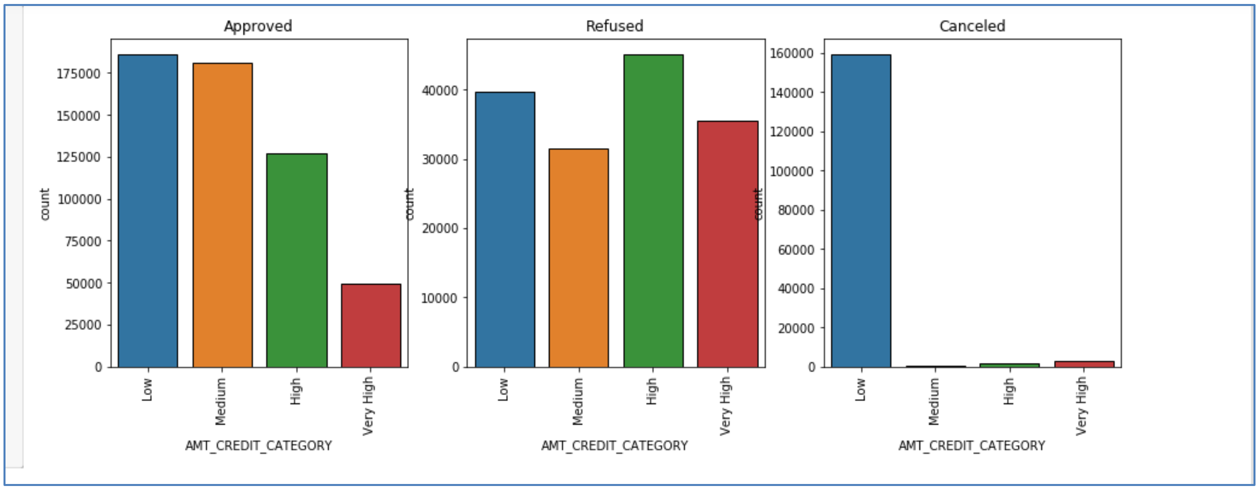
Percent of loans cancelled_percent: 18.81

Perform univariate analysis on the below Categorical variables

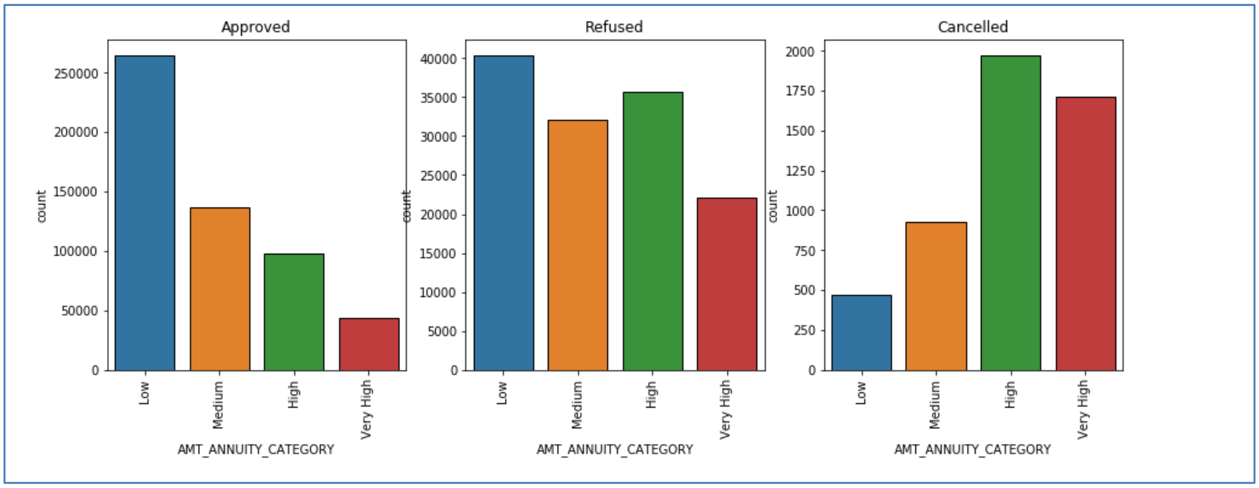
1. AMT_CREDIT

2. AMT_ANNUITY

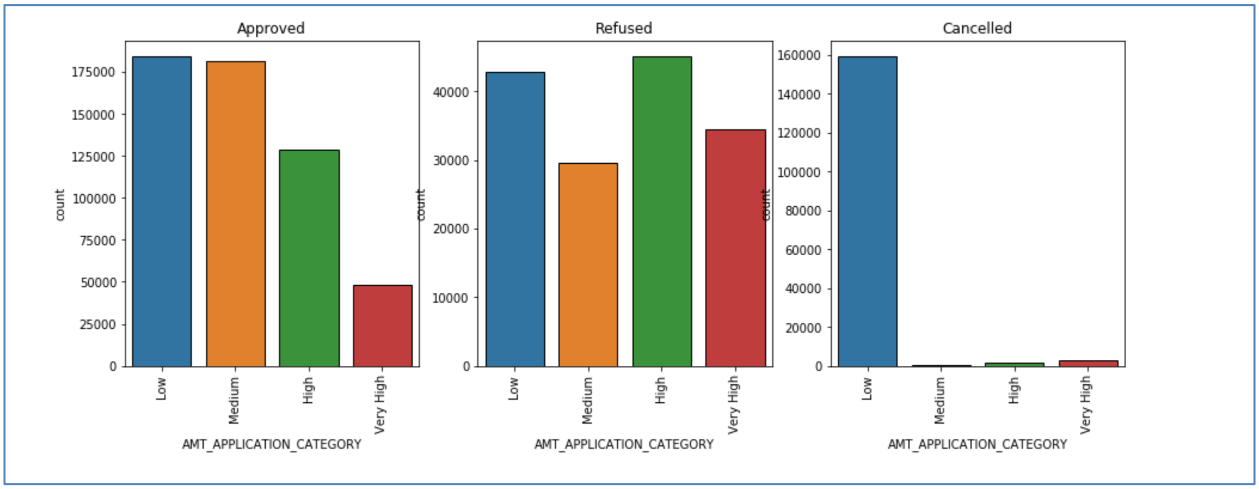
3. AMT_APPLICATION



From the above graph we can infer that Majority of Approved loans are for Low and Medium amount_credit, one reason could be high risk in very_high amount credits



From the above, we can infer that, For approved loans, majority of them have low Annuity amounts



From the above, we can infer that,

1. Loans with low and medium application amounts are approved.

2. Refused loans has high percentage of High and Very high application amounts

# Univariate analysis of Categorical variables against approved, refused and cancelled loans

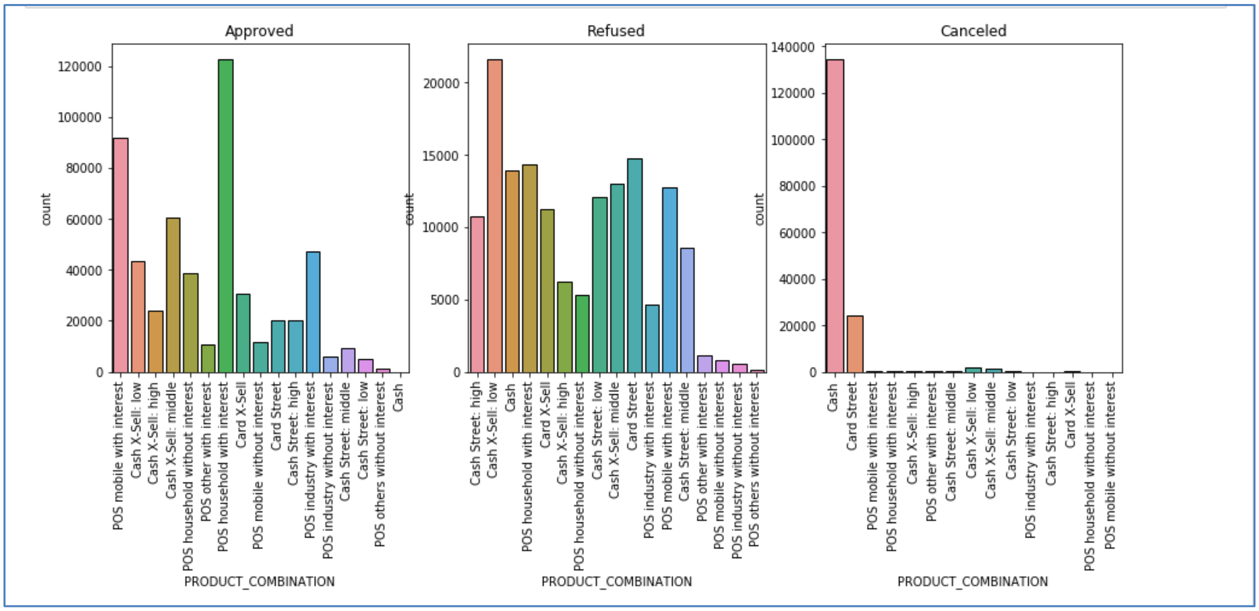
1. PRODUCT_COMBINATION

2. NAME_PORTFOLIO

3. NAME_SELLER_INDUSTRY

4. NAME_CONTRACT_TYPE

5. NAME_PAYMENT_TYPE

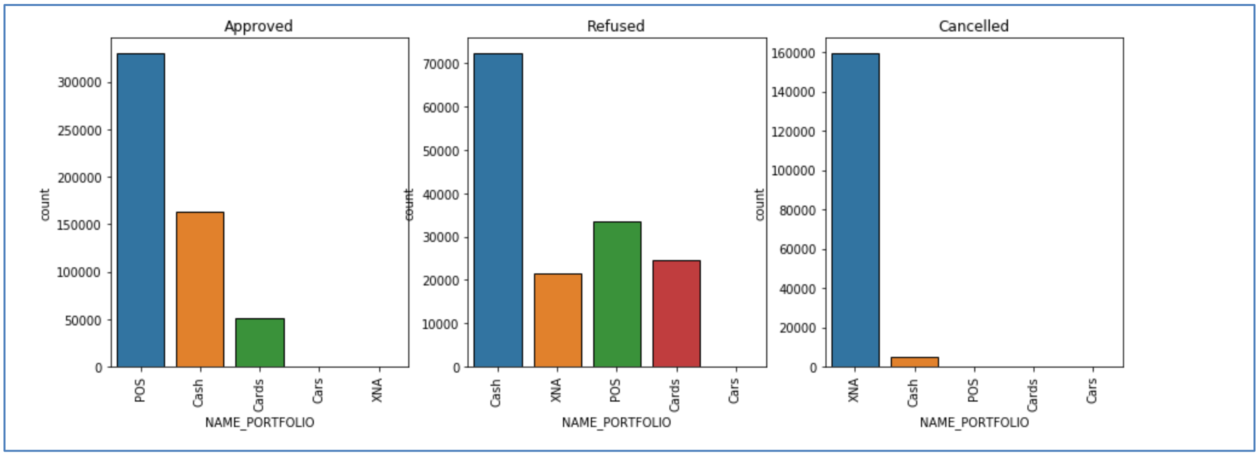


From the above chart, we can infer the below,

1. We observe most number of loans were approved for POS household with interest.

2. Most number of refused loans were of Cash X-Sell: Low Product combination

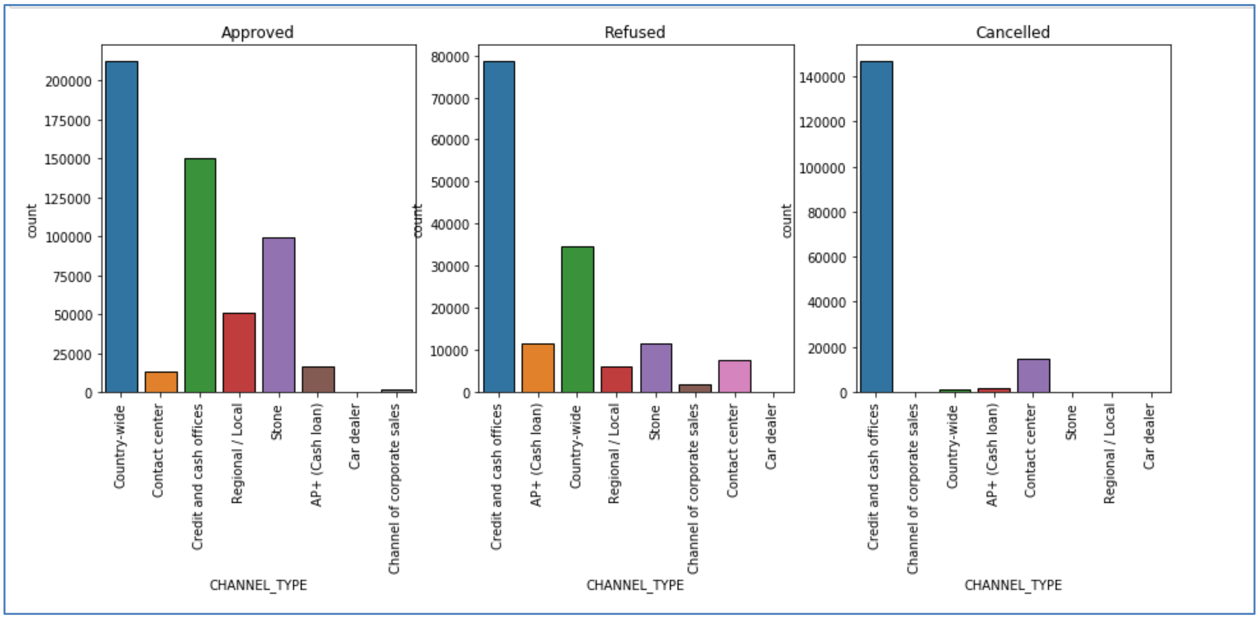
3. Most Canceled loans were Cash loans



From the above chart, we can infer the below,

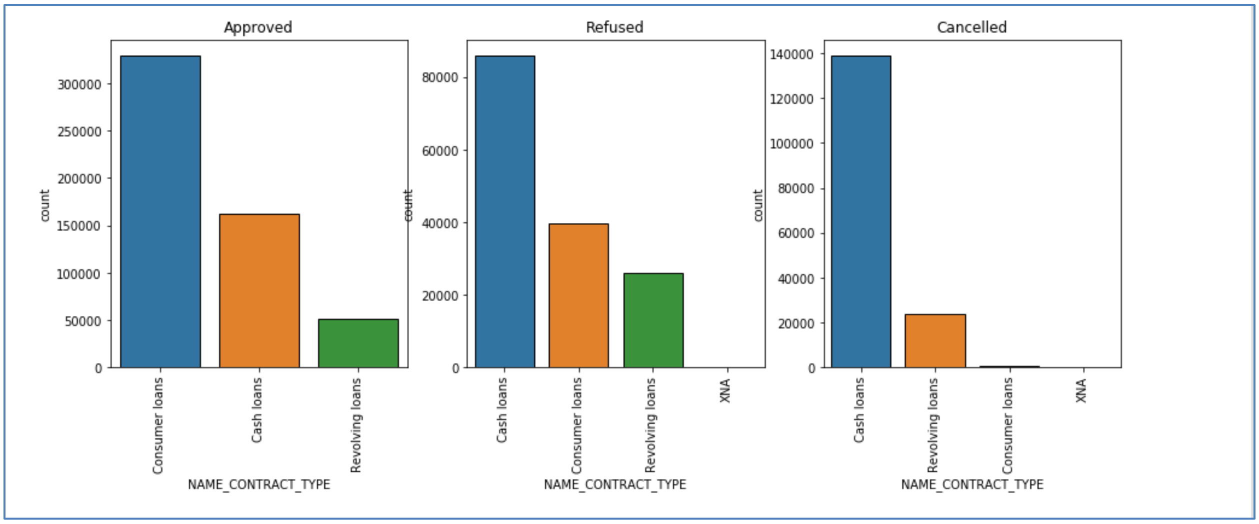
1. Most approved loans are for POS

2. Most rejected loans are for Cash



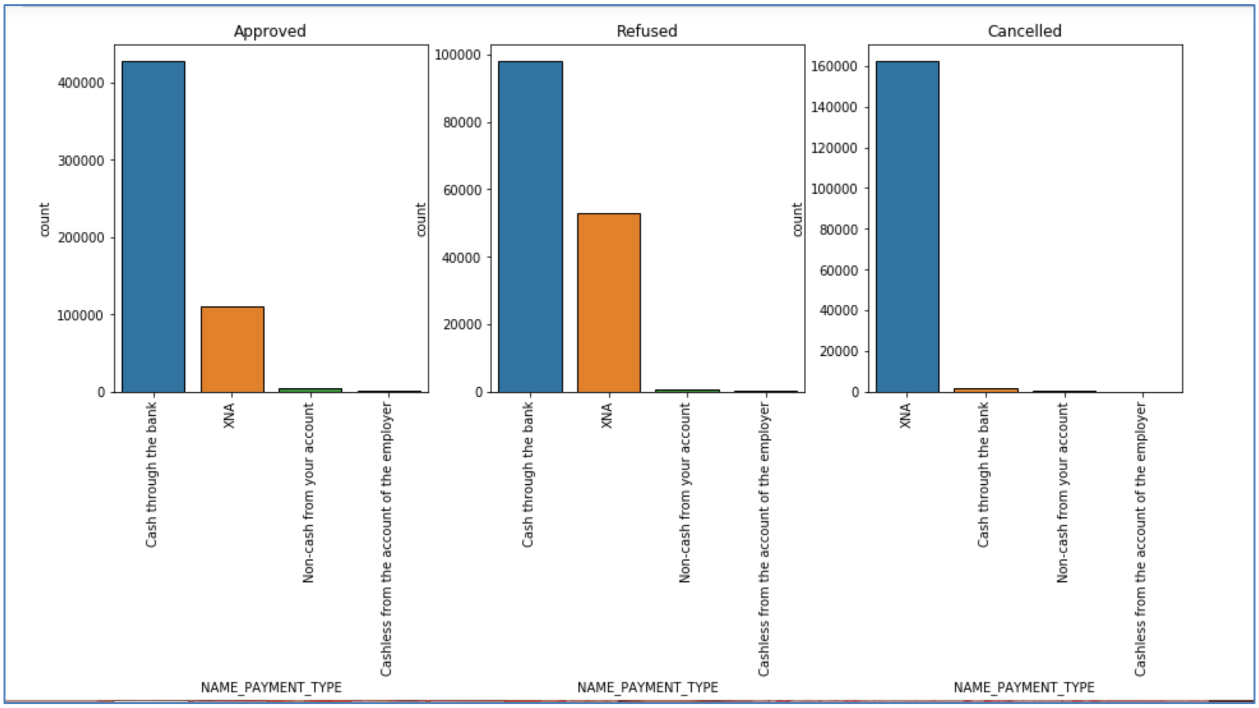
1. Most approved loans were from Country-wide Channel

2. Most refused loans were from Credit and Cash Offices Channel



1. Most of the approved contract types are Consumer loans

2. Most of the rejected contract types are Cash loans



1. We can see that Payment Types are independent of loans approved or rejected.

# Bivariate analysis and finding correlation of float columns

1) Approved loans category

Top correlated variables for approved data frame with their correlation factors

```
AMT_CREDIT        AMT_GOODS_PRICE      0.993364
AMT_APPLICATION   AMT_CREDIT           0.961762
DAYS_LAST_DUE     DAYS_TERMINATION     0.928684
AMT_ANNUITY       AMT_GOODS_PRICE      0.830909
AMT_CREDIT        AMT_ANNUITY          0.825471
AMT_ANNUITY       AMT_APPLICATION      0.814345
CNT_PAYMENT       AMT_APPLICATION      0.646532
                  AMT_GOODS_PRICE      0.636545
AMT_CREDIT        CNT_PAYMENT          0.626794
```

2) Refused loans category

Top correlated variables for approved data frame with their correlation factors

```
AMT_CREDIT        AMT_GOODS_PRICE      0.993364
AMT_APPLICATION   AMT_CREDIT           0.961762
DAYS_LAST_DUE     DAYS_TERMINATION     0.928684
AMT_ANNUITY       AMT_GOODS_PRICE      0.830909
AMT_CREDIT        AMT_ANNUITY          0.825471
AMT_ANNUITY       AMT_APPLICATION      0.814345
CNT_PAYMENT       AMT_APPLICATION      0.646532
                  AMT_GOODS_PRICE      0.636545
AMT_CREDIT
```

Conclusion:

From the above Analysis we can conclude that below variables should be considered for granting loan.

DAYS_BIRTH – This is the Age of the consumer and it is found that chances to be a defaulter are more for the consumers with lesser Age compare to Higher.

AMT_CREDIT – This is also very important factor, If Credit amount is higher then Risk from the defaulter is Highers

AMT_ANNUITY - As Credit amount is increased, Annuity amount also increased. Refusal of loan is higher in case of Higher Amount of Annuity

DAYS_EMPLOYED – This factor shows if consumer has regular and stable income, more the days employed less the Risk for company.

Refused Loans – There are high risk of being defaulter if the previous applications of consumer are refused.