EDA: Loan Credit

Handling Missing Values

- We have dropped the columns having null values greater than 45%
- In the next step errors in the data are treated like we have Changed the 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH' and 'DAYS_LAST_PHONE_CHANGE' which had negative or mixed values and imputed them with absolute values for our analysis.
- Some of the columns have 'XNA' as values in them, we assumed it is because the person didn’t want to disclose the gender so we fixed XNA by ‘Not disclosed’.

Outliers Handling

Step 1: Finding Outliers

- We found outliers in the numerical columns.
- Columns having unique value more than 100 are only considered.
- Ext_Source_2 and Ext_Source_2 are removed as their values are already normalized in the provided data. SK_ID_CURR has only individual IDs so we removed this too as this data is not numerical.

Outliers Handling (contd.)

Step 2: Handling outliers

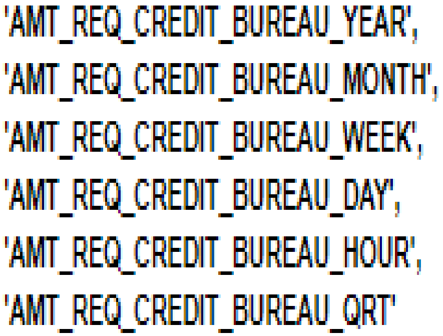
- We have removed outliers as per IQR theory.
- **Outlier = less than Q1-1.5xIQR and greater than Q3+1.5xIQR**. Since all of the boxplots have outliers on the right side, we removed values greater than Q3+1.5xIQR.
- After handling outliers few columns still had some percent of null values for which imputation was needed.

Imputations: Categorical

- **OCCUPATION_TYPE** and **NAME_TYPE_SUITE** are the two columns we have imputed.
- In case of **OCCUPATION_TYPE** we have imputed the missing values by creating a separate category ‘Unknown’. Imputing mode will not be correct here as 30% of the missing values can’t be assumed to be laborers.
- For **NAME_TYPE_SUITE** we have used mode for our imputations.

Imputations: Numerical

Below are the columns for which we have imputed missing values.

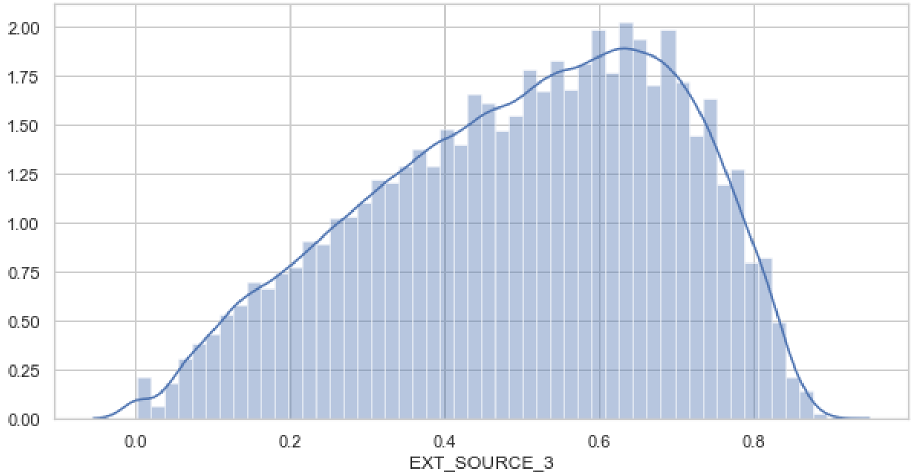


```
'AMT_REQ_CREDIT_BUREAU_YEAR',
'AMT_REQ_CREDIT_BUREAU_MONTH',
'AMT_REQ_CREDIT_BUREAU_WEEK',
'AMT_REQ_CREDIT_BUREAU_DAY',
'AMT_REQ_CREDIT_BUREAU_HOUR',
'AMT_REQ_CREDIT_BUREAU_QRT'
```

For the imputation of these columns we have used the mode, which is the value '0’. Since they all are number of credit enquiries (discrete variables), replacing missing values with the most frequent value rather than mean would make more sense.

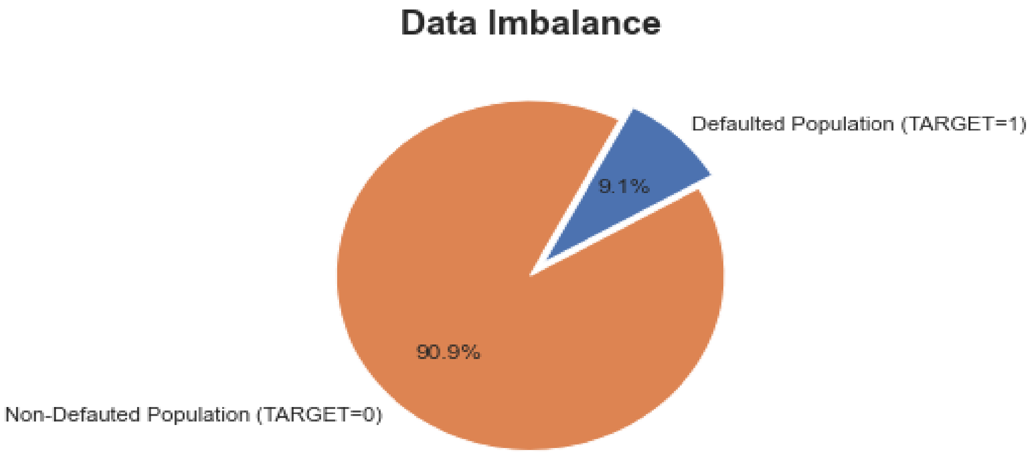
Imputations: Numerical (contd.)

EXT_SOURCE_3



It is clear that this data is 'right-skewed'. Hence, the mean of the data is shifted right due to the skewness. It will be better to impute this column by Median. Also, the difference between median and mean is low compared to mode.

Data Imbalance



The ratio of the count of defaulters to the count of all other cases in the given dataset is 0.0998.

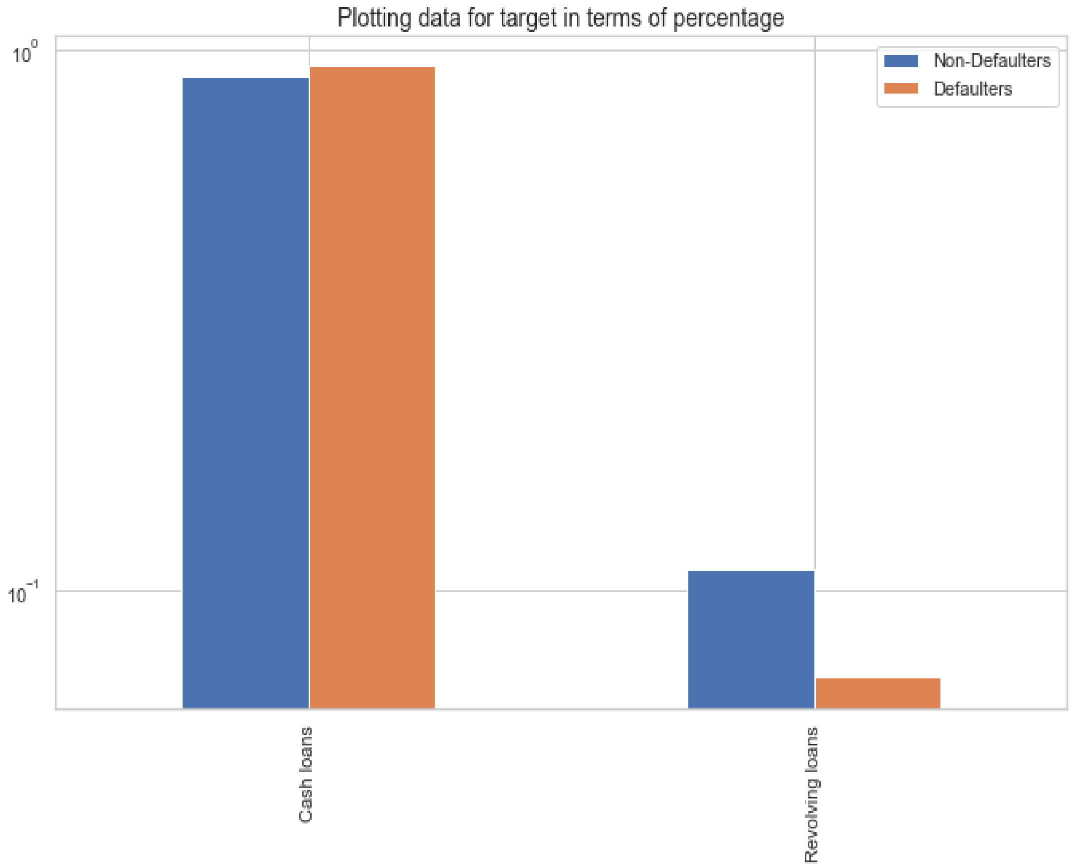
Thus, we can say that about 9.1% of people faced difficulties in repayment of the loan.

Univariate Analysis of Categorical Columns

Columns analyzed:

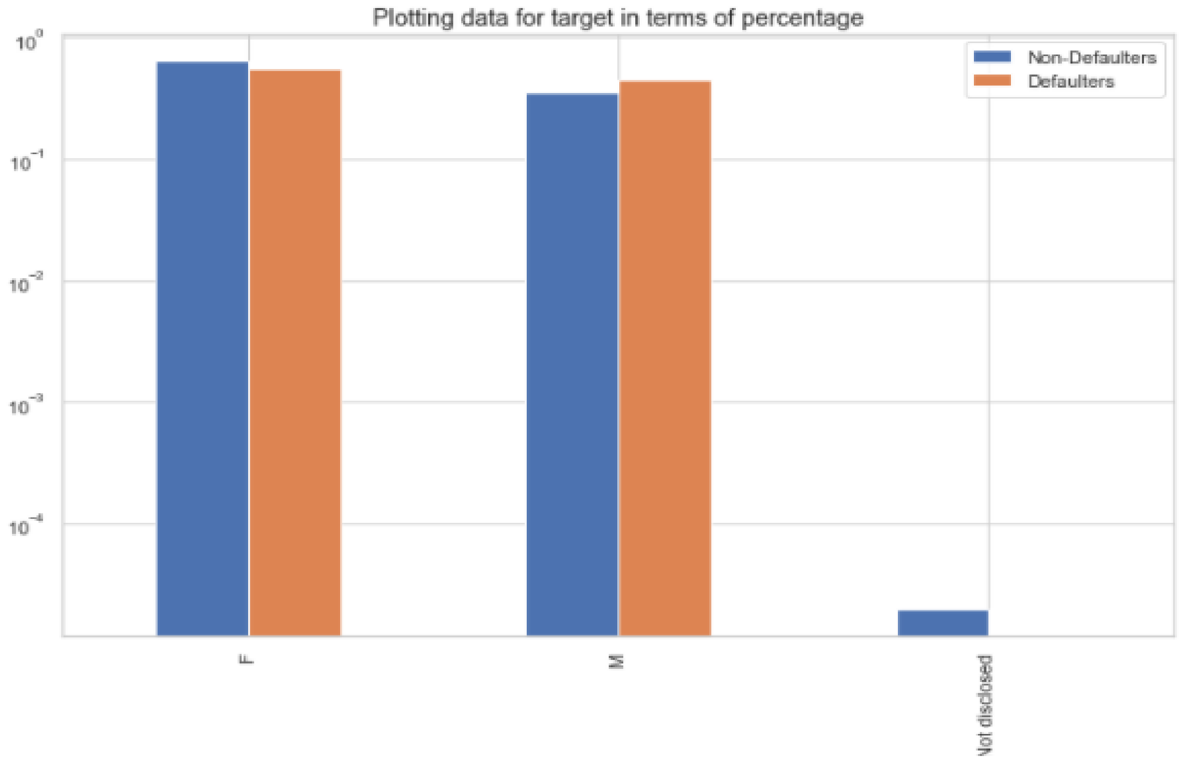
- Contract Type
- Gender
- Income Type
- Education Type
- Occupation Type
- Age Group
- Family Status Type
- Housing Type

Contract Type



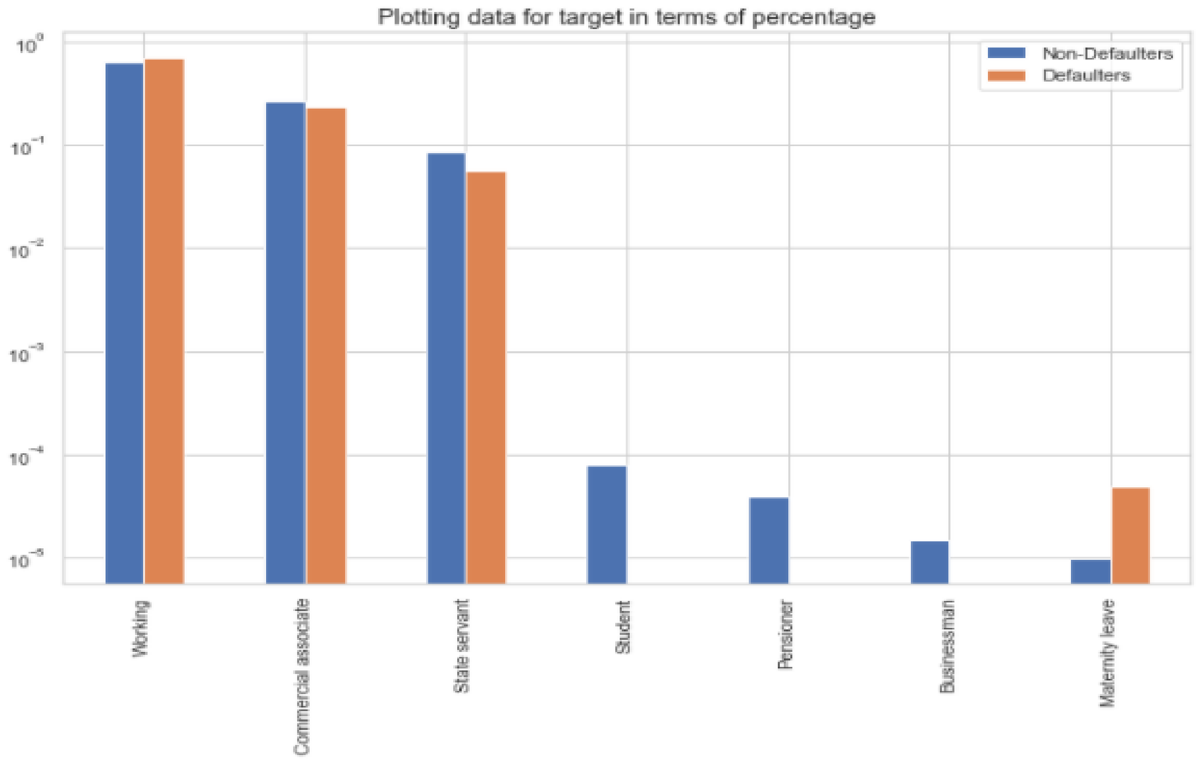
- People are likely to repay the loan if it is of revolving type as compared to cash loans.
- The number of cash loans is higher as compared to the count of revolving loans.

Gender



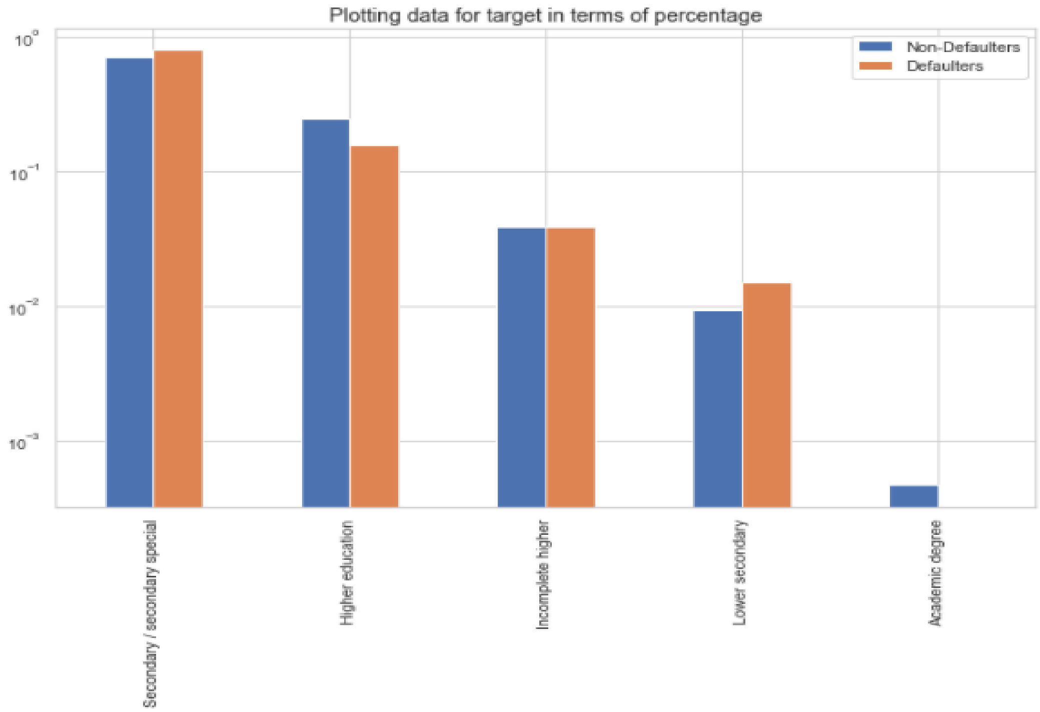
- We can say females applied for loan more than the males.
- Females are more sincere towards repaying the loans as defaulter percentage is lower as compared to males.

Income Type



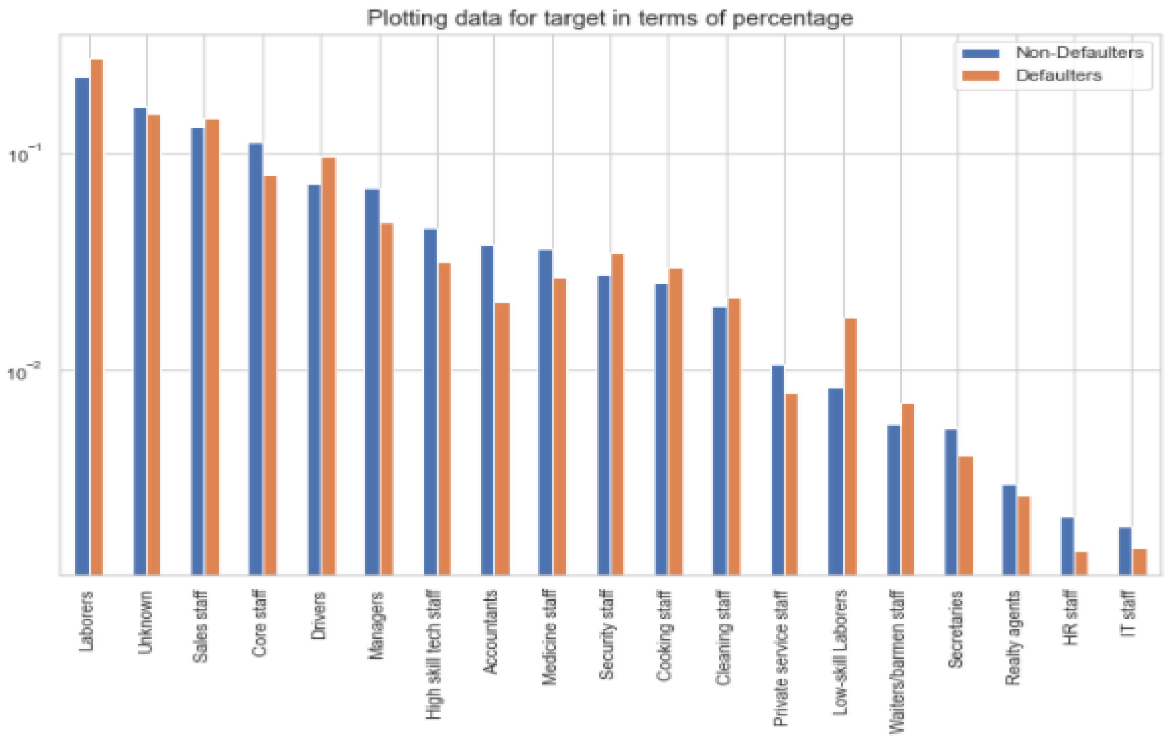
- Working class has higher count and percentage of defaulters.
- State servants have low chances of being in defaulter list as compared to working and commercial associates.

Education Type



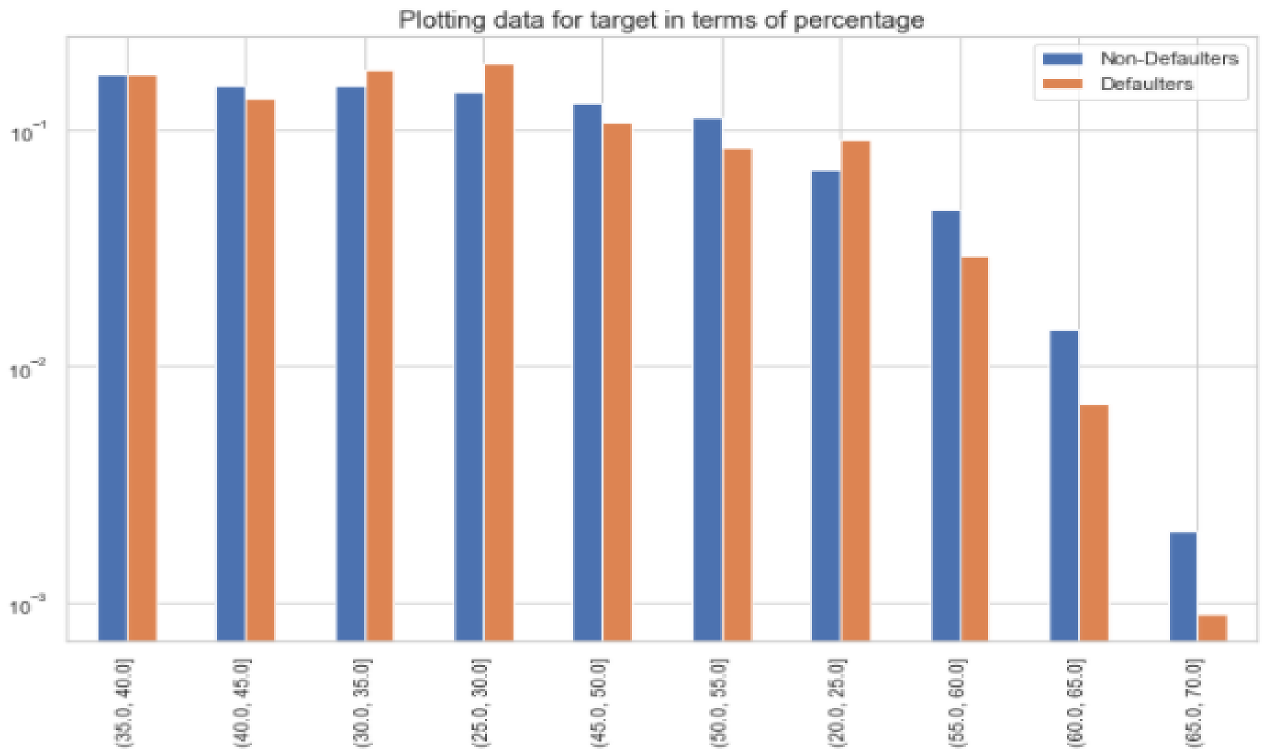
- Secondary education type people are likely to be defaulters we can say that the percentage is higher as more number of people are applying for loans in this category.

Occupation Type



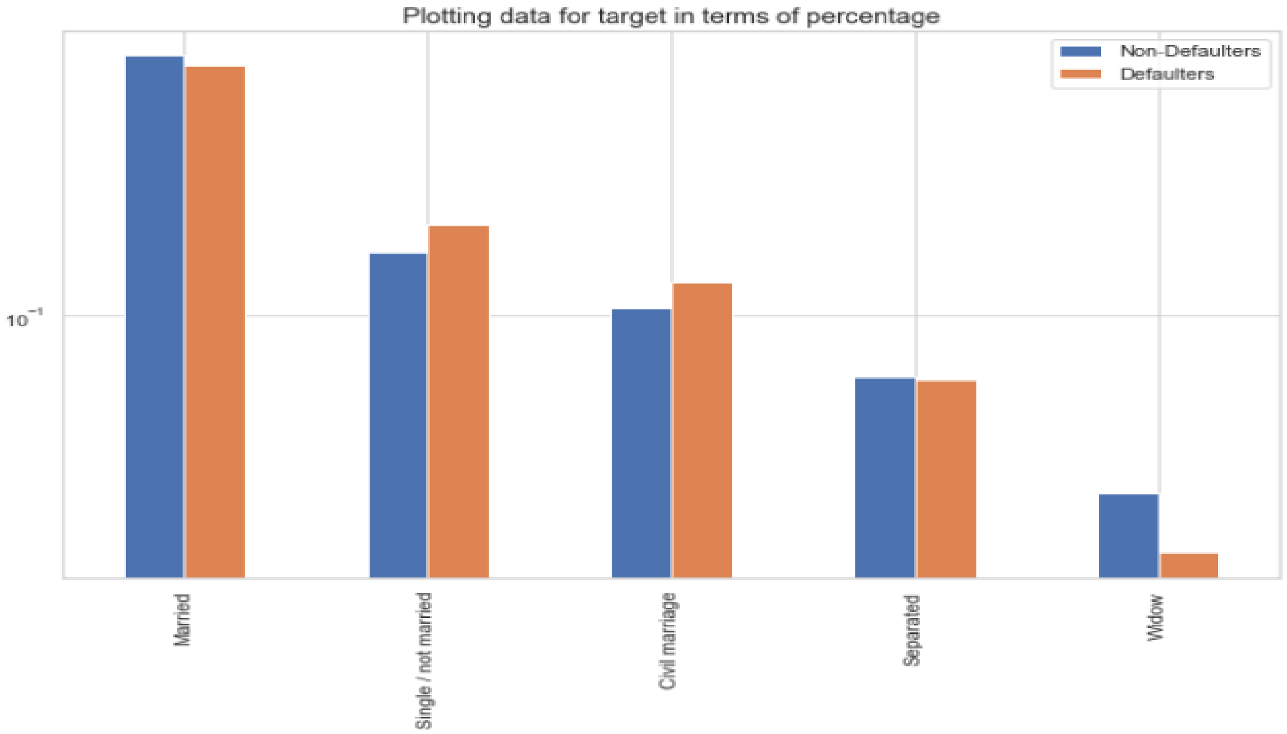
- Huge chunk of people that applied for loans were laborers.
- Laborers are more likely to be defaulters followed by sales and drivers.
- The percentage of Core staff and Accountants who were defaulters was low.

Age Group



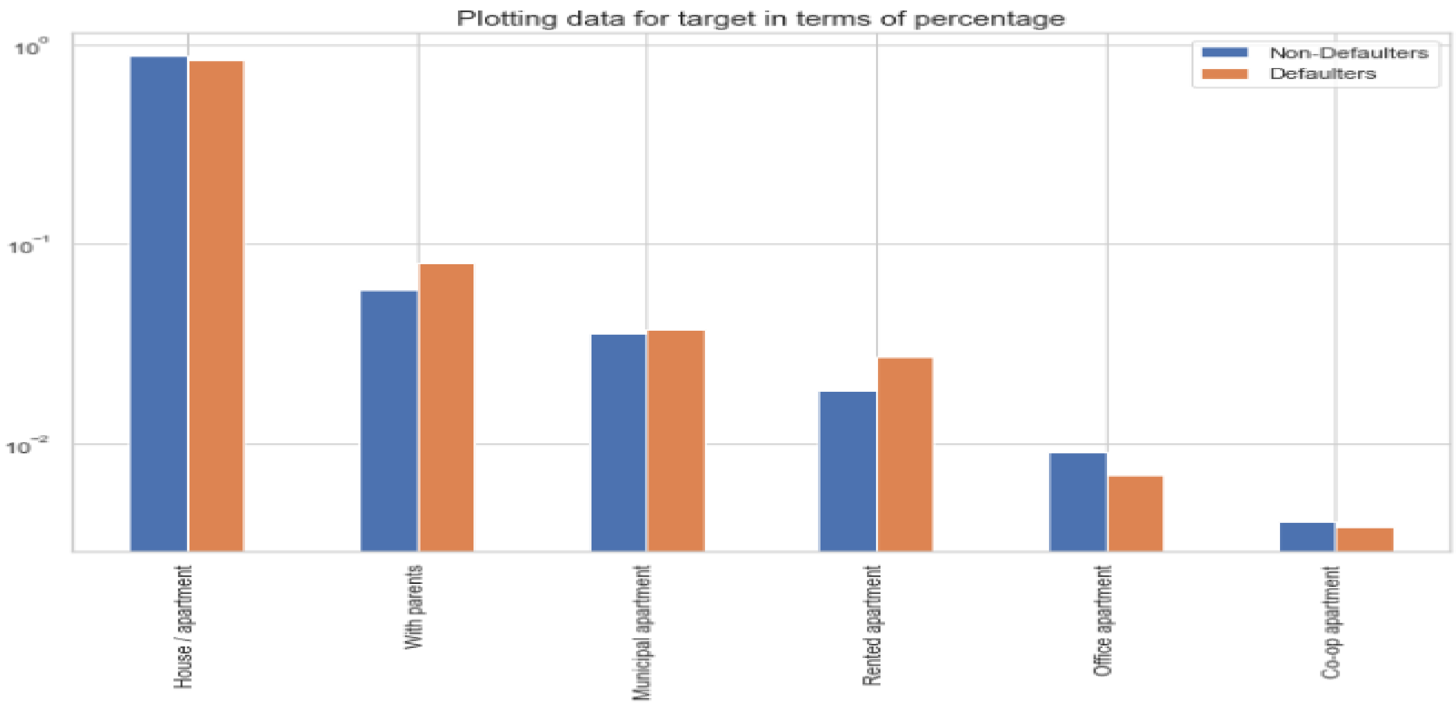
- Most of the people who required loan were between 20 to 50.
- People between the age of 20 to 35 are likely to be defaulters.
- The count of applicants older than 55 years of age was low.

Family Status Type



- Most of the people that applied for loans were married.
- High percentage of Single and Civil married people were defaulters.

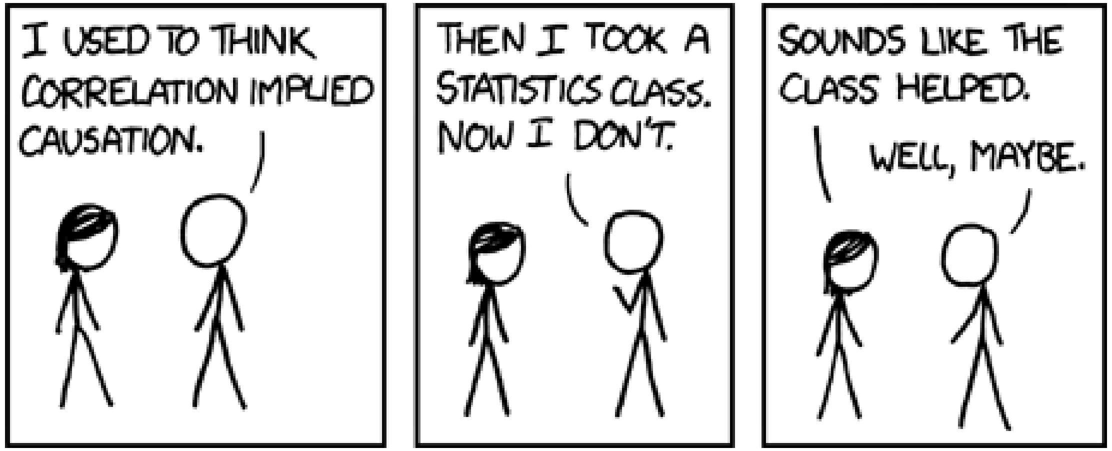
Housing Type



Housing Type (contd.)

- Population living in Rented apartments and those living with parents have higher default rate as they have higher proportion in the defaulted population as compared to non defaulted population.
- Living in rental apartment means a cash out flow towards rent and thus less cash left for repayment of loan.
- Living with parents may suggest that the income is not too high and thus difficulty in repayment of loan.
- Large amount of loans was taken by own house/apartment people and they were less likely to be defaulters.

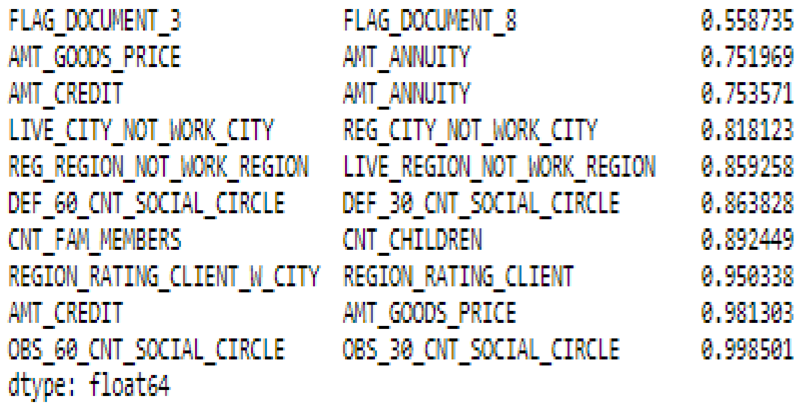
Correlations


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Top 10 variable pairs having the highest absolute correlation have been calculated along with their correlation coefficients.

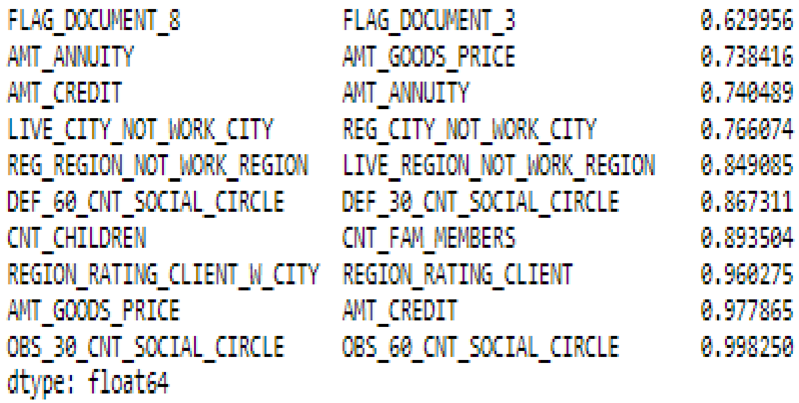
Top ten Correlations for Non-Defaulter and Defaulter

Non Defaulters



```
FLAG_DOCUMENT_3            FLAG_DOCUMENT_8              0.558735
AMT_GOODS_PRICE            AMT_ANNUITY                  0.751969
AMT_CREDIT                 AMT_ANNUITY                  0.753571
LIVE_CITY_NOT_WORK_CITY    REG_CITY_NOT_WORK_CITY       0.818123
REG_REGION_NOT_WORK_REGION LIVE_REGION_NOT_WORK_REGION  0.859258
DEF_60_CNT_SOCIAL_CIRCLE   DEF_30_CNT_SOCIAL_CIRCLE     0.863828
CNT_FAM_MEMBERS            CNT_CHILDREN                 0.892449
REGION_RATING_CLIENT_W_CITY REGION_RATING_CLIENT        0.950338
AMT_CREDIT                 AMT_GOODS_PRICE              0.981303
OBS_60_CNT_SOCIAL_CIRCLE   OBS_30_CNT_SOCIAL_CIRCLE     0.998501
dtype: float64
```

Defaulters



```
FLAG_DOCUMENT_8            FLAG_DOCUMENT_3              0.629956
AMT_ANNUITY                AMT_GOODS_PRICE              0.738416
AMT_CREDIT                 AMT_ANNUITY                  0.740489
LIVE_CITY_NOT_WORK_CITY    REG_CITY_NOT_WORK_CITY       0.766074
REG_REGION_NOT_WORK_REGION LIVE_REGION_NOT_WORK_REGION  0.849085
DEF_60_CNT_SOCIAL_CIRCLE   DEF_30_CNT_SOCIAL_CIRCLE     0.867311
CNT_CHILDREN               CNT_FAM_MEMBERS              0.893504
REGION_RATING_CLIENT_W_CITY REGION_RATING_CLIENT        0.960275
AMT_GOODS_PRICE            AMT_CREDIT                   0.977865
OBS_30_CNT_SOCIAL_CIRCLE   OBS_60_CNT_SOCIAL_CIRCLE     0.998250
dtype: float64
```

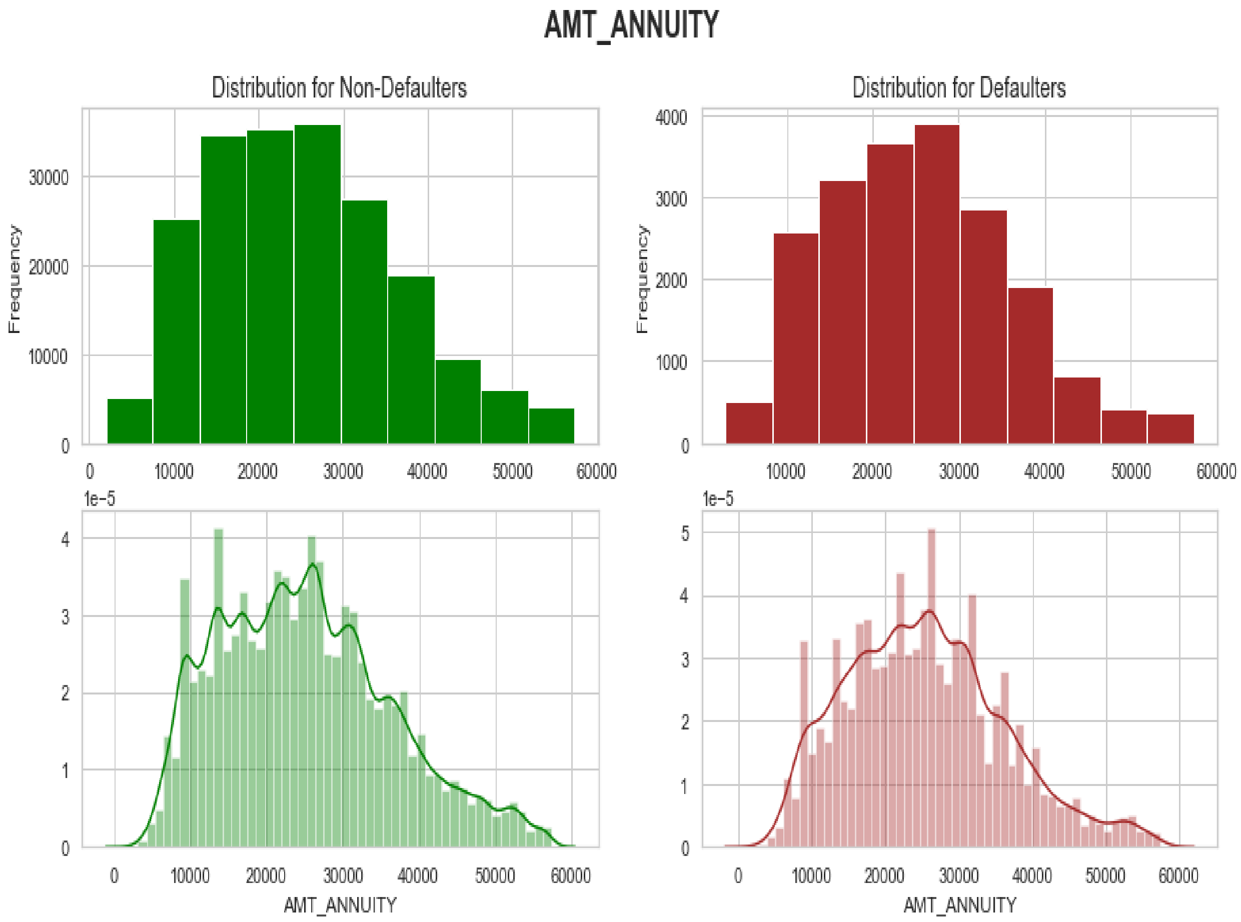
We can clearly see that top 10 pairs of variables for defaulters and non defaulters are the same.

Univariate Analysis of Numerical Columns

Columns analyzed:

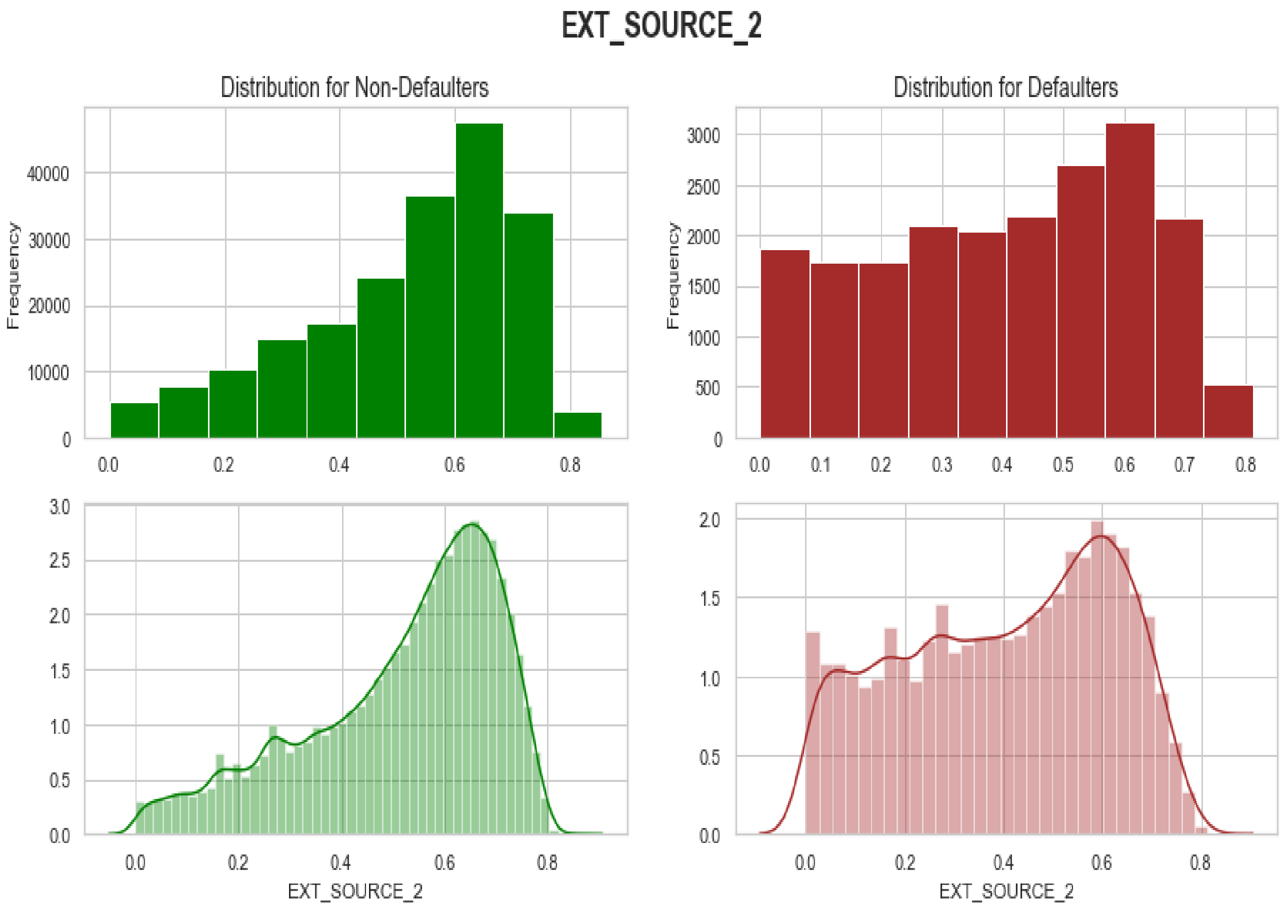
- Annuity
- Ext_Source_2
- Relative Population of the region
- Age

Annuity



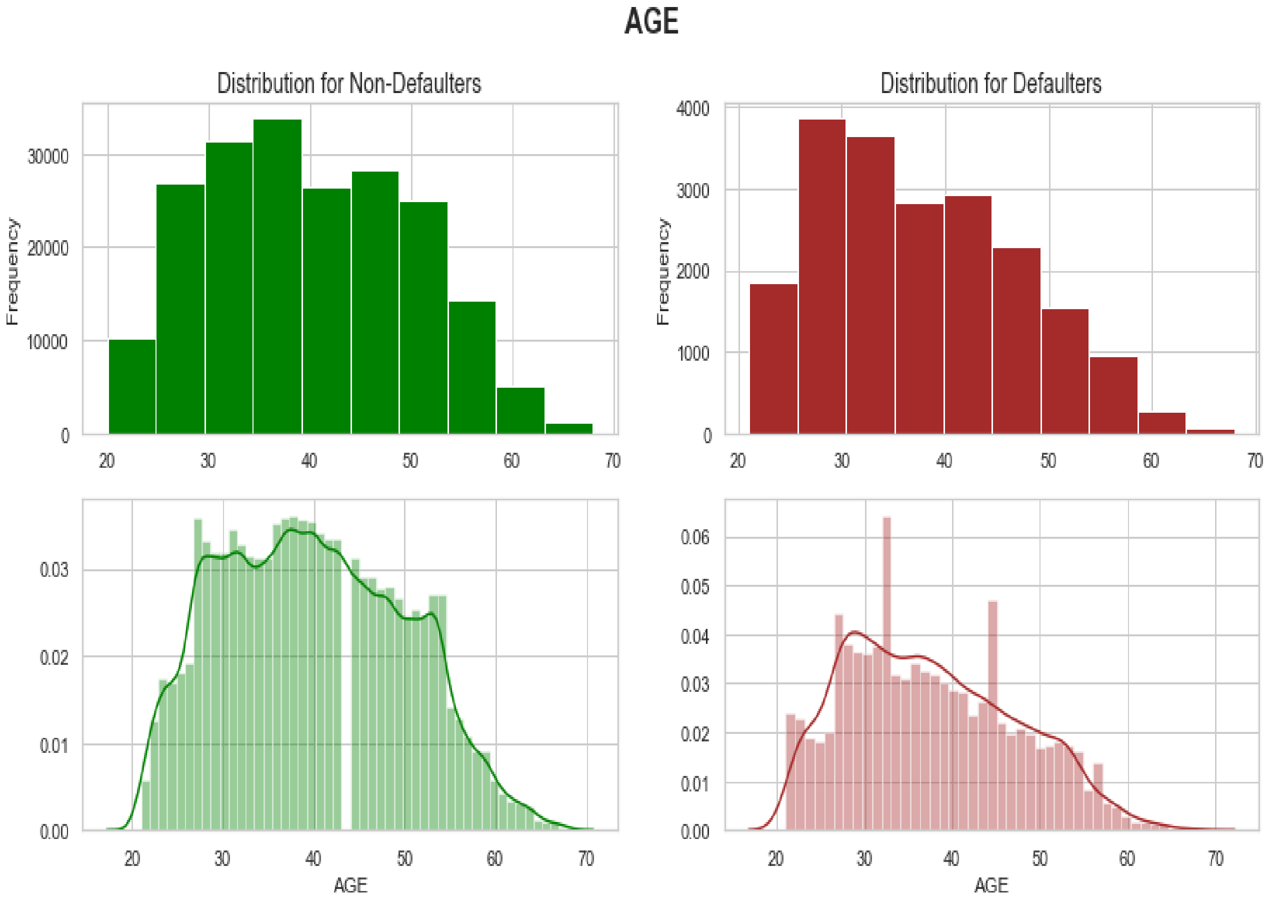
If Annuity is lower chances of being defaulter is less, as annuity increases the number of defaulters increase.

Ext_source_2



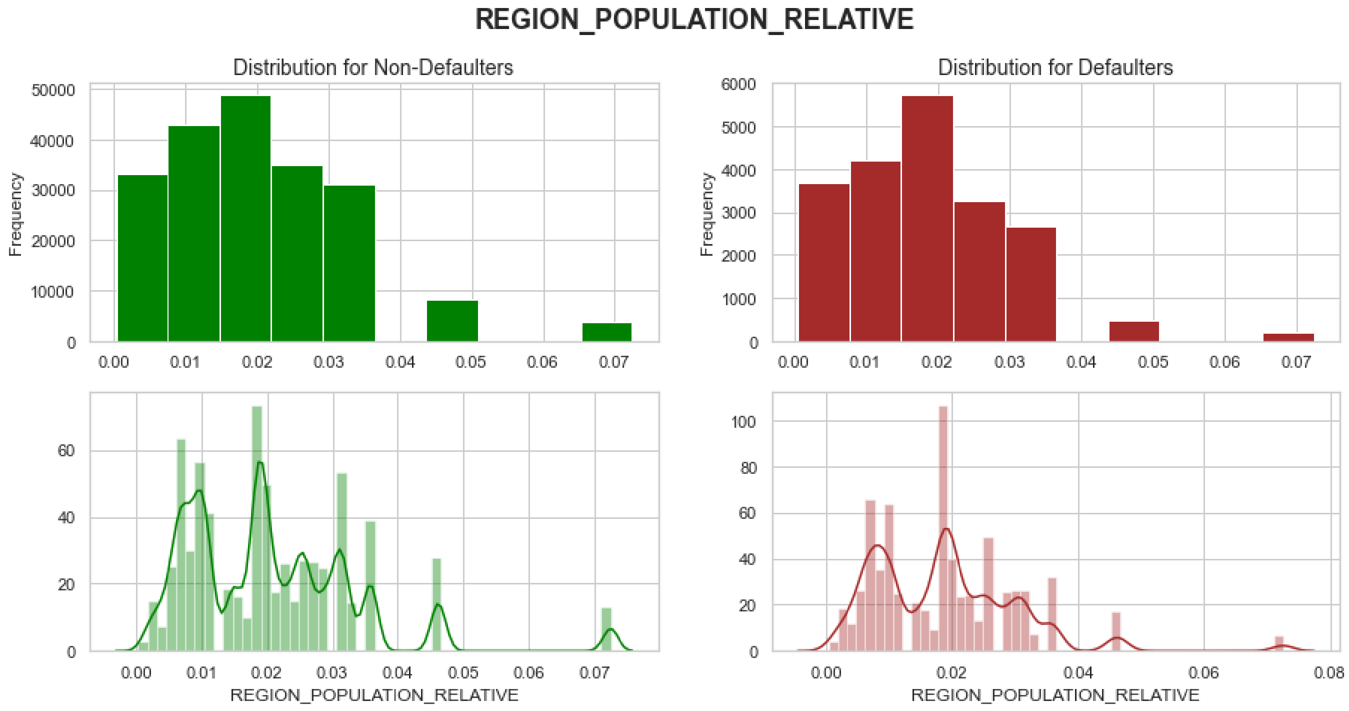
A lower value of EXT_SOURCE_2 score indicated high chances of defaults.

Age



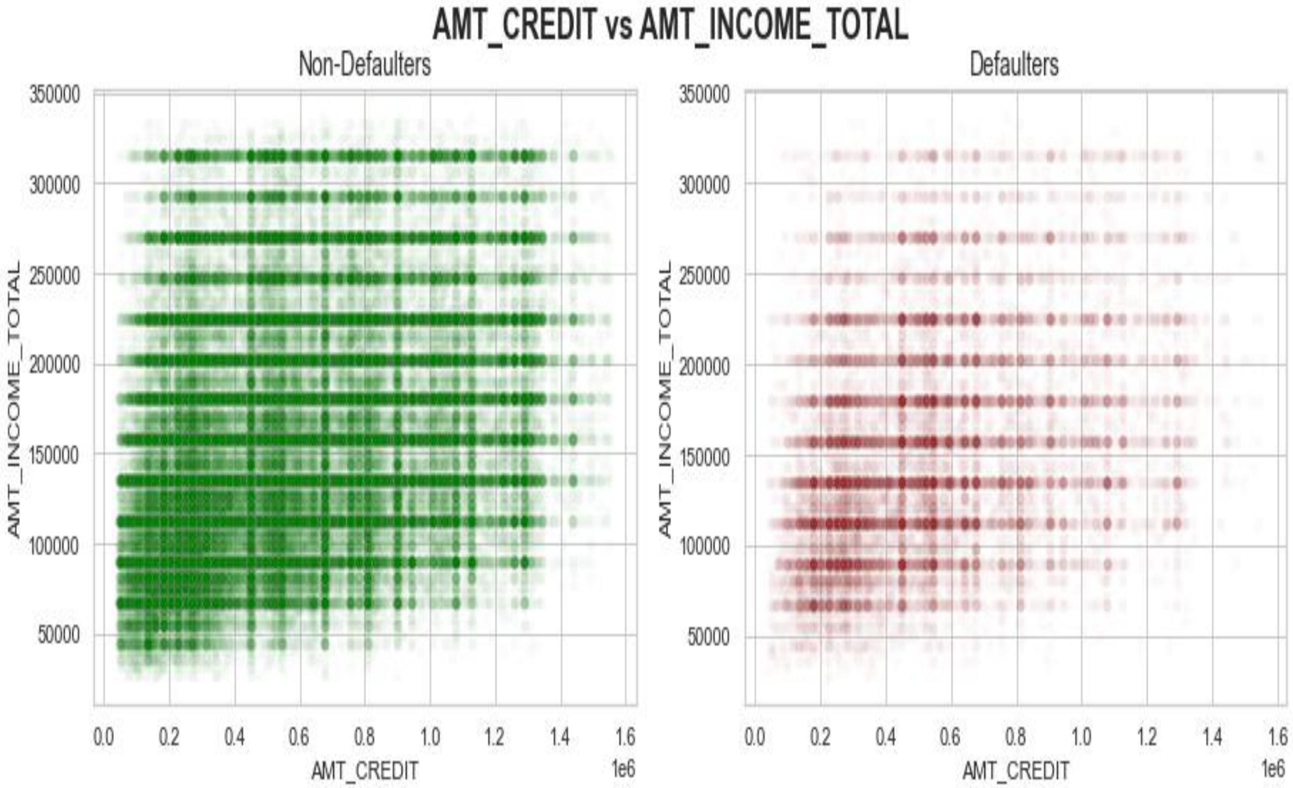
Defaulter distribution plot data is left skewed so younger people are likely to be defaulter as compared to older people

Relative population of the region

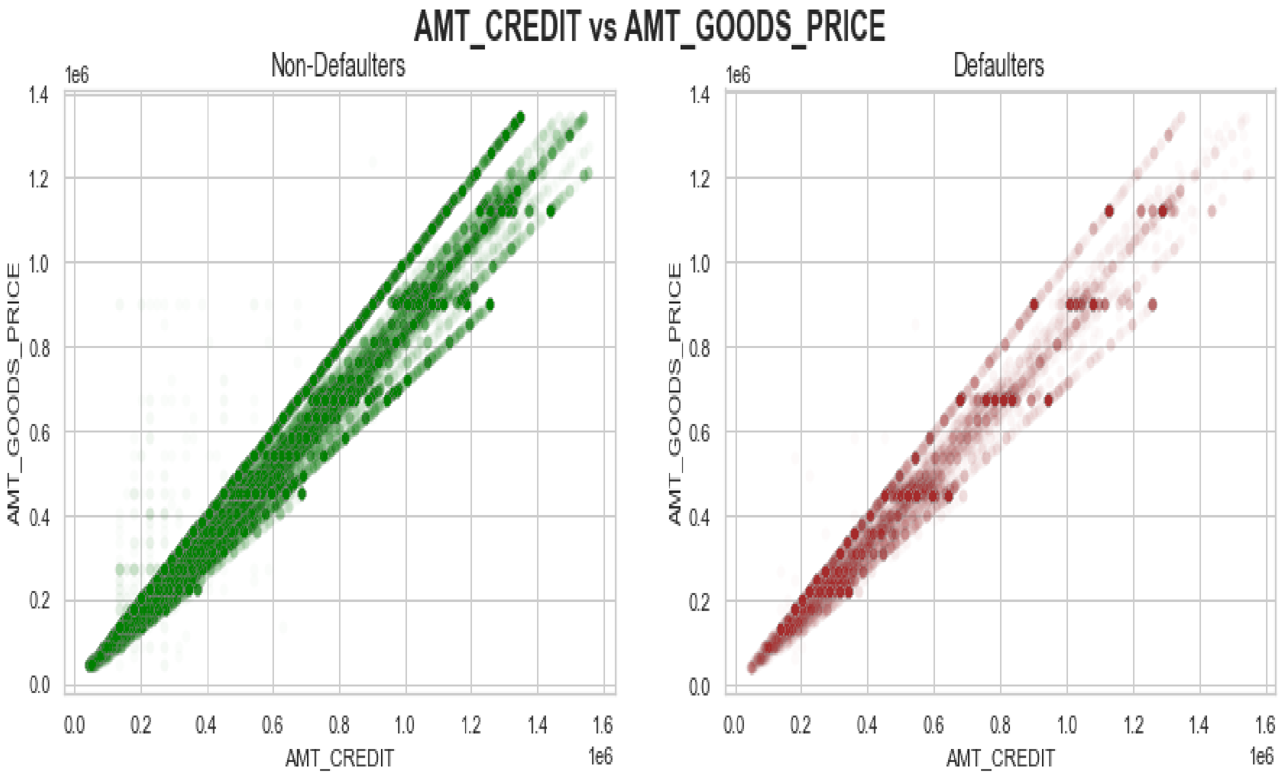


People living in higher density areas having lesser defaults

Bivariate Analysis of Numerical Columns



Lower density of defaults where income is higher than 300k or credit is lower than 200k.

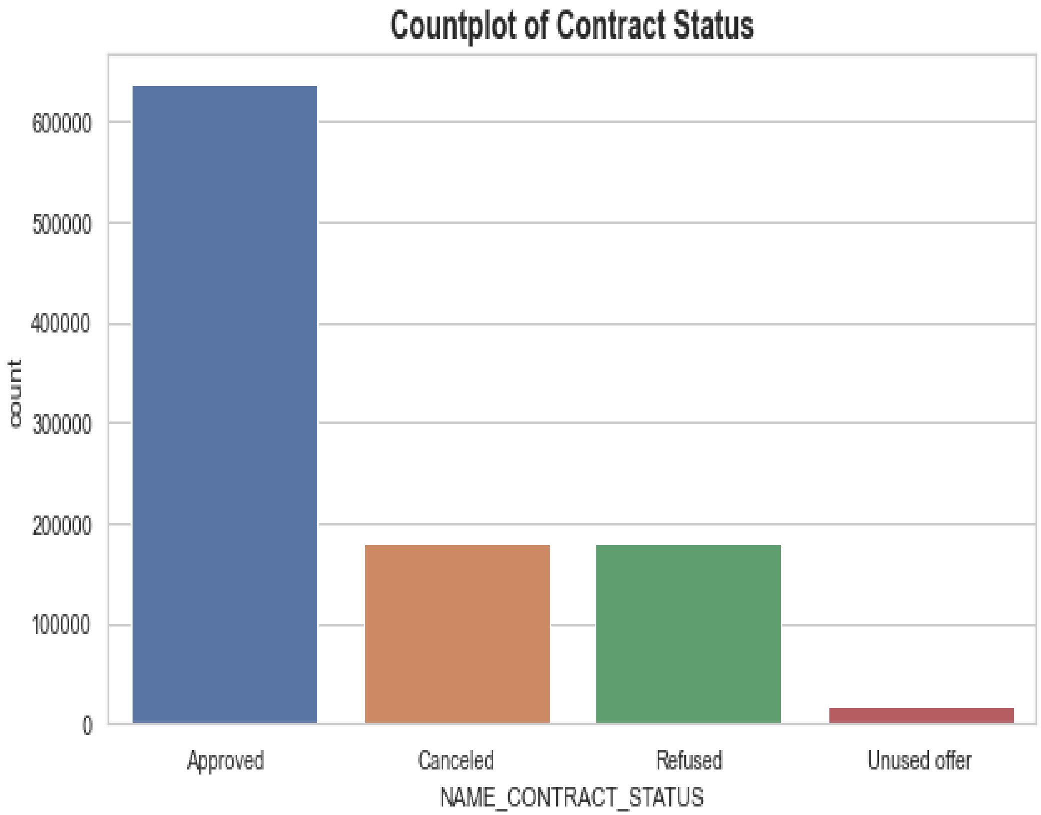


A high positive correlation found between credit amount and goods price as expected by correlation analysis.

Previous Application Analysis

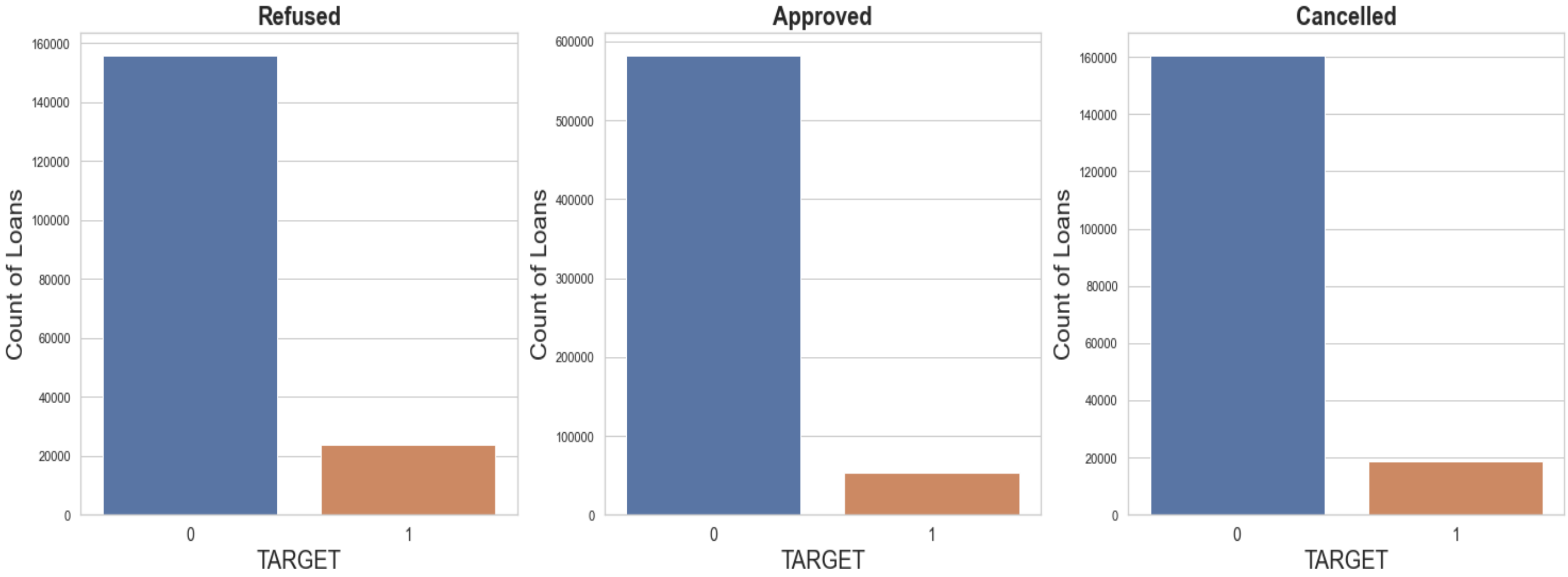
- Previous application data merged and treated.
- The merged data frame was split into four separate data frames as per Contract Status.
- The merged data frame was also split as per Target variable – Defaulters and Non-Defaulters.

Contract Status



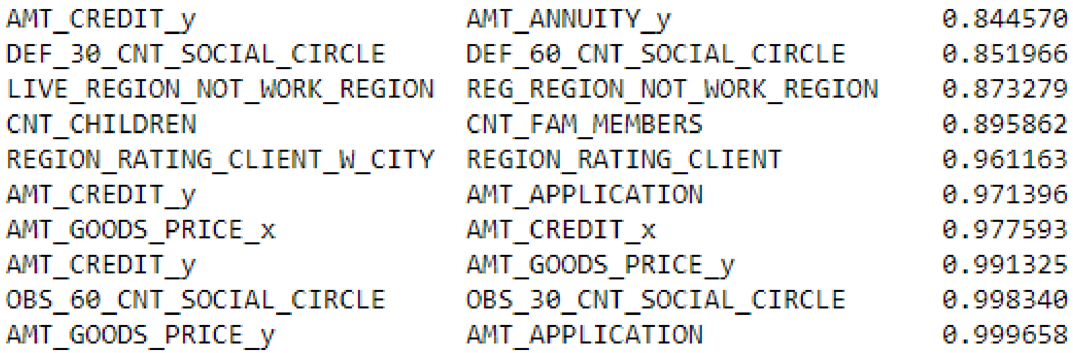
The approved applications seems to be on higher side as compared to other loan status.

Contract Status (contd.)



Loans which were previously refused had a higher default rate.

Bivariate Analysis on the Merged Dataset

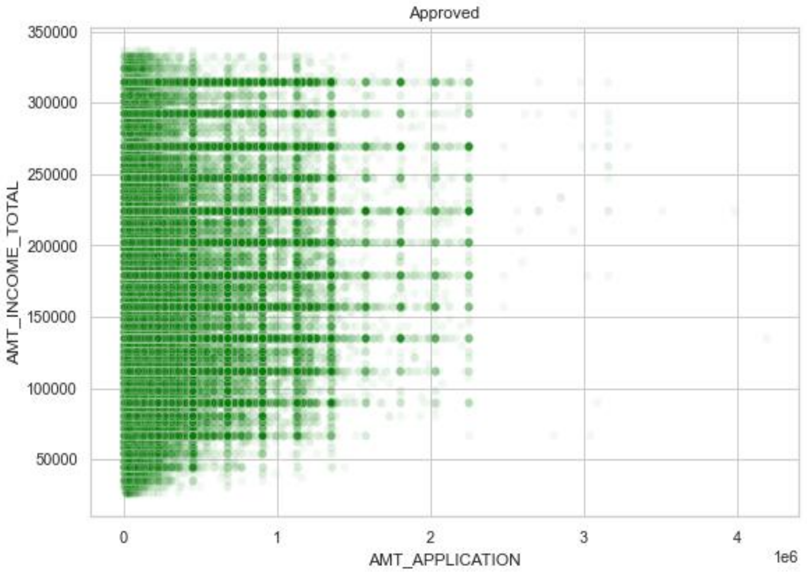


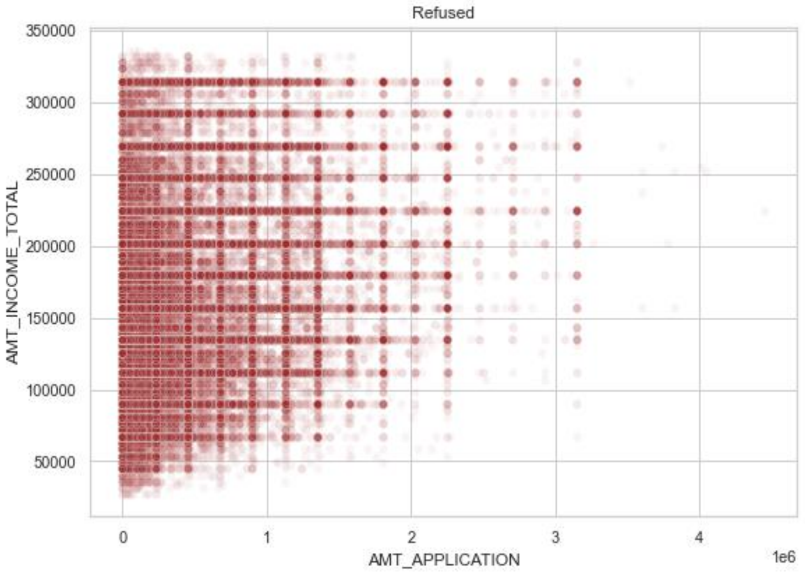
```
AMT_CREDIT_y                 AMT_ANNUITY_y                0.844570
DEF_30_CNT_SOCIAL_CIRCLE     DEF_60_CNT_SOCIAL_CIRCLE     0.851966
LIVE_REGION_NOT_WORK_REGION  REG_REGION_NOT_WORK_REGION   0.873279
CNT_CHILDREN                 CNT_FAM_MEMBERS              0.895862
REGION_RATING_CLIENT_W_CITY  REGION_RATING_CLIENT         0.961163
AMT_CREDIT_y                 AMT_APPLICATION              0.971396
AMT_GOODS_PRICE_x            AMT_CREDIT_x                 0.977593
AMT_CREDIT_y                 AMT_GOODS_PRICE_y            0.991325
OBS_60_CNT_SOCIAL_CIRCLE     OBS_30_CNT_SOCIAL_CIRCLE     0.998340
AMT_GOODS_PRICE_y            AMT_APPLICATION              0.999658
```

Top 10 absolute correlations in the merged dataset along with their correlation coefficients.

Please refer to the Python notebook for detailed heatmaps on the same.

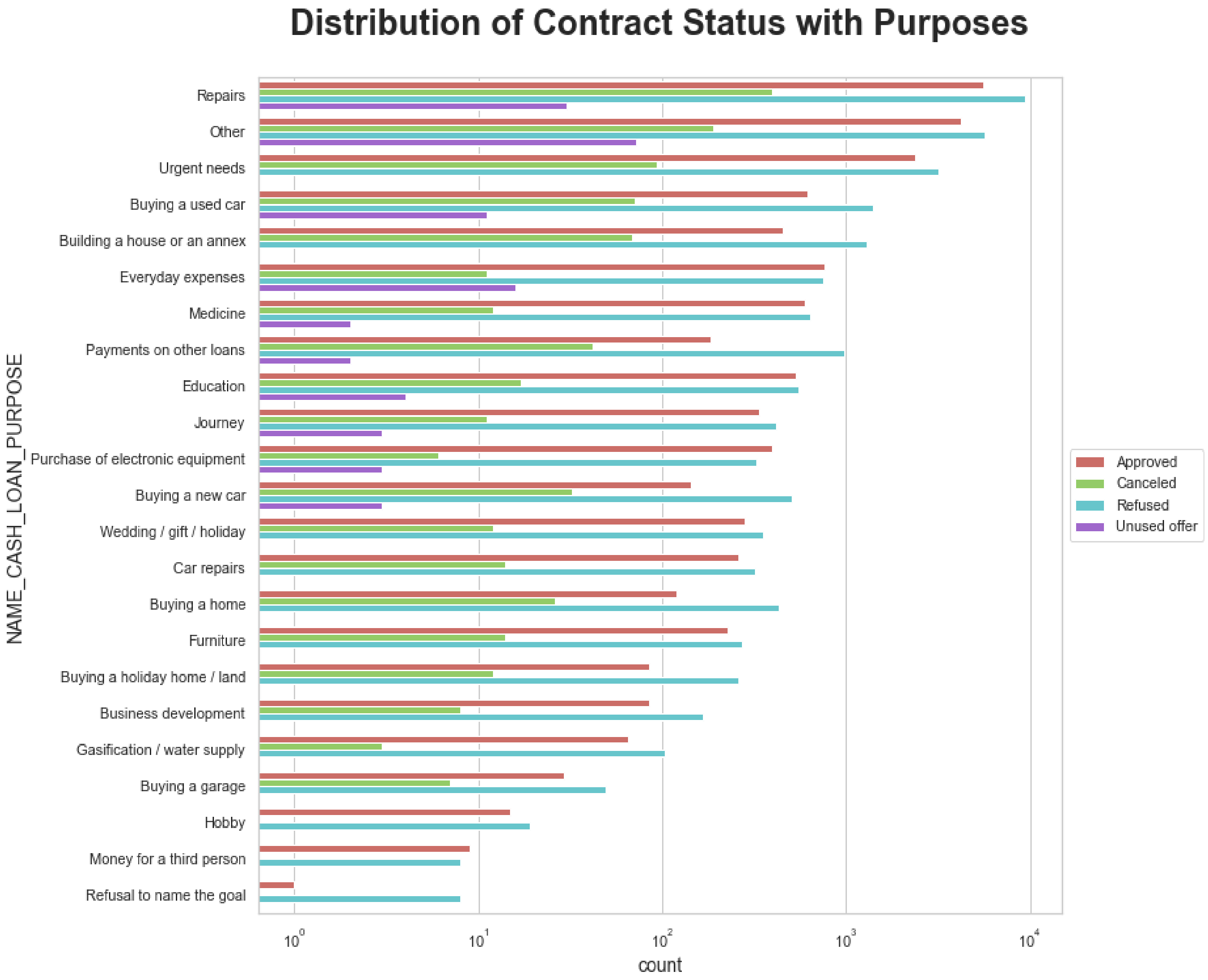
Loan Rejection Rate





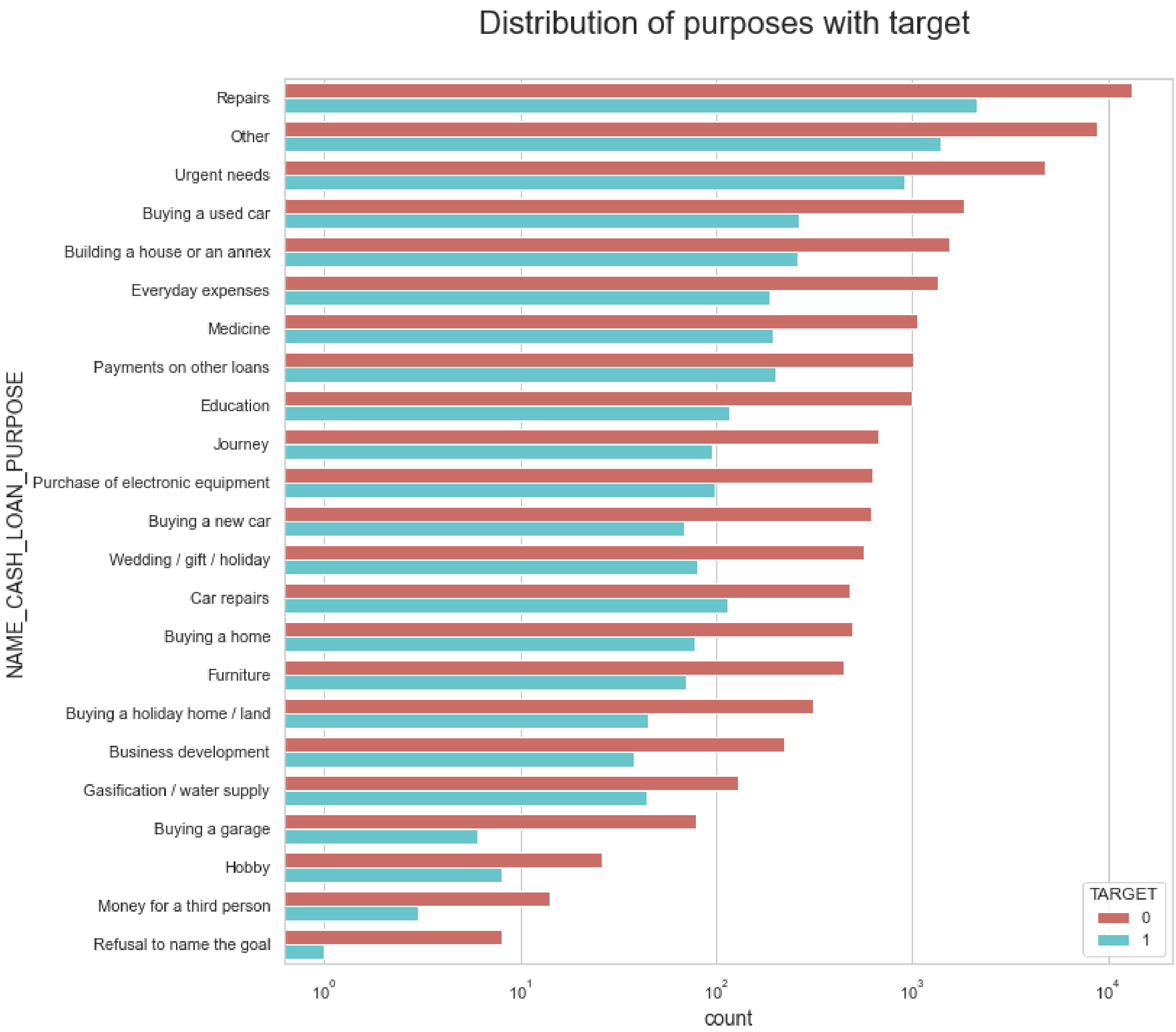
Loan request higher than 200k had a higher rejection rate compared to approval rate. Also, loan rejection rate was much lower if the income was higher than 300k.

Contract Status with Purposes



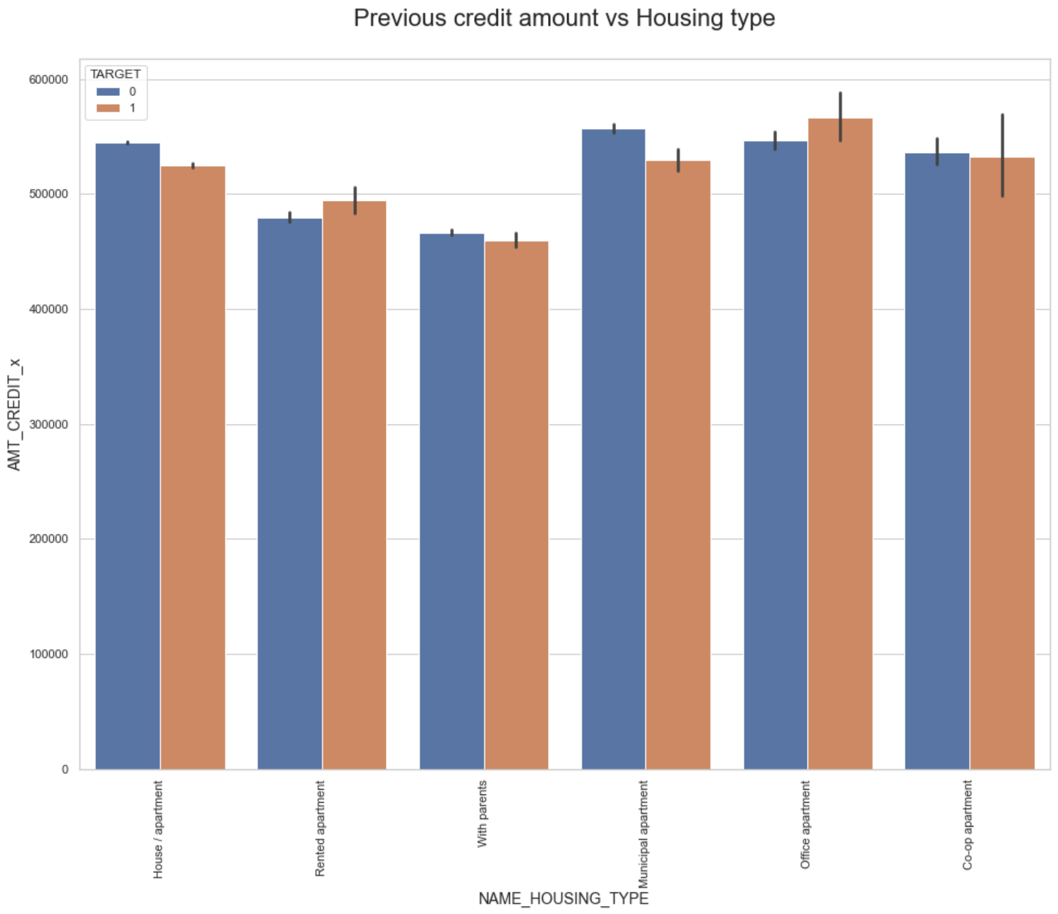
- Most rejection of loans came from purpose 'repairs'.
- For education purposes we have equal number of approves and rejection.
- Paying other loans , buying a home and buying a new car is having significant higher rejection than approves.

Contract Status with Targets



- Loan purposes with 'Repairs' are facing more difficulties in payment on time.
- There are few places where loan payment is significant higher than facing difficulties - they are 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car' and 'Education'

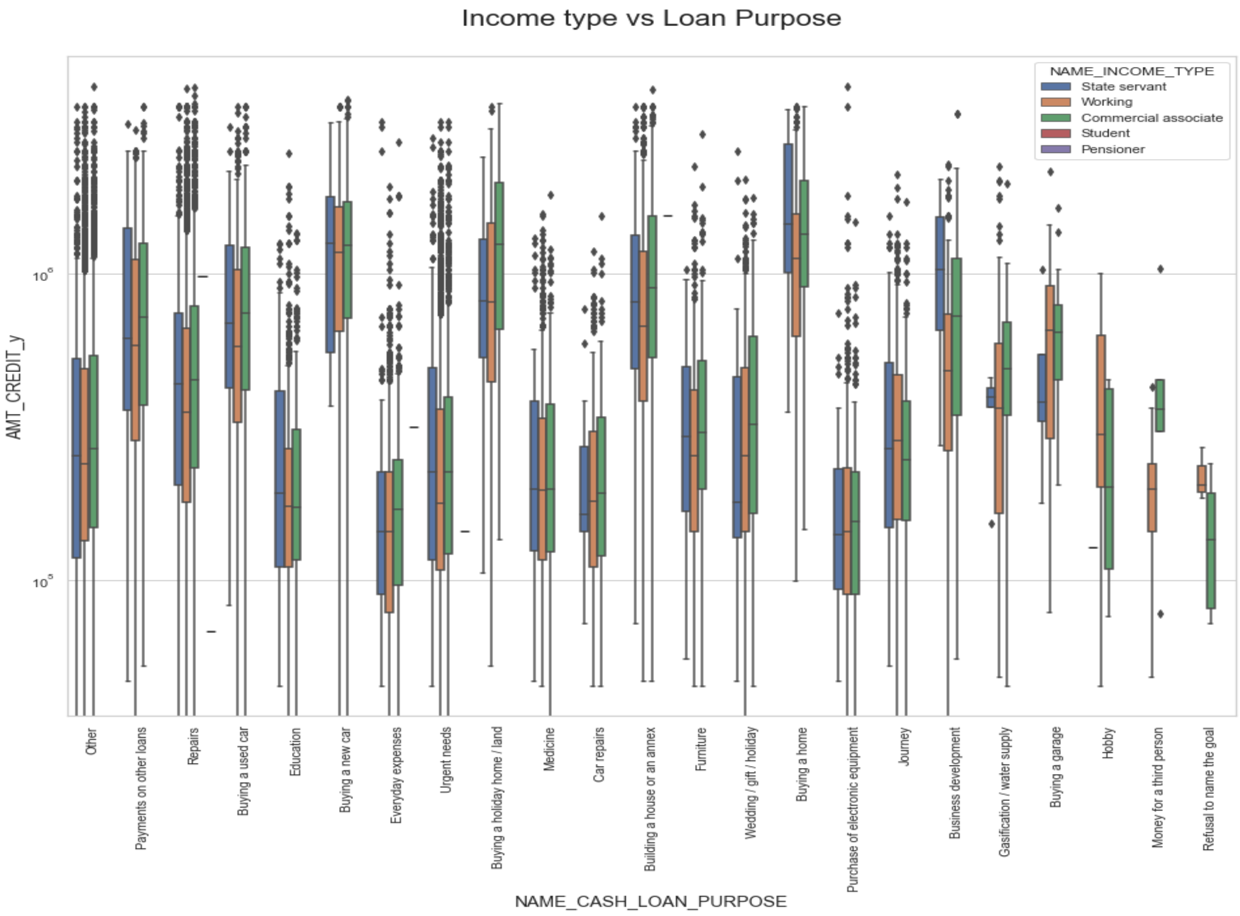
Previous credit amount vs Housing type



Here for Housing type, office apartment is having higher credit of target 0 and co-op apartment is having higher credit of target 1. So, we can conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment.

Bank can focus mostly on housing type with House\apartment or municipal apartment for successful payments.

Income type vs Loan Purpose



- The credit amount of Loan purposes like 'Buying a home', 'Buying a land', 'Buying a new car' and 'Building a house' is higher.
- Income type of state servants have a significant amount of credit applied
- Money for third person or a Hobby is having less credits applied for.

Thank You!