PREDICTING INDICATORS FOR BANK LOAN DEFAULTS THROUGH EDA

SUBMITTED BY MRINMOY THOKDAR

EDA : BANK LOAN DEFAULT RISK ANALYSIS

We have been given two datasets, one of them describing current loan applications with the TARGET variable being whether the clients are having payment difficulties or not, and the other one is historical data of previous applications and their approval/rejection status.

Business Objective : To gather useful and actionable business insights to predict which factors can contribute to the loan default cases, so that lender can improve its loan approval/rejection process.

01

From the ‘application_data’ dataset, there are a total of 122 attributes per application id. Among which we see a lot of unnecessary columns, which are completely irrelevant to our analysis. We will first remove them to get a better grip of the dataset.

02

During null value analysis, we see a no of attributes are having more than 45 % missing values, imputation of these values should not be a preferred approach. So , we will drop these columns as well. For some columns we have used to replace null values for continuous type and mode for categorical type. Data has been standardised and data types have been checked thereafter.

03

For the TARGET variable, almost 91% of entries are for clients not having payment difficulties, so there is definitely skewness to this data. The imbalance ratio has been found to be around 9. We tried to take a sample of the data by randomly choosing certain no of rows, bur the imbalance ratio remained same.

04

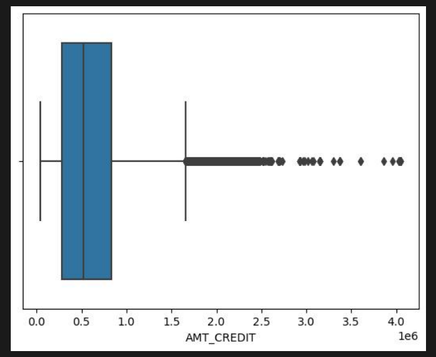
So for better analysis we have bifurcated the dataset into two datasets, one having entries with TARGET variable 1 which means clients having payment difficulties, and the other having entries with TARGET variable 0 which signifies all other cases. We have then tried to find association between different variables taking each dataset separately.

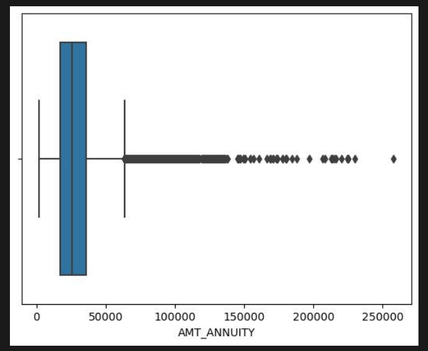
05

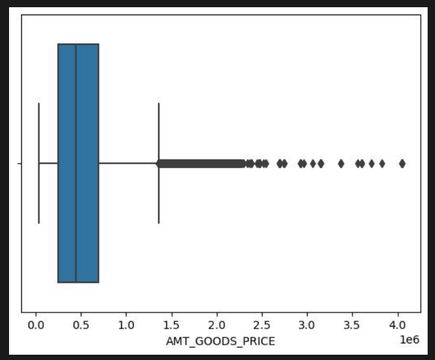
We have also separated the remaining columns into continuous and categorical columns based on no of unique values present in them. Some of the continuous columns have been converted to categorical columns to get better understanding from them, for ex, total income range , credit amount range, age groups etc to understand the variation er.

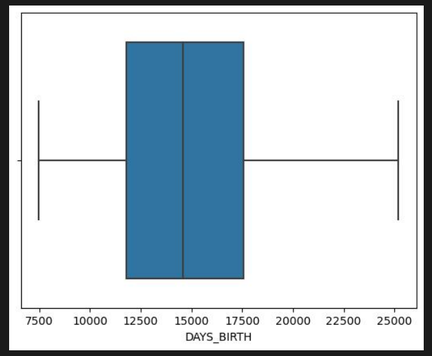
UNIVARIATE ANALYSIS FOR CONTINUOUS COLUMNS

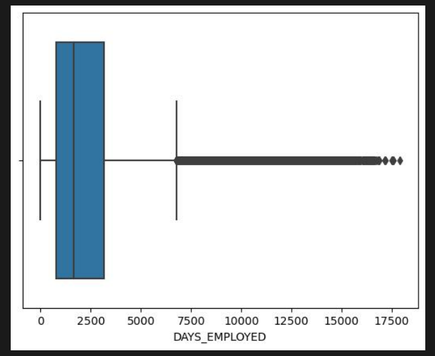
By plotting boxplots and distribution plots , we see there are a lot of outliers (positively skewed data) for the 'AMT_INCOME_TOTAL' , 'AMT_CREDIT' , 'AMT_ANNUITY' , 'DAYS_EMPLOYED' columns, which are expected in such datasets, we will not drop or cap the outliers since no of outliers are significant. While calculating aggregating features we will take median or calculate different quartiles to get insights , thus ignoring the outliers.







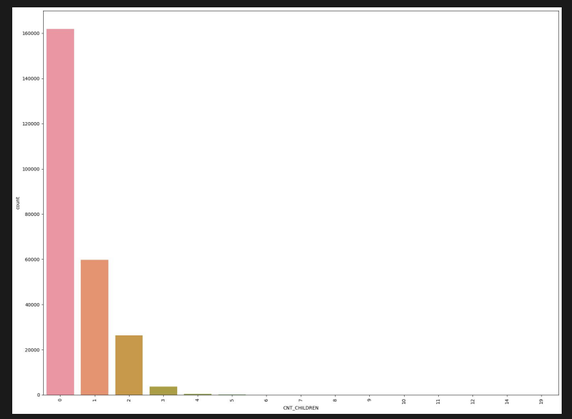


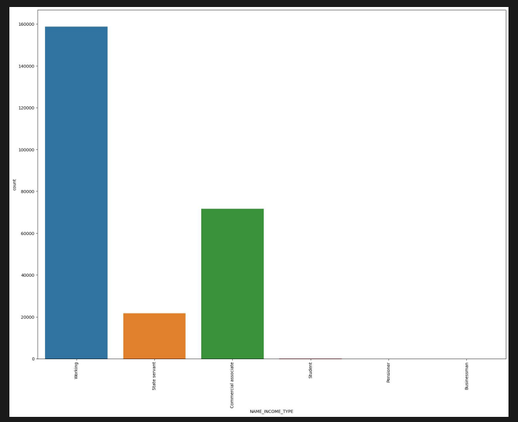


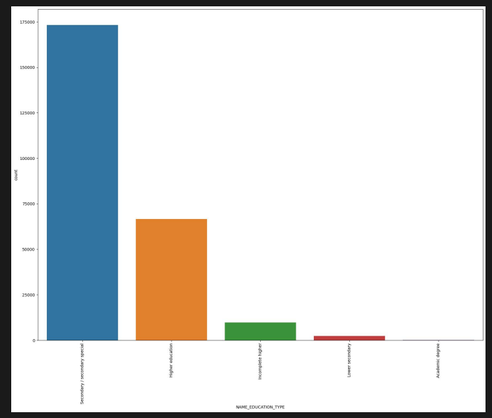
For 'AMT_INCOME_TOTAL' , 'AMT_CREDIT' , 'AMT_ANNUITY' , 'DAYS_EMPLOYED' , 'DAYS_BIRTH' , we have binned to small groups to understand the trend groupwise.

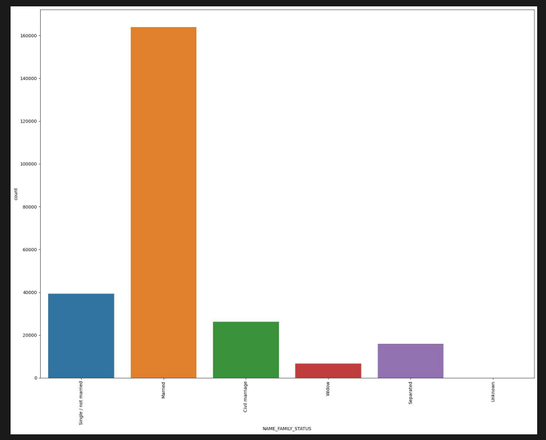
SEGMENTED UNIVARIATE ANALYSIS FOR CATEGORICAL COLUMNS

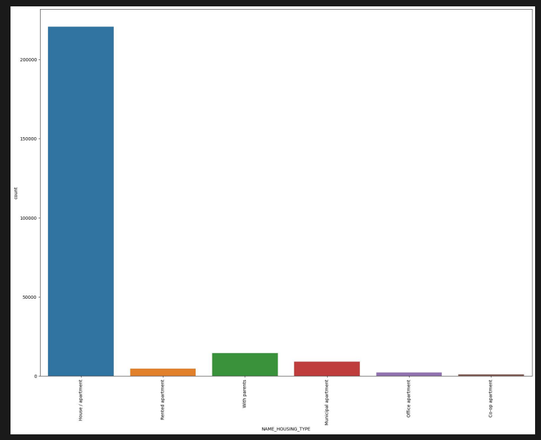
For categorical columns we have used count plots to understand the frequencies of different categories.











DEFINING A NEW PARAMETER TO UNDERSTAND DATA BETTER

It has been observed from the univariate analysis of the columns, certain categories has very high frequency than the rest in each column. For ex, Laborers in OCCUPATION_TYPE attribute, which means data will tend to lean towards this category.

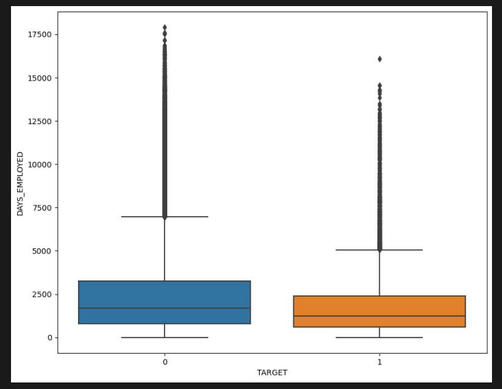
So we will use, **default percentage** which has been calculated as below:

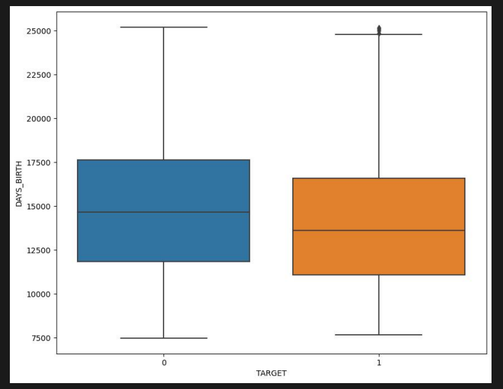
default percentage = (no of entries of specific category in specific column having repayment difficulties) /(no of entries of specific category in specific column having repayment difficulties + no of entries of specific category in specific column having no repayment difficulties)

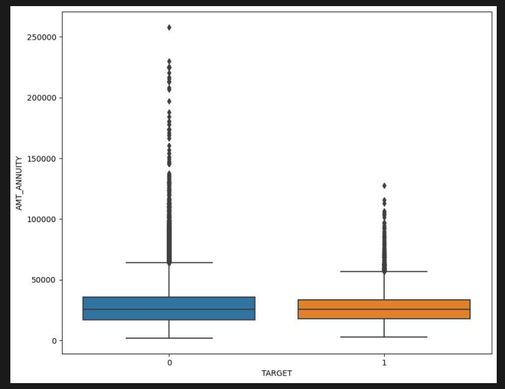
This is nothing but to normalize the data and to offset the skewness.

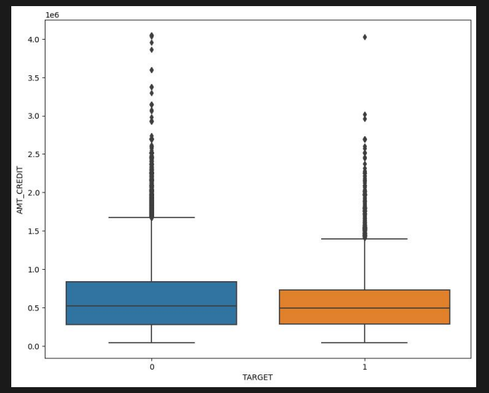
VARIATION OF CONT. COLUMNS WRT TARGET VARIABLE

We will now examine continuous variables wrt TARGET variable. Here we have tried to understand the spread of the data for each column.







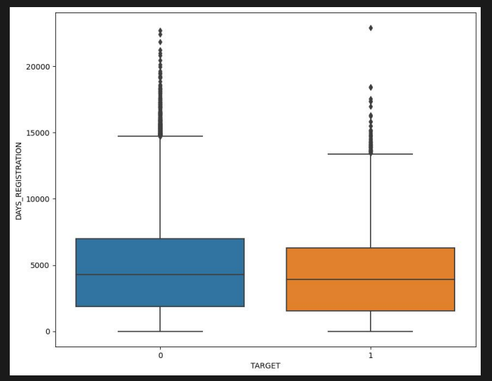


Based on the above plots, following can be observed :

1. People with more days with employment in the current job are more likely to repay which is expected due to stability factor.
2. Also generally older people are more likely to repay.
3. There are a lot of outliers (except in 'DAYS_BIRTH' and 'DAYS_ID_PUBLISH') in the columns, which are expected in this kind of dataset.

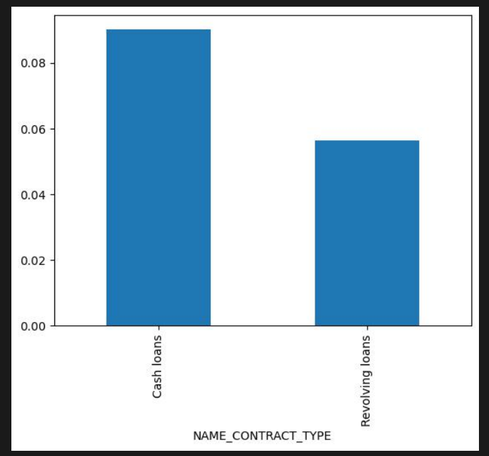
Less obvious observations :

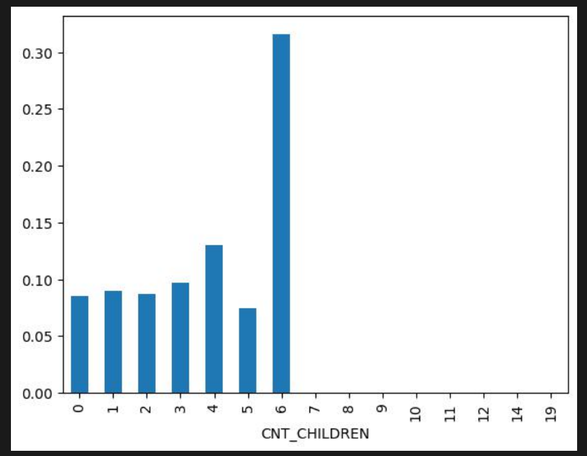
1. People having higher annuity and higher credit amount are more likely to repay.
2. People who changed their identity document or registration recenly are less likely to repay.

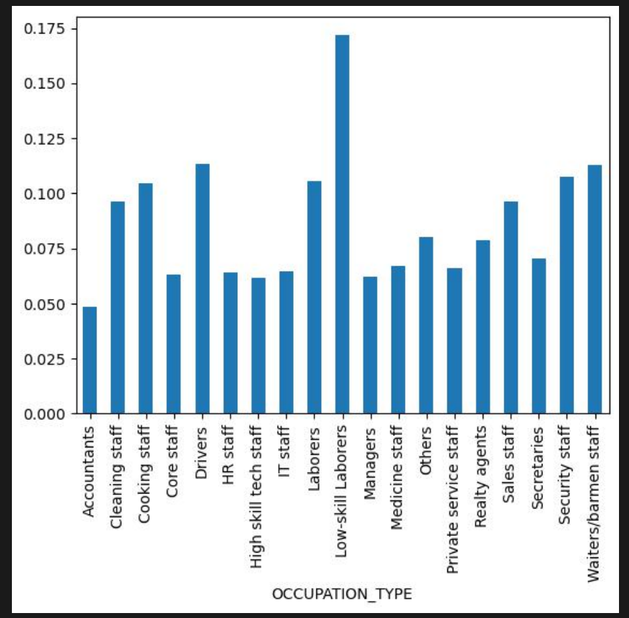


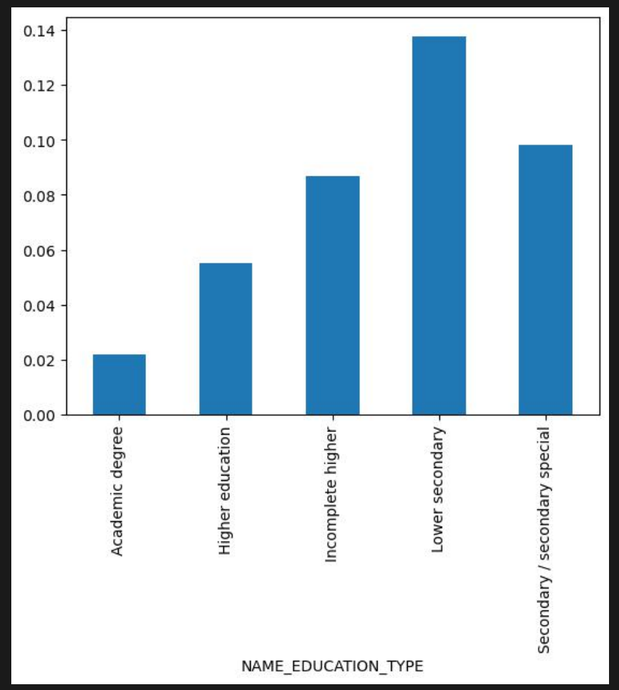
UNDERSTANDING CATEGORICAL COLUMNS WITH NEW PARAMETER

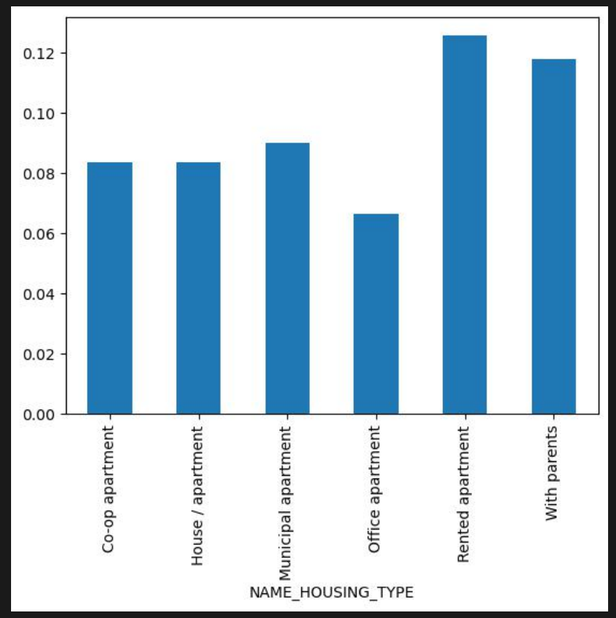
Now for the categorical variables we will use default and / repay percentage indicator to understand the association with TARGET variable:

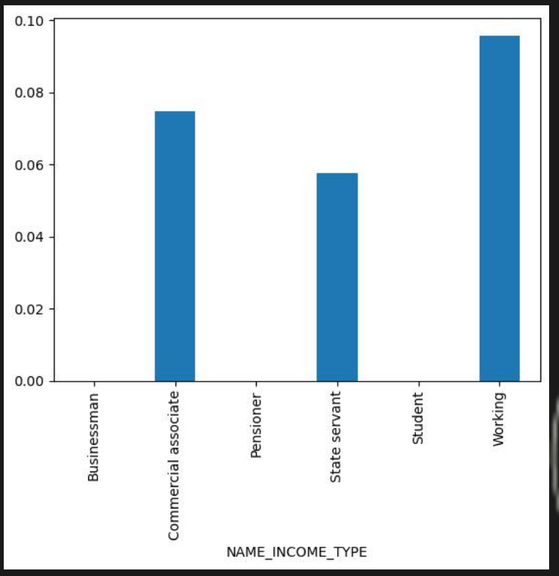












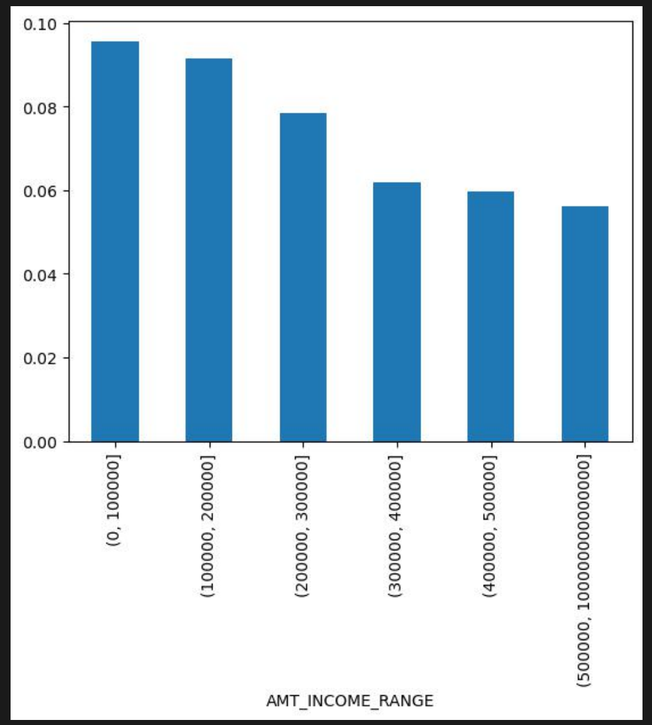
OBSERVATIONS FROM ABOVE PLOTS

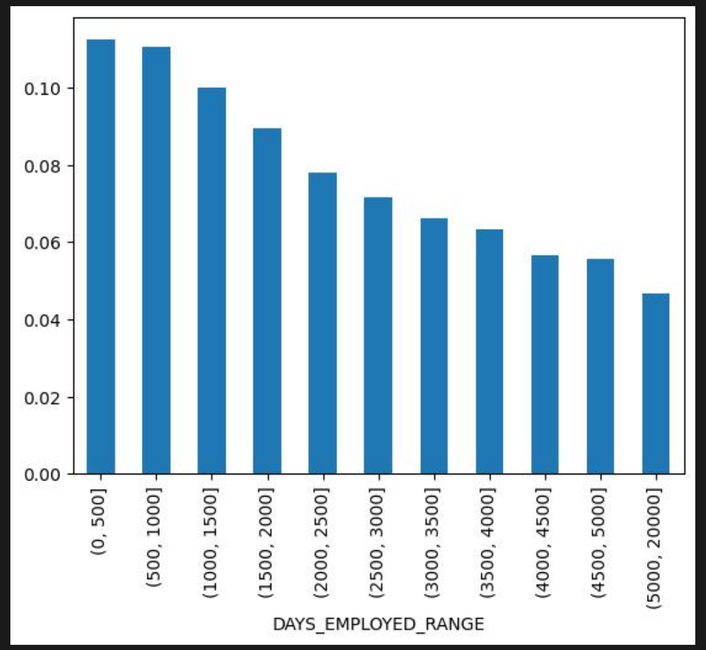
- Cash loans are at a higher risk of default.
- With more children the chance of default is generally higher.
- People in 'Working' category less likely to repay.
- People with lower levels of education and with low skills have high default rate.
- Student, Pensioner and Businessman are less likely to default.
- People living in rented apartments have high % of default rate and lowest default rate observed for people living in office apartments.

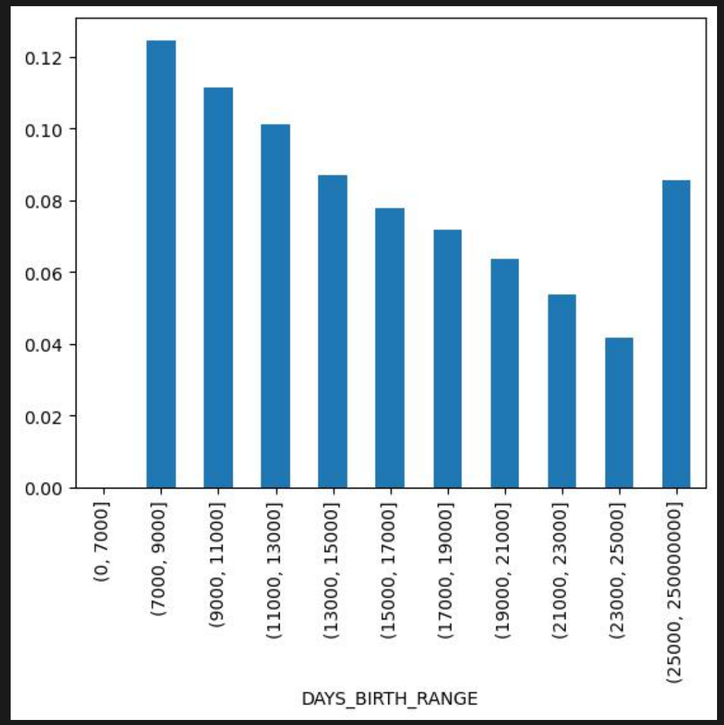
SEGMENTED UNIVARIATE ANALYSIS FOR BINNED COLUMNS

Now for the binned columns :

We have observed the following:





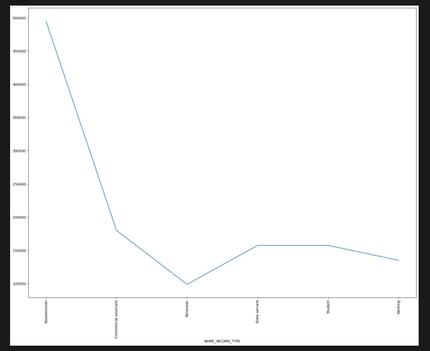


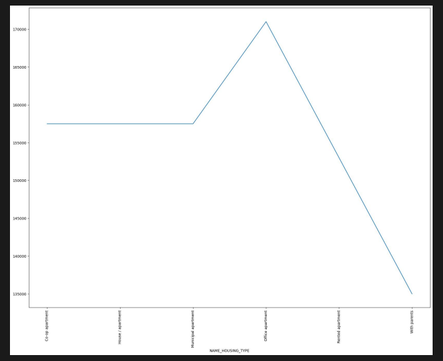
- People with higher income tend to repay and less likely to default.
- We see people with more days of current employment tend to repay due to more stability.
- For AMT_CREDIT_RANGE and AMT_ANNUITY_RANGE, we see a peak in the distribution. For Credit range, the default percentage peaks at 500000-600000 and for annuity range , its at 25000-35000.

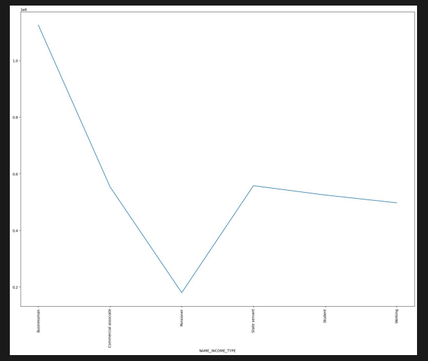
BI/MULTIVARIATE ANALYSIS

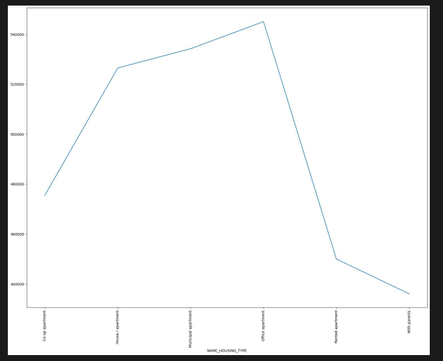
For continuous columns there is no obvious association, as can be seen in the scatter plots.

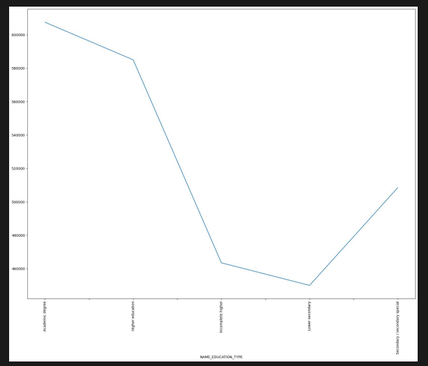
For categorical variables :







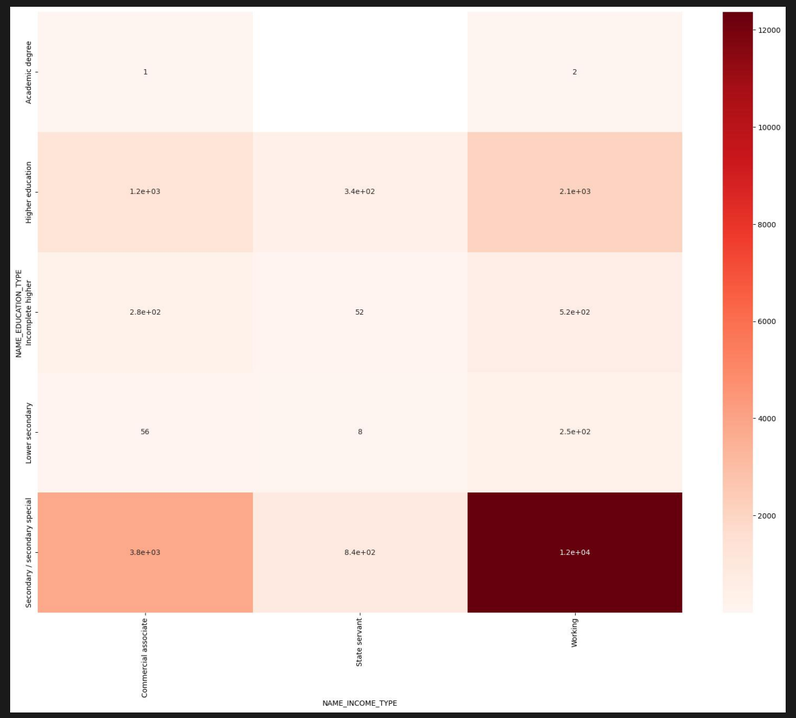


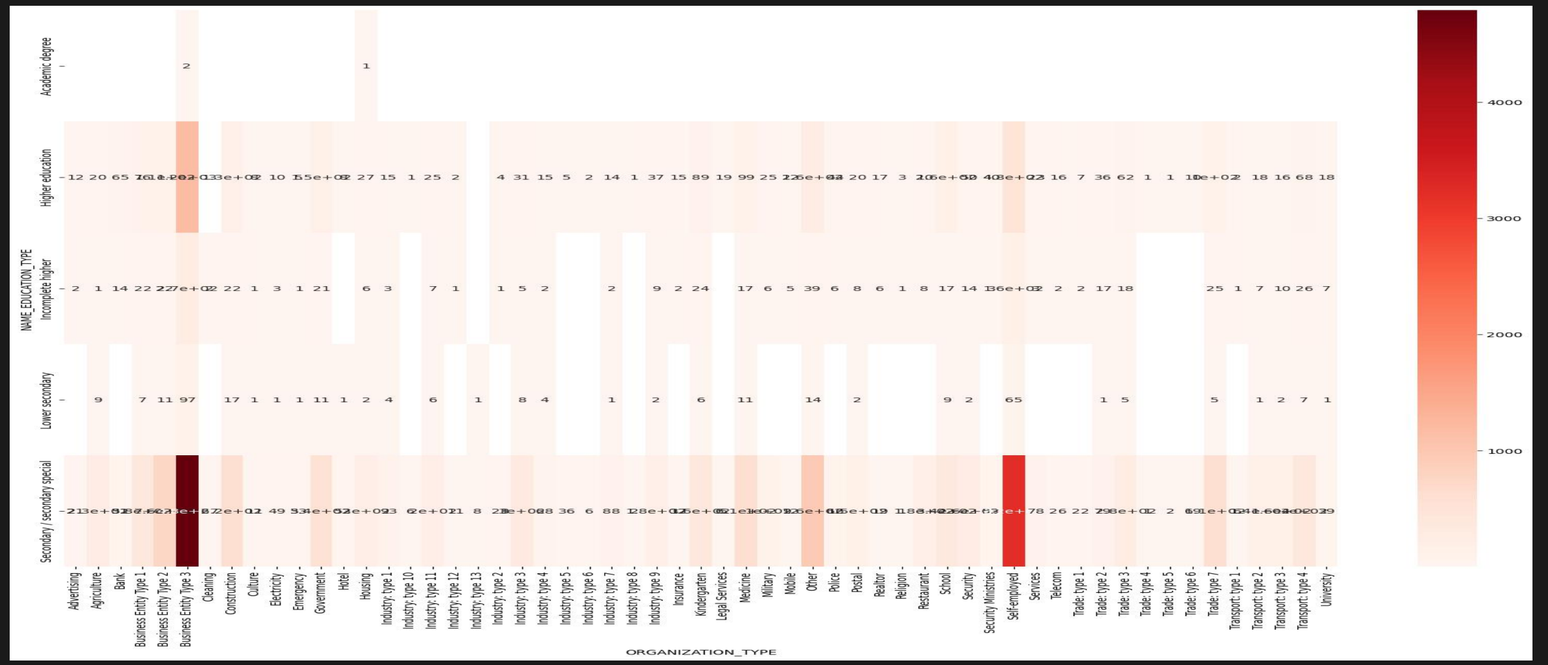


- Business people tend to have higher income and applies for higher credit amount.
- the same goes for people living in office apartments.
- as usual with higher education , income also increases. But credit amount seems to be highest for people having secondary education.

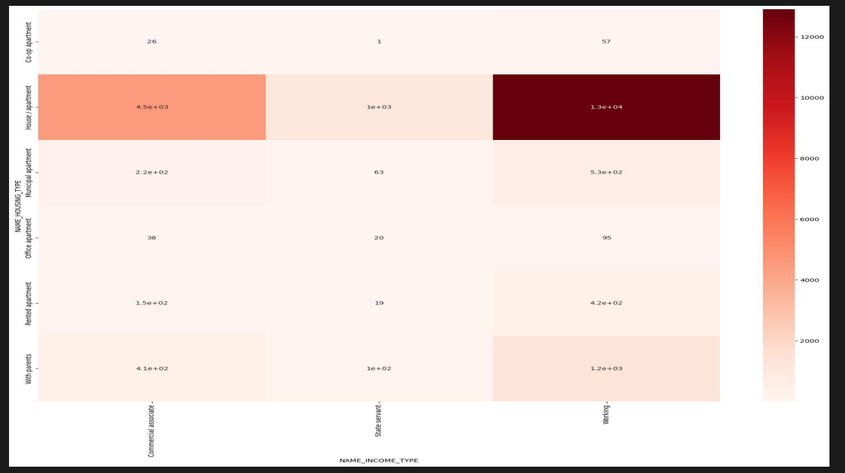
BI/MULTIVARIATE ANALYSIS

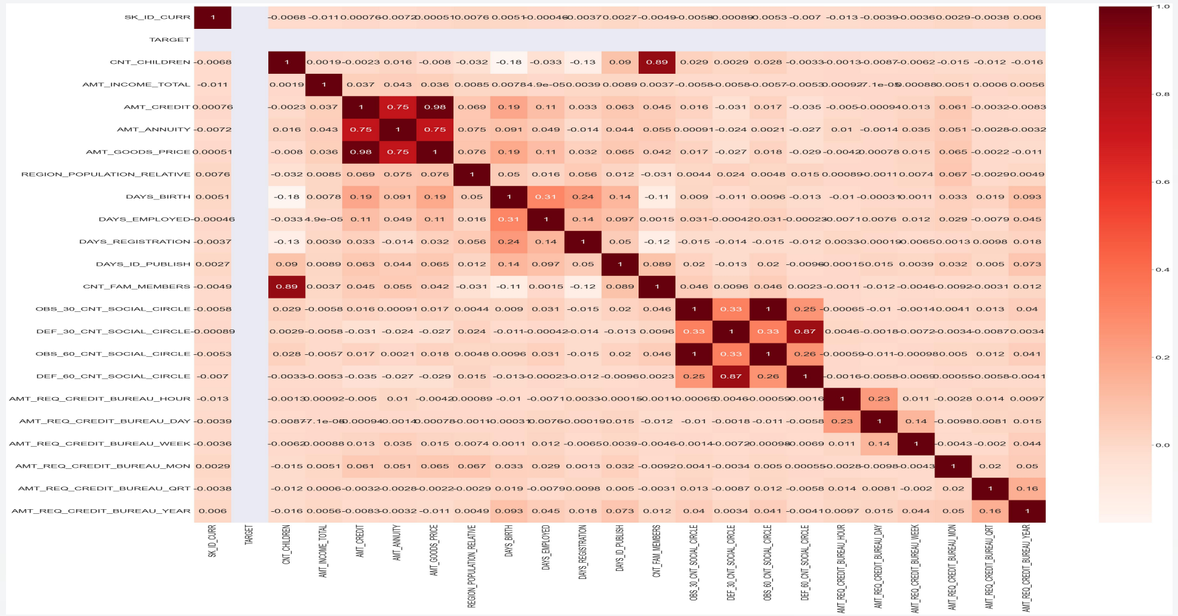
By analysing different categorical variables with different education type, we observe :





- 'Working' people with 'secondary/secondary special' are more in number to default.
- Similarly self employed type people are more in numbers to default across education levels.
- people having academic degree seems to have lowest rate of defaults.
- By analysing , we see business entity type 3 people are more likely to default.



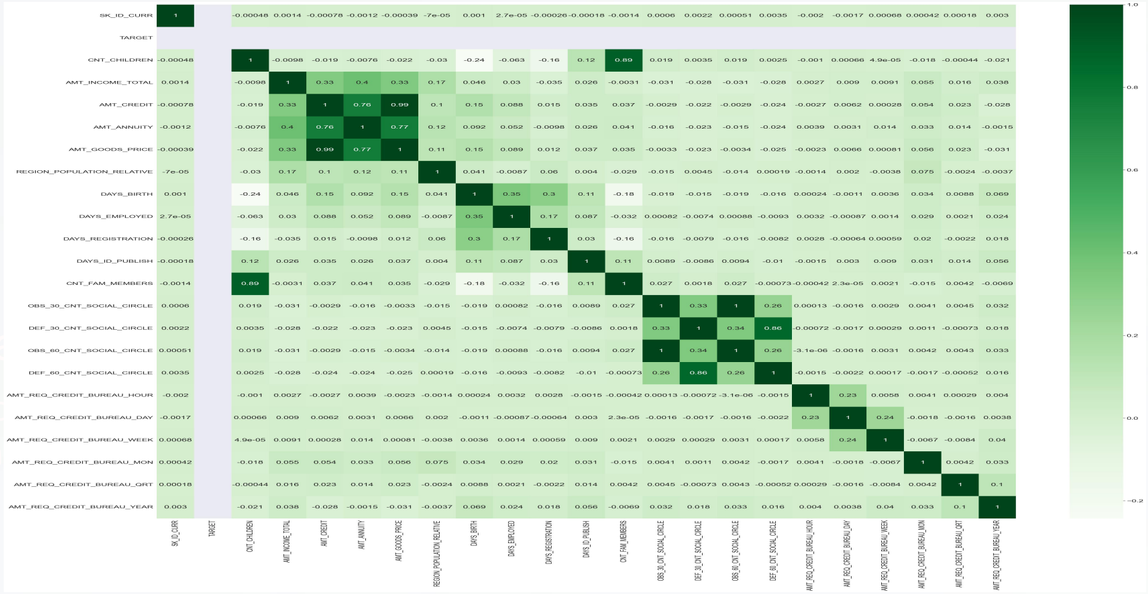


Top 10 correlations between variables for ‘Clients having payment difficulties’

1 AMT_GOODS_PRICE VS AMT_CREDIT - 0.98
2 CNT_FAM_MEMBERS VS CNT_CHILDREN - 0.89
3 DEF_60_CNT_SOCIAL_CIRCLE VS DEF_30_CNT_SOCIAL_CIRCLE - 0.76
4 DEF_30_CNT_SOCIAL_CIRCLE VS OBS_30_CNT_SOCIAL_CIRCLE - 0.33
5 OBS_60_CNT_SOCIAL_CIRCLE VS DEF_30_CNT_SOCIAL_CIRCLE - 0.33
6 DAYS_EMPLOYED VS DAYS_BIRTH - 0.31
7 DEF_60_CNT_SOCIAL_CIRCLE VS OBS_60_CNT_SOCIAL_CIRCLE - 0.26
8 DEF_60_CNT_SOCIAL_CIRCLE VS OBS_30_CNT_SOCIAL_CIRCLE - 0.25
9 DAYS_BIRTH VS DAYS_REGISTRATION - 0.24
10 AMT_REQ_CREDIT_BUREAU_DAY VS AMT_REQ_CREDIT_BUREAU_HOUR - 0.23

Top 10 correlations between variables for ‘all other cases’.

1 AMT_GOODS_PRICE VS AMT_CREDIT - 0.99
2 CNT_FAM_MEMBERS VS CNT_CHILDREN - 0.89
3 DEF_CNT_SOCIAL_CIRCLE VS DEF_30_CNT_SOCIAL_CIRCLE - 0.86
4 AMT_CREDIT VS AMT_ANNUITY - 0.76
5 AMT_ANNUITY VS AMT_INCOME_TOTAL - 0.4
6 DAYS_EMPLOYED VS DAYS_BIRTH - 0.35
7 OBS_60_CNT_SOCIAL_CIRCLE VS DEF_30_CNT_SOCIAL_CIRCLE - 0.34
8 DEF_30_CNT_SOCIAL_CIRCLE VS OBS_30_CNT_SOCIAL_CIRCLE - 0.33
9 AMT_INCOME_ VS AMT_CREDIT - 0.33
10 DAYS_BIRTH VS DAYS_REGISTRATION - 0.3

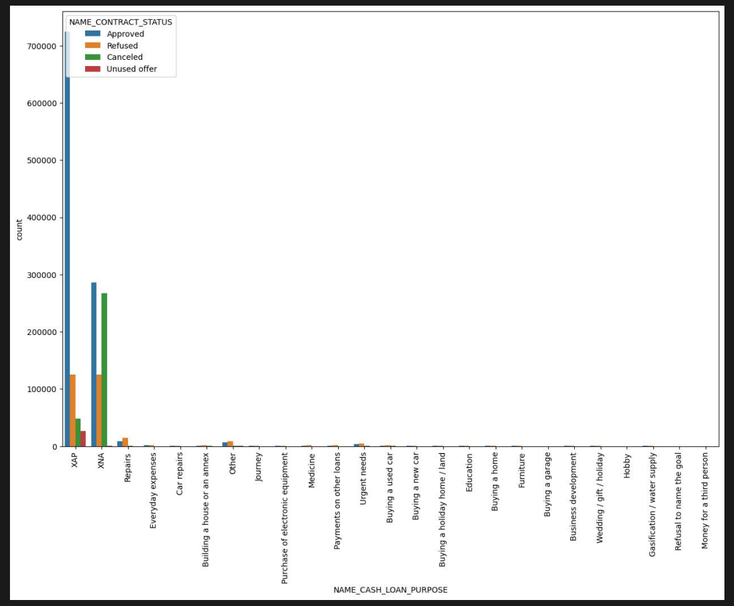


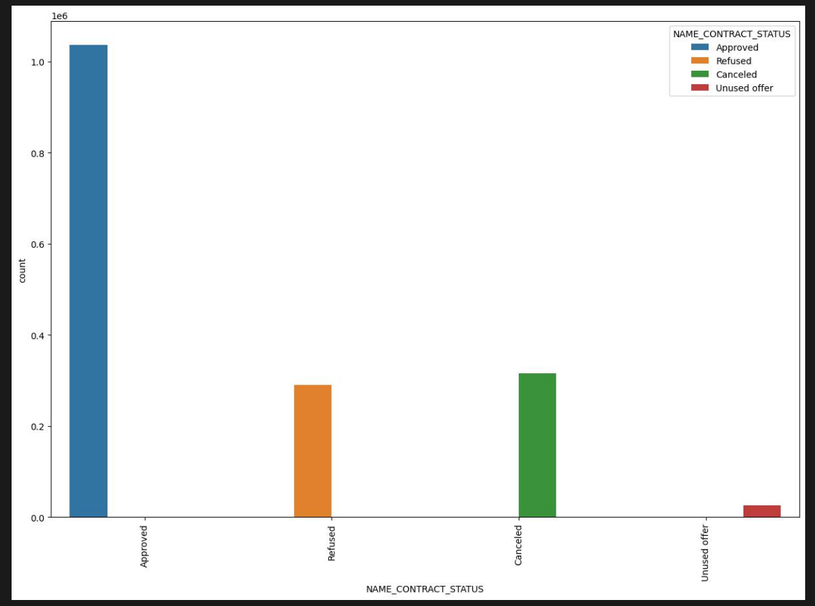
There are only obvious insights such as with increase in goods price, credit amount will increase or with more children, count of family members will increase.

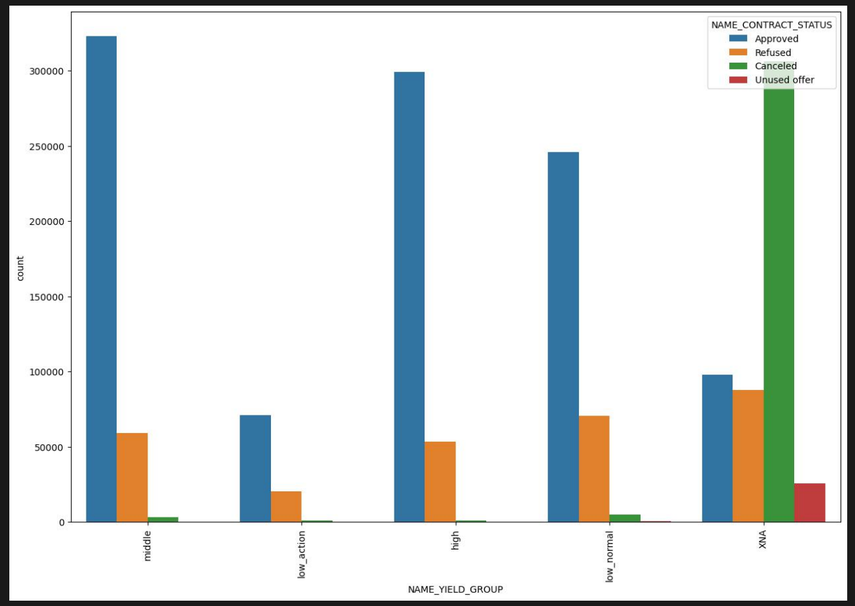
UNDERSTANDING DATA FOR PREVIOUS APPLICATION

For the ‘previous_application’ dataset, we follow the same approach of removing irrelevant columns and columns having more than 40% have been removed from dataset for efficient analysis. Also, null values for the rest have been kept as it is. Outliers have been observed as expected.

By analysing Credit amount with contract status, the obvious observation is bank tend to approve loans for repeaters. Also, a lot of applications have been cancelled. We have to see if those represent any opportunity lost.





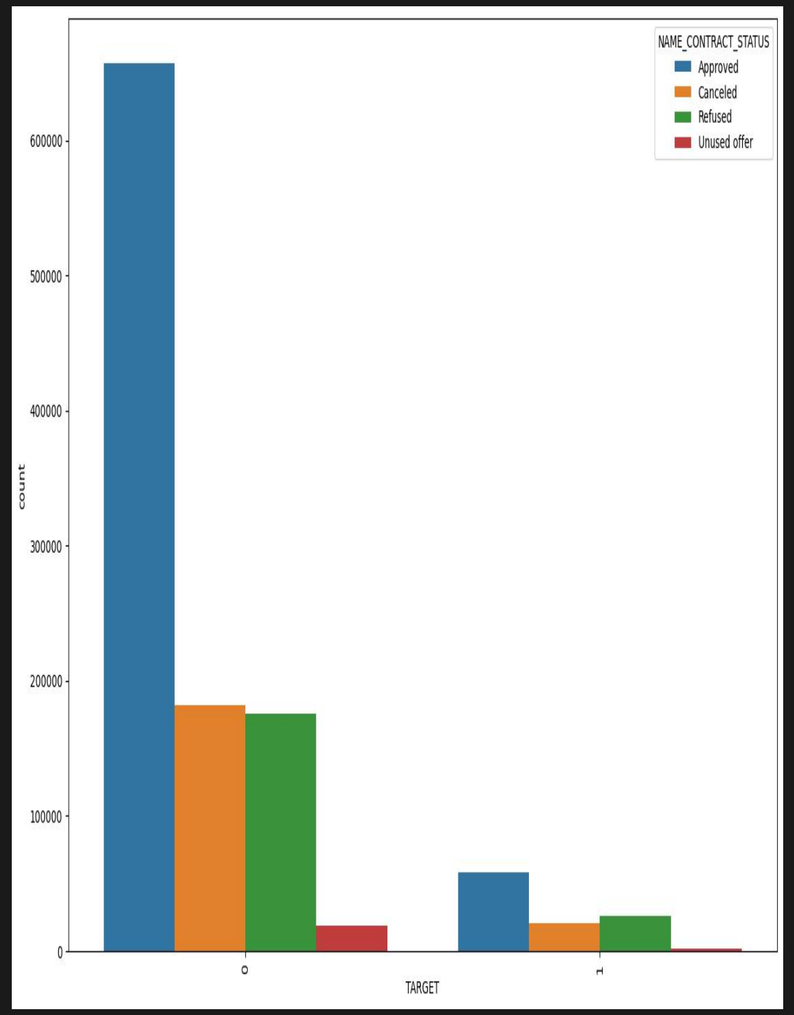


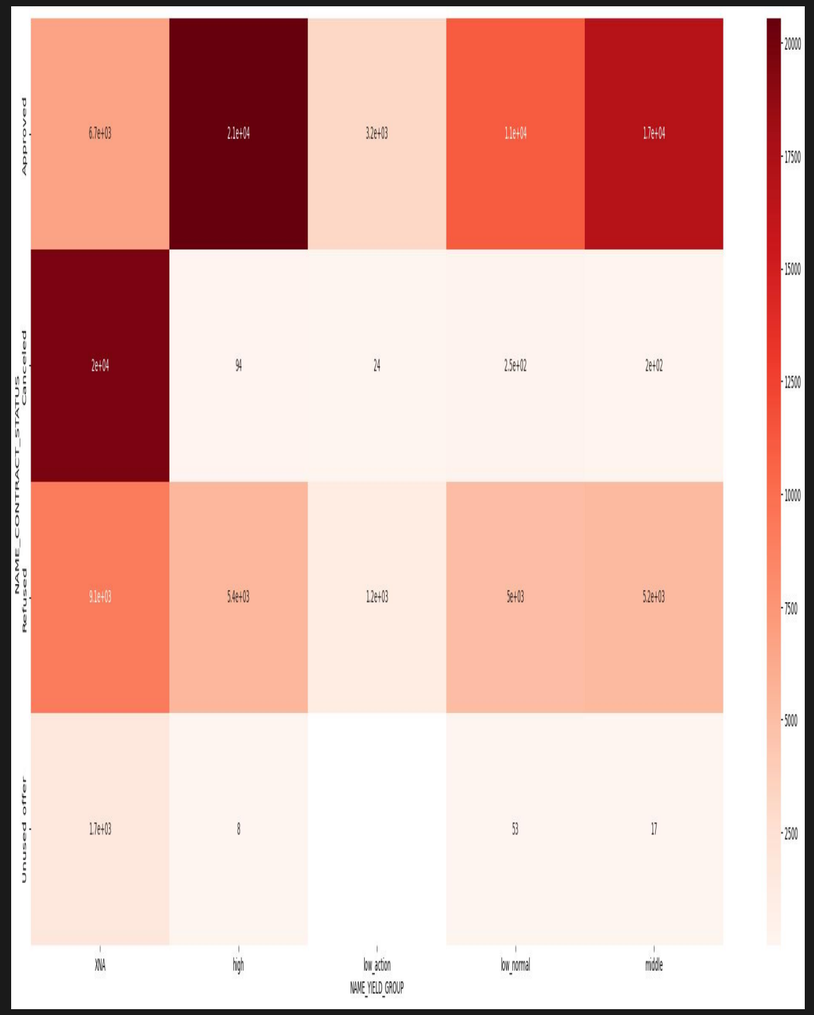
The following observations can be made :

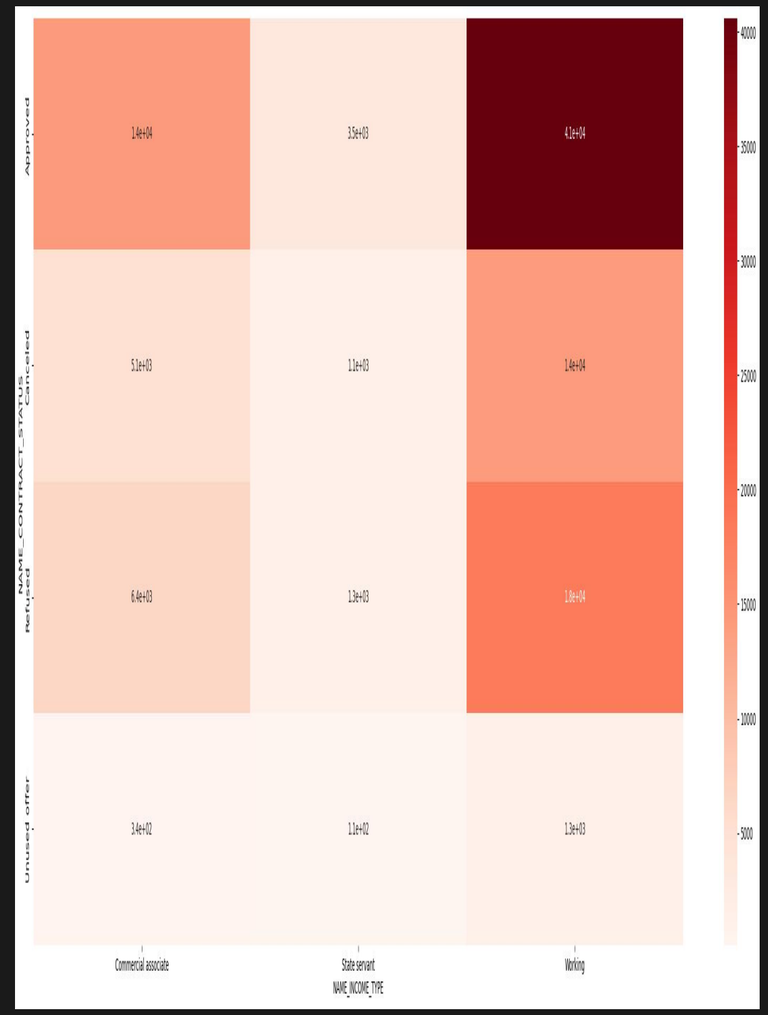
1. more rejection are for cash loans type and related to repairs.
2. Repeaters type of customers having most of the rejections.
3. Loans related to computers having more rejections.
4. Low_normal type of interest rate are having more rejections.
5. Loans for repairs are having lot of rejections.

ANALYSING COMBINED DATASET

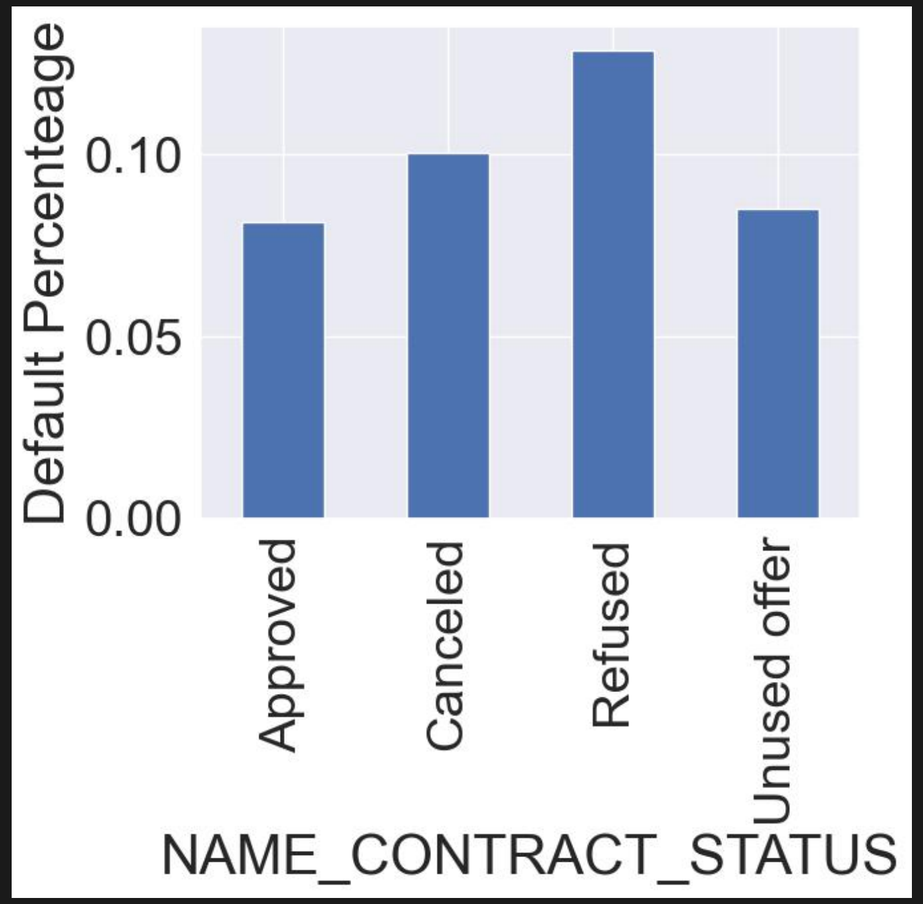
- Now, we have combined the two datasets based on current id. We have to investigate which approved categories are likely to be default.
- Repairs category are most likely to default after approval.
- It seems loans with higher interest rates are less likely to be repaid after getting approved.

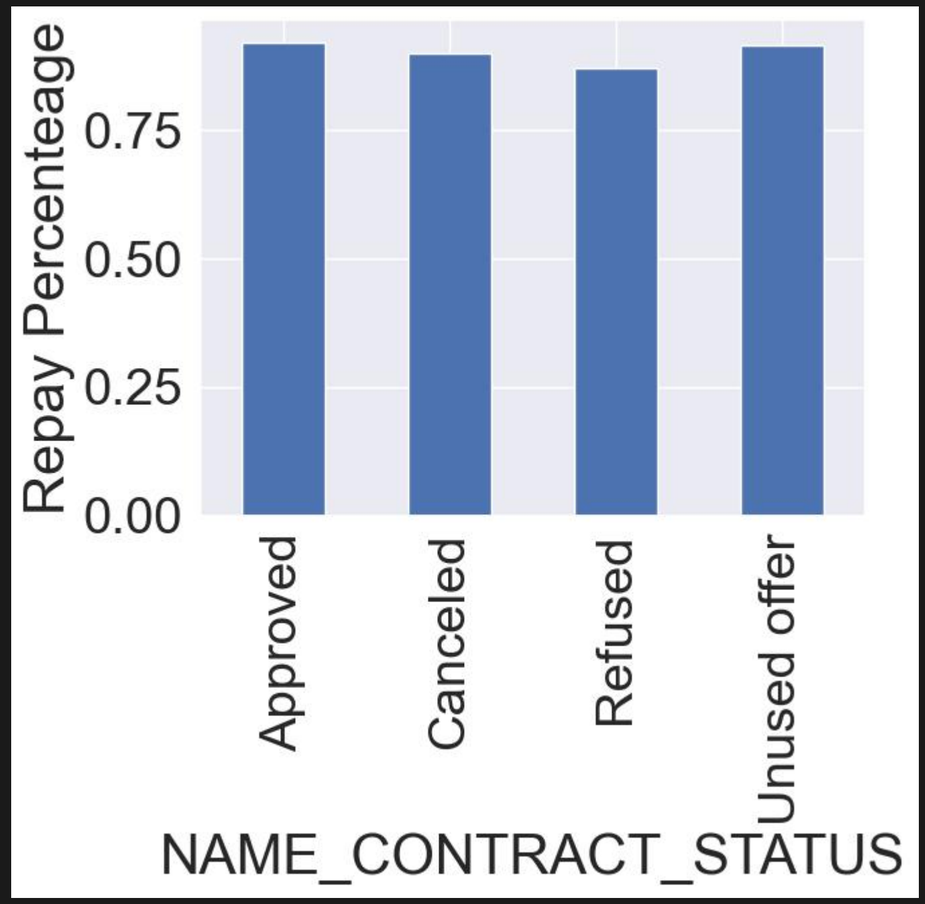






ANALYSING COMBINED DATASET





- It can be concluded, percentage of earlier refused applications tend to default more.
- So refusal mechanism of the bank is effective in that sense.
- It can also be noted, a significant percentage of earlier cancelled applications tend to repay as well,
- So bank should work on how to convince the consumers to accept the loan , by either reducing interest rate or providing other benefits.

OBSERVATIONS AND SUGGESTIONS

AMT_INCOME_TOTAL = People with higher income (> 500000) less likely to default.

DAYS_BIRTH : Older people (40+) are generally more likely to repay.

DAYS_EMPLOYED : people with more days of current employment tend to repay due to more stability.

CNT_CHILDREN : People with zero to two babies are more likely to pay

NAME_EDUCATION_TYPE : Academic degree holders are safest bet in terms of repaying the loan. People with lower secondary education are more likely to default.

NAME_INCOME_TYPE : Student, businessmen and pensioner have higher chance of repaying while people under working category has highest chance of default.

NAME_HOUSING TYPE : People living in office apartments tend to repay while people in rented apartments tend to default.

OCCUPATION_TYPE: Low skill laborers are more likely to default, while accountants are most reliable at repaying.

CNT_FAM_MEMBERS : With increasing family members (beyond 5) chances of default increases significantly.

ORGANISATION_TYPE : Transport: type 3 segment has higher default rate.

NAME_FAMILY STATUS : People having civil marriage and single tend to default more.

OBSERVATIONS AND SUGGESTIONS

It can also be noted, percentage of earlier refused applications tend to default more.
So refusal mechanism of the bank is effective in that sense.

But , a significant percentage of earlier cancelled applications tend to repay as well, so bank should work on how to convince the consumers to accept the loan , by either reducing interest rate or providing other benefits.

THANK YOU