Risk Analysis Of Loan Approval

BHARATH KEDARNATH KURRA

Problem Statement:

- *The problem is to determine whether a loan applicant will default or repay.*
- *The primary objective of approval process is not to reject or accept all applications, as this would lead to, rejecting capable applicants who can repay or approve for customer who cannot repay.*
- *Based on the given data, identify if the client has any difficulty paying the installments.*
- *If a loan is to be approved, do we approve a full requested amount or reduce the amount or offer a better interest rate.*
- *Figure out the driving factors for a loan default.*

Approach: How do we go about the problem

- *Looking at the basic dataset information, I had the question of why we have 122 columns, and I quickly opened the descriptor data to understand what kind of data we have.*
- *After skimming through the data, I realized there are quiet a lot of columns that doesn't add value. For ex: Documents related columns, apartment related information, escalator info, all these have no intrinsic value in determining defaulters. If a customer doesn't provide required documents, the application will not get to consideration stage and escalators, no.of entrances practically doesn't define repayment capability of a customer.*
- *Data like, income, occupation, family details, property details, tax information, education, investment details, past credit activity. All these data may help us better understand the applicant's motive.*
- *Check for missing values, outliers, identify object columns, int columns, float columns, check for data imbalance, create plots in loop as suggested and draw insights.*
- *For application data I followed top-down approach instead of directly dropping columns to understand the difference and what I will be left with.*

application_data: After careful consideration, I picked 24 columns which are challenging . Next task is to Identify missing and invalid data. Most of the variables with missing values doesn't pose great threat to our analysis. Those records are happy to be dropped from dataset. However, 'OCCUPATION_TYPE' has big chunk of data close to 32% missing. In our case, it would be ideal to assign a category that would make sense as deleting 32% records is not ideal and imputing the value based on max or mode value may create imbalanced data. So, assign new category "unknown" in this case.

previous_application: For previous data, lets keep all the columns and proceed with the analysis instead of removing columns upfront. Missing percent ranges from 22% to 99%, so I sequentially dropped columns to see how many rows I will be left with, Interestingly I was left with 100 records. I then dropped some more columns and left with 25 columns. Which is like application dataset. The difference between both datasets I finally created is, they have few differing columns.

Some of the missing values are hard to replace in this case because, we cannot decide what amount was disbursed, what was the first due date and data points like this may need a model to predict missing data.

Data Quality

Datasets before cleaning

	Application_data	Previous_application
Columns	122	37
Rows	307511	1670214

Datasets after cleaning

	Application_data	Previous_application
Columns	26	25
Rows	307217	648964

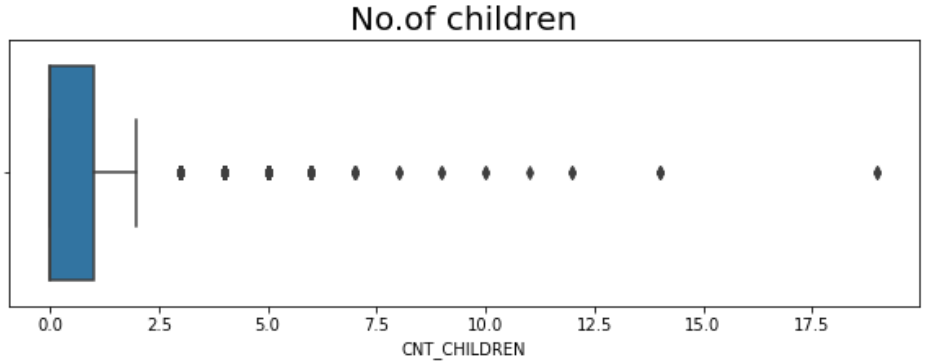
Missing percentage of application and previous datasets

Variable	Percentage
AMT_ANNUITY	0.004
AMT_GOODS_PRICE	0.090
OCCUPATION_TYPE	31.346
CNT_FAM_MEMBERS	0.001

AMT_ANNUITY	22.29
AMT_DOWN_PAYMENT	53.64
AMT_GOODS_PRICE	23.08
RATE_DOWN_PAYMENT	53.64
RATE_INTEREST_PRIMARY	99.64
RATE_INTEREST_PRIVILEGED	99.64
NAME_TYPE_SUITE	49.12
CNT_PAYMENT	22.29
DAYS_FIRST_DRAWING	40.30
DAYS_FIRST_DUE	40.30
DAYS_LAST_DUE_1ST_VERSION	40.30
DAYS_LAST_DUE	40.30
DAYS_TERMINATION	40.30
NFLAG_INSURED_ON_APPROVAL	40.30

NOTE >1

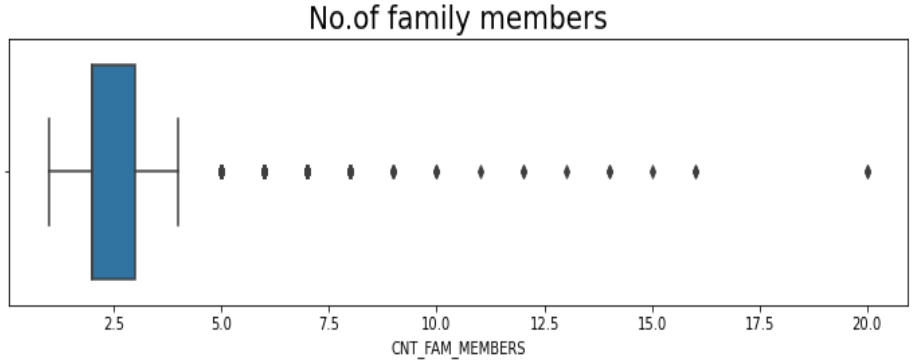
OUTLIER'S DETECTION



PERCENTILES

0.500	0.0
0.700	0.0
0.900	2.0
0.950	2.0
0.990	3.0
0.999	4.0

Although the child count seems to have outliers, they are continuous higher numbers and very few applicants have more children and majority of applicants doesn't have children.

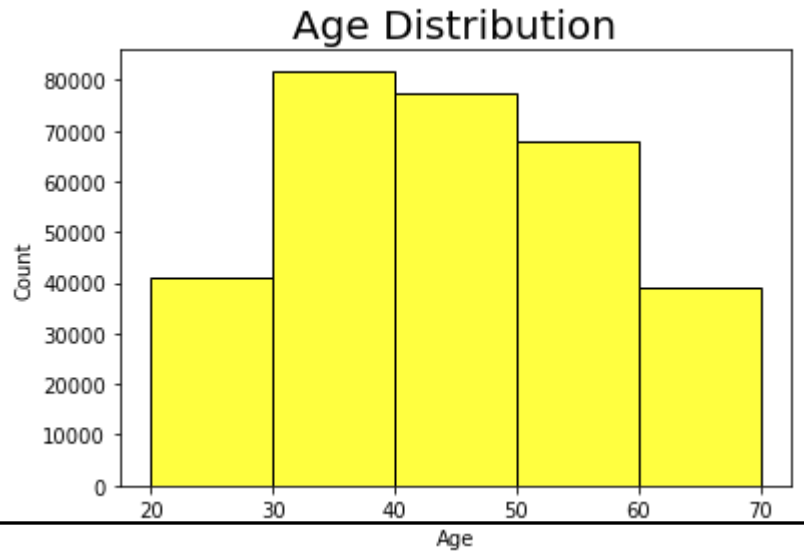


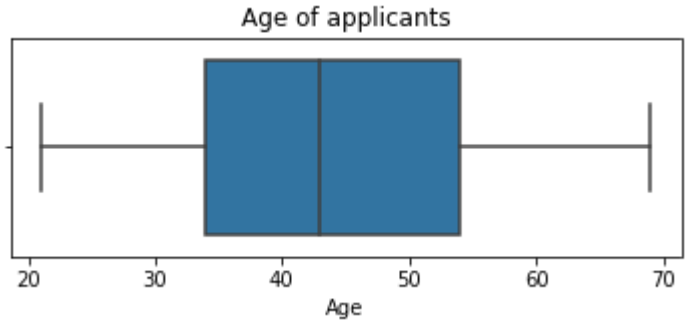
PERCENTILES

0.500	2.0
0.700	2.0
0.900	3.0
0.950	4.0
0.990	5.0
0.999	6.0

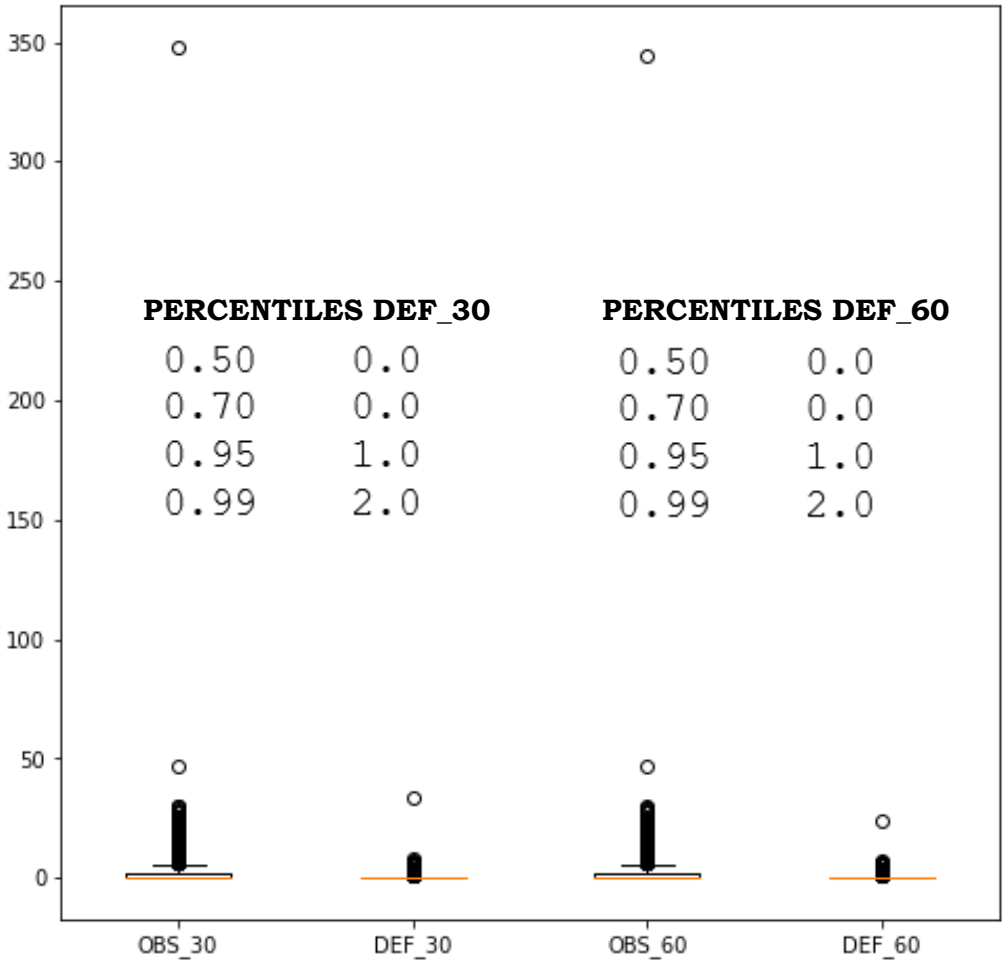
Similarly, not all the applicants have a big family, if we look at 95th percentile, majority of applicants families are =< 4. One can bucket records based on family count and perform further analysis to identify how their repayment behavior would be.

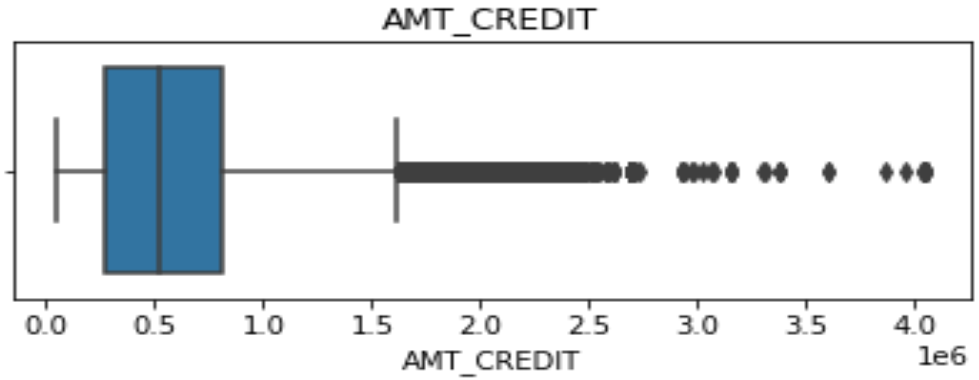
From the 'Age' plots we can see there are no outliers and data is well distributed.
We can also see a decrease in loan seekers above 40.
From age 30 applications drastically increases.
Majority of applicants seem to be working class between the age range 30-60.



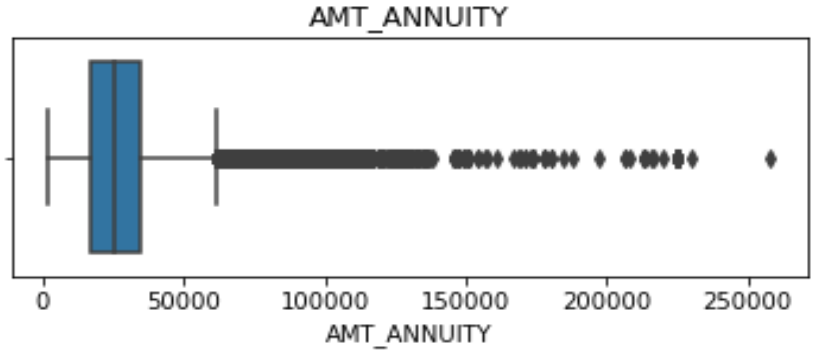


Social Surroundings With Defaulters

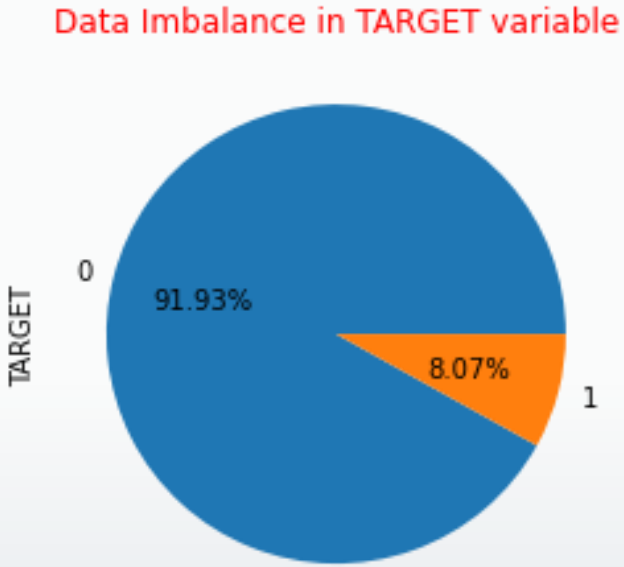




The Credit and Annuity columns have values higher than the upper fence in both cases. The values seem to be continuous and not many users fall under larger amount ratios. However, we will keep these records as they don't seem to harm our analysis.



Interestingly, the social surroundings defaulters seem to have some outliers, at the same time this is valuable information about defaults which might give us some insights. Keeping in mind these are outliers, our analysis must focus on percentiles instead of mean and medians. For instance, in DEF_30&60, the 99th percentile is still at 2 and outliers are way beyond the 99th percentile.



DATA IMBALANCE

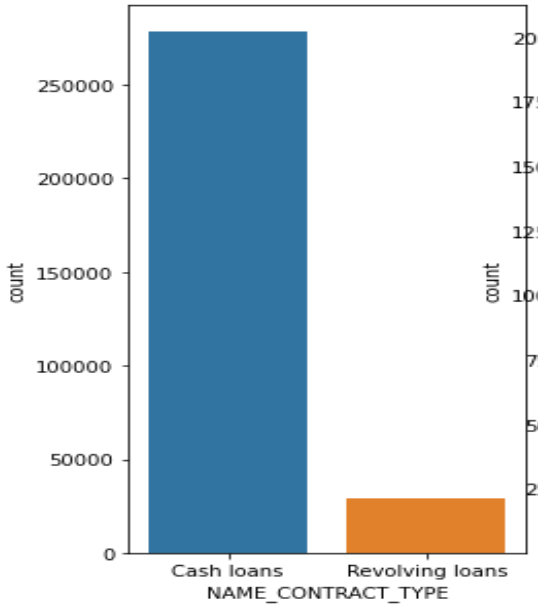
1. The target variable is highly imbalanced, this shows us the percentage of applicants with payment difficulties is comparatively less in these applications, but there are whopping percentage of people with difficulties.
2. 8 percent of defaulters may increase the NPS percentage significantly.
3. The ratio is 11, which means 1 out of 11 applicants has difficulty to repay.

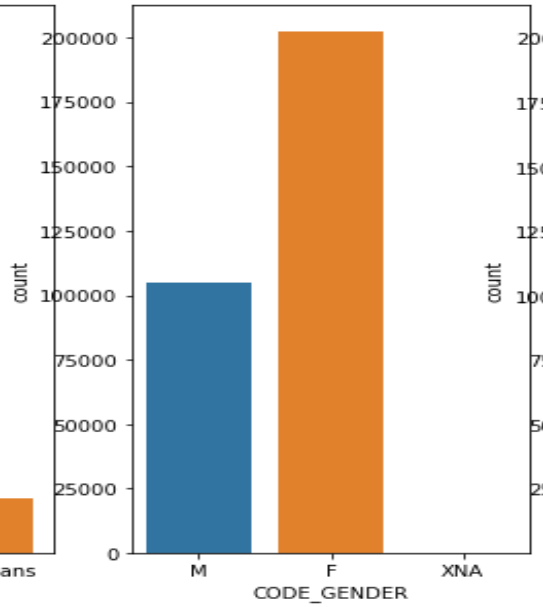
Dtype="object"

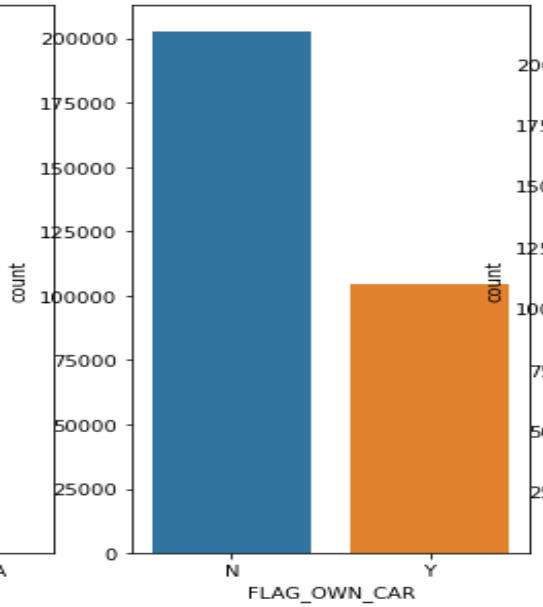
After plotting all our object type columns together, lets draw some insights from the plots.

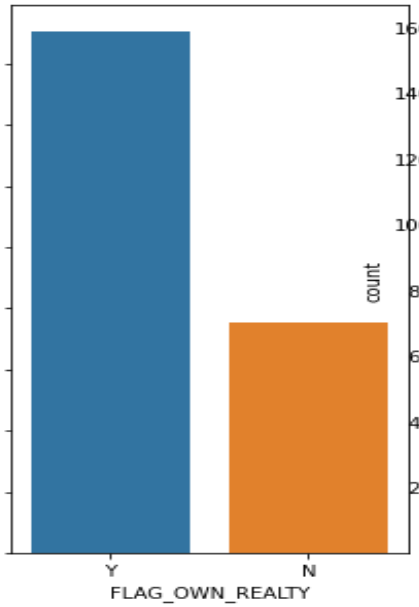
- From contract type we can see 90% of applicants are looking for cash loans.
- Interestingly women are highest loan seekers compared to men.
- People who doesn't own a car are in much need of loan, contrary to this people with own house or property has high applications.
- Looking at the plots it is obvious that married and working-class people need more loan than retired people.
- Although the plot looks good to put all the columns together, we should further analyze occupation and organization for better understanding.

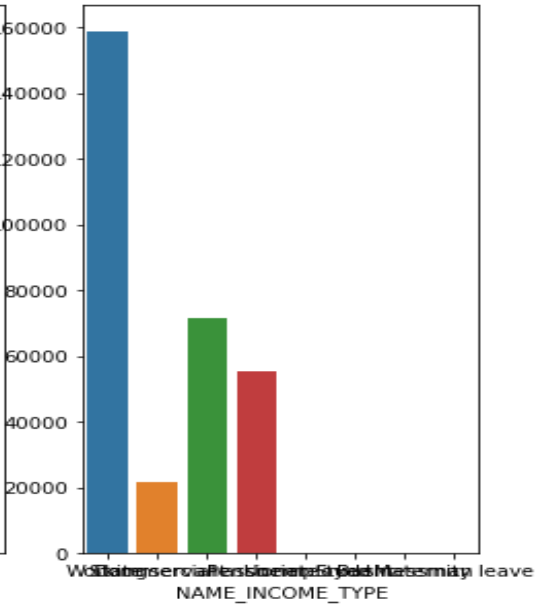
UNIVARIATE ANALYSIS

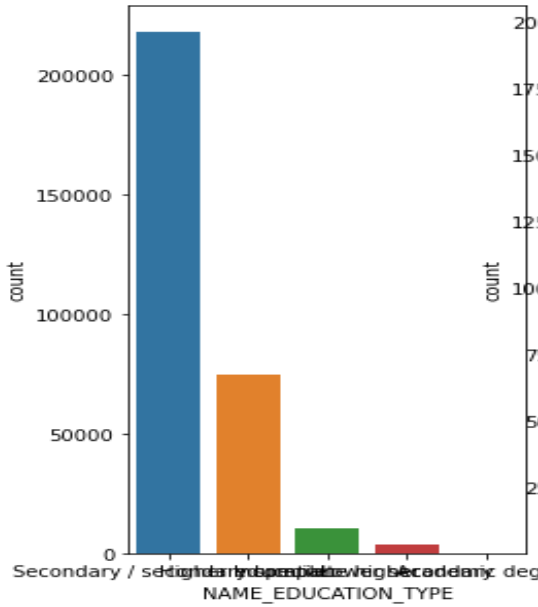


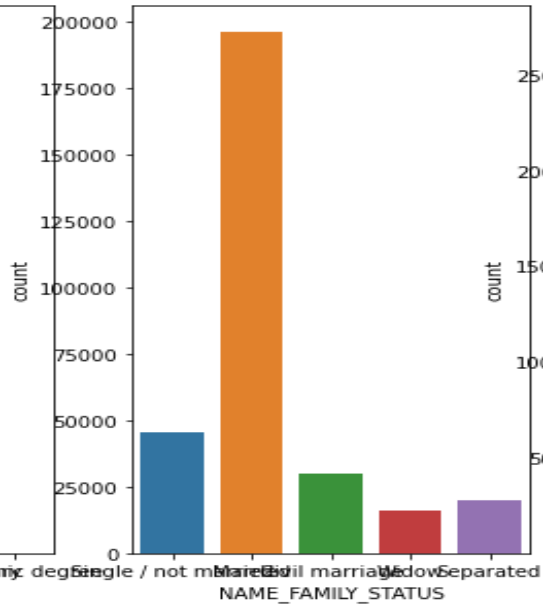


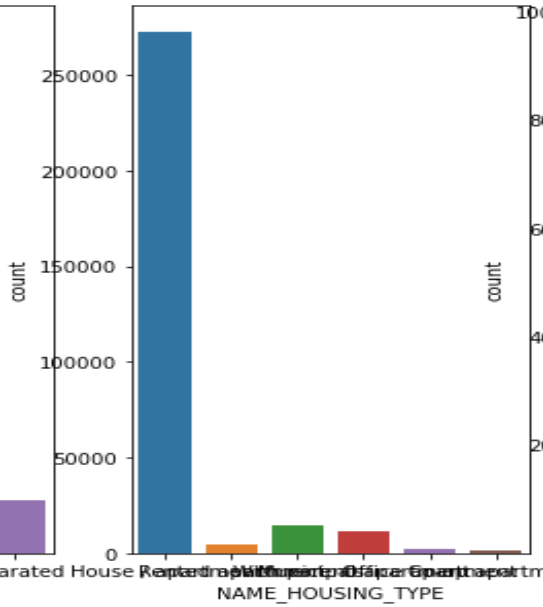


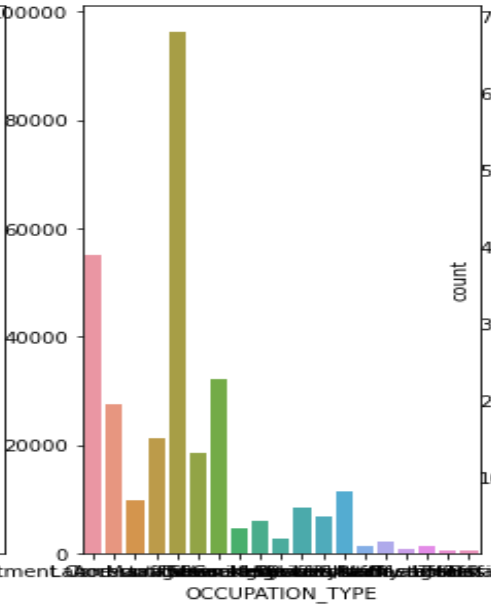


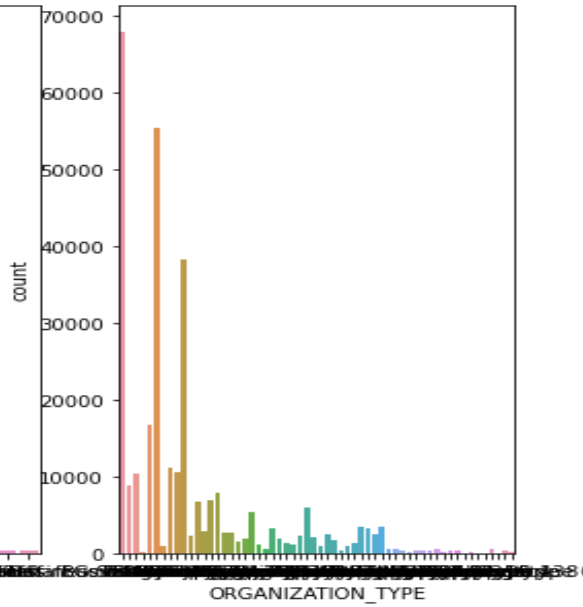


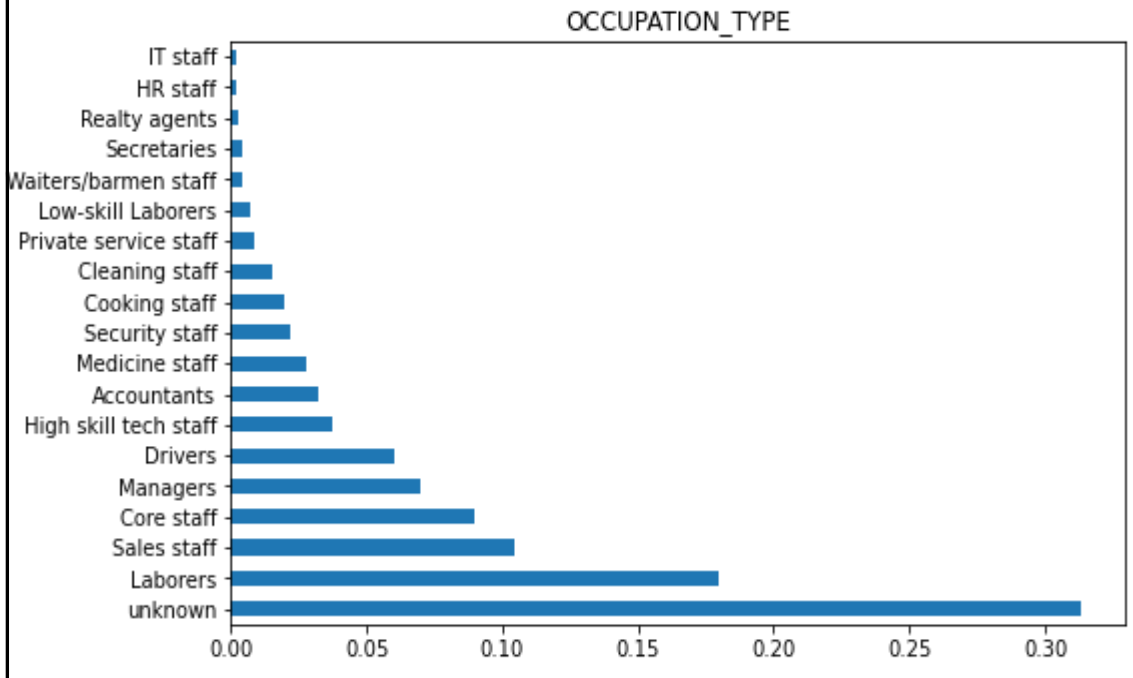


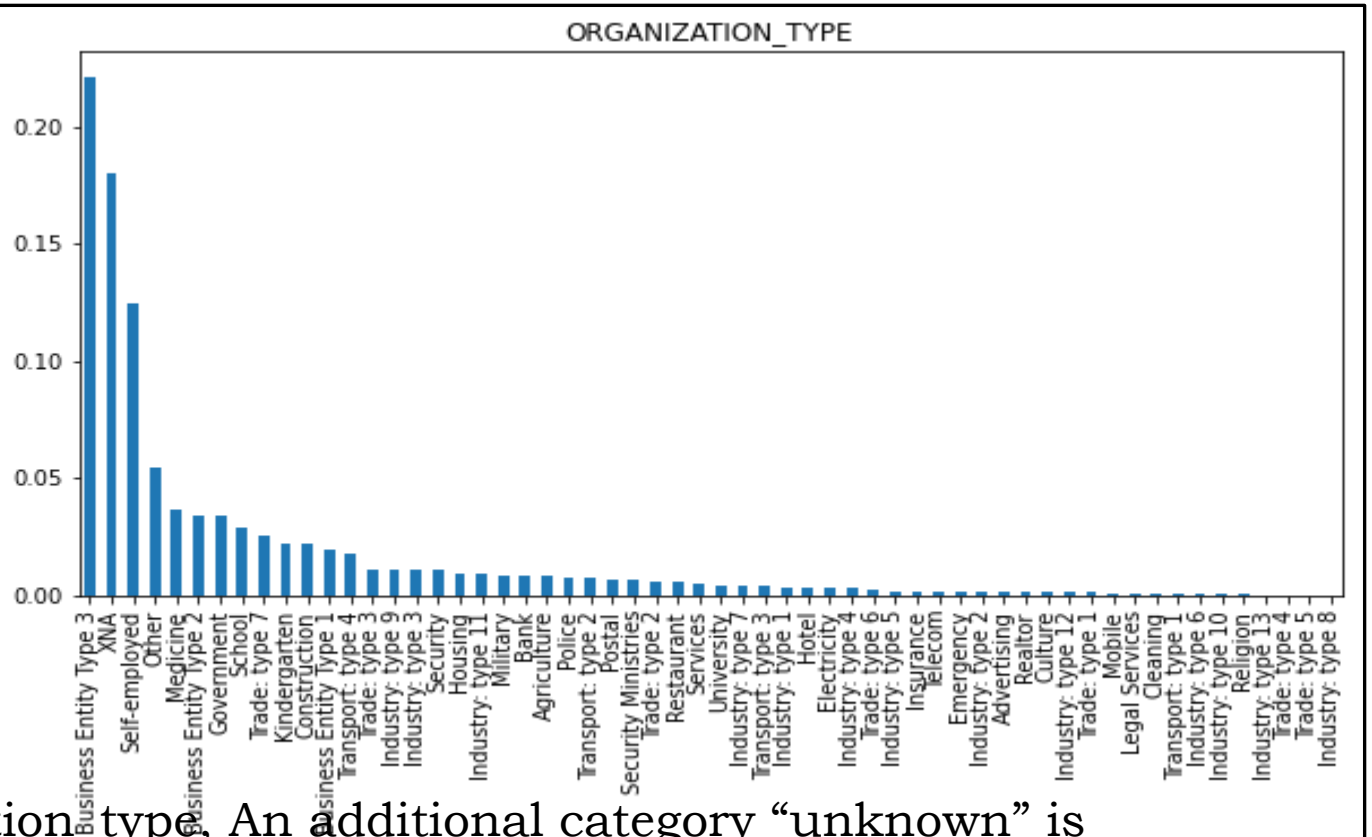






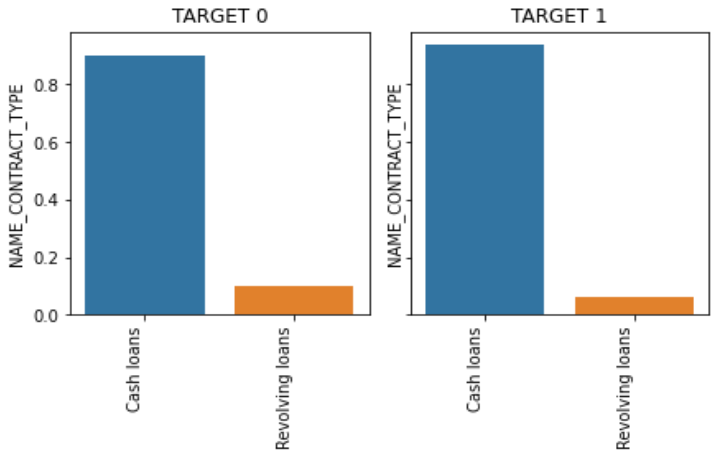


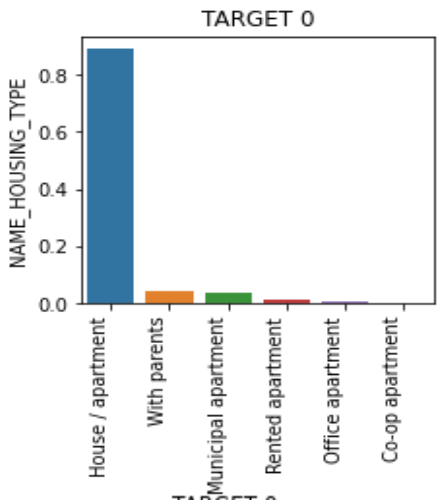


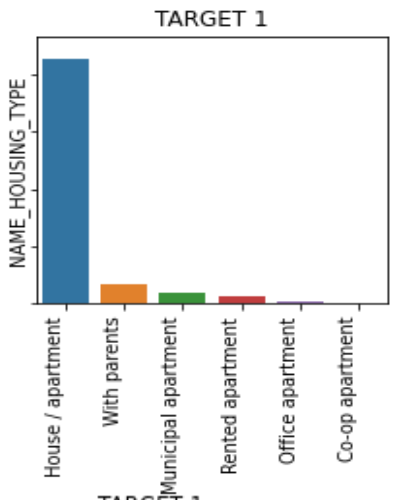


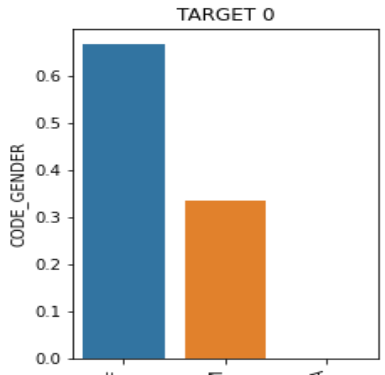
- In order to handle missing data in occupation_type, An additional category “unknown” is introduced. Applicants in this category seem to be high.
- If we look at the pattern, people in underpaid jobs seem to be >70%.
- From our first impressions we understood business and self-employed people are highest loan seekers.
- To summarize key data from the plots, cash loans, females, no car, own realty, working class, Secondary education, married people, underpaid employees are highest amongst the loan seekers, which is a natural phenomenon.

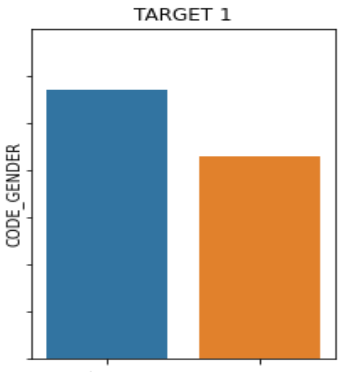
SEGMENTED UNIVARIATE ANALYSIS

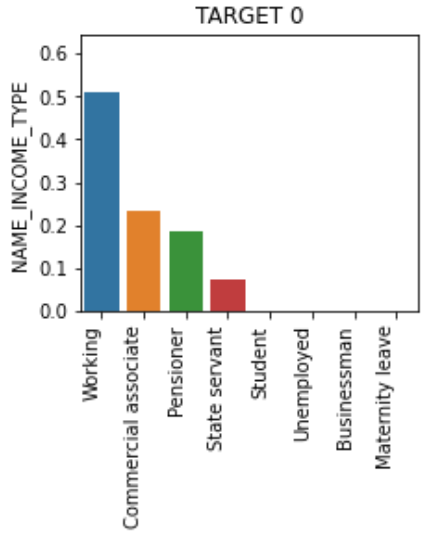


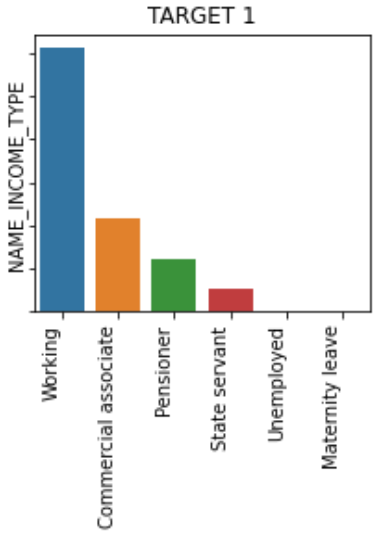


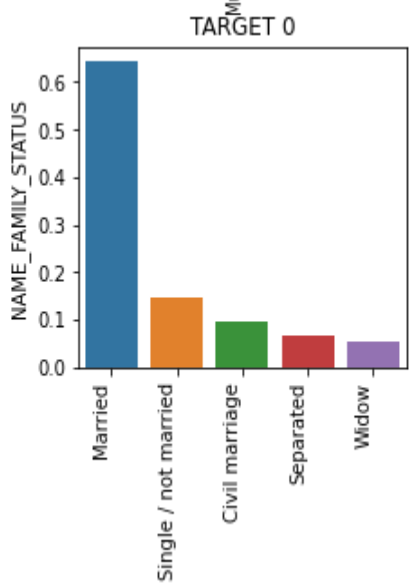


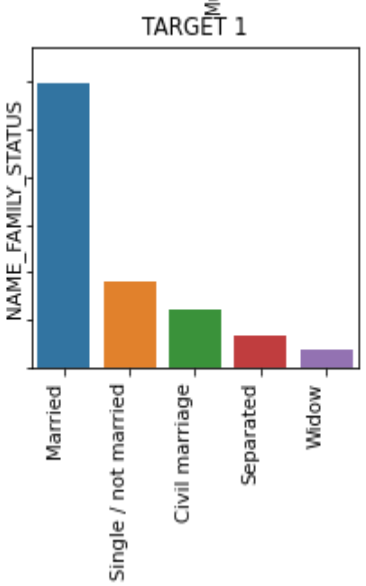


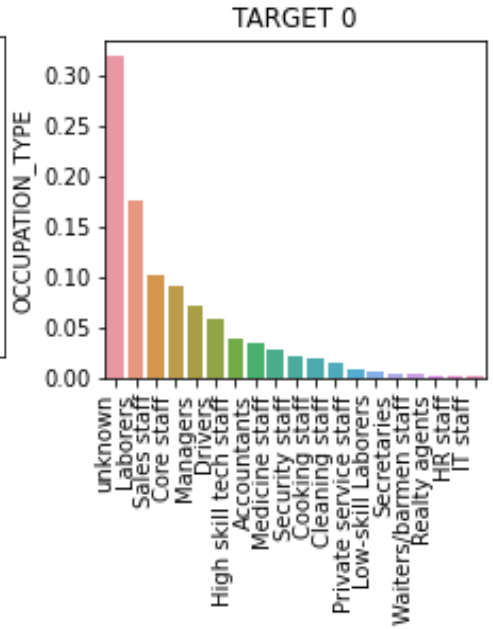


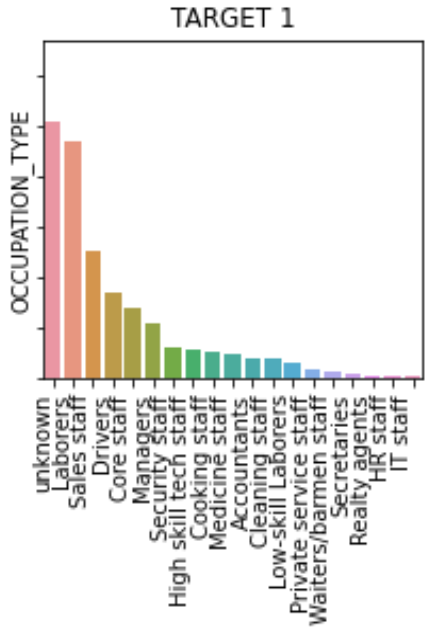






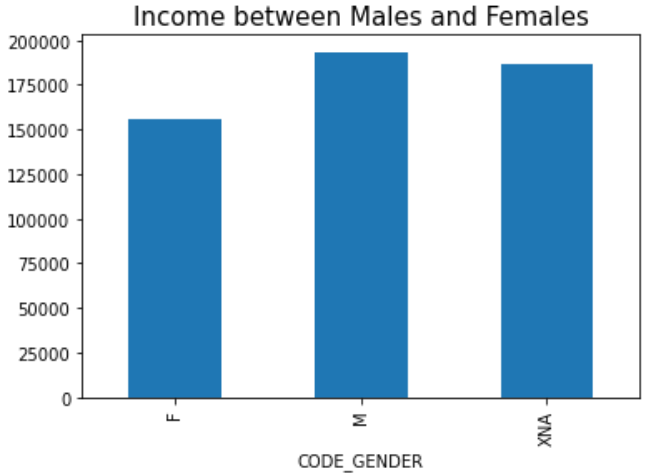


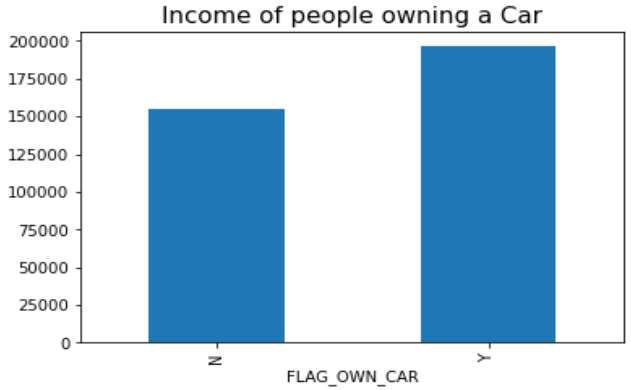


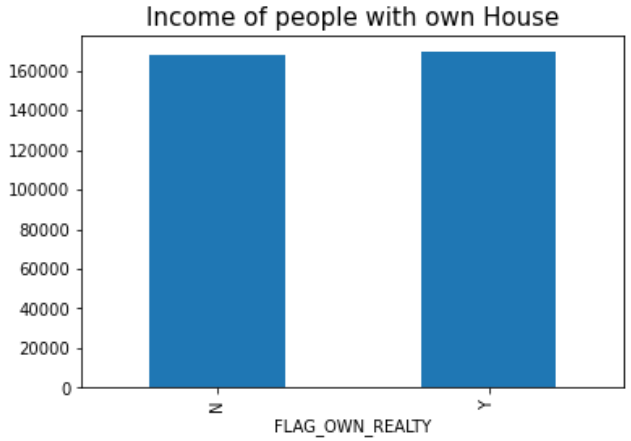


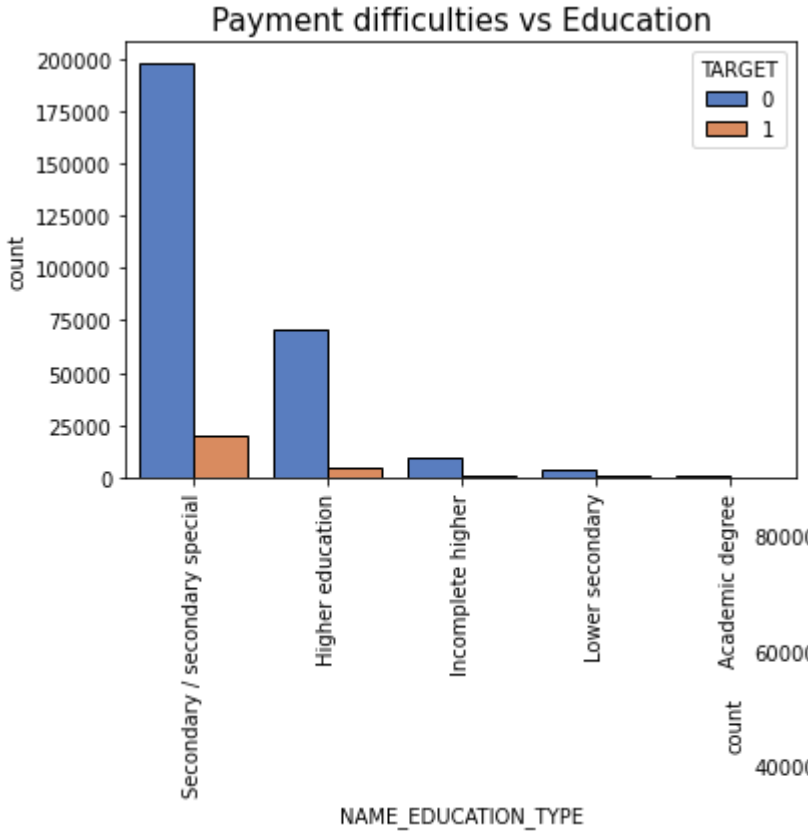
- 45% Males has repayment difficulty. 90% Need Cash loans. 60% Working class has repayment difficulty
- 60% Married people has payment difficulty. Sales staff and Laborers seem to have greater difficulty.

BIVARIATE ANALYSIS



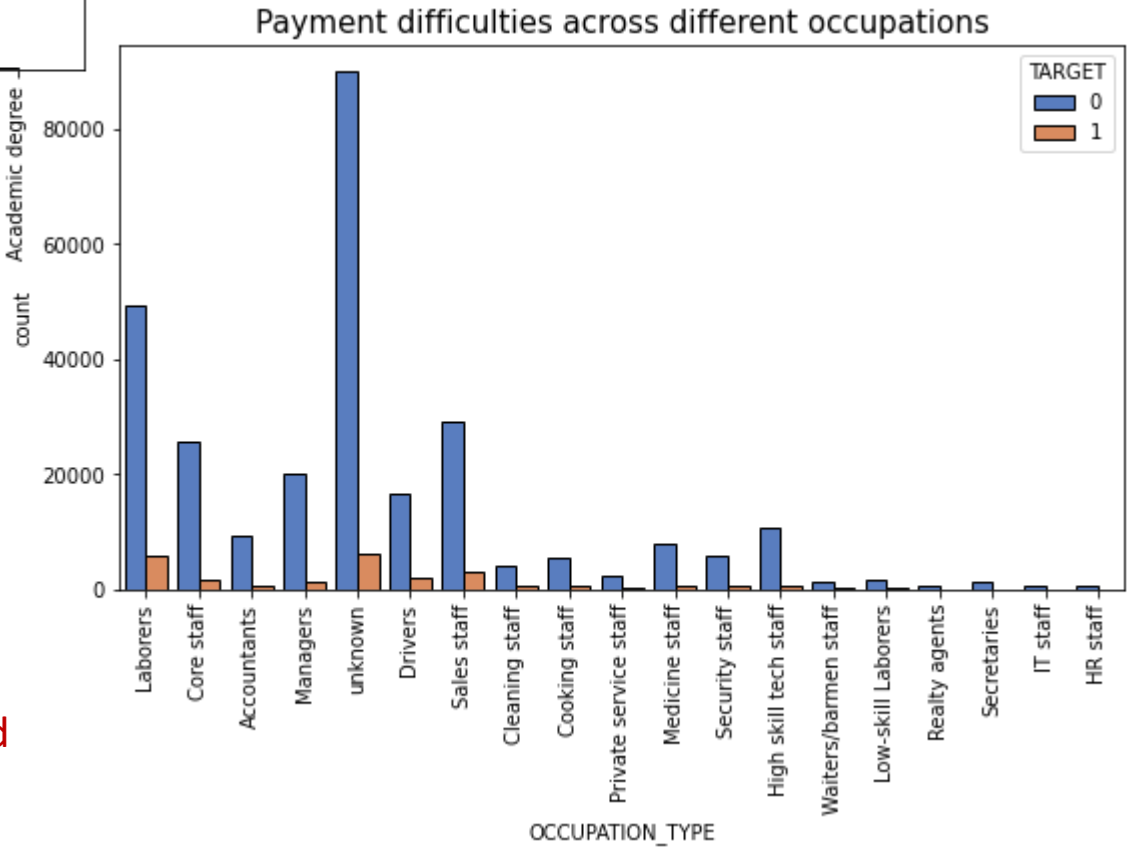






- Education has great impact on repayment.
- Males earn higher than females, however from previous analysis we found Males has higher repayment difficulties compared to Females.

- People with car has high salary
- Income of own reality and rented has same income levels



S.no	corr1	corr2	Corr_val
1	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998493
2	AMT_CREDIT	AMT_GOODS_PRICE	0.986968
3	CNT_FAM_MEMBERS	CNT_CHILDREN	0.879159
4	DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.860519
5	AMT_ANNUITY	AMT_GOODS_PRICE	0.775109
6	AMT_CREDIT	AMT_ANNUITY	0.76994
7	OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.332011
8	CNT_CHILDREN	Age	0.33093
9	DEF_30_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.32978
10	Age	CNT_FAM_MEMBERS	0.278911