



CREDIT EDA CASE STUDY

By:


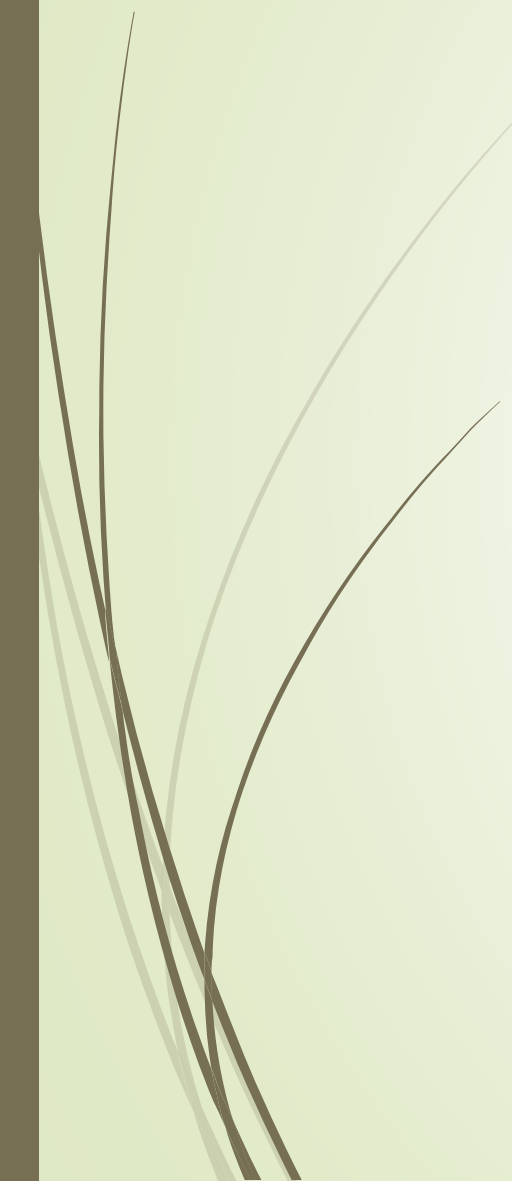
Anil Kumar Soren

Sagar Kumar Behera



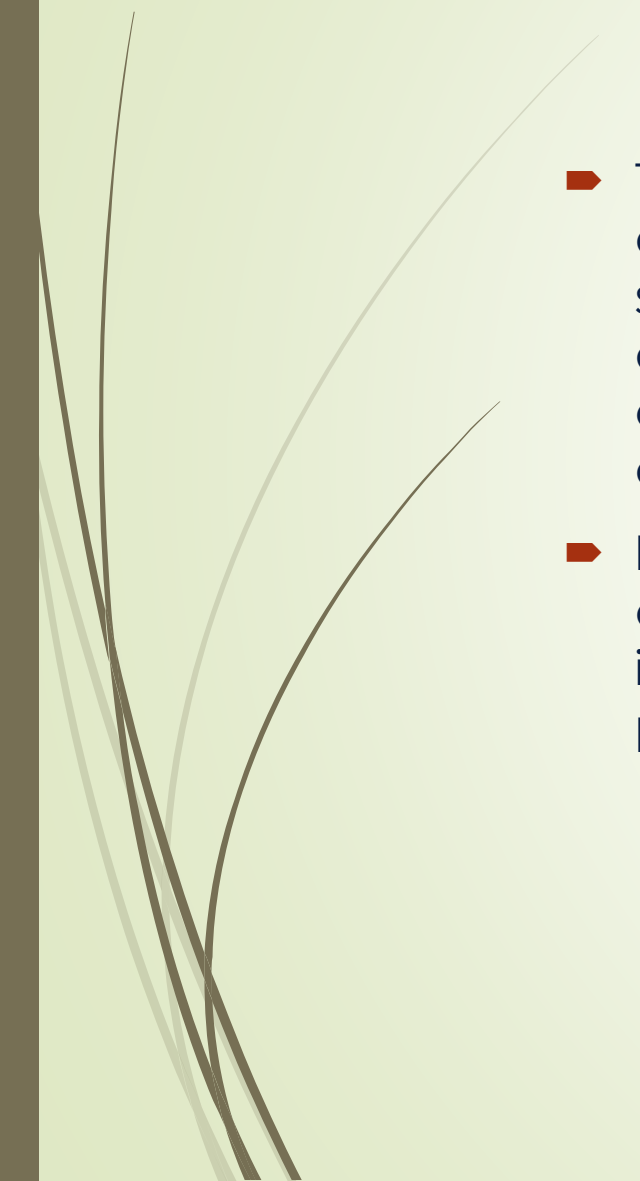
Business Understanding

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.
- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
 - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
 - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

- 
- 
- The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:
 - **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
 - **All other cases:** All other cases when the payment is paid on time.
 - When a client applies for a loan, there are four types of decisions that could be taken by the client/company):
 1. **Approved:** The Company has approved loan Application
 2. **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.
 3. **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).
 4. **Unused offer:** Loan has been cancelled by the client but on different stages of the process.



Business Objective

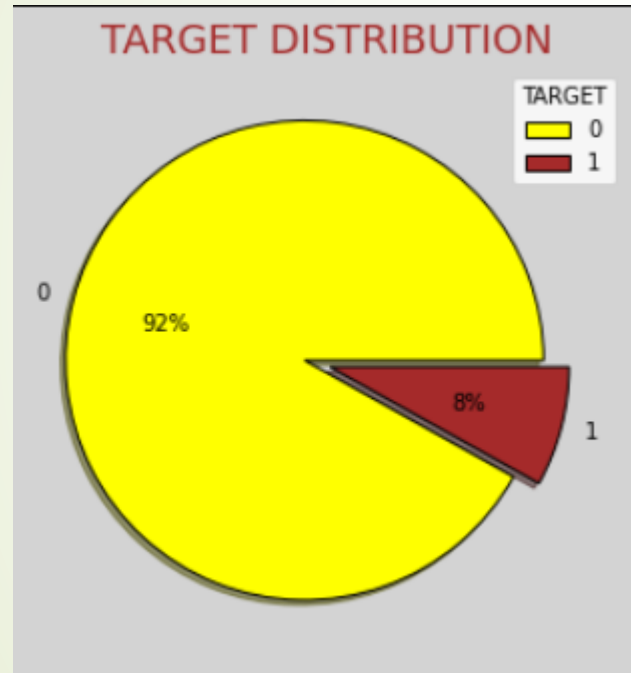
- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.
 - In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.
- 



Problem Statement

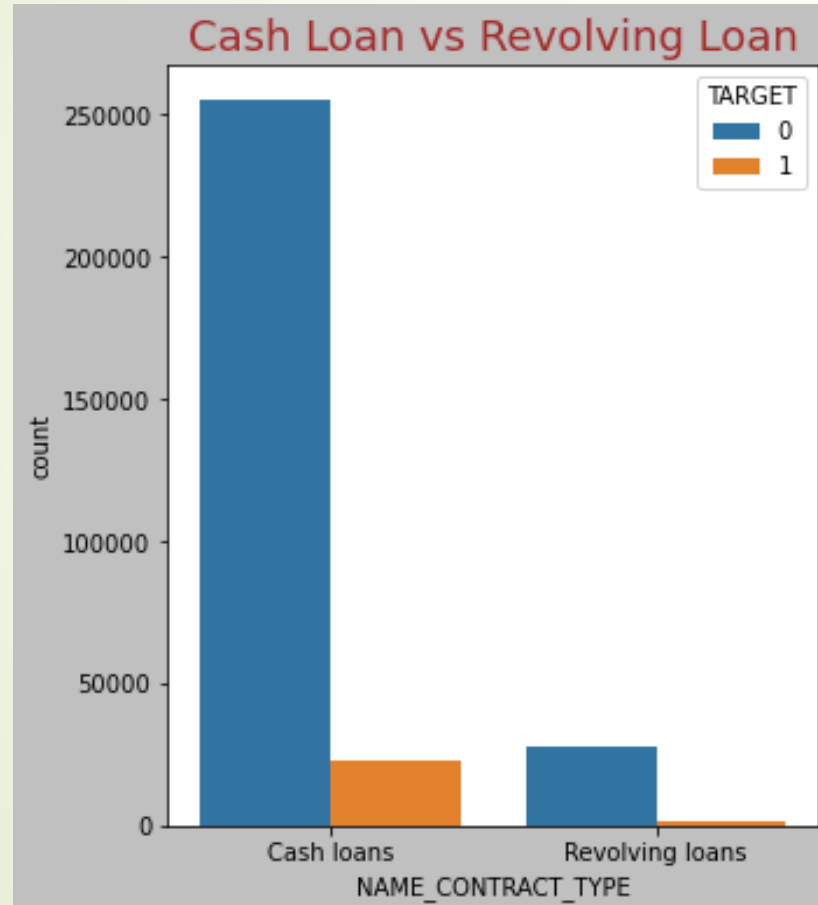
- Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)
- Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.
- Identify if there is data imbalance in the data. Find the ratio of data imbalance.
- Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.
- Find the top 10 correlation for the **Client with payment difficulties** and **all other cases** (Target variable).

Distribution of Target Variable



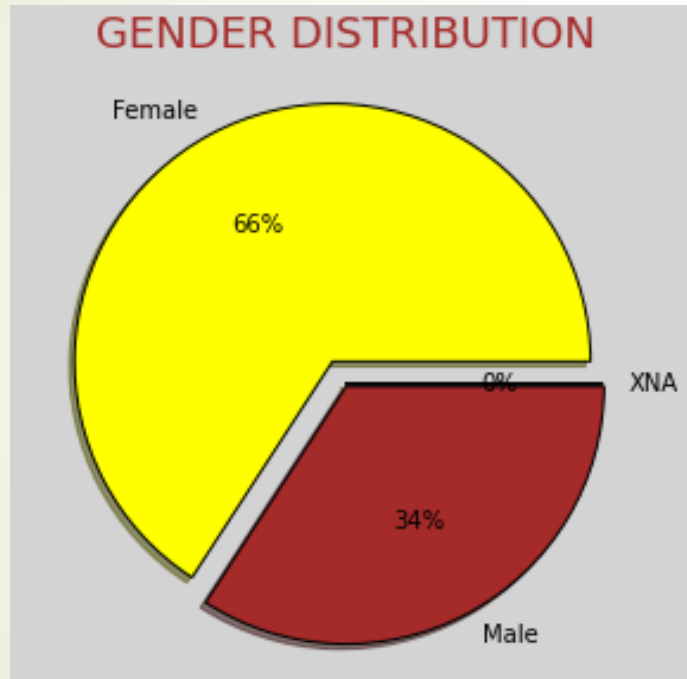
1. We can see its a highly imbalanced dataset of ratio of 2:23
2. 8% of clients are those with payment difficulties and may have the chances of defaulting loan. And the rest 92% are able to pay their loan on time

Distribution of Contract types



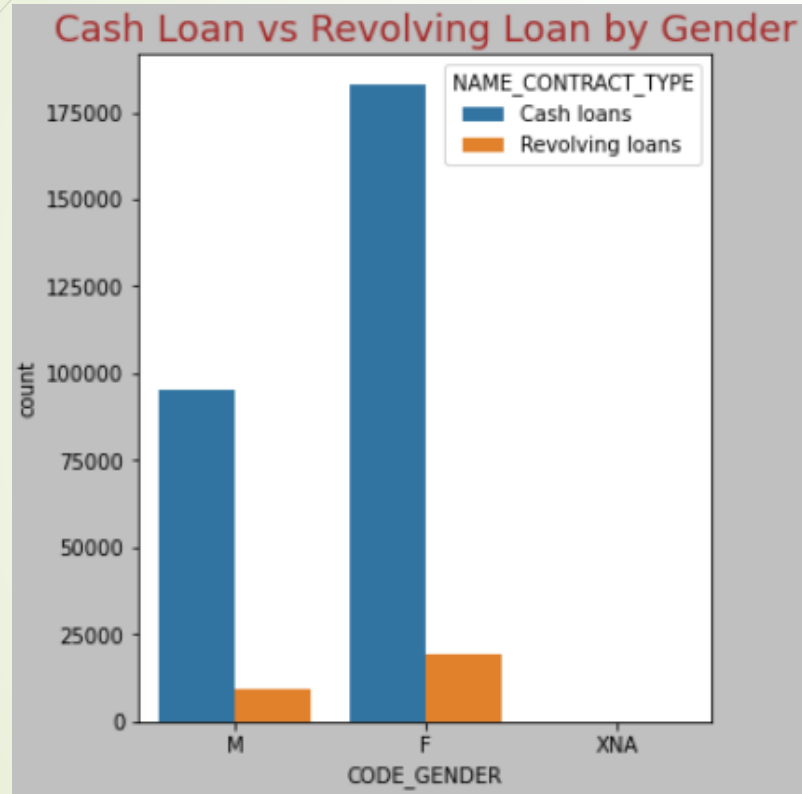
1. Number of the clients applied for cash loans (90%) is much higher than Revolving loans(10%)
2. Client with cash loans have the higher chances of defaulting loan.

Gender distribution



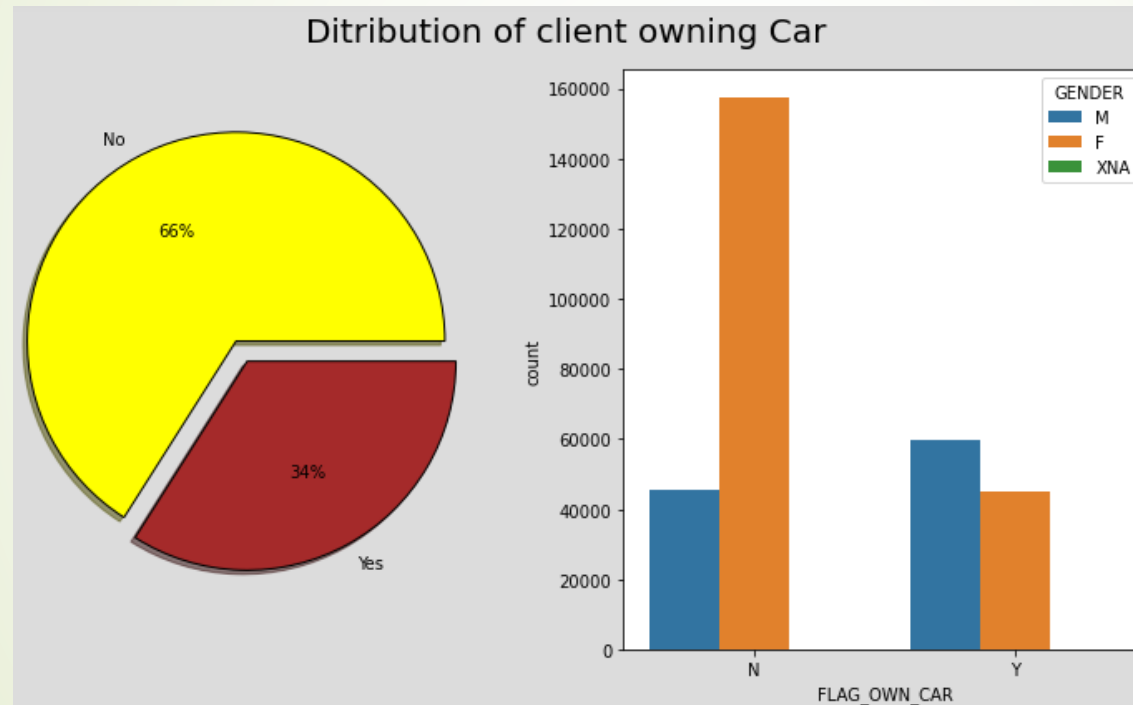
1. Number of female clients is 64% and number of male clients is 34%
2. Female clients tend to take more loans than male clients

Contract type vs Gender



Both gender prefer Cash loan over Revolving loan

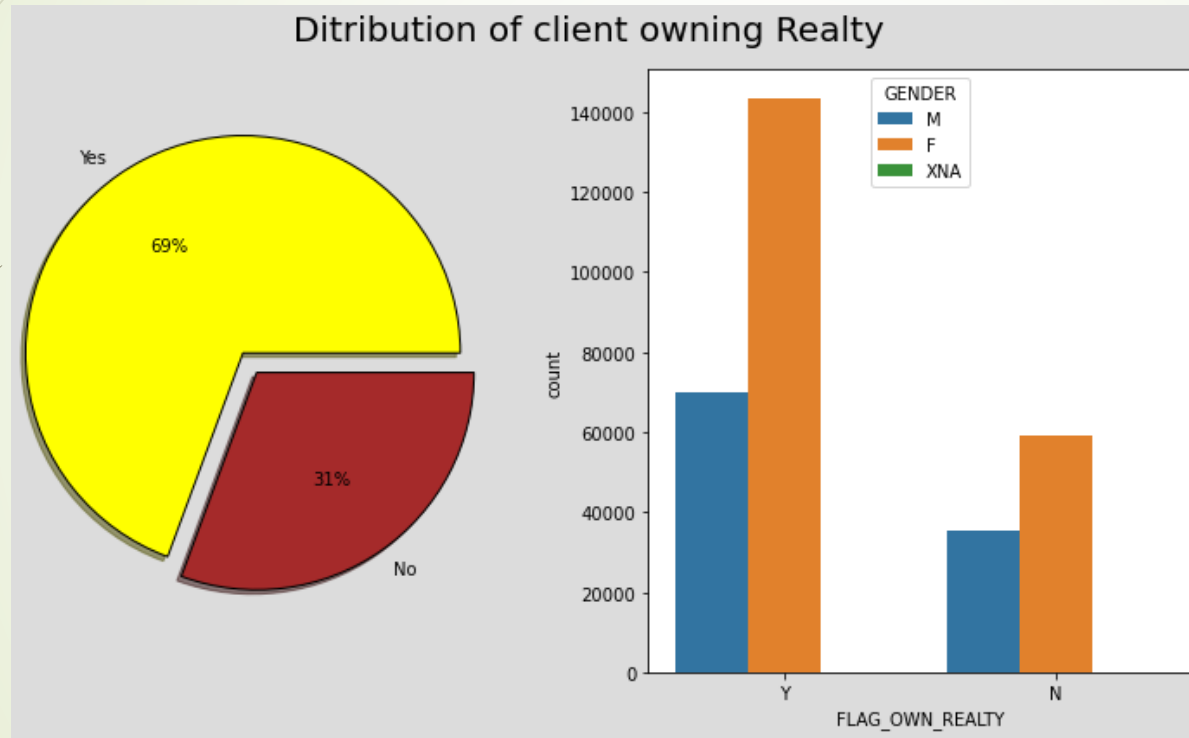
Distribution of client owning Car



1. 66% percent of clients do not own car

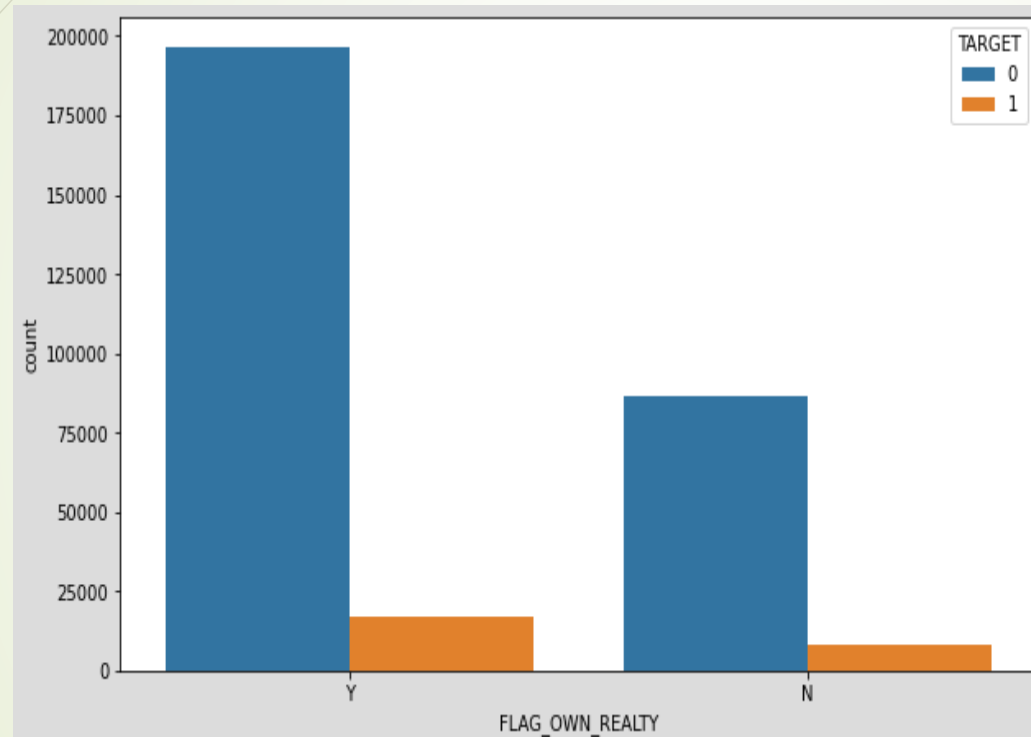
2. Male clients own more car than female clients

Ditribution of client owning Realty



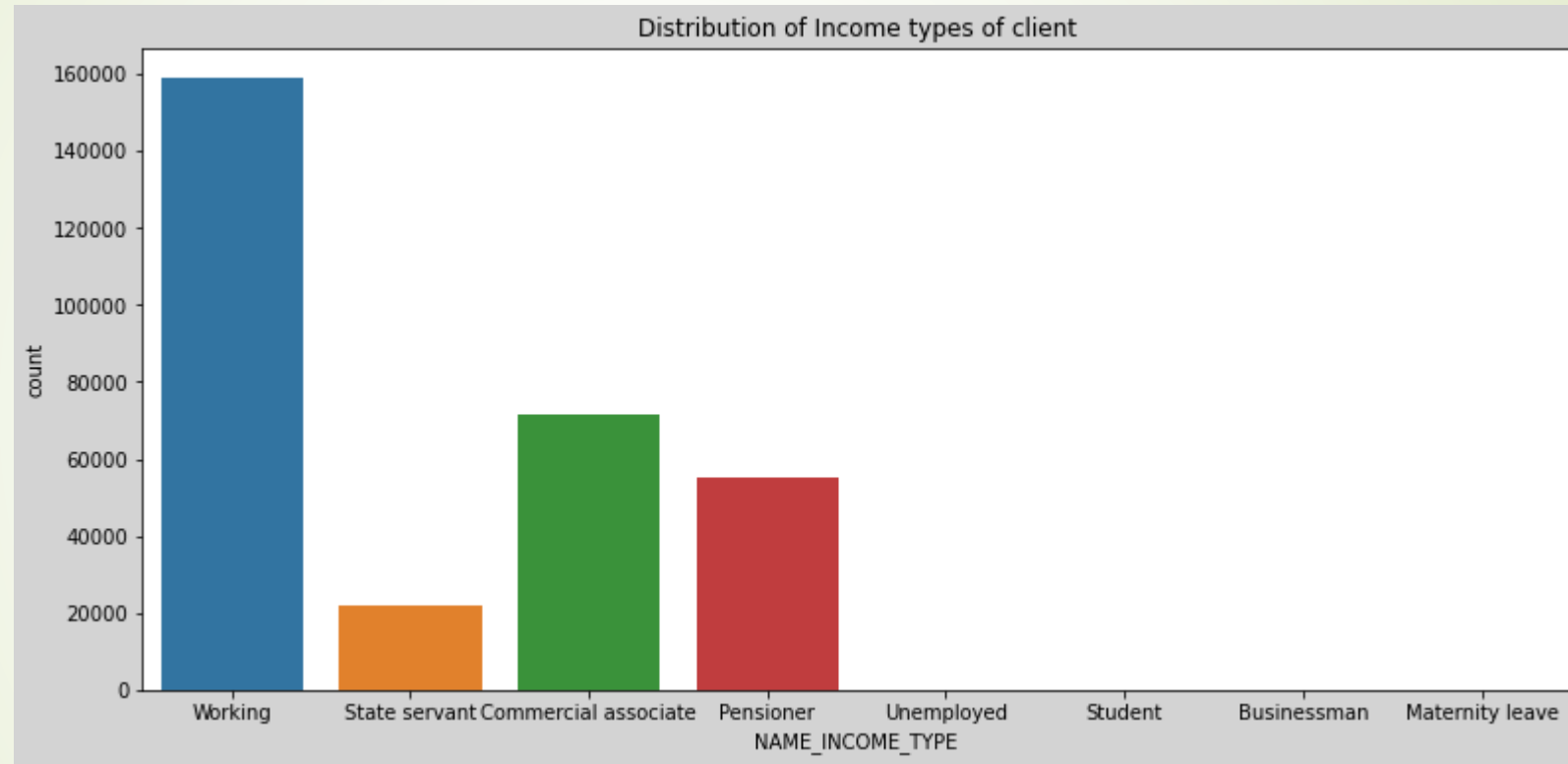
1. 69% percent of clients own realty property
2. Female clients own more realty than male clients

Realty vs Target variables



Clients owning realty properties tend to make their loan payments on time

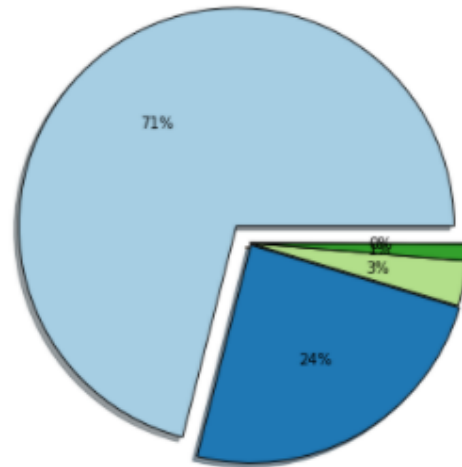
Distribution of Income types of client



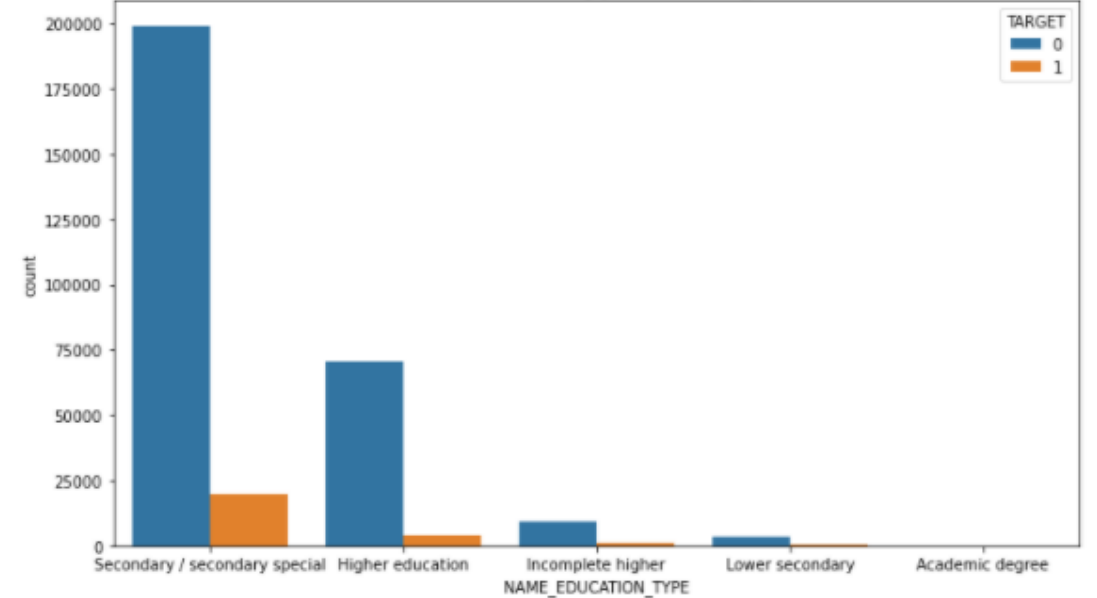
1. Working people are among the highest who are applying for loan.
2. Businessmen, Unemployed people, students, and ones with maternity leave are those whose numbers are very low.

Education types vs Target

Ditribution of client with different Education type

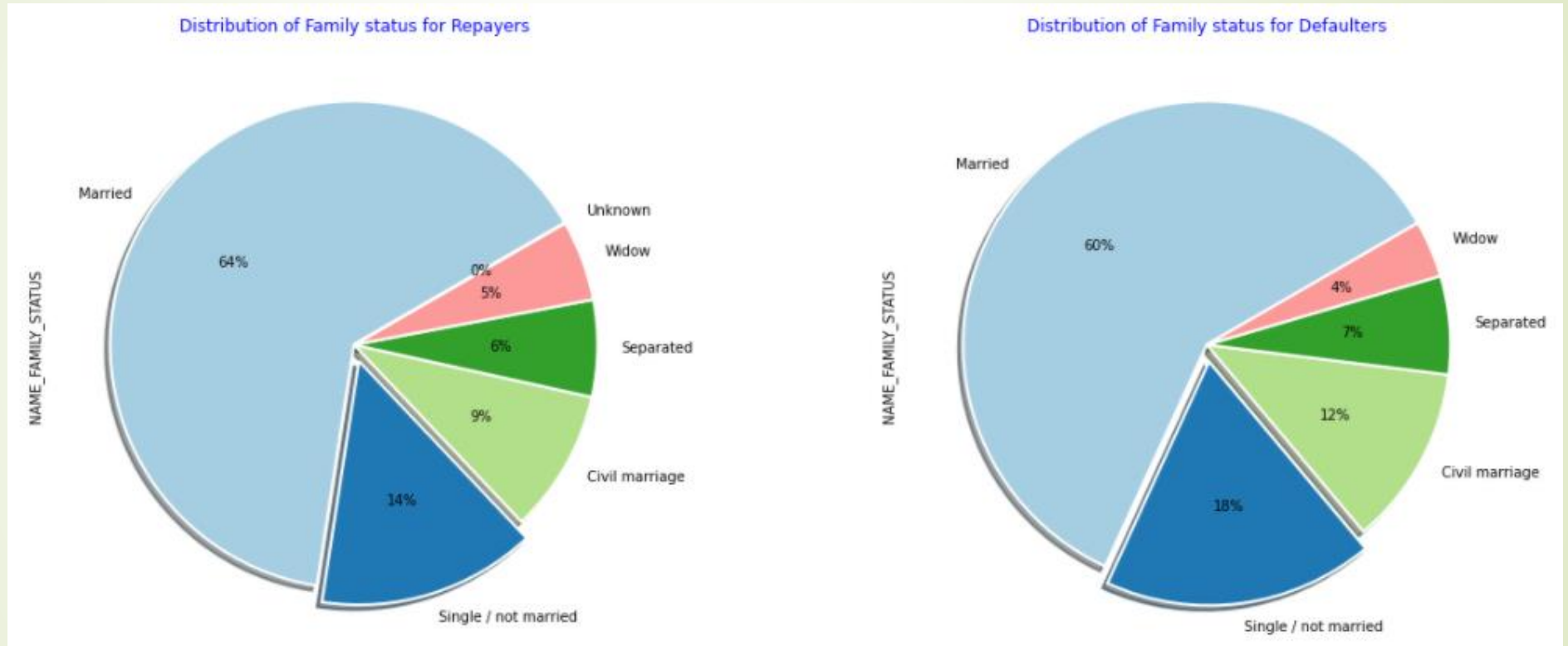


Education Type vs Target



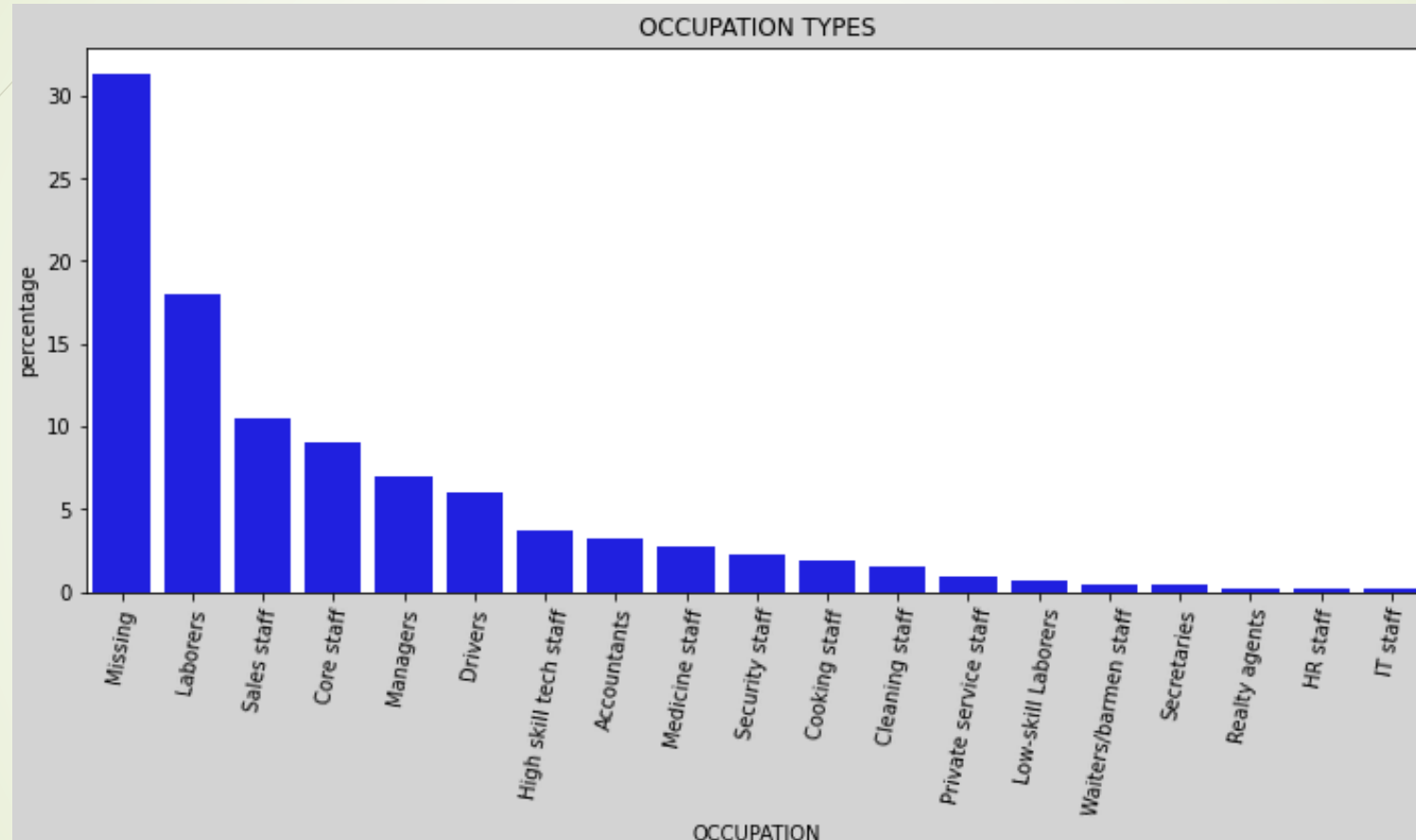
1. 71% of clients with secondary education are applying for loan and they are also those with higher chances of defaulting.
2. 24% of clients are those with higher education

Distribution of Family status vs Target variable



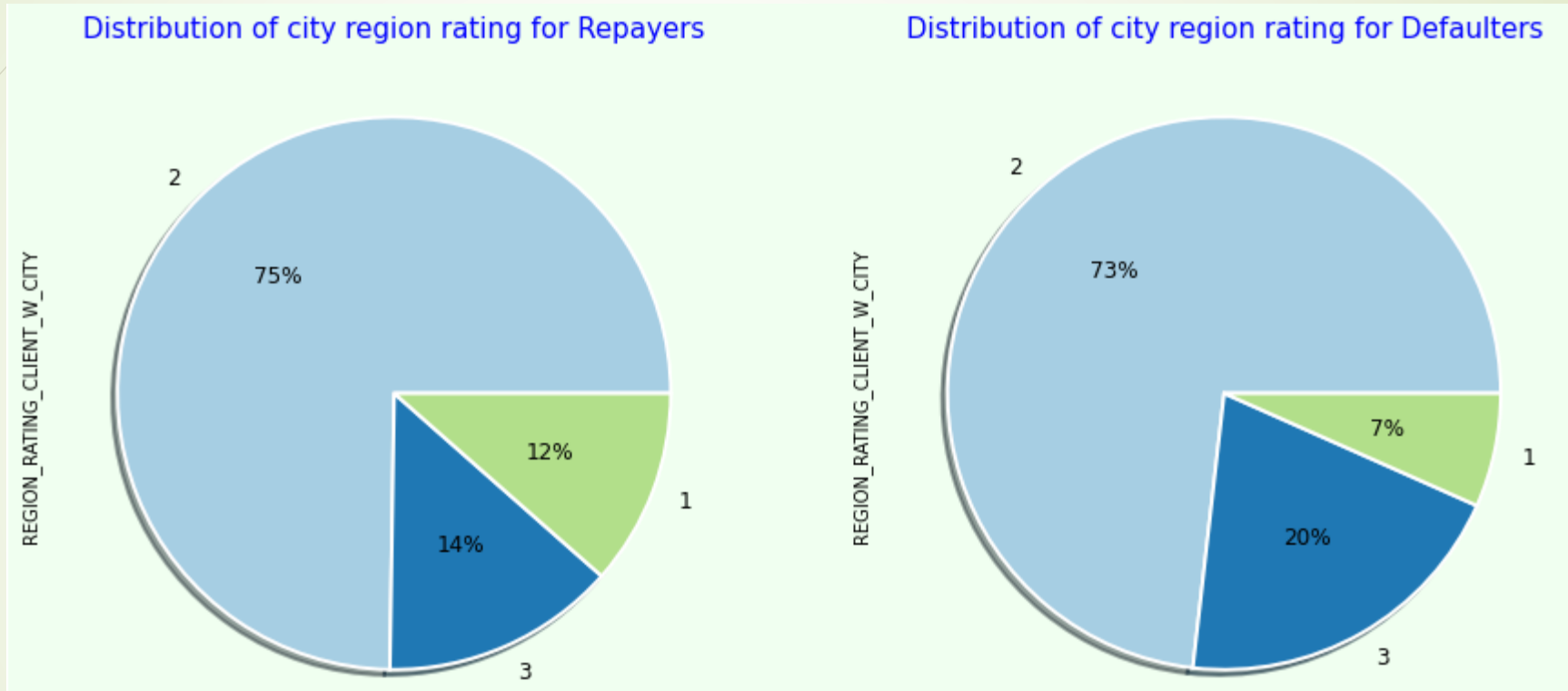
1. In both cases number of married clients are highest who are repayers as well as defaulters
2. Defaulters percentage is more in case of single, civil marriage and separated cases

Distribution of Occupation Types



1. Most of the applicant's occupation is missing
2. Top applicants are Laborers, Sales Staff, Core staff

Region Rating by city vs Target variable



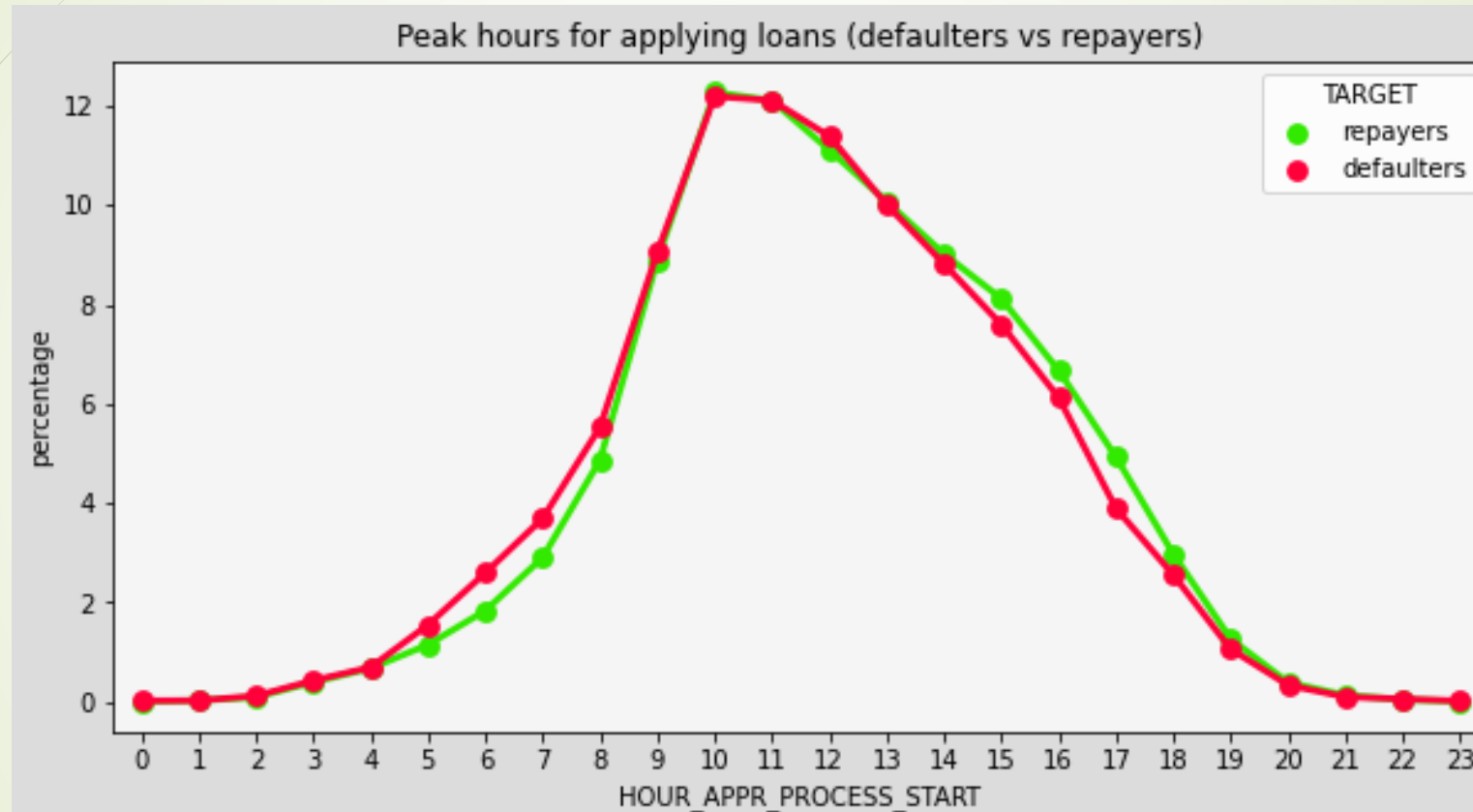
1. Most of the city region of repayers and defaulters have been rated 2 out of 3.
2. Percentage of defaulters are less in 1 rated city region compared to repayers.
3. Percentage of defaulters are higher in 3 rated city region compared to repayers.

Weekdays vs Target Variable



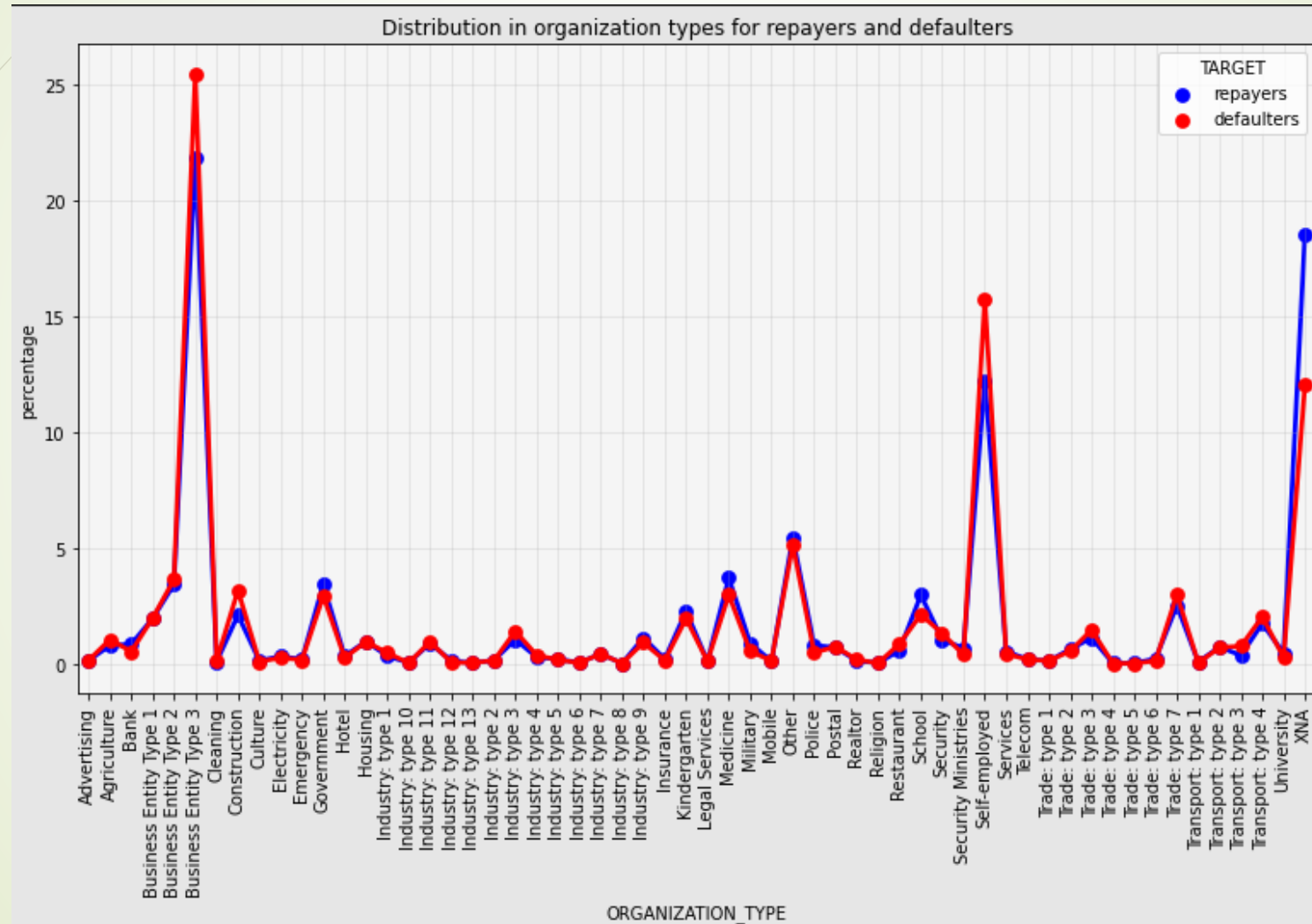
Loan application processes are highest on Tuesday

Peak hours for applying loans



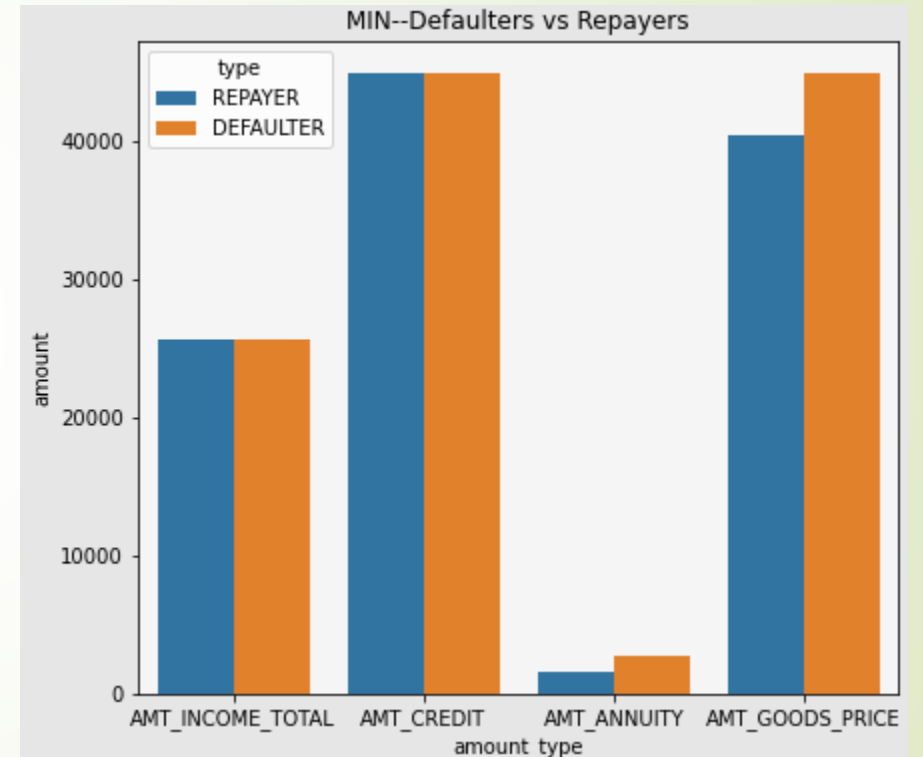
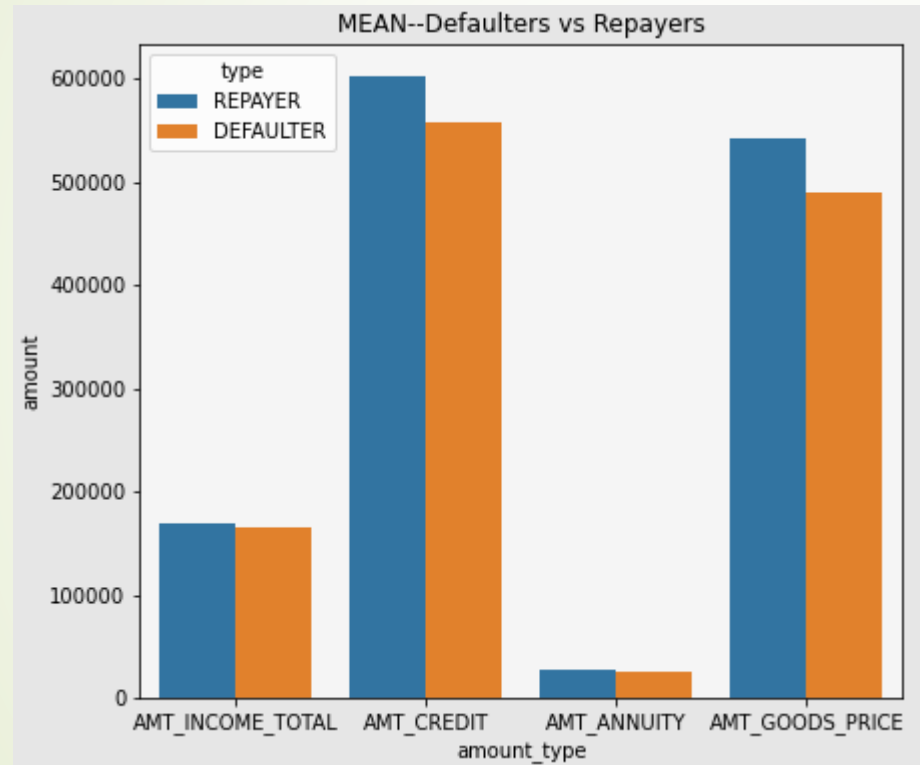
Around 10:00 AM to 12:00 PM, loan application processes are highest

Distribution in organization types

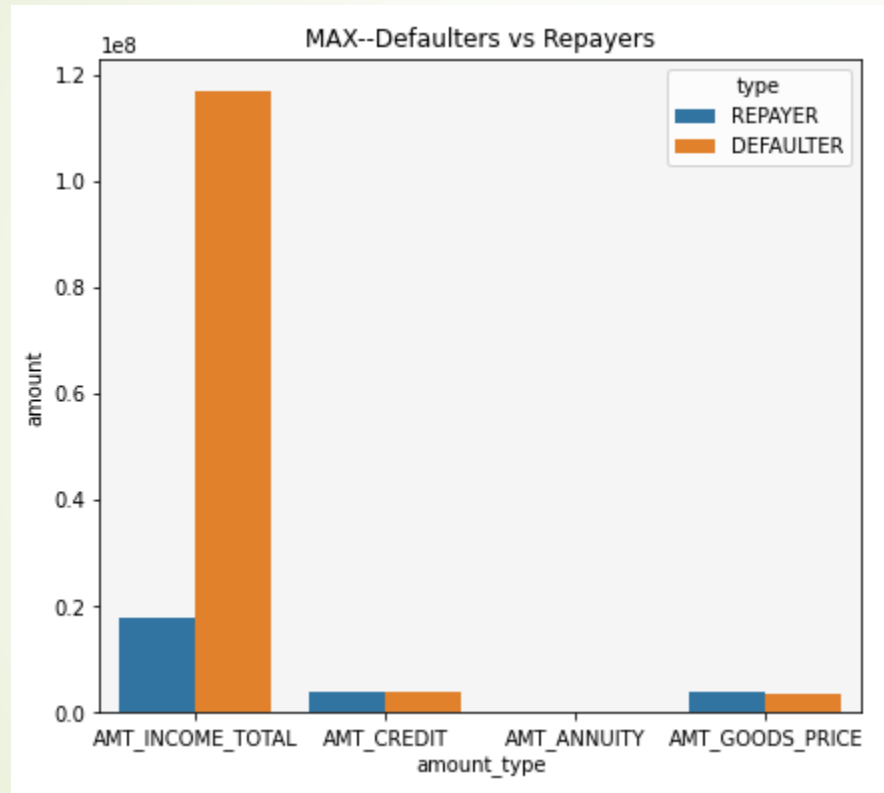


Organizations like Business Entity Type 3, Construction, Industry-type 3, Self-employed have percentage of defaulters higher than repayers

Amount columns vs Target variable

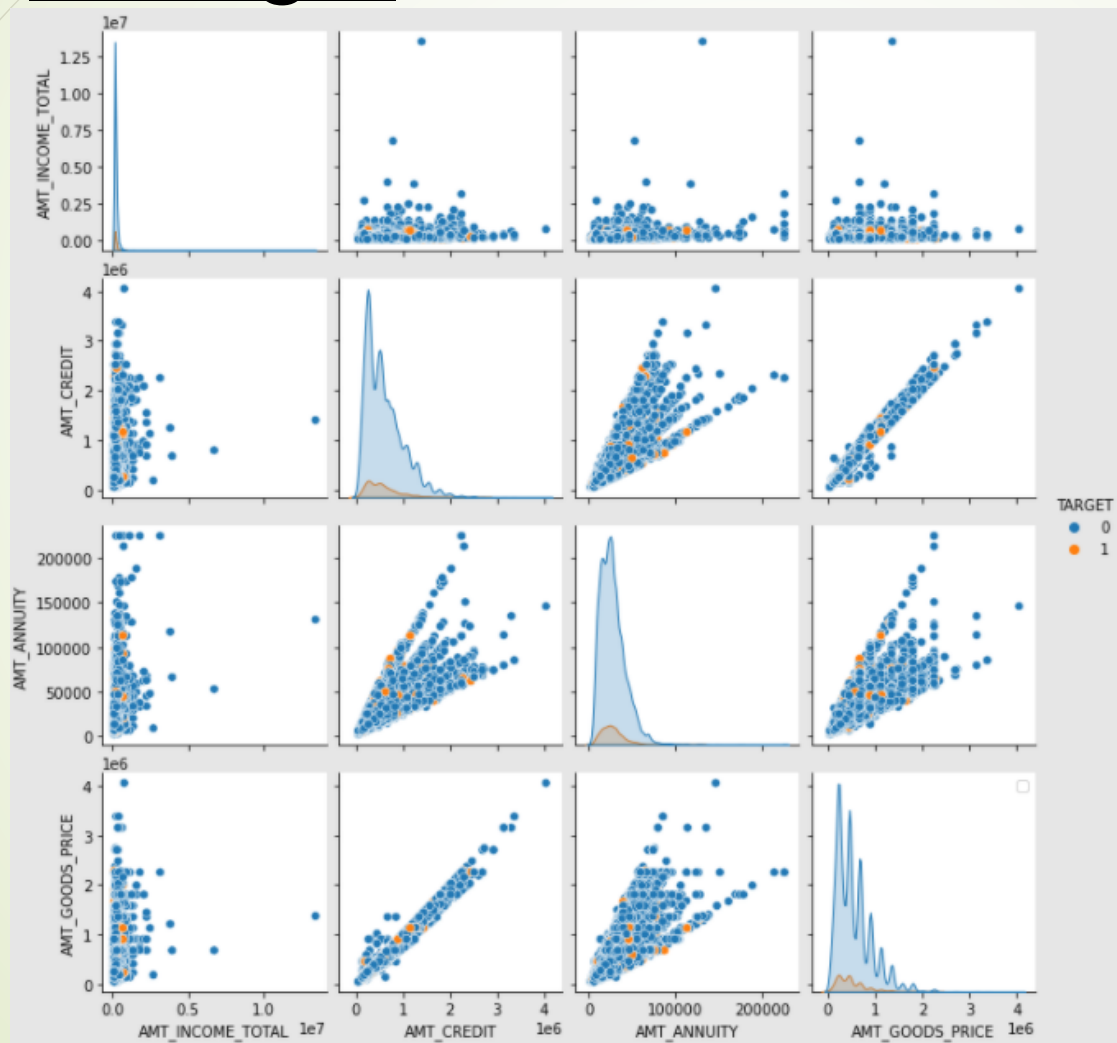


Amount columns vs Target variable



1. Average income of defaulters and repayers are almost same
2. Maximum income is of the defaulter
3. Average credit loan given to repayers is higher than defaulters
4. Average price of goods in case of repayers is higher than defaulters

Income vs Credit vs Annuity vs Goods Price vs Target



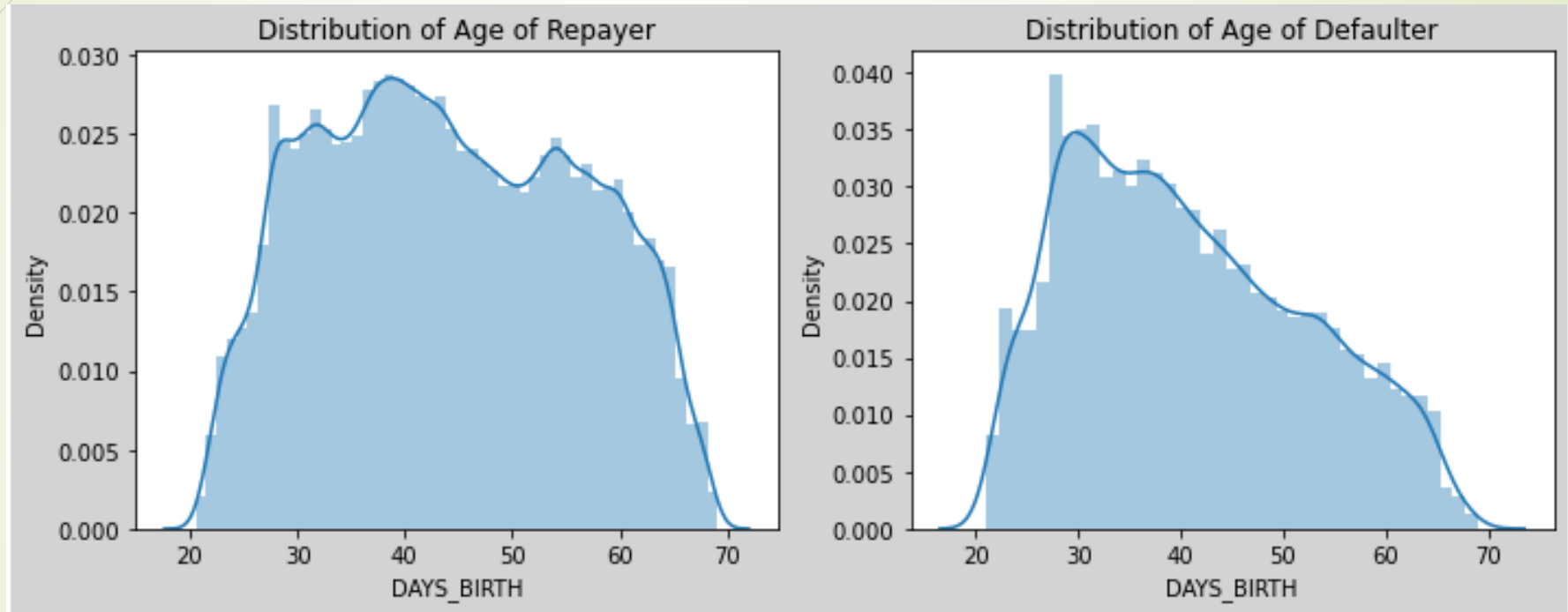
1. Goods price is directly proportional to credit amount and loan annuity. As price of the goods increases, credit amount and loan annuity also increases.
2. Loan annuity, and credit amount is similarly related to income

Distribution Of Age of the Employees



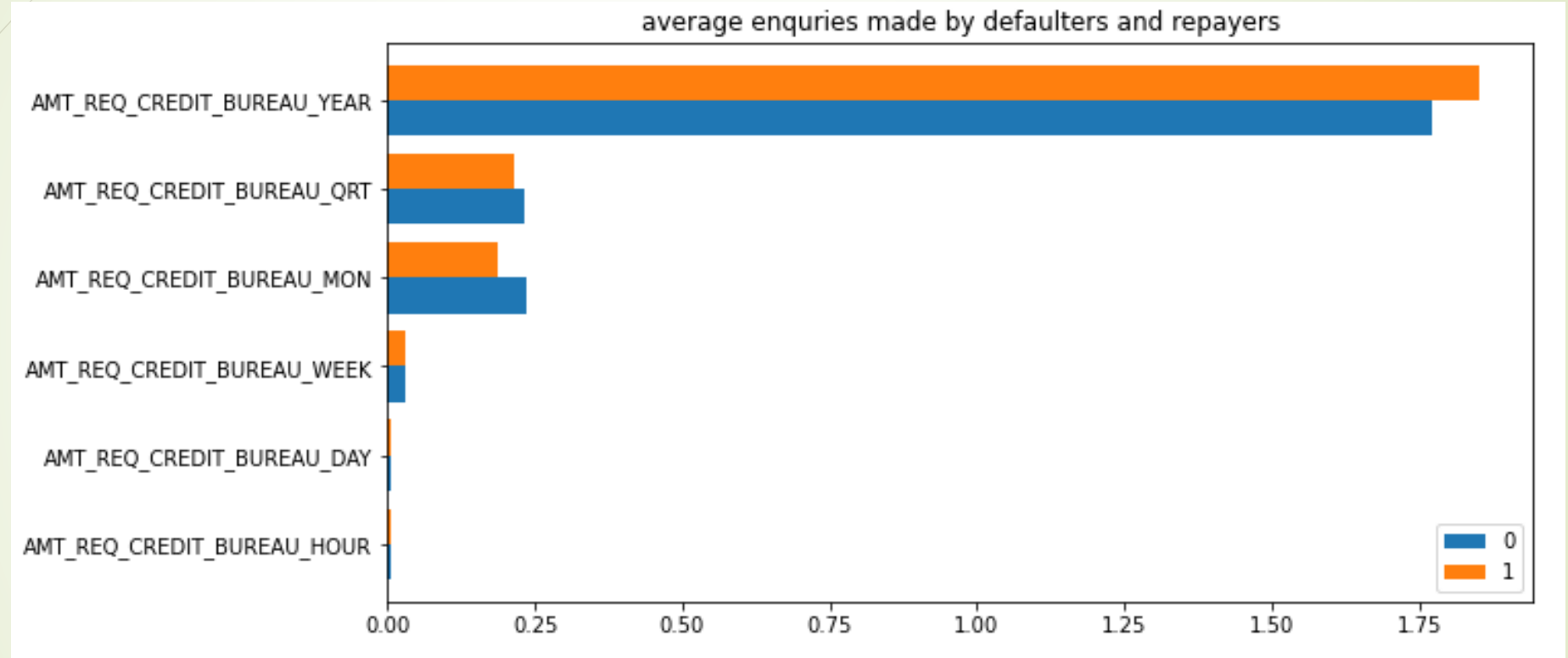
Most of the applicant's age lie between 35 and 55

Distribution of Age vs Target



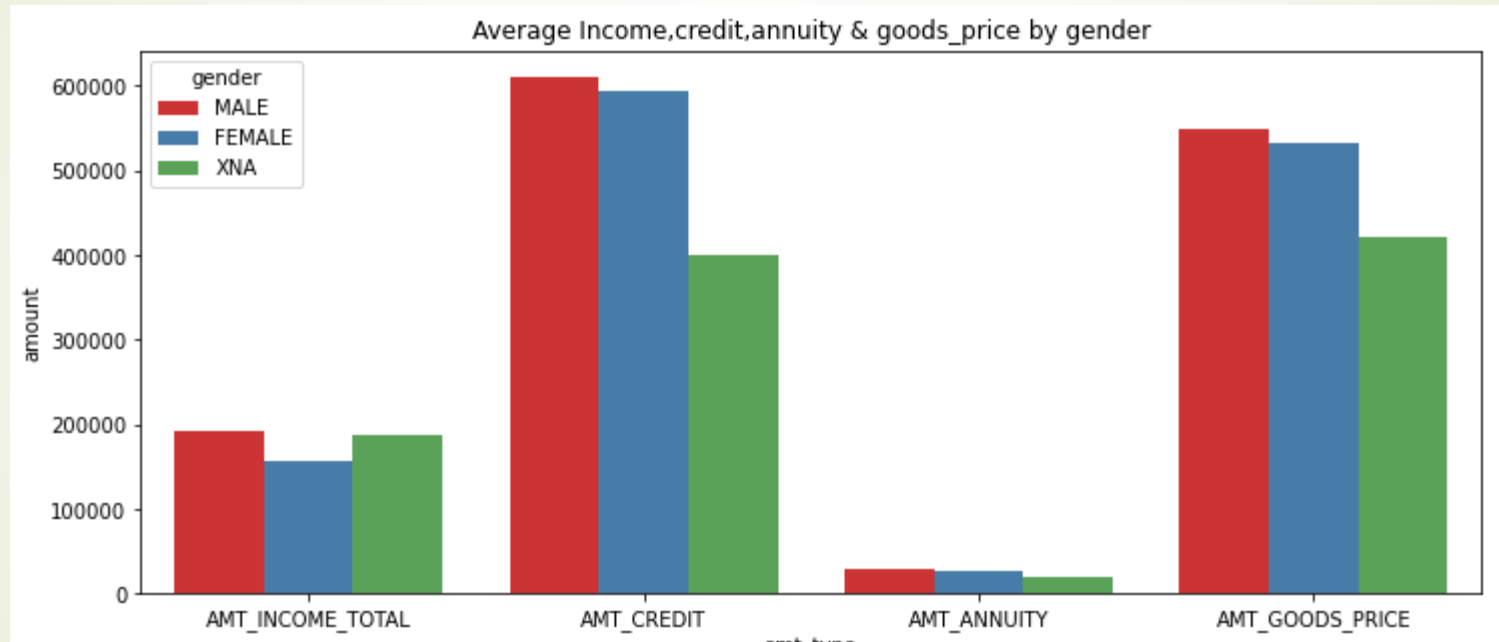
Most of the defaulters age lie around 30

Distribution to Enquiries to Credit Bureau



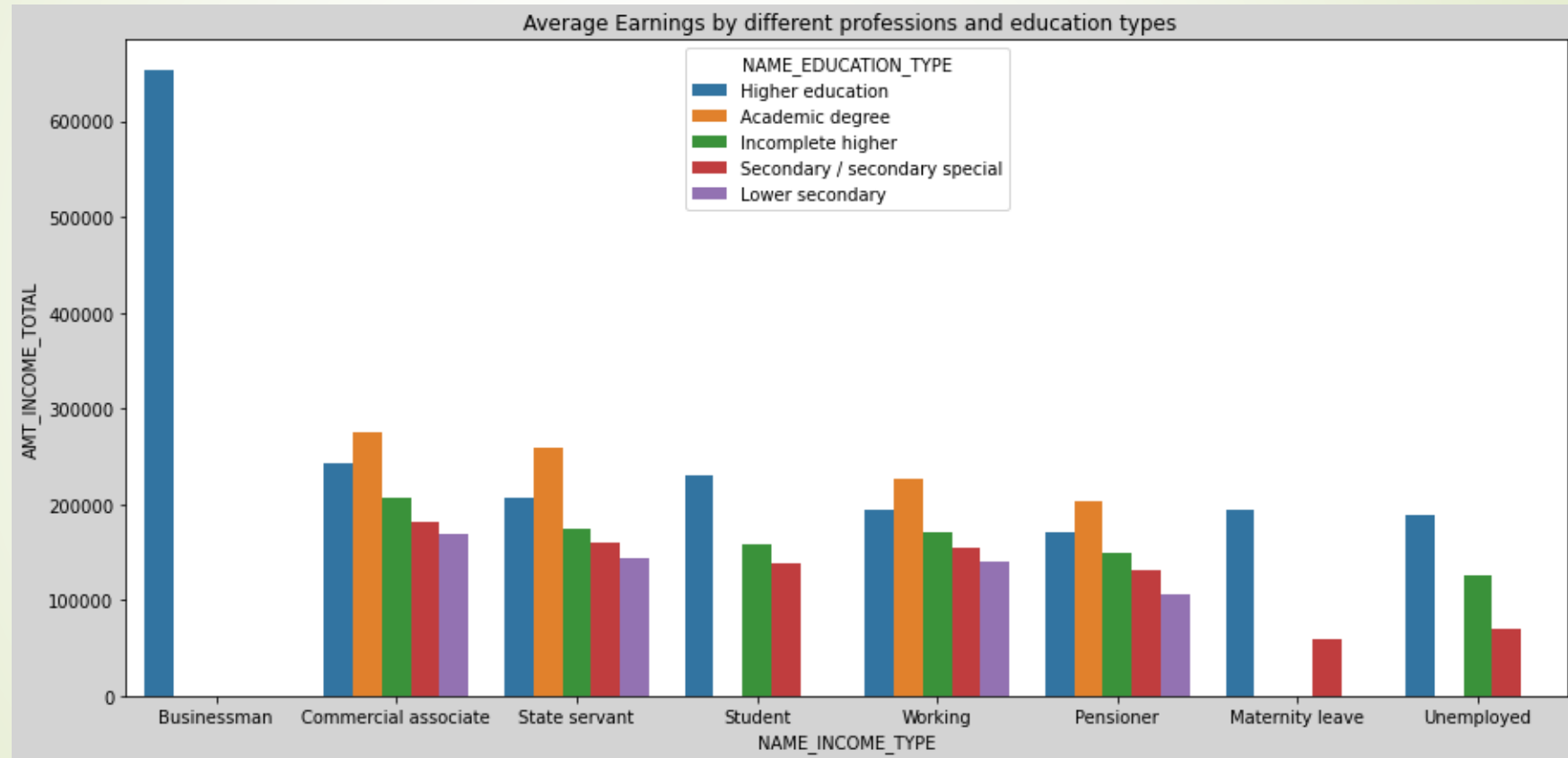
Average number of enquiries to Credit Bureau about the client 1 year ago before application is maximum for both non defaulters and defaulters

Average Income, credit, annuity & goods_price by gender



1. Male clients have higher average income
2. Average credit is given higher to male clients
3. Average loan annuity is higher for male clients
4. Average price of goods is loaned by male clients are higher than female clients

Average Earnings by different professions and education types



1. Businessman with higher education has the highest income.
2. After higher education those with academic degree has the highest salary

Top 10 correlation for the Client with payment difficulties

OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998270
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998270
AMT_GOODS_PRICE	AMT_CREDIT	0.982783
AMT_CREDIT	AMT_GOODS_PRICE	0.982783
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956637
CNT_CHILDREN	CNT_FAM_MEMBERS	0.885484
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.869016
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.869016
dtype: float64		

Number of observation of client's social surroundings with observable 30 DPD (days past due) default and that of 60 DPD has the highest correlation for the clients having payment difficulties

Top 10 correlation for the Client with all other cases

OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998510
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998510
AMT_GOODS_PRICE	AMT_CREDIT	0.987022
AMT_CREDIT	AMT_GOODS_PRICE	0.987022
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950149
CNT_CHILDREN	CNT_FAM_MEMBERS	0.878571
CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.861861
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.861861

dtype: float64

Number of observation of client's social surroundings with observable 30 DPD (days past due) default and that of 60 DPD has the highest correlation for the clients having no payment difficulties



THANK YOU