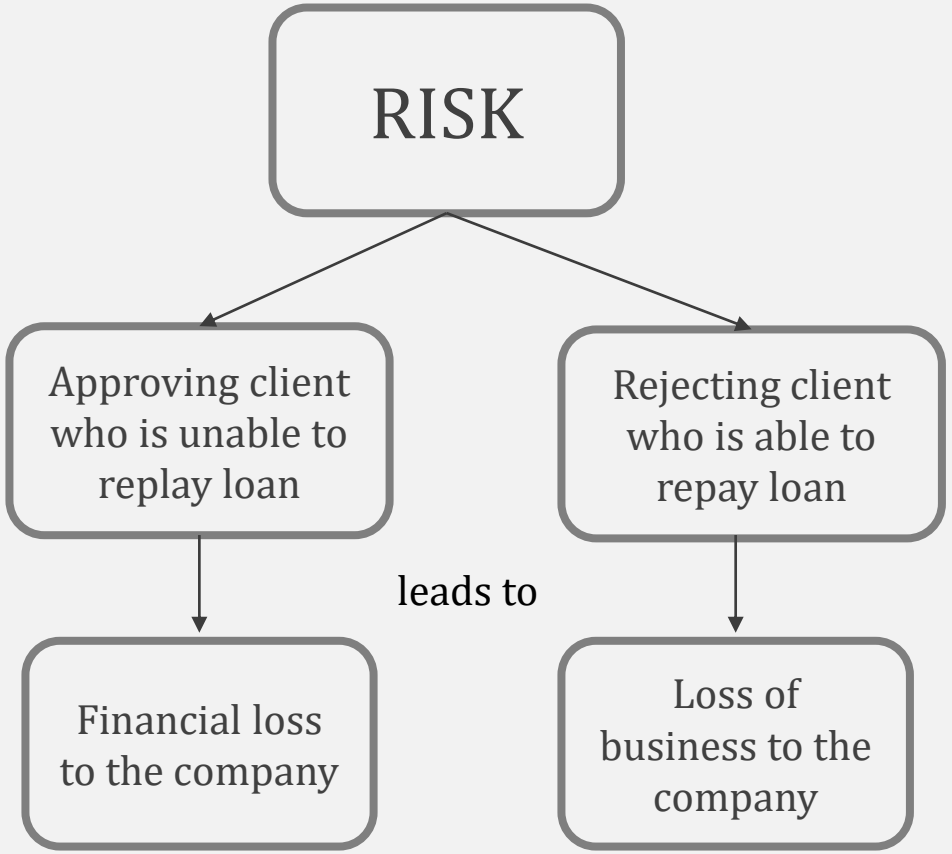
EDA Assignment

Abhishek
Thilakan

INDEX

Problem Statement



To avoid the risks mentioned in the above flow chart, patterns needs to be analyzed to identify potential clients and driving factors has to be determined for the benefit of bank.

Limitations

- Only a few columns or variables from the whole data set have been considered for the analysis.
- The visualization graphs have been prepared in the Jupyter notebook for every considered variable.
- Only a few of them have been presented in this document which I found useful.

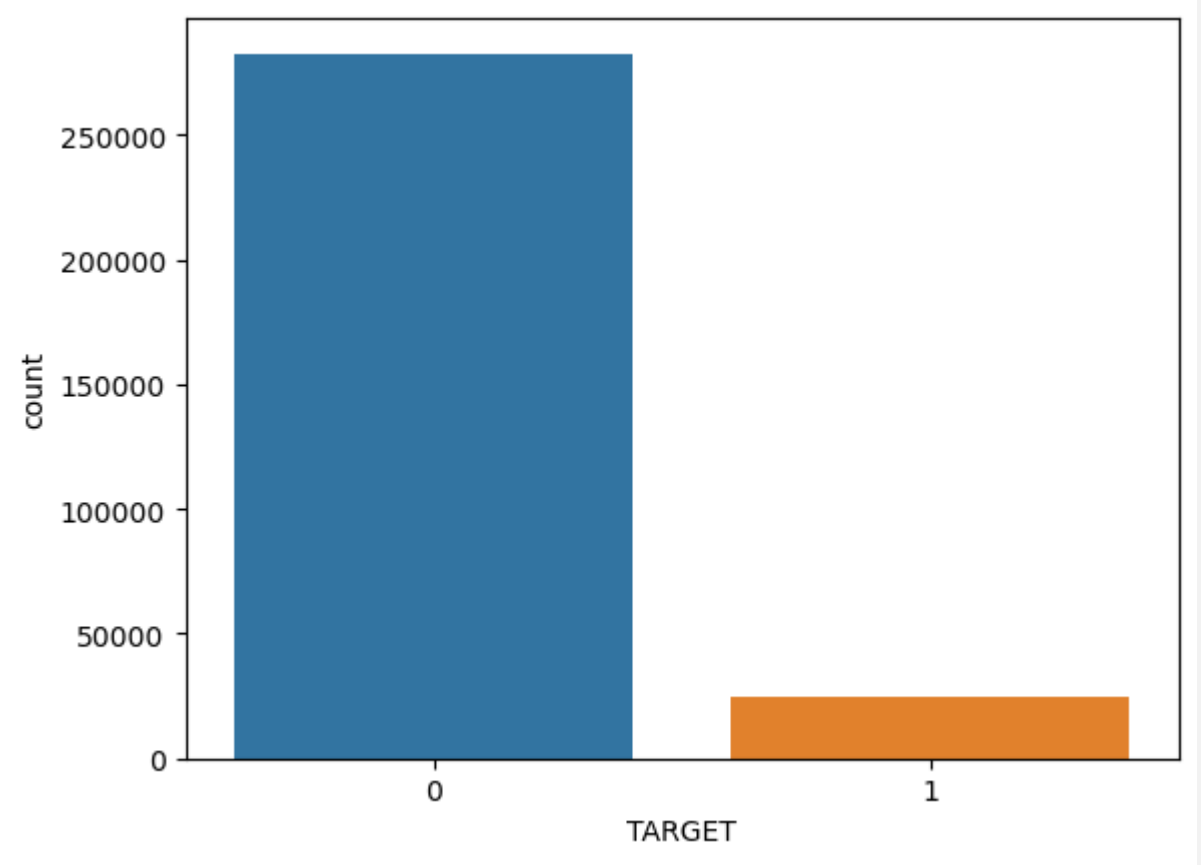
Ratio of Data Imbalance

The ratio of data imbalance is calculated by dividing the length of a data by another.

In our case,

ratio = len(targeto)/ len(target1)

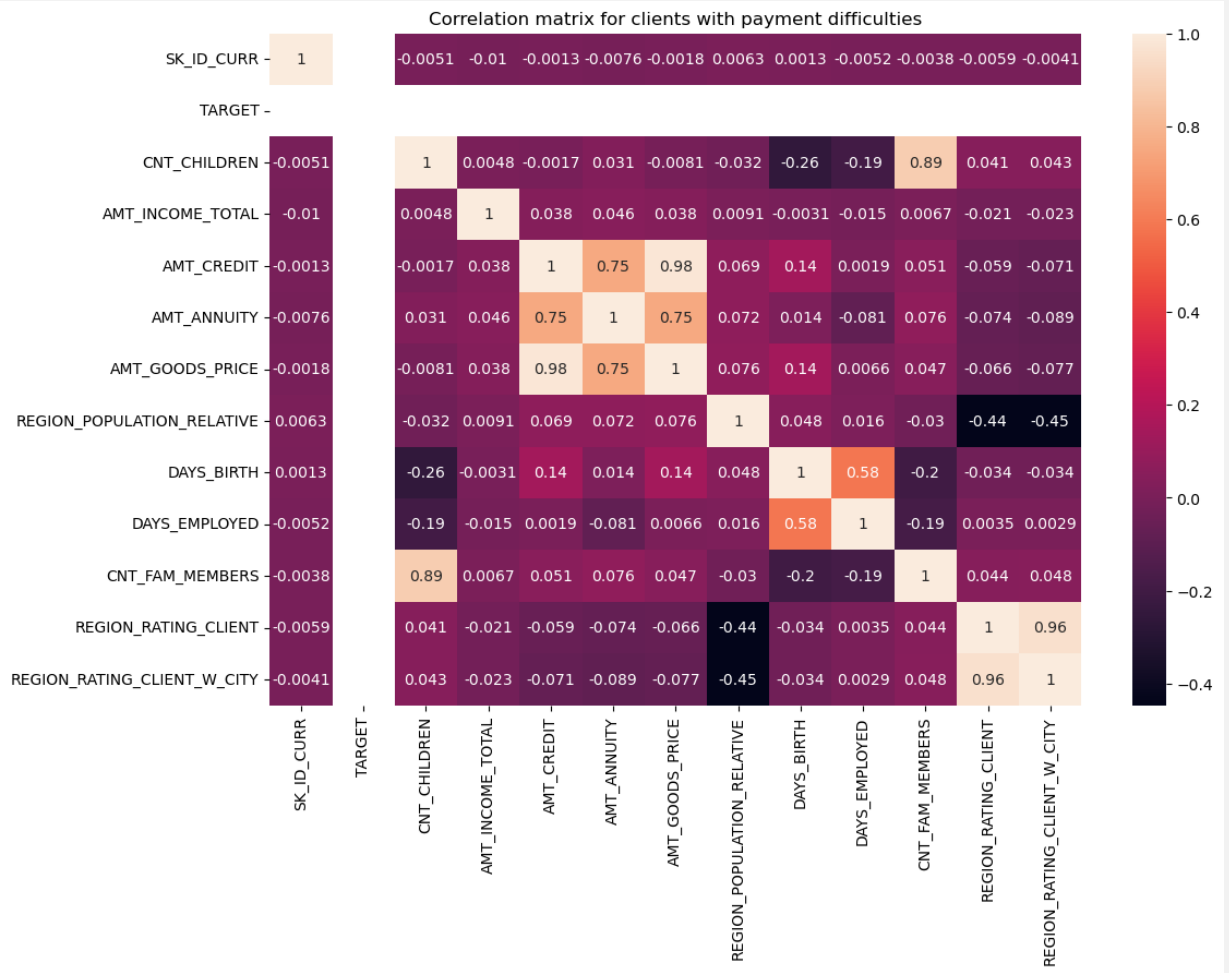
The ratio of data imbalance is 11.39



Top Correlation

The top 10 correlation with respect to the clients with payment difficulties can be seen here.

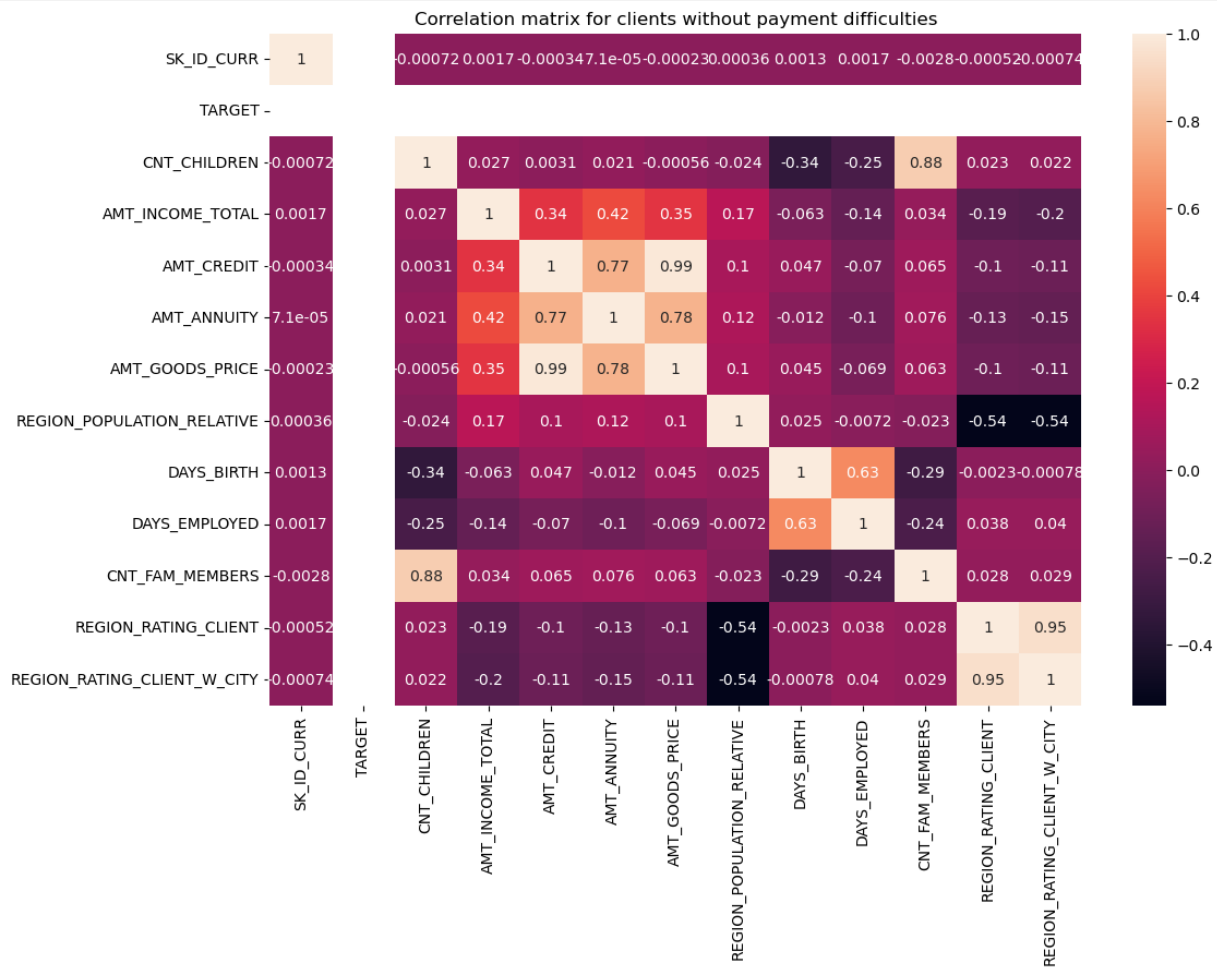
	VAR1	VAR2	Correlation_Value	Corr_abs
82	AMT_GOODS_PRICE	AMT_CREDIT	0.982783	0.982783
167	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637	0.956637
132	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484	0.885484
83	AMT_GOODS_PRICE	AMT_ANNUITY	0.752295	0.752295
69	AMT_ANNUITY	AMT_CREDIT	0.752195	0.752195
125	DAYS_EMPLOYED	DAYS_BIRTH	0.582185	0.582185
163	REGION_RATING_CLIENT_W_CITY	REGION_POPULATION_RELATIVE	-0.446977	0.446977
150	REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	-0.443236	0.443236
106	DAYS_BIRTH	CNT_CHILDREN	-0.259109	0.259109
138	CNT_FAM_MEMBERS	DAYS_BIRTH	-0.203267	0.203267



Top Correlation

The top 10 correlation with respect to the clients without payment difficulties can be seen here.

	VAR1	VAR2	Correlation_Value	Corr_abs
82	AMT_GOODS_PRICE	AMT_CREDIT	0.987022	0.987022
167	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.950149	0.950149
132	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571	0.878571
83	AMT_GOODS_PRICE	AMT_ANNUITY	0.776421	0.776421
69	AMT_ANNUITY	AMT_CREDIT	0.771297	0.771297
125	DAYS_EMPLOYED	DAYS_BIRTH	0.626114	0.626114
150	REGION_RATING_CLIENT	REGION_POPULATION_RELATIVE	-0.539005	0.539005
163	REGION_RATING_CLIENT_W_CITY	REGION_POPULATION_RELATIVE	-0.537301	0.537301
68	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418948	0.418948
81	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349426	0.349426

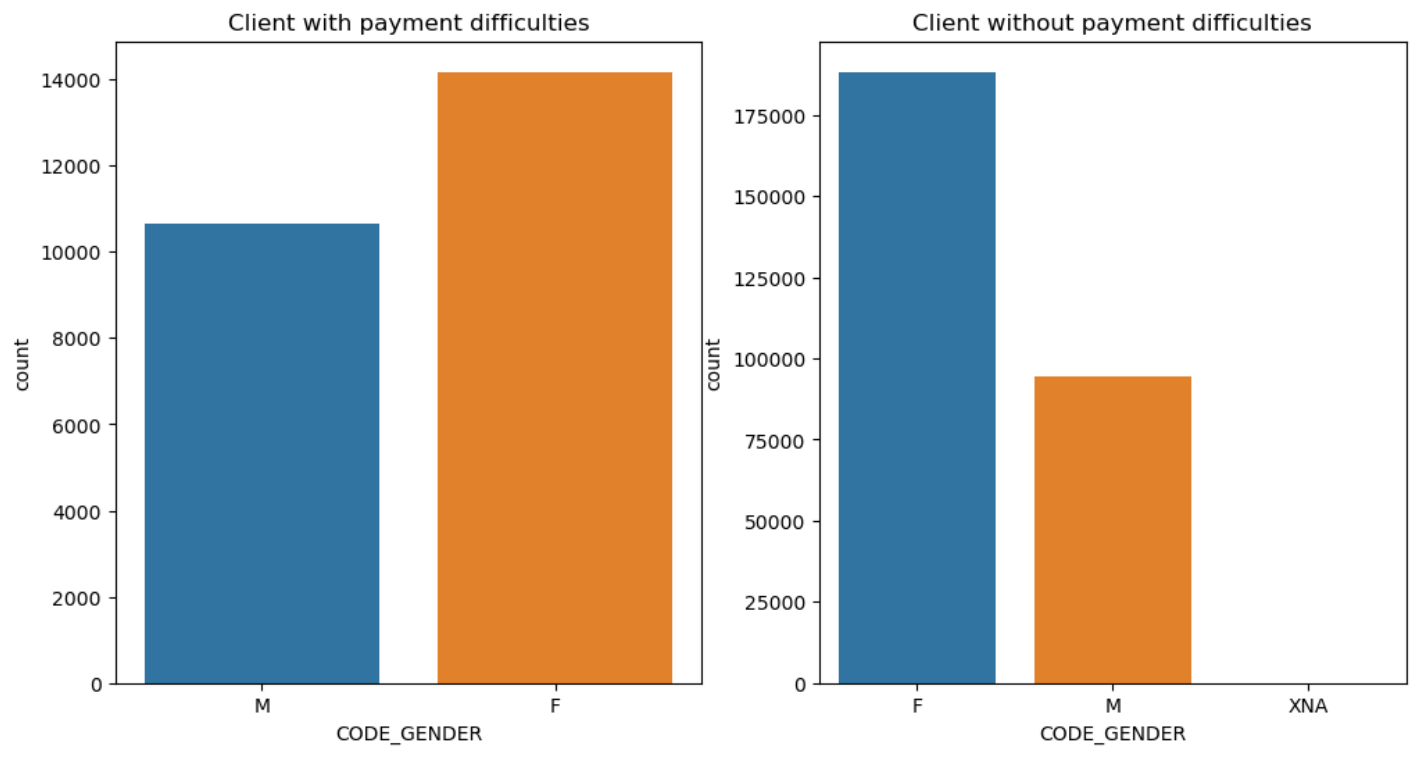


Top Correlation

In both the cases, it can be observed that the top 3 correlation are between the below variables.

VAR1	VAR2
AMT_GOODS_PRICE	AMT_CREDIT
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT
CNT_FAM_MEMBERS	CNT_CHILDREN

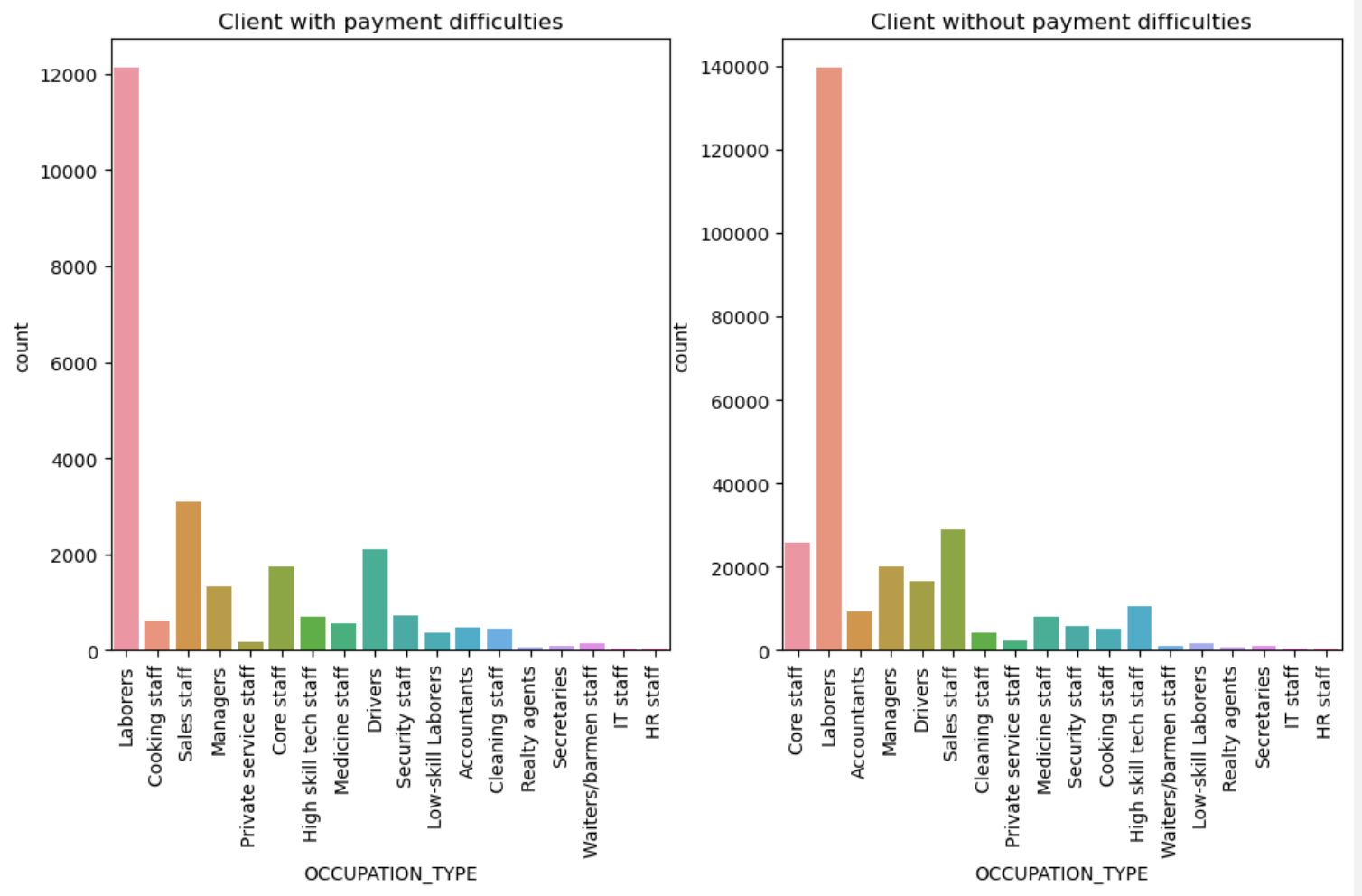
Univariate Analysis



Females are more tend to take loans compared to men.

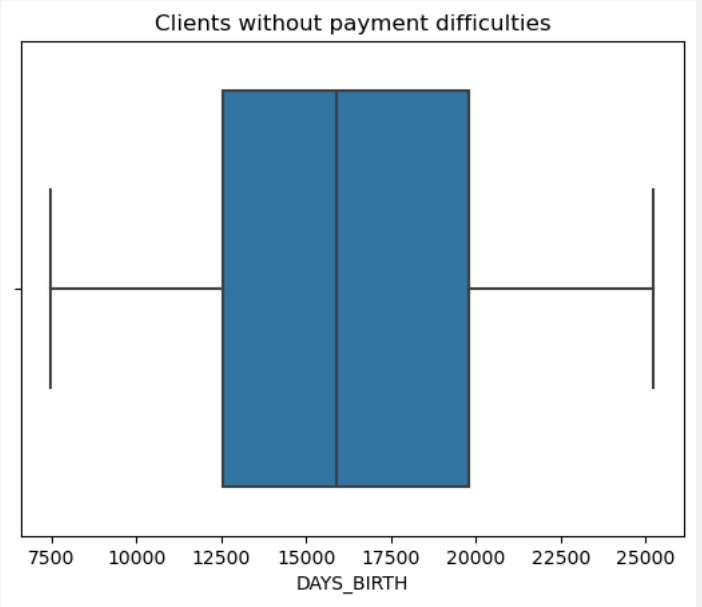
Univariate Analysis

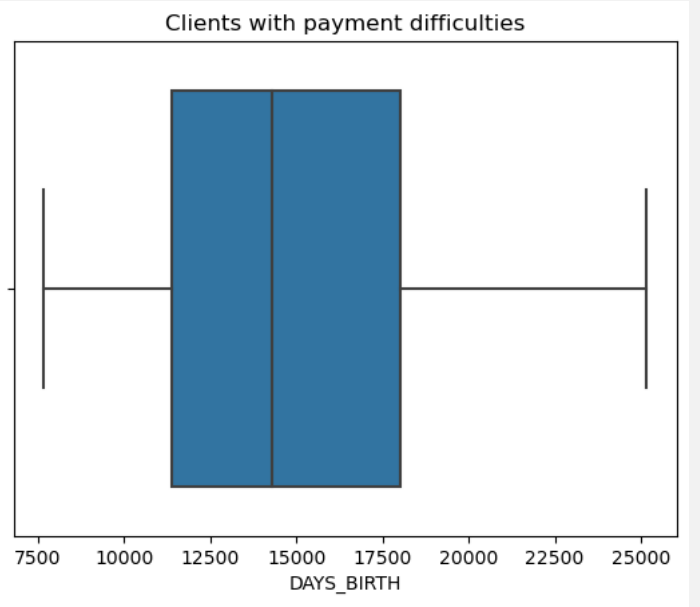
- Laborers, Sales Staff, Drivers and Core Staff are the occupation types majorly involved in taking a loan.
- Out of these four, laborers could be a good choice for approving loans as the number of laborers who can repay loans is more than 11 times of the laborers who cannot.



Univariate Analysis

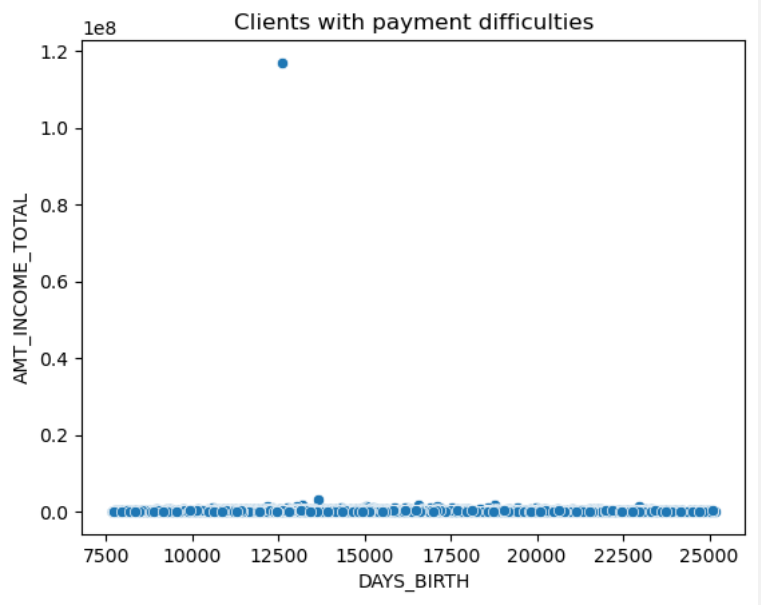
- The graphs show that clients within the age range of 35 – 55 were able to repay the loans easily whereas clients within the age range of 33 – 49 faced some difficulties.
- This seems unclear due to similarity in the age range. Hence, will try to analyse this with another variable for better understanding in the next slide.

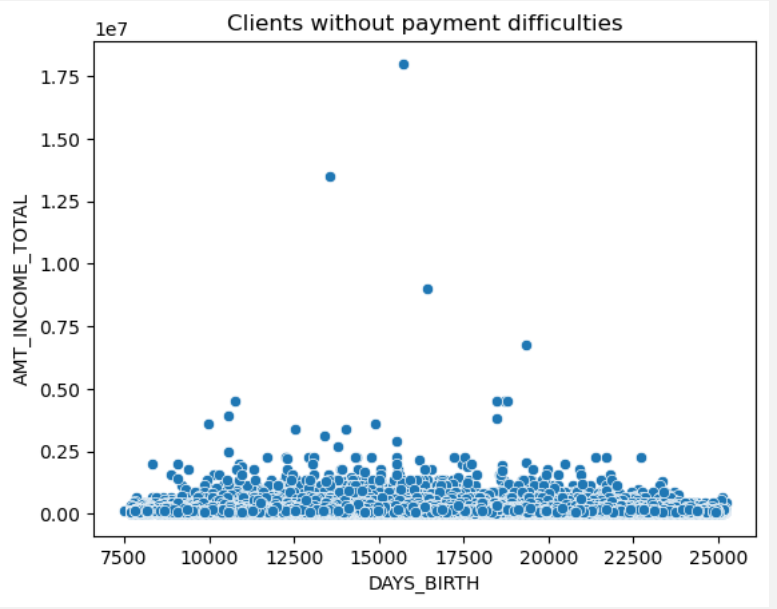




Bivariate Analysis

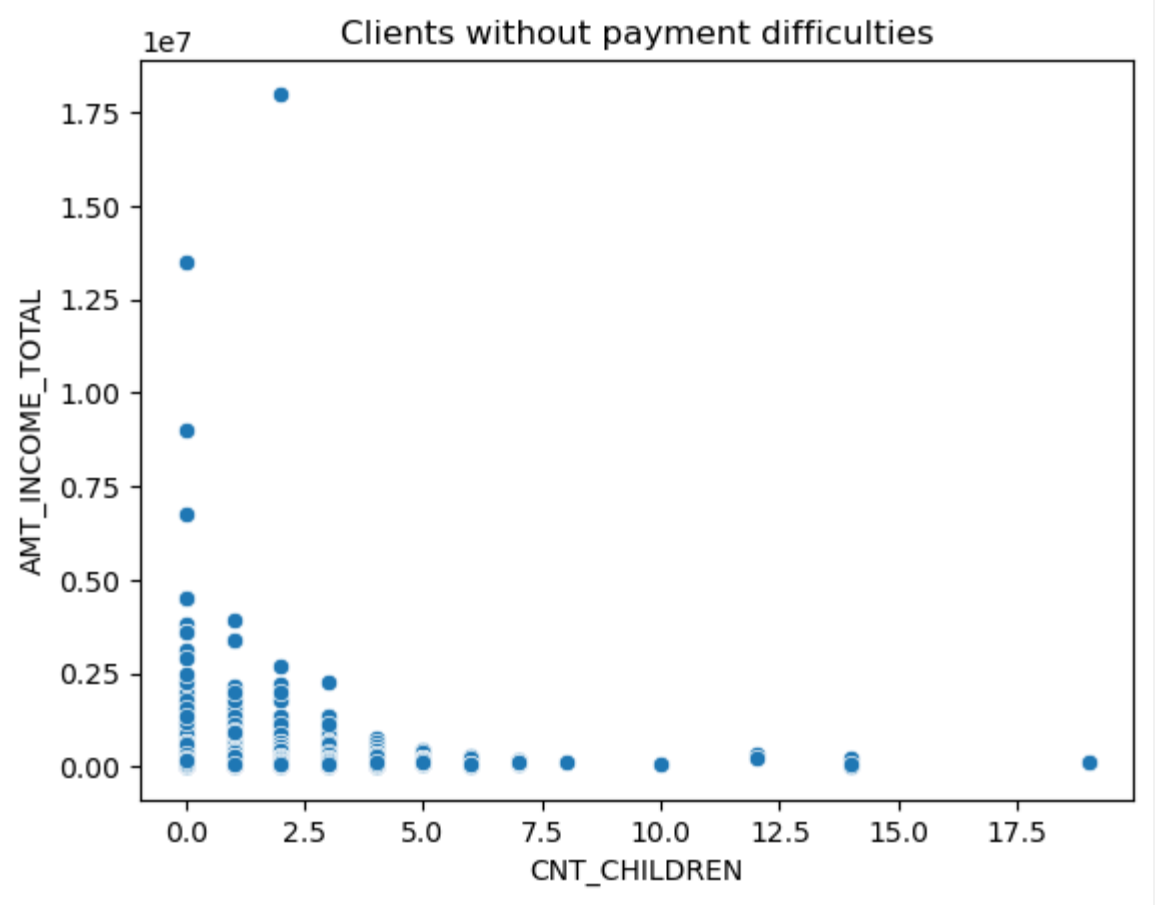
- For a better clarity on age of clients, will also consider the income of the clients.
- From the graphs beside, it is evident that age actually does not form a major driving force for selecting potential clients but income does.
- As we can see that, clients from every age range face difficulties in repaying the loan if their income is less.





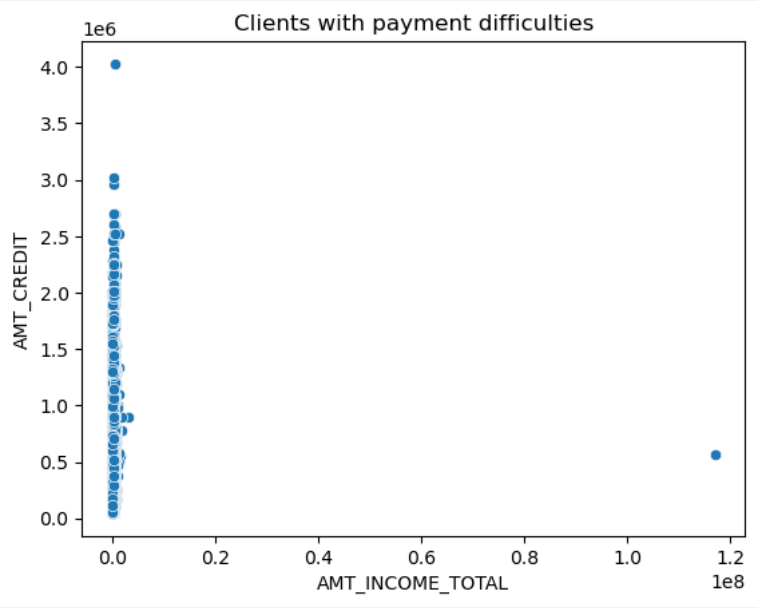
Bivariate Analysis

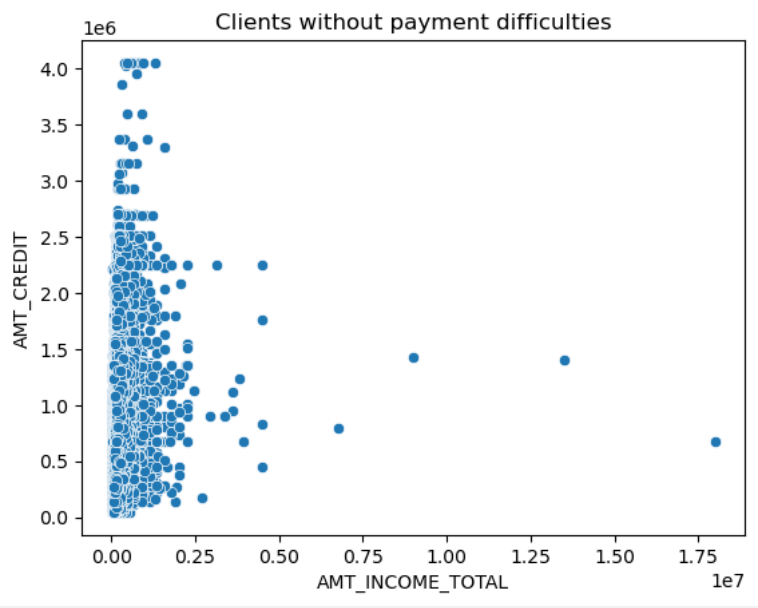
Clients with higher income and less number of children can always be a good choice as we can see that such cases have high chances of repaying the loan easily.



Bivariate Analysis

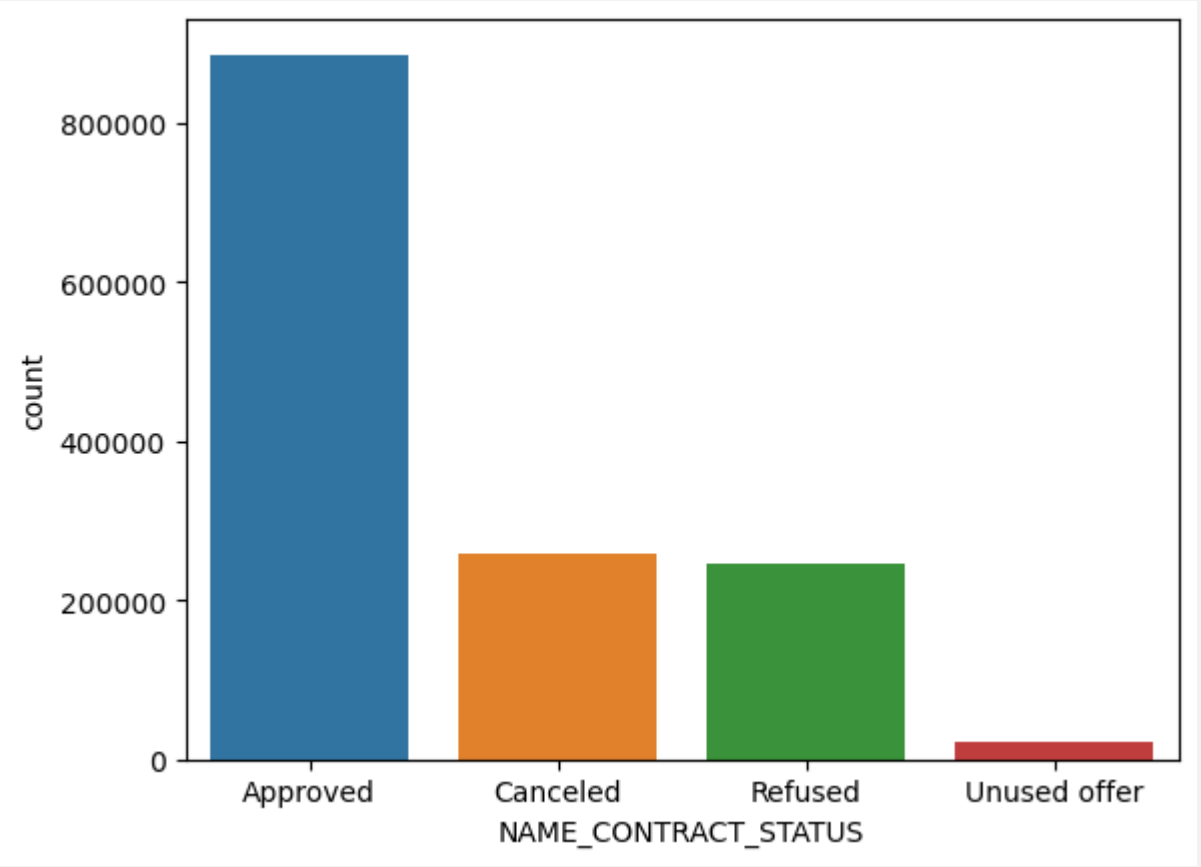
- Another variable which supports income of the client is the credit amount here.
- We can observe that regardless of the credit amount being high or less, clients with less income always face difficulties to repay the loan.
- Hence, avoiding lower income ranges could be a good choice for the loan providers.





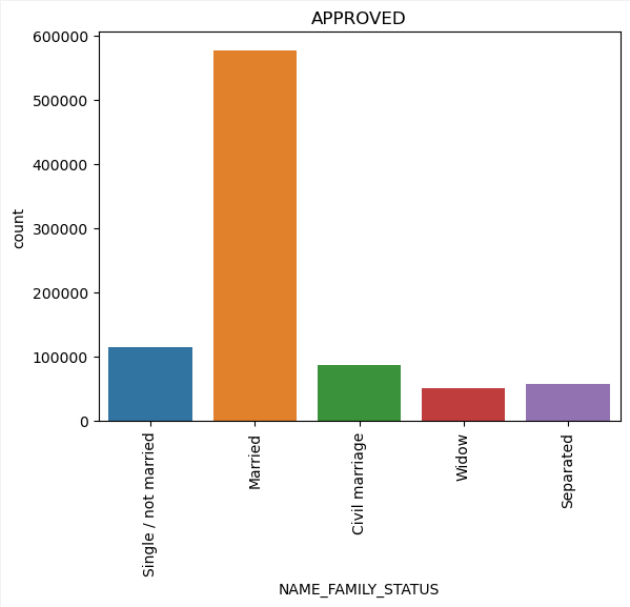
Merged Data Analysis

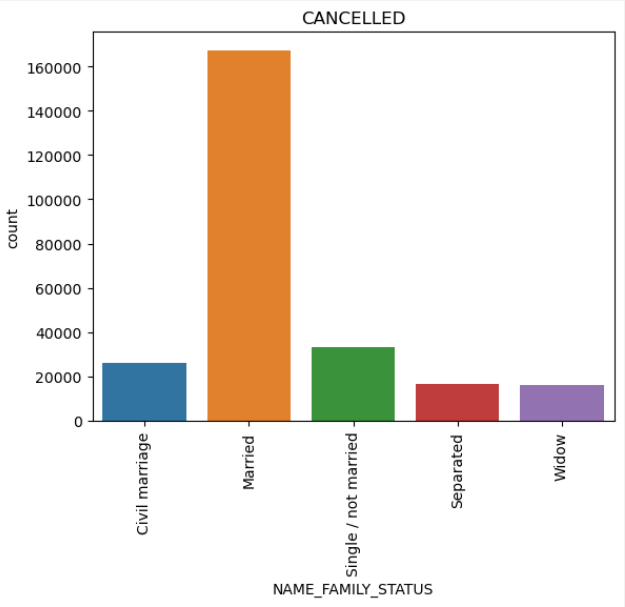
From the merged dataset, we observe that major portion of the loan applicants are approved whereas hardly a slight portion is unused as compared to others.

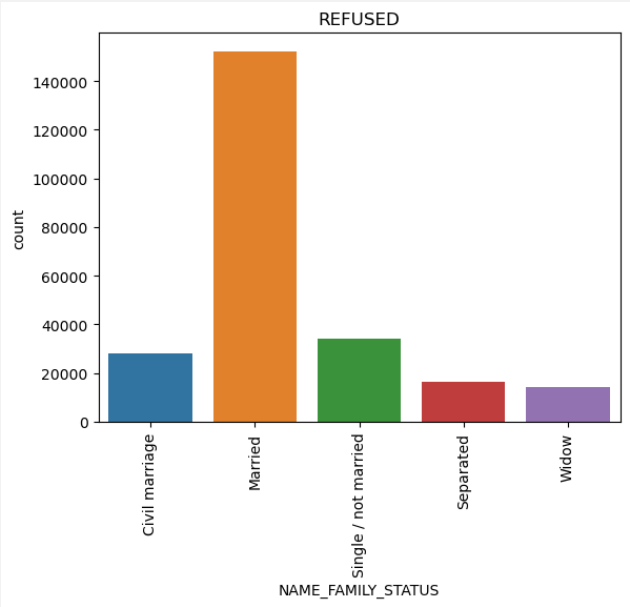


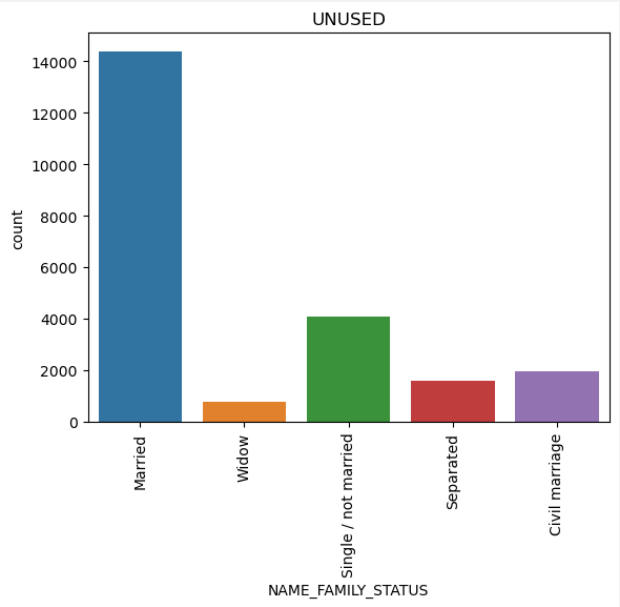
Merged Data Analysis

- Married clients contribute to a major share of loan applicants.
- Married clients who were approved are huge in number than the married clients who were rejected or cancelled.
- Hence, even married clients can be considered as a driving force to choose potential clients but it has to be analyzed with other variables too.









Suggestions / Recommendations

- Banks should focus majorly on married clients with good income as they are huge in number with lesser risk of payback difficulties.
- The above categories if managed along with number of children and assets could help even more.
- Income type working could be a risky choice due to high number of client who face difficulties in repaying the loan.

ThankYou