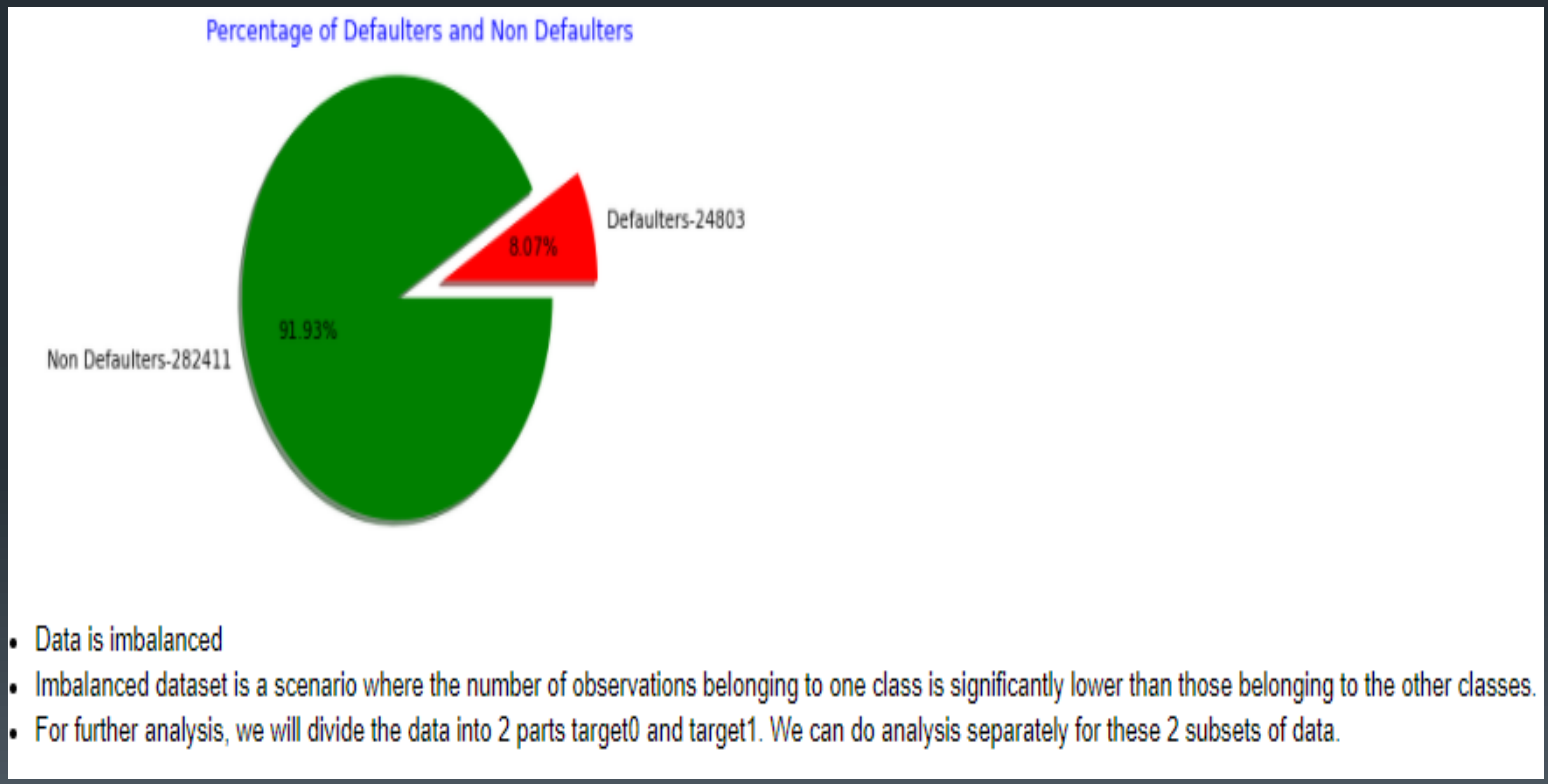
CREDIT EDA CASE STUDY

SUBMITTED BY:
DIVIT KARMIANI
DHWANI SISODIYA

Following steps were involved to reach at a conclusion:

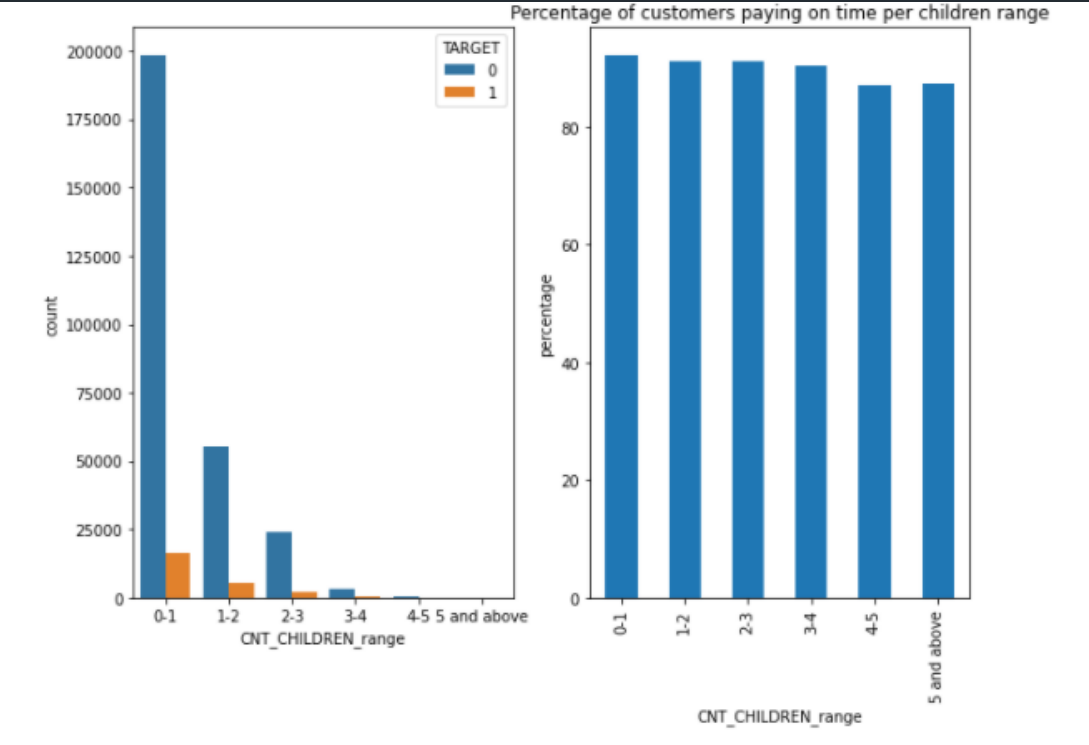
- Loading the application_data.csv
- Cleaning the data by dropping the null values
- Standardising values
- Performing sanity checks
- Finding imbalance percentage of the data
- Performing univariate and bivariate analysis
- Loading previous_data.csv
- Merging both the data
- Performing univariate and bivariate analysis

IMBALANCE RATIO

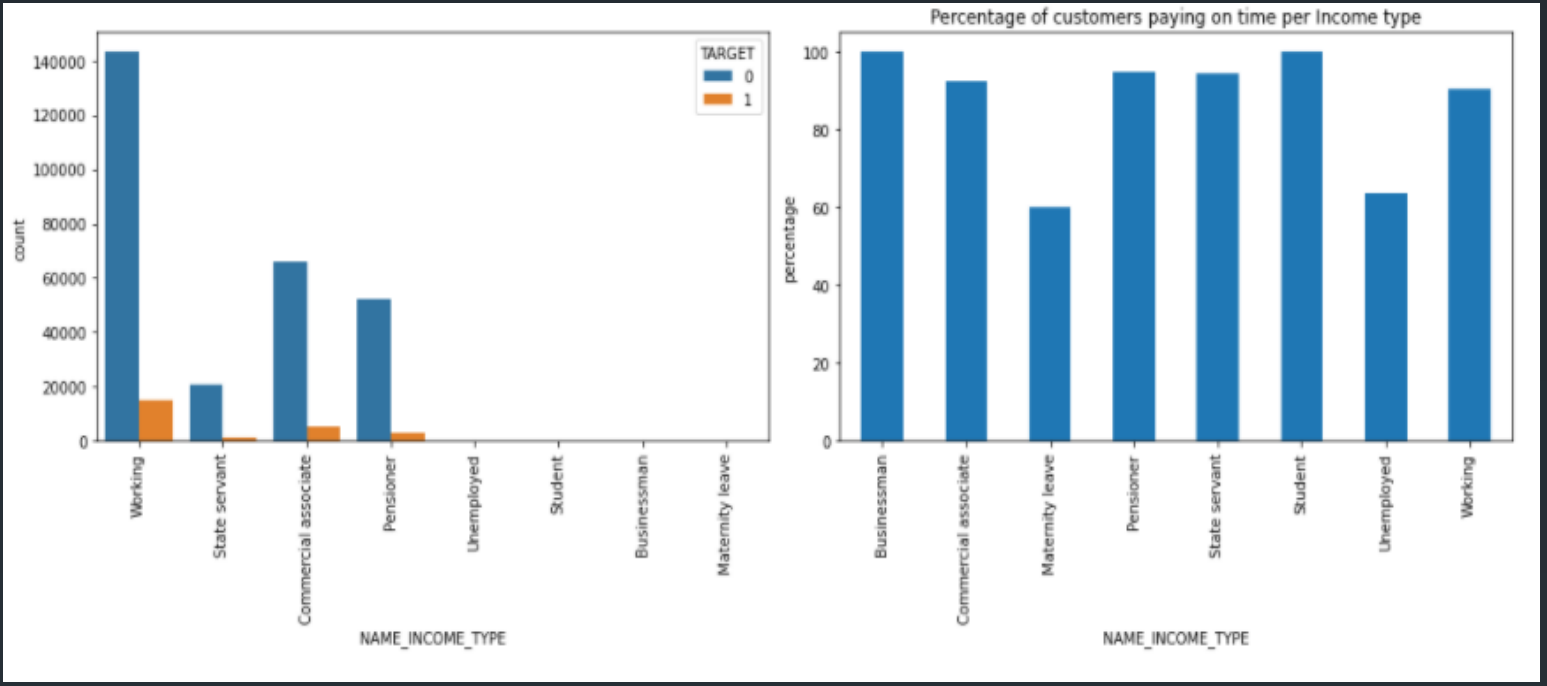


- Data is imbalanced
- Imbalanced dataset is a scenario where the number of observations belonging to one class is significantly lower than those belonging to the other classes.
- For further analysis, we will divide the data into 2 parts target0 and target1. We can do analysis separately for these 2 subsets of data.

UNIVARIATE ANALYSIS OF CATAGORICAL VARIABLES

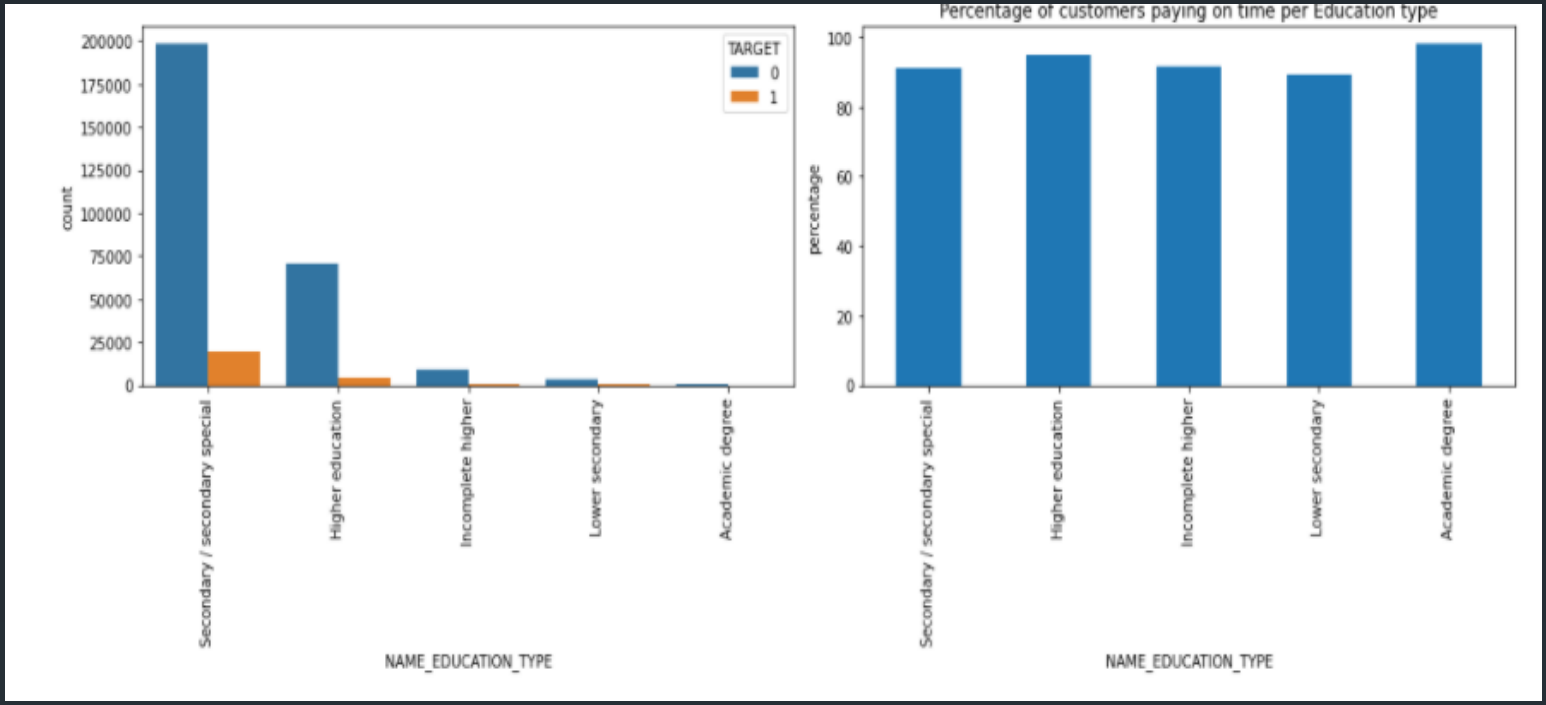


People with 0 children have taken more loans and are higher in terms of percentage in terms of paying on time.
People with 4 children or 5 and above children have difficulty in paying on time as compared to others.



Working, commercial associate and pensioners tend to take more loans as implied in this data

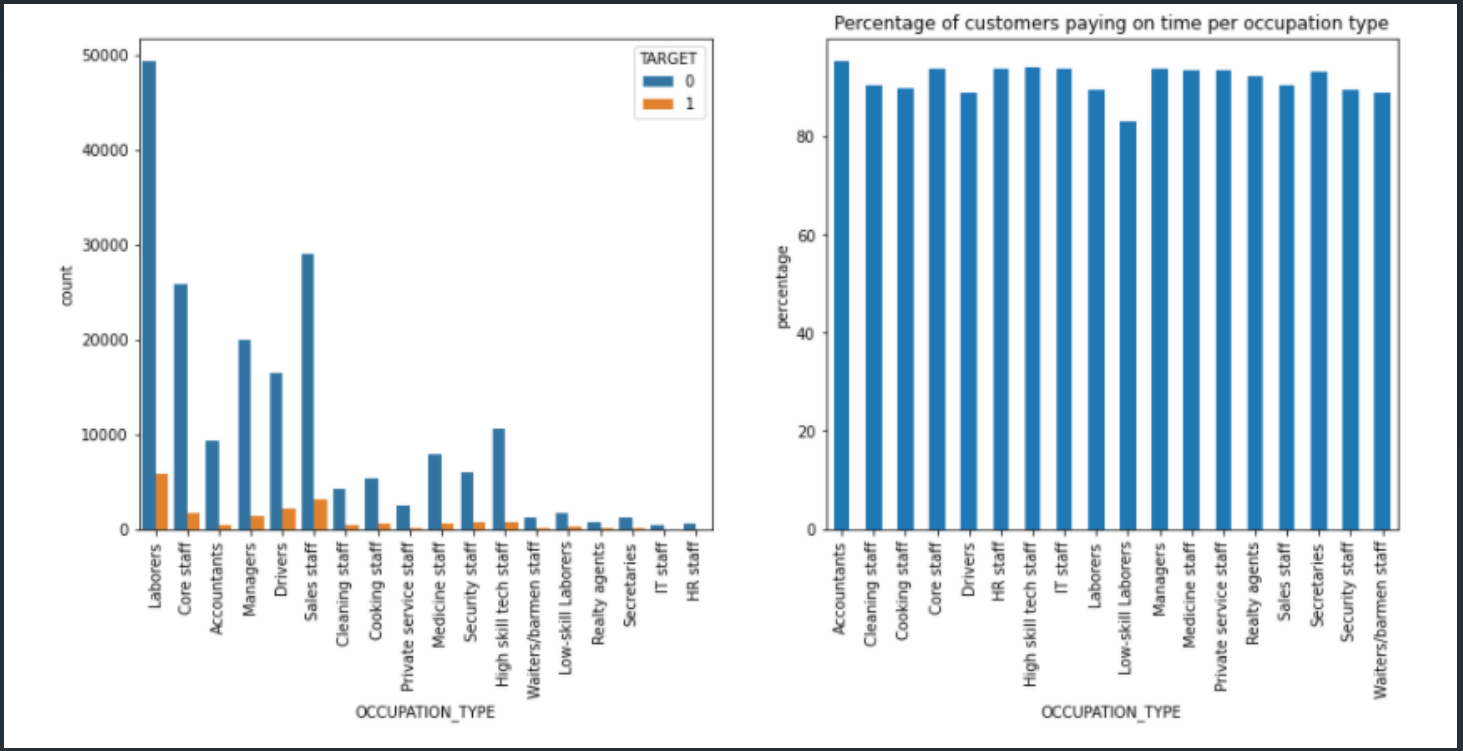
Businessman and Students have highest percentage in terms of paying on time while maternity leave and unemployed customers have lowest percentages. This may so happen because since they are unemployed or on maternity leave they might not be earning much and hence difficulty in paying on time.



Secondary/ secondary special and Higher education customers have taken more loans.

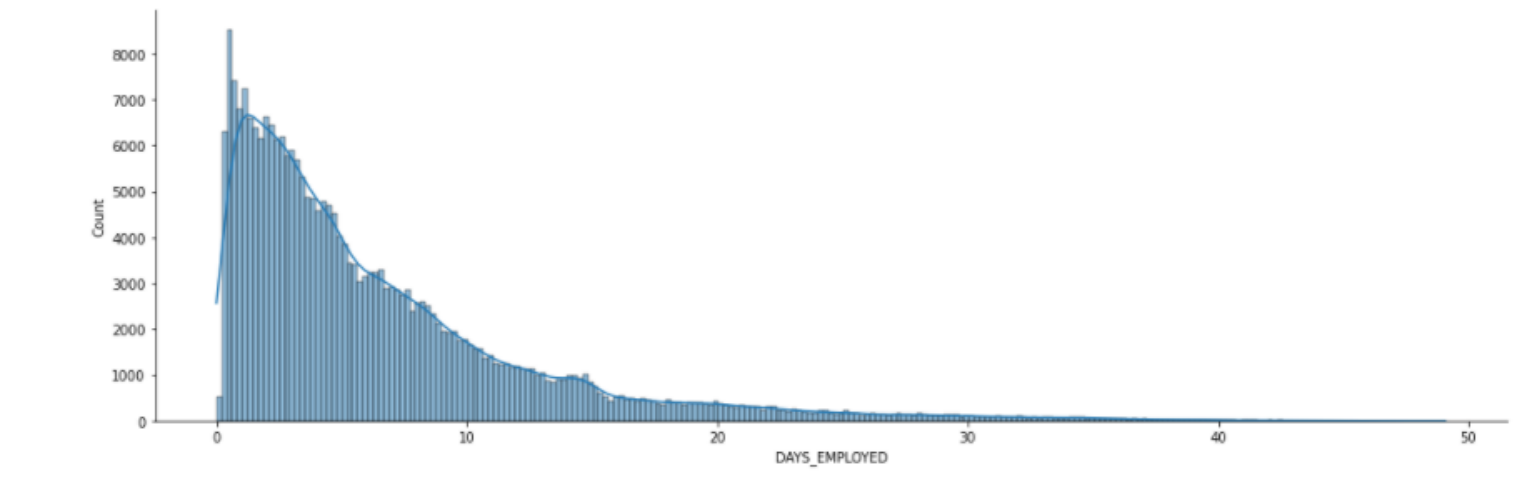
If we talk in percentages per education type, Academic degree holders have higher percentage in terms of paying on time whereas lower secondary customers have the lowest percentage in terms of paying on time.

This may be because, degree holders might get better jobs and hence higher pay and more chances of paying loans on time. But this is something to be verified. The above can be formed as a hypothesis.

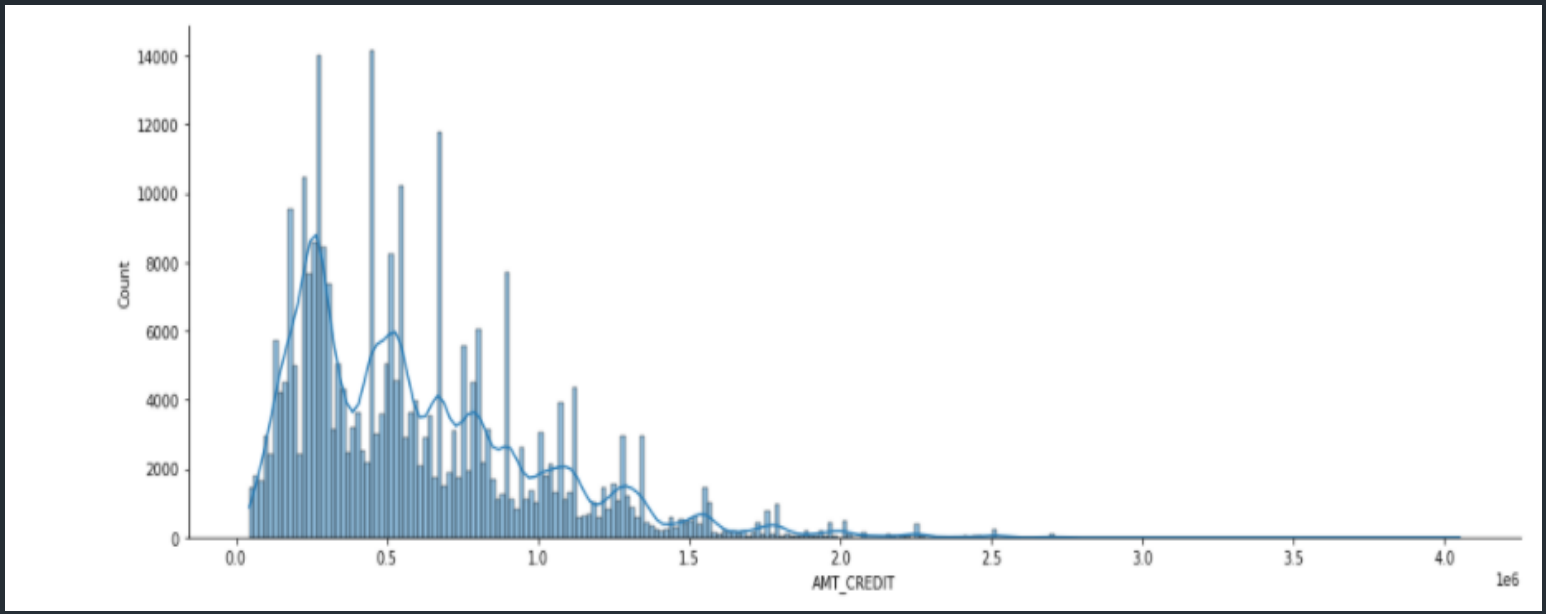


Labourers and Sales staff have taken more loans.
Low-skill labourers have most difficulty in paying the loan on time, while Accountants have the least difficulty.

UNIVARIATE ANALYSIS OF NUMERICAL VARIABLES

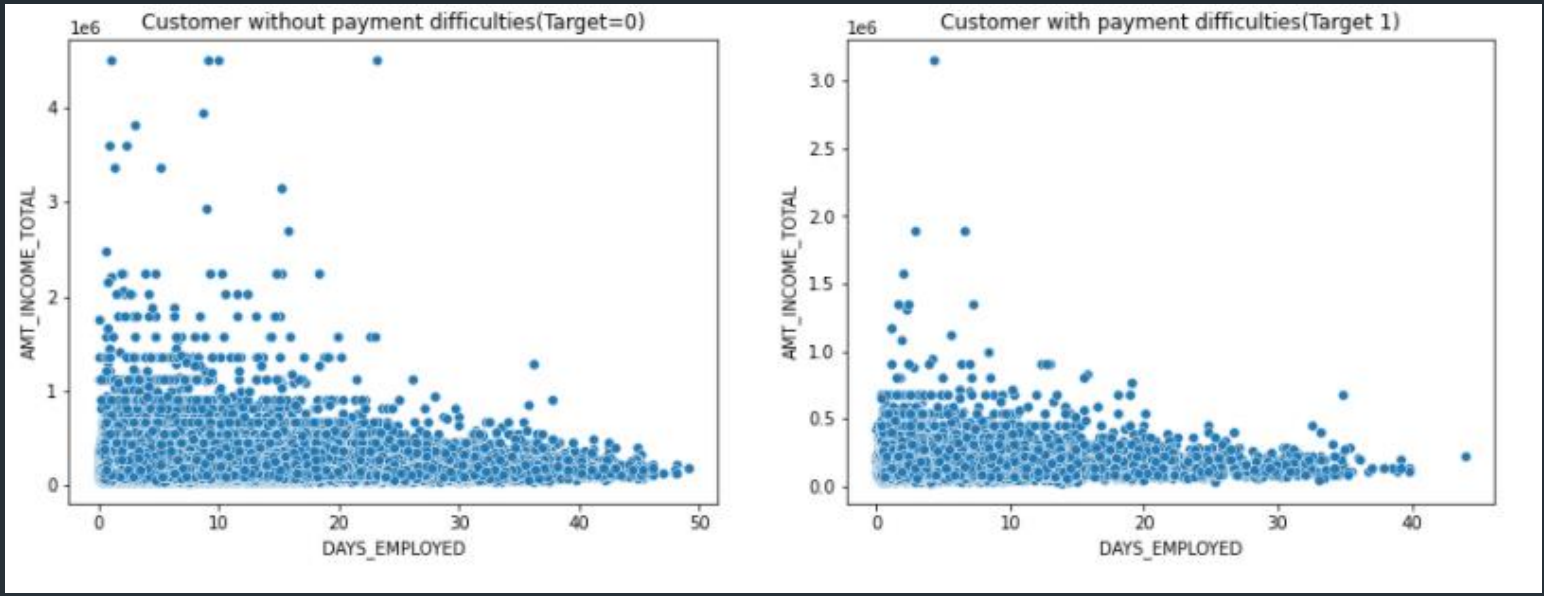


More people have taken loans at the start of their professional career than in the later stages of their career where they are more settled and hence less need of loans

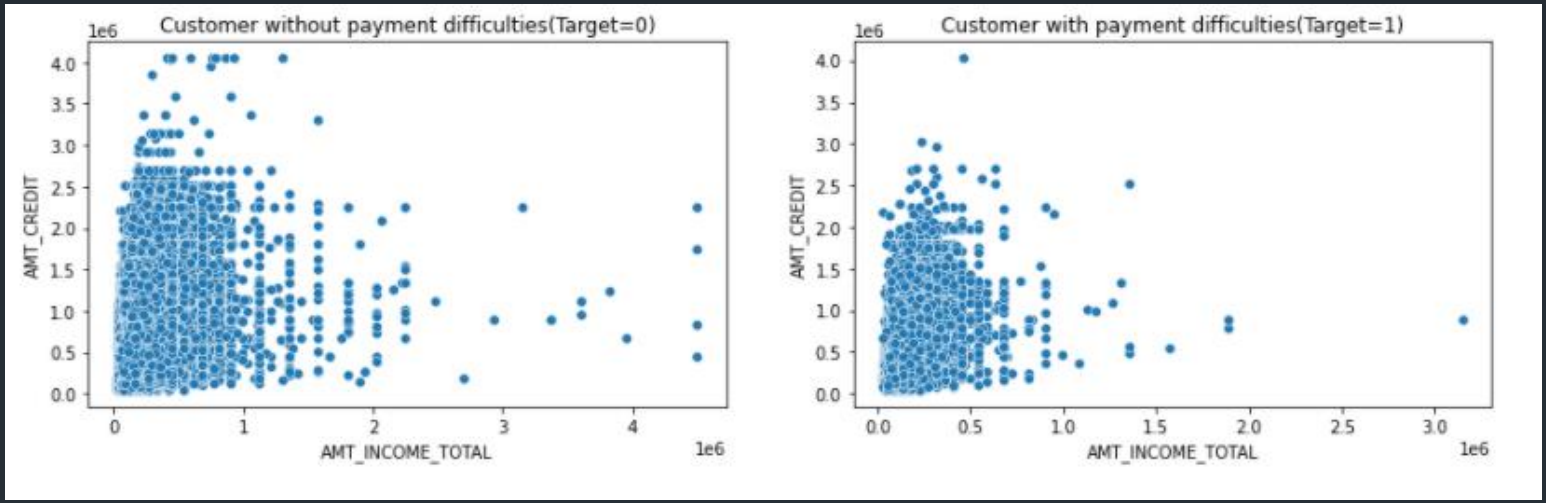


Credit amount taken by customers mostly lie in the range 45000-20,00,000
Though people have mostly preferred lower credit amounts.

BIVARIATE ANALYSIS FOR NUMERICAL VARIABLES



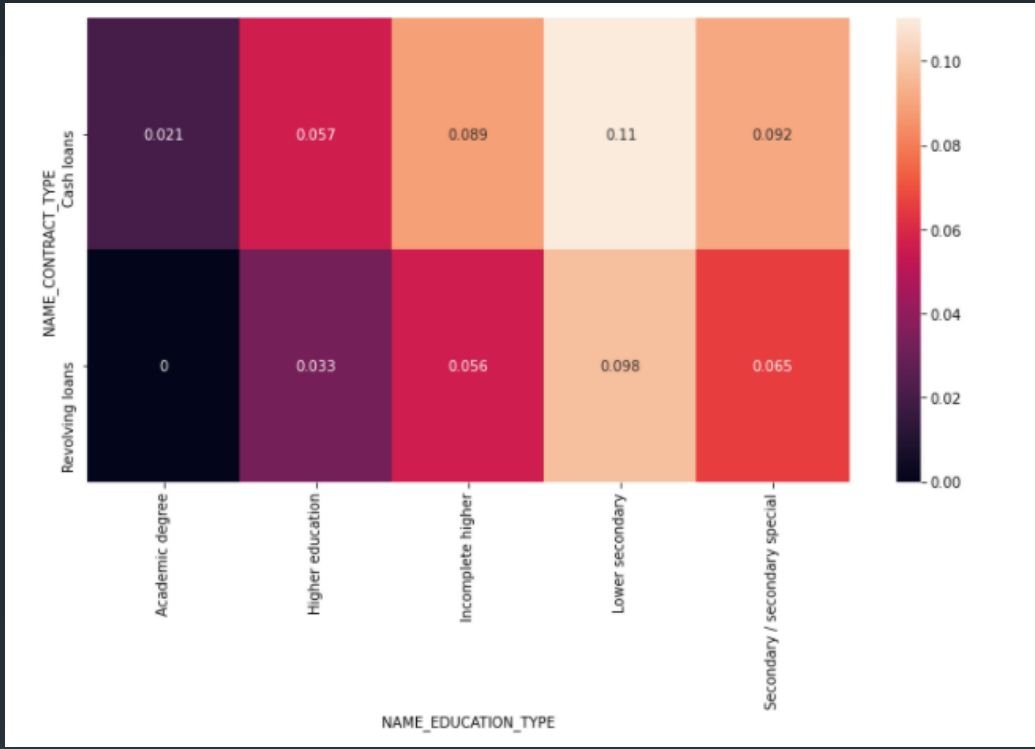
In general, Customers without having payment difficulties have more income than customers having payment difficulties.
For customers with more than 40 years of work experience, they usually don't have any difficulty in paying their loans.



Those who have greater income(2000000) are able to take loans and pay them on time(hence no data points observed in customer with payment difficulties graph)

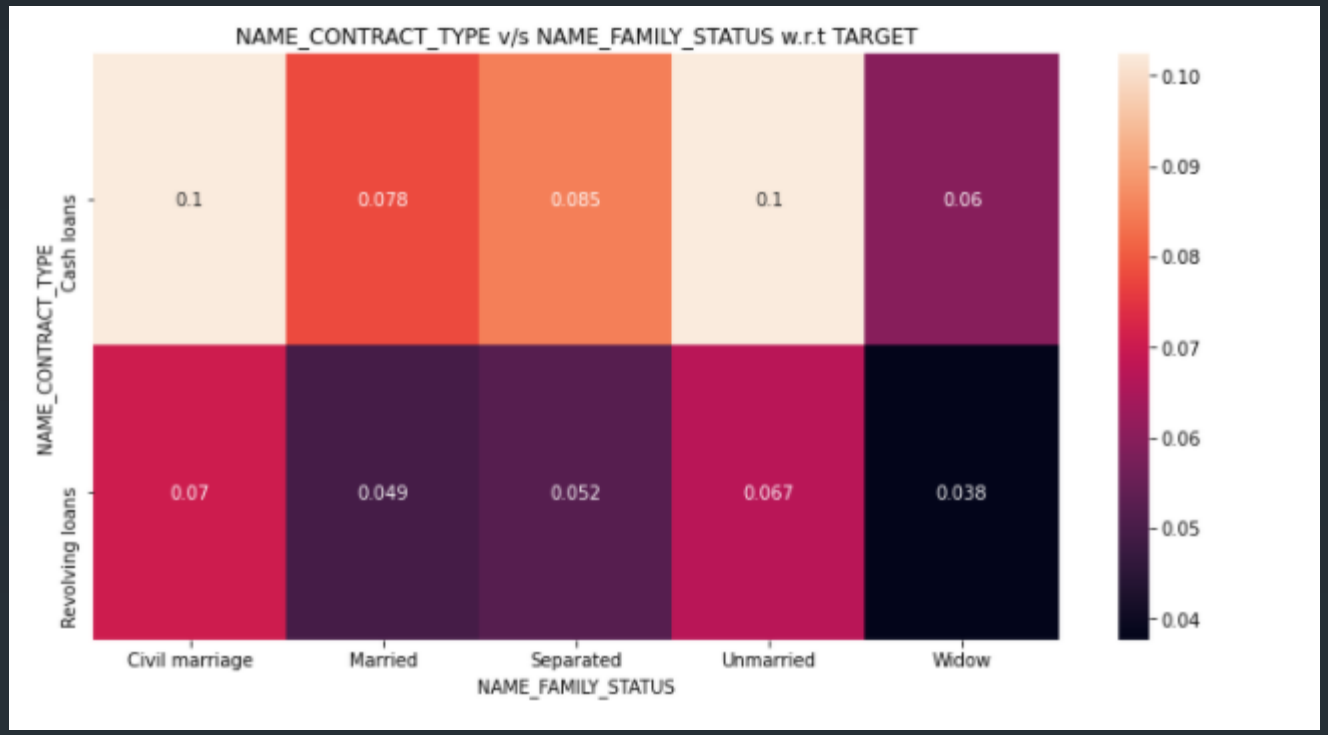
Those who have income on the lower side tend to take more loans

BIVARIATE ANALYSIS OF CATAGORICAL VARIABLES



From the plot above we can observe that clients with Lower secondary and cash loans have highest probability of defaulting while Academic degree holders and revolving loans have no defaulters

Academic degree holders have the lowest chance of defaulting across all education groups



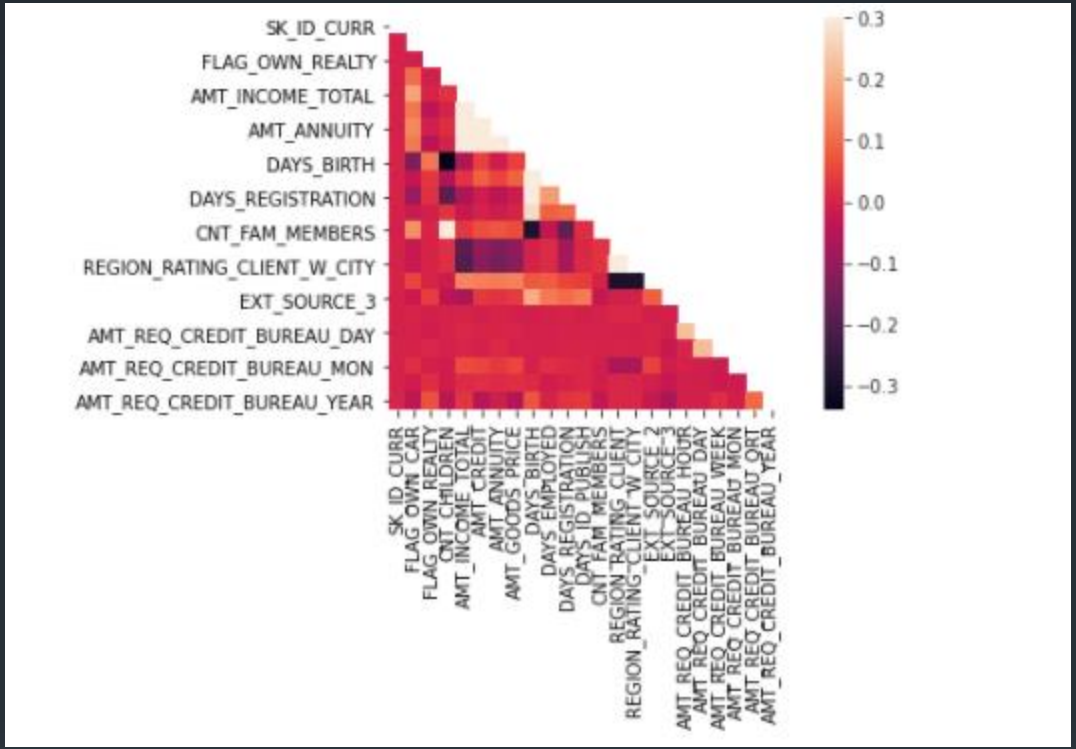
From the above plot, in cash loans, civil marriage and unmarried people have higher chance of defaulting, while in revolving loans, married and widowed clients have lowest chance of defaulting.
For widows, for both cash and revolving loans, they are able to pay on time.
For civil marriage and unmarried, for both cash and revolving loans, they have difficulty in payment of loans on time.

The customers living in office apartment and credit range 2500000 - 2750000 have the highest chance of defaulting

For people usually living in house/apartment, for all credit ranges, there are some defaulters.

For credit ranges 250000 - 1000000, there are more defaulters as compared to other credit ranges

CORRELATION MATRIX FOR TARGET=0

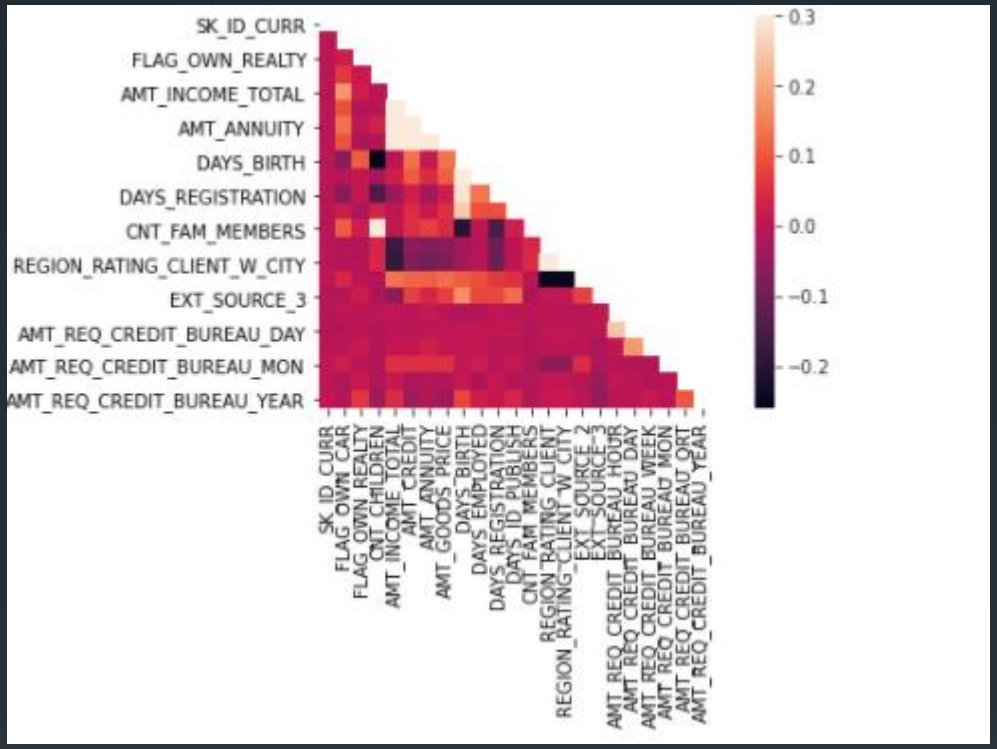


If we see most of the top 10 correlations are pretty obvious. For example

- AMT_GOOD_PRICE and AMT_CREDIT :- For consumer loans the credit amount is equal to the goods price and hence they will be almost equal
- CNT_FAM_MEMBERS and CNT_CHILDREN :- The family members include the count of children, hence a linear relationship between them
- AMT_ANNUITY and AMT_CREDIT :- Greater the amount of credit, greater will be the amount of annuity most of the times.
- DAYS_EMPLOYED and DAYS_BIRTH :- Greater the age, the work experience tends to increase.

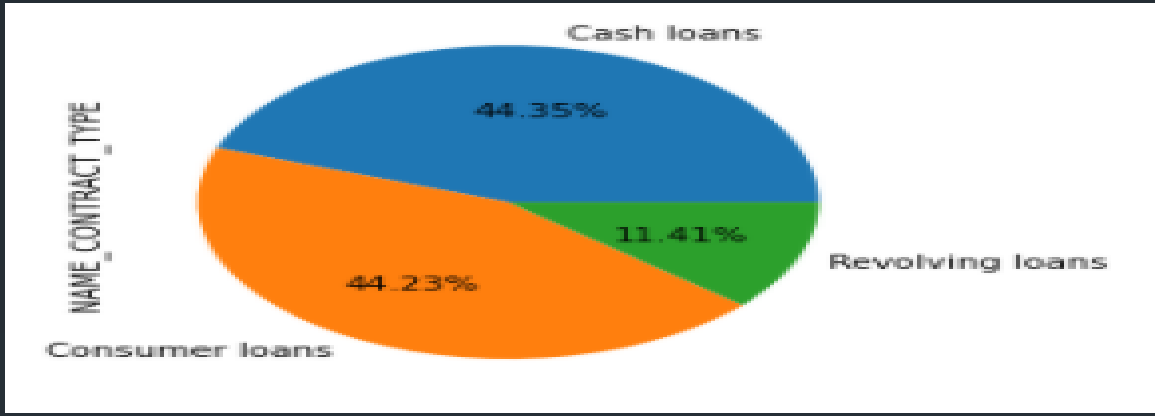
The correlations that matter will be the ones after these top 10 correlations where it isn't obvious.

CORRELATION MATRIX FOR TARGET=1

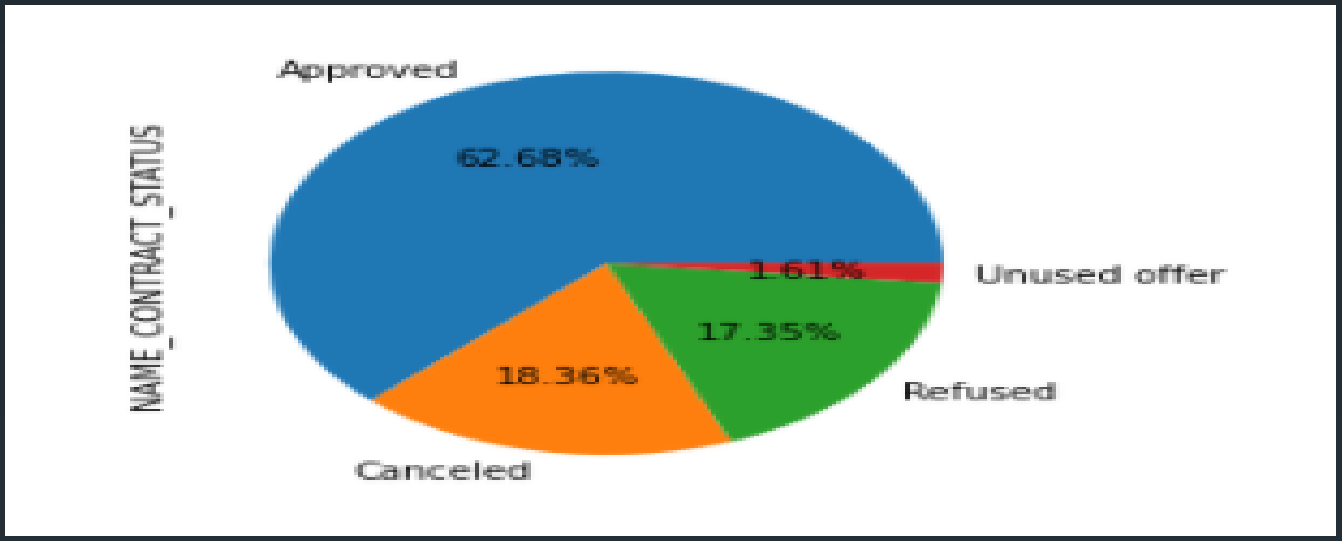


Similar kinds of insights can be drawn for target1 data as done in target0

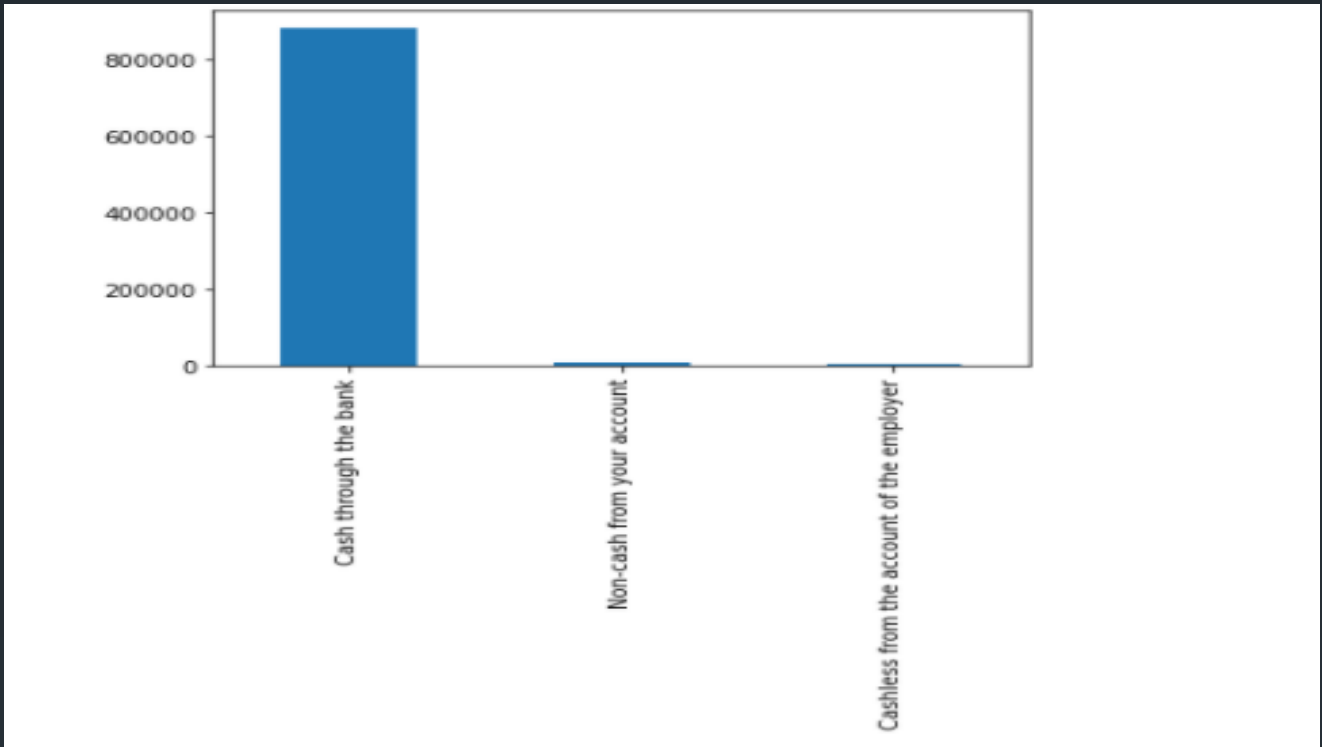
UNIVARIATE ANALYSIS OF MERGED DATA



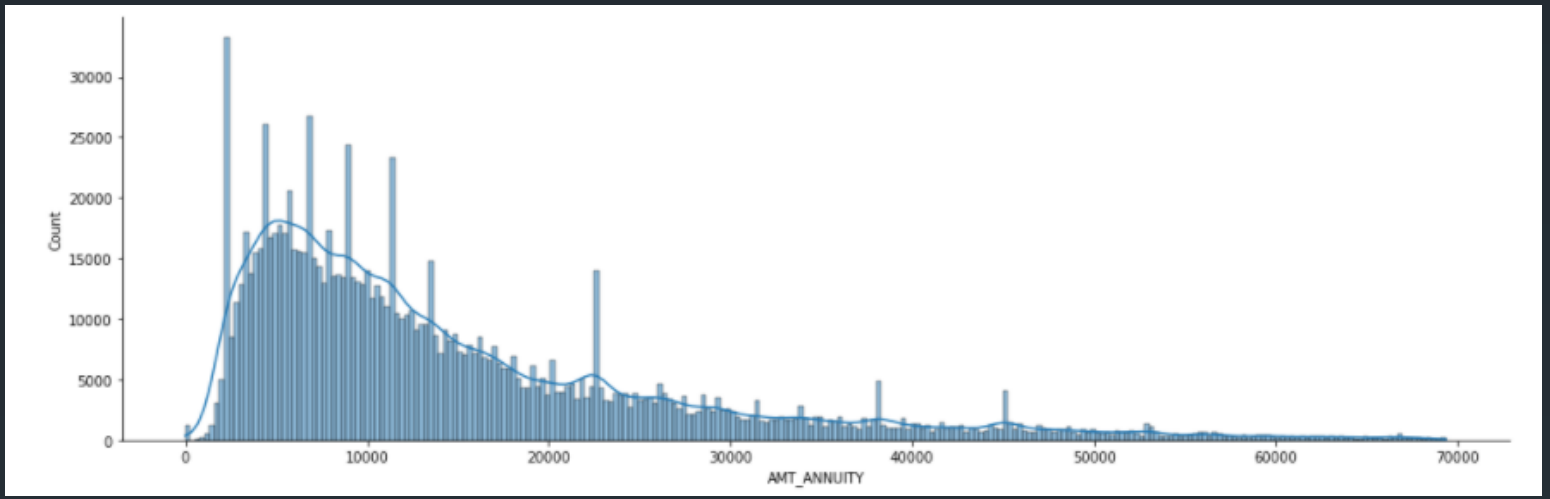
Cash and consumer loans are preferred over revolving loans in previous data as well



Majority of the loans have been approved but a significant portion belongs to Cancelled and Refused category

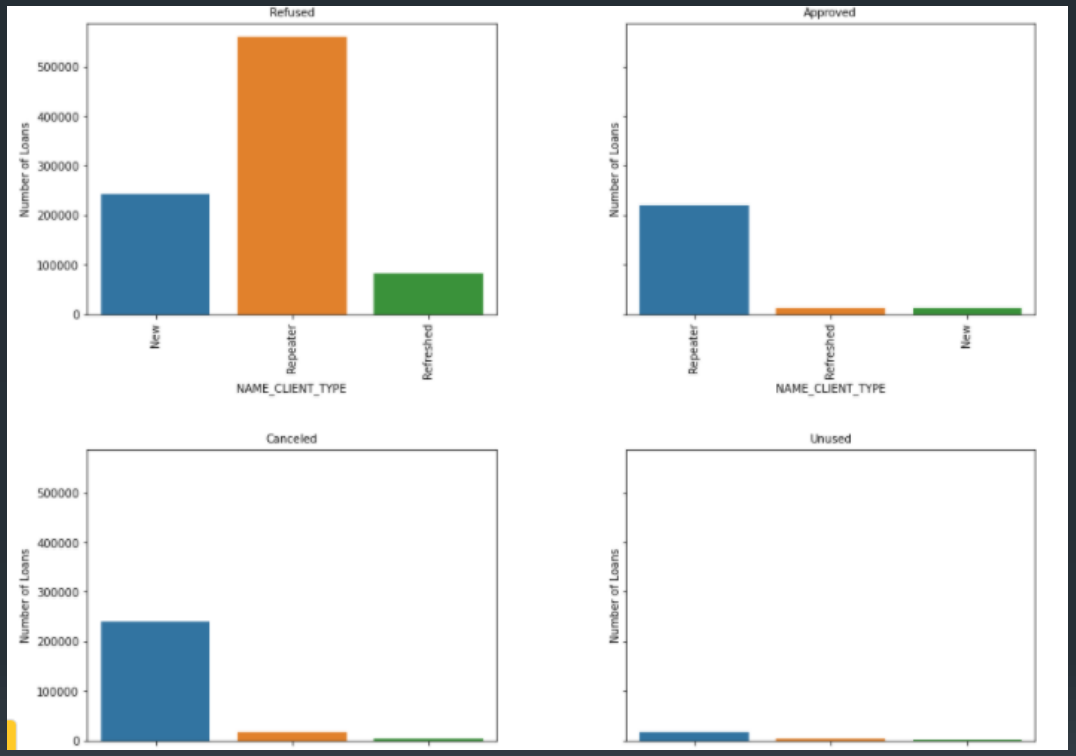


Cash through the bank is the most preferred method of paying loans

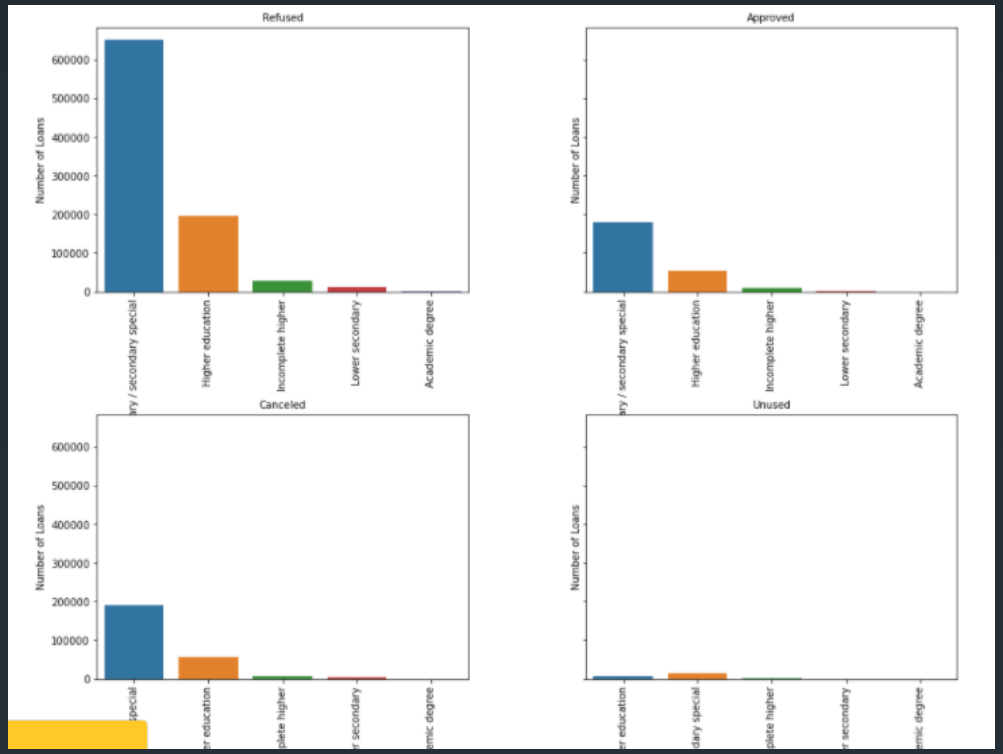


AMT_ANNUITY follows a skewed normal distribution
The AMT_ANNUITY is concentrated on the lower side of range 3400-60000

BIVARIATE ANALYSIS OF MERGED DATA



Most of the Repeaters have their loans being refused.
Very few New applications were approved and most of them were refused



For Secondary / secondary special people, majority of their loan applications were refused.

CONCLUSION

- Females usually have taken more loans and pay on time as compared to males. Hence we can promote some special type of loans or lower the interest rates for them.
- Those who have Academic degree have higher chance on paying the loans on time. Other education type groups can be scrutinized more carefully. For e.g. for Secondary / secondary special people, majority of their loan applications were refused.
- Labourers usually take more loans but should be given lower credit amount as they can then have a higher chance of defaulting. While job types like Managers or accountant which are higher in hierarchy can be given higher credit amount.
- People with less than 2 children are likely to pay on time. People with more number of children hence should be given loans with lower credit amount or introduce some special type of loans for them to handle such cases.
- Generally people in higher income ranges have lower chance of defaulting than people in lower income ranges. We can promote more loans to people with higher income ranges.