PROBLEM STATEMENT

- **Understanding the Problem:** The task involves analyzing loan application data to identify patterns indicating repayment difficulties. If loans are given to borrowers who default, it leads to financial losses for the company. Conversely, rejecting loans to deserving applicants results in a loss of potential business opportunities.

- In Exploratory Data Analysis (EDA), I'll use simple techniques like looking at numbers, making charts, and checking how different things are connected.
- The main thing I'll look at is whether a person had trouble paying back their loan or not. If they did, it's marked as '1', and if they didn't, it's marked as '0'. After I finish looking at the data, I want to find out what things make it more likely for someone to have trouble paying back their loan. This will help us make better decisions about who to give loans to and who to be careful with.

- *Dataset for analysis –*

- *'application_data.csv'* contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.
- *'previous_application.csv'* contains information about the client's previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer

Approach & Methodology:

Step 1: Understand the problem statement and data
Step 2: Started with analyzing the 'application_data.csv' file:

- Check if there are any missing values – Handle missing values
  - Removed columns with more than 40% missing values to ensure data reliability.
  - Normalized the 'OCCUPATION_TYPE' column to ensure consistency and imputed null values with "Unknown" to maintain data integrity, considering the presence of a large number of missing values that could distort the mode.
  - Imputed null values in the 'NAME_TYPE_SUITE' column with the mode since it is a categorical variable.
  - Droped missing value records where missing value percentage is very low i.e. less than 0.1% - 'AMT_ANNUITY' , 'DAYS_LAST_PHONE_CHANGE' , 'AMT_GOODS_PRICE' and CNT_FAM_MEMBERS

- Select columns which seem useful for analysis and drop the remaining columns – got 36 columns for further analysis
- Standardizing the dataset - changing data types, Converting days to absolute values, calculating age in years, calculating income in lakhs and creating buckets.
- Identify if there are outliers in the dataset – check the mean and 50%, plot boxplot.
- Check the data imbalance for the ‘Target variable’ in the dataset
- Perform univariate and bivariate analysis.

Step 3: Similarly, analyzing the 'previous_application.csv' file.
Step 4: Merge both the dataframes and analyze the merged dataframe.

Some examples of the Standardization done in the datasets:

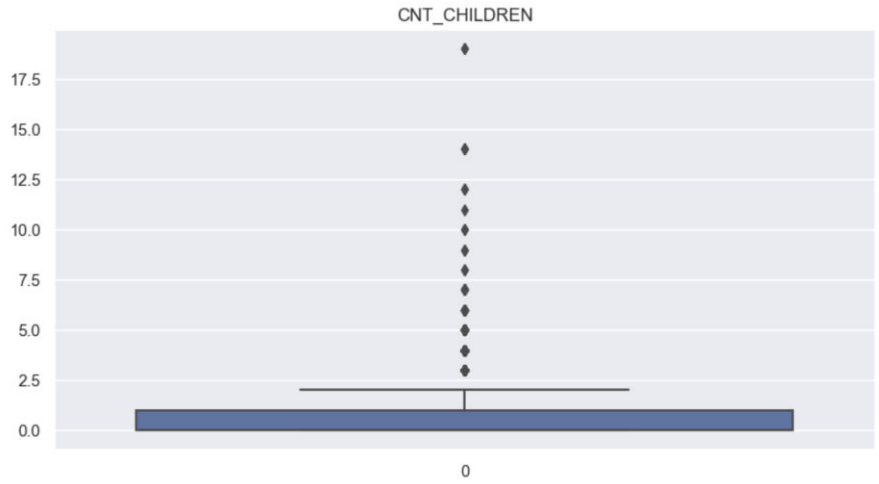
1. The columns"DAYS_BIRTH," "DAYS_EMPLOYED," "DAYS_REGISTRATION," "DAYS_ID_PUBLISH," and "DAYS_LAST_PHONE_CHANGE" were converted to absolute values. This transformation was applied to ensure that negative values representing the number of days relative to a certain event are represented as positive values, making them more intuitive to interpret and analyze.

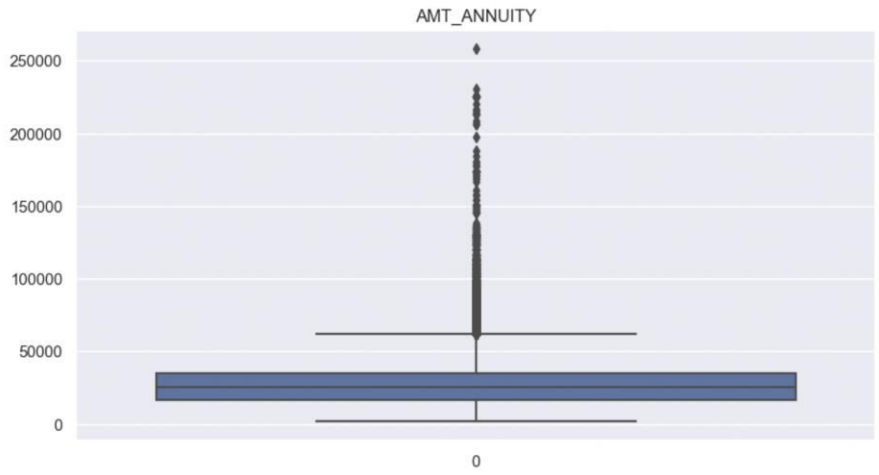
2. Converting "EMPLOYMENT_YEARS" and "AGE_IN_YEARS" from days to years simplifies the data, as days are a smaller unit of time, leading to larger numbers that can be harder to interpret. By converting to years

3. The "AGE_IN_YEARS" & EMPLOYMENT_YEARS_RANGE column was transformed into age buckets for analysis, categorized as "0-20," "20-25," "25-30," "30-35," "35-40," "40-45," "45-50," "50-55," "55-60," "60-65," and "above 65." This grouping simplifies the age data, making it easier to analyze and interpret trends

4. The columns "AMT_INCOME_TOTAL" and "AMT_CREDIT" were transformed into lakhs and grouped into bins: "0-5L," "5-10L," "10-15L," "15-20L," "20-25L," "25-30L," "30-35L," "35-40L," and "Above 40L." This conversion simplifies the amounts, making it easier to compare and analyze income and credit levels across different ranges.

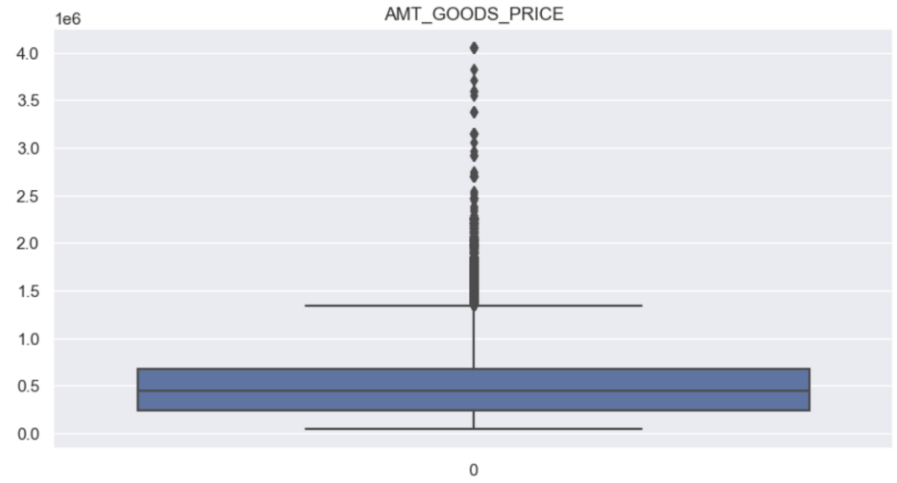
Identify Outliers

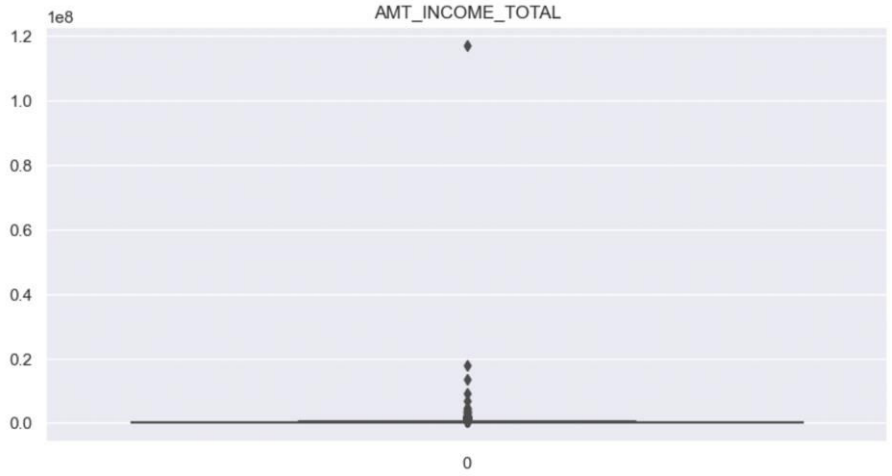




- The graphs show presence of outliers in CNT_CHILDREN AND AMT_ANNUITY.
- Outliers are marked as individual points beyond the whiskers.

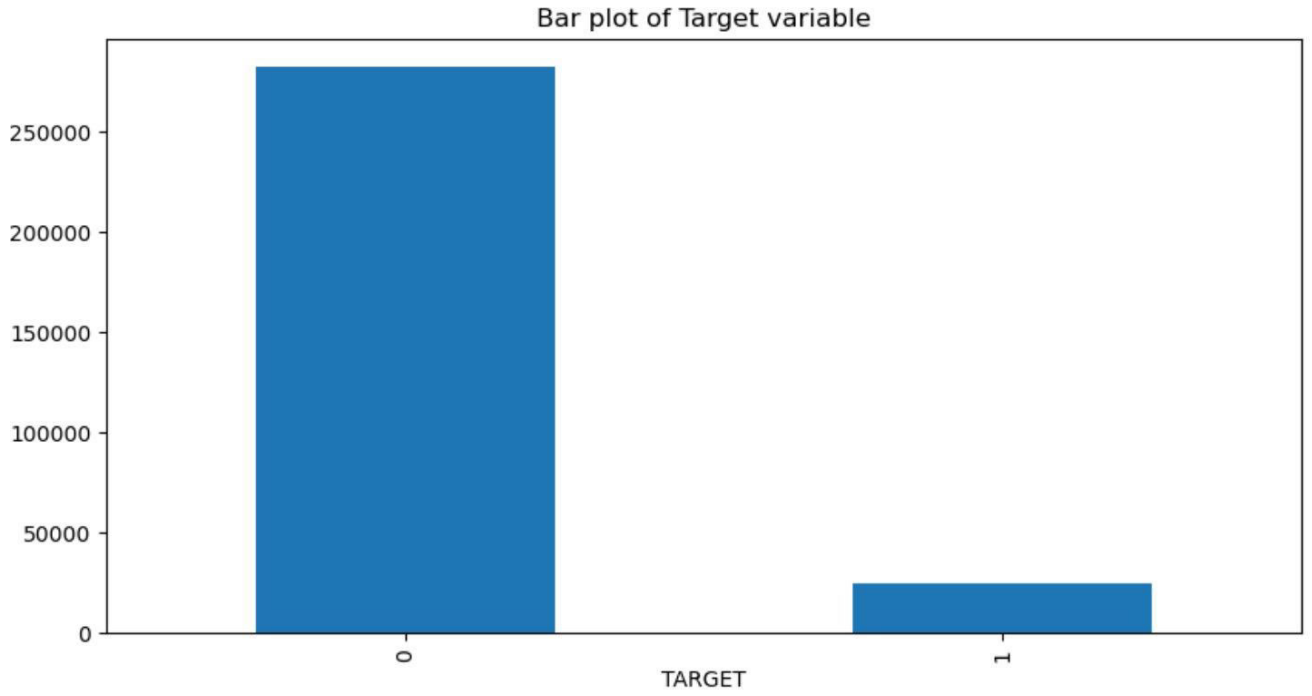
Identify Outliers





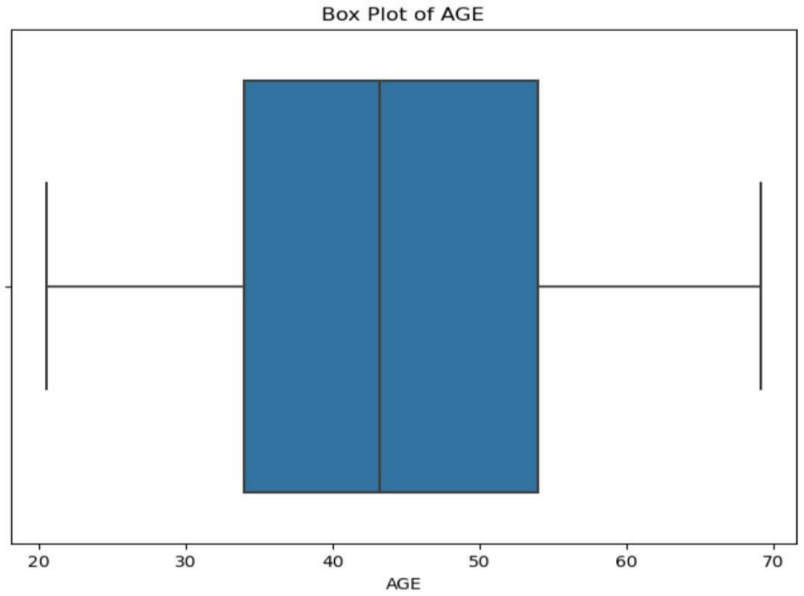
- The graphs show presence of outliers in AMT_GOODS_PRICE AND AMT_INCOME_TOTAL.
- Outliers are marked as individual points beyond the whiskers.

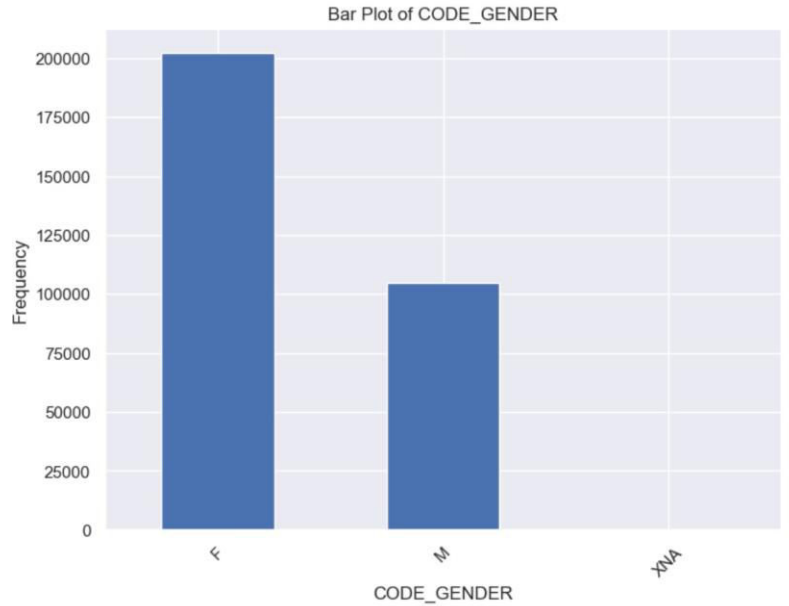
Data imbalance



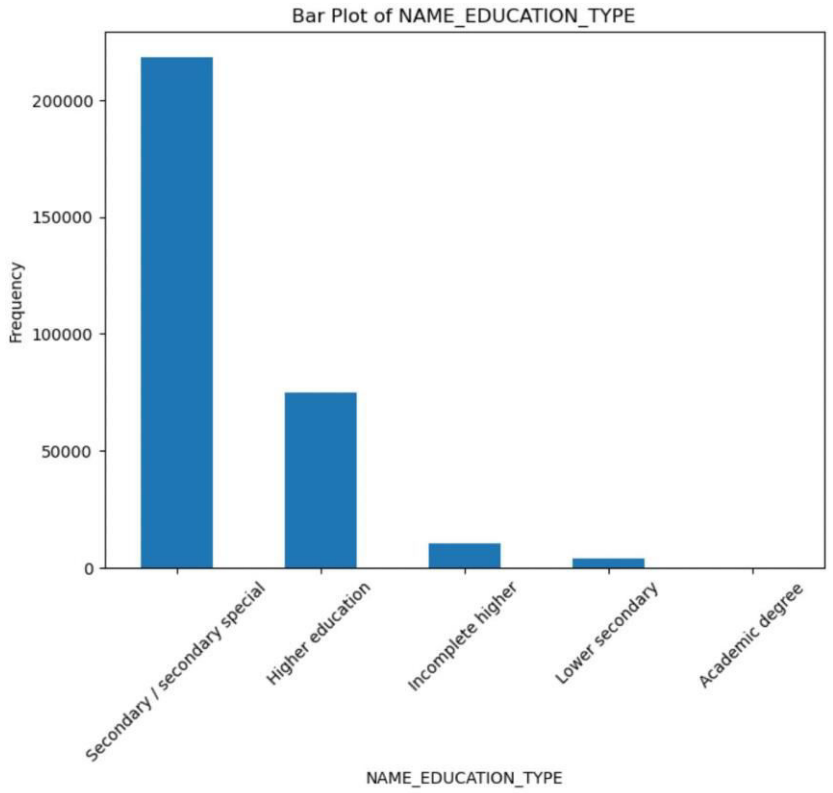
- The graph shows high data imbalance as the number of people having payment difficulties are more than in total population.

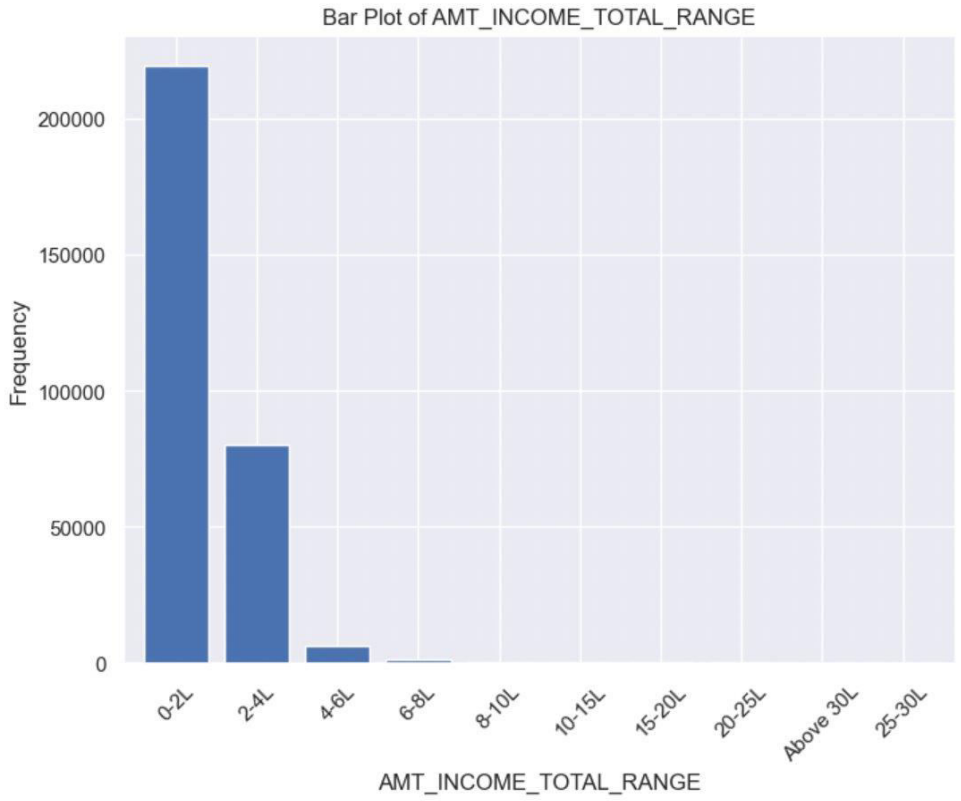
Univariate analysis





- The graphs show that the age group of the population who have applied for loans lie between 30-60.
- The number of females who have applied for loans are more than the number of males



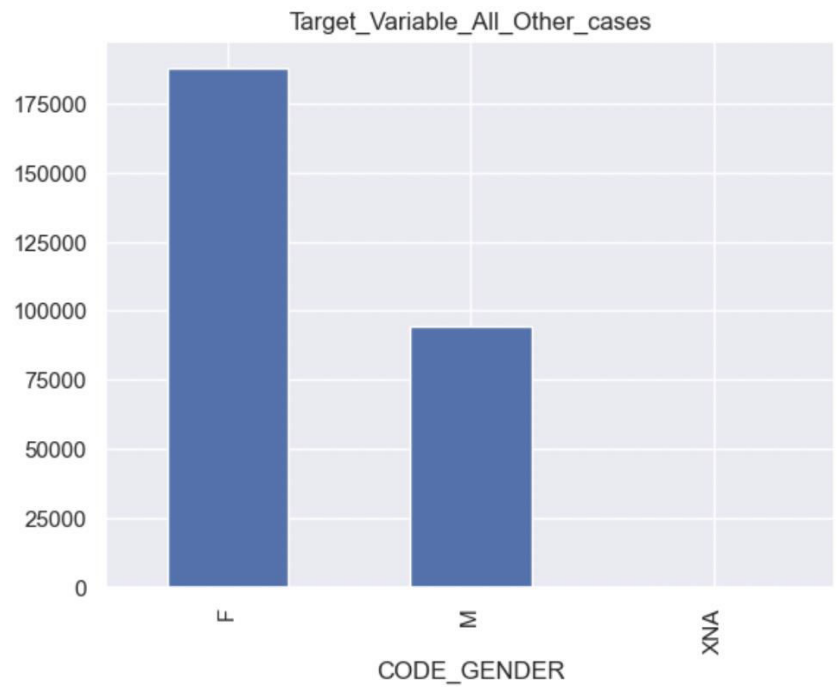


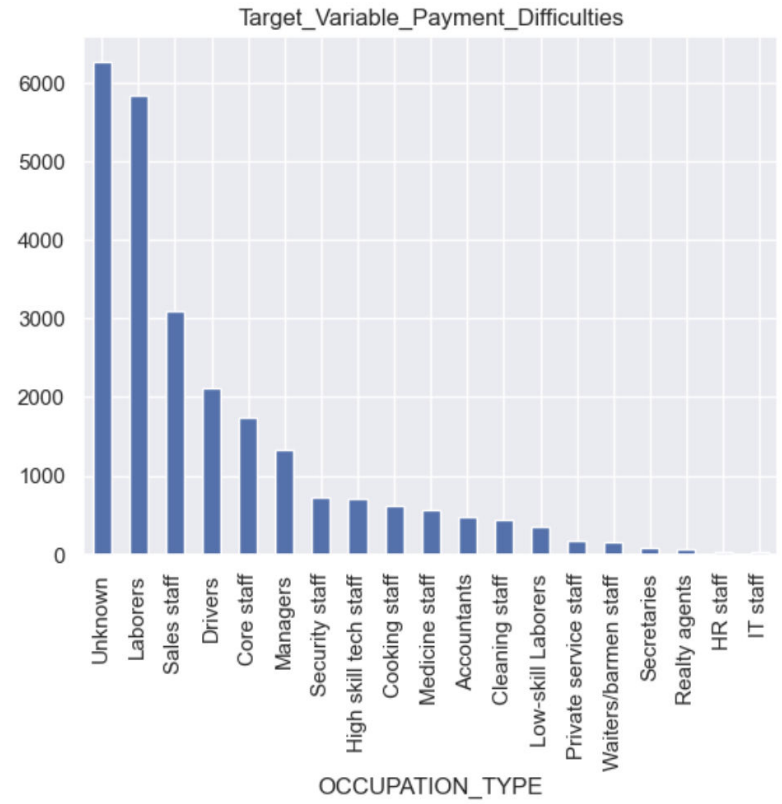
- The graph show that the people having education level as secondary have applied for loans majorly.
- The graph show that the people with low income have applied for loans majorly.
- This indicates people with lower income levels are more likely to seek loans, possibly due to insufficient funds for larger expenditure.

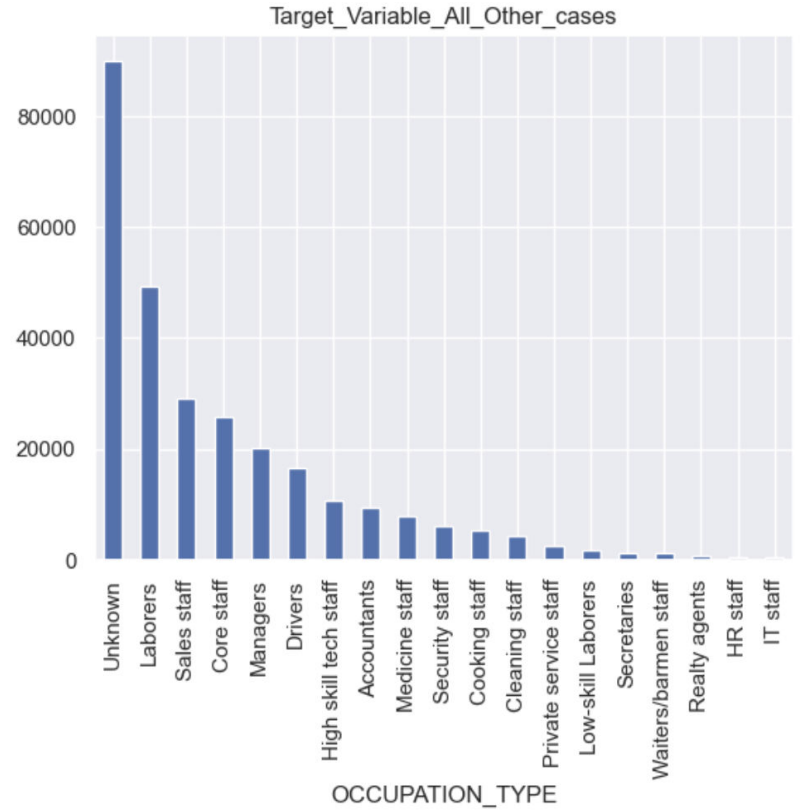
Analysis of variables with respect to target variable:

- Formed two different dataframes where one is when target variable is 1 - clients with payment difficulties and second is when target variable is 0 - all other cases

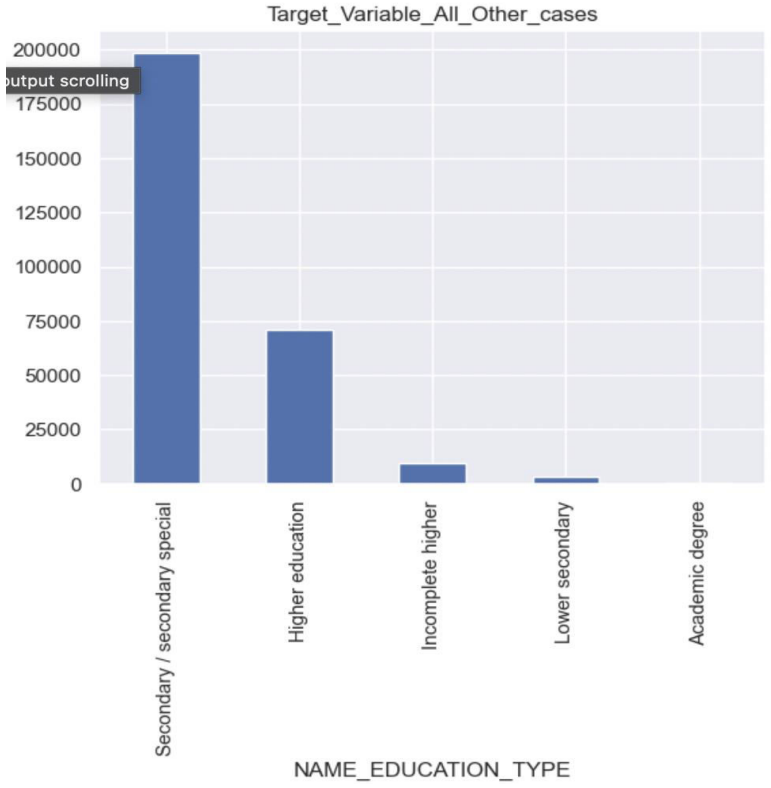
- OBSERVATIONS:

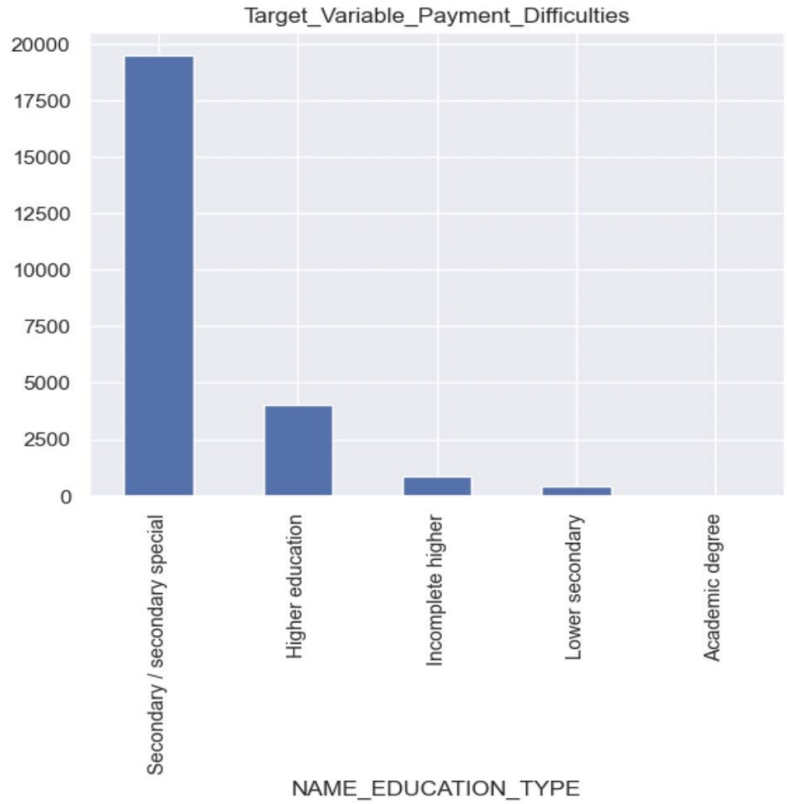






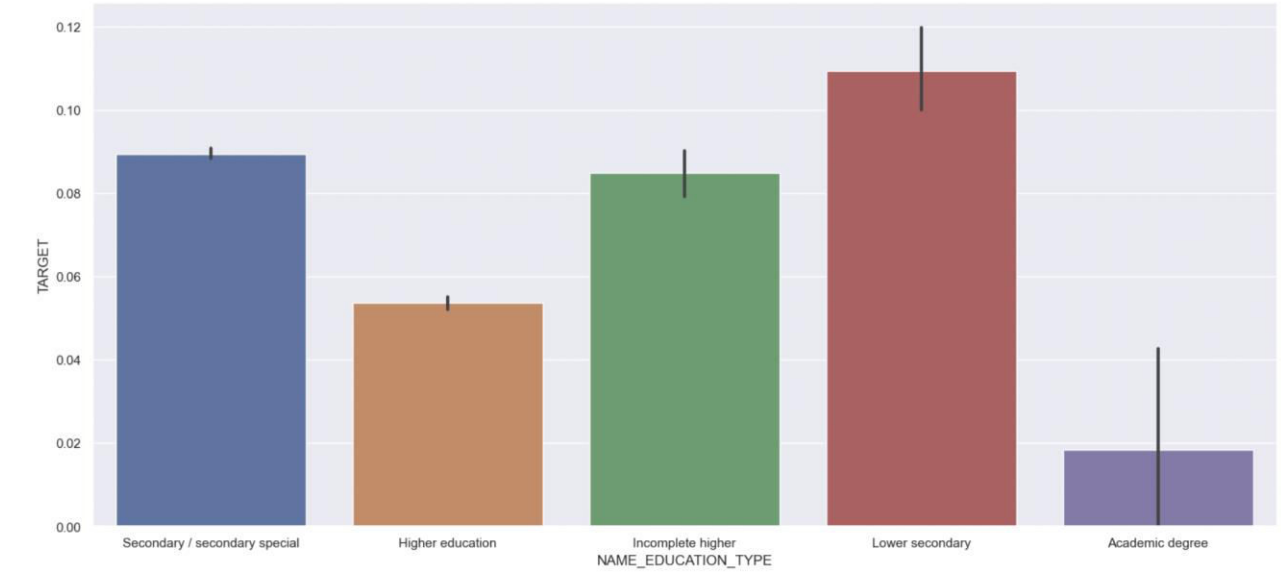
- The graphs show that laborers, sales staff and drivers are the groups facing the most significant challenges in repaying loans, this suggests that these occupations might have less financial stability or lower incomes compared to other professions in dataset.

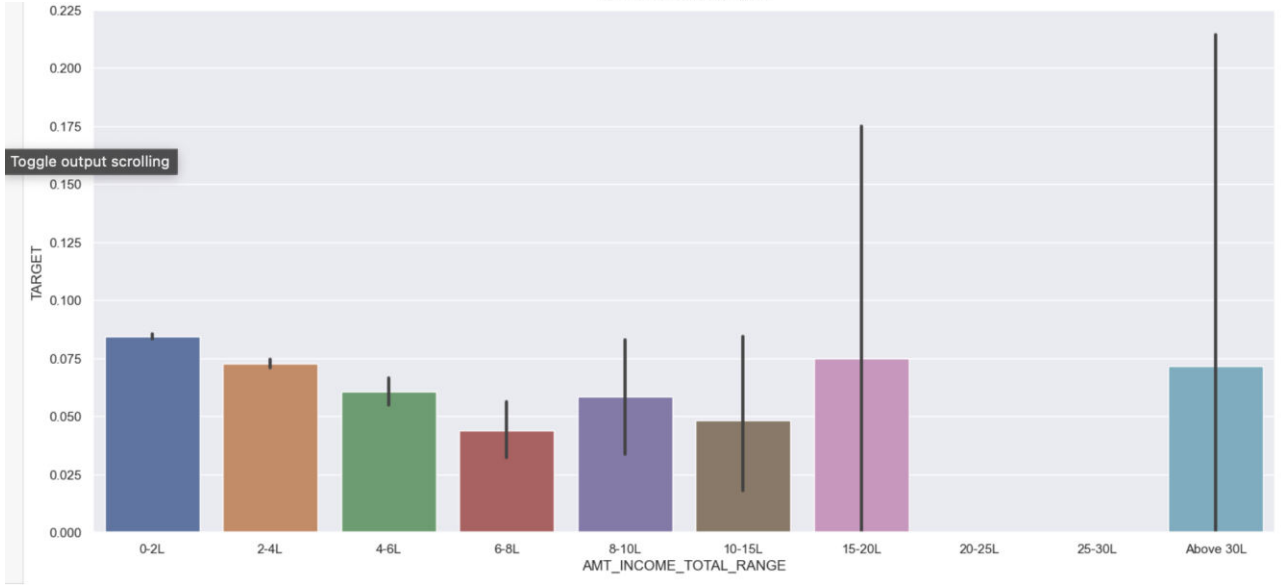




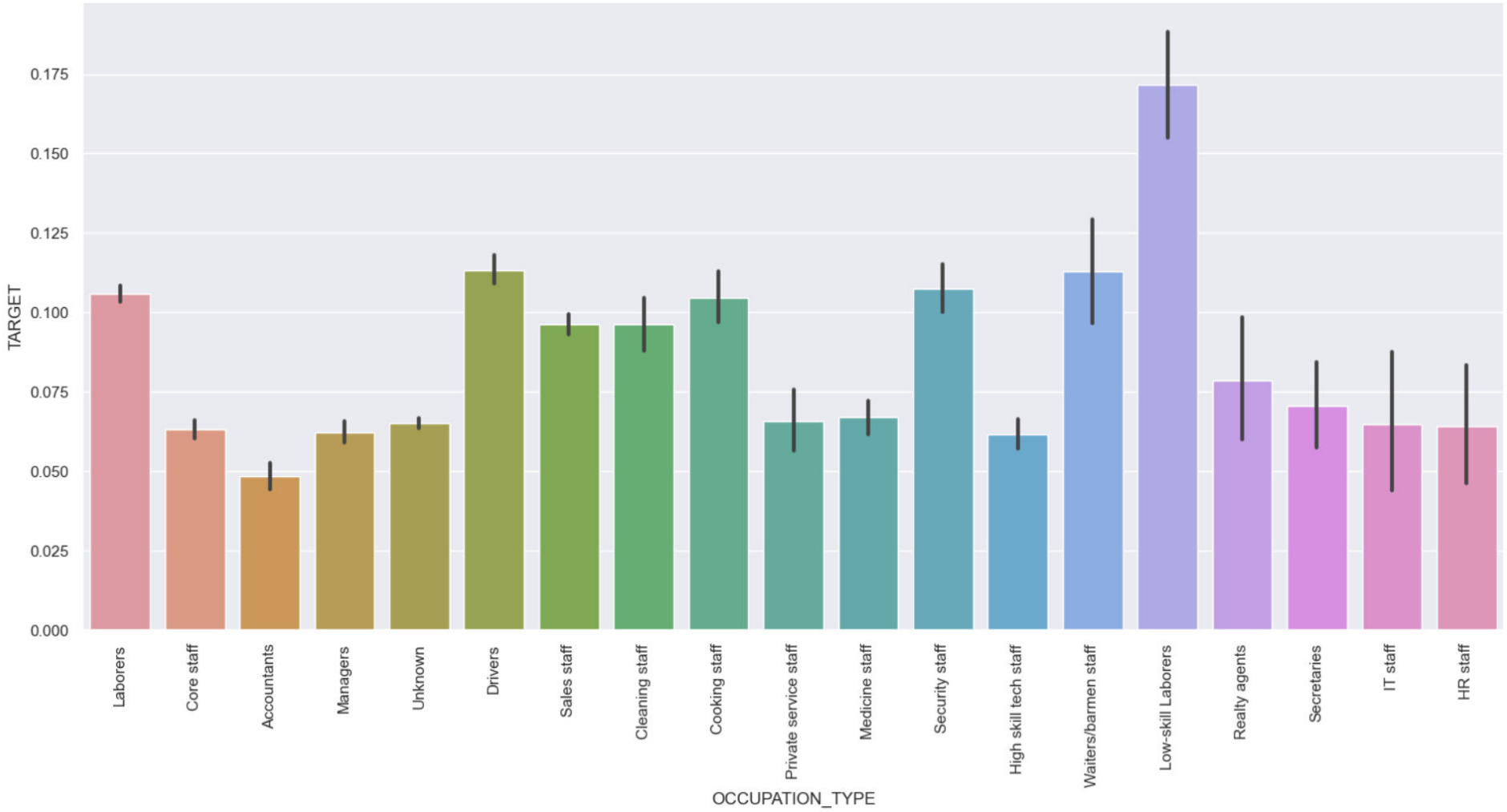
- The graphs show that the people with education level as secondary are facing difficulties in payments.
- This suggests that people with this education level may have lower income or less stable employment leading to financial difficulties.

Bivariate analysis





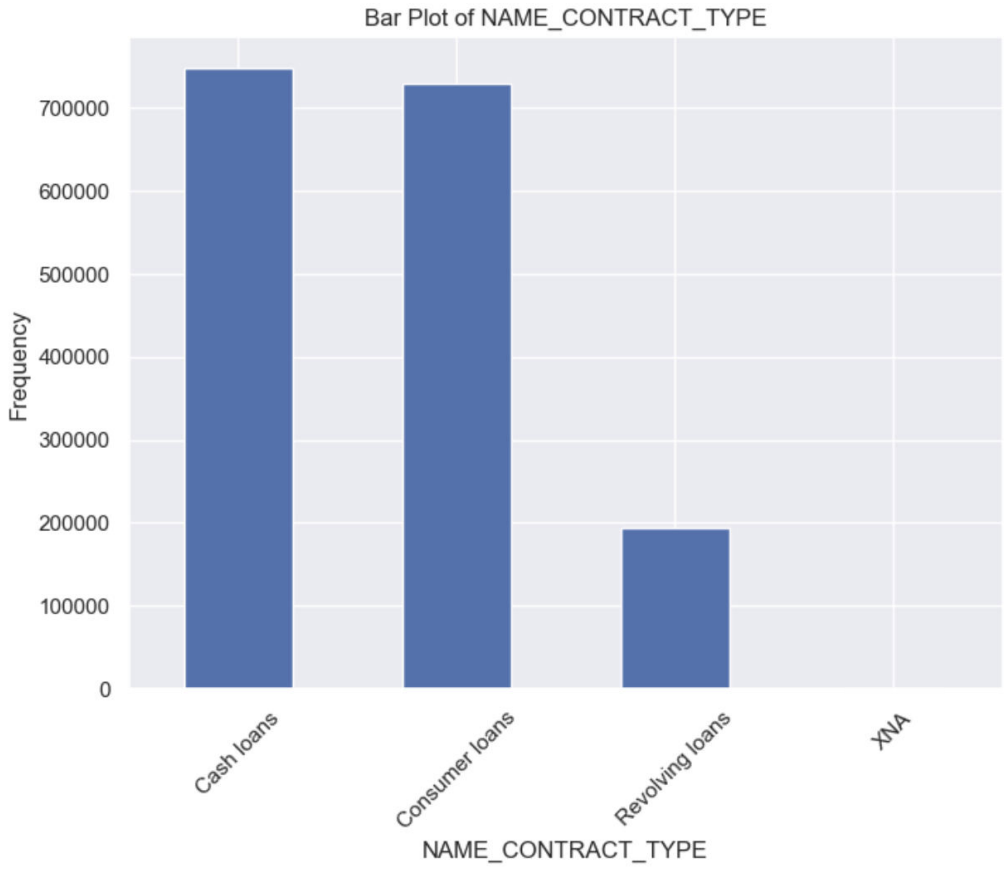
- The graph show that the people with education level as lower secondary are facing difficulties in payments.
- People having low income are facing difficulties in payments.
- It also show that people having high income are also facing difficulty in repaying loans.



Observation :

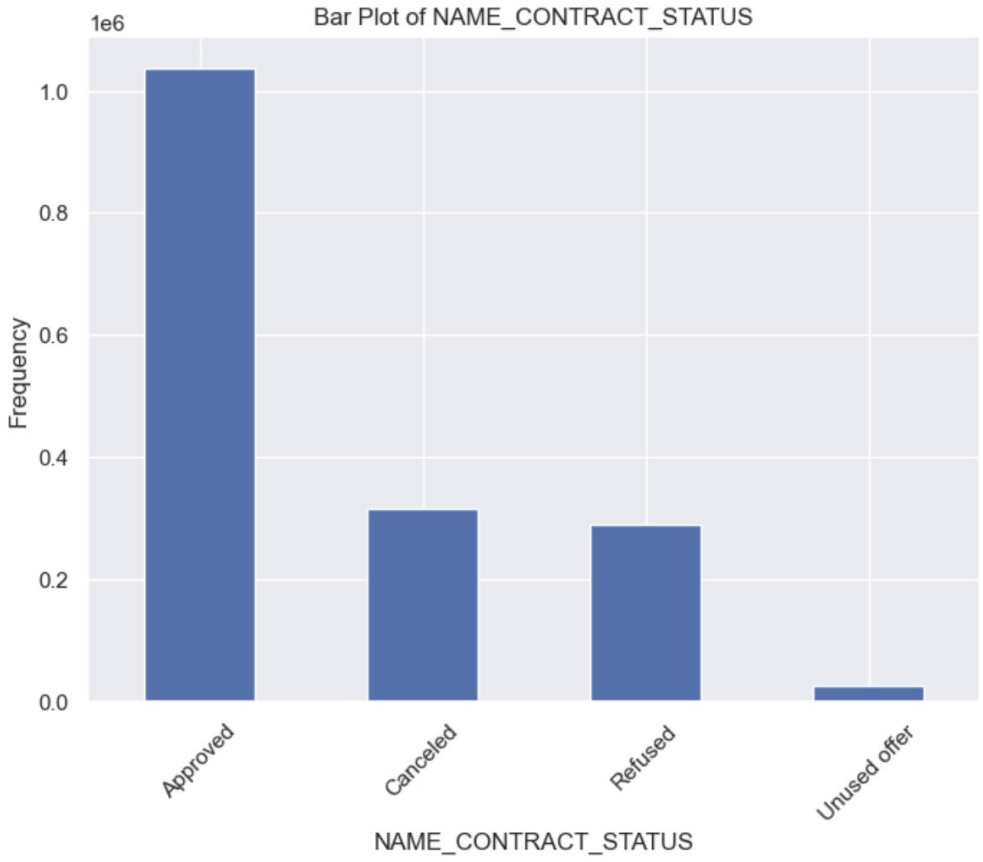
People with occupation type as low-skill laborers are facing difficulty in repaying loans.

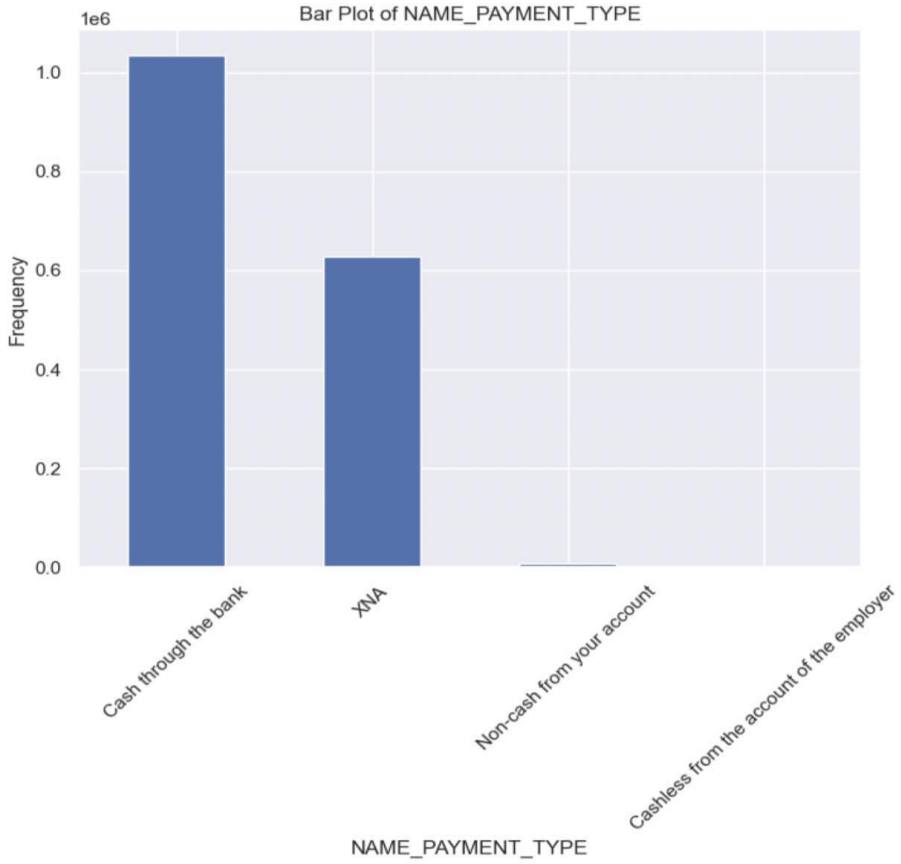
Analysis of variables in p*revious_application.csv' file:*



Observation :-

- The graph shows that people having applied for more cash loans.

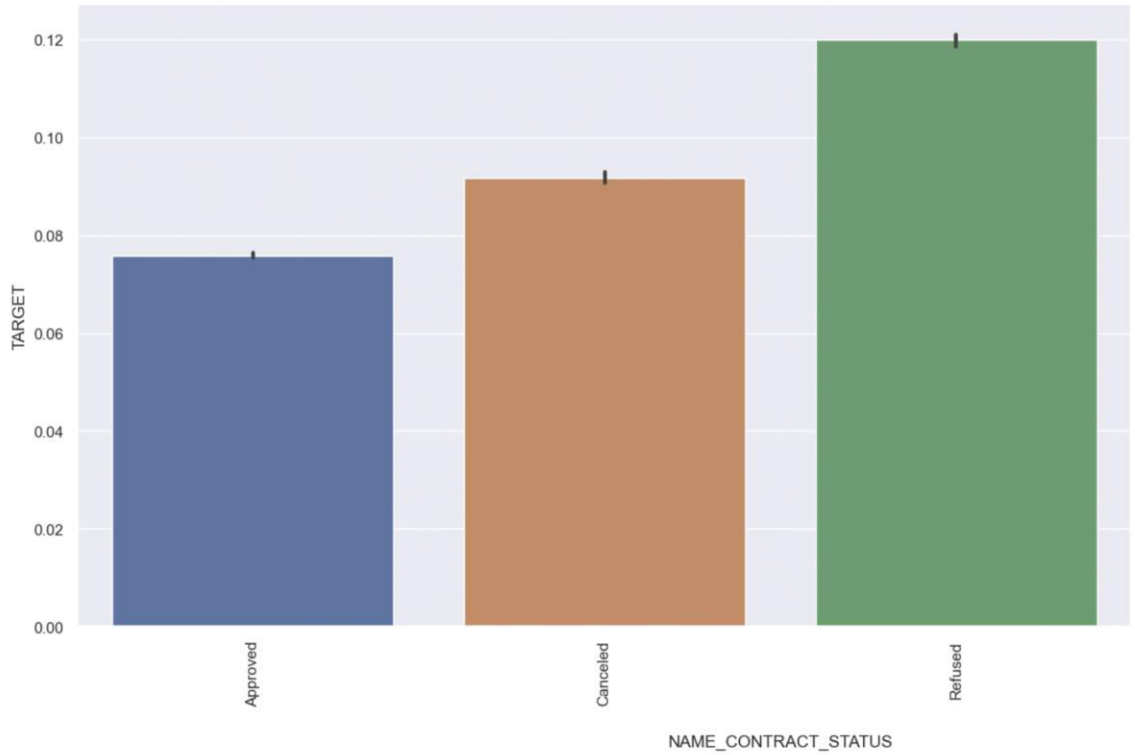




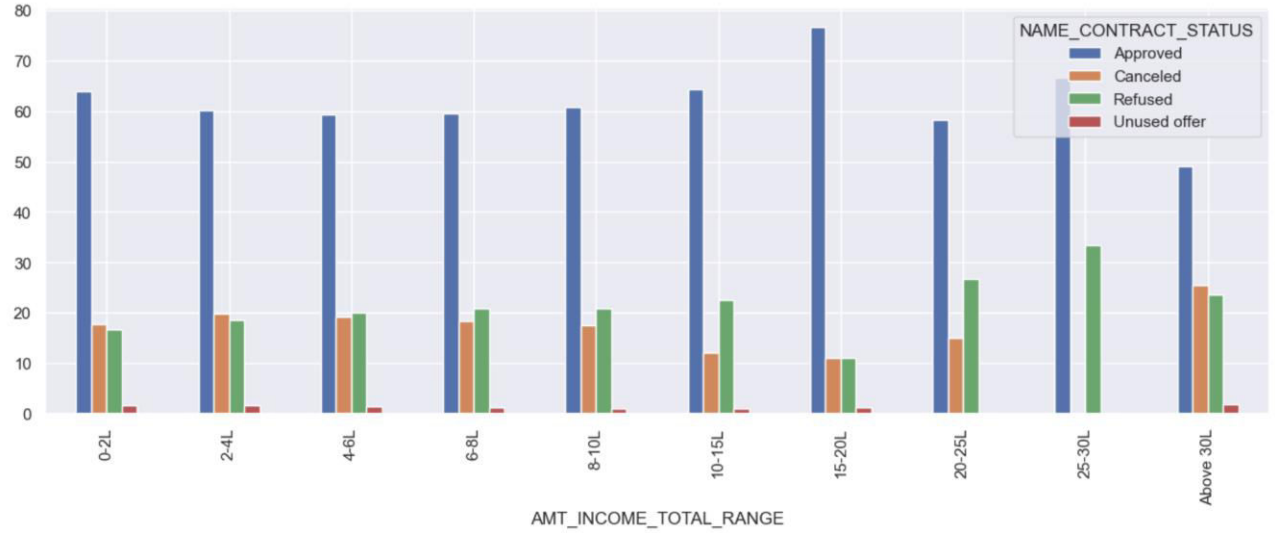
Observation :-

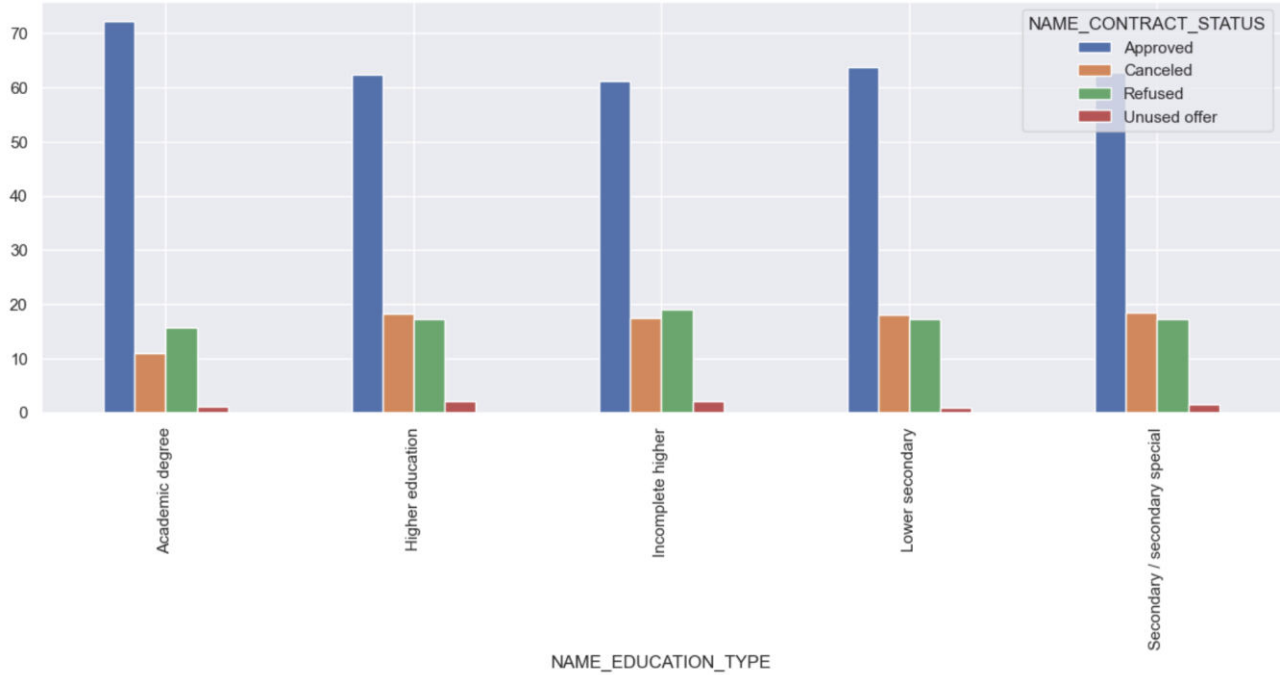
- The majority of loan applications were approved.
- The graph shows that people have preferred to pay cash through the bank

Merged Data:



The graph shows that the applicants which were refused previously are likely to face challenge in paying loans.

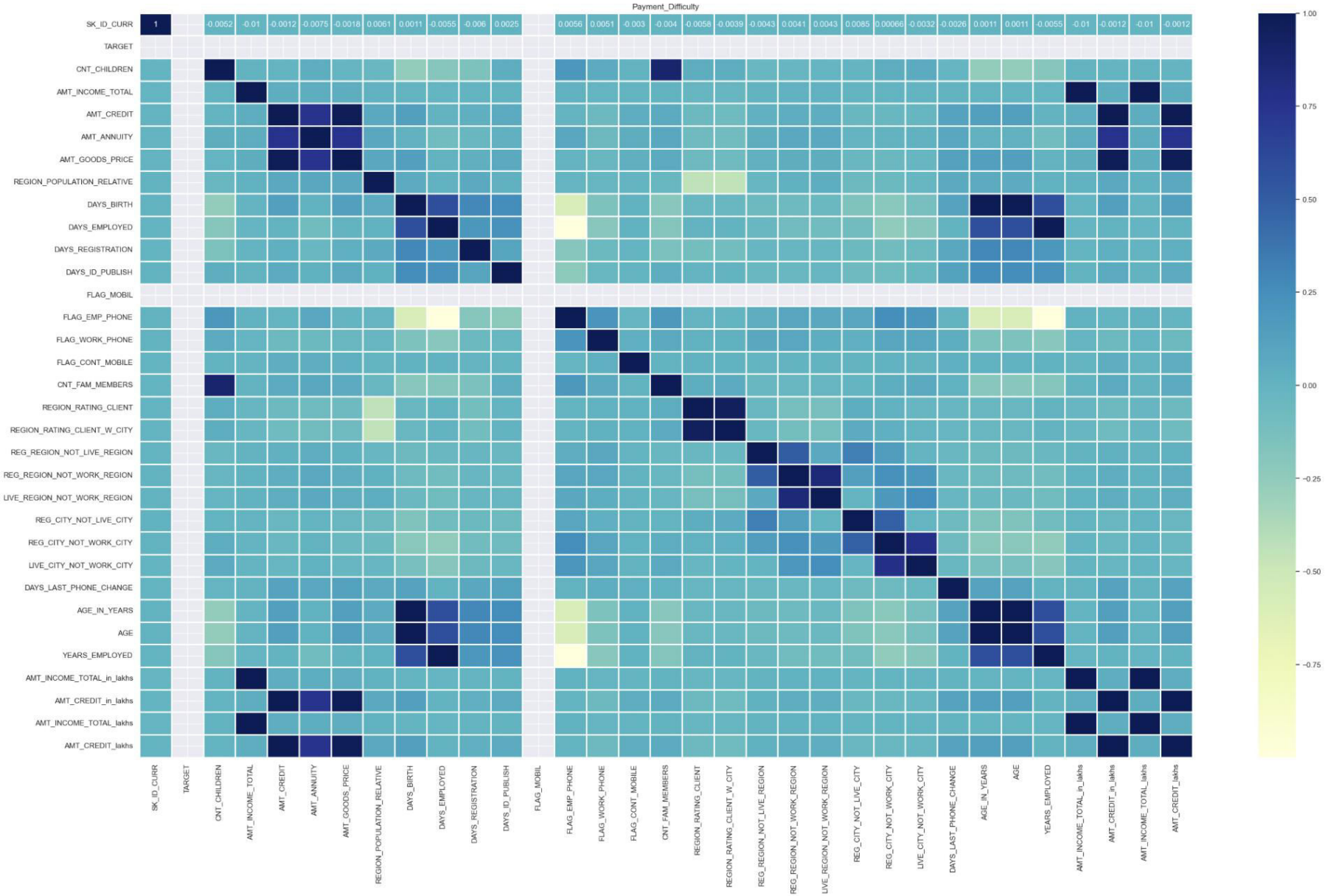




- Applicants having income range between 15-20 Lakhs have a higher rate of approved loan applications compared to other income brackets.
- This graph also highlights unexpected trends where applicants having incomes exceeding 30 lakhs have more cancelled loan applications compared to other income brackets
- The applicants having academic degree have a higher rate of approved and cancelled loan applications compared to other education types.

Correlation

- Credit amount is
- Credit Ratio has



Observations from heatmap:

- Credit amount has strong correlation with Goods Price Amount and Annuity Amount
- Goods Price Amount is highly correlated with Annuity Amount
- DAYS_EMPLOYED is highly correlated with DAYS_BIRTH
- CNT_FAM_MEMBERS is highly correlated with CNT_CHILDREN

Conclusion:

- Good clients - people with higher education and people with higher income
- Bad clients – people with education type as secondary and having low income
- Laborers, sales staff and drivers are the groups facing the most significant challenges in repaying loans, this suggests that these occupations might have less financial stability or lower incomes compared to other professions in dataset.
- The people with education level as secondary are facing difficulties in payments which suggests that people with this education level may have lower income or less stable employment leading to financial difficulties.
- Men are at relatively higher default rate
- Applicants from age group 25-40 are more likely to face difficulty in re-paying loans.