Credit EDA Assignment

PRAGYA BHARGAVA

DSC 45

Problem Statement

Let us consider a scenario where you work for a consumer finance company that specializes in providing urban customers with several kinds of loans. To analyze the patterns found in the data, you must perform EDA. By doing this, it will be ensured that only those applicants who can repay the loan would be accepted.

In order to take steps like denying the loan, lowering the loan amount, lending an amount to relatively risky applicants at a higher interest rate, etc., it is possible to uncover patterns that show whether a client has trouble paying their installments. By doing this, it will be ensured that only borrowers who can repay the loan will be accepted.

Overall approach

- Understanding the Problem Statement and the Data Sets Provided
- Importing Relevant Warnings and Libraries
- Data Loading
- Basic Metadata Check (shape, info, dtypes, describe)
- Data Cleaning
- Handling Outliers
- Analysis
- Univariate Analysis
- Segmented Univariate Analysis
- Bivariate Analysis
- Draw top 10 correlation heat maps and gather insights from it

Understanding the Problem Statement and the Data Sets Provided

We are provided with two .csv files in form of data sets namely – application_data and previous_applications.

- ‘application_data.csv’ : contains all the information of the client at the time of application.
- '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.

Along with the data sets, another file has been provided which acts as a dictionary for all the column present in both the data sets ('columns_description.csv’)

By looking at the data set and the problem statement carefully, we have found out that the most important variable on which we have base our analysis on is the Target Variable

Target variable is a flag type and determines if a client will pay the loan on time or not.

- 1 :The client with payment difficulties
- 0 : All other cases.

Importing Relevant Warnings and Libraries

Warnings:

```
import warnings
warnings.filterwarnings("ignore")
```

Libraries:

Mathematical Analysis:

- Pandas as pd
- Numpy as np

Visualization

- Matplotlib as plt
- Seaborn as sns

Data Loading and Metadata Check

In this step, we imported files and ran a quick data check to get a quick idea of how much data we're dealing with

Data check can be run through syntaxes like shape(), info, dtypes, describe() etc.

```
inp0= pd.read_csv (r'C:\Users\pragya.bhargava\Downloads\application_data.csv')
inp1= pd.read_csv (r'C:\Users\pragya.bhargava\Downloads\previous_application.csv')
```

Data Cleaning

Finding missing percentage in each column and dropping them based on the predetermined threshold

- First, we determined columns which were irrelevant to the data set and then removed it
- Then, we found the percentage null values in each column. We treated those columns first which have more than 50% null values

Removing missing values and replacing them with appropriate data

Standardizing data

Data Cleaning

Finding missing percentage in each column and dropping them based on the predetermined threshold

Removing missing values and replacing them with appropriate data

- For columns that have numerical values, we will use the describe syntax and visualize the values using a boxplot to find out any outliers. Then, the missing cell in the data will be replaced by the median of the values in that column if there is a clear outlier in the data. Otherwise, it will be replaced by the mean of all values.
- For columns that have object values, if the percentage of missing values is relatively significant, we will be adding another category for missing values. Else, based on the relevance and the meaning of the missing value, the missing value will be dropped and be imputed based on the relevant data categories in a column.

Standardizing data

Data Cleaning

Finding missing percentage in each column and dropping them based on the predetermined threshold

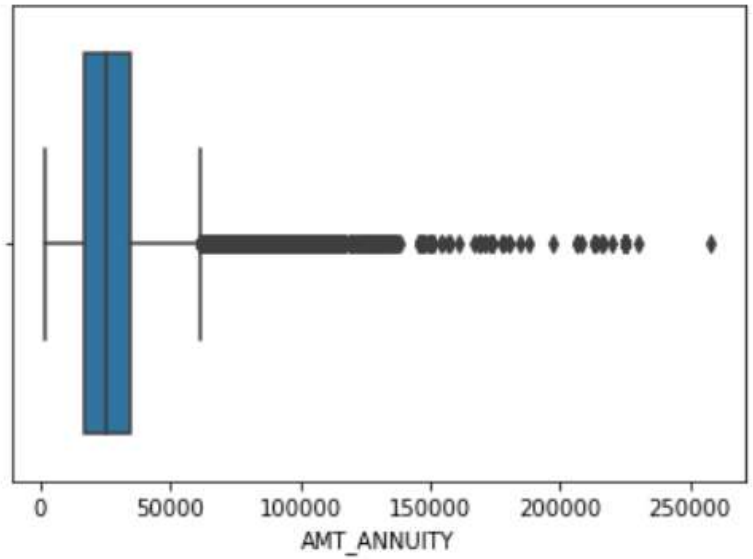
Removing missing values and replacing them with appropriate data

Standardizing data

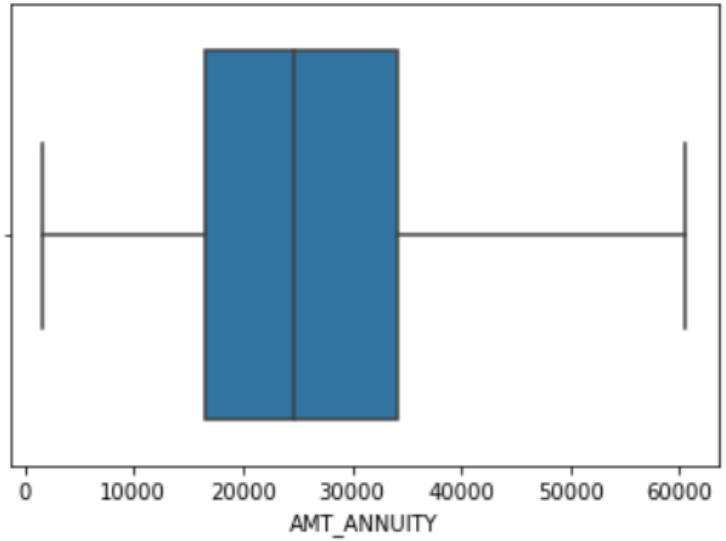
- Fixed:
  - Columns which had data in days and converted them to years
  - Changed data type of the columns to Object

Handling Outliers

- Found out Outliers using the describe syntax and then visualized it using boxplot.
- Treated Outliers via capping and flooring



Before treatment



After treatment

Analysis

Following are the steps followed for analysis:

Analysis of Application_data

- Segmented analysis : by diving the Target column values into two- defaulter for 1 and non-defaulter for 0
- Identification of Continuous and Categorical Variables
- Univariate Analysis of Continuous Variables – using line chart
- Univariate Analysis of Categorical Variables – using bar chart
- Bivariate Analysis – using scatter plots

Analysis of previous_application

- Combined both previous application and cleaned application_data data sets using the common column SK_ID_CURR and performed analysis to see how the previous applications of clients affect the decision making process of the bank

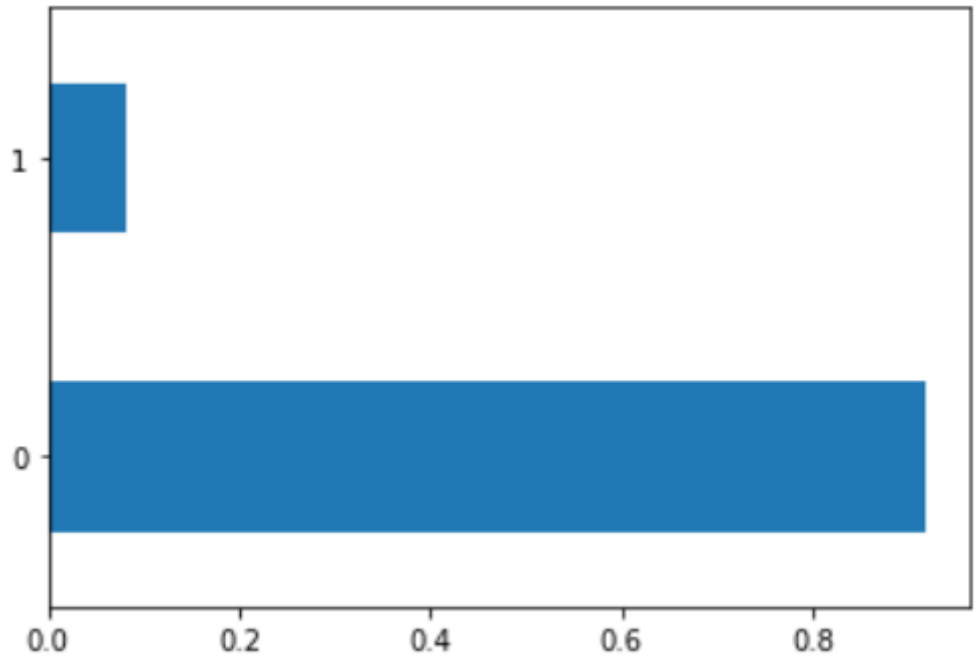
Performed Correlation

- Found Top 10 Correlations in the data

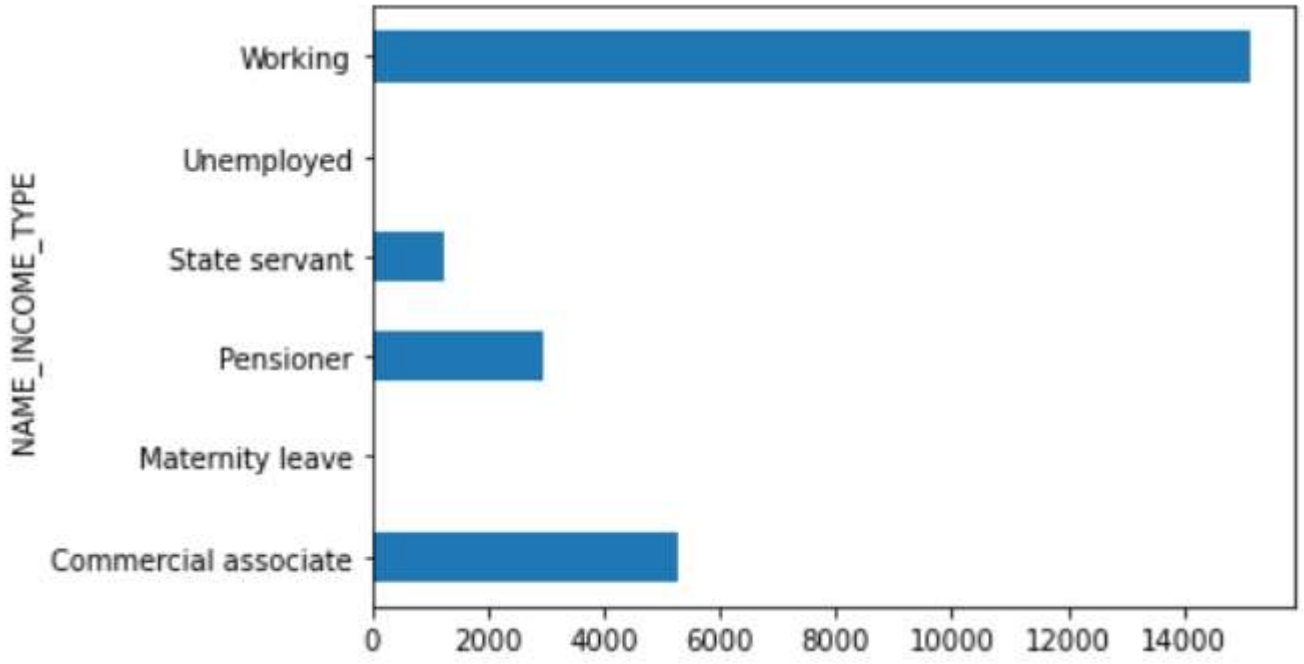
Characteristics of a defaulting client

- Lies between the age bracket 30-40
- Either has work experience of 0-5 years or approx. 35 years
- Are mostly married
- Live in a self-owned house/apartment
- Most belong to Region 2
- Most belong to Region with City rated 2
- Are mostly labourers by profession
- Most client default when they take cash loans
- Majority of the defaulters are females

Data imbalance

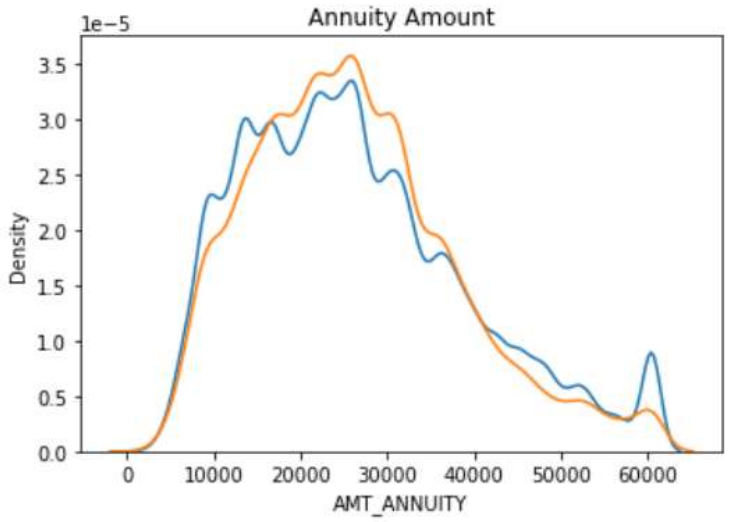


Income type vs defaulter

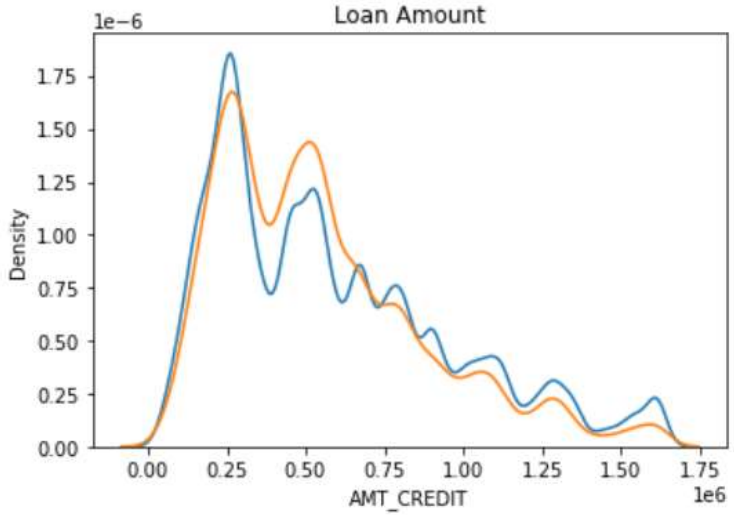


working people default more

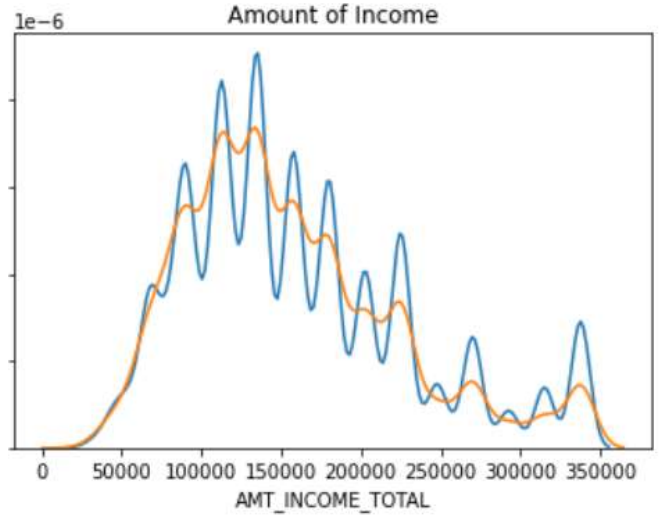
Continuous



Annuity amount for defaulting clients is relatively more

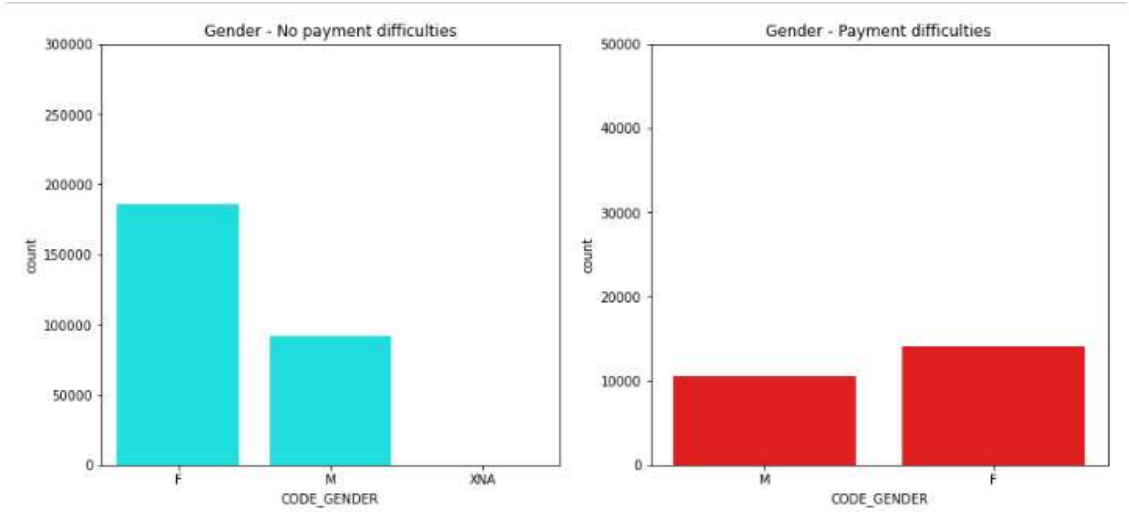


Loan Amount for non-defaulting clients is relatively more

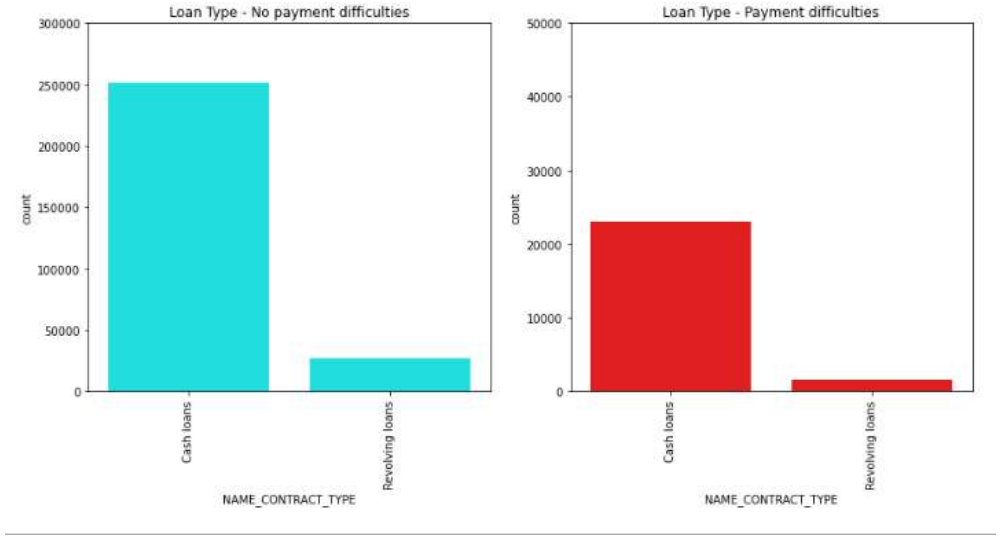


Income of non-defaulting clients is relatively more

Categorical

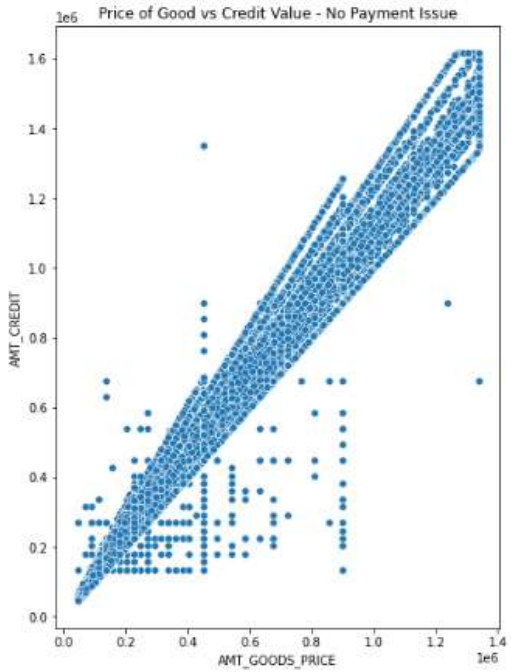


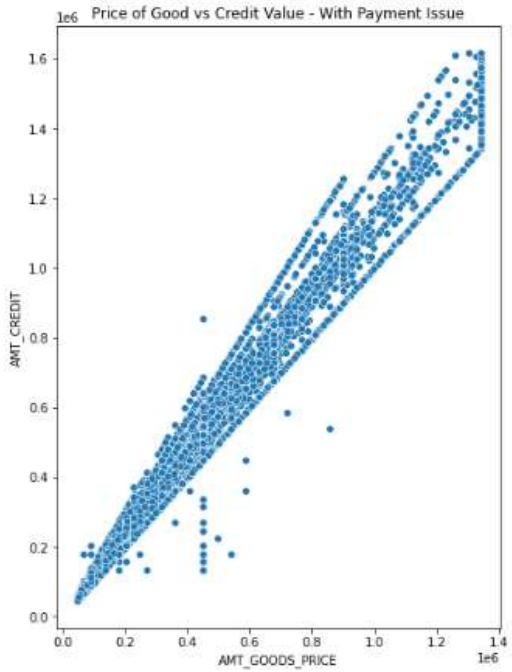
Most males are non-defaulters and most females are defaulters



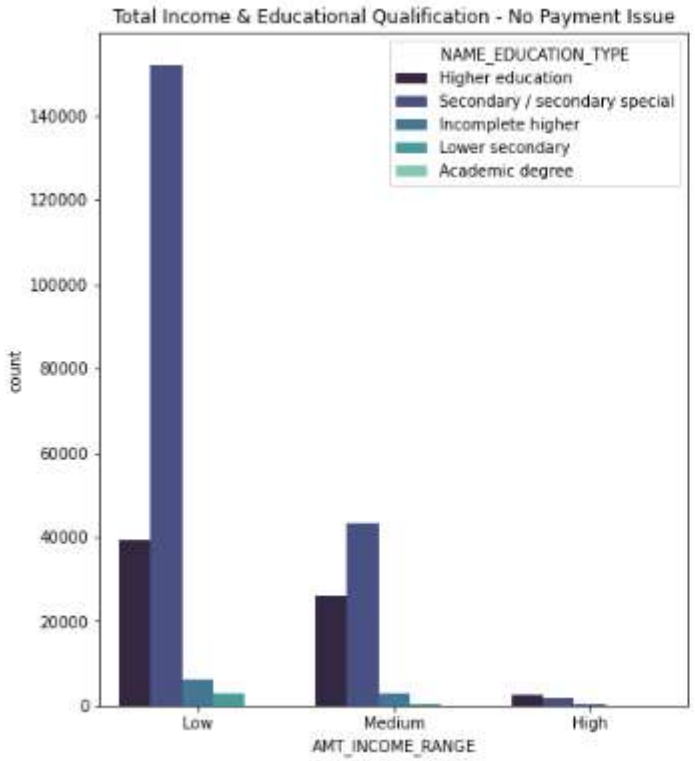
People default more when the loan method is cash payments

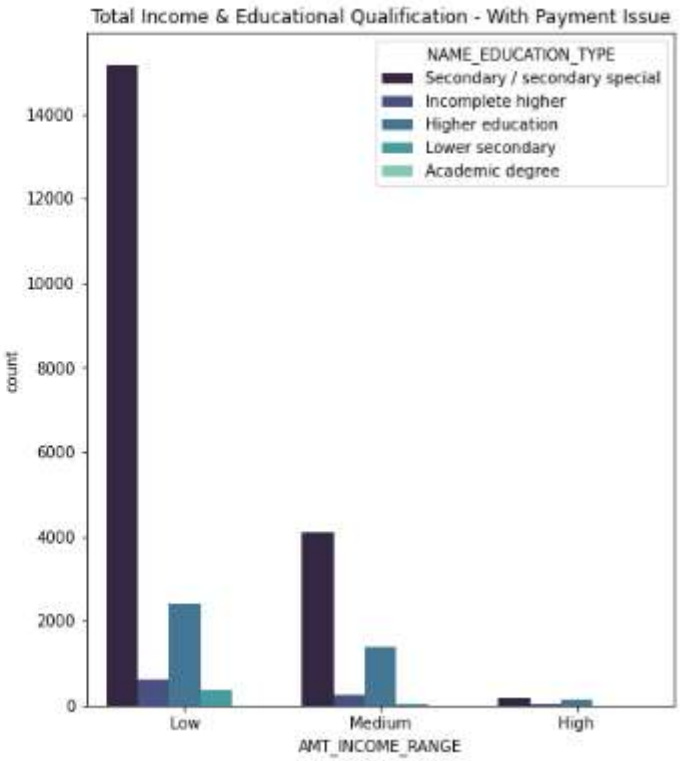
Clients who have repayed the loan on time have a higher chance of getting the loan again for more expensive goods and also have a high probable opportunity to get credit for particular goods value.



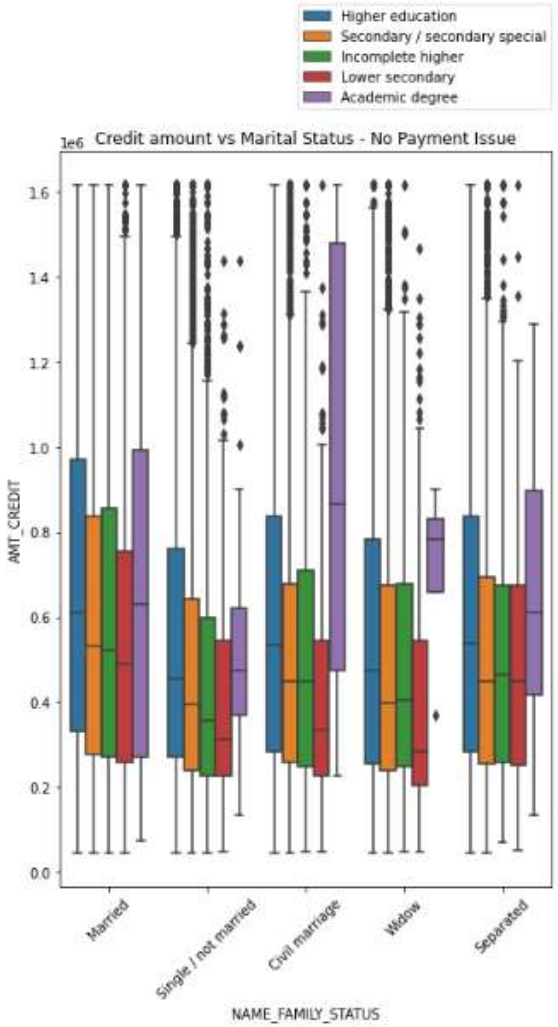


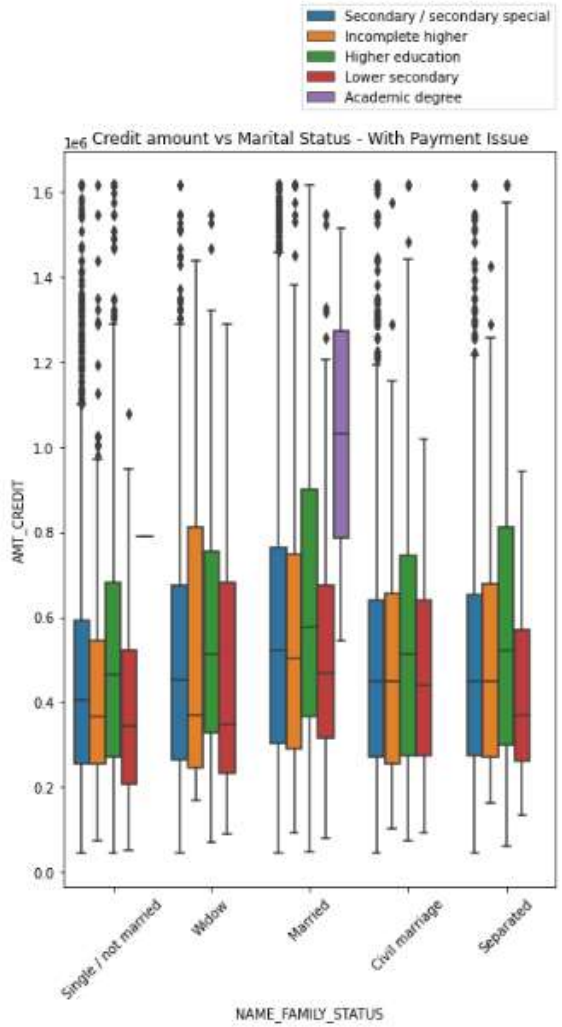
People most likely to repay the loan have secondary/Secondary special education status and a lower income range



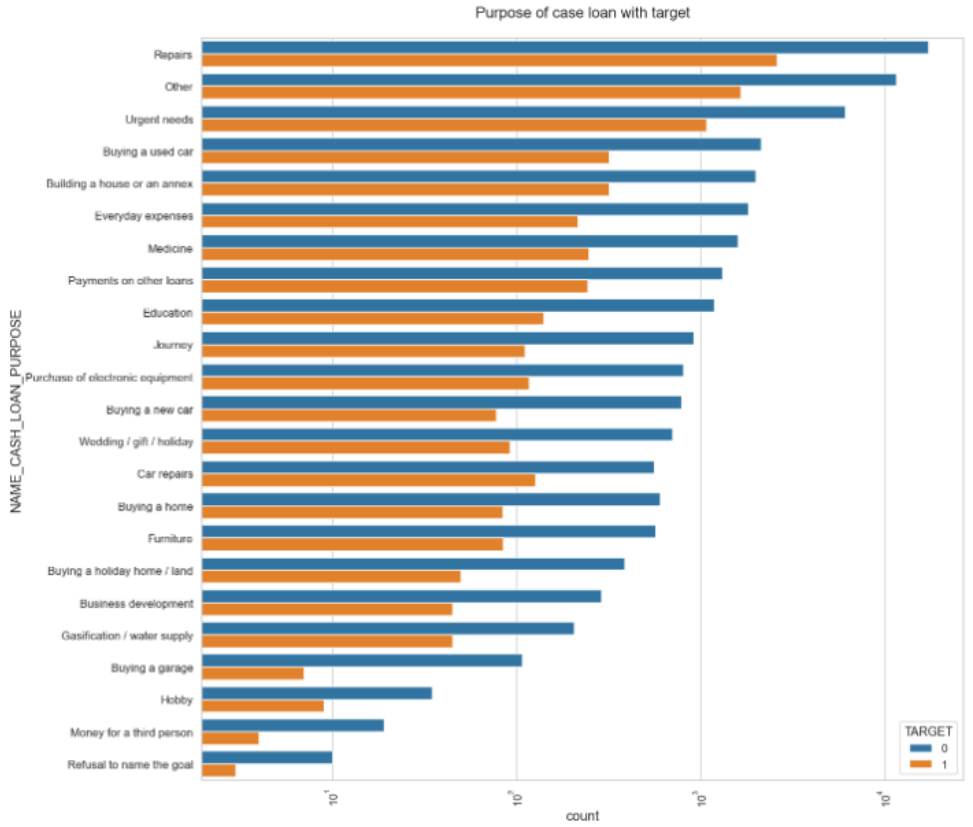


Married people with higher education get higher credit as compared to widows with low educations status

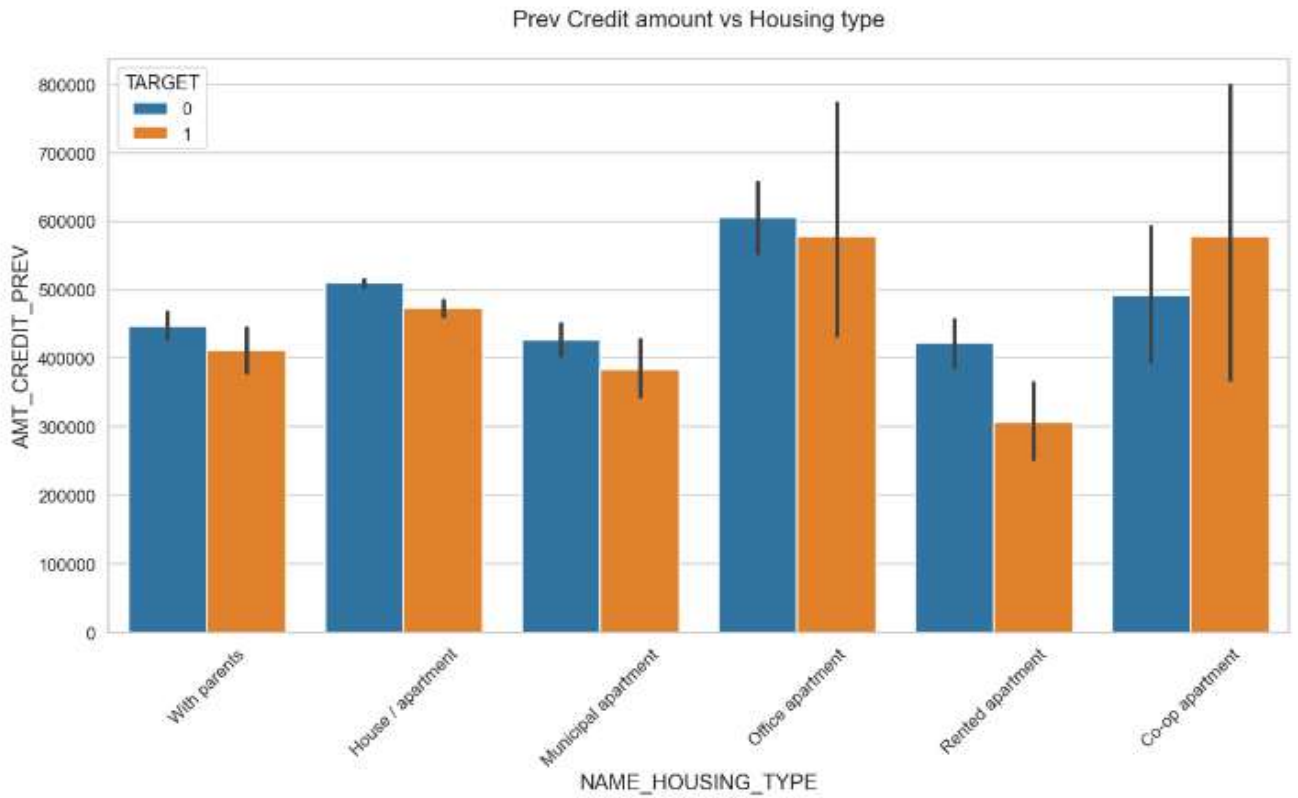




Categories like 'Buying a garage', 'Money for a third person', etc are the categories who have a high chance of successful loan repayment and should be prioritized.



When it comes to the ease of payment, "office apartment" category has the higher credit as compared to others. Also, bank should be careful while approving loans for 'co-op apartment' categories.



Top 10 Correlations

	Col_1	Col_2	Corr
0	DAYS_EMPLOYED	FLAG_EMP_PHONE	0.999705
1	FLAG_EMP_PHONE	DAYS_EMPLOYED	0.999705
2	DAYS_REGISTRATION	registration_change	0.999319
3	registration_change	DAYS_REGISTRATION	0.999319
4	age	DAYS_BIRTH	0.999056
5	DAYS_BIRTH	age	0.999056
6	OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998274
7	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998274
8	DAYS_ID_PUBLISH	ID_change	0.997651
9	ID_change	DAYS_ID_PUBLISH	0.997651
10	DAYS_LAST_PHONE_CHANGE	PHONE_CHANGE	0.991828

Conclusions and Suggestions

- A defaulter is more likely to have a loan application rejected or refused than someone who has no problems making payments. It is advised to look up the client's past application status.
- Customers who change their phone numbers regularly are more prone to default.It has been shown that clients who have a significant likelihood of defaulting if they changed their number in a year.
- Only customers who have linked their phone numbers to an Aadhar or PAN must be granted a loan.
- Clients with little professional expertise run a greater risk of default. Banks have two options for loan reduction: money or use a high-interest rate.
- It has been noted that clients who work as managers and real estate agents typically earn a high mean income and will probably default.
- Avg External Score of a Defaulter is less compared to a non defaulter. These scores are given by external data