Case-Study EDA Assignment

PALAK GARG

Understanding the Problem Statement

- The data provided contains information about the loan application of the clients at the time of applying for the loan.
- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile.
- However, loan-providing companies find it difficult to approve loan applications easily because there are many points to consider before taking further steps.
- If the applicant is likely to repay the loan, then not approving the loan will result in a loss of business to the company. But, 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.

DATA UNDERSTANDING

We have been provided with ‘3’ different datasets which are inter-related to one another:

1. ‘**application_data**’ : this dataset contains all the information of the client at the time of application; including whether he/she has payment difficulties or not.
2. ‘**previous_application**’ : gives information about the client’s previous loan details and application, if any, and if yes, then whether they had been approved, canceled, or refused, etc. This will help in analyzing whether to approve a client’s application or not, interest rate, etc.
3. ‘**columns_description**’ : is basically a data dictionary that describes the meaning of the variables or columns given.

OBJECTIVE

- ✓ Develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.
- ✓ Use EDA to analyze and interpret the data to make it easier to understand and draw conclusions.
- ✓ Identifying patterns and relations to decide whether the company should process the application, charge a high interest rate, reduce the amount of the loan, or directly deny it.
- ✓ Clean the data and make it look simpler and easy to draw insights.

APPROACH

1. **Data Loading and Understanding** - got familiarized with the data using different steps, such as shape, info, describe, and head.
2. **Data cleaning** –

- *Interpreted columns with NULL VALUE RATIO more than 45% and removed them.*
- *Identified and dropped columns of no use for further analysis.*

3. **Handling Missing Values** – imputed missing values of 'Numerical category with median' and 'Category columns with mode'.
4. **Unique Values** – identified and created a separate column for unique values and converted columns with less unique values into Categorical data type columns.

5. **Negative Values** - converted negative values to ABSOLUTE VALUES and units of duration or age from days to years, so that they don’t impact the analysis.
6. **Binning** – created bins for necessary columns and converted some numerical columns with very high values to Categorical columns with the help of bin.
7. **Outliers** – identified and treated outliers. Columns with a substantial gap between the 75th percentile and the maximum value might contain outliers. Outliers are data points that deviate significantly from the rest of the dataset. Therefore, it's important to identify and handle them appropriately.

DATA ANALYSIS

1. **UNIVARIATE ANALYSIS**: involves the examination and interpretation of a single variable or feature in isolation. The focus is on understanding the distribution, central tendency, and spread of that variable.

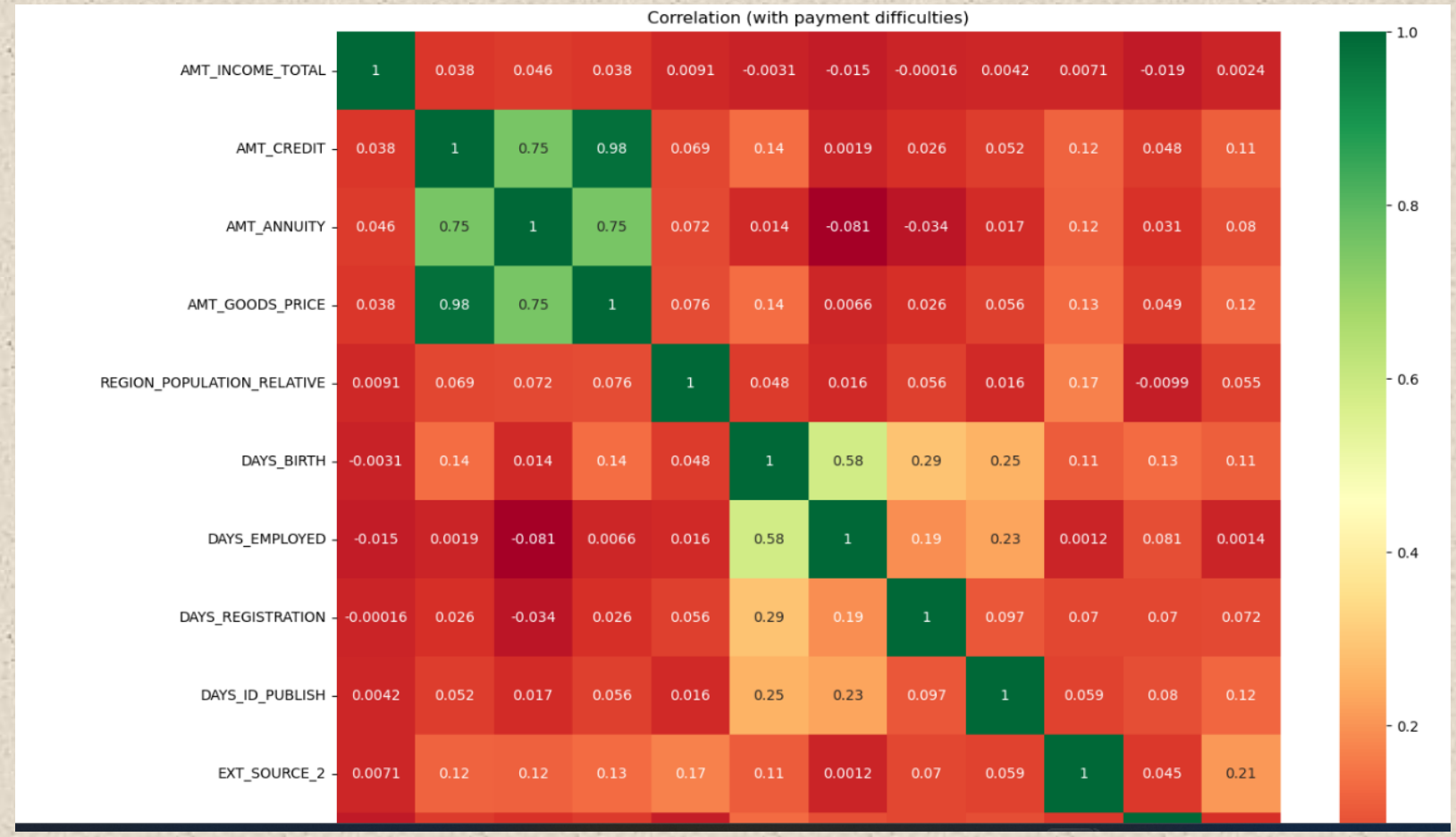
- *Countplot - frequency distribution of categorical variables.*
- *Boxplot - distribution of numerical variables.*

2. **BIVARIATE ANALYSIS:** involves the simultaneous analysis of two variables to understand the relationship between them. The goal is to explore how changes in one variable relate to changes in another.

- *Categorical variables create barplot with hue of TARGET variable to see which category has payment difficulties more.*

3. **MULTIVARIATE ANALYSIS:** involves the simultaneous analysis of three or more variables. The objective is to understand the complex relationships and interactions among multiple variables.

- *Heatmap – created pivot table.*



INSIGHTS

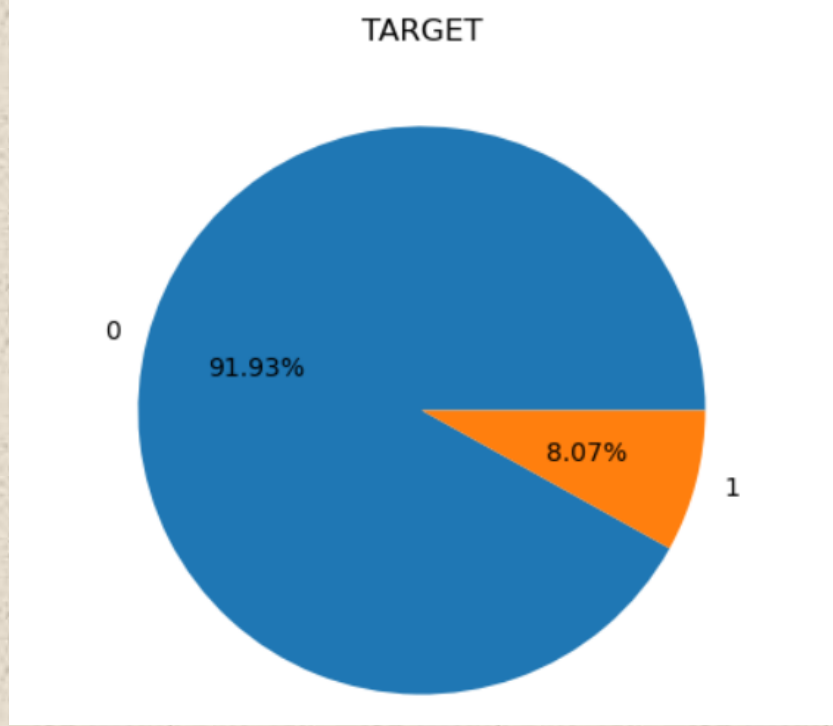
1. **TARGET VARIABLE –**

Target column has 2 variables;

'0' indicating non-defaulters, that is, clients, having no payment difficulties

&

'1' indicates defaulters, that is, clients having payment difficulties.



INSIGHTS

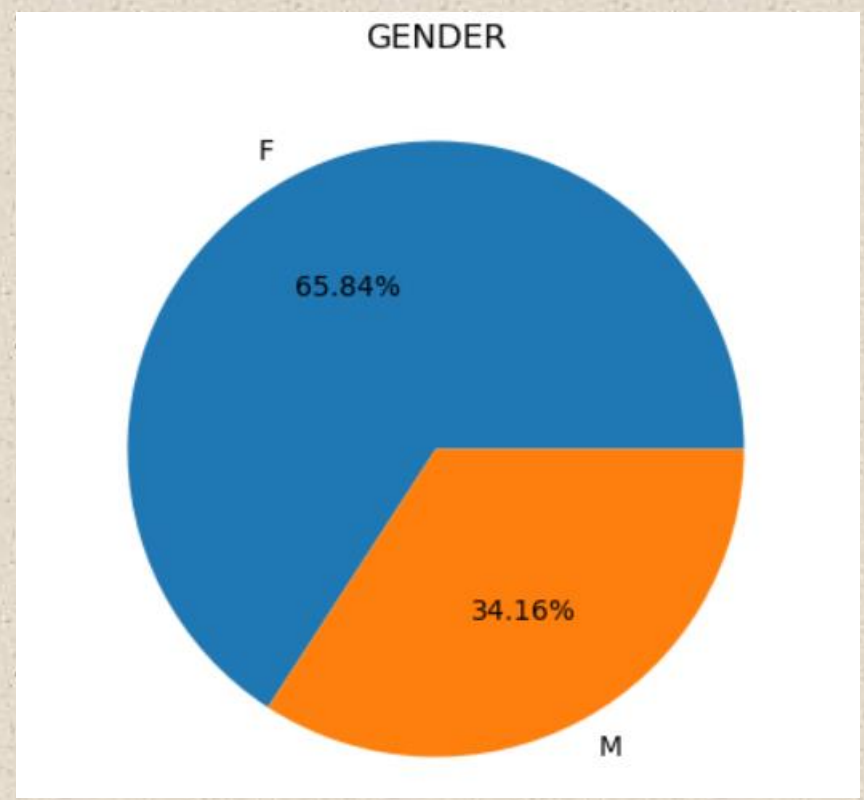
2. **GENDER VARIABLE –**

Loan application on the basis of Gender;

'F' indicating Females – 65.84%

&

'M' indicating Males – 34.16%

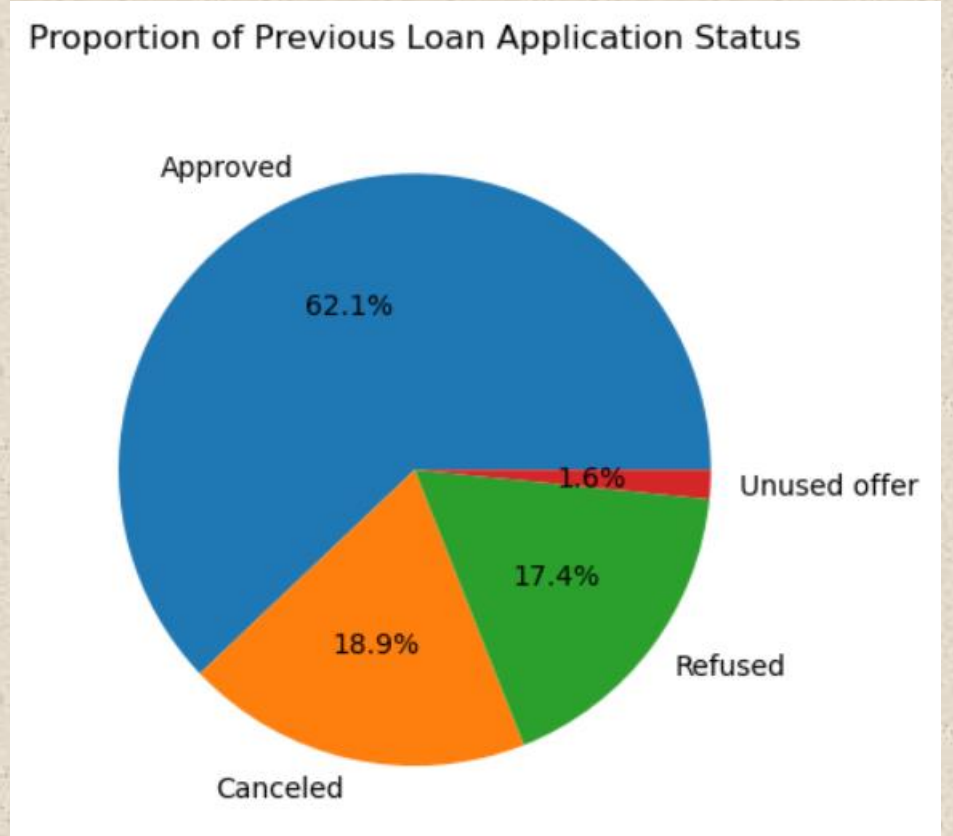


INSIGHTS

3. Previous loan Application status –

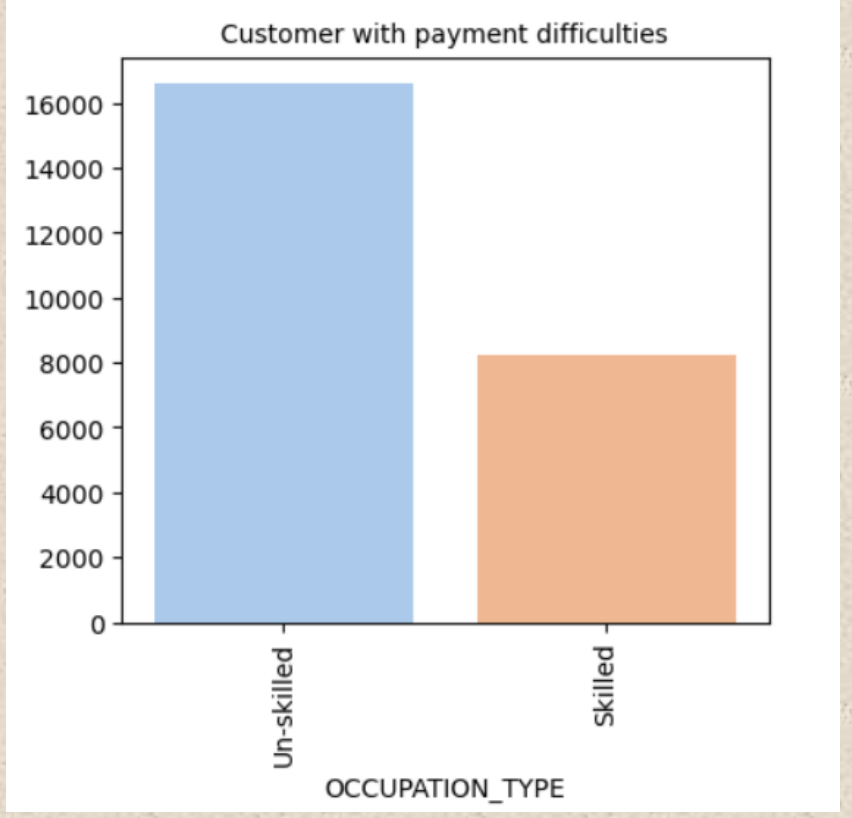
Almost 62% of previous applications of clients got approved, with almost 19% canceled and 17% refused.

This data helps in analyzing current applications for loans more promptly.



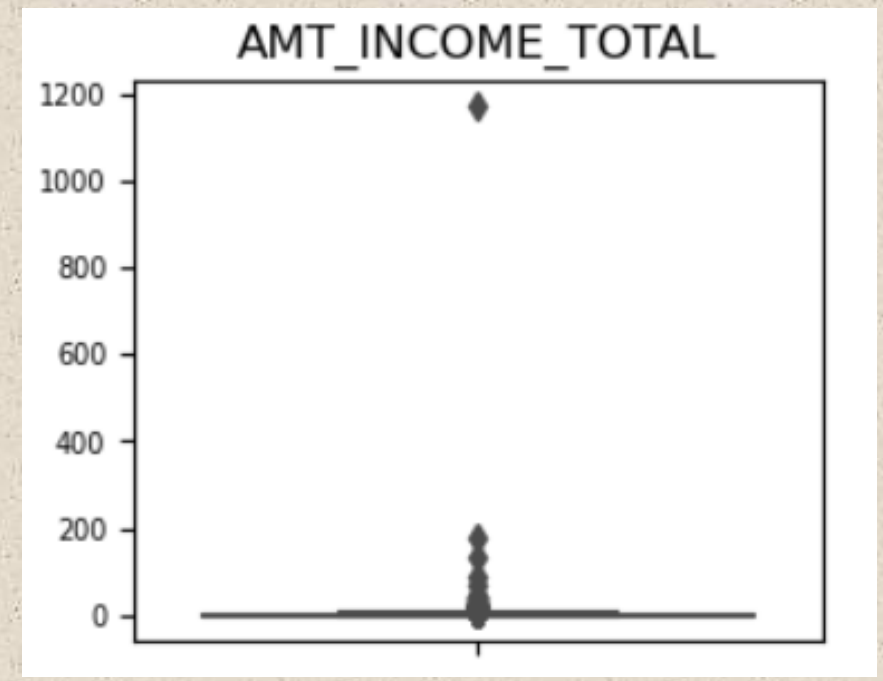
INSIGHTS

4. People from the UN-SKILLED occupation category have more difficulty with payment of loans as compared to the SKILLED category.



INSIGHTS

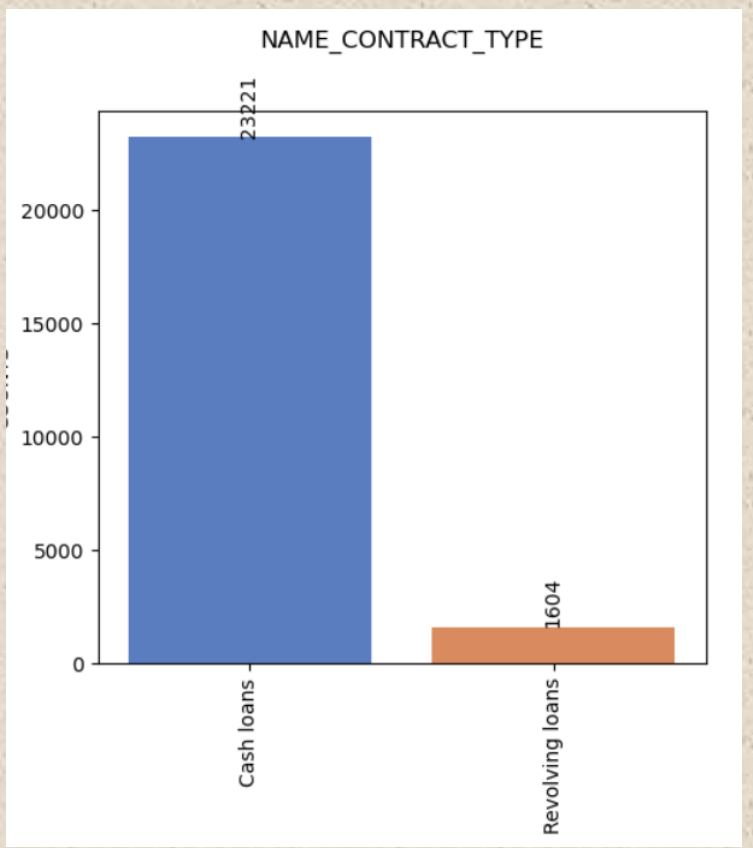
5. **AMT_INCOME_TOTAL has a high number of outliers indicating some people with higher income levels, therefore, we can replace them with lower limit or upper limit values.**



INSIGHTS

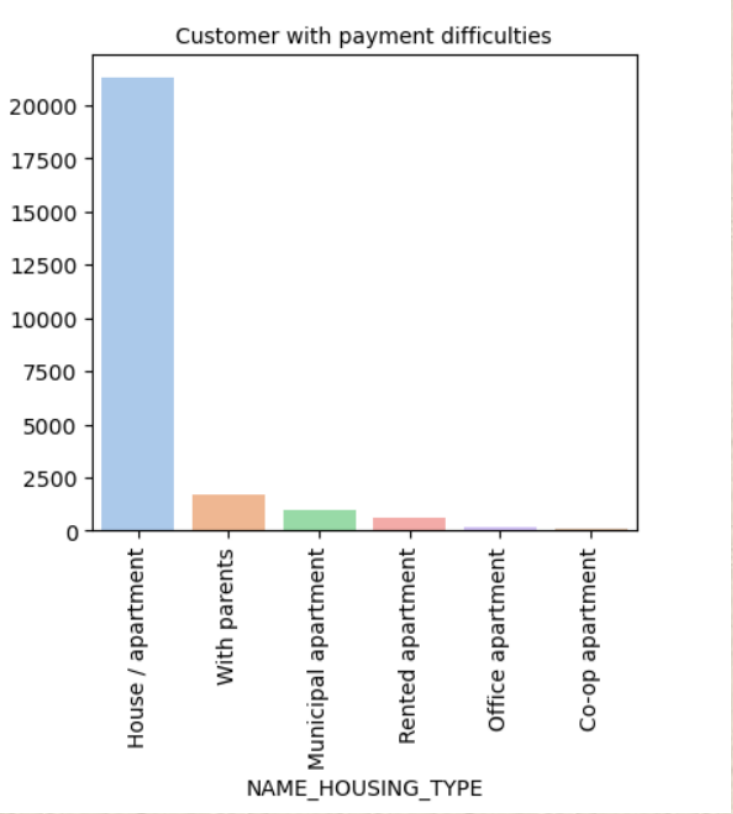
6. Distribution for contract type

For contract type, ‘cash loans’ have a higher number of credits than, ‘Revolving loans’ contract type.



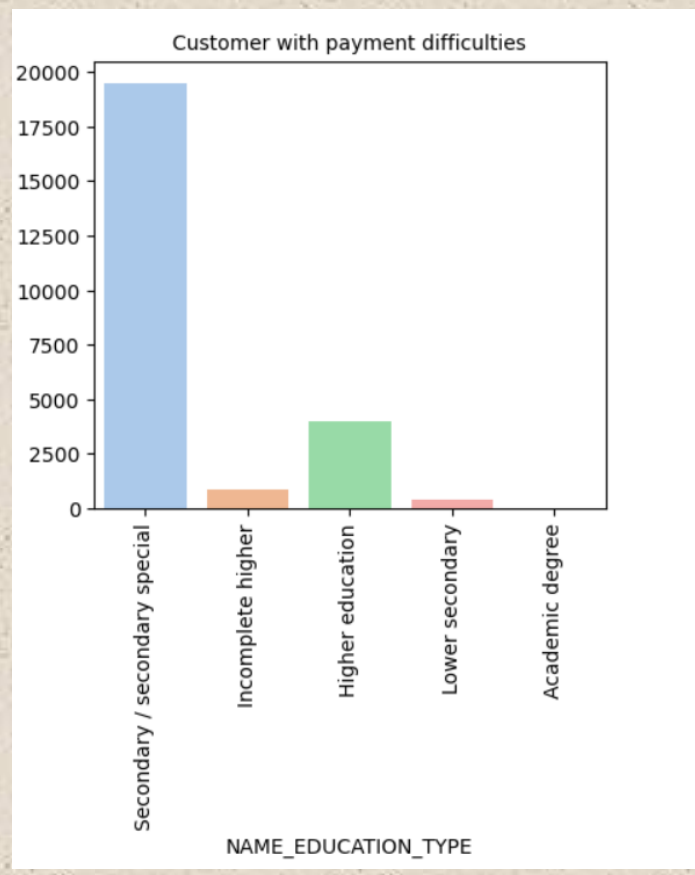
INSIGHTS

7. Clients with houses or apartments are having more difficulty with payment of loan as compared to others.



INSIGHTS

8. People with a secondary level of education are the ones with a higher count of defaulters as compared to others.



CONCLUSION

- Unskilled laborers, unemployed people, or people with less stable jobs and incomes are likely to have more difficulty with payment of loans.
- People having weak educational backgrounds are less likely to pay the loans.
- People from the previously approved application category can be considered for giving the loan but after a thorough analysis.
- Frequent borrowers can bring problems with the payment of loans.
- People having higher incomes can be considered for loans, if any.