Data Understanding:
This dataset has 3 files as explained below:

1. '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.

2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Canceled, Refused or Unused offer.

3. 'columns_description.csv' is a data dictionary which describes the meaning of the variables.

Business Understanding

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

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.

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:
The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample,

All other cases: All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

Approved: The Company has approved loan Application

Canceled: The client canceled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.

Refused: The company had rejected the loan (because the client does not meet their requirements etc.).

Unused offer: Loan has been canceled by the client but on different stages of the process.
In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Business Objective:

The main objective of this case study is to conduct an in-depth analysis of the loan dataset to identify patterns and uncover factors that indicate whether a client is likely to experience difficulties in repaying their installments. By examining the client's profile and loan attributes, the study aims to understand the driving factors behind loan default, providing valuable insights for the company's risk analytics in the banking and financial services sector.
The analysis will enable the company to take proactive measures, such as denying loans, reducing loan amounts, or offering loans to risky applicants at higher interest rates, in order to minimize financial losses and optimize the loan portfolio. By utilizing Exploratory Data Analysis (EDA) techniques, the company can gain a comprehensive understanding of the types of variables and their significance in predicting loan default. Furthermore, the study seeks to evaluate the potential of logistic regression as a predictive model to forecast loan approval or refusal for individual clients. By training the logistic regression model on historical loan data, the company can estimate the

probability of payment difficulties for new loan applicants. This predictive capability will empower the company to make informed decisions about loan applications, ensuring that deserving clients capable of repaying the loan are not rejected while mitigating the risk of default.

Overall, the objectives of this case study encompass enhancing risk assessment methodologies, improving loan approval processes, and leveraging data-driven insights to make accurate lending decisions that align with the company's financial objectives and minimize potential losses.

Results Expected by Learners

- Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.

- You need to upload one/two Ipython notebooks which clearly explains the thought process behind your analysis (either in comments or markdown text), code and relevant plots.

- The Observation file will contain the observation as well as recommendation that you must have achieved after doing this EDA.

- In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also 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.
- Identify the missing data and use appropriate methods to deal with it. (Remove columns/or replace it with an appropriate value)

- Hint: Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach.

- Identify if there are outliers in the dataset.

- Also, mention why you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.

- Identify if there is data imbalance in the data.

- Find the ratio of data imbalance.

- Hint: How will you analyze the data in case of data imbalance? You can plot more than one type of plot to analyze the different aspects due to data imbalance. For example, you can choose your own scale for the graphs, i.e. one can plot in terms of percentage or absolute value. Do this analysis for the 'Target variable' in the dataset (clients with payment difficulties and all other cases). Use a mix of univariate and bivariate analysis etc.

- Hint: Since there are a lot of columns, you can run your analysis in loops for the appropriate columns and find the insights.

- Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.

- Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find the top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.

- Can logistic regression effectively predict the likelihood of payment difficulties for loan applicants? By analyzing client profiles and loan attributes, logistic regression can assess the risk associated with loan applications, aiding in decision-making processes such as loan approval, denial, or modification.

- Include visualizations and summarize the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases.