# Objective

The goal of this case study is to use Exploratory Data Analysis (EDA) to tackle a real-world problem, uncover valuable insights, and present them in a way that's easy for business stakeholders to understand.

## Benefits of the Case Study:

* **Practical EDA Application:** Learn how EDA is applied to real-life business challenges.
* **Risk Analytics Understanding:** Gain a foundational knowledge of risk analytics in banking and finance.
* **Financial Decision-Making:** Discover how data helps minimize financial losses when lending to clients.
* **Improved Visualization Skills:** Enhance your ability to choose and create effective visualizations for real-world data.

## Business Understanding

The business goal is to make informed decisions on loan application whether to approve or reject the based on specific variables.

## Dataset Details:

* **Content:** The dataset contains information about past loan applicants and indicates whether they defaulted on their loans.
* **Scope:** The data includes details about approved loans, not rejected ones.
* **Loan Statuses:** The loans have three statuses: Fully Paid, Current, and Charged-Off.

# Data Cleaning

* There were no header, footers, summary or Total rows found.
* There were 57 columns which is having more than 50% rows values as null/blank and doesn’t participate in analyse has been removed.
* We excluded 21 columns of behavioral data from the analysis, as this information is captured but will not be available during the loan approval process
* 8 columns whose values were 1, and is uniqueness in nature has been dropped from analysis.
* Deleted all the columns which value is uniqe in nature.
* Deleted 'member\_id', and 'url' as it doesn't count in EDA. Will keep the id columns as represent the row data unique for further analysis
* Deleted descriptive or textual informations columns as doesn't participate in EDA analysis.
* Removed funded\_amnt\_inv as it is a internal data and is calculated after loan approval thus cannot be used as input for the
* Removed zip\_code as it is a masked data and cannot be used as input for the analysis
* After all the Data cleaning process we are left with 39717 rows and 18 columns.

## Deleting and fixing the null values

* Removed the null values from the emp\_length and pub\_rec\_bankruptcies column
* Filtered the completed and defaulted loan entries as ‘current’ does not participate in analysis.
* There were no duplicates rows found.

## Correcting Data Types and Deriving New Columns

* Converted term column from string to int. Additional ‘months’ has been trimmed.
* ‘int\_rate’ has been converted from string to flot. Additional ‘%’ has been trimmed.
* issue\_d has been converted to datatype.
* Creating a derived columns for ‘issue\_year’ and ‘issue\_month ’ from ‘issue\_d’ which will be using for further analysis.
* Clean 'emp\_length' column and convert to numeric.

## Removing the outliers

* Outliers exits for numeric data 'loan\_amnt','int\_rate', ‘annual\_inc'.
* Outliers treatment has been done for above fields using quantile mechanism.