## Objective

The objective of this case study is to perform Exploratory Data Analysis (EDA) on the given dataset to identify patterns whether a customer might default on Loan repayment based on a Consumer Finance Company dataset.

### EDA Steps

1. Loading of Dataset
2. Removal of Columns with all values as Null / Same
3. Analyzing Columns basis the Data Dictionary
4. Removing Outliers
5. Correcting Column Data Types
6. Univariate Analysis
7. Segmented Univariate Analysis

These steps are discussed as under :

## Loading of Dataset

We unzip the dataset and then load the resulting csv file using the read_csv command.

## Removal of Columns with values as Null / Same

This is part of data cleaning where we remove all the columns with a significant percentage of empty rows (containing null values). We start by removing columns with all values as null and then progress to other columns with significant null value percentages.
We also remove columns that contain identical values across all rows as they don't add any value to our data.

## Analyzing Columns basis the Data Dictionary

From the data dictionary provided, we find that some of the columns (specifically fields that have id values) don't add any value to our analysis. We find such fields using the description in the data dictionary and then drop them from the data frame.

## Removing Outliers

We have created a generic method that removes the outliers basis the IQR (Interquartile Range). We use this for removing the outliers in the loan_amnt field as we don't want to exclude very high and very low loan amounts in our analysis.

## Correcting Column Data Types

For some fields that are inherently numbers, we do text manipulation and then conversion to their numeric values. This converts them into quantitative variables.

## Univariate Analysis

Univariate analysis is done on certain fields that might provide insights into the customer population. These fields include term, emp_length, home_ownership

## Segmented Univariate Analysis

The column that needs to be analyzed with respect to the impact that the other fields have on it is loan_status. This has 3 possible values :

1. Charged Off
2. Fully Paid
3. Current

“Charged Off” means that the customer has defaulted on loan payment.
We segment other columns based on this categorical variable. The variables considered for this analysis are :

1. int_rate : This denotes the interest rate being charged on the loan. We find that loans that were "Charged Off" had a higher median interest rate of 13.59% as against a rate of 11.49% for loans that were fully paid.

2. pub_rec_bankruptcies : This denotes the public record bankruptcies. We see that with increase in the pub_rec_bankruptcies field, the percentage of people defaulting on loan payment rises.

3. annual_inc : We see that the median annual income of people who have defaulted on loan payment is lower than that of people who have not.

4. delinq_2yrs : We see that if the number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years is 7 or 8, there is a high risk of default, with the highest default rate at 8 delinquencies.

5. total_acc : We have grouped this field into categories. Basically converted this quantitative value field to an ordered categorical value field. We find that in the highest grouping of total_acc field, the percentage of defaulting customers is the least.