The primary libraries of python that we'll be using are:

* **NumPy**: Provides a fast-numerical array structure and helper functions.
* **pandas**: Provides a Data Frame structure to store data in memory and work with it easily and efficiently.
* **matplotlib**: Basic plotting library in Python; most other Python plotting libraries are built on top of it.

Problem Statement:

A Card issuing Bank has over issued its cash and credit card in-order-to increase its market share, even to unqualified applicants. At the same time, most cardholders, irrespective of their repayment ability, overused credit card for consumption and accumulated heavy credit and cash– card debts. The crisis caused the blow to consumer finance confidence and it is a big challenge for both banks and cardholders. From the perspective of risk control, estimating the probability of default will be more meaningful than classifying customers into the binary results – risky and non-risky.

Description of the Data:

This research employed a binary variable – default payment, Y (Yes = 1, No = 0), as the response variable.

This study used the following 23 variables as explanatory variables:

X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

X2: Gender (1 = male; 2 = female).

X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

X4: Marital status (1 = married; 2 = single; 3 = others).

X5: Age (year).

X6–X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows:

X6 = the repayment status in September, 2005;

X7 = the repayment status in August, 2005;

X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: 0 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; ...; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

X12–X17: Amount of bill statement (NT dollar).

X12 = amount of bill statement in September,2005;

X13 = amount of bill statement in August, 2005;

X17 = amount of bill statement in April, 2005.

X18–X23: Amount of previous payment (NT dollar).

X18 = amount paid in September,2005;

X19 = amount paid in August,2005;

X23 = amount paid in April, 2005.

**Analytics tasks:**

* Analyze the trend on outstanding amount for the bank
  + Overall outstanding amount trends
  + Number of customers with outstanding amount (in different outstanding amount buckets)
  + Age of outstanding amount analysis
* Is there any relationship between in outstanding amount / trend with respect to age, education, marriage, credit limit
* Does outstanding amount / trend affect the default behavior in next month.
* Identify the errors in dataset
* Check for normal Distribution for important columns and provide me valuable insights from the visualization.
* Check for outliers and guess the reason behind outliers and say whether they are worthy to have it in our dataset or not required.