**Rapport Machine Learning**

1. **Description :**
2. ID: ID of each client
3. LIMIT\_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
4. SEX: Gender (1=male, 2=female)
5. EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
6. MARRIAGE: Marital status (1=married, 2=single, 3=others)
7. AGE: Age in years
8. PAY\_0: Repayment status in September, 2005 (-1=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)
9. PAY\_2: Repayment status in August, 2005 (scale same as above)
10. PAY\_3: Repayment status in July, 2005 (scale same as above)
11. PAY\_4: Repayment status in June, 2005 (scale same as above)
12. PAY\_5: Repayment status in May, 2005 (scale same as above)
13. PAY\_6: Repayment status in April, 2005 (scale same as above)
14. BILL\_AMT1: Amount of bill statement in September, 2005 (NT dollar)
15. BILL\_AMT2: Amount of bill statement in August, 2005 (NT dollar)
16. BILL\_AMT3: Amount of bill statement in July, 2005 (NT dollar)
17. BILL\_AMT4: Amount of bill statement in June, 2005 (NT dollar)
18. BILL\_AMT5: Amount of bill statement in May, 2005 (NT dollar)
19. BILL\_AMT6: Amount of bill statement in April, 2005 (NT dollar)
20. PAY\_AMT1: Amount of previous payment in September, 2005 (NT dollar)
21. PAY\_AMT2: Amount of previous payment in August, 2005 (NT dollar)
22. PAY\_AMT3: Amount of previous payment in July, 2005 (NT dollar)
23. PAY\_AMT4: Amount of previous payment in June, 2005 (NT dollar)
24. PAY\_AMT5: Amount of previous payment in May, 2005 (NT dollar)
25. PAY\_AMT6: Amount of previous payment in April, 2005 (NT dollar)
26. Y: Default payment (1=yes, 0=no)

* Total Variables: 23 Features , 1 Target

2) Problem Definition :

The incidence of the recent financial crisis has shown that the general public is not familiar with the elementary financial concepts to make proper financial decisions. In order to understand the relationship between customer income and financial failures, the minimum monthly credit card payment of individuals should be studied deeply. Hence, in this paper, we investigate the failing to pay the minimum credit card balance which is a transaction to be regularly carried out every month. In order to increase consumer finance confidence, avoiding delinquency is a big challenge for cardholders and banks as well. In a well-established financial system, risk estimation is more important than crisis management. The significant reason for hazard expectation is to utilize money-related data, for example, business financial statements, client transaction and reimbursement records, etc. To predict business execution or individual clients' credit chance, and to lessen the harm and instability.

In order to prevent these financial complications including decisions that are easier and more frequent in the unintentional failure for paying monthly credit card balances, we proposed a data mining-based failure prevention system from the view of risk management.

3) Objectif:

The project is being prepared to understand whether the credit card customer will make a payment default in the next month or not based on a number of features.

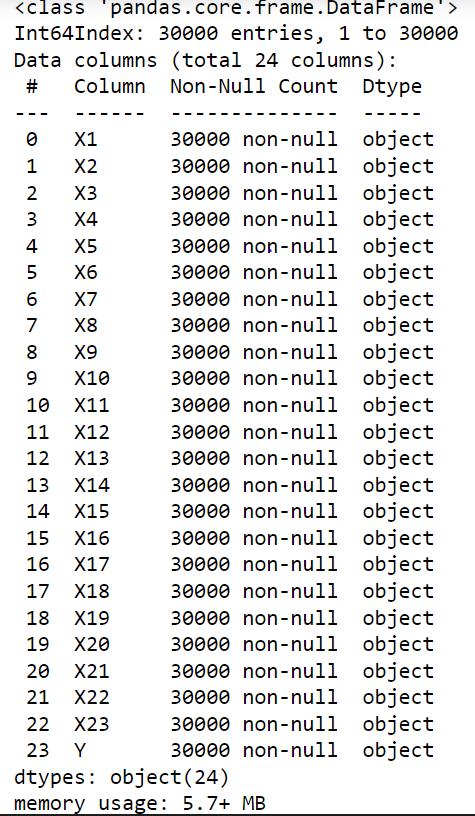
# 4) Data Understanding

We used an anonymized dataset of Taiwanese credit card holders from October 2006 (Chou, 2006). This dataset was made available to the public and posted on the UC Irvine Machine Learning Repository website.

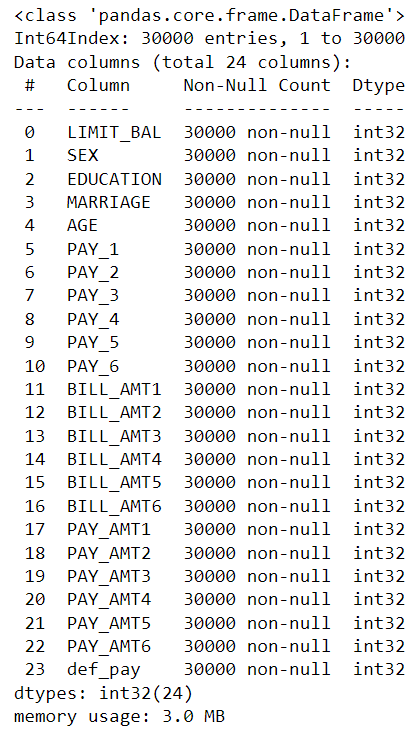
The dataset contains:

* \*30,000 observations. Each of these observations corresponds to an individual credit cardholder.

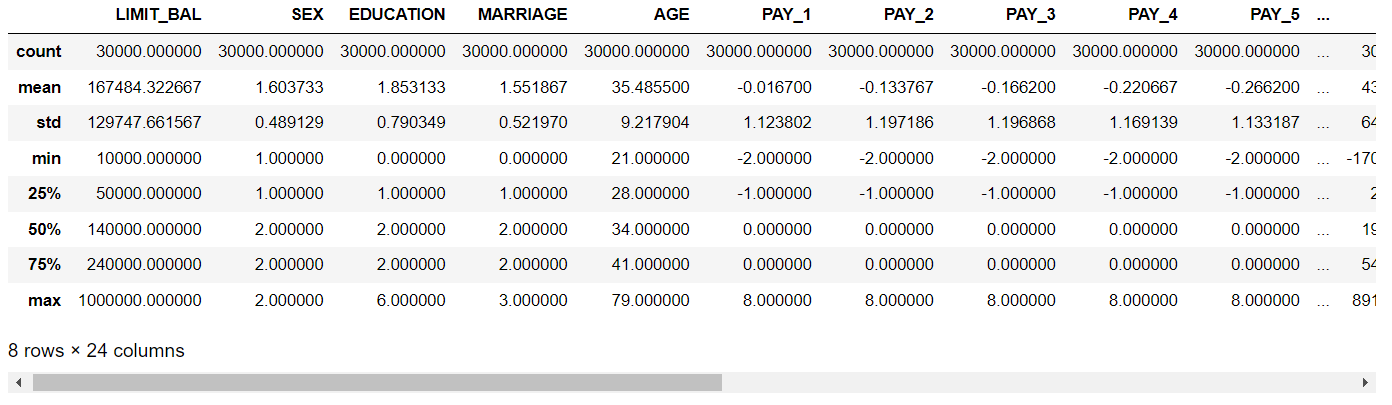
**Reading dataset and viewing info**



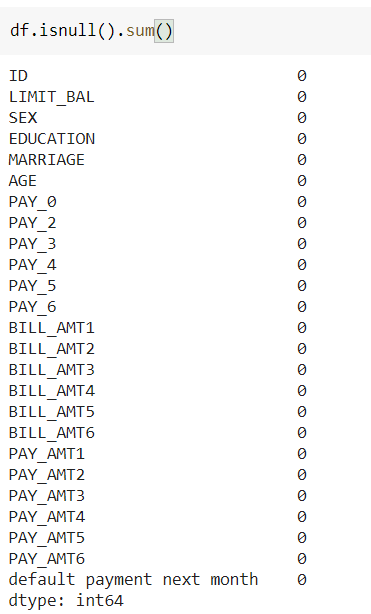
**We noticed that all variables are object, so we made the change to Integer.**



**Dataset description**



**The values equal to zero**

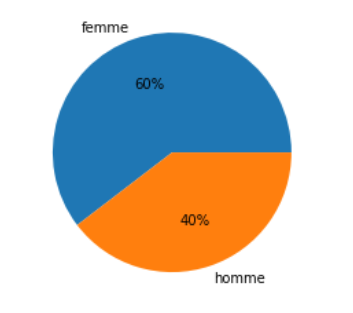


**From the description of the dataset we see that:**

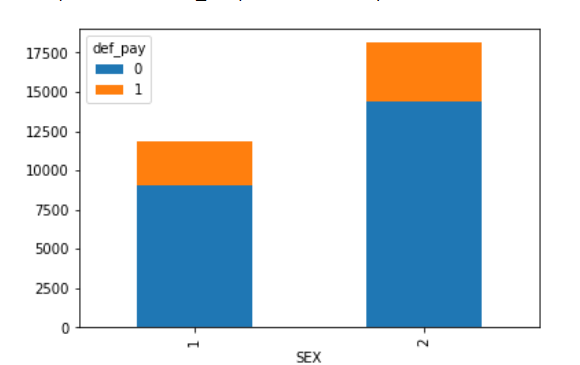
* **the number of women is higher> Men (average (sex) = 1,603)**
* **the average age is 35 years or we can say that the majority of people have completed their university studies**
* **the simple number of Pay\_M = -2 which does not exist in the documents**
* **50% of people are single**
* **min(BillAmt\_m) is Negative which is not logical, it must be studied**

**Categorical Variables**

* **SEX:**

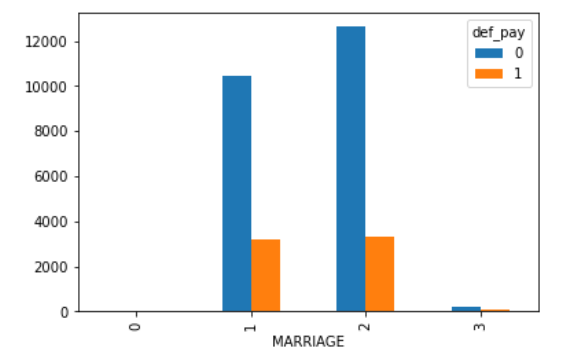


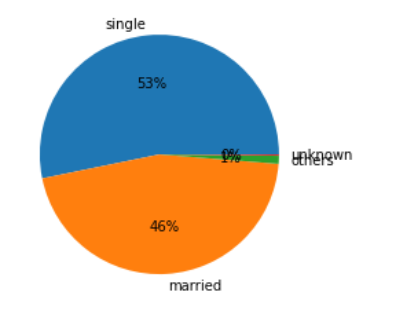
**Our dataset is consisted of 60% Women and 40% Men**



**The non-defaulters with the high count are the women**

* **Marriage :**

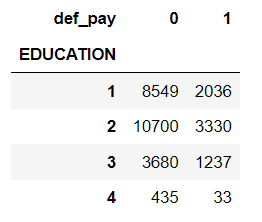


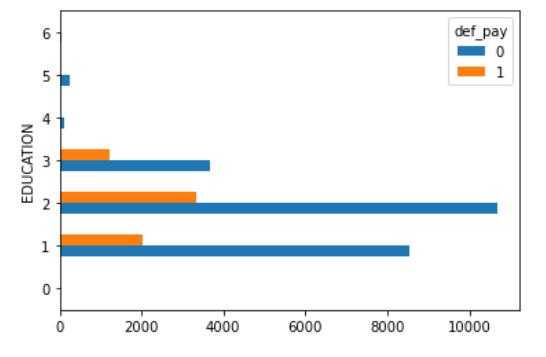


**We see the presence of anonymous type we will replace it with others.**

**As expected, most people fall either on the 'Married' or 'Single' category.**

* **Education**



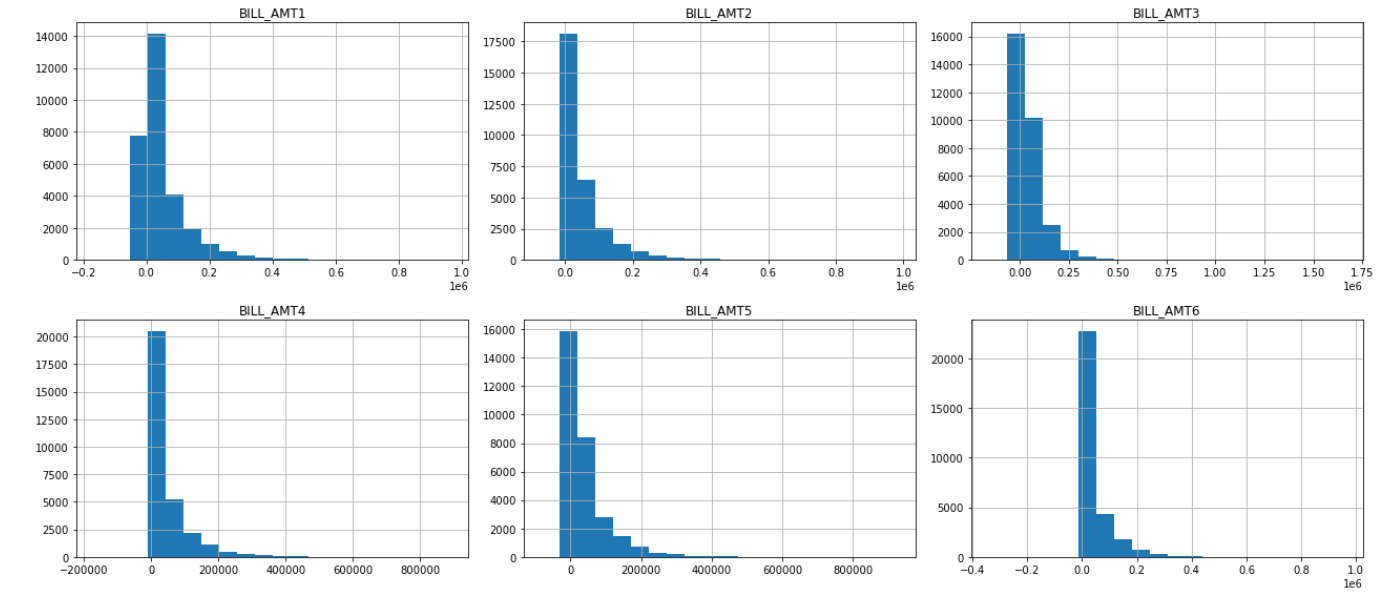


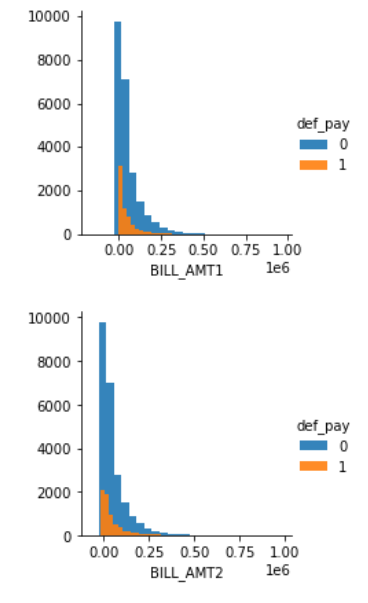
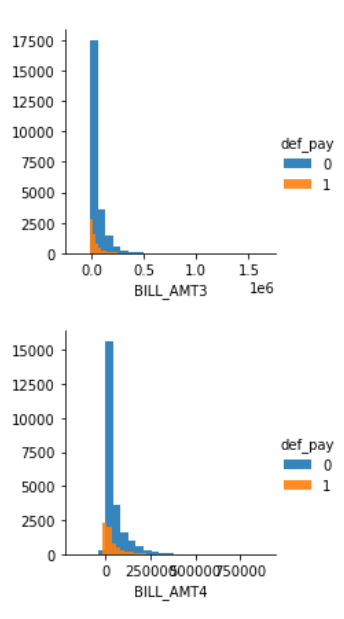
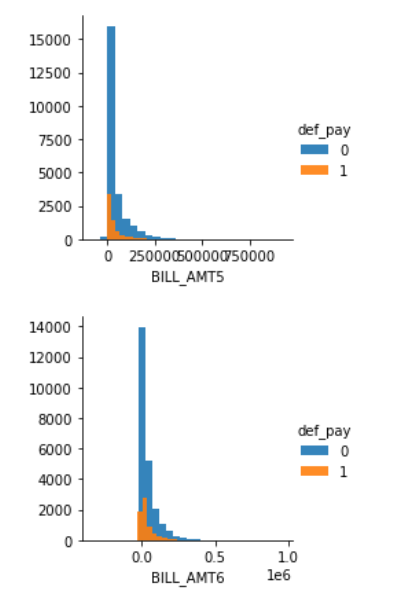
**We should replace the other types (4,5,6) with the type other**

**It seems that the higher is the education, the lower is the probability of defaulting the next month.**

### **Numerical Variables**

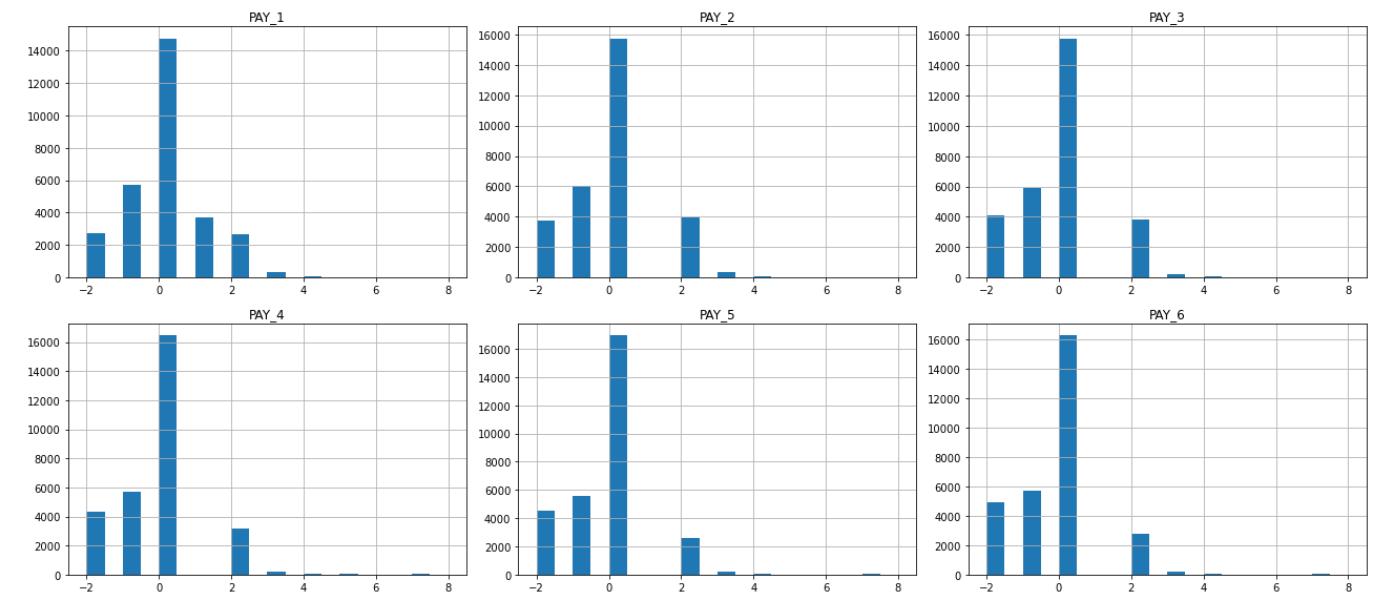
* **Bill Amount**





**As expected, those who have a negative bill statement have a lower chance of default than the rest.**

* **PAY\_X**



**Most customers are duly paying their credit card bills. And it's pretty clear that their likelihood of default is much lower than the rest.**

**we see that those who have credit cards pay maximum after 2 months, so we can work only with PAY\_1 and PAY\_2**

### **Data Cleaning**

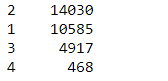
**As seen previously, some categories are mislabeled or undocumented. Before proceeding, it is time to fix it.**

**The 0 in MARRIAGE can be safely categorized as 'Other' (thus 3).**

**The 0 (undocumented), 5 and 6 (label unknown) in EDUCATION can also be put in a 'Other' cathegory (thus 4)**

**Categorical Variables**

Regrouping Education 5,6,0 In one category called Other(4)



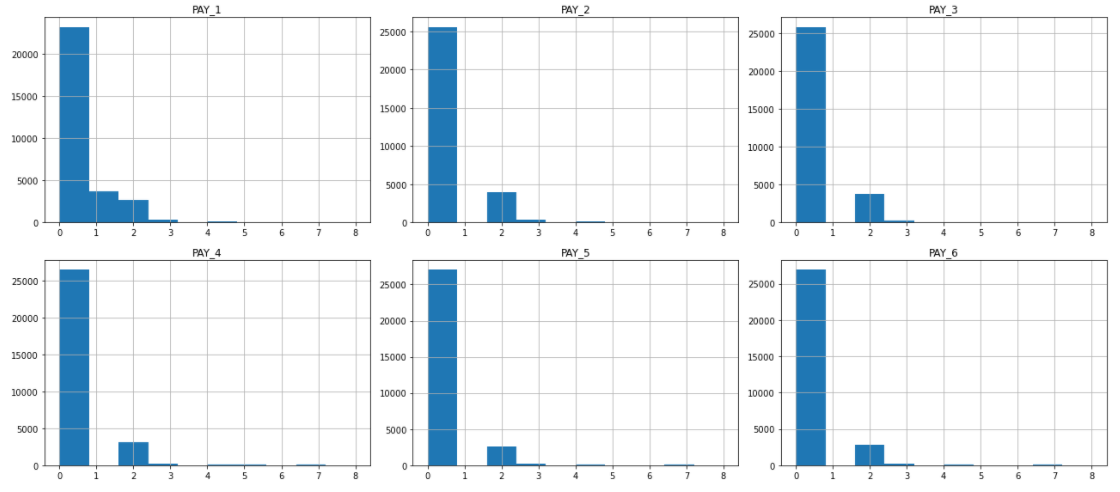
Regrouping Marriage 0 In one category called Other(3)



**Numerical Variables**

**According to our documentation, the PAY\_X variables indicate the number of months of delay and indicates "pay duly"with -1. Then what is -2 for ? And what is 0? We will adjust the label to 0 for pay duly.**

Regrouping -2,-1 to 0 which represents the pay duly



**Correlation**

