Data wrangling for Capstone project

#### DataSet Information:

This research is aimed at the case of customers default payments in Taiwan in 2005. The Credit Card Default Data on the UCI Machine Learning Repository can be found directly below:

<http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>

Attribute Information:

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

This study reviewed the literature and used the following 23 variables as explanatory variables:  
**LIMIT\_BAL**: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.  
**SEX** : Gender (1 = male; 2 = female).  
**EDUCATION**: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).  
**MARRIAGE** : Marital status (1 = married; 2 = single; 3 = others).  
**AGE** : Age (year).

*(X6 - X11)*  
**PAY\_0 - PAY\_6**: 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: -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.

*X12-X17*  
**BILL\_AMT1 - BILL\_AMT6**: 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*  
**PAY\_AMT1 - PAY\_AMT6**: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

Data wrangling steps:

1. Import the csv file in pandas dataframe and inspect the dataframe
2. Draw boxplot and violinplot for the missed payment columns to understand the gravity of missed payments
3. Change the values of categorical variables to more understandable value names:
   1. SEX: from 1,2 to ‘male’,’female’
   2. EDUCATION: from1,2,3,4 to ‘graduate school’, ‘university', 'high school', 'others'
   3. MARRIAGE: from 1,2,3 to 'married', 'single', 'others'
4. Create a new column called AGE\_RANGE: '20-29','30-39','40-49','50-59','60-69','70-79'.
5. Calulate the average monthly balance(**AVG\_6MNTH\_BAL)** for the 6 month period of April to September 2005
6. Calculate the maximum number of missed payments(**MISSED\_PYMTS)**

during that 6 month period.

1. Create a dataframe with limited columns for future analysis:
   1. **ID**
   2. **LIMIT\_BAL**
   3. **SEX**
   4. **EDUCATION**
   5. **MARRIAGE**
   6. **AGE\_Range**
   7. **MISSED\_PYMTS**
   8. **BALANCE\_TO\_LIMIT**
   9. **default payment next month**
2. Try to answer the questions:
   1. # Which group has the highest average credit limit?
   2. # Which group has the lowest average credit limit?
   3. # Which group is comprised of highest percentage of people who have a balance-to-limit rating less than or equal to 30%?
   4. # Which group has the lowest utilization or balance-to-limit rating?
   5. # Which group has the highest amount of debt, is the most likely to default, and is the most likely to miss a payment?
   6. # Which group has the lowest amount of debt, is the least predicted to default, and is not likely to miss a payment?