Data Wrangling

First, I imported required packages that we would need for Data Wrangling. Then, I used pandas read_excel module to import excel file directly from the website of UCI Machine Learning Repository. The excel file had two column labels for each column, so we skipped the first row to ignore the label that we do not need. Then, I converted that imported OrderedDict to a pandas DataFrame. The column label/names were inconsistent, so i changed the column labels with better column labels. Columns for specific month included month name, and space in between from column default payment next month was changed to default_payment. Then, I checked to see if there were any missing value and outliers.

In column 'education', the value 1 represent graduate school, 2 represent university, 3 represent high school, and 4 represent others. However, there were values 5 and 6 which would be a mistake or those value does not represent any level of education. I located those values and replaced those value with 4 and kept them under the category 4 representing others.

In column 'marital_status' the value 1 represent married, 2 represent single, and 3 represent others. However, there was value 0 present in the column which did not represent any category. That value may be a missing value or a mistake or others. Here, I located those value and replaced those value with 3 which represent others.

- **What kind of cleaning steps did you perform?**
  Checked for missing data
  Checked for inconsistent column names
  Checked for Outliers
  Checked column data types

- **How did you deal with missing values, if any?**
  There are no missing values for this dataset.

- **Were there outliers, and how did you handle them?**
  There were outliers in the age column, the balance amount columns for each month, and the amount paid columns for each month. We checked for outliers using Box Plot Diagram and 1.5*IQR rule. Both of them proved that outliers were present. I created a copy of current dataframe to update and modify the outlier values from those columns. Data higher or greater than Q3+1.5*IQR is replaced by Q3+1.5*IQR, and data lower than Q1-1.5*IQR is replaced by Q3-1.5*IQR. I did not change the outliers from the age column.