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.

First Approach (update and modify bad data 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, 6 and 0 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. Here, I assumed that the best possible value for these bad data would be others. Others meaning any other education level except 1 graduate, 2 university, and 3 high school. This is the reason why i choose others would be the best possible value.

Addition to this, I will deal with the data by using other techniques like using mode and median to replace the bad data and outliers, and drop/delete data to exclude such bad data from affecting our study. The first and important approach would be to consult or ask with other team members if they know exactly why we have such bad data.

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 assuming other to be the best possible value.

There are six column for repayment status for month september, august, july, june, may, and april. In each column, value -1 represent pay duly, 1 for payment delay one month, 2 for payment delay two months,.....8 for delay 8 months and 9 for delay 9 month or above. There are value 0 and -2, which are bad data. I checked to see what those value were in previous months. Then, located those bad data to change them to -1, which represent duly paid.

- **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.

  Second Approach (Remove/Drop bad data and outliers):
  In the second approach, I dropped all the rows with bad data and outliers instead of estimating the best possible value to replace them. First, I would locate any bad data and outliers, then I drop them to exclude them from the analysis.

  Third Approach (Replacing bad data with mode value and Outliers with median value) :
  In the third approach, I located any bad data and replaced them with their mode value. Mode value is the most repeated and majority of the value. I did not use mean and median, because mode is a better choice for categorical data. Then I replaced any outlier values to their median value. Median value is the middle value or mid-point of the data set.