**Data Wrangling for Credit Card Default Data**

1. What kind of cleaning steps did you perform?

In the project, I first check the shape of the data. There are 30000 records and 24 different variables. I visually inspect the first five and last five rows of the data.

• I check the column names. I rename some of the variables to make it easy to understand.

• If I have string categorical variable, I would check if they use a consistent way to represent things. For example, ‘Friday’, ‘fri’, and ‘Fri’ could all be there for Friday. ‘United States’, ‘USA’, and ‘US’ could be the same thing.

• I check the data type of each column and the data type matches the content.

• I count the frequency for each categorical variables. I find a few strange things:

* For Education, there are three undocumented categories (0, 5, 6). A future investigation could be done if I have access to people who created this data. Since I do not have that access, for now, I will group 0, 5, 6 to others (which is 4).
* For Marital status, class 0 is not documented. For now, I group 0 to others (which is 3).

2. How did you deal with missing values, if any?

I also check if there are any missing values. I did not find any missing values in the current dataset. However, if I have missing values, I would need to investigate the nature of the missing values.

• If the records with missing values are only a small fraction of the dataset, we can remove all the cases with missing values.

• If we are dealing with a small data set, we need to see which imputation methods we could use for the data set. If the missing value is numerical and missing at random, we can use mean if we are going to fit the data with linear regression since mean is an unbiased estimator. If we want a simple measure that is robust to outliers, we could use the median.

• Another good choice for imputation is to use models to predict the missing values. A good example is MICE (multivariate Imputation via Chained Equations). In Python, FancyImpute has a function IterativeImputer, which is similar to MICE but it returns a single imputation. We can also use interpolation with linear regression, random forest, and so on.

• If the missing value is categorical, one could impute value ”missing” to make missing values as a category. We can also impute with the most frequent class.

3. Were there outliers, and how did you handle them?

I applied median-absolute-deviation (MAD) based outlier detection for all numerical features. I used a threshold of 3.5. A data point with Z score whose absolute value larger than 3.5 is labeled as an outlier. I also used boxplot to check for outliers.

However, outliers do not mean errors. It could be a valid point as well. By checking the data for balance limit records, they look valid points.