Data Cleaning

There are 86 variables in the raw data, many of which are highly correlated and hard to interpret. Therefore, we group data by project code, and calculate 8 variables based on original variables. Most of calculations use given instruction (‘calcs.pdf’) as inference.

We notice that PTD (Period to Date) is the accumulation of MTD (Month to Date). For example, for project with code ‘120400100’, the Labor Cost PTD for 2004/03 is 676.92 and the Labor Cost MTD for 2004/04 is 233.27, then the labor cost PTD for 2004/04 is 910.19, the sum of 676.92 and 233.27. To reduce data redundancy, we tend to choose the maximum value in PTD, which means the total value for each project. For the project mentioned above, we choose the maximum value in the Labor Cost PTD 910.19 as the total labor cost of this project.

We remove project starting with code ‘15’, which are not profitable according to instruction.

# Detect NANs

There is no missing values in the dataset.

# Create new variables

Labor Cost:

Value = the maximum (Labor Cost PTD) for each project

Overhead:

*[Overhead expenses are indirect costs, meaning they are not related to specific business activities that generate money.]*

Value = the maximum (Overhead PTD) for each project

Overhead Rate:

According to the instruction, overhead rate changes monthly so it need to be calculated for each month. For each project, we take the average of monthly overhead rates as overhead rate of each project.

Value = the mean of (labor Overhead MTD / Labor Cost MTD) for each project

If Labor Cost MTD = 0, we define overhead rate 0.

Labor Revenue:

Value = the maximum (Labor Revenue PTD) for each project

Complete Labor:

Value = maximum (Labor Revenue PTD)/ maximum (Total Labor Budget Effort)

WIP\_PTD :

*[The term work-in-progress (WIP) is a production and supply-chain management term describing partially finished goods awaiting completion.]*

Value = maximum (Revenue PTD) –maximum (Billing PTD)

Profit PTD:

Value = maximum (Revenue PTD) – maximum (Cost PTD) + maximum Overhead

Profit Percentage:

Value = Profit PTD /maximum (Revenue PTD)

If Revenue PTD = 0, we define profit percentage 0.

# Subsequent processing

* We delete the project with Profit Percentage < -1 or > 1.
* We delete outliers in Overhead rate.
* According to instruction, a positive WIP means “unbilled revenue” and negative WIP means “billings in excess”. To reduce redundancy, we transform WIP\_PTD into binary value. We define it 0 for positive value and 1 for negative value.