One-minute analysis: EDA

As seen in Figure 1a., income have a cluster around 5,000 to 15,000. This suggests a skewness in income data with a few values on the right tails. This will impact the regression and relationship with target variable by enlarging the variance. With more spread-out values in the income variable, the mode will be able to predict more accurate with a linear regression. Therefore, Figure 1a. suggests a transformation for income.

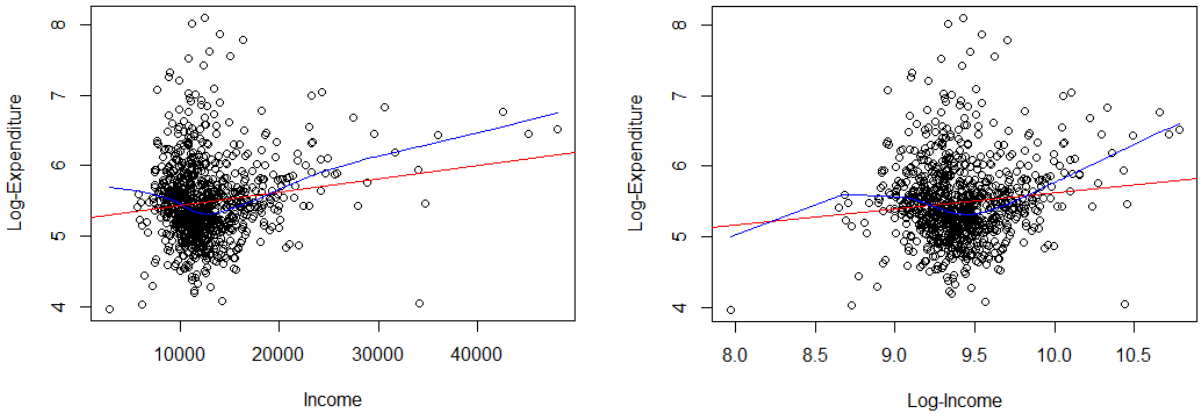

Figure 1. a: Scatter plot of log-expenditure and income with regression line (red) and LOWESS smooth (blue), b: Scatter plot of log-expenditure and log-income with both regression line and LOWESS smooth.

After the log-transformation was applied to income, Figure 1b. shows a plot with values spread out. However, cluster between 9 and 10 in log-income can still be seen implying another transformation should be made. Looking at Figure 1b. LOWESS smooth in blue, there are two peaks approximately at 8.7 and 9.6 log-income. This indicates that log-income follows a polynomial cubic function. Therefore, a cubic polynomial transformation on log-income was considered.

There are a few trade-offs of complexity vs. functional fit. With complexity transformation, the pattern would more likely to be linear. However, this complex transformation has multiple steps and transformation which would confuse the results of the regression if we don't interpret it carefully. The functional fit alone would not solve the clustering issue. However, with one transformation, the regression would be simpler.

During the model development process, we can access the significant of the variable income by fitting the regression and check p-values and confident interval. If the p-value and confident interval indicate that income is significant, then we would also validate the model by using R-square or MSE to see how accurate income would predict.