Descriptive Statistics & Data Quality Check

The dataset consists of 5,526 records with 17 explanatory variables and 1 target variable. As part of the quality check, the data was analyzed for missing values and outliers. The Interquartile Range (IQR) technique was employed for outlier detection. For each variable, observations lying outside the interval (Lower Limit, Upper Limit) were treated as outliers. The limits are defined as:

* **Lower Limit** = Q1 - 1.5 × IQR
* **Upper Limit** = Q3 + 1.5 × IQR

The table below (Table Number) presents the distribution of missing values and the number of outliers detected for each variable.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Since the data contains outliers, missing values were imputed using the median of the respective variable. Outliers were then addressed by replacing them with the corresponding lower and upper limits based on the following conditions:

1. If the original data point is **greater than or equal to the Upper Limit**, it is replaced with the Upper Limit.
2. If the original data point is **less than or equal to the Lower Limit**, it is replaced with the Lower Limit.

To assess multicollinearity in the dataset, both bivariate and multivariate statistical tests, including Variance Inflation Factor (VIF) analysis and Correlation Analysis, were conducted. In the Correlation Analysis, explanatory variables with an absolute correlation coefficient greater than 0.30 were identified for further investigation.

Additionally, the VIF test was applied to all variables to quantify multicollinearity. A higher VIF score indicates a greater likelihood of multicollinearity. A VIF score exceeding 10 was used as the threshold for identifying multicollinearity. None of the variables in the dataset exhibited a VIF score greater than 10, indicating that multicollinearity is not a concern.

For each variable, a binning method was applied, and the Weight of Evidence (WoE) and Information Value (IV) were calculated accordingly. The table below provides the IV values for each variable.

The following table (Table Number) presents the characteristics of the IV values. Based on the magnitude of the IV values, weak variables such as *"Operating Cashflow/CY Repayment Obligation"* and *"Trend in Operating Cash Flow"* were removed from the dataset.

Additionally, the shape of the WoE for each variable was analyzed, and variables with suspicious patterns were excluded. For example, the *Fixed Asset Ratio (Turnover/Net Fixed Assets)* had an IV value of 0.21; however, it exhibited a V-shaped trend in default rates, which is inconsistent with expected behavior. Therefore, it was removed from the dataset.

In addition to IV values and the shape of WoE, expert judgment was also considered during the variable selection process. For instance, despite *Sales Growth* having a strong IV value, the occurrence of negative sales growth does not align with business logic in binning. As a result, *Sales Growth* was excluded from the dataset.

The remaining set of variables is as follows: