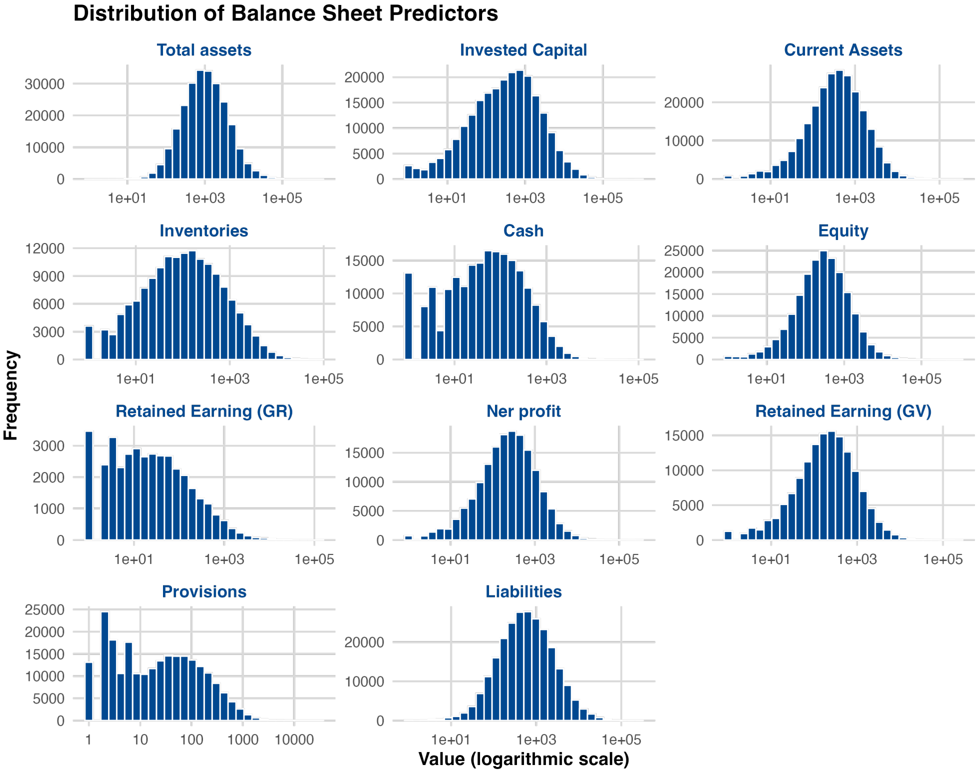
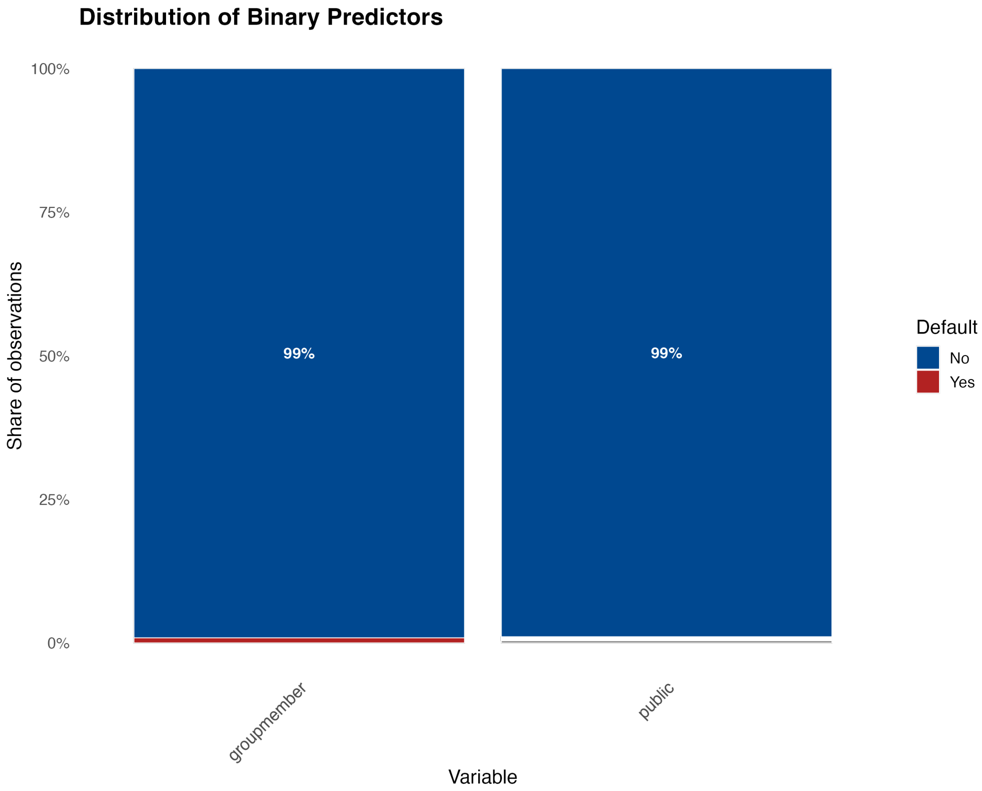
Data exploration

1.2 Distribution and bivariate Data Analysis

1.2.1 Distributions

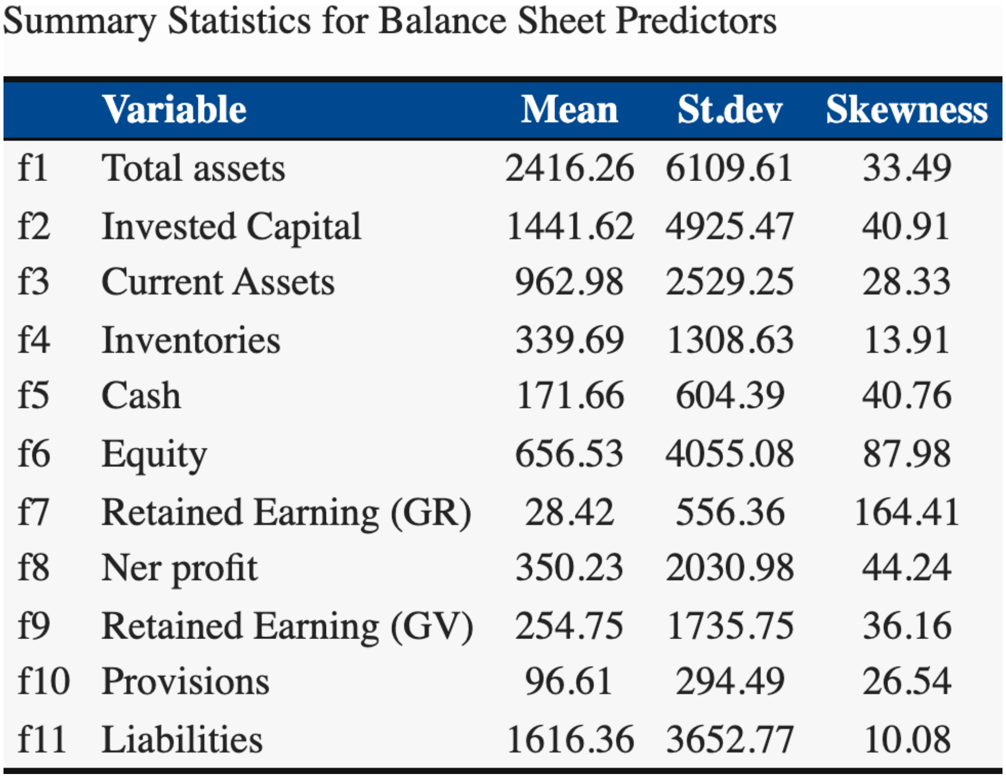


The histograms on the logarithmic scale show that all Balance Sheet Predictors are heavily right-tailed and spread across several orders of magnitude. Without the log scale, the values would be too uneven to visualize properly. After applying the log scale, the distribution of most variables reminds a log-normal distribution. Most observations are relatively small, while a small number of very large values stretch the distributions far to the right.



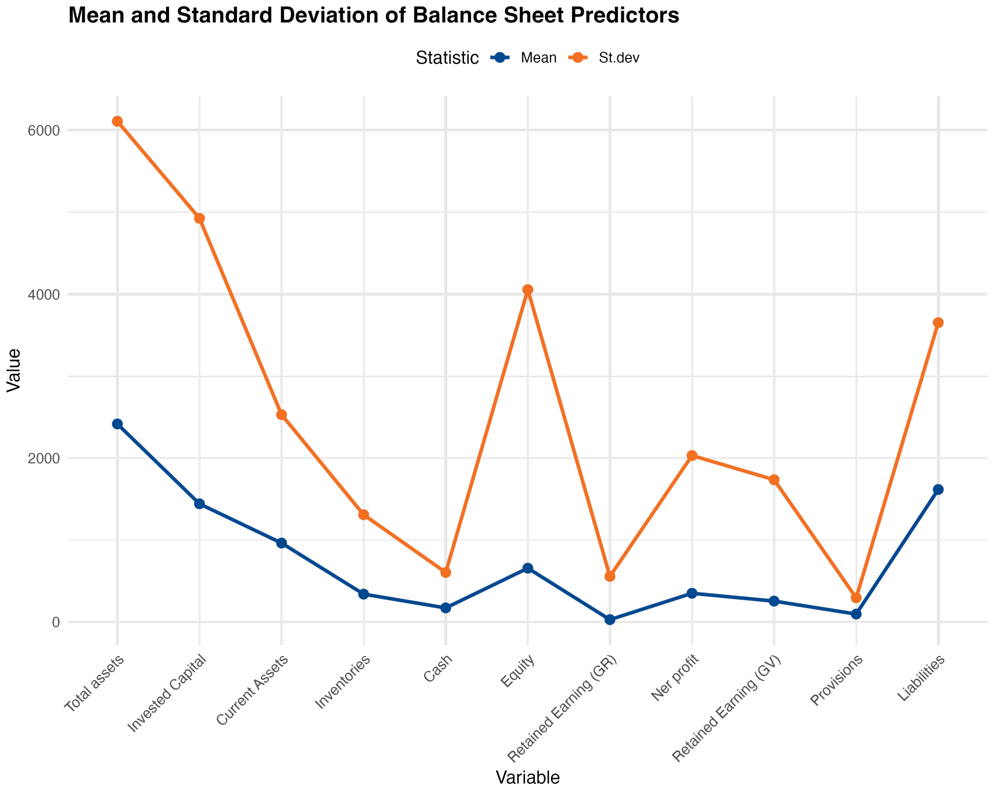
From the distribution plot of binary variables, we see that both group member and public are extremely imbalanced.

2. Summary statistics



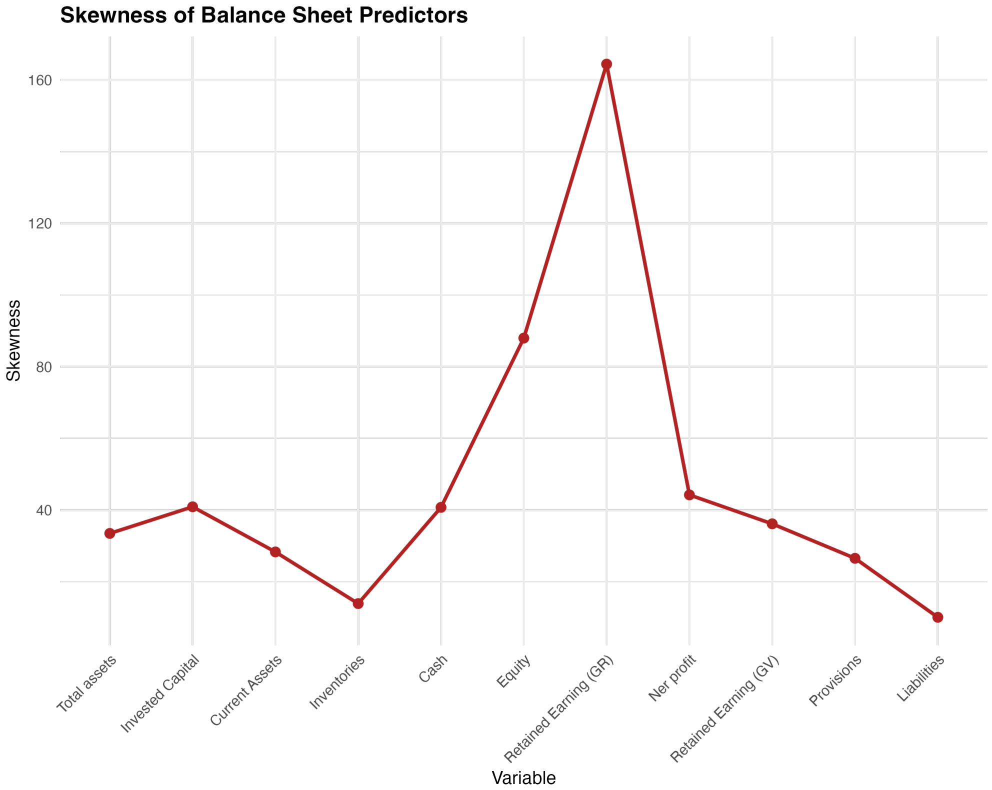
Summary Statistics for Balance Sheet Predictors

	Variable	Mean	St.dev	Skewness
f1	Total assets	2416.26	6109.61	33.49
f2	Invested Capital	1441.62	4925.47	40.91
f3	Current Assets	962.98	2529.25	28.33
f4	Inventories	339.69	1308.63	13.91
f5	Cash	171.66	604.39	40.76
f6	Equity	656.53	4055.08	87.98
f7	Retained Earning (GR)	28.42	556.36	164.41
f8	Ner profit	350.23	2030.98	44.24
f9	Retained Earning (GV)	254.75	1735.75	36.16
f10	Provisions	96.61	294.49	26.54
f11	Liabilities	1616.36	3652.77	10.08



Mean and Standard Deviation of Balance Sheet Predictors

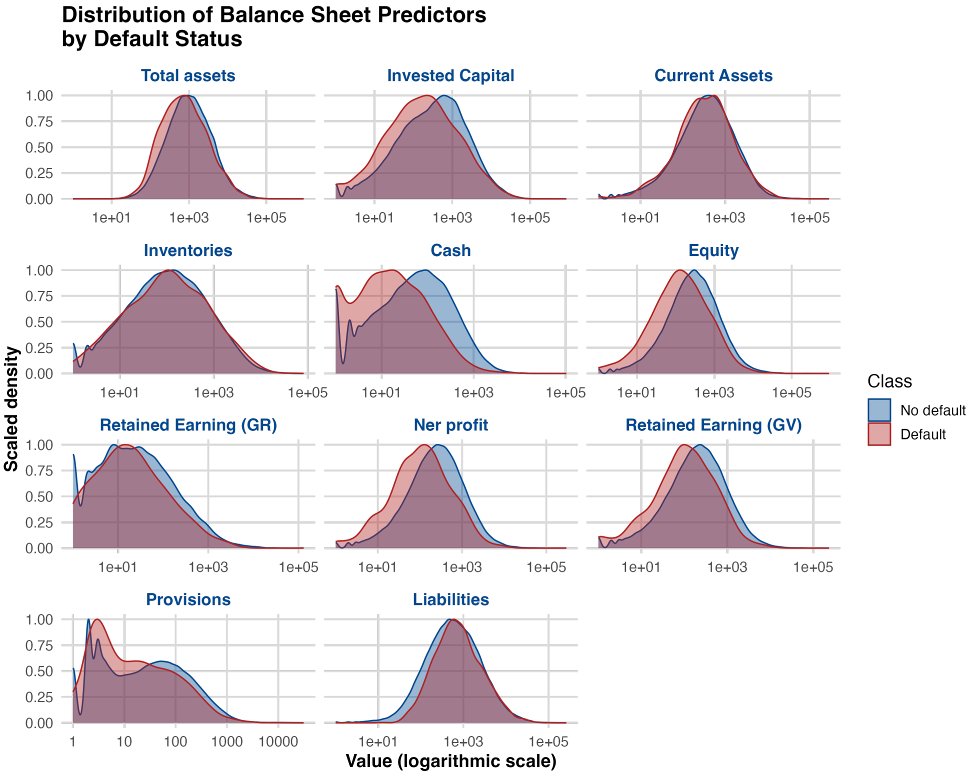
Mean and Standard Deviation Since the standard deviation is larger than the mean for every Predictor from the Balance Sheet. there is high variability within each variable. Variables: total assets (f1), invested capital (f2), equity (f6), net profit (f8), retained earning (f9) and Liabilities (f11) demonstrate especially large spreads, pointing to extreme values in the upper tail. Variables, such as cash (f5), retained earning (GR) (f7), and provisions (f10) show considerably lower difference between mean and standard deviation.The data is not centered around a typical value and includes many extreme observations that increase the overall spread.



Skewness

All variables show positive skewness, meaning they have long right tails. Equity (f6) and retained earnings (GV) (f7) stand out with very high skewness values, indicating that a few very large observations dominate the shape of the distribution. Inventories (f4) and liabilities (f11) are the least skewed but still clearly asymmetric. This consistent right-skewness across all predictors matches what we see in the plots: most values are concentrated near the lower end, with a small number of extreme values pushing the distributions far to the right.

1.2.2 Data Dependency



The plot shows the distribution comparison for Balance Sheet Predictors. We applied scaled density to properly visualize the data, without scaling the distribution of the default class would be mostly invisible due to the class imbalance.

Obvious separation:

1. Invested capital (f2). Defaulted companies tend to have lower invested capital on average compared to non-default companies. Default distribution has a slightly heavier right tail, meaning a few defaulting customers have unusually high values of invested capital.

2. Cash (f5), Net profit (f8), Retained Earning (GV) (f9). Defaulted companies tend to have lower cash, net profit and retained earning on average compared to non-default companies.

1.3 Feature selection

1.3.2 Significance test

Numerical variables

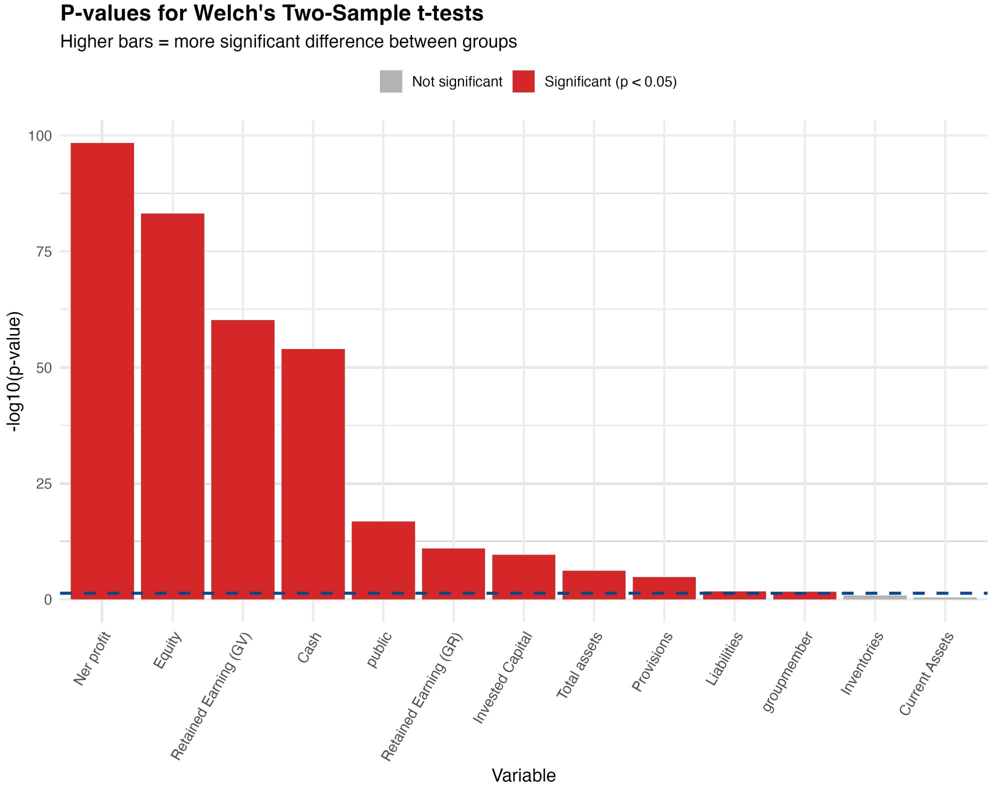
By calculating two group means and doing Welch's Two-Sample Test, we check weather each variables differs systematically between the two groups y=0 and y=1. For example if the group mean differ substantially, it indicated that x_j is associated with the response variable y and it could potentially be a useful predictor in a logistic regression, tree or any classification model.

To determine the difference in the group mean, we conduct two-sample t-test

H0: $E[x_j|y = 0] = E[x_j|y = 1]$

H1: $E[x_j|y=0] \neq E[x_j|y=1]$

We create a table and order the features by the lowest p-value. It means that its group means differ the most relative to the within-group variability.



The plots evaluated how well each predictor distinguish between the two groups. We use -log(10) in the charts for better visualization, taller bars indicate stronger difference between the group means.Low p-values indicate predictors that separate the classes and are useful in a model; high p-values indicate weak predictors that do not contribute meaningful information. Low p-value means that the variable contains a useful signal (information that meaningfully differs between groups) and can potentially improve the model's ability to predict the target.

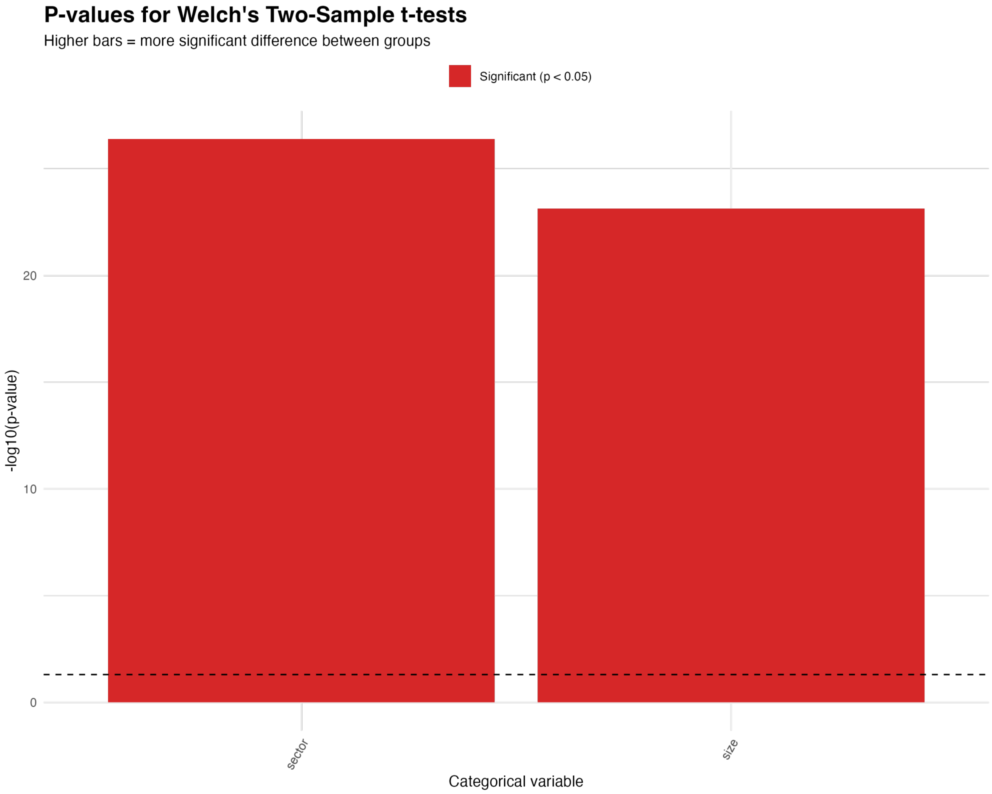
Variables with substantially different group means are Net profit, Equity, Retained Earnings (GV) asd Cash.

Categorical Variables

The analysis helps us understand whether the categorical variable is associated with default.For each categorical predictor, construct a contingency table against the binary outcome (default vs non-default) and performed a chi-squared test of independence. Variables with p-values below 0.05 were considered significantly associated with default, indicating that the distribution of default events differs across their categories.

To check whether category distributions differ between groups, we conduct a Chi-squared test

From the results (p-values being < 0.5) we see that there's strong association of categorical variables and the outcome y. Distribution differs significantly between the two groups.



The distribution of classes differs significantly across sectors and across size categories, which means that both variables have a useful signal for distinguishing between default and non-default observations.

1.3.3.b VIF

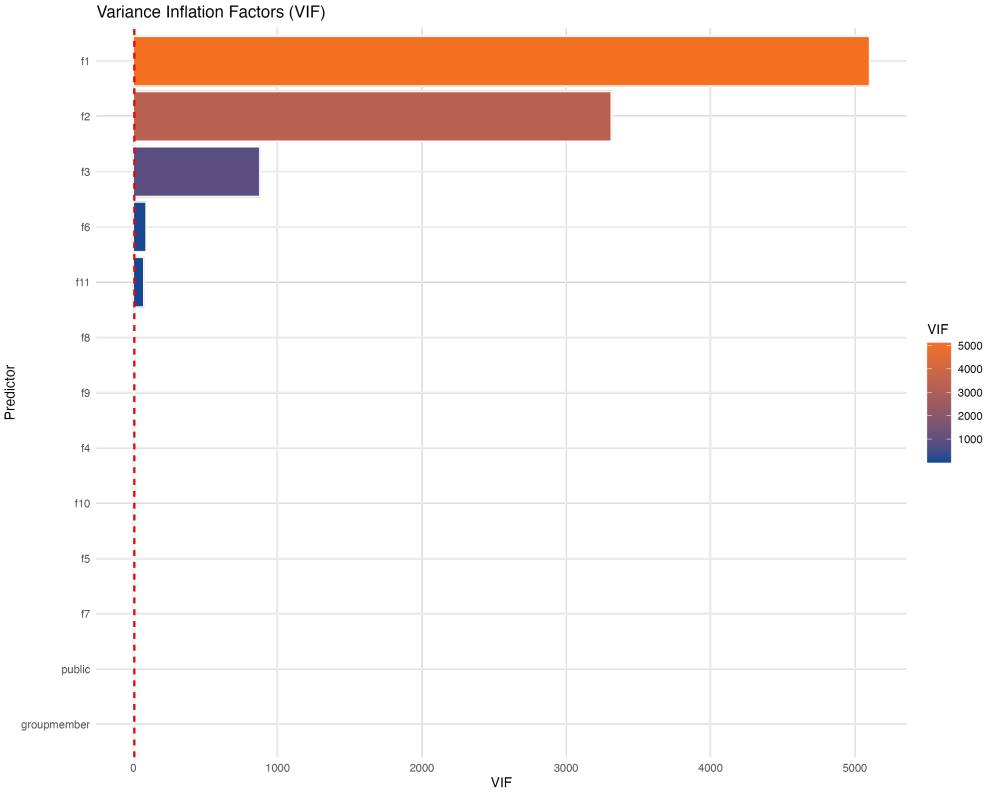
1.Regress each predictor X_j on all the other predictors:

$$X_j = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon.$$

2. Obtain the R_j^2 of this regression:
   - If $R_j^2 = 0$: the predictor is independent (no multicollinearity).
   - If $R_j^2 = 1$: the predictor is perfectly explained by the other predictors (full multicollinearity).
3. Compute the variance inflation factor (VIF):

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}.$$

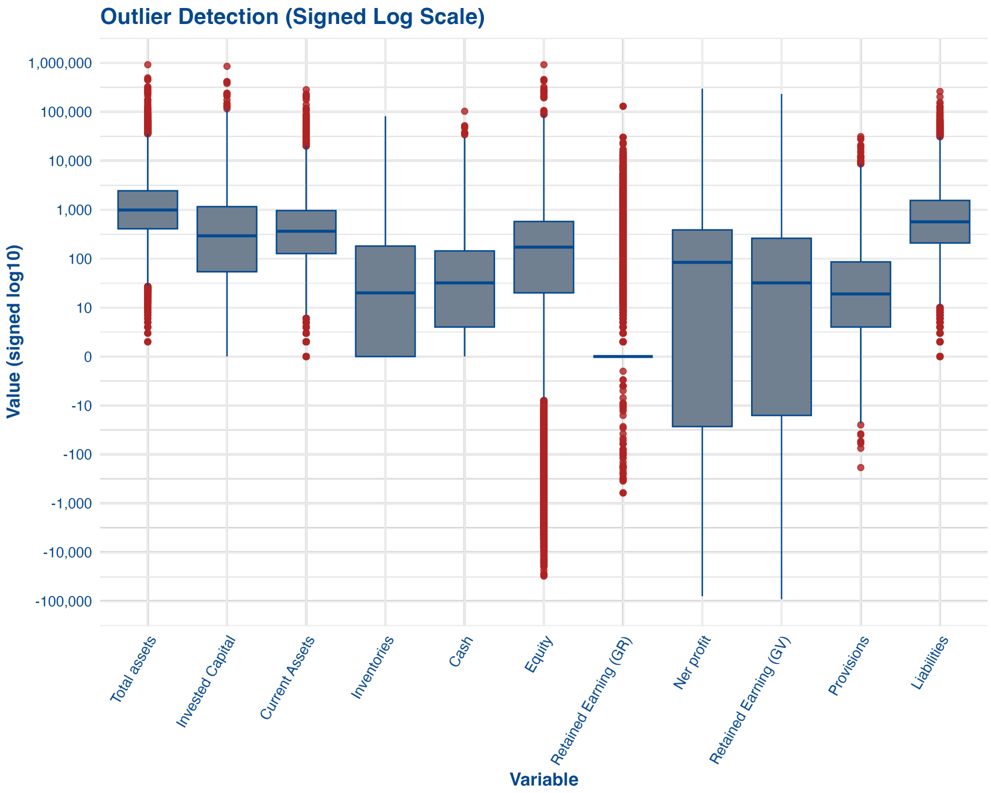
- Variance Inflation Factor (VIF) quantifies how much the variance of a regression coefficient is increased due to multicollinearity.
- A VIF of 1 indicates no correlation with other predictors, while higher values indicate that the predictor can be linearly explained by other predictors.
- VIFs above 5 suggest moderate multicollinearity, and values above 10 indicate severe redundancy and unstable parameter estimates.



Total assets, invested capital,current assets and cash show the highest VIF, which is obvious beacause total assets is a linear combination of all others.

1.5 Outliers

1.5.1 Outliers detection



Since the data set has negative values and zeros, we used signed log transformation to visualize the outliers.

- $\text{signedlog}(x) = sign(x) \cdot \log_{10}(1 + |x|)$.

The transformation doesn't change the rank or ordering of the data, it only rescales it. Signed log keeps negative values negative, but compresses their magnitude in the way as positive values. **Boxplots interpretation:**

All predictors from the Balance Sheet are highly skewed and heavy-tailed distributed. Equity (f6), Retained Earnings GR (f7) have dense clusters of extreme values. Presence of numerous outliers suggests considerable variability in the underlying Balance Sheet Data.

1.5.2 Robust scaling

One of the ways to handle the outliers is to use **Robust scaling method**.

1. Robust Scaler transforms centers the variable and IQR (interquartile range) scales the spread.

$$x_{scaled} = \frac{x - \text{median}(x)}{\text{IQR}(x)}$$

. ‘ This way, it reduces the influence of extreme values without removing them, preserves the underlying structure. The method is suitable for using in tree-based models or boosting.

1.5.3 Weight of Evidence

WOE is a binning-based transformation primarily used in binary classification.

WOE is calculated as:

$$\mathrm{WOE} = \ln\left(\frac{\%\mathrm{Good}}{\%\mathrm{Bad}}\right)$$

.

A large positive WOE means that the bin is strongly associated with non-dafeults, while a large negative value demonstrates the opposite. After calculating WOE, the original value is replaced by WOE of the bin it belongs to. The method is suitable for using in glm models and iff the outcome is binary.