Corporate Finance Feature Engineering (1996–2022)

Vidit Pokharna

September 9, 2025

Contents

1 Data and Preprocessing

- **Source:** Compustat annual fundamentals; identifiers `GVKEY` (firm), `FYEAR` (fiscal year)
- **Sample:** FY 1996–2022 (inclusive)
- **Engineering (examples):** TOTAL_DEBT = `DLTT` + `DLC`, QUICK_ASSETS = `ACT` − `INVT`; liquidity, leverage, profitability, efficiency, and cash-flow coverage ratios
- **Recession flag:** Annual NBER recessions set to 1 in {2001, 2008, 2009, 2020} (`USREC`); 0 otherwise
- **Missingness:** Pairwise deletion for correlations; median imputation for PCA
- **Outliers:** Light winsorization (1%/99%) when scaling; log1p for highly skewed nonnegative features

2 Qualitative Feature Selection

Feature	Rationale
Liquidity	
`cash` (CHE)	Liquidity buffer; firms hoard cash in stress episodes. Heavy right tail.
`short_term_investments` (IVST)	Near-duplicate of cash for many firms; often zero-inflated.
`quick_ratio`	Short-term solvency excluding inventory; preferred to current ratio.
`current_ratio`	Redundant with quick ratio (near-perfect correlation); candidate to drop.
Leverage / Solvency	
`total_liabilities` (LT)	Size & leverage scale; very right-skewed.
`debt_to_assets`	Balance-sheet leverage share; near-bounded in practice.
`debt_to_equity`	Sensitive to low/negative equity; long tails capture stress.
`interest_expense` (XINT)	Direct debt-service burden; rises with leverage and rates.
Profitability	
`sales` (SALE)	Scale of operations; right-skewed.
`cogs` (COGS)	Tracks sales almost 1:1; prefer margin; candidate to drop.
`oibdp`	Operating performance before D&A.
`net_income` (NI)	Bottom-line profitability; thick left tail in recessions.
`roa`	Profitability normalized by assets; cyclically sensitive.
Efficiency	
`asset_turnover`	Revenue per dollar of assets; sector-style dispersion.
Cash-flow Health	
`operating_cf` (OANCF)	Cash earnings; co-moves with resilience in downturns.
`capex` (CAPX)	Investment intensity; retracts in downturns.
`ocf_to_debt`	Coverage proxy; left tail signals distress.
Working Capital & Structure	
`accounts_receivable` (RECT)	Credit exposure to customers; expands with sales.
`inventory` (INVT)	Stock build/clear cycles; write-down risk.
`ppe_net` (PPENT)	Capital intensity; industry structure.
`shareholders_equity` (SEQ)	Capitalization; negative values flag stress.
`dividends` (DV)	Payout policy; omissions/cuts in recessions.

Collinearity decisions. Empirically, `current_ratio` $\approx$ `quick_ratio` ($r \sim 1.00$), `sales` $\approx$ `cogs` ($r \sim 0.97$), and `cash` $\approx$ `short_term_investments` ($r \sim 0.91$). We therefore **drop** `current_ratio`, `cogs`, and `short_term_investments` for modeling.

3 Faceted Histograms (1996–2022)

Each figure shows per-year histograms. Across time, most variables remain strongly right-skewed; spreads widen in 2001, 2008–2009, and 2020. Within year, mass clusters at low values with long right tails.

Feature-specific notes:

- *Liquidity (`cash`, `IVST`)*: higher centers post-crises (cash hoarding); `IVST` zero-inflated
- *Profitability (`net_income`, `roa`)*: left tails thicken in recessions
- *Leverage (`debt_to_equity`)*: very long tails when equity ≈ 0 or < 0
- *Payout (`dividends`)*: spike at zero in recessions (cuts/omissions)

Figure 1: Faceted histograms — `cash` (1996–2022)

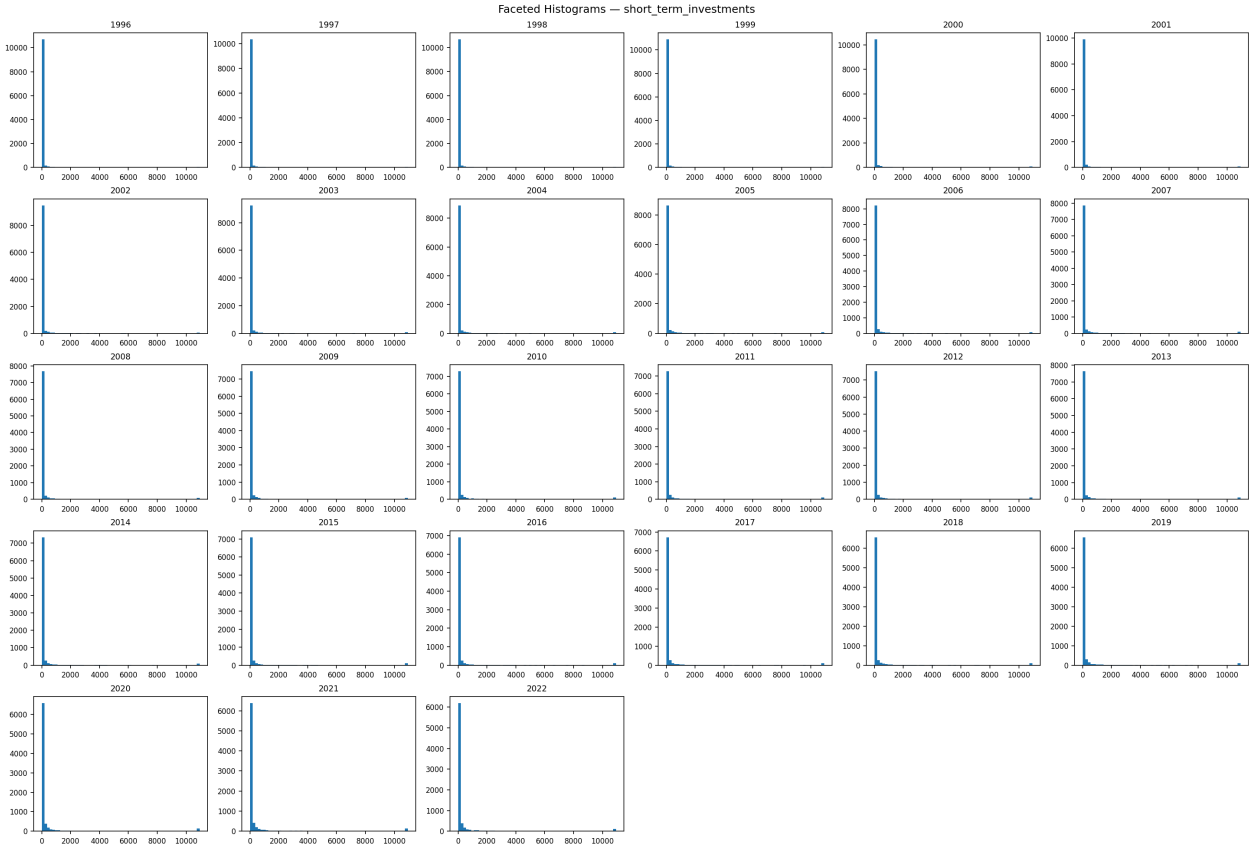

Figure 2: Faceted histograms — $\texttt{short}_{t}erm_{i}nvestments(1996 - -2022)$

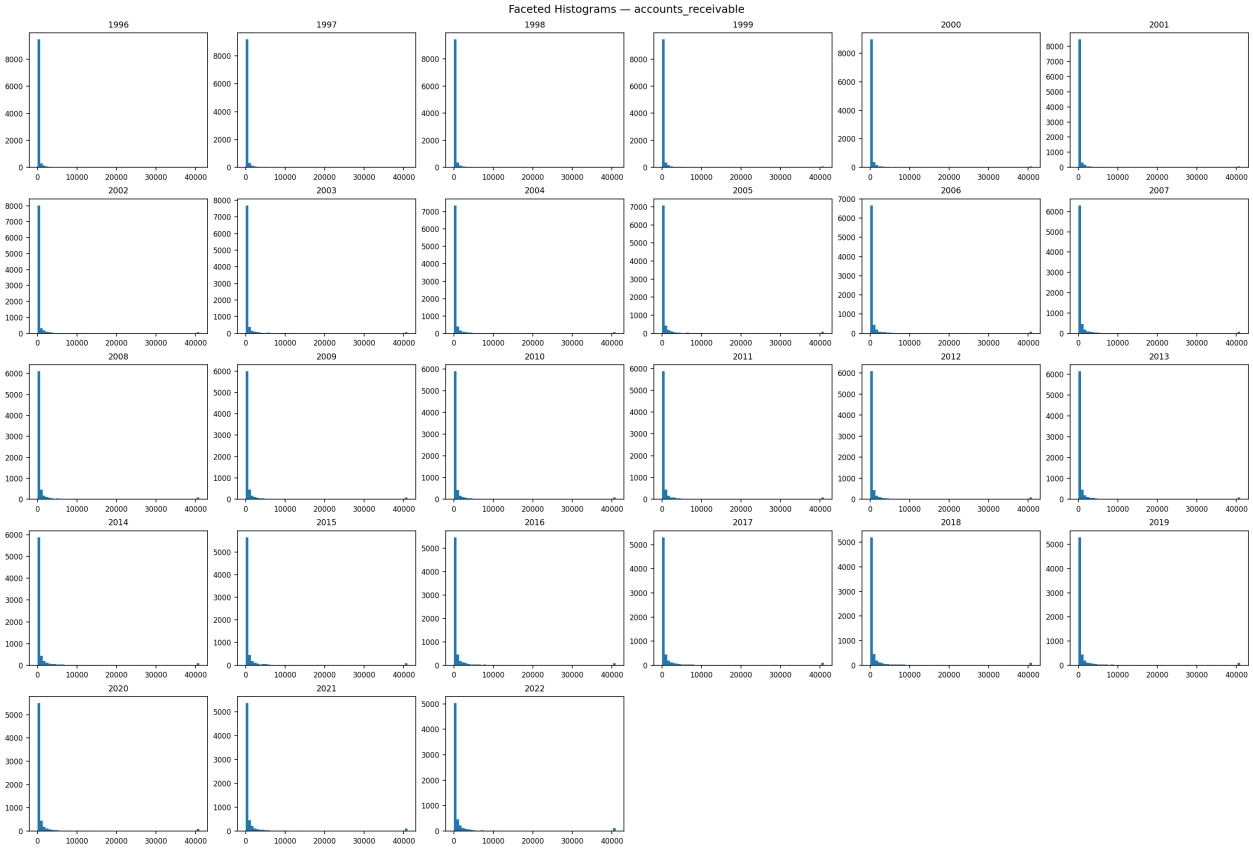

Figure 3: Faceted histograms — $\texttt{accounts}_r eceivable(1996 - -2022)$

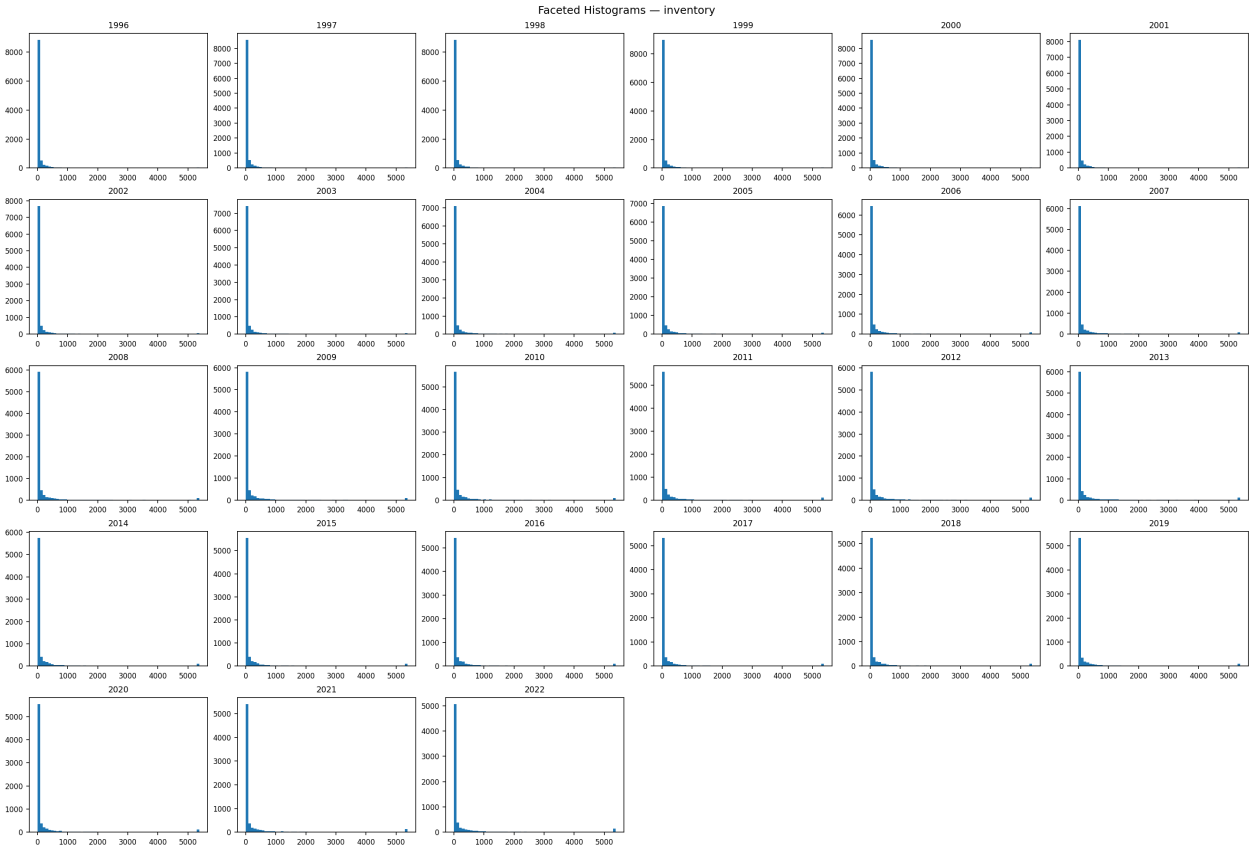

Figure 4: Faceted histograms — `inventory` (1996–2022)

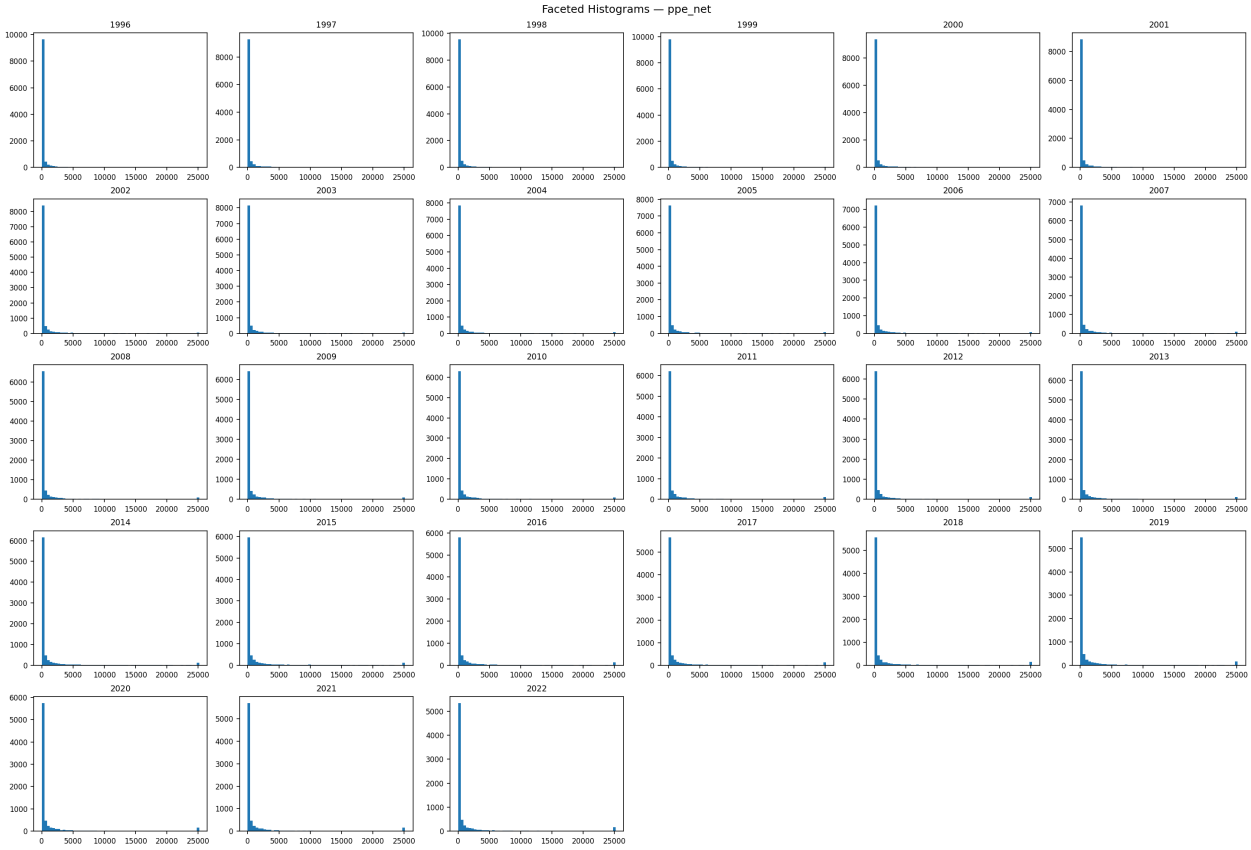

Figure 5: Faceted histograms — $\texttt{ppe}_net(1996 - -2022)$

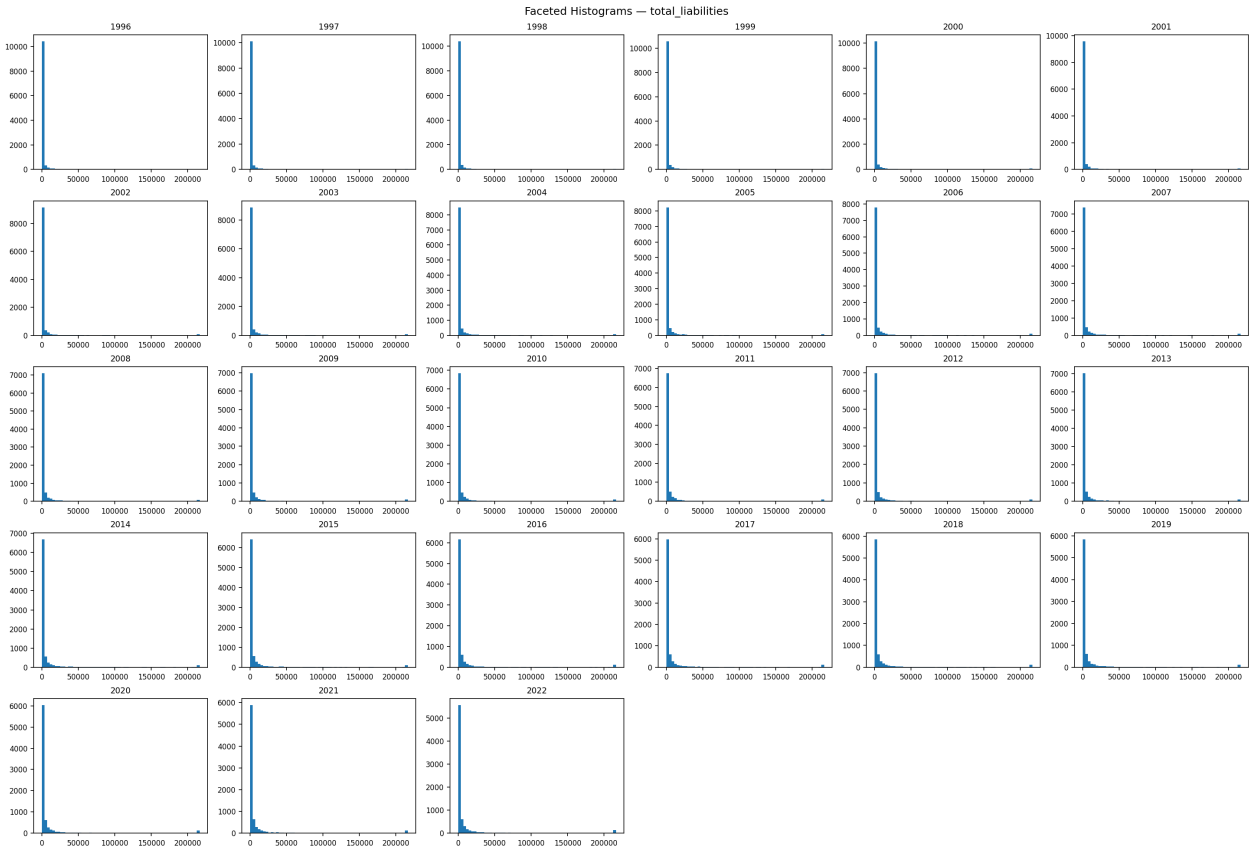

Figure 6: Faceted histograms — `total`$liabilities$(1996 – −2022)

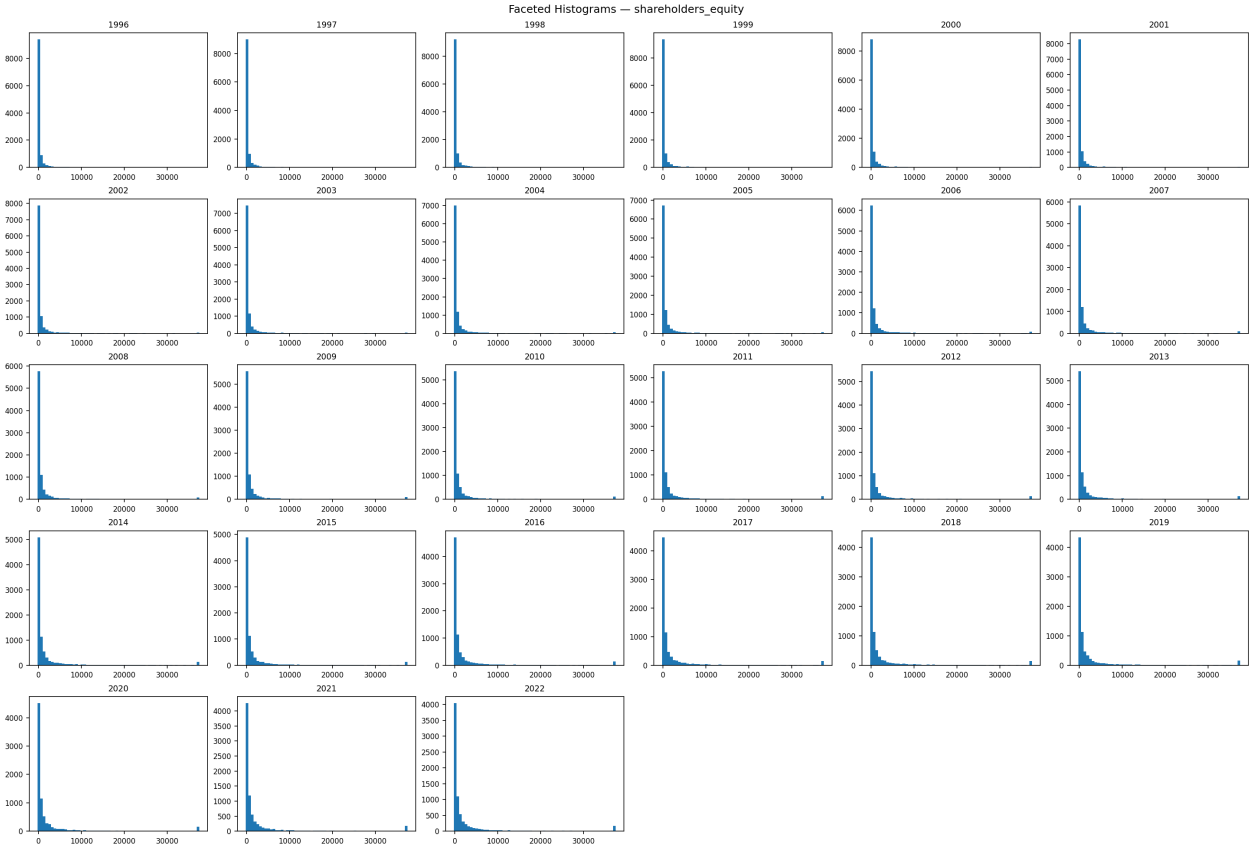

Figure 7: Faceted histograms — $\texttt{shareholders}_{e}quity(1996 - -2022)$

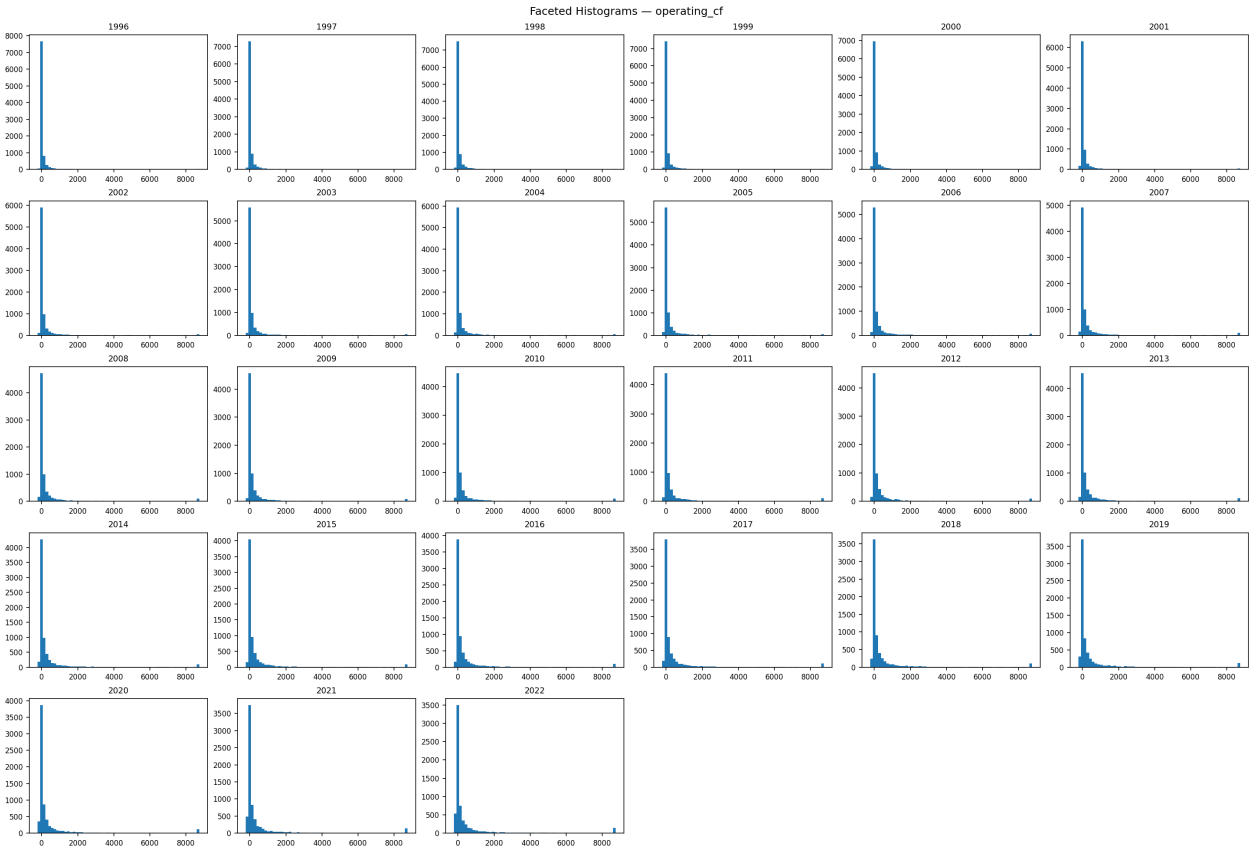

Figure 8: Faceted histograms — $\texttt{operating}_cf (1996 - -2022)$

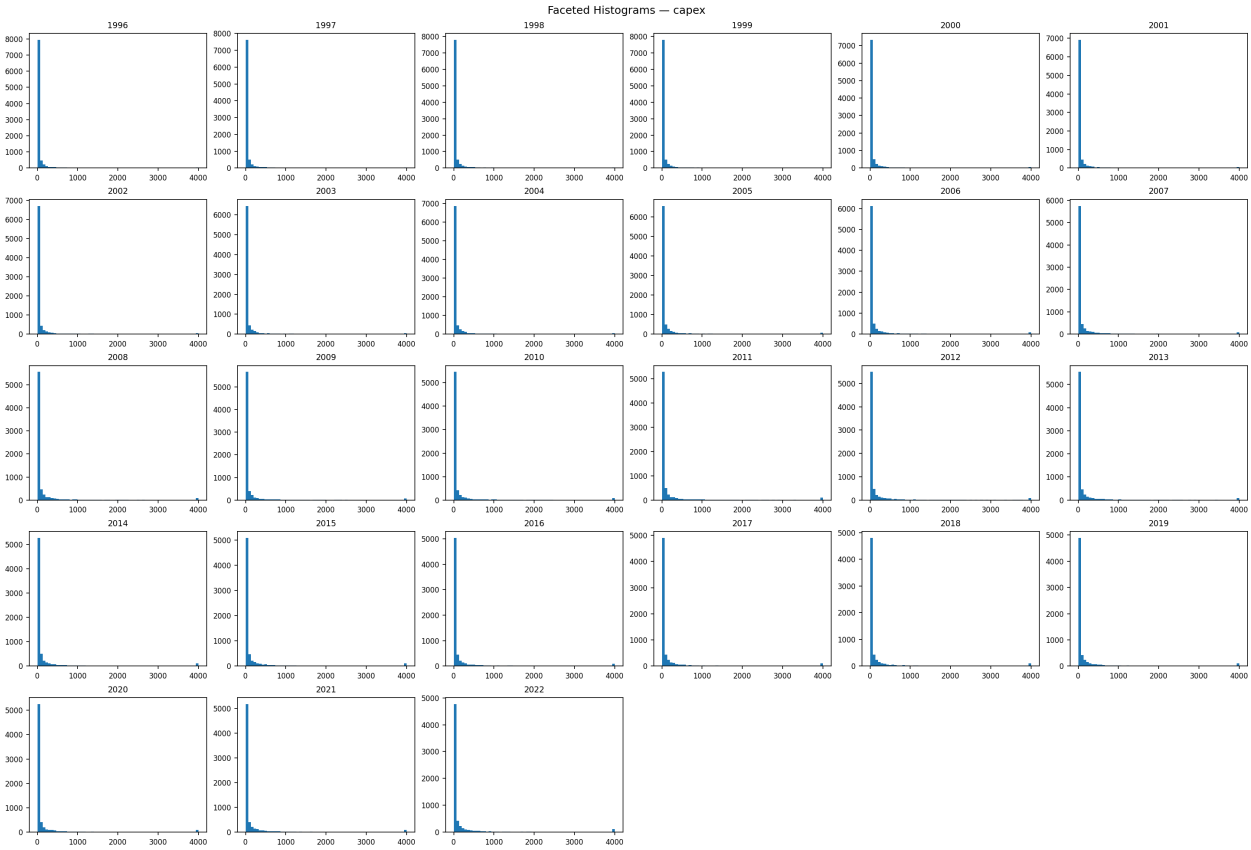

Figure 9: Faceted histograms — `capex` (1996–2022)

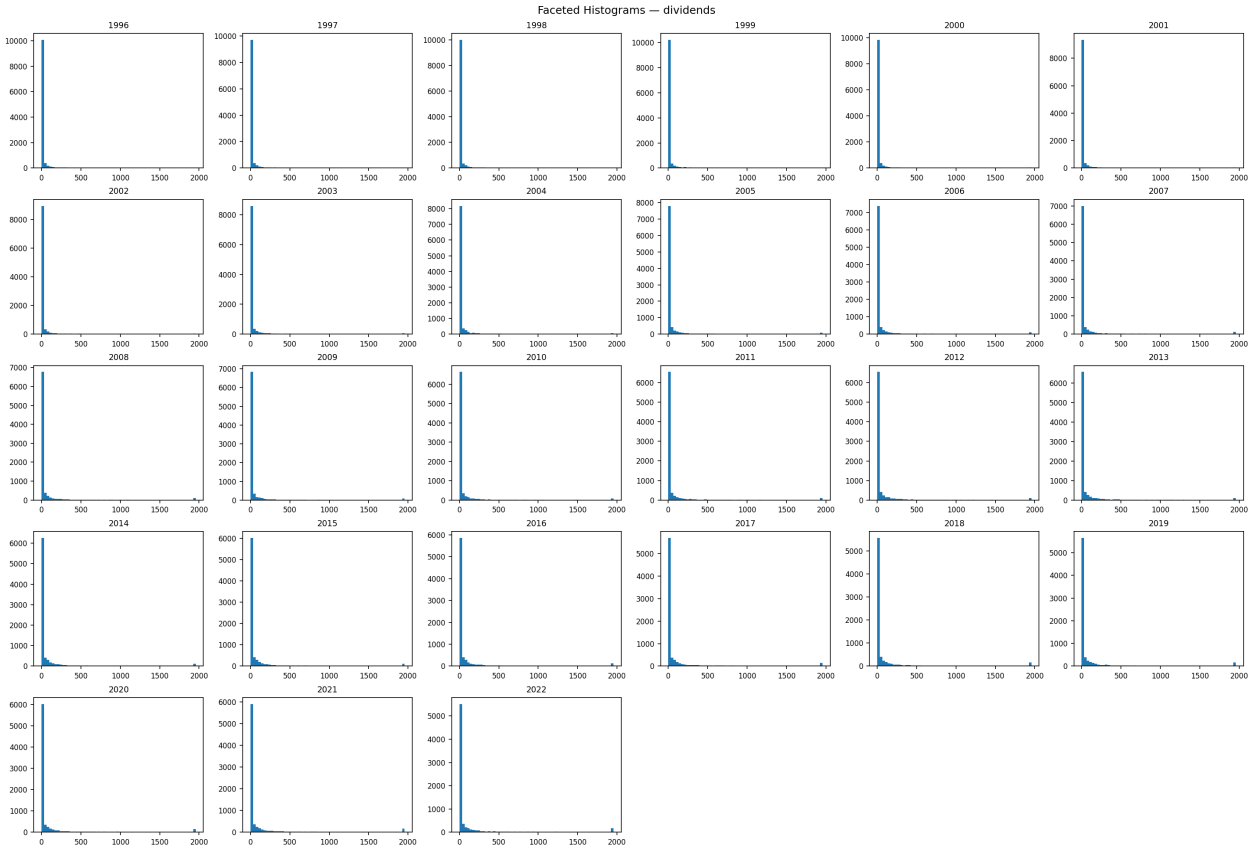

Figure 10: Faceted histograms — `dividends` (1996–2022)

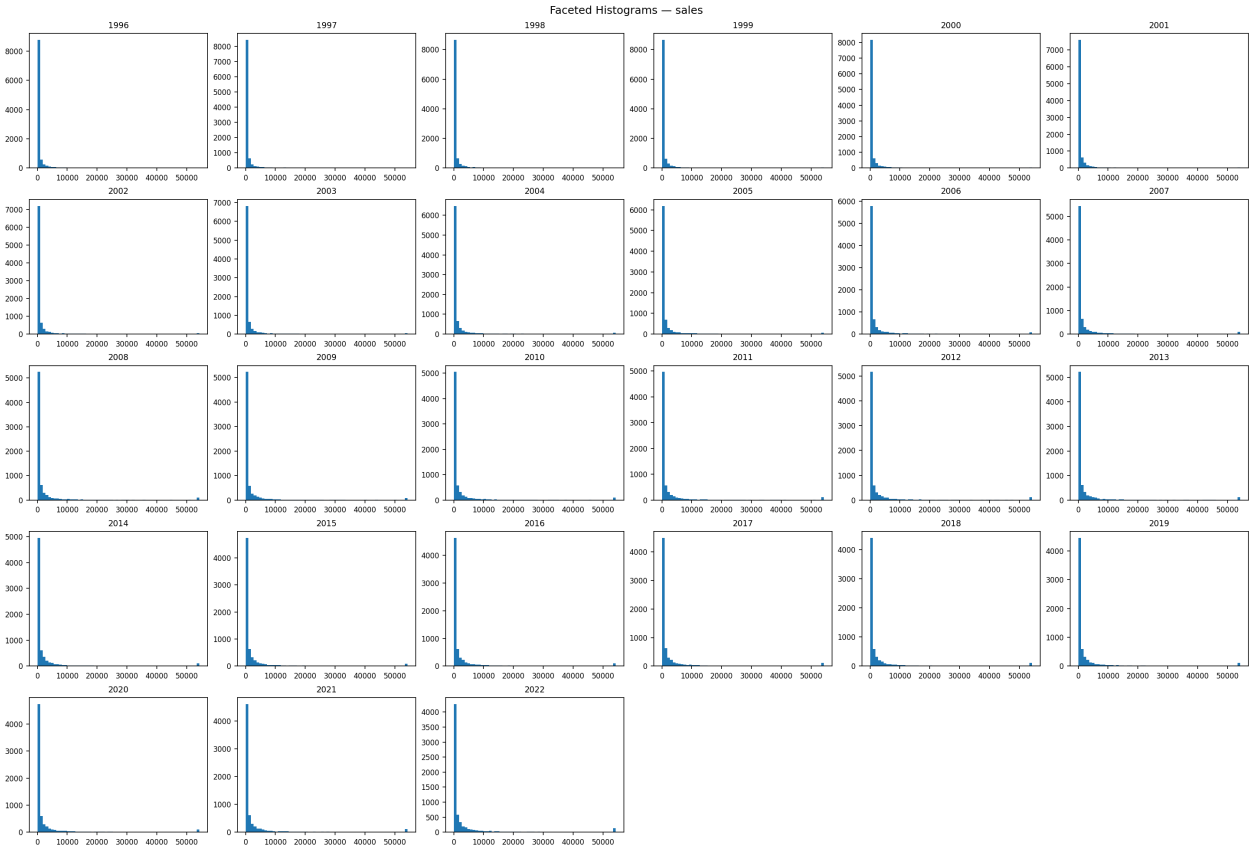

Figure 11: Faceted histograms — `sales` (1996–2022)

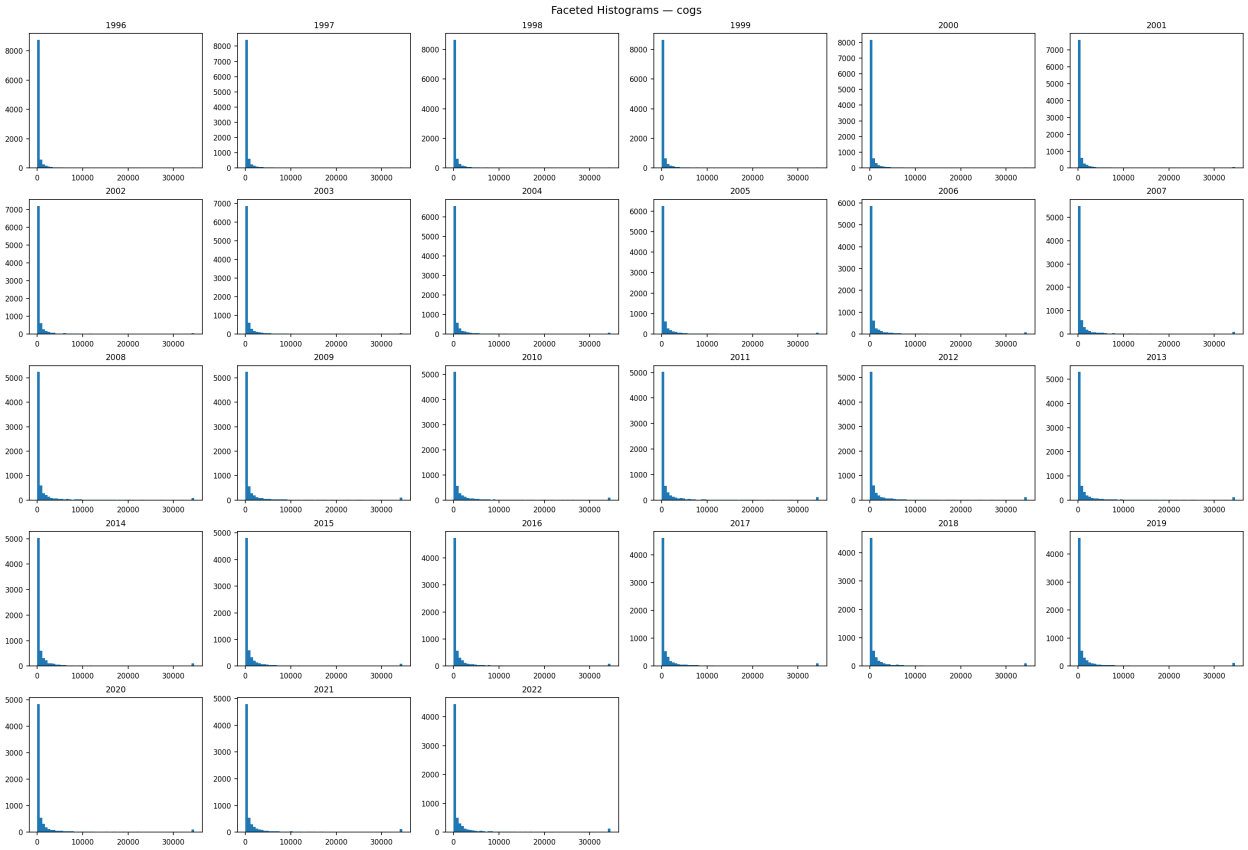

Figure 12: Faceted histograms — `cogs` (1996–2022)

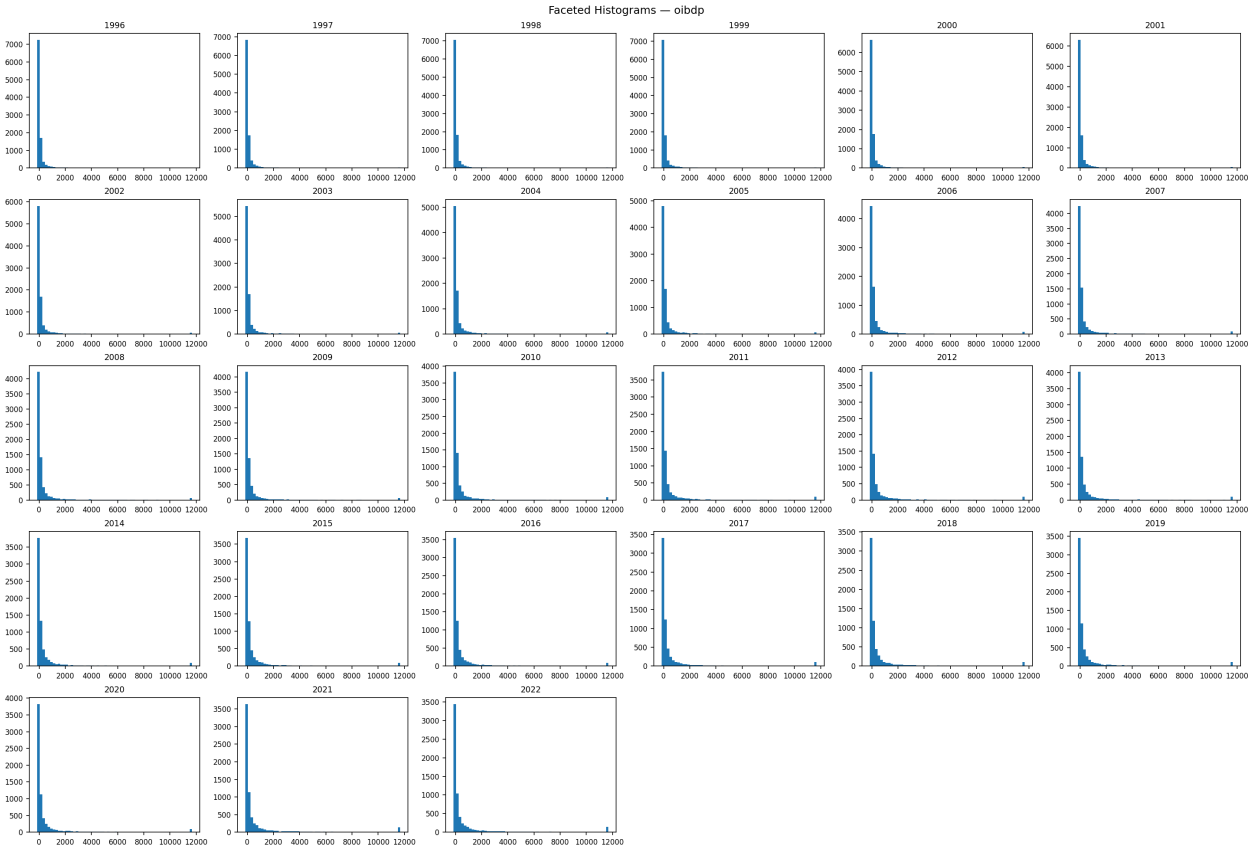

Figure 13: Faceted histograms — `oibdp` (1996–2022)

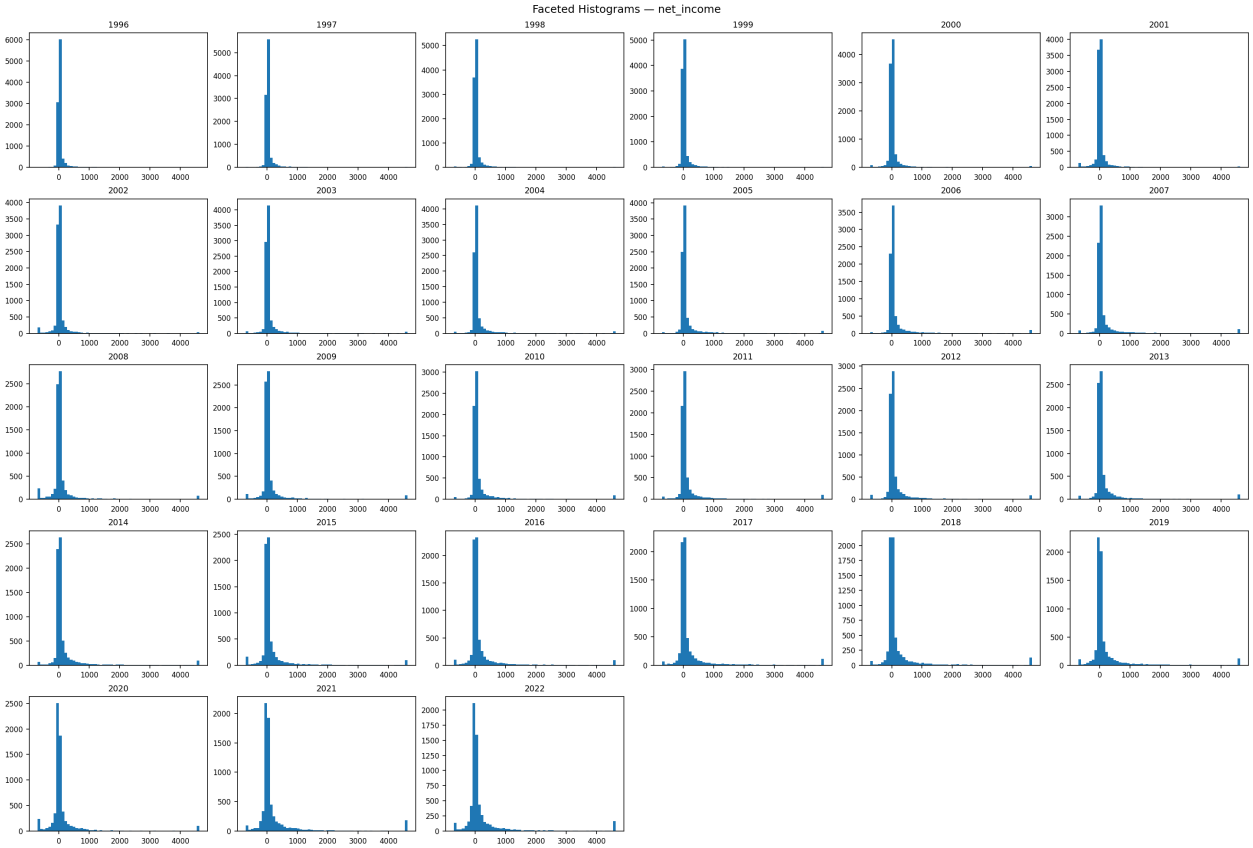

Figure 14: Faceted histograms — $\texttt{net}_i ncome (1996 - -2022)$

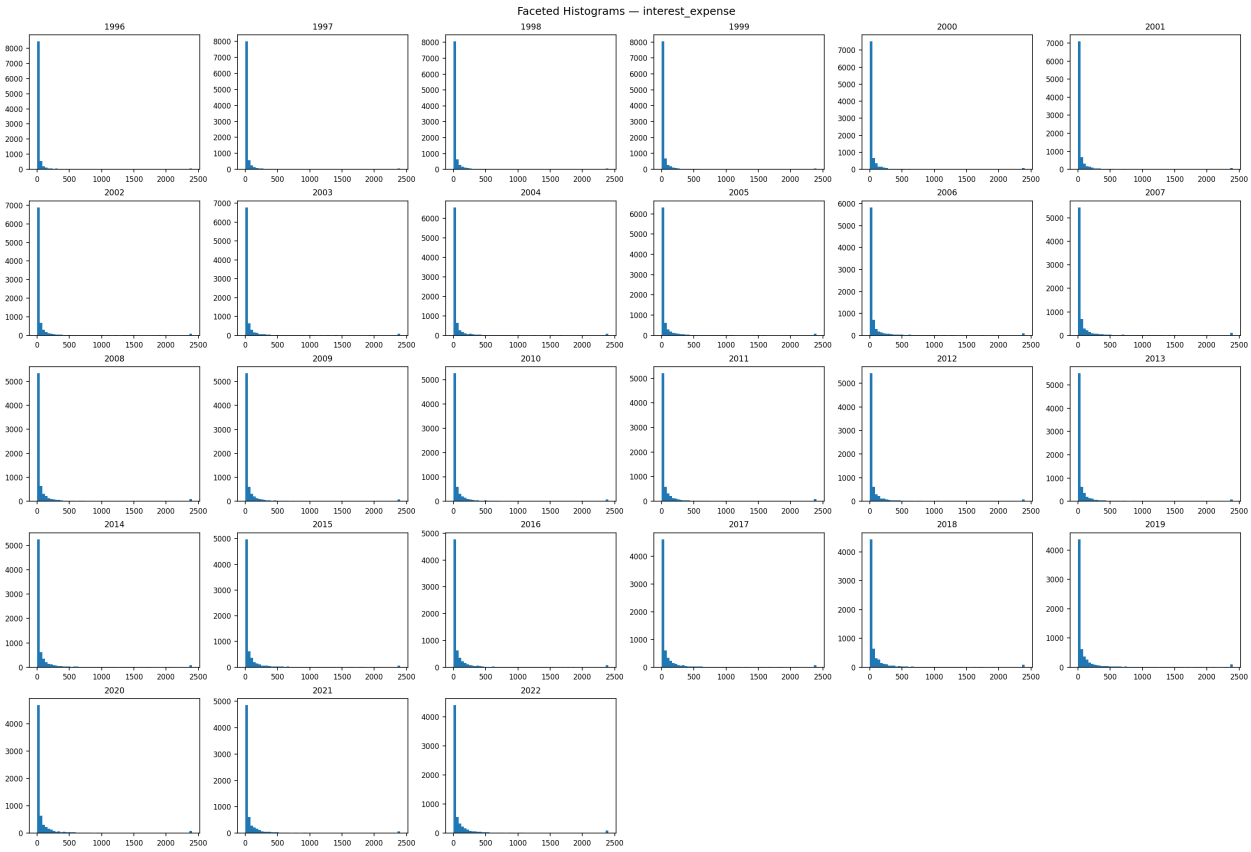

Figure 15: Faceted histograms — $\texttt{interest}_e xpense(1996 - -2022)$

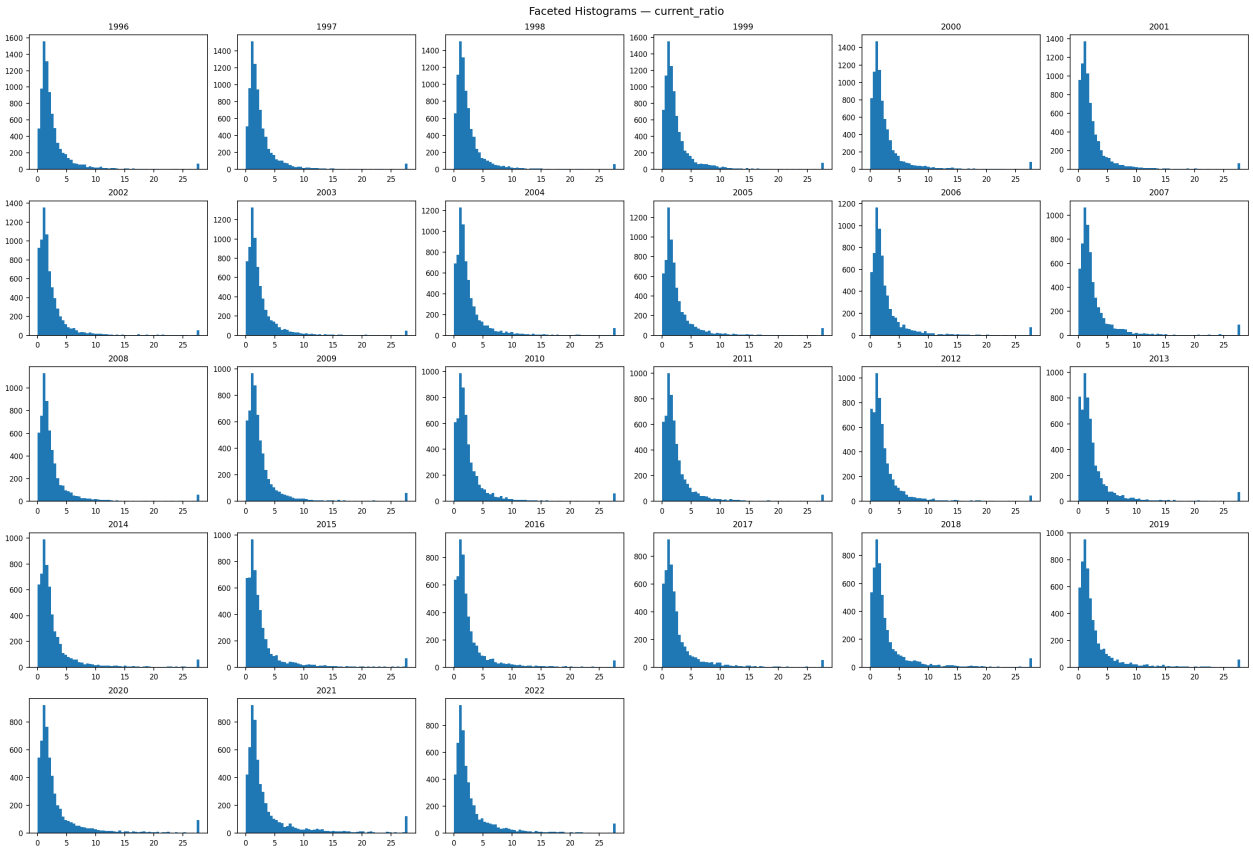

Figure 16: Faceted histograms — $\texttt{current}_ratio(1996 - -2022)$

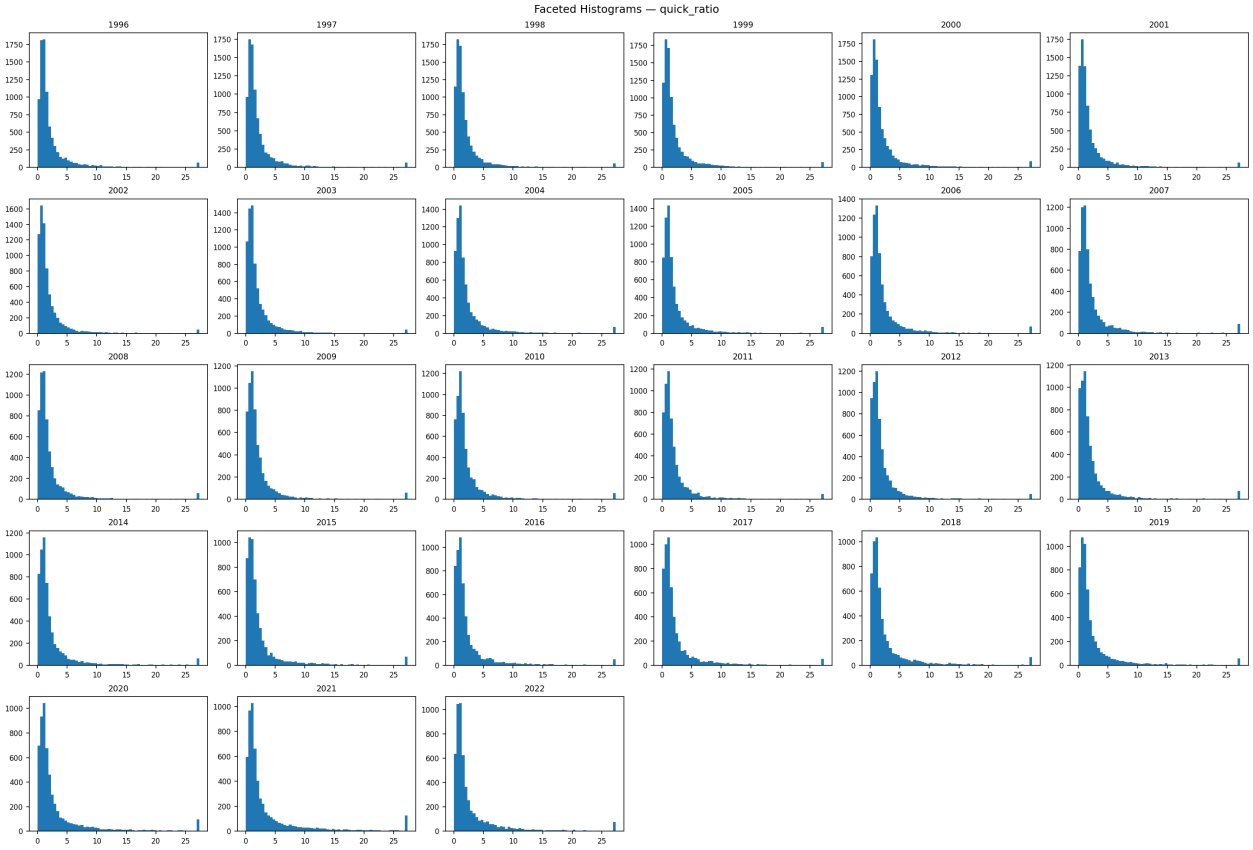

Figure 17: Faceted histograms — $\texttt{quick}_{r}atio(1996 - -2022)$

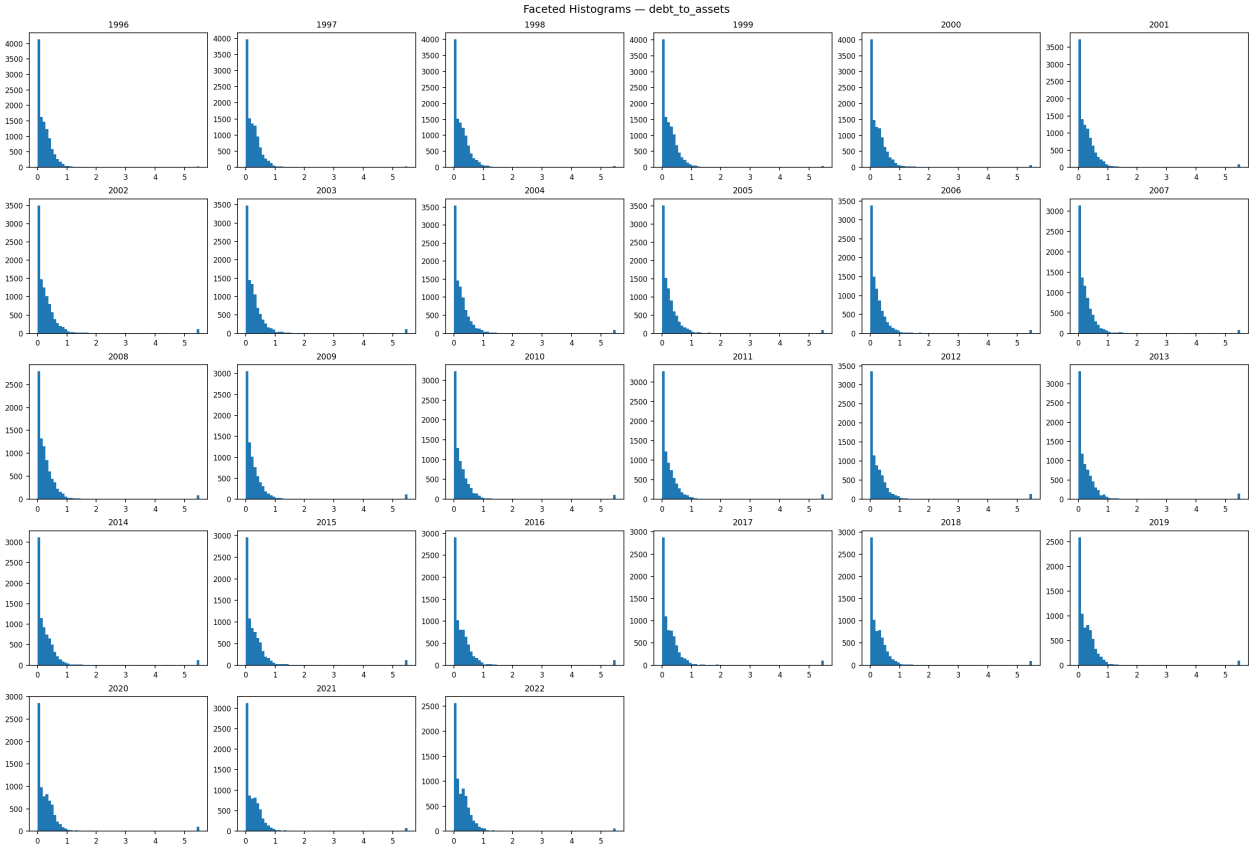

Figure 18: Faceted histograms — $\texttt{debt}_t o_a ssets (1996 - -2022)$

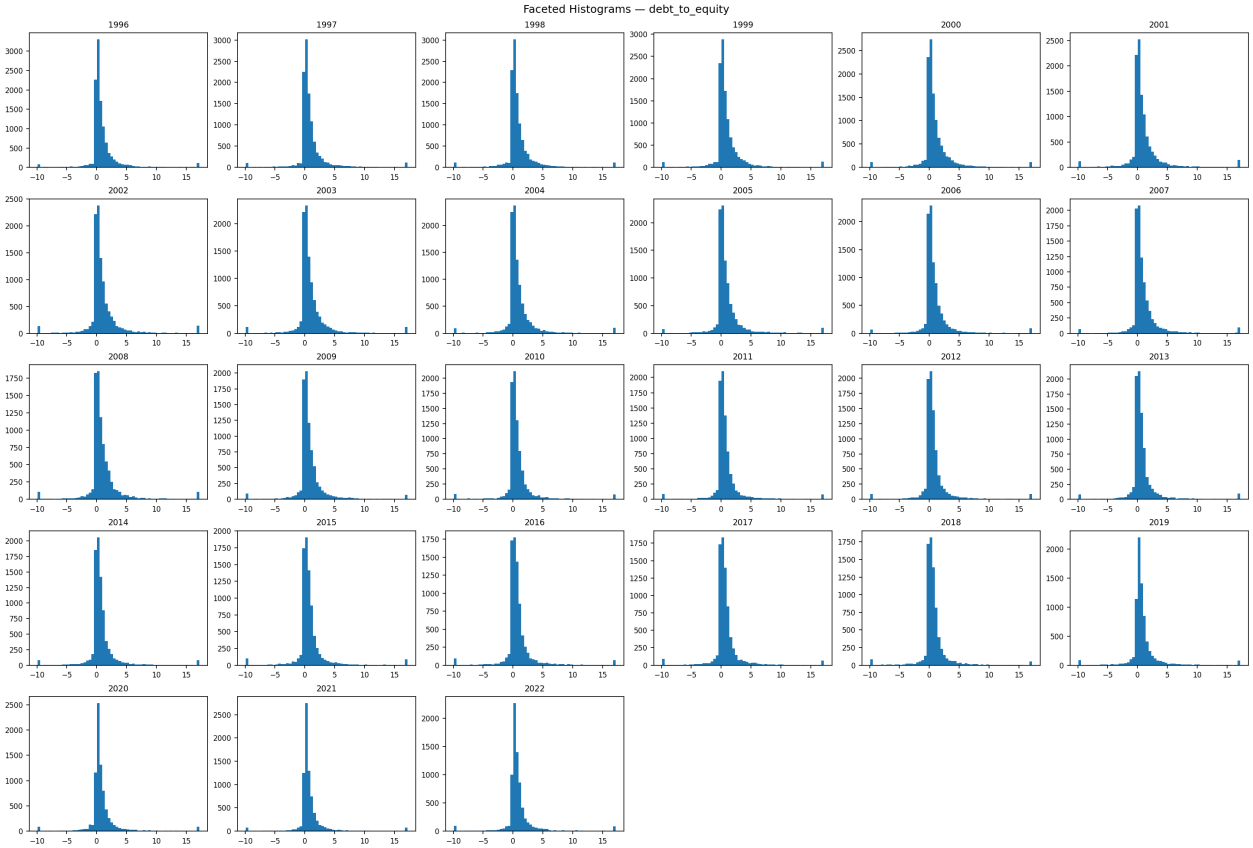

Figure 19: Faceted histograms — $\texttt{debt}_t o_e quity(1996 - -2022)$

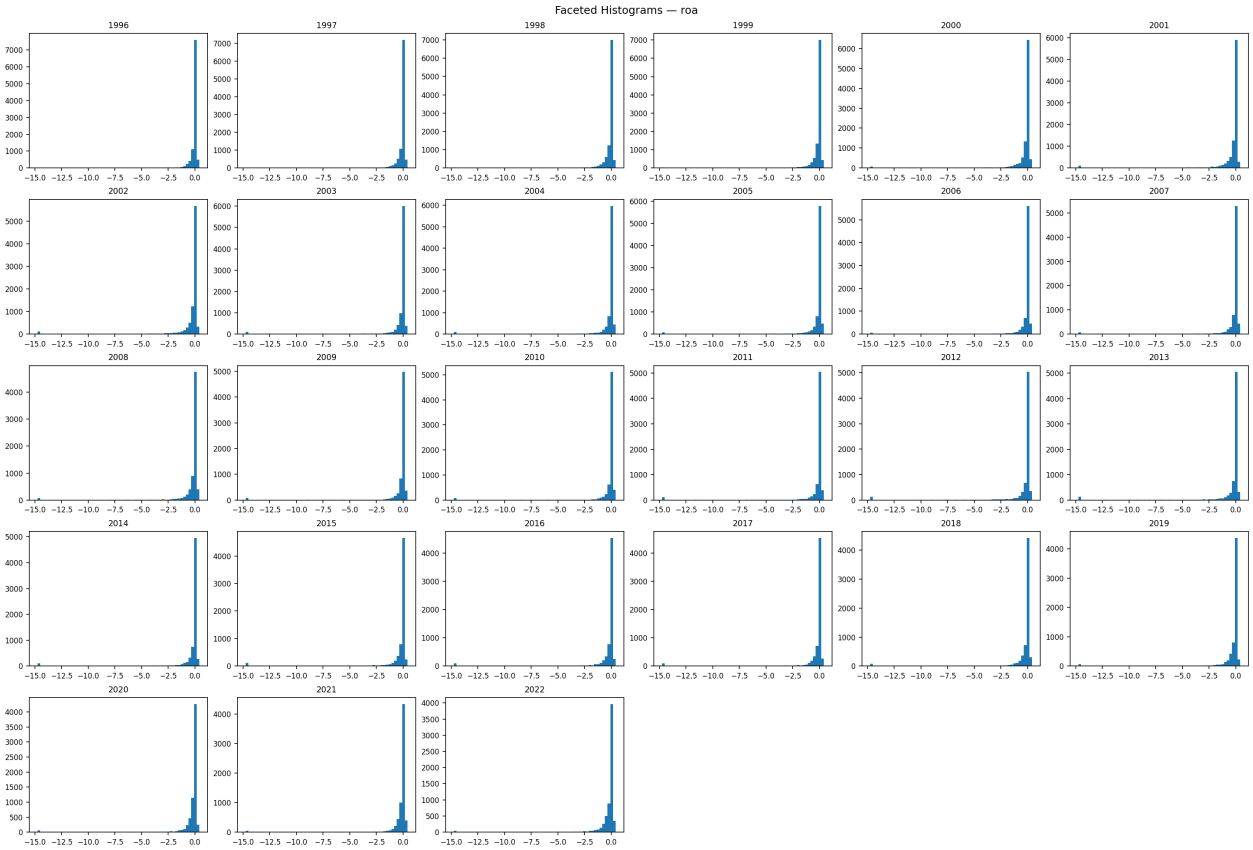

Figure 20: Faceted histograms — `roa` (1996–2022)

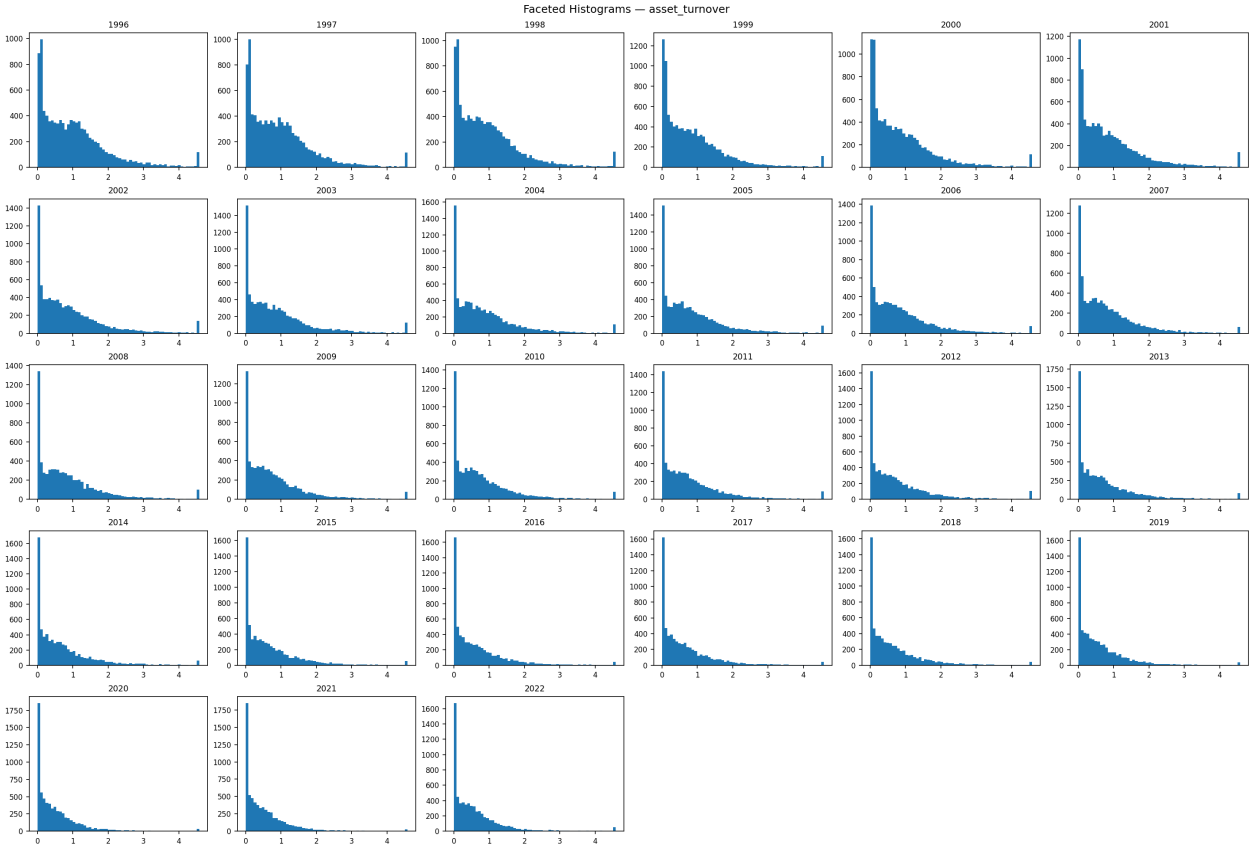

Figure 21: Faceted histograms — $\texttt{asset}_t urnover(1996 - -2022)$

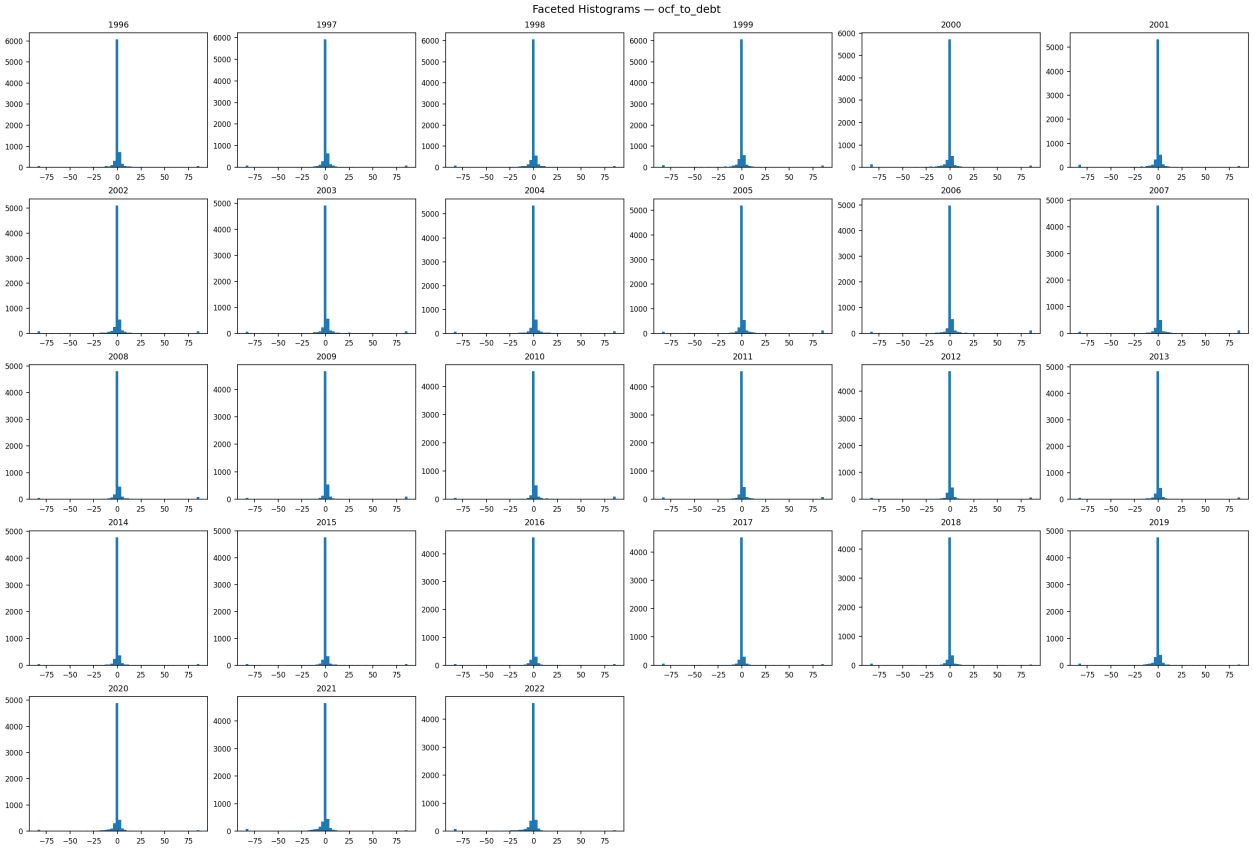

Figure 22: Faceted histograms — $\texttt{ocf}_t o_d ebt (1996 - -2022)$

4 Correlation Analysis

4.1 Overall correlation (all years)

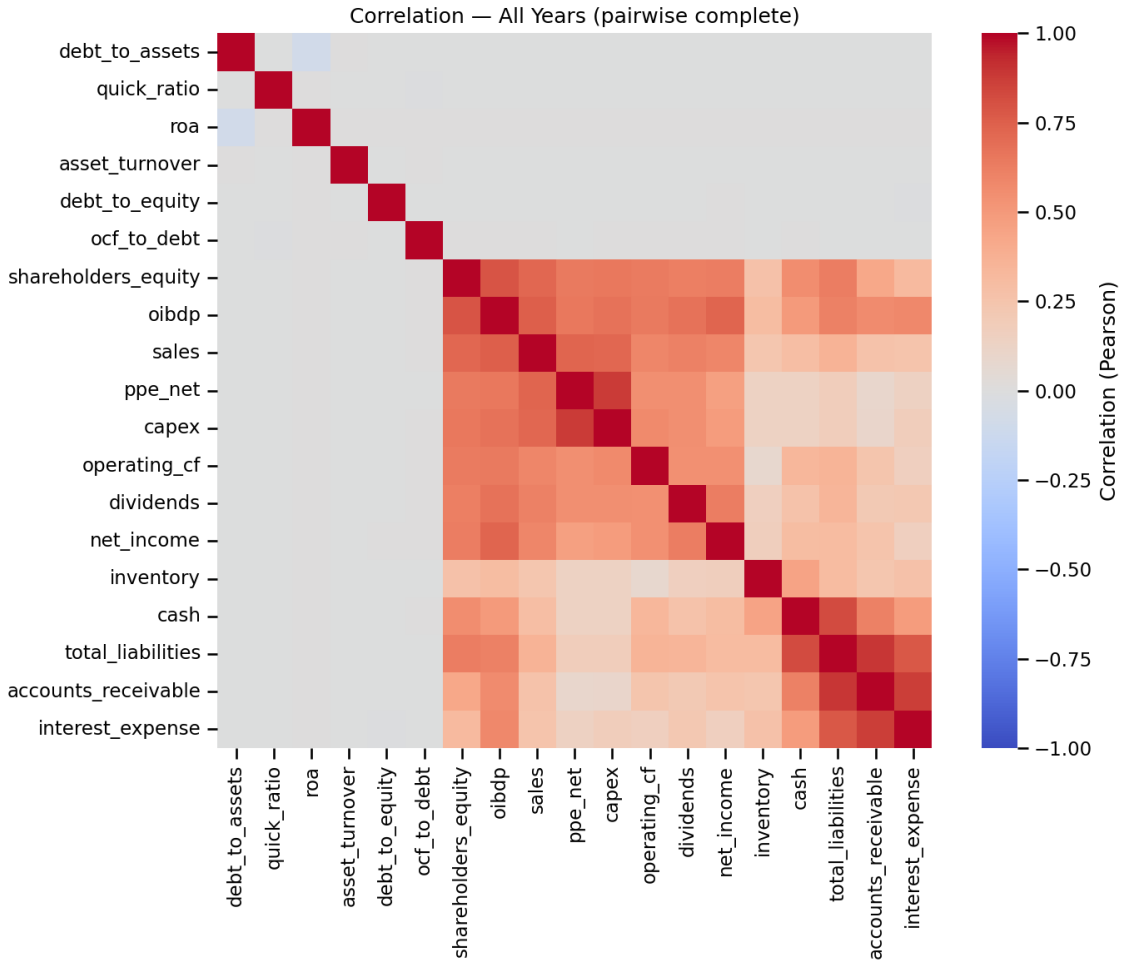

Figure 23: Overall Pearson correlation among selected features (pairwise complete)

Key observations:

- `current_ratio` ∼ `quick_ratio` ($|r| \approx 1.00$) ⇒ drop one
- `sales` ∼ `cogs` ($r \approx 0.97$) ⇒ use margins rather than both levels
- `cash` ∼ `short_term_investments` ($r \approx 0.91$) ⇒ keep `cash` or define "pure cash"

4.2 Recession vs. Expansion

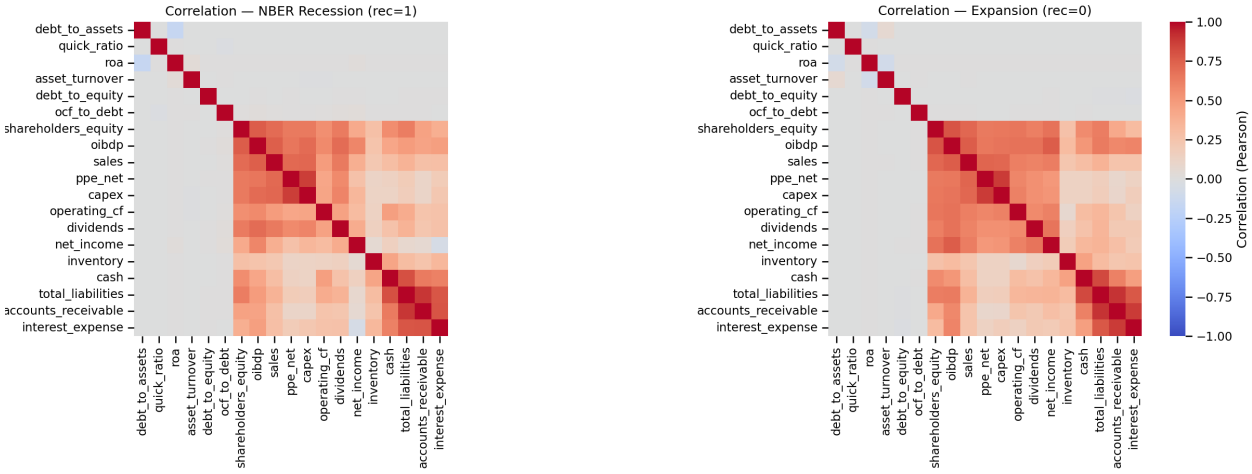

Figure 24: Correlation structure in NBER recessions (left) vs. expansions (right)

What changes:

- **Tighter in recessions:** `cash` with `operating_cf` and `interest_expense`; `roa` with `asset_turnover`
- **Weaker/more negative in recessions:** `net_income` vs. many scale variables; larger negative ties with `interest_expense`, `dividends`, `total_liabilities`, etc.

Illustrative top pairs

Top highly correlated pairs (overall)		
`current_ratio`	`quick_ratio`	$r = 1.000$
`sales`	`cogs`	$r = 0.968$
`cash`	`short_term_investments`	$r = 0.909$

Suggested drops: {`cogs`, `current_ratio`, `short_term_investments`}

Pairs with largest recession shifts

Table 2: More positive in recessions ($\Delta r > 0$)

var_a	var_b	Δr
cash	operating_cf	+0.171
cash	interest_expense	+0.138
roa	asset_turnover	+0.127
operating_cf	interest_expense	+0.121
shareholders_equity	interest_expense	+0.077

Table 3: More negative in recessions ($\Delta r < 0$)

var_a	var_b	Δr
operating_cf	net_income	−0.305
total_liabilities	net_income	−0.274
shareholders_equity	net_income	−0.272
dividends	net_income	−0.267
net_income	interest_expense	−0.258

5 Feature Scaling Plan

Based on observed shapes:

Method	Features
Log1p + Standardize	cash, accounts_receivable, inventory, ppe_net, total_liabilities, shareholders_equity, operating_cf, capex, dividends, sales, oibdp, interest_expense, quick_ratio, asset_turnover
Robust (median/IQR)	net_income, roa, debt_to_equity, ocf_to_debt
MinMax (0–1)	debt_to_assets

Rationale: log1p tames heavy right tails for nonnegative variables; robust scaling resists outliers for signed/heavy-tailed metrics; MinMax preserves bounded interpretation for leverage shares.

6 Principal Component Analysis (last two years)

- Preprocessing: light log1p on nonnegative features; median impute; standardize
- Using FY of the last two years in the sample (e.g., 2021–2022)
- **Result:** the minimum number of components for $\geq 90\%$ cumulative variance is **12**

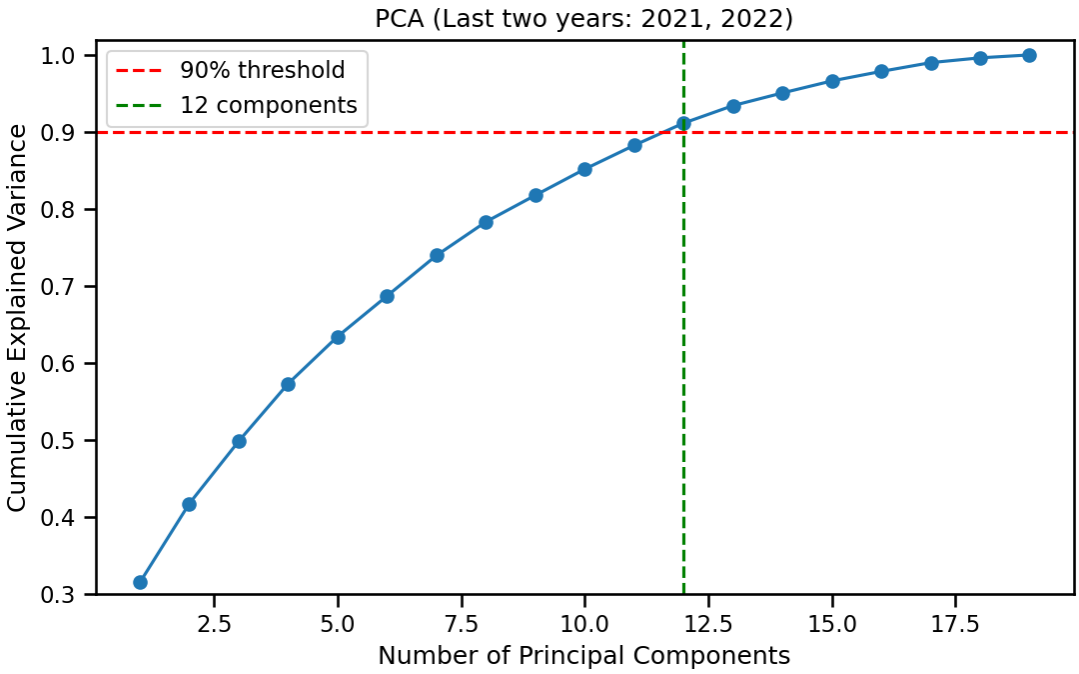

Figure 25: Cumulative explained variance vs. number of principal components (vertical line at k for 90%)

Interpretation: early PCs load on size/liquidity/leverage blocks; mid PCs separate profitability from scale; later PCs capture coverage and payout nuances. Use PCs for linear models or as diagnostics; tree/boosting models can directly exploit the raw (scaled) features.

7 Limitations and Next Steps

- **Missingness/Survivorship:** pairwise correlations rely on varying sample sizes across pairs
- **Industry mix:** sector heterogeneity drives level dispersion; consider industry-neutral z-scores
- **Outliers:** heavy tails persist even after clipping; robust losses may help
- **Next:** add margin variants (e.g., EBITDA, net profit), market-based signals (returns/volatility), and evaluate feature importance on a labeled distress proxy