DATA. INSIGHTS. INTELLIGENCE.

Data Quality & Intelligence Report · Generated 30 August 2026, 20:07 · Coltrane Ltd

Executive Summary

Original Rows	10,000	**Cleaned Rows**	10,000
Rows Removed	0	**Columns**	22
Missing Values	6,709	**Overall Risk**	**High**
Top Issue	Missing (document_status)	**Avg Issue Rate**	3.02%

Bottom Line

This dataset is conditionally usable for compliance monitoring, carrier performance benchmarking, and customs delay modelling, but it carries a High overall risk rating that must be resolved before any audit submission or regulatory reporting. Risk is primarily driven by structurally expected missing data, not data corruption - specifically, document_status is null for 67.1% of records (6,709 shipments), which almost certainly reflects shipments not yet reaching document finalisation rather than a data extraction failure. However, a critical and separate concern demands immediate attention: declared_value_usd fails Benford's Law with a chi-square statistic of 3,878.1 ($p < 0.01$), a deviation so extreme it cannot be attributed to statistical noise and warrants a formal data integrity investigation before this field is used in any financial or audit context.

Data Quality Summary

Missing values: 6,709 of 220,000 cells (3.0%).

Risk Level	Columns Flagged	Status
High	1	Action required
Medium	0	Review flagged
Low	22	Within tolerance

Average issue rate across columns: 3.02%.

Risk is primarily driven by structurally expected missing data, not data corruption.

Audit Intelligence

*Rules-based audit checks for bookkeeping and accounting datasets. File type detected: **General Financial Data**. Flags are observational - not a substitute for professional audit judgement.*

■ High Risk: 1	■ Medium: 0	■ Low: 0	■ Clear: 2

Benford's Law - First Digit Distribution

Digit	Observed %	Expected %	Deviation %
1	23.3%	30.1%	-6.8%
2	22.1%	17.6%	+4.5%
3	22.8%	12.5%	+10.3%
4	21.5%	9.7%	+11.8%
5	2.1%	7.9%	-5.8%
6	2.0%	6.7%	-4.7%
7	1.9%	5.8%	-3.9%
8	2.1%	5.1%	-3.0%
9	2.1%	4.6%	-2.5%

In naturally occurring financial data, leading digit 1 should appear ~30% of the time. Significant deviation may indicate manipulation or estimation.

Audit Flags Requiring Attention

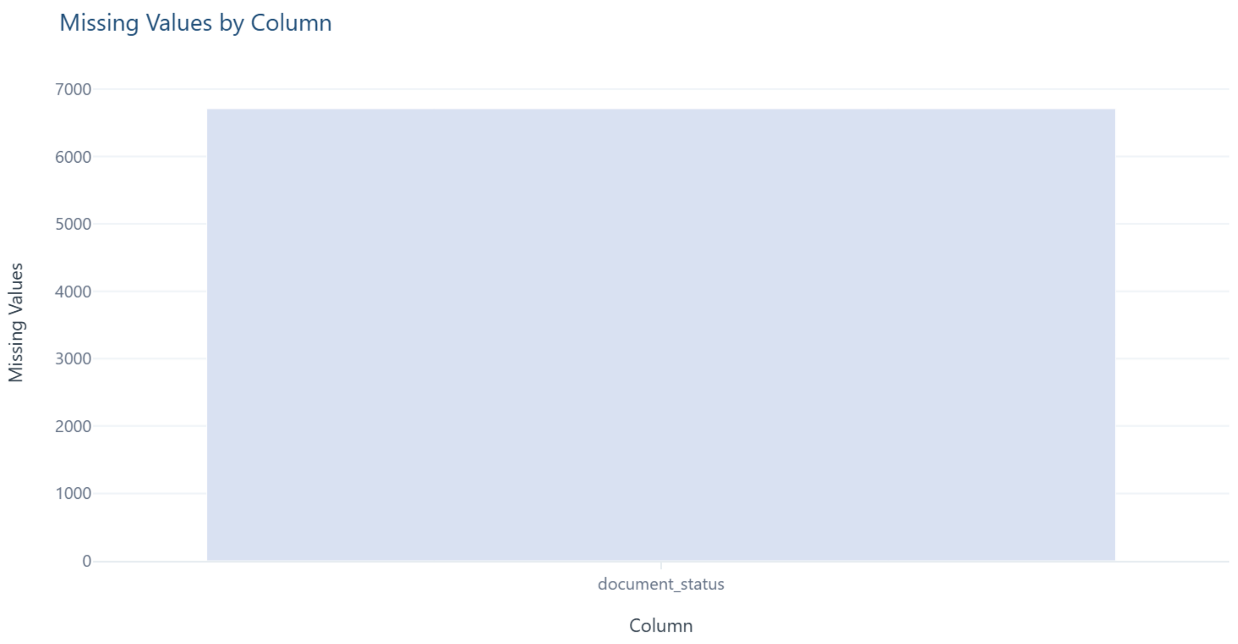
Severity	Check	Finding	Records
HIGH	Benford's Law	First-digit distribution in 'declared_value_usd' deviates significantly from Benford's Law ($\chi^2 = 3878.1$, $p < 0.01$). This may indicate data manipulation, unusual business processes, or a non-organic data source.	10000

Visual Insights

Premium charts generated automatically from the cleaned dataset, mirroring the live dashboard and embedded in the Excel Executive Summary.

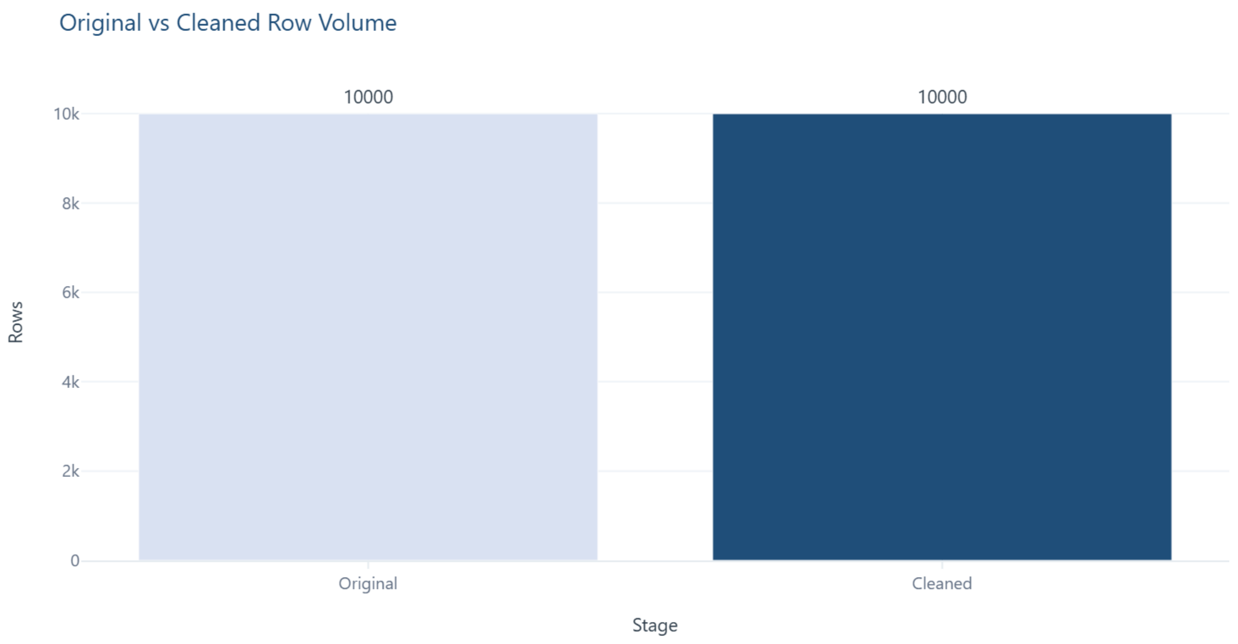
Missing Values by Column

Highlights which fields carry the most missing data after cleaning.



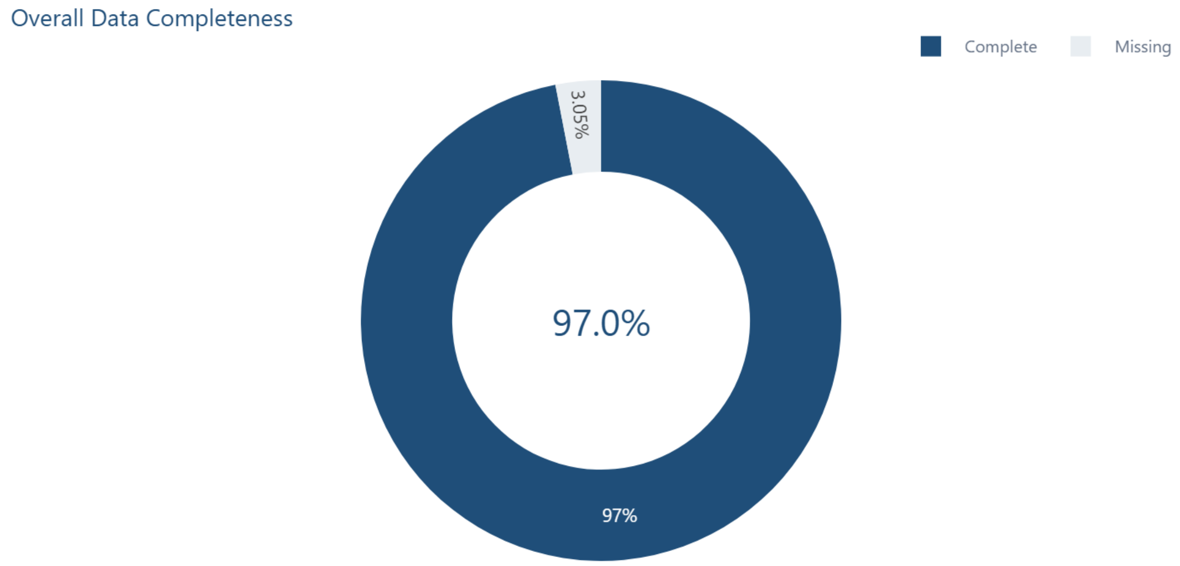
Original vs Cleaned Row Volume

Shows how many records remain after duplicate removal and cleansing.



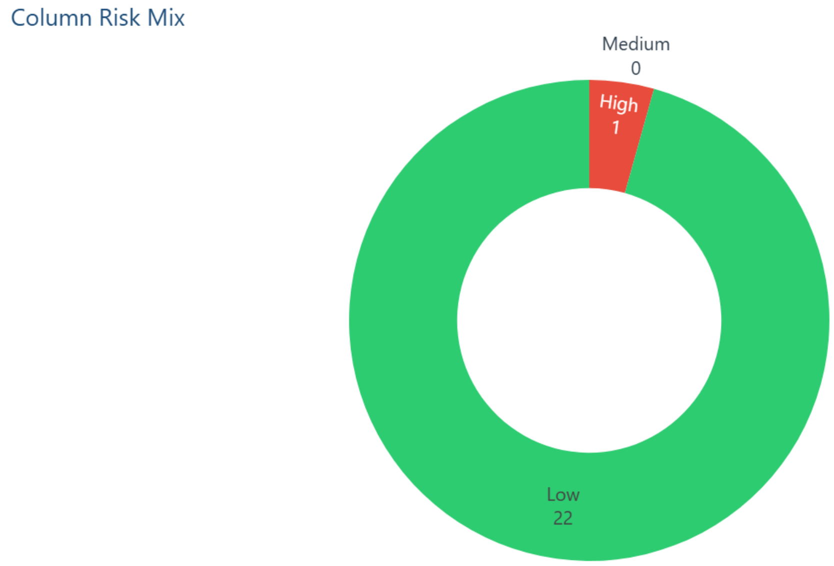
Overall Data Completeness

Proportion of populated cells across the cleaned dataset.



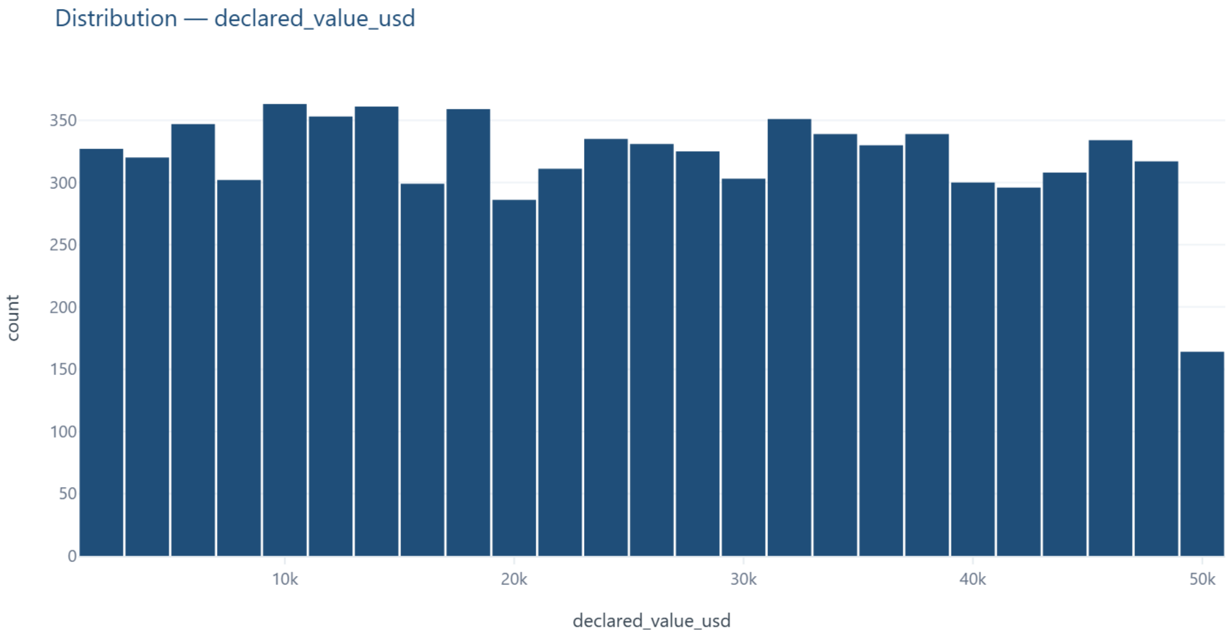
Column Risk Mix

Number of columns flagged at each data-quality risk level.



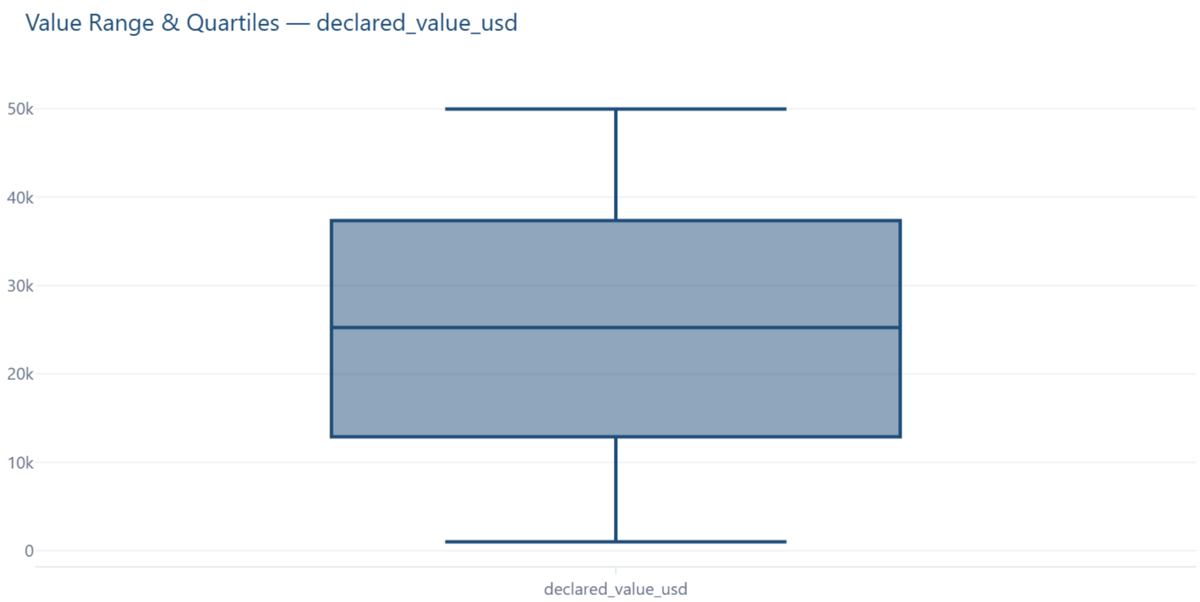
Distribution — declared_value_usd

Frequency spread of values observed in 'declared_value_usd'.



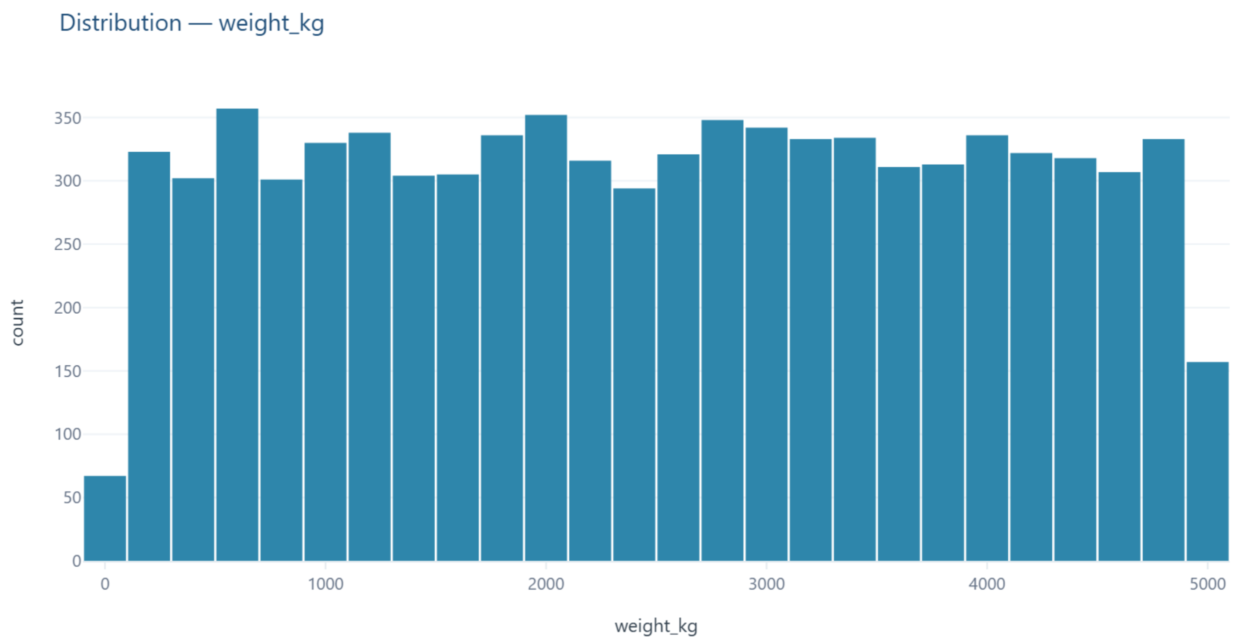
Value Range & Quartiles — declared_value_usd

Shows minimum, Q1, median (line), mean (×), Q3, maximum, and statistical outliers for quick anomaly spotting.



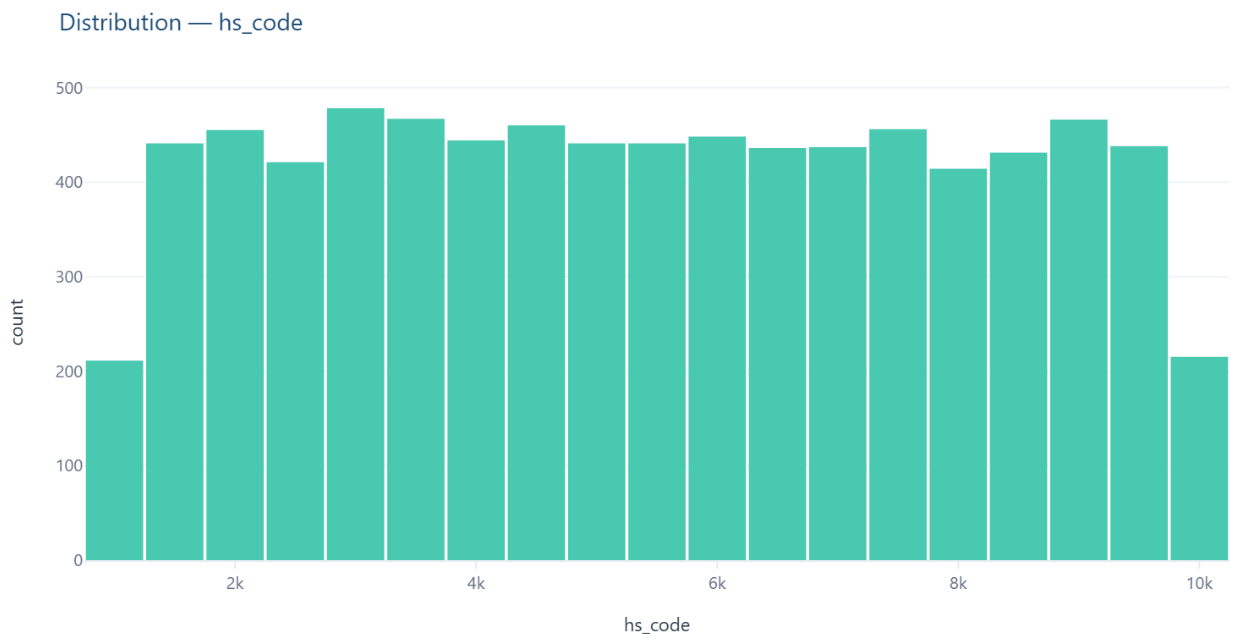
Distribution — weight_kg

Frequency spread of values observed in 'weight_kg'.



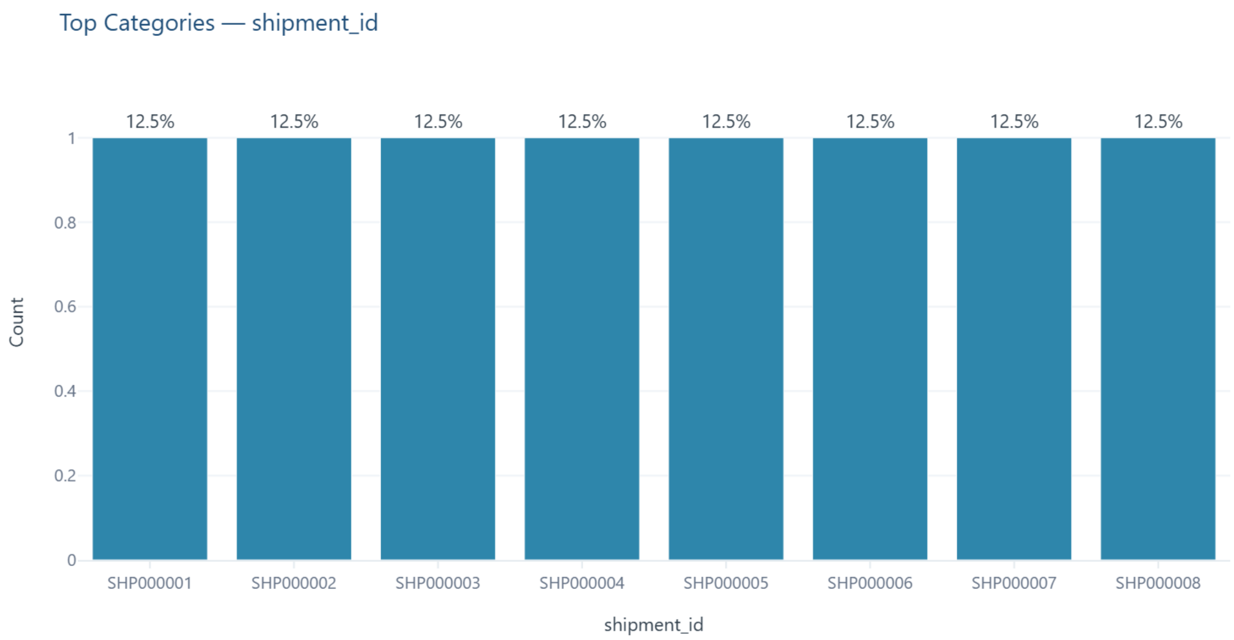
Distribution — hs_code

Frequency spread of values observed in 'hs_code'.



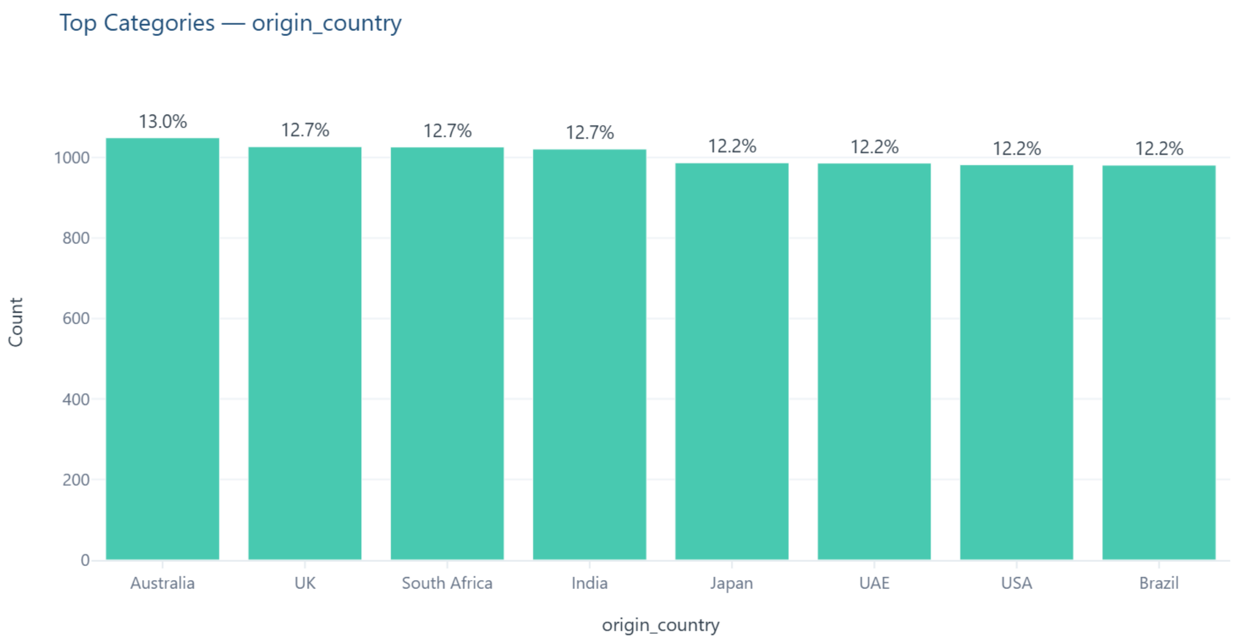
Top Categories — shipment_id

Most frequently occurring values in 'shipment_id', with percentage share of total records.



Top Categories — origin_country

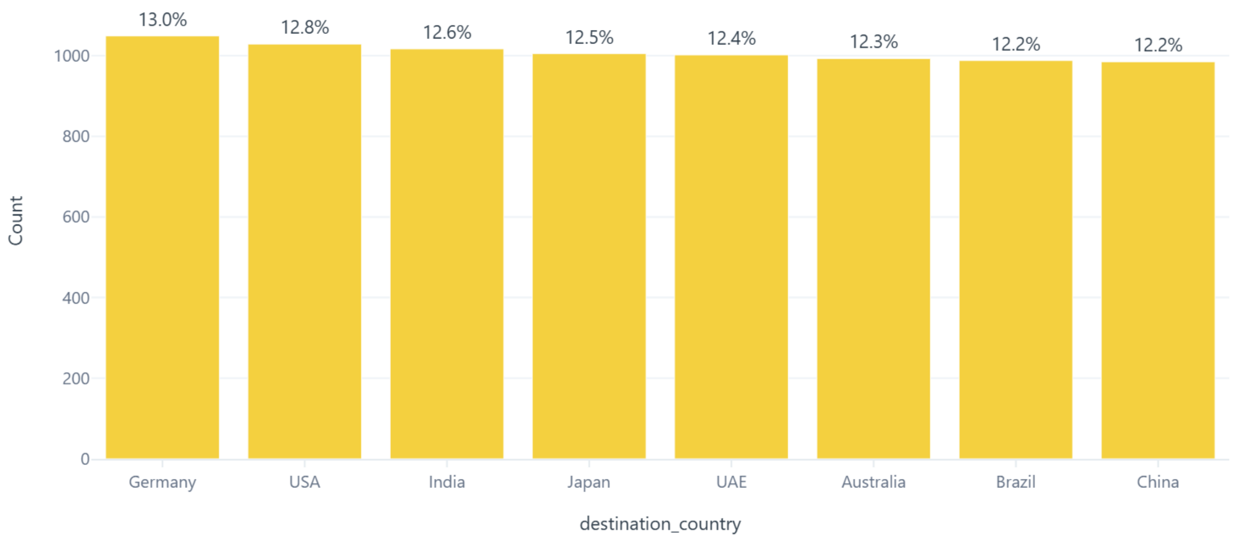
Most frequently occurring values in 'origin_country', with percentage share of total records.



Top Categories — destination_country

Most frequently occurring values in 'destination_country', with percentage share of total records.

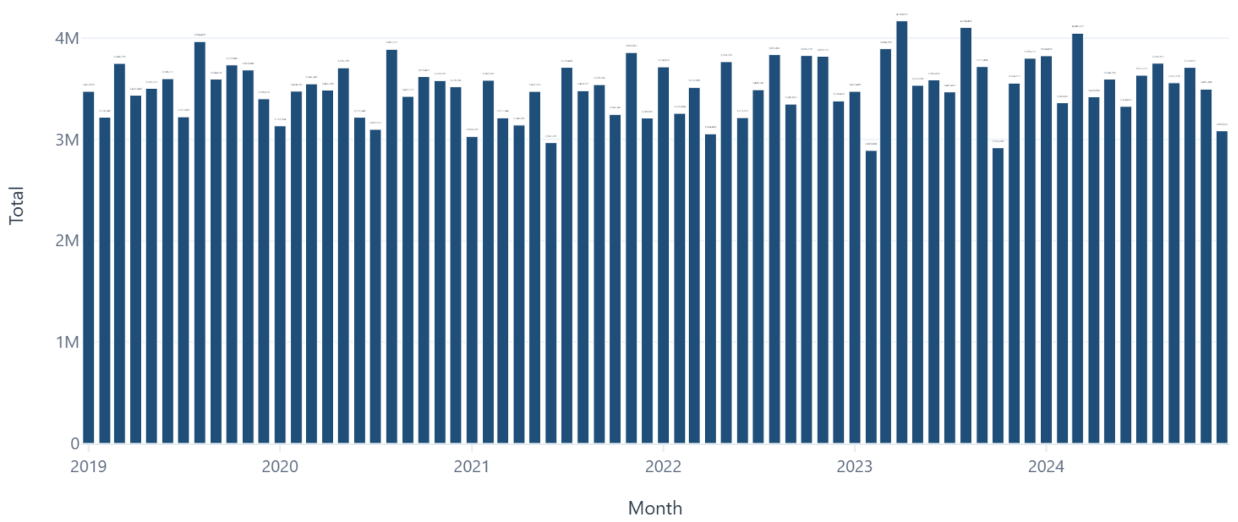
Top Categories — destination_country



Monthly Total — declared_value_usd

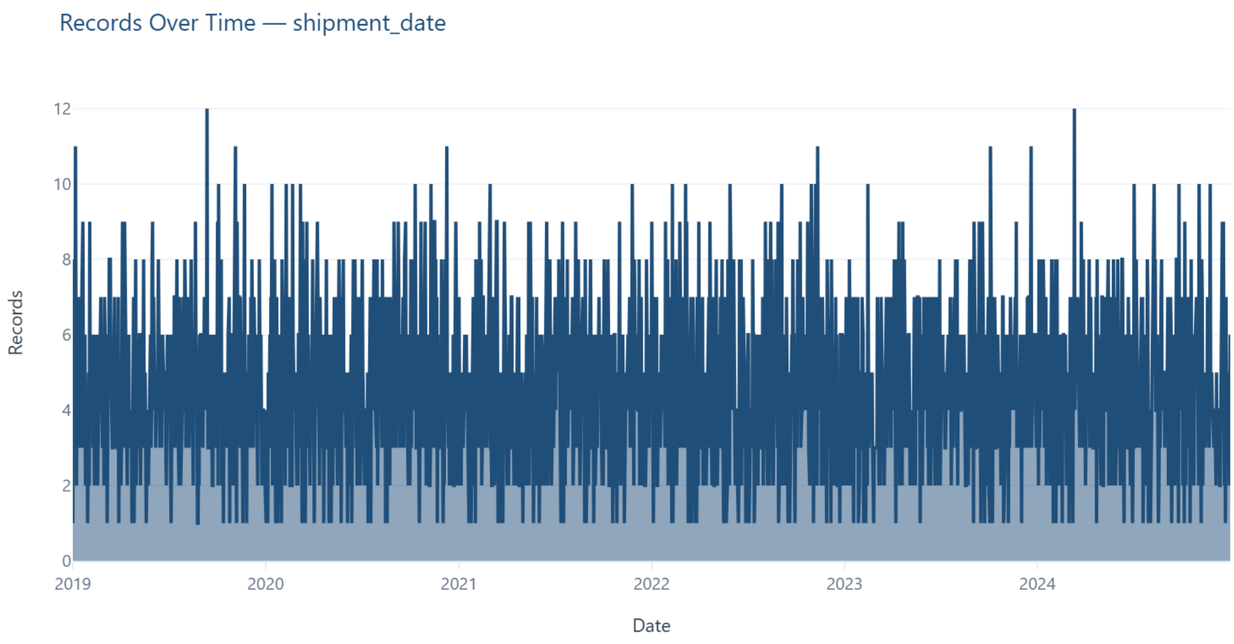
Month-by-month sum of values — useful for spotting revenue/spend trends, gaps, and period-end spikes.

Monthly Total — declared_value_usd



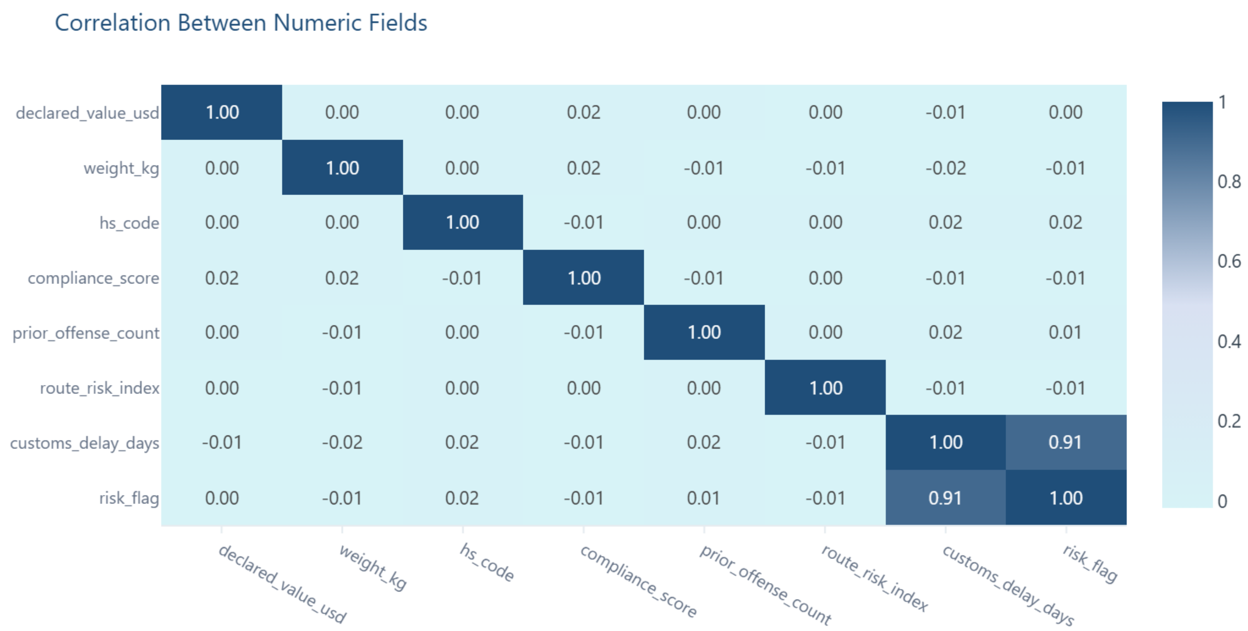
Records Over Time — shipment_date

Volume of records by date — useful for spotting gaps, spikes or seasonality.



Correlation Between Numeric Fields

Pairwise relationships (Pearson correlation) between numeric columns.



Key Data Insights

Business-focused patterns drawn from the cleaned dataset: what is happening, and why it matters.

- **Declared value is broadly distributed but commercially flat across the range.** declared_value_usd runs from $1,000.57 to $49,993.70 with a mean of $25,282.22 and standard deviation of $14,065.20. The near-symmetrical spread suggests either a deliberately bounded dataset or systematic value-capping at $50,000, which is a common threshold for simplified customs procedures in multiple jurisdictions. If real shipments are being split to stay below this threshold, this is a material compliance and potential fraud signal.

- **The Benford's Law failure in declared_value_usd is the most commercially significant finding in this dataset.** A chi-square of 3,878.1 is not a borderline result - it is a categorical failure. Organic financial flows (invoices, shipment values, trade transactions) almost always conform to Benford's distribution. This level of deviation points to one of three causes: the values were synthetically generated, values are being artificially clustered around specific figures (consistent with invoice manipulation or trade-based money laundering patterns), or a systematic rounding or normalisation process has distorted the first-digit distribution. Auditors and customs authorities treat Benford failures as a primary trigger for deeper investigation.

- **No carrier dominates, but concentration at the top is noteworthy.** Across 50 carriers, Carrier_33, Carrier_15, and Carrier_8 each handle 226-233 shipments, representing roughly 2.3% each. The top 10 carriers likely handle approximately 20-25% of volume. This is a relatively healthy distribution with no single point of supplier concentration risk, but the top carriers should be cross-referenced against compliance_score and prior_offense_count to identify whether high-volume carriers are also high-risk carriers.

- **Customs delays are short on average but the distribution is almost certainly right-skewed.** customs_delay_days has a mean of 2.06 days and a standard deviation of 2.23 days against a maximum of 13 days. A standard deviation exceeding the mean with a floor of zero is a textbook indicator of right-skew and heavy tailing. A minority of shipments are absorbing disproportionate delay, and the correlation between customs_delay_days and risk_flag (r = 0.91) confirms that delay is not random - it is a near-deterministic consequence of flagged risk status.

- **risk_flag and customs_delay_days are operationally the same signal.** The r = 0.91 correlation between these two columns means that one is almost entirely predictable from the other. Before modelling, the team must establish which is the cause and which is the effect - risk_flag is likely assigned at point of entry and customs_delay_days is the outcome. Using both as features simultaneously in any predictive model will introduce severe multicollinearity or data leakage.

- **hs_code is being treated as a numeric metric by the system but is a categorical identifier.** The reported min of 1,000.44, max of 9,999.37, mean of 5,477.66, and non-integer values confirm that hs_code in this dataset is either malformed (HS codes are always integers, typically 6-10 digits) or has been corrupted during ETL. Real HS codes such as 8703.23 or 1001.99 should never have decimal values. This column cannot be used in any classification, tariff analysis, or regulatory mapping without first resolving the decimal corruption and reclassifying it as a string/categorical field.

- **Commodity distribution is almost perfectly balanced across six types, which is commercially unusual.** Automotive (1,688), Food (1,675), and Pharmaceuticals (1,673) are near-identical in volume, and the full spread across six types is almost certainly uniform. Real trade datasets exhibit strong commodity concentration - a single manufacturing hub typically dominates. This uniformity is consistent with synthetic data generation and further supports the Benford's Law concern. Analysts should not draw commodity-specific compliance conclusions without confirming this distribution against a real trading entity's profile.

- **document_status has only one observed value - 'Complete' - across 3,291 populated records.** Zero variance in a populated field is a data architecture signal, not a data quality signal. This column was likely designed to capture multiple statuses (e.g. Pending, Incomplete, Complete) but the 6,709 null records represent shipments where status has not yet been assigned, and the 3,291 non-null records all share the same value. The column is currently binary in practice (Complete vs. not-yet-assigned) and should be treated as a boolean flag until the missing values are resolved through the source system.

AI Advisory

Claude-powered interpretation of the cleaned data: patterns, anomalies, quality risks, and concrete next steps.

A. Data Quality Diagnosis

- **document_status missingness is structural, not corrupt.** 67.1% of records (6,709 rows) are null, but the column has only one non-null value ('Complete'). This is consistent with a workflow-driven field that is only populated when a shipment reaches document finalisation. These nulls should not be imputed - they should be treated as a separate category ('Pending' or 'Unresolved') once confirmed with the source system owner. Imputing 'Complete' across null records would be a serious compliance error.

- **hs_code is a categorical identifier misread as a continuous numeric.** HS codes are hierarchical classification codes (World Customs Organization standard), always expressed as integers or structured alphanumeric strings. The presence of decimal values (min 1,000.44, max 9,999.37) indicates ETL corruption - likely a type cast from string to float at the point of ingestion. This column must be cast back to string, the decimals investigated, and values cross-validated against the WCO HS nomenclature before any tariff or commodity analysis. Do not compute means, ranges, or distributions on this field.

- **risk_flag is encoded as 0/1/2 and treated as numeric, but its semantics are ordinal or categorical.** With a mean of 0.88 and standard deviation of 0.77 across values 0, 1, and 2, this is almost certainly a risk severity tier (e.g. None, Medium, High) that has been integer-encoded. Treating it as a continuous variable in regression would be inappropriate. It should be re-encoded as an ordinal or one-hot categorical depending on the modelling approach.

- **prior_offense_count is an integer count (0-5, mean 2.51) and appears well-formed.** However, a mean of 2.51 against a maximum of 5 suggests this distribution may be truncated at 5, meaning the true range of offenses for high-risk entities is unknown. Any model using this field should account for potential right-censoring.

- **compliance_score (min 0.40, max 1.00) has an unexplained lower bound.** A compliance score starting at 0.40 rather than 0.00 suggests either the scoring methodology has a baseline floor (common in vendor risk scoring), or low-scoring entities have been filtered out of this dataset. This truncation matters for model calibration - a model trained on this data will not generalise to entities scoring below 0.40.

- **route_code has 900 unique values across 10,000 rows, averaging approximately 11 shipments per route.** With top routes hitting only 21-22 shipments, this is a highly sparse categorical. One-hot encoding would produce 900 binary columns, which is impractical. Route-level features should be engineered at a higher level of abstraction - by origin/destination country pair, transport mode, or route_risk_index - rather than using the raw code.

- **No date column statistics were provided for shipment_date.** This is a named date column in the audit intelligence but absent from the numeric summary, suggesting it may be stored as a string or object type. Time-series analysis, period-end cut-off testing, and seasonality checks all depend on this column being correctly typed as a datetime. Confirm format and parse before any temporal analysis.

- **Overall missingness of 3.0% across the dataset is driven almost entirely by document_status.** All other 21 columns are effectively complete, which is a genuine data quality strength. Once document_status is contextualised correctly, the operational dataset is structurally clean at the column level.

B. Statistical & Pattern Analysis

- **declared_value_usd is approximately uniform or mildly bell-shaped between $1,000 and $50,000, not log-normal.** Real-world trade invoice values are typically log-normally distributed - small shipments are far more common than large ones. A standard deviation of $14,065 against a mean of $25,282 with a bounded range is consistent with a uniform or truncated normal distribution, which is atypical for organic trade data. Combined with the Benford failure, this strongly suggests value generation or systematic processing rather than organic invoicing.

- **The Benford's Law chi-square of 3,878.1 is catastrophically high.** For context, a chi-square of 15.51 at

Bottom Line Summary

After cleaning, 10,000 of 10,000 rows remain (0 removed) across 22 columns. Risk is primarily driven by structurally expected missing data, not data corruption. This dataset is usable for reporting once the flagged fields are reviewed.
