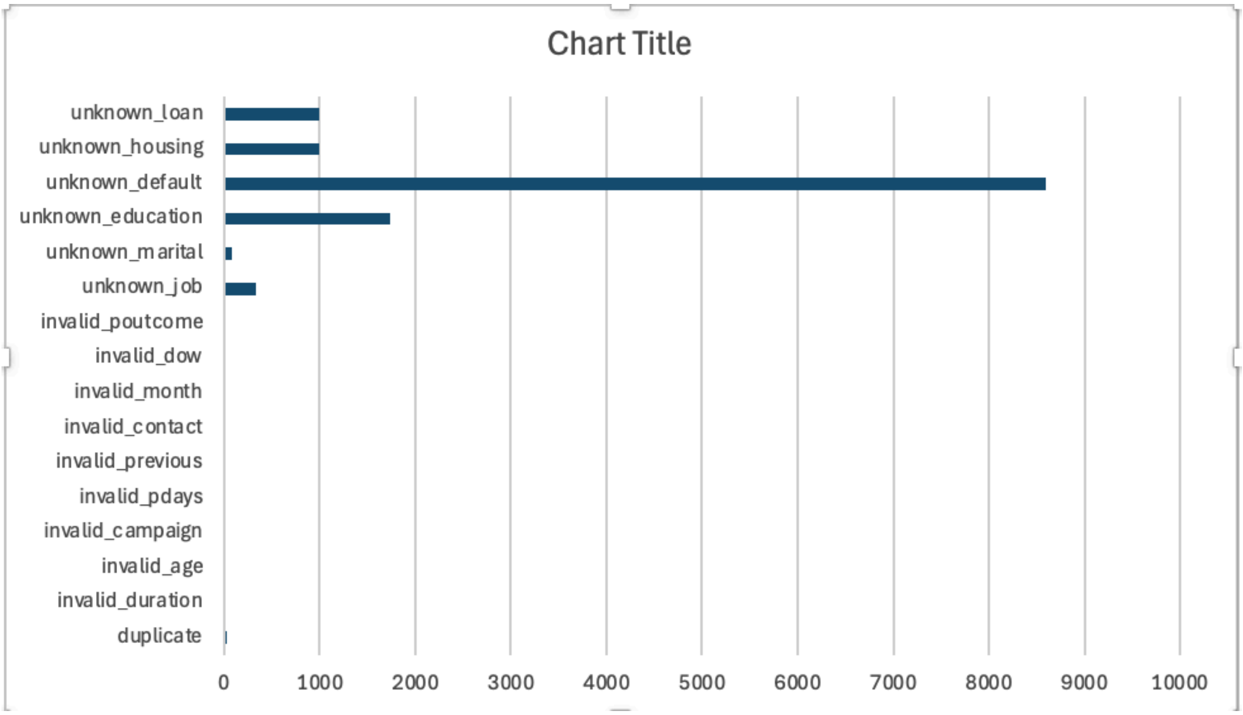
P1 — Data Quality Audit (Records → Insights → Recommendation)

Problem: The bank-marketing table contains exact duplicate rows, non-positive call durations, under-18 ages, and `"unknown"` values in key attributes.
Why it matters: Ambiguous/invalid data skews campaign reporting, model training, and segment performance.
Metric (success): Remove exact duplicates; flag all non-positive durations and under-18 ages; quantify and isolate `"unknown"` values; publish exception log + SOP.
Data: `bank-additional-full.csv` — **41,188 rows × 21 columns**.

Baseline (from initial scan)
- **Exact duplicates to drop:** **12** (keep one exemplar per identical set)
- **Non-positive `duration`:** **4**
- **Under-18 `age`:** **5** (min 17, max 98)
- **"unknown" counts:**
  - `job` **330** (0.80%) · `marital` **80** (0.19%) · `education` **1,731** (4.20%)
  - `default` **8,597** (20.87%) · `housing` **990** (2.40%) · `loan` **990** (2.40%)



Scope & Rules (what was checked)
Duplicates — exact match across **all 21** columns → keep one, drop the rest, log examples.

Required fields — `age, job, marital, education, contact, month, day_of_week, duration, campaign, pdays, previous, poutcome, y`.
Numeric ranges — `age ∈ [18,99]`, `duration > 0`, `campaign ≥ 1`, `previous ≥ 0`, `pdays = −1` or `pdays ≥ 0`.
Valid categories — `contact ∈ {cellular, telephone}`, `month ∈ {jan…dec}`, `day_of_week ∈ {mon…fri}`, `poutcome ∈ {nonexistent, failure, success}`.
Treat `"unknown"` as missing — `job, marital, education, default, housing, loan`.

Actions
1. Added repeatable **flag columns** (duplicate / range / category / `"unknown"`).
2. Counted issues (COUNTIF/Pivot) and created an **issues bar chart**.
3. **Removed exact duplicates**, logged representative rows in `checks/exception_log.csv`.
4. Wrote/committed a simple **SOP** (`checks/SOP_checklist.md`) for intake → validation → update.

Result
- Duplicates removed; remaining issues **isolated and traceable** via flags + exception log.
- Baseline quality levels are now **quantified**, enabling consistent monitoring across ingests.

Recommendation
- **Run this SOP on every ingest**; publish counts + chart each refresh.
- Owners resolve **open exceptions weekly** (collect missing values; decide policy on under-18 rows).
- Revisit thresholds and add domain-specific rules **quarterly**.

Deliverables
- `checks/exception_log.csv` (running log of issues, actions, status)
- `checks/SOP_checklist.md` (intake → validation → update)
- `docs/P1_QualityBrief.pdf` (this brief exported to PDF) + issues chart image in `dashboards/`