Data Quality Summary Report

Dataset Context: Banking campaign data for customer outreach and product marketing.

Objective: Evaluate the dataset's integrity and usability for business applications by conducting a data quality profiling workflow.

1. **Overview & Structure**

Total records: 45,211

Columns analyzed: 17

No schema issues (all columns are aligned with the data types).

2. **Completeness**

Explicit missing values: 0 (0.0%)

Disguised nulls ("unknown," "N/A"): 52,124 fields (approximately 6.8% of all data points)

Due to the sheer number of placeholders, the impact on customer data segmentation, model building and/or compliance reporting could be significant.

3. **Uniqueness & Duplicates**

- Exact duplicate rows: 0 (0.0%) – No risk of double counting.

Composite key uniqueness: ~99. 95% interactions with distinct customers

4. **Domain Validity**

Age range: 18 – 95 (within expected bounds)

Categorical Fields (default, loan, housing): Contain only valid values (yes/no).

5. **Business Rule Violations**

Duration ≤ 0: 3 records (<0.01%) - Possible logging issues

Campaign < 1: 0 records (0.0%) - Clean

- Default = “yes” and balance > 0: 301 records (0.7%) - Potential inconsistency or elevated risk profile

6. **Outlier Detection (IQR Method)**

Column	**Outlier**	**% of Records**
Age	487	1.1%
Balance	4,729	10.5%
Duration	3,235	7.2%
Campaign	3,064	6.8%
Pdays	8,257	18.3%
Previous	8,257	18.3%

There is a significant skew with long tails in the financial and contact history fields.

7. **Correlation Analysis**

High correlations (|r| > 0.8): none

Low multicollinearity – good for predictive modeling.

8. **Timeliness**

Customers never contacted (pdays = 999): 0 (0.0%)

There are no customers without any contact history.

9. Overall Evaluation

Strengths:

- ✓ Schema and structure
- ✓ No duplicates or explicit missing values
- ✓ Uniqueness
- ✓ Valid values

Areas to Address:

- The high percentage of disguised nulls (6.8%)
- Up to 18% of employees in important financial and behavioral categories.
- Potential discrepancies between loan status and default flag (0.7%).

10. **Recommendations**

- Impute or flag disguised nulls for downstream modeling and compliance
- Investigate outliers to distinguish genuine long tail product behavior from data entry errors.
- Review loan–default overlaps for risk modeling or regulatory reporting.
- Integrate profiling checks in dashboards/notebooks for repeatability and auditability.