AI Data Quality Audit Checklist

A thorough audit of your data quality readiness for AI and machine learning projects — covering completeness, accuracy, governance, integration, and bias assessment across 50+ critical evaluation items.

6	50+	Score	Free
AUDIT SECTIONS	ITEMS TO CHECK	EACH SECTION OUT OF 10	PRINT & USE NO STRINGS

How to Use This Checklist

Complete this audit before beginning any AI or machine learning project. Work with your data owners, database administrators, and business analysts to assess each item honestly. Sections scoring below 6 indicate significant data quality risks that should be remediated before model training begins.

1 Data Completeness & Availability

Assess whether the data you need for your AI initiative actually exists, is accessible, and has sufficient volume and coverage to train reliable models.

- [ ] All **required data fields** for your target AI use case have been identified and documented in a data requirements specification (feature list)
- [ ] The percentage of **missing values** in each required field has been measured — fields with more than 20% missing data require remediation before use in model training
- [ ] Data covers a **sufficient time period** to capture seasonal patterns, cyclical trends, and edge cases relevant to your business domain (minimum 12–24 months recommended)
- [ ] The **volume of labelled training examples** is adequate for your chosen AI approach — simple classifiers need thousands; complex NLP or vision models need tens of thousands or more
- [ ] Data from **all relevant business units and systems** has been identified and access permissions secured, including legacy systems that may hold critical historical records
- [ ] There are **no critical data gaps** that would prevent the model from learning key patterns — for example, customer churn data without the preceding behaviour signals is insufficient
- [ ] A **data collection plan** is in place for any missing data, with realistic timelines for accumulating sufficient volume before model training begins
- [ ] Sample data has been **extracted and reviewed manually** by domain experts to confirm it contains the information expected and is representative of real-world conditions

Section Score: _____ /10

2 Data Accuracy & Consistency

Evaluate whether your data is truthful, consistently formatted, and free from errors that would cause AI models to learn incorrect patterns.

- [ ] A **random sample of records** (minimum 5%) has been manually validated against source documents to measure the baseline accuracy rate across key fields
- [ ] Data entry standards and **validation rules** are enforced at the point of capture — free-text fields where structured data is expected are a common source of inconsistency
- [ ] **Duplicate detection** has been performed across all datasets, with a documented deduplication strategy for exact matches and fuzzy matches (e.g., "Ltd" vs "Limited")
- [ ] Date and time fields use a **consistent format** across all data sources — mixed formats (DD/MM/YYYY, MM/DD/YYYY, YYYY-MM-DD) cause silent errors in model training
- [ ] Numerical fields use **consistent units and scales** — revenue in pounds vs pence, distances in miles vs kilometres, temperatures in Celsius vs Fahrenheit must be standardised
- [ ] Categorical fields use a **controlled vocabulary** rather than free text — "United Kingdom", "UK", "U.K.", and "Great Britain" should all map to a single canonical value
- [ ] Data from merged systems or acquisitions has been **reconciled and harmonised** — legacy data often uses different coding schemes, field names, and business rules
- [ ] **Outlier detection** has been performed on numerical fields to identify values that are implausible (e.g., age 999, negative quantities, dates in the future) and may distort model training
- [ ] A **data quality scorecard** has been produced summarising accuracy, completeness, and consistency metrics for each dataset, with red/amber/green status indicators

Section Score: _____/10

3 Data Freshness & Timeliness

AI models trained on stale data produce stale predictions. Assess whether your data is current enough to drive reliable AI outputs.

- [ ] The **last update timestamp** for each data source has been documented, confirming data is current enough for your use case (real-time, daily, weekly, or monthly refresh)
- [ ] You have defined the **maximum acceptable data latency** for your AI application — a fraud detection model needs near real-time data; a quarterly forecasting model can tolerate daily updates
- [ ] Data refresh schedules are **automated and monitored** — manual data exports that rely on individuals remembering to run them are unreliable and will eventually be forgotten
- [ ] Historical data has **not been retroactively modified** in ways that would distort model training — backdated corrections should be flagged and handled appropriately
- [ ] You have assessed whether the **business environment has changed significantly** since the training data was collected — post-pandemic behaviour patterns may not match pre-pandemic data
- [ ] A **data freshness monitoring system** alerts the team when data pipelines fail or data has not been updated within expected timeframes
- [ ] The **training/serving data gap** has been measured — if your model was trained on data up to last month but serves predictions on today's data, understand what might have changed in between
- [ ] There is a defined **retraining trigger** based on data freshness — when new data accumulates beyond a threshold or performance metrics degrade, retraining is initiated automatically

Section Score: ____/10

4 Data Governance & Documentation

Well-governed data with clear documentation is essential for reproducible, auditable, and trustworthy AI systems.

- [ ] Every dataset used for AI has a **designated data owner** who is accountable for its quality, accuracy, and appropriate use within the organisation
- [ ] A **data dictionary** exists documenting every field, its definition, data type, allowed values, and business meaning — this is essential for new team members and auditors
- [ ] The **lineage of each dataset** is documented — where it originated, how it was transformed, and what business rules were applied during processing
- [ ] Data access is governed by **role-based access controls (RBAC)** ensuring that only authorised personnel can read, modify, or export data used for AI training
- [ ] A **data classification scheme** categorises all datasets by sensitivity level (public, internal, confidential, restricted) with corresponding handling requirements
- [ ] Changes to data schemas, business rules, or source systems are managed through a **formal change control process** with impact assessment and stakeholder notification
- [ ] All datasets containing personal data have been **assessed under UK GDPR**, with lawful basis documented, retention periods defined, and data subject rights procedures in place
- [ ] A **metadata management system** or catalogue tracks dataset versions, quality metrics, usage history, and dependencies across the organisation
- [ ] There is a **documented process for decommissioning datasets** that are no longer needed, including secure deletion and removal from all downstream pipelines and models

Section Score: ____/10

5 Data Integration & Accessibility

AI projects often require combining data from multiple systems. Assess how easily your data can be integrated and made available for model training.

- [ ] All relevant data sources can be **accessed programmatically** via APIs, database connections, or file exports — data locked in proprietary systems without export capability is a significant blocker
- [ ] A **common identifier** (customer ID, employee number, product code) exists across systems to enable reliable joins and lookups between different data sources
- [ ] Data from different systems has been **test-joined** to verify that matching works correctly — mismatched keys, orphaned records, and many-to-many relationships have been identified and resolved
- [ ] Your data warehouse or lakehouse can **handle the query volumes** required for model training — extracting millions of rows for feature engineering should not cripple production database performance
- [ ] Real-time or near-real-time data access is available for use cases that require it — **streaming data pipelines** (Kafka, Event Hubs) have been evaluated where batch processing is insufficient
- [ ] Data is available in **formats compatible with your AI platform** — CSV, Parquet, JSON, or direct database connections, with appropriate character encoding (UTF-8) for international data
- [ ] There is a **sandbox or development environment** where data scientists can explore and experiment with data without risking production systems or violating data protection policies
- [ ] Network connectivity between data sources and the AI platform has been tested, with **sufficient bandwidth and acceptable latency** for transferring large training datasets

Section Score: _____ /10

6 Bias & Fairness Assessment

AI models amplify patterns in data — including biases. Proactively assess your data for sources of bias before training begins.

- [ ] You have **analysed demographic representation** in your training data to identify whether certain groups are over- or under-represented relative to the population your model will serve
- [ ] Historical decisions encoded in your data have been **reviewed for systemic bias** — if past decisions were biased, training a model on them will perpetuate and potentially amplify that bias
- [ ] Data collection methods have been assessed for **selection bias** — data gathered only from certain channels, time periods, or customer segments may not represent the full picture
- [ ] Proxy variables that **correlate with protected characteristics** (postcode as a proxy for ethnicity, name as a proxy for gender) have been identified and their potential impact on model fairness evaluated
- [ ] A **fairness metric** has been selected for your use case — equal opportunity, demographic parity, or predictive equality — and baseline measurements have been established
- [ ] You have a plan for **ongoing bias monitoring** in production, including regular audits of model outputs across different demographic groups to detect emerging disparities
- [ ] The ICO's guidance on **AI and data protection** has been reviewed, with particular attention to automated decision-making provisions under UK GDPR Article 22
- [ ] Your team includes individuals with **diverse perspectives** who can identify potential biases that a homogeneous team might overlook during data review and model evaluation
- [ ] There is a **documented remediation process** for when bias is detected — including model retraining, data rebalancing, algorithmic constraints, or removal of problematic features

Section Score: _____/10

7 Audit Summary & Action Plan

#	AUDIT AREA	SCORE	PRIORITY
1	Data Completeness & Availability	/ 10	H / M / L
2	Data Accuracy & Consistency	/ 10	H / M / L
3	Data Freshness & Timeliness	/ 10	H / M / L
4	Data Governance & Documentation	/ 10	H / M / L
5	Data Integration & Accessibility	/ 10	H / M / L
6	Bias & Fairness Assessment	/ 10	H / M / L
TOTAL SCORE		**/ 60**	

Score Interpretation

80–100: Excellent. Your IT setup is well-managed. Focus on continuous improvement and emerging threats.
60–79: Good foundation but gaps exist. Prioritise areas scoring below 6 and create an action plan.
Below 60: Significant gaps that put your business at risk. Consider an urgent review with an IT specialist.

Top 3 Priority Actions:

1. _____
2. _____
3. _____

Additional Notes

_____

_____

_____

_____

_____

_____

Audit completed by: _____ Date: _____ Next review due: _____