

Assignment 1

Data Quality assessment Report

Client: Chief Data Officer

Client Dataset (411K records)

- **Client IDs:** 0 duplicates, unique and clean.
- **Registration Date:** Range from Oct 2011 to Mar 2025 , no missing values.
- **Gender:** 94% missing ('-'), only 6% (male, female).
- **Age:** 66% blank or invalid, only 34% usable.
- **Number of Receipts:** Median = 4, max = 454. 9.3% are outliers (>19).
- **Total Revenue:** Median = 17339, max = 3067388. 8.1% are outliers (>115135).

Transaction Dataset (2.6M records)

- **Receipts:** 62261 duplicates (match on receipt_id + date + revenue).
- **Sale vs Return:** 51166 returns (1.91%), 2569803 sales.
- **Channels:** Retail = 67.5%, Internet = 32.5%.
- **Platforms:** 67.5% blank (' '), only 32.5% split between web (23%) and app (9%).
- **Credit Transactions:** 16.4% involve credit (binary field).
- **Revenue:**
 - Missing in 24522 records (0.92%)
 - Negative in 51166 (1.91%)
 - Outliers in 15.4% (e.g., >7K or <-3.8K)
 - Max value = 539909999 (needs validation)
- **Quantity:**
 - Negative values in 52202 rows (1.95%)
 - Max = 100 units, possibly bulk sales.
- **Top Businesses:** Accessories (29.5%), Small Household Appliances (16.8%), COMFY HOME (12.5%)
- **Top Categories:** Paid Delivery, Smartphone, Smartphone Services, Packages.
- **Top Brands:** 2,120 total, with '-' as brand for 11.8% of records.
- **Cities:** 3,786 unique; top = Kyiv (15.9%), Dnipro (10.1%), Kharkiv (6.5%).
- **Regions:** 26 unique; top = Kyivska (23.7%), Dnipropetrovska (20%), Kharkivska (6.8%).

What the Company Can Do

- Clean gender, age, and platform fields.
- Remove or investigate duplicate receipts.
- Validate negative and missing revenue entries.
- Enrich demographics via surveys, loyalty programs, or third-party data.
- Segment and target high-revenue, high-frequency customers.
- Tag platform sources accurately (web vs. app).
- Analyze and optimize digital channel performance.
- Analyze high-return brands/categories.
- Improve product descriptions, return policies, or quality control.