

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

Data Quality, Cleaning & Integration

Lecture 6 · Friday 25 September 2026

Dr. Innocent Nyalala | BS Semester 5 (DSAI) | IIT Madras Zanzibar

Session Outline

IIT Madras Zanzibar

1. (0–5 min) Recap; why cleaning comes first
2. (5–12 min) Six dimensions of data quality
3. (12–25 min) Missing values: strategies compared on our 24 missing yields
4. (25–33 min) Noise and outliers: binning and the IQR rule
5. (33–42 min) Integration and redundancy: χ^2 and correlation
6. (42–50 min) Exercise, summary, Week 3 preview

Learning Outcomes

IIT Madras Zanzibar

By the end of this session, you will be able to:

Name	the six dimensions of data quality and spot which one a problem breaks
Compare	five ways to handle missing values, with numbers from our data
Smooth	noisy values with binning, and find outliers with the IQR rule — by hand
Merge	two sources: match entities, convert units, resolve conflicts
Test	for redundant attributes with χ^2 (categories) and correlation (numbers)

Recap: Every Formula Assumed Clean Data

IIT Madras Zanzibar

Lectures 4–5

correlation, distance — assumed perfect data

Today

find and fix the problems first

Week 3

transform and reduce the clean data

Garbage in, garbage out: a perfect algorithm on dirty data gives confident, wrong answers. Cleaning is not optional.

IIT Madras Zanzibar

→ **Foundations: What 'Dirty Data' Means**

Handling Missing Values

Noise and Outliers

Data Integration

Redundancy: χ^2 and Correlation

In-Class Exercise

Six Dimensions of Data Quality

IIT Madras Zanzibar

Dimension	Question	Example
Accuracy	is the value correct?	yield 500 t typed instead of 5.00
Completeness	is anything missing?	24 of our 500 farms have no yield
Consistency	do records and sources agree?	'Zanzibar Urban' vs 'Zanz Urban'
Timeliness	is it recent enough?	2019 rainfall used to plan 2026 planting
Believability	do we trust the source?	self-reported yields when a subsidy is at stake
Interpretability	can we understand it?	a column named 'v7' with no description

Use it as a checklist: for every new dataset, ask all six questions before computing anything.

Data Quality, Cleaning & Integration

IIT Madras Zanzibar

Foundations: What 'Dirty Data' Means

→ **Handling Missing Values**

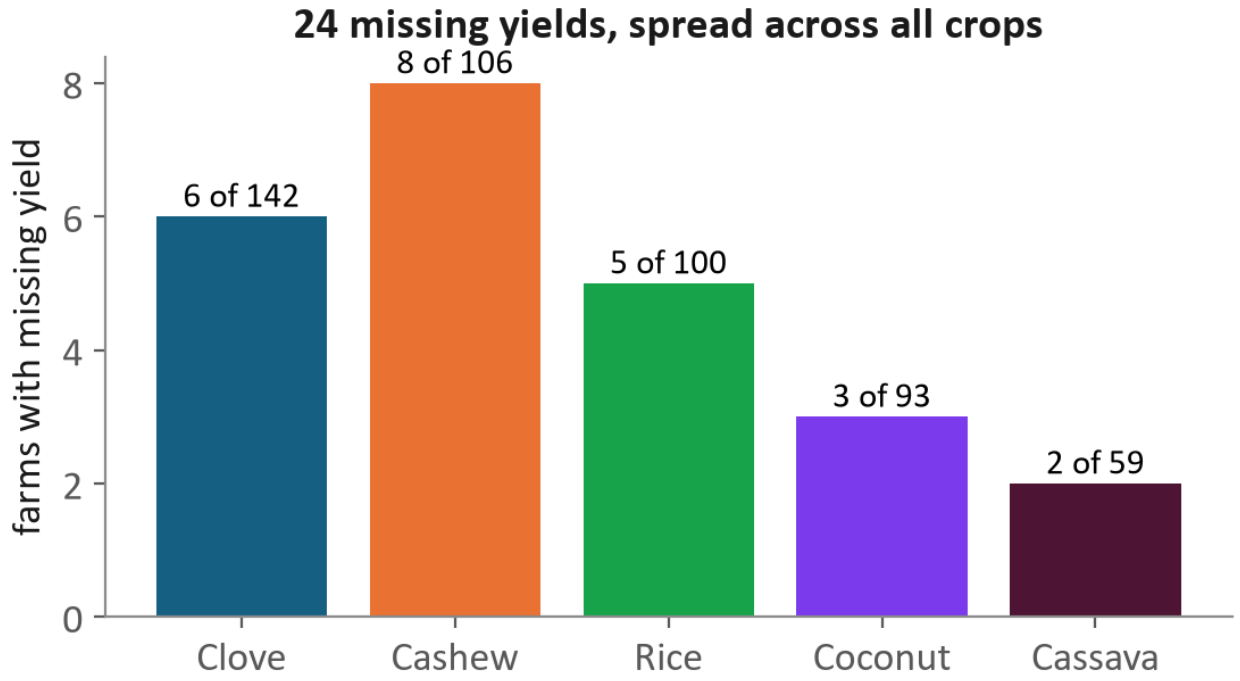
Noise and Outliers

Data Integration

Redundancy: χ^2 and Correlation

In-Class Exercise

Our Data's Missing Values



1 **How many?**
24 of 500 farms (4.8%), only in yield_tonnes

2 **Where?**
every crop — no obvious pattern

3 **Why?**
always ask: random (lost form) or systematic (failed harvests not recorded)?

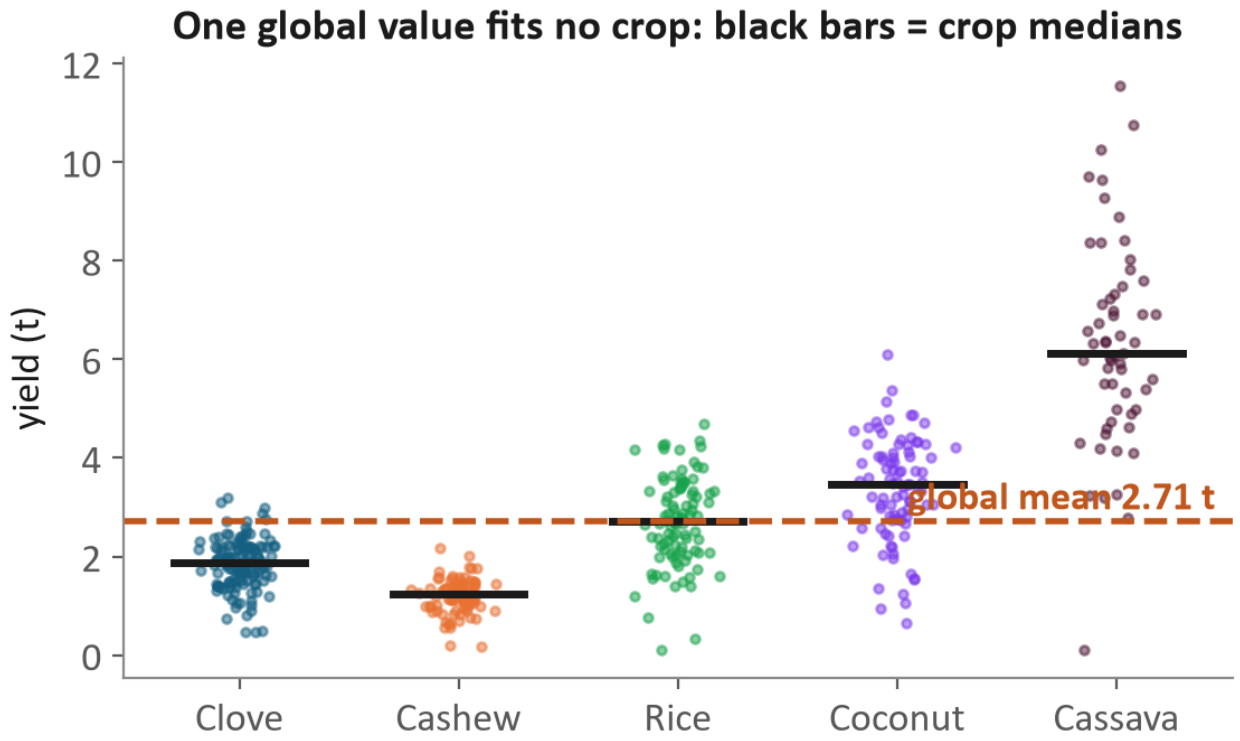
Key question: missing at random → simple fixes are safe. Missing for a reason (e.g. failed harvests left blank) → every simple fix is biased.

Five Strategies for Missing Values

IIT Madras Zanzibar

Strategy	How	Risk
Delete the row	drop farms with a gap	loses data; biased if not missing at random
Fill by hand	an expert checks each record	only works for a few rows
Constant	fill with 0 or 'Unknown'	0 t means crop failure — a false fact
Column mean / median	one value for every gap	ignores crop; shrinks the spread
Group median	median of the same crop	best simple choice; assumes crops are alike inside
Predict it	from related attributes (Weeks 6–7)	more work; best when relations are strong

Why One Global Value Fits No Crop



crop	median yield	global mean
Clove	1.87 t	2.71 t
Cashew	1.25 t	2.71 t
Rice	2.71 t	2.71 t
Coconut	3.46 t	2.71 t
Cassava	6.11 t	2.71 t

Lesson: a missing cassava yield filled with 2.71 t is less than half its typical 6.11 t. Fill within the group.

The Strategies Compared on Our 500 Farms

IIT Madras Zanzibar

strategy	farms kept	mean yield	spread (σ)
delete rows	476	2.706 t	1.808
global mean	500	2.706 t	1.764
constant 0	500	2.576 t	1.857
crop median	500	2.691 t	1.790

1. **Constant 0**
   mean falls to 2.58: 24 fake failed harvests

2. **Global mean**
   spread shrinks to 1.76: 24 identical values

3. **Crop median**
   keeps all 500, realistic mean and spread

Our choice (used from Week 3 on): fill each missing yield with the median yield of its crop.

Bonus: Recover Values From Related Attributes

IIT Madras Zanzibar

farm	crop	revenue ($)	price ($/kg)	yield = revenue / (price × 1000)	crop median
1	Cassava	1,841	0.26	**7.08 t**	6.11 t
2	Rice	2,526	0.69	**3.66 t**	2.71 t
11	Coconut	1,220	0.32	**3.81 t**	3.46 t
20	Cashew	2,549	1.99	**1.28 t**	1.25 t
88	Rice	1,780	0.55	**3.24 t**	2.71 t

Lesson: if a rule links the missing attribute to recorded ones (revenue = yield × 1000 × price), RECOVER the value instead of guessing it.

Data Quality, Cleaning & Integration

IIT Madras Zanzibar

Foundations: What 'Dirty Data' Means

Handling Missing Values

→ **Noise and Outliers**

Data Integration

Redundancy: χ^2 and Correlation

In-Class Exercise

Noise: Smoothing by Binning

IIT Madras Zanzibar

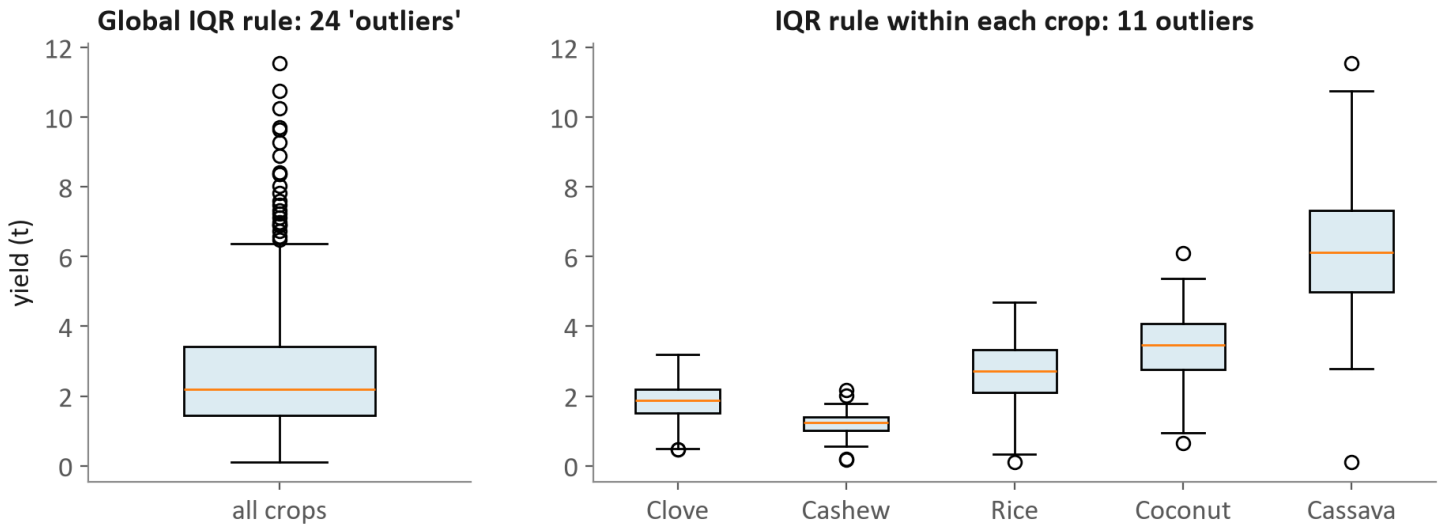
	bin 1	bin 2	bin 3
sorted values ($100s)	4, 8, 9	15, 21, 21	24, 25, 26
smoothed by bin mean	7, 7, 7	19, 19, 19	25, 25, 25
smoothed by bin boundary	4, 9, 9	15, 21, 21	24, 24, 26

1 **Equal-frequency**
same NUMBER of values in each bin

2 **Equal-width**
same RANGE width per bin (Week 3 compares both)

Effect: the trend 7 → 19 → 25 survives; the small wobbles inside each bin disappear.

Outliers: The IQR Rule — and a Trap



$$\text{outlier if} > Q3 + 1.5 \cdot IQR$$

or below Q1 − 1.5·IQR

1 **All crops together**
fence 6.39 t → 24 flagged, ALL cassava

2 **Within each crop**
11 flagged

Trap: the 'outliers' were normal cassava farms. An outlier only means something relative to the right group — investigate before deleting.

Data Quality, Cleaning & Integration

IIT Madras Zanzibar

Foundations: What 'Dirty Data' Means

Handling Missing Values

Noise and Outliers

→ **Data Integration**

Redundancy: χ^2 and Correlation

In-Class Exercise

Integration: Merging Two Sources

IIT Madras Zanzibar

Cooperative log	
id	F-001
farmer	Amina Juma
region	Zanz Urban
revenue	4,602,500 TZS

Government census	
id	001
farmer	A. Juma
region	Zanzibar Urban
revenue	1,900 USD

1. **Same farm?**
   match on a shared key, or name + location

2. **Different units / names**
   convert TZS → USD; one region code list

3. **Values disagree**
   rule: prefer audited, or newer, or average — and write it down

Week 3 Friday: data warehouses do exactly this integration, every night, in the ETL step.

Data Quality, Cleaning & Integration

IIT Madras Zanzibar

Foundations: What 'Dirty Data' Means

Handling Missing Values

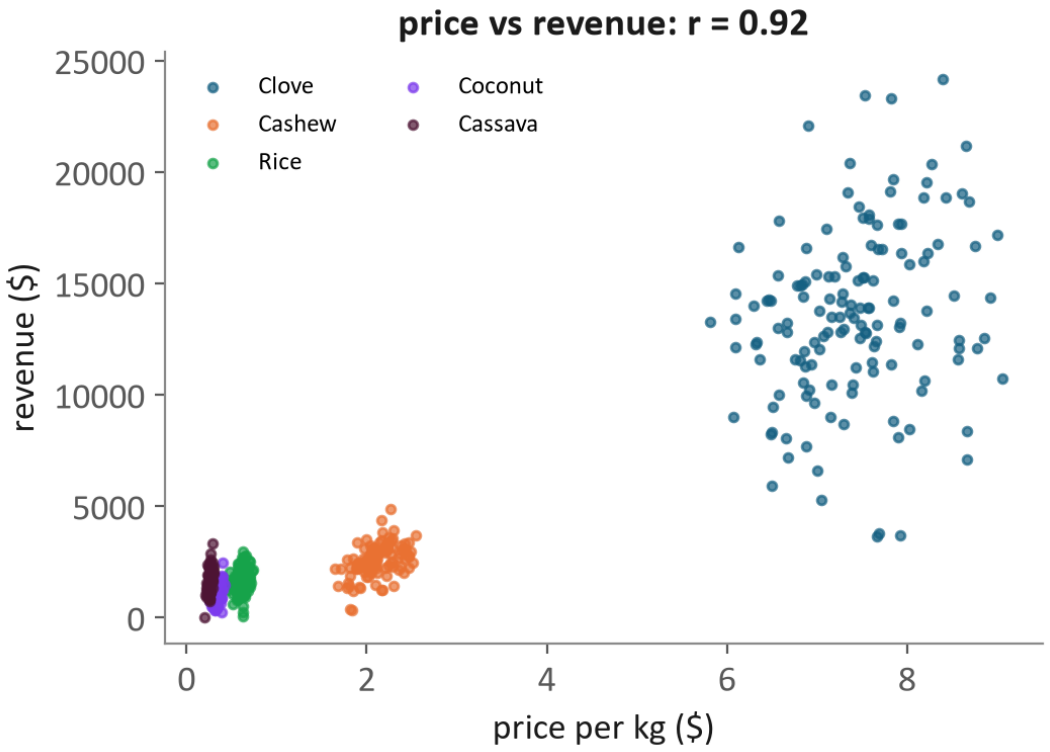
Noise and Outliers

Data Integration

→ **Redundancy: χ^2 and Correlation**

In-Class Exercise

Redundancy Between Numbers: Correlation



1 **price ↔ revenue**
r = 0.92 — nearly the same information

2 **yield ↔ price**
r = -0.42 — related, not redundant

3 **rainfall ↔ others**
|r| ≤ 0.02 — independent

Rule of thumb: |r| > 0.9 → consider dropping one attribute or combining them (PCA, Week 3).

Redundancy Between Categories: The χ^2 Test

IIT Madras Zanzibar

200 farms	irrigated: Yes	irrigated: No	total
fertilizer: Yes	60 (exp. 50)	40 (exp. 50)	100
fertilizer: No	40 (exp. 50)	60 (exp. 50)	100
total	100	100	200

$$\chi^2 = \Sigma\,(O - E)^2 / E$$

E = row total × column total / grand total

1 **Expected**
100 × 100 / 200 = 50

2 **Each cell**
$(60 - 50)^2 / 50 = 2$

3 **Total**
$\chi^2 = 8 > 3.84 \rightarrow$ associated

Critical values: degrees of freedom = (rows − 1)(columns − 1). 1 → 3.84; 4 → 9.49 (5% level). Above it: the attributes are related.

χ^2 on Our Data: Fertilizer Use and Crop

IIT Madras Zanzibar

fertilizer	Clove	Cashew	Rice	Coconut	Cassava
Yes	65 (54.0)	33 (40.3)	38 (38.0)	34 (35.3)	20 (22.4)
No	77 (88.0)	73 (65.7)	62 (62.0)	59 (57.7)	39 (36.6)

observed count (expected if independent)

1 **χ^2 and degrees of freedom**

$\chi^2 = 6.27$, df = (2−1)(5−1) = 4

2 **Compare with 9.49**

6.27 < 9.49 → not related; keep both

Lesson: 'no association' is a real finding: fertilizer use and crop carry different information, so neither is redundant.

Data Cleaning in the Real World

IIT Madras Zanzibar

1 National censuses

duplicate households removed; missing ages imputed by group

2 Mobile money

transactions in TZS and USD unified; duplicates removed

3 Health records (DHIS2)

facility names matched across reporting systems

4 Weather stations

sensor error codes (e.g., 9999) flagged as missing

5 E-commerce

customer records merged across apps

6 Our mini-project

the cleaning section is graded

Common thread: most of a real project's time is spent here — and it decides whether results can be trusted.

Data Quality, Cleaning & Integration

IIT Madras Zanzibar

Foundations: What 'Dirty Data' Means

Handling Missing Values

Noise and Outliers

Data Integration

Redundancy: χ^2 and Correlation

→ **In-Class Exercise**

In-Class Exercise (6 minutes, individually)

IIT Madras Zanzibar

IN-CLASS TASK

Farm sizes (ha), sorted: 0.3, 0.5, 1.0, 1.2, 3.0, 4.5. Two coconut farms have missing yield; coconut median = 3.46 t; global mean = 2.71 t.

1. Make 2 equal-frequency bins and smooth by bin mean.
2. Which value would you use for the missing coconut yields, and why?
3. Q1 = 0.5, Q3 = 3.0. Is 4.5 ha an outlier by the IQR rule?
4. Farmers self-report sizes for a subsidy. Which quality dimension worries you most?

Solution

IIT Madras Zanzibar

1 **Q1**
bins {0.3, 0.5, 1.0} → 0.6 each; {1.2, 3.0, 4.5} → 2.9 each

2 **Q2**
coconut median 3.46 t — the global mean 2.71 ignores that coconut yields differ

3 **Q3**
IQR = 2.5; fence = 3.0 + 3.75 = 6.75; 4.5 < 6.75 → not an outlier

4 **Q4**
believability (and accuracy): a subsidy gives a reason to over-report

Summary & Exam Checkpoints

IIT Madras Zanzibar

Six dimensions

Accuracy, completeness, consistency, timeliness, believability, interpretability.

Missing values

Fill within the group (crop median) — or recover from related attributes. Constant 0 and the global mean distort.

Noise and outliers

Smooth by binning. IQR rule within the right group: all 24 global 'outliers' were normal cassava farms.

Integration

Match entities, convert units, document conflict rules. Redundancy: r (price–revenue 0.92) and χ^2.

What's Next

IIT Madras Zanzibar

Wed — **Data transformation**
normalization and discretization

Thu — **Data reduction**
sampling and PCA; Tutorial 3 and the mini-project

Fri — **Data warehousing**
concepts and architecture

References & Further Reading

- Han, J., Pei, J. & Tong, H. (2023). Data Mining: Concepts and Techniques, 4th ed. — Sections 2.1–2.3 (quality, cleaning, integration).
- Little, R. J. A. & Rubin, D. B. (2019). Statistical Analysis with Missing Data, 3rd ed. Wiley.
- Pearson, K. (1900). On the criterion that a given system of deviations … Philosophical Magazine, 50, 157–175. (χ^2 test.)