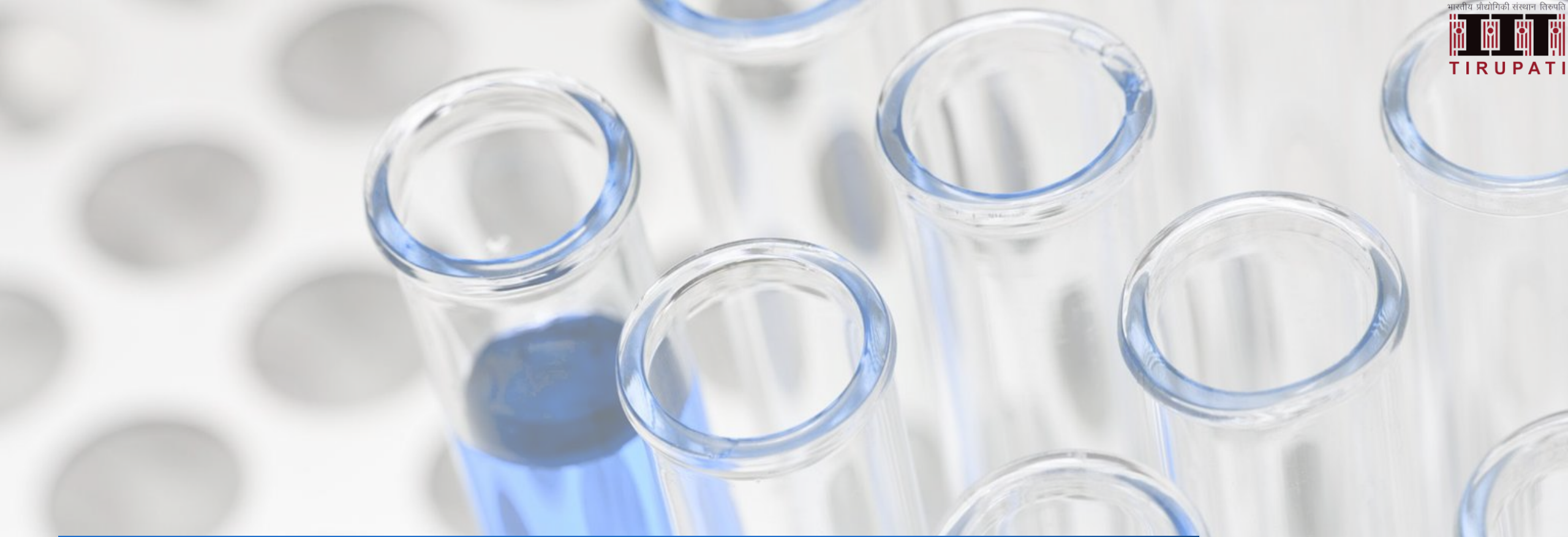


Data Cleaning

Dr. Panchatcharam M

Associate Professor, Department of Mathematics and Statistics, IIT Tirupati

Data Cleaning

From messy, raw data to reliable, analysis-ready datasets

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT
TIRUPATI

Why Data Cleaning Matters

Real-world data is messy — surveys, sensors and manual entry all introduce errors.



Surveys
- Missing answers
- Inconsistent options
- Human bias



Sensors
- Noisy readings
- Calibration drift
- Missing data

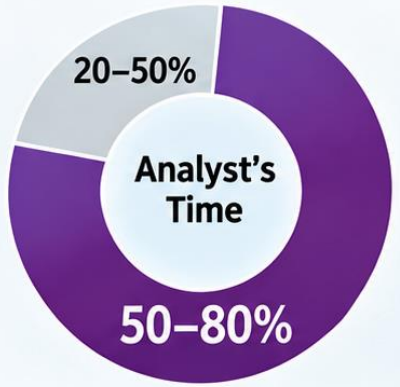


Manual Entry
- Typing mistakes
- Wrong units/formats
- Duplicates

Models are only as good as their inputs: garbage in, garbage out.



Analysts spend 50–80% of their time preparing and cleaning data.



- Data preparation & Cleaning (50–80%)
- Analysis, modeling & insights (20–50%)

Clean data builds trust — results can be reproduced and audited.



Clean, well-documented data



- ✅ Trustworthy results
- ✅ Reproducible workflow
- ✅ Auditable and transparent



भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

What is Data Cleaning?

Raw Data (Messy)

ID	Age	Income	City
1	25	50,000	Tirupati
2	—	70,000	Tirupathi
3	200	60,000	Tirupati
4	28	NA	TPT
5	28	50,000	tirupati



Data Cleaning

Clean Data (Trustworthy)

ID	Age	Income	City
1	25	50,000	Tirupati
2	27	70,000	Tirupati
3	65	60,000	Tirupati
4	28	62,000	Tirupati
5	28	50,000	Tirupati

Detecting and correcting — or removing — inaccurate, incomplete or inconsistent records.

It is not about making data look nicer; it is about **making it trustworthy.**

Cleaning decisions must be **documented** so the analysis can be reproduced.

Every choice (drop, impute, cap) changes the data — always **record what and why.**

The four pillars: quality issues, missing values, outliers, inconsistencies.

The Four Pillars of Data Cleaning

Quality Issues

Errors, invalid values, wrong formats or out-of-range entries.

Missing Values

Absent information (e.g., NA, blank) — remove, impute or collect.

Outliers

Unusually large or small values that can distort analysis — check, cap or treat.

Inconsistencies

Conflicting or non-standard entries (e.g., Tirupati, Tirupathi, TPT) — standardize and align.

Dirty Data vs Clean Data

Dirty data

- Duplicate customer records
- Dates as text, mixed formats
- "N/A", "-999", blanks for missing
- Ages of 0 and 250 together
- "UK" vs "United Kingdom"

Clean data

- One row per real entity
- All dates in ISO format
- One explicit missing marker
- Values within valid ranges
- One agreed label per category

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Data quality issues

Dimensions, common errors and profiling

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

What is Data Quality?

Data quality is the degree to which data is **fit for its intended purpose.**



Quality is relative — a dataset can be good enough for one task, useless for another.

It is measured across **several dimensions**, not by a single score.

Quality must be **assessed before analysis**, not discovered afterwards.

Poor quality is usually **invisible** until a result looks wrong.

Key Dimensions of Data Quality

Accuracy
Correct and free from errors

Completeness
All required data is present

Consistency
Values are uniform across sources and time

Timeliness
Up to date and available when needed

Validity
Conforms to rules, ranges and formats

Uniqueness
No unintended duplicates

High data quality builds **trust** and leads to **reliable, reproducible** results.

Data Quality Dimensions (1 of 2)

Accuracy

- Values reflect reality
- Needs a source of truth
- Passes format checks
- Hardest dimension to test
- Checked by sampling

Completeness

- No required values missing
- All expected records present
- Measured as % non-null
- Blanks may be meaningful
- Compare against totals

Consistency

- No internal contradictions
- Same meaning everywhere
- Agrees across systems
- Units match across columns
- Totals reconcile

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Data Quality Dimensions (2 of 2)

Validity

- Conforms to defined rules
- Correct data type
- Correct format
- Values within range
- Enforced by constraints

Timeliness

- Data is current enough
- Updated as expected
- Lag is documented
- Stale data misleads
- Matters most for decisions

Uniqueness

- No duplicate entities
- One record per real thing
- Keys are truly unique
- Merges create duplicates
- Deduplication required

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Dimension 1: Accuracy

- Accuracy asks: does the recorded value match the real-world value?
- Example: a customer's address on file is their previous address
- Hardest dimension to test — needs an external reference or source of truth
- Checks: compare against master data, verify samples manually, cross-reference systems
- Inaccurate data passes every format check, so it hides in plain sight

Both wrong addresses are perfectly formatted — only an external reference finds them

Customer address file, 6 records

ID	Address on file	Verified address	Status
C-1041	12 Oak St	12 Oak St	Match
C-1042	5 Mill Rd	88 Elm Ave	**Mismatch**
C-1043	7 Fern Cl	7 Fern Cl	Match
C-1044	33 Pine Rd	6 Harbour Vw	**Mismatch**
C-1045	3 Kiln Way	3 Kiln Way	Match
C-1046	9 Vale St	9 Vale St	Match

Illustrative data.

What's wrong

C-1042 and C-1044 moved; the file still holds the old address. Both entries are well-formed, so no format or validity check flags them.

How to catch it

Match against the postal address file, verify a sample by phone, or cross-check recent deliveries.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Dimension 2: Completeness

- Completeness asks: is all the required data present?
- Two levels: missing values within records, and missing records entirely
- Example: 30% of survey respondents left the income field blank
- Measure it as the percentage of non-null values per column
- Missing records are harder to spot — compare counts against an expected total

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Date of birth is filled for only half of the 5,000 customer records

Fill rate by field

Field	Filled	Missing	Fill rate
Postcode	4,995	5	99.9%
Email	4,812	188	96.2%
Consent flag	4,100	900	**82.0%**
Phone	3,950	1,050	**79.0%**
Date of birth	2,600	2,400	**52.0%**

What's wrong

Three fields sit below the 90% threshold. A churn model needing date of birth would drop 2,400 customers.

How to catch it

Count nulls per field, then check whether the missing rows differ from the rest before filling or dropping them.

Illustrative data. n = 5,000 customer records; threshold set at 90%.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Dimension 3: Consistency

- Consistency asks: does the data agree with itself and with other systems?
- Example: order date is later than the delivery date on the same record
- Example: the CRM says a customer is active, billing says the account is closed
- Cross-field rules catch contradictions inside a single record
- Cross-system reconciliation catches disagreements between sources
- Inconsistency often appears after merging data from different teams

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

One customer is Active in two systems and Cancelled in two others

Customer C-1042 across four systems

System	Status	Plan	Last updated
CRM	**Active**	Premium	12 Mar 2026
Billing	**Cancelled**	Premium	03 Feb 2026
Support	**Active**	**Standard**	28 Jan 2026
Marketing	**Cancelled**	Premium	11 Mar 2026

Illustrative data. Status read on 14 Mar 2026.

What's wrong

Every record is valid on its own. The contradiction appears only when the systems are compared: billing stopped invoicing while the CRM still shows a renewal due.

How to catch it

Reconcile the same key across systems on a schedule and name one system the source of truth.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Dimension 4: Validity

- Validity asks: does the value conform to the rules defined for that field?
- Rules cover data type, format, allowed range and permitted category values
- Example: an email field containing "n/a", or a month value of 13
- Validity is the easiest dimension to automate with constraints and schemas
- A value can be perfectly valid and still be inaccurate
- Enforce validity at entry to prevent problems downstream

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Four of five fields break their own rule and can be rejected at entry

Order form, one submitted record

Field	Rule	Value submitted	Valid
Email	Contains @ and domain	priya.sharma#mail.com	**No**
Date of birth	Between 1900 and 2008	14/02/2045	**No**
Postcode	UK format	LS1 4DT	**Yes**
Order quantity	Integer, 1 or more	-3	**No**
Phone	10 to 11 digits	0771234	**No**

Illustrative data.

What's wrong

Each value breaks a stated rule: no @ sign, a birth date in the future, a negative quantity, a phone too short.

How to catch it

Validity is machine-checkable. Enforce the rules at entry; a valid postcode can still be the wrong one, which is accuracy.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

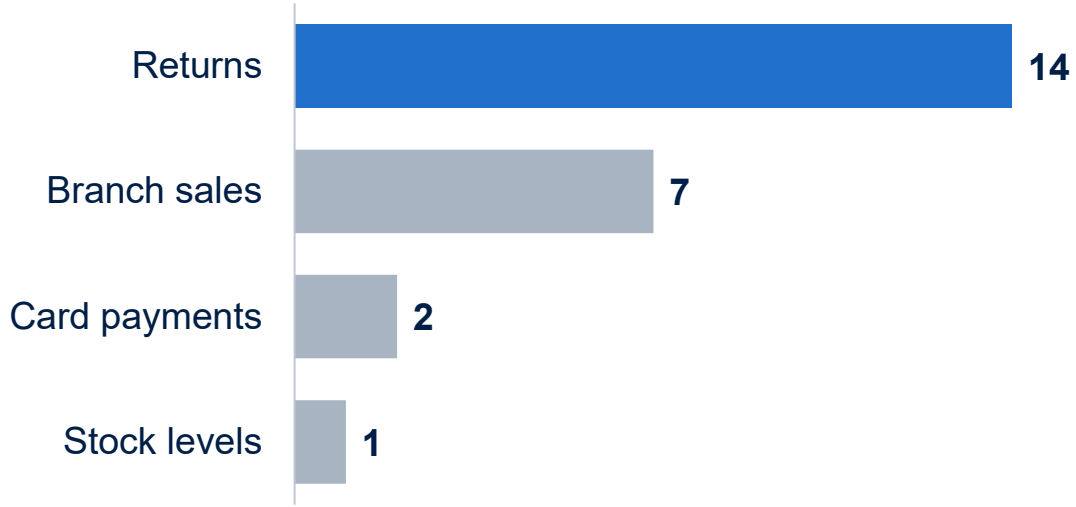
Dimension 5: Timeliness

- Timeliness asks: is the data current enough for the decision at hand?
- Every dataset has a lag between the real event and its record
- Example: stock levels refreshed nightly cannot support live order promising
- Document the refresh frequency and the age of the latest record
- Stale data is dangerous because it looks complete and valid
- Add a last-updated timestamp so freshness can always be checked

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Returns reach the warehouse 14 days late, so a reorder run on 14 April cannot see them

Lag from event date to warehouse load, days



What's wrong

Stock levels land the next day, but the returns feed is two weeks behind. Stock on hand therefore excludes returned units already back on the shelf.

How to catch it

Track load lag per feed against the decision it supports, and stamp every table with its as-of date.

Illustrative data. Event date 1 April 2026 for all four feeds.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Dimension 6: Uniqueness

- Uniqueness asks: is each real-world entity represented exactly once?
- Duplicates inflate counts, distort averages and double-count revenue
- Example: the same customer entered twice with slightly different spellings
- Exact duplicates are easy to find; near-duplicates need fuzzy matching
- Merging datasets is the most common source of duplication
- Define a business key that identifies an entity uniquely

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Three customer IDs are one person, and an email match finds only two of them

Customer records for one person

ID	Name	Email	Phone
C-2011	Priya Sharma	p.sharma@mail.com	**07700 900123**
C-2044	Priya Sharma	p.sharma@mail.com	**07700 900123**
C-2098	P. Sharma	psharma@mail.com	**07700 900123**

Illustrative data.

What's wrong

The customer count reads 3 instead of 1, lifetime value splits three ways, and the person gets three copies of every mailing.

How to catch it

Exact email matching joins only the first two. Match on normalised phone or name plus address to reach C-2098.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Where Quality Problems Come From

People

- Manual typing errors
- Inconsistent conventions
- Misunderstood fields
- Workarounds and free text
- Missing training

Systems

- Faulty sensors
- Failed data transfers
- Schema changes
- Software defects
- Silent truncation

Process

- No validation at entry
- Merging mismatched sources
- Undocumented transforms
- No ownership of data
- Legacy migrations

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Human Data-Entry Errors

- Typographical slips: transposed digits, doubled letters, stray characters
- Inconsistent conventions: "Street" vs "St", uppercase vs lowercase
- Free-text fields where a controlled list should have been used
- Placeholder values entered to bypass mandatory fields, such as 99999
- Data entered into the wrong column or wrong record entirely
- Prevention beats correction: dropdowns, validation and clear field labels

Measurement and Sensor Errors

- Instruments drift out of calibration and record a consistent offset
- Sensors fail intermittently, producing gaps or frozen repeated readings
- Rounding and limited precision hide small but important variation
- Values clipped at the sensor's maximum look like a ceiling in the data
- Environmental interference produces sudden implausible spikes
- Always check sensor metadata: units, precision, calibration date

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Integration and Merge Errors

- Joining on a non-unique key silently multiplies rows
- Mismatched keys drop records that should have matched
- The same field means different things in two source systems
- Different units or currencies combined without conversion
- Overlapping time periods double-count the same events
- Always compare row counts before and after every join

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Processing and Transformation Errors

- Incorrect formulas applied during aggregation or derivation
- Filters that quietly exclude more rows than intended
- Type conversions that truncate decimals or overflow integers
- Sorting a single column and breaking the row alignment
- Reusing an old script after the source schema has changed
- Version-control your cleaning code and test it on known samples

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Common Issue: Duplicate Records

How they appear

- Exact copies of a full row
- Same entity, different spelling
- Repeated imports of a file
- Merge without deduplication
- Multiple systems, one customer

Why they matter

- Counts are inflated
- Revenue is double-counted
- Averages are skewed
- Models over-weight duplicates
- Customer contact is repeated

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Common Issue: Typos and Misspellings

Typical forms

- Transposed characters: Jhon
- Missing or doubled letters
- Wrong keyboard neighbours
- Extra whitespace
- Inconsistent capitalisation

How to tackle

- List distinct values and counts
- Rare values are often typos
- Fuzzy string matching
- Map to a controlled vocabulary
- Fix at source where possible

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Four spellings of one city split 1,312 orders into four rows of the sales report

City field, order table

Value as entered	Orders	Standardised
Manchester	1,042	Manchester
Manchestor	148	Manchester
manchester	97	Manchester
Manchster	25	Manchester
Total	1,312	Manchester

What's wrong

A group-by on city returns four cities. The three misspellings hold 270 orders that never reach the Manchester total.

How to catch it

Rank values by frequency: rare strings close to a common one are almost always typos. Then match against a reference list.

Illustrative data. Edit distance of 1 to 2 characters from the dominant spelling.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Common Issue: Inconsistent Formats

- Dates written as 03/04/2025, 2025-04-03 and 3 April 2025 in one column
- Phone numbers with and without country codes, spaces and dashes
- Currency stored as "£1,200", "1200 GBP" and 1200 in the same field
- Names as "Surname, First" in some rows and "First Surname" in others
- Mixed formats force text storage and block arithmetic and sorting
- Fix by parsing to a proper type and re-serialising to one standard

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

The same date and phone number arrive in four different shapes from four channels

One customer, four source systems

Source	Date as stored	Phone as stored
Web form	2026-03-04	+44 7700 900123
Call centre	04/03/2026	07700900123
Partner feed	**03/04/2026**	**44-7700-900123**
Legacy CRM	4 Mar 26	(07700) 900123

Illustrative data. All four rows describe the same customer and the same day.

What's wrong

04/03/2026 and 03/04/2026 are the same day in two conventions. Parsed naively, one of them becomes 3 April.

How to fix it

Convert on the way in, not on the way out: ISO 8601 for dates, E.164 for phone numbers, one canonical form per field.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Common Issue: Invalid or Impossible Values

- Ages of 0, negative values or 250 in a human population
- Dates in the future for events that have already happened
- Percentages above 100 or probabilities outside zero to one
- Category codes that do not appear in the reference list
- Sentinel values such as -999, 9999 or 1900-01-01 standing in for missing
- Define plausible ranges for every numeric field and test against them

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Six impossible values survive in a table where every field holds the right data type

Values that pass type checks

Field	Value	Why impossible
Age	**214**	No human lives that long
Date of birth	**2045-02-14**	In the future
Order quantity	**-3**	Negative units ordered
Discount	**140%**	Above 100%
Temperature	**-273.9 °C**	Below absolute zero
Delivery date	**Before order date**	Sequence impossible

What's wrong

Each value is a valid number or date, so the loader accepts it. Only a range or a cross-field rule rejects it.

How to catch it

Define a minimum and maximum per column and the handful of rules that link fields, then run them at ingest.

Illustrative data.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Common Issue: Structural Errors

- Header rows repeated in the middle of the data
- Multiple values packed into one cell, such as "Red; Large; 2kg"
- One variable spread across many columns, such as Jan, Feb, Mar
- Merged cells and blank spacer rows imported from spreadsheets
- Footnotes and totals mixed in with the observations
- Reshape to tidy data: one variable per column, one observation per row

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Two values in one cell and one record spread over two rows both break a simple count

Order extract as received

Row	Customer	Products
1	A. Khan	Laptop
2	R. Patel	Monitor; Keyboard
3	S. Ahmed	Printer
4	(blank)	Toner

What's wrong

Row 2 holds two products in one cell; rows 3 and 4 are one order split across lines. A count of orders returns 4 instead of 3.

How to fix it

Make one row one observation and one cell one value before any cleaning: split multi-value cells into rows, fill the carried-down keys.

Illustrative data.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Common Issue: Wrong Data Types

- Numbers stored as text, so sums and comparisons silently fail
- Dates stored as text or as Excel serial numbers
- Identifiers read as numbers, dropping leading zeros from postcodes
- Booleans stored variously as True, T, 1, Yes and Y
- Categorical codes treated as continuous numbers by models
- Always inspect and set the type of every column after loading

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Postcodes read as numbers lose the leading zero and a text price column cannot be summed

Columns loaded with inferred types

Column	Read as	Result
Postcode	Number	01234 becomes 1234
Order ID	Number	Long IDs round to 1.00E+15
Price	Text	"1,299.00" cannot be summed
Order date	Text	No date arithmetic
Is active	Text	"TRUE" is not a boolean

Illustrative data.

What's wrong

The loader guessed each type from the first rows. Nothing errors; the values are simply wrong or unusable from then on.

How to fix it

Declare the schema on read rather than letting the loader infer it, and keep identifiers as text.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Common Issue: Truncation and Padding

Truncation

- Field length limits cut text
- Long names silently shortened
- Decimals lost on conversion
- IDs cut, breaking joins
- Often fails without warning

Padding and whitespace

- Leading and trailing spaces
- "UK " does not equal "UK"
- Non-breaking spaces from web
- Tabs inside text fields
- Trim every text column first

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

A 20-character limit cut 8% of email addresses and trailing spaces broke every join

Values as stored, with lengths

Stored value	Length	Issue
"rachel.okonkwo@exa"	18	Cut at the field limit
"R. Okonkwo "	14	Trailing spaces
" 07700 900123"	13	Leading space
"00123" vs "123"	5 / 3	Zero-padded key mismatch

What's wrong

None of these rows look damaged in a viewer. The truncated address bounces, and the padded keys match nothing on join.

How to catch it

Trim on load, compare field lengths against the source system's limits, and join on normalised keys.

Illustrative data. Quotation marks show the stored string; they are not part of the value.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Common Issue: Encoding Problems

- Accented characters rendered as mojibake, such as "José" becoming "JosÃ©"
- Mixed encodings when files come from different systems or countries
- Smart quotes and dashes pasted from word processors
- Invisible control characters breaking string comparisons
- Emoji and symbols that some tools cannot store
- Standardise on UTF-8 at ingestion and declare the encoding explicitly

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Read in the wrong encoding, one customer name becomes three different strings

Names as displayed after loading

As displayed	Intended	Cause
JosÃ© GarcÃa	José García	UTF-8 read as Latin-1
Zo?? Baker	Zoë Baker	Characters replaced
Müller	Müller	HTML entity not decoded
ï»¿Name	Name	Byte-order mark kept

Illustrative data.

What's wrong

The bytes are intact; the reader is using the wrong character set. Joins on name fail and the damage is often saved back.

How to fix it

Declare UTF-8 end to end, strip the byte-order mark on read, and test every loader with a row of accented names.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

The Cost of Poor Data Quality

Analytical

- Biased estimates
- Wrong conclusions
- Unstable models
- Failed reproducibility
- Lost credibility

Operational

- Wasted rework time
- Duplicate mailings
- Failed deliveries
- Manual reconciliation
- Slower reporting

Strategic

- Poor decisions
- Regulatory breaches
- Lost customer trust
- Missed opportunities
- Higher project costs

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Data Profiling: What It Is

- Profiling is the systematic examination of a dataset before any analysis
- It answers: what is in here, what shape is it in, and what can I trust?
- Produces summary statistics, value distributions and rule violations
- Turns vague worries about quality into a concrete, measurable list
- Should be the first step on every new dataset, without exception
- Repeat profiling after each cleaning step to confirm it worked

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Column-Level Profiling Metrics

For every column

- Data type and storage format
- Count and % of missing values
- Number of distinct values
- Most and least frequent values
- Length of shortest, longest text

For numeric columns

- Minimum, maximum, range
- Mean, median, standard deviation
- Quartiles and interquartile range
- Count of zeros and negatives
- Skewness of the distribution

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Dataset-Level Profiling Checks

- Total row count against the expected or previously reported count
- Number of fully duplicated rows and duplicated key values
- Primary key uniqueness and referential integrity across tables
- Correlations between columns that reveal redundancy or derivation
- Cross-field rule violations, such as end date before start date
- Coverage over time: are any periods missing or unusually sparse?

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Building a Data Quality Report

- Record one row per issue: column, description, severity and row count
- Classify severity: blocking, material, or cosmetic
- State the proposed action for each issue and who must approve it
- Quantify the impact: how many records and which analyses are affected
- Share the report with data owners before making irreversible changes
- Re-run the report after cleaning to show what was resolved

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Worked Example: Profiling a Customer Table

What we found

- 12,480 rows loaded
- payment_date stored as text
- 3.1% of emails blank
- age ranges 0 to 214
- country has 19 spellings

Issues raised

- Type error, blocking
- Completeness, material
- Validity, blocking
- Consistency, material
- 842 duplicate customer IDs

Actions agreed

- Parse dates to ISO format
- Flag blank emails, keep rows
- Cap age at 0 to 110, review
- Map country to ISO codes
- Deduplicate on business key

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Recap and Check Your Understanding

Key takeaways

- Quality has six dimensions
- Problems come from people, systems, process
- Valid is not the same as accurate
- Profiling makes quality measurable
- Document issues before fixing them

Questions to try

- Name the six quality dimensions
- Give an example of valid but inaccurate data
- Why do joins create duplicates?
- Which metrics would you profile first?
- How would you rank issue severity?

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Handling missing values

Mechanisms, detection and imputation

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

What Are Missing Values?

- A missing value is an entry where data was expected but none was recorded
- Stored as NULL, NaN, None or an empty string, depending on the tool
- Often disguised as codes: -999, "N/A", "unknown", 0 or a blank space
- Missing is not the same as zero — zero is a measured value
- The first task is to find every convention used to mean "missing"
- Convert all of them to one explicit missing marker before analysis

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

How Missing Data Appears in Practice

- Survey respondents skip questions they find sensitive or unclear
- Sensors drop out during power loss or network failure
- Fields added to a system later are blank for all historical records
- Joins leave gaps where no matching record existed in the other table
- Conditional questions are blank because they did not apply
- Data entry staff leave optional fields empty under time pressure

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Why Missing Values Are Dangerous

- Many algorithms simply fail or silently drop rows containing gaps
- Dropping rows shrinks the sample and reduces statistical power
- If missingness relates to the outcome, results become biased, not just noisy
- Naive imputation shrinks variance and overstates confidence
- Different tools handle gaps differently, so results stop being reproducible
- The pattern of missingness is itself information worth examining

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Three Mechanisms of Missingness

MCAR

- Missing completely at random
- Unrelated to any variable
- Like a random deletion
- Rare in real data
- Safe to drop rows

MAR

- Missing at random
- Explained by observed data
- Depends on other columns
- Most common assumption
- Imputation works well

MNAR

- Missing not at random
- Depends on the missing value
- Cannot be tested from data
- Most problematic case
- Needs domain knowledge

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

MCAR: Missing Completely at Random

- The probability of a value being missing is the same for every record
- Missingness is unrelated to both observed and unobserved variables
- Example: a laboratory sample is dropped and broken purely by accident
- Example: a page of paper questionnaires is lost in the post
- The remaining data is a genuine random subsample of the full data
- Deletion loses precision but does not introduce bias

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

MCAR: Examples and Implications

Examples

- Sample tube broken in transit
- Random equipment failure
- Questionnaire page lost
- Random subset not processed
- Barcode scan failed at random

Implications

- Complete cases stay unbiased
- Sample size falls, power drops
- Simple deletion is defensible
- Mean imputation still shrinks variance
- Rare in practice — test it

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

MAR: Missing at Random

- Missingness depends on other variables you have observed, not on the missing value
- Example: younger respondents skip the income question more often than older ones
- Because age is recorded, the pattern can be explained and modelled
- The name is misleading — nothing is truly random here
- This is the assumption most imputation methods rely on
- Include the explaining variables in your imputation model

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

MAR: Examples and Implications

Examples

- Income missing more for the young
- Weight missing for a hospital ward
- Sensor gaps during night shifts
- New field blank for old records
- Optional field skipped by region

Implications

- Complete-case analysis is biased
- Imputation can recover structure
- Use the explaining variables
- Multiple imputation preferred
- Document the assumption made

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

MNAR: Missing Not at Random

- The probability of being missing depends on the unobserved value itself
- Example: people with very high incomes decline to state their income
- Example: a scale cannot read weights above its maximum, so they are blank
- Cannot be confirmed or ruled out using the data alone
- Every standard imputation method will produce biased results
- Requires domain knowledge, sensitivity analysis or new data collection

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

MNAR: Examples and Implications

Examples

- High earners hide income
- Severe cases drop out of study
- Sensor saturates above a limit
- Unhappy customers skip surveys
- Failed results not reported

Implications

- Bias cannot be removed by imputation
- Untestable from the data alone
- Model the missingness explicitly
- Run sensitivity analyses
- State the risk in your write-up

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

How to Tell Which Mechanism Applies

- Compare records with and without the missing field on other variables
- If groups differ systematically, the data is unlikely to be MCAR
- Little's MCAR test gives a formal check of the MCAR assumption
- Fit a model predicting missingness from the observed columns
- MNAR can never be proven from the data — it must be argued from context
- Talk to the people who collected the data; they usually know why

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Detecting Missing Values

- Count nulls per column and per row as the very first check
- Search text columns for disguised codes: "NA", "none", "?", "-", "unknown"
- Look for sentinel numbers such as -999, 9999 or 0 used as placeholders
- Check for whitespace-only strings that look empty but are not null
- Compare row counts against the expected number of records
- Standardise every convention to a single explicit missing marker

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Quantifying Missingness

Per column

- Percentage of values missing
- Rank columns worst to best
- Below 5% is usually manageable
- Above 40% questions the column
- Check if pattern is time-related

Per row

- Count of missing fields per row
- Rows missing almost everything
- Drop rows below a completeness threshold
- Check whether gaps cluster
- Report both views together

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Only 58% of rows are complete, though no single column is worse than 21% missing

Missingness in a 5,000-row customer table

Column	Missing	% missing
Income	1,050	21.0%
Date of birth	600	12.0%
Phone	400	8.0%
Email	150	3.0%
Rows with any gap	**2,100**	**42.0%**

Illustrative data. n = 5,000 rows.

What the numbers say

Per-column rates look survivable. The row-level figure is what a complete-case model actually sees: 2,100 rows carry at least one gap.

What to report

Both figures, always: the rate per column and the share of complete rows. One without the other hides the problem.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Visualising Missingness

- A missingness matrix shows gaps as stripes across rows and columns
- A bar chart of missing percentage per column ranks the worst offenders
- A heatmap of correlations between missing indicators reveals linked gaps
- A dendrogram groups columns that tend to be missing together
- Plotting missingness against time exposes outages and collection changes
- A picture makes the pattern obvious far faster than a table of counts

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Missingness Patterns Across Columns

- Univariate: only one column has gaps — the simplest case to handle
- Monotone: once a record drops out it is missing everything afterwards
- Arbitrary: gaps scattered unpredictably across many columns
- Structural: blank because the question did not apply to that record
- Structural gaps should be labelled "not applicable", never imputed
- The pattern determines which methods are usable

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Three Strategies for Missing Data

Delete

- Remove rows or columns
- Simple and transparent
- Loses information
- Biased unless MCAR
- Use when gaps are few

Impute

- Fill with estimated values
- Keeps sample size
- Understates uncertainty
- Quality depends on method
- The usual choice

Model

- Handle gaps within the model
- Some algorithms accept nulls
- Multiple imputation pools results
- Most statistically sound
- More work to implement

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Listwise Deletion (Complete-Case Analysis)

How it works

- Drop any row with a gap
- Analyse only complete records
- The default in many tools
- Simple and reproducible
- Easy to describe in a report

Trade-offs

- Unbiased only if MCAR
- Can discard most of the data
- Power falls sharply
- Biased under MAR and MNAR
- Always report how many rows went

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Dropping every incomplete row costs 42% of the sample and tilts it toward higher earners

Full sample against complete cases

Measure	Full sample	Complete cases
Rows	5,000	2,900
Mean income	$47,000	**$52,400**
Share under 30	28%	**19%**
Default rate	5.6%	**4.1%**

What's wrong

The 2,100 dropped rows are not a random subset. The surviving sample earns more, skews older and defaults less than the real population.

When it is acceptable

Gaps are few and missing completely at random. Compare the dropped rows against the kept ones before deciding.

Illustrative data. n = 5,000 rows, of which 2,900 are complete.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Pairwise Deletion

How it works

- Use all available data per calculation
- Each statistic uses its own rows
- Common for correlation matrices
- Retains more information
- No single analysis sample

Trade-offs

- Sample size differs per statistic
- Results can be inconsistent
- Correlation matrix may be impossible
- Hard to report clearly
- Use with caution

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Dropping Columns Entirely

- Sometimes the sensible action is to remove the variable, not the rows
- Consider dropping when well over half the values are missing
- Consider dropping when the column duplicates another, more complete field
- Never drop a column simply because it is inconvenient to impute
- Check first whether the column is essential to the research question
- Record the decision and the missing percentage that justified it

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Dropping the 64%-missing column removes 3,200 gaps and the second strongest predictor

Candidate columns for dropping

Column	% missing	Rank in model
Income	**64%**	**2 of 20**
Tenure	9%	15 of 20
Channel	2%	14 of 20
Last login	31%	6 of 20

The trade-off

Income is the emptiest column and the second most useful one. Tenure is almost full and barely matters.

How to decide

Weigh the gap rate against the column's contribution. A missing-indicator flag plus imputation often beats dropping it outright.

Illustrative data. Rank by permutation importance in a 20-feature model, n = 5,000.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

When Deletion Is Acceptable

Reasonable

- Gaps well under 5%
- Mechanism is plausibly MCAR
- Large dataset, power to spare
- Rows are almost entirely empty
- Column is redundant

Risky

- Large fraction of rows affected
- Missingness relates to outcome
- Small or precious sample
- Subgroups lost disproportionately
- No record of what was removed

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Imputation: The Basic Idea

- Imputation replaces each gap with a plausible estimated value
- The goal is to preserve relationships, not to invent facts
- Every imputed value carries uncertainty that must be acknowledged
- Simple methods distort variance; sophisticated methods cost effort
- Always fit the imputation on training data only, never the test set
- Keep a flag column recording which values were imputed

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Mean Imputation

How it works

- Replace gaps with column mean
- Numeric columns only
- Fast and easy to explain
- Preserves the overall mean
- Common naive baseline

Problems

- Shrinks variance artificially
- Weakens correlations
- Distorts the distribution shape
- Very sensitive to outliers
- Understates standard errors

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Mean imputation holds the mean at 40 but shrinks the standard deviation by 16%

Customer age, 10 records

Record	Raw	Imputed
1	22	22
2	28	28
3	34	34
4	40	40
5	46	46
6	52	52
7	58	58
8	—	**40**
9	—	**40**
10	—	**40**

Population standard deviation; illustrative data.

Effect on the statistics

Statistic	Raw	Imputed
Mean	40.0	40.0
Std deviation	12.0	10.0
Variance	144.0	100.8

Blue values are imputed. All three sit exactly at the centre, so the mean is unchanged while the spread collapses: standard deviation falls 16% and variance 30%. A model fitted on the filled column reports standard errors that are too small.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Median Imputation

How it works

- Replace gaps with column median
- Numeric columns only
- Robust to extreme values
- Better for skewed data
- Preferred over mean by default

Problems

- Still shrinks variance
- Still ignores other columns
- Creates a spike at the median
- Does not preserve correlations
- Only a reasonable baseline

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Median imputation leaves the centre at 40; a mean fill moves it to 44

Annual income, $000, 10 records

Record	Raw	Median fill
1	28	28
2	32	32
3	35	35
4	38	38
5	42	42
6	46	46
7	55	55
8	184	184
9	—	40
10	—	40

Illustrative data. Median of the 8 observed values = 40.0.

Where the centre lands

Statistic	Observed	Median fill	Mean fill
Fill value	—	40.0	57.5
Median	40.0	40.0	44.0
Mean	57.5	54.0	57.5

Blue values are imputed. One high earner at 184 drags the mean to 57.5, above 7 of the 8 observed records. Filling with that mean shifts the median from 40 to 44; filling with the median of 40 leaves it where it was.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Mode Imputation for Categories

How it works

- Fill with most frequent category
- The standard categorical baseline
- Simple and quick
- Keeps the column usable
- No new categories created

Problems

- Over-represents the majority class
- Hides genuine minority groups
- Ignores other columns entirely
- Poor when classes are balanced
- Consider a 'Missing' category instead

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Filling 40 blanks with ‘Email’ lifts its share from 58% to 66% and shrinks every other channel

Preferred contact channel, 200 customers

Channel	Recorded	Share before	Share after
Email	92	57.5%	**66.0%**
Phone	38	23.8%	19.0%
Post	20	12.5%	10.0%
SMS	10	6.3%	5.0%

What's wrong

Forty blanks all become Email. The majority class grows while the three minority channels shrink, though not one customer changed channel.

What to do instead

Use an explicit ‘Unknown’ category when the blank may mean something; keep mode fill as a baseline only.

Illustrative data. 160 recorded responses, 40 blanks; shares of recorded and of all 200.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Constant and Indicator Imputation

- Fill gaps with an explicit constant such as "Unknown" or "Not recorded"
- Treats missingness as a category in its own right, which is honest
- Very useful when the fact of being missing is itself informative
- Tree-based models handle such a category naturally and often well
- Avoid numeric constants like 0 or -1, which distort statistics badly
- Pair with an indicator flag so the model can learn from the gap

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Labelling the gap ‘Not recorded’ keeps the rows and lets the model use the blank itself

Employment status, five records

ID	Employment	Filled value	Flag
C-1	Employed	Employed	0
C-2	—	Not recorded	1
C-3	Student	Student	0
C-4	—	Not recorded	1
C-5	Retired	Retired	0

Illustrative data. Flag is 1 where the original value was blank.

What it buys

The category and the flag both survive into the model. A tree can split on ‘Not recorded’ directly instead of guessing a status.

What to avoid

Numeric constants. Filling a missing income with 0 or -1 drags the mean down and invents a cluster that is not there.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Forward and Backward Fill

How it works

- Carry the last observation forward
- Or pull the next value backward
- Designed for ordered time series
- Assumes values persist
- Simple to apply and explain

Cautions

- Wrong for fast-changing series
- Long gaps propagate stale values
- Never use on unordered rows
- Set a maximum fill length
- Can create artificial flat lines

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Forward fill holds the sensor at 20.1 for 45 minutes while the true reading climbed to 22.4

Furnace temperature, °C, 15-minute readings

Time	Reading	Forward fill	Backward fill
09:00	20.1	20.1	20.1
09:15	—	**20.1**	**22.4**
09:30	—	**20.1**	**22.4**
09:45	—	**20.1**	**22.4**
10:00	22.4	22.4	22.4

Illustrative data. Readings either side of the gap are the only observed values.

What's wrong

Forward fill draws a flat line through the gap and then a jump at 10:00. Backward fill makes the same error in reverse, too early.

How to use it safely

Cap the fill at one or two steps, and leave longer gaps missing rather than carrying a stale value.

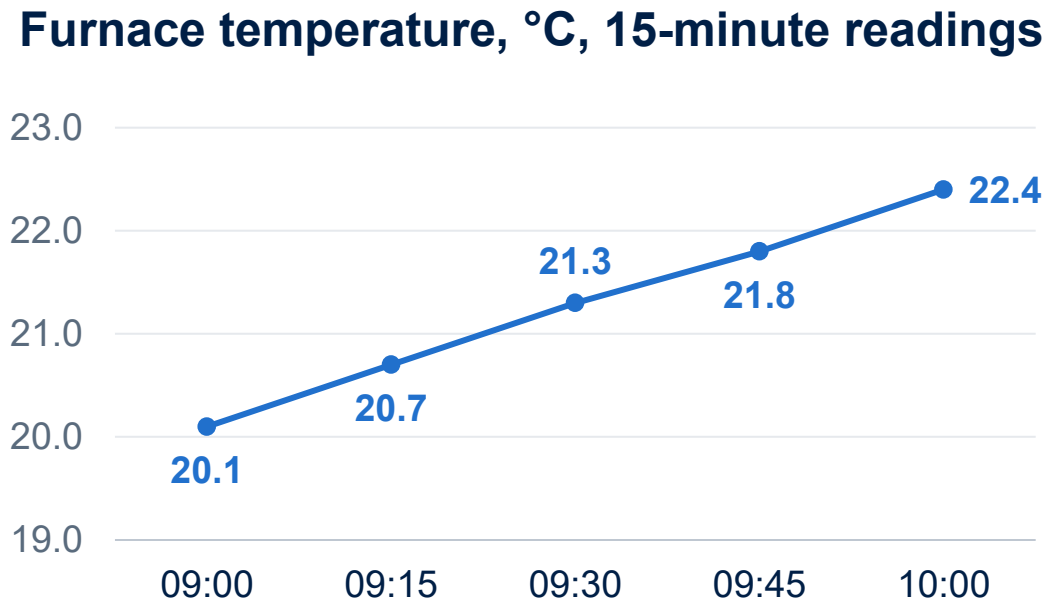
भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Interpolation for Time Series

- Estimate gaps from the values on either side rather than repeating one
- Linear interpolation draws a straight line between the two neighbours
- Spline and polynomial methods follow smooth curves through the data
- Seasonal decomposition respects repeating daily or yearly patterns
- Works well for short gaps in smoothly varying measurements
- Never interpolate across a known outage or a change in instrument

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Linear interpolation recovers the climb that a forward fill flattens



What it does

The gap is estimated from the values either side: 20.7, 21.3, 21.8 instead of three copies of 20.1. The slope through the gap is preserved.

When not to use it

Never across a known outage or an instrument change. If the furnace was off, the true readings were nowhere near that line.

Illustrative data. Observed readings at 09:00 and 10:00 only; 09:15 to 09:45 estimated.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Regression Imputation

How it works

- Predict the gap from other columns
- Fit a model on complete rows
- Uses relationships in the data
- Far better than a column mean
- Works under the MAR assumption

Cautions

- Predictions are too perfect
- Correlations become inflated
- Add random noise to correct this
- Risk of leakage from the target
- Refit within cross-validation folds

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Every imputed income lands exactly on the regression line; real ones scatter ±$4,800

Income fitted from age and years in job, $

ID	Age	Years	Income	Residual
C-1	34	6	48,000	−4,300
C-3	29	3	39,000	−4,850
C-5	45	15	79,000	+4,750
C-2	41	12	66,650	0
C-4	52	20	87,200	0

Illustrative data. Fitted model: Income = 15,000 + 850 × age + 1,400 × years in job.

What's wrong

Observed rows have residuals of several thousand. Imputed rows have a residual of exactly zero, so the fit looks tighter than reality.

The correction

Add random noise drawn from the residual spread to each filled value, and fit the imputation model inside the cross-validation folds.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

KNN Imputation

How it works

- Find the k most similar rows
- Average or vote on their values
- Captures local structure
- Handles numeric and categorical
- Popular and easy to apply

Cautions

- Requires scaled features
- Slow on very large datasets
- Sensitive to the choice of k
- Struggles in high dimensions
- Distance needs complete columns

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

The three nearest rows put the missing income at $58,300, not the column mean of $47,000

Nearest neighbours of record C-7

Record	Age	Spend	Tenure	Income	Dist.
C-7 (gap)	38	420	4	—	—
C-2	39	410	4	58,000	0.21
C-5	37	435	5	61,000	0.34
C-9	40	405	3	56,000	0.38
Fill for C-7				**58,300**	

How the fill is formed

C-7 has no income. Its three closest rows on age, spend and tenure average $58,300, which replaces the gap.

What to watch

Features must be scaled or monthly spend swamps age and tenure. A larger k drags the fill back toward the column mean.

Illustrative data. k = 3, Euclidean distance on standardised features.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Multiple Imputation (MICE)

- Create several complete datasets, each with different plausible fills
- Each column is imputed in turn using all the others, cycled repeatedly
- Analyse every dataset separately, then pool the results
- Rubin's rules combine the estimates and their uncertainty correctly
- This is the only common method that reflects imputation uncertainty
- Standard practice in medical and social research under MAR

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Pooling five imputations widens the standard error from 119 to 143 — the honest figure

Age coefficient across five imputed datasets

Dataset	Imputed income	Age coefficient	Std error
1	64,200	607	118
2	69,800	662	121
3	61,500	702	115
4	67,100	742	119
5	72,300	797	123
Pooled	**—**	**702**	**143**

Illustrative data. m = 5; pooled standard error by Rubin’s rules.

What MICE does

Each dataset draws a different plausible income for the same gap, so the five estimates disagree. That disagreement is real uncertainty.

Why it matters

Rubin’s rules add the spread between datasets to the error within them. A single fill would report 119 and overstate confidence.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Comparing Imputation Methods

Simple

- Mean, median, mode
- Fast, easy to explain
- Shrinks variance
- Ignores other columns
- Baseline only

Model-based

- Regression, KNN
- Uses relationships
- More accurate fills
- Still one value per gap
- Needs care with leakage

Multiple

- MICE and variants
- Several datasets pooled
- Honest uncertainty
- Gold standard under MAR
- Most computation

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

On the same 200 held-out gaps, KNN and regression cut the error by a third against a mean fill

Imputation error, income column

Method	Fill for C-2	RMSE	Variance kept
Mean	47,000	14,800	71%
Median	44,500	14,300	73%
Regression	66,650	9,100	88%
KNN (k = 5)	58,300	9,600	92%
MICE (m = 5)	**varies**	**9,400**	**101%**

Reading the table

A mean fill keeps only 71% of the original variance. MICE is the only method whose spread matches the complete data.

How to choose

Pick on the job: a quick baseline, a single accurate point prediction, or an estimate that carries its own uncertainty.

Illustrative data. 200 values hidden at random and re-imputed; RMSE in $, variance against the complete column.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Missing Indicator Flags

- Add a binary column recording whether the original value was missing
- Lets a model learn from the fact of missingness itself
- Essential when data may be MNAR and the gap carries meaning
- Keeps your cleaning transparent and fully auditable
- Costs one extra column per affected variable
- Combine with any imputation method rather than replacing it

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Customers who left income blank defaulted at 11.8%, against 4.2% for everyone else

Default rate by whether income was recorded

Measure	Income recorded	Income missing
Records	4,100	900
Default rate	4.2%	**11.8%**
Model AUC	0.71 without flag	**0.76 with flag**

Why the flag helps

The blank is not random. Dropping those 900 rows or filling them silently throws away the strongest single signal in the file.

What it costs

One binary column per affected variable. Add it alongside any imputation method rather than instead of one.

Illustrative data. n = 5,000 accounts; AUC from the same model with and without the flag.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Choosing a Strategy: A Decision Framework

- Step 1: quantify how much is missing, per column and per row
- Step 2: investigate why — talk to the data owners about the mechanism
- Step 3: if gaps are tiny and plausibly MCAR, deletion is defensible
- Step 4: under MAR, use model-based or multiple imputation
- Step 5: under MNAR, add indicators and run a sensitivity analysis
- Step 6: document every decision and compare results with and without it

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Worked Example: A Survey Dataset

The situation

- 4,200 respondents
- income missing for 22%
- age and region complete
- Young skip income most
- Mechanism judged MAR

What we did

- Ruled out simple deletion
- Would have lost 22% of rows
- Chose multiple imputation
- Used age, region, education
- Added a missing flag

The outcome

- Full sample retained
- Age gradient preserved
- Wider, honest confidence intervals
- Sensitivity check run
- Decisions documented

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Recap and Check Your Understanding

Key takeaways

- Find every disguised missing code
- MCAR, MAR and MNAR differ
- Deletion is safe only under MCAR
- Mean imputation shrinks variance
- Multiple imputation is the gold standard

Questions to try

- Define MAR in your own words
- Why is MNAR untestable?
- When is listwise deletion fine?
- What does a missing flag add?
- Which method for a time series gap?

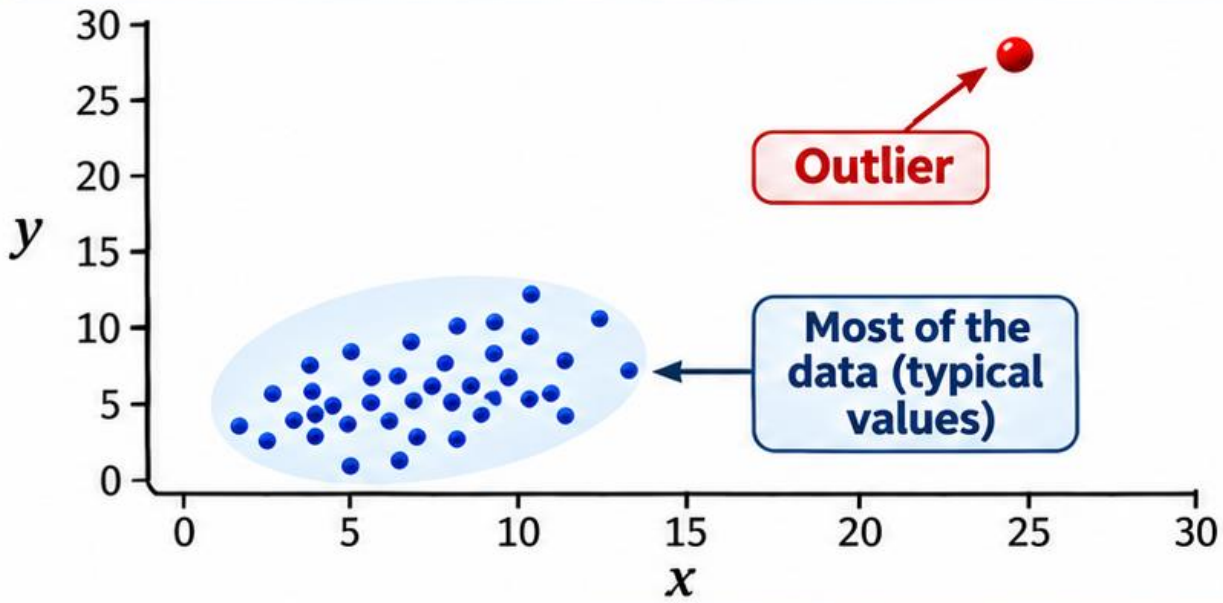
Outlier detection and treatment

Finding and handling extreme values

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT
TIRUPATI

What is an Outlier?

An **outlier** is an observation that lies far away from the rest of the data.



Distance can be in **one variable**, or in the **combination of several**.

Outliers are **not automatically errors** — some are the **most important records.**

Fraud, equipment failure and rare disease all appear as outliers.

The question is never **"is it extreme?"** but **"why is it extreme?"**

Investigate before you decide to remove anything.

Examples of Outliers

Fraud
(unusual transaction)

Equipment failure
(abnormal reading)

Rare disease
(very unusual case)

Outliers can be **problems, signals** or **opportunities.**

An outlier is not just an extreme value — it is a **valuable clue** that deserves investigation.

Three Kinds of Outlier

Global

- Extreme against whole dataset
- A height of 2.6 metres
- Easiest type to spot
- Found by simple thresholds
- Univariate methods work

Contextual

- Extreme only in context
- 30°C is normal in July
- Same value odd in January
- Needs the context variable
- Common in time series

Collective

- A group of points together
- Each value looks normal
- The pattern is anomalous
- A flat line from a sensor
- Needs sequence analysis

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Error or Genuine Extreme Value?

Likely an error

- Physically impossible value
- Decimal point shifted
- Sentinel code such as 9999
- Unit mixed up: kg versus lb
- Contradicts other fields

Likely genuine

- Plausible but rare
- Consistent with other fields
- Confirmed by a second source
- Repeats across time
- Known real phenomenon

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Why Outliers Matter

- The mean and standard deviation are pulled strongly by extreme values
- Linear regression coefficients can be driven by a single point
- Distance-based methods such as KNN and k-means are badly distorted
- Scaling is distorted: one extreme value squashes everything else
- Correlations can be created or destroyed by a handful of points
- Some methods are robust: medians, trees and rank-based statistics

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

A Workflow for Handling Outliers

- Step 1: detect candidates using statistical and visual methods
- Step 2: inspect the actual records rather than just the numbers
- Step 3: classify each as data error, genuine extreme, or uncertain
- Step 4: choose a treatment that matches the classification
- Step 5: analyse with and without them to test how much they matter
- Step 6: document every point removed or altered, and why

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Detection Approach 1: Statistical Rules

- Compare each value against a summary of the distribution
- The Z-score measures distance from the mean in standard deviations
- The IQR rule uses quartiles and is robust to extreme values
- The modified Z-score uses the median and is more robust still
- Fast, transparent and easy to justify to a reviewer
- Assume a roughly known shape, so check the distribution first

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

The Z-Score Method

- Z equals the value minus the mean, divided by the standard deviation
- Flag any observation with an absolute Z-score above 3 as a candidate
- A threshold of 3 captures about 99.7% of a normal distribution
- Only meaningful when the data is approximately normally distributed
- The mean and standard deviation are themselves distorted by outliers
- This masking effect can hide the very points you are looking for

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

One transaction scores 3.58 standard deviations above the mean and is flagged

Transaction amounts against the z-score rule

Transaction	Amount	Z-score	Flagged
T-1	$142	-0.13	No
T-2	$168	0.30	No
T-3	$95	-0.92	No
T-4	$318	2.80	No
T-5	$365	3.58	Yes

Illustrative data. Column mean $150, standard deviation $60, threshold |z| > 3.

How it works

Each amount is restated as its distance from the column mean in standard deviations. Anything past 3 is flagged.

The weakness

The mean and the standard deviation are themselves inflated by the extreme value, so a single outlier can hide a second one.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

The IQR Rule

How it works

- IQR equals Q3 minus Q1
- Lower fence: Q1 minus 1.5 × IQR
- Upper fence: Q3 plus 1.5 × IQR
- Anything outside is flagged
- Basis of the box plot

Why it is preferred

- Quartiles resist extreme values
- No normality assumption needed
- Works on skewed data
- Use 3 × IQR for extreme outliers
- Simple to compute and explain

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

The IQR rule sets the fence at $313 and catches the $318 transaction the z-score missed

Quartiles and fences, same transaction column

Statistic	Value
Q1	$118
Q3	$196
IQR	$78
Lower fence	$1
Upper fence	**$313**
Flagged	**$318 and $365**

How it works

The fences sit 1.5 interquartile ranges beyond the quartiles. Quartiles ignore the extremes, so the fence stays where the bulk of the data is.

What to watch

The 1.5 multiplier is a convention, not a law. On heavily skewed data it flags a long tail that is perfectly genuine.

Illustrative data. Same column as the z-score example; fences at Q1 − 1.5 × IQR and Q3 + 1.5 × IQR.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

The Modified Z-Score

- Replaces the mean with the median and uses the median absolute deviation
- MAD is the median of the absolute distances from the median
- The score is 0.6745 times the deviation, divided by the MAD
- A common flagging threshold is an absolute score above 3.5
- Far more robust than the standard Z-score for small samples
- Fails when more than half the values are identical, giving a MAD of zero

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Switching to the median and MAD lifts the $318 transaction to 3.78 and flags it

Z-score against modified z-score

Amount	Z-score	Modified z	Flagged by
$142	-0.13	-0.18	Neither
$210	1.00	1.35	Neither
$318	**2.80**	**3.78**	**Modified only**
$365	3.58	4.83	Both

What changes

The median and the MAD are not moved by the extreme value, so the $318 transaction no longer hides behind the $365 one.

When to use it

Whenever the column already contains extremes, which is most real data. It is the safer default of the two rules.

Illustrative data. Modified z = 0.6745 × (x − median) / MAD; median $150, MAD $30; threshold 3.5.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Comparing Statistical Methods

Z-score

- Mean and standard deviation
- Assumes normality
- Threshold about 3
- Distorted by outliers
- Simple to explain

IQR rule

- Quartiles only
- No distribution assumption
- 1.5 × IQR fences
- Robust and popular
- Default choice

Modified Z

- Median and MAD
- Highly robust
- Threshold about 3.5
- Good for small samples
- Fails if MAD is zero

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Detection Approach 2: Visual Inspection

- Plots reveal shapes and clusters that summary statistics conceal
- Box plots show the quartiles and flag points beyond the fences
- Histograms expose skew, gaps and unexpected secondary peaks
- Scatter plots reveal points that are odd only in combination
- Time-series line plots make spikes, drops and flat lines obvious
- Always plot the data before trusting any automatic threshold

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Reading a Box Plot

What it shows

- Box spans Q1 to Q3
- Line inside is the median
- Whiskers reach 1.5 × IQR
- Points beyond are flagged
- Width shows the spread

How to read it

- Off-centre median means skew
- Long whisker on one side: tail
- Many flagged points: heavy tails
- Compare boxes across groups
- Always pair with a histogram

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Histograms and Scatter Plots

- A histogram shows where the mass of the distribution sits
- Isolated bars far from the main body indicate candidate outliers
- Two separate peaks may mean two populations mixed together
- A scatter plot exposes points that break the general relationship
- A point can be typical in each variable yet extreme in combination
- Colour by group to see whether extremes belong to one subgroup

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Detection Approach 3: Model-Based Methods

- Statistical rules examine one column; models examine many at once
- They learn what "normal" looks like and score how unusual each row is
- Essential when outliers are only visible across several variables
- Isolation Forest, Local Outlier Factor and DBSCAN are common choices
- Outputs are scores, so you still have to choose a cut-off
- Scale your features first — distances are meaningless otherwise

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Isolation Forest

The intuition

- Split the data at random
- Outliers separate quickly
- Few splits to isolate a point
- Average path length is the score
- Builds many random trees

In practice

- Handles many dimensions well
- Fast on large datasets
- Set the expected contamination
- No distance calculations
- Little tuning required

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Local Outlier Factor (LOF)

The intuition

- Compares local densities
- A point in a sparse pocket scores high
- Density judged by k neighbours
- Finds local, not just global, outliers
- Score near 1 means normal

In practice

- Good for clustered data
- Sensitive to the choice of k
- Requires scaled features
- Slower on large datasets
- Struggles in high dimensions

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

DBSCAN for Outlier Detection

The intuition

- A density-based clustering method
- Groups points packed closely together
- Points in no cluster are noise
- Noise points are your outliers
- No need to set cluster count

In practice

- Two parameters: eps and minPts
- Results depend heavily on eps
- Requires scaled features
- Handles odd cluster shapes
- Struggles with varying density

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Comparing Model-Based Methods

Isolation Forest

- Random partitioning
- Scales to big data
- Handles many features
- Set contamination rate
- Robust default choice

LOF

- Local density ratio
- Finds local anomalies
- Needs scaled features
- Sensitive to k
- Slower on large data

DBSCAN

- Density clustering
- Noise points are outliers
- Two parameters to tune
- Finds odd shapes
- Struggles with mixed density

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Multivariate Outliers

- Some records are unremarkable in every column but impossible in combination
- Example: a height of 1.9 metres with a weight of 40 kilograms
- Univariate rules will never flag either value on its own
- Mahalanobis distance measures distance while accounting for correlation
- Scatter plot matrices and PCA projections make such points visible
- Always check pairs of related variables, not just each column alone

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Outliers in Time Series

- Point anomalies: a single spike far from its neighbours
- Contextual anomalies: normal overall, but wrong for that season or hour
- Collective anomalies: a run of values forming an impossible pattern
- Flat lines repeating one value usually mean a stuck sensor
- Compare each point against a rolling median rather than the global mean
- Decompose into trend, season and residual, then examine the residual

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Setting Sensible Thresholds

- Every method needs a cut-off, and the cut-off is a judgement call
- Tighter thresholds flag more points, including many false alarms
- Looser thresholds miss genuine problems in the data
- Balance the cost of investigating against the cost of missing an error
- Domain limits often beat statistical rules: a human age above 120
- Whatever you choose, state the threshold explicitly in your write-up

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Four Ways to Treat an Outlier

Remove

- Delete the record
- Only for confirmed errors
- Loses information
- Can bias the sample
- Always document it

Correct

- Fix the underlying error
- Best option when possible
- Needs a source of truth
- Recheck the units
- Preferred over deletion

Transform or cap

- Winsorise at percentiles
- Log or square-root transform
- Keeps every record
- Reduces the influence
- Changes interpretation

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Treatment 1: Removal

When it is right

- Value is physically impossible
- Confirmed data-entry error
- Duplicate or test record
- Cannot be corrected
- Few records affected

Cautions

- Never remove just for being extreme
- Removal can bias the sample
- Repeated removal chases the tail
- Keep a copy of removed rows
- Report the count and reason

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Treatment 2: Correction

- Always the best outcome — the record is kept and made accurate
- Look for the signature of a decimal point in the wrong place
- Check whether the value is right but recorded in the wrong unit
- Cross-reference other fields on the same record for the true value
- Return to the source system, the paper form or the data owner
- Log the original value alongside the correction for auditability

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Treatment 3: Capping and Winsorising

How it works

- Set a floor and a ceiling
- Pull extremes back to the limit
- Often the 1st and 99th percentile
- Or use the IQR fences
- Every record is retained

Trade-offs

- Reduces outlier influence
- Distorts the true distribution
- Creates a spike at each limit
- Understates real variability
- State the caps you applied

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Capping at the 99th percentile pulls the mean from $186 to $154 and keeps all 5,000 rows

Transaction column before and after capping

Measure	Raw	Capped at p99
Mean	$186	$154
Maximum	$4,210	$480
Standard deviation	$240	$96
Rows kept	**5,000**	**5,000**

What it does

Every value above the 99th percentile is set to it. The extreme rows stay in the dataset; only their magnitude is limited.

How to apply it

Fit the threshold on the training set only, record the value, and apply the same number to new data.

Illustrative data. n = 5,000 transactions; cap at the 99th percentile, $480.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Treatment 4: Transformation

- A log transform compresses a long right-hand tail dramatically
- Square-root and Box-Cox transforms are gentler alternatives
- Rank transforms discard magnitude entirely and keep only order
- No data is deleted and no value is altered arbitrarily
- Results are expressed on the transformed scale, so interpretation changes
- Log transforms cannot handle zero or negative values without a shift

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

A log transform pulls the skew from 4.8 to 0.3 without discarding a single transaction

Transaction column, raw against log scale

Measure	Raw	Log_{10}
Skewness	**4.8**	**0.3**
Mean	$186	2.14
Maximum ÷ median	28×	1.6×
Rows kept	5,000	5,000

Illustrative data. n = 5,000 transactions; base-10 logarithm.

What it does

Compressing the long right tail makes the distribution near-symmetric, so means, distances and linear models behave again.

What it costs

Results come back in log units and must be converted before reporting. Only valid for positive, right-skewed quantities.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Treatment 5: Keep and Flag

- Often the most honest option: retain the value and mark it
- Add an indicator column recording that the record was flagged
- Run the analysis with and without the flagged records
- If conclusions are unchanged, the outliers did not matter
- If conclusions flip, that finding belongs in your report
- Prefer robust methods over deleting inconvenient observations

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Choosing a Treatment

Confirmed error

- Correct it if possible
- Otherwise remove
- Or set to missing
- Then impute
- Document the change

Genuine extreme

- Keep the record
- Flag it clearly
- Use robust methods
- Consider transforming
- Report sensitivity

Uncertain

- Investigate the source
- Keep and flag by default
- Cap only if justified
- Run both analyses
- State the assumption

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Common Mistakes with Outliers

- Deleting every flagged point automatically without ever looking at it
- Repeating the detection rule until the data looks conveniently tidy
- Applying a Z-score rule to strongly skewed data
- Removing outliers computed on the whole dataset before splitting for a model
- Confusing rare-but-real observations with recording errors
- Failing to report how many points were altered, and on what basis

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Robust Alternatives to Removal

- Use the median and interquartile range instead of mean and standard deviation
- Fit robust regression methods such as Huber or RANSAC
- Use tree-based models, which split on order and ignore magnitude
- Apply rank-based tests such as Spearman rather than Pearson
- Use a robust scaler that centres on the median for machine learning
- Robust methods let you keep every record and still get stable results

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Outliers and Machine Learning

Sensitive models

- Linear and logistic regression
- K-means clustering
- KNN and distance methods
- Principal component analysis
- Neural networks with scaling

Robust models

- Decision trees
- Random forests
- Gradient boosting
- Rank-based methods
- Median-based scalers

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Avoiding Data Leakage

- Split into training and test sets before computing any thresholds
- Calculate fences, caps and scalers on the training data only
- Apply the stored training thresholds unchanged to the test set
- Never remove test-set rows because they look extreme
- Fold-level recomputation is required inside cross-validation
- Leakage produces optimistic scores that collapse in production

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Worked Example: Retail Transactions

What we found

- 86,400 transactions
- Values from £0 to £480,000
- IQR flags 2.1% of rows
- Three rows above £100,000
- Twelve rows of exactly £0

What we did

- Inspected the flagged records
- £0 rows were cancelled orders
- Large rows were bulk B2B orders
- Decimal error found in two rows
- Corrected the two errors

The outcome

- Cancelled orders excluded, flagged
- B2B orders kept and flagged
- Log transform for modelling
- Results stable either way
- All decisions documented

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Documenting Outlier Decisions

- Record the detection method and the exact threshold you used
- State how many points were flagged and what percentage that represents
- Classify each treated group: corrected, removed, capped or flagged
- Give the reasoning that justified each treatment choice
- Keep the original values so any decision can be reversed
- Report the sensitivity analysis alongside the headline result

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Recap and Check Your Understanding

Key takeaways

- Outliers are not always errors
- IQR beats Z-score on skewed data
- Multivariate outliers need models
- Investigate before you remove
- Robust methods often beat deletion

Questions to try

- Write down the IQR fence formulas
- Why is the Z-score self-defeating?
- Give a contextual outlier example
- When would you winsorise?
- What causes outlier data leakage?

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Resolving data inconsistencies

Standardisation, deduplication and validation

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIIT
TIRUPATI

What Is a Data Inconsistency?

- An inconsistency is a disagreement between values that should agree
- Within a record: the delivery date precedes the order date
- Within a column: the same country written five different ways
- Across tables: the customer's address differs between two systems
- Across time: a category was renamed halfway through the year
- Inconsistencies block grouping, joining and reliable aggregation

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Types of Inconsistency

Representation

- Different formats
- Different spellings
- Different letter case
- Abbreviations mixed in
- Whitespace differences

Semantic

- Different units
- Different currencies
- Different definitions
- Different time zones
- Different reference periods

Logical

- Contradictory fields
- Impossible date orders
- Totals that do not add up
- Broken foreign keys
- Overlapping ranges

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Why Inconsistencies Are So Common

- Free-text fields let every user invent their own convention
- Systems built by different teams encode the same concept differently
- Business rules and category definitions change over time
- Mergers and acquisitions bring incompatible legacy data together
- International data brings different date, decimal and address formats
- Nobody owns the definition, so every team makes its own assumption

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

The Standardisation Process

- Step 1: list every distinct value in the column with its frequency
- Step 2: decide the single canonical form for each real concept
- Step 3: build an explicit mapping from every variant to the canonical form
- Step 4: apply the mapping and re-list the distinct values to check
- Step 5: store the mapping so the same rules apply next time
- Step 6: push the standard upstream so the problem stops recurring

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Standardising Text Fields

Basic cleanup

- Trim leading and trailing spaces
- Collapse repeated inner spaces
- Apply a consistent letter case
- Remove invisible characters
- Normalise accented characters

Then map values

- Count distinct values first
- Expand or fix abbreviations
- Map synonyms to one label
- Use a controlled vocabulary
- Re-check the distinct list

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Case, spacing and punctuation alone split one company into five customer records

Company name field, 561 records

As entered	Records	After standardising
"ACME Ltd"	412	acme ltd
" Acme Ltd."	88	acme ltd
"acme limited"	35	acme ltd
"ACME LTD"	19	acme ltd
"Acme-Ltd"	7	acme ltd

What's wrong

Nobody typed a different company. Five strings differ only in case, spaces, a full stop and a hyphen, and each becomes its own account.

The standard recipe

Trim, collapse inner spaces, fold case, strip punctuation, then map legal suffixes to one form.

Illustrative data. Quotation marks show the stored string.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Standardising Dates and Times

- Parse every date into a true date type; never leave dates as text
- Beware the ambiguity of 03/04/2025 — British and American order differ
- Adopt the ISO 8601 standard: YYYY-MM-DD for every stored date
- Store timestamps in UTC and record the original time zone separately
- Watch for Excel serial numbers and Unix epoch values in the same column
- Validate the parsed result: check the minimum and maximum dates make sense

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

03/04/2026 is 3 April in London and 4 March in New York, and the file does not say which

One stored string, two readings

As stored	Read day-first	Read month-first
03/04/2026	**3 Apr 2026**	**4 Mar 2026**
2026-03-04	4 Mar 2026	4 Mar 2026
4 Mar 26	4 Mar 2026	4 Mar 2026
1772582400	4 Mar 2026	4 Mar 2026

What's wrong

A day-first and a month-first parser both succeed on the same string and disagree by a month. Nothing errors, so nobody notices.

How to fix it

Store ISO 8601 with a time zone, and keep the raw string beside the parsed value until the source confirms its convention.

Illustrative data. Epoch value is seconds since 1 January 1970, UTC.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Standardising Units and Currencies

The problem

- Kilograms mixed with pounds
- Metres mixed with feet
- Celsius mixed with Fahrenheit
- Currencies added together
- Unit only stated in the header

The fix

- Choose one unit per column
- Convert with explicit factors
- Put the unit in the column name
- Record the conversion date for rates
- Sanity-check the resulting range

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Summing a mixed-unit column reports 1,100 kg for a pallet that weighs 745 kg

Weight column with units in a second field

Item	Value	Unit	In kg
A	450	kg	450.0
B	180	lb	81.6
C	320	lb	145.1
D	150	lb	68.0
Total	**1,100**	**mixed**	**744.7**

Illustrative data. 1 lb = 0.4536 kg.

What's wrong

The numbers are correct and the unit column is correct. Only the sum is wrong, because it adds pounds to kilograms.

How to fix it

Convert to one base unit on load and keep the unit in the field name. For currency, store the rate and the date it applied.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Standardising Categorical Values

- List distinct values with counts — rare values are usually variants
- Define the authoritative list of permitted categories first
- Use published standards where they exist, such as ISO country codes
- Map every variant explicitly rather than relying on automatic guessing
- Decide how to handle values that match nothing in your list
- Group very rare categories into an explicit "Other" bucket

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Nine labels in the status column describe three real states

Status column, 3,700 records

Raw labels	Records	Mapped to
Active, ACTIVE, active	2,410	Active
A	180	Active
Cancelled, cancelled	760	Cancelled
Canc., CXL	95	Cancelled
Suspended, On hold	220	Suspended
(blank)	35	Unknown

Illustrative data. Raw labels grouped; blank treated as Unknown.

What's wrong

A dashboard grouped on this column shows nine statuses. Two of them, A and CXL, are internal shorthand nobody outside the team can read.

How to fix it

Keep the allowed values in a reference list, map every raw label to it, and reject unknown labels at ingest rather than in the report.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Standardising Names and Addresses

- Split full names into components only when the order is reliable
- Handle titles, middle initials, suffixes and hyphenated surnames explicitly
- Addresses need parsing into street, locality, region and postcode
- Validate postcodes against an official reference list where available
- Specialised address-matching libraries beat hand-written rules
- Never assume a Western name or address structure applies globally

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Cross-Field Consistency Rules

Date rules

- Start before end
- Birth before hire
- Order before delivery
- No future event dates
- Age matches birth date

Arithmetic rules

- Parts sum to the total
- Quantity × price = line total
- Percentages sum to 100
- Balances reconcile
- Discount below list price

Logical rules

- Status matches dates
- Country matches postcode
- Child fields need a parent
- Codes exist in reference list
- No mutually exclusive flags

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Cross-Source Reconciliation

- Decide which system is the authoritative source for each field
- Compare overlapping records and quantify how often they disagree
- Classify disagreements: timing lag, definition difference, or genuine error
- Set a written precedence rule for resolving each type of conflict
- Reconcile totals before merging, not after the analysis has run
- Record every unresolved discrepancy rather than quietly picking one

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Duplicate Detection: Exact Matches

- Start with fully identical rows across every column — the easy case
- Then check for duplicated primary key or business key values
- Normalise text before comparing: trim, lowercase and strip punctuation
- Decide which copy to keep: the newest, the most complete, or the first
- Count duplicates before deleting so the loss is documented
- Check whether repeated rows are genuine repeated events, not errors

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Duplicate Detection: Fuzzy Matching

How it works

- Compare strings by similarity
- Levenshtein counts edit distance
- Jaro-Winkler favours prefixes
- Phonetic codes match sound
- Score every candidate pair

In practice

- Set a similarity threshold
- Block on postcode to cut comparisons
- Review borderline pairs by hand
- Combine several fields
- Expect false matches

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

A similarity of 0.92 joins the pair exact matching misses; 0.71 is a false pair

Candidate pairs by similarity score

Record A	Record B	Score	Decision
Jon Smith, LS1 4DT	John Smith, LS1 4DT	0.94	**Merge**
Acme Ltd	Acme Limited	0.92	**Merge**
R Patel, M1 2AA	**R Patel, M1 3BB**	**0.78**	**Review**
A Khan	A Kaur	0.68	Keep apart

How it works

Each pair gets a score between 0 and 1. Above the upper threshold they merge, below the lower one they stay apart, and the middle band goes to a person.

How to set it

Tune the thresholds on labelled pairs, and block on postcode or email domain first so the comparison stays affordable.

Illustrative data. Normalised similarity on name plus postcode; merge above 0.90, review 0.75 to 0.90.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Record Linkage and Entity Resolution

- The task: decide which records across sources describe the same entity
- Deterministic linkage matches on agreed keys and exact rules
- Probabilistic linkage scores agreement across several fields at once
- Blocking reduces the comparison space to manageable candidate pairs
- A surviving "golden record" is built from the best fields of each copy
- Always keep the links so a merge can be traced or undone

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Merging Duplicate Records Safely

- Define survivorship rules field by field before merging anything
- Common rules: keep the most recent, the most complete, or the trusted source
- Preserve every original identifier so downstream links do not break
- Sum or recount transactional children rather than discarding them
- Keep an audit table mapping merged identifiers to the surviving record
- Review a sample of merges manually before running the whole batch

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Validation Rules and Constraints

Type rules

- Correct data type
- Correct format pattern
- Fixed length where required
- Numeric precision
- Valid date parsing

Range rules

- Minimum and maximum
- Non-negative quantities
- Percentages 0 to 100
- Plausible human ages
- Dates within a window

Set rules

- Value in reference list
- Unique key constraint
- Foreign key exists
- Required field not null
- Mutually exclusive flags

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Automating Validation Checks

- Write validation rules as code so they run on every new data load
- Fail loudly: a broken rule should stop the pipeline, not warn quietly
- Separate blocking failures from warnings that need human review
- Version the rules alongside the data so history stays interpretable
- Report a pass rate per rule to track quality over time
- Add a new rule every time a fresh class of error is discovered

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Tidy Data: A Consistent Structure

- Each variable forms a column, and only one variable per column
- Each observation forms a row, at a single consistent level of detail
- Each type of observational unit forms its own table
- Split cells that pack several values into one field
- Pivot columns that are really values, such as Jan, Feb and Mar
- Tidy structure makes every later cleaning step dramatically easier

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Schemas and Data Contracts

What they define

- Column names and order
- Data type of each field
- Allowed values and ranges
- Which fields are required
- Expected update frequency

Why they help

- Errors caught at the boundary
- Producers know their obligations
- Consumers can rely on structure
- Changes become deliberate
- Fewer surprises downstream

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Master Data and Reference Lists

- Master data is the agreed single version of core business entities
- Reference lists hold the permitted values for categorical fields
- One owner per list, with a documented process for adding values
- Store lists centrally so every system resolves against the same source
- Version the lists: category definitions genuinely change over time
- Most inconsistency disappears once everyone shares one reference list

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Worked Example: Merging Two Customer Systems

What we found

- CRM: 48,000 records
- Billing: 51,000 records
- No shared identifier
- Country spelled 23 ways
- Dates in two formats

What we did

- Standardised names and case
- Parsed dates to ISO format
- Mapped country to ISO codes
- Blocked on postcode
- Fuzzy matched name and address

The outcome

- 41,200 confident matches
- 2,600 pairs reviewed by hand
- Golden records built
- Audit table of all links
- Reference list published

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Preventing Inconsistency at Source

- Replace free-text fields with dropdown lists wherever possible
- Validate format and range at the point of entry, not weeks later
- Publish clear definitions for every field so users share one meaning
- Give each dataset a named owner accountable for its quality
- Automate ingestion instead of relying on manual re-typing
- Cleaning is a symptom; prevention is the cure

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Module D Recap and Check Your Understanding

Key takeaways

- Inconsistency has three types
- Standardise to one canonical form
- Fuzzy matching finds near-duplicates
- Encode rules as automated checks
- Fix the source, not just the copy

Questions to try

- Why is 03/04/2025 ambiguous?
- What does blocking achieve?
- Give a cross-field rule example
- When is fuzzy matching risky?
- What belongs in a data contract?

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Reshaping: Wide and Long Formats

Wide format

- One row per entity
- A column per time period
- Readable for humans
- Hard to filter or group
- Breaks when periods are added

Long format

- One row per observation
- Period becomes a value
- Preferred for analysis
- Easy to filter and aggregate
- Scales to new periods

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Two month columns become one month column and two rows per customer

Wide form: one row per customer

Customer	Jan	Feb
C-1	120	135
C-2	90	104

Long form: one row per observation

Customer	Month	Spend
C-1	Jan	120
C-1	Feb	135
C-2	Jan	90
C-2	Feb	104

Illustrative data. Monthly spend in $.

What changes

The numbers are identical. In long form the month is data rather than a column name, so a new month adds rows instead of changing the schema.

Which to use

Long for analysis, charting and storage; wide for reading on a page. Reshape at the point of use, not in the warehouse.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Splitting and Combining Columns

- Split fields that pack several facts into one, such as "Red; Large; 2kg"
- Use a reliable delimiter and check every row has the expected part count
- Rows with too many or too few parts need manual review, not silent truncation
- Combine columns only when the pieces genuinely belong together
- Keep the original column until the split has been fully validated
- Name the resulting columns explicitly so their meaning is unambiguous

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

One address field splits into four, and two name fields combine into one key

Split and combine operations

Operation	Before	After
Split	12 Oak St, Leeds, LS1 4DT	**street / city / postcode**
Combine	"Priya" + "Sharma"	**Priya Sharma**
Split	2026-03-04T09:15	**date + time**
Combine	"LS1" + "4DT"	**LS1 4DT**

Illustrative data.

Why it matters

You cannot group by city while the city is buried inside a single address string, and you cannot join on a name held in two columns.

How to do it safely

Split on a reliable delimiter, keep the original column until the result is checked, and count the pieces per row to catch the odd ones.

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Consistency in Derived Fields

- Derived fields such as age, tenure or margin must be recomputed, not trusted
- A stored age becomes wrong the moment the reference date changes
- Check that stored totals still equal the sum of their components
- Recalculate derived values from the cleaned source columns
- Document the exact formula and the reference date used
- Delete stale derived columns rather than letting them contradict the data

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Consistency Across Time

What changes

- Category definitions revised
- Products renamed or merged
- Regions redrawn
- Survey questions reworded
- Measurement method changed

How to handle it

- Record when each change happened
- Map old codes to new ones
- Keep both versions where needed
- Flag breaks in any trend line
- Never compare across a break silently

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Tools of the Trade

Profiling

- Summary statistics
- Distinct value counts
- Missingness matrices
- Automated profile reports
- Distribution plots

Cleaning

- Spreadsheet find and replace
- Regular expressions
- Scripted transformations
- Mapping and lookup tables
- Rule engines

Matching

- Edit distance measures
- Phonetic algorithms
- Blocking strategies
- Record linkage libraries
- Manual review queues

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

An Inconsistency Checklist

- Have you listed the distinct values of every categorical column?
- Are all dates parsed to a single, unambiguous format?
- Does every numeric column have one stated unit?
- Have text fields been trimmed and case-standardised?
- Have duplicate entities been identified using a business key?
- Are cross-field rules written down and automatically tested?

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Putting it all together

An end-to-end cleaning case study and best practices

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT
TIRUPATI

The End-to-End Cleaning Workflow

1. Understand

- Read the data dictionary
- Clarify the analysis question
- Profile every column
- Record first impressions
- Agree quality thresholds

2. Repair

- Fix structure and types
- Standardise representations
- Handle missing values
- Investigate outliers
- Resolve inconsistencies

3. Confirm

- Re-profile the cleaned data
- Run validation rules
- Compare before and after
- Document every decision
- Hand over with a summary

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Case Study: The Dataset and the Brief

- A retailer's transaction table: 240,000 rows across 18 months
- Columns include order date, store, product, quantity, unit price and total
- The brief: report monthly revenue per store and identify top products
- The data comes from three tills that were installed in different years
- No data dictionary exists, and two staff have entered records by hand
- We will work through profiling, missing values, outliers and inconsistencies

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Case Study: Profiling Findings

Structure

- Order date stored as text
- Quantity stored as text
- Store name free-text
- 18 columns, 3 unused
- 2,100 exact duplicate rows

Content

- Unit price 6% missing
- Quantity has negative values
- Totals disagree in 4% of rows
- Store spelled 14 ways
- Dates in two formats

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Case Study: Handling the Missing Values

- Unit price was missing in 14,400 rows — 6% of the dataset
- Missingness was concentrated in one till, so it was not MCAR
- Deleting those rows would have removed one store almost entirely
- Where total and quantity were present, price was derived arithmetically
- The remaining 1,900 rows used the median price for that product
- An indicator column records which prices were imputed and how

भारतीय प्रौद्योगिकी संस्थान तिरुपति
IIT TIRUPATI

Case Study: Investigating the Outliers

- Quantity ranged from -40 to 9,999, which is clearly implausible
- Negative quantities turned out to be returns — genuine, and kept but flagged
- The 9,999 values were a till default when the operator skipped the field
- Those 340 rows were set to missing and imputed from the product median
- Three legitimate bulk orders above the IQR fence were verified and retained
- Revenue was recomputed after treatment and fell by 2.1%

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Case Study: Resolving the Inconsistencies

- Store names were mapped from 14 spellings to 3 canonical values
- Dates in two formats were parsed separately, then unified to ISO 8601
- Quantity and price were converted from text to proper numeric types
- The 4% of rows where total disagreed were recalculated from quantity × price
- 2,100 exact duplicate rows were removed after confirming they were re-scans
- Three unused columns were dropped and the change was documented

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Case Study: Results and Impact

Before cleaning

- Revenue overstated by 2.1%
- One store split across 14 labels
- Monthly trend uninterpretable
- Top-product list wrong
- No confidence in the figures

After cleaning

- Revenue reconciles to the ledger
- Three clean store totals
- Clear 18-month trend
- Top products verified
- Every decision documented

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Reproducible Cleaning Pipelines

- Never clean by hand-editing the raw file — write the steps as code
- Keep the raw data immutable; every output is regenerated from it
- Order the steps deliberately: structure, then types, then values
- Version-control the cleaning script alongside the analysis
- Re-running the pipeline on new data must give consistent results
- A reproducible pipeline turns cleaning from an art into a process

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Documenting Your Cleaning

- Record what was wrong, what you did, and why you chose that treatment
- Report row counts before and after every step that removes data
- Keep indicator columns showing which values were imputed or capped
- State the assumptions a reader would need to challenge your results
- List the issues you found but chose not to fix, with your reasoning
- Undocumented cleaning is indistinguishable from data manipulation

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Best Practices Checklist

Before

- Keep the raw data intact
- Profile before deciding
- Define quality thresholds
- Understand the analysis need
- Agree who owns the data

During

- Work in small, tested steps
- Investigate before deleting
- Prefer flags over removal
- Re-check after each change
- Write it as code

After

- Re-profile the result
- Run validation rules
- Document every decision
- Report what you could not fix
- Push fixes upstream

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Common Pitfalls to Avoid

Do not

- Delete rows without counting
- Impute before understanding why
- Remove every statistical outlier
- Trust stored derived fields
- Clean silently and undocumented

Instead

- Quantify what you remove
- Diagnose the missingness mechanism
- Investigate each extreme value
- Recompute from source columns
- Publish a cleaning log

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI

Course Summary and Further Reading

- Data quality has measurable dimensions — profile before you clean
- Missing values need a diagnosis of mechanism before a treatment
- Outliers are candidates for investigation, not automatic deletion
- Inconsistency is solved by standardisation, matching and validation rules
- Recommended reading: Wickham's "Tidy Data"; Little and Rubin on missing data

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI