DegreeLive

Bachelor of Commerce (B2001) | Level 1 • Part B Major | Monash University

Introduction to data analysis

Course Code: ETC1010 • Module 1 Notes

Module 1: Reproducible Data Workflows, Tidy Data & Exploratory Data Visualisation

1.1 Reproducible Analytical Workflows and Project Architecture

Modern empirical commerce and policy analysis relies on auditable, repeatable computational pipelines rather than ad-hoc spreadsheet edits. When an analyst manually sorts columns, deletes anomalous rows, or types formulas into an unrecorded spreadsheet cell, the provenance of every downstream figure is lost. In Australian corporate and regulatory environments—such as capital adequacy reporting to the Australian Prudential Regulation Authority (APRA), pricing reviews for the Australian Competition and Consumer Commission (ACCC), or treasury forecasting—analysts must demonstrate exactly how raw source records were transformed into summary tables and visualisations.

A **reproducible workflow** guarantees that an independent reviewer, given the same raw inputs and analysis codebase, can regenerate identical tables, statistical summaries, and figures without manual intervention. Reproducibility rests on four operational pillars:

- **Immutability of Raw Data:** Source files downloaded from external providers—such as Australian Bureau of Statistics (ABS) tables, ASX market feeds, or internal point-of-sale exports—are stored in a dedicated read-only directory (for example, `data-raw/`) and never overwritten by cleaning scripts.
- **Scripted Transformations:** Every filtering, type conversion, recoding, and aggregation step is encoded in version-controlled source files (such as R or Python scripts) rather than executed through point-and-click menus.
- **Project-Oriented Directory Structure:** Analysis projects are anchored to a self-contained root directory using relative paths (for instance, `here::here("data", "retail.csv")`) rather than machine-specific absolute paths (such as `C:\\Users\\Analyst\\Desktop\\`) that fail when shared across collaborators.
- **Literate Programming and Environment Pinning:** Tools such as Quarto and R Markdown weave narrative prose, executable code blocks, and rendered outputs (tables and charts) into a single source document. Package versions and random number generator seeds (`set.seed()`) are recorded so stochastic routines such as resampling or clustering yield identical outputs across runs.

Workflow Component	Non-Reproducible Practice	Reproducible Standard	Audit Risk Mitigated
Raw Data Storage	Overwriting original CSV/Excel files after manual edits	Read-only `data-raw/` folder with dated metadata	Silent corruption or loss of original records
File Referencing	Hardcoded local desktop paths	Project root markers and relative paths	Broken scripts when executed on another machine

DegreeLive

Workflow Component	Non-Reproducible Practice	Reproducible Standard	Audit Risk Mitigated
Data Cleaning	Manual deletion of rows or find-and-replace in spreadsheets	Versioned scripts with explicit assertions	Untraceable exclusions and irreproducible metrics
Reporting	Copy-pasting numbers and charts into slide decks	Dynamic inline computation in Quarto / R Markdown	Transcription errors and stale figures after data updates
Software Environment	Implicit reliance on globally installed library versions	Lockfiles (e.g., `renv.lock`) and session manifests	Breaking changes across package updates

In team-based analytics, version control systems such as Git complement literate programming by maintaining a chronological commit log. Each commit records who altered a script, when the alteration occurred, and which exact lines changed. Combined with automated checks that verify row counts, missing-value bounds, and key uniqueness before downstream modelling begins, reproducible project architecture prevents cascading analytical failures.

Literate Programming Anatomy: Quarto and R Markdown

A literate programming source file (such as `report.qmd` or `analysis.Rmd`) integrates three distinct syntactic components within a single plain-text file:

1. **YAML Metadata Header:** Enclosed between triple hyphens (`---`) at the top of the document, declaring output formats (HTML, PDF, Word), execution defaults (such as `echo: false` for executive audiences vs. `echo: true` for technical audits), figure dimensions, and bibliography references.
2. **Markdown Narrative Prose:** Plain-text headings, paragraphs, ordered lists, tables, and inline dynamic expressions (for example, `` \`r nrow(df)\` ``) that pull computed statistics directly into sentences so reported sample sizes never drift out of synchronisation with underlying tables.
3. **Executable Code Chunks:** Fenced blocks executed sequentially in a clean, isolated session upon rendering. Because rendering spins up a fresh interpreter environment from top to bottom, hidden dependencies on variables created interactively in a console session are immediately exposed and eliminated.

To prevent unnecessary recomputation during iterative report drafting while preserving strict reproducibility, analysts separate computationally expensive ingestion and model-fitting scripts (writing validated intermediate binary formats such as Parquet or RDS into `data-clean/`) from lightweight presentation scripts. When source data change, a pipeline orchestration tool or master build script re-executes the entire dependency graph in topological order.

DegreeLive

1.2 Data Types, Measurement Scales, and Parsing "Data from the Wild"

Before visualising or modelling a dataset, an analyst must distinguish between the *statistical measurement scale* of a variable and its *computational storage type*. Confusing the two leads to nonsensical operations, such as computing the arithmetic mean of Australian postcodes or treating Likert survey ranks as continuous interval measurements.

Under Stanley Smith Stevens' classical taxonomy, empirical variables occupy four measurement scales:

- **Nominal Scale:** Unordered categorical labels used strictly for classification and identity. Examples include Australian states and territories (`NSW`, `VIC`, `QLD`, `WA`, `SA`, `TAS`, `ACT`, `NT`), ANZLIC industry classifications, and payment methods. Only equality comparisons and mode/frequency counts are meaningful.
- **Ordinal Scale:** Categories possessing an intrinsic rank ordering, though distances between adjacent ranks are not necessarily equal. Examples include credit ratings (`AAA`, `AA`, `A`, `BBB`) and customer satisfaction tiers (`Low`, `Medium`, `High`). Medians and quantiles are valid, whereas arithmetic means rely on unverified equal-spacing assumptions.
- **Interval Scale:** Numerical measurements with uniform unit intervals but an arbitrary zero point. Calendar years and Celsius temperatures are interval variables; a difference between 2020 and 2025 equals 5 years, yet ratios such as "twice as hot" in Celsius are meaningless.
- **Ratio Scale:** Numerical measurements with both equal intervals and a true, non-arbitrary zero indicating complete absence of the quantity. Weekly household income in Australian dollars (AUD), retail turnover, age in years, and electricity consumption in kilowatt-hours are ratio variables supporting valid ratios and multiplicative comparisons.

Computationally, modern analytical environments map these scales onto concrete storage classes:

Computational Type	Underlying Representation	Typical Measurement Scale	Australian Dataset Example
`logical` (Boolean)	Binary truth value (`TRUE`, `FALSE`, `NA`)	Nominal (binary indicator)	Whether an ASX firm paid a franked dividend
`integer`	Exact 32-bit whole numbers	Ratio (counts) or Ordinal	Number of dwellings approved in a Victorian LGA
`double` (Numeric)	64-bit IEEE floating-point numbers	Interval or Ratio	RBA cash rate target (4.35%) or CPI index
`character` (String)	Variable-length UTF-8 text sequences	Nominal or Unstructured	Suburb names, ABN strings, product descriptions
`factor`	Integer codes paired with a level lookup table	Nominal or Ordinal	ABS labour force status (Employed, Unemployed, NILF)
`Date` / `POSIXct`	Days since 1970-01-01 / seconds since UTC epoch	Interval	ABS quarterly reference period end dates

DegreeLive

When ingesting "data from the wild"—spreadsheets published by government agencies, vendor CSV exports, or sensor logs—raw files rarely arrive with clean column headers or uniform types. Common ingestion hazards include:

1. **Metadata Preamble and Footer Rows:** Public statistical workbooks (such as ABS Time Series Spreadsheets) frequently place titles, series IDs, units, and footnotes in rows 1 through 9 above the actual data table, plus explanatory notes at the bottom. Importing without specifying `skip` and maximum row bounds coerces numeric columns into character strings because the column contains text headers.
2. **Sentinel Missingness Codes:** Legacy systems encode missing or suppressed values using strings like `".."`, `"np"` (not published for confidentiality), `"-"`, or numeric sentinels such as `-999` or `9999`. Unless declared explicitly via `na = c("", "NA", "np", "..")` during parsing, numeric sentinels distort means and regressions while text sentinels trigger character coercion.
3. **Formatted Numbers and Locale Conflicts:** Currency symbols (`$`, `A$`), thousands separators (`1,450,200`), and percentage signs (`4.35%`) cause parsers to read quantities as text. Similarly, European or Australian day-first dates (`12/03/2026` meaning 12 March) are misparsed as 3 December under US month-first defaults.
4. **Character Encoding Mismatches:** Text saved in Windows-1252 or ISO-8859-1 but read as UTF-8 corrupts non-ASCII characters (em-dashes, curly quotes, or accented characters in customer records), requiring explicit locale specification during import.

1.3 Tidy Data Principles and Structural Diagnostics

Data analysts spend a disproportionate share of project time restructuring tables so that visualisation libraries, grouping verbs, and statistical modelling engines can operate consistently. Hadley Wickham formalised the **tidy data** standard—an adaptation of E. F. Codd's third normal form for statistical tables—around three structural invariants:

1. Each variable forms its own column.
2. Each observation forms its own row.
3. Each type of observational unit forms a separate table.

A dataset is called *messy* whenever its physical layout violates one or more of these three rules. Recognising the five canonical messy data structures enables systematic diagnosis before writing transformation pipelines:

- **Column Headers Are Values, Not Variable Names:** Financial and demographic spreadsheets routinely place calendar years, quarters, or Australian states across column headers (for example, columns named `2022_Q1`, `2022_Q2`, `2022_Q3`). Here, time period is a single variable whose values have been spread horizontally.
- **Multiple Variables Stored in One Column:** Demographic tables often concatenate two attributes into a single column header or cell, such as `NSW_Female` or `VIC_Male`, combining state and sex in one string.
- **Variables Stored in Both Rows and Columns:** Weather and macroeconomic tables sometimes include an indicator column (`Metric` with values `"Tmax"` and `"Tmin"`) alongside day-of-month

DegreeLive

columns (`Day1`, `Day2`, ..., `Day31`). Resolving this requires pivoting days longer and indicators wider.

- **Multiple Observational Units in One Table:** Transaction logs frequently repeat customer address and demographic attributes on every line item of an invoice, mixing invoice-level, product-level, and customer-level attributes in a single flat table and risking update anomalies.
- **A Single Observational Unit Split Across Multiple Tables:** Annual snapshots stored in separate files (`sales_2023.csv`, `sales_2024.csv`, `sales_2025.csv`) require row-binding with an explicit year identifier column added during ingestion.

Worked Example: Tidying Wide ABS Quarterly Retail Turnover Data

Suppose an analyst extracts quarterly retail turnover figures (in AUD millions) across three Australian states from a published spreadsheet in wide presentation format:

State	2025_Q1	2025_Q2	2025_Q3
VIC	27,410	27,890	28,340
NSW	33,120	33,450	34,010
QLD	21,600	22,050	22,490

Step 1 (Diagnose Violations): The table contains three underlying variables—`State`, `Quarter`, and `Turnover_AUD_M`—yet displays four columns because quarter values (`2025_Q1`, `2025_Q2`, `2025_Q3`) occupy column headers. Furthermore, comma-formatted strings prevent immediate arithmetic aggregation.

Step 2 (Parse and Pivot Longer): Stripping commas to parse turnover as numeric doubles and pivoting the three quarter columns into two columns (`Quarter` and `Turnover_AUD_M`) transforms the 3×4 matrix into a 9×3 tidy table. If desired, `Quarter` can be split into separate integer `Year` and ordered factor `Qtr` columns.

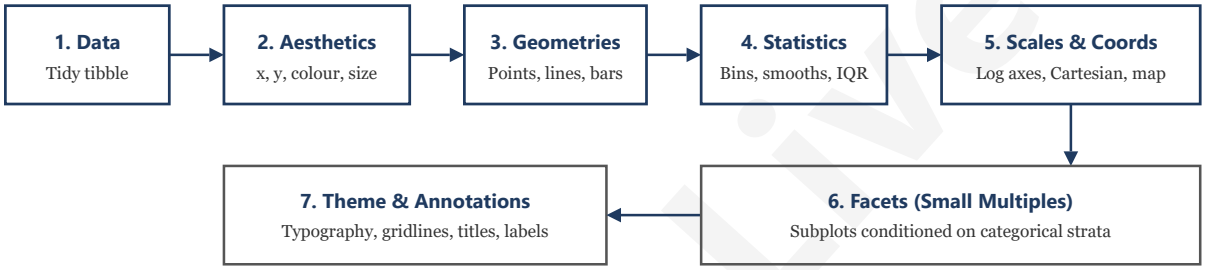
State	Year	Qtr	Turnover_AUD_M
VIC	2025	Q1	27410
VIC	2025	Q2	27890
VIC	2025	Q3	28340
NSW	2025	Q1	33120
NSW	2025	Q2	33450
NSW	2025	Q3	34010
QLD	2025	Q1	21600
QLD	2025	Q2	22050
QLD	2025	Q3	22490

DegreeLive

Step 3 (Verify Tidy Invariants): Every column now represents exactly one attribute, every row represents a unique State–Quarter observational unit ($3 \text{ states} \times 3 \text{ quarters} = 9 \text{ rows}$), and downstream grouping by `State` or plotting `Turnover_AUD_M` against time works without manual restructuring.

1.4 The Grammar of Graphics and Visual Mapping

Traditional charting software treats graphs as a fixed catalogue of disconnected chart types ("bar chart", "pie chart", "scatter plot"). By contrast, Leland Wilkinson's **Grammar of Graphics**—operationalised in R via `ggplot2`—decomposes any statistical graphic into orthogonal, composable layers. Under this grammar, a graphic is a mathematical mapping from tidy data attributes to aesthetic properties of geometric objects rendered within a coordinate system.



Each layer performs a distinct, non-overlapping role:

- **Data and Aesthetic Mapping (`aes()`):** Binds variables in a tidy data frame to visual channels such as horizontal position (`x`), vertical position (`y`), stroke colour (`colour`), interior shading (`fill`), marker geometry (`shape`), transparency (`alpha`), and line width (`linewidth`). Crucially, mapping an aesthetic inside `aes()` links it to data variation and generates a legend, whereas setting a constant outside `aes()` (such as `colour = "navy"`) fixes the visual appearance uniformly across all observations.
- **Geometric Objects (`geom_*`):** Determines the visual marks drawn on the canvas—points for bivariate scatter plots, lines for ordered time series, bars for discrete counts or aggregates, boxplots for quantile summaries, and density curves for smoothed distributions.
- **Statistical Transformations (`stat_*`):** Computes derived quantities prior to rendering. For example, a histogram computes bin counts (stat_bin), a boxplot extracts the median and interquartile hinges (stat_boxplot), and a regression overlay fits conditional expectations via ordinary least squares or LOESS (stat_smooth).
- **Scales:** Translates data-space domains into perceptual ranges (pixels, hex colour codes, point radii) and constructs axes and legends. Logarithmic scales ($\log_{10}$) are essential for right-skewed commercial quantities such as firm market capitalisation on the ASX or suburb median house prices, converting multiplicative ratios into equal visual intervals.
- **Coordinate Systems:** Combines two position scales into a 2D plane. Cartesian coordinates preserve orthogonal distances; flipped coordinates improve legibility for long categorical labels; polar coordinates map Cartesian bars onto angles and radii (producing pie or Coxcomb charts, often at the cost of perceptual accuracy).

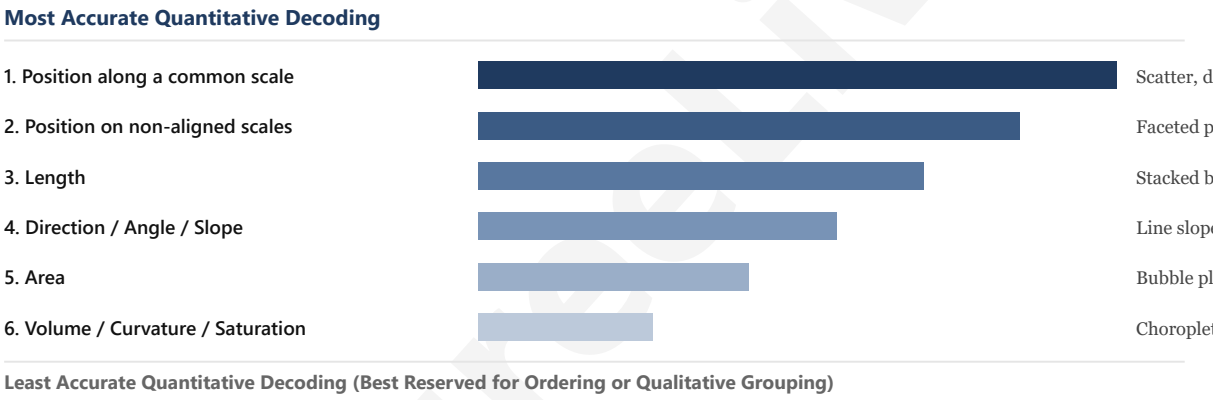
DegreeLive

- **Faceting:** Partitions the dataset by one or two discrete conditioning variables into aligned panels ("small multiples"). Faceting avoids overplotting when comparing five or more subgroups while maintaining shared axis scales for valid cross-panel comparisons.
- **Themes:** Controls non-data ink—background shading, gridline density, font hierarchy, and legend placement—without altering the underlying statistical mapping.

1.5 Perceptual Principles, Preattentive Processing & Chart Selection

Effective exploratory and explanatory visualisation aligns visual encodings with the neurophysiology of human vision. Visual perception operates in two stages: rapid, unconscious **preattentive processing** (occurring in under 200 milliseconds across the visual field for attributes such as spatial position, hue contrast, and orientation) followed by slower, serial **attentive comparison** (reading text labels, decoding multi-attribute legends, or comparing areas).

Building on empirical psychophysics, William Cleveland and Robert McGill established an accuracy hierarchy ranking the visual channels through which readers decode quantitative magnitudes:



Several practical design rules follow directly from perceptual psychology and Edward Tufte's principle of maximising the data-ink ratio:

- **Zero Baselines for Length Encodings:** Because bar charts encode magnitude through bar length from the baseline, truncating the y-axis at a non-zero floor exaggerates relative differences. By contrast, line plots and scatter plots encode values via position rather than bar length, so zooming the vertical axis around the observed range is often appropriate to reveal subtle variation.
- **Banking to 45 Degrees:** Human judgement of slope differences between line segments is sharpest when segments orient near 45°. Choosing an aspect ratio that stretches a steep time series horizontally or compresses a flat series vertically prevents visual misinterpretation of growth rates.
- **Principled Colour Selection:** Colour palettes must match the underlying variable structure. *Qualitative palettes* (distinct hues, uniform luminance) suit nominal categories such as states; *sequential palettes* (monotonically varying luminance, such as Viridis or single-hue blues) suit ordered magnitudes; *diverging palettes* (two contrasting hues radiating from a neutral midpoint) suit deviations around a meaningful threshold such as zero budget balance or RBA

inflation targets. Palettes must remain distinguishable under red-green colour vision deficiency (affecting roughly 8% of males) and retain monotonic luminance in greyscale printing.

- **Avoiding Overplotting and Distortion:** In large scatter plots (such as thousands of Melbourne property sales), overlapping points obscure density structure. Remedies include alpha transparency, hexagonal binning (geom_hex), 2D kernel density contours, or faceting. Dual y-axes with arbitrary independent scales and 3D perspective bars introduce severe visual distortion and should be replaced by small multiples or indexed relative series.

Analytical Task	Primary Variable Types	Recommended Geometry	Perceptual Rationale
Distribution of 1 continuous variable	1 Ratio / Interval	Histogram, density plot, ECDF	Position along common x- and y-axes reveals modality and skew
Comparison across categories	1 Nominal + 1 Ratio	Sorted horizontal bar or dot plot	Position along common axis; avoids rotated text labels
Distribution across categories	1 Nominal + 1 Ratio	Boxplot + jittered points or ridgeline	Exposes quantiles, spread, and multimodal subgroups
Relationship between 2 continuous variables	2 Ratio / Interval	Scatter plot + smooth trend	Common-scale 2D position exposes non-linearity and outliers
Temporal trajectory	Date + 1 Ratio	Line chart (faceted by group)	Connected segments encode temporal order and rate of change
Part-to-whole composition	1 Nominal + 1 Proportion	Sorted bar chart or 100% stacked bar	Length beats pie angles; avoids area/curvature error

1.6 Exploratory Data Analysis (EDA) and Missing Data Diagnostics

Exploratory Data Analysis (EDA), pioneered by John Tukey, is an iterative cycle of formulating questions, computing robust numerical summaries, plotting distributions, and uncovering anomalies before committing to formal statistical models. A rigorous univariate and bivariate audit examines location, spread, shape, outliers, and missingness simultaneously.

For a numerical sample $x_1, x_2, \ldots, x_n$, standard location and dispersion statistics include the sample mean $\bar{x}$, sample variance s^2, sample standard deviation s, median Q_2, and interquartile range $\mathrm{IQR} = Q_3 - Q_1$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad s = \sqrt{s^2}$$

Where:

- n is the number of non-missing observations in the sample.
- x_i is the i-th observed numerical value.
- $\bar{x}$ is the sample arithmetic mean (sensitive to extreme values).
- s^2 and s are the sample variance and standard deviation using Bessel's correction $(n-1)$ in the denominator.

Because $\bar{x}$ and s are sensitive to extreme skewness or recording errors, Tukey's five-number summary $(\min, Q_1, Q_2, Q_3, \max)$ and IQR fences provide resistant diagnostics. Observations lying outside the lower or upper inner fences are flagged for inspection (never automatically deleted without domain verification):

$$\text{Lower Fence} = Q_1 - 1.5 \times \text{IQR}, \qquad \text{Upper Fence} = Q_3 + 1.5 \times \text{IQR}$$

Where:

- Q_1 is the 25th percentile (first quartile) and Q_3 is the 75th percentile (third quartile).
- $\text{IQR} = Q_3 - Q_1$ is the interquartile range spanning the middle 50% of ordered values.

Worked Numerical Example: Outlier & Spread Diagnostics on Melbourne Weekly Rents

Consider a small sample of $n = 9$ advertised weekly rents (in $A\$$) for two-bedroom apartments in inner Melbourne suburbs, ordered ascending:

Ordered sample: $520, 540, 560, 580, 600, 620, 640, 680, 1420$.

Step 1 (Sample Mean and Standard Deviation):

Sum of observations: $\sum x_i = 520 + 540 + 560 + 580 + 600 + 620 + 640 + 680 + 1420 = 6160$. Thus $\bar{x} = 6160/9 \approx 684.44$ AUD.

Computing squared deviations $(x_i - 684.44)^2$: ((-164.44)^2 + (-144.44)^2 + (-124.44)^2 + (-104.44)^2 + (-84.44)^2 + (-64.44)^2 + (-44.44)^2 + (-4.44)^2 + (735.56)^2 = 27041.98 + 20864.20 + 15486.42 + 10908.64 + 7130.86 + 4153.09 + 1975.31 + 19.75 + 541042.00 = 628622.22.

Dividing by $n - 1 = 8$: $s^2 = 628622.22/8 = 78577.78$, so $s = \sqrt{78577.78} \approx 280.32$ AUD.

Step 2 (Resistant Quartiles and Tukey Fences):

The median Q_2 is the 5th observation: 600 AUD. Using inclusive lower/upper halves excluding the median (Tukey hinges): lower half $\{520, 540, 560, 580\}$ gives $Q_1 = (540 + 560)/2 = 550$ AUD; upper half $\{620, 640, 680, 1420\}$ gives $Q_3 = (640 + 680)/2 = 660$ AUD. Therefore, $\text{IQR} = 660 - 550 = 110$ AUD.

Applying Tukey's inner fences:

$\text{Lower Fence} = 550 - 1.5(110) = 550 - 165 = 385$ AUD; $\text{Upper Fence} = 660 + 1.5(110) = 660 + 165 = 825$ AUD.

Step 3 (Interpretation): Notice how the single penthouse listing at $\$1,420$ pulls the arithmetic mean ($\$684.44$) above 8 of the 9 observations and inflates s to $\$280.32$, whereas the median ($\$600$) and IQR ($\$110$) accurately characterise the typical market segment while cleanly flagging $\$1,420 > \$825$ as an upper outlier for verification.

Missing Data Mechanisms and Diagnostic Workflows

Missing values (`NA`) are rarely uniform across real-world datasets. Under Donald Rubin's classification, missingness arises via three distinct statistical mechanisms:

- **Missing Completely at Random (MCAR):** The probability that a value is missing is completely independent of both observed and unobserved data (for example, a laboratory sensor randomly drops packets due to power glitches). Complete-case analysis remains unbiased, though statistical power decreases.
- **Missing at Random (MAR):** The probability of missingness depends systematically on other *observed* variables, but not on the missing value itself once observed covariates are conditioned on (for example, younger ABS survey respondents are less likely to answer superannuation balance questions, where age is fully recorded). Listwise deletion produces biased population estimates unless missingness is modelled or imputed conditionally on observed covariates.
- **Missing Not at Random (MNAR):** The probability of missingness depends directly on the unobserved value itself (for example, ultra-high-income earners declining to disclose taxable income, or distressed borrowers failing to report financial stress). Standard imputation using observed covariates cannot eliminate MNAR bias without external auxiliary information or sensitivity modelling.

To diagnose missingness systematically prior to modelling, analysts apply visual and tabular summaries (implemented in packages such as `naniar` and `visdat`):

1. **Dataset-Wide Typology and Missingness Matrix:** Plotting observations as rows and variables as columns highlights whether missingness is sporadic, concentrated in a specific subset of variables, or clustered in blocks across contiguous rows (often indicating a discontinued survey wave or merged table mismatch).
2. **Shadow Matrix Tracking:** Creating parallel binary indicator columns (e.g., `income_NA` taking values `"NA"` or `"!NA"`) enables bivariate boxplots and scatter plots comparing the distribution of observed variables (such as age or suburb median rent) across missing and non-missing strata—directly testing MCAR against MAR.
3. **UpSet Plots for Co-Missingness:** When multiple columns exhibit missing values, an UpSet intersection chart quantifies how frequently subsets of variables are missing simultaneously, guiding decisions on whether to drop sparse features, perform single or multiple imputation (such as predictive mean matching), or retain explicit `"Missing"` factor levels for categorical predictors.

Bivariate EDA, Correlation Limits, and Simpson's Paradox

When exploring pairs of numerical variables (x_i, y_i), analysts routinely compute Pearson's sample correlation coefficient r to quantify linear association strength and direction:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where:

- $r \in [-1, 1]$ is the dimensionless sample Pearson correlation coefficient.
- x_i, y_i are paired numerical observations for unit i.
- $\bar{x}, \bar{y}$ are the respective sample means.

As demonstrated by Francis Anscombe's Quartet and the Datasaurus Dozen, four or more radically different bivariate structures—a clean linear trend, a quadratic curve, a linear cluster with one high-leverage outlier, and an almost vertical cluster shifted by a single extreme point—can yield identical means, variances, and Pearson correlations ($r = 0.816$). Relying solely on summary correlation matrices without scatter visualisation obscures non-linearity, bimodal clusters, and influential data errors.

Equally critical in commercial and social data is **Simpson's Paradox**, wherein an aggregate trend observed across pooled observations reverses when data are stratified by a confounding categorical variable. Consider a hypothetical Australian retail bank reviewing small-business loan approval rates across two metropolitan regions:

Applicant Segment	Melbourne Approvals	Sydney Approvals	Subgroup Difference
Low-Risk Established Firms	85 / 100 (85.0%)	320 / 400 (80.0%)	Melbourne higher (+5.0 pp)
High-Risk Startups	180 / 400 (45.0%)	40 / 100 (40.0%)	Melbourne higher (+5.0 pp)
Pooled Aggregate Total	**265 / 500 (53.0%)**	**360 / 500 (72.0%)**	**Sydney higher (+19.0 pp)**

Although Melbourne approves a strictly higher proportion of applicants in *both* risk tiers ($85\% > 80\%$ and $45\% > 40\%$), Sydney exhibits a higher aggregate approval rate ($72.0\% > 53.0\%$) because 80% of Sydney's applicant pool consists of low-risk established firms whereas 80% of Melbourne's pool consists of high-risk startups. Visualising conditional facets or stratified tables protects decision-makers from drawing inverted causal conclusions from aggregated dashboards.

Imputation Mechanics and Variance Distortion

When missing data cannot be dropped without losing substantial sample size or introducing MAR bias, analysts must choose an imputation strategy carefully:

- **Unconditional Mean or Median Imputation:** Replacing missing x_i with $\bar{x}$ preserves the sample mean under MCAR but artificially shrinks the sample variance and attenuates covariance with other variables, distorting downstream regression slopes and underestimating standard errors.

- **Conditional Regression and Predictive Mean Matching (PMM):** Predicting missing values from observed covariates $\hat{x}_i = \hat{f}(\mathbf{z}_i)$ restores conditional structure under MAR, while adding stochastic residual draws or sampling from nearest observed donors (PMM) preserves realistic dispersion and bounds.
- **Multiple Imputation (MICE):** Generating $M \geq 5$ plausible completed datasets, fitting the target model on each, and combining parameter estimates via Rubin's Rules propagates imputation uncertainty into final confidence intervals rather than treating imputed guesses as observed facts.

DegreeLive