DegreeLive

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

Introduction to data analysis

Course Code: ETC1010 • Module 4 Notes

Module 4: Dimension Reduction, Clustering, Text/Spatial Data & Data Communication

4.1 Dimension Reduction and Principal Component Analysis (PCA)

Modern commercial and demographic datasets frequently track dozens or hundreds of correlated variables simultaneously—for instance, 40 financial ratios across ASX-listed corporations or 60 Australian Bureau of Statistics (ABS) Socio-Economic Indexes for Areas (SEIFA) indicators across Melbourne suburbs. High dimensionality creates severe analytical challenges: visualising more than three axes directly is impossible, redundant features inflate collinearity in downstream models, and pairwise Euclidean distances grow sparse (the *curse of dimensionality*).

Principal Component Analysis (PCA) is an unsupervised linear dimension-reduction method that transforms p correlated numerical variables $X_1, X_2, \dots, X_p$ into a new set of p mutually orthogonal (uncorrelated) composite variables called **principal components** ($Z_1, Z_2, \dots, Z_p$), ordered strictly by the proportion of total dataset variance they capture.

Mathematical Formulation and Standardisation

Before extracting principal components, analysts must decide whether to decompose the *covariance matrix* $\mathbf{\Sigma}$ or the *correlation matrix* $\mathbf{R}$ (equivalent to standardising each column to mean 0 and variance 1 via $z_{ij} = (x_{ij} - \bar{x}_j)/s_j$). Whenever features carry incompatible physical units—such as household income in $A\$$, unemployment rate in percent, and population density per km^2—unstandardised PCA is dominated artificially by whichever variable has the largest raw numeric variance. Standardising puts every variable on equal unit variance prior to rotation.

The first principal component Z_1 is the normalised linear combination of standardised predictors $\widetilde{\mathbf{x}}_i = (\widetilde{x}_{i1}, \dots, \widetilde{x}_{ip})^\top$ that achieves maximum sample variance:

$$Z_{i1} = w_{11}\widetilde{x}_{i1} + w_{21}\widetilde{x}_{i2} + \cdots + w_{p1}\widetilde{x}_{ip}, \qquad \text{subject to} \quad \sum_{j=1}^{p} w_{j1}^2 = 1$$

Where:

- Z_{i1} is the first principal component **score** for observation i.
- w_{j1} is the **loading** of standardised variable $\widetilde{x}_j$ on the first principal component.
- The unit-norm constraint $\|\mathbf{w}_1\|_2^2 = 1$ prevents variance from being inflated trivially by scaling loadings upward.

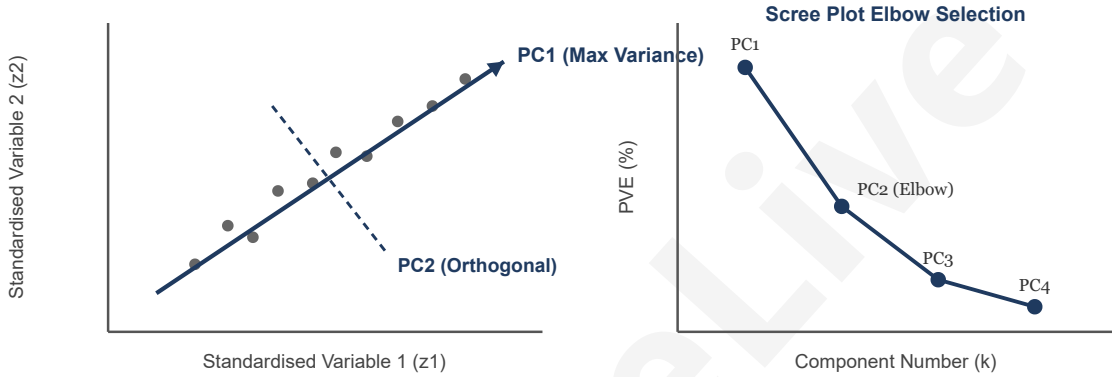
Each subsequent component Z_k maximises remaining variance subject to unit norm and strict orthogonality ($\mathbf{w}_k^\top \mathbf{w}_m = 0$ for all $m < k$). Algebraically, the loading vectors $\mathbf{w}_1, \dots, \mathbf{w}_p$ are the eigenvectors of the correlation matrix $\mathbf{R}$, and the variance of component Z_k equals its corresponding eigenvalue λ_k:

Degreelive

$$\mathbf{R}\mathbf{w}_k = \lambda_k \mathbf{w}_k, \qquad \text{PVE}_k = \frac{\lambda_k}{\sum_{m=1}^{p} \lambda_m} = \frac{\lambda_k}{p}$$

Where:

- $\mathbf{R}$ is the $p \times p$ sample correlation matrix of the p standardised variables.
- $\lambda_k \geq 0$ is the k-th largest eigenvalue, equal to $\text{Var}(Z_k)$.
- PVE_k is the Proportion of Variance Explained by the k-th principal component (in correlation PCA, $\sum_{m=1}^{p} \lambda_m = p$).



Worked Numerical Example: 2×2 Standardised PCA

Suppose we analyse two standardised socioeconomic variables across Australian suburbs: median household income (z_1) and tertiary degree share (z_2), with sample correlation $r = 0.80$:

$$\mathbf{R} = \begin{pmatrix} 1 & 0.80 \\ 0.80 & 1 \end{pmatrix}$$

Where:

- $\mathbf{R}$ is the 2×2 correlation matrix with unit diagonal variances and off-diagonal correlation 0.80.

Step 1 (Eigenvalues via Characteristic Equation): Solving $\det(\mathbf{R} - \lambda \mathbf{I}) = (1-\lambda)^2 - 0.80^2 = 0$ gives $1 - \lambda = \pm 0.80$, hence $\lambda_1 = 1 + 0.80 = 1.80$ and $\lambda_2 = 1 - 0.80 = 0.20$. Check trace conservation: $\lambda_1 + \lambda_2 = 1.80 + 0.20 = 2.00 = p$.

Step 2 (Proportion of Variance Explained):

- $\text{PVE}_1 = 1.80/2.00 = 0.90$ (90.0% of total standardised variance captured by PC1 alone).
- $\text{PVE}_2 = 0.20/2.00 = 0.10$ (10.0% captured by PC2).

Step 3 (Eigenvector Loadings and Component Interpretation): Substituting $\lambda_1 = 1.80$ into $(\mathbf{R} - 1.80\mathbf{I})\mathbf{w}_1 = \mathbf{0}$ yields $-0.80 w_{11} + 0.80 w_{21} = 0 \implies w_{11} = w_{21}$. Enforcing unit norm $w_{11}^2 + w_{21}^2 = 1$ gives $\mathbf{w}_1 = (1/\sqrt{2}, 1/\sqrt{2})^\top \approx (0.707, 0.707)^\top$ (an overall socio-economic

advantage index), while the orthogonal second eigenvector is $\mathbf{w}_2 = (0.707, -0.707)^\top$ (contrasting income against education). In higher dimensions, analysts inspect **scree plots** for an "elbow" where marginal PVE flattens, apply the Kaiser rule ($\lambda_k > 1$ in correlation PCA), and visualise observations alongside loading arrows in a **biplot**.

PCA in Predictive Pipelines: Strengths and Caveats

When PCA scores $Z_1, \dots, Z_M$ ($M \ll p$) serve as predictors in Principal Component Regression (PCR), orthogonal components eliminate predictor multicollinearity completely. However, because PCA is strictly unsupervised—maximising variance among $\mathbf{X}$ without consulting response y—there is no guarantee that the highest-variance principal components are the directions most strongly associated with y. If predictive signal resides along a low-variance minor component, discarding $Z_{M+1}, \dots, Z_p$ degrades predictive accuracy; hence validation error (CV RMSE) must confirm component retention choices rather than relying solely on cumulative PVE thresholds.

4.2 Unsupervised Learning: K-Means and Hierarchical Clustering

Whereas supervised regression predicts a labelled target y, **clustering** is an unsupervised task aimed at partitioning unlabelled observations $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ into internally cohesive, mutually distinct subgroups. In Australian retail banking and marketing, clustering segments stores by sales-mix profiles, classifies suburbs by commuting patterns, and detects anomalous operational regimes.

Distance Metrics and Feature Scaling

All distance-based clustering algorithms require a formal dissimilarity measure between observations $\mathbf{x}_i$ and $\mathbf{x}_{i'}$:

- **Euclidean Distance:** $d_{\text{Euc}}(\mathbf{x}_i, \mathbf{x}_{i'}) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}$, sensitive to scale differences and thus requiring z-score standardisation prior to clustering.
- **Manhattan (L1) Distance:** $d_{\text{Man}}(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} |x_{ij} - x_{i'j}|$, less sensitive to isolated extreme coordinates.
- **Cosine Dissimilarity and Gower Distance:** Cosine similarity compares vector orientation (standard for sparse high-dimensional text TF-IDF vectors), while Gower distance accommodates mixed numeric and categorical attributes.

K-Means Clustering and Lloyd's Algorithm

K-means clustering partitions n observations into K disjoint clusters $C_1, \dots, C_K$ by minimising total **Within-Cluster Sum of Squares (WCSS)**:

> $$\min_{C_1,\dots,C_K} \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \mu_{kj})^2$$

Where:

- K is the user-specified number of clusters.
- $\mu_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$ is the centroid (mean) of feature j across observations assigned to cluster C_k.

Finding the exact global minimum across all partitions is NP-hard, so **Lloyd's iterative algorithm** converges to a local optimum in alternating steps: (1) initialise K centroids randomly; (2) assign each observation to its nearest centroid by squared Euclidean distance; (3) recompute each cluster's centroid as the mean of its assigned points; and (4) repeat steps 2–3 until assignments stabilise. Because poor random starts trap Lloyd's algorithm in suboptimal local minima, analysts run multiple random initialisations (for example, `nstart = 25`) and retain the partition achieving minimum WCSS.

Selecting K relies on complementary diagnostics: plotting total WCSS against $K = 1, 2, \ldots$ to locate an **elbow** where marginal variance reduction slows, and maximising average **silhouette width** $s_i \in [-1, 1]$:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

Where:

- a_i is the average distance from observation i to all other observations in its own assigned cluster.
- b_i is the minimum average distance from observation i to all observations in any other cluster (its nearest neighbouring cluster).

Values near $+1$ indicate tight, well-separated cluster membership; values near 0 sit on cluster boundaries; negative values indicate likely misassignment.

Worked Numerical Example: Two Iterations of K-Means ($K = 2$)

Consider $n = 4$ Victorian retail outlets described by two standardised metrics (z_1: foot traffic, z_2: average basket size): Store A $(0, 0)$, Store B $(1, 1)$, Store C $(5, 4)$, and Store D $(6, 5)$.

Initialisation: Suppose initial centroids are placed arbitrarily at $\boldsymbol{\mu}_1^{(0)} = (0, 0)$ and $\boldsymbol{\mu}_2^{(0)} = (1, 1)$.

Iteration 1 (Assignment): Computing squared Euclidean distances $d^2(\mathbf{x}, \boldsymbol{\mu})$:

- Store A $(0, 0)$: $d^2(\text{A}, \boldsymbol{\mu}_1) = 0, d^2(\text{A}, \boldsymbol{\mu}_2) = 2 \implies C_1$.
- Store B $(1, 1)$: $d^2(\text{B}, \boldsymbol{\mu}_1) = 2, d^2(\text{B}, \boldsymbol{\mu}_2) = 0 \implies C_2$.
- Store C $(5, 4)$: $d^2(\text{C}, \boldsymbol{\mu}_1) = 25 + 16 = 41, d^2(\text{C}, \boldsymbol{\mu}_2) = 16 + 9 = 25 \implies C_2$.
- Store D $(6, 5)$: $d^2(\text{D}, \boldsymbol{\mu}_1) = 36 + 25 = 61, d^2(\text{D}, \boldsymbol{\mu}_2) = 25 + 16 = 41 \implies C_2$.

Iteration 1 (Centroid Update): $C_1 = \{\text{A}\}$ gives $\boldsymbol{\mu}_1^{(1)} = (0, 0)$; $C_2 = \{\text{B}, \text{C}, \text{D}\}$ gives $\boldsymbol{\mu}_2^{(1)} = (\frac{1+5+6}{3}, \frac{1+4+5}{3}) = (4.00, 3.33)$.

DegreeLive

Iteration 2 (Reassignment): Re-evaluating Store B $(1,1)$: $d^2(\mathrm{B}, \boldsymbol{\mu}_1^{(1)}) = 1^2 + 1^2 = 2$, whereas $d^2(\mathrm{B}, \boldsymbol{\mu}_2^{(1)}) = (1-4)^2 + (1-3.33)^2 = 9 + 5.44 = 14.44$. Thus Store B switches to C_1, leaving $C_1 = \{\mathrm{A}, \mathrm{B}\}$ and $C_2 = \{\mathrm{C}, \mathrm{D}\}$. Updating centroids yields $\boldsymbol{\mu}_1^{(2)} = (0.5, 0.5)$ and $\boldsymbol{\mu}_2^{(2)} = (5.5, 4.5)$. A third check confirms assignments remain unchanged, achieving convergence with $\mathrm{WCSS} = [0.5 + 0.5] + [0.5 + 0.5] = 2.0$.

Computing the silhouette score for Store A $(0,0)$: its within-cluster distance to Store B $(1,1)$ is $a_\mathrm{A} = \sqrt{1^2 + 1^2} = \sqrt{2} \approx 1.414$. Its distances to $C_2 = \{\mathrm{C}, \mathrm{D}\}$ are $d(\mathrm{A}, \mathrm{C}) = \sqrt{25 + 16} \approx 6.403$ and $d(\mathrm{A}, \mathrm{D}) = \sqrt{36 + 25} \approx 7.810$, giving $b_\mathrm{A} = (6.403 + 7.810)/2 = 7.107$. Hence $s_\mathrm{A} = (7.107 - 1.414)/7.107 = 0.801$, confirming strong cluster separation.

Agglomerative Hierarchical Clustering and Dendrograms

Unlike K-means, **agglomerative hierarchical clustering** does not require pre-specifying K. Starting with n singleton clusters, it iteratively merges the two closest clusters until a single root encompasses all observations, recording the entire nested merge history in a tree diagram called a **dendrogram**. Cluster-to-cluster dissimilarity is governed by the chosen **linkage criterion**:

- **Complete Linkage:** Maximum pairwise distance between points in two clusters; favours compact, roughly spherical clusters.
- **Single Linkage:** Minimum pairwise distance; detects elongated shapes but suffers from "chaining" noise bridges.
- **Average Linkage:** Mean distance across all inter-cluster pairs; compromises between single and complete linkage.
- **Ward's Minimum Variance (`ward.D2`):** Merges the pair producing the smallest increase in total within-cluster sum of squares, yielding balanced, variance-compact partitions analogous to K-means.



Cutting the dendrogram horizontally at a chosen height h determines the resulting cluster count K; large vertical gaps between consecutive merges indicate natural cluster structure in the data.

Clustering Method	Cluster Shape Assumption	Key Hyperparameters	Strengths & Commercial Limitations
K-Means (Lloyd's)	Convex, roughly equal-radius spheres	Number of clusters K; `nstart` restarts	Fast $O(nKT)$ scaling to large customer bases; sensitive to outliers and non-spherical clusters

DegreeLive

Clustering Method	Cluster Shape Assumption	Key Hyperparameters	Strengths & Commercial Limitations
Agglomerative Hierarchical	Depends on linkage (Ward favours spherical)	Distance metric, linkage rule, cut height h	Reveals full nested taxonomy via dendrogram; requires $O(n^2)$ pairwise distance matrix memory
Density-Based (DBSCAN)	Arbitrary connected high-density shapes	Neighbourhood radius ε; minimum points	Detects spatial noise/outliers natively without forcing all points into clusters; struggles with varying densities

4.3 Text Data Mining: Tokenisation, Stop Words, and TF-IDF

Substantial business intelligence resides in unstructured text—ASX continuous disclosure filings, Reserve Bank of Australia (RBA) monetary policy statements, product reviews, and parliamentary transcripts. Transforming raw text into tidy quantitative representations follows the **tidy text** pipeline:

1. **Tokenisation:** Splits raw document strings into discrete analytical units called *tokens*—typically individual words (unigrams), consecutive word pairs (bigrams) capturing phrases such as `"cash rate"` or `"not profitable"`, or sentences.
2. **Normalisation and Stop-Word Filtering:** Converts text to lowercase, strips punctuation and extraneous digits where appropriate, and removes high-frequency grammatical function words (*stop words* such as `"the"`, `"and"`, `"of"`) that carry minimal topical discrimination.
3. **Stemming vs. Lemmatisation:** *Stemming* heuristically truncates suffixes (`"investing"`, `"invested"` → `"invest"`), whereas *lemmatisation* maps inflected forms to valid dictionary roots using part-of-speech context.
4. **Document-Term Matrix (DTM) or Tidy Token Table:** Represents a corpus of N documents and V unique vocabulary terms either as a sparse $N \times V$ matrix of term counts or as a tidy long table with columns `(document_id, term, count)`.

Term Frequency–Inverse Document Frequency (TF-IDF)

Raw word counts favour common boilerplate terms that appear across almost every document in a domain (for example, `"board"` or `"million"` in ASX annual reports). **TF-IDF** reweights each term so that words occurring frequently within a specific document yet rarely across the broader corpus receive the highest salience:

$$\text{TF}(t,d) = \frac{n_{t,d}}{\sum_k n_{k,d}}, \qquad \text{IDF}(t) = \ln\left(\frac{N}{df_t}\right), \qquad \text{TF-IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t)$$

Where:

- $n_{t,d}$ is the raw occurrence count of term t in document d, and $\sum_k n_{k,d}$ is the total token length of document d.
- $\text{TF}(t,d)$ is the normalised Term Frequency (proportion of tokens in document d equal to t).

DegreeLive

- N is the total number of documents in the corpus.
- df_t is the document frequency (number of documents containing term t at least once).
- $\text{IDF}(t)$ is the Inverse Document Frequency; any term present in every document ($df_t = N$) receives $\ln(1) = 0$, eliminating corpus-wide boilerplate automatically without manual stop-word lists.

Worked Numerical Example: TF-IDF Across Four Australian Corporate Disclosures

Consider a mini-corpus of $N = 4$ short Australian corporate announcements, after stop-word removal, each containing 10 tokens total:

- **Doc 1 (Retail):** contains `"profit"` (2 times), `"retail"` (3 times), `"store"` (5 times).
- **Doc 2 (Banking):** contains `"profit"` (3 times), `"mortgage"` (4 times), `"rate"` (3 times).
- **Doc 3 (Mining):** contains `"profit"` (1 time), `"lithium"` (4 times), `"export"` (5 times).
- **Doc 4 (Energy):** contains `"profit"` (2 times), `"solar"` (5 times), `"grid"` (3 times).

Computing TF-IDF for `"profit"` vs. `"lithium"` in **Doc 3 (Mining)**:

- Term `"profit"` appears in all $df_{\text{profit}} = 4$ documents: $\text{TF}(\text{"profit"}, \text{Doc 3}) = 1/10 = 0.10$, $\text{IDF}(\text{"profit"}) = \ln(4/4) = \ln(1) = 0$, hence $\text{TF-IDF} = 0.10 \times 0 = 0.000$.
- Term `"lithium"` appears only in $df_{\text{lithium}} = 1$ document (Doc 3): $\text{TF}(\text{"lithium"}, \text{Doc 3}) = 4/10 = 0.40$, $\text{IDF}(\text{"lithium"}) = \ln(4/1) = \ln(4) \approx 1.386$, hence $\text{TF-IDF} = 0.40 \times 1.386 = 0.554$.

TF-IDF cleanly elevates `"lithium"` and `"export"` as signature descriptors of Doc 3 while suppressing ubiquitous corporate vocabulary like `"profit"`.

For tone analysis, lexicon-based **sentiment scoring** joins tidy unigrams against curated polarity dictionaries (such as Bing binary positive/negative lists, AFINN integer scores from -5 to $+5$, or Loughran-McDonald financial sentiment dictionaries). Domain verification is essential: general-purpose lexicons misclassify neutral accounting nouns such as `"liability"`, `"tax"`, or `"risk"` as negative sentiment, and unigram matching fails when preceded by negation modifiers (`"not profitable"`) unless bigrams or dependency negators are handled explicitly.

Document Similarity via Cosine Distance

When comparing document lengths that vary dramatically—such as a 2-page RBA media release against a 60-page Statement on Monetary Policy—Euclidean distance is distorted by document length. Instead, text analytics measures angular alignment between sparse TF-IDF vectors $\mathbf{u}$ and $\mathbf{v}$ using **cosine similarity**:

$$\cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|_2 \, \|\mathbf{v}\|_2} = \frac{\sum_{t=1}^{V} u_t v_t}{\sqrt{\sum_{t=1}^{V} u_t^2} \, \sqrt{\sum_{t=1}^{V} v_t^2}}$$

Where:

- $u_t, v_t \geq 0$ are the TF-IDF weights of vocabulary term t in documents $\mathbf{u}$ and $\mathbf{v}$.
- $\cos(\theta) \in [0, 1]$ equals 1 when two documents share identical term proportions and 0 when their vocabularies are completely disjoint.

For instance, two disclosures with TF-IDF vectors $\mathbf{u} = (0.55, 0.20, 0)^\top$ and $\mathbf{v} = (0.40, 0.30, 0.50)^\top$ across terms $(\text{lithium}, \text{export}, \text{solar})$ yield dot product $0.55(0.40) + 0.20(0.30) = 0.28$, norms $\sqrt{0.3425} \approx 0.585$ and $\sqrt{0.50} \approx 0.707$, and cosine similarity $0.28/(0.585 \times 0.707) \approx 0.677$.

4.4 Spatial Data Foundations and Mapping Australian Regions

Geographic data link quantitative indicators to physical locations across Australia. Modern spatial pipelines represent vector geography under the Open Geospatial Consortium **Simple Features (`sf`)** standard, storing each geographic observation as a row in a tidy data frame paired with an explicit geometry list-column supporting `POINT` (store addresses, weather stations), `LINESTRING` (rail corridors, rivers), and `POLYGON` / `MULTIPOLYGON` (suburb, LGA, or island state boundaries such as Tasmania and Victoria).

Coordinate Reference Systems (CRS) and Projections

Because the Earth is an oblate spheroid rather than a flat plane, every spatial dataset requires an explicit **Coordinate Reference System (CRS)** specifying how coordinates map onto physical space:

- **Geographic (Unprojected) CRS:** Measures angular longitude and latitude in decimal degrees on a 3D ellipsoid—principally `WGS84` (`EPSG:4326`) used by GPS devices globally, and Australia's official geodetic datum `GDA2020` (`EPSG:7844`). Angular degrees do not maintain equal metric area or distance across latitudes, making raw degree plotting distorted for continental area comparisons.
- **Projected CRS:** Projects curved ellipsoid coordinates onto a flat 2D metric plane (metres), trading off between preserving area, shape (conformality), or distance. For national Australian thematic maps comparing regional densities, **Australian Albers Equal Area (`EPSG:3577`)** preserves true relative land areas across the continent. For city-scale engineering or Melbourne municipal distances, metric UTM zones such as Map Grid of Australia 2020 Zone 55 (`EPSG:7855`) minimise local distortion.

Australian Statistical Geography Standard (ASGS) and Choropleth Pitfalls

When mapping demographic, retail, or health data in Australia, analysts rely on the ABS **Australian Statistical Geography Standard (ASGS)** nested hierarchy rather than irregular postal codes (Australia Post postcodes are delivery routes that cross state borders and change arbitrarily without fixed population bounds):

ASGS / Administrative Unit	Approximate Australian Count	Typical Population Size	Primary Analytical Role
Mesh Block (MB)	~368,000	30–60 dwellings	Smallest building block; land-use classification
Statistical Area Level 1 (SA1)	~61,800	200–800 persons	Finest unit for Census small-area release

ASGS / Administrative Unit	Approximate Australian Count	Typical Population Size	Primary Analytical Role
Statistical Area Level 2 (SA2)	~2,473	3,000–25,000 persons	Represents functional suburbs and towns
Statistical Area Level 3 (SA3)	~359	30,000–130,000 persons	Regional labour markets and service catchments
Statistical Area Level 4 (SA4)	~108	100,000–500,000+ persons	ABS Labour Force Survey sub-state reporting
Local Government Area (LGA)	~547	Highly variable	Municipal council policy and local infrastructure

Two severe analytical pitfalls arise in spatial mapping:

1. **Area-Size Bias in Choropleths:** Standard choropleth maps shade polygons by value. Across Australia, sparsely populated outback SA4s in Western Australia, South Australia, and the Northern Territory span hundreds of thousands of square kilometres, whereas densely populated Melbourne and Sydney SA2s occupy tiny pixel footprints. Mapping raw totals (e.g., total retail spend) rather than rate-normalised metrics (e.g., spend per capita), or failing to use inset metropolitan panels, tile cartograms, or dot-density overlays, causes vast rural regions to visually dominate national patterns.
2. **Modifiable Areal Unit Problem (MAUP):** Summary statistics and correlations computed over aggregated spatial polygons shift purely as a function of boundary zoning and aggregation scale (SA1 vs. SA3 vs. LGA). Analysts must interpret ecological correlations at the spatial scale of aggregation and avoid attributing regional averages directly to individual residents (the *ecological fallacy*).

Spatial Joins and Geometric Predicates

Unlike tabular key joins, a **spatial join** (`st_join()`) links two datasets through topological geometry predicates—most commonly point-in-polygon containment (`st_within` or `st_intersects`). For example, joining geocoded Melbourne retail store coordinates (`POINT` geometry in `EPSG:7844` GDA2020) onto ABS SA2 polygons (ensuring both layers share an identical CRS via `st_transform()` and fixing self-intersecting rings via `st_make_valid()`) enriches each store record with official Census SA2 median income, population density, and commute share without relying on error-prone suburb name string matching.

Choropleth Binning Schemes for Skewed Spatial Distributions

When shading discrete polygons using a sequential colour ramp, the choice of numerical class intervals substantially alters visual perception:

- **Equal Interval Binning:** Divides the full data range $[\min, \max]$ into equal-width spans. On right-skewed Australian spatial metrics (such as dwelling price per m^2 or population density across SA2s), a handful of inner-Sydney or inner-Melbourne outliers compress 95% of regions into the lowest colour bin, masking genuine regional heterogeneity.
- **Quantile (Equal-Count) Binning:** Assigns an equal number of polygons to each bin (e.g., quartiles or quintiles), maximising visual contrast across dense suburban areas, though wide gaps

Degreelive

can separate adjacent values near the upper tail.

- **Fisher-Jenks Natural Breaks:** Minimises within-class variance while maximising between-class variance, isolating natural clustering thresholds in skewed geographic distributions.

4.5 Reproducible Data Communication, Narratives & Decision Support

Technical rigour yields zero institutional impact unless analytical findings are communicated accurately, concisely, and transparently to decision-makers. Effective commercial and policy data communication integrates structural narrative discipline, visual clarity, and ethical restraint.

Structuring Executive and Technical Narratives

Business and policy readers differ sharply from academic peers in how they consume analytical reports. Rather than burying key findings at the end of a chronological diary of cleaning steps, decision-oriented reports apply the **inverted pyramid structure**:

- **Executive Lead (Decision First):** Opens immediately with the headline empirical answer, quantified business or policy magnitude (in clear units such as AUD millions or percentage points), and key recommendation.
- **Evidence Synthesis and Visual Hierarchy:** Supports each claim with a tightly edited figure or summary table. Replacing generic descriptive titles (`"Retail Sales by State, 2021–2025"`) with declarative takeaway titles (`"Victorian Retail Turnover Grew 6.4% Faster Than National Average in FY2025"`), direct-labelling key series lines instead of forcing eye movement to distant legends, and annotating structural breaks (such as RBA rate cycles or border reopenings) accelerates reader comprehension.
- **Methodology, Uncertainty, and Limitations:** Explicitly reports sample scope, missing-data treatment, out-of-sample error bounds (confidence intervals or CV RMSE), and model assumptions in a transparent technical appendix or collapsible code record so external auditors can verify every transformation end-to-end.

Ethical Data Storytelling and Epistemic Integrity

Analysts carry professional responsibility for ensuring visualisations and models inform rather than manipulate:

- **Avoiding Selective Baselines and Cherry-Picked Windows:** Starting a time-series chart at a cyclical trough exaggerates growth, while truncating bar-chart zero baselines distorts relative proportions.
- **Guarding Against p-Hacking and Post-Hoc Hypothesising:** Testing dozens of subgroup splits or transformations until one achieves $p < 0.05$ without multiple-testing correction or holdout validation mistakes random noise for real signal.
- **Distinguishing Correlation from Causation:** Observational regression slopes and clustering profiles describe empirical association rather than guaranteed policy intervention effects whenever confounding variables or selection bias remain uncontrolled.
- **End-to-End Provenance Auditability:** Linking every published figure back through versioned scripts to immutable ABS, RBA, ASX, or internal source extracts ensures that strategic commercial decisions rest on verifiable, reproducible evidence.

Degreelive

Chapter 4: End-to-End Predictive Analytics & Statistical Decision Architecture

4.1 Binary Classification & Logistic Regression in Commercial Analytics

While linear regression models continuous monetary or physical outcomes, business decisions frequently involve binary categorical responses: whether an Australian SME defaults on a commercial bank loan, whether a retail subscriber churns from a telecommunications contract, or whether a credit card transaction represents fraudulent activity. Fitting an ordinary least squares (OLS) linear probability model to a binary indicator outcome $Y \in \{0, 1\}$ produces estimated probabilities that can fall outside the logically admissible $[0, 1]$ probability interval, violates the assumption of constant error variance (causing inherent heteroskedasticity), and imposes unrealistic constant marginal effects across the entire range of predictor variables.

Logistic regression resolves these structural limitations by modelling the natural logarithm of the odds (the log-odds or logit transformation) as a linear combination of explanatory predictors:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

Where:

- $p_i = P(Y_i = 1 \mid \mathbf{X}_i)$ is the conditional probability that individual observation i experiences the target event.
- $\frac{p_i}{1-p_i}$ represents the odds of the event occurring relative to not occurring.
- $\ln\left(\frac{p_i}{1-p_i}\right)$ is the continuous logit link function mapping probabilities from $(0, 1)$ onto $(-\infty, \infty)$.
- β_0 is the baseline log-odds when all explanatory covariates equal zero.
- β_j represents the marginal change in the log-odds of the outcome associated with a one-unit increase in predictor X_j, holding other variables constant.

Inverting the logit transformation yields the logistic cumulative distribution function, which guarantees that predicted probabilities remain bounded strictly between 0 and 1 for any covariate vector:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{k} \beta_j X_{ji})}} = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j X_{ji}}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j X_{ji}}}$$

Where $e \approx 2.71828$ is Euler's mathematical constant. The parameters of the logistic regression model cannot be solved analytically via closed-form OLS formulas; instead, they are estimated numerically using maximum likelihood estimation (MLE) algorithms, which iteratively maximise the log-likelihood function across all observed data points.

DegreeLive

4.2 Interpreting Logistic Regression: Odds Ratios and Marginal Effects

Because the log-odds scale is non-intuitive for commercial stakeholders, analysts transform logistic coefficients into odds ratios (OR) by exponentiating the estimated slope parameters:

$$\mathrm{OR}_j = \exp(\beta_j) = e^{\beta_j}$$

Where:

- An odds ratio $\mathrm{OR}_j > 1$ indicates that a one-unit increase in covariate X_j multiplies the baseline odds of the target event by a factor greater than one, increasing the likelihood of occurrence.
- An odds ratio $\mathrm{OR}_j < 1$ indicates that an increase in covariate X_j reduces the odds of the event occurring.
- An odds ratio $\mathrm{OR}_j = 1$ indicates that the predictor has no association with the odds of the response.

For small percentage changes in continuous predictors or for categorical factor variables, the percentage change in the odds of the outcome is computed directly as $(\mathrm{OR}_j - 1) \times 100\%$. Furthermore, average marginal effects (AME) compute the derivative of the probability curve averaged over all observed observations, translating non-linear logit relationships into actionable percentage-point shifts suitable for executive briefings.

4.3 Confusion Matrix, Classification Thresholds & Diagnostic Performance

Converting predicted continuous probabilities $\hat{p}_i$ into discrete operational decisions requires selecting a decision threshold $c \in (0, 1)$, typically initialised at $c = 0.50$:

$$\hat{Y}_i = \begin{cases} 1 & \text{if } \hat{p}_i \geq c \\ 0 & \text{if } \hat{p}_i < c \end{cases}$$

Cross-tabulating predicted classifications against actual empirical outcomes across the test dataset produces a 2×2 confusion matrix containing four foundational counts:

- **True Positives (TP):** Actual positives correctly classified as positive by the decision rule.
- **True Negatives (TN):** Actual negatives correctly classified as negative.
- **False Positives (FP):** Actual negatives incorrectly classified as positive (Type I error).
- **False Negatives (FN):** Actual positives incorrectly classified as negative (Type II error).

Standard evaluative performance metrics derived from these counts include:

$$\text{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

DegreeLive

$$\text{Sensitivity (Recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Precision (Positive Predictive Value)} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$F_1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

Where the F_1-score represents the harmonic mean of precision and recall, providing a balanced diagnostic measure that remains robust even when classes are highly imbalanced.

4.4 Worked Numerical Example: Credit Risk Default Classification for Australian SMEs

An Australian commercial credit assessment team evaluates a validation holdout sample of $n = 1{,}000$ commercial SME loan accounts. The estimated logistic model uses two normalised predictors: the debt-to-equity ratio X_1 and the interest coverage ratio X_2. The fitted logit equation is:

$$\ln\left(\frac{\hat{p}_i}{1 - \hat{p}_i}\right) = -2.40 + 0.85X_{1i} - 0.60X_{2i}$$

For a prospective borrower with debt-to-equity ratio $X_1 = 2.0$ and interest coverage ratio $X_2 = 1.5$:

$$\text{Logit Score} = -2.40 + 0.85(2.0) - 0.60(1.5) = -2.40 + 1.70 - 0.90 = -1.60$$

$$\hat{p}_i = \frac{1}{1 + e^{-(-1.60)}} = \frac{1}{1 + e^{1.60}} = \frac{1}{1 + 4.9530} = \frac{1}{5.9530} \approx 0.1680 \quad (16.80\%)$$

Because $\hat{p}_i = 16.80\% < 50\%$, applying standard threshold $c = 0.50$ classifies the borrower as non-default ($\hat{Y} = 0$).

Across the entire validation holdout sample of 1,000 firms, the observed contingency results are:

- True Positives (TP) $= 70$ defaulted firms correctly flagged.
- False Positives (FP) $= 30$ healthy firms mistakenly flagged.
- False Negatives (FN) $= 30$ defaulting firms overlooked.

Degreelive

- True Negatives (TN) $= 870$ healthy firms correctly approved.

Step-by-step diagnostic calculations:

1. **Overall Classification Accuracy:**

$$\text{Accuracy} = \frac{70 + 870}{1{,}000} = \frac{940}{1{,}000} = 94.0\%$$

2. **Sensitivity (Recall / Hit Rate):**

$$\text{Sensitivity} = \frac{70}{70 + 30} = \frac{70}{100} = 70.0\%$$

   The model detects 70% of all actual defaulters.

3. **Specificity (True Negative Rate):**

$$\text{Specificity} = \frac{870}{870 + 30} = \frac{870}{900} = 96.67\%$$

4. **Precision (Positive Predictive Value):**

$$\text{Precision} = \frac{70}{70 + 30} = \frac{70}{100} = 70.0\%$$

5. **Harmonic Mean F_1-Score:**

$$F_1 = 2 \times \frac{0.70 \times 0.70}{0.70 + 0.70} = \frac{0.98}{1.40} = 0.7000 \quad (70.0\%)$$

Varying the decision threshold c from 0.0 to 1.0 traces out the Receiver Operating Characteristic (ROC) curve, plotting Sensitivity on the vertical axis against $(1 - \text{Specificity})$ (the False Positive Rate) on the horizontal axis. The Area Under the Curve (ROC-AUC) provides a threshold-independent measure of discriminative capacity: an AUC of 0.50 denotes pure random guessing, while an AUC exceeding 0.85 confirms strong commercial discriminative utility.

DegreeLive

4.5 Summary Table: Evaluative Diagnostic Metrics in Business Data Analytics

Metric	Formula	Core Focus	Commercial Significance
Accuracy	$\frac{\text{TP}+\text{TN}}{\text{Total}}$	Overall correct proportion	Misleading when target class is rare (e.g. 1% fraud).
Sensitivity (Recall)	$\frac{\text{TP}}{\text{TP}+\text{FN}}$	Positive capture rate	Critical when missing an event is costly (credit default).
Specificity	$\frac{\text{TN}}{\text{TN}+\text{FP}}$	Negative capture rate	Measures ability to avoid false alarms on healthy accounts.
Precision	$\frac{\text{TP}}{\text{TP}+\text{FP}}$	Reliability of positive flags	Ensures commercial investigations target genuine risks.
F_1**-Score**	$\frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}$	Harmonic mean of P & R	Robust benchmark for imbalanced business classifications.
ROC-AUC	$\int_0^1 \text{Sensitivity}\, d(1-\text{Spec})$	Threshold-invariant separation	Standard benchmark for comparing competing algorithms.

DegreeLive