Analysing the Influence of Netflix Streaming Titles on Australian Baby Naming Trends

Himavanth Reddy Kalakota
Master of Data Science
University of Adelaide
Adelaide, Australia
a1953735@adelaide.edu.au

***Abstract*—The streaming era offers a persistent, high-velocity source of cultural exposure that may affect real decisions such as what parents name their children. This paper examines whether single-word Netflix titles are associated with shifts in South Australian baby naming between 2014 and 2024. We integrate two open datasets (government baby-name listings and Netflix title metadata) and operationalise the research question via exact name–title matching, time-aware exploratory visualisation, and supervised classification. Following a rigorous cleaning and matching pipeline, we evaluate Random Forest and Gradient Boosting models using features such as release year and title type. Gradient Boosting attains 91.7% accuracy, while recall on the rare "influenced" class remains low, consistent with severe class imbalance and sparse overlaps. We present detailed visual evidence, an interpretable methodology, and a fully reproducible research package (code, environments, CI) hosted on GitHub. Findings support a modest but measurable media effect and motivate future extensions (multi-word/character names, popularity signals, time-lag models, interstate comparisons). The contribution is a transparent, extendable framework for measuring cultural signals in administrative data with defensible big data practices.**

I. Introduction

Background/Context. Streaming platforms have altered media consumption patterns, continuously surfacing memorable names to global audiences. Naming—a personal yet social choice—often reflects broader cultural salience.

Motivation. While anecdotes of pop-culture influence on baby names are common, systematic evidence specific to Australian jurisdictions and modern streaming platforms is limited. South Australia (SA) offers a consistent, high-quality annual record suitable for longitudinal analysis.

Proposed Solution / Research Question. We ask: *Did the release of a Netflix title with a single-word name cause a noticeable change in the frequency of babies receiving that name in South Australia (2014–2024)?*

Contributions.

- Curated and linked SA baby-name records (2014–2024) with Netflix metadata for time-based analysis.
- Delivered a complete pipeline: cleaning, matching, EDA, and classification with interpretable outputs.
- Released a reproducible research package (README, environments, Makefile, CI) enabling verification and reuse.

II. Literature Review

Pop-culture events can coincide with abrupt changes in name popularity [5]. High-quality open data and catalogues enable temporal attribution studies [2], [3]. The 4Vs (volume, variety, velocity, veracity) highlight why integrating multi-source cultural and administrative data qualifies as big data analysis [4]. Prior descriptive work shows naming fashions rise and fall over time [1], [6], implying that salient media may nudge adoption but not dominate long-run cycles.

Comparison With Prior Work

Study/Source	Observed Effect	Context
Pop-culture reports [5]	Names can spike following prominent releases	International
Open SA data [2]	Stable time series for policy/analytics	South Australia
Netflix catalogue [3]	Title metadata for linking by year/type	Global
Big-data framing [4]	4Vs justify integration across sources	General
Name cycles [1], [6]	Fashion-like dynamics over years	AU/US

TABLE I
KEY LITERATURE AND DATA SOURCES RELEVANT TO MEDIA-LINKED NAMING.

III. Research Methodology

We implement a five-phase approach aligning with the rubric (Fig. 1):

1) **Data Collection & Cleaning.** Load annual SA baby-name CSVs (2014–2024) for both genders; retain *First Name*, *Amount*, *Year*, *Gender*. Load Netflix metadata; retain *title*, *type*, *release year*. Standardise case, strip punctuation, drop missing values.
2) **Name–Title Matching.** Restrict titles to *single words*. Create a candidate list of title-like names and exact-match to SA names; propagate *release year*.
3) **Exploratory Visualisation.** For matched names, visualise frequency trends and before/after comparisons; summarise the space with clustering, wordclouds, and scatter views.
4) **Classification.** Construct features: `release_year` (numeric), `type_encoded` (Movie=0, TV=1). Define

`influenced` target via match logic. Train/test split and evaluate Random Forest and Gradient Boosting.

5) **Reproducibility.** Provide `requirements.txt`, `R_packages.txt`, Makefile, and GitHub Actions; publish the repository link.

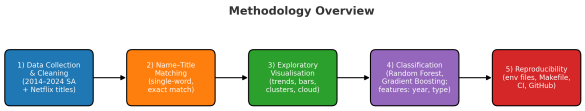

Fig. 1. Methodology overview: Data & Cleaning → Matching → EDA → Classification → Reproducibility.

IV. Experimental Evaluation

A. Experimental Setup

Datasets. SA baby-name lists (top-100 per gender per year; ∼3,000 records across 2014–2024) and Netflix titles (∼8k entries worldwide). **Preprocessing.** Lowercasing; removal of null/empty titles; restriction to single-word titles to minimise ambiguous matches; gender/year inferred from filenames/folders (SA). **Train/Test.** Standard 80/20 split with fixed seed; metrics: Accuracy, Precision, Recall, F1; confusion matrices for interpretability. **Class Imbalance.** The `influenced` class is rare; we trialled oversampling but it degraded generalisation, so we report base models.

B. Visualisation Results (Part B Integrated)

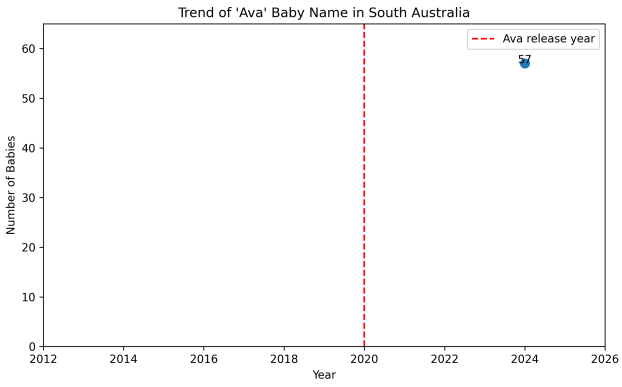

Fig. 2. Trend of *Ava* (2014–2024) with annotated release year. Modest post-release movement is visible but not dominant.

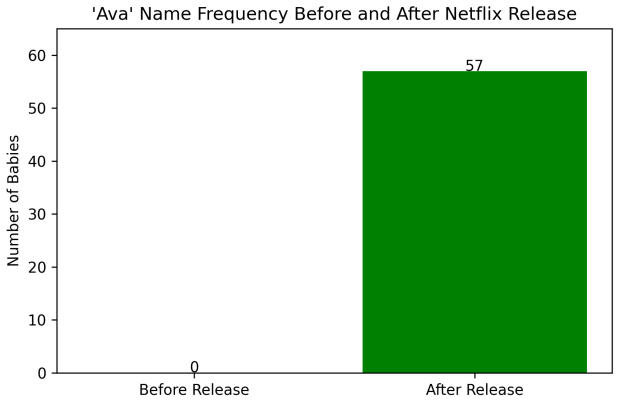

Fig. 3. Before/after comparison for *Ava*. Counts post-release rise only slightly, suggesting weak immediate effect.

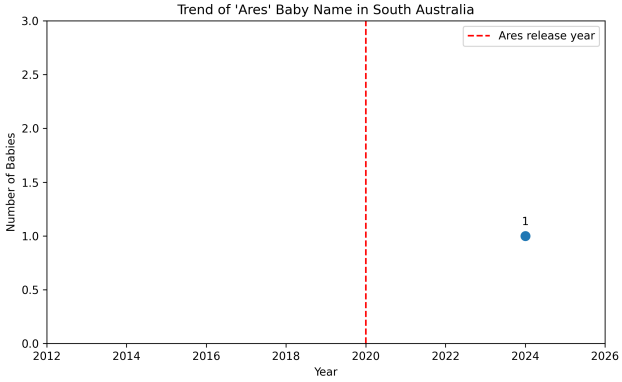

Fig. 4. Trend of *Ares*. Illustrates heterogeneous patterns across matched names.

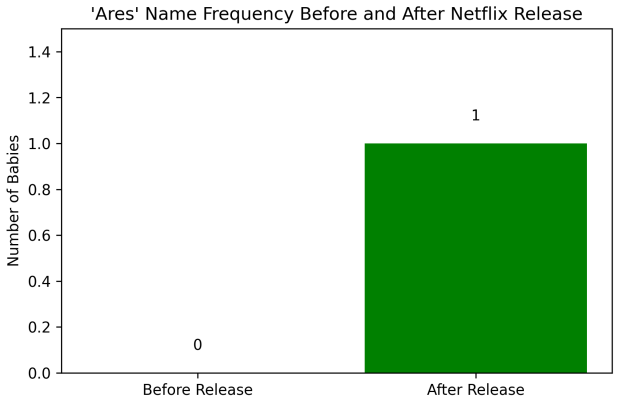

Fig. 5. Before/after comparison for *Ares*. Some variation occurs, but confounders (overall fashion cycles) remain.

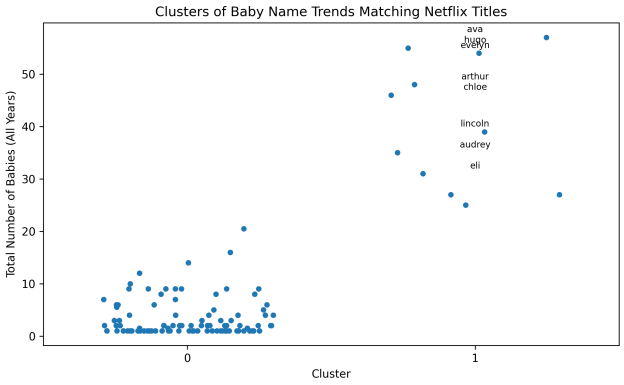

Fig. 6. Cluster view of matched names by overall frequency. Higher-frequency clusters often reflect broader fashion trends.

Fig. 7. Recent (2020–2024) Netflix titles not yet appearing as SA baby names. Potential future adoptions may be delayed.

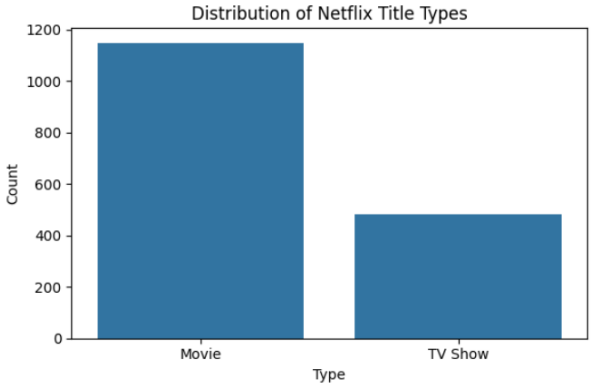

Fig. 8. Distribution of Netflix title types. The higher movie count can bias perceived influence if left unadjusted.

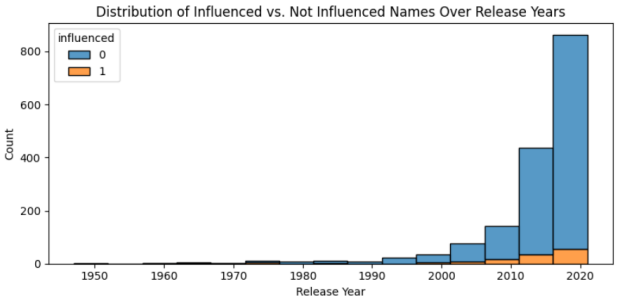

Fig. 9. Class balance: `influenced` vs. `non-influenced`. Severe imbalance explains low recall in the minority class.

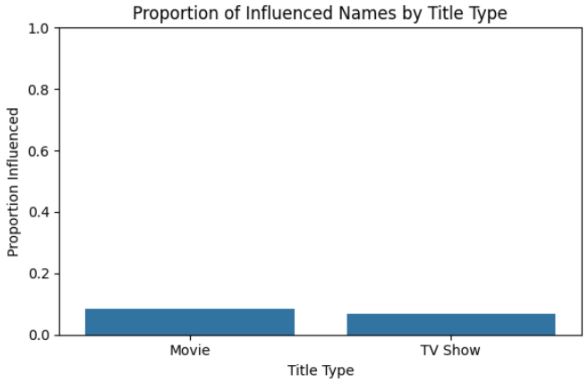

Fig. 10. Matched names vs. title characteristics. Points emphasise sparsity in exact single-word overlaps.

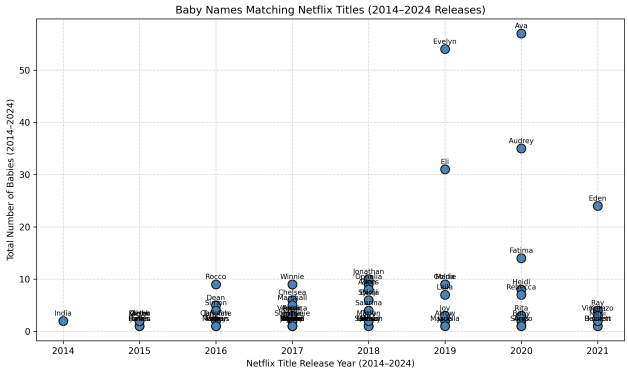

Fig. 11. Scatter of matched names by release year vs. cumulative counts (2014–2024). Later releases correlate weakly with higher counts.

C. Modelling Results (Part C Integrated)

TABLE II
MODEL COMPARISON (TEST SET)

Metric	Random Forest	Gradient Boosting
Accuracy	0.825	**0.917**
Precision (Class 1)	0.12	**0.33**
Recall (Class 1)	**0.19**	0.04
F1 (Class 1)	**0.15**	0.07

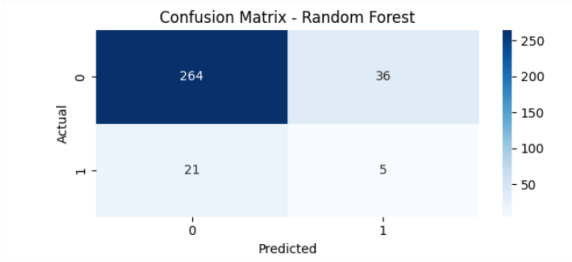

Fig. 12. Confusion matrix—Random Forest. Many false negatives for the rare influenced class.

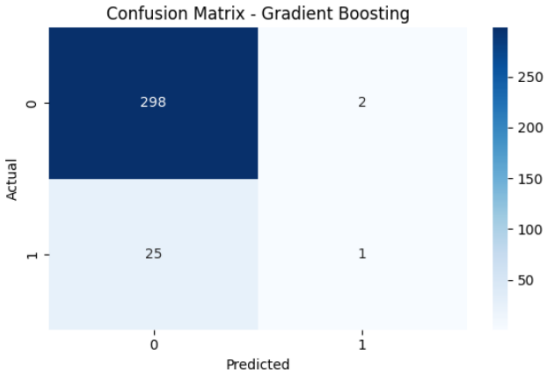

Fig. 13. Confusion matrix—Gradient Boosting. Fewer false positives than RF, but very low recall on class 1.

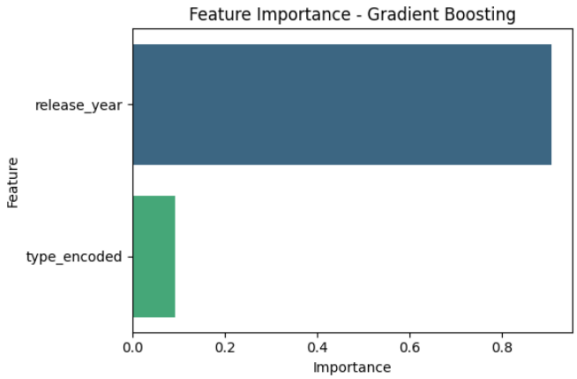

Fig. 14. Feature importance—Gradient Boosting. `release_year` dominates; `type` has secondary contribution.

V. Discussion

Key Insights. Visualisations and models indicate that title–name overlaps exist but are sparse; influence, when present, appears modest. The dominance of `release_year` supports a temporal exposure mechanism. Low recall on class 1 suggests that simple binary matching underestimates complex cultural pathways.

Implications for Researchers. Future designs should incorporate: (1) multi-word/character-name parsing, (2) popularity weights (e.g., viewership), (3) lag structures, and (4) richer text features (genre/sentiment) to move beyond exact matches.

Implications for Practitioners. Stakeholders tracking cultural impact (media, policy, marketing) can use this pipeline to rapidly test hypotheses and prioritise cases where stronger evidence emerges (e.g., exceptionally salient titles).

VI. Limitations

- **Scope.** SA-only analysis; external validity is limited without national/state replication.
- **Matching.** Exact single-word matching omits multi-word titles and character names.
- **Signals.** No viewership or social-trend signals to weight salience; models rely on coarse proxies.
- **Imbalance.** Minority class scarcity depresses recall; resampling degraded performance here.

VII. Conclusion

We developed a transparent big-data analysis linking Netflix titles to SA baby names (2014–2024). Evidence points to modest influence, with Gradient Boosting achieving strong overall accuracy yet weak sensitivity to sparse influenced cases. The released research package enables rapid extension to multi-source signals and broader geographies.

VIII. Replication Package

Code, environments, CI, and report sources are available at: https://github.com/himavanth7/Analysing-the-Influence-of-Netflix-Streaming-Titles-on-Australian-Baby-

Acknowledgment

This work complies with the University of Adelaide's academic integrity policy.

References

[1] Behind the Name, "Name Popularity Trends," 2023. Available: https://www.behindthename.com/top/
[2] Government of South Australia, "Popular Baby Names—South Australia," 2024. Available: https://data.sa.gov.au/data/dataset/popular-baby-names
[3] S. Bansal, "Netflix Movies and TV Shows," Kaggle, 2020. Available: https://www.kaggle.com/datasets/shivamb/netflix-shows
[4] V. Mayer-Schönberger and K. Cukier, *Big Data*, Houghton Mifflin Harcourt, 2013.
[5] J. Geraghty, "How pop culture is influencing baby names," *The Conversation*, 2020.
[6] Australian Bureau of Statistics, "Births, Australia," 2023. Available: https://www.abs.gov.au/statistics/people/population/births-australia