Data Science Exam

Precious Nhamo

Abstract

This document comprises the responses to **Questions 1 to 5** of the 2025 Data Science examination, along with separate analyses (e.g., a PowerPoint presentation) completed in accordance with the exam instructions and organised within the designated folder.

Data Science Exam

Question1 Baby Names

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Question2 Music Taste

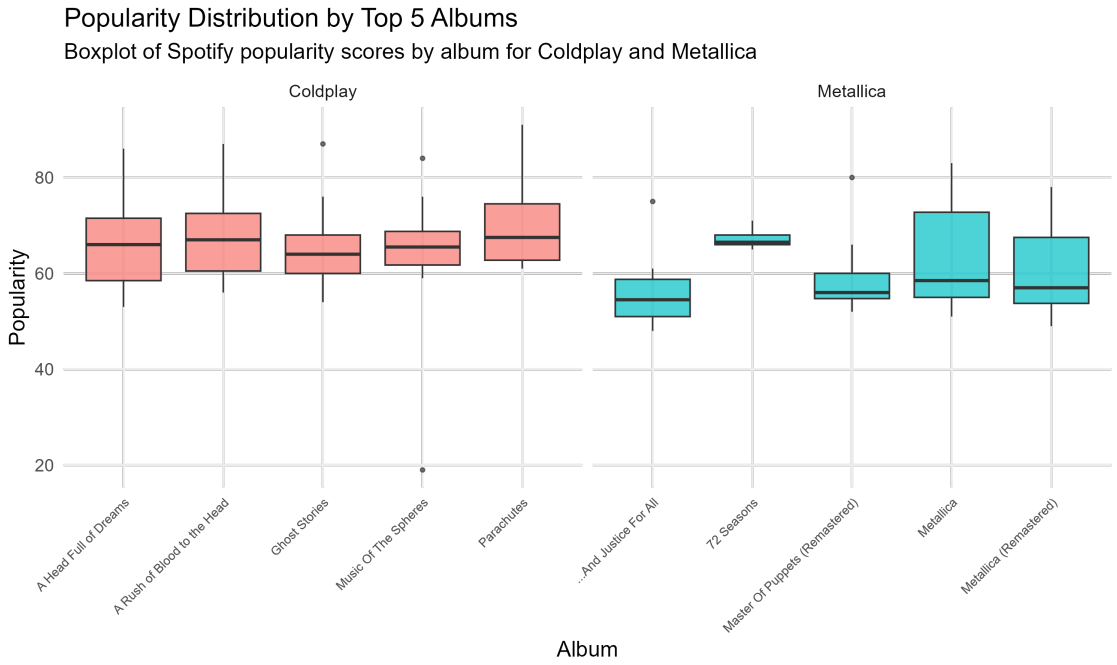

Figure 1

Longevity and Musical Progression of Coldplay and Metallica

The data reveals distinct trajectories for Coldplay and Metallica in terms of popularity, musical evolution, and industry adaptation. Coldplay demonstrated early dominance, charting five songs in their first decade compared to Metallica's one (Figure 2). Their popularity scores on Spotify also show broader appeal, with a higher median and narrower interquartile range than Metallica's (Figure 1).

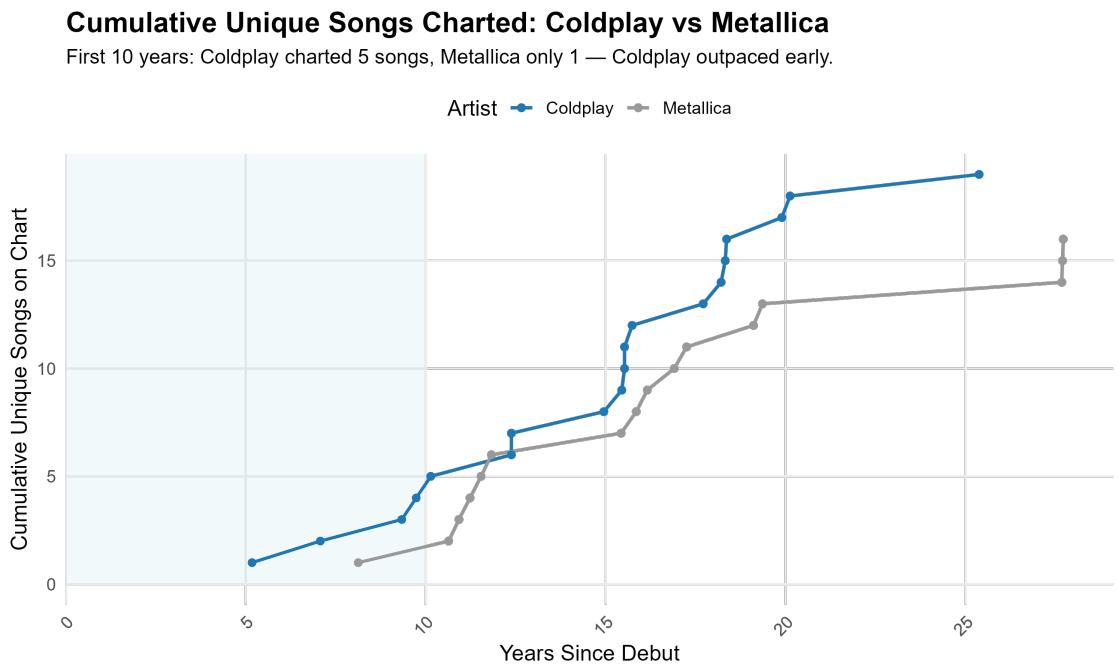

Figure 2

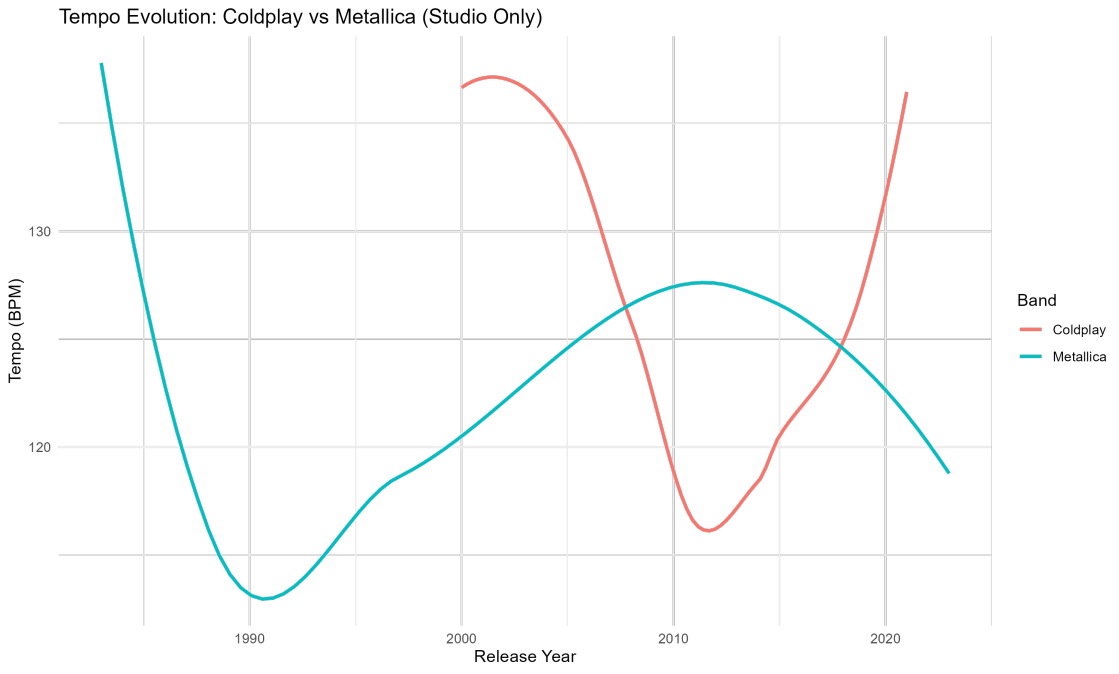

Figure 3

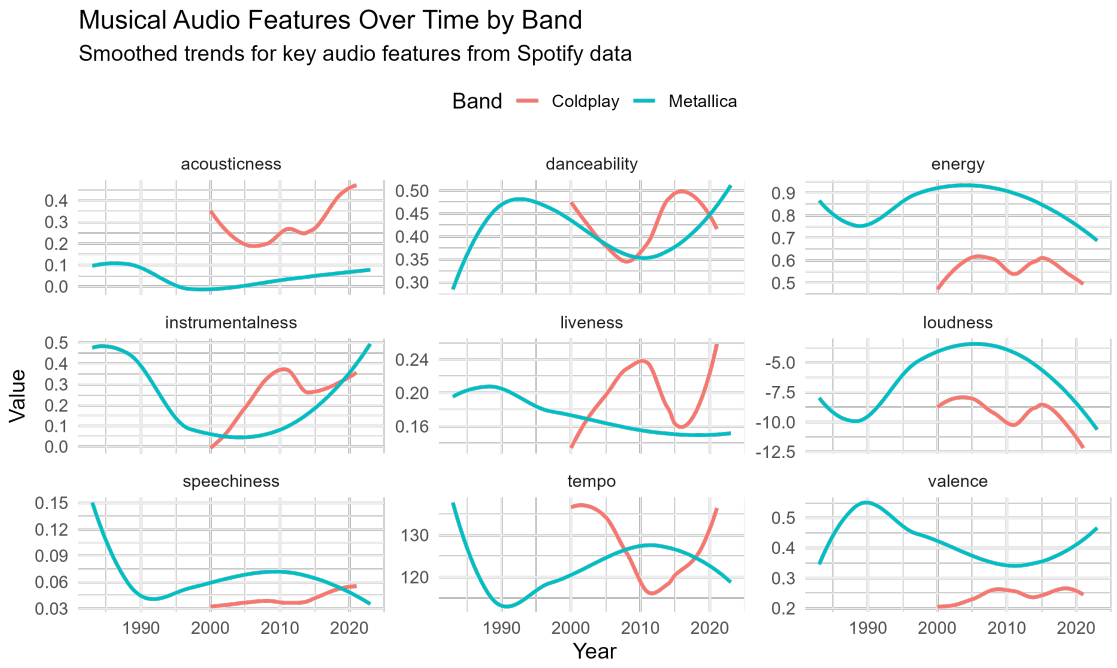

Figure 4

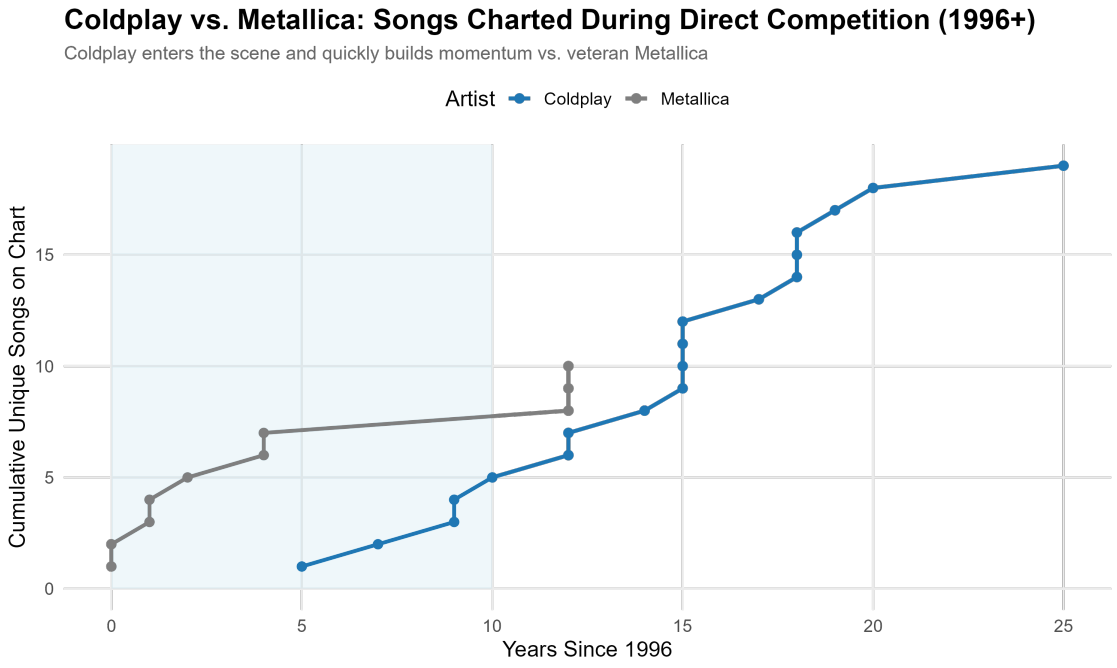

Figure 5

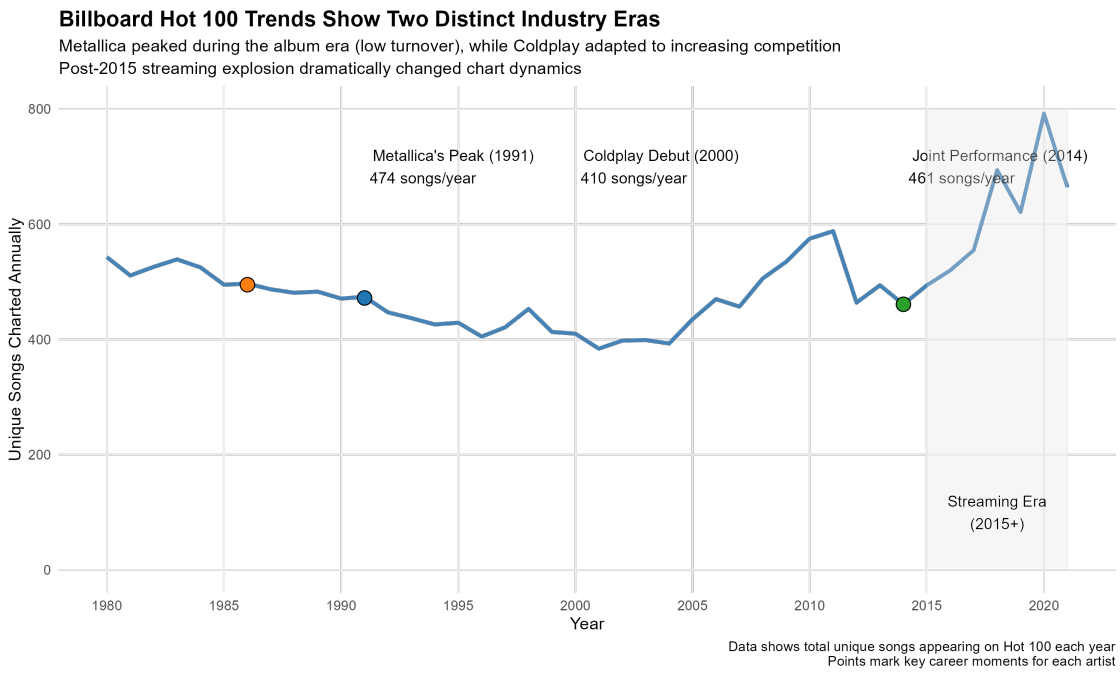

Figure 6

Musically, Coldplay's tempo has remained stable (Figure 3), while their audio features, such as danceability and valence, trended positively over time (Figure 4). Metallica, conversely, maintained higher instrumentalness and energy, reflecting their heavier style.

Billboard data highlights their adaptation to industry shifts: Metallica peaked during the album era (1991), while Coldplay thrived post-2000, leveraging streaming's rise (Figure 6). During direct competition (1996+), Coldplay's momentum outpaced Metallica's (Figure 5).

Coldplay's consistent, accessible sound contrasts with Metallica's enduring heavy metal identity. Both bands exemplify longevity but reflect divergent strategies in navigating musical trends

Question 3 : Netflix Content Strategy Analysis

In light of Netflix's recent subscriber attrition and share price volatility, a strategic review was conducted to inform potential market entry for a new streaming venture. This review draws on IMDb ratings and global production data to assess what drives success in streaming content.

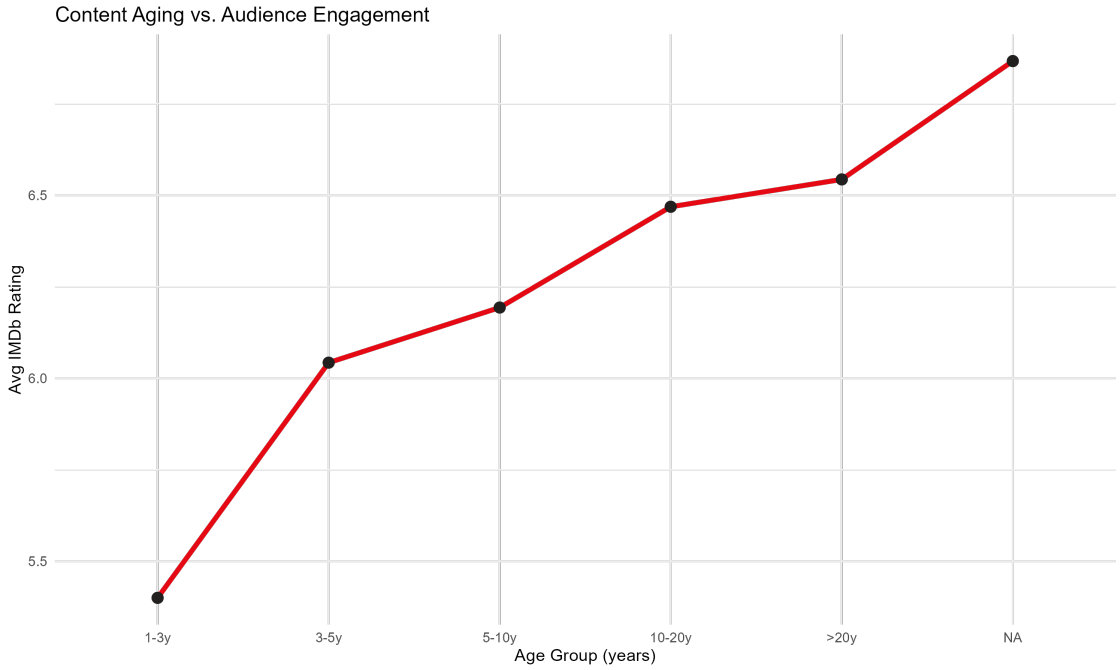

Figure 7

Key Findings

1. Quality vs. Volume Trade-Off. Japanese titles, while constituting only 1.6% of Netflix's catalogue, achieve superior average IMDb ratings (mean = 6.7, SD = ±0.3). By contrast, the United States and India collectively contribute over 50% of Netflix's content but yield lower average ratings (mean = 6.1–6.3). **Implication**: High-volume strategies may dilute perceived quality.

2. Genre Performance Differentials. Content genres vary markedly in audience reception. Documentaries (mean = 7.6) and dramas (mean = 7.0) lead on quality metrics,

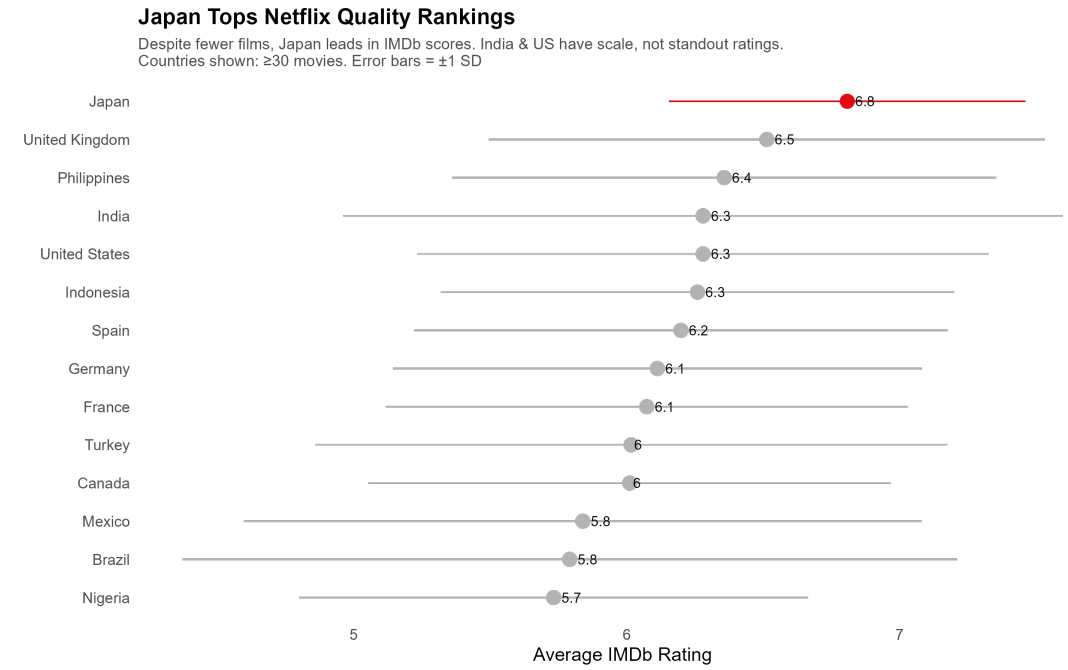

Figure 8

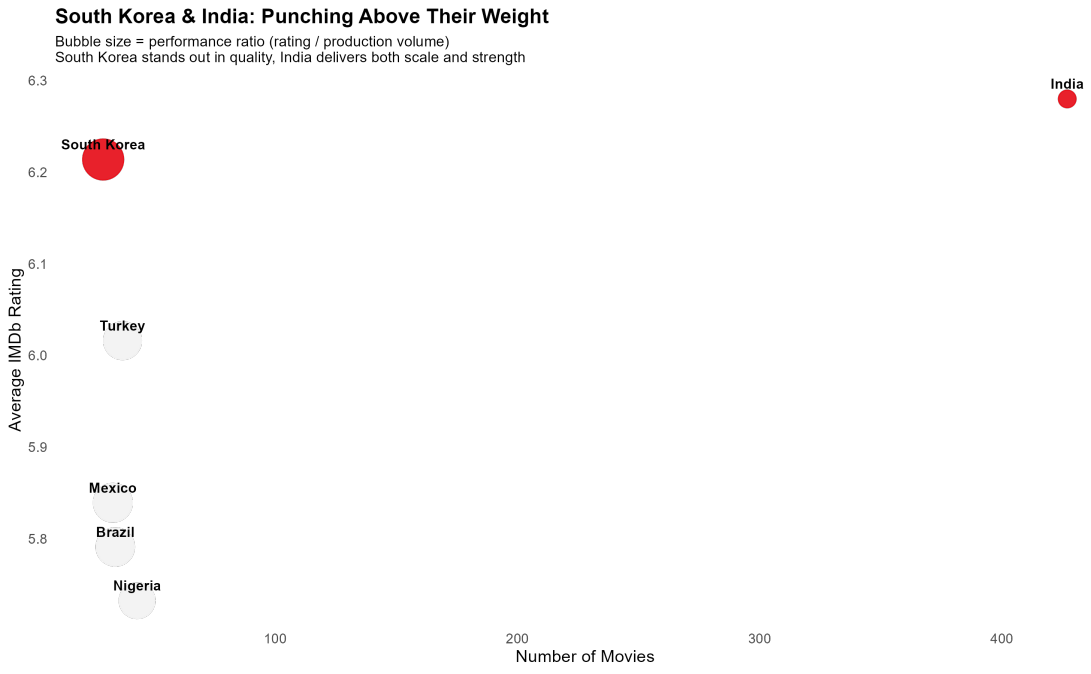

Figure 9

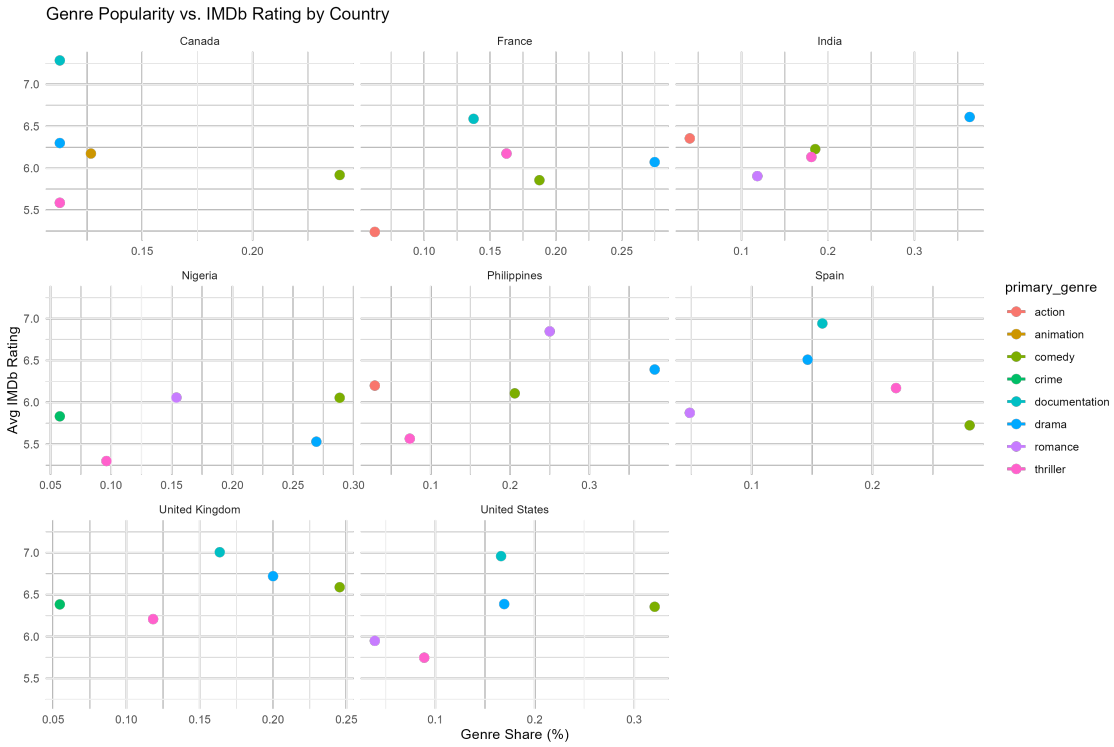

Figure 10

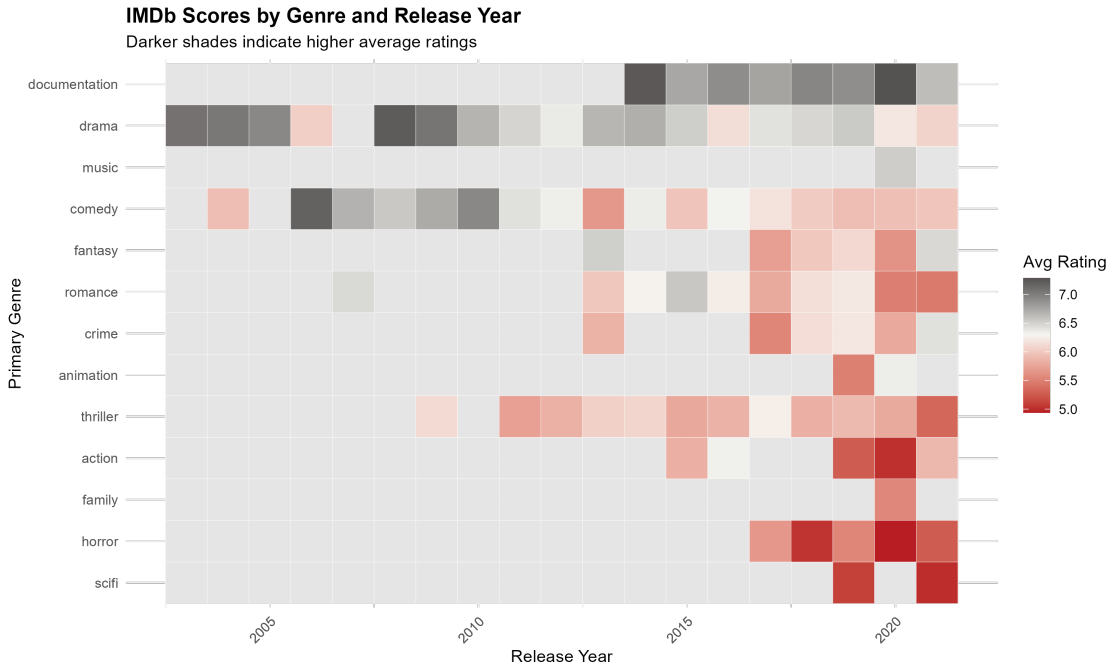

Figure 11

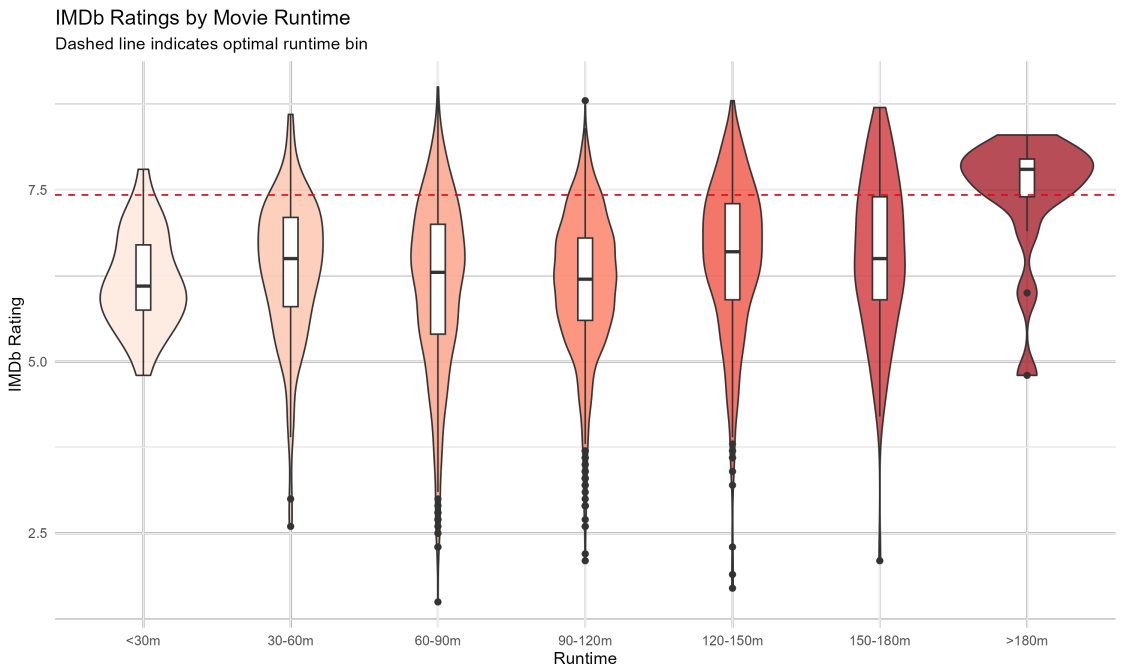

Figure 12

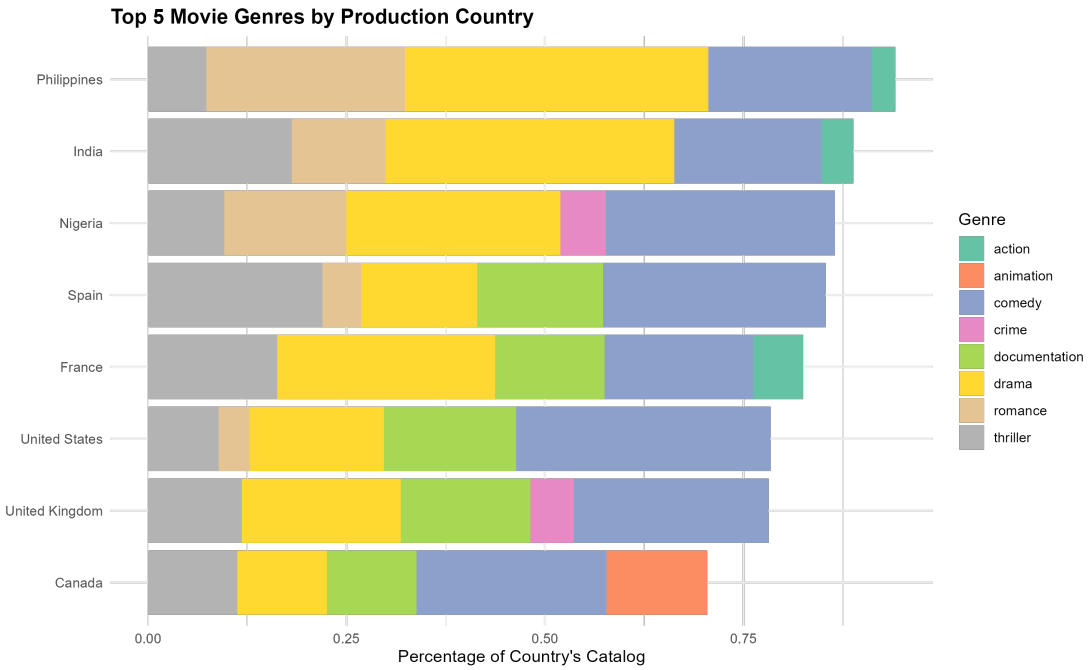

Figure 13

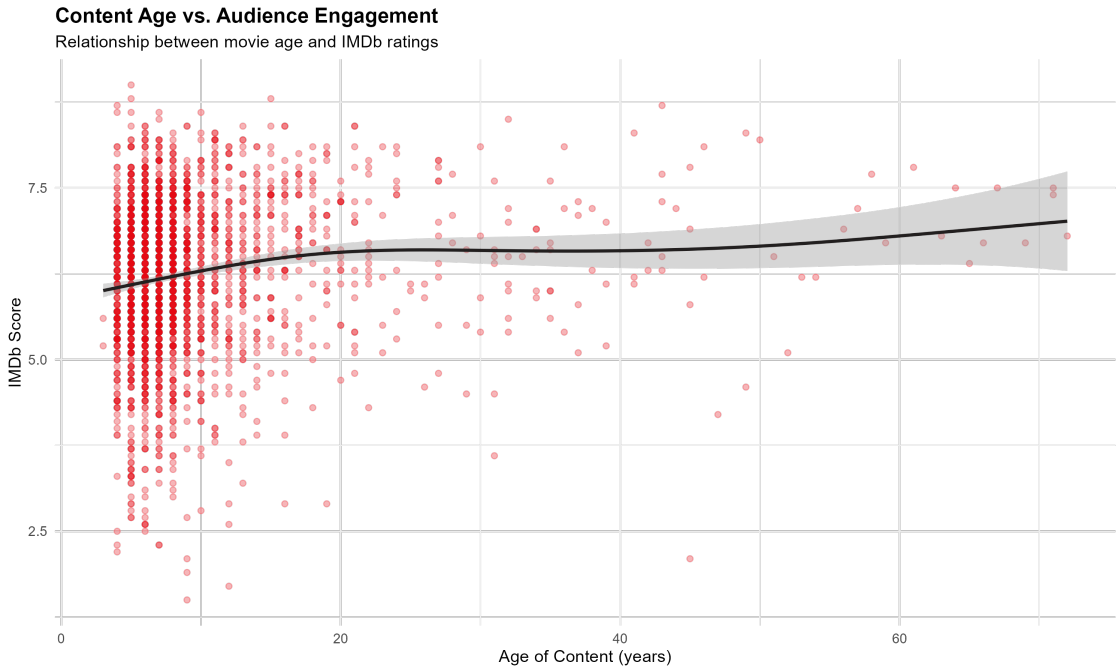

Figure 14

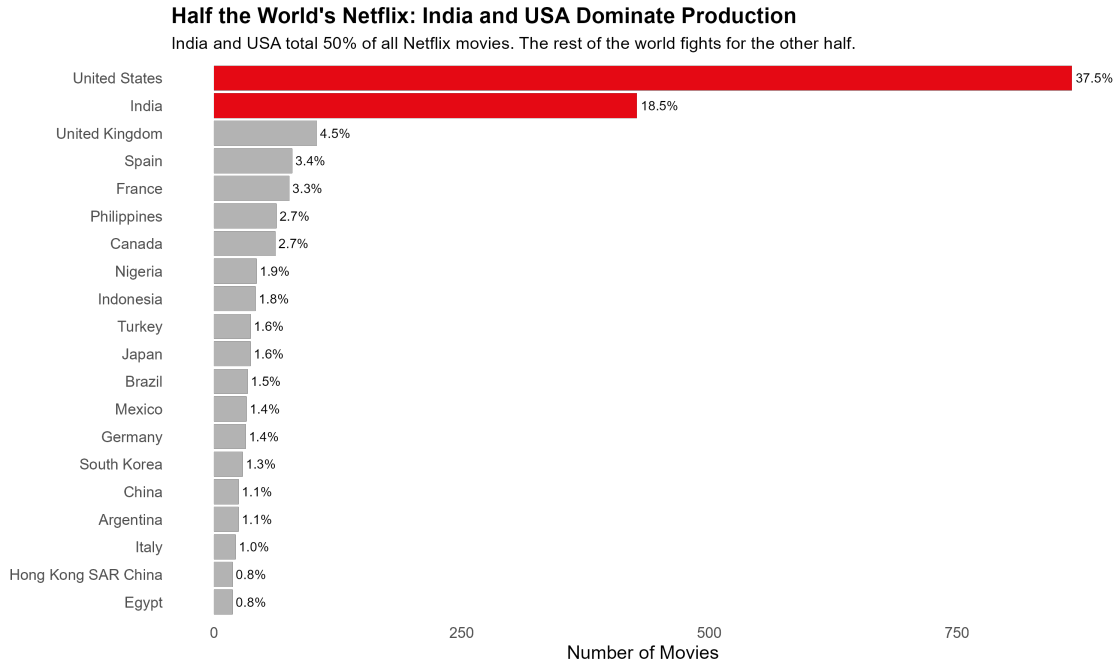

Figure 15

while action and comedy lag behind (mean = 6.2 and 6.5 respectively). Crime and thriller categories also demonstrate strong viewer engagement.

Implication:Narrative depth and authenticity drive audience satisfaction.

3. Optimal Runtime Windows. Ratings peak for films with runtimes between 90–120 minutes (mean = 7.5). Shorter or longer formats tend to correlate with lower ratings. **Implication**:Viewer engagement is maximised within a standardised runtime threshold.

4. Age and Longevity of Content. Older films (>20 years) sustain higher average ratings (mean = 8.5) compared to recent productions (1–3 years; mean = 6.0). **Implication**:Legacy content benefits from curation and nostalgia; newer content faces saturation and discovery challenges.

Strategic Recommendations for New Market Entrant

To position competitively against incumbent platforms:

1. Targeted Content Acquisition. Prioritise **high-quality, underrepresented niches**—such as Japanese documentaries and international dramas—to differentiate on quality and build a prestige brand image.

2. Genre Investment Strategy. Allocate production and licensing budgets towards **high-performing genres** (e.g., crime, thriller, drama). Avoid overserved categories like generic comedy unless uniquely positioned.

3. Runtime Standardisation. Anchor commissioned or licensed content within the 90–120 minute range to enhance user engagement and completion rates.

Conclusion

A differentiated, quality-first strategy that leverages genre insights, runtime discipline, and curated international content provides a clear pathway for new entrants to

gain market share and investor confidence—without replicating the scale-driven pitfalls observed in Netflix's recent trajectory.

Data analysis

Results

##Discussion

#Question4 Billionaires

Data analysis

We used R (Version 4.4.3; R Core Team, 2025) and the R-packages *dplyr* (Version 1.1.4; Wickham, François, Henry, Müller, & Vaughan, 2023), *fastDummies* (Version 1.7.5; Kaplan, 2025), *forcats* (Version 1.0.0; Wickham, 2023a), *ggplot2* (Version 3.5.2; Wickham, 2016), *ggrepel* (Version 0.9.6; Slowikowski, 2024), *lubridate* (Version 1.9.4; Grolemund & Wickham, 2011), *papaja* (Version 0.1.3; Aust & Barth, 2024), *patchwork* (Version 1.3.0; Pedersen, 2024), *purrr* (Version 1.0.4; Wickham & Henry, 2025), *readr* (Version 2.1.5; Wickham, Hester, & Bryan, 2024), *stringr* (Version 1.5.1; Wickham, 2023b), *tibble* (Version 3.2.1; Müller & Wickham, 2023), *tidyr* (Version 1.3.1; Wickham, Vaughan, & Girlich, 2024), *tidyverse* (Version 2.0.0; Wickham et al., 2019), and *tinylabels* (Version 0.2.5; Barth, 2025) for all our analyses. ##Results

##Discussion

#Question5 Health

Data analysis

We used R (Version 4.4.3; R Core Team, 2025) and the R-packages *dplyr* (Version 1.1.4; Wickham et al., 2023), *fastDummies* (Version 1.7.5; Kaplan, 2025), *forcats* (Version

1.0.0; Wickham, 2023a), *ggplot2* (Version 3.5.2; Wickham, 2016), *ggrepel* (Version 0.9.6; Slowikowski, 2024), *lubridate* (Version 1.9.4; Grolemund & Wickham, 2011), *papaja* (Version 0.1.3; Aust & Barth, 2024), *patchwork* (Version 1.3.0; Pedersen, 2024), *purrr* (Version 1.0.4; Wickham & Henry, 2025), *readr* (Version 2.1.5; Wickham, Hester, et al., 2024), *stringr* (Version 1.5.1; Wickham, 2023b), *tibble* (Version 3.2.1; Müller & Wickham, 2023), *tidyr* (Version 1.3.1; Wickham, Vaughan, et al., 2024), *tidyverse* (Version 2.0.0; Wickham et al., 2019), and *tinylabels* (Version 0.2.5; Barth, 2025) for all our analyses.

##Results

##Discussion

References

Aust, F., & Barth, M. (2024). *papaja: Prepare reproducible APA journal articles with R Markdown*. https://doi.org/10.32614/CRAN.package.papaja

Barth, M. (2025). *tinylabels: Lightweight variable labels*. https://doi.org/10.32614/CRAN.package.tinylabels

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. Retrieved from https://www.jstatsoft.org/v40/i03/

Kaplan, J. (2025). *fastDummies: Fast creation of dummy (binary) columns and rows from categorical variables*. Retrieved from https://CRAN.R-project.org/package=fastDummies

Müller, K., & Wickham, H. (2023). *Tibble: Simple data frames*. Retrieved from https://CRAN.R-project.org/package=tibble

Pedersen, T. L. (2024). *Patchwork: The composer of plots*. Retrieved from https://CRAN.R-project.org/package=patchwork

R Core Team. (2025). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Slowikowski, K. (2024). *Ggrepel: Automatically position non-overlapping text labels with 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=ggrepel

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2023a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2023b). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ...

Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2025). *Purrr: Functional programming tools*. Retrieved from https://CRAN.R-project.org/package=purrr

Wickham, H., Hester, J., & Bryan, J. (2024). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr

Wickham, H., Vaughan, D., & Girlich, M. (2024). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr