Netflix Movies and TV Shows Analysis (Executive Summary)

Project title: Netflix Movies and TV Shows Analysis (Using PostgreSQL)

Analyst / Role: Data Analyst (SQL-focused)

Project overview

This project cleans, transforms, and analyzes a Netflix catalogue dataset to produce actionable business insights about content composition, creator and talent distribution, genre coverage, regional strengths (with emphasis on India), temporal trends, and simple content-safety signals. The goal is to equip product, programming, and marketing stakeholders with clear, SQL-powered answers that support content acquisition, promotion prioritization, metadata enrichment, and regional investment decisions.

Data & approach

- **Primary dataset:** netflix_titles.csv — metadata for Netflix content (show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, description).
- **Data engineering:** Created a canonical table Tb_netflix. Key cleaning steps: normalized date_added, parsed duration to numeric minutes, trimmed/standardized text fields, removed malformed rows, and expanded multi-valued fields (country, listed_in) using string-splitting + unnesting to enable accurate aggregations.
- **Techniques used:** DDL for table creation, data cleaning (REPLACE/TRIM/DELETE), date parsing, string splitting UNNEST, case-insensitive pattern matching (ILIKE), aggregations (GROUP BY), ranking (ORDER BY ... LIMIT), conditional CASE classification for content-safety, and windowing where appropriate.
- **Deliverables:** Clean Tb_netflix and a suite of reproducible SQL queries answering 15 business-focused questions.

Key findings

1. **Catalogue composition:** A clear split between *Movies* and *TV Shows* that informs content scheduling and homepage prioritization.
2. **Content guidance mix:** The most common ratings per content type were identified — useful for age-targeted marketing and parental controls.
3. **Timely releases:** All movies released in 2020 were enumerated — helpful for curated seasonal collections or retrospectives.
4. **Regional volume leaders:** Top 5 countries by content volume were identified, highlighting which markets contribute the most catalogue items.

5. **Outlier durations:** The single longest movie was found — useful for spotlight features or "epic" programming tags.
6. **Recently added content:** Items added within the last five years were isolated for freshness-driven promotions.
7. **Director-level insight:** All titles by **Rajiv Chilaka** were listed, enabling director-focused campaigns.
8. **Binge-worthy shows:** TV shows with more than 5 seasons were flagged — prime candidates for binge playlists.
9. **Genre distribution:** Counts per genre (by splitting listed_in) reveal genre concentrations and gaps for content acquisition strategy.
10. **India release trends:** Per-year averages of content releases in India were computed; top 5 years with the highest averages were identified — guiding regional investment and release timing.
11. **Documentaries:** All documentary movies were listed — useful for non-fiction bundles and documentary channels.
12. **Metadata gaps:** Content records missing a director were flagged — priority targets for metadata enrichment.
13. **Actor-specific counts:** How many movies **Salman Khan** appeared in over the last 10 years was calculated — useful for talent-focused licensing choices.
14. **Top Indian actors:** Top 10 actors by number of movies produced in India were identified — supports influencer and cross-promotion strategies.
15. **Content-safety taxonomy:** Content classified as **Bad** (contains keywords kill or violence in description) vs **Good**, with counts for each — a quick content-safety filter for moderation/labeling.

Business impact

- **Programmatic curation:** Build data-driven playlists (e.g., “Binge Classics”, “India: Top Years”, “Documentary Spotlight”) using the lists produced.
- **Acquisition prioritization:** Identify genres/countries under-represented in the catalogue for targeted licensing.
- **Marketing & personalization:** Use rating and release-year signals to refine campaigns for different user cohorts (age, regional preference).
- **Metadata quality uplift:** Target records missing directors for enrichment to improve search/recommendation accuracy.
- **Talent & partnership strategy:** Quantify which actors/directors produce the most regional content to inform partnership and promotion deals.

- **Safety & compliance:** Quick flagging of potentially violent content to feed tagging and parental-control paths.

Key recommendations

1. **Enrich missing metadata** (directors/cast/countries) to improve search and recommendation signals — prioritize high-visibility titles first.
2. **Leverage high-season years** (top India release years) for region-focused campaigns and retrospectives.
3. **Create genre gap analyses** (next step) to identify acquisition opportunities by comparing catalogue distribution to viewer demand.
4. **Refine content-safety logic** beyond keyword matching — incorporate NLP-based sentiment/violence classifiers for higher precision.
5. **Operationalize queries** into scheduled views or dashboards (Power BI / Looker) for continuous monitoring and stakeholder access.

Technical notes & reproducibility

- All SQL queries are written against the curated Tb_netflix table and use standard SQL constructs compatible with most RDBMSs (Postgres-style STRING_TO_ARRAY/UNNEST used for multi-valued fields — minor syntax tweaks may be required for other engines).
- Reusable SQL patterns included: date normalization, duration extraction, multi-value unnesting, ILIKE-based keyword searches, and CASE-based classification.
- The project is reproducible: the raw CSV → cleaning SQL → analytical queries form a repeatable pipeline that can be scheduled or scaled.