# Netflix Movies and TV Shows Analysis (Problem Statement)

**Background**

I worked with the **netflix_titles.csv** dataset (~**8,809** content records plus header) containing Netflix catalogue metadata: show_id, type (Movie/TV Show), title, director, cast(s), country, date_added, release_year, rating, duration, listed_in (genres), and description.

I created a relational table **Tb_netflix**, performed data-cleaning (removed rows with misplaced values, trimmed text, checked duplicates), converted and parsed fields for analysis (e.g., parsed duration numeric minutes, converted date_added to dates, split multi-valued country and listed_in using STRING_TO_ARRAY + UNNEST), and fixed or identified NULL/missing director entries.

Key SQL techniques used (from your files):

- DDL: CREATE TABLE Tb_netflix
- Data cleaning: DELETE bad rows, TRIM/REPLACE, simple validation checks
- Date parsing: TO_DATE(date_added, 'Month DD, YYYY')
- Text parsing / multi-valued fields: STRING_TO_ARRAY(..., ',') + CROSS JOIN LATERAL UNNEST(...)
- Type casting: CAST(REPLACE(duration, ' min', '') AS INTEGER)
- Pattern matching and case-insensitive searches: ILIKE '%...%'
- Aggregations and windowing: GROUP BY, COUNT, ORDER BY, EXTRACT(YEAR FROM ...)
- Conditional categorization: CASE on description for content-safety labels

**Role**

**Data Analyst (SQL-focused)** — responsible for end-to-end exploratory data analysis and preparatory data engineering on a single-source streaming catalogue to produce business-friendly insights for content strategy, catalogue curation, and regional marketing decisions.

**Objective / Problem Statement**

Design and implement an SQL-driven analysis of the Netflix catalogue to answer business questions that inform content strategy and regional programming decisions. The project objective is:

**Using Tb_netflix as the canonical table, clean and transform the dataset, then produce actionable insights that describe the catalogue composition, content safety signals, regional production strength (with emphasis on India), creative contributors (directors/actors), and temporal trends in releases — enabling product, programming, and marketing teams to prioritize content acquisition, promotion, and regional investment.**

Concretely, the project solves these business needs (as implemented in your SQL file):

- Measure catalogue composition: count Movies vs TV Shows.

- Understand audience guidance mix: most common ratings per content type.
- Find timely/targeted content: list movies released in 2020; content added in last 5 years.
- Identify regional strengths: top 5 countries by content volume; top actors in Indian productions; per-year content release trends in India (top years by average releases).
- Content duration & scheduling: identify the longest movie.
- Creator & genre curation: list content by a given director; list TV shows with >5 seasons; list documentary movies.
- Data quality checks: identify content records missing director metadata.
- Talent analytics: count appearances of a named actor (e.g., Salman Khan) over a timeframe; list top 10 actors by number of Indian films.
- Genre and content distribution: count items per genre (by splitting listed_in).
- Content-safety flagging: categorize items as **Bad** or **Good** based on presence of keywords (kill, violence) in description, and report counts by category.

**Deliverables (what this SQL project produces):**

- A clean Tb_netflix table ready for analytics.
- SQL queries answering the 15 business questions above (aggregation tables, filtered lists, ranked results).
- Reusable SQL patterns for parsing multi-valued text fields (country, listed_in), date normalization, and content classification that can be extended for further analyses (e.g., sentiment on descriptions, deeper keyword taxonomies, or time-series of content additions).

**Business value:** The outputs enable stakeholders to:

- Prioritize content licensing and marketing by country/genre.
- Identify gaps in metadata (missing directors) for catalog enrichment.
- Track talent-centric investment (which actors/directors drive volume).
- Apply quick content-safety filters for age-appropriate catalog curation.
- Make India-focused decisions using per-year release trends and top actors.