Netflix Movies & TV Shows End-to-End Big Data Analytics

Python | SQL | BI Dashboard | Business Insights

PROJECT OVERVIEW

Transforming Raw Data into Strategic Insights

Project Objective

Comprehensive analysis of Netflix's global content catalogue using modern data analytics tools and methodologies to uncover actionable insights for content strategy optimisation.

Key Technologies

- Python for data preprocessing and feature engineering
- SQL for analytical querying and aggregations
- Power BI/Lovable for interactive visualisation

Expected Outcomes

Data-driven recommendations for content acquisition, regional expansion strategies, and audience targeting based on rigorous statistical analysis and trend identification.

DATASET

Dataset Overview & Structure

8.8K

Total Records

Comprehensive catalogue of titles

12

Key Attributes

Rich metadata fields

2

Content Types

Movies and TV Shows

Core Data Attributes

Content Classification

Content Type (Movie/TV Show), Genre, Rating

Temporal Data

Date Added, Release Year, Duration

Geographical & Production

Country of Origin, Director, Cast Members

Initial Data Quality Challenges

Raw datasets in real-world scenarios rarely arrive analysis-ready. The Netflix dataset presented several data quality issues requiring systematic preprocessing before meaningful analysis could commence.

Date Format Issues

Date fields stored as unstructured text strings requiring parsing and conversion to datetime objects

Missing Values

Significant gaps in director, cast, and country fields necessitating imputation strategies

Text Inconsistencies

Irregular formatting, capitalisation, and whitespace issues across multiple text columns

Duration Parsing

Mixed format duration values combining numeric and text (e.g., "90 min", "2 Seasons")

Raw data quality assessment revealed that systematic preprocessing was essential before conducting any meaningful analytical queries or visualisations.

DATA ENGINEERING

Data Cleaning & Feature Engineering

Systematic Preprocessing Pipeline

Data Ingestion

Load raw CSV dataset using pandas

Data Cleaning

Remove duplicates, standardise formatting, handle nulls

Feature Engineering

Extract temporal features, parse duration fields

Validation & Export

Quality checks and export clean dataset

Key Transformations Implemented

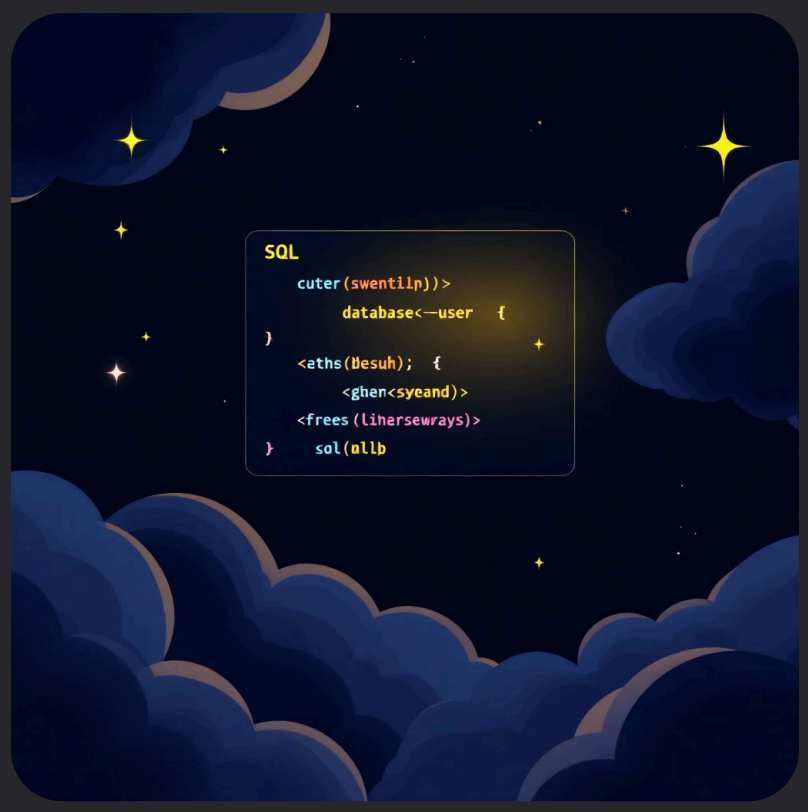
- Converted `date_added` from text to datetime format
- Created derived features: `year_added` and `month_added`
- Parsed duration into separate numeric value and unit columns
- Standardised text fields with consistent capitalisation
- Applied domain-specific imputation for missing values
- Generated clean dataset: netflix_titles_clean.csv

SQL ANALYTICS

SQL-Based Analytical Framework

Structured Query Analysis

Leveraged SQL for complex aggregations, temporal analysis, and cross-dimensional queries to extract meaningful patterns from the cleaned dataset.



Key Analytical Questions

01

Content Type Distribution

Movies versus TV Shows composition analysis

02

Temporal Growth Patterns

Year-over-year content addition trends

03

Geographic Analysis

Top content-producing countries and regions

04

Content Classification

Rating categories and genre distribution

05

Production Insights

Most prolific directors and production patterns

06

Seasonal Trends

Monthly and quarterly release patterns

Interactive Business Intelligence Dashboard

Comprehensive Visual Analytics Platform

KPI Summary Cards

Total Titles, Movies Count, TV Shows Count, Percentage Split by Content Type

Content Composition

Visual breakdown of Movies vs TV Shows distribution

Growth Trajectory

Year-over-year content addition trends and acceleration

Geographic Distribution

Country-wise content production and availability analysis

Ratings & Genres

Content rating categories and genre popularity metrics

Seasonality Patterns

Monthly and quarterly release timing analysis

KEY INSIGHTS

Content Mix & Growth Dynamics

Rapid Catalogue Expansion

Netflix's content library experienced exponential growth in recent years, with particularly aggressive acquisition and production strategies post-2015, reflecting the platform's shift towards content abundance as a competitive differentiator.

Movies Dominate Volume

While movies constitute the majority of total titles by count, the growth rate of TV show additions has been accelerating significantly, indicating a strategic pivot towards episodic content formats.

Binge-Worthy Strategy

The increasing proportion of TV shows reflects evolving audience consumption patterns favouring serialised storytelling and the platform's distinctive binge-watching culture, supported by original series investments.

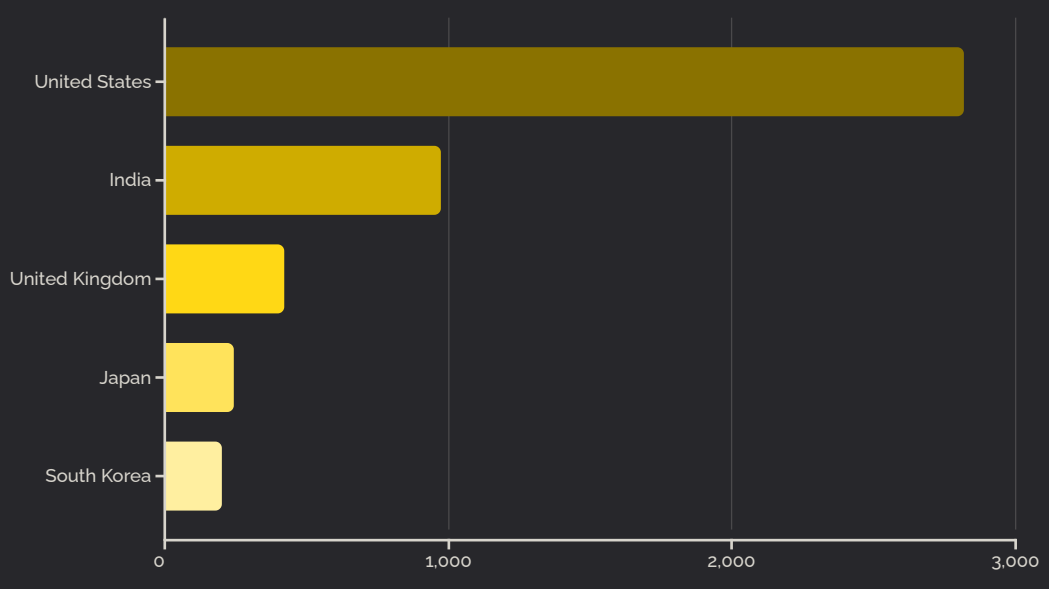
Strategic Implication

These trends suggest Netflix's recognition that episodic content drives sustained engagement and subscription retention more effectively than standalone films, informing future content investment priorities.

Catalogue Composition Analysis

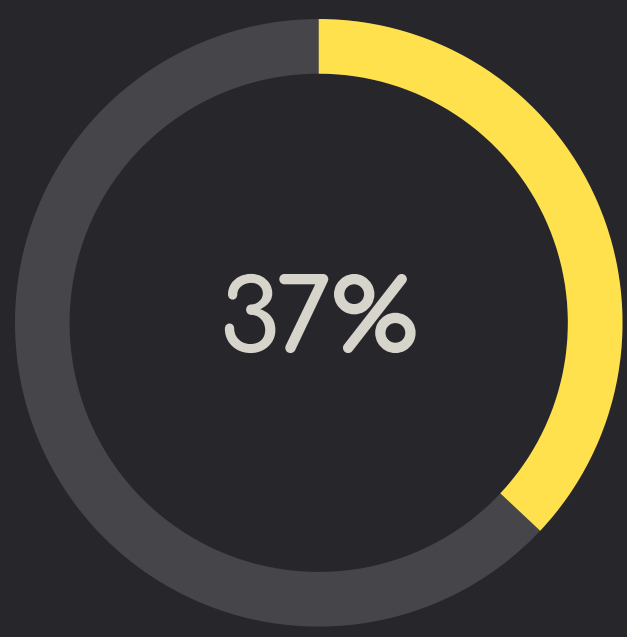
Geographic Concentration

Content production is heavily concentrated among a small number of countries, with the United States, India, and United Kingdom dominating the catalogue.



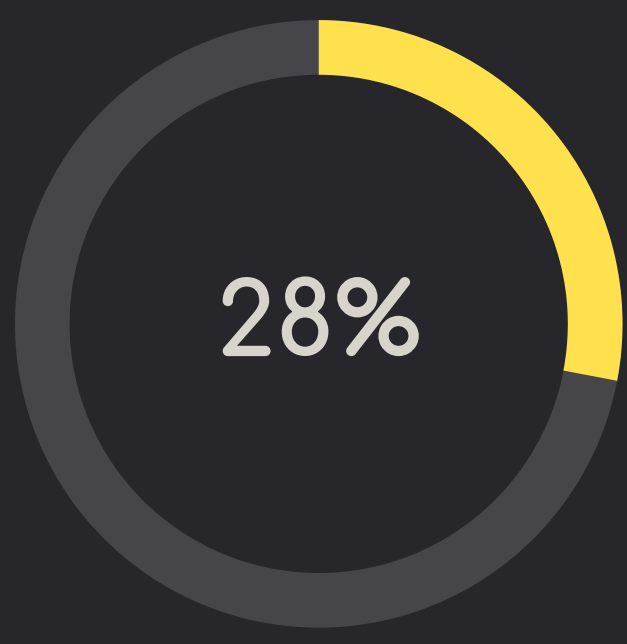
Geographic dependency creates supply chain risks and limits regional appeal in underserved markets.

Content Rating Distribution



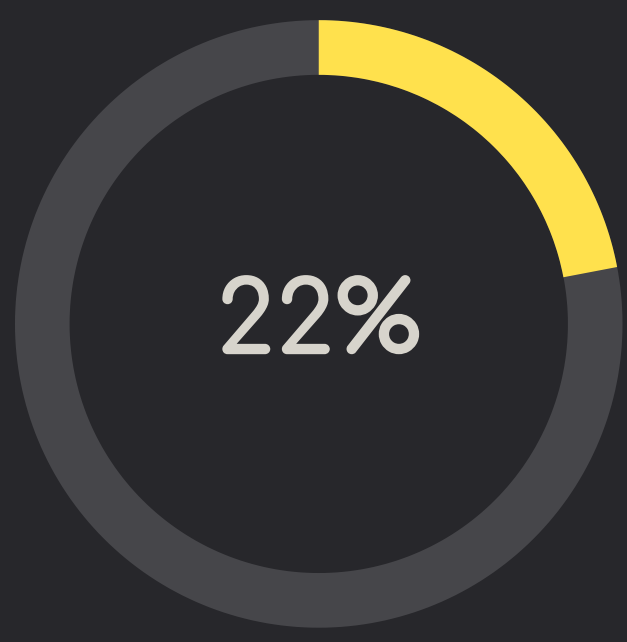
TV-MA

Mature audiences content dominates



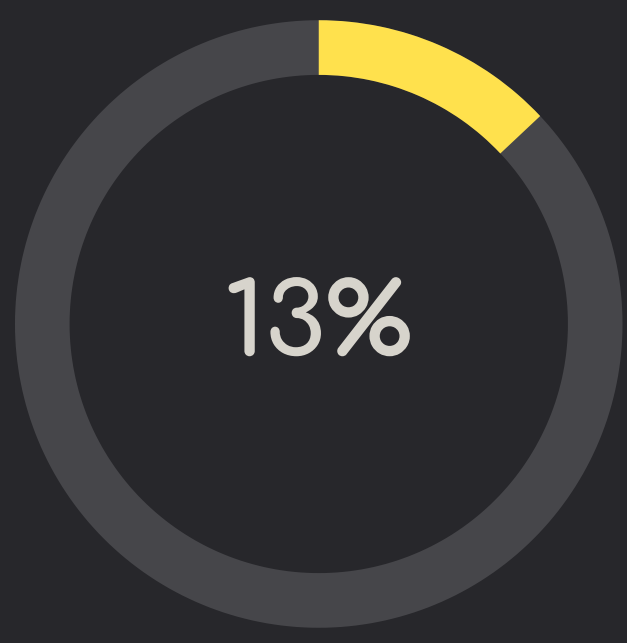
TV-14

Teen-appropriate programming



R-Rated

Restricted movie content



Family-Friendly

TV-G, TV-Y, PG content

Genre Concentration

A small number of genres—Drama, Comedy, Documentary, and Action—account for the vast majority of titles, revealing significant opportunities for diversification into underrepresented categories.

RECOMMENDATIONS

Strategic Business Recommendations

Data-Driven Content Strategy

Accelerate TV Show Investment

Increase production and acquisition of original TV series to capitalise on superior engagement and retention metrics compared to standalone films.

Diversify Regional Sourcing

Expand content partnerships beyond heavily represented markets to reduce geographic concentration risks and appeal to underserved international audiences.

Balance Content Ratings

Invest in family-friendly and children's programming to address the current skew towards mature content and capture broader household viewership.

Explore Niche Genres

Develop content in underrepresented genres such as science fiction, thriller, and fantasy to differentiate offerings and attract specialised audience segments.

Optimise Release Timing

Align high-value content releases with identified seasonal demand patterns to maximise initial viewership and subscription conversion opportunities.