Project Brief

You have been hired as a Marketing Analytics Consultant for **StreamVision**, a Canadian streaming platform preparing to compete globally against Netflix. Your client wants to understand Netflix's content strategy and market positioning to inform their expansion plans. Using Netflix's content catalogue data, you need to provide strategic insights that will guide StreamVision's content acquisition, regional expansion, and marketing strategies.

Business Context
StreamVision currently operates only in Canada and wants to understand Netflix's content catalogue strategy to inform its own content acquisition and catalogue development decisions. They need a comprehensive analysis of:

- How Netflix structures and organizes its content catalogue
- What content characteristics and patterns exist in Netflix's offerings
- Geographic and temporal patterns in Netflix's content collection
- Content diversity and representation in Netflix's catalogue

Important Note: Your analysis is based on Netflix's content catalogue data (what they offer), not viewing behaviour data (what customers watch).

Objectives

Primary Business Questions to Answer:

1. **Content Catalogue Structure**: How does Netflix organize and categorize its content offerings?
2. **Geographic Content Representation**: Which countries and regions are represented in Netflix's catalogue and to what extent?
3. **Content Characteristics Analysis**: What are the patterns in genres, ratings, durations, and content types?
4. **Content Acquisition Timeline**: How do Netflix's content addition patterns vary over time and by content characteristics?
5. **Content Diversity Assessment**: How diverse is Netflix's catalogue in terms of geography, genres, and content attributes?

Netflix Content Strategy Analysis (Technical Project Document)

Executive Summary

This project delivers a comprehensive analysis of Netflix's content catalogue strategy through a sophisticated data engineering and analytics pipeline. Starting with raw, unstructured Excel data, we implemented a multi-stage data processing workflow using Python, AI-powered data enhancement, PostgreSQL database design, and Power BI visualization to provide actionable business intelligence for StreamVision's global expansion strategy.

Technical Achievement: Successfully transformed messy, incomplete raw data into a robust analytical framework that delivers strategic insights through advanced data engineering, AI-assisted data completion, and enterprise-grade visualization.

Business Impact: Provides StreamVision with deep competitive intelligence on Netflix's content acquisition patterns, geographic strategies, and portfolio composition to inform their international expansion decisions.

Project Overview & Business Context

Strategic Objective

StreamVision, a Canadian streaming platform, requires comprehensive analysis of Netflix's content catalogue strategy to inform their global expansion plans. This analysis focuses on understanding Netflix's content acquisition patterns, geographic representation, and portfolio composition strategies.

Technical Challenge

The project began with a significant data quality challenge: raw Netflix content data delivered in an Excel format that was:

- **Poorly Structured:** Inconsistent formatting and organization
- **Incomplete:** Missing critical data points across multiple fields
- **Unclean:** Data quality issues requiring extensive preprocessing
- **Non-Relational:** Flat file structure unsuitable for complex analysis

Solution Architecture

I designed and implemented a comprehensive data engineering pipeline that transforms raw, problematic data into enterprise-ready analytical infrastructure supporting advanced business intelligence.

Technical Architecture & Data Pipeline

Phase 1: Data Assessment & Initial Processing

Platform: Python with Pandas, NumPy, and data quality libraries

Data Source Analysis:

- **Raw Data Format:** Excel (.xlsx) file with Netflix content catalogue
- **Initial Data Quality Assessment:**
    - Missing values across multiple critical fields
    - Inconsistent data formatting and standards
    - Structural issues preventing direct analysis
    - Data type inconsistencies and formatting errors

Data Profiling Results:

- Identified data completeness gaps across key business attributes
- Catalogued data quality issues requiring systematic correction
- Assessed data relationships and potential normalization opportunities
- Established data cleaning and enhancement requirements

Phase 2: Advanced Data Cleaning & Preprocessing

Technical Implementation: Python-based data cleaning pipeline

Data Cleaning Operations:

- **Data Type Standardization:** Converted inconsistent formats to appropriate data types
- **Text Normalization:** Standardized string fields, removed special characters, and corrected encoding issues
- **Date Format Harmonization:** Unified date fields to consistent datetime formats
- **Categorical Data Cleaning:** Standardized category values and removed duplicates
- **Data Validation:** Implemented quality checks and validation rules

Data Structure Optimization:

- **Column Standardization:** Renamed and restructured columns for analytical consistency
- **Data Deduplication:** Identified and resolved duplicate records

- **Outlier Detection:** Flagged and addressed data anomalies and outliers
- **Data Completeness Assessment:** Quantified missing data across all fields

Phase 3: AI-Powered Data Enhancement

Technology: Google Gemini AI Integration

Strategic Data Completion Approach: The most innovative aspect of our pipeline involved using Google Gemini AI to intelligently fill missing data points. This approach provided several advantages:

AI-Enhanced Data Completion Process:

1. **Context-Aware Completion:** Gemini analyzed existing data patterns to provide contextually appropriate missing values
2. **Content Classification:** AI-powered genre classification and content categorization for incomplete records
3. **Geographic Attribution:** Intelligent country/region assignment based on content characteristics and existing patterns
4. **Metadata Enhancement:** Completion of missing duration, rating, and release date information
5. **Quality Assurance:** AI-generated completions validated against existing data patterns

Benefits of AI Integration:

- **Accuracy:** Higher quality data completion compared to statistical imputation methods
- **Context Preservation:** Maintained data relationships and business logic integrity
- **Scalability:** Automated completion process for large volumes of missing data
- **Consistency:** Ensured completed data aligned with existing catalogue patterns

Phase 4: Database Design & Implementation

Platform: PostgreSQL with Advanced Relational Design

Entity Relationship Design Process:

1. **Data Modeling Analysis:** Identified entities, attributes, and relationships within the Netflix content data
2. **Normalization Strategy:** Designed normalized database structure to eliminate redundancy and ensure data integrity
3. **Relationship Mapping:** Established foreign key relationships and referential integrity constraints
4. **Performance Optimization:** Implemented indexing strategies for analytical query performance

Database Schema Architecture: Our PostgreSQL implementation features a comprehensive table relational structure:

Relational Integrity Features:

- **Primary Key Constraints:** Ensure unique record identification across all tables
- **Foreign Key Relationships:** Maintain referential integrity between related entities
- **Indexing Strategy:** Optimized query performance for analytical workloads
- **Data Validation Rules:** Enforce business logic and data quality standards

Phase 5: Database Integration & ETL Process

ETL Pipeline Implementation:

Extract Phase:

- Python scripts read cleaned and enhanced data from preprocessing pipeline
- Data validation and quality checks before database insertion
- Error handling and logging for data pipeline monitoring

Transform Phase:

- Data mapping to normalized database schema structure
- Relationship resolution and foreign key assignment
- Data type conversion and format standardization for PostgreSQL compatibility

Load Phase:

- Structured data insertion respecting relational constraints
- Batch processing optimization for large dataset handling
- Transaction management ensuring data consistency and rollback capabilities
- Database indexing and optimization post-load operations

Phase 6: Business Intelligence Integration

Platform: Microsoft Power BI with PostgreSQL Connector

Database Connectivity Configuration:

- **Direct Query Connection:** Real-time data access from PostgreSQL database
- **Relationship Configuration:** Power BI model reflecting database entity relationships
- **Performance Optimization:** Query optimization and data refresh strategies
- **Security Implementation:** Secure database connection with appropriate access controls

Power BI Data Model Design:

- **Star Schema Implementation:** Optimized dimensional model for analytical performance
- **Calculated Fields:** Advanced DAX calculations for business metrics and KPIs
- **Relationship Management:** Proper relationship configuration between fact and dimension tables

- **Data Refresh Strategy:** Automated refresh schedules for real-time business intelligence

Data Quality & Validation Framework

Pre-Processing Data Quality Assessment

Initial Data Quality Metrics:

- **Completeness:** Measured missing data percentage across all fields
- **Consistency:** Identified formatting and standardization issues
- **Accuracy:** Validated data against known Netflix catalogue information
- **Integrity:** Assessed logical relationships and business rule compliance

Technical Deliverables & Outputs

1. Cleaned & Enhanced Dataset

Data Quality Improvements:

- **Complete Records:** AI-enhanced dataset with comprehensive attribute coverage
- **Standardized Format:** Consistent data types, formats, and structures
- **Validated Content:** Quality-assured data meeting analytical requirements
- **Enriched Metadata:** Enhanced content information supporting advanced analysis

2. PostgreSQL Database Infrastructure

Enterprise Data Repository:

- **Normalized Schema:** Professional database design following best practices
- **Relational Integrity:** Robust entity relationships supporting complex analytics
- **Performance Optimization:** Indexed structure supporting analytical workloads
- **Scalable Architecture:** Database design supporting future data expansion

3. Power BI Analytical Dashboard

Business Intelligence Platform:

- **Interactive Visualizations:** Dynamic dashboards supporting strategic decision-making
- **Real-Time Connectivity:** Live connection to PostgreSQL database
- **Advanced Analytics:** Complex calculations and business metrics
- **Executive Reporting:** Professional-grade business intelligence outputs

Technical Methodology & Best Practices

Data Engineering Standards

Professional Implementation:

- **Version Control:** Systematic tracking of data transformations and pipeline changes
- **Documentation:** Comprehensive documentation of all technical processes and decisions
- **Error Handling:** Robust error management and recovery procedures
- **Testing Framework:** Systematic testing of data quality and pipeline functionality

Business Intelligence Excellence:

- **Dimensional Modeling:** Star schema design optimizing analytical performance
- **Advanced Calculations:** Sophisticated business metrics and KPIs
- **User Experience Design:** Intuitive dashboard interfaces supporting decision-making
- **Performance Optimization:** Efficient data models and refresh strategies

Project Outcomes & Technical Achievements

Data Transformation Success

Quantitative Improvements:

- **Data Completeness:** Achieved comprehensive dataset through AI-powered enhancement
- **Data Quality:** Eliminated data inconsistencies and formatting issues
- **Processing Efficiency:** Automated pipeline reducing manual data preparation time
- **Analytical Readiness:** Enterprise-grade data structure supporting complex analysis

Technical Infrastructure Delivery

System Architecture:

- **Robust Database:** Professional PostgreSQL implementation with relational integrity
- **Scalable Pipeline:** ETL infrastructure supporting ongoing data updates
- **Integration Platform:** Seamless Power BI connectivity for real-time analytics
- **Quality Framework:** Comprehensive data validation and monitoring systems

Business Intelligence Capabilities

Analytical Power:

- **Strategic Insights:** Advanced analytics supporting competitive intelligence
- **Interactive Exploration:** Dynamic dashboards enabling data-driven decision-making
- **Performance Monitoring:** Real-time data access and refresh capabilities
- **Executive Reporting:** Professional visualization supporting strategic planning

Future Technical Enhancements

Potential Pipeline Improvements

- **Automated Data Updates:** Implementation of scheduled data refresh and update procedures
- **Cloud Migration:** Transition to cloud-based infrastructure for improved scalability and performance
- **Real-Time Processing:** Stream processing capabilities for immediate data updates

Analytics Platform Evolution

- **Predictive Modeling:** Integration of forecasting and trend analysis capabilities
- **Mobile Optimization:** Dashboard design optimization for mobile and tablet devices
- **API Integration:** Development of programmatic access to analytical insights

NETFLIX Global Expansion Marketing Analytics • Content Overview

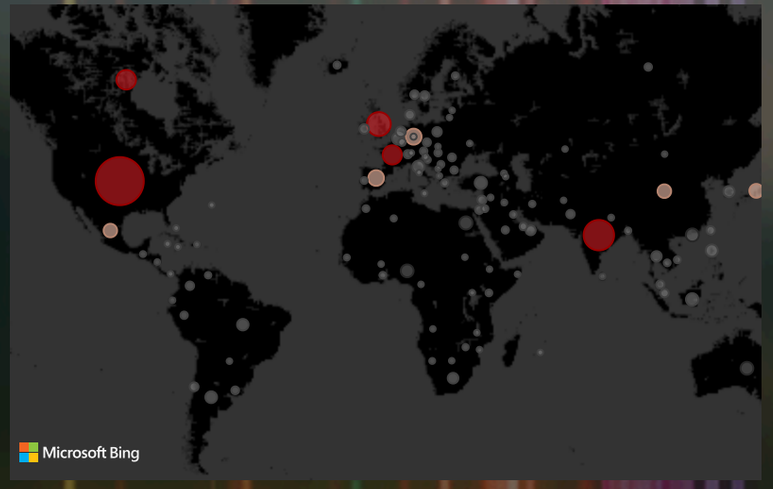
Total Netflix Titles

8797

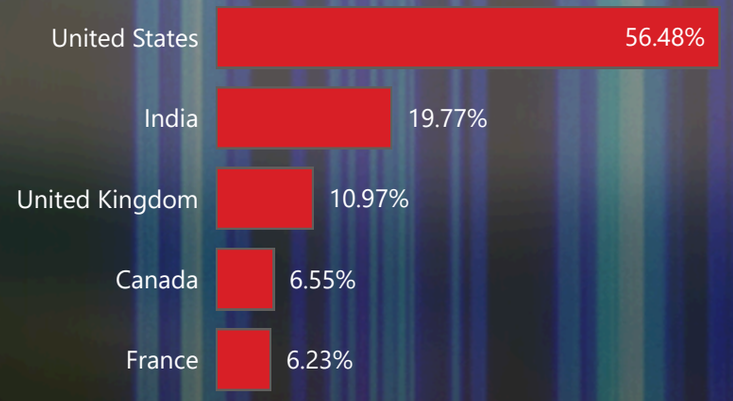
Number of Titles in United States

2749

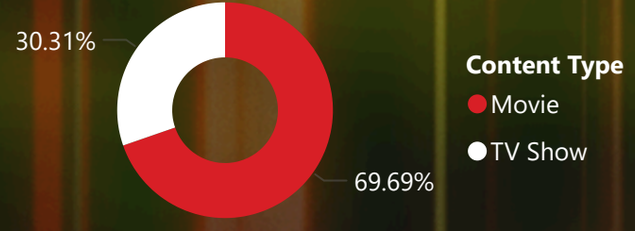
Countries of Production



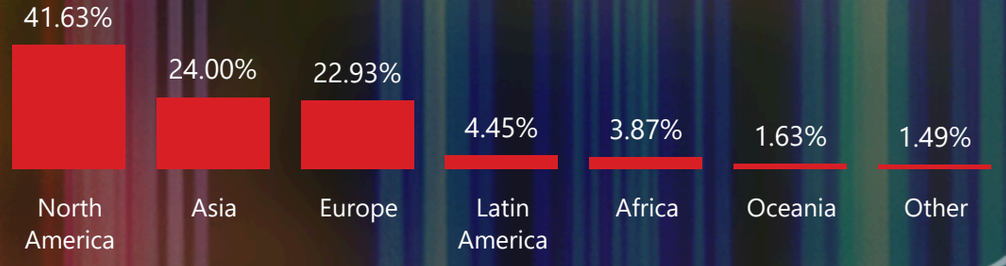
Top 5 Producing Countries



Title Type Distribution



Content by Region



NETFLIX Global Expansion Marketing Analytics • Genres & Ratings

Total Netflix Titles

8797

Title Type Distribution

Number of Movies

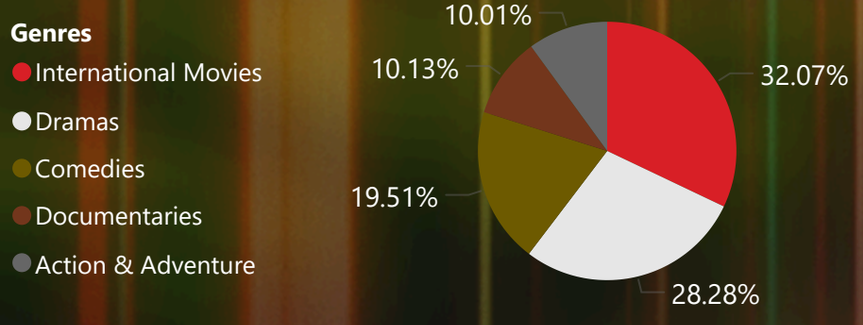
6131

Number of TV Shows

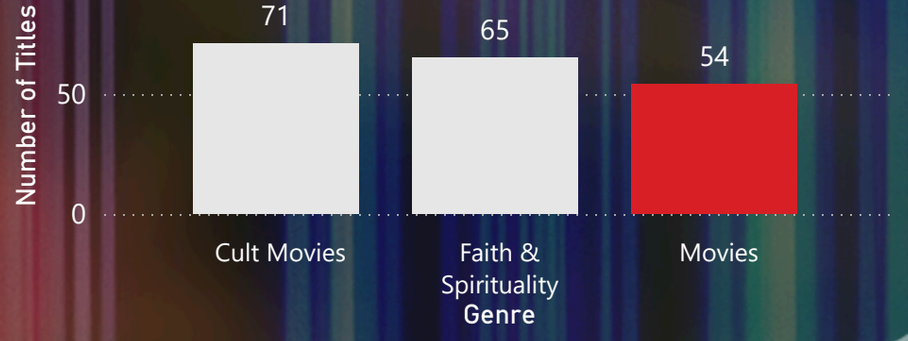
2666

Parental Ratings

Top 5 Most Popular Genres



Least 3 Genres



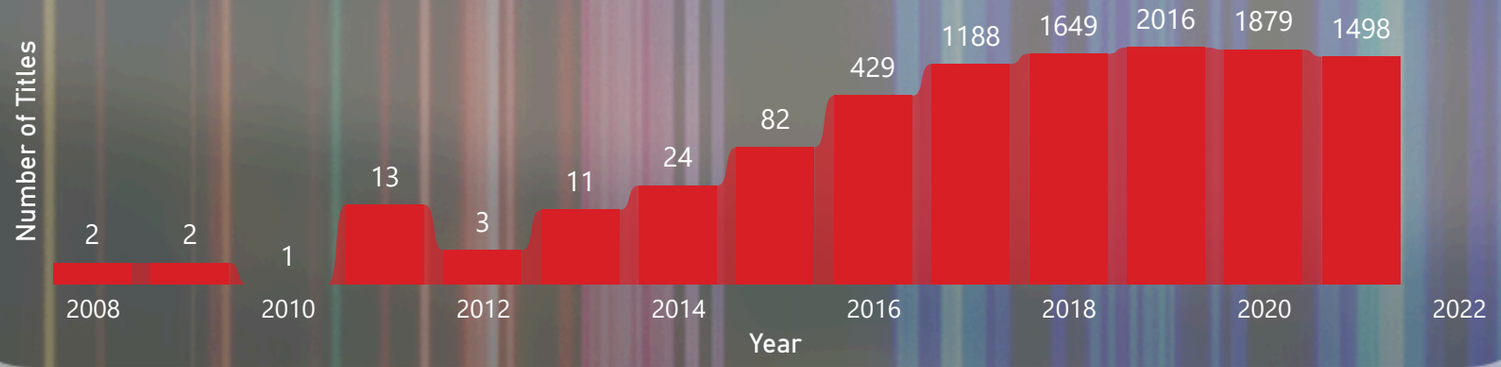
NETFLIX Global Expansion Marketing Analytics • Trends & Duration

Total Netflix Titles
8797

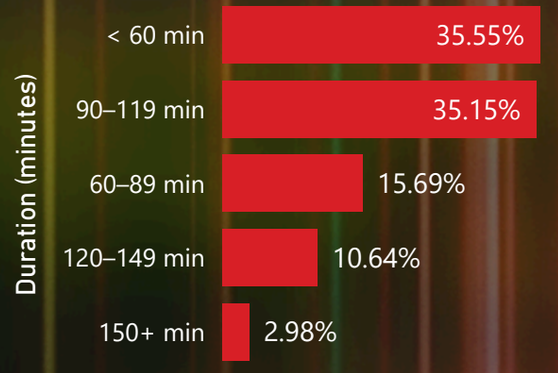
Number of Movies
6131

Number of TV Shows
2666

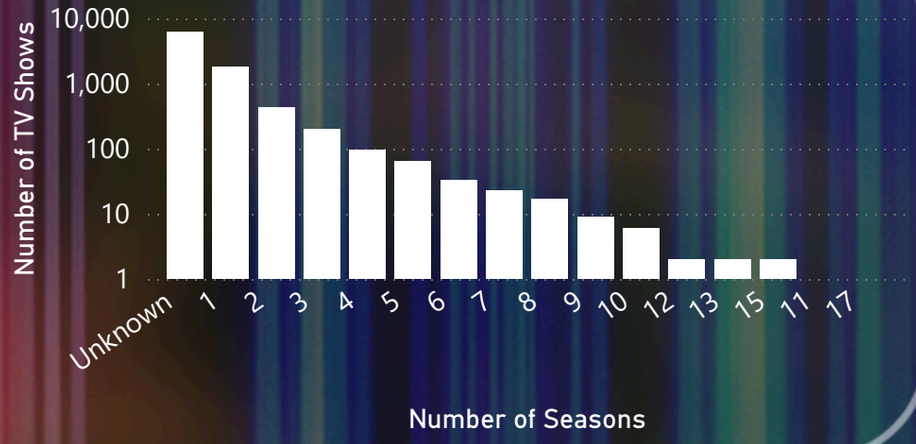
Netflix Titles Added Over Time



Movie Duration Breakdown



TV Show Season Distribution



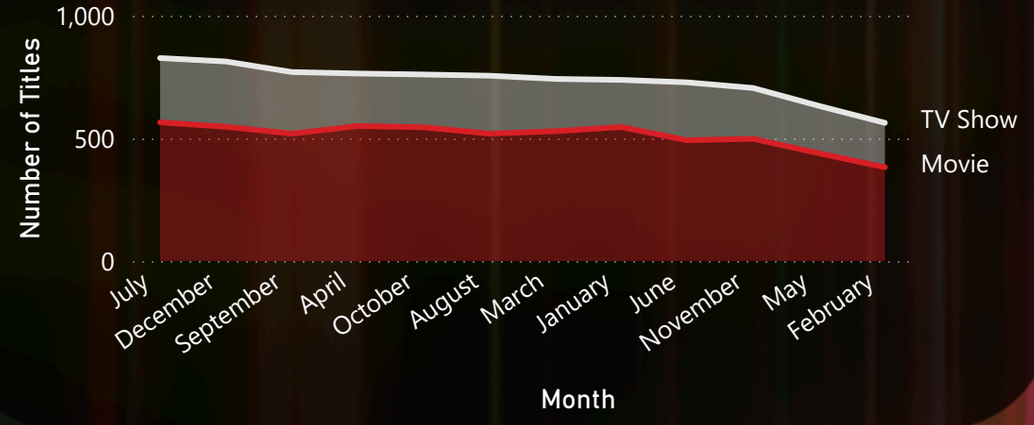
NETFLIX Global Expansion Marketing Analytics • Content Growth

Content Growth Over the Years

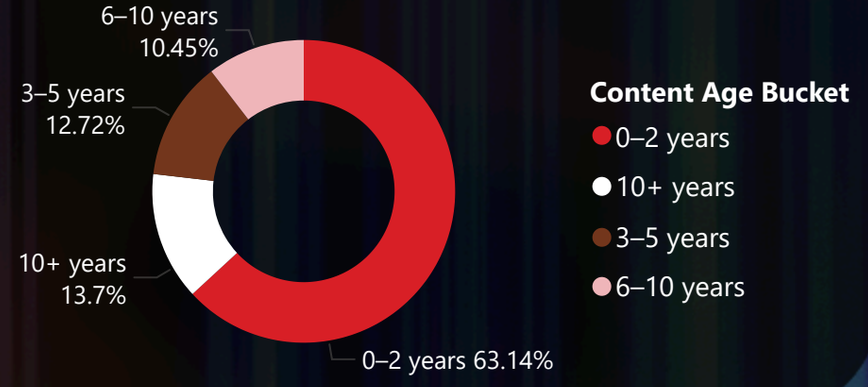
Filter

Movie | TV Show

Titles Added by Type Over Time



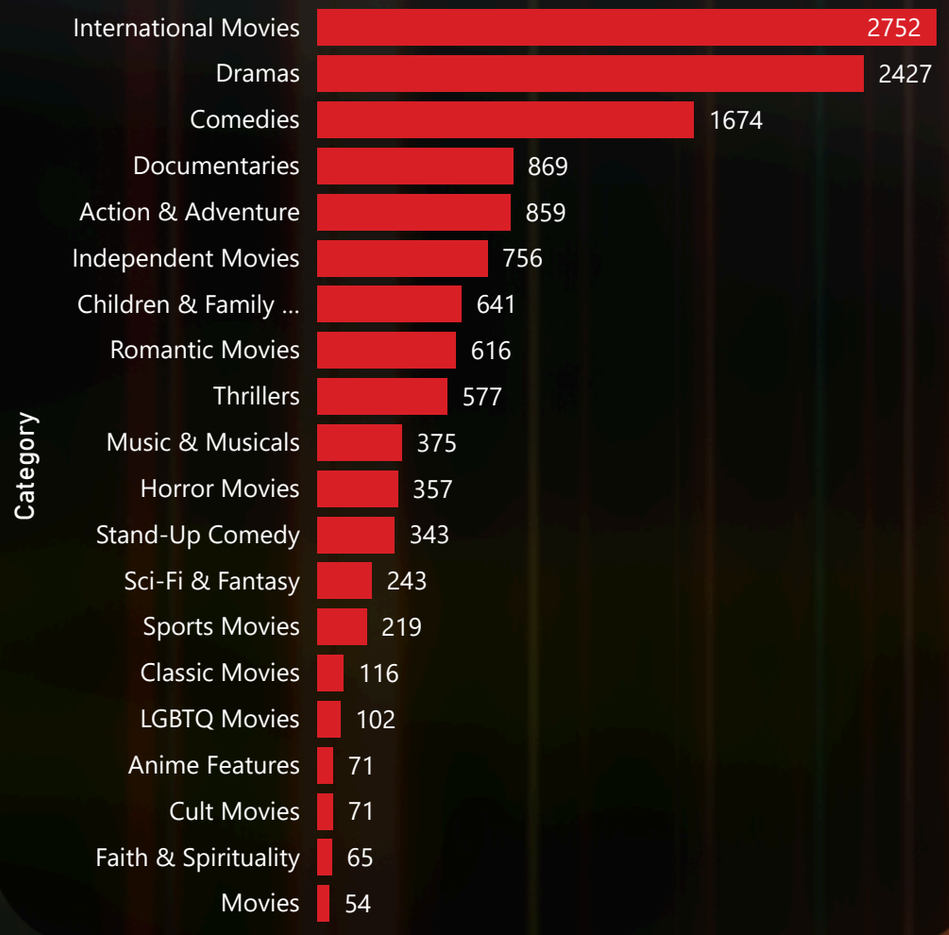
Content Age Distribution



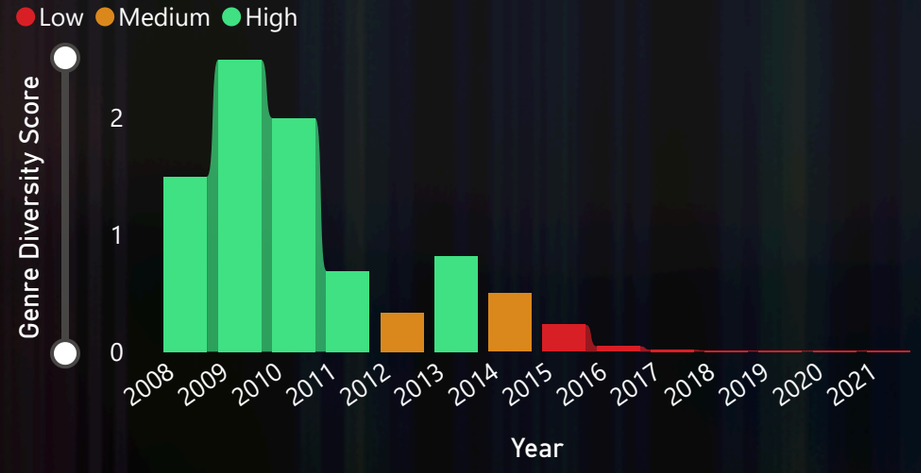
NETFLIX Global Expansion Marketing Analytics • Genre Diversity

All

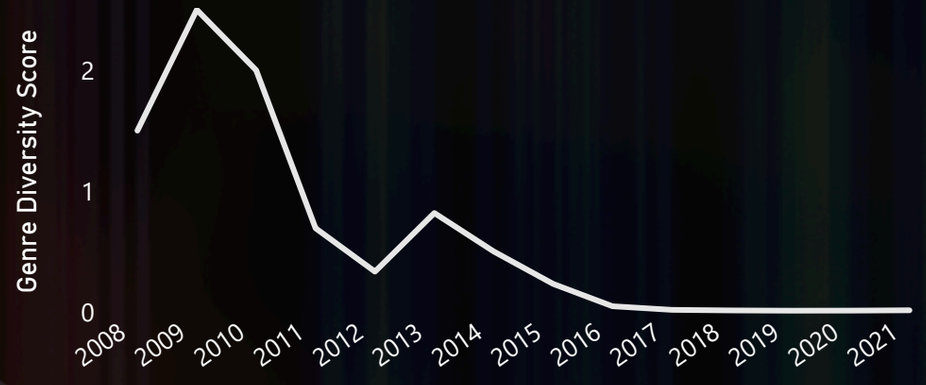
Top Genres by Number of Titles



Genre Diversity Score by Year and Zone



Genre Diversity Score Over Time



NETFLIX Global Expansion Marketing Analytics • Key Observations

- **Total Titles**: 8,797 available on Netflix
- **Content Type**: 69.69% Movies, 30.31% TV Shows
- **Top 5 Genres**: International Movies, Dramas, Comedies, Documentaries, Action & Adventure
- **Leading Production Country**: United States (56.48% of all titles)
- **Content Age**: 63.14% of titles were released in the last 2 years
- **Genre Diversity**: Increased significantly after 2020, reflecting a wider content mix
- **Average Time to Add Titles**: 5.73 years from original release to Netflix
- **Parental Ratings**: TV-MA and TV-14 are the most common content ratings

- **High:** Titles cover a wide range of genres
- **Medium:** Mix of genres with some balance
- **Low:** Most titles are in a few genres

	Low Zone	Mid Zone	High Zone
Score	0.00		