Netflix Data Analysis(EDA)

--Kshitij Thorat

Summary- This document provides a comprehensive analysis of Netflix's content catalog using PySpark on Databricks Serverless platform. The analysis explores content growth patterns, distribution metrics, and geographic contributions to understand Netflix's content strategy and catalog evolution.

Table of Contents

1.Introduction

Purpose

This analysis aims to uncover patterns in Netflix's content acquisition and production strategies by examining temporal trends, content characteristics, and geographic distribution of titles available on the platform.

Methodology

1. Platform: Databricks Serverless with PySpark
2. Approach: Distributed data processing with aggregation and visualization
3. Tools: PySpark SQL, Matplotlib, Seaborn, Pandas

2. Dataset Overview

Data Source

- Location: /Volumes/workspace/default/netflix/cleaned_netflix_titles/
- Format: CSV with header and inferred schema
- Processing: Loaded using PySpark DataFrame API

Data Quality Metrics

- Total records analyzed
- Column completeness assessment
- Schema validation with automatic type inference
- Null value handling for critical fields

3. Analysis Framework

Task 1: Temporal Analysis

Examines content growth patterns over time to understand platform evolution

Task 2: Distribution Metrics

Analyzes content characteristics including type, ratings, and genres

Task 3: Geographic Intelligence

Explores country-level contributions and regional content diversity

4. Task 1: Content Growth Over Time

4.1 Content Growth by Release Year

Objective

Track the volume of content available on Netflix by the year titles were originally released.

Methodology

1. Filter records with valid release year
2. Group by year and count titles
3. Analyze trends from 1950 onwards for meaningful patterns

Key Metrics

1. Time Range: Historical content distribution from 1950 to present
2. Growth Pattern: Identification of acceleration phases
3. Recent Trends: Focus on 2015-2024 period for current strategy

Importance

1. Content Diversity: Shows Netflix's investment in both classic and contemporary content
2. Strategic Positioning: Reveals whether Netflix favors recent or timeless titles
3. Library Depth: Demonstrates catalog breadth across decades

Visualizations

1. Line Plot (1950+): Shows long-term growth trajectory with exponential increase in recent years
2. Bar Chart (2000+): Highlights modern content explosion and year-over-year patterns

Business Insights

1. Exponential growth in recent years indicates aggressive content acquisition
2. Historical content provides catalog depth and nostalgia value
3. Peaks in specific years correlate with major licensing deals or production investments

4.2 Content Added to Netflix Platform by Year

Objective

Analyze when content was added to the Netflix platform (regardless of release year).

Methodology

1. Filter records with valid added_year
2. Calculate year-over-year (YoY) growth rates
3. Identify platform expansion phases

Key Metrics

1. Annual Additions: Volume of titles added each year
2. YoY Growth Rate: Percentage change between consecutive years
3. Platform Activity Timeline: Period of Netflix's catalog building

Importance

1. Acquisition Strategy: Shows Netflix's content investment timeline
2. Platform Growth: Correlates with subscriber growth and market expansion
3. Strategic Shifts: Identifies changes in content addition velocity

Calculation Example

YoY Growth = ((Current Year - Previous Year) / Previous Year) × 100

Business Insights

1. Steep increases indicate market expansion periods
2. Plateaus or declines suggest strategic pivots or market maturity
3. Recent trends inform future content budget allocations.

5. Task 2: Distribution Analysis

5.1 Content Type Distribution (Movies vs TV Shows)

Objective

Determine the split between Movies and TV Shows in Netflix's catalog.

Methodology

1. Group content by type field
2. Calculate counts and percentages
3. Compare relative proportions

Key Metrics

1. **Movie Count & Percentage**: Total movies and their share

2. **TV Show Count & Percentage**: Total series and their share
3. **Content Mix Ratio**: Movies to TV Shows ratio

Importance

1. **Content Strategy**: Reveals whether Netflix is movie-focused or series-focused
2. **Viewer Engagement**: Different content types have varying engagement patterns
3. **Production Investment**: Indicates where Netflix allocates production resources

Visualizations

1. **Bar Chart**: Direct count comparison
2. **Pie Chart**: Percentage distribution for quick understanding

Business Implications

1. Higher movie count suggests one-time viewing content strategy
2. Higher TV show count indicates focus on long-term engagement
3. Balance reflects audience preference and retention strategy.

5.2 Rating Distribution

Objective

Analyze content ratings to understand audience targeting strategy.

Methodology

1. Group content by rating category
2. Rank by frequency
3. Analyze top 15 ratings for clarity

Key Metrics

1. **Rating Categories**: TV-MA, TV-14, TV-PG, PG-13, R, etc.
2. **Distribution**: Count of titles per rating
3. **Top 5 Ratings**: Most common rating classifications

Importance

1. **Audience Segmentation**: Shows which demographic Netflix targets
2. **Content Appropriateness**: Indicates family-friendly vs. mature content mix
3. **Regulatory Compliance**: Different markets have different rating requirements

Ratings Interpretation

1. **TV-MA / R**: Mature audiences (17+)
2. **TV-14 / PG-13**: Teenagers and young adults
3. **TV-PG / PG**: Family-friendly content
4. **TV-Y / G**: Children's content

Business Insights

1. Predominance of mature ratings indicates adult-focused strategy
2. Family ratings show investment in household viewing
3. Distribution informs content acquisition and production decisions

5.3 Genre Distribution

Objective

Identify the most prevalent genres across Netflix's catalog.

Methodology

1. Split listed_in field (comma-separated genres)
2. Explode into individual genre records
3. Trim whitespace and aggregate counts
4. Rank top 20 genres

Key Metrics

1. **Genre Count**: Number of titles per genre
2. **Genre Diversity**: Total unique genres
3. **Top Genres**: Most popular content categories

Importance

1. **Content Positioning**: Shows Netflix's genre specialization
2. **Market Demand**: Reflects audience preferences
3. **Competitive Advantage**: Identifies niche vs. mainstream focus

Common Genre Categories

1. International content (e.g., International Movies/TV Shows)
2. Dramas (various types)
3. Comedies (Stand-up, Romantic, Dark)
4. Documentary & Reality content
5. Children & Family content
6. Thrillers & Action

Business Insights

1. International content popularity indicates global expansion success
2. Drama and comedy dominance reflects universal appeal
3. Niche genres support diverse audience segments
4. Documentary presence shows educational content investment

6. Task 3: Country-Level Analysis

6.1 Top Content-Producing Countries

Objective

Identify which countries contribute the most content to Netflix's global catalog.

Methodology

1. Split country field (comma-separated for co-productions)
2. Explode into individual country records
3. Clean and aggregate counts
4. Rank top 20 countries

Key Metrics

1. **Content Count by Country**: Number of titles from each nation
2. **Geographic Diversity**: Total countries represented
3. **Top 5 Contributors**: Leading content-producing nations

Importance

1. **Global Strategy**: Shows Netflix's international expansion approach
2. **Content Localization**: Indicates regional content investment
3. **Production Partnerships**: Reveals key production markets
4. **Cultural Diversity**: Demonstrates catalog inclusivity

Expected Leaders

1. **United States**: Largest single contributor
2. **India**: Bollywood and regional content
3. **United Kingdom**: English-language productions
4. **South Korea**: K-drama phenomenon
5. **Japan**: Anime and Asian content

Business Insights

1. US dominance reflects home market and Hollywood partnerships
2. Emerging market content shows localization strategy
3. Co-productions appear multiple times due to international collaborations
4. Regional content attracts global audiences (e.g., Squid Game, Money Heist)

6.2 Content Type by Country (Bonus Analysis)

Objective

Examine whether countries specialize in movies or TV shows.

Methodology

1. Filter top 10 content-producing countries
2. Cross-tabulate country and content type
3. Create stacked visualization for comparison

Key Metrics

1. **Movies vs. TV Shows by Country**: Type distribution per nation
2. **Country Specialization**: Whether countries focus on one type
3. **Comparative Analysis**: Cross-country content type preferences

Importance

1. **Regional Patterns**: Different markets have different production traditions
2. **Partnership Strategy**: Informs which countries to approach for specific content types
3. **Market Insights**: Cultural preferences for episodic vs. film content

Expected Patterns

1. Some countries may specialize in film (e.g., film industries)
2. Others may produce more series (e.g., soap opera traditions)
3. US likely shows balanced production across both types

7. Conclusions & Recommendations

Key Findings Summary

1. Content Growth: Netflix has dramatically expanded its catalog, especially in recent years
2. Type Distribution: Balance between movies and series reflects diverse audience needs
3. Rating Strategy: Distribution shows targeted demographic approach
4. Genre Diversity: Wide genre coverage ensures broad market appeal
5. Geographic Expansion: International content demonstrates global strategy success