Netflix Dataset Analysis and Insights

- Kshitij Thorat

Dataset: Netflix Titles Dataset (8,809 initial records)

Summary: This document presents a comprehensive analysis of the Netflix dataset, including detailed data cleaning procedures and key insights derived from 8,784 cleaned records. The analysis reveals content distribution patterns, temporal trends, and content characteristics across Netflix's catalog.

1.Data Cleaning Process:

Initial Dataset Assessment:

Initial Records: 8,809 rows with 12 columns.

Missing Values Identified.

Duplicate Check: No duplicate records found.

2.Data Cleaning Steps:

1. Missing Value are filled by “Unknown”.

2. Text Standardization and Cleaning.

3. Data Type Conversions: Extracted added_year from date strings using regex and added_month by parsing month names to numeric values and maintained original date_added string for reference.

4. Data Validation: Filtered to range 1900-2025 (current year)

5. Final Dataset Structure:
Final Records: 8,784 rows.

Final Columns: 16 columns (6 new engineered features)

Data Quality: All critical missing values handled, invalid records removed.

3.Key Metrics and Insights:

1.Content Distribution:

Content Type Breakdown: MOVIES are in majority.

Content Ratings Distribution:

TV-14:	Popular teen and adult rating
PG-13:	Common movie rating
TV-Y7, TV-G, G:	Family-friendly content segment
Not Rated/UR/NR:	Unrated content category
TV-MA:	Most common rating for mature content

Average Addition Year: 2018.87

2.Duration Characteristics: Typical movie duration clustering around 90-120 minutes and Single-season shows are common.

3.Geographic and Content Diversity:

Genre Distribution (Listed_in): Diverse content categories spanning multiple genres.

4.Conclusion:

The Netflix dataset cleaning process successfully transformed 8,809 raw records into 8,784 high-quality, analysis-ready records with enhanced features. The comprehensive cleaning approach addressed missing values, standardized formats, engineered new features, and applied rigorous quality filters.

The cleaned dataset reveals Netflix's diverse, global content strategy spanning multiple decades, ratings, and content types. The data now supports advanced analytics for content strategy, viewer preference analysis, and business intelligence applications.

Key Achievements:

- 99.7% data retention rate
- 100% missing value treatment
- 6 new analytical features created
- Comprehensive data validation implemented
- Analysis-ready dataset produced