Netflix Dataset Insights

1. Introduction

This document provides a comprehensive analysis of the Netflix dataset, detailing the data cleaning and preprocessing steps performed using Pandas in a Databricks environment. The primary objective of this analysis is to produce a cleaned and well-structured dataset that enables accurate insights, meaningful visualizations, and supports further modeling or predictive analytics.

2. Dataset Information

File: netflix_analysis.csv
Source: Netflix Movies and TV Shows — Shivamb's dataset

Columns:

Column	Description
show_id	Unique ID for each title
type	Movie or TV Show
title	Title of the content
director	Director of the title
cast	Cast members
country	Country of production
date_added	Date added to Netflix
release_year	Year of release
rating	Rating based on age group
duration	Duration of movie (minutes) or TV Show (seasons)
listed_in	Genres (comma-separated)
description	Text description of the title

Notes:

- Some columns had missing values (director, cast, country, rating).
- Some columns were multi-valued (listed_in → multiple genres).
- date_added had inconsistent date formats.

2. Data Cleaning Steps

2.1 Remove Duplicates

- Removed duplicate rows to ensure unique records.
- Specifically dropped duplicates based on 'title' and 'release_year'.

Function:

```
df.drop_duplicates(), df.drop_duplicates(subset=['title','release_year'])
```

2.2 Handle Missing Values

- Replaced missing values in 'director', 'cast', 'country', and 'rating' with 'Unknown'.

Function:

```
df[col].fillna("Unknown")
```

2.3 Converted Dates

- Converted 'date_added' to datetime format, coercing errors to NaT.

Function:

```
pd.to_datetime(df['date_added'], errors='coerce')
```

- Created a binary column 'date_missing' to indicate missing dates.

Function:

```
df['date_added'].isna().astype(int)
```

2.4 Exploded Multi-Value Columns

- Split listed_in (genres) into multiple rows for better analysis of each genre separately.

2.5 Handled Outliers in Duration

- Cleaned duration by removing text like "min" and "Season(s)".

Cleaned dataset:

```
import pandas as pd
df_cleaned = pd.read_csv("/Volumes/workspace/default/netflix/cleaned_netflix.csv")
df_cleaned.head()
```

2.6 Column Transformation and Normalization

- Removed extra spaces in type.

- df_cleaned[‘type’].str.strip()

- Mapped ratings into groups: *Kids, Family, Teens, Adults, Unknown*.

- df_cleaned['rating'].map(rating_map)

- Standardized country names (e.g., USA → United States).

- df_cleaned['country'].replace({'USA':'United States'})

- Used **one-hot encoding** for type.

- pd.get_dummies(df_cleaned['type'],prefix='type')

- Used **Frequency Encoding** for High-Cardinality Columns

- freq_encoding = df_normalized['director'].value_counts().to_dict()
- df_normalized['director_freq'] = df_normalized['director'].map(freq_encoding)

- Used **Ordinal Encoding** for Rating Groups

- ord_enc = OrdinalEncoder(categories=rating_order)
- df_normalized['rating_group_encoded'] = ord_enc.fit_transform(df_cleaned[['rating_group']]).astype(int)

- Grouped rare countries (appearing < 20 times) into "Other".

- df_cleaned['country'].replace(rare_countries, 'Other')

Normalized dataset

normalized_file_path = "/Volumes/workspace/default/netflix/normalized_netflix.csv"

df_normalized = df_cleaned.copy()

df_normalized.to_csv(normalized_file_path, index=False)

print(f"Normalized dataset saved at: {normalized_file_path}")

2.7 Exploratory Data Analysis (EDA)

Plot	Function	Insight
Movies vs TV Shows	df_cleaned['type'].value_counts().plot(kind='bar')	Movies dominate Netflix content library
Content growth over time	df_cleaned['release_year'].value_counts().sort_index().plot(kind='line')	Number of releases increased significantly post-2015
Top 10 countries	df_cleaned['country'].value_counts().head(10).plot(kind='barh')	USA & India are top content producers
Ratings distribution	df_cleaned['rating_group'].value_counts().plot(kind='bar')	Most content is targeted at Teens and Adults
Top 10 genres	df_exploded['genre'].value_counts().head(10).plot(kind='bar')	Drama, International Movies, Comedies dominate

3. Insights from the Dataset

1. **Netflix Library Has More Movies than TV Shows**
    - Movies dominate the Netflix library, indicating a focus on quick-to-release content.
    - Suggests recommendations may be heavily movie-oriented.
2. **United States and India Are the Largest Content Contributors**
    - Majority of content is produced in the U.S. and India, reflecting strong production capacity.
    - English and Hindi content likely dominate the platform.
3. **Drama, Comedies, and International Movies Are the Most Common Genres**
    - These genres are the most frequent, showing Netflix targets broad audience appeal.
    - International Movies indicate diverse content offerings beyond major markets.
4. **Majority of Content Is Aimed at Teens and Adults**
    - Most content is suitable for Teens and Adults, smaller portion for Kids/Family.
    - Netflix primarily targets older audiences for engagement and retention.
5. **The Number of Releases Increased Sharply After 2015**
    - Significant growth in content post-2015 aligns with Netflix's global expansion.
    - Users after 2015 have access to a larger, more varied library.

4. Cleaned Dataset Output

- The cleaned dataset was saved at:
/Volumes/workspace/default/netflix/cleaned_netflix.csv