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 (EDA) - Basics

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. Exploratory Data Analysis (EDA)

3.1Content Type Distribution

Movies dominate the Netflix library, showing a focus on movie content.

- df['type'].value_counts().plot(kind='pie', autopct='%1.1f%%')

3.2 Content Growth Over Time

Significant growth in content after 2015, reflecting Netflix's global expansion.

- df['release_year'].value_counts().sort_index().plot(kind='line', marker='o')

3.3 Country-Level Contributions

The United States and India are top contributors.

- df['country'].value_counts().head(10).plot(kind='bar')

3.4 Genre Distribution

Top genres: Drama, International Movies, and Comedies.

- genre_counts=df['listed_in'].str.split(',',expand=True).stack().value_counts().head(10)

3.5 Ratings Distribution

Most content targets Teens and Adults.

- sns.countplot(data=df,x='rating',order=df['rating'].value_counts().index)

3.6 Bivariate Analysis

Movie vs TV Show Ratings Comparison

- sns.countplot(data=df, x='rating', hue='type', palette='Set2')

Content Growth by Type

- recent_data = df[df['release_year'] >= 2010]
- content_by_year_type=recent_data.groupby(['release_year','type']).size().unstack(fill_value=0)
- content_by_year_type.plot(kind='bar', stacked=True)

4. Feature Engineering

4.1 Content Length Category

For Movies:

Classified movies based on duration.

```
def movie_duration_category(x):
  if pd.isna(x): return 'Unknown'
  elif x < 60: return 'Short'
  elif 60 <= x <= 120: return 'Medium'
  else: return 'Long'
```

```
movie_df['Content_Length_Category'] =
movie_df['duration'].apply(movie_duration_category)
```

For TV Shows:
Categorized based on number of seasons.

```
def tv_show_category(x):
   if pd.isna(x): return 'Unknown'
   elif x == 1: return 'Single Season'
   elif 2 <= x <= 4: return 'Mini Series'
   else: return 'Long Series'
tv_df['Content_Length_Category'] = tv_df['duration'].apply(tv_show_category)
```

Visualization:

```
sns.countplot(data=movie_df, x='Content_Length_Category', palette='pastel')
sns.countplot(data=tv_df, x='Content_Length_Category', palette='cool')
```

4.2 Audience Grouping (from Rating)

Created Age_Group_Category to segment content by audience.

```
def categorize_rating(rating):
   if rating in ['G', 'TV-Y', 'TV-G']:
      return 'Kids'
   elif rating in ['PG', 'TV-PG', 'TV-Y7', 'TV-Y7-FV']:
      return 'Family'
   elif rating in ['PG-13', 'TV-14']:
      return 'Teens'
   elif rating in ['R', 'NC-17', 'TV-MA']:
      return 'Adults'
   else:
      return 'Unknown'
df['Age_Group_Category'] = df['rating'].apply(categorize_rating)
```

Visualization:

```
sns.countplot(data=df, x='Age_Group_Category', hue='type', palette='Set2')
```

4.3 Cast Size Feature

Added a numeric column for cast count.

```
df['cast_count'] = df['cast'].apply(lambda x: len(x.split(',')) if pd.notnull(x) else 0)
```

Visualization:

```
top_casts = df[['title','cast_count']].sort_values(by='cast_count',ascending=False).head(10)
sns.barplot(data=top_casts, x='cast_count', y='title', palette='viridis')
```

4.4 Content Origin (Original vs Licensed)

Determined if content is Netflix Original or Licensed using description text.

```
origin = []
for desc in df['description']:
  if isinstance(desc, str) and ('netflix' in desc.lower() or 'original' in desc.lower()):
    origin.append('Netflix Original')
  else:
    origin.append('Licensed')
df['content_origin'] = origin
```

Visualization:

```
sns.countplot(data=df, x='content_origin', hue='type', palette='pastel')
```

5. 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.

6. Cleaned Dataset Output

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