DATA CLEANING FOR POWERBI:

In [4]:

```
# ===============================
# □ STEP 1: Import Libraries
# ===============================
import pandas as pd
import re
import warnings


#Ignore all warnings:
warnings.filterwarnings('ignore')

# ===============================
# □ STEP 2: Load Dataset Safely
# ===============================
raw_df = pd.read_csv("netflix_titles.csv")
df = raw_df.copy()

print("Original shape:", df.shape)

# ===============================
# □ STEP 3: Handle Missing Values (Soft Approach)
# ===============================
df["director"].fillna("No Director", inplace=True)
df["cast"].fillna("No Cast", inplace=True)
df["country"].fillna("No Country", inplace=True)
df["rating"].fillna("No Rating", inplace=True)
df["date_added"].fillna("Unknown", inplace=True)

# ===============================
# □ STEP 4: Explode Columns Safely (Optional)
# ===============================
# Exploding after filling NaNs prevents losing rows
df["cast"] = df["cast"].str.split(',')
df = df.explode("cast").reset_index(drop=True)

df["country"] = df["country"].str.split(',')
df = df.explode("country").reset_index(drop=True)

df["listed_in"] = df["listed_in"].str.split(',')
df = df.explode("listed_in").reset_index(drop=True)

df["director"] = df["director"].str.split(',')
df = df.explode("director").reset_index(drop=True)

print("After explode shape:", df.shape)

# ===============================
# □ STEP 5: Clean and Extract Date Columns
# ===============================
df['date_added'] = df['date_added'].replace('Unknown', pd.NA)
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')

df['day_added'] = df['date_added'].dt.day
df['month_added'] = df['date_added'].dt.month
df['year_added'] = df['date_added'].dt.year

# Fill NA in these numeric columns with 0 (or choose another strategy)
df['day_added'].fillna(0, inplace=True)
df['month_added'].fillna(0, inplace=True)
df['year_added'].fillna(0, inplace=True)

df['day_added'] = df['day_added'].astype(int)
df['month_added'] = df['month_added'].astype(int)
```

```
df['year_added'] = df['year_added'].astype(int)

# ===============================
# □ STEP 6: Duration Cleaning
# ===============================
def extract_minutes(duration):
    if "min" in str(duration):
        return int(re.sub(r"[^0-9]", "", duration))
    return None

def extract_seasons(duration):
    if "Season" in str(duration):
        return int(re.sub(r"[^0-9]", "", duration))
    return None

df['duration_minutes'] = df['duration'].apply(extract_minutes)
df['duration_tvshows'] = df['duration'].apply(extract_seasons)

# Replace missing with 0 instead of dropping rows
df['duration_minutes'].fillna(0, inplace=True)
df['duration_tvshows'].fillna(0, inplace=True)

# ===============================
# □ STEP 7: Content Age Calculation
# ===============================
df['content_age'] = pd.Timestamp.now().year - df['release_year']

# ===============================
# □ STEP 8: Remove Duplicates (Keep One Copy)
# ===============================
df = df.drop_duplicates(subset=["show_id"])

# ===============================
# □ STEP 9: Final Null Check
# ===============================
print("Null counts after cleaning:")
print(df.isnull().sum())

print("Final shape:", df.shape)



mean_date = df['date_added'].mean()
df['date_added'].fillna(mean_date, inplace=True)



# Final cleanup before saving
df = df[df['year_added'].notna()]   # remove rows with missing year
df = df[df['year_added'] != 0]      # remove rows with 0

# ===============================
# □ STEP 10: Save Cleaned Dataset
# ===============================
df.to_csv('Updated_cleaned_netflix_titles_final.csv', index=False)
print("□ Cleaned file saved successfully as 'Updated_cleaned_netflix_titles_final.csv'")

# ===============================
# □ STEP 11: Preview Top 5 Rows
# ===============================
df.head(5)
```

```
Original shape: (6234, 12)
After explode shape: (139984, 12)
Null counts after cleaning:
show_id                0
type                   0
title                  0
director               0
cast                   0
country                0
date_added           651
```

```
release_year         0
rating               0
duration             0
listed_in            0
description          0
day_added            0
month_added          0
year_added           0
duration_minutes     0
duration_tvshows     0
content_age          0
dtype: int64
Final shape: (6234, 18)
□ Cleaned file saved successfully as 'Updated_cleaned_netflix_titles_final.csv'
```

Out[4]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in
0	81145628	Movie	Norm of the North: King Sized Adventure	Richard Finn	Alan Marriott	United States	2019-09-09	2019	TV-PG	90 min	Children & Family Movies
160	80117401	Movie	Jandino: Whatever it Takes	No Director	Jandino Asporaat	United Kingdom	2016-09-09	2016	TV-MA	94 min	Stand-Up Comedy
161	70234439	TV Show	Transformers Prime	No Director	Peter Cullen	United States	2018-09-08	2013	TV-Y7-FV	1 Season	Kids' TV
173	80058654	TV Show	Transformers: Robots in Disguise	No Director	Will Friedle	United States	2018-09-08	2016	TV-Y7	1 Season	Kids' TV
181	80125979	Movie	#realityhigh	Fernando Lebrija	Nesta Cooper	United States	2017-09-08	2017	TV-14	99 min	Comedies

In [5]:

```
df.shape
```

Out[5]:

```
(5583, 18)
```

In [6]:

```
df.isnull().sum()
```

Out[6]:

```
show_id             0
type                0
title               0
director            0
cast                0
country             0
date_added          0
release_year        0
```

```
rating              0
duration            0
listed_in           0
description         0
day_added           0
month_added         0
year_added          0
duration_minutes    0
duration_tvshows    0
content_age         0
dtype: int64
```

In [ ]:

```

```

In [ ]:

```

```

In [ ]:

```

```

In [ ]:

```

```