Data Cleaning Report: Netflix Dataset

Objective

The purpose of this data cleaning process is to ensure that the Netflix dataset is structured, consistent, and optimized for analysis. The main focus areas include trimming spaces, handling missing values, normalizing data formats, and unnesting multi-valued fields.

Cleaning Steps and Transformations

1. Removing Extra Spaces

- Applied `TRIM()` function to remove leading and trailing spaces from text-based columns:
  - `title`
  - `director`
  - `casts`
  - `country`
  - `rating`
  - `listed_in` (genre)
  - `description`
- Ensured that NULL values replace empty strings (`NULLIF(TRIM(column_name), '')`).**2. Handling Missing Values**

- Replaced empty values with NULL for `director`, `casts`, `country`, `rating`, and `description`.
- `title` column is mandatory, so rows with missing titles were excluded.

3. Splitting Multi-Valued Columns

- `country`: Contains multiple values separated by commas (e.g., "USA, Canada"). Applied `STRING_TO_ARRAY()` and `UNNEST()` to store each country in a separate row.
- `listed_in` (genre): Contains multiple values separated by commas. Applied `STRING_TO_ARRAY()` and `UNNEST()` to ensure each genre has its own row.

4. Date Formatting

- `date_added`: Originally stored as a string in "Month DD, YYYY" format.
- Converted to `DATE` format using:
- `CASE`
- `    WHEN date_added ~ '^[A-Za-z]+ \d{1,2}, \d{4}$' THEN TO_DATE(date_added, 'Month DD, YYYY')`
- `    ELSE NULL`
- `END AS date_added`

5. Duration Normalization

- `duration` column contains values like "90 min" (for movies) and "3 Seasons" (for TV shows).
- Split the numeric values using `SPLIT_PART()` and stored separately:
    - `movie_duration_minutes`: Extracted for movies.
    - `tv_show_seasons`: Extracted for TV shows.

6. Removing Duplicates

- Applied `DISTINCT` to eliminate redundant rows caused by unnesting `country` and `listed_in` columns.

-

Final Query

```
WITH cleaned_data AS (
    SELECT
        show_id,
        type,
        TRIM(title) AS title,
        NULLIF(TRIM(director), '') AS director,
        NULLIF(TRIM(casts), '') AS casts,
        -- Unnesting 'country'
        UNNEST(STRING_TO_ARRAY(NULLIF(TRIM(country), ''), ','))::TEXT AS
country,
        -- Date conversion
        CASE
            WHEN date_added ~ '^[A-Za-z]+ \d{1,2}, \d{4}$' THEN
TO_DATE(date_added, 'Month DD, YYYY')
            ELSE NULL
        END AS date_added,
        release_year,
        NULLIF(TRIM(rating), '') AS rating,
        -- Duration normalization
        CASE
            WHEN type = 'Movie' THEN NULLIF(SPLIT_PART(duration, ' ', 1),
'')::INTEGER
            ELSE NULL
        END AS movie_duration_minutes,
        CASE
            WHEN type = 'TV Show' THEN NULLIF(SPLIT_PART(duration, ' ', 1),
'')::INTEGER
            ELSE NULL
        END AS tv_show_seasons,
        -- Unnesting 'listed_in' (genres)
        UNNEST(STRING_TO_ARRAY(NULLIF(TRIM(listed_in), ''), ','))::TEXT AS
genre,
        NULLIF(TRIM(description), '') AS description
    FROM netflix
    WHERE title IS NOT NULL
)
SELECT DISTINCT * FROM cleaned_data;
```

Results and Improvements

Issue Fixed	Description
Extra Spaces	Removed leading/trailing spaces from text fields
Missing Values	Handled NULL replacements for empty values
Multi-Valued Columns	`country` and `listed_in` were unnested into separate rows
Date Format	Converted `date_added` to `DATE` format
Duration Standardization	Split `duration` into separate `movie_duration_minutes` and `tv_show_seasons` columns
Duplicate Entries	Applied `DISTINCT` to remove redundant rows

Conclusion

The cleaned Netflix dataset is now structured, optimized, and ready for analysis. These transformations ensure better query performance and data integrity while maintaining accuracy.

✅ **Trimmed text fields**
✅ **Handled missing values**
✅ **Unnested multi-valued columns**
✅ **Standardized date formats**
✅ **Normalized duration data**
✅ **Removed duplicate records**

This structured dataset will enable more efficient querying and analysis for insights into Netflix content trends. 🚀