```
-- ####################################################################
-- ################### Netflix Data Preprocessing (PostgreSQL) ##################
-- ####################################################################
-- ######################################################
-- # 1. Drop and Create the Database
-- ######################################################
-- Drop the database if it already exists to start fresh
DROP DATABASE IF EXISTS netflix_pre_processing;
-- Create a new database
CREATE DATABASE netflix_pre_processing;
-- Connect to the newly created database in CLI
-- ####################################################################
-- ################### Create Table for Netflix Data ################
-- ####################################################################
-- Drop table if it already exists
DROP TABLE IF EXISTS netflix2021;
CREATE TABLE netflix2021 (
show_id VARCHAR(100) PRIMARY KEY,
category VARCHAR(255),
title VARCHAR(500) NOT NULL,
director VARCHAR(255),
cast_members TEXT,
country VARCHAR(255),
release_date DATE,  -- Changed to DATE type for better handling
rating VARCHAR(100),
duration VARCHAR(100),
type VARCHAR(100),
description TEXT
);
-- ######################################################
-- # 2. Handle NaN Values and Convert to NULL
-- ######################################################
UPDATE netflix2021
SET
director = NULLIF(director, 'NaN'),
country = NULLIF(country, 'NaN'),
release_date = NULLIF(release_date, 'NaN'),
rating = NULLIF(rating, 'NaN'),
duration = NULLIF(duration, 'NaN'),
type = NULLIF(type, 'NaN'),
description = NULLIF(description, 'NaN');
-- ######################################################
-- # 3. Check for Duplicate `show_id`
-- ######################################################
-- Identify duplicate show_id values
SELECT show_id, COUNT(*)
FROM netflix2021
GROUP BY show_id
HAVING COUNT(*) > 1;
-- ######################################################
-- # 4. Remove Duplicates Using Window Function
-- ######################################################
-- Drop the temporary table if it exists
DROP TABLE IF EXISTS netflix2021_temp;
-- Create a temporary table with unique rows using ROW_NUMBER()
```

```
CREATE TABLE netflix2021_temp AS
SELECT * FROM (
SELECT *, ROW_NUMBER() OVER (PARTITION BY title, director, country, release_date ORDER
BY show_id) AS row_num
FROM netflix2021
) AS temp
WHERE row_num = 1;
-- Replace original table with the cleaned table
DROP TABLE netflix2021;
ALTER TABLE netflix2021_temp RENAME TO netflix2021;
-- ########################################################
-- # 5. Check for NULL Values
-- ########################################################
-- Count NULL values in each column
SELECT
COUNT(*) FILTER (WHERE show_id IS NULL) AS showid_nulls,
COUNT(*) FILTER (WHERE category IS NULL) AS category_nulls,
COUNT(*) FILTER (WHERE title IS NULL) AS title_nulls,
COUNT(*) FILTER (WHERE director IS NULL) AS director_nulls,
COUNT(*) FILTER (WHERE cast_members IS NULL) AS cast_members_nulls,
COUNT(*) FILTER (WHERE country IS NULL) AS country_nulls,
COUNT(*) FILTER (WHERE release_date IS NULL) AS release_date_nulls,
COUNT(*) FILTER (WHERE rating IS NULL) AS rating_nulls,
COUNT(*) FILTER (WHERE duration IS NULL) AS duration_nulls,
COUNT(*) FILTER (WHERE type IS NULL) AS type_nulls,
COUNT(*) FILTER (WHERE description IS NULL) AS description_nulls
FROM netflix2021;
-- ########################################################
-- # 6. Populate NULL `director` Values Using `cast_members`
-- ########################################################
UPDATE netflix2021 AS n1
SET director = n2.director
FROM (
SELECT cast_members, MAX(director) AS director
FROM netflix2021
WHERE director IS NOT NULL
GROUP BY cast_members
) AS n2
WHERE n1.cast_members = n2.cast_members
AND n1.director IS NULL;
-- ########################################################
-- # 7. Fill Remaining NULL `director` Values
-- ########################################################
UPDATE netflix2021
SET director = 'Not Given'
WHERE director IS NULL;
-- ########################################################
-- # 8. Populate NULL `country` Values Using `director`
-- ########################################################
UPDATE netflix2021 AS n1
SET country = n2.country
FROM (
SELECT director, MAX(country) AS country
FROM netflix2021
WHERE country IS NOT NULL
```

```
GROUP BY director
) AS n2
WHERE n1.director = n2.director
AND n1.country IS NULL;
-- ######################################################
-- # 9. Fill Remaining NULL `country` Values
-- ######################################################
UPDATE netflix2021
SET country = 'Not Given'
WHERE country IS NULL;
-- ######################################################
-- # 10. Delete Rows Where Critical Columns Have NULL Values
-- ######################################################
DELETE FROM netflix2021 WHERE release_date IS NULL OR rating IS NULL OR duration IS
NULL;
-- ######################################################
-- # 11. Standardize Text Formatting
-- ######################################################
UPDATE netflix2021
SET title = TRIM(title),
director = TRIM(director),
country = TRIM(country);
-- ######################################################
-- # 12. Drop Unnecessary Columns
-- ######################################################
ALTER TABLE netflix2021
DROP COLUMN cast_members;
-- ######################################################
-- # 13. Final Data Validation
-- ######################################################
-- Confirm all NULL values are handled
SELECT
COUNT(*) FILTER (WHERE show_id IS NULL) AS showid_nulls,
COUNT(*) FILTER (WHERE category IS NULL) AS category_nulls,
COUNT(*) FILTER (WHERE title IS NULL) AS title_nulls,
COUNT(*) FILTER (WHERE director IS NULL) AS director_nulls,
COUNT(*) FILTER (WHERE country IS NULL) AS country_nulls,
COUNT(*) FILTER (WHERE release_date IS NULL) AS date_added_nulls,
COUNT(*) FILTER (WHERE rating IS NULL) AS rating_nulls,
COUNT(*) FILTER (WHERE duration IS NULL) AS duration_nulls,
COUNT(*) FILTER (WHERE type IS NULL) AS type_nulls,
COUNT(*) FILTER (WHERE description IS NULL) AS description_nulls
FROM netflix2021;
-- ######################################################
-- # Final Check: Ensure Data is Cleaned Properly
-- ######################################################
SELECT * FROM netflix2021 LIMIT 10;
```