# **SQL Data Cleaning**

Here is the complete SQL Documentation for data cleaning.

1. **Netflix\_titles**

This is a detailed analysis of the netflix\_titles table that contains 8,809 rows and 12 columns.

| **Name** | **Data type** | **Description** | **Values (example)** |
| --- | --- | --- | --- |
| show\_id | Primary key  nvarchar(50) | ID for every show | s1  s10 |
| type | nvarchar(50) | Content classification | Movie  TV Show |
| title | nvarchar(250) | Content title | Dick Johnson Is Dead  On the Verge |
| country | nvarchar(max) | Countries where the title is available | United States  India |
| date\_added | date | Date when the content was added to Netflix | 2021-09-25  2021-09-24 |
| release\_year | int | Release year of content | 2020  2021 |
| rating | nvarchar(50) | Content rating | PG-13  TV-MA |
| listed\_in | nvarchar(max) | Genres associated with the content | Documentaries  Comedies, Dramas |
| director | nvarchar(250) | Director's name | Kirsten Johnson  Theodore Melfi |
| cast | text | Cast of the content | Melissa McCarthy, Chris O'Dowd, etc |
| duration | nvarchar(50) | Content duration | 90 min  1 Season |
| description | text | Brief description of the content | As her father nears the end of his life, filmmaker Kirsten Johnson stages his death in inventive and comical ways to help them both face the inevitable. |

1. **Back up my Netflix\_titles table.**

****

1. **Data Cleaning.**

**3.1 Drop unnecessary columns.**

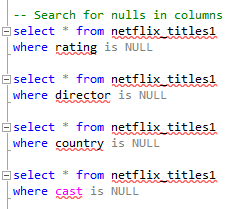
****

**3.2 Correct errors in “duration” and “rating”**

There were 3 rows where the duration value was incorrectly placed in the rating column.

****

**3.3 Search for nulls values in the data**

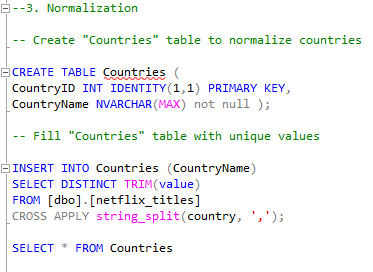
****

1. **Countries Normalization**

**4.1 Countries table**

It was necessary to create the "Countries" table to normalize the country data. This table contains only two columns:

* CountryID INT, Primary Key
* CountryName NVARCHAR(MAX), not null

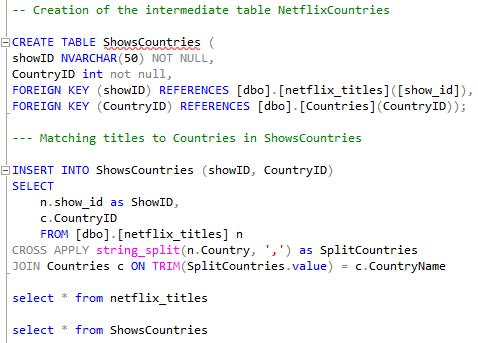


**4.2 ShowsCountries table**

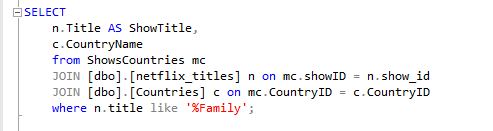
To connect our netflix\_titles to the Countries table, I created an intermediary table containing the IDs from both tables:

-showID int not null.

-CountryID int not null.



**4.3 Double-check relationships between ShowCountry - Countries**

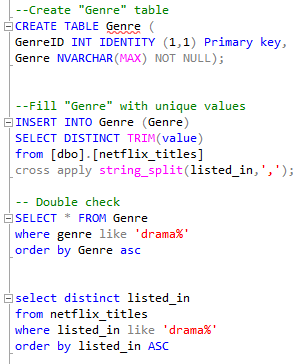
****

1. **Genres Normalization**

**5.1 Create Genre table**

This table contains unique genre values with their corresponding IDs:

* GenreID int, Primary Key
* Genre NVARCHAR(MAX) Not null,

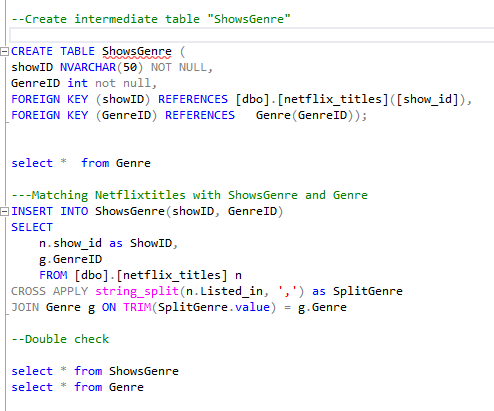


**5.2 Create intermediate table “ShowsGenre”**

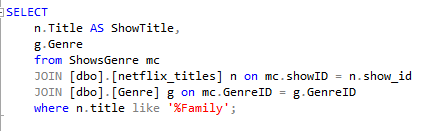
This table contains IDs from netflix\_titles and Genre

-showID NVARCHAR(50) not null,

-GenreID INT not null,



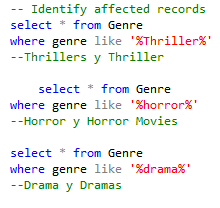
**5.3 Double-check relationships between ShowGenre - Genre**

****

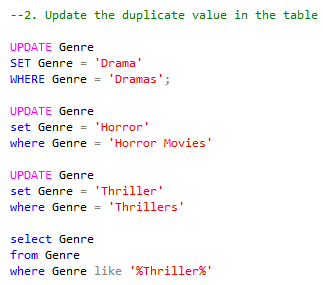
1. **Deleting repeated values**

**6.1 Identify affected records**

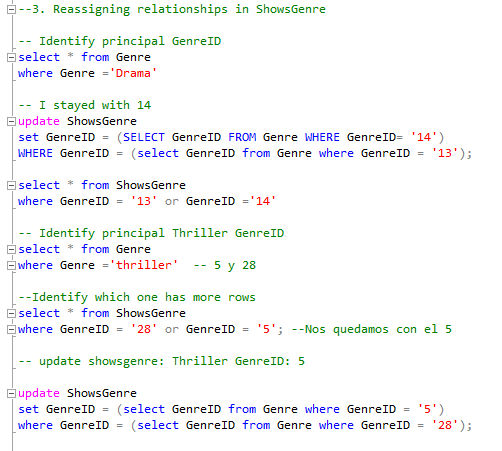
In this case, there were only 3 repeated values “Thriller”, “Drama” and, “Horror”

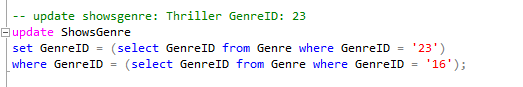
****

**6.2 Update duplicate values**

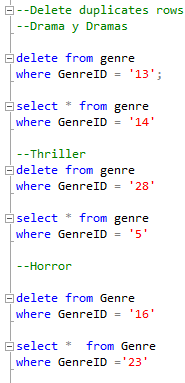
****

**6.3 Reasign relationships**

****

****

**6.4 Delete duplicate values**

****

Thanks to these transformations, we now have 1 fact table and 4 additional tables related to the netflix\_title table.

Now, we are ready to start our analysis in Power BI.