**Data Cleaning and EDA in PostgreSQL**

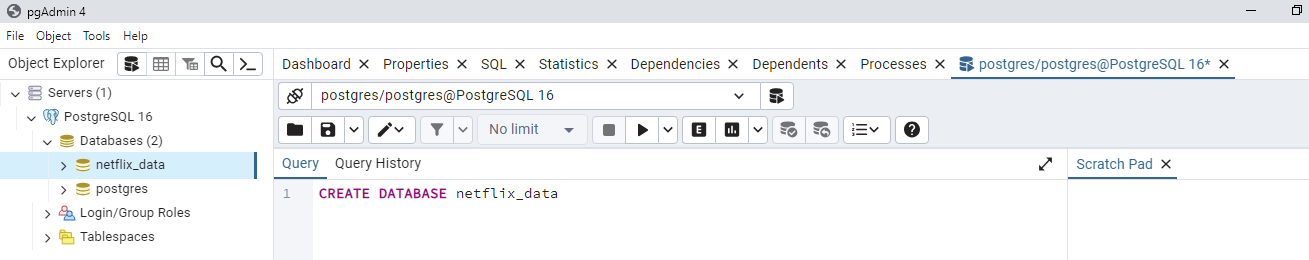
**Objective :** Netflix is one of the most popular media and video streaming platforms. They have over 8000 movies or tv shows available on their platform, from 2008 to as of mid-2021. This tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc. The purpose of this dataset is to clean the data with PostgreSQL.

**Steps followed for Data Cleaning :**

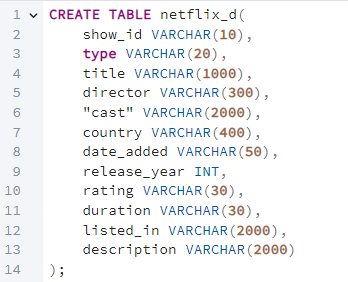
1. Creating database and Table in pgAdmin4
2. Importing dataset
3. Remove duplicates
4. Remove Null values
5. Split Column
6. Drop not needed columns

**Code Comments:**

* The ‘**netflix\_data’** Database is created .



* The table ‘**netflix\_d**’ is created.

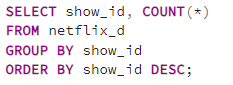


* **Importing the data** to table ‘netflix\_d’.

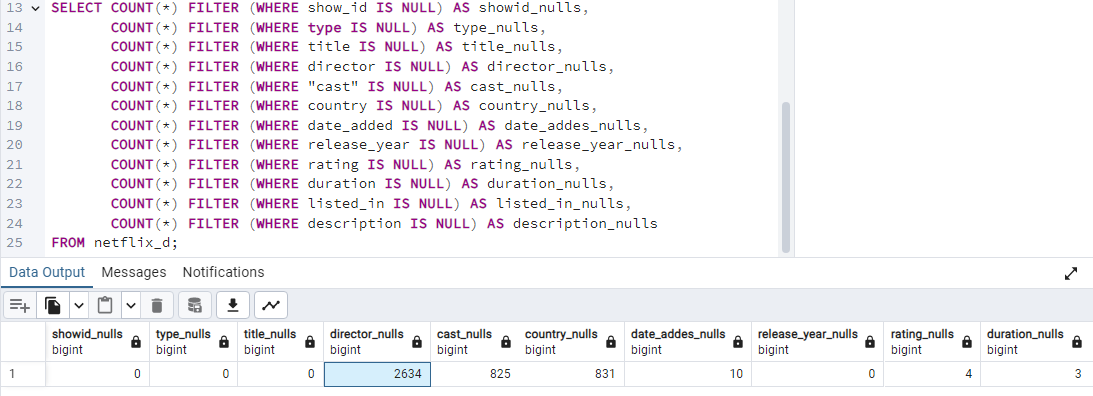
A screenshot of a computer

Description automatically generated

* Checking for **duplicates** in show\_id.



* Verifying **Null** Values across columns.

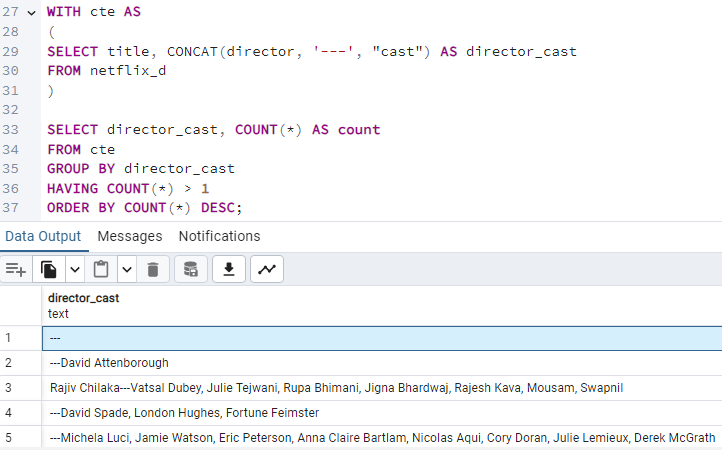


There are NULLS.

1. director\_nulls = 2634
2. cast\_nulls = 825
3. country\_nulls = 831
4. date\_added\_nulls = 10
5. rating\_nulls = 4
6. duration\_nulls = 3

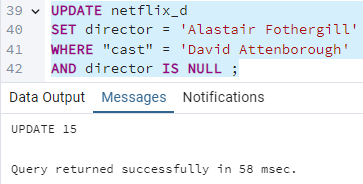
I will not delete the director column as it contain about 30% of nulls. Let’s check another column to populate it. To populate the director column, let’s find out if there is relationship between cast column and director column.

* Check if some directors are likely to work with particular cast



We can now populate NULL rows in directors using their record with cast.

* Repeat this step to populate the rest of the director nulls.



* Populate the rest of the NULL in director as "Not Given".

A screenshot of a computer error message

Description automatically generated

* Let’s not delete the nulls in country. Since the country column is related to director and movie, populate the country column using the director column.



* Confirm if there are still directors linked to country that refuse to update.

A close up of words

Description automatically generated

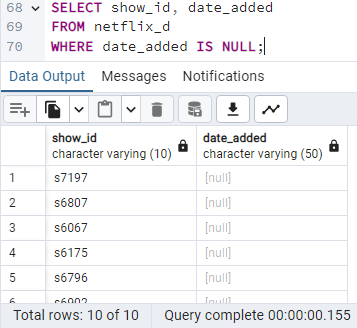
* Populate the rest of the NULL in director as "Not Given".

A screenshot of a computer screen

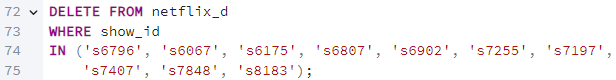
Description automatically generated

* Deleting the date\_added nulls .

Show date\_added



Delete nulls.



* Deleting the ratings nulls.

Show rating nulls

A screenshot of a computer

Description automatically generated

Delete the nulls and show deleted fields

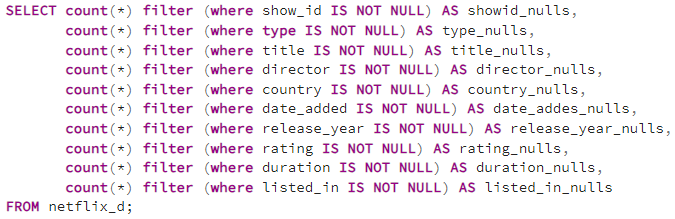
A screenshot of a computer

Description automatically generated

* Delete duration nulls.

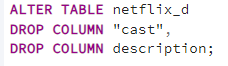


* Execute the below query to display the number of nulls in each column and to confirm if there are still nulls. After this, execute the query to confirm the row number in each column is the same.

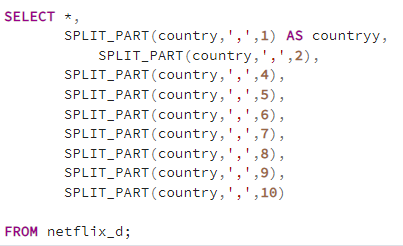


Total number of rows are same in all columns.

* Dropping the not needed columns in table.



* Some of the rows in country column has multiple countries. So, split the country column and retain the first country by the left.



* Create a column named country1 and Update it with the first split country.





* Delete the country column



* Rename the country1 column to country

