1. Проверка на дубликаты:

SELECT show\_id, COUNT(\*) as row\_counter

FROM netflix\_original

GROUP BY show\_id

ORDER BY show\_id;

SELECT \*, COUNT(\*) OVER (PARTITION BY title) as row\_counter

FROM netflix\_original;

Найдены только уникальные id – дубликатов нет.

При проверке на дубликаты по title были найдены три пары задублированных строк с уникальными id – удалим дубликаты по id:

DELETE

FROM netflix\_original

WHERE (SELECT \*, ROW\_NUMBER() OVER(PARTITION BY title) AS "dup"

FROM netflix\_original

WHERE "dup" > 1);

1. Проверка на NULL:

SELECT COUNT(\*) FILTER (WHERE show\_id IS NULL) AS showid\_nulls,

COUNT(\*) FILTER (WHERE type IS NULL) AS type\_nulls,

COUNT(\*) FILTER (WHERE title IS NULL) AS title\_nulls,

COUNT(\*) FILTER (WHERE director IS NULL) AS director\_nulls,

COUNT(\*) FILTER (WHERE movie\_cast IS NULL) AS movie\_cast\_nulls,

COUNT(\*) FILTER (WHERE country IS NULL) AS country\_nulls,

COUNT(\*) FILTER (WHERE date\_added IS NULL) AS date\_addes\_nulls,

COUNT(\*) FILTER (WHERE release\_year IS NULL) AS release\_year\_nulls,

COUNT(\*) FILTER (WHERE rating IS NULL) AS rating\_nulls,

COUNT(\*) FILTER (WHERE duration IS NULL) AS duration\_nulls,

COUNT(\*) FILTER (WHERE listed\_in IS NULL) AS listed\_in\_nulls,

COUNT(\*) FILTER (WHERE description IS NULL) AS description\_nulls

FROM netflix\_original;

NULL обнаружены только в столбцах director и country (в небольшом количестве), что не повлияет на анализ и визуализацию. Заменим их на выражения Not Given:

UPDATE netflix\_original

SET director = 'Not Given'

WHERE director IS NULL;

1. Удалим ненужные для анализа столбцы:

ALTER TABLE netflix\_original

DROP COLUMN movie\_cast

DROP COLUMN description;

1. Было замечено, что некоторые строки в поле страны содержат несколько стран через запятую (2 или 3) – удалим не первые страны:

SELECT \*,

SPLIT\_PART(country, ',', 1) AS right\_country

FROM netflix\_original;

ALTER TABLE netflix\_original

DROP COLUMN country;

Выгружаем данные из SQLite и загружаем в Jupyter для дальнейшей обработки (питоновский код приложен).

После обработки данных в Jupyter, выгружаем их в формате excel для небольшой корректировки (ее удобнее сделать именно в excel).

1. В столбце type содержатся значения: Movie и TV Show. Заменим их на более простые к пониманию для русофонов Фильм и Сериал (с помощью CTRL+F).
2. Добавим столбцы dur\_series и dur\_movies для удобства фильтрации в Power BI и в столбце movies добавим значения с промежутками минут (например, 90-120 минут) вместо целых значений (87, 115 минут и т.д.)
3. Добавим новый столбец genre, в котором присвоим более лаконичные значения жанров из столбца listed\_in1 для дальнейшей фильтрации в Power BI.
4. Для фильтрации по году добавления контента на Netflix вычленим из полной даты добавления только год в столбце year\_added.