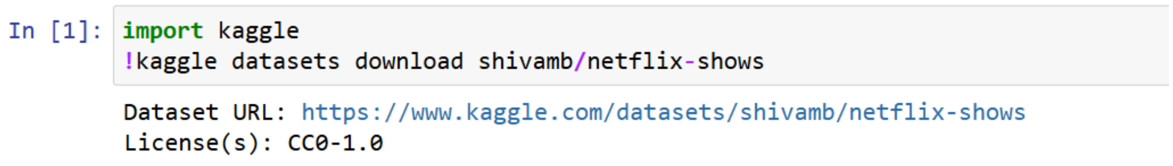
NETFLIX DATA CLEANING AND ANALYTICS

In this analytics project , mainly focuses on data cleaning process and later analytics on cleaned data.
It involves ELT process (Extract Load and Transform). It firstly extract the data from kaggle and loaded it in postgresql database and then transformed the data in database.

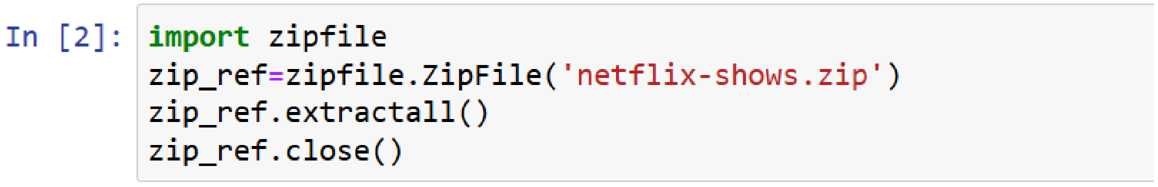
Techstack used : Python, SQL

Extract:

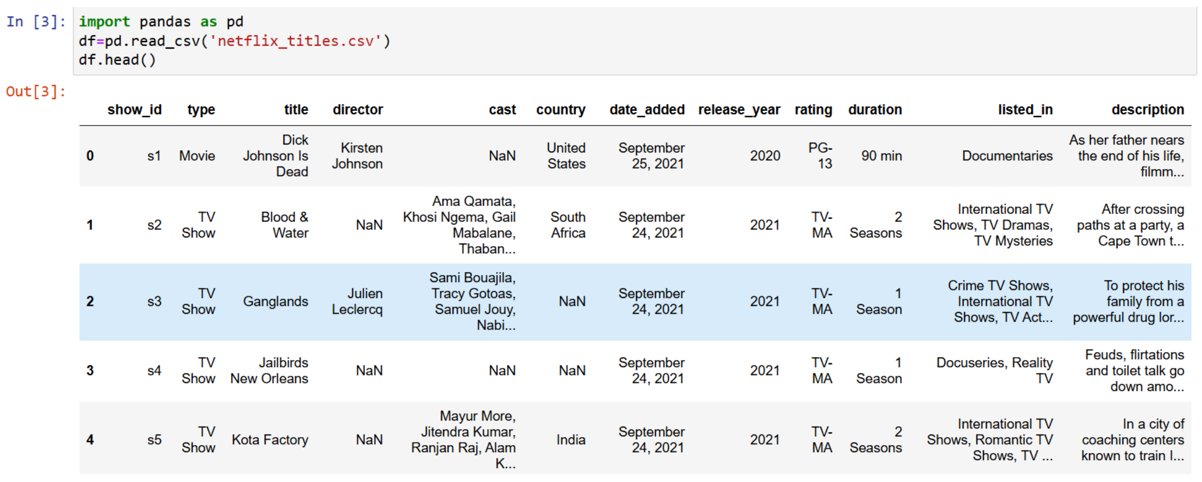
The netflix data is available on Kaggle website using python library, I've downloaded the dataset to local machine.
For that I've used kaggle library in python. It will download the dataset in the format of zip.
Used zipfile library in python to extract the data.
Through the help of Python pandas I've read the csv data and done some analytics of finding null values in each column using python.



```
In [1]: import kaggle
        !kaggle datasets download shivamb/netflix-shows

        Dataset URL: https://www.kaggle.com/datasets/shivamb/netflix-shows
        License(s): CC0-1.0
```



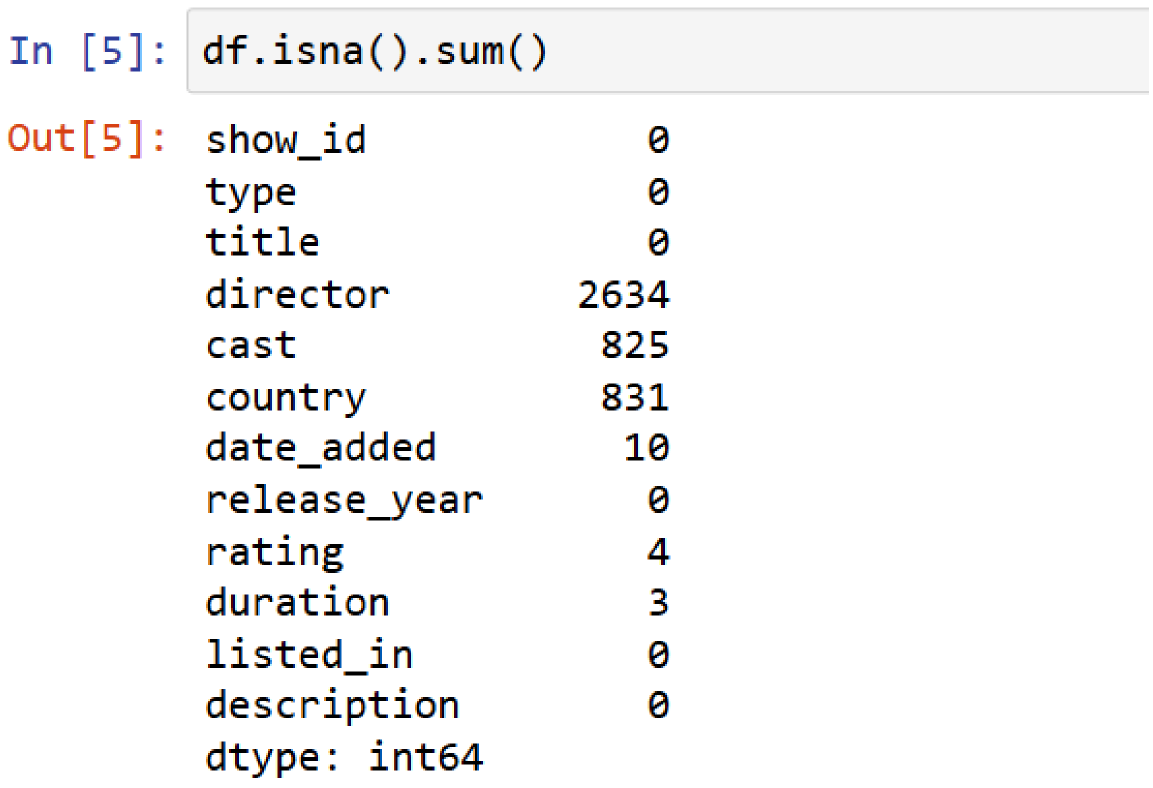
```
In [2]: import zipfile
        zip_ref=zipfile.ZipFile('netflix-shows.zip')
        zip_ref.extractall()
        zip_ref.close()
```



```
In [3]: import pandas as pd
        df=pd.read_csv('netflix_titles.csv')
        df.head()
```

Out[3]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

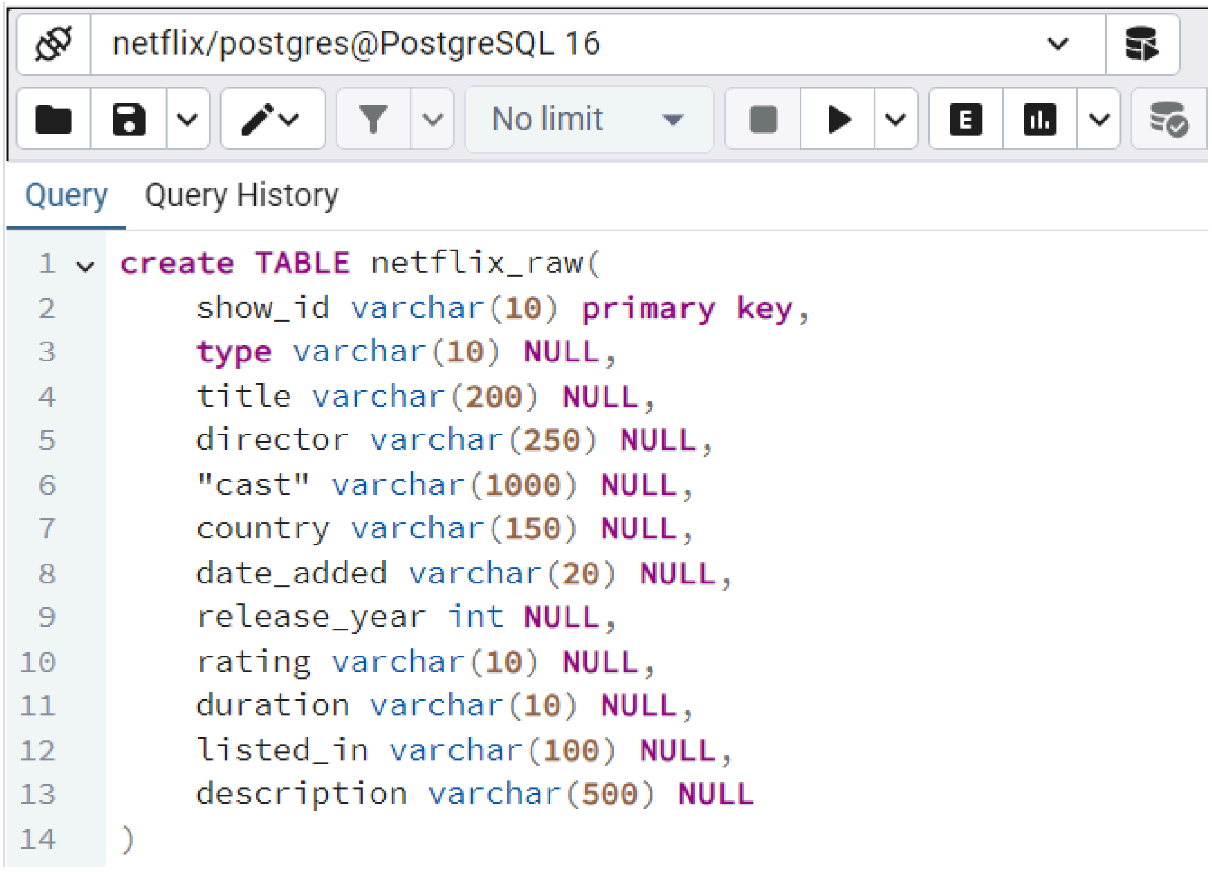


```
In [5]: df.isna().sum()

Out[5]: show_id           0
        type              0
        title             0
        director       2634
        cast            825
        country         831
        date_added       10
        release_year      0
        rating            4
        duration          3
        listed_in         0
        description       0
        dtype: int64
```

Loading the data in Postgres DB:

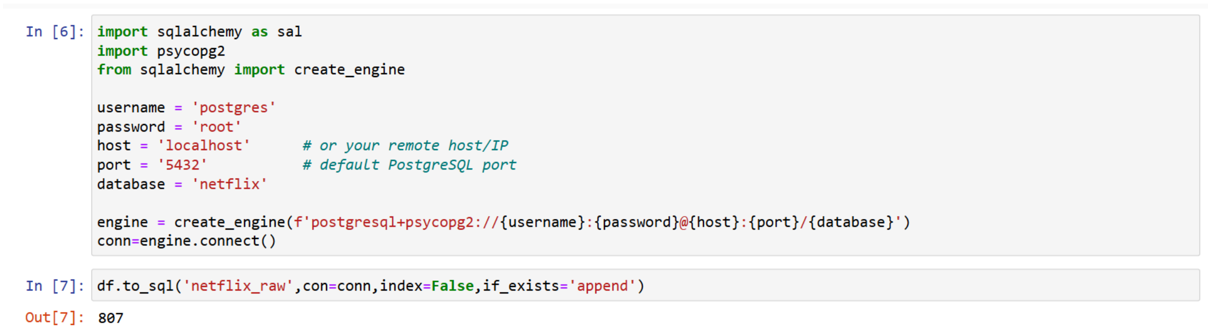
From Pandas dataframe to SQL table.

Firstly, I’ve created the table manually in pgadmin.
Why I choose manually because It can create automatically table but it will create the memory to max of it. So it can cause memory issues.
I’ve manually allocated space for each column by manually checking the max length of each column.

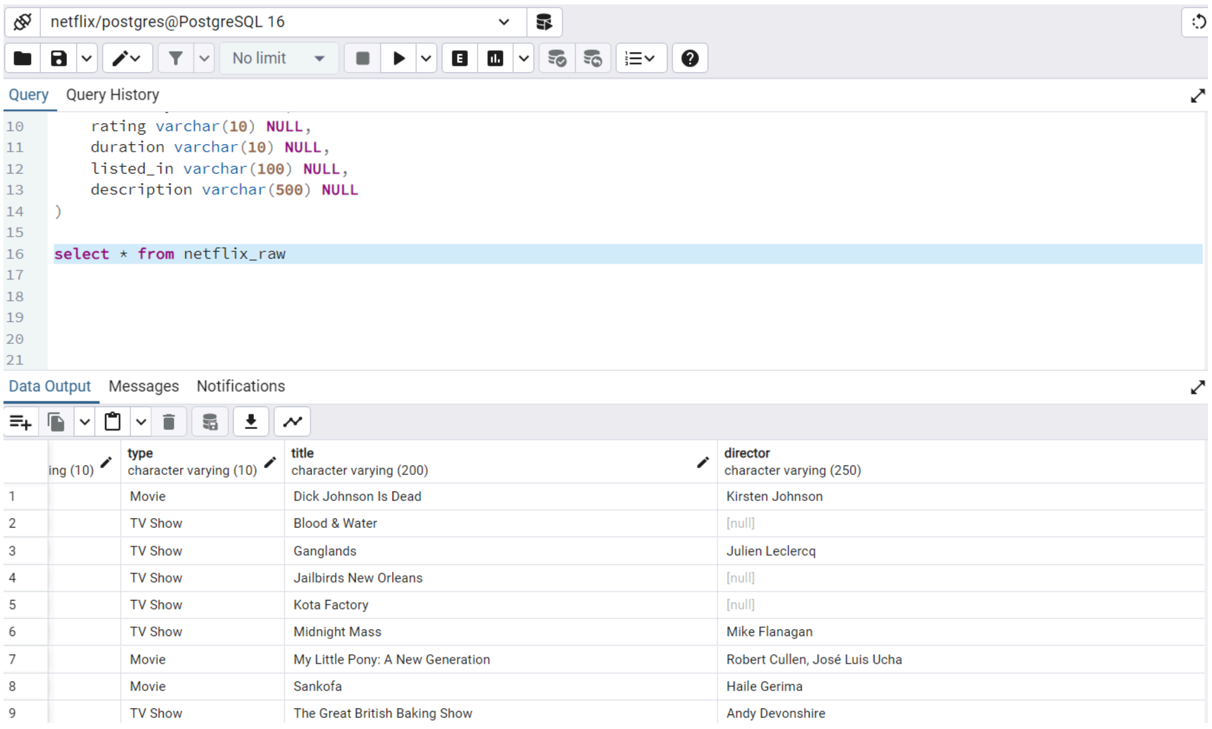


```sql
create TABLE netflix_raw(
    show_id varchar(10) primary key,
    type varchar(10) NULL,
    title varchar(200) NULL,
    director varchar(250) NULL,
    "cast" varchar(1000) NULL,
    country varchar(150) NULL,
    date_added varchar(20) NULL,
    release_year int NULL,
    rating varchar(10) NULL,
    duration varchar(10) NULL,
    listed_in varchar(100) NULL,
    description varchar(500) NULL
)
```

Then using python sqlachemy library , which is used for inserting the data into DB. Psycopg2 library is for postgres DB.



```
In [6]: import sqlalchemy as sal
        import psycopg2
        from sqlalchemy import create_engine

        username = 'postgres'
        password = 'root'
        host = 'localhost'      # or your remote host/IP
        port = '5432'           # default PostgreSQL port
        database = 'netflix'

        engine = create_engine(f'postgresql+psycopg2://{username}:{password}@{host}:{port}/{database}')
        conn=engine.connect()

In [7]: df.to_sql('netflix_raw',con=conn,index=False,if_exists='append')

Out[7]: 807
```

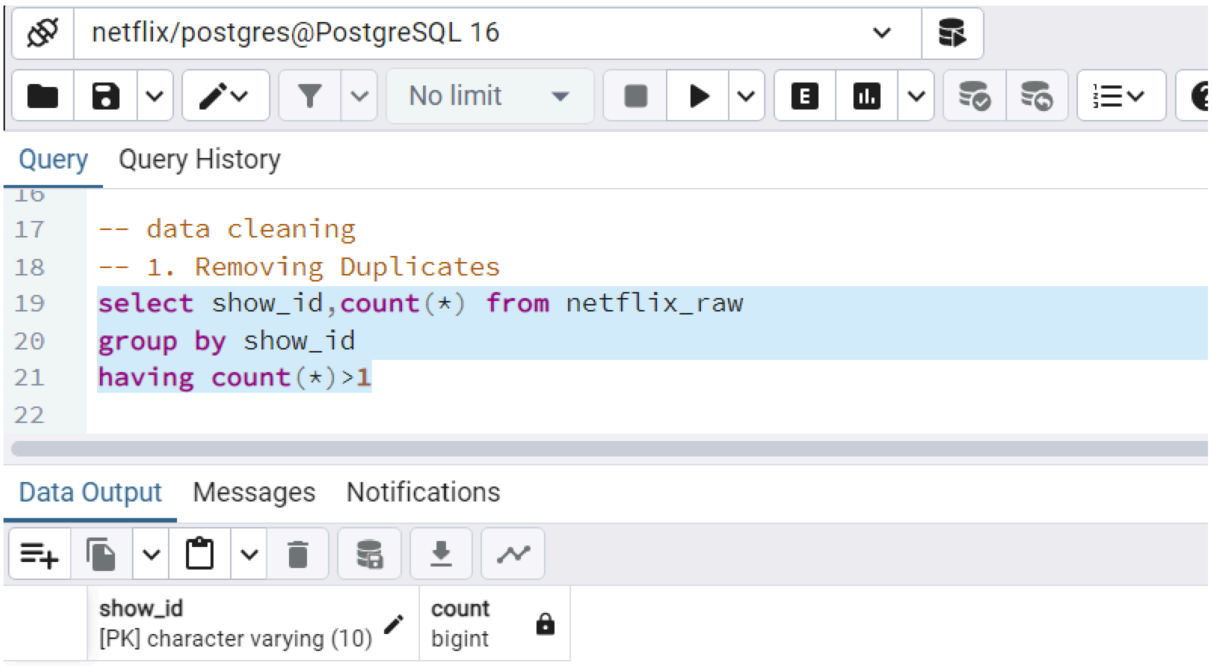


```
    rating varchar(10) NULL,
    duration varchar(10) NULL,
    listed_in varchar(100) NULL,
    description varchar(500) NULL
)

select * from netflix_raw
```

	ing (10)	type character varying (10)	title character varying (200)	director character varying (250)
1		Movie	Dick Johnson Is Dead	Kirsten Johnson
2		TV Show	Blood & Water	[null]
3		TV Show	Ganglands	Julien Leclercq
4		TV Show	Jailbirds New Orleans	[null]
5		TV Show	Kota Factory	[null]
6		TV Show	Midnight Mass	Mike Flanagan
7		Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha
8		Movie	Sankofa	Haile Gerima
9		TV Show	The Great British Baking Show	Andy Devonshire

Transforming the data:

Firstly, focusing on removing duplicate values in raw_data.
Checked column wise,

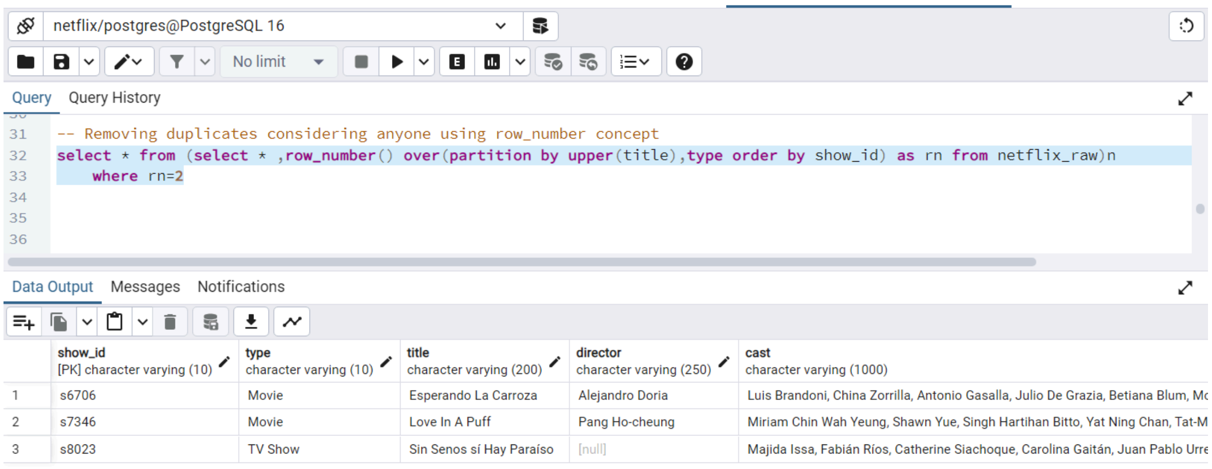


```
-- data cleaning
-- 1. Removing Duplicates
select show_id,count(*) from netflix_raw
group by show_id
having count(*)>1
```

show_id [PK] character varying (10)	count bigint

I could see no duplicates found for the show_id column as it is primary key, it does not allow the duplicate value.

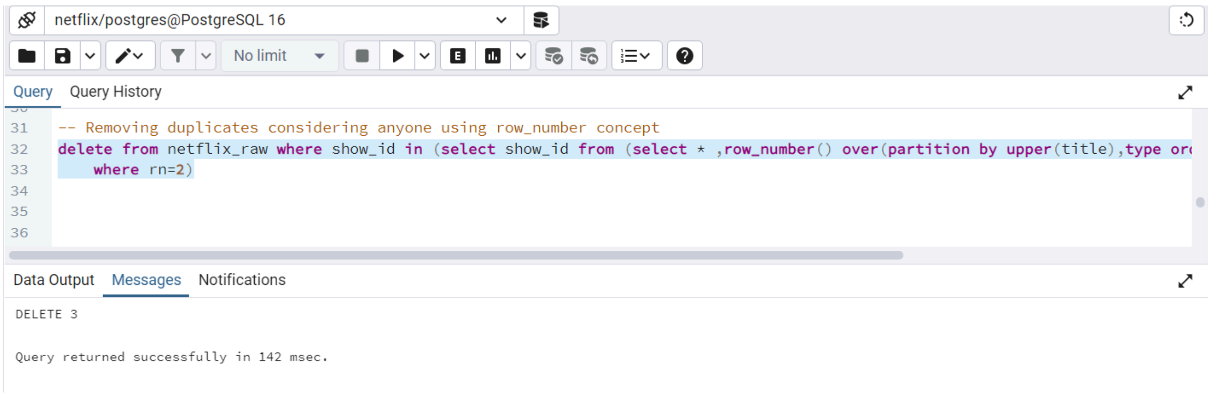
Next performed for title column



```
-- Removing duplicates considering anyone using row_number concept
select * from (select * ,row_number() over(partition by upper(title),type order by show_id) as rn from netflix_raw)n
    where rn=2
```

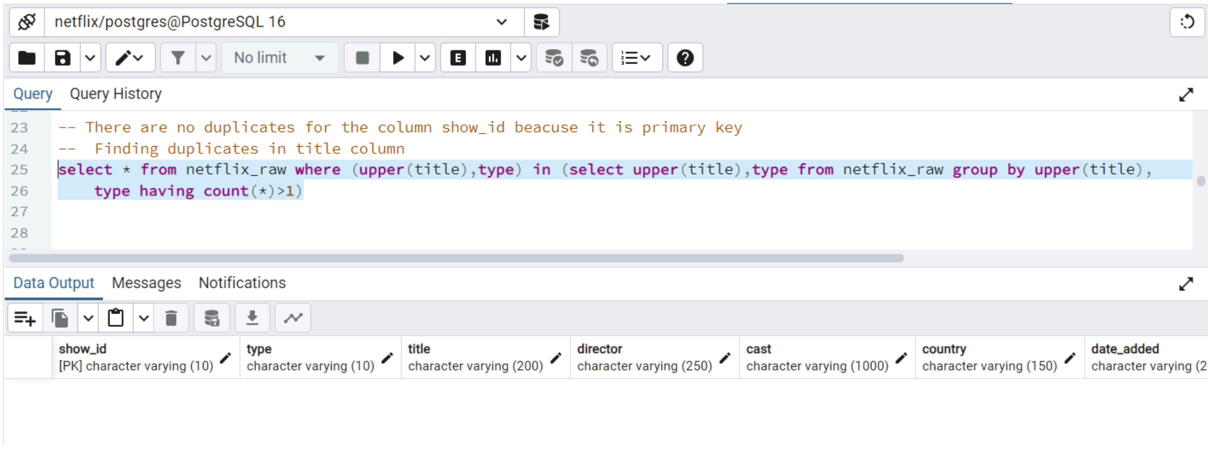
	show_id [PK] character varying (10)	type character varying (10)	title character varying (200)	director character varying (250)	cast character varying (1000)
1	s6706	Movie	Esperando La Carroza	Alejandro Doria	Luis Brandoni, China Zorrilla, Antonio Gasalla, Julio De Grazia, Betiana Blum, Mc
2	s7346	Movie	Love In A Puff	Pang Ho-cheung	Miriam Chin Wah Yeung, Shawn Yue, Singh Hartihan Bitto, Yat Ning Chan, Tat-M
3	s8023	TV Show	Sin Senos sí Hay Paraíso	[null]	Majida Issa, Fabián Ríos, Catherine Siachoque, Carolina Gaitán, Juan Pablo Urre

I could see 3 movies having duplicate values with different show_id
I removed the duplicate values by using row_number function



```
-- Removing duplicates considering anyone using row_number concept
delete from netflix_raw where show_id in (select show_id from (select * ,row_number() over(partition by upper(title),type ord
    where rn=2)
```

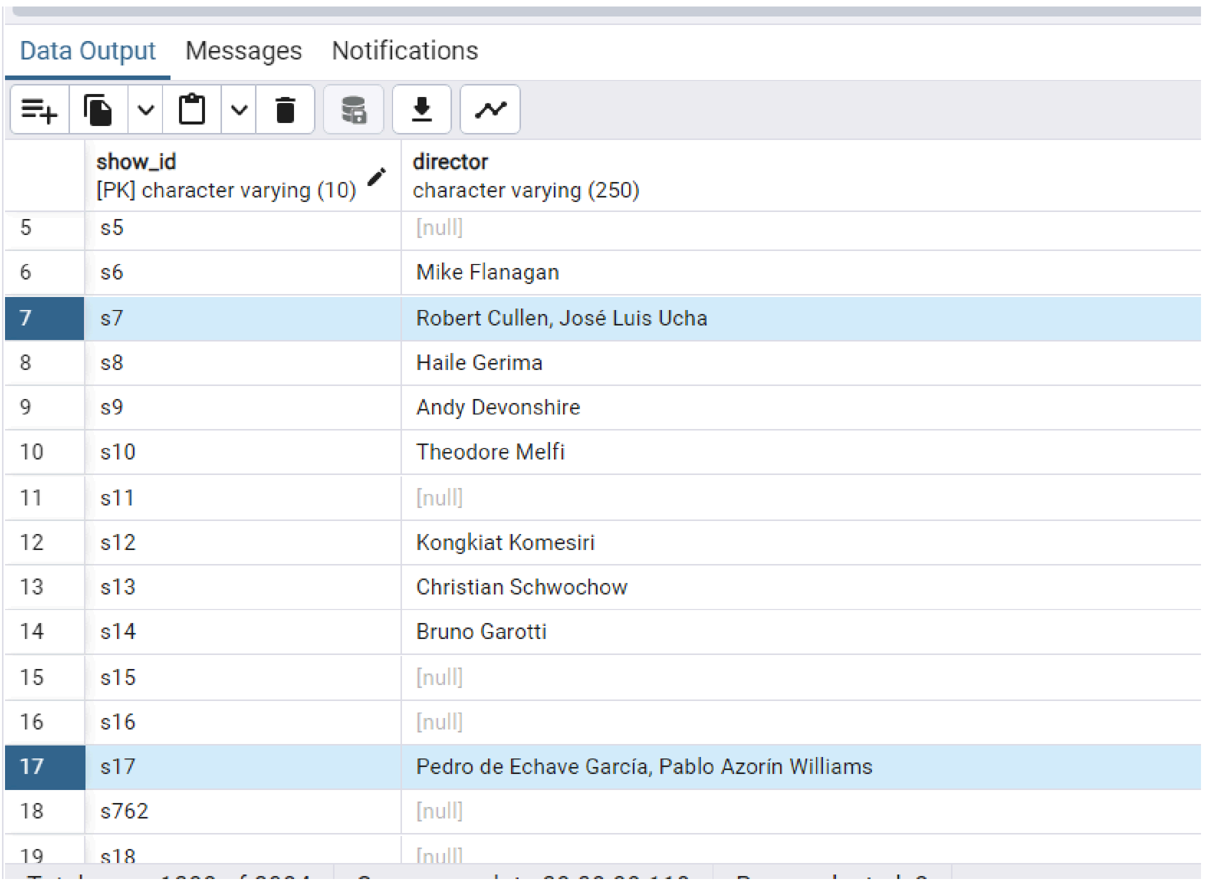
```
DELETE 3

Query returned successfully in 142 msec.
```



```
-- There are no duplicates for the column show_id beacuse it is primary key
--  Finding duplicates in title column
select * from netflix_raw where (upper(title),type) in (select upper(title),type from netflix_raw group by upper(title),
    type having count(*)>1)
```

show_id [PK] character varying (10)	type character varying (10)	title character varying (200)	director character varying (250)	cast character varying (1000)	country character varying (150)	date_added character varying (2

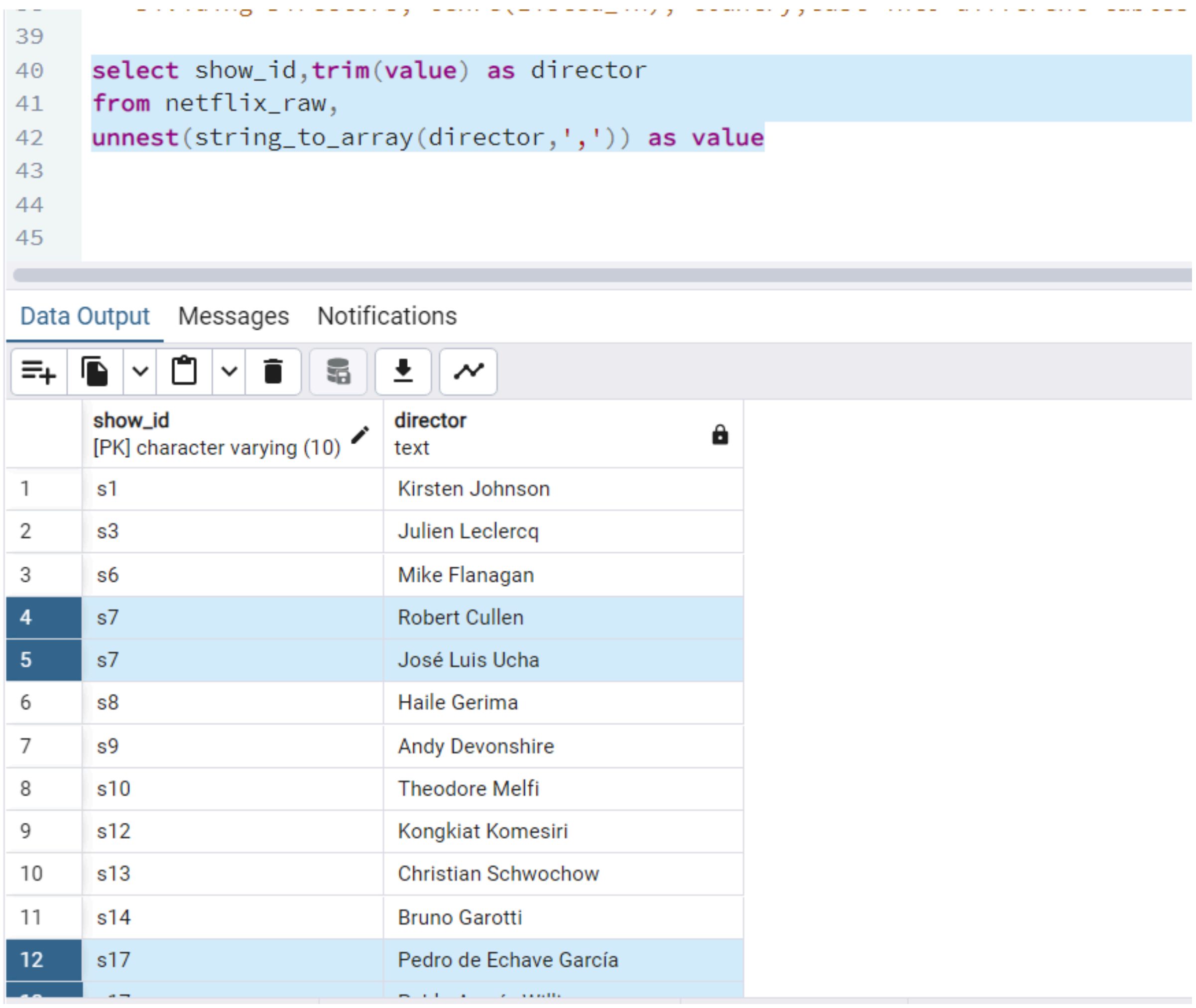
Dividing Directors, Genre(Listed_in), Country,Cast into different tables


Data Output Messages Notifications

	show_id [PK] character varying (10)	director character varying (250)
5	s5	[null]
6	s6	Mike Flanagan
7	s7	Robert Cullen, José Luis Ucha
8	s8	Haile Gerima
9	s9	Andy Devonshire
10	s10	Theodore Melfi
11	s11	[null]
12	s12	Kongkiat Komesiri
13	s13	Christian Schwochow
14	s14	Bruno Garotti
15	s15	[null]
16	s16	[null]
17	s17	Pedro de Echave García, Pablo Azorín Williams
18	s762	[null]
19	s18	[null]

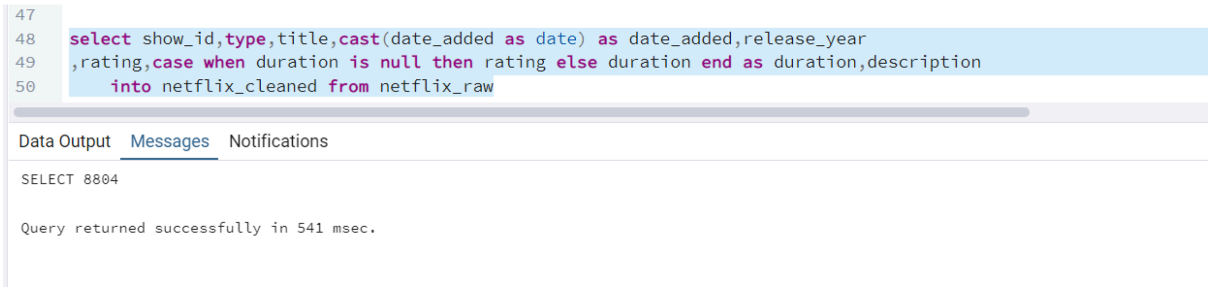
I could see for one show_id having multiple directors and I've decided to breakdown the directors into separate rows inorder to achieve normalization and store the separate rows into separate tables.

Not only for director column, Genre(Listed_in), Country,Cast columns having multiple things so divided into separate tables.



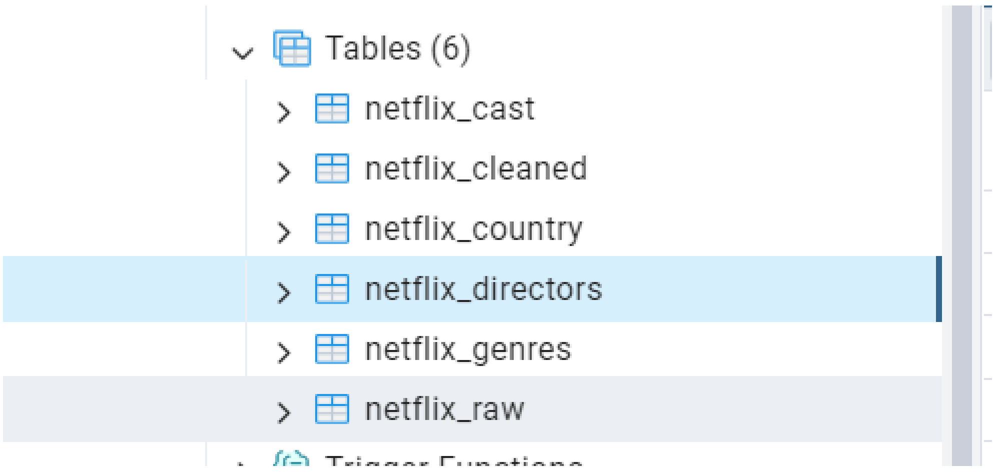
```sql
select show_id,trim(value) as director
from netflix_raw,
unnest(string_to_array(director,',')) as value
```

	show_id [PK] character varying (10)	director text
1	s1	Kirsten Johnson
2	s3	Julien Leclercq
3	s6	Mike Flanagan
4	s7	Robert Cullen
5	s7	José Luis Ucha
6	s8	Haile Gerima
7	s9	Andy Devonshire
8	s10	Theodore Melfi
9	s12	Kongkiat Komesiri
10	s13	Christian Schwochow
11	s14	Bruno Garotti
12	s17	Pedro de Echave García



```sql
select show_id,type,title,cast(date_added as date) as date_added,release_year
,rating,case when duration is null then rating else duration end as duration,description
   into netflix_cleaned from netflix_raw
```

```
SELECT 8804

Query returned successfully in 541 msec.
```



Analysing the data using SQL:

/*1 for each director count the no of movies and tv shows created by them in separate columns
for directors who have created tv shows and movies both */

--2 which country has highest number of comedy movies

--3 for each year (as per date added to netflix), which director has maximum number of movies released

--4 what is average duration of movies in each genre

--5 find the list of directors who have created horror and comedy movies both.

Netflix data analysis

/*1 for each director count the no of movies and tv shows created by them in separate columns
for directors who have created tv shows and movies both */

```
	select nd.director,count(distinct case when type='Movie' then n.show_id end) as
no_of_movies,
	count(distinct case when type='TV Show' then n.show_id end) as no_of_shows from
netflix_directors nd join netflix_cleaned n on n.show_id = nd.show_id
	group by nd.director
	having count(distinct n.type)>1

	select type from netflix_cleaned where show_id in(select show_id from
netflix_directors where director='Abhishek Chaubey')
```

--2 which country has highest number of comedy movies

```
select nc.country,count(ng.show_id) as no_of_movies from netflix_genres ng join
netflix_country nc on ng.show_id=nc.show_id
	join netflix_cleaned n on n.show_id=ng.show_id
	where ng.genre='Comedies' and n.type='Movie'
	group by nc.country
	order by no_of_movies desc limit 1
```

Comedies

--3 for each year (as per date added to netflix), which director has maximum number of movies released

```
select * from(
SELECT nd.director,extract(year from n.date_added) AS yr,count(n.show_id) as cnt from
netflix_directors nd
join netflix_cleaned n on n.show_id=nd.show_id
where n.type='Movie'
group by nd.director,extract(year from n.date_added)
order by nd.director,yr) t order by cnt desc

select * from netflix_cleaned where show_id in (select show_id from netflix_directors where
director='S.S. Rajamouli')

--4 what is average duration of movies in each genre

select ng.genre,round(avg(cast(replace (n.duration,' min','')as int)),0) as avg_duration from
netflix_cleaned n join netflix_genres ng on n.show_id=ng.show_id where type='Movie'
group by ng.genre order by ng.genre


--5  find the list of directors who have created horror and comedy movies both.
-- display director names along with number of comedy and horror movies directed by them

select nd.director,count(distinct case when ng.genre='Horror Movies' then ng.show_id end)
as no_of_horror_movies,
          count(distinct case when ng.genre='Comedies' then ng.show_id end) as
no_of_comedy_movies
          from netflix_genres ng join netflix_directors nd on ng.show_id=nd.show_id
          join netflix_cleaned n on n.show_id=ng.show_id where n.type='Movie' and ng.genre
in ('Horror Movies','Comedies')
group by nd.director having count(distinct ng.genre)=2
```