**PROJECT BRIEF –**

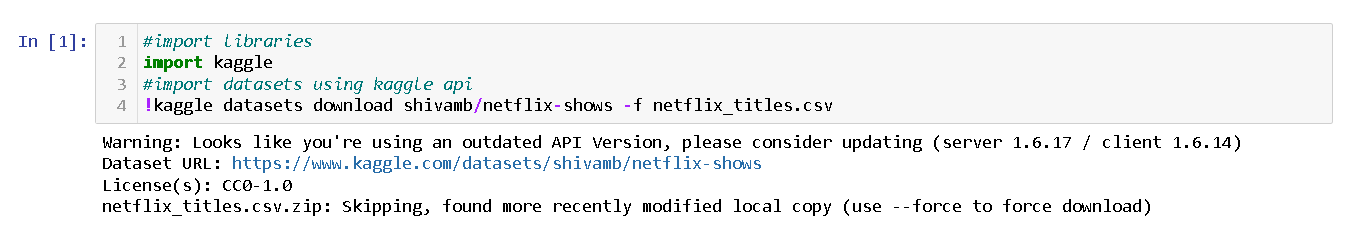
Netflix data need to be imported from kaggle api to pandas framework and load the data to MySQL and perform data cleaning and finding below objectives.

Requirements

1. Remove Duplicates
2. Handling foreign characters
3. New table for listed in director, country, cast
4. Populate missing values in country, duration columns
5. Populate the rest of the nulls as not\_available
6. Drop column director, listed\_in country, cast

**Solution Approach:-**

1. Generated new token and saved in the user file in C drive under .kaggle folder
2. Workings in Jupyter notebook where we use python libraries to fetch the data from kaggle.
   1. Install kaggle (!pip install kaggle )
   2. import kaggle
   3. Download dataset using kaggle api



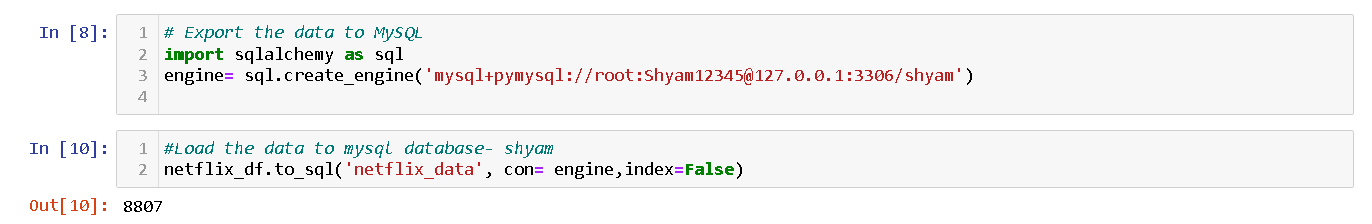
* 1. Extract file from zip file



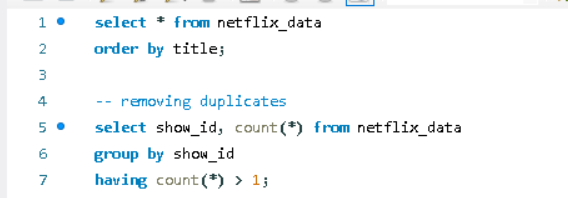
* 1. Import pandas data frame, load & read the data

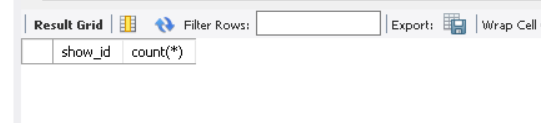


* 1. Export the data into MySQL



1. **My SQL**
   1. Remove Duplicates



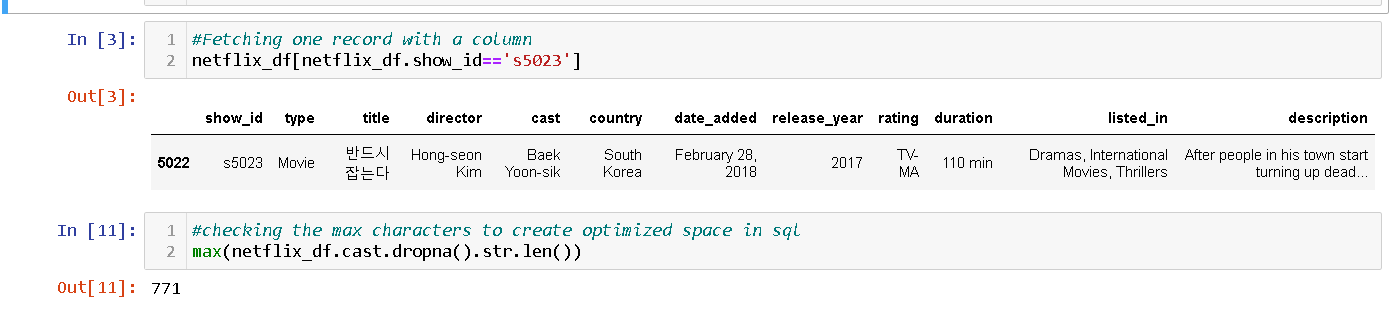


Checking at other column eg- title uppper lower making duplicate movies/shows as same.

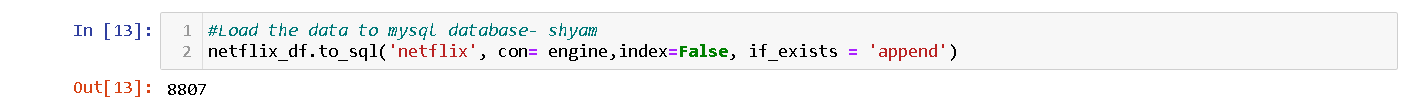
But below issue occurred so trying to create new table with same columns but optimized spaces then importing the data from python by using if\_exists =’append’



Checking the max characters in each column



Loading using append method to sql db



Create table netflix(

Show\_id varchar(10),

Type varchar(10),

Title varchar(200),

Director varchar(250),

Cast varchar(1000),

Country varchar (150),

Date\_added varchar(20),

Release\_year int,

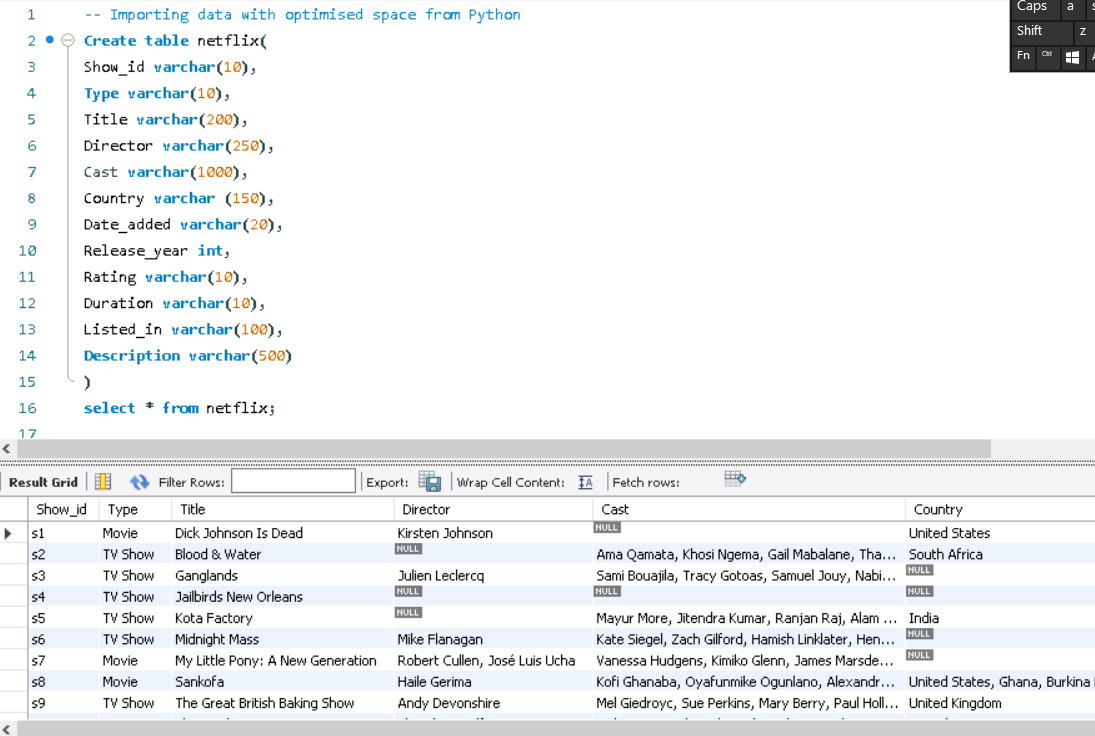
Rating varchar(10),

Duration varchar(10),

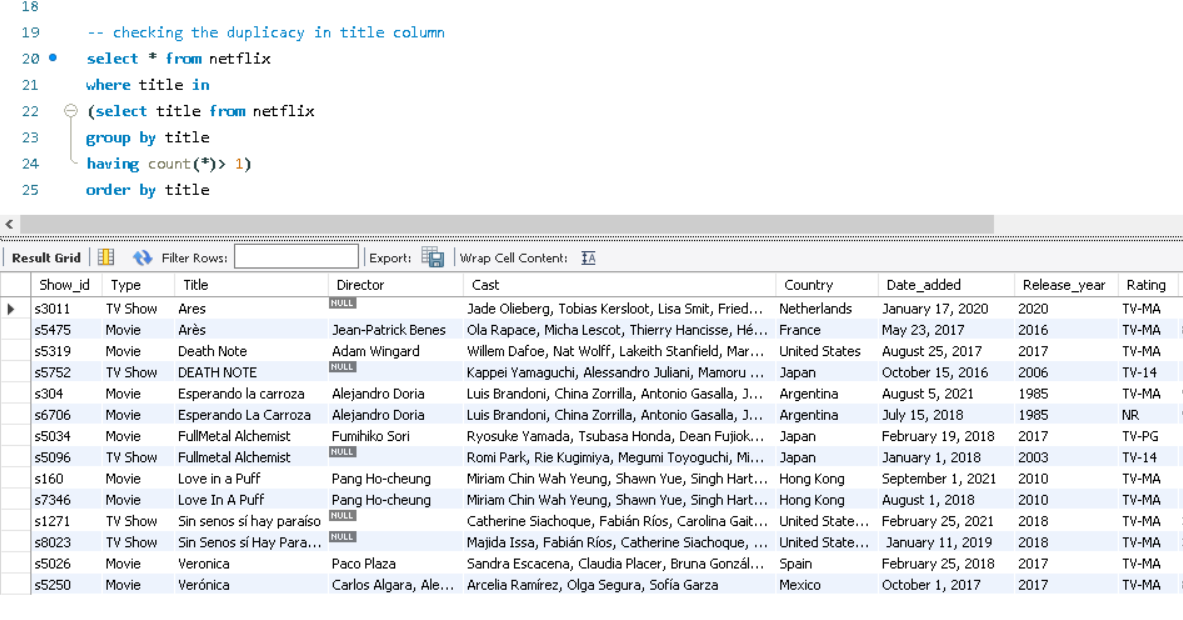
Listed\_in varchar(100),

Description varchar(500)

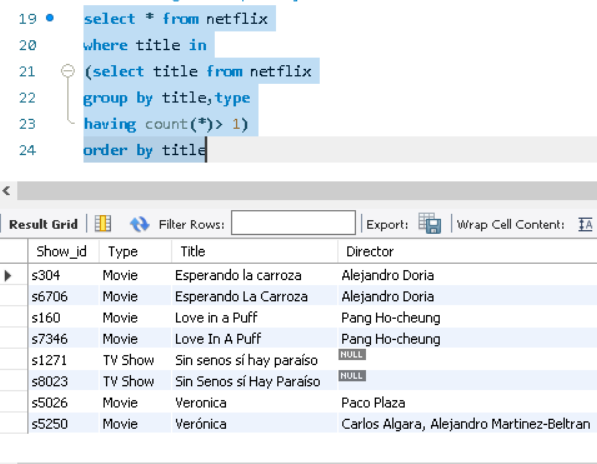
)



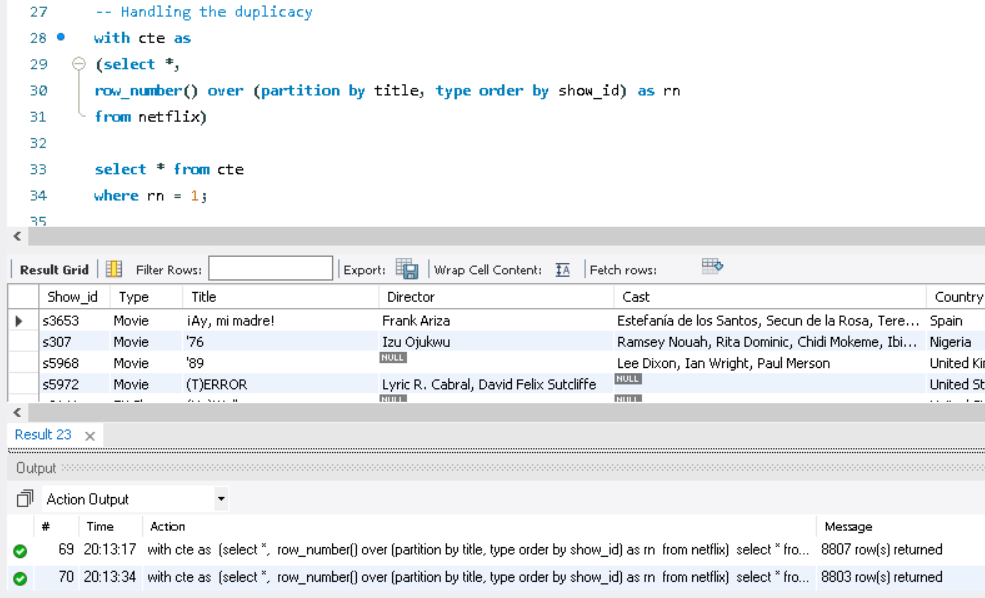
Title level duplicacy check

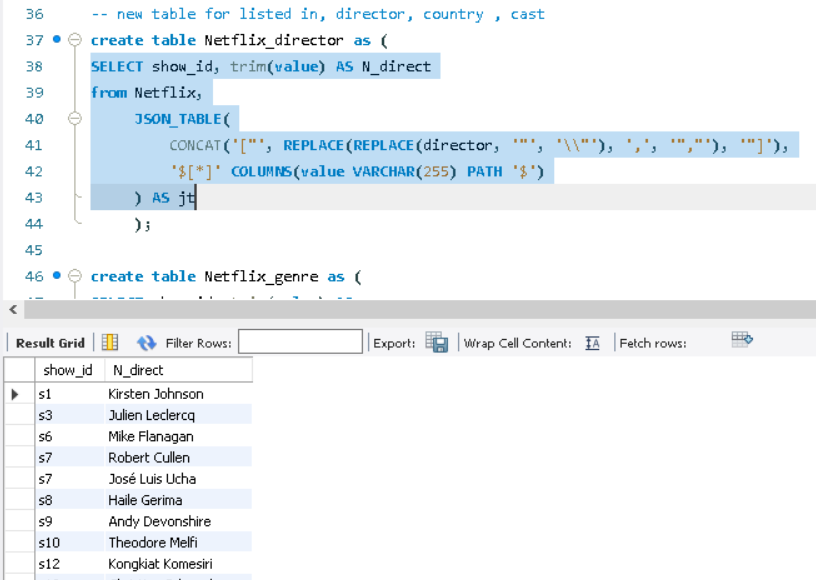


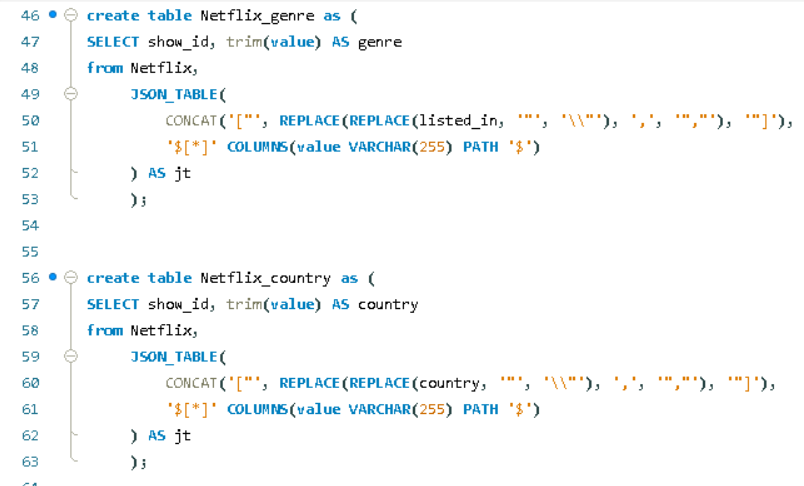
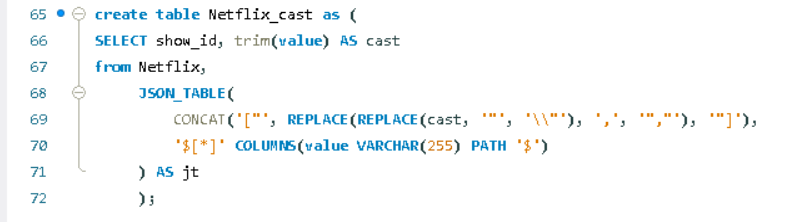
Now grouping by title and type

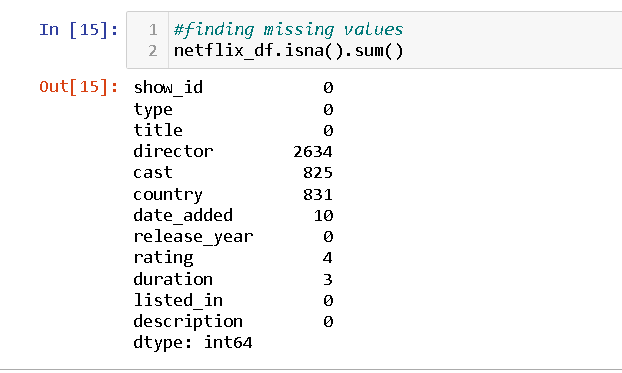


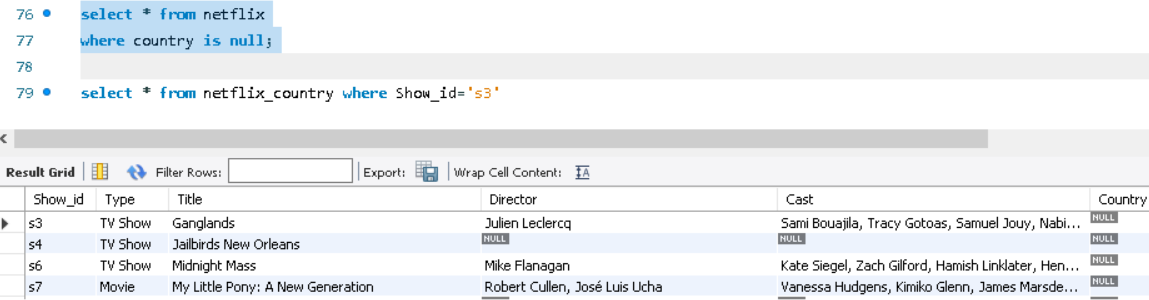
Finally removing the duplicacy – 4 duplicates removed

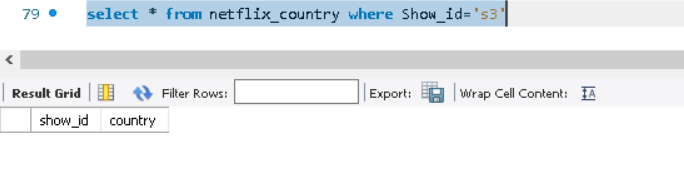




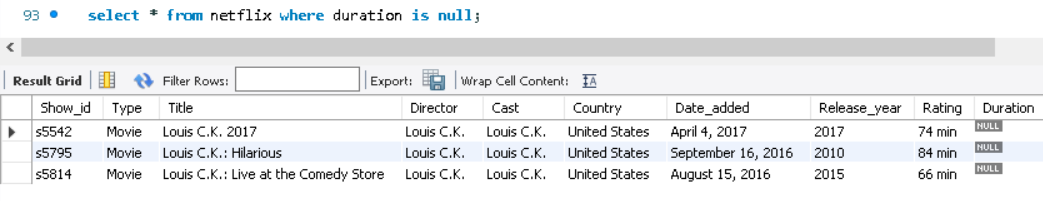
 

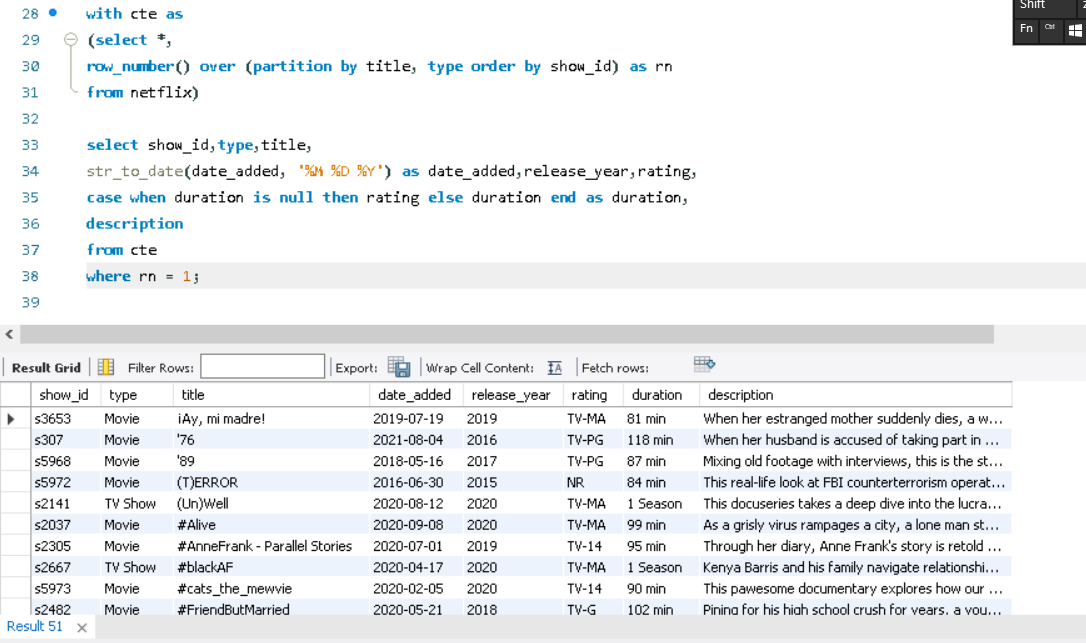


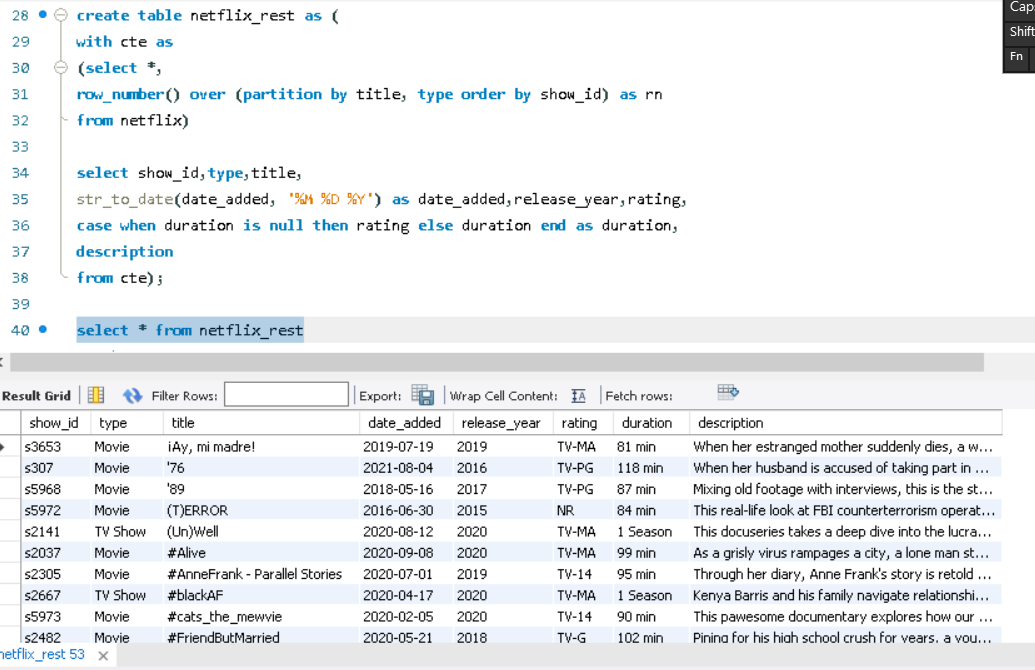




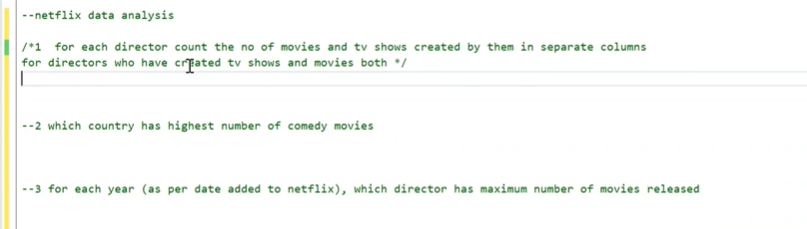






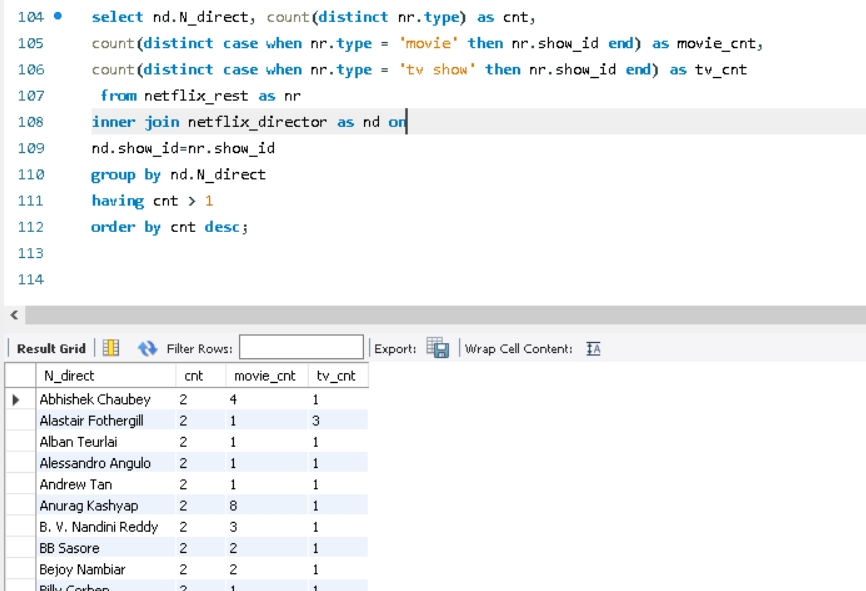


* 1. Handling foreign characters
  2. New table for listed in director, country, cast
  3. Populate missing values in country, duration columns
  4. Populate the rest of the nulls as not\_available
  5. Drop column director, listed\_in country, cast

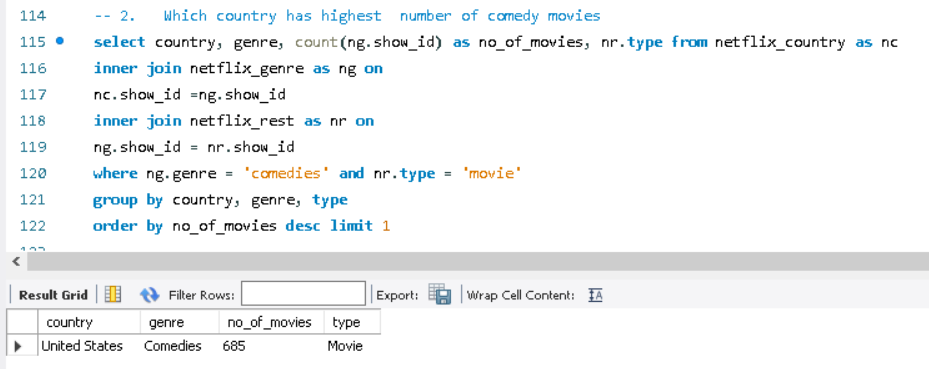
****

**Netflix data analysis**

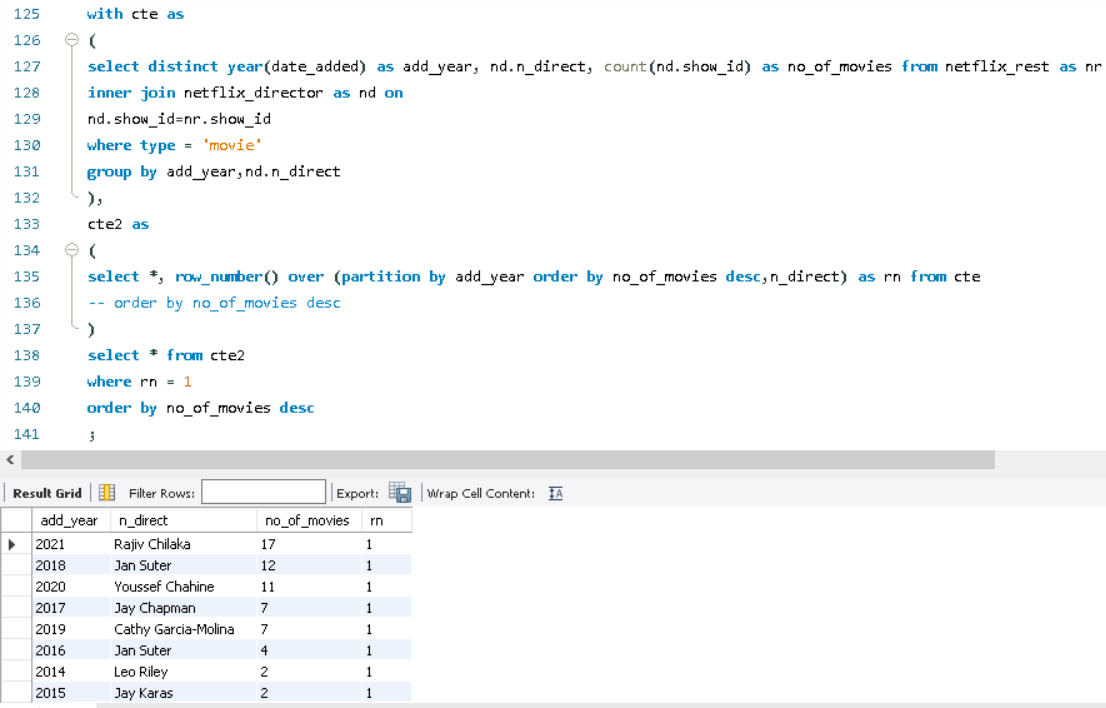
1. **For each director count the no of movies and tv shows created by then in separate columns for directors who have created tv shows and movies both**

****

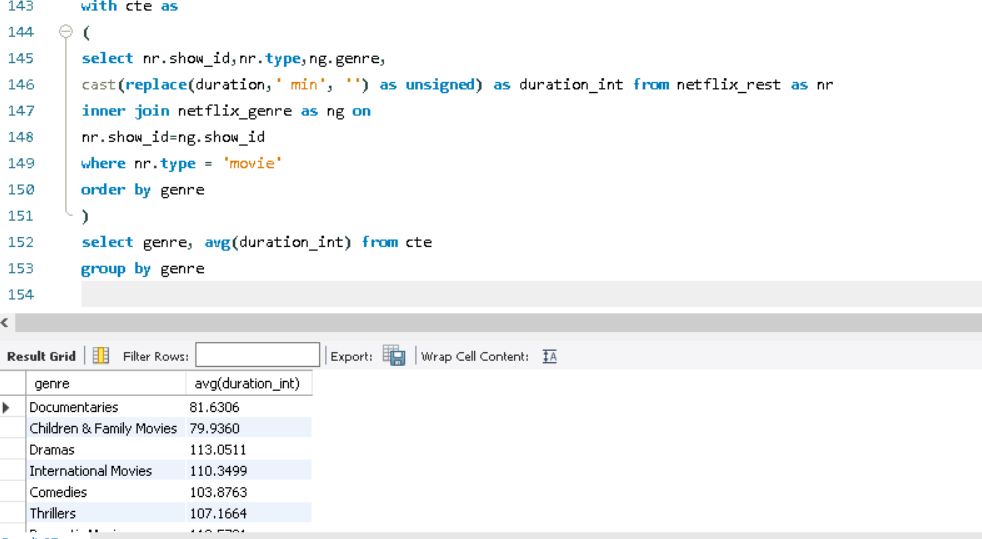
1. **Which country has highest number of comedy movies**

****

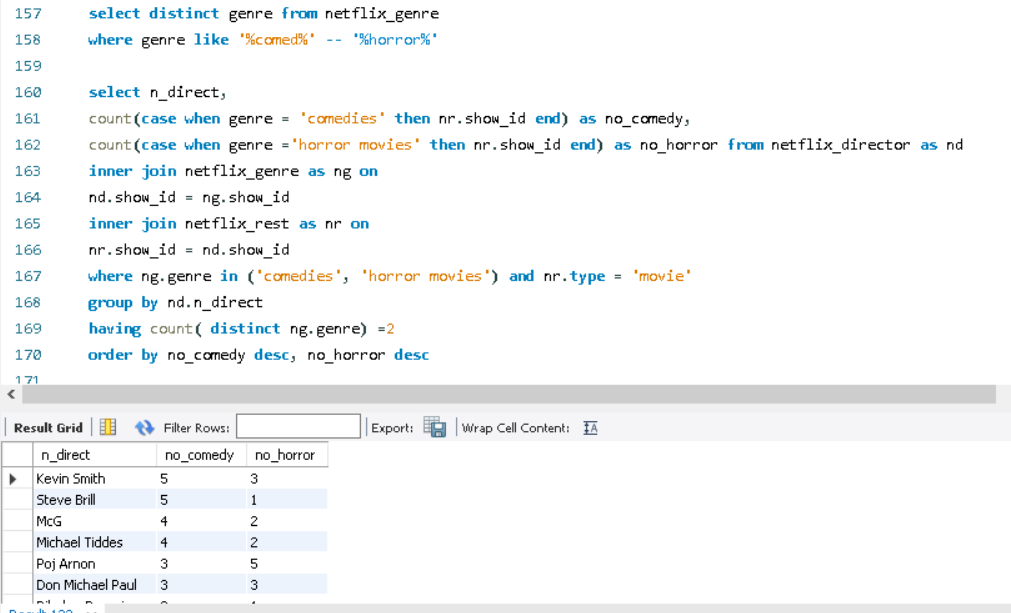
1. **For each yeaer ( as per date added to netflix), which director has max number of movies released**

****

1. **What is the average duration of movies in each genre**

****

1. **Find the list of directors who have created both comedy and horror movies. Display director names along with the no. of comedy and horror movies directed by them**

****