**Colab NoteBook Link  
  
https://colab.research.google.com/drive/1DanRiqj0DPM4n\_1W75iRPgmYUBhBAg-k?usp=sharing  
  
PDF Notebook Link**[**https://drive.google.com/file/d/1sFx2Ez\_2QDFNAkL212CyzDAXLXvK6CvQ/view?usp=sharing**](https://drive.google.com/file/d/1sFx2Ez_2QDFNAkL212CyzDAXLXvK6CvQ/view?usp=sharing) **About NETFLIX**

Netflix is one of the most popular media and video streaming platforms. They have over 10000 movies or tv shows available on their platform, as of mid-2021, they have over 222M Subscribers globally. 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.

**Business Problem**

Analyze the data and generate insights that could help Netflix ijn deciding which type of shows/movies to produce and how they can grow the business in different countries

**Dataset**

The dataset provided consists of a list of all the TV shows/movies available on Netflix:

**Show\_id:** Unique ID for every Movie / Tv Show  
**Type:** Identifier - A Movie or TV Show  
**Title:** Title of the Movie / Tv Show  
**Director:** Director of the Movie  
**Cast:** Actors involved in the movie/show  
**Country:** Country where the movie/show was produced  
**Date\_added:** Date it was added on Netflix  
**Release\_year:** Actual Release year of the movie/show  
**Rating:** TV Rating of the movie/show  
**Duration:** Total Duration - in minutes or number of seasons  
**Listed\_in:** Genre  
**Description:** The summary description

**1. Defining Problem Statement and Analysing basic metrics.**

 We basically have to predict what kind of TV shows/movies to produce in order to increase the company's business.

 Following can be some approaches:

1.  In the last 10 years, which Genre(will be splitted from listed\_in column ) movies/TV Shows is **watched**more on Netflix (can be analyzed by duration)

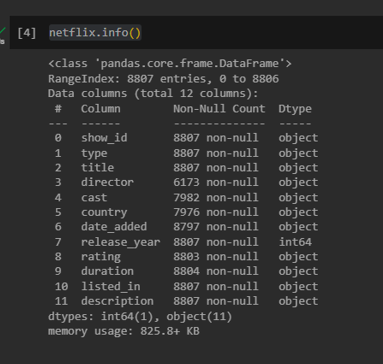
* For Movies in minutes
* For TV shows in terms of the number of seasons (Assumption all Season have the same number of episodes and each episode is of the same time interval. Hence MORE NUMBER of SEASONS mean MORE WATCH TIME)
* Who are the directors, cast worked in this genre and in which country were the movies/TV Shows produced.

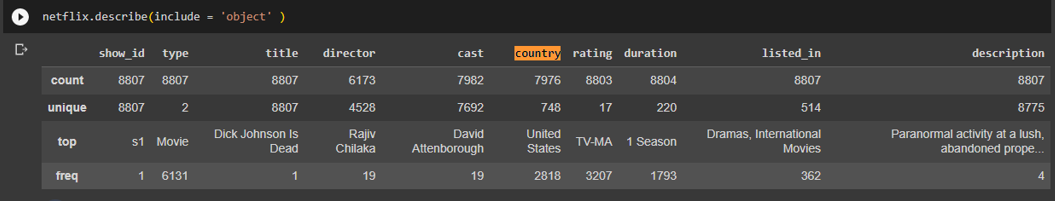
2.  In the last 10 years, which Genre(will be splitted from listed\_in column ) movies/TV Shows is**added more** by Netflix.

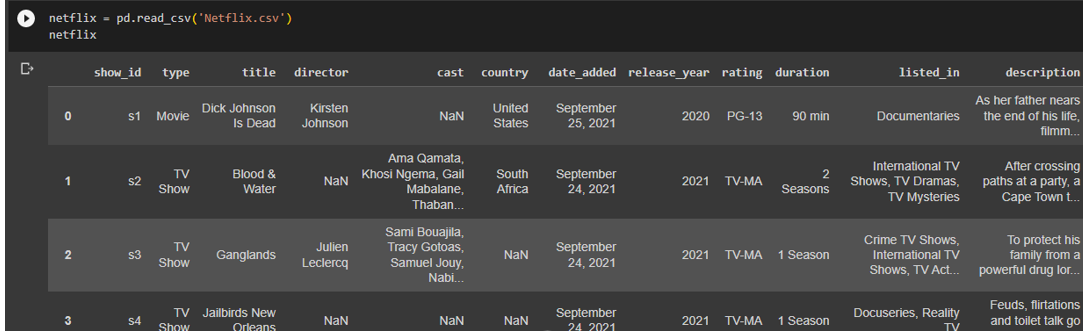
* Who are the directors, cast worked in this genre and in which country were the movies/TV Shows produced.

**Basic Metrics**:

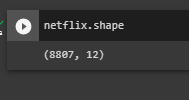
* There are 8807 distinct shows data present in the dataset.
* Shape of the data is (8807,12)
* There are null values present in Column director, cast, country, date\_added, rating, duration
* There are comma-separated values present in column director, cast, listed\_in, country
* There are two types of shows present **TV Shows and Movie**



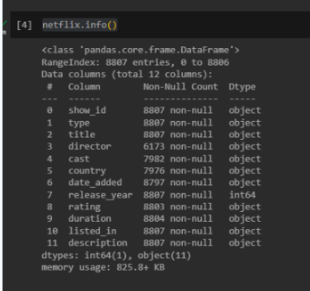


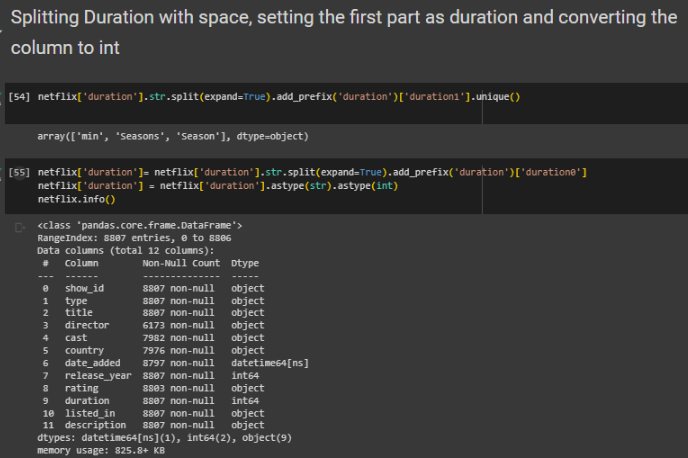


**222. Observations on the shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), missing value detection, statistical summary**

* Shape of the data is (8807,12). There are 8807 distinct shows data present in the dataset.  
    
  
* All are **object**types (**as most**contain **string**) except **release\_year**. Also **date\_added**is an object type but we have converted it into **datetime**for ease of use by the below code.

netflix['date\_added'] = pd.to\_datetime(netflix['date\_added'].str.strip(),format='%B %d, %Y')



* Made **duration**column **int**after processing.  
    
  
* For Column **listed\_in , s**plit values on comma, there are no missing values as per .info(), melt it, strip individual genre names to remove whitespaces.

netflix['listed\_in'] = netflix['listed\_in'].str.split(',')

#IMPORTANT Didnot used fillna because there is no null for listed\_in column as infered by .info method in previous cell

spgenredf = pd.DataFrame(netflix['listed\_in'].tolist()).add\_prefix('genre\_')

len(max(netflix[~netflix['listed\_in'].isna()]['listed\_in'], key=len))

# Number of columns is as per max number of listed\_in we have for a movie in the dataset.

netflixWithspGenre = pd.concat([netflix,spgenredf],axis =1)

netflixWithspGenre = pd.melt(netflixWithspGenre,id\_vars=netflix.columns.tolist())

netflixWithspGenre.rename({'value':'genre'},axis = 1,inplace=True)

# IMPORTANT : For movies where we have less number of listed\_in than max number of listed\_in we have for a movie, None is added

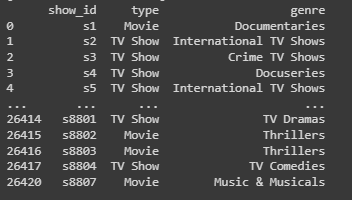
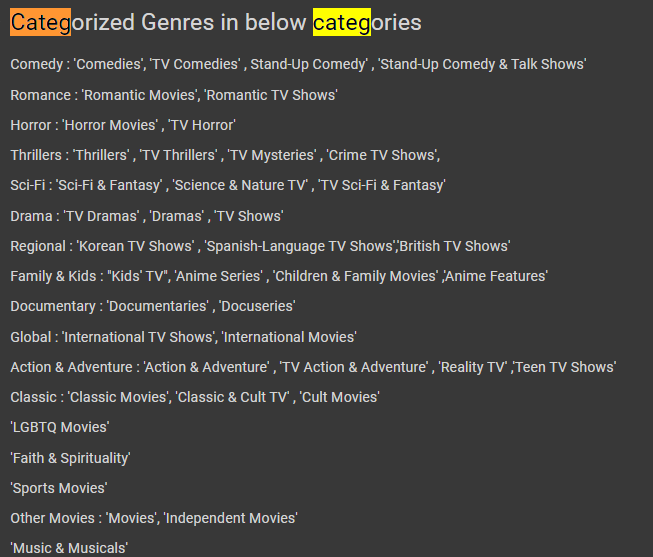
# IMPORTANT : After melting (multiple genre column will be transformed in a single column named genre) we will remove the rows having None as value for listed\_in because they were added due to above step and is not needed

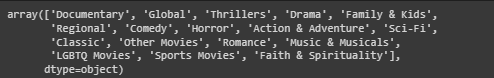
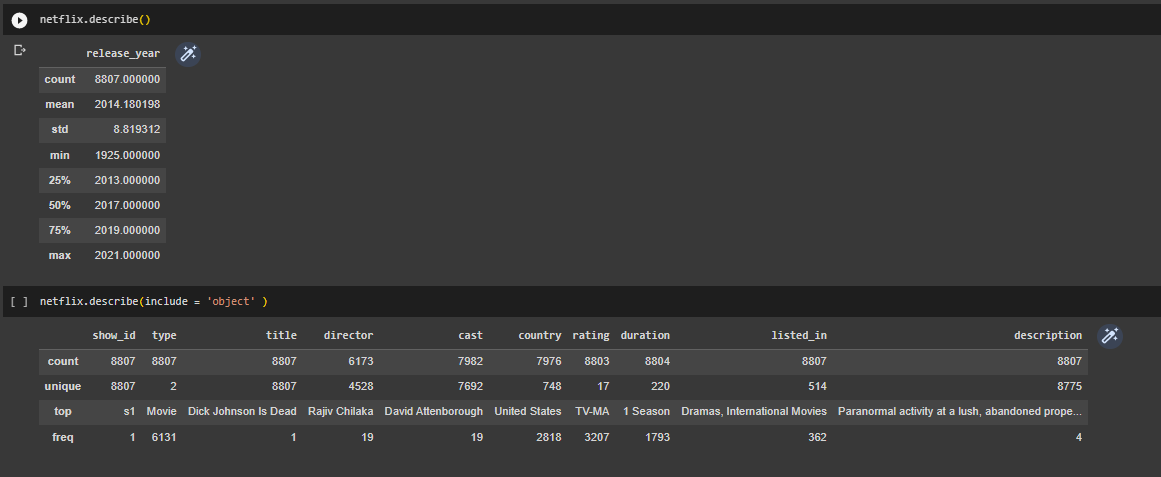
netflixWithspGenre = netflixWithspGenre[~netflixWithspGenre['genre'].isna()]

netflixWithspGenre = netflixWithspGenre.loc[:,['show\_id','type','genre']]

# Strip out Starting and Trailing WhiteSpaces

netflixWithspGenre['genre'] = netflixWithspGenre['genre'].str.strip()

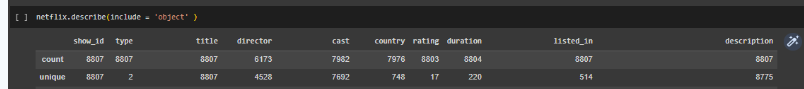
  
  
After this operation, **listed\_in**which is converted into **genre**column name is categorized like below. **The code to convert is present in the attached PDF.**  
  


* **Unique Converted Genres** will look something like this.  
    
  
* As per .info()and isna() there are**missing values** present in Column**director, cast, country, date\_added, rating, duration.**
* **Statistical summary**  
    
  

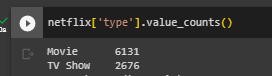
â€‹â€‹3. **Non-Graphical Analysis: Value counts and unique attributes**

     Before doing any preprocessing below is the count and unique attributes count in the dataset.

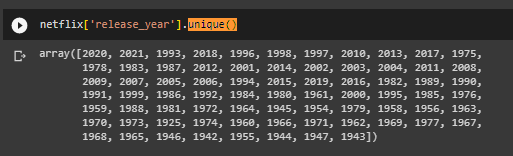
* There are 2 types of show**'TV Show'**and '**Movie'**



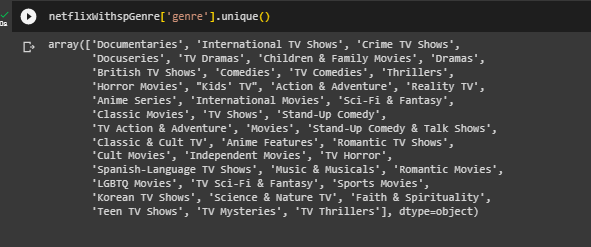
* There are **6131**Movies and **2676**TV Shows in the dataset.



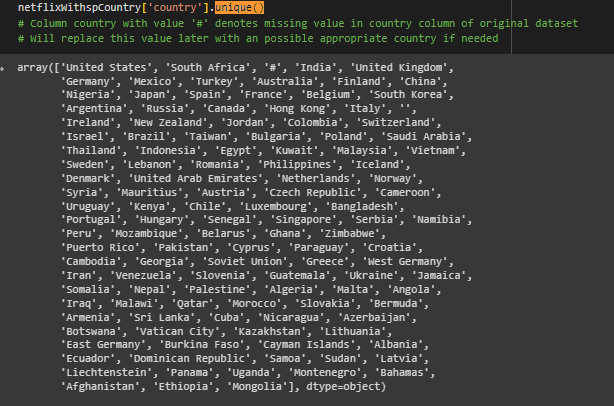
* **Unique release year**



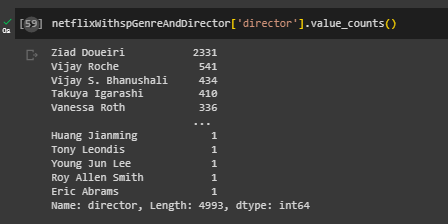
* **Unique Genres** after processing comma separated listed\_in( Renamed column to the genre )



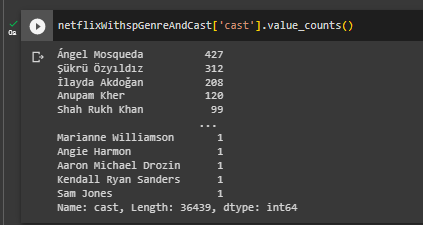
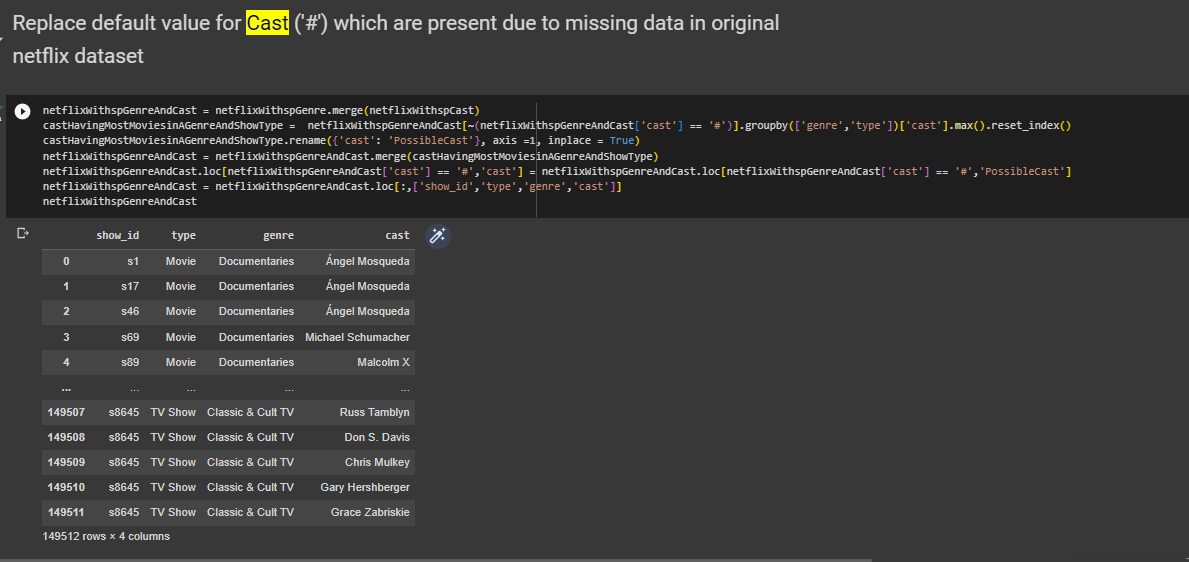
* **Unique Countries, # denote the missing value**



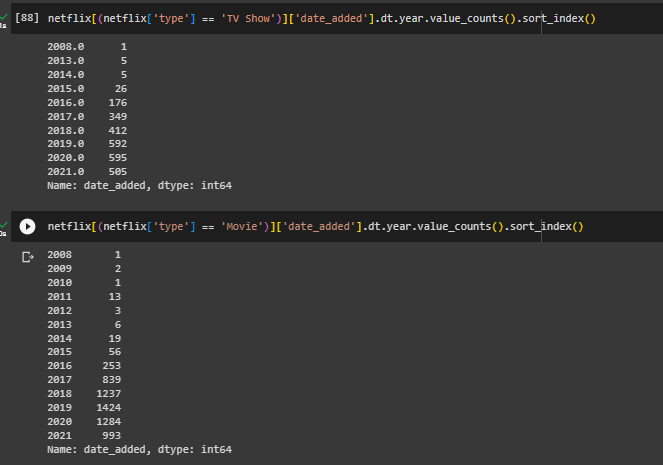
* After **processing the director column** and filling in the missing values with the director having max movies in the missing data row show genre and show Type, below is the**director's name and the count of shows they have directed**



* After **processing the cast column** and filling in the missing values with the cast having max movies in the missing data row show genre and show Type, below is the **cast name and the count of shows they have acted in**

* Most of the TV Shows and Movies is added after 2014 on Netflix.

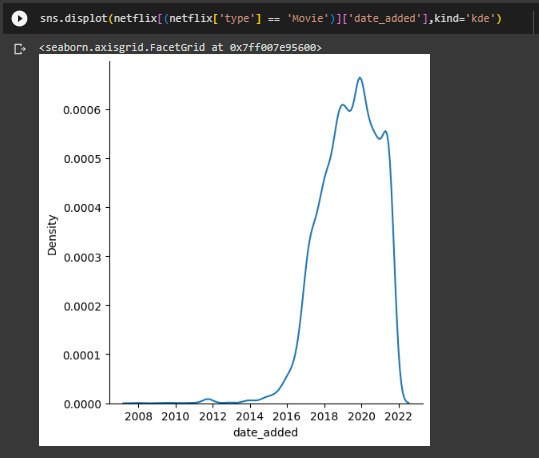


4.  **Visual Analysis - Univariate, Bivariate after pre-processing of the data**

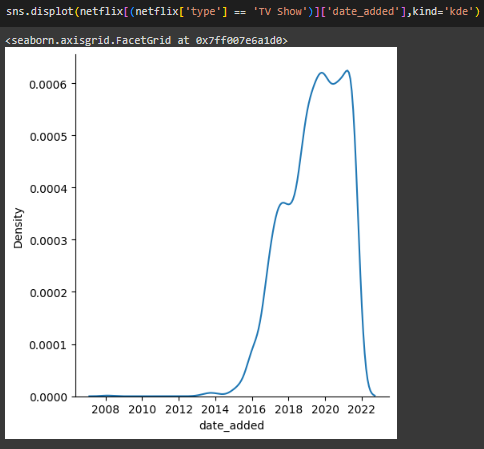
**Note: Pre-processing involves unnesting the data in columns like Actor, Director, Country**

4.1 For continuous variable(s): Distplot, countplot, histplot for univariate analysis   
\*Covering **6.1, 6.2. 6.3** as well.

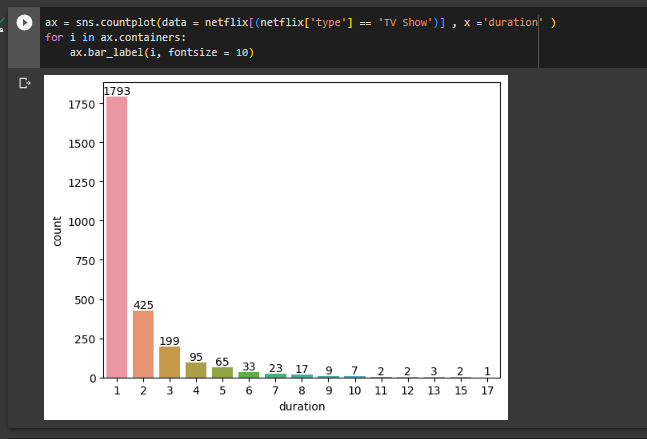
* **Movies are added between the years 2008-2021**
* **Most of the Movies are added in the year range 2014 - 2021**
* **There is a dip in the number of movies in the year 2021**

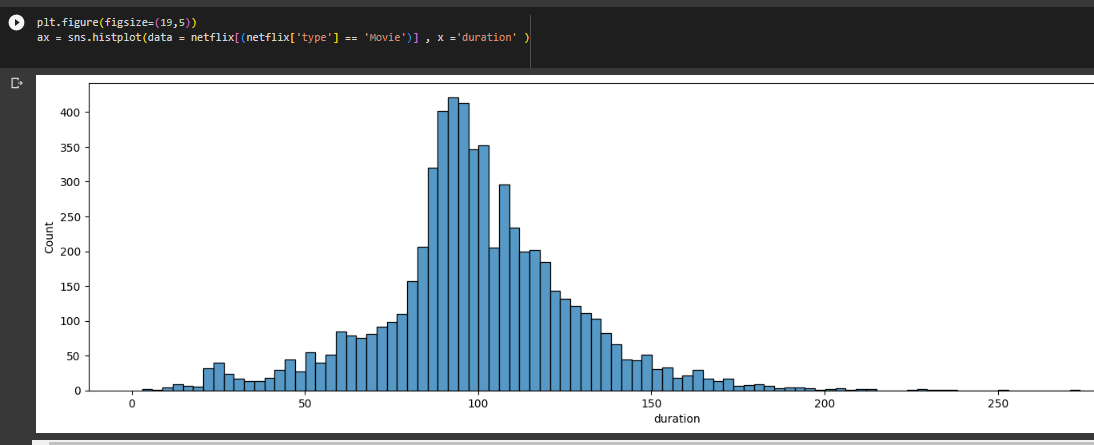


* **TV Shows are added between the years 2008-2021**
* **Most of the TV Shows are added in the year range 2014 - 2021**



* **Most TV Shows are having seasons from 1 - 5. Max Seasons are 17**

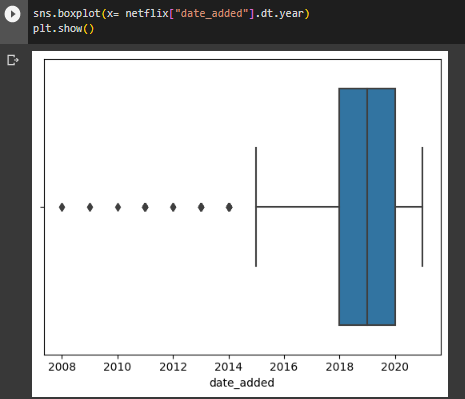


* **Movies duration range is around 20-200 mins. Most movies are in the duration range of 50-150 mins.**
* 

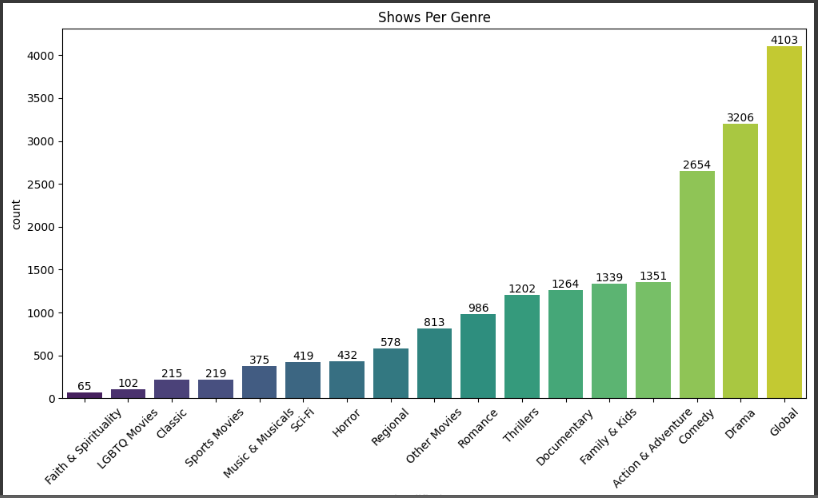
4.2 For categorical variable(s): Boxplot

Covering **6.1, 6.2. 6.3** as well

* Shows are added on the platform in the year range of 2008-2021 with a median of the total number of shows around 2018-2019.



* **Shows Per Genre**  
  **The top 3 Genres** (Movie and TV Show) are **Global**('International TV Shows', 'International Movies'), **Drama**('TV Dramas', 'Dramas', 'TV Shows'), **Comedy** ('Comedies', 'TV Comedies', 'Stand-Up Comedy', 'Stand-Up Comedy & Talk Shows')  
    
  **NOTE**: One show can belong to multiple genres since the listed\_in column had multiple comma-separated genre categories (Which we have split and categorized for simplicity)

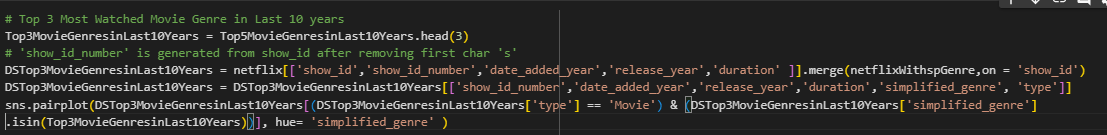
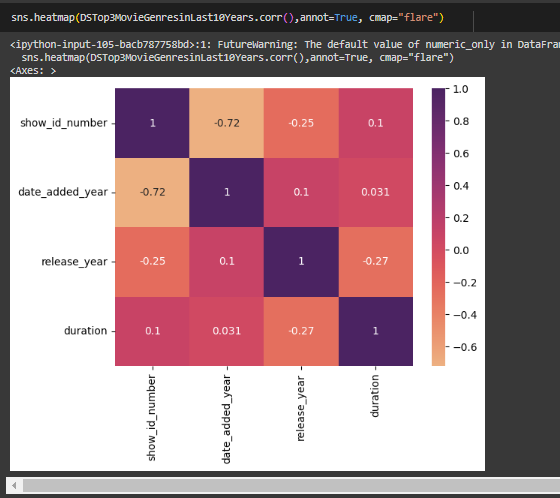


4.3 For correlation: Heatmaps, Pairplots

Covering **6.1, 6.2. 6.3** as well  
  
**Movies Details as per Top 3 Most Watched Movie Genre in Last 10 years using PairPlot**

* **Global**Category Movie is watched most, followed by**Drama and Comedy.**
* **Movies** are added in Year Rang**e 2008-2021 on netflix**
* **Movies a**re released in Year Range**1942-2021**
* **Comedy** Movie Duration is less than **100 min mostly**

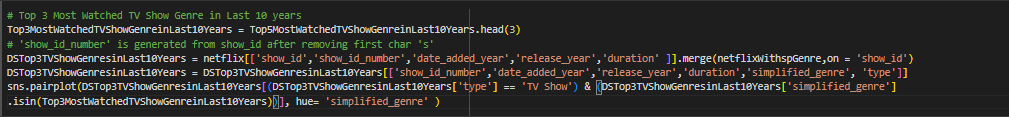
**NOTE**: One show can belong to multiple genres since the listed\_in column had multiple comma-separated genre categories (Which we have split and categorized for simplicity)

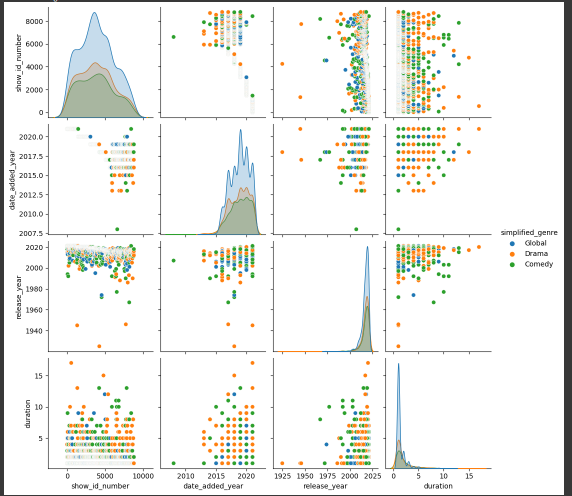
  
  


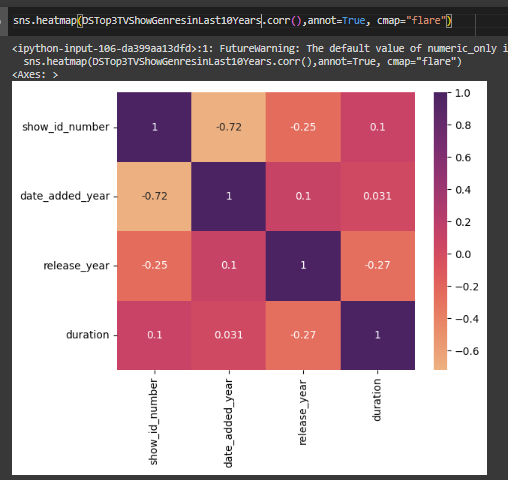
**TV Show Details as per Top 3 Most Watched TV Show Genre in Last 10 years using PairPlot**

* + **Global Category Movie is watched most, followed by Drama and Comedy.**
  + **TV Shows are added in Year Range 2008-2021 on netflix**
  + **TV Shows are released in Year Range 1925-2021**
  + **Most TV Show have 7 or less seasons.**

**NOTE**: One show can belong to multiple genres since the listed\_in column had multiple comma-separated genre categories (Which we have split and categorized for simplicity)

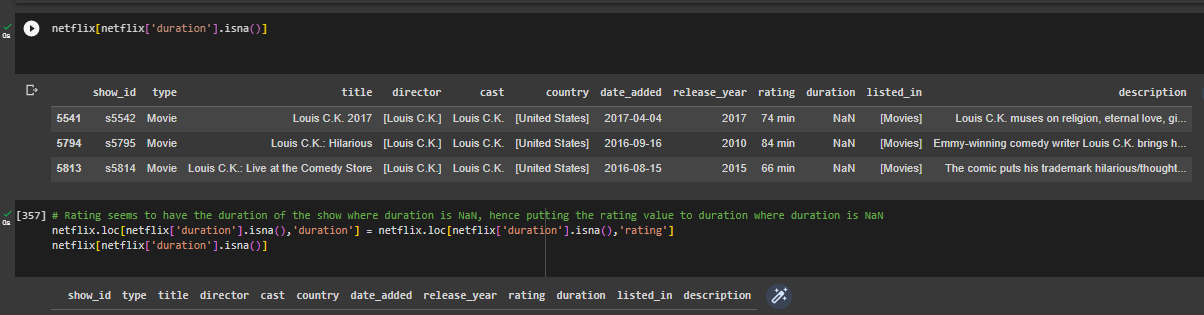
* 



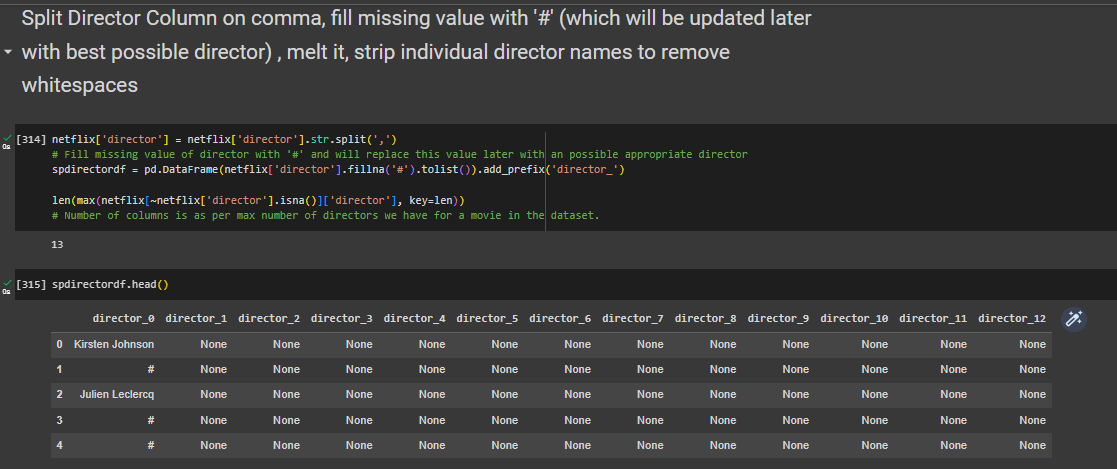
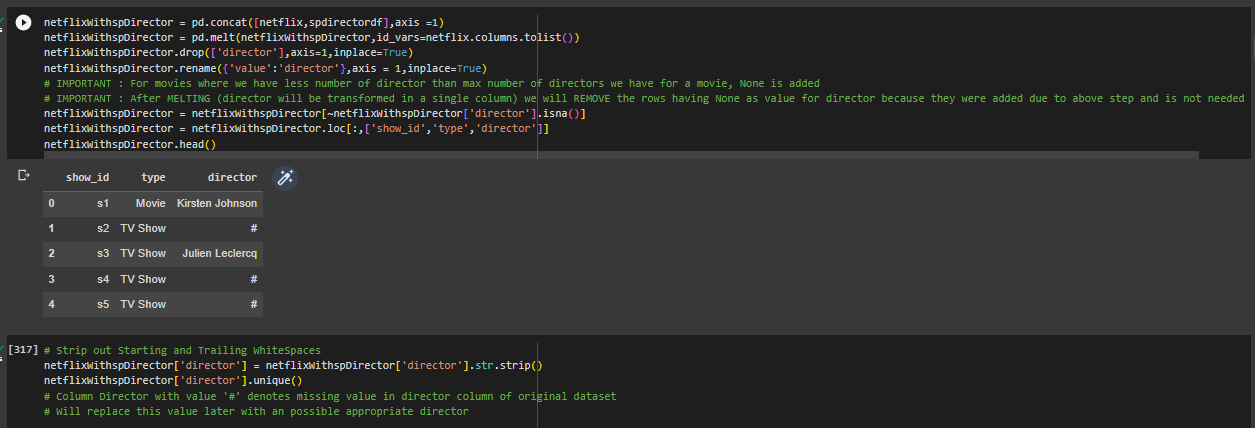
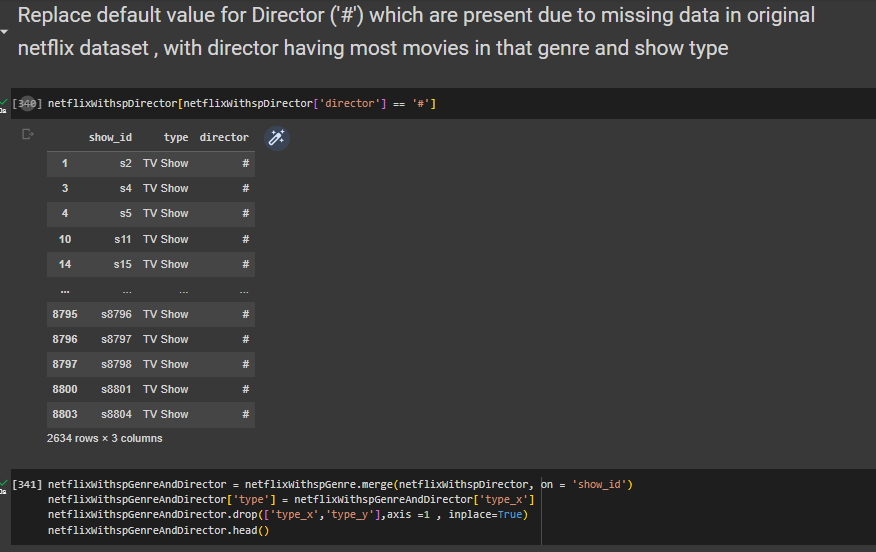
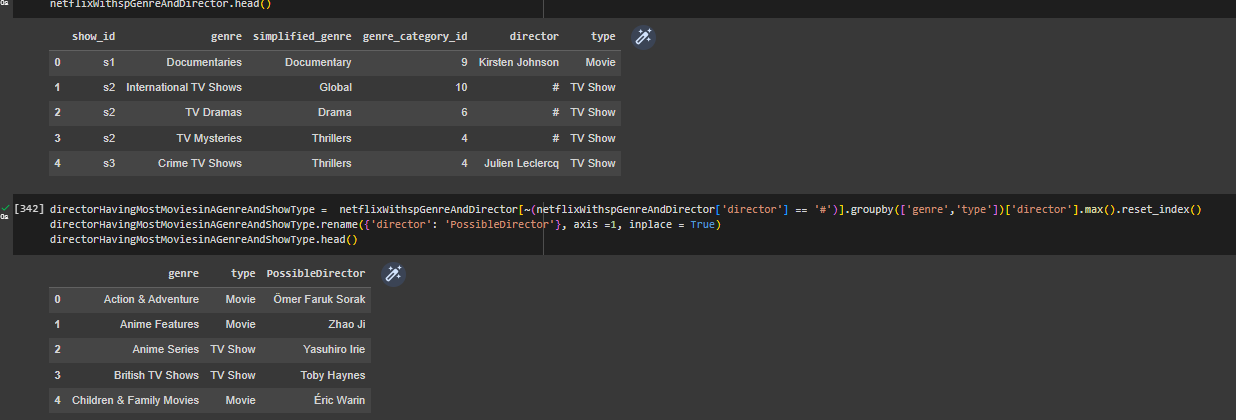
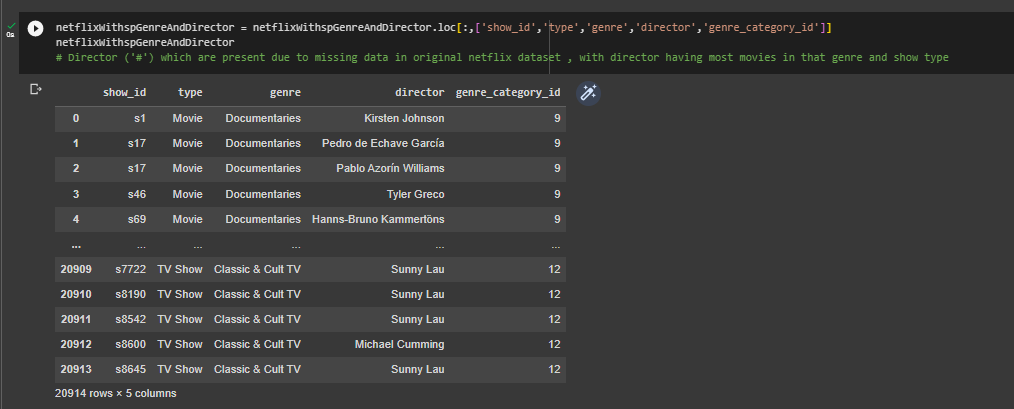
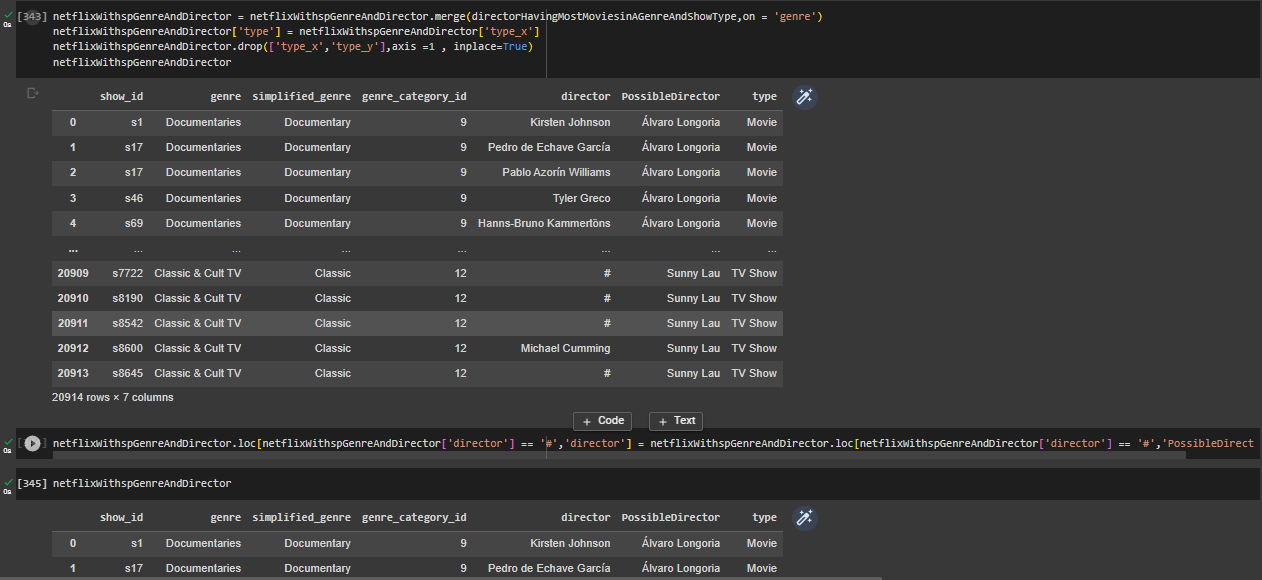
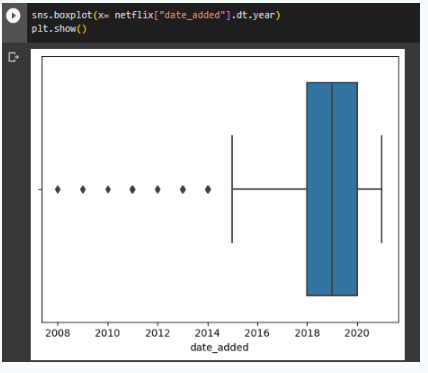


**5. Missing Value & Outlier check (Treatment optional)**

As per .info()and isna() there are**missing values** present in Column**director, cast, country, date\_added, rating, duration.**



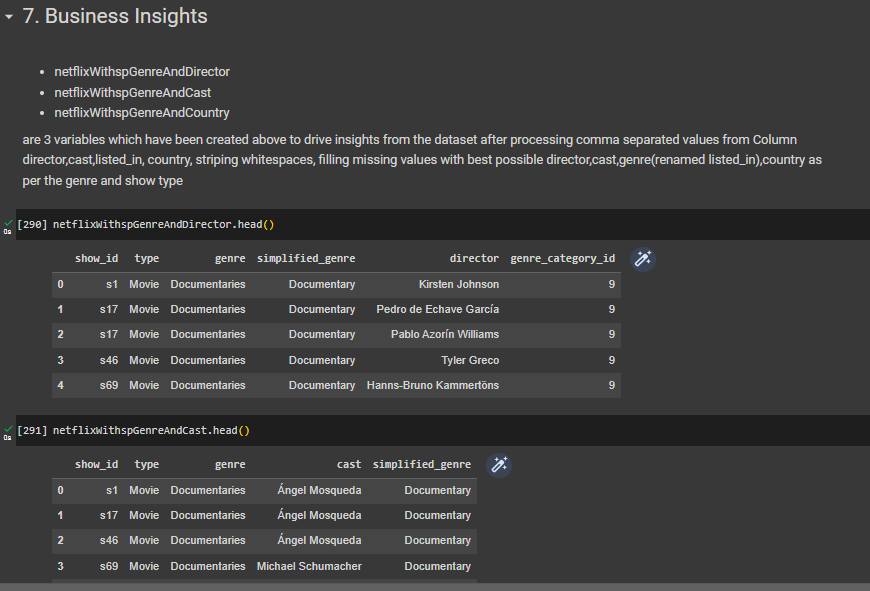
* **Missing Value is populated for column director, cast, country (Details present in attached pdf)**

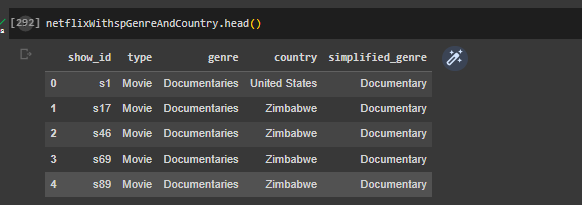
For **director, cast, country column** , after processing the comma-separated values, the **missing value**s which are initially replaced with #,  are replaced by the top director / cast member/country name as per the missing data row show genre, and show type.  
  
e,g, For the director  
  
  
  
  
  
  
  
Similarly, it's done for**cast and country** columns. **Details are available in the attached PDF and in the Colab Notebook.**  
  
**Outlier**  
  
As we can see the shows are added mostly after the year 2014   


6.

* Insights based on Non-Graphical Analysis covered in point 3
* Insights based on Visual Analysis covered in point 4.

7. **Business Insights**





* **TOP 5 Genre in Which Movies Are Added In the Last 10 Years**

temp = netflix[['show\_id','date\_added\_year']].merge(netflixWithspGenre,on = 'show\_id')

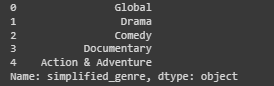
temp = temp.loc[:,['show\_id','date\_added\_year','type','simplified\_genre']]

genreinWhichMostMoviesAreAddedInLast10Years  = temp[(temp['type'] == 'Movie') & (temp['date\_added\_year'].isin(temp['date\_added\_year'].value\_counts().sort\_index(ascending=False).reset\_index()['index'].head(10)))]

genreinWhichMostMoviesAreAddedInLast10Years = genreinWhichMostMoviesAreAddedInLast10Years.groupby('simplified\_genre')['show\_id'].count().sort\_values(ascending=False).reset\_index()

Top5MovieGenresinLast10Years = genreinWhichMostMoviesAreAddedInLast10Years.head(5)['simplified\_genre']

Top5MovieGenresinLast10Years



* **TOP 5 Genre in Which TV Shows Are Added In the Last 10 Years**

temp = netflix[['show\_id','date\_added\_year']].merge(netflixWithspGenre,on = 'show\_id')

temp = temp.loc[:,['show\_id','date\_added\_year','type','simplified\_genre']]

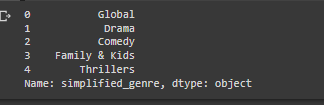
genreinWhichMostTVShowAreAddedInLast10Years  = temp[(temp['type'] == 'TV Show') & (temp['date\_added\_year'].isin(temp['date\_added\_year'].value\_counts().sort\_index(ascending=False).reset\_index()['index'].head(10)))]

genreinWhichMostTVShowAreAddedInLast10Years.groupby('simplified\_genre')['show\_id'].count().sort\_values(ascending=False).reset\_index()

# Top 3 TV Show Generes in Last 10 years as per TV Show added count on netflix in last 10 years

Top5TVShowGenresinLast10Years = genreinWhichMostTVShowAreAddedInLast10Years.groupby('simplified\_genre')['show\_id'].count().sort\_values(ascending=False).reset\_index().head(5)['simplified\_genre']

Top5TVShowGenresinLast10Years



mergedData = netflixWithspGenreAndDirector[['show\_id','simplified\_genre','director','type']].merge(netflixWithspGenreAndCast[['show\_id','cast']]).merge(netflixWithspGenreAndCountry[['show\_id','country']])

mergedData.head()



directorCastCountryAsPerTop5GenreinMovies =  mergedData[(mergedData['simplified\_genre'].isin(Top5MovieGenresinLast10Years)) & (mergedData['type'] == 'Movie')].groupby(['director','cast','country'])['show\_id'].count()

result = directorCastCountryAsPerTop5GenreinMovies.sort\_values(ascending = False).reset\_index().head(15)

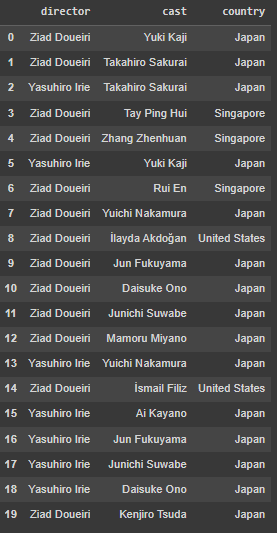
result[['director','cast','country']]

* **DIRECTORS , CAST worked in this genre and in which COUNTRY were the movies produced**  
  
* **DIRECTORS , CAST worked in this genre and in which COUNTRY was the TV Show produced.**

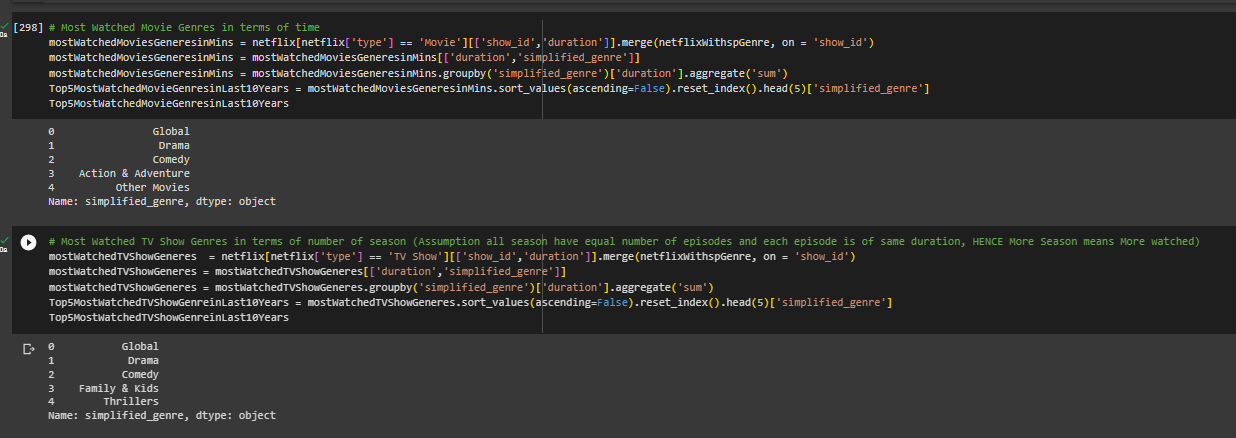
directorCastCountryAsPerTop5GenreinTVShow =  mergedData[(mergedData['simplified\_genre'].isin(Top5TVShowGenresinLast10Years)) & (mergedData['type'] == 'TV Show')].groupby(['director','cast','country'])['show\_id'].count()

result = directorCastCountryAsPerTop5GenreinTVShow.sort\_values(ascending = False).reset\_index().head(20)

result[['director','cast','country']]



* **Most Watched Movie Genres in terms of time**
* **Most Watched TV Show Genres**(Assumption all seasons have an equal number of episodes and each episode is of the same duration, HENCE More Season means More watched)



* **Directors, CAST worked in these most watched genres and in which country were the MOVIES produced**

directorCastCountryAsPerTop5MostWatchedMovieGenres =  mergedData[(mergedData['simplified\_genre'].isin(Top5MostWatchedMovieGenresinLast10Years)) & (mergedData['type'] == 'Movie')].groupby(['director','cast','country'])['show\_id'].count()

result = directorCastCountryAsPerTop5GenreinMovies.sort\_values(ascending = False).reset\_index().head(15)

result[['director','cast','country']]



* **Directors, CAST worked in these most watched genres and in which country were the TV Show produced**

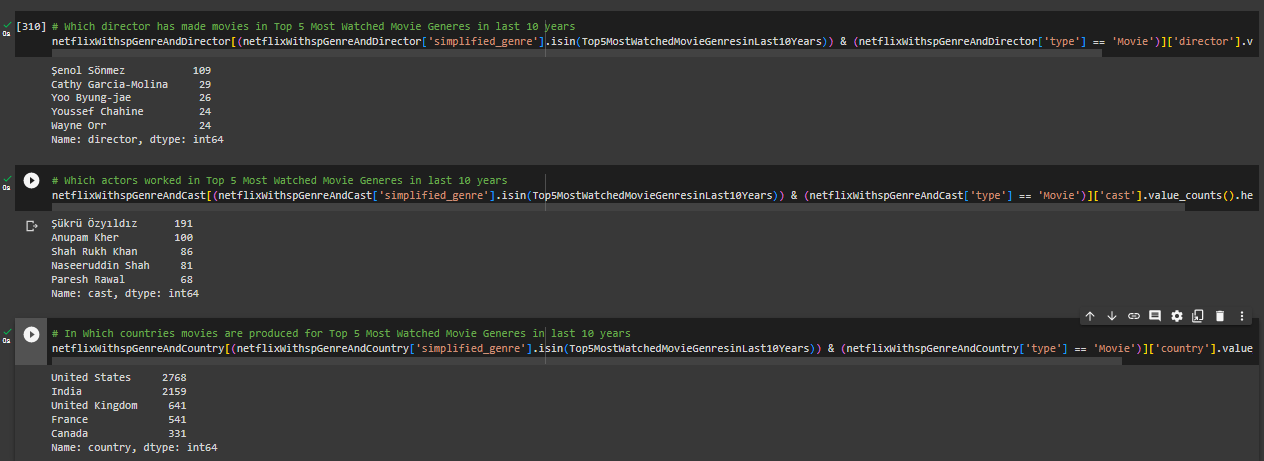
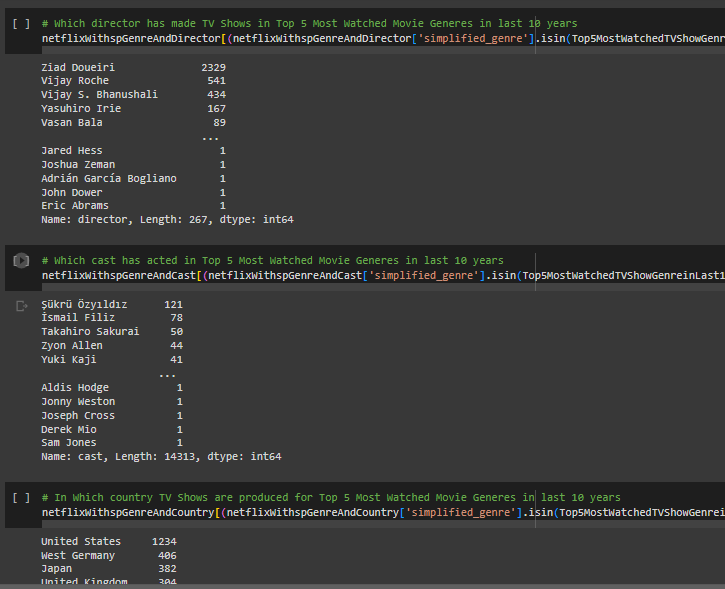
directorCastCountryAsPerTop5MostWatchedTVShowGenres =  mergedData[(mergedData['simplified\_genre'].isin(Top5MostWatchedTVShowGenreinLast10Years)) & (mergedData['type'] == 'TV Show')].groupby(['director','cast','country'])['show\_id'].count()

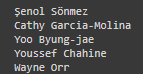
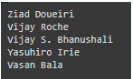
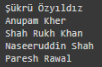
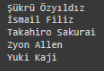
result = directorCastCountryAsPerTop5MostWatchedTVShowGenres.sort\_values(ascending = False).reset\_index().head(15)

result[['director','cast','country']]



* **The DIRECTORS, CAST, COUNTRY present Most in the Top  5 Movie/TV Show Genre (\* One show can have more than one director, cast, and country)  
    
  The code is available in the attached PDF.**

  
  
  
8. **Recommendations**- Actionable items for business. No technical jargon. No complications. Simple action items that everyone can understand

* The top 5 Most Watched movie genres are**Global**( 'International Movies'), **Drama**('Dramas'), **Comedy** ('Comedies')**,  Action & Adventure (**'Action & Adventure' **), Other Movies (**'Movies', 'Independent Movies')  
    
  **REC 1: Movies should be made in these genres.**
* The top 5 Most Watched TV show genres are **Global**('International TV Shows'), **Drama**('TV Dramas', 'TV Shows'), **Comedy** ('Comedies', 'TV Comedies', 'Stand-Up Comedy', 'Stand-Up Comedy & Talk Shows'), **Family & Kids** ("Kids' TV", 'Anime Series', 'Children & Family Movies' ,' Anime Features'), **Thrillers**('Thrillers' , 'TV Thrillers' , 'TV Mysteries' , 'Crime TV Shows')  
    
  **REC 2: TV Shows should be made in these genres.**
* Top 5 Directors Movie directors for Most Watched movie genre   
    
    
    
  **REC 3: These directors can be preferred for making movies in these genres**
* Top 5 Directors TV Show directors for Most Watched TV Show Genre  
    
    
    
  **REC 4: These directors can be preferred for making TV Shows in these genres**
* Top 5 Cast Member for Most Watched movie genre   
    
    
    
  **REC 5: These cast members can be preferred for acting in movies of the above genres**
* Top 5 Cast Member for Most Watched TV Show Genre   
    
    
    
  **REC 6: These cast members can be preferred for acting in TV Shows of the above genres**
* Some of the directors, and cast who worked in this top 5 most watched Movie genre, and the country they were produced in.  
    
  
* Some of the directors, and cast worked in this top 5 most watched TV Show genre, and the country they were produced in  
    
  