TEMIRLAN KADYR

# INF-323 PROJECT

Id : 190103490

# Agenda

## What goes in the slides

DATA DESCRIPTION

PREPROCESSING STEPS

ANALYSIS FOR CATEGORICAL / NUMERICAL

ANALYSIS FOR SET OF 2 VARIABLE

ANALYSIS FOR DATETIME

OWN ANALYSIS

GROUPBY AND FILTRATION

# Data Description

**Dataset that consists of listings of all the movies and tv shows available on Disney+ streaming platform.**

**Columns:**

- ID
- Type
- Title
- Director
- Cast
- Country
- Date added
- Release year
- Rating
- Duration
- Listed in
- Description

# COLUMNS DESCRIPTION

① **Id**
Unique ID of Movie or Tv Show

② **Type**
Is it a Movie or Tv Show

③ **Title**
The name of Movie / Tv Show

④ **Director**
Director of Movie / Tv Show

⑤ **Cast**
Main Cast of Movie / Tv Show

⑥ **Country**
Country of production

⑦ **Date added**
Date added on Disney+

⑧ **Release year**
Original Release year of Movie / Tv Show

⑨ **Rating**
Rating of the Movie / Tv Show

⑩ **Duration**
Total duration of Movie / Tv Show

⑪ **Listed in**
Listed In - Genere

⑫ **Description**
Short decription of the Movie / Tv Show

**G** – All ages admitted – General audiences.

**PG** – Parental guidance suggested – Some material may not be suitable for pre-teenagers.

**PG-13** – Parents strongly cautioned – Some material may be inappropriate for children under 13.

**TV-14** – Parents Strongly Cautioned. This program contains some material that many parents would find unsuitable for children under 14 years of age.

**TV-G** – General Audience. Most parents would find this program suitable for all ages.

**TV-PG** – (Parental Guidance Suggested) This program contains material that parents may find unsuitable for younger children.

**TV-Y** – (All Children) This program is designed to be appropriate for all children.

**TV-Y7** – (Directed To Older Children) This program is designed for children age 7 and above.

**TV-Y7-FV** – TV-Y7-rated program contains behavior that, while violent and often combative, is fictional and can be shown to children who understand the difference between fantasy and reality.

# PREPROCESSING DATA

## COLUMNS THAT CONSIST MISSING VALUES

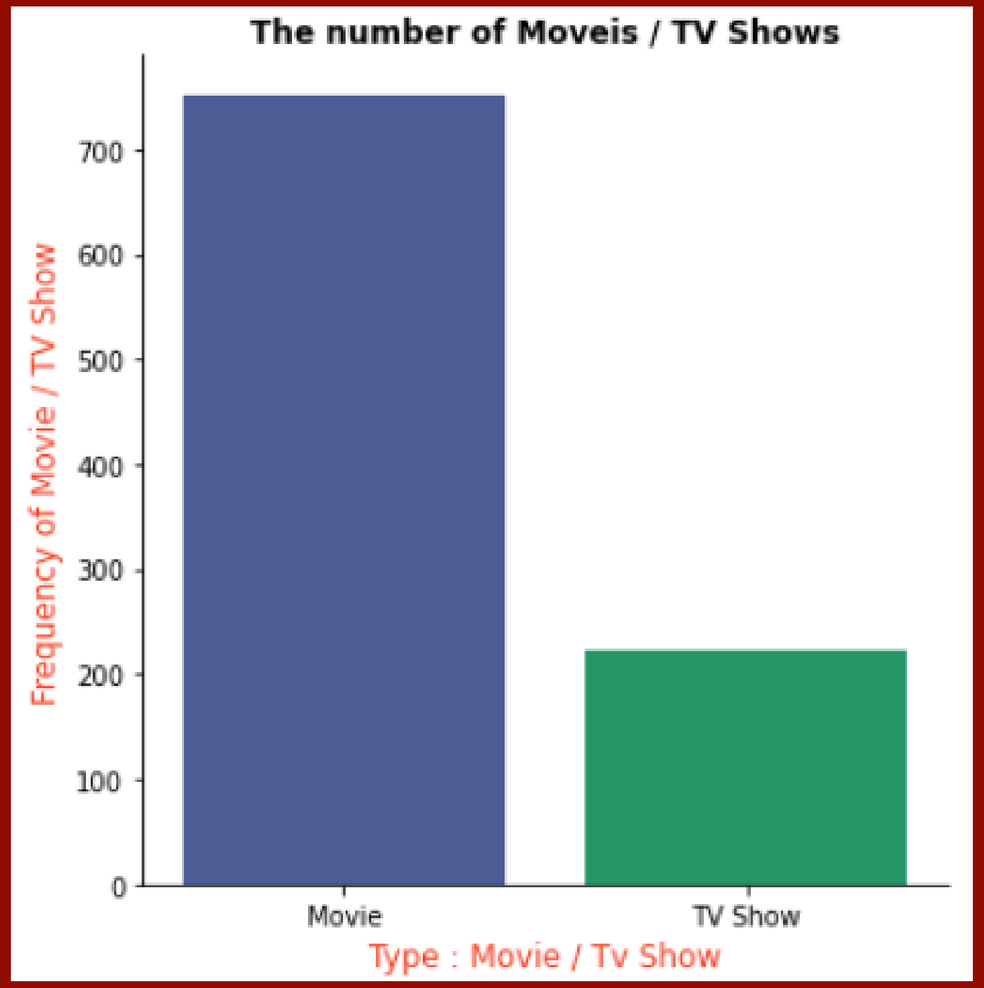
Columns such as : ***director***, ***cast*** and ***date_added*** have some missing values. Missing values for director and cast filled with Unknown value.

## DUPLICATE VALUES

Dataset does not has any duplicate values,

## RESHAPING DATASET

For future analysis touched only Movies / TV Shows which country production is United States.




# Analysis for categorical variable

## Grouping data by its type

Here we can see that the number of Movies is much higher than Tv Show programs in Disney+ platform.

Overall Number of Movie / Tv Show:

- Movies : 752
- TV Shows : 222

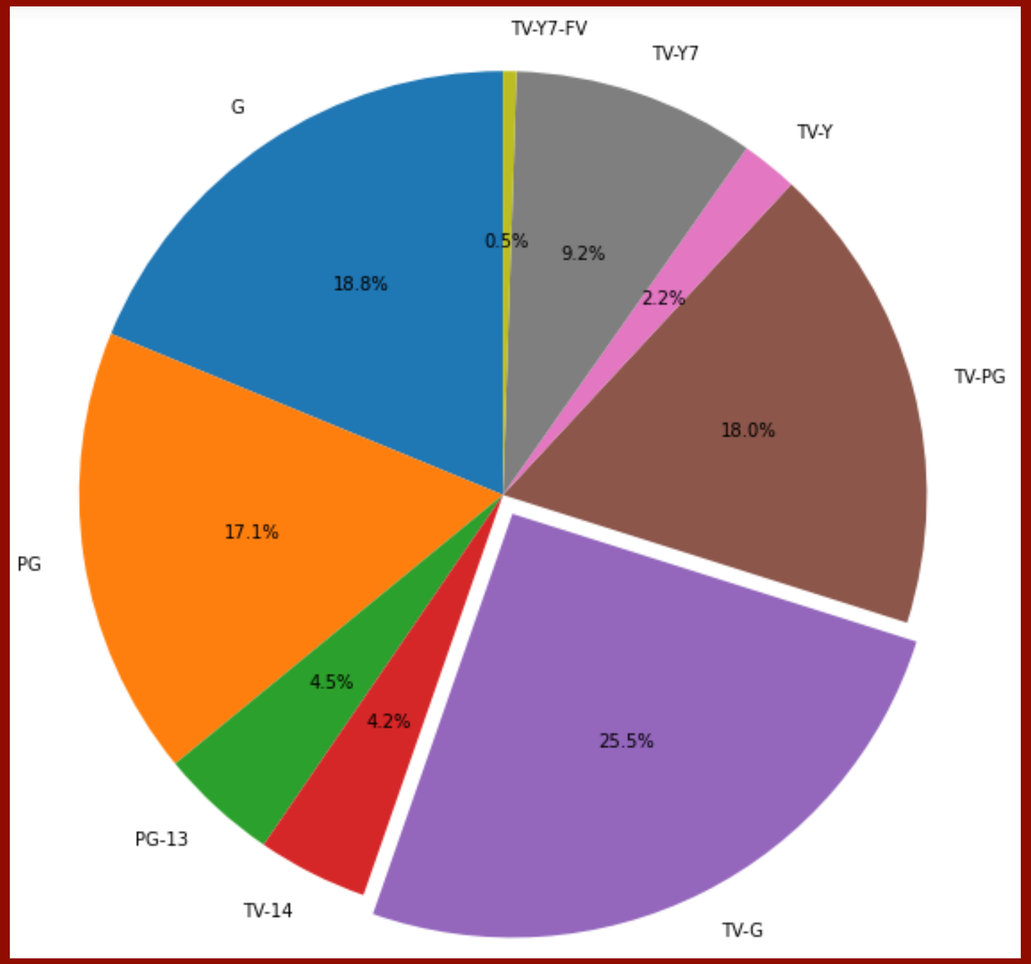
# Analysis for categorical variable

## Grouping data by its rating

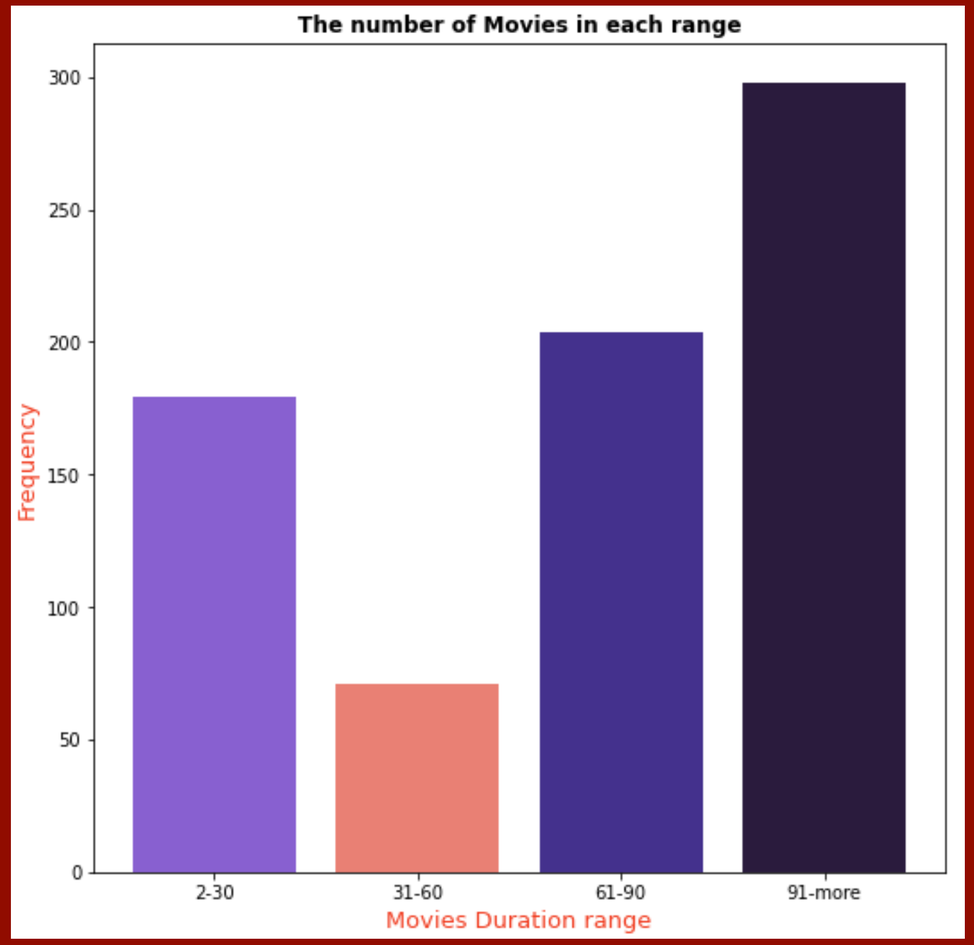
Here we can see that both Movies and Tv Shows with rating 'TV-G'  streams the most time. (25.5% of all rating)

Number of Movie / Tv Show by rating:

- rating G : 183
- rating PG : 167
- rating PG-13 : 44
- rating TV-14 : 41
- rating TV-G : 248
- rating TV-PG : 175
- rating TV-Y : 21
- rating TV-Y7 : 90
- rating TV-Y7-FV : 5



# Analysis for numerical variable




## Grouping Movies by its duration

Firstly, i divide durations to category of duration between 2-30, 31-60, 61-90, 91-more mins.
We can see that Movies with duration between 91-more streams the most time.

Number of Movies by duration range:

- duration between 2-30 : 179
- duration between 31-60 : 71
- duration between 61-90 : 204
- duration between 91-more: 298

# Analysis for set of two variable

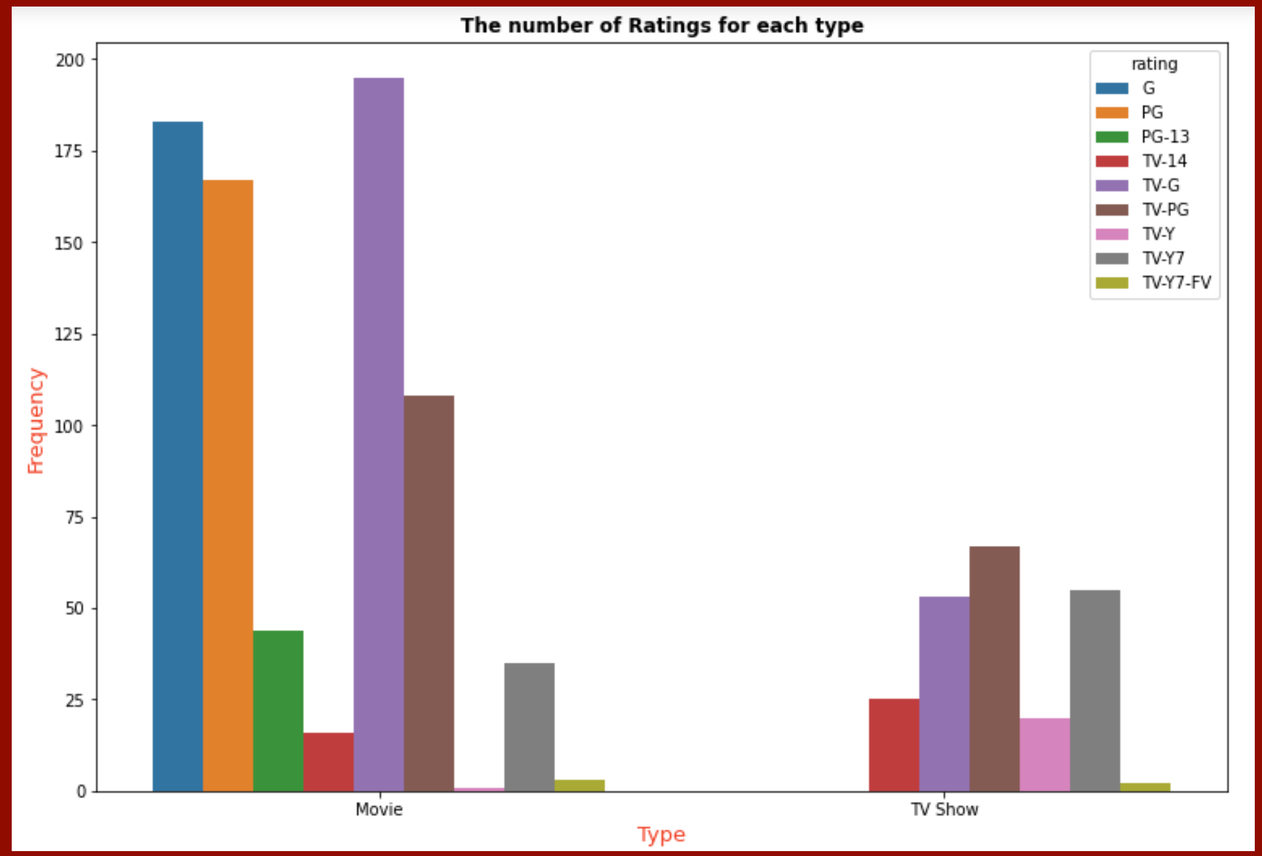
## Grouping data by its type and rating

Number of ***Movie*** grouped by rating:

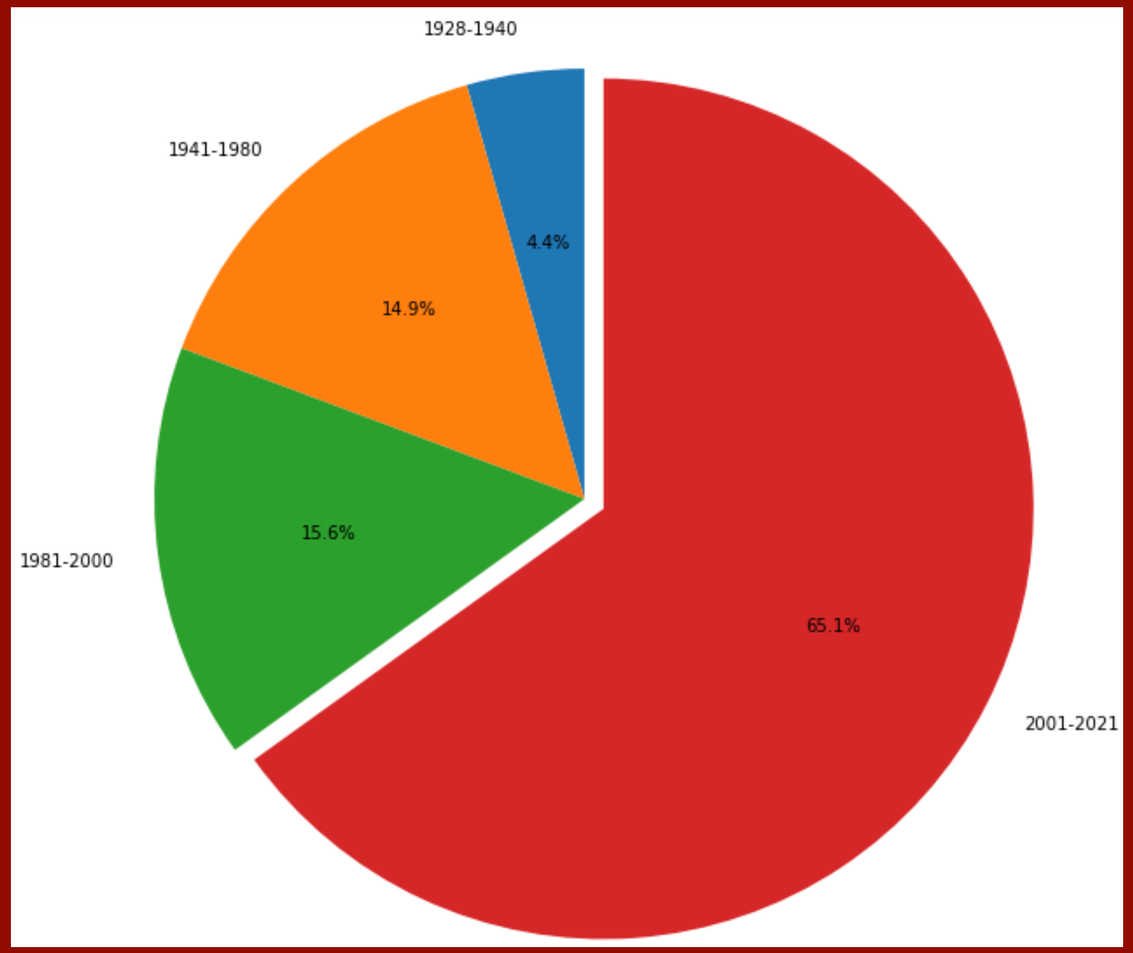
- rating TV-14 : 16
- rating TV-G : 195
- rating TV-PG : 108
- rating TV-Y : 1
- rating TV-Y7 : 35
- rating TV-Y7-FV : 3

Number of ***TV Shows*** grouped by rating:

- rating TV-14 : 25
- rating TV-G : 53
- rating TV-PG : 67
- rating TV-Y : 20
- rating TV-Y7 : 55
- rating TV-Y7-FV : 2



# Analysis for datetime variable




## Grouping Movies / TV Shows by its release year

Divide release years into category : 1928-1940, 1941-1980, 1981-2000 and 2001-2021.
We can see that most Movies release year was between 2001-2021.

Number of Movies / TV Shows by release year range:

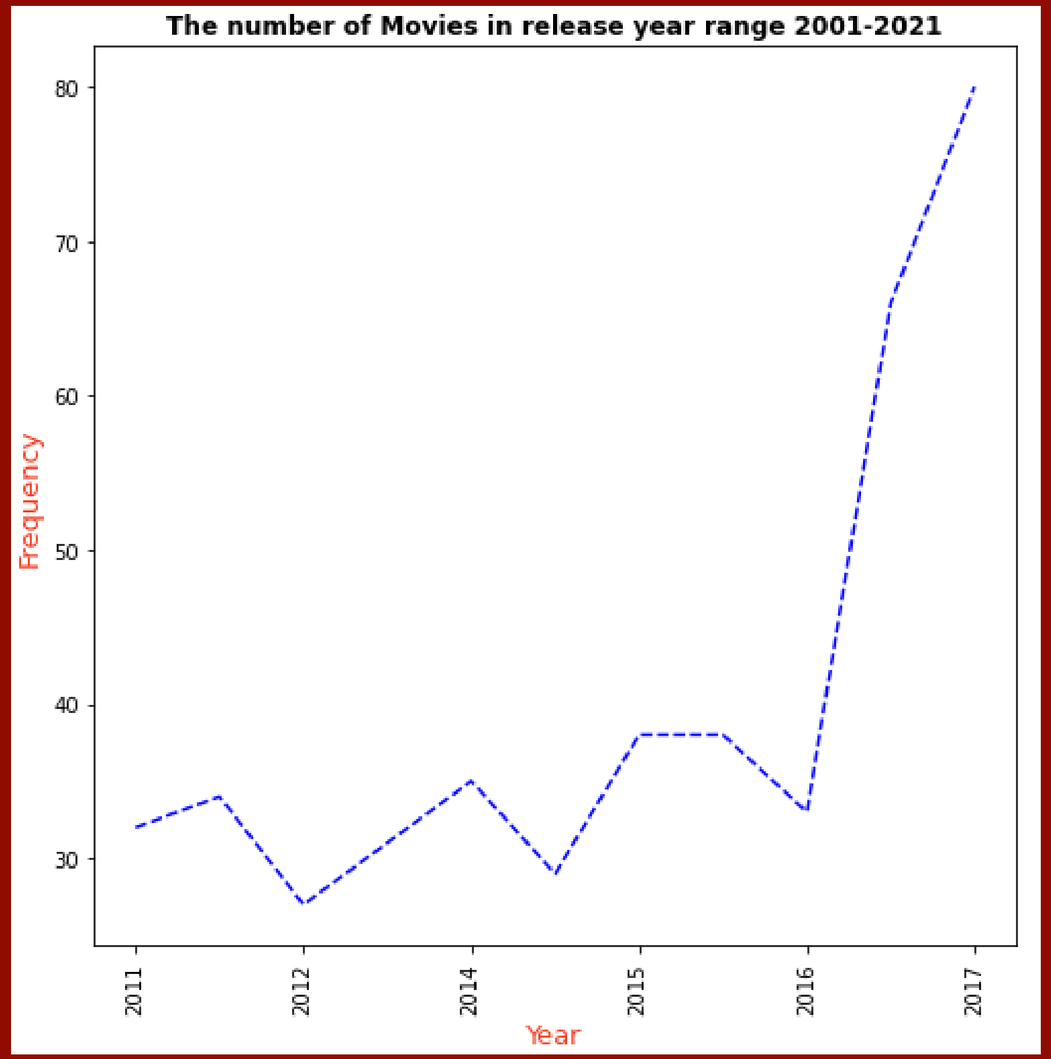
- release year between 1928-1940 : 43
- release year between 1941-1980 : 145
- release year between 1981-2000 : 152
- release year between 2001-2021 : 634

# Continue Analysis for datetime

## Grouping Movies / TV Shows by most release year

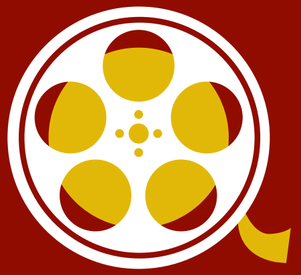
In previous analysis Movies / Tv Shows in range 2001-2021 was most popular. In this range we can see that Movies / Tv Show most append in 2020.

- release year 2010 : 32
- release year 2011 : 34
- release year 2012 : 27
- release year 2014 : 35
- release year 2015 : 29
- release year 2016 : 38
- release year 2017 : 38
- release year 2018 : 33
- release year 2019 : 66
- release year 2020 : 80

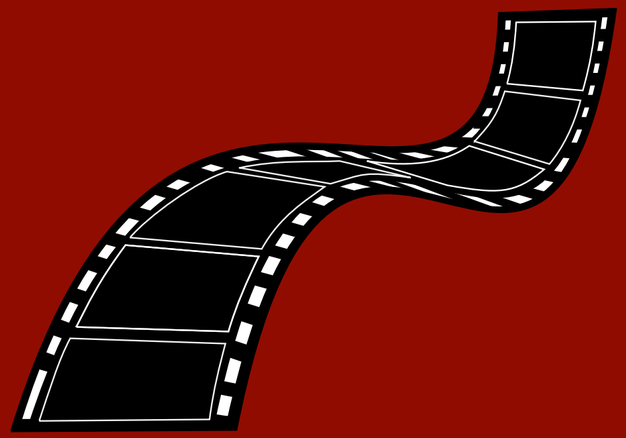


**What are Movies that have average duration over than 100 min?**

# To deal with this question we will use groupby() + filter()



Grouping data by rating and apply filtration which will take only the movies with average duration over than 100 minutes.



As a result of analysis we will see that Movies that have average duration more that 100 are in 'PG-13' rating.

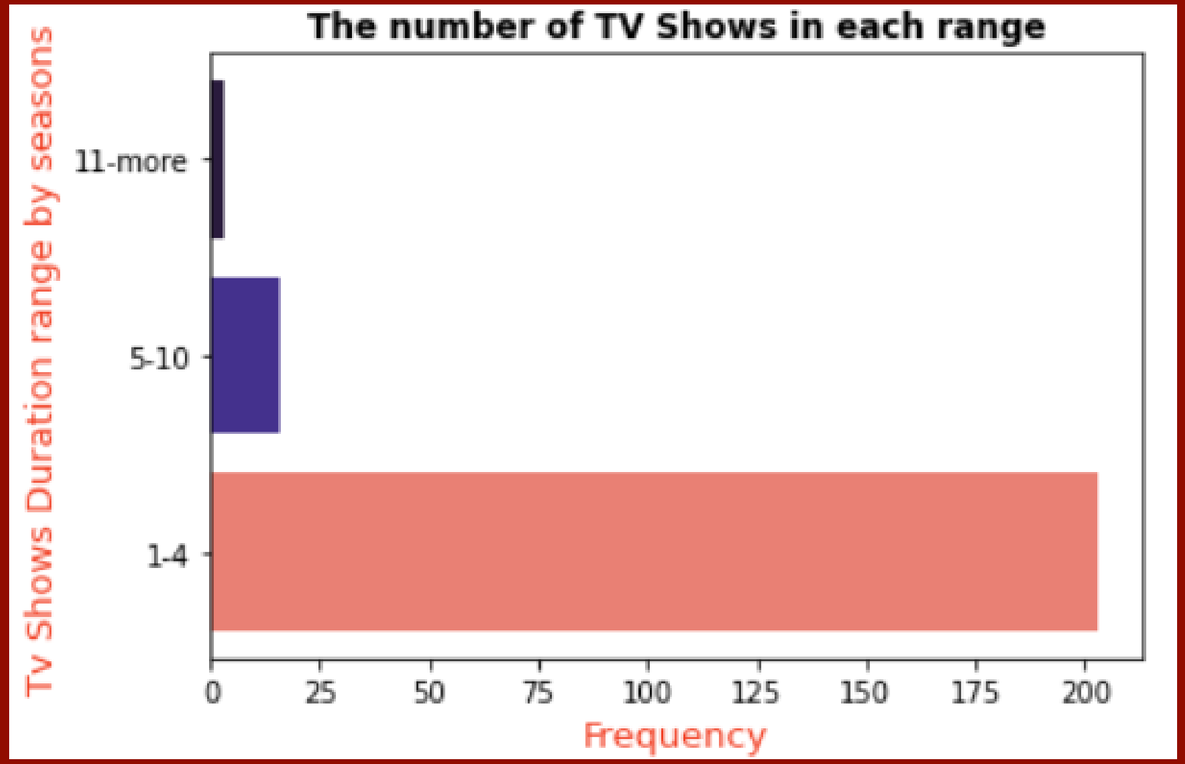
# Own Analysis

## Grouping TV Shows by its seasons duration

Firstly, i divide durations to category of duration between 1-4, 5-10, 11-more seasons.
We can see that TV Shows with 1-4 seasons duration streams the most time.

Number of TV Show by seasons duration :

- 1-4 seasons : 203
- 5-10 seasons : 16
- 11-more seasons : 3



# THANKS FOR ATTENTION!