

# Исследование датасета **Netflix** по загруженным данным на конец **2021** года

**DA1022**

**Телеш Елизавета**

**NETFLIX**

# История изменений проекта

Версия	Дата	Примечания
0.1	12.05.2022	Выбор датасета, план
0.2	14.05.2022	Очистка и анализ данных в Python Pandas
0.3	17.05.2022	Детализированный анализ данных в Python Pandas
0.4	24.05.2022	Визуализация в Python Pandas
0.5	28.05.2022	Построение ДБ в Power BI
0.6	31.05.2021	Добавление слайдов в презентацию

# Этапы реализации проекта

## 1. Загрузка и описание данных в Python Pandas

2. Очистка и анализ данных в Python Pandas

2.1. Очистка данных

2.2 Анализ данных

2.3. Детализированный анализ датасета: статистические группировки

3. Визуализация в Python Pandas

3.1. Аналитика ТОР-10

3.2. Визуализация в разрезе стран по жанрам

3.3. Анализ динамики загруженной видеопродукции в разрезе по фильмам и ТВ-шоу

3.4 Рейтинг по возрастному ограничению

3.5. Map chart

4. Построение ДБ в Power BI



# 1. Загрузка и описание данных

Данные состоят из контента, добавленного в Netflix с 2008 по 2021 год. Самый старый контент датирован 1925 годом, а самый новый – 2021 годом. Датасет включает 8791 строку.

## Описание полей:

- show\_id: идентификационный номер видео
- title: название
- country: страна-производитель фильма
- release\_year: год выпуска
- listed\_in: жанр видео
- type: тип Тип: Movie/TV Show
- director: имя режиссера
- date\_added: дата и время добавления
- rating: возрастное ограничение

```
In [183]: Netflix=pd.read_csv('archive.zip', parse_dates=['date_added'])
Netflix.head(5)
```

Out[183]:

	show_id	type	title	director	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2021-09-25	2020	PG-13	90 min	Documentaries
1	s3	TV Show	Ganglands	Julien Leclercq	France	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	2021-09-24	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	2021-09-22	2021	TV-PG	91 min	Children & Family Movies, Comedies
4	s8	Movie	Sankofa	Haile Gerima	United States	2021-09-24	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies



# Этапы реализации проекта

1. Загрузка и описание данных в Python Pandas
2. Очистка и анализ данных в Python Pandas
  - 2.1. Очистка данных
  - 2.2 Анализ данных
  - 2.3. Детализированный анализ датасета: статистические группировки
3. Визуализация в Python Pandas
  - 3.1. Аналитика ТОР-10
  - 3.2. Визуализация в разрезе стран по жанрам
  - 3.3. Анализ динамики загруженной видеопродукции в разрезе по фильмам и ТВ-шоу
  - 3.4 Рейтинг по возрастному ограничению
  - 3.5. Map chart
4. Построение ДБ в Power BI



## 2. Очистка и анализ данных

### 2.1. Очистка данных

#### 1. Поиск нулей

```
In [211]: #search of null values  
Netflix.isna().sum()
```

```
Out[211]: show_id      0  
type          0  
title         0  
director      0  
country        0  
date_added    0  
release_year   0  
rating         0  
duration       0  
listed_in      0  
dtype: int64
```

#### 2. Переименование, удаление нулевых значений и дубликатов

```
In [212]: #rename of columns  
Netflix = Netflix.rename(columns={'listed_in': 'genres'})
```

```
In [213]: #remove of null values  
Netflix=Netflix.dropna()
```

```
In [214]: #remove of duplicates  
Netflix = Netflix.drop_duplicates()
```

### 3. Выделение ключевого жанра

Этот набор данных содержит видеопродукты по жанрам, которые могут содержать несколько жанров одновременно. Однако основным жанром является название, которое идет первым. Для анализа видеопродуктов по жанрам мы будем выделять только те жанры, которые являются ключевыми:

```
In [215]: #selecting a key video genre for analysis  
Netflix[['genres', 'genres_2']] = Netflix['genres'].str.split(',', 1, expand=True)  
Netflix.head(5)
```

	show_id	type	title	director	country	date_added	release_year	rating	duration	genres	genres_2
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2021-09-25	2020	PG-13	90 min	Documentaries	None
1	s3	TV Show	Ganglands	Julien Leclercq	France	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, TV Action & Adventure	



### 2.1. Очистка данных

#### 4. Выделение годов из даты

Для анализа набора данных по годам необходимо разделить дату и выбрать годы.

```
In [216]: #adding a column with years
Netflix['year_added']=Netflix['date_added'].dt.to_period('Y')
Netflix.head(5)
```

Out[216]:

	show_id	type	title	director	country	date_added	release_year	rating	duration	genres	genres_2	year_added
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2021-09-25	2020	PG-13	90 min	Documentaries	None	2021
1	s3	TV Show	Ganglands	Julien Leclercq	France	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, TV Action & Adventure	International TV Shows, TV Action & Adventure	2021

#### 5. Удаление ненужных колонок

```
In [217]: Netflix = Netflix.drop('genres_2', axis=1)
Netflix = Netflix.drop('date_added', axis=1)
```

```
In [219]: Netflix.head(5)
```

Out[219]:

	show_id	type	title	director	country	release_year	rating	duration	genres	year_added
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2020	PG-13	90 min	Documentaries	2021
1	s3	TV Show	Ganglands	Julien Leclercq	France	2021	TV-MA	1 Season	Crime TV Shows	2021
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	2021	TV-MA	1 Season	TV Dramas	2021
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	2021	TV-PG	91 min	Children & Family Movies	2021
4	s8	Movie	Sankofa	Haile Gerima	United States	1993	TV-MA	125 min	Dramas	2021

## 2. Очистка и анализ данных

### 2.2. Анализ данных датасета

Функция first\_check:

Input:

```
In [8]: def first_check(dataset):

    print()
    print('Первые 5 строк таблицы')
    display(dataset.head(5))

    print()
    print('Последние 5 строк таблицы')
    display(dataset.tail(5))

    print()
    print('Информация о таблице')
    print(dataset.info())

    print()
    print('Наименование колонок')
    print(dataset.columns)

    print()
    print('Типы колонок')
    print(Netflix.shape)

    print()
    print('Типы колонок')
    print(Netflix.dtypes)
```

Output:

```
In [9]: first_check(Netflix)
```

Первые 5 строк таблицы						
show_id	type	title	director	country	date_added	rating
0	#1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2021-09-25
1	#3	TV Show	Ganglands	Julien Leclercq	France	2021-09-24
2	#6	TV Show	Midnight Mass	Mike Flanagan	United States	2021-09-24
3	#14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	2021-09-22
4	#8	Movie	Sankofa	Haile Gerima	United States	2021-09-24

Последние 5 строк таблицы						
show_id	type	title	director	country	date_added	rating
8785	#8787	TV Show	Yunus Emre	Not Given	Turkey	2017-01-17
8786	#8788	TV Show	Zak Storm	Not Given	United States	2018-09-13
8787	#8801	TV Show	Zindagi Gulzar Hai	Not Given	Pakistan	2018-12-15
8788	#8784	TV Show	Yoko	Not Given	Pakistan	2018-06-23
8789	#8788	TV Show	YOM	Not Given	Pakistan	2018-06-07

Информация о таблице  
<class 'pandas.core.frame.DataFrame'>  
Int64Index: 8790 entries, 0 to 8789  
Data columns (total 11 columns):  
 # Column Non-Null Count Dtype  
 ---  
 0 show\_id 8790 non-null object  
 1 type 8790 non-null object  
 2 title 8790 non-null object  
 3 director 8790 non-null object  
 4 country 8790 non-null object  
 5 date\_added 8790 non-null datetime64[ns]  
 6 release\_year 8790 non-null int64  
 7 rating 8790 non-null object  
 8 duration 8790 non-null object  
 9 genres 8790 non-null object  
 10 genres\_2 6778 non-null object  
 dtypes: datetime64[ns](1), int64(1), object(9)  
 memory usage: 824.1+ KB  
None

Наименование колонок  
Index(['show\_id', 'type', 'title', 'director', 'country',  
 'release\_year', 'rating', 'duration', 'genres',  
 'genres\_2'],  
 dtype='object')

Типы колонок  
(8790, 11)

Типы колонок

show_id	object
type	object
title	object
director	object
country	object
date_added	datetime64[ns]
release_year	int64
rating	object
duration	object
genres	object
genres_2	object

dtype: object

## 2.3. Детализированный анализ датасета: статистические группировки

### 1. Количество фильмов и ТВ-шоу по жанрам

```
In [220]: movie_genres=Netflix.groupby(['type','genres'], as_index=False).agg({'duration':'count'}).sort_values('duration', ascending=False)
movie_genres.head(2)

Out[220]:
   type      genres  count
0  Movie    Dramas    1599
1  Movie   Comedies    1210

In [222]: movie_genres['perc']=movie_genres['count']/Netflix.release_year.count()*100
movie_genres.head(2)

Out[222]:
   type      genres  count      perc
0  Movie    Dramas    1599  18.101126
1  Movie   Comedies    1210  13.765643
```

Наибольшее количество фильмов по жанру **драма** – 1599 (18,2%), **комедии** – 1210 (13,7%) и **приключения** – 859 (9,8%). Наибольшее количество сериалов и тв-шоу по жанру международные ТВ-шоу – 773 (8,8%), криминальные – 399 (4,5%) и для детей – 385 (4,4%).

### 2. Количество фильмов и ТВ-шоу по годам

```
In [19]: year_added=Netflix.year_added.value_counts().loc[lambda x: x>50].to_frame()
year_added['perc']=year_added['year_added']/Netflix.release_year.count()*100
year_added

Out[19]:
   year_added      perc
0      2019  22.935154
1      2020  21.376564
2      2018  18.748578
3      2021  17.042093
```

Наибольшее количество фильмов было добавлено на сайт Netflix в **2019 году – 2016 (22,9% всех фильмов)**, в **2020 году – 1879 (21,4%)**, в **2018 году – 1648 фильмов (18,7%)**, в 2017 – 1498 фильма (17,1%).

### 3. Количество фильмов и сериалов по странам

```
In [20]: df_countries=Netflix['country'].value_counts().loc[lambda x : x > 100].to_frame()
df_countries.rename
df_countries = df_countries.rename(columns={'country': 'count'})
df_countries['perc']=df_countries['count']/Netflix['country'].count()*100
df_countries.head(6)

Out[20]:
   count      perc
0  United States  3240  36.860068
1           India  1057  12.025028
2  United Kingdom   638   7.258248
```

Наибольшее количество фильмов, размещенных на Netflix было выпущено в **США – 36,8%**, в **Индии – 12%**, Великобритании – 7,3%.

### 4. Количество фильмов и ТВ-шоу по возрастному ограничению

```
In [22]: ratings.sort_values('Movie', ascending=False).head(5)

Out[22]:
   type      Movie  TV Show    All  Movie_perc  TV show_perc
rating
0      All       6126     2664    8790  100.000000  100.000000
1     TV-MA      2062     1143    3205   33.659811  42.905405
2     TV-14      1427      730    2157   23.294156  27.402402
```

Наибольшее количество фильмов с рейтингом только **для взрослых – 2062 (33,65%)**, сериалов и ТВ-шоу – с **рейтингом только для взрослых – 1143 (42,9%)**.

# Этапы реализации проекта

1. Загрузка и описание данных в Python Pandas
2. Очистка и анализ данных в Python Pandas
  - 2.1. Очистка данных
  - 2.2 Анализ данных
  - 2.3. Детализированный анализ датасета: статистические группировки

## 3. Визуализация в Python Pandas

- 3.1. Аналитика ТОР-10
- 3.2. Визуализация в разрезе стран по жанрам
- 3.3. Анализ динамики загруженной видеопродукции в разрезе по фильмам и ТВ-шоу
- 3.4 Рейтинг по возрастному ограничению
- 3.5. Map chart

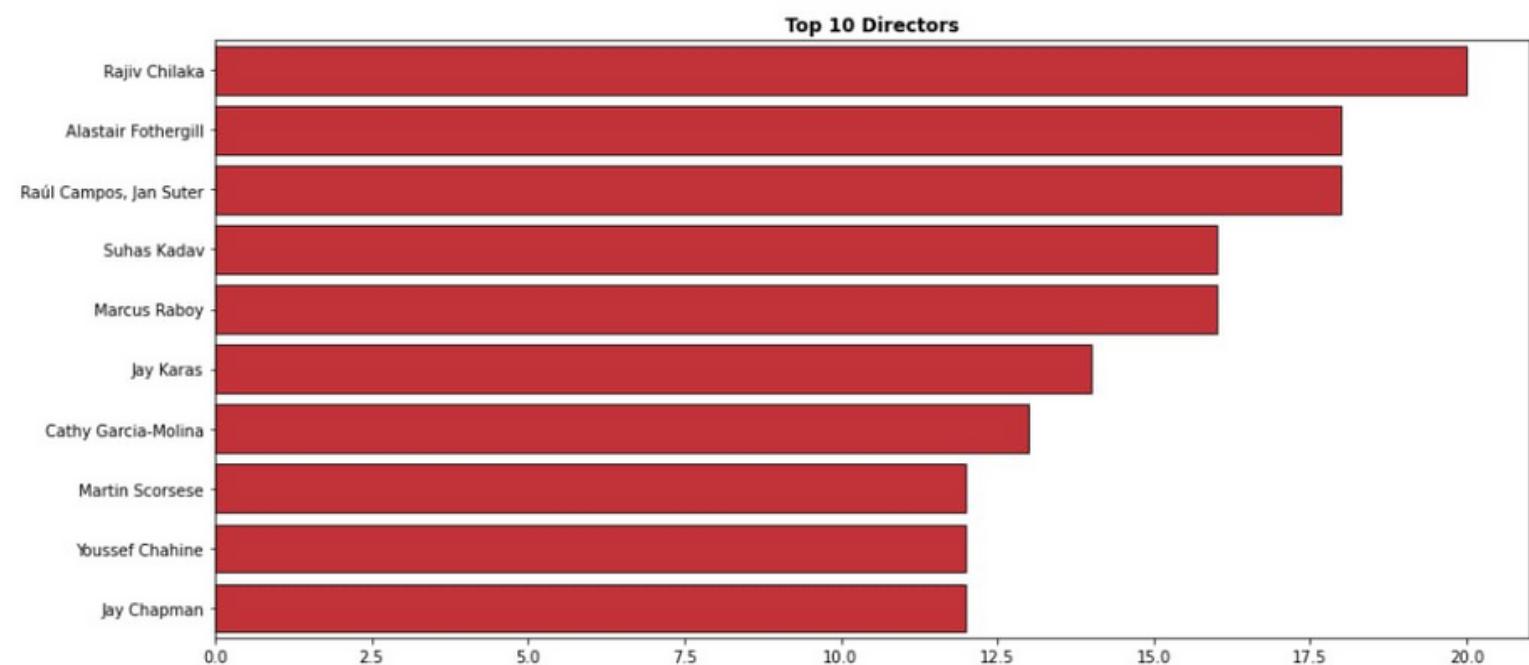
4. Построение ДБ в Power BI

NETFLIX

## 2. Визуализация

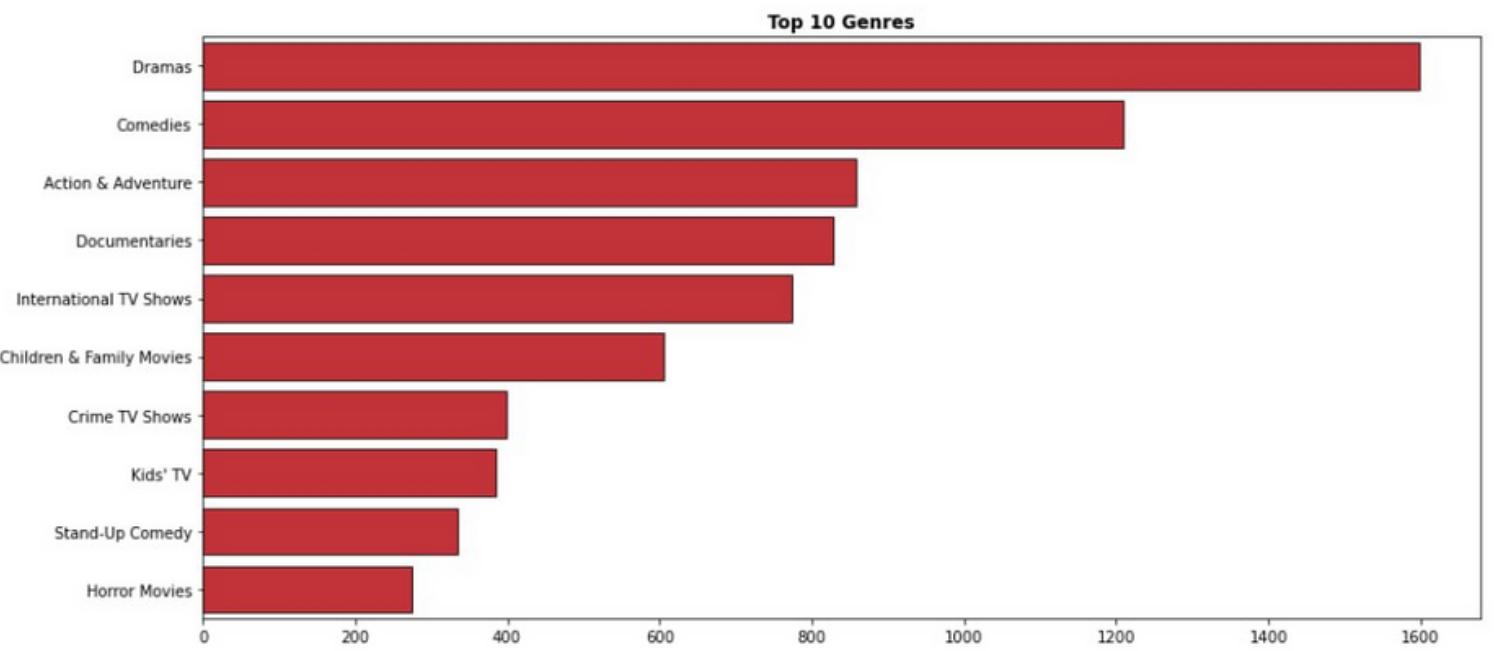
### 3.1. Аналитика ТОР-10

```
In [96]: plt.figure(figsize = (15,7))
top_10 = Netflix['director'].value_counts().drop('Not Given').head(10)
sns.barplot(x = top_10.values, y = top_10.index, edgecolor = 'k', linewidth = 1, saturation = 11, color = '#b30007', alpha = 0.8)
plt.title("Top 10 Directors", fontsize = 12, fontweight = 'heavy')
plt.show()
```

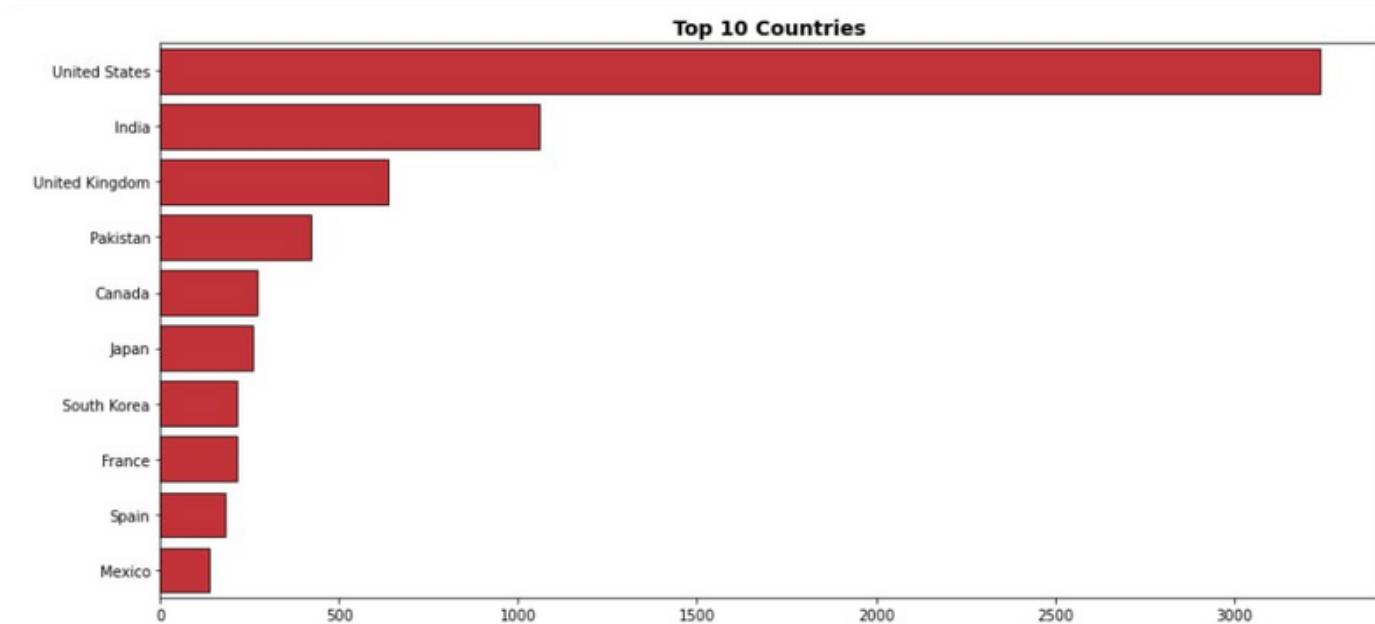


На первом месте по количеству выпущенной видеопродукции по режиссерам – **Раджив Чилака, Аластер Фортергилл, Рауль Кампос и Джан Сутер.**

**Раджив Чилака** – это индийский кинорежиссер и продюсер анимационных фильмов и телесериалов. **Аластер Фортергилл** – это известный британский продюсер и режиссер документальных фильмов о природе. Он является соучредителем компании Silverback Films, которая производит такие документальные фильмы, как "Планета Земля II", "Наша планета" и "Дикие острова". **Рауль Кампос (Raul Campos)** и **Джан Сутер (Jan Suter)** – это радиоведущие на радиостанции KCRW в Лос-Анджелесе, США.



Наибольшее количество видеопродукции, загруженной на Нетфликс, – это **драмы, комедии и приключения.**

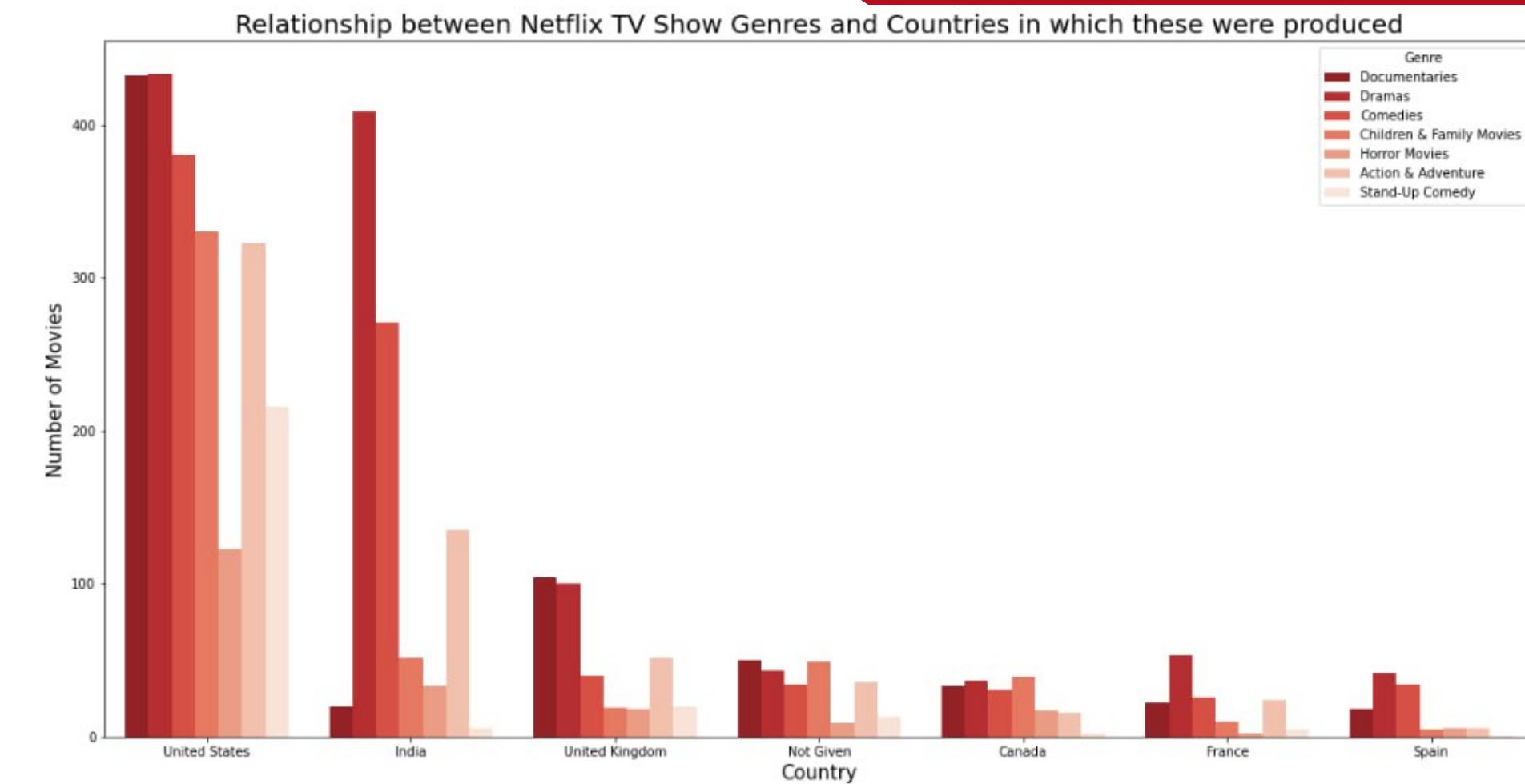
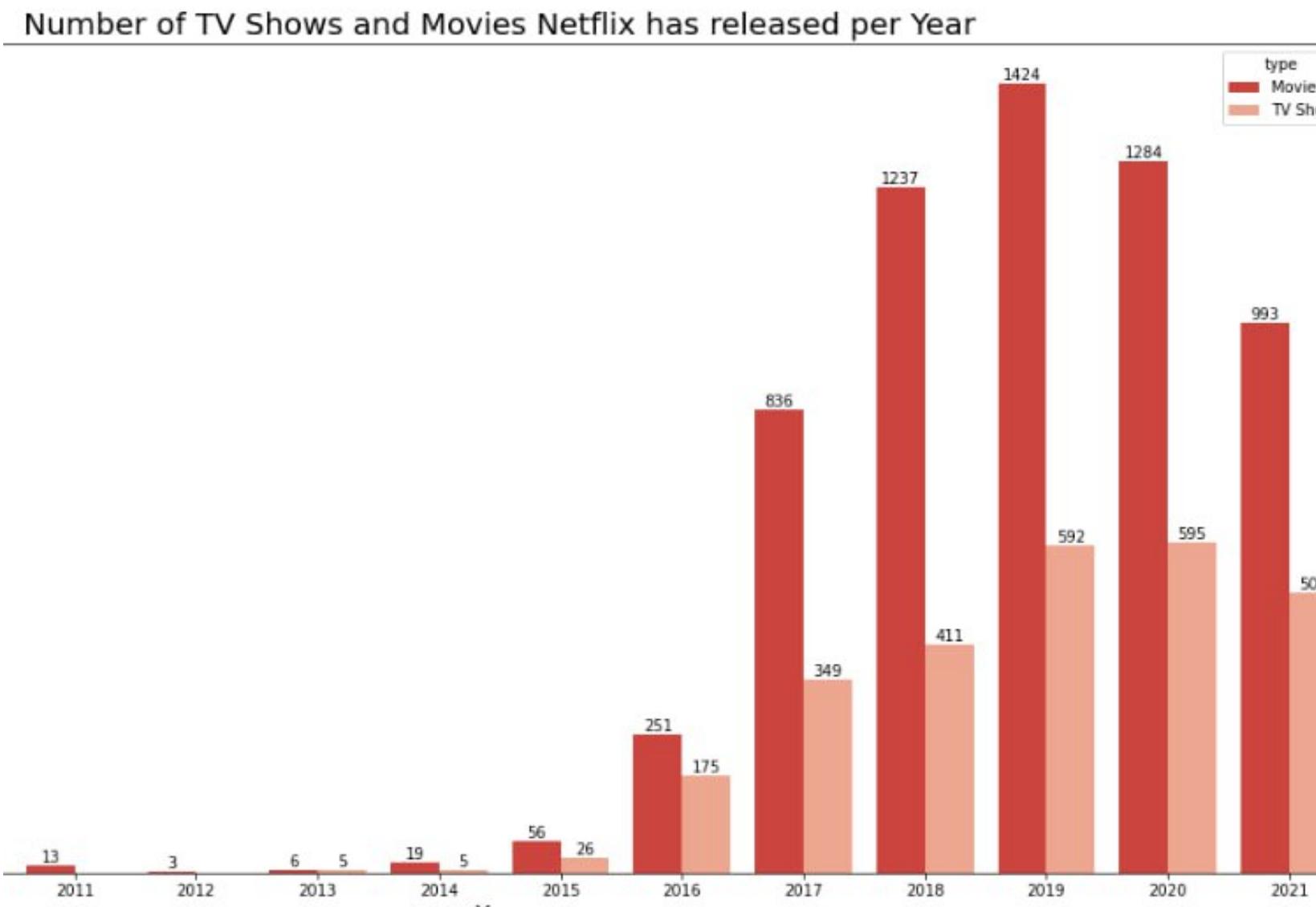


Наибольшее количество фильмов, размещенных на Netflix было выпущено в **США – 3240 фильмов (36,8%), в Индии – 1057 (12%), Великобритании – 638 (7,3%).**

## 3.2. Визуализация в разрезе стран по жанрам

В **США** лидирующие позиции по выпущенными фильмам занимают драмы и документальные фильмы, комедии, в **Индии** – драмы, в **Британии** – документальные фильмы. По ТВ-шоу в **США** - детские ТВ-шоу и документальные сериалы, в **Пакистане** – международные ТВ-шоу, в **Великобритании** – реалити шоу.

## 3.3. Анализ динамики загруженной видеопродукции в разрезе по фильмам и ТВ-шоу



```
In [124]: plt.figure(figsize=[20,10])
base_color = sns.color_palette('coolwarm',n_colors=5)
tv_movie = sns.countplot(x=Netflix.date_added.dt.year, data=Netflix, hue='type', palette = "Reds_r")
tv_movie.set_title("Number of TV Shows and Movies Netflix has released per Year",fontsize = 20)
tv_movie.set_xlabel('Year',fontsize = 15)
tv_movie.set_ylabel('Number of Movies/TV Shows',fontsize = 15)
for container in tv_movie.containers:
    tv_movie.bar_label(container)
```

Количество фильмов на Netflix в 2018 году было в 3 раза больше, чем количество телешоу. Количество фильмов, загруженных на Нетфликс имело **положительную динамику до 2019 года включительно, которая затем пошла на снижение.**

Это может быть связано с пандемией, либо с тем фактом, что первые годы активной деятельности Нетфликс постоянно догружал фильмы, которых не было ранее.

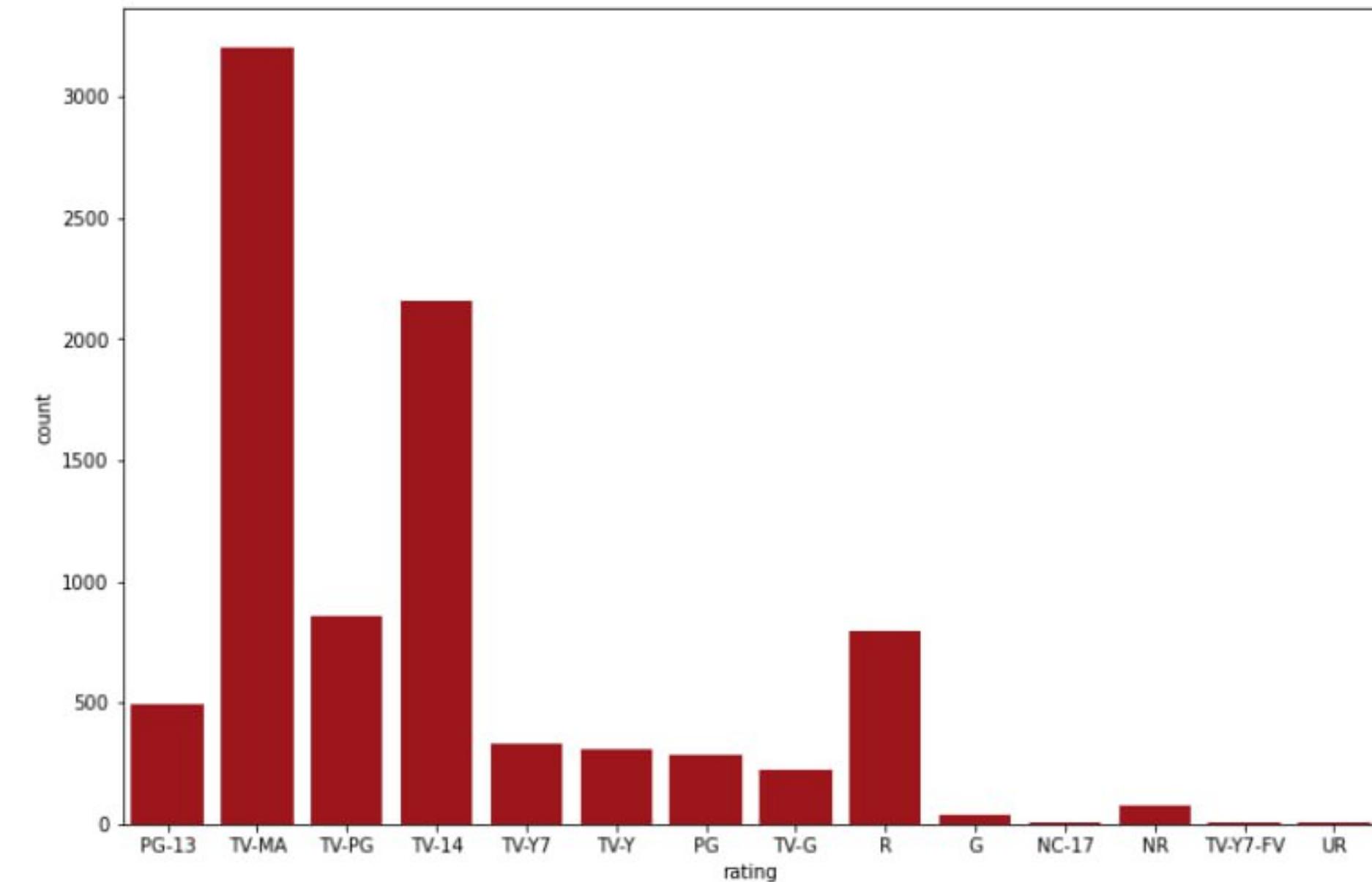
## 3.4 Рейтинг по возрастному ограничению

На графике расположены значения по количеству фильмов по рейтингам, которые используются для классификации контента в кино, телевидении и других медиаформатах в зависимости от того, подходит ли он для детей и какой возрастной категории.

По возрастному ограничению преобладает видеопродукция, которая подходит для просмотра только взрослым, на втором месте – контент для детей 14+ для просмотра с родителями.

```
plt.figure(figsize = (12,8))
sns.countplot(x='rating', data = Netflix, linewidth = 2, color = "#b30007")
```

<AxesSubplot:xlabel='rating', ylabel='count'>

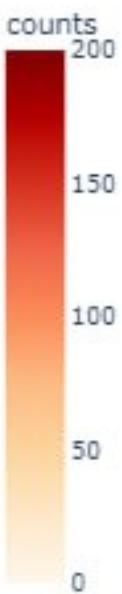
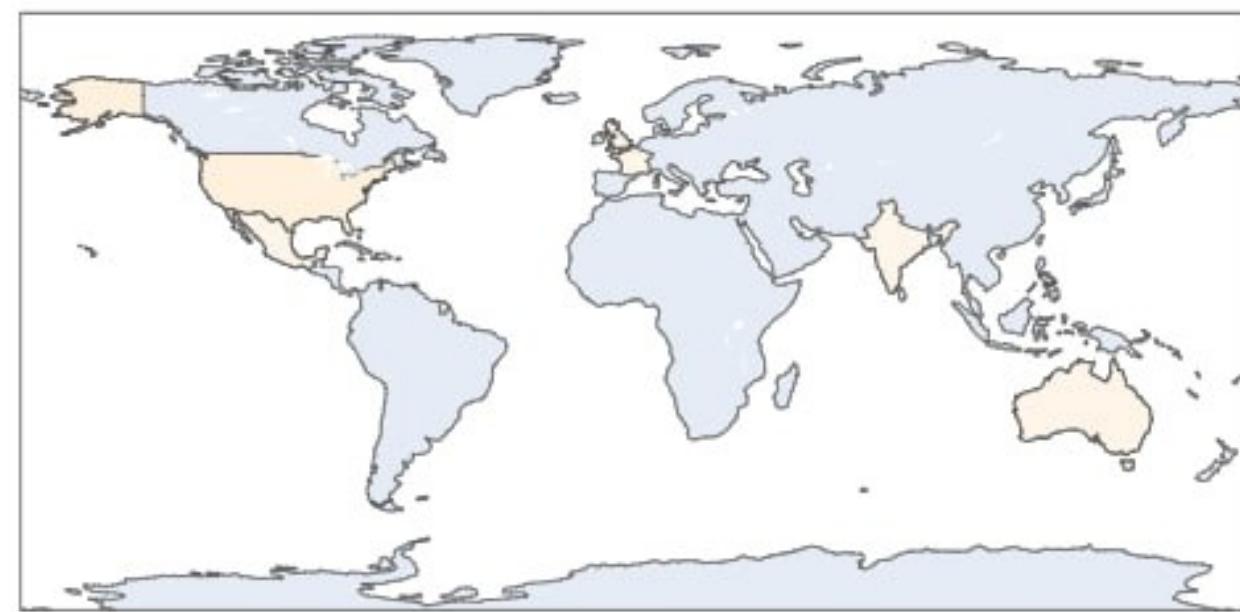


## 3.5. Мар chart

На данном графике необходимо выбрать год выпуска фильма. Чем ярче цвет страны, тем больше количество выпущенных фильмов за данный год.

```
ввод [30]: import plotly.express as px  
  
fig = px.choropleth(year_country2, locations="country", color="counts",  
                     locationmode='country names',  
                     animation_frame='release_year',  
                     range_color=[0,200],  
                     color_continuous_scale=px.colors.sequential.OrRd  
)  
  
fig.update_layout(title='Comparison by country')  
fig.show()
```

Comparison by country



# Этапы реализации проекта

1. Загрузка и описание данных в Python Pandas
2. Очистка и анализ данных в Python Pandas
  - 2.1. Очистка данных
  - 2.2 Анализ данных
  - 2.3. Детализированный анализ датасета: статистические группировки
3. Визуализация в Python Pandas
  - 3.1. Аналитика ТОР-10
  - 3.2. Визуализация в разрезе стран по жанрам
  - 3.3. Анализ динамики загруженной видеопродукции в разрезе по фильмам и ТВ-шоу
  - 3.4 Рейтинг по возрастному ограничению
  - 3.5. Map chart
4. Построение ДБ в Power BI

# NETFLIX

Quantity of movies/TV Shows

8651

Choose the  
filter:

2016

2021



Type

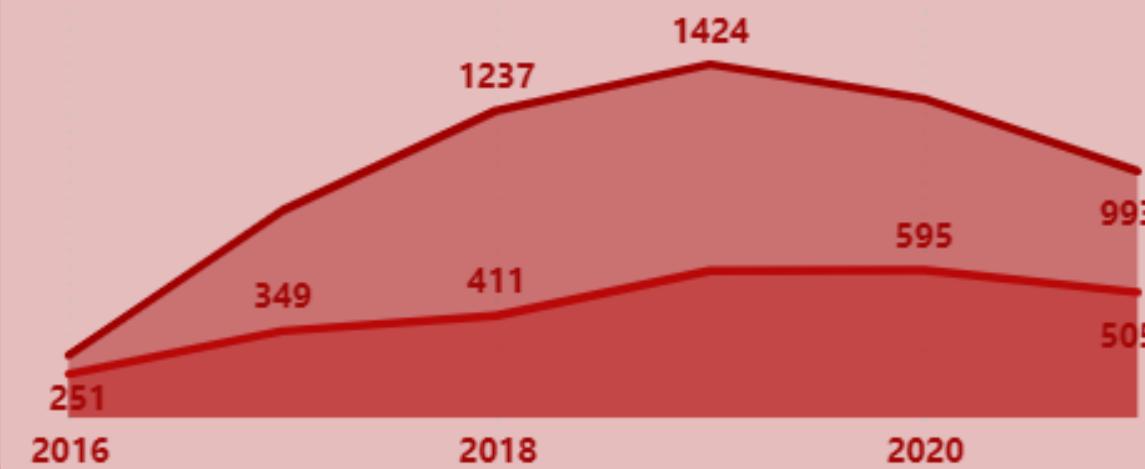
Multiple selections



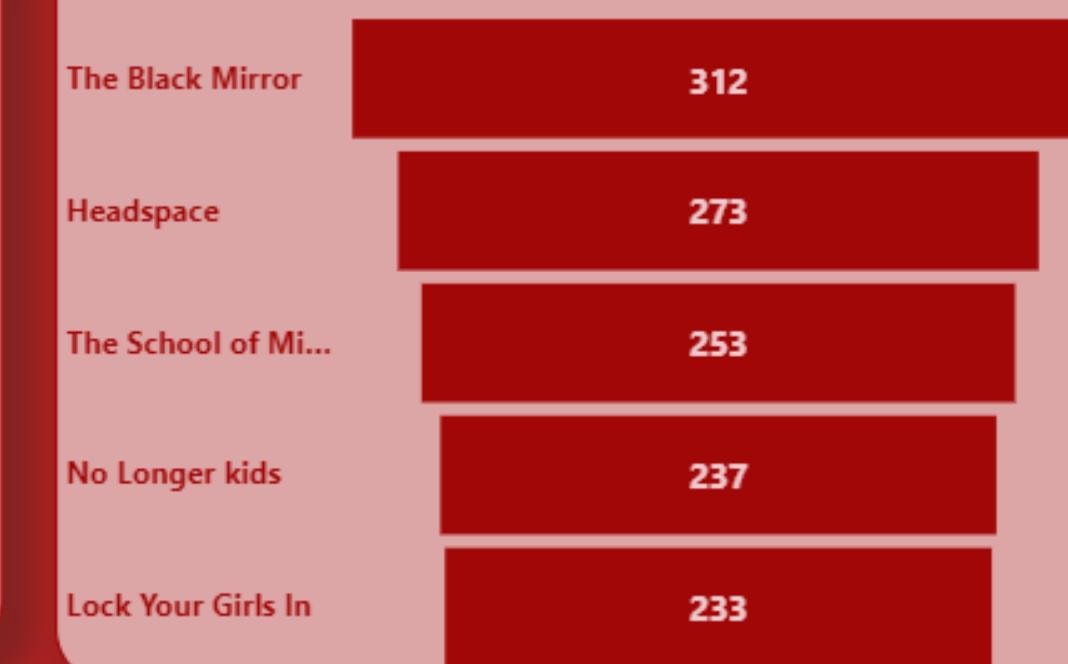
The quantity of films/videos in different countries



● Movie ● TV Show

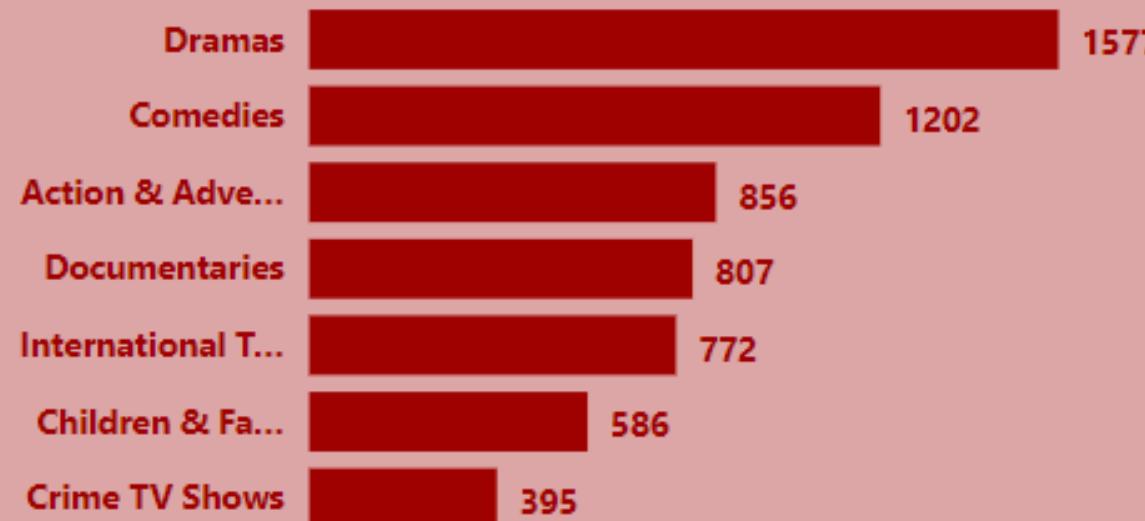


Duration of films, min.

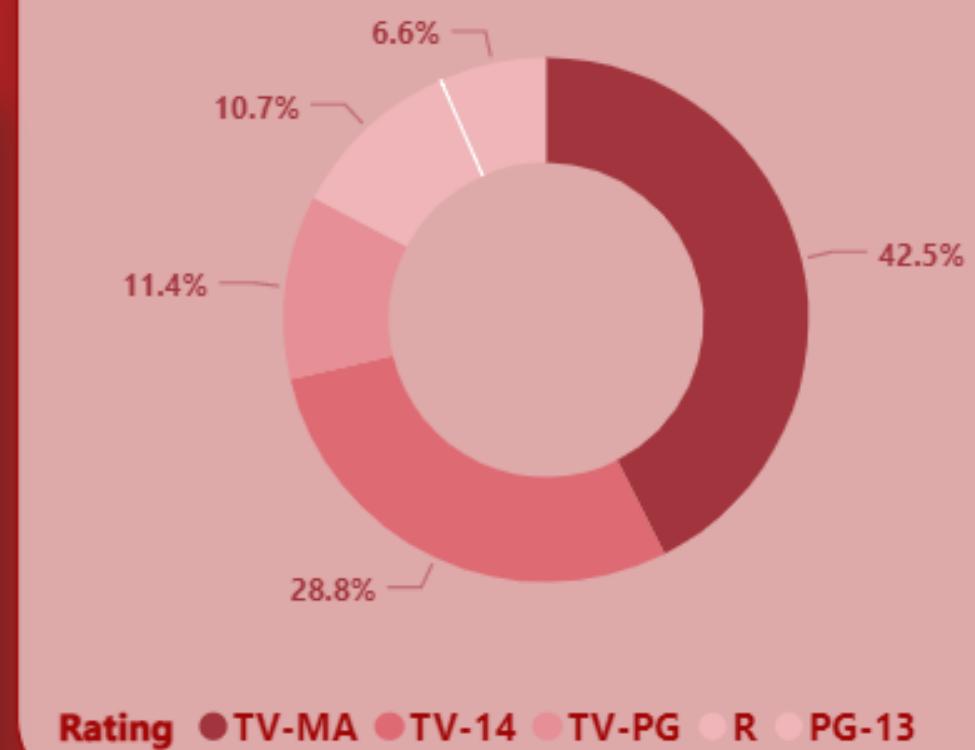
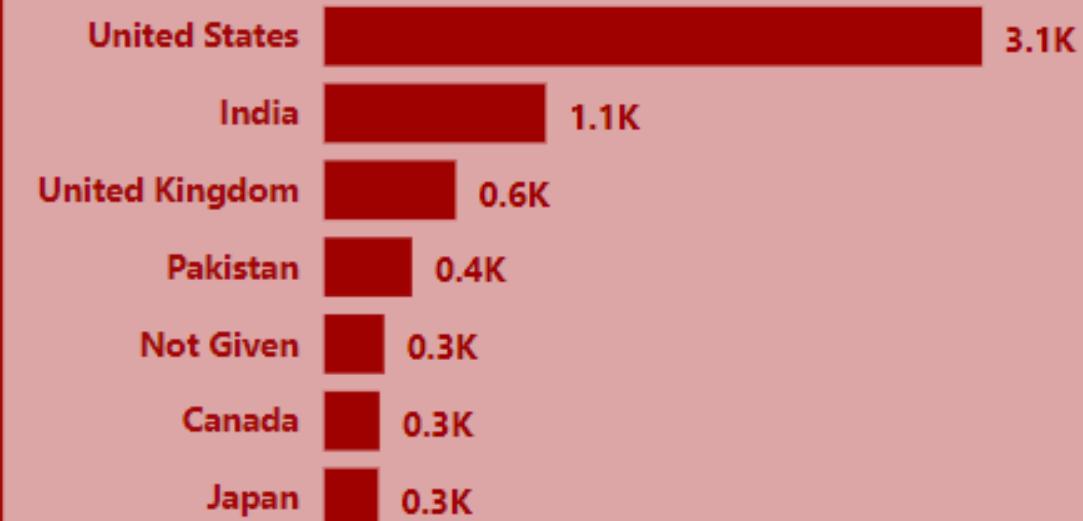


## TOP Films

### TOP-7 by Genres



### TOP-7 by Countries



# Выводы



Наибольшее количество фильмов по жанрам:

- драма – **18,2%**
- комедии – **13,7%**
- приключения – **9,8%**



Наибольшее количество фильмов было добавлено на сайт Netflix в:

- 2019 году – **2016 (22,9% всех фильмов)**
- в 2020 году – **1879 (21,4%).**



- **Наибольшее количество фильмов, размещенных на Netflix:**
- США - **3240 (36,8%)**
- Индия – **1057 (12%)**
- Великобритании – **638 (7,3%).**



По режиссерам с наибольшим количеством фильмов на нетлифкс лидируют: **Раджив Чилака, Аластер Фортергилл, Рауль Кампос и Джан Сутер.**



Наиболее длительными фильмами на Нетфликсе являются **Black mirror, Headspace, the School of Michief.**

NETFLIX

Спасибо за  
внимание!