Final Project :

Data Science

ساميه عبدالله	افنان مفرح عسيري	احلام عامر
442807640	441816447	441816446

Data Name :

- Netflix

Project Target :

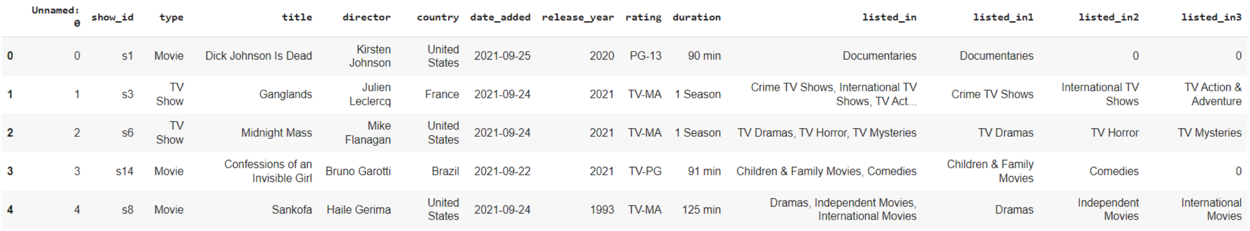
- what years were the largest number of series in it, and what country produces films, and when did the number of series begin to increase (by what date).

Problems :

- The data has missing values and these values affect the graph and the understanding of the data, in addition to that the type of data for the dates needed to reformat.

1- Data Overview :

This data set consists of contents added to Netflix from 2008 to 2021.



	Unnamed: 0	show_id	type	title	director	country	date_added	release_year	rating	duration	listed_in	listed_in1	listed_in2	listed_in3
0	0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2021-09-25	2020	PG-13	90 min	Documentaries	Documentaries	0	0
1	1	s3	TV Show	Ganglands	Julien Leclercq	France	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	Crime TV Shows	International TV Shows	TV Action & Adventure
2	2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	2021-09-24	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries	TV Dramas	TV Horror	TV Mysteries
3	3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	2021-09-22	2021	TV-PG	91 min	Children & Family Movies, Comedies	Children & Family Movies	Comedies	0
4	4	s8	Movie	Sankofa	Haile Gerima	United States	2021-09-24	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies	Dramas	Independent Movies	International Movies

2- Data properties (columns) :

All the data columns type is string , The variables of this data set are:

1. show_id: Netflix ID of the media.
2. Type: Movie or TV Show.
3. title: Title of the media.
4. director: Director of the media.
5. country: Country in which the movie was made.
6. date_added: Date in which the media was added.
7. release_year: Year in which the media was released.
8. rating: Age rating of the media.
9. duration: Duration of the media.
10. listen_in: Classification given by Netflix.

3- Mechanism of Analysis :

- (reading data) .

  The data was read by a pandas library by function (read_csv()),
  And then displayed all the information about our data .

  ```
  pd.read_csv('Netflix_data.csv', header=0)
  ```

- (data preparation) .

  1- We checked for nan values by column_name.isnull().values.any() And then we treat these nan values in two columns .
  2- check for duplicate value by DataFram.duplicated().value_counts() 3- By browsing the data we notice that :

  Variable 'date_added' has the wrong data type.
  Variable 'duration' has the wrong data type.

  we convert the format of (date_added) to data_time and make (duration) an integer value .

  ```
  df['country'].isnull().values.any()
  df['director'].isnull().values.any()
  ```

```
df = df.drop(['Unnamed: 0','listed_in', 'listed_in1',
'listed_in2', 'listed_in3'],axis=1)
df['country'] = df['country'].fillna(df['country'].mode()[0])

df['director'].replace(np.nan, 'No Data',inplace = True)

df.dropna(inplace=True)

df.drop_duplicates(inplace= True)

df_clean.date_added = pd.to_datetime(df_clean.date_added)
```

- (Data exploration) .

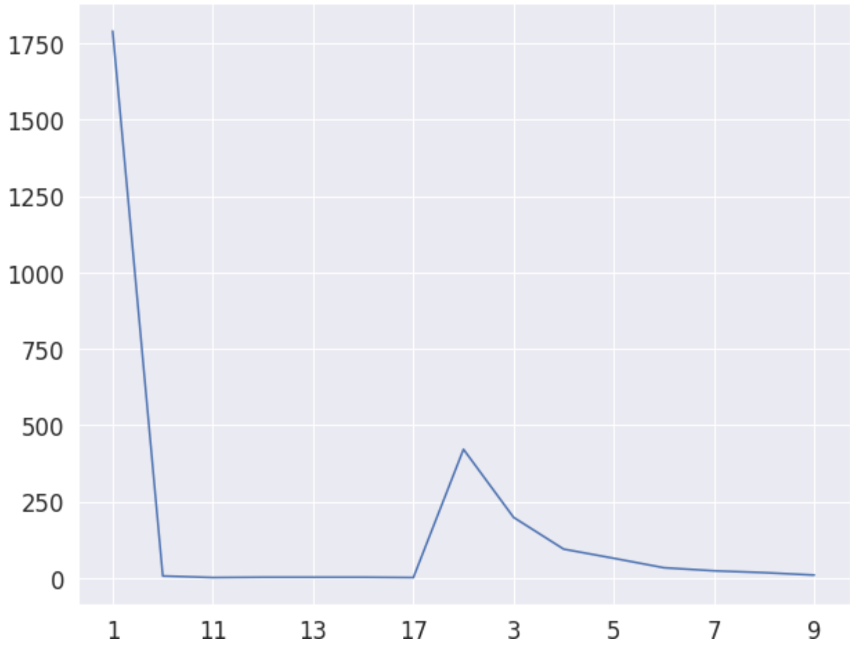
We noticed that our data is divided into two parts (TV Show ,Movie) .
So that we talk one part and deal with it .

```
df_tv = df_clean[df_clean.type == 'TV Show']


df.info()


df.describe()
df.duplicated().value_counts()
```

After that we explore the data by visualizing the data and the relationships between some columns .

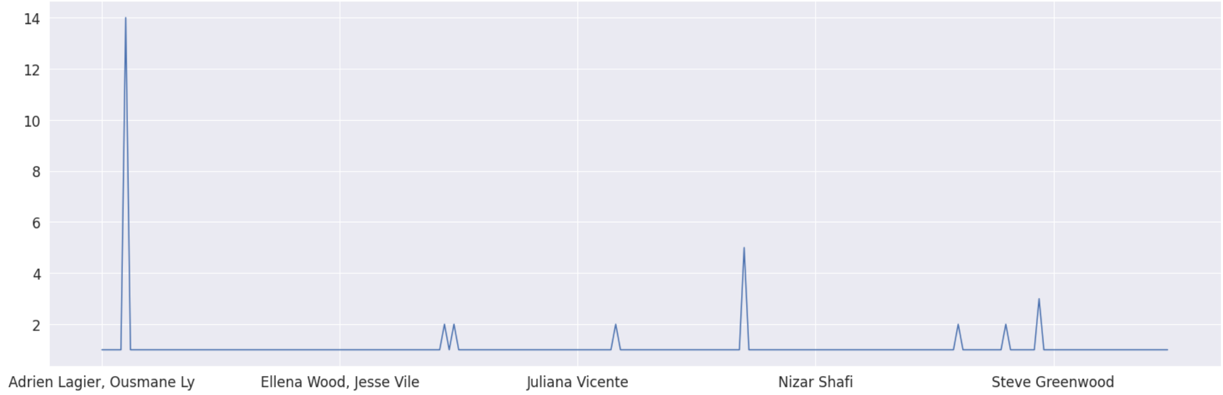
1- Show what is the common ratio of the number of parts in a Netflix series.

```
df_tv['duration_season'].value_counts().sort_index().plot.line()
```



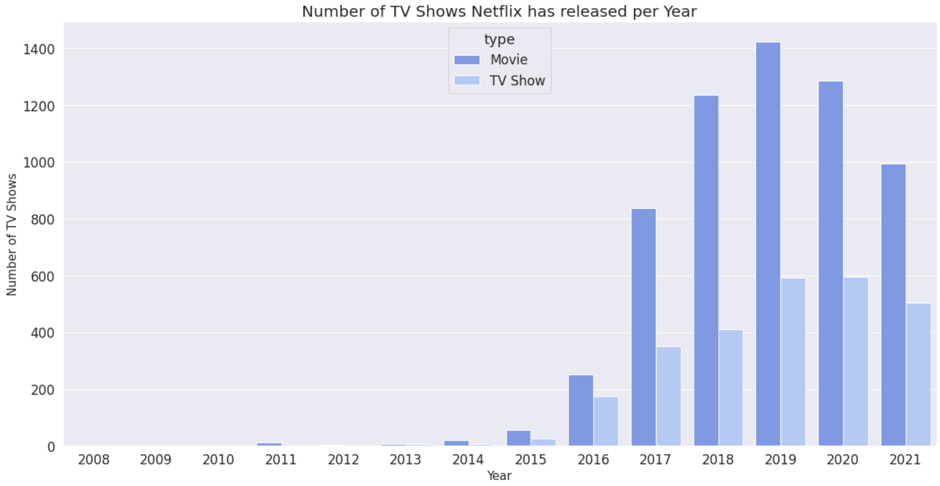
2- Show what is the common directors in a Netflix series

```
directores.sort_index().plot.line()
```



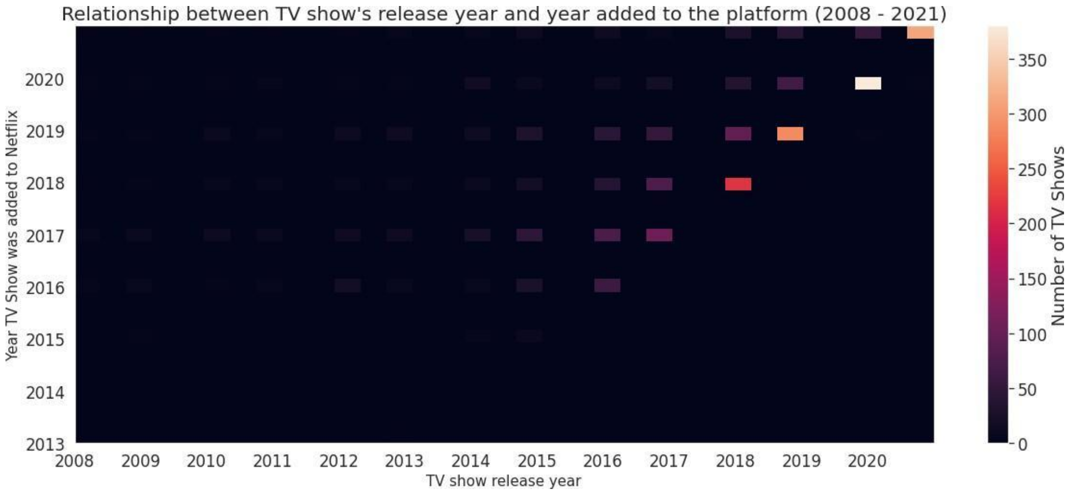
3- Show the Number of TV Shows Netflix has released per Year

```
base_color = sns.color_palette('coolwarm',n_colors=5)
tv_movie = sns.countplot(x=df_clean.date_added.dt.year, data=df_clean, hue='type', palette = base_color)
tv_movie.set_title("Number of TV Shows Netflix has released per Year",fontsize = 20)
tv_movie.set_xlabel('Year',fontsize = 15)
tv_movie.set_ylabel('Number of TV Shows',fontsize = 15)
```



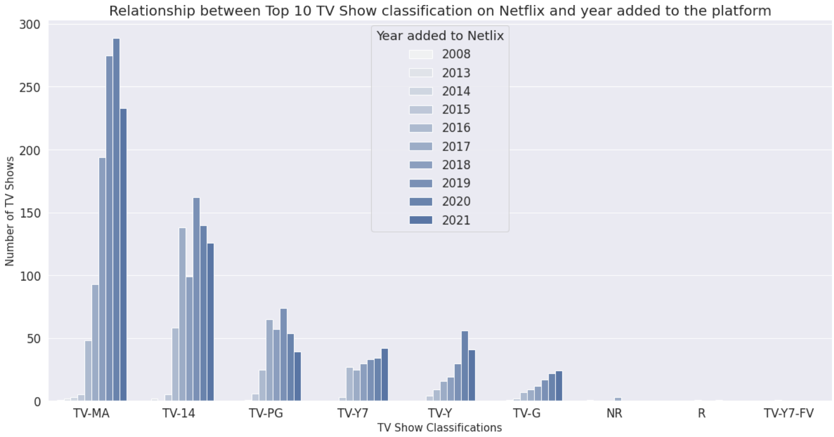
4- show the Relationship between TV show's release year and year added to the platform (2008 - 2021):

```
ry_f = df_tv.release_year>2007
da_f = df_tv.date_added.dt.year>2008
df_tv_f = df_tv[ry_f][da_f]
tv_rd1 = plt.hist2d(data=df_tv_f,x='release_year',y=df_tv_f.date_added.dt.year , bins=33)
plt.xticks(np.arange(2008,2021,1));
plt.yticks(np.arange(2013,2021,1));
plt.xlabel('TV show release year',fontsize = 15)
plt.ylabel('Year TV Show was added to Netflix',fontsize = 15)
plt.title("Relationship between TV show's release year and year added to the platform (2008 - 2021)",fontsize = 20)
plt.colorbar(label = 'Number of TV Shows')
```



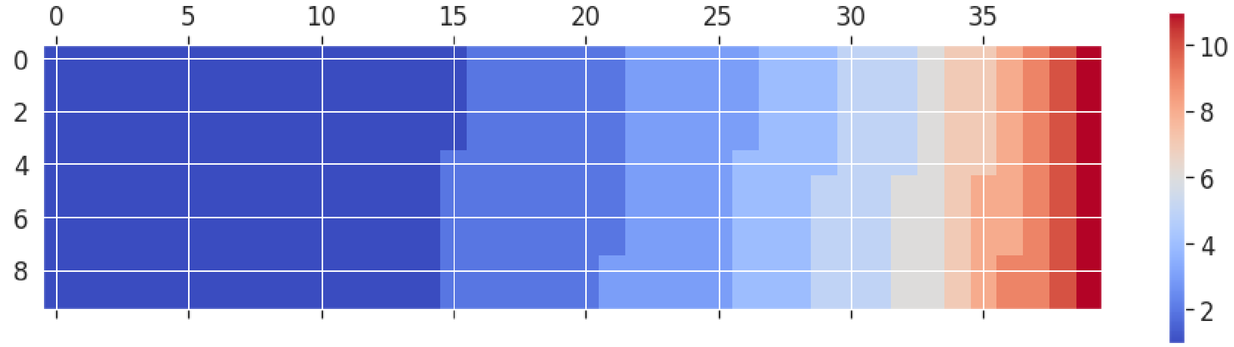
5- show the the most common rating in tv_show by netflix:

```
order = list(df_tv['rating'].value_counts().index)
base_color = base_color = sns.color_palette()[0]
a=df_tv.date_added.dt.year
tv_g = sns.countplot(data=df_tv,x='rating',hue=a, order=order, color=base_color)
tv_g.set_xlabel('Rating',fontsize = 15)
tv_g.set_ylabel('year',fontsize = 15)
tv_g.set_title("the most common rating in tv_show by netflix",fontsize = 20)
plt.legend(title = 'Year added to Netlix',)
```



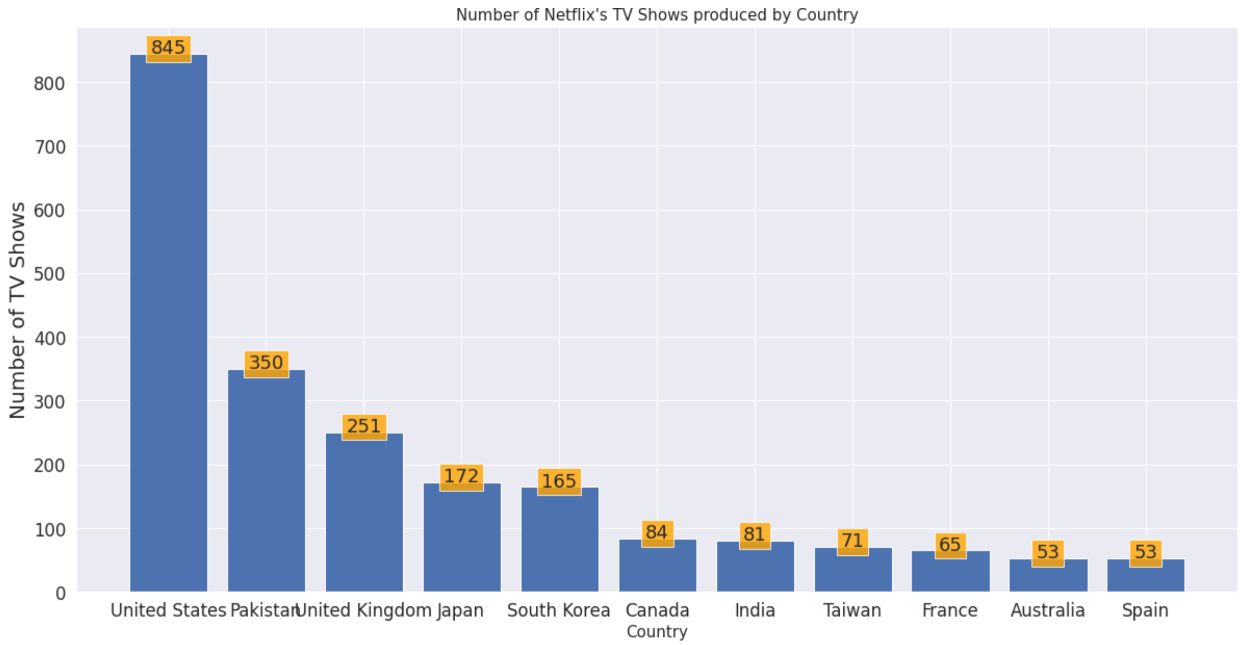
6- we use Waffle Chart to show Which the most countries produce films:

Take states that have been repeated more than 50 times and show it by Waffle Charts



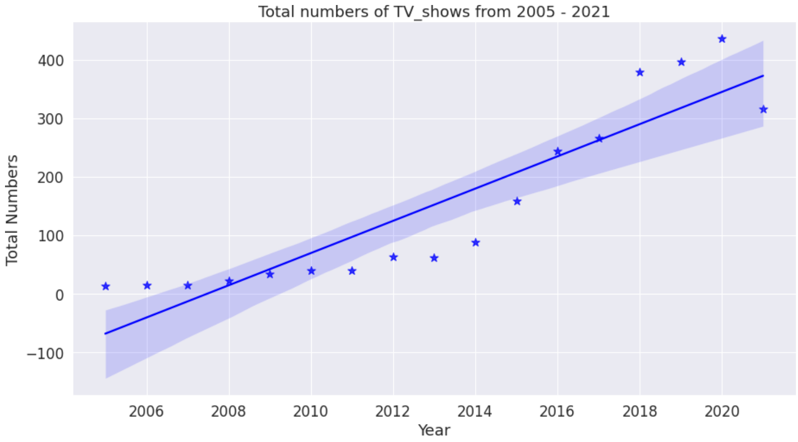
7- show Number of Netflix's TV Shows produced by Country :

```
def addlabels(x,y):
    for i in range(len(x)):
        plt.text(i, y[i], y[i], ha = 'center', Bbox = dict(facecolor = 'orange',
alpha =.8))

x = countries.index.tolist()
y = countries.values
plt.figure(figsize=[20,10])
plt.bar(x, y)
addlabels(x, y)
```

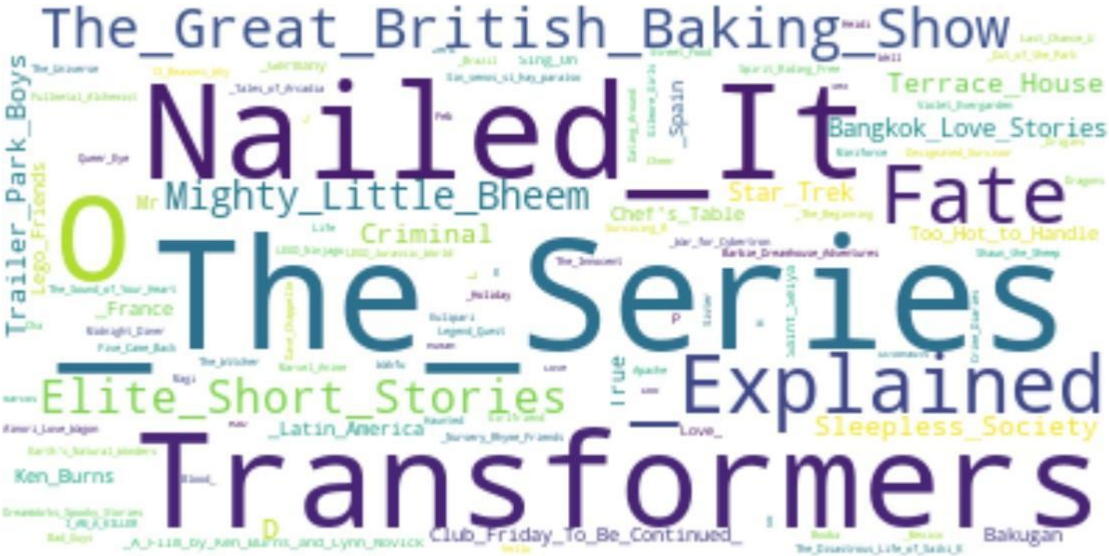


8- show the changing in Total numbers of TV_shows from 2005 - 2021 :

```
sns.set(font_scale=1.5)
ax = sns.regplot(x='year', y='total', data=df_tot, color='blue',
marker='*', scatter_kws={'s': 100})
ax.set(xlabel='Year', ylabel='Total Numbers')
ax.set_title('Total numbers of TV_shows from 2005 - 2021')
```



9- display the titles of the TV_shows by word clouds :



- (Data Analysis) .

  1- The common number of parts in a Netflix series is between 2 and 5 seasons .
  2- The common director in a Netflix series is (Alastair Fothergill) .
  3- Between 2018 and 2020 is The largest percentage of movies .
  4- In the Relationship between TV show's release year and year added to the platform (after 2018 is the best) .
  5- The most common rating in tv_show by netflix is (TV_MA) .
  6- The most countries producing films are the United States .
  7- After 2015 the increase in the number of films is noticeable