S04_T02

October 13, 2021

1 S04 T02: Visualització gràfica de Múltiples variables

[608]:
```
import numpy as np
import pandas as pd

import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
import matplotlib.dates as mdates

warnings.filterwarnings('ignore')

%matplotlib inline
```

1.1 Exercici 2: Repeteix l'exercici 1 amb el dataset que disposem en el repositori de GitHub PRE-PROCESSING-DATA, movies.dat

[609]:
```
movies = pd.read_csv("/Users/deliagonzalezmata/Documents/IT_Academy/Sprint_4/
 ↪S04_T02/Data-Science/Pre-processing-data/movies.dat",
                     sep="::", header=None,
                     names=["Num","Títol", "Gènere"])

# preparem el data set:
movies = movies.drop("Num", 1)

movies['Any'] = movies['Títol'].str.extract(r'\((\d{4})\)')
movies.replace('\(\d{4}\)', '', regex=True, inplace=True)
movies['Any'] = pd.to_numeric(movies['Any'])



movies.head()
```

[609]:

	Títol	Gènere	Any
0	Toy Story	Animation|Children's|Comedy	1995
1	Jumanji	Adventure|Children's|Fantasy	1995

```
2              Grumpier Old Men                 Comedy|Romance  1995
3             Waiting to Exhale                   Comedy|Drama  1995
4  Father of the Bride Part II                         Comedy  1995
```

[610]:
```
movies.shape
```

[610]:
```
(3883, 3)
```

[611]:
```
movies.ndim
```

[611]:
```
2
```

[612]:
```
movies.columns
```

[612]:
```
Index(['Títol', 'Gènere', 'Any'], dtype='object')
```

[613]:
```
movies.dtypes
```

[613]:
```
Títol     object
Gènere    object
Any        int64
dtype: object
```

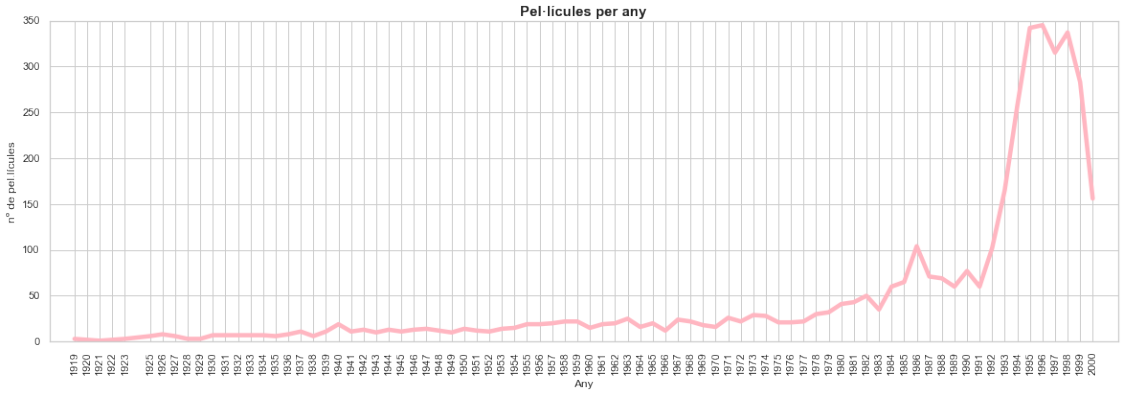
1.1.1 Pel • lícules per any

[614]:
```
pelis_per_any = movies['Any'].value_counts()
pelis_per_any = pelis_per_any.sort_index()

pelis_per_any
```

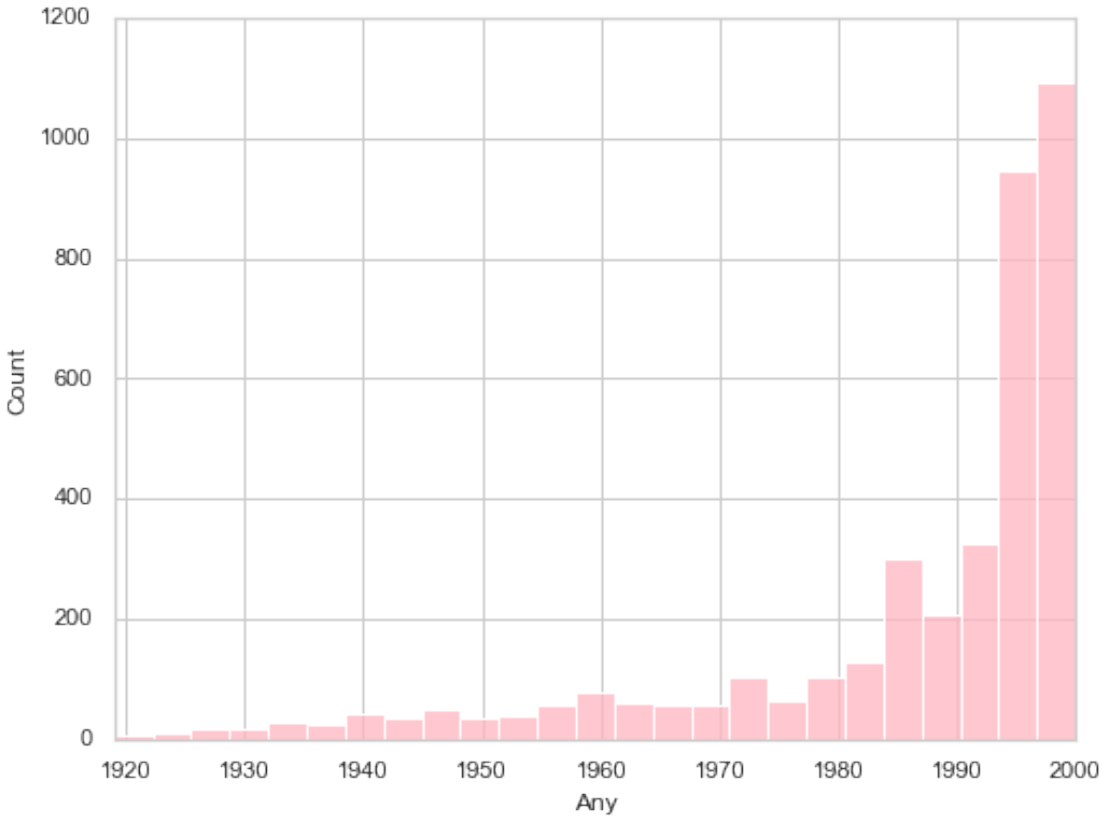
[614]:
```
1919      3
1920      2
1921      1
1922      2
1923      3
        ...
1996    345
1997    315
1998    337
1999    283
2000    156
Name: Any, Length: 81, dtype: int64
```

[615]:
```
fig,ax = plt.subplots(figsize=(20,6))

sns.lineplot(data = pelis_per_any, x=pelis_per_any.index,
             y=pelis_per_any.values, color = 'lightpink',
             linewidth = 4.5, ax=ax)
```

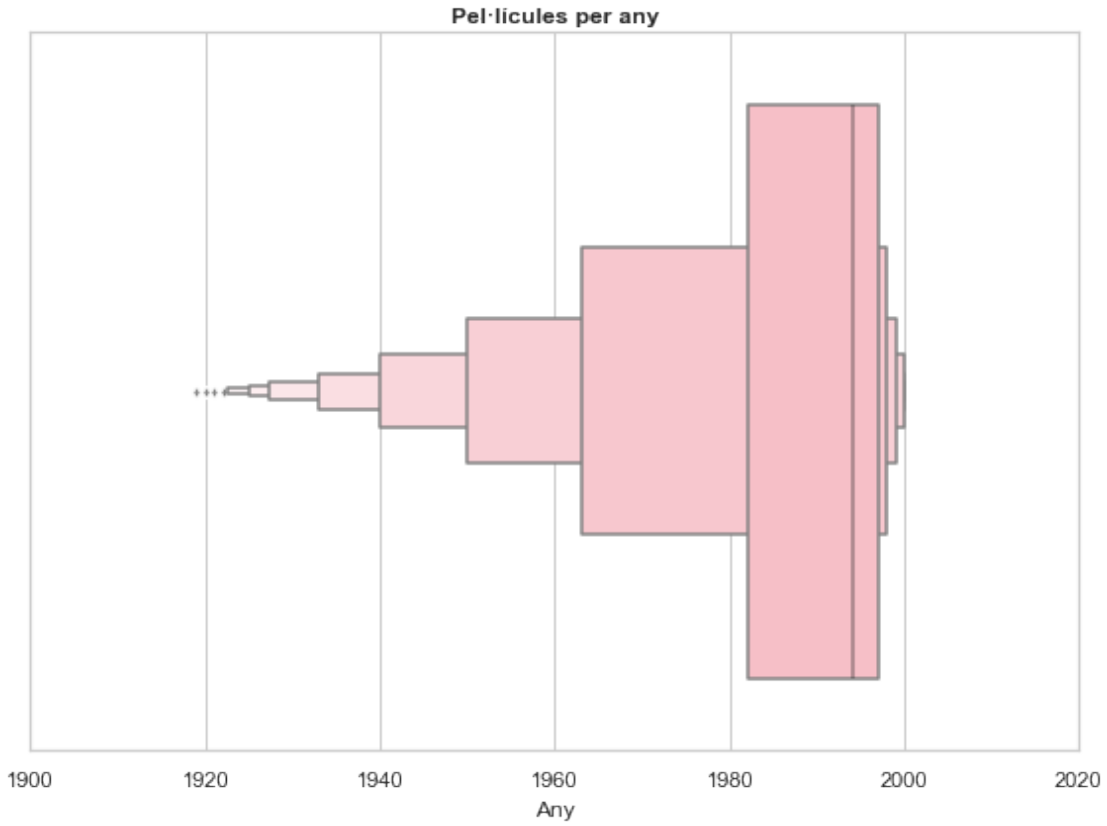
```
ax.ticklabel_format(useOffset=False, style='plain')
plt.title('Pel·lícules per any', fontweight = 'bold', size = 15)
ax.ticklabel_format(axis='both',useOffset=None)
plt.xlim(1917, 2002)
plt.xticks(pelis_per_any.index, rotation =90)
ax.set_xlabel('Any')
ax.set_ylabel('nº de pel.lícules')
plt.show()
```



[616]:
```
# HISTOGRAMA:
fig,ax = plt.subplots(figsize=(8,6))

sns.histplot(data=pelis_per_any, x=movies['Any'], color ='lightpink', bins = 25)
ax.ticklabel_format(useOffset=False, style='plain')
plt.xlim(1919, 2000)
plt.show()
```



[617]:
```
fig, ax = plt.subplots(figsize=(8,6))

sns.boxenplot(data=movies, x="Any", ax=ax, color = 'lightpink')

plt.title('Pel·lícules per any', fontweight = 'bold')
ax.set_xlabel('Any')

plt.tight_layout()
plt.show()
```



1.1.2 Pel • lícules per gènere

[618]:
```
def checkForGenre(series, name):
    list1 = []
    for i in series:
        if name in i:
            list1.append(1)
        else:
            list1.append(0)
    return list1

def getListOfUniqueGenres(series):
    listGenres = set()
    for i in series.str.split('|'):
        for j in i:
            listGenres.add(j)
    return listGenres

listGenres = getListOfUniqueGenres(movies.Gènere)
for genre in listGenres:
```

```
    movies[genre] = checkForGenre(movies.Gènere, genre)

df_genre = movies.iloc[:,3:]
df_genre.head()
```

[618]:

	War	Action	Animation	Mystery	Comedy	Drama	Film-Noir	Western	\
0	0	0	1	0	1	0	0	0	
1	0	0	0	0	0	0	0	0	
2	0	0	0	0	1	0	0	0	
3	0	0	0	0	1	1	0	0	
4	0	0	0	0	1	0	0	0	

	Children's	Fantasy	Sci-Fi	Romance	Documentary	Adventure	Musical	\
0	1	0	0	0	0	0	0	
1	1	1	0	0	0	1	0	
2	0	0	0	1	0	0	0	
3	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	

	Horror	Thriller	Crime
0	0	0	0
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0

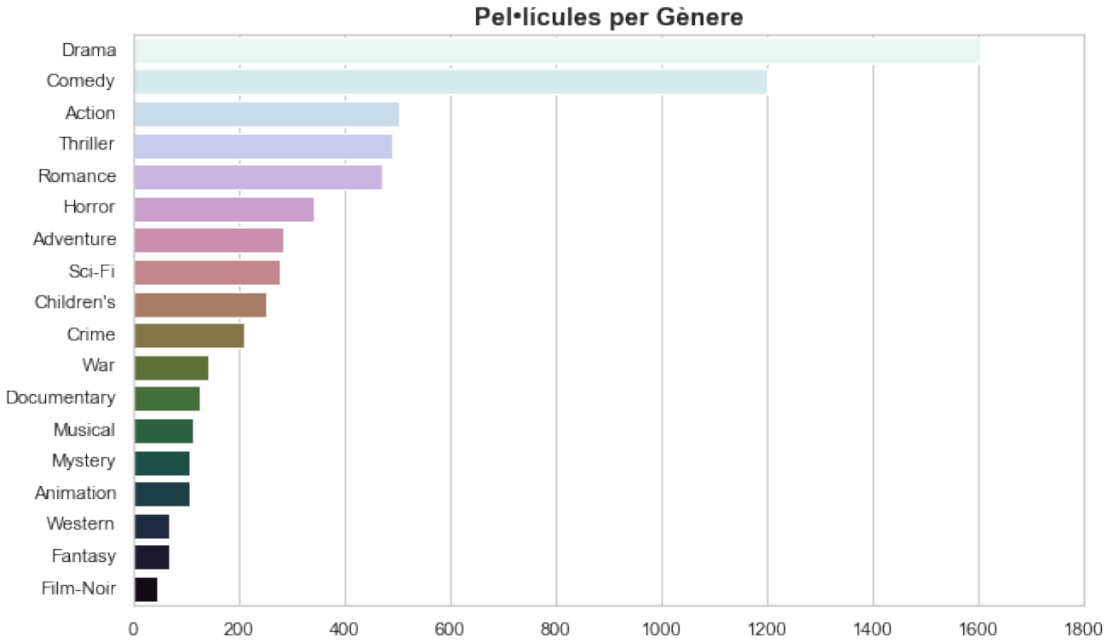
[619]:
```
pelis_per_genere = df_genre.sum().sort_values(ascending=False)
pelis_per_genere
```

[619]:
```
Drama          1603
Comedy         1200
Action          503
Thriller        492
Romance         471
Horror          343
Adventure       283
Sci-Fi          276
Children's      251
Crime           211
War             143
Documentary     127
Musical         114
Mystery         106
Animation       105
Western          68
Fantasy          68
Film-Noir        44
```

```
dtype: int64
```

grafiquem els gèneres per ordre d'importància

[620]:
```
fig, ax = plt.subplots(figsize=(10,6))

sns.barplot(x=pelis_per_genere.values, y=pelis_per_genere.index,
            ax=ax, orient ='h', palette = 'cubehelix_r');

plt.title('Pel•lícules per Gènere', fontweight = 'bold', size = 15);
```



1.1.3 Número de gèneres per película

[621]:
```
movies['num_generes']= movies.iloc[:, 3:].sum(axis=1)
movies.head()
```

[621]:

	Títol	Gènere	Any	War	\
0	Toy Story	Animation|Children's|Comedy	1995	0	
1	Jumanji	Adventure|Children's|Fantasy	1995	0	
2	Grumpier Old Men	Comedy|Romance	1995	0	
3	Waiting to Exhale	Comedy|Drama	1995	0	
4	Father of the Bride Part II	Comedy	1995	0	

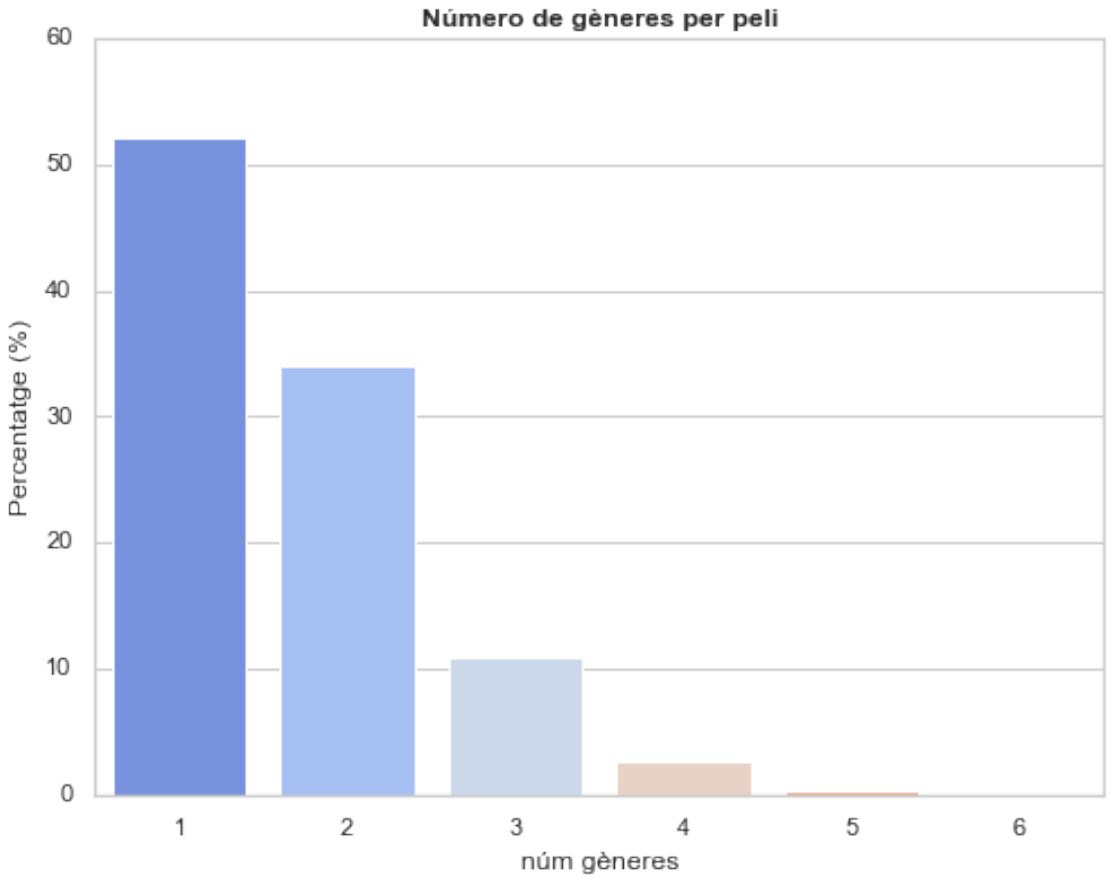
	Action	Animation	Mystery	Comedy	Drama	Film-Noir	…	Fantasy	Sci-Fi	\
0	0	1	0	1	0	0	…	0	0	
1	0	0	0	0	0	0	…	1	0	

```
2       0          0        0       1      0          0  …        0       0
3       0          0        0       1      1          0  …        0       0
4       0          0        0       1      0          0  …        0       0

   Romance  Documentary  Adventure  Musical  Horror  Thriller  Crime  \
0        0            0          0        0       0         0      0
1        0            0          1        0       0         0      0
2        1            0          0        0       0         0      0
3        0            0          0        0       0         0      0
4        0            0          0        0       0         0      0

   num_generes
0            3
1            3
2            2
3            2
4            1

[5 rows x 22 columns]
```

[622]:
```
movies_by_genere = movies.groupby('num_generes')['Títol'].count()
movies_by_genere
```

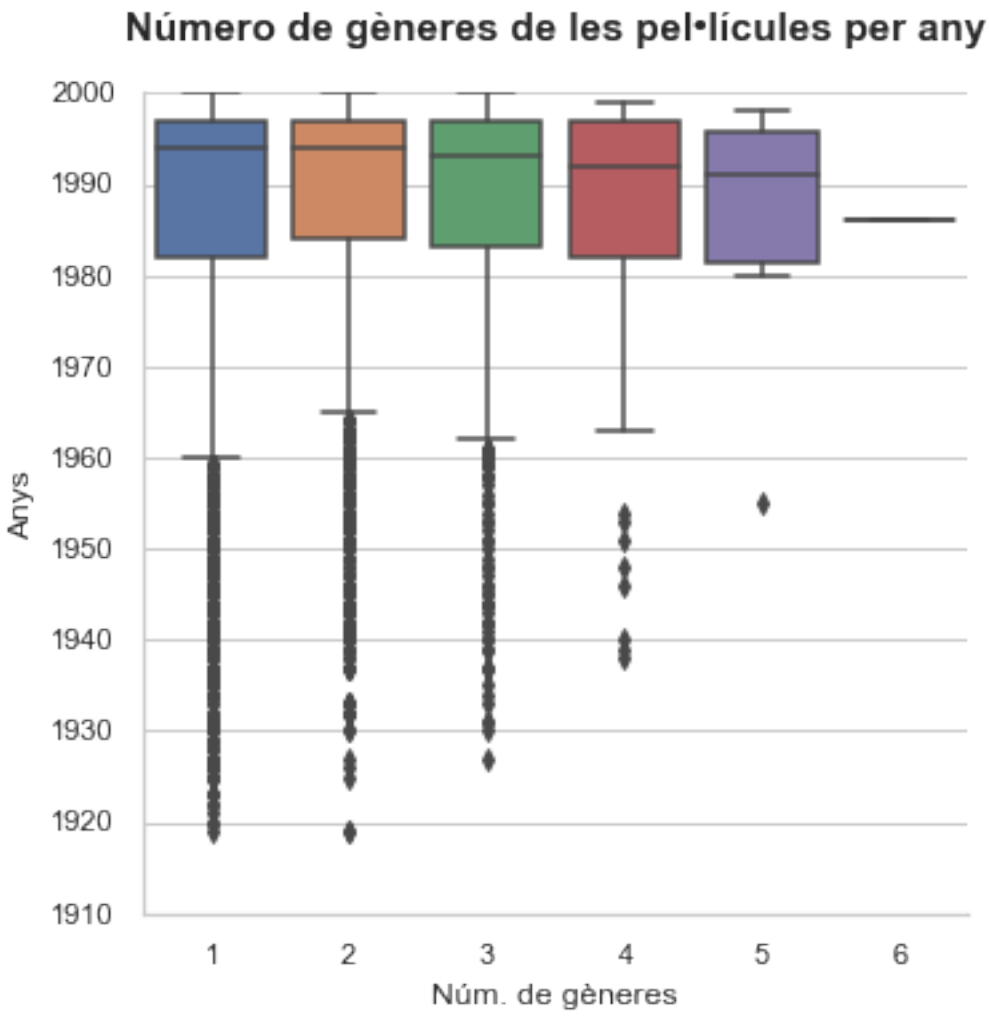
[622]:
```
num_generes
1    2025
2    1322
3     421
4     100
5      14
6       1
Name: Títol, dtype: int64
```

[623]:
```
percent_genere = (movies_by_genere / movies_by_genere.sum())*100


fig,ax = plt.subplots(figsize=(8,6))

x = percent_genere.index
y = percent_genere.values

sns.barplot(x = x, y=y, ax=ax, palette = 'coolwarm')
plt.title('Número de gèneres per peli', fontweight = 'bold')
ax.set_xlabel('núm gèneres')
ax.set_ylabel('Percentatge (%)')
plt.show()
```

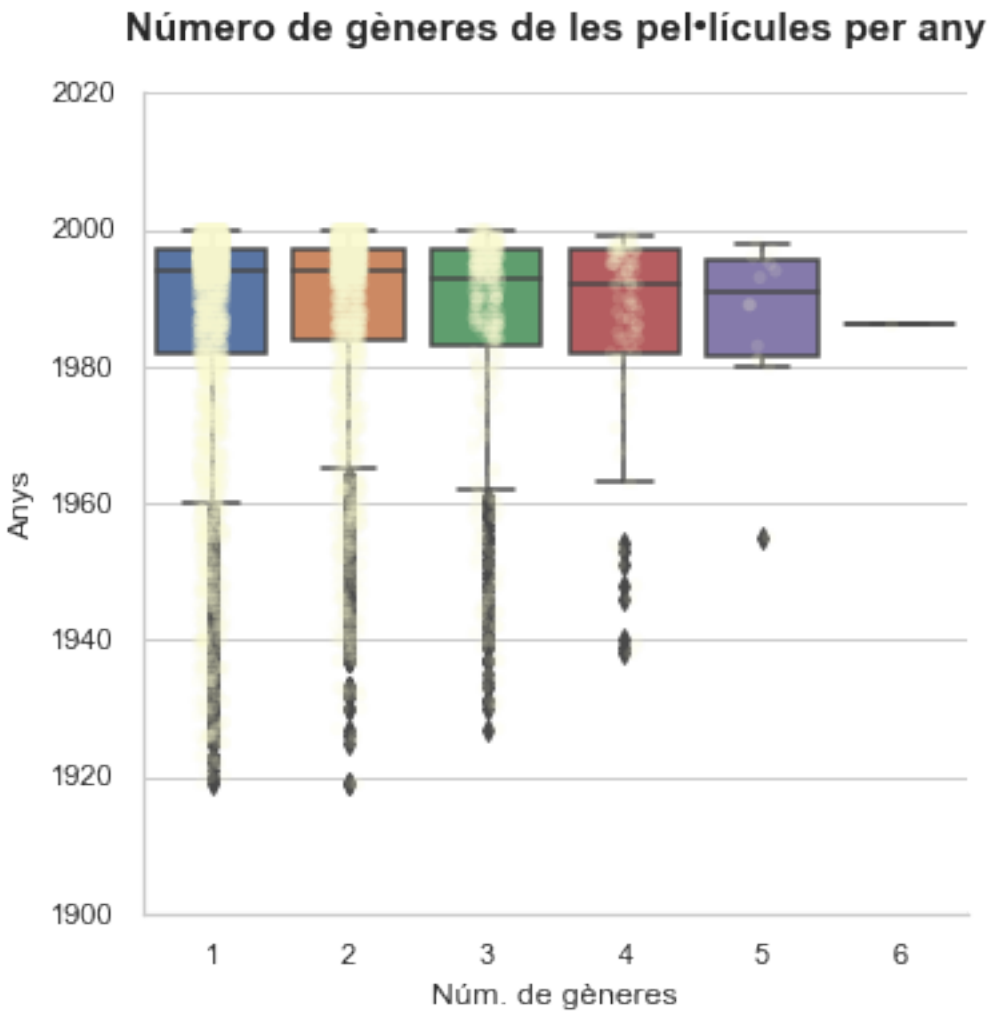


1.1.4 Evolució del número de gèneres vs. anys

[624]:
```
ax = sns.catplot(y='Any', x='num_generes', kind='box',
                 data=movies)

ax.set(xlabel = 'Núm. de gèneres', ylabel = 'Anys')
plt.ticklabel_format(axis="y", style="plain", useOffset = False)
plt.title("Número de gèneres de les pel•lícules per any",
                  fontsize=15, fontdict={"weight": "bold"}, pad = 20);
```



[625]:
```
ax = sns.catplot(y='Any', x='num_generes', kind='box',
                 data=movies)


plt.ticklabel_format(axis="y", style="plain", useOffset = False)
plt.title("Número de gèneres de les pel•lícules per any",
                  fontsize=15, fontdict={"weight": "bold"}, pad = 20)


ax = sns.stripplot(y='Any', x='num_generes', data=movies,␣
 ↪color='lightgoldenrodyellow', alpha= 0.15)

ax.set(xlabel = 'Núm. de gèneres', ylabel = 'Anys');
```


Número de gèneres de les pel•lícules per any
2020
2000
1980
1960
1940
1920
1900
Anys
1
2
3
4
5
6
Núm. de gèneres