1 Exercici 1

Realitza la pràctica del notebook a GitHub "03 EXAMINING DATA" amb seaborn i el dataset

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
df = pd.read_csv("https://raw.githubusercontent.com/IT-Academy-BCN/Data-Science/ma
```

In [4]:

```
df.shape
```

```
(244, 7)
```

In [5]:

```
df.columns
```

```
Index(['total_bill', 'tip', 'sex', 'smoker', 'day', 'time', 'size'], dtype='object')
```

In [24]:

```
df.head()
```

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

In [27]:
```
df.time.unique()
```

```
array(['Dinner', 'Lunch'], dtype=object)
```

In [28]:
```
df.time.nunique()
```

```
2
```

In [37]:
```
df.smoker.unique()
```

```
array(['No', 'Yes'], dtype=object)
```

In [36]:
```
df.sex.unique()
```

```
array(['Female', 'Male'], dtype=object)
```

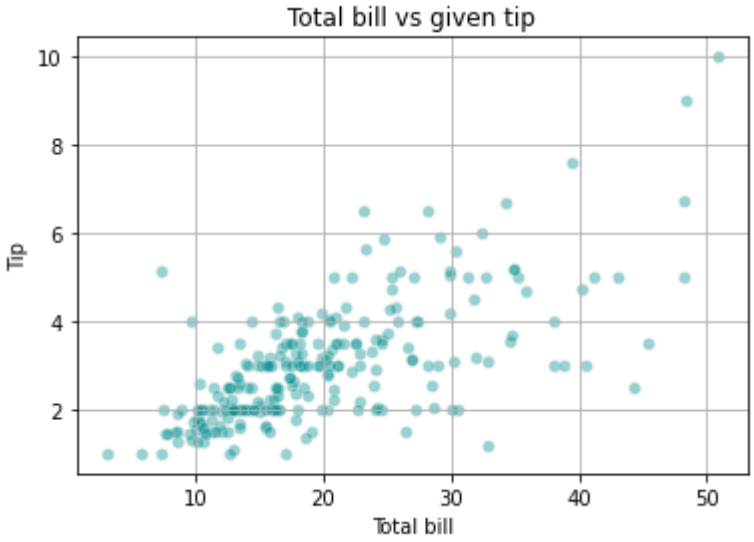
In [67]:
```
df.describe().round(2)
```

	total_bill	tip	party
count	244.00	244.00	244.00
mean	19.79	3.00	2.57
std	8.90	1.38	0.95
min	3.07	1.00	1.00
25%	13.35	2.00	2.00
50%	17.80	2.90	2.00
75%	24.13	3.56	3.00
max	50.81	10.00	6.00

1.1 Scatterplots

In [23]:

```
#plot
sns.scatterplot(x = df.total_bill, y = df.tip, color = "DarkCyan", alpha = 0.4)

plt.title("Total bill vs given tip")
plt.ylabel("Tip")
plt.xlabel("Total bill")
plt.grid()
plt.show()
```



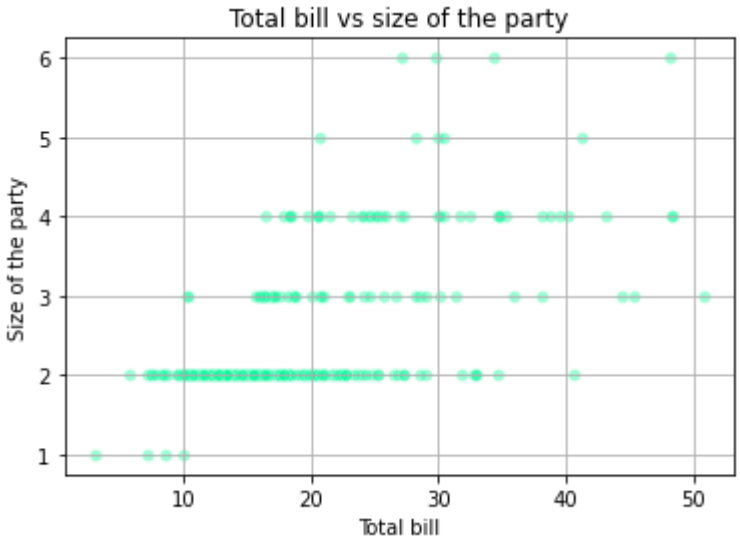
There is linear correlation between the total bill and the tip.

In [3]:

```
df.rename(columns = {"size" : "party"}, inplace = True)
```

In [41]:

```
sns.scatterplot(x = df.total_bill, y = df.party, color = "MediumSpringGreen", alph

plt.title("Total bill vs size of the party")
plt.ylabel("Size of the party")
plt.xlabel("Total bill")
plt.grid()
plt.show()
```



In [42]:

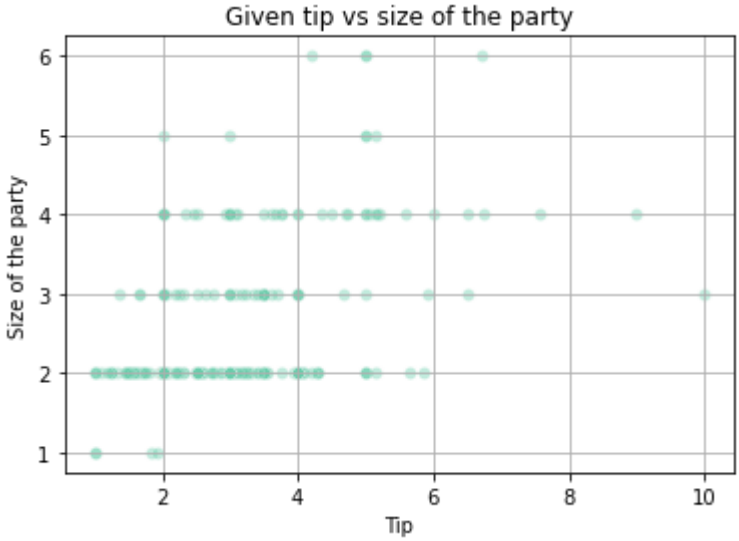
```
df.total_bill.corr(df.party)
```

```
0.5983151309049022
```

There is some correlaion between the total bill and the size of the party.

In [43]:

```
sns.scatterplot(x = df.tip, y = df.party, color = "MediumAquaMarine", alpha = 0.4)

plt.title("Given tip vs size of the party")
plt.ylabel("Size of the party")
plt.xlabel("Tip")
plt.grid()
plt.show()
```



In [44]:

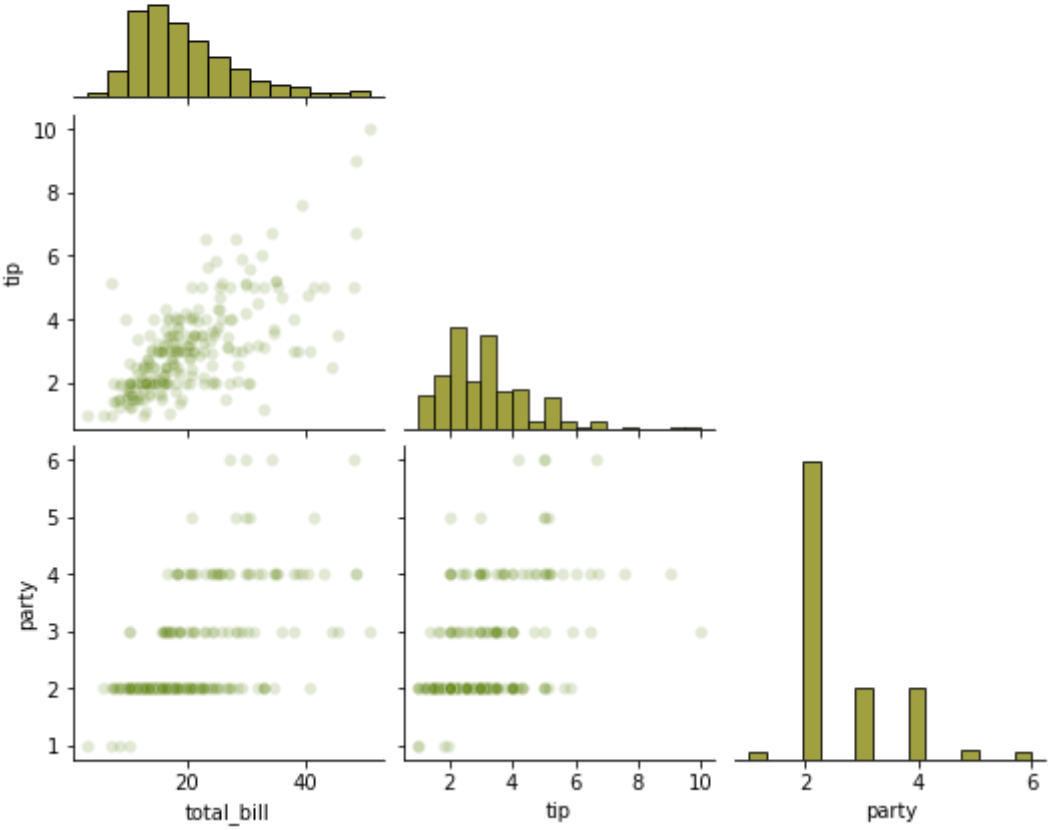
```
df.tip.corr(df.party)
```

```
0.4892987752303577
```

There is less correlation between the size of the party and the tip given. It is not a good pred

In [102]:

```
#matrix plot
sns.pairplot(df, diag_kind = "hist", plot_kws = {"alpha":0.2, "color":"OliveDrab"}
             diag_kws = { "color":"Olive"})
```



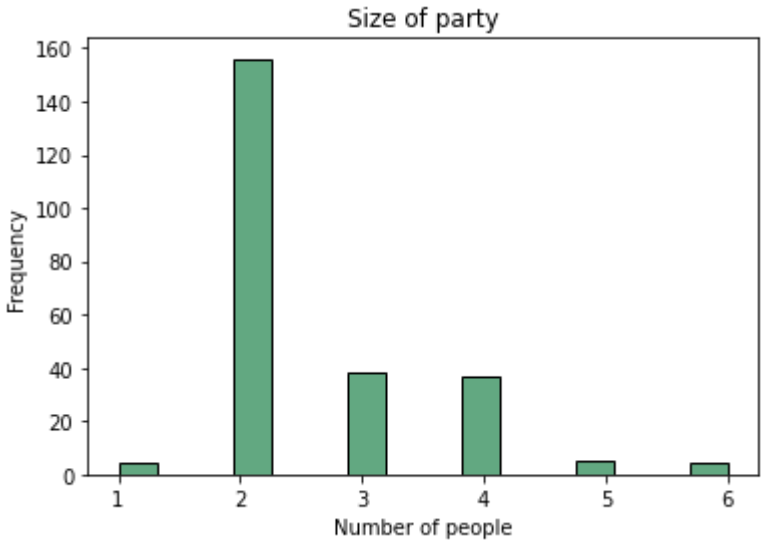
Matrix plot that gathers the above plots.

1.2 Histograms

In [104]:

```
sns.histplot(data = df, x = "party", color = "SeaGreen")
plt.title("Size of party")
plt.ylabel("Frequency")
plt.xlabel("Number of people")
```

```
Text(0.5, 0, 'Number of people')
```



The majority of people going to the restaurant go in groups of 2. That means it is unimodal a

In [78]:

```
df.total_bill.mean() #tip average
```

```
19.785942622950824
```

In [77]:

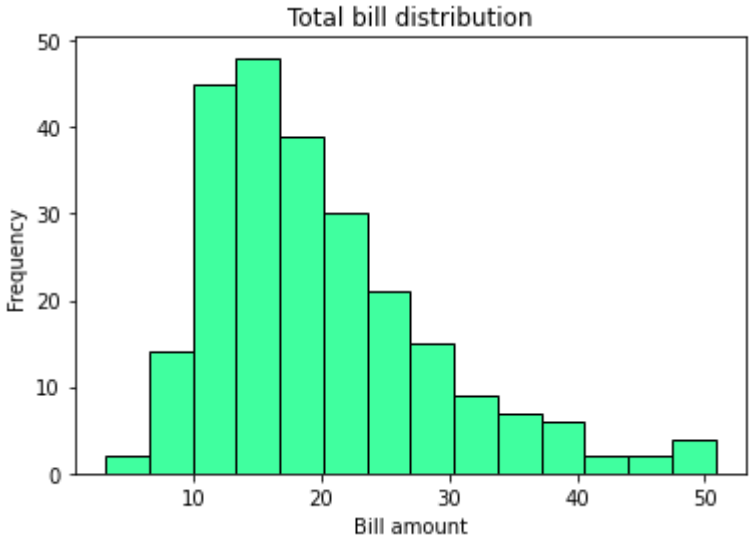
```
df.total_bill.std()  #tip standard deviation
```

```
8.902411954856856
```

In [82]:

```
sns.histplot(data = df, x = "total_bill", color = "SpringGreen")
plt.title("Total bill distribution")
plt.ylabel("Frequency")
plt.xlabel("Bill amount")
```

```
Text(0.5, 0, 'Bill amount')
```



This distribution is unimodal, and is very skewed to the left. Most of the counts are close to t within less than a standard deviation away from it.

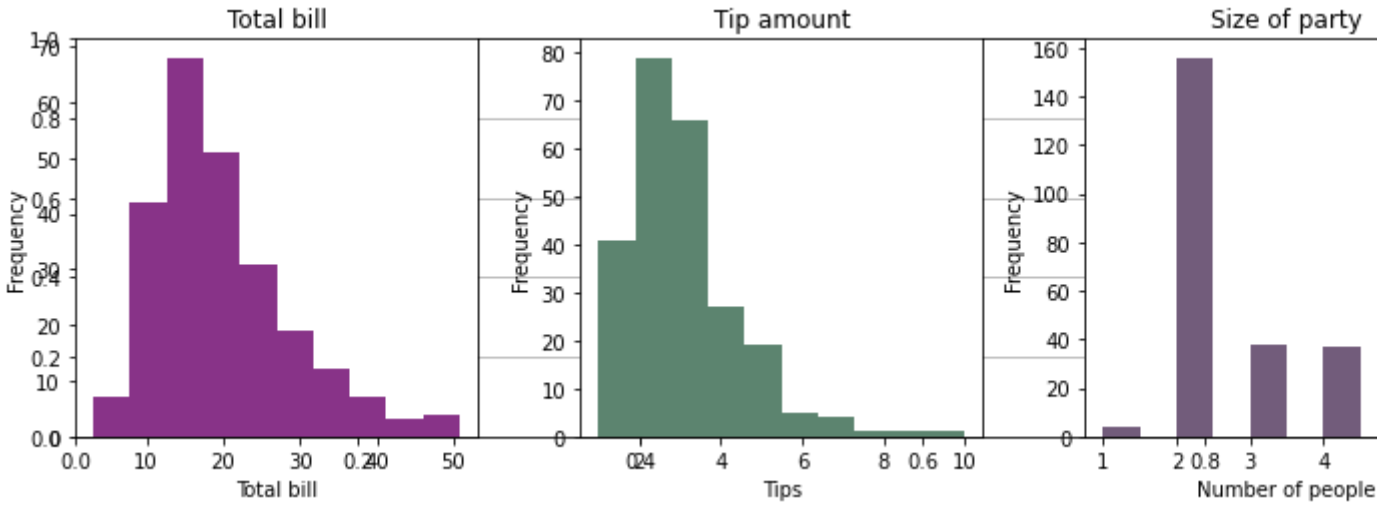
In [79]:

```
df.head()
```

	total_bill	tip	sex	smoker	day	time	party
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

In [106]:

```
fig = plt.figure(figsize = (11, 4))

ax1 = fig.add_subplot(1, 3, 1)

ax1.hist(df["total_bill"], color = "#883388")
plt.title("Total bill")
plt.ylabel("Frequency")
plt.xlabel("Total bill")

ax2 = fig.add_subplot(1, 3, 2)

ax2.hist(df["tip"], color = "#5C846F")
plt.title("Tip amount")
plt.ylabel("Frequency")
plt.xlabel("Tips")

ax3 = fig.add_subplot(1, 3, 3)

ax3.hist(df["party"], color = "#725C7B")
plt.title("Size of party")
plt.ylabel("Frequency")
plt.xlabel("Number of people")

plt.tight_layout()
```



All of the above plots show that they are all very skewed to the left and unimodal. The mode

cases and the great majority of values are comprised in the range of the mean ± one standa

In [96]:

```
df.shape
```

```
(244, 7)
```

In [99]:

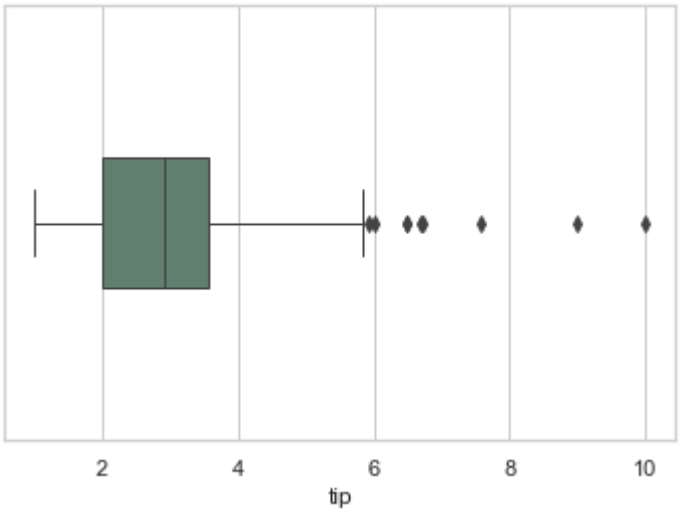
```
df["tip"].describe().round(2)
```

```
count    244.00
mean       3.00
std        1.38
min        1.00
25%        2.00
50%        2.90
75%        3.56
max       10.00
Name: tip, dtype: float64
```

In [127]:

```
sns.set(style = "whitegrid")
sns.boxplot(x = df["tip"], color = "#5C846F", fliersize = 5,  linewidth = 1, width
```

```
<AxesSubplot:xlabel='tip'>
```

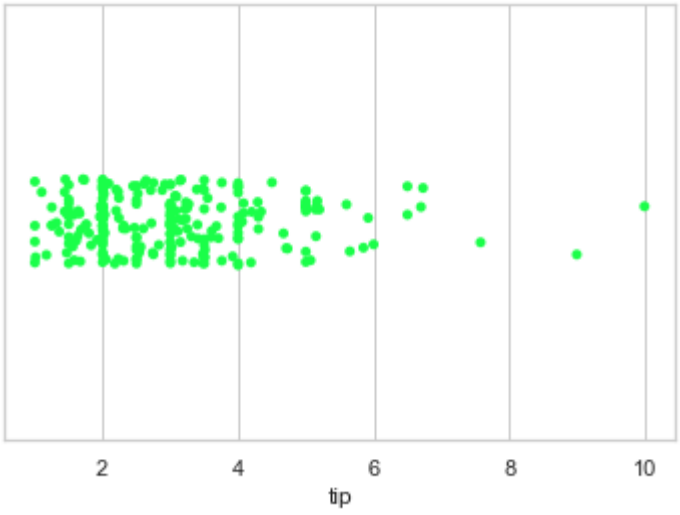


50 % of all values go between 2 and 3.56. Although the median is a bit closer to the third qua on the interquartile range tend to be below it, not above it, as evidenced by the shorter whisk there are outliers that are very high and far from the other values.

In [131]:

```
sns.stripplot(x = df["tip"], color = "#1AFE49")
```

```
<AxesSubplot:xlabel='tip'>
```

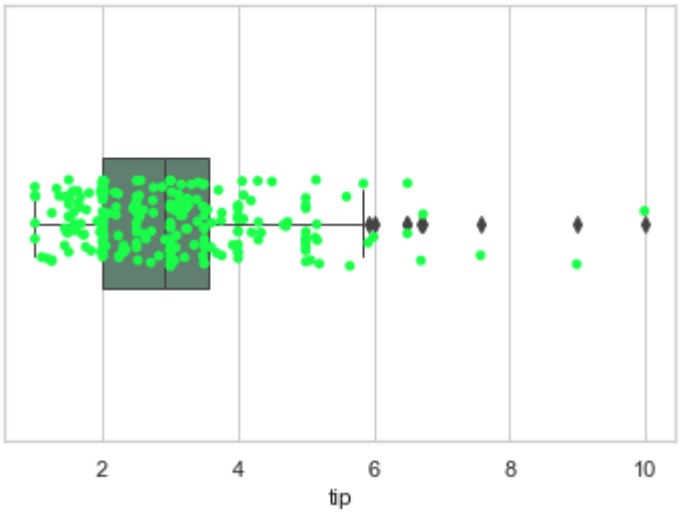


As the boxplot, this scatter shows that the majority of values are in a small range and there a

In [128]:

```
sns.boxplot(x = df["tip"], color = "#5C846F", fliersize = 5,  linewidth = 1, width
sns.stripplot(x = df["tip"], color = "#1AFE49")
```

```
<AxesSubplot:xlabel='tip'>
```

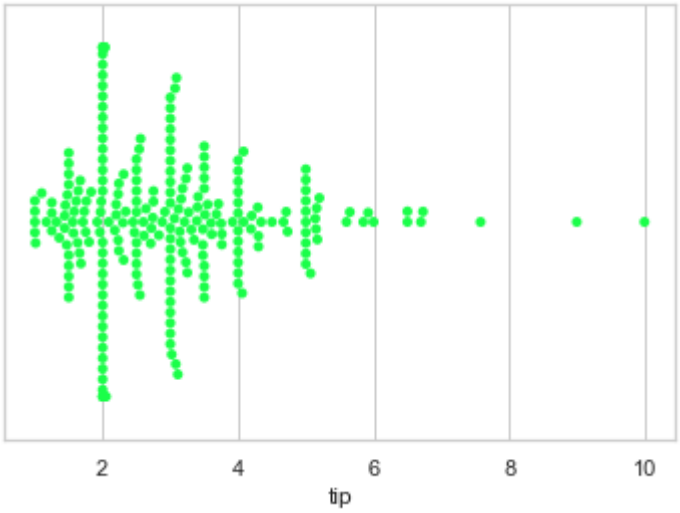


Both plots together confirm that they overlap and show that most of the values are in the inte

In [133]:

```
sns.swarmplot(x = df["tip"], color = "#1AFE49")
```

```
<AxesSubplot:xlabel='tip'>
```

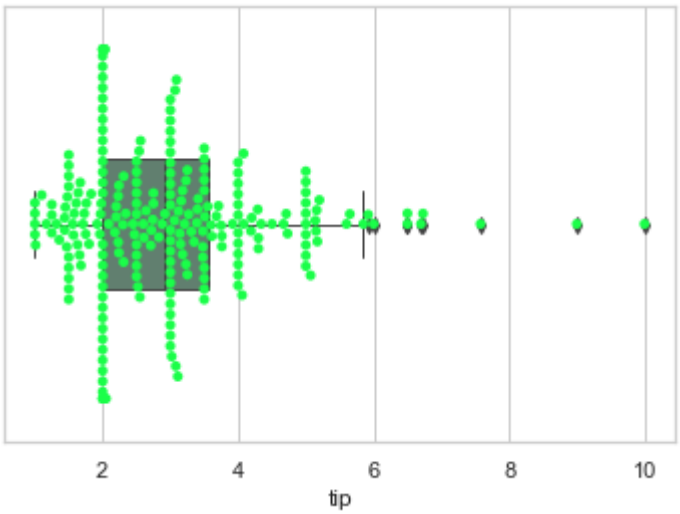


This plot allows us to see more clearly wear the majority of the values fall, and shows a rathe
the median.

In [134]:

```
sns.boxplot(x = df["tip"], color = "#5C846F", fliersize = 5,  linewidth = 1, width
sns.swarmplot(x = df["tip"], color = "#1AFE49")
```

```
<AxesSubplot:xlabel='tip'>
```



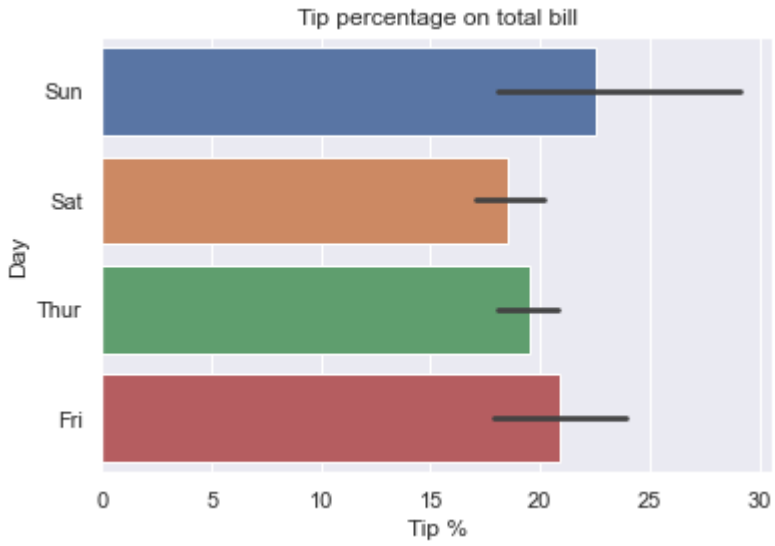
In [7]:

```
df['tip_pct'] = round((df['tip'] / (df['total_bill'] - df['tip'])) * 100, 2)
df.head()
```

	total_bill	tip	sex	smoker	day	time	party	tip_pct
0	16.99	1.01	Female	No	Sun	Dinner	2	6.32
1	10.34	1.66	Male	No	Sun	Dinner	3	19.12
2	21.01	3.50	Male	No	Sun	Dinner	3	19.99
3	23.68	3.31	Male	No	Sun	Dinner	2	16.25
4	24.59	3.61	Female	No	Sun	Dinner	4	17.21

In [11]:

```
sns.barplot(x = 'tip_pct', y = 'day', data = df, orient = "h")
plt.title("Tip percentage on total bill")
plt.ylabel("Day")
plt.xlabel("Tip %")
plt.show()
sns.set(style = "darkgrid")
```



In [13]:

```
df.describe().round(2)
```

	total_bill	tip	party	tip_pct
count	244.00	244.00	244.00	244.00
mean	19.79	3.00	2.57	20.21
std	8.90	1.38	0.95	16.34
min	3.07	1.00	1.00	3.70
25%	13.35	2.00	2.00	14.83
50%	17.80	2.90	2.00	18.31
75%	24.13	3.56	3.00	23.68
max	50.81	10.00	6.00	245.24

In [16]:

```
round(df.describe(include = 'all'), 2)
```

	total_bill	tip	sex	smoker	day	time	party	tip_pct
count	244.00	244.00	244	244	244	244	244.00	244.00
unique	NaN	NaN	2	2	4	2	NaN	NaN
top	NaN	NaN	Male	No	Sat	Dinner	NaN	NaN
freq	NaN	NaN	157	151	87	176	NaN	NaN
mean	19.79	3.00	NaN	NaN	NaN	NaN	2.57	20.21
std	8.90	1.38	NaN	NaN	NaN	NaN	0.95	16.34
min	3.07	1.00	NaN	NaN	NaN	NaN	1.00	3.70
25%	13.35	2.00	NaN	NaN	NaN	NaN	2.00	14.83
50%	17.80	2.90	NaN	NaN	NaN	NaN	2.00	18.31
75%	24.13	3.56	NaN	NaN	NaN	NaN	3.00	23.68
max	50.81	10.00	NaN	NaN	NaN	NaN	6.00	245.24

In [17]:

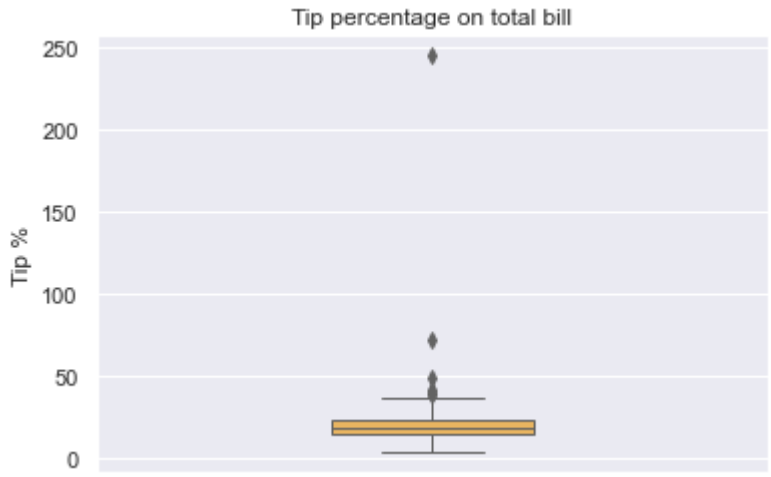
```
df.isnull().sum() / len(df) # % of null values
```

```
total_bill    0.0
tip           0.0
sex           0.0
smoker        0.0
day           0.0
time          0.0
party         0.0
tip_pct       0.0
dtype: float64
```

1.3 One variable

In [22]:

```
sns.boxplot(y = "tip_pct", data = df[df.tip < 10],  color = "#FFBA49", fliersize =
            orient = 'v', linewidth = 1 , width = 0.3)
plt.title("Tip percentage on total bill")
plt.ylabel("Tip %")
plt.show()
```

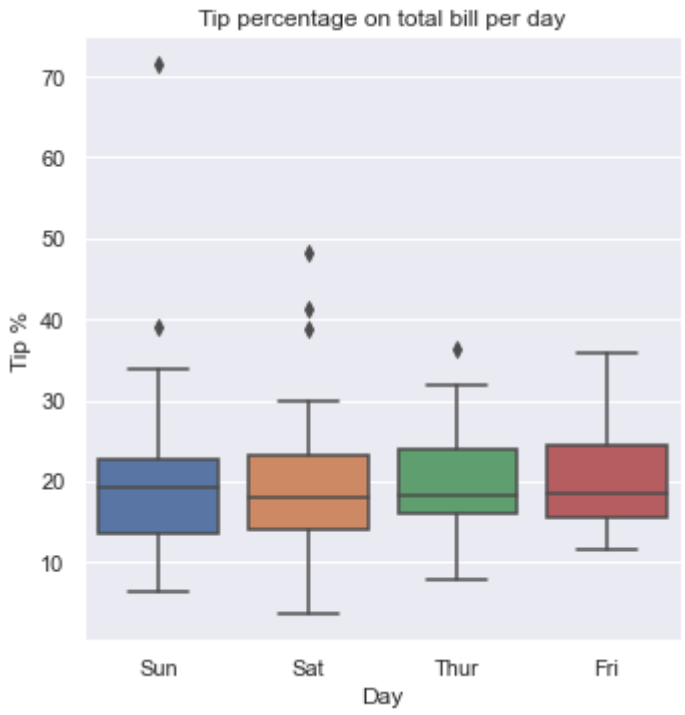


There's a value that is extremely far off from the rest, including the other outliers. Since it say bill, it could mean it was wrongly introduced into the dataset.

1.4 Two variables

In [25]:

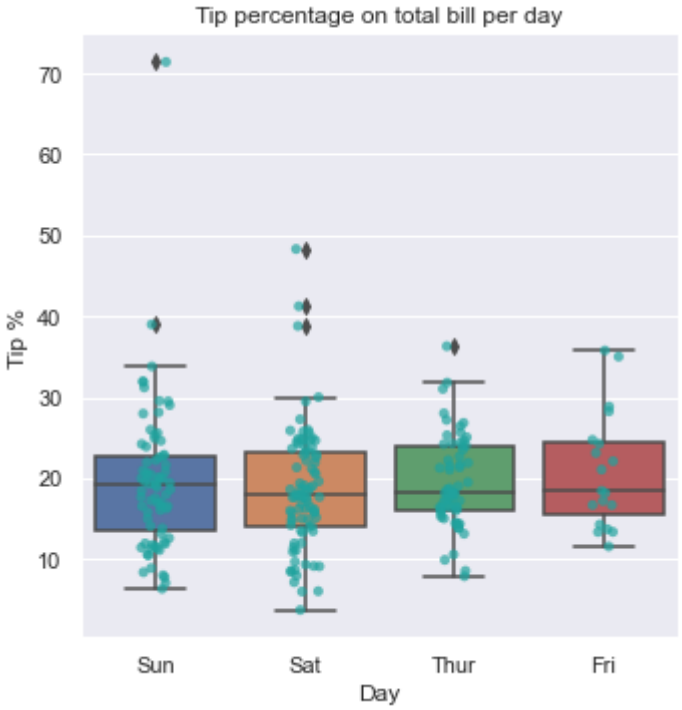
```
sns.catplot(x = 'day', y = 'tip_pct', kind = 'box', data = df[df.tip_pct < 245])
plt.title("Tip percentage on total bill per day")
plt.ylabel("Tip %")
plt.xlabel("Day")
plt.show()
```



Having removed that extreme outlier, the values are closer to the median, although there are
day, the median does not vary much, but the bottom whisker, that means the lower values, d
Saturdays and the highest is on Fridays.

In [28]:

```
sns.catplot(x = 'day', y = 'tip_pct', kind = 'box',data = df[df.tip_pct < 245])
sns.stripplot(x = 'day', y = 'tip_pct', data = df[df.tip_pct < 245], orient = 'v',
              color = "#20A39E", alpha = 0.7)
plt.title("Tip percentage on total bill per day")
plt.ylabel("Tip %")
plt.xlabel("Day")
plt.show()
```

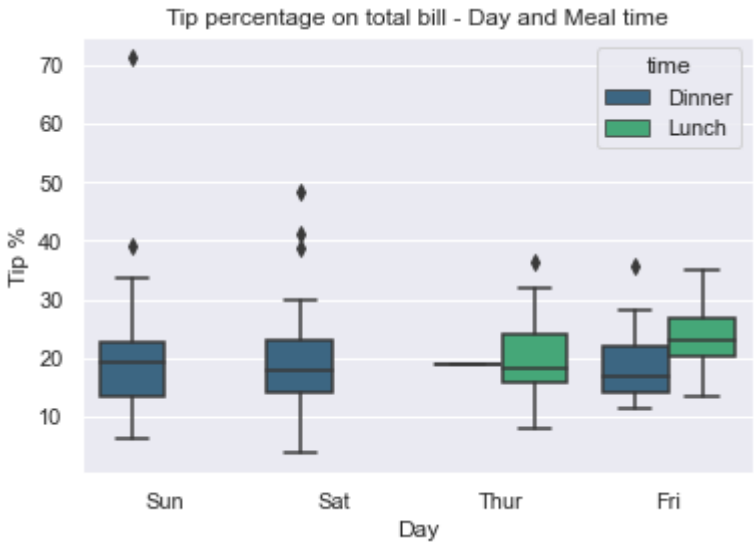


We can see the scatter shows that most of the values fall between the whiskers of the box pl

1.5 Three variables

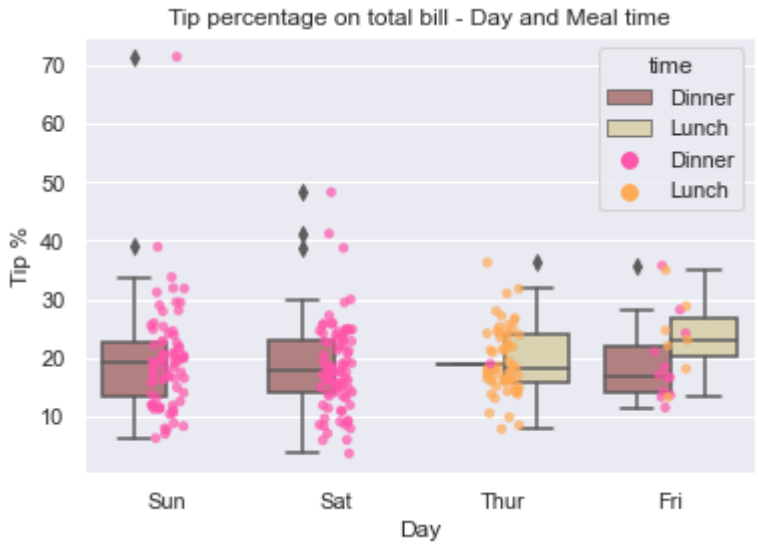
In [33]:

```
sns.boxplot(x = 'day', y = 'tip_pct', hue = 'time', data = df[df.tip_pct < 245], p
plt.title("Tip percentage on total bill - Day and Meal time")
plt.ylabel("Tip %")
plt.xlabel("Day")
plt.show()
```



In [41]:

```
sns.boxplot(x = 'day', y = 'tip_pct', hue = 'time', data = df[df.tip_pct < 245], p
sns.stripplot(x = 'day', y = 'tip_pct', hue = 'time', data = df[df.tip_pct < 245],
              palette = "spring", alpha = 0.7)
plt.title("Tip percentage on total bill - Day and Meal time")
plt.ylabel("Tip %")
plt.xlabel("Day")
plt.show()
```

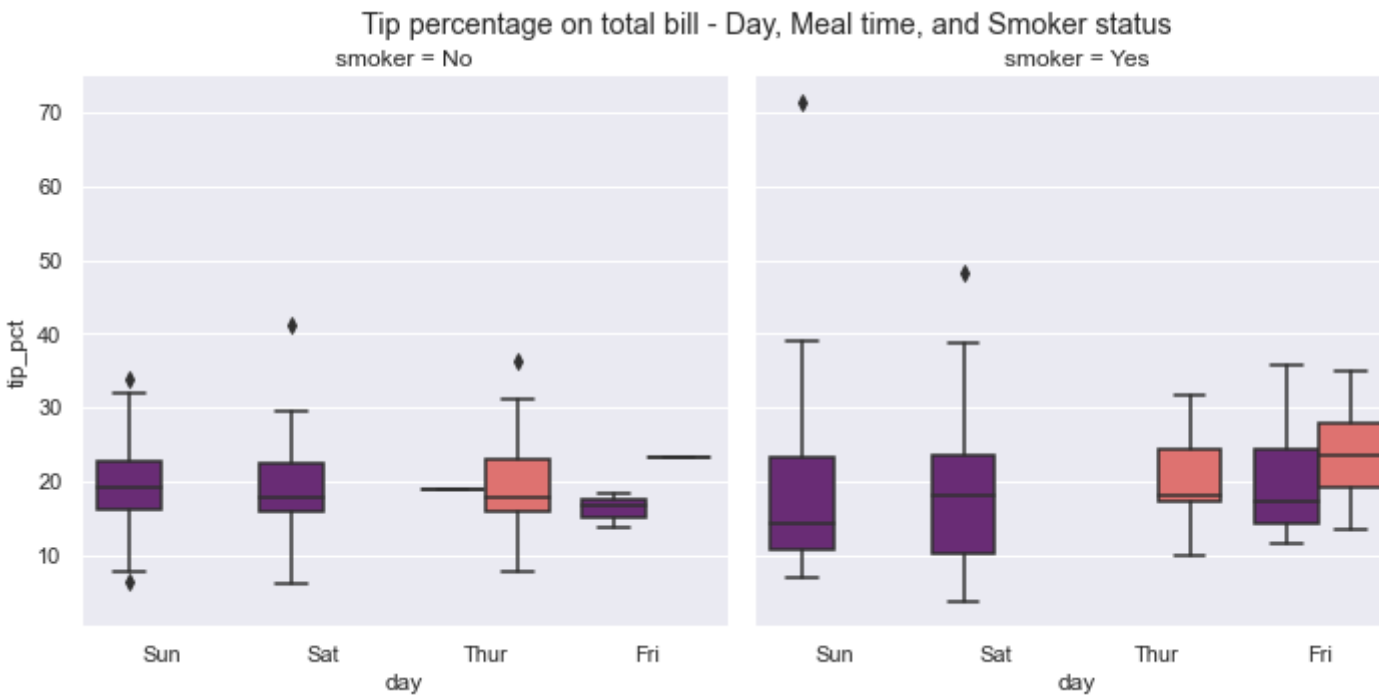


It seems there were no lunches on Sunday and Saturday, and hardly any dinners on Thursda were both meal times is also the day where there is less count of either. On average, people generous on a time, compared to another, but the outliers tend to be higher on dinner time.

1.6 Four variables

In [53]:

```
plot = sns.catplot(x = 'day', y = 'tip_pct', hue = 'time', col = 'smoker', kind =
            data = df[df.tip_pct < 245], palette = "magma")
plot.fig.suptitle("Tip percentage on total bill - Day, Meal time, and Smoker statu
plot.fig.subplots_adjust(top = 0.89)
plt.show()
```



There is more variability on tip generosity on smokers than in the the non smoker group.

2 Exercici 2

Repeteix l'exercici 1 amb el dataset que disposem en el repositori de GitHub PRE-PROCES

In [13]:

```
movies = pd.read_table(
    "https://raw.githubusercontent.com/IT-Academy-BCN/Data-Science/main/Pre-proces
    sep = "::", engine = "python", encoding = "ISO-8859-1", header = None,
    names = ["id", "title", "genres"], index_col = 0)
```

In [4]:

```
movies.head()
```

id	title	genres
1	Toy Story (1995)	Animation|Children's|Comedy
2	Jumanji (1995)	Adventure|Children's|Fantasy
3	Grumpier Old Men (1995)	Comedy|Romance
4	Waiting to Exhale (1995)	Comedy|Drama
5	Father of the Bride Part II (1995)	Comedy

In [14]:

```
#split name and year into two columns
movies[["title", "year"]] = movies["title"].str.rsplit("(", expand = True, n = 1)
```

In [15]:

```
#remove end parenthesis
movies["year"] = movies["year"].str.rstrip(")")
```

In [16]:

```
#create dummy list for genre
dmovies = movies.genres.str.get_dummies(sep = "|")
dmovies.head()
```

id	Action	Adventure	Animation	Children's	Comedy	Crime	Documentary	Drama	Fantasy	Film Noi
1	0	0	1	1	1	0	0	0	0	0
2	0	1	0	1	0	0	0	0	1	0
3	0	0	0	0	1	0	0	0	0	0
4	0	0	0	0	1	0	0	1	0	0
5	0	0	0	0	1	0	0	0	0	0

In [17]:

```
#add dummy list to main dataset
movies = movies.join(dmovies)
```

In [15]:

```
movies.head()
```

id	title	genres	year	Action	Adventure	Animation	Children's	Comedy	Cri
1	Toy Story	Animation|Children's|Comedy	1995	0	0	1	1	1	0
2	Jumanji	Adventure|Children's|Fantasy	1995	0	1	0	1	0	0
3	Grumpier Old Men	Comedy|Romance	1995	0	0	0	0	1	0
4	Waiting to Exhale	Comedy|Drama	1995	0	0	0	0	1	0
5	Father of the Bride Part II	Comedy	1995	0	0	0	0	1	0

5 rows × 21 columns

In [9]:

```
movies.shape
```

```
(3883, 21)
```

2.1 One variable

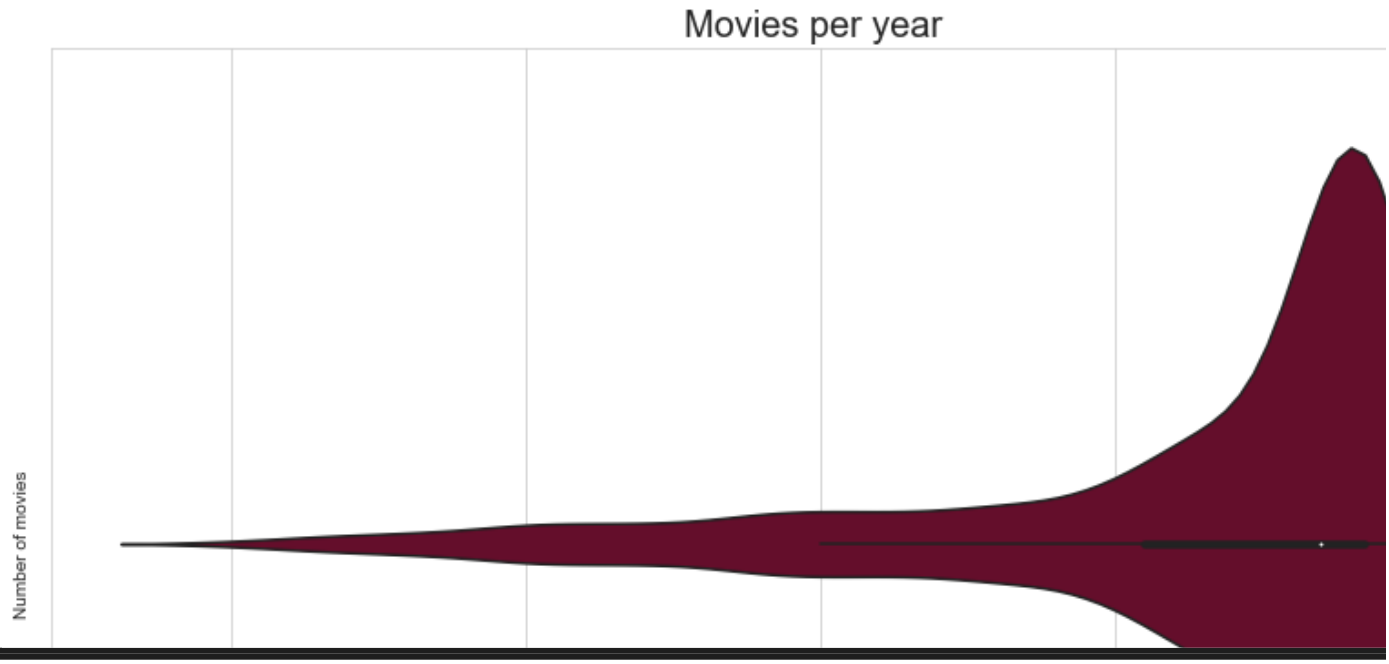
In [18]:

```
#make year variables into int
movies["year"] = movies["year"].astype(str).astype(int)
```

In [49]:

```
sns.set_style("whitegrid")
plt.figure(figsize = (15, 10))
sns.violinplot(x = "year", data = movies, color = "#720026")
plt.title("Movies per year", fontsize = 20)
plt.ylabel("Number of movies")
plt.xlabel("Year")
```

```
Text(0.5, 0, 'Year')
```



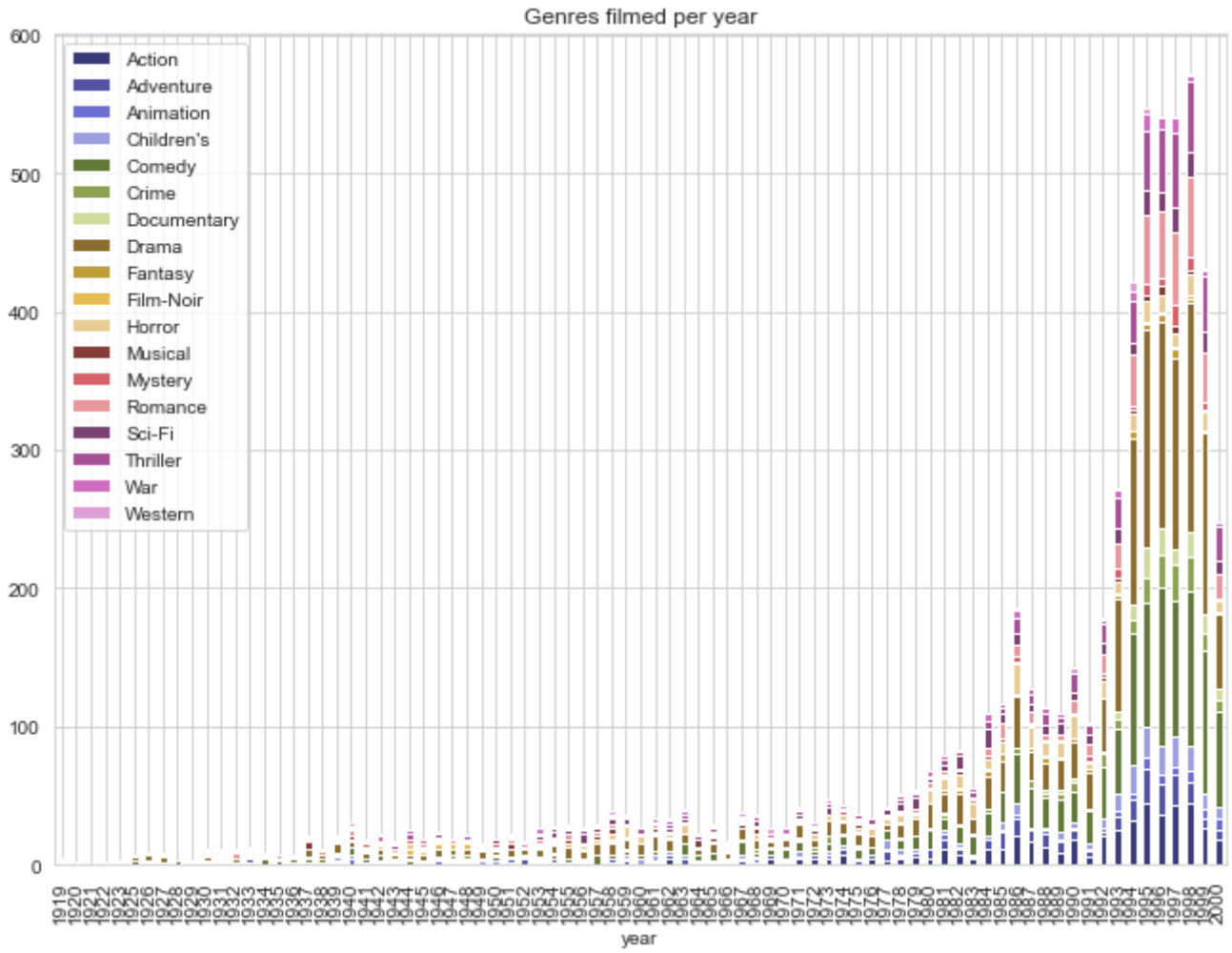
The majority of the movies in the dataset were made in the 1990s.

2.2 Two variables

In [32]:

```
movies.groupby("year").sum().plot(kind = "bar", stacked = True, title = "Genres fi
                                  figsize = (11, 8), colormap = "tab20b")
```

```
<AxesSubplot:title={'center':'Genres filmed per year'}, xlabel='year'>
```



It seems that throughout the years, Drama is the most filmed genre, followed by Comedy, wh
1980s. There's also a rise in action movies in the 1990s.

3 Exercici 3

En aquest exercici no us donarem gaires indicacions perquè volem que ens mostreu la vostr

gràfiques i interpretacions del dataset "movies.dat" del exercici anterior.

In [21]:

```
df = movies.groupby("year").sum().reset_index()
```

In [23]:

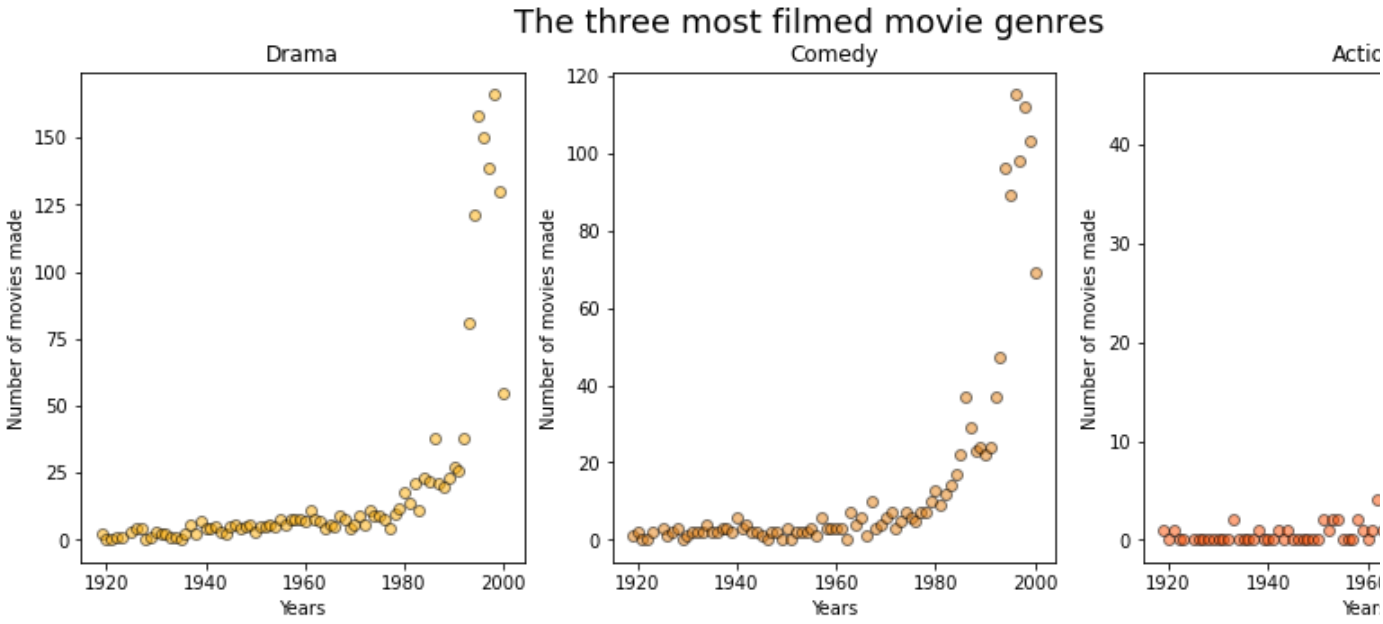
```
df.tail()
```

ren's	Comedy	Crime	Documentary	Drama	Fantasy	Film-Noir	Horror	Musical	Mystery	Romance	Sci F
	115	23	19	150	5	2	12	7	6	48	14
	98	26	11	139	6	2	10	5	15	52	18
	112	25	18	166	2	3	15	3	10	58	17
	103	12	15	130	2	0	14	1	5	37	15
	69	8	8	55	1	0	8	1	1	17	10

In [27]:

```
fig = plt.figure(figsize = (15, 5))
plt.suptitle("The three most filmed movie genres", size = 18)

ax1 = fig.add_subplot(1, 3, 1)
ax1.scatter(x = df.year, y = df.Drama, c = "#FFAE03", alpha = 0.5, edgecolors = "b
plt.title("Drama")
plt.ylabel("Number of movies made")
plt.xlabel("Years")

ax2 = fig.add_subplot(1, 3, 2)
ax2.scatter(x = df.year, y = df.Comedy, c = "#E67F0D", alpha = 0.5, edgecolors = "
plt.title("Comedy")
plt.ylabel("Number of movies made")
plt.xlabel("Years")

ax3 = fig.add_subplot(1, 3, 3)
ax3.scatter(x = df.year, y = df.Action, c = "#FE4E00", alpha = 0.5, edgecolors = "
plt.title("Action")
plt.ylabel("Number of movies made")
plt.xlabel("Years")
```

```
Text(0.5, 0, 'Years')
```

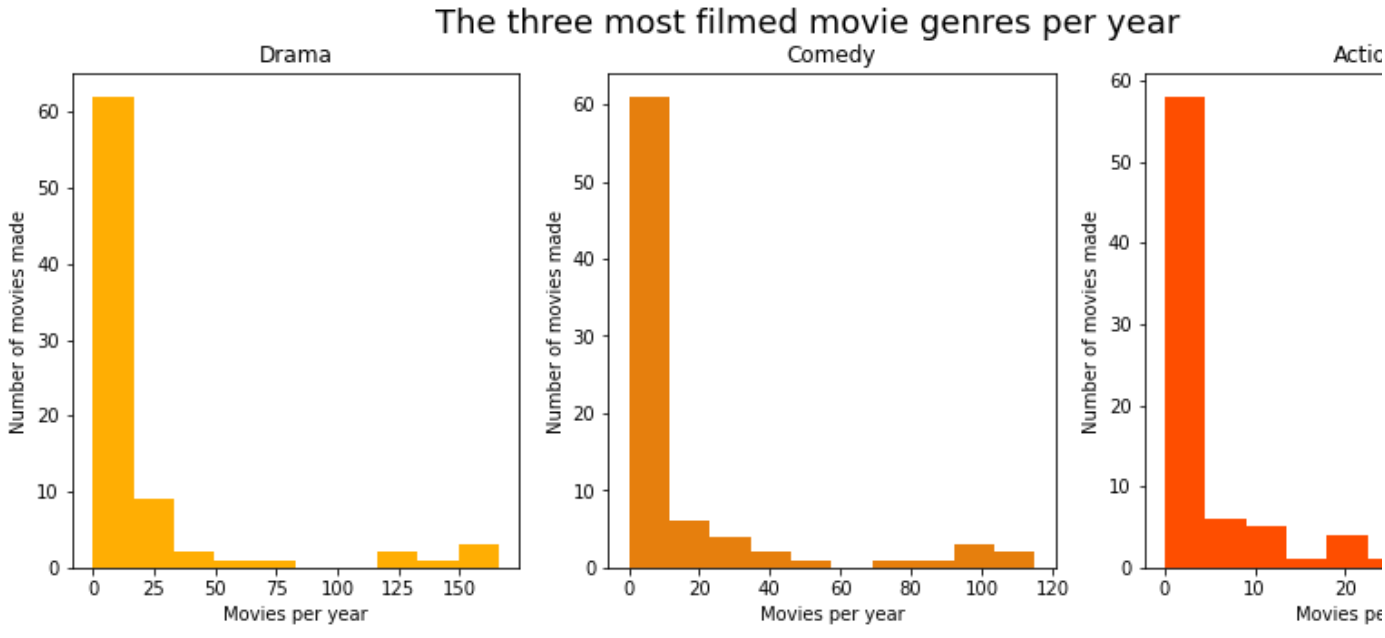


In all of these genres there's an exponential rise of the number of movies, although this coul
are more movies from the 80s and 90s in the dataset.

In [30]:

```
fig = plt.figure(figsize = (15, 5))
plt.suptitle("The three most filmed movie genres per year", size = 18)

ax1 = fig.add_subplot(1, 3, 1)
ax1.hist(x = df.Drama, color = "#FFAE03")
plt.title("Drama")
plt.ylabel("Number of movies made")
plt.xlabel("Movies per year")

ax2 = fig.add_subplot(1, 3, 2)
ax2.hist(x = df.Comedy, color = "#E67F0D")
plt.title("Comedy")
plt.ylabel("Number of movies made")
plt.xlabel("Movies per year")

ax3 = fig.add_subplot(1, 3, 3)
ax3.hist(x = df.Action, color = "#FE4E00")
plt.title("Action")
plt.ylabel("Number of movies made")
plt.xlabel("Movies per year")
```

```
Text(0.5, 0, 'Movies per year')
```



The high number of movies made for each genre per year is not the norm. In the Drama and less than 20 movies with those genres made. As for Action, that number falls to 10.