In [134…

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

EXPLORANDO VARIABLES CON GRÁFICOS DE DISPERSIÓN

In [60]:

```
# Primero importamos el archivo. Cómo ha dado problemas para pasarlo a un DF, hago lo sigu
from io import open #  elimino las comillas para la lectura
#correcta.
fichero = open ( "tips.csv", "r", encoding= "utf-8")
texto= fichero.read()
texto=texto.replace ( '"','')
fichero.close()
```

In [58]:

```
fich = open ( "tips2.csv", "w", encoding= "utf-8")
fich.write(texto)# paso todo el archivo nuevo(sin comillas a tips2.csv
fich.close()
```

In [61]:

```
df= pd.read_csv("tips2.csv",sep=",")
df
```

Out[61]:

	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
...	...	...	...	...	...	...	...
239	29.03	5.92	Male	No	Sat	Dinner	3
240	27.18	2.00	Female	Yes	Sat	Dinner	2
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

244 rows × 7 columns

In [64]:

```
# Analizamos un poco la tabla, pero a simple vista se identifica todo, el coste de la com
# si es fumador o no, el sexo, las personas por cuenta, el día que se hace la comida.

df.shape# las filas son número de cuentas
```

Out[64]:

```
(244, 7)
```

In [66]:

```
df.columns
```

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

Out[66]:

In [67]:
```
# podemos ver que a excepcción de total_bill, tip y size, son variables categóricas.  A su
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   total_bill  244 non-null    float64
 1   tip         244 non-null    float64
 2   sex         244 non-null    object
 3   smoker      244 non-null    object
 4   day         244 non-null    object
 5   time        244 non-null    object
 6   size        244 non-null    int64
dtypes: float64(2), int64(1), object(4)
memory usage: 13.5+ KB
```

In [71]:
```
#Podemos ver que no hay valores nulos
df[df.duplicated(keep="last")]
```

Out[71]:

	total_bill	tip	sex	smoker	day	time	size
198	13.0	2.0	Female	Yes	Thur	Lunch	2

In [72]:
```
# bien podría ser una casualidad y no ser  un duplicado así que no lo elimino
```

In [116…
```
df.nunique()# nos informa de cuantos elementos distintos hay en cada columna
```

Out[116…
```
total_bill    229
tip           123
sex             2
smoker          2
day             4
time            2
size            6
dtype: int64
```

In [117…
```
df.isnull().sum()#observamos que no  hay valores nulos
```

Out[117…
```
total_bill    0
tip           0
sex           0
smoker        0
day           0
time          0
size          0
dtype: int64
```

In [118…
```
# miramos información estadística de las cuentas
df["total_bill"].describe()
```

Out[118…
```
count    244.000000
mean      19.785943
std        8.902412
min        3.070000
25%       13.347500
```

```
50%       17.795000
75%       24.127500
max       50.810000
Name: total_bill, dtype: float64
```

In [120…

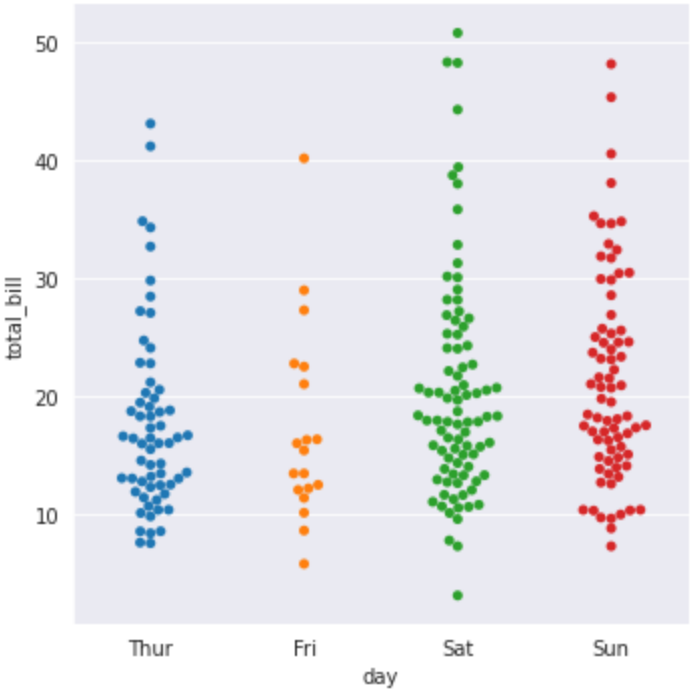
```
df["tip"].describe()# así como miramos info de las propinas
```

Out[120…

```
count    244.000000
mean       2.998279
std        1.383638
min        1.000000
25%        2.000000
50%        2.900000
75%        3.562500
max       10.000000
Name: tip, dtype: float64
```

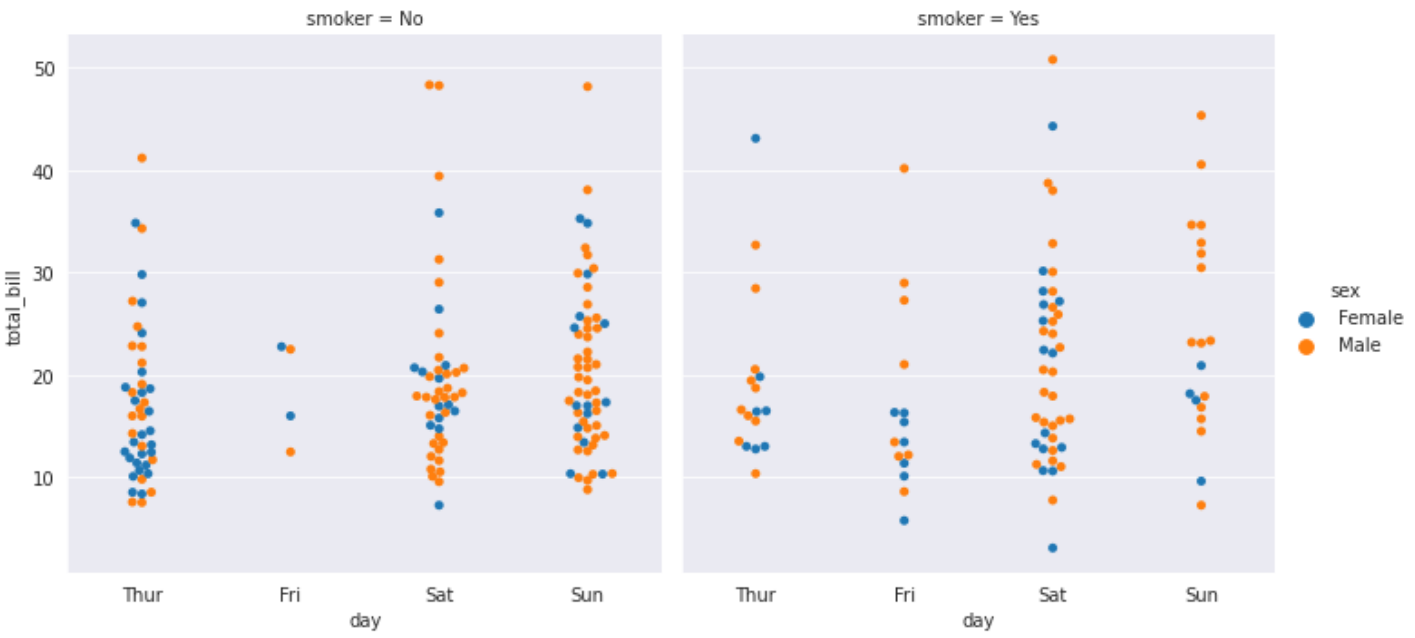
In [97]:

```
# Una vez mirada la informacón estadíastica vamos a ver que información nos dan los diagra
# Primero veremos los ingresos en función del día de la semana

with sns.axes_style("darkgrid"):
    sns.catplot( x= "day", y = "total_bill", data = df, kind= "swarm",order= ["Thur","Fri"
                "Sun"])
# 1. Vemos que hay una mayor concentración de comandas con precios de  10 y 20 unidades, a
# y más consumiciones el fin de semana
```



In [98]:

```
# Podemos diferenciarlos por sexos y por fumadores
# 2.Observamos que los Jueves y Domingos  hay muchos más no-fumadores que fumadores, mient


with sns.axes_style("darkgrid"):
    sns.catplot( x= "day", y = "total_bill", col= "smoker" , hue= "sex" , data = df, kind=
                "Sun"])
```



In [243…

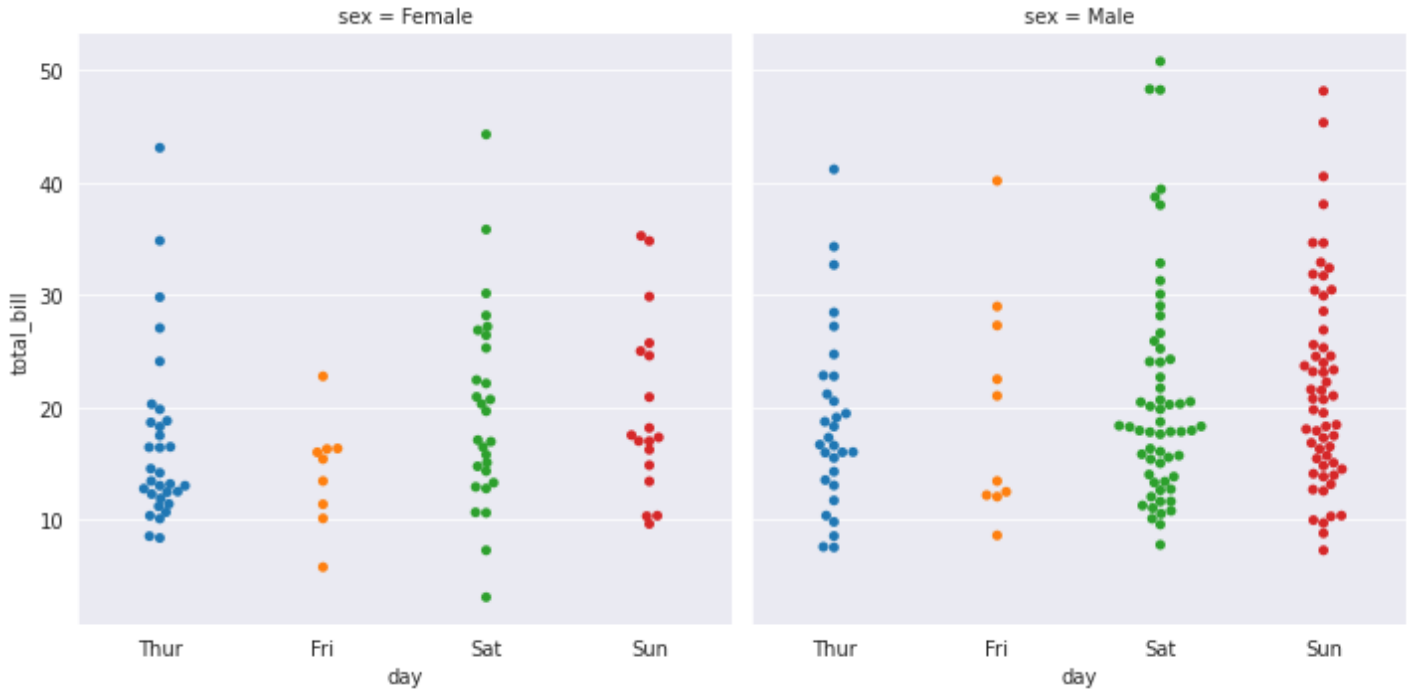
```
df["smoker"].value_counts()# número de fumadores
```

Out[243…

```
No     151
Yes     93
Name: smoker, dtype: int64
```

In [179…

```
# Si hacemos sólo la observación por sexo
with sns.axes_style("darkgrid"):
    sns.catplot( x= "day", y = "total_bill" , col= "sex" , data = df, kind= "swarm", order
                "Sun"])
# 3.Conclumimos que hay más presencia de hombres que de mujeres en el fin de semana.
```



In [223…

```
#miramos la proporcion de hombres  y mujeres.
df["sex"].value_counts()
# vemos que la cantidad de hombres es el doble.
```

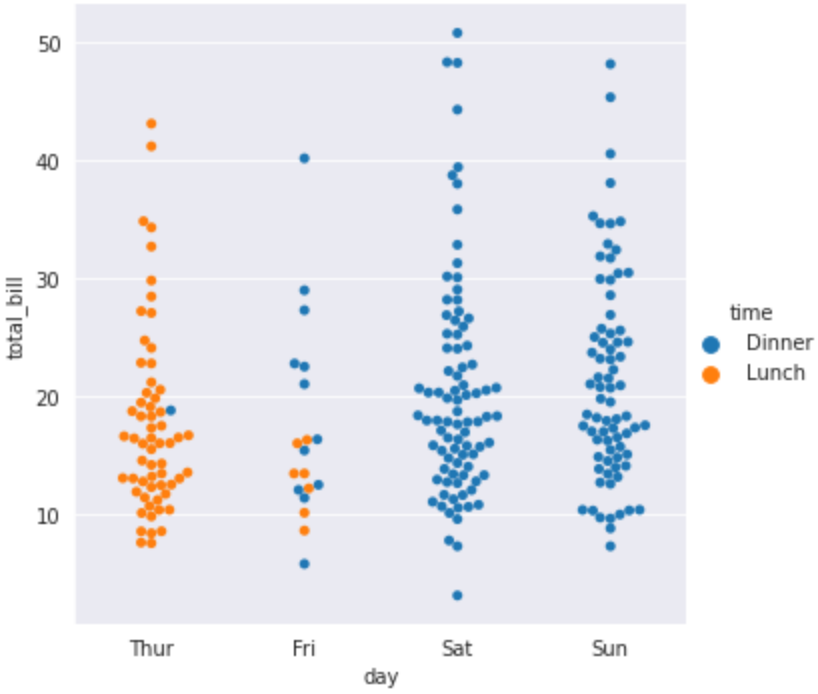
Out[223…

```
Male      157
Female     87
```

```
Name: sex, dtype: int64
```

In [101…

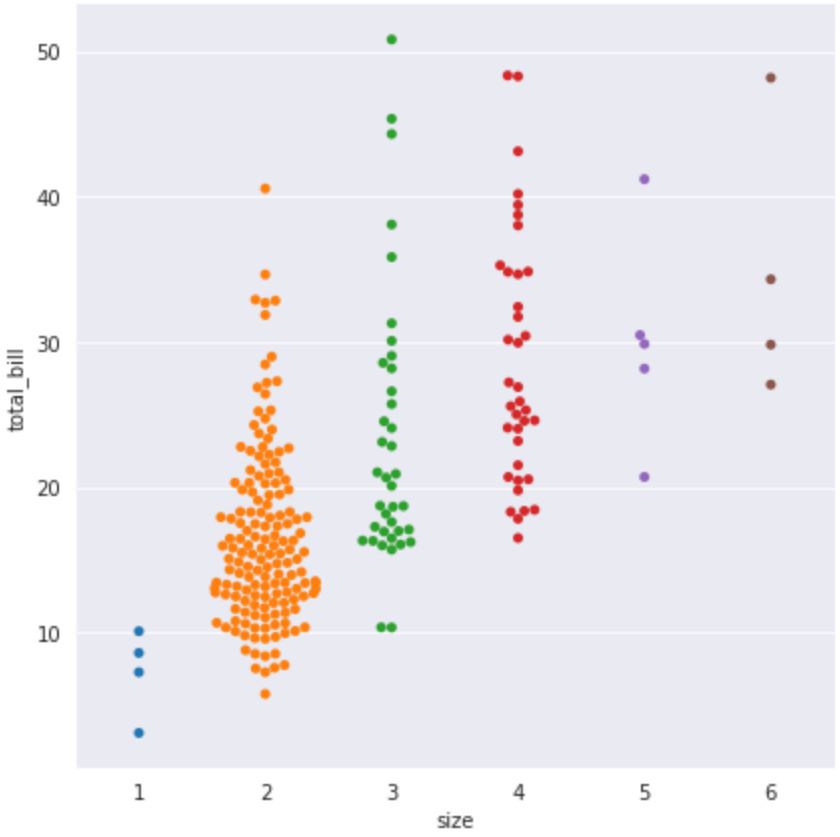
```
# Si miramos los ingresos de cuentas hechos en función de que comida del día ( "time"),
with sns.axes_style("darkgrid"):
    sns.catplot( x= "day", y = "total_bill" , hue= "time" , data = df, kind= "swarm", orde
                "Sun"])
# 4.se ve rápidamente que al llegar el fin de semana los ingresos son exclusivamente de ce
# el fin de semana no sirven comidas, o como segunda posibilidad, nadie consume  comida.
```



In [244…

```
# vamos a ver la relación entre el tamaño de la mesa y los ingresos.
with sns.axes_style("darkgrid"):
    sns.catplot( x= "size", y = "total_bill" , data = df, kind= "swarm", order= [1,2,3,4,5
# 5.Observamos que casi todas las cuentas por debajo de 25 son de 2 personas por mesa.
# la tendencia es a más comensales por mesa, más gasto en la cuenta, la mayoria de meseas
#pertenecen a una mesa de tamaño de 2, minetras que las cuentas entre 25 y 35 se distribuy
# entre mesas de tamaño 2, 3 y 4
```

```
C:\Users\walte\anaconda3\lib\site-packages\seaborn\categorical.py:3750: UserWarning: The `
size` parameter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```

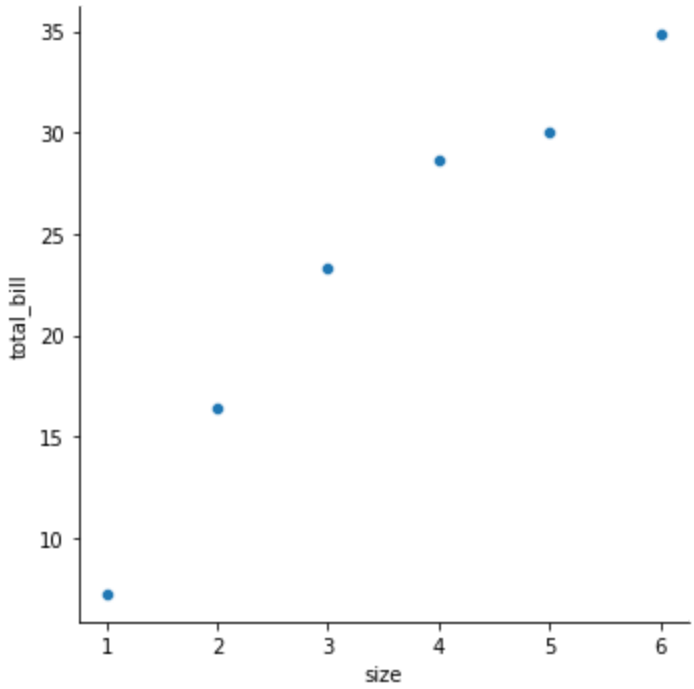


In [176…

```
# Vamos a indagar todavía más en esto, vamos a calcular la media de cuenta por mesa en fu
df2 = df[["size", "total_bill"]].groupby(["size"]).mean()
sns.relplot( data = df2, x= "size", y = "total_bill")
# 6 Vemos que  la media de gasto por mesa, tiende a una relación lineal entre el número de
```

Out[176…

```
<seaborn.axisgrid.FacetGrid at 0x1aa9f5887f0>
```

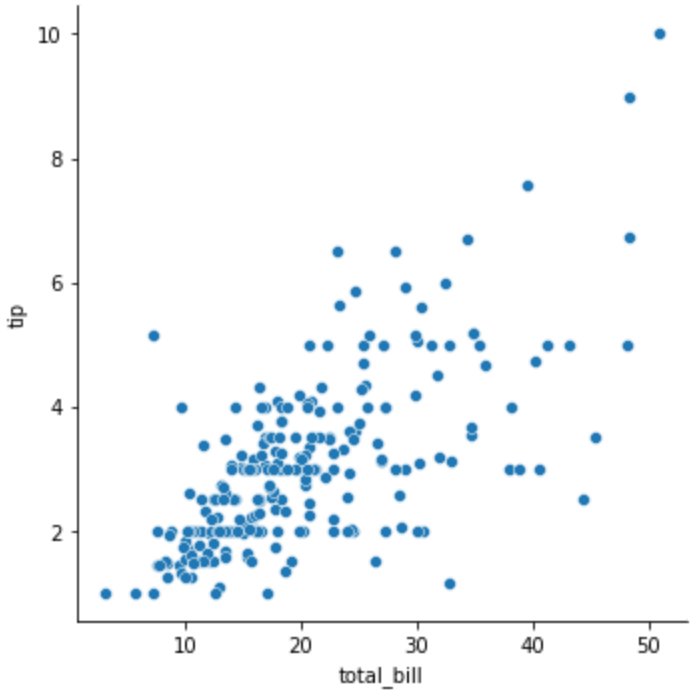


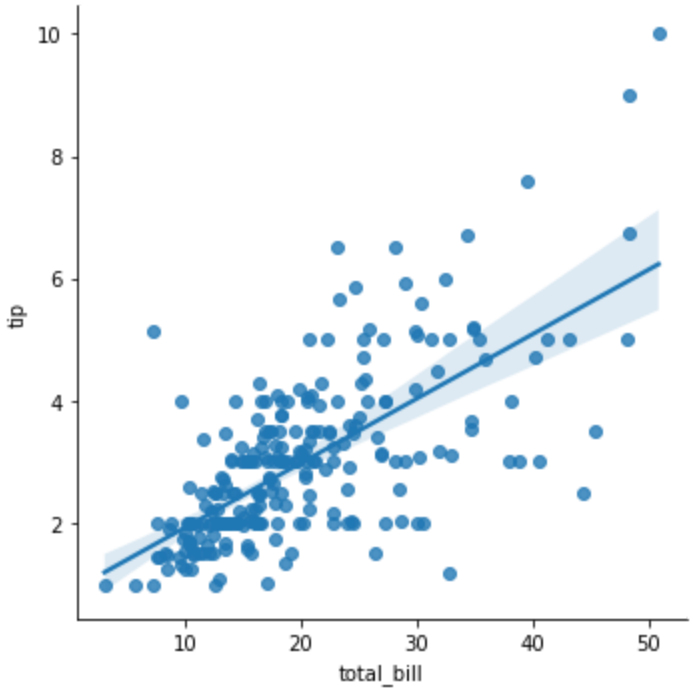
In [245…

```
# para analizar las propinas, vamos a ver que relación tienen en función de la cuentas. Po
# podremos deducir que lo que se cumple para "total_bill" se cumplirá para "tip"
sns.relplot(  data = df,  x = "total_bill", y = "tip")
sns.lmplot(  data = df,  x = "total_bill", y = "tip", robust= True)
# Al ver el gráfico vemos tiende a una relación lineal, hay demasiada dispersión respecto
# Así que analizamos las propinas aparte.
```

Out[245…

```
<seaborn.axisgrid.FacetGrid at 0x1aaaadc7040>
```

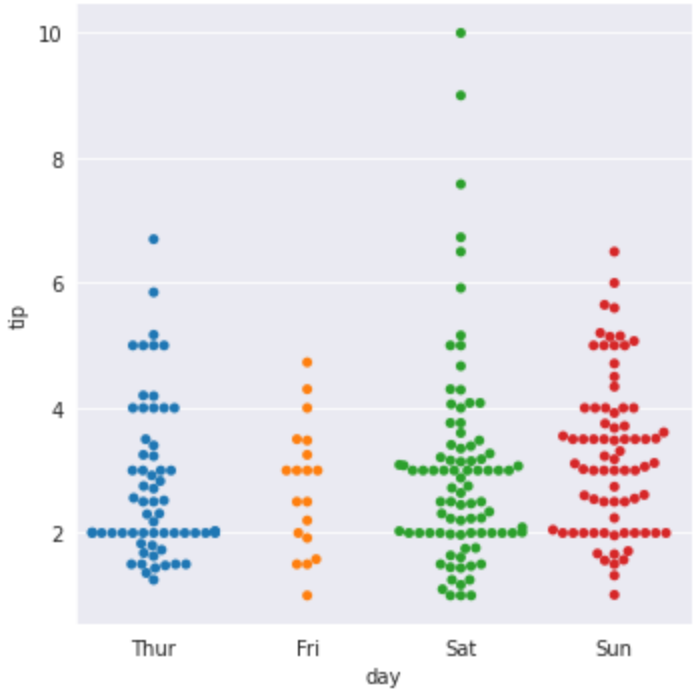




In [165…

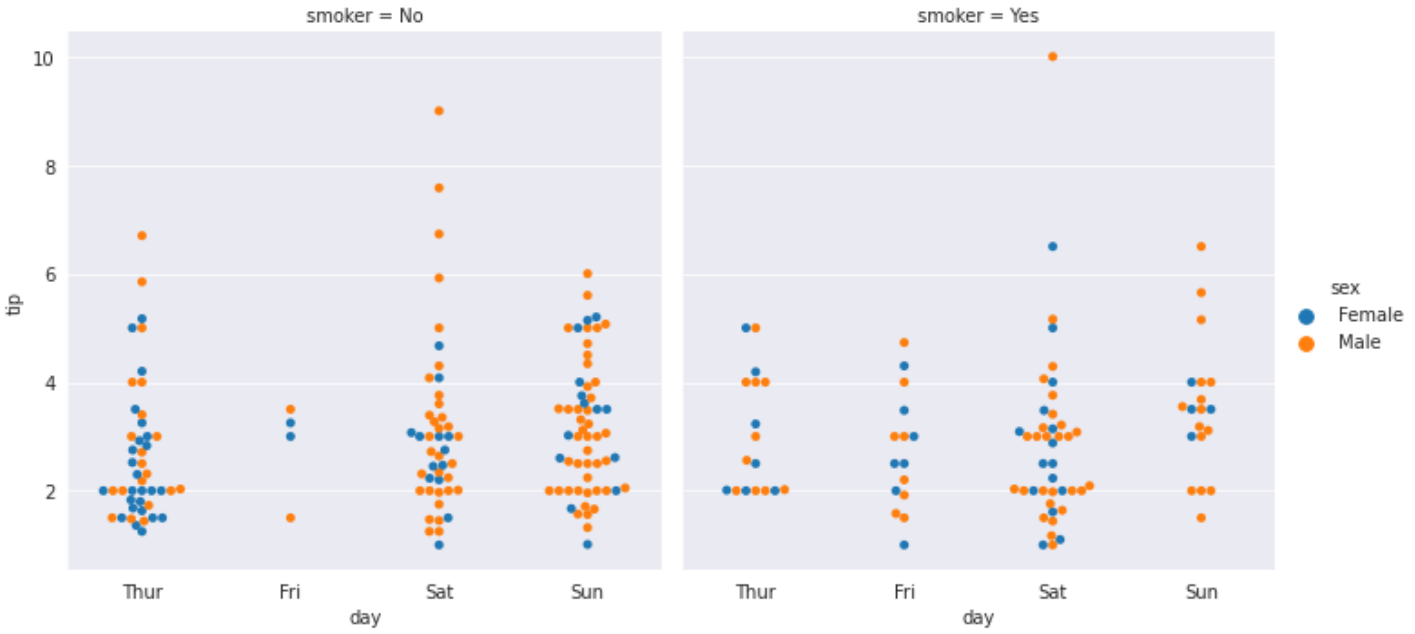
```
# Analizamos las propinas en función de los días de las semanas

with sns.axes_style("darkgrid"):
    sns.catplot( x= "day", y = "tip", data = df, kind= "swarm",order= ["Thur","Fri", "Sat"
                "Sun"])

# Tras hacer el gráfico se observa una relación parecida a al total de la cuenta
```

```
C:\Users\walte\anaconda3\lib\site-packages\seaborn\categorical.py:1296: UserWarning: 8.1%
of the points cannot be placed; you may want to decrease the size of the markers or use st
ripplot.
  warnings.warn(msg, UserWarning)
```



In [168…

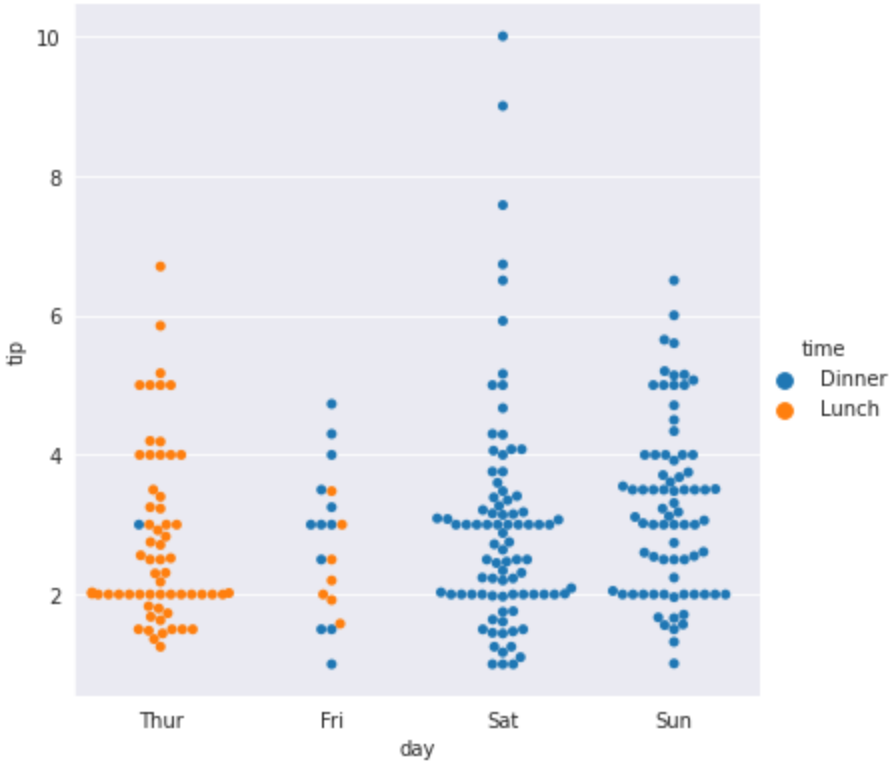
```
# si diferenciamos por fumador y sexo, vemos unas relaciones parecidas  al gasto de cuenta
with sns.axes_style("darkgrid"):
    sns.catplot( x= "day", y = "tip" , col= "smoker" , hue= "sex" , data = df, kind= "swar
                "Sun"])
```



In [177…

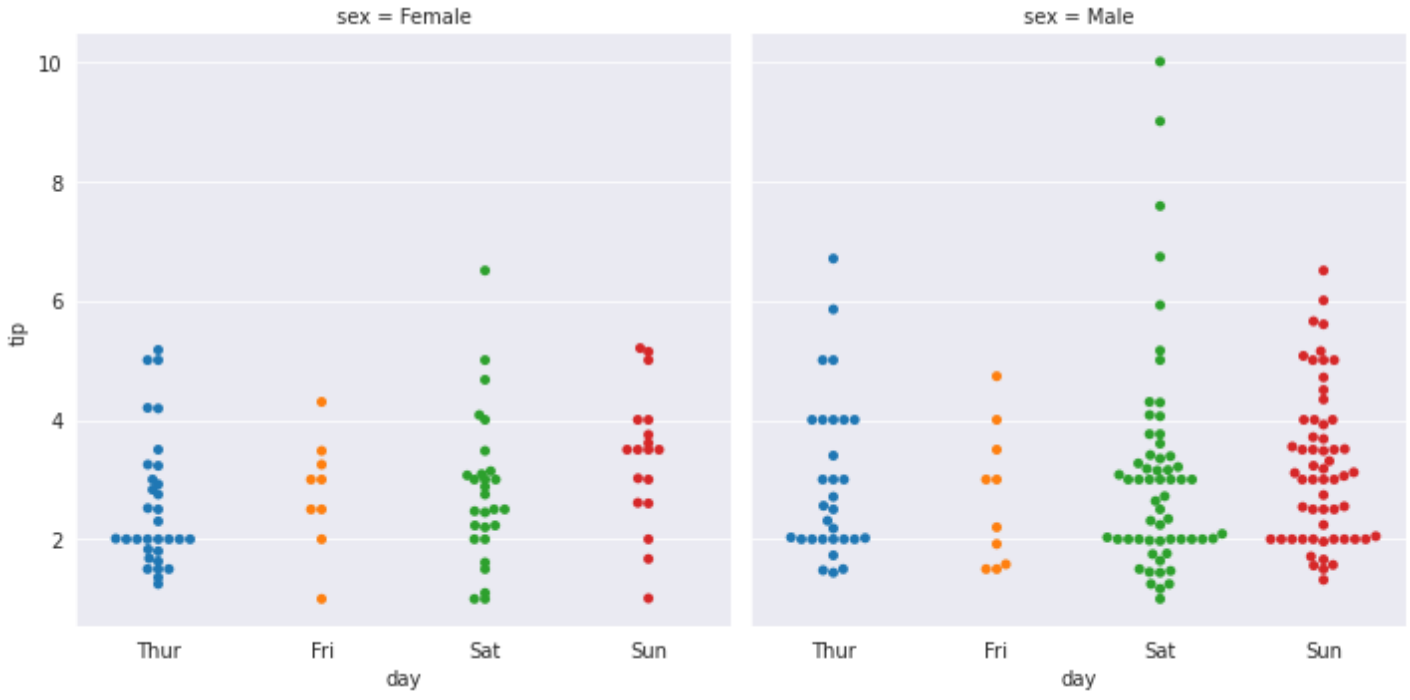
```
# si diferenciamos por si es comida o cena, se sigue manteniendo el parecido con Total_bi
with sns.axes_style("darkgrid"):
    sns.catplot( x= "day", y = "tip" , hue= "time" , data = df, kind= "swarm", size= 5.5,
                "Sun"])
```

```
C:\Users\walte\anaconda3\lib\site-packages\seaborn\categorical.py:3750: UserWarning: The `
size` parameter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```



In [178…

```
with sns.axes_style("darkgrid"):
    sns.catplot( x= "day", y = "tip" , col= "sex" , data = df, kind= "swarm", order= ["Thu
                "Sun"])

# si diferenciamos la relación por sexos, vemos que se mantiene la proporcion en el gasto
```

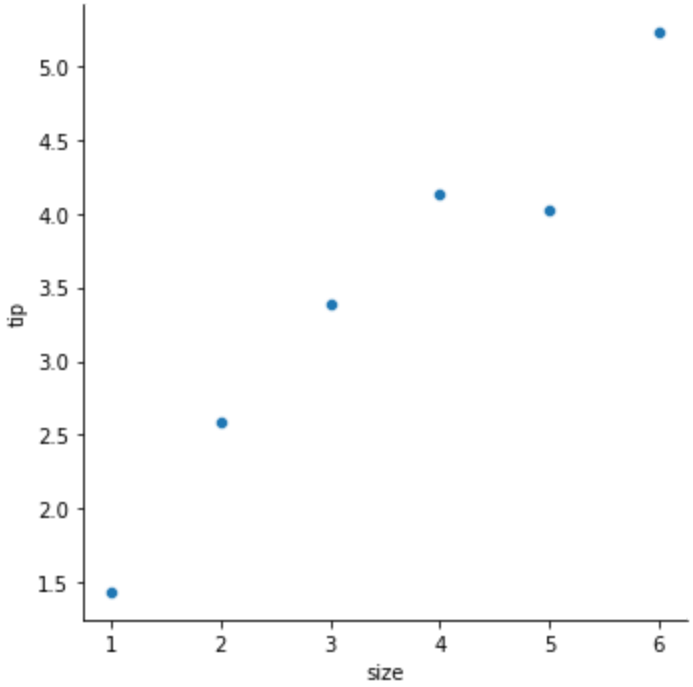


In [175…

```
# por último miramos, la media de propina en función dle tamaño de la mesa.
df3 = df[["size", "tip"]].groupby(["size"]).mean()
sns.relplot( data = df3, x= "size", y = "tip")
# no sólo oibservamos un relación casi lineal, si no que tiene un cierto parecido al hecho
```

Out[175…

```
<seaborn.axisgrid.FacetGrid at 0x1aa9f806130>
```



1. Conclusiones de las visyualizaciones gráficas:

- El gasto en propinas es proporcinal al gasto de la cuenta.
- El gasto de cada mesa(tanto en propinas como en la cuenta) es proporcional al tamaño de la mesa
- El fin de semanas sólo se sirven cenas, y hay más consumiciones el fin de semana que el jueves o el Viernes.
- La mayoría de consumiciones están entre 10 y 20 .
- hay más presencia de hombres que mujeres en las cuentas el fin de semana
- Hay más presencia de no fumadores los Jueves y Domingos, mientras que hay más presencia de fumadores el Viernes

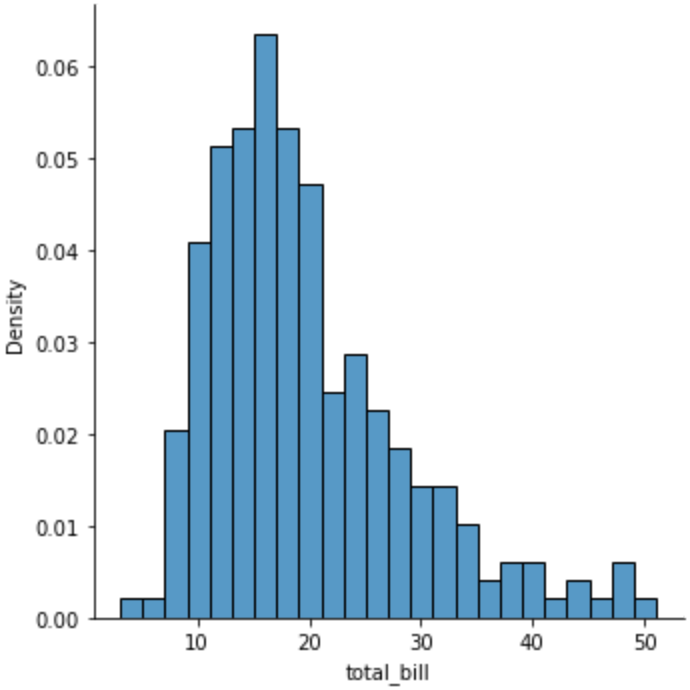
EXPLORANDO VARIABLES CON HISTOGRAMAS

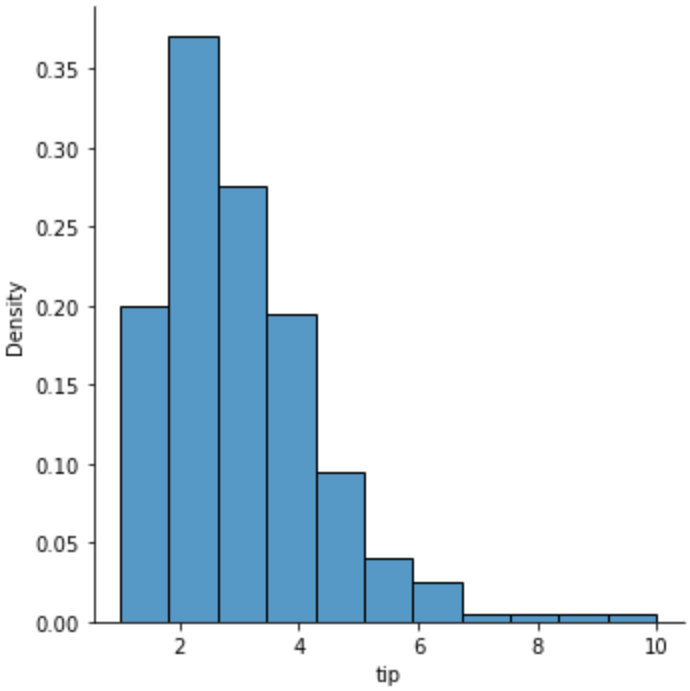
In [204…

```
# Los histogramas nos ayudarán a confirmar algunas conclusiones sacadas en el apartado enu
# Primero vemos el gráfico de las cuenta y propinas
sns.displot( df, x= "total_bill", binwidth= 2, stat="density")
sns.displot( df, x= "tip", bins=11, stat="density")
```

Out[204…

```
<seaborn.axisgrid.FacetGrid at 0x1aaa2642fd0>
```





In [229…

```
# miramos información estadística de las cuentas, cómo hemos hecho antes.
df["total_bill"].describe()
#Podemos ver que la media se encuentra en 20 con una Desvación estandar de caso 9
```

Out[229…

```
count    244.000000
mean      19.785943
std        8.902412
min        3.070000
25%       13.347500
50%       17.795000
75%       24.127500
max       50.810000
Name: total_bill, dtype: float64
```

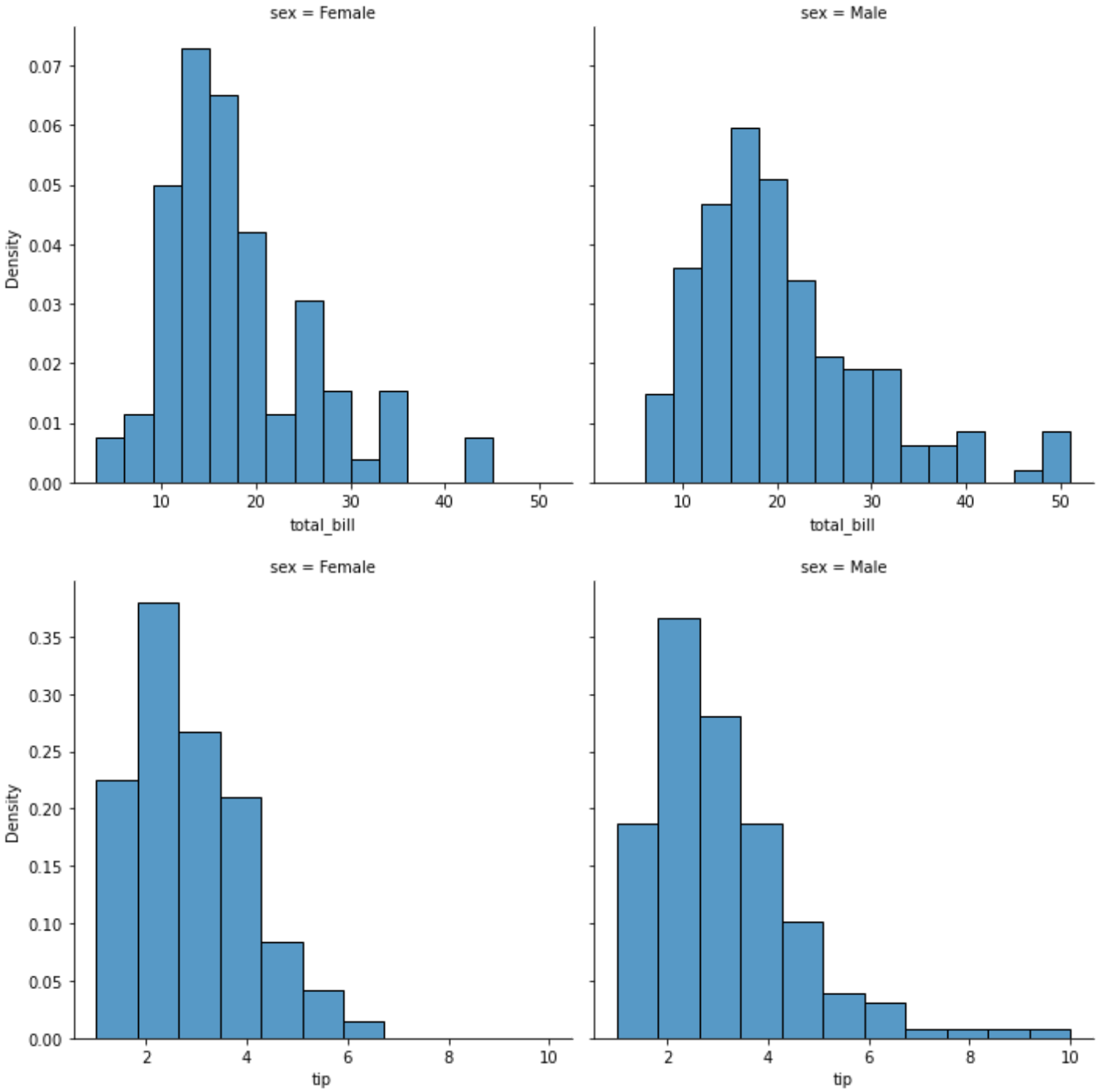
- Podemos ver que aproximadamente el 60% de las cuentas son entre 10 y 20 euros, y un 20% gastará entre 20 y 30.
- Mientas que de propinas, un 85% paga entre 2 y 4

In [213…

```
# Si observamos, la relación entre el gasto, por sexos.
sns.displot( df, x= "total_bill", binwidth= 3, col = "sex", multiple="dodge", stat="densit
sns.displot( df, x= "tip", bins=11, col = "sex", multiple="dodge",stat="density",common_nc
```

Out[213…

```
<seaborn.axisgrid.FacetGrid at 0x1aa9f5a8670>
```



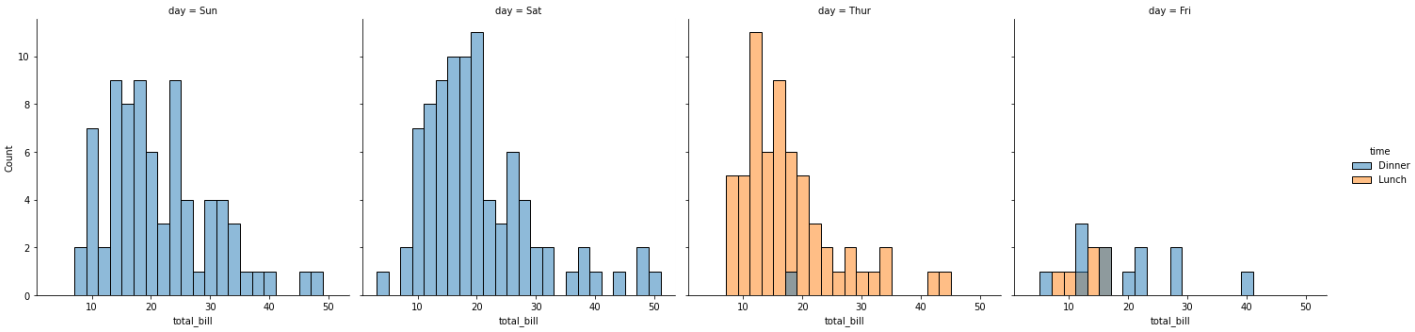
- El gasto en propinas por sexos, es muy similar, mientras que el gasto de cuenta, los hombres gastan un poco más.

In [218…

```
#Si miramos los gastos por días y comida
sns.displot( df, x= "total_bill", binwidth= 2, col = "day", hue= "time" )
```

Out[218…

```
<seaborn.axisgrid.FacetGrid at 0x1aaa4f96550>
```



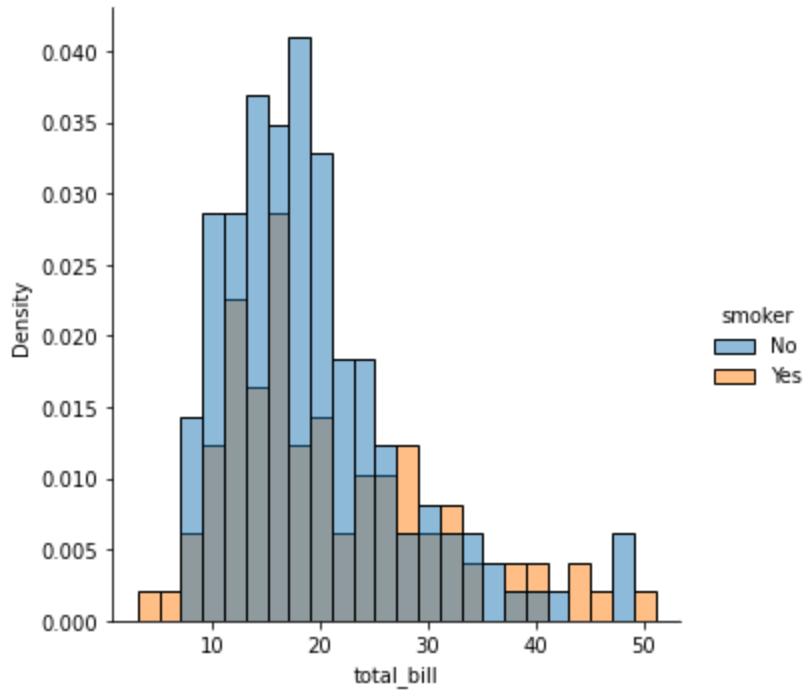
- La mayoría de ingresos se obtienen en la cena del sábado así como en la comida del Jueves.

In [227…

```
# Si miramos el gasto entre fumadores y no fumadores,
sns.displot( df, x= "total_bill", binwidth= 2, hue = "smoker", stat= "density")
```

Out[227…

```
<seaborn.axisgrid.FacetGrid at 0x1aaa39c2070>
```



- Se conluye que los no fumadores, que són mas cómo vimos antes, gastan más. Aunque en las cuentas de 38 para arriba,

hay más fumadores.

- Conclusiones:

1. Podemos ver que aproximadamente el 60% de las cuentas son entre 10 y 20 euros, y un 20% gastará entre 20 y 30.
2. Mientas que de propinas, un 85% paga entre 2 y 4
3. El gasto en propinas por sexos, es muy similar, mientras que el gasto de cuenta, los hombres gastan un poco más.
4. La mayoría de ingresos se obtienen en la cena del sábado así como en la comida del Jueves.
5. Se conluye que los no fumadores, que són mas cómo vimos antes, gastan más

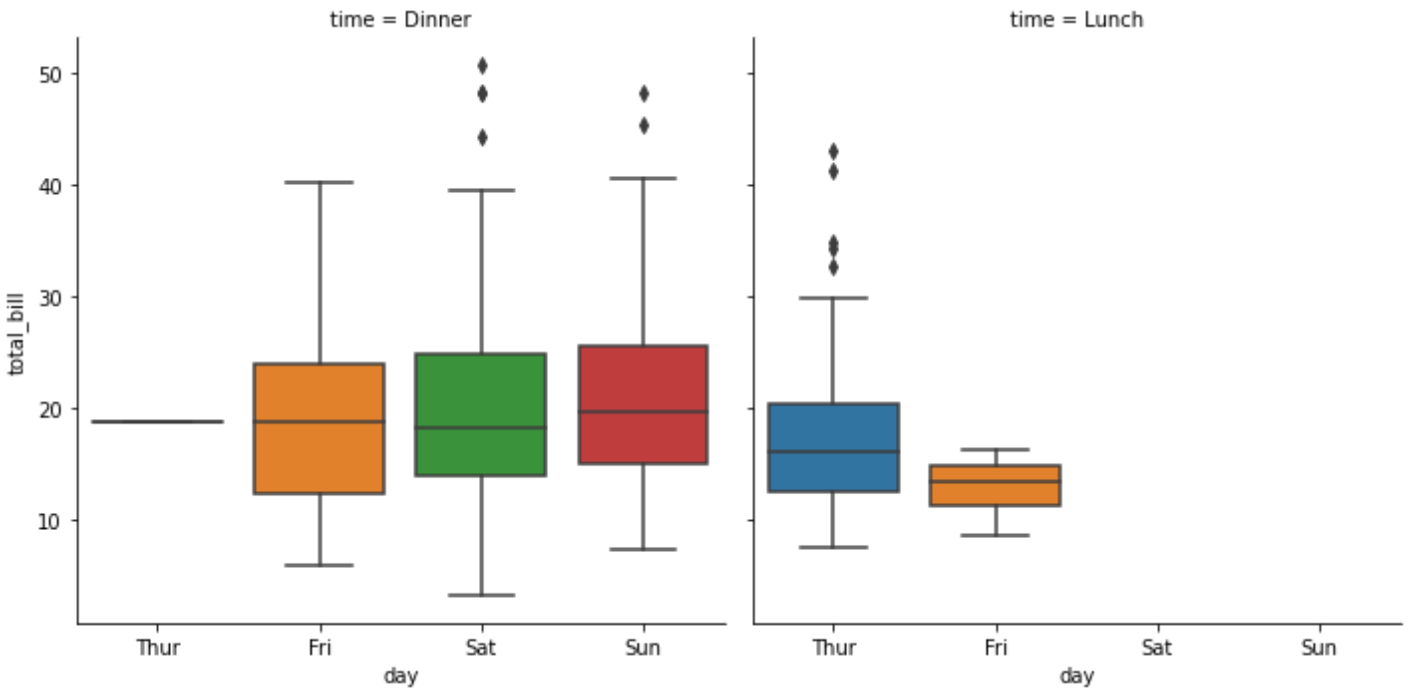
6. la cuenta "total_bill" es unimodal, sesgada a la derecha

Boxplot.

- Estos últimos nos pueden ser muy útiles para observar la media de gasto para variables categóricas.

In [238…
```
# empezamos por los días.
sns.catplot(x= "day", y = "total_bill", data = df, kind= "box", col="time",order= ["Thur",
```

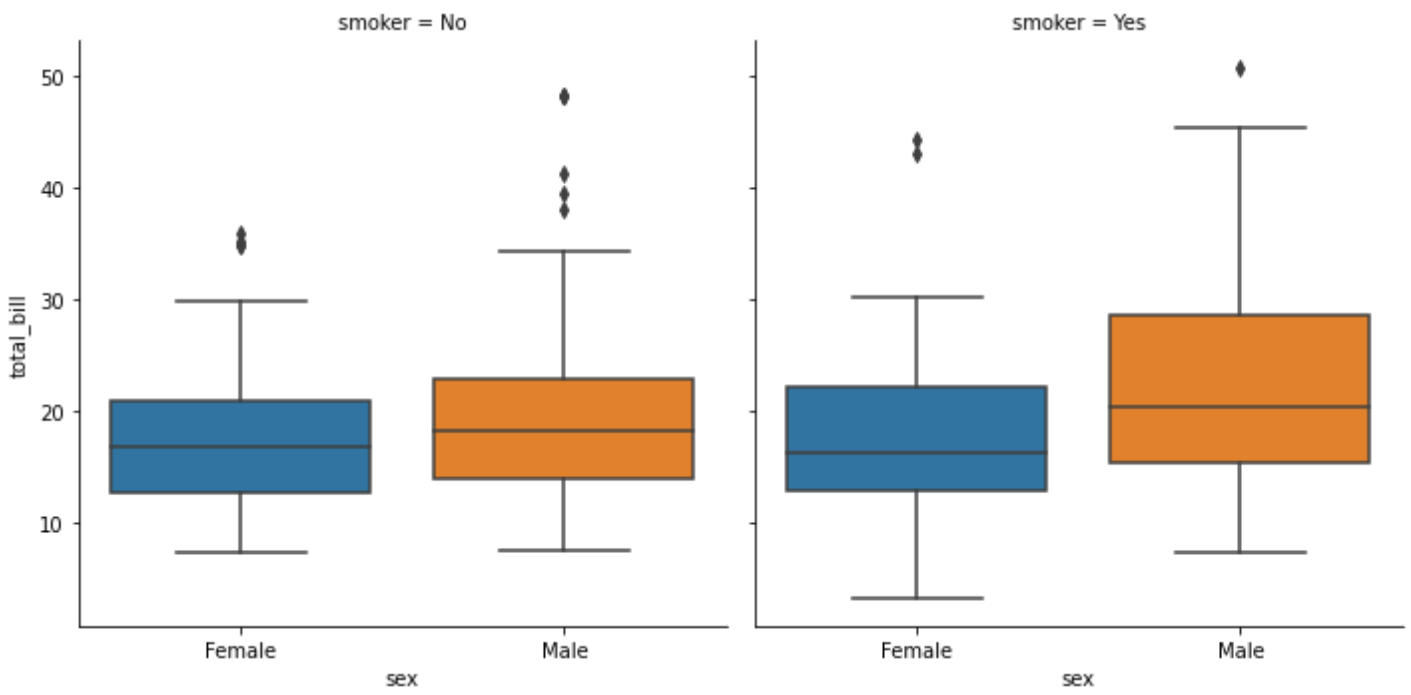
Out[238…
```
<seaborn.axisgrid.FacetGrid at 0x1aaa97e81f0>
```



- Confirmamos los dicho anteriormente, pero con algun matiz sobre el Jueves-
- La media de las cenas se encuentra en 20, teniendo un gasto medio ligeramente superior el Domingo.
- A pesar de que hubierna más cuentas los jueves, el gasto es menor,

In [241…
```
# Si miramos por sexos, y fumadores
sns.catplot(x= "sex", y = "total_bill",col = "smoker", data = df, kind= "box")
```

Out[241…
```
<seaborn.axisgrid.FacetGrid at 0x1aaa3d0edf0>
```



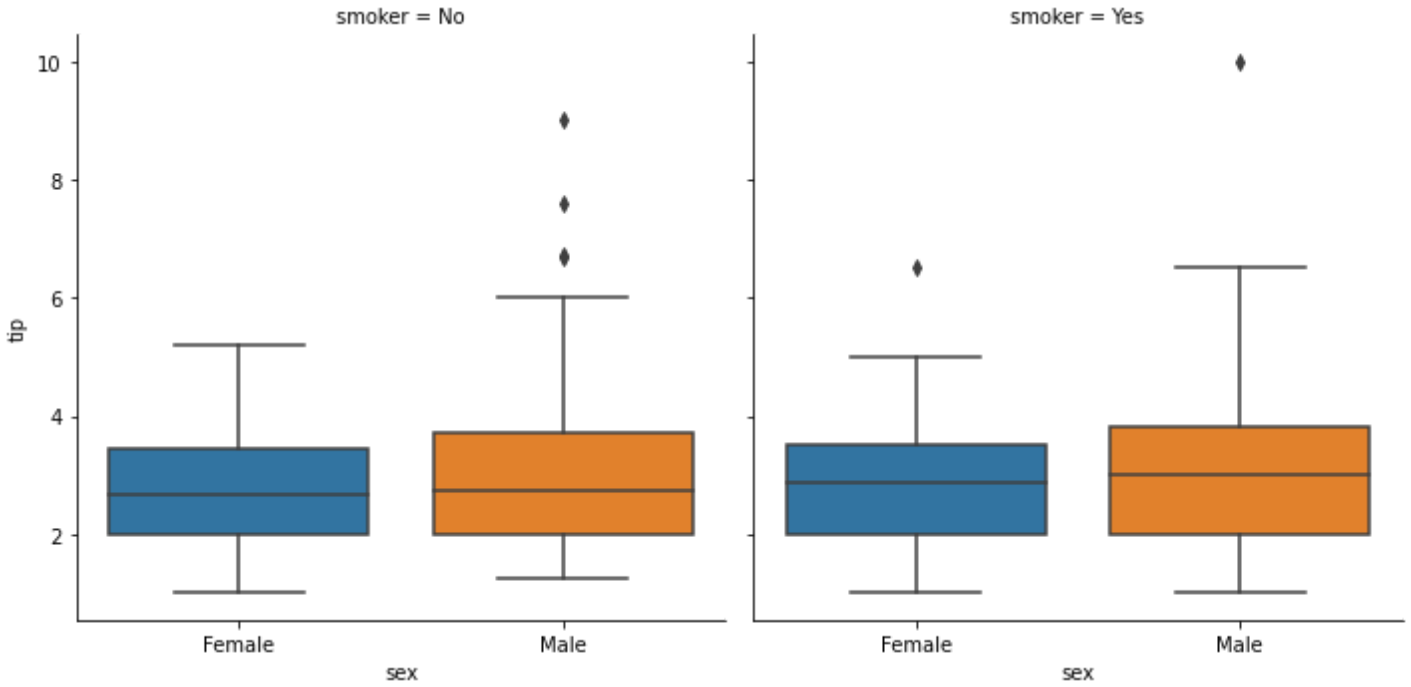
- Vemos que el gasto medio de los hombres es mayor. en el caso de no fumadores es parecido por sexos, pero en el caso hombres fumadores, el gasto es significativamente superior

In [246…

```
sns.catplot(x= "sex", y = "tip",col = "smoker", data = df, kind= "box")
# observamos que mantiene la linealidad con las propinas
```

Out[246…

```
<seaborn.axisgrid.FacetGrid at 0x1aaaad7bc70>
```

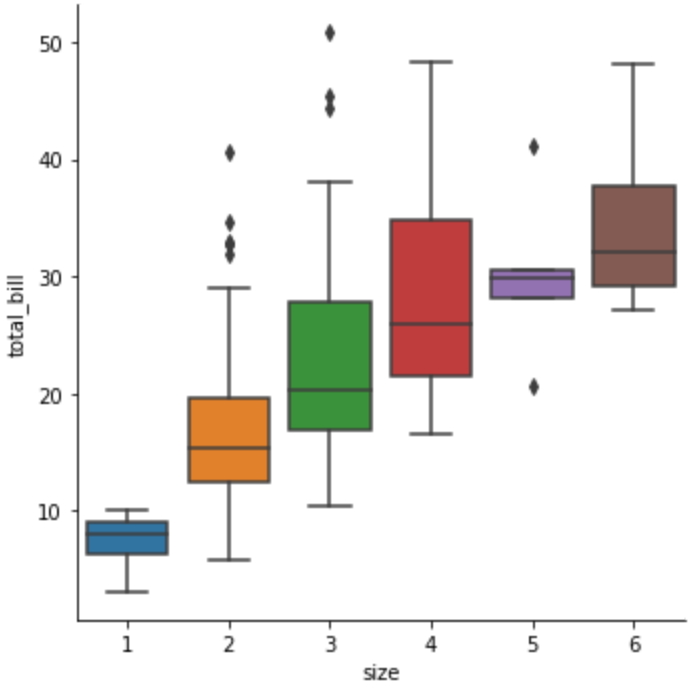


In [248…

```
# y por último miramos el gasto por el tamaño de mesa

sns.catplot(x= "size", y = "total_bill", data = df, kind= "box")

# vemos que mantiene la linealidad por mesa.
```

Out[248…

```
<seaborn.axisgrid.FacetGrid at 0x1aaabfbeaf0>
```



- Conclusiones

    1. La media de las cenas se encuentra en 20, teniendo un gasto medio ligeramente superior el Domingo, el gasto medio el jueves es menor, pero tiene muchos clientes.

    2. Los hombres fumadores, gastan más que los mujeres o los hombres no fumadores.

In [ ]: