In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import csv
from IPython.display import display, HTML


import seaborn as sns
# %matplotlib qt
```

In [2]:

```
from IPython.display import display_html
from itertools import chain,cycle

def display_side_by_side(*args,titles=cycle([''])):
    html_str=''
    for df,title in zip(args, chain(titles,cycle(['</br>'])) ):
        html_str+='<th style="text-align:center"><td style="vertical-
        html_str+=f'<h2 style="text-align: center;">{title}</h2>'
        html_str+=df.to_html().replace('table','table style="display:
        html_str+='</td></th>'
    display_html(html_str,raw=True)
```

Pandas Dataframes

dataframes → manipulation → results/visualisation

Outline:

- Group-by recap
- look-up tables / relational databases
- apply functions to dataframe
- Plotting with seaborn

A problem from the aerospace industry

Our boss has asked us to calculate how much money each airline spent on aircraft parts last year. The data that we have available are:

- **fleet data**: what types of aircraft each airline has;
- **aircraft type to part number**: a look-up table that indicates which part number fits to which aircraft type
- **cost of each part**: a look-up table that indicates how much each part costs to buy

In [3]:

```
fl = pd.read_csv("airlines_2.csv")
fc = fl.copy()
fl
```

Out[3]:

	airline	ac_type	variant	number
0	Lufthansa	Boeing	737-100	4
1	Lufthansa	Boeing	737-100	3
2	Lufthansa	Boeing	737-100	1
3	Lufthansa	Boeing	737-200	5
4	Lufthansa	Airbus	A380	3
5	Lufthansa	Airbus	A380	6
6	KLM	Airbus	A380	1
7	KLM	Airbus	A380	3
8	KLM	Airbus	A320	3
9	KLM	Airbus	A320	4
10	KLM	Airbus	A320	2
11	Air France	Airbus	A380	2
12	Air France	Airbus	A380	3
13	Air France	Boeing	747	4

In [4]:

```
ac_pn = pd.read_csv("ac_pn.csv")
ac_pn
```

Out[4]:

	variant	pn
0	747	PN-1
1	737-100	PN-2
2	737-200	PN-3
3	A320	PN-4
4	A380	PN-5

In [5]:

```
pn_cst = pd.read_csv("pn_cost.csv")
pn_cst
```

Out[5]:

	pn	cost
0	PN-1	2174
1	PN-2	3925
2	PN-3	1529
3	PN-4	4926
4	PN-5	987

In [6]:
```
display_side_by_side(fl,ac_pn, pn_cst, titles = ['fleet_data', 'look
```

fleet_data

	airline	ac_type	variant	number
0	Lufthansa	Boeing	737-100	4
1	Lufthansa	Boeing	737-100	3
2	Lufthansa	Boeing	737-100	1
3	Lufthansa	Boeing	737-200	5
4	Lufthansa	Airbus	A380	3
5	Lufthansa	Airbus	A380	6
6	KLM	Airbus	A380	1
7	KLM	Airbus	A380	3
8	KLM	Airbus	A320	3
9	KLM	Airbus	A320	4
10	KLM	Airbus	A320	2
11	Air France	Airbus	A380	2
12	Air France	Airbus	A380	3
13	Air France	Boeing	747	4

look-up table 1

	variant	pn
0	747	PN-1
1	737-100	PN-2
2	737-200	PN-3
3	A320	PN-4
4	A380	PN-5

look-up table 2

	pn	cost
0	PN-1	2174
1	PN-2	3925
2	PN-3	1529
3	PN-4	4926
4	PN-5	987

Grouping-by

- slow way to do it:

In [7]:
```
airln = "Lufthansa"

lh = fl.loc[fl.airline == airln] # locating entries
lh
```

Out[7]:

	airline	ac_type	variant	number
0	Lufthansa	Boeing	737-100	4
1	Lufthansa	Boeing	737-100	3
2	Lufthansa	Boeing	737-100	1
3	Lufthansa	Boeing	737-200	5
4	Lufthansa	Airbus	A380	3
5	Lufthansa	Airbus	A380	6

In [8]:
```
lh['variant'].to_list()
```

Out[8]:
```
['737-100', '737-100', '737-100', '737-200', 'A380', 'A380']
```

In [9]:
```
set(lh['variant'].to_list())
```

Out[9]:
```
{'737-100', '737-200', 'A380'}
```

In [10]:
```
lh_vars = list(set(lh['variant'].to_list()))
lh_vars
```

Out[10]:
```
['737-100', '737-200', 'A380']
```

In [11]:
```
var = '737-100'
lh.loc[lh.variant == var]
```

Out[11]:

	airline	ac_type	variant	number
0	Lufthansa	Boeing	737-100	4
1	Lufthansa	Boeing	737-100	3
2	Lufthansa	Boeing	737-100	1

In [12]:
```
lh.loc[lh.variant == var]['number'].sum()
```

Out[12]:
```
8
```

In [13]:
```
var = '737-200'
lh.loc[lh.variant == var]
```

Out[13]:

	airline	ac_type	variant	number
3	Lufthansa	Boeing	737-200	5

In [14]:
```
lh.loc[lh.variant == var]['number'].sum()
```

Out[14]:
```
5
```

- fast way to do it:

In [15]:
```
fl
```

Out[15]:

	airline	ac_type	variant	number
0	Lufthansa	Boeing	737-100	4
1	Lufthansa	Boeing	737-100	3
2	Lufthansa	Boeing	737-100	1
3	Lufthansa	Boeing	737-200	5
4	Lufthansa	Airbus	A380	3
5	Lufthansa	Airbus	A380	6
6	KLM	Airbus	A380	1
7	KLM	Airbus	A380	3
8	KLM	Airbus	A320	3
9	KLM	Airbus	A320	4
10	KLM	Airbus	A320	2
11	Air France	Airbus	A380	2
12	Air France	Airbus	A380	3
13	Air France	Boeing	747	4

In [16]:
```
fl_gr = fl.groupby(['airline', 'ac_type', 'variant']).sum(numeric_onl
fl_gr
```

Out[16]:

	airline	ac_type	variant	number
0	Air France	Airbus	A380	5
1	Air France	Boeing	747	4
2	KLM	Airbus	A320	9
3	KLM	Airbus	A380	4
4	Lufthansa	Airbus	A380	9
5	Lufthansa	Boeing	737-100	8
6	Lufthansa	Boeing	737-200	5

Look-up tables

- adding a PN column

In [17]:
```
display_side_by_side(fl_gr,ac_pn)
```

	airline	ac_type	variant	number
0	Air France	Airbus	A380	5
1	Air France	Boeing	747	4
2	KLM	Airbus	A320	9
3	KLM	Airbus	A380	4
4	Lufthansa	Airbus	A380	9
5	Lufthansa	Boeing	737-100	8
6	Lufthansa	Boeing	737-200	5

	variant	pn
0	747	PN-1
1	737-100	PN-2
2	737-200	PN-3
3	A320	PN-4
4	A380	PN-5

In [18]:
```
ac_lu = dict(zip(ac_pn.variant, ac_pn.pn))
ac_lu
```

Out[18]:
```
{'747': 'PN-1',
 '737-100': 'PN-2',
 '737-200': 'PN-3',
 'A320': 'PN-4',
 'A380': 'PN-5'}
```

In [19]:
```
variant = "A380"
fl_gr.loc[fl_gr.variant == variant, "PN"] = ac_lu[variant]
fl_gr
```

Out[19]:

	airline	ac_type	variant	number	PN
0	Air France	Airbus	A380	5	PN-5
1	Air France	Boeing	747	4	NaN
2	KLM	Airbus	A320	9	NaN
3	KLM	Airbus	A380	4	PN-5
4	Lufthansa	Airbus	A380	9	PN-5
5	Lufthansa	Boeing	737-100	8	NaN
6	Lufthansa	Boeing	737-200	5	NaN

- alternative: for-loop

In [20]:
```
for vrnt in ["737-100", "737-200", "A380", "A320", "747"]:
    fl_gr.loc[fl_gr.variant == vrnt, "PN"] = ac_lu[vrnt]
```

In [ ]:
```
```

In [21]:
```
for vrnt in ["737-100", "737-200", "A380", "A320", "747"]:
    fc.loc[fc.variant == vrnt, "PN"] = ac_lu[vrnt]
```

In [22]:
```
fl_gr
```

Out[22]:

	airline	ac_type	variant	number	PN
0	Air France	Airbus	A380	5	PN-5
1	Air France	Boeing	747	4	PN-1
2	KLM	Airbus	A320	9	PN-4
3	KLM	Airbus	A380	4	PN-5
4	Lufthansa	Airbus	A380	9	PN-5
5	Lufthansa	Boeing	737-100	8	PN-2
6	Lufthansa	Boeing	737-200	5	PN-3

- adding cost column

In [23]:
```
pn_lu = dict(zip(pn_cst.pn, pn_cst.cost))
pn_lu
```

Out[23]:
```
{'PN-1': 2174, 'PN-2': 3925, 'PN-3': 1529, 'PN-4': 4926, 'PN-5': 98
7}
```

In [24]:
```
pn = "PN-5"
fl_gr.loc[fl_gr.PN == pn, "cost/part"] = pn_lu[pn]
fl_gr
```

Out[24]:

	airline	ac_type	variant	number	PN	cost/part
0	Air France	Airbus	A380	5	PN-5	987.0
1	Air France	Boeing	747	4	PN-1	NaN
2	KLM	Airbus	A320	9	PN-4	NaN
3	KLM	Airbus	A380	4	PN-5	987.0
4	Lufthansa	Airbus	A380	9	PN-5	987.0
5	Lufthansa	Boeing	737-100	8	PN-2	NaN
6	Lufthansa	Boeing	737-200	5	PN-3	NaN

- alternative: for-loop

In [25]:
```
for pn in ["PN-1", "PN-2", "PN-3", "PN-4", "PN-5"]:
    fl_gr.loc[fl_gr.PN == pn, "cost/part"] = pn_lu[pn]
```

In [26]:
```
for pn in ["PN-1", "PN-2", "PN-3", "PN-4", "PN-5"]:
    fc.loc[fc.PN == pn, "cost/part"] = pn_lu[pn]
```

In [27]:
```
fl_gr
```

Out[27]:

	airline	ac_type	variant	number	PN	cost/part
0	Air France	Airbus	A380	5	PN-5	987.0
1	Air France	Boeing	747	4	PN-1	2174.0
2	KLM	Airbus	A320	9	PN-4	4926.0
3	KLM	Airbus	A380	4	PN-5	987.0
4	Lufthansa	Airbus	A380	9	PN-5	987.0
5	Lufthansa	Boeing	737-100	8	PN-2	3925.0
6	Lufthansa	Boeing	737-200	5	PN-3	1529.0

In [28]:
```
fl_gr['total cost'] = fl_gr['cost/part']*fl_gr['number']
fc['total cost'] = fc['cost/part']*fc['number']
```

In [ ]:

In [29]:
```
fc
```

Out[29]:

	airline	ac_type	variant	number	PN	cost/part	total cost
0	Lufthansa	Boeing	737-100	4	PN-2	3925.0	15700.0
1	Lufthansa	Boeing	737-100	3	PN-2	3925.0	11775.0
2	Lufthansa	Boeing	737-100	1	PN-2	3925.0	3925.0
3	Lufthansa	Boeing	737-200	5	PN-3	1529.0	7645.0
4	Lufthansa	Airbus	A380	3	PN-5	987.0	2961.0
5	Lufthansa	Airbus	A380	6	PN-5	987.0	5922.0
6	KLM	Airbus	A380	1	PN-5	987.0	987.0
7	KLM	Airbus	A380	3	PN-5	987.0	2961.0
8	KLM	Airbus	A320	3	PN-4	4926.0	14778.0
9	KLM	Airbus	A320	4	PN-4	4926.0	19704.0
10	KLM	Airbus	A320	2	PN-4	4926.0	9852.0
11	Air France	Airbus	A380	2	PN-5	987.0	1974.0
12	Air France	Airbus	A380	3	PN-5	987.0	2961.0
13	Air France	Boeing	747	4	PN-1	2174.0	8696.0

In [30]:
```
display_side_by_side(fl,ac_pn, pn_cst, titles = ['fleet_data', 'look
```

fleet_data

	airline	ac_type	variant	number
0	Lufthansa	Boeing	737-100	4
1	Lufthansa	Boeing	737-100	3
2	Lufthansa	Boeing	737-100	1
3	Lufthansa	Boeing	737-200	5
4	Lufthansa	Airbus	A380	3
5	Lufthansa	Airbus	A380	6
6	KLM	Airbus	A380	1
7	KLM	Airbus	A380	3
8	KLM	Airbus	A320	3
9	KLM	Airbus	A320	4
10	KLM	Airbus	A320	2
11	Air France	Airbus	A380	2
12	Air France	Airbus	A380	3
13	Air France	Boeing	747	4

look-up table 1

	variant	pn
0	747	PN-1
1	737-100	PN-2
2	737-200	PN-3
3	A320	PN-4
4	A380	PN-5

look-up table 2

	pn	cost
0	PN-1	2174
1	PN-2	3925
2	PN-3	1529
3	PN-4	4926
4	PN-5	987

In [31]:
```
fl_al = fl_gr.groupby('airline').sum(numeric_only = True).reset_inde
fl_al
```

Out[31]:

	airline	number	cost/part	total cost
0	Air France	9	3161.0	13631.0
1	KLM	13	5913.0	48282.0
2	Lufthansa	22	6441.0	47928.0

In [ ]:
```
fl_al.pop('cost/part')
```

In [50]:
```
fl_al
```

Out[50]:

	airline	number	total cost
0	Air France	9	13631.0
1	KLM	13	48282.0
2	Lufthansa	22	47928.0

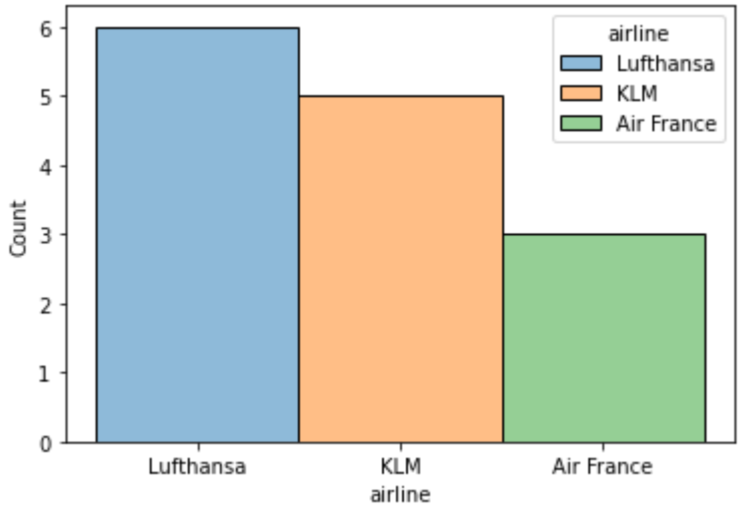
Seaborn plotting

- distplot (distributions)
- relplot (relational)
- catplot (categorical)

distplot (distributions)

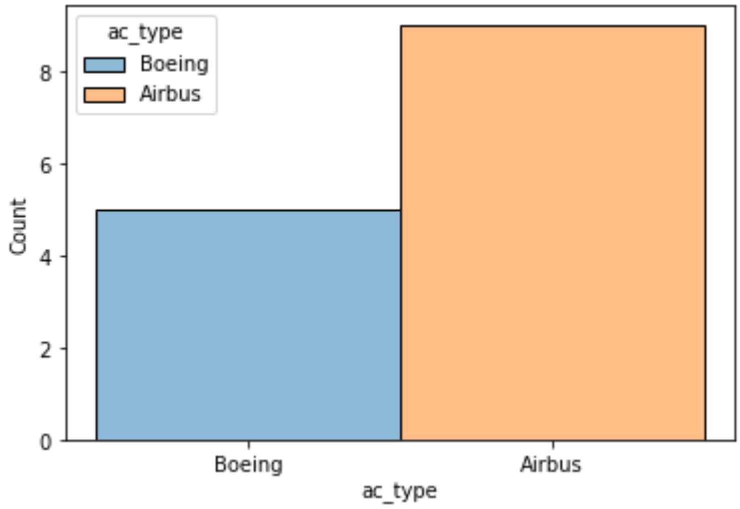
```
In [32]: sns.histplot(data=fc, x="airline", hue="airline")
```

```
Out[32]: <AxesSubplot:xlabel='airline', ylabel='Count'>
```



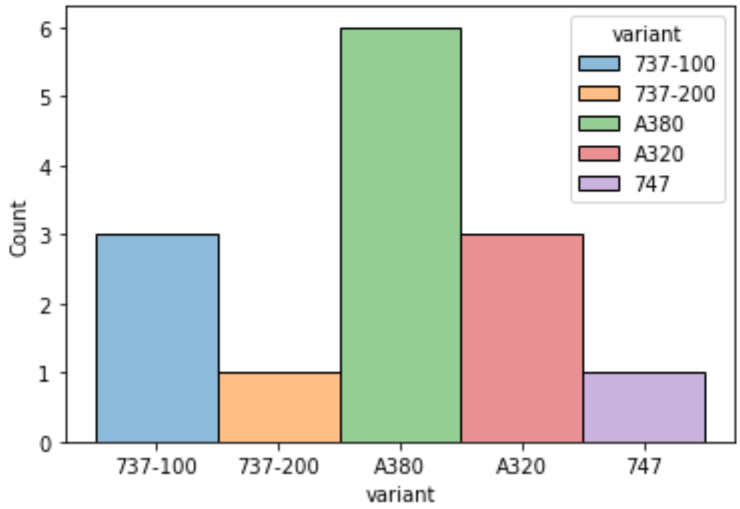
```
In [33]: sns.histplot(data=fc, x="ac_type", hue="ac_type")
```

```
Out[33]: <AxesSubplot:xlabel='ac_type', ylabel='Count'>
```



```
In [34]: sns.histplot(data=fc, x="variant", hue="variant")
```

```
Out[34]: <AxesSubplot:xlabel='variant', ylabel='Count'>
```



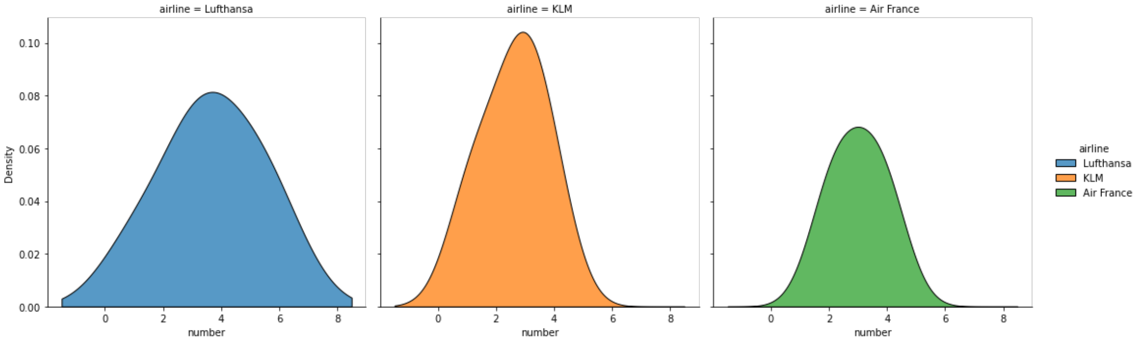
```
In [35]: fc
```

Out[35]:

	airline	ac_type	variant	number	PN	cost/part	total cost
0	Lufthansa	Boeing	737-100	4	PN-2	3925.0	15700.0
1	Lufthansa	Boeing	737-100	3	PN-2	3925.0	11775.0
2	Lufthansa	Boeing	737-100	1	PN-2	3925.0	3925.0
3	Lufthansa	Boeing	737-200	5	PN-3	1529.0	7645.0
4	Lufthansa	Airbus	A380	3	PN-5	987.0	2961.0
5	Lufthansa	Airbus	A380	6	PN-5	987.0	5922.0
6	KLM	Airbus	A380	1	PN-5	987.0	987.0
7	KLM	Airbus	A380	3	PN-5	987.0	2961.0
8	KLM	Airbus	A320	3	PN-4	4926.0	14778.0
9	KLM	Airbus	A320	4	PN-4	4926.0	19704.0
10	KLM	Airbus	A320	2	PN-4	4926.0	9852.0
11	Air France	Airbus	A380	2	PN-5	987.0	1974.0
12	Air France	Airbus	A380	3	PN-5	987.0	2961.0
13	Air France	Boeing	747	4	PN-1	2174.0	8696.0

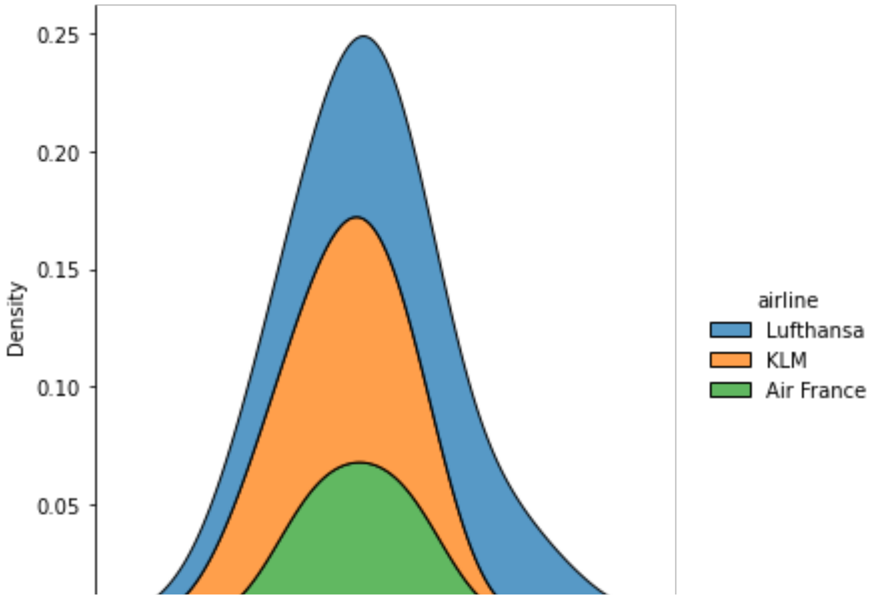
```
In [36]: sns.displot(data=fc, x="number", hue="airline", col="airline",  multi
```

```
Out[36]: <seaborn.axisgrid.FacetGrid at 0x7f5edb3bd2b0>
```



```
In [37]: sns.displot(data=fc, x="number", hue="airline", multiple="stack", kir
```

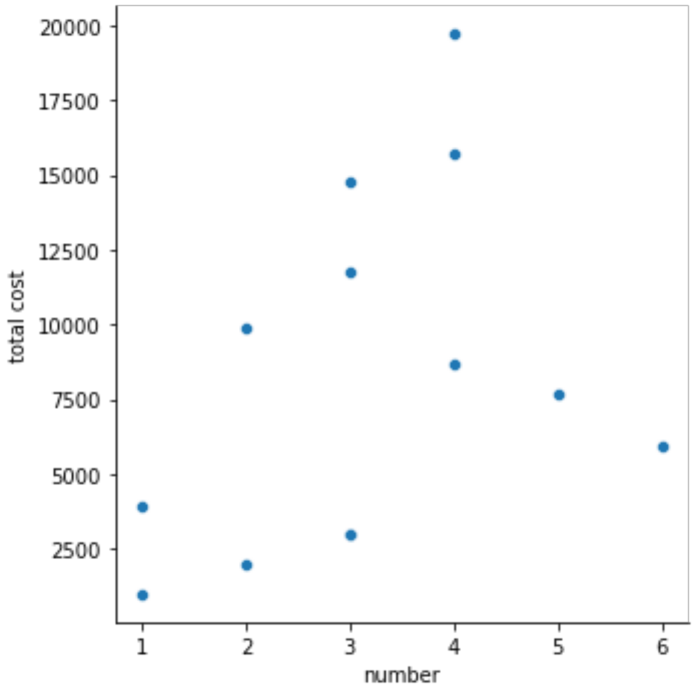
```
Out[37]: <seaborn.axisgrid.FacetGrid at 0x7f5edb42c670>
```



relplot (relational)

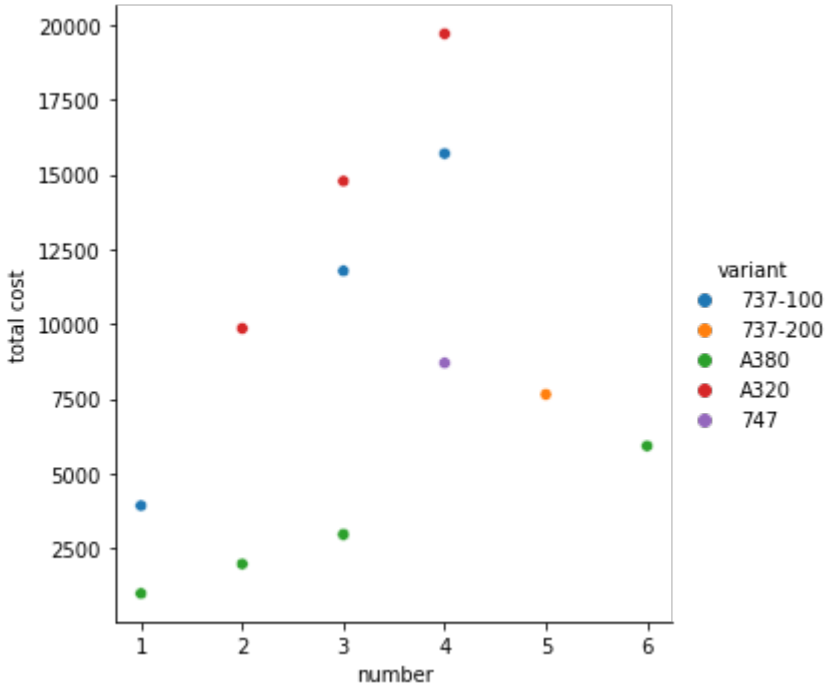
```
In [38]: sns.relplot(data = fc, x = "number", y = "total cost")

Out[38]: <seaborn.axisgrid.FacetGrid at 0x7f5edb0e9d60>
```



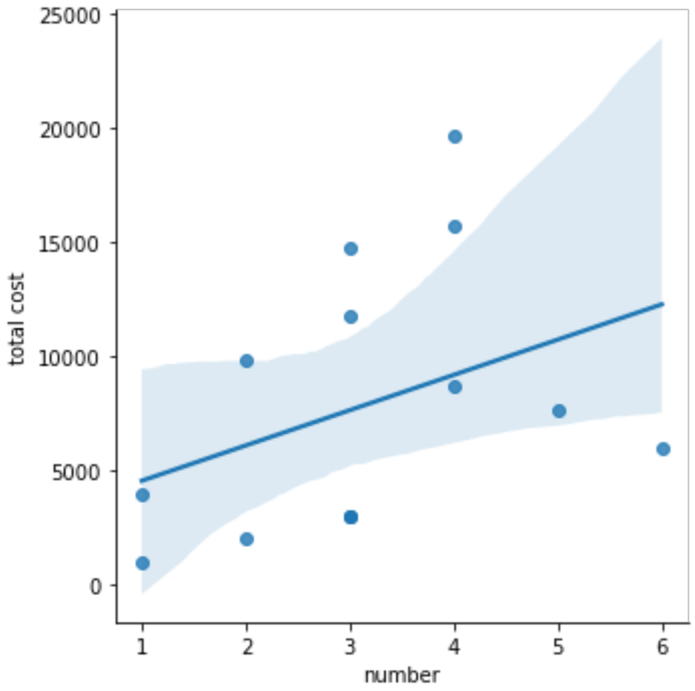
```
In [39]: sns.relplot(data = fc, x = "number", y = "total cost", hue = 'variant

Out[39]: <seaborn.axisgrid.FacetGrid at 0x7f5ed8f77eb0>
```



```
In [40]: sns.lmplot(data=fc, x="number", y="total cost")
```

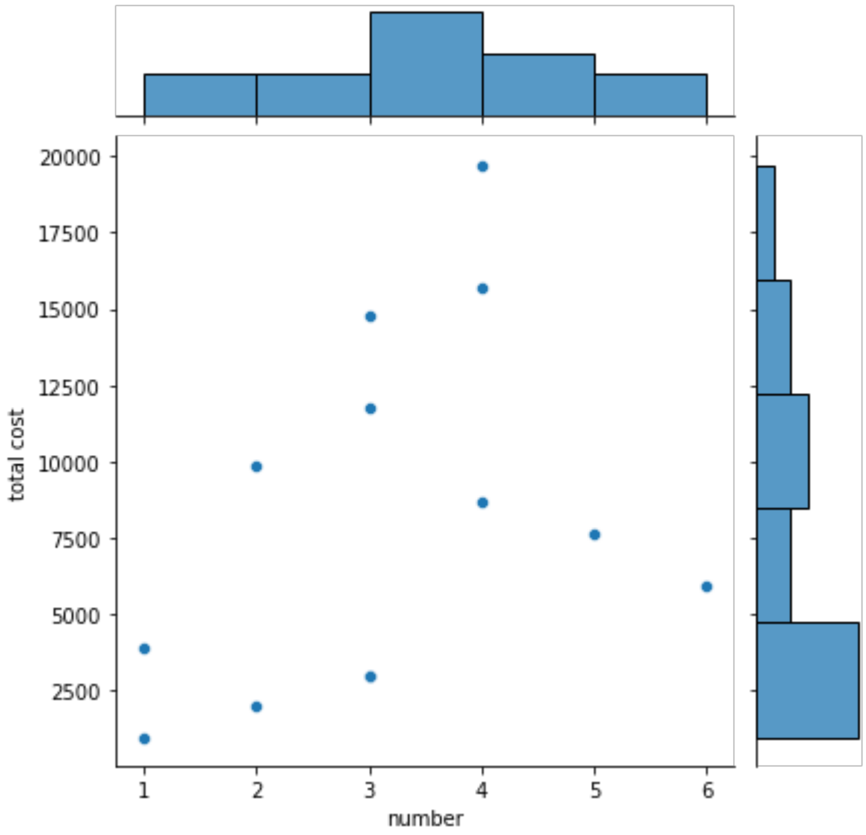
```
Out[40]: <seaborn.axisgrid.FacetGrid at 0x7f5edb465cd0>
```



relational + distributions

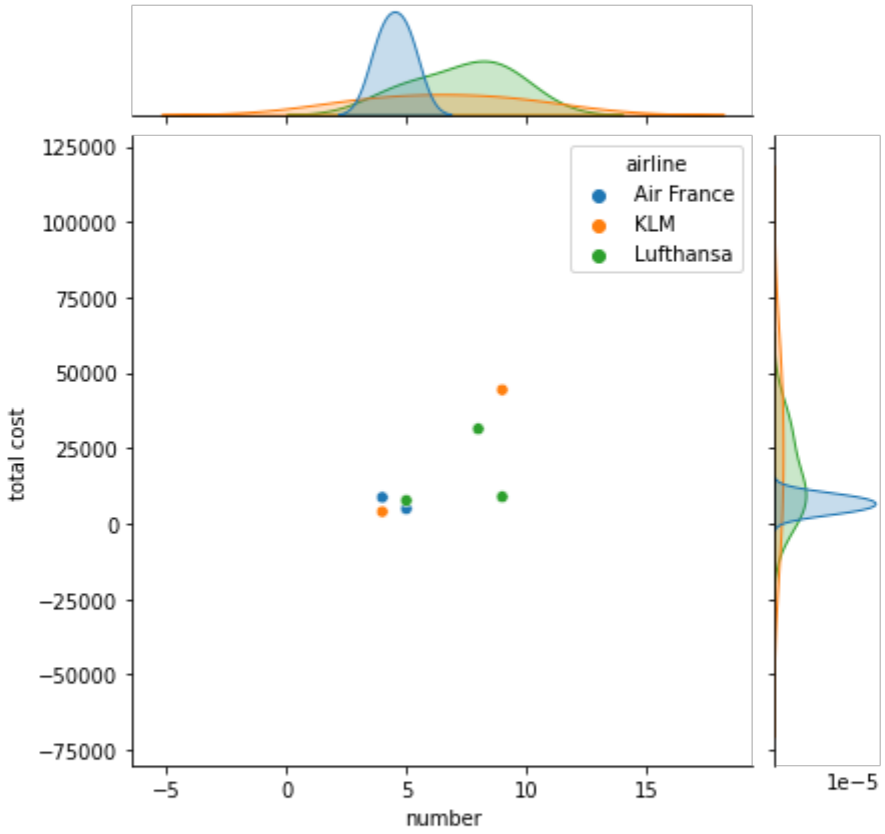
```
In [41]: sns.jointplot(data=fc, x="number", y="total cost") #, kind = "reg"
```

```
Out[41]: <seaborn.axisgrid.JointGrid at 0x7f5ed8cdf5e0>
```



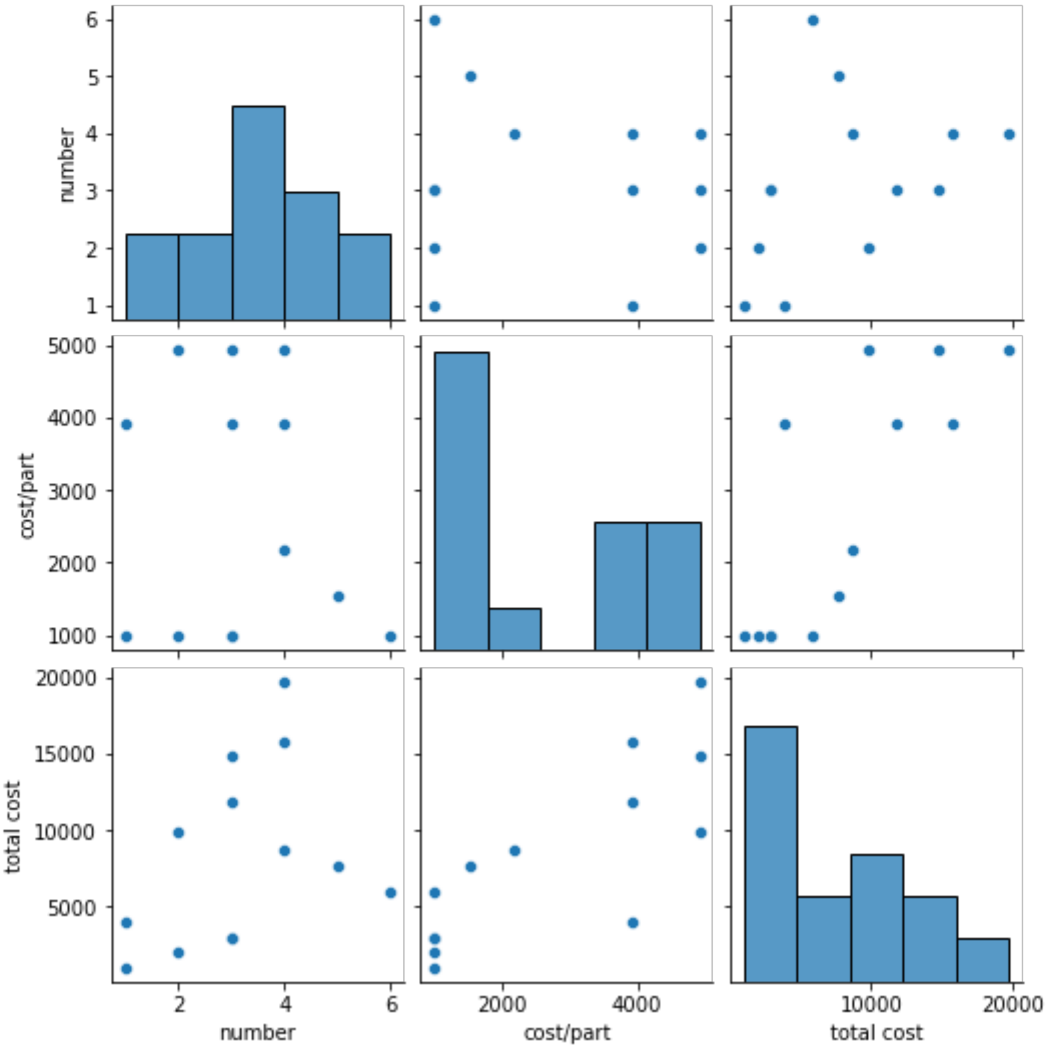
```
In [42]: sns.jointplot(data=fl_gr, x="number", y="total cost", hue="airline")
```

```
Out[42]: <seaborn.axisgrid.JointGrid at 0x7f5ed8c00070>
```



```
In [45]: sns.pairplot(data=fc) #,  kind = "reg", hue="airline"
```

```
Out[45]: <seaborn.axisgrid.PairGrid at 0x7f5ed81d6be0>
```



```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```