COMP7035

Python for Data Analytics and Artificial Intelligence

Seaborn

Renjie Wan

29/10/2024

香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

Numpy Array Padding

- **numpy.pad(*array*, *pad_width*, *mode='constant'*, ***kwargs*)**
- This function is used to pad an array.

```
import numpy as np
a = [[1, 2], [3, 4]]
print(a)
a_pad = np.pad(a, ((2, 3), (3, 3)), 'constant')
print(a_pad)
```

Pad three columns of zero along the horizontal direction, before the second axis

Pad three columns of zero along the horizontal direction, after the second axis

Pad two rows of zero along the vertical direction, before the first axis

```
[[0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 1 2 0 0 0]
 [0 0 0 3 4 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]]
```

Pad three rows of zero along the vertical direction, after the first axis

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

A small exercise for you

1. Create a matrix like the left

```
[[0 0 0 0 0 0 0 0]
 [0 0 0 1 2 0 0 0]
 [0 0 0 3 4 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]]
```

A small exercise for you

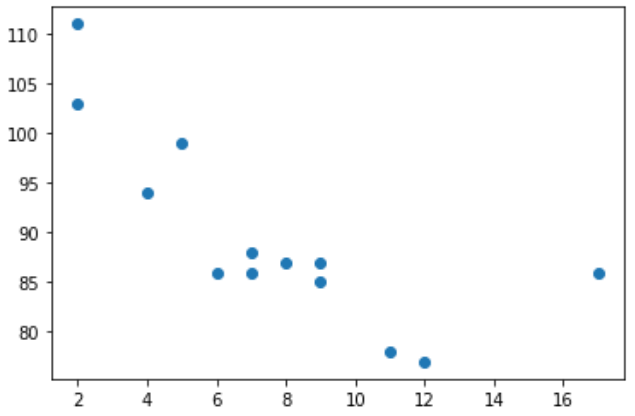
```
[[0 0 0 0 0 0 0 0]
 [0 0 0 1 2 0 0 0]
 [0 0 0 3 4 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 0]]
```

Create a matrix like the left

PyPlot: Scatter

- The scatter() function plots one dot for each observation. It needs two arrays of the same length, one for the values of the x-axis, and one for values on the y-axis:

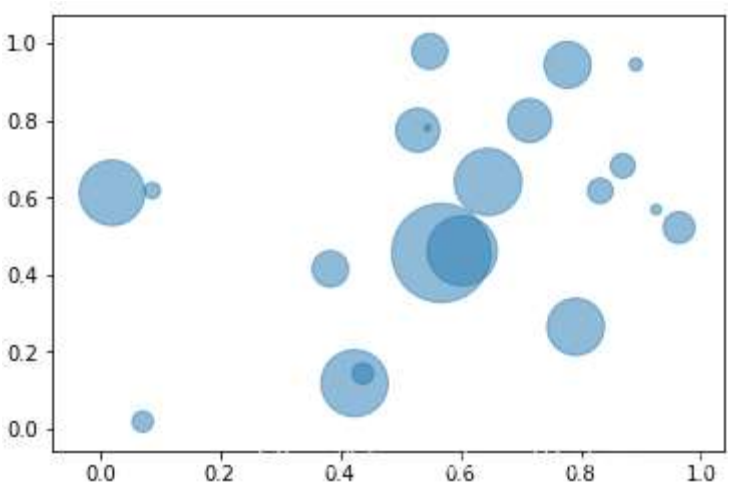
```
import matplotlib.pyplot as plt
import numpy as np

x = np.array([5,7,8,7,2,17,2,9,4,11,12,9,6])
y = np.array([99,86,87,88,111,86,103,87,94,78,77,85,86])

plt.scatter(x, y)
plt.show()
```



PyPlot: Scatter

- We can also control the areas and their colors of each dot
- In this page, area is a random number array

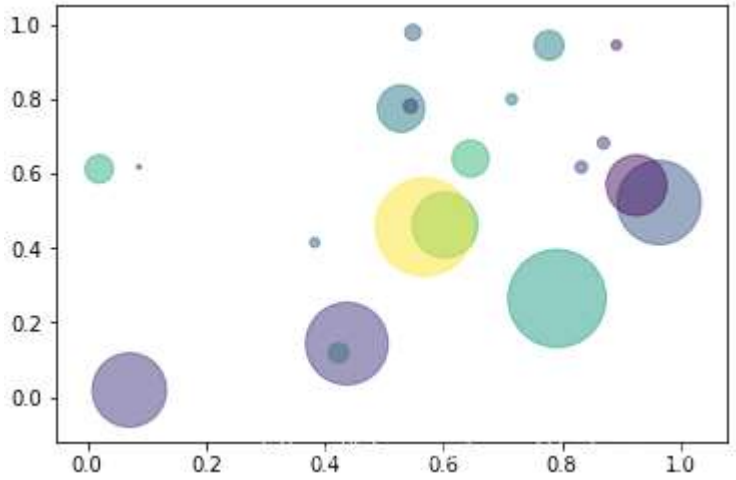
```
import numpy as np
import matplotlib.pyplot as plt

np.random.seed(0)
x=np.random.rand(20)
y=np.random.rand(20)

area=(50*np.random.rand(20))**2

plt.scatter(x,y,s=area,alpha=0.5)
plt.show()
```



PyPlot: Scatter

- In this page, color also becomes a random number array.

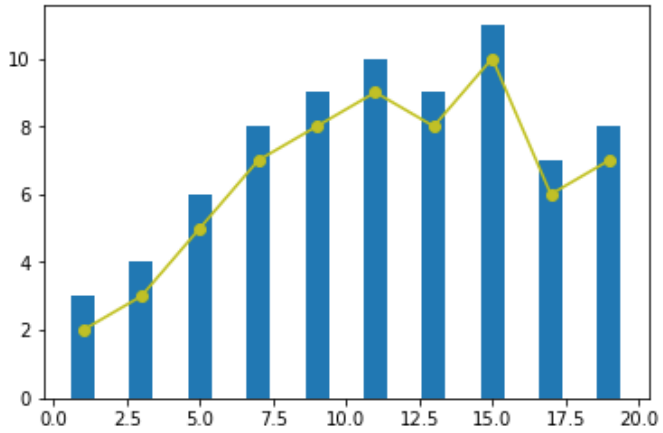
```
import numpy as np
import matplotlib.pyplot as plt

np.random.seed(0)
x=np.random.rand(20)
y=np.random.rand(20)

colors=np.random.rand(20)
area=(50*np.random.rand(20))**2

plt.scatter(x,y,s=area,c=colors,alpha=0.5)
plt.show()
```



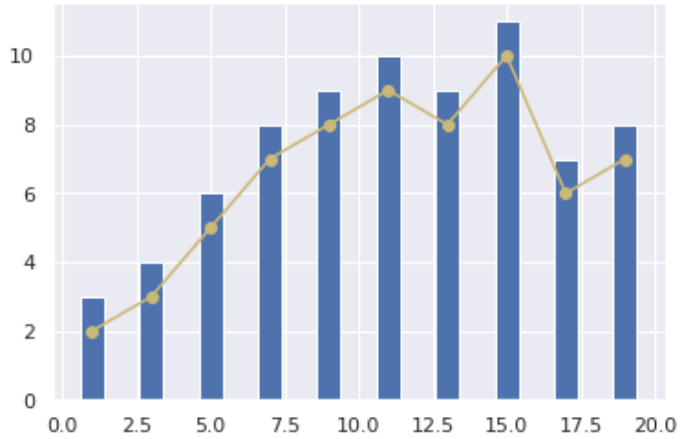
What is seaborn?

- You can simply declare seaborn style in your code

```
import matplotlib.pyplot as plt

x = [1, 3, 5, 7, 9, 11, 13, 15, 17, 19]
y_bar = [3, 4, 6, 8, 9, 10, 9, 11, 7, 8]
y_line = [2, 3, 5, 7, 8, 9, 8, 10, 6, 7]

plt.bar(x, y_bar)
plt.plot(x, y_line, '-o', color='y')
```



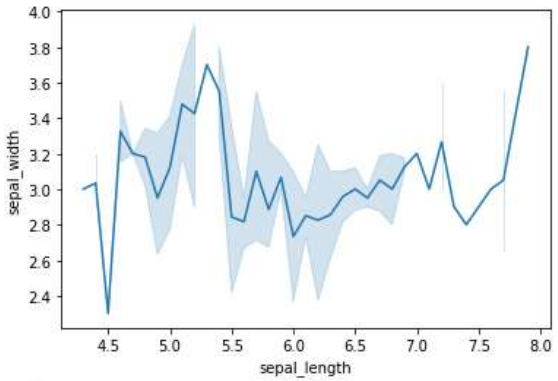

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.set()  # Declare seborn style

x = [1, 3, 5, 7, 9, 11, 13, 15, 17, 19]
y_bar = [3, 4, 6, 8, 9, 10, 9, 11, 7, 8]
y_line = [2, 3, 5, 7, 8, 9, 8, 10, 6, 7]

plt.bar(x, y_bar)
plt.plot(x, y_line, '-o', color='y')
```



What is seaborn?

- Seaborn is a library for making statistical graphics in Python.
- Seaborn comes with many built-in datasets.
- You can find more in this website:
  - https://github.com/mwaskom/seaborn-data
- That means you don't have to spend a whole lot of your time finding the right dataset and cleaning it up to make Seaborn-ready

```
import seaborn as sns

# loading dataset
data = sns.load_dataset("iris")

# draw lineplot
sns.lineplot(x="sepal_length", y="sepal_width", data=data)
```



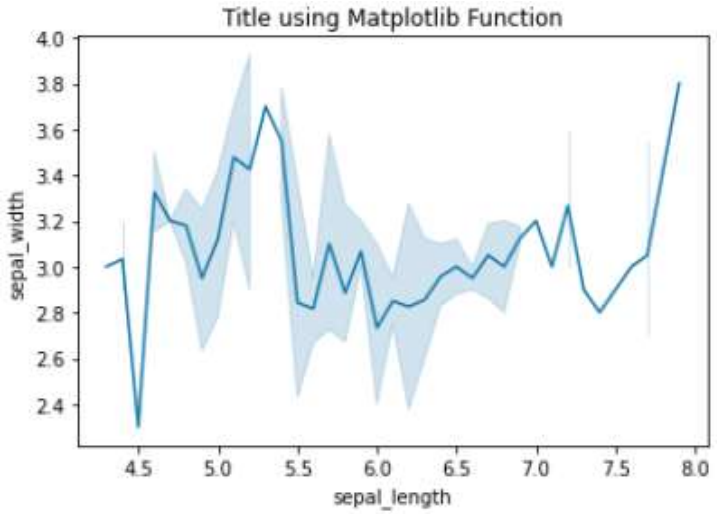
sepal_length	sepal_width	petal_length	petal_width	species

What is seaborn?

Iris flower data set

From Wikipedia, the free encyclopedia

The ***Iris* flower data set** or **Fisher's *Iris* data set** is a multivariate data set used and made famous by the British statistician and biologist Ronald Fisher in his 1936 paper *The use of multiple measurements in taxonomic problems* as an example of linear discriminant analysis.[1] It is sometimes called **Anderson's *Iris* data set** because Edgar Anderson collected the data to quantify the morphologic variation of *Iris* flowers of three related species.[2] Two of the three species were collected in the Gaspé Peninsula "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".[3]




```
# loading dataset
data = sns.load_dataset("iris")

# draw lineplot
sns.lineplot(x="sepal_length", y="sepal_width", data=data)

# setting the title using Matplotlib
plt.title('Title using Matplotlib Function')

plt.show()
```

Working with the seaborn dataset

```
import seaborn as sns
import matplotlib.pyplot as plt

iris = sns.load_dataset("iris")
iris.head()
```

Check the details of initial data

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

```
iris = sns.load_dataset("iris")
iris.head(3)
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa

Working with the seaborn dataset

```
iris = sns.load_dataset("iris")
iris['sepal_width']
```

Check the details of each data type

```
0      3.5
1      3.0
2      3.2
3      3.1
4      3.6
      ...
145    3.0
146    2.5
147    3.0
148    3.4
149    3.0
Name: sepal_width, Length: 150, dtype: float64
```

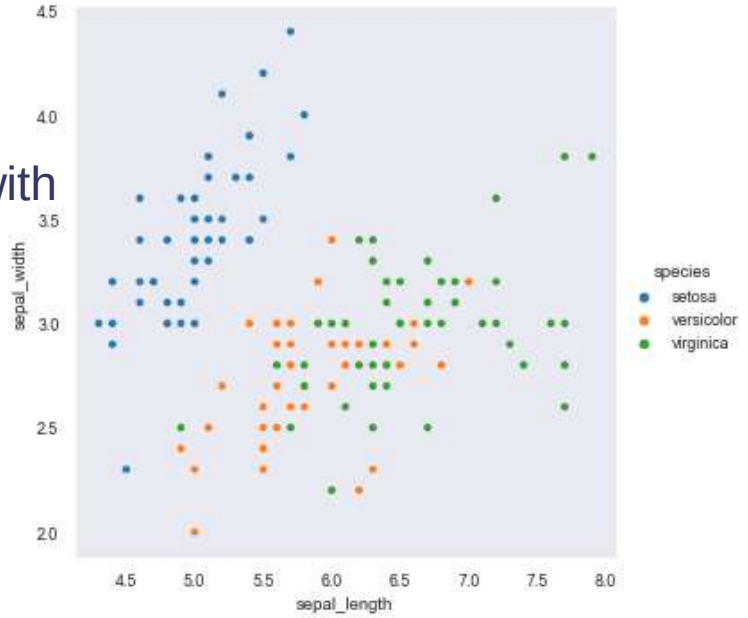
```
iris = sns.load_dataset("iris")
iris['petal_width']
```

```
0      0.2
1      0.2
2      0.2
3      0.2
4      0.2
      ...
145    2.3
146    1.9
147    2.0
148    2.3
149    1.8
Name: petal_width, Length: 150, dtype: float64
```

Working with the seaborn dataset

relplot provides access to several different axes-level functions that show the relationship between two variables with semantic mappings of subsets.
The kind parameter selects the underlying axes-level function to use:

seaborn.**relplot**(*data*=*None*, *, *x*=*None*, *y*=*None*, *hue*=*None*)
x,y: Variables that specify positions on the x and y axes.
hue*:* Grouping variable that will produce elements with different colors.

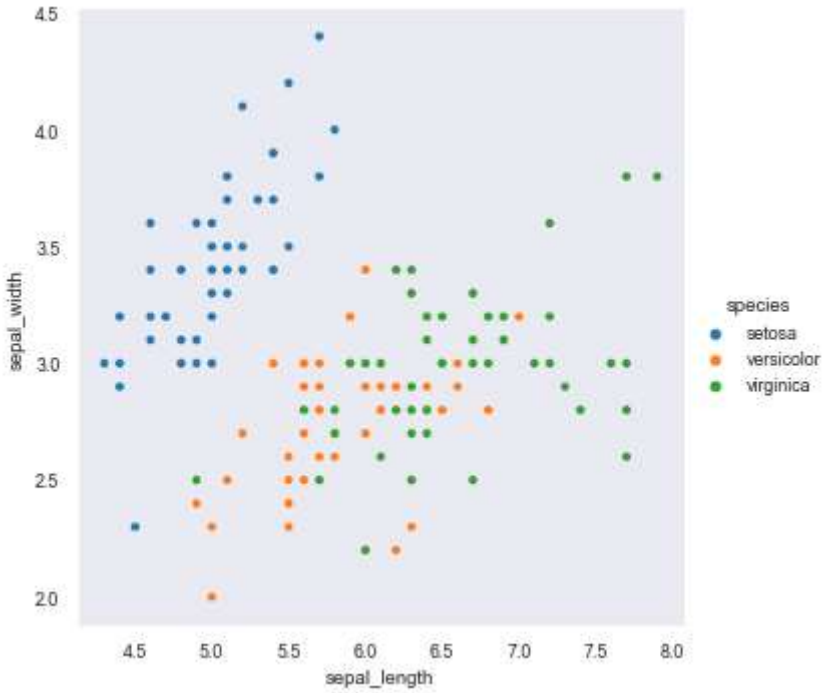


Working with the seaborn dataset

relplot provides access to several different axes-level functions that show the relationship between two variables with semantic mappings of subsets.
The kind parameter selects the underlying axes-level function to use:

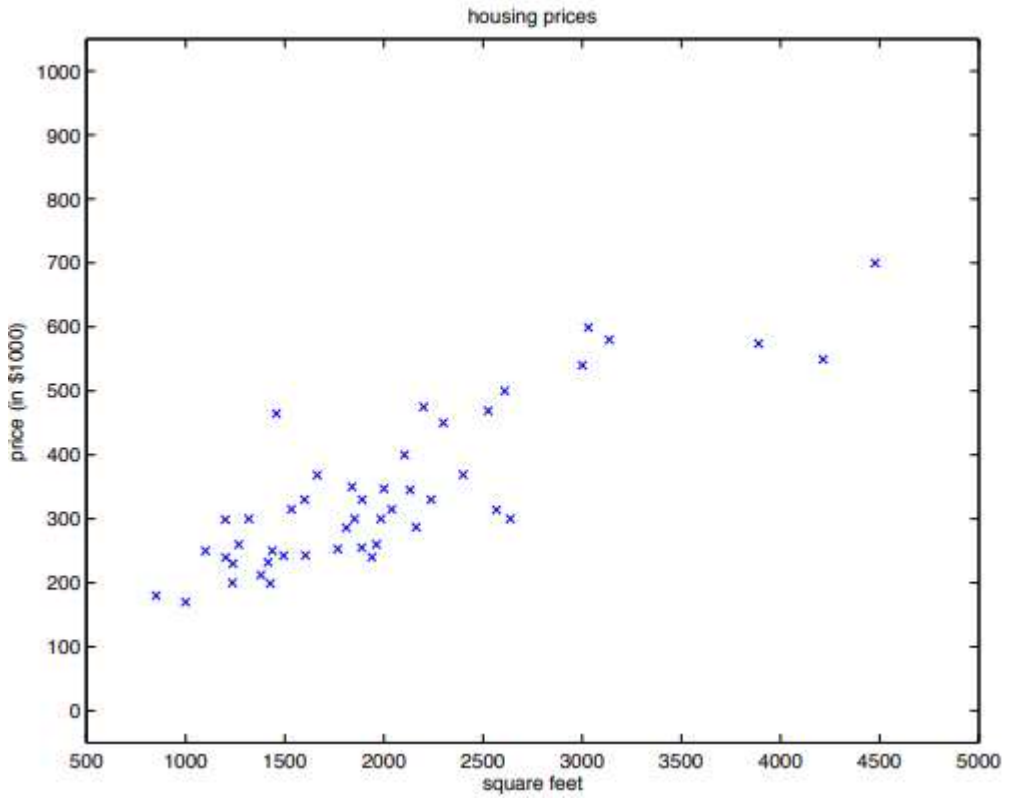
```
import seaborn as sns
import matplotlib.pyplot as plt

iris = sns.load_dataset("iris")

sns.relplot(
    data=iris,
    x="sepal_length", y="sepal_width",
kind='scatter', hue = 'species')
```

line or scatter



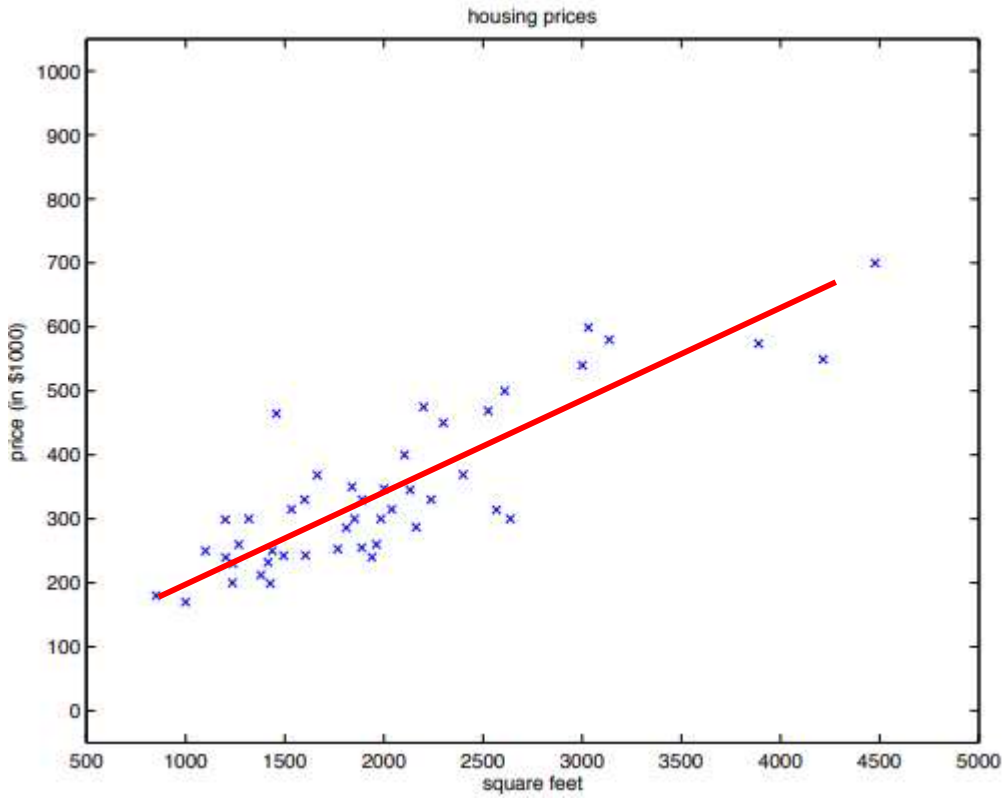
Regression analysis

Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮



Regression analysis

Living area (feet^2)	Price (1000$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮



A small exercise for you

```
['anagrams', 'anscombe', 'attention', 'brain_networks', 'car_crashes', 'diamonds',
'dowjones', 'exercise', 'flights', 'fmri', 'geyser', 'glue', 'healthexp', 'iris',
'penguins', 'planets', 'seaice', 'taxis', 'tips', 'titanic']
```

	total	speeding	alcohol	not_distracted	no_previous	ins_premium
0	18.8	7.332	5.640	18.048	15.040	784.55
1	18.1	7.421	4.525	16.290	17.014	1053.48
2	18.6	6.510	5.208	15.624	17.856	899.47
3	22.4	4.032	5.824	21.056	21.280	827.34
4	12.0	4.200	3.360	10.920	10.680	878.41

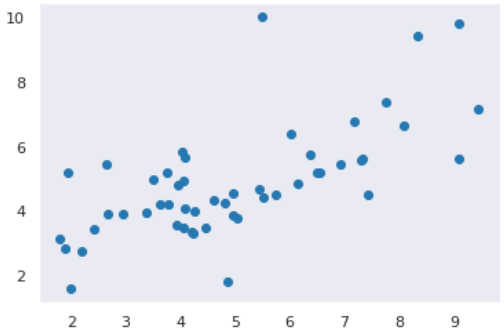
	ins_losses	abbrev
0	145.08	AL
1	133.93	AK
2	110.35	AZ
3	142.39	AR
4	165.63	CA

Use replot to show the correlation between speeding and alcohol

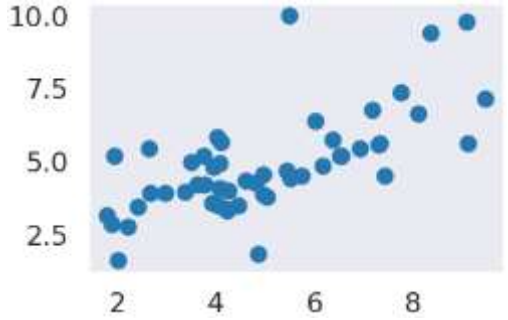
Control individual elements

- Seaborn also allows us to control individual elements of our graphs and thus we can control the scale of these elements or the plot by using the set_context() function. We have four preset templates for contexts, based on relative size, the contexts are named as follows
  - paper
  - notebook
  - talk
  - poster

```
from matplotlib import pyplot as plt
import seaborn as sns
plt.scatter(df.speeding,df.alcohol)
sns.set_style("dark")
sns.set_context("notebook")
plt.show()
```



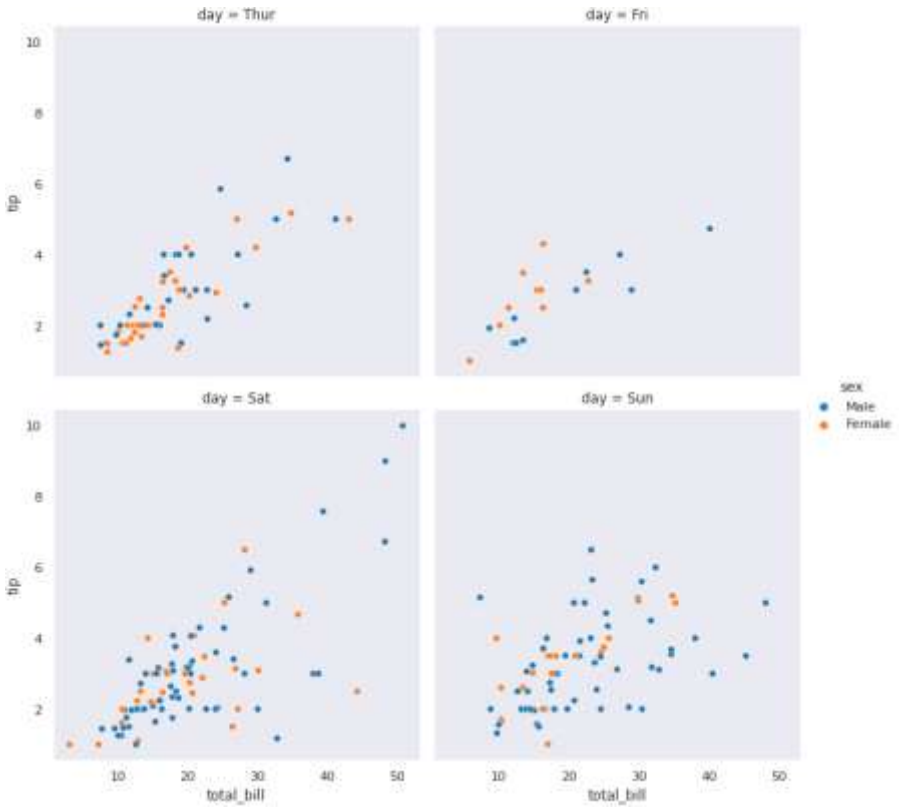

```
from matplotlib import pyplot as plt
import seaborn as sns
plt.scatter(df.speeding,df.alcohol)
sns.set_style("dark")
sns.set_context("poster")
plt.show()
```



Seaborn relplot

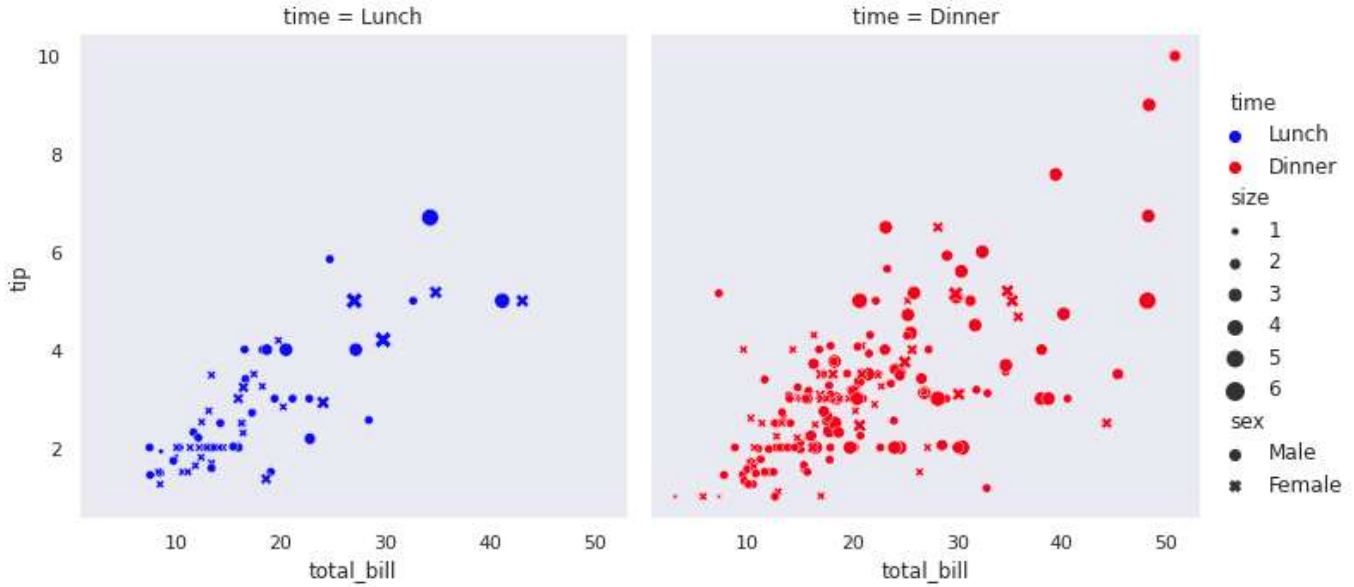
About tip dataset: One waiter recorded information about each tip he received over a period of a few months working in one restaurant. He collected several variables:

```
tips = sns.load_dataset("tips")
sns.relplot(data=tips,
x="total_bill", y="tip",
hue="sex", col="day", col_wrap=2)
```



More power functions of relplot

```
tips = sns.load_dataset("tips")
sns.relplot(data=tips, x="total_bill", y="tip", col="time", hue="time",
size="size", style="sex", palette=["b", "r"], sizes=(10, 100))
```



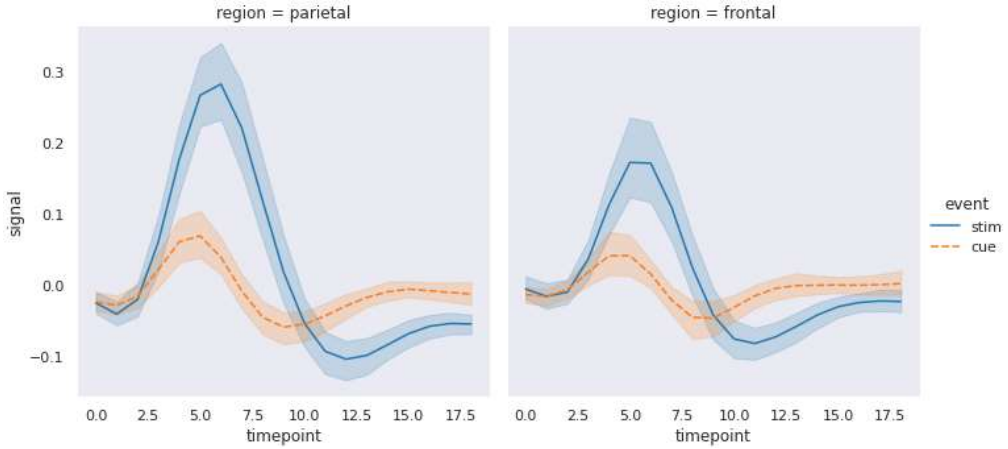
Seaborn relplot

```
fmri = sns.load_dataset("fmri")
fmri.head()
```

fmri is a popular dataset in neuroscience

	subject	timepoint	event	region	signal
0	s13	18	stim	parietal	-0.017552
1	s5	14	stim	parietal	-0.080883
2	s12	18	stim	parietal	-0.081033
3	s11	18	stim	parietal	-0.046134
4	s10	18	stim	parietal	-0.037970

```
sns.relplot(
    data=fmri, x="timepoint", y="signal", col="region",
    hue="event", style="event", kind="line",
)
```

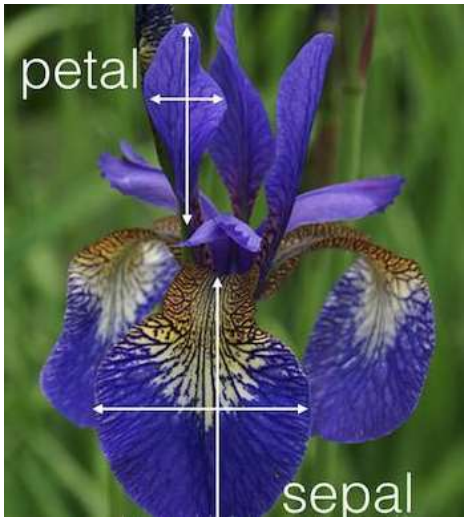


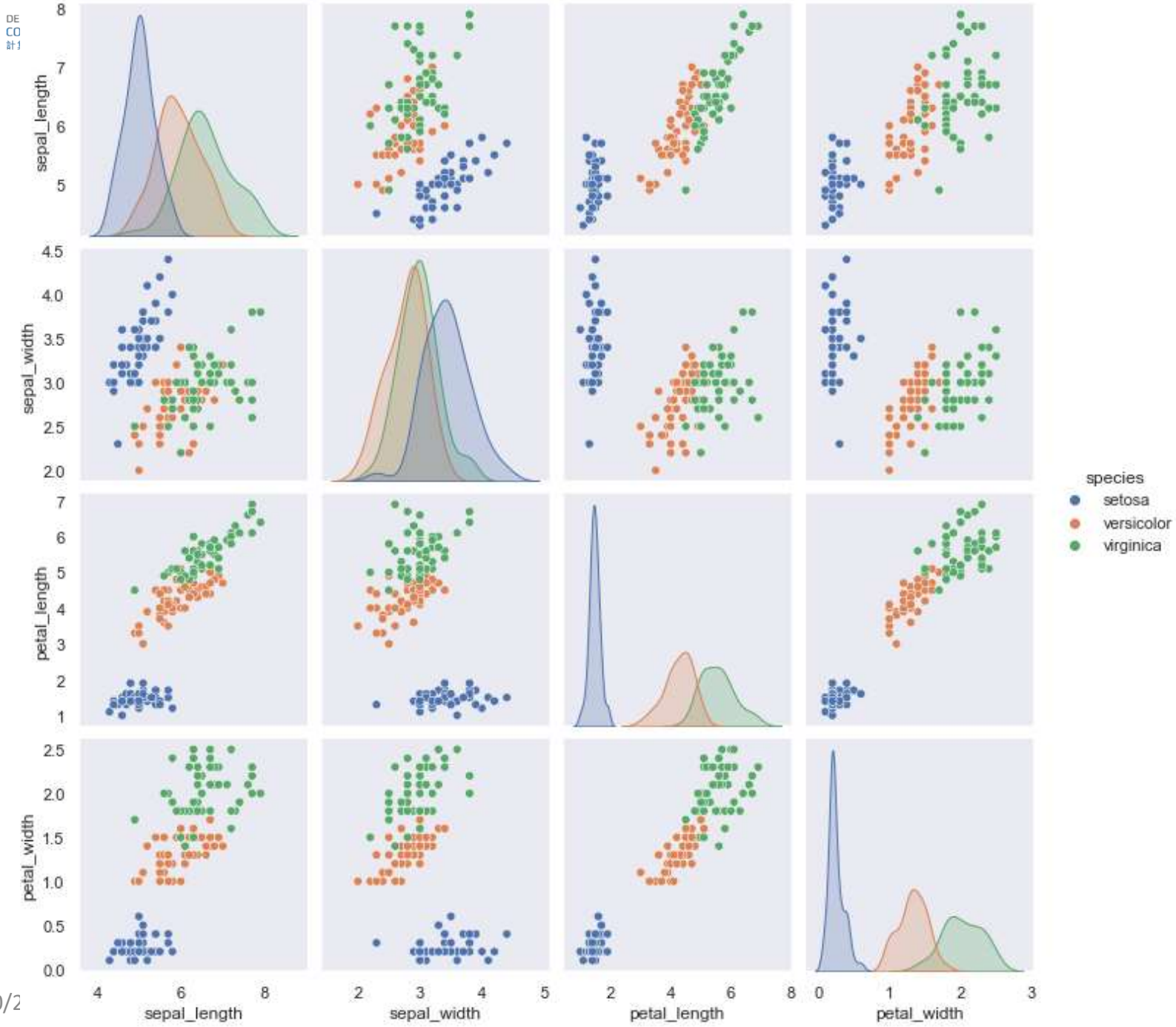
Seaborn pairplot

- Plot pairwise relationships in a dataset.
- How can we find the correlation between petal and sepal?

```
import seaborn as sns
iris = sns.load_dataset("iris")
sns.pairplot(iris, hue = 'species')
```

Different species will
be shown with different color

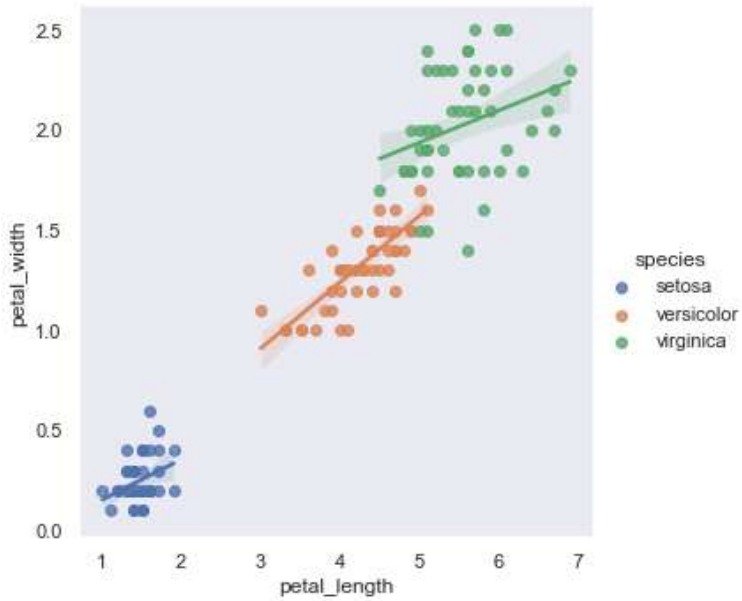



sepal_length
sepal_width
petal_length
petal_width
species
setosa
versicolor
virginica

Seaborn lmplot

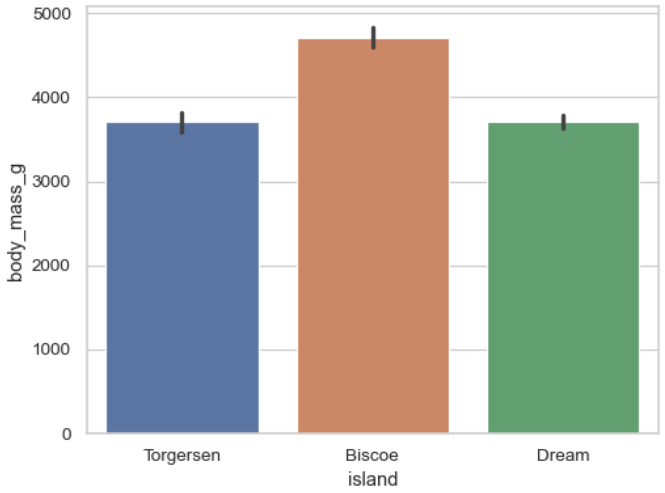
- Show the data regression
- Show the trend across a number of data samples

```
import seaborn as sns
iris = sns.load_dataset("iris")
sns.lmplot(data=iris, x="petal_length", y="petal_width", hue="species");
```



More functions about seaborn

```
df = sns.load_dataset("penguins")
sns.barplot(data=df, x="island",
y="body_mass_g")
```



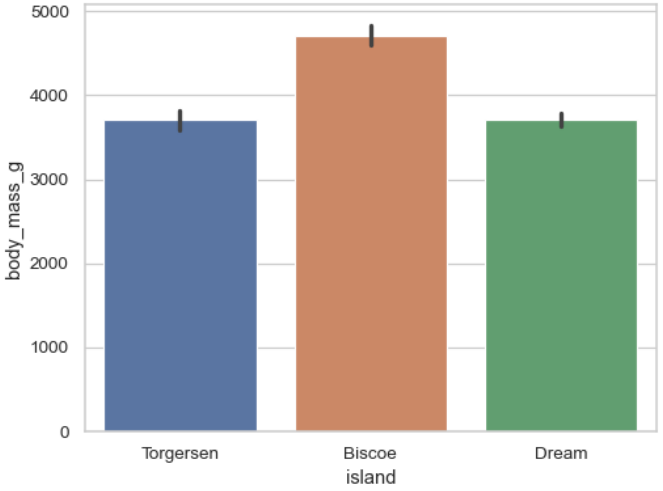

```
df = sns.load_dataset("titanic")
sns.violinplot(x=df["age"])
```

titanic dataset: about the passenges on titantic

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

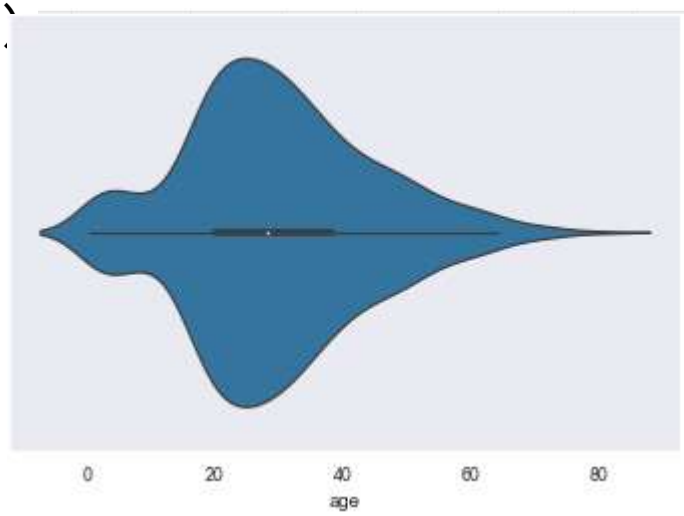
More functions about seaborn

```
df = sns.load_dataset("penguins")
sns.barplot(data=df, x="island",
y="body_mass_g")
```




```
df = sns.load_dataset("titanic")
sns.violinplot(x=df["age"])
```

titanic dataset: about the passenges on titantic



Parch	Ticket	Fare	Cabin	Embarked
0	A/5 21171	7.2500	NaN	S
0	PC 17599	71.2833	C85	C
0	STON/O2. 3101282	7.9250	NaN	S
0	113803	53.1000	C123	S
0	373450	8.0500	NaN	S