

STAT3622 Data Visualization (with Python)

Lecture 3

Wenbin Du

The University of Hong Kong

1 February 2021

Pandas: Visualization with Python

- pandas is an open source library providing high-performance, easy-to-use data structures and data analysis tools for Python

```
import pandas as pd
```

```
iris = pd.read_csv("iris.data")
```



iris setosa



iris versicolor



iris virginica

Pandas: Data Structures

- There are two primary data structures in pandas
 - Series

The series constitutes the data structure designed to accommodate a sequence of one-dimensional data,

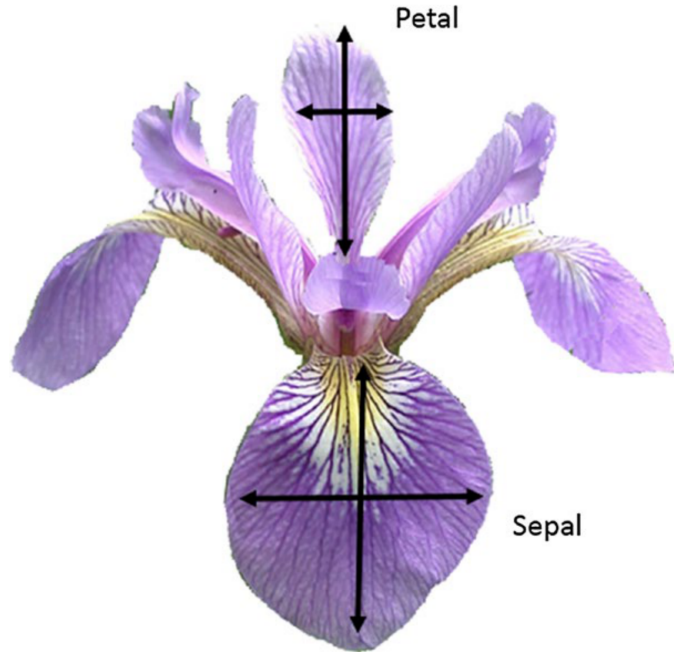
- Dataframes

Designed to contain cases with several dimensions

Pandas: Series

Series	
index	value
0	2
1	4
2	-3
3	7

Pandas: DataFrame



	SepalLength	SepalWidth	PetalLength	PetalWidth	Name
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Nelli F. Python data analytics: with pandas, numpy, and matplotlib[M]. Apress, 2018.

Pandas: DataFrame

```
type(iris)
```

pandas.core.frame.DataFrame

```
iris.index
```

RangeIndex(start=0, stop=150, step=1)

```
iris.columns
```

Index(['SepalLength', 'SepalWidth', 'PetalLength', 'PetalWidth', 'Name'], dtype='object')

```
iris.values[:3,:4]
```

```
array([[5.1, 3.5, 1.4, 0.2],  
       [4.9, 3.0, 1.4, 0.2],  
       [4.7, 3.2, 1.3, 0.2]], dtype=object)
```

Pandas: DataFrame

```
iris.index
```

```
RangeIndex(start=0, stop=150, step=1)
```

```
iris.columns
```

```
Index(['SepalLength', 'SepalWidth', 'PetalLength', 'PetalWidth', 'Species'], dtype='object')
```

```
iris.values
```

```
array([[5.1, 3.5, 1.4, 0.2, 'Iris-setosa'],  
       [4.9, 3.0, 1.4, 0.2, 'Iris-setosa'],  
       [4.7, 3.2, 1.3, 0.2, 'Iris-setosa'],  
       [4.6, 3.1, 1.5, 0.2, 'Iris-setosa'],  
       [5.0, 3.6, 1.4, 0.2, 'Iris-setosa'],  
       [5.4, 3.9, 1.7, 0.4, 'Iris-setosa'],  
       [4.6, 3.4, 1.4, 0.3, 'Iris-setosa'],  
       [5.0, 3.4, 1.5, 0.2, 'Iris-setosa'],  
       [4.4, 2.9, 1.4, 0.2, 'Iris-setosa'],  
       [4.9, 3.1, 1.5, 0.1, 'Iris-setosa'],  
       [5.4, 3.7, 1.5, 0.2, 'Iris-setosa'],  
       [4.8, 3.4, 1.6, 0.2, 'Iris-setosa'],  
       [4.8, 3.0, 1.4, 0.1, 'Iris-setosa'],  
       [4.3, 3.0, 1.1, 0.1, 'Iris-setosa'],  
       [5.8, 4.0, 1.2, 0.2, 'Iris-setosa'],  
       [5.7, 4.4, 1.5, 0.4, 'Iris-setosa'],  
       [5.4, 3.9, 1.3, 0.4, 'Iris-setosa'],  
       [5.1, 3.5, 1.4, 0.3, 'Iris-setosa'],  
       [5.7, 3.8, 1.7, 0.3, 'Iris-setosa']])
```

Pandas: DataFrame attributes

```
iris.head()
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Selecting Elements

```
iris['SepalLength']
```

```
iris.SepalLength
```

```
0      5.1
1      4.9
2      4.7
3      4.6
4      5.0
...
145    6.7
146    6.3
147    6.5
148    6.2
149    5.9
Name: SepalLength, Length: 150, dtype: float64
```

Selecting Elements

```
iris.iloc[1]
```

```
SepalLength      4.9
SepalWidth        3
PetalLength       1.4
PetalWidth        0.2
Species           Iris-setosa
Name: 1, dtype: object
```

```
iris.iloc[[2,4]]
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species
2	4.7	3.2	1.3	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Selecting Elements

```
iris[1:10]
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa

```
iris['Species'][2]
```

'Iris-setosa'

Filtering

```
setosa = iris[iris['Species']=='Iris-setosa']
```

```
iris_part = iris[iris['SepalWidth']>3.5]
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
10	5.4	3.7	1.5	0.2	Iris-setosa
14	5.8	4.0	1.2	0.2	Iris-setosa
15	5.7	4.4	1.5	0.4	Iris-setosa
16	5.4	3.9	1.3	0.4	Iris-setosa
18	5.7	3.8	1.7	0.3	Iris-setosa
19	5.1	3.8	1.5	0.3	Iris-setosa
21	5.1	3.7	1.5	0.4	Iris-setosa
22	4.6	3.6	1.0	0.2	Iris-setosa
32	5.2	4.1	1.5	0.1	Iris-setosa
33	5.5	4.2	1.4	0.2	Iris-setosa
44	5.1	3.8	1.9	0.4	Iris-setosa
46	5.1	3.8	1.6	0.2	Iris-setosa
48	5.3	3.7	1.5	0.2	Iris-setosa
109	7.2	3.6	6.1	2.5	Iris-virginica
117	7.7	3.8	6.7	2.2	Iris-virginica
131	7.9	3.8	6.4	2.0	Iris-virginica

Filtering

```
setosa = iris[(iris['Species']=='Iris-setosa') & (iris['SepalLength']>5)]  
setosa
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species
0	5.1	3.5	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
10	5.4	3.7	1.5	0.2	Iris-setosa
14	5.8	4.0	1.2	0.2	Iris-setosa
15	5.7	4.4	1.5	0.4	Iris-setosa
16	5.4	3.9	1.3	0.4	Iris-setosa
17	5.1	3.5	1.4	0.3	Iris-setosa
18	5.7	3.8	1.7	0.3	Iris-setosa
19	5.1	3.8	1.5	0.3	Iris-setosa
20	5.4	3.4	1.7	0.2	Iris-setosa
21	5.1	3.7	1.5	0.4	Iris-setosa
23	5.1	3.3	1.7	0.5	Iris-setosa
27	5.2	3.5	1.5	0.2	Iris-setosa
28	5.2	3.4	1.4	0.2	Iris-setosa

Filtering

```
flower = iris[(iris['Species']=='Iris-setosa')|(iris['SepalLength']>6)]  
flower
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...
144	6.7	3.3	5.7	2.5	Iris-virginica
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica

111 rows x 5 columns

Adding a column

```
sepal_size = iris['SepalLength']*iris['SepalWidth']  
print(sepal_size)  
iris['SepalSize']=sepal_size  
iris.head()
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species	SepalSize
0	5.1	3.5	1.4	0.2	Iris-setosa	17.85
1	4.9	3.0	1.4	0.2	Iris-setosa	14.70
2	4.7	3.2	1.3	0.2	Iris-setosa	15.04
3	4.6	3.1	1.5	0.2	Iris-setosa	14.26
4	5.0	3.6	1.4	0.2	Iris-setosa	18.00

Statistics Functions

```
iris.iloc[:, :3].sum()
```

```
SepalLength      876.5  
SepalWidth       458.1  
PetalLength      563.8  
dtype: float64
```

```
iris.iloc[:, :3].mean()
```

```
SepalLength      5.843333  
SepalWidth       3.054000  
PetalLength      3.758667  
dtype: float64
```

Statistics Functions

```
iris.describe()
```

	SepalLength	SepalWidth	PetalLength	PetalWidth
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Sorting

```
iris.sort_index()
```

```
iris.sort_index(ascending=False)
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species		SepalLength	SepalWidth	PetalLength	PetalWidth	Species
0	5.1	3.5	1.4	0.2	Iris-setosa	149	5.9	3.0	5.1	1.8	Iris-virginica
1	4.9	3.0	1.4	0.2	Iris-setosa	148	6.2	3.4	5.4	2.3	Iris-virginica
2	4.7	3.2	1.3	0.2	Iris-setosa	147	6.5	3.0	5.2	2.0	Iris-virginica
3	4.6	3.1	1.5	0.2	Iris-setosa	146	6.3	2.5	5.0	1.9	Iris-virginica
4	5.0	3.6	1.4	0.2	Iris-setosa	145	6.7	3.0	5.2	2.3	Iris-virginica
...
145	6.7	3.0	5.2	2.3	Iris-virginica	4	5.0	3.6	1.4	0.2	Iris-setosa
146	6.3	2.5	5.0	1.9	Iris-virginica	3	4.6	3.1	1.5	0.2	Iris-setosa
147	6.5	3.0	5.2	2.0	Iris-virginica	2	4.7	3.2	1.3	0.2	Iris-setosa
148	6.2	3.4	5.4	2.3	Iris-virginica	1	4.9	3.0	1.4	0.2	Iris-setosa
149	5.9	3.0	5.1	1.8	Iris-virginica	0	5.1	3.5	1.4	0.2	Iris-setosa

Sorting

```
iris.sort_values(by='SepalLength')
```

```
iris.sort_values(by=[ 'SepalLength', 'SepalWidth' ])
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species		SepalLength	SepalWidth	PetalLength	PetalWidth	Species
13	4.3	3.0	1.1	0.1	Iris-setosa	13	4.3	3.0	1.1	0.1	Iris-setosa
42	4.4	3.2	1.3	0.2	Iris-setosa	8	4.4	2.9	1.4	0.2	Iris-setosa
38	4.4	3.0	1.3	0.2	Iris-setosa	38	4.4	3.0	1.3	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa	42	4.4	3.2	1.3	0.2	Iris-setosa
41	4.5	2.3	1.3	0.3	Iris-setosa	41	4.5	2.3	1.3	0.3	Iris-setosa
...
122	7.7	2.8	6.7	2.0	Iris-virginica	118	7.7	2.6	6.9	2.3	Iris-virginica
118	7.7	2.6	6.9	2.3	Iris-virginica	122	7.7	2.8	6.7	2.0	Iris-virginica
117	7.7	3.8	6.7	2.2	Iris-virginica	135	7.7	3.0	6.1	2.3	Iris-virginica
135	7.7	3.0	6.1	2.3	Iris-virginica	117	7.7	3.8	6.7	2.2	Iris-virginica
131	7.9	3.8	6.4	2.0	Iris-virginica	131	7.9	3.8	6.4	2.0	Iris-virginica

Sorting

```
iris.sort_index(axis=1)  
##sort the data by col name in alphabetical order
```

```
iris.sort_index(axis=1,ascending=False)
```

	Species	SepalWidth	SepalSize	SepalLength	PetalWidth	PetalLength
0	Iris-setosa	3.5	17.85	5.1	0.2	1.4
1	Iris-setosa	3.0	14.70	4.9	0.2	1.4
2	Iris-setosa	3.2	15.04	4.7	0.2	1.3
3	Iris-setosa	3.1	14.26	4.6	0.2	1.5
4	Iris-setosa	3.6	18.00	5.0	0.2	1.4
...
145	Iris-virginica	3.0	20.10	6.7	2.3	5.2
146	Iris-virginica	2.5	15.75	6.3	1.9	5.0
147	Iris-virginica	3.0	19.50	6.5	2.0	5.2
148	Iris-virginica	3.4	21.08	6.2	2.3	5.4
149	Iris-virginica	3.0	17.70	5.9	1.8	5.1

	PetalLength	PetalWidth	SepalLength	SepalSize	SepalWidth	Species
0	1.4	0.2	5.1	17.85	3.5	Iris-setosa
1	1.4	0.2	4.9	14.70	3.0	Iris-setosa
2	1.3	0.2	4.7	15.04	3.2	Iris-setosa
3	1.5	0.2	4.6	14.26	3.1	Iris-setosa
4	1.4	0.2	5.0	18.00	3.6	Iris-setosa
...
145	5.2	2.3	6.7	20.10	3.0	Iris-virginica
146	5.0	1.9	6.3	15.75	2.5	Iris-virginica
147	5.2	2.0	6.5	19.50	3.0	Iris-virginica
148	5.4	2.3	6.2	21.08	3.4	Iris-virginica
149	5.1	1.8	5.9	17.70	3.0	Iris-virginica

Sorting: sort values

```
iris.sort_values(by='SepalLength')
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species	SepalSize
13	4.3	3.0	1.1	0.1	Iris-setosa	12.90
42	4.4	3.2	1.3	0.2	Iris-setosa	14.08
38	4.4	3.0	1.3	0.2	Iris-setosa	13.20
8	4.4	2.9	1.4	0.2	Iris-setosa	12.76
41	4.5	2.3	1.3	0.3	Iris-setosa	10.35
...
122	7.7	2.8	6.7	2.0	Iris-virginica	21.56
118	7.7	2.6	6.9	2.3	Iris-virginica	20.02
117	7.7	3.8	6.7	2.2	Iris-virginica	29.26
135	7.7	3.0	6.1	2.3	Iris-virginica	23.10
131	7.9	3.8	6.4	2.0	Iris-virginica	30.02

Sorting: sort values

```
iris.sort_values(by=[ 'SepalLength' , 'SepalWidth' ] )
```

```
iris.sort_values(by=[ 'SepalLength' , 'SepalWidth' ] , ascending=[ True , False ] )
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species	SepalSize
13	4.3	3.0	1.1	0.1	Iris-setosa	12.90
8	4.4	2.9	1.4	0.2	Iris-setosa	12.76
38	4.4	3.0	1.3	0.2	Iris-setosa	13.20
42	4.4	3.2	1.3	0.2	Iris-setosa	14.08
41	4.5	2.3	1.3	0.3	Iris-setosa	10.35
...
118	7.7	2.6	6.9	2.3	Iris-virginica	20.02
122	7.7	2.8	6.7	2.0	Iris-virginica	21.56
135	7.7	3.0	6.1	2.3	Iris-virginica	23.10
117	7.7	3.8	6.7	2.2	Iris-virginica	29.26
131	7.9	3.8	6.4	2.0	Iris-virginica	30.02

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species	SepalSize
13	4.3	3.0	1.1	0.1	Iris-setosa	12.90
42	4.4	3.2	1.3	0.2	Iris-setosa	14.08
38	4.4	3.0	1.3	0.2	Iris-setosa	13.20
8	4.4	2.9	1.4	0.2	Iris-setosa	12.76
41	4.5	2.3	1.3	0.3	Iris-setosa	10.35
...
117	7.7	3.8	6.7	2.2	Iris-virginica	29.26
135	7.7	3.0	6.1	2.3	Iris-virginica	23.10
122	7.7	2.8	6.7	2.0	Iris-virginica	21.56
118	7.7	2.6	6.9	2.3	Iris-virginica	20.02
131	7.9	3.8	6.4	2.0	Iris-virginica	30.02

Rename

```
mini_iris = iris[:3]
mini_iris = mini_iris.rename(columns={'SepalLength': 'SL'})
mini_iris = mini_iris.rename(index={0: 'first', 1: 'second', 2: 'third'})
```

	SL	SepalWidth	PetalLength	PetalWidth	Species	SepalSize
first	5.1	3.5	1.4	0.2	Iris-setosa	17.85
second	4.9	3.0	1.4	0.2	Iris-setosa	14.70
third	4.7	3.2	1.3	0.2	Iris-setosa	15.04

del & drop

```
iris_cp = iris.copy()
del iris_cp[ 'SepalLength' ]
iris_cp
```

	SepalWidth	PetalLength	PetalWidth	Species	SepalSize
0	3.5	1.4	0.2	Iris-setosa	17.85
1	3.0	1.4	0.2	Iris-setosa	14.70
2	3.2	1.3	0.2	Iris-setosa	15.04
3	3.1	1.5	0.2	Iris-setosa	14.26
4	3.6	1.4	0.2	Iris-setosa	18.00
...
145	3.0	5.2	2.3	Iris-virginica	20.10
146	2.5	5.0	1.9	Iris-virginica	15.75
147	3.0	5.2	2.0	Iris-virginica	19.50
148	3.4	5.4	2.3	Iris-virginica	21.08
149	3.0	5.1	1.8	Iris-virginica	17.70

del & drop

```
iris_cp = iris.copy()
iris_cp=iris_cp.drop(range(10))
iris_cp
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species	SepalSize
10	5.4	3.7	1.5	0.2	Iris-setosa	19.98
11	4.8	3.4	1.6	0.2	Iris-setosa	16.32
12	4.8	3.0	1.4	0.1	Iris-setosa	14.40
13	4.3	3.0	1.1	0.1	Iris-setosa	12.90
14	5.8	4.0	1.2	0.2	Iris-setosa	23.20
...
145	6.7	3.0	5.2	2.3	Iris-virginica	20.10
146	6.3	2.5	5.0	1.9	Iris-virginica	15.75
147	6.5	3.0	5.2	2.0	Iris-virginica	19.50
148	6.2	3.4	5.4	2.3	Iris-virginica	21.08
149	5.9	3.0	5.1	1.8	Iris-virginica	17.70

140 rows × 6 columns

Groupby

```
group=iris.groupby(iris.Species)
#<pandas.core.groupby.generic.DataFrameGroupBy object at 0x7fc0ad445ac0>
group.mean()
group.max()
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	SepalSize
Species					
Iris-setosa	5.006	3.418	1.464	0.244	17.2088
Iris-versicolor	5.936	2.770	4.260	1.326	16.5262
Iris-virginica	6.588	2.974	5.552	2.026	19.6846

	SepalLength	SepalWidth	PetalLength	PetalWidth	SepalSize
Species					
Iris-setosa	5.8	4.4	1.9	0.6	25.08
Iris-versicolor	7.0	3.4	5.1	1.8	22.40
Iris-virginica	7.9	3.8	6.9	2.5	30.02

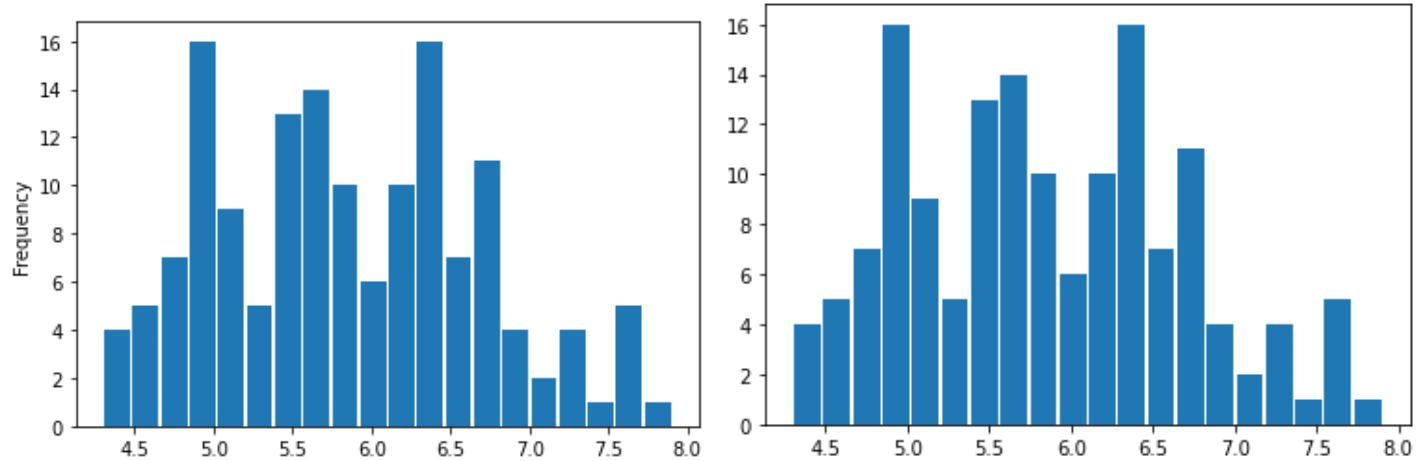
Pandas: DataFrame attributes

```
iris.describe()
```

	SepalLength	SepalWidth	PetalLength	PetalWidth
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

DataFrame: hist

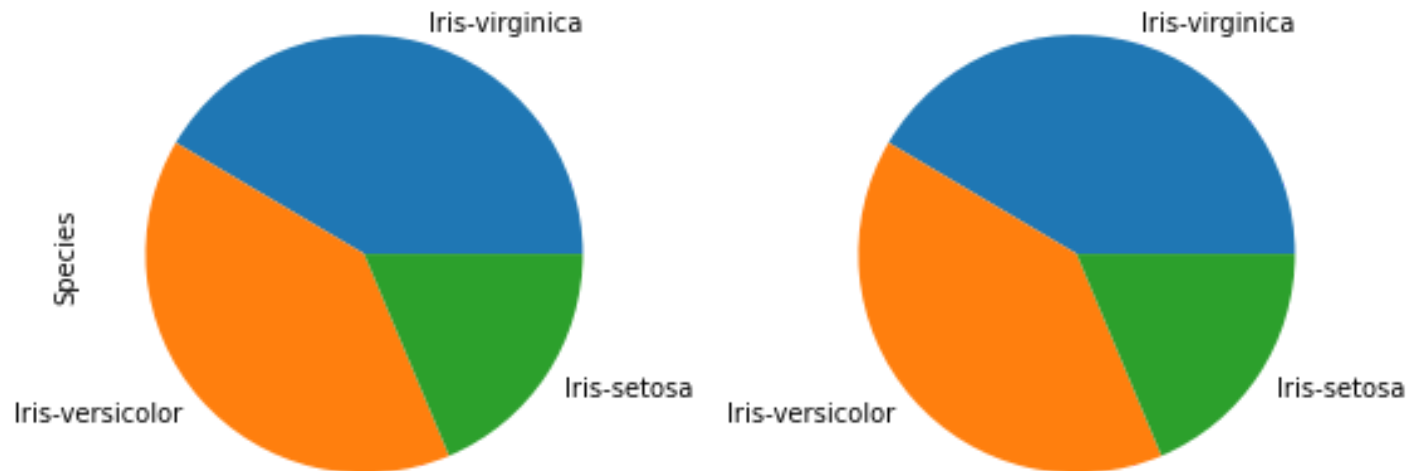
```
x = iris["SepalLength"]  
x.plot.hist( bins=20,rwidth=0.9) # pandas  
plt.hist(sepal_length,bins=20,rwidth=0.9)# matplotlib
```



Pandas: pie&bar

```
flag = sepal_length>5  
species = iris["Species"]  
species = species[flag]  
data = species.value_counts()
```

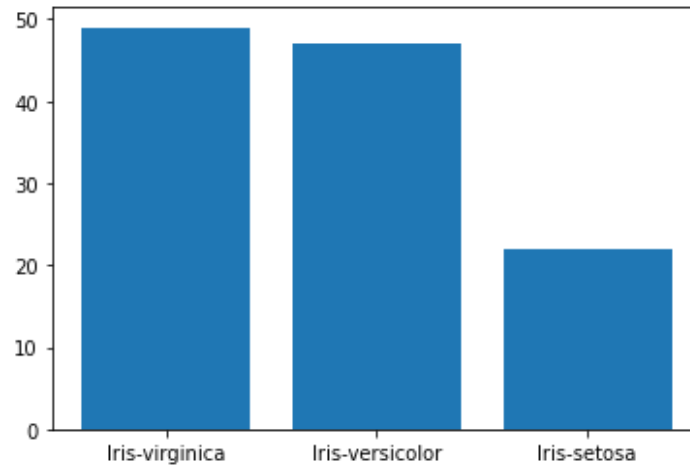
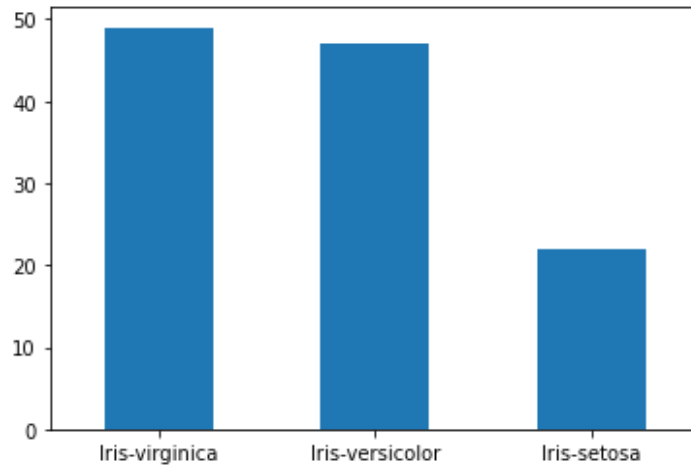
```
data.plot.pie(y="Species",figsize=(5,5)) # pandas  
plt.pie(x=data,labels=data.index) # matplotlib
```



Pandas: pie&bar

```
flag = sepal_length>5  
species = iris["Species"]  
species = species[flag]  
data = species.value_counts()
```

```
data.plot.bar(x=data.index,y=data.values,rot=0) # pandas  
plt.bar(data.index,data.values)# matplotlib
```



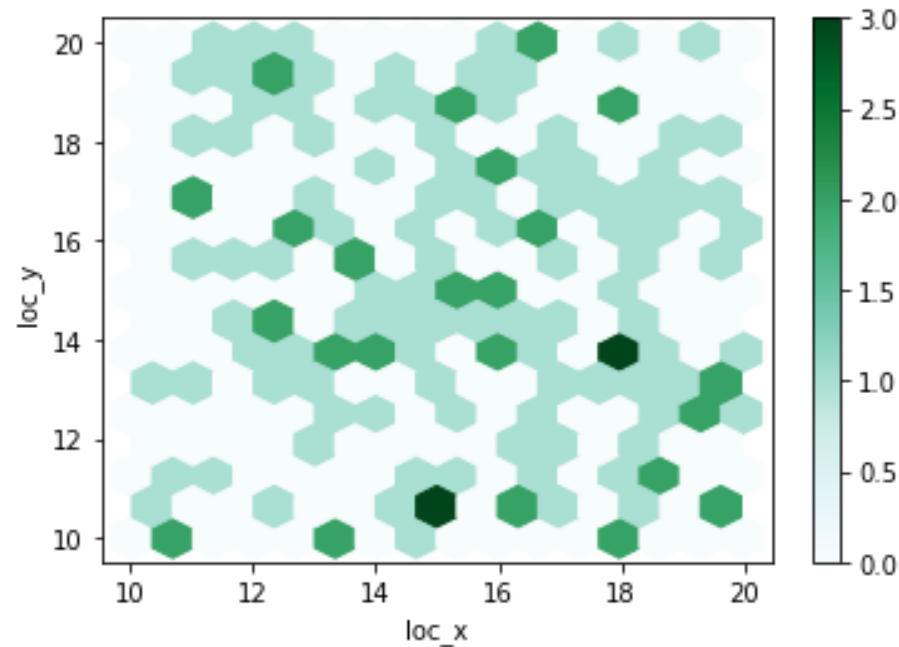
Bivarible: dataframe.plot.hexbin

```
iris['loc_x']=np.random.uniform(10,20,150)  
#Draw samples from a uniform distribution  
#Lower/Upper boundary of the output interval  
iris['loc_y']=np.random.uniform(10,20,150)  
iris['value']=np.random.uniform(0,3,150)  
iris.head()
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	Species	value	loc_x	loc_y
0	5.1	3.5	1.4	0.2	Iris-setosa	2.936846	17.486732	13.581169
1	4.9	3.0	1.4	0.2	Iris-setosa	2.310197	13.625390	18.725476
2	4.7	3.2	1.3	0.2	Iris-setosa	2.720488	11.824858	13.616035
3	4.6	3.1	1.5	0.2	Iris-setosa	2.171980	17.359638	11.216383
4	5.0	3.6	1.4	0.2	Iris-setosa	1.174683	19.189866	19.643002

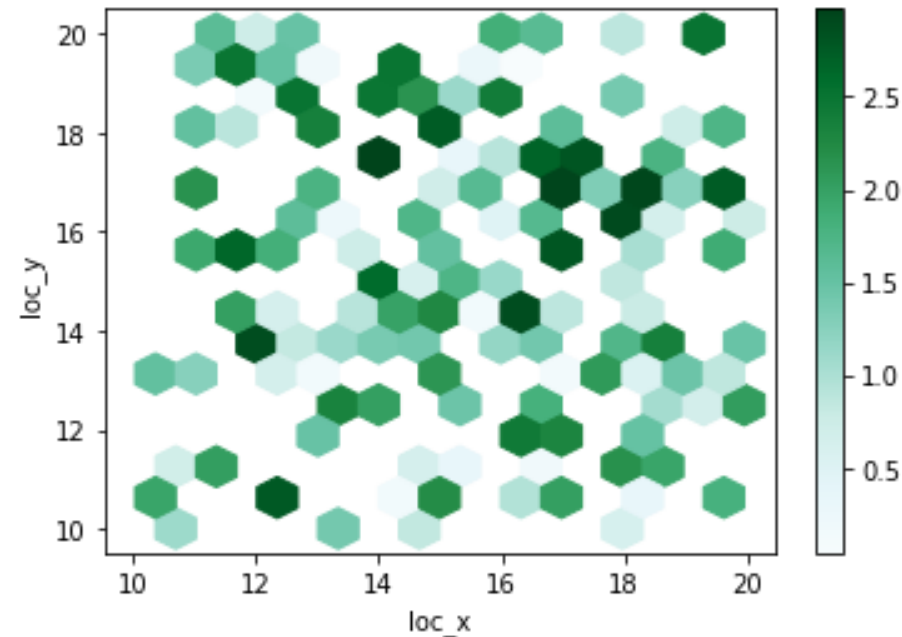
Bivarible: dataframe.plot.hexbin

```
iris['loc_x']=np.random.uniform(10,20,150)  
#Draw samples from a uniform distribution  
#Lower/Upper boundary of the output interval  
iris['loc_y']=np.random.uniform(10,20,150)  
iris['value']=np.random.uniform(0,3,150)  
iris.plot.hexbin(x="loc_x", y="loc_y", sharex=False,gridsize=15)
```



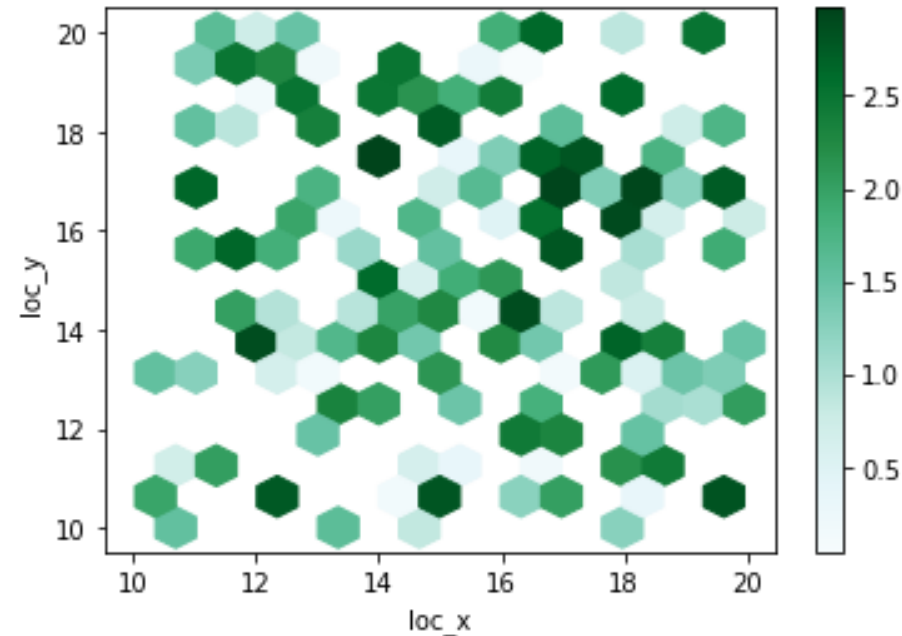
Bivarible: dataframe.plot.hexbin

```
iris['loc_x']=np.random.uniform(10,20,150)  
#Draw samples from a uniform distribution  
#Lower/Upper boundary of the output interval  
iris['loc_y']=np.random.uniform(10,20,150)  
iris['value']=np.random.uniform(0,3,150)  
iris.plot.hexbin(x="loc_x", y="loc_y",C="value", sharex=False,gridsize=15)
```



Bivarible: dataframe.plot.hexbin

```
iris['loc_x']=np.random.uniform(10,20,150)
iris['loc_y']=np.random.uniform(10,20,150)
iris['value']=np.random.uniform(0,3,150)
iris.plot.hexbin(x="loc_x", y="loc_y",C="value", sharex=False,
                reduce_C_function=np.max,gridsize=15)
```



Thank you!

Q&A or Email wbd@hku.hk