Data Exploration

January 5, 2019

In [3]:
```
# load data
import pandas as pd
iris_file = 'iris.csv'
iris = pd.read_csv(iris_file, header=None, names=['sepal_length', 'sepal_width',
                                                  'petal_length','petal_width',
                                                  'target'])

iris.head()
```

Out[3]:

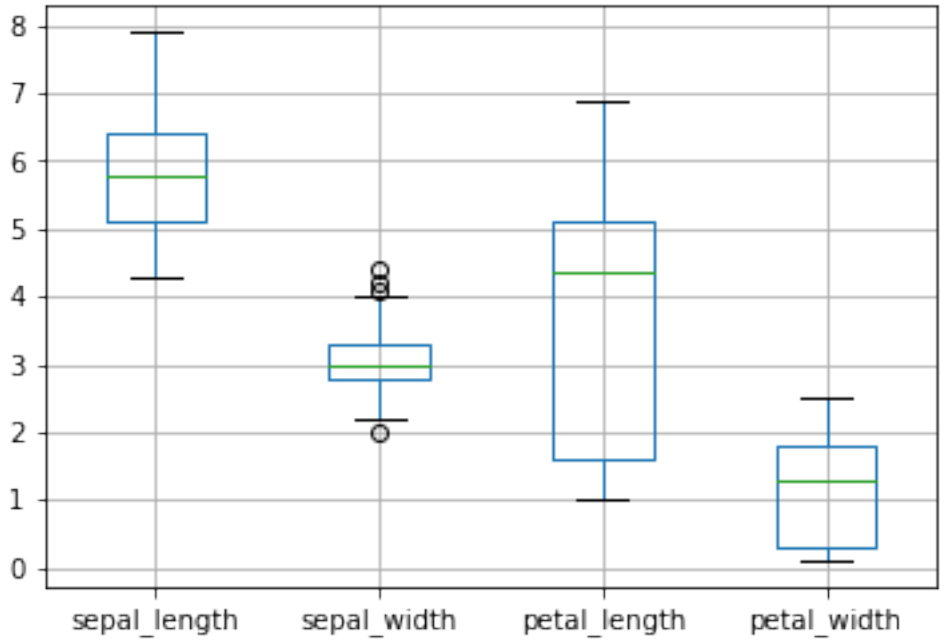
	sepal_length	sepal_width	petal_length	petal_width	target
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

In [4]:
```
iris.describe()
```

Out[4]:

	sepal_length	sepal_width	petal_length	petal_width
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

In [7]:
```
#box plote
box = iris.boxplot(return_type='axes')
```




```
In [8]: # quantile 1
        iris.quantile([0.1,0.9])
```

Out[8]:

	sepal_length	sepal_width	petal_length	petal_width
0.1	4.8	2.50	1.4	0.2
0.9	6.9	3.61	5.8	2.2

```
In [9]: # quantile 2
        iris.quantile([0.01,0.02])
```

Out[9]:

	sepal_length	sepal_width	petal_length	petal_width
0.01	4.4	2.2	1.149	0.1
0.02	4.4	2.2	1.200	0.1

```
In [10]: # quantile 3
         iris.quantile([0.99])
```

Out[10]:

	sepal_length	sepal_width	petal_length	petal_width
0.99	7.7	4.151	6.7	2.5

```
In [11]: # get categorical features
         iris.target.unique()
```

```
Out[11]: array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

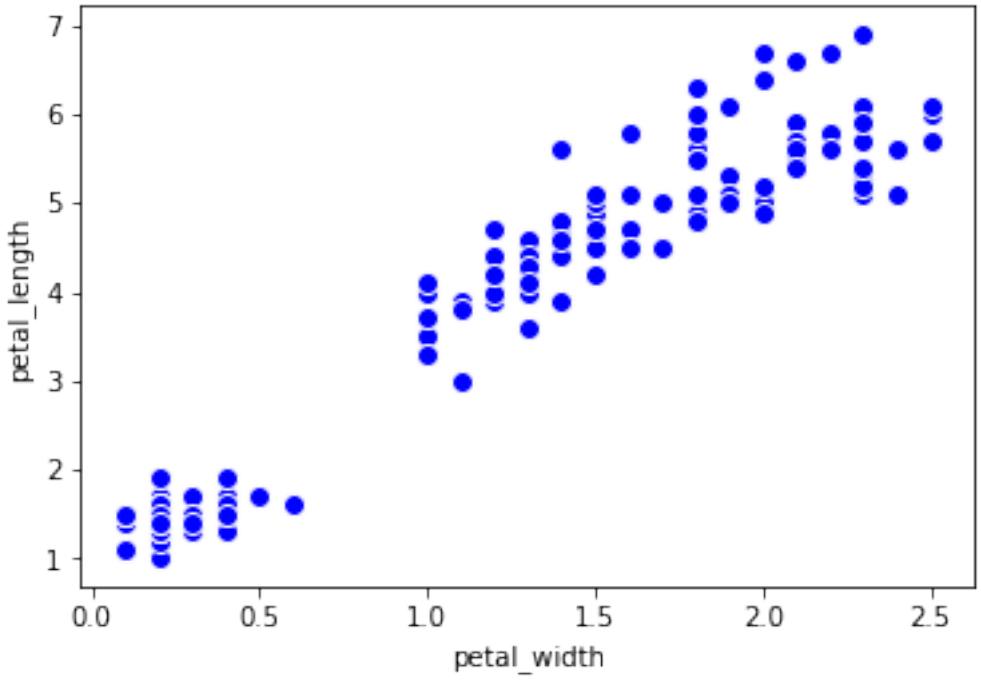
```
In [15]: #similarity matrix
         # 3.758667 and 1.198667 are mens
         pd.crosstab(iris['petal_length'] > 3.758667,iris['petal_width'] > 1.198667)
```

```
Out[15]: petal_width   False  True
         petal_length
         False            56      1
         True              4     89
```

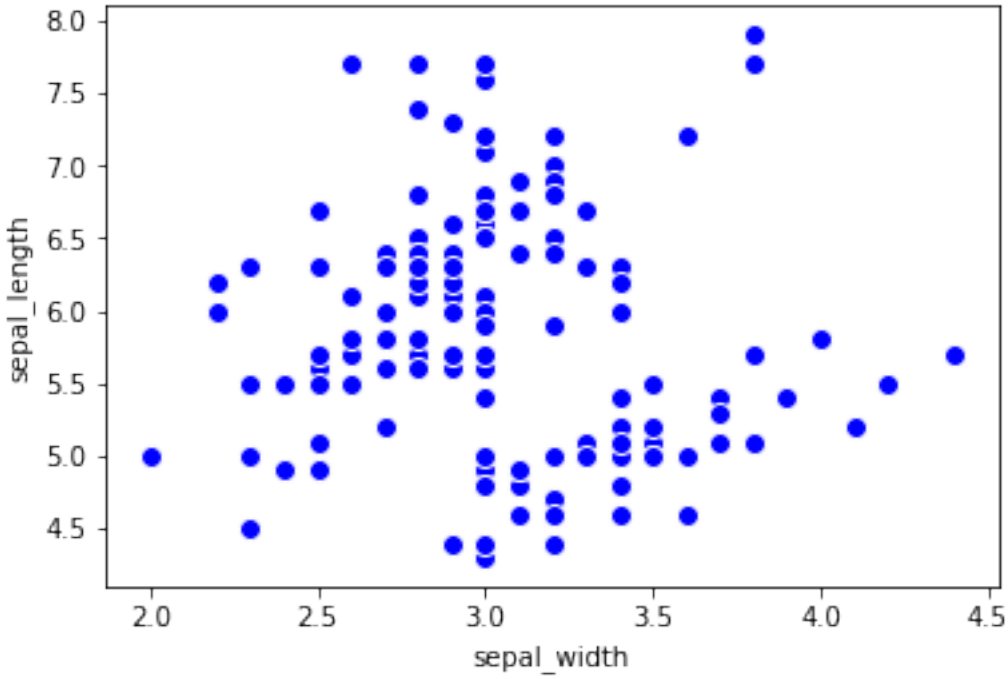
```
In [23]: #similarity matrix 2
         # 5.843333 and 1.198667 are mens
         pd.crosstab(iris['sepal_length'] > 5.843333        ,iris['sepal_width'] > 1.198667)
```

```
Out[23]: sepal_width   True
         sepal_length
         False           80
         True            70
```

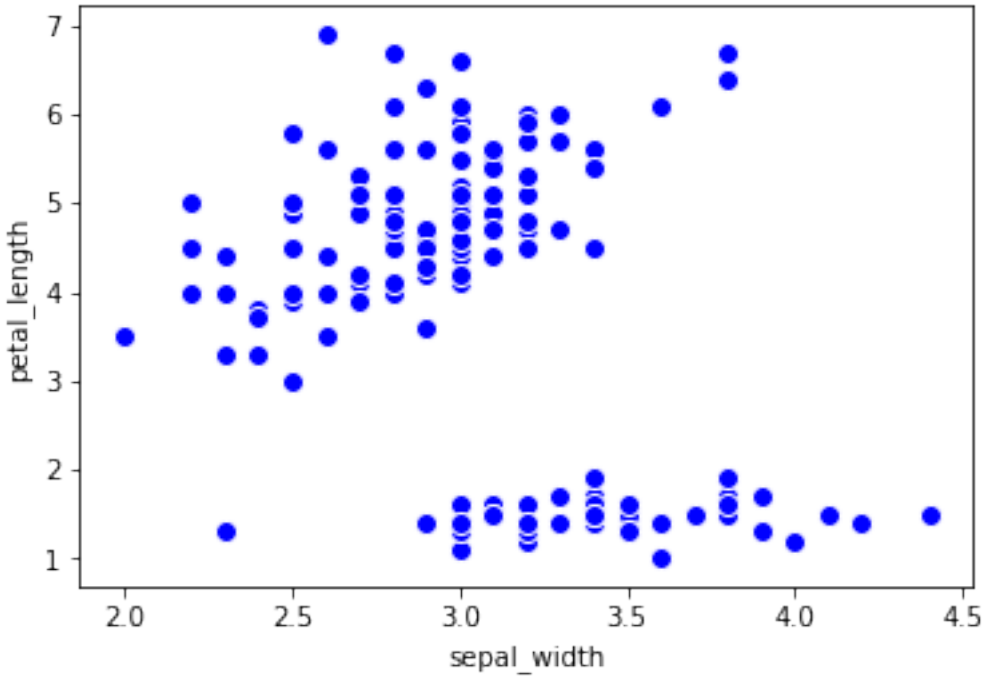
```
In [17]: #scatter plot
         # same data sa in similarity matrix 1
         scatter_plot = iris.plot(kind='scatter', x='petal_width', y = 'petal_length', s = 64,
```



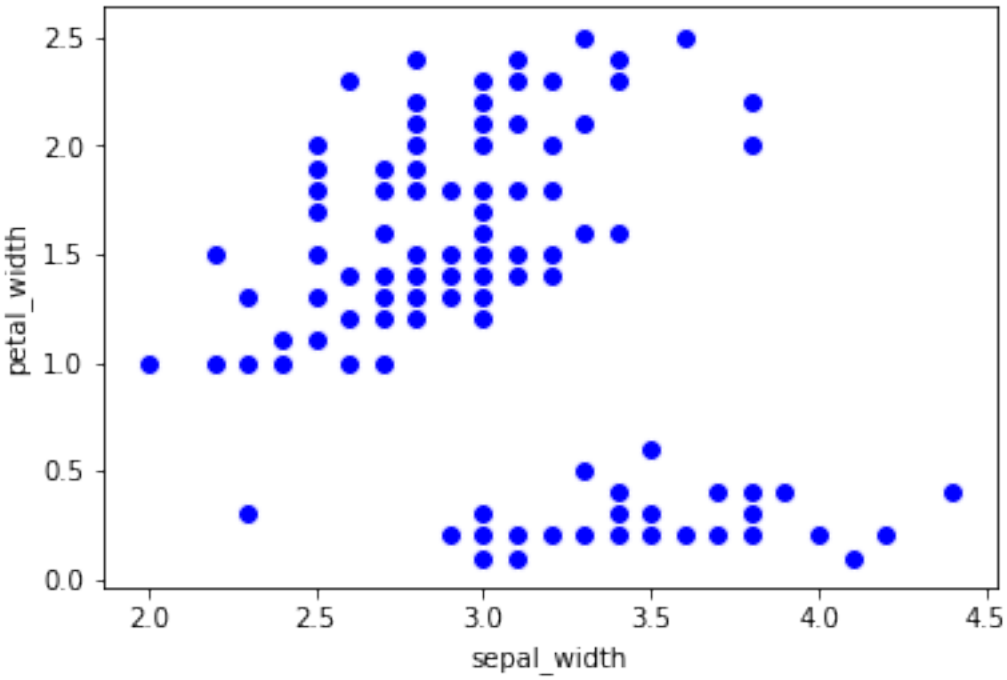

```
In [25]: #scatter plot
         # same data sa in similarity matrix 2
         scatter_plot = iris.plot(kind='scatter', x='sepal_width', y = 'sepal_length', s = 64,
```



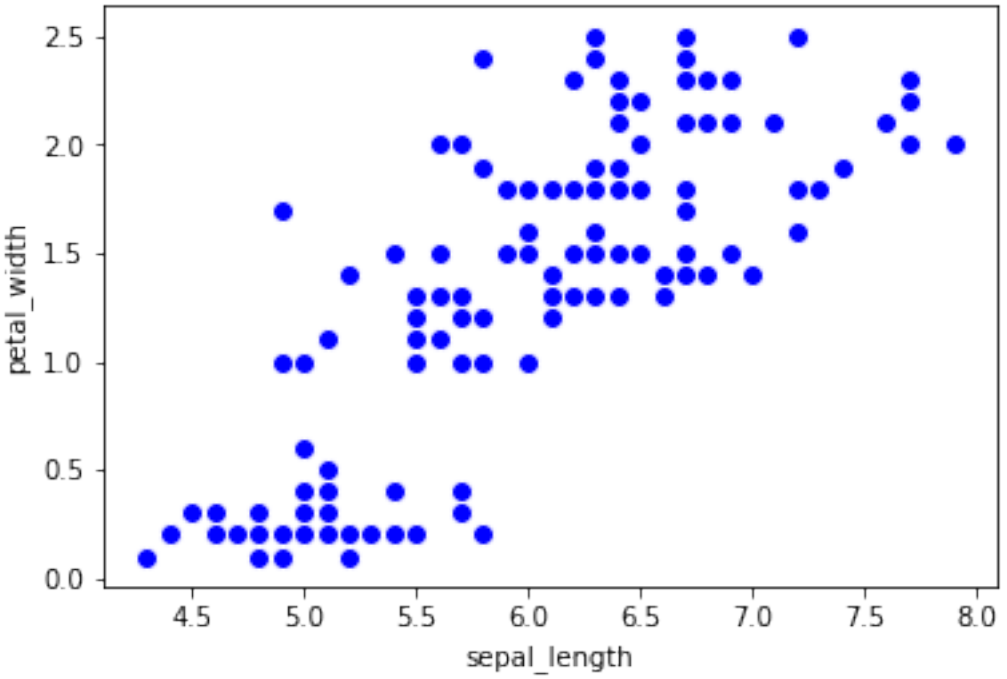
In [26]:
```
#scatter plot
scatter_plot = iris.plot(kind='scatter', x='sepal_width', y = 'petal_length', s = 64,
```



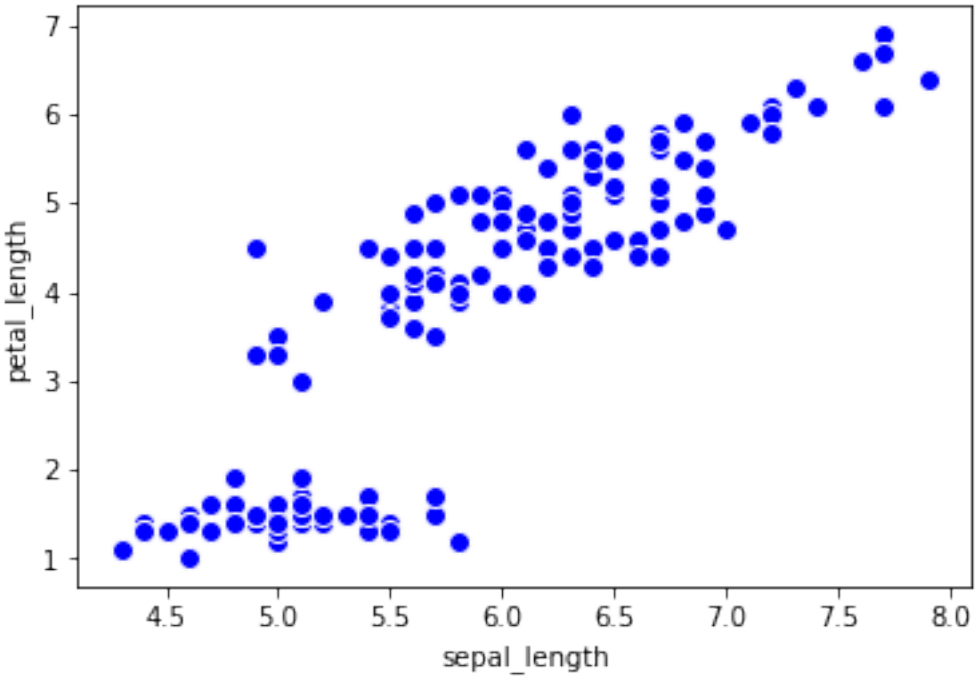
In [27]:
```
#scatter plot
scatter_plot = iris.plot(kind='scatter', x='sepal_width', y = 'petal_width', s = 64, c
```



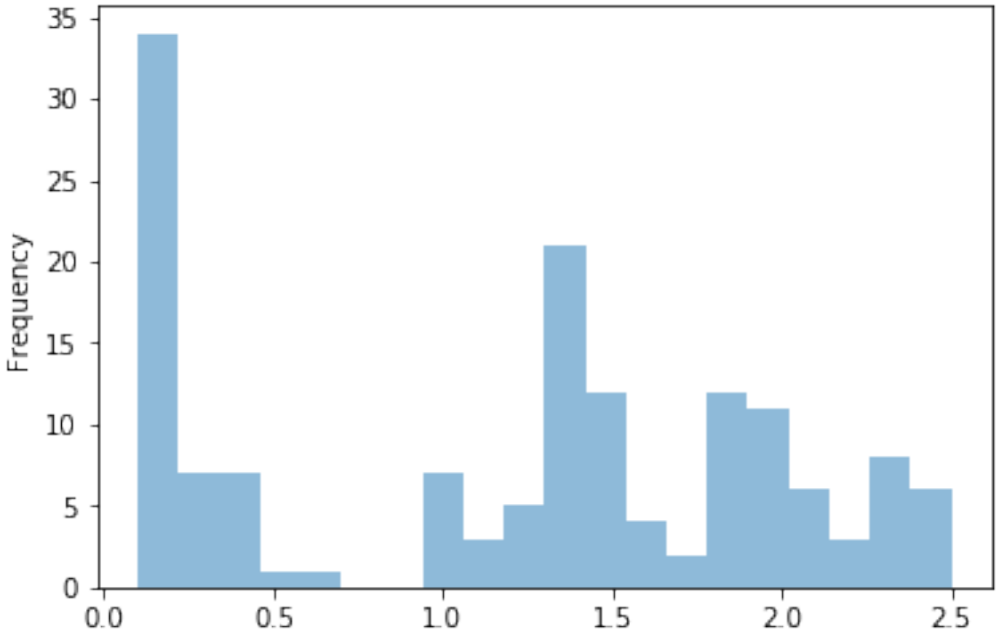
In [28]:
```
#scatter plot
scatter_plot = iris.plot(kind='scatter', x='sepal_length', y = 'petal_width', s = 64,
```



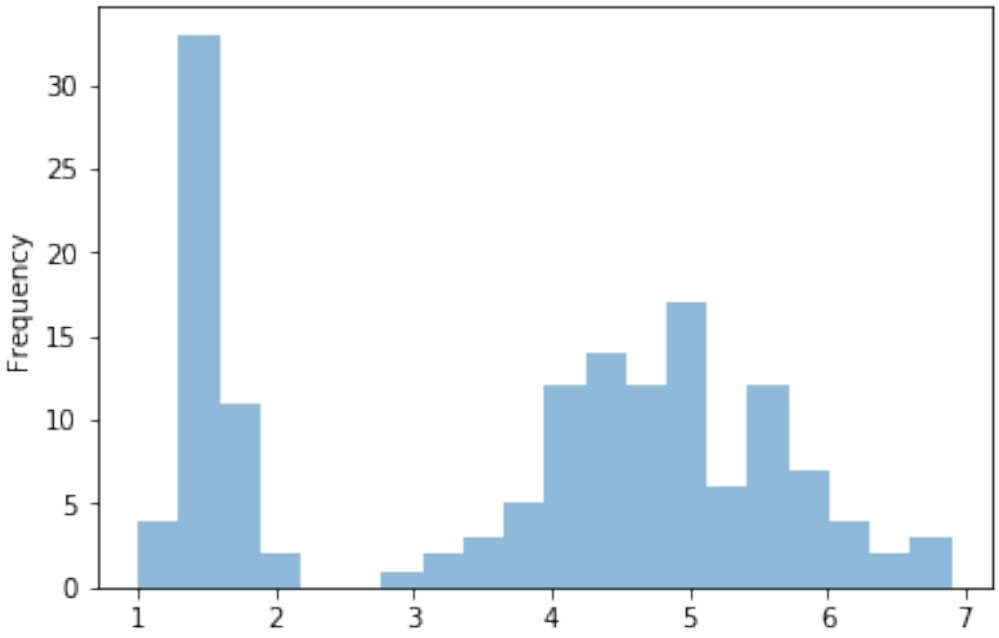
In [31]:
```
#scatter plot
scatter_plot = iris.plot(kind='scatter', x='sepal_length', y = 'petal_length', s = 64
```



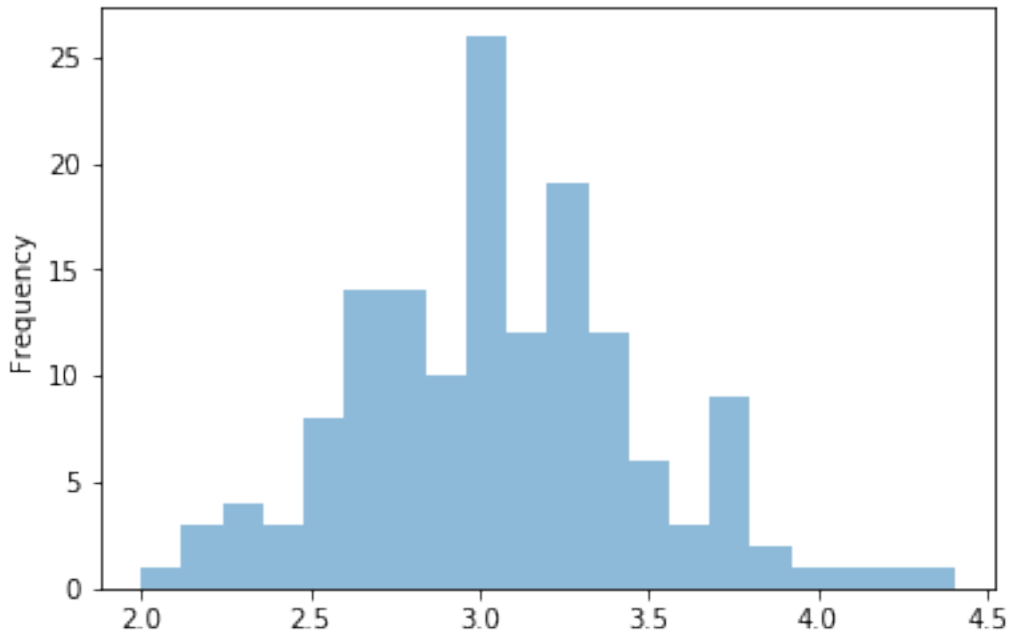
In [19]:
```
# distribution of features 1
distr = iris.petal_width.plot(kind='hist', alpha=0.5, bins=20)
```



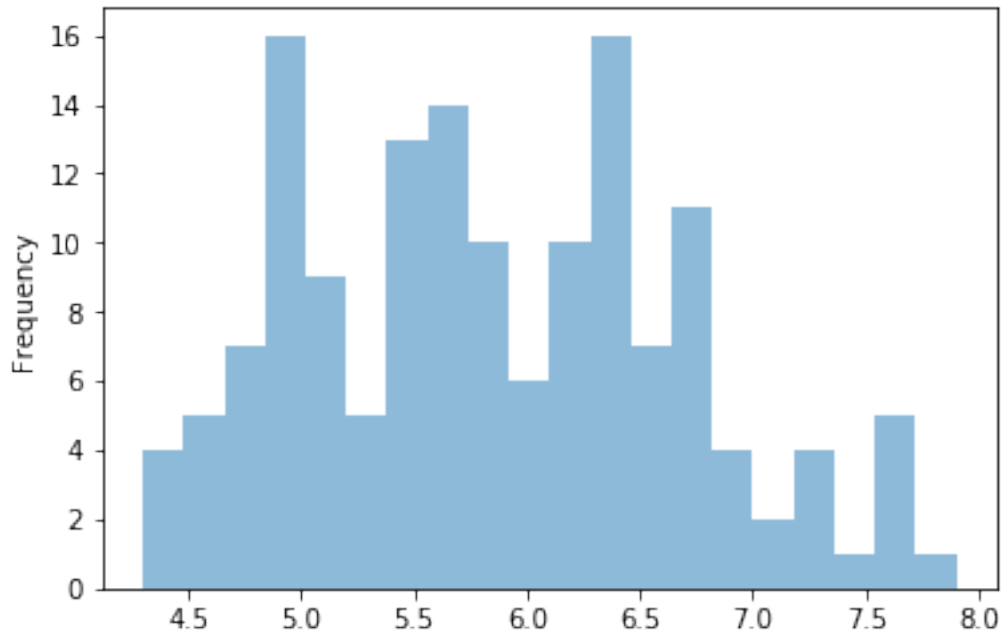

```
In [20]: # distribution of features 2
         distr = iris.petal_length.plot(kind='hist', alpha=0.5, bins=20)
```



In [21]:
```
# distribution of features 3
distr = iris.sepal_width.plot(kind='hist', alpha=0.5, bins=20)
```



In [22]:
```
# distribution of features 4
distr = iris.sepal_length.plot(kind='hist', alpha=0.5, bins=20)
```



In [ ]: