

### **Week 3 Hw**

Script:

```
from sklearn import datasets
```

```
iris = datasets.load_iris()
```

```
import pandas as pd
```

```
data = { "weight": [4.17, 5.58, 5.18, 6.11, 4.50, 4.61, 5.17, 4.53, 5.33, 5.14, 4.81, 4.17, 4.41,  
3.59, 5.87, 3.83, 6.03, 4.89, 4.32, 4.69, 6.31, 5.12, 5.54, 5.50, 5.37, 5.29, 4.92, 6.15, 5.80,  
5.26], "group": ["ctrl"] * 10 + ["trt1"] * 10 + ["trt2"] * 10 }
```

```
PlantGrowth = pd.DataFrame(data)
```

```
# Iris Data Set
```

```
# a. Make a histogram of the variable Sepal.Width.
```

```
df = pd.DataFrame(data=iris.data,  
                  columns=iris.feature_names)
```

```
df.head()
```

```
import matplotlib.pyplot as plt
```

```
plt.hist(df['sepal width (cm)'])
```

```
plt.show()
```

```
import seaborn as sns
```

```
sns.histplot(df['sepal width (cm)'],kde=True)
```

```
plt.show()
```

# b. Based on the histogram from #1a, which would you expect to be higher, the mean or the median? Why?

```
print("I expect the mean to be higher because there appears to be a very slight right skew")
```

# c. Confirm your answer to #1b by actually finding these values.

```
sepal= df["sepal width (cm)"]
```

```
import numpy as np
```

```
np.mean(sepal)
```

```
np.median(sepal)
```

# d. Only 27% of the flowers have a Sepal.Width higher than \_\_\_\_\_ cm.

```
np.percentile(sepal,[73])
```

# Only 27% of the flowers have a Sepal.Width higher than 3.3 cm.

# e. Make scatterplots of each pair of the numerical variables in iris (There should be 6 pairs/plots).

```
sns.pairplot(df)
```

```
plt.show()
```

# f. Based on #1e, which two variables appear to have the strongest relationship? And which two appear to have the weakest relationship?

```
# Petal Length and Petal Width seem to have a strong positive correlation.
```

```
# Sepal Length and Sepal Width Seem to have no correlation at all
```

```
# PlantGrowth DataSet
```

```
#Histogram
```

```
PlantGrowth.head()
```

```
weight=PlantGrowth['weight']
```

```
bins = np.arange(3.3, weight.max() + 0.3, 0.3)
```

```
sns.histplot(data=PlantGrowth, x='weight', bins=bins)
```

```
plt.show()
```

```
#Boxplot
```

```
sns.boxplot(x='group', y='weight', data=PlantGrowth)
```

```
plt.show()
```

```
# Almost all the Trt1 weights are below the TRT2 Minimum
```

```
T2 = PlantGrowth[PlantGrowth['group'] == "trt2"]
```

```
T2.head()
```

```
T2['weight'].min()
```

```
# min trt2 is 4.92
```

```
T1=PlantGrowth[PlantGrowth['group'] == "trt1"]
```

```
T1['weight'].max()
```

```
np.percentile(T1['weight'],[79])
```

```
# about 78 to 79% of the weights in the trt1 group are below the minimum trt 2 weight.
```

```
# Only including plants with a weight above 5.5, make a barplot of the variable group. Make  
the barplot colorful using some color palette
```

```
filtered = PlantGrowth[PlantGrowth['weight'] > 5.5]
```

```
groupcounts = filtered['group'].value_counts().reset_index()
```

```
groupcounts.columns = ['group', 'count']
```

```
sns.barplot(x='group', y='count', data=groupcounts, palette='viridis')
```

```
plt.show()
```

Terminal:

```
PS C:\Users\ssark> from sklearn import datasets
```

```
At line:1 char:1
```

```
+ from sklearn import datasets
```

```
+ ~~~~
```

The 'from' keyword is not supported in this version of the language.

```
+ CategoryInfo      : ParserError: (:) [], ParentContainsErrorRecordException
```

```
+ FullyQualifiedErrorId : ReservedKeywordNotAllowed
```

```
PS C:\Users\ssark> iris = datasets.load_iris()
```

```
At line:1 char:27
```

```
+ iris = datasets.load_iris()
```

```
+ ~
```

An expression was expected after '('.

```
+ CategoryInfo          : ParserError: (:) [], ParentContainsErrorRecordException
```

```
+ FullyQualifiedErrorId : ExpectedExpression
```

```
PS C:\Users\ssark> import pandas as pd
```

import : The term 'import' is not recognized as the name of a cmdlet, function, script file, or operable program. Check the spelling of the name, or if a path was included, verify that the path is correct and try again.

At line:1 char:1

```
+ import pandas as pd
```

```
+ ~~~~~
```

```
+ CategoryInfo          : ObjectNotFound: (import:String) [], CommandNotFoundException
```

```
+ FullyQualifiedErrorId : CommandNotFoundException
```

```
PS C:\Users\ssark> data = { "weight": [4.17, 5.58, 5.18, 6.11, 4.50, 4.61, 5.17, 4.53, 5.33,
5.14, 4.81, 4.17, 4.41, 3.59, 5.87, 3.83, 6.03, 4.89, 4.32, 4.69, 6.31, 5.12, 5.54, 5.50, 5.37,
5.29, 4.92, 6.15, 5.80, 5.26], "group": ["ctrl"] * 10 + ["trt1"] * 10 + ["trt2"] * 10}
```

At line:1 char:5

```
+ data = { "weight": [4.17, 5.58, 5.18, 6.11, 4.50, 4.61, 5.17, 4.53, 5 ...
```

```
+ ~
```

The Data section is missing its statement block.

At line:1 char:18

```
+ data = { "weight": [4.17, 5.58, 5.18, 6.11, 4.50, 4.61, 5.17, 4.53, 5 ...
```

```
+ ~
```

Unexpected token ':' in expression or statement.

```
+ CategoryInfo      : ParserError: (:) [], ParentContainsErrorRecordException
+ FullyQualifiedErrorId : MissingStatementBlockForDataSection
```

```
PS C:\Users\ssark> PlantGrowth = pd.DataFrame(data)
```

data : The term 'data' is not recognized as the name of a cmdlet, function, script file, or operable program. Check the spelling of the name, or if a path was included, verify that the path is correct and try again.

At line:1 char:28

```
+ PlantGrowth = pd.DataFrame(data)
+
+ CategoryInfo      : ObjectNotFound: (data:String) [], CommandNotFoundException
+ FullyQualifiedErrorId : CommandNotFoundException
```

```
PS C:\Users\ssark> &
```

```
C:/Users/ssark/AppData/Local/Programs/Python/Python313/python.exe
```

```
Python 3.13.7 (tags/v3.13.7:bcee1c3, Aug 14 2025, 14:15:11) [MSC v.1944 64 bit (AMD64)]
on win32
```

Type "help", "copyright", "credits" or "license" for more information.

```
>>> from sklearn import datasets
```

```
... iris = datasets.load_iris()
```

```
... import pandas as pd
```

```
... data = { "weight": [4.17, 5.58, 5.18, 6.11, 4.50, 4.61, 5.17, 4.53, 5.33, 5.14, 4.81, 4.17,
4.41, 3.59, 5.87, 3.83, 6.03, 4.89, 4.32, 4.69, 6.31, 5.12, 5.54, 5.50, 5.37, 5.29, 4.92, 6.15,
5.80, 5.26], "group": ["ctrl"] * 10 + ["trt1"] * 10 + ["trt2"] * 10}
```

```
... PlantGrowth = pd.DataFrame(data)
```

```
...
```

Traceback (most recent call last):

File "<python-input-0>", line 1, in <module>

```
from sklearn import datasets
```

```
ModuleNotFoundError: No module named 'sklearn'
```

```
>>> exit()
```

```
PS C:\Users\ssark> pip install sklearn
```

```
Defaulting to user installation because normal site-packages is not writeable
```

```
Collecting sklearn
```

```
Downloading sklearn-0.0.post12.tar.gz (2.6 kB)
```

```
Installing build dependencies ... done
```

```
Getting requirements to build wheel ... error
```

```
error: subprocess-exited-with-error
```

```
× Getting requirements to build wheel did not run successfully.
```

```
| exit code: 1
```

```
└> [15 lines of output]
```

```
The 'sklearn' PyPI package is deprecated, use 'scikit-learn'
rather than 'sklearn' for pip commands.
```

Here is how to fix this error in the main use cases:

- use 'pip install scikit-learn' rather than 'pip install sklearn'
- replace 'sklearn' by 'scikit-learn' in your pip requirements files (requirements.txt, setup.py, setup.cfg, Pipfile, etc ...)
- if the 'sklearn' package is used by one of your dependencies, it would be great if you take some time to track which package uses 'sklearn' instead of 'scikit-learn' and report it to their issue tracker
- as a last resort, set the environment variable

SKLEARN\_ALLOW\_DEPRECATED\_SKLEARN\_PACKAGE\_INSTALL=True to avoid this error

More information is available at

<https://github.com/scikit-learn/sklearn-pypi-package>

[end of output]

note: This error originates from a subprocess, and is likely not a problem with pip.  
error: subprocess-exited-with-error

× Getting requirements to build wheel did not run successfully.

| exit code: 1

↳ See above for output.

note: This error originates from a subprocess, and is likely not a problem with pip.

PS C:\Users\ssark> Python

Python 3.13.7 (tags/v3.13.7:bcee1c3, Aug 14 2025, 14:15:11) [MSC v.1944 64 bit (AMD64)]  
on win32

Type "help", "copyright", "credits" or "license" for more information.

>>> from sklearn import datasets

... iris = datasets.load\_iris()

... import pandas as pd

... data = { "weight": [4.17, 5.58, 5.18, 6.11, 4.50, 4.61, 5.17, 4.53, 5.33, 5.14, 4.81, 4.17,  
4.41, 3.59, 5.87, 3.83, 6.03, 4.89, 4.32, 4.69, 6.31, 5.12, 5.54, 5.50, 5.37, 5.29, 4.92, 6.15,  
5.80, 5.26], "group": ["ctrl"] \* 10 + ["trt1"] \* 10 + ["trt2"] \* 10}

... PlantGrowth = pd.DataFrame(data)

...



Traceback (most recent call last):

File "<python-input-0>", line 1, in <module>

from sklearn import datasets

ModuleNotFoundError: No module named 'sklearn'

>>> exit()

PS C:\Users\ssark> pip install scikit-learn

Defaulting to user installation because normal site-packages is not writeable

Collecting scikit-learn

Downloading scikit\_learn-1.7.2-cp313-cp313-win\_amd64.whl.metadata (11 kB)

Requirement already satisfied: numpy>=1.22.0 in

c:\users\ssark\appdata\local\packages\pythonsoftwarefoundation.python.3.13\_qbz5n2kfr  
a8p0\localcache\local-packages\python313\site-packages (from scikit-learn) (2.3.2)

Collecting scipy>=1.8.0 (from scikit-learn)

Downloading scipy-1.16.2-cp313-cp313-win\_amd64.whl.metadata (60 kB)

Collecting joblib>=1.2.0 (from scikit-learn)

Downloading joblib-1.5.2-py3-none-any.whl.metadata (5.6 kB)

Collecting threadpoolctl>=3.1.0 (from scikit-learn)

Downloading threadpoolctl-3.6.0-py3-none-any.whl.metadata (13 kB)

Downloading scikit\_learn-1.7.2-cp313-cp313-win\_amd64.whl (8.7 MB)

---

—— 8.7/8.7 MB 27.5 MB/s 0:00:00

Downloading joblib-1.5.2-py3-none-any.whl (308 kB)

Downloading scipy-1.16.2-cp313-cp313-win\_amd64.whl (38.5 MB)

---

—— 38.5/38.5 MB 38.5 MB/s 0:00:00

Downloading threadpoolctl-3.6.0-py3-none-any.whl (18 kB)

Installing collected packages: threadpoolctl, scipy, joblib, scikit-learn

Successfully installed joblib-1.5.2 scikit-learn-1.7.2 scipy-1.16.2 threadpoolctl-3.6.0

PS C:\Users\ssark> Python

Python 3.13.7 (tags/v3.13.7:bcee1c3, Aug 14 2025, 14:15:11) [MSC v.1944 64 bit (AMD64)]  
on win32

Type "help", "copyright", "credits" or "license" for more information.

```
>>> from sklearn import datasets
```

```
... iris = datasets.load_iris()
```

```
... import pandas as pd
```

```
... data = {"weight": [4.17, 5.58, 5.18, 6.11, 4.50, 4.61, 5.17, 4.53, 5.33, 5.14, 4.81, 4.17,  
4.41, 3.59, 5.87, 3.83, 6.03, 4.89, 4.32, 4.69, 6.31, 5.12, 5.54, 5.50, 5.37, 5.29, 4.92, 6.15,  
5.80, 5.26], "group": ["ctrl"] * 10 + ["trt1"] * 10 + ["trt2"] * 10}
```

```
... PlantGrowth = pd.DataFrame(data)
```

```
...
```

```
>>>
```

```
>>>
```

```
>>> iris.head()
```

Traceback (most recent call last):

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\sklearn\utils\\_bunch.py",  
line 57, in \_\_getattr\_\_

```
    return self[key]
```

```
~~~~~^^^^^
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\sklearn\utils\\_bunch.py",  
line 42, in \_\_getitem\_\_

```
    return super().__getitem__(key)
```

```
~~~~~^
```

KeyError: 'head'

During handling of the above exception, another exception occurred:

Traceback (most recent call last):

File "<python-input-3>", line 1, in <module>

iris.head()

^^^^^^^^

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\sklearn\utils\\_bunch.py", line 59, in \_\_getattr\_\_

raise AttributeError(key)

AttributeError: head

>>> iris

{'data': array([[5.1, 3.5, 1.4, 0.2],

[4.9, 3. , 1.4, 0.2],

[4.7, 3.2, 1.3, 0.2],

[4.6, 3.1, 1.5, 0.2],

[5. , 3.6, 1.4, 0.2],

[5.4, 3.9, 1.7, 0.4],

[4.6, 3.4, 1.4, 0.3],

[5. , 3.4, 1.5, 0.2],

[4.4, 2.9, 1.4, 0.2],

[4.9, 3.1, 1.5, 0.1],

[5.4, 3.7, 1.5, 0.2],

[4.8, 3.4, 1.6, 0.2],

[4.8, 3. , 1.4, 0.1],  
[4.3, 3. , 1.1, 0.1],  
[5.8, 4. , 1.2, 0.2],  
[5.7, 4.4, 1.5, 0.4],  
[5.4, 3.9, 1.3, 0.4],  
[5.1, 3.5, 1.4, 0.3],  
[5.7, 3.8, 1.7, 0.3],  
[5.1, 3.8, 1.5, 0.3],  
[5.4, 3.4, 1.7, 0.2],  
[5.1, 3.7, 1.5, 0.4],  
[4.6, 3.6, 1. , 0.2],  
[5.1, 3.3, 1.7, 0.5],  
[4.8, 3.4, 1.9, 0.2],  
[5. , 3. , 1.6, 0.2],  
[5. , 3.4, 1.6, 0.4],  
[5.2, 3.5, 1.5, 0.2],  
[5.2, 3.4, 1.4, 0.2],  
[4.7, 3.2, 1.6, 0.2],  
[4.8, 3.1, 1.6, 0.2],  
[5.4, 3.4, 1.5, 0.4],  
[5.2, 4.1, 1.5, 0.1],  
[5.5, 4.2, 1.4, 0.2],  
[4.9, 3.1, 1.5, 0.2],  
[5. , 3.2, 1.2, 0.2],  
[5.5, 3.5, 1.3, 0.2],  
[4.9, 3.6, 1.4, 0.1],

[4.4, 3. , 1.3, 0.2],  
[5.1, 3.4, 1.5, 0.2],  
[5. , 3.5, 1.3, 0.3],  
[4.5, 2.3, 1.3, 0.3],  
[4.4, 3.2, 1.3, 0.2],  
[5. , 3.5, 1.6, 0.6],  
[5.1, 3.8, 1.9, 0.4],  
[4.8, 3. , 1.4, 0.3],  
[5.1, 3.8, 1.6, 0.2],  
[4.6, 3.2, 1.4, 0.2],  
[5.3, 3.7, 1.5, 0.2],  
[5. , 3.3, 1.4, 0.2],  
[7. , 3.2, 4.7, 1.4],  
[6.4, 3.2, 4.5, 1.5],  
[6.9, 3.1, 4.9, 1.5],  
[5.5, 2.3, 4. , 1.3],  
[6.5, 2.8, 4.6, 1.5],  
[5.7, 2.8, 4.5, 1.3],  
[6.3, 3.3, 4.7, 1.6],  
[4.9, 2.4, 3.3, 1. ],  
[6.6, 2.9, 4.6, 1.3],  
[5.2, 2.7, 3.9, 1.4],  
[5. , 2. , 3.5, 1. ],  
[5.9, 3. , 4.2, 1.5],  
[6. , 2.2, 4. , 1. ],  
[6.1, 2.9, 4.7, 1.4],

[5.6, 2.9, 3.6, 1.3],  
[6.7, 3.1, 4.4, 1.4],  
[5.6, 3. , 4.5, 1.5],  
[5.8, 2.7, 4.1, 1. ],  
[6.2, 2.2, 4.5, 1.5],  
[5.6, 2.5, 3.9, 1.1],  
[5.9, 3.2, 4.8, 1.8],  
[6.1, 2.8, 4. , 1.3],  
[6.3, 2.5, 4.9, 1.5],  
[6.1, 2.8, 4.7, 1.2],  
[6.4, 2.9, 4.3, 1.3],  
[6.6, 3. , 4.4, 1.4],  
[6.8, 2.8, 4.8, 1.4],  
[6.7, 3. , 5. , 1.7],  
[6. , 2.9, 4.5, 1.5],  
[5.7, 2.6, 3.5, 1. ],  
[5.5, 2.4, 3.8, 1.1],  
[5.5, 2.4, 3.7, 1. ],  
[5.8, 2.7, 3.9, 1.2],  
[6. , 2.7, 5.1, 1.6],  
[5.4, 3. , 4.5, 1.5],  
[6. , 3.4, 4.5, 1.6],  
[6.7, 3.1, 4.7, 1.5],  
[6.3, 2.3, 4.4, 1.3],  
[5.6, 3. , 4.1, 1.3],  
[5.5, 2.5, 4. , 1.3],

[5.5, 2.6, 4.4, 1.2],  
[6.1, 3. , 4.6, 1.4],  
[5.8, 2.6, 4. , 1.2],  
[5. , 2.3, 3.3, 1. ],  
[5.6, 2.7, 4.2, 1.3],  
[5.7, 3. , 4.2, 1.2],  
[5.7, 2.9, 4.2, 1.3],  
[6.2, 2.9, 4.3, 1.3],  
[5.1, 2.5, 3. , 1.1],  
[5.7, 2.8, 4.1, 1.3],  
[6.3, 3.3, 6. , 2.5],  
[5.8, 2.7, 5.1, 1.9],  
[7.1, 3. , 5.9, 2.1],  
[6.3, 2.9, 5.6, 1.8],  
[6.5, 3. , 5.8, 2.2],  
[7.6, 3. , 6.6, 2.1],  
[4.9, 2.5, 4.5, 1.7],  
[7.3, 2.9, 6.3, 1.8],  
[6.7, 2.5, 5.8, 1.8],  
[7.2, 3.6, 6.1, 2.5],  
[6.5, 3.2, 5.1, 2. ],  
[6.4, 2.7, 5.3, 1.9],  
[6.8, 3. , 5.5, 2.1],  
[5.7, 2.5, 5. , 2. ],  
[5.8, 2.8, 5.1, 2.4],  
[6.4, 3.2, 5.3, 2.3],

[6.5, 3. , 5.5, 1.8],  
[7.7, 3.8, 6.7, 2.2],  
[7.7, 2.6, 6.9, 2.3],  
[6. , 2.2, 5. , 1.5],  
[6.9, 3.2, 5.7, 2.3],  
[5.6, 2.8, 4.9, 2. ],  
[7.7, 2.8, 6.7, 2. ],  
[6.3, 2.7, 4.9, 1.8],  
[6.7, 3.3, 5.7, 2.1],  
[7.2, 3.2, 6. , 1.8],  
[6.2, 2.8, 4.8, 1.8],  
[6.1, 3. , 4.9, 1.8],  
[6.4, 2.8, 5.6, 2.1],  
[7.2, 3. , 5.8, 1.6],  
[7.4, 2.8, 6.1, 1.9],  
[7.9, 3.8, 6.4, 2. ],  
[6.4, 2.8, 5.6, 2.2],  
[6.3, 2.8, 5.1, 1.5],  
[6.1, 2.6, 5.6, 1.4],  
[7.7, 3. , 6.1, 2.3],  
[6.3, 3.4, 5.6, 2.4],  
[6.4, 3.1, 5.5, 1.8],  
[6. , 3. , 4.8, 1.8],  
[6.9, 3.1, 5.4, 2.1],  
[6.7, 3.1, 5.6, 2.4],  
[6.9, 3.1, 5.1, 2.3],



```
[5.8, 2.7, 5.1, 1.9],  
[6.8, 3.2, 5.9, 2.3],  
[6.7, 3.3, 5.7, 2.5],  
[6.7, 3. , 5.2, 2.3],  
[6.3, 2.5, 5. , 1.9],  
[6.5, 3. , 5.2, 2. ],  
[6.2, 3.4, 5.4, 2.3],  
[5.9, 3. , 5.1, 1.8]]), 'target': array([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, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,  
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,  
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,  
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,  
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]), 'frame': None, 'target_names':  
array(['setosa', 'versicolor', 'virginica'], dtype='<U10'), 'DESCR': '._iris_dataset:\n\nIris  
plants dataset\n-----\n\n**Data Set Characteristics:**\n\nNumber of  
Instances: 150 (50 in each of three classes)\nNumber of Attributes: 4 numeric, predictive  
attributes and the class\nAttribute Information:\n - sepal length in cm\n - sepal width in  
cm\n - petal length in cm\n - petal width in cm\n - class:\n      - Iris-Setosa\n      - Iris-Versicolour\n      - Iris-Virginica\nSummary Statistics:\n\n=====   
=====  
Min Max Mean SD Class  
Correlation\n===== \nsepal  
length: 4.3 7.9 5.84 0.83 0.7826\nsepal width: 2.0 4.4 3.05 0.43 -0.4194\npetal  
length: 1.0 6.9 3.76 1.76 0.9490 (high!)\npetal width: 0.1 2.5 1.20 0.76 0.9565  
(high!)\n===== \nMissing  
Attribute Values: None\nClass Distribution: 33.3% for each of 3 classes.\nCreator: R.A.  
Fisher\nDonor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)\nDate: July,  
1988\n\nThe famous Iris database, first used by Sir R.A. Fisher. The dataset is taken\nfrom  
Fisher's paper. Note that it's the same as in R, but not as in the UCI\nMachine Learning  
Repository, which has two wrong data points.\n\nThis is perhaps the best known database  
to be found in the\npattern recognition literature. Fisher's paper is a classic in the field  
and\nis referenced frequently to this day. (See Duda & Hart, for example.) The\ndata set
```

contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

dropdown:: References

- Fisher, R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950).
- Duda, R.O., & Hart, P.E. (1973) Pattern Classification and Scene Analysis. (Q327.D83) John Wiley & Sons. ISBN 0-471-22361-1. See page 218.
- Dasarathy, B.V. (1980) "Nosing Around the Neighborhood: A New System Structure and Classification Rule for Recognition in Partially Exposed Environments". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No. 1, 67-71.
- Gates, G.W. (1972) "The Reduced Nearest Neighbor Rule". IEEE Transactions on Information Theory, May 1972, 431-433.
- See also: 1988 MLC Proceedings, 54-64. Cheeseman et al's AUTOCLASS II conceptual clustering system finds 3 classes in the data.
- Many, many more ...

'feature\_names': ['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)'], 'filename': 'iris.csv', 'data\_module': 'sklearn.datasets.data'}

```
>>> df = pd.DataFrame(data=iris_data.data,
...                    columns=iris_data.feature_names)
```

...

Traceback (most recent call last):

File "<python-input-5>", line 1, in <module>

```
df = pd.DataFrame(data=iris_data.data,
```

```
^^^^^^^^^^
```

NameError: name 'iris\_data' is not defined

```
>>> df = pd.DataFrame(data=data.data,
...                    columns=data.feature_names)
```

...

Traceback (most recent call last):

File "<python-input-6>", line 1, in <module>

```
df = pd.DataFrame(data=data.data,
```

```
^^^^^^^^^^
```

AttributeError: 'dict' object has no attribute 'data'

```
>>> df = pd.DataFrame(data=iris.data,  
...                     columns=iris.feature_names)
```

```
...
```

```
>>> df.head()
```

|   | sepal length (cm) | sepal width (cm) | petal length (cm) | petal width (cm) |
|---|-------------------|------------------|-------------------|------------------|
| 0 | 5.1               | 3.5              | 1.4               | 0.2              |
| 1 | 4.9               | 3.0              | 1.4               | 0.2              |
| 2 | 4.7               | 3.2              | 1.3               | 0.2              |
| 3 | 4.6               | 3.1              | 1.5               | 0.2              |
| 4 | 5.0               | 3.6              | 1.4               | 0.2              |

```
>>> import matplotlib.pyplot as plt
```

```
>>> plt.hist[iris['Sepal.Width']]
```

Traceback (most recent call last):

File "<python-input-10>", line 1, in <module>

```
plt.hist[iris['Sepal.Width']]
```

```
~~~~^^^^^^^^^^^^^^^^^^^^
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\sklearn\utils\\_bunch.py",  
line 42, in \_\_getitem\_\_

```
return super().__getitem__(key)
```

```
~~~~~^
```

KeyError: 'Sepal.Width'

```
>>> plt.hist[iris['sepal width']]
```

Traceback (most recent call last):

File "<python-input-11>", line 1, in <module>

```
plt.hist[iris['sepal width']]
```

```
~~~~^^^^^^^^^^^^^^^^^^
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\sklearn\utils\\_bunch.py", line 42, in \_\_getitem\_\_

```
return super().__getitem__(key)
```

```
~~~~~^
```

KeyError: 'sepal width'

```
>>> plt.hist[iris['sepal width (cm)']]
```

Traceback (most recent call last):

File "<python-input-12>", line 1, in <module>

```
plt.hist[iris['sepal width (cm)']]
```

```
~~~~~^
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\sklearn\utils\\_bunch.py", line 42, in \_\_getitem\_\_

```
return super().__getitem__(key)
```

```
~~~~~^
```

KeyError: 'sepal width (cm)'

```
>>> plt.hist(df['sepal width (cm)'])
```

```
(array([ 4.,  7., 22., 24., 37., 31., 10., 11.,  2.,  2.]), array([2. , 2.24, 2.48, 2.72, 2.96, 3.2 , 3.44, 3.68, 3.92, 4.16, 4.4 ]), <BarContainer object of 10 artists>)
```

```
>>> plt.show
```

```
<function show at 0x00000292F5E63740>
```

```
>>> plt.show()
```

```
>>> import seaborn as sns
```

```
>>> sns.histplot(df['sepal width (cm)'],kde=True)
<Axes: xlabel='sepal width (cm)', ylabel='Count'>
>>> plt.show()
>>> plt.show()
>>> sns.histplot(df['sepal width (cm)'],kde=True)
... plt.show()
...
>>> print("I expect the mean to be higher because there appears to be a very slight right
skew")
```

I expect the mean to be higher because there appears to be a very slight right skew

```
>>> sepal= df["sepal width (cm)"]
... mean(sepal)
... median(sepal)
...
```

Traceback (most recent call last):

```
File "<python-input-22>", line 2, in <module>
    mean(sepal)
    ^^^^
```

NameError: name 'mean' is not defined

```
>>> np.mean(sepal)
... np.median(sepal)
...
```

Traceback (most recent call last):

```
File "<python-input-23>", line 1, in <module>
    np.mean(sepal)
    ^^
```

NameError: name 'np' is not defined

```
>>> import numpy as np
```

```
... np.mean(sepal)
```

```
... np.median(sepal)
```

```
...
```

```
np.float64(3.0)
```

```
>>> np.mean(sepal)
```

```
np.float64(3.0573333333333337)
```

```
>>> np.median(sepal)
```

```
np.float64(3.0)
```

```
>>> np.percentile(sepal,[73])
```

Traceback (most recent call last):

File "<python-input-27>", line 1, in <module>

```
    np.percentile(sepal,[73])
```

```
~~~~~^
```

TypeError: 'numpy.\_ArrayFunctionDispatcher' object is not subscriptable

```
>>> np.percentile(sepal,[73])
```

```
array([3.3])
```

```
>>> sns.pairplot(df)
```

```
... plt.show()
```

```
...
```

```
>>>
```

```
>>>
```

```
>>> PlantGrowth.head()
```

```
weight group
```

```
0  4.17  ctrl
```

```

1 5.58 ctrl
2 5.18 ctrl
3 6.11 ctrl
4 4.50 ctrl

>>> weight=PlantGrowth['weight']
>>> PlantGrowth.head()

... weight=PlantGrowth['weight']
... bins = np.arange(3.3, weight.max() + 0.3, 0.3)
... sns.histplot(data=PlantGrowth, x='weight', bins=bins)
... plt.show()

...

>>> sns.boxplot(x='group', y='weight', data=PlantGrowth)
... plt.show()

...

>>> sns.boxplot(x='group', y='weight', data=PlantGrowth)
... plt.show()

...

>>> T2= PlantGrowth[PlantGrowth['group']=="trt2"]

```

Traceback (most recent call last):

```

File
"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13_qbz5n
2kfra8p0\LocalCache\local-packages\Python313\site-
packages\pandas\core\indexes\base.py", line 3812, in get_loc

```

```

    return self._engine.get_loc(casted_key)

```

```

~~~~~^

```

```

File "pandas/_libs/index.pyx", line 167, in pandas._libs.index.IndexEngine.get_loc

```

```

File "pandas/_libs/index.pyx", line 196, in pandas._libs.index.IndexEngine.get_loc

```

File "pandas/\_libs/hashtable\_class\_helper.pxi", line 7088, in  
pandas.\_libs.hashtable.PyObjectHashTable.get\_item

File "pandas/\_libs/hashtable\_class\_helper.pxi", line 7096, in  
pandas.\_libs.hashtable.PyObjectHashTable.get\_item

KeyError: False

The above exception was the direct cause of the following exception:

Traceback (most recent call last):

File "<python-input-37>", line 1, in <module>

T2= PlantGrowth[PlantGrowth['group']=="trt2"]

~~~~~^

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n  
2kfra8p0\LocalCache\local-packages\Python313\site-packages\pandas\core\frame.py",  
line 4107, in \_\_getitem\_\_

indexer = self.columns.get\_loc(key)

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n  
2kfra8p0\LocalCache\local-packages\Python313\site-  
packages\pandas\core\indexes\base.py", line 3819, in get\_loc

raise KeyError(key) from err

KeyError: False

>>> T2 = PlantGrowth[PlantGrowth['group'] == "trt2"]

>>> np.min(T2)

Traceback (most recent call last):

File "<python-input-39>", line 1, in <module>

np.min(T2)



~~~~~^

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\numpy\\_core\fromnumeric.py", line 3302, in min

```
    return _wrapreduction(a, np.minimum, 'min', axis, None, out,
                           keepdims=keepdims, initial=initial, where=where)
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\numpy\\_core\fromnumeric.py", line 84, in \_wrapreduction

```
    return reduction(axis=axis, out=out, **passkwargs)
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\pandas\core\frame.py", line 11650, in min

```
    result = super().min(axis, skipna, numeric_only, **kwargs)
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\pandas\core\generic.py", line 12407, in min

```
    return self._stat_function(
```

~~~~~^

```
    "min",
```

~~~~~

...<4 lines>...

```
    **kwargs,
```

~~~~~

```
)
```

^

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\pandas\core\generic.py",  
line 12396, in \_stat\_function

```
    return self._reduce(  
        ~~~~~~^  
        func, name=name, axis=axis, skipna=skipna, numeric_only=numeric_only  
        ^^^  
    )  
    ^
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\pandas\core\frame.py",  
line 11509, in \_reduce

```
    return func(df.values)
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\pandas\core\frame.py",  
line 11461, in func

```
    return op(values, axis=axis, skipna=skipna, **kwds)
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\pandas\core\nanops.py",  
line 147, in f

```
    result = alt(values, axis=axis, skipna=skipna, **kwds)
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\pandas\core\nanops.py",  
line 404, in new\_func

```
    result = func(values, axis=axis, skipna=skipna, mask=mask, **kwargs)
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\pandas\core\nanops.py", line 1098, in reduction

```
result = getattr(values, meth)(axis)
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\numpy\\_core\\_methods.py", line 47, in \_amin

```
return umr_minimum(a, axis, None, out, keepdims, initial, where)
```

TypeError: '<=' not supported between instances of 'float' and 'str'

```
>>> T2.head()
```

```
weight group
```

```
20  6.31 trt2
```

```
21  5.12 trt2
```

```
22  5.54 trt2
```

```
23  5.50 trt2
```

```
24  5.37 trt2
```

```
>>> T2['weight'].min()
```

```
np.float64(4.92)
```

```
>>> T1=PlantGrowth[PlantGrowth['group'] == "trt1"]
```

```
... T1['weight'].max()
```

```
...
```

```
np.float64(6.03)
```

```
>>> np.percentile(T1,[75])
```

Traceback (most recent call last):

File "<python-input-43>", line 1, in <module>

```
np.percentile(T1,[75])
```

~~~~~^ ^ ^ ^ ^ ^ ^ ^ ^ ^

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\numpy\lib\\_function\_base\_impl.py", line 4292, in percentile

```
    return _quantile_unchecked(
        a, q, axis, out, overwrite_input, method, keepdims, weights)
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\numpy\lib\\_function\_base\_impl.py", line 4569, in \_quantile\_unchecked

```
    return _ureduce(a,
        func=_quantile_ureduce_func,
        ...<5 lines>...
        overwrite_input=overwrite_input,
        method=method)
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\numpy\lib\\_function\_base\_impl.py", line 3914, in \_ureduce

```
    r = func(a, **kwargs)
```

File

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\numpy\lib\\_function\_base\_impl.py", line 4744, in \_quantile\_ureduce\_func

```
    result = _quantile(arr,
        quantiles=q,
        ...<2 lines>...
        out=out,
        weights=wgt)
```

"C:\Users\ssark\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.13\_qbz5n2kfra8p0\LocalCache\local-packages\Python313\site-packages\numpy\lib\\_function\_base\_impl.py", line 4859, in \_quantile

TypeError: '<' not supported between instances of 'str' and 'float'

```
>>> np.percentile(T1['weight'],[75])
array([4.87])
>>> np.percentile(T1['weight'],[76])
array([4.8772])
>>> np.percentile(T1['weight'],[80])
array([5.086])
>>> np.percentile(T1['weight'],[77])
array([4.8844])
>>> np.percentile(T1['weight'],[78])
array([4.9096])
>>> np.percentile(T1['weight'],[79])
array([4.9978])
>>> filtered = PlantGrowth[PlantGrowth['weight'] > 5.5]
```

```
... groupcounts = filtered['group'].value_counts().reset_index()
... groupcounts.columns = ['group', 'count']
... sns.barplot(x='group', y='count', data=groupcounts, palette='viridis')
... plt.show()
...
```

<python-input-50>:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0.  
Assign the `x` variable to `hue` and set `legend=False` for the same effect.