Live 2025-10-20a

October 22, 2025

1 Clustering on Iris Data

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
X = sns.load_dataset('iris')
```

[3]:
```
y = X.species
X = X.drop('species',axis=1)
```

[4]:
```
X.head()
```

[4]:

	sepal_length	sepal_width	petal_length	petal_width
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

[5]:
```
y.head()
```

[5]:
```
0    setosa
1    setosa
2    setosa
3    setosa
4    setosa
Name: species, dtype: object
```

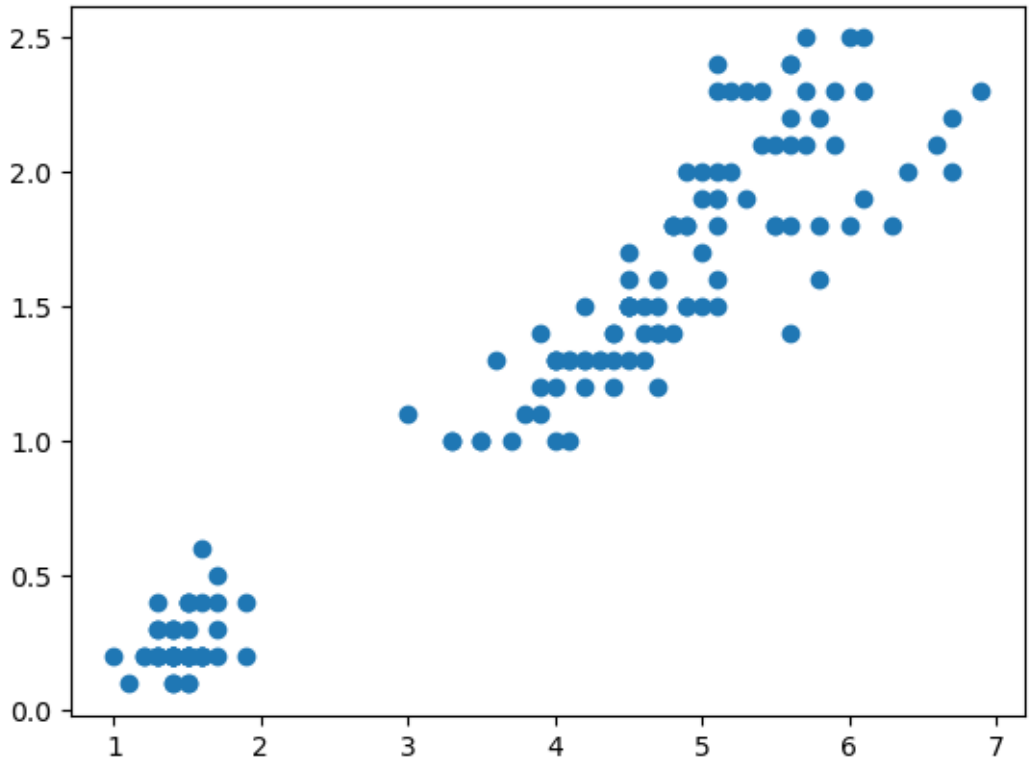
X is now the matrix with the feature values

y is a series with the target value for each set of features

Let's start with unsupervised learning and not use y

[6]:
```
plt.scatter(X.petal_length,X.petal_width)
```

[6]:
```
<matplotlib.collections.PathCollection at 0x7f9859acdd20>
```



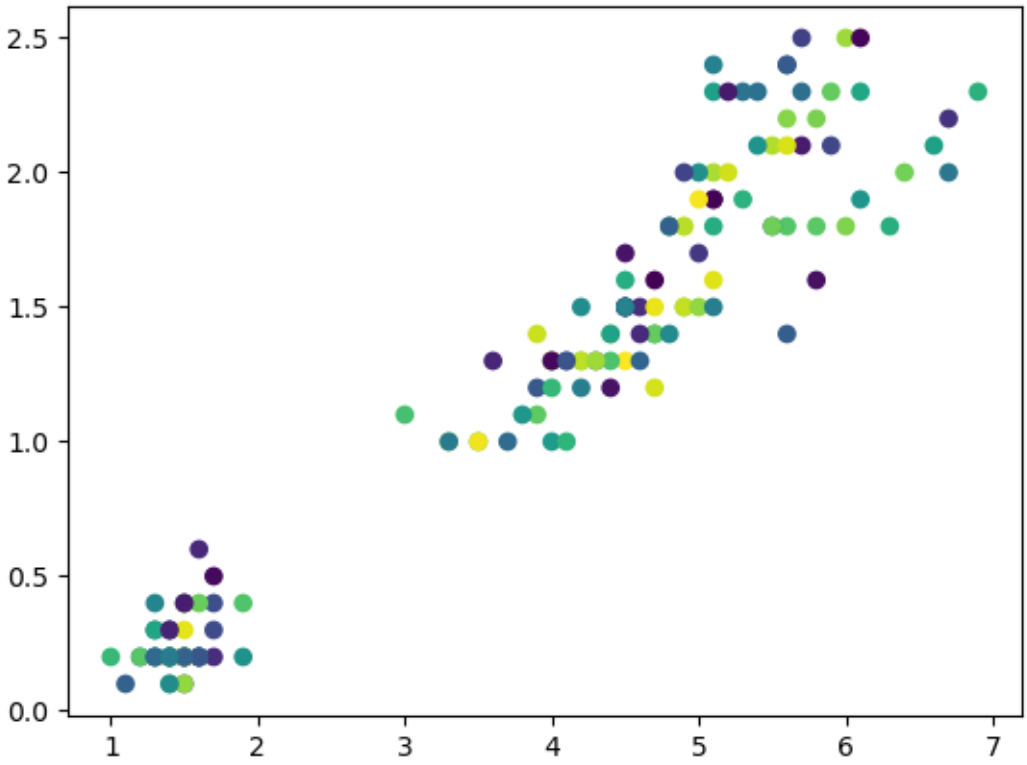
Let's find some clusters with sklearn

[16]:
```
from sklearn.cluster import KMeans

clusterer = KMeans(
    n_clusters=150,
    init='random',
    random_state=42)

clusterer.fit(X)

y_pred = pd.Series(clusterer.predict(X))
```

```
/opt/conda/lib/python3.10/site-packages/sklearn/base.py:1365:
ConvergenceWarning: Number of distinct clusters (149) found smaller than
n_clusters (150). Possibly due to duplicate points in X.
  return fit_method(estimator, *args, **kwargs)
```

[17]:
```
plt.scatter(X.petal_length,X.petal_width,c=y_pred)
```

[17]:
```
<matplotlib.collections.PathCollection at 0x7f9837266860>
```


2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7

[ ]: