assignment 19 (kmeans)

March 8, 2022

1 Creating a cluster model on Penguins dataset

[7]:
```
import pandas as pd
import seaborn as sns
from sklearn.cluster import KMeans
```

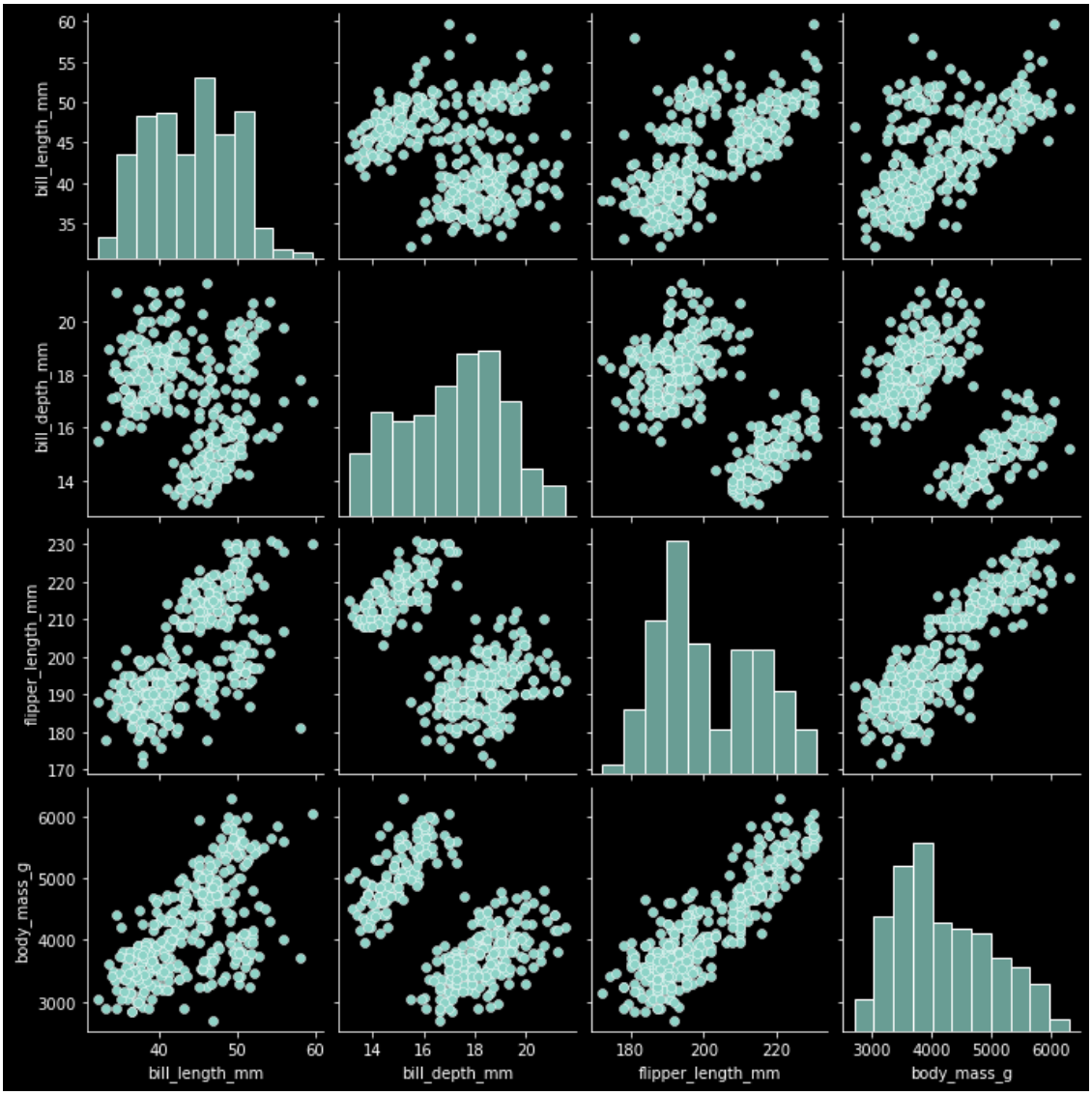
[8]:
```
penguins = sns.load_dataset("penguins")
penguins.head()
```

[8]:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	Male
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	Female
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	Female
3	Adelie	Torgersen	NaN	NaN	NaN	NaN	NaN
4	Adelie	Torgersen	36.7	19.3	193.0	3450.0	Female

[9]:
```
sns.pairplot(penguins)
```

[9]:
```
<seaborn.axisgrid.PairGrid at 0x1abf8848fa0>
```



I see 2 to maybe 3 clusters

[35]:
```
penguins = penguins.dropna()
features = ['body_mass_g','bill_depth_mm'] # , 'flipper_length_mm'
km = KMeans(n_clusters=2, random_state=42).fit(penguins[features])
```

Features do not really influence the silhouette score, upping the amount of clusters to 3 significantly decreases the score

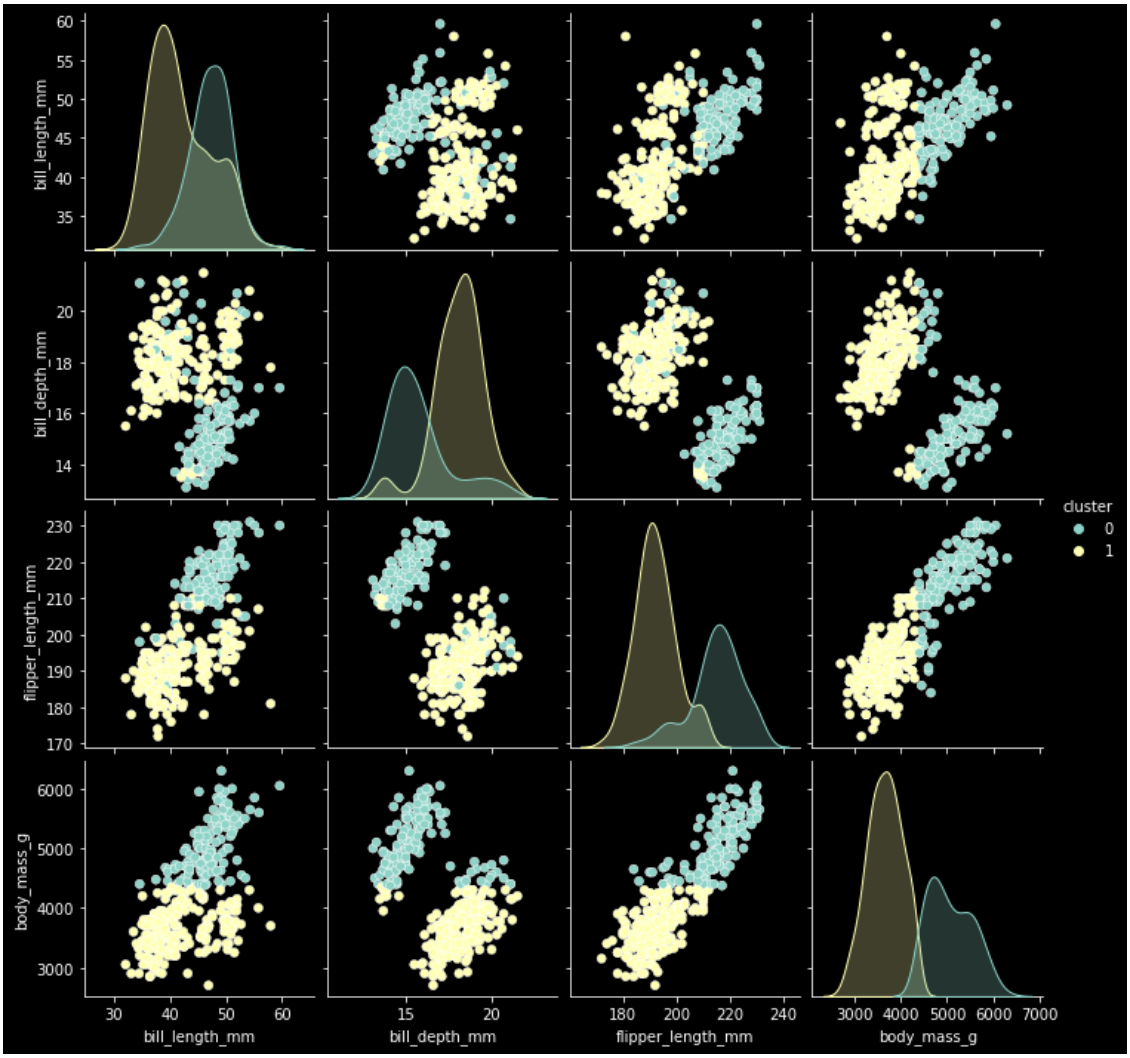
[31]:
```
penguins['cluster'] = km.predict(penguins[features])
```

[27]:
```
penguins.cluster.value_counts()
```

[27]:
```
1    203
0    130
Name: cluster, dtype: int64
```

[33]:
```
sns.pairplot(penguins, hue="cluster")
```

[33]:
```
<seaborn.axisgrid.PairGrid at 0x1abfdcadc10>
```



2 Evaluating the clustering

[18]:
```
from sklearn import metrics
from sklearn.metrics import pairwise_distances
```

[36]:
```
metrics.silhouette_score(penguins[features], km.labels_, metric='euclidean')
```

[36]:
```
0.5759423377017066
```

[34]:
```
contingency_table = penguins.groupby(['species','cluster']).size().
↪unstack('cluster', fill_value=0)
contingency_table
```

[34]:

cluster	0	1
species		
Adelie	14	132
Chinstrap	5	63
Gentoo	111	8