Portfolio assignment 13

10 min: Do a bivariate analysis on the penguins dataset for the following combination of columns:

- species VS sex
- island VS sex

For this bivariate analysis, at least perform the following tasks:

- Do you expect their to be a correlation between the two columns?
- Create a contingency table. Do you observe different ratios between categories here?
- Create a bar plot for this contingency table. Do you observe different ratios between categories here?
- Do a chi-squared test. What does the result say? What's the chance of there being a correlation between the two columns?

In [1]:

```
import seaborn as sns
from scipy.stats import chi2_contingency
penguins = sns.load_dataset("penguins")
```

In [2]:

```
contingencyTable = penguins.groupby(['species','sex']).size().unstack('species', fill_value
contingencyTable
```

Out[2]:

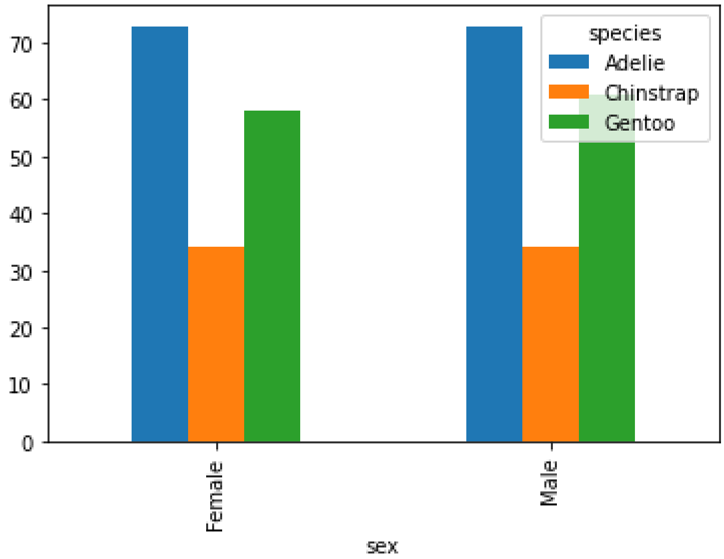
species	Adelie	Chinstrap	Gentoo
sex			
Female	73	34	58
Male	73	34	61

In [3]:

```
contingencyTable.plot(kind='bar')
```

Out[3]:

```
<AxesSubplot:xlabel='sex'>
```



It's what I expected to be, the chances of being born male or female is 50%, now it's proven that the dataset is almost devided in 50% male and 50% female.

In [4]:

```
chi2_contingency(contingencyTable)
```

Out[4]:

```
(0.04860717014078319,
 0.9759893689765846,
 2,
 array([[72.34234234, 33.69369369, 58.96396396],
        [73.65765766, 34.30630631, 60.03603604]]))
```

It's a extreme low correlation, it's safe to say that there is no link between those two colums

In [5]:

```
contingencyTable = penguins.groupby(['island','sex']).size().unstack('island', fill_value=0
contingencyTable
```

Out[5]:

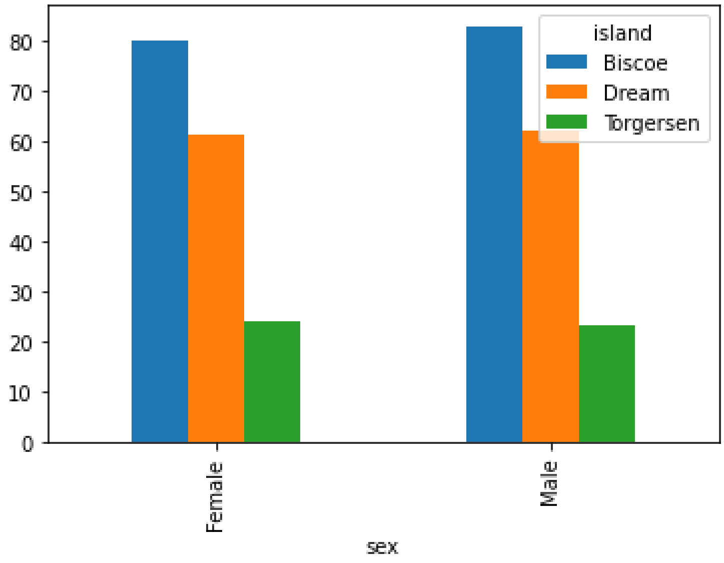
island	Biscoe	Dream	Torgersen
sex			
Female	80	61	24
Male	83	62	23

In [6]:

```
contingencyTable.plot(kind='bar')
```

Out[6]:

```
<AxesSubplot:xlabel='sex'>
```



It's what I expected to be, a male needs a female and the other way around too, so it makes sense that the population of the islands is devided in almost 50% male and 50% female

In [7]:

```
chi2_contingency(contingencyTable)
```

Out[7]:

```
(0.05759904881286206,
 0.971611229281065,
 2,
 array([[80.76576577, 60.94594595, 23.28828829],
        [82.23423423, 62.05405405, 23.71171171]]))
```

It's a extreme low correlation, it's safe to say that there is no relation between those two colums