Bivariate Analysis

Category vs category

In [2]:

```
import pandas as pd
import seaborn as sns
```

In [3]:

```
penguins = sns.load_dataset("penguins")
penguins.head()
```

Out[3]:

	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

In [4]:

```
def create_contingency_table(dataset, column1, column2):
    return dataset.groupby([column1, column2]).size().unstack(column1, fill_value=0)
```

Species vs Sex

Ik verwacht geen correlatie tussen species en sexe.

In [5]:

```
speciesVSsex = create_contingency_table(penguins, 'sex','species')
speciesVSsex
```

Out[5]:

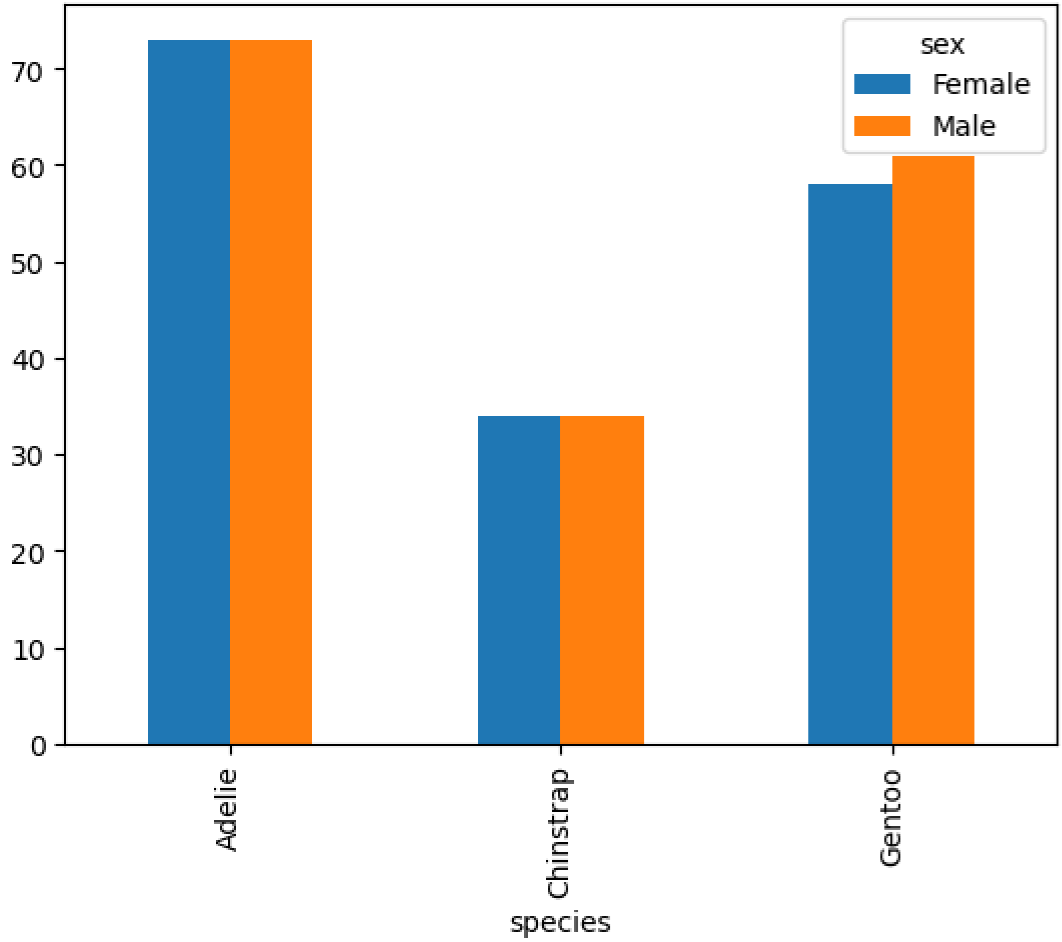
sex	Female	Male
species		
Adelie	73	73
Chinstrap	34	34
Gentoo	58	61

In [6]:

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

Out[6]:

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



Op het eiland Gentoo is de ratio male to female net iets anders dan de andere twee eilanden, waar de ratio 1 is.

In [7]:
```
from scipy.stats import chi2_contingency
def check_cat_vs_cat_correlation(dataset, column1, column2):
    contingency_table = create_contingency_table(dataset, column1, column2)
    chi2 = chi2_contingency(contingency_table)
    p_value = chi2[1]
    odds_of_correlation = 1 - p_value
    print(f"The odds of a correlation between {column1} and {column2} is {odds_of_correl
    print("This percentage needs to be at least 95% for a significant correlation.")
```

In [8]:
```
check_cat_vs_cat_correlation(penguins, 'sex','species')
```

```
The odds of a correlation between sex and species is 2.4010631023415385% (Based on a p v
alue of 0.9759893689765846).
This percentage needs to be at least 95% for a significant correlation.
```

Zoals verwacht is er maar een kleine kans (2.40%) dat er een correlatie is tussen species en sexe.

Island vs Sex

Ook hier verwacht ik dat er weinig correlatie zal zijn tussen de twee categorieën.

In [9]:
```
islandVSsex = create_contingency_table(penguins, 'sex','island')
islandVSsex
```

Out[9]:

sex	Female	Male
island		
Biscoe	80	83

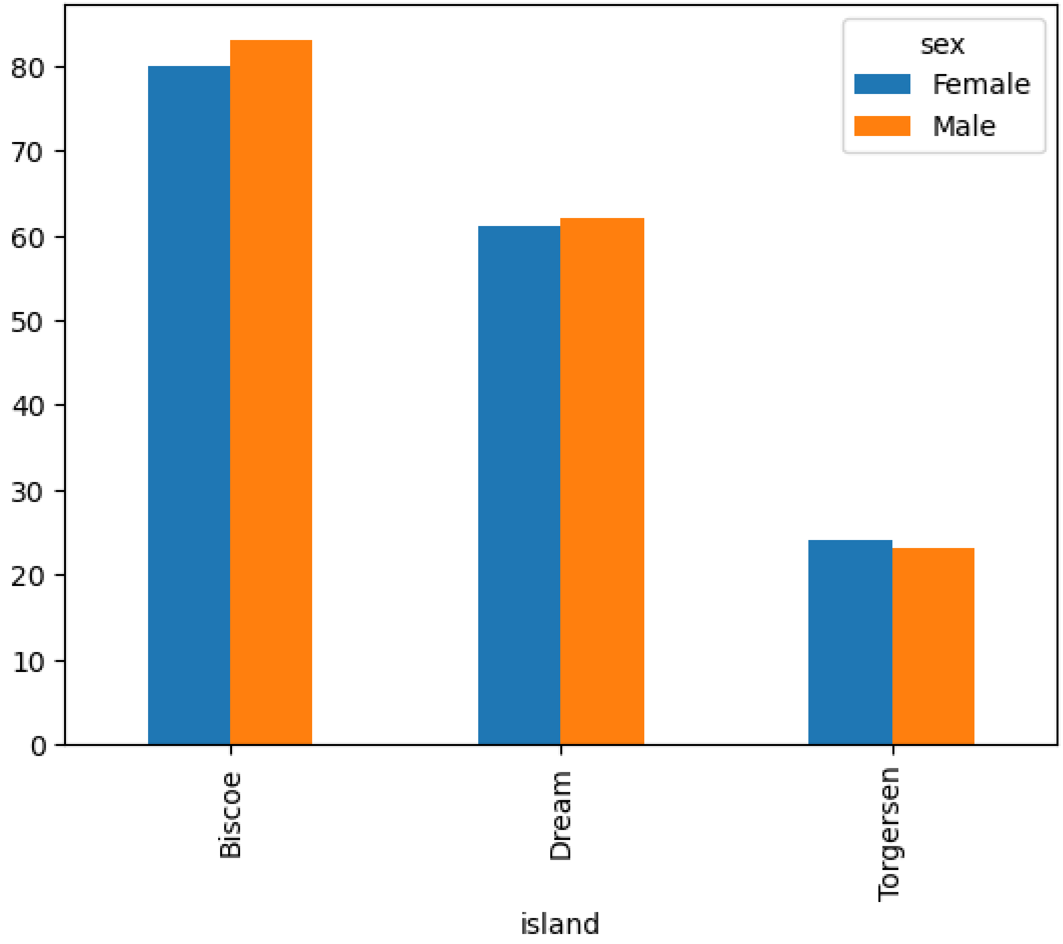
Dream	61	62
Torgersen	24	23

In [10]:

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

Out[10]:

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



Er is hier iets meer verschil te zien dan bij species, maar het lijkt met niet significant genoeg om te zeggen dat er een correlatie is.

In [11]:

```
check_cat_vs_cat_correlation(penguins, 'sex','island')
```

```
The odds of a correlation between sex and island is 2.83887707189349755% (Based on a p va
lue of 0.971611229281065).
This percentage needs to be at least 95% for a significant correlation.
```

En zoals verwacht is de kans dat er een correlatie is erg klein (2.84%).