Bivariate Analysis

Numerical vs Categorical

In [1]:
```
import pandas as pd
import seaborn as sns
```

In [2]:
```
penguins = sns.load_dataset("penguins")
```

In [3]:
```
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]:
```
category = "island"
penguins.groupby(category).mean()
```

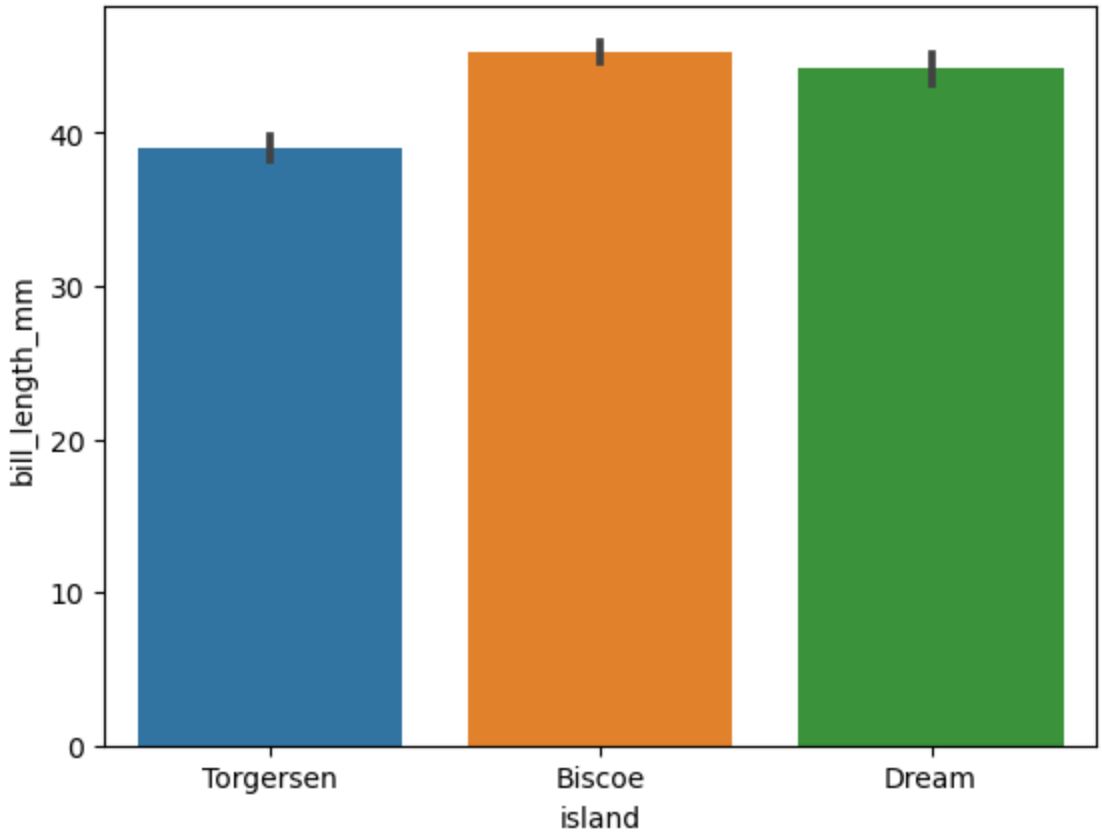
Out[4]:

island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g
Biscoe	45.257485	15.874850	209.706587	4716.017964
Dream	44.167742	18.344355	193.072581	3712.903226
Torgersen	38.950980	18.429412	191.196078	3706.372549

Er lijken wel verschillen te zitten tussen penguins van verschillende eilanden, voornamelijk dat penguins van Biscoe waarschijnlijk groter zijn dan van de andere twee eilanden. Om te kijken of de verschillen significant zijn plotten we de waarden met hun 95% confidence interval.

In [5]:
```
sns.barplot(y="bill_length_mm", x=category, data=penguins)
```

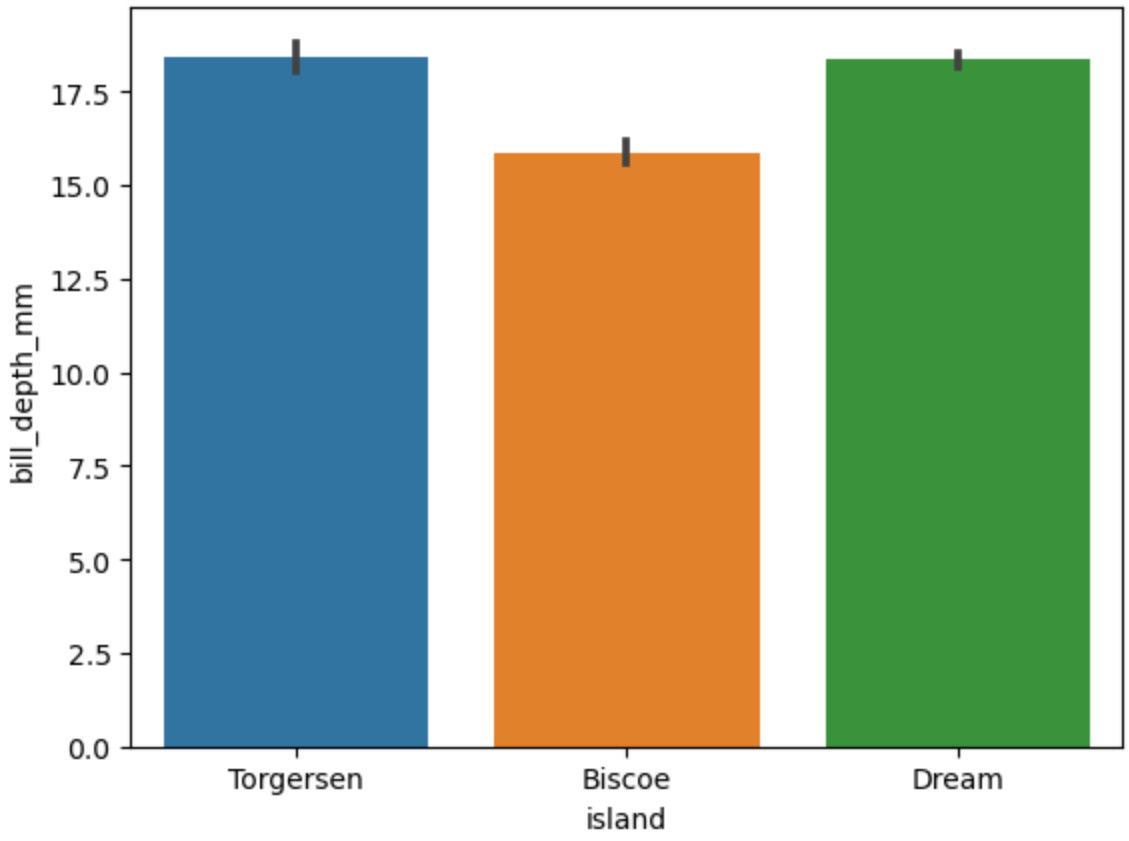
Out[5]:
```
<AxesSubplot:xlabel='island', ylabel='bill_length_mm'>
```



Voor bill length lijkt er tussen Biscoe en Dream geen significant verschil te zijn, de confidence intervallen overlappen. Torgersen heeft echter significant kleinere waarden.

In [6]:
```
sns.barplot(y="bill_depth_mm", x=category, data=penguins)
```

Out[6]:
```
<AxesSubplot:xlabel='island', ylabel='bill_depth_mm'>
```

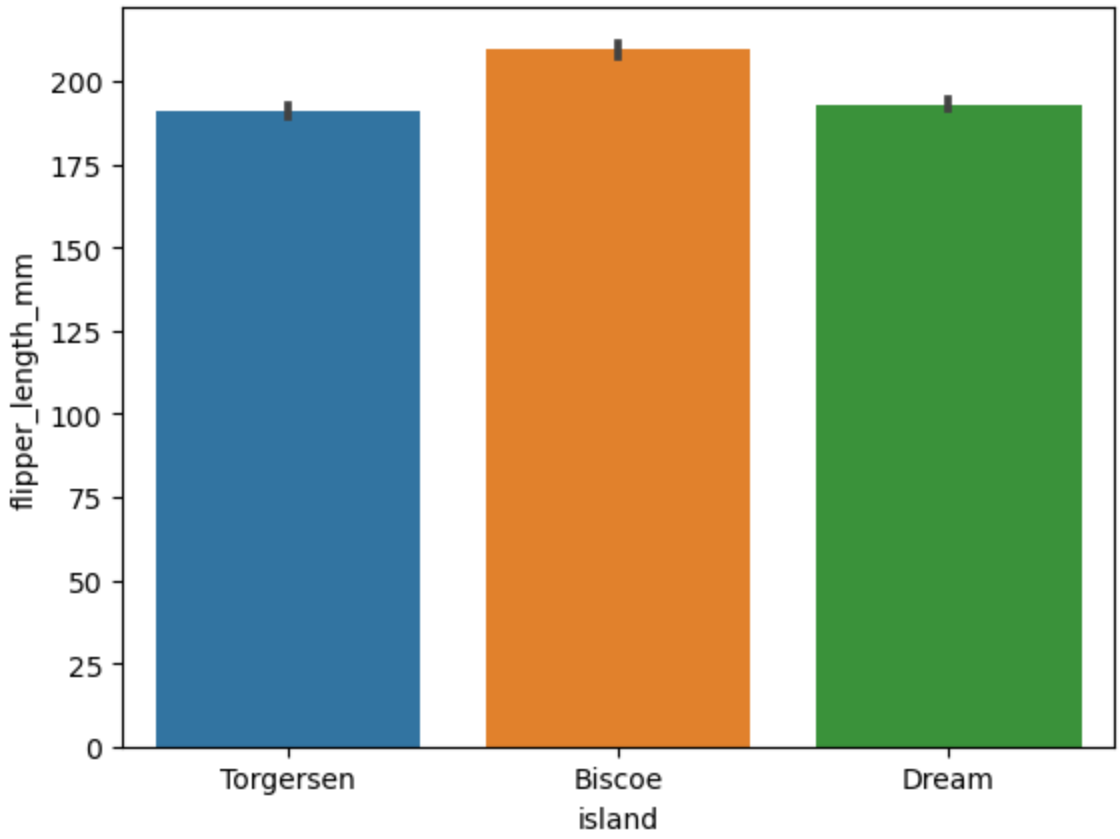


Voor bill depth lijken Torgersen en Dream vergelijkbare waarden te hebben, terwijl Biscoe significant kleinere

waarden heeft.

In [7]:
```
sns.barplot(y="flipper_length_mm", x=category, data=penguins)
```

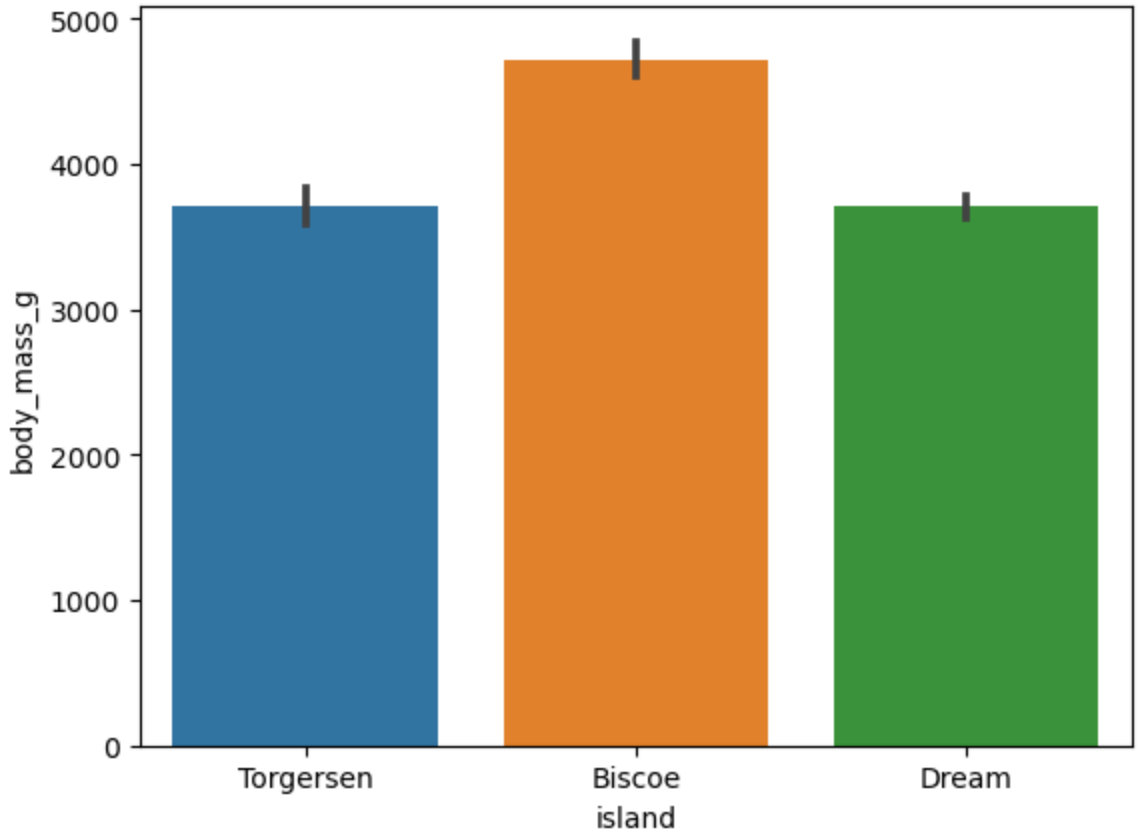
Out[7]:
```
<AxesSubplot:xlabel='island', ylabel='flipper_length_mm'>
```



Flipper length heeft een erg klein confidence interval, waardoor het bijna moeilijk te zien is of ze overlappen, maar het lijkt erop dat Torgersen en Dream weer vergelijkbare waarden hebben, terwijl Biscoe er duidelijk buiten ligt.

In [8]:
```
sns.barplot(y="body_mass_g", x=category, data=penguins)
```

Out[8]:
```
<AxesSubplot:xlabel='island', ylabel='body_mass_g'>
```



De body mass is vergelijkbaar aan de flipper lenght, Biscoe is hier een duidelijke outlier.