assignment 11

March 8, 2022

1 Bivariate analysis of numerical vs categorical columns using Penguins dataset

[15]:
```
import pandas as pd
import seaborn as sns
```

[16]:
```
penguins = sns.load_dataset("penguins")
sns.set_style("dark")
```

[17]:
```
penguins.head()
```

[17]:
```
  species     island  bill_length_mm  bill_depth_mm  flipper_length_mm  \
0  Adelie  Torgersen            39.1           18.7              181.0
1  Adelie  Torgersen            39.5           17.4              186.0
2  Adelie  Torgersen            40.3           18.0              195.0
3  Adelie  Torgersen             NaN            NaN                NaN
4  Adelie  Torgersen            36.7           19.3              193.0

   body_mass_g     sex
0       3750.0    Male
1       3800.0  Female
2       3250.0  Female
3          NaN     NaN
4       3450.0  Female
```

[18]:
```
penguins.groupby(by="species").mean()
```

[18]:
```
           bill_length_mm  bill_depth_mm  flipper_length_mm  body_mass_g
species
Adelie          38.791391      18.346358         189.953642  3700.662252
Chinstrap       48.833824      18.420588         195.823529  3733.088235
Gentoo          47.504878      14.982114         217.186992  5076.016260
```

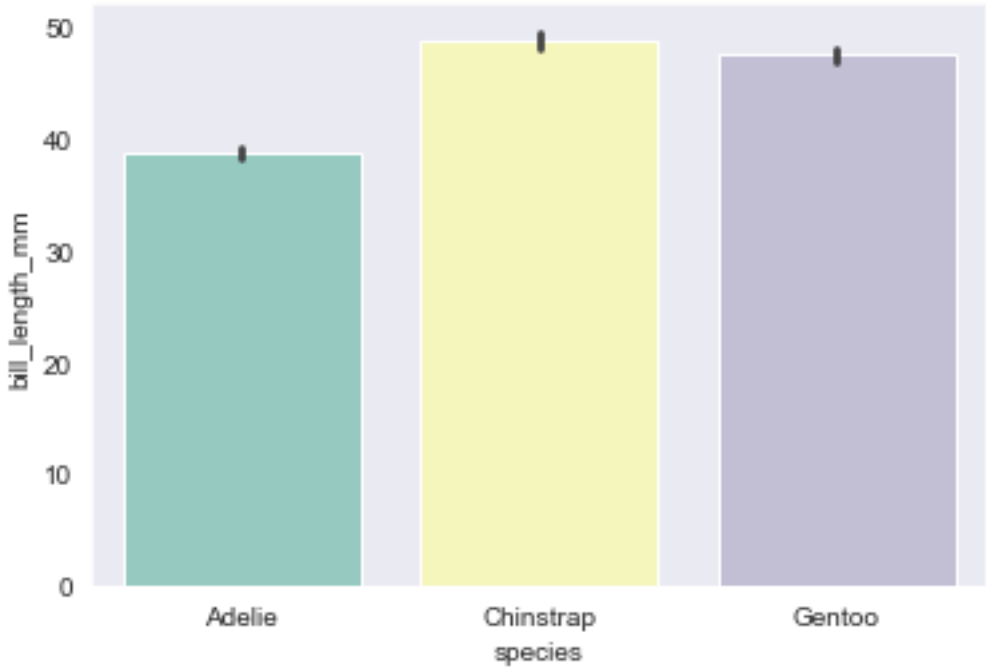
I chose to compare species and island to each of the numerical columns

2 Species

It seems there are less differences between Adelie and Chinstrap than between Adelie, Chinstrap and Gentoo.

[19]:
```
sns.barplot(y="bill_length_mm",x="species",data=penguins)
```

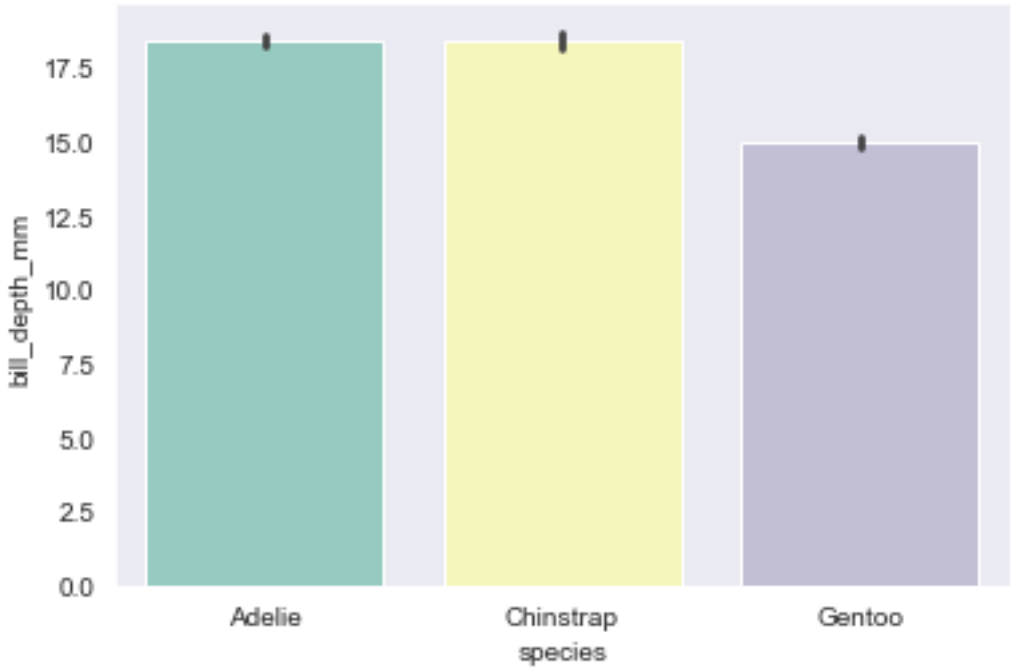
[19]:
```
<AxesSubplot:xlabel='species', ylabel='bill_length_mm'>
```



There is no statistically significant difference between Chinstrap and Gentoo, but there is between Adelie and the others

[20]:
```
sns.barplot(y="bill_depth_mm",x="species",data=penguins)
```

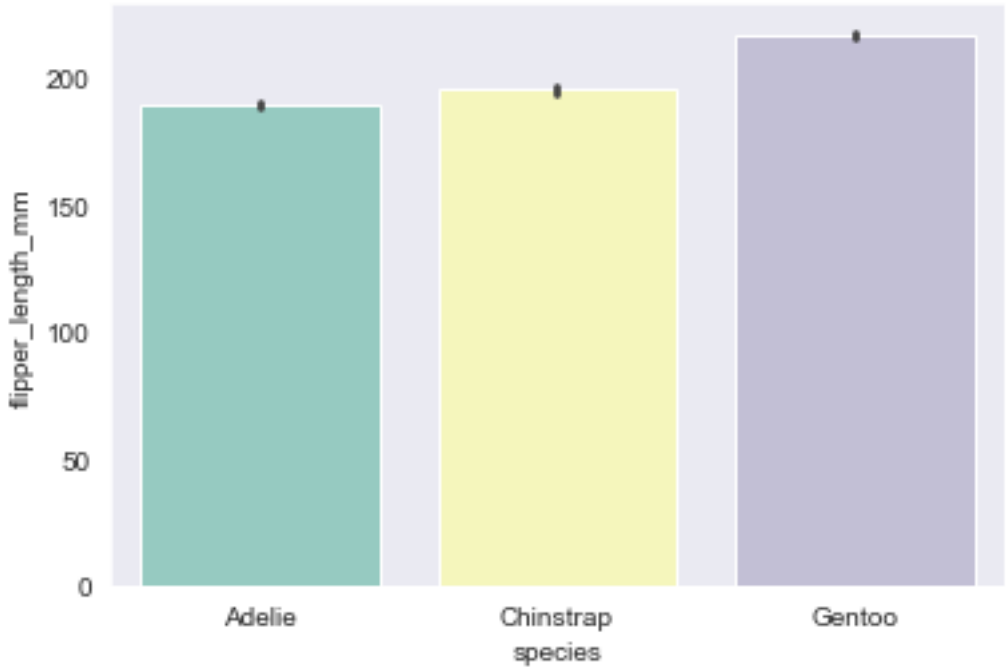
[20]:
```
<AxesSubplot:xlabel='species', ylabel='bill_depth_mm'>
```



There is no statistically significant difference between Adelie and Gentoo, but there is between Gentoo and the others

[21]:
```
sns.barplot(y="flipper_length_mm",x="species",data=penguins)
```

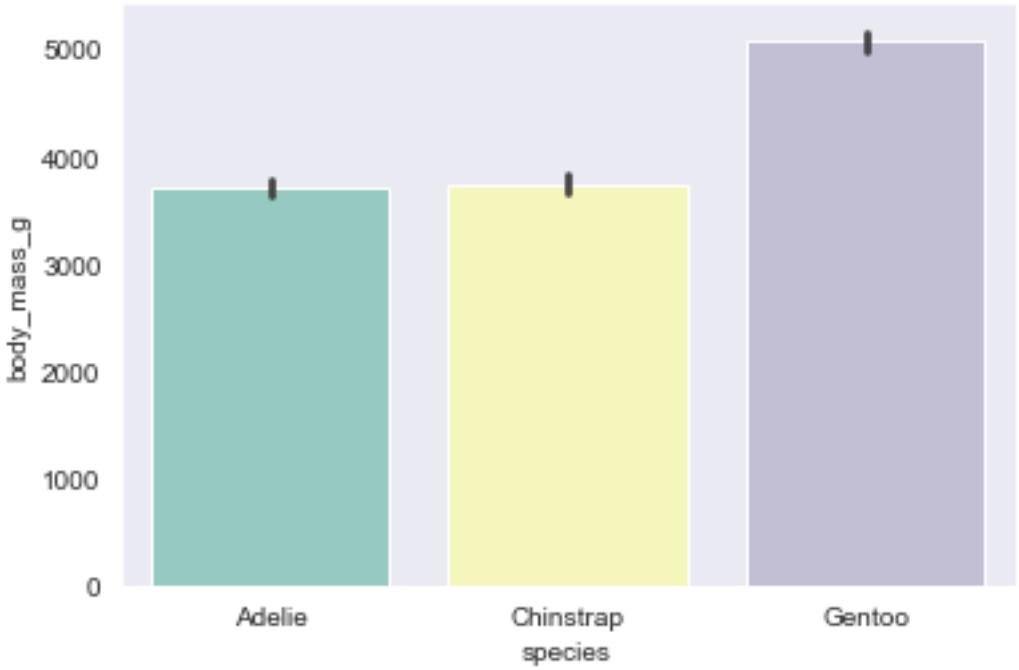
[21]:
```
<AxesSubplot:xlabel='species', ylabel='flipper_length_mm'>
```



There is no statistically significant difference between Adelie and Gentoo, but there is between Gentoo and the others

[22]:
```
sns.barplot(y="body_mass_g",x="species",data=penguins)
```

[22]:
```
<AxesSubplot:xlabel='species', ylabel='body_mass_g'>
```

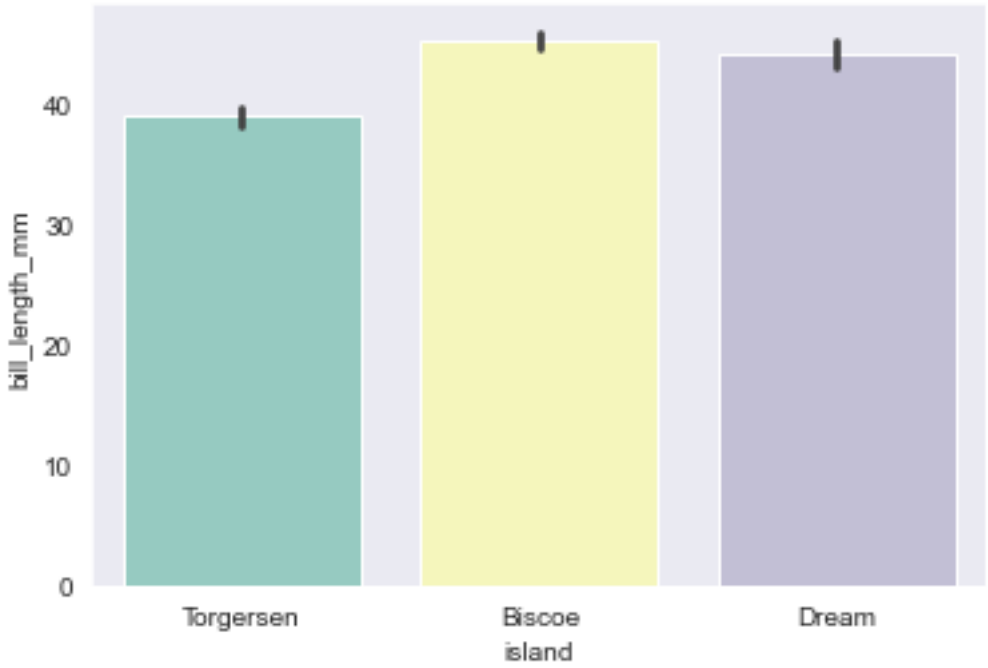


There is no statistically significant difference between Adelie and Gentoo, but there is between Gentoo and the others

3 Island

[23]:
```
sns.barplot(y="bill_length_mm",x="island",data=penguins)
```

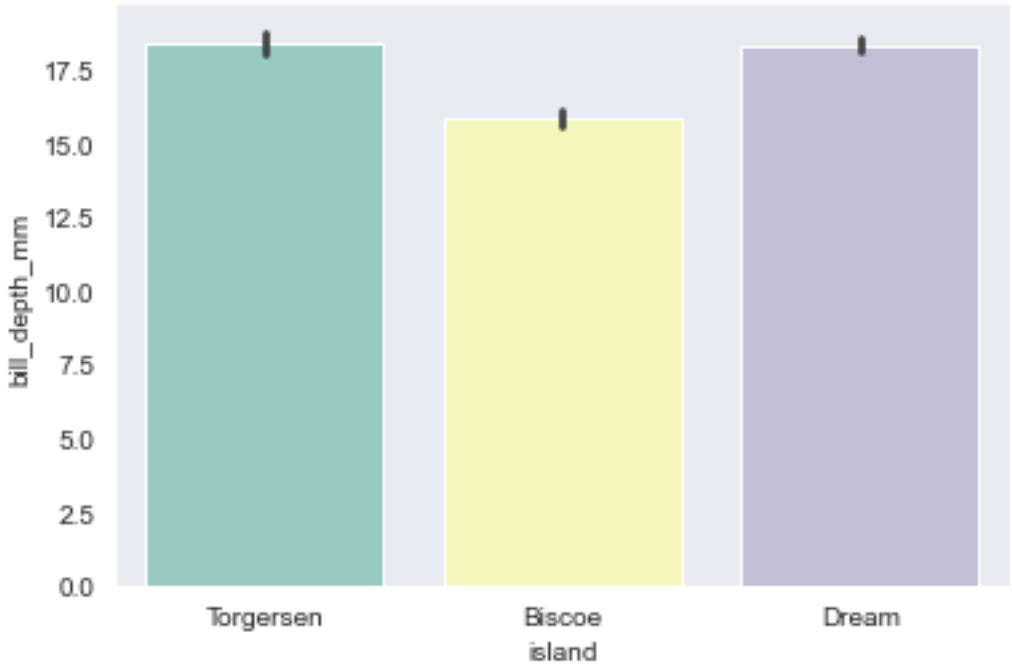
[23]:
```
<AxesSubplot:xlabel='island', ylabel='bill_length_mm'>
```



There is no statistically significant difference between Biscoe and Dream, but there is between Togersen and the others

[24]:
```
sns.barplot(y="bill_depth_mm",x="island",data=penguins)
```

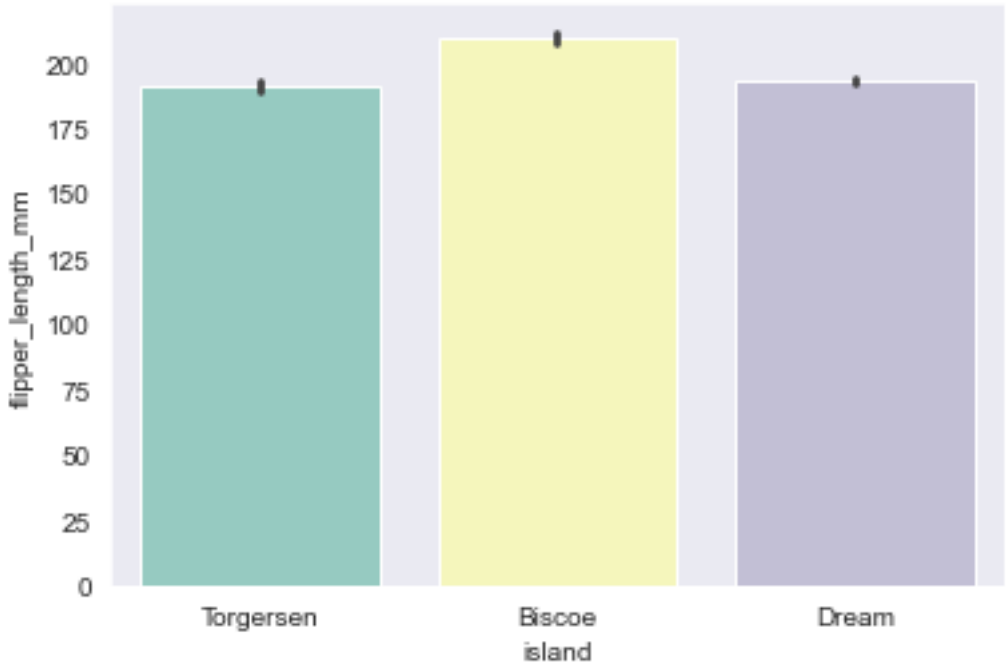
[24]:
```
<AxesSubplot:xlabel='island', ylabel='bill_depth_mm'>
```



There is no statistically significant difference between Togersen and Dream, but there is between Biscoe and the others

[25]:
```
sns.barplot(y="flipper_length_mm",x="island",data=penguins)
```

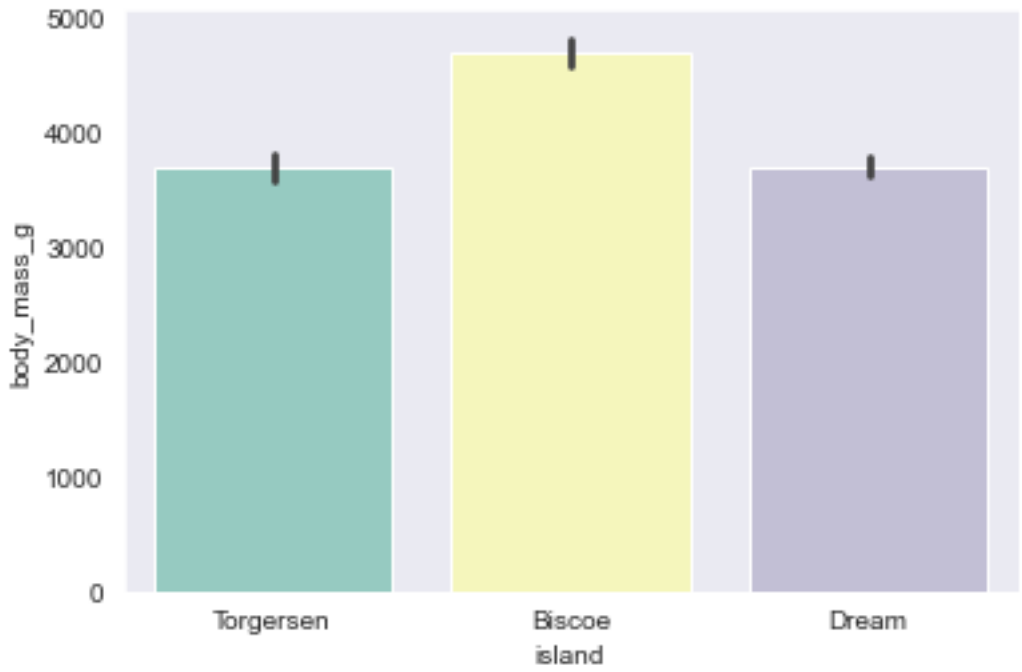
[25]:
```
<AxesSubplot:xlabel='island', ylabel='flipper_length_mm'>
```



There is no statistically significant difference between Togersen and Dream, but there is between Biscoe and the others

[26]:
```
sns.barplot(y="body_mass_g",x="island",data=penguins)
```

[26]:
```
<AxesSubplot:xlabel='island', ylabel='body_mass_g'>
```



There is no statistically significant difference between Togersen and Dream, but there is between Biscoe and the others