Exploratory Data Analysis on the penguins_size.csv Dataset

Introduction

This is a dataset made of 344 rows and 7 columns and the topic is related to penguins. The columns in the dataset are species, island, culmen_lenght_mm, culmen_depth_mm, flipper_lenght_mm, body_mass_g, and sex.

DATA CLEANING

In the Penguin file, I checked for duplicated rows, but there weren't any.

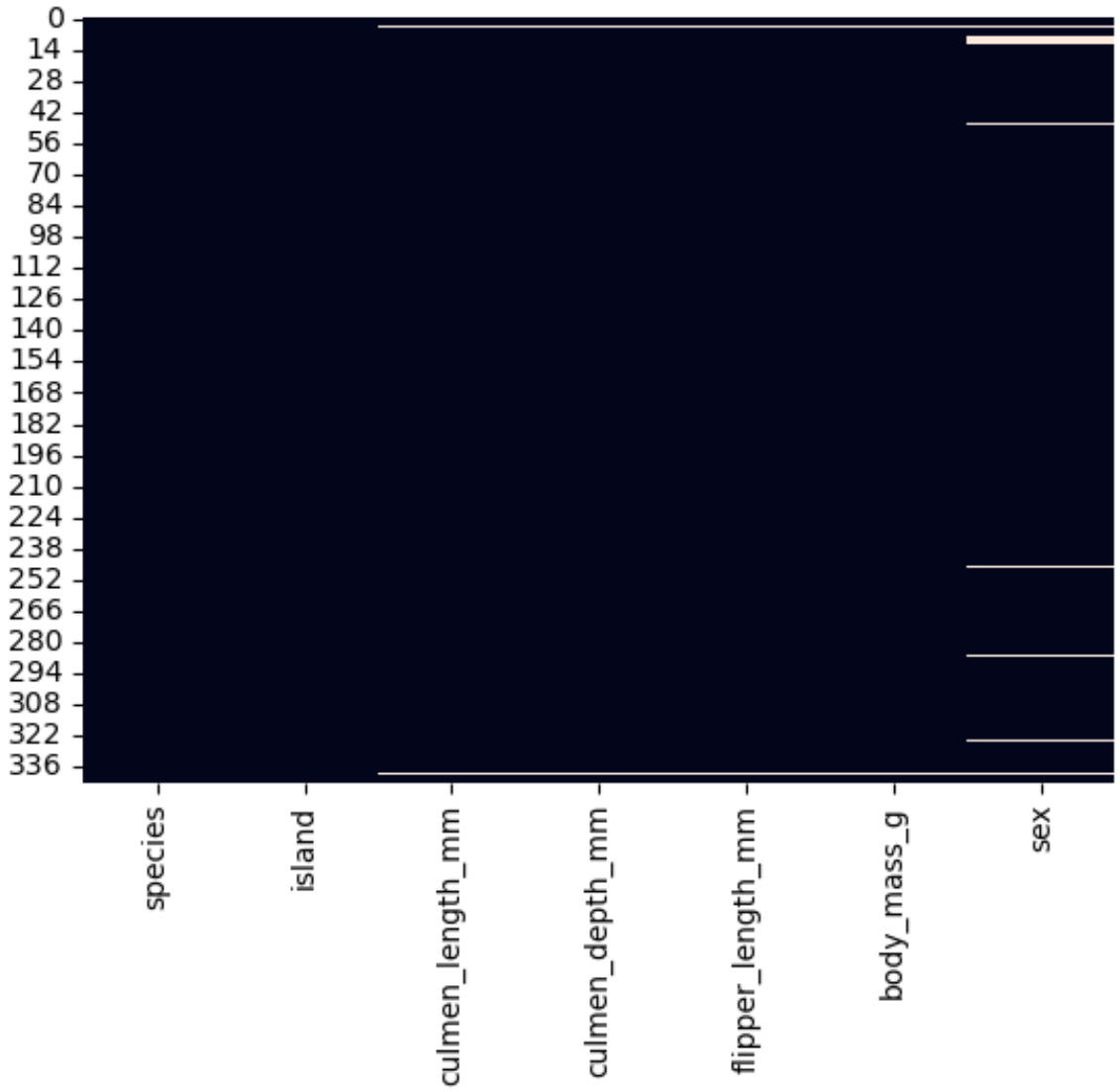
I checked for null values, and I found 2 belonging to the columns culmen_lenght_mm, culmen_depth_mm, flipper_lenght_mm, and body_mass_g.
I found another 10 missing values inside the column sex plus one “.” value in the same column.

MISSING DATA

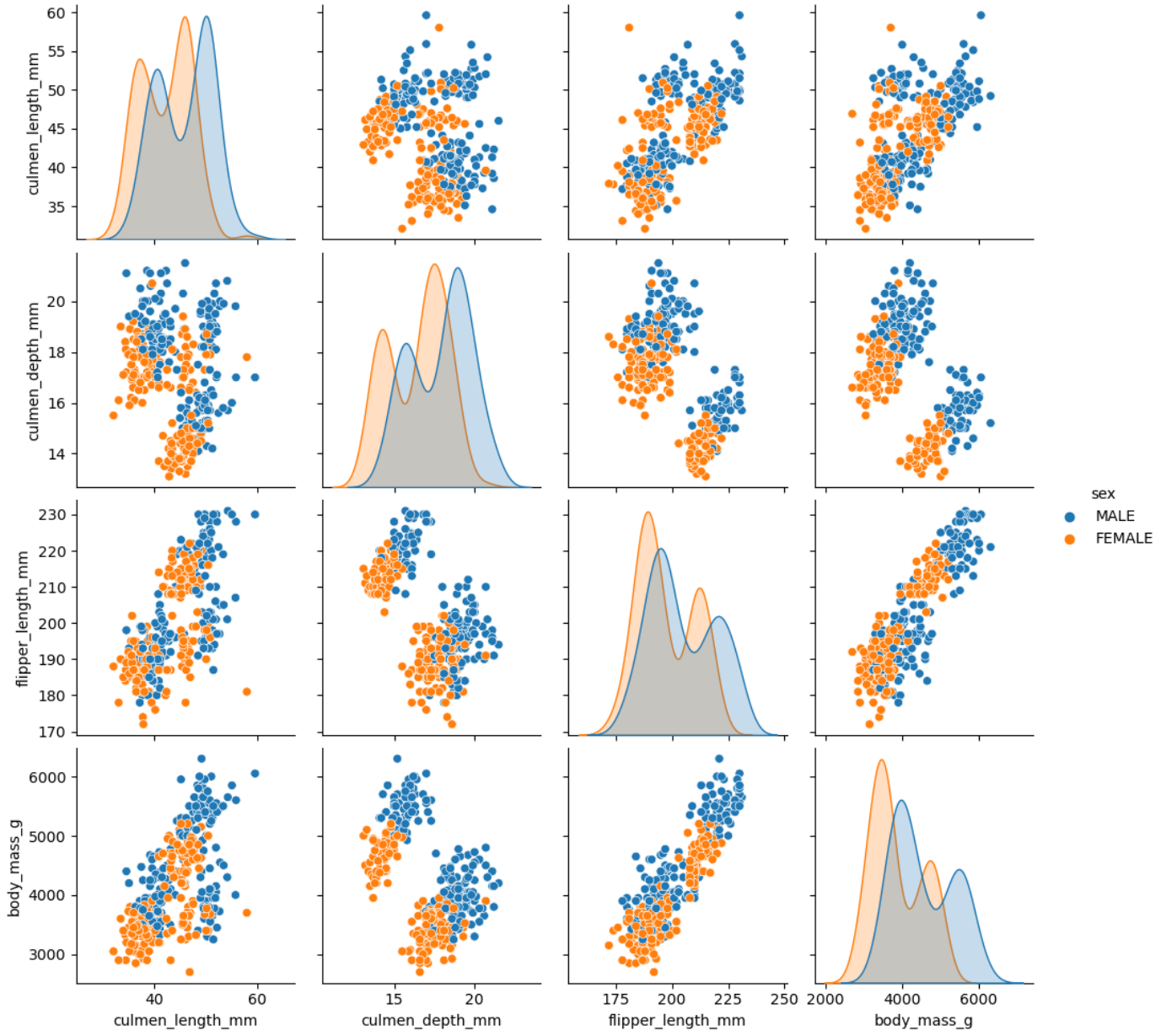
I decided to convert the “.” value to a Nan value in the categorical column sex.
I dropped the rows indexed 3 and 339 because they were many missing values.
After doing this, I had only 9 missing values in the column sex.
I tried to find a relationship between the dataset and the column sex for replacing those missing values with the opposite gender.

DATA STORIES AND VISUALISATIONS

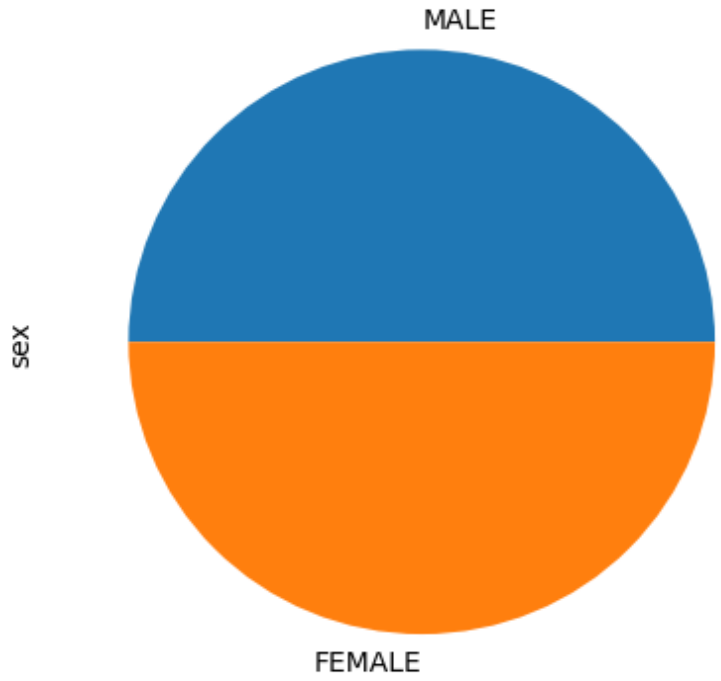
Missing value in the dataset:
Through this heatmap graph, you can visualize that most of the missing values are located in the column sex.



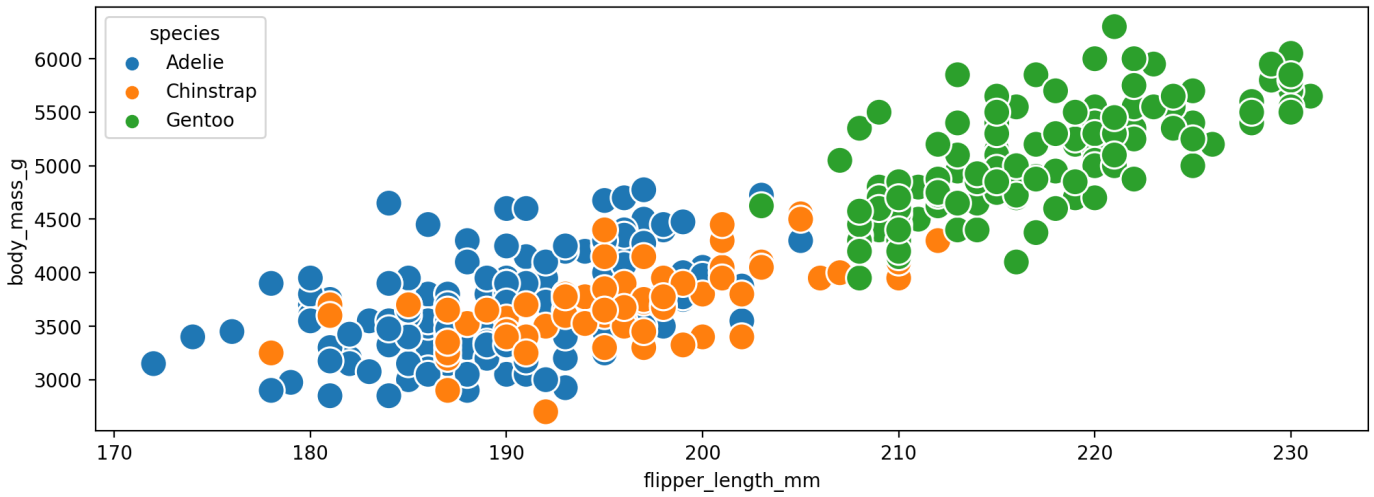
Discover the relationship between sex and the dataset:
From the pair plot below, we visualize that the Male sex has higher body mass, flipper length, culmen depth, and also culmen length.
This also helped me to replace the missing value in sex.



Genres distribution:
From this apple pie, we can easily notice that the distribution between males and females is equal.

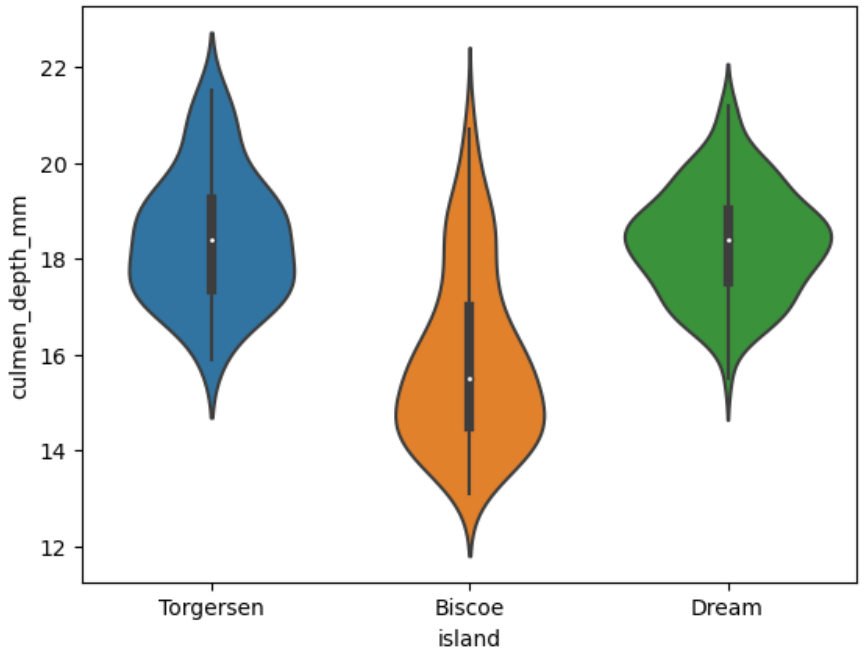


Relationship between flipper length, body mass, and species:
From the scatterplot graph below, we can visualize that the relationship between flipper length and body mass is directly proportional.
The bigger the size of the penguin, the longer the flipper.
We also visualize that penguins that belong to the Gentoo species are much bigger.



Culmen depth dimensions to each island

From this violin plot, we can visualize that in Biscoe island, there is a wider variation between the culmen depth of the penguins'.



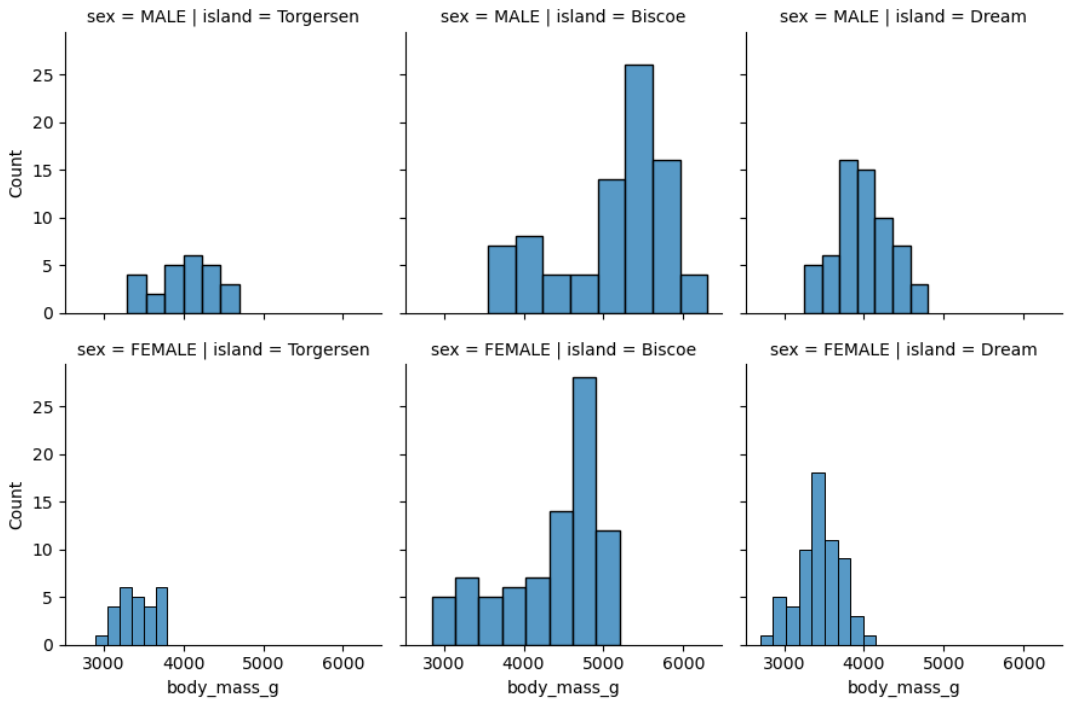
Correlation between island and sex for each body mass

From the face grid graph above, we can visualize different things.

The island of Biscoe contains bigger sizes penguins

The majority of the male genre that belongs to this island, has a body mass of around 5200 g.

The majority of the female genre that belongs to this island, has a body mass of around 4800 g.



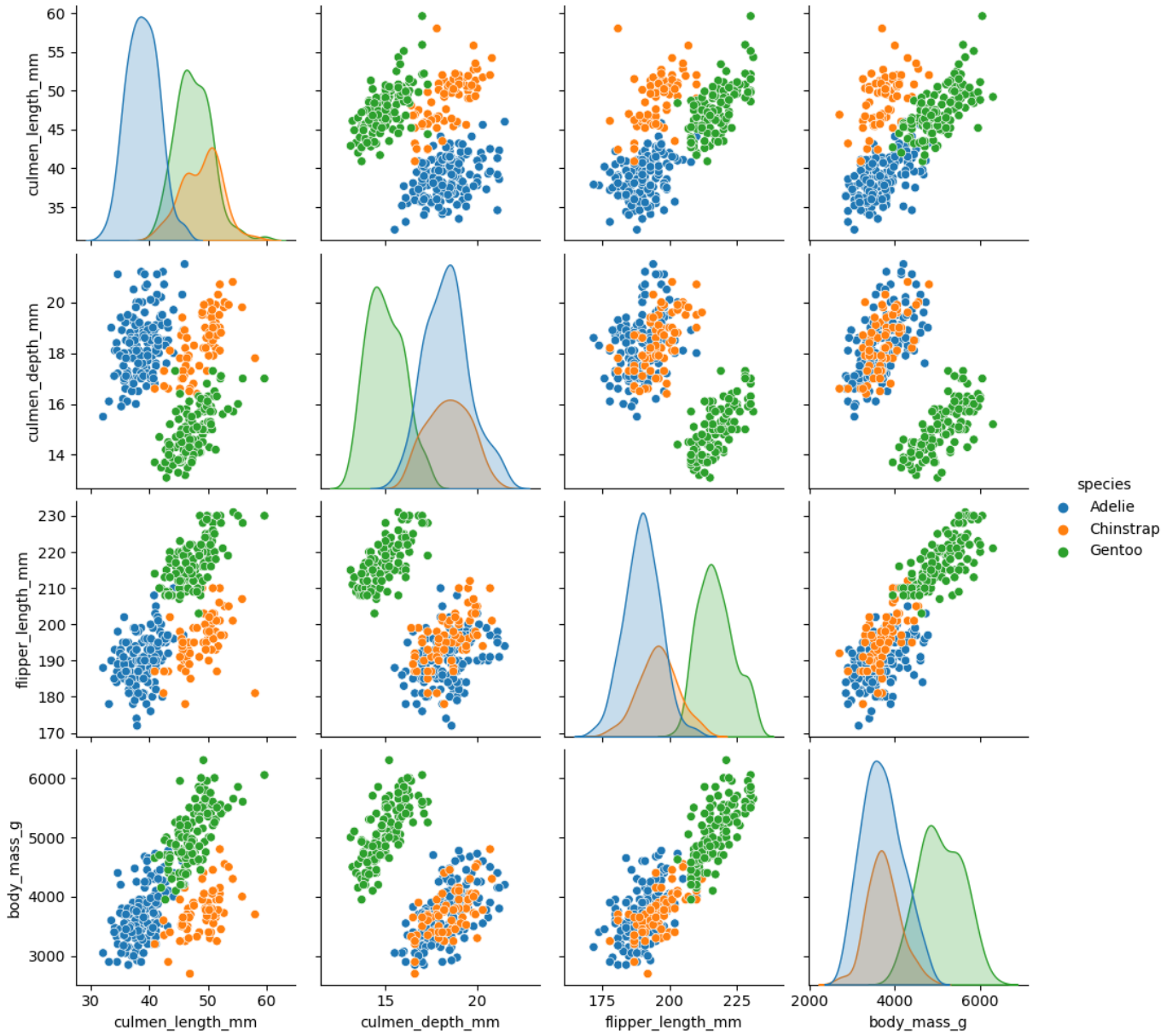
Discover the relationship between species:

From the above pair plot graph, we notice something interesting.

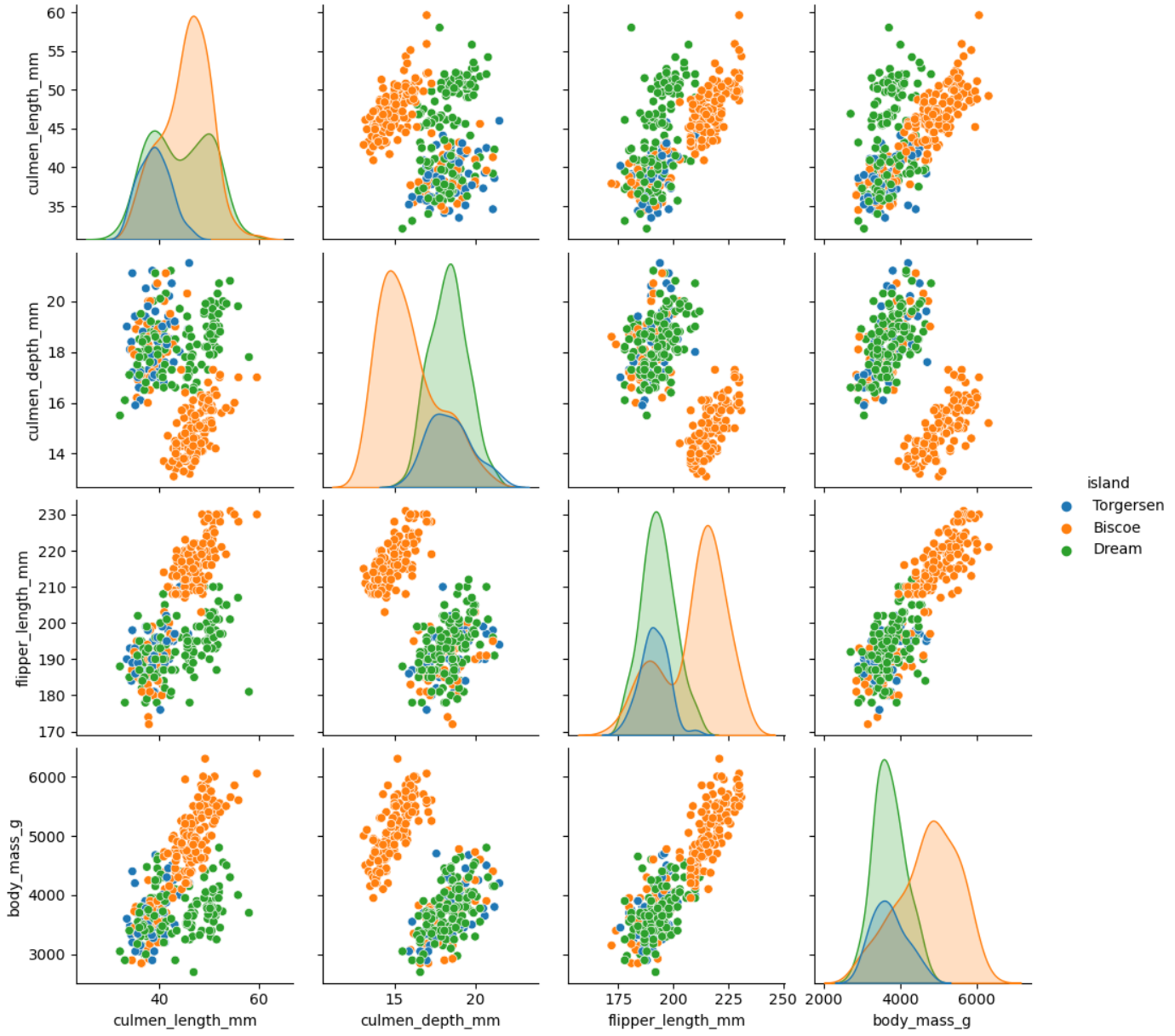
We can notice the sizes of the penguin belonging to the species Chinstrap and Adelie are very similar.

They have the same culmen depth, and flipper length and they differ only in the length of the culmen.

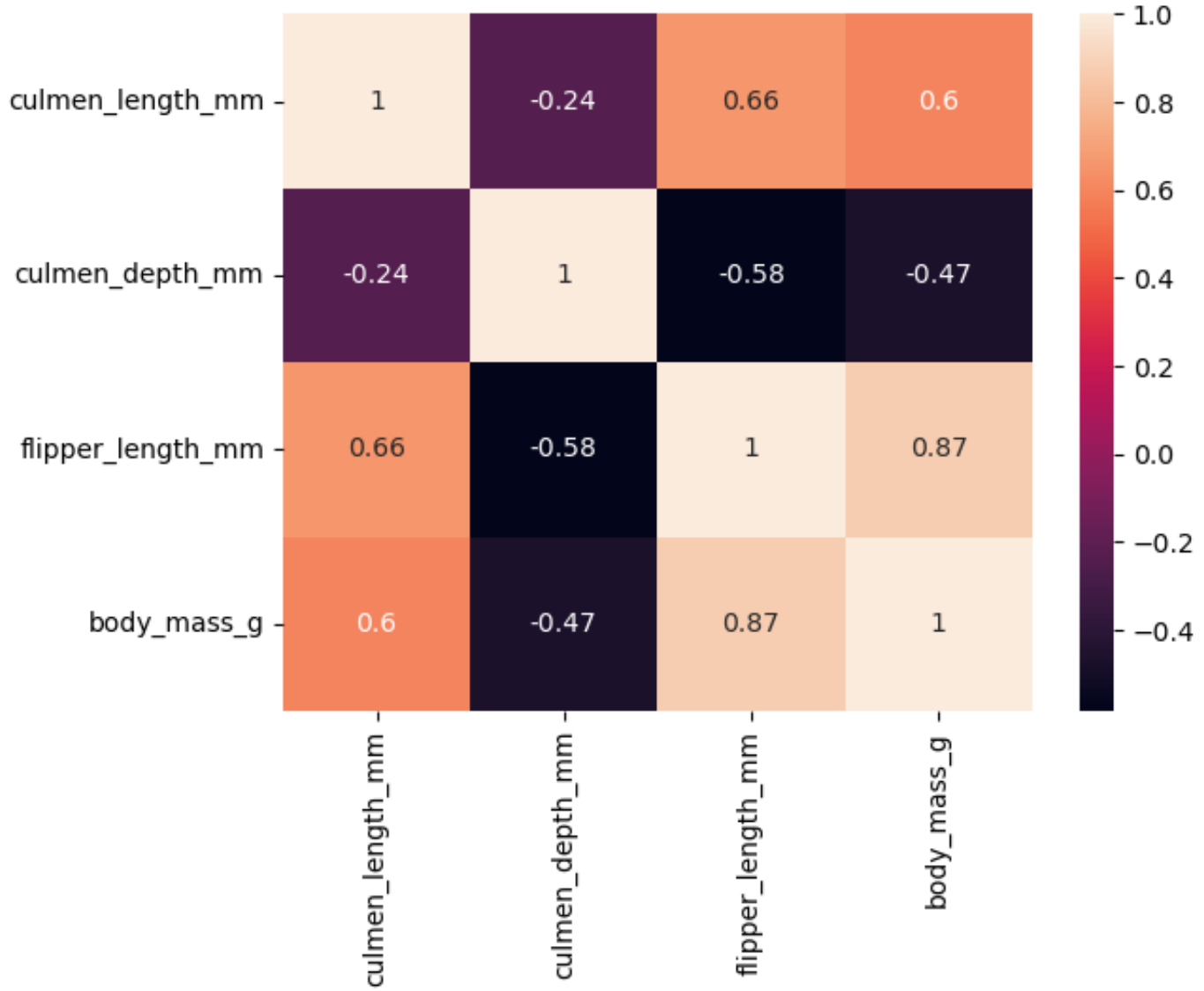
Differently, Gentoo species have different sizes.



Discover the relationship between islands:
Like the example above, we can visualize the relationship between the different islands.
It seems that in the islands of Biscoe and dream, the penguins have similar numeric variables.
Instead, on Torgersen island, we visualize that penguins have different sizes.



Correlation between numerical values:
In conclusion, we visualize from the heat graph below that only the culmen depth is negatively correlated with the other variables.



THIS REPORT WAS WRITTEN BY: GAETANO LOPEZ
DATE: 20/12/2022