**Assignment 7: Multivariate Data – Parallel Coordinate Plot**

**Task 1a**

1. First, we can clearly identify where do each of the species lives. Adelie penguins live in the three islands, tend to have a rather short but deep culmens, and their flipper length and body mass are also, in comparison to the other species, lower. Gentoos seem to be quite homogeneous as species, indicated by the density of lines in the plot. These penguins are larger than the ones from other species, since we can observe their flipper length and body mass are higher. Chinstrap penguins, on the other hand, are quite heterogeneous where body measurements concern. It’s not so easy to identify a clear pattern from the PCP.
2. Given the current order of the dimensions, we can really only visualize the correlation between gender and body mass, which are next to each other. There is a big overlap in body mass between females and males, but the largest penguins are males (from the Gentoo species), and the smallest penguins are females (mostly Adelie).
3. The pairwise correlation depends on the species. For Gentoos, there is a negative correlation between culmen length and depth, and a positive correlation between culmen depth and flipper length. For Adelies these correlations are in the opposite direction (positive between culmen length and depth, negative between culmen depth and flipper length). For Chinstraps it is difficult to visualize any clear correlation.
4. By looking at less areas in the extreme values, we can identify some outliers with rather large culmen length (Gentoo and Chinstrap), some with rather low values in flipper length (Adelie), and a rather large penguin in body mass from Gentoo species.

**Task 1b**

Advantages:

* Easier to identify cluster characteristics (for example with gender and body mass), or location of each species.
* Adding the categorical variables gives a more comprehensive outlook on the dataset.

Disadvantages:

* Generate a lot of overlapping lines.
* Makes the interpretation more complex, the viewer has to pay careful attention to the variable labels and reading them gets more difficult the more categories and observations there are. Is not intuitive as with numerical categories.
* Makes the plot more complex by adding more dimensions that could be added to the analysis in a different way – for example as it is done in the PCP with color for each species.

**Task 1c**

* We could use the interactive filter by species, which would reduce the overlap and may allow us to visualize new patterns. We could also filter by the other categorical variables, e.g. for sex or island. In this way we could remove them from the plot, reducing the complexity of it while also providing a more intuitive visualization.