GDA of *finch* (from **dynRB**)

Background The data are morphological measurements of Darwin's finches. There are quantitative measurements of nine traits characterizing five species of finches and each trait was measured on at least 10 individuals per species.

Aims How do the species differ? Which trait measurements are most important in differentiating species? How are the trait measurements related to one another?

Source Snodgrass R and Heller E (1904) Papers from the Hopkins-Stanford Galapagos Expedition, 1898-99. XVI. Birds. Proceedings of the Washington Academy of Sciences 5: 231-372. The book is available at https://archive.org/details/cu31924000035349. From this it is clear that only birds from the main island (then called Albemarle, now called Isabela) are included in the dataset, that cases with missing values have been excluded, and that the variable sex has been unaccountably left out.

Structure 146 observations on 10 variables (9 numeric, 1 factor)

The distribution of species is uneven:

```r
data(finch, package="dynRB")
ggplot(finch, aes(Species)) + geom_bar() +
  coord_flip() + xlab("") + ylab("Counts")
```

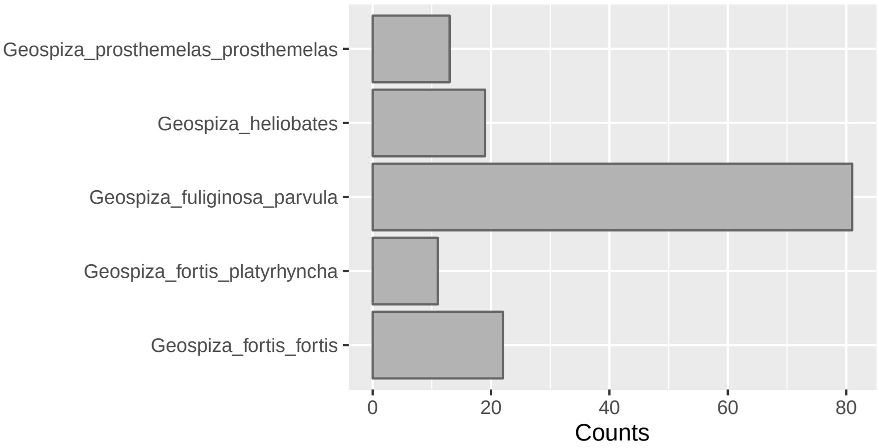

Fig 1: The numbers of birds of the five different species in the dataset. The distribution is unbalanced with over half the birds being from one species, but that is real data for you.

Individual features/variables can be compared by species in a number of ways. Figure 2 shows boxplots by group for body length on the left hand side and wing length on the right hand side. Note that first a new species variable has been constructed to provide shortened labels:

```r
finch$Sp <- gsub("Geospiza_(\\w+)", "\\1", finch$Species)
finch$Sp <- gsub("fuliginosa_(\\w+)", "fuliginosa", finch$Sp)
finch$Sp <- gsub("prosthemelas_(\\w+)", "prosthemelas", finch$Sp)
```

```
library(gridExtra)
a1 <- ggplot(finch, aes(Sp, BodyL)) + geom_boxplot() + xlab("")
a2 <- ggplot(finch, aes(Sp, WingL)) + geom_boxplot() + xlab("")
grid.arrange(a1,a2, nrow=1)
```

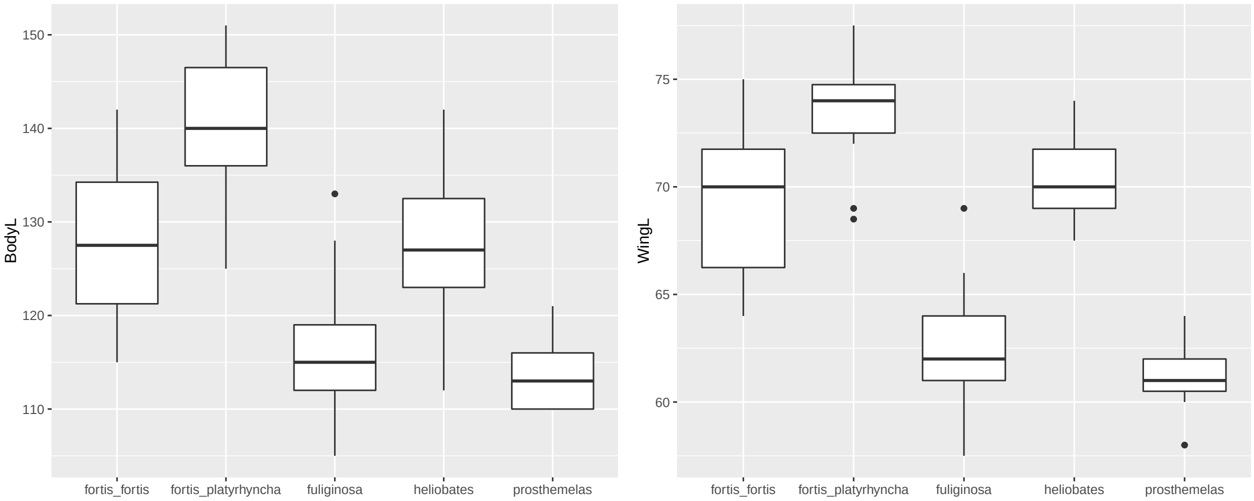

Fig 2: Body and wing lengths for the five species. The patterns are consistent. Two species are generally shorter than the others, two are in the middle, and one is bigger.

The small sizes of the groups and the overlaps make firm conclusions based on Figure 2 difficult. Perhaps displaying the two variables together in a scatter plot would separate the groups better. Figure 3 shows both the wing and body lengths with the individuals coloured by species.

```
ggplot(finch, aes(BodyL, WingL)) + geom_point(aes(colour=Sp), size=6) +
  theme(legend.position = "bottom", legend.title=element_blank())
```

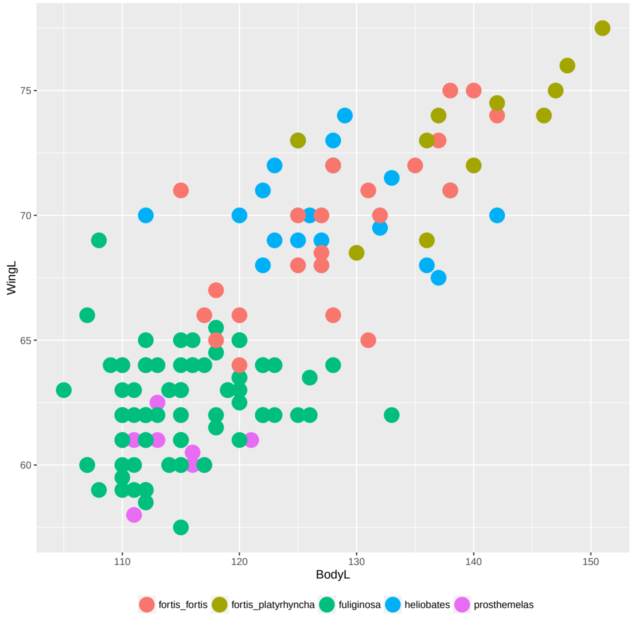

Fig 3: Wing length and body length for the five species: there is no clear separation between them.

In fact, a key feature differentiating the species is the beak. This can be seen in the next display, plotting all the features together.

```
library(GGally)
ggparcoord(finch, columns=2:10, groupColumn="Sp") + xlab("") + ylab("") +
  theme(legend.title=element_blank())
```

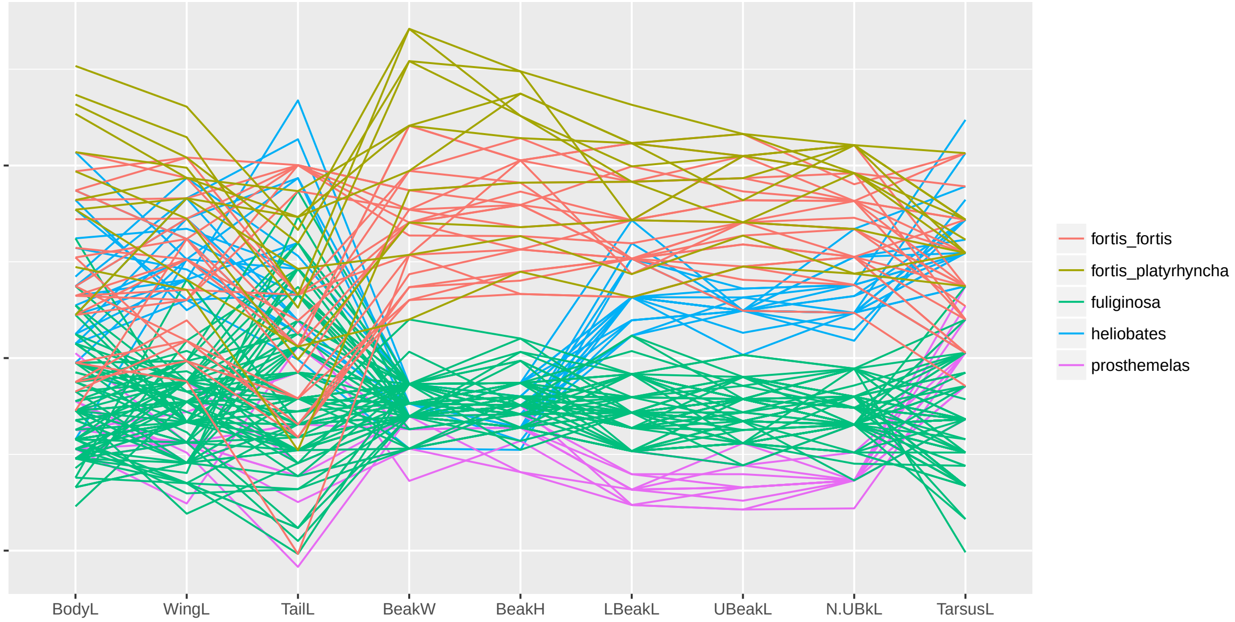

Fig 4: A parallel coordinate plot of the nine measurements made on each bird with the five species distinguished by colour. The first two beak variables (BeakW and BeakH) separate the two bigger species from the other three. The following three variables (LBeakL, UBeakL, and N.UBkL) separate the smaller species from one another.

Some of the variable names are self-explanatory (WingL must be wing length), others (e.g., N.UBkL, which turns out to be the distance from nostril to upper beak) are not so clear. The original reference or ornithologists can help.

The complete data from the 1904 reference for all the species (there are not just 5 taxa, but 32) for all the Galapogos islands, with the sex variable, and with the cases with missing values is available at the Dryad Digital Repository (http://datadryad.org/resource/doi:10.5061/dryad.152).