Palmer Archipelago Penguins Data in the palmerpenguins R Package - An Alternative to Anderson's Irises

by Allison M. Horst, Alison Presmanes Hill, and Kristen B. Gorman

Abstract In 1935, Edgar Anderson collected size measurements for 150 flowers from three species of *Iris* on the Gaspé Peninsula in Quebec, Canada. Since then, Anderson's *Iris* observations have become a classic dataset in statistics, machine learning, and data science teaching materials. It is included in the base R **datasets** package as **iris**, making it easy for users to access without knowing much about it. However, the lack of data documentation, presence of non-intuitive variables (e.g. "sepal width"), and perfectly balanced groups with zero missing values make **iris** an inadequate and stale dataset for teaching and learning modern data science skills. Users would benefit from working with a more representative, real-world environmental dataset with a clear link to current scientific research. Importantly, Anderson's *Iris* data appeared in a 1936 publication by R. A. Fisher in the *Annals of Eugenics* (which is often the first-listed citation for the dataset), inextricably linking **iris** to eugenics research. Thus, a modern alternative to **iris** is needed. In this paper, we introduce the **palmerpenguins** R package, which includes body size measurements collected from 2007 - 2009 for three species of *Pygoscelis* penguins that breed on islands throughout the Palmer Archipelago, Antarctica. The **penguins** dataset in **palmerpenguins** provides an approachable, charismatic, and near drop-in replacement for **iris** with topical relevance for polar climate change and environmental impacts on marine predators. Since the release on CRAN in July 2020, the **palmerpenguins** package has been downloaded over 402,000 times, highlighting the demand and widespread adoption of this viable **iris** alternative. We directly compare the **iris** and **penguins** datasets for selected analyses to demonstrate that R users, in particular teachers and learners currently using **iris**, can switch to the Palmer Archipelago penguins for many use cases including data wrangling, visualization, linear modeling, multivariate analysis (e.g., PCA), cluster analysis and classification (e.g., by k-means).

Introduction

In 1935, American botanist Edgar Anderson measured petal and sepal structural dimensions (length and width) for 50 flowers from three *Iris* species: *Iris setosa*, *Iris versicolor*, and *Iris virginica* (**?**). The manageable but non-trivial size (5 variables and 150 total observations) and characteristics of Anderson's *Iris* dataset (hereinafter referred to as **iris**), including linear relationships and multivariate normality, have made it amenable for introducing a wide range of statistical methods including data wrangling, visualization, linear modeling, multivariate analyses, and machine learning. The **iris** dataset is built into a number of software packages including the auto-installed datasets package in R (**?**), Python's scikit-learn machine learning library (**?**), and the SAS Sashelp library (SAS Institute, Cary NC), which has facilitated its widespread use. As a result, eighty-six years after the data were initially published, **iris** remains ubiquitous in statistics, computational methods, software documentation, and data science courses and materials.

There are a number of reasons that modern data science practitioners and educators may want to move on from **iris**. First, the dataset lacks metadata (**?**), which does not reinforce best practices and limits meaningful interpretation and discussion of research methods, analyses, and outcomes. Of the five variables in **iris**, two (*Sepal.Width* and *Sepal.Length*) are not intuitive for most non-botanists. Even with explanation, the difference between *petal* and *sepal* dimensions is not obvious. Second, **iris** contains equal sample sizes for each of the three species (n = 50) with no missing values, which is cleaner than most real-world data that learners are likely to encounter. Third, the single factor (*Species*) in **iris** limits options for analyses. Finally, due to its publication in the *Annals of Eugenics* by statistician R.A. Fisher (**?**), **iris** is burdened by a history in eugenics research, which we are committed to addressing through the development of new data science education products as described below.

Given the growing need for fresh data science-ready datasets, we sought to identify an alternative dataset that could be made easily accessible for a broad audience. After evaluating the positive and negative features of **iris** in data science and statistics materials, we established the following criteria for a suitable alternative:

- Available by appropriate license (ideally, CC0 "no rights reserved")
- Feature intuitive subjects and variables that are interesting and understandable to learners across disciplines
- Complete metadata and documentation

- Manageable (but not trivial) in size
- Minimal data cleaning and pre-processing required for most analyses
- Real-world (not manufactured) modern data
- Provides similar opportunities for teaching and learning R, data science, and statistical skills
- Can easily replace **iris** for most use cases

Here, we describe an alternative to **iris** that largely satisfies these criteria: a refreshing, approachable, and charismatic dataset containing real-world body size measurements for three *Pygoscelis* penguin species that breed throughout the Western Antarctic Peninsula region, made available through the United States Long-Term Ecological Research (US LTER) Network. By comparing data structure, size, and a range of analyses side-by-side for the two datasets, we demonstrate that the Palmer Archipelago penguin data are an ideal substitute for **iris** for many use cases in statistics and data science education.

Data source

Body size measurements (bill length and depth, flipper length - flippers are the modified "wings" of penguins used for maneuvering in water, and body mass), clutch (i.e., egg laying) observations (e.g., date of first egg laid, and clutch completion), and carbon (^{13}C/^{12}C, δ^{13}C) and nitrogen (^{15}N/^{14}N, δ^{15}N) stable isotope values of red blood cells for adult male and female Adélie (*P. adeliae*), chinstrap (*P. antarcticus*), and gentoo (*P. papua*) penguins on three islands (Biscoe, Dream and Torgersen) within the Palmer Archipelago were collected from 2007 - 2009 by Dr. Kristen Gorman in collaboration with the Palmer Station LTER, part of the US LTER Network. For complete data collection methods and published analyses, see **?**. Throughout this paper, penguins species are referred to as "Adélie", "Chinstrap", and "Gentoo".

The data in the **palmerpenguins** R package are available for use by CC0 license ("No Rights Reserved") in accordance with the Palmer Station LTER Data Policy and the LTER Data Access Policy, and were imported from the Environmental Data Initiative (EDI) Data Portal at the links below:

- Adélie penguin data (**?**): KNB-LTER Data Package 219.5
- Gentoo penguin data (**?**): KNB-LTER Data Package 220.5
- Chinstrap penguin data (**?**): KNB-LTER Data Package 221.6

R package: palmerpenguins

R users can install the **palmerpenguins** package from CRAN:

```
install.packages("palmerpenguins")
```

Information, examples, and links to community-contributed materials are available on the **palmerpenguins** package website: allisonhorst.github.io/palmerpenguins/. See the Appendix for how Python and Julia users can access the same data.

The **palmerpenguins** R package contains two data objects: **penguins_raw** and **penguins**. The **penguins_raw** data consists of all raw data for 17 variables, recorded completely or in part for 344 individual penguins, accessed directly from EDI (**penguins_raw** properties are summarized in Appendix B). We generally recommend using the curated data in **penguins**, which is a subset of **penguins_raw** retaining all 344 observations, minimally updated (Appendix A) and reduced to the following eight variables:

- *species:* a factor denoting the penguin species (Adélie, Chinstrap, or Gentoo)
- *island:* a factor denoting the Palmer Archipelago island in Antarctica where each penguin was observed (Biscoe Point, Dream Island, or Torgersen Island)
- *bill_length_mm:* a number denoting length of the dorsal ridge of a penguin bill (millimeters)
- *bill_depth_mm:* a number denoting the depth of a penguin bill (millimeters)
- *flipper_length_mm:* an integer denoting the length of a penguin flipper (millimeters)
- *body_mass_g:* an integer denoting the weight of a penguin's body (grams)
- *sex:* a factor denoting the sex of a penguin sex (male, female) based on molecular data
- *year:* an integer denoting the year of study (2007, 2008, or 2009)

The same data exist as comma-separated value (CSV) files in the package ("penguins_raw.csv" and "penguins.csv"), and can be read in using the built-in `path_to_file()` function in **palmerpenguins**. For example,

```
library(palmerpenguins)
df <- read.csv(path_to_file("penguins.csv"))
```

Table 1: Overview comparison of **penguins** and **iris** dataset features and characteristics.

Feature	iris	penguins
Year(s) collected	1935	2007 - 2009
Dimensions (col x row)	5 x 150	8 x 344
Documentation	minimal	complete metadata
Variable classes	double (4), factor (1)	double (2), int (3), factor (3)
Missing values?	no (n = 0; 0.0%)	yes (n = 19; 0.7%)

Table 2: Grouped sample size for **iris** (by species; n = 150 total) and **penguins** (by species and sex; n = 344 total). Data in **penguins** can be further grouped by island and study year.

iris sample size (by species)		penguins sample size (by species and sex)			
Iris species	Sample size	Penguin species	Female	Male	NA
setosa	50	Adélie	73	73	6
versicolor	50	Chinstrap	34	34	0
virginica	50	Gentoo	58	61	5

will read in "penguins.csv" as if from an external file, thus automatically parsing *species*, *island*, and *sex* variables as characters instead of factors. This option allows users opportunities to practice or demonstrate reading in data from a CSV, then updating variable class (e.g., characters to factors).

Comparing iris and penguins

The **penguins** data in **palmerpenguins** is useful and approachable for data science and statistics education, and is uniquely well-suited to replace the **iris** dataset. Comparisons presented are selected examples for common **iris** uses, and are not exhaustive.

Data structure and sample size

Both **iris** and **penguins** are in tidy format (**?**) with each column denoting a single variable and each row containing measurements for a single iris flower or penguin, respectively. The two datasets are comparable in size: dimensions (columns × rows) are 5 × 150 and 8 × 344 for **iris** and **penguins**, respectively, and sample sizes within species are similar (Tables 1 & 2).

Notably, while sample sizes in **iris** across species are all the same, sample sizes in **penguins** differ across the three species. The inclusion of three factor variables in **penguins** (*species*, *island*, and *sex*), along with *year*, create additional opportunities for grouping, faceting, and analysis compared to the single factor (*Species*) in **iris**.

Unlike **iris**, which contains only complete cases, the **penguins** dataset contains a small number of missing values (n_{missing} = 19, out of 2,752 total values). Missing values and unequal sample sizes are common in real-world data, and create added learning opportunity to the **penguins** dataset.

Continuous quantitative variables

Distributions, relationships between variables, and clustering can be visually explored between species for the four structural size measurements in **penguins** (flipper length, body mass, bill length and depth; Figure 1) and **iris** (sepal width and length, petal width and length; Figure 2).

Both **penguins** and **iris** offer numerous opportunities to explore linear relationships and correlations, within and across species (Figures 1 & 2). A bivariate scatterplot made with the **iris** dataset reveals a clear linear relationship between petal length and petal width. Using **penguins** (Figure 3), we can create a uniquely similar scatterplot with flipper length and body mass. The overall trend across all three species is approximately linear for both **iris** and **penguins**. Teachers may encourage students to explore how simple linear regression results and predictions differ when the species variable is omitted, compared to, for example, multiple linear regression with species included (Figure 3).

Notably, distinctions between species are clearer for iris petals - particularly, the much smaller petals for *Iris setosa* - compared to penguins, in which Adélie and Chinstrap penguins are largely

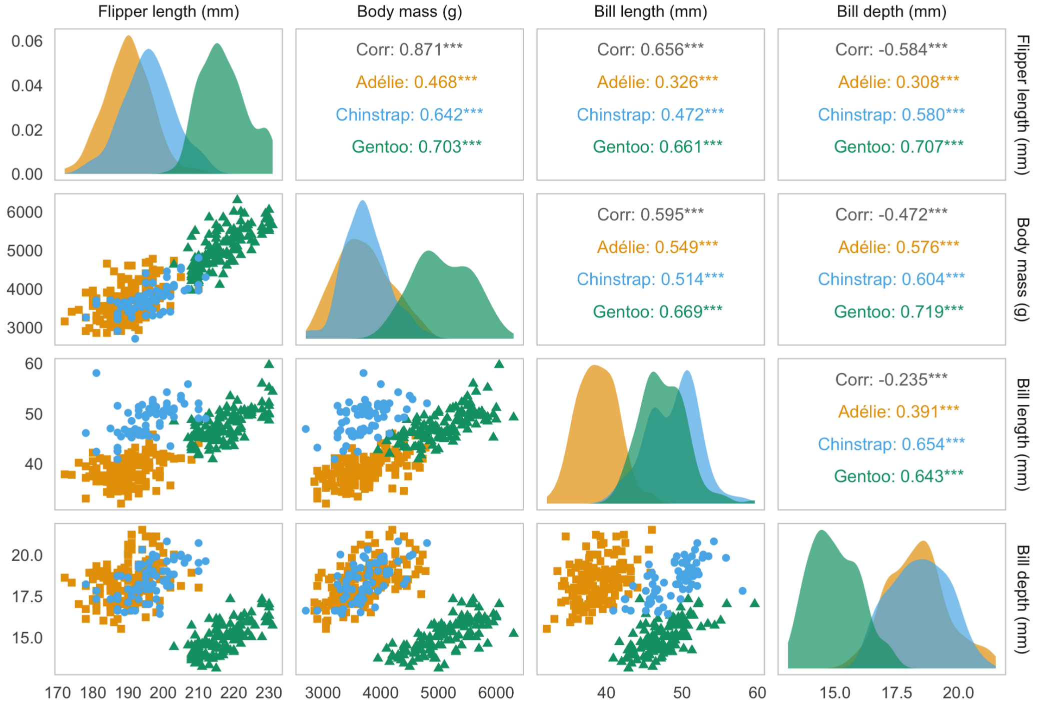

Figure 1: Distributions and correlations for numeric variables in the **penguins** data (flipper length (mm), body mass (g), bill length (mm) and bill depth (mm)) for the three observed species: Gentoo (green, triangles); Chinstrap (blue, circles); and Adélie (orange, squares). Significance indicated for bivariate correlations: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

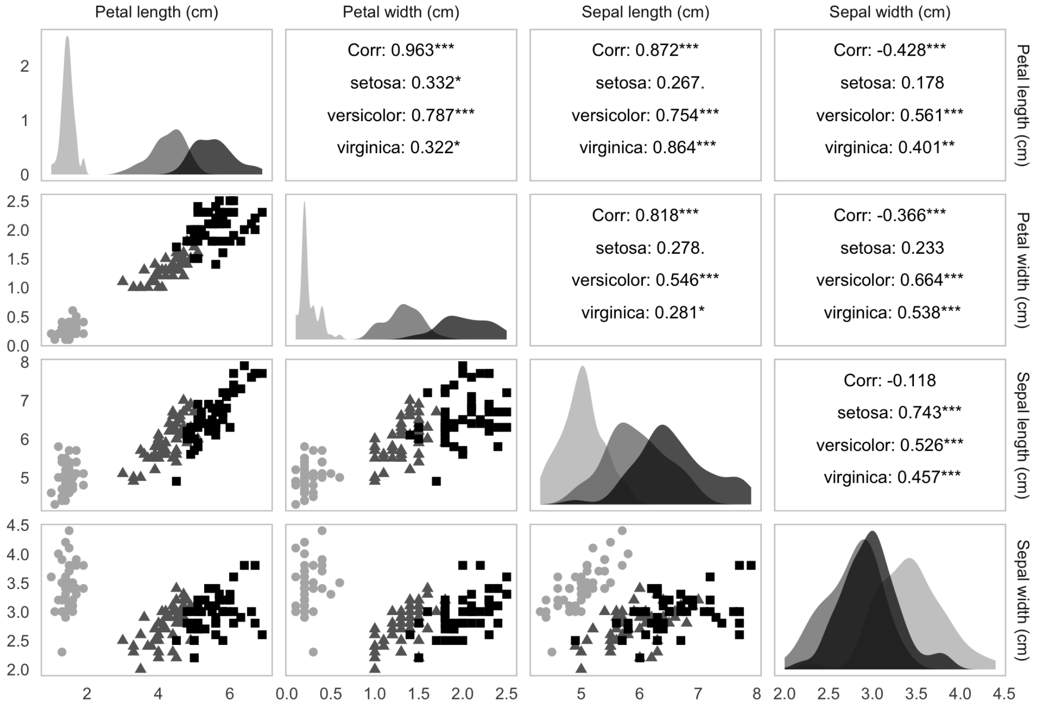

Figure 2: Distributions and correlations for numeric variables in **iris** (petal length (cm), petal width (cm), sepal length (cm) and sepal width (cm)) for the three included iris species: *Iris setosa* (light gray, circles); *Iris versicolor* (dark gray, triangles); and *Iris virginica* (black, squares). Significance indicated for bivariate correlations: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

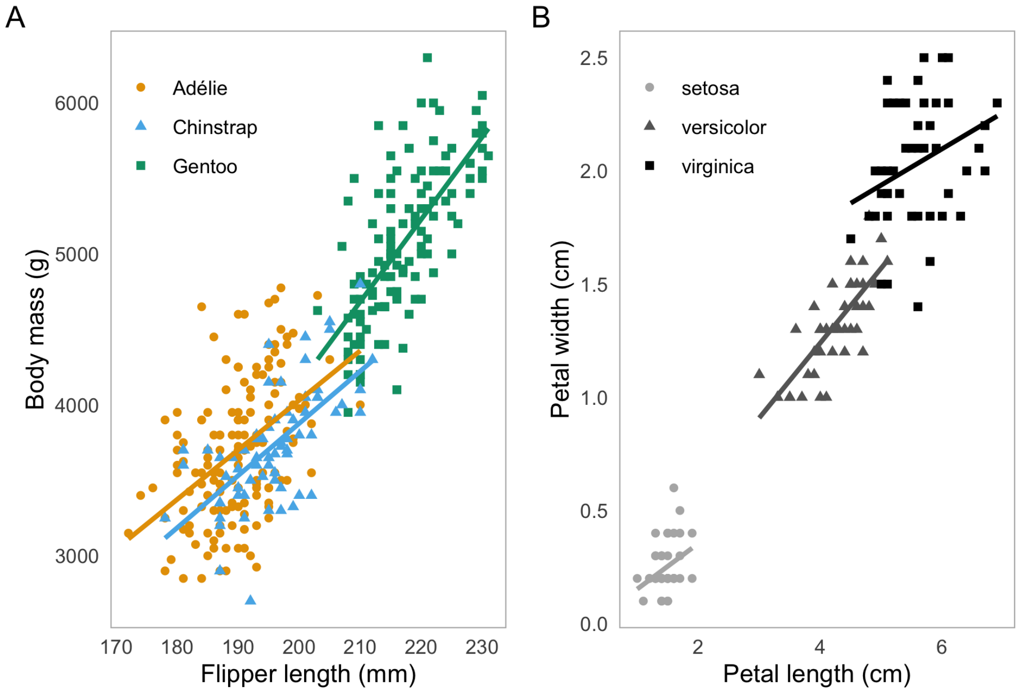

Figure 3: Representative linear relationships for (A): penguin flipper length (mm) and body mass (g) for Adélie (orange circles), Chinstrap (blue triangles), and Gentoo (green squares) penguins; (B): iris petal length (cm) and width (cm) for *Iris setosa* (light gray circles), *Iris versicolor* (dark gray triangles) and *Iris virginica* (black squares). Within-species linear model is visualized for each penguin or iris species.

overlapping in body size (body mass and flipper length), and are both generally smaller than Gentoo penguins.

Simpson's Paradox is a data phenomenon in which a trend observed between variables is reversed when data are pooled, omitting a meaningful variable. While often taught and discussed in statistics courses, finding a real-world and approachable example of Simpson's Paradox can be a challenge. Here, we show one (of several possible - see Figure 1) Simpson's Paradox example in **penguins**: exploring bill dimensions with and without species included (Figure 4). When penguin species is omitted (Figure 4A), bill length and depth appear negatively correlated overall. The trend is reversed when species is included, revealing an obviously positive correlation between bill length and bill depth within species (Figure 4B).

Principal component analysis

Principal component analysis (PCA) is a dimensional reduction method commonly used to explore patterns in multivariate data. The **iris** dataset frequently appears in PCA tutorials due to multivariate normality and clear interpretation of variable loadings and clustering.

A comparison of PCA with the four variables of structural size measurements in **penguins** and **iris** (both normalized prior to PCA) reveals highly similar results (Figure 5). For both datasets, one species is distinct (Gentoo penguins, and *setosa* irises) while the other two species (Chinstrap/Adélie and *versicolor*/*virginica*) appear somewhat overlapping in the first two principal components (Figure 5 A,B). Screeplots reveal that the variance explained by each principal component (PC) is very similar across the two datasets, particularly for PC1 and PC2: for **penguins**, 88.15% of total variance is captured by the first two PCs, compared to 95.81% for **iris**, with a similarly large percentage of variance captured by PC1 and PC2 in each (Figure 5 C,D).

K-means clustering

Unsupervised clustering by k-means is a common and popular entryway to machine learning and classification, and again, the **iris** dataset is frequently used in introductory examples. The **penguins**

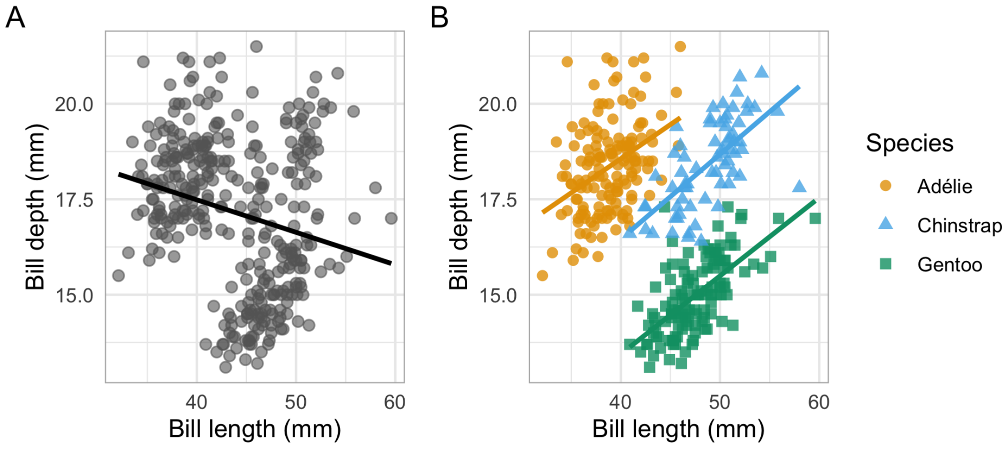

Figure 4: Trends for penguin bill dimensions (bill length and bill depth, millimeters) if the species variable is excluded (A) or included (B), illustrating Simpson's Paradox. Note: linear regression for bill dimensions without including species in (A) is ill-advised; the linear trendline is only included to visualize trend reversal for Simpson's Paradox when compared to (B).

Table 3: K-means cluster assignments by species based on penguin bill length (mm) and depth (mm), and iris petal length (cm) and width (cm).

Penguins cluster assignments				Iris cluster assignments			
Cluster	Adélie	Chinstrap	Gentoo	Cluster	setosa	versicolor	virginica
1	0	9	116	1	0	2	46
2	4	54	6	2	0	48	4
3	147	5	1	3	50	0	0

data provides similar opportunities for introducing k-means clustering. For simplicity, we compare k-means clustering using only two variables for each dataset: for **iris**, petal width and petal length, and for **penguins**, bill length and bill depth. All variables are scaled prior to k-means. Three clusters ($k = 3$) are specified for each, since there are three species of irises (*Iris setosa*, *Iris versicolor*, and *Iris virginica*) and penguins (Adélie, Chinstrap and Gentoo).

K-means clustering with penguin bill dimensions and iris petal dimensions yields largely distinct clusters, each dominated by one species (Figure 6). For iris petal dimensions, k-means yields a perfectly separated cluster (Cluster 3) containing all 50 *Iris setosa* observations and zero misclassified *Iris virginica* or *Iris versicolor* (Table 3). While clustering is not perfectly distinct for any penguin species, each species is largely contained within a single cluster, with little overlap from the other two species. For example, considering Adélie penguins (orange observations in Figure 6A): 147 (out of 151) Adélie penguins are assigned to Cluster 3, zero are assigned to Cluster 1, and 4 are assigned to the Chinstrap-dominated Cluster 2 (Table 3). Only 5 (of 68) Chinstrap penguins and 1 (of 123) Gentoo penguins are assigned to the Adélie-dominated Cluster 3 (Table 3).

Conclusion

Here, we have shown that structural size measurements for Palmer Archipelago *Pygoscelis* penguins, available as **penguins** in the **palmerpenguins** R package, offer a near drop-in replacement for **iris** in a number of common use cases for data science and statistics education including exploratory data visualization, linear correlation and regression, PCA, and clustering by k-means. In addition, teaching and learning opportunities in **penguins** are increased due to a greater number of variables, missing values, unequal sample sizes, and Simpson's Paradox examples. Importantly, the **penguins** dataset encompasses real-world information derived from several charismatic marine predator species with regional breeding populations notably responding to environmental change occurring throughout the Western Antarctic Peninsula region of the Southern Ocean (see **?**, **?**, **?**, **?**). Thus, the **penguins** dataset can facilitate discussions more broadly on biodiversity responses to global change - a contemporary and critical topic in ecology, evolution, and the environmental sciences.

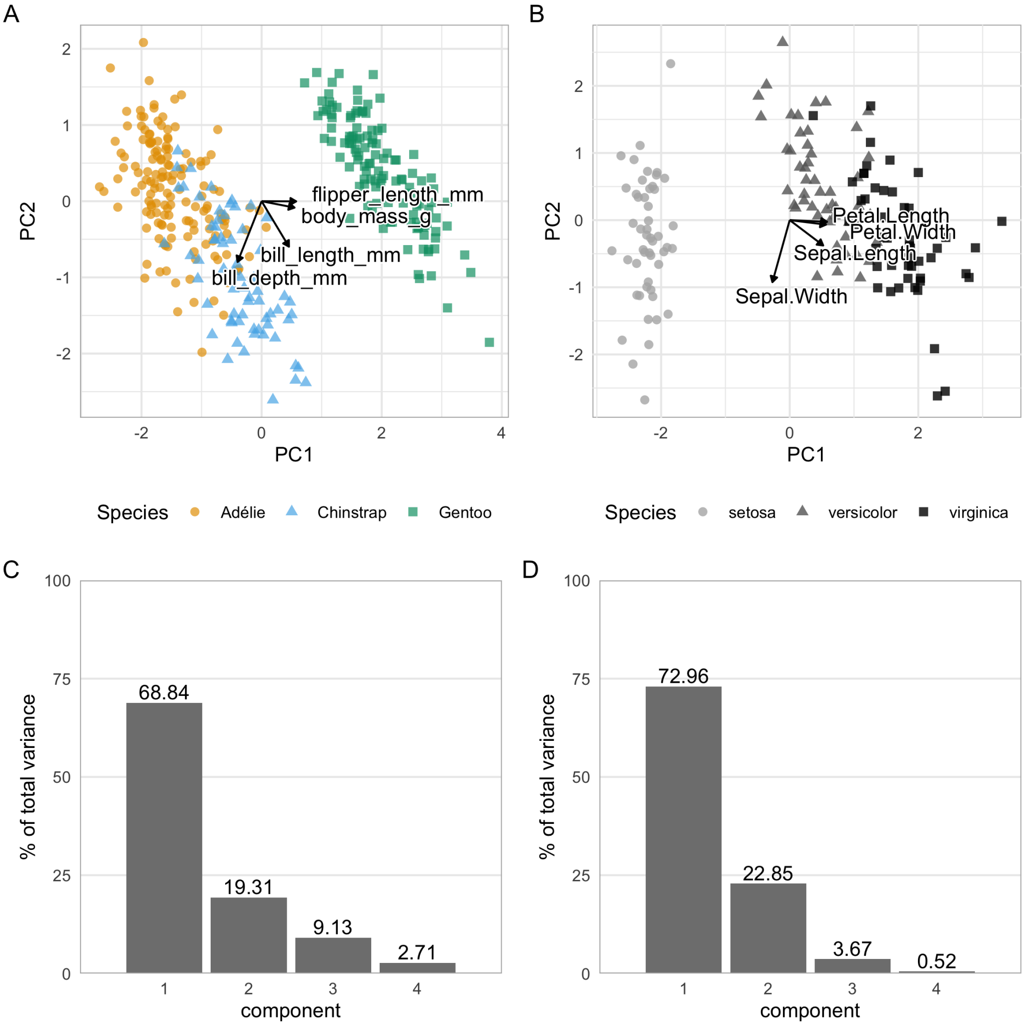

Figure 5: Principal component analysis biplots and screeplots for structural size measurements in **penguins** (A,C) and **iris** (B,D), revealing similarities in multivariate patterns, variable loadings, and variance explained by each component. For **penguins**, variables are flipper length (mm), body mass (g), bill length (mm) and bill depth (mm); groups are visualized by species (Adélie = orange circles, Chinstrap = blue triangles, Gentoo = green squares). For **iris**, variables are petal length (cm), petal width (cm), sepal length (cm) and sepal width (cm); groups are visualized by species (*Iris setosa* = light gray circles, *Iris versicolor* = dark gray triangles, *Iris virginica* = black squares). Values above screeplot columns (C,D) indicate percent of total variance explained by each of the four principal components.

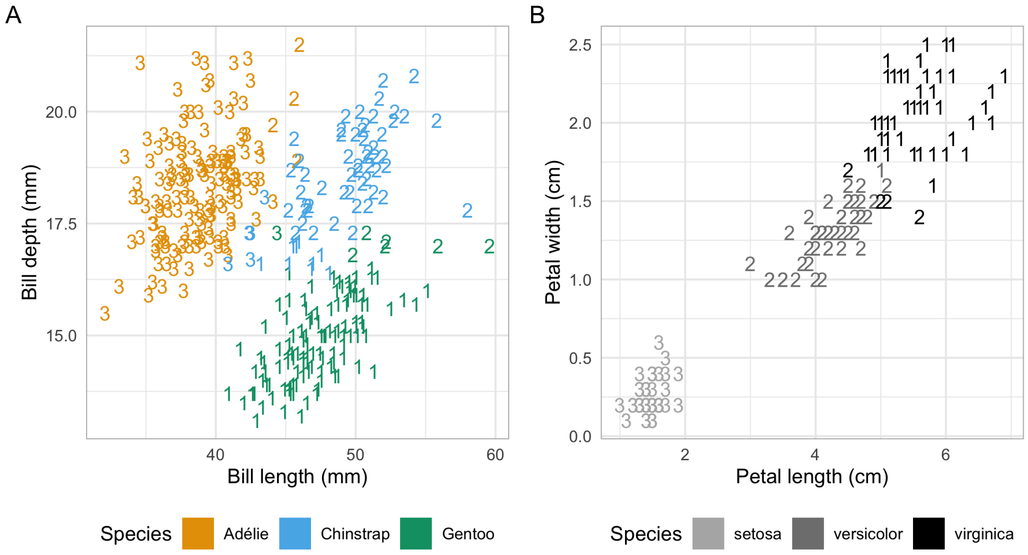

Figure 6: K-means clustering outcomes for penguin bill dimensions (A) and iris petal dimensions (B). Numbers indicate the cluster to which an observation was assigned, revealing a high degree of separation between species for both **penguins** and **iris**.

Appendix

Penguins data processing

Data in the **penguins** object have been minimally updated from **penguins_raw** as follows:

- All variable names are converted to lower snake case (e.g. from *Flipper Length (mm)* to *flipper_length_mm*)
- Entries in *species* are truncated to only include the common name (e.g. "Gentoo", instead of "gentoo penguin (*Pygoscelis papua*)")
- Recorded sex for penguin N36A1, originally recorded as ".", is updated to `NA`
- *culmen_length_mm* and *culmen_depth_mm* variable names are updated to *bill_length_mm* and *bill_depth_mm*, respectively
- Class for categorical variables (*species*, *island*, *sex*) is updated to factor
- Variable *year* was pulled from clutch observations

Summary of the penguins_raw dataset

Feature	penguins_raw
Year(s) collected	2007 - 2009
Dimensions (col x row)	17 x 344
Documentation	complete metadata
Variable classes	character (9), Date (1), numeric (7)
Missing values?	yes (n = 336; 5.7%)

palmerpenguins for other programming languages

Python: Python users can load the palmerpenguins datasets into their Python environment using the following code to install and access data in the palmerpenguins Python package:

```
pip install palmerpenguins
from palmerpenguins import load_penguins
penguins = load_penguins()
```

Julia: Julia users can access the penguins data in the **PalmerPenguins.jl** package. Example code to import the penguins data through **PalmerPenguins.jl** (more information on **PalmerPenguins.jl** from David Widmann can be found here):

```
julia> using PalmerPenguins
julia> table = PalmerPenguins.load()
```

TensorFlow: TensorFlow users can access the penguins data in TensorFlow Datasets. Information and examples for **penguins** data in TensorFlow can be found here.

Acknowledgements

All analyses were performed in the R language environment using version 3.6.2 (**?**). Complete code for this paper is shared in the Supplemental Material. We acknowledge the following R packages used in analyses, with gratitude to developers and contributors:

- **tidyverse** (**?**): for data import and cleaning
- **ggplot2** (**?**): for data visualizations
- **kableExtra** (**?**): for finalized tables
- **GGally** (**?**): for pairs plots
- **patchwork** (**?**): for compound figures
- **shadowtext** (**?**): to add a background color to text labels
- **recipes** (**?**): for data pre-processing
- **base** and **stats** (**?**): for various analyses throughout

Allison M. Horst
University of California Santa Barbara
Bren School of Environmental Science and Management
Santa Barbara, CA 93106-5131
ahorst@ucsb.edu

Alison Presmanes Hill
AI Strategy & Innovation, IBM

apreshill@gmail.com

Kristen B. Gorman
University of Alaska Fairbanks
College of Fisheries and Ocean Sciences
2150 Koyukuk Drive
245 O'Neill Building
Fairbanks, AK 99775-7220
kbgorman@alaska.edu