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

by Allison M. Horst, Alison P. Hill, Kristen B. Gorman

Abstract The famous *Iris* dataset, which exists in the base R **datasets** package as **iris**, is ubiquitous in statistics and data science education. While the **iris** measurements were originally collected by botanist Edgar Anderson, their use in Ronald A. Fisher's eugenics research has created momentum to address the dataset's racist use and/or replace **iris** with alternate data in teaching materials. Penguin morphological measurements for three species of *Pygoscelis* penguins that breed on islands throughout the Palmer Archipelago, Antarctica, provide an approachable, charismatic and near drop-in replacement for the **iris** data. Here, we introduce the **palmerpenguins** R package and describe the included **penguins** data. We directly compare the two datasets for selected analyses to demonstrate that R users, in particular teachers and learners currently using the **iris** data, can switch to the Palmer Archipelago penguins for many use cases including data wrangling and visualization, correlation, regression, hypothesis testing, 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* (Anderson, 1935). The manageable but non-trivial size (5 variables and 150 total observations) and characteristics of the data make it amenable for introducing a wide range of statistical methods including data wrangling, visualization, regression, multivariate exploration, and machine learning. Anderson's *Iris* dataset is built into a number of software packages including the auto-installed datasets package in R (R Core Team, 2019), Python's scikit-learn machine learning library (Pedregosa et al., 2011), and the SAS Sashelp library (SAS Institute, Cary NC), which have facilitated widespread use of the data. As a result, eighty-five years after the data were initially published, Anderson's *Iris* measurements are ubiquitous in statistics and data science course materials and tutorials. **Sounds great. Why consider replacing iris?**

In 1936, one year after being described in the Bulletin of the American Iris Society, Anderson's *Iris* measurements were published in full by eugenicist and statistician Ronald A. Fisher in *Annals of Eugenics* (Fisher, 1936). Regardless of Anderson's initial motivation, the data today remain inextricably linked to Fisher's eugenics research and are even commonly, if unfairly, referred to as "Fisher's iris data." For example, an August 2020 search for "Iris flower data set" in Wikipedia returns an article beginning with "The iris flower data set or Fisher's iris data…", then later in the article: "It is sometimes called Anderson's Iris data set" (Wikipedia, 2020). The *Iris* dataset is similarly credited to Fisher in statistical computing literature (Trendafilov and Vines, 2009, Wang (2015), Woods et al. (2015), Chen et al. (2018)).

There is recent momentum to address Fisher's racist legacy in statistics. In June 2020, the Committee of Presidents of Statistical Societies' R.A. Fisher Award and Lectureship was replaced by the Distinguished Achievement Award and Lectureship, to "advance a more just, equitable, diverse, and inclusive statistical community" (noa, 2020a). Cambridge University's Gonville and Caius College recently announced plans to remove a stained-glass window celebrating Fisher from a campus dining hall (noa, 2020b). In the R community, there are calls to (1) address the *Iris* dataset's use in Fisher's eugenics work, and/or (2) consider using a different dataset (Poisot, 2020, Aden-Buie (2020)). This paper describes (2): a suitable alternative dataset for educators wanting to replace the *Iris* data in their teaching materials.

Considering *Iris* data usage in data science and statistics materials, we established the following criteria for a suitable replacement dataset:

- Available by appropriate license
- Features intuitive subjects and variables understandable to learners across fields
- Manageable (but not trivial) in size
- Minimal data cleaning and pre-processing required for most analyses
- Real-world (not manufactured) data
- Similar opportunities for teaching and learning R, data science, and statistical skills

Here, we describe an alternative to Anderson's *Iris* data that largely satisfies these criteria: an

approachable and charismatic dataset containing real-world morphological data for three *Pygoscelis* penguin species that breed throughout the Western Antarctic Peninsula region, made available through the Long-Term Ecological Research Network (US LTER). By comparing data structure, size, and a range of analyses side-by-side for the two datasets, we demonstrate that the Palmer Archipelago penguin measurements can replace Anderson's *Iris* data for many use cases in statistics and data science education.

Data source

Body size measurements, clutch (i.e., egg laying) observations (e.g., date of first egg laid, and clutch completion), and carbon ($^{13}C/^{12}C$, $\delta^{13}C$) and nitrogen ($^{15}N/^{14}N$, $\delta^{15}N$) stable isotope values of red blood cells for male and female adult Adélie (*P. adeliae*), chinstrap (*P. antarcticus*), and gentoo (*P. papua*) penguins on three islands (Biscoe, Dream and Torgersen) in 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 Gorman et al. (2014).

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 (LTER and Gorman 2020a): KNB-LTER Data Package 219.5
- Chinstrap penguin data (LTER and Gorman 2020b): KNB-LTER Data Package 221.5
- Gentoo penguin data (LTER and Gorman 2020c): KNB-LTER Data Package 220.5

R package: palmerpenguins

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

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

Alternatively, the development version of the package can be installed from GitHub:

```
`remotes::install_github("allisonhorst/palmerpenguins")`
```

Information, examples, and links to community-contributed materials are available on the **palmerpenguins** package website: `https://allisonhorst.github.io/palmerpenguins/`.

The **palmerpenguins** R package contains two data objects: **penguins_raw** and **penguins**. The **penguins_raw** data consists of all raw data for 17 variables (Appendix Table 1), recorded completely or in part for 344 individual penguins, accessed directly from EDI. As a direct alternative to Anderson's *Iris* data we recommend using the curated data in **penguins**, which is a subset of **penguins_raw** retaining all 344 observations, minimally updated (Appendix B) and reduced to the following eight variables:

- *species:* a factor denoting the penguin species (Adélie, chinstrap, or gentoo)
- *island:* a factor denoting the island (in Palmer Archipelago, Antarctica) where observed (Biscoe, Dream or Torgersen)
- *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(tidyverse)
library(palmerpenguins)
df <- read_csv(path_to_file("penguins.csv"))
```

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

Table 1: Grouped sample size for iris (by species; n = 150 total) and penguins (by species and sex; n = 344 total). Penguins can be further grouped by variables for island and study year.

Iris sample size (by species)		Penguin 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

Other data access options

Python: Python users can access the penguins data in the **seaborn** data visualization library (Waskom et al. 2017). Example code to load the data in Python:

```
import seaborn as sns
df = sns.load_dataset('penguins')
```

Julia: Julia users can access the penguins data in the **PalmerPenguins.jl** package. Example code to import the penguins data through **PalmerPenguins.jl**:

```
julia> using DataFrames
julia> df = DataFrame(table)
```

Software and code

All analyses were performed in the R language environment using version 3.6.2 (R Core Team 2019). 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** (Wickham et al. 2019), for data import and cleaning
- **ggplot2** (Wickham 2016:2), for data visualizations
- **here** (Müller 2017), for file path control
- **kableExtra** (Zhu 2019) for finalized tables
- **gt** (Iannone, Cheng, and Schloerke 2020) for finalized tables
- **GGally** (Schloerke et al. 2020), for pairs plots
- **patchwork** (Pedersen 2019), for compound figures
- **shadowtext** (Yu 2019), to add a background color to text labels
- **recipes** (Kuhn and Wickham 2020), for data pre-processing
- **base** and **stats** (R Core Team 2019) for various analyses throughout
- **pkgdown** (ref) was used to build the package website

Selected comparisons between iris and penguins

The **penguins** data in **palmerpenguins** is generally 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 [ref] with each column denoting a single variable and each row containing measurements for a single *Iris* flower or penguin. 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 (Table 1). Notably, sample sizes differ for the three penguin species, while sample sizes in **iris** are equal (n = 50 for each Iris species). Multiple factor variables in **penguins** (*species*, *island* and *sex*) along with *year* create additional opportunities for grouping, 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 we believe add learning value to **penguins**.

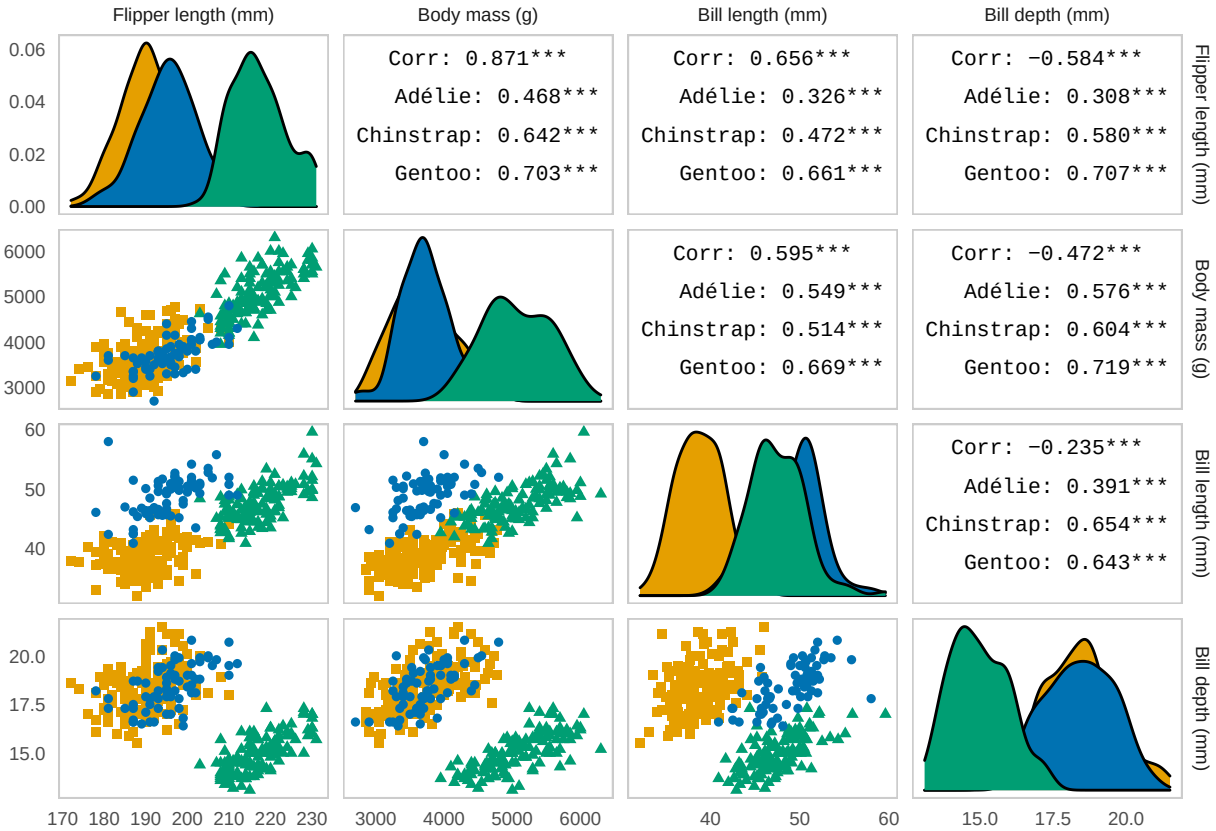

Figure 1: Distribution and correlations for numeric variables in the penguins data (flipper length (mm), body mass, (g) bill length (mm) and bill depth (mm)) for the three observed species: gentoo (green, triangles); chinstrap (blue, circles); and Adélie (orange, squares). Correlations are Pearson's r (*p < 0.05; **p < 0.01; ***p < 0.001).

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).

There exist numerous opportunities to explore linear relationships and correlations in **penguins** and **iris** both within and across species (Figures 1 & 2). Here, we highlight one comparison that is uniquely similar: petal dimensions (petal length versus petal width) from iris, and penguin size (flipper length versus body mass) in penguins (Figure 3). The overall trend across all three species is approximately linear for both iris and penguins (Figure 3A,B). Teachers may encourage students to explore how simple linear regression results and predictions differ when the species variable is omitted (Figure 3A,B), compared to multiple linear regression with species included (Figure 3C,D).

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

Penguins bonus: Simpson's Paradox

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 XX) Simpson's Paradox examples in **penguins**: exploring bill dimensions with and without *species* included (Figure XX). When penguin *species* is omitted (Figure XX), the bill length and depth appear negatively correlated overall. The trend is reversed when *species* is included, revealing a positive correlation between bill length and bill depth (Figure XX)

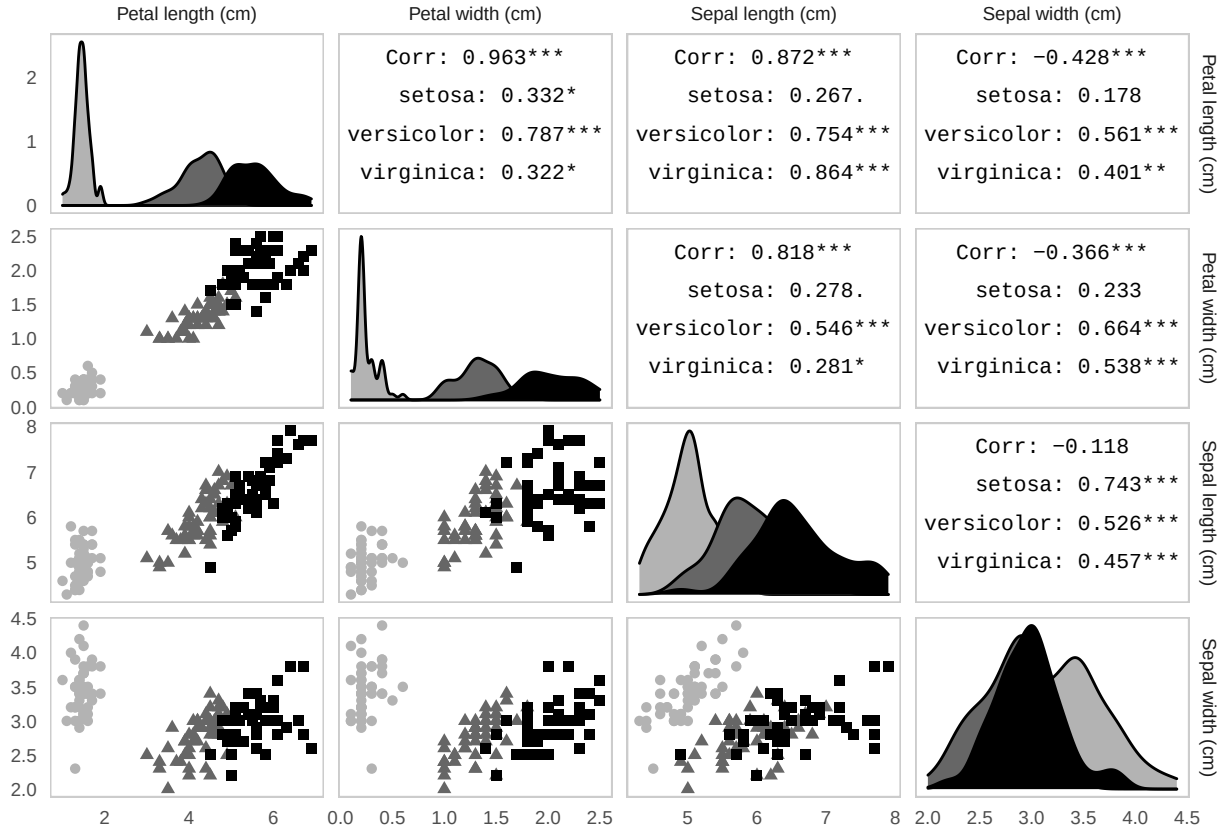

Figure 2: Distribution and correlations for numeric variables in the iris data (petal length (cm), petal width (cm), sepal length (cm) and sepal width (cm)) for the three included Iris species: setosa (light gray, circles); versicolor (dark gray, triangles); and virginica (black, squares). Correlations are Pearson's r (*p < 0.05; **p < 0.01; ***p < 0.001).

Principal component analysis

Principal component analysis (PCA) is a dimensional reduction method commonly used to explore patterns in multivariate data. PCA tutorials frequently employ the **iris** dataset, which is useful for teaching the method due to multivariate normality, and clear, approachable outcomes for 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 XX). 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 XX). Variance explained by each principal component (PC) is similar, 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 XX).

K-means clustering

Unsupervised clustering by k-means is a common and popular entryway to machine learning and classification, and frequently employs the **iris** data for introductory examples. The **penguins** 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 both *Iris* (*setosa*, *versicolor*, and *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 XX, Table XX). For iris petal dimensions, k-means yields a perfectly separated cluster (Cluster 1) containing all 50 setosa iris observations and zero misclassified virginica or versicolor irises (Table 4). 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 1, zero are assigned to Cluster 2, and 4 are assigned to the chinstrap-dominated Cluster 3 (Table 4). Only 5 (of 68) chinstrap penguins and 1 (of 123) gentoo penguins are assigned to the Adélie-dominated Cluster 1 (Table 4).

Other alternatives to iris

Discussion

Educators have three options regarding **iris** use in course materials:

1. Use **iris** without addressing its use in Fisher's eugenics research
2. Use **iris** *and* address its use in Fisher's eugenics research
3. Replace **iris** with an alternative dataset offering similar teaching and learning opportunities

Here, we have shown that structural size measurements for Palmer Archipelago *Pygoscelis* penguins, available as **penguins** in the **palmerpenguins** R package, can replace **iris** for a number of common use cases in 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.

The **iris** dataset is widespread today because it was employed by Fisher to advance statistical methods for eugenics research. While the *Irises* were described in Anderson's 1935 article in *The Bulletin of the American Iris Society* (Anderson, 1935), the measurements were published in full in Fisher's 1936 *Annals of Eugenics* publication (Fisher, 1936) - a journal for which Fisher served as editor before becoming chair of the Department of Eugenics at University College London.

Conclusion

Appendix

Appendix A

Appendix B

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

- All variable names are converted to lower snake case
- 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

Bibliography

Institute of Mathematical Statistics, July 2020a. URL https://imstat.org/2020/07/16/copss-statement-on-r-a-fisher-award/. [p1]

Sir Ronald Fisher: Cambridge college window for eugenicist to be removed. *BBC News*, June 2020b. URL https://www.bbc.com/news/uk-england-cambridgeshire-53192576. [p1]

G. Aden-Buie. Let's move on from iris, June 2020. URL https://www.garrickadenbuie.com/blog/lets-move-on-from-iris/. [p1]

E. Anderson. The irises of the Gaspé Peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935. [p1, 6]

C.-C. Chen, H.-H. Juan, M.-Y. Tsai, and H. H.-S. Lu. Unsupervised Learning and Pattern Recognition of Biological Data Structures with Density Functional Theory and Machine Learning. *Scientific Reports*, 8(1):557, Dec. 2018. ISSN 2045-2322. doi: 10.1038/s41598-017-18931-5. URL http://www.nature.com/articles/s41598-017-18931-5. [p1]

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, Sept. 1936. ISSN 20501420. doi: 10.1111/j.1469-1809.1936.tb02137.x. URL http://doi.wiley.com/10.1111/j.1469-1809.1936.tb02137.x. [p1, 6]

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. [p1]

T. Poisot. It's time to retire the iris dataset, Jan. 2020. URL https://armchairecology.blog/iris-dataset/. [p1]

R Core Team. R: a language and environmental for statistical computing, 2019. URL https://www.R-project.org/. [p1]

N. T. Trendafilov and K. Vines. Simple and interpretable discrimination. *Computational Statistics & Data Analysis*, 53(4):979–989, Feb. 2009. ISSN 01679473. doi: 10.1016/j.csda.2008.11.018. URL https://linkinghub.elsevier.com/retrieve/pii/S0167947308005495. [p1]

C.-C. Wang. A MATLAB package for multivariate normality test. *Journal of Statistical Computation and Simulation*, 85(1):166–188, Jan. 2015. ISSN 0094-9655, 1563-5163. doi: 10.1080/00949655.2013.808638. URL http://www.tandfonline.com/doi/abs/10.1080/00949655.2013.808638. [p1]

Wikipedia. Iris flower data set, July 2020. URL https://en.wikipedia.org/wiki/Iris_flower_data_set. [p1]

R. P. Woods, L. K. Hansen, and S. Strother. How Many Separable Sources? Model Selection In Independent Components Analysis. *PLOS ONE*, 10(3):e0118877, Mar. 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0118877. URL https://dx.plos.org/10.1371/journal.pone.0118877. [p1]

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

Alison P. Hill
RStudio, PBC
250 Northern Ave
Boston, MA 02210
alison@rstudio.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